Bharat Bhushan
Editor

Encyclopedia of Nanotechnology

Second Edition

Springer Reference

Encyclopedia of Nanotechnology

Bharat Bhushan
Editor

Encyclopedia of Nanotechnology

Second Edition

With 2761 Figures and 168 Tables

Springer Reference

Editor
Bharat Bhushan
Nanoprobe Laboratory for
Bio- and Nanotechnology and Biomimetics
The Ohio State University
Columbus, OH, USA

ISBN 978-94-017-9779-5 ISBN 978-94-017-9780-1 (eBook)
ISBN 978-94-017-9781-8 (print and electronic bundle)
DOI 10.1007/978-94-017-9780-1

Library of Congress Control Number: 2009004646

© Springer Science+Business Media Dordrecht 2012, 2016
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media B.V.
The registered company address is: Van Godewijckstraat 30, 3311 GX Dordrecht, The Netherlands

Preface of Second Edition

On December 29, 1959, at the California Institute of Technology, Nobel Laureate Richard P. Feynman gave a talk at the Annual Meeting of the American Physical Society that has become one of the classic science lectures of the twentieth century. His talk "There's Plenty of Room at the Bottom" presented a technological vision of extreme miniaturization several years before the word "chip" became part of the lexicon. He talked about the problem of manipulating and controlling things on a small scale. Extrapolating from known physical laws, Feynman envisioned a technology using the ultimate toolbox of nature, building nanoobjects atom by atom or molecule by molecule. Since the 1980s, many inventions and discoveries in the fabrication of nanoobjects have been testaments to his vision.

In a January 2000 speech, also at the California Institute of Technology, President William J. Clinton talked about the exciting promise of nanotechnology and the importance of expanding research in nanoscale science and technology more broadly. Later that month in his State of the Union Address, he announced an ambitious $497 million federal, multiagency National Nanotechnology Initiative (NNI) in the FY 2001 budget and made the NNI a top science and technology priority (Anonymous, 2000). The objective of this initiative was to form a broad-based coalition in which academia, the private sector, and local, state, and federal governments work together to push the envelope of nanoscience and nanoengineering to reap nanotechnology's potential social and economic benefits.

In January 2003, following the creation of the NNI, the US Senate in the 108th Congress introduced a bill to establish a National Nanotechnology Program. On December 3, 2003, President George W. Bush signed into law the 21st Century Nanotechnology Research and Development Act (NRDA) (Public Law 108–153). The legislation put into law programs and activities supported by the NNI. The bill gave nanotechnology a permanent home in the federal government and authorized $3.7 billion to be spent in the 4-year period beginning in October 2005. The funds were used to provide grants to researchers, coordinate R&D across six agencies, establish interdisciplinary research centers, and accelerate technology transfer into the private sector.

Regarding international activities in early stages, in 2002 the European Union (EU) made nanosciences and nanotechnologies a priority in the Sixth Framework Program (FP6) for the period 2003–2006. They had also dedicated small funds in the FP4 and FP5 programs before. FP6 was tailored to

help better structure European research and to cope with the strategic objectives set out in the Lisbon Strategy Development Plan of 2000. Japan identified nanotechnology as one of its main research priorities in 2001. The funding levels increased from $400 million in 2001 to around $950 million in 2004. In 2003, South Korea embarked on a 10-year program with around $2 billion in public funding, and Taiwan had committed around $600 million in public funding over 6 years. Russia, Singapore, and China also started to invest on a large scale.

Through FY 2015 in the USA, federal R&D investment has been about $20 billion, with annual investment in FY 2015 of about $1.5 billion, and more than double that by the private sector. The revenues from nano-enabled products continue growing, with over $200 billion in FY 2012 in the USA alone and over $700 billion worldwide. This represents an impressive return on investment.

Nanotechnology refers to any technology done on a nanoscale that has applications in the real world. It is defined as the control or restructuring of matter at the atomic and molecular levels in the size range of about 1–100 nm. The underlying science is referred to as nanoscience. The properties of matter at the nanoscale are different from those at a larger scale. When the dimensions of a material are reduced from a large size, the properties remain the same at first, then small changes occur. Finally, when the size drops below 100 nm, dramatic changes in properties can occur. The unique physical and chemical properties of nanomaterials can be exploited for commercial applications and for novel performance that benefits society. The discovery of novel materials, processes, and phenomena at the nanoscale and the development of new experimental and theoretical techniques for research at the end of the twentieth century provided fresh opportunities for the development of innovative nanosystems and nanomaterials. This field is opening new venues in science and technology.

Nanotechnology encompasses nanomanufacturing and application of physical, chemical, and biological systems at scales ranging from individual atoms or molecules to submicron dimensions, as well as the integration of the resulting nanostructures into larger systems. It spans across scientific fields, including chemistry, physics, material science, engineering, and manufacturing. Its impact on our society and economy in the twenty-first century is comparable to that of semiconductor technology, information technology, or cellular and molecular biology in the twentieth century. The integration of nanotechnology into larger systems has given breakthrough solutions to many current environmental, medical, and industrial problems, including smart materials, nanomanufacturing, electronics, drug delivery, energy and water, biotechnology, information technology, and national security. Nanotechnology will have a profound impact on our economy and society; it is a modern industrial revolution. Nanotechnology represents a megatrend, bringing disruptive innovation. It has become a general purpose technology, being applicable across various industrial sectors.

There has existed a need for a multidisciplinary, system-oriented approach to design and manufacturing of micro/nanodevices which function reliably. This can only be achieved through the cross-fertilization of ideas from

different disciplines and the systematic flow of information and people among research groups. A broad-based *Springer Handbook of Nanotechnology* was published with the first edition in April 2004, second edition in 2007, and the third edition in 2010. It presents an overview of nanomaterial synthesis, micro/nanofabrication, micro- and nanocomponents and systems, scanning probe microscopy, reliability issues (including nanotribology and nanomechanics) for nanotechnology, and various industrial applications including biomedical applications. A two-volume *Springer Handbook of Nanomaterials Properties* was launched in 2014.

A four-volume *Encyclopedia of Nanotechnology* was launched in 2012 with 325 essays written by experts in the field from academia, national research labs, and industry, and from all over the world. The objective of this encyclopedia was to introduce a large number of terms, devices, and processes. For each entry or essay, a brief description is provided by experts in the field. The entries were written by a large number of internationally recognized experts in the field, from academia, national research labs, and industry, and from all over the world. The *Encyclopedia of Nanotechnology* has provided a comprehensive and multidisciplinary reference to the many fields relevant to the general field of nanotechnology. The encyclopedia focuses on engineering and applications with some coverage of science of nanotechnology. It is a comprehensive and genuinely international reference work and is aimed at graduate students, researchers, and practitioners. Given the explosive growth in nanoscience and nanotechnology, we offer an updated and expanded second edition.

The development of the encyclopedia was undertaken by an Editorial Board, who were responsible for the focus and quality of contributions, and an Advisory Board, who were responsible for advising about the selection of topics.

The *Encyclopedia of Nanotechnology* is available both in a print version and an online version. The print version of the encyclopedia contains four volumes with a total of 426 entries and about 4944 pages. The online version of the encyclopedia is expected to be updated periodically. The editor-in-chief and all editors thank the large number of authors for making contributions to this major reference work. Thanks are also due to the referees who meticulously read the entries and made their recommendations.

Powell, Ohio **Bharat Bhushan**
August 2016

Preface of First Edition

On December 29, 1959, at the California Institute of Technology, Nobel Laureate Richard P. Feynman gave a speech at the Annual meeting of the American Physical Society that has become one of the twentieth-century classic science lectures, titled "There's Plenty of Room at the Bottom." He presented a technological vision of extreme miniaturization in 1959, several years before the word "chip" became part of the lexicon. He spoke about the problem of manipulating and controlling things on a small scale. Extrapolating from known physical laws, Feynman envisioned a technology using the ultimate toolbox of nature, building nanoobjects atom by atom or molecule by molecule. Since the 1980s, many inventions and discoveries in the fabrication of nanoobjects have been testament to his vision. In recognition of this reality, National Science and Technology Council (NSTC) of the White House created the Interagency Working Group on Nanoscience, Engineering and Technology (IWGN) in 1998. In a January 2000 speech at the same institute, former President W. J. Clinton spoke about the exciting promise of "nanotechnology" and the importance of expanding research in nanoscale science and technology, more broadly. Later that month, he announced in his State of the Union Address an ambitious $497 million federal, multiagency National Nanotechnology Initiative (NNI) in the fiscal year 2001 budget and made the NNI a top science and technology priority. The objective of this initiative was to form a broad-based coalition in which the academe, the private sector, and local, state, and federal governments work together to push the envelope of nanoscience and nanoengineering to reap nanotechnology's potential social and economic benefits.

The funding in the USA has continued to increase. In January 2003, the US Senate introduced a bill to establish a National Nanotechnology Program. On December 3, 2003, President George W. Bush signed into law the 21st Century Nanotechnology Research and Development Act. The legislation put into law programs and activities supported by the (NNI). The bill gave nanotechnology a permanent home in the federal government and authorized $3.7 billion to be spent in the 4-year period beginning in October 2005, for nanotechnology initiatives at five federal agencies. The funds have provided grants to researchers, coordinated R&D across five federal agencies (National Science Foundation (NSF), Department of Energy (DOE), NASA, National Institute of Standards and Technology (NIST), and Environmental Protection Agency (EPA)), established interdisciplinary research centers, and accelerated

technology transfer into the private sector. In addition, the Department of Defense (DOD), Homeland Security, Agriculture and Justice, as well as the National Institutes of Health (NIH) also fund large R&D activities. They currently account for more than one-third of the federal budget for nanotechnology.

The European Union (EU) made nanosciences and nanotechnologies a priority in the Sixth Framework Program (FP6) in 2002 for the period 2003–2006. They had dedicated small funds in FP4 and FP5 before. FP6 was tailored to help better structure European research and to cope with the strategic objectives set out in Lisbon in 2000. Japan identified nanotechnology as one of its main research priorities in 2001. The funding levels increased sharply from $400 million in 2001 to around $950 million in 2004. In 2003, South Korea embarked upon a 10-year program with around $2 billion of public funding, and Taiwan has committed around $600 million of public funding over 6 years. Singapore and China are also investing on a large scale. Russia is well funded as well.

Nanotechnology literally means any technology done on a nanoscale that has applications in the real world. Nanotechnology encompasses production and application of physical, chemical, and biological systems at scales, ranging from individual atoms or molecules to submicron dimensions, as well as the integration of the resulting nanostructures into larger systems. Nanotechnology is likely to have a profound impact on our economy and society in the early twenty-first century, comparable to that of semiconductor technology, information technology, or cellular and molecular biology. Science and technology research in nanotechnology is leading to breakthroughs in areas such as materials and manufacturing, nanoelectronics, medicine and healthcare, energy, biotechnology, information technology, and national security. It is widely felt that nanotechnology will be the next industrial revolution.

There is an increasing need for a multidisciplinary, system-oriented approach toward designing and manufacturing micro/nanodevices which function reliably. This can only be achieved through the cross-fertilization of ideas from different disciplines and the systematic flow of information and people among research groups. Reliability is a critical technology for many micro- and nanosystems and nanostructured materials. The first edition of a broad-based *Handbook of Nanotechnology* from Springer was published in April 2004, the second edition in 2007, and the third edition in 2010. It presents an overview of nanomaterial synthesis, micro/nanofabrication, micro- and nanocomponents and systems, scanning probe microscopy, reliability issues (including nanotribology and nanomechanics) for nanotechnology, and various industrial including biomedical applications.

The field of nanotechnology is getting a strong foothold. It attracts people from various disciplines including science and engineering. Given the explosive growth in nanoscience and nanotechnology, this *Encyclopedia of Nanotechnology* is being launched with essays written by experts in the field from academia and industry.

The objective of this encyclopedia is to introduce a large number of terms, devices, and processes. For each entry, a brief description is provided by experts in the field. The entries have been written by a large number of

internationally recognized experts in the field, from academia, national research labs, and industry.

The *Encyclopedia of Nanotechnology* is expected to provide a comprehensive and multidisciplinary reference to the many fields relevant to the general field of nanotechnology. The encyclopedia focuses on engineering and applications with some coverage of the science of nanotechnology. It aims to be a comprehensive and genuinely international reference work and is aimed at graduate students, researchers, and practitioners.

The development of the encyclopedia was undertaken by an Editorial Board, who were responsible for the focus and quality of contributions, and an Advisory Board, who were responsible for advising about the selection of topics.

The print version of the encyclopedia contains four volumes with a total of **325** entries and about **3068** pages. The encyclopedia is also available online. Editors expect to update it periodically. The editor-in-chief and all the editors thank a large number of authors for making contributions to this major reference work. We also thank the referees who meticulously read the entries and made their recommendations.

Powell, Ohio
USA
May 2012

Bharat Bhushan

About the Editor



Dr. Bharat Bhushan received an M.S. in Mechanical Engineering from the Massachusetts Institute of Technology in 1971; an M.S. in Mechanics and a Ph.D. in Mechanical Engineering from the University of Colorado at Boulder in 1973 and 1976, respectively; an MBA from Rensselaer Polytechnic Institute at Troy, NY, in 1980; Doctor Technicae from the University of Trondheim at Trondheim, Norway, in 1990; a Doctor of Technical Sciences from the Warsaw University of Technology at Warsaw, Poland, in 1996; and Doctor Honouris Causa from the National Academy of Sciences at Gomel, Belarus, in 2000, and University of Kragujevac, Serbia, in 2011. He is a registered professional engineer. He is presently an Ohio Eminent Scholar and The Howard D. Winbigler Professor in the College of Engineering, and the Director of the Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics (NLB^2) and affiliated faculty in John Glenn College of Public Affairs at the Ohio State University, Columbus, Ohio. In 2013–2014, he served as an ASME/AAAS Science and Technology Policy Fellow, House Committee on Science, Space and Technology, United States Congress, Washington, DC. His research interests include fundamental studies with a focus on scanning probe techniques in the interdisciplinary areas of bio/nanotribology, bio/nanomechanics and bio/nanomaterials characterization and applications to bio/nanotechnology, and biomimetics. He is an internationally recognized expert of bio/nanotribology and bio/nanomechanics using scanning probe microscopy and is one of the most prolific authors. He is considered by some a pioneer of tribology and mechanics of magnetic storage devices. He has authored 8 scientific books, 90+ handbook chapters, 800+ scientific papers

(Goggle Scholar h-index – 105+ with 50k+ citations; Web of Science h-index – 80+; ISI Highly Cited Researcher in Materials Science since 2007 and in Biology and Biochemistry since 2013; ISI Top 5 % Cited Authors for Journals in Chemistry since 2011), and 60+ technical reports. He has also edited 50+ books and holds 20 US and foreign patents. He is coeditor of Springer NanoScience and Technology Series and coeditor of Microsystem Technologies, and Member of Editorial Board of PNAS. He has given more than 400 invited presentations on six continents and more than 200 keynote/plenary addresses at major international conferences.

Dr. Bhushan is an accomplished organizer. He organized the first Symposium on Tribology and Mechanics of Magnetic Storage Systems in 1984 and the first International Symposium on Advances in Information Storage Systems in 1990, both of which are now held annually. He organized two international NATO institutes in Europe. He is the founder of an ASME Information Storage and Processing Systems Division founded in 1993 and served as the founding chair during 1993–1998. His biography has been listed in over two dozen *Who's Who* books including *Who's Who in the World*. He has received more than two dozen awards for his contributions to science and technology from professional societies, industry, and US government agencies including Life Achievement Tribology Award and Institution of Chemical Engineers (UK) Global Award. His research was listed as the top ten science stories of 2015. He is also the recipient of various international fellowships including the Alexander von Humboldt Research Prize for Senior Scientists, Max Planck Foundation Research Award for Outstanding Foreign Scientists, and Fulbright Senior Scholar Award. He is a foreign member of the International Academy of Engineering (Russia), Byelorussian Academy of Engineering and Technology, and the Academy of Triboengineering of Ukraine; an honorary member of the Society of Tribologists of Belarus and STLE; a fellow of ASME, IEEE, and the New York Academy of Sciences; and a member of ASEE, Sigma Xi, and Tau Beta Pi.

Dr. Bhushan has previously worked for Mechanical Technology Inc., Latham, NY; SKF Industries Inc., King of Prussia, PA; IBM, Tucson, AZ; and IBM Almaden Research Center, San Jose, CA. He has held visiting professorship at University of California at Berkeley; University of Cambridge, UK; Technical University Vienna, Austria; University of Paris, Orsay; ETH Zurich; EPFL Lausanne; Univ. of Southampton, UK; Univ. of Kragujevac, Serbia; Tsinghua Univ., China; Harbin Inst., China; and KFUPM, Saudi Arabia.

http://www.mecheng.osu.edu/nlbb/

Section Editors



1. Section: Definitions of Terms

Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



2. Section: Microfabrication Processes and Materials for Fabrication

Chunlei (Peggy) Wang Department of Mechanical and Materials Engineering, Florida International University, Miami, FL, USA



Marc Madou Department of Mechanical and Aerospace and Biomedical Engineering, University of California at Irvine, Irvine, CA, USA



Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



3. Section: Nanofabrication Processes and Materials for Fabrication

Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



4. Section: Nanoscale Measurement Techniques – SFA, STM, AFM, Nanoindenter, Electrical, Etc.

Enrico Gnecco Friedrich Schiller University of Jena, Otto Schott Institute of Materials Research, Jena, Germany



Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

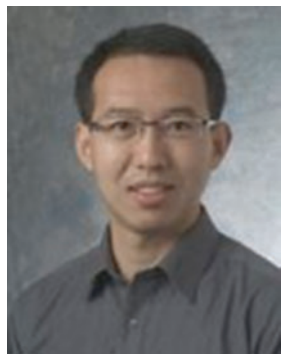


5. Section: Nanostructures

Fabrizio Pirri Dipartimento di Scienza dei Materiali e Ing. Chimica, Laboratorio Materiali e Microsistemi/LATEMAR, Politecnico di Torino, Torino, Italy



Ahmed A. Busnaina NSF Science and Engineering Center for High-rate Nanomanufacturing, Northeastern University, Boston, MA, USA



Philip Feng Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA

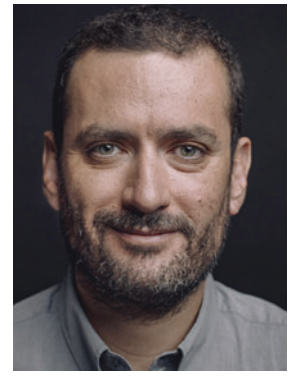


6. Section: Nanomaterials

Apparao M. Rao Clemson Nanomaterials Center, Department of Physics and Astronomy, Center for Optical Materials Science and Engineering Technologies, Clemson University, Anderson, SC, USA



Ramakrishna Podila Department of Physics and Astronomy, Clemson University, Clemson, SC, USA



Ernest Mendoza Centre de Recerca en Nanoenginyeria, Universitat Politècnica de Catalunya, Barcelona, Spain



Eberhard Burkel Institute of Physics, University of Rostock, Rostock, Germany



7. Section: Properties: Mechanical, Electrical, Tribology, Etc.

Lorenzo Valdevit Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA



Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



8. Section: Molecular Modeling

Alessandra Catellani CNR-NANO_S3, Modena, Italy



Giancarlo Cicero Department of Applied Science and Technology, Politecnico di Torino, Turin, Italy



Carlo Massimo Casciola Mechanical and Aerospace Engineering Department, Sapienza University of Rome, Roma, Italy

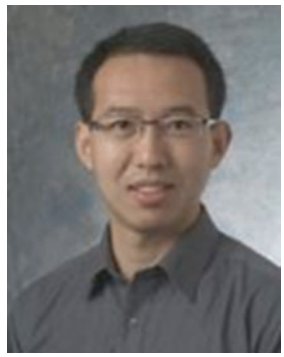


Simone Meloni Mechanical and Aerospace Engineering Department, Sapienza University of Rome, Roma, Italy



9. Section: MEMS/NEMS

Matteo Rinaldi Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA



Philip Feng Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA



Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



10. Section: Microfluidics and Nanofluidics

Shaurya Prakash E347 Scott Laboratory, Department of Mechanical Engineering, The Ohio State University, Columbus, OH, USA



Frederic Restagno Laboratoire de Physique des Solides, Université Paris Sud – CNRS, ORSAY cedex, France



Vincent Senez CNRS/IEMN UMR 8520, Villeneuve d'Ascq cedex, France



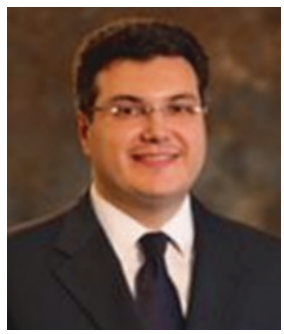
11. Section: Biomedical Engineering and Biodevices

Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



12. Section: Bio/Nanotechnology and Nanomedicine

Jared M. Brown Department of Pharmaceutical Sciences, University of Colorado | Anschutz Medical Campus, Skaggs School of Pharmacy, Aurora, CO, USA



13. Section: Nanoparticles and Other Methods for Drug Delivery and Biomedical Imaging

Paolo Decuzzi Translational Imaging Department and Nanomedicne Department, Houston Methodist Research Institute, Houston, TX, USA



Tony Y. Hu Department of Nanomedicine, The Methodist Hospital Research Institute, Houston, TX, USA



14. Section: Environment, Health, and Safety Issues in Nanomaterials

Emmanuel Flahaut Institut Carnot CIRIMAT, UMR 5085, Bâtiment CIRIMAT, Université de Toulouse, UPS, INP, Toulouse cedex 9, France



Candace S.-J. Tsai Department of Environmental and Radiological Health Science, Colorado State University, Fort Collins, CO, USA

Birck Nanotechnology Center, Purdue University, West Lafayette, IN, USA

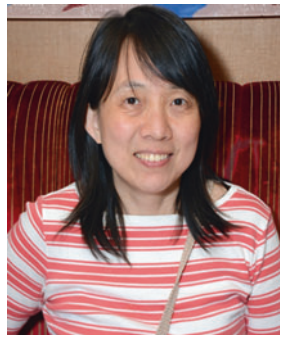


15. Section: Nano-optics and Nano-optical Devices

Othmar Marti Abteilung Experimentelle Physik, Universitaet Ulm, Ulm, Germany



Fabrizio Giorgis Department of Applied Science and Technology, Politecnico di Torino, Torino, Italy



Lih Y. Lin Department of Electrical Engineering, University of Washington, Seattle, WA, USA

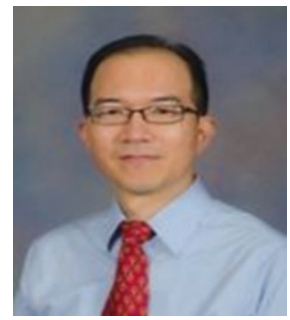


16. Section: Micro/Nano Integration

Huikai Xie Biophotonics and Microsystems Lab, Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA



Philippe Lutz The Université de Franche-Comté, Besançon, France



Y. K. Yoon Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA



17. Section: Materials, Coatings, and Surface Treatments for Nanotribology

Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



18. Section: Biomimetics: Plants and Biominerals

Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



19. Section: Biomimetics: Animals

W. Jon P. Barnes Centre for Cell Engineering, University of Glasgow, Glasgow, UK



20. Section: Biomimetics: Bioinspiration and Products

Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA



21. Section: Nanomanufacturing and Self-Assembly

Ahmed A. Busnaina NSF Science and Engineering Center for High-rate Nanomanufacturing, Northeastern University, Boston, MA, USA



Erik Svedberg The National Academies of Sciences, Engineering, and Medicine, National Materials and Manufacturing Board, Washington, DC, USA



22. Section: Nano-Plasmonics

Andrea Toma Istituto Italiano di Tecnologia, Genova, Italy



Remo Proietti Zaccaria Istituto Italiano di Tecnologia, Genoa, Italy



Rabah Boukherroub Institut d'Electronique, de Microélectronique et de Nanotechnologie (IEMN), UMR CNRS8520, Université Lille1, Villeneuve d'Ascq, France



Sabine Szunerits Institut d'Electronique, de Microélectronique et de Nanotechnologie (IEMN), UMR CNRS8520, Université Lille1, Villeneuve d'Ascq, France



23. Section: Nanoelectronics

Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Contributors

Ahmad Nabil Abbas Department of Electrical and Computer Engineering, Florida International University, Miami, FL, USA

Patrick Abgrall Formulaction, L'Union, France

Angelo Accardo Lab. BIONEM, Dipartimento di Medicina Sperimentale e Clinica, Università "Magna Grecia" di Catanzaro, Catanzaro, Italy

Soft Matter Structures Group ID13 – MICROFOCUS Beamline, European Synchrotron Radiation Facility, Grenoble, France

Wafa Achouak iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

Laboratoire d'Ecologie Microbienne de la Rhizosphère et d'Environnements Extrême, UMR 6191 CNRS–CEA–Aix–Marseille Université de la Méditerranée, CEA Cadarache, St Paul lez Durance, France

Véronique Adam Laboratoire d'Hydrologie et de Géochimie de la Surface, EOST, UDS (Université de Strasbourg), Strasbourg, France

Solomon Adera Device Research Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Masoud Agah Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA

Ranjeet Agarwala Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC, USA

Richa Agrawal Department of Mechanical and Materials Engineering, Florida International University, Miami, FL, USA

Daniel Aili Division of Molecular Physics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden

Alexey V. Akimov Department of Chemistry, University of South California, Los Angeles, CA, USA

Abdellatif Akjouj Institut d'Electronique, de Microélectronique et de Nanotechnologie, UMR CNRS 8520, Université Lille 1, Sciences et Technologies, Villeneuve d'Ascq, France

Alessandro Alabastri Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Seyedhamidreza Alaie Radiology, Weill Cornell Medicine, New York, NY, USA

Muhammad A. Alam School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

Tuncay Alan Mechanical and Aerospace Engineering Department, Monash University, VIC, Australia

Paula A. Algarin Department of Electrical Engineering, University of South Florida, Tampa, FL, USA

Antonio Aliano Department of Physics, Politecnico di Torino, Torino, Italy

Wafa' T. Al Jamal Nanomedicine Laboratory, Centre for Drug Delivery Research, The School of Pharmacy, University of London, London, UK

Paolo Allia Materials Science and Chemical Engineering Department, Politecnico di Torino, Torino, Italy

N. R. Aluru Department of Mechanical Science and Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana – Champaign, Urbana, IL, USA

Nabil Amari INSA Centre Val de Loire, Université d'Orléans, Bourges, France

Giampiero Amato Quantum Research Laboratory, Electromagnetism Division, Istituto Nazionale di Ricerca Metrologica, Torino, Italy

Benoy Anand Department of Physics, Sri Sathya Sai Institute of Higher Learning, Vidyagiri, Puttaparthi, Andhra Pradesh, India

Luca Anghinolfi Dipartimento di Fisica, Università di genova, Genova, Italy

Laboratory for Micro- and Nanotechnology, Paul Scherrer Institut (PSI), Villigen, Switzerland

José V. Anguita Nano Electronics Center, Advanced Technology Institute, University of Surrey, Guildford, Surrey, UK

Umberto Anselmi-Tamburini Department of Chemistry, University of Pavia, Pavia, Italy

Wadih Arap David H. Koch Center, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Departments of Genitourinary Medical Oncology and Cancer Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Thomas Arnold Leibniz-Institute of Surface Modification, Leipzig, Germany

Walter Arnold Department of Material Science and Technology, Saarland University, Saarbrücken, Germany

Physikalisches Institut, Göttingen University, Göttingen, Germany

Christopher Arntsen Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, USA

Eduard Arzt INM – Leibniz Institute for New Materials, Saarbrücken, Germany

Saarland University, Saarbrücken, Germany

Burcu Aslan Department of Experimental Therapeutics, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

Elena Astanina Department of Oncological Sciences, University of Torino, Candiolo, Torino, Italy

Orlando Auciello Materials Science Division, Argonne National Laboratory, Argonne, IL, USA

Mélanie Auffan CEREGE, UMR 6635 CNRS/Aix–Marseille Université, Aix–en–Provence, France

iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

M. Autore Department of Physics, Sapienza University of Rome, Rome, Italy

Thomas Bachmann Institute for Fluid Mechanics and Aerodynamics, Technische Universität Darmstadt, Darmstadt, Germany

Armelle Baeza-Squiban Unit of Functional and Adaptive Biology (BFA), Laboratory of Molecular and Cellular Responses to Xenobiotics, UMR CNRS 8251, Univ Paris Diderot, (Sorbonne Paris Cité), France

Darren M. Bagnall School of Photovoltaic and Renewable Energy Engineering, The University of New South Wales, Sydney, NSW, Australia

Gaurav Bahl Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Xuedong Bai Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Zhongguancun, Beijing, China

Anisullah Baig Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

David J. Bakewell Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, UK

Tommaso Baldacchini Technology and Applications Center, Newport Corporation, Irvine, CA, USA

Department of Electrical Engineering, University of Nebraska–Lincoln, Lincoln, NE, USA

Belén Ballesteros ICN2 – Institut Català de Nanociència i Nanotecnologia, Bellaterra, Barcelona, Spain

Antoine Barbier CEA-Saclay, DSM/IRAMIS/SPCSI, Gif-sur-Yvette, France

Giovanni Barcaro CNR IPCF, Consiglio Nazionale delle Ricerche, Pisa, Tuscany, Italy

Robert Barchfeld Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Yoseph Bar-Cohen Jet Propulsion Laboratory (JPL), California Institute of Technology, Pasadena, CA, USA

Charles B. Barker Department of Mechanical Engineering and Materials Science, Rice University, Houston, TX, USA

W. Jon. P. Barnes University of Glasgow, Institute of Molecular Cell and Systems Biology, Centre for Cell Engineering, Joseph Black Building, University Avenue, Glasgow, Scotland, UK

Larry R. Barnett Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Friedrich G. Barth Life Sciences, Department of Neurobiology, University of Vienna, Vienna, Austria

Michael H. Bartl Department of Chemistry, University of Utah, Salt Lake City, UT, USA

R. Baskaran Department of Electrical Engineering, University of Washington, Seattle, WA, USA

Components Research, Intel Corporation, Santa Clara, CA, USA

Buzz Baum Medical Research Council – Laboratory for Molecular Cell Biology, University College London, London, UK

Werner Baumgartner Johannes Kepler University, Linz, Austria

Pierre Becker Laboratoire de Biologie des Organismes Marins et Biomimétisme, Université de Mons – UMONS, Mons, Belgium

Allison M. Beese Department of Materials Science and Engineering, Pennsylvania State University, University Park, PA, USA

Novid Beheshti Swedish Biomimetics 3000® Ltd, iBIC – Birmingham Science Park Aston, Birmingham, UK

Rachid Belkhou Synchrotron SOLEIL, L'Orme des Merisiers, Gif-sur-Yvette, France

Laila Benameur Department of Biomedical Engineering, McGill University, Montreal, QC, Canada

Rüdiger Berger Max Planck Institute for Polymer Research, Mainz, Germany

Magnus Berggren Department of Science and Technology, Linköpings University, Norrköping, Sweden

Pierre Berini School of Information Technology and Engineering (SITE), University of Ottawa, Ottawa, ON, Canada

Valentin G. Bessergenev CCMar – Centre of marine science and FCT, Universidade do Algarve, Faro, Portugal

Shekhar Bhansali Department of Electrical and Computer Engineering, Florida International University, Miami, FL, USA

Vikram Bhatia Science and Technology, Corning Incorporated SP-PR-02-1, Corning, NY, USA

Gauri Bhave Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, USA

Rustom B. Bhiladvala Department of Mechanical Engineering, University of Victoria, Victoria, BC, Canada

Bharat Bhushan Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Xin Bian Division of Applied Mathematics, Brown University, Providence, RI, USA

Stefano Bianco Center for Space Human Robotics, Fondazione Istituto Italiano di Tecnologia, Torino, Italy

L. Biferale Department of Physics and INFN, University of "Tor Vergata", Rome, Italy

Francesco Bisio CNR-SPIN, Genova, Italy

Ion Bita Qualcomm MEMS Technologies, Inc., San Jose, CA, USA

Martin G. Blaber Department of Chemistry and International Institute for Nanotechnology, Northwestern University, Evanston, IL, USA

Luca Boarino Istituto Nazionale di Ricerca Metrologica, Nanofacility Piemonte, Torino, Italy

Jorge Boczkowski INSERM U955 Eq04, University Paris Est Val de Marne (UPEC), Créteil, France

Stuart A. Boden Nanoelectronics and Nanotechnology Research Group, Electronics and Computer Science, University of Southampton, Highfield, Southampton, UK

Wiktor Bodnar Institute of Physics, University of Rostock, Rostock, Germany

Mauro Boero Institut de Physique et Chimie des Materiaux de Strasbourg (IPCMS), University of Strasbourg and CNRS, UMR 7504, Strasbourg, France

Peter Bøggild DTU Nanotech – Department of Micro- and Nanotechnology, Technical University of Denmark, Lyngby, Denmark

Ardemis A. Boghossian Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Paul W. Bohn Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, USA

Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, USA

K. F. Böhringer Department of Electrical Engineering, University of Washington, Seattle, WA, USA

Sonja Boland Unit of Functional and Adaptive Biology (BFA), Laboratory of Molecular and Cellular Responses to Xenobiotics, UMR CNRS 8251, Univ Paris Diderot, (Sorbonne Paris Cité), France

Robert D. Bolskar TDA Research, Inc., Wheat Ridge, CO, USA

Valentina Bonanni Department of Chemistry, University of Florence, Florence, Italy

James C. Bonner Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

Nicephore Bonnet National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

Alexander Booth Energy and Resources Research Institute, University of Leeds, Leeds, West Yorkshire, UK

Garry J. Bordonaro Cornell NanoScale Science and Technology Facility, Cornell University, Ithaca, NY, USA

Edward Bormashenko Laboratory of Polymers, Physics Department, Ariel University, Ariel, Israel

Céline Botta Aix-Marseille Université, CNRS, IRD, CEREGE UM34, Aix-en- Provence, France

Jean-Yves Bottero Aix-Marseille Université, CNRS, IRD, CEREGE UM34, Aix en Provence, France

Alessia Bottos Department of Oncological Sciences, University of Torino, Candiolo, Torino, Italy

Rabah Boukherroub Institut d'Electronique, de Microélectronique et de Nanotechnologie (IEMN), UMR CNRS 8520, Université Lille 1, Villeneuve d'Ascq, France

Floriane Bourdiol EcoLab – Laboratoire d'Écologie Fonctionnelle et Environnement, Université de Toulouse, INP, UPS, Castanet Tolosan, France

CNRS UMR 5245, EcoLab, Castanet Tolosan, France

Institut Carnot Cirimat, Université de Toulouse, UPS, INP, Toulouse, France

Olivier Bourgeois Institut Néel CNRS-UJF, Grenoble, France

Herbert Bousack Peter Grünberg Institut, Forschungszentrum Jülich GmbH, Jülich, Germany

Michael T. Bowser Department of Chemistry, University of Minnesota, Minneapolis, MN, USA

Thomas Braschler Laboratory of Stem Cell Dynamics, SV-EPFL, Lausanne, Switzerland

Graham Bratzel Laboratory for Atomistic and Molecular Mechanics, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Donald W. Brenner Department of Materials Science and Engineering, North Carolina State University, Raleigh, NC, USA

Victor M. Bright Department of Mechanical Engineering, University of Colorado, Boulder, CO, USA

Sandrine Brochard Department of Physics and Mechanics of Materials, Institut P′, CNRS – Université de Poitiers ENSMA, Futuroscope Chasseneuil, France

Lawrence F. Bronk David H. Koch Center, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Joseph J. Brown Department of Mechanical Engineering, University of Colorado, Boulder, CO, USA

Dorothea Brüggemann The Naughton Institute, School of Physics, Trinity College Dublin, CRANN, Dublin, Ireland

D. M. Bubb Department of Physics, Rutgers University – Camden, Camden, NJ, USA

Markus J. Buehler Laboratory for Atomistic and Molecular Mechanics, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Sven Burger Zuse Institute Berlin, Berlin, Germany

Eberhard Burkel Institute of Physics, University of Rostock, Rostock, Germany

Federico Bussolino Department of Oncological Sciences, University of Torino, Candiolo, Torino, Italy

Donald P. Butler Nanotechnology Research and Teaching Facility, The Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX, USA

Hans-Jürgen Butt Max Planck Institute for Polymer Research, Mainz, Germany

Javier Calvo Fuentes Nanogap Sub-Nm-Powder S.A., Milladoiro – Ames (A Coruña), Spain

Michael Caminale Dipartimento di Fisica, Università di genova, Genova, Italy

CEA, INAC–SPINTEC, Grenoble, France

Giulio Campo Department of Chemistry, University of Florence, Florence, Italy

Maurizio Canepa Dipartimento di Fisica, Università di genova, Genova, Italy

Mary Cano-Sarabia ICN2 – Institut Catala de Nanociencia i Nanotecnologia, Bellaterra, Barcelona, Spain

Andrés Cantarero Materials Science Institute, University of Valencia, Valencia, Spain

Paola Carbone School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester, UK

Francesca Carpino Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, USA

Adrian Carretero-Genevrier Institute des Nanotechnologies de Lyon (INL-CNRS), Lyon, France

Marie Carrière Laboratoire Lésions des Acides Nucléiques, Commissariat à l'Energie Atomique, SCIB, UMR-E 3 CEA/UJF-Grenoble 1, INAC, Grenoble, France

Jérôme Casas Institut de Recherche en Biologie de l'Insecte, IRBI UMR CNRS 6035, Université of Tours, Tours, France

Francesca Casoli CNR-IMEM (Institute of Materials for Electronics and Magnetism – Italian National Research Council), Parma, Italy

Oscar Castaño Biomaterials for Regenerative Therapies, Institute for Bioengineering of Catalonia (IBEC), Barcelona, Spain

Zeynep Celik-Butler Nanotechnology Research and Teaching Facility, The Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX, USA

Frederik Ceyssens Department ESAT-MICAS, KULeuven, Leuven, Belgium

Nicolas Chaillet FEMTO-ST/UFC-ENSMM-UTBM-CNRS, Besançon, France

Peggy Chan Micro/Nanophysics Research Laboratory, RMIT University, Melbourne, VIC, Australia

Munish Chanana Departamento de Química Física, Universidade de Vigo, Vigo, Spain

Pen-Shan Chao Chung-Shan Institute of Science and Technology, Taoyuan, Taiwan, ROC

Satish C. Chaparala Science and Technology, Corning Incorporated SP-PR-02-1, Corning, NY, USA

Changyao Chen Argonne National Laboratory, Lemont, IL, USA

Haorong Chen School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Jian Chen State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing, People's Republic of China

Wei Chen CAS Key Laboratory of Low-Carbon Conversion Science and Engineering, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China

Alessandro Chiolerio Physics Department, Politecnico di Torino, Torino, Italy

Letizia Chiodo Center for Life Nano Science CLNS@Sapienza, Fondazione Istituto Italiano di Tecnologia and European Theoretical Spectroscopy Facility (ETSF), Rome, Italy

Angelica Chiodoni Center for Space Human Robotics, Fondazione Istituto Italiano di Tecnologia, Torino, Italy

Alessandro Chiolerio Physics Department, Politecnico di Torino, Torino, Italy

Fred Chiou Department of Engineering Technology, Weber State University, Ogden, UT, USA

Yoon-Kyoung Cho School of Nano-Bioscience and Chemical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Ulju-gun, Republic of Korea

Jungwook Choi School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Jong Hyun Choi School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Michael Chu Advanced Pharmaceutics and Drug Delivery Laboratory, Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada

Han-Sheng Chuang Department of Biomedical Engineering, National Cheng Kung University, Taiwan

Medical Device Innovation Center, National Cheng Kung University, Taiwan

Lucio Colombi Ciacchi Hybrid Materials Interfaces Group, Faculty of Production Engineering and Bremen Center for Computational Materials Science, University of Bremen, Bremen, Germany

Giancarlo Cicero Department of Physics, Politecnico di Torino, Torino, Italy

Materials Science and Chemical Engineering Department, Politecnico di Torino, Torino, Italy

Christian Cierpka Institute of Fluid Mechanics and Aerodynamics, Universität der Bundeswehr München, Neubiberg, Germany

Katherine Clayton Birck Nanotechnology Center, School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Dominique Collard LIMMS/CNRS-IIS (UMI 2820), Institute of Industrial Science, The University of Tokyo, Meguro-ku, Tokyo, Japan

Maria Laura Coluccio Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Lab. BIONEM, Dipartimento di Medicina Sperimentale e Clinica, Università "Magna Grecia" di Catanzaro, Catanzaro, Italy

A. T. Conlisk Department of Mechanical Engineering, The Ohio State University, Columbus, OH, USA

Andrew Copestake Swedish Biomimetics 3000® Ltd, University of Southampton Science Park, Southampton, UK

Marta Corno Department of Chemistry, University of Torino, Torino, Italy

Miguel A. Correa-Duarte Departamento de Química Física, Universidade de Vigo, Vigo, Spain

Giovanni Costantini Department of Chemistry, The University of Warwick, Coventry, UK

Claire Coutris Department of Plant and Environmental Sciences, Norwegian University of Life Sciences, Ås, Norway

Thomas M. Crawford Department of Physics and Astronomy, Smart State Center for Experimental Nanoscale Physics, University South Carolina, Columbia, SC, USA

Eduardo Cruz-Silva Department of Polymer Science and Engineering, University of Massachusetts Amherst, Amherst, MA, USA

Clotilde S. Cucinotta School of Physics, Trinity College, Dublin, Ireland

Steven Curley Department of Surgical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Christian Dahmen Division Microrobotics and Control Engineering (AMiR), Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Lu Dai The Key Laboratory of Remodeling-related Cardiovascular Diseases, Ministry of Education, Capital Medical University, Beijing, China

F. D'Apuzzo Istituto Italiano di Tecnologia and Sapienza University of Rome, Rome, Italy

Gobind Das Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Bakul C. Dave Department of Chemistry and Biochemistry, Southern Illinois University Carbondale, Carbondale, IL, USA

Francesco De Angelis Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Maarten P. De Boer Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

César de Julián Fernández CNR-IMEM (Institute of Materials for Electronics and Magnetism – Italian National Research Council), Parma, Italy

Marc Lamy de la Chapelle Sorbonne Paris Cité, Laboratoire CSPBAT, CNRS, (UMR 7244), Université Paris 13, Bobigny, France

Roberto de la Rica Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, Scotland, UK

Francesco Buatier de Mongeot Dipartimento di Fisica, Università di genova, Genova, Italy

Amadeo L. Vázquez de Parga Department of Física de la Materia Condensada, Universidad Autónoma de Madrid and Instituto Madrileño de Estudios Avanzados en Nanociencia (IMDEA-Nanociencia), Madrid, Spain

Enrica De Rosa Nanomedicine Department, The Methodist Hospital Research Institute, Houston, TX, USA

Marco De Vivo Department of Drug Discovery and Development - CompuNet, Istituto Italiano di Tecnologia, Genoa, Italy

Paolo Decuzzi Department of Translational Imaging, and Nanomedicine, The Methodist Hospital Research Institute, Houston, TX, USA

Christian L. Degen Department of Physics, ETH Zurich, Zurich, Switzerland

Ada Della Pia Department of Chemistry, The University of Warwick, Coventry, UK

Gregory Denbeaux College of Nanoscale Science and Engineering, University at Albany, Albany, NY, USA

Mingge Deng Division of Applied Mathematics, Brown University, Providence, RI, USA

Parag B. Deotare Electrical Engineering, Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA

Emiliano Descrovi DISAT – Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, Torino, Italy

Joseph M. DeSimone Department of Chemistry, University of North Carolina, Chapel Hill, NC, USA

Department of Pharmacology, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA

Carolina Center of Cancer Nanotechnology Excellence, University of North Carolina, Chapel Hill, NC, USA

Institute for Advanced Materials, University of North Carolina, Chapel Hill, NC, USA

Institute for Nanomedicine, University of North Carolina, Chapel Hill, NC, USA

Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA

Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC, USA

Sloan–Kettering Institute for Cancer Research, Memorial Sloan–Kettering Cancer Center, New York, NY, USA

Hans Deyhle Biomaterials Science Center (BMC), University of Basel, Basel, Switzerland

Charles L. Dezelah IV Picosun USA, LLC, Detroit, MI, USA

Enzo Di Fabrizio Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Lab. BIONEM, Dipartimento di Medicina Sperimentale e Clinica, Università "Magna Grecia" di Catanzaro, Catanzaro, Italy

Rosa Di Felice Department of Physics and Astronomy, University of Southern California, Los Angeles, CA, USA

Center S3, CNR Institute of Nanoscience, Modena, Italy

P. Di Pietro INSTM Udr Trieste-ST and Sincrotrone Trieste, Trieste, Italy

Jan-Henning Dirks Department of New Materials and Biosystems, Max Planck Institute for Intelligent Systems, Stuttgart, Germany

Nathan Doble The New England College of Optometry, Boston, MA, USA

Mitchel J. Doktycz Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Fabrizio Dolcini Department of Applied Science and Technology, Politecnico di Torino, Torino, Italy

Calvin Domier Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Avinash M. Dongare Department of Materials Science and Engineering, and Institute of Materials Science, University of Connecticut, Storrs, CT, USA

Jingyan Dong Edward P. Fitts Department of Industrial and System Engineering, North Carolina State University, Raleigh, NC, USA

Lixin Dong Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA

Jakub Dostalek BioSensor Technologies, AIT-Austrian Institute of Technology GmbH, Vienna, Austria

Emmanuel M. Drakakis Department of Bioengineering, The Sir Leon Bagrit Centre, Imperial College London, London, UK

Wouter H. P. Driessen David H. Koch Center, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Carlos Drummond Centre de Recherche Paul Pascal, CNRS–Université Bordeaux 1, Pessac, France

Jie Du Beijing Institute of Heart Lung and Blood Vessel Diseases, Beijing Anzhen Hospital, Beijing, China

Jean-Marie Dupret Unit of Functional and Adaptive Biology (BFA), Laboratory of Molecular and Cellular Responses to Xenobiotics, UMR CNRS 8251, Univ Paris Diderot, (Sorbonne Paris Cité), Paris cedex 13, France

Julianna K. Edwards David H. Koch Center, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Volkmar Eichhorn Division Microrobotics and Control Engineering (AMiR), Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Thomas Endlein University of Glasgow, Institute of Molecular Cell and Systems Biology, Centre for Cell Engineering, Joseph Black Building, University Avenue, Glasgow, Scotland, United Kingdom

Clifford J. Engel Department of Chemistry, Northwestern University, Evanston, IL, USA

Masayoshi Esashi The World Premier International Research Center Initiative for Atom Molecule Materials, Tohoku University, Aoba-ku Sendai, Japan

Hashem Etayash University of Alberta, Edmonton, AB, Canada

Mikael Evander Department of Biomedical Engineering, Faculty of Engineering, Lund University, Lund, Sweden

Michael Evangelopoulos Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

D. L. Fan Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX, USA

Materials Science and Engineering Program, The University of Texas at Austin, Austin, TX, USA

Yubo Fan Center for Bioengineering and Informatics, Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston, TX, USA

Zheng Fan Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA

Sergej Fatikow Division Microrobotics and Control Engineering (AMiR), Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Henry O. Fatoyinbo Centre for Biomedical Engineering, University of Surrey, Surrey, UK

Jiansheng Feng Device Research Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Jing Feng State Key Laboratory on Integrated Optoelectronics, College of Electronic Science and Engineering, Jilin University, Changchun, China

Philip Feng Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA

Joseph Fernandez-Moure Nanomedicine Department, The Methodist Hospital Research Institute, Houston, TX, USA

Mauro Ferrari Department of NanoMedicine, Houston Methodist Research Institute, Houston, TX, USA

Antoine Ferreira INSA Centre Val de Loire, Université d'Orléans, Bourges, France

Tobin Filleter Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

Benjamin M. Finio School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Emmanuel Flahaut Institut Carnot Cirimat, Université de Toulouse, UPS, INP, Toulouse, France

CNRS, Institut Carnot Cirimat, Toulouse, France

Patrick Flammang Laboratoire de Biologie des Organismes Marins et Biomimétisme, Université de Mons – UMONS, Mons, Belgium

David Folio INSA Centre Val de Loire, Université d'Orléans, Bourges, France

Richard G. Forbes Advanced Technology Institute and Department of Electrical and Electronic Engineering, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, UK

Alessandro Fortunelli Istituto per la Chimica dei Composti Organometallici (Institute for the Chemistry of Organometallic Compounds, ICCOM), Consiglio Nazionale delle Ricerche (National Research Council, CNR), Pisa, PI, Italy

Isabelle Fourquaux CMEAB, Centre de Microscopie Electronique Appliquée à la Biologie, Université Paul Sabatier, Faculté de Médecine Rangueil, Toulouse, France

Marco Francardi Lab. BIONEM, Dipartimento di Medicina Sperimentale e Clinica, Università "Magna Grecia" di Catanzaro, Catanzaro, Italy

International School for Advanced Studies (SISSA), Trieste, Italy

Thomas Franz NTG Neue Technologien GmbH & Co. KG, Gelnhausen, Germany

Francesca Frascella DISAT – Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, Torino, Italy

Andrea Fratalocchi PRIMALIGHT, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Roger H. French Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, OH, USA

James Friend Micro/Nanophysics Research Laboratory, RMIT University, Melbourne, VIC, Australia

Frank Frost Leibniz-Institute of Surface Modification, Leipzig, Germany

Hiroyuki Fujita Center for International Research on MicroMechatronics (CIRMM), Institute of Industrial Science, The University of Tokyo, Meguro-ku, Tokyo, Japan

Katsumasa Fujita Department of Applied Physics, Osaka University, Osaka, Japan

Toshio Fukuda School of Mechatronic Engineering, Beijing Institute of Technology, Beijing, China

Kenji Fukuzawa Department of Micro System Engineering, Nagoya University, Chikusa-ku, Aichi, Nagoya, Japan

Naama E. Toledano Furman Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Diana Gamzina Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Xuefeng Gao Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences, Suzhou, People's Republic of China

Alberto García-Cristóbal Universidad de Valencia, Valencia, Spain

Pablo García-Sánchez Departamento de Electrónica y Electromagnetismo, Universidad de Sevilla, Sevilla, Spain

Jean-Luc Garden Institut Néel CNRS-UJF, Grenoble, France

Paolo Gasco Nanovector srl, Torino, Italy

Laury Gauthier EcoLab – Laboratoire d'écologie fonctionnelle et environnement, Université de Toulouse, INP, UPS, Castanet Tolosan, France

CNRS UMR 5245, EcoLab, Castanet Tolosan, France

Shady Gawad MEAS Switzerland, Bevaix, Switzerland

Denis Gebauer Department of Chemistry, Physical Chemistry, University of Konstanz, Konstanz, Germany

Ille C. Gebeshuber Institute of Applied Physics, Vienna University of Technology, Wien, Austria

Francesco Gentile Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Lab. BIONEM, Dipartimento di Medicina Sperimentale e Clinica, Università "Magna Grecia" di Catanzaro, Catanzaro, Italy

Céline Gérard Department of Physics and Mechanics of Materials, Institut P′, CNRS – Université de Poitiers ENSMA, Futuroscope Chasseneuil, France

Claudio Gerbaldi Department of Applied Science and Technology – DISAT, Institute of Chemistry Politecnico di Torino, Group for Applied Materials and Electrochemistry – GAMELab, Torino, Italy

Amitabha Ghosh Bengal Engineering and Science University, Shibpur, Howrah, India

Ranajay Ghosh Department of Mechanical, Aerospace and Nuclear Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA

Larry R. Gibson II Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, USA

Maria Caterina Giordano Dipartimento di Fisica, Università di genova, Genova, Italy

F. Giorgianni Department of Physics, Sapienza University of Rome, Rome, Italy

Giacomo Giorgi Department of Chemical System Engineering, School of Engineering, The University of Tokyo, Tokyo, Japan

CREST–JST, Tokyo, Japan

Jason P. Gleghorn Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, USA

Julien Godet Department of Physics and Mechanics of Materials, Institut P′, CNRS – Université de Poitiers ENSMA, Futuroscope Chasseneuil, France

Biana Godin Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Juan Sebastian Totero Gongora PRIMALIGHT, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Irene González-Valls Laboratory of Nanostructured Materials for Photovoltaic Energy, Escola Tecnica Superior d Enginyeria (ETSE), Centre d'Investigació en Nanociència i Nanotecnología (CIN2, CSIC), Bellaterra (Barcelona), Spain

Ashwini Gopal Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, USA

Claudia R. Gordijo Advanced Pharmaceutics and Drug Delivery Laboratory, Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada

Reuven Gordon Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

Yann Le Gorrec FEMTO-ST/UFC-ENSMM-UTBM-CNRS, Besançon, France

Alok Govil Qualcomm MEMS Technologies, Inc., San Jose, CA, USA

Richard Gowers School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester, UK

Paul Graham Centre for Computational Neuroscience and Robotics, School of Life Sciences, University of Sussex, Brighton, UK

Dmitri K. Gramotnev Nanophotonics Pty Ltd, Brisbane, QLD, Australia

Nicolas G. Green Nano Research Group, Faculty of Physical and Applied Sciences, University of Southampton, Southampton, UK

School of Electronics and Computer Science, University of Southampton, Highfield, Southampton, UK

Robert J. Greenberg Second Sight Medical Products (SSMP), Sylmar, CA, USA

Julia R. Greer Division of Engineering and Applied Sciences, California Institute of Technology, Pasadena, CA, USA

Dane A. Grismer Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, USA

Jeffrey C. Grossman Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Cordula Grüttner Micromod Partikeltechnologie GmbH, Rostock, Germany

Petra Gruber Transarch – Biomimetics and Transdisciplinary Architecture, Vienna, Austria

Vladimir Gubala Biomedical Diagnostics Institute, Dublin City University, Glasnevin, Dublin, Ireland

Congzhong Guo ECE Department, Carnegie Mellon University, Pittsburgh, PA, USA

Pablo Gurman Materials Science Division, Argonne National Laboratory, Argonne, IL, USA

Evgeni Gusev Qualcomm MEMS Technologies, Inc., San Jose, CA, USA

MEMS Research and Innovation Center, Qualcomm MEMS Technologies, Inc., San Jose, CA, USA

Maria Laura Habegger Department of Integrative Biology, University of South Florida, Tampa, FL, USA

Yassine Haddab FEMTO-ST/UFC-ENSMM-UTBM-CNRS, Besançon, France

Charles-Henri Hage Laboratoire de Physique des Lasers, Atomes et Molécules (PhLAM, UMR 8523), Université Lille 1, Villeneuve d'Ascq cedex, France

Neal A. Hall Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, USA

Moon-Ho Ham Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Xiaodong Han Institute of Microstructure and Property of Advanced Materials, Beijing University of Technology, Chaoyang District, Beijing, People's Republic of China

Hee Dong Han Gynecologic Oncology, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

Center for RNA Interference and Non–coding RNA, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

Zhili Hao Department of Mechanical and Aerospace Engineering, Old Dominion University, Norfolk, VA, USA

Judith A. Harrison Department of Chemistry, United States Naval Academy, Annapolis, MD, USA

Nadine Harris Department of Chemistry and International Institute for Nanotechnology, Northwestern University, Evanston, IL, USA

Kelly A. Hartman Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Achim Hartschuh Department Chemie and CeNS, Ludwig-Maximilians-Universität München, Munich, Germany

Jian He Department of Physics and Astronomy, Clemson University, Clemson, SC, USA

Martin Hegner The Naughton Institute, School of Physics, Trinity College Dublin, CRANN, Dublin, Ireland

Michael G. Helander Department of Materials Science and Engineering, University of Toronto, Toronto, ON, Canada

Laurent Heliot Laboratoire de Physique des Lasers, Atomes et Molécules (PhLAM, UMR 8523), Université Lille 1, Villeneuve d'Ascq cedex, France

Michael J. Heller Department of Nanoengineering, University of California San Diego, La Jolla, CA, USA

Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

Simon J. Henley Nano Electronics Center, Advanced Technology Institute, University of Surrey, Guildford, Surrey, UK

Elise Henneberd Laboratoire de Biologie des Organismes Marins et Biomimétisme, Université de Mons – UMONS, Mons, Belgium

Joseph P. Heremans Department of Mechanical and Aerospace Engineering, Department of Physics and Department of Materials Science and Engineering, The Ohio State University, Columbus, OH, USA

Simone Hieber Biomaterials Science Center (BMC), University of Basel, University Hospital Basel, Basel, Switzerland

Dale Hitchcock Department of Physics and Astronomy, Clemson University, Clemson, SC, USA

Hendrik Hölscher Karlsruher Institut für Technologie (KIT), Institut für Mikrostrukturtechnik, Karlsruhe, Germany

Sarah Hmaidan Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

David Holmes London Centre for Nanotechnology, University College London, London, UK

J. H. Hoo Department of Electrical Engineering, University of Washington, Seattle, WA, USA

Bart W. Hoogenboom London Centre for Nanotechnology and Department of Physics and Astronomy, University College London, London, UK

Kazunori Hoshino Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA

Larry L. Howell Department of Mechanical Engineering, Brigham Young University, Provo, UT, USA

Hou-Jun Hsu C.C.P. Contact Probes CO., LTD, New Taipei, Taiwan, ROC

Jung-Tang Huang National Taipei University of Technology, Taipei, Taiwan, ROC

C. Huck Kirchhoff Institute for Physics, Heidelberg University, Heidelberg, Germany

Michael P. Hughes Centre for Biomedical Engineering, University of Surrey, Surrey, UK

Yu Hui Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA

Shelby B. Hutchens Mechanical Science and Engineering, University of Illinois Urbana-Champaign, Champaign, IL, USA

John W. Hutchinson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Gilgueng Hwang Laboratoire de Photonique et de Nanostructures (LPN-CNRS), Site Alcatel de Marcoussis, Paris, France

Hyundoo Hwang School of Nano-Bioscience and Chemical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Ulju-gun, Republic of Korea

Barbara Imhof LIQUIFER Systems Group, Wien, Austria

Hiromi Inada Hitachi High-Technologies Corporation, Hitachinaka, Japan, Japan

Salman B. Inayat University of California, Davis, CA, USA

Klaus Michael Indlekofer Hochschule RheinMain University of Applied Sciences, Institute of Microtechnologies (IMtech), Rüsselsheim, Germany

Rita Claudia Iotti Department of Applied Science and Technology, Politecnico di Torino, Torino, Italy

Electrical and Computer Engineering, University of California – Davis Integrated Nanodevices and Nanosystems Lab, Davis, CA, USA

Monsur Islam Multiscale Manufacturing Laboratory, Department of Mechanical Engineering, Clemson University, Clemson, SC, USA

M. Saif Islam University of California, Davis, CA, USA

Electrical and Computer Engineering, University of California – Davis Integrated Nanodevices and Nanosystems Lab, Davis, CA, USA

Mitsumasa Iwamoto Department of Physical Electronics, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan

Esmaiel Jabbari Biomimetic Materials and Tissue Engineering Laboratory, Department of Chemical Engineering, Swearingen Engineering Center, University of South Carolina, Columbia, SC, USA

Laurent Jalabert LIMMS/CNRS-IIS (UMI 2820), Institute of Industrial Science, The University of Tokyo, Meguro-ku, Tokyo, Japan

Yogesh Jaluria Department of Mechanical and Aerospace Engineering, Rutgers University, Piscataway, NJ, USA

Dongchan Jang Department of Applied Physics and Materials Science, 230 Steele Laboratory, California Institute of Technology, Pasadena, CA, USA

Daniel Jasper Division Microrobotics and Control Engineering (AMiR), Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Debdeep Jena Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA

Lixin Jia The Key Laboratory of Remodeling-related Cardiovascular Diseases, Ministry of Education, Capital Medical University, Beijing, China

Lei Jiang Center of Molecular Sciences, Institute of Chemistry Chinese Academy of Sciences, Beijing, People's Republic of China

Lijia Jiang Department of Electrical Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA

Taeksoo Ji School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea

Samsung Mobile Display Co., Ltd, Young–in, South Korea

Zhaoxia Ji California NanoSystems Institute, University of California, Los Angeles, CA, USA

Dilip S. Joag Centre for Advanced Studies in Materials Science and Condensed Matter Physics, Department of Physics, Savitribai Phule Pune University, Pune, Maharashtra, India

Stefan Johansson Department of Engineering Sciences, Uppsala University, Uppsala, Sweden

Erik J. Joner Bioforsk Soil and Environment, Ås, Norway

Suhas S. Joshi Department of Mechanical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

Gabriela Juarez-Martinez Centeo Biosciences Limited, Dumbarton, UK

Soyoun Jung Samsung Mobile Display Co., Ltd, Young-in, South Korea

Christian J. Kähler Institute of Fluid Mechanics and Aerodynamics, Universität der Bundeswehr München, Neubiberg, Germany

Sergei V. Kalinin Oak Ridge National Laboratory, Oak Ridge, TN, USA

Ping Kao The Pennsylvania State University, University Park, PA, USA

Swastik Kar Department of Physics, Northeastern University, Boston, MA, USA

Rassul Karabalin TowerJazz Semiconductors, Newport Beach, CA, USA

Mustafa Karabiyik Department of Electrical and Computer Engineering, Florida International University, Miami, FL, USA

Sinan Karaveli School of Engineering, Brown University, Providence, RI, USA

David Karig Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA

George Em Karniadakis Division of Applied Mathematics, Brown University, Providence, RI, USA

Michael Karpelson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Andreas G. Katsiamis Toumaz Technology Limited, Abingdon, UK

Anupama B. Kaul Department of Metallurgical and Materials Engineering, Department of Electrical and Computer Engineering (Joint), University of Texas, El Paso, College of Engineering, El Paso, TX, USA

Christine D. Keating Department of Chemistry, Penn State University, University Park, PA, USA

Pamela L. Keating Department of Chemistry, United States Naval Academy, Annapolis, MD, USA

John B. Ketterson Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA

S. M. Khaled Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Arash Kheyraddini Mousavi Department of Mechanical Engineering, New Mexico Tech, Socorro, NM, USA

Andrei L. Kholkin Institute of Natural Sciences, Ural Federal University, Ekaterinburg, Russia

Department of Materials and Ceramic Engineering and CICECO, University of Aveiro, Aveiro, Portugal

Jian-Wei Khor Birck Nanotechnology Center, School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Boris Khusid Department of Chemical, Biological and Pharmaceutical Engineering, New Jersey Institute of Technology, Newark, NJ, USA

Bongsang Kim Advanced MEMS, Sandia National Laboratories, Albuquerque, NM, USA

Changho Kim Division of Applied Mathematics, Brown University, Providence, RI, USA

C. J. Kim Mechanical and Aerospace Engineering Department, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Kwanoh Kim Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX, USA

Moon Suk Kim Department of Molecular Science and Technology, Ajou University, Suwon, South Korea

Pilhan Kim Graduate School of Nanoscience and Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea

Seong H. Kim Department of Chemical Engineering, Pennsylvania State University, University Park, PA, USA

Uday Kishore Centre for Infection, Immunity and Disease Mechanisms, College of Health and Life Sciences, Brunel University London, London, UK

M. Todd Knippenberg Department of Chemistry, High Point University, High Point, NC, USA

Joachim Knoch Institute of Semiconductor Electronics, RWTH Aachen University, Aachen, Germany

Yee Kan Koh Department of Mechanical Engineering, National University of Singapore, Singapore, Singapore

Helmut Kohl Physikalisches Institut, Westfälische Wilhelms-Universität Münster, Münster, Germany

Mathias Kolle Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA

Susan Köppen Hybrid Materials Interfaces Group, Faculty of Production Engineering and Bremen Center for Computational Materials Science, University of Bremen, Bremen, Germany

Monica Kosa Department of Chemistry, Faculty of Exact Sciences, Bar Ilan University, Ramat Gan, Israel

Kostas Kostarelos Nanomedicine Laboratory, Centre for Drug Delivery Research, The School of Pharmacy, University of London, London, UK

Abhay Kotnala Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

Roman Krahne Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Gijs Krijnen Transducers Science and Technology group, MESA + Research Institute for Nanotechnology, University of Twente, Enschede, The Netherlands

Rajaram Krishnan Biological Dynamics, Inc., University of California San Diego, San Diego, CA, USA

Florian Krohs Division Microrobotics and Control Engineering (AMiR), Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Elmar Kroner INM – Leibniz Institute for New Materials, Saarbrücken, Germany

Tom N. Krupenkin Department of Mechanical Engineering, The University of Wisconsin-Madison, Madison, WI, USA

Aloke Kumar Department of Mechanical Engineering, University of Alberta, AB, Canada

Satish Kumar G. W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Momoko Kumemura LIMMS/CNRS-IIS (UMI 2820), Institute of Industrial Science, The University of Tokyo, Meguro-ku, Tokyo, Japan

Harry Kwok Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

Jae-Sung Kwon Department of Chemistry, University of Minnesota, Minneapolis, MN, USA

Jérôme Labille Aix-Marseille Université, CNRS, IRD, CEREGE UM34, Aix en Provence, France

Jean Christophe Lacroix Interfaces, Traitements, Organisation et Dynamique des Systemes Universite Paris 7-Denis Diderot, UMR 7086, Paris, France

Nicolas Lafitte LIMMS/CNRS-IIS (UMI 2820), Institute of Industrial Science, The University of Tokyo, Meguro-ku, Tokyo, Japan

Akhlesh Lakhtakia Department of Engineering Science and Mechanics, Pennsylvania State University, University Park, PA, USA

Andrea Lamberti Department of Applied Science and Technology, Politecnico di Torino, Torino, Italy

Périne Landois Institut Carnot Cirimat, Université de Toulouse, UPS, INP, Toulouse, France

Amy Lang Department of Aerospace Engineering and Mechanics, University of Alabama, Tuscaloosa, AL, USA

Sophie Lanone Inserm U955, Équipe 4, Université Paris Est Val de Marne (UPEC), Créteil, France

Service de pneumologie et pathologie professionnelle, Hôpital Intercommunal de Créteil, Créteil, France

Gregory M. Lanza C-TRAIN Labs, Washington University, St. Louis, MO, USA

Lars Uno Larsson Swedish Biomimetics 3000® AB, Stockholm, Sweden

Camille Larue Laboratoire Lésions des Acides Nucléiques, Commissariat à l'Energie Atomique, SCIB, UMR-E 3 CEA/UJF-Grenoble 1, INAC, Grenoble, France

Michael J. Laudenslager Department of Materials Science and Engineering, University of Florida, Gainesville, FL, USA

Thomas Laurell Department of Biomedical Engineering, Lund University, Lund, Sweden

M. Laver Laboratory for Neutron Scattering, Paul Scherrer Institut, Villigen, Switzerland

Materials Research Division, Risø DTU, Technical University of Denmark, Roskilde, Denmark

Nano–Science Center, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA

Falk Lederer Institute of Condensed Matter Theory and Solid State Optics, Abbe Center of Photonics, Friedrich-Schiller-Universität Jena, Jena, Germany

Jaeho Lee Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

Joshua E-Y Lee Department of Electronic Engineering, City University of Hong Kong, Kowloon Tong, Hong Kong

Jaesung Lee Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA

David W. Lee Department of Biological Sciences, Florida International University Modesto Maidique Campus, Miami, FL, USA

Kuo-Yu Lee National Taipei University of Technology, Taipei, Taiwan, ROC

Andreas Lenshof Department of Biomedical Engineering, Lund University, Lund, Sweden

Donald J. Leo Mechanical Engineering, Center for Intelligent Material Systems and Structures, Virginia Tech, Virginia Polytechnic Institute and State University, Arlington, VA, USA

Zayd Chad Leseman Department of Mechanical Engineering, New Mexico Tech, Socorro, NM, USA

Kenneth Mei Yee Leung The Swire Institute of Marine Sciences and School of Biological Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China

Gaëtan Lévêque Institut d'Electronique, de Microélectronique et de Nanotechnologie, UMR CNRS 8520, Université Lille 1, Sciences et Technologies, Villeneuve d'Ascq, France

Nastassja A. Lewinski Department of Bioengineering, Rice University, Houston, TX, USA

Chen Li Department of Chemistry and Biochemistry, University of Bern, Bern, Switzerland

Chun Li Department of Experimental Diagnostic Imaging-Unit 59, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Feiran Li School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Huan Li Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA

Jason Li Advanced Pharmaceutics and Drug Delivery Laboratory, Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada

King C. Li Department of Bioengineering, Rice University, Houston, TX, USA

Mo Li Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA

Sheng-Shian Li Institute of NanoEngineering and MicroSystems, National Tsing Hua University, Hsinchu, Taiwan

Weicong Li Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

Wenzhi Li Department of Physics, Florida International University, Miami, FL, USA

Xuejin Li Division of Applied Mathematics, Brown University, Providence, RI, USA

Zhang Li Institute of Robotics and Intelligent Systems, ETH Zurich, Zurich, Switzerland

Zhen Li Division of Applied Mathematics, Brown University, Providence, RI, USA

Zhenyu Li Department of Biomedical Engineering, The George Washington University, Washington, DC, USA

Meng Lian Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Carlo Liberale Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

O. Limaj EPFL IBI-STI-BIOS, Lausanne, Switzerland

Chung-Yi Lin FormFactor, Inc, Hsinchu, Taiwan, ROC

Feng Lin Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA

Lih Y. Lin Department of Electrical Engineering, University of Washington, Seattle, WA, USA

Ling Lin Beijing National Laboratory for Molecular Sciences (BNLMS), Key Laboratory of Organic Solids, Institute of Chemistry Chinese Academy of Sciences, Beijing, People's Republic of China

Zhoumeng Lin Institute of Computational Comparative Medicine, Department of Anatomy and Physiology, Kansas State University, Manhattan, KS, USA

Mónica Lira-Cantú Laboratory of Nanostructured Materials for Photovoltaic Energy, Escola Tecnica Superior d Enginyeria (ETSE), Centre d'Investigació en Nanociència i Nanotecnología (CIN2, CSIC), Bellaterra (Barcelona), Spain

Shawn Litster Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

Caihong Liu Wanger Institute for Sustainable Energy Research, Illinois Institute of Technology, Chicago, IL, USA

Department of Mechanical, Materials and Aerospace Engineering, Illinois Institute of Technology, Chicago, IL, USA

Chang Liu Tech Institute, ME/EECS, Northwestern University, Evanston, IL, USA

Gang Logan Liu Micro and Nanotechnology Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Ying Liu School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Yufei Liu Department of Physics and Astronomy, Clemson University, Clemson, SC, USA

Yuyun Liu Department of Materials Science and State Key Laboratory of Molecular Engineering of Polymers, Fudan University, Shanghai, China

Matthew T. Lloyd National Renewable Energy Laboratory, Golden, CO, USA

I. Lo Vecchio Department of Physics, Sapienza University of Rome, Rome, Italy

Sarah B. Lockwood Department of Chemistry and Biochemistry, Southern Illinois University Carbondale, Carbondale, IL, USA

V. J. Logeeswaran University of California, Davis, CA, USA

Electrical and Computer Engineering, University of California – Davis Integrated Nanodevices and Nanosystems Lab, Davis, CA, USA

Mariangela Lombardi IIT – Italian Institute of Technology @ POLITO – Centre for Space Human Robotics, Torino, Italy

Marko Loncar Electrical Engineering, Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA

Kenneth A. Lopata William R. Wiley Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA

Gabriel Lopez-Berestein Department of Experimental Therapeutics, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

Cancer Biology, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

Center for RNA Interference and Non–coding RNA, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

The Department of Nanomedicine and Bioengineering, UTHealth, Houston, TX, USA

Alejandro Lopez-Bezanilla National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA

M. Arturo López-Quintela Laboratory of Magnetism and Nanotechnology, Institute for Technological Research, University of Santiago de Compostela, Santiago de Compostela, Spain

Jun Lou Department of Materials Science and NanoEngineering, Rice University, Houston, TX, USA

Jia Grace Lu Departments of Physics and Electrophysics, University of Southern California, Los Angeles, CA, USA

Michael S.-C. Lu Department of Electrical Engineering, Institute of Electronics Engineering, and Institute of NanoEngineering and MicroSystems, National Tsing Hua University, Taiwan, Republic of China

Wei Lu Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI, USA

Yang Lu Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Kowloon, Hong Kong, P. R. China

Yongfeng Lu Department of Electrical Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA

Zheng-Hong Lu Department of Physics, Yunnan University, Kunming, Yunnan, People's Republic of China

Department of Materials Science and Engineering, University of Toronto, Toronto, ON, Canada

Vanni Lughi DI3 – Department of Industrial Engineering and Information Technology, University of Trieste, Trieste, Italy

Neville C. Luhmann, Jr. Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Lorenzo Lunello Biofunctional Surfaces and Interfaces, FBK-CMM Bruno Kessler Foundation and CNR-IBF, Povo, TN, Italy

S. Lupi IOM-CNR and Department of Physics, Sapienza University of Rome, Rome, Italy

Richard F. Lyon Google Inc, Santa Clara, CA, USA

Kuo-Sheng Ma Kennesaw State University, Marietta, GA, USA

Yingnan Ma Electrical Mechatronics Engineering, Southern Polytechnic State University, Marietta, GA, USA

Marc Madou Department of Mechanical & Aerospace & Biomedical Engineering, University of California at Irvine, Irvine, CA, USA

Imran Mahboob NTT Basic Research Laboratories, NTT Corporation, Atsugi-shi, Kanagawa, Japan

Daniele Malleo Fluxion Biosciences, South San Francisco, CA, USA

Supone Manakasettharn National Nanotechnology Center (NANOTEC), National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand

Maria Grazia Manera CNR-IMM (Institute for Microelectronics and Microsystems – Italian National Research Council), Lecce, Italy

Liberato Manna Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Richard P. Mann Centre for Interdisciplinary Mathematics, Uppsala University, Uppsala, Sweden

Shengcheng Mao Institute of Microstructure and Property of Advanced Materials, Beijing University of Technology, Chaoyang District, Beijing, People's Republic of China

Francelyne Marano Unit of Functional and Adaptive Biology (BFA), Laboratory of Molecular and Cellular Responses to Xenobiotics, UMR CNRS 8251, Univ Paris Diderot, (Sorbonne Paris Cité), Paris cedex 13, France

Sylvain Martel NanoRobotics Laboratory, Department of Computer and Software Engineering, and Institute of Biomedical Engineering, École Polytechnique de Montréal (EPM), Montréal, QC, Canada

Christian Martella Dipartimento di Fisica, Università di genova, Genova, Italy

Pascal Martin University Paris 7-Denis Diderot, ITODYS, Nanoelectrochemistry Group, UMR CNRS 7086, Paris, France

Rodrigo Martinez-Duarte Multiscale Manufacturing Laboratory, Department of Mechanical Engineering, Clemson University, Clemson, SC, USA

Paola Martino Politronica Inkjet Printing Technologies S.r.l, Torino, Italy

Erik Martinsson Division of Molecular Physics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden

Armand Masion Aix-Marseille Université, CNRS, IRD, CEREGE UM34, Aix en Provence, France

iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

Daniel Maspoch ICN2 – Institut Catala de Nanociencia i Nanotecnologia, Bellaterra, Barcelona, Spain

Institució Catalana de Recerca i Estudis Avançats (ICREA), Bellaterra, Barcelona, Spain

Kyoko Masui Applied Physics of Engineering, Osaka University, Osaka, Japan

Cintia Mateo Departamento de Química Física, Universidade de Vigo, Vigo, Spain

Shinji Matsui Laboratory of Advanced Science and Technology for Industry, University of Hyogo, Hyogo, Japan

Giovanni Mattei Department of Physics and Astronomy, University of Padova, Padova, Italy

Alessandro Mattoni Istituto Officina dei Materiali, Consiglio Nazionale delle Ricerche, CNR-IOM Cagliari (SLACS), Monserrato, CA, Italy

Jeffrey S. Mayer Department of Electrical Engineering, Penn State University, University Park, PA, USA

Theresa S. Mayer Department of Electrical Engineering and Materials Science and Engineering, The Pennsylvania State University, University Park, PA, USA

Chimaobi Mbanaso College of Nanoscale Science and Engineering, University at Albany, Albany, NY, USA

Eva McGuirea Department of Materials, Imperial College London, London, UK

Andy C. McIntosh Energy and Resources Research Institute, University of Leeds, Leeds, West Yorkshire, UK

Rachel McKendry London Centre for Nanotechnology and Department of Medicine, University College London, London, UK

Federico Mecarini Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Olin T. Mefford IV Department of Materials Science and Engineering, Clemson University, Clemson, SC, USA

Ernest Mendoza Centre de Recerca en Nanoenginyeria, Universitat Politècnica de Catalunya, Barcelona, Spain

Christoph Menzel Institute of Condensed Matter Theory and Solid State Optics, Abbe Center of Photonics, Friedrich-Schiller-Universität Jena, Jena, Germany

Timothy J. Merkel Department of Chemistry, University of North Carolina, Chapel Hill, NC, USA

Narcis Mestres Consejo Superior de Investigaciones Científicas (CSIC), Institut de Ciencia de Materials de Barcelona (ICMAB), Barcelona, Spain

Vincent Meunier Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, Troy, NY, USA

Wen-Tian Mi Institute of Microelectronics, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, Haidian, China

Agostino Migliore Department of Chemistry, Duke University, Durham, NC, USA

Paul T. Mikulski Department of Physics, United States Naval Academy, Annapolis, MD, USA

Hwall Min The Pennsylvania State University, University Park, PA, USA

Rodolfo Miranda Dep. Física de la Materia Condensada, Universidad Autónoma de Madrid and Instituto Madrileñó de Estudios Avanzados en Nanociencia (IMDEA-Nanociencia), Madrid, Spain

M. Missori Institute for Complex Systems, National Research Council, Rome, Italy

Sushanta K. Mitra Micro and Nano-scale Transport Laboratory, Department of Mechanical Engineering, University of Alberta, Edmonton, AB, Canada

Cristian Mocuta Synchrotron SOLEIL, L'Orme des Merisiers, Gif-sur-Yvette, France

Mohammad R. K. Mofrad Molecular Cell Biomechanics Lab, Department of Bioengineering, University of California, Berkeley, CA, USA

Seyed Moein Moghimi Department of Pharmaceutics and Analytical Chemistry, Centre for Pharmaceutical Nanotechnology and Nanotoxicology, University of Copenhagen, Copenhagen, Denmark

Farghalli A. Mohamed Department of Chemical Engineering and Materials Science, University of California, Irvine the Henry Samueli School of Engineering, Irvine, CA, USA

Alejandro Molina-Sánchez University of Luxembourg, Luxembourg, Luxembourg

Nancy A. Monteiro-Riviere Nanotechnology Innovation Center of Kansas State, Department of Anatomy and Physiology, Kansas State University, Manhattan, KS, USA

Roberto Morandotti Institut National de la Recherche Scientifique – Énergie Matériaux et Télécommunications, Université du Québec, Varennes, QC, Canada

Mahendra A. More Centre for Advanced Studies in Materials Science and Condensed Matter Physics, Department of Physics, Savitribai Phule Pune University, Pune, India

A. Mosca Conte ETSF, and Department of Physics, University of Rome Tor Vergata, Rome, Italy

Philip Motta Department of Integrative Biology, University of South Florida, Tampa, FL, USA

Florence Mouchet EcoLab – Laboratoire d'Écologie Fonctionnelle et Environnement, Université de Toulouse, INP, UPS, Castanet Tolosan, France

CNRS UMR 5245, EcoLab, Castanet Tolosan, France

Catherine Mouneyrac LUNAM Université, Université Catholique de l'Ouest, Angers, France

Weiqiang Mu Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA

Stefan Mühlig Institute of Condensed Matter Theory and Solid State Optics, Abbe Center of Photonics, Friedrich-Schiller-Universität Jena, Jena, Germany

Partha P. Mukherjee Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Bert Müller Biomaterials Science Center (BMC), University of Basel, Basel, Switzerland

Claudia Musicanti Nanovector srl, Torino, Italy

V. Sai Muthukumar Department of Physics, Sri Sathya Sai Institute of Higher Learning, Vidyagiri, Puttaparthi, Andhra Pradesh, India

Jit Muthuswamy School of Biological and Health Systems Engineering, Arizona State University, Tempe, AZ, USA

Rafik Naccache Institut National de la Recherche Scientifique – Énergie Matériaux et Télécommunications, Université du Québec, Varennes, QC, Canada

J. J. Naddeo Department of Physics, Rutgers University – Camden, Camden, NJ, USA

Shrikant C. Nagpure Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Jijeesh R. Nair Department of Applied Science and Technology – DISAT, Institute of Chemistry Politecnico di Torino, Group for Applied Materials and Electrochemistry – GAMELab, Torino, Italy

Vishal V. R. Nandigana Department of Mechanical Science and Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana – Champaign, Urbana, IL, USA

Roger H. French Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, OH, USA

Rucha Natu Multiscale Manufacturing Laboratory, Department of Mechanical Engineering, Clemson University, Clemson, SC, USA

Avinash P. Nayak University of California, Davis, CA, USA

Electrical and Computer Engineering, University of California – Davis Integrated Nanodevices and Nanosystems Lab, Davis, CA, USA

Suresh Neethirajan Schlool of Engineering, University of Guelph, Guelph, ON, Canada

Bradley J. Nelson Institute of Robotics and Intelligent Systems, ETH Zurich, Zurich, Switzerland

Celeste M. Nelson Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, USA

Department of Molecular Biology, Princeton University, Princeton, NJ, USA

Gilbert Daniel Nessim Chemistry Department, Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA), Bar-Ilan University, Ramat Gan, Israel

Fabrice Nesslany Laboratory of Genetic Toxicology, Institut Pasteur de Lille, Lille Cedex, France

F. Neubrech 4th Physics Institute, University of Stuttgart, Stuttgart, Germany

Daniel Neuhauser Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, USA

J. Tanner Nevill Fluxion Biosciences, South San Francisco, CA, USA

Nam-Trung Nguyen Queensland Micro and Nanotechnology Centre, Griffith University, Brisbane, QLD, Australia

Hossein Nili Nano Research Group, University of Southampton, Highfield, Southampton, UK

School of Electronics and Computer Science, University of Southampton, Southampton, UK

Nano Research Group, Faculty of Physical and Applied Sciences, University of Southampton, Southampton, UK

Vincent Niviere GDRI ICEINT: International Center for the Environmental Implications of Nanotechnology, CNRS–CEA, Aix–en–Provence, France

Laboratoire de Chimie et Biologie des Métaux, UMR 5249, iRTSV–CEA Bat. K', Grenoble, France

Michael Nosonovsky Department of Mechanical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Thomas Nowotny Centre for Computational Neuroscience and Robotics, School of Engineering and Informatics, University of Sussex, Brighton, UK

Teri W. Odom Department of Chemistry, Northwestern University, Evanston, IL, USA

Seajin Oh Department of Mechanical and Aerospace Engineering and Biomedical Engineering, University of California at Irvine, Irvine, CA, USA

Murat Okandan Advanced MEMS and Novel Silicon Technologies, Sandia National Laboratories, Albuquerque, NM, USA

Mathew Ombaba University of California, Davis, CA, USA

Brian E. O'Neill Department of Radiology Research, The Methodist Hospital Research Institute, Houston, TX, USA

Takahito Ono Department of Mechanical Systems and Design, Graduate School of Engineering, Tohoku University, Aoba-ku Sendai, Japan

C. H. Raymond Ooi Department of Physics, University of Malaya, Kuala Lumpur, Malaysia

Atsushi Oshiyama Department of Applied Physics, The University of Tokyo, Tokyo, Japan

Clifford W. Padgett Chemistry and Physics, Armstrong Atlantic State University, Savannah, GA, USA

Christine Paillès CEREGE, UMR 6635 CNRS/Aix–Marseille Université, Aix–en–Provence, France

iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

Nezih Pala Department of Electrical and Computer Engineering, Florida International University, Miami, FL, USA

Manuel L. B. Palacio Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Giulia Palermo Department of Drug Discovery and Development – CompuNet, Istituto Italiano di Tecnologia, Genoa, Italy

Laboratory of Computational Chemistry and Biochemistry, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Maurizia Palummo Dipartimento di Fisica, Universitá di Roma Tor Vergata, European Theoretical Spectroscopy Facility (ETSF), Rome, Italy

Jing Pan School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Xudong Pan Harbin Institute of Technology, Harbin, Heilongjian, People's Republic of China

Jeong Young Park Graduate School of EEWS (WCU), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

K. S. Park Department of Electrical Engineering, University of Washington, Seattle, WA, USA

Kinam Park Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA

Woo-Tae Park Seoul National University of Science and Technology, Seoul, South Korea

Andrew R. Parker Department of Zoology, The Natural History Museum, London, UK

Alessandro Parodi Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Renata Pasqualini David H. Koch Center, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Laura Pasquardini Biofunctional Surfaces and Interfaces, FBK-CMM Bruno Kessler Foundation, Povo, TN, Italy

Melissa A. Pasquinelli Fiber and Polymer Science, Textile Engineering, Chemistry and Science, North Carolina State University, Raleigh, NC, USA

Siddhartha Pathak Chemical and Materials Engineering, University of Nevada, Reno, Reno, NV, USA

Nicolas Pavillon Biophotonics Lab, Immunology Frontier Research Center (IFReC), Osaka University, Osaka, Japan

Néstor O. Pérez-Arancibia School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Cecilia Pederzolli Biofunctional Surfaces and Interfaces, FBK-CMM Bruno Kessler Foundation, Povo, TN, Italy

Dmitry Pelegov Institute of Natural Sciences, Ural Federal University, Ekaterinburg, Russia

Natalia Pelinovskaya Aix-Marseille Université, CNRS, IRD, CEREGE UM34, Aix en Provence, France

Giovanni Pellegrini Department of Physics and Astronomy, University of Padova, Padova, Italy

Jean-Philippe M. Peraud Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Dimitrios Peroulis School of Electrical and Computer Engineering, Birck Nanotechnology Center, Purdue University, West Lafayette, IN, USA

Vinh-Nguyen Phan School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, Singapore

Reji Philip Light and Matter Physics Group, Raman Research Institute, Bangalore, India

Andrew Philippides Centre for Computational Neuroscience and Robotics, Department of Informatics, University of Sussex, Brighton, UK

Gianluca Piazza Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA

Remigio Picone Harvard Medical School and Dana-Farber Cancer Institute, Boston, MA, USA

Francesco Pineider Department of Chemistry, University of Florence, Florence, Italy

Laurent Pizzagalli Department of Physics and Mechanics of Materials, Institut P′, CNRS – Université de Poitiers ENSMA, Futuroscope Chasseneuil, France

Ilya V. Pobelov Department of Chemistry and Biochemistry, University of Bern, Bern, Switzerland

Ryan M. Pocratsky Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

Ramakrishna Podila Department of Physics and Astronomy, Clemson University, Clemson, SC, USA

Martino Poggio Department of Physics, University of Basel, Basel, Switzerland

R. G. Polcawich US Army Research Laboratory, Adelphi, MD, USA

Jan Pomplun Zuse Institute Berlin, Berlin, Germany

Kirsten M. Pondman Centre for Infection, Immunity and Disease Mechanisms, College of Health and Life Sciences, Brunel University London, London, UK

Neuro Imaging, MIRA Institute, University of Twente, Enschede, NB, The Netherlands

Samuele Porro Applied Science and Technology Department, Politecnico di Torino, Torino, Italy

Alexandra Portera Department of Materials, Imperial College London, London, UK

Cristina Potrich Biofunctional Surfaces and Interfaces, FBK-CMM Bruno Kessler Foundation and CNR-IBF, Povo, TN, Italy

Siavash Pourkamali Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, USA

Prachya Mruetusatorn Department of Electrical Engineering and Computer Science, University of Tennessee Knoxville, Knoxville, TN, USA

Oakridge National Laboratory, Oak Ridge, TN, USA

Shaurya Prakash Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, OH, USA

Lucas Prater Electrical Mechatronics Engineering, Southern Polytechnic State University, Marietta, GA, USA

Oleg V. Prezhdo Department of Chemistry, University of South California, Los Angeles, CA, USA

Luigi Preziosi Dipartimento di Matematica, Politecnico di Torino, Torino, Italy

Luca Primo Department of Clinical and Biological Sciences, University of Torino, Candiolo, Torino, Italy

R. M. Proie US Army Research Laboratory, Adelphi, MD, USA

Olivier Proux GDRI ICEINT: International Center for the Environmental Implications of Nanotechnology, CNRS–CEA, Aix–en–Provence, France

OSUG, Grenoble, France

A. Pucci Kirchhoff Institute for Physics, Heidelberg University, Heidelberg, Germany

Pascal Puech CEMES, Toulouse, France

Robert Puers Department ESAT-MICAS, KULeuven, Leuven, Belgium

O. Pulci Institute for Complex Systems, National Research Council, Rome, Italy

J. S. Pulskamp US Army Research Laboratory, Adelphi, MD, USA

Aisha Qi Micro/Nanophysics Research Laboratory, RMIT University, Melbourne, VIC, Australia

Yabing Qi Energy Materials and Surface Sciences (EMSS) Unit, Okinawa Institute of Science and Technology, Onna-son, Kunigami-gun, Okinawa, Japan

Yongfen Qi The Key Laboratory of Remodeling-related Cardiovascular Diseases, Ministry of Education, Capital Medical University, Beijing, China

Qiquan Qiao Department of Electrical Engineering and Computer Science, South Dakota State University, Brookings, SD, USA

Hongwei Qu Department of Electrical and Computer Engineering, Oakland University, Rochester, MI, USA

Marzia Quaglio Center for Space Human Robotics, Fondazione Istituto Italiano di Tecnologia, Torino, Italy

Gaetana Quaranta Laboratoire d'Hydrologie et de Géochimie de la Surface, EOST, UDS (Université de Strasbourg), Strasbourg, France

Marta Quintanilla Institut National de la Recherche Scientifique – Énergie Matériaux et Télécommunications, Université du Québec, Varennes, QC, Canada

Regina Ragan Chemical Engineering and Materials, The Henry Samueli School of Engineering, Science University of California, Irvine, CA, USA

Mina Rais-Zadeh Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA

Melur K. Ramasubramanian Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC, USA

Antonio Ramos Departamento de Electrónica y Electromagnetismo, Universidad de Sevilla, Sevilla, Spain

Hyacinthe Randriamahazaka University Paris 7-Denis Diderot, ITODYS, Nanoelectrochemistry Group, UMR CNRS 7086, Paris, France

Kaushik K. Rangharajan Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, OH, USA

Apparao M. Rao Department of Physics and Astronomy, Clemson University, Clemson, SC, USA

Center for Optical Materials Science and Engineering Technologies, Clemson University, Clemson, SC, USA

Luca Razzari Institut National de la Recherche Scientifique – Énergie Matériaux et Télécommunications, Université du Québec, Varennes, QC, Canada

Stéphane Régnier Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie, Paris, France

E. Reina-Remo School of Engineering, University of Seville, Seville, Spain

Roberto Rella CNR-IMM (Institute for Microelectronics and Microsystems – Italian National Research Council), Lecce, Italy

Tian-Ling Ren Institute of Microelectronics, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, Haidian, China

Philippe Renaud Microsystems Laboratory, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland

Scott T. Retterer Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Department of Electrical Engineering and Computer Science, University of Tennessee Knoxville, Knoxville, TN, USA

Agneta Richter-Dahlfors Department of Neuroscience, Swedish Medical Nanoscience Center, Karolinska Institutet, Stockholm, Sweden

Michèle Riesen Department of Genetics Evolution and Environment, Institute of Healthy Ageing, University College London, London, UK

Matteo Rinaldi Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA

Aditi Risbud Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

José Rivas Laboratory of Magnetism and Nanotechnology, Institute for Technological Research, University of Santiago de Compostela, Santiago de Compostela, Spain

Paola Rivolo Dipartimento di Scienza dei Materiali e Ingegneria Chimica, Politecnico di Torino, Torino, Italy

Stephan Roche CIN2 (ICN–CSIC), Catalan Institute of Nanotechnology, Universidad Autónoma de Barcelona, Barcelona, Spain

Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Carsten Rockstuhl Institute of Condensed Matter Theory and Solid State Optics, Abbe Center of Photonics, Friedrich-Schiller-Universität Jena, Jena, Germany

Fernando Rodrigues-Lima Unit of Functional and Adaptive Biology (BFA), Laboratory of Molecular and Cellular Responses to Xenobiotics, UMR CNRS 8251, Univ Paris Diderot, (Sorbonne Paris Cité), Paris cedex 13, France

Brian J. Rodriguez University College Dublin, Dublin, Ireland

Shadi Rohani Institut National de la Recherche Scientifique – Énergie Matériaux et Télécommunications, Université du Québec, Varennes, QC, Canada

Giuseppe Romano Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Jérôme Rose CEREGE UMR 6635–CNRS–Université Paul Cézanne Aix–Marseille III, Aix–Marseille Université, Aix–en–Provence, France

GDRI ICEINT: International Center for the Environmental Implications of Nanotechnology, CNRS–CEA, Aix–en–Provence, France

Yitzhak Rosen Superior NanoBioSystems LLC, Washington, DC, USA

Fausto Rossi Department of Applied Science and Technology, Politecnico di Torino, Torino, Italy

Massimiliano Rossi Institute of Fluid Mechanics and Aerodynamics, Universität der Bundeswehr München, Neubiberg, Germany

Marina Ruths Department of Chemistry, University of Massachusetts Lowell, Lowell, MA, USA

Kathleen E. Ryan Department of Chemistry, United States Naval Academy, Annapolis, MD, USA

Malgorzata J. Rybak-Smith Department of Engineering Science, University of Oxford, Begbroke, UK

Maria Ilenia Saba Istituto Officina dei Materiali, Consiglio Nazionale delle Ricerche, CNR-IOM Cagliari (SLACS), Monserrato, CA, Italy

S. Siva Sankara Sai Department of Physics, Sri Sathya Sai Institute of Higher Learning, Vidyagiri, Puttaparthi, Andhra Pradesh, India

Ophélie Saison Institut d'Electronique, de Microélectronique et de Nanotechnologie, UMR CNRS 8520, Université Lille 1, Sciences et Technologies, Villeneuve d'Ascq, France

Verónica Salgueirino Departamento de Física Aplicada, Universidade de Vigo, Vigo, Spain

Meghan E. Samberg Arteriocyte, Inc., Cleveland, OH, USA

Florence Sanchez Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA

Claudio Sangregorio CNR-ICCOM (Institute of Chemistry of OrganoMetallic Compounds – Italian National Research Council), Florence, Italy

Catherine Santaella iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

Laboratoire d'Ecologie Microbienne de la Rhizosphère et d'Environnements Extrême, UMR 6191 CNRS–CEA–Aix–Marseille Université de la Méditerranée, CEA Cadarache, St Paul lez Durance, France

J. A. Sanz-Herrera School of Engineering, University of Seville, Seville, Spain

Stephen Andrew Sarles Mechanical Aerospace and Biomedical Engineering, University of Tennessee, Knoxville, TN, USA

M. Sbragaglia Department of Physics and INFN, University of "Tor Vergata", Rome, Italy

J. David Schall Department of Mechanical Engineering, Oakland University, Rochester, MI, USA

George C. Schatz Department of Chemistry and International Institute for Nanotechnology, Northwestern University, Evanston, IL, USA

Norbert Schell Institute of Materials Research, Helmholtz-Zentrum Geesthacht, Geesthacht, Schleswig-Holstein, Germany

Axel Schindler NTG Neue Technologien GmbH & Co. KG, Gelnhausen, Germany

André Schirmeisen Institute of Applied Physics, Justus-Liebig-University Giessen, Giessen, Germany

Frank Schmidt Zuse Institute Berlin, Berlin, Germany

Helmut Schmitz Institute of Zoology, University of Bonn Poppelsdorfer Schloss, Bonn, Germany

Scott R. Schricker Division of General Practice and Materials Science, College of Dentistry, The Ohio State University, Columbus, OH, USA

Georg Schulz Biomaterials Science Center (BMC), University of Basel, Basel, Switzerland

Udo D. Schwarz Department of Mechanical Engineering, Yale University, New Haven, CT, USA

Praveen Kumar Sekhar Electrical Engineering, School of Engineering and Computer Science, Washington State University Vancouver, Vancouver, WA, USA

Luca Sementa CNR-ICCOM, Consiglio Nazionale delle Ricerche, Pisa, Tuscany, Italy

Rita E. Serda Michael E. DeBakey Department of Surgery, Baylor College of Medicine, Houston, TX, USA

Department of NanoMedicine, Houston Methodist Research Institute, Houston, TX, USA

Nika Shakiba Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

Karthik Shankar Department of Electrical and Computer Engineering, W2-083 ECERF, University of Alberta, Edmonton, AB, Canada

Jonathan H. Shannahan Department of Pharmaceutical Sciences, University of Colorado – Anschutz Medical Campus, Aurora, CO, USA

Leon Shaw Wanger Institute for Sustainable Energy Research, Illinois Institute of Technology, Chicago, IL, USA

Department of Mechanical, Materials and Aerospace Engineering, Illinois Institute of Technology, Chicago, IL, USA

Yajing Shen Department of Mechanical and Biomedical Engineering, College of Science and Engineering, City University of Hong Kong, Kowloon, Hong Kong, China

Zhijian Shen Department of Materials and Environmental Chemistry, Arrhenius Laboratory, Stockholm University, Stockholm, Sweden

Li Shi Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX, USA

Yunfeng Shi Department of Materials Science and Engineering, MRC RM114, Rensselaer Polytechnic Institute, Troy, NY, USA

Youngmin Shin Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Satoru Shoji Department of Engineering Science, The University of Electro-Communications, Chofu, Tokyo, Japan

Wolfgang M. Sigmund Department of Materials Science and Engineering, University of Florida, Gainesville, FL, USA

Department of Energy Engineering, Hanyang University, Seoul, Republic of Korea

S. Ravi P. Silva Nano Electronics Center, Advanced Technology Institute, University of Surrey, Guildford, Surrey, UK

Robert B. Sim Department of Pharmacology, University of Oxford, Oxford, UK

Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, UK

Nipun Sinha Department of Mechanical Engineering and Applied Mechanics, Penn Micro and Nano Systems (PMaNS) Lab, University of Pennsylvania, Philadelphia, PA, USA

Dunja Skoko The Naughton Institute, School of Physics, Trinity College Dublin, CRANN, Dublin, Ireland

Nicholas Smith Biophotonics Lab, Immunology Frontier Research Center (IFReC), Osaka University, Osaka, Japan

Craig Snoeyink Birck Nanotechnology Center, Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Konstantin Sobolev Department of Civil Engineering and Mechanics, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Helmut Soltner Zentralabteilung Technologie, Forschungszentrum Jülich GmbH, Jülich, Germany

Thomas Søndergaard Department of Physics and Nanotechnology, Aalborg University, Aalborg Øst, Denmark

Youngjun Song Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA, USA

Anil K. Sood Gynecologic Oncology, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

Cancer Biology, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

Center for RNA Interference and Non–coding RNA, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA

The Department of Nanomedicine and Bioengineering, UTHealth, Houston, TX, USA

Jolanda Spadavecchia Laboratoire de Réactivité de Surfaces, UMR CNRS 7197, Université Pierre & Marie Curie – Paris VI, Ivry-sur-Seine, France

Pratheev S. Sreetharan School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Arati Sridharan School of Biological and Health Systems Engineering, Arizona State University, Tempe, AZ, USA

Bernadeta Srijanto Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Vaishnavi Srinivasaraghavan Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA

Tomasz Stapinski Department of Electronics, AGH University of Science and Technology, Krakow, Poland

A. Starace Department of Physics, Sapienza University of Rome, Rome, Italy

Ullrich Steiner Department of Physics, Cavendish Laboratories, University of Cambridge, Cambridge, UK

Michael S. Strano Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Arunkumar Subramanian Department of Mechanical and Nuclear Engineering, Virginia Commonwealth University, Richmond, VA, USA

S. Succi Istituto per le Applicazioni del Calcolo CNR, Rome, Italy

Maxim Sukharev Department of Applied Sciences and Mathematics, Arizona State University, Mesa, AZ, USA

Bobby G. Sumpter Computer Science and Mathematics Division and Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Bingbing Sun Division of NanoMedicine, Department of Medicine, University of California, Los Angeles, CA, USA

Daoheng Sun School of Aerospace Engineering, Xiamen University, Fujian, China

Hong-Bo Sun State Key Laboratory on Integrated Optoelectronics, College of Electronic Science and Engineering, Jilin University, Changchun, China

Tao Sun Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Yu Sun Department of Mechanical and Industrial Engineering and Institute of Biomaterials and Biomedical Engineering and Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada

Vishnu-Baba Sundaresan Mechanical and Nuclear Engineering, Virginia Commonwealth University, Richmond, VA, USA

Sabine Szunerits Institut d'Electronique, de Microélectronique et de Nanotechnologie (IEMN), UMR CNRS 8520, Université Lille 1, Villeneuve d'Ascq, France

Srinivas Tadigadapa Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Talal Al Tahtamouni Department of Physics, Yarmouk University, Irbid, Jordan

Saikat Talapatra Department of Physics, Southern Illinois University Carbondale, Carbondale, IL, USA

Qingyuan Tan Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada

Hiroto Tanaka School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

School of Engineering and Applied Sciences, Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA, USA

Yu-Hang Tang Division of Applied Mathematics, Brown University, Providence, RI, USA

Xinyong Tao College of Chemical Engineering and Materials Science, Zhejiang University of Technology, Hangzhou, China

Ennio Tasciotti Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

J. Ashley Taylor Department of Mechanical Engineering, The University of Wisconsin-Madison, Madison, WI, USA

Alain Thiéry iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

IMEP, UMR 6116 CNRS/IRD, Aix–Marseille Université, Marseille, France

Antoine Thill iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

Laboratoire Interdisciplinaire sur l'Organisation Nanométrique et Supramoléculaire, UMR 3299, CEA/CNRS SIS2M, Gif–sur–Yvette, France

Sylvia W. Thomas Department of Electrical Engineering, University of South Florida, Tampa, FL, USA

Thomas Thundat University of Alberta, Edmonton, AB, Canada

He Tian Institute of Microelectronics, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, Haidian, China

Katarzyna Tkacz–Smiech Faculty of Materials Science and Ceramics, AGH University of Science and Technology, Krakow, Poland

Steve To Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada

Gerard Tobias Institut de Ciència de Materials de Barcelona (ICMAB-CSIC), Bellaterra, Barcelona, Spain

Andrea Toma Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Katja Tonisch Institut für Mikro- und Nanotechnologien, Technische Universität Ilmenau Fachgebiet Nanotechnologie, Ilmenau, Germany

Elka Touitou Institute of Drug Research, School of Pharmacy, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

Helen E. Townley Department of Engineering Science, University of Oxford, Begbroke, UK

Lesa A. Tran Department of Chemistry, The Richard E. Smalley Institute for Nanoscale Science and Technology, Rice University, Houston, TX, USA

Ilenia G. Tredici Department of Chemistry, University of Pavia, Pavia, Italy

Alexander A. Trusov Department of Mechanical and Aerospace Engineering, The University of California, Irvine, CA, USA

Soichiro Tsuda School of Chemistry, University of Glasgow, Glasgow, UK

Kostiantyn Turcheniuk Institut d'Electronique, de Microélectronique et de Nanotechnologie (IEMN), UMR CNRS 8520, Université Lille 1, Villeneuve d'Ascq, France

Piero Ugliengo Department of Chemistry, University of Torino, Torino, Italy

Aeraj ul Haque Biodetection Technologies Section, Energy Systems Division, Argonne National Laboratory, Lemont, IL, USA

Federico Urzi Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Lorenzo Valdevit Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

Ana Valero Microsystems Laboratory, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland

Jao van de Lagemaat National Renewable Energy Laboratory, Golden, CO, USA

Renewable and Sustainable Energy Institute, Boulder, CO, USA

Pablo Varona Dpto. de Ingenieria Informatica, Universidad Autónoma de Madrid, Madrid, Spain

Daniele Varsano Center S3, CNR Institute of Nanoscience, Modena, Italy

K. Venkataramaniah Department of Physics, Sri Sathya Sai Institute of Higher Learning, Vidyagiri, Puttaparthi, Andhra Pradesh, India

Nuria Vergara-Irigaray Department of Genetics Evolution and Environment, Institute of Healthy Ageing, University College London, London, UK

Georgios Veronis Department of Electrical and Computer Engineering and Center for Computation and Technology, Louisiana State University, Baton Rouge, LA, USA

Fiorenzo Vetrone Institut National de la Recherche Scientifique – Énergie Matériaux et Télécommunications, Université du Québec, Varennes, QC, Canada

Achini K. Vidanapathirana Australian Research Council (ARC) Centre for Nanoscale Biophotonics, University of Adelaide, Adelaide, SA, Australia

Heart Health, South Australian Health and Medical Research Institute, Adelaide, SA, Australia

J. Vogt Kirchhoff Institute for Physics, Heidelberg University, Heidelberg, Germany

Alexey N. Volkov Department of Mechanical Engineering, University of Alabama, Tuscaloosa, AL, USA

Frank Vollmer Laboratory of Biophotonics and Biosensing, Max Planck Institute for the Science of Light, Erlangen, Germany

Fritz Vollrath Department of Zoology, University of Oxford, Oxford, UK

Prashant R. Waghmare Micro and Nano-scale Transport Laboratory, Department of Mechanical Engineering, University of Alberta, Edmonton, AB, Canada

Hermann Wagner Institute for Biology II, RWTH Aachen University, Aachen, Germany

Lucas K. Wagner Department of Physics, University of Illinois, Urbana, IL, USA

Richard Walker ICON plc, Marlow, Buckinghamshire, UK

Thomas Wandlowski Department of Chemistry and Biochemistry, University of Bern, Bern, Switzerland

Chunlei (Peggy) Wang Department of Mechanical and Materials Engineering, Florida International University, Miami, FL, USA

Enge Wang International Center for Quantum Materials, School of Physics, Peking University, Beijing, China

Evelyn N. Wang Device Research Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Feng-Chao Wang State Key Laboratory of Nonlinear Mechanics (LNM), Institute of Mechanics, Chinese Academy of Sciences, Beijing, China

Guoxing Wang School of Microelectronics, Shanghai Jiao Tong University (SJTU), Minhang, Shanghai, People's Republic of China

Jing Wang Department of Electrical Engineering, University of South Florida, Tampa, FL, USA

Max Zenghui Wang Department of Electrical Engineering and Computer Science, Case School of Engineering, Case Western Reserve University, Cleveland, OH, USA

Shengqin Wang Institute of Materials Research and Engineering, A*STAR (Agency for Science, Technology and Research), Singapore, Singapore

Szu-Wen Wang Chemical Engineering and Materials Science, The Henry Samueli School of Engineering, University of California, Irvine, CA, USA

Wenlong Wang Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Zhongguancun, Beijing, China

Xin Wang Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Xinwei Wang Department of Mechanical Engineering, Iowa State University, Ames, IA, USA

Yanchen Wang Electrical Mechatronics Engineering, Southern Polytechnic State University, Marietta, GA, USA

Yanwei Wang Mechanical Engineering, Harbin Institute of Petroleum, Purdue University, West Lafayette, IN, USA

Yu-Feng Wang Institute of Microelectronics, Tsinghua University, Beijing, China

Zhibin Wang Department of Materials Science and Engineering, University of Toronto, Toronto, ON, Canada

Zhong Lin Wang School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Reinhold Wannemacher Madrid Institute for Advanced Studies, IMDEA Nanociencia, Madrid, Spain

Benjamin L. J. Webb Division of Infection and Immunity, University College London, London, UK

Michael Weigel-Jech Division Microrobotics and Control Engineering (AMiR), Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Steven T. Wereley School of Mechanical Engineering, Purdue University, Birck Nanotechnology Center, West Lafayette, IN, USA

Tad S. Whiteside Savannah River National Laboratory, Aiken, SC, USA

John P. Whitney School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Samuel A. Wickline C-TRAIN Labs, Washington University, St. Louis, MO, USA

Mark Wiesner iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

CEINT, Center for the Environmental Implications of NanoTechnology, Duke University, Durham, NC, USA

Stuart Williams Department of Mechanical Engineering, University of Louisville, Louisville, KY, USA

Kerry Allan Wilson London Centre for Nanotechnology, University College London, London, UK

Lon J. Wilson Department of Chemistry, The Richard E. Smalley Institute for Nanoscale Science and Technology, Rice University, Houston, TX, USA

Patrick M. Winter Department of Radiology, Imaging Research Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Kerstin Witte Institute of Physics, University of Rostock, Rostock, Germany

Micromod Partikeltechnologie GmbH, Rostock, Germany

Stephen TC Wong Center for Bioengineering and Informatics, Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston, TX, USA

Robert J. Wood School of Engineering and Applied Sciences, Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA, USA

Matthew Wright Malvern Instruments, Malvern, Worcestershire County, UK

Dezhi Wu Department of Aeronautics, Xiamen University, Fujian, People's Republic of China

Xiao Yu Wu Advanced Pharmaceutics and Drug Delivery Laboratory, Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada

Yan Wu Department of Engineering Physics, University of Wisconsin – Platteville, Platteville, WI, USA

Tian Xia Division of NanoMedicine, Department of Medicine, University of California, Los Angeles, CA, USA

California NanoSystems Institute, University of California, Los Angeles, CA, USA

Guoqiang Xie Institute for Materials Research, Tohoku University, Sendai, Japan

H. Xie The State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, Harbin, China

Huikai Xie Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

Wei Xiong Department of Electrical Engineering, University of Nebraska-Lincoln, Lincoln, NE, USA

Yan Xiong School of Materials Science and Engineering, Hubei Province Key Laboratory of Green Materials for Light Industry, School of Materials, Hubei University of Technology, Wuhan, China

Department of Materials and Environmental Chemistry, Arrhenius Laboratory, Stockholm University, Stockholm, Sweden

Didi Xu Institute of Robotics and Intelligent Systems, ETH Zurich, Zurich, Switzerland

Shang Xu Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Kowloon, Hong Kong, P. R. China

Shen Xu Department of Mechanical Engineering, Iowa State University, Ames, IA, USA

Tingting Xu Department of Applied Chemistry, School of Science, Northwestern Polytechnical University, Xi'an, Shaanxi, People's Republic of China

Hiroshi Yamaguchi NTT Basic Research Laboratories, NTT Corporation, Atsugi-shi, Kanagawa, Japan

Christophe Yamahata Microsystem Lab., Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Chun Yang Division of Thermal Fluids Engineering, Nanyang Technological University, School of Mechanical and Aerospace Engineering, and School of Chemical and Biomedical Engineering, Singapore, Singapore

Jinling Yang Institute of Semiconductors, Chinese Academy of Sciences, Beijing, People's Republic of China

State Key Laboratory of Transducer Technology, Shanghai, People's Republic of China

Keqin Yang Department of Physics and Astronomy, Clemson University, Clemson, SC, USA

Yi Yang Institute of Microelectronics, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, Haidian, China

Yuehai Yang Department of Physics, Florida International University, Miami, FL, USA

Zhan Yang Jiangsu Key Laboratory for Advanced Robotics Technologies and Collaborative Innovation Center of Suzhou, Nano Science and

Technology, Soochow University, 215021, Suzhou, China

Yoke Khin Yap Department of Physics, Michigan Technological University, Houghton, MI, USA

Takashi Yatsui Department of Electrical Engineering and Information Systems School of Engineering, University of Tokyo, Tokyo, Japan

Alireza Yazdani Division of Applied Mathematics, Brown University, Providence, RI, USA

Iman K. Yazdi Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Leslie Yeo Micro/Nanophysics Research Laboratory, RMIT University, Melbourne, VIC, Australia

Junghoon Yeom Department of Mechanical Engineering, Michigan State University, East Lansing, MI, USA

Zheng Yin Center for Bioengineering and Informatics, Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston, TX, USA

Yaroslava G. Yingling Materials Science and Engineering, North Carolina State University, Raleigh, NC, USA

Minami Yoda G. W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Sang-Hee Yoon Department of Mechanical Engineering, Inha University, Incheon, Republic of Korea

Lidan You Department of Mechanical and Industrial Engineering and Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada

Yanlei Yu Department of Materials Science and State Key Laboratory of Molecular Engineering of Polymers, Fudan University, Shanghai, China

Wei Yu The Key Laboratory of Remodeling-related Cardiovascular Diseases, Ministry of Education, Capital Medical University, Beijing, China

Yanan Yue School of Power and Mechanical Engineering, Wuhan University, Wuhan, Hubei, People's Republic of China

Seok H. Yun Graduate School of Nanoscience and Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea

Wellman Center for Photomedicine, Department of Dermatolog, Harvard Medical School and Massachusetts General Hospital, Boston, MA, USA

The Harvard–Massachusetts Institute of Technology Division of Health Science and Technology, Cambridge, MA, USA

Mana Man Na Yung The Swire Institute of Marine Sciences and School of Biological Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China

Remo Proietti Zaccaria Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy

Dongyan Zhang Department of Physics, Michigan Technological University, Houghton, MI, USA

F. Zhang School of Materials Science and Engineering, Southeast University, Nanjing, China

Jin Z. Zhang Department of Chemistry and Biochemistry, University of California, Santa Cruz, CA, USA

John Xiaojing Zhang Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

Lijun Zhang Department of Mechanical Engineering, Shanghai Ocean University, Shanghai, People's Republic of China

Xiaobin Zhang Department of Materials Science and Engineering, Zhejiang University, Hangzhou, China

Yi Zhang Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai, China

Ze Zhang Institute of Microstructure and Property of Advanced Materials, Beijing University of Technology, Chaoyang District, Beijing, People's Republic of China

State Key Laboratory of Silicon Materials and Department of Materials Science and Engineering, Zhejiang University, Hangzhou, China

Jinfeng Zhao Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Ya-Pu Zhao State Key Laboratory of Nonlinear Mechanics (LNM), Institute of Mechanics, Chinese Academy of Sciences, Beijing, China

Gaofeng Zheng Department of Mechanical and Electrical Engineering, Xiamen University, Fujian, People's Republic of China

Leonid V. Zhigilei Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA, USA

Ying Zhou Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

Wenguang Zhu Department of Physics and Astronomy, The University of Tennessee, Knoxville, TN, USA

Yimei Zhu Center for Functional Nanomaterials, Brookhaven National Lab, Upton, NY, USA

Yingxi Zhu Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, USA

Rashid Zia School of Engineering, Brown University, Providence, RI, USA

A

Ab Initio DFT Simulations of Nanostructures

Antonio Aliano[1] and Giancarlo Cicero[1,2]
[1]Department of Physics, Politecnico di Torino, Torino, Italy
[2]Materials Science and Chemical Engineering Department, Politecnico di Torino, Torino, Italy

Synonyms

Electronic structure calculations; First principles calculations; Mean field approaches to many-body systems

Definition

Density Functional Theory (DFT) is an exact theory in which the electronic charge density is the basic quantity used to determinate the ground state properties of a many-body quantum system [1, 2]. Such a theory is an alternative formulation of the Schrödinger picture of quantum mechanics and represents a fundamental breakthrough in the application of computer simulation to material science and to nanosystems in particular. In the last decades, DFT-based analysis showed a quantitative predictive role and has been successfully employed in the study of molecules, crystals, and nanostructured materials. DFT calculations permit to investigate phenomena at an atomic scale hardly accessible by other techniques, emerging as a fundamental tool for complementing experimental investigations and for addressing open issues regarding nanomaterials and novel design processes [3, 4]. The DFT methodology is an ab initio (or first principles) approach, since no parameter characterizing the Hamiltonian of the system is tuned to empirical data and simulations results are thus free from empirical hints.

Introduction and DFT Foundations

The chemical and physical properties of materials both at the macro- and nano-scale strongly depend on the fundamental interactions holding together atoms and electrons present in the structure. A complete understanding of the material properties rests upon being able to accurately describe the electronic structure of the system, or, in other words, to solve the Schrödinger equation for the Hamiltonian characterizing the system. Unfortunately, solving this equation is not possible without making some physical assumptions that decrease the mathematical complexity of the problem. One of the most important approximations considered in the quantum treatment of materials consists of the so-called Born-Oppenheimer or adiabatic approximation: considering the three orders of magnitude difference between nuclei and electron masses (and De Broglie wavelengths), only the electrons are regarded as quantum particles and it is accepted

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

that the motion of the nuclei does not influence the electron dynamics (nuclei are considered fixed in space). Within this hypothesis, the description of correlated nuclei and electrons can be decoupled so that the wave function Ψ of an N-electron system depends only on the electron coordinates, namely, Ψ (r_1, r_2, …, r_N) = Ψ ({r}), and the nuclei positions enter only as parameters. The Hamiltonian operator $\hat{H}$ for the electrons at a fixed nuclear configuration can be written in the compact form:

$$\hat{H} = \hat{T} + \hat{V}_{ee} + \hat{V}_{ext} \tag{1}$$

where $\hat{T}$ is the kinetic energy operator, $\hat{V}_{ee}$ the operator corresponding to the electron–electron interactions, and $\hat{V}_{ext}$ the external potential operator typically defined by the Coulomb attractive potential due to nuclei. Despite the simplifications introduced, when the form of the interaction potentials are known, the many-body Schrödinger equation $\hat{H}\Psi = E\Psi$ for the electrons cannot be solved analytically and it is not numerically manageable because of the prohibitive number of degrees of freedom. Density Functional Theory, which was elaborated in 1960s by Hohenberg and Kohn, together with the ansatz made by Kohn and Sham has provided a way to obtain the ground state properties for real systems of many electrons, starting from the general Hamiltonian reported in Eq. 1.

Theoretical Background

Density Functional Theory (DFT) takes its historical root in the Thomas-Fermi model [1] and has its solid mathematical justification in the Hohenberg-Kohn theorems [1, 2]. These theorems and their corollaries prove rigorously that the electronic charge density, $\rho(r)$, can be used as the only basic variable to obtain the ground state properties of a system of N interacting electrons moving in an external potential $v_{ext}(r)$. In particular, it can be demonstrated that the total energy of a quantum system is a functional of the electron density $\rho(r)$ only: knowing the electron density explicitly implies having a complete physical description of the system and it is equivalent to know its ground state wave function ($\psi_0[\rho]$). The energy functional can be written as:

$$\begin{aligned} E[\rho] &= E[\rho] + \int v_{ext}(r)\rho(r)dr \\ \text{with } F[\rho] &= T[\rho] + V_{ee}[\rho] \end{aligned} \tag{2}$$

In Eq. 2, the functional $E[\rho]$ is organized in two parts: a term depending on the external potential or, in other words, by the specific material and structure under investigation, and the functional $E[\rho]$ which is the same for all the physical systems (atoms, molecules, solids, or nanostructures) and it is called universal functional. In quantum mechanics the electron charge density $\rho(r)$ is defined as:

$$\rho(r) = N\int \dots \int |\Psi(\{r\})|^2 dr_2 \dots dr_N \tag{3}$$

and obeys the condition $\int \rho(r)dr = N$ in order to represent the number of electrons per unit volume for a given state (in Eq. 3 and throughout the chapter atomic units are used). The second Hohenberg-Kohn theorem states that the exact ground state energy of a system is the global minimum value of the functional $E[\rho]$ and that the density that minimizes this functional corresponds to the exact ground state density ρ_0. In this context, if ρ_0 is unknown, the best estimate of the ground state properties can be obtained following a variational approach by minimizing the energy functional $E[\rho]$ calculated with respect to a trial density. The latter condition can be expressed considering that also the wave function of the system is a functional of the density, and the total energy of the system can be written as:

$$E[\rho] = \langle \Psi[\rho]\hat{H}|\Psi|[\rho]\rangle \tag{4}$$

the ground state charge density ρ_0 results from:

$$\rho_0 = \min_{\rho} E[\rho]. \tag{5}$$

The Hohenberg and Kohn formulation has in principle a very general validity but it is impractical since the exact form of the universal

functional $E[\rho]$, is not known. A decisive step toward its practical applications was made possible thanks to the ansatz by Kohn-Sham (KS); according to their hypothesis, the interacting electron systems can be replaced with an auxiliary system of independent electrons characterized by the same ground state charge density, ρ_0, as the original one. For the auxiliary system the energy functional, $E[\rho]$, can be written as:

$$E[\rho] = T_s[\rho] + \frac{1}{2}\int\int \frac{\rho(r)\rho(r')}{|r-r'|} dr dr' + \int v_{ext}\rho(r)dr + E_{xc}[\rho] \tag{6}$$

where $T_s[\rho]$ is the kinetic energy term of noninteracting electrons, $1/2\int\int \rho(r)\rho(r')/|r-r'|drdr'$ is a classical term representing the Coulomb repulsion between two continuous charge distributions (Hartree term), while $E_{xc}[\rho]$ is the "exchange and correlation" term in which all the many-body effects of the original interacting system are included. For the noninteracting electrons, the wave function is expressed by $\Psi = \psi(r_1)\psi(r_2)\ldots\psi(r_N)$ and the charge density $\rho(r)$ becomes

$$\rho(r)\sum_i f_i|\psi_i|^2, \mathrm{i} = 1, \ldots, \mathrm{N} \tag{7}$$

where ψ_i are the single-particle wave functions and f_i are the occupation numbers. Once the density of the auxiliary system is obtained, one knows also the ground state density of the original interacting system. The minimization of Eq. 6 with respect to ρ yields a set of N ground state single-particle Schrödinger-like equations that reads:

$$\left(-\frac{1}{2}\nabla^2 + \int\frac{\rho(r')}{|r-r'|}dr' + v_{ext}(r) + \frac{\delta E_{xc}[\rho]}{\delta\rho(r)}\right)\psi_i^{KS} = \varepsilon_i^{KS}\psi_i^{KS}, i = 1, \ldots, N \tag{8}$$

It is worth to remark that the eigenvalues (ε_i^{KS}), and the wave functions (ψ_i^{KS}) calculated in Eq. (8) are referred to a Kohn-Sham system and are not those relative to the real system of interacting electrons, yet, following Eq. 7, ψ_i^{KS} are used to obtain the correct density. The main intricacy of Eqs. 6 and 8 is related to the fact that the explicit dependence of $E_{xc}[\rho]$ on the electron density is not known, thus, even if the KS ansatz has reduced the complexity of the original many-body problem, the correct ground state is still not achievable. Following the work by KS, many approximations to $E_{xc}[\rho]$ have been proposed in literature and these are usually classified as a "Jacob's ladder of DFT" [5] which goes from lower accuracy (lower steps) to the exact unknown form (highest step). The simplest construction of $E_{xc}[\rho]$ is represented by the Local Density Approximation (LDA) in which $E_{xc}[\rho]$ is taken so that $\delta E_{xc}[\rho]/\delta\rho$ is equal to the exchange and correlation energy density of a homogeneous electron gas, $\varepsilon_{xc}^{\mathrm{hom}}$, at the density of the auxiliary system at point r of space. The quantity $\varepsilon_{xc}^{\mathrm{hom}}(\rho(r))$ has an analytical expression, that, once substituted in Eq. 8 allows finding the eigenfunctions of the auxiliary system. Many approximations other than LDA are available for $E_{xc}[\rho]$; one of the most commonly used consists of the Generalized Gradient Approximation (GGA) in which the exchange and correlation energy density term has an explicit dependance also on the density gradient $|\nabla\rho(r)|$. Several expressions for such dependence have been proposed in literature. Both LDA and GGA functionals suffer from being approximations of the true energy functional and, for this reason, they fail in describing physical phenomena in which electron correlation has an important role as for example Van der Waals forces or strongly correlated systems such as the d-shells of transition metals. To overcome the limits of KS-DFT formulation, several post-DFT methods have been developed; for example, the LDA + U (or GGA + U) approximation in which an on-site Hubbard-like correction (U) is added to the strongly correlated electrons [6], or the GW approach in which Green functions are used to evaluate self-energy corrections on single particle eigenvalues [7]. Finally, we mention that DFT, which is a ground state theory, has been extended to a time dependent formulation (TD-DFT) to provide a solid methodology to study the excited state properties of a many-body system in the presence of an external time dependent potential [8].

Numerical Details

Once the Exchange and Correlation term is made explicit, the KS wave functions ψ_i^{KS} (and the ground state charge density ρ_0) can be obtained by solving self-consistently the set of nonlinear differential equations (8). In the iterative solution procedure, the wave function is generally represented as a linear combination of basis set functions yielding a matrix representation of the Hamiltonian. This choice transforms the solution of partial integro-differential equations into a diagonalization problem, for which many standard algorithms have been largely tested and developed. Traditionally many DFT codes, implemented to calculate the electronic structure of materials, have been developed to deal with crystalline solids, that is, with periodic arrangements of atoms in a Bravais lattice and consequently yielding $v_{ext}(r)$ appearing in Eq. 8 also periodic. In this case, according to the Bloch's theorem [9], the KS eigenfunctions can be expressed as Bloch states denoted by a wave vector $\vec{k}$ within the first Brillouin Zone (BZ). A convenient basis set to expand these wave functions consists of the plane waves (PWs) basis set. Since the system is periodic, not all the PWs appear in the expansion but only those having a wave vector belonging to the reciprocal lattice of the crystal. In principle one would have to include an infinite number of functions to express the solution of the KS eigenvalue equation; in practice it is enough to limit the expansion to a finite subset of the basis set space once the accuracy of the subset has been proved. In the case of nonperiodic systems, like molecules, surfaces, or nanostructures, one can still effectively adopt a finite PWs basis set and use all the established numerical strategies deviced for 3-D periodic structures, if one adopts a supercell approach [2]. With the term "supercell" one indicates a 3-D periodic cell that contains the nonperiodic (or partially periodic) structure and a vacuum region. The cell sizes along the periodic directions correspond to the lattice parameters of the structure in those directions, while in the other directions the cell size has to be large enough to accommodate the finite structure and the vacuum. Periodic boundary conditions are applied to the supercell, in this way the structure is repeated in space even if it is not fully periodic; if the vacuum size is large enough the replicas along the nonperiodic directions can be considered to be isolated. Indeed, the increment of the cell size to create a reliable system increases the computational cost, thus a compromise has to be reached to obtain the desired accuracy. One of the most used methods, employed to decrease the size of the basis set (and thus reducing the computational cost) is represented by the so-called "pseudopotential" approximation [10]. According to this approximation the electrons tightly bound to the nuclei and not involved in bond formation (core electrons) are "frozen" into the core and their effect on the other electrons (valence electrons) is included in an effective way into the external potential ($v_{ext}(r)$). The pseudopotential represents an electrostatic potential that contains the effect of the bare ionic charge screened by the core electrons and it is characterized by pseudo-wavefunctions that, in the core region, are smoother than the wave functions of the bare nuclear potential and, for this reason, require less PWs to be expanded.

Once the solutions of the KS equations are obtained, it is possible to calculate any physical quantity of interest for the investigated system. For the case of periodic structures many properties such as the counting of electrons in bands, total energy, density, etc., are obtained by summing over all the states labeled by $\vec{k}$ in the BZ. For a general function $F_i(\vec{k})$ associated with a certain physical quantity, the average value $\overline{F}_i$ is defined as:

$$\overline{F}_i = \frac{1}{\Omega_{BZ}} \int_{BZ} F_i\left(\vec{k}\right) d\,\vec{k} \tag{9}$$

where i denotes discrete states for each $\vec{k}$, Ω_{BZ} is the BZ volume, and the integral is extended over the entire BZ. In computer codes, accurate integration is performed by summation over discrete set of points in the BZ. One of the methods that has been largely used to sample the BZ is the one proposed by Monkhorst and Pack [2]. This method consists of constructing a uniform grid of points indicated by three integer numbers (n_1, n_2, n_3) that define the mesh discretization

along the reciprocal space axis. If these points are chosen carefully and the grid is fine enough, the error in the estimate of $\overline{F}_i$ is small and the integral is well approximated by the summation. Finally once the ground state density of the KS system is known, according to the Hellmann-Feynman theorem [2], it is possible to calculate analytically the forces acting on each atom of the structure. Thus, if the structure is not at its equilibrium geometry, atoms can be moved and systems relaxed following the calculated forces until a selected threshold value is reached.

DFT Applications to Nanostructures

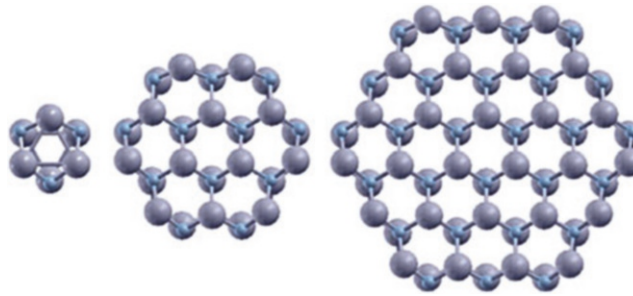
In the following the application of DFT to the investigation of the structural and electronic properties of a particular type of nanostructures, namely, one-dimensional nanowires (NWs) is presented. Understanding how the properties of a material change when it is reduced to a nanostructure is important for a wide range of scientific and technological applications as well as for basic science. The reduction to nanometers in the size of a semiconductor [11] demands theoretical approaches based on an atomistic description of a material; first principles–based approaches have demonstrated to be capable of exploring the electronic structure of bulk materials and small nanostructures with highly predictive character reaching a degree of detail that can hardly be achieved with experimental techniques. In particular, DFT simulations have played a major role in the study of electron confinement, and in determining how the band gap, the optical properties, and the electronic transport change at the nanoscale. In the last years, many different materials have been studied ranging from low (e.g., Ge, Si, InN) to wide band gap (e.g., ZnO, SiC) semiconductor, in the form of both simple and core/shell wires. Hereinafter the results for simple nanowires are discussed for a particular type of a technologically important material: Indium Nitride (InN). More in detail the effects of size reduction on the energy gap of the material are presented and the changes induced by surface termination on the electronic properties of the nanostructure are analyzed.

DFT Simulations of InN Nanowires

Group-III nitrides represent a material class with promising electronic and optical properties [12]. Among these, InN exhibits the narrowest band gap of about 0.67 eV at low temperatures, the lowest effective electron mass, and the highest peak drift velocity and electron mobility. InN nanowires have been proposed for applications in new generation integrated systems like quantum wire transistors with the aim of reducing the power consumption in large-scale integrated circuits. Due to their high surface to volume ratio, nanowires are also potential candidates for sensor applications and they have been proposed as key elements in new generation solar energy harvesting devices such as dye sensitized or other hybrid solar cells. A crucial role for applications of InN nanowires as novel nanoelectronic devices is played by a quasi-two-dimensional electron accumulation layer around the wire. To fully exploit the potential of InN nanowires, their physical properties have to be understood in depth.

Here the LDA and LDA + U results are discussed for wurtzite InN NWs [13] of increasing diameter with both clean and hydrogenated surfaces of three sizes (see Fig. 1): NW1 has one atom ring and a diameter d ≈ 4.1 Å, NW2 has two rings and a diameter d ≈ 10.6 Å, while NW3 has three rings and a diameter d ≈ 17.6 Å. By construction, these NWs have hexagonal cross-section and present nonpolar $(1\overline{1}00)$ facets, in agreement with experimental observation. Typically for a DFT simulation, due to the numerical approximations adopted to solve the KS equations presented previously, accuracy of the results has to be checked carefully once the following computational conditions have been chosen:

- Pseudopotentials needed to describe the external potential felt by the valence electrons.
- Size of the plane waves basis set employed to expand the one-electron wave function. This is expressed by an energy cutoff that corresponds to the highest kinetic energy (lowest wave vector) of the PWs included in the expansion.
- Convergence criteria on the total energy of the system and on the forces acting on the atoms in



Ab Initio DFT Simulations of Nanostructures, Fig. 1 Ball and stick representation of the investigated InN nanowires: *left* – 1 ring NW (d ≈ 4.1 Å), *middle* – 2 rings NW (d ≈ 10.6 Å), *right* – 3 rings NW (d ≈ 17.6 Å). *Large balls* represent In atoms, *small balls* represent N atoms

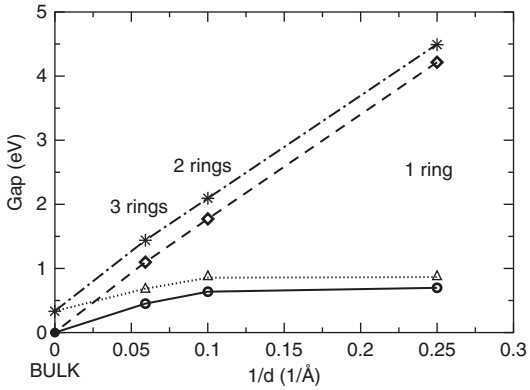
the case of structural relaxation based on Hellman-Feynman forces.
- Accuracy of the Monkhorst-Pack grid used to sample the Brillouin Zone of the periodic structure.
- Vacuum size in the supercell, in the case of structures that are nonperiodic in one or more dimensions (e.g., surfaces, nanowires, quantum dots, etc.).

To simulate InN 1-D nanostructures previously mentioned, the supercell approach was used: in this case the nanostructures are periodic only along the NW axis (*z*), thus large supercells have to be employed to accommodate enough vacuum space surrounding the wire (i.e., ∼ 10 Å) in the directions perpendicular to the NW axis. The integration over the Brillouin Zone was performed using a (1 × 1 × 6) k-point grid. All the structures were fully relaxed until the forces acting on the atoms were less than 10^{-3} Ry/bohr. The lattice parameter along the direction of the NW axis can be optimized by varying the cell size along that direction and relaxing all the atomic positions. The minimum energy structure is then obtained by fitting the total energy values with a polynomial fit. It has been found that in nanowires two main relaxation mechanisms occur: the lattice parameter along the NW axis changes with respect to the bulk value and the atoms at the NW surfaces relax.

The lattice constant along the NW axis increases decreasing the NW diameter and, except for the smallest diameter NW1, the behavior appears to be linear. While the NWs elongate along their axis, the lattice contracts in the perpendicular directions. This behavior has been observed also for other more ionic compound such as ZnO [14]. After relaxation, changes of InN bonds occur at the nanowire surface: the InN bond length becomes 2.05 Å (shorter than InN length in bulk of ∼ 0.1 Å) and it tilts by about 9.5°, with the In atoms relaxing inward. Upon hydrogenation the In–N bond length recovers its bulk value (∼2.18 Å), but the In–N bond tilting is now opposite to the clean case (∼5–6°), due to the moving out of the In atom.

In Fig. 2, the band gap dependence on the inverse of the NW diameter (1/d) for NWs with clean and hydrogenated surfaces is reported. It can be noted that in both LDA and LDA + U approaches the dependence of quantum confinements is almost linear and that the inclusion of the Hubbard (U) correction determines a rigid shift of the band gap (about 0.2 eV for NWs with clean surfaces and 0.35 eV for NWs with hydrogenated surfaces) that depends only slightly on the NW diameter. The larger energy gaps observed in the case of nanowires with hydrogenated surfaces imply that quantum confinement effects are more noticeable in passivated NWs and that one might expect to see spectroscopic differences in experiments performed in different environments (e.g., solution vs ultra high vacuum). LDA and LDA +U predict similar confinement effects also in terms of charge density distribution when comparing nanowires with clean and hydrogenated surfaces.

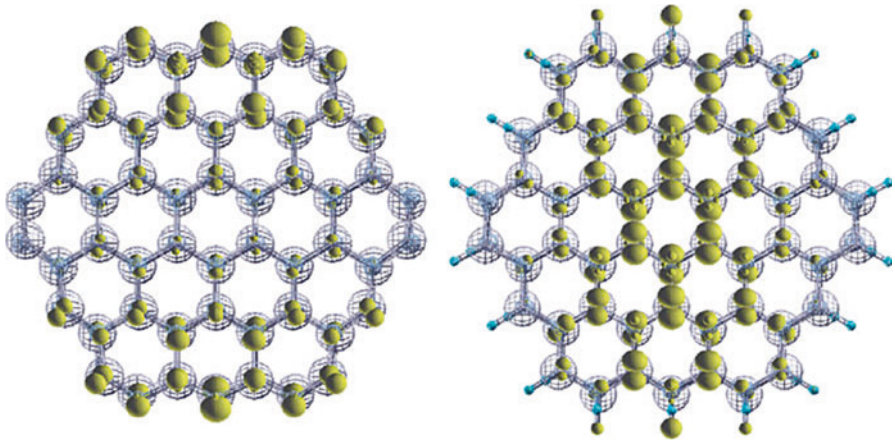
In the latter case, due to surface dangling bonds saturation, the electronic charge density accumulates in the center of the wire, while in nanowires with clean surfaces the density is more delocalized on the surface atoms. This effect is visible in Fig. 3 where the electronic charge density isosurface of the highest occupied state for NW3 is shown for clean and hydrogenated surfaces.



Ab Initio DFT Simulations of Nanostructures, Fig. 2 Band gap versus inverse diameter (1/d) of InN nanowires with clean surface indicated by *solid* (LDA) and *dotted* (LDA + U) *lines*, and for hydrogenated nanowires indicated by *dashed* (LDA) and *dashed-dotted* (LDA + U) *lines*

Conclusion

In conclusion, first principles DFT electronic structure calculations can be effectively used to systematically investigate the relative stability and electronic properties of nanostructures with clean, reconstructued, or passivated surfaces. In the case of Nanowires one can also consider the effects of the growth direction and of the diameter. Electronic properties of NWs, including band gaps, band structures, and effective masses, are usually found to depend sensitively on all the NW structural parameters. In many cases, the size dependence of the band gap depends on the growth direction of the NW, and the band gaps for a given size also depend on surface structure; moreover NWs with clean surfaces have HOMO orbitals localized on the facets and exhibit smaller band gaps. DFT calculations demonstrate the extremely rich nature of the electronic properties of nanostructures and NWs in particular and have a unique role in describing confinement effects in these small dimension systems. By changing the growth direction and NW diameter, or tuning the surface structure via chemical or physical methods, it is possible to engineer NWs with a wide range of technologically important electronic properties.



Ab Initio DFT Simulations of Nanostructures, Fig. 3 Electronic charge density isosurface (denoted by shaded *yellow-green* area) for the highest occupied state: *left panel* – NW3, *right panel* – NW3 + H. Sketched globes represent atoms

Cross-References

- ▶ Hybrid Solar Cells
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanostructured Materials for Sensing
- ▶ Optical and Electronic Properties
- ▶ Surface Electronic Structure
- ▶ Theory of Optical Metamaterials

References

1. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. Oxford University Press, Oxford, UK (1989)
2. Martin, R.M.: Electronic Structure-Basic Theory and Practical Methods. Cambridge University Press, Cambridge, UK (2008)
3. Hafner, J., Wolverton, C., Ceder, G.: Towards computational material design: the impact of density functional theory on material research. MRS Bull. **31**, 659 (2006)
4. Marzari, N.: Realistic modeling of nanostructures using density functional theory. MRS Bull. **31**, 681 (2006)
5. Perdew, J.P., Shmidt, K.: Density Functional Theory and Its Application to Materials, p. 1. American Institute of Physics, New York (2001)
6. Anisimovy, V.I., Aryasetiawanz, F., Lichtenstein, A.I.: First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA + U method. J. Phys. Condens. Matter **9**, 768 (1997)
7. Onida, G., Reining, L., Rubio, A.: Electronic excitations: density-functional versus many-body Green's-function approaches. Rev. Mod. Phys. **74**, 601 (2002)
8. Marques, M.A.L., Ullrich, C.A., Nogueira, F., Rubio, A., Burke, K., Gross, E.K.U.: Time-Dependent Density Functional Theory. Springer, Berlin/Heidelberg (2006)
9. Ashcroft, N.W., Mermin, N.D.: Solid State Physics. Saunders College Publishing, Philadelphia (1976)
10. Bachelet, G.B., Hamann, D.R., Schlüter, M.: Pseudopotentials that work: from H to Pu. Phys. Rev. B **26**, 4199 (1982)
11. Law, M., Goldberger, J., Yang, P.: Semiconductor nanowires and nanotubes. Annu. Rev. Mater. Res. **34**, 83 (2004)
12. Wu, J.: When group-III nitrides go infrared: new properties and perspectives. J. Appl. Phys. **106**, 011101 (2009)
13. Terentjevs, A., Catellani, A., Prendergast, D., Cicero, G.: Importance of on-site corrections to the electronic and structural properties of InN in crystalline solid, nonpolar surface, and nanowire forms. Phys. Rev. B **82**, 165307 (2010)
14. Cicero, G., Ferretti, A., Catellani, A.: Surface induced polarity inversion in ZnO nanowires. Phys. Rev. B **80**, 201304(R) (2009)

Ab Initio Modeling and Simulation of Metal Nanoparticles

▶ Metal Nanoparticles from First Principles

Ab Initio Molecular Dynamics

▶ Car–Parrinello Molecular Dynamics

AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation

Boris Khusid
Department of Chemical, Biological and Pharmaceutical Engineering, New Jersey Institute of Technology, Newark, NJ, USA

Synonyms

Dielectrophoretic phenomena; Dipolar particle chaining; Dipole–dipole particle interaction; Field-driven particle assembly; Suspension interaction with external AC field gradient

Definition

The total force exerted on a polarizable particle by a spatially nonuniform electrical field is the sum of the electrophoretic force (product of the net particle charge and the field strength) and the dielectrophoretic force (product of the particle dipole moment and the field strength gradient). Electrophoretic force effects vanish in an alternating current (AC) field of a sufficiently high frequency due to the zero time average over the field oscillations. In contrast, dielectrophoretic force operates in AC fields as its averaging yields a

nonzero value whose magnitude is the product of the particle volume, the gradient of the time-average squared field strength, and the relative particle polarization at the field frequency. Depending on whether the particle is more or less polarizable than the suspending fluid, it migrates toward the regions of high field strength (positive dielectrophoresis) or low field strength (negative dielectrophoresis).

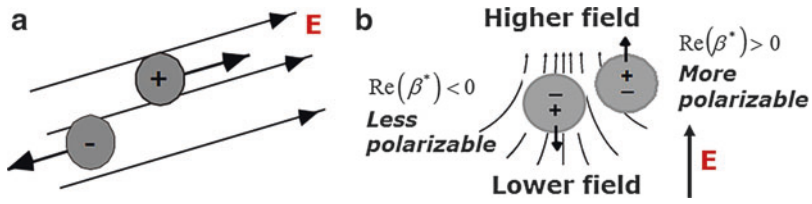
AC dielectrophoresis is becoming one of the major techniques for micro- and nanoscale systems. As one of its major features is the ability to concentrate particles, AC dielectrophoresis in suspensions is controlled not only by the interaction of particles with an externally applied field but also by the dipole–dipole forces between the polarized particles which acquire dipole moments as exposed to an electrical field. The main effect of the interparticle dipolar interactions is that they, if sufficiently strong, cause a suspension to undergo reversible transitions from a random arrangement of particles into a variety of ordered aggregation patterns. This trend has been long recognized and exploited in electrorheology to initiate the fluid–solid transition by applying a sufficiently strong field to a suspension. Control of multiparticle interactions driven by the interplay of dielectrophoretic and dipolar forces was shown to provide a powerful way to manipulate suspensions.

Overview

Dielectrophoresis, the term coined by Pohl [1], is the motion of an object under the force resulting from an electrical field gradient. Application of AC dielectrophoresis is based on differentiation of dielectric and conducting properties of objects to allow their separation and identification. To enable the dielectrophoretic manipulation of particles and macromolecules, the applied AC field should be sufficiently strong to dominate the thermal Brownian motion. However, a high electrical field also gives rise to undesirable processes in the suspending fluid, such as electrolysis, electroosmosis, thermal convection, and electrothermal flow caused by the variation in the fluid conductivity, permittivity, and density due to Joule heating. Estimation of the orders of magnitude of the various forces experienced by a particle and a fluid in dielectrophoretic micro-fluidics indicates that AC dielectrophoresis would prevail in small-scale devices and for sufficiently high field frequencies.

Compared to other available methods, AC dielectrophoresis is becoming one of the major methods for control and manipulations of particles and macromolecules in micro- and nanofluidics. Its main advantages include the use of an electrical field which requires no moving parts and employing polarization forces acting on a fine particle that are insensitive to the particle charge which is difficult to control. Therefore, dielectrophoresis-based techniques offer easy adaptability to electronics and can be incorporated more favorably into micro-systems to integrate multiple analytical methods down to chip format. A wide variety of dielectrophoretic micro-devices have been developed to address challenges in materials science, analytical chemistry, and life science, such as assembling nanoparticles, separating polydisperse particle suspensions into homogeneous subpopulations, manipulating and concentrating biologically relevant molecules, distinguishing dead and living cells or ill and healthy cells, etc. [2].

Since particles undergoing dielectrophoresis acquire a dipole moment, they interact with one another through dipole–dipole forces. It is well known that, because of dipolar interactions, polarized particles have a tendency to align rapidly along the field direction and to form chains that subsequently coalesce into thicker columns. This phenomenon causes a colloidal dispersion to undergo a reversible fluid–solid transition and exhibit a yield stress at low shear rates (often referred to as electrorheological effect) [3]. Recent studies reveal that, due to the particle confinement at the microscale, synergism of dielectrophoretic and dipole forces provides a powerful way to control and manipulate particles in micro-devices and extends the use of dielectrophoresis to the nanoscale for various applications ranging from nano-biotechnology to molecular electronics.



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 1 (**a**) Electrophoresis. (**b**) Dielectrophoresis

Basic Methodology

Electrical Force Acting on an Individual Particle

The force exerted by an electrical field on a particle consists of two terms:

$$\mathbf{F} = Q\mathbf{E} + (\mathbf{P} \cdot \nabla)\mathbf{E} \tag{1}$$

where Q is the particle charge and $\mathbf{P}$ is the particle dipole moment. The first term in Eq. 1 is the *electrophoretic force*, which causes a charged particle subjected to a direct current (DC) field to travel along the electrical lines (Fig. 1a), whereas the other term is referred to as the *polarization force* or very often as the *dielectrophoretic force*. In contrast to the electrophoretic force, the latter acts on both neutral and charged particles. Equation 1 represents the first terms in the expansion of the net electrical force acting on the induced charges in a polarized particle as a power series in the field gradient. When the particle dimensions become comparable to a characteristic length scale of an array of microelectrodes, the presence of the particle modifies the overall electrical field, and the dipole approximation in Eq. 1 is unreliable. The significance of higher-order terms in Eq. 1 for different electrode configurations is analyzed through comparison against the net force computed by integrating the Maxwell stress tensor over the particle surface [4]. Higher-order terms in the net electrical force exerted on a nonspherical particle (which includes the majority of biological particles) appear to be more significant due to the shape dependence of the particle polarization.

The polarization of a not perfectly insulating or perfectly conducting particle depends on the time variation of the applied electrical field over some characteristic time interval in the recent past. In particular, for a suspended sphere when the speed with which charge carriers rearrange themselves in response to an applied field is independent of the field strength, it becomes

$$\mathbf{P} = 3v_p\varepsilon_0\varepsilon_f \int_{-\infty}^{t} \overline{\beta}(t-t')\mathbf{E}(t')dt' \tag{2}$$

where v_p is the particle volume, ε_0 is the vacuum permittivity, ε_f is the fluid dielectric constant, and $\overline{\beta}(t)$ describes the dielectric relaxation. The real and imaginary parts of the Fourier transform of $\overline{\beta}(t)$ describe the interaction of a particle with an oscillating (AC) field of the frequency $\omega = 2\pi v$ where v is given in Hz:

$$\mathrm{Re}(\beta^*) + i\mathrm{Im}(\beta^*) = \int_0^{\infty} \overline{\beta}(t)e^{-i\omega t}dt \tag{3}$$

A model of a perfect dielectric assumes that the polarization responds instantaneously to an applied field so that $\overline{\beta}(t) = C_1\delta(t)$ with some constant C_1 and $\delta(t)$ being the Dirac delta function. In this case, Eq. 3 yields $\mathrm{Re}(\beta^*) = C_1$ and $\mathrm{Im}(\beta^*) = 0$. The opposite extreme of a perfectly conducting material corresponds to the polarization proportional to the rate of the time variation of an applied field so that $\overline{\beta}(t) = C_2\delta'(t)$ with some constant C_2 and $\delta'(t)$ being the derivative of the delta function, in this case, $\mathrm{Re}(\beta^*) = 0$ and $\mathrm{Im}(\beta^*) = \omega C_2$. When the electrical properties of the particle and that of the suspending fluid (i.e., their dielectric constants, ε_p and ε_f, and conductivities, σ_p and σ_f) are frequency independent, the Maxwell–Wagner theory [5, 6] yields the following expression for β^* often referred to as the Clausius–Mossotti factor:

$$\mathrm{Re}(\beta^*) = \frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_2 + 2\varepsilon_1} + \frac{3(\varepsilon_1\sigma_2 - \varepsilon_2\sigma_1)(\sigma_2 + 2\sigma_1)}{\left[(\sigma_2 + 2\sigma_1)^2 + \omega^2(\varepsilon_2 + 2\varepsilon_1)^2\right](\varepsilon_2 + 2\varepsilon_1)} \quad (4)$$

$$\mathrm{Im}(\beta^*) = -\frac{3\omega(\varepsilon_1\sigma_2 - \varepsilon_2\sigma_1)}{(\sigma_2 + 2\sigma_1)^2 + \omega^2(\varepsilon_2 + 2\varepsilon_1)^2}. \quad (5)$$

Similar expressions are available for ellipsoidal particles [5, 6]. Equations 4 and 5 indicate that a particle responds according to its dielectric constant at sufficiently high frequencies and conductivity at sufficiently low frequencies and DC field:

$$\mathrm{Re}(\beta^*) \to \frac{\varepsilon_2 - \varepsilon_1}{\varepsilon_2 + 2\varepsilon_1} \text{ and } \mathrm{Im}(\beta^*) \to 0 \text{ as } \omega \to \infty$$

$$\mathrm{Re}(\beta^*) \to \frac{\sigma_2 - \sigma_1}{\sigma_2 + 2\sigma_1} \text{ and } \mathrm{Im}(\beta^*) \to 0 \text{ as } \omega \to 0$$

There have been extensive theoretical and experimental studies of polarization of particles and macromolecules subjected to spatially uniform electrical fields [5]. It is found that the polarization of a fine particle and a macromolecule dispersed in a fluid strongly depends on its structure, fluid properties, and interface characteristics. There are no simple models containing a few adjustable parameters capable to characterize the dependence of Re(β*) and Im(β*) on the applied frequency and electrical and physical properties of the particle and the fluid. Dielectric relaxation spectroscopy that employs spatially uniform low fields (~1 V/mm) is used to measure the polarization of a colloidal dispersion to obtain data on the frequency dependence of an individual particle polarizability. Specifically, β* can be calculated from experimental data on the concentration and frequency dependence of the complex dielectric permittivity of the colloidal dispersion, ε_s^*, and the suspending fluid, ε_f^*, using the Maxwell-Garnett expression (Eq. 6) (often referred to as the Lorentz–Lorenz model) for a dilute suspension of spheres [5]:

$$\left(\varepsilon_s^* - \varepsilon_f^*\right) / \left(\varepsilon_s^* + 2\varepsilon_f^*\right) = \beta^*(\omega)\, c \quad (6)$$

where c is the particle volume fraction. Polarizability measurements enable quantitative predictions of dielectrophoretic phenomena in a colloidal dispersion subjected to strong fields (~ kV/mm) since the field dependence of β*(ω) usually appears to be insignificant.

As the dipole moment **P** is proportional to the field strength in Eq. 2, the second term in Eq. 1 appears to be proportional to the square of the applied field. This implies that under the action of an AC electrical field, the dielectrophoretic force yields a nonzero time-average value [1, 6]:

$$\left\langle \mathbf{F}_{dep} \right\rangle = \frac{3}{2}\varepsilon_0\varepsilon_f v_p \mathrm{Re}[\beta^*(\omega)]\nabla E_{rms}^2(\mathbf{r}) \quad (7)$$

where $\langle\,\rangle$ denotes averaging over the period of the field oscillations and E_{rms} is the root mean square of the local electrical field strength. This force causes a particle to move to the regions of high field strength (*positive* dielectrophoresis) or low field strength (*negative* dielectrophoresis), depending on whether the particle is more ($\mathrm{Re}(\beta^*) > 0$) or less ($\mathrm{Re}(\beta^*) < 0$) polarizable than the suspending fluid (Fig. 1b).

Single-Particle Model of Dielectrophoresis

A single-particle model ignores the interparticle interactions and considers only the dielectrophoretic, Eq. 7, and hydrodynamic forces exerted on a polarized particle. The Reynolds numbers for the fluid flow $\mathrm{Re}_f = \rho_f |\mathbf{v}_f| d/\eta_f$ and for the particle motion relative to the fluid $\mathrm{Re}_p = \rho_p |\mathbf{u} - \mathbf{v}_f| a/\eta_f$, where $\mathbf{v}_f$ is the local fluid velocity, $\mathbf{u}$ is the local particle velocity, and ρ_p and ρ_f are, respectively, the particle and fluid densities, are usually very small. The hydrodynamic force exerted on a spherical particle is given by the Stokes equation since the contribution of the inertial and Basset forces to the transient regimes of the particle motion relative to the fluid is typically negligible in comparison with the viscous forces [5]. With these assumptions, the force balance for a spherical particle flowing through a field region yields

$$\frac{d\mathbf{r}}{dt} = \mathbf{u} \text{ with } \mathbf{r}|_{t=0} = \mathbf{r}_0 \tag{8}$$

$$\text{with } 6\pi\eta_f a\left[\mathbf{u} - \mathbf{v}_f(\mathbf{r})\right] = 2\pi\varepsilon_0\varepsilon_f a^3 \mathrm{Re}(\beta^*)\nabla E^2_{rms}(\mathbf{r}) + \frac{4}{3}\pi\left(\rho_p - \rho_f\right)a^3\mathbf{g}$$

where $\mathbf{g}$ is the gravitational acceleration and $\mathbf{r}(t)$ is the displacement of a particle initially at $\mathbf{r}_0$. To find the local values of the fluid velocity $\mathbf{v}_f(\mathbf{r})$ and the field gradient $\nabla E^2_{rms}(\mathbf{r})$, Navier–Stokes' equations for the fluid flow and Laplace's equation for the electrical field are solved for the dielectrophoretic device without particles. A gradient of an electrical field in Eq. 8 can be estimated as $\nabla E^2_{rms} \sim E^2_{rms}/d$, where d is the characteristic length of the field variation. Equation 8 provides a scale of time it takes a particle to travel the distance d relative to the fluid under the action of the dielectrophoretic force.

$$\tau_d = \frac{3\eta_f d^2}{a^2\varepsilon_0\varepsilon_f|\mathrm{Re}(\beta^*)|E^2_{rms}} \tag{9}$$

Three nondimensional parameters of Eq. 8 characterize the particle motion: the ratio of the dielectrophoretic time, Eq. 9, to the average fluid residence time, $\tau_d Q/(AL)$, where L is the characteristic length along the flow streamlines, Q is the volumetric flow rate, and A is the channel cross-sectional area; the ratio L/d, and the ratio of the gravitational and dielectrophoretic forces $G = 2\left(\rho_p - \rho_f\right)gd^3/3\varepsilon_0\varepsilon_f|\mathrm{Re}(\beta^*)|V^2_{rms}$. Predictions of Eq. 8 for the motion of polarized particles are limited to very dilute suspensions when the electrical and hydrodynamic interactions between particles are negligible [7].

Predictions of the dielectrophoretic migration and rotation of a nonspherical polarizable particle require considering the balance of hydrodynamic and electrical forces and torques exerted on the particle by an electrical field and a flowing fluid.

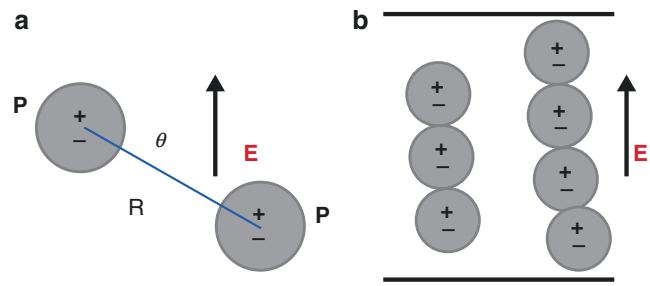
Structural Transitions in Suspension Subjected to Spatially Uniform Electrical Field

When an electrical field is applied to a polarized suspension, each particle acquires a dipole moment given by Eq. 2. The dipole–dipole interparticle interactions decay with the inverse cube of separation distance and, depending on the relative orientation of the dipoles, vary from attraction to repulsion [5, 6]. When the dielectric constants and conductivities of the particles and the suspending fluid are frequency independent, the time-average energy of the dipole–dipole interaction between two spheres subjected to an AC field is [8]

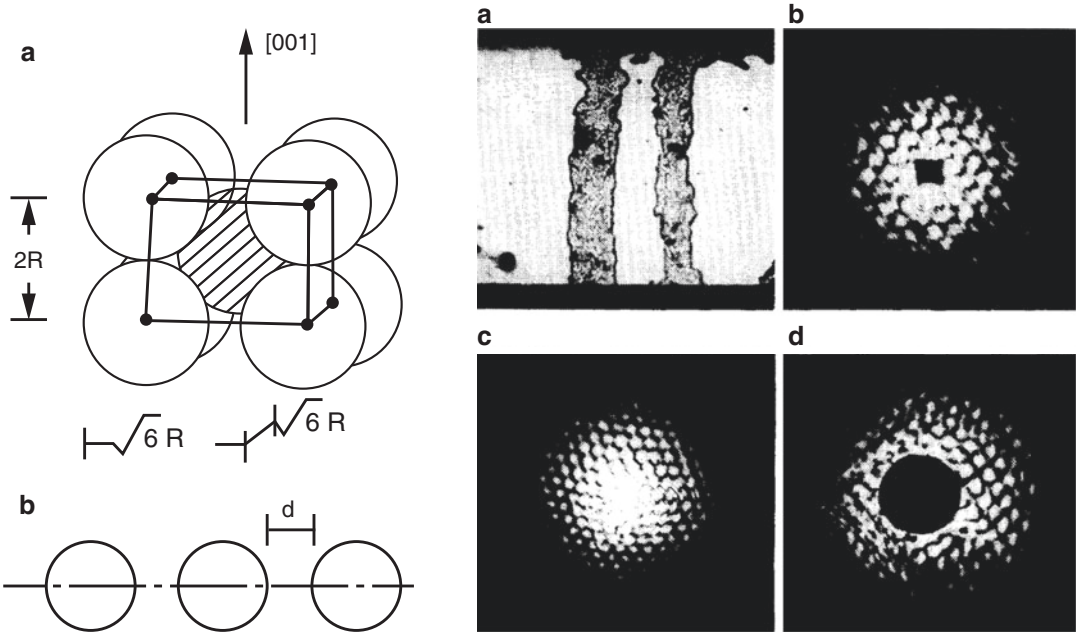
$$W_{dd} = \varepsilon_0\varepsilon_f v_p \Psi_\omega E^2_{rms}(a/\mathrm{R})^3\left(1 - 3\cos^2\theta\right) \tag{10}$$

where a is the particle radius, R is the center-to-center separation between the particles, θ is the angle between the line connecting the particle centers and the electrical field line (Fig. 2a), E_{rms} is the root mean square of the electrical field strength, and the function Ψ_ω characterizes the electrical energy of the interparticle interaction that depends on the field frequency and electrical properties of the particles and the suspending fluid. A rough estimate yields $\Psi_\omega \sim 3\mathrm{Re}(\beta^*)^2$ for a relatively small $\mathrm{Im}(\beta^*)$.

When the dipolar interparticle interactions exceed the randomizing Brownian motion,



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 2 (**a**) Interacting spherical particles with induced dipole moments. (**b**) Polarized particles aligned along the field direction



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 3 A low-viscosity silicone oil containing size-selected glass microspheres of diameter either 20.0 ± 1.8 or 40.7±1.7 μm. (*Left*) (**a**) Unit cell of the body-centered tetragonal (bet) structure lattice for spheres of radius R, where the [001] axis is the applied field direction. (**b**) A linear periodic array of glass spheres of radius R separated by a distance d. (*Right*) (**a**) View of typical columns formed between the two electrodes which are separated by 3 mm. (**b**) Diffraction pattern of a (110) plane for 20.0-μm-diam spheres. (**c**) Pattern of a (110) plane for 40.7-μm-diam spheres. (**d**) Pattern of a (100) plane for 40.7-μm-diam spheres. In (**b**) and (**d**), the centers are masked to suppress overexposure (Reprinted with permission from Ref. [9]. Source: http://journals.aps.org/prl/abstract/10.1103/PhysRevLett.68.2555. Copyright by the American Physical Society)

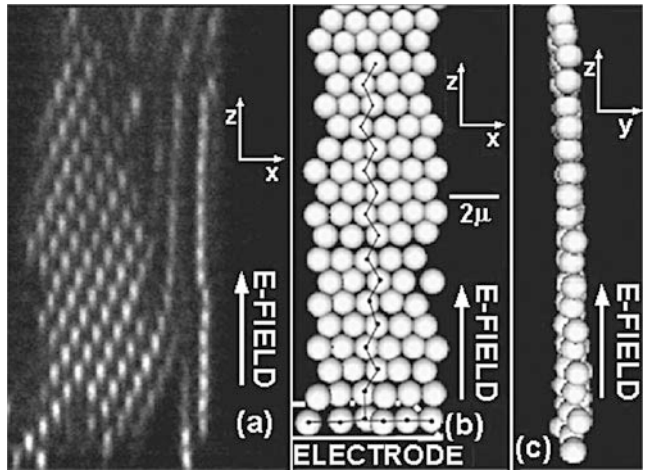
electrically induced particle rearrangement would occur. A head-to-tail configuration with $\theta = 0$ or $\theta = \pi$ in Eq. 10 is the lowest energy configuration of two polarized particles. Therefore, dipolar interactions in suspensions exposed to a spatially uniform electrical field cause particles to form chains along the applied field direction (Fig. 2b). Computer simulations for the electrical energy of a suspension of spherical particles treated as a system of point dipoles demonstrate that the zero-temperature ground state of the suspension (i.e., neglecting entropy) is a body-centered tetragonal (bct) structure. This prediction was validated in experiments (Fig. 3; [9]). However, the long-range nature of dipole–dipole interactions, Eq. 10, leads to coarsening of the chains/columns into sheets (Fig. 4; [10]) that eventually transform into a bct lattice of strings [3, 11, 12].

A thermodynamic theory [8, 13] provides an estimate of conditions under which polarizable particles would aggregate. This theory yields the following expressions for the chemical potential μ_p and the osmotic pressure Π_p of the particles averaged over the period of the field oscillations [8]:

$$\mu_p = \left(k_B T / v_p\right)\left(df_0/dc\right) - \varepsilon_0 \left(\partial \varepsilon_s'/\partial c\right)_{\omega t_c} \left\langle E^2/2 \right\rangle \tag{11}$$

$$\Pi_p = k_B T c Z / v_p + \varepsilon_0 \left[\varepsilon_s' - c\left(\partial \varepsilon_s'/\partial c\right)_{\omega t_c}\right] \left\langle E^2/2 \right\rangle \tag{12}$$

where $f_0 = c \ln \frac{c}{e} + c \int_0^c \frac{Z-1}{c} dc$, $k_B T$ is the thermal energy, Z is the suspension compressibility factor, and $\left(\partial \varepsilon_s'/\partial c\right)_{\omega t_c}$ is the derivative of the real part of



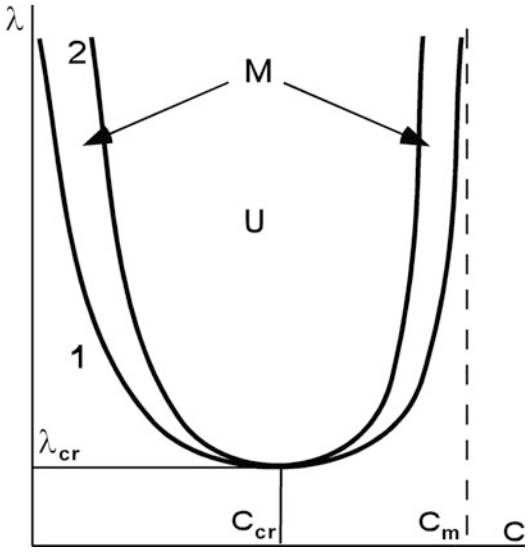
AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 4 Silica spheres (0.525 μm radius, polydispersity in size 1.8 %) in a mixture of water and glycerol. Sheet structure. The field strength is 1.2 kV/mm and the field has been on for ~30 min. (**a**) *x*–*z* Confocal micrograph (raw data). There is a sheet of particles seen in the left side of the image, with a single chain of particles to the right of it. The bottom of the image is closest to the objective lens of the microscope. It should be noted that since only the cores of the silica spheres fluoresce, touching spheres are seen as separated. It is the anisotropic point spread function that causes the spherical fluorescent cores to appear ellipsoidal. Note also that the intensity and resolution diminish away from the lens due to spherical aberration. (**b**) View of a sheet seen face on. The image is taken from a rendering of reconstructed 3D raw data such as seen in (**a**). The white scale bar is 2 μm. The bottom row of spheres is touching the glass electrode (the electrode is perpendicular to the image), and the electric field is upwards. The spheres are hexagonally close packed within a sheet. The smallest angle between nearest neighbors and the field is +/−30°, except for the first two layers adjacent to the electrode, where nearest neighbors are aligned along the field. (**c**) Side view of the same sheet object (Reprinted with permission from Ref. [10]. Source: http://scitation.aip.org/content/aip/journal/jcp/112/8/10.1063/1.480933. Copyright 2014 AIP Publishing)

the suspension dielectric permittivity ε_s^* taken at a constant value of ωt_s with t_s being the relaxation time of the particle polarization. The first terms in Eqs. 11 and 12 refer to the free energy and the osmotic pressure in the absence of an external electrical field. The other terms represent the interaction of particles with the external electrical field and the particle–particle dipole interactions. The concentration dependence of Z is described by the Carnahan–Starling equation for a suspension of spheres in a disordered state [5]. Equations 11 and 12 express the critical conditions beyond which a random arrangement of particles becomes unstable in a spatially uniform AC field in terms of the particle volume fraction c and the dimensionless parameter $\lambda = \varepsilon_0 \varepsilon_f \langle E^2 \rangle v_p / k_B T$ that characterizes the ratio of electrical and thermal energies. The phase diagram of a polarizable suspension in terms of particle concentration versus relative field strength is similar to the phase diagram of concentration versus temperature of two liquids which are completely miscible only above a certain temperature. The simplest suspension phase diagram corresponding to the case of small Im(β^*) is sketched in Fig. 5 [7], in which the spinodal separates the metastable regime **M** (slow phase transition) from the unstable regime **U** (fast phase transition) in the two-phase region. On the curve separating two-phase and single-phase regions in Fig. 5, the chemical potentials and the osmotic pressures of particles in the coexisting phases are equal. The critical point on the spinodal curve, c_{cr} and λ_{cr}, coincides with the inflection point of μ_p as a function of c [8]:

$$\frac{\partial \mu_p}{\partial c} = \frac{\partial^2 \mu_p}{\partial c^2} = 0 \text{ with } \frac{\partial^3 \mu_p}{\partial c^3} > 0$$

The single-phase region of the phase diagram (at $\lambda < \lambda_{cr}$) represents a suspension with the

random spatial arrangement of particles, whereas the two-phase region (at $\lambda > \lambda_{cr}$) corresponds to the appearance of the field-induced separation of a suspension in regions enriched with and depleted of particles. The critical point and the coexistence and spinodal lines on the phase diagram are found to be sensitive to the concentration dependence of the second derivative of the suspension dielectric permittivity $\partial^2 \varepsilon_s' / \partial c^2$ [14]. Describing structures formed by particles is beyond the scope of the thermodynamic theory [8, 13] and requires molecular dynamics simulations [11, 12].



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 5 Phase diagram of particle concentration c versus relative field strength λ for a suspension. Curves 1 and 2 are, respectively, coexistence and spinodal curves (Reprinted with permission from Ref. [7]. Source: http://journals.aps.org/pre/abstract/10.1103/PhysRevE.69.021402. Copyright 2014 by the American Physical Society)

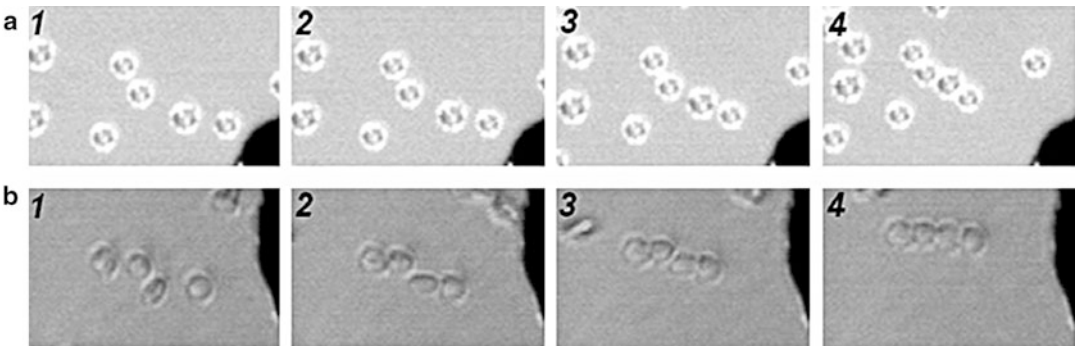
Key Research Findings

Synergism of AC Dielectrophoresis and Dipolar Interactions

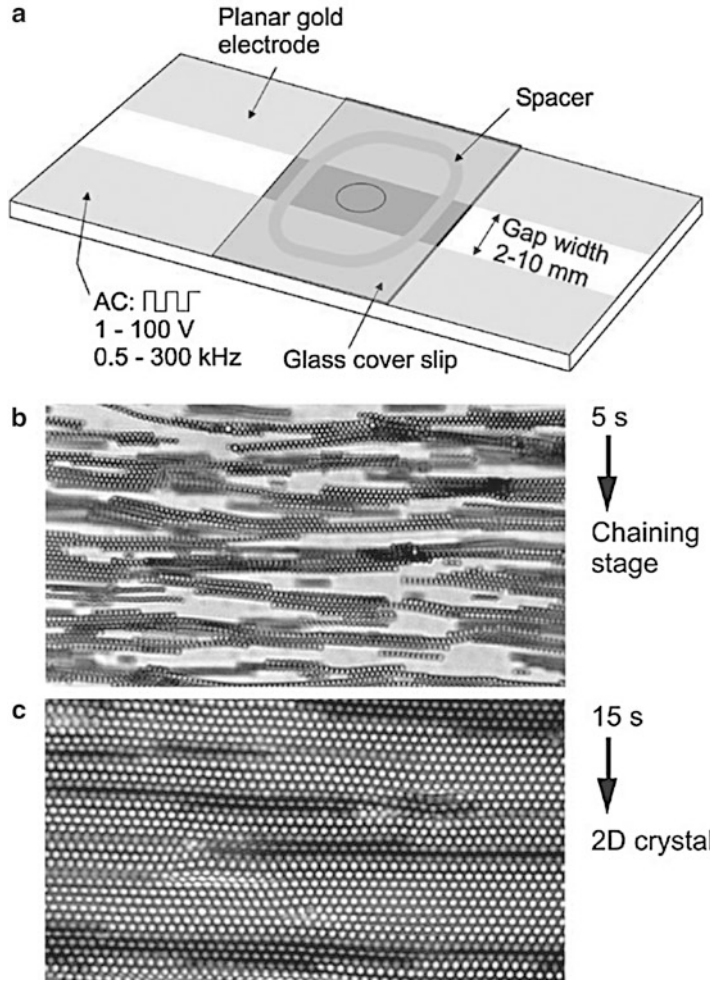
Figures 6, 7, 8, and 9 from Refs. [15–18] demonstrate that the synergism of the dielectrophoretic and dipole–dipole forces provides an efficient tool for the organization and assembly of particles into functional structures. Under such conditions, the interaction of particles with an external field, Eq. 7, in a suspension is exposed to a spatially nonuniform AC electrical field that is accompanied with the interparticle dipole–dipole interactions, Eq. 10, as each particle acquires a dipole moment.

As the dipolar interparticle force is $\langle \mathbf{F}_{dd} \rangle = -\nabla W_{dd}$, Eqs. 7 and 10 yield the following estimate for the ratio of the dipole–dipole and dielectrophoretic forces:

$$\frac{\langle \mathbf{F}_{dd} \rangle}{\langle \mathbf{F}_{dep} \rangle} \sim \frac{6|\mathrm{Re}(\beta^*)|a^3 d}{R^4} \tag{13}$$

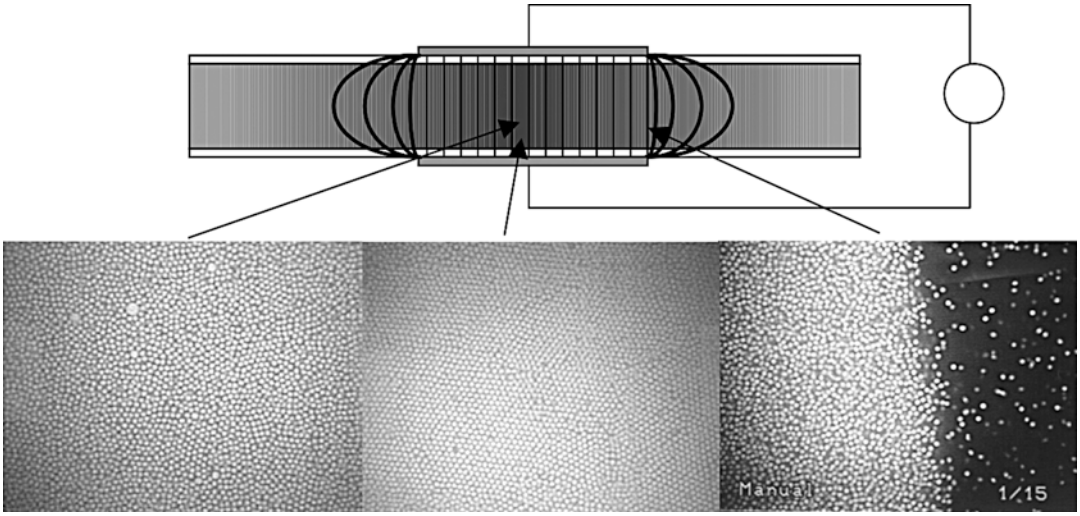


AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 6 Polystyrene beads with a diameter of 6.08 μm and red blood cells in buffer solutions. (**a**) Polystyrene beads and (**b**) red blood cells form a chain when exposed to an electric field in a quadruple electrode micro-device (Reprinted with permission from Ref. [15]. Source: http://onlinelibrary.wiley.com/doi/10.1002/elps.201300636/abstract. Copyright 2014 John Wiley and Sons)



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 7 Assembly process of polystyrene microspheres in a thin experimental cell under an AC electric field. (**a**) Schematic of a coplanar electrode experimental cell. The spacer can be created using Teflon tape or hydrophobic PAP pen yielding a cell height of 60–100 μm or 10–20 μm, respectively. (**b**, **c**) Optical micrographs illustrating the two-stage mechanism of crystallization for latex particles. (**b**) Shortly after the field is applied, the particles align in chains due to dipolar attraction. Simultaneously, the dielectrophoretic force due to the field gradient attracts the particles to the high-field intensity region. (**c**) The particle chains confined on the surface form 2D hexagonal crystals aligned with one axis in the direction of the field (Reprinted with permission from Ref. [16]. Source: http://pubs.rsc.org/en/content/articlelanding/2009/pc/b803015b#!divAbstract. Copyright 2014 Royal Society of Chemistry)

where the field gradient is estimated as $\nabla E_{rms}^2 \sim E_{rms}^2/d$ and R can be estimated as $a/c^{1/3}$. The interparticle interactions can be neglected if the ratio in Eq. 13 is small that occurs at a very low particle concentration. This situation corresponds to the single-phase region in the suspension phase diagram in Fig. 5. When the local values of the particle concentration and the field strength in certain regions of a dielectrophoretic device fall into the two-phase region of the suspension phase diagram in Fig. 5, dielectrophoresis of the particles would be accompanied by field-induced structural transitions caused by the interparticle dipolar interactions [13]. The kinetics of these transitions depends on the ratio of two time scales: τ_a and τ_d, given in Eq. 9. The former refers to the



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 8 A suspension of 1-μm PMMA particles in decalin in the "dielectrophoretic bottle." The schematic diagram shows the capacitor plates (outside the cover slips) and the electric field lines. The particle concentration can be varied from less than 1 % to above the crystallization concentration ~0.54. The confocal images on the left and right illustrate the density profile about ten particle layers up from the bottom cover slip. The middle image was taken at ~5 layers in, where the density is slightly higher due to gravity sedimentation (Reprinted with permission from Ref. [17]. Source: http://iopscience.iop.org/article/10.1088/0953-8984/15/1/302/meta. Copyright © IOP Publishing)

characteristic time for the field-induced structure formation in a suspension subjected to a spatially uniform field [8]:

$$\tau_a \approx \frac{\eta_f}{\varepsilon_0 \varepsilon_f [\mathrm{Re}(\beta^*) E_{rms}]^2} \xi \text{ and } \frac{\tau_d}{\tau_a} \approx \frac{3d^2 |\mathrm{Re}(\beta^*)|}{a^2 \xi(c)} \tag{14}$$

where η_f is the viscosity of the suspending fluid and $\xi \approx 0.4\left[(\pi/6c)^{5/3} - 1\right]$ is a coefficient that depends on the volume fraction of the particles.

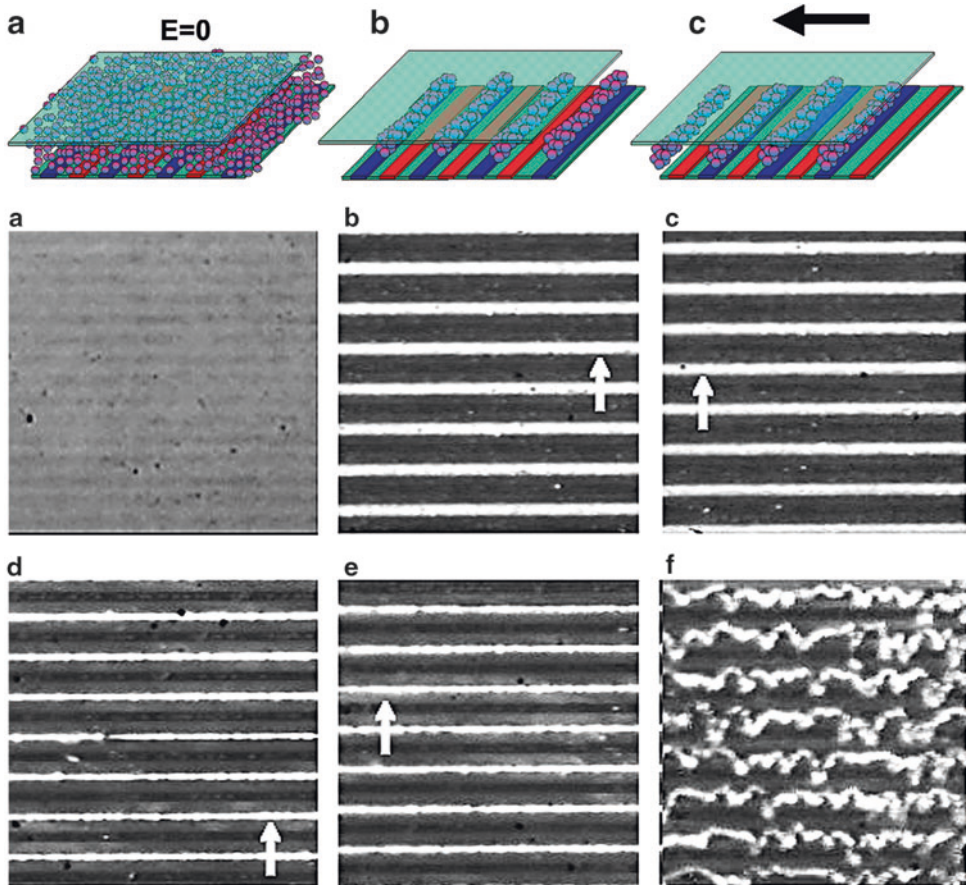
The regime $\tau_d << \tau_a$ corresponds to *heterogeneous aggregation*. Under these conditions typical of dilute suspensions, the particles are forced to accumulate in certain areas of the dielectrophoretic device and then to undergo field-induced structural transitions. The regime $\tau_d >> \tau_a$ corresponds to *homogeneous aggregation* in which macroscopic redistribution of particles during field-induced structural transitions is negligible. This regime is typical of applications of electrorheological effect [3].

Macroscopic Models

The mean-field model [7, 19] expresses the electrical and hydrodynamic interparticle interactions in terms of the instantaneous local electrical fields, particle concentrations, and flow velocities. It is based on coupled quasi-steady equations for an electrical field, Eq. 15, plus momentum balance and continuity equations for a suspension, Eq. 16, and the equation for the field and flow-driven particle motion, Eq. 17. These equations are averaged over the field oscillations. The suspension is approximated as a Newtonian fluid with a concentration-dependent viscosity. When local values of the particle concentration and electrical field lie within the single-phase region of the suspension phase diagram in Fig. 5, the electrical force exerted on the particles and the particle velocity relative to the suspending fluid in these equations are given by the particle chemical potential μ_p, Eq. 11.

Field equations

$$\nabla \cdot \varepsilon_s^*(\omega, c) \nabla \phi_\omega^* = 0 \text{ with } \mathbf{E}_\omega^* = -\nabla \varphi_\omega^* \tag{15}$$



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 9 The schematic diagram shows conveyor belt building of large-scale multiparticle structures. Formation [(**a**) and (**b**)] and transportation [(**b**) and (**c**)] of the particle structures by energizing (*red*) and grounding (*blue*) electrodes. Photos show 1 % (*v*/*v*) suspension of 3.1-μm latex beads in a chamber equipped with two-electrode array dielectrophoretic chip energized with 5 V_{rms}/1 MHz, chamber top grounded. (**a**) Initially, the electrodes were barely visible through the suspension. (**b**) Beads were concentrated into columns (seen as *white*) in the low-field regions caused by energizing electrode array 1 and grounding array 2 and then (**c**) were moved to other low-field regions by grounding array 1 and energizing array 2. Columns were formed by energizing array 1 and grounding array 2 and (**e**) were translated by moving the chip with velocity 4.3 μm/s using the stage but (**f**) were destabilized at a chip velocity of 700 μm/s. *White arrows* on (**b**) and (**c**) as well as (**d**) and (**e**) indicate the sequential positions of the same column, respectively. Chamber height: [(**a**)–(**c**)] ~100 μm and [(**d**)–(**f**)] ~50 μm. (Reprinted with permission from Ref. [18]. Source: http://scitation.aip.org/content/aip/journal/apl/90/15/10.1063/1.2721238. Copyright 2014 AIP Publishing)

Momentum balance and continuity equations

$$\rho_s\left(\frac{\partial \mathbf{v}}{\partial t}+\mathbf{v}\cdot\nabla\mathbf{v}\right)=-\nabla p+\nabla\cdot\eta_s(c)\left(\nabla\mathbf{v}+\nabla\mathbf{v}^T\right)$$
$$-c\nabla\mu_p+c\left(\rho_p-\rho_f\right)\mathbf{g}\ \text{and}\ \nabla\cdot\mathbf{v}=0 \qquad (16)$$

Particle motion

$$\frac{\partial c}{\partial t}+\nabla\cdot\left(c\mathbf{v}+\mathbf{j}_p\right)=0\ \text{with}$$
$$\mathbf{j}_p=\frac{c(1-c)^2 v_p}{6\pi a\eta_s(c)}\left[-\nabla\mu_p+\left(\rho_p-\rho_f\right)\mathbf{g}\right] \qquad (17)$$

where $\varepsilon_s^*(\omega, c)$ is the complex dielectric permittivity of a suspension as a function of the field

frequency ω; the particle concentration c, $\mathbf{E}^*_\omega$, and φ^*_ω are, respectively, the Fourier components of the electrical field and potential; v and p are the suspension velocity and pressure; and $\rho_s = (1-c)\rho_f + c\rho_p$ and $\eta_s = \eta_f\hat{\eta}(c)$ are the suspension density and viscosity with $\hat{\eta} = (1-c/c_m)^{-n}$ where $n \approx 2$ and $c_m \sim 0.68$ is the concentration of spheres at random close packing. The particle chemical potential μ_p, Eq. 11, is a function of the particle concentration c and oscillation-averaged field strength $\langle E^2\rangle = \left|\mathbf{E}^*_\omega(\mathbf{r})\right|^2/2$. The factor $(1-c)^2$ in the expression for the particle flux $\mathbf{j}_p$ in Eq. 17 represents the hindrance function that slows down the particle mobility.

The effects of the particle arrangement on the local field strength are included in the model via the particle concentration dependence of $\varepsilon^*_s(\omega, c)$ in Eq. 15. Equations 16 and 17 incorporate the interparticle electrical interactions in terms of the field and concentration dependence of the particle chemical potential, Eq. 11. The interparticle hydrodynamic interactions are included in terms of the concentration dependence of the suspension viscosity and the hindrance function.

Hydrodynamic equations for local values of the particle concentration and electrical field in the two-phase region of the phase diagram in Fig. 5 are formulated for relatively fast local phase separation, i.e., $\tau_d >> \tau_a$ in Eq. 14 at a local particle concentration. Under these conditions, the value of the chemical potential is the same in coexisting phases and is specified as a function of λ by the coexistence curve of the phase diagram. The particle concentrations in both phases, c_1 and c_2, are also specified as functions of λ by expressions [8]

$$\mu_p(c_1) = \mu_p(c_2) \text{ and } \Pi_p(c_1) = \Pi_p(c_2) \quad (18)$$

The volume fraction of the high concentration phase ψ is given by the lever rule $c_2\psi + c_1(1-\psi) = c$ as a function of c and λ. For simplicity, the model ignores the explicit dependence on ψ of the complex dielectric permittivity, viscosity, and hindrance function for this zone. These parameters are obtained by the same way as in the single-phase region. Under these conditions, the two-phase model consists of Eqs. 15, 16, and 17 in which the chemical potential is now specified as a function of the field strength, $\mu_p = \mu_p(\lambda)$, through solution of Eq. 18. Notice that the spatial gradient of the particle chemical potential $\nabla\mu_p$, considered as a function of the local particle concentration and the relative field strength, experiences an abrupt change when the suspension undergoes the local field-driven phase transition, i.e., when the point $c(\mathbf{r}, t)$, $\lambda(\mathbf{r}, t)$ representing the local suspension composition and field strength in the phase diagram, Fig. 5, crosses the coexistence curve.

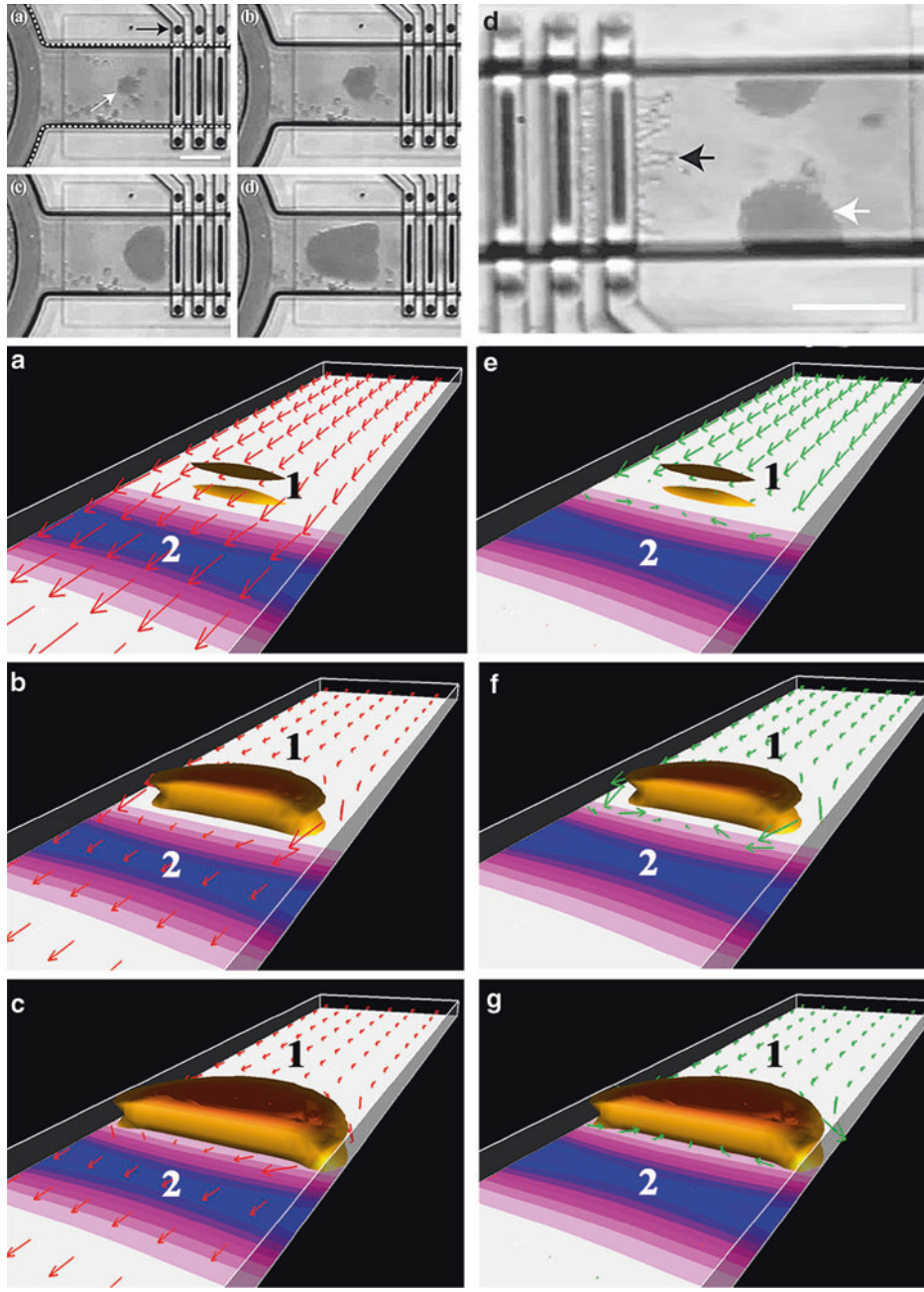
For the limiting case of a dilute suspension for which $c \to 0$, Eqs. 15 and 16 reduce to the field and flow equations for the suspending fluid in a dielectrophoretic device, while Eq. 17 yields

$$\frac{\partial c}{\partial t} + \nabla\cdot\left[c\left(\mathbf{v}_f + \mathbf{u}_e\right)\right] = D_B\Delta c \text{ with}$$

$$\mathbf{u}_e = \frac{v_p}{6\pi a\eta_f}\left[\frac{3}{2}\varepsilon_0\varepsilon_f \mathrm{Re}(\beta)\nabla E^2_{rms} + \left(\rho_p - \rho_f\right)\mathbf{g}\right] \quad (19)$$

where $\mathbf{u}_e$ is the velocity of a particle relative to the fluid under the action of the dielectrophoretic and gravity forces and $D_B = k_BT/6\pi a\eta_f$ is the particle Brownian diffusion coefficient. When Brownian motion is insignificant, Eq. 19 leads to the single-particle model, Eq. 8. Another limiting case is the equilibrium state of neutrally buoyant particles subjected to a spatially nonuniform AC field [13], for which Eqs. 15, 16, and 17 yield $\mathbf{v} = \nabla p = 0$ and $\nabla\mu_p(\mathbf{r}) = 0$ where μ_p is given by Eq. 11. This spatial distribution of the particles is reached when the chemical potential of a particle becomes independent of position. The equation $\nabla\mu_p(\mathbf{r}) = 0$ represents c as a function of $\langle E^2\rangle$ along the curve of constant μ_p on the suspension phase diagram, Fig. 5.

Equations 15, 16, 17, and 18 require no fitting parameters provided that particle properties are measured independently. Simulations of front formation and propagation using independently measured data for the fluid and particle properties are quantitatively consistent with experimental data (Figs. 10 and 11). Numerical simulations and



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 10 (*Top left*) The photo shows dielectrophoresis (10 V pp, 30 MHz) of 1-μm polystyrene beads in deionized water accompanied by a phase separation (*white arrow*). The *dashed lines* denote the outline of the fluidic channel, and the *black arrow* indicates the energized electrode. The time is (**a**) 10, (**b**) 70, (**c**) 120, and (**d**) 180 s. The flow was from left to right, with the flow rate ranging from 0.24 to 9.6 pL/s. Scale bar = 20 μm. (*Top right*) The photo shows dielectrophoretic separation of bacterial cells and 1-μm polystyrene beads in deionized water (10 V pp, 15 MHz). The cells adhere to the energized electrode (*black arrow*), while the beads experience negative dielectrophoresis accompanied by a phase separation (*white arrow*). The flow is from right to left: 6pL/s. Scale bar = 20 μm. (*Bottom*) Numerical simulations show the concentration contours (1) for the 0.1 % (v/v)-suspension, 8.64 pL/s, 10 V pp. In the panels, the fluid flow is from the

asymptotic analysis of these equations reveal two limiting regimes for the dielectrophoretic formation of a sharp concentration front seen in Figs. 9, 10, and 11. One of them is the quasi-equilibrium regime in which dielectrophoresis is accompanied by field-driven phase separation leading to an abrupt change in particle concentration [13]. The position and the magnitude of the concentration front are then determined by the balance of the dielectrophoretic, dipole–dipole, and steric forces. The other is a hydrodynamic mechanism which operates through the balance of the dielectrophoretic and hydrodynamic forces. The position and the magnitude of the concentration jump are governed by the rapid local growth of the suspension viscosity due to the particle accumulation near the front which slows down the field-driven particle motion by decreasing their mobility, proportional to $(1-c)^2/\hat{\eta}(c)$. The concentration front forms in both the single-phase and the two-phase states of a suspension and travels with the speed that increases with increasing applied voltage. In contrast to the front formation due to the field-driven phase transition, the hydrodynamic mechanism can operate only in an unsteady regime of dielectrophoresis and will cease when the particle spatial distribution reaches the equilibrium.

Microscopic Models

Direct numerical simulations of the motion of interacting particles require the solution of Laplace's equation for the electrical field and Navier–Stokes' equations for the fluid flow in a dielectrophoretic device that consider the presence of all particles. The electrical and hydrodynamic forces and torques exerted on each particle are then computed by integrating the Maxwell stress tensor and the hydrodynamic stress tensor over the particle surface. As the particle positions change continuously, the solution of Laplace's and Navier–Stokes' equations is carried out at each simulation time step. The capabilities of current computers limit direct numerical simulations to a system of two or three particles.

A microscopic approach [20] for modeling the motion of interacting spherical particles is implemented by presenting the total electrical field $\mathbf{E}_{tot}(\mathbf{r}_i)$ exerted on a particle in the location $\mathbf{r}_i$ as the external field in this point $\mathbf{E}_{ext}(\mathbf{r}_i)$ plus the contributions of dipoles induced in other particles:

$$\mathbf{E}_{tot}(\mathbf{r}_i) = \mathbf{E}_{ext}(\mathbf{r}_i) + \sum_{j\neq i} \mathbf{E}_j(\mathbf{r}_i) \text{ for } i = 1, \cdots, N \tag{20}$$

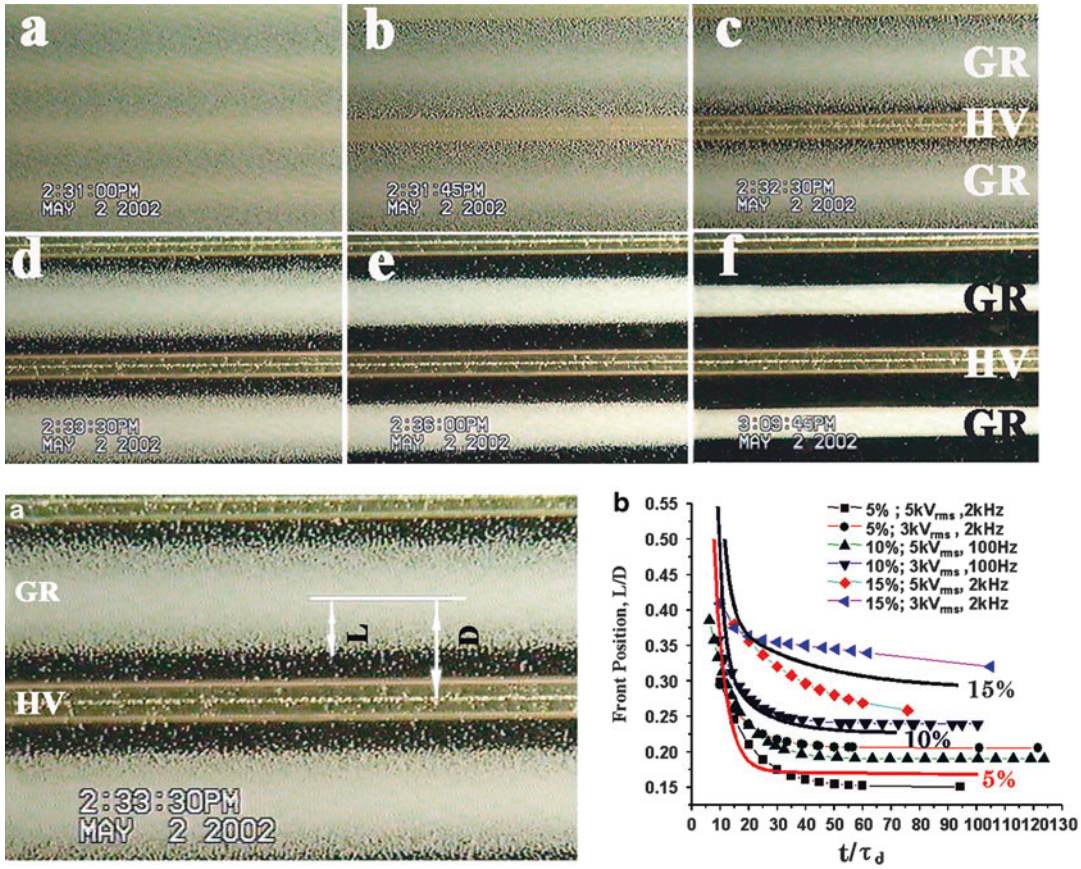
$$\text{with } \mathbf{E}_j(\mathbf{r}_i) = \frac{1}{4\pi\varepsilon_0\varepsilon_f r_{ij}^3}\left[3\mathbf{n}\left(\mathbf{n}\cdot\mathbf{P}_j\right) - \mathbf{P}_j\right] \text{ and}$$

$$\mathbf{P}_i = 3v_p\varepsilon_0\varepsilon_f\beta^*\mathbf{E}_{tot}(\mathbf{r}_i)$$

where $\mathbf{P}_i$ and $\mathbf{P}_j$ are, respectively, the dipoles of particles at the locations $\mathbf{r}_i$ and $\mathbf{r}_j$, $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$, $\mathbf{n} = \mathbf{r}_{ij}/r_{ij}$, and $\mathbf{E}_j(\mathbf{r}_i)$ is the electrical field generated by the particle at the location $\mathbf{r}_j$. The external field $\mathbf{E}_{ext}(\mathbf{r}_i)$ is computed for a dielectrophoretic device without particles. The effects of the particle dipole–dipole interactions are simulated by solving coupled Eq. 20 for all particles. The electrical energy of the instantaneous particle configuration is then computed as

$$U(\{\mathbf{r}_1, \cdots, \mathbf{r}_N\}) = \sum_i U_{eff}(\mathbf{r}_i) + \sum_{j\neq i} U_{eff}\left(\mathbf{r}_i, \mathbf{r}_j\right)$$

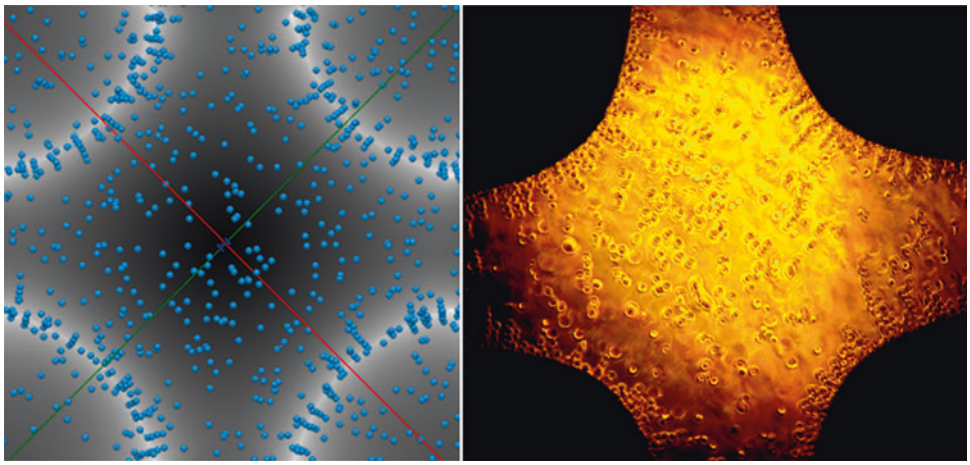
AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 10 (continued) upper right-hand corner to the lower left-hand corner. The color band (2) indicates the variation of E^2 along the channel bottom (*white* for up to 1.3 kV^2/cm^2, *pink* for 1.3–2.7 kV^2/cm^2, and *blue* for 10.6–11.9 kV^2/cm^2), whose maximum is located near the first electrode. The time is (**a**, **e**) 10, (**b**, **f**) 70, (**c**, **g**) 180, and (**d**, **h**) 120 s. The computed values for c_{max} in the bolus [in %(v/v)] are (**a**) 8.73, (**b**) 45.4, (**c**) 56.4, and (**d**) 54.1. In (**a**–**d**), the *red arrows* show the relative magnitude of the flow velocity [$v_{max\ at\ t=0}$ = 36 μm/s; v_{max} = (**a**) 49.8, (**b**) 106.3, (**c**) 110.4, and (**d**) 139.9 μm/s]. In (**e**–**h**), the green arrows show the relative magnitude of the particle velocity, $v_{p,max\ at\ t=0}$ = 36 μm/s, $v_{p,max}$ = (**e**) 45.2, (**f**) 78.8, (**g**) 151.4, and (**h**) 171.6 μm/s (Reprinted with permission from Ref. [19]. Source: http://scitation.aip.org/content/aip/journal/apl/83/23/10.1063/1.1629789. Copyright 2014 AIP Publishing)



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 11 (*Top*) Negatively polarized neutrally buoyant spherical polyalphaolefin 87-μm particles in Mazola corn oils. The particle distribution in a suspension with 10 % (v/v) particle concentration (**a**) before and (**b**)–(**f**) following the application of a field 5 kV rms, 100 Hz at t= (**b**) 45 s, (**c**) 90 s, (**d**) 150 s, (**e**) 300 s, and (**f**) final state, ~39 min. The electrode width is 1.6 mm. HV and GR refer to the high-voltage and grounded electrodes, respectively. (*Bottom*) (**a**) The photograph illustrates how the front position L was measured; D =3.6 mm. (**b**) The experimental data (symbols) and computational results (*solid lines*) of the electrohydrodynamic model Eqs. 15, 16, 17, and 18 for the front propagation in suspensions with 5 %, 10 %, and 15 % (v/v) particle concentrations for different voltages and frequencies of the applied fields. The experimental data and the simulation results are plotted against a nondimensional time, t/τ_d, with τ_d given by Eq. 9 with $E_{rms}^2 = V_{rms}^2/d$ where d is the electrode width and V_{rms} is the root mean square of the applied AC voltage (Reprinted with permission from Ref. [7]. Source: http://journals.aps.org/pre/abstract/10.1103/PhysRevE.69.021402. Copyright 2014 by the American Physical Society)

where $U_{eff}(\mathbf{r}_i)$ is the energy of the dipole $\mathbf{P}_i$ in the electrical field $\mathbf{E}_{tot}(\mathbf{r}_i)$ and $U_{eff}(\mathbf{r}_i, \mathbf{r}_j)$ is the electrical energy of two dipoles $\mathbf{P}_i$ and $\mathbf{P}_j$. A Monte Carlo method that employs a stochastic sequence of single-particle displacement events is used for simulating the equilibrium arrangement of N particles subjected to a spatially nonuniform external field [20]. Specifically, a randomly selected particle displacement $\Delta\mathbf{r}_i$ is accepted if the difference between the energies of the new and old configurations $\Delta U = U(\{\mathbf{r}_1, \cdots, \mathbf{r}_i + \Delta\mathbf{r}_i, \cdots, \mathbf{r}_N\}) - U(\{\mathbf{r}_1, \cdots, \mathbf{r}_N\})$ is negative. Equation 20 includes the effects of the particle arrangement on the local field strength. However, the model predictions are limited to the equilibrium arrangement of particles, Fig. 12 [20], since the dynamics of the field-driven particle motion is not considered.

In approach [21], modeling the motion of interacting spherical particles is implemented by



AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation, Fig. 12 (*Left*) Top view of the final spatial distribution of the Monte Carlo simulation with superimposed intensity map of the electric field $|\mathbf{E}_{rms}(\mathbf{r})|$. (*Right*) Top view of experimental spatial distribution of the human breast cancer cell line MDA-MB-231 after 180 s of dielectrophoretic manipulation. *Black areas* are associated to the metal electrodes. The chains are clearly present all around the electrodes (Reprinted with permission from Ref. [20]. Source: http://www.biomedical-engineering-online.com/content/13/1/71. Copyright from BioMed Central)

presenting the total force exerted on each spherical particle as

$$\mathbf{F}_{tot}(\mathbf{r}_i) = \mathbf{F}_{ext}(\mathbf{r}_i) + \sum_{j \neq i} \mathbf{F}_{i,j}(\mathbf{r}_i) \text{ for } i = 1, \cdots, N \tag{21}$$

where the dielectrophoretic force $\mathbf{F}_{ext}(\mathbf{r}_i)$ is calculated using Eq. 7 with the external electric field $\mathbf{E}_{ext}(\mathbf{r}_i)$ and the forces $\mathbf{F}_{i,j}(\mathbf{r}_i)$ acting on this particle by other particles are computed as

$$\mathbf{F}_{i,j}(\mathbf{r}_i) = \frac{3}{4\pi\varepsilon_0\varepsilon_f r_{ij}^4}\left[\mathbf{n}_{ij}\left(\mathbf{P}_i \cdot \mathbf{P}_j\right) + \left(\mathbf{n}_{ij} \cdot \mathbf{P}_i\right) \cdot \mathbf{P}_j + \left(\mathbf{n}_{ij} \cdot \mathbf{P}_j\right)\mathbf{P}_j - 5\mathbf{n}_{ij}\left(\mathbf{n}_{ij} \cdot \mathbf{P}_i\right)\left(\mathbf{n}_{ij} \cdot \mathbf{P}_j\right)\right]$$

with $\mathbf{P}_i = 3v_p\varepsilon_0\varepsilon_f\beta^*\mathbf{E}_{ext}(\mathbf{r}_i)$. The external field $\mathbf{E}_{ext}(\mathbf{r}_i)$ is also computed for a dielectrophoretic device without particles. This model does not include the effects of the particle arrangement on the local field strength. If the interparticle hydrodynamic interactions are neglected, the velocity of each particle is calculated using the equation for Stokes drag force and Eq. 21 for the total electrical force $\mathbf{F}_{tot}(\mathbf{r}_i) = 6\pi\eta_f a\mathbf{u}_i$. The simultaneous displacements of interacting particles are then computed using the following coupled equations:

$$\frac{d\mathbf{r}_i}{dt} = \mathbf{u}_i \text{ with } \mathbf{r}_i|_{t=0} = \mathbf{r}_{0,i} \text{ for } i = 1, \ldots, N$$

where $\mathbf{r}_{0,i}$ is the initial particle position.

Simulations [20, 21] within the framework of pairwise models, Eqs. 20 and 21, demonstrate that the effects of the dipole–dipole interactions play a crucial role in the evolution of particles even at a concentration lower than concentrations typically used in dielectrophoretic devices. Microscopic simulations can be used for evaluation of experimental conditions (e.g., electrode configuration, applied voltage and frequency, particle concentration, etc.) under which the contribution of interparticle interactions is insignificant.

Future Direction for Research

Dielectrophoresis is becoming one of the major methods for control and manipulations of micro- and nano-sized particles in various applications. Rapid advances in manufacturing technologies and new materials open the avenue for the fabrication of sophisticated dielectrophoretic devices to address challenges in materials science, analytical chemistry, and life science. The greatest

challenge is to understand the role of electrical and hydrodynamic interparticle interactions in directing and controlling dielectrophoresis in a colloidal dispersion.

Cross-References

- ▶ Dielectrophoresis
- ▶ Lab-on-a-Chip

References

1. Pohl, H.A.: Dielectrophoresis: The Behavior of Neutral Matter in Nonuniform Electric Fields. Cambridge University Press, Cambridge (1978)
2. Pethig, R.: Review article-dielectrophoresis: status of the theory, technology, and applications. Biomicrofluidics **4**, 022811 (2010)
3. Sheng, P., Wen, W.: Electrorheological fluids: mechanisms, dynamics, and microfluidics applications. Ann. Rev. Fluid Mech. **44**, 143–174 (2012)
4. Nili, H., Green, N.G.: Higher-order dielectrophoresis of nonspherical particles. Phys. Rev. E **89**, 063302 (2014)
5. Russel, W.B., Saville, D.A., Schowalter, W.R.: Colloidal Dispersions. Cambridge University Press, Cambridge (1989)
6. Jones, T.B.: Electromechanics of Particles. Cambridge University Press, Cambridge (1995)
7. Kumar, A., Qiu, Z., Acrivos, A., Khusid, B., Jacqmin, D.: Combined negative dielectrophoresis and phase separation in nondilute suspensions subject to a high-gradient ac electric field. Phys. Rev. E **69**, 021402 (2004)
8. Khusid, B., Acrivos, A.: Effects of conductivity in electric-field-induced aggregation in electrorheological fluids. Phys. Rev. E **52**, 1669–1693 (1995)
9. Chen, T.-J., Zitter, R.N., Tao, R.: Laser diffraction determination of the crystalline structure of an electrorheological fluid. Phys. Rev. Lett. **68**, 2555–2558 (1992)
10. Dassanayake, U., Fraden, S., van Blaaderen, A.: Structure of electrorheological fluids. J. Chem. Phys. **112**, 3851–3858 (2000)
11. Dobnikar, J., Snezhko, A., Yethiraj, A.: Emergent colloidal dynamics in electromagnetic fields. Soft Matter **9**, 3693–3704 (2013)
12. van Blaaderen, A., Dijkstra, M., van Roij, R., Imhof, A., Kamp, M., Kwaadgras, B.W., Vissers, T., Liu, B.: Manipulating the self assembly of colloids in electric fields. Eur. Phys. J. Spec. Top. **222**, 2895–2909 (2013)
13. Khusid, B., Acrivos, A.: Effects of interparticle electric interactions on dielectrophoresis in colloidal suspensions. Phys. Rev. E **54**, 5428–5435 (1996)
14. Wood, J.A., Docoslis, A.: Electric-field induced phase transitions of dielectric colloids: impact of multiparticle effects. J. Appl. Phys. **111**, 094106 (2012)
15. Moncada-Hernandez, H., Nagler, E., Minerick, A.R.: Theoretical and experimental examination of particle–particle interaction effects on induced dipole moments and dielectrophoretic responses of multiple particle chains. Electrophoresis **35**, 1803–1813 (2014)
16. Velev, O.D., Gangwala, S., Petsev, D.N.: Particle-localized AC and DC manipulation and electrokinetics. Ann. Rep. Prog. Chem. Sect. C **105**, 213–246 (2009)
17. Sullivan, M., Zhao, K., Harrison, C., Austin, R.H., Megens, M., Hollingsworth, A.D., Russel, W.B., Cheng, Z., Mason, T., Chaikin, P.M.: Control of colloids with gravity, temperature gradients, and electric fields. J. Phys. Condens. Matter **15**, S11–S18 (2003)
18. Kumar, A., Acrivos, A., Khusid, B., James, C.D., Jacqmin, D.: Conveyor-belt method for assembling microparticles into large-scale structures using electric fields. Appl. Phys. Lett. **90**, 154104 (2007)
19. Bennett, D., Khusid, B., Galambos, P.C., James, C.D., Okandan, M., Jacqmin, D., Acrivos, A.: Combined field-induced dielectrophoresis and phase separation for manipulating particles in microfluidics. Appl. Phys. Lett. **83**, 4866–4868 (2003)
20. Camarda, M., Fisicaro, G., Anzalone, R., Scalese, S., Alberti, A., La Via, F., La Magna, A., Ballo, A., Giustolisi, G., Minafra, L., Cammarata, F.P., Bravatà, V., Forte, G.I., Russo, G., Gilardi, M.G.: Theoretical and experimental study of the role of cell-cell dipole interaction in dielectrophoretic devices: application to polynomial electrodes. Biomed. Eng. **13**, 71 (2014). Online
21. Kadaksham, J., Singh, P., Aubry, N.: Dielectrophoresis induced clustering regimes of viable yeast cells. Electrophoresis **26**, 3738–3744 (2005)

AC Electrokinetic Manipulation

▶ Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly

AC Electrokinetics

▶ Dielectrophoresis

AC Electrokinetics of Colloidal Particles

► AC Electrokinetics of Nanoparticles

AC Electrokinetics of Nanoparticles

Hossein Nili[1,2,3] and Nicolas G. Green[2,3]
[1]Nano Research Group, University of Southampton, Highfield, Southampton, UK
[2]School of Electronics and Computer Science, University of Southampton, Highfield, Southampton, UK
[3]Nano Research Group, Faculty of Physical and Applied Sciences, University of Southampton, Southampton, UK

Synonyms

AC electrokinetics of colloidal particles; AC electrokinetics of sub-micrometer particles

Definition

AC electrokinetics is the name given to a group of techniques that utilize alternating (AC) electric fields to move dielectric particles in suspension. AC electrokinetics of nanoparticles refers to methods of exerting electrical force and/or torque on particles of nanometer dimensions, examples of which are viruses, macromolecules, and colloidal particles.

Introduction

Dielectrics do not bear a net charge but rather polarize when subjected to electric fields. When dielectric particles in suspension are subjected to electric fields, polarization results in charge accumulation at the particle/electrolyte interface. Electrode polarization may also occur, giving rise to the buildup of a double layer of ions and counterions at the electrode/electrolyte interface. Particle polarization is represented by effective dipole and higher-order moments. AC electrokinetic forces and torques on particles result from interactions of an applied electric field with the effective moments. Electric field-induced fluid motion is also studied under AC electrokinetics due to the important effect it can have on particle behavior. Electric field interactions with the double layer and gradients in fluid permittivity and conductivity (resulting from localized heating of the medium) are the main factors causing fluid motion under AC electric fields.

AC electrokinetic techniques are particularly fitted to lab-on-a-chip applications where multiple processes involving manipulation, separation, or characterization of biological particles, mostly dielectrics, are integrated onto a single chip. The direction and magnitude of AC electrokinetic forces and torques can be easily controlled by varying electric field frequency and geometry. The techniques are advantageous over alternative means of moving particles in their noninvasive nature and easy integration onto micro-devices, not relying on any moving parts.

It was long believed that for sub-micrometer particles, AC electrokinetic forces would be overwhelmed by thermal effects such as Brownian motion. The strong electric fields required to exert sufficient electrical force to dominate particle behavior at the nanometer scale were believed to generate excessive heat giving rise to strong fluid motion that would hinder AC electrokinetic interactions. Thanks to fabrication techniques that realized micro- and nano-electrode geometries, electric fields of sufficient strength to overcome Brownian motion of nanoparticles could be generated with the application of modest voltages, avoiding excessive heating of the suspension.

AC Electric Field Interactions with Nanoparticle Suspensions

When an electric field is applied to dielectric particles in suspension, surface charge accumulates

at interfaces between the dielectrics due to the differences in electrical properties. Since the polarizabilities of each dielectric are frequency dependent, the magnitude of the surface charge is also frequency dependent, and the total (complex) permittivity of the system exhibits dispersions solely due to the polarization of the interfaces. This is referred to as the Maxwell-Wagner interfacial polarization. For a spherical particle of radius R, the effective dipole moment is given by

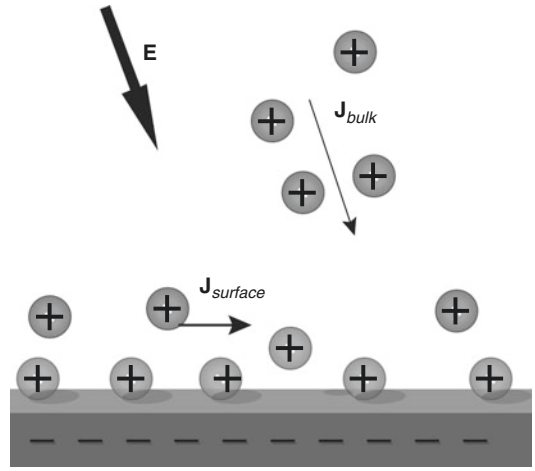
$$\boldsymbol{p} = 4\pi\epsilon_m R^3 [K(\omega)]\boldsymbol{E} \tag{1}$$

where ω is the frequency of the applied electric field $\boldsymbol{E}$ and $K(\omega)$, known as the Clausius-Mossotti factor, describes a relaxation in the polarizability of the particle with a relaxation time

$$\tau_{MW} = \frac{\epsilon_p + 2\epsilon_m}{\sigma_p + 2\sigma_m} \tag{2}$$

The angular frequency $\omega_{MW} = 2\pi f_{MW} = 1/\tau_{MW}$ is often referred to as the Maxwell-Wagner relaxation frequency.

During the first half of the twentieth century, dielectric spectroscopic measurements of suspensions of nanoparticles identified that the Maxwell-Wagner relaxation frequency for charged nanoparticles was higher than expected. Later, O'Konski [1] showed that the dielectric properties of nanoparticles in this frequency regime were dominated by surface conductance effects. The high value of particle conductivity determined from dielectric measurements was explained by the inclusion of a surface conductance component in the derivation of the particle's dipole moment. The model developed by O'Konski is shown in Fig. 1. He assumed that the flux due to the transport of charge carriers, associated with the fixed charge on the particle surface, could be added to the flux due to the transport of bulk charge to and from the surface. With this assumption, he derived the potential around the spherical particle and derived the equation for the dipole moment of a dielectric sphere (Eq. 1) but with the particle conductivity given by the sum of the bulk conductivity of the particle and a surface conductivity term



AC Electrokinetics of Nanoparticles, Fig. 1 Schematic of the mechanism with which surface conductance affects interfacial polarization of nanoparticles in suspension. The AC electric field produces both a bulk flow of ions and a surface flow around the particle

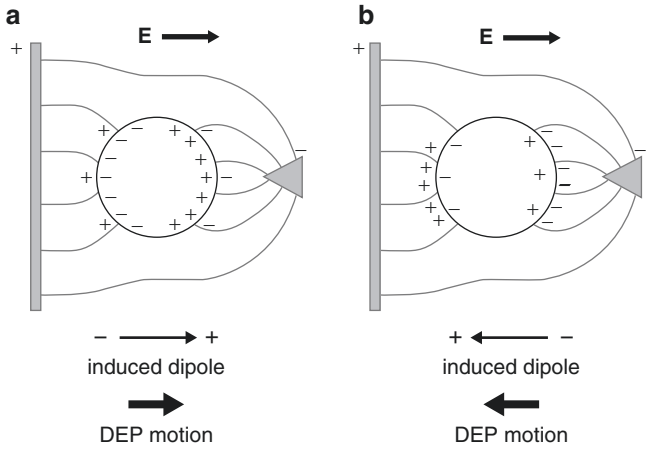
$\sigma_p = \sigma_{p,\,\text{bulk}} + \sigma_{p,\,\text{surface}}$. The surface conductivity of the particle is given by

$$\sigma_{p,\,\text{surface}} = \frac{2K_s}{R} \tag{3}$$

where K_s is the surface conductance.

AC Electrokinetic Techniques

Dielectrophoresis (DEP): Fig. 2 shows the underlying principle of dielectrophoresis as the most widely used of AC electrokinetic techniques. The gradient in electric field strength gives rise to unequal forces experienced by polarization charges at the particle/electrolyte interface, hence a net force on the particle moving it toward or away from regions of high field intensity depending on whether the particle is more or less polarizable than the suspending medium. Polarization charges at dielectric interfaces form induced dipole and higher-order moments. As shown in Fig. 2, the notion of an induced dipole helps better visualize the direction of dielectrophoretic force on particles. The particle would move toward or away from regions of high electric field strength when the induced dipole is



AC Electrokinetics of Nanoparticles, Fig. 2 Dielectrophoresis – interaction of a nonuniform electric field with a dielectric particle in suspension gives rise to unequal forces experienced by polarization charges at the particle/electrolyte interface. The result is a net dielectrophoretic force that moves the particle (**a**) toward or (**b**) away from regions of high electric field intensity depending on whether the particle is more or less polarizable than its suspending medium

aligned with or against the electric field, respectively. In more complex geometries, particularly those involving nonspherical particles subjected to highly divergent electric fields, higher-order moments will need to be accounted for as they could contribute notably to the DEP force on particles [2, 3].

If particle dimensions are smaller than a characteristic length scale of electric field nonuniformity, higher-order terms can be neglected, and the time-averaged dielectrophoretic force on a spherical particle of radius R suspended in a fluid of dielectric constant ϵ_m is given by [4]

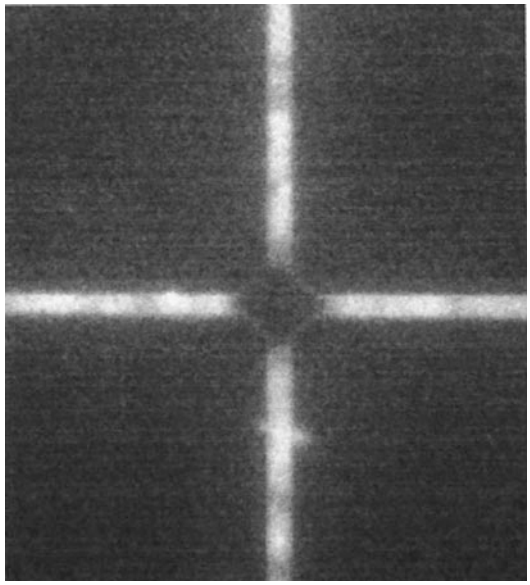
$$\langle \boldsymbol{F}_{DEP} \rangle = \pi \epsilon_m R^3 Re[K(\omega)] \nabla |\boldsymbol{E}|^2 \qquad (4)$$

It is understood from Eq. 1 that the dielectrophoretic force scales with particle volume and also inversely with the cube of a characteristic dimension of the electrode geometry. It is due to the latter proportionality that micro- and nanoelectrode geometries are capable of exerting electric fields of sufficient strength to move particles as small in volume as nanoparticles, overcoming Brownian and field-induced fluid motion.

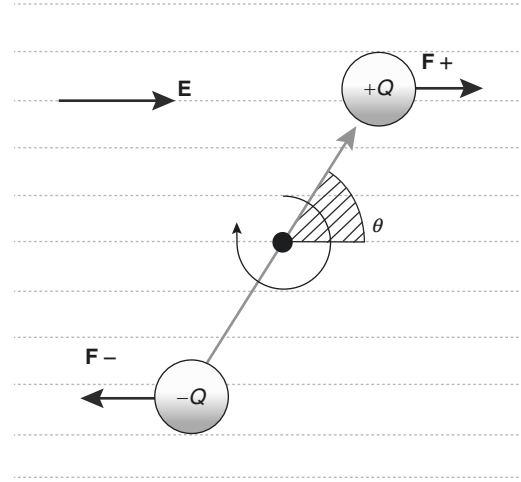
Although first observations of the effect date back to earlier times, Pohl was the first to use the term dielectrophoresis for the force exerted by a nonuniform electric field on a dielectric [5]. He observed coagulation of carbon particles from a polymer solution upon application of a DC or AC electric field. Pohl used the term "dielectrophoresis" to distinguish the effect from electrophoresis, which describes motion of charged particles under DC (and not AC) electric fields.

Since its advent, dielectrophoresis has been used in a broad range of applications. Separation, as a fundamental part of many processes in micro-total-analysis systems (μTAS), has been diversely and controllably accomplished using dielectrophoresis. The dependence of the dielectric properties of particles on their structure and composition has broadened the applicability of DEP-based separation techniques. In electrode geometries with well-defined regions of electric field maxima and minima, particles of different properties can be separated into subpopulations. Figure 3 shows the collection of 280 nm tobacco mosaic viruses (TMV) in a polynomial electrode configuration [6]. As the TMV particles are more polarizable than the suspending medium, they gather at electrode edges where the electric field is strongest.

Dielectrophoresis has been used extensively for manipulation and characterization of



AC Electrokinetics of Nanoparticles, Fig. 3 Dielectrophoretic collection of nanoparticles – applying a 6 MHz AC electric field leads to the collection of 280 nm tobacco mosaic viruses (*TMV*) at the electrode edges where the field is strongest [4]



AC Electrokinetics of Nanoparticles, Fig. 4 Principle of electro-orientation – in a uniform field **E** (indicated by the vector and the *dotted field lines*), the two charges experience equal and opposite forces resulting in a torque about the center point of the dipole

nanoparticles. One of the most notable application areas is diagnostics and healthcare, toward which a huge amount of effort has been directed involving sub-micrometer biological particles such as viruses, DNA, and chromosomes [7]. Dielectrophoresis has also been used for separation of metallic from semiconducting carbon nanotubes [8], assembly of nanoparticles into micro-wires [9], and three-dimensional focusing of nanoparticles in microfluidic channels [10]. With the emergence of nano-electrode geometries, there is huge prospect for further applications involving dielectrophoresis of nanoparticles including the fabrication of a new generation of electronic devices and sensors [7].

Electro-orientation: As shown in Fig. 4, when a dipole sits in a uniform field, each charge on the dipole experiences an equal and opposite force tending to align the dipole with the electric field. The effect can also be observed in nonuniform and rotating electric fields and is the basis of a phenomenon known as electro-orientation.

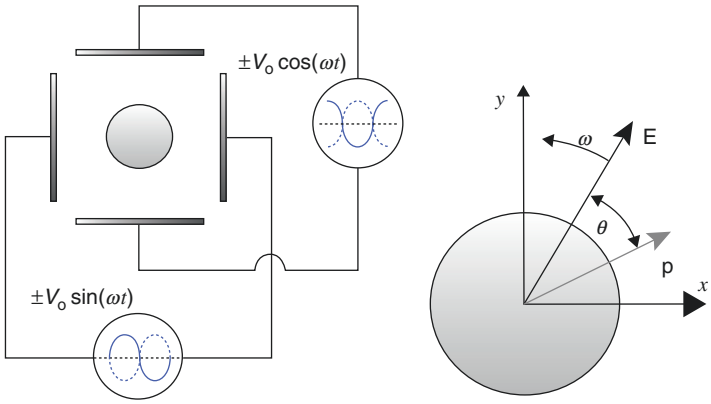
A dielectric particle of no net charge in a uniform electric field will be subject to no net force but will experience a torque given by

$$\Gamma = \boldsymbol{p} \times \boldsymbol{E} \tag{5}$$

This torque always tends to align the dipole with the electric field. However, small dipoles, such as those of nanoparticles, will not completely align with the field due to the dominating effect of Brownian motion. In typical dielectric spectroscopy measurements, the electric fields are only of sufficient magnitude to align individual molecules by fractions of degrees, but the total effect of alignments of many molecules gives rise to a large net or average alignment [11].

In AC electric fields, electro-orientation is frequency dependent with nonspherical particles orienting along different axes in different frequency ranges. The set of crossover frequencies is referred to as the orientation spectrum. The orientation spectra of elongated biological particles such as erythrocytes and bacteria have been used to study the dielectric properties of these particles [12].

Electro-rotation (**ROT**): Dielectric particles subjected to rotating electric fields experience a



AC Electrokinetics of Nanoparticles, Fig. 5 Electro-rotation – (**a**) A schematic diagram of an electro-rotation setup. Four signals, successively 90° out of phase, are applied to four electrodes encircling the particle. (**b**) Schematic diagram showing how the induced dipole moment of a particle lags behind a rotating applied electric field

torque based on a principle which is somewhat different from that of electro-orientation. As mentioned previously, applying an electric field to a dipole will result in the exertion of a (electro-orientational) torque that tends to align the dipole with the electric field. Alignment of the dipole with the electric field vector will not be immediate as it takes a finite amount of time for polarization charges to move toward the particle/electrolyte interface and form a dipole. In a rotating electric field, where the field vector changes direction, this time delay gives rise to a (electro-rotational) torque, as shown in Fig. 5. The figure also shows an example electro-rotation setup, where voltages of 90° phase differences are applied to successive electrodes encircling the particle to generate a rotating electric field.

The first order electro-rotational torque on a spherical particle of radius R is given by [13]

$$\Gamma_{ROT} = -4\pi\epsilon_m R^3 Im[K(\omega)]|\boldsymbol{E}|^2 \qquad (6)$$

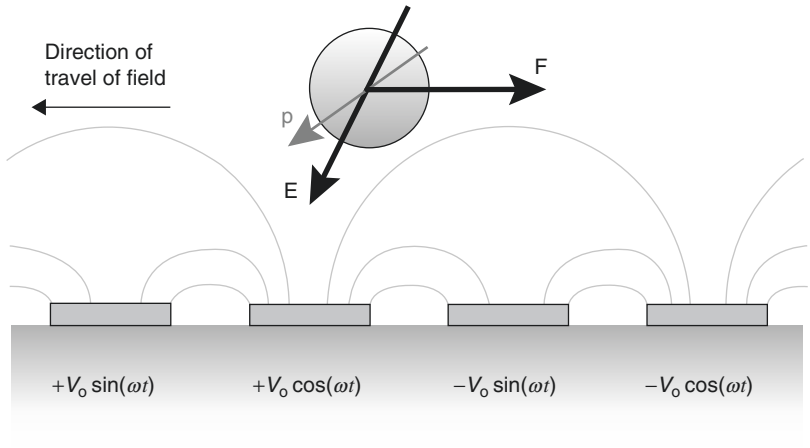
where $Im[K(\omega)]$ denotes the imaginary part of the Clausius-Mossotti factor. The particle will rotate with or counter to the applied field depending on whether $Im[K(\omega)]$ is negative or positive, respectively. Accounting for viscous drag force from the suspending fluid, the rotation rate of the particle is given by [13]

$$\Omega = -\frac{\epsilon_m Im[K(\omega)]|\boldsymbol{E}|^2}{2\eta} \qquad (7)$$

where η denotes fluid viscosity. Variations with frequency of the rotation rate are referred to as electro-rotation spectra.

Two important distinctions can be readily identified from a comparison of Eqs. 1 and 3 for dielectrophoretic force and electro-rotational torque. Firstly, the ROT torque is proportional to the square of the electric field magnitude, while the DEP force is a function of the gradient of the square of the field magnitude – and is therefore zero in uniform electric fields. Secondly, the electro-rotational torque depends on the imaginary rather than the real part of the Clausius-Mossotti factor. As a result, particles may experience both dielectrophoresis and electro-rotation with the relationship between the two determined by the dielectric properties of the particles and their suspending media.

Initially observed as a nuisance in cell electrofusion procedures, electro-rotation has since developed into a useful means of measuring dielectric properties and membrane electrical parameters of biological particles. Rotational spectra are analyzed to determine fundamental cell properties and to monitor changes in these properties upon different types of treatments [12]. Among the major works done on electro-rotation of nanoparticles are investigations by Washizu and co-workers into the torque-speed behavior of the flagellar motor mechanism of bacterial cells [14]. In a more recent work, electro-rotation and electro-orientation of metal nanowires were studied experimentally and theoretically [15].



AC Electrokinetics of Nanoparticles, Fig. 6 Traveling-wave dielectrophoresis – schematic diagram of a linear traveling-wave dielectrophoresis array and the consecutive phase-shifted signals required to generate the traveling electric field. Also shown are the approximate field lines for time $t = 0$, the electric field, and the dipole moment induced in the particle together with the force on the particle

Traveling-wave dielectrophoresis (twDEP): Traveling-wave DEP can be considered as the linear analog of electro-rotation where voltages of 90° phase difference are applied to successive electrodes that are laid out as tracks, rather than being arranged in a circle. This generates an electric field wave which travels along the electrodes. The dipole induced in the particle moves with the electric field but lags behind the field, as in electro-rotation. As shown in Fig. 6, the result is the induction of a force, rather than a torque, given – for a spherical particle of radius R – by [16]

$$\boldsymbol{F}_{twDEP} = -\frac{4\pi\epsilon_m R^3 Im[K(\omega)]|\boldsymbol{E}|^2}{\lambda} \qquad (8)$$

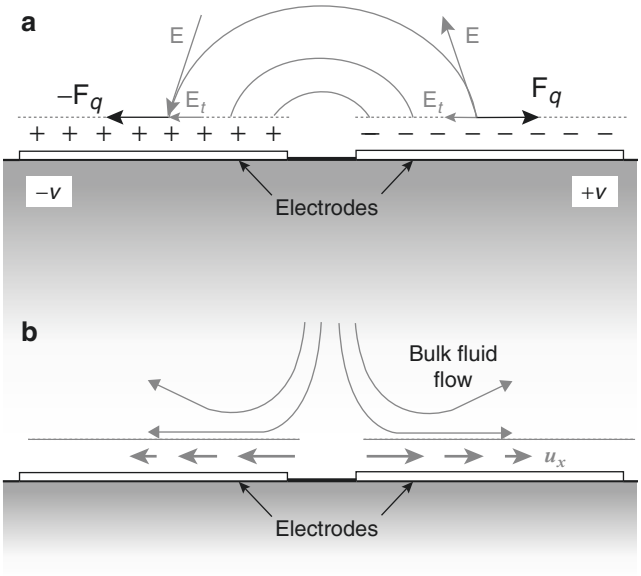
where λ is the wavelength of the traveling wave. The negative sign in Eq. 8 indicates that, as shown in Fig. 6, the twDEP force propels particles in the opposite direction to the moving field vector. For a finite twDEP force to be exerted on a particle, two criteria need to be met: (a) the particle must experience a DEP force that levitates it above the electrode array, and (b) some loss mechanism needs to be present for the imaginary part of the Clausius-Mossotti factor to be nonzero.

Masuda et al. [17] were the first to describe the principles of traveling-wave DEP. Since then, twDEP has been used for characterizing and separating cells and microorganisms [18]. Nanoparticle manipulation and separation has also been accomplished using the traveling electric field above interdigitated electrodes [19].

AC Electrokinetic Fluid Motion

AC electroosmosis (ACEO): AC electric fields applied to a suspension of dielectric particles can give rise to fluid motion, in an effect known as AC electroosmosis [20]. The effect arises from interaction of the electric field with charges at the double layer and is only observed if the field is nonuniform.

The mechanism for AC electroosmosis is shown in Fig. 7. Applying equal and opposite voltages to successive electrodes gives rise to the electric field $\boldsymbol{E}$ with tangential component $\boldsymbol{E}_t$ outside the double layer and an induced charge on each electrode. The induced charge experiences a force $\boldsymbol{F}_q$ due to the action of the tangential field, resulting in fluid flow. Figure 7a shows the system for one half-cycle of an AC field. In the other half-cycle, the sign of the potential, that of the induced charge, and the direction of the tangential field are all opposite. As a result, the direction of the force vector remains the same giving a nonzero



AC Electrokinetics of Nanoparticles, Fig. 7 AC electroosmosis – (**a**) schematic diagram outlining the mechanism of AC electroosmosis. (**b**) The interaction of the tangential field at the surface with the charge in the double layer gives rise to a surface fluid velocity and a resulting bulk flow

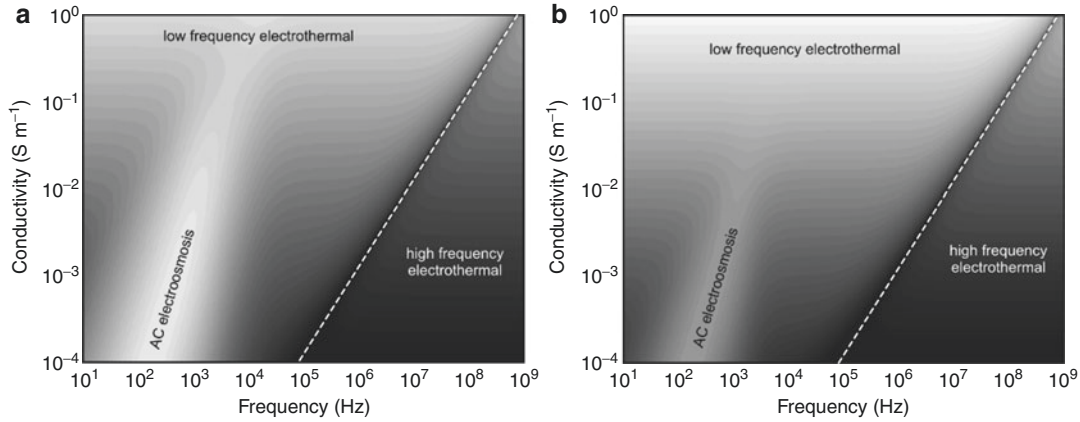
time-averaged force and steady-state fluid flow occurs, as shown in Fig. 7b.

Electrothermal fluid flow: At higher frequencies (>100 kHz), AC electroosmotic flow is negligible and the dominant fluid flow is due to electrothermal effects. This type of flow requires temperature gradients in the fluid and these can be generated both by internal and/or external sources. The internal source is Joule heating, where the electric field causes power dissipation in the fluid and the corresponding temperature rise diffuses through the system. This gives rise to gradients in the conductivity and permittivity; the electric field acts on these gradients to give a body force on the fluid and consequently a flow. In this case, the velocity of the flow is proportional to the temperature rise in the fluid, which is in turn proportional to the conductivity of the electrolyte and the magnitude of the electric field. As a result, this type of fluid flow occurs mainly at high electrolyte conductivities.

Apart from the conductivity dependence of the electrothermal effect, the magnitude of the flow is also frequency dependent. The ratio of low- to high-frequency limiting values is approximately 10:1, and the flow changes magnitude and direction at a frequency of the order of the charge relaxation frequency of the medium. The change in the velocity of the fluid as a function of the two parameters, frequency of the applied field and conductivity of the electrolyte, is complicated and is summarized in Fig. 8. In these plots, the logarithm of the magnitude of the flow velocity is indicated by a gray scale as a function of the conductivity of the electrolyte and frequency of the field, for two applied voltages. Both figures demonstrate the regions in which strong fluid flow occurs and also which regions might be "safe" from fluid motion.

Review of AC Electrokinetic Forces on Nanoparticles

As described through the different mechanisms responsible, particle behavior under AC electric fields is determined by a variety of factors. The total force on any particle is given by the sum of many forces including Brownian, dielectrophoretic, and hydrodynamic forces, the latter arising from fluid motion. These forces can be comparable to, or in certain circumstances much stronger than, the dielectrophoretic force exerted on the particle. At certain combinations of medium conductivity and electric field frequency and strength, the hydrodynamic forces can dominate particle behavior. Under other experimental conditions, these forces are negligible, and as a result,



AC Electrokinetics of Nanoparticles, Fig. 8 Schematic frequency/conductivity maps of the two types of electric field-driven fluid flow: AC electroosmosis and electrothermal. The *white dotted line* indicates the frequency at which electrothermal flow changes direction. The velocities are plotted on a log gray scale and were determined for the same position (10 μm from the edge of the electrode) and applied signals: (**a**) 1 V and (**b**) 10 V on each electrode. Note that the ratio of AC electroosmotic flow is much less at 10 V

small changes in the dielectric properties of the particles can be detected and measured by DEP.

Ramos and co-workers have conducted order-of-magnitude calculations for the forces on nanoparticles in AC electric fields [21]. It has been shown that for low-conductivity suspensions, Joule heating has little effect on system temperature and is therefore negligible. With ±10 V applied to a microelectrode system with a medium conductivity of 10 mS/m, temperature rises of the order of only 1 °C have been observed. In contrast, electrothermal forces have been shown to be of sufficient strength to compete with the dielectrophoretic force. However, it is important to note that the DEP force varies much more rapidly than electrothermal forces upon proximity with electrode edges. Applying ±5 V to a system of parallel finger microelectrodes and a 10 ms/m-conductivity suspension of 282 nm particles has been shown to generate a fluid velocity of 5 μm/s against a DEP-induced velocity of 1.8 μm/s. Near electrode edges, DEP-induced and electrothermal velocities have been reported to be 200 μm/s and 50 μm/s, respectively. Changes in fluid density arising from field-induced temperature gradients have been shown to be on the order of 0.01 % per degree. As a result, natural convection has been found to be very small compared to electrothermal and DEP forces.

Brownian motion was for long considered the biggest obstacle in AC electrokinetic motion of sub-micrometer particles. Pohl had estimated that electric field strengths required to overcome Brownian motion of nanoparticles would (a) not be realizable and (b) lead to excessive heating, and hence electrothermal motion, of the fluid, thereby hinders DEP motion of particles [2]. Experimental findings have shown that Pohl had largely overestimated the electric field strength required to overcome thermal effects, the reason being his comparison of thermal energy with dielectrophoretic *potential* rather than *force*, the latter being the gradient of the former [6]. It has been shown that DEP forces required to overcome Brownian motion of nanoparticles are in the sub-pico-newton range and can be easily generated using micro- and nano-electrode geometries by application of voltages modest enough to avoid excessive heating of the medium. The observable deterministic force required to move a 282 nm particle has been found to be of the order of 0.01 pN over a 1-s time frame of observation, requiring an electric field on the order of 100 kV/m, which can be generated by applying voltages on the order of 1 V across a 10 μm gap.

In summary, although hydrodynamic effects can complicate AC electrokinetic motion of nanoparticles, it has been shown that the forces

can be predicted and therefore controlled. Consequently, it is expected that recent technological advances in AC electrokinetics could be applied, together with field-induced fluid motion, to develop new methods for the characterization, manipulation, and separation of nanoparticles.

Cross-References

- ▶ AC Electroosmosis: Basics and Lab-on-a-Chip Applications
- ▶ Dielectrophoresis
- ▶ Dielectrophoresis of Nucleic Acids
- ▶ Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes

References

1. O'Konski, C.T.: Electric properties of macromolecules V: theory of ionic polarization in polyelectrolytes. J. Phys. Chem. **64**, 605–619 (1960)
2. Nili, H., Sun, T., Green, N.G.: Higher-order dielectrophoretic force characterisation of non-spherical particles. J. Phys. Conf. Ser. **301**, 012061 (2011)
3. Nili, H., Green, N.G.: Higher-order dielectrophoresis of nonspherical particles. Phys. Rev. E **89**, 063302 (2014)
4. Pohl, H.A.: Dielectrophoresis: The Behavior of Neutral Matter in Nonuniform Electric Fields. Cambridge University Press, Cambridge (1978)
5. Pohl, H.A.: The motion and precipitation of suspensoids in divergent electric fields. J. Appl. Phys. **22**, 869–871 (1951)
6. Green, N.G.: Dielectrophoresis of Sub-micrometre particles. Thesis, University of Glasgow, Glasgow. (1998)
7. Pethig, R.: Review article – dielectrophoresis: status of the theory, technology, and applications. Biomicrofluidics **4**, 022811 (2010)
8. Krupke, R., Hennrich, F., von Lohneysen, H., Kappes, M.M.: Separation of metallic from semiconducting single-walled carbon nanotubes. Science **301**, 344–347 (2003)
9. Hermanson, K.D., Lumsdon, S.O., Williams, J.P., Kaler, E.W., Velev, O.D.: Dielectrophoretic assembly of electrically functional microwires from nanoparticle suspensions. Science **238**, 1082–1086 (2001)
10. Morgan, H., Holmes, D., Green, N.G.: 3D focusing of nanoparticles in microfluidic channels. IEE Proc. Nanobiotechnol **150**, 76–81 (2003)
11. Morgan, H., Green, N.G.: AC Electrokinetics of Colloids and Nanoparticles. Research Studies Press, Baldock, Herts, UK (2003)
12. Jones, T.B.: Electromechanics of Particles. Cambridge University Press, Cambridge (1995)
13. Arnold, W.M., Zimmermann, U.: Electro-rotation– development of a technique for dielectric measurements on individual cells and particles. J. Electrost. **21**, 151–191 (1988)
14. Washizu, M., Shikida, M., Aizawa, S., Hotani, H.: Orientation and transformation of flagella in electrostatic field. IEEE Trans. IAS **28**, 1194–1202 (1992)
15. Arcenegui, J.J., Garcia-Sanchez, P., Morgan, H., Ramos, A.: Electro-orientation and electrorotation of metal nanowires. Phys. Rev. E **88**, 063018 (2013)
16. Hughes, M.P.: AC electrokinetics: applications for nanotechnology. Nanotechnology **11**, 124–132 (2000)
17. Masuda, S., Washizu, M., Iwadare, M.: Separation of small particles suspended in liquid by nonuniform traveling field. IEEE Trans. Ind. Appl. **23**, 474–480 (1987)
18. Hagedorn, R., Fuhr, G., Muller, T., Schnelle, T., Schnakenberg, U., Wagner, B.: Design of asynchronous dielectric micromotors. J. Electrost. **33**, 159–185 (1994)
19. Li, W.H., Du, H., Chen, D.F., Shu, C.: Analysis of dielectrophoretic electrode arrays for nanoparticle manipulation. Comp. Mater. Sci. **30**, 320–325 (2004)
20. Ramos, A., Morgan, H., Green, N.G., Castellanos, A.: AC electric-field-induced fluid flow in microelectrodes. J. Colloid Interface Sci. **217**, 420–422 (1999)
21. Ramos, A., Morgan, H., Green, N.G., Castellanos, A.: AC electrokinetics: a review of forces in microelectrode structures. J. Phys. D. Appl. Phys. **31**, 2338–2353 (1998)

AC Electrokinetics of Sub-micrometer Particles

- ▶ AC Electrokinetics of Nanoparticles

AC Electroosmosis: Basics and Lab-on-a-Chip Applications

Pablo García-Sánchez and Antonio Ramos
Departamento de Electrónica y Electromagnetismo, Universidad de Sevilla, Sevilla, Spain

Synonyms

Induced charge electroosmosis

Definition

Microelectrode structures subjected to AC voltages can generate flow of aqueous solutions by the action of the AC field on the charges induced by itself at the electrode-electrolyte interface, i.e., induced charges in the electrical double layer (EDL). The phenomenon is called *AC electroosmosis* (ACEO) in analogy with the "classical" electroosmosis, where fluid flow is generated by the action of an applied electric field on the mobile charge of a given solid-liquid interface [1]. The main difference is that the applied field in ACEO is responsible for both inducing the charge and pulling on it. The term induced-charge electrokinetics (ICEK) has been proposed to refer to all phenomena where the electric field acts on the electrical double layer induced by itself [2].

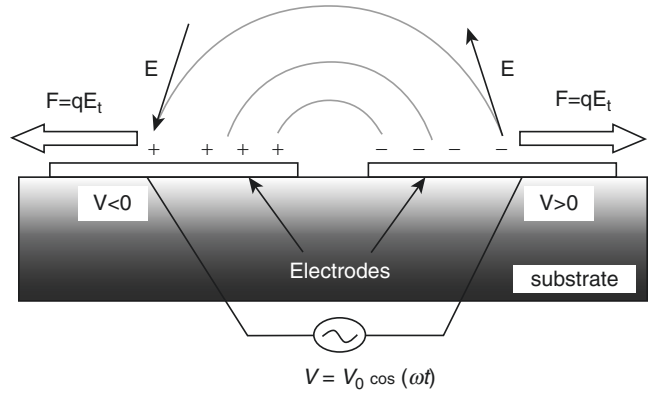
Basic Mechanism

Possibly, the simplest system for the study of AC electroosmosis consists of a couple of coplanar electrodes covered by an aqueous electrolyte. The electrodes are subjected to a harmonic potential difference of amplitude V_0 and frequency ω, $\boldsymbol{V(t) = V_0 \cos(\omega t)}$. Figure 1 shows a sketch of the physical system at a given time of the AC signal cycle. The applied field attracts counterions at the solid-liquid interfaces with a certain delay, and since the electric field is nonuniform, there appears an electrical force pulling the liquid outward $\boldsymbol{F = qE_t}$ ($\boldsymbol{E_t}$ is the component of the electric field tangential to the electrode surfaces). Note that in the other half cycle both the tangential field and the induced charge change sign and, therefore, the force direction remains the same and the time-averaged force is nonzero.

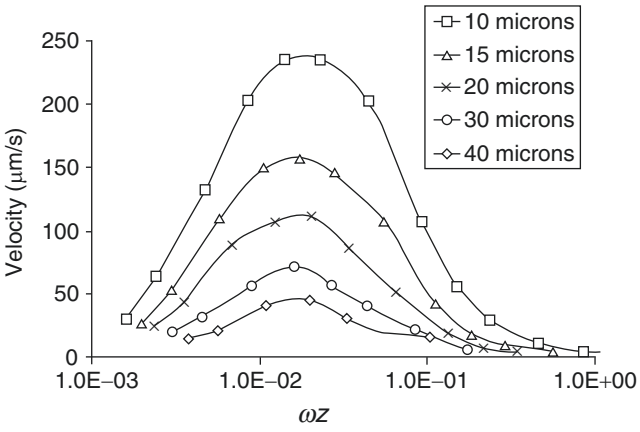
The fluid velocity generated by this mechanism is frequency dependent. At low frequencies the electrical charges have sufficient time to completely screen the electric field at the electrodes; the electric field in the electrolyte bulk vanishes and so does the induced velocity. In this view, the electrodes are supposed to be perfectly polarizable, i.e., no faradaic currents occur. At high frequencies the electrical charges have no time to follow the electric field, the induced charge at the electrode tends to zero, and therefore, the electrical force is negligible. The theoretical expectation for the velocity versus frequency plot is a bell-shaped curve, in accordance with the experimental observations (Fig. 2).

Simple Model for Small Voltages

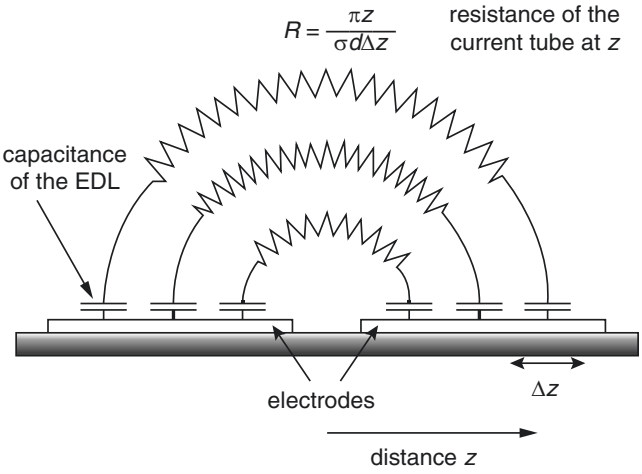
The electroosmotic velocity generated at metal surfaces can be computed from the Helmholtz-Smoluchowski formula, $\boldsymbol{v_{HS} = \varepsilon \zeta E_t / \eta}$, where ε is the dielectric constant of the electrolyte, ζ is the zeta potential of the solid-liquid interface (i.e., voltage drop across the electrical double layer), and η is the liquid viscosity. For small voltage amplitudes, the relation between ζ and the charge



AC Electroosmosis: Basics and Lab-on-a-Chip Applications,
Fig. 1 Basics of the ACEO mechanism. Electrical charge is induced at the electrode-electrolyte interface. The tangential component of the electric field acts on this charge giving rise to a surface fluid velocity



AC Electroosmosis: Basics and Lab-on-a-Chip Applications, Fig. 2 Fluid velocity at several positions on the electrode surface. The velocity is plotted against the product of the angular frequency and the distance of measurement from the center of the two electrodes (Reprinted with permission from [1], © 1999 Elsevier)



AC Electroosmosis: Basics and Lab-on-a-Chip Applications, Fig. 3 RC-circuit model for computing the charge density in the electrical double layer as a function of the position on the electrodes. The resistance of the liquid is modeled by the resistors. The capacitance of the electrical double layer is represented by the capacitors

in the EDL is linear, and the formula reads $v_{HS} = \sigma_q \lambda_D E_t / \eta$, with λ_D the Debye length and σ_q the charge per unit area in the EDL.

In ACEO, σ_q is the induced charge in the EDL by the AC potential. A constant (non-oscillating) term in the charge on the electrode-electrolyte interface is not considered although it might exist, the so-called intrinsic surface charge. However, in this linear model, the time-average velocity due to that charge is zero. In the system of symmetric electrodes, σ_q can be computed from the RC-circuit model shown in Fig. 3. In this model the electrical current in the electrolyte is discretized in current tubes of semicircular shape and width Δz. The resistance of one of these tubes is $R = \pi z / \sigma d \Delta z$, where z is the distance to the interelectrode gap center, σ the electrolyte conductivity, and d the depth of the electrodes. The electrical current in the tubes charges the electrical double layer, represented in this model by a distributed capacitor with a capacitance per surface area that, at low amplitudes, can be estimated from the Debye-Hückel theory as $\varepsilon d \Delta z / \lambda_D$. The value of σ_q at any position can be computed from $\sigma_q = \varepsilon V_d / \lambda_D$, where V_d is the voltage drop across the EDL, and it is a function of the position z:

$$V_d(z) = \frac{V_0}{1 + j\omega\pi\varepsilon z / \sigma\lambda_D} \tag{1}$$

and the tangential electric field is obtained from $E_t = -\partial V_d / \partial z$. The time average of the Helmholtz-Smoluchowski formula reads:

$$\langle v_{HS} \rangle = \frac{1}{2}\mathrm{Re}\left[\frac{\sigma_q E_t^* \lambda_D}{\eta}\right] = \frac{\varepsilon V_0^2 \Omega^2}{8\eta z\left(1+\Omega^2\right)^2} \quad (2)$$

where * indicates the complex conjugate and the nondimensional frequency Ω is written as $\Omega = \omega\pi\varepsilon z/2\sigma\lambda_D$. Maximum velocity is predicted for $\Omega = 1$, i.e., $\omega_{\max} = 2\sigma\lambda_D/\pi\varepsilon z$. The frequency for maximum velocity is dependent on position on the electrode z, and the electroosmotic velocity vanishes for either $\Omega << 1$ or $\Omega >> 1$, as expected. The model is in qualitative agreement with experimental measurements [3].

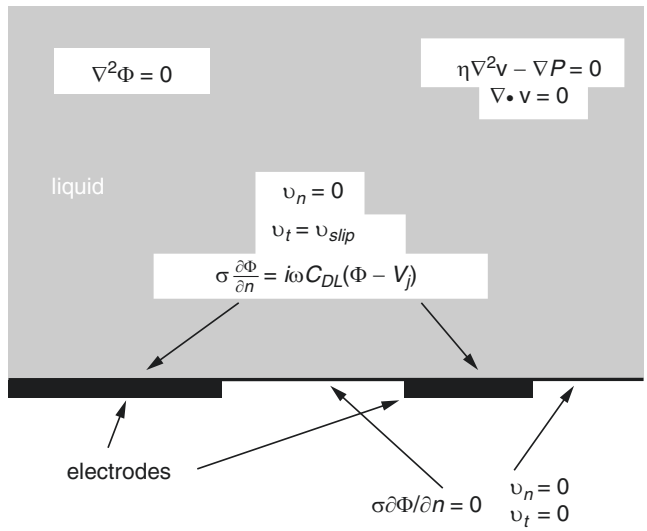
A rigorous study of ACEO flows for small potentials can be found in Ref. [4], where the electrokinetic equations are solved in the limit of thin electrical double layer, i.e., the Debye length is much smaller than the size of the electrodes. In this limit, the fluid velocity in the liquid can be obtained from Stokes equation with boundary condition of slip velocity on the electrodes, the electroosmotic velocity. For electrode j, subjected to a harmonic potential $V(t) = \mathrm{Re}\left[V_j \exp(i\omega t)\right]$, the slip velocity is computed according to

$$v_{\mathrm{slip}} = -\frac{\varepsilon}{4\eta}\Lambda\frac{\partial\left|\Phi - V_j\right|^2}{\partial s} \quad (3)$$

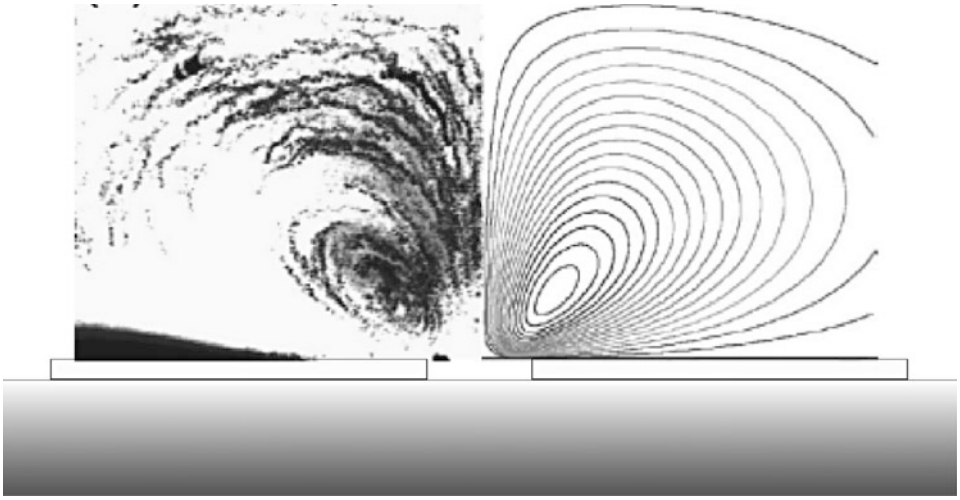
where s is the coordinate tangential to the electrode surface and Φ is the electric potential phasor evaluated at the level of the electrodes (the electric potential in the liquid is written as $\phi(t) = \mathrm{Re}[\Phi\exp(i\omega t)]$). The parameter Λ appears to account for the effect of a compact layer on the electrodes, acting as a capacitor in series with the Debye layer and, as a consequence, diminishing the voltage drop across the latter [5], $\Lambda = C_s/(C_s + C_d) \leq 1$, where C_s and C_d are the capacitances of the compact and Debye layers, respectively.

Thus, it is first required to solve the electric potential in the bulk of the liquid, which is solution of Laplace equation with specific boundary conditions, as shown in Fig. 4. Charge accumulation at the electrode-electrolyte interface yields the following boundary condition:

$$\sigma\frac{\partial\Phi}{\partial n} = i\omega C_{DL}\left(\Phi - V_j\right) \quad (4)$$



AC Electroosmosis: Basics and Lab-on-a-Chip Applications, Fig. 4 Equations and boundary conditions for computing the fluid velocity generated by ACEO. Indexes n and t indicate normal and tangential component, respectively. The electric potential is obtained from Laplace equation with specific boundary conditions at the electrodes. From the solution of the potential at the level of the electrodes, the slip velocity is computed with Eq. 3 and introduced as boundary condition in solving Stokes equation for the fluid velocity



AC Electroosmosis: Basics and Lab-on-a-Chip Applications, Fig. 5 Side view of the streamlines in ACEO with two coplanar electrodes. (*Left*) Experiments with AC signals at 100 Hz (electrode gap 50 μm, electrolyte conductivity 2.1 mS/m). Fluorescent particles are used as flow tracers. (*Right*) Numerical Simulation (Reprinted with permission from [5], © 2002 APS)

where **n** is the coordinate normal to the electrode surface and σ is the liquid conductivity and the total capacitance of the double layer is given by $C_{DL} = C_s C_d/(C_s + C_d)$. Zero normal current, $\partial\Phi/\partial\mathbf{n} = 0$, is imposed at the other boundaries of the domain. Figure 5 shows a comparison of experimental streamlines obtained using fluorescent particles as tracers and streamlines computed with this model [5].

Limiting Effects

The simple model presented in the previous section is rigorous for voltages of the order of $k_B T/e \sim 25$ mV or smaller. For these voltages, the Debye-Hückel approximation is valid, and the capacitance of the EDL per unit area is given by ε/λ_D. This does not hold true at higher voltages, and analytical models become more complicate, especially when the finite size effect of ions must be considered [6].

Another simplification is that the electrodes are considered to be perfectly polarizable, meaning that there is no charge transfer from the electrode to the electrolyte or vice versa. However, faradaic currents appear for increasing voltage, and it has been found to have an important influence in the flow [7].

Finally, there are more limitations not related to the amplitude of the applied voltage. The simple model does not consider the possible existence of an oxide layer on top of the electrode, the so-called compact layer. Very good fit to experimental results was obtained when this layer was taken into account [5]. In addition, the effect of ionic adsorption on the compact layer can reduce the slip velocity [8, 9]. In general, ACEO flow of electrolytes with high ionic strength (greater or equal to 100 mM) has not been reported. It seems that there are fundamental limitations around this ionic strength related to steric effects at high electrolyte concentrations as well as problems with the generation of faradaic reactions.

Applications for the Lab-on-a-Chip

Current clean room technologies allow for the fabrication of complex microelectrode structures and their integration in microfluidic devices. Therefore, AC electrokinetic phenomena represent an opportunity for performing standard operations in lab-on-a-chip systems. In particular, AC

electroosmosis has been demonstrated to be useful for the following applications:

ACEO Pumping and Mixing

The example of two coplanar electrodes of the same size is perhaps the easiest system to illustrate the mechanism of ACEO. As a consequence of the symmetry of the system, no net pumping of fluid is produced. The flow rolls on top of the two electrodes are identical and cancel each other. However, as predicted in [10], any electrode array with some asymmetry will lead to net pumping of the electrolyte.

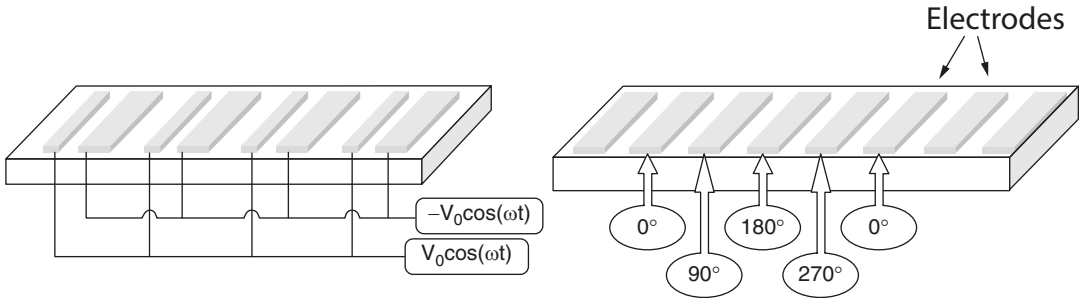
Two different electrode arrays have been broadly explored for achieving net pumping of fluid; see Fig. 6. When pairs of asymmetric electrodes are used, the flow roll on top of the larger electrode dominates and net flow is created (from left to right in the figure) [11, 12]. Ref. [13] demonstrates a microfluidic pump consisting of a long serpentine microchannel lined with 3D stepped electrode arrays, allowing to achieve high pressures with voltages around 1 V_{rms}. Because the ACEO pump employs low voltage, it has potential applications in portable and/or implantable biomedical microfluidic devices. A different strategy to break the symmetry is the application of a traveling-wave electric potential [14, 15]. In experiments, this is implemented by using an array of equal size electrodes and applying a sinusoidal potential to each electrode but with a phase lag of 90° between neighbors. Net flow is achieved in the same direction of the traveling wave (left to right in the figure). A common feature of both arrays is that the flow direction is reversed for sufficiently high voltages [16]. This observation cannot be explained by the standard ACEO model, and limiting effects like ion crowding and faradaic currents should be accounted for.

The predominant laminar character of the flow in microfluidics makes mixing a slow process. Refs. [17, 18] show how AC electroosmosis can be exploited for generating micro-vortices locally and enhancing the mixing.

Transport and Aggregation of Particles and Molecules

AC electroosmotic flows have also been used in combination with other forces (dielectrophoresis in most cases) for concentration of bio-particles. The flow is used to bring and concentrate the particles on top of the electrodes, while the other forces are used for holding them [19]. For example, ref. [20] demonstrates ACEO for concentrating DNA molecules; the generated bulk flow was able to transport DNA molecules from a large effective region to the electrode surface. In ref. [21], ACEO combined with dielectrophoresis is employed for the sorting of ZnO nanoparticles with two different morphologies (nanowires and cubic shape).

The detection times for heterogeneous immunoassays can be very long if analytes have to travel only by diffusion to the device surface, where they then bind or react with surface-bound receptors. Ref. [22] demonstrates that the fluid flow generated by ACEO can enhance the analyte



AC Electroosmosis: Basics and Lab-on-a-Chip Applications, Fig. 6 Microelectrode structures for pumping of electrolytes. (*Left*) Array of asymmetric couples of electrodes subjected to a harmonic potential. (*Right*) Array of equal size electrodes subjected to a 4-phase traveling-wave potential

transport and reduce the detection time. Ref. [23] shows how a thin metal layer on a cantilever can generate ACEO vortices and increase concentration of bio-particles before detection.

Promising results have recently been obtained by using transparent semiconductors that become conductors when light is projected on them. In this way, light patterns created at will on the substrate will function as electrodes and, for example, allow for dynamic and in-situ concentrations of particles. Ref [24] showed the light patterning of 31,000 microfluidic ACEO vortices on a featureless photoconductive surface which were able to concentrate and transport micro- and nanoscale particles.

Cross-References

- ▶ Electric Double Layer Capacitor
- ▶ Induced Charge Electroosmosis

References

1. Ramos, A., Morgan, H., Green, N.G., Castellanos, A.: AC electric-field-induced fluid flow in microelectrodes. J. Colloids Interface Sci. **217**, 420–422 (1999)
2. Bazant, M.Z., Squires, T.M.: Induced-charge electrokinetic phenomena: theory and microfluidic applications. Phys. Rev. Lett. **92**, 066101 (2004)
3. Green, N.G., Ramos, A., González, A., Morgan, H., Castellanos, A.: Fluid flow induced by nonuniform ac electric fields in electrolytes on microelectrodes. III. Experimental measurements. Phys. Rev. E **61**, 4011–4018 (2000)
4. González, A., Ramos, A., Green, N.G., Castellanos, A., Morgan, H.: Fluid flow induced by nonuniform ac electric fields in electrolytes on microelectrodes. II. A linear double-layer analysis. Phys. Rev. E **61**, 4019–4028 (2000)
5. Green, N.G., Ramos, A., González, A., Morgan, H., Castellanos, A.: Fluid flow induced by nonuniform ac electric fields in electrolytes on microelectrodes. III. Observations of streamlines and numerical simulation. Phys. Rev. E **66**, 026305 (2002)
6. Bazant, M.Z., Kilic, M.S., Storey, B.D., Ajdari, A.: Towards an understanding of induced-charge electrokinetics at large applied voltages in concentrated solutions. Adv. Colloid Interface Sci. **152**, 48–88 (2009)
7. González, A., Ramos, A., García-Sánchez, P., Castellanos, A.: Effect of the combined action of Faradaic currents and mobility differences in ac electroosmosis. Phys. Rev. E **81**, 016320 (2010)
8. Pascall, A.J., Squires, T.M.: Induced charge electroosmosis over controllably contaminated electrodes. Phys. Rev. Lett. **104**, 088301 (2010)
9. Suh, Y.K., Kang, S.: Numerical prediction of ac electro-osmotic flows around polarized electrodes. Phys. Rev. E **79**(4), 046309 (2009)
10. Ajdari, A.: Pumping liquids using asymmetric electrode arrays. Phys. Rev. E **61**, R45–R48 (2000)
11. Brown, A.B..D., Smith, C.G., Rennie, A.R.: Pumping of water with ac electric fields applied to asymmetric pairs of microelectrodes. Phys. Rev. E **63**, 016305 (2000)
12. Ramos, A., González, A., Castellanos, A., Green, N. G., Morgan, H.: Pumping of liquids with ac voltages applied to asymmetric pairs of microelectrodes. Phys. Rev. E **67**, 056302 (2003)
13. Huang, C., Bazant, M.Z., Thorsen, T.: Ultrafast high-pressure AC electro-osmotic pumps for portable biomedical microfluidics. Lab Chip **10**, 80–85 (2010)
14. Cahill, B.P., Heyderman, L.J., Gobrecht, J., Stemmer, A.: Electro-osmotic streaming on application of traveling-wave electric fields. Phys. Rev. E **70**, 036305 (2004)
15. Ramos, A., Morgan, H., Green, N.G., González, A., Castellanos, A.: Pumping of liquids with traveling-wave electroosmosis. J. Appl. Phys. **97**, 084906 (2005)
16. García-Sánchez, P., Ramos, A., Green, N.G., Morgan, H.: Experiments on AC electrokinetic pumping of liquids using arrays of microelectrodes. IEEE Trans. Dielectr. Electr. Insul. **13**, 670–677 (2006)
17. Harnett, C.K., Templeton, J., Dunphy-Guzman, K.A., Sensousy, Y.M., Kanouff, M.P.: Model based design of a microfluidic mixer driven by induced charge electroomosis. Lab Chip **8**, 565–572 (2008)
18. Sasaki, N., Kitamori, T., Kim, H.B.: AC electroosmotic micromixer for chemical processing in a microchannel. Lab Chip **6**(4), 550–554 (2006)
19. Wong, P.K., Chen, C.Y., Wang, T.H., Ho, C.M.: Electrokinetic bioprocessor for concentrating cells and molecules. Anal. Chem. **76**, 6908–6914 (2004)
20. Lei, K.F., Cheng, H., Choy, K.Y., Chow, L.: Electrokinetic DNA concentration in microsystems. Sensors Actuators A: Phys. **156**, 381–387 (2009)
21. Riahifar, R., Marzbanrad, E., Raissi, B., Zamani, C., Kazemzad, M., Aghaei, A.: Sorting ZnO particles of different shapes with low frequency AC electric fields. Mater. Lett. **65**, 632–635 (2011)
22. Hart, R., Ergezen, E., Lec, R., et al.: Improved protein detection on an AC electrokinetic quartz crystal microbalance (EKQCM). Biosens. Bioelectron. **26**, 3391–3397 (2011)
23. Islam, N., Lian, M., Wu, J.: Enhancing microcantilever capability with integrated AC electroosmotic trapping. Microfluid. Nanofluid. **3**, 369–375 (2007)
24. Chiou, P.Y., Ohta, A.T., Jamshidi, A., Hsu, H.Y., Wu, M.C.: Light-actuated AC electroosmosis for nanoparticle manipulation. Microelectromech. Syst. **17**, 525–531 (2008)

ac-Calorimetry

▶ Nanocalorimetry

Accumulator

▶ Nanomaterials for Electrical Energy Storage Devices

Acoustic Contrast Factor

Andreas Lenshof and Thomas Laurell
Department of Biomedical Engineering, Lund University, Lund, Sweden

Definition

The acoustic contrast factor (Φ) governs whether an object will be affected by the acoustic radiation force in an acoustic standing wave field [1]. It is comprised of the density and speed of the sound of the object (ρ_p and c_p) and the surrounding medium (ρ_0 and c_0), as seen in the equation below. A positive contrast factor means that the object will move to a pressure node, while a negative factor will move the object toward a pressure antinode:

$$\Phi = \frac{\rho_p + \frac{2}{3}(\rho_p - \rho_0)}{2\rho_p + \rho_0} - \frac{1}{3}\frac{\rho_0 c_0^2}{\rho_p c_p^2}$$

Cross-References

▶ Acoustic Trapping
▶ Acoustophoresis
▶ Integrated Micro-acoustic Devices

References

1. Gorkov, L.P.: On the forces acting on a small particle in an acoustical field in an ideal fluid. Sov. Phys. Dokl. **6**(9), 773–775 (1962)

Acoustic Enrichment

▶ Acoustic Trapping

Acoustic Nanoparticle Synthesis for Applications in Nanomedicine

Aisha Qi, Peggy Chan, Leslie Yeo and James Friend
Micro/Nanophysics Research Laboratory, RMIT University, Melbourne, VIC, Australia

Synonyms

Drug delivery and encapsulation; Nanocarriers; Nanomedicine; Polymer nanocapsules; Sound propagation in fluids; Ultrasonic atomization

Definition

Sound wave propagation arising from the high-frequency acoustic irradiation of a fluid can generate considerable stresses at the free surface of the fluid leading toward its destabilization and subsequent breakup. If the fluid comprises a polymer solution, the evaporation of the solvent from the aerosols that are generated as a consequence of the atomization process leaves behind a solidified polymer core with submicron dimensions. Here, any discussion of nanoparticle synthesis due to chemical reactions driven by sound waves, i.e., sonochemistry, is omitted and the discourse is limited to the physical synthesis of such nanoparticles due to the atomization of polymer solutions into a gas, typically dry air.

Overview

The synthesis of functional nanoparticles is an area of tremendous interest, particularly from the standpoint of many industry applications from catalysis, optics, and electronics to biomolecular

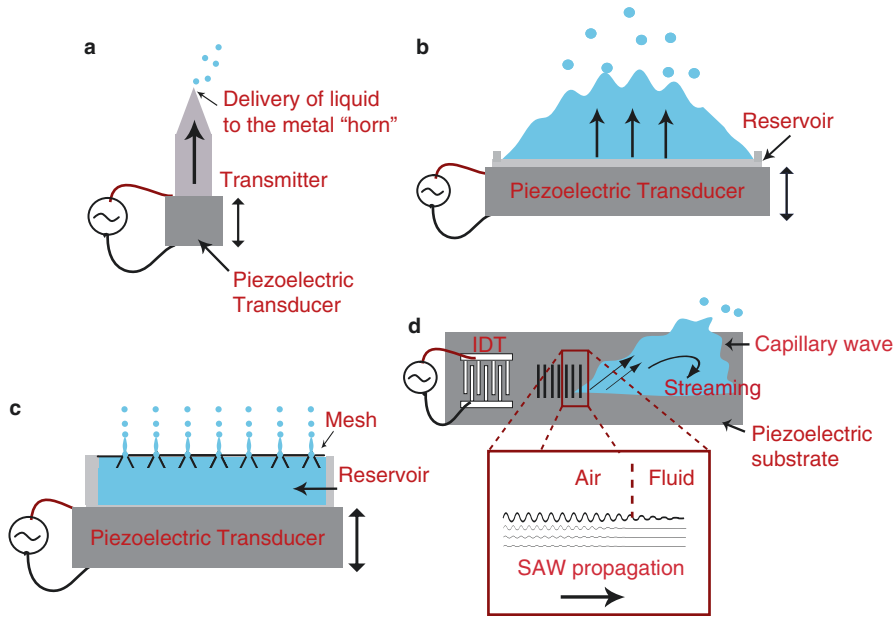
sensing, regenerative medicine, and pharmaceutical science. In the latter, there are several challenges that current drug delivery technologies and systems face. For example, the low solubility of many drugs in biological environments prevents their absorption and distribution in vivo [1, 2]. While drug solubility limitations can be circumvented through chemical structure modification or the introduction of surfactant, the former often requires expensive and complicated procedures whereas the latter may not only increase the dosage volume but is also associated with toxicity problems [1]. In addition, the delivery of most drugs cannot be localized to target a site of interest, leading to undesirable side effects or immune responses when taken up by uninfected tissues and organs. Further, drugs are often susceptible to decomposition, enzymatic degradation, aggregation, or denaturation, thus either reducing their shelf life or their efficacy in vivo.

Nanoparticle-based delivery systems, however, offer the possibility of addressing some of these issues. For example, the drug dissolution rate can be increased by reducing the particle size [1, 2]. Tumor tissues are characterized by leaky vasculature and poor lymphatic clearance; due to their subcellular dimensions, nanoparticles are able to preferentially accumulate in tumor tissue through an enhanced permeation and retention (EPR) effect [3]. As an attractive alternative to conventional invasive surgical procedures, nanoparticles have also been reported to be able to cross the blood–brain barrier to be administrated into the central nervous system [1].

The encapsulation of drugs within the polymer nanoparticle not only acts as a protective layer surrounding the drug from hostile in vivo environments but also enables the drug to diffuse out slowly over extended periods. Such controlled release as well as direct local targeting of diseased regions can also be further tuned by judicious choice of the polymeric excipient chemistry, comonomer ratios, and their degradability in certain regions. Further site-specific targeting can be achieved by functionalizing the nanoparticle surface with tissue-specific or cell-specific ligands to increase uptake into receptor expressing cells [3]. Alternatively, it is straightforward to chemically modify the nanoparticle surface to actively target a diseased site by attaching a surface-bound ligand, e.g., monoclonal antibodies or polymeric conjugates that specifically bind to target cells [1].

There are numerous routes for nanoparticle synthesis, all of which can be broadly delineated into three general approaches. Mechanical milling techniques are widely used commercially but are typically confined to industrial applications such as ceramics and paints due to contamination issues. While the level of impurities can be reduced, for example, by carrying out the process in vacuum, this is expensive and the powders are often polydispersed. Wet chemical deposition and precipitation techniques involve phase separation and include crystallization and sol–gel processing. Polymeric nanoparticles can be synthesized using emulsion (solvent extraction/evaporation) and self- or directed-assembly methods. These are usually batch operations and mass production can lead to scale up in complexity and cost. Moreover, while these methods allow for size and size distribution reproducibility, controlling these parameters is difficult. Gas phase evaporation and condensation techniques, on the other hand, encompass a range of methods that include combustion flame pyrolysis, plasma chemical vapor deposition, and laser ablation as well as spray drying. As with all aerosol processing strategies, spray drying is typically straightforward, fast, and allows high throughput, producing dried nanoparticles by atomizing a solution into micron or submicron-sized aerosols which then rapidly pass through a drying configuration. By removing the solvent either through evaporation or extraction using a hardening agent [4–6], the material of interest, e.g., proteins, peptides, and a wide range of polymers, can solidify into dried particles usually with desired dimensions and morphologies governed by the parent aerosol [7].

Other aerosol processing techniques include hydrodynamic flow focusing and electrospraying. Hydrodynamic flow focusing generates droplets by forcing liquid through an orifice using direct pressure or a fast moving air stream [8], although this typically results in an uneven and large particle size distribution. As such, it is typically used



Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 1 Different conceptual configurations for ultrasonic atomization. (**a**) Indirect bulk vibration driven by a piezoelectric transducer via a metallic horn on which the working fluid drop is deposited. (**b**) Direct bulk vibration transmitted to the fluid drop placed on the piezoelectric transducer. (**c**) Use of micron-sized nozzles and orifices to drive ultrasonic microjets. (**d**) Transmission of surface vibration energy in the form of SAWs into a fluid drop to drive interfacial destabilization

by the food industry, for example, in the manufacture of milk powder, in which stringent particle size requirements are not essential. Electrospraying, on the other hand, employs large voltages to stretch and pinch off aerosol droplets from a liquid meniscus at the tip of a metal capillary [9]. While this produces aerosol droplets and polymeric nanoparticles with a uniform size distribution, the throughput is generally low and the large electric fields required not only poses safety hazards but could also result in molecular lysis, although these could potentially be circumvented using high-frequency AC fields [10].

New technologies for nanoparticle synthesis and drug encapsulation, however, have not kept up with the rapid advances achieved in nanomedicine, particularly, from the standpoint of drug discovery and formulation. The rest of this entry is concerned with atomization processes via acoustic means for polymeric nanoparticle synthesis and recent advances in this field that allow it to be exploited as a powerful tool for drug delivery.

Ultrasonic Atomization

Ultrasonic atomization generally refers to a method for aerosol production induced by irradiating a fluid in order to destabilize its interface with acoustic energy at driving frequencies between 20 kHz to a few MHz [11], although novel technologies such as surface acoustic waves (SAWs) allow operation at higher frequencies above 10 MHz [12]. The source of the acoustic energy is usually a piezoelectric transducer driven by an alternating electrical signal. Ultrasonic atomization therefore typically encompasses both indirect and direct vibration-induced atomization [11, 13], ultrasonic microjetting [14], and, more recently, SAW atomization [15]. Figure 1 illustrates these various conceptual strategies that are generally classified as ultrasonic atomization methods, the difference between these arising through the mechanism by which the acoustic energy is introduced or the way in which the aerosol droplets are produced. For example, *bulk* vibration energy is transferred

indirectly from the piezoelectric transducer to the fluid drop to be atomized through a metallic horn (Fig. 1a) or directly to the sessile fluid drop placed on the piezoelectric substrate (Fig. 1b). Nozzles and orifice plates from which fluid jets emanate that subsequently break up into aerosol droplets can also be used in an extension of the indirect method to provide finer control over the final droplet size (Fig. 1c). Alternatively, a nozzle can assume the place of the horn in the direct method in Fig. 1a, with the fluid issuing through the needle instead of placing the fluid on the horn. Nozzles or orifices however tend to clog over time and require constant cleaning; the pressure drop is also significantly increased, and hence more power is typically required to drive the atomization process. Alternatively, the *surface* vibration in the form of SAWs can couple acoustic energy into a fluid drop sitting on the piezoelectric substrate to effect interfacial destabilization (Fig. 1d). The operating principle and governing mechanisms that underpin ultrasonic atomization (cavitation or capillary wave destabilization) are discussed in detail in Yeo et al. [12].

The average diameter D of the aerosol droplets produced depends on the most unstable wavelength λ of the capillary wave instability that is induced, specified by the Kelvin equation, which essentially arises from a dominant force balance between the stabilizing capillary stresses and the destabilizing stress imposed by the vibration:

$$D \approx C_1\lambda \approx C_2\left(\frac{2\pi\gamma}{\rho f_c^2}\right)^{\frac{1}{3}}, \tag{1}$$

where γ and ρ are the surface tension and density of the fluid, respectively, and f_c is the capillary wave frequency. C_1 and C_2 are empirically determined coefficients commonly used to fit the experimental data to the predictions, and vary widely in the literature, although they are typically of order unity. It should be noted that Kelvin's theory does not provide a way in which the capillary wave frequency is related to the forcing frequency due to the acoustic vibration, and the usual assumption made (see, for example, [15]) is that the capillary waves are excited at a subharmonic frequency that is one half of the forcing frequency, i.e., $f_c = f/2$. Recent studies have however shown that this may not always be true and that the capillary wave frequency can be predicted from a dominant balance between capillary and viscous stresses, at least for a sessile drop or thick liquid film [12, 15]:

$$f_c \sim \frac{\gamma}{\mu L}, \tag{2}$$

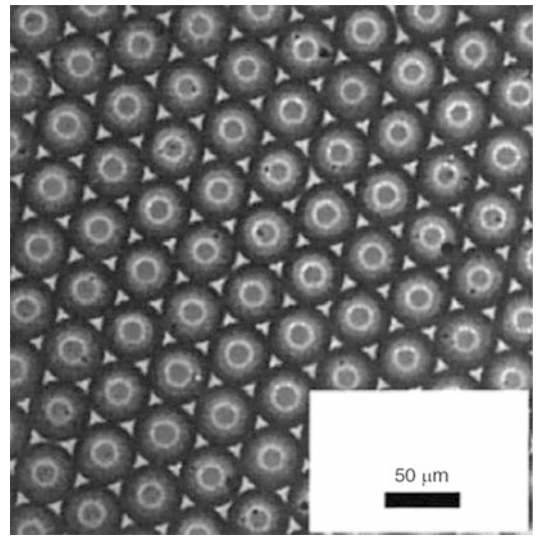
where μ is the viscosity of the fluid and L is the characteristic length scale of the drop or film. A correction factor $(H/L)^2$, wherein H is the characteristic height scale, can be included to account for the geometry of the parent drop, in particular, the axial capillary stress, when substituting Eq. 2 into Eq. 1 [15].

SAWs, generated by applying an oscillating electrical signal to the interdigital transducer electrodes patterned onto a piezoelectric substrate (Fig. 1d), the gap and width of which specifies the SAW wavelength and hence the resonant frequency (typically 10–100 MHz), have also been shown as an effective mechanism for atomizing drops. Essentially, SAWs are nanometer order amplitude ultrasonic waves that are confined to and propagate along the substrate surface in the form of a Rayleigh wave. Owing to high MHz order frequencies and substrate displacement velocities, on the order of 1 m/s in the vertical direction perpendicular to the substrate, substrate accelerations on the order of 10^7 m^2/s arise. This, together with the strong acoustic streaming that is induced within the fluid drop placed on the substrate, leads to fast destabilization and breakup of the drop free surface to produce 1–10 μm dimension aerosol drops [15], as illustrated in Fig. 1d. One advantage of the SAW atomization over its ultrasonic counterparts is the efficient energy transfer mechanism from the substrate to the liquid – unlike bulk vibration, most of the acoustic energy of the SAW is localized within a region on the surface of the substrate about 3–4 wavelengths thick (a SAW wavelength is typically on the order of 100 μm) and is transferred to the liquid drop to drive the atomization. Consequently, it is possible to atomize fluids using the SAW at input powers of

around 1 W, which is one to two orders of magnitude smaller than that required using other ultrasonic methods. Another advantage of the SAW is the high frequencies (>10 MHz) that can be assessed. The timescale associated with the period (inverse frequency) of the oscillating acoustic and electromechanical field at these frequencies is much shorter than the hydrodynamic timescale $\mu L/\gamma \sim 10^{-4}$ s as well as the 10^{-5}–10^{-6} s relaxation timescales (inverse of the strain rate) associated with shear-induced molecular lysis [12]. In addition, it is not possible to induce cavitation, which is known to cause molecular lysis, at the low powers and high frequencies associated with SAW atomization [15].

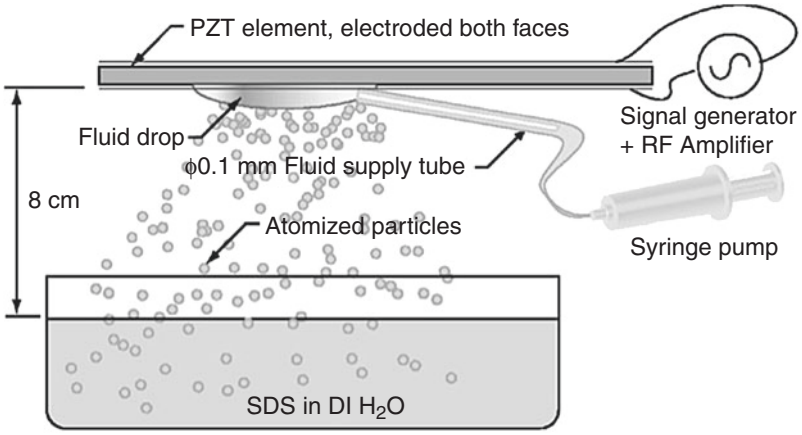
Synthesis of Polymeric Microparticles and Nanoparticles

Early work to demonstrate the possibility of synthesizing polymeric particles, albeit with micron dimensions, was reported by Tsai et al. [16] and Berkland et al. [4]. The apparatus in both studies is similar, based on concept known as two-fluid atomization. A piezoelectric transducer is used to vibrate a nozzle or orifice from which a jet comprising the working fluid issues and which subsequently suffers from Rayleigh-Plateau instabilities to break up into individual aerosol droplets. In a manner similar to the Kelvin equation given in Eq. 1, the droplet diameter is controlled by the most unstable wavelength of the axisymmetric instability, which can be predicted from a dominant force balance between the inertia imposed on the jet and the capillary stresses that stabilize it, and is a function of the diameter of the undisturbed jet, which is slightly larger than the nozzle or orifice diameter. In both cases, an external annular stream surrounding the nozzle through which the liquid jet issues is employed. The annular sheath fluid consists of air in the former and an immiscible carrier liquid that does not dissolve the polymer solution in the latter. In both cases, the sheath fluid, whether air or a second immiscible liquid, is flowed much faster than the jetting fluid although the underlying reason provided for its necessity differs. In the former, the airflow is suggested to vibrate in resonance with the vibrating nozzle due to the coaxial annular arrangement, resulting in a magnification of the amplitude of the capillary waves on the liquid jet. In the latter, the interfacial shear imposed by the carrier fluid surrounding the jet is suggested to aid the primary breakup of the jet away from its parent fluid at the orifice, allowing the production of droplets one order of magnitude smaller. 40 μm diameter xanthan gum particles were synthesized in the former when driven at a fundamental resonant frequency of 54 kHz through a 0.93 mm nozzle whereas 5–500 μm diameter poly(d,l-lactic-*co*-glycolic acid) (PLGA) particles (Fig. 2) were synthesized in the latter when driven at frequencies between 19 and 70 kHz through 60 and 100 μm diameter orifices. The carrier fluid, poly (vinyl alcohol), in which the PLGA particles were collected and solidified within, however, remained on the surface of the particles, which could potentially affect physical and cellular uptake [3].

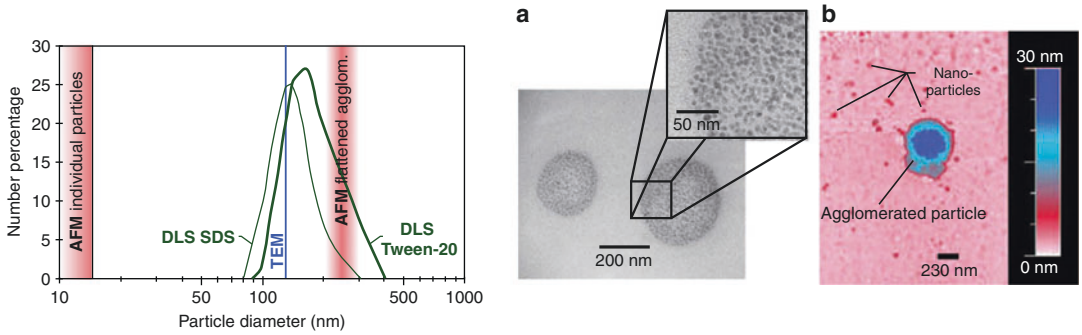


Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 2 Forty five micro meter PLGA microparticles synthesized by ultrasonically atomizing the polymer solution through a 60 μm orifice which is indirectly vibrated using a piezoelectric transducer at frequencies between 19 and 70 kHz (Reprinted with permission from Berkland et al. [4]. Copyright (2001) Elsevier)

Forde et al. [17] subsequently showed the possibility of synthesizing poly(ε-caprolactone)



Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 3 Schematic illustration of the direct ultrasonic atomization setup of Forde et al. in which a sessile drop is vibrated at 1.645 MHz or 5.345 MHz in a piston-like manner using a hard PZT disk (Reprinted with permission from Forde et al. [17]. Copyright (2006) American Institute of Physics)



Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 4 The *left panel* shows dynamic light scattering measurements of the PCL nanoparticles produced using SAW atomization. The *right panel* shows (**a**) transmission electron microscopy and (**b**) atomic force microscopy images showing 150–200 nm clusters of 5–10 nm nanoparticles (Reprinted with permission from Friend et al. [19]. Copyright (2008) Institute of Physics)

(PCL) *nano*particles around 200 nm in dimension by atomizing a PCL/acetone solution using the direct method shown in Fig. 1b involving a piston-vibrated hard lead zirconate titanate piezoelectric disk between 1 and 5 MHz, as illustrated in Fig. 3. The nanoparticle diameter D_p can be estimated from volume considerations [18]:

$$D_p \approx D\left(\frac{C}{\rho}\right)^{\frac{1}{3}}, \tag{3}$$

wherein C denotes the initial polymer concentration, such that it is possible to tailor the nanoparticle size by tuning the aerosol size through its physical properties as well as the geometry of the parent drop, as suggested by Eq. 2.

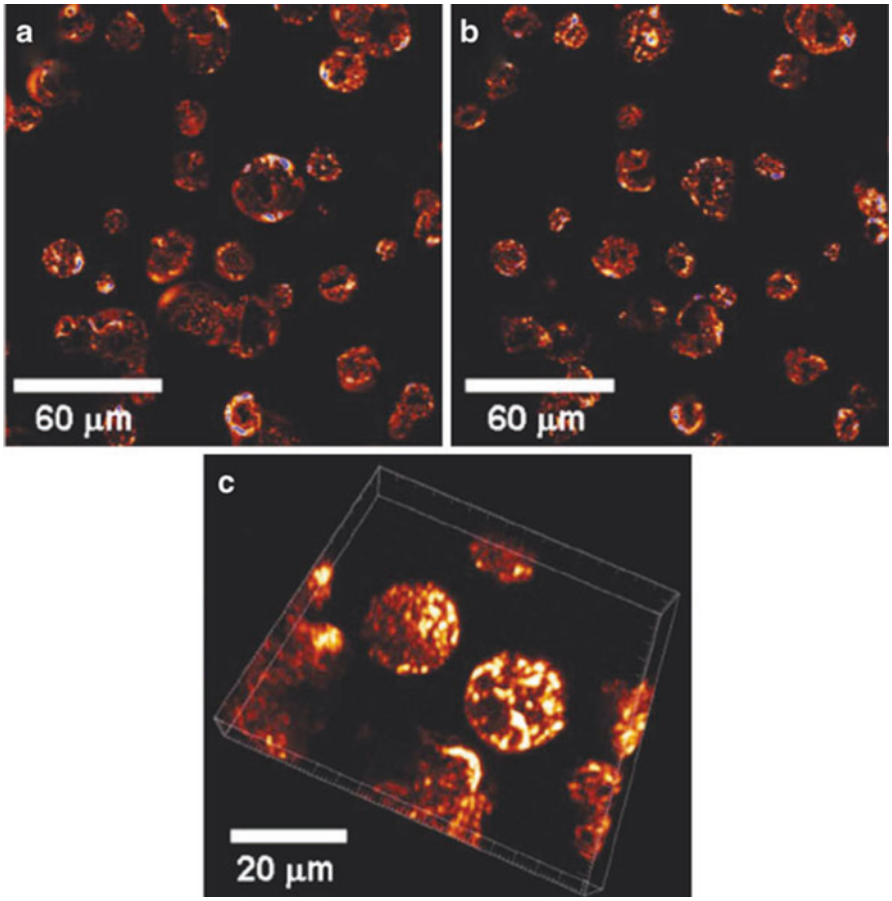
Polymeric nanoparticles have also been produced using SAW atomization. Friend et al. [19] demonstrated the synthesis of 150–200 nm clusters of 5–10 nm PCL nanoparticle aggregates, as shown in Fig. 4, the cluster size being weakly dependent on several parameters such as the surfactant used and the drying length. The grape-bunch-like cluster morphology was attributed to nonuniform solvent evaporation during in-flight

drying of the droplets that drive a thermodynamic instability. This results in spinodal decomposition and phase separation, which gives rise to separate regions that are polymer-rich and solvent-rich. The polymer-rich regions then solidify more rapidly, creating nucleation sites that lead to the formation of a cluster of PCL molecules until a critical nucleation size is attained, estimated from classical nucleation theory to be around 10 nm and consistent with the 5–10 nm nanoparticles observed. In addition to PCL, 1–10 μm dimension protein (bovine serum albumin and insulin) aerosols have been generated using the SAW, which evaporate in flight to form 50–100 nm nanoparticles [18].

Drug Encapsulation

Early work on drug encapsulation via ultrasonic atomization was carried out by Felder et al. [5], who showed the possibility of encapsulating protein and peptides into poly(lactic acid) (PLA) and PLGA directly using ultrasonic atomization. The polymer solution was sonicated with a protein (bovine serum albumin; BSA) or peptide solution (2–10 % w/w concentration) to create a stable water-in-oil emulsion, which was then atomized into a beaker in which hardening agent (octamethylcyclotetrasil, hexane, isopropyl myristate, or water) was agitated to yield 10–100 μm order microparticles. This solidification method relies on *solvent extraction* and hence polymer desolvation within the hardening agent as opposed to *solvent evaporation* in-flight used in the other methods reported here. While smaller nanoparticles on the order of 100 nm diameter were also reported, it is not clear whether these contained any encapsulated material, as the authors only reported size distribution data for the naked PLA and PLGA particles. It would, however, be consistent with the other studies reported below that encapsulated particles are generally of micron-order dimension due to the size of the peptides and proteins within. The frequency at which the ultrasonic atomization was carried out was also unspecified. Reported encapsulation efficiencies, obtained by dissolving the polymer particles in the original solvent (or an acetonitrile/chloroform mixture) to release the encapsulated material which is then recovered on a 200 nm cellulose filter followed by elution with phosphate buffered saline and quantitative characterization using a spectrofluorometer (protein) or through a HPLC assay (peptides), were typically low, between 10 % and 35 %. This was attributed to comparable extraction rates between that of the aqueous phase in which the protein or peptide resides with that of the polymer solvent through diffusion and partitioning. Consequently, the protein or peptide is removed during washing. When a different peptide which is only weakly soluble in water was encapsulated, the efficiency was observed to increase to 63 % and 93 %. As such, it is instructive to note here that encapsulation efficiencies are therefore likely to be higher if solvent evaporation is used as the polymer desolvation method instead of solvent extraction, since the solvent is much more likely to be removed rapidly through evaporation while in flight at a rate much faster than that of water, therefore resulting in more efficient encapsulation of the aqueous phase and the proteins or peptides within the fast solidifying polymer shell.

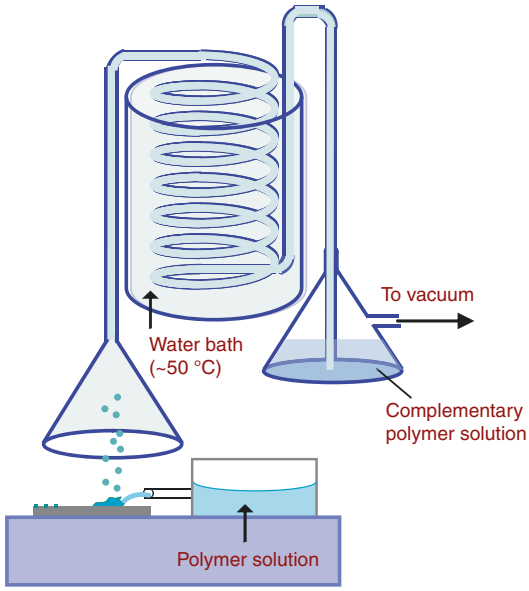
BSA was also encapsulated in PLGA using 100 kHz ultrasonic atomization of a similar water-in-oil emulsion in a study by Freitas et al. [6]. The solvent was evaporated and the solidified particles were collected and further desolvated in an agitated aqueous solution and subsequently recovered using a filter. To allow for aseptic encapsulation conditions, the conventional spray dryer unit was shortened by running the unit under near-vacuum conditions instead of using hot air, and the cyclone unit that is usually present in spray dryers was replaced by an aqueous collection bath; in that way, the entire unit can be placed within a laminar flow chamber. Again, the particles were in the micron dimension (13–24 μm mean diameter). As expected due to the use of solvent evaporation over solvent extraction, the BSA encapsulation efficiencies were higher, around 50–60 % compared to 10–35 %, which the authors claim could be improved if BSA loss in the aqueous collection solution was replaced by a fluid which is a nonsolvent for



Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 5 Confocal microscope images of the cross-sectional slice (*top, bottom, and center*) across a PCL microparticle showing the encapsulation of fluorescent biotin within (Reprinted with permission from Alvarez et al. [18]. Copyright (2009) American Institute of Physics)

BSA. Alternatively, they suggest reducing the polymer concentration.

Plasmid DNA (pDNA) and poly(ethyleneimine) (PEI) complexes were encapsulated within 10–20 μm diameter PLGA microparticles using a 40 kHz ultrasonic atomizer [20]. Solvent extraction via a poly(vinyl alcohol) hardening agent was used in this case. Encapsulation efficiencies in a wide range between 20 % and 90 % were reported, with lower polymer concentrations and higher pDNA-PEI volume fractions giving higher efficiencies; contrary to prior claims, the authors however did not find significant dependence of the encapsulation efficiency on the N/P ratio (N referring to the nitrogen content and P referring to the DNA phosphate content of the pDNA-PEI complex). These values were however obtained by subtracting the mass of pDNA remaining in the hardening agent, measured using a spectrofluorometer, from that in the feedstock, which does not take into account loss of pDNA in the environment during atomization, the amount of unatomized pDNA left on the atomizer as well as the pDNA that resides on the surface of the particles instead of being encapsulated within. The authors also investigated the post-atomization structural integrity of the pDNA using gel electrophoresis due to the susceptibility of pDNA to shear degradation, and found 80 % retention of the pDNA in supercoiled

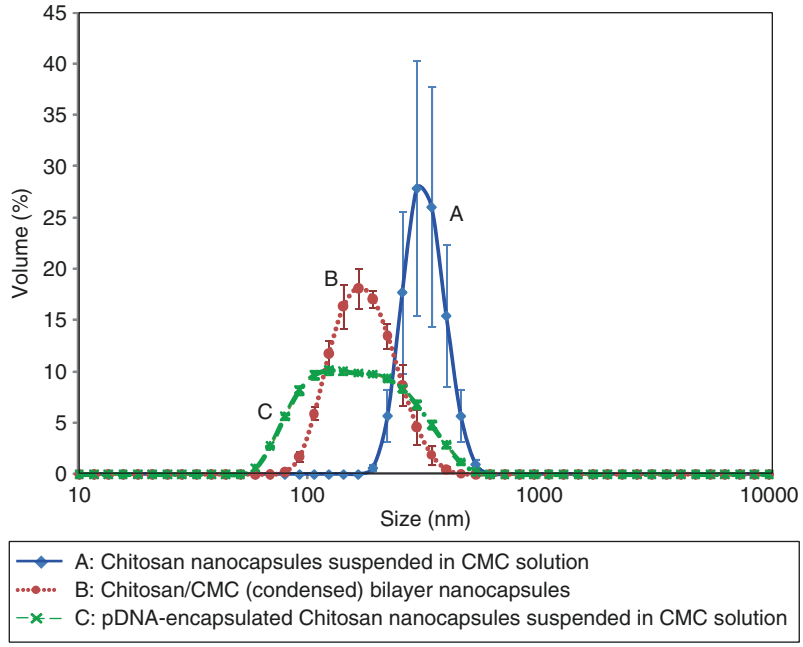


Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 6 Schematic illustration of the SAW atomization setup involving a single atomization–evaporation–resuspension step used to deposit a single polymer layer and to resuspend it in a complementary polymer solution. The step is repeated to deposit subsequent layers for as many layers as required

conformation. This was attributed to cationic complexation between pDNA and PEI, which reduces its size and hence the possibility of shear-induced degradation (when naked pDNA was assessed, only 8 % of this remained supercoiled). Nevertheless, PEI has been known to induce cytotoxic effects and hence its use may be limited for gene transfection and delivery. An alternative mitigation strategy involving multilayer polymer nanoparticles will be discussed in the next section.

Alvarez et al. [18] later demonstrated the encapsulation of BSA in PCL particles using SAW atomization. With 10 MHz SAWs, the particle sizes had a mean of 23 μm whereas this decreased to around 6 μm with 20 MHz SAWs. Proof of encapsulation was acquired through confocal image slices of the particles showing the fluorescently tagged BSA appearing not only at the top and bottom of the cross-sectional slices across the particle but also in the particle center, thus verifying that the BSA was entrapped within and not simply bound to the surface of the particle (Fig. 5). The encapsulation efficiency value reported was slightly different compared to other

Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 7 Dynamic light scattering measurements of (**a**) single-layer chitosan nanoparticles, (**b**) bilayer chitosan and CMC nanoparticles, and (**c**) pDNA encapsulated chitosan nanoparticles

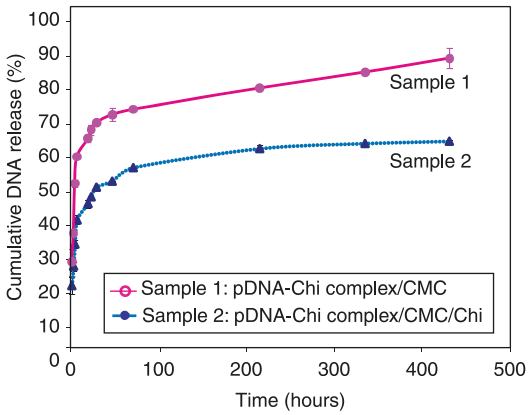


studies in that 54 % of BSA from the feedstock was found to be encapsulated with its chemical structure intact (as opposed to the total BSA content encapsulated reported in previous studies). Thus, the total encapsulation efficiency value could be significantly higher (if a comparison were to be made with the values in Felder et al. [5] and Freitas et al. [6]). As discussed above, the high-frequency operation of the SAW limits the amount of shear and cavitation damage caused to the molecules, and constitutes a considerable advantage of using SAWs over conventional ultrasonic atomization.

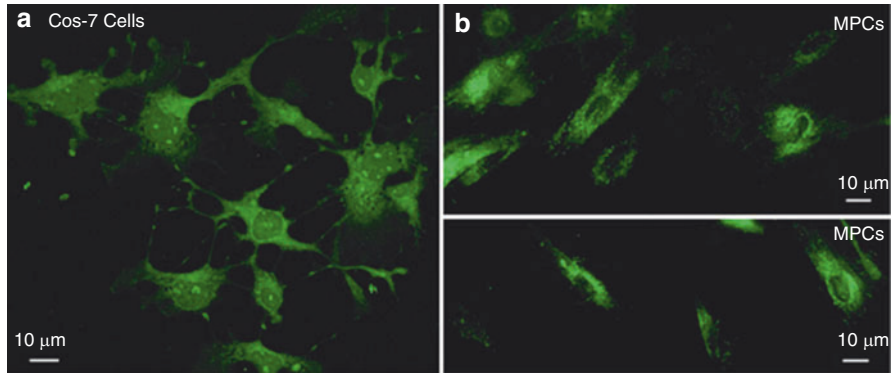
Multilayer Nanoparticle Synthesis and Encapsulation

Multilayered polymer particles offer significant advantages over a particle comprising a single layer. Layers of different polymers with varying chemistry and hence degradation profiles offer the possibility of tuning the drug release over time and in different physiological regions, therefore affording tremendous opportunities for targeted and controlled release delivery. Very recently, SAW atomization has been exploited to demonstrate the potential for rapidly synthesizing nanoparticles with alternating layers of complementary polymers of opposing charge [21]. This is done through a variation of the usual procedure of atomizing an initial polymer solution followed by evaporated-assisted solidification to produce a single-layer polymer nanoparticle. In addition, however, the solidified polymer particle is collected in a solution in which the second polymer is dissolved (Fig. 6), which is then re-atomized and dried to deposit the second polymer layer over the initial polymer core. By collecting the two-polymer-layer nanoparticle in a solution comprising the polymer to form the third layer and re-atomizing, a further layer can be deposited. The atomization–evaporation–resuspension procedure is then repeated for as many times as the number of layers desired. The requirement of the polymers comprising alternating layers is that each successive polymer must be complementary to the previous polymer, i.e., they must have opposing charges. In addition, the polymer making up the subsequent layer must be soluble in a solvent that cannot dissolve the polymer making up the previous layer. Up to eight layers of alternating chitosan (or PEI) and CMC layers were synthesized; proof of the deposition of successive layers was provided by visual inspection (atomic force microscopy), charge characterization (reversal of the zeta-potential after the deposition of each successive layer), Fourier transform infrared spectrometry showing ionic complexation between the deposited layers, and fluorescence measurements of labeled polymers. In addition, pDNA was also encapsulated within the multilayer polymeric nanoparticles to demonstrate the therapeutic capability of the nanoparticles in particular for gene therapy.

Figure 7 shows the size distribution of negatively charged chitosan and positively charged CMC polymer bilayer particles. The decrease in size upon deposition of the second CMC layer over the chitosan core or the encapsulated pDNA can be attributed to ionic complexation that tends to compact the particle size. Even with encapsulation, however, the size of the particles remains in the 100–200 nm range, which is a significant advance over the micron-sized particles obtained whenever a therapeutic molecule is encapsulated



Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 8 In vitro release profile showing the diffusion of pDNA out of chitosan/CMC bilayer nanoparticles (*sample 1*) and chitosan/CMC/chitosan trilayer nanoparticles (*sample 2*) acquired through fluorescence absorbance measurements of fluorescently labeled pDNA. The lines are added to aid visualization



Acoustic Nanoparticle Synthesis for Applications in Nanomedicine, Fig. 9 Confocal microscopy images of (**a**) COS-7 cells and (**b**) human mesenchymal progenitor cells (MPCs) transfected with pDNA encoded with a yellow fluorescent protein (the expression is depicted in the images in *green*) encapsulated in PEI/CMC bilayer nanoparticles

(see previous section). In targeted cancer therapy, for example, the mean vascular pore size of most human tumors is around 400 nm [1], and hence extravasation is likely to be more effective with the multilayer nanoparticle drug carriers synthesized through this technique. Figure 8 shows in vitro release profiles of the pDNA, showing the possibility for slowing and hence controlling the release with the deposition of an additional layer. Good in vitro DNA transfection is also demonstrated in COS-7 and human mesenchymal progenitor cells, as seen in Fig. 9.

Cross-References

- ▶ Nanoencapsulation
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Polymer Coatings

References

1. Kumar, M.N.V.R. (ed.): Handbook of Particulate Drug Delivery. American Scientific, California (2008)
2. Mehnert, W., Mäder, K.M.: Solid lipid nanoparticles: production, characterization and applications. Adv. Drug Deliv. Rev. **47**, 165–196 (2001)
3. Panyam, J., Labhasetwar, V.: Biodegradable nanoparticles for drug and gene delivery to cells and tissue. Adv. Drug Deliv. Rev. **55**, 329–347 (2003)
4. Berkland, C., Kim, K., Pack, D.W.: Fabrication of PLG microspheres with precisely controlled and monodisperse size distributions. J. Control. Release **73**, 59–74 (2001)
5. Felder, C., Blanco-Prieto, M., Heizmann, J., Merkle, H., Gander, B.: Ultrasonic atomization and subsequent polymer desolvation for peptide and protein microencapsulation into biodegradable polyesters. J. Microencapsul. **20**, 553–567 (2003)
6. Freitas, S., Merkle, H., Gander, G.: Ultrasonic atomisation into reduced pressure atmosphere – envisaging aseptic spray-drying for microencapsulation. J. Control. Release **95**, 185–195 (2004)
7. Alvarez, M., Friend, J., Yeo, L.Y.: Rapid generation of protein aerosols and nanoparticles via surface acoustic wave atomization. Nanotechnology **19**, 455103 (2008)
8. Gañán-Calvo, A.M.: Enhanced liquid atomization: from flow-focusing to flow-blurring. Appl. Phys. Lett. **86**, 214101 (2005)
9. Grace, J., Marijnissen, J.: A review of liquid atomization by electrical means. J. Aerosol. Sci. **25**, 1005–1019 (1994)
10. Yeo, L.Y., Gagnon, Z., Chang, H.-C.: AC electrospray biomaterials synthesis. Biomaterials **26**, 6122–6128 (2005)
11. Friend, J., Yeo, L.Y.: Microscale acoustofluidics: microfluidics driven via acoustics and ultrasonics. Rev. Mod. Phys. **83**, 647–704 (2011)
12. Yeo, L.Y., Friend, J.R., McIntosh, M.P., Meeusen, E. N.T., Morton, D.A.V.: Ultrasonic nebulization platforms for pulmonary drug delivery. Expert Opin. Drug Deliv. **7**, 663–679 (2010)
13. James, A.J., Vukasinovic, B., Smith, M.K., Glezer, A.: Vibration-induced drop atomization and bursting. J. Fluid Mech. **476**, 1–28 (2003)
14. Meacham, J.M., Varady, M.J., Degertekin, F.L., Fedorov, A.G.: Droplet formation and ejection from a micromachined ultrasonic droplet generator: visualization and scaling. Phys. Fluids **17**, 100605 (2005)

15. Qi, A., Yeo, L., Friend, J.: Interfacial destabilization and atomization driven by surface acoustic waves. Phys. Fluids **20**, 074103 (2008)
16. Tsai, S.C., Luu, P., Song, Y.L., Tsai, C.S., Lin, H.M.: Ultrasound-enhanced atomization of polymer solutions and applications to nanoparticles synthesis. Trans. Ultrason. Symp. **1**, 687–690 (2000)
17. Forde, G., Friend, J., Williamson, T.: Straightforward biodegradable nanoparticle generation through megahertz-order ultrasonic atomization. Appl. Phys. Lett. **89**, 064105 (2006)
18. Alvarez, M., Yeo, L.Y., Friend, J.R., Jamriska, M.: Rapid production of protein-loaded biodegradable microparticles using surface acoustic waves. Biomicrofluidics **3**, 014102 (2009)
19. Friend, J.R., Yeo, L.Y., Arifin, D.R., Mechler, A.: Evaporative self-assembly assisted synthesis of polymeric nanoparticles by surface acoustic wave atomization. Nanotechnology **19**, 145301 (2008)
20. Ho, J., Wang, H., Forde, G.: Process considerations related to the microencapsulation of plasmid DNA via ultrasonic atomization. Biotechnol. Bioeng. **101**, 172–181 (2008)
21. Qi, A., Chan, P., Ho, J., Rajapaksa, A., Friend, J., Yeo, L.: Template-free synthesis and encapsulation technique for layer-by-layer polymer nanocarrier fabrication. ACS Nano (2011). doi:10.1021/nn202833n

Acoustic Particle Agglomeration

► Acoustic Trapping

Acoustic Trapping

Mikael Evander[1] and Thomas Laurell[2]
[1]Department of Biomedical Engineering, Faculty of Engineering, Lund University, Lund, Sweden
[2]Department of Biomedical Engineering, Lund University, Lund, Sweden

Synonyms

Acoustic enrichment; Acoustic particle agglomeration; Acoustic tweezers

Definition

Acoustic trapping is the immobilization of particles and cells against a flow in the node of an ultrasonic standing wave field.

Overview

Acoustic trapping has been shown to be a gentle way of performing noncontact immobilization of cells and particles in microfluidic systems. Localized ultrasonic standing waves are created in microfluidic channels, cavities, or other small, confined spaces creating pressure nodes that attract and hold particles and cells. Commonly, structures in the range of a couple of 100 μm are used, corresponding to acoustic frequencies in the MHz range.

Background

Some 30 years ago, NASA and ESA started to develop containerless processing, and one of several possible techniques was acoustic trapping/levitation [1]. These open-air levitators, still in use today, are rather large instruments that create a standing wave in air and make it possible to position and levitate liquid droplets or small and light solid matter. The use of radiation forces on particles and cells in liquid suspensions has been a more common approach than the open-air levitators however. Baker showed in 1972 that a band formation owing to the radiation forces could be observed when subjecting erythrocytes to a 1 MHz standing wave in polystyrene containers [2]. Since then, the technique has been refined and successfully incorporated into the lab-on-a-chip field as one of several methods to control cells and particles in laminar flows.

Theory

In order to create an efficient standing wave, the geometry of the channel/cavity must match the actuation frequency of the transducer. To create a single pressure node, the length of the resonance cavity should be $\lambda/2$, where λ is the wavelength of the ultrasound in the media. To increase the trapping capacity, this can then be scaled so that the number of pressure nodes where objects can be trapped is equal to $n*\lambda/2$. For water-based buffers, the sound velocity is typically 1,500 m/s. So in

order to create a standing wave at 2 MHz, the resonance cavity should be 375 μm (v/2f).

The most dominant force in play is the *primary radiation force* (Eq. 1) that is proportional to the particle volume (V_c) and is thus strongly size dependent [3]. The force also depends on the acoustic wavelength (λ), the acoustic pressure amplitude (p_0), and the particles' position in the acoustic wave (x). The *acoustic contrast factor* (Eq. 2) is dependent on the density and compressibility of both the particle (ρ_c, β_c) and the medium (ρ_w, β_w):

$$F_r = -\left(\frac{\pi p_0^2 V_c \beta_w}{2\lambda}\right) \cdot \phi(\beta, \rho) \cdot \sin(2kx) \quad (1)$$

$$\phi(\beta, \rho) = \frac{5\rho_c - 2\rho_w}{2\rho_c + \rho_w} - \frac{\beta_c}{\beta_w} \quad (2)$$

The acoustic contrast factor can change the direction of the force so that, depending on the material parameters, an object may be pushed toward the pressure node or the pressure antinode. For most cells and particles, the force will be directed toward the pressure node however. *Acoustophoresis* uses this factor to enable separation between different cell types that may differ in density and/or compressibility.

While the *primary radiation force* is responsible for moving an object into the pressure node or antinode, it is the lateral component of it that enables acoustic trapping. The acoustic radiation force is based on both the pressure and the velocity gradient in the standing wave, and in order to create an efficient trapping system, a large gradient is required. The force resulting from the lateral velocity gradient for a dense particle with a radius, R, situated in the pressure node of a standing wave with constant amplitude gradient was expressed by Gröschl as [4]

$$F_{\text{Lat}} = \pi\rho\omega^2 R^2 \hat{u}_0 \hat{u}_m \quad (3)$$

where ρ is the medium density and the angular frequency of the ultrasound, u_0, is the displacement amplitude at the center of the particle, while u_m is the difference in amplitude between the center and the edge of the particle. As can be seen, the larger the amplitude difference between the center and the edge of the particle, the larger the trapping force.

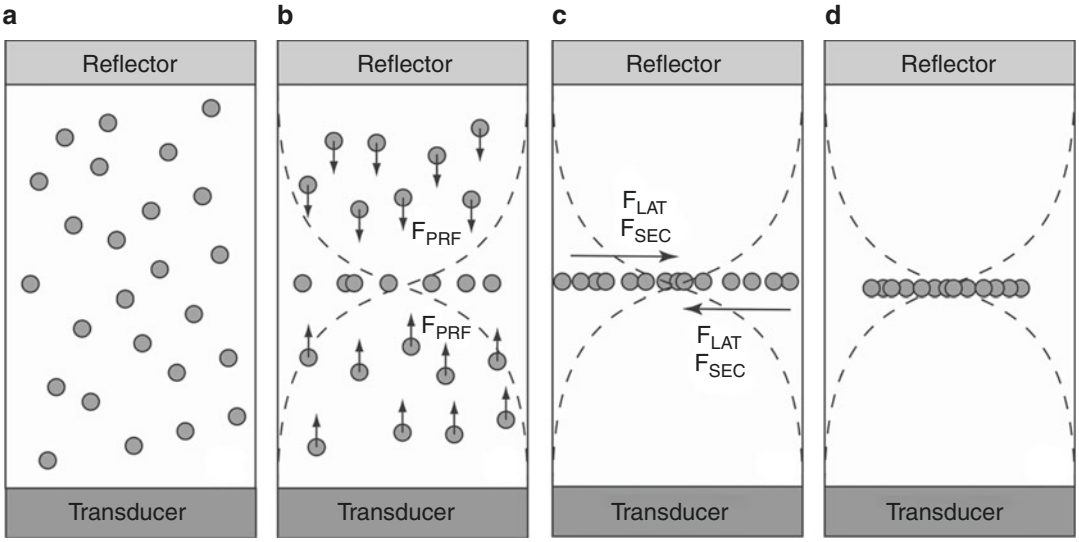
There are also *secondary interparticular forces* that arise when the incident sound wave is scattered on the objects in the standing wave. These forces only act on short distances but help to form a cluster of the objects and keep them together (Fig. 1).

System Design

An acoustic trap is usually designed to either use a localized pressure field or create a localized resonance. An example of a localized pressure field would be a small transducer that couples sound to a fluid only in a small portion of a channel, also called a layered resonator. The alternative approach is to actuate the entire chip and design a resonance cavity on the chip that matches the actuation frequency. The standing wave will then only exist where the geometry matches the standing wave criteria. Another alternative that has not been as common as the other two designs is the use of focused transducers. By focusing the ultrasound into a very narrow beam and then reflecting the wave in the focal plane, a hemispheric standing wave that can trap and hold objects can be created [6] (Fig. 2).

Designs Using Local Resonance Cavities

Designs that use a transducer coupled directly or through a matching layer to the fluid typically have a larger trapping capacity than the cavity approach. However, since the cavity trap is usually created using photolithographic techniques, it is easy to array and to integrate with other sample processing steps on the chip. An example of a cavity array was presented by Vanherberghen et al. where a 10 × 10 array was used for aggregating cells for isolated cell studies using microscopy [7] (Fig. 3).



Acoustic Trapping, Fig. 1 A schematic image demonstrating how different acoustic forces act on particles in a standing wave. (**a**) Before activating the ultrasound, particles will be evenly distributed in the channel. (**b**) When activating the ultrasound, the particles will be moved into the pressure node by the primary radiation force. (**c**) Lateral forces will then focus the particles to the center of the sound field. Once the distance between the particles is short enough, secondary interparticular forces will create an attractive force between the particles and pull the cluster firmer together. (**d**) The end result is a centered, levitated cluster of particles situated in the middle of the resonance cavity (Reprinted from Evander [5] with permission from the author)

Designs Using a Localized Transducer

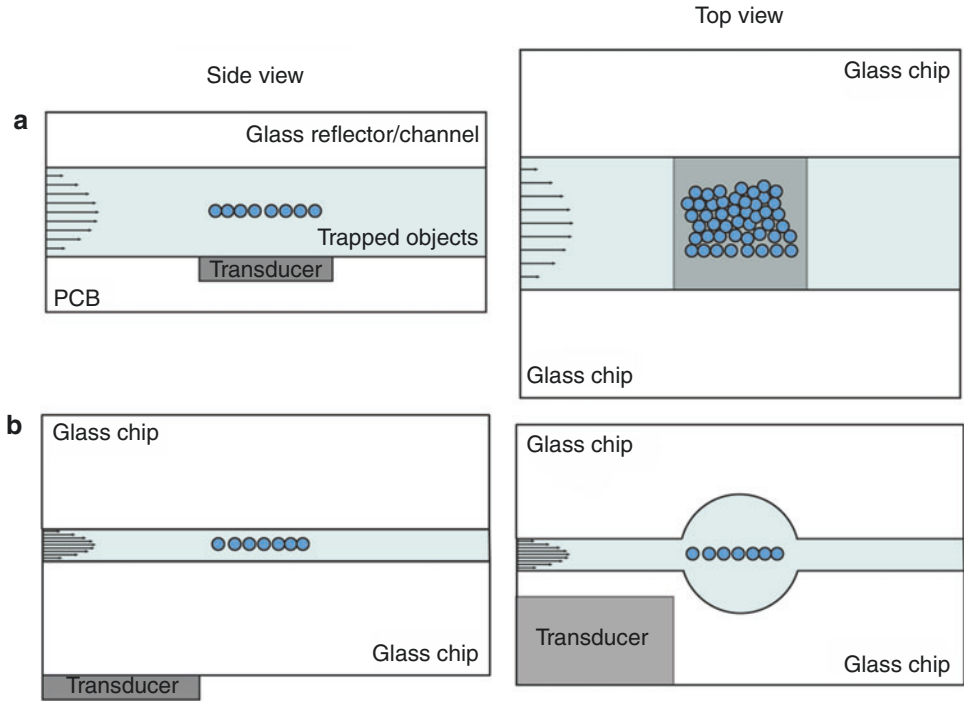
Spengler and Coakley presented a trapping system that used a layered design in 2000 [8]. By using a PTFE spacer and a glass reflector, they were able to study cell aggregates in a standing wave at 1.93 MHz. This system was later redesigned to a stainless steel system using a matching layer to couple the ultrasound into the resonator and used for studies of cell membrane spreading [9].

An alternative to using a matching layer is to integrate the transducer in the bottom of a microfluidic channel. A system based around a miniature transducer that was embedded in a printed circuit board was presented by Evander et al. in 2007 [10]. A glass lid with a microfluidic channel was placed on top of the printed circuit board, and a standing wave was formed above the transducer in the channel. To test the system, yeast cells were grown in the standing wave, and an online viability assay was performed on trapped neural stem cells (see Fig. 4).

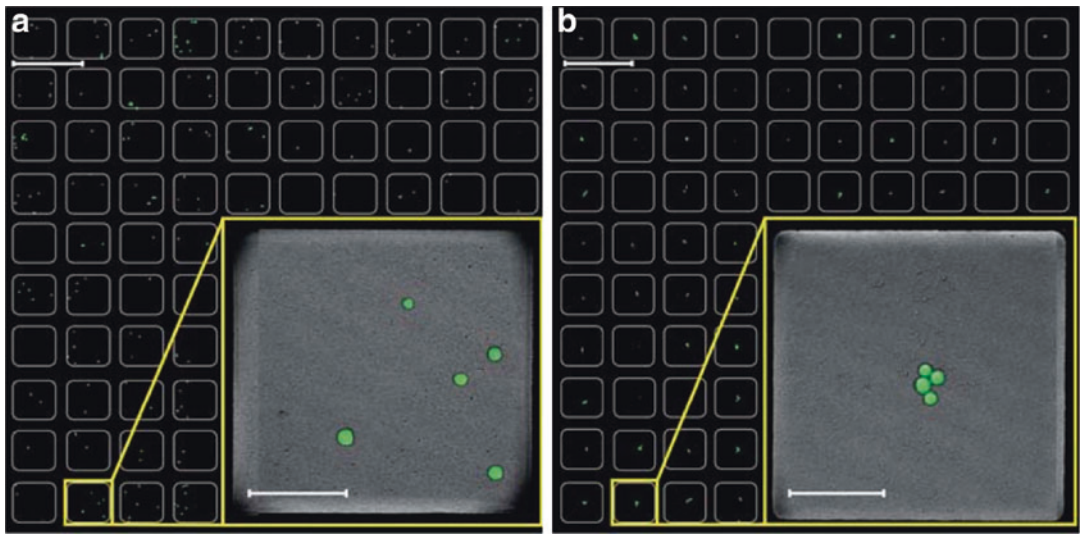
A further development of the trapping system was presented by Hammarström et al. [11]. Instead of etched microfluidic channels, a square borosilicate capillary was coupled to an external miniature transducer through a thin glycerol layer (see Fig. 5). Using commercially available capillaries both lowers the production costs and allows for disposable channels for more sensitive analysis.

Cell Viability

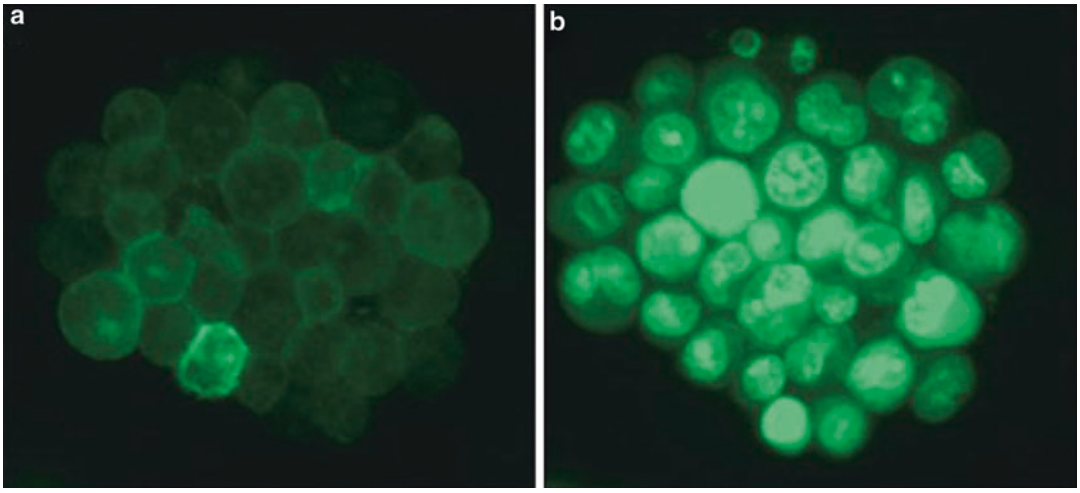
Several studies have shown that ultrasonic levitation does not seem to have any negative effects on live cells. Hultström et al. cultured cells that had been levitated for over an hour and studied the doubling times of the cells [12]. No direct or delayed damage could be detected on the cells. Bazou et al. performed a study on mouse embryonic stem cells and confirmed that they could see no changes in gene expression after ultrasonic treatments up to 1 h and the cells maintained their pluripotency [13].



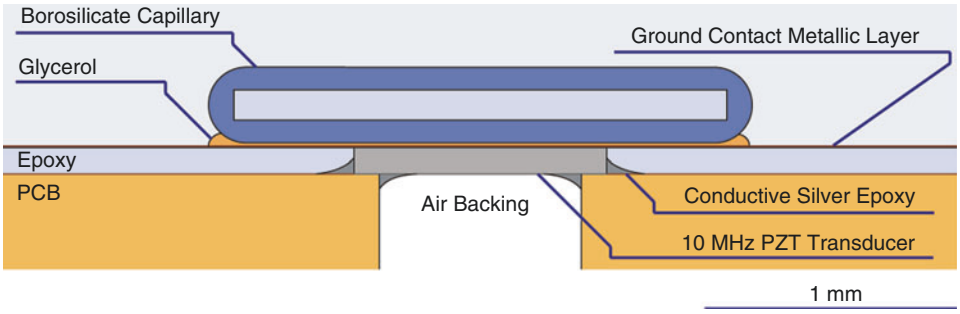
Acoustic Trapping, Fig. 2 (**a**) An acoustic trapping approach using a small transducer that creates a very localized pressure field in a channel. The transducer is embedded into the bottom of the fluidic channel and creates a large pressure gradient over a small area; (**b**) An example of an acoustic trapping cavity. Here, the entire chip is actuated, and a standing wave is created only where the geometry matches the ultrasound



Acoustic Trapping, Fig. 3 The 10 × 10 ultrasonic cavity array presented by Vanherberghen et al. [7]. In (**a**), human B cells are shown in the cavities without ultrasound, and in (**b**) the cells have been aggregated using standing waves (Reproduced with permission from The Royal Society of Chemistry: http://dx.doi.org/10.1039/c004707d/)



Acoustic Trapping, Fig. 4 Neural stem cells, HiB5-GFP, trapped in an acoustic standing wave. In (**a**) the cells have just been trapped in the standing wave, and in (**b**) the cells have been levitated for 15 min after which they were perfused with acridine orange to test for viability. The increase in fluorescence indicates that the cells are still viable (Reprinted and modified with permission from Evander [10]. Copyright 2007 American Chemical Society)



Acoustic Trapping, Fig. 5 The trapping platform presented in [11] uses a square borosilicate capillary as microfluidic channel and resonance cavity. A system with an external transducer and off-the-shelf capillaries has a lower production cost and is also compatible with more sensitive assays where disposable channels must be used (Reproduced with permission from The Royal Society of Chemistry: http://dx.doi.org/10.1039/c004504g)

Future Directions of the Field

The main research focus has so far been on technology development, but as the robustness of the systems is steadily increasing, more application-driven development can be expected. Further work aimed at integrating the trapping systems with other lab-on-a-chip systems can also be expected as well as an ambition to increase the particle size range that can be manipulated, in particular toward smaller objects such as bacteria. A lot of effort is also focused on controlling trapped clusters in three dimensions through the use of transducer matrix with phase and frequency control. These platforms will open up for exciting new applications where it will be able to have full control over clusters and perhaps single cells in microfluidic platforms.

Cross-References

- ▶ Acoustic Contrast Factor
- ▶ Acoustic Tweezers
- ▶ Acoustophoresis
- ▶ Integrated Micro-acoustic Devices

References

1. Lierke, E.G.: Akustische Positionierung – Ein umfassender Überblick über Grundlagen und Anwendungen. Acustica **82**(2), 220–237 (1996)
2. Baker, N.V.: Segregation and sedimentation of red blood-cells in ultrasonic standing waves. Nature **239**(5372), 398–399 (1972)
3. Gorkov, L.P.: On the forces acting on a small particle in an acoustic field in an ideal fluid. Sov. Phys. Dokl. **6**(9), 773–775 (1962)
4. Groschl, M.: Ultrasonic separation of suspended particles – part I: Fundamentals. Acustica **84**(3), 432–447 (1998)
5. Evander, M.: Cell and particle trapping in microfluidic systems using ultrasonic standing waves. Dissertation, Department of Electrical Measurements and Industrial Engineering and Automation, Lund University, Lund (2008).
6. Wiklund, M., Nilsson, S., Hertz, H.M.: Ultrasonic trapping in capillaries for trace-amount biomedical analysis. J. Appl. Phys. **90**(1), 421–426 (2001)
7. Vanherberghen, B., et al.: Ultrasound-controlled cell aggregation in a multi-well chip. Lab. Chip. **10**(20), 2727–2732 (2010)
8. Spengler, J.F., et al.: Observation of yeast cell movement and aggregation in a small-scale MHz-ultrasonic standing wave field. Bioseparation **9**(6), 329–341 (2000)
9. Coakley, W.T., et al.: Cell-cell contact and membrane spreading in an ultrasound trap. Colloids Surf. B: Biointerfaces **34**(4), 221–230 (2004)
10. Evander, M., et al.: Noninvasive acoustic cell trapping in a microfluidic perfusion system for online bioassays. Anal. Chem. **79**(7), 2984–2991 (2007)
11. Hammarstrom, B., et al.: Non-contact acoustic cell trapping in disposable glass capillaries. Lab. Chip **10**(17), 2251–2257 (2010)
12. Hultstrom, J., et al.: Proliferation and viability of adherent cells manipulated by standing-wave ultrasound in a microfluidic chip. Ultrasound Med. Biol. **33**(1), 145–151 (2007)
13. Bazou, D., et al.: Gene expression analysis of mouse embryonic stem cells following levitation in an ultrasound standing wave trap. Ultrasound Med. Biol. **37**(2), 321–330 (2011)

Acoustic Tweezers

- ▶ Acoustic Trapping

Acoustophoresis

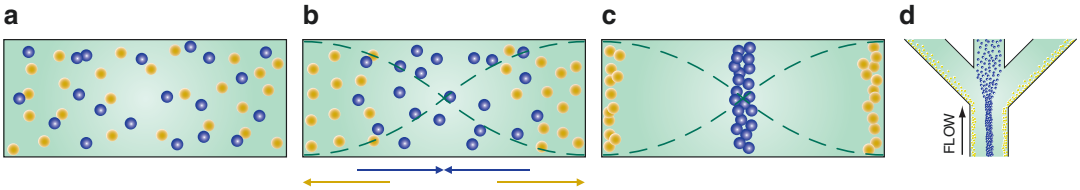
Andreas Lenshof and Thomas Laurell
Department of Biomedical Engineering, Lund University, Lund, Sweden

Definition

"Acoustophoresis" means migration with sound, i.e., "phoresis" (migration) and "acousto" (sound waves) are the executors of the movement. In related concepts, electric forces move particles in electrophoresis and magnetic forces in magnetophoresis [1]. Acoustophoresis is a non-contact and label-free mode of manipulating particles and cell populations and allows for implementation of several separation modes [2]. The technology is currently finding increased applications in bioanalytical and clinical applications of cell handling and manipulation. An extensive tutorial series on acoustophoresis was recently published by Royal Society of Chemistry [3].

Theory

Particles in suspension exposed to an acoustic standing wave field will be affected by an acoustic radiation force [4]. The force will cause the particle to move in the sound field if the acoustic properties of the particle differ from the surrounding medium. The magnitude of the movement depends on factors, such as the size of the particle, the acoustic energy density, and the frequency of the sound wave. The direction of the particle movement is dependent on the density and speed of sound of the particle as well as the liquid media. In an acoustic standing wave generated in a liquid (e.g., water)-filled channel with wall boundaries of dense materials such as metal, silicon, or glass, a standing wave pressure maxima will form at the walls/fluid interface. If the width of the channel is matched to half a wavelength, a pressure node will form in the center of the channel.



Acoustophoresis, Fig. 1 Illustration of particles with different sign of the acoustic contrast factor being exposed to an acoustic stranding wave field. The radiation force move the particles with positive contrast factor (*blue*) to the pressure node located in the center of the channel, while the particles with negative contrast factor (*yellow*) move to the antinodes at the channel side walls, (**a-c**). The laminar flow enables separation of the two particle types using a trifurcation at the end of the channel (**d**)

The most predominant acoustic force acting on microparticles in an acoustic standing wave is the primary axial acoustic radiation force (PRF), Eq. 1. The magnitude of the PRF is dependent on the acoustic energy density, E_{ac}, and the radius, a, of the particle. The direction of the particle movement depends on the acoustic contrast factor Φ, which comprises the inherent physical properties of the cell or particle, such as the density, ρ_p, and speed of sound, c_p, relative to the properties of the surrounding medium, ρ_0 and c_0. The sign of the acoustic contrast factor defines the direction of the movement.

$$F_y^{rad} = 4\pi k a^3 E_{ac} \Phi \sin(2ky) \tag{1}$$

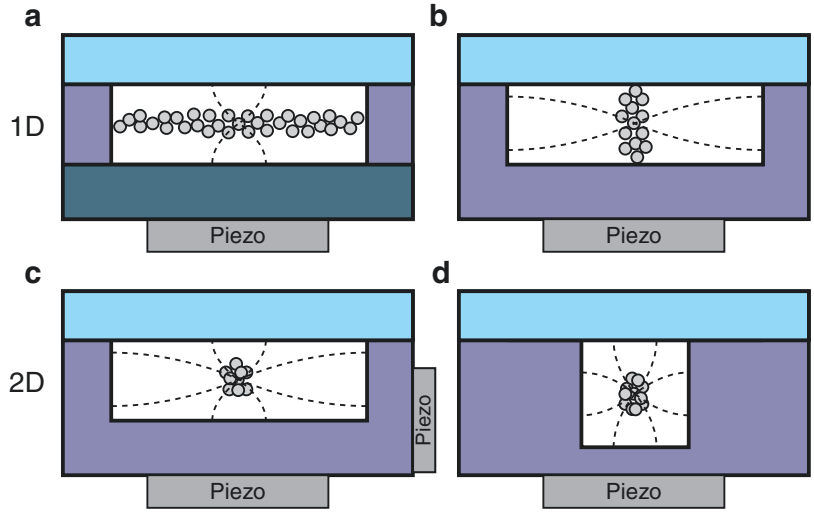
$$\Phi = \frac{\rho_p + \frac{2}{3}(\rho_p - \rho_0)}{2\rho_p + \rho_0} - \frac{1}{3}\frac{\rho_0 c_0^2}{\rho_p c_p^2} \tag{2}$$

As shown in Fig. 1, most rigid cells and particles (blue particles) have a positive contrast factor and are moved to the pressure node located in the center of the channel. Liquid vesicles or air bubbles will move to the antinodes at the wall (yellow particles). If the size of the flow channel is in microdomain, the flow conditions are generally laminar, and the particles passing through the standing wave field will be moved to their nodal position and remain in that position throughout the flow channel even after exiting the sound field. By terminating the flow channel with a trifurcation, it is possible to separate and/or concentrate the particles from the medium; see Fig. 1d.

Beside the primary axial radiation force, also secondary forces act on particles in a standing wave field [5]. These interparticle forces can mostly be negligible as they are only effective when particles are very close to each other. The secondary forces assist in particles forming clusters which are less common in continuous flow systems but are useful in stagnant systems which rely on agglomeration and sedimentation to clarify medium [6].

When designing flow channels with dimensions that match a half wavelength in the MHz frequency regime, conventional microfabrication offers simple means of fabricating acoustophoresis chips typically in glass or silicon or other high Young's modulus materials.

There are several ways of actuating an acoustophoresis chip. The most straightforward approach is to set up the standing wave between a coupling layer bonded to the transducer and the opposing channel wall, commonly referred to as a layered acoustic resonator (Fig. 2a). This creates an acoustic nodal plane perpendicular to the wave propagation. Another way is to place the transducer underneath the entire separation system or at any location on the chip surface where space is available. This causes the whole acoustophoresis chip to be actuated, not just the separation channel, but this mode of actuation requires the chip material allow waves to propagate without too much losses. By matching the actuation frequency to the channel dimension, it is possible to obtain a standing wave horizontally or vertically or in both directions in the microchannel (Fig. 2b). Two-dimensional focusing can be obtained either using two transducers of different frequencies or using a single transducer if the cavity has height to width proportions (Fig. 2c, d).



Acoustophoresis, Fig. 2 Two ways of actuation of the resonator system. (**a**) Layered resonator where the standing wave is generated in the direction of the primary direction of actuation. (**b**) Transversal resonance. The resonance is created perpendicular to the primary direction of actuation. (**c**) Two-dimensional focusing in a rectangular channel using transducers of two different frequencies. (**d**) Two-dimensional focusing in a square channel using a single transducer

The use of two-dimensional acoustic standing wave systems to pre-align sample before affecting them in a secondary acoustic separation or concentration step has shown to give superior performance over one-dimensional systems [7, 8]. This is due to the particles being forced to a common position in the flow and are thus all exposed to the same flow profile and Stoke's drag, which isolates the particles from the effect of most of the external flow factors, and the acoustic effects become prominent [9].

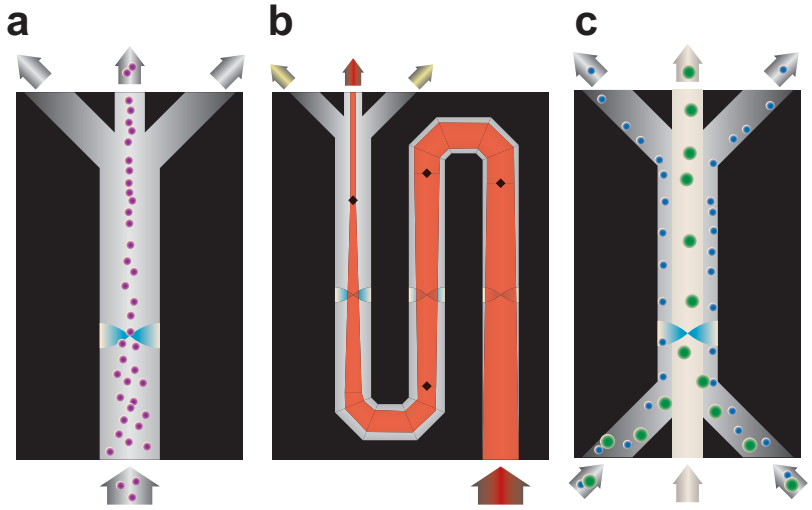
Applications

Acoustophoretic Enrichment and Depletion of Cells and Particles

Acoustophoretic enrichment/concentration of a sample is the most straightforward operation as most cells and particles easily focus on pressure nodes and thus deplete the surrounding medium of solid matter. By using flow splitters such as a trifurcation, the particle dense fraction in the central pressure node can be collected via the middle outlet of the trifurcation (Fig. 3a). Likewise, cell and particle depletion from a fluid is also obtained with the chip configuration of Fig. 3a if the fluid of interest is collected from the side branches.

However, like many microfluidic devices presented in the literature, acoustophoretic systems display problems in handling fluids of high concentrations, such as whole blood. Plasmapheresis, i.e., the removal of the cellular content from the blood plasma, is a quite common microfluidic procedure where free plasma is desired for further diagnostic purposes. Commonly, these devices are forced to work with dilute blood samples, which deteriorate the plasma composition from an analytical point of view. The acoustical chips presented earlier are no exception, and the plasma separation efficiency is known to decrease significantly when the hematocrit is increased above 5–10 %).

Means to handle samples with higher amounts of cellular content/hematocrit can be realized by a sequential acoustophoretic enrichment and partial removal of enriched cells from the standing wave node region [10]. This requires a microchannel configuration that extends over a longer distance such that the duration in the acoustic field is sufficiently long to focus cells between each point of cell removal. Figure 3b shows a schematic of a



Acoustophoresis, Fig. 3 (**a**) An acoustophoretic concentrator. The particles are concentrated in the central pressure node, while particle-depleted fluid exits to the sides. (**b**) Depletion of high cell or particle contents. Acoustically concentrated particles leave outlets in the bottom of the chip (*black diamonds*) thus lowering the particle load until eventually all particles can be removed through the central duct of the trifurcation outlet. This mode of operation is suitable for plasmapheresis of whole blood. (**c**) Wash or fractionation of cells. Sample enters the separation channel from the side branches and becomes laminated against the side walls. Clean buffer enters through the central inlet. The acoustic radiation force moves the large cells into the clean buffer and thus completes a wash in form of a medium switch

chip surface conservative meander style layout, where focused blood cells are partially removed at each outlet in the center of the channel bottom indicated as dark squares. At the end of the meander structure, the concentrations of blood cells are sufficiently low for all remaining cells to be removed via the central outlet, and the clean plasma fraction can be extracted from the side branches of the trifurcation.

Acoustophoretic Cell Washing

Microchip acoustophoresis can be configured to perform buffer exchange operations by adding two extra inlets on each side of an acoustophoresis channel. The sample solution is supplied via the side inlets, while a clean buffer is provided via the central channel inlet (Fig. 3c). Cells/particles are acoustophoretically moved from their original sample stream along the channel side wall into the central, particle-free medium stream thus exchanging the medium the particles are suspended in [11]. The typical buffer exchange efficiency displays performance data equal or better than conventional centrifugation steps and allows continuous flow-based operation, where several sequential wash steps can be implemented depending on the requirements on the wash efficiency [12]. Acoustophoretic cell and particle washing has been demonstrated for red blood cell washing in glass chips [13] and in washing of microbeads used to selectively affinity capture bacteriophages in a large phage library [14] as well as extraction of phosphopeptides from protein digests using metal ion affinity specific microbeads [12] and more recently also as a means of performing stain and wash unit operations in flow cytometry applications [15].

Acoustophoretic Separation/Fractionation

Acoustophoretic separation can be performed in two modes where:

(a) The species to be separated display different signs of the acoustic contrast factor and hence gather in the acoustic pressure node and antinode, respectively, so-called binary separation.
(b) The rate of acoustophoretic transport of a cell or particle into a pressure node determines the

lateral position in the flow stream of a given cell or particle type as they leave the separation zone, and thus, the acoustophoretic mobility determines the outcome of the separation, so-called free flow acoustophoresis – FFA.

An early clinical application of microchip-integrated acoustophoretic binary separation targeted blood wash in open-heart surgery. Patient bloodshed during the surgery is contaminated by lipid microemboli from tissue undergoing surgery and has to be removed prior to autologous re-transfusion to the patient. Jönsson et al. [16] demonstrated that acoustophoresis can remove the lipid microemboli from blood, utilizing the fact that erythrocytes display a positive acoustic contrast factor ($\sim +0.05$), while lipid particles have negative contrast factor (~ -0.18). Hence, the blood cells and the lipid vesicles will move to different lateral positions in the acoustic standing wave field and can thus be separated from each other as they pass through the acoustic separation channel (Fig. 1d) following the principles of binary separation.

Binary separation has also been demonstrated as an efficient sample preprocessing step in milk quality control where the lipid emulsion was removed by binary acoustophoresis, enabling online Fourier transform infrared spectroscopic analysis of lactose and protein content of the milk [17].

When performing separation based on free flow acoustophoresis (FFA), the major factor that influences separation outcome is the primary axial radiation force which is dependent on the size, density, and speed of sound of the particle; see Eq. 1. This commonly means that larger particles will experience a larger acoustic force than smaller-sized particles, and thus, size fractionation can be performed by acoustophoresis. Petersson et al. demonstrated successful FFA of suspensions with mixed particle sizes by tuning the acoustic force and the flow rate such that the largest particles reached the pressure node shortly before entering the center outlet, and hence, laterally from the center to the side wall, a gradient of particle sizes were located, which could be routed to individual outlets of the acoustophoresis chip [18].

By manipulating the density of the medium through the addition of a buffer that alters the fluid density, it is also possible to improve separation conditions between species that display similar acoustophoretic mobility. Buffer density manipulation in combination with the FFA fractionation device demonstrated different separation profiles of fractionation of erythrocytes, platelets, and leucocytes [18].

Free flow acoustophoresis has also been demonstrated in the preparation of peripheral blood progenitor cells from apheresis product. Platelets are an unwanted contaminant when the apheresis instrument continuously separates the buffy coat from the whole blood as they compromise the subsequent immunomagnetic extraction of stem cells from the apheresis product. Acoustophoresis can be used to solve this problem using the same chip design as in Fig. 3c where sample enters the channel at the side inlets and clean buffer enters centrally [19]. The acoustophoretic force on the platelets is very low and thus proceeds with the laminar flow along the channel side walls, while the larger leukocytes and the few remaining erythrocytes present in the apheresis product are focused on the center of the channel and exited through the central outlet as a platelet-depleted fraction [19].

Some cells, however, do not differ enough in acoustic properties to be separated by acoustic standing waves alone. By introducing affinity beads that target certain cell types, the combined bead-cell complex will experience a larger acoustic force than the unbound cells which enables a separation of the targeted cells [20].

Summary

Microchip acoustophoresis offers simple means to non-contact handling/processing of cells in continuous flow mode. Acoustophoresis is to a large extent independent of surface charge, ionic strength, and pH variations (within physiological conditions). It is also a non-perturbing technique

in the sense that cells are not experiencing any major stress when undergoing acoustophoresis as seen in several viability studies [19–21].

Cross-References

- Acoustic Contrast Factor
- Acoustic Trapping
- Acoustic Tweezers
- Integrated Micro-Acoustic Devices

References

1. Lenshof, A., Laurell, T.: Continuous separation of cells and particles in microfluidic systems. Chem. Soc. Rev. **39**(3), 1203–1217 (2010)
2. Laurell, T., Petersson, F., Nilsson, A.: Chip integrated strategies for acoustic separation and manipulation of cells and particles. Chem. Soc. Rev. **36**(3), 492–506 (2007)
3. Bruus, H., et al.: Forthcoming Lab on a chip tutorial series on acoustofluidics: acoustofluidics—exploiting ultrasonic standing wave forces and acoustic streaming in microfluidic systems for cell and particle manipulation. Lab Chip **11**, 3579–3580 (2011)
4. Gorkov, L.P.: On the forces acting on a small particle in an acoustical field in an ideal fluid. Sov. Phys. Dokl. **6**(9), 773–775 (1962)
5. Groschl, M.: Ultrasonic separation of suspended particles – part I: fundamentals. Acustica **84**(3), 432–447 (1998)
6. Trampler, F., et al.: Acoustic cell filter for high-density perfusion culture of hybridoma cells. Bio/Technology **12**(3), 281–284 (1994)
7. Augustsson, P., et al.: Microfluidic, label-free enrichment of prostate cancer cells in blood based on acoustophoresis. Anal. Chem. **84**, 7954–7962 (2012)
8. Nordin, M., et al.: Two-hundredfold volume concentration of dilute cell and particle suspensions using chip integrated multistage acoustophoresis. Lab Chip **12**, 4610–4616 (2012)
9. Jakobsson, O., et al.: Acoustic actuated fluorescence activated sorting of microparticles. Lab Chip **14**, 1943–1950 (2014)
10. Lenshof, A., et al.: Acoustic whole blood plasmapheresis chip for prostate specific antigen microarray diagnostics. Anal. Chem. **81**(15), 6030–6037 (2009)
11. Petersson, F., et al.: Carrier medium exchange through ultrasonic particle switching in microfluidic channels. Anal. Chem. **77**(5), 1216–1221 (2005)
12. Augustsson, P., et al.: Decomplexing biofluids using microchip based acoustophoresis. Lab Chip **9**(6), 810–818 (2009)
13. Evander, M., et al.: Acoustophoresis in wet-etched glass chips. Anal. Chem. **80**(13), 5178–5185 (2008)
14. Persson, J., et al.: Acoustic microfluidic chip technology to facilitate automation of phage display selection. FEBS J. **275**(22), 5657–5666 (2008)
15. Lenshof, A., Warner, B., Laurell, T. Acoustophoretic pretreatment of cell lysate prior to FACS analysis. In: Micro Total Analysis Systems 2010. Groningen (2010)
16. Jonsson, H., et al.: Particle separation using ultrasound can radically reduce embolic load to brain after cardiac surgery. Ann. Thorac. Surg. **78**(5), 1572–1578 (2004)
17. Grenvall, C., et al.: Harmonic microchip acoustophoresis: a route to online raw milk sample precondition in protein and lipid content quality control. Anal. Chem. **81**(15), 6195–6200 (2009)
18. Petersson, F., et al.: Free flow acoustophoresis: microfluidic-based mode of particle and cell separation. Anal. Chem. **79**(14), 5117–5123 (2007)
19. Dykes, J., et al.: Efficient removal of platelets from peripheral blood progenitor cell products using a novel microchip based acoustophoretic platform. PLoS One **6**, e23074 (2011)
20. Lenshof, A., et al.: Efficient purification of CD41 lymphocytes from peripheral blood progenitor cell products using affinity bead acoustophoresis. Cytom. A **85**, 933–941 (2014)
21. Hultstrom, J., et al.: Proliferation and viability of adherent cells manipulated by standing-wave ultrasound in a microfluidic chip. Ultrasound Med. Biol. **33**(1), 145–151 (2007)

Active Carbon Nanotube-Polymer Composites

Jian Chen
State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing, People's Republic of China

Synonyms

Smart carbon nanotube-polymer composites

Definition

Active carbon nanotube-polymer composites are carbon nanotube-polymer composites that display active material functions such as actuation and sensing.

Introduction

Carbon nanotubes (CNTs) represent a rare class of materials, which exhibit a number of outstanding properties in a single material system, such as high aspect ratio, small diameter, light weight, high mechanical strength, high electrical and thermal conductivities, and unique optical and optoelectronic properties. CNTs are recognized as the ultimate carbon fibers for high-performance, multifunctional polymer composites, where an addition of only a small amount of CNTs, if engineered appropriately, could lead to simultaneously enhanced mechanical strength and electrical conductivity [1, 2]. While most efforts in the field of CNT-polymer composites have been focused on passive material properties such as mechanical, electrical, and thermal properties, there is growing interest in harnessing active material functions such as actuation, sensing, and power generation in designed CNT-polymer composites. The synergy between CNTs and the polymer matrix has been judiciously exploited to create highly desirable active material functions. In this entry, recent progress in active CNT-polymer composites is briefly highlighted with a focus on smart materials and infrared (IR) sensors.

Shape-Memory CNT-Polymer Composites

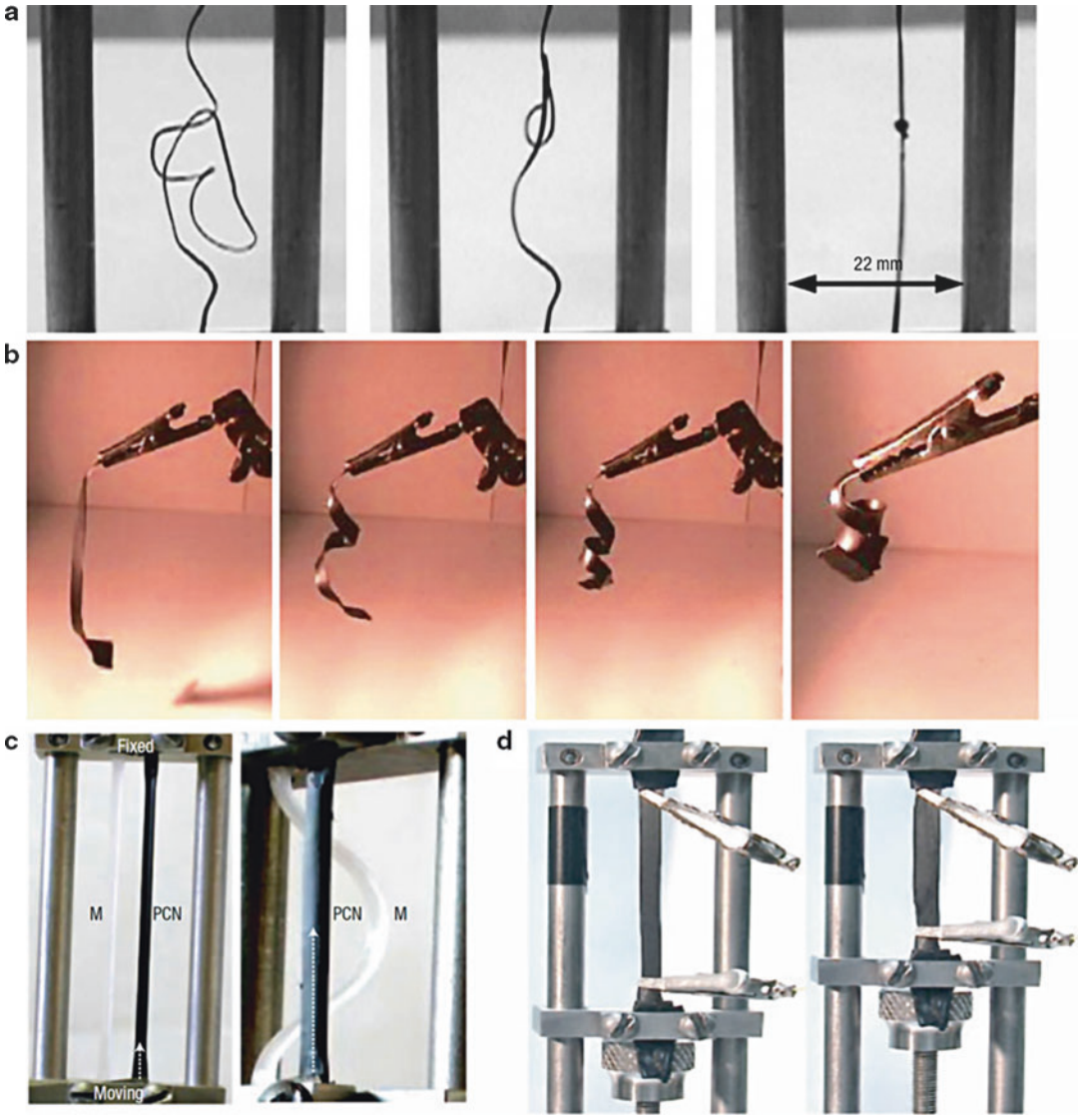
Shape-memory polymers (SMPs) are polymeric smart materials that can memorize a temporary shape and are able to return from a temporary shape to their permanent shape upon exposure to an external stimulus such as heat [3]. Compared with shape-memory alloys and ceramics, SMPs offer a number of distinctive advantages, which include high recoverable strain (up to 400 %), low density, ease of processing and the ability to tailor the recovery temperature, programmable and controllable recovery behavior, and low cost. Such advantages could enable a broad spectrum of demanding applications including deployable space structures, morphing wings, information storage, smart textiles, biomedical devices, and drug delivery.

Although SMPs show promising shape-memory effects, several major issues remain to be addressed: (1) slow recovery speed; (2) low recovery stress; nnn lack of remote control. Vaia and coworkers demonstrate that the addition of multiwalled CNTs (MWNTs) into a thermoplastic polyurethane matrix (Morthane) could address these issues simultaneously (Fig. 1) [4]. The low recovery speed (up to several minutes) of thermal-responsive SMPs originates from their intrinsically low thermal conductivity (<0.3 $Wm^{-1}K^{-1}$) [3]. Therefore, the heat transfer from an external heating source to the core of a SMP sample will take considerable amount of time. Vaia and coworkers show that non-radiative decay of IR photons absorbed by the nanotubes raises the internal temperature, melting strain-induced polymer crystallites (which act as physical cross-links that secure the deformed shape), and remotely trigger the release of the stored strain energy. The 1 wt% MWNT-Morthane nanocomposite displays a rapid shape-recovery within 5 s upon exposure to an IR light (Fig. 1b). Comparable effects occur for electrically induced actuation associated with Joule heating of the matrix when a current is passed through the conductive percolative network of the nanotubes within the resin in a 16.7 wt% MWNT-Morthane nanocomposite (Fig. 1d).

The low recovery stress of thermal-responsive SMPs originates from their intrinsically low modulus [3]. CNTs have proven to enhance the mechanical properties, particularly modulus, of various polymers. For example, 8.5 wt% MWNTs could increase the room temperature rubbery modulus of Morthane by a factor of 5 [4]. In fact, incorporation of 5 wt% of MWNTs into polyurethane SMPs results in an increase of the recovery stress by $\sim$ 230 % [4].

CNT-Polymer Composite Actuators

Polymeric actuators are polymers that deform in response to an external stimulus such as heat, light, or electric field, and they normally recover their original shapes after the stimulus is terminated [5]. Potential applications of polymeric



Active Carbon Nanotube-Polymer Composites, Fig. 1 (**a**) Stretched (800 %) Morthane ribbon containing 1 wt% MWNTs, tied into a loose knot and heated at 55 °C. The knot closes on strain recovery. (**b**) Strain recovery and curling of the 1 wt% MWNT-Morthane nanocomposite ribbon upon IR irradiation within 5 s. (**c**) Comparison of the stress recovery before (*left*) and after (*right*) remote actuation by IR irradiation. Neat Morthane (*M*) bends and does not recover. In contrast, a 1 wt% MWNT- Morthane nanocomposite (*PCN*) contracts on exposure to IR irradiation (*arrow* indicates moving direction). (**d**) Electrically stimulated stress recovery of a 16.7 wt% MWNT-Morthane nanocomposite (Reprinted by permission from Macmillan Publishers Ltd [4])

actuators include artificial muscles, mini- and microrobots, "smart skins," pumps, and valves in microfluidic systems for drug delivery, ventricular assist devices for failing hearts, and sensors for mechanical strain, humidity, and gases. Researchers have recently demonstrated that the synergistic combination of CNTs and a suitable polymer matrix could lead to promising CNT-polymer composites with either new or significantly enhanced actuation properties.

Nematic liquid crystalline elastomers (LCEs) have fascinated scientists and engineers

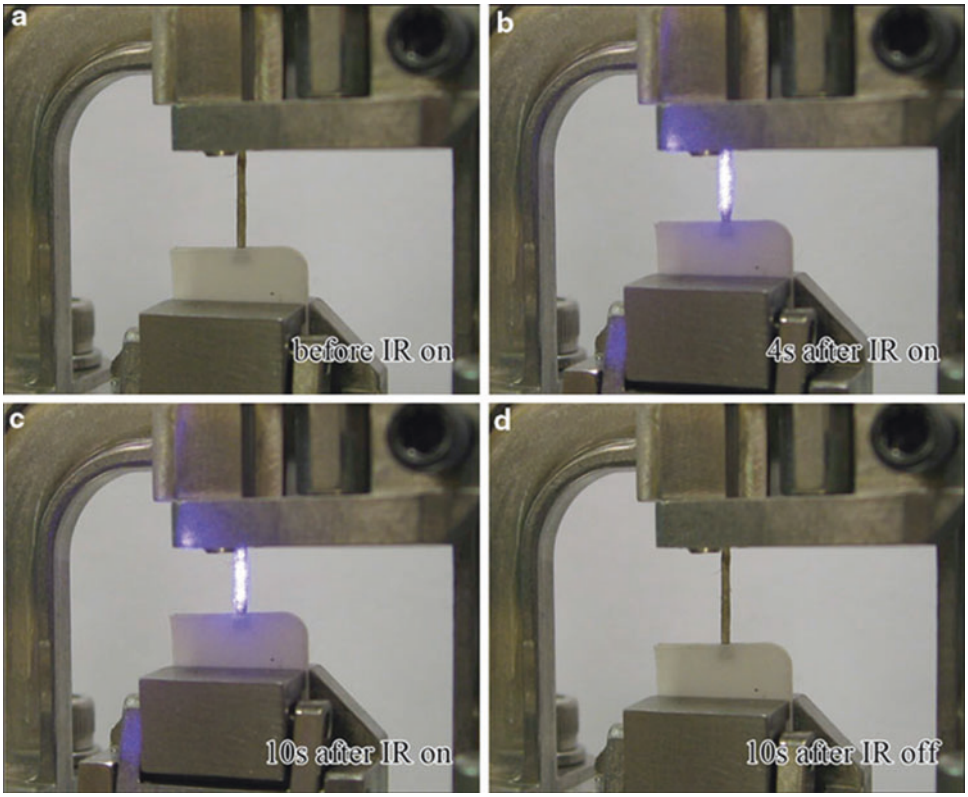
continually since 1981 when they were first synthesized by Finkelmann and coworkers [6]. LCEs bring together, as no other polymer, three important features: orientational order exhibited by the mesogenic units in amorphous soft materials; topological constraints via cross-links; and responsive molecular shape due to the strong coupling between the orientational order and the mechanical strain [7]. Acting together, they create many new physical phenomena that can lead to many potential applications. Nematic LCEs can dramatically and reversibly elongate or contract in response to temperature changes; however, this type of LCE actuator tends to have a slow response time due to the low thermal conductivity of LCEs and cannot be actuated remotely, thereby limiting its potential applications.

Terentjev and coworkers have studied the rich photomechanical behavior of three types of MWNT-elastomer nanocomposites irradiated with near-IR light, including polydimethylsiloxane, styrene–isoprene–styrene, and a nematic LCE with a polysiloxane backbone [8, 9]. They show that these composite materials have the novel ability to change their actuation direction, from expansive to contractive response, as greater imposed strain is applied to the sample.

Unlike MWNTs, which have featureless visible/near-IR absorption, single-walled CNTs (SWNTs) show strong and specific absorptions in the visible/near-IR region owing to bandgap transitions [10]. Recent advances in nanotube separation make it possible to obtain enriched single-chirality SWNTs with strong and narrow optical absorptions, which may ultimately offer unique opportunities for the wavelength-selective IR actuation of LCE nanocomposites. The effective use of SWNTs in polymer-composite applications strongly depends on the ability to disperse them uniformly in a polymer matrix without destroying their integrity. Pristine SWNTs are incompatible with most solvents and polymers, which lead to poor dispersion of the nanotubes in solvents and polymer matrices. Chen and coworkers have recently developed a versatile, nondamaging chemistry platform that enables them to modify specific CNT surface properties, while preserving the CNT's intrinsic properties [11]. They have discovered that rigid, conjugated macromolecules, for example, poly(*p*-phenyleneethynylene) (PPE), can be used to noncovalently functionalize and solubilize CNTs and disperse CNTs homogeneously in polymer matrices [2, 11].

Chen and coworkers have recently reported the reversible IR actuation behavior of a new type of nematic SWNT-LCE nanocomposite using PPE-functionalized SWNTs (PPE-SWNTs) as a filler in a LCE matrix with side-on mesogenic units [12]. The SWNT-LCE nanocomposites have been prepared through a two-stage photopolymerization process coupled with a "hotdrawing" technique, which enable the fabrication of relatively thick films (∼160–260 μm). The excellent dispersion of PPE-SWNTs in the LCE matrix has allowed them to observe a significant and reversible IR-induced strain (∼30 %) at very low SWNT loading levels (0.1–0.2 wt%, Fig. 2). The effects of various parameters have been investigated, including the degree of pre-alignment, the loading level of the SWNTs, and the curing time. In SWNT-LCE composites, semiconducting SWNTs can efficiently absorb and transform IR light into thermal energy, thereby serving as numerous nanoscale heaters being uniformly embedded in the LCE matrix. The absorbed thermal energy then induces the LCE nematic-isotropic phase transition, which leads to the shape change of the nanocomposite film. The IR strain response of SWNT-LCE nanocomposites increases with an increasing SWNT loading level and a higher degree of hot-drawing, and decreases with longer photocuring [12].

Electric field driven pure nematic LCE actuators have not been developed because the characteristic rotation energy density is too low compared with the characteristic resistance energy density from the rubbery elastic network due to the low anisotropic dielectric permittivity of the LCEs, even under a very high electric field [13]. However, adding CNTs with high anisotropic polarizability into the LCEs, aligned along the uniaxial director of a monodomain nematic elastomer network, the CNT-LCE nanocomposites below the conductivity percolation threshold can achieve an effective dielectric anisotropy many

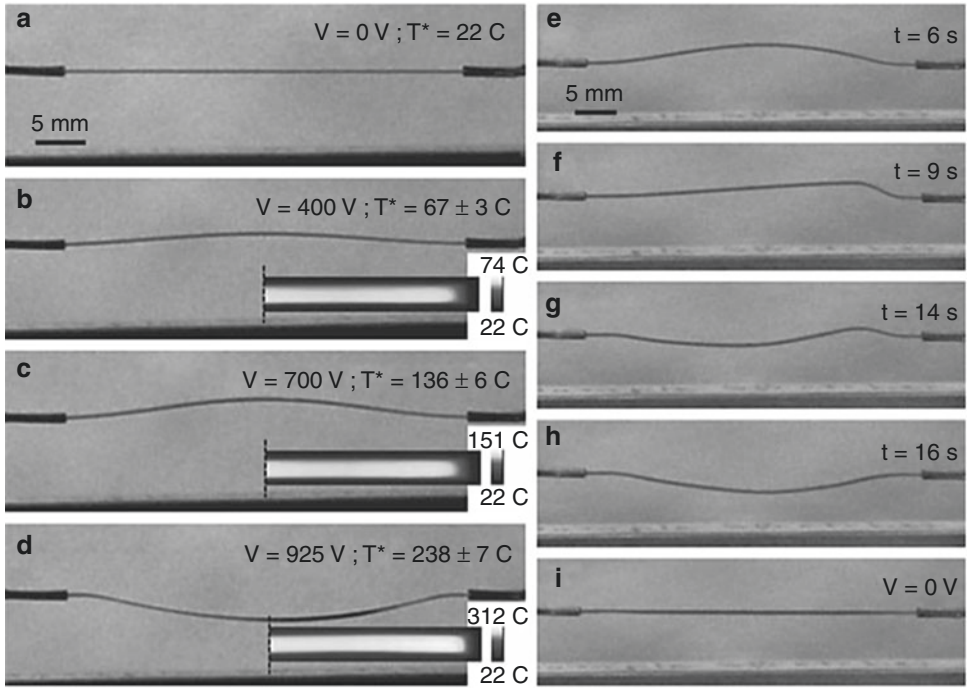


Active Carbon Nanotube-Polymer Composites, Fig. 2 IR actuation of a 0.2 wt% SWNT-LCE nanocomposite film: (**a**) before the IR was turned on; (**b**) after the IR was turned on for 4 s; (**c**) after the IR was turned on for 10 s; (**d**) the film returned to its original length with the IR beam turned off (Reprinted by permission from Wiley-VCH Verlag GmbH & Co. KGaA [12])

orders of magnitude higher than the LCEs alone [13]. Therefore, introducing highly polarizable anisotropic CNTs into the LCE will lead to a significantly enhanced electromechanical effect and make electrical-field driven LCE actuators possible. Terentjev and coworkers have demonstrated, for the first time, a large electromechanical response (uniaxial stress of $\sim$ 1 kPa in response to a constant field of $\sim$ 1 MV/m) in nematic LCEs filled with a very low ($\sim$0.0085 wt%) concentration of MWNTs, pre-aligned along the nematic director at preparation [13]. However, the response stress in the isostrain conditions shows a linear dependence on the applied field E, instead of E^4 that would be expected from a CNT rotation-induced electromechanical response, a puzzling finding to the researchers as stated in the original paper [13].

Park and coworkers have investigated the electromechanical properties of a 0.05 wt% SWNT/LaRC-EAP (Langley Research Center-ElectroActive Polyimide) composite, which forms an intrinsic unimorph during the fabrication process to actuate without the need for additional inactive layers [14]. The 0.05 wt% SWNT/LaRC-EAP exhibits a large strain ($\sim$2.6 %) at a low driving voltage ($<$1 MV/m) while possessing excellent mechanical and thermal properties. The out-of-plane strain is proportional to the square of the electric field E, which suggests that the actuation mechanism of the SWNT/LaRC-EAP composite is primarily due to electrostriction [14].

The addition of CNTs also enables electrically stimulated thermo-mechanical actuation via electrical Joule heating when the loading level of CNTs is above its electrical conductivity

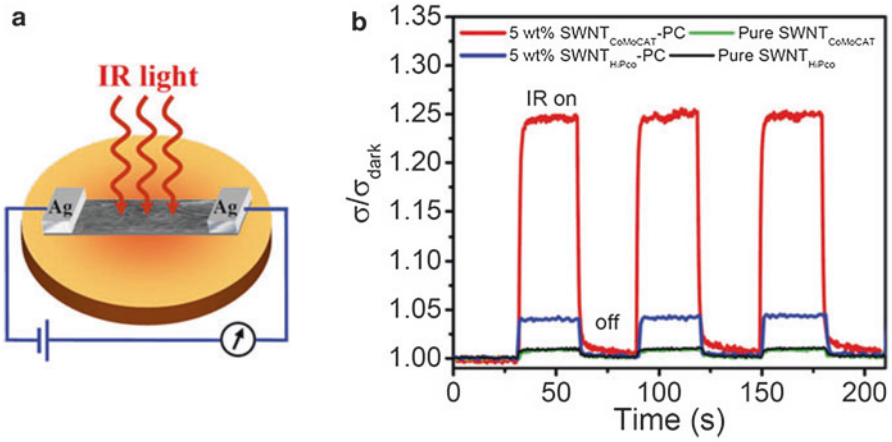


Active Carbon Nanotube-Polymer Composites, Fig. 3 (**a**–**d**) Deflection of a 0.5 vol% SWNT-CP2 beam (L = 4 cm; d = 140 μm; w = 1.9 mm) as a function of applied voltage and temperature. For $T^* < T_g$ (~220 °C), thermal expansion induces a positive deflection in the beam due to a buckling instability (**b**, **c**). At $T^* \approx T_g$, mechanical softening promotes a deflection inversion as gravity becomes the dominant force acting on the beam (**d**). Sample temperature was determined at each voltage using the thermal images shown inset; (**e**–**i**) Deflection of a 0.5 vol% SWNT-CP2 beam as a function of time at V_{th} (voltage where T* is first observed to exceed T_g). The sample initially buckles as a result of thermal expansion (**e**); however, as the sample temperature continues to increase, the sample kinks (**f**, **g**), and then eventually sags under its own weight at T_g (h). Once the voltage is removed, the beam reversibly actuates back to its original form (**i**) (Reprinted by permission from Wiley-VCH Verlag GmbH & Co. KGaA [15])

percolation threshold. Vaia and coworkers have studied the electrothermal properties of SWNT-polyimide (CP2) nanocomposites (Fig. 3) [15]. They show that the reversible softening associated with temperature transitions through T_g can provide substantial strain increases ($\Delta S \approx 4$ %) over relatively small changes in applied electric field ($\Delta E \approx 0.01$ MV/m).

CNT-Polymer Composite IR Sensors

Organic electronic materials offer ease of materials processing and integration, low cost, physical flexibility, and large device area as compared to traditional inorganic semiconductors. Optoelectronic materials that are responsive at wavelengths in the near-infrared (NIR) region (e.g., 800–2,000 nm) are highly desirable for various demanding applications such as telecommunication, thermal imaging, remote sensing, thermal photovoltaics and solar cells [16]. SWNTs have strong absorptions in the NIR region owing to the first optical transition (S_{11}) of semiconducting nanotubes [10]. The inverse diameter dependence of S_{11} optical transition energy enables the wavelength-tuning of SWNT's NIR absorptions, which, coupled with the rapid progress in nanotube separation, will ultimately allow the development of SWNT IR sensors tailored to specific



Active Carbon Nanotube-Polymer Composites, Fig. 4 (**a**) Schematic experimental setup. IR light covers all sample area. (**b**) Relative conductivity (σ/σ_{dark}) response of SWNT-PC nanocomposites to the on/off IR illumination (power intensity: 7 mW/mm^2)

regions of the NIR spectrum. There are a number of reports on the IR photoelectrical property of SWNT films. Levitsky and coworker have demonstrated that the arc-produced SWNT ($SWNT_{arc}$) film is capable of generating a very weak photocurrent upon continuous-wave IR illumination (12 mW/mm^2) in the air at room temperature, and the current increase upon IR illumination is only about 0.2 % [17]. In addition, they have observed a dark current drift due to oxygen adsorption and a relatively slow rise/decay (~4–5 s relaxation time) of the photocurrent in response to the on/off illumination. Haddon and coworkers have reported that the IR photoresponse in the electrical conductivity of a $SWNT_{arc}$ film is dramatically enhanced when the nanotube film is suspended in vacuum at low temperature (e.g., 50 K) [18].

Chen and coworkers have discovered that the IR photoresponse in the electrical conductivity of SWNTs is dramatically enhanced by embedding SWNTs in an insulating polymer matrix such as polycarbonate (PC) in the air at room temperature (Fig. 4) [19, 20]. In contrast to the gradual photoresponse and weak conductivity change (1.10 %) observed in a HiPco®-produced SWNT ($SWNT_{HiPco}$) film in the air at room temperature, the 5 wt% $SWNT_{HiPco}$-PC nanocomposite under the same IR illumination (power intensity: 7 mW/mm^2) shows a sharp photoresponse and strong conductivity change (4.26 %) [19]. The dark current drift owing to oxygen adsorption is also minimized by embedding nanotubes in the polymer matrix. They have shown that semiconducting SWNTs are critical to the IR photoresponse of nanotube materials [19, 20]. They have also demonstrated that both SWNT types and nanotube-matrix polymer-nanotube junctions have profound impact on the IR photoelectrical property of SWNT-polymer composites [20]. Composite IR sensors based on CoMoCAT®-produced SWNTs ($SWNTs_{CoMoCAT}$) significantly outperform those based on $SWNTs_{HiPco}$. The 5 wt% $SWNT_{CoMoCAT}$-PC nanocomposite demonstrates a very strong conductivity change of 23.45 % upon the IR illumination (7 mW/mm^2) in the air at room temperature [20], which is 5.5 times of that (4.26 %) observed in the 5 wt% $SWNT_{HiPco}$-PC composite film [19], nearly 27 times of that (0.88 %) observed in the pure $SWNT_{CoMoCAT}$ film [20], 21 times of that (1.10 %) observed in the pure $SWNT_{HiPco}$ film [19], and 117 times of that (0.2 %) observed in the pure $SWNT_{arc}$ film in the air at room temperature (12 mW/mm^2) [17]. In addition, the 5 wt% $SWNT_{CoMoCAT}$-PC composite film shows a

detectable IR photoresponse at a light intensity as low as 23.4 μW/mm^2 [20].

Acknowledgments JC is greatly indebted to his students and collaborators and his colleagues in this field, whose names are cited in the references. Financial support from the National Science Foundation (DMI-06200338, CMMI-0625245, and CMMI-0856162), UWM start-up fund, UWM Research Growth Initiative award, and the Lynde and Harry Bradley Foundation is gratefully acknowledged.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Functionalization of Carbon Nanotubes
- ▶ Insect Infrared Sensors
- ▶ Light-Element Nanotubes and Related Structures
- ▶ Nanorobotics
- ▶ Nanostructured Materials for Sensing
- ▶ Organic Actuators
- ▶ Synthesis of Carbon Nanotubes
- ▶ Thermal Actuators

References

1. Winey, K.I., Kashiwagi, T., Mu, M.: Improving electrical conductivity and thermal properties of polymers by the addition of carbon nanotubes as fillers. MRS Bull. **32**, 348–353 (2007)
2. Chen, J., Ramasubramaniam, R., Xue, C., Liu, H.Y.: A versatile, molecular engineering approach to simultaneously enhanced, multifunctional carbon nanotube-polymer composites. Adv. Funct. Mater. **16**, 114–119 (2006)
3. Leng, J., Lu, H., Liu, Y., Huang, W.M., Du, S.: Shape-memory polymers – a class of novel smart materials. MRS Bull. **34**, 848–855 (2009)
4. Koerner, H., Price, G., Pearce, N.A., Alexander, M., Vaia, R.A.: Remotely actuated polymer nanocomposites – - stress-recovery of carbon-nanotube-filled thermoplastic elastomers. Nat. Mater. **3**, 115–120 (2004)
5. Madden, J.D., Vandesteeg, N.A., Anquetil, P.A., Madden, P.G.A., Takshi, A., Pytel, R.Z., Lafontaine, S.R., Wieringa, P.A., Hunter, I.W.: Artificial muscle technology: physical principles and naval prospects. IEEE J. Ocean. Eng. **29**, 706–728 (2004)
6. Finkelmann, H., Kock, H.-J., Rehage, G.: Investigation on liquid crystalline siloxanes: 3. liquid crystalline elastomers. Makromol. Chem. Rapid Commun. **2**, 317–322 (1981)
7. Warner, M., Terentjev, E.M.: Liquid Crystal Elastomers. Oxford University Press, Oxford (2003)
8. Ahir, S.V., Terentjev, E.M.: Photomechanical actuation in polymer-nanotube composites. Nat. Mater. **4**, 491–495 (2005)
9. Ahir, S.V., Squires, A.M., Tajbakhsh, A.R., Terentjev, E.M.: Infrared actuation in aligned polymer-nanotube composites. Phys. Rev. B **73**, 085420 (2006)
10. Hamon, M.A., Itkis, M.E., Niyogi, S., Alvaraez, T., Kuper, C., Menon, M., Haddon, R.C.: Effect of rehybridization on the electronic structure of single-walled carbon nanotubes. J. Am. Chem. Soc. **123**, 11292–11293 (2001)
11. Chen, J., Liu, H., Weimer, W.A., Halls, M.D., Waldeck, D.H., Walker, G.C.: Noncovalent engineering of carbon nanotube surfaces by rigid, functional conjugated polymers. J. Am. Chem. Soc. **124**, 9034–9035 (2002)
12. Yang, L., Setyowati, K., Li, A., Gong, S., Chen, J.: Reversible infrared actuation of carbon nanotube-liquid crystalline elastomer nanocomposites. Adv. Mater. **20**, 2271–2275 (2008)
13. Courty, S., Mine, J., Tajbakhsh, A.R., Terentjev, E.M.: Nematic elastomers with aligned carbon nanotubes: new electromechanical actuators. Europhys. Lett. **64**, 654–660 (2003)
14. Park, C., Kang, J.H., Harrison, J.S., Costen, R.C., Lowther, S.E.: Actuating single wall carbon nanotube-polymer composites: intrinsic unimorphs. Adv. Mater. **20**, 2074–2079 (2008)
15. Sellinger, A.T., Wang, D.H., Tan, L.-S., Vaia, R.A.: Electrothermal polymer nanocomposite actuators. Adv. Mater. **22**, 3430–3435 (2010)
16. McDonald, S.A., Konstantatos, G., Zhang, S., Cyr, P. W., Klem, E.J.D., Levina, L., Sargent, E.H.: Solution-processed PbS quantum dot infrared photodetectors and photovoltaics. Nat. Mater. **4**, 138–142 (2005)
17. Levitsky, I.A., Euler, W.B.: Photoconductivity of single-wall carbon nanotubes under continuous-wave near-infrared illumination. Appl. Phys. Lett. **83**, 1857–1859 (2010)
18. Itkis, M.E., Borondics, F., Yu, A., Haddon, R.C.: Bolometric infrared photoresponse of suspended single-walled carbon nanotube films. Science **312**, 413–416 (2006)
19. Pradhan, B., Setyowati, K., Liu, H., Waldeck, D.H., Chen, J.: Carbon nanotube-polymer nanocomposite infrared sensor. Nano Lett. **8**, 1142–1146 (2008)
20. Pradhan, B., Kohlmeyer, R.R., Setyowati, K., Owen, H. A., Chen, J.: Advanced carbon nanotube/polymer composite infrared sensors. Carbon **47**, 1686–1692 (2009)

Active Nanoantenna System

▶ Active Plasmonic Devices

Active Plasmonic Devices

Jean Christophe Lacroix[1], Pascal Martin[2] and Hyacinthe Randriamahazaka[2]
[1]Interfaces, Traitements, Organisation et Dynamique des Systemes Universite Paris 7-Denis Diderot, UMR 7086, Paris, France
[2]University Paris 7-Denis Diderot, ITODYS, Nanoelectrochemistry Group, UMR CNRS 7086, Paris, France

Synonyms

Active nanoantenna system

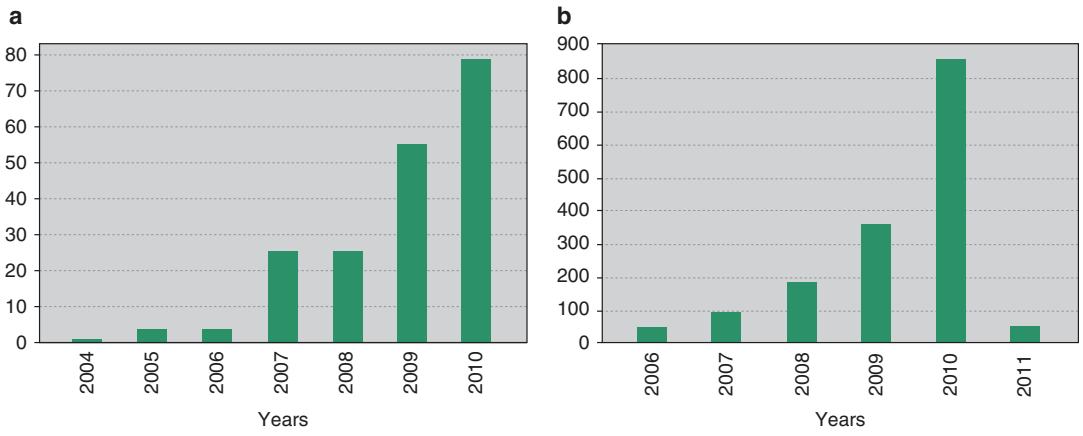
Definition

Active plasmonic devices combine plasmonic systems and a physical or chemical control input.

Main Text

Published papers using the term "active plasmonics" and citations of those papers are shown in Fig. 1a, b. The first papers explicitly devoted to this new field were published in 2004 and 2005, even though many previous publications had already pointed out the need for such devices. The growth of these two curves clearly demonstrates that breakthroughs have been achieved in the last 5 years and that many new devices will briefly be demonstrated.

Active plasmonic systems generally use two components: one of them supporting *Localized Surface Plasmons* (LSP) or *Surface Plasmon Polaritons* (SPPs) [1]. One needs to distinguish two types of approaches. In the first approach, the plasmonic component acts as an input in order to control or tune the properties of the second component. Such active plasmonic systems are widely studied but do not belong to the field of active plasmonic devices as described in this Essay. Surface Enhanced Raman Spectroscopy (SERS) is an example of such approaches using an extreme light concentration in order to develop a highly sensitive analysis that can now even detect single molecules [2]. In the second approach, the second component is used in order to control, switch, or tune the properties of the plasmonic component, and the plasmonic response to the external stimulus is the output of the device. This review will be devoted only to this second approach, even though some examples of plasmonic systems



Active Plasmonic Devices, Fig. 1 Histogram of (**a**) published papers and (**b**) number of citations of papers including terms "active plasmonics"

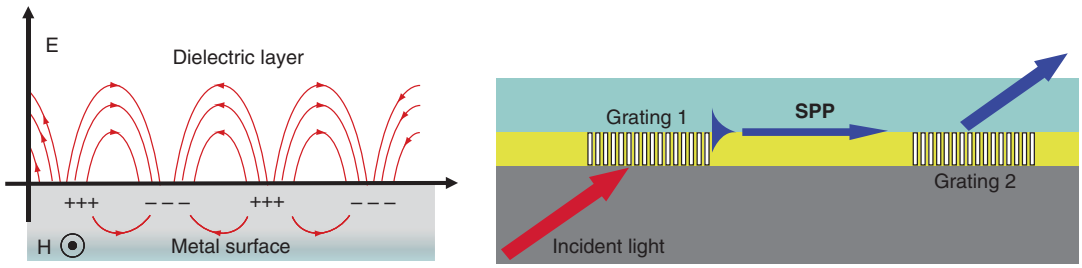
using the first approach will be given when they are of interest in the design of active devices.

Surface Plasmons: Qualitative Description

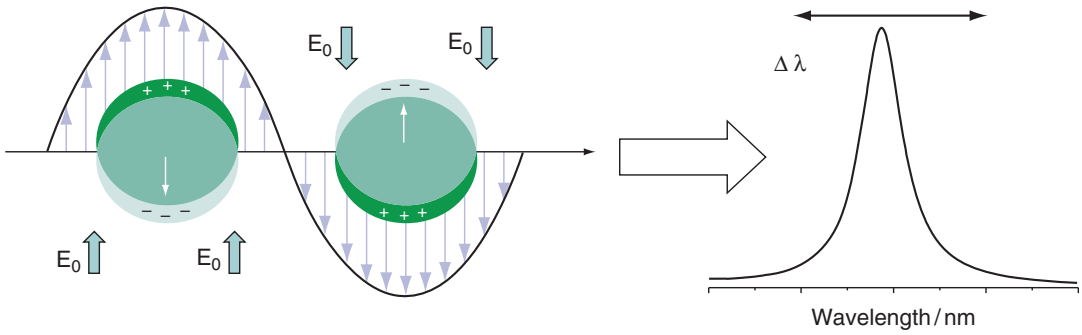
In order to understand the evolution depicted in Fig. 1 and the tremendous research and technological opportunities of this new field, one has first to have in mind why in the last decade passive plasmonic devices have been studied so much and what kinds are of interest.

Briefly, they can be divided in two classes.

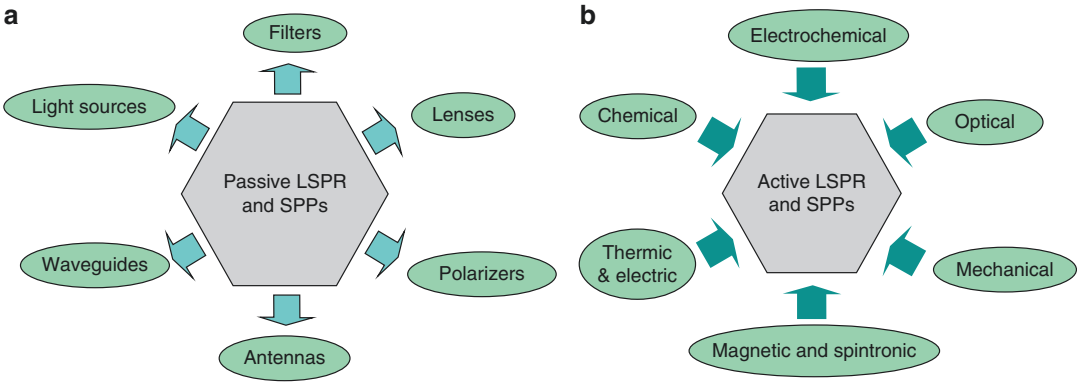
- Those using SPPs, i.e., electromagnetic waves coupled to the coherent oscillation of free-charge carriers at the interface between a dielectric and a conductor (Fig. 2). SPPs propagate along subwavelength nanostructured metals or semiconductors and are often referred to as low-dimensional waves due to the strong confinement of the electromagnetic field on the surface [3].
- Those using LSPs, some metal NanoParticles (NPs) exhibit coherent but confined oscillations of the quasi-free electrons in the conduction bands (Fig. 3). When the characteristic frequency of these oscillations coincides with that of the light excitation, the response of the metallic NPs becomes resonant and strong absorption in the visible and near-infrared range occurs (so-called localized surface plasmon resonance, LSPR). The frequency of LSPR strongly depends on the size, shape, spacing of the particles, and on the dielectric constants of the substrate and surrounding medium [4].



Active Plasmonic Devices, Fig. 2 SPs propagating along the interface between metal and dielectric materials. These waves have a combined electromagnetic and surface charge character. Schema of a plasmonic signal, coupled to and from the waveguide by gratings on a metallic layer



Active Plasmonic Devices, Fig. 3 Schema of the localized oscillations of electrons exposed to an electric field. The electrons are shifted, and the polarization charges on opposite surface elements apply a restoring force on the electrons. The extinction spectra of NPs shows the localized surface plasmons resonance peak which shift with the shape, size, and spacing of particles



Active Plasmonic Devices, Fig. 4 (**a**) Major plasmonic devices. (**b**) Overview of major routes for the development of active plasmonics

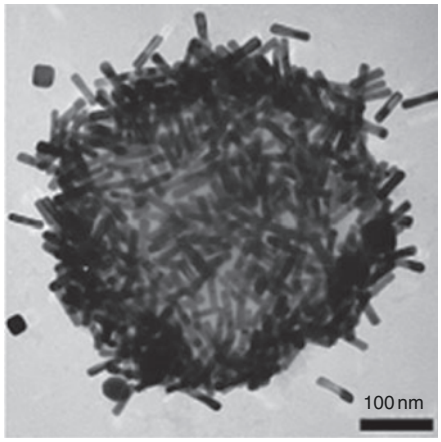
Because of surface plasmons capability to enhance electric fields in the very close vicinity of NP structures or to localize and guide light in metallic structures, plasmonics allows the manipulation of the flow of light and its interaction with matter at the nanoscale. In this sense, NPs work in a similar way to that of antennas in radio and telecommunication systems, but at optical frequencies, i.e., at frequencies corresponding to typical electronic excitations in matter. Plasmonics also offer an opportunity to merge photonics and electronics at nanoscale dimensions to obtain unusual optical properties and opportunities for unprecedented levels of synergy between optical and electronic functions. Figure 4a depicts the main passive devices already demonstrated, including light sources, filters, waveguides, lenses, plasmonic antennas, and polarizers [5].

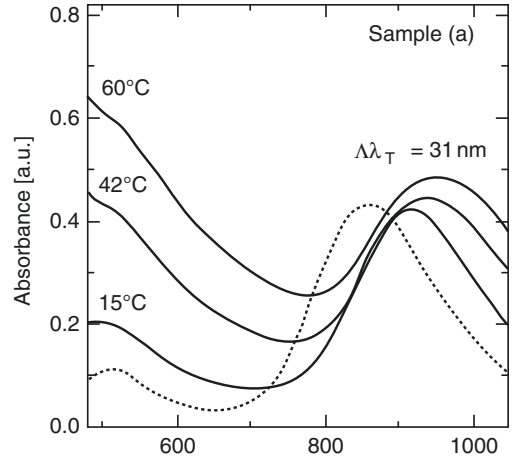
Nevertheless, in order to become a high valuable technology, plasmonics needs switches, modulators, and routers. Indeed, such active plasmonic systems are needed to achieve functional Surface Plasmons (SP) circuits that will, for instance, first convert light into SPs using nanoantennas; SPP would then propagate along SP waveguides and be processed by logic elements before being converted back to light using nanoscopic sources. Active plasmonic devices can thus be the basic building blocks for an alternative technology capable of writing, reading, storing, and processing information at the nanoscale. Figure 4b show the various types of physical and chemical external stimuli already demonstrated to control passive plasmonic devices. Currently, their number is expanding.

Many recent reviews [1] have been devoted to passive plasmonics, and this contribution will only describe active plasmonic devices, and few aspects of theoretical background will be used to define and understand strategies to reversibly switch or modulate plasmons.

Controlling Localized Surface Plasmons

The frequency of LSPR depends mainly on the size, shape, and spacing of the particle and on the dielectric constants of the substrate and surrounding medium. Adding to a LSPR-based plasmonic system, an external stimulus that will tune the size, the shape, the spacing of the particle or the dielectric functions of the surrounding medium can therefore lead to active plasmonic devices. Some strategies will easily allow reversibility; others based on tuning the size and shape of the NPs will be less successful for reversible control. Of particular interest is the recent discovery of mixed exciton-plasmon states in metal-molecule complexes, which makes them promising candidates for active molecular plasmonic devices.





Active Plasmonic Devices, Fig. 5 TEM image of a single hybrid particle based on Gold nanorod-loaded poly-(NIPAM-co-AAA) microgel with high surface coverage. UV-Vis spectra of hybrid sample at 15 °C (swollen state), at 42 °C, and at 60 °C (collapsed state) (Reprinted with permission from Karg et al. [6])

Thermally Driven Active Plasmonic Devices

The LSPR is not tuned by changing the temperature of a substrate. A thermoresponsive layer surrounding NPs could be use to design *thermally driven active plasmonic devices*. Most examples use thermoresponsive polymers, inorganic layers, and liquid crystals in which the thermal effect induces a variation of the spacing between the NPs or of the dielectric function of the surrounding medium. For all these systems, a temperature-driven phase transition is observed and induces a plasmon resonance shift.

Several systems based on the thermally induced and reversible collapse of poly (N-isopropylacrylamide), pNIPAM have been reported (Fig. 5) [6]. Inorganic composites were also used. For example, Metal–VO_2 nanocomposites have shown LSPR shifts of about 50 nm for a temperature variation of 95 °C.

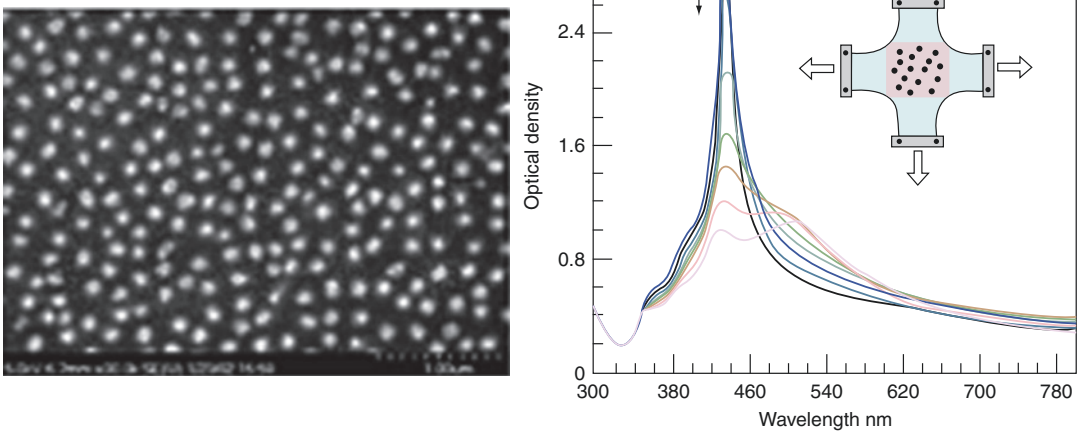
Other thermoresponsive active plasmonic devices have been developed using liquid crystals (LCs). LCs are unique materials that show such characteristic properties as self-organization, fluidity with long-range order, cooperative motion and alignment change with various external stimuli (temperature, electric fields, etc.) at surfaces and interfaces. Thereby, the modulation of alignment of LCs gives rise to a change in optical properties, which forms the basis of LC displays (LCDs) [7]. The LSPR in anisotropic media such as liquid crystals are strongly dependent on the local ordering of the media, and that induced orientational transitions of the liquid crystal may be used to tune the optical properties of metallic NPs.

Mechanical Input

An alternative approach is to fabricate plasmonic structures on mechanically tunable substrates. In most of the cases, small metallic nanospheres have been deposited on stretchable elastomeric films like polydimethylsiloxane (PDMS). As the film is stretched to different degrees, the intensity of plasmon resonance rapidly decreases, and a new band appears at higher wavelength (Fig. 6).

Chemically Driven Plasmonic Devices

LSPR modulation using molecular recognition is a mature technology that is widely used to design sensors and biosensors. The LSPR of isolated NP shows limited sensitivity of absorption spectra (for small colloidal 5–20 nm nanoparticles), to molecular or biomolecular recognition events due to the small changes of the refractive index of the surrounding medium induced by the molecular recognition events. Additionally, the LSPR wavelength and the UV-Vis absorption of the molecules do not match in most cases.

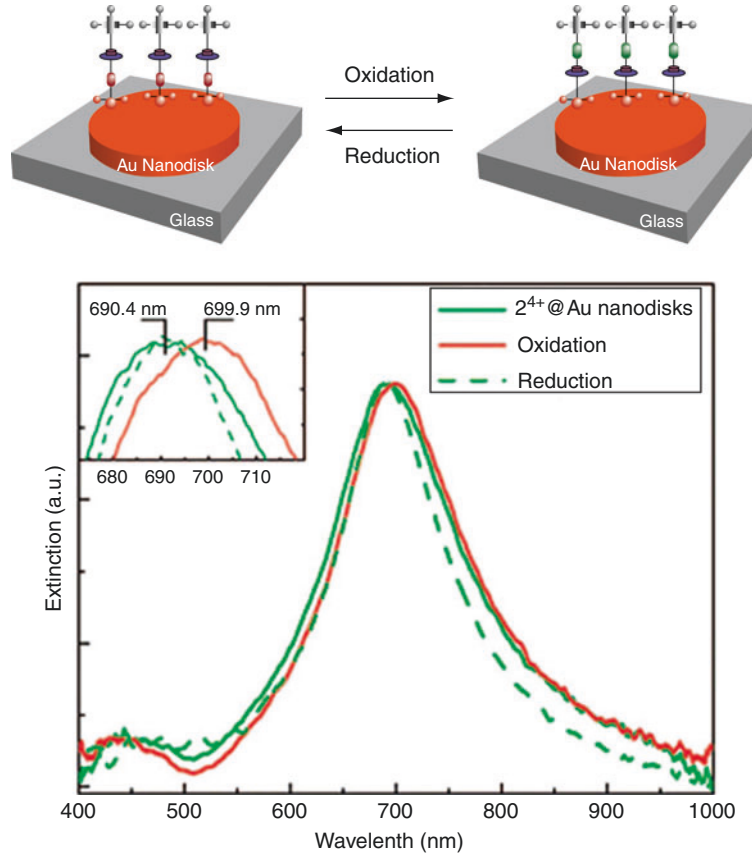


Active Plasmonic Devices, Fig. 6 SEM image of the 2D array of Ag nanoparticles into PDMS. Extinction spectra of the array as a function of the stretching force. *Red curve* represents the film stretched 50 % from its original, nonstretched state (*black curve*) (Reprinted with permission from Journal of American Chemical Society, Malynych and Chumanov [8])

However, LSPRs are highly sensitive to interparticle spacing, originated from the near-field plasmon coupling of nanoparticles. The strength of the effect falls rapidly with distance; at separations above 2.5 times the particle size, the nanoparticles behave essentially as isolated clusters. In particular, particle aggregation results in a pronounced red shift of the LSPR peaks, and most of the applications for sensors are based on this distance variation, resulting in the change of their plasmon coupling. Recent reviews [9] have described the large body of research on plasmon-based sensors. Indeed, even though such devices can be considered as active plasmonic devices, their purpose is not to control LSPR but to use LSPR as a transduction of the molecular events occurring at the NP surface. This part will only focus on some recent systems illustrating the various chemical stimuli which tuned LSPR.

As first examples, the pH of a solution could act as a chemical input. In a typical configuration, metal nanoparticles are immobilized as a monolayer on a glass substrate, while a polymer material is either deposited on the top of the nanoparticles monolayer or between nanoparticles. PH-induced conformational changes will alter the interparticle distance and, hence, the strength of plasmon coupling. The variation of the nanoparticles–island distance was limited to 1–20 nm, which corresponds to the optimal distance between metal nanoparticles for electromagnetic coupling. Reversible operation was achieved by returning the pH and optical properties of the system to the initial state.

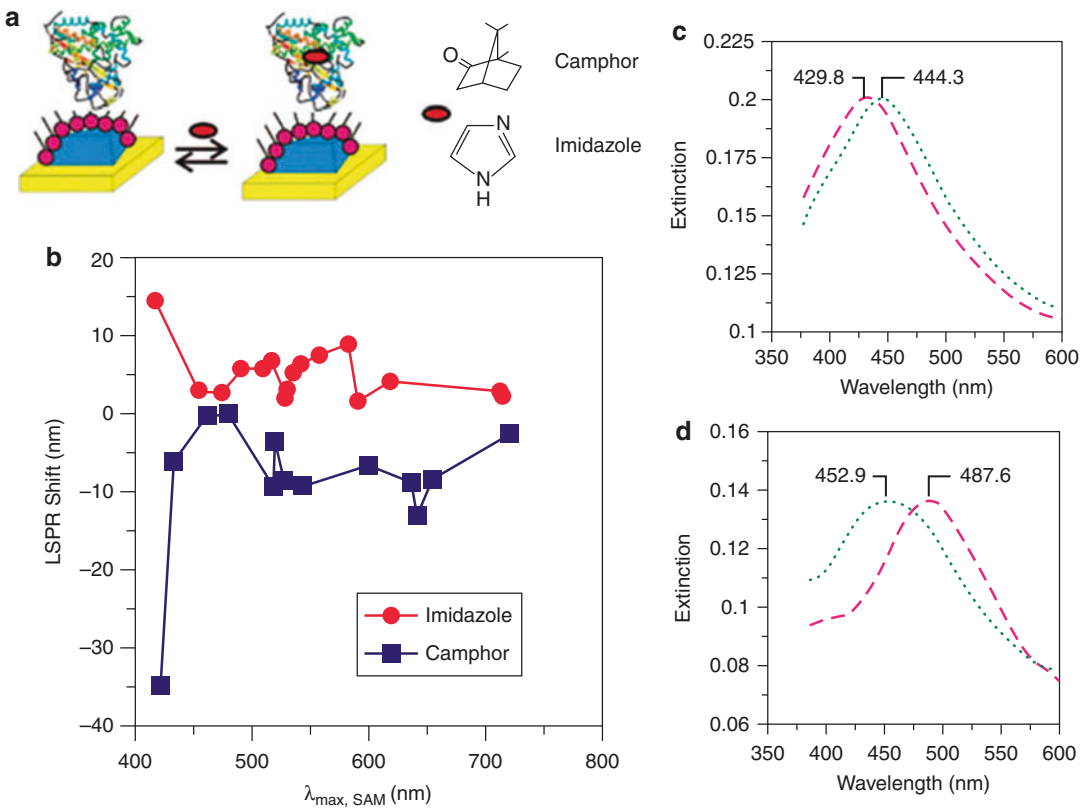
Electroactive molecular machines with grafted NPs have also been used but, in this case, the input was a chemical chosen to reduce or oxidize the electroactive machine. Zheng et al. recently showed that a gold nanodisc array, coated with bistable, redox-controllable [2]rotaxane molecules, when exposed to chemical oxidants and reductants, undergoes reversible switching of its plasmonic properties (Fig. 7) [10]. These observations suggest that the nanoscale movements within surface-bound "molecular machines" can be used as the active components in plasmonic devices. One has, nevertheless, to note that the observed LSPR shifts were rather small (10 nm) due to the small change of monolayer's effective refractive index and the very thin molecular film.



Active Plasmonic Devices, Fig. 7 Schematic of working principle of the molecular-machine-based active plasmonics. Extinction spectra recorded from Au nanodisks in air during the redox process (Reprinted with permission from Zheng et al. [10])

In these previous examples, the molecules used did not absorb light near the nanoantenna extinction. New phenomena occur when LSPR and molecular absorption coincide. When the peak position of the NPs LSPR did not overlap with the molecular resonance, a small shift (20 nm for example) was observed upon monolayer deposition, whereas a higher shift is observed when the LSPR peak partially overlaps with the molecular resonance [11]. The oscillatory dependence of the peak shift on wavelength was shown to track with the wavelength dependence of the real part of the refractive index of the monolayer, as determined by a Kramers–Koenig transformation of the molecular resonance absorption spectrum.

The influence of surrounding adsorbates is important even for monolayers as explained above or at low-coverage adsorption. In such cases, it is difficult to estimate the dielectric constant and an atomistic description is needed [12]. Thus, the influence of molecular adsorbates on the absorption spectra of spherical NPs was investigated using arbitrary coupling to the surface and changing the molecular excitation characteristics through its HOMO–LUMO gap. For the weak adsorption regime, the adsorbates influenced the resonance width; the resonance energy, however, is increased by less than 10 nm. For strong coupling to the surface, and when the excitation of the adsorbates becomes lower in energy than that of the plasmon, it was shown that the peak splits into two excitations. The higher-energy broad peak was shown to be located at the particle surface and was, therefore, attributed to plasmonic excitation. The lower-energy excitation was located over the adsorbates and was attributed to an in-phase oscillation of molecular dipoles on opposite sites of the NP. This new excitation is similar to a completely symmetric molecular exciton, in which molecular dipole oscillations couple



Active Plasmonic Devices, Fig. 8 Schematic representation of small molecule binding to receptors on functionalized Ag nanoparticles (*below*). The wavelength-dependent LSPR shift induced by inhibitor (*red line with dots*) and substrate (*blue line with squares*). LSPR spectra of NPs with cytochrome before (*dashed line*) and after (*dotted line*) substrate (*up*) or inhibitor (*down*) binding (Reprinted with permission from Zhao et al. [13])

in-phase. However, the electrostatic coupling between molecular dipoles occurs through the metallic particle, allowing very significant interaction at large distances.

Strong coupling of chromophore and plasmon resonances has been exploited for molecular sensors. Zhao et al. showed an extreme LSPR shift for isolated NPs due to small changes in the resonances of a chromophore that are induced by binding an additional analyte [13]. In this case, the binding of substrate or inhibitor molecules to a cytochrome caused its absorption band to shift to shorter or longer wavelengths, respectively, and to be *off-resonance* with the plasmon, which was at the origin of a huge LSPR shift for isolated NP (Fig. 8).

The creation of such responsive plasmonic devices by coupling a chromophore to the plasmon through absorption overlap with the wavelength of the LSPR is an important new research trend for applications in molecular sensing and active plasmonics. Excitons and plasmons are the two main excited states of matter. Understanding their interactions is likely to become a very important research field. It yields to mixed exciton-plasmon states, so-called hybidized states. Indeed, when excitons of molecules, such as in J-aggregates, resonate with the plasmon modes of metal NPs, new peaks appear in the extinction spectra. These peaks are governed by the strength of the coupling between excitons and plasmons, and strategies based on tuning this strength, through some physical or chemical input, have recently been reported and considerably widen the opportunities to reach active molecular plasmonic devices. Change of the

angle of incident light or change of the geometry of the NPs was used to control the coupling strength. For strong coupling, the LSPR splits into two bands upon adsorption of the J-aggregates. Reversible tuning of the coupling strength was not obtained since tuning was generated using the flexible geometry or changing the incident angle. Based on these recent results, the reversible tuning of exciton–plasmon coupling using other stimuli will be an easy and flourishing means to develop active plasmonic devices.

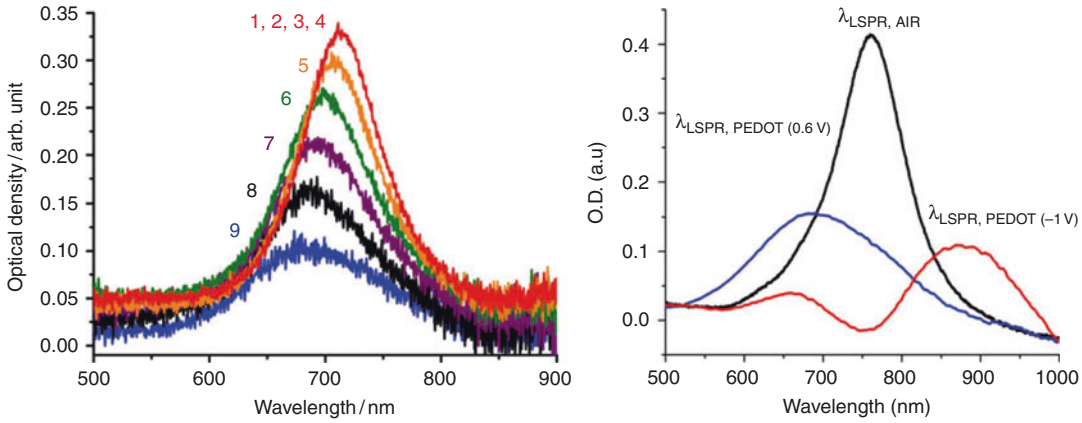
Electrochemically Driven Active Plasmonic Devices

Nanoelectrochemistry is rapidly expanding its frontiers toward many new aspects including redox-gated molecular junctions and photovoltaic integrated systems. In this context, electrochemical switching is a useful tool to reversibly control the properties of metallic NPs.

Electrochemistry on nanostructured electrodes supporting LSPR, used as the working electrode, has been widely investigated. Sweeping or changing the potential of such an electrode in an electrolyte usually yields very small LSPR wavelength shifts. Ung et al. were the first to monitor the position of the surface plasmon band of silver colloids as a function of the applied potential [14]. The SP shift amounted to 10 nm/V in the potential regime dominated by double-layer charging. With various electroinactive organic shells a higher blue shift could be obtained. The LSPR shift was attributed to the increased of metal plasma frequency due to the excess stored electrons.

Changing the spacing of the particle through electrochemistry has not yet been widely used, probably because reversibility is not easy to obtain. Electrochemical etching induces LSPR wavelength modulation but is unlikely to be used for reversible active plasmonic devices. On the contrary, reversible control through electrochemistry of the surrounding dielectric medium is easy to obtain. Deposition on the NPs of an electroactive thin film or of an electroactive monolayer makes it possible to trigger LSPR through the oxidoreduction properties of the deposited layer and thus through the applied potential to the electrode. For example, Ag NPs, arranged in closely spaced two-dimensional arrays, were coated by a thin film of tungsten oxide [15]. WO_3 is a well-known electro-active electrochromic material that undergoes a reversible redox process to form reduced tungsten oxide. Reduced and oxidized forms of WO_3 have different electrical conductivities and thus different dielectric functions. By switching the externally applied potential, the optical extinction signal was reversibly modulated and a blue shift was observed.

Following that, Lacroix et al. reported the use of electroactive conductive polymer (CP) thin layers deposited on gold gratings [16]. In a first study, they used polyaniline (PANI) because such films exhibit ultrafast switching (around 1 μs, RC limited) between their reduced non-conductive state and their oxidized conductive state. Blue shift was observed from $\lambda_{LSP} = 633–571$ nm, below that of λ_{LSP} observed in air (593 nm), accompanied by a strong damping of LSPR. This reversible effect was observed repeatedly for more than 1 h. The interpretation lies in the switching of the dielectric constant ε of PANI. When PANI is in its reduced state, ε is real, whereas in its oxidized state, ε is complex ($\varepsilon = \varepsilon_{Re} + i\,\varepsilon_{Im}$). Moreover, oxidized metallic PANI samples follow a Drude-like behavior [1] with a plasma frequency $\omega_p = 2$ eV yielding a simplified expression of ε_{Re} given by $\varepsilon_{Re} = \varepsilon\propto - (\omega_p/\omega)$ [2]. The imaginary part of the dielectric constant can be expressed as $\varepsilon_{Im} = (2\sigma\lambda/c)$, where λ is the vacuum wavelength of incident light, σ the wavelength-dependent conductivity, and c the speed of light. The electrochemical switching of PANI thus generates a large change in both the real and the imaginary part of the dielectric function of the material around the nanoparticles. The large blue shift can be attributed to a decrease in the real part of the dielectric constant at wavelength around 632.8 nm, whereas switching in the imaginary part of the dielectric constant (from zero to a finite value) induces quenching and damping of the LSP. In a second step, they showed that the LSPR in such a PANI/gold NP hybrid material exhibits a progressive damping as a function of the applied potential, as evidenced in



Active Plasmonic Devices, Fig. 9 (*Left*) Extinction spectra of prolate gold grating under x-polarized light and polarization from −500 to 500 mV (Reprinted with permission from Leroux et al. [17]). (*Right*) Extinction spectra of prolate gold particle grating (X polarized light) in air (*black curve*); overcoated with a PEDOT/SDS film in its oxidized state (*blue curve*); overcoated with a PEDOT/SDS film in its reduced state (*red curve*) (Reprinted with permission from Stockhausen et al. [18])

Fig. 9 (left). These results clearly show that the observed modulations of the LSPR are closely related to the charge-carrier density injected into the conductive polymer through electrochemical doping.

More recently, poly(3,4-ethylenedioxythiophene) (PEDOT) was used to shift the LSPR of gold NP arrays [18]. Figure 9 shows the giant red shift from 685 to 877 nm under polarization at −1.0 V. The reversible shift between reduced and oxidized states is 192 nm. Such a giant plasmon resonance shift, covering almost half of the visible spectrum, is, to the best of our knowledge, unique.

Bringing together two systems used in two emerging scientific areas with important applications, i.e., plastic electronics and plasmonics, can thus yield new active plasmonic devices. In this context, it is important to evaluate the potentialities and the flexibility of such devices. CPs can be tailored at will in order to change their switching potential or the variation of their dielectric constant upon switching which will give electrochemically driven plasmonic devices with tunable optical properties. Switching speed is another key point for photonics application and in the present case, switching times of the LSPR in the millisecond range have been observed. Electrode size reduction allows switching times in the microsecond range, and it can thus be predicted that switching frequencies in the MHz range could be easily obtained. Moreover, the external stimuli inducing CP switch is not restricted to the electrochemical input. CPs can also be doped using light, and all-optical solid-state devices using CPs for controlling plasmonic systems are likely to be demonstrated in the near future. Finally, other plasmonic systems, such as waveguides or nanoscopic light sources, can also be combined with plastic electronic components providing a wide range of active plasmonic components for solid-state nanophotonics.

Electric Field-Driven Active Lasmonic Devices

Another input for controlling LSPR is the application of an external electric field. The applied voltage can then be used to exert direct control over an electrooptically active dielectric medium, changing its refractive index and thus the SPR modes at the metal/dielectric interface. Using an electric field as an input could lead to faster switches than thermally, mechanically, chemically, or electrochemically driven systems, and is especially significant because both plasmonic and electric signals can be guided in the same metallic circuitry.

Coupling LSPR devices and organic field-effect transistors has not yet been reported but is

likely to be developed in the next couple of years. On the contrary, many studies have used liquid crystals (LC) or ferroelectric films. In all these cases, shifts have been observed, and the spectral position of the surface plasmon resonances depends on the applied voltage. The variations of the local density of the LC molecules and/or changes in the refractive index in the active films provide a sufficient effect to modulate the LSPR of gold nanostructures. The switching time for the system was in the order of ms.

Optically Driven Active Plasmonic Devices

Light-driven plasmonic devices have several potential advantages: they could operate faster, light can be used for inducing (writing) as well as detecting (reading) the behavior of plasmonic devices; and they can be integrated in future all-optical plasmonic circuits. Efficient methods to achieve optically driven plasmonic devices have been developed and are based on tuning the spacing between particles or the dielectric constant of the surrounding medium.

In this context, it is worth mentioning that there are a number of reports on photosensitive Au NPs with surface coating with small molecules bearing a photochromic moiety. In the case of colloidal metallic NPs, azobenzene trans-cis photoisomerization was used to induce NP aggregation, resulting in the change of the interparticle distance and a shift of the surface plasmon resonance. Sidhaye et al. reported the red shift (close assembly of particles) and the blue shift (loose assembly of the gold nanoparticles) of the surface plasmon peak position ($\Delta\lambda = 80$ nm) induced by the reversible manipulation of interparticle spacing in gold nanoparticle networks connected with azobenzene derivatives [19]. Isomerization yields and switching speeds in such devices remain clear limitations, even though lowering the surface coverage of azobenzene provided more free space for isomerization to take place and enhanced the LSPR shift. Reversibility was clearly demonstrated.

Finally, several studies described the use of specific photoresponsive liquid crystals due to their large and controllable birefringence [7]. The surrounding media was homogeneous mixtures of a liquid crystal and a photoswitchable molecule such as azobenzene. Upon photoirradiation, the LCs undergo a phase transition and a change in their refractive index, leading to a LSPR shift of few of tens nm (for Gold arrays). Reversibility was demonstrated, and a response time in the second range was observed.

Magnetic and Magneto-optic Driven Active Plasmonic Devices

The study of the excitation of SPs on ferromagnetic materials has recently attracted much interest not only from a fundamental point of view but also because of the possibility of realizing spin-plasmonic devices, bringing together the rapidly developing plasmon technology and the field of spintronics. Active spin-plasmonic devices have been developed for controlling SPP and will be briefly reviewed in the next part of this essay. To the best of our knowledge, the only study, combining LSPR and magneto-active materials, have been recently reported by Du et al., using spherical and elliptical nanocylinders [20]. However, they focused on how plasmons can enhance the magneto-optical properties of a thin magnetic metal layer (Co, Ni or Fe). Active control of the LSPR using magnetic layers remains to be demonstrated.

Controlling Surface Plasmon Polaritons

Surface plasmon polaritons (SPPs) are electromagnetic waves coupled to the coherent oscillation of free charge carriers at the interface between a dielectric and a conductor. The propagation length of SPPs depends on the metal, even though resistive heating losses in metals can severely limit the performance of devices. Several passive plasmonic devices, such as waveguides, and nanoscopic light sources have been demonstrated. In the latter system, the light could be transmitted through holes with lateral dimensions smaller than half the incident wavelength, while classical theory predicts orders-of-magnitude less transmission (extraordinary or enhanced optical transmission [EOT]) [3]. For a better efficiency, the metal should be corrugated with a periodic

structure such as an array of holes or a single hole surrounded by grooves.

In order to fully exploit the possibilities of such SPP-based plasmonic devices, it is necessary to achieve active control. Since SPP fields are strongly confined, they are again very sensitive to refractive index changes within a few tens of nanometers of the surface. As a consequence, many strategies used to control SPPs are similar to those used for LSPR tuning.

Most systems are solid-state devices avoiding the use of liquid, and when molecules are used, they are deposited as thin films and not used in solution which excludes, to some extent, PH or redox input. A notable exception lies in Surface Plasmon Resonance (SPR) spectroscopy which is very sensitive and widely used to design sensors and biosensors [9]. Readers are directed toward other reviews for many examples of such system. One of the considered configuration replaces the "dielectric" layer in the structure with an active thin film which can be artificially doped by suitable inputs and allows the dynamic modulation of the SPP wavevector [5]. One of the first examples was demonstrated by Krasavin and Zheludev in [21] and will be described in the optical input section.

Inorganic semiconducting materials supporting SPP are also used, whereas metal NP remains the main basic blocks for LSPR devices. Indeed, as the carrier densities in semiconductors are much lower than those in metals, the plasma frequency is much smaller, being typically at mid- or far-infrared frequencies. The main advantage of semiconductors is that their carrier density and mobility, and consequently the SPPs, can be easily controlled by many external stimuli (thermal, electrical, and optical excitation).

Thermally Driven Active Plasmonic Devices

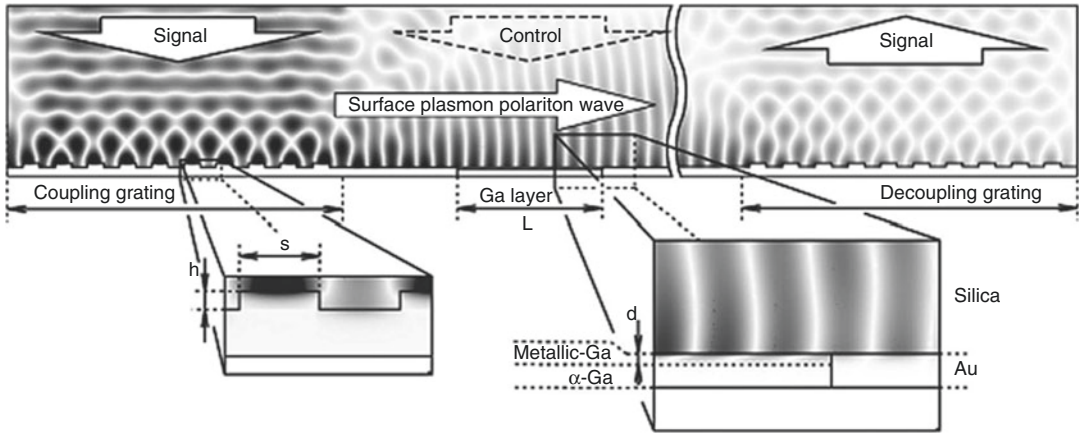
Nikolajsen's group fabricated and characterized thermo-optic Mach–Zender interferometric modulators (MZIMs) and directional-coupler switches (DCSs) at telecom wavelengths [22]. MZIMs and DCSs operation utilize the SPP waveguiding along thin gold stripes embedded in polymer and heated by electrical signal currents. The operation of a thermo-optic MZIM is based on changing the SPP propagation constant in a heated arm resulting in the phase difference of two SPP modes that interfere in the output Y-junction. By sandwiched, the stripes between a thick layer of benzocyclobutene, they observed the modulation of the SPP propagation (extinction ratio of 35 dB for a 8 mW electrical power). This achieved driving power was considerably lower than that of conventional thermo-optic MZIMs.

In the field of EOT, ordered arrays of nanoapertures have been considered as field enhancers and controllers for near-field electromagnetic waves. The EOT transmission takes place in a double-layer structure (a metal-oxide layer). The characteristics of the transmission spectra are fixed by the dielectric constants of the materials that bound the array and by the hole's depth and pitch. Using a specific oxide layer in which semiconductor–metal transition is observed; the transmission of light incident on a metallic nanoholes array is modulated by switching, at transition temperature, the oxide layer between metallic and semiconducting phases with distinctly different dielectric functions.

Electric Field-Driven Active SPP Devices

Replacing the "dielectric" layer in the structure with an electro active thin film (for example barium titanate or GaAs layer) allows the dynamic modulation of the SPP wavevector. Under an electric field, the orientation of ferroelectric domains is affected by film stresses. Thus, the SPP waves are modulated by the switch of the in-plane to out-of-plane domains.

Coupling field-effect transistor with plasmonic devices supporting SPP has not yet been widely developed. In 2009, a Metal-Oxide-Semiconductor (MOS) field-effect plasmonic modulator has nevertheless been demonstrated with a device operating at 1.55 μm [23]. A four layer metal-MOS–metal (Ag-SiO_2-Si-Ag) waveguide structure supporting both photonic and plasmonic modes was fabricated. Its transmission coefficient is determined by interference between the two modes. When the MOS is driven into accumulation under a bias of 0.75 V, the Si index changes and the photonic mode is cut off, which leads to transmission modulation. Switching time of 10 ns has been demonstrated



Active Plasmonic Devices, Fig. 10 Representation of gold waveguide containing a Ga switching layer (Reprinted with permission from Krasavin and Zheludev [21])

(limited by the experimental apparatus), but gigahertz frequencies could be reachable.

Optically Field-Driven Active SPP Devices

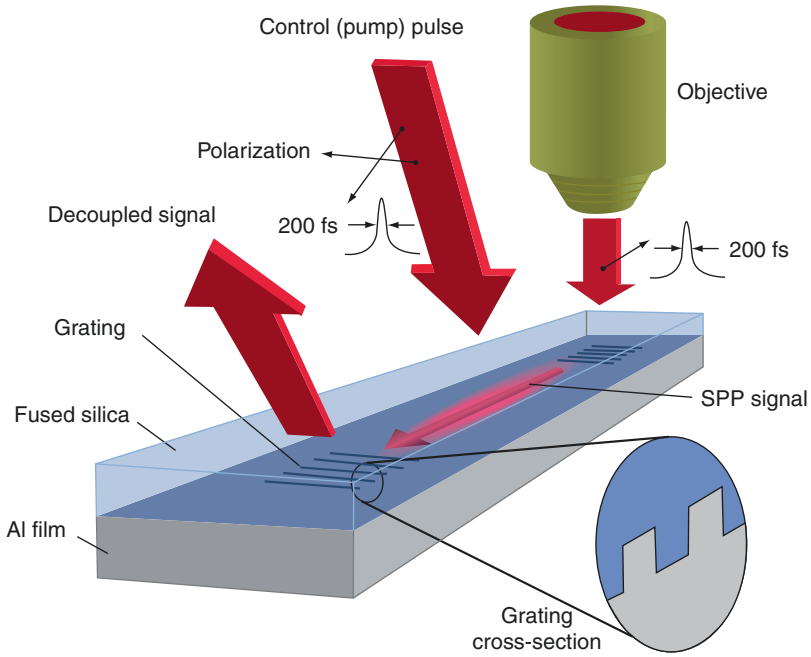
Due to the optical properties of dielectric layer, the switching may be achieved by external optical excitation through a phase transition in the layer section. For example, Krasavin et al. investigated the control of a dielectric waveguides containing a gallium section a few microns long (Fig. 10).

In such a waveguide, the SPP waves propagated at the interface between the metal film and silica substrate, through the gold and gallium sections, but the transmitted SPP wave was attenuated due to the mismatch of dielectric characteristics. The transition of the gallium layer from the semiconductor solid phase to the metallic liquid phase changes the local refractive index and provides a high-contrast modulation of the SPP propagation. This system could also be thermally driven. A fluence about 10 mJ/cm [2] was needed to induce the Ga-phase transition. Since this first work, many studies have been done in order to implement such system in different active plasmonic devices.

The use of photochromic (PC) molecules was developed by several groups. Among many interesting examples is the use of spyropyran, which are mixed with PMMA and deposited on top of the aluminum layer. Spiropyrans are PC molecules that can undergo a ring-opening reaction and can reversibly be switched between transparent and absorbing states using a free space optical pump. In the transparent (signal "on") state, the SPPs freely propagate through the molecular layer, and in the absorbing (signal "off") state, the SPPs are strongly attenuated (26 % of the original intensity). Upon exposure with 532 nm green light, the signals go back up to 90 %. Therefore the switch was very reproducible but slow, and the incomplete recovery of the signal was attributed to degradation.

With sub-microsecond or nanosecond response times at best, these techniques are likely to be too slow for future applications in data transport and processing fields. Few groups are working to develop some solutions to control or modulate a femtosecond optical frequency plasmon pulses by direct ultrafast optical excitation of the metal. MacDonald's group have discovered a nonlinear interaction between a propagating SPP and light that takes place in the skin layer of the metal surface along which the plasmon wave is propagating [24]. A femtosecond optical pulse incident on the metal surface disturbs the equilibrium in the energy–momentum distribution of electrons, thereby influencing SPP propagation along the surface (Fig. 11).

The transient effect of control pulse excitation on the propagation of the SPP signal was monitored by varying the time delay between the SPP excitation and optical pump pulses.



Active Plasmonic Devices, Fig. 11 Schematic representation of SPP modulation by ultrafast pulse in a Al-silica waveguide (Reprinted with permission from MacDonald et al. [24])

Two components were observed in the transient pump-probe interaction. The "fast" component was only seen where the pump polarization has a component parallel to the direction of SPP propagation with a magnitude of around 7.5 % for pump fluence of 10 mJ cm^{-2}. On the contrary, the "slow" component was a characteristic decay time of about 60 ps and reaches about 4 % of magnitude at the same fluence. The ultrafast femtosecond switching times and modest switching energy reported here open the gates to the exploration of what can ultimately be achieved in nonlinear plasmonics and active plasmonic switching, in particular for the purposes of high-bandwidth interconnects, plasmon–polariton modulation, all-plasmonic switching.

Spin Coupling SPP

A first proposal for a spin-plasmonic device was reported by Chau et al. [25]. The properties of a gigahertz magnetoplasmon optical modulator based on a bismuth-substituted yttrium iron garnet system are studied, showing the possibility of achieving a high-speed magneto-optic modulation of improved efficiency compared to bulk devices. This system is characterized by relevant magneto-optical properties (with Faraday rotations as high as 1°/μm). However, the high absorptions of this material in the visible range of wavelengths represent one of the major limitations in its application in devices. In 2007, the first experimental evidence of a spin-dependent transport in spin-plasmonic media was given by the same group, showing a magnetic manipulation of near-field-mediated light transport via the electron spin. They demonstrated that ferromagnetic particles coated with nonmagnetic metal nanolayers exhibit an enhanced magnetic field controlled attenuation of the electromagnetic field propagated through the sample. The mechanism is related to dynamic, electromagnetically induced electron spin accumulation in the non-magnet. Since 2007, many studies have been done based on core shell metal/magnetic material/metal or magnetic material/metal/magnetic material nanosandwichs system demonstrating a strong coupling between the plasmon resonance and the magneto-optics activity.

Future Directions for Research

Active plasmonic is still in its infancy, even though controlling LSPR and SPP by various

external stimuli has been widely successful. Many new devices are likely to be fabricated shortly. Strategies are similar, but some of them have only been used in order to control one kind of plasmonic device. Switching speed measurements in LSPR-based systems are still lacking, whereas SPP control does not yet use plastic electronic components. It can be anticipated that works on LSPR and SPP control will feed each other, and that active molecular plasmonic devices will lead to a wealth of new important results.

Cross-References

- Biosensors
- Gold Nanorods
- Nanoparticles
- Nanophotonic Structures for Biosensing
- Nanopolaritonics
- Nanostructures for Photonics
- Optical properties of Metal Nanoparticles
- Plasmon Resonance Energy Transfer Nanospectroscopy
- Surface Plasmon-Polariton-Based Detectors

References

1. Brongersma, M.L., Kik, P.G.: Surface Plasmon Nanophotonics. Springer, New York (2007)
2. Baia, M., Astilean, S., Iliescu, T.: New developments in SERS-active substrates. In: Baia, M., Astilean, S., Iliescu, T. (eds.) Raman and SERS Investigations of Pharmaceuticals, pp. 187–205. Springer, Berlin (2008)
3. Barnes, W.L., Dereux, A., Ebbesen, T.W.: Surface plasmon subwavelength optics. Nature **424**, 824–830 (2003)
4. Jain, P.K., Huang, X., El-Sayed, I.H., El-Sayed, M.A.: Noble metals on the nanoscale: optical and photothermal properties and some applications in imaging, sensing, biology, and medicine. Acc. Chem. Res. **41**, 1578–1586 (2008)
5. MacDonald, K.F., Zheludev, N.I.: Active plasmonics: current status. Laser Photonics Rev. **4**, 562–567 (2010)
6. Karg, M., Pastoriza-Santos, I., Perez-Juste, J., Hellweg, T., Liz-Marzan, L.M.: Nanorod-coated PNIPAM microgels: thermoresponsive optical properties. Small **3**, 1222–1229 (2007)
7. Ikeda, T.: Photomodulation of liquid crystal orientations for photonic applications. J. Mater. Chem. **13**, 2037–2057 (2003)
8. Malynych, S., Chumanov, G.: J. Am. Chem. Soc. **125**(10), 2896–2898 (2003)
9. Homola, J.: Surface Plasmon Resonance Based Sensors. Springer, Berlin (2006)
10. Zheng, Y.B., Yang, Y.W., Jensen, L., Fang, L., Juluri, B.K., Flood, A.H., Weiss, P.S., Stoddart, J.F., Huang, T.J.: Active molecular plasmonics: controlling plasmon resonances with molecular switches. Nano Lett. **9**, 819–825 (2009)
11. Haes, A.J., Zou, S., Zhao, J., Schatz, G.C., Van Duyne, R.P.: Localized surface plasmon resonance spectroscopy near molecular resonances. J. Am. Chem. Soc. **128**, 10905–10914 (2006)
12. Negre, C.F.A., Sánchez, C.G.: Effect of molecular adsorbates on the plasmon resonance of metallic nanoparticles. Chem. Phys. Lett. **494**, 255–259 (2010)
13. Zhao, J., Das, A., Schatz, G.C., Sligar, S.G., Van Duyne, R.P.: Resonance localized surface plasmon spectroscopy: sensing substrate and inhibitor binding to cytochrome P450. J. Phys. Chem. C **112**, 13084–13088 (2008)
14. Ung, T., Giersig, M., Dunstan, D., Mulvaney, P.: Spectroelectrochemistry of colloidal silver. Langmuir **13**, 1773–1782 (1997)
15. Wang, Z.C., Chumanov, G.: WO3 sol-gel modified Ag nanoparticle arrays for electrochemical modulation of surface plasmon resonance. Adv. Mater. **15**, 1285–1289 (2003)
16. Leroux, Y.R., Lacroix, J.C., Chane-Ching, K.I., Fave, C., Felidj, N., Levi, G., Aubard, J., Krenn, J.R., Hohenau, A.: Conducting polymer electrochemical switching as an easy means for designing active plasmonic devices. J. Am. Chem. Soc. **127**, 16022–16023 (2005)
17. Leroux, Y., et al.: ACS Nano **2**(4), 728–732 (2008)
18. Stockhausen, V., Martin, P., Ghilane, J., Leroux, Y., Randriamahazaka, H., Grand, J., Felidj, N., Lacroix, J. C.: Giant plasmon resonance shift using poly (3,4-ethylenedioxythiophene) electrochemical switching. J. Am. Chem. Soc. **132**, 10224–10226 (2010)
19. Sidhaye, D.S., Kashyap, S., Sastry, M., Hotha, S., Prasad, B.L.V.: Gold nanoparticle networks with photoresponsive interparticle spacings. Langmuir **21**, 7979–7984 (2005)
20. Du, G.X., Mori, T., Saito, S., Takahashi, M.: Shape-enhanced magneto-optical activity: degree of freedom for active plasmonics. Phys. Rev. B Condens. Matter **82**, 4 (2010)
21. Krasavin, A.V., Zheludev, N.I.: Active plasmonics: controlling signals in Au/Ga waveguide using nanoscale structural transformations. Appl. Phys. Lett. **84**(8), 1416–1418 (2004)
22. Nikolajsen, T., Leosson, K., Bozhevolnyi, S.I.: Surface plasmon polariton based modulators and switches operating at telecom wavelengths. Appl. Phys. Lett. **85**, 2 (2004)

23. Dionne, J.A., Diest, K., Sweatlock, L.A., Atwater, H. A.: PlasMOStor: a metal-oxide-Si field effect plasmonic modulator. Nano Lett. **9**, 897–902 (2009)
24. MacDonald, K.F., Samson, Z.L., Stockman, M.I., Zheludev, N.I.: Ultrafast active plasmonics. Nat. Photonics **3**, 55–58 (2009)
25. Chau, K.J., Irvine, S.E., Elezzabi, A.Y.: A gigahertz surface magneto-plasmon optical modulator. IEEE J. Quantum Electron. **40**, 571–579 (2004)

Adhesion

► Nanotribology

Adhesion in Insects

Jan-Henning Dirks
Department of New Materials and Biosystems, Max Planck Institute for Intelligent Systems, Stuttgart, Germany

Synonyms

Insect adhesion

Definition

Mechanisms of adhesion in insects.

The Fascination of Insect Adhesion

For hundreds of millions of years, insects have been inhabiting almost every part of the world, and different species have developed adaptations to very successfully survive in almost all ecosystems.

The ability to survive in ecosystems obviously depends on many different and complex biological and non-biological factors. One very important skill however is the ability to safely walk and stick to many different types of surfaces. Insects have brought this ability to almost perfection, and their feet have inspired authors and scientists for centuries.

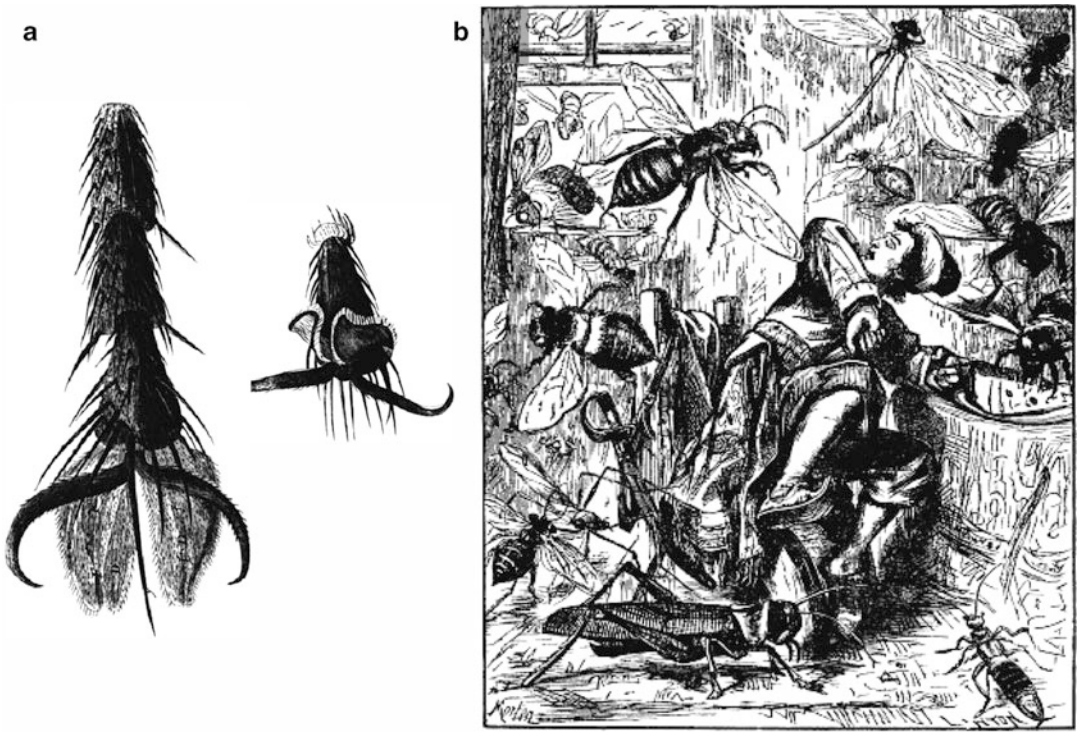
In the beginning of the seventeenth century, soon after the invention of the microscope, researchers were able to study the details of adhesive mechanisms for the first time. Quickly, a lively discussion started into how insects are able to stick to and walk on many kinds of surfaces. Many hypotheses about insect adhesion, from "gluten-filled sponges," the interlocking of fine hairs, suction cups to adhesive secretions, have been proposed over the last two centuries (see Fig. 1a and 1b) [1, 3–7].

In recent years more elaborate microscopes and better analytical tools have become available, and the interest of biologists and engineers in biological and biomimetic adhesives has increased [8–13]. Imaging techniques such as atomic force microscopy (AFM) and various different types of scanning electron microscopy (SEM) now allow biologists to analyze sensitive biological samples on a nanometer scale. Nevertheless, the detailed mechanisms allowing insects to safely adhere to surfaces are still not fully understood.

Adhesive Organs in Insects

Despite the insects' ecological importance and astounding diversity, insect adhesion is surprisingly still understudied, compared to the natural adhesives of a small number of animals such as geckos and mussels. Although at first glance the wide variety of insect adhesive organs appears very different, many of them are functionally quite similar, combining versatility with reliability and controllability.

In their natural environment the surfaces which climbing insects encounter can change fundamentally from one step to the next. Insect foot pads are however not only versatile and work on many surfaces, but can also produce remarkable high attachment forces with only moderate adhesive stresses. Some insects, such as the Asian weaver ant *Oecophylla smaragdina*, can carry loads of more than 100 times body weight when walking upside down on a smooth substrate (see Fig. 2a).

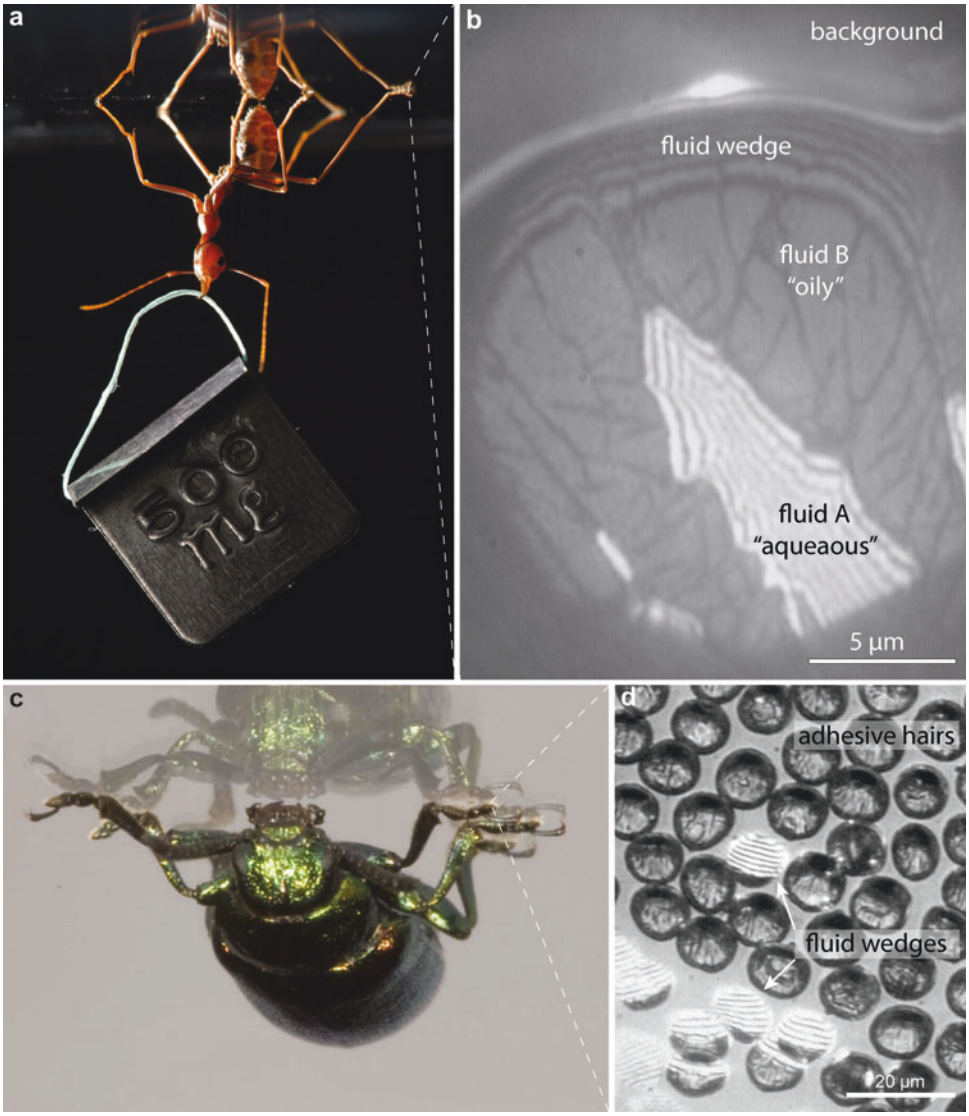


Adhesion in Insects, Fig. 1 (**a**) Probably one of the first drawings of the adhesive feet of a fly published by Robert Hooke in 1665. "The foot of a Fly [. . .] is of a most admirable and curious contrivance" (Figure adapted from [1], public domain). (**b**) When traveling the land of Brobdingnag, Captain Lemuel Gulliver is attacked by giant wasps on the breakfast table and notes "I could easily trace that viscous matter, which, our naturalists tell us, enables those creatures to walk with their feet upwards upon a ceiling" (Figure adapted from [2], public domain)

Although such high adhesive forces of many insects are fascinating, to really understand the principles of insect attachment forces, it is necessary to (i) measure the forces generated by a single adhesive pad, (ii) relate these forces to the respective contact area (to calculate the stresses), and (iii) take into account the chemical and mechanical properties of the surface (wettability, roughness). Biomechanical measurements of the attachment forces also require a precise real-time control of the normal force and, in case of dynamic measurements, also of the adhesive pad's sliding speed. A schematic drawing of a setup used to comprehensibly characterize the attachment forces of a single adhesive pad is shown in Fig. 3.

Although many modern artificial adhesives can outperform insect attachment forces, the feet of insects are still unmatched in respect to their rapid controllability. Insects use their dynamic adhesive organs to attach and detach their feet countless times during their lifetime. Looking at the fast locomotion of many insects, it becomes clear that their adhesive organs thus must be well adapted to rapid attachment and detachment on many different surfaces. Each of these steps also bears the risk of not only damaging but also contaminating the adhesive organ, which results in a loss of adhesion.

One of the most basic biological structures for adhering to a substrate is a pair of sickle-shaped claws at the end of each foot. Such claws can be found in a variety of different sizes and shapes in many vertebrates and most arthropods. Claws are used to grip and interlock on rough surfaces, where their function is determined by their stiffness and shape and in particular by the radius of their tip.

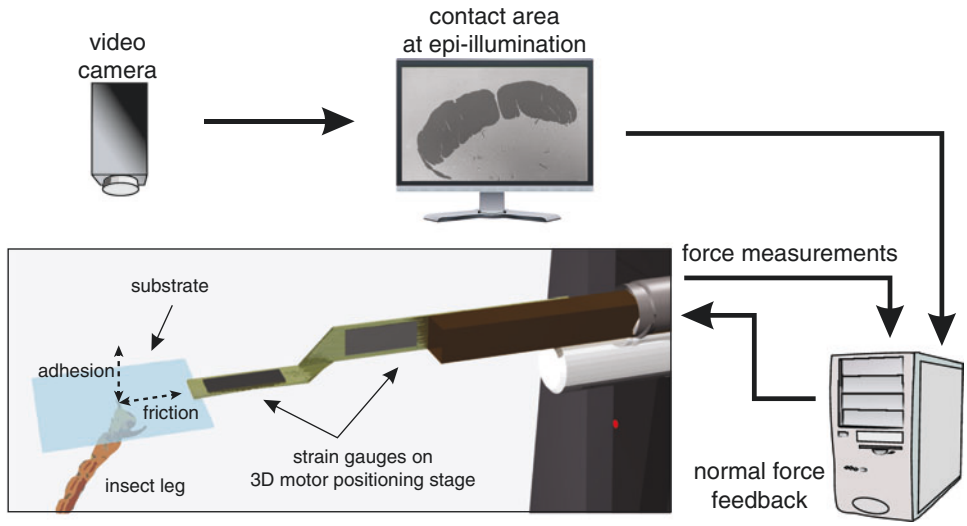


Adhesion in Insects, Fig. 2 (**a**) Asian weaver ant (*Oecophylla smaragdina*) holding more than 100 times its body weight upside down on a smooth surface. (**b**) IRM of adhesive contact zone of the ant's pad with the two-phasic adhesive fluid showing aqueous droplets (fluid A) within a hydrophobic continuous fluid (fluid B). (**c**) Leaf beetle (*Gastrophysa viridula*) and (**d**) IRM of discoidal adhesive setae in contact with glass. Interference fringes show hairs in partial contact. Setae appear darker on the side of the stalk (Insect photographs with kind permission by T. Endlein. Figure adapted with permission from own work [14])

However, if the surface asperities are smaller than the diameter of the claw tip and the claws cannot insert into the material, the adhesive pads will fail. Hence, to stick to smooth and stiff natural substrates, such as stones or leaves, insects and other arthropods have to use adhesive pads.

Hairy Versus Smooth Adhesive Pads

The adhesive organs of insects, like all parts of their exoskeleton, are almost entirely made of cuticle. Although cuticle is a material known for its variability, exoskeletal parts are mostly orders of magnitude stiffer than required by the



Adhesion in Insects, Fig. 3 Schematic illustration of an experimental setup for in vivo measurement of adhesion and shear stress (friction) on single adhesive pads of insects. A force transducer with attached surfaces can be moved in three dimensions using a computer-controlled translation stage. A feedback system controls normal forces in real time during friction measurements. The contact areas of the adhesive pads are imaged from above under reflected light using digital cameras synchronized with the force measurements. This allows the calculation of shear stress from force measurements (Figure adapted with permission from own work [15])

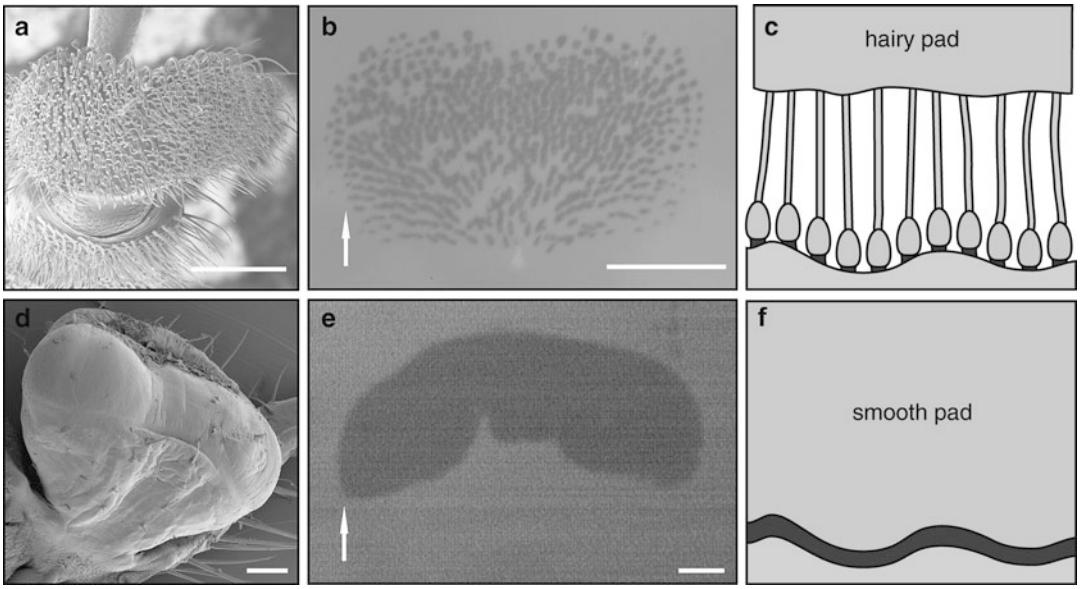
"Dahlquist criterion" (a Young's modulus of 100 kPa), which is usually considered as the upper limit for "tacky" materials.

To reduce the effective stiffness of the adhesive organ cuticle, two types of adhesive systems have evolved several times in insects: "hairy" (fibrillar) and "smooth" pads (see Fig. 4a and 4d). Both systems provide attachment to most rough and smooth surfaces by maximizing the contact area [16, 17].

The hairy adhesive pads on the feet of flies, beetles (see Fig. 2c), and spiders are densely covered with arrays of flexible *setae*, with each of the setae making close contact to the substrate separately (see Fig. 4b and 4c) [18, 19]. In most hairy pads these setae only end in a single terminal element; however, in some beetles they can branch toward the tip. Even within one tarsus or between the sexes of one species, these terminal elements can vary in shape and size [20]. Recently, it has been shown that in beetles the setae show a decreasing stiffness of the cuticle toward the tip of the setae [21].

Smooth pads of insects increase the contact area by adapting as a whole to the surface roughness (see Fig. 4e and 4f) [22, 23]. They can be found in many groups of insects such as ants, bees, stick insects, grasshoppers, true bugs, and cockroaches [24]. A soft cuticular sac (arolium) is located between the claws on the pretarsus of the insect leg and shows the presence of a specialized ultrastructure, which can deform and mold to rough surface profiles (see Fig. 4f). Within the arolium, cuticular rods are oriented almost perpendicular to the surface plane and branch out into finer fibers toward the pad surface. Again, this ultrastructure can be found in many insect orders and even in the toe pad epidermis of tree frogs. Some insect orders, such as Hymenoptera, Thysanoptera, or Mecoptera, also show additional adaptations, such as an unfoldable arolium.

In addition to the pre-tarsal arolium, some insect orders such as Blattodea, Heteroptera, Hymenoptera, Orthoptera, Phasmatodea, and Plecoptera have tarsal pads called euplantulae, which are specialized for fundamentally different functions.



Adhesion in Insects, Fig. 4 (**a**) Adhesive pad morphology of a "hairy" pad from a male dock beetle (*G. viridula* and **d**) and a "smooth" adhesive pad from an Indian stick insect (*C. morosus*). Epi-illumination can be used to visualize the contact area of the adhesive pads in contact with glass (**b**, **e**). *Arrows* indicate distal direction. Subfigures **c** and **f** show a schematic view of the close contact of the hairy and smooth pads. Note that in both cases the contact zone is mediated by an adhesive fluid (dark). On rough surfaces many single adhesive hairs adapt to the surface asperities and allow the adhesive pad to form close contact. The soft arolium of smooth adhesive pads adapts as a whole to the surface asperities on rough substrates (Figure adapted with permission from own work [14])

In stick insects (*Carausius morosus*), for example, arolia are only in contact with the surface when the insect is hanging upside down. When standing upright, the insects rested on their proximal euplantulae. Single-pad force measurements suggest that, strictly speaking, only arolia are "true" adhesive pads that stick to the substrate, while euplantulae are specialized "friction pads" [23].

The large number of insect groups with smooth pads, compared to the smaller number of groups with hairy pads, and the structural diversity of the smooth pads in different insect orders, led to the suggestion that smooth pads probably appeared earlier in insect evolution and represent a more "basic" evolutionary adhesive structure and have independently evolved several times [19]. "Hairy" structures however can also be found in many other biological adhesive pads such as lizards and spiders, indicating a general evolutionary "favorable" attachment design for surface attachment [8].

Insect Adhesive Fluid

Despite their different morphology, hairy and smooth adhesive pads of insects both secrete adhesive fluid into the contact zone (see Fig. 2b and 2d). Within the last three centuries, many scientific hypotheses about the role of the adhesive fluid were proposed and tested.

In 1833 Blackwall placed insects in evacuated air cylinders and showed that air pressure does not affect the insects' ability to adhere to smooth surfaces [3]. This experiment was initially thought to disprove the popular "suction cup" hypothesis; however, Blackwall also noticed that small quantities of condensed water vapor, oil or wheat, pulverized chalk, and gypsum prevented insects and spiders from sticking to clean glass surfaces. This observation leads him to the conclusion that ". . .these facts, far from corroborating the mechanical theory, appeared quite inexplicable, except on the supposition that an adhesive

secretion is emitted by the instruments employed in climbing." This appears to be one of the first experimental scientific approaches to understand the function of the insects' adhesive fluid.

Until now adhesive secretions have been found in all insect groups studied to date, including smooth pads of cockroaches, ants, and stick insects, as well as in hairy pads of flies, bugs, and beetles [14]. Adhesive fluids are also known from other organisms such as tree frogs and bats, and recent work has demonstrated that even the adhesive pads of spiders, mites, and other arachnids can secrete a fluid. This leaves geckos and anoles currently as the only known representatives of "dry" adhesive pads in nature.

Although the adhesive fluids secreted by the attachment organs of insects were discovered and described several hundred years ago, their chemical composition, mechanisms of secretion, and biological function are still not fully understood.

Composition and Secretion of Adhesive Fluid

For the smooth pads of stick insects, cockroaches, and ants, it has been shown that the adhesive fluid is a two-phasic microemulsion, consisting of a hydrophilic, volatile dispersive phase within a hydrophobic and evaporation-persistent continuous phase (see Fig. 2b). Chemical analyses of hydrophobic footprint droplets in smooth and hairy systems show the presence of long-chained hydrocarbons (mainly C_{22} to C_{29} n-alkanes and n-alkenes [25, 26], as well as more polar fatty acids, carbohydrates, and amino acids. As many of these substances are found in other parts of the insects' cuticle, it is still unclear whether and to what extent the adhesive fluid differs in composition from the general lipid cover of the cuticle. While a similarity between footprint chemicals and cuticular waxes was reported for beetles, differences were shown for locusts and flies [26, 27].

Deposited footprint droplets of stick insects and weaver ants are highly persistent against evaporation, remain present even after rinsing them with water, and show high refractive indices typical of hydrocarbons ($n_{fluid} = 1.47$, fluid B) [24]. When in contact with glass, droplets of the emulsion's disperse phase (fluid A) are noticeable as bright spots with a much lower refractive index. These small droplets rapidly evaporate when approaching the edge of the contact zone. Although the two-phasic nature of adhesive fluid has been found for several insect groups and mites with smooth adhesive pads, there is still no clear experimental evidence for the presence of two fluid phases in the hairy pads of insects and arachnids [28].

The secretion mechanism of the adhesive fluid is still unclear for both hairy and smooth pads. In smooth pads, it is so far only known that the secretion relies on simple physical principles, allowing for passive control of fluid volume within the contact zone. Footprints left by freely climbing insects are much smaller than those left during consecutive experimental press-downs, suggesting that insects can conserve secretion by minimizing contact area and by recovering fluid during detachment [29].

Biomechanical Function of the Adhesive Fluid: Self-Cleaning, Adhesion, and Friction

Typical cuticle secretions form an important protective barrier against evaporation and are used for intra- and interspecies communication. The main functions of the fluid secreted by adhesive pads however are self-cleaning of the pads and improvement of attachment forces [29].

Adhesion Forces

Insects climbing on a surface have to generate both adhesion (perpendicular to the substrate) and friction forces (parallel to the substrate). For simplicity, fluid-based adhesion is often modeled using two flat, undeformable surfaces mediated by a thin layer of fluid. The adhesion observed between the surfaces of such a model is the sum of three components: the fluid's surface tension, the Laplace pressure, and the *Stefan* adhesion:

$$F_{\text{adh}} = \underbrace{F_{\text{surface tension}} + F_{\text{Laplace pressure}}}_{\text{capillary forces}} + F_{\text{viscous forces}} \tag{1}$$

Using a few experimentally accessible parameters such as mediating fluid height, viscosity, contact

angle, etc., this model can be used to discuss and qualitatively predict a few characteristic "features" of fluid-based adhesion. For example, one can predict that an increasing fluid height h between adhesive pad and surface would result in decreasing capillary adhesive forces. As a consequence, insects and all organisms with fluid-mediated attachment organs (including tree frogs) should minimize the secretion of adhesive fluid into the contact area to increase capillary adhesion on smooth surfaces.

However, only very few natural surfaces are actually smooth. Experimental studies indeed show that the adhesive fluid actually plays a more important role in filling gaps between the pad and the surface, thus maximizing contact area and adhesion to rough substrates (see Fig. 3 and Refs. [18, 30]).

Besides the limitation to smooth surfaces, this simple adhesion model also bears several additional notable problems, in particular in light of the stiffness and deformability of the adhesive pads [31]. For example, at certain ratios of adhesive pad size and stiffness, the Young's modulus of the adhesive pad can play an important role in determining the capillary adhesion. In these models a softer adhesive pad or substrate with a lower elasticity deforms more easily at a given external force, which results in a larger contact area. This larger contact area again increases the radius of the mediating liquid's meniscus, which then increases the capillary force.

Friction Forces

In a fluid-mediated system with a continuous Newtonian fluid film, the friction forces between the substrates are often described using two basic principles: the surface tension of the mediating fluid and the laws of *hydrodynamic* lubrication.

However, the predicted contribution of surface tension toward the generation of friction forces is several orders of magnitude smaller than the shear stresses measured in smooth adhesive pads and thus is unlikely to explain the high friction forces generated by the adhesive pads of insects alone [18, 32, 33].

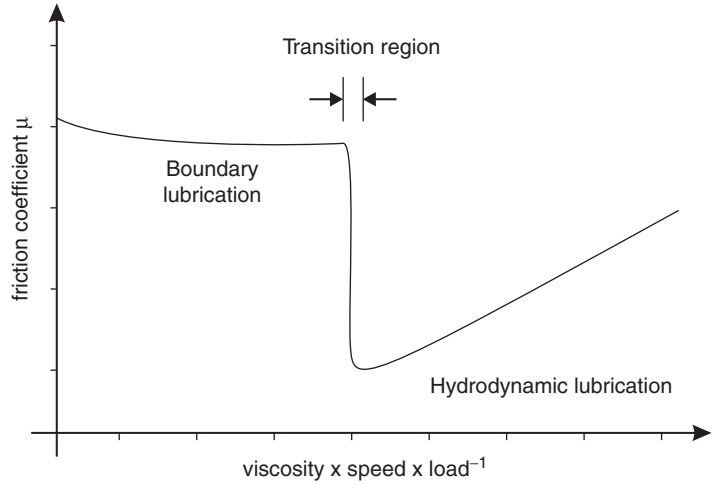
Similar to the *Stefan* adhesion, the "hydrodynamic lubrication" friction term takes into account the *viscosity* of the mediating fluid layer. Two parallel smooth surfaces with a distance of h sliding at a velocity v relative to each other generate the friction force

$$F_F = \eta_{\text{eff}} A \frac{v}{h} \tag{2}$$

where η_{eff} is the effective viscosity of the mediating fluid layer and A the size of the contact area. Again, similar to the time-dependent *Stefan* adhesion, this term shows that a simple fluid-mediated system at rest should not be able to generate any static friction. Also, hydrodynamic friction forces generated by a continuous fluid film of 90 nm height would be one order of magnitude smaller than the shear stresses observed in adhesive organs [24].

Thus, to date it is still not entirely clear how insects with a fluid-mediated adhesive pad generate the observed high friction forces, preventing them from sliding down smooth vertical substrates. Experimental results suggest that thixotropic (shear thinning) non-Newtonian properties of the adhesive emulsion could explain the presence of high static friction forces [15, 30].

A second, yet to be tested, hypothesis to explain the observed high friction forces in the adhesive pads of insect is the formation of local "dry" rubberlike direct contacts between the adhesive pad and the substrate. The classic *hydrodynamic* model only describes the friction observed with relative "thick" layers of lubricant ($\geq$0.5 μm), where neither the specific surface properties (roughness, surface energy) nor *Amonton's* law is involved. Friction forces generated by fluid layers thinner than five to ten molecular layers are usually modeled using the more complex *boundary lubrication* theory, where a decreasing film height and increasing number and area of direct contacts between the two substrates actually *increase* the friction coefficient. The range between the hydrodynamic model and the boundary lubrication model, where the mediating fluid layer is still lubricating the contact, however, allows a weak interaction between the surfaces and is mostly referred to as *elasto-hydrodynamic lubrication* (see Fig. 5).



Adhesion in Insects, Fig. 5 Schematic illustration showing regions of hydrodynamic and boundary lubrication of two fluid-mediated smooth surfaces and the resulting friction forces. With decreasing sliding velocity and increasing load, the mediating fluid film becomes thinner, allowing regions of direct contact between the two surfaces. Within the transition region, the friction coefficient rapidly increases, usually to a value smaller than unlubricated surfaces. This region is also referred to as *elasto-hydrodynamic* lubrication. It is not yet fully understood in which of the two "regions" the adhesive organs of insects operate

Within this range, areas with "dry" solid-like contact can form as a result of local "dewetting" of the adhesive fluid. Indeed, a similar mechanism has been proposed for the fluid-mediated adhesive toes found in tree frogs, where local "dewetting" might be responsible for the high friction forces. However, so far there is no experimental evidence for the occurrence of dewetting or direct contacts between adhesive pad and surface in smooth or hairy adhesive organs of insects.

Conclusion

Even after many years of research, several of the fundamental physical properties enabling insects to safely adhere to surfaces are still not fully understood. Simple fluid-mediated models based on capillary and viscosity are still often used to estimate the attachment forces generated by insect feet. However, newer experimental data and more comprehensive tribological models suggest that many aspects of these classic "microscopic" attachment models fall short in fully explaining the forces generated by the fluid-mediated adhesive pads of insects. To fully understand the physical principles, one thus has to extend these models to incorporate non-Newtonian properties of the adhesive fluid [15, 30] and also take into account possible more complex nano-tribological models, with boundary lubrication or other mechanisms resulting in points of "dry" contact between the pad and the surface.

Further fundamental experimental work, in particular a more accurate in vivo measurement of the mediating fluid's height and high-resolution single-leg force measurements on smooth substrates with well-known physical properties, is required to answer these questions and help to solve a century-old scientific question.

Cross-References

▶ Adhesion
▶ Bioadhesion
▶ Biomimetics

- Cuticle
- Exoskeleton
- Fibrillar Adhesion
- Friction
- Gecko Adhesion
- Insects
- Microscale Fluid Mechanics
- Reversible Adhesion

References

1. Hooke, R.: Micrographia. Oxford Press, London (1665). Project Gutenberg, Public Domain
2. Swift, J.: Travels into Several Remote Nations of the World, in Four Parts. By Lemuel Gulliver, First a Surgeon, and then a Captain of Several Ships. Benjamin Motte, London (1726). Project Gutenberg, Public Domain
3. Blackwall, J.: Letter addressed to the secretary. Extracts from the Minute-Book of the Linnean Society **21**:767 (1833)
4. Burmeister, H.: Handbuch der Entomologie. G. Reimer, Berlin (1832)
5. Dewitz, H.: Über die Fortbewegung der Thiere an senkrechten Flächen vermittels eines Secretes. Zool. Anz. **7**, 400–405 (1884)
6. Power, H.: Experimental Philosophy. Martin and Allestry, London (1664)
7. West, T.: The foot of the fly, its structure and action elucidated by comparison with the feet of other insects. Trans. Linn. Soc. **23**, 393–421 (1862)
8. Federle, W.: Why are so many adhesive pads hairy? J. Exp. Biol. **209**(Pt 14), 2611–2621 (2006). doi:10.1242/jeb.02323
9. Geim, A., Dubonos, S., Grigorieva, I., Novoselov, K., Zhukov, A., Shapoval, S.: Microfabricated adhesive mimicking gecko foot-hair. Nat. Mater. **2**(7), 461–463 (2003)
10. Kovalev, A.E., Varenberg, M., Gorb, S.N.: Wet versus dry adhesion of biomimetic mushroom-shaped microstructures. Soft Matter **8**(29), 7560 (2012). doi:10.1039/c2sm25431j
11. Seidl, T., Vidoni, R.: Adhesion to flat surfaces: from spiders to stickers. In: Spider Ecophysiology, pp. 463–473. Springer, Heidelberg (2013)
12. Waite, J.H., Andersen, N., Jewhurst, S., Sun, C.: Mussel adhesion: finding the tricks worth mimicking. J. Adhes. **81**(3–4), 297–317 (2005)
13. Zhou, M., Pesika, N., Zeng, H., Tian, Y., Israelachvili, J.: Recent advances in gecko adhesion and friction mechanisms and development of gecko-inspired dry adhesive surfaces. Friction **1**(2), 114–129 (2013). doi:10.1007/s40544-013-0011-5
14. Dirks, J.H., Federle, W.: Fluid-based adhesion in insects – principles and challenges. Soft Matter **7**, 11047–11053 (2011). doi:10.1039/C1SM06269G
15. Dirks, J.H., Clemente, C.J., Federle, W.: Insect tricks: two-phasic foot pad secretion prevents slipping. J. R. Soc. Interface **7**(45), 587–593 (2010). doi:10.1098/rsif.2009.0308
16. Beutel, R.G., Gorb, S.N.: Ultrastructure of attachment specializations of hexapods, (arthropoda): evolutionary patterns inferred from a revised ordinal phylogeny. J. Zool. Syst. Evol. Res. **39**(4), 177–207 (2001)
17. Gorb, S.N.: Structural design and biomechanics of friction-based re-leasable attachment devices in insects. Integr. Comp. Biol. **42**(6), 1127–1139 (2002)
18. Bullock, J.M.R., Drechsler, P., Federle, W.: Comparison of smooth and hairy attachment pads in insects: friction, adhesion and mechanisms for direction-dependence. J. Exp. Biol. **211**, 3333–3343 (2008)
19. Gorb, S. N.: Smooth attachment devices in insects: Functional morphology and biomechanics. Adv. Insect Physiol. **34** (2008). Doi: 10.1016/S0065-2806(07)34002-2
20. Bullock, J.M.R., Federle, W.: Division of labour and sex differences between fibrillar, tarsal adhesive pads in beetles: effective elastic modulus and attachment performance. J. Exp. Biol. **212**, 1876–1888 (2009)
21. Peisker, H., Michels, J., Gorb, S. N.: Evidence for a material gradient in the adhesive tarsal setae of the ladybird beetle coccinella septempunctata. Nat. Commun. **4**, 1608–7 (2013). doi: 10.1038/ncomms2576
22. Clemente, C., Dirks, J.H., Barbero, D., Steiner, U., Federle, W.: Friction ridges in cockroach climbing pads: anisotropy of shear stress measured on transparent, microstructured substrates. J. Comp. Physiol. A **195**(9), 805–814 (2009). doi:10.1007/s00359-009-0457-0
23. Labonte, D., Federle, W.: Functionally different pads on the same foot allow control of attachment: Stick insects have load-sensitive "heel" pads for friction and shear-sensitive "toe" pads for adhesion. PLOS ONE **8**(12), e81, 943 (2013). doi:10.1371/journal.pone.0081943.s002
24. Federle, W., Riehle, M., Curtis, A.S.G., Full, R.J.: An integrative study of insect adhesion: mechanics and wet adhesion of pretarsal pads in ants. Integr. Comp. Biol. **42**(6), 1100–1106 (2002)
25. Attygalle, A.B.., Aneshansley, D.J., Meinwald, J., Eisner, T.: Defense by foot adhesion in a chrysomelid beetle (*Hemisphaerota cyanea*): characterization of the adhesive oil. Zool. Anal. Complex Syst. **103** (1–2), 1–6 (2000)
26. Vötsch, W., Nicholson, G., Müller, R., Stierhof, Y., Gorb, S.N., Schwarz, U.: Chemical composition of the attachment pad secretion of the locust *Locusta migratoria*. Insect Biochem. Mol. Biol. **32**(12), 1605–1613 (2002)

27. Bauchhenss, E.: Die Pulvillen von *Calliphora erythrocephala* (Diptera, Brachycera) als Adhäsionsorgane. PhD thesis, University of Munich (1979)
28. Peattie, A. M., Dirks, J. H., Henriques, S., Federle, W.: Arachnids secrete a fluid over their adhesive pads. PLOS ONE **6**(5), e20, 485 (2011). doi: 10.1371/journal.pone.0020485
29. Dirks, J.H., Federle, W.: Mechanisms of fluid production in smooth adhesive pads of insects. J. R. Soc. Interface **8**(60), 952–960 (2011). doi:10.1098/rsif.2010.0575
30. Drechsler, P., Federle, W.: Biomechanics of smooth adhesive pads in insects: influence of tarsal secretion on attachment performance. J. Comp. Physiol. A **192**(11), 1213–1222 (2006). doi:10.1007/s00359-006-0150-5
31. Dirks, J.H.: Physical principles of fluid-mediated insect attachment – shouldn't insects slip? Beilstein J. Nanotechnol. **5**, 1160–1166 (2014). doi:10.3762/bjnano.5.127
32. Clemente, C., Federle, W.: Pushing versus pulling: division of labour between tarsal attachment pads in cockroaches. Proc. R. Soc. B **275**, 1329–1336 (2008)
33. Federle, W., Baumgartner, W., Hölldobler, B.: Biomechanics of ant adhesive pads: frictional forces are rate- and temperature-dependent. J. Exp. Biol. **207**, 67–74 (2004)

Adsorbate Adhesion

► Disjoining Pressure and Capillary Adhesion

Aerosol-Assisted Chemical Vapor Deposition (AACVD)

► Chemical Vapor Deposition (CVD)

AFM

► Robot-Based Automation on the Nanoscale

AFM Force Sensors

► AFM Probes

AFM in Liquids

Bart W. Hoogenboom
London Centre for Nanotechnology and Department of Physics and Astronomy, University College London, London, UK

Synonyms

Atomic force microscopy in liquids; Scanning force microscopy in liquids

Definition

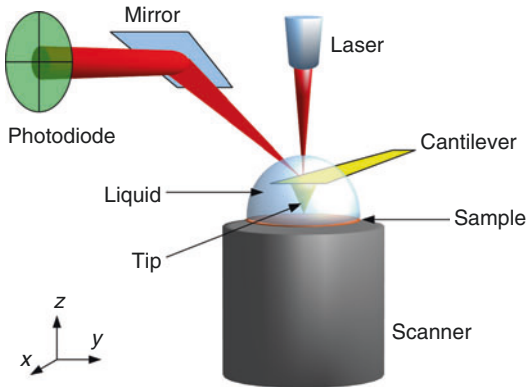
Atomic force microscopy (AFM) in liquids is the application of AFM in liquid environment, i.e., in which both the surface under investigation and the scanning probe are immersed in liquid.

Overview and Definitions

AFM is a microscopy technique that can provide three-dimensional images of virtually any surface at nanometer-scale resolution. It relies on the force between a sharp probe and the surface, which is detected while scanning the probe over the sample. Unlike many other microscopy techniques at such a resolution, it can readily be applied in liquid environment.

An atomic force microscope consists of a sharp probe ("tip") mounted on a microfabricated cantilever beam and a mechanism ("scanner") to scan the tip over the surface at subnanometer resolution [1], see Fig. 1. Typically, an optical detection scheme is used to detect the deflection of the cantilever. Via the spring constant of the cantilever, the cantilever deflection can be translated to a force between tip and sample. For a rectangular lever, the spring constant is given by

$$k = \frac{Et^3 w}{4l^3}, \tag{1}$$



AFM in Liquids, Fig. 1 Schematic of AFM in liquid. A sharp tip is scanned across the sample surface while the tip-sample interaction is monitored via the deflection of the cantilever to which the tip is attached. The sample is mounted on a (usually piezoelectric) scanner for three-dimensional positioning with sub-nanometer accuracy. The sample, the tip, and the cantilever are immersed in liquid. The bending of the cantilever is usually detected via a laser beam deflected on a position-sensitive detector (4-quadrant photodiode)

where E is the Young's modulus of the cantilever material (typically silicon or siliconnitride), and t, w, and l are its thickness, width, and length, respectively, which are in the micron range. Most cantilevers for AFM have spring constants between 0.01 and 100 N/m. In its most basic mode of operation, the deflection of the cantilever (and thus the tip-sample force) is kept constant by adjusting the vertical position of the sample with respect to the tip.

In most instruments, the cantilever deflection is recorded via the position of a laser beam that is deflected from the cantilever. Images of the surfaces are acquired by line-by-line scanning the surface, and tracing its surface contours on a false-color scale. For AFM in liquid, the surface and the cantilever are immersed in liquid. Typically, the minimum liquid volume is some tens of microliters, which can be either contained in a closed liquid cell, or in the shape of a droplet that is formed by capillary forces between the sample surface and the cantilever holder. In principle, the liquid can be of arbitrary nature, though the optical deflection detection in most microscopes limit their use to liquids that are transparent for (near infra-) red light.

AFM in Liquids, Table 1 Different modes of operation in AFM, with the parameter that is used to control the tip-sample distance while scanning over the surface

Mode of operation	Other common names	Control parameter
Static	Contact mode	Static deflection
Amplitude modulation (AM)	Tapping mode	Oscillation amplitude
	Intermittent contact, AC mode	
Frequency modulation (FM)	Noncontact (NC) AFM	Resonance frequency
Force spectroscopy	Force volume, quantitative imaging, peakforce tapping	Maximum force

There are many possible imaging modes of AFM and their names are rather confusing. They are divided into static mode and a variety of dynamic modes in which the cantilever is oscillating [2]. The latter are of particular importance when the lateral (drag) forces on the sample need to be minimized, e.g., for imaging DNA molecules that are loosely bound to a flat surface. The most common modes of operation in liquid are static or contact mode and amplitude modulation or tapping mode. In contact mode, the bending or deflection of the cantilever is used to detect the tip-sample force. In tapping mode, the cantilever is vertically oscillated above the sample, and the tip-sample interactions detected via the change (usually decrease) of amplitude of the cantilever oscillation. In variations along the same theme, the cantilever may be oscillated and tip-sample interactions are detected via changes in phase or resonance frequency of the cantilever [3]. Thus, the amplitude, phase, or frequency of the cantilever oscillation can be the control parameter that is used to keep the (average) tip-sample interaction constant, via continuous adjustment of the relative position of the tip with respect to the sample (or vice versa). An overview of the most relevant AFM modes in liquids is given in Table 1. Confusion may arise because the commonly used terms contact, intermittent-contact, and noncontact often – but not always – refer to a mode of operation (as referred to in Table 1), and not necessarily to the actual contact (or lack

of it) between the tip and the sample. As a particular example, FM ("noncontact") AFM in liquids is usually performed in the range of repulsive tip-sample interactions, with the tip in actual intermittent contact with the sample. In addition, if the amplitude is used to trace and control the tip-sample distance (i.e., tapping mode), the phase of the cantilever can be recorded simultaneously for phase imaging, which provides an additional means to measure tip-sample interactions. Or, in FM AFM, a frequency shift is used to trace elastic tip-sample interactions and control the tip-sample distance, while the oscillation amplitude can be kept constant by adjusting the cantilever drive signal. This cantilever drive signal then provides a simultaneous measure of the dissipative tip-sample interactions. In addition, the tip-sample force may be continuously probed by acquiring force-distance curves (force spectroscopy). Provided that these curves are acquired sufficiently fast, the maximum force in such a force curves (at smallest tip-sample separation) can be used as the control parameter for imaging the surface topography [4].

Physical and Chemical Principles

The spatial resolution of AFM critically depends on the sharpness of the tip and on the range of the tip-sample interactions. Ideally, only the very (nanometer-scale) end of the tip interacts with the surface, thus avoiding convolution effects due to interactions between the sample and the (micron-scale) bulk of the tip. The presence of long-range tip-sample forces is therefore decremental to the spatial resolution. On the other hand, the shorter the range of the tip-sample interaction, the more likely the tip is to enter in hard contact with the sample, which may damage the tip and/or the sample.

By immersing the tip and sample in liquid, it is not only possible to access scientifically and technologically interesting solid-liquid interfaces (including biological samples), but also to tune the tip-sample interactions by varying the ingredients of the liquid. In particular, the electrostatic sample interaction between the tip and a flat sample can be approximated by [5]:

$$F_{el} = g_\kappa \sigma_t \sigma_s e^{-\kappa z} + g_{2\kappa}\left(\sigma_t^2 + \sigma_s^2\right) e^{-2\kappa z}, \quad (2)$$

where g_κ and $g_{2\kappa}$ are constants that depend on the tip geometry and the dielectric properties of the medium, σ_t and σ_s are the surface charges of the tip and the sample, respectively, z is the tip-sample distance, and κ^{-1} the Debye screening length [6],

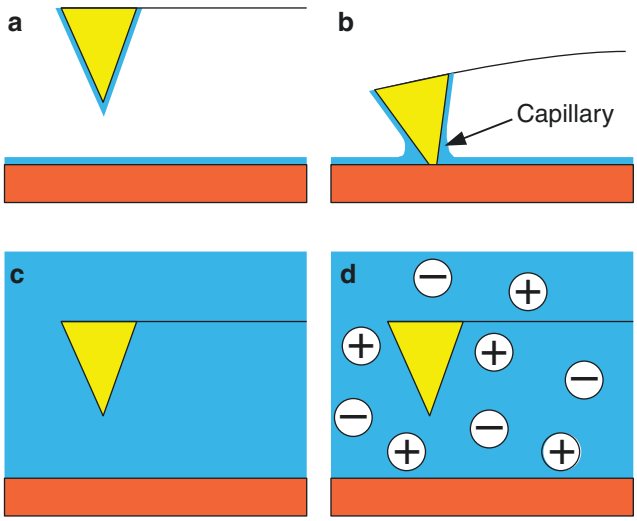
$$\kappa^{-1} = \sqrt{\frac{\varepsilon_0 \varepsilon_b k_B T}{2e^2 I}}. \quad (3)$$

In the latter, ε_0 is the permittivity of vacuum (8.854×10^{-23} F/m), ε_b the permittivity of the bulk solution (or the dielectric medium), k_B is the Boltzmann's constant (1.38×10^{-23} J/K), and T is the temperature in Kelvin. The parameter I is the ionic strength of the medium,

$$I = \frac{1}{2}\sum_i z_i^2 c_i. \quad (4)$$

where z_i is the valence of ion i and c_i is its concentration. The Debye length is historically written as an inverse length (i.e., κ has units of m^{-1}). In typical biological buffers, the Debye length is between 1 and 10 nm.

In a chemically reactive medium, σ_t and σ_s are generally not zero. For typical tip terminations such as silicon oxide or silicon nitride in water, $\sigma_t < 0$ at neutral pH. As can be seen from Eq. 2, the strength of the electrostatic tip-sample interaction depends on σ_t and σ_s. These can be changed by adding or removing solutes that react with or bind to the tip and sample surfaces. In aqueous solutions, this is most readily achieved by changing the pH (i.e., adding H_3O^+ or OH^-). Moreover, the dielectric properties – and in particular the ionic strength (Eq. 4) – of the medium can be tuned to control the range of the electrostatic tip-sample interaction. This provides a significant advantage over AFM in gaseous mediums and vacuum environment, and is one of the reasons why the first true atomic resolution (see below)



AFM in Liquids, Fig. 2 (**a**) Capillary forces for AFM in air arise due to the thin water layers that are usually present at both tip and sample under ambient conditions. (**b**) At small tip-sample distances, they pull the tip toward the surface to create a larger contact area and thus limit the lateral resolution. (**c**) In liquid environment, these capillary forces are absent. (**d**) Moreover, long-range electrostatic forces can be screened by ions in the liquid

AFM images were obtained, in contact mode, in aqueous environment rather than in vacuum or air [7].

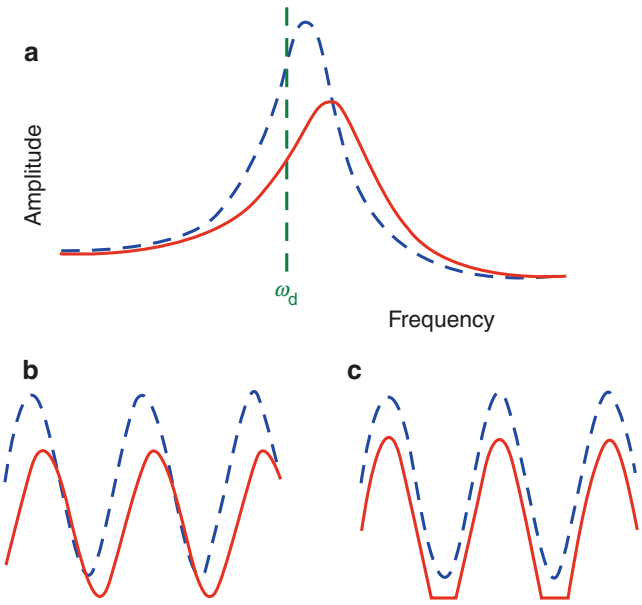
Compared to AFM in air, AFM in liquids has the additional advantage of preventing the capillary forces that arise from the humid coverage of both sample and tip under ambient conditions, as illustrated in Fig. 2. The capillary forces create a strong pull on the tip toward the sample at distances of many nanometers. On close approach, this results in a tip-sample contact area that covers many atoms at the surface, thus prohibiting high resolution. Under such conditions, AFM is likely to result in one of its most extensively documented artifacts: The apparent "atomic resolution" that results from the periodic interaction averaged over many atoms of tip and surface. This represents the atomic periodicity of the surface rather than individual atomic-scale features such as atomic defects (which are often used to demonstrate "true atomic resolution").

Apart from electrostatic and capillary forces, AFM in liquids is also (and not exclusively) sensitive to atomic bonding forces and hard-core repulsion, van der Waals interactions, dissolution/solvation forces, and general dissipative interactions.

The latter are particularly important in the dynamic modes of AFM operation. To illustrate this, it is helpful to depict the cantilever resonance for different elastic and dissipative interactions, see Fig. 3.

A repulsive (elastic) tip-sample interaction increases the effective spring constant of the cantilever and thus increases its resonance frequency, shifting the whole resonance curve to the right. Any dissipative interaction reduces the sharpness of the resonance curve and the maximum of the curve. The quality factor Q is the most common measure for this sharpness, defined as $\omega_0/\Delta\omega$, where ω_0 is the natural frequency of the resonator and $\Delta\omega$ the width of the resonance curve at $1/\sqrt{2}\times$ its maximum amplitude. In tapping mode, the driving frequency is usually set just below the resonance frequency, such that both elastic and dissipative interactions cause the amplitude to decrease when the tip comes into contact with the sample. In FM AFM, the shift in the resonance frequency provides a direct measure of the elastic interaction, and the conversion to a force is straightforward when the amplitude of the oscillation is (kept) constant [8]. The dissipated energy can be deduced from the energy that is required to maintain a fixed amplitude. Thus, tapping mode is sensitive to a mixture of elastic and dissipative tip-sample interactions, whereas FM AFM can track elastic and dissipative interactions separately (and simultaneously).

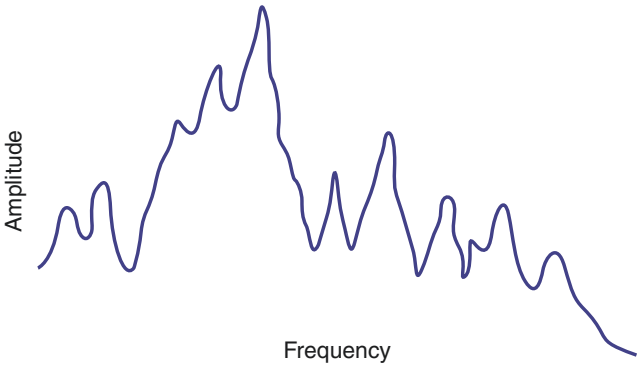
This analysis relies on the cantilever behaving as a simple harmonic oscillator, which is



AFM in Liquids, Fig. 3 (**a**) Schematic resonance curve of a freely oscillating cantilever (*dashed, blue curve*), based on a model of a simple harmonic oscillator. In the presence of a (repulsive) elastic tip-sample force as well as dissipative tip-sample interactions, the resonance shifts to higher frequencies, broadens and decreases in amplitude (*red curve*). At a fixed driving frequency ω_d, such as in tapping mode, the effects of elastic or dissipative interactions will cause the amplitude to decrease or increase, depending on the choice of ω_d. (**b**) Sinusoidal oscillation of a cantilever above the surface (*dashed, blue*). Close to the sample, the oscillation amplitude decreases and changes phase (*red*). (**c**) For heavily damped cantilever oscillations ($Q \sim 1$), however, the harmonic-oscillator analysis is not valid any more. The tip-sample interaction will only affect the bottom part of the oscillation, where the tip is closest to the sample

appropriate for $Q \gg 1$, typical for operation in vacuum $(Q \gtrsim 1000)$ and air $(Q \gtrsim 100)$. However, in liquid, viscous damping reduces Q to $\lesssim 10$. In particular, for softer cantilevers $k \lesssim 1\mathrm{N/m}$, $Q \approx 1$, and the contact between tip and sample will prevent the cantilever from following a harmonic, sinusoidal oscillation. The cantilever oscillation will simply be topped off at the bottom of the oscillation, at the closest tip-sample approach. In this case, significant information on the tip-sample interaction will be contained in higher harmonics of the oscillation. The interpretation of cantilever oscillation in terms of tip-sample forces thus becomes considerably more complicated.

The low quality factor has another undesirable side effect for AFM operation that involves oscillating the cantilever. Cantilevers are usually driven by a small piezoelectric element that drives not only the cantilever itself, but also its (macroscopic) support chip and often the whole cantilever holder. In vacuum and air, the high Q singles out the cantilever resonance from other mechanical resonances in the instrument. In liquid, this is not the case anymore, and a typical excitation spectrum contains a so-called forest of peaks, i.e., a convolution of the broad cantilever resonance with many sharp(er) mechanical resonances from the microscope, as illustrated in Fig. 4. Again, this complicates the interpretation of the cantilever behavior in terms of a simple harmonic oscillator. It also makes the choice of optimum driving frequency much less straightforward (usually, this is done by trial and error). Moreover, since this excitation spectrum depends on the macroscopic geometry of the fluid cell, it is much more dependent on drift than the cantilever alone. For these reasons, there are a number of alternative methods to only drive the microscopic cantilever (and no macroscopic parts), such as



AFM in Liquids, Fig. 4 Excitation spectrum such as can be obtained for a cantilever in water that is actuated by piezoelectrically driving the cantilever holder rather than the cantilever alone. Note the contrast with the ideal resonance shape in Fig. 3a

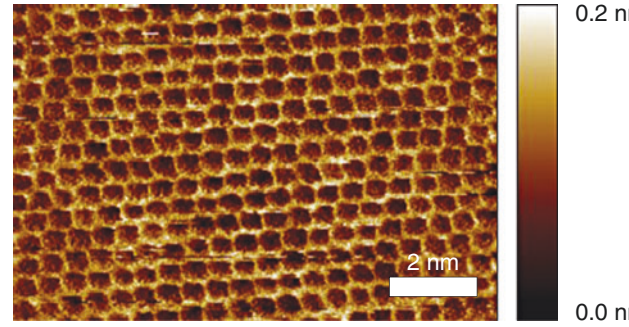
magnetic and optical actuation. In the former, the cantilever is coated with a magnetic material and driven by an AC magnetic field; in the latter, an AC-modulated actuation laser locally heats the metal-coated surface of cantilever, which acts as a bimetal and thus transduces laser power to cantilever deflection. These and other methods have the disadvantage of adding complexity to the instrument and eventually to the microfabricated cantilever itself.

These concerns can partly be addressed by so-called Q-control or self-oscillatory methods [9], in which the signal from the cantilever oscillation itself is phase-shifted, amplified, and added to the actuation signal. This positive feedback is very familiar in electronic oscillatory circuits. It causes any change to the oscillation to ring for a longer time, stretching its effect over many oscillations. As a result the effective Q of the cantilever oscillation is enhanced, and the cantilever is forced to follow a perfectly harmonic, sinusoidal oscillation. As for problems related to the forest of peaks in nonideal cantilever excitation, these are not resolved by Q control, since the positive feedback does not distinguish between the cantilever resonance and other mechanical resonances. Q control or self-oscillatory methods can be applied to each mode of operation (see Table 1), but for tapping mode has the side effect of slowing down the measurement: Changes in amplitude occur at a timescale of $\sim Q/\omega_0$, unlike changes in phase or resonance frequency, which are instantaneous. In FM AFM, the cantilever can even entirely be driven by its phase-shifted and amplified thermal noise.

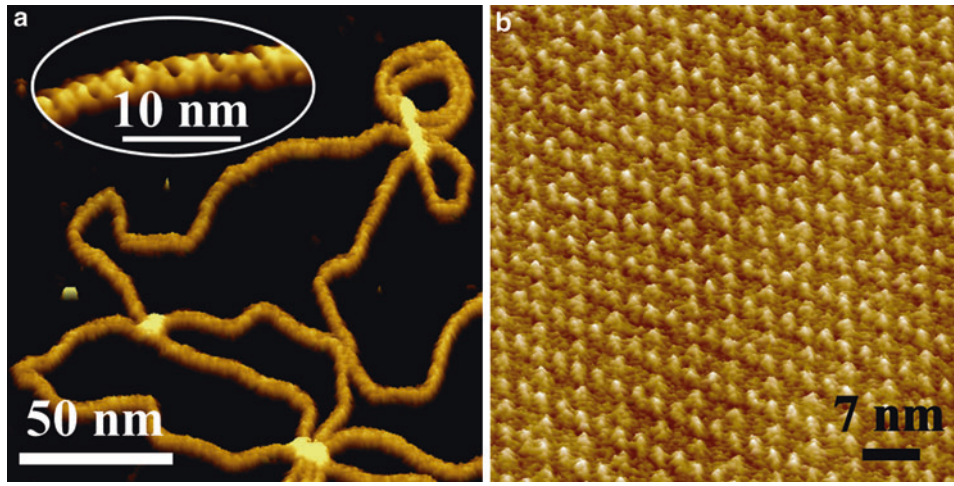
Applications

AFM has been used to probe a vast amount of interfaces between (hard and soft) condensed matter and liquids. It is of particular importance for the study of surfaces that are instable or show distinctly different behavior when taken out of the liquid, and for the study of dynamic processes that depend on an exchange of material (ions, macromolecules) between the surface and the liquid. AFM in liquids has been important for the study of, among others, crystal growth/dissolution, polymer science, and the behavior of liquids in confined geometries. The vast majority of its applications, however, lie in the life sciences [10]: For most biomolecular structures, aqueous solutions represent their natural environment. Their structures and functions are strongly dependent on the presence of water, a number of ions, and other macromolecules. The structure and function of biological samples can thus be studied at (sub)nanometer resolution while they are still "alive," and for varying liquid contents. The following represents some illustrative examples of the use of AFM in aqueous solutions on easily accessible samples, with a strong focus on biological applications.

Up to now, highest resolution AFM images in liquid have been obtained on atomically flat and inorganic surfaces, with calcite and muscovite mica as most common examples. On such surfaces, the lateral resolution can be less than an Ångström and the vertical resolution a few picometers. Though initially such resolution was only obtained in contact mode [7], more recently FM AFM has become popular as a method for atomic-resolution



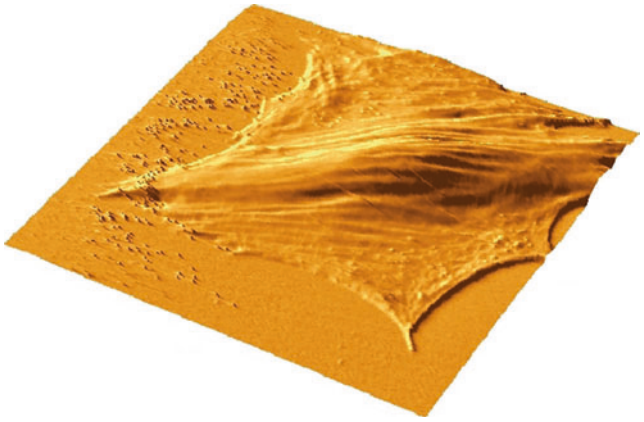
AFM in Liquids, Fig. 5 Atomic-resolution image of mica in (aquaeous) buffer solution, obtained by FM AFM (Reproduced from Khan et al. [13], with permission)



AFM in Liquids, Fig. 6 (**a**) AFM image of DNA adsorbed on mica [15]. The *inset* shows a higher-magnification image with the major and minor grooves of the DNA double helix structure clearly resolved. Image courtesy Alice Pyne. (**b**) Two-dimensional crystal of the membrane protein bacteriorhodopsin, obtained by FM AFM on the extracellular side of the protein (Image courtesy Carl Leung)

imaging that is significantly less susceptible to thermal drift. One of the key steps toward high-resolution imaging in liquids was the development of low-noise deflection sensors [11, 12]. This has brought the measurement noise down to the thermal noise of the cantilevers, even for stiff and heavily damped cantilevers in liquid. As a result, surfaces can now be stably measured with Ångström amplitudes of cantilever oscillation. This enhances the sensitivity to short-range forces and yields images such as depicted in Fig. 5. Because it is flat, hydrophilic, and easily cleaved, mica is one of the substrates of choice for adsorbing molecules in AFM experiments in aqueous solutions.

AFM has been successful in imaging DNA as well as DNA-protein complexes. DNA is routinely adsorbed on mica, where it is generally assumed to adopt a two-dimensional projection of its original three-dimensional configuration [14]. Under optimum imaging conditions, the major and minor grooves of the DNA double helix can be resolved [15]. The rather flat geometry required for high-resolution AFM is a natural one for proteins that are embedded in a lipid membrane. In particular when two-dimensional crystals are available, such as for bacteriorhodopsin, high-resolution images can readily be obtained in the various modes of AFM operation [16]. DNA and bacteriorhodopsin represent some of the biomolecules that have been most often imaged at high resolution by AFM in liquid (Fig. 6).



AFM in Liquids, Fig. 7 $100 \times 100 \times 3\ \mu m^3$ contact-mode AFM topograph of an osteoblast (bone cell) adsorbed on a glass coverslip (Image courtesy Guillaume Charras)

On a completely different scale, AFM can image objects as large as whole cells (Fig. 7). Due to thermal fluctuations, vibrations, and the overall softness of the cell, the spatial resolution is considerably lower than that obtained on flat surfaces or molecules directly adsorbed on a hard substrate. As, for cell imaging, alone, the advantage of AFM over optical techniques is therefore limited.

AFM can also be used to deliberately probe sample properties other than structure [17]. Single molecules that are tethered between the tip and the substrate can be deliberately stretched, unfolded, and allowed to refold at controlled load force [18], similar to optical-tweezers experiments. Using the same force spectroscopy techniques, force-distance curves can be obtained on whole cells, yielding information about their local elasticity or – when using chemically modified tips – local binding sites for specific ligands [19].

Finally, fast AFM methods [20] have improved the image rate from the typical frame per minute up to many frames per second. This gives direct access to the kinetics of molecular-scale processes at nanometer resolution.

Cross-References

- ▶ AFM Probes
- ▶ AFM, Noncontact Mode
- ▶ AFM, Tapping Mode
- ▶ Atomic Force Microscopy
- ▶ Force Modulation in Atomic Force Microscopy
- ▶ Friction Force Microscopy
- ▶ Imaging the Human Body Down to the Molecular Level
- ▶ Kelvin Probe Force Microscopy
- ▶ Magnetic Resonance Force Microscopy
- ▶ Mechanical Properties of Hierarchical Protein Materials
- ▶ Nanotechnology
- ▶ Nanotribology
- ▶ Optical Tweezers
- ▶ Optomechanical Resonators
- ▶ Surface Forces Apparatus

References

1. Sarid, D.: Scanning Force Microscopy with Applications to Electric, Magnetic, and Atomic Forces, 2nd edn. Oxford University Press, Oxford (1994)
2. Garcia, R., Perez, R.: Dynamic atomic force microscopy methods. Surf. Sci. Rep. **47**, 197–301 (2002)
3. Giessibl, F.J.: Advances in atomic force microscopy. Rev. Mod. Phys. **75**, 949–983 (2003)
4. Dufrêne, Y.F., Martínez-Martín, D., Medalsy, I., Alsteens, D., Müller, D.J.: Multiparametric imaging of biological systems by force-distance curve–based AFM. Nat. Methods **10**, 847–854 (2013)
5. Butt, H.-J.: Electrostatic interaction in atomic force microscopy. Biophys. J. **60**, 777–785 (1991)

6. Israelachvili, J.N.: Intramolecular and Surface Forces, 3rd edn. Academic, Oxford (2011)
7. Ohnesorge, F., Binnig, G.: True atomic resolution by atomic force microscopy through repulsive and attractive forces. Science **260**, 1451–1456 (1993)
8. Sader, J.E., Jarvis, S.P.: Accurate formulas for interaction force and energy in frequency modulation force spectroscopy. Appl. Phys. Lett. **84**, 1801–1803 (2004)
9. Bhushan, B., Fuchs, H., Hosaka, S. (eds.): Applied Scanning Probe Methods, vol. I. Springer, Berlin/Heidelberg (2004)
10. Morris, V.J., Gunnig, A.P., Kirby, A.R.: Atomic Force Microscopy for Biologists. World Scientific, River Edge (1999)
11. Hoogenboom, B.W., Frederix, P.L.T.M., Yang, J.L., Martin, S., Pellmont, Y., Steinacher, M., Zäch, S., Langenbach, E., Heimbeck, H.-J., Engel, A., Hug, H.J.: A Fabry-Perot interferometer for micrometer-sized cantilevers. Appl. Phys. Lett. **86**, 074101 (2005)
12. Fukuma, T., Kimura, M., Kobayashi, K., Matsushige, K., Yamada, H.: Development of low noise cantilever deflection sensor for multienvironment frequency-modulation atomic force microscopy. Rev. Sci. Instrum. **76**, 053704 (2005)
13. Khan, Z., Leung, C., Tahir, B., Hoogenboom, B.W.: Digitally tunable, wide-band amplitude, phase, and frequency detection for atomic-resolution scanning force microscopy. Rev. Sci. Instrum. **81**, 073704 (2010)
14. Hansma, H.G.: Surface biology of DNA by atomic force microscopy. Annu. Rev. Phys. Chem. **52**, 71–92 (2001)
15. Pyne, A., Thompson, R., Leung, C., Roy, D., Hoogenboom, B.W.: Single-molecule reconstruction of oligonucleotide secondary structure by atomic force microscopy. Small **10**, 3257–3261 (2014)
16. Müller, D.J., Engel, A.: Atomic force microscopy and spectroscopy of native membrane proteins. Nat. Protoc. **2**, 2191–2197 (2007)
17. Müller, D.J., Dufrêne, Y.F.: Atomic force microscopy as a multifunctional molecular toolbox in nanobiotechnology. Nat. Nanotechnol. **5**, 261–269 (2008)
18. Fisher, T.E., Marszalek, P.E., Fernandez, J.M.: Stretching single molecules into novel conformations using the atomic force microscope. Nat. Struct. Biol. **7**, 719–724 (2000)
19. Hinterdorfer, P., Dufrêne, Y.F.: Detection and localization of single molecular recognition events using atomic force microscopy. Nat. Methods **3**, 347–355 (2006)
20. Yamamoto, D., Uchihashi, T., Kodera, N., Yamashita, H., Nishikori, S., Ogura, T., Shibata, M., Ando, T.: High-speed atomic force microscopy techniques for observing dynamic biomolecular processes. In: Walter, N.G. (ed.) Methods in Enzymology, vol. 475, pp. 541–564. Academic, Burlington (2010)

AFM Probes

Kenji Fukuzawa
Department of Micro System Engineering, Nagoya University, Chikusa-ku, Aichi, Nagoya, Japan

Synonyms

AFM force sensors; AFM tips

Definition

AFM probes are transducers that convert the interaction force with a sample surface into a deformation or a change of the vibrational state of the probe. Most probes consist of a sharp microtip and a force transducer. The former determines the lateral resolution of the AFM and the latter provides the force sensitivity.

Overview

A typical AFM probe consists of a sharp tip and a microcantilever, which plays the role of a force transducer. In this case, the interaction force between the tip and sample deflects the cantilever. The deflection can be detected by a displacement measurement method, such as an optical lever or interference techniques. Assuming that the deflection and the spring constant of the cantilever are Δz and k, respectively, the interaction force F is given by

$$F = k\Delta z. \tag{1}$$

In contact AFM mode, the probe height is controlled so that the probe deflection is constant, which means that the interaction force does not change during the probe scanning. Mapping the controlled height can thus provide the topography of the sample.

In the early days, a metal foil with a glued microparticle or sharpened metal wire was used

as an AFM probe. Here, the particle or wire acted as a tip [1]. Improving the force sensitivity and reproducibility was not easy with those probes. Then, microfabricated cantilever probes were introduced [2]. These probes are made of silicon nitride or silicon and are produced by microfabrication techniques such as anisotropic chemical etching. Microfabricated probes enabled to improve the force sensitivity and expand the freedom of the probe design. They were also suitable for mass production. Therefore, microfabricated probes are the most common at present and various types of them are commercially available nowadays. Undoubtedly, the microfabricated probes have been playing an important role in AFM's becoming one of the most widely used methods in nanotechnology.

Probe Design

Since the specifications of the tip are rather limited by the fabrication method, the main part of the probe design is the force transducer, such as the rectangular micro cantilever sketched in Fig. 1. In addition to rectangular cantilevers, V-shaped cantilevers are widely used, especially in contact AFM mode. The force sensitivity of the probe should be high enough to detect weak interaction forces. From Eq. 1, the cantilever should convert a small force F into a large displacement Δz, which requires a small spring constant $k = F/\Delta z$. This means that higher sensitivity requires softer cantilevers. In addition to the force sensitivity, the robustness against mechanical disturbances from the environment should be considered. If the vibrations induced by the disturbances exceed the deflection due to the interaction force, the AFM signal is buried in the noise. The vibrations due to disturbances are amplified when the disturbance frequency is around the resonance frequency (natural frequency) of the probe. Since the disturbance frequencies are generally low, the resonance frequency should be set at a high value, typically larger than about 10 kHz. If the force transducer of the probe is simplified as a point mass and spring system, the resonance frequency is given by

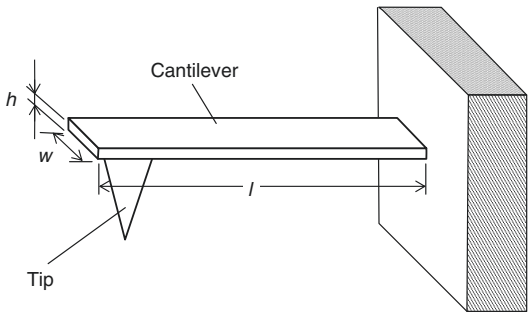
$$f_r = \frac{1}{2\pi}\sqrt{\frac{k}{m}}, \tag{2}$$

where m is the mass of the transducer. The sensitivity needs low values of k whereas the robustness against the environmental disturbances requires high values of f_r. This means that low values of m are required. Therefore, the force transducer of the probe should be soft and small. If a rectangle cantilever is chosen as a force transducer, it has to be thin and small. Microfabricated cantilevers meet this demand and, for this reason, they became standard probes for AFM.

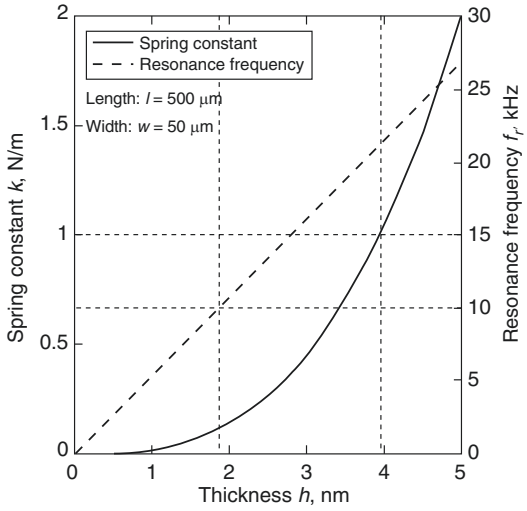
The design details of a rectangular cantilever are considered below. If the length, width, and thickness of the cantilever are l, w, and h, respectively, the spring constant k is given by

$$k = \frac{Ewh^3}{4l^3} \tag{3}$$

where E is the Young's modulus of the cantilever. It should be noted that the spring constant is proportional to $(h/l)^3$. This means that longer and thinner cantilevers can provide higher force sensitivity. The force F to be measured depends on the interaction between the tip and sample and the deflection Δz depends on the sensitivity of the displacement measurement method. Therefore, the spring constant k should be determined by considering the typical values of F and Δz for the targeted samples and the measurement system.



AFM Probes, Fig. 1 Schematic of a microcantilever probe



AFM Probes, Fig. 2 Thickness dependencies of the spring constant and resonance frequency of a microcantilever

On the other hand, the resonance frequency of a rectangular cantilever is given by

$$f_r = \frac{\lambda^2}{4\pi}\frac{h}{l^2}\sqrt{\frac{E}{3\rho}}, \tag{4}$$

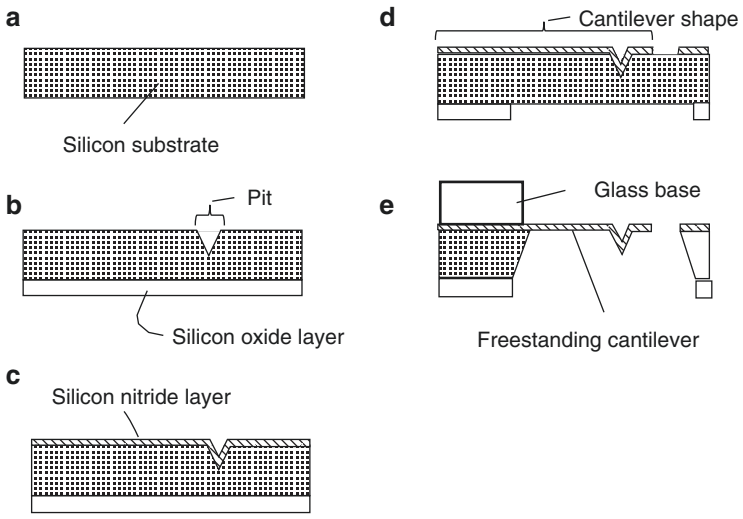
where ρ is the density of the cantilever and λ is a constant that depends on the vibrational mode. For the first mode λ is about 1.875. It should be noted that the cantilever width is not included in Eq. 4. Since f_r is proportional to h/l^2, higher resonance frequencies f_r require shorter and thicker cantilevers. However, as described above, high sensitivity requires opposite conditions. Thus, a balance between the spring constant and the resonance frequency has to be found in designing the probe. An example for contact mode cantilevers is shown in Fig. 2, where the thickness dependencies of the spring constant and resonance frequency are plotted using Eqs. 3 and 4 for a length $l = 500$ μm and a width $w = 50$ μm, respectively. If a spring constant of less than 1 N/m and a resonance frequency larger than 10 kHz are selected, the former curve requires a thickness of more than 1.9 μm and the latter curve requires a thickness of less than 4.0 μm. Therefore, a thickness in the overlapping range should be selected. A probe with these dimensions is not easy to fabricate by conventional machining and requires micromachining techniques.

Measurements of the spring constant k and resonance frequency f_r are important for converting the measured deflection signal into a force signal. The resonance frequency f_r can be measured by varying the drive frequency as it is routinely done in commercial AFM equipments. However, the accurate measurement of the spring constant k is not easy. An approximate value can be calculated from the dimensions of the cantilever using Eq. 3. An experimental method using the thermal fluctuation was presented in Ref. [3].

Many types of AFMs using a vibrating cantilever have been developed, such as cyclic contact mode and non-contact mode AFMs. The probe vibrations can be induced using a quartz-oscillator. In this case, the resonance properties, especially the quality factor Q, are very important. The quality factor Q is defined as $Q = 1/2\zeta$, where ζ is the damping coefficient. A probe with a high Q is able to resonate effectively and its vibration is hard to damp. The minimum detectable force gradient in the frequency-modulation detection, which is widely used in highly sensitive AFMs such as non-contact AFMs and magnetic force microscopy, is given by

$$\left(\frac{\partial F}{\partial z}\right)_{\min} = \frac{2}{A}\sqrt{\frac{kk_BTB}{\omega_r Q}}, \tag{5}$$

where B, ω_r, and A are the measurement band width, resonance angular frequency, and vibration amplitude, respectively [4]. Higher force sensitivity requires a higher Q. Since the vibration damping of the probe is caused by the internal friction and by friction between the probe and environmental gas such as air, the quality factor Q usually depends on the probe structure and environment. Typical values of Q of silicon cantilevers are on the order of 10–100 in air whereas they can increase up to more than 10,000 in vacuum. A method to increase Q electronically was presented, where the deflection signal is fed back to the drive signal after the phase of the deflection signal is shifted by 90° [5].



AFM Probes, Fig. 3 Fabrication process of a microfabricated probe: (**a**) silicon substrate, (**b**) formation of pit for a microtip, (**c**) deposition of silicon nitride layer, (**d**) patterning of cantilever, and (**e**) removal of substrate

Probe Fabrication

At present, most of the probes are fabricated by microfabrication techniques and various types of probes have been designed. Moreover, a considerable number of them are commercially available because microfabricated probes are suitable for mass production. The fabrication methods have also been improved, corresponding to a general improvement of the probe quality. Here, the fabrication methods for two typical types of probes are outlined. We will first address the probes made of silicon nitride. These probes are widely used in contact AFM mode. The fabrication process is shown in Fig. 3 [6]. In this method, a small pit that is formed in a silicon substrate by anisotropic wet etching is used as a mold. The silicon oxide layer is formed as an etching mask for the later substrate removal (Fig. 3a, b). A silicon nitride film is deposited onto the substrate and the cantilever is patterned by lithography. At this point, the cantilever with tip is shaped (Fig. 3c, d). By attachment of the glass base by anodic bonding and removal of the silicon substrate by wet etching, a freestanding cantilever is obtained (Fig. 3e). Another type of probes are silicon cantilevers [7]. Silicon cantilevers are the most widely used probes in AFM and they are operated in various modes. By using the undercut etching of the underneath of a circular mask, a tip is formed. Then the cantilever is patterned. The substrate silicon is removed by wet etching with protection of the fabricated tip and cantilever. Sharpening the tip end can be achieved by thermal oxidation followed by oxide removal [6].

Future Directions for Research

In addition to standard silicon or silicon nitride cantilever probes, various types of probes have been recently introduced. In a piezoresistive probe a piezoresistive strain sensor is embedded with a silicon cantilever [8]. The embedded sensor detects the cantilever deflection and makes the detection system such as an optical lever unnecessary. This not only makes the setup compact but also provides easier operation in the special case of liquid or vacuum environments. Along the growth of microfabricated cantilever probes, probes that use a quartz-oscillator as a force transducer have also been developed. Various types of oscillators such as tuning fork or linear-extension resonators are used [9]. In many probes, a microtip is manually glued to the quartz-oscillator although the oscillator is microfabricated. Quartz-oscillator probes have advantages of high Q and large spring constant. The latter allows to avoid

the jump-in of the probe. In addition, quartz-oscillator probes provide self-sensing as piezoresistive probes.

As stated above, the improvement of the probes is essential for the development of AFM. Recent advances in micro/nanomachining are promoting this improvement, which is expected to open up new applications of AFM.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Friction Force Microscopy
- ▶ MEMS
- ▶ Micromachining
- ▶ NEMS
- ▶ Scanning Probe Microscopy

References

1. Binnig, C., Quate, C.F., Gerber, C.: Atomic force microscope. Phys. Rev. Lett. **56**(9), 930–933 (1986)
2. Binnig, G., Gerber, C., Stoll, E., Albrecht, T.R., Quate, C.F.: Atomic resolution with atomic force microscope. Europhys. Lett. **3**(12), 1281–1286 (1987)
3. Hutter, J.L., Bechhoefer, J.: Calibration of atomic-force microscope tips. Rev. Sci. Instrum. **64**(7), 1868–1873 (1993)
4. Albrecht, T.R., Grutter, P., Horne, D., Rugar, D.: Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity. J. Appl. Phys. **69**(2), 668–673 (1991)
5. Anczykowski, B., Cleveland, J.P., Krüger, D., Elings, V., Fuchs, H.: Analysis of the interaction mechanisms in dynamic mode SFM by means of experimental data and computer simulation. Appl. Phys. **A66**, S885–S889 (1998)
6. Albrecht, T.R., Akamine, S., Carver, T.E., Quate, C.F.: Microfabrication of cantilever styli for the atomic force microscope. J. Vac. Sci. Technol. **A8**, 3386–3396 (1990)
7. Wolter, O., Bayer, T., Greschner, J.: Micromachined silicon sensors for scanning force microscopy. J. Vac. Sci. Technol. **B9**, 1353–1357 (1991)
8. Tortonese, M., Barrett, R.C., Quate, C.F.: Atomic resolution with an atomic force microscope using piezoresistive detection. Appl. Phys. Lett. **62**(8), 834–836 (1993)
9. Giessibl, F.J.: Atomic resolution on Si(111)-(7 × 7) by noncontact atomic force microscopy with a force sensor based on a quartz tuning fork. Appl. Phys. Lett. **76**(11), 1470–1472 (2000)

AFM Tips

▶ AFM Probes

AFM, Noncontact Mode

Hendrik Hölscher
Karlsruher Institut für Technologie (KIT), Institut für Mikrostrukturtechnik, Karlsruhe, Germany

Definition

The *noncontact atomic force microscopy* (NC-AFM) is a specific AFM technique primarily developed for application in vacuum where standard AFM cantilevers made from silicon or silicon nitride exhibit very high quality factors Q, what makes the response of the system slow if driven in AM or tapping mode (▶ AFM, Tapping Mode). The technique to oscillate the cantilever also in high Q environments is called *frequency-modulation* (FM) mode. In contrast to the tapping or AM mode typically applied in air or liquids, this approach features a so-called self-driven oscillator, which uses the cantilever deflection itself as drive signal, thus ensuring that the cantilever instantaneously adapts to changes in the resonance frequency. The NC-AFM technique is the method of choice to obtain true atomic resolution on non-conduction surfaces with an atomic force microscope.

Overview

To obtain high resolution images with an atomic force microscope it is most important to prepare clean sample surfaces free from unwanted adsorbates. Therefore, these experiments are usually performed in ultra-high vacuum with pressures below 1×10^{-10} mbar. As a consequence most dynamic force microscope experiments in vacuum utilize the so-called *frequency modulation* (FM) detection scheme introduced by Albrecht

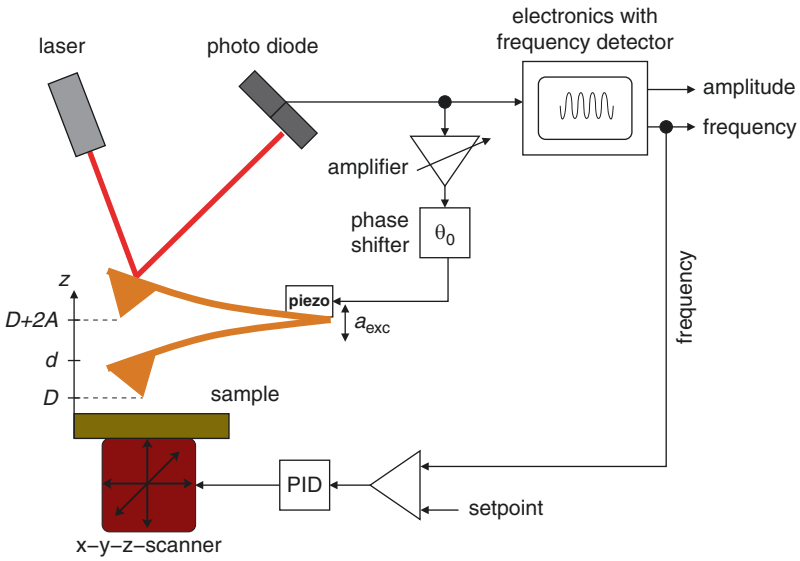
et al. [1]. In this mode the cantilever is self-oscillated, in contrast to the AM- or tapping-mode (▶ AFM, Tapping Mode). The FM-technique enables the imaging of single point defects on clean sample surfaces in vacuum and its resolution is comparable with the scanning tunneling microscope, while not restricted to conducting surfaces [2–6]. In the years after the invention of the FM-technique the term *non-contact atomic force microscopy* (NC-AFM) was established, because it is commonly believed that a repulsive, destructive contact between tip and sample is prevented by this technique.

Set-Up of FM-AFM

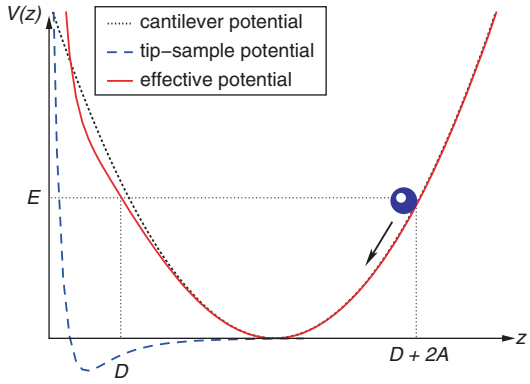
In vacuum applications, the Q-factor of silicon cantilevers is in the range of 10,000–30,000. High Q-factors, however, limit the acquisition time (bandwidth) of a dynamic force microscopy, since the oscillation amplitude of the cantilever needs a long time to adjust. This problem is avoided by the FM-detection scheme based on the specific features of a self-driven oscillator.

The basic set-up of a dynamic force microscope utilizing this driving mechanism is schematically shown in Fig. 1. The movement of microfabricated cantilevers is typically measured with the laser beam deflection method or an interferometer. A self-detecting sensor like a tuning fork do not need an additional detection sensor. In any case the amplitude signal fed back into an amplifier with an *automatic gain control* (AGC) and is subsequently used to excite the piezo oscillating the cantilever. The time delay between the excitation signal and cantilever deflection is adjusted by a time ("phase") shifter to a value of $\approx 90°$, since this ensures an oscillation at resonance. Two different modes have been established: The *constant amplitude*-mode [1], where the oscillation amplitude A is kept at a constant value by the AGC, and the *constant excitation mode* [7], where the excitation amplitude is kept constant. In the following, however, only the constant amplitude mode is discussed.

The key feature of the described set-up is the positive feedback-loop which oscillates the cantilever always at its resonance frequency f [8]. The reason for this behavior is that the cantilever serves as the frequency determining element. This is in contrast to an external driving of the cantilever in tapping mode by a frequency generator (▶ AFM, Tapping Mode). If the cantilever oscillates near the sample surface, the tip-sample interaction alters its resonant frequency, which is then different from the eigenfrequency f_0 of the free cantilever. The actual value of the resonant frequency depends on the nearest tip-sample distance and the oscillation amplitude. The measured



AFM, Noncontact Mode, Fig. 1 The schematic set-up of a dynamic force microscope based on the frequency modulation technique often used in UHV. A significant feature is the positive feedback of the self-driven cantilever. The detector signal is amplified and phase shifted before it is used to drive the piezo. The measured quantity is the frequency shift due to the tip-sample interaction, which serves as the feedback signal for the cantilever-sample distance



AFM, Noncontact Mode, Fig. 2 The frequency shift in dynamic force microscopy is caused by the tip-sample interaction potential (*dashed line*), which alters the harmonic cantilever potential (*dotted line*). Therefore, the tip moves in an anharmonic and asymmetric effective potential (*solid line*)

quantity is the *frequency shift* Δf, which is defined as the difference between both frequencies ($\Delta f := f - f_0$). For imaging the frequency shift Δf is used to control the cantilever sample distance. Thus, the frequency shift is constant and the acquired data represents planes of constant Δf, which can be related to the surface topography in many cases.

Origin of the Frequency Shift

Before presenting experimental results obtained in vacuum the origin of the frequency shift is analyzed in mode detail. A good insight into the cantilever dynamics is given by looking at the tip potential displayed in Fig. 2. If the cantilever is far away from the sample surface, the tip moves in a symmetric parabolic potential (dotted line), and its oscillation is harmonic. In such a case, the tip motion is sinusoidal and the resonance frequency is given by the eigenfrequency f_0 of the cantilever. If, however, the cantilever approaches the sample surface, the potential – which determines the tip oscillation – is modified to an effective potential V_{eff} (solid line) given by the sum of the parabolic potential and the tip-sample interaction potential V_{ts} (dashed line). This effective potential differs from the original parabolic potential and shows an asymmetric shape.

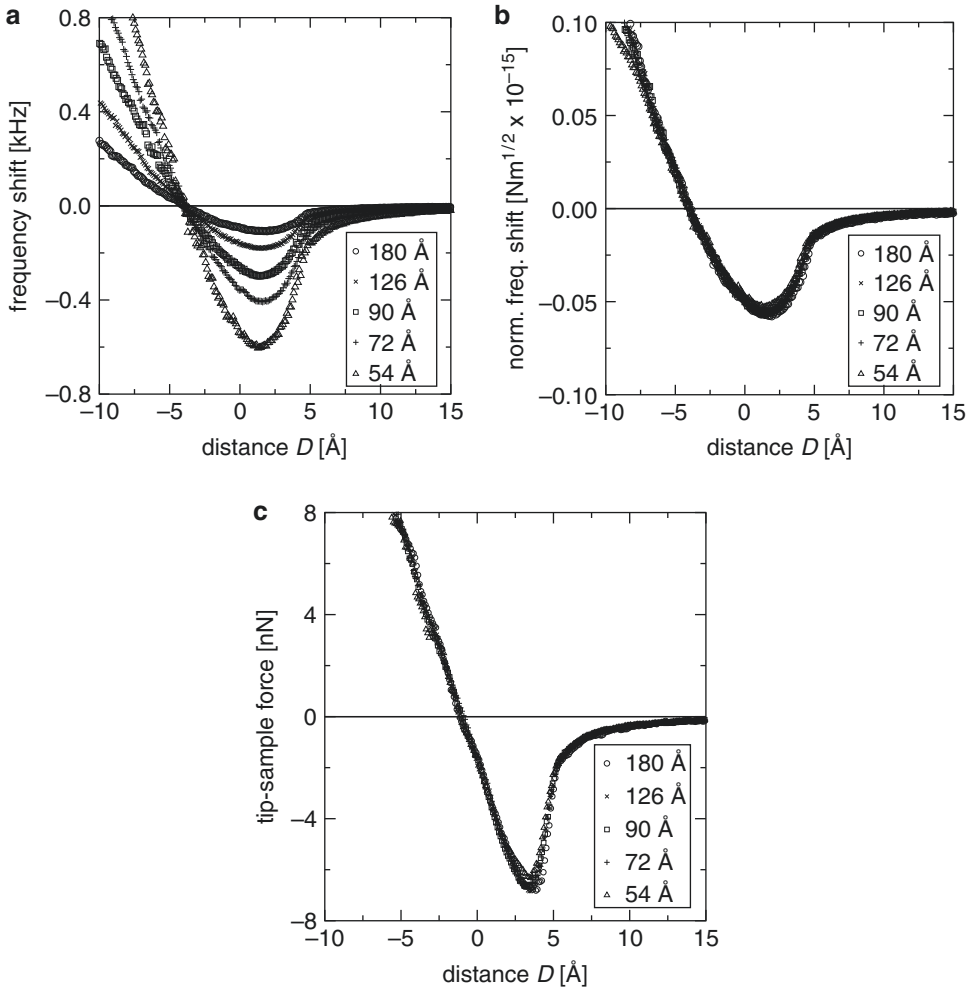
As a result of this modification of the tip potential the oscillation becomes anharmonic, and the resonance frequency of the cantilever depends now on the oscillation amplitude A. Since the effective potential experienced by the tip changes also with the nearest distance D, the frequency shift is a functional of both parameters ($\Rightarrow \Delta f := \Delta f(D, A)$).

Figure 3 displays some experimental frequency shift versus distance curves for different oscillation amplitudes [9]. The obtained experimental frequency shift versus distance curves show a behavior expected from the simple model explained above. All curves show a similar overall shape, but differ in magnitude in dependence of the oscillation amplitude and the nearest tip-sample distance. During the approach of the cantilever towards the sample surface, the frequency shift decreases and reaches a minimum. With a further reduction of the nearest tip-sample distance, the frequency shift increases again and becomes positive. For smaller oscillation amplitudes, the minimum of the $\Delta f(z)$-curves is deeper and the slope after the minimum is steeper than for larger amplitudes, i.e., the overall effect is larger for smaller amplitudes.

This can be explained by the simple potential model as well: A decrease of the amplitude A for a fixed nearest distance D moves the minimum of the effective potential closer to the sample surface. Therefore, the relative perturbation of the harmonic cantilever potential increases, which increases also the absolute value of the frequency shift.

Theory of FM-AFM

As already described in the previous subsection it is a specific feature of the FM-modulation technique that the cantilever is "self-driven" by a positive feedback loop. Due to this experimental set-up, the corresponding equation of motion is different from the case of the externally driven cantilever discussed for the tapping-mode. The external driving term has to be replaced in order to describe the self-driving mechanism correctly. Therefore, the equation of motion is given by



AFM, Noncontact Mode, Fig. 3 (**a**) Experimental frequency shift versus distance curves acquired with a silicon cantilever ($c_z = 38$ N/m; $f_0 = 171$ kHz) and a graphite sample for different amplitudes (54–180 Å) in UHV at low temperature ($T = 80$ K). (**b**) Transformation of all frequency shift curves shown in (**a**) to one universal curves using Eq. 6. The normalized frequency shift $\gamma(D)$ is nearly identical for all amplitudes. (**c**) The tip-sample force calculated with the experimental data shown in (**a**) and (**b**) using the formula Eq. 7

$$m\ddot{z}(t) + \frac{2\pi f_0 m}{Q}\dot{z}(t) + c_z(z(t) - d) + \underbrace{g c_z(z(t - t_0) - d)}_{\text{driving}} = F_{\text{ts}}[z(t), \dot{z}(t)]. \tag{1}$$

where $z := z(t)$ represents the position of the tip at the time t; cz, m, and Q are the spring constant, the effective mass, and the quality factor of the cantilever, respectively. $F_{\text{ts}} = -(\partial V_{\text{ts}})/(\partial z)$ is the tip-sample interaction force. The last term on the left describes the active feedback of the system by the amplification of the displacement signal by the *gain factor* g measured at the retarded time $t - t_0$.

The frequency shift can be calculated from the above equation of motion with the ansatz

$$z(t) = d + A\cos(2\pi f t) \tag{2}$$

describing the stationary solutions of Eq. 1. Again, it is assumed that the cantilever oscillations are more or less sinusoidal and develop the

tip-sample force F_{ts} into a Fourier-series. As a result the following equation for the frequency shift is obtained

$$\Delta f = \frac{1}{\pi c_z A^2} \int_{d-A}^{d+A} (F\downarrow + F\uparrow) \frac{z-d}{\sqrt{A^2-(z-d)^2}} dz \tag{3}$$

and the energy dissipation

$$\Delta E = \left(g - \frac{1}{Q}\frac{f}{f_0}\right) \pi c_z A^2. \tag{4}$$

Since the amplitudes in FM-AFM are often considerably larger than the distance range of the tip-sample interaction, apply the "large amplitude approximation" [10, 11] can be applied. This yields the formula

$$\Delta f = \frac{1}{\sqrt{2\pi}} \frac{f_0}{c_z A^{3/2}} \int_D^{D+2A} \frac{F_{ts}(z)}{\sqrt{z-D}} dz \tag{5}$$

It is interesting to note that the integral in this equation is virtually independent of the oscillation amplitude. The experimental parameters (cz, f_0, and A) appear as pre-factors. Consequently, it is possible to define the *normalized frequency shift* [10]

$$\gamma(z) := \frac{c_z A^{3/2}}{f_0} \Delta f(z) \tag{6}$$

This is a very useful quantity to compare experiments obtained with different amplitudes and cantilevers. The validity of Eq. 6 is nicely demonstrated by the application of this equation to the frequency shift curves already presented in Fig. 3a. As shown in Fig. 3b all curves obtained for different amplitudes result into one universal γ-curve, which depends only on the actual tip-sample distance D.

These equations help to calculate the frequency shift for a given tip-sample interaction law. The inverse problem, however, is even more interesting: *How can the tip-sample interaction be determined from frequency shift data?* Several solutions to this question have been presented by various authors and have lead to the *dynamic force spectroscopy* (DFS) technique, which is a direct extension of the FM-AFM mode [12].

Here the approach of Dürig [11] is presented, which is based on the inversion of the integral Eq. 5. It can be transformed to

$$F_{ts}(D) = \sqrt{2} \frac{c_z A^{3/2}}{f_0} \frac{\partial}{\partial D} \int_D^{\infty} \frac{\Delta f(z)}{\sqrt{z-D}} dz, \tag{7}$$

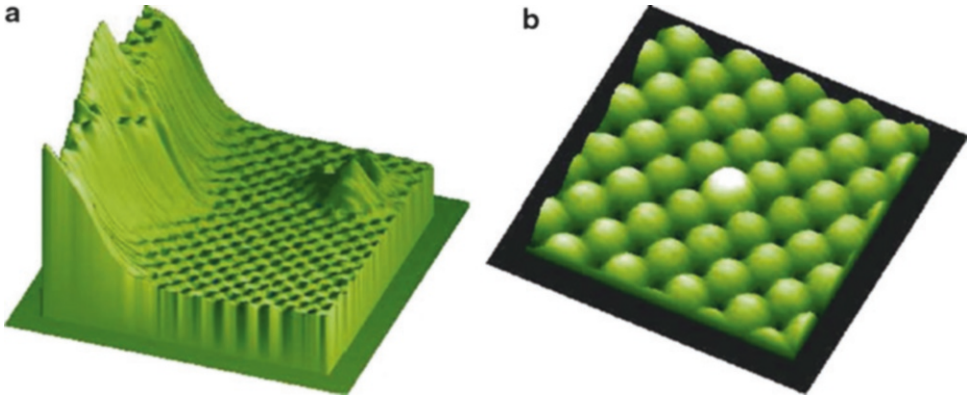
which allows a direct calculation of the tip-sample interaction force from the frequency shift versus distance curves.

An application of this formula to the experimental frequency shift curves already presented in Fig. 3a is shown in Fig. 3c. The obtained force curves are nearly identical although obtained with different oscillation amplitudes. Since the tip-sample interactions can be measured with high resolution, dynamic force spectroscopy opens a direct way to compare experiments with theoretical models and predictions.

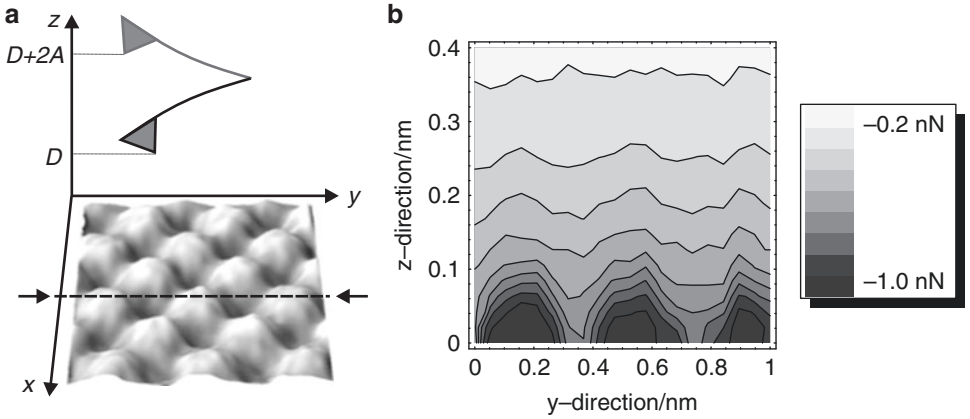
Applications of FM-AFM

The excitement about the NC-AFM technique in ultrahigh vacuum was driven by the first results of Giessibl [13] who achieved to image the true atomic structure of the Si(111)-7 × 7-surface with this technique in 1995. In the same year Sugawara et al. [14] observed the motion of single atomic defects on InP with true atomic resolution. However, imaging on conducting or semiconducting surfaces is also possible with the scanning tunneling microscope (STM) and these first NC-AFM images provided no new information on surface properties. The true potential of NC-AFM lies in the imaging of non-conducting surface with atomic precision. A long-standing question about the surface reconstruction of the technological relevant material Aluminium oxide could be answered by Barth et al. [15], who imaged the atomic structure of the high temperature phase of α-Al_2O_3(0001).

The high resolution capabilities of non-contact atomic force microscopy are nicely demonstrated



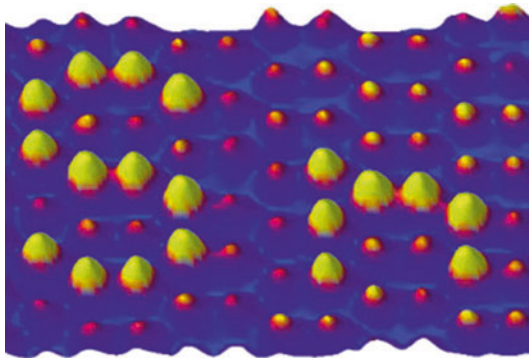
AFM, Noncontact Mode, Fig. 4 Imaging of a NiO(001) sample surface with a non-contact AFM. (**a**) Surface step and an atomic defect. The lateral distance between two atoms is 4.17 Å. (**b**) A dopant atom is imaged as a light protrusion about 0.1 Å higher as the other atoms (Images courtesy of W. Allers and S. Langkat, University of Hamburg; used with kind permission)



AFM, Noncontact Mode, Fig. 5 (**a**) Principle of 3D-force spectroscopy. The cantilever oscillates near the sample surface and measure the frequency shift in a x-y-z-box. The three dimensional surface shows the topography of the sample (image size: 10 Å × 10 Å) obtained immediately before the recording of the spectroscopy field. (**b**) The reconstructed force field of NiO(001) shows atomic resolution. The data is taken along the line shown in (**a**)

by the images shown in Fig. 4. Allers et al. [16] resolved atomic steps and defects with atomic resolution on Nickel oxide. Today such a resolution is routinely obtained by various research groups (for an overview see, e.g., Refs. [2–6]). Recent efforts have also been concentrated on the analysis of functional organic molecules, since in the field of nanoelectronics it is anticipated that in particular organic molecules will play an important role as the fundamental building blocks of nanoscale electronic device elements.

The concept of dynamic force spectroscopy can be also extended to 3D-force spectroscopy by mapping the complete force field above the sample surface [12]. Figure 5a shows a schematic of the measurement principle [17]. Frequency shift vs. distance curves are recorded on a matrix of points perpendicular to the sample surface. Using Eq. 7 the complete three-dimensional force field between tip and sample can be recovered with atomic resolution. Figure 5b shows a cut through the force field as a two-dimensional map.



AFM, Noncontact Mode, Fig. 6 Final topographic NC-AFM image of the process of rearranging single Sn-atoms on a Ge(111)-c(2 × 8) semiconductor surface at room temperature (Reproduced from [19])

If the NC-AFM is capable of measuring forces between single atoms with sub-nN precision, why should it not be possible to also exert forces with this technique? In fact, the new and exciting field of nanomanipulation would be driven to a whole new dimension, if defined forces can be reliably applied to single atoms or molecules. In this respect, Loppacher et al. [18] achieved to push on different parts of an isolated Cu-TBBP molecule, which is known to possess four rotatable legs. They measured the force-distance curves while one of the legs was pushed by the AFM tip and turned by 90°, thus being able to measure the energy which was dissipated during "switching" this molecule between different conformational states. The manipulation of single Sn-atoms with the NC-AFM was nicely demonstrated by Sugimoto et al. who manipulated single Sn-atoms on the Ge(111)-c(2 × 8) semiconductor surface (Fig. 6). By pushing single Sn-atoms from one lattice site to the other they finally succeeded to write the letter "Sn" with single atoms.

Cross-References

- ▶ AFM, Tapping mode
- ▶ Friction Force Microscopy
- ▶ Kelvin Probe Force Microscopy
- ▶ Magnetic Resonance Force Microscopy
- ▶ Piezoresponse Force Microscopy and Spectroscopy

References

1. Albrecht, T.R., Grütter, P., Horne, D., Rugar, D.: Frequency modulation detection using high-Q cantilevers for enhanced force microscope sensitivity. J. Appl. Phys. **69**, 668 (1991)
2. Morita, S., Wiesendanger, R., Meyer, E. (eds.): Noncontact Atomic Force Microscopy. Springer, Berlin (2002)
3. Garcia, R., Pérez, R.: Dynamic atomic force microscopy methods. Surf. Sci. Rep. **47**, 197 (2002)
4. Giessibl, F.-J.: Advances in atomic force microscopy. Rev. Mod. Phys. **75**, 949 (2003)
5. Meyer, E., Hug, H.J., Bennewitz, R.: Scanning Probe Microscopy – The Lab on a Tip. Springer, Berlin (2004)
6. Morita, S., Giessibl, F.J., Wiesendanger, R. (eds.): Noncontact Atomic Force Microscopy, vol. 2. Springer, Berlin (2009)
7. Ueyama, H., Sugawara, Y., Morita, S.: Stable operation mode for dynamic noncontact atomic force microscopy. Appl. Phys. A **66**, S295 (1998)
8. Hölscher, H., Gotsmann, B., Allers, W., Schwarz, U.D., Fuchs, H., Wiesendanger, R.: Comment on "Damping mechanism in dynamic force microscopy". Phys. Rev. Lett. **88**, 019601 (2002)
9. Hölscher, H., Schwarz, A., Allers, W., Schwarz, U.D., Wiesendanger, R.: Quantitative analysis of dynamic force spectroscopy data on graphite(0001) in the contact and non-contact regime. Phys. Rev. B **61**, 12678 (2000)
10. Giessibl, F.J.: Forces and frequency shifts in atomic-resolution dynamic-force microscopy. Phys. Rev. B **56**, 16010 (1997)
11. Dürig, U.: Relations between interaction force and frequency shift in large-amplitude dynamic force microscopy. Appl. Phys. Lett. **75**, 433 (1999)
12. Baykara, M.Z., Schwendemann, T.C., Altman, E.I., Schwarz, U.D.: Three-dimensional atomic force microscopy taking surface imaging to the next level. Adv. Mater. **22**, 2838 (2010)
13. Giessibl, F.-J.: Atomic resolution of the silicon (111)-(7 × 7) surface by atomic force microscopy. Science **267**, 68 (1995)
14. Sugawara, Y., Otha, M., Ueyama, H., Morita, S.: Defect motion on an InP(110) surface observed with noncontact atomic force microscopy. Science **270**, 1646 (1995)
15. Barth, C., Reichling, M.: Imaging the atomic arrangement on the high-temperature reconstructed α-Al_2O_3(0001) surface. Nature **414**, 54 (2001)
16. Allers, W., Langkat, S., Wiesendanger, R.: Dynamic low-temperature scanning force microscopy on nickel oxide (001). Appl. Phys. A **72**, S27 (2001)
17. Hölscher, H., Langkat, S.M., Schwarz, A., Wiesendanger, R.: Measurement of threedimensional force fields with atomic resolution using dynamic force spectroscopy. Appl. Phys. Lett. **81**, 4428 (2002)

18. Loppacher, C., Guggisberg, M., Pfeiffer, O., Meyer, E., Bammerlin, M., Luthi, R., Schlittler, R., Gimzewski, J.K., Tang, H., Joachim, C.: Direct determination of the energy required to operate a single molecule switch. Phys. Rev. Lett. **90**, 066107 (2003)
19. Sugimoto, Y., Abe, M., Hirayama, S., Oyabu, N., Custance, O., Morita, S.: Atom inlays performed at room temperature using atomic force microscopy. Nat. Mater. **4**, 156 (2005)

AFM, Tapping Mode

Hendrik Hölscher
Karlsruher Institut für Technologie (KIT), Institut für Mikrostrukturtechnik, Karlsruhe, Germany

Definition

The *tapping* or AM-mode is the most common dynamic mode used in atomic force microscopy. In dynamic mode AFM the cantilever is oscillated with (or near) its resonance frequency near the sample surface. Using a feedback electronic the cantilever sample distance is controlled by keeping either the amplitude or the phase of the oscillating cantilever constant. Since lateral tip–sample forces are avoided by this technique the resolution is typically higher compared to the classical contact mode AFM where tip and sample are in direct mechanical contact.

Overview

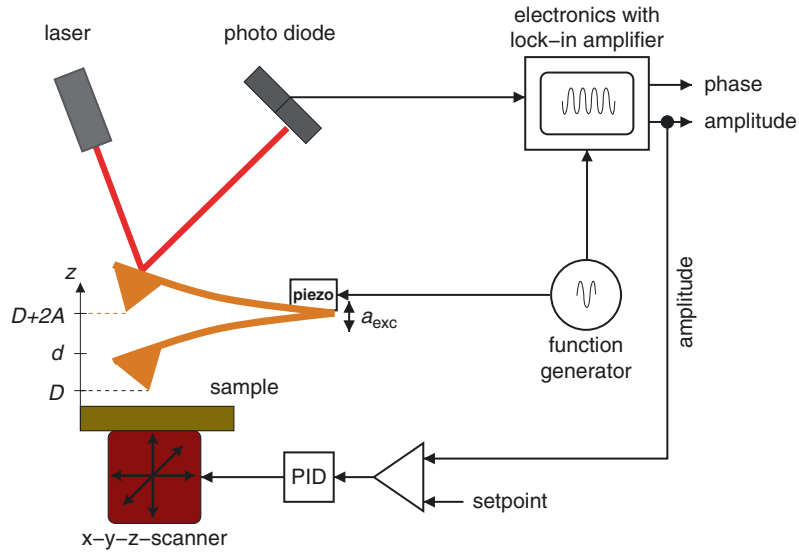
Since its introduction in 1986 [1], the *atomic force microscope* became a standard tool in nanotechnology. In early experimental setups, a sharp tip located at or near the end of a microstructured cantilever profiled the sample surface in direct mechanical contact (*contact mode*) to measure the force acting between tip and sample. Maps of constant tip–sample interaction force, which are usually regarded as representing the sample's "topography," were then recovered by keeping the deflection of the cantilever constant (▶ Friction Force Microscopy). This is achieved by means of a feedback loop that continuously adjusts the z-position of the sample during the scan process so that the output of the deflection sensor remains unchanged at a preselected set-point.

Despite the widespread success of contact mode AFM in various applications, the resolution was frequently found to be limited (in particular for soft samples) by lateral forces acting between tip and sample. In order to avoid this effect, it is advantageous to vibrate the cantilever in vertical direction near the sample surface. AFM imaging with an oscillating cantilever is often denoted as *dynamic force microscopy* (DFM) [2].

The cantilever dynamics are governed by the tip–sample interaction as well as by its driving method. For instance, the historically oldest scheme of cantilever excitation in DFM imaging is the external driving of the cantilever at a fixed excitation frequency chosen to be exactly at, or very close to, the cantilever's first resonance. For this driving mechanism, different detection schemes measuring either the change of the oscillation amplitude or the phase shift between driving signal and resulting cantilever motion were proposed. Over the years, the *amplitude modulation* (AM) or *tapping* mode, where the actual value of the oscillation amplitude is employed as a measure of the tip–sample distance, has been established as the most widely applied technique for ambient conditions and liquids.

Experimental Setup

A schematic of the experimental setup of an atomic force microscope driven in amplitude modulation mode is shown in Fig. 1. In this mode the cantilever is vibrated close to its resonant frequency near the sample surface. Due to the tip–sample interaction the resonant frequency (and consequently also amplitude A and phase ϕ) of the cantilever changes with the cantilever sample distance d. Therefore, the amplitude as well as the phase can be used as feedback channels. A certain set-point for example, the amplitude is given, and the feedback loop will adjust the tip–sample distance such that the amplitude remains constant. The cantilever sample distance is recorded as a function of the lateral position of



AFM, Tapping Mode, Fig. 1 Setup of a dynamic force microscope operated in the tapping (or AM-) mode. A laser beam is deflected by the back side of the cantilever, and its deflection is detected by a split photo-diode. The cantilever vibration is caused by an external frequency generator driving an excitation piezo. A lock-in amplifier is used to compare the cantilever driving with its oscillation. The amplitude signal is held constant by a feedback loop controlling the cantilever sample distance

the tip with respect to the sample and the scanned height essentially represents the surface topography.

The deflection of the cantilever is typically measured with the laser beam deflection method. During operation in conventional tapping mode, the cantilever is driven with a fixed frequency and a constant excitation amplitude using an external function generator, while the resulting oscillation amplitude and/or the phase shift are detected by a lock-in amplifier. The function generator supplies not only the signal for the dither piezo; its signal serves simultaneously as a reference for the lock-in amplifier.

This setup is mostly used in air and in liquids. A typical image obtained with this experimental setup in ambient conditions is shown in Fig. 2. The phase between excitation and oscillation can be acquired as an additional channel and gives information about the different material properties of DPPC and the mica substrate. As shown at the end of the next section, the phase signal is closely related to the energy dissipated in the tip–sample contact.

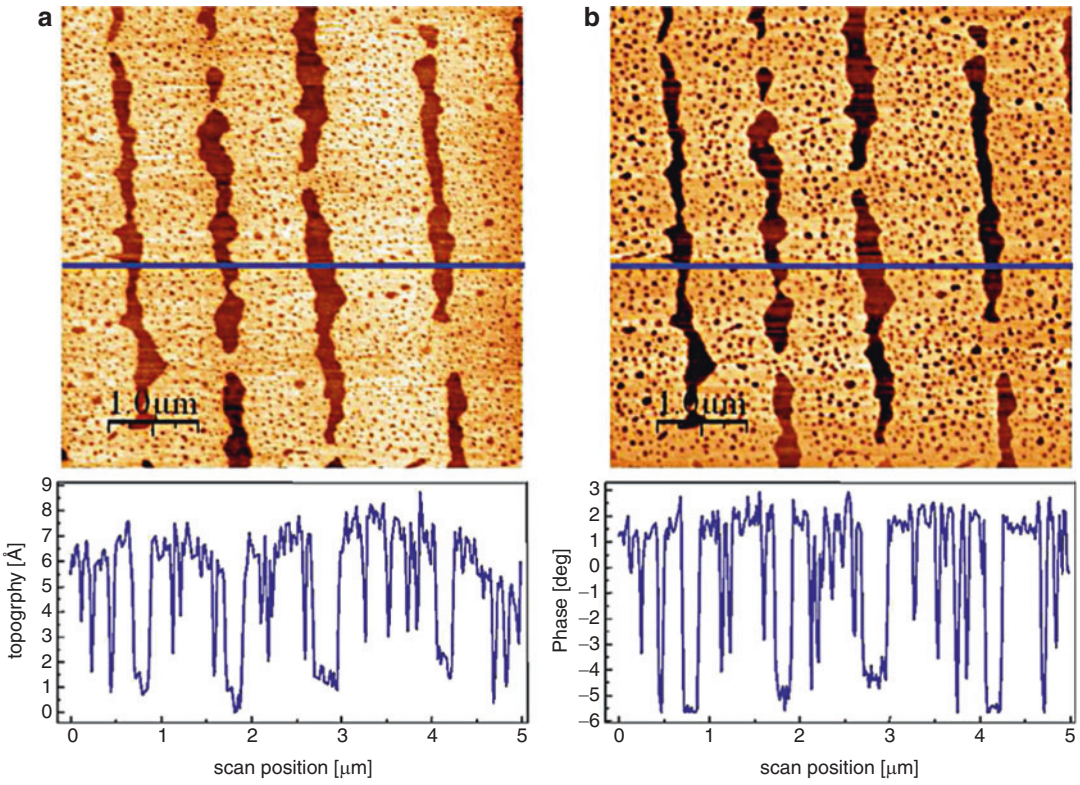
Due to its technical relevance the investigation of polymers has been the focus of many studies and high-resolution imaging has been extensively performed in the area of material science. Using specific tips with additionally grown sharp spikes Klinov et al. [3] obtained molecular resolution on a polydiacetylene crystal. Imaging in liquids opens up the avenue for the investigation of biological samples in their natural environment. For example, Möller et al. [4] have obtained high-resolution images of the topography of hexagonally packed intermediate (HPI) layer of *Deinococcus radiodurans* with tapping-mode AFM. A typical example for the imaging of DNA in liquid solution is shown in Fig. 3.

Theory of Tapping-Mode AFM

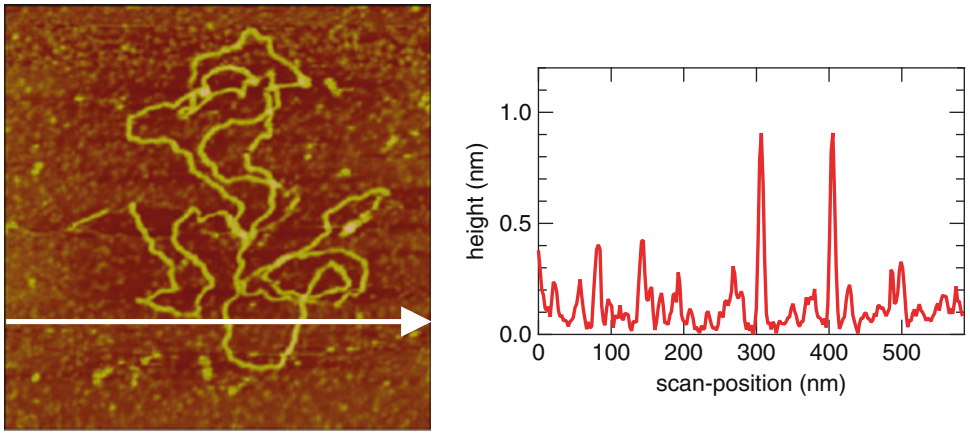
Many features observed in tapping-mode AFM can be described by a simple spring-mass model which includes the tip–sample interaction force.

$$m\ddot{z}(t) + \frac{2\pi f_0 m}{Q_0}\dot{z}(t) + c_z(z(t) - d) = \underbrace{a_d cz \cos(2\pi f_d t)}_{\text{external driving force}} + \underbrace{F_{ts}[z(t), \dot{z}(t)}_{\text{tip–sample force}}. \quad (1)$$

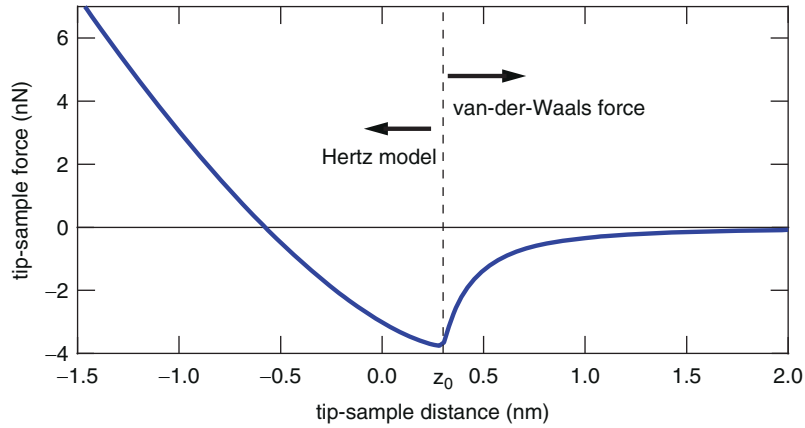
Here, $z(t)$ is the position of the tip at the time t; cz, m, and $f_0 = \sqrt{(c_z/m)}/(2\pi)$ are the spring constant, the effective mass, and the eigenfrequency of the cantilever, respectively. The quality factor Q_0 combines the intrinsic damping of the cantilever and all influences from surrounding media,



AFM, Tapping Mode, Fig. 2 (**a**) A dynamic force microscopy image of a monomolecular DPPC (l-α-dipalmitoyl-phophatidycholine) film adsorbed on mica. (**b**) The phase contrast is different between substrate and DPPC film. Equation 6 explains how this contrast is related to the energy dissipation caused by the tip–sample contact interaction



AFM, Tapping Mode, Fig. 3 Topography of DNA adsorbed on mica imaged in buffer solution by tapping mode AFM. The *right* graph shows a single scan line obtained at the position marked by an arrow in the *left* image



AFM, Tapping Mode, Fig. 4 Tip–sample model force after the DMT-M model for air Eq. 2 using the typical parameters: $A_H = 0.2$ aJ, $R = 10$ nm, $z_0 = 0.3$ nm, and $E^* = 1$ GPa. The *dashed line* marks the position z_0 where the tip touches the surface

such as air or liquid. The equilibrium position of the tip is denoted as d. The first term on the right-hand side of the equation represents the external driving force of the cantilever by the frequency generator. It is modulated with the constant excitation amplitude a_d at a fixed frequency f_d. The (nonlinear) tip–sample interaction force F_{ts} is introduced by the second term.

The actual tip–sample force is unknown in nearly all cases. However, in order to understand the most important effects it is often sufficient to apply the *DMT-M theory* [5]. In this approach it is assumed the tip is nearly spherical and that the noncontact forces are given by the long-range van-der-Waals forces while the contact forces are described by the well-known Hertz model. The resulting overall force law is given by

$$F_{\mathrm{DMT-M}}(z) = \begin{cases} -\dfrac{A_H R}{6z^2} & \text{for } z \geq z_0, \\ \dfrac{4}{3} E * \sqrt{R}(z_0 - z)^{3/2} & \text{for } z < z_0, \end{cases} \tag{2}$$

where AH is the Hamaker constant, E^* the effective modulus, and R the tip radius. At z_0 tip and sample come in contact. Figure 4 displays the resulting tip–sample force curve for this model.

For the analysis of dynamic force microscopy, experiments only on the steady states with sinusoidal cantilever oscillation are of interest. Consequently, the steady-state solution is given by the ansatz

$$z(t \gg 0) = d + A\ \cos(2\pi f_d t + \phi), \tag{3}$$

where ϕ is the phase difference between the excitation and the oscillation of the cantilever.

Here, the situation where the driving frequency is set *exactly* to the eigenfrequency of the cantilever ($f_d = f_0$) is analyzed. With this choice, which is also very common in actual DFM experiments, defined imaging conditions are given leading to a handy formula relationship between the free oscillation amplitude A_0, the actual amplitude A, and the equilibrium tip position d [6].

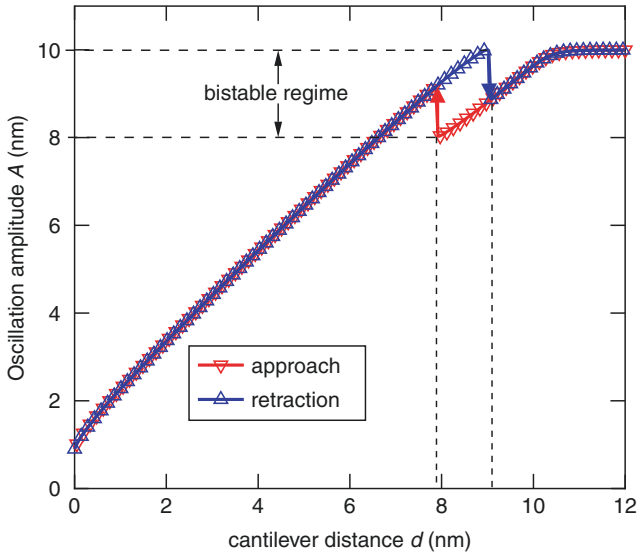
$$A_0 = A\sqrt{1 + (Q_0 I_+[d, A]^2}, \tag{4}$$

where

$$\begin{aligned} & I + (d, A) \\ &= \frac{2 f_d}{c_z A} \int_0^{1/f_d} F_{ts}[z(t), \dot{z}(t)] \cos(2\pi f_d t + \phi) dt \\ &= \frac{1}{p c_z A^2} \int_{d-A}^{d+A} (F_\downarrow + F_\uparrow) \frac{z - d}{\sqrt{A^2 - (z - d)^2}} dz. \end{aligned} \tag{5}$$

The function $I_+(d, A)$ depends on the actual oscillation amplitude A and cantilever–sample distance d and it is a weighted average of the tip–sample forces during approach and retraction ($F_\downarrow + F_\uparrow$).

The solution of this equation allows us to study amplitude versus distance curves as shown in Fig. 5 for a Q-factor of 300. Most noticeable, the

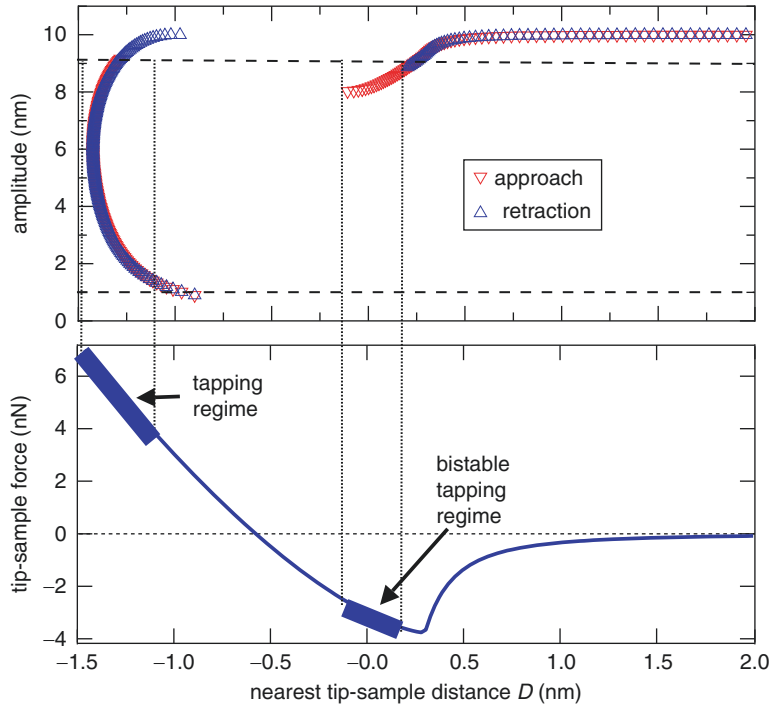


AFM, Tapping Mode, Fig. 5 Amplitude versus distance curve for tapping-mode AFM assuming the parameters $A_0 = 10$ nm, $f_0 = 300$ kHz, and the tip–sample interaction force shown in Fig. 4. The overall amplitude decreases the sample surface distance, but instabilities (indicated by *arrows*) occur during approach and retraction

tapping mode curve exhibits jumps between unstable branches, which occur at different locations for approach and retraction. The resulting bistable regime then causes a hysteresis between approach and retraction and divides the tip–sample interaction into two regimes. In order to identify the forces acting between tip and sample in these two regimes, the oscillation amplitude is plotted as a function of the nearest tip–sample distance in Fig. 6. In addition, the bottom graph depicts the corresponding tip–sample force (cf. Fig. 4). The origin of the nearest tip–sample position D is defined by this force curve. Since both, the amplitude curves and the tip–sample force curve, are plotted as a function of the nearest tip–sample position, it is possible to identify the resulting maximum tip–sample interaction force for a given oscillation amplitude.

During the approach of the vibrating cantilever toward the sample surface, there is discontinuity in the nearest tip–sample position D. This gap corresponds to the bi-stability and the resulting jumps in the amplitude versus distance curve. After the jump from the attractive to the repulsive regime has occurred, the amplitude decreases continuously. The nearest tip–sample position, however, does not reduce accordingly, remaining roughly between -0.8 and -1.5 nm. As a result, larger A/A_0 ratios do not necessarily result into lower tip–sample interactions, which is important to keep in mind while adjusting imaging parameters in tapping mode. For practical applications, it is reasonable to assume that the set-point of the amplitude used for imaging has been set to a value between 90 % and 10 % of the free oscillation amplitude. With this condition, one can identify the accessible imaging regimes indicated by the horizontal (dashed) lines and the corresponding vertical (dotted) lines. Therefore, in tapping mode, two imaging regimes are typically accessible: the *tapping regime* (left) and the *bistable tapping regime* (middle). The bistable imaging state is only accessible during approach. Imaging in this regime is indeed possible with the limitation that the oscillating cantilever might jump into the repulsive regime [7, 8].

In the above paragraphs, the influence of the tip–sample interaction on the cantilever oscillation was analyzed, the maximum tip–sample interaction forces based on the assumption of a specific model force was calculated, and subsequently possible routes for image optimization were discussed. However, in practical imaging, the tip–sample interaction is not a priori known. Therefore, several authors [9–14] suggested solutions to this inversion problem, but some of them need further technical equipment. However, most commercial systems give access to amplitude and



AFM, Tapping Mode, Fig. 6 A comparison between the maximum tip–sample forces (tip–sample forces acting at the point of closest tip–sample approach/nearest tip–sample position D) experienced by conventional "tapping mode" AM-AFM assuming the same parameters as in Fig. 5. The *upper* graph shows the nearest tip–sample position D versus the actual oscillation amplitude A for tapping mode. The graph at the *bottom* reveals the force regimes sensed by the tip. The maximal tip–sample forces in tapping mode are on the repulsive (tapping regime) as well as attractive (bistable tapping regime) part of the tip–sample force curve

phase versus distance data which already allows the reconstruction of the tip–sample force curve applying numerical procedures [11, 12, 14].

The energy dissipation caused by the tip–sample interaction can be easily calculated using the conservation of energy principle [15].

$$\Delta E = \left(\frac{1}{Q_0}\frac{f_d}{f_0} + \frac{a_\mathrm{d}}{A}\sin\phi\right)\pi c_z A^2. \qquad (6)$$

Since all parameters – except the phase ϕ – are constant during scanning, this formula shows that the energy dissipation is roughly proportional to $\sin\phi$ as already mentioned in the caption of Fig. 2.

Cross-References

- ▶ AFM, Noncontact Mode
- ▶ Friction Force Microscopy
- ▶ Kelvin Probe Force Microscopy
- ▶ Magnetic Resonance Force Microscopy
- ▶ Piezoresponse Force Microscopy and Spectroscopy

References

1. Binnig, G., Quate, C.F., Gerber, C.: Atomic force microscopy. Phys. Rev. Lett. **56**, 930–933 (1986)
2. Garcia, R., Pérez, R.: Dynamic atomic force microscopy methods. Surf. Sci. Rep. **47**, 197–301 (2002)
3. Klinov, D., Maganov, S.: True molecular resolution in tapping-mode atomic force microscopy with high-resolution probes. Appl. Phys. Lett. **84**, 2697–2698 (2004)
4. Möller, C., Allen, M., Elings, V., Engel, A., Müller, D. J.: Tapping-mode atomic force microscopy produces faithful high-resolution images of protein surfaces. Biophys. J. **77**, 1150–1158 (1999)
5. Schwarz, U.D.: A generalized analytical model for the elastic deformation of an adhesive contact between a sphere and a flat surface. J. Coll. Interf. Sci. **261**, 99–106 (2003)

6. Hölscher, H., Schwarz, U.D.: Theory of amplitude modulation atomic force microscopy with and without Q-control. Int. J. Nonlinear Mech. **42**, 608–625 (2007)
7. Paulo, A.S., García, R.: High-resolution imaging of antibodies by tapping-mode atomic force microscopy: attractive and repulsive tip-sample interaction regimes. Biophys. J. **78**, 1599–1605 (2000)
8. Stark, R.W., Schitter, G., Stemmer, A.: Tuning the interaction forces in tapping mode atomic force microscopy. Phys. Rev. B **68**, 085401 (2003)
9. Stark, M., Stark, R.W., Heckl, W.M., Guckenberger, R.: Inverting dynamic force microscopy: from signals to time-resolved interaction forces. Proc. Natl. Acad. Sci. U. S. A. **99**, 8473–8478 (2002)
10. Legleiter, J., Park, M., Cusick, B., Kowalewski, T.: Scanning probe acceleration microscopy (SPAM) in fluids: mapping mechanical properties of surfaces at the nanoscale. Proc. Natl. Acad. Sci. U. S. A. **103**, 4813–4818 (2006)
11. Lee, M., Jhe, W.: General solution of amplitude-modulation atomic force microscopy. Phys. Rev. Lett. **97**, 036104 (2006)
12. Hölscher, H.: Quantitative measurement of tip-sample interactions in amplitude modulation atomic force microscopy. Appl. Phys. Lett. **89**, 123109 (2006)
13. Sahin, O., Maganov, S., Chanmin, S., Quate, C., Solgaard, O.: An atomic force microscope tip designed to measure time-varying nanomechanical forces. Nat. Nanotechnol. **2**, 507–514 (2007)
14. Shuiqing, H., Raman, A.: Inverting amplitude and phase to reconstruct tip-sample interaction forces in tapping mode atomic force microscopy. Nanotechnology **19**, 375704 (2008)
15. Cleveland, J.P., Anczykowski, B., Schmid, A.E., Elings, V.B.: Energy dissipation in tapping-mode atomic force microscopy. Appl. Phys. Lett. **72**, 2613 (1998)

Aging

► Fate of Manufactured Nanoparticles in Aqueous Environment

ALD

► Atomic Layer Deposition

Alteration

► Fate of Manufactured Nanoparticles in Aqueous Environment

Alum

► Aluminum-Based Nano-adjuvants

Aluminum Salts

► Aluminum-Based Nano-adjuvants

Aluminum Vaccine Adjuvant

► Aluminum-Based Nano-adjuvants

Aluminum-Based Nano-adjuvants

Bingbing Sun[1], Zhaoxia Ji[2] and Tian Xia[1,2]
[1]Division of NanoMedicine, Department of Medicine, University of California, Los Angeles, CA, USA
[2]California NanoSystems Institute, University of California, Los Angeles, CA, USA

Synonyms

Alum; Aluminum salts; Aluminum vaccine adjuvant

Definition

Aluminum-based nano-adjuvants are insoluble aluminum salts used in human and animal vaccines for the effective stimulation of the immune system against infectious diseases.

Introduction

Vaccination remains one of the most effective tools to prevent infectious disease [1–3], and adjuvants based on aluminum salts are commonly

used in vaccines to provide an efficient and long-lasting protective immunity against infectious agents. In 1926, Glenny et al. demonstrated that injection of aluminum potassium sulfate with diphtheria toxoid could induce a stronger antibody response than soluble toxoid in guinea pigs [4]. Since then, aluminum-based adjuvants have been widely used in both human and veterinary vaccination against infectious diseases [5] and are being evaluated for immunotherapy against cancer. Aluminum-based adjuvants generally induce a strong Th2 immune response, with little or no Th1 immune response. While it is known that aluminum adjuvants tightly bind to antigens and provide slow release of antigens as a result of these materials' astringent characteristics, the specific mechanistic details of how the physicochemical properties of these materials lead to immunopotentiation has just begun to emerge [6].

Aluminum-Based Nano-adjuvants in Licensed Vaccines

Aluminum-based vaccine adjuvants have been in routine use in human vaccination against various infectious diseases including DTaP (diphtheria, tetanus, acellular pertussis), human papillomavirus and penumococcal disease, hepatitis A, and hepatitis B for over 80 years (Table 1) [7]. Depending on the commercial source, they may be composed of aluminum hydroxide (Alhydrogel), aluminum phosphate (Adju-Phos), aluminum potassium sulfate (Alum), aluminum hydroxyphosphate sulfate (AHSA), or a mixture of aluminum and magnesium hydroxides (Imject Alum, commonly used for experimental and basic immunological studies) (Table 1) [5, 7, 8]. These commercially available aluminum salts have different physicochemical properties [6]. For example, aluminum hydroxides are needlelike particles with diameters of 2 nm, while aluminum phosphates are platelike particles with a primary size of 50 nm. However, their names sometimes do not correctly describe their adjuvant structures. X-ray diffraction analysis, energy-dispersive spectroscopy and infrared spectroscopy have identified aluminum hydroxide adjuvant as crystalline aluminum oxyhydroxide, AlOOH; and aluminum phosphate is determined as amorphous aluminum hydroxyphosphate, $Al(OH)_x(PO_4)_y$ [6]. The amount of aluminum adjuvants used in the currently licensed vaccines varies from 0.85 to 0.125 mg per dose in the USA, and a 1-year-old baby who receives a complete series of recommended vaccines may have 1.6–4.1 mg of aluminum [7]. Although there are concerns on the high dose of aluminum adjuvants in vaccines for human use, they have been showing a safe profile up to date [9].

Mechanisms of Aluminum-Based Nano-adjuvant-Induced Immune Responses

Despite of the wide usage for over 80 years as adjuvants in vaccines, the detailed mechanisms of how aluminum-based adjuvants enhance adaptive immunity and exert their adjuvant effect remain unclear, and several mechanisms are proposed (Fig. 1). Firstly, upon mixing, antigens may be retained at the injection site either by being adsorbed by the aluminum containing adjuvant or by being trapped in the void spaces within the adjuvant aggregates to form a local higher concentration of antigen, which could allow retaining and a slower release of antigen in the body for prolonged and effective stimulation of the immune system, an effect referred to as the "depot effect" [10]. However, the depot effect has been challenged, with the evidence showing that the excision of injection site did not interfere with the development of immune response [11]. Recently, it has been shown that aluminum adjuvants could activate the nucleotide binding and oligomerization, leucine-rich repeat, pyrin domain-containing 3 (NLRP3) inflammasome, and IL-1β production, which could explain its ability to induce local inflammation, recruitment of antigen-presenting cells (APCs), enhanced antigen uptake, dendritic cell (DC) maturation, and stimulation of T-cell activation and differentiation [12, 13]. However, there is disagreement about the necessity of this pro-inflammatory response pathway in generating adjuvant effects in vivo [14], with NLRP3-deficient mice

Aluminum-Based Nano-adjuvants, Table 1 Types of alum salts used in licensed vaccines [7, 8]

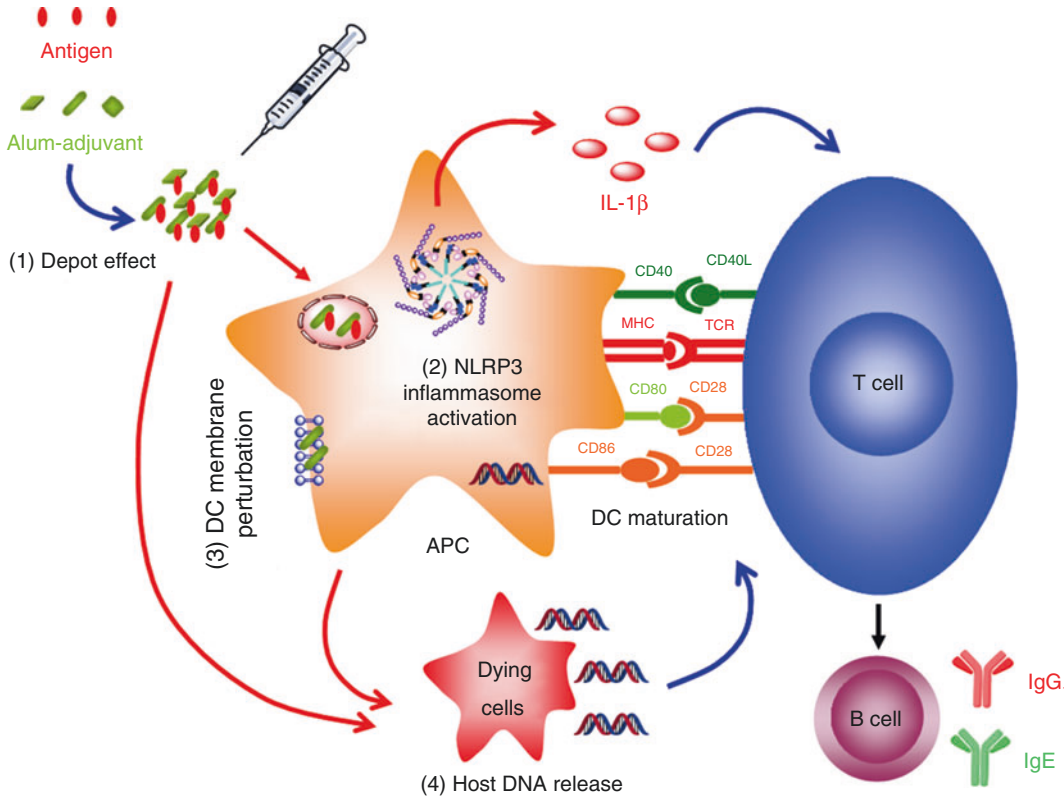
Vaccine (trade name)	Manufacturer	Type of aluminum salts
Anthrax (BioThrax)	Emergent BioSolutions, Inc.	Aluminum hydroxide
DT (Sanofi)	Sanofi Pasteur Limited	Aluminum potassium sulfate
DTaP (Daptacel)	Sanofi Pasteur Limited	Aluminum phosphate
DTaP-IPV (Kinrix)	GlaxoSmithKline Biologicals	Aluminum hydroxide
DTaP-HepB-IPV (PEDIARIX)	GlaxoSmithKline Biologicals	Aluminum hydroxide, aluminum phosphate
DTaP-IPV/Hib (Pentacel)	Sanofi Pasteur Limited	Aluminum phosphate
Hib (PedvaxHIB)	Merck & Co., Inc	Aluminum hydroxyphosphate sulfate
Hib/Hep B (Comvax)	Merck & Co., Inc	Aluminum potassium sulfate
Hep A (Havrix)	GlaxoSmithKline Biologicals	Aluminum hydroxide
Hep B (Engerix-B)	GlaxoSmithKline Biologicals	Aluminum hydroxide
Hep B (Recombivax)	Merck & Co., Inc	Aluminum hydroxyphosphate sulfate
Hep A/Hep B (Twinrix)	GlaxoSmithKline Biologicals	Aluminum hydroxide, aluminum phosphate
Human papillomavirus (HPV) (Cervarix)	GlaxoSmithKline Biologicals	Aluminum hydroxide
Human papillomavirus (HPV) (Gardasil)	Merck & Co., Inc	Aluminum hydroxyphosphate sulfate
Japanese encephalitis (IXIARO)	Intercell Biomedical Ltd.	Aluminum hydroxide
Pneumococcal (PCV13-Prevnar 13)	Wyeth Pharmaceuticals, Inc.	Aluminum phosphate
Td (Decavac)	Sanofi Pasteur Limited	Aluminum potassium sulfate
Td (Tenivac)	Sanofi Pasteur Limited	Aluminum phosphate
Td (MassBiologics)	MassBiologics	Aluminum phosphate
Tdap (Adacel)	Sanofi Pasteur Limited	Aluminum phosphate
Tdap (Boostrix)	GlaxoSmithKline Biologicals	Aluminum hydroxide

AP acellular pertussis, *D* diphtheria, *Hep A* hepatitis A, *Hep B* hepatitis B, *Hib* Haemophilus influenzae type b, *IPV* inactivated poliovirus, *T* tetanus

mounting a normal OVA-specific IgG_1 response [15]. Studies have demonstrated that aluminum hydroxide could induce cytotoxicity and host DNA release at sites of immunization [16, 17], which mediate the adjuvant effect for Th2-biased adaptive immune responses (Fig. 1). In addition, it has been shown that aluminum adjuvants are able to perturb the DC membrane as a way of providing immune stimulation [18]. One possible reason for these discrepancies is the existence of different adjuvant mechanisms by which aluminum adjuvants boost in vivo immune responses. Also, different types and/or structures of aluminum adjuvants were used in these studies. An adjuvant with rigorously controlled structure is needed to address the controversy regarding the mechanisms of the adjuvants' immunopotentiating effects in vivo.

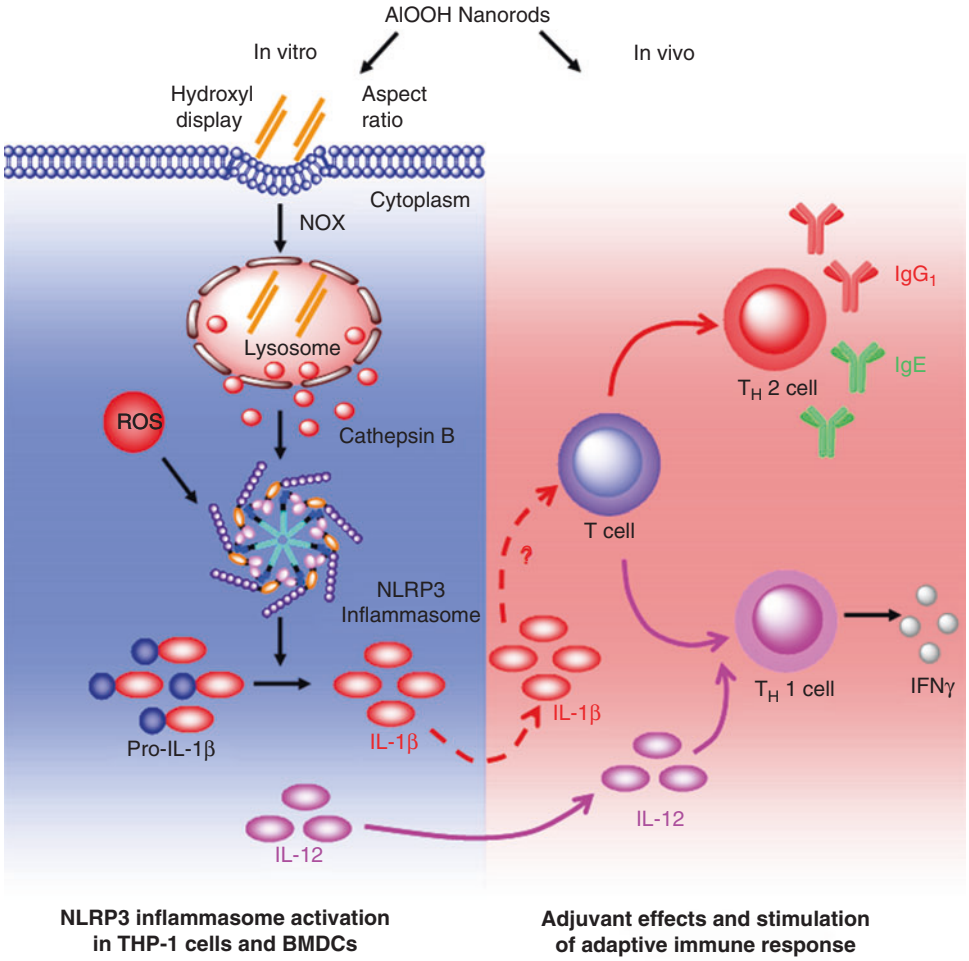
Engineered AlOOH Nanorods as Vaccine Adjuvants

A recent study sheds light on the detailed mechanism of aluminum adjuvant-induced immune responses using engineered aluminum oxyhydroxide (γ-AlOOH, boehmite) nanoparticles [19]. In this study, AlOOH nanoparticles were synthesized with precisely controlled shape, crystallinity, and hydroxyl content, which have been shown to play important roles in the induction of immune responses.



Aluminum-Based Nano-adjuvants, Fig. 1 Mechanisms of aluminum-based nano-adjuvant-induced immune responses, including (*1*) the depot effect, (*2*) NLRP3 inflammasome activation, (*3*) DC membrane perturbation, and (*4*) host DNA release, which promote DC maturation and enhance adaptive immune responses

AlOOH nanorods with higher hydroxyl contents lead to increased reactive oxygen species (ROS) generation that induced stronger NLRP3 inflammasome activation and IL-1β production compared to polyhedron- and plate-shaped particles as well as commercial Imject Alum in both THP-1 macrophages and bone marrow-derived dendritic cells (BMDCs) in vitro (Fig. 2) [19]. Interestingly, AlOOH nanorods were also able to induce the maturation of BMDCs as indicated by the expression of surface maturation markers including MHC II, CD80, and CD86. The in vitro results were corroborated by in vivo studies using ovalbumin (OVA) as a model antigen. It was shown that vaccination with AlOOH nanorods bounded with OVA exhibited higher antibody production including OVA-specific IgG_1 and IgE in mice compared to commercial Imject Alum. Adoptive transfer was also conducted to confirm the role of DCs in which ex vivo activation of BMDC by AlOOH nanorods could induce a stronger adjuvant effect than Alum in intact animals. It is interesting to note that in addition to the boosting of IL-1β production, AlOOH nanorods could also increase IL-12 production in BMDCs, which suggests that it is possible to induce Th1 cytotoxic immune responses including the generation of CD4- and CD8-positive T cells and the production of IFN-γ by these particles (Fig. 2) [19]. These results show that the engineered design of aluminum adjuvants in combination with property-activity analysis can be used to design more potent aluminum-based adjuvants to induce strong Th2 and potential Th1 immune responses. In addition, it is possible to use AlOOH nanorods as reference



Aluminum-Based Nano-adjuvants, Fig. 2 Engineered AlOOH nanorods as vaccine adjuvants

materials to study whether the paradoxical outcomes in different mechanistic studies could be due to the differences in the physicochemical properties of various aluminum-based materials that were used previously.

Future Directions

Aluminum vaccine adjuvants can generally induce strong Th2 immune responses with little or no Th1 immunity. However, Th1 cytotoxic responses are needed for clearance of viral infections and tumor cells. To achieve the goal of activating Th1 responses, additional stimuli including Toll-like receptor (TLR) agonists are needed. Triggering of certain TLRs (e.g., TLR4, 7, and 9) could stimulate the production of cytokines/chemokines and type I interferons (IFNs) that promote the immune system to activate Th1 responses to eliminate exogenous pathogens. Clinical trials involving individual TLRs suggest that TLR ligands can be safe and effective vaccine adjuvants, with vaccines already licensed in the USA and Europe containing such ligands [20]. Therefore, using engineered nanoparticles including AlOOH nanorods together with TLR agonists, it is possible to develop combinatorial

vaccine adjuvants to direct the T-cell differentiation toward Th1 cytotoxic responses for more effective prevention of infectious diseases and treatment of cancer in clinics.

Cross-References

► Nanoparticles

References

1. Pulendran, B., Ahmed, R.: Translating innate immunity into immunological memory: implications for vaccine development. Cell **124**, 849–863 (2006)
2. Ogra, P.L., Faden, H., Welliver, R.C.: Vaccination strategies for mucosal immune responses. Clin. Microbiol. Rev. **14**, 430–445 (2001)
3. Whitney, C.G., Farley, M.M., Hadler, J., Harrison, L. H., Bennett, N.M., Lynfield, R., et al.: Decline in invasive pneumococcal disease after the introduction of protein-polysaccharide conjugate vaccine. N. Engl. J. Med. **348**, 1737–1746 (2003)
4. Glenny, A.T., Pope, C.G., Waddington, H., Wallace, U.: Immunological notes XVLL.-XXIV. J. Pathol. Bacteriol. **29**, 31–40 (1926)
5. HogenEsch, H.: Mechanism of immunopotentiation and safety of aluminum adjuvants. Front Immunol. **3**, 1–13 (2013)
6. Hem, S.L., HogenEsch, H.: Relationship between physical and chemical properties of aluminum-containing adjuvants and immunopotentiation. Expert Rev. Vaccines **6**, 685–698 (2007)
7. Baylor, N.W., Egan, W., Richman, P.: Aluminum salts in vaccines – US perspective. Vaccine **20**, S18–S23 (2002)
8. http://www.cdc.gov/vaccines/vac-gen/additives.htm
9. Lindblad, E.B.: Aluminium adjuvants – in retrospect and prospect. Vaccine **22**, 3658–3668 (2004)
10. Glenny, A.T., Buttle, G.A.H., Stevens, M.F.: Rate of disappearance of diphtheria toxoid injected into rabbits and guinea-pigs: toxoid precipitated with alum. J. Pathol. Bacteriol. **34**, 267–275 (1931)
11. Holt, L.: Developments in Diphtheria Prophylaxis. Heinemann, London (1950)
12. Eisenbarth, S.C., Colegio, O.R., O'Connor Jr., W., Sutterwala, F.S., Flavell, R.A.: Crucial role for the Nalp3 inflammasome in the immunostimulatory properties of aluminium adjuvants. Nature **453**, 1122–1126 (2008)
13. Sun, B., Wang, X., Ji, Z., Li, R., Xia, T.: NLRP3 inflammasome activation induced by engineered nanomaterials. Small **9**, 1595–1607 (2013)
14. Franchi, L., Nunez, G.: The Nlrp3 inflammasome is critical for aluminium hydroxide-mediated IL-1 beta secretion but dispensable for adjuvant activity. Eur. J. Immunol. **38**, 2085–2089 (2008)
15. Kool, M., Pétrilli, V., De Smedt, T., Rolaz, A., Hammad, H., van Nimwegen, M., et al.: Cutting edge: alum adjuvant stimulates inflammatory dendritic cells through activation of the NALP3 inflammasome. J. Immunol. **181**, 3755–3759 (2008)
16. McKee, A.S., Burchill, M.A., Munks, M.W., Jin, L., Kappler, J.W., Friedman, R.S., et al.: Host DNA released in response to aluminum adjuvant enhances MHC class II-mediated antigen presentation and prolongs CD4 T-cell interactions with dendritic cells. Proc. Natl. Acad. Sci. U. S. A. **110**, E1122–E1131 (2013)
17. Marichal, T., Ohata, K., Bedoret, D., Mesnil, C., Sabatel, C., Kobiyama, K., et al.: DNA released from dying host cells mediates aluminum adjuvant activity. Nat. Med. **17**, 996–1002 (2011)
18. Flach, T.L., Ng, G., Hari, A., Desrosiers, M.D., Zhang, P., Ward, S.M., et al.: Alum interaction with Dendritic cell membrane lipids is essential for its adjuvanticity. Nat. Med. **17**, 479–487 (2011)
19. Sun, B., Ji, Z., Liao, Y.-P., Wang, M., Wang, X., Dong, J., et al.: Engineering an effective immune adjuvant by designed control of shape and crystallinity of aluminum oxyhydroxide nanoparticles. ACS Nano **7**, 10834–10849 (2013)
20. Steinhagen, F., Kinjo, T., Bode, C., Klinman, D.M.: TLR-based immune adjuvants. Vaccine **29**, 3341–3355 (2011)

Amphibian Larvae

► Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae

Amplitude-Phase-Distance (APD) Curves

► Dynamic Force Measurements on Modified Surfaces with AFM

Anamorphic Particle Image Tracking

► Astigmatic Microparticle Imaging

A

Anchor Loss in MEMS/NEMS

Zhili Hao
Department of Mechanical and Aerospace Engineering, Old Dominion University, Norfolk, VA, USA

Synonyms

Clamping loss; Support loss

Definition

Anchor loss is part of vibration energy in a MEMS/NEMS resonator that is dissipated through its anchor/anchors.

Background

Generally speaking, a MEMS/NEMS resonator is comprised of two fundamental elements: a micro/nano-mechanical structure and its integrated transducers for converting a mechanical signal to an electrical one or vice versa. Electrostatic transducers and piezoelectric transducers are the two types of transducers commonly employed for MEMS/NEMS resonators. In most cases, anchor loss is of great concern to a resonator with electrostatic transducers and thus needs to be examined. In contrast, interfacial damping is significant in a resonator with piezoelectric transducers, while its anchor loss is not a concern. Therefore, only anchor loss in a resonator with electrostatic transducers is examined here.

The mechanical behavior of a MEMS/NEMS resonator can be described by the second-order mechanical system. By operating a MEMS/NEMS resonator in one of its vibration modes, many functions have been realized, including sensing acceleration and rotation rate [1], processing wireless signals, and serving as a time/frequency keeper [2, 3]. From the mechanical design perspective, two key design parameters for a MEMS/NEMS resonator are the resonant frequency and the mechanical quality factor (Q). While the resonant frequency is application dependent, a high Q is highly desirable in most cases, simply because a high Q in these resonators directly translates to high performance of the final resonator products, including high sensitivities, better bias stability, and improved frequency selectivity.

Owing to its small size, it is feasible to package a MEMS/NEMS resonator in vacuum and thereby eliminate air damping. Then, anchor loss, thermoelastic damping (TED), and surface loss come to the fore and determine the Q of a resonator. The overall measured Q of a resonator operating in vacuum is given by [4]:

$$\frac{1}{Q_{\text{measured}}} \cong \frac{1}{Q_{\text{anchor}}} + \frac{1}{Q_{\text{TED}}} + \frac{1}{Q_{\text{surface}}} \tag{1}$$

Physical Mechanism

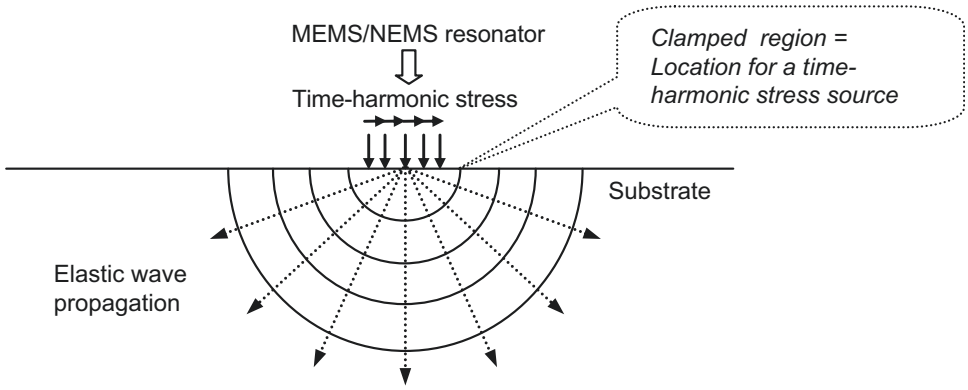
In a MEMS/NEMS resonator, its micro/nano-mechanical structure is fixed on a substrate. During its vibrations, the structure exerts a time-harmonic stress on the substrate through its clamped region (Fig. 1). Acting as an excitation source, this time-harmonic stress will excite elastic waves propagating into the substrate. Part of the vibration energy of the structure, which is dissipated through elastic wave propagation into the substrate, is commonly referred to as anchor loss [5–7]. Then, the Q related to anchor loss of a resonator is defined as:

$$Q_{\text{anchor}} = 2\pi \frac{W}{\Delta W_{\text{anchor}}} \tag{2}$$

where ΔW_{anchor} denotes the energy dissipated through the anchor per cycle of vibration and W denotes the maximum vibration energy stored per cycle of vibration in a resonator.

Theory

Anchor loss involves both the micro/nano-mechanical structure and the substrate, with the



Anchor Loss in MEMS/NEMS, Fig. 1 Physical mechanism of anchor loss in a MEMS/NEMS resonator – elastic wave propagation in the substrate with the clamped region as the excitation source [5]

latter being extremely large in relation to the former. Thus, the difficulty in calculating anchor loss lies in the fact that unlike the micro/nano-mechanical structure, the substrate is virtually a semi-infinite medium and cannot be easily incorporated into finite element analysis (FEA) software [5]. In order to calculate anchor loss in a resonator at the design stage, numerical algorithms have been developed that tackle the semi-infinite geometry of the substrate. However, such numerical algorithms are not only extremely computationally intensive but also not efficient at providing insights on how anchor loss is affected by resonator design parameters so as to modify a resonator design for achieving a higher Q. With the goal of providing the reader with the general concept of how anchor loss is calculated, the related theoretical analysis is described here.

To gain insights on how anchor loss in a resonator varies with resonator design parameters, a separation-and-transfer method has been developed to calculate anchor loss [5–7]. The two fundamental assumptions underlying this method are given below:

1. There is negligible interference between the elastic vibrations in a resonator and the elastic wave propagation in its substrate.
2. All the vibration energy of a resonator entering the substrate is not returned to the resonator (i.e., no reflection from the substrate boundaries) and therefore is considered to be lost.

Accordingly, the micro/nano-mechanical structure and the substrate are separated first, and then the time-harmonic stress from the structure is transferred to the substrate for elastic wave excitation through the clamped region. Thus, anchor loss is formulated as the integral of the product of the time-harmonic stress from the structure and the time-harmonic displacement on the substrate, across the clamped region over one cycle of vibration:

$$\Delta W_{\text{anchor}} = \pi \cdot \int_{\text{clamped region}} \text{stress} \times \text{displacement} \tag{3}$$

where π is due to the time-harmonic nature of vibrations; stress and displacement represent the amplitudes of the two time-harmonic parameters. To calculate anchor loss in a resonator, one needs to calculate the time-harmonic stress at the clamped region and the time-harmonic displacement on the substrate across the clamped region. Once the ΔW_{anchor} is known, the Q_{anchor} can be obtained using Eq. 2.

Based on the first assumption, the classic vibration theory can be utilized to describe the elastic vibrations in a resonator, where the clamped region has zero displacement but suffers from a time-harmonic stress. As to a resonator taking a regular geometry, a closed-form expression for the time-harmonic stress at the clamped region can be obtained, as well as the maximum vibration

energy stored in the resonator. In contrast, as to a resonator taking an irregular geometry, FEA software is needed to obtain the time-harmonic stress at the clamped region and the maximum vibration energy stored in the resonator. The second assumption translates to treating the substrate as a semi-infinite medium. Then, one can utilize the large body of literature on the investigation of the time-harmonic displacement on a semi-infinite or infinite elastic medium [8], which were originally developed for other applications (e.g., earthquake and acoustic emission) over the past century, to find out the time-harmonic displacement on the substrate, corresponding to the time-harmonic stress from the vibrations of a resonator. To illustrate how anchor loss is calculated using the separation-and-transfer method, an example on anchor loss in a cantilever beam resonator is given below [5].

Analytical Model of Anchor Loss in a Cantilever Beam Resonator

A cantilever beam resonator is fixed at one end (Fig. 2a). The length, width, and thickness of the resonator are denoted by L, b, and h, respectively. It operates in one of its in-plane (x-y plane) flexural modes. It is assumed that the dimensions of the substrate in the x-y plane are much larger than those of the beam resonator. For simplicity, it is also assumed that both the beam and the substrate are made of the same isotropic and homogeneous material. The in-plane flexural vibrations of the beam resonator are governed by the following equation:

$$\frac{\partial^4 y}{\partial x^4} = -\frac{\rho \cdot S}{E \cdot I} \cdot \frac{\partial^2 y}{\partial t^2} \tag{4}$$

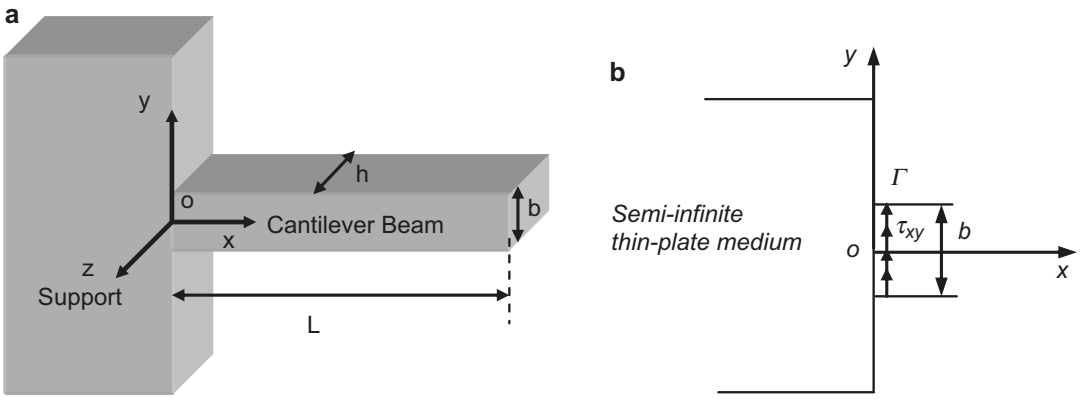
where E and ρ denote Young's modulus and density of the beam; I and S are the moment of inertia and cross-section area of the beam, respectively. When this beam resonator undergoes time-harmonic vibrations, one can assume:

$$y(x,t) = Y(x) \cdot e^{-\omega \cdot t} \tag{5}$$

where ω denotes the angular frequency of the vibrations. The angular frequency, ω, of the beam resonator is:

$$\omega = \frac{\pi^2 \cdot \beta^2}{L^2} \cdot \sqrt{\frac{E \cdot I}{\rho \cdot S}} \tag{6}$$

where β is the mode constant. The vibration mode shape of the resonator is:



Anchor Loss in MEMS/NEMS, Fig. 2 (**a**) Schematic of a cantilever beam resonator [6]; (**b**) the semi-infinite thin-plate medium of thickness, h, subjected to a time-harmonic shear stress from the cantilever beam resonator [5] (drawn out of proportion for clear illustration)

$$Y(x) = \frac{U}{2}\left\{\cosh\left(\beta\pi \cdot \frac{x}{L}\right) - \cos\left(\beta\pi \cdot \frac{x}{L}\right) + \chi \cdot \left(\sinh\left(\beta\pi \cdot \frac{x}{L}\right) - \sin\left(\beta\pi \cdot \frac{x}{L}\right)\right)\right\} \tag{7}$$

where $U/2$ denotes the vibration amplitude and χ denotes the mode shape factor.

Accordingly, the stored maximum vibration energy in the resonator can be expressed as:

$$W = \frac{1}{8}\rho \cdot S \cdot L \cdot \omega^2 \cdot U^2 \tag{8}$$

Through its clamped end, this beam resonator exerts a vibrating shear force, Γ, on its substrate over the clamped region, $x = 0, |y| \le b/2$ (Fig. 2b):

$$\Gamma = E \cdot I \cdot U \cdot \left\{\frac{\pi \cdot \beta}{L}\right\}^3 \cdot \chi \tag{9}$$

This shear force is assumed to be uniformly distributed across the clamped region, acting as a source to excite elastic waves propagating into the substrate. Hence, the excitation shear stress source, τ_{xy}, at the clamped end becomes:

$$\tau_{xy} = \frac{\Gamma}{b \cdot h}, \ |y| \le b/2; \ \tau = 0, \ |y| > b/2 \tag{10}$$

The substrate is treated as a semi-infinite thin-plate medium of thickness, h, and the elastic waves, excited by the shear stress source, are governed by the 2D in-plane elastic wave equations:

$$\frac{\partial^2 u_x}{\partial t^2} = c_L^2 \cdot \frac{\partial^2 u_x}{\partial x^2} + c_T^2 \cdot \frac{\partial^2 u_x}{\partial y^2} + \left(c_L^2 - c_T^2\right) \cdot \frac{\partial^2 u_y}{\partial x \partial y} \tag{11a}$$

$$\frac{\partial^2 u_y}{\partial t^2} = c_L^2 \cdot \frac{\partial^2 u_y}{\partial y^2} + c_T^2 \cdot \frac{\partial^2 u_y}{\partial x^2} + \left(c_L^2 - c_T^2\right) \cdot \frac{\partial^2 u_x}{\partial x \partial y} \tag{11b}$$

where u_x and u_y are the displacements along the x-axis and y-axis, respectively. The propagation velocities for longitudinal (c_L) and transverse waves (c_T) are given by:

$$c_L^2 = \frac{E}{\rho \cdot (1 - \upsilon^2)}, \ c_T^2 = \frac{E}{2\rho \cdot (1 + \upsilon)} \tag{12}$$

The Poisson's ratio of the thin plate is denoted by υ. The shear stress source, Eq. 10, serves as the boundary condition for the two governing equations, Eq. 11. Through the related mathematical manipulations on Eqs. 10 and 11, the average displacement amplitude along the y-axis (shear stress direction) over the clamped region can be obtained:

$$\begin{aligned} v_{x=0,|y|\le b/2} &= \frac{4 \cdot b \cdot \tau}{\pi \cdot E} \cdot \frac{1+\upsilon}{1-\upsilon} \cdot \Psi \text{ with } \Psi \\ &= \int_0^\infty \frac{\sqrt{\zeta^2 - (c_L/c_T)^2}}{\left\{2\zeta^2 - (c_L/c_T)^2\right\}^2 - 4\zeta^2 \cdot \sqrt{\zeta^2 - (c_L/c_T)^2} \cdot \sqrt{\zeta^2 - 1}} \cdot d\zeta \end{aligned} \tag{13}$$

where $\zeta = \xi c_L/\omega$, with ξ being the variable of the Fourier transform. The imaginary part of the integral of (13) contributes to anchor loss and is numerically calculated to be 0.336 (assuming $\upsilon = 0.28$).

After the average displacement amplitude incurred by the shear stress over the clamped region is determined, the amount of energy loss per cycle from the anchor of the beam resonator can be explicitly calculated as:

$$\Delta W_{\text{anchor}} = \pi \cdot \Gamma \cdot v_{x=0,|y|\le b/2} \tag{14}$$

By combining Eqs. 2, 8, and 14 a closed-form expression for the Q_{anchor} of a cantilever beam resonator is obtained:

$$Q_{\text{anchor}} = \left[\frac{0.24 \cdot (1-\upsilon)}{(1+\upsilon)\cdot \Psi}\right] \cdot \frac{1}{(\beta \cdot \chi)^2} \cdot \left[\frac{L}{b}\right]^3 \quad (15)$$

This expression clearly relates the Q_{anchor} of a cantilever beam resonator to its design parameters, indicating that a high ratio of length/width yields a high Q_{anchor}. Finally, it is worth mentioning that closed-form expressions for the time-harmonic displacement on different substrate configurations of MEMS/NEMS resonators under varying time-harmonic stress sources are available in the literature [5, 8].

Sometimes, a resonator, such as a tuning-fork gyroscope [1], is relatively large, and then the substrate cannot be effectively treated as a semi-infinite medium. Meanwhile, depending on how the substrate is mounted on its surroundings, its boundaries may reflect part of dissipated vibration energy into the resonator. Furthermore, the fabrication technology employed could severely affect how a resonator is fixed on a substrate through its anchor/anchors [9] and vary the time-harmonic stress at its anchor/anchors [10] and thus plays a critical role in determining anchor loss. As such, a theoretically calculated Q_{anchor} is perhaps best suited for providing the design guidelines for improving the Q_{anchor} of a resonator, rather than accurately estimating its value.

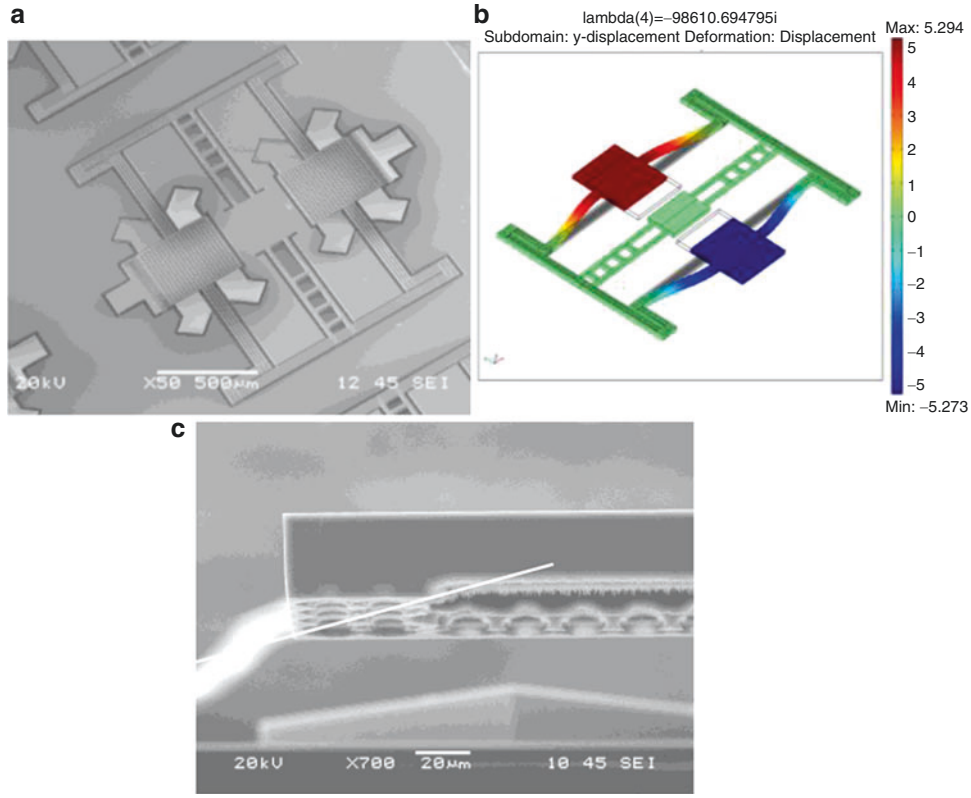
Key Research Findings

While surface loss is not much relevant to a resonator design, both anchor loss and TED can vary dramatically with a resonator design [6]. Thus, at the resonator design stage, both anchor loss and TED need to be examined for achieving a high Q. Different from anchor loss, TED is confined within a resonator and thus imposes the fundamental limit on the achievable Q of a resonator. As compared with anchor loss, TED is relatively easy to calculate with great accuracy in FEA software [11], in the sense that fabrication variations do not affect the calculated Q_{TED} in a resonator as much as the calculated Q_{anchor}. Accumulated from the analytical and experimental studies on both TED and anchor loss, the following two conclusions have been well established:

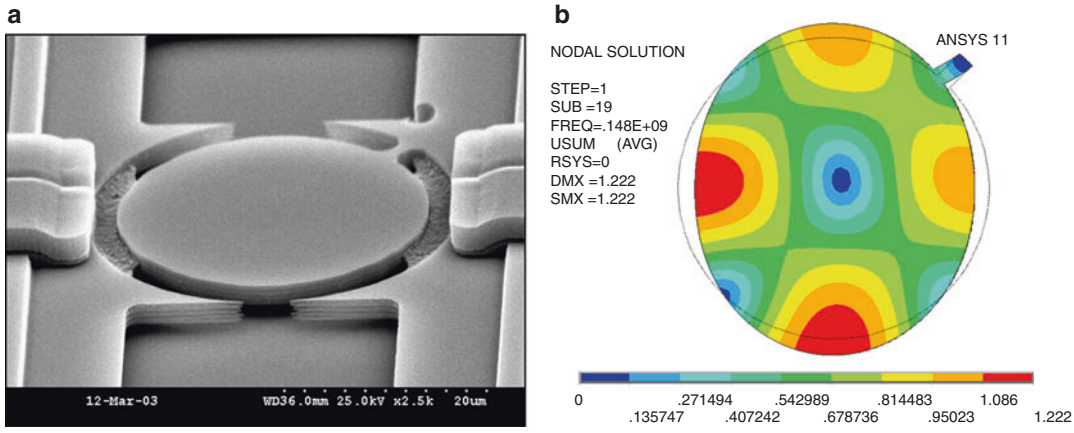
1. As to a resonator operating in one of its bulk modes, anchor loss is the sole dominant loss, while TED is completely negligible.
2. As to a resonator operating in one of its flexural modes, both anchor loss and TED could be dominant losses, depending on the resonator design and fabrication variations.

According to the theory of anchor loss, the essence for improving the Q_{anchor} of a resonator is to minimize the reaction force (or time-harmonic stress) at the clamped region of a resonator and minimize the displacement on the substrate across the clamped region. It is worth mentioning that the displacement on the substrate is caused by the reaction force and thus is proportional to the reaction force. Then, reducing the reaction force at the clamped region will coincidently reduce the displacement on the substrate.

To minimize the reaction force in a resonator, two common strategies are (1) to design a symmetrical mechanical structure operating in a vibration mode that is symmetric to its anchor and (2) to fix a resonator to one of its resonant nodes through a slim support beam. An example of the first strategy is a multiple-beam tuning-fork gyroscope (Fig. 3). In this gyroscope, a symmetric multiple-beam tuning-fork structure is fixed at its center and one of its operation modes is symmetric to the anchor. An example of the second strategy is a disk resonator operating in its elliptic bulk mode (Fig. 4) [12, 13]. In theory, a resonant node is a point with no dimension, and thus a support beam should be designed as slim as possible. Meanwhile, care needs to be taken to ensure that adding a support beam does not affect the operation mode of the main structure (disk). To minimize the displacement on the substrate across the anchor region, one strategy is to use a substrate with a higher elasticity so that a reaction force of the same amplitude gives rise to a smaller displacement on the substrate. Alternatively, using a



Anchor Loss in MEMS/NEMS, Fig. 3 (**a**) Scanning electron microscope (*SEM*) picture of a multiple-beam tuning-fork gyroscope [9]; (**b**) one of the operation modes of the gyroscope symmetric to its anchor [9].(**c**) SEM picture of the multiple-beam tuning-fork structure undermined by fabrication variations [9]



Anchor Loss in MEMS/NEMS, Fig. 4 (**a**) SEM picture of a disk resonator [13]; (**b**) the elliptic bulk mode of the disk resonator fixed at one of its resonant nodes through a slim support beam to the substrate [12]

material with high elasticity for a resonator will increase the maximum vibration energy in the resonator and consequently increase the Q_{anchor}, although not necessarily reducing anchor loss [14].

Although the abovementioned design strategies for improving the Q_{anchor} have been well established, practical obstacles for implementing them arise mainly from the fabrication technologies employed to produce MEMS/NEMS resonators. For instance, the multiple-beam tuning-fork structure can be designed with a perfect symmetric structure, but unavoidable fabrication variations render a fabricated gyroscope with a non-perfect symmetric structure (Fig. 3). Such a non-perfect symmetric structure gives rise to a relatively large reaction force at the anchor and consequently a relatively low Q_{anchor}. Moreover, fabrication variations vary with individual resonators of identical design and made from the same fabrication run. When anchor loss is a dominant loss in a resonator, the measured Qs of these individual resonators can vary dramatically. Thus, for those resonators with anchor loss being a dominant loss, fabrication technologies employed play a critical role in determining their ultimate Q, which might be quite different from the calculated value. Although electrostatic tuning has been introduced for compensating for the unsymmetry in a fabricated structure, implementing this compensation strategy has proven to be extremely challenging. Similarly, owing to fabrication variations, a support beam is not perfectly located at a resonant node, yielding a relatively large reaction force. Silicon is a commonly used building material for a resonator and its substrate, owing to its low cost and mature fabrication technologies. Replacing silicon with other materials, such as diamond, would lead to high cost of final resonator products. As such, minimizing the reaction force of a resonator is perhaps the most cost-effective and practical strategy to improve the Q_{anchor} of a MEMS/NEMS resonator.

Future Directions of the Field

Currently, the physical mechanism of anchor loss is well understood and the strategies for improving the Q_{anchor} of a MEMS/NEMS resonator are well established. The main research focus has so far been on fabrication technology development for reducing fabrication variations among fabricated MEMS/NEMS resonators so as to reduce performance variations and the Q_{anchor} variations among them and replacing silicon with other desirable materials for increasing the Q_{anchor} of MEMS/NEMS resonators.

References

1. Alexander, A., Trusov, A.: Micromachined rate gyroscope architecture with ultra-high quality factor and improved mode ordering. Sensor. Actuat. A **165**, 26–34 (2011)
2. Hao, Z.: VHF single crystal silicon elliptic bulk-mode capacitive disk resonators part I: design and modeling. J. Microelectromech. Syst. **13**, 1043–1053 (2004)
3. Wang, K.: VHF free–free beam high-Q micromechanical resonators. J. Microelectromech. Syst. **9**, 347–360 (2000)
4. Yasumura, K.: Quality factors in micron- and submicron-thick cantilevers. J. Microelectromech. Syst. **9**, 117–125 (2000)
5. Hao, Z.: An analytical model for support loss in micromachined beam resonators with in-plane flexural vibrations. Sensor. Actuat. A **109**, 156–164 (2003)
6. Hao, Z.: Support loss in the radial bulk-mode vibrations of center-supported micromechanical disk resonators. Sensor. Actuat. A **134**, 582–593 (2007)
7. Hao, Z.: Vibration displacement on substrate due to the time-harmonic stress sources from a micromechanical resonator. J. Sound Vib. **322**, 196–215 (2009)
8. Miller, G.: The field and radiation impedance of mechanical radiators on the free surface of a semi-infinite isotropic solid. Proc. R. Soc. Lond. V **223**, 521–541 (1954)
9. Xu, Y.: Numerical models and experimental investigation of energy loss mechanisms in soi-based tuning-fork gyroscopes. Sensor. Actuat. A **152**, 63–74 (2009)
10. Trusov, A.: Electrostatic regulation of quality factor in tuning fork MEMS. IEEE Sensors 2011 Conference. Limerick, Ireland, October 28–31 (2011)
11. Hao, Z.: Thermal-energy method for calculating thermoelastic damping in micromechanical resonators. J. Sound Vibrat. **322**, 870–882 (2009)
12. Hao, Z.: VHF single crystal silicon elliptic bulk-mode capacitive disk resonators part I: Design and modeling. J. Microelectromech. Syst. **13**, 1043–1053 (2004)
13. Pourkamali, Z.: VHF single crystal silicon elliptic bulk-mode capacitive disk resonators. Part II: Implementation and characterization. J. Microelectromech. Syst. **13**, 1054–1062 (2004)

14. Wang, J., Butler, J.E., Feygelson, T., Nguyen, C.T.-C.: 1.51-GHz nanocrystalline diamond micromechanical disk resonator with material-mismatched isolating support. Proc. MEMS. 641–644 (2004)

Anderson Localization

► Harnessing Disorder at the Nanoscale

Angiogenesis

Alessia Bottos[1], Elena Astanina[1], Luca Primo[2] and Federico Bussolino[1]
[1]Department of Oncological Sciences, University of Torino, Candiolo, Torino, Italy
[2]Department of Clinical and Biological Sciences, University of Torino, Candiolo, Torino, Italy

Definition

Originally, the term "angiogenesis" was strictly confined to describe the sprouting of blood vessels from preexisting capillaries or post-capillary venules. Presently, angiogenesis refers to the molecular and cellular mechanisms leading to the formation and remodeling of vascular tree occurring during embryo development and in adult life. However, this term often includes the concept of vasculogenesis. Vasculogenesis indicates the earlier events of the formation of the cardiovascular system and is characterized by the assembly of a primitive vascular plexus constituted by angioblasts, which are believed to arise from the hemangioblast, a common precursor for endothelial and hematopoietic lineages.

There are two other general concepts that strictly parallel angiogenesis. The first is arteriogenesis, which describes the growth of functional collateral arteries from preexisting arterio-arteriolar anastomoses in response to ischemic injuries. The second is lymphangiogenesis, which illustrates the formation of the lymphatic system. This forms a unidirectional network of blind-ended terminal lymphatic vessels that collect the protein-rich fluid that exudes from blood capillaries and then drain through a conduit system of collecting vessels, lymph nodes, lymphatic trunks and ducts into the venous circulation.

Introduction

The notion that blood circulation is indispensable for bringing nutrients to the organs and for organism growth has been known for millennia. However, the connected concept – that growth of blood vessels may promote disease – is more recent [1, 2].

The perception that tumor growth is linked to an increased vascularization took place in the mid-twentieth century. In 1971, Judah Folkman gave the first formal demonstration that tumors release soluble molecules able to recruit from surrounding tissue capillaries, which enable cancer to growth over few millions of cells, infiltrate the normal parenchyma, and colonize distant organs. These observations opened an extraordinary research area that is far to be completely exploited and is at the frontier of the molecular medicine. The greatest achievement of these studies has been made over the past years in targeting angiogenesis for human therapy. In particular, in the 2004 vascular-endothelial growth factor (VEGF)-A removal by the humanized monoclonal antibody bevacizumab and by the specific aptamer pegaptanib was, respectively, approved for the treatment of metastatic colorectal cancer and exudative retinopathy [3]. Currently more than 50 anti-angiogenic compounds are in clinical trials. However, clinical experience has shown that the antivascular effect of bevacizumab as well as of other anti-angiogenic compounds are often transient or even negligible [4, 5], indicating that new knowledge is required to better understand the complexity of vascular formation and to properly select responder patients.

Nevertheless the impact of angiogenesis in medicine overcomes oncology and is extended to other relevant pathologies including ischemic diseases, chronic inflammatory diseases,

Angiogenesis, Table 1 Human diseases characterized by abnormal vascularization

Organs	Diseases
Whole body	Cancer, autoimmune diseases, ischemic diseases, diabetes, infections
Vessels	Vascular malformations, cavernous hemangiomas, hemangiomas, atherosclerosis, hereditary hemorrhagic telangiectasia, Di George syndrome, glomerulonephritis, transplant arteriopathy
Adipose tissue	Obesity
Skin	Psoriasis, warts, blistering disease, allergic dermatitis, scar keloid, pyogenic granulomas
Eye	Persistent hyperplastic vitreous syndrome, Norrin disease, exudative retinopathy, retinopathy of prematurity
Lung	Primary pulmonary hypertension, asthma, nasal polyps
Intestines	Inflammatory bowel disease, ascites, periodontal diseases
Reproductive system	Endometriosis, ovarian cysts, ovarian hyperstimulation
Bone, joints	Arthritis, synovitis, osteomyelitis, osteophyte formation

metabolic disorders, obesity, some mendelian inherited diseases, reproductive disorders, and bone diseases (Table 1).

The Morpho-Functional Vascular Unit

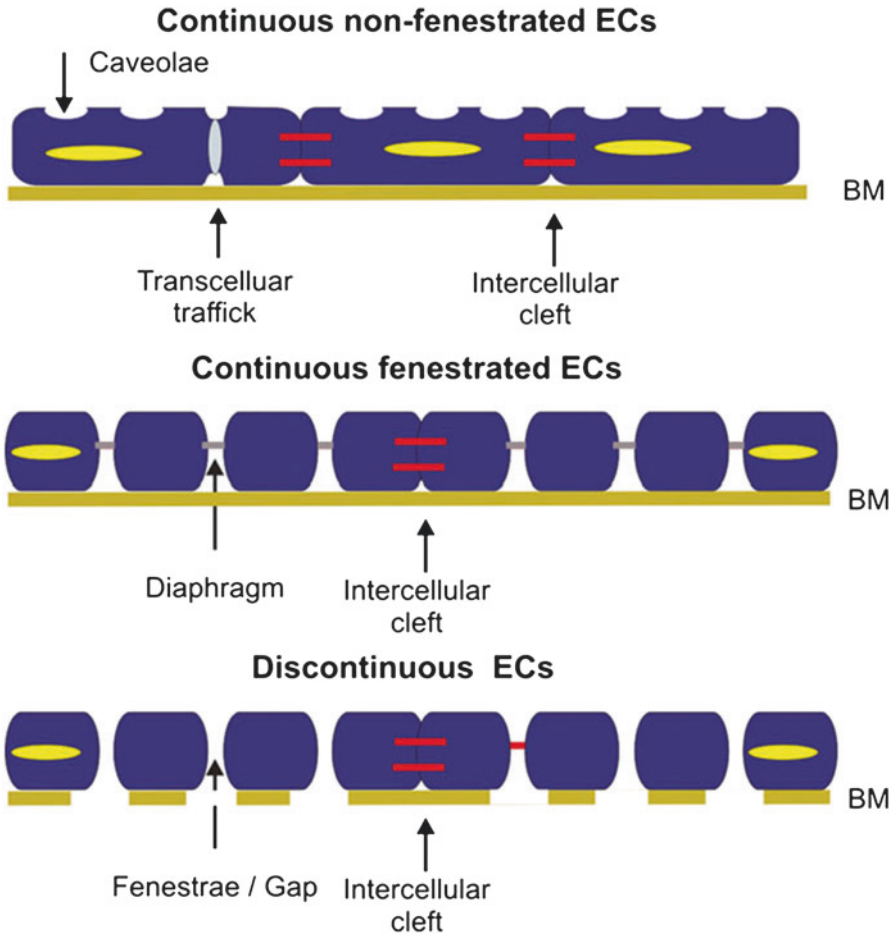
Arteries, veins, and capillaries share a common structure (tunica intima) characterized by endothelial cells (EC) that line vascular lumens and are surrounded by the basement membrane (BM) constituted by extracellular matrix proteins and complex sugars (i.e., heparan sulfates, proteoglycans). The external face of tunica intima is covered by vascular smooth muscle cells and pericytes (mural cells) that form a multilayer wall of different size (thinner in the veins and ticker in the arteries), named tunica media. In capillaries, pericytes do not form a continuous layer but they are scattered and embedded in the BM. Mural cells exert key roles in multiple microvascular processes, including: (1) endothelial cell proliferation and differentiation, (2) contractility and tone, (3) stabilization and permeability, and (4) morphogenesis and remodeling during disease onset [6, 7].

EC thickness varies from less than 0.1 μm in capillaries and veins to 1 μm in the aorta and in arteries they are aligned in the direction of blood flow and this remodeling occurs in response to hemodynamic shear stress. On the contrary in vasculature characterized by a slow blood stream, endothelial shape may be irregular, rounded, or elliptical. Endothelium may be continuous or discontinuous (Fig. 1). Continuous endothelium, in turn, is fenestrated or nonfenestrated. Nonfenestrated continuous endothelium is found in arteries, veins, and capillaries of the brain, skin, heart, and lung. Fenestrated continuous endothelium occurs in tissues that are characterized by increased filtration or increased transendothelial transport, including exocrine and endocrine glands, gastrointestinal mucosa, and kidney. Fenestrae are transcellular pores (∼70 nm in diameter) that develop through the whole thickness of the cell and have a thin 5–6 nm nonmembranous diaphragm across their aperture.

Discontinuous endothelium characterizes capillaries of sinusoidal vascular beds (i.e., liver and bone marrow). It is characterized by large fenestrations (100–200 nm in diameter) that lack a diaphragm and contain gaps (or large holes) within individual cells.

The regulation of these kinds of EC-EC interactions occurs at two types of intercellular junctions: the tight junctions (also named *zona occludens*), which are mainly located at the apical region of the intercellular cleft and the adherens junctions (*zona adherens*). Zonula occluden-1, occludin, claudin, and JAM are the major elements of tight junctions, while vascular-endothelial cadherin is a specific constituent of adherens junctions and its clustering at cell-cell contacts promotes the formation of multimolecular complexes that comprise signaling, regulatory, and scaffold proteins.

In physiologic and pathological conditions, the above-described structural features of endothelium contribute to maintain the fluid and the cellular homeostasis between blood and tissues. Actually, the vascular system needs to be



Angiogenesis, Fig. 1 Morphology of capillaries. Capillaries in different anatomical districts have different morphology. Continuous nonfenestrated EC is mainly found in skin, heart, muscle, brain, and lung. It regulates the passage of molecules through transendothelial channels or by caveolae-mediated transcytosis. Continuous fenestrated ECs characterize glomerulus, endocrine glands, and intestines and demonstrate greater permeability to water and small molecules but do not allow the passage of larger macromolecules by using diaphragms as molecular filters. Discontinuous endothelium is present in liver, bone marrow, and spleen and has fenestrae (without diaphragms), gaps, and partially structured BM. The intercellular cleft represents tight junctions or adherens junctions

sufficiently permeable to allow the ready exchange of small molecules (gases, nutrients, waste products) with the tissues. Plasma proteins also need to cross the normal vascular barrier, at least in small amounts. Albumin, for example, transports fatty acids and vitamins and immunoglobulin antibodies are required for host defense. The passage of small molecules across the endothelial layer is almost passive and has to take into account three parameters: hydraulic conductivity, reflection coefficient, and diffusion. Diffusion is the most important of these for the exchange of small molecules and is driven by the molecular concentration gradient across vascular endothelium (Fick equation). On the other hand, filtration is much more important than diffusion for the flux of large molecules such as plasma proteins and is determined by the Starling equation.

On structural point of view, endothelial permeability is mediated by the so-called transcellular and paracellular pathways – that is, solutes and cells can pass through (transcellular) or between

(paracellular) EC. While the paracellular route restricts the passage of solutes larger than 3 nm in radius, transcellular vesicle trafficking selectively transports macromolecules such as albumin across the endothelium. Transcellular passage (transcytosis) is an energy-dependent mechanism that requires either cell fenestration or a complex system of transport vesicles, which includes caveolae and organelles called vesiculo-vacuolar organelles (VVO). These organelles, or similar structures, can also fuse and appear as channels that traverse single cells and allow the passage of leukocytes and solutes through the endothelium. Caveolae are 70-nm membrane-bound, flask-shaped vesicles that usually open to the luminal or abluminal side of the EC membrane. They are enriched in cholesterol and some, but not all, caveolae possess a thin diaphragm similar to that found in fenestrae. The density of caveolae is far greater in capillary endothelium (up to 10,000 per cell) compared with other vessels and their number is highest in continuous nonfenestrated endothelium. A notable exception is the blood–brain barrier, where caveolae are rare, which reflects the protective role of this vascular bed for the brain.

VVOs are comprised of hundreds of cytoplasmic vesicles and vacuoles that together form an organelle that traverses venular endothelial cytoplasm from lumen to albumen. VVOs often extend to inter-endothelial cell interfaces and their individual vesicles commonly open to the inter-endothelial cell cleft. The vesicles and vacuoles comprising VVOs are linked to each other and to the luminal and abluminal plasma membranes by stomata that are normally closed by thin diaphragms that appear similar to those found in caveolae and fenestrae.

The paracellular pathway, by contrast, is mediated by the coordinated opening and closure tight junctions and adherens junctions, and is mainly operative during pathological conditions, in which this barrier is altered by the release of permeability factors (acute and chronic inflammatory injury) or by an abnormal and chaotic vascularization (i.e., tumor angiogenesis).

The Morpho-Functional Vascular Unit in Pathological Settings

The formation of vascular bed depends from well-regulated circuits that involve soluble molecules with inducing and inhibiting activity, mechanical forces, different cell types and features of the extracellular matrix. The final result of these dynamic interactions is the ordered distribution of capillaries (the intercapillary distance ranged from ∼100 to ∼150 μm) allowing the appropriate cell oxygenation on the basis of local pO_2 and O_2 coefficient diffusion [6–8].

In pathological settings (i.e., tumor, chronic inflamed tissues, wound healing, postischemic revascularization) this homeostatic balance is disturbed leading to the formation of vessels with functional and structural defects.

In pathological tissues and in particular in tumors, vessels are dilated, tortuous, and show a chaotic spatial distribution. Normal vascular tree is characterized by vessels with decreasing diameter and by dichotomous branching, but tumor vessels are unorganized with trifurcations and irregular diameters. Endothelial phenotype is also modified in tumors. Large discontinuities characterize inter-endothelial junctions, as well as the increased numbers of fenestrae and VVOs and alterations or lack of BM. Similarly, mural cells are poorly distributed and show defects in association with vessel wall. The molecular and cellular mechanisms causing these abnormal vascular architectures are largely unknown, but the imbalance of pro- and anti-angiogenic associated with mechanicals stresses is believed to be instrumental in achieving these features.

These structural alterations are accountable for some features of tumor microenvironment. In normal tissues permeation of hydrophilic solutes is progressively restricted with increasing molecular size, as by a sieve, with many openings 8 nm and a few 40–60 nm wide. In solid tumors vascular permeability is generally higher allowing (≥80–90 nm) or facilitating (≥10 nm) the extravasation of molecules with sizes that in general are retained. The blood flow is irregular because the flow resistance, which depends from vascular

geometry and blood viscosity, is increased. Focal interruption of vessel wall may also compromise the downstream blood flow. As a result, overall perfusion rates (blood flow rate per unit volume) in tumors are lower than in healthy organs. Blood flux is erratically distributed and fluctuates with time causing a jeopardized perfusion.

The increased permeability abolishes the differences between hydrostatic and colloid osmotic pressures that characterize normal tissues. As a result, tumors show interstitial hypertension and reduction of transmural pressure gradient and convection across tumor vessel walls. Consequently, the increased permeability connected with the interstitial hypertension reduces pressure difference between up and down stream of blood vessels and leads to blood flow stasis.

The final consequences of these functional abnormalities takes place at metabolic level leading to hypoxia and acidosis. Tumor vessels fail to adequately deliver oxygen and nutrient as well as to remove wastes. The imbalance of vascular network development and tumor cell proliferation results in the formation of hypovascular regions in tumors. Since tissue diffusion limit of oxygen is 80–200 μm, the regions far from blood vessels become hypoxic. Extracellular acidosis is a consequence of hypoxia and Warburg effect (cancer cells produce energy through glycolysis also in the presence of adequate amount of oxygen). When pO_2 is low, the metabolic process (glycolysis followed by oxidative phosphorylation) sustaining oxidative demolition of glucose and ATP production decreases its efficiency. Actually, in the absence of oxygen, the end product of glycolysis, pyruvic acid, does not proceed toward oxidative phosphorylation but it is reduced to lactic acid. This metabolite, together carbonic acid generated by growing cells, accounts for the reduction of extracellular pH in tumors.

The above-described metabolic and functional characteristics render the microenvironment hostile, leading to a genetic selection that favors the most aggressive and malignant cancer cells. Finally, the combination of acidosis (which may modify the polarity of hydrophilic compound) with the blood stasis and the increased interstitial pressure hinders the delivery and efficacy of therapeutic agents to tumors.

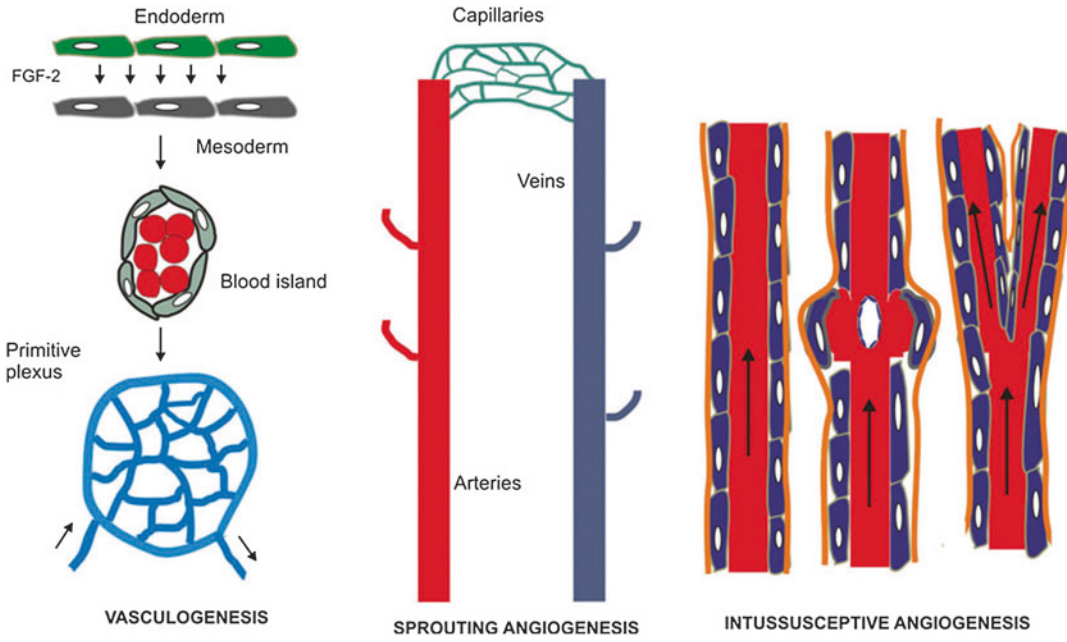
The Mechanisms of Angiogenesis

Mesenchymal cells differentiate in angioblasts, which undergo vasculogenesis where they aggregate and form a primitive vascular network firstly located into extraembryonic tissues and then invading the embryo. As vessels begin to be remodeled, they undergo localized proliferation and regression, as well as programmed branching and migration into different regions of the body. They need to be specified into different calibers and types of vessel, including division into arteries, veins, and lymphatics, with further subdivision into large vessels, venules, arterioles, and capillaries. In addition, they need to recruit supporting mural cells and create the optimal extracellular matrix to ensure the stability of the vessels formed. Pulsatile and shear stress forces mediated by heart beats are instrumental for capillary branching and pruning. The establishment of final vascular architecture is characterized by four different mechanisms, which are also present in adult life: (1) formation of primitive and immature plexus, (2) stabilization of primitive plexus, (3) remodeling toward the final architecture, and (4) vascular specialization. Molecules involved in these steps are summarized in Table 2 [9–12].

Formation of primitive and immature vascular plexus – During embryo development, the nascent capillary network is generated by combining vasculogenesis and angiogenesis. Vasculogenesis describes the *de novo* formation of vascular plexus. Mesenchymal cells are stimulated by soluble molecules (i.e., Indian hedgehog, basic fibroblast growth factor, bone morphogenic protein 4) released by endodermal cells to differentiate in vascular-endothelial growth receptor-2 (VEGFR-2)-positive cells. In extraembryonic yolk sac, these cells form hemangioblast, the common precursor of vascular and hematopoietic system. Primitive endothelial and hematopoietic cells coalesce to form blood islands. The outer cells of the blood islands are endothelial, whereas the inner cells give rise to hematopoietic progenitors.

Angiogenesis, Table 2 Most important molecules involved in angiogenesis

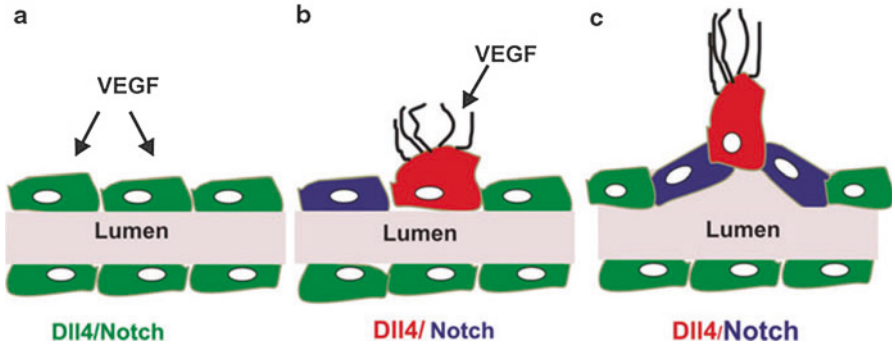
Angiogenic inducers (ligand/receptor)	
VEGF-A/VEGFR-1	Negative regulation of embryonic angiogenesis. Later, VEGFR-1 acts as decoy receptor. Recruitment of monocytes and macrophages
VEGF-A/VEGFR-2	Regulation of proliferation, permeability, migration, and survival of ECs. ECM remodeling. EC differentiation and specialization (i.e., fenestrae formation). The activity of VEGFR-2 is modulated by association with co-receptors (neuropilins, VE-cadherin, αvβ3 integrin). VEGF-A isoforms contribute to vascular patterning
Placental growth factor/VEGFR-1	Regulation of pathological angiogenesis (i.e., tumor, collateral formation, wound healing). Contribution to hematopoiesis. Recruitment of inflammatory cells
VEGF-B/VEGFR-1	Cardiac vascularization
VEGF-C or VEGF-D/VEGFR-2 and/or VEGFR-3	Control of lymphangiogenesis in physiologic and pathological settings. Modulators of vascular angiogenesis
Notch/Dll4 or Jagged	Control of EC tip formation in sprouting angiogenesis. Differentiation and recruitment of mural cells. Venous-arterial differentiation
Angiopoietins/Tie-2	Regulation of the relationships between mural cells and EC. Regulation of vessel stabilization. Accessory role in lymphangiogenesis. Activation of inflammatory cells
TGFβ/ALK or endoglin	Production of ECM and proteases. Venous-arterial differentiation. Mural cell differentiation. Control of EC proliferation and migration
PDGFs/PDGF receptors	Differentiation, survival, and recruitment of mural cells
Ephrin B2/EphB4	Venous-arterial differentiation. Vascular branching
Slits/Robo	Formation of primitive plexus. Vascular stabilization. Control of EC motility
Netrins/UNC5B	Negative control of sprouting angiogenesis. Regulation of vascular branching
Semaphorins/plexins-Neuropilins	Regulation of vascular morphogenesis. Inhibition of EC adhesion and migration. Anti-angiogenic effect in tumor angiogenesis
Hepatocyte growth factor/Met	Regulation of pathological angiogenesis. Collateral formation
Chemokines/Chemokine receptors	Members that contain the "ELR" motif are potent promoters of angiogenesis via binding and activating CXCR2 on ECs. In contrast, members, in general, those lack the ELR motif (ELR-) are potent inhibitors of angiogenesis, and bind to CXCR3 on ECs
Endogenous angiogenic inhibitors	
Tumstatin, arresten, and canstatin	They are cryptic angiostatic peptides released by collagen type IV during ECM proteolysis in sprouting angiogenesis
Endostatin	A cryptic angiostatic peptide released by collagen type XVIII during ECM proteolysis in sprouting angiogenesis
Angiostatin	A plasminogen fragment produced during sprouting angiogenesis. It Inhibits EC migration and proliferation and induces apoptosis
Vasostatin	Vasostatin is an N-terminal fragment of the calreticulin. It inhibits angiogenesis presumably by binding to laminin and blocking EC adhesion and proliferation
Thrombospondin	It inhibits EC proliferation and migration by interacting with CD36 molecule
2-methoxyestradiol	It inhibits angiogenesis by reducing EC proliferation and inducing EC apoptosis
Histidine-rich glycoprotein	It inhibits tumor angiogenesis by preventing release of angiogenic factors from the matrix and inhibiting growth factor-induced EC migration
Pigment epithelial-derived factor	It is released by retinal pigment epithelium and causes EC apoptosis
Maspin	It is a serpin inhibitor of serine-proteases with anti-angiogenic activity in tumor
PEX	It is a fragment of metalloproteinase-2 that prevents its collagenolytic activity during sprouting angiogenesis
Fibulin 5	It is a ECM protein that antagonizes VEGF activities and enhances that of thrombospondin
Interferons (α,β,γ)	They inhibit angiogenesis by multiple and direct effects on EC or by modulating innate immunity



Angiogenesis, Fig. 2 Vasculature assembly. *Vasculogenesis* starts when mesenchymal cells stimulated by soluble molecule released by endoderm differentiate into EC precursors (hemangioblasts and angioblasts) and form the blood islands (*left*). Fusion of blood islands leads to formation of primary capillary. *Sprouting angiogenesis*. Blood circulation is established and primary plexi are remodeled into a hierarchical network of arteries, capillaries, and veins. Sprouting angiogenesis starts when a gradient of VEGF triggers the detachment of EC-EC contacts leading EC to proliferate and migrate. The nascent capillaries undergo maturation characterized by reorganization of ECM and recruitment of mural cells. This step parallels the selection of the proper trajectory for the best blood supply. *Intussusception*. The division of vessels through the insertion of tissue pillars is an alternative mechanism that leads to the expansion of blood vessels in several organs. Little is known about the function or regulation of intussusceptive growth but the process seems to involve EC growth, motility, and remodeling of ECM

The endothelial-lined blood islands fuse to generate *de novo* a primary capillary plexus. This process is named vasculogenesis. In intra-embryonic tissues, VEGFR-2-positive mesenchymal cells differentiate into angioblasts which directly aggregate into the dorsal aorta or cardinal vein, without a plexus intermediate. These large vessels then undergo connection with the primitive vascular network.

Besides vasculogenesis, sprouting and intussusceptive angiogenesis take place (Fig. 2). Sprouting angiogenesis if facilitated by hypoxia, which upregulates the transcription of a large set of pro-angiogenic genes. EC loose the cell-cell contacts and vessels become leaky. The BM is digested by the release of specific proteases (metalloproteases) and a specific suppression of protease inhibitors allowing EC migration and proliferation. Intriguingly, BM cleavage can produce multiple anti-angiogenic molecules that counteract and balance the effect of angiogenic inducers thus allowing an accurate vascularization. The sprouts reorganize internally to form a vascular lumen and are finally connected to other capillary segments. VEGF signaling is critically important in this process, as it promotes endothelial cell proliferation and modulates migration. Angiopoietin-2, in the presence of VEGF, facilitates sprouting angiogenesis because it contributes in dismantling cell contacts. The pivotal effect of VEGF in sprouting angiogenesis is regulated by Notch receptors and their Delta-like-4 (DLL4) and Jagged-1 ligands to allow a specific focalization of capillary area from which a new capillary stems. The fine-tuning activity between VEGF and DLL4 allows the selection of the tip



Angiogenesis, Fig. 3 Tip cell selection in sprouting angiogenesis. (**a**) At the beginning of VEGF stimulation ECs most probably contain equal amount of Dll4 and its receptor Notch. (**b**) Tip selection occurs by increasing Dll4 expression in a specific EC (*red* cell) stimulated by VEGF. As consequence, Notch is upregulated in neighboring cells (*blue*), which inhibit the expression of VEGF receptors. (**c**) This modulation allows tip cells reacting strongest to VEGF signaling and acquiring sprouting phenotype whereas tip cell behavior is suppressed in Jag1-expressing (stalk) cells (*blue*), which form the base of the emerging sprout

EC, which is devoid of proliferative activity and drives the sprout by emitting filopodia. Initially, Dll4 and Notch signaling are thought to be balanced in EC until presumptive tip cells eventually increase Dll4 expression in response to VEGF signaling. Consequently, Notch is upregulated in neighboring cells, which inhibits the expression of VEGF receptors and the tip cell differentiation. Thus, Dll4-expressing tip cells react strongest to VEGF signaling and acquire a motile, invasive, and sprouting phenotype whereas tip cell behavior is suppressed in Jag1-expressing (stalk) cells, which form the base of the emerging sprout. Reduced levels of DLL4 or blocking of Notch signaling enhance the formation of tip cells, resulting in dramatically increased sprouting and chaotic angiogenesis. The principle in which a differentiating cell represses the differentiation of adjacent cells is called lateral inhibition (Fig. 3).

More recently ephrins and wnt pathway has been demonstrated to modulate tip cell dynamics. Although little is known about the lumenization of blood vessels, there is strong evidence for the involvement of pinocytosis and vacuole formation. High-resolution time-lapse imaging has established that the lumen in these ECs is formed by intracellular and, subsequently, intercellular fusion of large vacuoles.

Intussusception (growth within itself) is an alternative to the sprouting mode of angiogenesis. The protrusion of opposing microvascular walls into the capillary lumen creates a contact zone between ECs. The endothelial bilayer is perforated, intercellular contacts are reorganized, and a transluminal pillar with an interstitial core is formed, which is soon invaded by myofibroblasts and pericytes leading to its rapid enlargement by the deposition of collagen fibrils. When a primitive capillary plexus is generated by vasculogenesis or sprouting, intussusception is triggered and is responsible for rapid vascular growth and optimal remodeling.

Stabilization of primitive plexus – Blood perfusion of nascent capillaries requires their maturation. This step is mainly characterized by recruitment of mural cells and remodeling of extracellular matrix (ECM). Mural cells are recruited by molecules released by EC and in particular by platelet-derived growth factors (PDGF) and sphingosine-1 phosphate, which, respectively, bind the tyrosine kinase PDGF receptor and the seven spanning domains endothelial differentiation sphingolipid G-protein-coupled receptor-1. Compelling support for this hypothesis comes from mouse genetic models in which the modification of the expression of these molecules impair the correct relationships between mural cells and EC and alter production of ECM. Mural and stroma cells produce angiopoietin-1 (Ang-1), which binds the tyrosine

kinase receptor, Tie2 on endothelial cells, resulting in enhanced EC-mural cell interactions leading the vessels leak-resistant. A homolog, Ang-2, appears to antagonize Ang-1, relaxing these interactions and triggering BM degradation in preparation for angiogenic sprout. Transforming-growth factor β promotes the differentiation of mesenchymal cells into mural cells and contributes to ECM production. Mutations in endoglin and activin-like kinase, two receptors of transforming-growth factor β, cause hereditary hemorrhagic telangiectasia, a disease that is characterized by capillary dilatation, microaneurysms, and bleeding.

Interactions between ECs and their surrounding BM and ECM are critical for vascular maturation and occur through specialized adhesive receptors named integrins. The importance of these connections are underscored by the vascular defects seen in mice deficient in some types of integrins and various BM constituents such as type IV collagen, fibronectin, and laminins. Interestingly, the adhesion of EC to BM through integrins mediates anti-apoptotic signals that are instrumental to avoid vessel regression. Actually, anti-integrin molecules, including specific antibodies, have anti-angiogenic effects. Besides an adhesive role, integrins act as co-receptors of several tyrosine kinase receptors involved in angiogenesis modulating their activities. Following BM assembly, vessels are stabilized by cross-linking of ECM components and associated proteins by transglutaminase-2-mediated lysine bonds. The ECM cross-linking activity leads to inhibition of angiogenesis by increasing ECM stiffness and deposition.

Remodeling toward the final architecture – The transition from the plexus to the vascular tree is characterized by regression and pruning of some vascular segments, new branching, remodeling, and navigation to build the most appropriate network to distribute oxygen, nutrients, and wash catabolites off. This step is regulated by the combined activities of physical forces, axon guidance cues, and ECM features. Using in vivo molecular imaging associated with genetic and vascular grafting studies it has been established that physical cues drive guidance of lumenized vessel sprouts. ECs are subjected to tension created by cell-cell and cell-ECM interactions or by blood flow (shear stress) and pressure (circumferential wall stress). Interestingly, the absence of blood flow or sufficient tensions promotes vascular regression. Physical stimuli must be sensed by cells and transmitted through intracellular transduction pathways to the nucleus, resulting in gene expression. However, how cells sense biomechanical stimuli is still under debate. Several hypotheses experimentally supported have been proposed. They include the capacity of heterotypic (integrin-mediated cell-ECM) and homotypic (VE-cadherin mediated EC-EC interaction) adhesive receptors to link the internal cytoskeleton and provide mechanical coupling across the cell surface, including ion- channel opening [13].

As above mentioned, proteases released by angiogenic EC modify the structure of ECM during the formation of primitive plexus. Furthermore, this mechanism occurs also during the remodeling phase leading to pro- or anti-angiogenic activities. This stimulatory effect may be due to decreasing the density of extracellular matrix proteins, and/or by exposing cryptic binding sites within matrix molecules that promote migration. In contrast, the action of proteases can be anti-angiogenic, due to their ability to generate fragments of matrix molecules that have pro-apoptotic properties not possessed by the intact molecule. Therefore the characteristics of surrounding ECM may favor the progression or the regression of primitive capillary network. On the other hand ECM elements, and in particular fibronectin, collagens, laminins, and heparan sulfates, participate in vascular stabilization and patterning as demonstrated by specific vascular defects (increase in lumen size, edema, leakage, vessel rupture) in mouse genetic models.

ECM also affects the availability of angiogenic regulators. For instance, different isoforms of VEGF-A, characterized by increased affinity to ECM, form specific gradients required for vessel patterning. Similarly, fibroblast growth factor-2 may be sequestered by ECM and released to cell surface by proteolysis. Generally, ECM and proteolytic activities adapt the strength and the

duration of angiogenic signals and train ECs to initiate fine-tuned and specific signal genetic programs regulating vascular patterning.

During the remodeling step, vessels select specific trajectories leading to an optimal spatial distribution instrumental for the molecular exchanger in the tissues. This strategy is shared with the nervous system, in which axon guidance molecules are responsible for guiding the navigating axon toward its target area. Indeed ECs and neurons share some fundamental mechanisms during the formation of their networks. On an anatomical point of view, in several tissues (skin, muscles, intestine) nerves and larger blood vessel align into two parallel structures and genetic studies indicate that peripheral nerves in the dermis provide a spatial template for the growth and differentiation of arteries. During axon navigation, growth cones project numerous filopodia as shown in endothelial tip cells that actively extend and retract in response to extracellular cues. Four ligand-receptor pairs of guidance cues are known to play attractive of repulsive effects in nervous and vascular systems: ephrins/Eph, semaphorins/plexins-neuropilins, slit/Robo, and netrins/UNC-DCC. Many studies performed in genetic modified mice or zebrafish clearly indicate that these molecules mainly act in defining the final shape and architecture of vessels in body tissues.

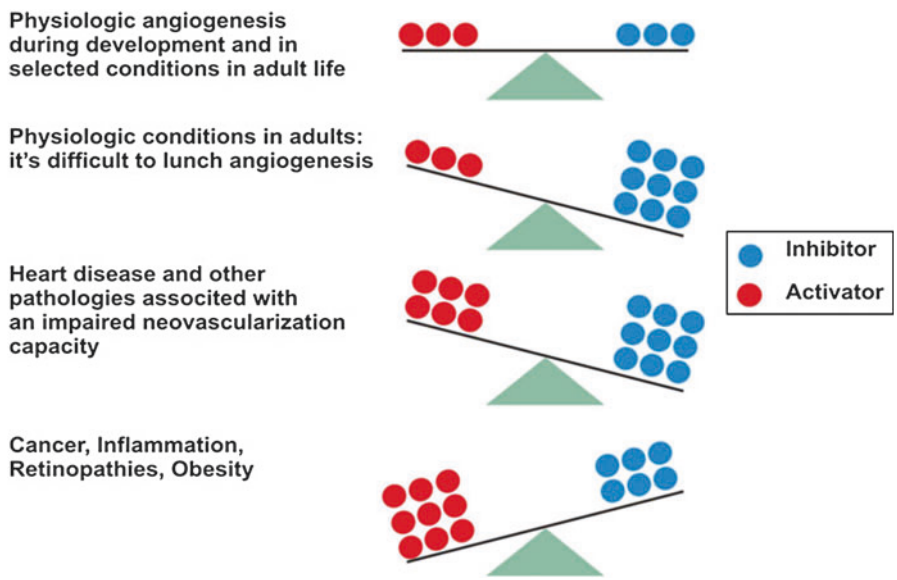
Vascular specialization – The paradigmatic example of this concept is the differentiation of arteries and veins. During vascular development cardiac work presses high-pressure blood flow through large diameter arterial vessels, which acquire an extensive supporting system of mural cells. Blood comes back to the heart with low pressure through veins, which develop *venous valves* to prevent blood from flowing back. Therefore hemodynamic forces have a crucial role in defining arterial and venous circulation. These mechanisms are strictly intermingled with early genetic determinant. The landmark of arterial–venous cell fate is the expression of ephrinB2 and EphB4, respectively, in plasma membrane of arterial and venous ECs. The interactions between these two markers are essential for proper vascular development. However, the expression of these molecules is differently controlled. VEGF upregulates the Notch pathway that through hesr1/hesr2 transcription factors leads to arterial specification via EphrinB2 activation. In veins, COUP-TFII is required for strong expression of EphB4 and venous differentiation. Therefore it seems that molecular differences between arteries and veins are predetermined but the final fate is contributed by epigenetic and physical cues indicating a prominent plasticity of the system.

The vascular architecture has definite features in different organs to respond to specific needs in terms of nutrient requirements and transport [14]. For instance, the filtration barrier of glomerular endothelial cells in kidney results from the presence of continuous fenestrations allowing bulk fluid and anionically charged molecules to be filtered and facilitates movement of cationically charged molecules. The sinusoidal hepatic system is characterized by discontinuous endothelium with sieve-like pores (∼100 nm diameter) and nondiaphragmatic fenestrae, which regulate the permeability and the blood flow in hepatic sinusoids. Endocrine organs have fenestrated capillaries that contain pores sometimes lined with thin diaphragms. These structures are thought to allow increased transport of long-range hormones between extravascular cells and blood. In contrast to the fenestrated vascular beds present in high-transport organs, brain capillaries are linked by extensive tight junctions that restrict transendothelial movement.

These differences are furtherer envisaged by specific patterns of gene expression orchestrated by mechanisms largely unexplored. It is emerging the existence of angiogenic inducers specific for each organ. They include members of Wnt family and endocrine gland-VEGF.

The Mechanisms of Pathological Angiogenesis

In adult life, physiologic angiogenesis employs the same mechanisms illustrated above and occurs in endometrium during female cycle, in *corpus luteum* during reproductive cycle, in hypertrophic glands (i.e., breast during lactation), and in



Angiogenesis, Fig. 4 Homeostatic hypothesis in pathological angiogenesis. During embryonic development and in some process in adult life (i.e., female cycle, reproduction, lactation, muscle exercise), tissues express balanced amounts of activators and inhibitors of angiogenesis, which allow the formation of regular vessels. In other adult tissues, angiogenesis is repressed by an increased amount of inhibitors. Vascular regeneration may occur in ischemic tissues but usually it is insufficient to recover tissue oxygenation. On the other hand, tumors, obesity, chronic inflammatory diseases, and retinopathies are characterized by a prominent but chaotic angiogenesis. These different scenarios are associated with unbalanced ratios between activators and inhibitors

exercise-trained skeletal muscles. Several diseases are connected with and sustained by an abnormal angiogenesis. In some pathological settings, including tumors and chronic inflammatory diseases, abnormal angiogenesis results from an altered homeostasis between angiogenic inducers and inhibitors driven and sustained by specific pathogenic events. On the contrary, ischemic diseases are characterized by attempts to revascularize the necrotic area through angiogenesis and arteriogenesis (Fig. 4) [2, 15, 16].

Tumor angiogenesis – Oncogene-mediated cell transformation is necessary but not sufficient to promote metastatic and aggressive cancer. Tumor cell proliferation alone, in the absence of angiogenesis, can give rise to dormant, microscopic tumors of $\sim$1 mm^3 or less, but these in situ cancers are harmless to the host. Oncogene somatic mutations deeply modify the transcription profile of transformed cells, including the upregulation of VEGF and other growth factors often associated with the down-modulation of inhibitors of vessel formation (i.e., thrombospondin). Therefore the accumulation of genetic mutation along the tumor span-life progressively modifies the angiogenic homeostasis leading to the formation of a chaotic vasculature (see above). However the process of tumor angiogenesis is not limited to the altered production of factors in tumor cells and other mechanisms may occur. The angiogenic process may be partially sustained by EC precursors mobilized by bone marrow and migrated to the angiogenic site. Bone marrow participates to the angiogenic process by producing myeloid cells that differently modulate angiogenesis. For example, macrophages have long been characterized as a highly plastic cell type capable of tumor-suppressive or tumor-promoting effects, depending on the polarization state (M1 vs. M2). Tie2-expressing monocytes have also been reported to promote tumor growth through secretion of angiogenic factors, as also demonstrated for neutrophils and platelets. Interestingly the accumulation of $CD11b^{+}Gr1^{+}$ cells in mice has been associated with the refractoriness to anti-VEGF therapies.

At least in their early phase, some tumors do not mediate sprouting angiogenesis but they utilize preexisting vessels. For instance, brain tumors are highly vascularized and vessels have a similar phenotype than normal brain. Vessels are surrounded, co-opted by tumor cells, and no sprouts are observed (co-option mechanism). Later, when the tumor mass collapses normal capillaries, hypoxia takes places and activates a VEGF-mediated sprouting angiogenesis.

In some tumors and in particular in colon-rectal cancers, the capillary wall is often constituted by EC and tumor cells (mosaic vessel). This scenario may be induced by the endothelial migration during rapid vessel growth without sufficient proliferation to complete the endothelial lining, leaving cancer cells exposed to the lumen. Second, ECs could be shed from the vessel lining, leaving exposed tumor cells. Third, migrating tumor cells could invade the vessel wall, displacing endothelial cells from the lining.

Finally vascular mimicry describes the potential of tumor cells (i.e., in melanomas) or cancer stem cells (at least in glioblastoma) to differentiate in EC with a mixed phenotype and be able to form functional vessels. Despite a significant body of literature, vascular mimicry is not yet a universally accepted mechanism, as the concept has been challenged by several investigators.

Arteriogenesis – The development of an arterial circulation circumventing the occlusion of a large artery from preexistent small arterioles is named arteriogenesis and occurs exclusively in adult injured tissues. It shares some features with angiogenesis, but the pathways leading to it are different. Furthermore arteriogenesis is potentially able to fully replace an occluded artery whereas angiogenesis cannot. As demonstrated in embryonic arterial differentiation, a pivotal role in arteriogenesis is played by physical forces and in particular by shear stress. It is proportional to blood flow velocity and inversely related to the cube of the collateral vessel radius Therefore, since growth increases the collateral vessel radius, shear stress falls quickly. This may be the reason why arteriogenesis arrests without a completed vascularization of ischemic tissue. Collateral vessels of an occluded artery show activated ECs, able to release molecules involved in the recruitment of circulating monocytes, in vascular proliferation and remodeling (proteases, Notch ligands, integrins, VEGF, angiopoietins, fibroblast growth factors) and in control of vascular tone (nitric oxide). Furthermore these collaterals are leaky. A similar activation is observed in smooth muscle cells, which undergo a massive proliferation. Therefore it seems to be conceivable to hypothesize that high fluid shear stress starts at least two signaling mechanisms: one that attracts bone marrow-derived cells for remodeling and another that causes endothelial and smooth muscle cells to enter the cell cycle, leading to proliferation.

Acknowledgments Some of the concepts here reported originate from experimental papers supported by "Associazione Italiana per la Ricerca sul Cancro" (AIRC).

Cross-References

- ▶ Biosensors
- ▶ Cell Adhesion
- ▶ Intravital Microscopy Analysis
- ▶ Nanomedicine
- ▶ Nanotechnology in Cardiovascular Diseases
- ▶ Synthetic Biology

References

1. Carmeliet, P.: Angiogenesis in health and disease. Nat. Med. **9**, 653–660 (2003)
2. Folkman, J.: Angiogenesis: an organizing principle for drug discovery? Nat. Rev. Drug Discov. **6**, 273–286 (2007)
3. Ferrara, N., Kerbel, R.S.: Angiogenesis as a therapeutic target. Nature **438**, 967–974 (2005)
4. Bergers, G., Hanahan, D.: Modes of resistance to anti-angiogenic therapy. Nat. Rev. Cancer **8**, 592–603 (2008)
5. Jain, R.K.: Lessons from multidisciplinary translational trials on anti-angiogenic therapy of cancer. Nat. Rev. Cancer **8**, 309–316 (2008)
6. Aird, W.C.: Phenotypic heterogeneity of the endothelium: II. Representative vascular beds. Circ. Res. **100**, 174–190 (2007)
7. Aird, W.C.: Phenotypic heterogeneity of the endothelium: I Structure, function, and mechanisms. Circ. Res. **100**, 158–173 (2007)
8. Fukumura, D., Jain, R.K.: Tumor microvasculature and microenvironment: targets for anti-angiogenesis and normalization. Microvasc. Res. **74**, 72–84 (2007)

9. Jain, R.K.: Molecular regulation of vessel maturation. Nat. Med. **9**, 685–693 (2003)
10. Serini, G., Napione, L., Arese, M., Bussolino, F.: Besides adhesion: new perspectives of integrin functions in angiogenesis. Cardiovasc. Res. **78**, 213–222 (2008)
11. Roca, C., Adams, R.H.: Regulation of vascular morphogenesis by Notch signaling. Genes Dev. **21**, 2511–2524 (2007)
12. Coultas, L., Chawengsaksophak, K., Rossant, J.: Endothelial cells and VEGF in vascular development. Nature **438**, 937–945 (2005)
13. Hoffman, B.D., Grashoff, C., Schwartz, M.A.: Dynamic molecular processes mediate cellular mechanotransduction. Nature **475**(7356), 316–23 (2011)
14. Aitsebaomo, J., Portbury, A.L., Schisler, J.C., Patterson, C.: Brothers and sisters: molecular insights into arterial-venous heterogeneity. Circ. Res. **103**, 929–939 (2008)
15. Schaper, W.: Collateral circulation: past and present. Basic Res. Cardiol. **104**, 5–21 (2009)
16. Coffelt, S.B., Lewis, C.E., Naldini, L., Brown, J.M., Ferrara, N., De Palma, M.: Elusive identities and overlapping phenotypes of proangiogenic myeloid cells in tumors. Am. J. Pathol. **176**, 1564–1576 (2010)

Animal Coloration

▶ Structural Color in Animals

Anodic Arc Deposition

▶ Physical Vapor Deposition

Antifogging Properties in Mosquito Eyes

Xuefeng Gao
Suzhou Institute of Nano-Tech and Nano-Bionics, Chinese Academy of Sciences, Suzhou, People's Republic of China

Definition

Fogging occurs when vapor condenses into droplets on condensation nuclei (e.g., dust) or exposed objects (e.g., glass and plastics) with diameters larger than 190 nm, namely, half of the shortest wavelength of visible light. These droplets can firmly stick to the surface of optical transparent materials so as to reduce the visibility by light scattering and reflection, which are undesired. Antifogging is to avoid condensed water droplets on a surface. Antifogging properties in mosquito eyes present a novel surface nonstick superhydrophobicity for micrometer-size water droplets [1].

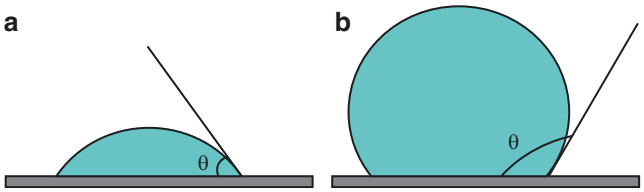
Overview

Surface wettability is a very common but important surface property closely related to the chemical composition and microstructures of material surfaces. According to the affinity of water to material chemicals, the wettability may be simply divided into the hydrophobicity and hydrophilicity. The intrinsic contact angle θ of a water droplet on a smooth solid surface is given by the classical Young's equation:

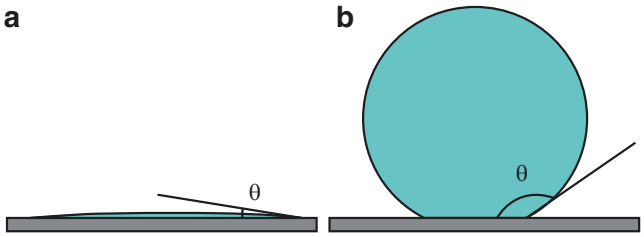
$$\cos\theta = \frac{\gamma_{SV} - \gamma_{SL}}{\gamma_{LV}} \quad (1)$$

where γ_{SL}, γ_{SV}, γ_{LV} are the interfacial free energies per unit area of the solid–liquid, solid–gas, and liquid–gas interfaces, respectively. The higher surface energy of solid materials means a lower contact angle with water, and vice versa. Generally, one defines that a surface with $\theta < 90^\circ$ is called a hydrophilic surface; conversely, a surface with $\theta > 90^\circ$ is called the hydrophobic (Fig. 1). Note that glass and PMMA are two kinds of most widely used optical transparent materials, both of which are hydrophilic with water contact angles of about 30° and 68°, respectively.

Water vapor easily condenses in the form of tiny droplets on any flat surface made of hydrophobic and hydrophilic materials once its temperature is lower than the dew point. Strictly speaking, the water condensed on exposed objects such as plastic and glass should be called dew although the formation of fog and dew virtually follows the same thermodynamic principles. However, for material sciences and engineering,



Antifogging Properties in Mosquito Eyes, Fig. 1 Contact angle and wetting states. (**a**) A hydrophilic surface with a water contact angle less than 90°. (**b**) A hydrophobic surface with a water contact angle higher than 90°



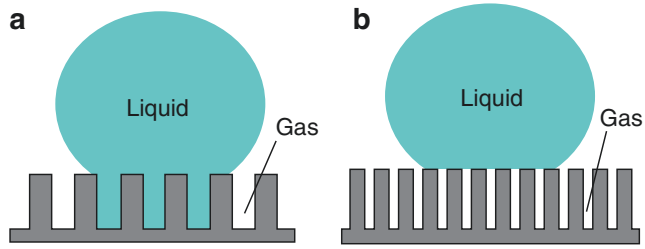
Antifogging Properties in Mosquito Eyes, Fig. 2 (**a**) A superhydrophilic surface with a water contact angle close to zero. (**b**) A superhydrophobic surface with a water contact angle larger than 150°

the measures of preventing the moisture condensation on optical transparent materials are customarily called antifogging while those aiming at other opaque materials such as metals and paints are called antidewing. It is known that the ordinary wettability makes the condensed water droplets firmly stick to the surfaces once contacting, which generates light scattering and reflection so as to greatly blur the visual field. It often occurs that the eyeglasses fog up when one comes inside on a cold day. This is because the eyeglasses have been cooled outside down to a temperature that is below the dew point and the moisture in the inside air will come in contact with the cooled eyeglasses and condense into tiny droplets. Similar phenomena can frequently occur on the aluminum heat exchange fins in air conditioners, resulting in higher energy consumption and mildew wind. Besides, other side effects caused by moisture condensation are often neglected, for example, the corrosion of metal, the swelling and peeling of paints, and the attachment of dust particles.

Up to now, most antifogging coatings reported are highly hydrophilic, aiming at protecting the optical transparency. A known example is the use of photocatalytic TiO_2 nanoparticle coatings that become superhydrophilic under UV irradiation [2]. Besides, a three-dimensional capillary effect was adopted to achieve superhydrophilic properties by constructing porous nanostructures from layer-by-layer assembled nanoparticles [3]. The key to these two wet-style antifogging strategies is to make these condensed droplets rapidly spread into a uniform thin film, which can avoid the light scattering and reflection. However, many other problems caused by adsorption of moisture are hard to solve by the superhydrophilic approach due to its inherently wet nature. Thus, a dry-style antifogging technique that can prevent the condensed drops from sticking to the material surface is highly desired, which may be developed into multifunctional coatings such as antifogging, antirust, and antifouling.

Basic Principles

Fogging of material surfaces generally forms via the impact of fog drops suspending in air or the direct condensation of moisture. The antifogging property can be achieved by changing the water affinity of the material surfaces toward the two extremities, as shown in Fig. 2. One is the superhydrophilic approach, with a water contact



Antifogging Properties in Mosquito Eyes, Fig. 3 Two types of superhydrophobic states. (**a**) Wenzel's model with water trapping into the intervals of structures on surface. (**b**) Cassie's model without any water trapping

angle close to 0° [2, 3]. The other is the nonstick superhydrophobic approach, with a contact angle close to 180° and extremely low adhesion for micrometer-size water droplets [1]. The first strategy is now well understood. Thus, we mainly intend to introduce the possibility of the superhydrophobic strategy in this text.

With the advance of nanofabrication techniques, one can easily construct micro- and nanostructures on the surfaces of different materials, which provide the possibility to study and modulate surface wettability toward the nonstick superhydrophobicity. The real material surfaces are actually not ideally smooth at atomic and molecular scale, which means that the classic Young's equation is not suitable in analyzing their water contact angles. Thus, Wenzel [4] proposed a theoretical model describing the apparent contact angle θ_W on a rough surface and modified Young's equation as follows:

$$\cos\theta_W = \frac{r(\gamma_{SV} - \gamma_{SL})}{\gamma_{LV}} = r\cos\theta \qquad (2)$$

where r is a roughness factor, defined as the ratio of the actual area of a rough surface to the geometric projected area. Apparently, this factor is always larger than unity and the hydrophobicity of a rough surface can be augmented by the increase of the solid–liquid contact area. However, such mechanism brings about very strong water adhesion, as shown in Fig. 3a. Thus, Wenzel model may be effective in the case of superhydrophobic surfaces with microscopic large interspaces or hollows enabling the water penetration but inapplicable to the development of nonstick and antifogging materials.

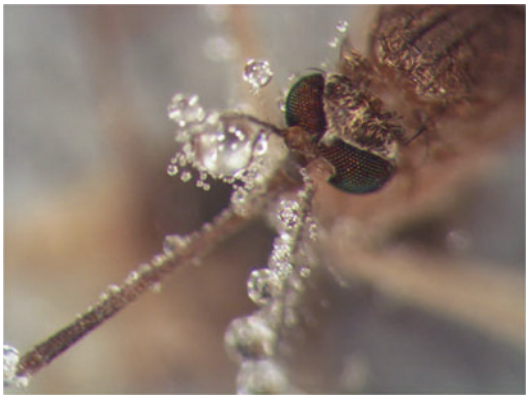
Air is an effective hydrophobic medium with a water contact angle ~180° [5]. As the air phase can be stably trapped into the microscopic hollows, a rough hydrophobic surface may be considered a composite surface composed of air and solid. Accordingly, the contact angle θ_C of a water droplet on the composite surface can be given by Cassie and Baxter's equation [5] as follows:

$$\begin{aligned}\cos\theta_C &= f\cos\theta + (1-f)\cos 180^\circ \\ &= F(\cos\theta + 1) - 1\end{aligned} \qquad (3)$$

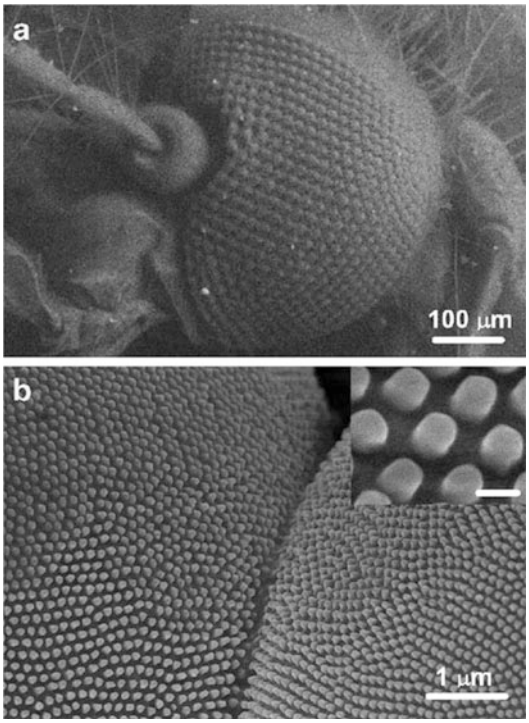
where f and $(1 - f)$ represent the area fraction of solid and trapped air, respectively; θ is the intrinsic contact angle of solid materials. Clearly, the hydrophobicity of a rough surface can be amplified by the decrease of the solid–liquid contact area (Fig. 3b). It has been reported that the surface adhesion of a superhydrophobic surface is governed by van der Waals attraction at the solid–liquid interfaces and the capillary force caused by the curving of the liquid–air interfaces [6]. Thus, it is theoretically possible that the nonstick superhydrophobic property can be obtained by fine surface nanoengineering for the antifogging purpose.

Key Research Findings

Lotus leaves are a sort of classic examples of natural nonstick superhydrophobic with a water contact angle >160° and a sliding angle <5°, exhibiting remarkable raindrop self-cleaning ability. Recent bionic research has clearly revealed that such property is attributed to the cooperative



Antifogging Properties in Mosquito Eyes, Fig. 4 A photograph of antifogging mosquito eyes (*C. pipiens*). Even though they are exposed to moisture, the surface of the eyes remains dry and clear while the surrounding hairs nucleate many drops



Antifogging Properties in Mosquito Eyes, Fig. 5 SEM images of (**a**) a single mosquito eye (*C. pipiens*) and (**b**) two neighboring ommatidia covered by hexagonally non-close-packed nanonipples. The inset is a high-resolution image. The scale bar: 100 nm (Adapted from Ref. [1]. Copyright © 2007 Wiley-VCH)

effects of the hierarchical micropapillae and branch-like nanostructures and the low-surface-energy wax [7, 8]. Up to now, much efforts have been made toward the bio-mimetic fabrication of the lotus-like nonstick superhydrophobic materials, which was designed for various applications such as self-cleaning and drag-reduction [9]. The emergence of such nonstick superhydrophobic technique was expected to be able to effectively solve the fogging problem. However, there are no reports on man-made dry-style antifogging coatings. Recently, researchers from General Motors have reported that the surface of lotus leaves becomes wet with moisture because the micrometer-size fog drops can easily trap in the interspaces among micropapillae [10]. Thus, the lotus-like microstructures are unsuitable for the creation of superhydrophobic antifogging coatings, and a new inspiration from nature is desired for solving this problem.

Recently, Gao et al. reported a novel dry-style antifogging strategy, inspired by the eyes of mosquitoes, *Culex pipiens*, which have ideal superhydrophobic properties (Fig. 4), providing an effective protective mechanism for maintaining clear vision in a humid habitat. As shown in Fig. 5, scanning electronic microscopic (SEM) observation reveals that the mosquito eye is a compound structure composed of hundreds of micro-hemispheres, which act as individual sensory units called ommatidia. These ommatidia are uniform with a diameter of ca. 26 μm and organize in a hexagonal close-packing (hcp) arrangement. The surface of each ommatidium was covered with numerous nanoscale nipples. These uniform nanonipples, with diameters of (101.1 ± 7.6) nm and interparticle spacings of (47.6 ± 8.5) nm, are organized in an approximately hexagonal non-close-packed (ncp) array.

According to Cassie's principle for surface wettability, such hierarchical micro- and nanostructures may be considered as heterogeneous curved surfaces composed of air and solids. Similar to the surfaces of lotus leaves [7, 8], air may be trapped in the spaces between the nanonipples and micro-hemispheres to form a stable air cushion that acts as an effective water barrier, greatly reducing the contact of tiny fog

drops with the eye surface. Chandler and co-workers have reported that it is theoretically difficult for a hydrogen-bonded network of water molecules to invade the voids in nanomaterials such as parallel hydrophobic plates with a critical separation of ca. 100 nm at room temperature and atmospheric pressure [11]. As a result, air is firmly trapped in the small voids between these neighboring nanonipples, which possess an average interparticle spacing of 47.6 nm. In this case, water surface tension is strong enough to cause the micrometer-sized fog drops to contract, assuming a perfect spherical shape due to their negligible weight. Moreover, the loose arrangement of nanonipples makes an extremely discrete and nonplanar triple-phase (liquid–air–solid) contact line, which is energetically favorable for driving the spherical fog drops effortlessly from the surface [12–15]. Thus, the combination of the ncp nipples at the nanoscale and hcp ommatidia at the microscale on the surface of mosquito eyes may induce ideal antifogging properties via a superhydrophobic approach.

Although mosquito eyes and lotus leaves are both superhydrophobic, the difference in their surface microstructures gives rise to different biological behavior: the former is antifogging for microscale fog drops while the latter is self-cleaning for millimeter-scale raindrops. It has been reported that the arrangement of lotus papillae at the microscale is random and diffuse with a spacing larger than fog drop diameters. Accordingly, the tiny fog drops are easily trapped in the spaces between the papillae so that the lotus leaf surface may become wet during a long exposure to moisture. In comparison, the mosquito ommatidia at the microscale assume a compact hcp arrangement with triangular voids less than 3 μm, which effectively prevent the plunge of fog drops, while the hexagonal ncp nanonipples trap a stable air cushion, thus preventing microscale fog drops from condensing on the corneal surface.

It is noteworthy that similar nanonipple array structures have also been found on other insects, such as moth eyes and cicada wings [16–19]. Detailed microscopic investigations of these nanonipples indicate that they have in common a characteristic tapered configuration, with periods in the range of 180–250 nm [16–18]. Interestingly, such nanostructures can have other biological functions such as anti-adhesive and broadband antireflective properties [16–19]. Inspired by these, one can develop advanced (multi-) functional interfacial materials. For example, the bio-inspired self-cleaning and zero-reflective coatings can greatly enhance the photoelectric conversion efficiency of solar cells. Many efforts have been made in fabricating nanotaper (e.g., nanonipple and nanocone) arrays by self-masked and masked reactive ion etching techniques [20]. However, the dry-style antifogging properties of these nanostructures have not been reported so far.

Future Directions for Research

The superhydrophobic antifogging technique can keep material surfaces dry, and thus is very attractive as comparing with other antifogging techniques. Nowadays, how to quantitatively measure the interaction of nanostructure surfaces with moisture and condensed drops remains a great challenge. In the near future, one should develop a widely accepted method. There are no successful reports on man-made dry-style antifogging coatings based on the current nonstick superhydrophobic techniques. Accordingly, continuous efforts should be made in designing and fabricating the nonstick superhydrophobic surfaces for microscale water drops. At the same time, their potential physical mechanism is under scrutiny, which is helpful to further guide the experimental research. Finally, when the dry-style antifogging technique is developed for practical applications, it becomes increasingly important to build up low-cost and large-scale nanomanufacture techniques and endow the robust nanofilms with excellent anti-wear, anti-aging properties as well as durability, in order to ensure functioning over longer operation periods.

Cross-References

- ▶ Lotus Effect
- ▶ Moth-eye Antireflective Structures
- ▶ Nanoscale Properties of Solid–Liquid Interfaces
- ▶ Nanostructured Functionalized Surfaces
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Nanotechnology

References

1. Gao, X., Yan, X., Yao, X., Xu, L., Zhang, K., Zhang, J., Yang, B., Jiang, L.: The dry-style antifogging properties of mosquito compound eyes and artificial analogues prepared by soft lithography. Adv. Mater. **19**, 2213–2217 (2007)
2. Wang, R., Hashimoto, K., Fujishima, A., Chikuni, M., Kojima, E., Kitamura, A., Shimohigoshi, M., Watanabe, T.: Light-induced amphiphilic surfaces. Nature **388**, 431–432 (1997)
3. Lee, D., Rubner, M.F., Cohen, R.E.: All-nanoparticle thin-film coatings. Nano Lett. **6**, 2305–2312 (2006)
4. Wenzel, R.N.: Resistance of solid surfaces to wetting by water. Ind. Eng. Chem. **28**, 988–994 (1936)
5. Cassie, A.B..D., Baxter, S.: Wettability of porous surfaces. Trans. Faraday Soc. **40**, 546–551 (1944)
6. Lai, Y., Gao, X., Zhuang, H., Huang, J., Lin, C., Jiang, L.: Designing superhydrophobic porous nanostructures with tunable water adhesion. Adv. Mater. **21**, 3799–3803 (2009)
7. Barthlott, W., Neinhuis, C.: Purity of the sacred lotus, or escape from contamination in biological surfaces. Planta **202**, 1–8 (1997)
8. Feng, L., Li, S., Li, Y., Li, H., Zhang, L., Zhai, J., Song, Y., Liu, B., Jiang, L., Zhu, D.: Superhydrophobic surfaces: from natural to artificial. Adv. Mater. **14**, 1857–1860 (2002)
9. Bhushan, B., Jung, Y.C.: Natural and biomimetic artificial surfaces for superhydrophobicity, self-cleaning, low adhesion and drag reduction. Prog Mater. Sci. **56**, 1–108 (2011)
10. Cheng, Y.T., Rodak, D.E., Angelopoulos, A., Gacek, T.: Microscopic observations of condensation of water on lotus leaves. Appl. Phys. Lett. **87**, 194112 (2005)
11. Lum, K., Chandler, D., Weeks, J.D.: Hydrophobicity at small and large length scales. J. Phys. Chem. B **103**, 4570–4577 (1999)
12. Chen, W., Fadeev, A.Y., Hsieh, M.C., Öner, D., Youngblood, J., McCarthy, T.J.: Ultrahydrophobic and ultralyophobic surfaces: some comments and examples. Langmuir **15**, 3395–3399 (1999)
13. Yoshimitsu, Z., Nakajima, A., Watanabe, T., Hashimoto, K.: Effects of surface structure on the hydrophobicity and sliding behavior of water droplets. Langmuir **18**, 5818–5822 (2002)
14. Gao, X., Yao, X., Jiang, L.: Effects of rugged nanoprotrusions on the surface hydrophobicity and water adhesion of anisotropic micropatterns. Langmuir **23**, 4886–4891 (2007)
15. Zheng, Y., Gao, X., Jiang, L.: Directional adhesion of superhydrophobic butterfly wings. Soft Matter **3**, 178–182 (2007)
16. Bernhard, C.G., Gemne, G., Sällström, J.: Comparative ultrastructure of corneal surface topography in insects with aspects on phylogenesis and function. Z Versgl Physiol. **67**, 1–25 (1970)
17. Parker, A.R., Hegedus, Z., Watts, R.A.: Solar-absorber antireflector on the eye of an eocene fly (45 Ma). Proc. R. Soc. Lond. B **265**, 811–815 (1998)
18. Sun, T.L., Feng, L., Gao, X.F., Jiang, L.: Bioinspired surface with special wettability. Acc. Chem. Res. **38**, 644–652 (2005)
19. Peisker, H., Gorb, S.N.: Always on the bright side of life: anti-adhesive properties of insect ommatidia grating. J. Exp. Biol. **213**, 3457–3462 (2010)
20. Min, W., Jiang, B., Jiang, P.: Bioinspired self-cleaning antireflection coatings. Adv. Mater. **20**, 3914–3918 (2008)

Anti-icing

▶ Micro-/Nanostructured Icephobic Materials

Anti-Stokes Luminescence

▶ Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics

Aperture Trap

▶ Laser Tweezers Using Nanoapertures in Metal Films

Applications of Nanofluidics

Vinh-Nguyen Phan[1], Nam-Trung Nguyen[2], Chun Yang[3] and Patrick Abgrall[4]
[1]School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore, Singapore
[2]Queensland Micro and Nanotechnology Centre, Griffith University, Brisbane, QLD, Australia
[3]Division of Thermal Fluids Engineering, Nanyang Technological University, School of Mechanical and Aerospace Engineering, and School of Chemical and Biomedical Engineering, Singapore, Singapore
[4]Formulaction, L'Union, France

Synonyms

nanofluidics; nanoscale fluid mechanics; nanoscale fluid dynamics; nanoscale mass transport

Definition

Nanofluidics studies the properties and transport phenomena of fluid in nanoscale, which typically ranges from 1 to 100 nm. The applications of nanofluidics employ those properties and transport phenomena either for real-life needs or to support other fields of research.

Overview

The transport of fluid confined in nanoscale structures (1–100 nm) possesses many interesting unique characteristics so that it promises various applications in science and technology. In history, people have applied the principle of transport in nanoscale to perform chromatography with porous materials. In chromatography, a mobile phase, usually liquid or gas, is driven through a static phase, or matrix, usually in solid. Due to difference in affinity of the particles in the mobile phase to the surface of the matrix, different kinds of particles will move to a different distance. By observing the pattern that the particles form, one can point out the composition of the particles in the mobile phase. Because the surface affinity plays a key role in chromatography, it is preferable for the matrix to have large area-to-volume ratio. Natural nanoporous materials satisfy well this criterion. The driving force in chromatography may be of various kinds, such as pressure driven, capillary, and electrophoresis. In techniques such as paper chromatography, the capillary is preferred because of its simplicity. However, in nanoporous material, where viscous force becomes more important, the capillary and pressure-driven chromatography turn out to be less effective. In such case, the electrophoresis is more favorable. Electrophoresis phenomenon describes the movement of particles submerged in fluids under the influence of the electrostatic field. The movement happens because of the electric charge associated with the intersurface of fluid and solid. Due to the chemical affinity, the defects in the crystalline structure of the solid surface, and the adsorption, the ions of the fluid attach to the solid surface. This process is generally selective. Therefore, the intersurface between the fluid and the solid gains a net charge. The thickness of this charge layer is usually represented by the Debye's length $\kappa^{-1} = \sqrt{\epsilon\epsilon_0 kT/e^2 \sum_i z_i^2 n_i}$. For large objects, the effect of this charge layer is negligible. However, for small particles, where the surface-over-volume ratio is high, the charged surface becomes more important. When an external electric field is applied, the charged particles move.

Chromatography

Gel electrophoresis is traditionally used to analyze the biomolecules such as DNA, RNA, or proteins. The gel is a porous matrix, usually made from polymer. Several wells are punched in the gel to hold the analyte. When the external electric field is applied, the analyte moves across the gel. Because of the difference in size, mass, shape, and charge of the molecules, different kinds of the molecules migrate to different distances. The molecules are

then stained and then observed to obtain the information about the molecular distribution. For small molecules such as proteins or shorter DNA, the polyacrylamide gel is preferred. For the molecules whose size is about 100 bases, the carbohydrate-based matrix agarose is used. However, conventional gel electrophoresis shows to be less effective for large molecules.

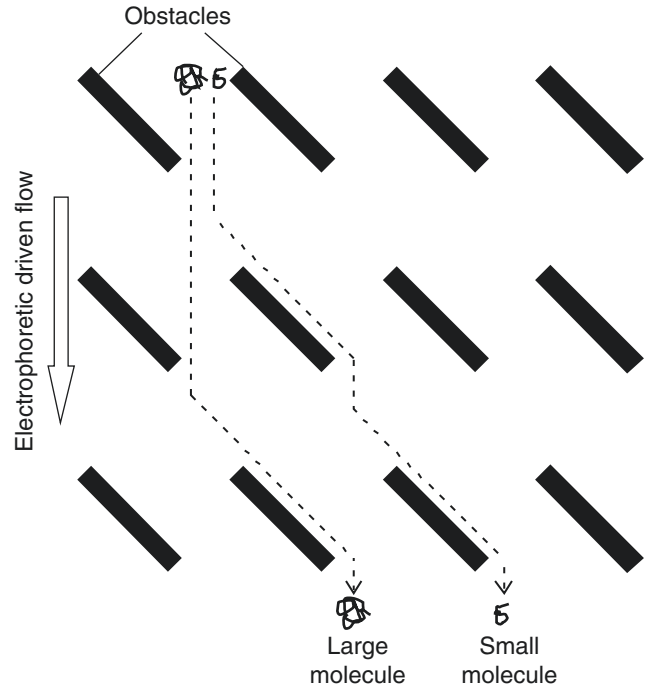
With the introduction of new devices, which is better controlled and fabricated, the applications of nanofluidics are widely expanded. General perspectives of these new devices and their application were given by Abgrall and Nguyen [1]. Random patterns of the porous material can be substituted by more regular fabricated pillar arrays. Bakajin et al. introduced an improving technique to separate DNA strands using switching electric fields [2] along different orientations. The DNA strands react with the electric field by first realigning their orientation and then migrating along the direction of the electric field. Longer strands take more time to realign; therefore, they have less time to migrate in a cycle. The shorter strands therefore migrate through longer distances than the long ones. The authors demonstrated the separation of 100-kbp DNA in 10 s.

However, in the pillar array design, the large molecules tend to clog the array and easily break up. Baba et al. introduced a new structure to solve the problem [3]. The structure consists of parallel wide channels connected with narrower channels run perpendicular from the wide channels. Electrophoresis is used to drive molecules along the wide channels. Molecules smaller than the narrow channel can enter the narrow channel by Brownian motion, while the large molecules remain in the wide channels. As a result, the large molecules migrate in faster than the small molecules. Han et al. performed a series of studies in DNA separation in entropic trap [4]. The entropic trap is a pattern of alternative deep (1.5–3-μm) and shallow (75–100-nm) regions. The molecules were driven across the pattern by electrophoresis technique. The system showed effective separation of long-chain DNA. The depth value can be changed to optimize for different size range of molecules. The devices showed to be effective in separating long DNA (5–200 kbp) within 30 min. This principle is also applied to separate protein or smaller DNA in similar devices [1].

Chromatography methods can also be performed in well-defined microchannels or nanochannels rather than in porous material. Various studies have been carried out to establish the relation between the species retention and coefficient of dispersion as functions of Debye's length and the species charge [1]. The channel size is crucial in electrophoresis separation effect. Pennathur and Santiago [5] introduced an investigation on the electrokinetic transport in nanochannels. The investigation described a complete dependence of the effective mobility of species in electrophoretic motion on the bulk electrolyte mobility values, zeta potential, ion valence, and background electrolyte concentration. From this theory, the authors suggested a method called electrokinetic separation by ion valence. This method determines the ion valence and the ion mobility by comparing transport motion of species in micro- and nanochannels.

Beside electrophoresis, pressure-driven flow can also be used in DNA chromatography, as demonstrated by Wang et al. [6]. The DNA sample was injected into the 500-nm channel. The sample was pushed along the channel with pressure. The longer molecules migrate faster than the shorter ones. Real demonstration with DNA sample from *Arabidopsis* showed that this method is effective. The method, therefore, also promises a great potential to be used for protein and other biomolecules.

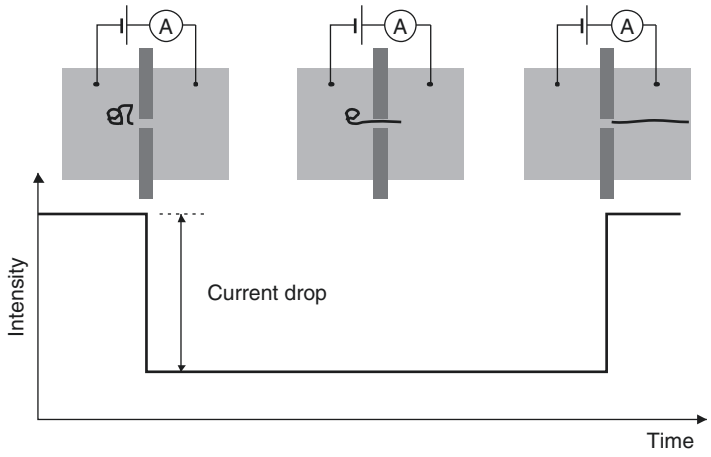
Other concepts such as shear-driven separation and anisotropic continuous-flow separation are also used to separate DNA. Shear-driven separation employs the relative movement of two parallel substrates separated by a spacer. This method does not require a high pressure as in pressure-driven separation. It also does not use electric field as in electrokinetic-driven method; therefore, it is suitable for samples which are sensitive to electric field. It is also faster than electrokinetic-driven method to reach the optimal velocity for separation. Experiments with dyes showed effective separation within only 0.2 s and over a distance of 1.8 mm [7].



Applications of Nanofluidics, Fig. 1 Array of oblong obstacle used to sieve biomolecules. The smaller molecules suffer more from diffusion and deflect further from the electrophoretic direction

Continuous-flow separation technique offers advantage such as real-time monitoring or integration of signal over time. Duke and Austin [8] introduced a sieving mechanism for biological molecules. A stream of molecules is driven by electrophoresis along a sieve consisting of oblong obstacle in oblique direction (Fig. 1). Besides the electrophoretic movement, the molecules also move orthogonally depending on their diffusion coefficient. Therefore, the molecules with different sizes follow different trajectories. This principle was tapped to sieve the DNA molecules of different sizes [1]. Huang et al. also used an array of pillars for sieving, but instead of natural diffusion process, a weaker electric field pulse is used to sieve the molecules [9]. The sieving action is done by alternating strong and weak electric field pulses in different directions. Because of the apparent similarity between the sieving phenomenon and optical dispersion, the author proposed the term "DNA prism" for the device. Using this DNA prism, separation of DNA with sizes in the range of 61–209 kbp was completed within 15 s. The principle can be applied with different shapes and arrangements of the obstacles to optimize the sieving effect. Fu et al. [10] introduced a device with parallel deep channels connected by perpendicular nanochannels. The macromolecules such as DNA and proteins are driven by diagonal electric field. One component of the electric field drives the molecules along the deep channels, while the other component leads the molecules to jump from one deep channel to another through the nanochannels. The molecules jump more or less frequently depending on their size. As a result, channels with different depths contain molecules of different size ranges.

Separation effect such as in chromatography is usually used for analysis purpose. However, it is also useful in preparation of solution, thanks to the electrokinetic mechanism. Kuo et al. demonstrated the delivery of analyte from one microchannel (injection) to another (collection) [11]. The rate of delivery is controlled by bias voltage across two microchannels and the type of analyte. Different configurations of voltages (specified, ground, or float) at four ends of the injection and collection channels shows different levels of control. Karnik et al. [12] realized the nanofluidic transistor circuit to control the



Applications of Nanofluidics, Fig. 2 Translocation of DNA molecules through a nanopore, causing a drop in electric current intensity

transport of protein between microchannels by connecting nanochannels, serving as the gate. The electric field across the gate is controlled with electrode. The gate then can be switched on or off to control the delivery. Microfabricated porous membrane and nanochannels were used as electrical current passages incorporated in the microchannel system; preconcentration with a factor of up to millions can be reached [1].

Single-Molecule Analysis

The size of the nanoscale structures allows them to capture less and less molecules, approaching the goal of handling individual macromolecules for analysis. With recent advances in nanotechnology, the single-molecule sensing technique has improved significantly. Conventionally, molecules such as DNA are handled in porous material. The properties of the molecules are studied statistically on a large population, usually obtained after amplification processes, such as polymerase chain reaction (PCR). The well-defined nanostructure allows the study of the motion of individual DNA. The obtained information is then more specific than that with statistical approach.

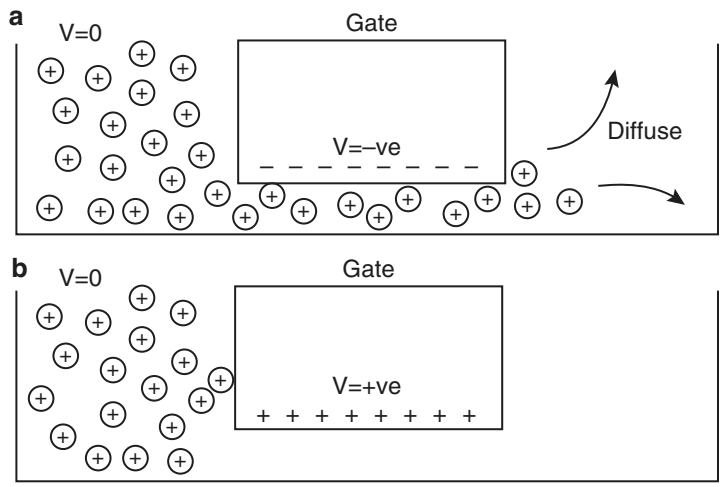
When driven into the nanochannels, the DNA molecules separate from one another. Fluorescent-labeled technique was used to capture the individual images of each molecule. By correlating the photon burst size obtained with the length of the DNA, the length of each molecule can be measured. This technique also can be used to count the number of protein bound to DNA using integrated photon molecular counting method. It was shown that the minimum channel size required to stretch a DNA molecule not only depends on the length of the DNA but also on the ionic strength of the buffer [1]. In a study performed by Krishnan et al. [13], the DNA molecules introduced by capillary filling in a nanochannel were observed to extend at the edge of the channel. This phenomenon promises a simple way to stretch the DNA molecules. However, the physics of this phenomenon is still unknown. Zevenbergen et al. [14] introduced a device to perform redox reaction of molecules in a nanodevice. The solution is filled in a nanochannel with two electrodes on top and at the bottom. The potential of each electrode can be adjusted individually with respect to a reference Ag/AgCl electrode immersed in the solution. The product generated at one electrode becomes the reactant at the other electrode and vice versa. The process is called redox cycling. The redox cycling enables one molecule to contribute many electrons to the induced current, making it possible to detect a small number of molecules by measuring the induced current. The amplification factor up to 400 and a resolution down to 70 molecules were detected.

Figure 2 explains the principle of reading information of individual DNA strand using its translocation through a nanopore. The DNA moves

through the nanopore under an applied electric field. During the movement, the DNA partially blocks the pore. The ionic movement along the pore reduces, leading to a drop in the electric current. If the molecules move across the pore individually, each down peak of the current associates with one translocation. With the nanopore small enough, the DNA passes across the nanopore linearly. Then, the DNA length can be deduced from the time of the translocation. However, the traveling speed depends both on the composition (A, C, G, or T) and the orientation (3′ or 5′) of the DNA [1]. Therefore, measuring the length of DNA is not a simple problem, unless the interaction between the nanopore and the molecules is well understood. To analyze such an interaction, Keyser et al. [15] placed a double-stranded RNA molecule attached to a bead by optical tweezers near the nanopore. The strand was pulled into the pore by electric field, but the optical restoring prevents the complete translocation of the molecules. The optical restoring force causes a dislocation of the beads which can be measured. The translocation also caused a drop in electric current. Correlation between the dislocation of the bead and the electric current drops reveals the interaction between the molecules and the nanopore. The information of such interactions provided tools for further studies in single-molecule DNA sequencing.

Nanofluidic Electronics

Applications of nanofluidics require the fluid flow controlling in nanoscale, where most of conventional fluid controlling methods are impossible to implement. Electrokinetic properties of the fluid in nanoscale inspire the use of electric field to control the fluid, similar to its use in controlling the electric current in diodes and transistors; this idea leads to the concept of nanofluidic electronics [1]. In a nanochannel, it is easier to create a strong electric field with a relatively small value of electric potential. The electric field in the nanochannel attracts or repulses the counterions or coions. Therefore, the flow of counterions and coions through the channels can be switched on or off. This principle is illustrated in Fig. 3. Gajar et al. [16] introduced the concept of ionic liquid-channel field-effect transistor, which uses the electric field at the gate to control the concentration ratio of cation and anion. For demonstration with solution of KCl in glycerol in 88-nm-thick channel, when the gate voltage changed from 0 to 25 V, the anion/cation ratio increases by 17.2, the conductance increases by 5.2. This device is not meant to substitute traditional transistor in electronic circuit but to inspire the possibility that the flow of chemical can be logically controlled, which gives great advantage in process automation. The same principle was also employed by

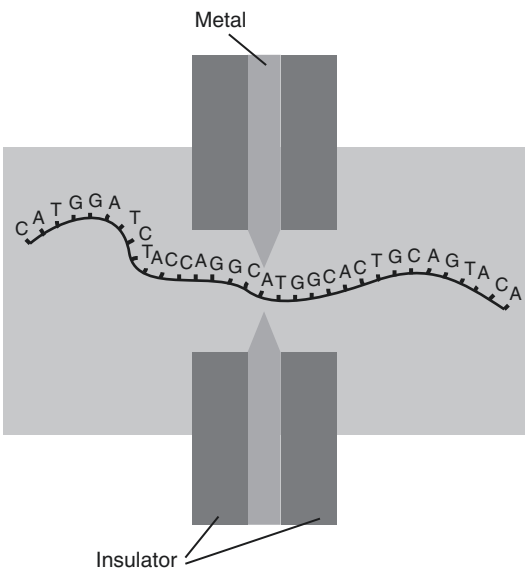


Applications of Nanofluidics, Fig. 3 Working principle of field-effect nanofluidic transistor. (**a**) A negative gate potential attracts cations in the nanochannel, making them possible to pass through the channel. (**b**) A positive gate potential repulses the cations from the nanochannels, blocking the flow of cations through the channel

other authors to control the transport of proteins or in the field-effect transistor biosensors [1]. The transistor principle can also be realized with carbon nanotube instead of nanochannels [17]. Especially, surface treatment of the nanotube can modify the polarity of the transistor; therefore, many types of transistor structures can be made in a similar way. Besides the transistor, the idea of nanofluidic diodes was also demonstrated with both nanochannels and nanotubes [1].

Future Directions

Apart from these broadly developed applications, nanofluidics suggests many other applications. The ability to control the diffusivity makes it greatly advantageous in drug delivery in vivo [1]. The diffusivity of the chemicals can be controlled by changing the size and shape of the porous pore. Therefore, it is possible to determine the lag time after which the drug will release, as well as the drug-releasing rate. Gersborg-Hansen et al. [18] integrated the microfluidic channels in a resonator for dye laser. The laser wavelength can be alternated by changing the dye or solvent. Theoretical models have been developed to directly read the sequence of a DNA molecule as it moves linearly through the nanopore like the magnetic tape reading in a tape deck [1], as illustrated in Fig. 4. The nanostructures for energy emerge as a promising field, as the need for new source of energy is increasing. It was showed that conversion of hydrostatic to electrical energy has a maximum efficiency of 12 % with aqueous-based solutions of lithium ions [19]. In practice, a lower efficiency was achieved in preliminary experiments but the conditions to improve the efficiency were also sketched out [1]. Current prototypes of fuel cell are mainly based on proton exchange across the membranes. However, because of the enrichment of counterions in a nanochannel, Liu et al. [20] proposed the use of fabricated nanochannels to substitute the porous membranes in fuel cell.



Applications of Nanofluidics, Fig. 4 Reading of DNA sequence as the DNA moves through a nanopore. Preliminary theoretical principles for such an application have been already sketched out

Cross-References

- ▶ Nanostructure Field Effect Transistor Biosensors
- ▶ Nanostructures for Energy
- ▶ Nanotechnology Applications in Polymerase Chain Reaction (PCR)
- ▶ Optical Tweezers

References

1. Abgrall, P., Nguyen, N.T.: Nanofluidic devices and their applications. Anal. Chem. **80**(7), 2326–2341 (2008)
2. Bakajin, O., Duke, T.A.J., Tegenfeldt, J., Chou, C.F., Chan, S.S., Austin, R.H., Cox, E.C.: Separation of 100-kilobase DNA molecules in 10 seconds. Anal. Chem. **73**(24), 6053–6056 (2001)
3. Baba, M., Sano, T., Iguchi, N., Lida, K., Sakamoto, T., Kawaura, H.: DNA size separation using artificially nanostructured matrix. Appl. Phys. Lett. **83**(7), 1468–1470 (2003)
4. Han, J., Craighead, H.G.: Characterization and optimization of an entropic trap for DNA separation. Anal. Chem. **74**(2), 394–401 (2002)
5. Pennathur, S., Santiago, J.G.: Electrokinetic transport in nanochannels. 1. Theory. Anal. Chem. **77**(21), 6772–6781 (2005)
6. Wang, X., Wang, S., Veerappan, V., Chang, K.B., Nguyen, H., Gendhar, B., Allen, R.D., Liu, S.: Bare

nanocapillary for DNA separation and genotyping analysis in gel-free solutions without application of external electric field. Anal. Chem. **80**(14), 5583–5589 (2008)
7. Clicq, D., Vankrunkelsven, S., Ranson, W., De Tandt, C., Baron, G.V., Desmet, G.: High-resolution liquid chromatographic separations in 400 nm deep micro-machined silicon channels and fluorescence charge-coupled device camera detection under stopped-flow conditions. Anal. Chim. Acta **507**(1), 79–86 (2004)
8. Duke, T.A.J., Austin, R.H.: Microfabricated sieve for the continuous sorting of macromolecules. Phys. Rev. Lett. **80**(7), 1552–1555 (1998)
9. Huang, L.R., Tegenfeld, J.O., Kraeft, J.J., Sturm, J.C., Austin, R.H., Cox, E.C.: A DNA prism for high-speed continuous fractionation of large DNA molecules. Nat. Biotechnol. **20**(10), 1048–1051 (2002)
10. Fu, J., Schoch, R.B., Stevens, A.L., Tannenbaum, S. R., Han, J.: A patterned anisotropic nanofluidic sieving structure for continuous-flow separation of DNA and proteins. Nat. Nanotechnol. **2**(2), 121–128 (2007)
11. Kuo, T.C., Cannon Jr., D.M., Chen, Y., Tulock, J.J., Shannon, M.A., Sweedler, J.V., Bohn, P.W.: Gateable nanofluidic interconnects for multilayered microfluidic separation systems. Anal. Chem. **75**(8), 1861–1867 (2003)
12. Karnik, R., Castelino, K., Majumdar, A.: Field-effect control of protein transport in a nanofluidic transistor circuit. Appl. Phys. Lett. **88**(12), 123114 (2006)
13. Krishnan, M., Mönch, I., Schwille, P.: Spontaneous stretching of DNA in a two-dimensional nanoslit. Nano Lett. **7**(5), 1270–1275 (2007)
14. Zevenbergen, M.A.G., Krapf, D., Zuiddam, M.R., Lemay, S.G.: Mesoscopic concentration fluctuations in a fluidic nanocavity detected by redox cycling. Nano Lett. **7**(2), 384–388 (2007)
15. Keyser, U.F., Koeleman, B.N., Van Dorp, S., Krapf, D., Smeets, R.M.M., Lemay, S.G., Dekker, N.H., Dekker, C.: Direct force measurements on DNA in a solid-state nanopore. Nat. Phys. **2**(7), 473–477 (2006)
16. Gajar, S.A., Geis, M.W.: Ionic liquid-channel field-effect transistor. J. Electrochem. Soc. **139**(10), 2833–2840 (1992)
17. Fan, R., Yue, M., Karnik, R., Majumdar, A., Yang, P.: Polarity switching and transient responses in single nanotube nanofluidic transistors. Phys. Rev. Lett. **95**(8), 1–4 (2005)
18. Gersborg-Hansen, M., Balslev, S., Mortensen, N.A., Kristensen, A.: A coupled cavity micro-fluidic dye ring laser. Microelectron. Eng. **78–79**(1–4), 185–189 (2005)
19. Van Der Heyden, F.H.J., Bonthuis, D.J., Stein, D., Meyer, C., Dekker, C.: Electrokinetic energy conversion efficiency in nanofluidic channels. Nano Lett. **6**(10), 2232–2237 (2006)
20. Liu, S., Pu, Q., Gao, L., Korzeniewski, C., Matzke, C.: From nanochannel-induced proton conduction enhancement to a nanochannel-based fuel cell. Nano Lett. **5**(7), 1389–1393 (2005)

Arc Discharge

► Physical Vapor Deposition

Arrays

► Moth-Eye Antireflective Structures

Arthropod Strain Sensors

Friedrich G. Barth
Life Sciences, Department of Neurobiology, University of Vienna, Vienna, Austria

Synonyms

Biological sensors; Biosensors; Cuticle; Exoskeleton; Fiber reinforced composite; Insects; Mechanoreceptors; Sense organ; Sensors; Spiders

Definition

Arthropods have highly refined sensors which provide them with a detailed picture of the mechanical events going on in their exoskeleton. These exoskeletal strain sensors are innervated holes embedded in the stiff cuticle that locally increase compliance when the exoskeleton is loaded by internal or external forces. Their deformation in the nanometer range suffices to set off nervous impulses. Arthropod strain sensors are called slit sensilla in arachnids and campaniform sensilla in insects. Their function is particularly closely linked to locomotion.

Introduction

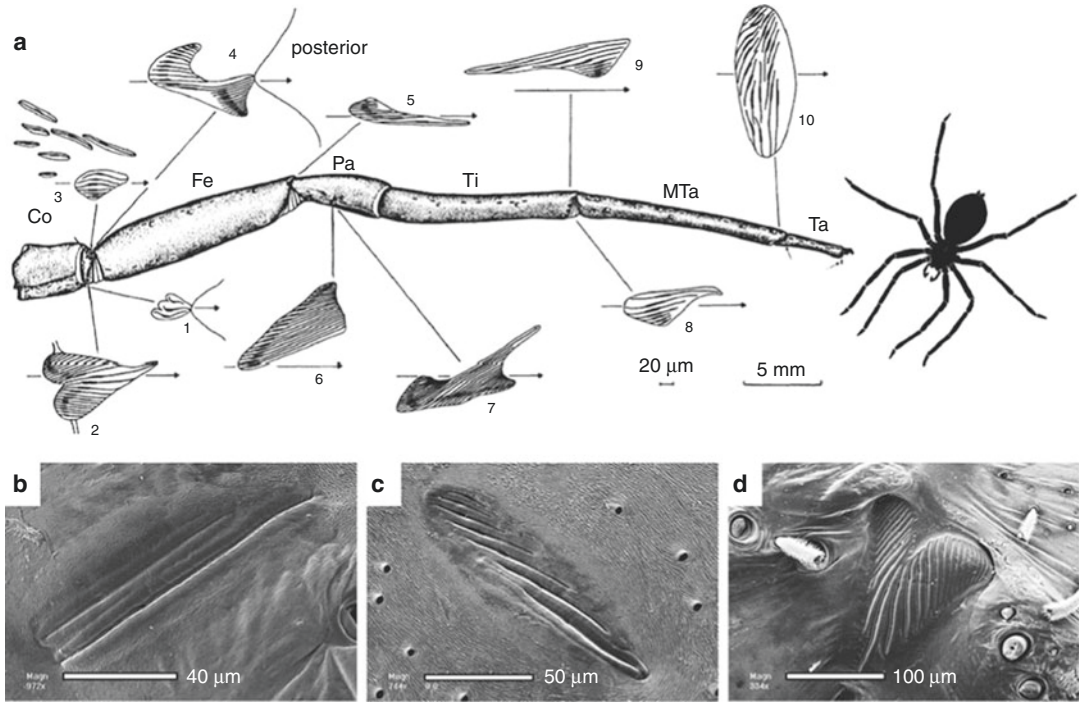
Arthropods are well informed about the mechanical events in their outside and inside worlds. They have numerous highly developed sense

organs which respond to a large variety of mechanical events such as tactile, acoustic, vibratory, and airflow stimuli. Whereas their proprioreceptive sensors tell them about self-generated mechanical stimuli like the position and movements of joints and hemolymph pressure, exteroreceptive sensors keep them informed about the external world. In both cases it is in particular changing conditions as opposed to stationary ones which they respond to and use for the guidance of their behavior.

This chapter deals with a sense alien to us humans and specifically linked to the cuticular *exoskeleton*, a structure most typical of arthropods and of dominant importance in regard to their behavior and way of life. The exoskeletons of insects and spiders and related animals not only provide mechanical support and protection and attachment sites for the musculature. They also advertise the animal to mates and conspecifics or conceal it from predators if adequately shaped and colored. In addition to all that, the exoskeleton carries tens or even hundreds of thousands of sensory organs, called sensilla, and has a particular sensory function not known in vertebrates. Embedded into their exoskeleton, insects and arachnids have *strain sensors* which may be considered analogous to technical strain gauges (reviews [1, 2]). These biological strain gauges form tiny holes in the cuticular exoskeleton which represent areas of increased mechanical compliance and locally amplified deformation. When the exoskeleton is loaded by internal or external forces, the resulting minute strains in the cuticle deform the sensors. This deformation in turn elicits nervous responses by straining the dendritic membrane of sensory cells attached to them [1, 3]. Apart from actually measuring strains, the biological strain sensors have control functions as parts of reflex and more complex nervous circuits which help the animal to detect substrate engagement, to temporally and spatially control its movements and posture, to adapt them to different environments, and to avoid overloading of the exoskeleton [2, 4, 5]. In night-active spiders it was shown that exoskeletal strain sensors even help the animal to remember motion sequences which it can later use to find back to a starting point where they might have lost a prey animal or where they usually return to as their home site or retreat. The zoologists refer to such a remarkable way of orientation as kinesthetic or idiothetic orientation [1, 6].

The largest *number* of strain sensors so far known in an arthropod is that found in the spider *Cupiennius salei* [1, 7]. The presence of some 3300 strain sensors suggests that there is much detail in the picture the spider receives about the spatiotemporal patterns of the mechanical events going on in its exoskeleton. The vast majority of the sensors (86 %) are embedded in the stiff sclerotized cuticle of the walking legs and pedipalps whose mechanical properties are close to that of bone with a Young's modulus of around 18 GPa [8]. The *slit sense organs or slit sensilla*, as the strain sensors are called in spiders and their arachnid relations because of their slit-like form, have aspect ratios (length *l*/width *b*) of up to more than 100, being from 8 to 200 μm long and 1–2 μm wide. They come in different configurations as individual slits, loose groups of slits, and so-called lyriform organs (Figs. 1 and 2). The latter are the most intriguing type of slit sensilla. Lyriform organs, of which *Cupiennius* has 144 (the majority on the legs, Figs. 1a and 2a), consist of arrays of up to about 30 slits of different length closely arranged in parallel. Some of these organs show bizarre outlines which substantially impact their mechanical behavior and as a consequence the information made available to the central nervous system.

In insects, the equivalent (analogous) strain sensors are called *campaniform sensilla* (Fig. 3). Their strain sensors usually form oval, more rarely round, holes in the cuticle which are covered by a dome-shaped cuticular membrane to which the dendrite of a mechanosensory cell attaches. The organs are called campaniform sensilla because the dendrite and the covering membrane remind of a bell and its clapper. In different campaniform sensilla the aspect ratios range from 1 (round) to rarely more than 3 [1]. The most complete mapping of campaniform sensilla is available for a fly (*Calliphora vicina*), where a total of about 1200 such campaniform sensilla were found [9, 10]. Similar to arachnid slit sensilla campaniform

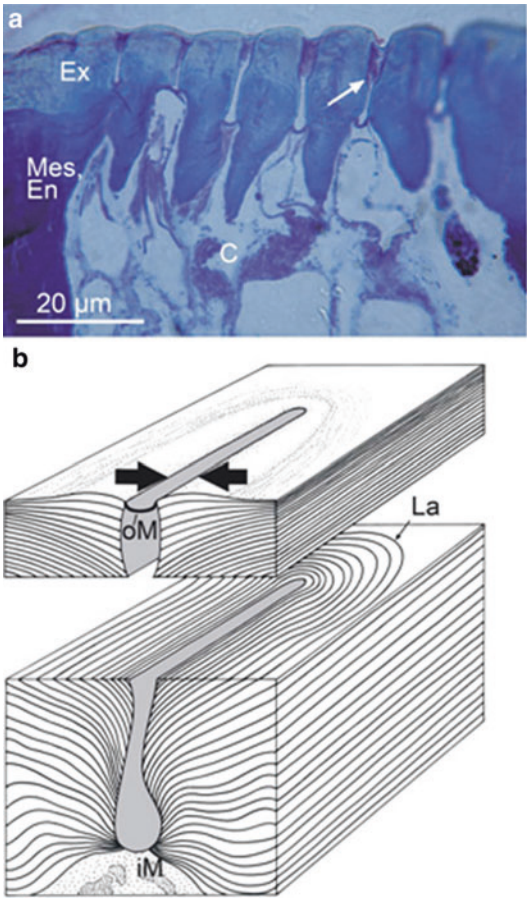


Arthropod Strain Sensors, Fig. 1 (**a**) Arachnid slit sensilla. Posterior aspect of a spider leg (*Cupiennius salei*) showing the occurrence of slit sense organs. Enlarged drawings of lyriform organs with details of slit arrangements (from Ref. [7], with permission) (**b**) Scanning electron micrographs of three different lyriform organs (courtesy R. Müllan) *Co* coxa, *Fe* femur, *Pa* patella, *Ti* tibia, *MTa* metatarsus, *Ta* tarsus

sensilla occur as single sensilla, loose groups, and what Grünert and Gnatzy [10] referred to as fields. In *Calliphora vicina* there were 86 single campaniform sensilla, 52 groups consisting of 3–32 sensilla with a total of 350 sensilla, and 12 fields with a total of 730 sensilla. As much as 55 % (about 730) of all campaniform sensilla were located on the halteres (see Fig. 3a and below) arranged in 10 of the 12 sensilla fields, about 250 sensilla were on the wings and ca. 210 on the walking legs underlining their central role in the coordination of locomotion.

Like in the case of technical strain gauges, the *mechanical parameter measured* depends to a large extent on the sensor's location and arrangement in the exoskeleton. In arthropods, strain sensors tell the animals about sources of load such as muscle contraction, hemolymph pressure, vibration of the substrate they are sitting on, wing deformation during flight due to aerodynamic forces and – in the case of dipteran flies – the Coriolis forces generated during flight (see section "Sensor Arrays"). A special case are insect strain sensors closely associated with the base of wind-detecting sensory hairs and responding when these are strongly stimulated.

After a short description of some functional principles of individual arthropod strain sensors, the main focus of this chapter will be on the intriguing questions of what the mechanical implications of the strange patterns of sensor arrangement in the lyriform organs may be and how nature managed to successfully maintain the mechanical intactness of the exoskeleton using holes as sensory structures (which an engineer always tends to avoid because of stress concentration problems leading to cracks and fracture). As experimental analyses and Finite Element modeling show, quantitative predictions can be made in regard to the mechanical effects of a number of morphological features which are



Arthropod Strain Sensors, Fig. 2 (**a**) Histological cross section through the slits of a lyriform organ. The outer and the inner covering membranes are clearly seen. *Ex* stiff, mechanically dominating exocuticle; *Mes, En* meso- and endocuticle, respectively; *c* cellular components. (**b**) Arrangement of cuticular laminae (*La*) around a slit according to 3D reconstruction. *oM* outer membrane, *iM* inner membrane which one of the two sensory cells penetrates to end at *oM*, whereas the dendrite of the other sensory cell ends at *iM* (**a** and **b** courtesy R. Müllan)

indeed seen to vary in nature. These predictions will be a valuable basis for the design of bio-inspired synthetic sensors used for technical applications.

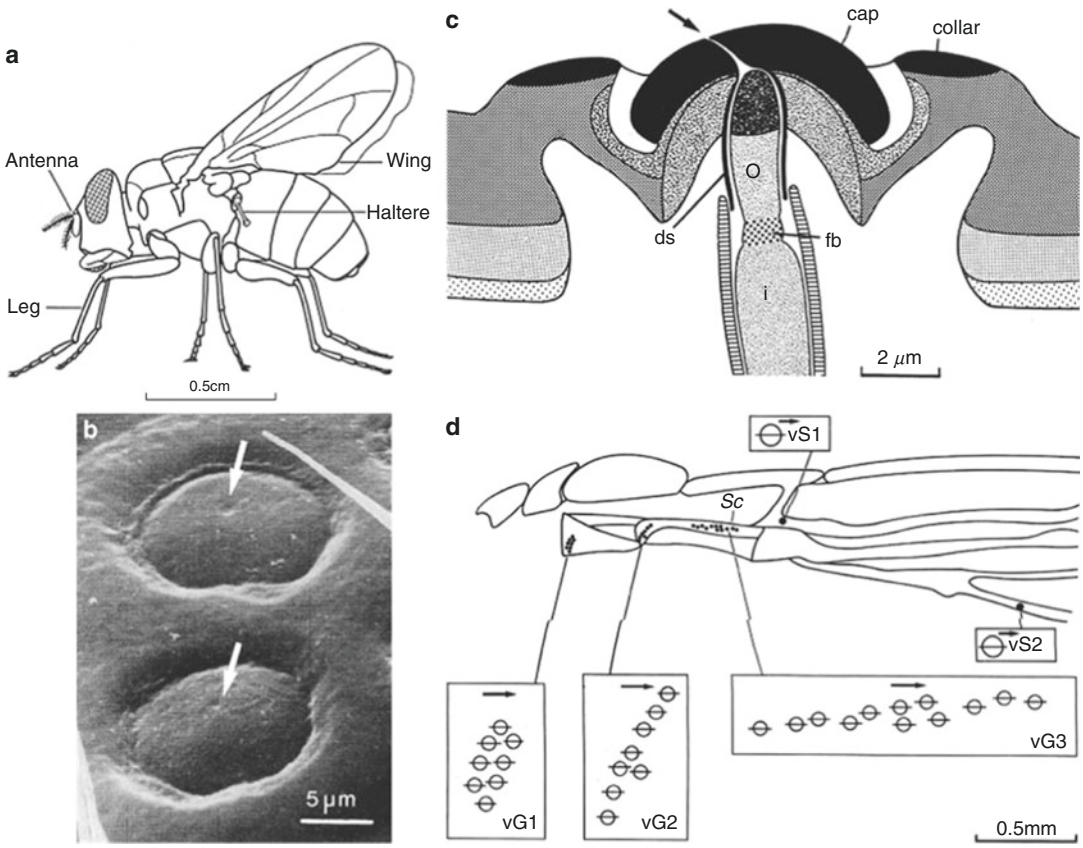
Form and Function

Individual Sensors

According to experimental studies the adequate stimulus of slit sensilla is compressive loading [1]. As demonstrated by Finite Element analyses [11] and earlier pioneering experiments with polymer models [1, 12], several different geometrical parameters affect stimulus transformation and thus the mechanical sensitivity of a slit. One of these parameters is (a) *orientation* in regard to load direction. Single slits are highly directional and best suited for detecting compressive stresses acting normal to their long axis. The same applies to oval campaniform sensilla responding to compression at right angles to their long axes [13]. The other parameter to which slit compression is most sensitive is (b) *slit length l* which indeed varies between about 8 and 200 µm in *Cupiennius* and is the most appropriate parameter to adjust slit compression. Slit length linearly scales with slit face displacement, that is, magnitude of stimulation. A slit's directionality increases with aspect ratio and slit length as well. Remarkably, the (c) *aspect ratio* (*l*/*b*) of slits plays a minor role in this context when assuming values between 20 and 100. However, when comparing a slit of aspect ratio 1 (reminding of some insect campaniform sensilla) with a reference slit of aspect ratio 100 (well within the range of arachnid slit sensilla) the deformation in the former exceeds that of the latter by 50 %. This is mainly due to the area covered by such round sensilla, which is about 80 times larger than in case of a sensor with an aspect ratio of 100. Similarly, (d) capped rectangular slits are compressed more than elliptic holes of identical aspect ratio due to their larger area. Apart from straight slits there are (e) C- and S-shaped slit geometries which, as a rule of thumb, are less mechanically sensitive than the straight slit under normal compressive loading. If one had to adjust/modify the mechanical sensitivity in a synthetic slit sense organ the primary parameter to adjust is slit length. Since slit deformation, which ultimately induces strains in the dendritic membrane and thus nervous activity, is largest in the middle along the slit's long axis mechanical sensitivity could also be adjusted (reduced) by moving the (f) dendrite attachment site toward the slit's ends.

Sensor Arrays

By forming arrays of slits in the lyriform organs, the measurement capabilities at a particular

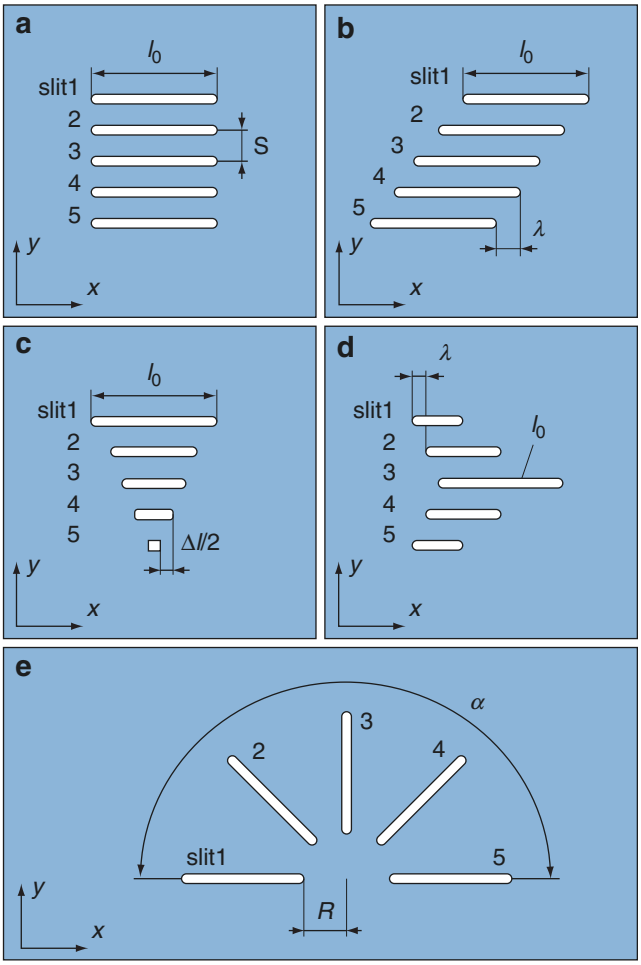


Arthropod Strain Sensors, Fig. 3 Insect campaniform sensilla. (**a**) The fly *Calliphora vicina*. Note halteres which are modified wings. (**b**) Scanning electron micrograph of two campaniform sensilla on the fly leg. (**c**) Schematic drawing of section through stimulus conducting structures of a campaniform sensillum. The different components seen in transmission electron micrographs are shaded differently. *o, i* outer and inner segment of the sensory cell's dendrite, with a so-called fibrillar body (fb) in between; *ds* cuticular dendritic sheath. (**d**) Arrangement of campaniform sensilla on the ventral aspect of the fly's wing base (**a** and **d** from Ref. [9]; **b** and **c** from Ref. [10], with permission)

location of the exoskeleton can be substantially increased. What is it that a lyriform organ can do more and better than a single slit?

Due to limited spatial resolution, it is quite difficult to reveal the mechanical tricks underlying the various arrangement patters found in spider lyriform organs by direct observation. Modeling again has helped to better understand the principles [11, 12, 14]. Whereas the analytical approach of Kachanov, developed within fracture mechanics to calculate stress intensity factors at the tips of interacting cracks, can be applied to single slits and loose groups of slits (where the distance S between slits is larger than about ½ slit length l), there are difficulties with lyriform organs where typically $S < 0.05\ l$ [15]. A more efficient way to study the mechanical properties of a large variety of slit shapes, positions, and orientations in lyriform organs is finite element analysis. Thus the effects of lateral spacing (S), longitudinal shift (λ), and length gradation (Δl) on slit deformation could be studied under uniaxial normal load (in plane, normal to long axis of longest slit in the array) (Fig. 4). In short, some basic results are the following.

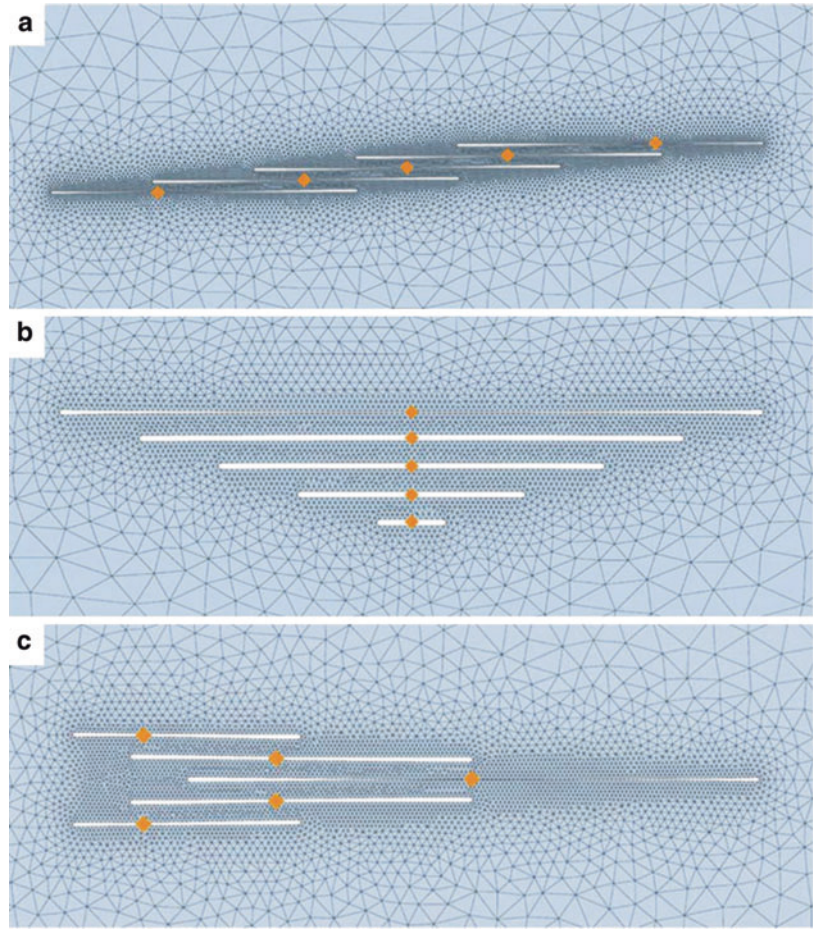
(1) For a group of parallel slits of equal length and as closely spaced as in real lyriform organs the inner slits are much less compressed than the



Arthropod Strain Sensors, Fig. 4 Some of the patterns of planar models studied by FE analysis. (**a**) Non-staggered array of five slits, (**b**) oblique bar arrangement, (**c**) triangular arrangement, (**d**) heart shaped arrangement, and (**e**) fan like arrangement. l_0 length of longest slit in array; S distance to neighboring slit; λ longitudinal shift between neighboring slits (From Ref. [11], with permission)

peripheral ones. The reason for this is the shielding effect of the outer slits which are deformed much more. (2) Slit compression may be larger by up to several hundred percent than that of a single slit if there is a longitudinal shift between slits. A longitudinal shift between neighboring slits is typical of natural lyriform organs; actually no case without it is known. (3) By arranging slits with different length in a group, mechanical sensitivity can be fractionated and the range of absolute and directional mechanical sensitivity of the entire organ extended. Each of the individual slits in one lyriform organ is innervated independently. Details are found in the papers by Hößl et al. [11, 14].

Considering different typical outlines of lyriform slit arrays (represented by five slits in the FE models) provides some answers to the question initially asked: What can a lyriform array do more or better than a single slit? A few examples shall illustrate this (Figs. 4 and 5). (1) In the "oblique bar" array the slits have the same length l but are longitudinally shifted by λ. The level of slit face displacement is similar in all slits and considerably larger than that of a single slit as long as $0.25 \leq \lambda/l \leq 2.5$. One might speculate that by such an arrangement and the central nervous convergence of the corresponding nervous information the signal-to-noise ratio (an ever-present problem in both biological and technical sensing) may be enhanced. (2) Another typical array has a "triangular outline" with slits of decreasing length where the shorter slits are strongly shielded at length gradations $\Delta l/l0 \geq 0.1$ (Δl, length

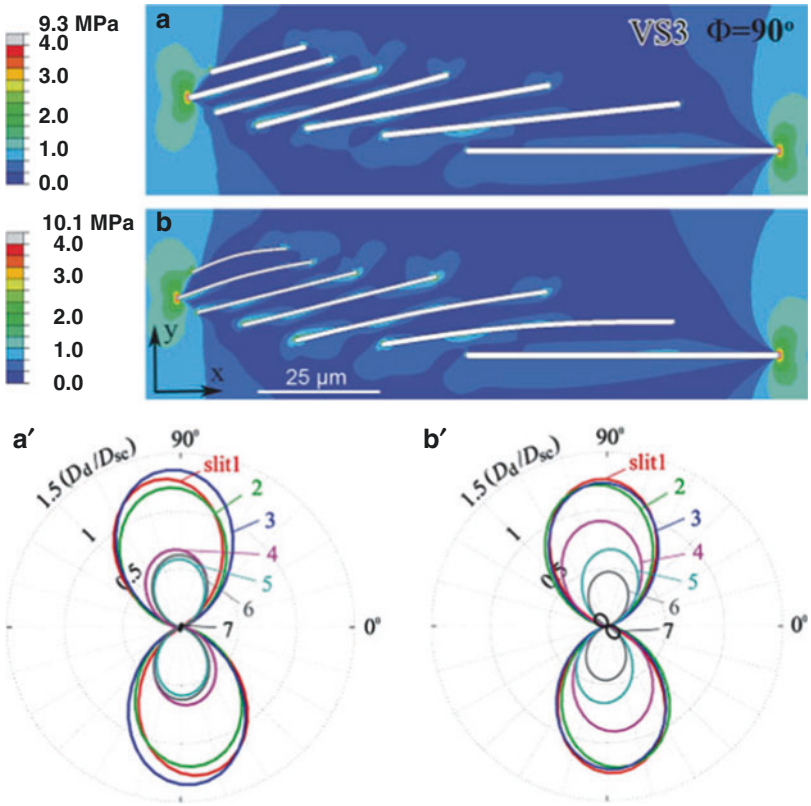


Arthropod Strain Sensors, Fig. 5 Deformed configurations of three finite element models subjected to uniaxial compressive load at right angle to slit long axes. The displacements are scaled up. (**a**) Oblique bar array with similar mid slit compression for all of the slits. (**b**) Triangular formation with linear gradation of slit length; note considerable shielding effect. (**c**) Heart-shaped array, where maximum slit deformation is not in the middle along the slits. Orange diamonds mark position of maximum compression (From Ref. [11], with permission)

difference of neighboring slits; l_0, length of longest slit) and will respond to relatively large strains (loads, respectively) only. The widely differing mechanical response in a triangular array suggests that it serves to maximize the range of signal magnitude the organ can cope with and exhibits a pronounced range fractionation. (iii) A third type of array is referred to as "heart-shaped" (Figs. 4 and 5). Similar to fan-like configurations with slit axes diverging in different directions, which are indeed known to occur in nature, one achieves sensitivity over a wide range of loading angles with heart-shaped arrays. This implies that even with particular types of parallel slit arrangements the range of effective load directions can be considerably enlarged. It also implies that such an array fractionates the stimulus both in regard to magnitude and direction. As Fig. 6 demonstrates, small differences in the shape and arrangement of slits in a lyriform array may affect the deformation pattern substantially.

The most spectacular arrays of insect campaniform sensilla are those at the base of the halteres of flies (Diptera) [9]. The halteres are club-shaped structures derived from the second pair of wings which oscillate during flight and work like gyroscopes, stimulated by Coriolis forces due to body rotations and providing inertial sensory information for flight control with very high temporal resolution [16] (Fig. 3a).

Vincent et al. [17] used finite element modeling to see stress concentration levels in order to analyze the reduction of the material strength by the "campaniform" holes. Their model was a multiple hole arrangement derived from a sensilla field on the base of the fly haltere and consisting of 4 rows



Arthropod Strain Sensors, Fig. 6 FE simulation of the directional sensitivities of the seven slits making up a lyriform organ on the spider leg of the oblique bar type. (**a**, **b**): Von Mises equivalent stresses (*MPa*) under loads normal to the long axis of slit 1, the longest slit in the array. (**a'**, **b'**): Directional mechanical sensitivity to uniaxial compressive loads in the simple (**a**) and more detailed (**b**) model formed according to scanning electron micrographs of the original organ. The mechanical sensitivity is given as the ratio between the deformation at the position of the dendrite (*Dd*) and the deformation in the middle of an isolated single slit (*Dsc*). Note the even fractionation of sensitivity among the slits in **b'** (From Ref. [14], with permission)

of 5 sensilla each (aspect ratio of sensilla 6.5). They found the maximum von Mises stress to be about 10 times higher than in case of the plate with no hole (stress concentration factor 11.1), whereas in case of the single hole it was only 4. Strain amplification in the multiple hole model was predicted to be $\times 8.75$ parallel to the loading. The effect of the aspect ratio, orientation of the holes, and the influence of other holes in the vicinity (distance between holes) were not studied.

Naturally Occurring Loads and Strains

An intriguing question is that of naturally occurring strains in the arthropod exoskeleton. It is these strains which must have shaped the evolution of the sensors receiving them. Again it is large spiders (*Cupiennius salei*; *Aphonopelma* spec.) where the corresponding quantitative measurements were done in some detail on the legs (where most of the sensilla are found) in living animals. Here the focus will be on just a few values. More details are found in the original literature [1, 8].

The three main sources of load are (1) *hemolymph pressure* which spiders use to extend their legs at joints lacking extensor muscles, (2) contraction of the *musculature* which is concentrated proximally in the legs so that the forces of inertia are kept low in these long and slender structures, and (3) *external sources* such as gravity and

substrate vibration. Under the conditions of slow walk at speeds of 1–10 cm/s the strains found at the site of lyriform organs varied between −13 and −20 με in a particular leg but differed in time course at different positions, relative to the walking pattern. Whereas one organ was stimulated during the stance or load bearing phase signaling muscle force (up to 200 mN) during joint flexion (strain due to resisted muscle action) another one responded to the increased hemolymph pressure (up to 5.3 kPa) during leg extension (swing phase). Since in areas without slit sensilla the strains occurring simultaneously may be significantly higher the conclusion is that lyriform organs are not necessarily located at sites of particularly high strain. In the given case the cyclical change of leg loading and thus the sensory timing of leg movements and the phases of walking seem to be a biologically more relevant aspect.

During situations like rapid starts and stops and jumps one would expect particularly high strain values. At a walking speed of 30 cm/s strains as high as −360 με were indeed measured in the cuticle of the joint region during the stance phase. During starts and stops and jumps the high accelerations imposed loads similar to those found during rapid walks. Clearly at least many of the lyriform organs are in regions with strongly varying strains and high peak loadings. According to both the model studies referred to above and electrophysiological analyses (see below), the lyriform organs are well adapted to monitor such variable load conditions and to avoid dangerous peaks by eliciting synergic reflexes [4].

Physiological Properties

For a lyriform organ on the spider leg tibia identified as HS-8 [7], which belongs to the "triangular" type (Fig. 3), the available analyses include FE-modeling, studies of organ deformation in the living animal using white light interferometry [14] and electrophysiological experiments [18]. By a comparison of all these data it turns out that the values for slit compression predicted by the FE analysis are very close to those obtained by direct measurement. At the electrophysiologically determined thresholds, where action potentials are elicited in the sensory cells, slit face displacements are as small as 1.7 nm. Clearly, physiologically relevant measurements down to the nanometer level are possible even in a material with a modulus of elasticity close to that of bone. The sequence of slits in the organ in regard to the magnitude of their deformation (and therefore sensitivity) follows slit length and is the same both in the FE models and in the interferometric measurements. Electrophysiologically determined threshold values differed by more than 40 dB (×100) among the slits of organ HS-8 and the linear parts of their working ranges complemented each other, thereby considerably increasing the linear range and range of high increment sensitivity of the entire lyriform organ as compared to a single slit [18].

Readers interested in the cellular mechanisms of mechanotransduction in slit sensilla are referred to reviews by French and Torkkeli [1, 3]. Zill et al. [2] reviewed the role of insect campaniform sensilla in load sensing and the control of posture and locomotion in some detail.

The Problem of Embedding Holes and Synthetic Sensors

Considering the basically simple design of arthropod strain sensors and the enormous range of their properties achieved by adjusting just a few parameters (whose relevance is understood quantitatively at least in slit sensilla), they are indeed of interest as models for the design of bio-inspired synthetic strain sensors for technical applications. Bio-inspired arrays would allow a highly resolved analysis of strains at small areas of particular interest and help the process of economizing (which is what natural evolution does) by saving material, weight, and energy. A network of sensors could help the decentralized steering of a robot and its feedback control in adaptive locomotion in particular in uneven terrain.

As embedded sensors [19] they would not cause a problem by being attached to the outside of a piece of material under load and thereby changing its mechanical properties in particular if the material studied is very thin. There are attempts to design high-performance MEMS sensors inspired by fly campaniform sensilla but the gap between the refinement and performance of the natural sensor and the engineered one is still considerable [17, 20].

In engineering one would normally try to avoid micro-holes like those of arthropod strain sensors because they are sources of stress concentration and the initiation of cracks. Obviously nature has found ways to effectively use holes for sensory purposes without endangering the intactness and provoking fracture of the material they are embedded in. There are several reasons why in nature the intactness of the material obviously is not a problem. (1) As already said the strain sensors not only measure strains but are also part of feedback loops and nervous circuits which can be used to avoid dangerous loads. (2) The resistance to tensile and pressure forces is increased close to the slits by an increased thickness of the stiff exocuticle around the opening (Fig. 2a). (3) The arthropod cuticle is a fiber-reinforced laminate material with the fibers arranged in layers parallel to the surface. Near the slits the fibers deviate from their arrangement parallel to the surface and follow a course reminiscent of the stress lines near a notch subjected to tensile force or pressure. They curve inward like the lines of greatest principal stress conforming to the contours of the slit. Thus the load-bearing capacity of the cuticle is largest in the direction of the greatest principal stress (Fig. 2b). (4) The sensory holes are not holes drilled into a homogeneous material but shaped within a fibrous laminate composite material while the exoskeleton is laid down. In line with this the chitinous microfibers are not interrupted by the hole but molded around it (Fig. 2b). (5) The ends of the notches are not sharp-edged but rounded (Fig. 2b).

The stress concentration factors SCF (local maximum equivalent stress/far-field stress) seen in the models of slit sense organs (which do not fully describe the natural case because they are homogeneous, in-plane isotropic plates of constant thickness) are about 22 at the ends of an isolated slit of aspect ratio 100 and only 4.2 when the aspect ratio is reduced to 2.5. For ellipses with aspect ratios of 2.5 as found in insect campaniform sensilla the SCF is about 6.5. For the lyriform groups highest SCFs were in the range between about 15 and 25 [14, 17].

Cross-References

- ▶ Biomimetics
- ▶ Biosensors
- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Insect Infrared Sensors
- ▶ Nanostructured Materials for Sensing
- ▶ Organic Sensors

References

1. Barth, F.G.: A Spider's World. Senses and Behavior. Springer, Berlin/Heidelberg (2002)
2. Zill, S., Schmitz, J., Büschges, A.: Load sensing and control of posture and locomotion. Arthropod Struct. Dev. **33**, 273–286 (2004)
3. French, A.S., Torkkeli, P.H.: Mechanotransduction in spider slit sensilla. Can. J. Physiol. Pharmacol. **82**, 541–548 (2004)
4. Seyfarth, E.-A.: Spider proprioception: receptors, reflexes and control of locomotion. In: Barth, F.G. (ed.) Neurobiology of Arachnids, pp. 230–248. Springer, Berlin/Heidelberg (1985)
5. Zill, S., Keller, B., Chaudry, S., Duke, E., Neff, D., Quinn, R., Flannigan, C.: Detecting substrate engagement: responses of tarsal campaniform sensilla in cockroaches. J. Comp. Physiol. A **196**, 407–420 (2010)
6. Seyfarth, E.-A., Barth, F.G.: Compound slit sense organs on the spider leg: mechanoreceptors involved in kinaesthetic orientation. J. Comp. Physiol. **78**, 176–191 (1972)
7. Barth, F.G., Libera, W.: Ein Atlas der Spaltsinnesorgane von *Cupiennius salei* keys. Chelicerata (Araneae). Z. Morph. Tiere **68**, 343–369 (1970)
8. Blickhan, R., Barth, F.G.: Strains in the exoskeleton of spiders. J. Comp. Physiol. A **157**, 115–147 (1985)

9. Gnatzy, W., Grünert, U., Bender, M.: Campaniform sensilla of *Calliphora vicina* (Insecta, Diptera) I. Topography. Zoomorphology **106**, 312–319 (1987)
10. Grünert, U., Gnatzy, W.: Campaniform sensilla of *Calliphora vicina* (Insecta, Diptera). II. Typology. Zoomorphology **106**, 320–328 (1987)
11. Hößl, B., Böhm, H.J., Rammerstorfer, F.G., Barth, F. G.: Finite element modeling of arachnid slit sensilla – I. The mechanical significance of different slit arrays. J. Comp. Physiol. A **193**, 445–459 (2007)
12. Barth, F.G., Ficker, E., Federle, H.U.: Model studies on the mechanical significance of grouping in spider slit sensilla (Chelicerata, Araneida). Zoomorphology **104**, 204–215 (1984)
13. Spinola, S.M., Chapman, K.M.: Proprioceptive indentation of the campaniform sensilla of cockroach legs. J. Comp. Physiol. **96**, 257–272 (1975)
14. Hößl, B., Böhm, H.J., Schaber, C.F., Rammerstorfer, F.G., Barth, F.G.: Finite element modelling of arachnid slit sensilla: II. Actual lyriform organs and the face deformations of the individual slits. J. Comp. Physiol. A **193**, 881–894 (2009)
15. Hößl, B., Böhm, H.J., Rammerstorfer, F.G., Müllan, R., Barth, F.G.: Studying the deformation of arachnid slit sensilla with a fracture mechanical approach. J. Biomech. **39**, 1761–1768 (2006)
16. Daniel, T.L., Dieugonne, A., Fox, J., Myhrvold, C., Sane, S., Wark, B.: Inertial guidance systems in insects: from the neurobiology to the structural mechanics of biological gyroscopes. J. Inst. Navig. **55**, 235–240 (2008)
17. Vincent, J.F.V., Clift, S.E., Menon, C.: Biomimetics of campaniform sensilla: measuring strain from the deformation of holes. J. Bionic Eng. **4**(2), 63–76 (2007)
18. Bohnenberger, J.: Matched transfer characteristics of single units in a compound slit sense organ. J. Comp. Physiol. **142**, 391–401 (1981)
19. Calvert, P.: Embedded mechanical sensors in artificial and biological systems. In: Barth, F.G., Humphrey, J. A.C., Secomb, T.W. (eds.) Sensors and Sensing in Biology and Engineering, pp. 359–378. Springer, Wien/New York (2003)
20. Wicaksono, D.H.B., Vincent, J.F.V., Pandraud, G., Craciun, G., French, P.J.: Biomimetic strain-sensing microstructure for improved strain sensor: fabrication results and optical characterization. J. Micromech. Microeng. **15**, S72–S81 (2005)

Artificial Muscles

► Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP)

Artificial Retina: Focus on Clinical and Fabrication Considerations

Pablo Gurman[1], Yitzhak Rosen[2] and Orlando Auciello[1]
[1]Materials Science Division, Argonne National Laboratory, Argonne, IL, USA
[2]Superior NanoBiosystems LLC, Washington, DC, USA

List of Acronyms

DOE	Department of energy
EP	Electric potentials
MD	Macular degeneration
OCT	Optical coherence tomography
RP	Retinitis pigmentosa
VPU	Video processing unit

Synonyms

Artificial vision; Bionic eye; Electronic visual prosthesis; Epiretinal implant; Retina implant; Retina silicon chip; Sub-retinal implant; Visual prosthesis

Definition

A device intended to restore the vision in patients suffering from vision loss caused by retinal disorders by replacing or improving the natural retina function or the vision pathway.

Visual Prosthesis for Retinal Disorders

Visual prostheses are currently under development by the scientific and industrial communities aimed to restore the vision for people suffering from retinal disorders such as retinitis pigmentosa (RP) and age-related macular degeneration (AMD). Several types of visual prosthesis have been developed to date by different groups using different approaches. All approaches focus on developing a so called "artificial retina."

The artificial retina includes a device intended to mimic the function of the photoreceptors in the natural retina. The natural retina is made of several layers. In retinal disorders, such as RP and AMD, the photoreceptor layer is compromised. The photoreceptor layer comprises photoreceptor cells that are responsible for capturing the incoming light and transforming it into electrical pulses that will travel to the brain where the image will be created. The artificial retina approach is based on the finding that, although the photoreceptor layer is dysfunctional, the rest of the retina (bipolar and ganglion cells) and the visual pathway is still capable of transmitting the visual information, as electrical pulses, to the brain.

This entry focuses on the Department of Energy's (DOE) artificial retina because it has been successfully developed and commercialized, although other prototypes using different approaches are being developed by other groups worldwide [1].

DOE Epiretinal Device Function

As with the natural retina, an artificial retina should be able to transform the light (a flow of photons) into an electrical pulse (a flow of electrons). This is performed by a camera mounted on a glass connected to a video processing unit (VPU) that is located on the belt of the patient. The camera receives the visual information and transmits this information to a VPU that converts this information into a digital format in the form of pixels that are sent to the internal unit. The internal unit uses a stimulator system to generate electrical pulses that activate the neuron cells through a microelectrode array placed in contact with the bodies of these neurons (the group of neurons that are thought to be activated are the ganglion cells). These cells trigger an electrical wave called action potential that transports the visual information codified in electric pulses through the optic nerve to the occipital cortex of the brain where the image percept is created.

Main Components of the DOE Artificial Retina

The device can be divided in two parts, an external component and an internal component:

External Components

Camera: A CCD camera, located on glasses used by the patient, collects the visual information in the form of pixels sending this information to a microprocessor located in the patient belt.

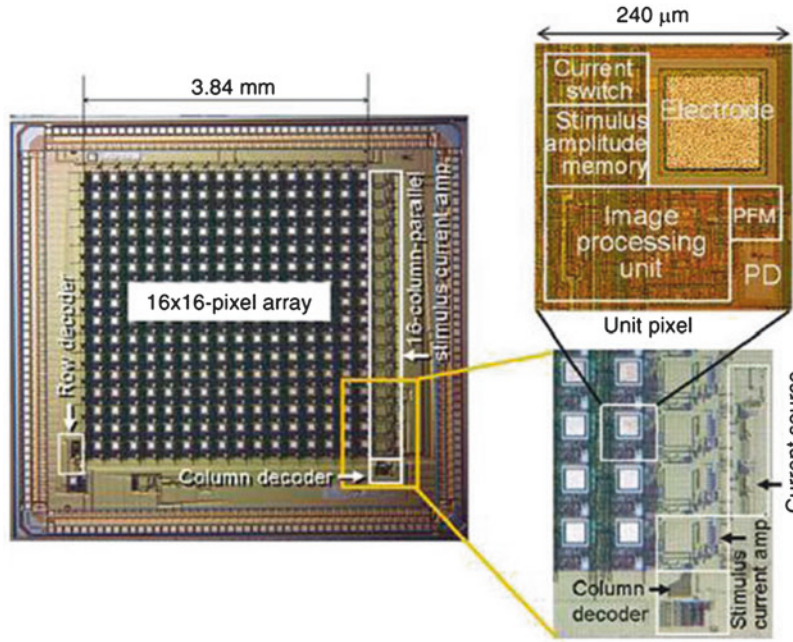
Microprocessor: A microprocessor is placed on the patient's belt in order to capture the visual information (pixels) and transforms it into electronic signals. This part of the device is located externally allowing adjustments of the visual information to send the most accurate data to the internal unit. In this manner, the ophthalmologist could evaluate the patient's vision and modify it accordingly.

Battery: A battery is placed externally to provide energy to the external microprocessor, but also to send energy wirelessly to the internal unit. Having an external battery is an important requirement since a battery located inside would have more constraints such as being of small size, and its replacement could demand frequent surgeries plus the fact that a battery inside the eye would require stringent bio-isolation from the eye to prevent eye rejection.

Internal Components

Transmitting Unit: Using a special algorithm, the internal unit codifies the electronic signals sent by the microprocessor and send electrical pulses through a stimulator to the microelectrode array located at the retina surface via wire connection. The power to supply this unit is provided wireless by the external component.

Microelectrode array: Microelectrodes are the interface between the device and the visual system that are responsible for stimulating the neurons that will transport the visual stimulus to the brain. Microelectrodes are micrometer size structures ranging from flat circular platinum discs to elongated sharp pyramidal structures, in the size ranging between 250 and 500 μ in diameter, responsible for transmitting the electrical pulses



Artificial Retina: Focus on Clinical and Fabrication Considerations, Fig. 1 Photograph of the internal unit with the microelectrode array and the stimulator circuit (With permission from Springer, from Ref. [4])

coming from the microstimulators to the neurons, which become activated and transport the electrical pulses to the brain. To date, microelectrode arrays have been developed, featuring 16, 64, and 200 microelectrodes. The 64 microelectrode array is currently the first artificial retina implanted commercially in blind people, first in Europe, where Second Sight has obtained approval for commercial implantation of the epiretinal device. Several kinds of microelectrodes have been developed for retina prosthesis. Desired features for retina microelectrodes are low impedance which means easy transfer of electrical pulses to the neurons, biocompatibility, or an acceptable tissue response to the implant, long lifetime to avoid additional surgeries, and flexibility (that means, adaptation of the implant to the curvature of the eyeball without hindering the electrode-neuron distance that could promote heating effects if they are too close or could not activate the neurons if they are too separate). Miniaturization of electrodes is highly desirable and currently under development as the larger number of electrodes allows to achieving higher resolution of the image. The problem, which a tough one, is that impedance increases as electrode size becomes smaller. As impedance increase, voltage should increase to deliver the same amount of charge to the neurons, resulting in an increase in energy consumption, and concurrently an increase in heat dissipation, which could damage the surrounding tissue.

Wireless Transfer of Information and Energy between External and Internal Units: In the current epiretinal devices, both information and energy communication between the external and internal units transmitted via electromagnetic signals wirelessly and coupled to the microchip via inductively coupling coils. Inductively coupling coils are conductive wires built up as coils that work based on the physical phenomena of electromagnetic induction by which an electric current flowing in one coil induces a voltage on the other coil by the electromagnetic field created in the former by the current flow. One of the coils is located in the external component while the other is located in the internal component (Fig. 1).

Materials Used in the DOE Artificial Retina

Silicon: Silicon is the current material used in most of the microelectronic components and

microfabrication processes. In the artificial retina, microprocessors are necessary for translating the visual stimulus into an electric signal. Also, the internal unit has a microchip stimulator to deliver the electric signal to the microelectrodes.

Diamond: In the DOE artificial retina prototype, a novel diamond material known as Ultrananocrystalline Diamond (UNCD) is being investigated in thin film form to provide a hermetic/bioinert/biocompatible coating to encapsulate the silicon microchip to enable implantation inside the eye. The UNCD film provides a coating extremely resistant to the corrosive environment of the eye saline, which otherwise would attack the silicon material underneath dissolving it [5].

Polymers: Some polymers including polyimide, parylene, and silicone have been tested to provide a flexible platform to deposit microelectrodes. In this manner, it is possible to adjust the microelectrode array to the curvature of the eye reducing the distance between the device and the neurons which results in lower voltage thresholds to activate the neurons decreasing power consumption.

Metals

Gold: gold is being used as one of the noble metals for wires to connect the microelectrode array to the pads where the current is delivered from the internal electronic unit that drives the energy provided wireless by the external battery.

Platinum: Platinum is being used as the material for microelectrodes. Key properties that make platinum an optimum candidate as microelectrode material includes its capacity to transfer electrical charge to the neurons (low impedance), its biocompatibility, and its durability. A potential problem of Pt, which has not been investigated in detail, is the catalytic properties of Pt. As the size of the Pt electrodes becomes smaller, the catalytic properties of Pt becomes bigger. Under the liquid environment of the eye, this could increase the risk of unwanted electrochemical reactions.

Indium Tin Oxide: material used for leads

Silicon Nitride: material used as insulator layer

Silicon Oxide: material used as insulator layer

Fabrication Process of the DOE Artificial Retina

Due to the tiny and delicate space that occupies the retina inside the eyeball, miniaturization technologies are deemed extremely necessary to create complex devices with integrated functionality. Some of the miniaturization techniques currently used to develop an artificial retina are:

Photolithography: Photolithography is a technique developed to imprint micro and nano patterns using a photosensitive polymer that when exposed to UV light it could reproduce a pattern imprinted in a mask over a substrate like silicon.

Reactive Ion Etching (RIE): Once a photolithography pattern is created, in order to etch away the parts that do not correspond to the pattern, an etching technique is needed. RIE is a technique that uses high-energy ions, radicals, and electrons created in plasma and accelerated toward the substrate to enhance chemical reactions of the reactive atoms in the system to produce directional (anisotropic) etching of the substrate.

Electroplating: Electroplating is a technique that allows growing metals to create microstructures such as the platinum discs microelectrode array.

Thin Film Deposition (Sputtering, Physical Evaporation, Chemical Vapor Deposition): Sputtering, physical evaporation, and chemical vapor deposition techniques are used to produce films with thickness from a few nanometers ($1\ \text{nm} = 10^{-7}$ cm) to a few microns ($1\ \mu\text{m} = 10^{-4}$ cm). For the components of the artificial retina, different type of films are used for different purposes. In the hermetic encapsulation of the microchip to provide bioinertness (UNCD films), in the metal contacts to allow current passage to the microelectrodes (platinum films), and insulation layers to minimize or eliminate leakage currents that could promote unwanted electrochemical reactions within the biological media of the eye.

Most of these techniques are carried out in special facilities known as clean room designed to keep the air free from dust particles, preserving the surface of the retina components free from contaminants (dust and other species).

Laboratory Studies for Assessing Retina Structure and Function

In Vitro and In Vivo Testing

In vitro and vivo testing of biomedical devices becomes of critical importance to accomplish the goal of developing a safe and effective commercial device. In vitro systems could provide valuable information regarding electrical requirements that are needed to activate ganglion cells and thus the visual pathway. Electrophysiological techniques are used to study and assess the response of the artificial retina components to electrical stimulus and are some of the laboratory techniques to perform analysis of the artificial retina performance. Electrophysiological studies are based on recording electric potentials of the retina neurons. Electric potentials (EP) are electric waves that travel through the neurons carrying information. In this case, visual information is codified by frequency and amplitude changes of the potentials. In retina diseases such as RP and AMD, it is found that electrical potentials are altered and therefore become an important tool to perform a diagnosis of retina disease. In the same manner, the EPs allow an electrical characterization of artificial retina performance by analyzing the effect of microelectrodes stimulation on the activation of retina neurons. The technique developed to record EPs on retina cells is called *Electroretinogram. In order to perform an Electroretinogram* it is necessary to mount electrodes in the ocular surface to be able to record electric potentials in the neurons of the retina. A more invasive technique, called intraocular electroretinogram, consists of placing an electrode inside the eye, near the retina surface. In spite of providing useful information during the stage of animal testing, the clinical utility of Electroretinogram remains controversial.

In addition to the microelectrode system, the experimental setup to perform these types of experiments includes isolating the retina from the eyeball of an animal (i.e., salamander). The retina tissue is stimulated by a microelectrode array displayed in such a way to contact the retina neurons. In this manner, it is possible to record the electrical activity of a group of neurons (i.e., bipolar cells, ganglion cells) after an electric pulse is applied through a microelectrode array. Before performing studies in human subjects, it is necessary to evaluate the safety and effectiveness of the artificial retina in an *animal model*. An animal model mimics the human disease in an animal. Animal models of RP and AMD have been developed to assess the safety and effectiveness of an artificial retina [6].

Clinical and Surgical Considerations

Device Implantation

The eye is divided in two main compartments, the anterior compartment composed of the cornea, iris, and lens, and the posterior segment composed of the sclera, choroid, retina vitreous, and optic nerve. In order to implant the artificial retina, it becomes mandatory to access the posterior compartment of the eye, which is filled with a gelatinous substance known as vitreous. Being able to place the device on the retina surface (epiretinal) requires to remove the vitreous out of the posterior chamber while simultaneously keeping the pressure of the eye constant to avoid the collapse of the eyeball. The procedure is called vitreoretinal surgery and is done by performing three incisions on the sclera, one for introducing an illuminating system, another one for introducing a cannula to insert the device and cut the vitreous (it is not possible to suction it because it could track the retina and produce a retinal detachment, therefore it should be cut first and then aspirated), and the last incision for infusing liquid to keep the intraocular pressure constant while the small pieces of vitreous are aspirated. The device is fixed to the surface of the inner part of the retina using a tack that keeps the device in place avoiding any displacement [7].

Clinical Assessment of a Patient with an Artificial Retina

Vision Before Surgery

Patients with RP initially suffer from night blindness. As the disease progress, patients start to relate symptoms associated with peripheral loss

of rods such as inability to see pedestrians while they are driving. In addition, a difficulty in seeing the light impedes the patient from reading. At more advances stages, the patient loses independency because they are only able to see a small center region of the visual field (tunnel vision) and they are not able to walk by themselves. The patient with MD on the other hand consults for central vision impairment such as vision blurring, or a scotoma, which is a patch in the center of the visual field. At later stages, in both diseases, blindness could be the final outcome. Therefore, an artificial retina could restore partial vision for these patients enhancing their quality of life significantly. Computer simulations indicate that in order for a blind person to recover basic visual capabilities such as face recognition, reading, and independent mobility, the artificial retina should have at least 1,000 electrodes in order to create an image of 1,000 pixels. Current technologies have not overcome this challenge yet because miniaturization of electrodes is difficult. However, improvement of everyday life activities performance has been observed in patients that were implanted with the device with 60 electrodes [1].

Clinical Studies to Assess Retina Structure and Function

OCT (Optical Coherence Tomography): OCT is a clinical technique that allows the measurement of the thickness of the retina. When an artificial retina is implanted, OCT brings about useful information about distance between microelectrodes and the retina surface which is an important parameter, since in order to obtain a neural response the microelectrode array should be placed in close proximity to the group of neurons. In addition, if the microelectrode array is placed too far apart from the retina neurons the system consumes more power and dissipates more heat, increasing the chances of tissue damage.

Fundus Photo: Fundus refers to the inside back part of the eye where the retina is located. A special tool called ophthalmoscope allows the physician to take pictures of this part of the eye obtaining valuable information of the status of the retina tissue affected by RP or AMD.

Fluorescein Angiogram

This technique allows visualization of retina vessels by injecting a fluorophore inside the artery vessels. It is a useful technique for diagnosis of RP.

Clinical Performance After Artificial Retina Implantation

After implantation with an artificial retina containing 16 microelectrodes, patients were capable of discerning light and dark patterns (called phospenes) and localizing the spatial origin of the phospenes (which microelectrode was stimulated), whereas with the new generation of 64 microelectrodes the patients were capable of performing spatial motor task where visual input results are of critical importance to perform motor activities. In order to assess visual performance in these patients, several specific test including psychophysical studies, specific questionnaires, and subjective perceptions are being used, and new strategies to objectively assess visual function in patients implanted with an artificial retina are being pursued.

Cross-References

- ▶ CMOS MEMS Fabrication Technologies
- ▶ Epiretinal Prosthesis
- ▶ MEMS Neural Probes
- ▶ MEMS on Flexible Substrates

References

1. Artificial Retina Project, DOE website. http://artificialretina.energy.gov/gpra2010.shtml. Accessed 1 Nov 2011
2. Auner, G.W., You, R., Siy, P., McAllister, J.P., Takluder, M., Abrams, G.W.: Development of a high frequency microarray implant for retina stimulation. In: Artificial Sight: Basic Research, Biomedical Engineering, and Clinical Advances. Springer, New York (2007)
3. Liu, W., Mohanasankar, S., Wang, G., Mingcui, Z., Weiland, J.D., Humayun, M.S.: Challenges in realizing a chronic high-resolution retinal prostheses. In: Artificial Sight: Basic Research, Biomedical Engineering, and Clinical Advances. Springer, New York (2007)
4. Ohta, J., Tokuda, T., Kagawa, K., Terasawa, Y., Ozawa, M., Fujikado, T., Yasuo, T.: Large scale integration base stimulus electrodes for retinal prosthesis. In: Artificial

Sight: Basic Research, Biomedical Engineering, and Clinical Advances. Springer, New York (2007)
5. Auciello, O., Sumant, A.V.: Status review of the science and technology of ultrananocristalline diamond (UNCD™) films and application to multifunctional devices. Diamond Relat. Mater. **19**, 699–718 (2010)
6. Fletcher, E.L., Jobling, A.I., Vessey, K.A., Luu, C., Guymer, R.H., Baird, P.N.: Animal models of retinal disease. Prog. Mol. Biol. Transl. Sci. **100**, 211–286 (2011)
7. Hameri, H., Weiland, J.D., Humayun, M.S.: Biological considerations for an intraocular retinal prosthesis. In: Artificial Sight: Basic Research, Biomedical Engineering, and Clinical Advances. Springer, New York (2007)

Artificial Synapse

► Dynamic Clamp

Artificial Vision

► Artificial Retina: Focus on Clinical and Fabrication Considerations

Arts-Based Research

► Biornametics: Architecture Defined by Natural Patterns

Astigmatic Microparticle Imaging

Christian J. Kähler, Christian Cierpka and Massimiliano Rossi
Institute of Fluid Mechanics and Aerodynamics, Universität der Bundeswehr München, Neubiberg, Germany

Synonyms

Anamorphic particle image tracking; Wavefront deformation particle image tracking

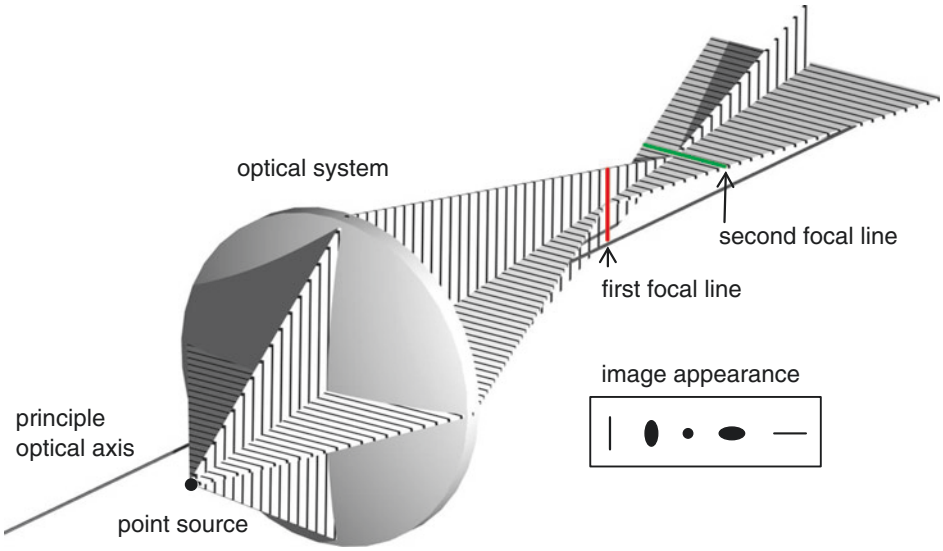
Definition

Astigmatism particle imaging is an optical method to determine the instantaneous volumetric position of spherical particles in a volume of transparent medium using only a single camera view. The method is mainly applied in microfluidics to measure the velocity and trajectories of tracer particles that follow the fluid motion. However, recently it could be shown that this measurement principle without bias errors due to the depth of correlation and spatial averaging is well suited to reconstruct the arbitrary-shaped interface of mixing layers in a micromixer with extended high precision.

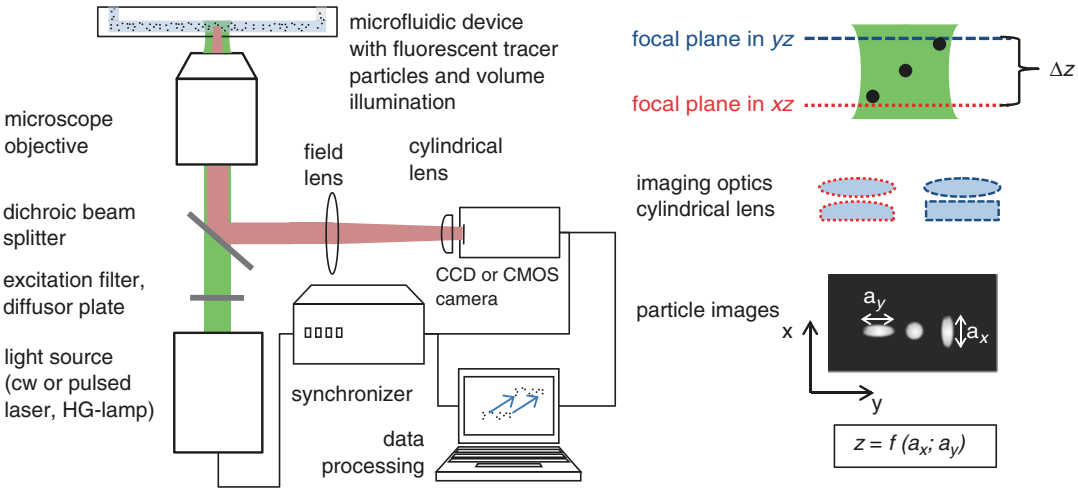
Optical Principle

Astigmatic imaging is a way to break the axis symmetry of an optical system, thus allowing for the depth coding of particle positions in 2D images. Among the manifold of 3D3C methods recently developed, it can be classified as a method that makes use of the particle defocus without using an extra aperture [13]. The word astigmatism is based on the Greek language description for not (a-) point-shaped (-stigma). An optical system that features astigmatism has two focal planes, and the astigmatic image of a point source is thus an ellipse as shown in Fig. 1. Astigmatism is commonly used for technical applications, e.g., autofocus systems in CD, DVD, or Blu-ray players, to measure the distance between the rotating disk and the reader head.

One approach producing an astigmatic image aberration is the use of a tilt angle between the measurement plane or volume and the camera, which results in off-axis imaging. This method is often not applicable due to the fixed optical axis in microscopes and requires complex calibration procedures. However, another more elegant method is the use of cylindrical lenses. Already in 1994, Kao and Verkman [7] applied this technique to measure the position of fluorescent particles in living cells. The basic principle and the implementation to a standard epi-fluorescence microscope are shown in Fig. 2. The only difference from a standard microparticle image



Astigmatic Microparticle Imaging, Fig. 1 Schematic of astigmatism caused by off-axis imaging and the appearance of the image at various distances between the two focal planes, according to [5]



Astigmatic Microparticle Imaging, Fig. 2 Schematic of the astigmatism technique integrated in an inverted epi-fluorescence microscope (*left*) and principle schematic of the imaging with the cylindrical lens, according to [13]. The width a_x and height a_y of the particle images can be related to the axial (depth) position z (*right*)

velocimetry (μPIV) system [12] is the cylindrical lens, which is installed in front of the camera chip as shown in Fig. 2. As for standard μPIV, the transparent flow of interest is usually seeded with monodisperse spherical fluorescent tracer particles that follow the flow faithfully. The whole microfluidic channel is then illuminated twice by a laser or another bright light source (continuous or pulsed). To homogenize the beam profile often, a diffuser plate is used. A dichroic beam splitter is applied to allow only short wavelength light to pass through (indicated in green in Fig. 2), if a white light source is used. This short wavelength light is then absorbed by the tracer particles' fluorescent dye, and longer wavelength light (indicated in red) is emitted. The dichroic

filter is now used to reject direct reflections of the microfluidic device and only allow the emitted red light to contribute to the signal. A combination of a field lens and the cylindrical lens is used to image the particles on a digital CCD (charge-coupled device), CMOS or sCMOS ((scientific) complementary metal-oxide semiconductor), or intensified camera. On the right side of Fig. 2, the imaging of the particles is schematically shown. The curvature of the cylindrical lens acts in one direction only. Therefore, two distinct focal planes are produced. In the left (red dotted) schematic, the cylindrical lens causes a shortening of the distance between the focal plane and the objective lens. Particles that are close to this focal plane show a sharp and small extension in the x-direction (a_x) and a larger stretched one in the y-direction (a_y), thus producing an oblate-shaped particle image. On the right side of the schematic (blue dashed), the cylindrical lens is rotated by 90°. The focal plane is now further away from the objective lens. a_y is thus smaller than a_x, resulting in a prolate-shaped particle image. The particle image width and height can now be related to the depth position z by a proper calibration.

Data Evaluation and Calibration

Since the volumetric particle position is used later to estimate the velocity or fluid interfaces, the accuracy of the method relies on the precise determination of the particle's position. For the in-plane, or x- and y-direction, an in-plane calibration by means of a grid, or a known object, can be performed. The in-plane position can later be determined with sub-pixel accuracy by either a two-dimensional Gauss fit or via a correlation with a Gaussian, where the latter method is much faster. The width and height of the particle images are identified by either a two-dimensional Gauss fit or the autocorrelation function of the particle image. These algorithms have shown accurate results, even for high noise levels for particle images that extend at least three pixels in both directions [2]. Since the technique relies on the exact determination of the particle images' shapes, at least an 8-bit or better sampling is necessary to yield reliable results. Once a_x and a_y are known, a relationship with the depth distance is necessary. If the particles' diameter distribution is very coarse, the difference $a_x - a_y$ [1] or the ratio a_x/a_y [2] can be used as an independent criteria. The trust region is then limited to the linear region between the two focal planes. For a narrow particle diameter distribution, a direct fit of a_x and a_y can be used to substantially extend the measurement depth. A model of the particle image diameter, as a function of the distance z from the in-focus plane, can be applied for this purpose [9]. Under the assumption that the geometric image of the particles and the effect of diffraction and defocusing can be approximated by a Gaussian function, and that the working distance of the lens is significantly larger than z, the particle image diameter can be estimated by the following equation [14]:

$$a(z) = M\sqrt{d_p^2 + 1.49\lambda^2\left(\frac{n_0^2}{\mathrm{NA}^2} - 1\right) + 4z^2\left(\frac{n_0^2}{\mathrm{NA}^2} - 1\right)^{-1}} \tag{1}$$

where d_p denotes the particle diameter, λ is the wavelength of emitted light, n_0 the refractive index of the immersion medium of the lens, and M and NA are the magnification and numerical aperture of the lens, respectively. Equation 1 represents an arc of a hyperbola. The function can be used to extend the measurement depth beyond the two focal planes, and the fit is applied to the measured data itself:

$$\begin{aligned} a_x(z) &= \sqrt{c_1^2(z - F_{xz})^2 + c_2^2} + c_3 \quad a_y(z) \\ &= \sqrt{c_4^2(z - F_{yz})^2 + c_5^2} + c_6 \end{aligned} \tag{2}$$

F_{xz} and F_{yz} denote the positions of the respective in-focus planes for the x- and y-directions, and $c_{1\ldots6}$ are parameters for the fit function. Particles that are beyond the in-focus planes can be taken into account by this fit, and thus the measurement volume is not restricted to the linear region between the in-focus planes. The distance between F_{xz} and F_{yz} in physical space can be

determined automatically by scanning a grid reticule through the volume. The axial position of F_{xz} and F_{yz} can be measured by a fit of a focus function (e.g., rms of the intensity) with submicron resolution. If additional image aberrations are present and the focal planes are not completely flat, these aberrations will also be estimated by the procedure, and the final results can be corrected for this systematic error. The validity of the model was tested on simulated and real data [3]. The advantages of the intrinsic calibration are the following:

- The measurement volume depth is not limited by the distance between the two in-focus planes.
- The optical path through the different media does not change from calibration to experiment.
- All data points are taken into account, and thus the calibration is statistically relevant to a high degree.

Nevertheless, the distance between the two in-focus planes Δz is the quantity that primarily defines the depth of the measurement volume. It can be adjusted by the focal length of the cylindrical lens or the distance between the imaging lens of the CCD and the cylindrical lens [1]. The specific measurement volume is finally dependent on the specific optical arrangement of the whole microscope and the magnification [1]. Increasing the magnification gives lower Δz for the same cylindrical lens, whereas decreasing the focal length of the cylindrical lens increases the measurement volume by approximately $1/f_{\mathrm{cyl}}$ [3].

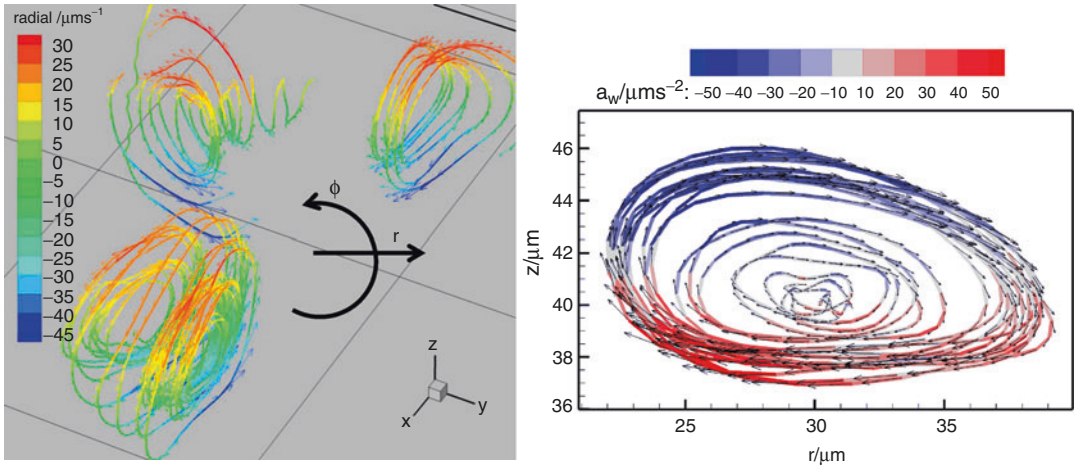
Applications

In general, the astigmatic imaging approach based on cylindrical lenses is very easy to apply and allows for the low cost extension of existing 2D measurement systems to fully 3D measurements without changing the illumination light path. Knowing the particle positions in the volume at two different time instants t and $t + \Delta t$ allows for an estimation of the first-order approximation of the particles' velocity. To find matching particles, several elaborate tracking schemes can be used [10, 16]. Despite the necessity for lower seeding concentration to avoid particle image overlapping, the major benefits of a 3D tracking scheme opposed to correlation-based methods are the following:

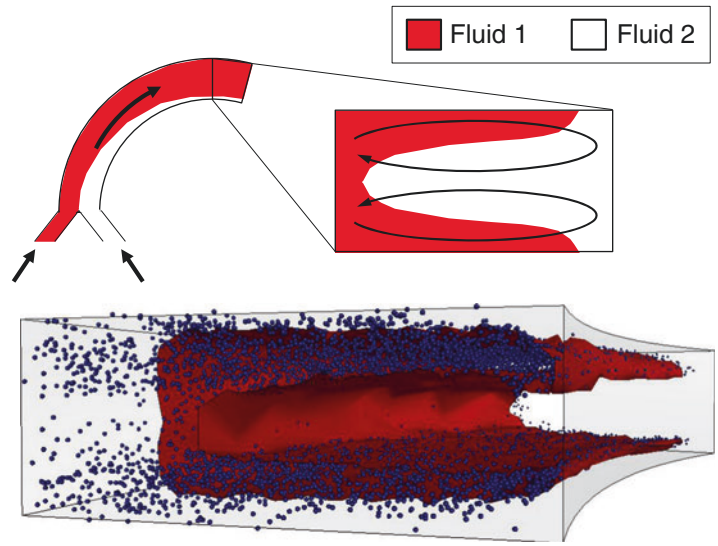
- Out-of-focus particles do not bias the velocity estimate by the depth of correlation effect [9].
- In-plane gradients and out-of-plane gradients are not averaged over interrogation volumes [6].
- The spatial resolution can be increased by taking a larger number of images.

The tracking approach was applied in recent investigations to measure flow velocities in micron-sized channels [1, 2] and close to the electrode during electrochemical metal deposition [15]. A comparison with standard and stereoscopic μPIV showed a similar uncertainty for the in-plane velocity components but an almost two times lower uncertainty for the out-of-plane component using A-μPTV [4]. Using time-resolved image sampling, the temporal and spatial evolution of the particle trajectories in an electrothermally generated microvortex was measured in a microvolume of 50 µm in depth without traversing [8]. Selected trajectories are shown in a volumetric representation on the left side of Fig. 3. The three-dimensional structure of the vortex and the mechanisms of particle trapping inside the vortex could be studied, and the results were used to validate the numerical boundary conditions for a simulation of such a scenario. Since the particle trajectories were sampled with sufficient temporal resolution, Lagrangian accelerations were also measured as shown on the right side of Fig. 3.

Astigmatic particle imaging can also be used to estimate the boundary of two mixing fluid regions of arbitrary shape or to reconstruct the three-dimensional shape of particle-seeded stream tubes. The volumetric topology of the interface can be estimated from the measured 3D positions of the particles. Algorithms based on Delaunay triangulation and numerical diffusion have been examined to perform the surface reconstruction [11].



Astigmatic Microparticle Imaging, Fig. 3 Selected trajectories in an electrothermally driven microvortex, color coded with the radial velocity (*left*). Representation of one trajectory in the *rz*-plane, color coded with the axial acceleration (*right*)



Astigmatic Microparticle Imaging, Fig. 4 Reconstruction of the mutual interface between two mixing fluids in a microfluidic mixer using the volumetric distribution of tracer particles. The flow is laminar and the convective mixing is induced by secondary flows established as a consequence of the channel curvature

The accuracy and resolution of the reconstruction are limited by the maximum particle density achievable and the error of the particle position determination. However, for steady flows (as typical in microfluidics), any desired accuracy can be achieved for a sufficient large number of samples. A classic example in which this approach is useful is the characterization of passive microfluidic mixers, where the topology of the mutual interface between the two mixing fluid streams plays an important role. In this case, one of the two mixing fluids is seeded with a homogeneous distribution of tracer particles. A typical result for the case of a micromixer, based on a curved microchannel, is shown in Fig. 4.

The cross section of the microchannel was $200 \times 70\ \mu m^2$, and the average particle concentration was set to 0.023 particles/μm^3, resulting in a resolution of 2.5 µm in the optical-axis direction with a maximum estimated error of 5.5 µm [11].

Conclusion and Future Aspects

It is obvious that astigmatic particle imaging is a technique that provides the instantaneous volumetric position of particles in a transparent medium with high precision. The displacement between two successive recordings can be used to estimate the volumetric three-component velocity field. In addition, the Lagrangian velocity and acceleration of the particle can be measured for sufficiently high sampling rates without systematic errors due to the depth of correlation and spatial averaging as typical for μPIV. The positions of the particles can also be used to estimate their volumetric distribution and to analyze the mixing of fluid streams [11], particle agglomeration, trapping, deposition, or other process engineering-relevant phenomena. Since not only the position but also the intensity of the particle images can be measured, other possible extensions of the technique are the measurement of temperature, pH value, or even pressure by particles that show a sensitivity in the fluorescence signal to these quantities.

Cross-References

► Micro/Nano Flow Characterization Techniques

References

1. Chen, S., Angarita-Jaimes, N., Angarita-Jaimes, D., Pelc, B., Greenaway, A.H., Towers, C.E., Lin, D., Towers, P.D.: Wavefront sensing for three-component three-dimensional flow velocimetry in microfluidics. Exp. Fluids **47**, 849–863 (2009). doi:10.1007/s00348-009-0737-z
2. Cierpka, C., Segura, R., Hain, R., Kähler, C.J.: A simple single camera 3C3D velocity measurement technique without errors due to depth of correlation and spatial averaging for micro fluidics. Meas. Sci. Technol. **21**, 045401 (2010). doi:10.1088/0957-0233/21/4/045401
3. Cierpka, C., Rossi, M., Segura, R., Kähler, C.J.: On the calibration of astigmatism particle tracking velocimetry for microflows. Meas. Sci. Technol. **22**, 015401 (2011). doi:10.1088/0957-0233/22/1/015401
4. Cierpka, C., Rossi, M., Segura, R., Mastrangelo, F., Kähler, C.J.: A comparative analysis of the uncertainty of astigmatism-μPTV, stereo-μPIV, and μPIV. Exp. Fluids (2011). doi:10.1007/s00348-011-1075-5. open access
5. Hain, R., Kähler, C.J., Radespiel, R.: Principles of a Volumetric Velocity Measurement Technique Based on Optical Aberration. Notes on Numerical Fluid Mechanics and Multidisciplinary Design, vol. 106, pp. 1–10. Springer, Berlin (2009). doi:10.1007/978-3-642-01106-1-1
6. Kähler, C.J. Scharnowski, S., Cierpka, C.: On the resolution limit of digital particle image velocimetry Experiments in Fluids, 52, 1629–1639 (2012), 10.1007/s00348-012-1280-x, open access
7. Kao, H.P., Verkman, A.S.: Tracking of single fluorescent particles in three dimensions: use of cylindrical optics to encode particle position. Biophys. J. **67**, 1291–1300 (1994). doi:10.1016/S0006-3495(94)80601-0
8. Kumar, A., Cierpka, C., Williams, S.J., Kähler, C.J., Wereley, S.T.: 3D3C velocimetry measurements of an electrothermal microvortex using wavefront deformation PTV and a single camera. Micro. Nano. **10**, 355–365 (2012). doi:10.1007/s10404-010-0674-4
9. Olsen, M.G., Adrian, R.J.: Out-of-focus effects on particle image visibility and correlation in microscopic particle image velocimetry. Exp. Fluids **29**, S166–S174 (2000). doi:10.1007/s003480070018
10. Oulette, N.T., Xu, H., Bodenschatz, E.: A quantitative study of three-dimensional Lagrangian particle tracking algorithms. Exp. Fluids **40**, 301–313 (2006). doi:10.1007/s00348-005-0068-7
11. Rossi, M., Cierpka, C., Segura, R., Kähler, C.J.: Volumetric reconstruction of the 3D boundary of stream tubes with general topology using tracer particles. Meas. Sci. Technol. **22**, 105405 (2011). doi:10.1088/09570233/22/10/105405. open access
12. Santiago, J.G., Wereley, S.T., Meinhart, C.D., Beebe, D.J., Adrian, R.J.: A particle image velocimetry system for microfluidics. Exp. Fluids **25**, 316–319 (1998). doi:10.1007/s003480050235
13. Cierpka, C., Kähler, C.J.: Particle imaging techniques for volumetric three-component (3D3C) velocity measurements in microfluidics. J. Vis. (2011). doi:10.1007/s12650-011-0107-9. open access
14. Rossi, M., Segura, R., Cierpka, C., Kähler, C.J.: On the effect of particle image intensity and image preprocessing on depth of correlation in micro-PIV. Exp. Fluids (2011). doi:10.1007/s00348-011-1194-z. open access
15. Tschulik, K., Cierpka, C., Uhlemann, M., Gebert, A., Kähler, C.J., Schultz, L.: In-situ analysis of three-dimensional electrolyte convection evolving during the electrodeposition of copper in magnetic gradient fields. Anal. Chem. **83**, 3275–3281 (2011). doi:10.1021/ac102763m
16. Cierpka, C. Lütke, B. Kähler, C.J.: Higher order multi-frame Particle Tracking Velocimetry, Experiment in Fluids, 54, 1533, doi 10.1007/s00348-013-1533-3

Atmospheric Pressure Chemical Vapor Deposition (APCVD)

▶ Chemical Vapor Deposition (CVD)

Atomic Cluster

▶ Synthesis of Subnanometric Metal Nanoparticles

Atomic Force Microscopy

Bharat Bhushan and Manuel L. B. Palacio
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Scanning probe microscopy

Definition

Atomic force microscopy (AFM) is a scanning probe technique used to measure very small forces (as low as <1 nN) and produce very high resolution, three-dimensional images of sample surfaces. The AFM is capable of investigating surfaces of both conductors and insulators on the atomic scale. Aside from topographic imaging, the AFM is used in a wide variety of applications, including, but not limited to, the measurement of mechanical properties, chemical forces, electrical properties, among others.

Overview

Gerd Binnig and his colleagues developed the AFM in 1985. It is capable of investigating surfaces of scientific and engineering interest on an atomic scale [1]. As shown in Table 1, it belongs to a general class of related instrumental techniques that are used in determining surface nanotopography, nanomechanical, and nanotribological properties, such as the surface force apparatus (SFA), the scanning tunneling microscopes (STM), and the friction force microscope (FFM). The SFA was developed in 1968 and is commonly employed to study both static and dynamic properties of molecularly thin films sandwiched between two molecularly smooth surfaces. The STM, developed in 1981, allows imaging of electrically conducting surfaces with atomic resolution, and has been used for imaging of clean surfaces. AFM/FFM techniques have the advantage of imaging all types of sample surfaces, regardless of its roughness or electrical properties, unlike the SFA or STM.

The AFM relies on a scanning technique to produce very high resolution, three-dimensional images of sample surfaces. It measures ultrasmall forces (less than 1 nN) present between the AFM tip surface mounted on a flexible cantilever beam and a sample surface. These small forces are obtained by measuring the motion of a very flexible cantilever beam having an ultrasmall mass, by a variety of measurement techniques including optical deflection, optical interference, capacitance, and tunneling current. The deflection can be measured to within 0.02 nm, so for a typical cantilever spring constant of 10 N/m, a force as low as 0.2 nN can be detected. To put these numbers in perspective, individual atoms and human hair are typically a fraction of a nanometer and about 75 μm in diameter, respectively, and a drop of water and an eyelash have a weight of about 10 and 100 nN, respectively.

In the operation of high-resolution AFM, the sample is generally scanned rather than the tip because any cantilever movement would add vibrations. AFMs are available for measurement of large samples, where the tip is scanned and the sample is stationary. To obtain atomic resolution with the AFM, the spring constant of the cantilever should be weaker than the equivalent spring between atoms. A cantilever beam with a spring constant of about 1 N/m or lower is desirable. For high lateral resolution, tips should be as

Atomic Force Microscopy, Table 1 Comparison of typical operating parameters in SFA, STM, and AFM/FFM used for micro/nanotribological studies

Operating parameter	SFA	STM[a]	AFM/FFM
Radius of mating surface/tip	~10 mm[b]	5–100 nm	5–100 nm
Radius of contact area	10–40 μm	N/a	0.05–0.5 nm
Normal load	10–100 mN	N/a	<0.1–500 nN
Sliding velocity	0.001–100 μm/s	0.02–200 μm/s (scan size ~1 nm × 1 nm–125 μm × 125 μm; scan rate <1–122 Hz)	0.02–200 μm/s (scan size ~1 nm × 1 nm–125 μm × 125 μm; scan rate <1–122 Hz)
Sample limitations	Typically atomically–smooth, optically transparent mica; opaque ceramic, smooth surfaces can also be used	Electrically conducting samples	None of the above

[a]Can be used for atomic-scale imaging
[b]Since stresses scale with the inverse of tip radius, SFA can provide very low stress measurement capabilities

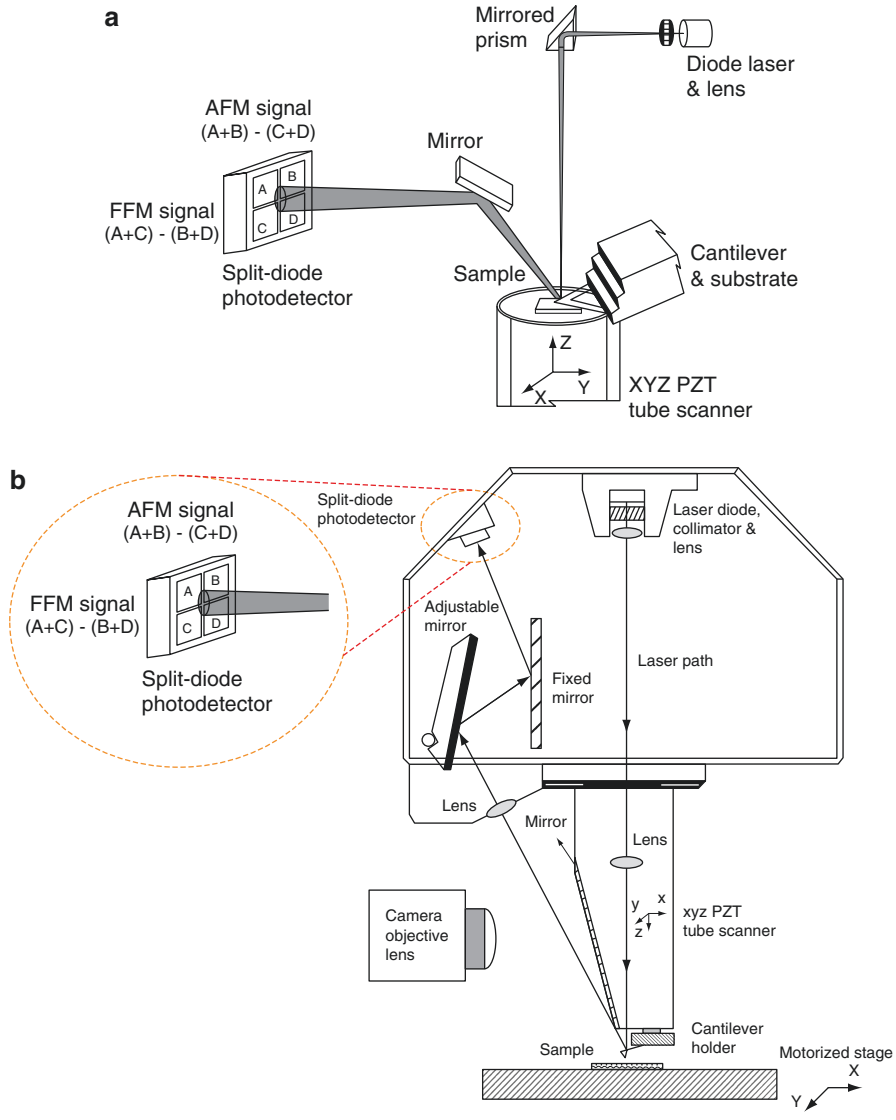
sharp as possible. Tips with a radius ranging from 5 to 50 nm are commonly available. Aside from topographic imaging, the AFM is also widely used for the characterization of surface/near-surface properties such as friction, adhesion, scratch, wear, fabrication/machining, electrical properties (such as surface potential, resistance, capacitance, tunneling current, electric field gradient distribution), local deformation, nanoindentation, local surface elasticity, viscoelasticity, and boundary lubrication.

Various measurement techniques using atomic force microscopy are described next.

Description of Various Measurement Techniques

Surface Roughness and Friction Force Measurements

The AFM is widely used for surface height imaging and roughness characterization down to the nanoscale. Commercial AFM/FFM is routinely used for simultaneous measurements of surface roughness and friction force [2–4]. These instruments are available for measurement of small samples and large samples. In a small sample AFM shown in Fig. 1a, the sample, generally no larger than 10 mm × 10 mm, is mounted on a piezoelectric crystal in the form of a cylindrical tube (referred to as a PZT tube scanner) which consists of separate electrodes to scan the sample precisely in the X–Y plane in a raster pattern and to move the sample in the vertical (Z) direction. A sharp tip at the free end of a flexible cantilever is brought in contact with the sample. Normal and frictional forces being applied at the tip-sample interface are measured using a laser beam deflection technique. A laser beam from a diode laser is directed by a prism onto the back of a cantilever near its free end, tilted downward at about 10° with respect to the horizontal plane. The reflected beam from the vertex of the cantilever is directed through a mirror onto a quad photodetector (split photodetector with four quadrants). The differential signal from the top and bottom photodiodes provides the AFM signal which is a sensitive measure of the cantilever vertical deflection. Topographic features of the sample cause the tip to deflect in the vertical direction as the sample is scanned under the tip. This tip deflection will change the direction of the reflected laser beam, changing the intensity difference between the top and bottom sets of photodetectors (AFM signal).



Atomic Force Microscopy, Fig. 1 Schematics (**a**) of a commercial small sample atomic force microscope/friction force microscope (*AFM/FFM*) and (**b**) of a large sample AFM/FFM

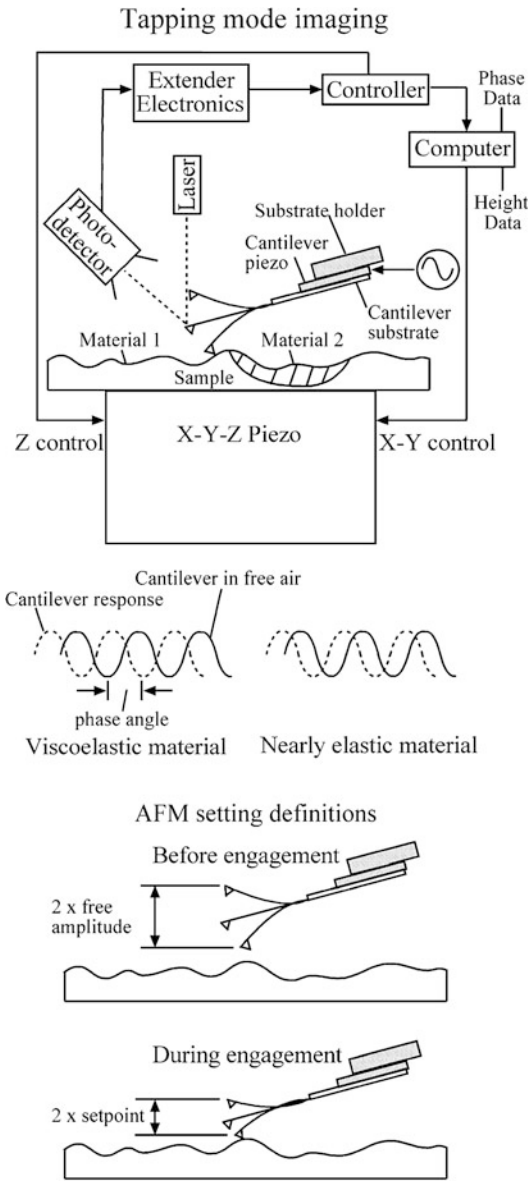
In the AFM operating mode called the contact mode, for topographic imaging or for any other operation in which the applied normal force is to be kept constant, a feedback circuit is used to modulate the voltage applied to the PZT scanner to adjust the height of the PZT, so that the cantilever vertical deflection (given by the intensity difference between the top and bottom detector) will remain constant during scanning. The PZT height variation is thus a direct measure of the surface roughness of the sample.

In a large sample AFM, both force sensors using optical deflection method and scanning unit are mounted on the microscope head, Fig. 1b. Because of vibrations added by cantilever movement, lateral resolution of this design can be somewhat poorer than the design in Fig. 1a in which the sample is scanned instead

of cantilever beam. The advantage of the large sample AFM is that large samples can be measured readily.

Aside from contact mode, most AFMs are equipped with dynamic operational mode capabilities, such as the amplitude modulation mode and the frequency modulated mode (also referred to as the self-excitation mode), which also can be used for surface roughness measurements. It should be noted that these dynamic modes could be implemented as an intermittent contact or noncontact imaging experiments. The amplitude modulation mode, also referred to as "tapping mode" will be described further below, as it is the most widely used among the dynamic modes of AFM operation.

During a surface scan in tapping mode, the cantilever/tip assembly with a normal stiffness of 20–100 N/m (tapping mode etched Si probe or TESP) is sinusoidally vibrated at its resonance frequency (350–400 kHz) by a piezo mounted above it, and the oscillating tip slightly taps the surface. The piezo is adjusted using feedback control in the Z-direction to maintain a constant (20–100 nm) oscillating amplitude (setpoint) and constant average normal force, Fig. 2 [2–4]. The feedback signal to the Z-direction sample piezo (to keep the setpoint constant) is a measure of surface roughness. The cantilever/tip assembly is vibrated at some amplitude, here referred to as the free amplitude, before the tip engages the sample. The tip engages the sample at some setpoint, which may be thought of as the amplitude of the cantilever as influenced by contact with the sample. The setpoint is defined as a ratio of the vibration amplitude after engagement to the vibration amplitude in free air before engagement. A lower setpoint gives a reduced amplitude and closer mean tip-to-sample distance. The amplitude should be kept large enough so that the tip does not get stuck to the sample because of adhesive attractions. Also, the oscillating amplitude applies less average (normal) load as compared to contact mode and reduces sample damage. The tapping mode is used in topography measurements to minimize effects of friction and other lateral forces, and is well suited for topographic imaging of soft surfaces.



Atomic Force Microscopy, Fig. 2 Schematic of tapping mode used to obtain height and phase data and definitions of free amplitude and setpoint. During scanning, the cantilever is vibrated at its resonance frequency and the sample X-Y-Z piezo is adjusted by feedback control in the Z-direction to maintain a constant setpoint. The computer records height (which is a measure of surface roughness) and phase angle (which is a function of the viscoelastic properties of the sample) data

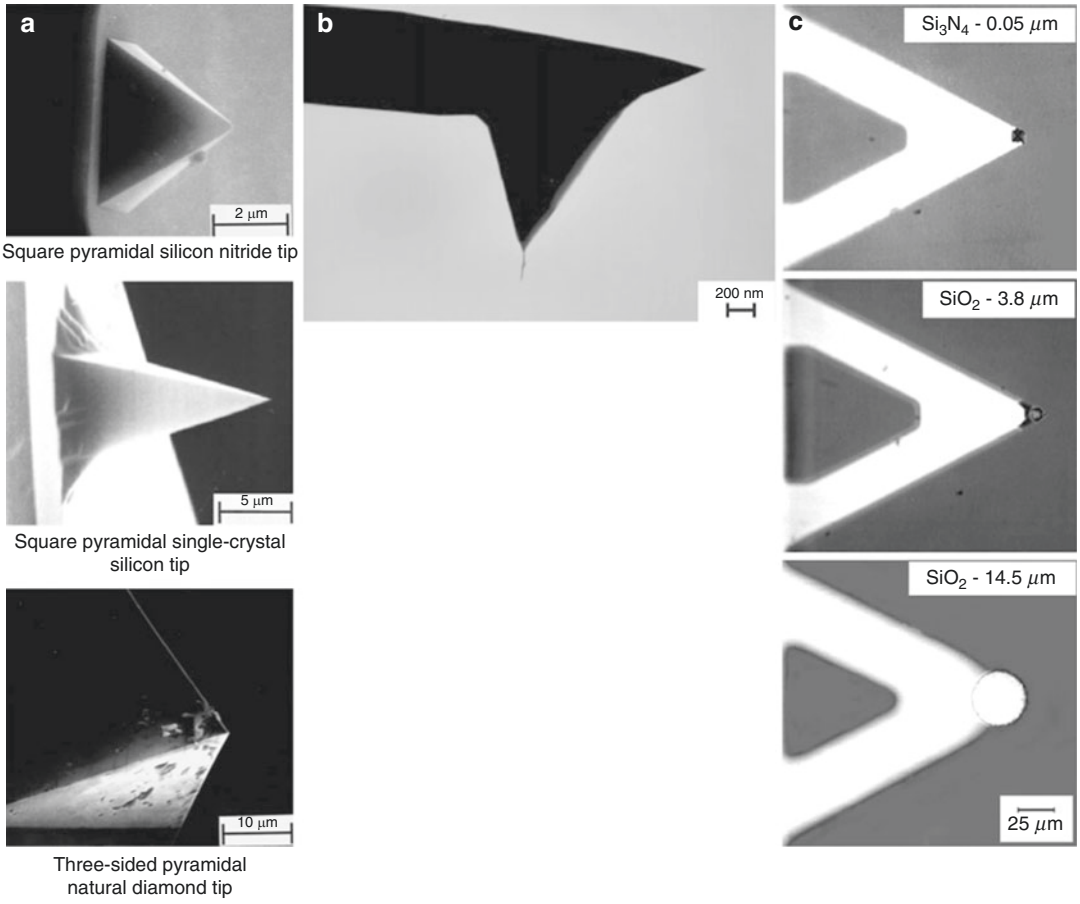
For measurement of friction force at the tip surface during sliding, left hand, and right hand sets of quadrants of the photodetector are used.

In the so-called friction mode, the sample is scanned back and forth in a direction orthogonal to the long axis of the cantilever beam. A friction force between the sample and the tip will produce a twisting of the cantilever. As a result, the laser beam will be reflected out of the plane defined by the incident beam and the beam reflected vertically from an untwisted cantilever. This produces an intensity difference of the laser beam received in the left hand and right hand sets of quadrants of the photodetector. The intensity difference between the two sets of detectors (FFM signal) is directly related to the degree of twisting and hence to the magnitude of the friction force. One problem associated with this method is that any misalignment between the laser beam and the photodetector axis would introduce error in the measurement. However, by following the procedures developed by Ruan and Bhushan [5], in which the average FFM signal for the sample scanned in two opposite directions is subtracted from the friction profiles of each of the two scans, the misalignment effect is eliminated. This method provides three-dimensional maps of friction force. By following the friction force, calibration procedures developed by Ruan and Bhushan [5], voltages corresponding to friction forces can be converted to force units [6]. The coefficient of friction is obtained from the slope of friction force data measured as a function of normal loads typically ranging from 10 to 150 nN. This approach eliminates any contributions due to the adhesive forces [2–4]. For calculation of the coefficient of friction based on a single-point measurement, friction force should be divided by the sum of applied normal load and intrinsic adhesive force. Furthermore it should be pointed out that for a single asperity contact, the coefficient of friction is not independent of load.

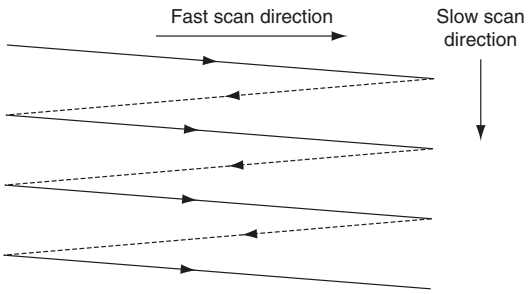
Surface roughness measurements in the contact mode are typically made using a sharp, microfabricated square-pyramidal Si_3N_4 tip with a radius of 30–50 nm on a triangular cantilever beam (Fig. 3a) with normal stiffness on the order of 0.06–0.58 N/m with a normal resonant frequency of 13–40 kHz (silicon nitride probe or NP) at a normal load of about 1–50 nN, and friction measurements are carried out in the load range of 1–100 nN. Surface roughness measurements in the tapping mode utilize a stiff cantilever with high resonance frequency; typically a square-pyramidal etched single-crystal silicon tip, with a tip radius of 5–10 nm, integrated with a stiff rectangular silicon cantilever beam (Fig. 3a) with a normal stiffness on the order of 17–60 N/m and a normal resonance frequency of 250–400 kHz (TESP), is used. Multiwalled carbon nanotube tips having a small diameter (few nm) and a length of about 1 μm (high aspect ratio) attached on the single-crystal silicon, square-pyramidal tips are used for high- resolution imaging of surfaces and of deep trenches in the tapping mode (noncontact mode) (Fig. 3b) [3, 4]. The MWNT tips are hydrophobic. To study the effect of the radius of a single asperity (tip) on adhesion and friction, microspheres of silica with radii ranging from about 4 to 15 μm are attached at the end of cantilever beams. Optical micrographs of two of the microspheres at the ends of triangular cantilever beams are shown in Fig. 3c.

The tip is scanned in such a way that its trajectory on the sample forms a triangular pattern, Fig. 4. Scanning speeds in the fast and slow scan directions depend on the scan area and scan frequency. Scan sizes ranging from less than 1 nm × 1 nm to 125 μm × 125 μm and scan rates from less than 0.5–122 Hz typically can be used. Higher scan rates are used for smaller scan lengths. For example, scan rates in the fast and slow scan directions for an area of 10 μm × 10 μm scanned at 0.5 Hz are 10 μm/s and 20 nm/s, respectively.

In an AFM measurement during surface imaging, the tip comes in intimate contact with the sample surface and leads to surface deformation with finite tip-sample contact area (typically few atoms). The finite size of contact area prevents the imaging of individual point defects, and only the periodicity of the atomic lattice can be imaged. Figure 5 shows the topography and friction force images of a freshly cleaved surface of highly oriented pyrolytic graphite (HOPG) [2–4]. The atomic-scale friction force of HOPG exhibited the same periodicity as the corresponding topography.



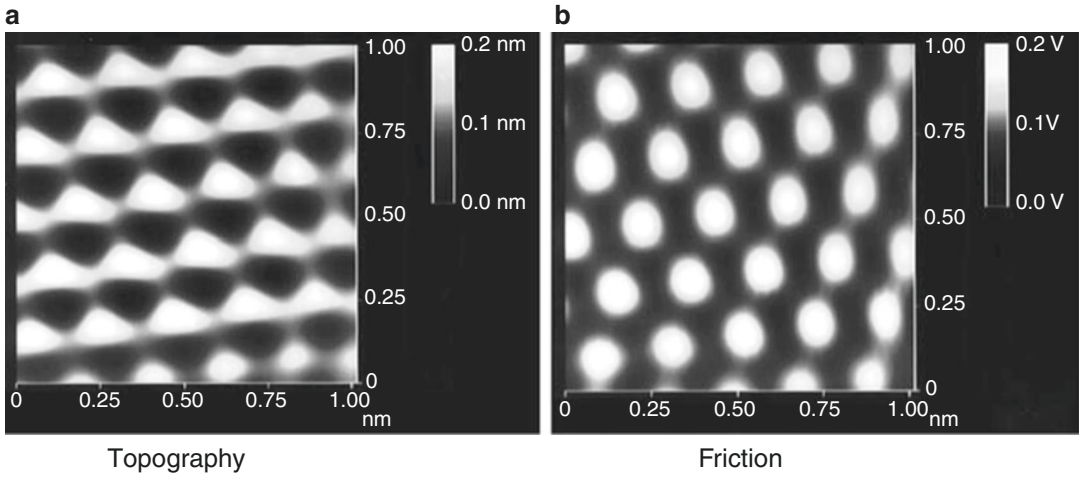
Atomic Force Microscopy, Fig. 3 (**a**) SEM micrographs of a square-pyramidal plasma-enhanced chemical vapor deposition (PECVD) Si_3N_4 tip with a triangular cantilever beam, a square-pyramidal etched single-crystal silicon tip with a rectangular silicon cantilever beam, and a three-sided pyramidal natural diamond tip with a square stainless steel cantilever beam, (**b**) SEM micrograph of a multiwalled carbon nanotube (MWNT) physically attached on the single-crystal silicon and square-pyramidal tip, and (**c**) optical micrographs of commercial Si_3N_4 tip and two modified tips showing SiO_2 spheres mounted over the sharp tip, at the end of the triangular Si_3N_4 cantilever beams (radii of the tips are given in the figure)



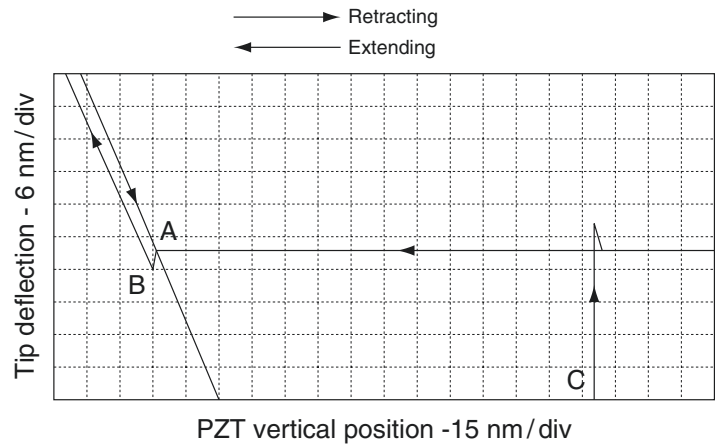
Atomic Force Microscopy, Fig. 4 Schematic of triangular pattern trajectory of the tip as the sample (or the tip) is scanned in two dimensions. During scanning, data are recorded only during scans along the solid scan lines

Adhesion Measurements

Adhesive force measurements are performed in the so-called force calibration mode. In this mode, force-distance curves are obtained. An example is shown in Fig. 6 for the contact between a Si_3N_4 tip and a single-crystal silicon surface. The horizontal axis gives the distance the piezo (and hence the sample) travels, and the vertical axis gives the tip deflection. As the piezo extends, it approaches the tip, which is at this point in free air and hence shows no deflection. This is indicated by the flat portion of the curve. As the tip approaches the sample within a few nanometers



Atomic Force Microscopy, Fig. 5 (**a**) Gray-scale plots of surface topography and friction force maps (2D spectrum filtered), measured simultaneously, of a 1 nm × 1 nm area of freshly cleaved HOPG, showing the atomic-scale variation of topography and friction and (**b**) schematic of superimposed topography and friction maps from (**a**); the symbols correspond to maxima. Note the spatial shift between the two plots [3, 4]

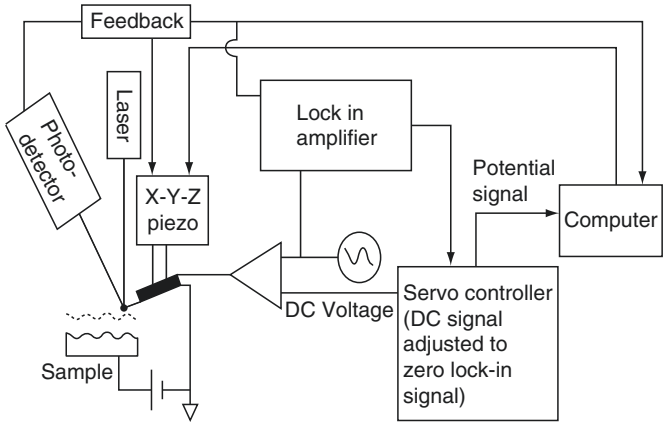


Atomic Force Microscopy, Fig. 6 Typical force-distance curve for a contact between Si_3N_4 tip and single-crystal silicon surface in measurements made in the ambient environment. Snap-in occurs at point A; contact between the tip and silicon occurs at point B; tip breaks free of adhesive forces at point C as the sample moves away from the tip

(point A), an attractive force exists between the atoms of the tip surface and the atoms of the sample surface. The tip is pulled toward the sample and contact occurs at point B on the graph. From this point on, the tip is in contact with the surface and as the piezo further extends, the tip gets further deflected. This is represented by the sloped portion of the curve. As the piezo retracts, the tip goes beyond the zero deflection (flat) line because of attractive forces (van der Waals forces and long-range meniscus forces), into the adhesive regime. At point C in the graph, the tip snaps free of the adhesive forces and is again in free air. The horizontal distance between points B and C along the retrace line gives the distance moved by the tip in the adhesive regime. This distance multiplied by the stiffness of the cantilever gives the adhesive force. Incidentally, the horizontal shift between the loading and unloading curves results from the hysteresis in the PZT tube [2–4].

Scratching, Wear, and Fabrication/Machining

For microscale scratching, microscale wear, nanofabrication/nanomachining, and nano-indentation hardness measurements, an extremely hard tip is required. A three-sided pyramidal



Atomic Force Microscopy, Fig. 7 Schematic of lift mode used to make surface potential measurement. The topography is collected in tapping mode in the primary scan. The cantilever piezo is deactivated. Using topography information of the primary scan, the cantilever is scanned across the surface at a constant height above the sample. An oscillating voltage at the resonant frequency is applied to the tip and a feedback loop adjusts the DC bias of the tip to maintain the cantilever amplitude at zero. The output of the feedback loop is recorded by the computer and becomes the surface potential map

single-crystal natural diamond tip with an apex angle of 80° and a radius of about 100 nm mounted on a stainless steel cantilever beam with normal stiffness of about 25 N/m is used at relatively higher loads (1–150 μN), Fig. 3a. For scratching and wear studies, the sample is generally scanned in a direction orthogonal to the long axis of the cantilever beam (typically at a rate of 0.5–50 Hz) so that friction can be measured during scratching and wear. The tip is mounted on the cantilever such that one of its edges is orthogonal to the long axis of the beam; therefore, wear during scanning along the beam axis is higher (about 2–3 x) than that during scanning orthogonal to the beam axis. For wear studies, an area on the order of 2 μm × 2 μm is scanned at various normal loads (ranging from 1 to 100 μN) for a selected number of cycles [2, 4].

Scratching can also be performed at ramped loads and the coefficient of friction can be measured during scratching [2–4]. A linear increase in the normal load approximated by a large number of normal load increments of small magnitude is applied using a software interface (such as the lithography module in Nanoscope) that allows the user to generate controlled movement of the tip with respect to the sample. The friction signal is tapped out of the AFM and is recorded on a computer. A scratch length on the order of 25 μm and a velocity on the order of 0.5 μm/s are used and the number of loading steps is usually taken to be 50.

Nanofabrication/nanomachining is conducted by scratching the sample surface with a diamond tip at specified locations and scratching angles. The normal load used for scratching (writing) is on the order of 1–100 μN with a writing speed on the order of 0.1–200 μm/s [2–4, 7].

Electrical Properties Measurements

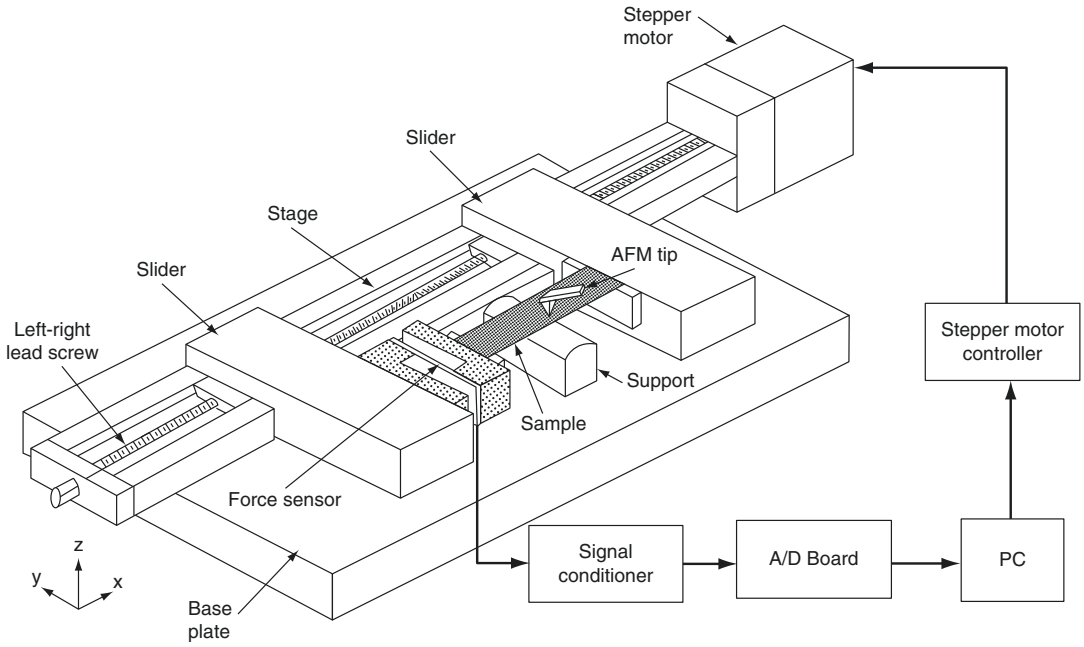
The use of a conductive tip in AFM experiments enables the mapping of surface or subsurface electrical properties of materials. Various AFM-based electrical characterization methods have been developed, such as scanning Kelvin probe force microscopy, scanning impedance microscopy, scanning capacitance microscopy, piezoresponse force microscopy, conductive AFM, scanning spreading resistance microscopy, and scanning gate microscopy. As an example, scanning Kelvin probe force microscopy, which is used for the measurement of the surface potential, is described below.

Mapping of the surface potential is made in the so-called lift mode (Fig. 7). These measurements are made simultaneously with the topography

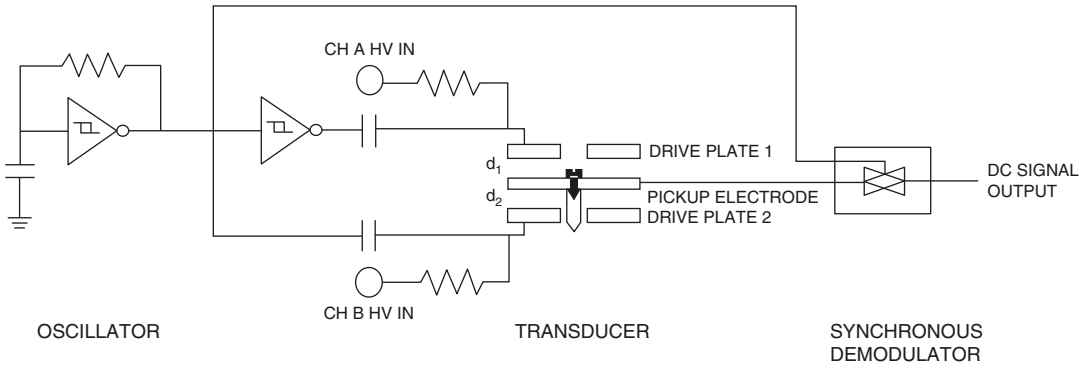
scan in the tapping mode, using an electrically conducting (nickel-coated single-crystal silicon) tip. After each line of the topography scan is completed, the feedback loop controlling the vertical piezo is turned off, and the tip is lifted from the surface and traced over the same topography at a constant distance of 100 nm. During the lift mode, a DC bias potential and an oscillating potential (3–7 V) is applied to the tip. The frequency of oscillation is chosen to be equal to the resonance frequency of the cantilever (~80 kHz). When a DC bias potential equal to the negative value of surface potential of the sample (on the order of ±2 V) is applied to the tip, it does not vibrate. During scanning, a difference between the DC bias potential applied to the tip and the potential of the surface will create DC electric fields that interact with the oscillating charges (as a result of the AC potential), causing the cantilever to oscillate at its resonance frequency, as in tapping mode. However, a feedback loop is used to adjust the DC bias on the tip to exactly nullify the electric field, and thus the vibrations of the cantilever. The required bias voltage follows the localized potential of the surface [8].

In Situ Characterization of Local Deformation Studies

In situ characterization of local deformation of materials can be carried out by performing tensile, bending, or compression experiments inside an AFM and by observing nanoscale changes during the deformation experiment [2–4]. In these experiments, small deformation stages are used to deform the samples inside an AFM. In tensile testing of the polymeric films carried out by Tambe and Bhushan [9], a tensile stage was used (Fig. 8). The stage with a left-right combination lead screw (that helps to move the slider in the opposite direction) was used to stretch the sample to minimize the movement of the scanning area, which was kept close to the center of the tensile specimen. One end of the sample was mounted on the slider via a force sensor to monitor the tensile load. The samples were stretched for various strains using a stepper motor and the same control area at different strains was imaged. In order to better locate the control area for imaging, a set of four markers was created at the corners of a 30 μm × 30 μm square at the center of the sample by scratching the sample with a sharp silicon tip.



Atomic Force Microscopy, Fig. 8 Schematic of the tensile stage to conduct in situ tensile testing of the polymeric films in AFM



Atomic Force Microscopy, Fig. 9 Schematic of a nano/picoindentation system with three-plate transducer with electrostatic actuation hardware and capacitance sensor [10]

The scratching depth was controlled such that it did not affect cracking behavior of the coating. A minimum displacement of 1.6 μm could be obtained. This corresponded to a strain increment of 8×10^{-3} % for a sample length of 38 mm. The maximum travel was about 100 mm. The resolution of the force sensor was 10 mN with a capacity of 45 N. During stretching, a stress–strain curve was obtained during the experiment to study any correlation between the degree of plastic strain and propensity of cracks.

Nanoindentation Measurements

For nanoindentation hardness measurements the scan size is set to zero, and then a normal load is applied to make the indents using the diamond tip. During this procedure, the tip is continuously pressed against the sample surface for about 2 s at various indentation loads. The sample surface is scanned before and after the scratching, wear, or indentation to obtain the initial and the final surface topography, at a low normal load on the order of 100–500 nN using the same diamond tip. An area larger than the indentation region is scanned to observe the indentation marks. Nanohardness is calculated by dividing the indentation load by the projected residual area of the indents [2–4].

Direct imaging of the indent allows one to quantify piling up of ductile material around the indenter. However, it becomes difficult to identify the boundary of the indentation mark with great accuracy. This makes the direct measurement of contact area somewhat inaccurate. A technique with the dual capability of depth-sensing as well as in-situ imaging, which is most appropriate in nanomechanical property studies, is used for accurate measurement of hardness with shallow depths [2–4, 10]. This nano/picoindentation system is used to make load-displacement measurement and subsequently carry out in-situ imaging of the indent, if required. The indentation system, shown in Fig. 9, consists of a three-plate transducer with electrostatic actuation hardware used for direct application of a normal load and a capacitive sensor used for measurement of vertical displacement. The AFM head is replaced with this transducer assembly while the specimen is mounted on the PZT scanner, which remains stationary during indentation experiments. The transducer consists of a three (Be-Cu) plate capacitive structure, and the tip is mounted on the center plate. The upper and lower plates serve as drive electrodes, and the load is applied by applying appropriate voltage to the drive electrodes. Vertical displacement of the tip (indentation depth) is measured by measuring the displacement of the center plate relative to the two outer electrodes using capacitance technique. Indent area and consequently hardness value can be obtained from the load-displacement data. The Young's modulus of elasticity is obtained from the slope of the unloading curve.

Localized Surface Elasticity and Viscoelasticity Mapping

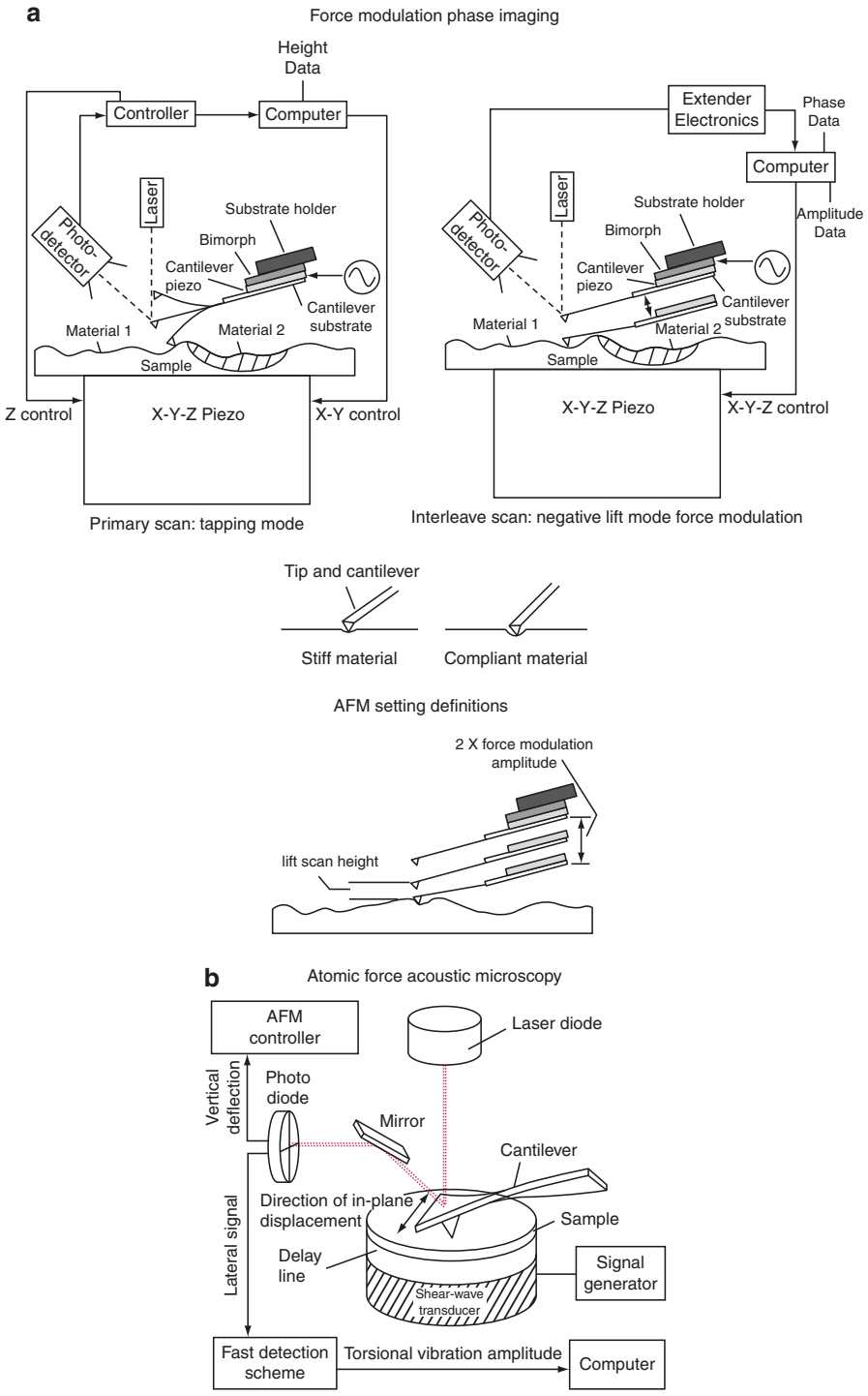
Localized Surface Elasticity. Indentation experiments provide a single-point measurement of the

Young's modulus of elasticity calculated from the slope of the indentation curve during unloading. Localized surface elasticity maps can be obtained using dynamic force microscopy, in which an oscillating tip is scanned over the sample surface in contact under steady and oscillating load. Lower frequency operation mode in the kHz range, such as force modulation mode [11] or pulsed force mode [12], are well suited for soft samples such as polymers. However, if the tip-sample contact stiffness becomes significantly higher than the cantilever stiffness, the sensitivity of these techniques strongly decreases. In this case, the sensitivity of the measurement of stiff materials can be improved by using high-frequency operation modes in the MHz range with a lateral motion, such as acoustic (ultrasonic) force microscopy, referred to as atomic force acoustic microscopy (AFAM) or contact resonance spectroscopy [13]. Inclusion of vibration frequencies other than only the first cantilever flexural or torsional resonance frequency also allows additional information to be obtained.

In negative lift mode force modulation technique, during primary scanning height data is recorded in tapping mode as described earlier. During interleave scanning, the entire cantilever/tip assembly is moved up and down at the force modulation holder's bimorph resonance frequency (about 24 kHz) at some amplitude, here referred to as the force modulation amplitude, and the Z-direction feedback control for the sample X-Y-Z piezo is deactivated, Fig. 10a [11]. During this scanning, height information from the primary scan is used to maintain a constant lift scan height. This eliminates the influence of height on the measured signals during the interleave scan. Lift scan height is the mean tip-to-sample distance between the tip and sample during the interleave scan. The lift scan height is set such that the tip is in constant contact with the sample, that is, a constant static load is applied. (A higher lift scan height gives a closer mean tip-to-sample distance.) In addition, the tip motion caused by the bimorph vibration results in a modulating periodic force. The sample surface resists the oscillations of the tip to a greater or lesser extent depending upon the sample's stiffness. The computer records amplitude (which is a function of elastic stiffness of the material). Contact analyses can be used to obtain a quantitative measure of localized elasticity of soft surfaces [3, 4]. Etched single-crystal silicon cantilevers with integrated tips (DI force modulation etched Si probe or FESP) with a radius of 25–50 nm, a stiffness of 1–5 N/m, and a resonant frequency of 60–100 kHz, are commonly used for the measurements. The scanning rate could be set between 0.1 and 2 Hz along the fast axis.

In the AFAM technique [13], the cantilever/tip assembly is moved either in the normal or lateral mode, and the contact stiffness is evaluated by comparing the resonance frequency of the cantilever in contact with the sample surface to those of the free vibrations of the cantilever. Several free resonance frequencies are measured. Based on the shift of the measured frequencies, the contact stiffness is determined by solving the characteristic equation for the tip vibrating in contact with the sample surface. The elastic modulus is calculated from contact stiffness using Hertz analysis for a spherical tip indenting a plane. Contact stiffness is equal to 8 multiplied with the contact radius and the reduced shear modulus in shear mode.

In the lateral mode using the AFAM technique, the sample is glued on cylindrical pieces of aluminum, which serves as ultrasonic delay lines coupled to an ultrasonic shear wave transducer, Fig. 10b [14]. The transducer is driven with frequency sweeps to generate in-plane lateral sample surface vibrations. These couple to the cantilever via the tip-sample contact. To measure torsional vibrations of the cantilever at frequencies up to 3 MHz, the original electronic circuit of the lateral channel of the AFM (using a lowpass filter with limited bandwidth to few hundred kHz) was replaced by a high-speed scheme which bypasses the lowpass filter. The high- frequency signal was fed to a lock-in amplifier, digitized using a fast A/D card, and fed into a broadband amplifier, followed by an rms-to-dc converter and read by a computer. Etched single-crystal silicon cantilevers (normal stiffness of 3.8–40 N/m) integrated tips are used.



Atomic Force Microscopy, Fig. 10 (**a**) Schematic of force modulation mode used to obtain amplitude (stiffness) and definitions of force modulation amplitude and lift scan height. During primary scanning, height data is recorded in tapping mode. During interleave scanning, the entire cantilever/tip assembly is vibrated at the bimorph's resonance frequency and the Z-direction feedback control for the sample X-Y-Z piezo is deactivated.

Viscoelastic Mapping. Another form of dynamic force microscopy, phase contrast microscopy, is used to detect the contrast in viscoelastic (viscous energy dissipation) properties of the different materials across the surface [15–18]. In these techniques, both deflection amplitude and phase angle contrasts are measured, which are measures of the relative stiffness and viscoelastic properties, respectively. Two phase measurement techniques – tapping mode and torsional resonance (TR) mode – have been developed, and are described below.

In the tapping mode (TM) technique, as described earlier, the cantilever/tip assembly is sinusoidally vibrated at its resonant frequency, and the sample X-Y-Z piezo is adjusted using feedback control in the Z-direction to maintain a constant setpoint, Fig. 3 [15]. The feedback signal to the Z-direction sample piezo (to keep the setpoint constant) is a measure of surface roughness. The extender electronics is used to measure the phase angle lag between the cantilever piezo drive signal and the cantilever response during sample engagement. As illustrated in Fig. 3, the phase angle lag (at least partially) is a function of the viscoelastic properties of the sample material. A range of tapping amplitudes and setpoints is used for measurements. Commercially etched single-crystal silicon tip (TESP) used for tapping mode, with a radius of 5–10 nm, a stiffness of 20–100 N/m, and a resonant frequency of 350–400 kHz, is normally used. Scanning is typically set between 0.1 and 2 Hz along the fast axis.

In the torsional mode (TR mode), a tip is vibrated in the torsional mode at high frequency at the resonance frequency of the cantilever beam. An etched single-crystal silicon cantilever with integrated tip (FESP) with a radius of about 5–10 nm, normal stiffness of 1–5 N/m, torsional stiffness of about 30 times normal stiffness and torsional resonant frequency of 800 kHz is typically used. A major difference between the TM and TR modes is the directionality of the applied oscillation – a normal (compressive) amplitude exerted for the TM and a torsional amplitude for the TR mode. The TR mode is expected to provide good contrast in the tribological and mechanical properties of the near-surface region as compared to the TM. Two of the reasons are as follows. (1) In the TM, the interaction is dominated by the vertical properties of the sample, so the tip spends a small fraction of its time in the near field interaction with the sample. Furthermore, the distance between the tip and the sample changes during the measurements, which changes interaction time and forces, and affects measured data. In the TR mode, the distance remains nearly constant. (2) The lateral stiffness of a cantilever is typically about two orders of magnitude larger than the normal (flexural) stiffness. Therefore, in the TM, if the sample is relatively rigid, much of the deformation occurs in the cantilever beam, whereas in the TR mode, much of the deformation occurs in the sample. A few comments on the special applications of the TR mode are made next. Since most of the deformation occurs in the sample, the TR mode can be used to measure stiff and hard samples. Furthermore, properties of thin films can be measured more readily with the TR mode. For both the TM and TR modes, if the cantilever is driven to vibrate at frequencies above resonance, it would have less motion (high apparent stiffness), leading to higher sample deformation and better contrast. It should be further noted that the TM exerts a compressive force, whereas the TR mode exerts torsional force, therefore normal and shear properties are measured in the TM and TR modes, respectively.

In the TR mode, the torsional vibration of the cantilever beam is achieved using a specially designed cantilever holder. It is equipped with a piezo system mounted in a cantilever holder, in which two piezos vibrate out-of-phase with

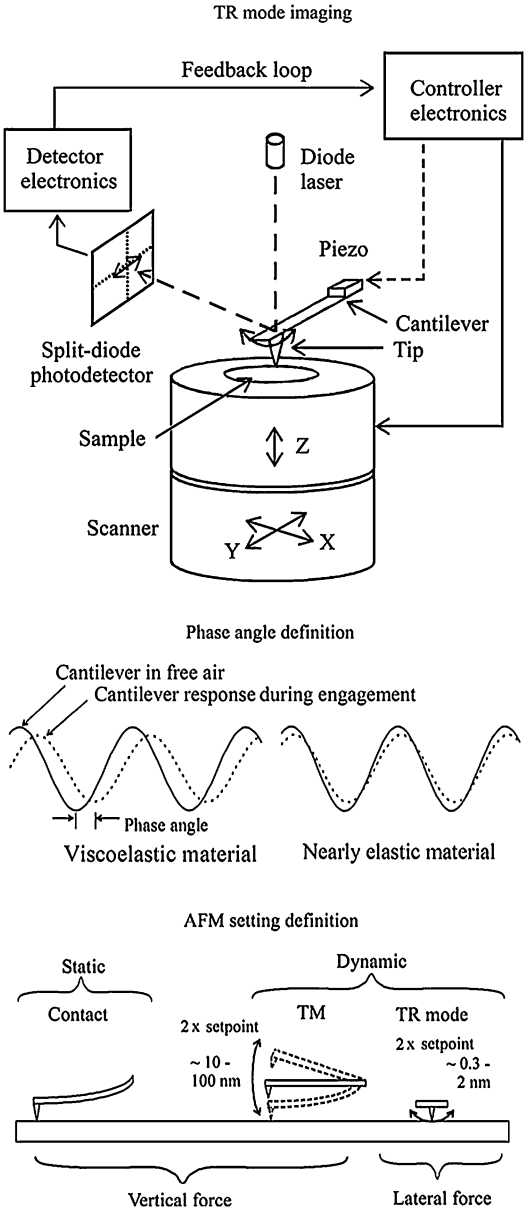
Atomic Force Microscopy, Fig. 10 (continued) During this scanning, height information from the primary scan is used to maintain a constant lift scan height. The computer records amplitude (which is a function of material stiffness) during the interleave scan, and (**b**) Schematic of an AFM incorporating shear wave transducer which generates in-plane lateral sample surface vibrations. Because of the forces between the tip and the surface, torsional vibrations of the cantilever are excited [14]. The shift in contact resonance frequency is a measure of contact stiffness

respect to each other. A tuning process prior to scanning is used to select the torsional vibration frequency. The piezo system excites torsional vibration at the cantilever's resonance frequency. The torsional vibration amplitude of the tip (TR amplitude) is detected by the lateral segments of the split-diode photodetector, Fig. 11 [18]. The TR mode measures surface roughness and phase angle as follows. During the measurement, the cantilever/tip assembly is first vibrated at its resonance at some amplitude dependent upon the excitation voltage, before the tip engages the sample. Next, the tip engages the sample at some setpoint. A feedback system coupled to a piezo stage is used to keep a constant TR amplitude during scanning. This is done by controlling the vertical position of the sample using a piezo moving in the Z direction, which changes the degree of tip interaction. The displacement of the sample Z piezo gives a roughness image of the sample. A phase angle image can be obtained by measuring the phase lag of the cantilever vibration response in the torsional mode during engagement with respect to the cantilever vibration response in free air before engagement. The control feedback of the TR mode is similar to that of tapping, except that the torsional resonance amplitude replaces flexural resonance amplitude [18].

Chen and Bhushan [16] used a variation to the approach just described (referred to as mode I here). They performed measurements at constant normal cantilever deflection (constant load) (mode II) instead of using constant setpoint in the mode I approach. Their approach overcomes the meniscus adhesion problem present in mode I, and reveals true surface properties.

Song and Bhushan [19] presented a forced torsional vibration model for a tip-cantilever assembly under viscoelastic tip-sample interaction. This model provides a relationship of torsional amplitude and phase shift with lateral contact stiffness and viscosity, which can be used to extract in-plane interfacial mechanical properties.

Various operating modes of AFM used for surface roughness, localized surface elasticity, and viscoelastic mapping, and friction force measurements are summarized in Table 2.



Atomic Force Microscopy, Fig. 11 Schematic of torsional resonance mode shown at the *top*. Two examples of the phase angle response are shown in the *middle*. One is for materials exhibiting viscoelastic (*left*) and the other nearly elastic properties (*right*). Three AFM settings are compared at the *bottom*: contact, tapping mode (*TM*), and TR modes. The TR mode is a dynamic approach with a laterally vibrating cantilever tip that can interact with the surface more intensively than other modes. Therefore, more detailed near-surface information is available

Atomic Force Microscopy, Table 2 Summary of various operating modes of AFM for surface roughness, stiffness, phase angle, and friction

Operating mode	Direction of cantilever vibration	Vibration frequency of cantilever (kHz)	Vibration amplitude (nm)	Feedback control	Data obtained
Contact	N/a			Constant normal load	Surface height, friction
Tapping	Vertical	350–400	10–100	Setpoint (constant tip amplitude)	Surface height, phase angle (normal viscoelasticity)
Force Modulation	Vertical	10–20 (bimorph)	10–100	Constant normal load	Surface height, amplitude (normal stiffness)
Lateral	Lateral (AFAM)	100–3,000 (sample)	~5 (sample)	Constant normal load	Shift in contact resonance (normal stiffness, friction)
TR mode I	Torsional	~800	0.3–2	Setpoint (constant tip amplitude)	Surface height, phase angle (lateral viscoelasticity)
TR mode II	Torsional	~800	0.3–2	Constant normal load	Surface height, amplitude and phase angle (lateral stiffness and lateral viscoelasticity)
TR mode III	Torsional	>800 in contact	0.3–2	Constant normal load	Shift in contact resonance (friction)

Boundary Lubrication Measurements

AFM can be used to study nanoscale boundary lubrication properties, mainly adhesion, friction, and wear (or durability). Adhesive forces at the nanoscale can be measured in the force calibration mode, as previously described. The adhesive forces can also be calculated from the horizontal intercept of friction versus normal load curves at a zero value of friction force [20]. For friction measurements, the unlubricated and lubricated surfaces (such as silicon wafers, bulk metal or metallic films, ceramics and polymer films, among others) are typically scanned using a Si_3N_4 or Si tip over scan length of 2–10 μm with the normal load ranging from 5 to 150 nN. Velocity effects on friction are studied by changing the scan frequency from 0.1 to 60 Hz. As an example, if the scan length is maintained at 2 μm, the sliding velocity then varies from 0.4 to 240 μm/s. To study the durability properties, the friction force and coefficient of friction are monitored during scanning at an arbitrary nanoscale normal load (on the order of 10–100 nN) and a predefined scanning speed for a desired number of cycles [20].

Cross-References

- ▶ AFM in Liquids
- ▶ AFM Probes
- ▶ AFM, Noncontact Mode
- ▶ AFM, Tapping Mode
- ▶ Robot-Based Automation on the Nanoscale

References

1. Binnig, G., Quate, C.F., Gerber, Ch.: Atomic force microscopy. Phys. Rev. Lett. **56**, 930–933 (1986)
2. Bhushan, B.: Handbook of Micro/Nanotribology, 2nd edn. CRC Press, Boca Raton (1999)
3. Bhushan, B.: Springer Handbook of Nanotechology, 3rd edn. Springer, Heidelberg (2010)
4. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
5. Ruan, J., Bhushan, B.: Atomic-scale friction measurements using friction force microscopy: part I – general principles and new measurement techniques. ASME J. Tribol. **116**, 378–388 (1994)
6. Palacio, M., Bhushan, B.: Normal and lateral force calibration techniques for AFM cantilevers. Crit. Rev. Solid State Mater. Sci. **35**, 73–104 (2010)

7. Bhushan, B., Israelachvili, J.N., Landman, U.: Nanotribology: friction, wear and lubrication at the atomic scale. Nature **374**, 607–616 (1995)
8. DeVecchio, D., Bhushan, B.: Use of a nanoscale kelvin probe for detecting wear precursors. Rev. Sci. Instrum. **69**, 3618–3624 (1998)
9. Tambe, N., Bhushan, B.: In situ study of nanocracking of multilayered magnetic tapes under monotonic and fatigue loading using an AFM. Ultramicroscopy **100**, 359–373 (2004)
10. Bhushan, B., Kulkarni, A.V., Bonin, W., Wyrobek, J. T.: Nano/picoindentation measurement using a capacitance transducer system in atomic force microscopy. Philos. Mag. **74**, 1117–1128 (1996)
11. Maivald, P., Butt, H.J., Gould, S.A.C., Prater, C.B., Drake, B., Gurley, J.A., Elings, V.B., Hansma, P.K.: Using force modulation to image surface elasticities with the atomic force microscope. Nanotechnology **2**, 103–106 (1991)
12. Krotil, H.U., Stifter, T., Waschipky, H., Weishaupt, K., Hild, S., Marti, O.: Pulse force mode: a new method for the investigation of surface properties. Surf. Interface Anal. **27**, 336–340 (1999)
13. Rabe, U., Janser, K., Arnold, W.: Vibrations of free and surface-coupled atomic force microscope cantilevers: theory and experiment. Rev. Sci. Instrum. **67**, 3281–3293 (1996)
14. Reinstaedtler, M., Rabe, U., Scherer, V., Hartmann, U., Goldade, A., Bhushan, B., Arnold, W.: On the nanoscale measurement of friction using atomic-force microscope cantilever torsional resonances. Appl. Phys. Lett. **82**, 2604–2606 (2003)
15. Bhushan, B., Qi, J.: Phase contrast imaging of nanocomposites and molecularly-thick lubricant films in magnetic media. Nanotechnology **14**, 886–895 (2003)
16. Chen, N., Bhushan, B.: Morphological, nanomechanical and cellular structural characterization of human hair and conditioner distribution using torsional resonance mode in an AFM. J. Microsc. **220**, 96–112 (2005)
17. Garcia, R., Tamayo, J., Calleja, M., Garcia, F.: Phase contrast in tapping-mode scanning force microscopy. Appl. Phys. A **66**, S309–S312 (1998)
18. Kasai, T., Bhushan, B., Huang, L., Su, C.: Topography and phase imaging using the torsional resonance mode. Nanotechnology **15**, 731–742 (2004)
19. Song, Y., Bhushan, B.: Atomic force microscopy dynamic modes: modeling and applications. J. Phys. Condens. Matter **20**, 225012 (2008)
20. Liu, H., Bhushan, B.: Nanotribological characterization of molecularly-thick lubricant films for applications to MEMS/NEMS by AFM. Ultramicroscopy **97**, 321–340 (2003)

Further Reading

Bhushan, B.: Adhesion and stiction: mechanisms, measurement techniques, and methods for reduction (invited). J. Vac. Sci. Technol. B **21**, 2262–2296 (2003)
Bhushan, B.: Scanning Probe Microscopy in Nanoscience and Nanotechnology. Springer, Heidelberg (2010)
Bhushan, B.: Scanning Probe Microscopy in Nanoscience and Nanotechnology, vol. 2. Springer, Heidelberg (2011)
Bhushan, B., Fuchs, H.: Applied Scanning Probe Methods II-IV. Springer, Heidelberg (2006)
Bhushan, B., Fuchs, H.: Applied Scanning Probe Methods VII. Springer, Heidelberg (2007)
Bhushan, B., Fuchs, H.: Applied Scanning Probe Methods XI-XIII. Springer, Heidelberg (2009)
Bhushan, B., Kawata, S.: Applied Scanning Probe Methods VI. Springer, Heidelberg (2007)
Bhushan, B., Fuchs, H., Hosaka, S.: Applied Scanning Probe Methods. Springer, Heidelberg (2004)
Bhushan, B., Fuchs, H., Kawata, S.: Applied Scanning Probe Methods V. Springer, Heidelberg (2007)
Bhushan, B., Fuchs, H., Tomitori, M.: Applied Scanning Probe Methods VIII-X. Springer, Heidelberg (2008)
Fuchs, H., Bhushan, B.: Biosystems – Investigated by Scanning Probe Microscopy. Springer, Heidelberg (2010)
Magonov, S.N., Whangbo, M.H.: Surface Analysis with STM and AFM: Experimental and Theoretical Aspects of Image Analysis. VCH, Weinheim (1996)
Meyer, E., Hug, H.J., Bennewitz, R.: Scanning Probe Microscopy: The Lab on a Tip. Springer, Heidelberg (2004)
Morita, S., Wiesendanger, R., Meyer, E.: Noncontact Atomic Force Microscopy. Springer, Heidelberg (2002)
Quate, C.F.: The AFM as a tool for surface imaging. In: Duke, C.B. (ed.) Surface Science: The First Thirty Years. North-Holland, Amsterdam (1994)

Atomic Force Microscopy in Liquids

▶ AFM in Liquids

Atomic Layer Chemical Vapor Deposition

▶ Atomic Layer Deposition

Atomic Layer Chemical Vapor Deposition (ALCVD)

▶ Chemical Vapor Deposition (CVD)

Atomic Layer CVD (AL-CVD)

► Atomic Layer Deposition

Atomic Layer Deposition

Charles L. Dezelah IV
Picosun USA, LLC, Detroit, MI, USA

Synonyms

ALD; Atomic layer chemical vapor deposition; Atomic layer CVD (AL-CVD); Atomic layer epitaxy (ALE); Molecular layer deposition (MLD)

Definition

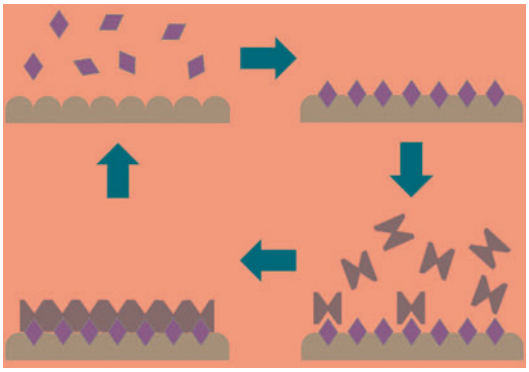
Atomic layer deposition is a method of thin film growth related to chemical vapor deposition whereby two or more vapor-phase precursor chemicals are individually and sequentially introduced to a substrate surface via discrete precursor pulses that are separated by purge steps. Each precursor pulse must yield saturative, self-limiting growth of a monolayer or sub-monolayer of the desired material if an ALD mechanism is operant. The resultant film exhibits excellent thickness uniformity and conformality, even on highly complex and demanding nanoscale substrate features, with the process affording an excellent degree of thickness control at the sub-nanometer level.

Introduction

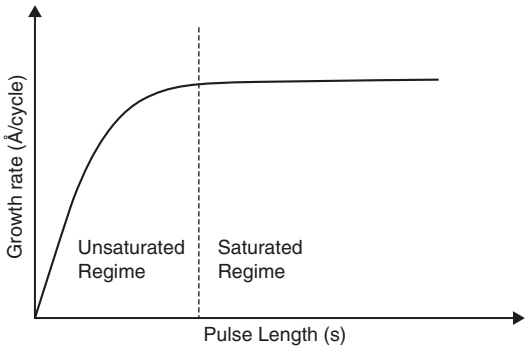
Atomic layer deposition (ALD) is a technique for the deposition of thin films of solid-state or polymeric materials onto solid substrates from vapor-phase precursor chemicals. While it is closely related to the chemical vapor deposition (CVD) technique and is similar to it in important ways, it also has substantial differences that distinguish it from CVD. Unlike CVD, ALD achieves film growth via the introduction of two or more precursor chemicals separately in discrete pulses into a reaction chamber. As a result, gas-phase reactions are avoided and only surface chemisorbed species are present at the time of exposure to the subsequent precursor. Growth with each exposure is self-limited and saturative, with each deposition cycle giving controlled deposition and layer-by-layer growth. As a result ALD is able to coat difficult substrate features and has been identified with its ability to overcome technological problems in the microelectronics industry, as well as potential for contributing to new areas in nanotechnology. ALD was first developed in Finland in the mid-1970s by Tuomo Suntola for thin film electroluminescent display applications, but experienced an enormous resurgence in the 2000s, mostly due to the need to replace SiO_2 gate dielectrics with high-permittivity alternatives to allow for future scaling of capacitor structures in microelectronics [1]. ALD continues to find new possibilities for future applications as the need for high-quality coatings on ever-smaller nano-sized features expands.

Key Features of an ALD Process

A number of key features are principal to the understanding of ALD [1–4]. Foremost is the concept of the ALD growth cycle. An ALD process is necessarily comprised of a sequence of pulses by which doses of precursor are delivered to a reaction chamber (see Fig. 1). In each pulse, a precursor is supplied in a dose sufficient to ensure reaction with every available surface site and over a time period that is long enough for the precursor vapor to effectively diffuse into substrate features and for reactions to occur on the surfaces that are present. The reaction between the precursor and the substrate must be self-limiting; the presence of excess precursor beyond that which is necessary for the reaction of all surface sites must not lead to additional film growth (Fig. 2). This behavior is often described as “saturative” growth and is the defining feature of ALD. A dose that has less than the required precursor amount to achieve this type of growth is considered “sub-saturative” and the



Atomic Layer Deposition, Fig. 1 A depiction of the ALD cycle. A precursor is alternately exposed to precursors that undergo self-limiting reactions with the surface. Repetition of the deposition cycle leads to layer-by-layer growth mechanism inherent to ALD



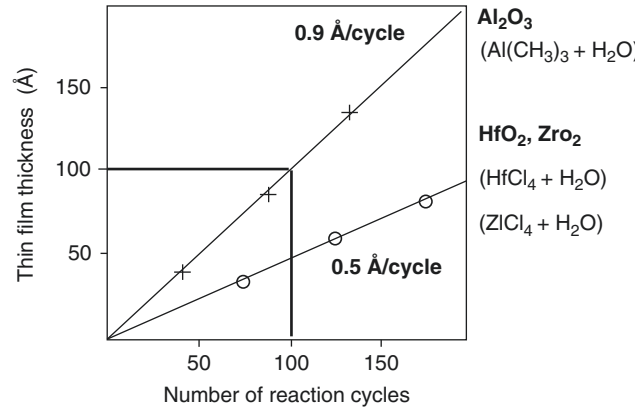
Atomic Layer Deposition, Fig. 2 ALD saturation curve. Each exposure of the substrate to precursor vapor leads to self-limiting growth, which is the defining feature of ALD. Any dose of sufficient size (here indicated by pulse length, which is one means of determining dose) leads to saturative growth whereby additional exposure does not lead to additional growth

resulting film is likely to not have the full characteristics expected of ALD. Additionally, it is crucial that no significant self-reaction or decomposition of the precursor occurs, as this may lead to an uncontrolled component of film growth.

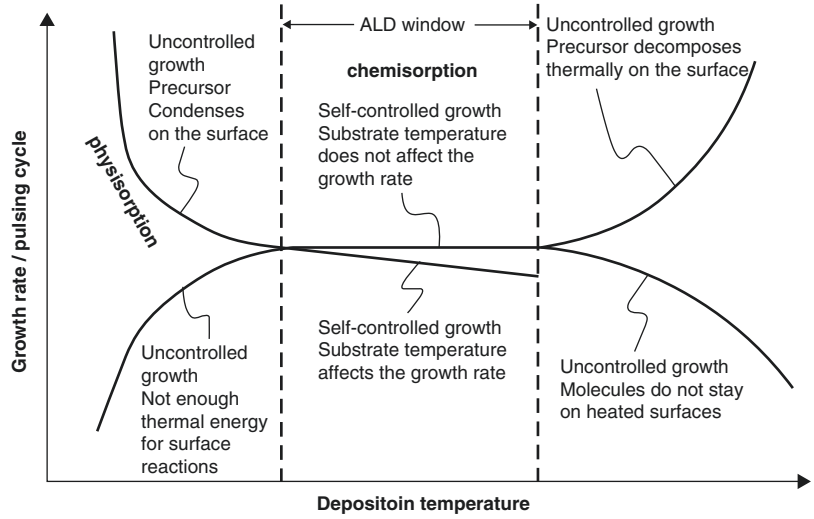
Upon the conclusion of a precursor pulse, the deposition chamber is purged, typically by the combined treatment of the deposition chamber by an inert gas flow and evacuation. Depending on the design of the ALD system being used, the application of inert gas flow and vacuum may be stopped and resumed to create a purge phase, or may be continuous, with precursor vapor introduced to the inert gas flow intermittently, so as to make the inert gas flow function alternately as a carrier gas and purge gas. The former of these approaches may be advantageous in that the pulse period can be lengthened to increase the residence time and effectively incubate the substrate without wasting precursor. The latter method, however, often exhibits shorter deposition cycles, faster deposition times, and more effective purging. The purge step must be long enough and effective enough to ensure that gas-phase reactions between precursors are avoided. The failure to avoid gas-phase mixing of the precursors can result in uncontrolled poor-quality dual-source-type CVD film growth and the formation of particulates in the deposition chamber, both of which can have deleterious effects on the intended film function.

Following the above-described pulse and purge steps, the substrate is exposed to a second precursor chemical chosen to react rapidly and completely with the first precursor, which is now already reacted with the surface, to achieve the desired material. This is followed again by a purging step to remove excess precursor vapor and any gas-phase reaction by-products, thus completing the second "half reaction" of the ALD cycle. By the repetition of the ALD cycle, the layer-by-layer growth of the thin film is achieved. A constant growth rate per deposition cycle is a feature of every ALD process, and as a result, the thickness of deposited film varies linearly with the number of cycle repetitions. The number of necessary cycles for a given film thickness for a particular process operated at a specific temperature is easily determined and is the basis for the degree of thickness control inherent in ALD (see Fig. 3).

The influence of deposition temperature on the growth rate is another fundamental ALD concept. Depending on the mechanistic phenomena taking place during the deposition of the film, the growth rate may change considerably, sometimes with loss of the self-limiting growth necessary for ALD. The common modes of growth behavior are outlined in Fig. 4. The concept of an ALD



Atomic Layer Deposition, Fig. 3 ALD exhibits excellent control of film thickness. By selecting a number of deposition cycles appropriate to the desired thickness, precise control can be achieved. For a given process at a given temperature, the growth per cycle is constant. As a result, film thickness increases linearly with the applied number of deposition cycles



Atomic Layer Deposition, Fig. 4 Depiction of how growth rate per cycle varies with temperature. Some ALD processes have an "ALD window" or temperature range over which the growth rate is constant. At high or low temperatures, a number of phenomena can occur leading to a large increase or decrease in growth rate and the loss of ALD behavior

window, a temperature range over which the growth rate does not change with temperature, is important to ALD. It is a common misunderstanding that an ALD window is a required attribute of every ALD process and that ALD only exists within an ALD window. This is not true, however, and in fact many well-known ALD processes totally lack an ALD window. These processes instead will display a modest increase or decrease in growth rate over a temperature range where self-limiting growth is observed. Nevertheless, an ALD window is considered an attractive feature, as it affords a greater degree of reproducibility by mitigating the influence of small variations in temperature. Regardless of the presence or absence of an ALD window, there is frequently a more dramatic change in growth rate as the temperature moves to a regime where the growth is no longer self-limiting. At low temperatures, a process may undergo a sizable decrease in growth rate as the degree of thermal activation becomes insufficient to sustain effective surface reactions. Alternatively, an increase in growth behavior may be observed as a result of precursor condensation and/or physisorption leading to the association of multiple monolayers of precursor with the substrate surface, all of which become involved with film growth upon exposure to the following precursor. At high temperatures a similar departure from normal ALD conditions can occur as well. Most often an excessive temperature will lead to a large increase in growth rate as precursor

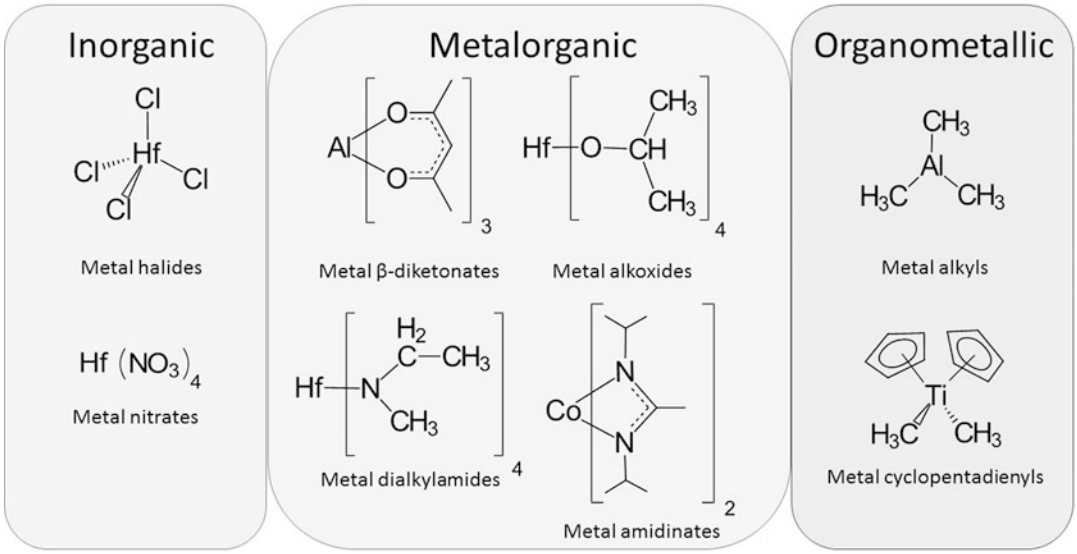
decomposition or self-reaction becomes a significant mode of film growth. High temperatures can also lead to a substantial drop in growth rate if precursor desorption from the surface becomes favorable. Both of these situations are likely to lead to loss of ALD growth.

Precursor Considerations

A number of factors are vital to the consideration of which precursor should be used for an ALD process [1–3]. Of primary importance is volatility. If a precursor compound cannot be effectively vaporized and delivered to the substrate surface in the gas phase, it will be impossible for the ALD process to proceed. A vapor pressure curve can be a very helpful tool in determining whether the volatility of a particular compound is suitable for effective delivery. Many precursors require heating to generate adequate vapor pressure for its delivery to the ALD chamber. The exact temperature required depends considerably on the configuration of the source delivery system used and the minimum vapor pressure requirement afforded by its design. A few precursors possess high enough vapor pressure at room temperature to ensure adequate dosing for ALD. These include some well-known examples such as trimethylaluminum (TMA), $Al(CH_3)_3$; titanium tetrachloride, $TiCl_4$; and diethylzinc, $Zn(CH_2CH_3)_2$. These provide the advantage not only of simplicity of delivery but also in allowing for low-temperature depositions of the corresponding metal oxides. Because the deposition temperature in a typical ALD reactor must exceed the temperature of the accompanying source delivery systems, low volatilization temperatures can help facilitate lower deposition temperatures. A further necessity of such low-temperature depositions – and for that matter, all ALD – is sufficient reactivity. In order for surface reactions to proceed via an ALD mechanism, both precursors must react rapidly and completely with the chemisorbed species left on the surface from the previous reaction. Insufficient reactivity may result in reduced growth rates, increased impurity content due to the inclusion of unreacted surface groups, and compromised film conformality and thickness uniformity. A third essential feature of an ALD precursor is thermal stability, that is, the inability of the precursor to undergo decomposition or self-reaction at a sufficient rate at the temperatures to which it is exposed to noticeably affect the quality of film growth. The lack of optimal thermal stability may result in uncontrolled single-source CVD-like growth and loss of ALD-type growth, oftentimes along with the loss of associated advantages of ALD. In addition to the three critical features of an ALD precursor, a number of other factors should be considered, such as availability, cost, ease of handling, the presence of undesired elements, the steric bulk of precursor molecules and its impact on film growth rate, reactivity and volatility of surface reaction by-products and their effects on the film and substrate, as well as environmental health and safety concerns.

Precursor Types

A variety of precursors exists for ALD and can be classified according to their key features [2, 3]. These features, as well as any associated advantages or disadvantages, must be considered carefully when determining the best process for a given application. Metal-containing precursors (i.e., for the delivery of the metal atoms to the growing film surface) can be divided into three categories: inorganic, metal-organic, and organometallic. Inorganic precursors are defined as those that do not contain carbon whatsoever. Typically inorganic precursors are restricted to metal halides, such as $TiCl_4$ for the growth of TiO_2 or TiN, but other examples exist as well, that is, the growth of HfO_2 from hafnium nitrate, $Hf(NO_3)_4$. Metal-organic precursors are those that contain one or more metal atoms and also contain at least one carbon-containing ligand bound to the metal through a heteroatom such as nitrogen or oxygen. Figure 5 shows common examples of metal-organic precursors. Classes of metal-organic precursors include alkoxides, amides, and amidinates. Such precursors are noted for their high volatility and good reactivity, although



Atomic Layer Deposition, Fig. 5 An overview of ALD precursor types and examples of each

many suffer from suboptimal thermal stability and exhibit a loss of ALD growth when the thermal budget has been exceeded. Due to this combination of characteristics, metal-organic precursors can be expected oftentimes to give ALD growth at the lower end of the range typical for ALD processes. Organometallic precursors are those that contain direct metal–carbon bond. Figure 5 shows common examples of organometallic precursors. These precursors primarily consist of metal alkyls, such as TMA, or cyclopentadienyl (Cp) compounds, such as dichlorohafnocene, $HfCp_2Cl_2$, or compounds containing both ligand types, such as dimethylhafnocene, $HfCp2(CH_3)_2$. Organometallic compounds typically have fairly good to excellent volatility, good to excellent reactivity, and excellent thermal stability. Their primary drawback is the limited availability of these compounds for many metals. For some organometallic compounds, high cost can also be an additional disadvantageous factor.

Besides metal-containing precursors, one must give proper consideration to the nonmetal precursor as well. A good nonmetal precursor must abide by all of the conditions and requirements for a good metal-containing precursor. Although many of the most commonly used nonmetal precursors are compressed gases where volatility is not an issue of concern, some nonmetal precursors are liquids, or perhaps even solids. In this case, good volatility is essential. Likewise, good thermal stability and reactivity are necessary, but due to the nature of the simple and robust nonmetal precursors commonly employed, only the second of these properties is usually a constraint. The reactivity of the nonmetal precursor must be appropriately matched to the metal-containing precursor and ideally should react in a vigorous and complete way that precludes the incorporation of impurities, as well as having oxidative or reductive properties adequate for achieving the desired oxidation state of material. Typically, metal oxide processes employ either H_2O or ozone. The use of H_2O commonly affords the metal oxide product by means of protonolysis and loss of the ligands bound to the metal of the adsorbed metal-containing precursor, along with replacement of the ligands by a hydroxyl group. The hydroxyl groups will then dominate the surface, making it amenable to reaction with the subsequent reintroduction of the metal-containing precursor. Ozone operates by a different mechanism, however, where the ligands undergo oxidation to CO_2 and H_2O, the latter of which is presumed to regenerate a hydroxyl-terminated surface. Ozone also brings with it a much greater

oxidative power than water, and its use typically assures that the resulting material will be in the highest accessible oxidation state. There are other precursors that have seen use for oxide growth, such as H_2O_2 and N_2O, but these are only used rarely. Oxygen (O_2) is typically not used in the growth of oxides, as its oxidative power and capability of promoting oxide growth are low at the temperatures used for ALD. Oxygen has been used extensively, however, in the growth of ALD films of the noble metals, where the combustion of organic ligands and the resulting reduction of the metal atoms are featured. The deposition of nitrides and sulfides is achieved most frequently using ammonia and hydrogen sulfide, respectively, which are analogous to water in that they are "hydrides" of their respective nonmetals and undergo protonolysis and ligand replacement reactions to achieve film growth.

Materials Grown by ALD

An ever-increasing variety of materials are available for growth by ALD. Comprehensive review articles exist that list and describe all known processes in detail [1–5]. It is beyond the scope of this entry to exhaustively list these processes, describe, and give references for them. Instead, a general overview of the materials according to type will be given, and the reader is encouraged to consult the recommended reviews for leading articles.

The large majority of ALD-grown materials are binary, that is, comprised of one metallic and one nonmetallic element. Metal oxides are easily the most common binary materials for which ALD processes are known and encompass a considerable portion of the periodic table. The oxides of the group 3 metals (Sc, Y, and lanthanide elements), group 4 metals (Ti, Zr, Hf), and several main group metals (Al, Mg, Zn) are particularly well represented, with many published studies and multiple processes developed from a variety of ligand types. Given the choice of available processes for these materials, ALD users can often select the process that best matches their needs and can choose optimal volatility, reactivity, and thermal stability from among the available precursors. Beyond these well-established processes are numerous others for which at least preliminary process studies have been published. These include nearly all of the first-row transition metal oxides, a few second- and third-row transition metal oxides (e.g., Nb_2O_5, Ta_2O_5, WO_3, RuO_2, RhO_2, IrO_2), as well as additional main group oxides beyond those mentioned earlier (e.g., B_2O_3, Ga_2O_3, In_2O_3, SiO_2, GeO_2). A representative list of known material processes is given in Table 1. Besides binary metal oxide processes, there are numerous examples of ternary (two metals and oxygen) and quaternary (three metals and oxygen) oxides that have been investigated. The metal titanates are the most

Atomic Layer Deposition, Table 1 An overview of known ALD materials (Based on information in Ref. [1])

Types	Materials
Metal oxides	Al_2O_3, TiO_2, ZrO_2, HfO_2, Ta_2O_5, Nb_2O_5, Sc_2O_3, Y_2O_3, MgO, B_2O_3, SiO_2, GeO_2, La_2O_3, CeO_2, PrO_x, Nd_2O_3, Sm_2O_3, EuO_x, Gd_2O_3, Dy_2O_3, Ho_2O_3, Er_2O_3, Tm_2O_3, Yb_2O_3, Lu_2O_3, $SrTiO_3$, $BaTiO_3$, $PbTiO_3$, $PbZrO_3$, Bi_xTi_yO, Bi_xSi_yO, $SrTa_2O_6$, $SrBi_2Ta_2O_9$,$YScO_3$, $LaAlO_3$, $NdAlO_3$, $GdScO_3$, $LaScO_3$, $LaLuO_3$, $Er_3Ga_5O_{13}$, In_2O_3, In_2O_3:Sn, In_2O_3:F, In_2O_3:Zr, SnO_2, SnO_2:Sb, ZnO, ZnO:Al, ZnO:B, ZnO:Ga, RuO_2, RhO_2, IrO_2, Ga_2O_3, V_2O_5, WO_3, W_2O_3, NiO, FeO_x, CrO_x, CoO_x, MnO_x, $LaCoO_3$, $LaNiO_3$, $LaMnO_3$, $La_{1-x}Ca_xMnO_3$
Metal nitrides	BN, AlN, GaN, InN, SiN_x, Ta_3N_5, Cu_3N, Zr_3N_4, Hf_3N_4, TiN, Ti–Si–N, Ti–Al–N, TaN, NbN, MoN, WN_x, WN_xC_y
Metal chalcogenides	ZnS, ZnSe, ZnTe, CaS, SrS, BaS, CdS, CdTe, MnTe, HgTe, La_2S_3, PbS, In_2S_3, Cu_xS, $CuGaS_2$, Y_2O_2S, WS_2, TiS_2, ZnS:M (M = Mn, Tb, Tm); CaS:M (M = Eu, Ce, Tb,Pb); SrS: M (M = Ce, Tb, Pb)
Metal pnictides	GaAs, AlAs, AlP, InP, GaP, InAs
Elements	Ru, Pt, Ir, Pd, Rh, Ag, W, Cu, Co, Fe, Ni, Mo, Ta, Ti, Al, Si, Ge
Others	CaF_2, SrF_2, MgF_2, LaF_3, ZnF_2, SiC, TiC_x, TaC_x, WC_x, $Ca_x(PO_4)_y$, $CaCO_3$, $Ge_2Sb_2Te_5$

frequently reported materials of this type and include examples such as $SrTiO_3$, $BaTiO_3$, $PbTiO_3$, and Bi_xTi_yO. Additionally, lanthanide-based ternary and quaternary materials are also well precedented, with accounts of the following in the literature: $YScO_3$, $LaAlO_3$, $NdAlO_3$, $GdScO_3$, $LaScO_3$, $LaLuO_3$, $Er_3Ga_5O_{13}$, $LaCoO_3$, $LaNiO_3$, $LaMnO_3$, and $La_{1-x}Ca_xMnO_3$. Besides stoichiometric multielement oxide films, doping of films by ALD is a frequently used approach. In fact due to its ability to precisely control film composition by setting a pulse sequence designed for the repeated introduction of a dopant-containing precursor, ALD is specially suited to the formation of thin films of such materials. Most of the work done to date has involved the doping of zinc, tin, or indium oxide with a single element, namely, ZnO:M, where M = B, Al, or Ga; SnO:M, where M = Sb; and In2O3:M, where M = Sn, F, or Zr.

After oxides, metal pnictides represent the second most explored group of ALD materials. Of these, nitrides have received by far the most attention. Metal nitride deposition by ALD generally falls into two categories: early transition metal nitrides and main group element nitrides. Many examples from the former group have metallic properties and have potential applications as an adhesion and barrier layers in microelectronics, thus creating a motivation for their study. Early transition metal nitrides include TiN, Ti–Si–N, Ti–Al–N, TaN, NbN, MoN, WN_x, and WN_xC_y. The remaining metal nitrides are semiconductor materials where the metallic element is from either group 13 or 14: BN, AlN, GaN, InN, and SiN_x. Other metal pnictides rather predictably fall into the category of III–V compound semiconductor materials: AlP, GaP, InP, AlAs, GaAs, and InAs.

In addition to the abovementioned material types that have been deposited by ALD, metal chalcogenides have been demonstrated for a number of metals. Upon examination, some trends become apparent: most of the grown films contain either group 2 or group 12 metals. Zinc is particularly well studied, as ZnS, ZnSe, ZnTe, and ZnS:M (M = Mn, Tb, Tm) are all known. Work on the group 2 metals, however, has been thus far restricted to sulfides and doped variants thereof, such as CaS, SrS, BaS, CaS:M, and SrS:M, where M = Ce, Tb, Pb. Other cases of metal chalcogenides are shown in Table 1.

The ALD of pure elements, particularly metallic elements, has been of increasing interest in recent years. Most of the attention has focused on processes for the "noble" metals, which are late transition metals which have a relatively unstable oxide and readily exist in the zero-valent metallic state. Although several notable noble metals have yet to be reported for ALD, elements such as Ru, Rh, Ir, Pd, and Pt have well-behaved processes. The noble metal ALD work has largely centered on the counterintuitive reduction of organometallic compounds using oxygen as the second precursor compound. Studies have indicated that the organic ligands undergo a combustion-like reaction in the presence of oxygen affording the reduction of the metal center and the production of CO_2 as a by-product. Other metal ALD processes have been based on the use of traditional reducing agents to ensure the formation of metallic species, including hydrogen, formaldehyde (as formalin), silanes, diborane, or the vapor of reductive metals such as zinc. Using such an approach, additional metals have been deposited, such as Cu and W, but with varying degrees of success. Recent advances have led to new processes that have started to make transition metal films feasible. Remote plasma ALD and radical-enhanced ALD are now actively being used to access metal films that were previously difficult or unknown by traditional thermal ALD methods. These include metals such as Ag, Ni, and others. This continues to be an active area of research and with time the number of ALD films in this category is likely to expand.

Molecular Layer Deposition

One variant of ALD that has gained considerable attention in recent years is what has become known as molecular layer deposition (MLD). MLD is essentially an ALD process where one or more of the sequential self-limiting surface reactions involve a molecular species that is deposited and itself becomes a layer in the

resulting film [4, 6]. MLD precursors are typically organic molecules that possess two or more reactive groups capable of undergoing ALD reactions. At least two such groups are needed in an MLD precursor, so that reaction can be achieved with the existing surface while generating a new surface that bears groups that are amenable to reaction with the precursor molecules of the subsequent precursor pulse. A precursor that contains only one such reactive group can react with the surface, but it will effectively terminate film growth due to the lack of ability to carry out subsequent reactions. Films grown by MLD generally fall into two categories: organic polymers and hybrid organic–inorganic materials.

To grow organic polymer films by MLD, many of the same principle apply for ALD. The precursors must be sufficiently volatile for gas-phase delivery. There must be no significant self-reaction or decomposition that leads to uncontrolled growth behavior – a fact that can become an even larger constraint for MLD than for ALD due to the nature of the precursors. Finally, the organic precursors must react with each other rapidly and completely within the time frame of an MLD growth cycle. This restricts MLD to those organic precursors that exhibit highly reactive functional groups. A common motif has been the reaction of a difunctional precursor molecule with two heteroatoms bearing protonizable hydrogen atoms (e.g., a diamine) with a second precursor that contains two functional groups that react vigorously with such functional groups (e.g., a diacid chloride or a dianhydride). An example of the diacid chloride and diamine MLD reaction is terephthaloyl chloride ($ClCOC_6H_4COCl$) and *p*-phenylenediamine ($NH_2C_6H_4NH_2$) to form the aromatic polyamide poly(*p*-phenylene terephthalamide) (PPTA). Similarly, films can be grown by pyromellitic dianhydride and the use of an amine such as ethylenediamine or *p*-phenylenediamine, to form the corresponding polyimide layers.

The MLD of hybrid organic–inorganic materials has been of increasing interest to researchers [4]. The approach for growing hybrid films is similar to that of other MLD films with one main difference: the precursor containing two or more groups with protonizable hydrogen atoms is reacted with a metal-containing compound having ligands that vigorously react with these functional groups. The obvious choice would be the use of TMA, which possesses methyl groups that are highly and irreversibly reactive toward such hydrogen atoms and, as has been described earlier, is amenable toward ALD growth. The reaction of TMA or other aluminum alkyl compounds with organic diols, such as ethylene glycol, leads to the formation of "alucones," which are the most well-studied class of MLD-grown hybrid inorganic–organic films. Alucone MLD processes have been described as robust and exhibiting the linear growth with a number of cycles expected of an ALD process [4]. By extension, other highly reactive metal alkyls should exhibit a similar ability to grow MLD organic–inorganic hybrids. "Zincones" have been formed by the MLD reaction of diethylzinc with organic diols with similar growth characteristics to those observed for the alucones [4]. The number and variety of available hybrid materials should be expected to increase rapidly in coming years as additional research is performed.

A number of obstacles exist for MLD processes [4, 6]. The vapor pressure requirements for an MLD precursor can substantially limit the available choice of organic precursor compounds that are suitable. Furthermore, the organic compounds typically under consideration nearly always have thermal stability limitations, which put an upper bound on the temperatures that can be used for volatilization, as well as putting an effective upper limit on the deposition temperature. Another problematic issue is the porous structure of the MLD-deposited films and the potential it creates for diffusion of precursor vapor into these small narrow features and – worse – the extra time required for the unreacted precursor to diffuse out during the purge step of the MLD process. This creates the possibility for gas-phase reactions that can lead to uncontrolled CVD and loss of the self-limiting growth mechanism. A further problem in MLD is the potential reaction of both functional groups of a precursor molecule with available surface reactive sites which can act to terminate or poison

growth and can perhaps alter the film characteristics. Double reactions of this type can be reduced or eliminated by selecting precursors with inherent conformational hindrances that greatly reduce the probability of this reaction type. Alternatively, the use of heterobifunctional precursors, where two chemically different reactive groups are present, has been used to limit the incidence of double reactions with the surface.

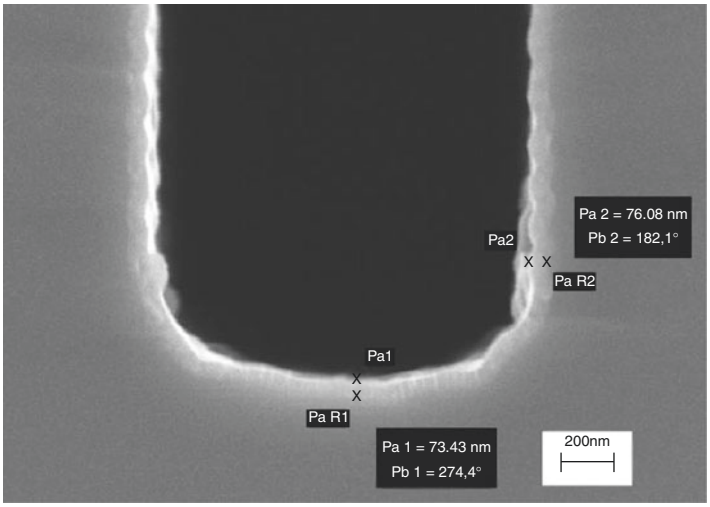
Plasma-Assisted ALD

Plasma-assisted ALD has experienced a trend toward increasing importance [1, 4, 7]. The central concept of plasma-assisted ALD differs from standard "thermal" ALD in that the reactivity of a precursor is increased by the application of plasma to generate radicals or other energetic species that can help promote reactions that might not otherwise progress by means of thermal activation alone. This advantage is commonly used to accomplish two ends: the reduction in deposition temperature of processes for which a thermal route is known and the ability to access difficult materials for which a thermal process may not be feasible or practical. In considering the former of these, the plasma-assisted process using TMA and O_2 plasma has been used to grow Al_2O_3 at low temperature, including down to room temperature, and has been shown to have improved dielectric properties relative to thermally grown Al_2O_3 ALD films [4]. An example of the latter type is TaN, which has not been grown in high quality by thermal-only methods, but has been achieved using an amido-/imido-based precursor and hydrogen radicals generated in a plasma source. The deposition of metal and metal nitride films is a common goal of plasma-assisted ALD processes, as these processes often require or benefit considerably from the advantages provided by plasma. Incidentally, it should be noted that plasma is always employed only for the generation of energetic nonmetal precursor species; it is not used with metal sources.

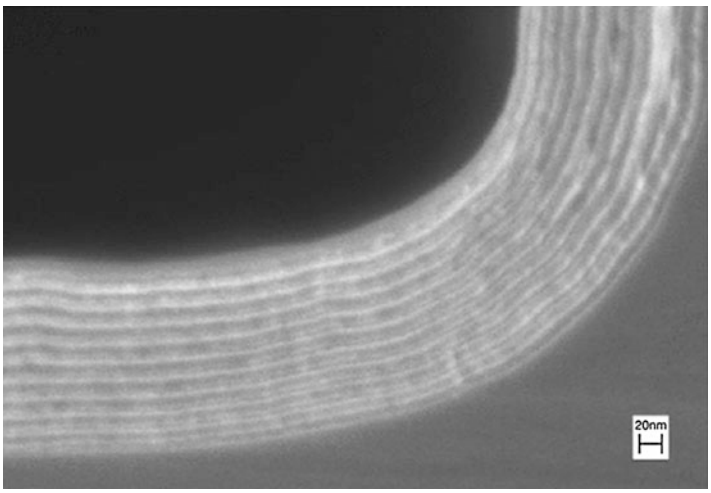
Plasma sources can be divided into two categories: direct plasma and remote plasma. Direct plasma sources are those where the substrate is directly exposed to the plasma, typically with the substrate resting on one of the plasma source electrodes [1]. Direct plasma is less commonly used in ALD due to drawbacks such as potential substrate damage. In remote plasma ALD, the substrate is situated some distance downstream from the plasma, with the energetic reactive species generated therein conveyed toward the substrate surface by the existing gas flow. The precursor pulse may contain a mixture of radicals and ions generated in the plasma, or in the case of a system that can be considered to be purely "radical-enhanced ALD," only radicals will be present as the reactive species.

Overview of Key Application Areas

Because of its ability to coat surfaces with exceptional control of thickness, excellent uniformity, and outstanding conformality and step coverage on complex and high aspect ratio surfaces (Fig. 6), ALD has found great potential for a range of applications [1, 8]. Foremost among these application areas is in microelectronics, where these three advantages are critical to the continued miniaturization of device features and the increase in performance demanded by Moore's law. Accordingly the ability to make coatings of advanced functional materials on increasingly challenging structures has proved valuable to the success of ALD, and microelectronics has been the major impetus for ALD technology development for the past two decades. Applications have generally fallen into three areas: (1) high-*k* dielectric gate oxide materials in metal oxide field effect transistors (MOSFET), (2) capacitor dielectrics in dynamic random access memory (DRAM), and (3) interconnect materials [1]. The driving force behind the use of high-*k* dielectrics in MOSFETs has been the need to replace SiO_2-based gate materials. The continual shrinkage of MOSFET features has led to a requirement that gate dielectric layers are only a few monolayers thick, which is a constraint that has allowed for tunneling and high leakage currents [1]. Alternate materials with high-permittivity values that can avoid the scaling problems inherent in SiO_2 have been a major focal



Atomic Layer Deposition, Fig. 6 ALD growth of Al_2O_3 in bottom of a 40:1 aspect ratio trench. ALD excels at coating challenging high aspect ratio features with uniform films and excellent conformality (Image: Picosun Oy proprietary)



Atomic Layer Deposition, Fig. 7 Nanolaminate thin film comprised of alternating layer of Al_2O_3 and TiO_2. Because of the stepwise methodology of ALD thin film growth, the deposition of nanolaminates and multicomponent composite materials is straightforward (Image: Picosun Oy proprietary)

point of ALD technology development. Hafnium oxide- and zirconium oxide-based materials have received the most attention, with phases containing silicon and nitrogen (HfSiON) considered the best solution for the near-term future [1, 8]. Active research continues on other materials that could become even better substitutes. DRAM production has also benefitted from the application of ALD. As with much of microelectronics, miniaturization has created a demand for DRAM features that are increasingly difficult to coat with the desired materials. Maintaining the required storage cell capacitance on a continually scaled down surface area and increasingly complex-shaped capacitor structure is a challenge to which ALD is particularly well suited. Capacitor dielectrics such as Hf–Al–O and nanolaminates of ZrO_2–Al_2O_3–ZrO_2 have seen recent use; however, it should be expected that the requirements of future technology nodes will necessitate the use of materials with a permittivity exceeding current benchmarks [1]. ALD is well suited to the deposition of such nanolaminate structures as shown in Fig. 7. Interconnects are another area where the application of ALD for microelectronics manufacture has been employed [1, 8]. Ruthenium and other noble metals have been grown by ALD and have been considered for seed layers for copper electroplating that can also act as an effective copper diffusion barrier. Early transition

metal nitrides have been evaluated also as copper diffusion barriers, but obstacles to their use remain. Tungsten processes based on the reduction of WF_6 have been used as seed layers for CVD tungsten contact plugs. Copper deposition for the formation of metallization seed layers is an attractive application area that continues to see active development and may eventually become a manufacturable process.

Optical applications represent a notable field where ALD has been used. ALD has the advantages of high uniformity, the ability to tailor refractive index, and film stress by combining several materials into one structure, all of which are appealing for optical devices [1, 8]. For planar optics, however, it seems unlikely that ALD will be helpful enough to overcome the traditional methods that have higher throughput. Optical designs that implement three dimensions into the structure, however, are more likely to gain from the use of ALD. For example, pillar structures of a selected arrangement have been conformally coated to produce microlens arrays, with integral interference filters incorporated via the deposition of an interference film stack [1]. Also ALD might be applicable to the fabrication of inverse opal photonic crystals, Fresnel zone plates for X-ray microscopy, and nanograting structures for polarizers and wave plates [1].

A variety of other application areas have also been explored. Due to the recent rapid miniaturization of magnetic read/write heads for inputting data onto hard disks, the deposition of conformal dielectrics by ALD has been implemented as insulating gap layers, as adhesion layers between the substrate and film stack, or as encapsulation layers [1, 8]. Thin film electroluminescent displays, which were the original application for which ALD was developed, continue to be a notable application area due to the necessity for defect-free films of uniform thickness. Doped metal sulfides, such as ZnS:Mn, are commonly used as the luminescent layer, whereas Al_2O_3 can be used as a passivation or protective layer [1]. Next-generation displays that incorporate organic light-emitting devices (OLEDs) have seen increasing use of ALD for thin film encapsulation of the device structure. The ALD films can be effective in limiting the rate of oxygen and water vapor permeation into the OLEDs, which can address one of the major factors limiting the useful lifetime of displays manufactured using this technology. Furthermore, MLD films can be used for flexible displays, where the ability to preserve the gas barrier properties over a surface that must endure repeated bending is required. ALD is also relevant to the development of photovoltaic applications. Transparent conductive oxides such as Al:ZnO have been deposited on complex or high surface area solar cells. Materials such as Al_2O_3 grown by ALD for passivation or charge recombination barrier layers have been of increasing importance. The application of ALD for solid-state 3-D Li-ion batteries has also been explored, where electrodes, current collectors, and electrolyte materials can all be potentially made using ALD thin films. Microelectromechanical systems (MEMS) are yet another area where ALD is experiencing a surge of interest. Because ALD can grow highly conformal and defect-free films on complex geometries, it can be used in MEMS components for wear-resistant, lubricating, anti-stiction, or charge-dissipating coatings [1, 9]. An emergent area of ALD research is in what could be described as nanotechnology: the coating of nano-sized features, such as those inside of porous materials, on the surface of nanowires, or onto nanotubes. Although the applications of this area are at this point largely prospective, the possibilities that could be realized are difficult to ignore. This topic has been the subject of a recent review to which the reader is referred for a detailed description [10].

Concluding Remarks

It is the purpose of this entry to give an overview of the basic principles, capabilities, and applications of ALD while referring the reader to several recent comprehensive review articles that treat specific areas in greater detail where relevant. In describing the unique qualities of ALD, it is apparent that the strong advantages of precise thickness control, thickness uniformity, and exceptional conformality lend themselves to a range of possible applications. Despite the fact

that ALD is an inherently slow deposition process, averaging around 1.0 Å/deposition cycle, the rapid miniaturization and increase in structural complexity of many devices has mitigated much of downside associated with its slowness. Correspondingly, these same developments have highlighted the need for ALD, as the limitations of traditional film growth methods have become an impediment to their future use in situations where challenging substrate features are involved. Additionally, in some areas, such as nanotechnology, ALD is likely to become a "critical enabler" that allows for unusual structures to be produced that would be unattainable by other methods. As such, the future for ALD is likely to yield new processes, novel research, and innovative use for an increasing number of applications.

Cross-References

- ▶ Chemical Vapor Deposition (CVD)
- ▶ Physical Vapor Deposition

References

1. Ritala, M., Niinistö, J.: Atomic layer deposition. In: Jones, A.C., Hitchman, M.L. (eds.) Chemical Vapor Deposition: Precursors, Processes and Application, pp. 158–206. Royal Society of Chemistry, Cambridge (2009)
2. Ritala, M., Leskelä, M.: Atomic layer deposition. In: Nalwa, H.S. (ed.) Handbook of Thin Film Materials, vol. 1, pp. 103–159. Academic, New York (2002)
3. Puurunen, R.L.: Surface chemistry of atomic layer deposition: a case study for the trimethylaluminum/water process. J. Appl. Phys. **97**, 121301 (2005)
4. George, S.M.: Atomic layer deposition: an overview. Chem. Rev. **110**, 111–131 (2010)
5. Kim, H.: Atomic layer deposition of metal and nitride thin films: current research efforts and applications for semiconductor device processing. J. Vac. Sci. Technol. B **21**, 2231–2261 (2003)
6. George, S.M., Yoon, B., Dameron, A.A.: Surface chemistry for molecular layer deposition of organic and hybrid organic–inorganic polymers. Acc. Chem. Res. **42**, 498–508 (2009)
7. Knoops, H.C.M., Langereis, E., van de Sanden, M.C. M., Kessels, W.M.M.: Conformality of plasma-assisted ALD: physical processes and modeling. J. Electrochem. Soc. **157**, G241–G249 (2010)
8. Ritala, M., Niinistö, J.: Industrial applications of atomic layer deposition. ESC Trans. **902**, 2564–2579 (2009)
9. Stoldt, C.R., Bright, V.M.: Ultra-thin film encapsulation processes for micro-electro-mechanical devices and system. J. Phys. D Appl. Phys. **39**, R-163–R-170 (2006)
10. Knez, M., Nielsch, K., Niinistö, L.: Synthesis and surface engineering of complex nanostructures by atomic layer deposition. Adv. Mater. **19**, 3425–3438 (2007)

Atomic Layer Deposition (ALD)

▶ Chemical Vapor Deposition (CVD)

Atomic Layer Epitaxial (ALE)

▶ Chemical Vapor Deposition (CVD)

Atomic Layer Epitaxy (ALE)

▶ Atomic Layer Deposition

Atomistic and Multiscale Modeling of Metal Nanoparticles

▶ Metal Nanoparticles from First Principles

Atomistic Simulations

▶ Simulations of Oxide/Polymer Hybrids

Attocalorimetry

▶ Nanocalorimetry

Automatic Data Analysis Workflow for RNAi

▶ Computational Systems Bioinformatics for RNAi

B

Bacterial Electrical Conduction

► Micro/Nano Transport in Microbial Energy Harvesting

Band Alignment

► Electrode–Organic Interface Physics

Band Pass Filters

► Tunable RF MEMS Filters: A Review

Bandpass Tunable Filter

► Tunable RF MEMS Filters: A Review

Basic MEMS Actuators

Arash Kheyraddini Mousavi[1], Seyedhamidreza Alaie[2] and Zayd Chad Leseman[1]
[1]Department of Mechanical Engineering, New Mexico Tech, Socorro, NM, USA
[2]Radiology, Weill Cornell Medicine, New York, NY, USA

Synonyms

Microactuators; MicroElectroMechanical systems; Micromachines

Definition

MEMS actuators are a type of MicroElectroMechanical System (MEMS) that convert energy into motion. MEMS are systems that integrate mechanical and electrical components with dimensions on the order of micrometers. Therefore, the typical motions achieved by MEMS actuators are on the order of micrometers as well.

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Overview

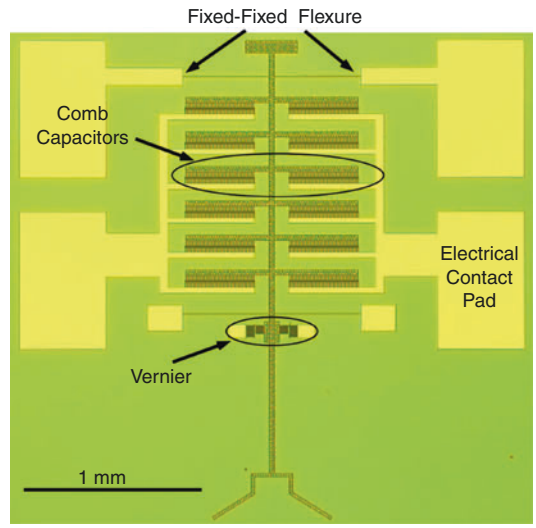
MEMS are systems that integrate mechanical and electrical components with dimensions on the order of micrometers. Though the concept of MEMS has existed since the 1960s, due to the advent of microfabrication techniques for miniaturizing electronic components, the term MEMS was not coined until 1986. Professors Jacobsen and Wood from the University of Utah devised this terminology in the course of writing a proposal to the Defense Advanced Research Projects Agency (DARPA) [1]. The term was then disseminated via a National Science Foundation (NSF) report, the Utah-held IEEE MEMS workshop in 1989, the IEEE/ASME *Journal of MEMS*, and subsequent DARPA MEMS funding solicitations. Since its inception, this term has gained wide acceptance as a catchall for microdevices in general.

MEMS actuators are a specific class of MEMS that convert energy into motion. The mechanisms by which energy is converted into motion are typically physical or chemical. Motion of the MEMS actuator can be used for positioning, open and closing valves, characterization of energy conversion processes, switching, and material characterization at the micro/nanoscale. A commercial application for positioning with MEMS actuators is that of Texas Instruments Digital Micromirror Device (DMD), wherein micromirrors are positioned (rotated) in order to direct light to create images in projectors for display applications. Opening and closing of (micro) valves is important for microfluidics and lab-on-a-chip applications. MEMS actuators can be configured to be energy storage devices or to detect quantities pertaining to energy conversion. Radio Frequency (RF) MEMS make use of MEMS actuators to create resonators for use in filters, reference oscillators, switches, switched capacitors, and varactors. Finally, MEMS actuators are commonly used to study material responses at the micro- and nanoscales. They are used to apply mechanical forces to materials in order to characterize their material properties mechanically.

Fabrication of MEMS actuators typically occurs via top-down fabrication methods. These methods begin with a larger piece of material, typically Si, and then shape the Si into the form of an actuator. Most MEMS actuators are extrusions of a 2-D pattern that are transferred into the Si via a photolithographic process. Three common methods by which MEMS actuators are fabricated are: Surface Micromachining, Silicon on Insulator (SOI) Surface Micromachining, and Single Crystal Reactive Etching and Metallization (SCREAM). Other processes exist, but many are a hybrid processes based in one of the prior mentioned processes.

This discussion of Basic MEMS Actuators will be broken into four parts. First, there will be a description of energy conversion mechanisms that generate forces; a specific discussion follows on comb capacitors and how they convert (electrical) energy into a force. Second, a discussion will ensue on how MEMS actuators convert the force into motion via a compliant structure. Third, methods for quantification of the motion will be discussed with specific emphasis on an integrated method for measurement of displacements. Finally, a brief description of the three main processes for fabrication of MEMS actuators will be given.

A MEMS actuator is depicted in Fig. 1. It uses comb capacitors for energy conversion into



Basic MEMS Actuators, Fig. 1 Basic MEMS Actuator that utilizes comb capacitors to generate force, a fixed-fixed flexure to guide the motion of the applied force, and a vernier to measure displacement

motion by deflecting a fixed-fixed beam flexure. Displacements of this actuator are measured using a vernier. All components mentioned will be discussed in detail here.

Energy Conversion into Forces and Motion

MEMS actuators convert different forms of energy into force and motion. For example, two electrically isolated parallel plates with an applied potential difference will develop an electrostatic force between the plates. If those plates are on flexible springs, then they will move toward one another. This type of energy conversion into motion is termed electrostatic actuation. Other types of actuators also use electrostatic actuation such as the nanotractor or inchworm actuator [2], scratch drive actuators [3], and the shuffle motor [4]. Beyond electrostatic actuators, many other actuators have been developed that use fluidic forces [5], magnetic forces [6], radiation pressure [7], piezoelectrics [8], shape-memory alloys [9], and thermal expansion [10].

Of particular importance for MEMS actuators is the way with which these forces scale with length; they scale differently due to the different physics involved. Table 1 displays a list of common forces and quantities that are of interest for MEMS actuators and gives how they scale with length.

Table 1 displays the way with which each force scales with length. Variables containing a geometric length pertinent to the physical system are in parentheses and moved to the right-hand side of the equation. It is these quantities that are used to determine quantities' sensitivity to length scale. Note each entry in Table 1 is highly dependent on the configuration of the system. In addition to common forces generated in MEMS actuators, this table also includes gravity, van der Waals forces, inertia, and the mass moment of inertia for reference.

Interpretation of Table 1 is accomplished by comparing the exponents of the fourth column. These exponents are used to compare the importance of a given energy conversion process at different length scales. l^0 indicates that the force-generating mechanism is invariant with length scale. l raised to a positive power means that the property grows with increasing scale, while a negative power implies that the property grows with decreasing scale.

Energy conversion mechanisms with negative exponents will behave more favorably at smaller length scales. Of the mechanisms listed, the easiest to employ and control are thermal (l^2) and the electrostatic forces generated by comb capacitors (l^0). Further discussion on thermal actuators is contained in the thermal actuators entry in this encyclopedia. Comb versus parallel plate capacitors are explained in more detail in the subsequent section. When compared to how inertia scales (l^3), both thermal and electrostatic actuation schemes behave more favorably at smaller length scales. The mass moment of inertia is also displayed in Table 1. The exponent of five (sphere) implies that it diminishes at shorter length scales more rapidly than any of the other forces. Because inertial force and mass moment of inertia are relatively small, compared to common forces used in MEMS actuators, their dynamic responses are much faster than their macroscopic counterparts.

Electrostatic and thermal actuators scale favorably with decreasing length scales and are more easily implemented than most other methods. Therefore, they have considerably more popularity than other energy conversion methods. In the remainder of this encyclopedia entry, MEMS actuators that use electrostatic forces, in particular capacitance due to comb capacitors, will be the topic of discussion. Thermal actuators are more thoroughly discussed in their entry in this encyclopedia.

Comb Capacitors

In order to generate the force applied to the MEMS actuator, one of the more common methods is to use capacitors. The common methods are parallel plate capacitors and comb capacitors. Parallel plate capacitors have a capacitance equal to:

Basic MEMS Actuators, Table 1 Sensitivity of physical quantities to length scale

Physical quantities	Examples	Governing equation	Sensitivity to length scale
van der Waals [11] $\breve{A}$:= *Hamaker constant* A:= *Area* d:= *Distance between objects* r:= *Radius* r_i:= *Radius of object "i"* l = *Facing lengths of two objects*	*Case 1:* Two spheres	$F = \frac{\breve{A}}{6}\left(\frac{r_1 r_2}{r_1+r_2}\frac{1}{d^2}\right)$	l^{-1}
	Case 2: A sphere and a surface	$F = \frac{A}{6}\left(\frac{r}{d^2}\right)$	l^{-1}
	Case 3: Two cylinders	$F = -\frac{A}{8\sqrt{2}}\left(\sqrt{\frac{r_1 r_2}{r_1+r_2}}\frac{l}{\sqrt{d^5}}\right)$	l^{-1}
	Case 4: Two crossed cylinders	$F = \frac{A}{6}\left(\frac{\sqrt{r_1 r_2}}{d^2}\right)$	l^{-1}
	Case 5: Two Surfaces	$F = \frac{A}{6\pi}\left(\frac{1}{d^3}\right)$	l^{-3}
Viscous forces μ:= *Dynamic viscosity* V_0:= *Relative velocity*	*Case 1:* Two infinite plates	$F = \mu V_0\left(\frac{1}{d}\right)$	l^{-1}
	Case 2: Two finite plates	$F = \mu V_0\left(\frac{A}{d}\right)$	l^{1}
Electrostatic force $\in_r$:= *Relative static permittivity* $\in_0$:= *Vacuum permittivity* V_e:= *Electrical potential* h:= *out-of-plane thickness*	*Case 1:* Finite parallel plates at distance d	$F = \frac{\in_0 \in_r V_e^2}{2}\left(\frac{A}{d^2}\right)$	l^0
	Case 2: Comb drive [12]	$F = \frac{\in_0 \in_r V_e^2}{2}\left(\frac{h}{d}\right)$	l^0
Thermal expansion E_y:= *Young modulus of elasticity* α_T:= *Therma expansion coefficient* ΔT:= *Temperature change*	*Case 1:* Constrained column	$F = E_y\ \alpha_T\ \Delta T A$	l^2
Magnetic forces [13] μ_0:= *Vacuum permeability* d:= *Distance between wires* l:= *Length along wire* I_i:= *Current in wire "i"* A_0:= *Cross sectional area* A_S:= *Surface area* $\dot{Q}_S$:= *Surface heat flow rate* I_e:= *Electrical current*	*Case 1:* Constant current density with the boundary condition $\frac{I_e}{A_0}$ = cons.	$F = \frac{\mu_0}{2\pi}\frac{l}{d} I_1 I_2$	l^4
	Case 2: Constant heat flow through the surface of the wire with the boundary condition $\frac{\dot{Q}_S}{A_S}$ = cons.	$F = \frac{\mu_0}{2\pi}\frac{l}{d} I_1 I_2$	l^3
	Case 3: Constant temperature rise of wire with the boundary condition ΔT = cons	$F = \frac{\mu_0}{2\pi}\frac{l}{d} I_1 I_2$	l^2
Piezoelectric force [14] $\in$:= *Mechanical strain* E_e:= *Electrical field* e_p:= *Piezoelectric constant* $\frac{\text{Charge density}}{\text{Applied strain}}$ E_y^E := *Young modulus at constant* E_e	*Case 1:* 1-D unconstrained actuation	$F = -e_p V_e\left(\frac{A_0}{d}\right) + E_y^E \in (A_0)$	l^0 and l^2
Drag force ρ:= *Density* C_d:= *Drag coefficient* A_p:= *Projected are normal to flow*	*Case 1:* Infinite cylinder ($C_d = 0.47$)	$F = \frac{1}{2}\rho V_0^2 C_d\left(A_p\right)$	l^2
	Case 2: Flat plate perpendicular to flow ($C_d = 1.28$)		
Surface tension force γ:= *Surface tension* p:= *Perimeter*		$F = \gamma(p)$	l^1
	Case 1: Weight	$F = \rho g(\forall)$	l^3

(*continued*)

Basic MEMS Actuators, Table 1 (continued)

Physical quantities	Examples	Governing equation	Sensitivity to length scale
Inertia and weight g:= *Gravity* α:= *Acceleration* $\forall$:= Volume	*Case 2:* Inertia force	$F = \rho a(\forall)$	l^3
Mass moment of inertia ρ_l:= *Mass per unit length* ρ_A:= *Mass per unit area* ρ_V:= *Mass per unit volume* r:= *Radius*	*Case 1:* Sphere	$\bar{I} = \frac{8\pi}{15}\rho_V(r^5)$	l^5
	Case 2: Thin circular disk	$\bar{I} = \frac{\pi}{2}\rho_A(r^4)$	l^4
	Case 3: Slender bar	$\bar{I} = \frac{1}{12}\rho_l(l^3)$	l^3
Shape-memory alloy [9] v:= *Poisson's ratio* h_f:= *Shape-memory alloy film thickness* h_s:= *Substrate thickness* R_i:= *Initial radius of curvature* R_i:= *Radius of curvature of the substrate with SAM film*	*Case 1:* Force caused by shape-memory alloy on a substrate	$F = -\frac{E_y}{1-v}\left(\frac{A_0 h_s^2 (r_1 - r_2)}{6h_f r_1 r_2}\right)$	l^2

$$C_{pp} = \frac{\epsilon_0 A}{d} \tag{1}$$

where ϵ_0 is the permittivity of the material between the plates (usually air or vacuum), A is the area of parallel plate capacitor, and d is the gap between the parallel plates. The electrostatic energy is $\frac{1}{2}C_{pp}\ V^2$, where V is the voltage between the parallel plate capacitors. Taking its derivative with respective to the gap, the force is calculated to be

$$F_{pp} = -\frac{\epsilon_0 A V^2}{2d^2} \tag{2}$$

Note that F_{pp} is always attractive and is nonlinear in not only the voltage, but also in the gap between the parallel plates. This geometric nonlinearity makes control of actuators difficult and led to the development of comb capacitors [15].

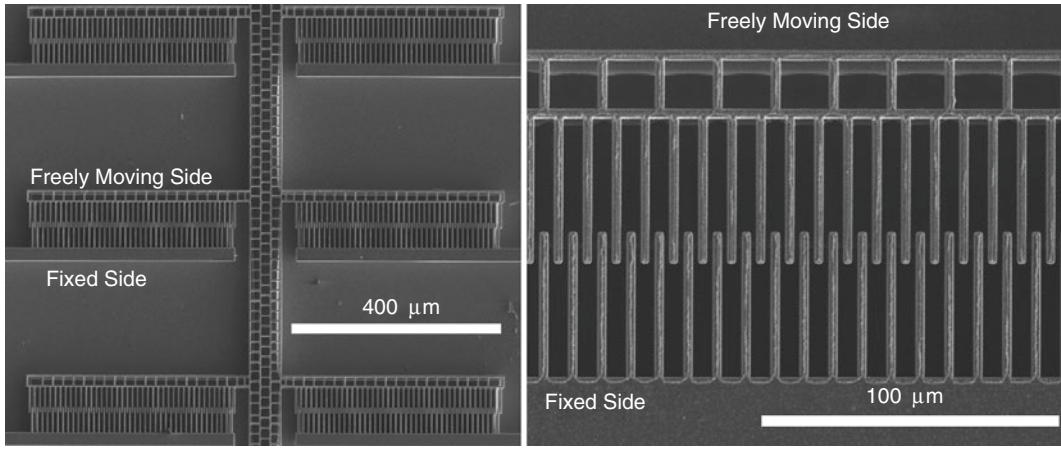
SEM images of a comb drive array are shown in Fig. 2. Comb capacitors are arrays of interdigitated fingers. One side of the array is rigidly attached to the substrate, while the other is allowed to move via a flexure structure which will be discussed in the next section.

Comb capacitor drives typically have hundreds to thousands of combs to generate a desired force. For analysis purposes, Fig. 3 shows one unit of a comb drive. When connected to an external source, opposite charges gather on the surfaces of the comb fingers which generate an electric field (see Fig. 4). Note that along the vertical line of symmetry of Fig. 4 that the horizontal components to the right and left are equal and opposite. Thus, the net force due to the horizontal components is zero. The vertical components of the fringing field at the tip of the comb finger, however, are not balanced and a net attractive force results. The capacitance of an array of comb capacitors is found to be:

$$C_{cc} = 2n\frac{\epsilon_0 h y}{d} \tag{3}$$

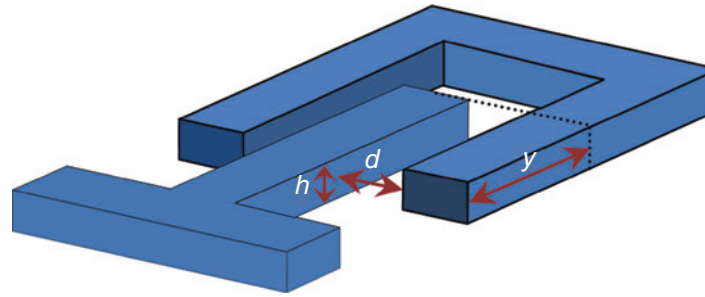
where n is the number of combs in the array, h is the depth of the comb, y is the overlap between the moving and fixed comb, and d is the gap between the moving and fixed comb. The attractive force generated by the comb capacitors when a voltage is applied is:

$$F_{cc} = \frac{n\epsilon_0 h}{d}V^2 \tag{4}$$



Basic MEMS Actuators, Fig. 2 Arrays of comb capacitors arranged symmetrically about the midline of a MEMS actuator (*left*). Closer view of a set of comb capacitors displaying the comb nature of the structure (*right*)

Basic MEMS Actuators, Fig. 3 A single unit of a comb drive actuator



The parameters in Eq. 4 are constant with the exception of the voltage. This type of relationship is more desirable than the relationship in Eq. 2, especially for dynamic excitation. Equations 3 and 4 are valid when the initial engagement of the comb is $>4d$ until the tip of the moving comb is within a distance $4d$ of the fixed side of the comb drive.

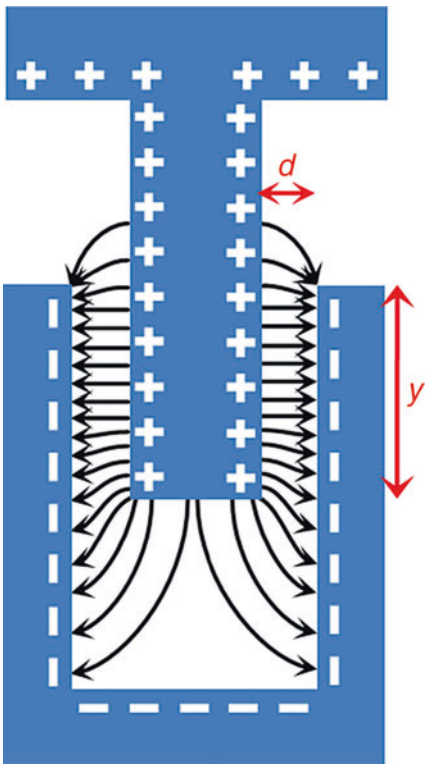
Comb drive fingers can be fabricated with various shapes to produce desired force-displacement relationships. For example, in [16] different comb capacitors have been designed, which exhibit linear, quadratic, and cubic response between force and displacement. The profiles of both stationary and moving fingers can be modified (as in Fig. 5), to obtain any desired force/displacement relationship. These types of responses between force and displacement are desirable because the spring constants used in many types of Basic MEMS actuators are nonlinear (see the subsequent section).

Spring Configurations for Basic MEMS Actuators

The force created by the conversion of energy is transmitted to a compliant structure (spring) via a stiff structure that traditionally comprises a majority of the mass – the proof mass. Though many different configurations have been used for the compliant structure, the two most common structures used are shown in Fig. 6. Figure 6a is a fixed-fixed (clamped-clamped) beam configuration, while Fig. 6b is referred to as the folded-beam flexure. Note that Fig. 6a is also referred to as a fixed-guided configuration. "*F*" in each of the figures represents the applied force onto the

structure by a means of energy conversion, as is described in the previous section.

One-dimensional motion is often desired. Therefore, one direction is designed to be considerably more compliant than the two orthogonal directions. This typically leads to beams that have slenderness ratios greater than 100. Additionally, structures must be designed such that the guided motion is parallel to the force. This requires a configuration that is symmetric about the axis of motion. Thus, only even numbers of beams can be connected to the proof mass. An additional concern is that asymmetric loading may cause the structure to rotate if there are only two connections. Therefore, many designs will have four or more connections at the extremities of the actuator to avoid rotations.



Basic MEMS Actuators, Fig. 4 Electric field of a comb capacitor

Analysis of a Fixed-Guided Beam Flexure

The linear response of fixed-guided configuration shown in Fig. 7 can be studied using the moment curvature relationship:

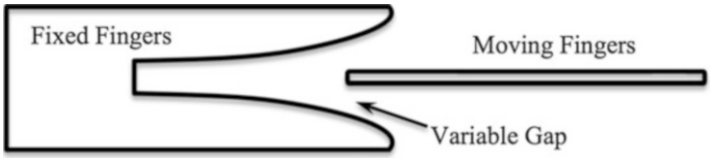
$$\frac{d^2v}{dx^2} = \frac{\mathrm{M}}{\mathrm{EI}} \tag{5}$$

where E is the elastic modulus, v is the deflection, and I is the second moment of inertia. The second moment of inertia for many MEMS structures is that of a rectangular cross section $\frac{hw^3}{12}$ where h and w are the height and width of the structure, respectively. The rectangular cross section is a result of the fabrication techniques used to make these devices (see "Fabrication Methods" section). Solving the ordinary differential equation, Eq. 5, one finds that:

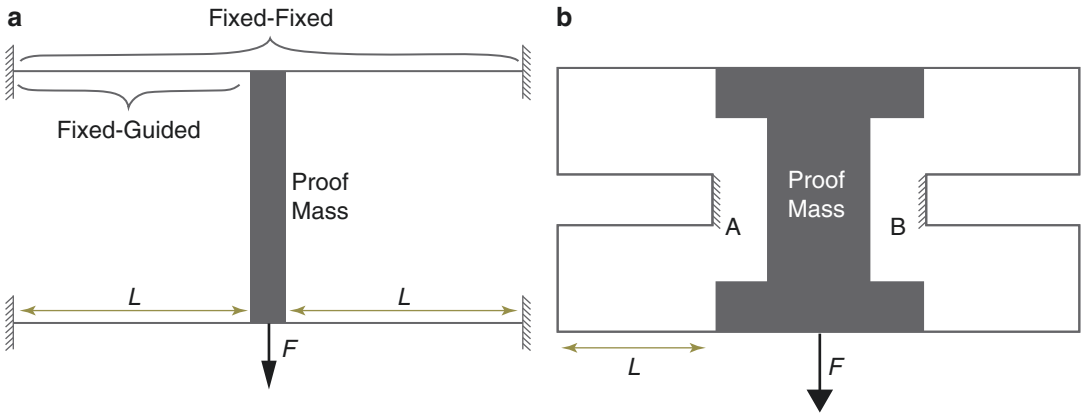
$$F = \frac{12\,\mathrm{EI}}{L^3}\delta \tag{6}$$

where δ is the maximum deflection of the structure. Note that this solution is for a single fixed-guided beam (one half of a fixed-fixed beam). In order to find the solution for multiple fixed-guided beams as in Fig. 6a, springs constants are added in parallel, as appropriate. For example, for the structure shown in Fig. 6a, the spring constant would be 48 EI/L^3.

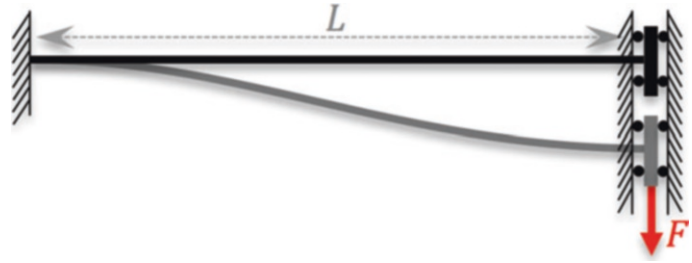
For small deflections, $\delta < w/4$, this relationship is sufficient for matching the true center line displacement of the fixed-guided beam with better



Basic MEMS Actuators, Fig. 5 Varying the shapes of the fixed and moving fingers of the comb capacitor results in modified force-displacement relationships



Basic MEMS Actuators, Fig. 6 Typical configurations for springs in MEMS actuators: (**a**) fixed-fixed or fixed-guided beam configuration and (**b**) folded-beam flexure configuration



Basic MEMS Actuators, Fig. 7 The fixed-guided beam fixed at one end (*left*) and guided on the other (*right*)

than 5 % accuracy. However, for larger deflections, stretching of the beam must be considered. Stretching of the beam induces an axial force F_x, which in turn results in an applied moment. This changes the moment in the ordinary differential equation to be solved. In order to solve this new equation, small displacement assumptions must be used in conjunction with Hooke's Law. These assumptions result in the following equations:

$$\delta = 2\left(\frac{2I}{A}\right)^{\frac{1}{2}}(u - \tanh u) \times \left(\frac{3}{2} - \frac{1}{2}\tanh^2 u - \frac{3}{2}\frac{\tanh u}{u}\right)^{-\frac{1}{2}} \tag{7}$$

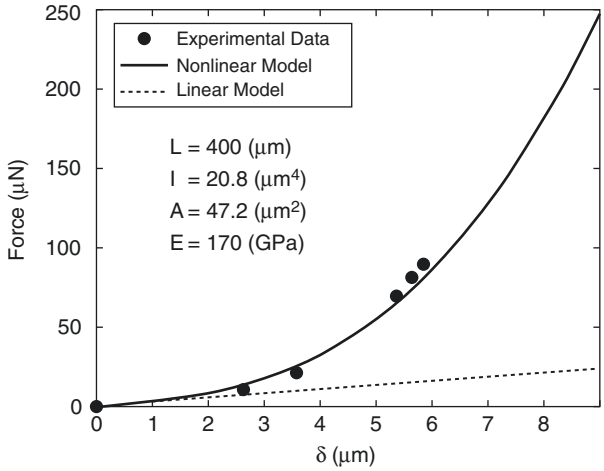
$$F = \frac{8EI}{L^3}\left(\frac{2I}{A}\right)^{\frac{1}{2}} u^3 \left(\frac{3}{2} - \frac{1}{2}\tanh^2 u - \frac{3}{2}\frac{\tanh u}{u}\right)^{-\frac{1}{2}} \tag{8}$$

The parameter $u = \sqrt{\frac{F_x L^2}{4EI}}$ in Eqs. 7 and 8. No analytical solution is possible, because F_x is not known a priori. In practice, δ is commonly found in experiments. With δ, Eq. 7 can be solved for u. Next, F is found by inserting the previously determined value of u into Eq. 8. A more detailed solution is located in Frisch-Fay [17].

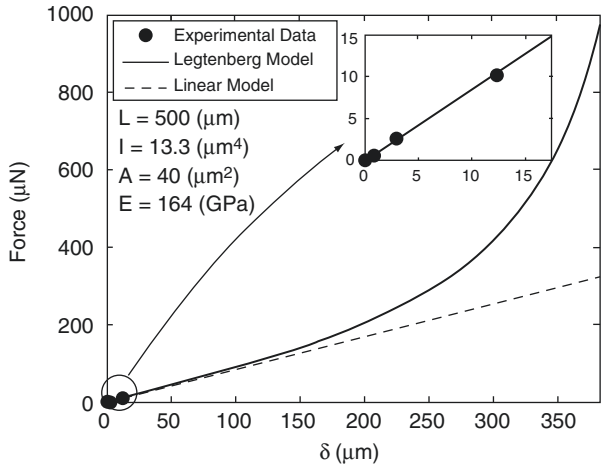
Figure 8 plots both the linear and nonlinear solution for the device shown in Fig. 1, which is schematically depicted in Fig. 6a. In addition to the derived solutions, experimental data are included on the plot. Data are taken by hanging calibrated weights from the device as is detailed in [18]. The width of the fixed-guided beams is 2 μm. Note that the nonlinear model diverges at about one-quarter of the value, that is, 500 nm.

Analysis of a Folded-Beam Flexure

The folded-beam flexure, Fig. 6b, is a variation of the fixed-guided beam configuration. Because the beams are folded, this structure experiences less axial stretch than the fixed-guided structure of Fig. 6a. However, for small deformations, the behavior is approximately the same and Eq. 6 can be used. Because there is less axial stretching in the folded-beam flexure, the linear range is



Basic MEMS Actuators, Fig. 8 Force versus displacement response for a fixed-guided beam that compares the linear and nonlinear models developed. Parameters used are for the actuator in Fig. 1. Experimental data are given as found using the method described in [18]



Basic MEMS Actuators, Fig. 9 Force versus displacement response for a folded-beam flexure that compares the linear and nonlinear models developed. Parameters used are for the actuator in Fig. 16. Experimental data are given as found using the method described in [18]

much larger for folded-beam flexures than for fixed-guided beam flexures. In order to find the solution for folded-beam flexure as in Fig. 6b, springs constants are added in series and parallel, as appropriate. For example, for the structure shown in Fig. 6b, the spring constant would be $24EI/L^3$.

For even larger deflections, elastica theory must be applied. Along the y-direction, the displacement and force are [19]:

$$\frac{L}{2}\sqrt{\frac{F_y}{EI}} = \int_{\phi_1}^{\frac{\pi}{2}} \frac{d\phi}{\sqrt{1-p^2\sin^2\phi}} \tag{9}$$

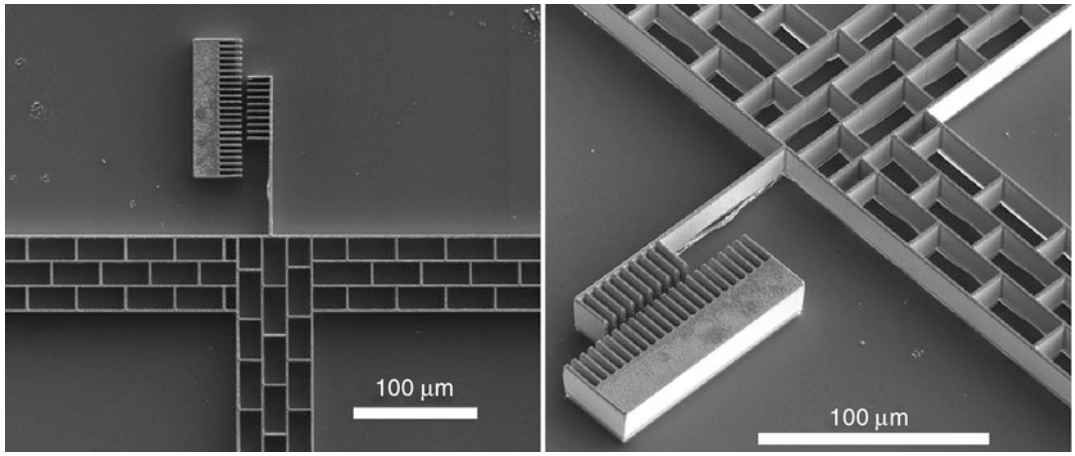
$$\delta = \sqrt{\frac{EI}{F_y}}\int_{\phi_1}^{\frac{\pi}{2}} \frac{2p^2\sin^2\phi - 1}{\sqrt{1-p^2\sin^2\phi}}\,d\phi \tag{10}$$

where:

$$\phi_1 = \sin^{-1}\left(\frac{1}{\sqrt{2p^2}}\right) \tag{11}$$

The solution method involves choosing p and finding ϕ_1 from Eq. 11. Then, the elliptic integrals of Eqs. 9 and 10, each of which depends only on p, are calculated.

Figure 9 plots both the linear and nonlinear solution for the device shown in Fig. 16. In addition to the derived solutions, experimental data are included on the plot. Data are taken by hanging calibrated weights from the device as is detailed in [18]. The width of the folded-beam flexure's



Basic MEMS Actuators, Fig. 10 SEM micrographs of a vernier for the measurement of displacement of a Basic MEMS Actuator

beams is 2 µm. Note that the nonlinear model diverges at almost 55 times the beam's thickness!

Motion Quantification

Many different schemes have been proposed for the quantification of the displacement of the MEMS actuators. The most common types include measuring capacitance, resistance, or interpreting an optical signal. Capacitance measurements are especially common when the method of actuation is by capacitors as well. By utilizing a high-frequency voltage on the capacitors, one can measure the impedance. Equation 1 is for the capacitance due to parallel plate capacitors and Eq. 3 is for comb capacitors. Both equations contain a term that relates the displacement to the capacitance of the MEMS actuator. Piezoresistance is also a common method because of the ease of measurement. Limitations of either capacitance or piezoresistance measurements include the sensitivity of the instrument and the ability to isolate the actuator from noise. Optical measurements can be broken into two main categories: inteferometric and vernier. Inteferometric techniques can make very sensitive measurements of displacement and (depending on the technique) may be limited only by thermal vibrations. A vernier is a simple method and can be co-fabricated into the device (see Fig. 10).

In order to measure the displacements of the springs, two sets of scales are incorporated in the design of the MEMS actuator. These are a fixed "main scale" that is attached to the substrate and a moving "vernier scale" that is integrated into the MEMS actuator (Figs. 10 and 11).

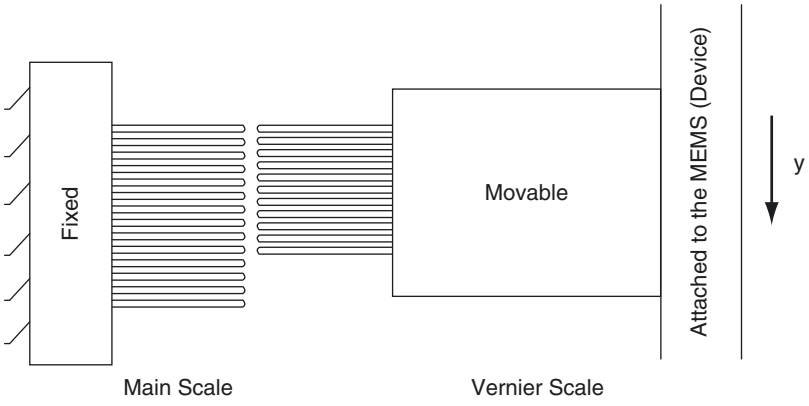
The vernier is an auxiliary scale, whose graduations are of different spacing from those of the main scale, but that bear a simple relation to them. The vernier scale of Fig. 11 has ten divisions that correspond in length to nine divisions on the main scale. Each vernier division is therefore shorter than a main-scale division by 1/10 of a main-scale division. Main-scale divisions are 5 µm apart. The Least Count (LC), which is defined as the smallest value that can be read on the vernier scale, is therefore:

$$LC = 5 \times \frac{1}{10} = 0.5\,\mu\text{m} \tag{12}$$

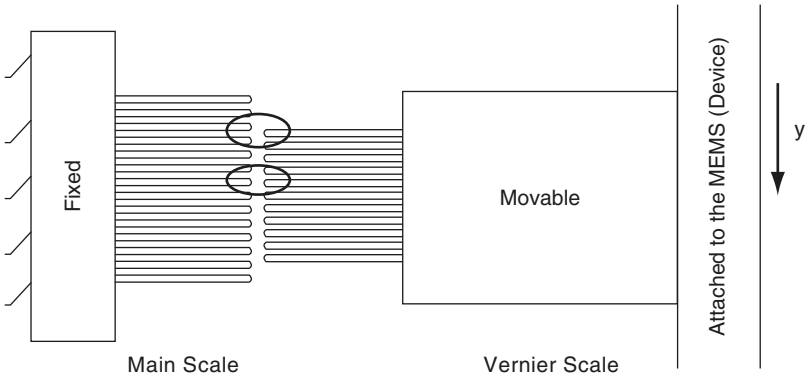
For example, in Fig. 12, the zero of the vernier scale is moved past two complete divisions of the main scale, and the fifth division of the vernier scale is aligned perfectly to the same fifth main-scale reading.

Therefore, the reading from the above setting would be:

- Main-scale divisions = 2
- Main-scale spacing = 5 µm



Basic MEMS Actuators, Fig. 11 Schematic of a vernier for measuring the displacement of a Basic MEMS Actuator



Basic MEMS Actuators, Fig. 12 Schematic of scale system, depicting a deflection of 12.0 μm

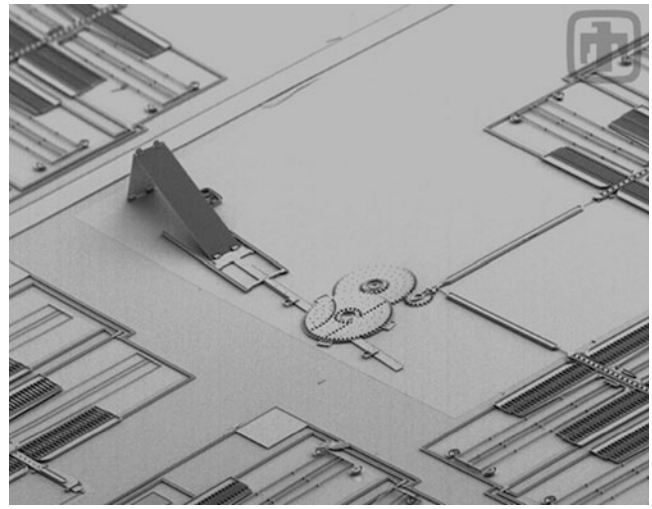
- Total main-scale reading = 2 × 5 μm = 10 μm
- Vernier-scale division that coincides with main scale = 4
- Least count of the vernier scale = 0.5 μm
- Total vernier-scale reading = 4 × 0.5 μm = 2.0 μm
- Total = main-scale reading + vernier-scale reading
- Total = 10 + 2.0 μm = 12.0 μm

Fabrication Methods

As mentioned previously, there are three main methods for fabrication of MEMS actuators: Surface Micromachining, SOI, and SCREAM [14]. It is important to emphasize that many other processes exist and new processes are being developed constantly. Modifications are typically introduced to accommodate for the actuators' ultimate application or for the type of energy conversion method that the actuator will be utilizing.

Surface micromachining is a process built by alternating the deposition and patterning of sacrificial layers and structural layers of material. Typically, the sacrificial layers are a type of silicate glass (Phospho-Silicate-Glass (PSG) or SiO_2). This layer is commonly referred to as the sacrificial oxide (sac-ox). The structural layers are often made of polycrystalline Si (poly-Si). After the deposition of each layer, that layer is patterned using a photolithographic process to define either holes through the sacrificial layer or physical features of the device in the case of a structural layer.

Figure 13 details a simple surface micromachining process for a cantilever beam. Each step (a–f) is a cross-sectional view for the structure. This is a common method of representing a microfabricated structure's process flow. This method typically revolves around an important feature of the final structure. Note that

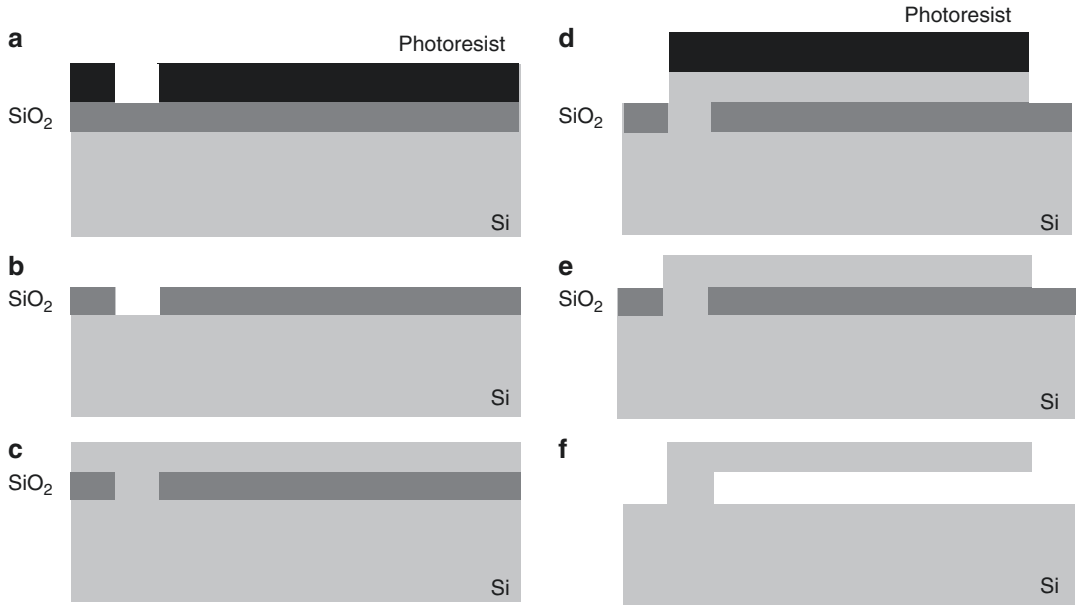


Basic MEMS Actuators, Fig. 13 A surface-micromachined set of Basic MEMS Actuators with comb capacitors and folded-beam flexure used to drive a gear train to raise a micromirror. Image courtesy of Sandia National Laboratories, SUMMiT™ Technologies, www.mems.sandia.gov

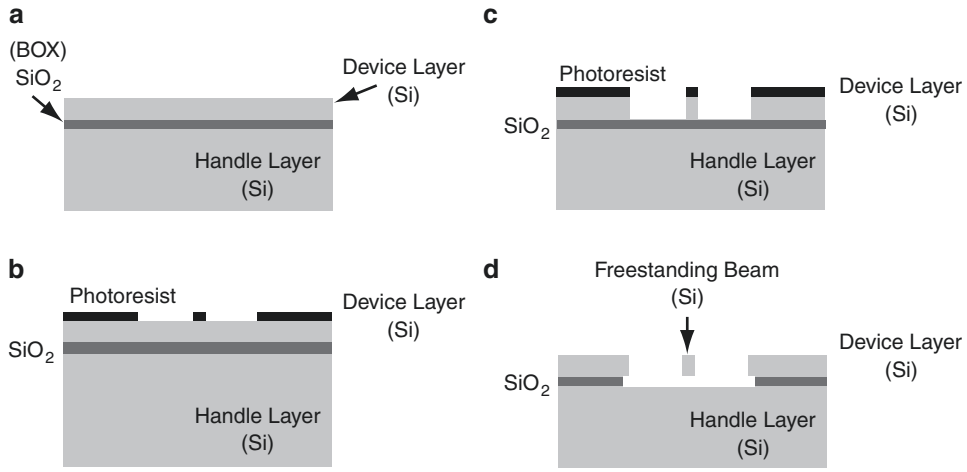
due to the nature of this process, numerous alternating layers of sac-ox and structural layers can be built up to form devices with multiple layers of functionality. One of the more common processes used for surface-micromachined MEMS is the SUMMiT process [20]. At the writing of this entry, the current process is entitled SUMMiT™ V, where V stands for five structural layers of poly-Si. An example of a SUMMiT surface-micromachined device is shown in Fig. 14. A device is displayed with a pair of orthogonal comb capacitors that turn gears and raise a micromirror out-of-plane.

SOI surface-micromachined devices are similar to the surface micromachining process previously described except that the process begins with an SOI wafer. SOI wafers are comprised of three layers: a handle layer, a buried oxide layer (BOX), and a device layer. The handle layer acts as a substrate for the BOX and device layer. BOX layers are the sacrificial layer to be removed by an oxide etch at the end of the process, whereas the device layer is the structural layer of Si for the actual MEMS. The device layer is a piece of single crystal Si, not poly-Si as in standard surface micromachining. Due to the nature of the SOI wafer, SOI surface micromachining is used to make single structural layer devices. The general process flow for an SOI surface-micromachined process is shown in Fig. 15. The freestanding beam in the Fig. 15d would be for any part of a MEMS actuator that is capable of movement. These freestanding beams however must be anchored to the substrate. This accomplished by having structures with larger cross sections that are not fully undercut by the HF etching process. The left and right side of the device layer in Fig. 15d are a partial view of cross section where the SiO_2 has not been fully undercut, thereby allowing for a physical connection between the device layer and handle layer. Note that the connections of the device to ground can be isolated electrically from one another. This is an important feature for placing biases on comb structures such as the one shown in Fig. 16. Although SOI surface micromachining and standard surface micromachining are similar, the device layer in the SOI process can have a broader range of thickness varying from 100's of nm to several 10's of μm. This is important because (1) layer thickness increases the actuator force and (2) a thick layer can make the out-of-plane stiffness of folded-beam suspension much greater, thereby minimizing unintended displacements.

The SCREAM process is a bulk processing method that utilizes a bulk piece of single crystal Si – a Si wafer. Typically, this process utilizes one mask (one photolithography step) as does SOI surface micromachining. In order to create free-standing structures, SCREAM utilizes the selectivity of etchants to certain deposited films and isotropic versus anisotropic etching.



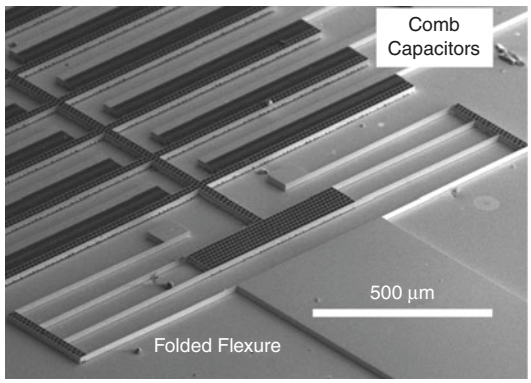
Basic MEMS Actuators, Fig. 14 Surface micromachining of a freestanding microcantilever: (**a**) deposition of a sacrificial SiO_2 layer on Si followed by photolithographic patterning, (**b**) removal of photo resist after etching, (**c**) poly-Si deposition, (**d**) photolithographic patterning and etching of Si to expose the sacrificial oxide layer, (**e**) microcantilever before release, and (**f**) etching the sacrificial oxide layer thus releasing the microcantilever



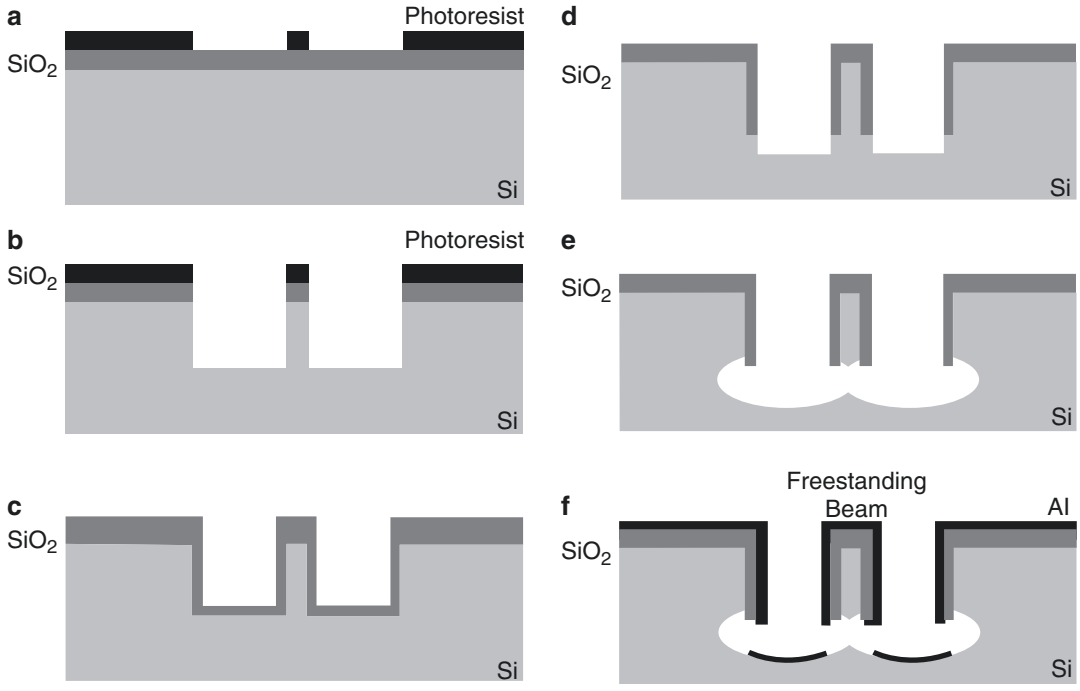
Basic MEMS Actuators, Fig. 15 Processing steps for the SOI surface micromachining process: (**a**) SOI wafer anatomy, (**b**) photolithographic patterning of device layer, (**c**) deep reactive ion etching of device layer, and (**d**) release of device by oxide etch

Figure 17 demonstrates the SCREAM process [21]. A Si wafer is the base substrate. SiO_2 is deposited on the surface of the Si. Photoresist is then patterned to define the 2D layout of the microactuator (Fig. 17a). Narrow features, approximately 1 μm, will be released (freed), while wider structures, approximately 100 μm or larger, will remain fixed to the substrate during the

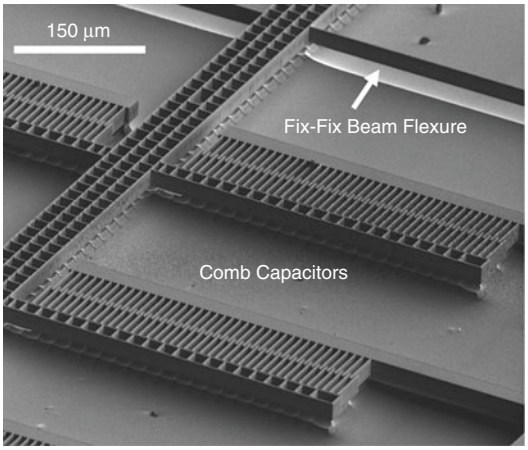
release process. After patterning, the Si is etched anisotropically (vertically) typically via a deep reactive ion etch (DRIE) (Fig. 17b). The photoresist is then removed and a conformal layer of SiO_2 is deposited (Fig. 17c). This step is important to the process in that it places an electrical insulator on the sidewalls of the structures. The SiO_2 on the floor of the wafer is then back etched to expose the Si. A subsequent Si etch then isotropically removes Si from the floor (Fig. 17d). This etch is allowed to continue to the point where the two etch fronts from either side of the narrower structures are allowed to merge, thereby releasing the narrow structures (Fig. 17e). The wider adjacent structures are also undercut, but not released. In the final step, a metal is deposited (Fig. 17f). Metal is allowed to coat the top horizontal surface and vertical surfaces of the device. Depending on the layout of the device, the islands on the left and right of the freestanding beam can be electrically isolated from the beam and from one another, connect to one or the other, or connect to both. The isotropic etch of Si in step (e) allowed for the SiO_2 layer to shield the metal (Al) from coming into contact with the Si. If the metal of each structure had touched the Si, then the entire device would be shorted and no potential could be set up for actuation of the device. Figure 18 is an SEM



Basic MEMS Actuators, Fig. 16 An SOI surface-micromachined device with comb capacitors and folded-beam flexures



Basic MEMS Actuators, Fig. 17 The SCREAM process: (**a**) Deposition of SiO_2 on a silicon wafer followed by photolithographic patterning, (**b**) deep reactive ion etching (DRIE) of Si, (**c**) conformal deposition of SiO_2, (**d**) etching of SiO_2 to expose the floors adjacent to the structure to be released, (**e**) isotropic etching of Si to release the freestanding structure, and (**f**) metallization



Basic MEMS Actuators, Fig. 18 Example of a Basic MEMS Actuator fabricated using the SCREAM process with comb capacitors and fixed-fixed beams. Note the scalloped floor beneath the actuators. This is indicative of isotropic Si etch in step (e) in Fig. 17

micrograph of a MEMS actuator fabricated with the SCREAM process.

Summary

This encyclopedia entry covered the basics of MEMS actuators. The discussion was broken into four parts covering: energy conversion mechanisms for generation of force, compliant structures, motion quantification, and fabrication techniques. The section covering energy conversion mechanisms discussed the common mechanisms used in MEMS actuators and went into detail on electrostatic actuation. In particular, the salient features of parallel plate and comb capacitors were presented. Converted forces are used to move MEMS actuators to perform a host of applications. This motion occurs in controlled manner via a compliant structure. Two of the more common compliant structures used for MEMS actuators are fixed-guided beams and the folded flexure structure. Both the linear and nonlinear force versus displacement relationships are given for both of these common structures. Motion of the actuator is quantified in many different ways. Use of a simple co-fabricated vernier is detailed in this entry. Finally, the three main processes for fabrication of MEMS actuators were detailed. These included: Surface Micromachining, SOI, and the SCREAM processes.

Cross-References

- ▶ Micropumps
- ▶ Thermal Actuators

References

1. Wood, J.: Personal communication (2011)
2. de Boer, M.P., Luck, D.L., Ashurst, W.R., Maboudian, R., Corwin, A.D., Walraven, J.A., Redmond, J.M.: High-performance surface-micromachined inchworm actuator. J. Microelectromech. Syst. **13**(1), 63–74 (2004)
3. Akiyama, T., Collard, D., Fujita, H.: Scratch drive actuator with mechanical links for self-assembly of three-dimensional MEMS. J. Microelectromech. Syst. **6**(1), 10–17 (1997)
4. Deladi, S., Sarajlic, E., Kuijpers, A.A., Krijnen, G.J. M., Elwenspoek, M.C.: Bidirectional electrostatic linear shuffle motor with two degrees of freedom. In: Proceedings of the IEEE MEMS Conference, Miami Beach, Florida, USA (2005)
5. Seidemann, V., Bütefisch, S., Büttgenbach, S.: Fabrication and investigation of in-plane compliant SU8 structures for MEMS and their application to micro valves and micro grippers. Sens. Actuator A Phys. **97–98**, 457–461 (2002)
6. Basantkumar, R.R., Hills Stadler, B.J., Robbins, W.P., Summers, E.M.: Integration of thin-film galfenol with MEMS cantilevers for magnetic actuation. IEEE Trans. Magn. **42**(10), 3102–3104 (2006)
7. Sulfridge, M., Saif, T., Miller, N., O'Hara, K.: Optical actuation of a bistable MEMS. J. Microelectromech. Syst. **11**(5), 574–583 (2002)
8. Kommepalli, H.K.R., Yu, H.G., Muhlstein, C.L., Trolier-McKinstry, S., Rahn, C.D., Tadigadapa, S.A.: Design, fabrication, and performance of a piezoelectric uniflex microactuator. J. Microelectromech. Syst. **18**(3), 616–625 (2009)
9. Gill, J.J., Ho, K., Carman, G.P.: Three-dimensional thin-film shape memory alloy microactuator with two-way effect. J. Microelectromech. Syst. **11**(1), 68–77 (2002)
10. Que, L., Park, J.-S., Gianchandani, Y.B.: Bent-beam electrothermal actuators – part I: single beam and cascaded devices. J. Microelectromech. Syst. **10**(2), 247–254 (2001)
11. Israelachvili, J.N.: Intermolecular and Surface Forces: With Applications to Colloidal and Biological Systems (Colloid Science), 2nd edn. Academic, London (2006)

12. Naraghi, M., Chasiotis, I.: Optimization of comb-driven devices for mechanical testing of polymeric nanofibers subjected to large deformations. J. Microelectromech. Syst. **18**(5), 1032–1046 (2009)
13. Trimmer, W.S.N.: Micro robots and micro mechanical systems. Sens. Actuator. **19**, 267–287 (1989)
14. Madou, M.J.: Fundamentals of Microfabrication: The Science of Miniaturization, 2nd edn. CRC Press, Boca Raton (2002)
15. Tang, W.C., Nguyen, T.-C.H., Howe, R.T.: Laterally driven polysilicon resonant microstructures. Sens. Actuator. **20**(1–2), 25–32 (1989)
16. Ye, W., Mukherjee, S., MacDonald, N.C.: Optimal shape design of an electrostatic comb drive in microelectromechanical systems. J. Microelectromech. Syst. **7**(1), 16–26 (1998)
17. Frisch-Fay, R.: Flexible Bars. Butterworths Scientific, Washington, DC (1962)
18. Abbas, K., Leseman, Z.C., Mackin, T.J.: A traceable calibration procedure for MEMS-based load cells. Int. J. Mech. Mater. Des. **4**(4), 383–389 (2008)
19. Legtenberg, R., Groeneveld, A.W., Elwenspoek, M.: Comb-drive actuators for large displacements. J. Micromech. Microeng. **6**, 320–329 (1996)
20. Sniegowski, J.J., de Boer, M.P.: IC-compatible polysilicon surface micromachining. Annu. Rev. Mater. Sci. **30**, 299–333 (2000)
21. Shaw, K.A., Zhang, Z.L., MacDonald, N.C.: SCREAM I: a single mask, single-crystal silicon, reactive ion etching process for microelectromechanical structures. Sens. Actuator A Phys. **A40**(1), 63–70 (1994)

Batteries

▶ Nanomaterials for Electrical Energy Storage Devices

Bendable Electronics

▶ Flexible Electronics

Bending Strength

▶ Nanomechanical Properties of Nanostructures

Bilayer Graphene

▶ Graphene

Binary Systems

▶ Nanoalloy Simulation

Bioaccessibility/Bioavailability

▶ Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms

Bioactive Materials

▶ Synthesis of Functional Materials for Bone Regeneration

Bioadhesion

Manuel L. B. Palacio and Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Cell adhesion; Mucoadhesion

Definition

Bioadhesion refers to the phenomenon where natural and synthetic materials adhere to biological surfaces. This adherence may or may not be associated with the use of a bioadhesive. Bioadhesion also refers to the incorporation of the biomaterial in the body, which is manifested by the formation of a biofilm on the biomaterial. A specific type of bioadhesion is referred to as mucoadhesion, where a mucous gel layer is formed on the surface of the biological surface during the adhesion process.

Occurrence

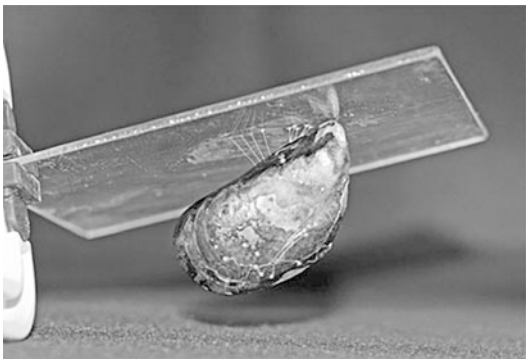
Depending on the chemical composition of the biomaterial and the function of the biological surface involved, bioadhesion may be desirable or undesirable. A few examples of the occurrence of bioadhesion interactions are described as follows.

Synthetic and Natural Biomaterial Interactions

The implantation of biomaterials to the body (in tissue engineering, orthopedics, or dental prostheses, among others) is an example of a situation where bioadhesion is desired. In this case, the bioadhesion between the biomaterial and the target organs lead to the formation of a biofilm (protein layer), which, in turn, facilitates the integration of the implant to the body. Studying the fundamental mechanisms that influence protein adhesion on biomaterials is of interest in the context of the "Vroman effect," which involves a series of adsorption and displacement of proteins on biomaterial surfaces. The proteins on the biomaterial surface interact with cells and influence their behavior. The formation of protein-containing adhesive complexes known as focal adhesions (or cell-matrix adhesions) determines cell migration, cell cycle progression, and eventual differentiation. The proteins involved in focal adhesions are sensitive to the properties of the underlying substrate, such as chemical composition and hydrophobicity [1, 2].

On the other hand, biofilm formation may not be considered as beneficial in some cases. For ocular (contact) lens applications, the formation of a biofilm on the surface of the polymeric lens leads to the accumulation of inflammatory cells, and is therefore undesired [3].

An example of bioadhesion between natural biomaterial surfaces involves lectins, which are proteins or glycoprotein complexes that can bind to polysaccharides (sugars). Lectins are found in both animals and plants. One of the common animal lectins is the calcium-dependent binding lectins (also referred to as "C-type"). One of these C-type lectins, the asialoglycoprotein lectins, is specific to liver cells and is involved in animal biological function [4]. Meanwhile, plant lectins are abundant and can be found in common plants such as tomato, soybeans, and kidney beans. Plant lectins are commonly regarded as toxins when they bind to animal cells. [4].



Bioadhesion, Fig. 1 Adhesion of a mussel on a glass slide, where multiple threads composed of the mussel adhesive protein were secreted by the animal to attach itself to the glass surface [5]

Mucoadhesion

The adherence of materials onto a mucous gel layer of a biological surface is referred to as mucoadhesion. A well-known example of mucoadhesion is the attachment of mussels to underwater surfaces (such as corals or ship surfaces). The animal's foot extends from its shell to secrete a pad of the mussel adhesive protein (MAP) on the underwater surface. The foot then retreats to the shell interior, leaving a thread that secures the mussel to the pad. As shown in Fig. 1, this process is repeated multiple times [5]. The MAP has been used as the model for the development of mucoadhesive materials.

Another important example of mucoadhesion pertains to drug-delivery systems, where the drug is carried in a bioadhesive system, which then adheres to the mucous gel layer located on the surface of the biological membrane. Examples of polymers used as bioadhesives include chitosan, poly(vinyl alcohol), poly(vinyl pyrrolidone), agarose, as well as the cellulose and poly(acrylic acid)-type polymers [4]. The mucous gel layer is composed mostly of water (approx. 95 % by weight), with the rest being the glycoprotein class known as mucins. The mucosal routes of

drug delivery include the eye, mouth, nose, vagina, rectum, and the gastrointestinal (GI) tract [6].

Medical Adhesives

Adhesives are used to close skin incisions, seal and strengthen tissues, stop the leakage of blood and other body fluids, and serve as an adjunct to staples or sutures. These adhesives can be used in various surgical operations such as cardiovascular, pulmonary, and gastrointestinal surgeries, among others. Medical adhesives could either be synthetic (such as cyanoacrylates) or naturally derived (such as fibrin sealants and bovine serum albumin-glutaraldehyde mixtures) [7]. In the case of cyanoacrylates, bonding occurs due to polymerization of the adhesive in the presence of hydroxyl groups in the environment. For the bovine serum albumin-glutaraldehyde glues, bonding takes place between the available lysine groups in the albumin and tissue surfaces and the aldehyde group in glutaraldehyde.

Dental Prostheses

Resin-based composites and polyacid-modified resin-based composites (compomers) are used to restore teeth damaged by cavities, fractures, erosion or aging. Moreover, adhesives are also used to repair fractured porcelain and composite restoratives. These adhesives are applied either to the enamel or to the dentin layer of the tooth requiring the restoration. Acrylic-based resins are used to restore human enamel. The enamel is pre-etched with acid (usually phosphoric acid) in order to create pores and increase surface wettability. The compositions of the enamel and dentin layers are widely different. Relative to enamel, dentin has much higher organic content, is less homogeneous, and more difficult to bond to restorative materials. As a consequence, the composition of adhesives intended for dentin is different from enamel adhesives. Dentin adhesives contain acrylic-based resins, such as bisphenol A-glycidyl methacrylate (BisGMA) and hydroxyethyl methacrylate (HEMA), but other components are present, such as a primer and a bonding agent, which are necessary to enhance bonding to dentin [8].

Key Research Findings

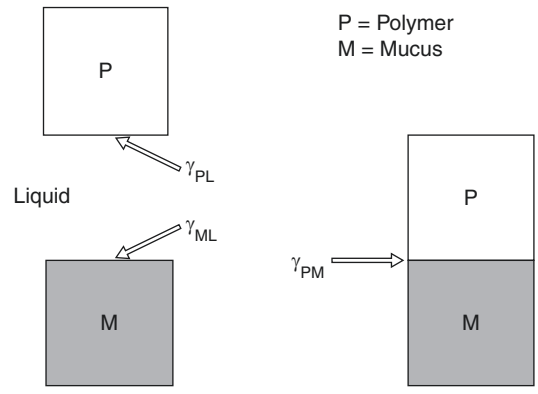
Bioadhesion Mechanisms

Interface Energetics

In systems exhibiting bioadhesion, the liquid environment influences the spreading of one material phase over another. Consider the case of a polymer surface adhering to the mucous gel layer on a biological membrane immersed in liquid medium (e.g., drug-delivery systems). Figure 2 is a schematic illustrating the interfacial energy components that should be considered in evaluating the thermodynamic work of adhesion of this system. During adhesion, a unit interface between the polymer and the liquid and between the mucus and the liquid vanishes, while an interface between the polymer and mucus forms. The thermodynamic work of adhesion (W_{PM}^{adh}) between the polymer and mucus is then defined as follows [6, 9]:

$$W_{PM}^{adh} = \gamma_{PM} - (\gamma_{PL} + \gamma_{ML}) \qquad (1)$$

where γ represent surface energies and the subscripts P, M, and L are for polymer, mucus, and liquid, respectively. A positive value for W_{PM}^{adh} is regarded to be a necessary condition to achieve successful bonding between the two surfaces.



Bioadhesion, Fig. 2 Schematic of the relevant interfacial energies (γ) that determine the formation of a bioadhesive bond, as illustrated for the adhesion between a polymer (*P*) and the mucus layer (*M*) on the biological membrane in the presence of a liquid medium (*L*) (Adapted from [6])

Chemical Interactions

Various interactions between two chemically active surfaces (i.e., not inert) facilitate the bioadhesion process. Strong adhesion can occur if the two surfaces are capable of forming either covalent, ionic or metallic bonds. At the same time, weaker interaction forces, such as polar (dipole–dipole), hydrogen bonding, or van der Waals interactions (induced-dipoles), also aid in bonding the two surfaces [4, 9].

The case of mussel adhesion on underwater surfaces is an example of how the chemical composition of the contacting surfaces affects the adhesion mechanism. The initial interaction between the mussel and the underwater surface involves the removal of weak boundary layers (mostly water). If the underwater surface is nonpolar, the water boundary layer interacts through weak dispersive forces. Since the mussel adhesive protein (MAP) is larger than a water molecule, the protein experiences greater dispersive interactions with the nonpolar surface relative to water, leading to the displacement of the water boundary layer, and adhesion of the protein. In the case of polar underwater surfaces, the water boundary layer cannot be easily displaced. In this case, the MAP uses its hydrophilic amino acid side chains, which contain aminoalkyl, hydroxyalkyl, and phenolic groups (such as 3,4-dihydroxyphenylalanine or Dopa), all of which are capable of forming strong hydrogen bonds. Hence, the MAP is able to displace water and the mussel is able to adhere to polar underwater surfaces as well [10]. The unique chemical composition and properties of the mussel adhesive proteins have led to the development of synthetic analogs for potential use as mucoadhesives for drug-delivery systems.

Mechanical Effects

While interfacial contact and chemical bonding interactions are needed for the initial stages of bioadhesion, the interpenetration or interdiffusion between the molecules of the two contacting surfaces will maintain the adhesive bond. If a polymer is one of the contacting systems involved in the bioadhesion process, the interpenetration process involves the mobility of the individual chains and its entanglement in the opposing biological membrane. A related concept is the swelling capacity of the polymer, which is the ratio of the wet to dry weights. A high swelling capacity for a given polymer implies that it has greater chain mobility, and a higher tendency toward interpenetration [6].

For bioadhesion in dental prostheses, mechanical interlocking or interpenetration is the primary adhesive mechanism. When bonding to enamel, surface etching is performed to create micropores on the surface that will increase the penetration of the restorative material. Dentin, on the other hand, contains tubules that radiate from the pulp. These tubular structures facilitate the penetration of resin monomers into the dentin and their retention once the resin has been polymerized [11].

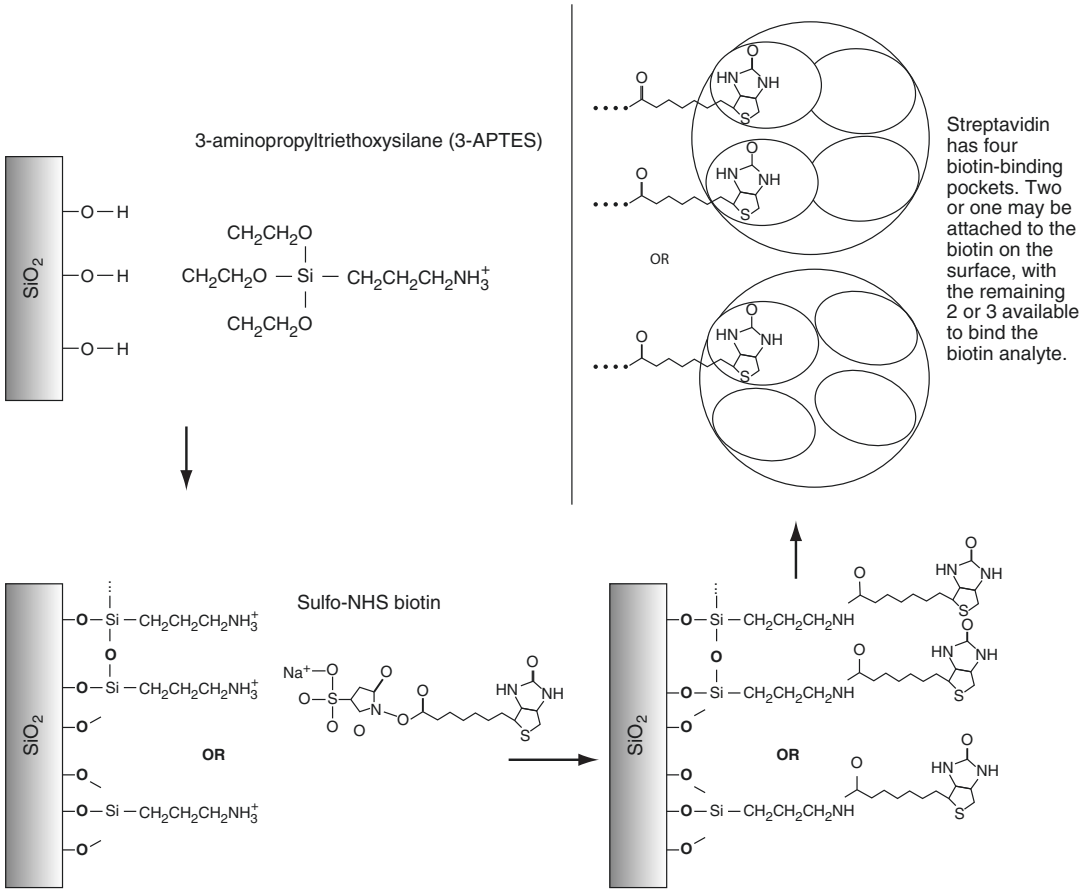
Interface Optimization

The contact between the two surfaces in contact during the bioadhesion process may be optimized using various physical, chemical, and mechanical methods.

Physical Processes

Plasma processing is a widely used technique for surface cleaning and modification of the surface chemical composition. The plasma, which is generated by applying an electric field to a low-pressure gas in vacuum, produces reactive species such as ions and free radicals. For surface cleaning purposes, the plasma breaks most organic bonds, such as C–H, C–C, C = C, C–O and C–N. In addition, if oxygen plasma is used, another cleaning mode is possible. The activated oxygen species (O•, O^+, O−, O_2^+, O_2−, O_3, among others) can combine with organic compounds to form H_2O, CO, CO_2, and low molecular weight hydrocarbons, thereby removing organic contaminants on the surface [12].

Surface modification through plasma processing can render the surface more hydrophilic or hydrophobic. If the gas used is oxygen or carbon dioxide, polar groups are introduced to the surface (such as –C–OH, −C = O and –COOH), enhancing surface hydrophilicity. The presence of more polar surface groups is valuable in enhancing adhesion, e.g., between proteins and



Bioadhesion, Fig. 3 Schematic showing the steps in the functionalization of silica surface with the protein streptavidin (Adapted from [13])

polymeric biomaterials used in tissue engineering applications. On the other hand, the use of fluorinated gases such as CF_4 in the plasma results in the introduction of a low surface energy polytetrafluoroethylene (PTFE)-like structure, enhancing surface hydrophobicity. This approach has been used to prevent the undesirable protein adhesion and the ensuing generation of inflammatory cells on poly(methyl methacrylate) contact lenses [3].

Chemical Functionalization

Specific functional groups can be attached to surfaces using chemical synthesis techniques to construct model systems for use in adhesion studies. An example is shown in Fig. 3, illustrating how a silica surface is chemically functionalized with the protein streptavidin [13]. In this illustration, the functionalization process consists of multiple steps, starting with the addition of the aminosilane 3-aminopropyltriethoxysilane (APTES) in order to introduce amine groups to the silica surface. This is then followed by reacting the surface with sulfo-*N*-hydroxysuccin-imido-biotin (abbreviated as biotin), which forms a covalent bond with the amine group of the APTES. In the final step, the protein streptavidin will bind to biotin. Streptavidin has a high affinity to biotin, and this interaction is regarded as one of the strongest non-covalent bonds known.

Mechanical Texturing

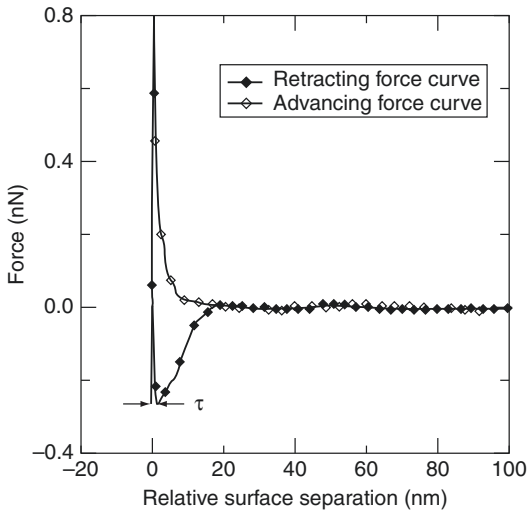
Surface modification through mechanical texturing is a simple and cost-effective method to roughen biomaterials and create anchor points

for enhanced adhesion. In this method, abrasive particles (such as silicon carbide or alumina) impinge on the biomaterial, and the resulting impact increases its surface roughness [14]. This method can be used on orthopedic implant materials such as titanium or cobalt/chrome.

Characterization Methods

Nanoscale

Fundamental studies of bioadhesion at the nanoscale are conducted with atomic force microscopy (AFM). Aside from its high-resolution surface imaging capability, the AFM can measure forces in the pico- to nanonewton range, allowing for the detection of single molecule interactions [15, 16]. In AFM, force-distance curves for the contact between the probe and the sample surface can be obtained. An example is shown in Fig. 4, which corresponds to the interaction between the protein streptavidin and its ligand biotin while immersed in phosphate buffered saline (PBS) [17]. In this example, the AFM probe consisted of a cantilever with a glass microsphere functionalized with biotin on its end, where the biotin is covalently bonded to bovine serum albumin (BSA). The streptavidin was deposited onto a mica surface.



Bioadhesion, Fig. 4 Atomic force microscopy (*AFM*) force-distance curve between biotinylated bovine serum albumin and streptavidin surface imaged in phosphate buffered saline (*PBS*) medium (Adapted from [17])

The force-distance curve consists of two segments, namely, the advancing curve, which shows how the probe approaches the surface, and the retracting curve, which illustrates how the probe detaches from the surface. In the retracting force curve, a distinct snap-off point is observed, which corresponds to the force necessary to separate the tip from the sample surface. This is the measured adhesive force.

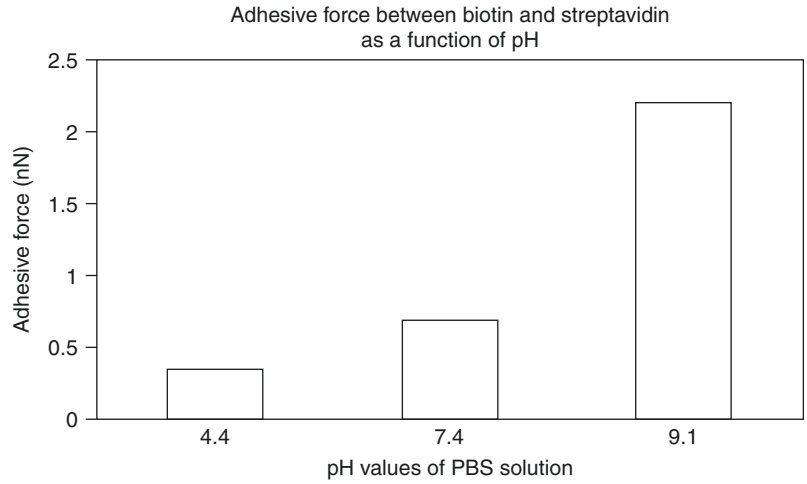
In Fig. 4, the quantity τ corresponds to the rupture period, the time over which the bond will break as related by the equation [17]:

$$F = [U - kT\ln(\tau/\tau_o)]\lambda \qquad (2)$$

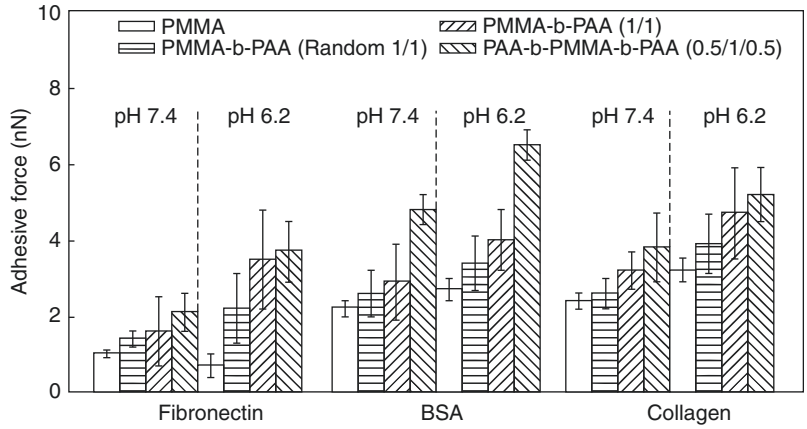
where F is the adhesive force, U is the bond energy, kT is the thermal energy, τ_o is the reciprocal of the natural frequency of oscillation, and λ is a parameter assumed to be equal to the effective rupture length.

The force resolution of the AFM enables it to detect differences on the net charge of biomolecules brought about by a change in the surrounding ionic environment. In another AFM study, Bhushan et al. [18] were able to demonstrate that the adhesion between streptavidin and biotin increases with the pH of the buffer solution, as shown in Fig. 5. This is related to the charge interactions between the two biomolecules. Streptavidin has an isolectric point of pH 5.5, such that it has a net positive charge at pH 4.4 and net negative charge at pH 7.4 and 9.1. The net negative charge in streptavidin facilitates adhesive interactions with the biotin molecule.

Force-distance curves can also be taken over a predefined area divided into an array of points, where the average adhesive force can be determined. This method is useful when analyzing the adhesive properties of next-generation biomaterials such as block copolymers, which have heterogeneous surface properties. Their morphology is a function of the composition and molecular weight of the individual blocks, as well as the spatial relationship of the blocks, for instance A-B block copolymers (or diblock) will have a different morphology than A-B-A block copolymers (or triblock), and so on. This variation in the morphology translates to observable differences



Bioadhesion, Fig. 5 Effect of pH on the adhesion between streptavidin and biotin as measured by AFM (Adapted from [18])



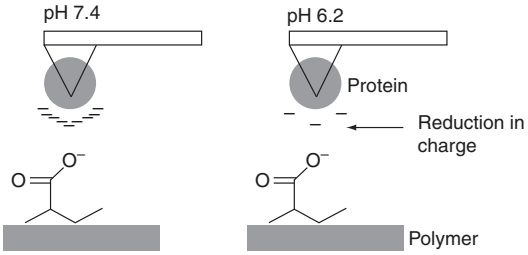
Bioadhesion, Fig. 6 AFM data on the average adhesive force for the interaction between block copolymers containing PMMA/PAA with the proteins fibronectin, BSA, and collagen (attached to functionalized AFM tips) in pH 7.4 (PBS) and pH 6.2 buffer media, with data for PMMA as a Ref. [2]

in the adhesive force. Figure 6 is a summary of the average adhesive force for the interactions between three proteins, namely, fibronectin, BSA, and collagen, on block copolymers composed of poly(methyl methacrylate) (PMMA) and poly(acrylic acid) (PAA), but with different block arrangements (random, diblock, and triblock) [2].

The data in Fig. 6 shows the variation of the adhesive force at two ionic environments, pH 7.4 (PBS) and pH 6.2, where it is seen that the liquid environment affects the adhesive force. At both buffer media, the highest average adhesive force was observed in the triblock copolymer (PAA-b-PMMA-b-PAA, where "b" denotes block copolymer), and the lowest force was obtained from the PMMA reference. The higher average adhesive force in the triblock copolymer surface can be due to greater ordering on its surface, which would expose more PAA-rich areas than it would for the diblock or random copolymer.

The average adhesive forces measured at pH 6.2 are higher throughout the entire series, relative to the PBS medium (pH 7.4) data. This is attributed to higher repulsion at pH 7.4 (or conversely, increased attractive forces at the lower pH). Chemical interactions between the protein and the polymer surface are influenced by the surface charges of the protein and the polymer itself. The amine groups of the proteins are protonated at both pH 7.4 and 6.2. But these proteins carry a net negative charge when immersed in either the pH 7.4 or 6.2 buffer media, since the isoelectric points of fibronectin,

BSA, and collagen are all lower than 6.2 [2]. There are more negative charges on the protein surface at pH 7.4 than at pH 6.2. Also, the acrylic acid chains are ionized (to acrylates) during the adhesive force mapping experiment as it is immersed in aqueous medium. The presence of negative charges on both the tip and sample surface leads to repulsive interactions. As illustrated in Fig. 7, the reduced repulsion between the acrylate groups in the acrylic acid and the smaller amount of surface negative charges in fibronectin, BSA and collagen at the lower pH increases the measured adhesion between the two surfaces [2]. At the higher pH (7.4), the proteins carry a greater net negative charge. Since the acrylic acid in the block copolymer is present in its ionized form, the repulsion from the negative charges present on both the protein and the polymer surface is more significant at pH 7.4 than at pH 6.2. This accounts for the higher adhesive force measured at pH 6.2. The results demonstrate how the AFM is able to measure the effect of the ionic environment on bioadhesion.

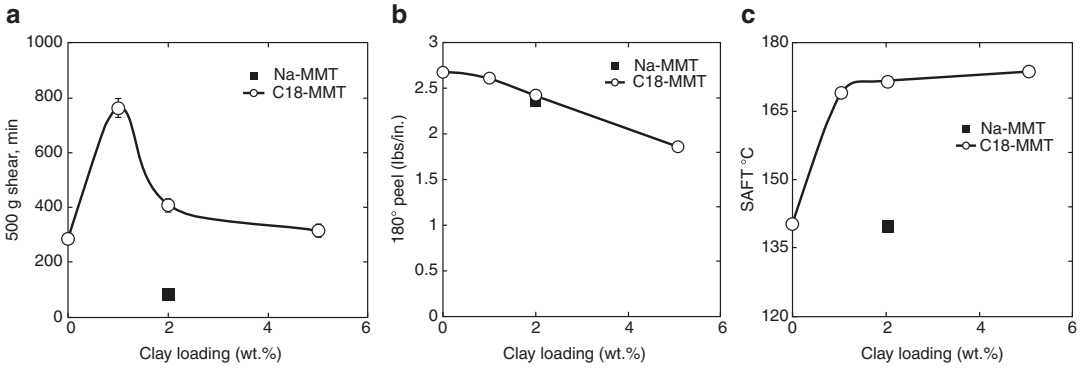


Bioadhesion, Fig. 7 Schematic illustrating the effect of pH on protein-block copolymer interactions (focusing on the effect of PAA, which is ionized at pH 7.4 and 6.2), showing negative charge reduction at lower pH conditions, which leads to decreased repulsion [2]

Macroscale

Standard ASTM characterization procedures, such as the peel test (ASTM D903-98), shear strength (ASTM D-3654), and shear adhesion failure temperature (SAFT, ASTM D4498-00), are conducted on interfaces where an adhesive is applied to the biomaterial. These tests are applicable to systems such as fibrin glues used in surgical operations and pressure-sensitive adhesives (PSA) intended for transdermal drug-delivery applications. Figure 8 shows examples of macroscale adhesion data on pressure-sensitive adhesives. The adhesive system investigated was poly(dimethylsiloxane) (PDMS). It was loaded with an organoclay based on montmorillonite (MMT) to form a composite material. The addition of the organoclay was shown to improve the shear strength and the SAFT, with a minimal reduction in the peel strength [19].

The flow-through method, which measures the flow rate needed to remove a bioadhesive-coated sphere, is a common macroscale experiment suited for the characterization of mucoadhesion of potential drug-delivery systems. A widely used



Bioadhesion, Fig. 8 Macroscale testing of bioadhesion. (**a**) Shear strength versus clay loading in pressure-sensitive adhesive (*PSA*) loaded with an organosilicate nanocomposite (C18-MMT); (**b**) 180° peel strength versus clay loading in the PSA; (**c**) shear adhesion failure temperature (*SAFT*) versus clay loading. Na-MMT denotes the pure clay used as the reference material [19]

biophysical assay involves the monitoring of molecular weight change through the variation in the sedimentation coefficient as measured by the analytical ultracentrifuge [20]. The Wilhelmy plate experiment is another technique for evaluating bioadhesion at the macroscale. Instead of having a liquid medium such as water (as used in surface tension measurements), it is replaced with either natural or synthetic mucus and the plate is coated with the polymer of interest [6, 14]. This method is useful for determining interfacial forces between mucus and polymer, as well as for the investigation of changes in bioadhesion properties over time.

Cross-References

▶ Bioadhesives
▶ BioPatterning

References

1. Keselowsky, B.G., Collard, D.M., Garcia, A.J.: Surface chemistry modulates focal adhesion composition and signaling through changes in integrin binding. Biomaterials **25**, 5947–5954 (2004)
2. Palacio, M.L.B., Schricker, S.R., Bhushan, B.: Bioadhesion of various proteins on random, diblock and triblock copolymer surfaces and the effect of pH conditions. J. R. Soc. Interface **8**, 630–640 (2011)
3. Legeay, G., Poncin-Epaillard, F.: Surface engineering by coating of hydrophilic layers: bioadhesion and biocontamination. In: Possat, W. (ed.) Adhesion: Current Research and Applications. Wiley, Weinheim (2005)
4. Yang, X., Robinson, J.R.: Bioadhesion in mucosal drug delivery. In: Okano, T. (ed.) Biorelated Polymers and Gels. Academic, San Diego (1994)
5. Burkett, J.R., Wojtas, J.L., Cloud, J.L., Wilker, J.J.: A method for measuring the adhesion strength of marine mussels. J. Adhes. **85**, 601–615 (2009)
6. Buckton, G.: Interfacial phenomena in drug delivery and targeting. Harwood Academic, Chur (1995)
7. Spotnitz, W.D., Burks, S.: Hemostats, sealants, and adhesives: components of the surgical toolbox. Transfusion **48**, 1502–1516 (2008)
8. van Noort, R.: Introduction to Dental Materials, 3rd edn. Elsevier, London (2007)
9. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
10. Waite, J.H.: Nature's underwater adhesive specialist. Int. J. Adhes. Adhes. **7**, 9–14 (1987)
11. Asmussen, E., Hansen, E.K., Peutzfeldt, A.: Influence of the solubility parameter of intermediary resin on the effectiveness of the Gluma bonding system. J. Dent. Res. **70**, 1290–1293 (1991)
12. Hozbor, M.A., Hansen, W.P., McPherson, M.: Plasma cleaning of metal surfaces. Precis. Clean. **2**, 46 (1994)
13. Bhushan, B., Tokachichu, D.R., Keener, M.T., Lee, S.C.: Morphology and adhesion of biomolecules on silicon based surfaces. Acta Biomater. **1**, 327–341 (2005)
14. Bhushan, B., Gupta, B.K.: Handbook of Tribology: Materials, Coatings and Surface Treatments. McGraw-Hill, New York (1991)
15. Bhushan, B.: Springer Handbook of Nanotechology, 3rd edn. Springer, Heidelberg (2010)
16. Bhushan, B.: Nanotribology and Nanomechanics, 3rd edn. Springer, Heidelberg (2011)
17. Lee, G.U., Kidwell, D.A., Colton, R.J.: Sensing discrete streptavidin-biotin interactions with atomic force microscopy. Langmuir **10**, 354–357 (1994)
18. Bhushan, B., Tokachichu, D.R., Keener, M.T., Lee, S.C.: Nanoscale adhesion, friction and wear studies of biomolecules on silicon based surfaces. Acta Biomater. **2**, 39–49 (2006)
19. Shaikh, S., Birdi, A., Qutubuddin, S., Lakatosh, E., Baskaran, H.: Controlled release in transdermal pressure sensitive adhesives using organosilicate nanocomposites. Ann. Biomed. Eng. **35**, 2130–2137 (2007)
20. Harding, S.E.: Trends in mucoadhesive analysis. Trends Food Sci. Technol. **17**, 255–262 (2006)

Bio-adhesion

▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Bioadhesives

Scott R. Schricker
Division of General Practice and Materials Science, College of Dentistry, The Ohio State University, Columbus, OH, USA

Synonyms

Clinical adhesives; Medical adhesives

Definition

Materials that are designed to bond interfaces for clinical applications.

Introduction

Bioadhesives can be broadly defined as materials that bond interfaces in biological or clinical situations. There are many well-known examples of biological adhesion including mussels or barnacles sticking to rocks, gecko lizards that walk on ceilings, and spider webs. Some examples of bioadhesion have detrimental consequences: Barnacles adhering to ship hulls are a significant problem for maritime operations, and bacteria will form adhesive biofilms that resist treatment resulting in significant harm to human health. These topics have been widely covered in the literature, and many excellent reviews on these topics have been written.

The main subject area of this entry is bioadhesives that are used in clinical situations to improve human health. While the proteins involved in mussel adhesion have been studied for clinical applications, most of the adhesives used clinically are synthetically derived. A notable exception to this is fibrin glue which will be discussed in detail. One of the reasons for this is that synthetic adhesives can offer a level of consistency and reproducibility that is not always possible from biologically derived materials. This entry will cover synthetic adhesives and fibrin glue and discuss some recent developments in nanostructured adhesives.

Applications and Clinical Problems

Bioadhesives are used in many situations: to adhere biological interfaces, a biological interface to a material or two material interfaces. Many types of materials are used as bioadhesives, and they can have a wide range of mechanical and physical properties depending on the application. For instance, methacrylate adhesives for dental and orthopedic applications are permanent and nondegradable, whereas most soft tissue adhesives such as cyanoacrylates and fibrin glue are temporary and degradable, while adhesives for maxillofacial prostheses and dentures are temporary, and the adhesives must allow for strong bonding until removal for daily cleaning is desired. In addition, bioadhesives must be resistant to biological attack such as hydrolysis, pH, temperature cycles, and enzymatic attack.

Adhesion is a multifaceted phenomenon involving mechanical locking and chemical interactions. For most hard tissue applications, mechanical locking is the dominant interaction that provides interfacial stability. The classic example is a dental amalgam that is bonded to the tooth almost exclusively by mechanical forces. Proper tooth preparation allows for a macroscale mechanical locking, and the malleable nature of the uncured amalgam allows for microscale mechanical locking with the tooth surface. The ability for an uncured adhesive to flow in-between the crevices and ridges of biological surface is important to establish a stable interface. Low viscosity is an important requirement for uncured bioadhesives and is the reason that most bioadhesives are lower-molecular-weight chemicals that are cured in vivo.

Chemical interactions play a significant role in determining the interfacial strength of a bioadhesive. The chemical composition of a bioadhesive will determine its ability to wet the surface of the substrate as well as its ability to form bonds with the substrate. Establishing a good mechanical interface requires the adhesive to wet the substrate, and poor wetting can result in voids/defects at the interface. As most biological surfaces are hydrophilic, low contact angle adhesives will flow into the crevices and ridges better than a hydrophobic adhesive. This will allow for good mechanical locking on the surface. Chemical interactions such as hydrogen bonding, polar interactions, ionic bonding, and covalent bonding will also play a role in determining the bond strength of a bioadhesive.

Recent Trends

Adhesives in Dentistry

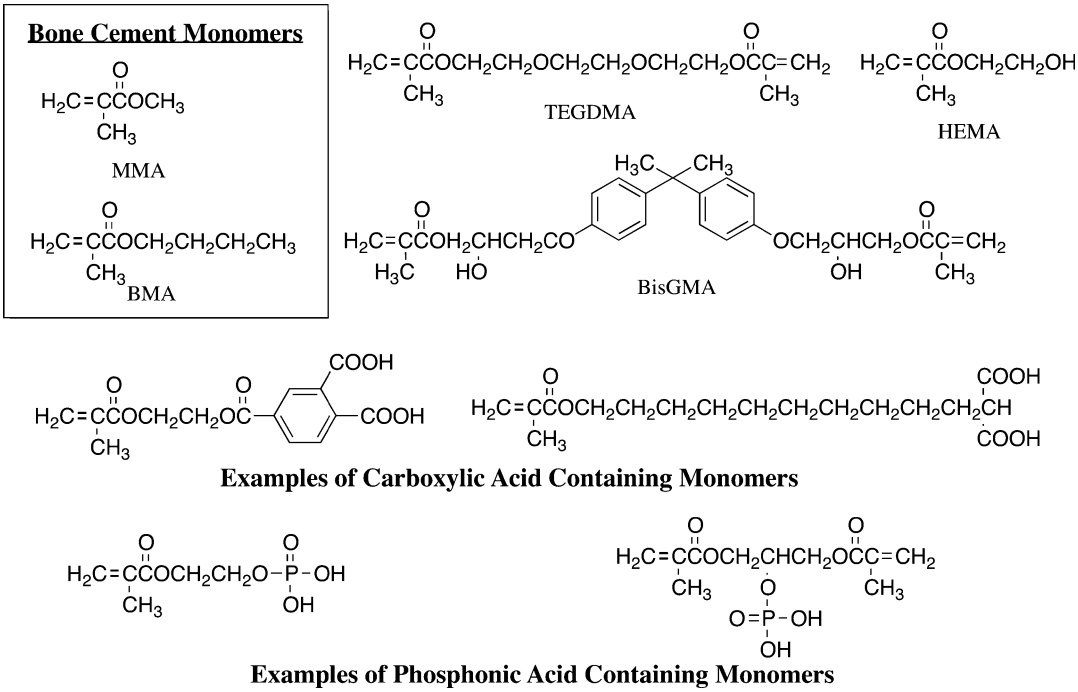
Bioadhesives play a very important role in many aspects of clinical dentistry, binding restorations

to tooth structure, repair of restorations, binding of orthodontic appliances (braces) to teeth, and retention of dentures and maxillofacial prostheses. In restorative dentistry it is estimated that between 50 % and 70 % of procedures are for failed restorations. A significant mechanism of failure is secondary caries or tooth decay underneath the restoration. Most of these failures are directly or indirectly due to poor interfacial bonding between the restorative and the tooth structure. For this reason a great deal of research has focused on improving the adhesive interface between a restorative and the tooth structure. As this constitutes the vast majority of dental adhesion research, this will be the focus of the following discussion. It is important to note that the adhesives described can also be used for orthodontic applications and the repair of restorations or dental appliances. The chemistry and principles of adhesion are the same, and often only minor changes in formulation are made in order to adapt an adhesive to a different dental application.

Most of the adhesives used in clinical dentistry are based on methacrylate polymers and oligomers. A generic outline of a restorative procedure will illustrate how a dental adhesive is used. The clinician will grind away the caries (tooth decay) and expose the underlying dentin. The exposed dentin is coated with ground dentin, known as the "smear layer" that will prevent good bonding. The smear layer is etched away with a strong acid, normally phosphoric acid. The dental adhesive is then placed on the dentin and polymerized with visible light. An uncured composite restorative is then placed on the adhesive and polymerized by visible light. Current dental adhesives incorporate phosphorous-based acid groups that can simultaneously etch the smear layer and bond to the dentin, thus eliminating a step. A visible light-cured adhesive is used in this example; however, if the restoration is preformed to fit the defect, such as crown, precured composite, or a veneer, then a chemically cured adhesive is used. The light cannot penetrate to the adhesive layer, and the adhesive must be polymerized while in contact with the cured restorative to achieve the full adhesive potential.

Methacrylate systems offer many advantages over other curing chemistries and meet most of the requirements for clinical dentistry. As previously mentioned, it is desirable to start with lower-molecular-weight precursors and cure them in vivo. Methacrylate monomers can be cured by visible light within a minute; a longer time frame is not efficient from a practitioner's standpoint and not well tolerated by the patient. Functional groups are well tolerated in methacrylate polymerizations, and hydroxyl-, carboxyl-, and phosphorous-based acid groups can be incorporated as part of dental adhesives. Current formulations of methacrylate monomers and oligomers are stable and have a good shelf life, curing on command with visible light. In addition, most methacrylate polymers and monomers are well tolerated from a biological standpoint. While some concerns about these systems will be addressed later, most methacrylate monomers are considered nontoxic and acceptable for human use.

Early generation dental adhesives are characterized by the use of hydrophilic monomers such as triethylene glycol dimethacrylate (TEGDMA) and 2-hydroxy-ethyl-methacrylate (HEMA). Often a hydrophobic monomer such as (BisGMA) or fillers are added to enhance the mechanical strength of the adhesive. Subsequent generations have incorporated ionic groups such as carboxylic acid, phosphinic acid, and phosphonic acid to enhance bonding to the tooth structure. These acidic groups bind calcium and adhere to calcium-rich surfaces. Enamel and dentin have high levels of the calcium-based mineral hydroxyapatite, 40–50 % and over 90 %, respectively. So these acidic groups contribute significantly to the bonding to the tooth surface. The varied chemistry of adhesive monomers is due to the desire to control properties such as viscosity, mechanical strength, and bonding to the organic components of the tooth structure. The structures and formulations of many commercial and experimental adhesives are found in the literature, and examples of adhesive components are shown in Fig. 1 [1]. The monomers and initiators are often dissolved in a volatile organic solvent such as



Bioadhesives, Fig. 1 Monomers used in dental adhesives and bone cements

ethanol or acetone or a solvent-water combination. The solvent aids in allowing the monomers to penetrate into the dentin to achieve maximum bonding.

A significant focus of dental adhesive research has been on controlling secondary caries and preventing the breakdown of dentin. As was previously discussed, secondary caries is a major factor in the failure of dental restoratives that occurs at the adhesive interface. Chemical modifications and loading the adhesive with antibacterial agents have been employed to increase interfacial stability. Incorporating quaternary amines, silver nanoparticles and monomers derived from antibiotics have all demonstrated the ability to reduce secondary caries. Fluoride-releasing materials have also been successfully incorporated in adhesive formulations. The incorporation of chlorhexidine has also been used to stabilize the adhesive interface. Chlorhexidine has been shown to inhibit MMPs (matrix metalloproteinases) which are known to degrade dentin. Intact dentin is necessary to maintain a good adhesive interface so this has become an active area of research [2].

Dental adhesives and their monomers are considered bioacceptable and are approved for clinical use. However, recent studies in cell culture and animal models have raised some concerns about the unpolymerized monomers and breakdown products that can elute from the adhesive. One of the concerns involves these chemicals acting as endocrine disruptors. Bisphenol A, a known endocrine disruptor, is a precursor to BisGMA and is often detectable in BisGMA-based materials in trace amounts. Additionally, some of the visible light curable initiators such as camphorquinone have been implicated as endocrine disruptors. Unpolymerized methacrylate monomers, in particular TEGDMA and HEMA, are known to induce apoptosis in cell culture. This suggests that as unpolymerized monomer leaches from the adhesive, it will be toxic to the surrounding tissue. In addition, some of these same monomers have been shown to cause genetic damage in cell lines [3–5]. While these are areas of concern,

no human studies have demonstrated any significant negative impact on human health. However, as the definition of biocompatibility evolves, such concerns may need to be addressed.

Bone Cements

There are two main categories of bone cements, methacrylate-based cements and water-based calcium phosphate or calcium sulfate. The chemistry and use of bone cements are similar to dental adhesives. Bone cements are primarily used to affix implants into live bone. Bone implants are normally made from titanium, ceramic, or derived from a human source. None of these implants are malleable and will not form a mechanical bond to the bone. Stress will be placed on the implant bone interface soon after placement, and the bone will not regrow quickly enough to establish a good interface to handle the stress. Unpolymerized methacrylate bone cements will mold to the implant and bone upon implantation. Polymerization will harden the cement and form a mechanical lock.

Bone cements are formed by mixing liquid and solid components which initiates a polymerization reaction. The liquid is a mixture of methacrylate monomers, typically methyl methacrylate (MMA) and butyl methacrylate (BMA) along with a polymerization accelerator. The solid component is a fine powder of polymerized methyl methacrylate polymers or copolymers. The most common powder composition is a mixture of methyl methacrylate homopolymer and a copolymer of methyl methacrylate and styrene. The powder also contains a polymerization initiator, and upon mixing, the accelerator and initiator start the curing process. In addition, the powder may contain fillers to improve mechanical properties or antibiotics to prevent infection. Upon mixing, the cement has a doughy consistency allowing the clinician to place and mold the cement. The set cement is an interpenetrating network of the powder and polymerized monomers. This provides good mechanical properties that are needed to withstand the stresses across the interface.

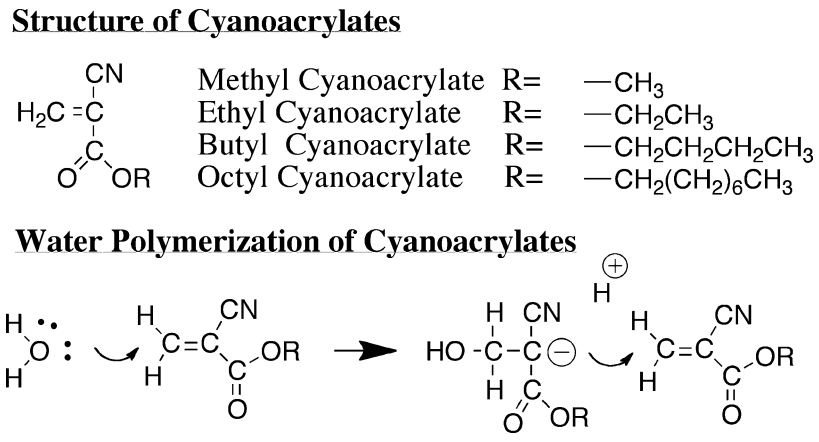
The biological properties of the bone cement are considered to be acceptable. The major drawback is that the polymerization is exothermic and the heat generated can cause tissue damage. There are some concerns about the unpolymerized monomers being cytotoxic, but the long history and extensive use of these materials provides good evidence that they are safe [6, 7].

Water-based bone cements such as calcium phosphate are widely used clinically. They have some advantages over methacrylate bone cements such as better biocompatibility and producing less heat during curing. Calcium phosphate (CP)-based cements are known to resorb in contrast to methacrylate materials which are considered permanent. This is due to the fact that calcium phosphate is similar to the mineral structure of the bone, and the CPs can exist in many different morphologies. These different morphologies will affect the mechanical and biological properties of these materials. The primary drawback to CPs is mechanical properties, which are weaker than the methacrylate cements particularly in tensile strength.

Bone cements, both methacrylate based and water based, have benefited from nanotechnology. Nanoscale fillers such as silica, magnesium oxide, and titania fibers have been used to reinforce methacrylate bone cements. Nanoscale silica demonstrated some physical property improvements but also a reduction in certain mechanical properties such as fatigue strength. The nanoscale magnesium oxide reinforcement demonstrated improved fracture toughness at the cement-bone interface, and it was demonstrated that the nanoscale magnesium oxide improved osteoblast adhesion. Both titania fibers and silicon carbide whiskers improved the fracture toughness and flexural strength at low loading levels. One of the important aspects of incorporating nanomaterials is having a good interface between the bone cement and nanomaterials.

Surgical and Wound Healing Applications

Clinical procedures often require the adhesion of tissue to close a wound or seal an incision after surgery. The applications can range from closing a wound on the skin to closing the small or large intestine after abdominal surgery. Suturing or stitching is a common technique to close wounds that is both labor intesive and requires skill and



Bioadhesives, Fig. 2 Structure and polymerization of cyanoacrylates

training. Surgical adhesives offer a quick, relatively simple alternative to suturing that is highly effective and in some cases superior to suturing. In some cases, sealing the wound immediately is important: For instance, athletes are typically not allowed to compete while they are bleeding. Adhesives offer a quick and convenient route to closing a wound or stopping bleeding.

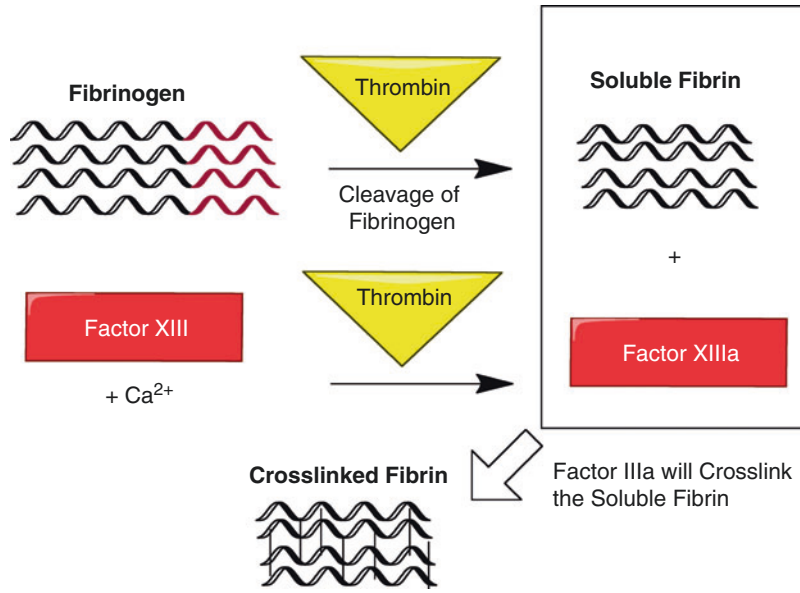
Cyanoacrylates have been widely used as tissue adhesives particularly for external applications. The chemical structures of the common alkyl derivatives and the polymerization process are shown in Fig. 2. Cyanoacrylates are commonly known as "superglue" and are characterized by a rapid set and a strong adhesive bond. The double bond is electron deficient due to the attached cyano and ester groups. The result is that weak nucleophiles such as water can initiate the polymerization. Hydroxyl, amine, and thiol groups that are often found on tissue can also initiate polymerization. Upon contact with blood or tissue, polymerization will start. This is one of its attractive features; no initiator is needed and the rapid polymerization will stop bleeding and quickly seal a wound.

The degradation chemistry of cyanoacrylates is the biggest barrier that prevents broader clinical use. Hydrolysis of polycyanoacrylates will result in the release of formaldehyde and cyanoacetate. Both by-products are considered toxic and as a result, internal use is limited. The chemistry of the cyanoacrylates can be manipulated to reduce this toxicity. As the alkyl group on the cyanoacrylates becomes longer, the polymer is more hydrophobic and the rate of degradation is slower. While the by-products are still toxic, the overall effects are diminished because the rate of release is slow. Toxicity is a combination of the agent and concentration, so slower release will result in diminished toxicity to the tissue. Another result of a longer alkyl chain is that the polymerization rate is reduced. So the handling properties and degradation rate can be manipulated by mixing different ratios of cyanoacrylate monomers.

In addition to external wound closure, there are several clinical applications for cyanoacrylates. A cyanoacrylate-based product known as IntegueSeal is used to seal surgical sites to prevent infections. Upon curing, a barrier is formed that prevents bacteria from migrating to the infection site. Typically, wound sites are treated with iodine solution to reduce infection, and the application of cyanoacrylate after the iodine solution forms a barrier that prevents the iodine sterilization solution from evaporating and increases it effectiveness. In one clinical study, the rates of infection dropped to 53 % compared to 68.7 % with iodine alone [8]. Cyanoacrylates have also been used in ophthalmic and cardiovascular procedures, and it is being explored as a sealant for gastrointestinal surgery.

Fibrin Glue

Fibrin glue is a surgical sealant that is based on the chemistry and components of the blood-clotting cascade. The entire clotting cascade is very



Bioadhesives, Fig. 3 Overview of the final steps of the clotting process and the basic steps involved the formation of fibrin glue. Many more steps and factors are involved in the clotting process, but this is relevant to fibrin glue

complex and is controlled by multiple feedback mechanisms. Figure 3 shows the last few biochemical steps in clotting that are relevant to fibrin glue. In commercial formulations, fibrinogen and factor XIII are isolated from whole blood as are the thrombin components. In most cases, the fibrinogen is isolated from human blood, and the thrombin can be isolated from a human or bovine source. Batches of the components are often pooled from multiple sources to reduce variability and make a more consistent product. There are other factors involved in the setting process that are isolated with the fibrinogen and thrombin; they are usually referred to as fibrinogen and thrombin components. A solution of the fibrinogen is mixed with a thrombin solution containing calcium salts to initiate the biochemical cascade and thus form the fibrin glue.

Fibrin glue was developed from the observation that fibrinogen itself could be used to stop bleeding and help to close wounds. As it became feasible to isolate more components of blood, thrombin was added to enhance the effectiveness of the glue. Because the components of the glue are naturally derived, it is well tolerated by the body. There is no inflammatory response, the fibrin glue can be resorbed by the body, and the breakdown components are nontoxic. The major drawback is the potential of contamination from the biological source. Viruses and other pathogens can be transmitted along with fibrinogen and thrombin components. While no evidence of serious contamination has been documented and many decontamination protocols are utilized, the potential for contamination remains. Prions represent an area of potential concern and also highlight the unknown factors that may be responsible for disease transmission. Until a few decades ago, prions were not known as infectious agents, but their ability to cause disease is well documented. Bovine spongiform encephalopathy (BSE) could potentially be transmitted from bovine sources, as could the human variant Creutzfeldt-Jakob disease. The fact that sources are often pooled to produce the components can increase the risk of transmission.

Fibrin glue was approved for clinical use in the United States in the late 1990s and has been approved for use in Europe and elsewhere for almost two decades prior to that. Many surgical applications use fibrin glue such as cardiovascular surgery, neurosurgery, reconstructive surgery, and gastrointestinal surgery. Though it is beyond the scope of this entry, it is worth noting that fibrin glue is considered useful as a drug delivery carrier and a tissue engineering scaffold. In addition to previously stated convenience of tissue adhesives, there is documented clinical benefit to the use of

fibrin glue. One of the complications of lung surgery is the leakage of air from a closure site on the lung. In a clinical study, one patient group had the surgical site closed with a suture, and in the second group, the site was closed with a suture plus fibrin glue. The percentage of patients without air leakage for the first group was 34 % while fibrin glue group was 61 %. In another example, fibrin glue was compared with other agents, based on collagen, cellulose, or gelatin, to stop emergency bleeding following surgery. After 5 min, 92.6 % of the fibrin glue patients had stopped bleeding compared to 12.4 % of patients using the other methods [9, 10].

Polyethylene Glycol Adhesives

Polyethylene glycol (PEG) is widely used as a biomaterial due to its excellent biocompatibility and ability to resist protein adhesion. PEG and PEG derivatives have been explored as tissue adhesives and sealants. A number of commercial PEG-based tissue sealants have been produced [11]. PEG-based sealants have many different formulations and chemistries to promote adhesion and to control curing. A PEG with a star architecture cured with a thiol-succinimide combination was evaluated as an artery sealant. The adhesive had desirable mechanical and gelling properties and successfully stopped arterial bleeding in a rabbit model. Blends of PEG and natural hydrogels such as gelatin and polyesters have also been evaluated as bioadhesives. A visible light cure formulation of PEG and polylactic acid (PLA) was evaluated in cardiac surgery as a sealant. Similar PEG-PLA block copolymer systems have been proposed in the literature as tissue adhesives.

Polyurethane Adhesives

Polyurethanes have achieved utility as a bioadhesive due to the ability to create biodegradable polyurethane formulations. Typically, urethane monomers or prepolymers having pendant isocyanate moieties cure in the presence of moisture to form a thermoset adhesive. As the isocyanate moieties contact water, they will begin to react losing carbon dioxide and producing an amine group. The amine will add to isocyanate groups, that have not yet reacted with water, forming carbamates. The competitive reaction between the water and carbamate formation is balanced so that a thermoset is formed. The prepolymers can be prepared in several ways including reacting diols and triols with di- or triisocyanates in stoichiometries that form low-molecular-weight prepolymers and not thermoset materials.

Polyurethanes are widely used as commercial adhesives. Isocyanate monomers based on amino acids and prepolymers formed with degradable polyesters have allowed polyurethanes to be used in clinical settings. Lysine di- and triisocyanates are the basis of many biodegradable polyurethanes. Lysine is an amino acid that allows for the gradual breakdown of the resulting urethanes under physiological conditions. In many instances lysine diisocyanate is allowed to react with polylactic acid diols or polycaprolactone diols to form urethane prepolymers. In this manner any bioacceptable diol or triol can be incorporated into the polymer matrix to modify the physical, biological, or degradation properties. The urethane prepolymer or monomer is cured by contact with water as described above. Molecular level adhesion to tissue can be achieved because the native alcohols and amines can react with the isocyanate groups.

Bioadhesives Derived from Nanotechnology

Ophthalmological Applications

Dendrimers have been utilized to develop bioadhesives for ophthalmological applications and represent one of the few examples of nanotechnology applied to adhesives for clinical applications. Dendrimers are a class of highly branched polymers that can adopt a spherical conformation at higher generations. Dendrimers are characterized by being unimolecular, having a well-defined architecture, and formed from a highly controlled synthesis. The unimolecular and defined architecture allows for structure-property relationships to be well defined. So, by changing the chemistry of the dendrimers, the resulting architecture will be modified and lead to different properties. Due to its spherical nature, the terminal groups will

determine the interactions of the dendrimer with solvent, other dendrimers, surfaces, and biological moieties. Complete or partial modification of these terminal groups will have a significant impact on dendrimer properties. The core group and branching units will determine the arrangement and architecture of the terminal groups as well as the relative stiffness or the dendrimer itself. While these principles are generally true, dendrimers are not rigid spheres but rather molecules that are flexible and can react to its external environment. So the branching units can potentially contribute to the solubility of a dendrimer, for example.

The advantageous properties of dendrimers have led Grinstaff and coworkers to develop ophthalmologic sealants with them. The dendrimers are based on polyester chemistry that could hydrolytically degrade over time. The terminal groups were functionalized with methacrylate groups that allowed them to be visible light cured. Animal models have demonstrated that these materials serve well as adhesives for eye surgery. The investigators note little inflammation and high success rate comparable to standard suturing technique. The dendrimer adhesives also provide advantages over suturing. Suturing can further damage the delicate tissue of the eye, and the visible light-cured adhesive is more rapid than suturing. Additionally, the Grinstaff group has developed dendrimers that have peptides as terminal groups instead of methacrylates. The peptides can chemically link to the tissue of the eye and avoid any potential damage from the heat and light of methacrylate curing [12, 13].

Nanostructured Biological Adhesives

The application of nanotechnology to bioadhesives is a growing but relatively new area. Other than the dendrimer-based adhesives mentioned previously, nanotechnology has not made significant inroads to clinical bioadhesives. One aspect of bioadhesion that has incorporated nanotechnology is the creation of surfaces that can control cell and protein adhesion. It has been recognized that micro- and nanotopography can affect cell and protein adhesion in vitro. While much of the nanotopography has focused on surface roughness, there is a growing body of literature that is examining the role of well-defined nanostructured surfaces in controlling cell and protein adhesion. In particular, self-assembled monolayers [14] and nanopatterned surfaces [15, 16] have been used to control cell and protein adhesion.

Block copolymers provide a platform for creating a variety of nanostructures. While their use for many different applications has been well established, recent work has demonstrated that block copolymer morphology can be used to control protein and cell adhesion on a surface. Work by Bhushan and Schricker [17, 18] has demonstrated that varying surface nanomorphology through block copolymer design can affect protein adhesion. As shown in Table 1, varying the arrangement of the blocks while keeping the composition constant can affect protein adhesion. The protein adhesion is measured by covalently attaching a protein to an AFM tip and measuring the amount of force required to pull the tip away from the surface. For the polymethyl methacrylate

Bioadhesives, Table 1 Summary of the measured adhesive force

Material	Adhesive force (nN)					
	Fibronectin	Fibronectin	BSA	BSA	Collagen	Collagen
	(pH 7.4)	(pH 6.2)	(pH 7.4)	(pH 6.2)	(pH 7.4)	(pH 6.2)
PMMA	1.0 ± 0.1	0.7 ± 0.3	2.2 ± 0.2	2.7 ± 0.3	2.4 ± 0.2	3.2 ± 0.2
PMMA-co-PAA (1/1)[a]	1.4 ± 0.2	2.2 ± 0.9	2.6 ± 0.6	3.4 ± 0.7	2.6 ± 0.4	3.9 ± 0.8
PMMA-b-PAA (1/1)	1.6 ± 0.9	3.5 ± 1.3	2.9 ± 1.0	4.0 ± 0.8	3.2 ± 0.5	4.7 ± 1.1
PAA-b-PMMA-b-PAA (0.5/1/0.5)	2.1 ± 0.5	3.7 ± 0.8	4.8 ± 0.4	6.5 ± 0.4	3.8 ± 0.9	5.2 ± 0.7
PMMA-co-PHEMA (1/1)[a]	1.6 ± 0.3	3.4 ± 0.4	1.6 ± 0.4	4.2 ± 0.6	3.6 ± 0.8	5.1 ± 0.6
PMMA-b-PHEMA (1/1)	3.0 ± 0.2	4.4 ± 0.9	3.8 ± 0.4	3.7 ± 0.7	3.9 ± 0.7	5.0 ± 0.6
PMMA-b-PHEMA-b-PMMA (0.5/1/0.5)	4.2 ± 0.3	4.6 ± 1.0	3.3 ± 0.2	3.7 ± 0.4	4.0 ± 0.8	4.5 ± 1.0

[a]Random copolymer

(PMMA) and polyacrylic acid (PAA) block copolymers, an effect in the adhesion of fibronectin and bovine serum albumin (BSA) is observed at low pH but not neutral. However, for the PMMA and poly(2-hydroxy-ethyl-methacrylate) (PHEMA), the effect is observed at neutral pH but not at low pH.

Work by Cooper-White et al. [19, 20] has also demonstrated the use of block copolymers in controlling cellular behavior. Block copolymers of polystyrene and polyethylene oxide have been utilized to effect cellular and protein adhesion. In particular, as the domain sizes of the block changed, the cellular adhesion was affected.

Another application of nanotechnology to bioadhesion is the use of nanoparticles to bond hydrogels and soft tissues. Silica nanoparticles suspended in basic water were applied to a hydrogel surface. By placing another hydrogel on top of the nanoparticle-modified surface and applying pressure, a bond is formed. The polymer chains of the hydrogel will adsorb on the nanoparticle surface and act as a dynamic linker between the chains and the surface. In addition, this technique was used to bond two pieces of calf's liver [21]. This is an interesting use of the intrinsic surface area and size of nanoparticles to create an adhesive bond that would not be possible with larger-scale materials.

Cross-References

► Bioadhesion

References

1. Van Landuyt, K.L., et al.: Systematic review of the chemical composition of contemporary dental adhesives. Biomaterials **28**, 3757–3785 (2007)
2. Montagner, A.F., Sarkis-Onofre, R., Pereira-Cenci, T., Cenci, M.S.: MMP Inhibitors on dentin stability: a systematic review and meta-analysis. J. Dent. Res. **93**, 733–743 (2014)
3. Janke, V., von Neuhoff, N., Schlegelberger, B., Leyhausen, G., Geurtsen, W.: TEGDMA causes apoptosis in primary human gingival fibroblasts. J. Dent. Res. **82**, 814–818 (2003)
4. Wada, H., Tarumi, H., Imazato, S., Narimatsu, M., Ebisu, S.: In vitro estrogenicity of resin composites. J. Dent. Res. **83**, 222–226 (2004)
5. Schweikl, H., Spagnuolo, G., Schmalz, G.: Genetic and cellular toxicology of dental resin monomers. J. Dent. Res. **85**, 870–877 (2006)
6. Kuehn, K.D., Ege, W., Gopp, U.: Acrylic bone cements: composition and properties. Orthop. Clin. North Am. **36**, 17–28 (2005)
7. Smith, D.C.: The genesis and evolution of acrylic bone cement. Orthop. Clin. North Am. **36**, 1–10 (2005)
8. Wilson, S.E.: Microbial sealing: a new approach to reducing contamination. J. Hosp. Infect. **70**(Suppl 2), 11–14 (2008)
9. Jackson, M.R.: Fibrin sealants in surgical practice: an overview. Am. J. Surg. **182**, 1S–7S (2001)
10. Spotnitz, W.D.: Fibrin sealant: past, present, and future: a brief review. World J. Surg. **34**, 632–634 (2010)
11. Mehdizadeh, M., Yang, J.: Design strategies and applications of tissue bioadhesives. Macromol. Biosci. **13**, 271–288 (2013)
12. Grinstaff, M.W.: Designing hydrogel adhesives for corneal wound repair. Biomaterials **28**, 5205–5214 (2007)
13. Mintzer, M.A., Grinstaff, M.W.: Biomedical applications of dendrimers: a tutorial. Chem. Soc. Rev. **40**, 173–190 (2010)
14. Raynor, J.E., Capadona, J.R., Collard, D.M., Petrie, T. A., Garcia, A.J.: Polymer brushes and self-assembled monolayers: versatile platforms to control cell adhesion to biomaterials (Review). Biointerphases **4**, FA3–FA16 (2009)
15. Lu, J., Rao, M.P., MacDonald, N.C., Khang, D., Webster, T.J.: Improved endothelial cell adhesion and proliferation on patterned titanium surfaces with rationally designed, micrometer to nanometer features. Acta Biomater. **4**, 192–201 (2008)
16. Dalby, M.J., Gadegaard, N., Wilkinson, C.D.: The response of fibroblasts to hexagonal nanotopography fabricated by electron beam lithography. J. Biomed. Mater. Res. A **84**, 973–979 (2008)
17. Schricker, S., Palacio, M., Thirumamagal, B.T., Bhushan, B.: Synthesis and morphological characterization of block copolymers for improved biomaterials. Ultramicroscopy **110**, 639–649 (2010)
18. Palacio, M., Schricker, S., Bhushan, B.: Bioadhesion of various proteins on random, diblock, and triblock copolymer surfaces and the effect of pH conditions. J.R. Soc. Interface **8**, 630–640 (2011)
19. George, P.A., Donose, B.C., Cooper-White, J.J.: Self-assembling polystyrene-block-poly(ethylene oxide) copolymer surface coatings: resistance to protein and cell adhesion. Biomaterials **30**, 2449–2456 (2009)
20. George, P.A., Doran, M.R., Croll, T.I., Munro, T.P., Cooper-White, J.J.: Nanoscale presentation of cell adhesive molecules via block copolymer self-assembly. Biomaterials **30**, 4732–4737 (2009)
21. Rose, S., Prevoteau, A., Elziere, P., Hourdet, D., Marcellan, A., Leibler, L.: Nanoparticle solutions as adhesives for gels and biological tissues. Nature **505**, 382–385 (2014)

Biocalorimetry

► Nanocalorimetry

Biocharacterization

► DNA Manipulation Based on Nanotweezers

Bioderived Smart Materials

Vishnu-Baba Sundaresan[1], Stephen Andrew Sarles[2] and Donald J. Leo[3]
[1]Mechanical and Nuclear Engineering, Virginia Commonwealth University, Richmond, VA, USA
[2]Mechanical Aerospace and Biomedical Engineering, University of Tennessee, Knoxville, TN, USA
[3]Mechanical Engineering, Center for Intelligent Material Systems and Structures, Virginia Tech, Virginia Polytechnic Institute and State University, Arlington, VA, USA

Definition

Bioderived smart materials are a category of ionic active materials that are fabricated from biological macromolecules and utilize the ion transport properties of cell membranes to couple multiple physical domains (chemical, electrical, mechanical, and optical). Owing to this coupling exhibited by biological macromolecules, bioderived smart materials are used as actuators, sensors, and energy harvesting applications.

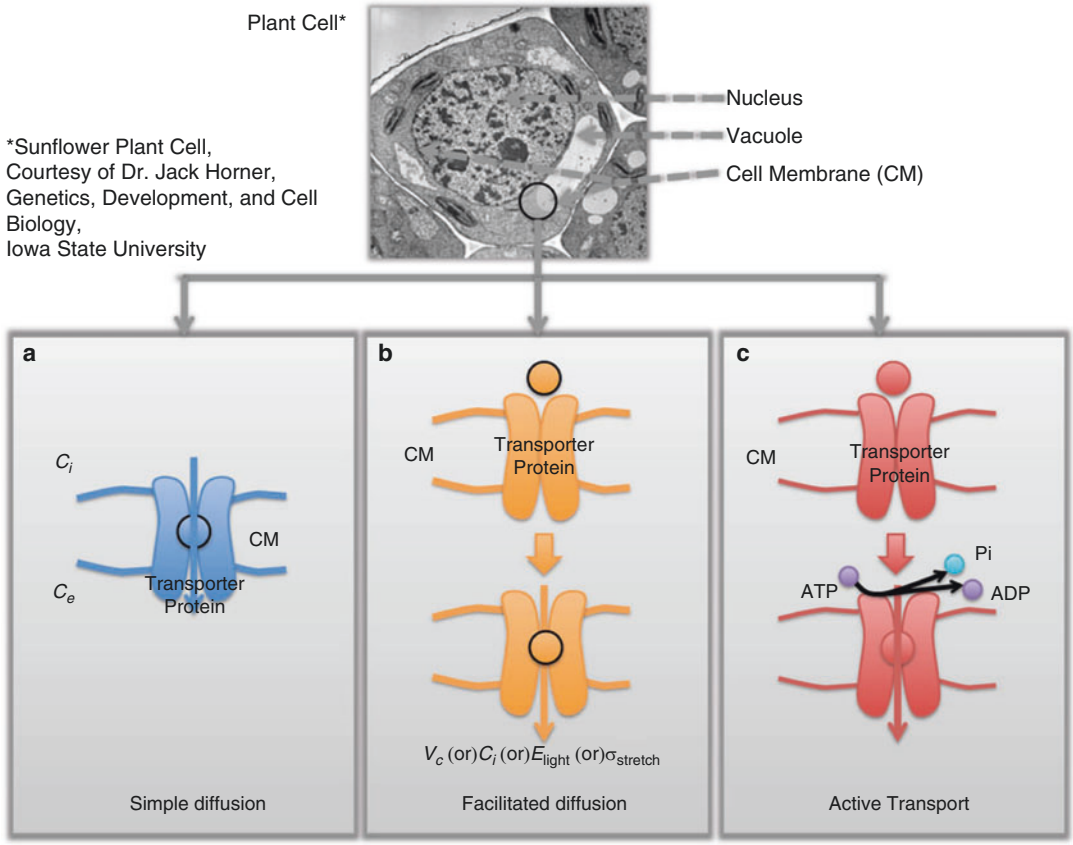
Cell Membrane Proteins: Transporter Macromolecules

The cell membranes of plant and animal cells are host to cholesterol and a variety of proteins that provide structural rigidity and serve as anchor points for cell attachment. In addition, some proteins serve as transporters and allow the cell to exchange ions and neutral molecules between the cell cytoplasm and its surroundings. The transport of species through a protein transporter in the cell membrane is driven by biological processes that have inspired the development of novel sensing, actuation, and energy harvesting concepts. This entry will provide the technical background on biological processes that have inspired the development of *bioderived smart materials* and discuss the system level concepts, fabrication, and characterization of these material systems. Specific emphasis will be given to novel actuation systems, sensing platforms, and recent advances in fabrication of these materials systems.

Transporter proteins are biological macromolecules embedded in cell membranes that convert electrical, mechanical, or chemical stimuli into ion transport and regulate chemoelectrical potentials across the membrane. It is this stimuli response of protein transporters that makes it viable for the development of a bioderived smart material using biological ion transport processes. The transport of ions through the transporter protein occurs via one of the following processes:

1. Simple diffusion
2. Facilitated diffusion
   (a) Voltage-gated diffusion
   (b) Ligand-gated diffusion
   (c) Mechanically gated diffusion
3. Active transport

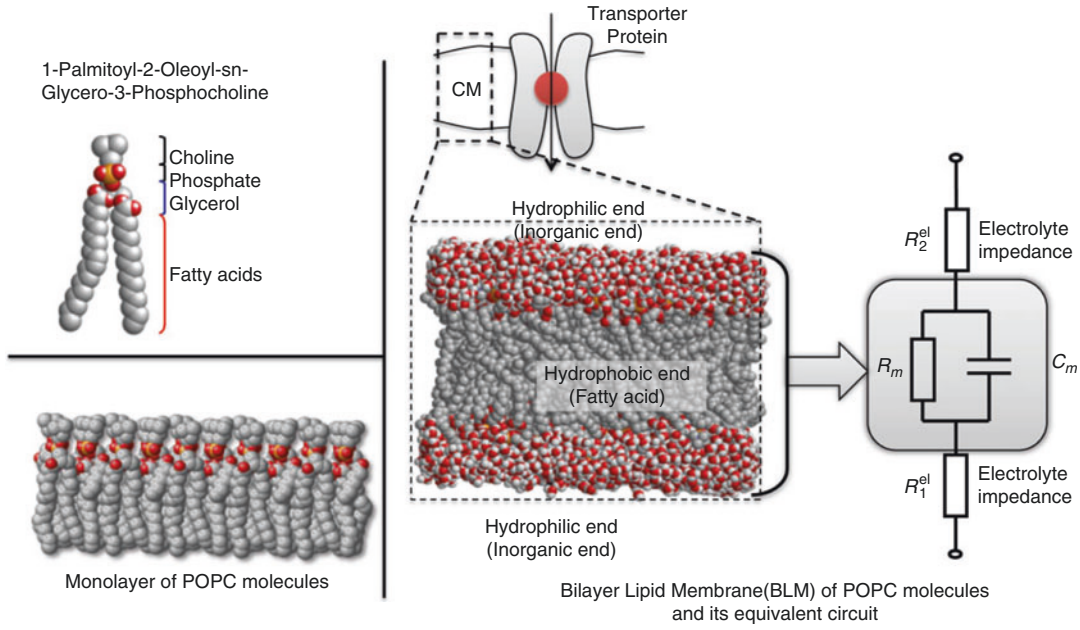
These ion transport processes are illustrated in Fig. 1 and are used to classify protein transporters found in cell membranes. In simple diffusion, the protein transporter behaves like an open channel and allows diffusion of a species along a concentration gradient across the membrane. In facilitated diffusion, the protein transporter uses a gating signal such as light, electrical field, mechanical stretch (or) an analyte to open and conduct an ion or neutral molecule across the membrane. In active transport, the protein transporter uses the energy from biosynthetic molecules such as adenosine triphosphate (ATP) to transport ions across the membrane. Since active transporters expend chemical energy available in triphosphates, ions are transported across the membrane against an



Bioderived Smart Materials, Fig. 1 Protein transporters in plant cell membrane

existing concentration gradient. In addition to ion transport, active transporters establish the necessary chemoelectrical gradients required for facilitated diffusion across the membrane and provide the driving force for long-range ion and solute transport in plant and animal cells.

The mechanism for ion transport through the protein transporter is explained by its molecular composition and structural arrangement. The structural composition of the protein transporter is unique to a plant (or) animal cell membrane and hence transporters are unique in their ion transport function. The diversity of plant and animal cell membranes and the uniqueness of protein transporters found in the cell membranes lead to a large number of protein transporters that can be used as an active component in a bioderived smart material. Protein transporters ion proteins are formed from amino acids that are connected by a polypeptide bond. There are some 20 amino acids that form the majority of the protein transporters and have clearly identified C-C backbone and polar, nonpolar, and aromatic side chains. The linear network of polypeptide bonds has a well-defined two-dimensional structure and is woven (folded) into a three-dimensional structure in the cell membrane. The arrangement of amino acid side chains in the two-dimensional and three-dimensional structures contained within the protein transporters leads to hydrogen bonding, van der Waals, and steric interactions in response to external stimuli. This response in structural changes for an applied optical, electrical, mechanical, or chemical stimuli results in the formation of a conductive pathway for a specific ion to migrate across the membrane and results in an ionic current through the protein transporter and the cell membrane carrying the protein transporter. As a



Bioderived Smart Materials, Fig. 2 L-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine molecule, monolayer, bilayer, and equivalent electrical circuit

consequence of the ionic current, the protein transporter maintains and regulates the electrical potential across the membrane in response to the applied stimulus.

Bilayer Lipid Membranes: A Breadboard for Bioderived Smart Material Concepts

The fabrication of a bioderived smart material requires a membrane that can serve as a host to transporter proteins and exhibit coupling between multiple physical domains (chemical, electrical, mechanical, and optical). This membrane should have the structural properties of a cell membrane and assume the ion transport function imparted by the transporter protein. In addition, this membrane should be sufficiently modular to accommodate multiple proteins and provide a composite ion transport function. The current state of the art in bioderived smart materials uses the same macromolecules that form the cell membranes of plant and animal cells. These molecules, referred to as glycerophospholipids (also referred to as phospholipids or lipids), are extracted and purified from cell membranes and are amphiphilic in nature. Phospholipids comprise of an organic alkyl end group with one or few double bonds and a polar end group connected to each other via a glycerol-phosphate group. An example phospholipid molecule L-*p*almitoyl-2-*o*leoyl-sn-glycero-3-*p*hospho*c*holine (POPC) with an organic palmitoyl- oleoyl- and polar -choline end group is shown in Fig. 2. These molecules, owing to their amphiphilic nature assemble into a bilayer formed from stacked monolayers with the hydrophilic ends of the monolayer facing outside as shown in Fig. 2. This stacked arrangement of hydrophilic-hydrophobic-hydrophilic groups in the bilayer membrane offers high impedance for transport of ions and neutral molecules through the membrane. The bilayer lipid membrane (BLM) measures 6–10 nm in height, spans 10–20/μm^2 area, and serves as the host to various transporter proteins and forms the basic structure of a bioderived smart material. Since various transporter proteins with different ion transport functions can be incorporated in the BLM to provide a composite ionic function, the BLM becomes the breadboard for various smart material concepts.

Owing to their biological origin, the BLM with the protein transporter is referred to in this entry as the bioderived membrane.

In order to quantify the ionic current through the bioderived membrane in the presence of the stimulus, it is required to establish a measure of baseline ion transport (current) through the BLM and charge separation (electrical potential) across the membrane without the protein transporter. The conductance of the membrane is represented using a frequency-dependent complex impedance function and the membrane is modeled using electrical equivalents as shown in Fig. 2. The complex impedance is measured using electrical impedance spectroscopy (EIS) and fitted to an equivalent circuit. In the absence of protein transporters in the BLM, the membrane has a high electrical impedance (~**O** 0.1–1 GΩ.cm^2) and a capacitance of 1/μF.cm^{-2} [1, 2]. In the presence of protein transporters, the membrane assumes the transport properties of the protein and has markedly different conductance states. The impedance of the membrane with the protein in the presence of the stimuli measured using EIS and the conductance states through cyclic voltammetry (CV) and chronoamperometry (CA) establish the incorporation and functioning of the protein transporter in the membrane.

Concepts for Ionic Active Materials Using Bioderived Membranes

The design of smart materials system using protein transporters employs one or several of the previously discussed ion transport processes. In order to understand the energy conversion processes in a protein transporter and its performance as an ionic active material, it is necessary to quantify the work performed by the bioderived membrane for an applied input energy. The electrical work done by bioderived membranes from ion transport is

$$\Delta U_E = \int_0^q E dQ \tag{1}$$

where E is the electrical field established by the protein and dQ is the charge displacement through the membrane. The electrical field established across the membrane is given by $E = V/t_m$ where V is the transmembrane potential and tm is the thickness of the membrane. The charge displacement dQ is due to the ionic current i_c through the protein transporter embedded in the membrane and can be obtained from electrochemical measurements on the membrane. The ionic current results in altering the concentration of the ion across the membrane that sets up an additional transmembrane potential V_c give by

$$V_c^{\infty} = \frac{RT}{zF} \ln\left(\frac{C_e^{\infty}}{C_i^{\infty}}\right) \tag{2}$$

where $C_e^{\infty}, C_i^{\infty}$ are the concentration of the ion on either sides of the membrane at the end of the process and z is the valency on the transported ion. Similarly, the chemical energy gradient across the membrane due to a concentration gradient is given by

$$\Delta\mu = RT \ln\left(\frac{y_e^{\infty}}{y_i^{\infty}}\right) \tag{3}$$

where $y_e^{\infty}, y_e^{\infty}$ are the mole fractions of species y on either side of the membrane at steady state. The interconversion between chemical potential and electrical potential across the membrane shown in Eqs. 1, 2, and 3 occur via ionic currents through the protein transporter.

The ionic currents through the bioderived membrane can be written for various transport process from *phenomenological equations* as a function of one or more of the following as shown in Sundaresan et al. [3]

1. Applied transmembrane potential (V)
2. Chemical energy gradient ($\Delta\mu$)
3. Chemical energy released from biochemical reaction ($\Delta\mu$)

The diffusion ionic current through the membrane of area A and permeability constant P_m in the presence of concentration gradients Ce, C_i, and transmembrane potential V is given by Goldman-Hodgkin-Katz equation as

$$i_{diff} = -AP_mF\left(\frac{zFV}{RT}\right)\left(\frac{[C]_i - [C]_e e^{-\frac{zFV}{RT}}}{1 - e^{-\frac{zFV}{RT}}}\right). \quad (4)$$

The ionic current through the protein transporter via voltage-gated diffusion i_{vg} is given by

$$i_{vg} = 2\psi zekT\sqrt{[C]_e[C]_i} \times \frac{A_p}{\in d}\sinh\left(\frac{e(V - V_C)}{2kT}\right), \quad (5)$$

where ψ is the voltage gating coefficient, z is the valency of the transported species, e is the charge on the ion, A_p is the area of the ion conducting channel, and d is the length of the pore in the protein transporter. The ionic current through the protein transporter via active transport i_{pump} is given by

$$i_{\text{pump}} = Me\lambda\tanh\left(\frac{[F(-V - V_{\text{ATP}} + V_C]}{2kT}\right), \quad (6)$$

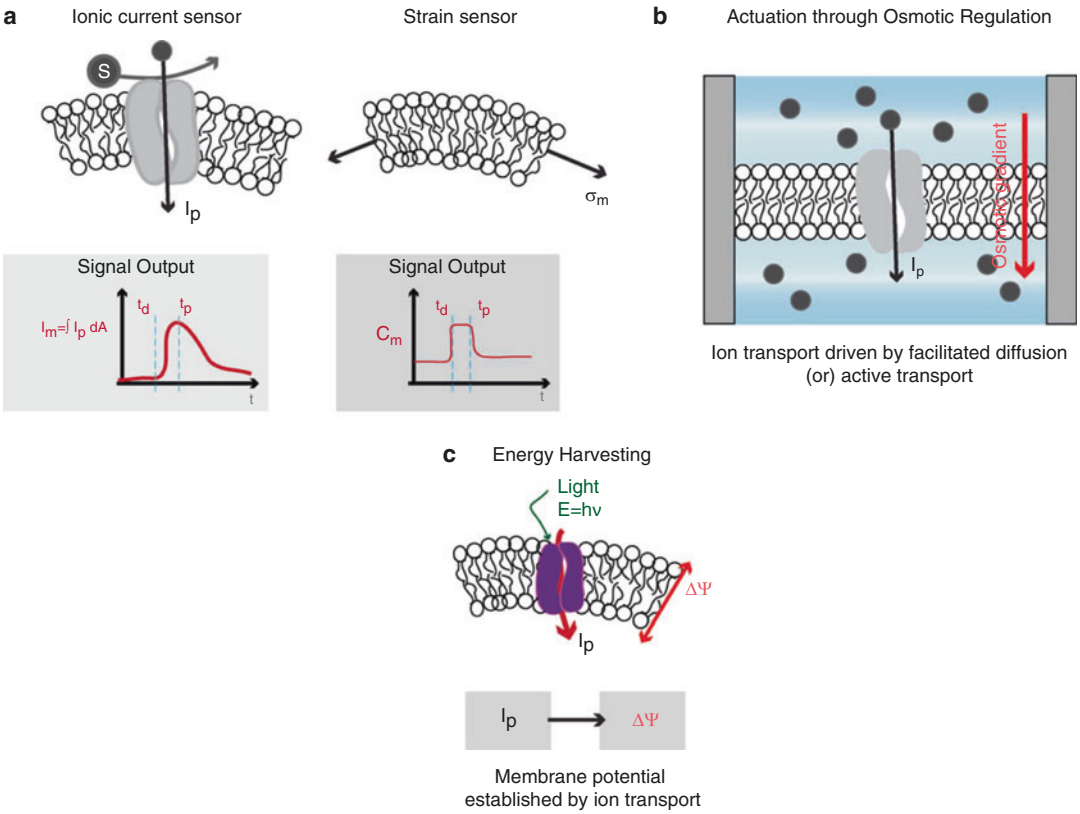
where M is the pump density in the membrane, z is the charge of the pumped ion, λ is the sum of the forward and backward reaction rates in the pump, V_c is the electrochemical gradient for the pumped ion given by Eq. 2, and k is Boltzmann constant.

The choice of transporter protein in bioderived materials depends on the nature of the stimulus to be detected for sensing, stimulus that is available for actuation and ambient energy source available for harvesting. This choice is also dependent on the ability to extract a protein transporter from a cell membrane and the ability to reconstitute the cell membrane-like structure on a synthetic platform. Thus, a bioderived membrane measuring 6–10 nm in thickness with finite number of transporter proteins supported or suspended on a solid substrate and membrane area of $\sim 10\ \mu m^2$ serves as the fundamental unit in a bioderived smart material. In an actuator, sensor, or energy harvesting device, a large number of such membranes with protein transporters are assembled to work in parallel as a smart material. An overview of concepts that use the bioderived membrane as the active component in sensors, actuators, and energy harvesting devices is shown in Fig. 3. The ionic currents through the membrane can be obtained from the ionic current through the transporter from expressions similar to Eqs. 4, 5, and 6 for the design of smart material systems.

Sensor: The bioderived membrane is used as the sensing element by monitoring the electrical response of the membrane in response to an applied stimulus. The bioderived membrane is formed with electrodes on either sides of the membrane on a synthetic substrate and packaged into a system for monitoring electrical current, transmembrane voltage, or electrical impedance. With the choice of an appropriate protein transporter, the sensor is tailored to respond to chemical analyte, bioelectrical sensors, mechanical strain, light, and temperature.

Actuator: A bioderived membrane is used to regulate the ionic concentration between two closed volumes in response to an applied stimulus. The ionic currents through the bioderived membrane sets up ionic and osmotic gradients across the membrane. This concentration gradient becomes an intermediate stimulus for osmotic regulation and movement of the solvent molecules (water) across the membrane. The bulk movement of water molecules leads to volumetric expansion of a closed volume that will lead to bulk stress in the material. The bulk stress generated in a closed volume encompassing the bioderived membrane is placed in a system that will lead to the development of a stack- or a bimorph-type actuator.

Energy Harvesting: The definition of an energy harvesting device has evolved over the last decade and is now defined as a system that converts ambient energy into electrical energy that can be stored locally and used later for powering electronic devices. In this context, a bioderived membrane generates a chemoelectrical potential from ion transport by varying the ionic concentration as shown in Eq. 2. The concentration gradient and the resulting chemoelectrical potential can be generated by active transport using incident light, consuming triphosphates (adenosine, guanosine), etc. Due to the nature of the ionic transport processes and the resulting gradient across the membrane, protein transporters behave like constant current power sources.

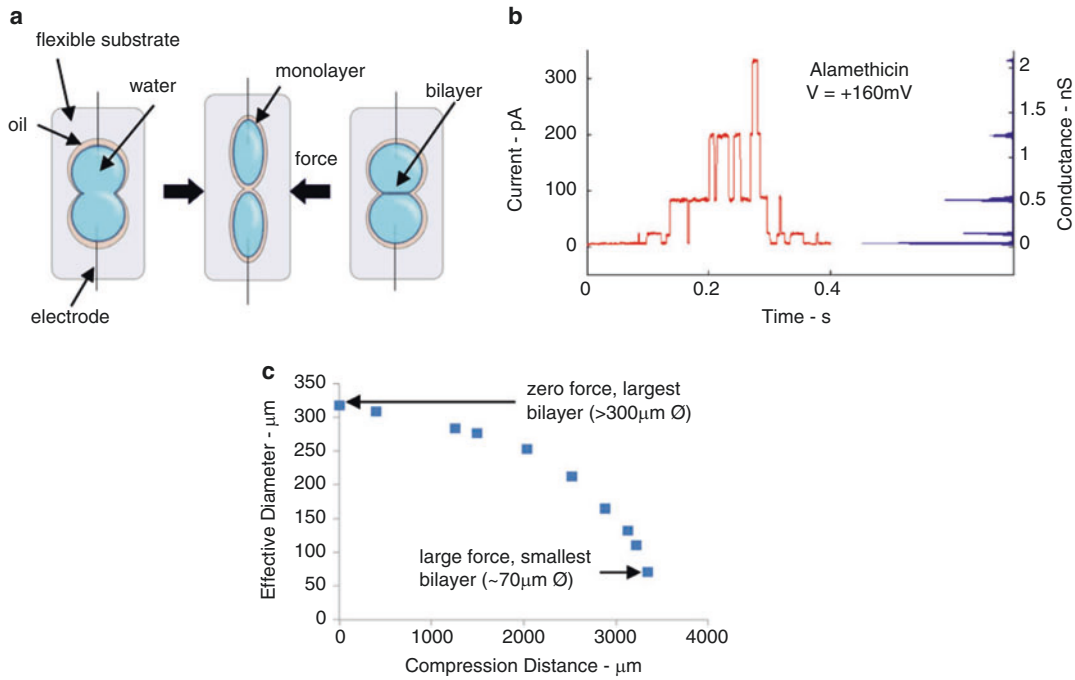


Bioderived Smart Materials, Fig. 3 Concepts for (**a**) sensor elements, (**b**) actuation units, and (**c**) energy harvesting cells using bioderived membranes

Fabrication Methods and Recent Advances

The methods to form BLM utilize the amphiphilic nature of the phospholipid molecule to produce a highly organized structure in the presence of an aqueous medium. In the past five decades, many techniques have emerged to assemble phospholipids into vesicles and planar BLM [1, 4–8]. Vesicles are hollow-spherical structures that are useful as carrier vehicles for chemical payloads. The planar configuration of the lipid membrane (planar BLM) permits access to both sides of the membrane and is preferred for the investigation of ion transport through transporter proteins. The bioderived smart materials typically use a planar BLM supported on a solid substrate (supported BLM) or suspended across the pores of a nanofabricated porous substrate (suspended BLM).

The tools and methods to fabricate a synthetic biological membrane trace its origin to the pioneering research by Mueller and Rudin and coworkers [1, 4] and Tien, Ottova, and coworkers [9]. The initial studies focused on forming planar BLMs from egg phospholipids across a 200-μm pore in a polymeric membrane with voltage-gated ion channels [4]. These initial experimental investigations and subsequent advances by Tien, Ottova, and coworkers were motivated by the need to create in vitro test platforms for analyzing the physiological and immunological function of cell membranes [10–12]. Recently, the experimental methods to form a BLM with proteins are motivated by the engineering applications of bioderived membranes. The anticipated applications for bioderived membranes have significantly influenced the fabrication methods and material choice in the past decade.



Bioderived Smart Materials, Fig. 4 (**a**) The regulated attachment method uses a flexible substrate and applied force to enable bilayer formation. (**b**) Membranes produced via the RAM enable single-channel measurements. (**c**) Modulating the applied force allows for reversible control on the size (area) of the bilayer

A consistent challenge in transitioning supported and suspended lipid bilayers into robust material platforms is the fragility and limited shelf life of the thin membrane. Various approaches have been investigated to stabilize lipid bilayers, including suspending bilayers across nanopores, tethering bilayers to solid surfaces, and sandwiching suspended membranes between water-swollen hydrogels. Among them, this entry will discuss a novel platform for reconstituting phospholipids into a BLM in a curable polymerizable structure. The method, referred to as droplet interface bilayer (DIB), utilizes the affinity of the hydrophobic tails to organic solvents (or) oils. The DIB technique first demonstrated by Funakoshi et al. [13] and then refined and expanded by Bayley et al. [14–16] showed that the stability of the bilayer can be improved by minimizing the direct interaction of phospholipids with a solid substrate. This leads to a BLM with improved longevity (DIBs lasted for days to weeks) and increased resistance to rupture. In this approach, a liquid-supported interface bilayer is formed by connecting two lipid-encased water droplets submerged in oil as shown in Fig. 4. The monolayers that make up the bilayer self-assemble at the oil–water interface surrounding each droplet and spontaneously "zip" together when the droplets are placed into contact. The formation of a lipid bilayer at this interface occurs by thinning when depletion flocculation removes excess oil from between the droplets. Using this principle, the amount of contact between two droplets can be modulated by the choice of oil.

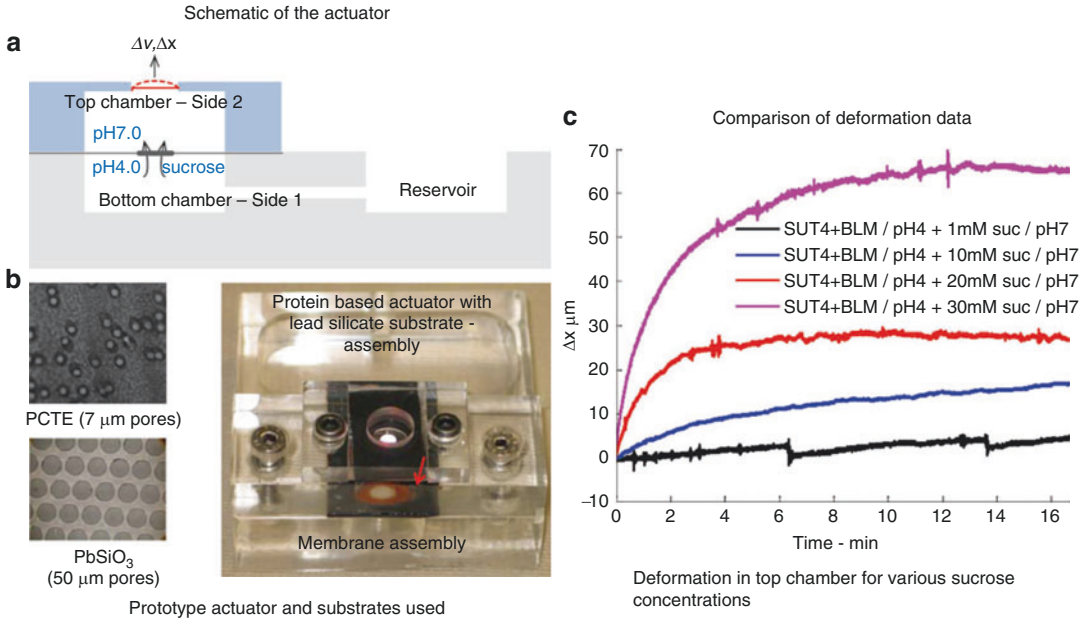
Aside from advantages of simplicity and stability, the DIB embodiment affords the ability to form a biomolecular network of lipid bilayers by connecting more than two droplets. The ability of the DIB membrane to accommodate proteins makes it feasible to demonstrate selective transport across multiple interfaces formed from a network of droplets. The primary disadvantage of this method, though, is the need to dispense and position individual water droplets a task that becomes increasingly difficult for small droplets

(<100 μm). To remedy this limitation and to provide added support to the droplets themselves, Sarles and Leo recently developed an alternative technique called the regulated attachment method (RAM) for interface bilayer formation within a flexible, nonwetting solid substrate [17, 18]. Instead of positioning droplets individually, the RAM uses mechanical force applied to the flexible substrate to control the attachment of lipid-coated aqueous volumes contained in neighboring compartments. The applied force regulates the dimensions of the aperture separating the compartment as shown in Fig. 4a. With the initial design, increasing the force closes the aperture and separates the volumes. Relaxing this force, then, opens the aperture and allows the volumes to come into contact. The results of this method proved that, like DIBs, interface bilayers formed using RAM lead to stable membranes that exhibit high electrical resistances (>10GΩ) necessary for measuring single-channel and multiple-channel ion currents through proteins such as alamethicin (Fig. 4b). Further, the RAM enables the size of the bilayer (i.e., the area of contact) to be reversibly modulated after membrane thinning. It has been shown through experiments that the effective diameter of the bilayer formed from RAM technique could be varied by a factor of 5× by merely modulating compressive force (Fig. 4c). The flexibility of the regulated attachment method (RAM) provides additional advantages over earlier methods to form lipid bilayers, including the DIB approach. First, this method provides control of the size of the bilayer, independent of the sizes or shapes of the aqueous volumes. Sarles and Leo later utilized this advantage in forming interface bilayers between nonspherical hydrogels [19]. Second, supporting the aqueous volumes within a flexible substrate helps to reduce relative motion between the connected volumes under vibration and shock, making for a more portable system. With this added support, additional contents in the form of discrete aqueous droplets can be added to an existing lipid-coated volume via a microchannel or syringe without causing the membrane to rupture. This capability enables biomolecules (or other species) to be added to the membrane at a designated time after bilayer formation.

Bioderived Microhydraulic Actuators

Motion in biomolecules, especially protein transporters, is typically associated with conformational change in proteins and is of the order of few angstroms of displacement and/or angular rotation of the subunits in the protein transporter [20]. There are numerous challenges in coupling this mechanical motion resulting from three-dimensional conformational change of proteins to an external system and hence cannot be used for performing mechanical work. Another interesting example of chemomechanical actuation is the demonstration of kinesin motility on microtubules [21]. In this system, a kinesin-functionalized nanoparticle traverses along the length of a microtubule resulting from the rotation of kinesin dimer from ATP hydrolysis. Various research groups have used nanobeads functionalized with kinesin as platform for demonstrating this novel actuation concept. While this concept is significantly advanced than molecular motors, coupling this linear motion to generate force and volumetric strain has challenges in interfacing the bimolecular motion with a structure. In order to demonstrate a system that uses biomolecules to generate volumetric strain similar to ferroelectrics, electroactive polymers, a membrane-based hydraulic actuation concept was developed by Sundaresan and Leo [3].

The membrane-based microhydraulic actuator uses a BLM with ion transporters as shown in Fig. 5a, b. The chemomechanical actuator uses osmotic regulation between two chambers resulting from ion and sucrose transport through the bioderived membrane. The bioderived membrane is suspended across the pores of a microporous substrate and separates the contents of the two chambers and is reconstituted with a sucrose transporter protein (SUT4) extracted from *Arabidopsis thaliana*. The SUT4 protein, grown in and extracted from yeast cell membrane, is a cotransporter that transports proton and sucrose from the side of the membrane with higher proton and sucrose concentration. On applying a buffered proton gradient, the cotransporter balances the pH gradient with sucrose concentration gradient. The sucrose concentration gradient generates an



Bioderived Smart Materials, Fig. 5 (**a**–**b**) Schematic of a membrane-based chemomechanical actuator. (**c**) Deformation of cover plate in the actuator for various initial sucrose concentrations

osmotic gradient across the membrane and produces water transport through the membrane. The additional volume of fluid from osmotic regulation, demonstrated in Sundaresan et al. [22], builds pressure in the chamber, balances the osmotic gradient established by sucrose transport, and deforms a flexible wall in the chamber.

The mechanics of ion transport leading to an osmotic gradient due to a cotransporter is shown using phenomenological equations. The forces in the system during proton-sucrose cotransport through the membrane for actuation are chemical potentials ($\Delta\mu$) due to

1. Concentration gradient of proton ($\Delta\mu_p$)
2. Concentration gradient of sucrose ($\Delta\mu_s$) – cotransported specie
3. Osmotic gradient due to concentration gradients of charged species and nonelectrolytes

The proton and sucrose fluxes (ϕ_p, ϕs) through the membrane due to the applied concentration gradients across a membrane can be represented by phenomenological equations for ion transport formulated by Katachalsky et al. [23, 24].

$$\begin{aligned}\phi_s &= L_{s,s}\Delta\mu_s + L_{s,p}\Delta\mu_p \\ \phi_p &= L_{p,s}\Delta\mu_s + L_{p,p}\Delta\mu_p.\end{aligned} \tag{7}$$

Assuming unity values for activity coefficients, the chemical potentials that contribute to the flux are [25]

$$\begin{aligned}\Delta\mu_p &= RT\ln\left(\frac{C_1^p}{C_2^p}\right) \\ \Delta\mu_s &= RT\ln\left(\frac{C_1^s}{C_2^s}\right).\end{aligned} \tag{8}$$

The coefficients in the flux equations in Eq. 7 represent the coupling between proton and sucrose transport through the membrane. The coefficient $L_{s,s}$ represents the coupling between sucrose chemical potential and sucrose flux. The coefficient $L_{s,p}$ relates sucrose flux to the chemical potential due to proton and similarly, $L_{p,s}$ relates the proton flux to sucrose chemical potential. The coefficients $L_{s,p}$, $L_{p,s}$ are assumed equal due to the Onsager symmetry relationship [26]. The chemical potentials due to sucrose and pH gradients are assumed to balance each other as shown in Sundaresan and Leo [22] and the equilibrium

concentration of sucrose is obtained for various pH and sucrose gradients. For an infinite source of sucrose on side 1 of the membrane and zero initial concentration of sucrose on side 2 of the membrane, it is assumed that the concentration on side 1 of the membrane remains unchanged from the initial concentration. The concentration of sucrose on side 2 balanced by a buffered pH gradient at equilibrium is given by the expression

$$C_2^s = C_1^s 10^{\Delta pH}, \tag{9}$$

where $\Delta pH = pH_2 - pH_1$. In the presence of a limited source of sucrose on side 1 of the membrane as in the actuator, sucrose is transported through the membrane until the concentration gradient of sucrose balances the applied proton gradient. The balancing concentration as a function of the volume ratios. The number of moles of the sucrose on side 1 and side 2 at equilibrium is denoted by $n_1^s|_{eq}$ and $n_2^s|_{eq}$, respectively. The volume of the chamber on side 1 and side 2 of the membrane is represented by V_1 and V_2 and the volume ratio by $Vr = V_1/V_2$. At equilibrium condition the number of moles in the system is constrained by

$$n_1^s\big|_{eq} + n_2^s\big|_{eq} = V_1 c_1^s \tag{10}$$

The equilibrium concentration on side 2 of the membrane from Eq. 9 is rewritten in terms of the number of moles of the sucrose on both the sides of the membrane,

$$n_2^s\big|_{eq} = n_1^s\big|_{eq} \frac{V_r}{10^{\Delta pH}} \tag{11}$$

The concentration on side 1 at equilibrium condition is obtained in terms of the initial sucrose concentration on side 1 of the membrane by substituting the mass balance expression in Eq. 10 into Eq. 11

$$c_1^s\big|_{eq} = c_1^s \frac{V_r}{V_r + 10^{\Delta pH}}. \tag{12}$$

Similarly, the concentration of sucrose at equilibrium on side 2 of the membrane is

Bioderived Smart Materials, Table 1 Equilibrium concentration of sucrose on side 2 of the membrane in the actuator for pH4/pH7 applied across the membrane assembly

c_1^s mM (Initial)	1	10	15	20
$c_2^s\|_{eq}$ mM	90	900	1,360	1,810

$$c_2^s\big|_{eq} = c_1^s \frac{10^{\Delta pH} V_r}{V_r + 10^{\Delta pH}}. \tag{13}$$

The equilibrium concentration of sucrose on side 2 of the membrane $c_2^s|_{eq}$ due to an imposed initial sucrose concentration (c_1^s) is computed for a volume ratio $V_r = 100$ in Table 1. It is observed from the table that the proton gradient concentrates sucrose on side 2 of the membrane. From the applied proton gradient and different initial concentration of sucrose on side 1 of the membrane, the membrane with cotransporter will develop different osmotic pressures on side 2 of the membrane within a finite duration. This analysis shows the feasibility to use concentration of sucrose as a control variable in the actuator with a cotransporter for modulating the force and deformation. A prototype chemomechanical actuator similar to the schematic in Fig. 5a was fabricated to demonstrate the concept of microhydraulic actuation using bioderived membranes. The porous glass plate is attached to a 500-μm Kapton film and sandwiched in between the two chambers with a rubber gasket. This serves as the supporting substrate for forming suspended BLM with SUT4 transporters. A clamping plate screwed onto the top chamber holds the PET (polyethylene terepthalate) coverplate that deforms out of the chamber due to increasing pressure from osmotic regulation across the two chambers. The volumes of the chambers are designed to be 0.54 ml and 50 ml so that the larger chamber resembles a reservoir. The procedure to assemble the actuator follows the description in Sundaresan and Leo [22]. The BLM is formed from POPS and POPE lipids by painting method and the proteins are reconstituted by vesicle fusion method. In this method, 10 μl of lipids (POPS:POPE mixed in 3:1 w/w ratio and dissolved in n-decane at 40 mg/ml) is painted on

the surface and allowed to dry for 10 min under a stream of nitrogen. This is followed by the addition of 10 μl of SUT4 transporters in liposomes suspended in pH7.0 medium. The porous substrate with added components are allowed to stand in air for 15 min and assembled into the actuator.

The assembled prototype actuator is characterized by applying different concentrations of sucrose in the bottom chamber and measuring the deformation of the coverplate. As a control study, different baseline tests with a key ingredient left out is tested. The different baseline tests performed on the actuator are:

1. BLM with pH7.0 buffer on both sides of the membrane
2. BLM with pH4.0/pH7.0 gradient without sucrose dissolved in the bottom chamber
3. BLM with pH4.0/pH7.0 gradient with 5 mM sucrose dissolved in the bottom chamber

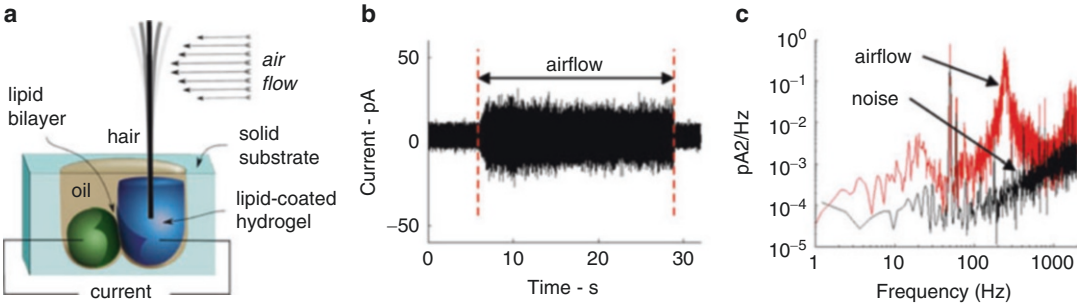
The results from baseline tests on the prototype shown in Sundaresan and Leo [22] demonstrate that there is no deformation in the coverplate for the case with pH7.0 buffer on both the sides of the membrane. Similarly, the coverplate does not show an appreciable deformation for a pH gradient applied to the membrane without sucrose. In the third baseline, pH gradient is applied across a BLM without SUT4 and 5 mM sucrose is added to the bottom chamber. It is observed that the coverplate deforms into the top chamber and is attributed to the migration of water molecules from the top chamber to the bottom chamber driven by the osmotic gradient.

The next set of experiments are performed on an actuator assembled with pH4.0/pH7.0 gradient applied across the BLM with SUT4 and different concentrations of sucrose. Representative results from these experiments are shown in Fig. 5c. From the experimental results in Fig. 5c, it is observed that the deformation in the actuator and rate of deformation increases with the initial sucrose concentration on side 1 of the membrane. The performance metrics of this actuator of this bioderived membrane-based actuator discussed in Sundaresan and Leo [22] compares well with soft polymeric electrochemomechanical actuators [27]. Thus, this demonstration follows the theoretical discussion that varying the sucrose concentration across the SUT4 cotransporter results in an osmotic gradient and leads to volumetric strain in the flexible wall of the actuator.

Hair Cell Sensors

Sensory hairs, or hair cells, are one of the most common forms of transducers found in nature. In general, a hair cell is type of sensory receptor that uses protruding structures, called cilia, to probe the surrounding environment. Sharing this basic trait, natural hair cells display tremendous diversity in their morphologies, mechanical properties, and physiological functions. As a result, animals use hair cells to detect sound, pressure, flow, vibration, chemical species, and even position (inertial tilt). Motivated by the diversity of stimuli-responsive behavior of natural hair cells, several groups have used synthetic active materials to engineer hair cell-inspired sensors. Most notably, Liu's research group [28] pioneered multiple generations of artificial cilia that used microfabricated cilia mounted on strain gauges or polymeric force-sensitive resistors (FSRs) to create an electrical response in response to flow or touch. However, living hair cells such as mammalian outer hair cells (OHCs) rely heavily on soft biological materials, namely, cell membranes and stretch-activated ion channels to sense deflections the cilia.

Building on recent advances to assemble and encapsulate an artificial cell membrane within a flexible substrate, Sarles et al. [29] demonstrated that a gel-supported lipid bilayer could be used to build a new type of membrane-based hair cell-inspired sensor. As shown in Fig. 6a, the hair cell consists of synthetic cilium (hair) that features an electrolyte-swollen polymeric gel at lower end and which is held vertically within one compartment of a nonswelling solid substrate. A second, lipid-encased aqueous volume contained in the neighboring compartment connects to the gel at the base of the hair to form a lipid bilayer. Both compartments also contain immiscible oil needed



Bioderived Smart Materials, Fig. 6 A membrane-based hair cell (**a**) uses the mechanoelectrical properties of a lipid bilayer to transduce airflow into current (**b**). (**c**) The frequency content of the measured signal correlates to the mechanical vibrations of the hair

for promoting phospholipid self-assembly at the oil–water interfaces. The performance of the hair cell sensor was characterized by measuring the current through the membrane (using conductive probes inserted into each aqueous volume) in response to the applied airflow across the hair. The results of this initial study demonstrated that the membrane-based hair cell generates an increase in the amplitude of the current through the bilayer in response to flow-induced vibration of the synthetic hair (Fig. 6b). Also, the power spectral density revealed that the frequency content of the sensing current (Fig. 6c) differs from that of the background noise evidence that the mechanical vibrations of the hair contribute to the measured current. Unlike living hair cells that use stretch-activated ion channels to provide both static and oscillatory sensing, this system only contained an Ac response (i.e., static deflection of the hair did not result in a change in the current). Yet, these initial responses did not explain the source of the current. By systematically varying the applied potential across the bilayer and holding the area of the membrane constant during the experiment, the authors discovered that the sensing current increased linearly with respect to the voltage. This result provided proof that a change in the electrical capacitance of the membrane, caused by bending in the bilayer, is the source of the measured current. In this work, the sensing currents ranged from 10 to 100 pA, depending on the speed of airflow, the length of the hair, and the transmembrane potential. Current efforts to increase the sensitivity of the sensor, provide static and directional responses, and fully encapsulate the liquid contents are underway.

Concluding Remarks

This entry presents a novel framework to develop smart materials using biomolecular components extracted from cell membranes. The bioderived smart materials use ion and fluid transport through protein transporters embedded in an impervious membrane and couple one or more of the following – chemical, electrochemical gradients across the membrane, mechanical stretch in the membrane and light energy incident on the membrane. This entry discusses a unique bioderived actuator that uses sucrose transport through a proton-sucrose transporter to generate volumetric strain. It is shown that the blocked force generated by these materials resemble soft polymeric electrochemomechanical actuators. We also present a novel electromechanical sensing concept using the BLM and demonstrate the potential of this device as a mechanical flow sensor. In the bioderived smart material systems described in this entry, two methods to fabricate the bioderived membrane are presented. Among the two methods, the first method utilizes the self-assembly of phospholipids in a pore to form a BLM with proteins. The second method, inspired by DIB, demonstrates an innovative method to fabricate a durable bilayer membrane, demonstrates the force and flow sensing, and provides a

rugged framework for future development in this field. The framework for building bioderived smart materials and their application as sensors, chemomechanical actuators, and energy harvesting devices discussed in this entry benefits from a wide variety of transporters that can be extracted from cell membranes and offers innovative approaches to couple various forms of energy.

Cross-References

- ▶ Biomimetic Flow Sensors
- ▶ Biomimetics
- ▶ Biosensors
- ▶ Molecular Dynamics Simulations of Nano-biomaterials
- ▶ Organic Actuators
- ▶ Organic Bioelectronics
- ▶ Synthetic Biology

References

1. Montal, M., Mueller, P.: Formation of bimolecular membranes from lipid monolayers and a study of their electrical properties. Proc. Natl. Acad. Sci. U. S. A. **69**(12), 3561–3566 (1972)
2. Ottova, A.L., Ti Tien, H.: Self-assembled bilayer lipid membranes: from mimicking biomembranes to practical applications. Bioelectrochem. Bioenerg. **42**(2), 141–152 (1997)
3. Sundaresan, V.B., Homison, C., Weiland, L.M., Leo, D. J.: Biological transport processes for micro-hydraulic actuation. Sens. Actuators B Chem. **123**, 685–695 (2007)
4. Mueller, P., Rudin, D.O., Tien, H.T., Wescott, W.C.: Reconstitution of excitable cell membrane structure in vitro. Circulation **26**(5), 1167–1171 (1962)
5. Needham, D., Haydon, D.A.: Tensions and free energies of formation of "solventless" lipid bilayers measurement of high contact angles. Biophys. J. **41**(3), 251–257 (1983)
6. Poulin, P., Bibette, J.: Adhesion of water droplets in organic solvent. Langmuir **14**(22), 6341–6343 (1998)
7. Haydon, D.: Properties of lipid bilayers at a water-water interface. J. Am. Oil Chem. Soc. **45**(4), 230–240 (1968)
8. Fettiplace, R., Andrews, D.M., Haydon, D.A.: The thickness, composition and structure of some lipid bilayers and natural membranes. J. Membr. Biol. **5**(3), 277–296 (1971)
9. Tien, H.T., Wurster, S.H., Ottova, A.L.: Electrochemistry of supported bilayer lipid membranes: background and techniques for biosensor development. Bioelectrochem. Bioenerg. **42**, 77–94 (1997)
10. Tien, H.T.: Self-assembled lipid bilayers as a smart material for nanotechnology. Mater. Sci. Eng. C **3**, 7–12 (1995)
11. White, S.H.: Analysis of the torus surrounding planar lipid bilayer membranes. Biophys. J. **12**(4), 432–445 (1972)
12. Singer, S.J., Nicolson, G.L.: The fluid mosaic model of the structure of cell membranes. Science **175**(4023), 720–731 (1972)
13. Funakoshi, K., Suzuki, H., Takeuchi, S.: Lipid bilayer formation by contacting monolayers in a microfluidic device for membrane protein analysis. Anal. Chem. **78**(24), 8169–8174 (2006)
14. Holden, M.A., Needham, D., Bayley, H.: Functional bionetworks from nanoliter water droplets. J. Am. Chem. Soc. **129**(27), 8650–8655 (2007)
15. Hwang, W.L., Holden, M.A., White, S., Bayley, H.: Electrical behavior of droplet interface bilayer networks: experimental analysis and modeling. J. Am. Chem. Soc. **129**(38), 11854–11864 (2007)
16. Bayley, H., Cronin, B., Heron, A., Holden, M.A., Hwang, W.L., Syeda, R., Thompson, J., Wallace, M.: Droplet interface bilayers. Mol. Biosyst. **4**(12), 1191–1208 (2008)
17. Sarles, S.A., Leo, D.J.: Regulated attachment method for reconstituting lipid bilayers of prescribed size within flexible substrates. Anal. Chem. **82**(3), 959–966 (2010)
18. Sarles, S.A., Leo, D.J.: Membrane-based biomolecular smart materials. Smart Mater. Struct. **20**(9), 094018 (2011)
19. Sarles, S.A., Stiltner, L.J., Williams, C.B., Leo, D.J.: Bilayer formation between lipid-encased hydrogels contained in solid substrates. ACS Appl. Mater. Interfaces **2**(12), 3654–3663 (2010)
20. Oster, G., Wang, H.: Rotary protein motors. Trends Cell Biol. **13**(3), 114–121 (2003)
21. Hess, H., Bachand, G.D., Vogel, V.: Powering nanodevices with biomolecular motors. Chem. A Eur. J. **10**(9), 2110–2116 (2004)
22. Sundaresan, V.B., Leo, D.J.: Modeling and characterization of a chemomechanical actuator using protein transporters. Sens. Actuators B Chem. **131**(2), 384–393 (2008)
23. Kedem, O., Katchalsky, A.: A physical interpretation of the phenomenological coefficients of membrane permeability. J. Gen. Physiol. **45**(1), 143–179 (1961). doi:10.1085/jgp.45.1.143.arXiv. http://www.jgp.org/cgi/reprint/45/1/143.pdf
24. Schultz, S.G., Curran, P.F.: Coupled transport of sodium and organic solutes. Physiol. Rev. **50**(4), 637–718 (1970). arXiv:http://physrev.physiology.org/cgi/reprint/50/4/637.pdf
25. Gyftopoulos, E.P., Berreta, G.P.: Thermodynamics: Foundations and Applications. Macmillan, New York (1996)
26. Onsager, L.: Reciprocal relations in irreversible processes. II. Phys. Rev. **38**(12), 2265–2279 (1931). doi:10.1103/PhysRev.38.2265
27. Akle, B., Bennett, M., Leo, D.J.: Development of a novel electrochemically active membrane and "smart" material based vibration sensor/damper. Sens. Actuators A Phys. **126**(1), 173–181 (2006)

28. Liu, C.: Micromachined biomimetic artificial haircell sensors. Bioinspir. Biomim. **2**(4), S162 (2007)
29. Sarles, S.A., Madden, J.D.W., Leo, D.J.: Hair cell inspired mechanotransduction with a gel-supported, artificial lipid membrane. Soft Matter **7**(10), 4644–4653 (2011)

Biodistribution

► In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles

Bio-FET

► Nanostructure Field Effect Transistor Biosensors

Biofilms in Microfluidic Devices

Suresh Neethirajan[1], David Karig[2], Aloke Kumar[3], Partha P. Mukherjee[4], Scott T. Retterer[5,6,7] and Mitchel J. Doktycz[5,6]
[1]School of Engineering, University of Guelph, Guelph, ON, Canada
[2]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[3]Department of Mechanical Engineering, University of Alberta, AB, Canada
[4]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[5]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[6]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[7]Department of Electrical Engineering and Computer Science, University of Tennessee Knoxville, Knoxville, TN, USA

Synonyms

Floccules; Microbial aggregations

Definition

Biofilms are aggregations of microbes that are encased by extracellular polymeric substances (EPS) and adhere to surfaces or interfaces. Biofilms exist in a very wide diversity of environments, and microfluidic devices are being increasingly utilized to study and understand their formation and properties.

Overview

Microbes often form aggregates on interfaces, and due to a production of EPS, the aggregates become encased in a matrix [1]. Though microbes in a biofilm are physiologically distinct from bacteria growing in a free swimming state (planktonic bacteria), biofilm growth is a complex process that is typically initiated by planktonic bacteria themselves. Biofilm growth is initiated with bacterial adhesion to a surface, followed by events such as growth, EPS secretion, and morphological and physiological changes. Microbial biofilms are excellent examples of multi-scale phenomena. Cell-to-cell communication, which plays a role in biofilm formation, is molecular in nature, but occurs over a scale of several cells. Adhesion events occur at the nanometer scale and are mediated by pili or flagella, while the cells themselves are typically micron-sized. Finally the biofilms themselves typically range between 10 and 1,000 μm in thickness. These length scales are compatible with the scale of microfluidic devices, thus making such tools useful for exploring the spatiotemporal properties of biofilms. Moreover, microfluidic devices are often optimized for online optical monitoring and/or incorporation of sensors. These factors make microfluidic devices appropriate for studying biofilms. For example, the effect of miniscule changes in molecular cues (such as nanomolar concentrations) can be characterized and studied easily with microfluidic devices. Another advantage of microfluidics is that they enable the precise control of the microenvironment, thereby allowing biofilms to be subjected to controlled external stimuli. Thus, when molecular cues or signals are externally applied, minute responses in the biofilm can be

effectively studied. The use of microfluidics offers distinct advantages for fundamental studies regarding the nature, properties, and evolution of microbial biofilms. Beyond being a platform for such studies, microfluidics is being increasingly applied toward miniaturized device creation. In this entry, some of the basic methodologies involved in biofilm studies in microfluidic devices are first discussed, followed by a discussion on some of the key findings reported in this area.

Methodology

Microfluidics

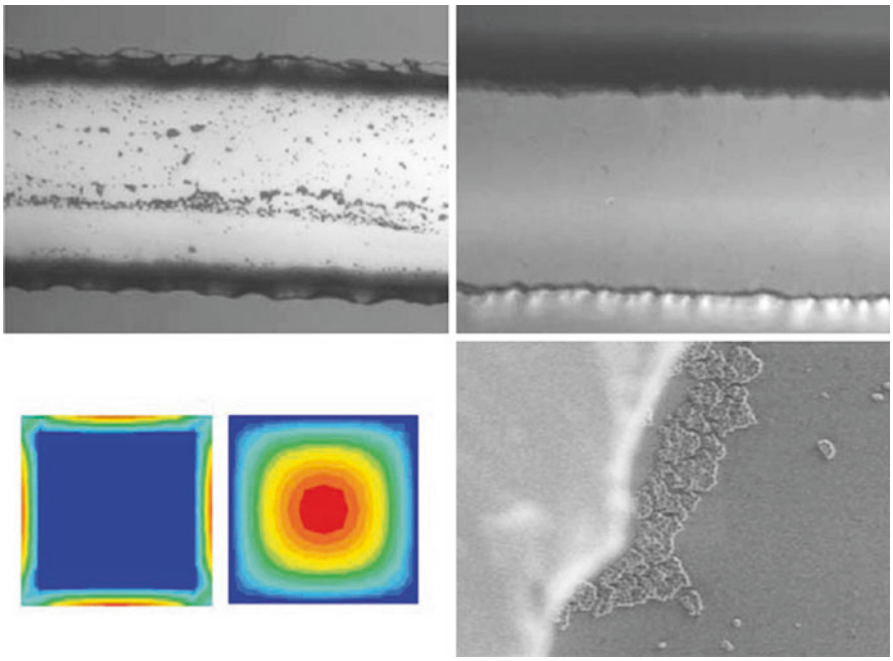
Microfluidics is the precise control and manipulation of fluids contained to miniaturized channels (typical length scale <100 μm). In the microfluidics regime, fluid flow is laminar, and flow is dominated by Stokes drag and surface tension effects. The study of biofilms in microfluidic devices typically requires device design and fabrication, controlled microbial growth, and analysis. Microfluidic device fabrication by itself is a significantly evolved science. There are several different techniques for device fabrication including, but not limited to, various micromachining processes and polymer-based soft-lithography. Complex structures such as micropumps, microvalves and mixers can also be incorporated on microfluidic devices with the help of different microfabrication techniques. The techniques of microfabrication are beyond the scope of this article and interested readers can refer to one or more manuscripts on this topic [2, 3].

Biofilms

A wide range of microbial species produce biofilms. Once the appropriate device is fabricated, controlled biofilm growth in the device may be desirable. For controlled in situ biofilm growth in a microfluidic device, a dilute microbial culture may be introduced into the microfluidic device. The diluted samples are made from liquid microbial cultures once the culture achieves a specified optical density (OD). A fluid port usually provides a means for the introduction of the inoculum into the device. Depending on the microorganism and its microenvironment and the specific experimental need, proper biofilm growth can require anywhere between a few hours to several days. During this interval, it might be desirable to control environmental conditions such as temperature and humidity. In such circumstances, the device is typically housed in an incubator. Depending on the organism, aerobic or anaerobic environment may be necessary. For microfluidic devices made from PDMS or similar permeable polymers, maintaining an aerobic environment is usually not an issue. On the other hand, to maintain anaerobic conditions typically necessitates more complex device fabrication techniques. Analysis of biofilm formation may be done by various forms of microscopy. Microfluidic devices usually lend themselves to optical microscopy with considerable ease – one of the reasons why it is popular as a diagnostic setup. Use of thin-walled chambers and optically clear materials in microfluidic devices allows probing through optical means and use of high-magnification lenses. Confocal laser scanning microscopy (CLSM) has been demonstrated as a tool for integrating microfluidics in studying the evolution and structural heterogeneity of biofilms [4]. Other forms of microscopy may also be employed such as scanning electron microscopy (SEM) and atomic force microscopy (AFM). Besides microscopy, micro- and nanosensors can also be easily incorporated into a microfluidic device for in situ monitoring. For example, micro-electrodes fabricated on a glass substrate can be incorporated in a microfluidic device for electrochemical impedance spectroscopy studies.

Key Findings

Microfluidic devices, in conjunction with microscopy and other analytical techniques, enable flexible and novel approaches for probing the multiple determinants of biofilm formation. Two factors shaping the dynamics of biofilm formation are fluid dynamics and cell phenotype. Fluid dynamics determine shear forces that govern cell attachment and detachment rates, which can alter whether or not a biofilm will form, biofilm depth, biofilm



Biofilms in Microfluidic Devices, Fig. 1 Microfluidic devices can be engineered to produce well-defined flow structures and shear rates. Such devices can be used to investigate hydrodynamic influences on biofilms (From Lee et al. [5]. Reproduced with kind permission from Springer Science and Business Media)

density, and surface coverage. On the other hand, cell phenotype dictates important processes such as EPS production, growth rate, and flocculation, all of which can also have profound effects on biofilm structure and function. Cell phenotype, in turn, is a function of both environmental cues and cell-cell communication. Accordingly, recent studies have harnessed microfluidic technology to probe the effects of fluidic dynamics, cell phenotype, and cell-cell communication. Mathematical modeling efforts have integrated these findings and guided further investigations.

Influence of Fluid Dynamics on Biofilms

Several studies have examined the effects of hydrodynamics on biofilm development. For example, microfluidic devices were used by Lee et al. [5] to study the influences of hydrodynamics of local microenvironments of *Staphylococcus epidermis* biofilm formation (Fig. 1). They observed that at high flow velocity, the cells formed an elongated biofilm morphology, and at low fluid velocity clump-like multilayered biofilms were produced. The results of this study indicate that microfluidic devices with embedded microvalves can perhaps be used for screening the effects of therapeutic reagents, and as novel tools for developing predictable in vitro models of biofilm-related infections. Also, Rusconi et al. [6] studied suspended filamentous biofilms in a microfluidic device. Experiments with several bacterial strains under high flow inside the miniaturized channels demonstrated the link between the extracellular matrix and the development of biofilm structures. Secondary vertical motion from the numerical simulations of the flow in the curved channels of the microfluidic device proved that the hydrodynamic forces were key in influencing the formation of suspended biofilm filamentous structures. Another study by Richter et al. [7] examined the influence of shear stress on the growth and structure of fungal biofilms using microfluidic systems. Electrode structures were incorporated into a microfluidic device for performing cellular dielectric spectroscopy, enabling real-time, noninvasive quantification of cell morphology changes. Increase in shear stress caused a significant change in the biofilm formation patterns, and the addition of

amphotericin B resulted in distinct dynamic behavior of the biofilm.

Effect of Cell Phenotype on Attachment

In addition to hydrodynamics, cell phenotype plays an important role in biofilm formation. For example, after initial attachment, pili, flagella, and adhesins can help cells adhere to surfaces. Characterizing these determinants of cell attachment is critical for understanding the early stages of biofilm development. Using a microfluidic device, De La Fuente et al. recently examined the roles of different *Xylella fastidiosa* pili in determining cell adhesion at different flow rates [8]. To do so, they compared different genetic variants: wild-type cells, cells with only type I pili, cells with only type IV pili, and cells with no pili. They subjected these variants to different drag forces by exposing the cells to different fluid flow rates. The results enabled quantification of the role of the different pili in attachment, and the adhesion force values were within the range of adhesion forces determined by AFM and by laser tweezers for other microbes.

Cellular-Communication Inside Microfluidic Systems

Several cell processes relevant to biofilm formation, including dispersion, EPS secretion, and lipid secretion, are often regulated by cell-cell communication. Thus, several studies have focused on studying intercellular communication in biofilm contexts. Specifically, many microbes communicate through the process of quorum sensing. Quorum sensing (QS) enables a group of cells to measure their local population density through the synthesis of and response to small signal molecules that can pass from cell to cell. While quorum sensing regulates several microbial behaviors that influence biofilm formation, biofilm characteristics in turn strongly impact the efficiency of quorum sensing and signal transport, thus setting up an interplay between quorum sensing and biofilm formation. Consequently, microfluidic studies have been used to examine quorum sensing in a biofilm context.

Timp et al. [9] studied genetically engineered cells in microfluidic devices to explore the fundamental principles of quorum sensing in biofilms. Engineered cells were flowed through a microfluidic channel into a region where optical traps were used to position cells into defined patterns. The patterned cells were then enclosed in a hydrogel to mimic a biofilm. "Transmitter" cells synthesized an N-Acyl Homoserine Lactone (AHL) communication signal which was detected by "receiver" cells that fluoresced in response. Activation of the receiver cells by the transmitters was found to be heavily dependent on the hydrodynamics around the biofilm mimic. In particular, low flow rates corresponding to diffusion-dominated transport yielded efficient communication between transmitters and receivers, but this communication began to break down at high flow rates corresponding to convection-dominated transport. These results indicate that the density of bacteria necessary to constitute a "quorum" depends on the hydrodynamic properties of the environment. A similar conclusion was reached by Connell et al. [10] who studied QS in fabricated picoliter-scale microcavities. Specifically, the authors found that they could achieve a sixfold increase in QS-dependent gene expression by bacteria trapped within their microcavities when they reduced the external flow rate from 250 to 5 $\mu L\ min^{-1}$.

Another interesting aspect of QS highlighted by the Connell study was the importance of cell density, rather than total cell count as a factor governing QS behavior. In particular, because the authors were able to capture small numbers of cells in small volumes, they were able to control cell density independent of cell population size. As a result, they were able to show that as few as 150 cells are capable of exhibiting QS behavior, provided that the cells are constrained such that cell density remains relatively high. This is important, since it demonstrates the potential for small bacterial communities to exhibit some of the properties (e.g., antibiotic resistance), which are typical of biofilms. In a related study, Boedicker et al. used microfluidics to examine the QS behavior of small numbers of bacteria trapped in poly(dimethylsiloxane) (PDMS) wells [11]. In this study, the authors were able to trap as few as one to two cells in each well, and while the

majority of cells needed several rounds of cell division to initiate QS, on occasion the authors found single cells expressing QS controlled genes. Again, this result showed that it is bacterial density, rather than bacterial cell numbers, that is of fundamental importance to the onset of QS behavior. Another aspect of QS and microbial aggregation was investigated by Park et al. [12], who devised a microfluidic device resembling a maze. They investigated the growth of *Escherichia coli* cells in the device. They found that the device geometry affected aggregation of the cells. Their results led them to conclude that self-attraction in microbes could allow them to easily exceed critical QS densities.

Collectively, these studies exemplify the flexibility offered by microfluidic systems to precisely define the cell's microenvironment, including cell position, confinement geometry, and fluid flow rate. These studies have begun to shed light on fundamental questions such as when quorum sensing is truly used to sense local population as opposed to merely sensing environmental conditions such as the local diffusion rate or degree of confinement. Thus, microfluidics is enabling the study of biological principles that would otherwise not be possible using conventional approaches in biology. We next focus on how data from these unique experimental studies can be used to validate mathematical models of biofilm formation and lead to new insights.

Mathematical Biofilm Modeling and Microfluidics

Thus far several determinants of the process of biofilm formation including fluid dynamics, cell phenotype, and cell-cell communication have been discussed. The complex nature of biofilm formation makes mathematical modeling challenging, yet critical for developing a fundamental quantitative understanding of the underpinnings of the process. Proper mathematical models take into account the multistaged process with its broad range of time and length scales. Mathematical development of biofilms has been long pursued with substantial developments having taken place in the past two decades. Klapper and Dockery present a comprehensive review of mathematical model development for biofilms [13]. Today, rapid developments in the biofilm community are pushing the need for better and comprehensive models; microfluidics presents itself as an invaluable aid to the modeling community. Microfluidics not only affords researchers to explore in real time biofilm growth and dynamics, but it also allows for a controlled microenvironment that can be adjusted to suit the need of the user. A controlled microfluidic environment amenable to probing by various sensors can allow researchers to pursue valuable validation experiments for proposed models. In one such study, Janakiraman et al. [14] used a microfuidic device to study biofilm development of *Pseudomonas aeruginosa* PA14, where mass and species transport effected biofilm development and vice versa. Their use of a microfluidic device allowed assessment of biofilm growth variation with shear rates, finally leading to model validation. The validated model, which took into account mass transport and its effect on quorum sensing, allowed valuable conclusions such as that flow rate could be used to turn on and turn off quorum sensing within the biofilm. Experimentally validated models like the one presented by Janakiraman et al. [14] allow a greater understanding of biofilm kinetics. In another study, Volfson et al. [15] used microfluidic devices to investigate spatial ordering and self-organization of microbes in a microfluidic device. The authors were investigating the role of contact biomechanics in the formation of dense colonies and toward this end they used a controlled microfluidic device to monitor the two-dimensional growth of motile microbes. The experimental investigations were combined with discrete element simulations (DES), and the authors showed that biomechanical interactions could lead to highly ordered structures in a microbial colony. Such investigations can help in the complete understanding of the role of various environmental and self-generated factors that play a role in biofilm formation. Deeper quantitative understanding of biofilm formation, resulting from the interplay between modeling and experimental validation with microfluidics, will drive progress in various applications.

Other Uses of Lab-on-a-Chip Technologies

Lab-on-a-chip (LoC) technologies refer to a suite of technologies that have evolved primarily over the last decade, where complex operations such as cell culture and sensing are integrated in a miniaturized platform. Apart from sensors, LoC technologies allow microfluidic systems to be interfaced with other systems such as micro-electro-mechanical (MEMs) and opto-electric systems. MEMs devices in LoC systems have been harnessed for various studies on microbial biofilms. Several different MEMs devices have been used to pursue biofilm-related studies, thus providing key insights into fundamental phenomena such as cellular self-organization in biofilms and the role of motility [16]. Purely electrical systems have also been incorporated into LoC systems and have been employed to investigate physiological heterogeneity within microbial biofilms [17]. Emerging lab-on-a-chip technologies can also benefit biofilm research. For example, novel opto-electric techniques [18] might find a use in biofilm engineering. Rapid electro kinetic patterning (REP) is an opto-electric technique, which can be used to manipulate micro- and nanosized objects noninvasively in a microfluidic device. REP has already been used to capture an aggregation of *Shewanella oneidensis* MR-1 in a microfluidic device and place it at user-defined intervals using an infrared laser and electric fields [18]. Such an approach could be used to trigger and organize the formation of beneficial biofilms.

Other than such appeal, LoC technologies are also being employed to fabricate miniaturized devices for applications such as current production and miniaturization of assays for clinical testing.

Future Perspectives

Microfluidic systems integrated with mechanical and/or electrical transducers and sensor components open up new ways to understand biofilm formation and bacterial surface interactions. Biofilms subjected to external stimuli through molecular cues during the growth and establishment stage can be characterized for their mechanical properties more accurately using microfluidic systems. Qualitative assessment of temporal and spatial patterns of biofilm formation against antimicrobial actions (chemical method) or electric field interference [19] or a physical disruption can be effectively studied using microfluidic systems.

Acknowledgments A. Kumar performed the work as a Eugene P. Wigner Fellow at the Oak Ridge National Laboratory (ORNL). The authors acknowledge research support from the US Department of Energy (US DOE) Office of Biological and Environmental Sciences. ORNL is managed by UT-Battelle, LLC, for the US DOE under Contract no. DEAC05-00OR22725.

Cross-References

▶ Micro- and Nanofluidic Devices for Medical Diagnostics
▶ Micro/Nano Transport in Microbial Energy Harvesting

References

1. Costerton, J.W., Lewandowski, Z., Caldwell, D.E., Korber, D.R., Lappinscott, H.M.: Microbial biofilms. Annu. Rev. Microbiol. **49**, 711–745 (1995)
2. Becker, H., Gartner, C.: Polymer microfabrication methods for microfluidic analytical applications. Electrophoresis **21**(1), 12–26 (2000)
3. Madou, M.: Fundamentals of Microfabrication: The Science of Miniaturization, 2nd edn. CRC Press, Boca Raton (2002)
4. Yawata, Y., Toda, K., Setoyama, E., Fukuda, J., Suzuki, H., Uchiyama, H., Nomura, N.: Bacterial growth monitoring in a microfluidic device by confocal reflection microscopy. J. Biosci. Bioeng. **110**(1), 130–133 (2010)
5. Lee, J.-H., Kaplan, J., Lee, W.: Microfluidic devices for studying growth and detachment of *Staphylococcus epidermidis* biofilms. Biomed. Microdevices **10**(4), 489–498 (2008)
6. Rusconi, R., Lecuyer, S., Guglielmini, L., Stone, H.A.: Laminar flow around corners triggers the formation of biofilm streamers. J. R. Soc. Interface **7**(50), 1293–1299 (2010)
7. Richter, L., Stepper, C., Mak, A., Reinthaler, A., Heer, R., Kast, M., Bruckl, H., Ertl, P.: Development of a microfluidic biochip for online monitoring of fungal biofilm dynamics. Lab Chip **7**(12), 1723–1731 (2007)
8. De La Fuente, L., Montanes, E., Meng, Y., Li, Y., Burr, T.J., Hoch, H.C., Wu, M.: Assessing adhesion forces

of type I and type IV pili of *Xylella fastidiosa* bacteria by use of a microfluidic flow chamber. Appl. Environ. Microbiol. **73**(8), 2690–2696 (2007)
9. Timp, W., Mirsaidov, U., Matsudaira, P., Timp, G.: Jamming prokaryotic cell-to-cell communications in a model biofilm. Lab Chip **9**(7), 925–934 (2009)
10. Connell, J.L., Wessel, A.K., Parsek, M.R., Ellington, A.D., Whiteley, M., Shear, J.B.: Probing prokaryotic social behaviors with bacterial "lobster traps". MBio **1**(4), e00202–e00210 (2010)
11. Boedicker, J.Q., Vincent, M.E., Ismagilov, R.F.: Microfluidic confinement of single cells of bacteria in small volumes initiates high-density behavior of quorum sensing and growth and reveals its variability. Angew. Chem. Int. Ed. Engl. **48**(32), 5908–5911 (2009)
12. Park, S., Wolanin, P.M., Yuzbashyan, E.A., Silberzan, P., Stock, J.B., Austin, R.H.: Motion to form a quorum. Science **301**(5630), 188 (2003)
13. Klapper, I., Dockery, J.: Mathematical description of microbial biofilms. SIAM Rev. **52**(2), 221–265 (2010)
14. Janakiraman, V., Englert, D., Jayaraman, A., Baskaran, H.: Modeling growth and auorum sensing in biofilms grown in microfluidic chambers. Ann. Biomed. Eng. **37**(6), 1206–1216 (2009)
15. Volfson, D., Cookson, S., Hasty, J., Tsimring, L.S.: Biomechanical ordering of dense cell populations. Proc. Natl. Acad. Sci. U. S. A. **105**(40), 15346–15351 (2008)
16. Ingham, C.J., Vlieg, J.: MEMS and the microbe. Lab Chip **8**(10), 1604–1616 (2008)
17. Stewart, P.S., Franklin, M.J.: Physiological heterogeneity in biofilms. Nat. Rev. Microbiol. **6**(3), 199–210 (2008)
18. Kumar, A., Williams, S.J., Chuang, H.-S., Green, N. G., Wereley, S.T.: Hybrid opto-electric manipulation in microfluidics-opportunities and challenges. Lab Chip **11**(13), 2135–2148 (2011)
19. Kumar, A., Mortensen, N.P., Mukherjee, P.P., Retterer, S.T., Doktycz, M.J.: Electric field induced bacterial flocculation of enteroaggregative *Escherichia coli* 042. Appl. Phys. Lett. **98**(25), 253701–253703 (2011)

Biognosis

▶ Biomimetics

Bioimpedance Sensing

▶ Impedance Sensing Techniques for Biomedical Microdevices

Bio-inspiration

▶ Biomimetics of Optical Nanostructures

Bioinspired and Bacterial Nanorobotics

▶ Nanorobotics for Bioengineering and Medical Interventions

Bioinspired CMOS Cochlea

Andreas G. Katsiamis[1], Emmanuel M. Drakakis[2] and Richard F. Lyon[3]
[1]Toumaz Technology Limited, Abingdon, UK
[2]Department of Bioengineering, The Sir Leon Bagrit Centre, Imperial College London, London, UK
[3]Google Inc, Santa Clara, CA, USA

Synonyms

Bionic ear; Cochlea implant; Gammatone filters; Log-domain; Low-power

Definition

This entry deals with the design and performance evaluation of a new analogue CMOS cochlea channel of increased biorealism. The design implements a recently proposed transfer function [12], namely the One–zero Gammatone Filter (or OZGF), which provides a robust foundation for modeling a variety of auditory data such as realistic passband asymmetry, linear low-frequency tail, and level-dependent gain. Moreover, the OZGF is attractive because it can be implemented efficiently in any technological medium – analogue or digital – using standard building blocks. The channel was synthesized using novel, low-power, Class-AB, log-domain,

biquadratic filters employing MOS transistors operating in their weak inversion regime. Furthermore, the entry details the design of a new low-power automatic gain-control circuit that adapts the gain of the channel according to the input signal strength, thereby extending significantly its input dynamic range. The performance of a fourth-order OZGF channel (equivalent to an eighth-order cascaded filter structure) was evaluated through both detailed simulations and measurements from a fabricated chip using the commercially available 0.35 μm AMS CMOS process. The whole system is tuned at 3 kHz, dissipates a mere 4.46 μW of static power, accommodates 124 dB (at <5 % THD) of input dynamic range at the center frequency, and is set to provide up to 70 dB of amplification for small signals.

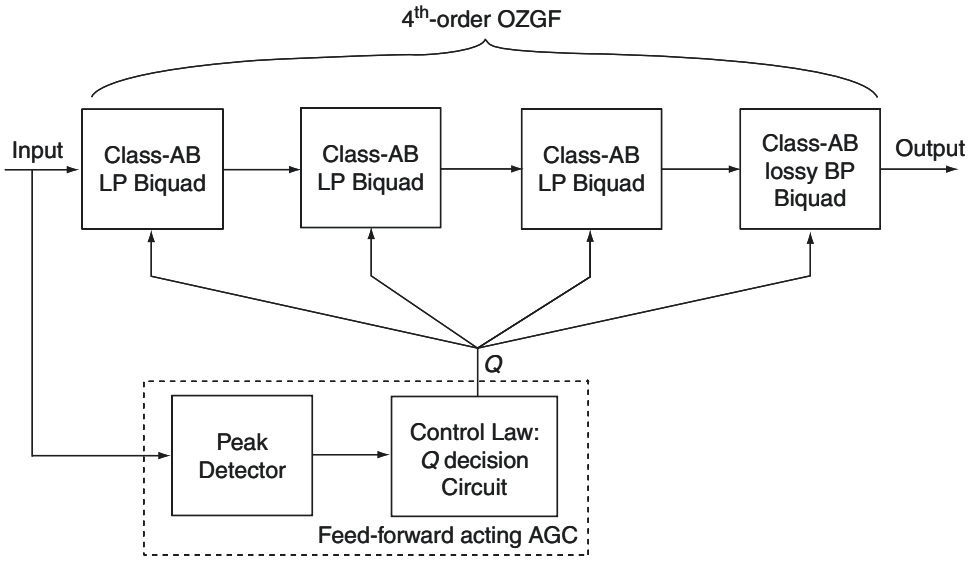
Introduction

The first generations of high-performance cochlea designs were synthesized using g_m-C filters and relied on power-hungry linearization techniques and/or the compressive action of the AGC for extending the input DR. However, recent advances in the field of analogue filter design have led to the development of inherently compressive systems that operate internally in the nonlinear domain while preserving overall input–output linearity. The application of *companding* in filter design resulted in the successful realization of topologies that were able to attain a wider input DR with a lower power-supply requirement compared to traditional g_m-C filters employing linearized transconductors. These companding filters or processors belong to the more general class of ELIN (Externally-Linear–Internally-Nonlinear) systems [21], and their systematic synthesis is articulated in the pioneering works of Frey [7] and Tsividis [20]. It is worth noting that the need for inherently compressive filters emerged very early in the development process of micropower, high DR cochlea designs and, in fact, a bit earlier than the first 1993 Log-domain paper by Frey [6].

Since Frey's [6] paper, Log-domain circuits progressed significantly with several contributions aiming at increasing the input DR and lowering the quiescent power dissipation. The two most thoroughly studied techniques are: (a) The use of two Class-A filters in a pseudo-differential Class-AB arrangement [8] that increases the DR without spending too much power and (b) the use of an AGC scheme that dynamically changes certain biasing levels of the filter (according to a particular measure of input signal strength) in order to optimize its output SNR and power dissipation – this technique is otherwise known as *syllabic companding* [9] and constitutes indeed a parallel engineering strategy to the one performed by the real cochlea to widen its input DR and maintain signal integrity. It is interesting to note that several mechanisms within the cochlea may include an adjustable "DC bias" in the BM caused by hair-cell interactions that affect the operating point on their detection nonlinearity and their sensitivity [15].

In the following sections, the proposed pseudo-differential, Class-AB, Log-domain implementation of a fourth-order (i.e., a cascade of four, second-order transfer functions) OZGF with the ability to automatically adapt its gain by means of a low-power, AGC circuit acting *in feed-forward* is discussed. The reason a feed-forward scheme was employed (like the one demonstrated in [17]), as opposed to a feedback (like in [5]), was in order to avoid any latency issues and to obtain faster adaptation times.

The AGC senses the strength of the input signal and according to a particular control law sets the Q values of the individual filter stages of the cascade comprising the channel. The action of the AGC helps to accommodate a wider input DR because when the input is small (say below or close to the noise floor), the OZGF adapts its gain and provides significant amplification for the output to be well above noise levels. On the other hand, when the input is large, the OZGF provides low or no amplification so that acceptable distortion levels are maintained at the output. Since, in effect, the AGC is changing the state variables or pole positions of the individual filters, the whole OZGF channel falls under the category of a *high-order, syllabically companding, ELIN, auditory (cochlea) amplifier/processor*.



Bioinspired CMOS Cochlea, Fig. 1 Block diagram of the active fourth-order OZGF channel

Figure 1 depicts the high-level block diagram of our *active* fourth-order OZGF channel architecture.

Clarification: Please note that the term "active" we refer to the channel together with its feed-forward AGC mechanism to distinguish it from the case where the AGC is disconnected (i.e., the OZGF gain is set manually off-chip); the latter case will be simply referred to as "open loop." It should be clarified here that strictly speaking, the term "feed-forward" always refers to an open-loop controller since it does not employ feedback to determine if its input has achieved the desired goal. This means that the system does not observe the output of the processes that it is controlling. Under this strict definition, the implemented OZGF is considered open-loop *irrespective of whether the AGC is active or not*, but during the course of this entry, we shall use the terms "active" and "open loop" to refer to an OZGF channel "with-" and "without-" AGC, respectively.

The Circuits

This section deals with the design of the circuits comprising the fourth-order OZGF channel with its AGC. All device sizes, biasing currents, capacitors, and power supplies are displayed in Table 1 in section "Active OZGF Channel Measured Results."

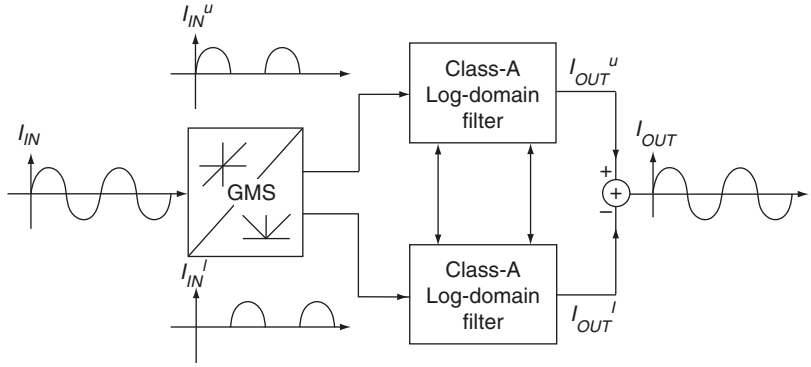
Bioinspired CMOS Cochlea, Table 1 Electrical and device parameters

Topology	Biquads	GMS	LPF	E-cell[a]	OTA[a]
$(W/L)_{PMOS}$	300 µm/1.5 µm				20 µm/1.5 µm
$(W/L)_{NMOS}$	60 µm/8 µm				10 µm/1.5 µm
I_o	20 nA				
I_Q	2 nA–24 nA (i.e., Q between 0.834 and 10)				
I_Z	2 nA (i.e., $0.1I_o$ for a–20 dB DC gain)				
I_{LPF}	20 nA				
$I_{o_control}$	42 nA				
I_{tail}	24 nA				
V_{DD}	1.8 V				
C	20 pF				
C_{LPF}	80 pF				

[a]The reported dimensions correspond to the minimum device sizes of Fig. 13

Pseudo-Differential Class-AB Log-Domain Biquads

Class-AB designs are offered as the most efficient solution for balancing DR and low quiescent power-consumption performance requirements. This is because for small signals, a Class-AB



Bioinspired CMOS Cochlea, Fig. 2 Block diagram of a general pseudo-differential Class-AB architecture. The two vertical *arrows* between the two Class-A filters indicate coupling that ensures correct Class-AB operation

design operates with the quality of a Class-A system, whereas for large signals with the efficiency of a Class-B system.

In Log-domain, one way to design a Class-AB filter is by connecting two Class-A filters in a pseudo-differential arrangement with a signal conditioner at the input. The signal conditioner (a geometric mean splitter (GMS)) ensures that a bidirectional input signal is split into two complementary, unidirectional, positive signals, which are then processed separately by the two Class-A filters. The respective unidirectional processed outputs are subsequently subtracted to form the total bidirectional output, which is a linearly filtered version of the input (see Fig. 2). However, this is not enough to guarantee Class-AB operation unless the following two rules (implied by the vertical, two-sided arrows in Fig. 2) are obeyed [8]:

1. Firstly, a nonzero DC-operating-point solution must always exist for all the state variables (these are currents in Log-domain filter implementations) for any strictly positive, static (i.e., DC) values of the complementary inputs. This can be achieved by ensuring that the derivatives of the state-variables can become equal to zero while the state-variables themselves remain strictly positive.
2. Secondly, all state-variables must remain strictly positive provided that the complementary inputs remain strictly positive and bounded for all time. This can be achieved by enforcing the derivative of a state-variable to be strictly positive in the limit as the variable approaches zero; hence, that variable can never reach zero and thus will stay strictly positive for all time.

In short, the above two rules state that in a transistor-level implementation of a Class-AB Log-domain topology, all devices should carry nonzero and strictly positive currents at all times.

One possible Class-AB-compatible states-space (SS) description for a differential biquad is given below [8]:

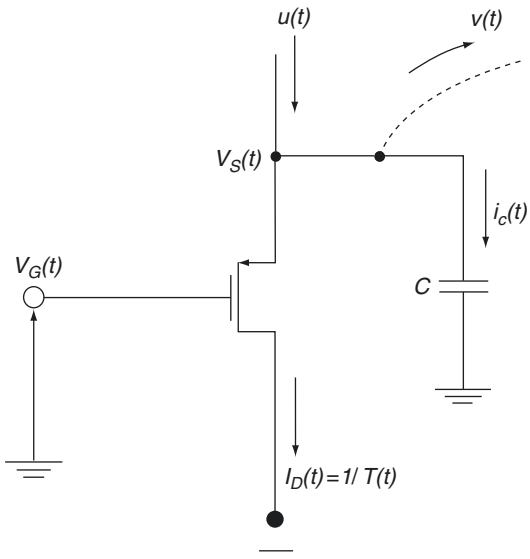
$$\begin{aligned} u &= u_u - u_l \\ \dot{x}_{u1} &= \omega_o\left(-\frac{1}{Q}x_{u1} + x_{l2} + u_u - \frac{x_{u1}x_{l1}}{8}\right) \\ \dot{x}_{u2} &= \omega_o\left(x_{u1} - \frac{x_{u2}x_{l2}}{g}\right) \\ \dot{x}_{l1} &= \omega_o\left(-\frac{1}{Q}x_{li} + x_{u2} + u_1 - \frac{x_{ul}x_{l1}}{g}\right) \\ \dot{x}_{l2} &= \omega_o\left(x_{l1} - \frac{x_{u2}x_{12}}{g}\right) \\ y_{BP} &= x_{u1} - x_{l1}; y_{LP} = x_{u2} - x_{l2} \end{aligned} \tag{1}$$

where u is the differential input, y_{LP}, y_{BP} are the LP and BP differential outputs, respectively, and g is a positive factor that depends on implementation and will be defined later on. The subscripts u and l denote "upper" and "lower" referring to the two individual Class-A filter branches. The interested reader could verify that the derivatives $\dot{x}_{uj}$, $\dot{x}_{lj}$ ($j = 1, 2$) always attain a positive value whenever their respective state-variables tend to zero, or in other words to guarantee the existence of a DC operating point (condition 1). Also, nonlinear cross-coupling terms of the form $x_{ui}x_{lj}$ ($j = 1, 2$) have

been added to ensure that the derivatives can reach a zero value while the state-variables themselves remain strictly positive. Since the outputs are formed differentially, the nonlinear cross-coupling terms cancel out ensuring an overall linear characteristic.

Even though there exist various ways to synthesize Log-domain circuits, in the next section, we describe the transistor-level Log-domain implementation of Eq. 1 using the Bernoulli Cell (BC) formalism.

The BC was introduced in [3] as a low-level operator for synthesizing Log-domain filters. It consists of an exponential transconductor and a grounded capacitor and, through a nonlinear change of its state-variable, see Eq. 2, linearizes and electronically solves a nonlinear differential equation of the well-known Bernoulli form, see Eq. 3. A cascade of BC is known as the "Bernoulli backbone" and implements a generic set of differential equations termed the Log-Domain State-Space or LDSS. Such a "backbone" is depicted in the Log-domain biquad implementation of Fig. 6 which realizes the SS equations shown in Eq. 1. In Figs. 4, 5, and 6, the BCs are denoted by circles, each consisting of a PMOS-WI transistor and a capacitor connected between its source terminal and ground (Fig. 3).



Bioinspired CMOS Cochlea, Fig. 3 The Bernoulli cell

$$I_D(t) = I_S \exp\left(\frac{V_S(t) - V_G(t)}{nV_T}\right) = \frac{1}{T(t)} \quad (2)$$

$$\dot{T}(t) - \left[\frac{\dot{V}_G(t)}{nV_T} - \frac{u(t) - v(t)}{nCV_T}\right] T(t) - \frac{1}{nCV_T} = 0 \quad (3)$$

The synthesis of the Log-domain biquad will now be described in more detail. The transistor-level synthesis procedure is based on the direct comparison of the required dynamics of the original prototype system (i.e., Eq. 1) with those codified by the LDSS equations; the scope of this comparison is to identify the necessary time-domain relations (implemented by means of static translinear loops (TL)) that modify the LDSS dynamics in such a way so that they become identical to the desired ones [3]. Having in mind a MOS-WI implementation, the two generic LDSS equations describing a Class-A 2nd-order system (i.e., the differential equations describing two interconnected, logarithmically driven BCs [3]) are:

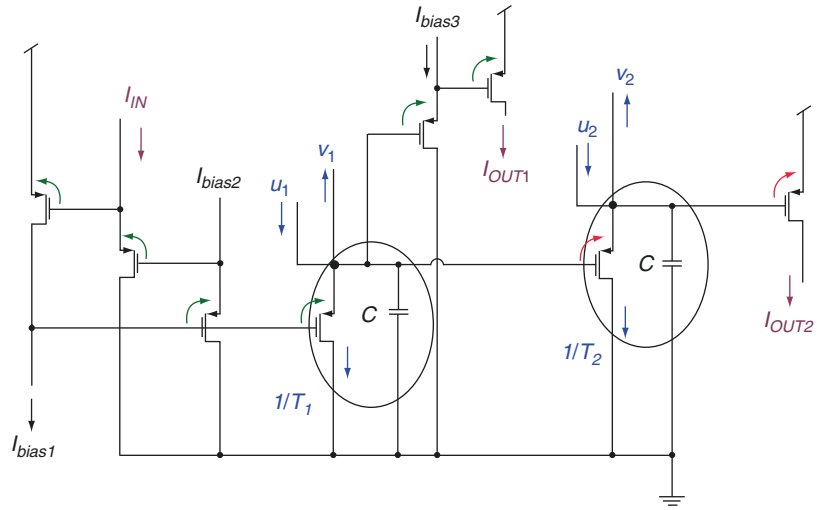
$$\begin{aligned} nCV_T\dot{w}_1 + (u_1 - v_1)w_1 &= I_{IN} \\ nCV_T\dot{w}_2 + (u_2 - v_2)w_2 &= w_1 \end{aligned} \quad (4)$$

with the w state variables being defined as:

$$\begin{aligned} w_1 &= T_1 I_{IN} (\text{dimensionless}) \\ w_2 &= T_2 T_1 I_{IN} = T_{2w1} (\text{units of amperes}^{-1}) \end{aligned} \quad (5)$$

and with C, n, and V_T denoting the grounded integrating capacitor, subthreshold slope parameter (usually between 1 and 2), and thermal voltage (25 mV at 300 K), respectively. At this point, the designer needs to decide on what kind of TL arrangement he/she wants to use for the interconnection of the BC. In this implementation, a used the "stacked" TL was used (as opposed to the "alternating" one), shown in Fig. 4 below. Note that the first BC is logarithmically driven because its gate voltage is essentially a level-shifted voltage arising from logarithmically compressing the input current I_{IN}.

With Fig. 4 as a reference, the w state variables can be "sensed" through two TL and extracted as the two output currents I_{OUT1} and I_{OUT2}. Note that even though w_1 is dimensionless and w_2 has



Bioinspired CMOS Cochlea, Fig. 4 Two interconnected logarithmically driven BCs representing the differential equations described in Eq. 4. The *green* TL is used to "sense" the w_1-state variable, whereas the *red* TL the w_2-state variable

dimensions of amperes^{-1}, both I_{OUT1} and I_{OUT2} have dimensions of amperes:

$$\begin{aligned} I_{OUT1} &= \left(\frac{I_{bias1}I_{bias2}}{I_{bias3}}\right)T_1 I_{IN} = \left(\frac{I_{bias1}I_{bias2}}{I_{bias3}}\right)w_1 \\ I_{OUT2} &= T_2(I_{bias1}I_{bias2})T_1 I_{IN} = T_2(I_{bias1}I_{bias2})w_1 \\ &= (I_{bias1}I_{bias2})w2 \end{aligned} \tag{6}$$

In the subsequent implementation $I_{\text{bias1}} = I_{\text{bias2}} = I_{\text{bias3}} = Io$, thus Eq. 6 becomes:

$$\begin{aligned} I_{OUT1} &= I_o w_1 \\ I_{OUT2} &= I_o w_2 \end{aligned} \tag{7}$$

By,

1. Splitting the input I_{IN} into two positive complementary inputs $I_{IN}{}^{u}$ and $I_{IN}{}^{l}$ (we will describe how to split I_{IN} later in section "The Input Signal Conditioner: The Geometric Mean Splitter (GMS)"
2. Duplicating the circuit topology in Fig. 4

The LDSS equations in Eq. 4 are now described in differential form as:

$$\begin{aligned} nCV_T\dot{w}_{u1} + (u_{u1} - v_{u1})w_{u1} &= I_{IN}{}^{u} \\ nCV_T\dot{w}_{u2} + (u_{u2} - v_{u2})w_{u2} &= w_{u1}n \\ CV_T\dot{w}_{l1} + (u_{l1} - v_{l1})w_{l1} &= I_{IN}{}^{l}n \\ CV_T\dot{w}_{l2} + (u_{l2} - v_{l2})w_{l2} &= w_{l1} \end{aligned} \tag{8}$$

where the subscripts u, l again indicate "upper" and "lower." Now, one needs to compare Eq. 8 with Eq. 1 and define the u, v currents via TL so that the two sets of relations become identical. As an example consider the relation of Eq. 1 and rearrange it slightly to:

$$\dot{x}_{u1} + \left(\frac{\omega_o}{Q} + \omega_o\frac{x_{l1}}{g}\right)x_{u1} = \omega_o x_{l2} + \omega_o u_u \tag{9}$$

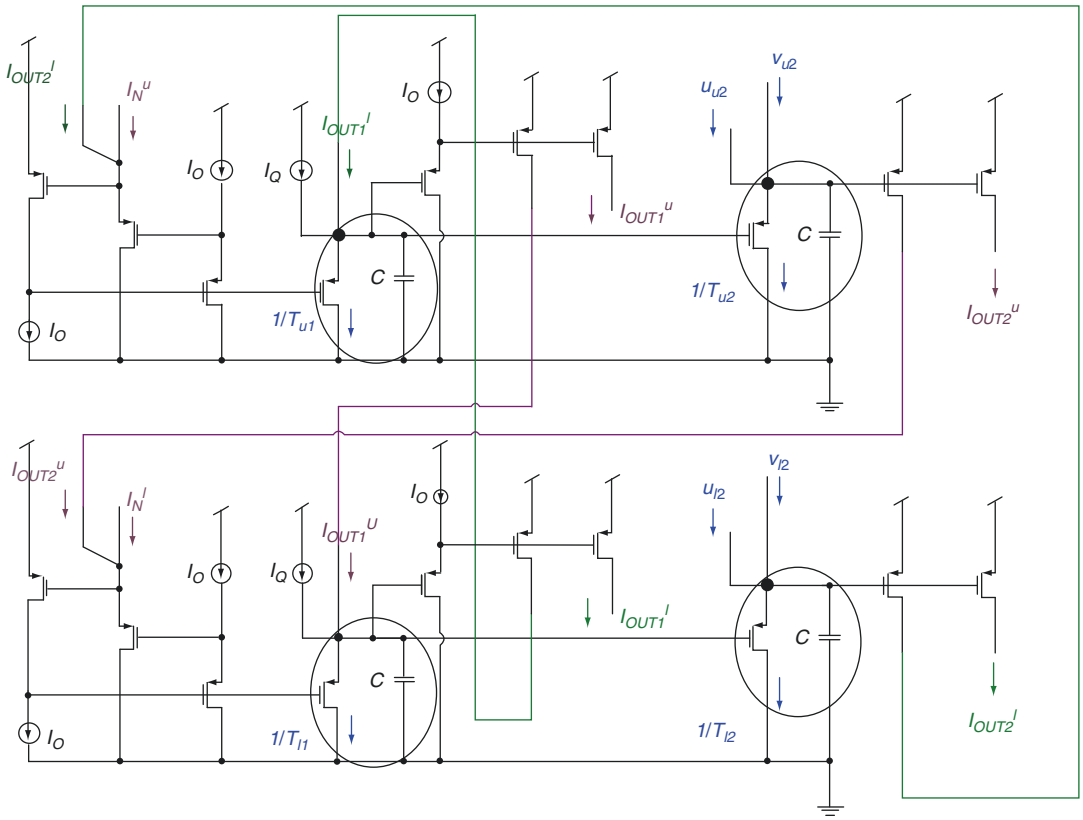
At this point, one needs to define the u, v current of the first relation in Eq. 8 (i.e., the u, v currents of the first BC) so that it "matches" that of Eq. 9. This comparison results in:

$$\begin{aligned} nCV_T\dot{w}_{u1} + \left(I_Q + I_o w_{l1}\right)w_{u1} \\ = I_p^2 w_{l2} + I_{IN}^{u} \end{aligned} \tag{10}$$

where I_Q, I_o will be defined later on. Please check the dimensional consistency of Eq. 10, given Eq. 5. Consequently, the same comparison is performed for the lower (complementary) part of the circuit.

$$\begin{aligned} nCV_T\dot{w}_{l1} + \left(I_Q + I_o w_{ul}\right)w_{l1} \\ = {I_o}^2 w_{u2} + {I_{IN}}^{l} \end{aligned} \tag{11}$$

The above comparison has defined the w state variables in Eq. 6 now to be:



Bioinspired CMOS Cochlea, Fig. 5 Circuit implementation of the *upper* and *lower* u_1 and v_1 currents

$$\begin{aligned} w_{u1} &= T_1 I_{IN} = \left[T_{u1}\left(I_{IN}{}^{u} + I_o{}^{2} w_{l2}\right)\right] \\ &= \left[T_{u1}\left(I_{IN}{}^{u} + I_{OUT2}{}^{l}\right)\right] \\ w_{l1} &= T_1 I_{IN} = \left[T_{l1}\left(I_{IN}{}^{l} + I_o{}^{2} w_{u2}\right)\right] \\ &= \left[T_{l1}\left(I_{IN}{}^{l} + I_{OUT2}{}^{u}\right)\right] \end{aligned} \tag{12}$$

The circuit implementation of Eqs. 10, 11, and 12 is shown in Fig. 5. Note the feeding of the "sensed" w state variables (i.e., I_{OUT1}, 2^{u}) from the upper to the lower portion of the circuit (purple-colored connections) and vice-versa (green-colored connections) which implement the nonlinear cross-coupling terms that guarantee true Class-AB operation.

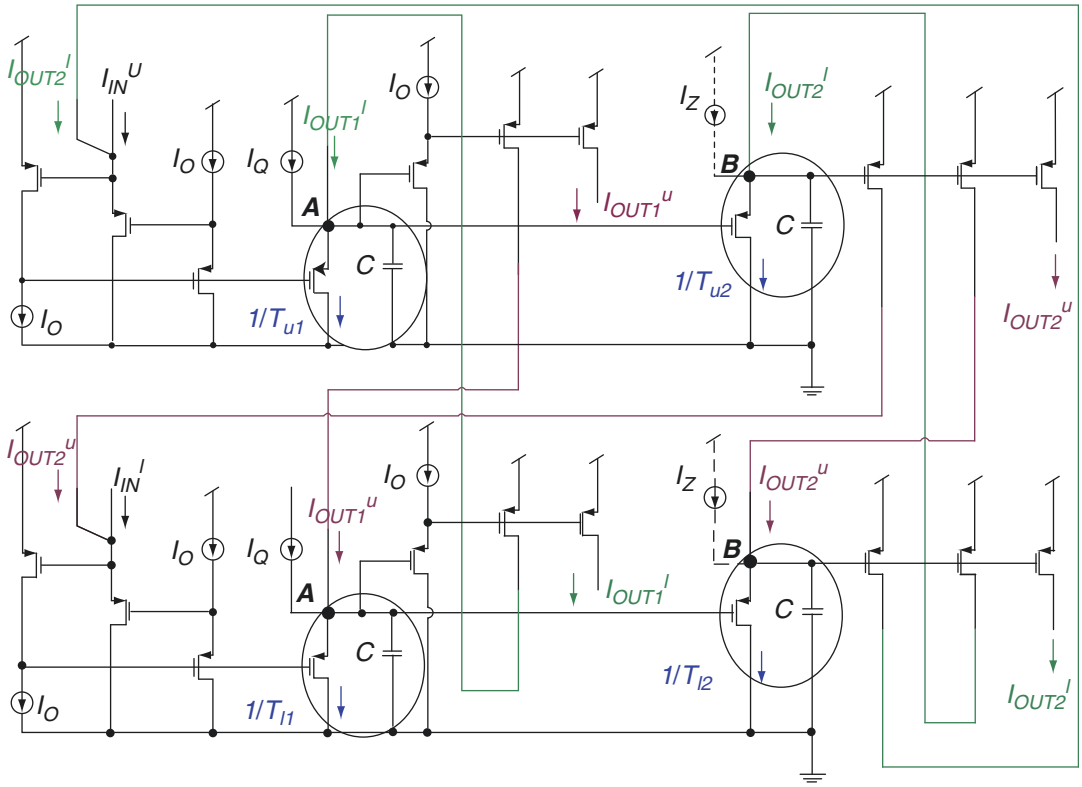
Similarly, by performing the same comparison for the other two sets of relations, the u, v currents for the second BC are defined.

$$\begin{aligned} nCV_T \dot{w}_{u2} + \left(I_o{}^{2} w_{l2}\right) w_{u2} &= w_{u1} n \\ CV_T \dot{w}_{l2} + \left(I_o{}^{2} w_{u2}\right) w_{l2} &= w_{l1} \end{aligned} \tag{13}$$

With the w state variables defined as:

$$\begin{aligned} w_{u2} &= T_2 T_1 I_{IN} = \left[T_{u2} T_{u1}\left(I_{IN}{}^{u} + I_o{}^{2} w_{l2}\right)\right] \\ &= \left[T_{u2} T_{u1}\left(I_{IN}{}^{u} + I_{OUT2}{}^{l}\right)\right] = T_{u2} w_{u1} \\ w_{l2} &= T_2 T_1 I_{IN} = \left[T_{u2} T_{u1}\left(I_{IN}{}^{u} + I_o{}^{2} w_{u2}\right)\right] \\ &= \left[T_{l2} T_{l1}\left(I_{IN}{}^{l} + I_{OUT2}{}^{u}\right)\right] = T_{l2} w_{l1} \end{aligned} \tag{14}$$

In relations (10, 11, 12, 13, and 14), $I_{IN}{}^{u,l}$ denote the two upper and lower complementary inputs (generated from the input signal conditioner described next), $1/T_{u,lj}$ ($j = 1, 2$) are the upper and lower BC drain currents, and I_Q, I_o are biasing currents that control the quality factor Q and pole frequency ω_o, respectively. Finally, the positive factor g appearing in Eq. 1 was set equal to the biasing current I_o. The final circuit implementation is shown in Fig. 5. Finally, the outputs are formed by taking the difference of the respective



Bioinspired CMOS Cochlea, Fig. 6 The pseudo-differential Class-AB Log-domain Biquad. The design implements the state-space equations shown in Eqs. 10, 11, 12, 13, and 14. The feeding of the currents $I_{OUT1,2}{}^{u}$ ($I_{OUT1,2}{}^{l}$) to the lower (*upper*) topology (implementing the linear and nonlinear cross-coupling terms to ensure correct Class-AB operation) is shown in colored connections. All biasing current sources, as well as the subtraction of the I_{OUT} currents to form the outputs, were facilitated by means of both PMOS and NMOS cascode current mirrors not shown for clarity

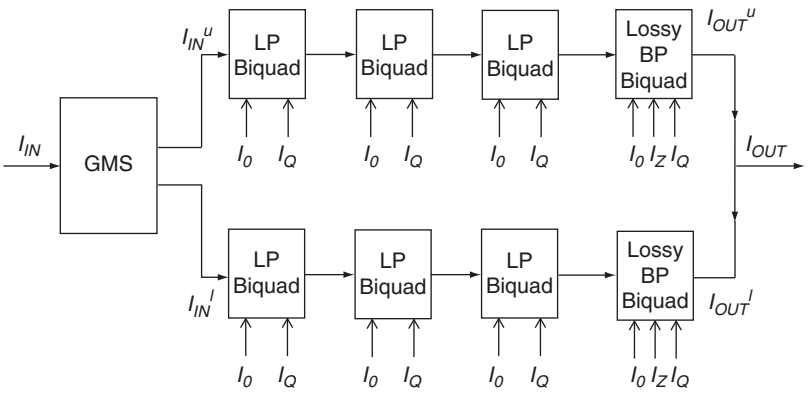
"sensed" w state variables (i.e., $I_{OUT1}{}^{u} - I_{OUT1}{}^{l}$ and $I_{OUT2}{}^{u} - I_{OUT2}{}^{l}$). The LP and BP transfer functions are given by Eqs. 16 and 17, whereas Eq. 18 describes the two-pole, one–zero transfer function required for the final OZGF channel. The transfer function in Eq. 18 is obtained by adding the biasing current I_Z at points B of Fig. 6. Note also that the addition of the biasing current I_Z alters Eq. 13 in the following way:

$$\begin{aligned} nCV_T\dot{w}_{u2} + \left(I_o{}^2 w_{l2} + I_Z\right)w_{u2} &= w_{u1}n \\ CV_T\dot{w}_{l2} + \left(I_o{}^2 w_{u2} + I_Z\right)w_{l2} &= w_{l1} \end{aligned} \tag{15}$$

By inspecting Eqs. 16, 17, and 18: $\omega_o = I_o/nCV_T$, $Q = I_o/I_Q$ and the zero position is controlled by the ratio I_o/I_Z. The final fourth-order OZGF pseudo-differential Class-AB channel architecture is depicted in Fig. 7.

$$\begin{aligned} \frac{I_{OUT_{LP}}}{I_{IN}} &= \frac{I_{OUT2}{}^{u} - I_{OUT2}{}^{l}}{I_{IN}{}^{u} - I_{IN}{}^{l}} \\ &= \frac{(I_o/nCV_T)^2}{S^2 + \dfrac{(I_o/nCV_T)}{(I_o/I_Q)}S + (I_o/nCV_T)^2} \end{aligned} \tag{16}$$

$$\begin{aligned} \frac{I_{OUT_{BP}}}{I_{IN}} &= \frac{I_{OUT1}{}^{u} - I_{OUT1}{}^{l}}{I_{IN}{}^{u} - I_{IN}{}^{l}} \\ &= \frac{(I_o/nCV_T)S}{S^2 + \dfrac{(I_o/nCV_T)}{(I_o/I_Q)}S + (I_o/nCV_T)^2} \end{aligned} \tag{17}$$



Bioinspired CMOS Cochlea, Fig. 7 The fourth-order OZGF pseudo-differential Class-AB Log-domain channel architecture

$$\frac{I_{OUT_{2PIZ}}}{I_{IN}} = \frac{{I_{OUT1}}^{u} - {I_{OUT1}}^{l}}{{I_{IN}}^{u} - {I_{IN}}^{l}} = \frac{\left(\frac{I_o}{nCV_T}\right)\left[S + \frac{I_Z}{nCV_T}\right]}{S^2 + \frac{a \times I_Q}{nCV_T}S + \frac{\beta \times {I_o}^2}{(nCV_T)^2}} \text{ with } a = \left(1 + \frac{I_Z}{I_Q}\right) \text{and } \beta = \left(1 + \frac{I_Q I_Z}{{I_o}^2}\right) \tag{18}$$

Clarification: Note that Eq. 18 has not the same denominator as Eq. 16 or 17. If one cascades Eq. 18 to a cascade of identical Eq. 16, then he/she will obtain a quasi-OZGF response which is not consistent with the transfer function discussed in [12] The choice for this discrepancy was deliberate for the following two reasons:

1. By adding a DC current *IZ* at point B in Fig. 6, a real zero in the transfer function is realized with the simplest possible way without adding extra circuitry. If one opts for Eq. 18 to have the same denominator as Eq. 16 or 17, then the various loops would have to be closed through exponential transconductors (E-cells) which would possibly degrade the performance of the resulting topology.
2. Since in this implementation I_Z is always fixed and 10 times smaller than I_o (in order to realize the −20 dB tail at DC), both a, β in Eq. 18 approach to unity when I_Q is maximum (i.e., for low gains – recall $Q = I_o/I_Q$). In that case, Eq. 18 approaches to a lossy BP transfer function with the same denominator as Eq. 16 or 17.

For smaller I_Q values (i.e., for higher gains) the variation in gain and pole frequency due to a and β increases but still does not affect the overall response noticeably due to the fact that in our fourth-order OZGF implementation, Eq. 18 is preceded by three identical LP biquads. This fact will become more apparent later on when we will discuss the measured frequency response of our OZGF implementation.

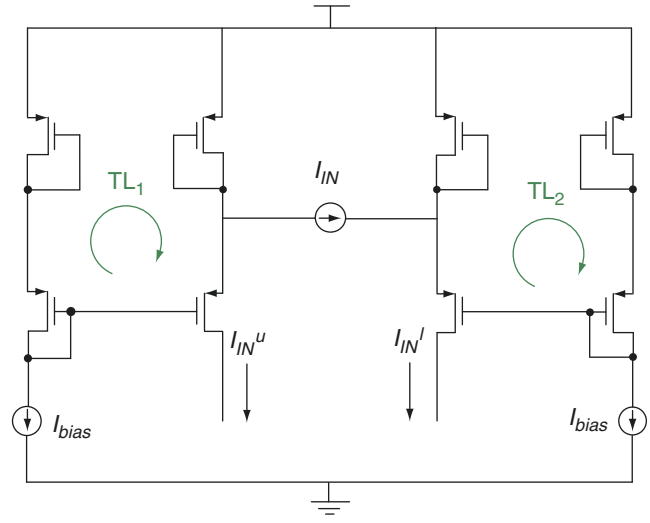
In other words, we chose to deviate from an ideal OZGF response due to the fact that it was easier to design Eq. 18 without sacrificing significantly any performance or modeling (like peak gain, CF, etc.) specifications.

The Input Signal Conditioner: The Geometric Mean Splitter (GMS)

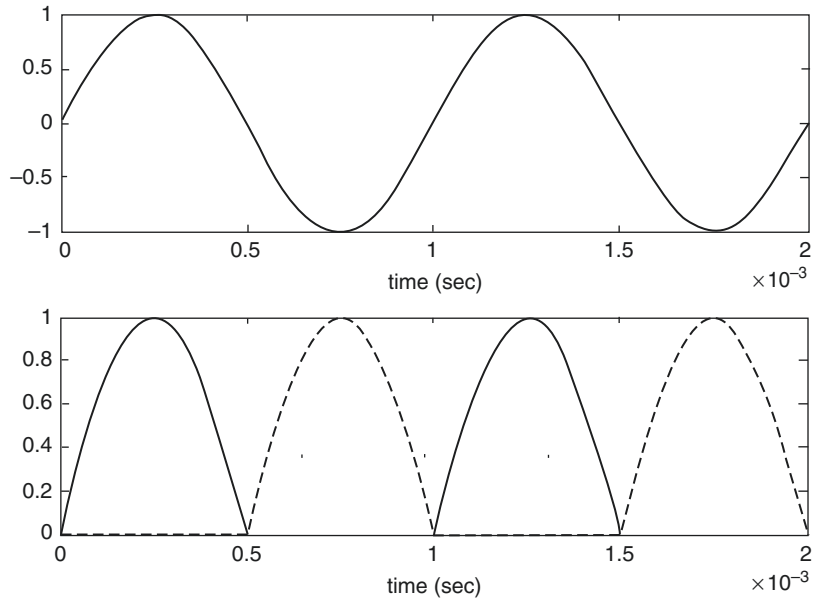
For the input signal conditioner, we chose to implement the GMS for three reasons:

1. It exhibits a good frequency response since it contains well-defined low-impedance points.
2. It can be realized with few noncritical components.
3. And for its low DC levels, ensuring lower static power consumption and noise, relative to other splitters employing, for example, the harmonic mean law.

Figure 8 depicts the PMOS-WI GMS implementation used in this work. From the two (left and right) TL, one may deduce that:



Bioinspired CMOS Cochlea, Fig. 8 PMOS-WI implementation of the geometric mean splitter



Bioinspired CMOS Cochlea, Fig. 9 Indicative current waveforms generated by the GMS. The upper (*solid*) and lower (*dotted*) complementary inputs resemble half-wave rectified signals

$$I_{IN}{}^{u} \times I_{IN}{}^{l} = I_{bias}{}^{2} \left(= I_{o}{}^{2}\right) \tag{19}$$

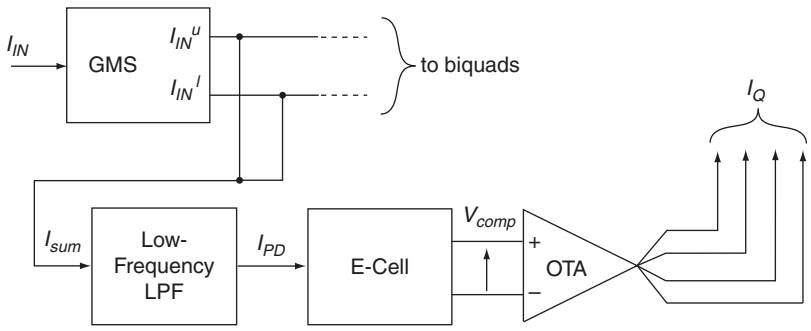
$$\text{It also holds that} \ : \ \mathrm{I_{IN}} = \mathrm{I_{IN}}^{\mathrm{u}} - \mathrm{I_{IN}}^{1} \tag{20}$$

Moreover, the large-signal expressions for the two positive, complementary, upper and lower, unidirectional currents are given by:

$$I_{IN}{}^{u} = \frac{I_{IN} + \sqrt{I_{IN}{}^{2} + 4I_{o}{}^{2}}}{2} \tag{21}$$

$$I_{IN}{}^{l} = \frac{I_{IN} + \sqrt{I_{IN}{}^{2} + 4I_{o}{}^{2}}}{2} \tag{22}$$

Relations (19) and (20) together describe the splitting action and the law with which the two positive complementary inputs are defined. Figure 9 illustrates typical current waveforms generated by the GMS used in this work. The biasing current I_{bias} could in principle be set arbitrarily small to reduce the noise and static power



Bioinspired CMOS Cochlea, Fig. 10 Block diagram describing the processing stages of the AGC

but, in the subsequent OZGF, implementation was set equal to I_o in order to ensure similar biasing conditions (and hence impedance levels) for all the transistors comprising the OZGF channel and reduce the overall number of external pin connections.

Finally, it should be noted that one of the potential disadvantages of the GMS topology of Fig. 8 is the implementation of the floating current source I_{IN}. In practice, this current source was realized via two grounded AC-current sources at an 180° phase difference with each other. The relative matching between the two sources will play a significant role in the linearity performance of the OZGF channel since the performance of the whole system cannot exceed that of its input stage. This point will be addressed in more detail in a later section.

The Automatic Gain-Control Circuits

The AGC should be designed in such a way so that a potentially wide DR at the input of the OZGF channel (greater than six orders of magnitude in this implementation) can be compressed to a small current range (typically within one order of magnitude) for the biasing current I_Q that sets the Q of the biquads. The proposed AGC acts in feed-forward and consists of a cascade of four circuit blocks: a GMS, a low-frequency LP filter (LPF), an exponential transconductor (or E-cell) operating as a logarithmic transimpedance amplifier, and a wide linear-range OTA.

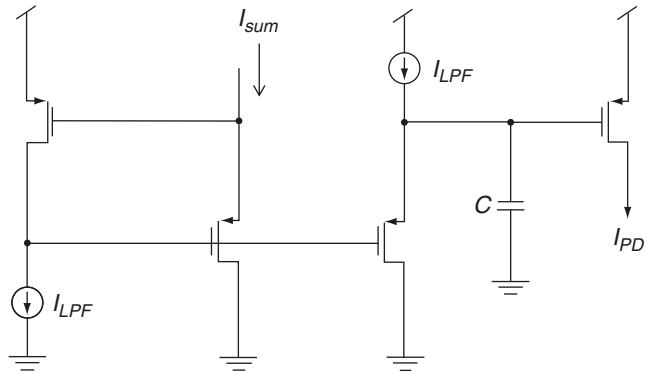
With the aid of Fig. 10, the processing stages can be summarized as follows:

- Stage 1: The GMS ensures that for small signals, the OZGF gain does not exceed a certain value or shoot to instability (i.e., it sets the highest OZGF peak gain for zero input) and also full-wave rectifies the bidirectional input I_{IN} (I_{sum}).
- Stage 2: The low-frequency current-input current-output LPF gives a quasi-DC output (*IPD*), whose value corresponds to the peak value of the full-wave rectified current-input I_{sum}. In other words, the GMS together with the LPF implement a peak detector (i.e., full-wave rectification plus averaging).
- Stage 3: The E-cell compresses the, still-wide, current range at the output of the LPF to a differential voltage range (V_{comp}) that can be easily accommodated by the OTA that follows. The compressive *I-V* transfer characteristic is electronically controlled by a single biasing current $I_{o_control}$.
- Stage 4: The OTA converts V_{comp} to a DC current range, corresponding to the Q range of the OZGF. The tail current of the OTA (I_{tail}) together with $I_{o_control}$ of the E-cell are the two main parameters that control the law with which the Q values are determined (i.e., which input strength value corresponds to which Q).

In the subsections that follow, the AGC circuits are described in detail.

AGC Stages 1 and 2: The GMS and Low-Frequency LPF

By adding the large-signal expressions of the two splitted complementary inputs (i.e., Eqs. 21 and 22) and by defining the peak values of the input as $\hat{I}_{IN} = m \times I_o$, with m being the modulation index, we arrive at the following result:



Bioinspired CMOS Cochlea, Fig. 11 The low-frequency LPF topology – A first-order Class-A Log-domain integrator

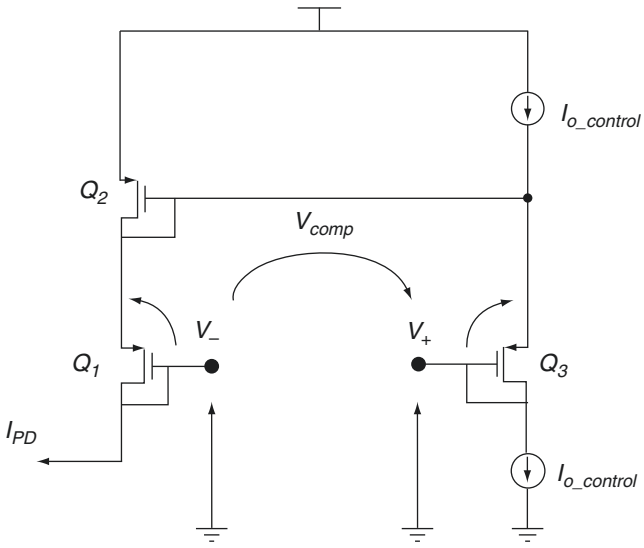
$$\begin{aligned}\hat{I}_{sum} &= \frac{\hat{I}_{IN} + \sqrt{\hat{I}_{IN}{}^2 + 4I_o{}^2}}{2} \\ &+ \frac{-\hat{I}_{IN} + \sqrt{\hat{I}_{IN}{}^2 + 4I_o{}^2}}{2} \\ &= \sqrt{\hat{I}_{IN}{}^2 + 4I_o{}^2} = I_o\sqrt{m^2+4}\end{aligned} \tag{23}$$

with m indicating how many times larger or smaller is the input zero-to-peak amplitude relative to the biasing current I_o. The above relation states that for small m and less than 1 (i.e., when the biquads operate in their Class-A mode), $\hat{I}_{sum} \approx 2I_o$, whereas for $m \gg 1$, $\hat{I}_{sum} \approx mI_o = \hat{I}_{IN}$. In other words, for zero or small signals, the GMS outputs a current which has a minimum peak value close to $2I_o$, whereas for larger signals, it follows the input peak almost exactly. Thus, by setting the control law to map the value $2I_o$ to the minimum I_Q value (corresponding to the maximum Q, since $Q = I_o/I_Q$), the OZGF channel will always have a bounded peak gain according to the required specifications and will never shoot up to large unwanted gain values or instability. That is indeed a "hidden" and very useful operation of the GMS because it can rectify the input and set the highest bound for the OZGF peak gain in a simple and elegant manner without needing to resort to additional circuitry.

The current I_{sum} is consequently filtered by the low-frequency LPF to extract its DC component I_{PD}. However, to accurately obtain the peak value, I_{sum} needs to be multiplied by an appropriate scale factor (the crest factor). This multiplication factor, which is waveform dependent and was set here to $\pi/2$, was implemented by means of scaled current mirrors in the GMS. Since I_{sum} is a rectified positive current, the LPF does not have to operate in Class-AB mode. The LPF topology used in this work and depicted in Fig. 11 is a simple Class-A, 1st-order, Log-domain integrator which implements the transfer function in Eq. 24. All symbols have their usual meaning.

$$H_{LPF}(s) = \frac{\omega_o'}{s+\omega_o'} = \frac{(I_{LPF}/nC_{LPF}V_T)}{s+(I_{LPF}/nC_{LPF}V_T)} \tag{24}$$

Clarification: It is known from physiology that the IHC perform half-wave rectification to extract the energy of the BM vibrations. So why did we opt for a full-wave rectification in this implementation since clearly this choice is inconsistent with biology? The reason is that IHC and OHC are two separate sensors in the real cochlea. The IHC do not control the gain of the cochlea as OHC do, but through the auditory nerves create action potentials to convey information to the brain. For this reason, both the phase and frequency information of the signal they detect is important. *Half-wave rectification enhances the zero-crossings of the signal while preserving its frequency information, whereas full-wave does not (it doubles the frequency)*. In this implementation, the silicon neurons that convert the output of the OZGF channel into spikes were not designed; only a circuit that models (at least qualitatively) the function of OHC for controlling the channel's gain. The proposed AGC operates virtually at DC and so both frequency and phase information is unimportant at this point. The choice for a full-wave rectification was made since the capacitor value used in



Bioinspired CMOS Cochlea, Fig. 12 The exponential transconductor – E-cell used for AGC action

the low-frequency LPF for smoothing the full-wave rectified signal can me made smaller than in the half-wave case, thereby saving area. That is one example of a choice that is not neuromorphic in its strict sense but, given the particular chip's area requirements, serves as a viable alternative.

AGC Stages 3 and 4: The E-cell and Wide Linear-Range OTA

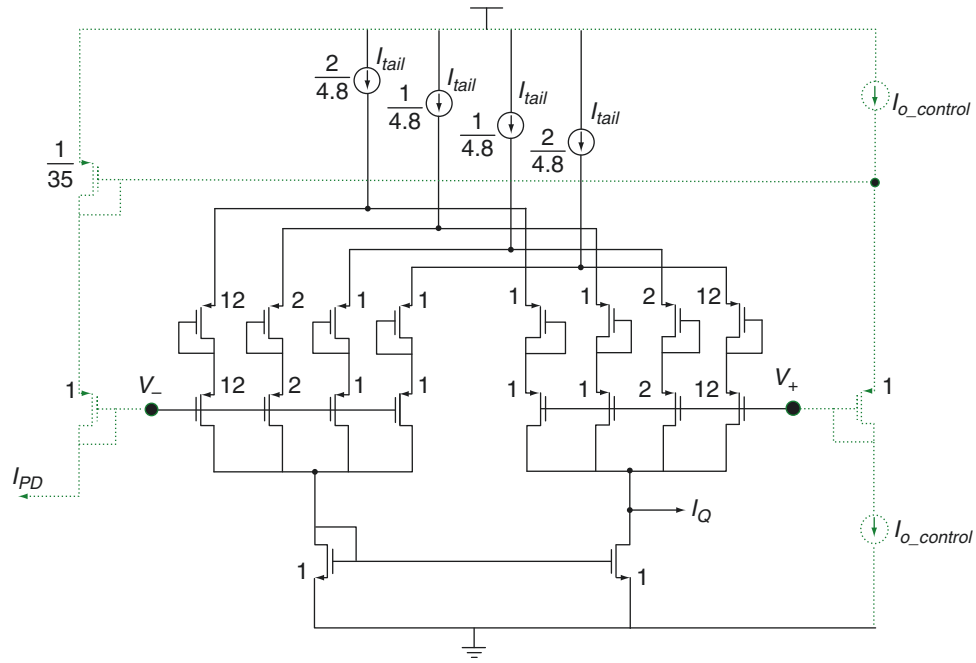
The E-cell, shown in Fig. 12, is an exponential transconductor operating as a nonlinear transimpedance amplifier. It consists of two low-impedance points (the drain-gates of devices Q_1 and Q_3), where the compressed differential voltage V_{comp} is generated when the current I_{PD} is sourced from the diode-connected drain of device Q_1. Device Q_2 serves as a degeneration (tail) resistor which, if appropriately sized, gives additional headroom for accommodating the upper biasing current source $I_{o_control}$ (implemented by means of cascoded mirrors not shown for simplicity). The reason for using an E-cell compared to a standard diode to perform the logarithmic compression is attributed to the fact that with an E-cell, the compressive *I-V* transfer characteristic can be electronically tuned by $I_{o_control}$ as described by Eq. 25. From the TL Q_1Q_3 and by assuming that $I_{Do1} = I_{Do3}$ (i.e., matched devices):

$$V+ - nV_T \ln\left(\frac{I_{PD}}{I_{Do1}}\right) + nV_T \ln\left(\frac{I_{o-control}}{I_{Do3}}\right) - V- = 0$$
$$\Rightarrow V_+ - V_- = V_{comp} = nV_T \ln\left(\frac{I_{PD}}{I_{o-control}}\right) \quad (25)$$

The compressed differential voltage V_{comp} is subsequently applied to an OTA whose linear range was widened by using two standard techniques:

1. Plurality of cross-coupled, source-coupled pairs [14, 19].
2. Source degeneration. More specifically, four OTAs, each with an appropriate offset voltage and tail current, were degenerated and connected in tandem in order to realize an overall flatter input–output slope. This was achieved by choosing the offset voltages and tail currents in such a way so that the higher-order even and odd derivatives of the total (combined) transconductance become zero.

The offset voltages were realized by scaling asymmetrically the dimensions of the devices that comprise the branches of each OTA pair (see Fig. 13). In [19], for maximally flat transconductance characteristics, these aspect ratios were calculated to be 13.4, 2.04, 1, 1 and with tail current multiplication factors of 1.83, 1, 1, 1.83.

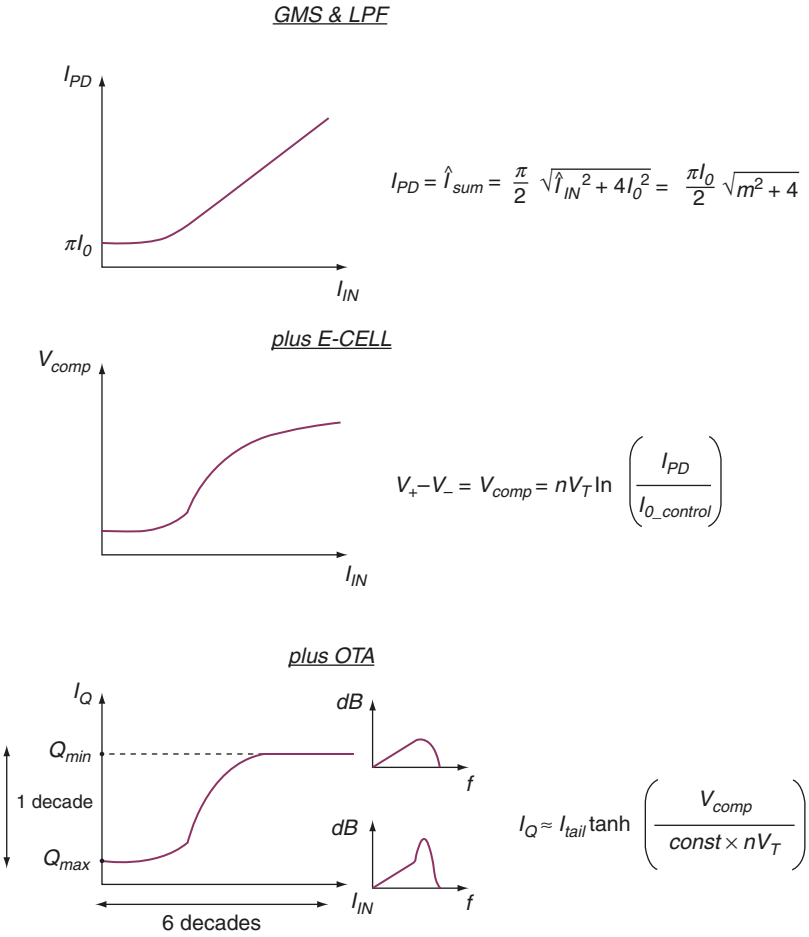


Bioinspired CMOS Cochlea, Fig. 13 The E-cell (*dotted green lines*) connected together with a wide linear-range OTA employing asymmetric cross-coupled, source-coupled pairs. The devices aspect ratios are indicated on the schematic

However, since accurate linearity is not a crucial performance requirement for the AGC, the device aspect ratios and tail current multiplication factors were set to 12, 2, 1, 1 and 2, 1, 1, 2, respectively, in order to simplify our layout matching efforts during fabrication phase. The scaling factor of 4.8, which all the tail currents are divided with, was chosen empirically so that the upper and lower saturating levels of the *V-I* tanh function coincide with the actual I_{tail} value. Figure 14 describes qualitatively the input to every AGC stage output transfer characteristics. The *x*-axes were exaggerated to indicate the various different ranges of operation. In essence, the AGC is able to compress a wide input range of 120 + dB to a much smaller range (∼20 dB) for the Q values setting effectively the peak gain of the OZGF response according to input level.

AGC Simulation and Measured Results

The simulation results presented in this section were obtained from Cadence IC Design Framework® and with the parameters presented in Table 1. Figure 15 shows the simulated *V-I* and transconductance transfer characteristics of the wide linear-range OTA for an I_{tail} of 24 nA. Strictly speaking, the achieved linear input range is around 200 mV_{peak}. However, since (a) the OTA operates in quasi-DC and (b) the overall AGC characteristic must be compressive, the practically usable range is around 400 mV_{peak}. It should be noted that since $I_Q{}^{max} \equiv I_{tail}$, the OTA sets the lowest bound of the OZGF peak gain according to $Q_{min} = I_o/I_{tail}$. The control law was calibrated (via the currents $I_{o_control}$ and I_{tail}) to map the minimum signal at the input of the OZGF (which results in a value of $2I_o \times (\pi/2)$ at the output of the LPF of the AGC) to a maximum biquad Q of 10. This corresponds to a value of 2 nA for the respective I_Q currents and a nominal OZGF peak gain of ∼80 dB. On the other hand, for a maximum allowable input of 10 μA (i.e., for $m = 500$), the AGC gives an I_Q equal to 20 nA corresponding to a Q of 1 (see Fig. 16). Since I_{tail} sets the upper level of the *V-I* transfer characteristic of the OTA, for very large input signals



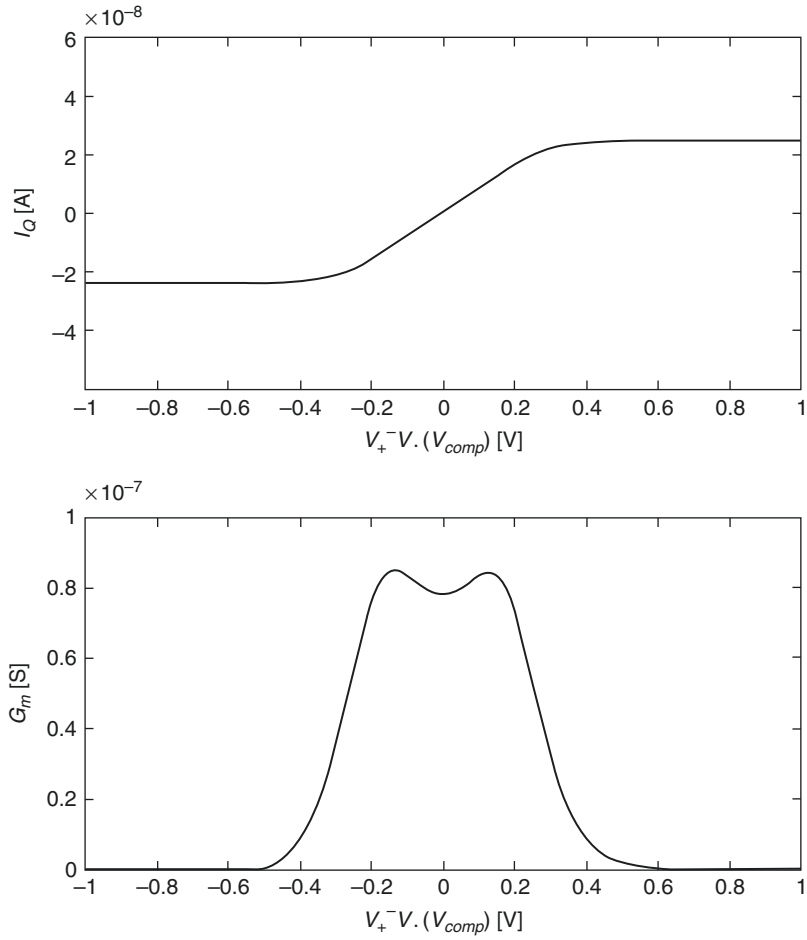
Bioinspired CMOS Cochlea, Fig. 14 Input to every AGC stage output transfer characteristics. Note that the actual I_Q expression is an approximate one since a closed form expression for the transconductance of the wide linear-range OTA cannot be obtained

($m > 500$) the maximum I_Q value saturates at 24 nA, corresponding to a lowest Q of 0.834 and a passive OZGF peak gain of 1.43 dB. Indicative waveforms at the outputs of each AGC stage are shown in Fig. 17.

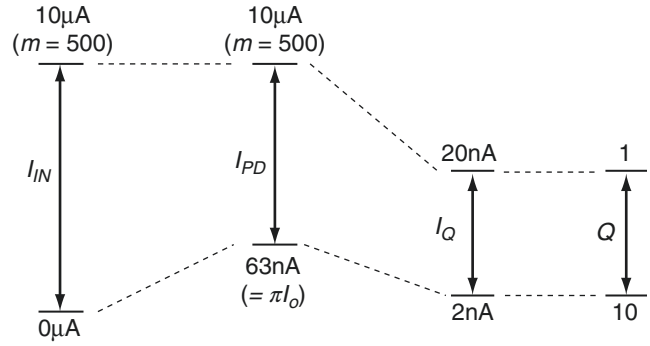
The law with which the Q values are determined was obtained by performing a DC-sweep analysis on the whole AGC system. Figure 17 depicts I_Q versus I_{IN} curves with $I_{o_control}$ and I_{tail} as the two implicit parameters. It can be observed that the overall input–output AGC transfer characteristic is logarithmically compressive in nature as suggested by Eq. 25 with $I_{o_control}$ controlling the "vertical shift" of the characteristic and with I_{tail} controlling the actual level of compression as expected from Eq. 25 and OTA operation. Also observe that a linear change in I_{tail} results in a logarithmic change on the compression level, whereas a linear change in $I_{o_control}$ results in a linear vertical shift of the whole characteristic. In conclusion, the above two electronically controllable degrees of freedom give a certain level of versatility to the AGC system because a variety of input DR values can be accommodated and mapped to a specific Q range according to the particular physiological (modelling) or design (performance) specifications (Figs. 18, 19, and 20).

Active OZGF Channel–Measured Results

The whole channel, together with its AGC, was integrated in the standard 0.35 μm AMS 2P/4 M CMOS process. The OZGF was tuned so that its



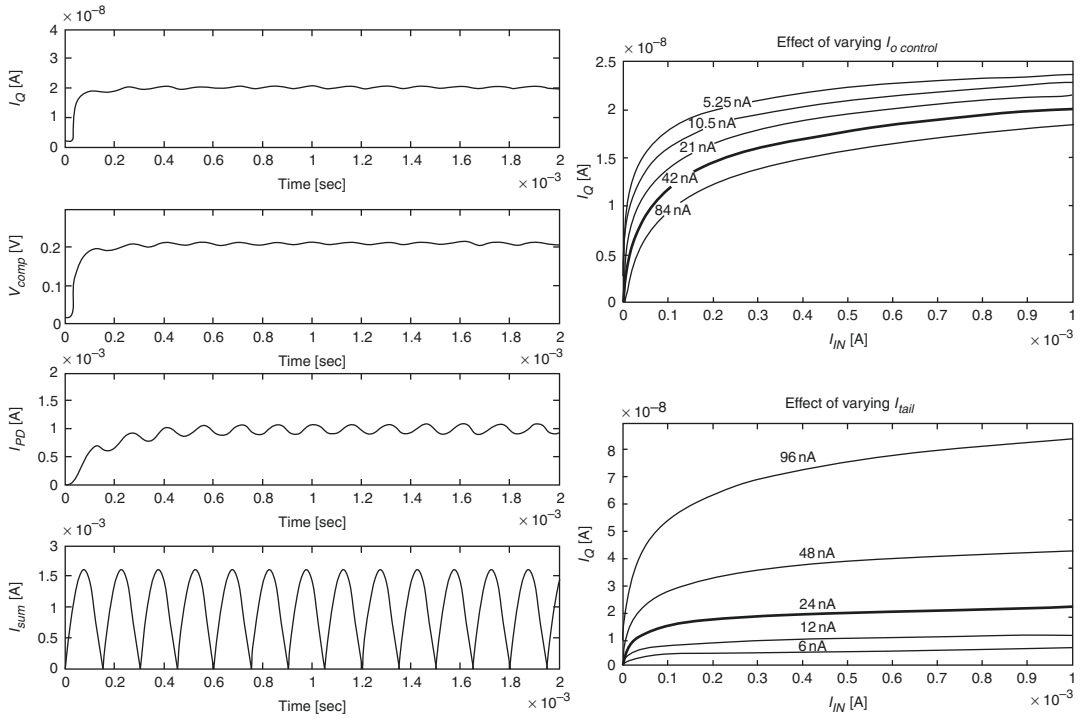
Bioinspired CMOS Cochlea, Fig. 15 *V-I-* (*upper*) and transconductance- (*lower*) simulated transfer characteristics of the wide linear-range OTA for $I_{tail} = 24$ nA



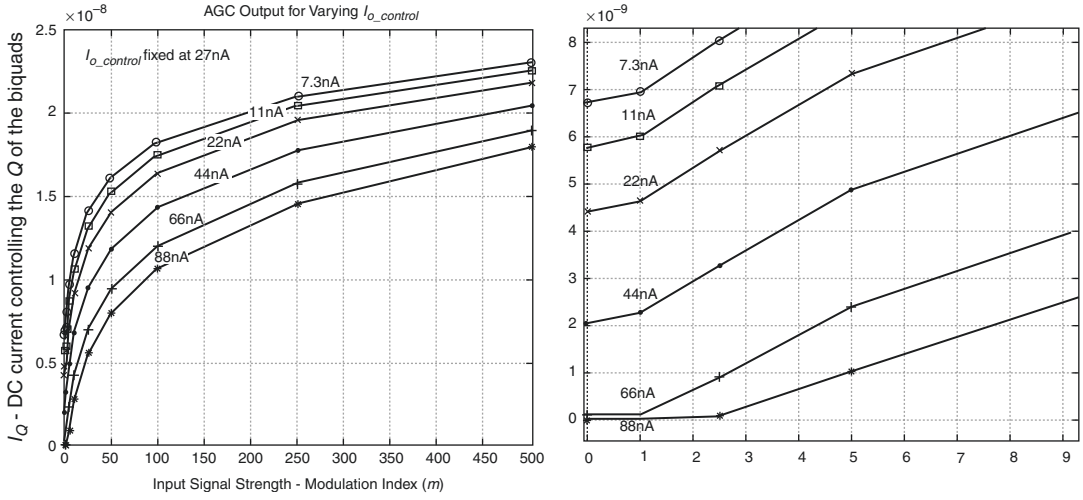
Bioinspired CMOS Cochlea, Fig. 16 Mapping of the input current range to OZGF Q values

nominal center frequency (CF) falls at 3.3 kHz for a Q of 1. This corresponds to $I_o = I_Q = 20$ nA.

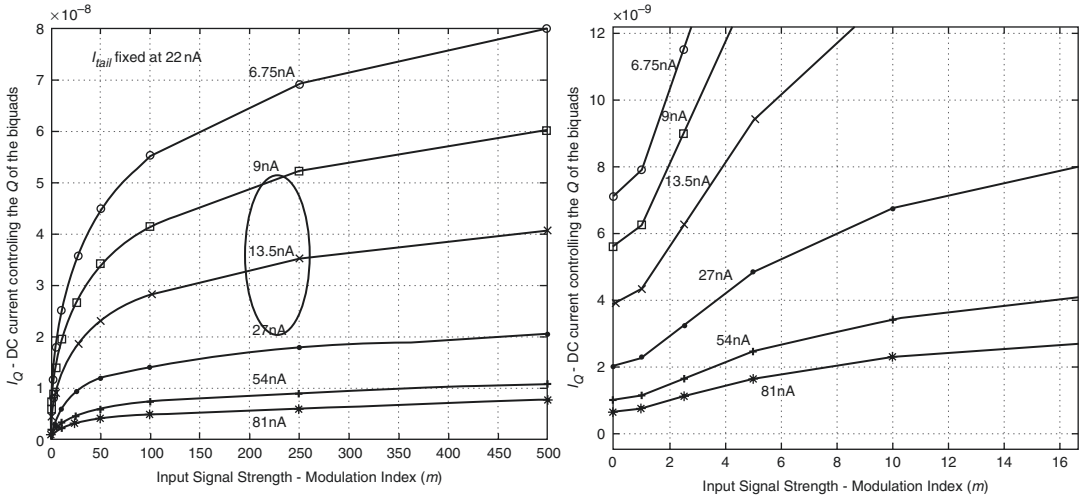
Measurement Setup: In section "The Input Signal Conditioner: The Geometric Mean Splitter (GMS)" it was mentioned that one of the disadvantages of the particular GMS topology was the need to accurately realize the floating current source I_{IN} in order to minimize distortion. In this case, this was achieved by using two 6221 precision AC/DC Keithley current sources. These were programmed using MatLab™ (via their GPIB interface) and triggered externally to output two



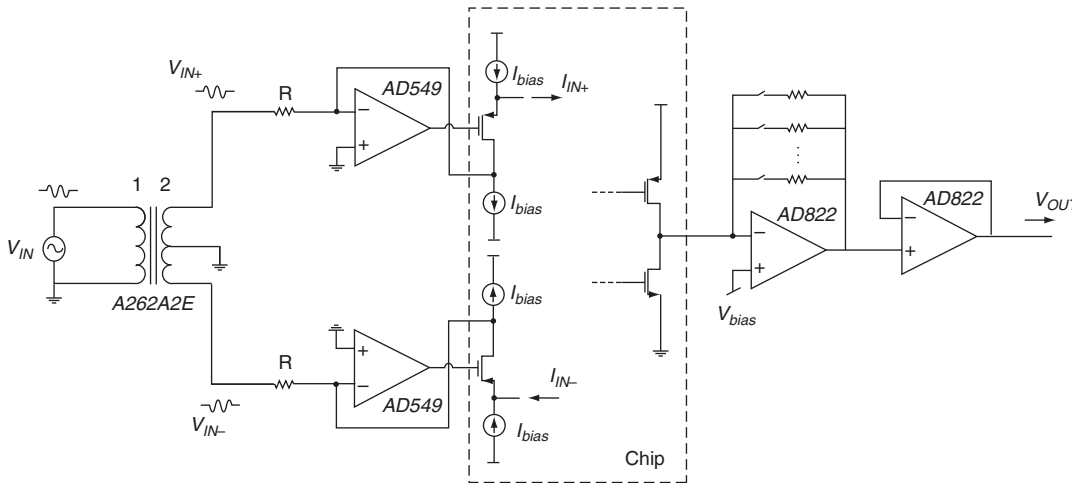
Bioinspired CMOS Cochlea, Fig. 17 (*Left*) Current waveforms from the output of each AGC stage for an input signal of $m = 500$. Observe that the peak of I_{sum} is at ∼16 μA instead of 10 μA (due to scaling by $\pi/2$), whereas the LP-filtered version of I_{sum}, I_{PD} is at 10 μA as expected. The I_Q value is at 20 nA corresponding to a Q of 1. (*Right*) Simulated parametric DC response plots of the Q-control law with varying $I_{o_control}$ (*upper*) and I_{tail} (*lower*). The *bold curves* are the ones corresponding to the values shown in



Bioinspired CMOS Cochlea, Fig. 18 (*Left*) Measured parametric DC response plots of the Q-control law with varying $I_{o_control}$. Note the characteristics for small-signals level of the *Right* plot (as expected from the operation of the GMS) with $I_{o_control}$ having a rather large effect on the y-intercept value but a smaller one on the large-signal saturating levels



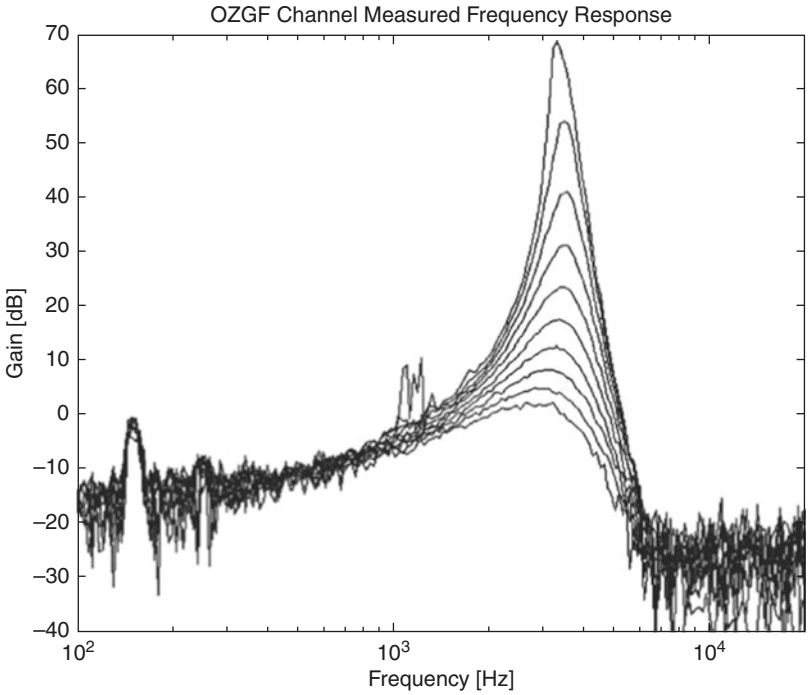
Bioinspired CMOS Cochlea, Fig. 19 (*Left*) Measured parametric DC response plots of the *Q*-control law with varying I_{tail}. Note that the characteristics level of for small signals (as expected from the operation of the GMS) with I_{tail} having a small effect on the *y*-intercept value (*Right*) but a much larger one on the large-signal saturating levels



Bioinspired CMOS Cochlea, Fig. 20 Experimental setup for measuring the OZGF channel

AC-current waveforms at an exact 180° phase difference with each other. Moreover, by fine-tuning their relative amplitudes and phases, we could partly overcome any mismatches between the two Class-A paths and, consequently, further optimize the linearity performance of the OZGF channel. The OZGF output current was measured via standard *I*-to-*V* converters (utilizing 100fA-leakage current AD549 Opamps) and through the 1 MΩ input resistance of a Stanford Research 1 mHz–100 kHz spectrum analyzer (the SRT785).

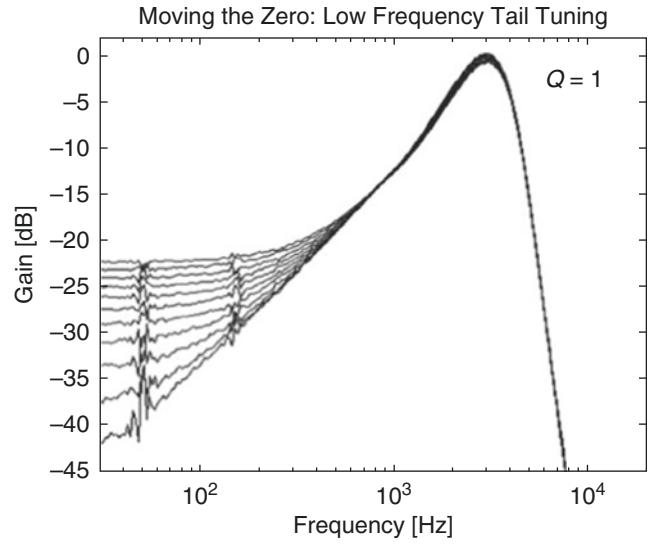
Table 1 shows the various dimensions of the devices comprising the channel and AGC. It should be emphasized that this choice occurred after careful deliberation and optimization. For the case of the Log-domain biquad (where high DR performance is of primary interest), a large *W* (such as 300 μm) shifts the $\ln I_{DS} - V_{GS}$



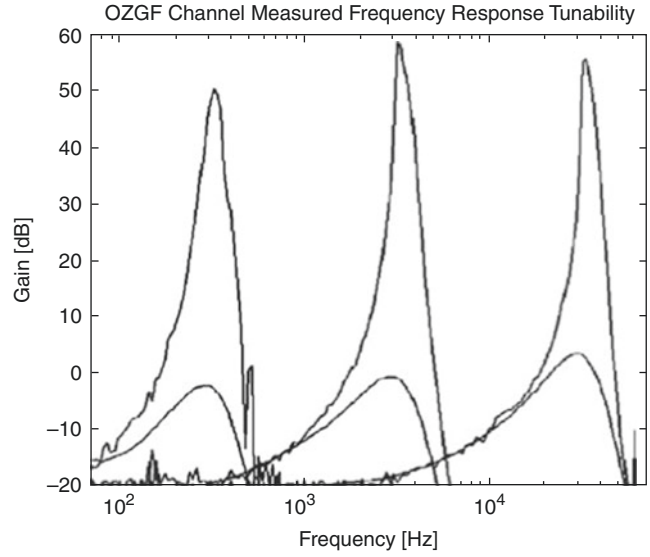
Bioinspired CMOS Cochlea, Fig. 21 Q-tunability of the active fourth-order OZGF frequency response. The maximum peak gain was measured at 70 dB

characteristic vertically so that the upper limit of the WI region extends to the μA-range, *while the leakage currents are still maintained below noise levels*. This is an important design choice to ensure that the full WI range, offered by the particular CMOS process, is exploited. In addition, the choice for L is equally important: For best THD performance, L should be made minimum since it affects the linearity of the WI slope, (i.e., the range of the flat portion of the $d(\ln I_{DS})/dV_{GS}$ versus V_{GS} curve) but a small L will give rise to channel-length modulation effects [1]. A larger L will definitely make the respective device's output impedance larger, but overall output THD will be reduced. For this reason, it was found that $L = 1.5$ μm (i.e., around four times the feature size) was a good compromise between driving capability and linearity. In addition, the device areas satisfied a 3σ-VT mismatch of about ± 2 mV which is somewhere in between moderate to precise matching.

Frequency Response: Figure 21 shows the Q-tunability of the active fourth-order OZGF frequency response. The current I_Q was changed automatically via the action of the AGC by changing accordingly the strength of the input signal from the spectrum analyzer. Observe also that a linear change in I_Q corresponds to a logarithmic change in the peak gain (or Q) due to (a) the compressive law of the AGC and (b) due to the fact that $Q = I_o/I_Q$; this is indeed another form of compressive behavior which is embedded in the frequency response and stems directly from the design equations of the Log-domain biquads. Figure 22 shows tuning of the OZGF's low-frequency tail by varying the current I_Z (note the quite small gain and pole frequency variation due to a and β terms in Eq. 18 – see again discussion at the end of section "Pseudo-Differential Class-AB Log-Domain Biquads"), whereas Fig. 23 shows the two orders of magnitude ω_o-tunability when changing the biasing current I_o to 2, 20 and 200 nA, successively. For each ω_o, two indicative Q setups are shown to demonstrate gain adaptation within the whole audio spectrum. Observe the quite stable characteristics over frequency. Also note that the low-ω_o, high-Q response is affected by input signal noise because of the low biasing current levels of the particular configuration. Finally, it should be emphasized



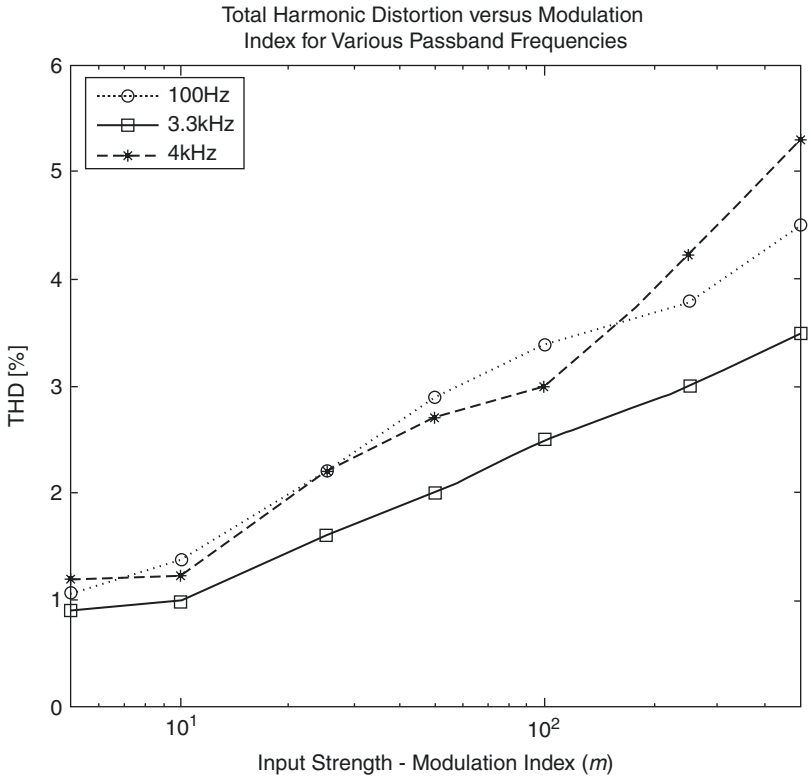
Bioinspired CMOS Cochlea, Fig. 22 Low-frequency tail tuning by means of varying the biasing current I_Z



Bioinspired CMOS Cochlea, Fig. 23 Gain adaptation and ω_o-tunability (CF from 330 to 33 kHz)

that all low-to-moderate Q responses were obtained with a large m ($>>1$) in order to validate the OZGF's Class-AB operation.

Total Harmonic Distortion (THD): The THD is a common measure for assessing the linearity performance of amplifier systems. However, when talking about audio-auditory processors, its value must be appropriately judged since the "shape" of a signal's spectrum is something that affects one's hearing perception. In other words, a signal with a certain THD value might sound "aesthetically more pleasing" to our ears than another of equal or lower THD value. In commercial and academic cochlea designs, a common upper limit for THD is 4–5 % (e.g., Sarpeshkar's neuromorphic cochlea: <5 % [17, 18], MED-EL's state-of-the-art DUET EAS hearing system: <5 % @500 Hz, Phonak Super-Front PPC-4 <7 % @500 Hz, etc.). This was also justified experimentally from a recent (and one of the first) Log-domain implementations of a CMOS cochlea, where the effect of distortion started to

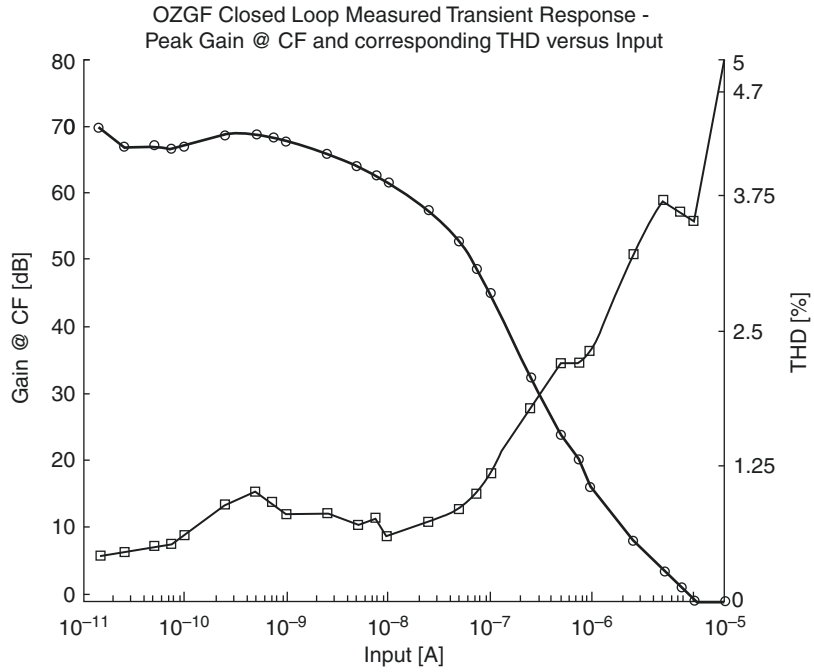


Bioinspired CMOS Cochlea, Fig. 24 Measured single-tone linearity of the fourth-order ($Q = 1$) OZGF for various passband frequencies

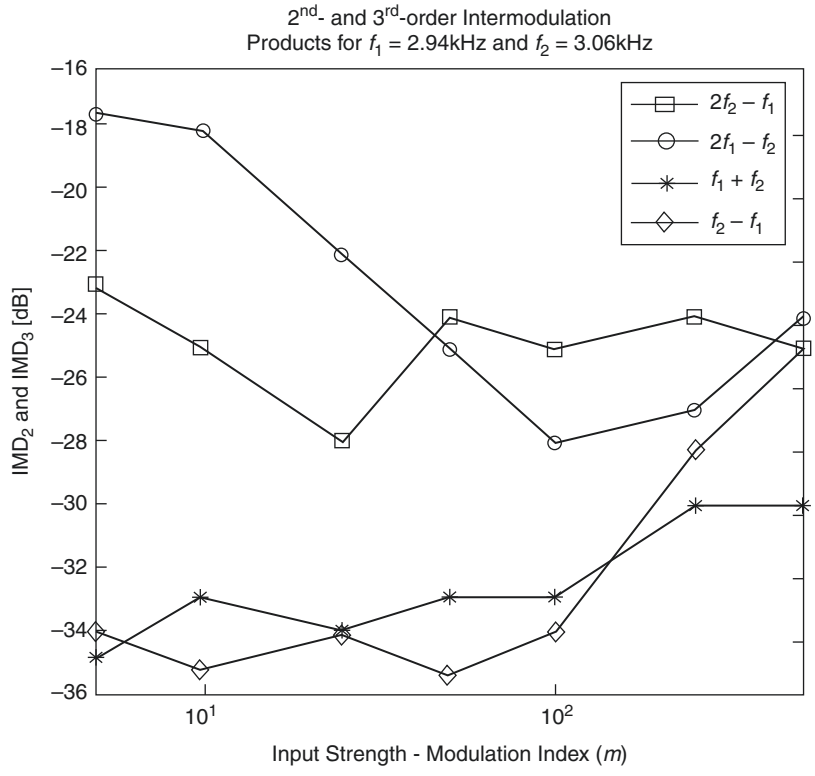
become evident to the ear after 5 % THD [10]. Finally, note that the 10 %@1 kHz THD value represents a desirable amount of allowable distortion for typical low-cost wide-band audio material such as FM broadcasts, CD-media or cassette tape media. For these reasons, it was decided that 5 % is a reasonable upper limit of THD for our OZGF cochlea channel.

Due to their differential nature, CMOS-WI Class-AB Log-domain filters exhibit mainly odd-harmonic distortion content. Their THD (for frequencies near the pole frequency) usually reaches the 1 % value quite early (for m values as low as 1, i.e., at the border between Class-A and Class-AB operation) but grows rather slowly thereafter; this behavior renders Log-domain filters as adequate candidates for applications where moderate THD values can be tolerated. Figure 24 shows THD versus input strength (m) for various tones near-and-at the CF and deep in the passband for the open-loop $Q = 1$ OZGF response. At the CF, the THD for an input of 10 μA ($m = 500$) is at 3.5 %. Figure 25 shows how the active OZGF adapts automatically its peak gain with input level together with the corresponding output THD. For a 14 pA input tone, the OZGF provides a peak gain of 70 dB at 0.4 % output THD, whereas for an input tone of 20 μA, the peak gain is at −0.85 dB with an output THD of 5 %.

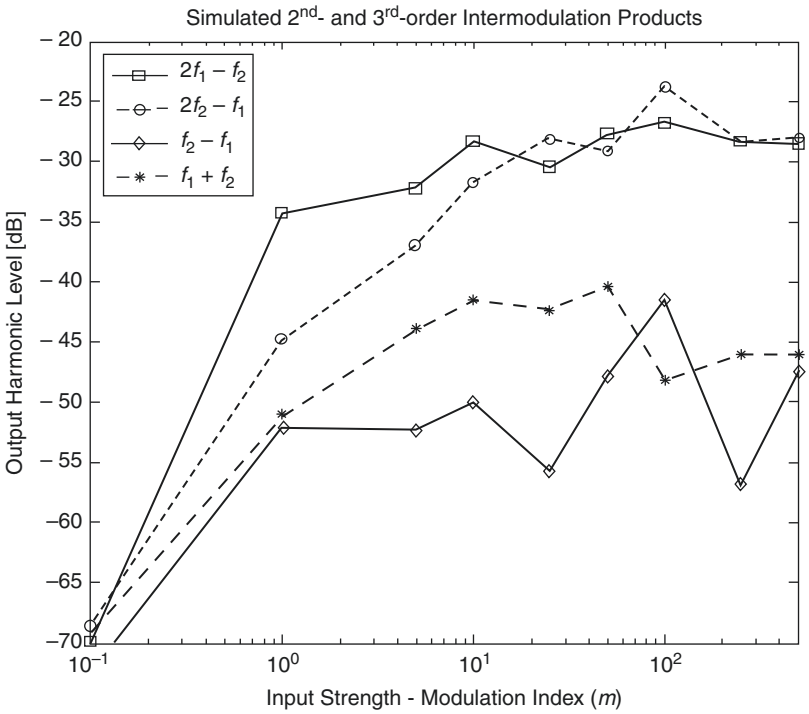
Two-Tone Intermodulation Tests: Figure 26 shows both even- and odd-order measured intermodulation distortion products (IMD) for 2 t placed at a frequency ±2 % away from the (CF) 3 kHz of the open-loop $Q = 1$ OZGF response. The IMD_2 starts at around −34 dB for small-signals and increases gradually to higher dB values with $f_2 - f_1$ increasing faster than $f_1 - f_2$. The IMD_3 is less attractive with a much unpredictable behavior toward smaller signals. These results unfortunately do not compare favorably to IMD figures reported from similar (biquadratic and not) prior CMOS-WI Log-domain efforts [11, 16] – even though those were (a) only IMD_2 (a quite unreasonable test for differential filters), (b) obtained at ±10 % of a given frequency, and (c) for lower-order filters with some gain. It is



Bioinspired CMOS Cochlea,
Fig. 25 Measured peak gain at CF and corresponding output THD versus I_{IN}



Bioinspired CMOS Cochlea,
Fig. 26 Measured two-tone third- and second-order intermodulation products of the fourth-order ($Q = 1$) OZGF for two frequencies at ±2 % away from 3 kHz

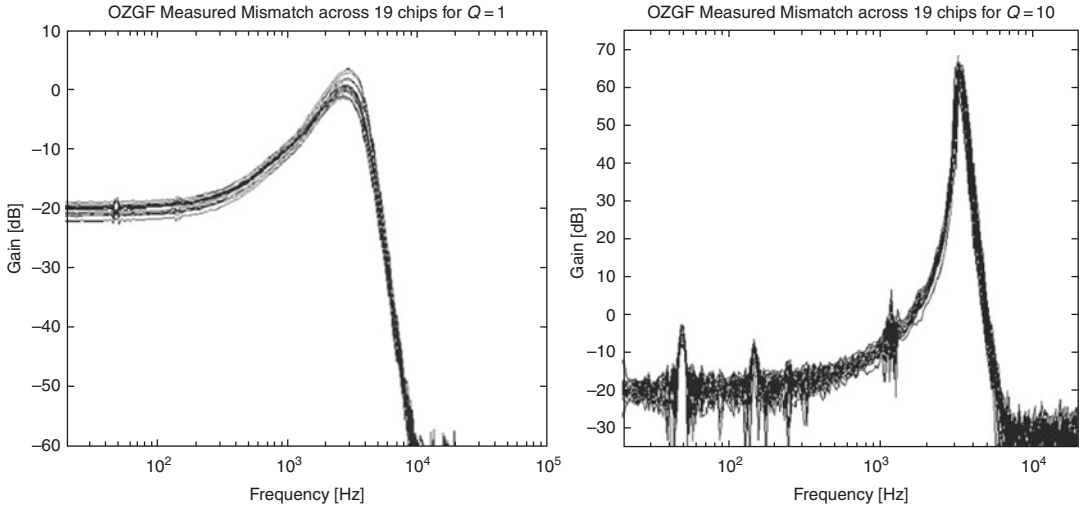


Bioinspired CMOS Cochlea, Fig. 27 Simulated two-tone third- and second-order intermodulation products of the second-generation, fourth-order ($Q = 1$) OZGF for two frequencies at ±2 % away from 3 kHz

believed that these set of results can be improved considerably in future implementations; nonetheless, one should appreciate that the real cochlea exhibits considerable intermodulation distortion, with both its $IMD_{2,3}$ levels around −20 dB for a large range of input intensities. In any case, we have managed recently to identify some potential optimization steps that may lead to an improved overall intermodulation linearity performance. A second-generation, fourth-order OZGF channel (one where its lossy BP stage is at the front of the cascade and not at the end) exhibited in simulation IMD_3 levels below −35 dB for small-signals, which also stayed below −25 dB across the whole input range (see Fig. 27).

Mismatch: One potential disadvantage of all pseudo-differential Class-AB filters of the type shown in Fig. 6 is that any potential mismatch between the two signal paths translates to distortion at the output. Therefore, considerable attention was given during layout in order to try and accurately match the two Class-A branches. Specifically, the two (upper and lower) TL comprising each Class-A biquad were nested together in such a way so that every upper transistor was interdigitized with its corresponding lower counterpart. Common centroid arrangements along both axes of symmetry were also employed. Figure 28 shows $Q = 1$ (upper plot) and $Q = 10$ (lower plot) measured OZGF frequency responses across 19 chips. Observe that in both cases the relative variation between responses does not exceed 6 dB.

Input Dynamic Range (DR): In this work, by input DR, we imply the ratio of the maximum input signal for a given allowable distortion at the output (measured in % of THD) over the noise floor for zero input. The measured input DR was calculated by taking into consideration the compressive action of the AGC by dividing the maximum signal at the input of the passive ($Q = 1$) OZGF response over the measured noise floor (integrated over the 3 dB bandwidth) of the fully active ($Q = 10$) OZGF response. The maximum input DR at CF was found to be 124 dB at <5 % THD, although at different frequencies, this figure changed since the distortion is not uniform across the whole bandwidth of the OZGF. More specifically, we observed an abrupt increase in THD around 1/3 of the CF; a problem partially



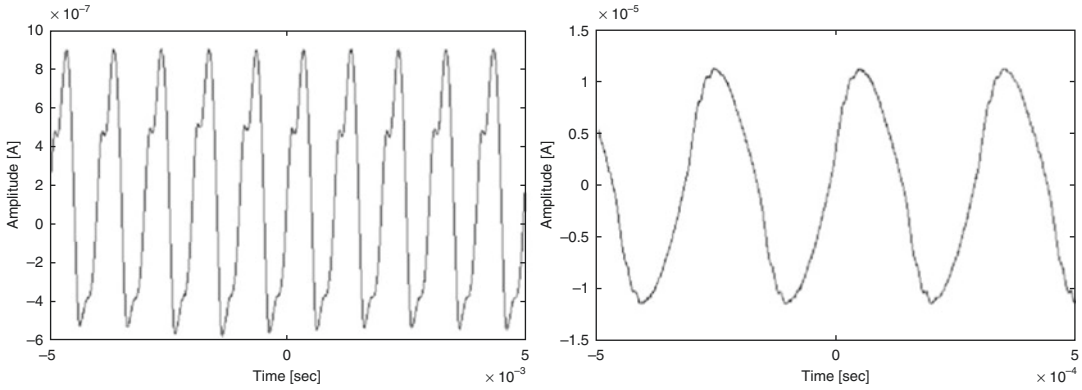
Bioinspired CMOS Cochlea, Fig. 28 OZGF mismatch across 19 chips for both extreme cases of Q

attributed to the fact that at 1/3 of the CF the input tone experiences a gain of ~ -10 dB, whereas the 3rd-harmonic (which is generated and propagated across the four Log-domain biquads) a gain of ~ 2 dB. *In other words, the fundamental gets suppressed, whereas the distortion gets amplified*. Moving deeper in the passband, one would expect that this particular problem would become more severe since both the third and fifth harmonics would get amplified while the fundamental would get suppressed even further. However, this is not entirely the case because for low frequencies, the whole structure operates virtually at DC, the capacitors draw smaller currents and, hence, the exerted nonlinearities are weaker.

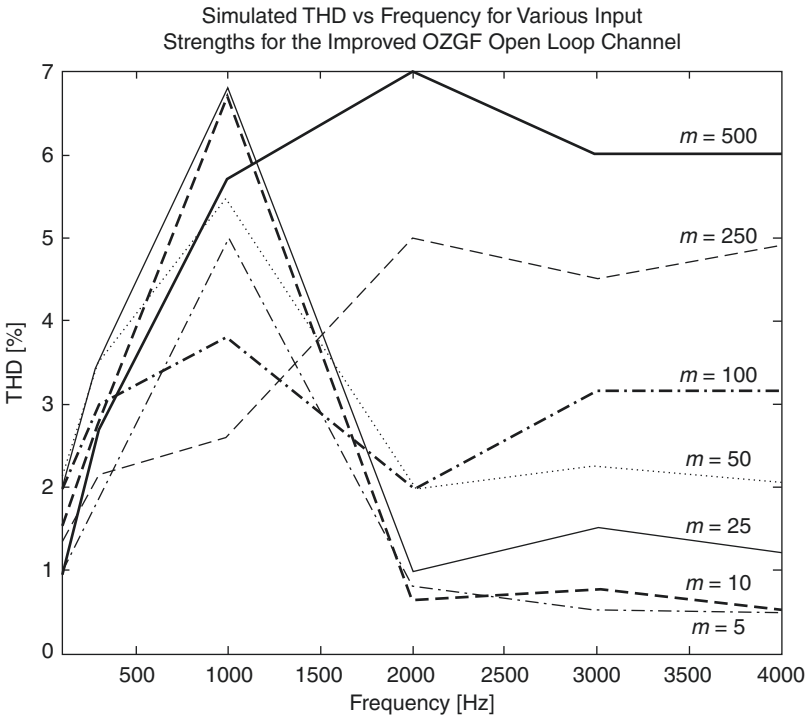
It is really interesting to notice that the aforementioned phenomenon (that of "amplified distortion" as opposed to the more common "distorted amplification") is also observed in the biological cochlea. Recent measurements in the basal and apical turns of the guinea-pig cochlea [2, 13] revealed that the real cochlea exhibits even-order harmonic content with maximum second-order harmonic distortion levels of 4 % and 28 % near and one octave below the CF, respectively. Since our implementation contains odd-harmonic content, we expect the maximum distortion to occur at 1/3 of the CF, which indeed was the case. Figure 29 illustrates large-signal transient waveforms at 1/3 of the CF and CF, respectively. Similar distortion occurred in the design of [17], where THD values of as high as 40 % were observed around one octave below the CF.

Figure 30 shows simulated results from the improved second-generation, fourth-order OZGF channel operating in open loop and with a $Q = 1$. The plot depicts THD values as a function of frequency with varying modulation index. Observe that for small m, the responses resemble HP filters, and BP filters for large m. The "sharpness" of these BP responses seem to increase with m, and their peak occurs approximately at 1/3 of the CF (around 1 kHz). Observe that for a wide range of input frequencies and strengths, the THD stays below 7 %.

Out-of-Band Interferer: A well-known problem of all ELIN topologies is that their noise floor is signal-dependent, resulting in a constant output SNR of about 60 dB for large m values. This observation was experimentally verified as well in our OZGF implementation. One manifestation of this problem is the ability of the filter to maintain good signal integrity at the presence of an out-of-bound interferer. Figure 31 shows the measured PSD of an 1 μA ($m = 50$) in-band signal at CF. An out-of-band interferer was placed at 20 kHz and for two distinct amplitudes, 1 μA



Bioinspired CMOS Cochlea, Fig. 29 Typical large-signal output waveforms from the OZGF channel at 1/3 of the CF (*upper*) and at the CF (*lower*), respectively. The THD in the *upper plot* is in excess of 20 %, whereas the distortion in the *lower plot* is less than 4 %

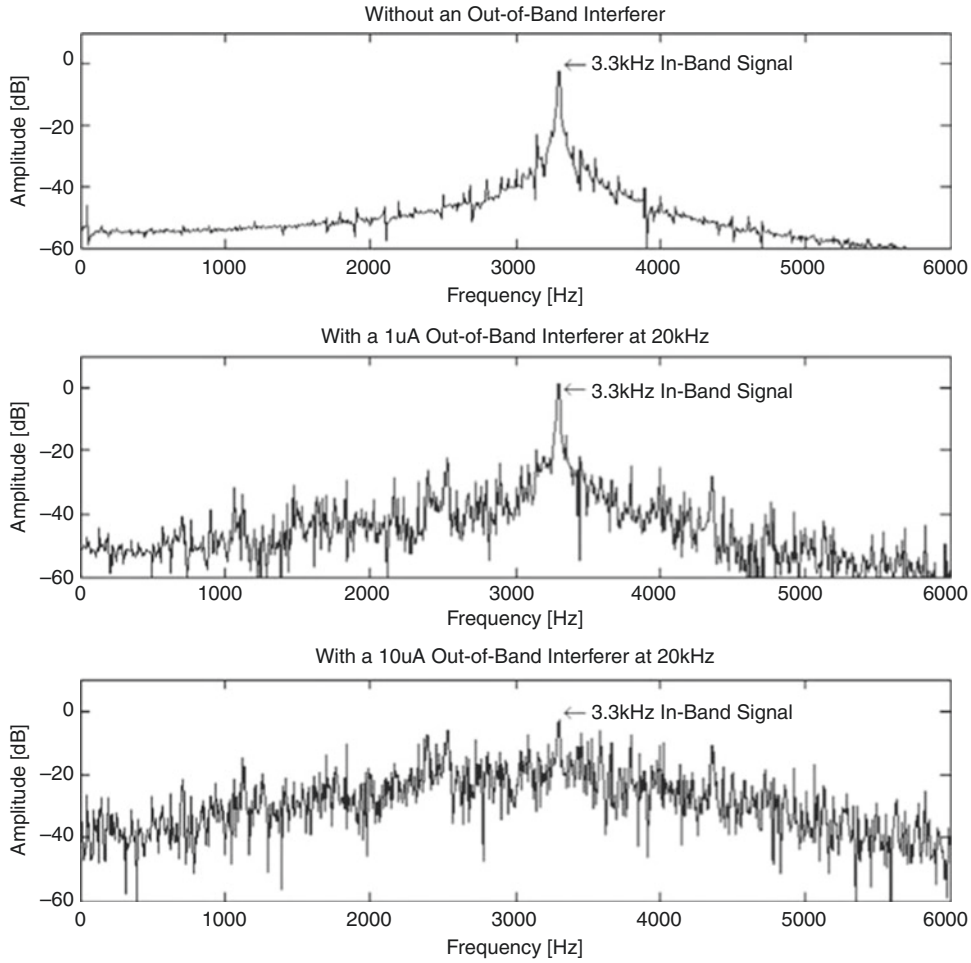


Bioinspired CMOS Cochlea, Fig. 30 Simulated THD versus frequency for varying modulation index *m*. For most of the characteristics, the maximum distortion occurs around 1 kHz (i.e., 1/3 of the CF)

and 10 μA. It was observed that the noise PSD increased from around −60 dB for the case of an absent interferer to about −40 dB for the case of the 10 μA interferer; that is about 10 dB per decade of interferer amplitude increase as predicted in [4] and experimentally verified in [16].

In conclusion, Table 2 summarizes the measured performance of the active fourth-order OZGF cochlea channel, whereas Fig. 31: Masking effect of an out-of-band interferer.

Figure 32 shows a chip micrograph of the channel's layout.



Bioinspired CMOS Cochlea, Fig. 31 Masking effect of an out-of-band interferer

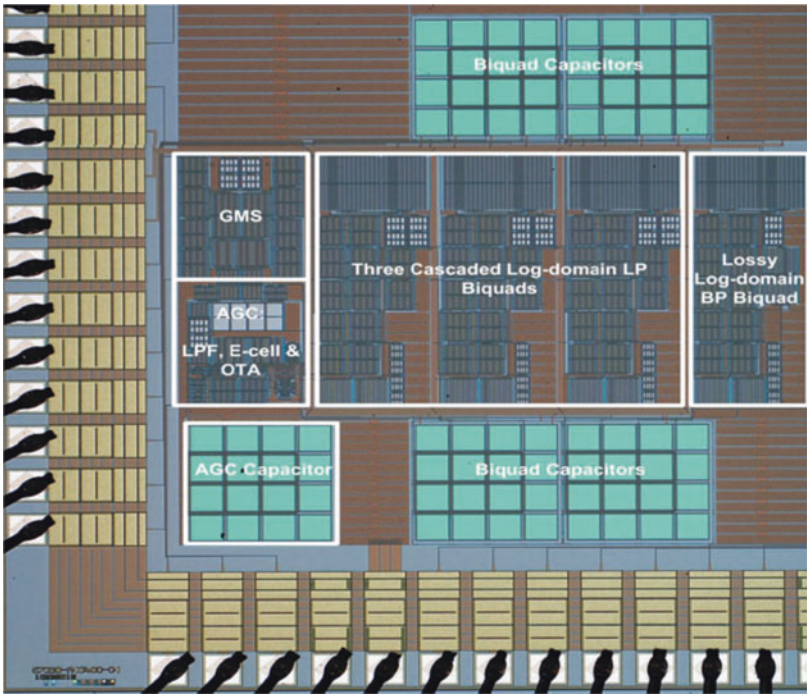
Bioinspired CMOS Cochlea, Table 2 OZGF-measured performance

Power consumption	3.4 μW (no AGC), 4.46 μW (with AGC)
Noise floor	67.5pA ($Q = 1$), 14pA ($Q = 10$)
Centre frequency	3.3 kHz ($Q = 1$)–3.7 kHz ($Q = 10$)
Peak gain	70 dB with <1 % THD
Input DR at CF	124 dB with <5 % THD (with AGC)
SNR at CF	∼60 dB for $m > 10$
Total on-chip capacitance	400 pF
Chip area	1.5 mm × 1.5 mm (2.25 mm^2)

Summary, Discussion, and Conclusion

The architectural choices and measured results of this cochlea channel implementation can be summarized in the following points:

1. Instead of opting for a filter-cascade implementation, which is prone to noise and offset accumulation, gain sensitivity, and yield, this design is based on separate filter-bank channels of short-cascades. In this way, it was possible to model the propagation of distortion products; conserve on computation; and



Bioinspired CMOS Cochlea, Fig. 32 Chip micrograph of the fourth-order OZGF channel with the AGC

obtain realistic roll-off slopes, amplitude, and group-delay responses without the associate disadvantages of long-cascade models.

2. The real cochlea exhibits 120 dB of input dynamic range at 3 kHz at <5 % THD. In this work, the measured results at the exact same frequency revealed an input DR of 124 dB at <5 % THD.
3. The BM in the real cochlea has an area of approximately 70 mm^2. This is equivalent to having a filter bank employing approximately thirty fourth-order OZGF channels.
4. A linear change in Q results in a logarithmic change in the peak gain (and in the temporal resolution) of our Log-domain OZGF response (see Fig. 21). Contrary to prior cochlea implementations, this is compatible with biology.
5. The implemented fourth-order OZGF frequency response achieves 70 dB of amplification for small-signals (even though in practice, we could get gains as high as 100 dB) while maintaining a linear low-frequency tail (20 dB/Dec) with steep high-frequency roll-off slopes (−160 dB/Dec).
6. The AGC dissipates 1 μW and compresses the six orders of magnitude of input signal intensity into one decade of Q-value tuning range. Moreover, it sets the lower and upper bounds of the peak-gain response avoiding excessive distortion or instability. In addition, asymmetric attack and release time constant programmability can be easily incorporated by modifying the LPF like in [22]. Lastly, in this implementation, the AGC acts in feed-forward and takes as input the global input. The downside of such a scheme is that the AGC does not contain any frequency information. The AGC can become frequency-dependent by modifying the architecture to accept as an input the output from the last BP OZGF stage or any other stage in between (in that case the AGC regulation will be closed-loop).

All the circuits presented in this work may serve as basic building blocks for the realization of new biorealistic cochlea processors designed in any of the two commonly used architectures; filter bank, or filter-cascade. Nonetheless, the fourth-order OZGF channel presented here is a good compromise between biorealism, circuit complexity, and yield and may be suitable in filter-bank

applications were biological fidelity and/or high DR performance is of primary interest. The choice for using CMOS-WI pseudo-differential Class-AB Log-domain biquads for low-power, high DR operation has proven successful at the expense of an increased chip area since the capacitor and transistor count increases by a factor of 2.

References

1. Binkley, D.M., Hopper, C.E., Tucker, S.D., Moss, B. C., Rochelle, J.M., Foty, D.P.: A CAD methodology for optimizing transistor current and sizing in analog CMOS design. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **22**, 225–237 (2003)
2. Cooper, N.P., Rhode, W.S.: Nonlinear mechanics at the apex of the guinea-pig cochlea. Hear. Res. **82**, 225–243 (1995)
3. Drakakis, E.M., Payne, A.J., Toumazou, C.: "Log-domain state-space": a systematic transistor-level approach for log-domain filtering. IEEE Trans. Circuits Syst. II Analog Digit. Signal Process. **46**, 290–305 (1999)
4. Enz, C., Punzenberger, M., Python, D.: Low-voltage log-domain signal processing in CMOS and BiCMOS. IEEE Trans. Circuits Syst. II Analog Digit. Signal Process. **46**, 279–289 (1999)
5. Fragniere, E., van Schaik, A., Vittoz, E.A.: Design of an analogue VLSI model of an active cochlea. Analog Integr. Circ. Signal Process. **13**, 19–35 (1997)
6. Frey, D.R.: Log-domain filtering: an approach to current-mode filtering. IEE Proc. G: Circuits Devices Syst. **140**, 406–416 (1993)
7. Frey, D.R.: Exponential state space filters: a generic current mode-design strategy. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. **43**, 34–42 (1996)
8. Frey, D.R., Tola, A.T.: A state-space formulation for externally linear class AB dynamical circuits. IEEE Trans. Circuits Syst. II Analog Digit. Signal Process. **46**, 306–314 (1999)
9. Frey, D.R., Tsividis, Y.P.: Syllabically companding log domain filter using dynamic biasing. Electron. Lett. **33**, 1506–1507 (1997)
10. Georgiou, J.: Personal Communication with Dr Julius Georgiou (2007)
11. Grech, I., Micallef, J., Vladimirova, T.: Low-power log-domain CMOS filter bank for 2-D sound source localization. Analog Integr. Circuits Signal Process. **36**, 99–117 (2003)
12. Katsiamis, A.G., Drakakis, E.M., Lyon, R.F.: Practical gammatone-like filters for auditory processing. EURASIP J. Audio Speech Music Process. 15 pp (2007)
13. Khanna, S.M., Hao, L.F.: Nonlinearity in the apical turn of living guinea pig cochlea. Hear. Res. **135**, 89–104 (1999)
14. Kimura, K.: The ultra-multi-tanh technique for bipolar linear transconductance amplifiers. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. **44**, 288–302 (1997)
15. Lyon, R.: A computational model of filtering, detection, and compression in the cochlea. In: Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'82, vol. 7, pp. 1282–1285 (1982)
16. Python, D., Enz, C.C.: A micropower class-AB CMOS log-domain filter for DECT applications. IEEE J. Solid-State Circuits **36**, 1067–1075 (2001)
17. Sarpeshkar, R., Lyon, R.F., Mead, C.: A low-power wide-dynamic-range analog VLSI cochlea. Analog Integr. Circuits Signal Process. **16**, 245–274 (1998)
18. Sarpeshkar, R., Salthouse, C., Ji-Jon, S., Baker, M.W., Zhak, S.M., Lu, T.K.T., Turicchia, L., Balster, S.: An ultra-low-power programmable analog bionic ear processor. IEEE Trans. Biomed. Eng. **52**, 711–727 (2005)
19. Tanimoto, H., Koyama, M., Yoshida, Y.: Realization of a 1-V active filter using a linearization technique employing plurality of emitter-coupled pairs. IEEE J. Solid-State Circuits **26**, 937–945 (1991)
20. Tsividis, Y.: On linear integrators and differentiators using instantaneous companding. IEEE Trans. Circuits Syst. II Analog Digit. Signal Process. **42**, 561–564 (1995)
21. Tsividis, Y.: Externally linear, time-invariant systems and their application to companding signal processors. IEEE Trans. Circuits Syst. II Analog Digit. Signal Process. **44**, 65–85 (1997)
22. Zhak, S.M., Baker, M.W., Sarpeshkar, R.: A low-power wide dynamic range envelope detector. IEEE J. Solid-State Circuits **38**, 1750–1753 (2003)

Bioinspired Microneedles

▶ Biomimetic Mosquito-Like Microneedles

Bioinspired Synthesis of Nanomaterials

Roberto de la Rica
Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, Scotland, UK

Synonyms

Biomimetic synthesis; Biomimetic synthesis of nanomaterials

Definition

Synthetic routes that mimic natural processes for the fabrication of nanomaterials in mild conditions and with improved control over key features such as size, shape, and hierarchical organization of the end product. These bioinspired processes frequently utilize biocatalytic templates, such as DNA, peptides, and proteins as well as cell bioreactors, whose performance can be enhanced via molecular biology tools.

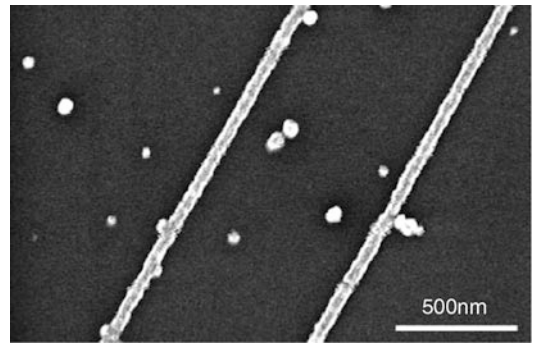
The transition from the microscale to the nanoscale has been a great endeavor for mankind, which has opened a new horizon in technological applications due to the superior features of nanomaterials compared to their bulk counterparts. Although major breakthroughs in nanotechnology have been accomplished, some areas of this field remain challenging for the widespread application of these materials in everyday life. For example, the organization of nanomaterials to yield hierarchical structures with defined shapes and sizes could generate materials with novel functions derived from the spatial arrangement of the nanometric building blocks. This feature is also essential for the easy integration of nanomaterials with complex circuits and microprocessors, a crucial step for electronic applications such as sensors and solar cells. The fabrication in mild conditions such as at room temperature and in aqueous solution would greatly help disseminating the utilization of nanomaterials in real applications, since less energy would be required for the fabrication and less toxic waste generated during the manufacture process, therefore reducing costs and environmental impact related to their synthesis. To address these issues, many scientists have searched for inspiration in biological systems. Living organisms have evolved the capability of growing inorganic nanomaterials with an exquisite control over the shape, size, crystal phase, and three-dimensional arrangement. These biological processes are orchestrated by a battery of biomolecules that act as biocatalytic templates for the synthetic process. Moreover, these biological phenomena take place under mild conditions, that is, at room temperature and in aqueous solution, which are not only advantageous from an economic and ecological standpoint but can also result in the fabrication of materials with improved properties such as defect-free crystals. Usually, a bioinspired route begins from the observation of a natural process and its replication ex vivo, for example, by recreating the fabrication in the presence of biomolecules found to be crucial in the biological synthesis. Subsequently, the synthesis can be refined by generating alterations on the biocatalytic templates via expression of mutant DNA sequences in microorganisms. This approach is particularly advantageous because it allows one to create a vast array of catalytic candidates, whose appropriateness for the fabrication of the target material can be tested with high-throughput methodologies. Furthermore, once a particular template has been recognized, its production can be easily scaled up by selecting the microorganism encoding the biomolecule and making it grow in adequate conditions. Subsequently, the target biomolecule can be isolated by well-known purification protocols. The utilization of biomolecules is not only advantageous for the synthesis of nanomaterials in benign conditions and with desired shapes and sizes; the intrinsic recognition capability of certain biomolecules can be also exploited to target them at particular sites on substrates for their integration with circuits and microprocessors, which is essential for the implementation of nanomaterials in real applications. Finally, bioinspired approaches have been described that mimic the biological mechanism of formation of inorganic materials rather than utilizing biomolecules as biocatalytic templates. Among these, biomimetic approaches for the fabrication of perfectly aligned arrays of nanomaterials, also called mesocrystals, are drawing much attention since the resulting superstructures could show novel physical properties derived from the intrinsic order of the collectives.

In the following sections, the relevance of several biocatalytic templates such as DNA, peptides, and proteins will be outlined and illustrated with key examples. The utilization of cells for the synthesis of nanomaterials as well as other biomimetic approaches for the

fabrication of mesocrystals will also be covered, followed by conclusions and future directions in the field.

DNA

DNA molecules have been extensively used as templates for the synthesis of nanomaterials due to their inherent biorecognition capabilities and excellent chemical properties for the deposition of metallic and semiconducting materials [1]. The recognition of DNA strands bearing complementary sequences has proved to be extremely useful for the programmable assembly of DNA templates on substrates, which is an important step for the integration of nanomaterials with common technologies. Moreover, the negatively charged phosphate groups of DNA molecules can concentrate metal ion precursors in the vicinity of the template for the selective deposition of materials. In certain cases, the coordination of these metal ions to the bases of DNA can result in the formation of nanocrystals with unique properties. An early example on the advantages of using DNA molecules as templates for the fabrication of nanomaterials can be found in the work of Ben-Yoseph et al. [2]. By immobilizing oligonucleotides bearing a complementary sequence on two adjacent gold electrodes, the positioning of the DNA template was programmed to happen at the gap between the electrodes. Subsequently, silver ions were concentrated in the vicinity of the DNA bridge via electrostatic interactions, and reduction with hydroquinone yielded silver nanowires of 12 μm length that bridged the electrodes. However, the metal nanowires prepared this way showed poor conductivity mainly due to the presence of granules and defects as a consequence of the fast-growth kinetics of metallization in the aforementioned conditions. More recently it was demonstrated that the conductivity of the metal coatings on DNA templates could be greatly improved by first depositing a metal oxide such as PdO and subsequently reducing the intermediate with hydrogen gas [3]. The resulting Pd nanowires are continuous and show higher conductivities (Fig. 1).



Bioinspired Synthesis of Nanomaterials, Fig. 1 Pd nanowires templated by DNA (Reprinted from Nguyen et al. [3], copyright Wiley-VCH Verlag GmbH & Co. KGaA)

Short DNA sequences have proven to be also very useful in the formation of extremely small metal nanoparticles, the so-called metal nanoclusters. When the particle size is further reduced and approaches the Fermi wavelength of electrons, the continuous density of states breaks up into discrete energy levels leading to the observation of dramatically different optical, electrical, and chemical properties compared to larger nanoparticles. For example, Ag nanoclusters are highly fluorescent, and their reduced dimensions compared to quantum dots make them ideal candidates for bioimaging provided that they are soluble in aqueous solution. By using cytosine-rich DNA sequence as templates, Dickson et al. demonstrated the formation of Ag nanoclusters with 2–4 atoms [4]. Contrary to former approaches where the interaction between the negative-charged phosphate backbone and metal ions was the driving force for the selective growth of the material, the sequence of the DNA templates plays a crucial role in the formation of metal nanoclusters. The utilization of biotechnology tools can greatly speed up the screening of DNA sequences that are useful for the fabrication of a particular material. For example, short DNA sequences can be chemically synthesized in a random fashion and spotted on a glass substrate to generate DNA microarrays. After incubation with silver and reduction, the spots containing DNA sequences capable of growing silver nanoclusters can be detected by measuring the

fluorescence of the whole substrate with a scanner. By this methodology, blue-, green-, yellow-, and red-emitting Ag nanoclusters could be synthesized from a DNA library [5].

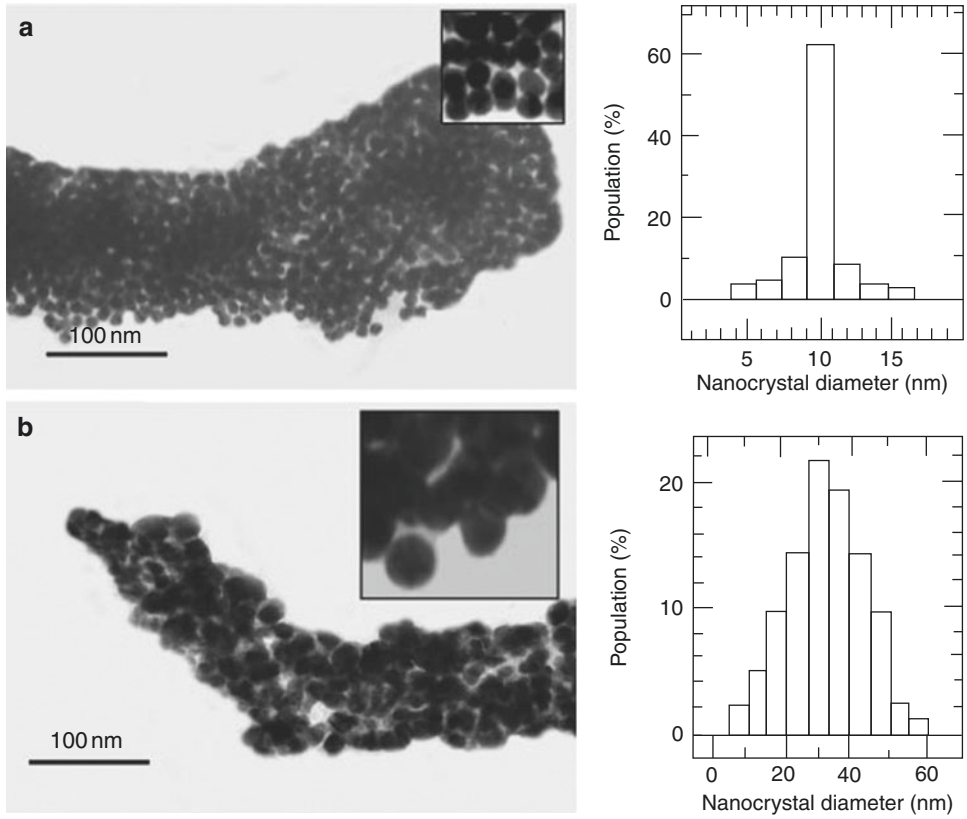
Peptides

Peptides have proved to be extremely useful tools for the fabrication of nanomaterials [6]. Compared to proteins, they possess shorter sequences, which facilitate their fabrication via synthetic methods. This feature also makes their 3D structure simpler and more predictable when synthesizing de novo sequences. Moreover, when carefully designed, peptides can assemble to yield supramolecular architectures with novel functions. A well-established area in nanomaterial fabrication synthesis is the application of peptides for growing nanocrystals with a particular crystalline structure, shape, and size. The selection of peptide sequences for the fabrication of nanocrystals is usually performed via molecular biology tools such as phage display. In this approach, random DNA sequences encoding peptides are inserted in bacteriophages. When the viruses replicate, they express the peptides on their capsids. Subsequently, the affinity of the peptide for a particular material is tested by applying stringent selection protocols. The peptides with the highest affinity for the target material are tested for the fabrication of nanocrystals in different conditions. Although the mechanism behind this bioinspired approach is not clear yet, it has been demonstrated that the amino acid composition and the 3D arrangement of chemical moieties dictated by the conformation of the peptides are crucial for the fabrication of nanocrystals. For example, Matsui et al. applied peptides with the sequence His-Gly-Gly-Gly-His-Gly-His-Gly-Gly-Gly-His-Gly for the fabrication of Cu nanocrystals on peptide nanotubes [7]. They found that the peptides could undergo a conformational change depending on the pH of the growing solution that could control the size and monodispersity of the resulting nanocrystals (Fig. 2). The same group demonstrated that peptides with the sequence Asn-Pro-Ser-Ser-Leu-Phe-Arg-Tyr-Leu-Pro-Ser-Asp could catalyze the growth of hexagonal Ag nanocrystals. The selectivity for growing a particular shape was provided by the high affinity of the peptide for the face (111) of Ag crystals.

Certain peptide sequences have been reported to self-assemble and yield complex architectures with catalytic functions. For example, peptides derived from naturally occurring leucine-zipper motifs have been described to fold in a coil-coil conformation that can assemble into templates for crystal growth. By changing key amino acids of the primary sequence for thiol-rich cysteine units, the supramolecular peptide aggregates can be used as templates for the fabrication of monodisperse Ag nanoparticles [8]. In another example, the self-assembly of bolaamphiphile peptides resulted in the formation of peptide "nanodoughnuts" that could catalyze the fabrication of Ga_2O_3 [9]. Chemical residues with affinity for the metal ions preconcentrated the precursors inside the cavity of the templates and generated a gallium hydroxide intermediate. The higher hydrophobicity inside the template played a crucial role in the dehydration of the intermediate to yield crystalline Ga_2O_3, a wide bandgap semiconductor of great interest in electronics. These examples highlight the ability of peptide assemblies to work as catalytic templates for crystal growth.

Proteins

When compared to peptides, proteins show a more complex tertiary structure and can perform highly specialized tasks such as catalyzing specific chemical reactions [6]. Therefore, the adaptation of a protein for the synthesis of a nanomaterial can be a very advantageous approach provided that it can be easily purified from natural sources in large amounts. Although the higher degree of complexity seen in proteins makes it more challenging to modify their structure to gain a new function, recent advances in molecular biology have enabled the fabrication of mutants with improved features for the fabrication of nanomaterials. The rich structural variety of proteins can be exploited for the fabrication of



Bioinspired Synthesis of Nanomaterials, Fig. 2 Peptide-templated Cu nanocrystals. The conformation of the peptide dictated by the pH of the growing solution controls the size and shape of the nanocrystals; (**a**) pH 6 and (**b**) pH 8 (Reprinted from Banerjee et al. [7], copyright The National Academy of Sciences of the USA)

nanomaterials with desired shape and dimensions when these proteins act as template for the synthetic process. For example, the protein ferritin is formed by 24 subunits that generate a cavity with a diameter of 7 nm. In cells, it contains iron crystallites and hence serves as an important intracellular iron storage unit. In vitro, iron-free ferritin (apoferritin) can be loaded with metal ions to generate nanoparticles of materials such as iron and cobalt oxide, which grow inside the cavity with uniform size and shape. This feature is especially advantageous compared to classical methods for nanoparticle synthesis, since in the absence of a template it is difficult to predict the final shape and size of the nanocrystals. For instance, in the fabrication of nanoparticle-based floating gates, it is crucial that all the building blocks have a precise size, shape, and surface density. These requirements could be met by synthesizing CoO_3 nanoparticles in apoferritin and immobilizing the protein-covered nanocrystals on the device via programmed electrostatic interactions [10]. The resulting floating gate memory showed excellent charge capacity, long charge retention, and stress resistance. Moreover, it has been demonstrated that the inner cavity of the cage can be modified via genetic engineering protocols to display peptide sequences that have high affinity for a certain metal ion precursor, which paves the way for the utilization of ferritin in the fabrication of nanocrystals of diverse technologically relevant materials.

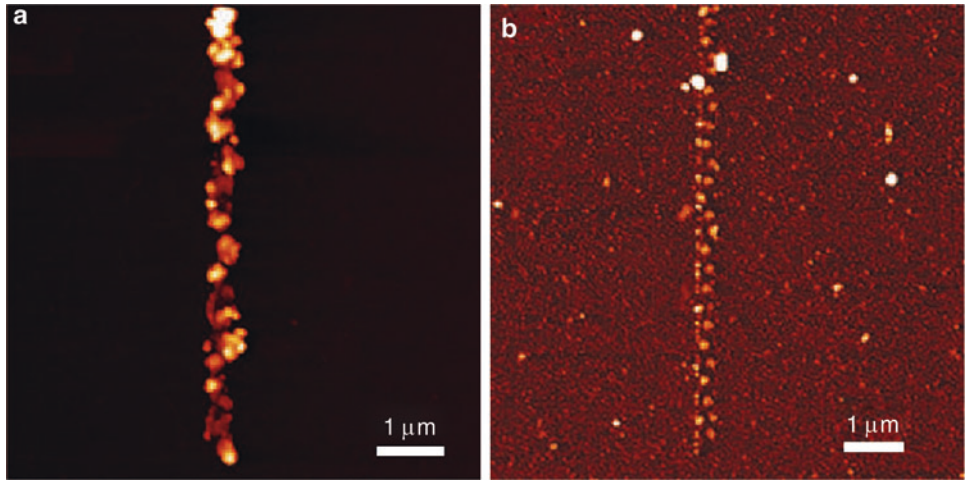
While protein cages have proved to be extremely useful for the fabrication of nanocrystals, nanostructures with a higher aspect ratio such as nanowires are required for a wide

variety of nanotechnology applications. To address this issue, a fibrillar protein such as collagen could be used as a template for the deposition of metallic or semiconducting materials. However, naturally occurring collagen may not be suitable for this application because the relatively harsh conditions for metallization can affect the conformation of the protein, therefore rendering nanomaterials of polydisperse size. This issue could be overcome by using a collagen-like triple helix protein that is genetically engineered to possess improved stability and rigidity, as well as uniform size. When used as a template for the electroless deposition of gold, the mutant protein could template the fabrication of gold nanowires of monodisperse size with a length of 40 nm and a width of 4 nm [11]. The fabrication of this protein via expression in a host microorganism such as *Escherichia coli* allows one to obtain large quantities of the biomolecular template by simply growing the bacteria and applying well-established protein purification protocols.

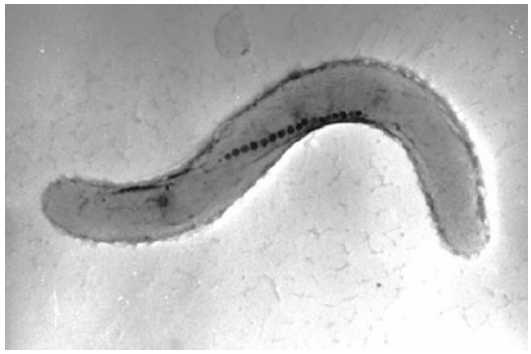
An emerging area in bionanotechnology is the adaptation of proteins with biocatalytic activity such as enzymes to the synthesis of nanomaterials because the mild conditions required for the biological catalysis make the manufacturing process more environmentally benign and, in some cases, with less side products. In a pioneering work by Samuelson et al., the enzyme peroxidase was utilized for the synthesis of the conductive polymer polyaniline [12]. In this approach, peroxidase catalyzed the oxidation of the monomer to trigger the polymerization in the presence of a polyelectrolyte template to yield water-soluble, high-molecular-weight polyaniline with minimal parasitic branching. The approach is particularly attractive in that it is simple (one step), uses very mild conditions, and requires minimal separation and purification. Another area that has benefited greatly from the use of enzymes for the fabrication of nanostructures is the room-temperature synthesis of oxide semiconductors. Oxide semiconductors such as ZnO are wide bandgap semiconductors with a myriad of applications in optics, electronics, solar cells, and catalysis. However, the high temperatures required for the fabrication of these materials with conventional methods increase the cost associated to the manufacturing process, which may hamper the widespread application of these materials. This issue could be solved by using the enzyme urease as biocatalytic template for the synthesis of ZnO nanoshells. In this approach, urease converted urea into ammonia and carbon dioxide, and the resulting fine-tuning of the pH at the enzyme-solution interphase could grow highly crystalline ZnO as nanoshells around the protein [13]. When immobilized on a substrate, the enzymatic production of ammonia could be harnessed to produce amorphous intermediates that could be used as ink to be patterned with an AFM tip for a pen in a procedure similar to dip-pen nanolithography [14]. By this technique called Biomimetic Crystallization Nanolithography, ZnO nanocrystals could be patterned on the surface with desired shapes and sizes, which is a crucial step for the integration of nanomaterials with electrical components of circuits and microprocessors (Fig. 3). Moreover, urease could be also applied for the fabrication of other relevant semiconductors such as metal sulfides by harnessing the production of S^{2-} via hydrolysis of thiourea. By this approach, the size, shape, and crystallinity of Ag_2S as a test material could be controlled to be as desired by fine-tuning the activity of the enzyme via inhibition with the metal ion precursor [15].

Cells

The utilization of cells such as bacteria for the fabrication of nanomaterials can be very advantageous since the fast-growth kinetics of these microorganisms in adequate conditions allows one to obtain large amounts of the target material with minimal resources. A well-known example of the application of cells in nanomaterials synthesis is the fabrication of magnetic nanoparticles by magnetotactic bacteria [16]. These microorganisms generate intracellular magnetic particles of iron oxide that allow them to migrate along oxygen gradients by acting as compass needles. Different bacterial strains have been reported to grow these bacterial magnetic particles in different shapes, which can be very advantageous for



Bioinspired Synthesis of Nanomaterials, Fig. 3 ZnO nanopatterns via biomimetic crystallization nanolithography with different tip velocities during the patterning step; (**a**) $v = 100$ nm/s; (**b**) $v = 500$ nm/s (Reprinted from de la Rica et al. [14], copyright Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim)

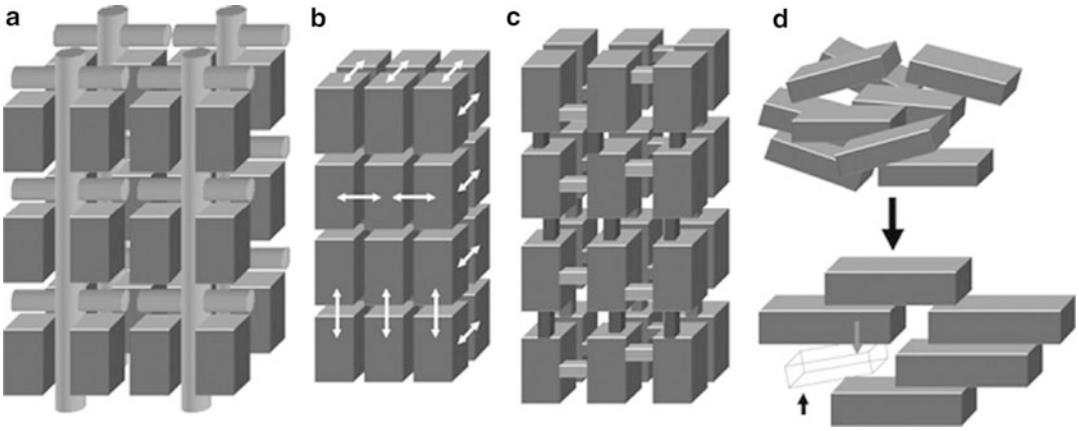


Bioinspired Synthesis of Nanomaterials, Fig. 4 Image of *Magnetospirillum magneticum* strain AMB-1 showing the magnetic nanoparticles (Reprinted from Arakaki et al. [16], copyright the Royal Society)

obtaining materials with unusual nanostructures (Fig. 4). For example, the *M. magneticum* strains AMB-1 and MGT-1 synthesize cubo-octahedral nanoparticles, whereas *Desulfovibrio magneticus* RS-1 produces irregular bullet-shaped particles. Bacterial magnetic particles also have the great advantage of being synthesized surrounded by an organic membrane, whose main components are phospholipids and proteins. These biomolecules confer better solubility to the particles, which disperse well in aqueous solvents without further surface treatment. Moreover, the biomolecules present at the surface of bacterial magnetic particles show chemical moieties that can be used as anchoring points for their functionalization with a wide variety of molecules such as enzymes and antibodies. Particularly, the amino groups of the phospholipids present at the membrane can be cross-linked to antibodies and oligonucleotides with well-known covalent chemistries, which confer the particles biomolecular recognition capabilities that are useful for medical and diagnostic applications. Alternatively, the protein of interest can be generated during the biological synthesis of bacterial magnetic particles by fusing its gene to proteins known to be present on the membrane of the nanoparticles so that when the bacteria express the resulting fusion protein, it is automatically incorporated on the surface of the nanoparticles during the biosynthesis.

Mesocrystals

Biomineralizing organisms such as sea urchins and marine stars have evolved the capability to generate nanocrystals and arrange them into highly ordered mesoscopic arrangements, also known as mesocrystals [17]. The perfect alignment of the constituent nanocrystals results in similar scattering pattern and behavior in



Bioinspired Synthesis of Nanomaterials, Fig. 5 Scheme of the different mechanisms to grow nanoparticle superstructures or mesocrystals (Reprinted from Song and Colfen [17], copyright Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim)

polarized light to those of a single crystal, although a close look into the ultrastructure reveals the presence of the small constituent nanoparticles. This phenomenon is responsible for the improved mechanical properties of the exoskeleton of crustaceans, and the replication ex vivo for the fabrication of assemblies of perfectly aligned nanocrystals of functional materials is extremely interesting to study the physical properties of nanoparticle collectives. Several mechanisms have pointed out to yield three-dimensional nanoparticle superstructures, as depicted in Fig. 5. One possibility is that crystal growth takes places inside an organic matrix such as a biopolymer, which guides the growth and alignment of the constituent nanocrystals (Fig. 5a). The presence of certain chemical groups with affinity for cationic precursors is important to initiate the nucleation of the nanoparticles inside the matrix and, in some cases, to stabilize a particular face of the crystal and direct the growth in a particular direction. Another possibility is the presence of physical fields as the driving force for nanoparticle aggregation, for example, the spontaneous alignment of dipoles (Fig. 5b). The growth of mesocrystals can also be permitted by the existence of the so-called mineral bridges, that is, the nanoparticles are held together by another inorganic structure in a "brick and mortar" fashion (Fig. 5c). Finally, when growing in a constraint reaction vessel, nanoparticles can be forced to interact and align with each other to reduce the entropy of the system (Fig. 5d). Although these basic rules for the formation of mesocrystals have been successfully replicated ex vivo by the correct choice of crystallization environment such as the presence of a polymer matrix or addition of soluble additives, the exact mechanism of this nanoparticle-mediated crystallization route remains still elusive. For example, nanospheres formed by ZnO nanocrystals oriented along the [001] direction could be obtained by using the polymer poly(vinylpyrrolidone) in a mixture of N,N-dimethylformamide (DMF) and water as a guiding matrix. The coordination of Zn^{2+} to active groups in the polymer and the stabilization of particular crystal planes play a crucial role for obtaining nanocrystal-based superstructures. The resulting porous arrangements of aligned nanocrystals are promising for photonic applications. Moreover, it is possible to fabricate composite thin films via self-assembly of the freestanding spheres by simply coating with the as-prepared suspensions, which facilitates the integration of these hybrid materials with planar devices. Recently, the biomimetic conditions for the fabrication of mesocrystals could be replicated by a novel approach in which a nanopore building block was self-assembled to yield a highly porous matrix. The macrocycle cucurbituril [7]

aggregated in the presence of Ag^+ and the ion-impregnated nanopore assemblies guided the growth of perfectly aligned Ag_2S quantum dots after the addition of S^{2-} to yield a nanoparticle superstructure with a pitch of 0.8 nm. This example represents an exquisite control over the crystallization of this material due to the extremely low solubility of Ag_2S in water, which usually results in the formation of particles of uncontrolled size and poor crystallinity. The key step of the process was to confine the metal ion precursor inside the porous matrix, which controls mass transport via diffusion of the anions through the pores of the assemblies [18]. Recently the enzyme glucose oxidase was utilized to grow mesocrystals containing gold nanoparticles under bioinspired conditions [19]. In this work the enzyme was patterned on a surface as lines separated by sub-micrometric distances. The lines were obtained by patterning the polymer polyethylene glycol with electron beam lithography. The polymer possessed carboxylic acid groups that allowed the covalent attachment of the enzyme on its surface. The enzyme generates hydrogen peroxide that tunes the kinetics of growth of gold nanoparticles. It was found that when the enzyme lines were separated by nanometric distances, the generation of hydrogen peroxide by the enzyme depended not only on the enzyme activity and mass transport but also on the nanoscale organization. This favored bioinspired nonclassical crystal growth conditions in which the nanoparticles grow by aggregation of nanometric building blocks. Clusters containing highly organized gold nanoparticles could be obtained with this method.

Conclusion and Outlook

The utilization of bioinspired approaches in the fabrication of nanomaterials is extremely useful to control key parameters such as size monodispersity, shape, and three-dimensional arrangement. These processes usually take place under mild biological conditions, which is advantageous for reducing costs and environmental impact of the manufacturing process. The utilization of microorganisms as bioreactors, whether to generate the biocatalytic template or to synthesize the target nanomaterial, facilitates the large-scale production due to the rapid growth kinetics of cells in adequate media. Moreover, in certain cases, the intrinsic recognition capabilities of the biomolecule templates can be exploited to locate the biomaterials at target places on substrates, which facilitate their integration with common technologies. While these features have been demonstrated in the last years, challenges in the field still remain. From a fundamental standpoint, it would be desirable to unravel the exact mechanism of formation of nanomaterials in biological conditions. A deeper understanding of the process would allow the formulation of rules for bioinspired synthesis so that the outcome of the manufacturing process is more predictable, and hence the fabrication of the nanomaterial requires less effort. For example, discerning the relationship between sequence and conformation in peptide and DNA templates for the fabrication of nanomaterials with specific shape, size, and polymorph could help designing de novo sequences for the synthesis of another material by applying lessons learnt from previous experiences. Another important step toward the widespread utilization of bioinspired routes would be to demonstrate the fabrication of real devices with the resulting nanomaterials. Although in the last years several key applications have been demonstrated such as the fabrication of floating memories [10] and the design of ultrasensitive sensors for the detection of heavy metal ions via target-specific bioinspired crystallization [20], the integration of nanomaterials obtained by bioinspired routes is still challenging in certain fields. For example, a common problem is that these nanomaterials are usually covered by the biological template. While this can be advantageous for their programmed assembly via biorecognition, the presence of an insulating organic layer can hamper certain applications such as in the fabrication of electronic components of circuits, in which the conductance of the nanomaterials and the electric contact with the circuitry is crucial for the performance of the device. Nevertheless, biocatalytic templates are extremely useful components of the

nanotechnology toolbox that show great potential to change classical routes for the synthesis of nanomaterials to be more environmentally benign and that offer improved control over fey features such as size, shape, and hierarchical organization.

Cross-References

► Dip-Pen Nanolithography

References

1. Becerril, H.A., Woolley, A.T.: DNA-templated nanofabrication. Chem. Soc. Rev. **38**, 329–337 (2009)
2. Braun, E., Eichen, Y., Sivan, U., Ben-Yoseph, G.: DNA-templated assembly and electrode attachment of a conducting silver wire. Nature **391**, 775–778 (1998)
3. Nguyen, K., Monteverde, M., Filoramo, A., Goux-Capes, L., Lyonnais, S., Jegou, P., Viel, P., Goffman, M., Bourgoin, J.P.: Synthesis of thin and highly conductive DNA-based palladium nanowires. Adv. Mater. **20**, 1099–1104 (2008)
4. Petty, J.T., Zheng, J., Hud, N.V., Dickson, R.M.: DNA-templated Ag nanocluster formation. J. Am. Chem. Soc. **126**, 5207–5212 (2004)
5. Richards, C.I., Choi, S., Hsiang, J.C., Antoku, Y., Vosch, T., Bongiorno, A., Tzeng, Y.L., Dickson, R.M.: Oligonucleotide-stabilized Ag nanocluster fluorophores. J. Am. Chem. Soc. **130**, 5038–5039 (2008)
6. de la Rica, R., Matsui, H.: Applications of peptide and protein-based materials in bionanotechnology. Chem. Soc. Rev. **39**, 3499–3509 (2010)
7. Banerjee, I.A., Yu, L., Matsui, H.: Cu nanocrystal growth on peptide nanotubes by biomineralization: size control of Cu nanocrystals by tuning peptide conformation. Proc. Natl. Acad. Sci. U. S. A. **100**, 14678–14682 (2003)
8. Ryadnov, M.G.: A self-assembling peptide polynanoreactor. Angew. Chem. Int. Ed. **46**, 969–972 (2007)
9. Lee, S.Y., Gao, X., Matsui, H.: Biomimetic and aggregation-driven crystallization route for room-temperature material synthesis: growth of β-Ga_2O_3 nanoparticles on peptide assemblies as nanoreactors. J. Am. Chem. Soc. **129**, 2954–2958 (2007)
10. Yamashita, I., Iwahori, K., Kumagai, S.: Ferritin in the field of nanodevices. Biochim. Biophys. Acta **1800**, 846–857 (2010)
11. Bai, H., Xu, K., Xu, Y., Matsui, H.: Fabrication of Au nanowires of uniform length and diameter using a monodisperse and rigid biomolecular template: collagen-like triple helix. Angew. Chem. Int. Ed. **46**, 3319–3322 (2007)
12. Liu, W., Kumar, J., Tripathy, S., Senecal, K.J., Samuelson, L.: Enzymatically synthesized conducting polyaniline. J. Am. Chem. Soc. **121**, 71–78 (1999)
13. de la Rica, R., Matsui, H.: Urease as a nanoreactor for growing crystalline ZnO nanoshells at room temperature. Angew. Chem. Int. Ed. **47**, 5415–5417 (2008)
14. de la Rica, R., Fabijanic, K., Matsui, H.: Biomimetic crystallization nanolithography: simultaneous nanopatterning and crystallization. Angew. Chem. Int. Ed. **49**, 1447–1450 (2010)
15. Pejoux, C., de la Rica, R., Matsui, H.: Biomimetic crystallization of sulfide semiconductor nanoparticles in aqueous solution. Small **6**, 999–1002 (2010)
16. Arakaki, A., Nakazawa, H., Nemoto, M., Mori, T., Matsunaga, T.: Formation of magnetite by bacteria and its application. J. R. Soc. Interface **5**, 977–999 (2008)
17. Song, R.Q., Colfen, H.: Mesocrystals – ordered nanoparticle superstructures. Adv. Mater. **22**, 1301–1330 (2010)
18. de la Rica, R., Velders, A.H.: Biomimetic crystallization of Ag_2S nanoclusters in nanopore assemblies. J. Am. Chem. Soc. **133**, 2875–2877 (2011)
19. de la Rica, R., Bat, E., Herpoldt, K.L., Xie, H., Bertazzo, S., Maynard, H., Stevens, M.M.: Nanoparticle growth via concentration gradients generated by enzyme nanopatterns. Adv. Funct. Mater. **24**, 3692–3698 (2014)
20. de la Rica, R., Mendoza, E., Matsui, H.: Bioinspired target-specific crystallization on peptide nanotubes for ultrasensitive Pb ion detection. Small **6**, 1753–1756 (2010)

Biological Breadboard Platform for Studies of Cellular Dynamics

Sang-Hee Yoon[1] and Mohammad R. K. Mofrad[2]
[1]Department of Mechanical Engineering, Inha University, Incheon, Republic of Korea
[2]Molecular Cell Biomechanics Lab, Department of Bioengineering, University of California, Berkeley, CA, USA

Synonyms

Biological platform; Cell adhesion and detachment; Cell migration; Cellular dynamics

Definition

A biological breadboard (BBB) is a platform that manipulates the adhesion and detachment of anchorage-dependent cells in a spatiotemporal

way, thus ultimately controlling a variety of cell behaviors such as cell migration, cell positioning, cell patterning, etc. The BBB is composed of an array of identical gold microelectrodes microfabricated on a Pyrex glass substrate. The gold microelectrodes are functionalized with arginine-glycine-aspartic acid (RGD)-terminated thiol (RTT) to create "cell-adhesive" surfaces; the Pyrex glass substrate is modified with polyethylene glycol (PEG) to achieve a "cell-repulsive" surface. The BBB manipulates cell adhesion and detachment by spatiotemporally controlling the reductive desorption of n-alkanethiolate self-assembled monolayers (SAMs) on gold with activation potential of -0.90 to -1.80 V. This biological platform can offer a new, versatile tool for studies of cellular dynamics, analogous to a breadboard in electronics.

Overview

Cell adhesion and detachment processes are mediated by complex biomolecules from both sides of the cell-matrix interface, fitting together like the pieces of a three-dimensional puzzle. Only biomolecules with the right shape can be involved in cell adhesion and detachment processes, and even very small changes in the shape may significantly affect cell behavior. When an anchorage-dependent cell adheres to extracellular matrix (ECM) components, integrins are activated; the activated integrins bind target ligands; the bound integrins cluster together by changing their conformation [1]. The cytoplasmic domain of the clustered integrins interacts with focal adhesion (FA) proteins (e.g., talin, focal adhesion kinase, vinculin, paxillin, etc.) to form FAs, followed by actin filament binding [2]. Likewise, cell detachment (or deadhesion) is also essential to many biological phenomena in cell biology (e.g., migration, metastasis, etc.). Cell detachment is known to result from an orchestrated process involving the molecular machinery composed of a host of extracellular, transmembrane, and cytoplasmic proteins.

Cell adhesion and detachment have profound effects on the behavior of anchorage-dependent cells. For example, cell adhesion and detachment are controlling parameters in a variety of biological phenomena such as embryonic development, cancer metastasis, wound healing, etc., and any abnormality in cell adhesion and detachment could lead to serious pathophysiological consequences [1]. The quantitative characterization of cell adhesion and detachment is therefore essential to understand a variety of pathophysiological phenomena.

Despite significant progresses over the past decade in characterizing biomolecules and signaling pathways related to cell adhesion and detachment, the biophysical and dynamic details of cell adhesion and detachment still remain elusive and challenging. Thus, diverse experimental methods which are able to characterize cell adhesion or cell detachment have been developed. The previous techniques for cell adhesion characterization can be classified into photolithography [3], e-beam lithography [4], dip-pen lithography [5], nanoimprint lithography [6], microcontact printing [7], elastomeric stencil [8], ink-jet printing [9], optical tweezer [10], electrophoresis [11], and switchable surface [12]. Although the conventional lithography techniques (i.e., photolithography, e-beam lithography, dip-pen lithography, and nanoimprint lithography), microcontact printing, elastomer stencil, and ink-jet printing technologies successfully characterized cell adhesion using specific proteins, they manipulated not cell adhesion but cell positioning. In terms of design flexibility, they altered cell adhesion only at the beginning of microfabrication and were not involved in cell detachment characterization. The optical tweezer and electrophoresis had a possibility of protein denaturation and cell electrolysis, respectively. The biocompatibility of the switchable surface was questionable. The existing techniques for cell detachment characterization can be categorized into hydrodynamic shear force assay [13, 14], centrifugal assay [15], and micropipette aspiration [16]. The hydrodynamic shear force assay was designed to characterize cell detachment with parallel, rotational, or radial flow. Although this method was able to quantify cell detachment with a wide range of shear force, it had several inherent drawbacks of cell rupture

during experiment and considerable local deformation of the cells. The centrifugal assay provided a relatively low centrifugal force for cell detachment, thus restricting its application to anchorage-dependent cells which were either cultured for short term (i.e., less than 1 h) or coupled with weak cell-to-substrate interactions. The micropipette aspiration also resulted in rapid cytoskeletal remodeling during experiment due to the direct contact between anchorage-dependent cell and micropipette. Recently, the electrochemical techniques have been developed to solve the problems of the previous approaches such as protein denaturation, cell electrolysis, local deformation, direct contact, etc. Together with microcontact printing technique, the reductive desorption of gold-alkanethiol SAM was used to quantify cell adhesion [17]. The success of this method was, however, restricted to only cell adhesion characterization at a cellular level and was involved in complicated preparation processes. The electrochemical reaction of several polyelectrolytes was also proposed to study cell adhesion as well as cell detachment [18]. The multiple deposition of the polyelectrolytes on a substrate and even on cells of interest disturbed experimental results due to unintended changes in the cells. The shortcomings of previous approaches, described above, motivate us to develop a promising biological platform for the quantitative analysis of cellular dynamics, similar to a breadboard in electronics.

A biological breadboard is developed for spatiotemporal manipulation of cell adhesion and detachment at cellular and subcellular levels. The BBB, inspired by an electrical breadboard, has the following features in characterizing cellular dynamics related to cell adhesion and detachment: addressability, multifunctionality, reusability, and ease of use. The BBB has addressability in manipulating cell adhesion and detachment due to its exquisite structure, that is, an array of independently operated gold microelectrodes microfabricated on a Pyrex glass substrate; the BBB is used to carry out a variety of biological characterizations (e.g., cell positioning, cell patterning, cell motility control, changes in material properties, etc.) related to cell adhesion and detachment; the electrochemistry involved in the BBB (reductive desorption of gold-thiol SAM) makes the BBB reusable; the BBB is easily incorporated with other conventional instruments such as optical, fluorescent, and confocal microscopes. Moreover, the BBB provides anchorage-dependent cells with microenvironment (RGD peptide) that is as similar as possible to in vivo microenvironment. This biological platform therefore makes it possible to characterize the adhesion and detachment of *living* and *intact* cells, thus meeting with solid and reliable results in cellular dynamics.

Structure and Working Principle

The BBB consists of an array of identical gold microelectrodes patterned on a Pyrex substrate, which secures a high degree of freedom in programmably manipulating cell adhesion and detachment, as shown in Fig. 1a. The gold microelectrodes and Pyrex substrate are modified with RTT and PEG, respectively. The RTT functionalization on the gold microelectrodes is to make cell-adhesive surfaces by tethering RGD peptide to the microelectrodes via thiol compound, following the spontaneous chemisorption

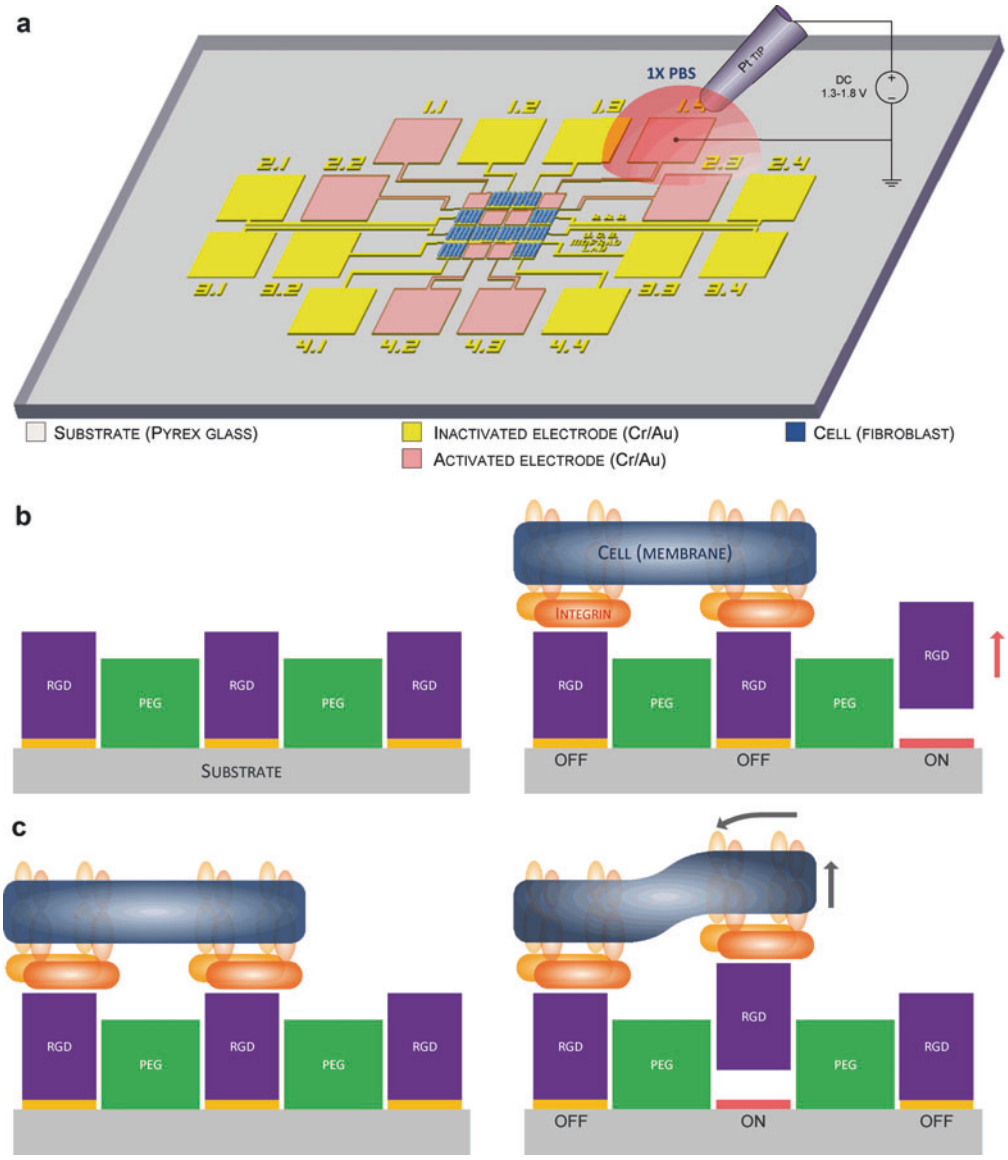
$$\mathrm{R-S-H+Au \rightarrow R-S-Au+1/2H_2} \quad (1)$$

where R is a substituent [19]. The RGD-ligand provides a binding site to integrins in cell adhesion process. The PEG treatment on the Pyrex substrate is performed to achieve a cell-resistive surface where the hydrated neutral PEG chains repulse cells sterically.

A spatiotemporal manipulation of cell adhesion is implemented by selectively detaching the RGD (and even an anchorage-dependent cell) from the gold microelectrodes with negative potential of -0.9 to -1.6 V, following the electrochemical reaction [20]

$$\mathrm{R-S-Au+H^+ + e^- \rightarrow R-S-H+Au.} \quad (2)$$

After the surface modifications, the RTT on a target area is detached by activating a



Biological Breadboard Platform for Studies of Cellular Dynamics, Fig. 1 A biological breadboard for spatiotemporal manipulation of cell adhesion and detachment at cellular and subcellular levels. (**a**) A schematic of the BBB consisting of a gold microelectrode array patterned on a Pyrex surface. The addressability, multifunctionality, reusability, and ease of use of the BBB are due to its exquisite structure and working principle. (**b**) Spatiotemporal manipulation of cell adhesion. Before cell adhesion (*left*), an RGD peptide is bound to all gold microelectrodes through RTT functionalization, whereas no cell adhesion site is formed on the Pyrex surface treated with PEG. On cell adhesion (*right*), the RGD peptide on a target microelectrode (third electrode from *left*) is detached from the target microelectrode with activation potential, followed by cell seeding. The loaded cells adhere only to inactivated microelectrodes (first and second electrodes from *left*) through RGD binding to integrin. (**c**) Spatiotemporal manipulation of cell detachment. Before cell detachment (*left*), cells adhere to the gold microelectrodes functionalized with RTT through RGD binding to integrin. On cell detachment (*right*), the cells or one part of the cell adhered to a target microelectrode (second electrode from *left*) is detached by activating the target electrode with activating potential which leads to the reductive desorption of gold-thiol SAM. The detached cells or one part of the cell retracts spontaneously

corresponding gold microelectrode with electrical potential, followed by sonication in cell culture media for 3 min and cell seeding. The loaded cells adhere only to the inactivated (still, RTT functionalized) gold microelectrodes (Fig. 1b), because the cells have substantially more affinity for cell adhesion to the RGD on the inactivated microelectrodes than to the gold of the activated microelectrodes and to the Pyrex glass substrate modified with PEG. The spatiotemporal manipulation of cell detachment is the same as that of cell adhesion except the sequential order of microelectrode activation and cell loading. For the spatiotemporal manipulation of cell detachment, the surfaces of the BBB are also modified with PEG and RTT, followed by cell seeding. The loaded cells graft and stretch to the RGD on gold microelectrodes. On cell detachment, cells or one part of the cell is detached (for subcellular detachment) from the BBB by activating a target microelectrode, and then the detached cell or cytoskeletal layer spontaneously retracts because chemical bonding between gold and thiol is broken by activation potential, as shown in Fig. 1c. When the detached cells or one part of the cell senses no external mechanical anchorage (focal adhesions), it begins to experience the liquefaction (i.e., gel–sol transition) of the cytoskeletal layer which accompanies cellular retraction by changing the length of actin filaments. This is how the BBB manipulates cell adhesion and detachment in a spatiotemporal way.

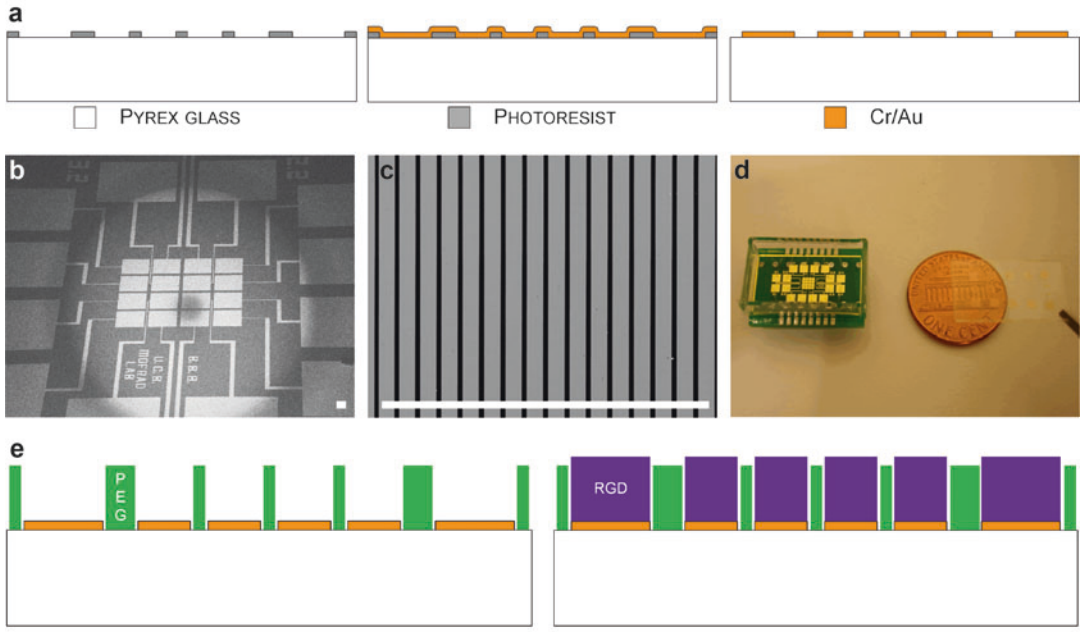
Materials and Methods

Microfabrication Process

The BBB was fabricated on a 4-in. Pyrex glass wafer with a thickness of 500 μm. After piranha cleaning with 1:1 (v:v) 96 % sulfuric acid (H_2SO_4) and 30 % hydrogen peroxide (H_2O_2) for 10 min, 1 μm-thick LOR resist (LOR 10A, MicroChem Corp.) was spin coated at 4000 rpm for 40 s, followed by soft baking at 170 °C for 5 min. A 2 μm-thick positive photoresist (S1818, Rohm and Haas Corp.) was spin coated on the LOR resist at 4000 rpm for 40 s for double-layer resist stack, followed by soft baking at 110 °C for 1 min. An optical lithography was done to micropattern the double-layer resist stack before e-beam evaporation process, as shown in Fig. 2a (left). The next was a deposition of 5 nm-thick chromium (Cr) adhesion layer and 100 nm-thick gold (Au) layer on the wafer, as shown in Fig. 2a (center). Because optical transparency in the visual spectrum is essential to make the BBB incorporate with other biological instruments (e.g., inverted optical and fluorescent microscopes), the Au layer thickness was reduced to 30 nm for experimental measurement [see Fig. 2d (right)]. Next, the Cr/Au-deposited wafer was immersed in an organic solvent mixture (BAKER PRS-3000 Stripper, Mallinckrodt Baker, Inc.) at 80 °C for 4 h to lift off the double-layer resist stack, thereby achieving two kinds of BBBs, as shown in Fig. 2a (right). One is for cell adhesion and detachment at a cellular level, each microelectrode of which is 500 μm in length and 500 μm in width, as shown in Fig. 2b. The other is for cell adhesion and detachment at a subcellular level, each gold microline of which is 10 μm in width and 3 μm in gap between two neighboring gold microlines, as shown in Fig. 2c. The microfabricated BBB was wire bonded in a chip carrier and then assembled with a cell-culture well made of polystyrene, as shown in Fig. 2d.

PEG Treatment on Pyrex Surface

Before PEG treatment, the microfabricated BBB was cleaned with an oxygen plasma chamber (PM-100 Plasma Treatment System, March Plasma Systems, Inc.) at 100 W for 30 s. The BBB was then incubated with 2 ml m-PEG silane (2 % v/v, Gelest, Inc.) and 1 ml hydrochloric acid (HCl, 1 % v/v, Fisher Scientific) dissolved in 97 ml anhydrous toluene (Fisher Scientific) for 2 h, as shown in Fig. 2e (left). This process was carried out in a nitrogen purge glove box to avoid atmospheric moisture. The incubated BBB was sequentially rinsed in fresh toluene and ethanol, dried with nitrogen, and cured at 120 °C for 2 h. The surface-modified BBB was stored in a vacuum desiccator until the next surface modification, RTT functionalization on gold surface.



Biological Breadboard Platform for Studies of Cellular Dynamics, Fig. 2 Microfabrication and surface modification of the BBB. (**a**) The BBB is fabricated by patterning an array of gold microelectrodes on a Pyrex substrate through patterning of photoresist by lithography (*left*), deposition of Cr and Au layers by e-beam evaporation (*center*), and patterning of the deposited Cr/Au layer by lift-off (*right*). (**b**) BBB for cell adhesion and detachment manipulation at a cellular level, each microelectrode of which is 500 μm in length and 500 μm in width. (**c**) BBB for cell adhesion and detachment manipulation at a subcellular level, each microelectrode of which is 10 μm in width and 3 μm in gap. (**d**) Photograph of the BBB before (*right*) and after (*left*) assembly, showing its transparency in the visual spectrum. (**e**) Surface modification process. The BBB is incubated with PEG solution to make the Pyrex surface cell repulsive (*left*), and then it is incubated with synthesized RTT solution to make gold microelectrodes cell adhesive (*right*). Scale bars of (**b**) and (**c**) are 100 μm

RTT Functionalization on Gold Surface

The gold microelectrodes of the BBB were functionalized with RTT whose solution was synthesized by chemically combining *cyclo* (Arg-Gly-Asp-D-Phe-Lys) ($C_{27}H_{41}N_9O_7$, Peptides International, Inc.) with dithiobis (succinimidyl undecanoate) ($C_{30}H_{48}N_2O_8S_2$, Dojindo Molecular Technologies, Inc.) as follows. The 3.02 mg *cyclo* (Arg-Gly-Asp-D-Phe-Lys) was dissolved in 5 ml dimethoxysulfoxide (DMSO, Sigma-Aldrich) to get 1 mM aliquot and stored at −20 °C. This reaction was made in a nitrogen purge glove box to protect the RGD peptide from exposure to atmospheric moisture. The 3.14 mg dithiobis(succinimidyl undecanoate) was also dissolved in 5 ml DMSO and then stored at −20 °C. This preparation was also done in moisture-free environment. Before a gold surface functionalization, both aliquots were warmed to room temperature in a desiccator. The 5 ml RGD peptide aliquot was mixed with 50 μl triethylamine (1 % v/v, Fisher Scientific) for 5 min to make all primary amines of a lysine amino acid unprotonated. The 5 ml dithiobis (succinimidyl undecanoate) was added to the 5 ml RGD peptide aliquot and then mixed well using a vortex mixer for 4 h to synthesize RTT solution. For the gold surface functionalization, the PEG-treated BBB was incubated with the solution for 1 h at room temperature to promote a spontaneous chemisorption between thiol and gold, followed by sonification in DMSO for 3 min, and rinsed in ethanol and phosphate-buffered saline (PBS, Sigma-Aldrich) to eliminate

all unbound RTTs from the gold surface, as shown in Fig. 2e (right). The thiol exposed to gold achieved a SAM which tethered an RGD peptide to the gold surface.

Contact Angle Measurement

The contact angles of a PEG-treated Pyrex surface and an RTT-functionalized gold surface were measured with a contact angle measurement system (KRÜSS582, KRÜSS), a goniometer. A sessile drop mode was used to estimate the wetting properties of the above two solid surfaces. The contact angle was averaged from 10 measurements. The contact angle of PEG-treated Pyrex surface was compared to that of pure Pyrex surface, and the contact angle of RTT-functionalized gold surface was compared to those of bare gold surface and thiol-treated gold surface.

X-ray Photoelectron Spectroscopy (XPS) Sample Preparation and Characterization

An XPS survey scan was used to confirm the existence of RGD peptide linked to a gold surface via thiol after RTT functionalization. An XPS sample was prepared on a 4-in. silicon wafer e-beam evaporated with 5 nm Cr adhesion layer and 50 nm Au layer. This wafer was immersed for 2 h in the prepared RTT solution for the RTT functionalization on a gold surface. A bare gold sample without RTT functionalization was run as a control experiment. The XPS analysis was carried out with a customized ESCA (Omicron NanoTechnology) at 1×10^{-8} Torr, and all measured spectra were referenced to the position of the Au 4*f* peaks. The scan was collected over a range of 20 eV around the peak of interest with pass energy of 23.5 eV.

Potentiodynamic Electrochemical Characterization of Reductive Desorption of Gold-Thiol SAM

A silicon wafer e-beam evaporated with 5 nm-thick Cr and 50 nm-thick Au was functionalized with RTT to prepare a cyclic voltammetry (CV) sample. This RTT-functionalized gold microelectrode was used as a working electrode while platinum and Ag/AgCl electrodes were used as counter and reference electrodes, respectively. A voltage supplied by a DC power source (B&K Precision Corporation, Yorba Linda, CA) was applied between the gold-thiol SAM (or Ag/AgCl electrode) and the platinum electrode. The CV was carried out in a Dubecco's phosphate-buffered saline solution [DPBS (pH 7.4), Sigma-Aldrich, St. Louis, MO] with an EG&G potentiostat model 362 (AMETEK Princeton Applied Research, Oak Ridge, TN). A scan started cathodically from 0.0 to −2.0 V, then annodically back to 0.0 V at a scan rate of 50 mV/s.

Cell Culture

NIH 3 T3 mouse embryonic fibroblast cell (NIH 3 T3 fibroblast) was cultured in Dulbecco's modified eagle medium (DMEM, GIBCO™) supplemented with 10 % fetal bovine serum (FBS, GIBCO™) and 1 % Penicillin-Streptomycin (GIBCO™) at 37 °C in humidified 5 % CO_2 atmosphere. The cell was passaged every 4 days as follows. The cell was washed 1 time in PBS and 1X trypsinized with Trypsin-EDTA solution 0.5 %. After centrifuging the cell, it was inoculated into a new Petri dish. The NIH 3 T3 fibroblasts with a passage number of 5–20 were used in the experiments. Before experiment, the surface-modified BBB was sterilized with 70 % ethanol, washed twice with 1X PBS, and placed into a Petri dish containing 5 ml cell culture medium with a cell suspension of about 1×10^6 cells/ml. For the subcellular detachment experiments, the cell concentration was changed into 1×10^6 cells/ml. After 1 h, unattached NIH 3 T3 fibroblasts were removed by additional wash in PBS, followed by cell culture medium replacement. All experiments were carried out 24 h after cell seeding in a self-designed chamber with humidified 5 % CO_2 atmosphere and at 37 °C.

AFM Indentation

The elastic strength of the motile cytoskeletal layer (retracting after subcellular detachment) of NIH 3 T3 fibroblasts was measured with an Autoprobe CP AFM (Park Science Instruments). All measurements were made at a low indentation speed of 10 nm/s to minimize a viscous damping

effect by a dashpot of the motile cytoskeletal layer in quantifying the elastic strength with AFM. The elastic modulus of the motile cytoskeletal layer was determined from the measurement on the deflection of an AFM microcantilever (HYDRA2R-100 N, Nanoscience Instruments, Inc.) at each measurement point of interest, located on the cytoskeletal layer. Although a nominal spring constant of the AFM microcantilever was 0.011 N/m, the spring constant of the AFM microcantilever was measured as 0.016 ± 0.005 N/m, which was used in calculating the elastic modulus of the motile cytoskeletal layer with AFM.

Key Research Findings

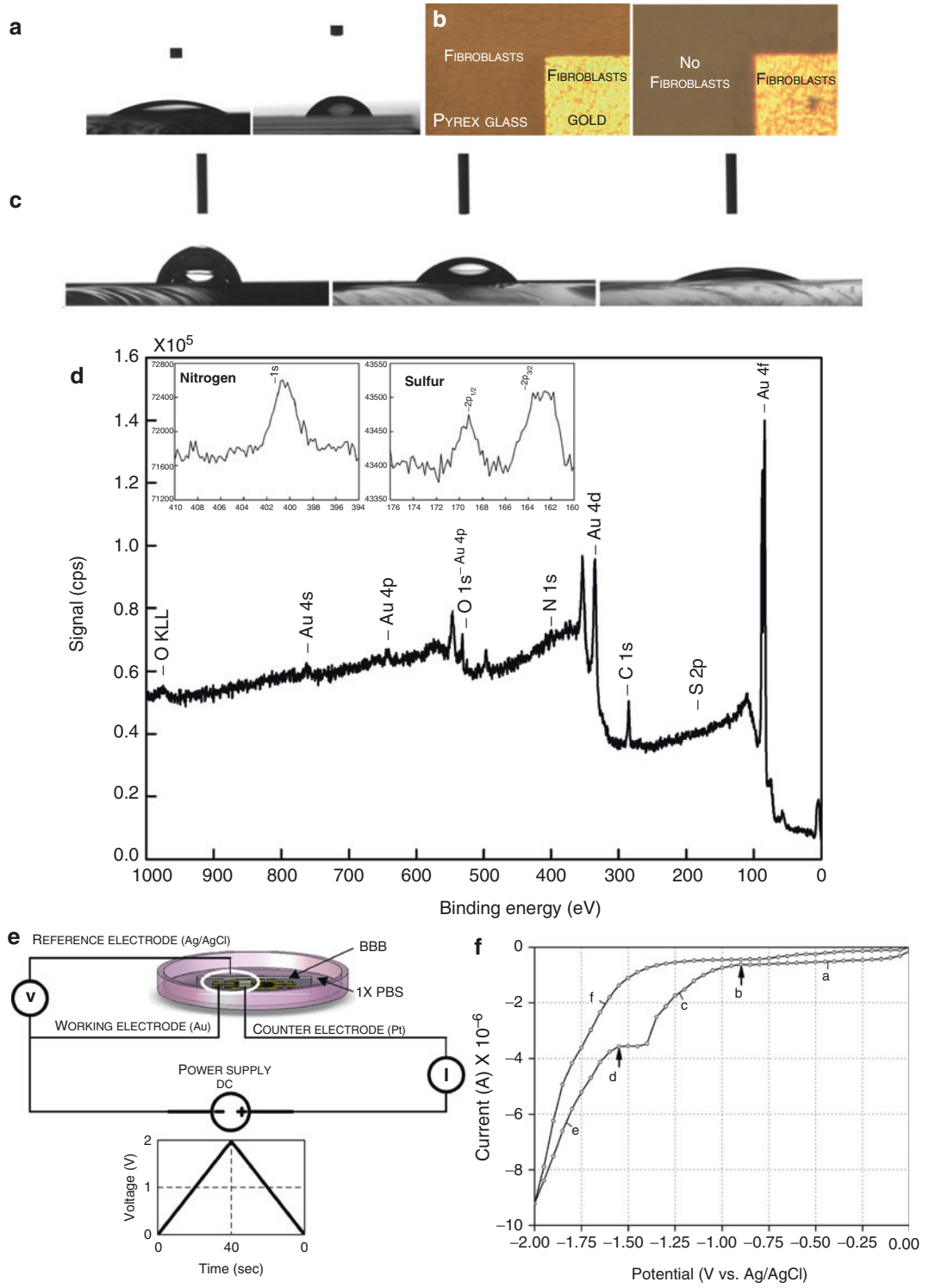
Verification of Surface Modifications

The surface modifications were verified by two methods – contact angle measurement and XPS survey. The contact angle of the PEG-treated Pyrex surface measured by a goniometer was 61.5 ± 3.8° (averaged from 10 measurements), whereas that of the untreated Pyrex surface was 25.7 ± 1.5°, as shown in Fig. 3a. This shows that a PEG-treated surface has strong hydrophobicity and consequently prevents cell adhesion and protein fouling. An NIH 3 T3 fibroblast load test was also made, as shown in Fig. 3b. The images obtained 24 h after cell seeding show the Pyrex surface is modified into cell-resistive through PEG treatment. To characterize the RTT functionalization, a contact angle was measured for bare gold, thiol-treated gold, and RTT-functionalized gold, each of which was 67.3 ± 2.5°, 53.3 ± 1.3°, and 24.6 ± 2.8°, respectively, as shown in Fig. 3c. This confirms the RTT functionalization is made as designed. The experimental results verify that our surface modifications are effective to achieve cell-repulsive and cell-adhesive surfaces in the BBB.

After RTT functionalization on gold surface, the existence of RGD linked to gold microelectrodes via thiol was also demonstrated by an XPS survey scan. A measured XPS survey spectrum of RGD/thiol/Au interface, as shown in Fig. 3d. The peaks of Au 4 *s*, Au 4*p*, Au 4*d*, and Au 4*f* indicate the presence of e-beam evaporated gold (Au(111)); the peaks of S $2p_{1/2}$ and S $2p_{3/2}$ (right inset) show that sulfur from thiol is in existence on the surface; the peaks of C 1 *s*, O 1 *s*, O *KLL*, and N 1 *s* (left inset) demonstrate that there are carbon, oxygen, and nitrogen from amine functional group ($–NH_2$) and carboxylic acid functional group (–COOH) of an RGD peptide. For reference, hydrogen was not detected due to XPS working principle. This XPS survey spectrum verifies that the RTT functionalization on the gold microelectrodes is well made as designed.

Reductive Desorption of Gold-Thiol SAM

The rapid desorption of a gold-thiol SAM under activation potential was also characterized with CV using a three-electrode system where the gold microelectrode (of the BBB), a platinum electrode, and an Ag/AgCl electrode work as a working electrode, a counter electrode, and a reference electrode, respectively, as shown in Fig. 3e. The measured CV shows the current measured at the working electrode as a function of applied voltage with respect to the Ag/AgCl electrode, as shown in Fig. 3f. At a potential range of 0.00 to −0.90 V (section a), the measured current was negligible, which means there is no reductive desorption of the SAM and consequently the SAM impedes electron transfer across an electrolyte-electrode interface. When the applied voltage increased (section c), the reductive desorption of the SAM started and finished at a potential of −0.90 V (point b) and −1.55 V (point d) respectively, breaking the chemical binding between thiol and gold. The reductive desorption of the SAM got maximized at a peak of −1.40 V. After finishing the rapid desorption of the gold-thiol SAM (sections e and f), the gold microelectrode worked as a resistor. This electrochemical result shows the activation potential needs to be larger than −0.90 V. This CV brings forward an indisputable evidence of the reductive desorption of the gold-thiol SAM and also demonstrates the negative bias potential is required to detach the RGD peptide from the gold microelectrode.



Biological Breadboard Platform for Studies of Cellular Dynamics, Fig. 3 Characterization of two surface modifications and potentiodynamic electrochemical characterization of the reductive desorption of gold-thiol SAM. (**a**) Contact angle of the Pyrex surface before (*left*) and after (*right*) PEG treatment. The contact angle is

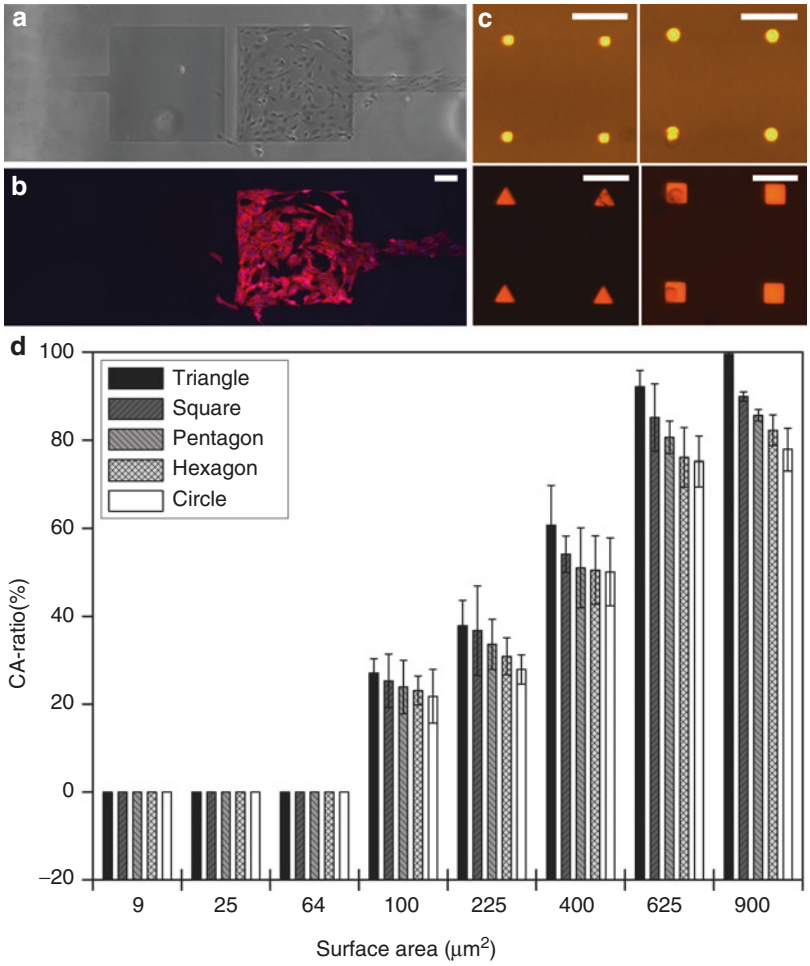
Cell Adhesion Manipulation

The adhesion of an anchorage-dependent cell (i.e., NIH 3 T3 fibroblast) was spatiotemporally manipulated with the surface-modified BBB composed of two-by-one gold microelectrodes where a left microelectrode was activated with activation potential of −1.20 V but a right one was inactivated. The BBB was then sonificated in a cell culture media and NIH 3 T3 fibroblasts loaded. An anchorage-dependent cell in general has much higher affinity for RGD peptide than gold and Pyrex glass when a cell makes adhesion to a substrate. The loaded cell therefore adheres only to the right electrode, as shown in optical and immunofluorescent images (Fig. 4a, b). The result demonstrates the BBB's addressability in cell adhesion manipulation. Next, the dependence of cell adhesion at a single cell level on the size and geometric shape of a gold microelectrode was studied to determine the geometry of the gold microelectrode for subcellular detachment. Forty types of gold microelectrode arrays were prepared, as shown in Fig. 4c. Each array was designed to have 25 identical gold surfaces whose size was 9 μm^2, 25 μm^2, 64 μm^2, 100 μm^2, 225 μm^2, 400 μm^2, 625 μm^2, or 900 μm^2 and their shape was an n-sided regular polygon ($n = 3,\ 4,\ 5,\ 6,\ \infty$ (circle)). NIH 3 T3 fibroblasts were loaded into the surface-modified assays at a cell suspension concentration of 1×10^6 cells/ml. As an index for quantifying cell adhesion at a single cell level, a CA-ratio which is the ratio of the number of gold microelectrodes with cell adhesion to the total number of gold microelectrodes was measured as a function of the size and geometric shape of a gold surface, as shown in Fig. 4d. The measured CA-ratio provides the following facts. First, the CA-ratio in a single cell level is proportional to the size of a gold microelectrode, and the minimum size of a gold microelectrode for single cell adhesion is the diameter of a cell in a floating state (10 μm for NIH 3 T3 fibroblast). Secondly, a cell wants to make its adhesion on the circumferential zone of a gold microelectrode rather than the central zone.

Cell Detachment Manipulation at a Cellular Level

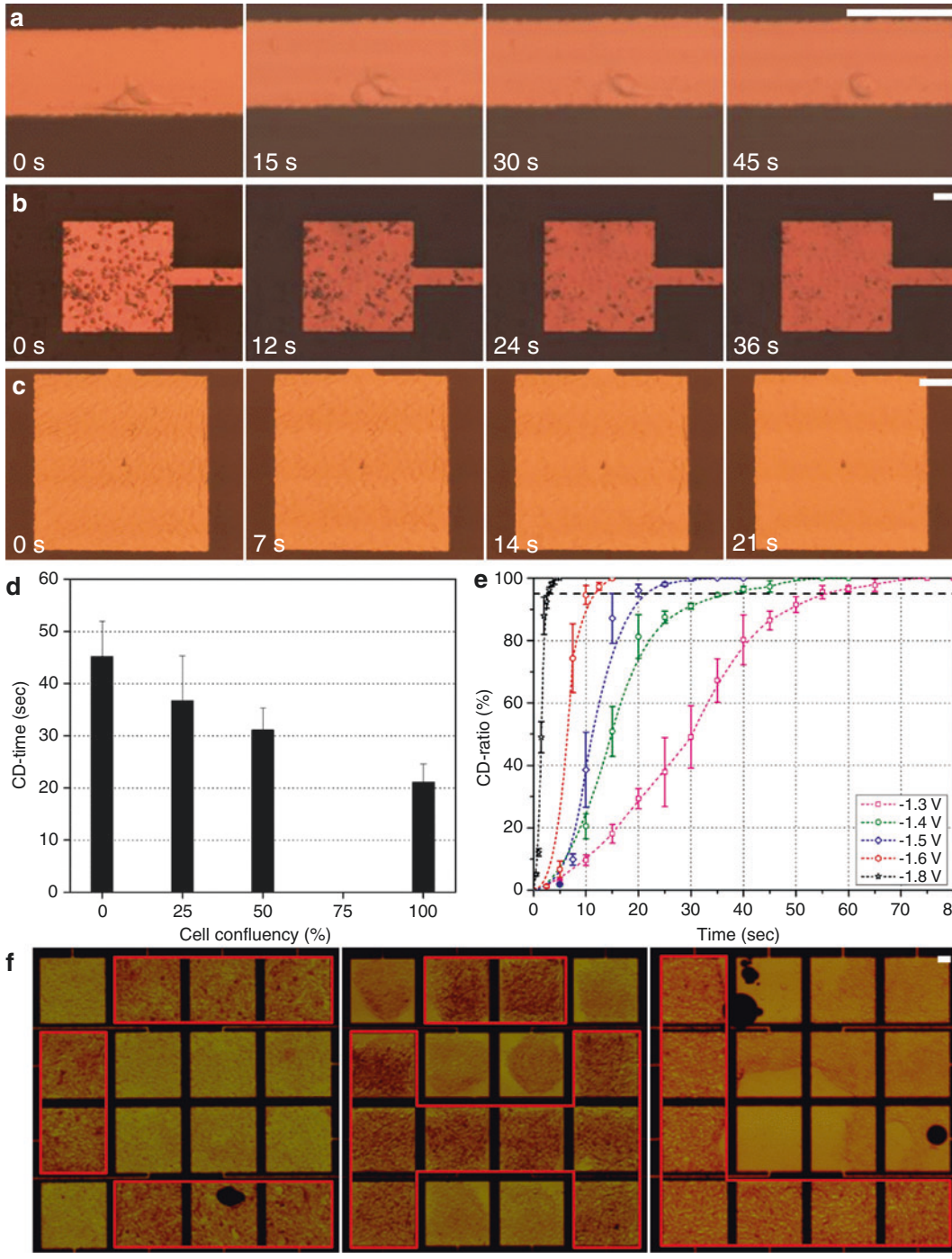
The cell detachment dynamics at a cellular level was characterized by disconnecting an entire living cell from substrate with the BBB. The cell detachment experiments were carried out for three cases: two fibroblasts (see Fig. 5a), sparse fibroblasts with about 25 % confluence (see Fig. 5b), and confluent fibroblasts with 100 % confluence (see Fig. 5c). The experiments at cellular and subcellular levels were performed at the same experimental conditions and apparatus as those of cell adhesion experiment. As an evaluation index for cell detachment at a cellular level, a time required to detach 95 % of the adhered fibroblasts (CD-time) was measured. The three cases had the CD-times of 45.2 ± 6.8 s, 36.7 ± 8.7 s, and 21.1 ± 3.5 s (averaged from at least 10 measurements) respectively at an activation potential of −1.50 V, showing more confluent cells have shorter CD-time, as shown in Fig. 5d. This is because all cells in confluent culture are connected to each other through cell-to-cell interaction. The detachment of one cell allows the neighboring cells to be easily detached by providing an external vertical force to them. The detached cell seems to

Biological Breadboard Platform for Studies of Cellular Dynamics, Fig. 3 (continued) changed from 25.7 ± 1.5 to 61.5 ± 3.8° by PEG treatment on the Pyrex surface. (**b**) Cell adhesion on the Pyrex surface before (*left*) and after (*right*) PEG treatment, showing the Pyrex surface is modified to be cell repulsive through PEG treatment. (**c**) Contact angles of bare gold (67.3 ± 2.5°, *left*), thiol-treated one (53.3 ± 1.3°, *middle*), and RTT-functionalized one (24.6 ± 2.8°, right), confirming the surface modification by RTT. (**d**) XPS survey spectrum of RTT-functionalized gold surface. Detected are the gold peak from gold element, sulfur peak from thiol, nitrogen peak from amine group of RGD peptide, and carbon and oxygen peaks from carboxylic acid group of RGD peptide. (**e**) Experimental setup for CV measurement where a gold microelectrode, a platinum electrode, and an Ag/AgCl electrode work as working, counter, and reference electrodes, respectively. (**f**) Measured CV showing that the reductive desorption of gold-thiol SAM starts and finishes at −0.90 V and −1.55 V respectively, and gets maximized at −1.40 V



Biological Breadboard Platform for Studies of Cellular Dynamics, Fig. 4 Spatiotemporal manipulation of cell adhesion using the BBB. (**a**) Optical image. The spatiotemporal manipulation of cell adhesion is demonstrated with a two-by-one BBB where a left electrode is activated but a right one is inactivated. (**b**) Immunofluorescent image. NIH 3 T3 fibroblasts are stained for actin with rhodamine phalloidin (*red*) and for cell nucleus with DAPI (*blue*). (**c**) Cell adhesion as a function of the shape and size of gold microelectrode. No cell adhesion is made on 64 μm^2-sized square gold microelectrodes (first from *left*) with the cell suspension of 1×10^6 cells/ml; cell adhesion is made on 25 % of 100 μm^2-sized regular hexagonal gold microelectrodes (second); cell adhesion is made on 25 % of 225 μm^2-sized equilateral triangular gold microelectrodes (third); cell adhesion is made on 50 % of 400 μm^2-sized square gold microelectrodes (fourth). (**d**) CA-ratio as a function of the size and geometric shape of gold microelectrodes. Scale bar of (**b**) is 100 μm, and those of (**c**) are 50 μm

continue to communicate with the neighboring cells through cell-to-cell interaction. The effect of activation potential on cell detachment was also investigated by counting the ratio of detached cells to total cells (i.e., CD-ratio) as a function of activation time and potential, as shown in Fig. 5e. The CD-ratio gets decreased as activation potential gets increased due to the reductive desorption of gold-thiol SAM which gets faster as negative potential gets increased (see Fig. 3f). The CD-ratio monotonically increases with two inflection points, s-shape curve, which clearly demonstrates there is a large deviation in integrin binding to extracellular matrix or other cells related to cell-to-cell interaction. A programmable cell patterning of NIH 3 T3 fibroblasts was also made with a single four-by-four to demonstrate the BBB's addressability and reusability in cell detachment manipulation at a cellular



Biological Breadboard Platform for Studies of Cellular Dynamics, Fig. 5 Spatiotemporal manipulation of cell detachment at a cellular level using the BBB. (**a**) Optical sequential images showing the spatiotemporal manipulation of cell detachment for two-cell case. Two cells are detached from a gold microelectrode activated with negative potential of −1.20 V. The average CD-time for two cells is 45.2 ± 6.8 s. (**b**) Cell detachment

level, as shown in Fig. 5f. The cells were sequentially patterned in the shape of "C," "A," and "L" by using a single device. The BBB was reused through simplified organic cleaning process after each usage, experimentally verifying the addressability, multifunctionality, and reusability of the BBB at the same time. The BBB is reusable due to its reductive desorption of gold-thiol SAM which yields original gold again.

Cell Detachment Manipulation at a Subcellular Level

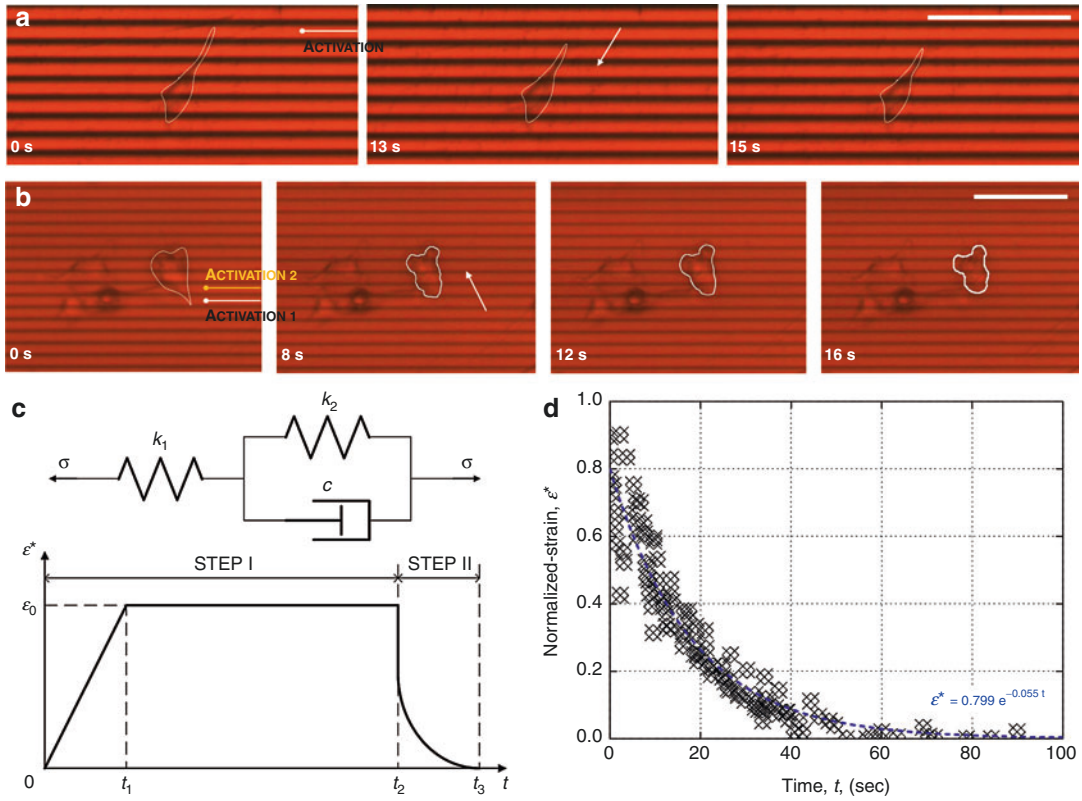
The cell detachment at a subcellular level was also characterized by disconnecting one part of an NIH 3 T3 fibroblast from a substrate using the BBB which consists of gold microelectrode lines with 10 μm width and 3 μm gap (see Fig. 2d). As shown in the optical sequential images (see Fig. 6a), the subcellular detachment was accomplished by retracting the detached cytoskeleton for 16 s, which is faster retraction than cell detachment at a cellular level. This is because a single cell which has no constraint in stretching itself with focal adhesions is always under higher strain than confluent cells. In the next subcellular detachment experiment, one part of NIH 3 T3 fibroblast was sequentially detached with a series of activations where the first activation (i.e., activation 1) was followed by the second one (i.e., activation 2) after 16 s. The optical sequential images reveal that repetitive activations within dozens of seconds do not have any influence on cell viability and a subcellular detachment amount is gradually increasing, as shown in Fig. 6b. These subcellular detachment experiments open a possibility for spatiotemporal manipulation of cell motility by disconnecting a part of the cell from a substrate. For example, when an anchorage-dependent cell is exposed to subcellular detachment at 3 o'clock direction, it will migrate into 9 o'clock direction.

In our biological breadboard platform, the spatiotemporal manipulation of cell adhesion and detachment is made by deliberately detaching an RGD peptide from a gold microelectrode, leading to adhere and detach an anchorage-dependent cell at cellular and subcellular levels. It is therefore necessary to check whether a detached cell is still alive rather than electrocuted. If a detached cell is electrocuted, this technology cannot be used for spatiotemporal manipulation of cell motility by disconnecting one part of the cell from a substrate and furthermore for cell adhesion and detachment manipulation for a better understanding of cellular dynamics. The viability of the detached cell was therefore characterized. In this experiment, NIH 3 T3 fibroblasts bound to the BBBs were detached by activation potential of −1.50 V. The detached cells collected with a centrifuge, and then it was loaded into a plastic Petri dish. Their viability was observed after 24 h of cell loading. This experiment shows 80 % of the detached cells are still alive. Compared to a general cell passage at which about 95 % of cells are viable, this result is comparably excellent and lays a technical foundation for spatiotemporal manipulation of cell adhesion and detachment using the BBB.

Changes in Viscoelastic Properties During Subcellular Detachment

The changes in the viscoelastic properties of a NIH 3 T3 fibroblast during its subcellular detachment were also quantified by detaching one part of the cell using the BBB. First of all, the detached cell was assumed as isotropic and viscoelastic. The retraction motion of the

Biological Breadboard Platform for Studies of Cellular Dynamics, Fig. 5 (continued) manipulation of 25 % confluent cells whose average CD-time is 36.7 ± 8.7 s. (**c**) Cell detachment manipulation of 100 % confluent cells whose average CD-time is 21.1 ± 3.5 s. (**d**) Measured CD-time as a function of cell confluency, showing CD-time is inversely proportional to cell confluency. (**e**) Measured CD-ratio as a function of activation time and potential for 100 % confluent cells. The CD-ratio is monotonically increasing with two inflection points (s-shaped curve), indicating there is a large deviation in integrin binding to ECM or other cells related to cell-to-cell interaction. (**f**) Programmable cell patterning (i.e., cell detachment manipulation) using a single four-by-four BBB. The cells are patterned into "C," "A," and "L" shapes, verifying the addressability and reusability of the BBBs. Scale bars are 100 μm



Biological Breadboard Platform for Studies of Cellular Dynamics, Fig. 6 Spatiotemporal manipulation of cell detachment at a subcellular level using the BBB and its applications to cellular dynamics characterizations. (**a**) Subcellular detachment using a single activation. One part of a cell is detached, and in turn the detached cytoskeleton starts to retract. (**b**) Subcellular detachment using a series of activations. One part of the cell is sequentially detached from the *gold lines*. (**c**) Continuum model to describe the retraction of a detached cytoskeleton where the cell is assumed as a homogeneous standard linear viscoelastic solid (*top*). Strain profile of an anchorage-dependent cell during cell adhesion and detachment (*bottom*). When a cell adheres to a substrate ($t = 0$), the cell extends its protrusion and adheres again ($0 < t \leq t_1$); the nucleus of the cell translocates ($t_1 < t \leq t_2$); one part of the cell is detached and retracts ($t_2 < t \leq t_3$). (d) Normalized strain as a function of time obtained from single cells which are detached at a subcellular level. *Arrows* of (**a**) and (**b**) indicate the retraction direction of detached cytoskeleton. Scale bars of (**a**) and (**b**) are 100 μm

detached cytoskeleton of a cell was described with a standard linear viscoelastic solid model composed of two springs, k_1 and k_2, and one dashpot, c [see Fig. 6c (top)]. From this model, a normalized strain, ε^* at retraction step [STEP II, Fig. 6c (bottom)] was mathematically expressed as

$$\varepsilon^* = \frac{\varepsilon(t - t_2)}{\varepsilon_0} = \left(1 - \frac{k_2}{k_1 + k_2}\right) e^{-\frac{k_2}{c}(t - t_2)}, \quad (3)$$

where ε_0 is the initial strain of the cell at its protrusion step (STEP I, $0 < t \leq t_1$).

The changes in the viscoelastic properties of the detached (and then retracting) cytoskeleton were measured by combining Eq. 3 with two experimental results: time-sequential images of the retraction motion of the detached cytoskeleton (obtained from 20 cells detached at a subcellular level) and AFM indentation results on the detached cytoskeleton. From the time-sequential images of subcellular detachment, the normalized strain of the detached cytoskeleton was described as $\varepsilon* = 0.799\, e^{-0.055\,t}$ (see Fig. 6d). Based on the AFM indentation results on 10 samples,

the total elastic modulus of the detached cytoskeleton ($k_{total} = k_1 k_2/(k_1 + k_2)$) was determined as 1320 ± 310 Pa. These experimental results with Eq. 3 determined the viscoelastic properties of the detached cytoskeleton ($k_1 = 6567$ Pa, $k_2 = 1652$ Pa, and $c = 30037$ Pa·sec). Compared to the previous results ($k_{total} > 4000$ Pa and $c < 100$ Pa·sec) obtained from the adhered (not detached) cytoskeleton of fibroblasts, the detached cytoskeleton showed a 3 times decrease in its elastic modulus and a 300 times increase in its damping coefficient. This measurement suggests that a detached cytoskeleton becomes softer and consequently has a remarkable increase in its damping coefficient after a few seconds of subcellular detachment. This phenomenon is likely owing to the gel–sol transition of actin filaments at cell detachment which changes the viscoelastic properties of the detached cytoskeleton. The subcellular detachment depolymerizes the cross-linked network of actin filaments. Thus, the structural strength of the detached cytoskeleton decreases, but its viscous damping capacity increases. This result shows that the BBB can be further exploited for studies of cellular rheology and for quantification of viscoelastic properties of the cytoskeleton to supplement computational modeling efforts.

Future Directions

Although we have developed the biological breadboard for spatiotemporal manipulation of cell adhesion and detachment at cellular and subcellular levels, the platforms are still at their earliest stage, i.e., there is no commercially available one. Four exquisite features of the BBB over the previous cell manipulation methods, however, make the BBB as a promising platform for studies of cellular dynamics: (i) addressability in the spatiotemporal manipulation of cell adhesion and cell detachment; (ii) multifunctionality in the characterization of cell biomechanics such as cell positioning, cell patterning, cell motility control, etc.; (iii) reusability due to the electrochemical reaction of a gold-thiol SAM; and (iv) no need for special accessories in incorporating the BBBs with other conventional instruments. Extrapolation of this technology to other anchorage-dependent cells might help us to investigate critical cellular function and behavior, thereby leading to a better understanding of cellular dynamics. Ongoing work is focusing on more in-depth control of cell motility by developing a large-scale assay to shed light on the dynamics of cell motility. Combined with molecular dynamics models, the proposed devices for programmable subcellular adhesion and detachment will offer a new platform for studies of molecular biomechanics of the cell, especially mechanotransduction at focal adhesions.

Cross-References

- ▶ Biological Breadboard Platform for Studies of Cellular Dynamics
- ▶ Cell Adhesion

References

1. Geiger, B., Bershadsky, A., Pankov, R., Yamada, K. M.: Transmembrane crosstalk between the extracellular matrix and the cytoskeleton crosstalk. Nat. Rev. Mol. Cell Biol. **2**, 793–805 (2001)
2. Galbraith, C.G., Yamada, K.M., Sheetz, M.P.: The relationship between force and focal complex development. J. Cell Biol. **159**, 695–705 (2002)
3. Clark, P., Connolly, P., Curtis, A.S., Dow, J.A., Wilkinson, C.D.: Topographical control of cell behavior. I. Simple step cues. Development **99**, 439–448 (1987)
4. Lussi, J.W., Tang, C., Kuenzi, P.-A., Staufer, U., Csucs, G., Vörös, J., Danuser, G., Hubbell, J.A., Textor, M.: Selective molecular assembly patterning at the nanoscale: a novel platform for producing protein patterns by electron-beam lithography on SiO_2/indium tin oxide-coated glass substrates. Nanotechnology **16**, 1781–1786 (2005)
5. Lee, K.-B., Park, S.J., Mirkin, C.A., Smith, J.C., Mrksich, M.: Protein nanoarrays generated by dip-pen nanolithography. Science **295**, 1702–1705 (2002)
6. Hoff, J.D., Cheng, L.-J., Meyhöfer, E., Guo, L.J., Hunt, A.J.: Nanoscale protein patterning by imprint lithography. Nano Lett. **4**, 853–857 (2004)
7. Chen, C.S., Mrksich, M., Huang, S., Whitesides, G. M., Ingber, D.E.: Geometric control of cell life and death. Science **276**, 1425–1428 (1997)
8. Folch, A., Jo, B.H., Hurtado, O., Beebe, D.J., Toner, M.: Microfabricated elastomeric stencils for

micropatterning cell cultures. J. Biomed. Mater. Res. A **52**, 346–353 (2000)
9. Roth, E.A., Xu, T., Das, M., Gregory, C., Hickman, J. J., Boland, T.: Inkjet printing for high-throughput cell patterning. Biomaterials **25**, 3707–3715 (2004)
10. Birkbeck, A.L., Flynn, R.A., Ozkan, M., Song, D., Gross, M., Esener, S.C.: VCSEL arrays as micromanipulators in chip-based biosystems. Biomed. Microdevices **5**, 47–54 (2003)
11. Rosenthal, A., Voldman, J.: Dielectrophoretic traps for single-particle patterning. Biophys. J. **88**, 2193–2205 (2005)
12. Lahann, J., Mitragotri, S., Tran, T.-N., Kaido, H., Sundaram, J., Choi, I.S., Hoffer, S., Somorjai, G.A., Langer, R.: A reversibly switching surface. Science **299**, 371–374 (2003)
13. van Kooten, T.G., Schakenraad, J.M., van der Mei, H. C., Dekker, A., Kirkpatrick, C.J., Busscher, H.J.: Fluid shear induced endothelial cell detachment from glass-influence of adhesion time and shear stress. Med. Eng. Phys. **16**, 506–512 (1994)
14. Kuo, S.C., Hammer, D.A., Lauffenburger, D.A.: Simulation of detachment of specifically bound particles from surfaces by shear flow. Biophys. J. **73**, 517–531 (1997)
15. Lotz, M.M., Burdsal, C.A., Erickson, H.P., McClay, D.R.: Cell adhesion to fibronectin and tenascin: quantitative measurements of initial binding and subsequent strengthening response. J. Cell Biol. **109**, 1795–1805 (1989)
16. Shao, J.-Y., Hochmuth, R.M.: Micropipette suction for measuring piconewton forces of adhesion and tether formation from neutrophil membranes. Biophys. J. **71**, 2892–2901 (1996)
17. Inaba, R., Khademhosseini, A., Suzuki, H., Fukuda, J.: Electrochemical desorption of self-assembled monolayers for engineering cellular tissues. Biomaterials **30**, 3573–3579 (2009)
18. Guillaume-Gentil, O., Gabi, M., Zenobi-Wong, M., Vörös, J.: Electrochemically switchable platform for the micro-patterning and release of heterotypic cell sheets. Biomed. Microdevices **13**, 221–230 (2011)
19. Karp, G.: Cell and Molecular Biology: Concepts and Experiments. Wiley, New York (2005)
20. Dalton, B.A., Walboomers, X.F., Dziegielewski, M., Evans, M.D., Taylor, S., Jansen, J.A., Steele, J.G.: Modulation of epithelial tissue and cell migration by microgrooves. J. Biomed. Mater. Res. A **56**, 195–207 (2001)

Biological Nano-crystallization

▶ Macromolecular Crystallization Using Nanovolumes

Biological Photonic Structures

▶ Structural Color in Animals

Biological Platform

▶ Biological Breadboard Platform for Studies of Cellular Dynamics

Biological Sensors

▶ Arthropod Strain Sensors

Biological Structural Color

▶ Structural Color in Animals

Biologicalcell Manipulation

▶ Nanomanipulation of Biocells

Biomedical Applications

▶ Magnetic Nanoparticles for Biomedical Applications

BioMEMS/NEMS

▶ Nanotechnology

Biomimetic

▶ Moth-Eye Antireflective Structures

Biomimetic Energy Conversion Devices

▶ Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components

Biomimetic Flow Sensors

Jérôme Casas[1], Chang Liu[2] and Gijs Krijnen[3]
[1]Institut de Recherche en Biologie de l'Insecte, IRBI UMR CNRS 6035, Université of Tours, Tours, France
[2]Tech Institute, ME/EECS, Northwestern University, Evanston, IL, USA
[3]Transducers Science and Technology group, MESA + Research Institute for Nanotechnology, University of Twente, Enschede, The Netherlands

Definition

Biomimetic flow sensors are biologically inspired devices that measure the speed and direction of fluids.

Introduction

This survey starts by describing the role and functioning of airflow-sensing hairs in arthropods and in fishes, carries on with the biomimetic MEMS implementations, both for air and water flow sensing, and ends up with some perspectives for bio-inspired micro-technologies based on the latest understanding of the biological sensors.

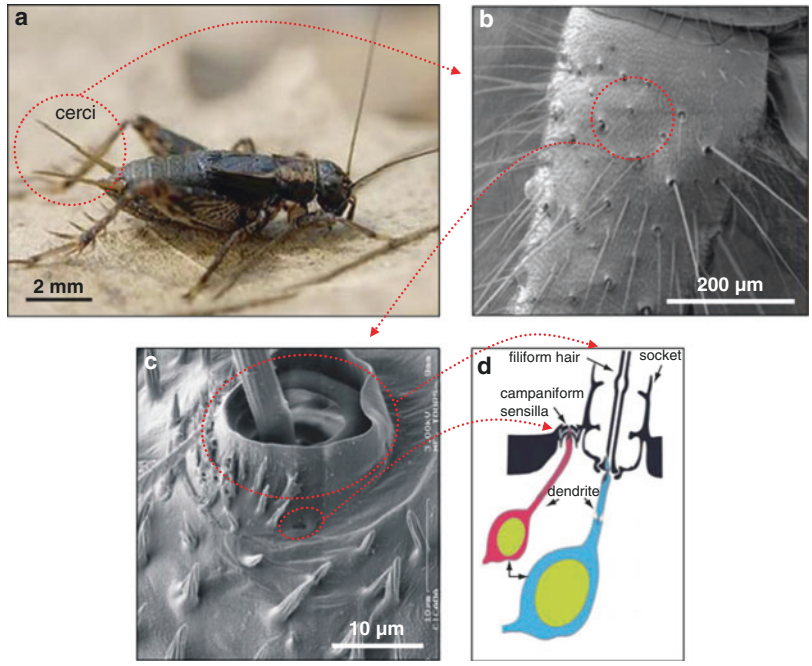
Inspiring Biological Systems

Flow sensing systems have been mainly studied on crickets and cockroaches, and to a lesser degree on spiders. The fact that the latter are predators of the former seems to matter little so far, and this entry will concentrate on the cricket for comprehensiveness. Flow sensors in crustaceans living in water have been studied to an even lesser degree but in a similar fashion, without any relationship with biomimetics, so that these works are not mentioned here.

Arthropod Hairs

The filiform hairs of many insects, spiders, and other invertebrates are among the most delicate and sensitive flow sensing cells: they measure displacements on the order of a fraction of the hydrogen atomic diameter (sensitivity ca. 10^{-10} m = 1 Å) and react to flow speeds down to 30 μm/s. If one considers the energy needed to elicit a neuronal spike, one finds that they react with a thousandth of the energy contained in a photon, so that they surpass photoreceptors. In fact, these mechanoreceptors work at the thermal noise level [19]. These hairs pick up air motion, implying that they are measuring both the direction and speed of air particles, in contrast to pressure receivers, i.e., ears. Many insects have ears, so that the extra information available in the flow field must be useful. The biomechanics of the filiform hairs have been studied with care since several decades by several groups worldwide, based on the analogy with a single degree of freedom inverted pendulum without bending (see the latest review of hair biomechanics in [8]).

Among insects, mainly cockroaches and crickets have been studied, because their airflow-sensitive hairs are put on two antenna-like appendages, the cerci (cercus in singular, see Fig. 1). Insect hairs have usually a high aspect ratio, with a length of a few hundreds of microns up to 2 mm, and with a diameter of less than a dozen of microns (Fig. 1). Their tapered shape have been found to have an influence on the drag forces. Hairs and sockets are ellipsoidal in cross section, which leads to a movement in a preferred direction. The hairs of spiders, called trichobothria, are often curved, are innervated by several sensory cells and have an ultrastructure which is favorable to drag while minimizing weight (a feathery structure which decreases the mass and hence their inertia considerably, while increasing the friction in air). The studied airflow-sensing trichobothria are located on the legs.



Biomimetic Flow Sensors, Fig. 1 Habitus of a wood cricket (**a**), with the two cerci highlighted, on which a large number of filiform hairs, intermixed with spines and chemoreceptive hairs, can be observed (**b**). The socket of a hair is a complex system made of several membranous stoppers. It deflection is sensed through another type of mechanosensor, the campaniform sensilla, which is a strain detector (**c**). Both act as a coupled system, the cuticula being deformed once the hair touches the border of the socket. Each sensor is innervated by a single hair cell (**d**)

These two aspects regarding the exact form of cricket and spider hairs show that natural selection is acting on the tiniest details of biological organization. The base of the hairs is much more complex, and its mechanics poorly understood. In crickets, only a single sensory cell is below the hair shaft, while spiders have several (Fig. 1). At the base of the socket, another mechanoreceptor type, the campaniform sensilla, reacts to deformation of the cuticula produced by the movement of the socket. Crickets possess often two sensilla around each hair. Thus the two sensors types act as a coupled system, extending thereby the range of forces it can measure.

Arthropods are very often quite hairy, and the high density of flow sensing hairs implies that they interact with each other in order to produce a high-resolution map of flow characteristics, acting like a flow camera (Fig. 1). Up to now, only the hydrodynamic type of interaction has been studied, called viscous coupling. It was found to be highly dependent on the geometrical arrangement of hairs, of their respective lengths and preferential planes of movement, as well as on the frequency content of the input signal [4]. Hairs often interact over long distances, up to 50 times their radius, and usually negatively. Short hairs in particular "suffer" substantially from the presence of longer hairs nearby, due to a shadowing effect. Positive interactions, where the flow velocity at one hair is increased by the presence of nearby hairs, have been however observed in real animals and reproduced computationally. The biological implications of these interactions have only very recently been addressed, and hint toward a coding of incoming signals which relies strongly on the specific sequence of hairs been triggered [16]. In other words, the signature of the incoming signal is mapped into a given sequence of recruited hairs, which in turn produces a typical sequence of action potentials.

Single hairs, or groups of hairs, are not placed at random on the body. For the same reasons that the

exact shape of hairs and their relative position within a group have been most likely molded by natural selection, the position of the hairs on the sensory organ must be shaped by selection. This aspect of mechanosensory research is however badly neglected. As for the positions along the cercus, the presence of a potential acoustic fovea (i.e., a location with particularly high acuity) at the base of the cercus has been hinted already twice, not only due to the highest hair density in this region, but also because it corresponds to the region with the largest flow velocities, the cercus being the largest there. Putting hairs radially around the cercus enables crickets also to pick up transversal flows, whose peak flow velocities are larger than in the situation where the hairs would be placed on a flat surface; in the later case, hairs are submitted to longitudinal flow with lower peak velocities. In summary, where you put your sensors relative to your body geometry matters a lot.

An action potential triggered by a moving hair ends up directly in the terminal abdominal ganglion (TAG), a local neuronal processing unit. Information from all the hairs, as well as from other sensors, converges there and is processed by interneurons. The compression and convergence of information at this stage is huge: About 1,500 afferent neurons of hairs are connected to only some 20 interneurons [10]. The fact that invertebrates possess few large, singly identifiable neurons enabling comprehensive mapping and repeated recordings of activities is a unique asset which explains the interest in such exotic systems. Information coming from the central brain as well as from the higher ganglia also descends into the TAG. Once processed, the combined information moves up quickly toward higher neuronal centers, in particular the ganglia in which the hindleg movements are being decided. This local feedback loop, with little input from the main brain, enables the animal to process vital information and act accordingly very quickly. As so often with invertebrates, what can be processed locally should be done so, a distributed processing scheme which explains why biomimetics has so much to gain from this group of animals.

The last level of integration is behavior, and flow sensing is known to be of importance in predator and prey perception, sexual selection, and most likely other context, such as noticing its own speed and movement. Predator avoidance is obviously a major selection force, where speed is of paramount importance. Jumping or running away is the behavior which is elicited using appropriate stimuli. The cricket possesses in the TAG an internal map of the direction of the stimuli from the outside world and the geometric computation of the direction of incoming flow by the cercus is one of the nicest case studies of spatial representation [10]. Computing the speed of an approaching predator is also carried out by the TAG, and has been only recently established using appropriate stimuli. Where to jump is a different question, in which directing stimuli and other conditions intervene.

Natural selection therefore acts along the full chain of information transfer, from acquisition and processing, up to actuation. This is important to restate in a biomimetic context, as the extreme sensitivity on the biomechanical side of the hair shaft, which has been the exclusive focus of the engineer attention, could be otherwise lost into an inefficient sequence of information transfer. As of today, one has however very little information about the constraints acting on the different parts of the chain, and hence no idea about their optimization levels.

Fish Neuromasts

Hair cells of the lateral-line system of fish are able to detect water displacement of the order of ca. 20 nm [2]. This ability is used in various ecological contexts, such as prey and predator localization, schooling, and most likely a host of other behaviors. Fish can also detect obstacle-created flow distortions of a flow field they have generated. The blind cave fish, *Astyanax fasciatus*, is an interesting model because it is believed that this species, having lost sight and having to maneuver in darkness, possesses particularly sensitive hair cells. The spatial discrimination ability, through self-induced water movements when approaching objects, of the blind cave fishes is about 1 mm^2. The hairs of

fish are located within specialized units, called neuromasts. These neuromasts, which are localized in many different parts of the body and which do not display a strong correlation with the flow structure of their environments, are located either just on the skin (called superficial neuromasts), or in specialized fluid-filled canals which produce a reticulated pattern on the head and along the body sides. A single neuromast consists of sensory cells that project into a jelly-like material, the cupula, in which several hairs (or cilia) of different lengths are embedded. The cupula is therefore the medium which connects the hairs and the external fluid.

An Engineering Perspective on Biomimetic Hairs

Hair-based flow sensing systems of arthropods and fish contain various levels, ranging from the mechanical structure, to the generation of action potentials in neurons, and signal transport, collecting and processing signals from many (hair-) sensors in the various neuronal centers. Taking inspiration from such complex systems is a non-trivial endeavor. The mechanics of these sensory systems are relatively easy to mimic in an engineering context, at least in comparison to the complexity of the neural system which requires delicate materials (e.g., membranes with ion-channels) and fluids with ion concentrations robustly and actively maintained by the organism. Consequently, up to date most of the bio-inspired flow sensors actually are defined merely by the fact that they use a hair, if needed covered by a cupula resembling cover, to capture viscous drag forces. Having no neural system in an engineering context implies that signal transduction, i.e., from the mechanical to the electrical domain, needs to be done by rather "un-bio-inspired" methods, e.g., using capacitive or piezoresistive interrogations methods. The next level up, namely the arrangement of the hair sensors into an array like structure (a cercal canopy for crickets and a lateral line for fish), is again open to engineering approaches. Mimicking the entire hair-based sensory system requires the ability to recreate dense hair-sensor arrays, including the associated electronic interfacing, asking for engineering solutions to the multiplexing problems. On the signal processing level, the biomimetic content can be high again, e.g., by exploitation of artificial neural networks to extract features and events from the multitude of array signals. The insight in the neuronal architecture of the animals needed for such exercise is however at best sketchy, with the exception of the architecture behind the directional acuity of crickets.

In view of the overall shape and sizes of the hairs in combination with the high densities and large numbers of hair sensors on the hair canopies, it is obvious that an artificial version of this kind of sensor arrays cannot be assembled but should be made monolithically. As it turns out, the range of dimensions is well attainable by micromachining techniques as used in the field of micro-electro-mechanical systems (MEMS). However, the materials generally used in MEMS are rather stiff compared to the materials found in nature, calling for some creative engineering.

Various types of hair-based flow sensors, for operation in air as well as in water, have been developed using both conventional machining and micromachining technologies. They are based on a variety of materials (including semiconductor, oxide, and polymers) and transduction principles (including electrostatic sensing, piezoresistive sensing, thermal sensing, etc.). However, artificial hair-cell (AHC) sensors represent unique challenges to microengineering design and fabrication. The high aspect ratio, vertical, hair must be fabricated using processing steps that are amenable to engineering, manufacturing and eventually scaled production. The overall process must be scalable to large areas as the AHC sensors are often used in arrays. Ideally, the sensors should be integrated with highly sensitive transduction mechanisms to couple mechanical input to electrical output. Further, it is desirable that the AHC sensors be made on a substrate that also houses the integrated electronics elements for signal conditioning and amplification.

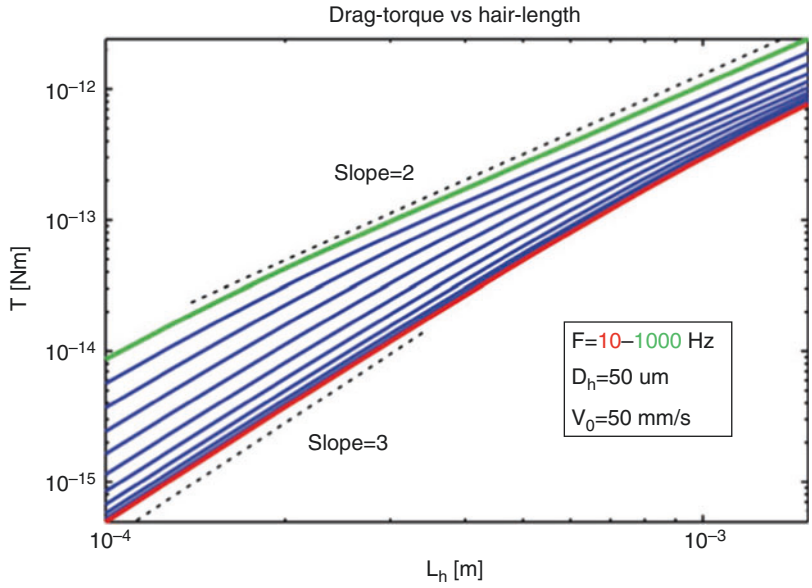
For reasons outlined above, most of the hair-based flow sensors reported in literature are made using MEMS fabrication technology. Ironically, the fabrication of the relative large hair structures poses considerable difficulties in this technology. Two basic types of artificial hairs can be distinguished, namely, hairs fabricated in the wafer plane and hairs fabricated perpendicular to the wafer plane. The first is straightforward since surface micromachining techniques can be used. However, surface-micromachined hairs cannot easily be combined into high-density arrays. An overview of the various reports in literature on ciliar inspired actuators and sensors is found in [22].

Ozaki et al. [17] were probably the first to provide flow sensors inspired by insects. Their piezoresistance-based sensors had either 400–800 μm long hairs in plane, being cantilevers exposed to flow after removing large part of the silicon substrate, or consisted of wires manually glued to a cross-shaped piezoresistive structure delivering two DOF sensitivity. Characterization by continuous flows in the m/s range proofed the functionality and directivity of the sensors. A comparable in-plane hair fabrication process was used in [12], but incorporated a stress gradient in the artificial hairs to make them curve out of plane. Readout is obtained by tracking the flow-induced vibrations of a set of dissimilar hairs resonating at different frequencies when exposed to specific flows. Argyrakis et al. [1] also use in-plane cantilevers with piezoresistive readout connected to an artificial neuronal circuit for spike generation. An alternative approach was proposed in [15] where the actual hair consisted of an optical fiber with reduced optical transmission when displaced by viscous drag. The sensor is meant for DC-flow measurements and achieves high volumetric sensitivity (μL/min range). Xueb et al. [20] more or less reverted to the structure of Ozaki using a cross-shaped frame with piezoresistors with a plastic hair of about 5 mm length. A nice twist on the hair sensor is presented in [14] by exploiting piezoelectric transduction by polyvinylidene fluoride in thermo-direct written clamped–clamped beam-type structures. Recently, interesting results have been obtained using a straightforward cantilever design in combination with appropriate electronic feedback which helped to incorporate adaptation of the sensor allowing for sharp filtering and gain [13].

Hair-Sensor Systems for Operation in Air: A Case Study

Over the course of several years, the possibilities to fabricate hair-based flow sensors and sensor arrays inspired by the hair sensors and canopies as found on the cerci of crickets have been explored. In several sensor iterations, understanding optimization criteria alongside with viable and robust fabrication technology and suitable interfacing was targeted, not only for single hair sensors but also for arrays.

The scaling of the hair sensors is clearly confined by various requirements. First of all the sensors are supposed to capture a sufficient amount of drag-force. Evidently, the longer and thicker the hairs, the larger the drag torque is (see Fig. 2). Even more so since air flowing over a substrate forms a boundary layer in which the flow-velocity transitions from zero, at the substrate, to the maximum (far-field) value. This boundary layer depends on viscosity and becomes thinner with increasing frequency. Next, for given drag torque the angular rotation will increase when the rotational stiffness decreases. Hence, a small rotational stiffness will be beneficial for the sensitivity and flow-velocity threshold. However, animals live in dynamic environments in which the amount of information derived from the flow sensors not only depends on the sensors sensitivity but also on the bandwidth in which they are sufficiently sensitive. Combining these observations one finds, unfortunately, that this bandwidth goes down with the square-root of the ratio of the spring stiffness and moment of inertia. In order for the animals to capture a large amount of drag torque while still maintaining a low spring-stiffness at a usable large bandwidth, it turns out that long and thin hairs are beneficial. Parametric studies on the drag-force as modeled by Stokes'



Biomimetic Flow Sensors, Fig. 2 Viscous drag torque on a cylinder of length L_h, diameter of 50 µm for various frequencies with flow amplitude of 10 mm/s

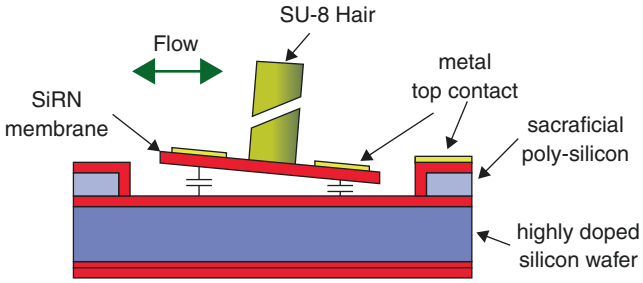
expressions showed that the dependence on hair diameter is far less than the dependence on hair length. Based on the mechanical responsivity and available bandwidth, one can propose a figure of merit (FoM) given by:

$$\frac{L^{1/2}}{\rho^{1/2} S^{1/2} D^{2/3}} \tag{1}$$

where L and D are the hair length and diameter, S is the rotational stiffness, and ρ is the specific mass density of the material from which the hairs are formed. The FoM is the product of the bandwidth of the sensors (expressed by the resonance frequency, i.e., square root of the rotational spring-stiffness S over the moment of inertia J) times their mechanical responsivity (i.e., the magnitude of angular rotation α per unit of flow-velocity V). Clearly it pays off to have long and thin hairs mounted on flexible suspensions.

The conditions under which hair sensors have to operate are such that the Reynolds (R) and Strouhal (St) numbers are relatively low. For a hair diameter of 25–50 µm, an air-oscillation frequency of 250 Hz and a flow-velocity amplitude of 10 mm/s, R varies between 0.008 and 0.016 and St between 1.96 and 3.92 (for the flow around the hairs). The rather small Reynolds numbers and the large hair-length to hair-diameter ratio allow using the Stokes expressions for the drag torque exerted by the airflow on the hairs. Furthermore, these hair sensors are mounted on flat substrates allowing the use of the Stokes expressions for a harmonic flow along the hairs. These expressions predict a viscous flow over an infinite substrate to be harmonic in time with zero flow velocity and 45° phase advance at the substrate interface and a boundary layer thickness (δ_b) proportional to the inverse of $\beta = (\omega/2\nu)^{0.5}$ where ν is the kinematic viscosity ($1.79 \cdot 10^{-5}$ m^2/s for air at room temperature) and ω is the radial frequency. The Stokes expressions can be usefully employed for this situation [7]. As an example, for a harmonic airflow of 100 Hz, the boundary layer is roughly 0.5 mm. Note that under most conditions the artificial hairs can be assumed infinitely stiff and the rotation angles are rather small (in the order of 1–100 mrad amplitude per m/s flow-velocity amplitude).

On a basal level the simple principle of a hair sitting on a torsional mount allowing for drag-induced rotation has been the starting point for these artificial hair sensors (see Fig. 3). In the implementation, the hairs are made of SU-8. Using two deposition-exposure cycles and



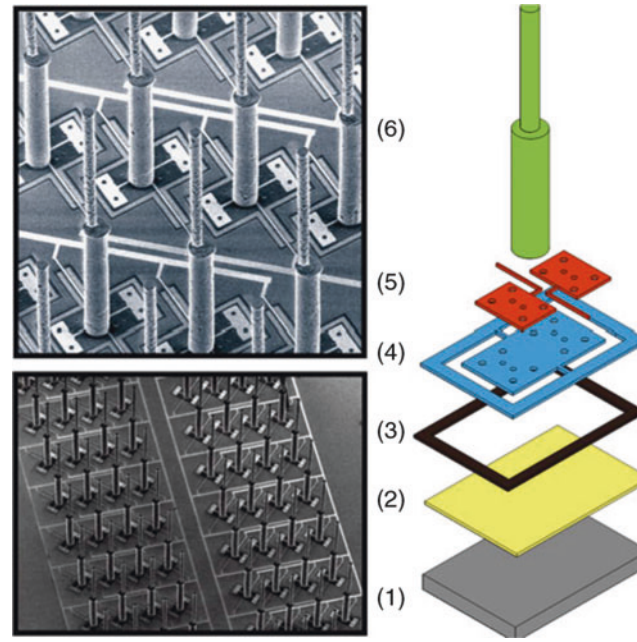
Biomimetic Flow Sensors,
Fig. 3 Schematic of artificial hair-based flow sensor with differential capacitive layout

subsequent development, these hairs can be made up to 1 mm long with diameters of about 50 μm for the bottom part and 25 μm for the top part. This configuration hardly reduces the drag torque but decreases the moment of inertia of the artificial hairs by about 65 %. The rotational freedom comes from two torsion beams. These beams as well as two membranes connected to the beams are made of silicon nitride. On the membranes an Al layer is deposited which forms a capacitor together with the underlying highly doped silicon substrate. On drag-induced tilting of hair and membranes, the capacitances change in opposite fashion allowing for a differential capacitive readout scheme. The hairs sit on the membranes as well. Rotational freedom is obtained when the structure is released by sacrificial etching of the poly-silicon layer (0.5–1.0 μm thick) that sits between the membrane and a protective silicon-nitride layer on the silicon substrate.

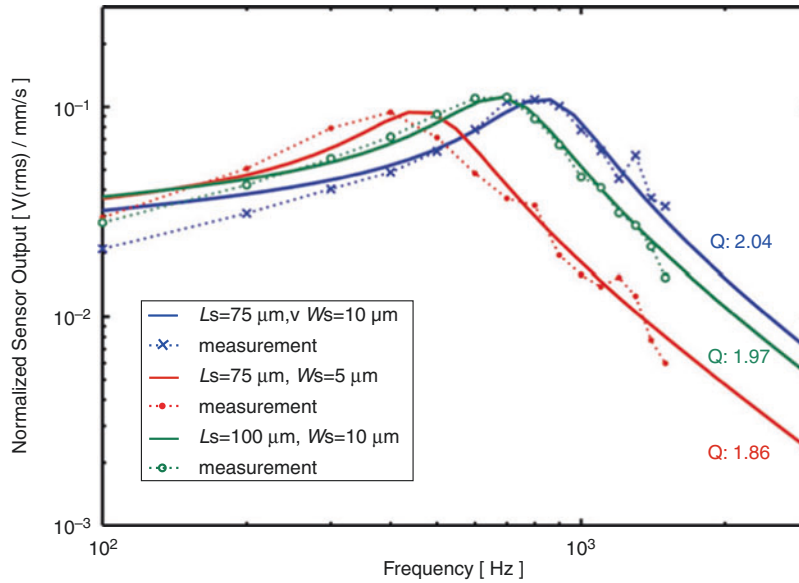
The choice for capacitive transduction is motivated by the fact that differential capacitive readout can be very sensitive and allows for rejection of various common mode signals (e.g., those that originate from flow in the direction perpendicular to the plane in which the hair is supposed to tilt). Moreover, contrary to e.g., thermal readout, it requires little power and does not thermally pollute the air around the sensors. This may not be important for a single hair but since the interest is in larger arrays, thermal effects could be far from negligible. Additionally capacitive readout allows for interrogation of many hair sensors in an array by frequency division multiplexing. There are also some clear drawbacks: the electronics required for capacitive readout are far more complex, certainly when compared to piezoresistive transduction and the method asks for rather complex mechanical structures when applied to sensors for use in liquids, especially in water. In the latter situation, the volume between the electrodes is ideally sealed in order to prevent liquid to get between since this could lead to electrolysis and large viscous damping. Figure 4 shows an exploded view of these capacitive, artificial hair-based flow sensors. Measurements of these sensors are shown in Fig. 5. Due to a combination of the boundary layer effect (high pass) and the second-order mechanics (i.e., its behavior is described by a second-order ordinary differential equation acting as low pass), the sensors show a band-pass characteristic.

The beauty of the neural system lies in its robustness as well as its largely localized generation of signals. In contrast, the capacitive measurement method employed in these sensors is plagued by parasitic capacitances that can hardly be avoided and have a far less localized nature. Changes in temperature, humidity, wiring geometry, etc. indeed cause fluctuations of the parasitics which are large relative to the small capacitive changes that the hair sensors induce. Therefore, control of the parasitics is of paramount importance for a proper device functioning. A reduction in parasitics can be obtained using much smaller electrode areas as well as exploiting shielding. Both have been shown to be feasible using silicon-on-insulator (SOI) technology. Results of the latest generation of sensors are shown in Fig. 6.

Successive iterations led to airflow sensors delivering sensitivities of about 400 μm/s (300 Hz bandwidth) with near-perfect figure of eight directivity. In order to do this, segmented cylinders, each up to 1 mm long, were used. Large arrays of sensors (over 100) can be fabricated and,



Biomimetic Flow Sensors, Fig. 4 *Right*: Exploded design view of a single hair sensor showing: (*1*) highly conductive silicon bulk (*bottom* electrode), n 200 nm thick SiRN layer for insulation and etch-stop, (*3*) Poly-silicon layer after final sacrificial etching, (*4*) 1 µm thick SiRN layer patterned into membranes, (*5*) 100 nm thick aluminum for *top* electrodes and (*6*) two 450 µm thick SU-8 layers patterned into a long hair. *Left*: SEM images of microfabricated hair sensor arrays with zoom of the two-stage artificial hairs

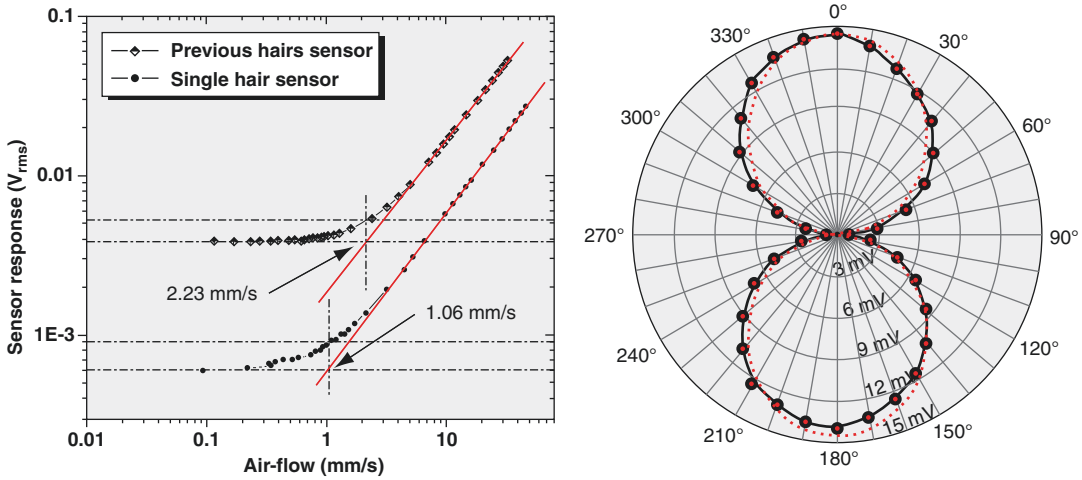


Biomimetic Flow Sensors, Fig. 5 Root mean square sensor output voltage normalized to 1 mm/s flow amplitude versus frequency for three sensor designs. *Solid lines* are model predictions, *symbols* measured responses. The values of the rotation beam length L_s and beam width W_s are indicated. The model predictions include fitted quality factors (Q) of 2.04 (*blue line*), 1.97 (*green line*), and 1.86 (*red line*) (Jaganatharaja et al. in preparation [11])

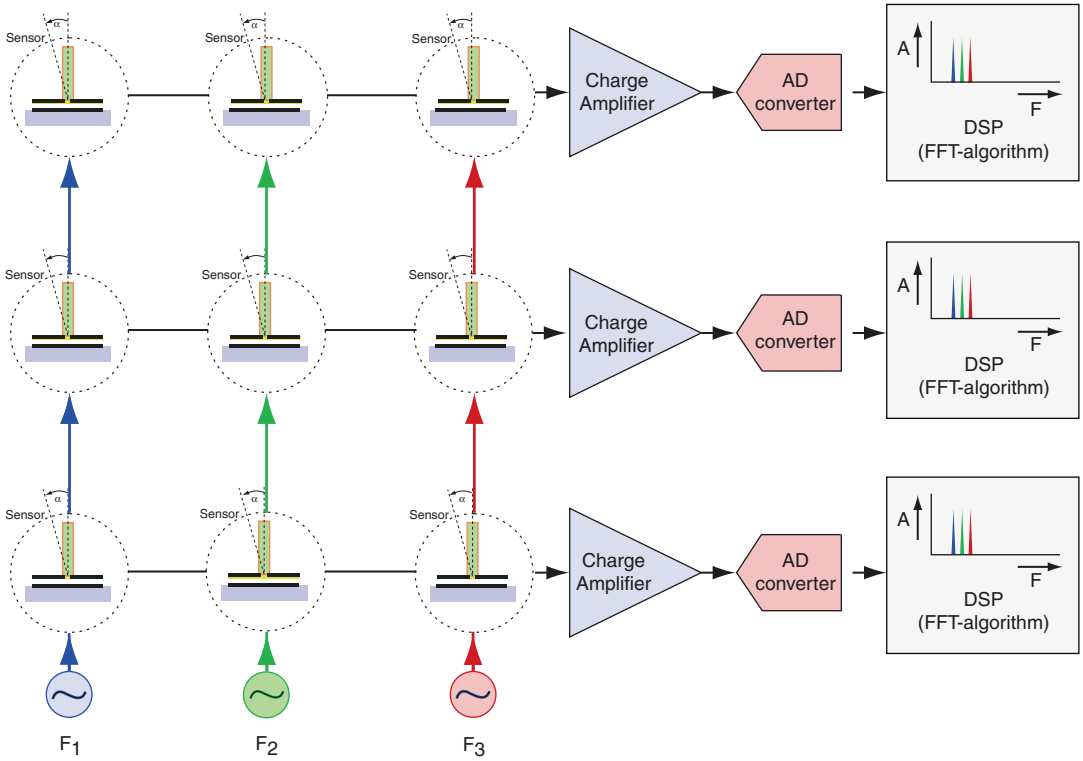
utilizing frequency division multiplexing, these sensors can be simultaneously and individually interrogated (see Fig. 7). This allows one to carry out measurements of spatiotemporal flow-patterns rather than single point measurements. Hence the aim is to develop a system delivering the functionality of a "flow camera" with sensor densities of about 25–100 sensors per mm^2 [6]. This development has largely benefitted from insights in the cricket cercal system.

Biomimetic Hairs in Water

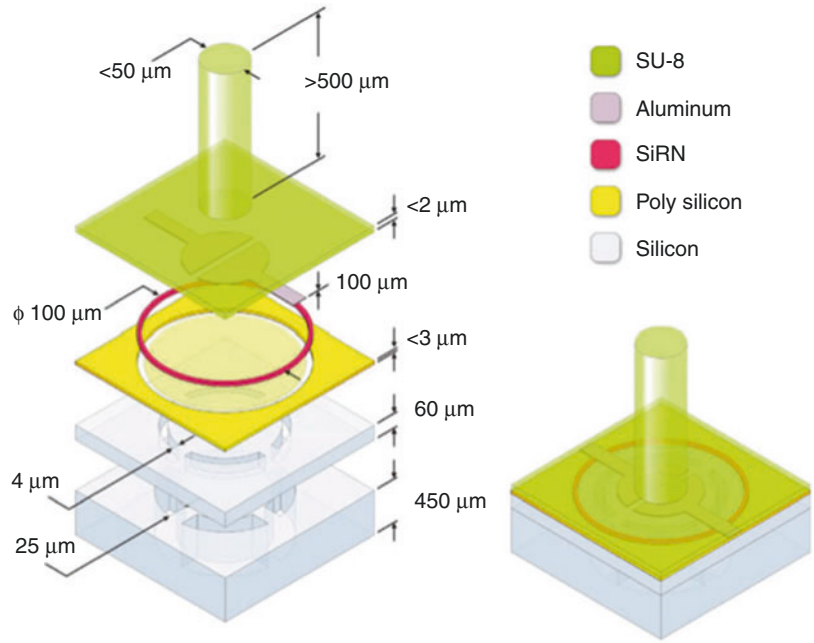
Though functionally not necessarily different, hair-based flow sensors in water are confronted with



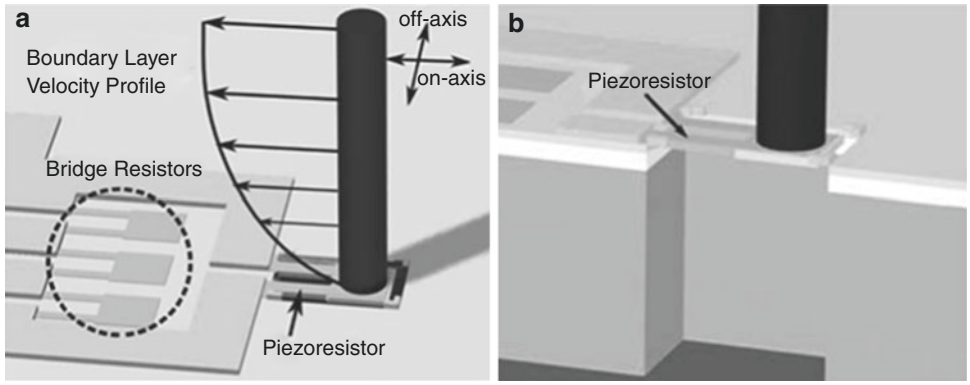
Biomimetic Flow Sensors, Fig. 6 *Left*: Output voltage of two generations of sensors measured in a 3 KHz bandwidth. The single hairs sensor is fabricated using a silicon on insulator substrate with a technology slightly different from the one indicated in Fig. 3. Measurements were done as a function of flow-amplitude for a 250 Hz sinusoidal flow. *Red lines* are asymptotic lines to determine the threshold flow. *Right*: directivity of the single hair sensor showing an almost ideal figure of eight (*red dashed lines*)



Biomimetic Flow Sensors, Fig. 7 Schematic of frequency division multiplexing as applied to hair-sensor array interfacing



Biomimetic Flow Sensors, Fig. 8 Exploded view of a capacitive aquatic hair-based flow sensor. The complex technology allows for sealed cavities (etched from the back side) safeguarding low damping and preventing electrolysis

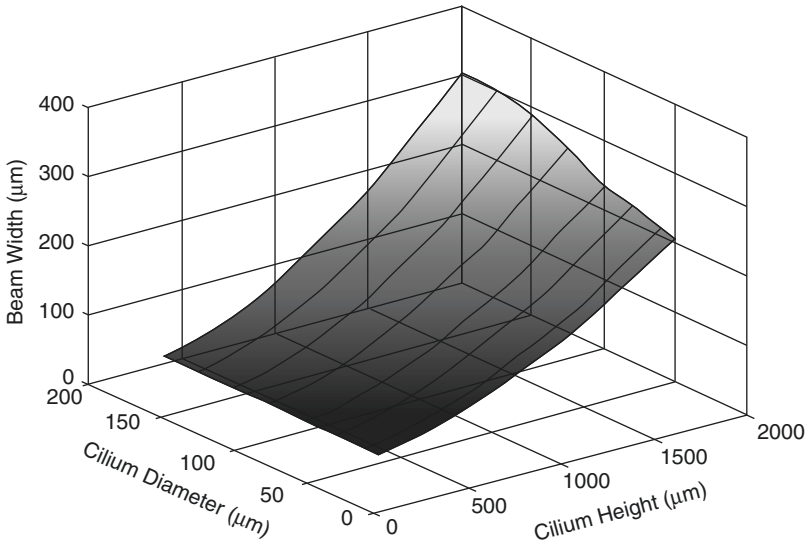


Biomimetic Flow Sensors, Fig. 9 (**a**) Schematic drawing of an individual AHC sensor with superimposed boundary layer flow velocity profile. (**b**) Cross-sectional perspective view of the AHC

some different scaling than sensors meant for air. The differences arise due to the fact that water has a higher mass density (about 1,000 times) and a smaller kinematic viscosity (about 20 times smaller). The first means that the effects of the moment of inertia of the hairs are smaller whereas the latter translates into boundary layers about 4.5 times smaller in water than in air. Hence, hairs for use in water can be both shorter and larger in diameter.

Hair-based sensing in water may exploit arbitrary transduction principles. However, the practical problems associated with capacitive readout caused by large viscous damping and necessary prevention of electrolysis may call for rather complicated technologies [9], as indicated in Fig. 8. This may be one of the reasons that up to now not much work has been presented on capacitive aquatic hair-based sensors.

The piezoresistance-based artificial hair cell device, shown schematically in Fig. 9, consists of a cilium located at the distal end of a paddle-shaped silicon cantilever [5]. Doped silicon strain



Biomimetic Flow Sensors, Fig. 10 3-D map showing the possible combinations of cilium height, diameter, and cantilever width that satisfy the design requirements

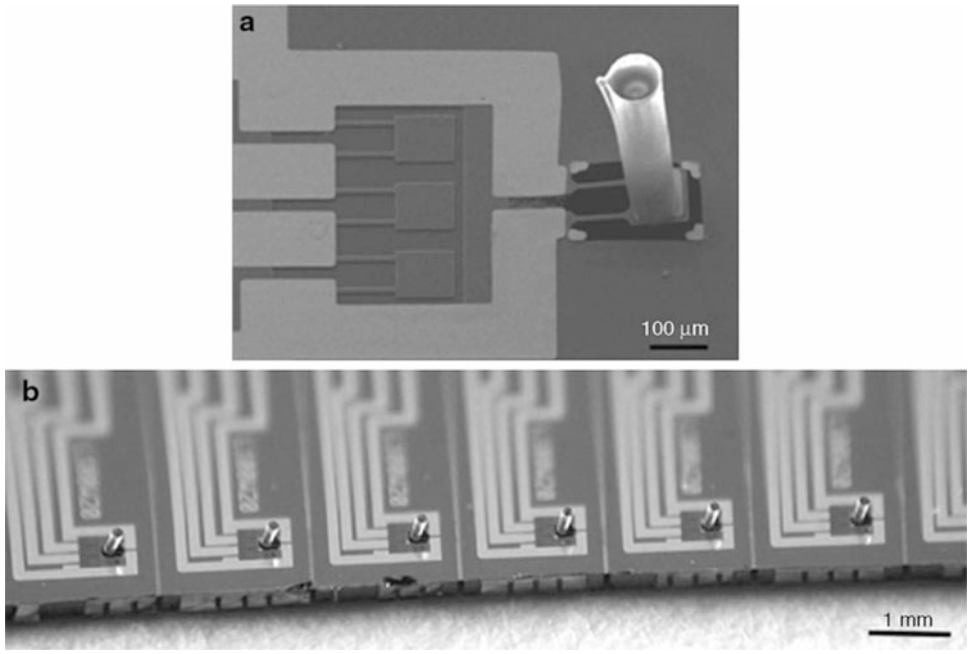
gauges are located at the base of the cantilever. The cilium is made of photodefinable SU-8 epoxy and is considered rigid. Lateral force along the on-axis acting on the cilium will create a bending moment (M), which is translated to the silicon beam through the stiff joint. The torque introduces a longitudinal strain and can be detected by piezoresistors at the base. The relation between the induced strain (ε) and the moment is given by

$$\varepsilon = \frac{6M}{Ewt^2} \tag{2}$$

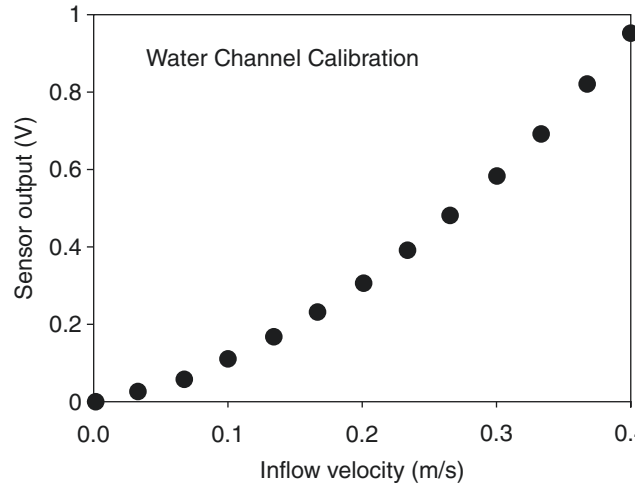
where E is the Young's modulus of silicon, w the cantilever width, and t the cantilever thickness. When used as a flow sensor, flow passing along the cilium introduces a bending moment (M) due to frictional and pressure drag. The device is primarily designed for underwater applications that require sensing of low-velocity and low-frequency flows. Structural and hydrodynamic model analyses suggest that the sensor's sensitivity is mainly determined by the following parameters: cantilever width (w), cantilever thickness (t), cilium height (h), cilium diameter (d), and sensor's distance form the leading edge (x).

For real-life underwater applications, good sensitivity in the sub-1 mm/s flow velocity range is desirable. Based on the existing signal conditioning circuitry capability, one can assume the minimum reliable voltage reading at 1,000 times signal gain as 1 mV. Taking 70 as a reasonable estimate for the gauge factor value, 1 mV output at 1,000 times signal gain translates into approximately 0.06 μ-strain using the quarter bridge model. In order to create and detect 0.06 μ-strain under 1 mm/s flow, the values of w, h, and d need to be carefully chosen. Figure 10 is a 3-D map with x, y, and z-axis being the three pertinent design variables, w, h, and d, respectively. A point within the plotted plane indicates a cantilever design that will satisfy the sensitivity requirement. The choice of w, h, and d are further narrowed down according to processing capabilities and frequency response considerations.

The cantilever thickness is chosen as 2 μm; this correlates to the thickness of the single crystal silicon epitaxial layer on the silicon-on-insulator (SOI) wafer used. For the analysis at hand, the sensor's distance from the leading edge (x) is assumed to be 2 mm, set by the die size. The devices are fabricated on SOI wafers with a 2 μm thick epitaxial silicon layer on top, 2 μm thick oxide, and 300 μm thick handle wafer. SU-8 epoxy is chosen for its ability to form rigid high aspect ratio structures. Figure 11 shows an SEM of an actual device, and an array of AHC sensors. A series of mechanical and electrical experiments were performed to characterize the sensor performance. The steady-state response to flow in water is presented in Fig. 12.



Biomimetic Flow Sensors, Fig. 11 Scanning electron micrograph of (**a**) an individual and of (**b**) an array of AHC sensors



Biomimetic Flow Sensors, Fig. 12 Experimental data of underwater steady-state flow velocity calibration of a single hair

The sensitivity of artificial hair sensors can be tailored up to some extent. However, as indicated by the FoM above there is a trade-off between bandwidth and sensitivity. There seem to be some indications that nature does not only, or not per se, maximize the mechanical response (e.g., the rotational displacement of the hairs) but also optimizes the energy captured from the environmental flow. The allometric scaling of cricket hairs seems to reflect the principle of mechanical impedance matching [19]. One may reason that capturing the maximum amount of energy will lead to the best signal-to-noise ratio whereas a high sensitivity may be obtained by proper adaptations, e.g., the lever ratio in the cricket hair-sensor system. Such systems are

found more often in nature; for example, the middle ear of the mammalian hearing system is merely an impedance transformer. It offsets the impedance difference by a factor of a few tens in order to capture maximum energy from air-borne sound waves and convert it to traveling waves in the cochlea despite a factor of a few thousand difference in acoustic impedance. Up to date little has been done in artificial hair sensors to address this aspect. Nevertheless some studies have shown the possibility to improve sensor performance by properly covering the sensory hairs by hydrogel structures, as discussed above. The high water content, reminiscent of the cupular material found in fish neuromast, alleviates the mechanical impedance difference, which together with the increased hydrodynamic surface, leads to a 40-fold improvement in responsivity [18, 21].

Perspectives

The performance of single flow detecting biomimetic MEMS hairs has steadily improved over the last two decades and the threshold flow velocity is now only about one order of magnitude higher than for the most spectacular biological sensors. This is quite an achievement, given that the signal transduction aspect of it is not carried out in a biomimetic approach. Some clever ideas might however originate from biology in this respect too. While active sensing, now found not only in humans and vertebrates but also in insects, has never been reported in these hairs yet, stochastic resonance has been already observed. This ability of biological systems to harness background or within-cell noise in order to improve signal capture needs to be seriously considered. Another as yet untapped possibility for increasing sensitivity and pattern recognition lies in the relative positioning of MEMS hairs, exploiting positive and negative viscous coupling [4]. Major challenges exist on the road until such sensors will be ready to use. They are brittle and break easily, in contrast to their natural counter parts which sustain very large deformations without harm. Furthermore, stiction problems (static friction) at the MEMS hair basis are occurring as soon as the flow velocity is too high, asking for incorporation of anti-stiction measures such as bumps and adhesion-reducing coatings. Nature is here again surprisingly resilient to extremes. Beyond potential technological applications, these hairs can now finally be turned into excellent physical models to tackle biological questions too difficult to address directly on biological materials: for example, the hydrodynamical interactions of tandem hairs moving within the same plane is nearly impossible to measure on real crickets, while MEMS of different lengths can be arranged at will.

Hair-based flow sensing is one of the few areas in which applied mathematics has been the common language of groups as diverse as field ecologists, materials scientists, aerodynamicists, theoretical physicists, computer scientists, and neurobiologists [3]. The mathematical toolbox, and its wide acceptance throughout the scientific community, is an unusual asset in biomimetics, as is the MEMS fabrication. Hair-based flow sensing is therefore *uniquely* suited for interactions among biologists and engineers working toward bio-inspired technologies.

Acknowledgments Without the many discussions we have had with our colleagues in the Cicada and Cilia projects this work would never have had the depth and breadth we have attained. These projects were the Cricket Inspired perCeption and Autonomous Decision Automata (CICADA) project (IST-2001-34718) and the Customized Intelligent Life Inspired Arrays (CILIA) project (FP6-IST-016039). Both projects were funded by the European Community under the Information Society Technologies (IST) Program, Future and Emergent Technologies (FET), Life-like Perception Systems action. We also like to acknowledge the contributions by our MSc and PhD students, PostDocs and technicians without which this paper would not exist.

Cross-References

- ▶ Bioinspired CMOS Cochlea
- ▶ Biosensors
- ▶ Integrated Micro-acoustic Devices
- ▶ Micro/Nano Flow Characterization Techniques
- ▶ Nanomechanical Resonant Sensors and Fluid Interactions

References

1. Argyrakis, P., Hamilton, A., Webb, B., Zhang, Y., Gonos, T., Cheung, R.: Fabrication and characterization of a wind sensor for integration with a neuron circuit. Microelectron. Eng. **84**, 1749–1753 (2007)
2. Bleckmann, H., Zelick, R.: Lateral line system of fish. Integr. Zool. **4**, 13–25 (2009)
3. Casas, J., Dangles, O.: Physical ecology of fluid flow sensing in arthropods. Annu. Rev. Entomol. **55**, 505–520 (2010)
4. Casas, J., Steinmann, T., Krijnen, G.: Why do insects have such a high density of flow-sensing hairs? Insights from the hydromechanics of biomimetic MEMS sensors. J. R. Soc. Interface **7**, 1487–1495 (2010)
5. Chen, J., Fan, Z., Zou, J., Engel, J., Liu, C.: Two dimensional micromachined flow sensor array for fluid mechanics studies. J. Aerosp. Eng. **16**, 85–97 (2003)
6. Dagamseh, A., Bruinink, C., Kolster, M., Wiegerink, R., Lammerink, T., Krijnen, G.: Array of biomimetic hair sensor dedicated for flow pattern recognition. In: Proceedings of the DTIP 2010, Seville (2010) ISBN: 978-2-35500-011-9
7. Dijkstra, M., van Baar, J., Wiegerink, R., Lammerink, T., de Boer, J., Krijnen, G.: Artificial sensory hairs based on the flow sensitive receptor hairs of crickets. J. Micromech. Microeng. **15**, S132–S138 (2005)
8. Humphrey, J.A., Barth, F.G.: Medium flow-sensing hairs: biomechanics and models. In: Casas, J., Simpson, S.J. (eds.) Insect Mechanics and Control. Advances in Insect Physiology, vol. 34, pp. 1–80. Elsevier, Amsterdam (2008)
9. Izadi, N., de Boer, M.J., Berenschot, J.W., Krijnen, G. J.M.: Fabrication of superficial neuromast inspired capacitive flow sensors. J. Micromech. Microeng. **20**, 085041 (2010)
10. Jacobs, G.A., Miller, J.P., Aldworth, Z.: Computational mechanisms of mechanosensory processing in the cricket. J. Exp. Biol. **211**, 1819–1828 (2008)
11. Jaganatharaja, R.K.: Cricket inspired flow-sensor arrays. PhD thesis, University of Twente, (2011), ISBN 978-90-365-3215-0, doi: 10.3990/1.9789036532150
12. Kao, I., Kumar, A., Binder, J.: Smart MEMS flow sensor: theoretical analysis and experimental characterization. IEEE Sensors J. **7**(5), 713–722 (2007)
13. Kim, H., Song, T., Kang-Hun, A.: Sharply tuned small force measurement with a biomimetic sensor. Appl. Phys. Lett. **98**, 013704 (2011)
14. Li, F., Liu, W., Stefanini, C., Dario, P.: A novel bioinspired PVDF micro/nano hair receptor for a robot sensing system. Sensors **10**, 994–1011 (2010)
15. Lien, V., Vollmer, F.: Microfluidic flow rate detection based on integrated optical fiber cantilever. Lab Chip **7**, 1352–1357 (2007)
16. Mulder-Rosi, J., Cummins, G.I., Miller, J.P.: The cricket cercal system implements delay-line processing. J. Neurophysiol. **103**(4), 1823–1832 (2010)
17. Ozaki, Y., Ohyama, T., Yasuda T., Shimoyama I.: An air flow sensor modeled on wind receptor hair of insects. In: Presented at International Conference on MEMS, Miyazaki (2000)
18. Peleshanko, S., Julian, M.D., Ornatska, M., McConney, M.E., LeMieux, M.C., Chen, N., Tucker, C., Yang, Y., Liu, C., Humphrey, J.A.C., Tsukruk, V. V.: Hydrogel-encapsulated microfabricated haircells mimicking fish cupula neuromast. Adv. Mater. **19**, 2903–2909 (2007)
19. Shimozawa, T., Murakami, J., Kumagai, T.: Cricket wind receptors: thermal noise for the highest sensitivity known. In: Barth, F.G., Humphrey, J.A., Secomb, T.W. (eds.) Sensors and Sensing in Biology and Engineering, pp. 145–159. Springer, Berlin (2003)
20. Xueb, C., Chena, S., Zhanga, W., Zhanga, B.: Design, fabrication, and preliminary characterization of a novel MEMS bionic vector hydrophone. Microelectron. J. **38**, 1021–1026 (2007)
21. Yang, Y.C., Nguyen, N., Chen, N.N., Lockwood, M., Tucker, C., Hu, H., Bleckmann, H., Liu, C., Jones, D. L.: Artificial lateral line with biomimetic neuromasts to emulate fish sensing. Bioinsp Biomim **5**(1), 16001 (2010)
22. Zhou, Z., Liu, Z.-W.: Biomimetic cilia based on MEMS technology. J. Bionic. Eng. **5**, 358–365 (2008)

Biomimetic Infrared Detector

► Insect Infrared Sensors

Biomimetic Mosquito-Like Microneedles

Melur K. Ramasubramanian and
Ranjeet Agarwala
Department of Mechanical and Aerospace Engineering, North Carolina State University, Raleigh, NC, USA

Synonyms

Bioinspired microneedles; Biomimicked microneedles; Bionic microneedles; Mosquito fascicle inspired microneedles; Nature inspired microneedle

Definition

Microneedles have evolved over recent years as a replacement for traditional hypodermic needles to minimize pain during transdermal drug delivery (TDD) or during blood draw for clinical analysis. Female mosquitoes have offered scientists an excellent insight into how an organism has evolved to drive a flexible microneedle into human skin and draw blood without causing pain to the host. Microneedle designs inspired by the anatomy of a mosquito fascicle and the mechanics of feeding by female mosquitoes have been developed by several researchers.

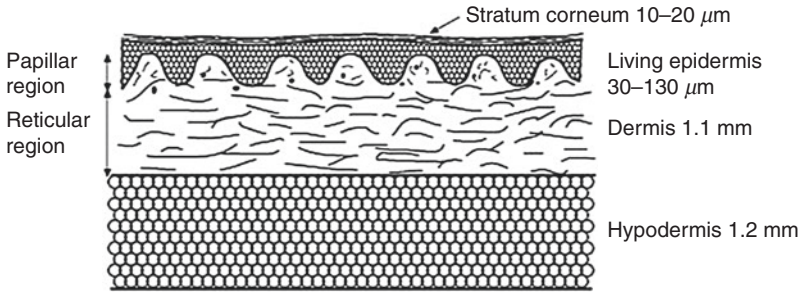
A microneedle is able to penetrate the skin painlessly when the critical load applied by it is greater than the load required to puncture the skin and the microneedle does not buckle. Once inserted, the microneedle draws blood or delivers therapeutic agents quickly into the blood stream.

Microneedles

Microneedles are used widely as a substitute for hypodermic needles for relatively painless blood draw, drug delivery, and for precise delivery of reagents in chemical and biological systems. Contrary to hypodermic needles, microneedles do not penetrate the skin far enough to contact, pinch, or pierce nerve endings. Khumpuang et al. [1] reported that penetration between 150 and 1,500 μm is painless due to the relative absence of nerve endings at these depths. This is further illustrated by viewing the human skin anatomy as depicted in Fig. 1 [2].

As shown in Fig. 1 [2], the human skin comprises three predominant layers. The outermost skin layer (living epidermis and stratum corneum) is approximately 150 μm thick and is devoid of blood vessels and nerve endings. The next skin layer (dermis) contains blood capillaries and is about 1,100 μm thick. Although the dermis has some sensory perception, it is mostly painless to penetration. The next layer of human skin (hypodermis or subcutaneous layer) consists of fat cells, large blood vessels, and nerve endings. Hence, penetration into this layer causes pain. However, with hypodermic needles, penetration into the subcutaneous layer is necessary to deliver drug into the larger blood vessels. Generally, a microneedle penetration is limited to the dermis layer and is effective in accessing blood capillaries for drug delivery or blood draw while minimizing if not eliminating pain in the process. Microneedles diameters generally range between 40 and 100 μm and can be solid or hollow, whereas the smallest size hypodermic needles are about 320 μm in outer diameter and 160 μm in inner diameter.

Researchers in the past have successfully created stainless steel and titanium microneedles claiming them to be painless [3–5]. Many microneedle models designed by researchers have been fabricated using silicon. Roxhed et al. [2] designed and built silicon microneedles that had side openings and ultrasharp tips. Needle tip radius has a direct bearing on the stress needed to puncture the skin. Their needles were cylindrical and tapered in shape, 400 μm long with a base diameter of 108 μm, and a tip radius below 100 nm. The bore of the needle was chosen to be



Biomimetic Mosquito-Like Microneedles, Fig. 1 Schematic representation of the anatomy of human skin [2]

eccentrically elliptic with major and minor axes dimensions of 40 and 62 μm, respectively. The side opening of the bore enables the formation of ultrasharp tips. Their needles exhibited insertion forces of 10 mN which was much smaller than previously reported data. The theoretical pressure required to penetrate human skin is given as 3.183×10^6 Pa. Silicon microneedles are generally very short, most less than 500 μm in length, and are generally painless when pushed against the skin.

Oka et al. [6] fabricated and tested silicon microneedles that were 40 μm in diameter, wall thickness of 1.6 μm, and about 1,000 μm long. They reported that their microneedle failed due to brittle fracture during insertion. Silicon microneedles which are brittle can fracture during skin penetration can be carried by blood to other parts of the body, potentially causing injury. Hence, efforts to make polymeric microneedles to prevent brittle fracture were attempted. Davis et al. [7] designed microneedles that were 720 μm long with a tip radius and thickness ranging from 30–80 to 5–58 μm, respectively. Their polymeric needles exhibited insertion forces between 0.1 and 3.0 N. Chaudhri et al. [8] fabricated high aspect ratio hollow micro needles made of SU-8 polymer based on the anatomy of a mosquito fascicle. Microneedles of about 1,600 μm in length, 100 μm in inner diameter, and about 15 μm in wall thickness were produced. The fabrication process was delineated but no results from performance studies were presented.

Reducing the diameter and wall thickness of hollow microneedles for painless insertion necessitates the use of high stiffness materials from a mechanics point of view. Although silicon microneedles are stiff, they are susceptible to fracture during penetration. On the other hand polymeric needles, although very tough, are not stiff enough to penetrate the skin painlessly at comparable dimensions without buckling. In nature, however, a mosquito uses its fascicle (the blood sucking tube) made of natural polymers and successfully penetrates the skin to draw blood. By studying the anatomy of a mosquito, the process a mosquito uses to insert its fascicle into the skin, and the process of blood draw, one can develop novel ideas leading to scientific breakthroughs in microneedle design, fabrication, and analysis.

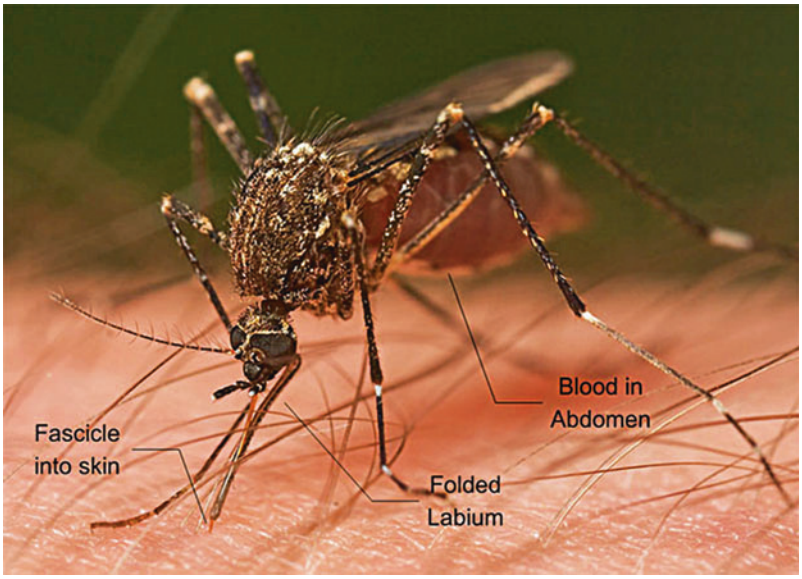
Bioinspiration and Biomimetics

Many of nature's organisms are involved in human skin penetration and blood extraction process as part of their life cycle. Meyers et al. [9] describe sharp cutting materials from plants, insects, fish, and mammals such as razor grass, bees, mosquitoes, piranhas, shark, rabbits, and rats. They report various biomimetic devices such as scissors inspired by piranha teeth, shredder cutting blades inspired by rabbit's incisors, and syringes inspired by a mosquito.

Why Mosquito?

Daniel et al. [10] compared blood feeding strategies of three common insects, mosquitoes (*Aedes aegypti*), bedbugs (*Cimex lectularis*), and body louse (*Pediculus humanus*) representing a wide range of sizes and feeding rates. Maximization of nutrient intake and minimization of blood feeding time seems to be the strategy used by all these insects given their anatomy. The results of their simulation model depicted that the mosquito fed at lower rates compared to that of a bedbug at a pressure drop that was one order of magnitude lower. The body louse on the other hand fed at one tenth the rate of a mosquito, but the pressure drop required was five times higher. The summary of their results showed that the tube radius was a critical variable that affected the feeding rate at a given pressure drop. Insects with larger food canal radius fed at faster rates at any applied pressure drop. The food canal radii for a bedbug, a mosquito, and a body louse were 8, 11, and 3.6 μm, respectively. Therefore, mosquitoes with their large syringe-like tubular food canal anatomy are an ideal model for bioinspiration and biomimetics for developing blood draw and drug delivery devices.

Figure 2 [11] depicts a mosquito that has just inserted its fascicle into the skin of a human host and has started to draw blood from the host.



Biomimetic Mosquito-Like Microneedles, Fig. 2 Mosquito drawing blood from a human host [11]

Mosquito Anatomy

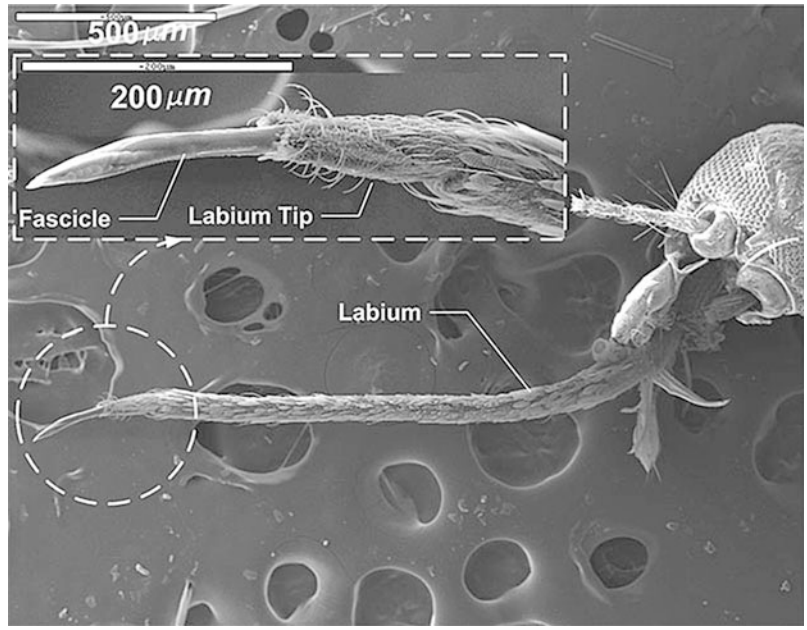
In order to understand how a mosquito draws blood from the host, it will be prudent to first introduce the anatomy of mosquito mouthparts, a compound structure, known as the proboscis.

The proboscis is 1,500–2,000 μm in length. The outer sheath is known as the labium that has a hairy structure at the end and has an inner diameter of 40 μm. The labium has a transverse slit for buckling out of the way during the insertion process. Inside the labium is the actual needle known as the fascicle with an outer diameter of 40 μm and inner diameter of 20 μm for blood draw. The fascicle of the mosquito which serves as a conduit for blood flow is generally retracted into the labium. It comprises of a labrum tube which is the main path for blood flow and hypopharynx which has a salivary duct. The fascicle has a maxilla with saw toothed tips. Figure 3 [12] depicts the tip of the fascicle of a mosquito protruding out of the labium. A fascicle extracted from the labium is shown in Fig. 4 [12]. The fascicle has a remarkable resemblance to that of a hypodermic needle. The fascicle's lateral dimension as shown in Fig. 4, decreases from about 10 μm to less than 1 μm at the tip in the last 50 μm length segment creating a very sharp tip.

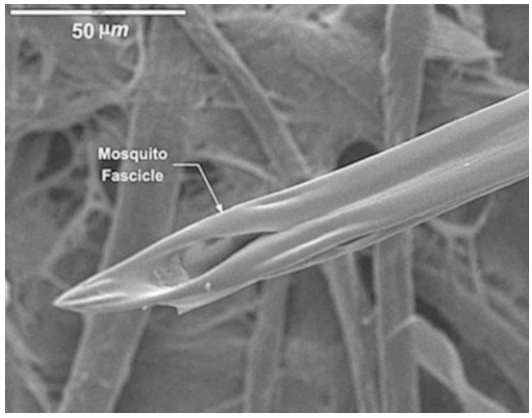
Mosquito-Inspired Microneedles

Many researchers have studied the mosquito anatomy and have designed microneedles by mimicking its morphology. Xin et al. [13] observed the surface morphology of mosquito and cicada fascicles using scanning electron microscopy. They drew inspiration from the fascicle and its jagged shape to design a biomimetic painless needle. The jagged shape of the microneedle design enabled drag reduction. They proposed that their needle design was able to pierce the skin without much resistance and the surface of their needle design exhibited smaller contact area and friction. Chakraborty et al. [14] researched a female mosquito's blood extraction activity to develop and conduct a fluid mechanics analyzis of a microneedle integrated with a painless blood extraction system. They used sputtering deposition method to fabricate titanium microneedles. They reported that their theoretical transport analysis agreed well with experimental observations in terms of flow rates.

Tsuchiya et al. [4] designed a titanium microneedle inspired by the anatomy of a female mosquito fascicle. They describe a self-monitoring of blood glucose device (SMBG) triggered by a shape memory alloy (SMA) actuator.



Biomimetic Mosquito-Like Microneedles, Fig. 3 (**a**) Mosquito proboscis attached to the head (**b**) Tip of the fascicle shown protruding from the labium [12]



Biomimetic Mosquito-Like Microneedles, Fig. 4 A mosquito fascicle tip extracted from the labium [12]

The system comprised a SMA-based indentation system, microelectric pump to extract blood, and a biosensor to evaluate the level of glucose. They evaluated the performance of the insertion system which required 0.1 N force to indent the skin and the pumping system extracted 2 µl/min of blood. Kosoglu et al. [15] developed pliant microneedles for transdermal lighting. These microneedles were inspired by the anatomy of a mosquito fascicle. The usage of these microcneedles as transdermal light probes or optical fibers has the potential to illuminate tissue at a greater depth for surgery or therapeutic procedures like treatment of cancer. They used a melt-drawing process to make various tip sized silica probes with an average diameter of 73 µm. They also experimentally investigated the needle performance in piercing porcine skins. They observed that sharper probe tips had a higher success rate of penetration compared to probes with more blunt tips.

Kong et al. [16, 17] studied the structure of Aedes albopictus mosquito fascicle using scanning electron microscope (SEM). They observed and documented the entire process from skin penetration to blood feeding. They experimentally measured the insertion force using a digital balance, a glass box, and a screen cloth. The mosquito was housed within the glass box and was free to move within it. The screen cloth connected to the bottom of the glass box deflected once the mosquito extended its fascicle to penetrate and draw blood from the volunteer's hands. The force causing the deflection of the skin cloth was recorded using a digital balance connected to the system. An average force value of 1.8 µN was required for the skin penetration, which is remarkably low. Tsuchiya et al. [3] designed and fabricated a painless needle inspired by the anatomy of

a female mosquito for blood extraction. The microneedles were made using sputtering deposition method wherein a rotary substrate was used to deposit the microneedle material, subsequently etched by a chemical solution, and finally heat treated to produce the microneedle. They also reported that the titanium microneedle could be inserted into the skin without buckling.

Aoyagi et al. [18] were able to create high aspect ratio microneedles using biodegradable polymers like PMMA, polyimide, SU-8, Parylene, and polycarbonate. They successfully created 1,000 μm long microneedles that were 10 μm in diameter. The microneedle was able to collect blood plasma with high reliability using capillary suction. Aoyagi et al. [19] used mosquito-inspired microneedles to investigate the resistance force during insertion into silicone rubber. They used a static and a vibrating needle for their experiment. Chichkov [20] proposed a two-photon-polymerization process to create a hollow microneedle and other medical devices using this process. The process used a hybrid material called ormocers; needles were able to penetrate porcine tissue surfaces without fracturing. McAllister et al. [21] used microfabrication techniques to produce microneedles motivated by the anatomy of female mosquitoes. They fabricated microhypodermic needles using silicon and assembled them with heat-controlled bubble pumps. These microneedles allowed flow of water through their bores under acceptable pressures and were able to permeate skin with a greater degree of ease. Suzuki et al. [22] devised a microglucose measuring system comprising a microneedle based on the feeding anatomy of a mosquito. Their system comprised a microneedle, a micro glucose sensor, and a sampling mechanism. Yoshida et al. [23] developed a microneedle design that mimicked a mosquito fascicle. They used human hair as a starting point for manufacturing microneedles. They drilled a hollow bore into the center of a human hair using a microdrilling machine. They created a pen-shaped micro needle which was 1.1 mm long, 80 μm base diameter and, 40 μm tip diameter. Finally, they bleached the drilled human hair using a diluted bleach solution.

Prow et al. [24] researched and developed the Nanopatch (NP). The NP design was inspired by mosquito anatomy. The NP consisted of projections of closely packed microneedles that are smaller than other available microneedles (less than 100 μm). The NP was coated with antigen, adjuvants, and DNA payloads which were transdermally delivered to mice in the lab. The NP exhibited good results with regards to a painless and efficient vaccine delivery system. Izumi et al. [25] used isotropic dry etching process to create long silicon microneedles with three-dimensional sharp tips. Their design did not impose any limit on either the length or the shape of the microneedle. They were able to fabricate straight, jagged, and harpoon like microneedles that mimicked the anatomy of a mosquito fascicle. They also reported that their microneedles with 3-D tips penetrated artificial skins relatively easier compared to prior microneedle designs.

Roxhed [5] investigated a patch-based microneedle drug delivery system. These microneedles were inspired by mosquitoes and interfaced with small dispenser mechanisms. The dispenser mechanisms were triggered by thermal actuators and contained highly expandable microspheres. The microspheres when actuated expanded into a liquid reservoir and dispensed liquid through the reservoir outlets. These microneedles were made of monocrystalline silicon and were chemically treated by a deep reactive ion etching process. The microneedles were hollow with opening along the side to have better penetration ability. The patch was tested for insulin delivery and exhibited promising results with comparable performance to other drug delivery devices.

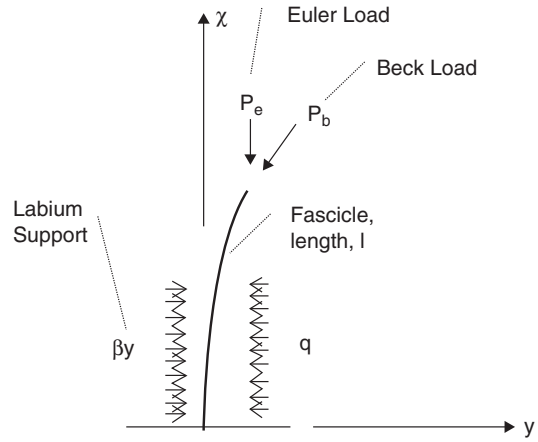
Mosquito Fascicle Insertion Dynamics

Almost all prior work to produce biomimetic microneedles were based on morphology as discussed in the earlier section. Very few published work exist to assess the significance of the dynamics of a mosquito penetration process in the design of bioinspired microneedles. It can be

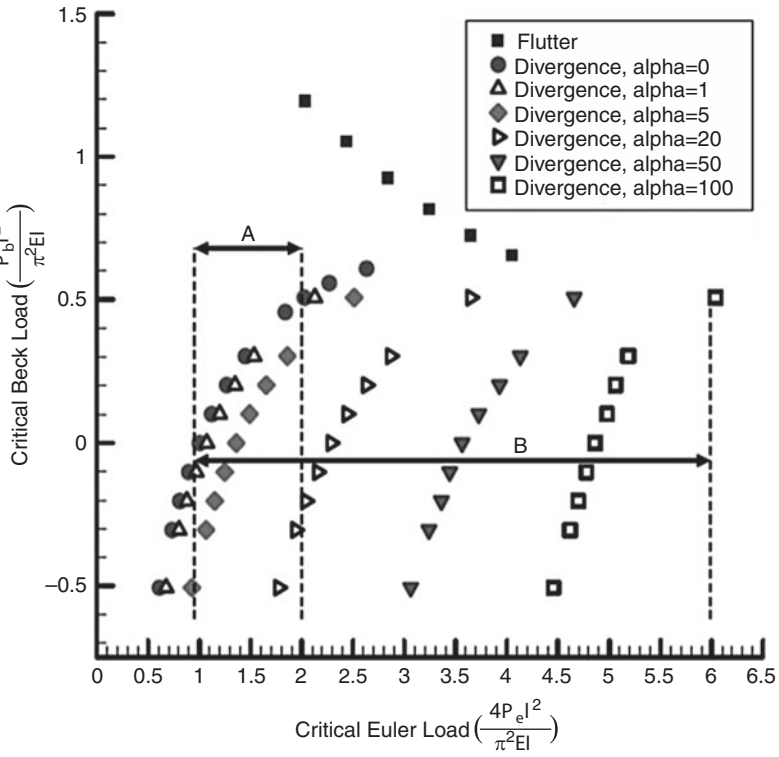
easily shown that if the morphology is reproduced without considering the material, it is possible to build a needle with high stiffness materials such as metal and silicon and penetrate the skin. If a polymeric needle is produced based on the morphology alone, the needle will not be able to penetrate the skin without bucking for comparable dimensions and material properties to that of a mosquito fascicle. Hence, studying the dynamics of the process of insertion of a fascicle into the skin by a mosquito is important. Ramasubramanian et al. [12] observed in detail how a mosquito (Aedes Aegypti) uses its fascicle to penetrate the skin, and have carried out detailed anatomical studies, and subsequently modeled the penetration process and provided insights into mosquito skin penetration process.

As observed in their research, the mosquito contacts the host to commence the feeding process. It hovers on the host from one spot to another until it is able to thrust its proboscis into the skin without bending it. Once the proboscis is thrust into the skin, the fascicle affixes itself into the epidermis of the skin with its tip and commences probing with irregular frequency. The mosquito keeps on applying axial force for some time by keeping its anatomy poised to prevent proboscis buckling and completes the penetration process by oscillating its head. The mosquito head is observed to oscillate at frequencies near 15 Hz at the commencement of the penetration process. The progressive insertion of the fascicle results in its shortening of the free length making it increasingly stable and consequently the mosquito head oscillates at relatively lower frequency, nearly 6 Hz. The labium with its axial slit is detached from the fascicle near the mouth and forms a loop that looks like a question mark (see Fig. 2). When the mosquito reaches a capillary in the dermis layer, it starts to draw blood until it is full. Figure 2 shows blood in the fascicle and collected in the abdomen.

Study of the penetration mechanics of mosquito fascicle into the skin is paramount as it is antecedent to the feeding behavior of the mosquito. Understanding the forces involved during the penetration process provides critical insights in the process of fascicle penetration as the mosquito tries to prevent the onset of buckling by keeping its posture perched appropriately. One of the movements is the forward tilt of the mosquito head between 10° and 20° with respect to vertical as shown in Fig. 2, which results in the application of a follower force which is a constituent of axial thrust. Such a feat may translate into considerable enhancement of buckling load once the fascicle is thrust into the skin. This effect of increasing the buckling load is further compounded by the labium with the labium providing lateral support during penetration. In Fig. 2, it can be seen that substantial part of the fascicle is supported by the labium and at the tip only a small part of the fascicle is visible before entering the skin. The penetration process and the mechanism of buckling force enhancement by follower force and lateral support has been described effectively by a mathematical model, quantifying their relative importance to the penetration process [12]. The fascicle (an inspiration for microneedle design described earlier) is subjected to an Euler load (P_e) and a nonconservative load known as Beck Load (P_b) supported laterally by the labium as shown in Fig. 5 [12]. The labium is treated as an elastic foundation which resists sideways motion of the fascicle. Here q is the distributed force per unit



Biomimetic Mosquito-Like Microneedles, Fig. 5 Schematic representation of fascicle-skin interaction with a mosquito applying the force components at the upper end [12]



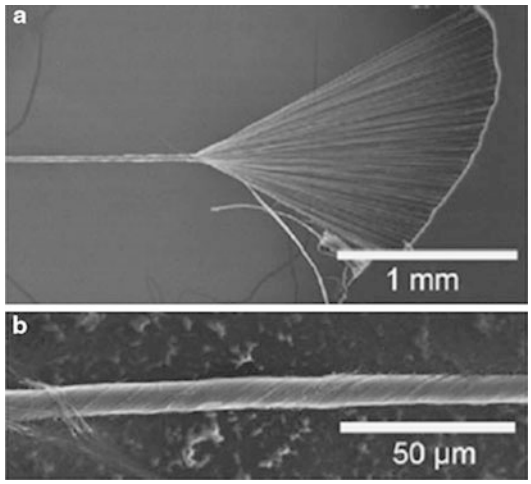
Biomimetic Mosquito-Like Microneedles, Fig. 6 Critical load pairs plot showing the effect of lateral support by the labium and the application of Beck load [12]

length and β is the foundation modulus (force per unit deflection per unit length).

The equation of motion as obtained is given by,

$$\frac{\partial^4 y}{\partial x^4} + \left(\frac{P_e + P_b}{EI}\right)\frac{\partial^2 y}{\partial x^2} + \frac{m}{EI}\frac{\partial^2 y}{\partial t^2} + \frac{\beta}{EI}y = 0 \quad (1)$$

With appropriate boundary conditions, namely, $y(0,t) = 0$; $y'(0,t) = 0$; $y''(l,t) = 0$; and $y'''(l,t) + \left(\frac{P_e}{EI}\right)y'(l,t) = 0$ where, m is mass per unit length, E is the Young's Modulus and I is the moment of inertia, the general solution for the differential equation can be obtained [12]. As depicted in Fig. 6, the distance labeled "A" illustrates the effect of nonconservative force as an increase in Euler load from 1 to 2 is attained solely by applying a Beck load of around 0.5 without substantial increment of lateral support. The distance labeled "B" illustrates the effect of foundation stiffness as an increase of Euler load from 1 to 6 is attained by increasing the foundation stiffness value from 0 to 100 for a value of nonconservative load of 0.5, representing the maximum enhancement with both factors. Figure 6 further illustrates that in the absence of nonconservative load, an increment of Euler load by a factor of nearly 5 is still attainable for foundation stiffness values of 100. It is also observed that in the absence of foundation stiffness, the Euler load can be increased by a factor of 3.5 for a Beck load of 0.7 before the onset of flutter instability thus representing the theoretical upper bound. Therefore, the lateral support plays an important role in helping the mosquito reach a force above the limit to penetrate the skin by increasing the critical buckling load by a factor of almost 5. From the mathematical model and numerical solution, it is evident that the lateral support along with Beck load has the effect of increasing the Euler buckling load by a factor of 6 or approximately 18 mN while the lateral support by itself increases the Euler buckling load by a factor of 5. Previous research has shown that the lowest penetration force exhibited by sharpest microneedles is around 10 mN and hence it can be seen that the presence of lateral support would



Biomimetic Mosquito-Like Microneedles, Fig. 7 Ultralong carbon nanotubes spun into micro-yarns [26]

suffice in aiding the microneedle to penetrate painlessly. Taking these findings into account, a more realistic biomimetic mosquito-like microneedle can be developed.

Future Directions

Current research in this area includes constructing carbon nanotube (CNT) based reinforced polymeric microneedles with high axial stiffness along with a lateral support system. Research to construct a synthetic mosquito fascicle based on CNT yarns developed by Zhao et al. [26] is in progress. As depicted in Fig. 7 [26], CNT yarns are spun from CNT arrays. Yarns with diameters of around 40 µm coated with polymers increases the axial stiffness. The coated yarns will be supported laterally and tested for buckling behavior while penetrating porcine skin during the application of an axial force. Successful construction of a polymeric microneedle with a lateral support system should be completed in the near future.

Conclusion

A mosquito is a sophisticated biological machinery that drives a flexible microneedle painlessly into the skin and draw blood successfully. Female mosquitoes of the type *Aedes aegypti* have offered scientists an excellent peek into how a nature's organism is involved in human skin penetration and blood extraction. Microneedle designs inspired by the feeding morphology and mechanics of female mosquito offer an excellent tool for researchers and medical practitioners. Fascicles of female mosquitoes of the type *Aedes aegypti* are one of nature's pliant polymeric microneedles that penetrate human skin effectively to draw blood painlessly by remaining within the bounds of epidermis and dermis, suitably maneuvring its anatomy to avoid fascicle breakage. Significant progress has been made to develop microneedles based on mosquito morphology. Further progress is possible when the dynamics and the importance of lateral support are incorporated in the design.

Cross-References

- ▶ Biomimicry
- ▶ Bionics
- ▶ Nanomechanics
- ▶ Nanomedicine

References

1. Khumpuang, S., Horde, M., Fujioko, K., Sugiyama, S.: Geometrical strengthening and tip-sharpening of a microneedle array fabricated by X-ray lithography. Microsyst. Technol. **13**, 209–214 (2007)
2. Roxhed, N., Gasser, T.C., Griss, P., Holzapfel, G.A., Stemme, G.: Penetration-enhanced ultrasharp microneedles and prediction on skin interaction for efficient transdermal drug delivery. J. Electromech. Syst. **16**(6), 1429–1440 (2007)
3. Tsuchiya, K., Isobata, K., Sato, M., Uetsuji, Y., Nakamachi, E., Kajiwara, K., Kimura, M.: Design of painless microneedle for blood extraction system. Proc. SPIE **67990Q**, 1–10 (2007)
4. Tsuchiya, K., Nakanishi, N., Uetsuji, Y., Nakamachi, E.: Development of blood extraction system for health monitoring system. Biomed. Microdevices **7**(4), 347–353 (2005)
5. Oki, A., Takai, M., Ogawa, H., Takamura, Y., Fukasawa, T., Kikuchi, J., Ito, Y., Ichiki, T., Horiike, Y.: Healthcare chip for checking health condition from analysis of trace blood collected by painless needle. J. Appl. Phys. **42**, 3722–3727 (2003)

6. Oka, K., Aoyagi, S., Arai, Y., Isono, Y., Hashiguchi, G., Fujita, H.: Fabrication of a micro needle for trace blood test. Sens. Actuators A **97–98**, 478–485 (2002)
7. Davis, P.S., Landis, B.J., Adams, Z.H., Allen, M.G., Prausnitz, M.R.: Insertion of microneedles into skin: measurement and prediction of insertion and needle fracture force. J. Biomech. **37**, 1155–1163 (2004)
8. Chaudhri, B.P., Ceyssens, F., Moor, P.D., Hoof, C.V., Puers, R.: A high aspect ratio SU-8 fabrication technique for hollow microneedles for transdermal drug delivery and blood extraction. J. Micromech. Microeng. **20**(6), 064006 (2010)
9. Meyers, M.A., Lin, A.Y.M., Lin, Y.S., Olevsky, E.A., Georgalis, S.: The cutting edge: sharp biological materials. J. Miner. Met. Mater. Soc. **60**(3), 19–24 (2008)
10. Daniel, T.L., Kingsolver, J.G.: Feeding strategy and the mechanics of blood sucking in insects. J. Theor. Biol. **105**, 661–672 (1983)
11. Harrison, J.J.: Wikipedia, http://www.wikipedia.com, (2009)
12. Ramasubramanian, M.K., Barham, O.M., Swaminathan, V.: Mechanics of a mosquito bite with applications to microneedle design. Bioinsp. Biomim. **3**, 046001 (2008) (10 pp)
13. Xin, Q.I., Ying-Chun, Q.I., Yan, L.I., Qian, C.: The insect fascicle morphology research and bionic needle pierced mechanical mechanism analysis. Adv. Nat. Sci. **3**(2), 251–257 (2010)
14. Chakraborty, S., Tsuchiya, K.: Development and fluidic simulation of microneedles for painless pathological interfacing with living systems. J. Appl. Phys. **103**(11), 114701-1–114701-9 (2008)
15. Kosoglu, M.A., Hood, R.L., Chen, Y., Xu, Y., Rylander, M.N., Rylander, C.G.: Fiber optic microneedles for transdermal light delivery: ex vivo porcine skin penetration experiments. J. Biomech. Eng. **132**(9), 091014 (2010)
16. Kong, X.Q., Wu, C.W.: Mosquito proboscis: an elegant biomicroelectromechanical system. Am. Phys. Soc. E **82**, 011910 (2010)
17. Kong, X.Q., Wu, C.W.: Measurement and prediction of insertion force for the mosquito fascicle penetrating into human skin. J. Bionic. Eng. **6**(2), 143–152 (2009)
18. Aoyagi, S., Izumi, H., Isono, Y., Fukuda, M., Ogawa, H.: Laser fabrication of high aspect ratio thin holes on biodegradable polymer and its application to a microneedle. Sens. Actuators **139**, 293–302 (2007)
19. Aoyagi, S., Izumi, H., Fukuda, M.: Biodegradable polymer needle with various tip angles and consideration on insertion mechanism of mosquito's proboscis. Sens. Actuators A **143**, 20–28 (2007)
20. Chichkov, B.: Two-photon polymerization enhances rapid prototyping of medical devices. Int. Soc. Opt. Eng. **7** (2007)
21. McAllister, D.V., Allen, M.G., Prausnitz, M.R.: Microfabricated microneedles for gene and drug delivery. Adv. Drug Deliv. Rev. **56**, 581–587 (2004)
22. Suzuki, H., Tokuda, T., Miyagishi, T., Yoshida, H., Honda, N.: A disposable on-line microsystem for continuous sampling and monitoring of glucose. Sens. Actuators B **97**, 90–97 (2004)
23. Yoshida, Y., Takei, T.: Fabrication of a microneedle using human hair. Jpn. J. Appl. Phys. **48**, 098007 (2009). 2 pp
24. Prow, T.W., Chen, X., Crichton, M., Tiwari, Y., Gradassi, F., Raphelli, K., Mahony, D., Fernando, G., Roberts, M.S., Kendall, M.A.F.: Targeted epidermal delivery of vaccines from coated micro-nanoprojection patches. Nanosci. Nanotechnol. **10287125**, 125–128 (2008)
25. Izumi, H., Aoyagi, S.: Novel fabrication method for long silicon microneedles with three-dimensional sharp tips and complicated shank shapes by isotropic dry etching. Trans. Electr. Electron. Eng. **2**, 328–334 (2006)
26. Zhao, H., Zhang, Y., Bradford, P.D., Zhou, Q., Jia, Q., Yuan, F.G., Zhu, Y.: Carbon nanotube yarn strain sensors. Nanotechnology **21**(5), 305502 (2010)

Biomimetic Muscles and Actuators

▶ Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP)

Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP)

Yoseph Bar-Cohen
Jet Propulsion Laboratory (JPL), California Institute of Technology, Pasadena, CA, USA

Synonyms

Artificial muscles; Biomimetic muscles and actuators; Electroresponsive polymers

Definition

Electroactive polymers (EAP) are also known as artificial muscles. They have the closest functional response similarity to biological muscles and therefore they gained the name artificial muscles. As such they are being considered for use to

mimic the mechanical performance of biological systems. Actuation via the use of electrical energy is an attractive activation method for causing elastic deformation in polymers and it offers great convenience and practicality.

Introduction

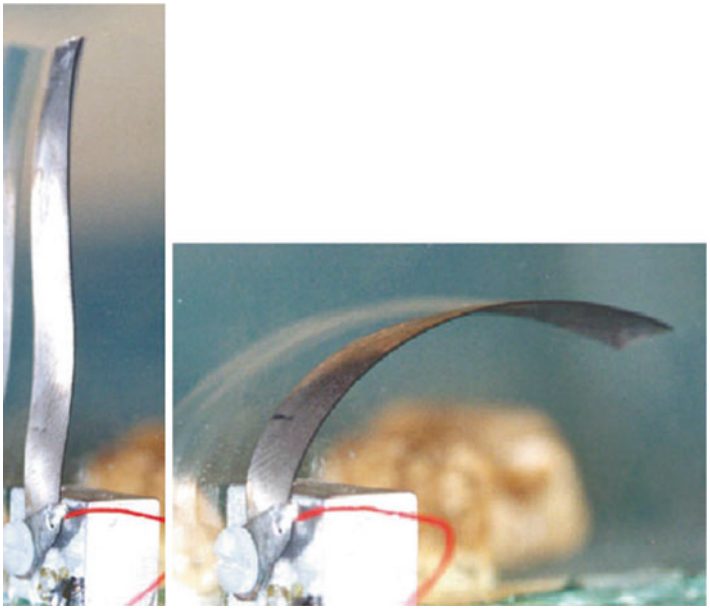
The ability to copy, adapt, and/or be inspired by nature is critically dependant on the availability of effective capabilities and tools. In most biological systems that are larger than bacteria, muscles are their actuators and they are able to lift loads as large as hundreds of kilograms with response time of milliseconds. Biological muscles are driven by a complex mechanism that is very difficult to mimic. Electroactive polymers (EAP) offer the closest functional similarity to biological muscles and therefore they gained the name artificial muscles. There are many polymers that vary their shape or size when subjected to electric, chemical, pneumatic, optical, or magnetic stimulation. However, electrical excitation is one of the most attractive activation methods for causing elastic deformation in polymers. The convenience and practicality of the electrical stimulation and the recent response improvements have made EAP materials the most preferred among the responsive polymers [1].

Polymers that respond to electrical stimulation were known for over 100 years; however, they received little attention since they responded with quite a small strain at the level of fraction of a percent. In the last 20 years, a series of EAP materials have emerged that respond with a very significant shape change. The impressive advances in improving their performance are attracting the attention of engineers and scientists from many different disciplines. Using the developed materials as actuators, many novel mechanisms and devices were already demonstrated including robot fish, fish-shaped blimp propelled by wagging its body and tail, refreshable braille displays, haptic interfaces, miniature gripper, catheter steering element, artificial eye lid, and dust-wiper.

The Available EAP Materials

The first milestone in developing EAP materials was marked by Roentgen [2] when he subjected to an electric field a rubber-band with fixed end and a mass attached to the free end. Other important steps in the development of these materials include the discoveries of the electrets by Eguchi [3] and the Polyvinylidene Fluoride (PVDF) by Kawai [4]. However, up until the 1990s, the generated strain by the known EAP materials was at the level of tenth of a percent [1]. Most of the EAP materials known today that generate large strain have emerged around 1990. One of the notable ones is the polyacrylate dielectric elastomer (VHB^{TM} 4910 from 3M^{TM}) that was reported to reach strain levels of about 380 % with a relatively short response time ($<$0.1 s) [5]. To simplify the categorization of the many EAP material types, the author divided them into two groups according to their characteristics, including the *ionic* and *electronic* (also known as the *field activated*) [1].

- *Ionic EAP*: As actuators, they consist of a polymer film with electrolyte and it is coated by two electrodes. Their activation mechanism involves mobility of ions under electrical excitation [6]. These polymers have the advantages of generating large bending displacement under activation by a relatively low voltage (1–2 V). Their disadvantages are the need to maintain electrolytes wetness, low energy conversion efficiency ($\sim$1 %), and difficulties sustaining constant displacement under activation of a DC voltage (except for conducting polymers). Since the strain in conducting polymers is proportional to charge and since they behave electrically as capacitors, they maintain a constant position under constant voltage [7]. Moreover, due to the related macroscopic motion of ions that is involved with their activation, these polymers respond relatively slow in the range of tens to a fraction of a second. Examples of ionic EAP materials include the ionomeric polymer-metal composites (IPMC), conducting polymers, carbon nanotubes, and ionic polymer gels.



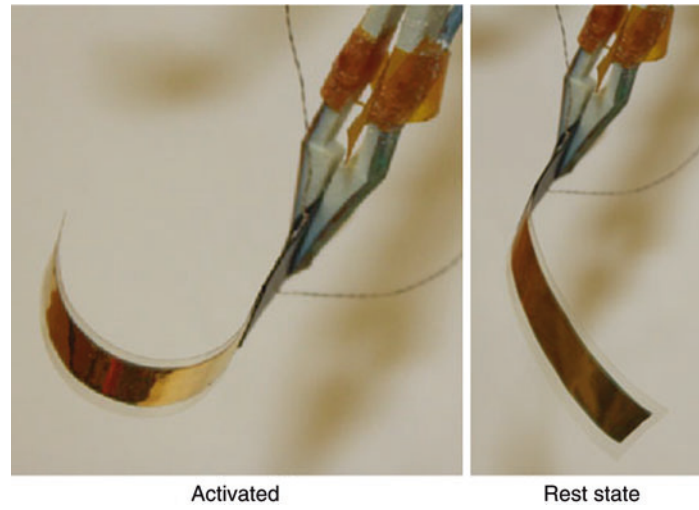
Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP), Fig. 1 IPMC in passive (*left*) and activated states (*right*)

Among the materials in this group, IPMC is one of the most studied EAP materials. The base polymer provides channels for mobility of positive ions in a fixed network of negative ions on interconnected clusters [6, 8]. In 1992, three different groups of researchers have reported independently the development of IPMC-based EAP, and they include: Oguro et al. [9] in Japan, as well as Shahinpoor [10] and Sadeghipour et al. [11] in the United States. The attractive characteristic of IPMC is the significant bending in response to a relatively low electrical voltage at the level of 1 V (Fig. 1). Generally, because of the fact that ions need to physically travel though the material, the response of IPMC is relatively slow (<10 Hz).

- *Electronic (also known as field-activated) EAP*: These materials are activated by Coulomb force that causes dimension change either due to direct thickness reduction (e.g., dielectric elastomers) or due to internal polarization with molecular alignment (e.g., ferroelectric EAP [12]). Generally, high electric field (>10-V/μm) is required, which may be close to the electric breakdown level that is the result of the low dielectric constant. To reduce the required high voltage two alternative approaches are used: (a) creating a stack of thin multilayers and (b) forming a composite using filler material with a high dielectric constant. Photographs of such a composite ferroelectric EAP in passive and activated states are shown in Fig. 2. The EAP material with the filler requires about an order of magnitude less voltage for its activation.

Since the materials in this group of EAP are activated by the electric field, the response is quite fast and can reach milliseconds. The advantages of the materials in this group include holding the generated displacement under a DC voltage making them attractive for robotic applications. Also, these EAP materials have a greater mechanical energy density than the ionic EAP and they can be activated in air with no major constraints. Their main disadvantage is the required high activation field that can be close to the electric breakdown level. Examples of these materials are dielectric elastomers, piezoelectric, and ferroelectric.

Currently, the dielectric elastomer EAP is considered to have the highest technology readiness with the least issues associated with the understanding of its operation. Generally,



Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP), Fig. 2 Photographs of a composite ferroelectric EAP in passive (*right*) and activated states (*left*). This EAP material was provided to the author as a courtesy of Qiming Zhang, Penn State University

this EAP type is based on the use of polymers with low elastic stiffness and high dielectric breakdown strength that are subjected to an electrostatic field and they generate a large actuation strain. An actuator that is made of dielectric elastomer EAP can be represented by a parallel plate capacitor [5]. The electrodes have to be highly compliant in order to avoid impeding the generated strain and conductive carbon grease that are widely used for this purpose. The grease provides the electrodes with the required resilience and the carbon provides the required high electrical conductivity. The required very high flexibility of the conductive carbon grease and high electrical conductivity make it best fit for the task. Application of an electric field results in a strain that is proportional to the square of the electric field and to the dielectric constant while inversely proportional to the elastic modulus. The first observation of the fact that dielectric elastomers sustain large strain (23 % in silicone films) when subjected to high electric field was reported in a 1992–1993 study by Pelrine and his coinvestigators [13, 14].

Since EAP materials are still relatively new it is necessary to establish the related scientific and engineering foundations, including improvement of the understanding of the basic principles that drive them [1]. These require developing effective computational chemistry models, comprehensive material science, electromechanical analytical tools, and material processing techniques. In order to maximize the actuation capability and operation durability, effective processing techniques are being developed for their fabrication, shaping, and electroding. Methods of reliably characterizing the response of EAP materials are being developed and efforts are underway to create databases with documented material properties (http://www.actuatorweb.org/). Bringing these materials to the level of making daily used products will necessitate finding a niche application that addresses critical needs.

Applications of EAP

Significant efforts are underway to develop practical EAP actuators and commercial products are starting to emerge. Eamax, Japan, is credited for developing the first EAP-related commercial product, and at the end of 2002 it announced the marketing of a fish robot that is actuated by IPMC [1]. Mechanisms and devices applicable to many fields are being considered or developed including zoom lens of cellular phones, valves, energy harvesting, pumps, and many others. EAP actuators offer many important capabilities for devices with biomimetic characteristics [1, 15]. One may produce such devices as artificial bugs that may



Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP), Fig. 3 The icon of the grand challenge for the development of EAP-actuated robotics

walk, swim, hop, crawl, and dig while reconfiguring themselves as needed. Mimicking nature would immensely expand the collection and functionality of robots allowing performance of tasks that are impossible with existing capabilities [15–17]. For example, the author and his coinvestigators constructed a miniature gripper with four IPMC-based fingers as bending actuators [1]. This gripper was demonstrated to grab rocks very similar to the human hand using hooks at the bottom emulating fingernails. Other recently reported robotic applications include a blimp with steering fins as well as a fish-shaped blimp propelled by wagging its body and tail (developed by EMPA, Switzerland).

Space applications are among the most demanding in terms of the harshness of the operating conditions (extreme temperatures, high pressure, or vacuum) and they require very high reliability and durability. Today's available materials are not applicable to handle the required actuators that need to operate down to as low as −200 °C as on Europa and Titan or +460 °C as on Venus. Another challenge is the need for large-scale EAP in the form of films, fibers, and others. The required dimensions can be as large as several meters or kilometers to produce large gossamer structures such as antennas, solar sails, and various large optical components. Making biomimetic capability using EAP material will potentially allow NASA and other space agencies to conduct missions in other planets using humanlike robots that emulate human operation before sending real humans.

Addressing the Challenges to EAP

In an effort to promote worldwide development toward the realization of the potential of EAP materials, the author posed in 1999 an arm-wrestling challenge (http://ndeaa.jpl.nasa.gov/nasa-nde/lommas/eap/EAP-armwrestling.htm). The challenge is to have an EAP-activated robotic arm win against a human in a wrestling match. The icon of this challenge is illustrated graphically in Fig. 3. The emphasis on arm wrestling with a human was chosen in order to use the intuitive comparison of our muscles as a baseline for gauging the advances in the capability of EAP actuators. Success will allow applying EAP materials to improve many aspects of our life including the development of products with unmatched capabilities and dexterity. The first arm-wrestling match was held in San Diego, California, USA, on March 7, 2005, as part of the EAP-in-Action Session of the SPIE's EAPAD Conference. Three robotic arms participated in the contest against a 17-year-old high school female student and the student won against all these arms (see Fig. 4).

The second Artificial Muscles Armwrestling Contest was held on February 27, 2006. Rather than wrestling with a human opponent the performance of the arms was measured and compared to the performance of the student (http://eap.jpl.nasa.gov). The measuring fixture (developed jointly by individuals from UCLA and the author's group at JPL) was strapped to the contest



Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP), Fig. 4 An EAP-driven arm made by students from Virginia Tech and the human opponent, 17-year-old student

table and the EAP-actuated arms were tested for speed and pulling force. Each competing arm pulled on the fixture cable that has a force gauge on its other end. To simulate wrestling action, the pulling on the cable lifted a 0.5-kg weight and the speed of lifting was measured from the travel time of the weight from the bottom to the top section of the fixture. To establish a baseline for performance comparison, the capability of the student (the same student who wrestled in 2005) was measured first and then the three participating robotic arms were tested. The results have shown two orders of magnitude lower performance of the arms compared to the student.

Challenges, Trend, and Potential Development

Since the early 1990s, new EAP materials have been developed that generate large strains making them highly attractive for biomimetic and many other applications. Their operational similarity to biological muscles, including resilience, damage tolerance, and ability to induce large actuation strains (stretching, contracting, or bending) makes them unique compared to other electroactive materials. Researchers are increasingly making improvements in the various related areas including a better understanding of the operation mechanism of the various EAP material types. The processes of synthesizing, fabricating, electroding, shaping, and handling are being refined to maximize the actuation capability and durability. Methods of reliably characterizing the response of these materials are being developed and efforts are being made to establish databases with documented material properties to support engineers that are considering the use of these materials. An initiative in this area has been taken by the University of British Columbia and they formed a Web database for viewing, comparing, and submitting EAP properties [18].

Applying EAP materials as actuators of manipulation, mobility, and robotic devices involves multidiscipline including materials, chemistry, electromechanics, computers, and electronics. Even though the actuation force of the existing materials requires further improvement, there has already been a series of reported successes in the development of mechanisms that are driven by EAP actuators. However, seeing EAP replace existing actuators in commercial devices and engineering mechanisms may be a difficult challenge and would require identifying niche applications where EAP materials would not need to compete with existing technologies. It is quite encouraging to see the growing number of researchers and engineers who are pursuing careers in EAP-related disciplines. Hopefully, the growth in the research and development activity will lead to making these materials becoming the actuators of choice.

Acknowledgment Some of the research reported in this entry was conducted at the Jet Propulsion Laboratory (JPL), California Institute of Technology, under a contract with National Aeronautics and Space Administration (NASA).

Cross-References

► Bioinspired Synthesis of Nanomaterials
► Biomimetic Flow Sensors
► Biomimetic Mosquito-like Microneedles

▶ Biomimetic Synthesis of Nanomaterials
▶ Biomimetics
▶ Biomimetics of Marine Adhesives
▶ Biosensors

References

1. Bar-Cohen, Y.: Electroactive polymer (EAP) Actuators as Artificial Muscles – Reality, Potential and Challenges, vol. PM136, 2nd edn, pp. 1–765. SPIE Press, Bellingham (2004). ISBN 0-8194-5297-1
2. Roentgen, W.C.: About the changes in shape and volume of dielectrics caused by electricity, section III. In: Wiedemann, G. (ed.) Annual Physics and Chemistry Series, vol. 11, pp. 771–786. John Ambrosius Barth Publisher, Leipzig (1880). In German
3. Eguchi, M.: On the permanent electret. Philos. Mag. **4**(9), 178 (1925)
4. Kawai, H.: Piezoelectricity of poly(viny1idene fluoride). Jpn. J. Appl. Phys. **8**, 975–976 (1969)
5. Pelrine, R., Kornbluh, R., Pei, Q., Joseph, J.: High-speed electrically actuated elastomers with strain greater than 100%. Science **287**(5454), 836–839 (2000)
6. Park, I.S., Jung, K., Kim, D.S.M., Kim, K.J.: Physical principles of ionic polymer-metal composites as electroactive actuators and sensors, special issue dedicated to EAP. Mater. Res. Soc. MRS Bull. **33**(3), 190–195 (2008)
7. Madden, J.D.W., Madden, P.G., Hunter, I.W.: Conducting polymer actuators as engineering materials. In: Bar-Cohen, Y. (ed.) Proceedings of the SPIE Smart Structures and Materials 2002: Electroactive Polymer Actuators and Devices (EAPAD), pp. 176–190. SPIE Press, Bellingham (2002). doi:10.1117/12.475163
8. Nemat-Nasser, S., Thomas, C.W.: Ionomeric polymer-metal composites, Chapter 6. In: Bar-Cohen, Y. (ed.) Electroative Polymer (EAP) Actuators as Artificial Muscles – Reality, Potential and Challenges, vol. PM136, 2nd edn, pp. 171–230. SPIE Press, Bellingham (2004). ISBN 0-8194-5297-1
9. Oguro, K., Kawami, Y., Takenaka, H.: Bending of an ion-conducting polymer film-electrode composite by an electric stimulus at low voltage. Trans. J. Micromachine Soc. **5**, 27–30 (1992)
10. Shahinpoor, M.: Conceptual design, kinematics and dynamics of swimming robotic structures using ionic polymeric gel muscles. Smart Mater. Struct. **1**(1), 91–94 (1992)
11. Sadeghipour, K., Salomon, R., Neogi, S.: Development of a novel electrochemically active membrane and 'smart' material based vibration sensor/damper. J. Smart Mater. Struct. **1**(1), 172–179 (1992)
12. Cheng, Z., Zhang, Q.: Field-activated electroactive polymers, special issue dedicated to EAP. Mater. Res. Soc. MRS Bull. **33**(3), 190–195 (2008)
13. Pelrine, R., Joseph, J.: FY 1992 Final report on artificial muscles for small robots, ITAD-3393-FR-93-063, SRI International, Menlo Park, California, submitted to Micro Machine Center, MBR99 Bldg. 6 F 67, Kanda-sakumagashi, Chiyoda-ku, Tokyo, 101–0026 Japan (1993)
14. Pelrine, R., Joseph, J.: FY 1993 final report on artificial muscles for small robots, ITAD-4570-FR-94-076, SRI International, Menlo Park, California, submitted to Micro Machine Center, MBR99 Bldg. 6 F 67, Kanda-sakumagashi, Chiyoda-ku, Tokyo, 101–0026 Japan (1994)
15. Bar-Cohen, Y. (ed.): Biomimetics – Biologically Inspired Technologies, pp. 1–527. CRC Press, Boca Raton (2005). ISBN 0849331633
16. Bar-Cohen, Y., Breazeal, C. (eds.): Biologically-Inspired Intelligent Robots, vol. PM122, pp. 1–393. SPIE Press, Bellingham (2003). ISBN 0-8194-4872-9
17. Bar-Cohen, Y., Hanson, D.: The Coming Robot Revolution – Expectations and Fears About Emerging Intelligent, Humanlike Machines. Springer, New York (2009). ISBN 978-0-387-85348-2
18. Madden, J.D.W.: Actuator selection tool. http://www.actuatorweb.org/. Accessed 12 Jan 2010

Biomimetic Synthesis

▶ Bioinspired Synthesis of Nanomaterials

Biomimetic Synthesis of Nanomaterials

▶ Bioinspired Synthesis of Nanomaterials

Biomimetics

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Biognosis; Biomimicry; Bionics; Lesson from nature; Nanotechnology

Definition

The emerging field of biomimetics allows one to mimic biology or nature to develop nanomaterials, nanodevices, and processes. Properties of biological materials and surfaces result from a complex interplay between surface morphology and physical and chemical properties. Hierarchical structures with dimensions of features ranging from macroscale to the nanoscale are extremely common in nature to provide properties of interest. Molecular scale devices, superhydrophobicity, self-cleaning, drag reduction in fluid flow, energy conversion and conservation, high adhesion, reversible adhesion, aerodynamic lift, materials and fibers with high mechanical strength, biological self-assembly, anti-reflection, structural coloration, thermal insulation, self-healing, and sensory aid mechanisms are some of the examples found in nature which are of commercial interest.

Introduction

Nature has gone through evolution over the 3.8 billion years, since life is estimated to have appeared on earth [1]. Nature has evolved objects with high performance using commonly found materials. These function on the macroscale to nanoscale. The understanding of the functions provided by objects and processes found in nature can guide researchers to imitate and produce nanomaterials, nanodevices, and processes. Biologically inspired design or adaptation or derivation from nature is referred to as "biomimetics." It means mimicking biology or nature. Biomimetics is derived from the Greek word biomimesis. The word was coined by polymath Otto Schmitt in 1957, who in his doctoral research, developed a physical device that mimicked the electrical action of a nerve. Other words used include bionics (coined in 1960 by Jack Steele of Wright-Patterson Air Force Base in Dayton, Ohio), biomimicry, and biognosis. The field of biomimetics is highly interdisciplinary. It involves the understanding of biological functions, structures, and principles of various objects found in nature by biologists, physicists, chemists, and material scientists and the design and fabrication of various materials and devices of commercial interest by engineers, material scientists, chemists, and others. The word biomimetics first appeared in Webster's dictionary in 1974 and is defined as "the study of the formation, structure or function of biologically produced substances and materials (as enzymes or silk) and biological mechanisms and processes (as protein synthesis or photosynthesis) especially for the purpose of synthesizing similar products by artificial mechanisms which mimic natural ones."

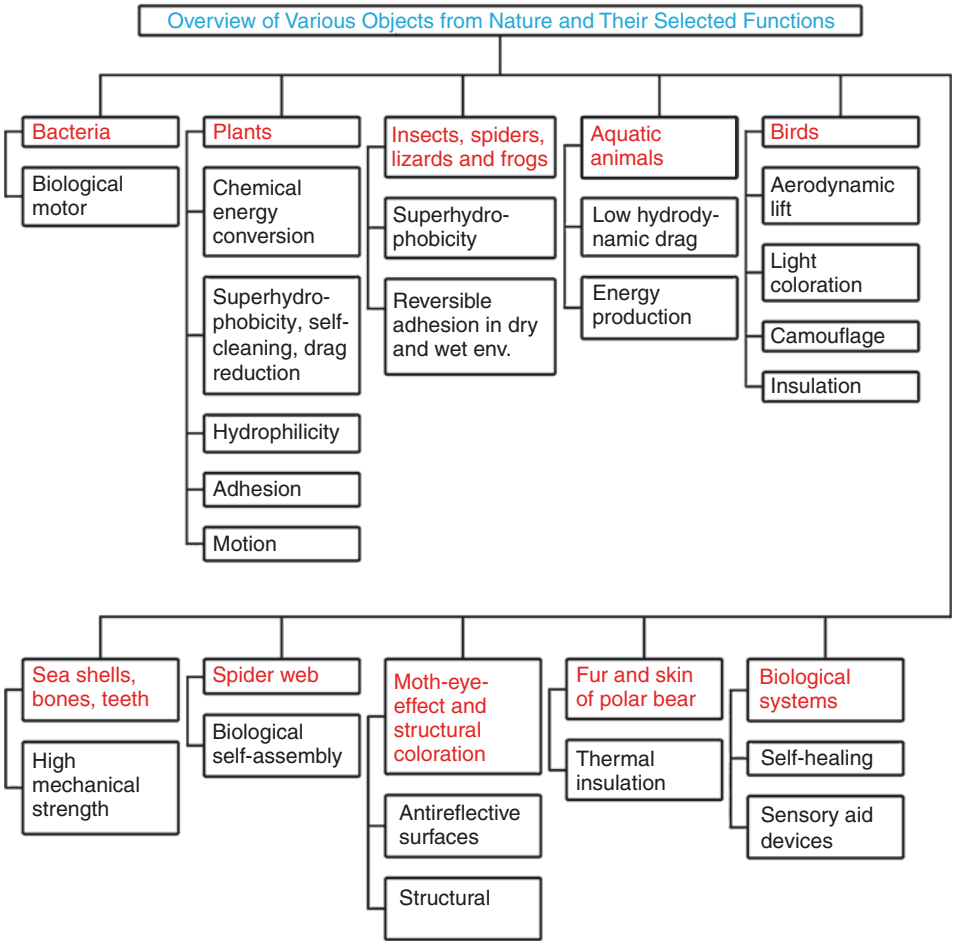
Biological materials are highly organized from the molecular to the nano-, micro-, and macroscales, often in a hierarchical manner with intricate nanoarchitecture that ultimately makes up a myriad of different functional elements [2]. Nature uses commonly found materials. Properties of materials and surfaces result from a complex interplay between surface structure and morphology and physical and chemical properties. Many materials, surfaces, and devices provide multifunctionality. Molecular scale devices, superhydrophobicity, self-cleaning, drag reduction in fluid flow, energy conversion and conservation, reversible adhesion, aerodynamic lift, materials and fibers with high mechanical strength, biological self-assembly, anti-reflection, structural coloration, thermal insulation, self-healing, and sensory aid mechanisms are some of the examples found in nature which are of commercial interest.

Various features found in nature objects are on the nanoscale. The major emphasis on nanoscience and nanotechnology since early 1990s has provided a significant impetus in mimicking nature using nanofabrication techniques for commercial applications [3]. Biomimetics has spurred interest across many disciplines.

It is estimated that the 100 largest biomimetic products had generated about US $1.5 billion over 2005–2008. The annual sales are expected to continue to increase dramatically.

Lessons from Nature and Applications

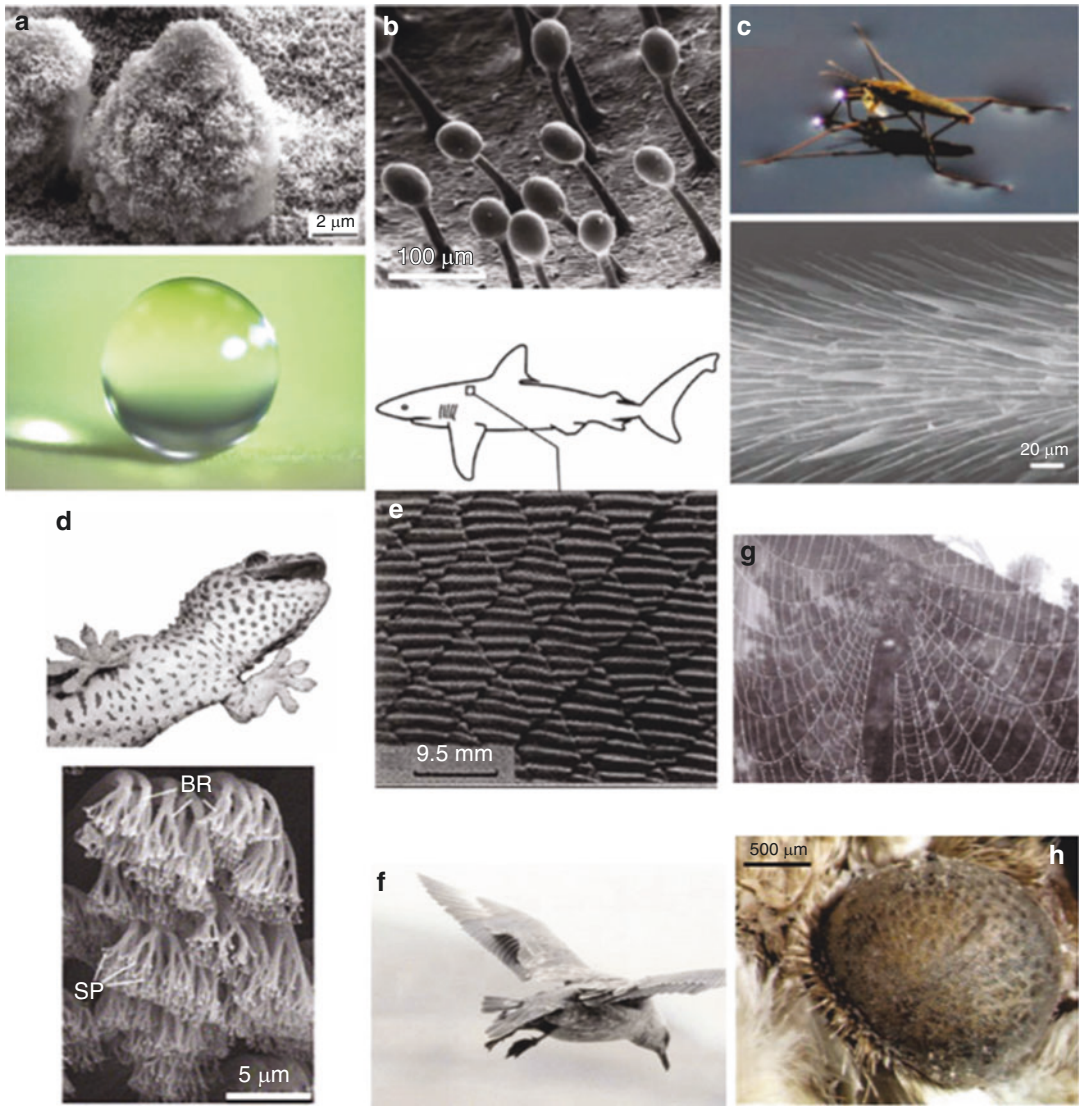
There are a large number of objects, including bacteria, plants, land and aquatic animals, and



Biomimetics, Fig. 1 An overview of various objects from nature and their selected functions [4]

sea shells, with properties of commercial interest. Figure 1 provides an overview of various objects from nature and their selected functions. Figure 2 shows a montage of some examples from nature. These serve as the inspiration for various technological developments [4].

Some leaves of water repellent plants such as *Nelumbo nucifera* (Lotus) and *Colocasia esculenta* are known to be superhydrophobic and self-cleaning due to hierarchical roughness (microbumps superimposed with nanostructure) and the presence of a hydrophobic coating [12–18]. Roughness-induced superhydrophobic and self-cleaning surfaces are of interest in various applications, including self-cleaning windows, windshields, and exterior paints for buildings, boats, ships, and aircrafts, utensils, roof tiles, textiles, solar panels, and applications requiring antifouling and a reduction of drag in fluid flow, for example, in micro/nanofluidics, boats, ships, and aircrafts. Superhydrophobic surfaces can also be used for energy conversion and conservation. Nonwetting surfaces also reduce stiction at a contacting interface in machinery [19, 20]. A model surface for superhydrophobicity and self-cleaning is provided by the leaves of the Lotus plant (*Nelumbo nucifera*) (Fig. 2a) [13–15, 21–24]. The so-called papillose epidermal cells form asperities or papillae and provide roughness on the microscale. The surface of the leaves is usually covered with a range of waxes made from a mixture of



Biomimetics, Fig. 2 Montage of some examples from nature. (**a**) Lotus effect [5], (**b**) glands of carnivorous plant secrete adhesive to trap insects [6], (**c**) pond skater walking on water [7], (**d**) gecko foot exhibiting reversible adhesion [8], (**e**) scale structure of shark reducing drag [9], (**f**) wings of a bird in landing approach, (**g**) spiderweb made of silk material [10], and (**h**) antireflective moth's eye [11]

long-chain hydrocarbon compounds that have a strong phobia of being wet. Sub-micron sized asperities composed of the 3-D epicuticular waxes are superimposed over microscale roughness, creating a hierarchical structure. The wax asperities consist of different morphologies, like tubules on Lotus or platelets on *Colocasia* [13–15]. The water droplets on these surfaces readily sit on the apex of nanostructures because air bubbles fill in the valleys of the structure under the droplet. Therefore, these leaves exhibit considerable superhydrophobicity. The water droplets on the leaves remove any contaminant particles from their surfaces when they roll off, leading to self-cleaning. A contact angle of 164° and a contact angle hysteresis of 3° have been reported for the Lotus leaf [5, 6].

Two strategies used for catching insects by plants for digestion are having sticky surfaces or sliding structures. As an example, for catching insects by using sticky surfaces, the glands of the carnivorous plants of the genus *Pinguicula* (butter-worts) and *Drosera* (sundew), shown in Fig. 2b, secrete adhesives and enzymes to trap and digest small insects like mosquitoes and fruit flies [14]. In *Pinguicula*, the stalked glands are the sticky ones which trap the insects by secreting an adhesive solution. The shorter ones are those which secrete digestive enzymes, including protease and phosphotase, and later resorb the digested material. In *Drosera*, the stalked glands can effectively enclose small flies by bringing numerous glands in contact with the prey.

Water striders (Gerris remigis) are insects which live on the surfaces of ponds, slow streams, and quiet waters. A water strider has the ability to stand and walk upon water surface without getting wet, Fig. 2c. Even the impact of rain droplets with a size greater than the strider's size does not make it immerse into water. Gao and Jiang [7] showed that the special hierarchical structure of strider legs, which are covered by large numbers of oriented tiny hairs (microsetae) with fine nanogrooves and covered with cuticle wax, makes the leg surfaces superhydrophobic and is responsible for the water resistance and enables them to stand and walk quickly on the water surface. They measured the contact angle of the insect's legs with water to be about 167°. Scanning electron microscope (SEM) micrographs revealed numerous oriented setae on the legs. The setae are needle-shaped hairs with diameters ranging from 3 μm down to several 100 nm. Most setae are roughly 50 μm in length and arranged at an inclined angle of about 20° from the surface of the leg. Many elaborate nanoscale grooves were found on each microseta, and these form a unique hierarchical structure. This hierarchical micro- and nanostructuring on the leg's surface seems to be responsible for its water resistance [16, 17] and the strong supporting force. Gao and Jiang [7] reported that a leg does not pierce the water surface until a dimple of 4.4 mm depth is formed. They found that the maximal supporting force of a single leg is 1.52 mN, or about 15 times the total body weight of the insect. The corresponding volume of water ejected is roughly 300 times that of the leg itself.

Leg attachment pads of several animals, including many insects (e.g., beetles and flies), spiders, and lizards (e.g., geckos) are capable of attaching to a variety of surfaces and are used for locomotion even on vertical walls or across ceilings [25–28]. Biological evolution over a long period of time has led to the optimization of their leg attachment systems. This dynamic attachment ability is referred to as reversible adhesion or smart adhesion [26, 27].

Attachment systems in various creatures such as insects, spiders, and lizards have similar structures. As the size (mass) of the creature increases, the radius of the terminal attachment elements decreases [29, 30]. This allows a greater number of setae to be packed into an area, hence increasing the linear dimension of contact and the adhesion strength. Based on surface energy approach, it has been reported that adhesion force is proportional to a linear dimension of the contact [19, 20]. Therefore, it increases with the division of contacts. The density of the terminal attachment elements strongly increases with increasing body mass. Flies and beetles have the largest attachment pads and the lowest density of terminal attachment elements. Spiders have highly refined attachment elements that cover their legs. Geckos have both the highest body mass and greatest density of terminal elements (spatula). Spiders and geckos can generate high dry adhesion, whereas beetles and flies increase adhesion by secreting liquids at the contacting interface.

Gecko is the largest animal that can produce high (dry) adhesion to support its weight with a high factor of safety. The gecko skin is comprised of a complex hierarchical structure of lamellae, setae, branches, and spatula [25, 26]. As shown in Fig. 2d, the gecko consists of an intricate hierarchy of structures beginning with lamellae, soft ridges that are 1–2 mm in length, that are located on the attachment pads (toes) that compress easily so that contact can be made with rough bumpy surfaces. Tiny curved hairs known as setae extend from the lamellae with a density of approximately 14,000 per square millimeter. These setae are

typically 30–130 μm in length and 5–10 μm in diameter and are composed primarily of β-keratin with some α-keratin component. At the end of each seta, 100–1,000 spatulae with a diameter of 0.1–0.2 μm branch out and form the points of contact with the surface. The tips of the spatula are approximately 0.2–0.3 μm in width, 0.5 μm in length, and 0.01 μm in thickness and get their name from their resemblance to a spatula.

The attachment pads on two feet of the Tokay gecko have an area of about 220 mm^2. About three million setae on their toes can produce a clinging ability of about 20 N (vertical force required to pull a lizard down a nearly vertical (85°) surface) and allow them to climb vertical surfaces at speeds of over 1 m/s with the capability to attach or detach their toes in milliseconds. It should be noted that a three-level hierarchical structure allows adaptability to surfaces with different magnitudes of roughness. The gecko uses peeling action to unstick itself [26, 27].

Replication of the structure of gecko feet would enable the development of a superadhesive polymer tape capable of clean, dry adhesion which is reversible (e.g., [26, 27, 31–33]). (It should be noted that common man-made adhesives such as tape or glue involve the use of wet adhesives that permanently attach two surfaces.) The reusable gecko-inspired adhesives have the potential for use in everyday objects such as tapes, fasteners, toys, and in high technology such as microelectronic and space applications. Replication of the dynamic climbing and peeling ability of geckos could find use in the treads of wall-climbing robots [34].

Many aquatic animals can move in water at high speeds with a low energy input. Drag is a major hindrance to movement. Most shark species move through water with high efficiency and maintain buoyancy. Through its ingenious design, their skin turns out to be an essential aid in this behavior by reducing drag by 5–10 % and auto-cleaning ecto-parasites from their surface [35, 36]. The very small individual tooth-like scales of shark skin, called dermal denticles (little skin teeth), are ribbed with longitudinal grooves (aligned parallel to the local flow direction of the water) which result in water moving very efficiently over their surface. An example of scale structure on the right front of a Galapagos shark (*C. Galapagensis*) is shown in Fig. 2e. The detailed structure varies from one location to another for a given shark. The scales are present over most of the shark's body. These are V-shaped, about 200–500 μm in height, and regularly spaced (100–300 μm) [37].

Due to the relatively high Reynolds number of a swimming shark, the turbulent flow occurs. The skin drag (wall shear stress) is not generally affected by the surface roughness. Longitudinal scales on the surface result in lower wall shear stresses than that on a smooth surface and control boundary layer separation. The longitudinal scales result in water moving more efficiently over their surface than it would were shark scales completely oriented differently. Over smooth surfaces, fast-moving water begins to break up into turbulent vortices, or eddies, in part because the water flowing at the surface of an object moves slower than water flowing further away from the object with so-called low boundary slip. This difference in water speed causes the faster water to get "tripped up" by the adjacent layer of slower water flowing around an object, just as upstream swirls form along riverbanks. The grooves in a shark's scales simultaneously reduce eddy formation in a surprising number of ways: (1) the grooves reinforce the direction of flow by channeling it, (2) they speed up the slower water at the shark's surface (as the same volume of water going through a narrower channel increases in speed), reducing the difference in speed of this surface flow and the water just beyond the shark's surface, (3) conversely, they pull faster water toward the shark's surface so that it mixes with the slower water, reducing this speed differential, and finally, (4) they divide up the sheet of water flowing over the shark's surface so that any turbulence created results in smaller, rather than larger, vortices [38].

It is also reported that longitudinal scales influence the fluid flow in the transverse direction by limiting the degree of momentum transfer. It is the difference in the protrusion height in the longitudinal and transverse directions which govern how much the scales impede the transverse flow.

Bechert et al. [35] reported that, based on their experimental data, thin, vertical scales result in the low transverse flow and low drag. They also reported that the ratio of the scale height to tip-to-tip spacing of 0.5 is the optimum value for low drag.

In addition to reduction of drag, the shark skin surface prevents marine organisms from being able to adhere to ("foul") it. It is not because of the Lotus effect, the shark skin is hydrophilic and wets with water. There are three factors which appear to keep the surface clean: (1) the accelerated water flow at a shark's surface reduces the contact time of fouling organisms, (2) the roughened nanotexture of shark skin reduces the available surface area for adhering organisms, and (3) the dermal scales themselves perpetually realign or flex in response to changes in internal and external pressure as the shark moves through water, creating a "moving target" for fouling organisms [38].

Speedo created the wholebody swimsuit called Fastskin bodysuit (TYR Trace Rise) in 2006 for elite swimming. The suit is made of polyurethane woven fabric with a texture based on shark scales. In the 2008 Summer Olympics, two thirds of the swimmers wore Speedo swim suits, and a large number of world records were broken. Boat, ship, and aircraft manufacturers are trying to mimic shark skin to reduce friction drag and minimize the attachment of organisms on their bodies. One can create riblets on the surface by painting or attaching a film (3 M). Skin friction contributes about half of the total drag between a solid surface and air in an aircraft. The transparent sheets with ribbed structure in the longitudinal direction have been used on the commercial Airbus 340 aircrafts. It is expected that riblet film on the body of the aircraft can reduce drag on the order of 10 % [39].

Mucus on the skin of aquatic animals, including sharks, acts as an osmotic barrier against the salinity of seawater and protects the creature from parasites and infections [36]. The mucus also operates as a drag-reducing agent on some fast predatory fishes, which allows the fish to attack more easily [40]. The artificial derivatives of fish mucus, that is, polymer additives for liquids, are used in drag reduction technology, for example, to propel crude oil in the Alaska pipeline [41].

Bird feathers perform multiple functions – make the body water repellant, create wing and tail for aerodynamic lift during flying, provide coloration for appearance as well as camouflage, and provide an insulating layer to keep the body warm. Many bird feathers exhibit hydrophobicity (an apparent contact angle of 100–140°). Bormashenko et al. [42] studied Feral Rock pigeon feathers. The morphology consists of a network formed by barbs and barbules made of keratin. It is the morphology which plays an important role in hydrophobicity. Birds consist of several consecutive rows of covering feathers on their wings which are flexible, Fig. 2f. These movable flaps develop the lift. When a bird lands, a few feathers are deployed in front of the leading edges of the wings which help to reduce the drag on the wings. Self-activated movable flaps (artificial bird feathers) have been shown to provide an increase in lift in flight experiments [36]. Birds serve as the inspiration for aircrafts and early development of wing design [43]. However, aircrafts do not flap their wings like birds to simultaneously produce lift and thrust. This is impractical in aircrafts due to limitations of scaling phenomena and the high speeds.

Spiders produce a variety of proteins, among which are major ampullate silks (MAS). MAS fibers are used by spiders as a scaffold upon which they attach other silk fibers during the formation of the web. The spider generates the silk fiber, and at the same time it is hanging on it. It has a sufficient supply of raw material for its silk to span great distances [10, 44]. Spider web is a structure built of a one-dimensional fiber, Fig. 2g. The fiber is very strong and continuous and is insoluble in water. The web can hold a significant amount of water droplets, and it is resistant to rain, wind, and sunlight [10, 45]. Spider silk is three times stronger than steel, having a tensile strength of about 1.2 GPa [46]. Some spider silks have high stiffness with a tensile modulus of about 10 GPa, while others are elastomeric with a modulus of about 1 GPa and extension to rupture of 200 %. The combination of strength and extensibility is primarily derived from the domains of crystalline β-sheets and flexible helixes within the polypeptide chain, imparting a

toughness that is greater than bone, Kevlar, and high strength steel. The web is designed to catch insects (food for the spider) that cross the net and get stuck to its stickiness and complex structure.

Optical reflection and anti-reflection is achieved in nature by using nanoscale architecture [47]. The eyes of moths are antireflective to visible light. The so-called moth-eye effect was discovered in the 1960s as a result of the study of insect eyes. For nocturnal insects, it is important not to reflect the light, since the reflection makes the insect vulnerable to predators. The eyes of moths consist of hundreds of hexagonally organized nanoscopic pillars, each about 200 nm in diameter and height, that result in a very low reflectance for visible light, Fig. 2h [11, 48]. These nanostructures' optical surfaces make the eye surface nearly antireflective in any direction.

Light reflection is avoided by a continuously increasing refractive index of the optical medium. The little protuberances upon the cornea surface increase the refractive index. These protuberances are very small microtrichia (about 200 nm in diameter). For an increase in transmission and reduced reflection, a continuous matching of the refraction index at the boundary of the adjacent materials (cornea and air) is required. If the periodicity of the surface pattern is smaller than the light wavelength, the light is not reflected [49]. If this condition is satisfied, it may be assumed that at any depth the effective refraction index is the mean of that of air and the bulk material, weighted in proportion to the amount of material present at that depth [16]. For a moth eye surface with the height of the protuberances of h and the spacing of d, it is expected that the reflectance is very low for wavelengths less than about 2.5 h and greater than d at normal incidence, and for wavelengths greater than 2d for oblique incidence. For protuberances with 220 nm depth and the same spacing (typical values for the moth eye), a very low reflectance is expected for the wavelengths between 440 and 550 nm [50].

This moth eye effect should not be confused with reduction of the specular reflectance by roughening of a surface. Roughness merely redistributes the reflected light as diffuse scattering. In the case of the moth eye there is no increase in diffuse scattering, the transmitted wavefront is not degraded, and the reduction in reflection gives rise to a corresponding increase in transmission [50].

Attempts are being made to incorporate microscopic corrugations in solar panels to reduce light reflection. Attempts are also being made to produce a glare-free computer screen by creating facets on a photosensitive lacquer using lasers. Some 25,000 dots of texture per square mm can essentially eliminate the glare on the screen [48]. Hadobás et al. [51] prepared patterned silicon surfaces with 300 nm periodicity and depth up to 190 nm. They found a significant reduction in reflectivity, partially due to the moth-eye effect. Gao et al. [52] used epoxy and resin to replicate the antireflective surface of a cicada's eye. It is also possible to create transparent surfaces using the moth-eye effect [53].

Hierarchical Organization in Biomaterials

Nature develops biological objects by means of growth or biologically controlled self-assembly adapting to the environmental condition and by using the most commonly found materials. Biological materials are developed by using the recipes contained in the genetic code. As a result, biological materials and tissues are created by hierarchical structuring at all levels in order to adapt form and structure to the function, which have the capability of adaptation to changing conditions and self-healing [16, 54]. The genetic algorithm interacts with the environmental condition, which provides flexibility. For example, a tree branch can grow differently in the direction of the wind and in the opposite direction. The only way to provide this adaptive self-assembly is a hierarchical self-organization of the material. Hierarchical structuring allows the adaptation and optimization of the material at each level.

It is apparent that nature uses hierarchical structures, consisting of nanostructures in many cases, to achieve the required performance [16–18]. Understanding the role of hierarchical structure and development of low cost and flexible fabrication techniques would facilitate commercial applications.

Outlook

The emerging field of biomimetics is already gaining a foothold in the scientific and technical arena. It is clear that nature has evolved and optimized a large number of materials and structured surfaces with rather unique characteristics. As the underlying mechanisms are understood, industry can begin to exploit them for commercial applications.

The commercial applications include new nanomaterials, nanodevices, and processes. As for devices, these include superhydrophobic self-cleaning and/or low drag surfaces, surfaces for energy conversion and conservation, superadhesives, robotics, objects which provide aerodynamic lift, materials and fibers with high mechanical strength, anti-reflective surfaces and surfaces with hues, artificial furs and textiles, various biomedical devices and implants, self-healing materials, and sensory aid devices, to name a few.

Cross-References

- ▶ Biornamentics: Architecture Defined by Natural Patterns
- ▶ Gecko Feet
- ▶ Lotus Effect
- ▶ Rose Petal Effect
- ▶ Shark Skin Effect

References

1. Gordon, J.E.: The New Science of Strong Materials, or Why You Don't Fall Through the Floor, 2nd edn. Pelican-Penguin, London (1976)
2. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P. (eds.): Molecular Biology of the Cell. Garland Science, New York (2008)
3. Bhushan, B. (ed.): Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
4. Bhushan, B.: Biomimetics: lessons from nature; an overview. Philos. Trans. R. Soc. A **367**, 1445–1486 (2009)
5. Bhushan, B., Jung, Y.C., Koch, K.: Micro-, nano-, and hierarchical structures for superhydrophobicity, self-cleaning, and low adhesion. Philos. Trans. R. Soc. A **367**, 1631–1672 (2009)
6. Koch, K., Bhushan, B., Jung, Y.C., Barthlott, W.: Fabrication of artificial lotus leaves and significance of hierarchical structure for superhydrophobicity and low adhesion. Soft Matter **5**, 1386–1393 (2009)
7. Gao, X.F., Jiang, L.: Biophysics: water-repellent legs of water striders. Nature **432**, 36 (2004)
8. Gao, H., Wang, X., Yao, H., Gorb, S., Arzt, E.: Mechanics of hierarchical adhesion structures of geckos. Mech. Mater. **37**, 275–285 (2005)
9. Reif, W.E.: Squamation and ecology of sharks. Cour. Forschungsinstitut Senckenberg **78**, 1–255 (1985) (Frankfurt am Main)
10. Bar-Cohen, Y. (ed.): Biomimetics: Biologically Inspired Technologies. Taylor & Francis, Boca Raton (2006)
11. Genzer, J., Efimenko, K.: Recent developments in superhydrophobic surfaces and their relevance to marine fouling: a review. Biofouling **22**, 339–360 (2006)
12. Bhushan, B., Jung, Y.C.: Wetting, adhesion and friction of superhydrophobic and hydrophilic leaves and fabricated micro/nanopatterned surfaces. J. Phys. Condens. Matter **20**, 225010 (2008)
13. Koch, K., Bhushan, B., Barthlott, W.: Diversity of structure, morphology, and wetting of plant surfaces (invited). Soft Matter **4**, 1943–1963 (2008)
14. Koch, K., Bhushan, B., Barthlott, W.: Multifunctional surface structures of plants: an inspiration for biomimetics (invited). Prog. Mater. Sci. **54**, 137–178 (2009)
15. Koch, K., Bhushan, B., Barthlott, W.: Multifunctional plant surfaces, smart materials. In: Bhushan, B. (ed.) Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
16. Nosonovsky, M., Bhushan, B.: Multiscale Dissipative Mechanisms and Hierarchical Surfaces: Friction, Superhydrophobicity, and Biomimetics. Springer, Heidelberg (2008)
17. Nosonovsky, M., Bhushan, B.: Roughness-induced superhydrophobicity: a way to design non-adhesive surfaces. J. Phys. Condens. Matter **20**, 225009 (2008)
18. Nosonovsky, M., Bhushan, B.: Biologically-inspired surfaces: broadening the scope of roughness. Adv. Funct. Mater. **18**, 843–855 (2008)
19. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
20. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
21. Barthlott, W., Neinhuis, C.: Purity of the sacred lotus, of escape from contamination in biological surfaces. Planta **202**, 1–8 (1997)
22. Bhushan, B., Jung, Y.C.: Micro and nanoscale characterization of hydrophobic and hydrophilic leave surface. Nanotechnology **17**, 2758–2772 (2006)
23. Neinhuis, C., Barthlott, W.: Characterization and distribution of water-repellent, self-cleaning plant surfaces. Ann. Bot. **79**, 667–677 (1997)
24. Wagner, P., Furstner, R., Barthlott, W., Neinhuis, C.: Quantitative assessment to the structural basis of water repellency in natural and technical surfaces. J. Exp. Bot. **54**, 1295–1303 (2003)

25. Autumn, K., Liang, Y.A., Hsieh, S.T., Zesch, W., Chan, W.P., Kenny, T.W., Fearing, R., Full, R.J.: Adhesive force of a single Gecko foot-hair. Nature **405**, 681–685 (2000)
26. Bhushan, B.: Adhesion of multi-level hierarchical attachment systems in Gecko feet, (invited). J. Adhesion Sci. Technol. **21**, 1213–1258 (2007)
27. Bhushan, B. (ed.): Nanotribology and Nanomechanics: An Introduction, 2nd edn. Springer, Heidelberg (2008)
28. Gorb, S.: Attachment Devices of Insect Cuticle. Kluwer, Dordrecht (2001)
29. Arzt, E., Gorb, S., Spolenak, R.: From micro to nano contacts in biological attachment devices. Proc. Natl. Acad. Sci. U. S. A. **100**, 10603–10606 (2003)
30. Federle, W.: Why are so many adhesive pads hairy? J. Exp. Biol. **209**, 2611–2621 (2006)
31. Bhushan, B., Sayer, R.A.: Surface characterization and friction of a bio-Inspired reversible adhesive tape. Microsyst. Technol. **13**, 71–78 (2007)
32. Geim, A.K., Dubonos, S.V., Grigorieva, I.V., Novoselov, K.S., Zhukov, A.A., Shapoval, S.Y.: Microfabricated adhesive mimicking Gecko Foot-hair. Nat. Mater. **2**, 461–463 (2003)
33. Gorb, S., Varenberg, M., Peressadko, A., Tuma, J.: Biomimetic mushroom-shaped fibrillar adhesive microstructure. J. R. Soc. Interface **4**, 271–275 (2007)
34. Cutkosky, M.R., Kim, S.: Design and fabrication of multi-materials structures for bio-inspired robots. Philos. Trans. R. Soc. A **367**, 1799–1813 (2009)
35. Bechert, D.W., Bruse, M., Hage, W., van der Hoeven, J.G.T., Hoppe, G.: Experiments on drag-reducing surfaces and their optimization with an adjustable geometry. J. Fluid Mech. **338**, 59–87 (1997)
36. Bechert, D.W., Bruse, M., Hage, W., Meyer, R.: Fluid mechanics of biological surfaces and their technological application. Naturwissenschaften **87**, 157–171 (2000)
37. Dean, B., Bhushan, B.: Shark skin surfaces for fluid drag reduction in turbulent flow: a review. Philos. Trans. R. Soc. A **368**, 4775–4806 (2010)
38. Anonymous, www.biomimicryinstitute.org
39. Fish, F.E.: Limits of nature and advances of technology: what does biomimetics have to offer to aquatic robots? ABBI **3**, 49–60 (2006)
40. Hoyt, J.W.: Hydrodynamic drag reduction due to fish slimes. In: Wu, T.Y.T., Brokaw, C.J., Brennen, C. (eds.) Swimming and Flying in Nature, vol. 2. Plenum, New York (1975)
41. Motier, J.F., Carrier, A.M.: Recent studies on polymer drag reduction in commercial Pipelines. In: Sellin, R. H.J., Moses, R.T. (eds.) Drag Reduction in Fluid Flows: Techniques for Friction Control. Ellis Horwood, Chichester (1989)
42. Bormashenko, E., Bormashenko, Y., Stein, T., Whyman, G., Bormashenko, E.: Why do pigeon feathers repel water? Hydrophobicity of pennae, cassie-baxter wetting hypothesis and cassie-wenzel capillarity-induced wetting transition. J. Colloid Interface Sci. **311**, 212–216 (2007)
43. Jakab, P.L.: Vision of a Flying Machine. Smithsonian Institution Press, Washington, DC (1990)
44. Jin, H.J., Kaplan, D.L.: Mechanism of silk processing in insects and spiders. Nature **424**, 1057–1061 (2003)
45. Sarikaya, M., Aksay, I.A. (eds.): Biomimetic Design and Processing of Materials. American Institute of Physics, Woodbury (1995)
46. Vogel, S.: Comparative Biomechanics: Life's Physical World. Princeton University Press, Princeton (2003)
47. Parker, A.R.: Natural photonics for industrial applications. Philos. Trans. R. Soc. A **367**, 1759–1782 (2009)
48. Mueller, T.: Biomimetics design by natures. Natl. Geogr. **2008**, 68–90 (2008)
49. Gorb, S.: Functional surfaces in biology: mechanisms and applications. In: Bar-Cohen, Y. (ed.) Biomimetics: Biologically Inspired Technologies, pp. 381–397. Taylor & Francis, Boca Raton (2006)
50. Wilson, S.J., Hutley, M.C.: The optical properties of the 'moth-eye' antireflective surfaces. J. Mod. Opt. **29**, 993–1009 (1982)
51. Hadobás, K., Kirsch, S., Carl, A., Acet, M., Wasserman, E.F.: Reflection properties of nanostructure-arrayed silicon surfaces. Nanotechnology **11**, 161–164 (2000)
52. Gao, H., Liu, Z., Zhang, J., Zhang, G., Xie, G.: Precise replication of antireflective nanostructures from biotemplates. Appl. Phys. Lett. **90**, 123115 (2007)
53. Clapham, P.B., Hutley, M.C.: Reduction of length reflection by moth eye principle. Nature **244**, 281–282 (1973)
54. Fratzl, P., Weinkamer, R.: Nature's hierarchical materials. Prog. Mater. Sci. **52**, 1263–1334 (2007)

Biomimetics of Marine Adhesives

Pierre Becker, Elise Henneberd and Patrick Flammang
Laboratoire de Biologie des Organismes Marins et Biomimétisme, Université de Mons – UMONS, Mons, Belgium

Synonyms

Marine bioinspired adhesives

Definition

A large diversity of marine organisms possesses adhesive mechanisms allowing strong attachment to various substrata in a wet and saline

environment. This remarkable capacity raised the interest of scientists for the development of bio-inspired underwater adhesives for various applications, notably in the biomedical field. Model organisms that have been used for the development of such biomimetic adhesives include mussels, barnacles, tubeworms, and algae. All attach permanently by means of solid cements mostly made up of specialized proteins. Synthesis of biomimetic marine adhesives is performed either by the recombinant DNA technology (i.e., the production of adhesive proteins by transformed host cells such as bacteria or yeasts) or by the chemical synthesis of polymers incorporating functional groups copying key amino acids from the natural cement proteins. Other remarkable properties of natural cements, such as the capacity to form complex coacervates, to self-assemble and to cross-link, have been mimicked for the underwater delivery and in situ curing of the biomimetic adhesives.

Introduction

Of all biological phenomena that have been investigated with a view to biomimetics, adhesion in nature has perhaps received the most interest. Indeed, biological adhesives often offer impressive performance in their natural context and, therewith, the potential to inspire novel, superior industrial adhesives for an increasing variety of high-tech applications [1, 2]. Yet, because of the complexity of biological adhesion and the multidisciplinarity needed to tackle it, there has been little visible progress in the development of bio-inspired adhesives and many technological challenges remain, such as the development of adhesives which can function underwater or in aqueous environments. The marine environment is a place of choice for the search of inspiration for such adhesives. Indeed, numerous invertebrates like molluscs, worms, or sea stars produce adhesive secretions that are able to form strong attachments even when totally immersed [1, 2]. These adhesive secretions contain specialized adhesive proteins sometimes associated with other macromolecules such as polysaccharides.

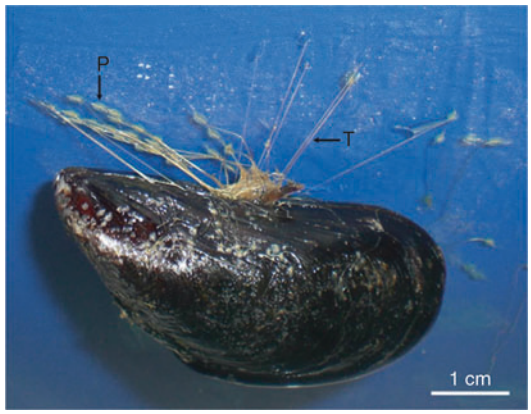
Three types of adhesions can be recognized in marine organisms [3]. Permanent adhesion is characteristic of algae and animals that attach strongly to a substratum with a solid glue possessing high adhesive and cohesive strength (e.g., mussel, barnacles, tube worms, tunicates [1, 2]). Nonpermanent adhesion occurs in invertebrates that attach firmly, but only temporarily, to the substratum through viscoelastic secretions and therefore retain the capacity to move (e.g., sea anemones, gastropods, cephalopods, sea stars, sea urchins [1, 2]). Finally, instantaneous adhesion comprises adhesive systems that rely on single-use organs or cells and are used in functions requiring a very fast formation of adhesive bonds like prey capture or defense reactions (e.g., comb jellies, sea cucumbers [1, 2]). All these systems allow underwater attachment but differ by their structure, functioning, and the physicochemical characteristics of their adhesives. Among them, however, a more intense research effort has been devoted to the elucidation of the mechanisms allowing permanent adhesion of sessile invertebrates and algae [4]. These organisms are indeed involved in the economically important problem of fouling in the marine environment. Because biomimetic efforts are only possible when composition and key molecular components of biological adhesives are understood, so far only a very limited number of organisms have been used for the development of bio-inspired adhesives. The best-characterized marine bioadhesive is that from the mussel and it has inspired most of the biomimetic adhesives currently available [5]. Recently, however, other organisms such as tubeworms, brown algae, and barnacles have also been used as models. The following paragraphs will focus on these four organisms, from the features of their adhesives to the imaginative applications that have been developed for their biomimetic counterparts.

Mussels

Natural Adhesive

To attach themselves to the substratum, mussels produce a byssus (Fig. 1), which consists of a

bundle of proteinaceous threads connected proximally to the base of the animal's foot, within the shell, and terminating distally with a flattened plaque which mediates adhesion to the substratum [5, 6]. These plaques are formed by the autoassembly of secretory products originating from four distinct glands enclosed in the mussel foot. These products comprise a collagenous substance, a mucous material, a mixture of polyphenolic proteins (known as foot proteins 2 to 6, abbreviated as fp-2-6) and an accessory protein (fp-1). Only the different foot proteins (fp-1-6), also known as mussel adhesive proteins (MAPs), will be considered here.



Biomimetics of Marine Adhesives, Fig. 1 A mussel of the species *Mytilus edulis* attached on a glass substratum by the mean of its byssus. The byssus consists of a bundle of threads (*T*), each terminating with a flattened plaque (*P*) which mediates adhesion to the surface

The Mussel Adhesive Proteins (MAPs)

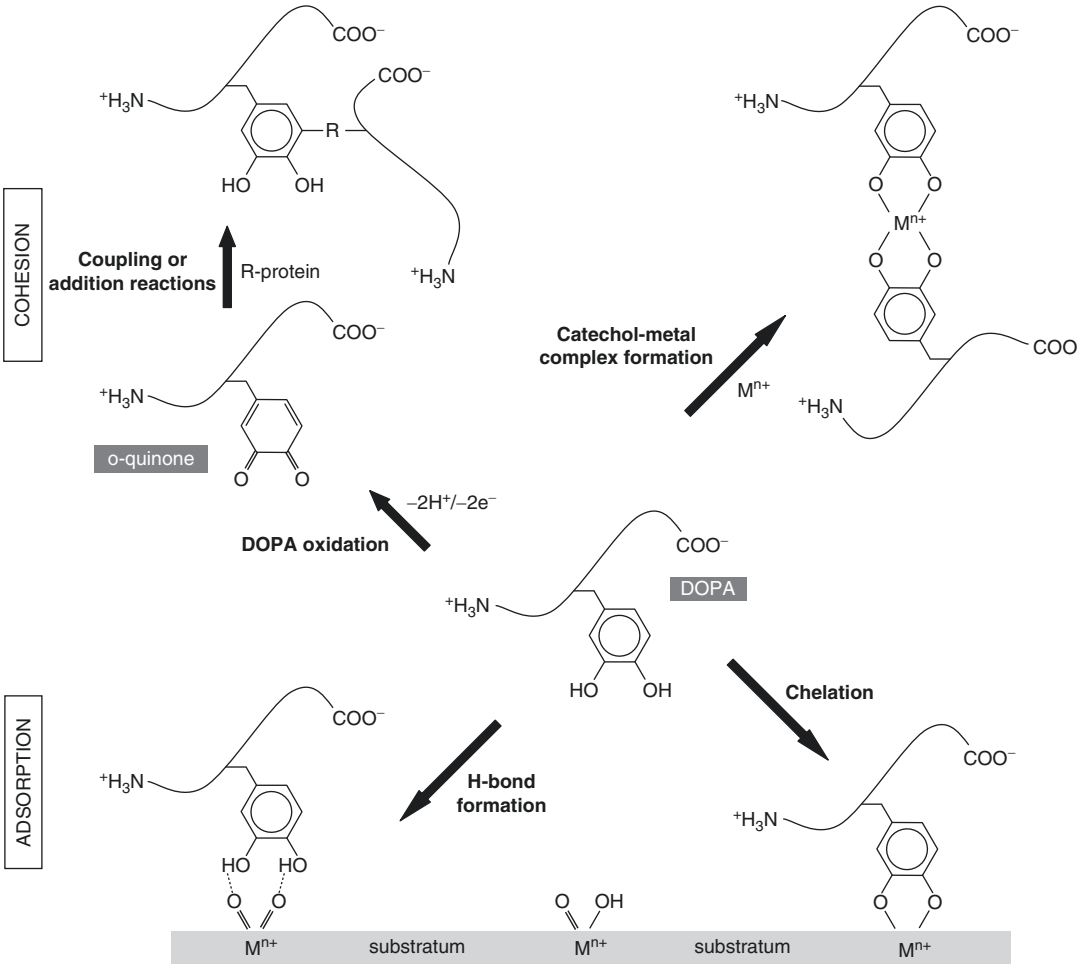
Since the characterization of fp-1 in the early 1980s, MAPs have been the subject of a very large number of studies leading to a detailed knowledge on their structures, functions, and interactions within the byssal attachment plaque. Proteins fp-2 and fp-4 form the central core of the plaque; fp-3, fp-5, and fp-6 are located at the interface between the plaque and the substratum (primer layer); and fp-1 forms a hard cuticle protecting the core from hydrolysis, abrasion, and microbial attack [5, 6].

The 6 f. proteins (fps) constituting the attachment plaques have been mostly characterized from mussels of the genus *Mytilus*. Their characteristics are summarized in Table 1. Most MAPs exhibit repeated sequence motifs, whose number and amino acid composition vary according to the species considered. Moreover, all the proteins identified in the plaque share a common distinctive feature: the presence of 3,4-dihydroxyphenylalanine (DOPA), a residue formed by the posttranslational hydroxylation of tyrosine [5, 6] (Fig. 2). This modified amino acid fulfills two important roles in the attachment plaque (Fig. 2): it is involved in the formation of cross-links between the different fps (cohesion) and it mediates physicochemical interactions with the surface (adhesion) [5]. It is generally accepted

Biomimetics of Marine Adhesives, Table 1 Characteristics of the proteins constituting the attachment plaques in the byssus of mussels from the genus *Mytilus*

Protein	Localization in the plaque	Mass (kDa)	Repeated unit (frequency)	DOPA (mol %)	Feature
fp1	Cuticle	110	AKPSYPPTYK (80)	10–15	Hyp, DiHyp
fp2	Core	47	EGF-like motif (11)	2–3	Cys
fp3	Primer layer	6	None	20–25	Arg-OH
fp4	Core	93	HVHTHRVLHK (36, in N-term half)	<2	His
			DDHVNDIAQTA (16, in C-term half)		
fp5	Primer layer	9.5	None	27	p-Ser
fp6	Primer layer	12	None	4	p-Ser, Cys

Arg-OH 4-hydroxyarginine, *Cys* cysteine, *DiHyp* 3,4 dihydroxyproline, *His* histidine, *Hyp* 4-hydroxyproline, *p-Ser* *O*-phosphoserine



Biomimetics of Marine Adhesives, Fig. 2 Dual functionality of peptidyl DOPA groups in mussel adhesive proteins. The catechol functions of DOPA-containing proteins contribute to adhesive adsorption through hydrogen bonding of the phenolic OH groups to the oxygen atoms of the surface (*lower left*) or by forming mono-bidentate complexes with metal ions at the surface of mineral, metal oxide or metal hydroxide substrata (*lower right*). Peptidyl DOPA groups also contribute to adhesive cohesion by forming bis- or tris-catechol-metal complexes (*upper right*) or by forming intermolecular cross-links. Cross-linking follows the oxidation of DOPA to an *o*-quinone that reacts with amino acid side chains (R) of other proteins, in which R can belong to cysteine, histidine, lysine, and other DOPA residues (*upper left*)

that cross-linking reactions are related to the oxidation of DOPA to DOPA-quinone, a reaction catalyzed by a catecholoxidase in the byssus. DOPA-quinones may also result from redox reactions involving transition metal ions or may form spontaneously at alkaline pH. Once formed, DOPA-quinone is capable of participating in a number of different reaction pathways leading to intermolecular cross-link formation [6]. DOPA also confers to proteins a capacity for intermolecular metal complexation through the formation of tris- and/or bis-catechol-metal complexes. On the other hand, DOPA is also involved in surface coupling, either through hydrogen bonds or by forming complexes with metal ions and metal oxides present in mineral surfaces [5, 6]. These complexes possess some of the highest known stability constants of metal-ligand chelates. Some of the MAPs also contain other posttranslationally modified amino

acids such as O-phosphoserine (fp-5 and fp-6), which could mediate adhesion to calcareous substrata, or 4-hydroxyproline (fp-1), 3,4 dihydroxyproline (fp-1) and 4-hydroxyarginine (fp-3) which could make hydrogen bonds with surfaces [5, 6].

MAP Extraction and Purification

The extraction and purification of MAPs directly from mussel tissues for purposes is feasible but is economically and ecologically problematic. Indeed, several thousands mussels are needed to obtain 1 g of proteins [4, 7, 8]. Harvesting the natural adhesive is therefore an unrealistic solution for most industrial applications. A commercial MAP extract is however commercially available under the name Cell-Tak™ (BD Biosciences, Franklin Lakes, NJ USA) and consists of a mixture of fp-1 and fp-2. It is sold as an adhesive for the in vitro immobilization of cells and tissue sections on various surfaces, including plastic, glass, metal, and biological materials [4]. In vivo transplantation experiments (e.g., corneal transplantation in rabbit) also showed that Cell-Tak appeared to be nontoxic and well tolerated by biological systems. In the laboratory, the adhesive properties of an extracted MAP mixture similar to Cell-Tak (i.e., containing ∼80 % fp-1 and ∼20 % fp-2) were investigated by bonding together strips of porcine tissues, either skin or small intestinal submucosa [8]. With skin, a satisfactory joint strength of about 1 MPa, similar to the commercial fibrin glue Tisseel™ (Baxter, Deerfield, IL USA) was obtained, but only for long curing time exceeding 12 h. With small intestine, an adhesive strength of almost 0.5 MPa was reported and the curing was reduced to 1 h by the addition of oxidizing transition metal ions such as Fe^{3+} and V^{5+}. It is usually admitted that adhesives suitable for medical applications should possess an adhesive and cohesive strength ranging from 0.01 to 6 MPa, depending on the particular bonding site (e.g., relatively unstressed soft tissues versus bones or tendons) [2]. In addition, the properties of the adhesive should be adapted to the "dry" (topical tissues such as teeth or skin) or "wet" (internal organs bathed in body fluids) state of the bonding site.

Production of Mussel Mimetic Adhesives and Their Applications

To bypass the problem of obtaining MAPs directly from mussels, biomimetic adhesives have been developed. These bio-inspired molecules are produced either in the form of recombinant preparations of the adhesive proteins or in the form of chemically synthesized polymers incorporating DOPA or catechol groups.

Production of Recombinant MAPs

Recombinant DNA technology has been used in order to obtain larger quantities of mussel adhesive proteins [6–9]. Attempts to produce recombinant forms of MAPs started in the early 1990s with the expression, production, and purification of complete fp-1 and of synthetic fp-1 protein analogs consisting of 6–20 repeats of the consensus fp-1 decapeptide (Ala-Lys-Pro-Ser-Tyr-Pro-Pro-Thr-Tyr-Lys) in the yeast or in the bacterium *Escherichia coli* [6, 9]. However, these first attempts failed to obtain large amounts of recombinant proteins (Table 2) for different reasons including the highly biased amino acid composition of fp-1 and differences in codon usage between mussels and the heterologous expression systems [6, 9]. Bacterial expression of recombinant fp-1 was enhanced by the fusion of fp-1 with an *E. coli* signal peptide (OmpASP), which boosts secretion by directing the expressed protein to the periplasm. Another problem with the heterologous expression of fp-1 and MAPs in general is that the recombinant proteins lack the posttranslational modifications such as the hydroxylation of proline, arginine, or most importantly, tyrosine residues. The production of functional recombinant MAPs therefore requires an additional in vitro modification step: the enzyme-catalyzed modification of tyrosine residues into DOPA [6]. This is usually done using a commercially available mushroom tyrosinase.

Fp-3 and fp-5 are other interesting adhesive proteins as they have a high DOPA content and are present at the interface between the byssus attachment plaques and the substratum. Both were cloned in expression vectors containing a histidine tag sequence which allows purification of the protein using metal affinity

Biomimetics of Marine Adhesives, Table 2 Comparison of several recombinant mussel adhesive proteins

Recombinant protein	Host	Production yield (mg/L) after purification	Solubility (g/L)
fp-1	*Saccharomyces cerevisiae* (Y)	nd	nd
	Escherichia coli (B)	4–10	nd
fp-2	*Saccharomyces cerevisiae* (Y)	nd	nd
fp-3	*Kluyveromyces lactis* (Y)	1	nd
	Escherichia coli (B)	0.8–3	1
fp-5	*Escherichia coli* (B)	2.8	1
fp-131	*Escherichia coli* (B)	nd	nd
fp-151	*Escherichia coli* (B)	1,000	330
fp-353	*Escherichia coli* (B)	39	90

B bacterium, *Y* yeast, *nd* not determined

chromatography [9]. The vectors were expressed in *E. coli* with subsequent in vitro treatment by mushroom tyrosinase for DOPA formation. Recombinant fp-5 presented a low purification yield (Table 2) but remarkable adsorption and adhesive abilities on various surfaces [6, 9]. Indeed adsorption and adhesion forces were comparable to – and sometimes exceeded – those of Cell-Tak. A better purification yield was achieved for recombinant fp-3 (Table 2). The adhesion capabilities of this protein were, however, lower than those of recombinant fp-5, but still comparable to those of Cell-Tak. Recombinant fp-5 was also tested successfully as a cell-adhesion material for in vitro cell culture [6, 9].

The successful production of recombinant MAPs and their good performance in microscale adhesion tests were encouraging. However, macroscale testing and large-scale applications were still prevented by poor production due to post-induction bacterial cell growth inhibition and by the low solubility of purified proteins in aqueous buffers (Table 2). To overcome these limitations, recombinant hybrid adhesive proteins were designed, the so-called fp-353, fp-151, and fp-131 [7, 9]. The former is a fusion protein with fp-3 at each terminus of fp-5. Because fp-353 formed inclusion bodies, host cell growth inhibition did not occur. In addition, the solubility of fp-353 was better than that of fp-3 or fp-5 alone, permitting the preparation of a viscous concentrated glue solution for large-scale adhesion strength measurements. The fp-151 and fp-131 hybrids resulted from the fusion of six fp-1 decapeptide repeats at each terminus of fp-5 and fp-3, respectively [6, 7, 9]. Recombinant fp-151 displayed the highest production yield reaching 1 g/L of batch-type flask culture (Table 2). In fed-batch-type bioreactor cultures, similar high production levels were maintained through a co-expression of fp-151 with bacterial hemoglobin that facilitates the utilization of cellular oxygen [6]. Of all recombinant MAPs, fp-151 possesses the highest solubility (up to 330 g/L), allowing the concentration of the protein solution into a sufficiently viscous liquid for practical adhesive application. Additionally, fp-151 showed a better adsorption and a similar adhesion force, compared to recombinant fp-5, and its adhesive strength (∼0.8 MPa on cowhide, ∼1.1 MPa on aluminum, and ∼1.8 MPa on poly (methyl methacrylate [Plexiglas])) is always largely in excess to that of fibrin glue. The recombinant fusion proteins, marketed in milligram quantities (Kollodis, Inc), are less expensive than extracted MAPs (Cell-Tak) [7].

Fp-151 proved to be biocompatible for the in vitro culture of various cell types including both anchorage-dependent and anchorage-independent cells [6, 9]. In this context, fp-151 was also fused to a RGD peptide, a sequence identified as a cell-adhesion recognition motif by integrins, and used in culture plates coating. The resulting hybrid, fp-151-RGD, maintained the high production yield of fp-151, but also presented superior spreading and cell-adhesion abilities compared to the commercially produced cell-adhesion materials poly-L-lysine (PLL) and

Cell-Tak, and this regardless of mammalian cell line used [6]. The excellent adhesion and spreading abilities of fp-151-RGD might be due to the fact that it combines three types of cell-binding mechanisms: DOPA adhesion of Cell-Tak, cationic binding force of PLL, and RGD sequence-mediated adhesion of fibronectin. These characteristics make the two hybrid proteins fp-151 and fp-151-RGD suitable for use as cell-adhesion material in cell culture or tissue engineering. In a totally different context, fp-151 has been proposed as a potential gene delivery material in view of its similar amino acid composition to histone proteins, which are known as effective mediators of transfection. The fusion protein displayed comparable transfection efficiency in human and mouse cells compared to the widely used transfection agent Lipofectamine™ (Invitrogen).

Production of MAP Mimetic Adhesive Polymers

Another way to mimic mussel adhesive proteins is by the chemical synthesis of important functional parts of these adhesives. This has been done either by peptide synthesis or by the functionalization of various other polymers with MAP reactive groups, mostly DOPA or catechol groups.

Various synthetic polypeptides inspired by MAPs were studied for adhesion strength [7, 8, 10]. This approach was used to experimentally identify the exact functions of the amino acids that are active in the chemistry of the adhesive proteins. The importance of particular MAP over-expressed amino acids, mainly Lys and Tyr/DOPA, was therefore tested through different polytripeptides and polydipeptides such as poly (X-Tyr-Lys), poly(Gly-Lys), poly(Tyr-Lys), and poly(Gly-Tyr), each peptide being subsequently incubated with tyrosinase for the enzymatic hydroxylation of tyrosines into DOPA. The adhesive strength of poly(X-Tyr-Lys) peptides, irrespective of whether X was Gly, Ala, Pro, Ser, Leu, Ile, or Phe, measured on pig skin (i.e., 0.01 MPa) was higher than that of poly(Gly-Lys) peptides but similar to the one of poly(Tyr-Lys) peptides and of a poly(fp-1 decapeptide) [10]. Tyrosine (DOPA) and lysine thus appears as key amino acids for adhesion efficiency, with little importance of the other residues and of the primary structure of the MAPs. A poly(Gly-Tyr-Lys) peptide was also used as a surgical glue to close a skin incision in a living pig. Good incision adhesion and reduced immunological response after 1 week were observed.

Peptide synthesis also allows the direct incorporation of DOPA into the polymers, thus avoiding the complications associated with the oxidation of precursor tyrosine residues [7, 8, 10]. Water soluble copolypeptides containing DOPA and Lys residues were prepared by polymerization of α-amino acid N-carboxyanhydride monomers. Using lap shear tensile adhesive measurements, these copolymers were found to form moisture-resistant adhesive bonds to a variety of substrata after cross-linking with different oxidizing agents. Adhesion was strongest to metals and polar surfaces such as glass (3–5 MPa) [10]. Adhesive strength of about 0.2 MPa was also measured on porcine skin and bone in vitro. Furthermore, outcomes of endothelial cell cultures demonstrated that the copolypeptide presented a good cell affinity, which would provide basic data for its application in the biomaterial field [10].

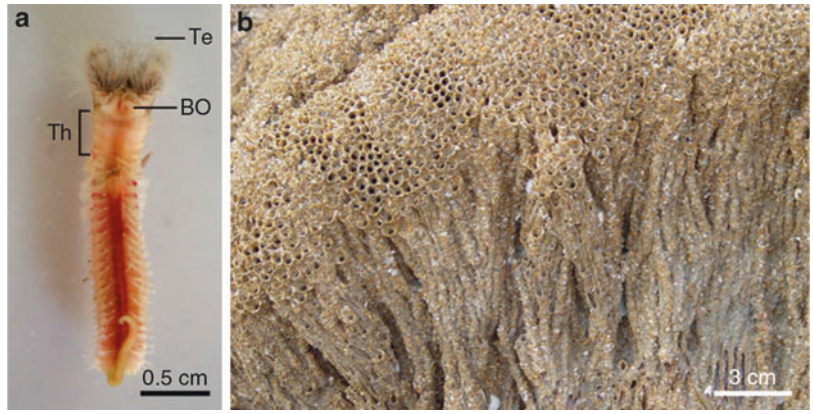
DOPA being considered as a key component of MAPs for adhesion, functionalization of synthetic polymers other than polypeptides with catechol groups has emerged as a promising strategy for the development of new biomimetic adhesives [1, 7, 11]. Following this approach, simplified polymer mimics of MAPs have been designed in which a polystyrene backbone is used to take the place of the protein polyamide chain. Hardening of this poly[(3,4-dihydroxystyrene)-*co*-styrene] by treatments with various oxidizing agents including dichromate and Fe^{3+} yielded adhesive strength of up to 1.2 MPa on aluminum [7]. Linear and branched poly(ethyleneglycol) (PEG) molecules were also chemically coupled to one to four DOPA endgroups. PEG has a low toxicity and is considered to be nonallergenic, therefore allowing its utilization as a convenient platform in medicine [1]. The addition of oxidizing agents resulted in the polymerization of PEG-DOPA via DOPA-quinone cross-links, forming polymer networks and rapid gelation. Although these gels were not

adhesive, it has been shown that unoxidized four-armed PEG polymers functionalized with a single DOPA residue at the extremity of each arm (PEG-$DOPA_4$) adsorbed strongly on mucin, this mucoadsorption being largely due to the presence of DOPA [11]. A tissue adhesive was investigated by using liposomes in order to compartmentalize an oxidizing reagent ($NaIO_4$) in a solution containing PEG-$DOPA_4$ [11]. While sodium periodate is sequestered in these vesicles at room temperature, it is released in the polymer solution when liposomes melted at wound site (37 °C). Oxidation of PEG-$DOPA_4$ resulted in hydrogel formation with interesting adhesive properties. Indeed, lap shear strength between two porcine dermal tissues was five times higher than that of fibrin glue [11]. A similar strategy was used for efficient immobilization of transplanted pancreatic islets in mice, with minimal inflammatory response. Polymers made up of PEG-DOPA coupled to polycaprolactone (PCL) were also used to coat the biologic meshes used for hernia repair.

Paradoxically, PEG-DOPA polymers can also be used for antifouling applications. In this last case, DOPA or DOPA-containing peptides are conjugated to PEG polymers to allow their adsorption to Au and Ti surfaces [1]. The surfaces modified with these polymers were shown to be resistant to cell attachment and protein adsorption suggesting that they might be used to design cell and protein-resistant surfaces for medical implants, anti-icing coating on aircraft wings and nonfouling marine structures (Kensey Nash Corporation, Exton, PA USA).

The use of toxic oxidizing reagents for solidification of hydrogels represents a medical concern for in vivo applications [1, 2, 7, 11]. To avoid such reagents, DOPA was then conjugated to poly(ethyleneoxide)-poly(propyleneoxide)-poly(ethyleneoxide) (PEO-PPO-PEO) block copolymers. These DOPA-modified block copolymers were soluble in cold water but self-aggregated into micelles at a higher temperature that was dependent on the block copolymer concentration. For instance, a 20 wt% solution formed gels when heated to body temperature. Photopolymerization is another approach to achieve oxidation-free cross-linking and monomers combining DOPA and a UV (or visible light)-polymerizable methacrylate group were therefore copolymerized with PEG diacrylate. In both heat and light-triggered polymerization, gelation did not result from DOPA-quinones cross-links. DOPA groups were thus available for adhesion but the adhesive strength of these hydrogels was not investigated. Self-assembly of amphiphilic triblock copolymers is also an alternative strategy to obtain oxidation-free polymerization. Triblock copolymers were made up of hydrophobic endblocks consisting of poly(methylmethacrylate) (PMMA) and of a poly(methacrylic acid) (PMAA) water-soluble midbl ock. DOPA was incorporated into the hydrophilic PMAA. Hydrogels were then obtained by exposing triblock copolymers solution in DMSO to water vapor. As water diffused into the solution, the hydrophobic endblocks formed aggregates that were bridged by water-soluble midblocks. Underwater adhesive properties of a DOPA-modified PMMA-PMAA-PMMA membrane on titanium and pig skin were assessed and showed strong adhesion in both cases.

The strategy centered on the functionality of the catechol group reaches its highest point with the report of a method to form multifunctional polymer coatings through simple dip-coating of objects in an aqueous solution of dopamine, a DOPA analog [12]. Indeed, dopamine self-polymerization forms thin, surface-adherent polydopamine films onto a wide range of inorganic and organic materials including metals, oxides, polymers, semiconductors, and ceramics. Secondary reactions can then be used to create a variety of ad-layers [12, 13]. On metals, these polydopamine films can provide corrosion protection properties. They can also promote cell adhesion on any type of material surfaces including the well-known anti-adhesive substrate, poly(tetrafluoroethylene), and therefore convert a variety of bioinert substrates into bioactive ones [12, 13]. Another important application of this method is the improvement of interfacial properties in composites. The macroscale properties of polymer-matrix composites depend immensely on the quality of the interaction between the reinforcement phase and the bulk polymer, and



Biomimetics of Marine Adhesives, Fig. 3 A polychaete of the species *Sabellaria alveolata* (**a**) and a reef fragment made of the imbricated tubes built by individuals of this species (**b**). *BO* building organ, *Te* tentacles, *Th* thorax

polydopamine films deposited on the reinforcement phase have been demonstrated to increase interfacial shear stress. For solar cell applications, binding photosensitizer onto semiconductor surfaces, especially TiO_2 surfaces, via catechol groups is a very promising method for creating stable interfaces [13].

Tubeworms

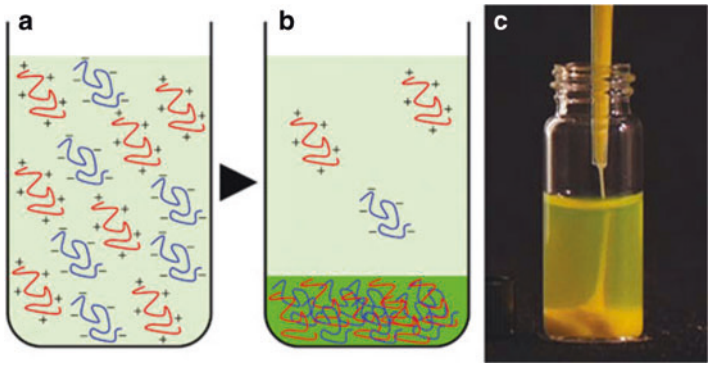
Some polychaete worms of the family Sabellariidae are tube-dwelling and live in the intertidal zone. They are commonly called honeycomb worms or sandcastle worms because they are gregarious and the tubes of all individuals are closely imbricated to form large reef-like mounds (Fig. 3). To build the tube in which they live, they collect particles such as sand grains or shell fragments with their tentacles from the water column and sea bottom. These particles are then conveyed to the building organ which is a crescent-shaped structure near the mouth. There, the particles are dabbed with spots of cement secreted by two types of unicellular glands located in the worm's thorax, and then added to the end of the preexisting tube by the building organ [2, 7, 14].

Cement composition has been investigated in the species *Phragmatopoma californica*. The adhesive consists mostly of three proteins (known as *Phragmatopoma* cement proteins 1–3, abbreviated as Pc-1, 2, and 3) and large amounts of Mg^{2+} and Ca^{2+} ions [7, 14]. The three cement proteins have highly repetitive primary structures with limited amino acid diversity (Table 3). As it is the case for plaque proteins in mussels, Pc-1 and 3 are present in the form of several variants. Pc-1 and 2 contain basic residues with amine side chains and DOPA residues, some of which are halogenated into 2-chloro-DOPA residues [14, 15]. DOPA residues presumably play the same functions as in MAPs (Fig. 2), contributing to both adsorptive interactions with the surface (adhesion) and cross-link formation (cohesion). Regarding the chloro-DOPA residues, they could be involved in the protection of the cement from microbial fouling and degradation. Pc-3 is particularly rich in serine (72.9 mol%) and careful calculations indicated that up to 90 % of

Biomimetics of Marine Adhesives, Table 3 Characteristics of the cement proteins of the tubeworm *Phragmatopoma californica*

Protein	Mass (kDa)	Repeated unit (frequency)	Features
Pc-1	18	VGGYGYGGKK (15)	DOPA, basic
Pc-2	21	HPAVHKALGGYG (8)	DOPA, basic
Pc-3	10–52[a]	Poly(S)	p-Ser (about 80 mol %), highly acidic
			p-Ser definition

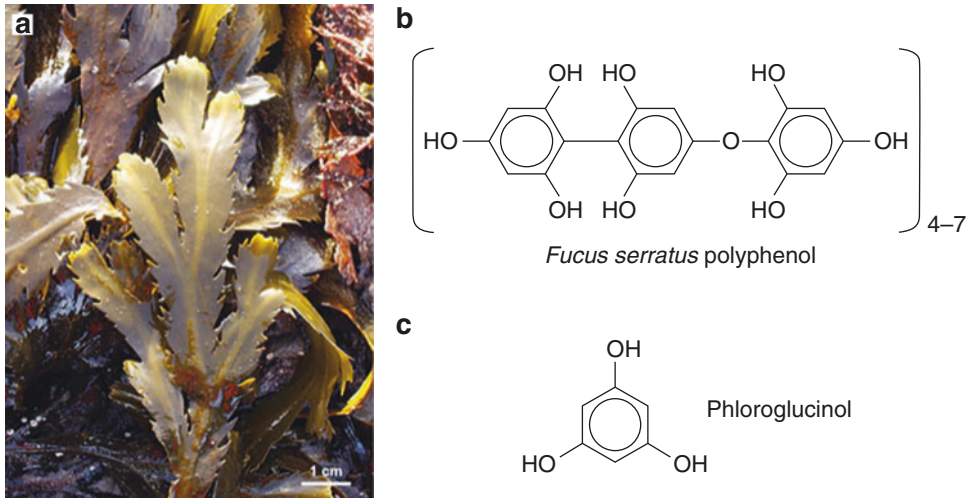
p-Ser *O*-phosphoserine
[a]The mass of Pc-3 varies according to the variant considered



Biomimetics of Marine Adhesives, Fig. 4 Complex coacervation is the associative phase separation in a solution of positively and negatively charged macroions. From a stable colloidal solution of polyelectrolyte complexes (**a**), a trigger such as a pH change leads to the formation of two immiscible aqueous phases, a dilute equilibrium phase and a denser solute-rich phase (**b**). The latter, called the coacervate phase, is an isotropic liquid containing amorphous associative particles that move freely relative to one another ((**c**) Courtesy of R. Stewart, University of Utah, USA)

these serine residues are posttranslationally phosphorylated [14, 15]. Pc-3 is therefore a remarkably acidic protein (pI ∼ 1). Phosphorylation is thought to impart a potential for both cohesive (by Ca^{2+} or Mg^{2+} bridging) and adhesive contributions to the cement [14].

The co-occurrence of the positively charged Pc-1 and Pc-2 proteins and the negatively charged Pc-3 protein in the tube worm cement led to the hypothesis that a phenomenon called complex coacervation could play a role in the condensation of the adhesive in the form of a dense water-immiscible fluid [14, 15]. Complex coacervation is the spontaneous separation of an aqueous solution of two oppositely charged polyelectrolytes into two immiscible aqueous phases, a dilute equilibrium phase and a denser solute-rich phase (Fig. 4). Coacervation occurs when the charges of the polyelectrolytes are balanced. This phenomenon is therefore pH dependent, occurring to the maximum extent at the pH where the solution is electrically neutral; ionic strength dependent, since shielding of charges can change the charge balance of the system [15]; and dependent on the ratio of polyelectrolyte concentrations. Whether complex coacervation is really involved in the formation of the natural cement or not is still under debate. However, the ideal material and rheological properties of complex coacervates provided a valuable blueprint for the synthesis of a biomimetic, water-borne, underwater adhesive [15]. Based on the tubeworm model, poly(meth) acrylate polymers with phosphate, amine, and catechol side chains in the same molar ratios as the natural adhesive proteins Pc1-3 were synthesized [7, 15]. Analogs of Pc-1 and 2 were created by copolymerization of *N*-(3-aminopropyl) methacrylamide hydrochloride (APMA) and acrylamide. Mimics of Pc-3 were synthesized by free radical copolymerization of monoacryloxyethyl phosphate (MAEP), acrylamide, and dopamine methacrylate (DMA), the latter being a DOPA analog monomer [15]. Both copolymers were mixed, with divalent cations, in similar proportions as the natural proteins. Around neutral pH, they condensed into a liquid complex coacervate (Fig. 4c) while at pH 10, they formed a hydrogel through cross-links between dopaquinone and amine side chains [7, 15]. The underwater bond strength of the biomimetic adhesive on aluminum was ∼0.8 MPa, about twice the value estimated for the natural glue [15]. Moreover, a biodegradable, temperature-triggered version of the adhesive coacervate was designed by replacing the APMA/acrylamide copolymer by an amine-modified collagen hydrolysate. This adhesive coacervate was used to repair rat calvarial bone defect and was capable of maintaining three-dimensional bone alignment in freely moving rats over a 12-week indwelling period



Biomimetics of Marine Adhesives, Fig. 5 The brown alga *Fucus serratus* ((**a**) Courtesy of H. Bianco-Peled, Technion – Israel Institute of Technology) and the chemical structure of brown algal polyphenol (**b**) and phloroglucinol monomer (**c**)

[15]. Histological evaluation demonstrated that the adhesive was gradually resorbed and replaced by new bone with an inflammatory response commensurate with normal wound healing. This noncytotoxic degradable coacervate therefore appears to be suitable for use in the reconstruction of craniofacial features [15].

The tubeworm cement-inspired complex coacervates prompted efforts to form complex coacervates with other biomimetic adhesives such as the mussel hybrid fusion proteins fp-151 and fp-131 [15]. The MAP hybrids are insoluble at physiological pH so coacervation was done at pH 3.8 with hyaluronic acid (HA) at HA:fp-151 (or fp-131) ratios of 3:7. Importantly, it was found that the highly condensed complex coacervates significantly increased the bulk adhesive strength of MAPs in both dry and wet environments (up to ~4 MPa on aluminum) [15]. In addition, oil droplets were successfully incorporated in the coacervate, forming a microencapsulation system that could be useful in the development of self-adhesive microencapsulated drug carriers, for use in biotechnological and biomedical applications. Coacervates were also prepared using HA and fp-151-RGD [15]. The low interfacial energy of the coacervate was exploited to coat titanium (Ti), a metal widely used in implant materials. The coacervate effectively distributed both HA and fp-151-RGD over the Ti surfaces and enhanced osteoblast proliferation.

Brown Algae

Brown algae live firmly attached on substrata in the subtidal and intertidal rocky shores where they are often exposed to high gradients of turbulence [1]. They have therefore developed adhesive strategies to attach themselves strongly and durably throughout their life cycle, from the microscopic reproductive cell stages to the large thalli of mature algae (Fig. 5a). Enzymatic extractions and cytochemical studies indicated that the adhesive mucilage produced by these algae contain polysaccharides and glycoproteins [1, 16]. Phenolic polymers also play a role in brown algae adhesion. They appear to be secreted a few hours after fertilization, allowing initial substratum adhesion and, after germination, are localized at the site of adhesive mucilage formation, the rhizoid tip in the case of *Fucus*. These polyphenols are composed of phloroglucinol units linked by carbon–carbon and ether bonds (Fig. 5b). They are reminiscent of the DOPA-containing proteins of mussels and tubeworms in that they can be oxidized and form

cross-links. The proposed mechanism of algal adhesion indeed postulates that the polyphenols are oxidized by an extracellular vanadium bromoperoxidase to enable their cross-linking with the polysaccharide components of the adhesive such as alginate [16].

The adhesive of the species *Fucus serratus* inspired a biomimetic glue composed of synthetic phloroglucinol units (Fig. 5c), alginate, and calcium for medical applications [17]. This synthetic adhesive was shown to be capable of adhering to a variety of both hydrophilic and hydrophobic surfaces, with strength similar to that of the algal-born natural adhesive (0.017–0.025 MPa). It also adhered well to porcine tissue and was shown to be safe for living cells. The biomimetic adhesive was also tested on porcine muscle tissue and presented a tensile strength similar to that of the commercially available fibrin glue Tisseel™. It is marketed as a vascular sealant for surgical adjunctive leakage control (SEAlantis Ltd., Haifa, Israel).

Barnacles

Barnacles are sessile crustaceans that attach firmly and massively not only to a large variety of underwater natural substrata such as rocks (Fig. 6), but also to man-made substrata such as ship hulls therefore causing major economical losses [2].



Biomimetics of Marine Adhesives, Fig. 6 Group of barnacles of the species *Elminius modestus* attached on a rock (Picture courtesy of N. Aldred, Newcastle University, UK)

In these organisms, attachment is mediated by the release of a permanent adhesive called cement [1, 2, 18]. In acorn barnacles, this cement is produced by large isolated secretory cells (the cement cells) joined together by ducts which open onto the base of the animal [18]. The cement is composed of approximately 90 % proteins with the remainder as carbohydrates (1 %), lipids (1 %), and inorganic ash (4 %, of which calcium accounts for 30 %). More than ten proteins have been found in the cement (cement proteins abbreviated as cp) [1, 19], of which six have been purified and characterized, originally from the species *Megabalanus rosa* and later from other species from the genera *Amphibalanus*, *Balanus*, *Fistulobalanus* and *Semibalanus* [1]. Their features are summarized in Table 4. Among these proteins, three (cp-19 k, cp-20 k, and cp-68 k) have a surface coupling function, two (cp-52 k and cp-100 k) have a bulk function, and the last one (cp-16 k) is an enzyme whose possible function is the protection of the cement from microbial degradation [18, 19]. Evidence to date suggests that barnacles display a contrasting adhesive system compared to mussels and tubeworms, two animals characterized by their reliance on post-translational modification of adhesive proteins, especially the formation of DOPA residues. On the other hand, little or no posttranslational modifications have been found so far in individual barnacle cement proteins [18]. Their strong attachment to surfaces does not involve therefore the "DOPA-system." Only one of the cement proteins, cp-68 k, appears to be posttranslationally modified. This protein is indeed glycosylated, although the glycosyl moiety is limited in amount.

As in mussels, the first attempts to exploit the potentialities of barnacle glue were undertaken in the early 1980s. Native cement was shown to be effective in tests of rabbit bone repair. However, the difficulty to collect sufficient quantities of adhesive material prevented further use of this material in concrete medical procedures [7]. Recently, the proteins cp-19 k and cp-20 k, have been produced in bacteria in the form of recombinant proteins at the laboratory scale. Both recombinant cps were demonstrated to present underwater irreversible adsorption activity to

Biomimetics of Marine Adhesives, Table 4 Characteristics of the cement proteins of the barnacle *Megabalanus rosa*

Protein[a]	Repeated unit	Feature	Proposed function
cp-16 k	None	Similar to lysozyme	Antimicrobial
cp-19 k	None	Bias toward Ser, Thr, Gly, Ala, Lys, Val (these residues amount for 67 % of the total amino acids)	Surface coupling
cp-20 k	Six Cys-rich repeated sequences	Cys, charged amino acids	Surface coupling
cp-52 k	Repetitive sequence	Hydrophobic	Cement structural framework
cp-68 k	None	Bias toward Ser, Thr, Ala, Gly (these residues amount for 60 % of the total amino acids); glycosylated	Surface coupling
cp-100 k	None	Hydrophobic	Cement structural framework

Ala alanine, *Cys* cysteine, *Gly* glycine, *Lys* lysine, *Ser* serine, *Thr* threonine, *Val* valine

[a]Denotes cement protein with its apparent molecular weight (kDa)

a variety of surface materials, including positively charged, negatively charged, and hydrophobic ones [19].

The protein cp-20 k was used for the design of self-assembling peptides [18]. This protein is characterized by its high content of charged amino acids and cysteines, and by the presence of six repetitive sequences deduced from the alignment of cysteines (Cys) [1]. A single repetitive sequence, containing four Cys, has been chemically synthesized to obtain a water-soluble peptide solution. The addition of salt at a final concentration of 1 M triggered an irreversible self-assembly into a macroscopic membrane made up of interwoven nanofilaments [18]. Treatment with dithiothreitol (DTT), a reducing agent, and the replacement of cysteines by serines in the peptide did not affect self-assembly. This means that the latter is not dependent on the formation of disulfide bonds but rather on noncovalent molecular interactions. However, extended incubation of the peptide solution in buffers at alkaline pH caused a conformational change and the formation of a three-dimensional mesh-like structure stabilized by disulfide bonds. It was proposed that this self-assembling peptide may present a novel source of inspiration for the design of new materials [18].

Perspectives

From the biologist's perspective, marine adhesives may serve one or more of the following functions: (a) the temporary or permanent attachment of an organism to a surface, including dynamic attachment during locomotion and maintenance of position; (b) the collection or capture of food items; or (c) the building of tubes or burrows. The evolutionary background and biology of the species, on one hand, and environmental constraints on the other hand, both influence the specific composition and properties of adhesive secretions in a particular organism. The diversity of marine adhesives is therefore huge and daunting for researchers but, in return, it will also allow a great deal of flexibility in applications. To date, only a very limited number of marine organisms have been used as models for the development of biomimetic adhesives and most of this diversity remains unexploited. The study of new model organisms representative of the three different types of adhesion is therefore required in order to extract essential structural, mechanical, and chemical principles from their adhesives. Simplification of natural systems while retaining the desired efficacy is, indeed, one of the foundation stones of biomimetics. As quoted by Herbert Waite [2], it is no longer absurd to predict that a compilation of bio-inspired adhesive designs will someday fill a textbook with recipes for a whole range of applications, from technical to medical.

Cross-References

- ▶ Bioadhesion
- ▶ Bioadhesives
- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Biomimetics
- ▶ Gecko Adhesion

References

1. Smith, A.M., Callow, J.A.: Biological Adhesives. Springer, Heidelberg (2006)
2. von Byern, J., Grunwald, I.: Biological Adhesive Systems – From Nature to Technical and Medical Application. Springer, Wien (2010)
3. Flammang, P., Santos, R., Haesaerts, D.: Echinoderm adhesive secretions: from experimental characterization to biotechnological applications. In: Matranga, V. (ed.) Marine Molecular Biotechnology: Echinodermata, pp. 201–220. Springer, Berlin (2005)
4. Taylor, S.W., Waite, J.H.: Marine adhesives: from molecular dissection to application. In: McGrath, K., Kaplan, D. (eds.) Protein-Based Materials, pp. 217–248. Birkhäuser, Boston (1997)
5. Waite, J.H., Andersen, N.H., Jewhurst, S., Sun, C.: Mussel adhesion: finding the tricks worth mimicking. J. Adhes. **81**, 297–317 (2005)
6. Silverman, H.G., Roberto, F.F.: Understanding marine mussel adhesion. Mar. Biotechnol. **9**, 661–681 (2007)
7. Stewart, R.J.: Protein-based underwater adhesives and the prospects for their biotechnological production. Appl. Microbiol. Biotechnol. **89**, 27–33 (2011)
8. Wilker, J.J.: Marine bioinorganic materials: mussels pumping iron. Curr. Opin. Chem. Biol. **14**, 276–283 (2010)
9. Cha, H.J., Hwang, D.S., Lim, S.: Development of bioadhesives from marine mussels. Biotechnol. J. **3**, 1–8 (2008)
10. Deming, T.J.: Synthetic polypeptides for biomedical applications. Prog. Polym. Sci. **32**, 858–875 (2007)
11. Lee, B.P., Dalsin, J.L., Messersmith, P.B.: Synthesis and gelation of DOPA-modified poly(ethylene glycol) hydrogels. Biomacromolecules **3**, 1038–1047 (2002)
12. Lee, H., Dellatore, S.M., Miller, W.M., Messersmith, P. B.: Mussel-inspired surface chemistry for multifunctional coatings. Science **318**, 426–430 (2007)
13. Ye, Q., Zhou, F., Liu, W.: Bioinspired catecholic chemistry for surface modification. Chem. Soc. Rev. **40**, 4244–4258 (2011)
14. Zhao, H., Sun, C., Stewart, R.J., Waite, J.H.: Cement proteins of the tube-building polychaete *Phragmatopoma californica*. J. Biol. Chem. **280**, 42938–42944 (2005)
15. Stewart, R.J., Wang, C.S., Shao, H.: Complex coacervates as a foundation for synthetic underwater adhesives. Adv. Colloid Interface Sci. **167**, 85–93 (2011)
16. Vreeland, V., Waite, J.H., Epstein, L.: Polyphenols and oxidases in substratum adhesion by marine algae and mussels. J. Phycol. **34**, 1–8 (1998)
17. Bitton, R., Bianco-Peled, H.: Novel biomimetic adhesives based on algae glue. Macromol. Biosci. **8**, 393–400 (2008)
18. Kamino, K.: Underwater adhesive of marine organisms as the vital link between biological science and material science. Mar. Biotechnol. **10**, 111–121 (2008)
19. Kamino, K.: Molecular design of banacle cement in comparison with those of mussel and tubeworm. J. Adhes. **86**, 96–110 (2010)

B

Biomimetics of Optical Nanostructures

Andrew R. Parker
Department of Zoology, The Natural History Museum, London, UK

Synonyms

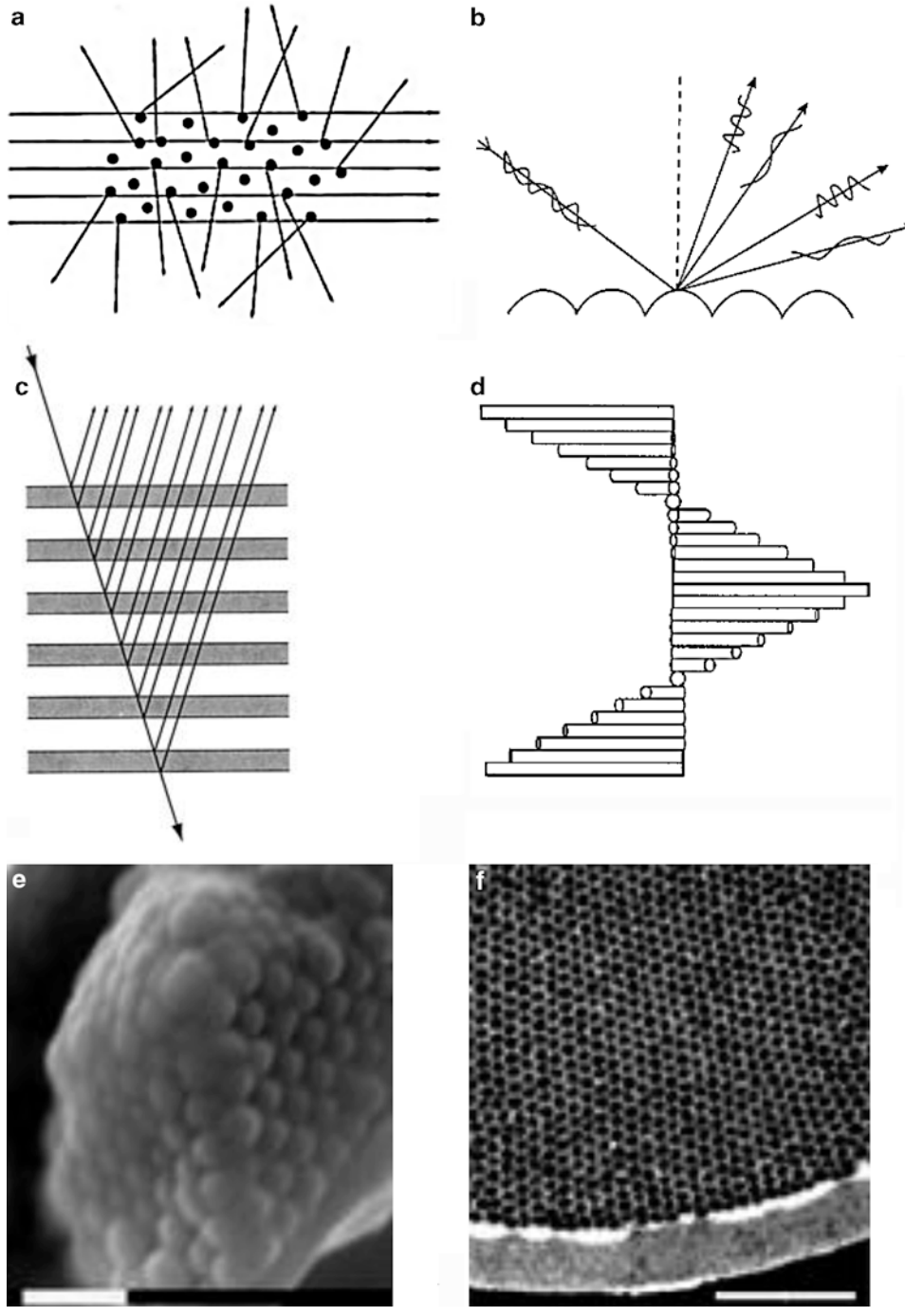
Bio-inspiration; Bio-photonics; Photonics

Definition

The study and industrial application of structural colour and antireflective devices in nature.

Introduction

Three centuries of research, beginning with Hooke and Newton, have revealed a diversity of optical devices at the nanoscale (or at least the submicron scale) in nature [1]. These include structures that cause random scattering, 2D diffraction gratings, 1D multilayer reflectors, and 3D liquid crystals (Fig. 1a–d). In 2001 the first photonic crystal was identified as such in animals [2], and since then, the scientific effort in this subject has accelerated. Now we know of a variety of 2D and 3D photonic crystals in nature (e.g., Fig. 1e, f), including some designs not encountered previously in physics.



Biomimetics of Optical Nanostructures, Fig. 1 Summary of the main types of optical reflectors found in nature: (**a**–**d**) where a light ray is (generally) reflected only once within the system (i.e., they adhere to the single scattering, or first Born approximation) and (**e**, **f**) where each light ray is (generally) reflected multiple times within the system. (**a**) An irregular array of elements that scatter incident light into random directions. The scattered (or reflected) rays do not interfere. (**b**) A diffraction grating, a surface structure, from which incident light is diffracted into specific angular directions resulting in a spatial separation of its angular wavelength/color

Biomimetics is the extraction of good design from nature. Some optical biomimetic successes have resulted from the use of conventional (and constantly advancing) engineering methods to make direct analogues of the reflectors and antireflectors found in nature. However, recent collaborations between biologists, physicists, engineers, chemists, and materials scientists have ventured beyond merely mimicking in the laboratory what happens in nature, leading to a thriving new area of research involving biomimetics via cell culture. Here, the nanoengineering efficiency of living cells is harnessed, and nanostructures such as diatom "shells" can be made for commercial applications via culturing the cells themselves.

Engineering of Antireflectors

Some insects benefit from antireflective surfaces, either on their eyes to see under low-light conditions or on their wings to reduce surface reflections in transparent (camouflaged) areas. Antireflective surfaces, therefore, occur on the corneas of moth and butterfly eyes [3] and on the transparent wings of hawkmoths [4]. These consist of nodules, with rounded tips, arranged in a hexagonal array with a periodicity of around 240 nm (Fig. 2b). Effectively, they introduce a gradual refractive index profile at an interface between chitin (a polysaccharide, often embedded in a proteinaceous matrix; r.i. 1.54) and air and hence reduce reflectivity by a factor of ten.

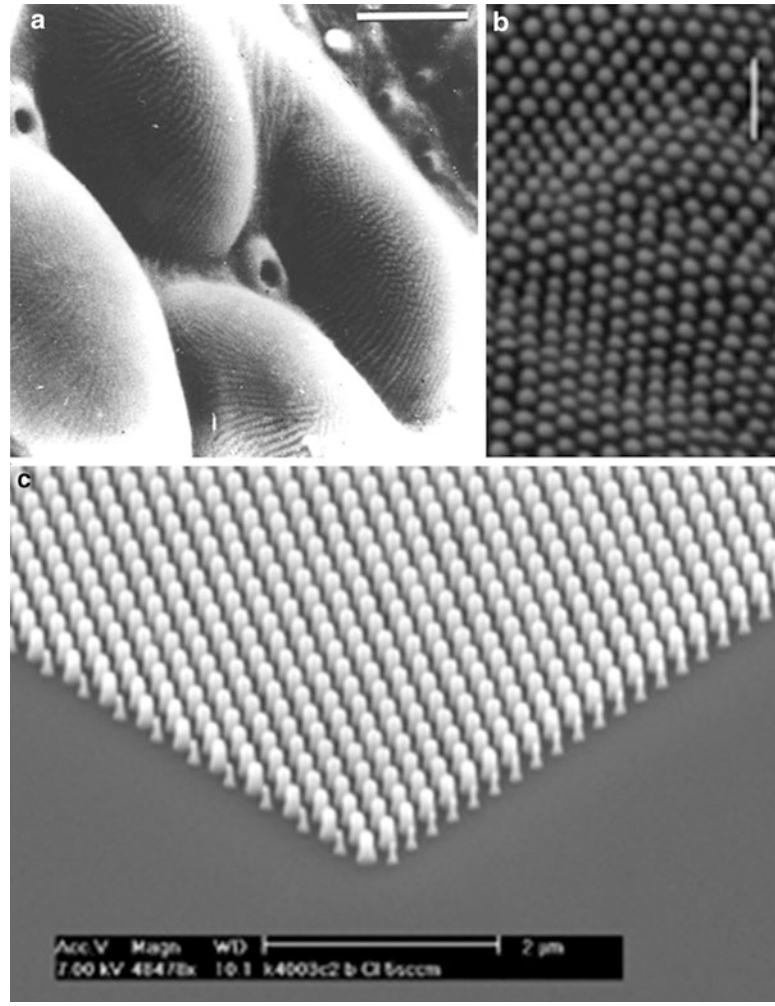
This "moth-eye structure" was first reproduced at its correct scale by crossing three gratings at 120° using lithographic techniques and employed as antireflective surfaces on glass windows in Scandinavia [5]. Here, plastic sheets bearing the antireflector were attached to each interior surface of triple-glazed windows using refractive-index-matching glue to provide a significant difference in reflectivity. Today, the moth-eye structure can be made extremely accurately using e-beam etching [6] and is also employed commercially on solid plastic and other lenses.

A different form of antireflective device, in the form of a sinusoidal grating of 250 nm periodicity, was discovered on the cornea of a 45 million-year-old fly preserved in amber [7] (Fig. 2a). This is particularly useful where light is incident at a range of angles (within a single plane, perpendicular to the grating grooves), as demonstrated by a model made in photoresist using lithographic methods [7]. Consequently, it has been employed on the surfaces of solar panels, providing a 10 % increase in energy capture through reducing the reflected portion of sunlight [8]. Again, this device is embossed onto plastic sheets using holographic techniques.

Engineering of Iridescent Devices

Many birds, insects (particularly butterflies and beetles), fishes, and lesser-known marine animals display iridescent (changing color with angle) and/or "metallic" colored effects resulting from photonic nanostructures. These appear comparatively brighter than the effects of pigments and often function in animals to attract the attention of a potential mate or to startle a predator. An obvious application for such visually attractive and optically sophisticated devices is within the anti-

Biomimetics of Optical Nanostructures, Fig. 1 (continued) components. Each corrugation is about 500 nm wide. Diffracted rays superimpose either constructively or destructively. (**c**) A multilayer reflector, composed of thin (ca. 100 nm thick) layers of alternating refractive index, where light rays reflected from each interface in the system superimpose either constructively or destructively. (**d**) A "liquid crystal" composed of nano-fibers arranged in layers, where the nano-fibers of one layer lie parallel to each other yet are orientated slightly differently to those of adjacent layers. Hence, spiral patterns can be distinguished within the structure. The height of the section shown here – one "period" of the system – is around 200 nm. (**e**) Scanning electron micrograph of the "opal" structure – a close-packed array of submicron spheres (a "3D photonic crystal") – found within a single scale of the weevil *Metapocyrtus* sp.; scale bar = 1 μm. (**f**) Transmission electron micrograph of a section through a hair (neuroseta) of the sea mouse *Aphrodita* sp. (Polychaeta), showing a cross section through a stack of submicron tubes (a "2D photonic crystal"); scale bar = 5 μm

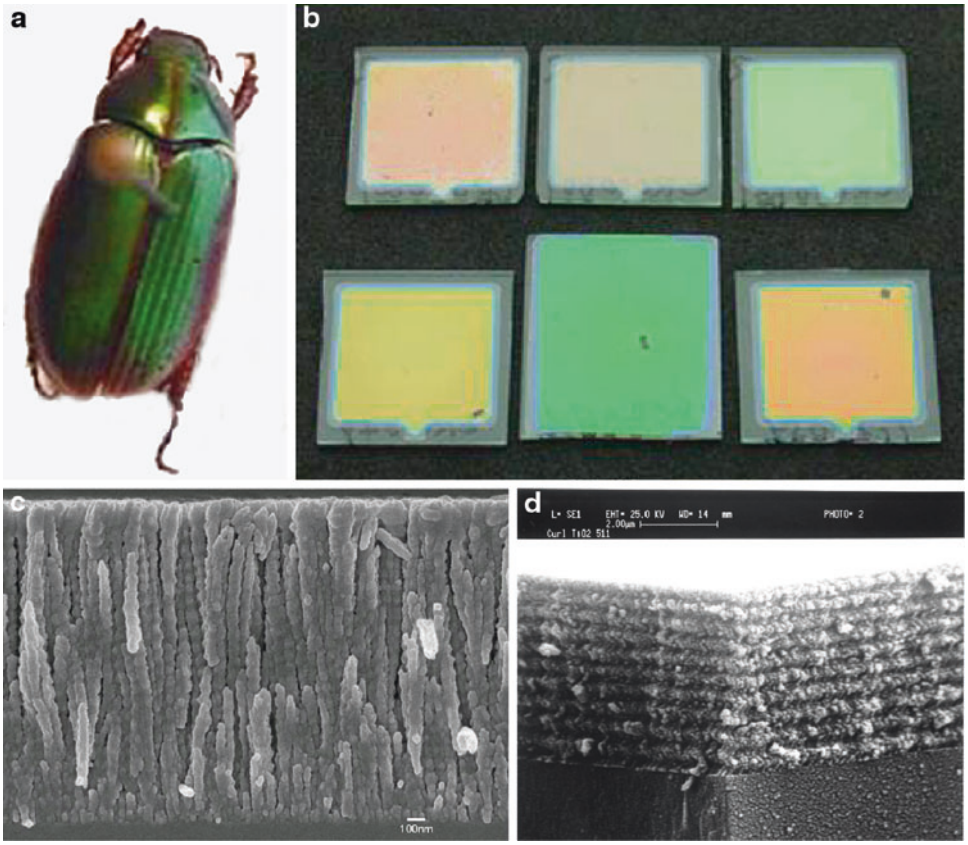


Biomimetics of Optical Nanostructures, Fig. 2 Scanning electron micrographs of antireflective surfaces. (**a**) Fly-eye antireflector (ridges on four facets) on a 45-million-year-old dolichopodid fly's eye. Micrograph by P. Mierzejewski, reproduced with permission. (**b**) Moth-eye antireflective surfaces. (**c**) Moth-eye mimic fabricated using ion beam etching (Micrograph by S. A. Boden and D. M. Bagnall, reproduced with permission. Scale bars = 3 μm (**a**), 1 μm (**b**), 2 μm (**c**))

counterfeiting industry. For secrecy reasons, work in this area cannot be described, although devices are sought at different levels of sophistication, from effects that are discernable by the eye to fine-scale optical characteristics (e.g., polarization and angular properties) that can be read only by specialized detectors. However, new research aims to exploit these devices in the cosmetics, paint, printing/ink, and clothing industries. They are even being tested in art to provide a sophisticated color change effect.

Original work on exploiting nature's reflectors involved copying the design but not the size, where reflectors were scaled up to target longer wavelengths. For example, rapid prototyping was employed to manufacture a microwave analogue of a *Morpho* butterfly scale that is suitable for reflection in the 10–30 GHz region. Here layer thicknesses would be in the order of 1 mm rather than 100 nm as in the butterfly, but the device could be employed as an antenna with broad radiation characteristics or as an antireflection coating for radar. However, today, techniques are available to manufacture nature's reflectors at their true size.

Nanostructures causing iridescence include photonic crystal fibers, opal and inverse opal, and unusually sculpted 3D architectures. Photonic crystals are ordered, often complex, sub-wavelength (nano)lattices that can control the propagation of light at the single wave scale in the manner that atomic crystals control



Biomimetics of Optical Nanostructures, Fig. 3 (**a**) A manuka (scarab) beetle with (**b**) titania mimetic films of slightly different pitches. (**c**) Scanning electron micrograph of the chiral reflector in the beetle's cuticle. Scale bar = 500 nm. (**d**) Scanning electron micrograph of the titania mimetic film. Scale bar = 2 μm (Images by L. DeSilva and I. Hodgkinson, reproduced with permission)

electrons [9]. Examples include opal (a hexagonal or cubic close-packed array of 250 nm spheres) and inverse opal (a hexagonal array of similar-sized holes in a solid matrix). Hummingbird feather barbs contain ultrathin layers with variations in porosity that cause their iridescent effects, due to the alternating nanoporous/fully dense ultrastructure [10]. Such layers have been mimicked using aqueous-based layering techniques [10]. The greatest diversity of 3D architectures can be found in butterfly scales, which can include micro-ribs with nano-ridges, concave multilayered pits, blazed gratings, and randomly punctate nano-layers [11, 12]. The cuticle of many beetles contains structurally chiral films that produce iridescent effects with circular or elliptical polarization properties [13]. These have been replicated in titania for specialized coatings [13], where a mimetic sample can be compared with the model beetle and an accurate variation in spectra with angle is observed (Fig. 3). The titania mimic can be nanoengineered for a wide range of resonant wavelengths; the lowest so far is a pitch of 60 nm for a circular Bragg resonance at 220 nm in a Sc_2O_3 film (Ian Hodgkinson, pers. comm.).

Biomimetic work on the photonic crystal fibers of the *Aphrodita* sea mouse is underway. The sea mouse contains spines (tubes) with walls packed with smaller tubes of 500 nm, with varying internal diameters (400–450 nm). These provide a bandgap in the red region and are to be manufactured via an extrusion technique. Larger glass tubes packed together in the proportion of the spine's nanotubes will be heated and pulled

through a drawing tower until they reach the correct dimensions. The sea mouse fiber mimics will be tested for standard PCF applications (e.g., in telecommunications) but also for anti-counterfeiting structures readable by a detector.

Analogues of the famous blue *Morpho* butterfly (Fig. 4a) scales have been manufactured [14, 15]. Originally, corners were cut. Where the *Morpho* wing contained two layers of scales – one to generate color (a quarter-wave stack) and another above it to scatter the light – the model copied only the principle [14]. The substrate was roughened at the nanoscale and coated with 80 nm thick layers alternating in refractive index [14]. Therefore, the device retained a quarter-wave stack centered in the blue region, but incorporated a degree of randomness to generate scattering. The engineered device closely matched the butterfly wing – the color observed changed only slightly with changing angle over 180°, an effect difficult to achieve and useful for a broad-angle optical filter without dyes.

A new approach to making the 2D "Christmas tree" structure (a vertical, elongated ridge with several layers of 70 nm thick side branches; Fig. 4b) has been achieved using focused-ion-beam chemical-vapor-deposition (FIB-CVD) [15]. By combining the lateral growth mode with ion beam scanning, the Christmas tree structures were made accurately (Fig. 4c). However, this method is not ideal for low-cost mass production of 2D and 3D nanostructures, and therefore, the ion-beam-etched Christmas trees are currently limited to high-cost items including nano- or micron-sized filters (such as "pixels" in a display screen or a filter). Recently, further corners have been cut in manufacturing the complex nanostructures found in many butterfly scales, involving the replication of the scales in ZnO, using the scales themselves as templates [16] (Fig. 4d, e).

Cell Culture

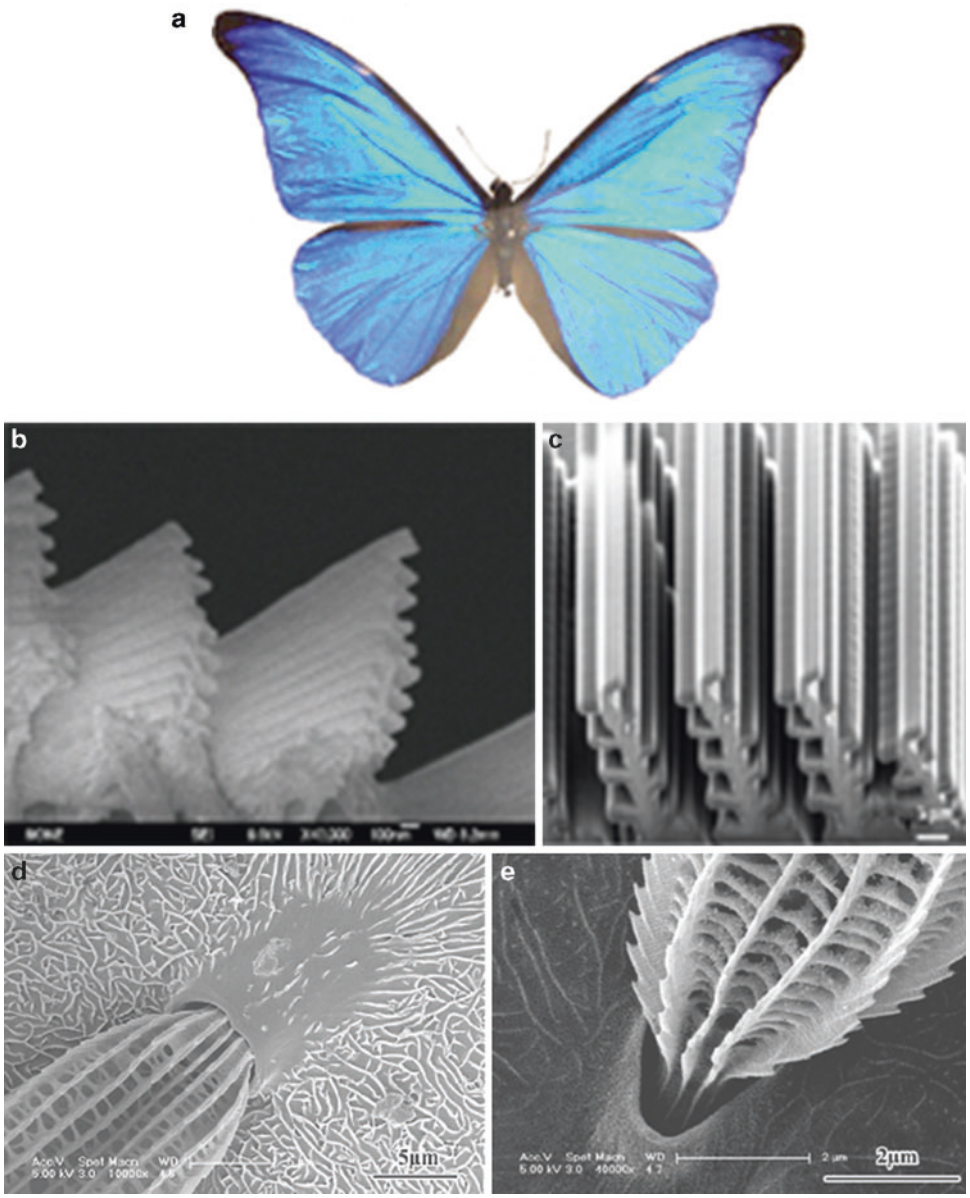
Sometimes nature's optical nanostructures have such an elaborate architecture at such a small (nano)scale that we simply cannot copy them using current engineering techniques. Additionally, sometimes they can be made as individual reflectors (as for the *Morpho* structure), but the effort is so great that commercial-scale manufacture would never be cost effective.

An alternative approach to making nature's reflectors is to exploit an aspect other than design – that the animals or plants can make them efficiently. Therefore, we can let nature manufacture the devices for us via cell culture techniques. Animal cells are in the order of 10 μm in size and plant cells up to about 100 μm and hence suitable for nanostructure production. The success of cell culture depends on the species and on the type of cell from that species. Insect cells, for instance, can be cultured at room temperature, whereas an incubator is required for mammalian cells. Cell culture is not a straightforward method, however, since a culture medium must be established to which the cells adhere, before they can be induced to develop to the stage where they make their photonic devices.

Current work in this area centers on butterfly scales. The cells that make the scales are identified in chrysalises, dissected, and plated out. Then the individual cells are separated, kept alive in culture, and prompted to manufacture scales through the addition of growth hormones. Currently, we have cultured blue *Morpho* butterfly scales in the lab that have identical optical and structural characteristics to natural scales. The cultured scales could be embedded in a polymer or mixed into a paint, where they may float to the surface and self-align. Further work, however, is required to increase the level of scale production and to harvest the scales from laboratory equipment in appropriate ways. A far simpler task emerges where the iridescent organism is single celled.

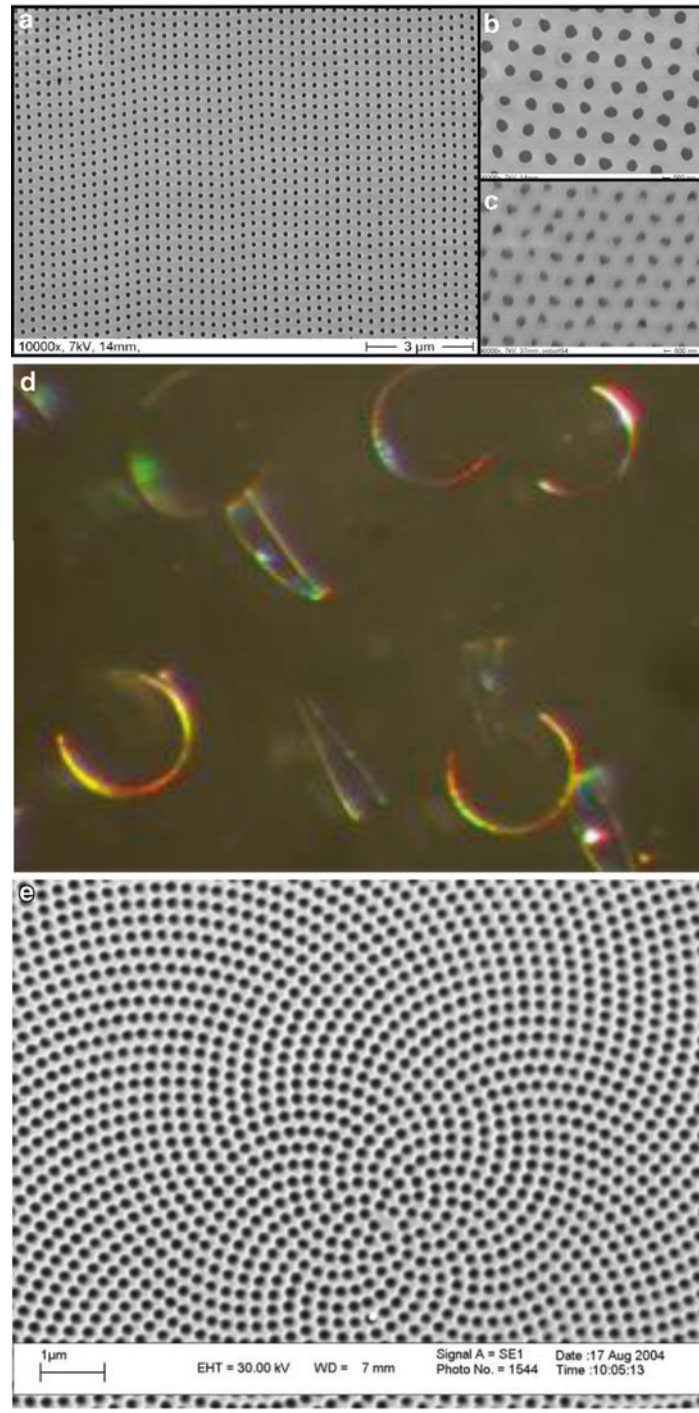
Diatoms and Coccolithophores

Diatoms are unicellular photosynthetic microorganisms. The cell wall is called the frustule and is made of the polysaccharide pectin impregnated with silica. The frustule contains pores



Biomimetics of Optical Nanostructures, Fig. 4 (**a**) A *Morpho* butterfly with (**b**) a scanning electron micrograph of the structure causing the blue reflector in its scales. (**c**) A scanning electron micrograph of the FIB-CVD-fabricated mimic. A Ga^+ ion beam (beam diameter 7 nm at 0.4 pA, 30 kV), held perpendicular to the surface, was used to etch a precursor of phenanthrene ($C_{14}H_{10}$). Both give a wavelength peak at around 440 nm and at the same angle (30°). (**d**) Scanning electron micrograph of the base of a scale of the butterfly *Ideopsis similis*. (**e**) Scanning electron micrograph of a ZnO replica of the same part of the scale in (**d**) ((**a**–**c**) by K. Watanabe and (**d**) and (**e**) by W. Zhang, all reproduced with permission of the authors. Scale bars: (**b**) and (**c**) = 100 nm, (**d**) = 5 μm, (**e**) = 2 μm)

(Fig. 5a–c) and slits which give the protoplasm access to the external environment. There are more than 100,000 different species of diatoms, generally 20–200 μm in diameter or length, but some can be up to 2 mm long. Diatoms have been proposed to build photonic devices directly in 3D [17]. The biological function of the optical property (Fig. 5d) is at present unknown, but may



Biomimetics of Optical Nanostructures, Fig. 5 (**a**–**c**) Scanning electron micrographs of the intercalary band of the frustule from two species of diatoms, showing the square array of pores from *Coscinodiscus granii* (**a**, **b**) and the hexagonal arrays of pores from *C. wailesii* (**c**). These periodic arrays are proposed to act as photonic crystal waveguides. (**d**) Iridescence of the *C. granii* girdle bands. (**e**) Southampton University mimic of a diatom frustule (patented for photonic crystal applications), scanning electron micrograph (By G. Parker, reproduced with permission)

affect light collection by the diatom. This type of photonic device can be made in silicon using a deep photochemical etching technique (initially developed by Lehmann [18]) (e.g., Fig. 5e). However, there is a new potential here since diatoms carry the added advantage of exponential growth in numbers – each individual can give rise to 100 million descendents in a month.

Unlike most manufacturing processes, diatoms achieve a high degree of complexity and hierarchical structure under mild physiological conditions. Importantly, the size of the pores does not scale with the size of the cell, thus maintaining the pattern. Fuhrmann et al. [17] showed that the presence of these pores in the silica cell wall of the diatom *Coscinodiscus granii* means that the frustule can be regarded as a photonic crystal slab waveguide. Furthermore, they present models to show that light may be coupled into the waveguide and give photonic resonances in the visible spectral range.

The silica surface of the diatom is amenable to simple chemical functionalization (e.g., Fig. 6a–c). An interesting example of this uses a DNA-modified diatom template for the control of nanoparticle assembly [19]. Gold particles were coated with DNA complementary to that bound to the surface of the diatom. Subsequently, the gold particles were bound to the diatom surface via the sequence-specific DNA interaction. Using this method, up to seven layers were added showing how a hierarchical structure could be built onto the template.

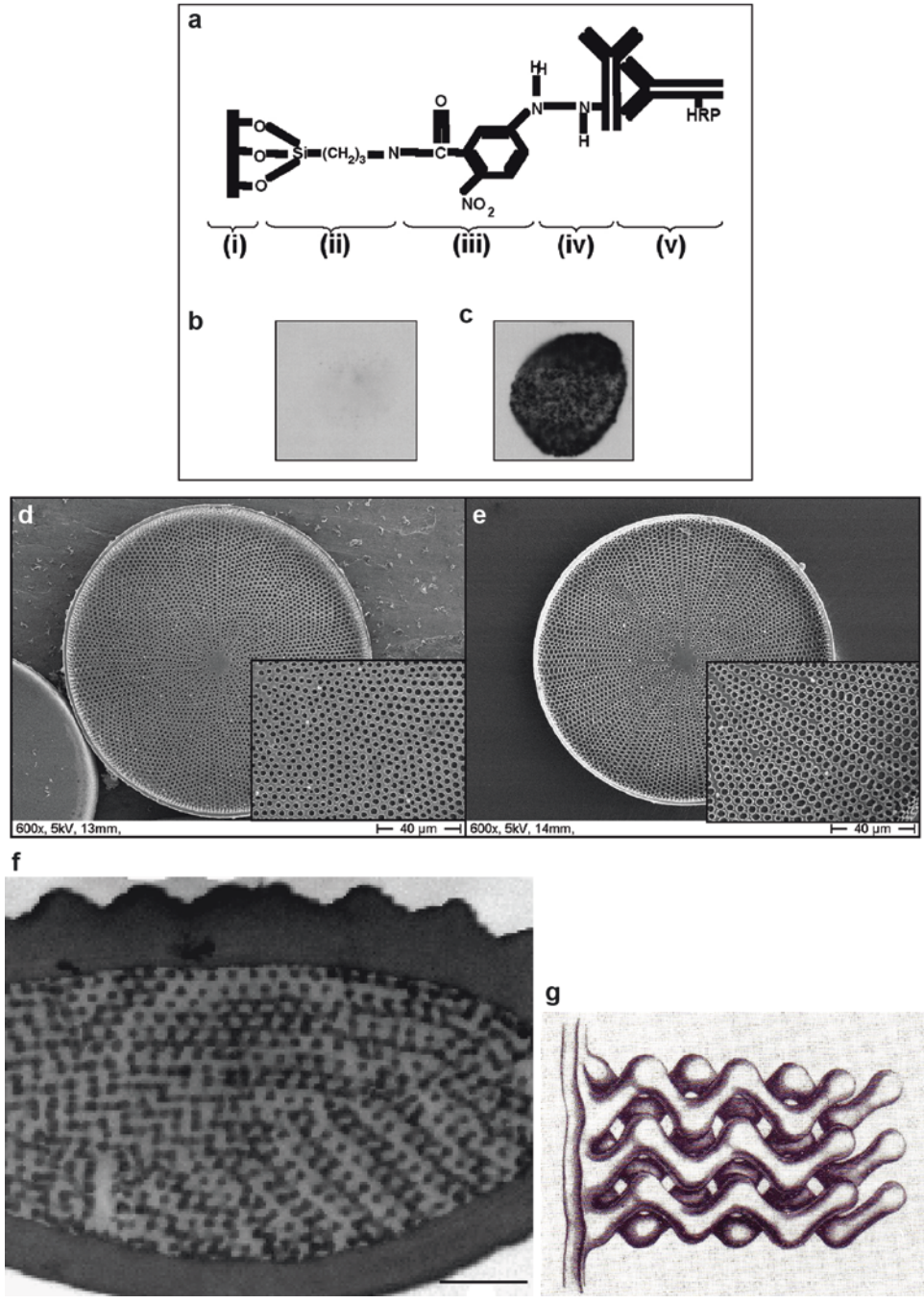
Porous silicon is known to luminesce in the visible region of the spectrum when irradiated with ultraviolet light [20]. This photoluminescence (PL) emission from the silica skeleton of diatoms was exploited by De Stefano [21] in the production of an optical gas sensor. It was shown that the PL of *Thalassiosira rotula* is strongly dependent on the surrounding environment. Both the optical intensity and peaks are affected by gases and organic vapors. Depending on the electronegativity and polarizing ability, some substances quench the luminescence, while others effectively enhance it. In the presence of the gaseous substances NO_2, acetone, and ethanol, the photoluminescence was quenched. This was because these substances attract electrons from the silica skeleton of the diatoms and hence quench the PL. Nucleophiles, such as xylene and pyridine, which donate electrons, had the opposite effect and increased PL intensity almost ten times. Both quenching and enhancements were reversible as soon as the atmosphere was replaced by air.

The silica inherent to diatoms does not provide the optimum chemistry/refractive index for many applications. Sandhage et al. [22] have devised an inorganic molecular conversion reaction that preserves the size, shape, and morphology of the diatom while changing its composition. They perfected a gas/silica displacement reaction to convert biologically derived silica structures such as frustules into new compositions. Magnesium was shown to convert SiO_2 diatoms by a vapor-phase reaction at 900 °C to MgO of identical shape and structure, with a liquid Mg_2Si by-product. Similarly, when diatoms were exposed to titanium fluoride gas, the titanium displaced the silicon, yielding a diatom structure made up entirely of titanium dioxide, a material used in some commercial solar cells.

An alternative route to silica replacement hijacks that native route for silica deposition in vivo. Rorrer et al. [23] sought to incorporate elements such as germanium into the frustule, a semiconductor material that has interesting properties that could be of value in optoelectronics, photonics, thin-film displays, solar cells, and a wide range of electronic devices. Using a two-stage cultivation process, the photosynthetic marine diatom *Nitzschia frustulum* was shown to assimilate soluble germanium and fabricate Si-Ge oxide nanostructured composite materials.

Porous glasses impregnated with organic dye molecules are promising solid media for tunable lasers and nonlinear optical devices, luminescent solar concentrators, gas sensors, and active waveguides. Biogenic porous silica has an open spongelike structure, and its surface is naturally OH terminated. Hildebrand and Palenik [24] have shown that rhodamine B and 6G are able to stain diatom silica in vivo and determined that the dye treatment could survive the harsh acid treatment needed to remove the surface organic layer from the silica frustule.

Now attention is beginning to turn additionally to coccolithophores – single-celled marine algae, also abundant in marine environments. Here, the cell secretes calcitic photonic crystal frustules which, like diatoms, can take a diversity of forms, including complex 3D architectures at the nano- and microscales.



Biomimetics of Optical Nanostructures, Fig. 6 Modification of natural photonic devices. (**a**–**c**) Diatom surface modification. The surface of the diatom was silanized and then treated with a heterobifunctional cross-linker, followed by attachment of an antibody via a primary amine group. (**a**) (i) Diatom exterior surface, (ii) APS (3-aminopropyltrimethoxysilane), (iii) ANB-NOS (*N*-5-azido-2-nitrobenzoyloxysuccinimide), (iv) primary antibody, and (v) secondary antibody with HRP (horseradish peroxide) conjugate. Diatoms treated with primary and secondary antibody with (**b**) no surface modification and (**c**) after surface modification. (**d**), (**e**) Scanning electron micrographs showing the pore pattern of the diatom *Coscinodiscus wailesii* (**d**) and after growth

Iridoviruses

Viruses are infectious particles made up of the viral genome packaged inside a protein capsid. The iridovirus family comprises a diverse array of large (120–300 nm in diameter) viruses with icosahedral symmetry. The viruses replicate in the cytoplasm of insect cells. Within the infected cell, the virus particles produce a paracrystalline array that causes Bragg refraction of light. This property has largely been considered aesthetic to date, but the research group of Vernon Ward (New Zealand), in collaboration with the Biomaterials Laboratory at Wright-Patterson Air Force Base, are using iridoviruses to create biophotonic crystals. These can be used for the control of light, with this laboratory undertaking large-scale virus production and purification as well as targeting manipulation of the surface of iridoviruses for altered crystal properties. These can provide a structural platform for a broad range of optical technologies, ranging from sensors to waveguides.

Virus nanoparticles, specifically *Chilo* and *Wiseana* invertebrate iridoviruses, have been used as building blocks for iridescent nanoparticle assemblies. Here, virus particles were assembled in vitro, yielding films and monoliths with optical iridescence arising from multiple Bragg scattering from close-packed crystalline structures of the iridovirus. Bulk viral assemblies were prepared by centrifugation followed by the addition of glutaraldehyde, a cross-linking agent. Long-range assemblies were prepared by employing a cell design that forced virus assembly within a confined geometry followed by cross-linking. In addition, virus particles were used as core substrates in the fabrication of metallodielectric nanostructures. These comprise a dielectric core surrounded by a metallic shell. More specifically, a gold shell was assembled around the viral core by attaching small gold nanoparticles to the virus surface using inherent chemical functionality of the protein capsid [25]. These gold nanoparticles then acted as nucleation sites for electroless deposition of gold ions from solution. Such nano-shells could be manufactured in large quantities and provide cores with a narrower size distribution and smaller diameters (below 80 nm) than currently used for silica. These investigations demonstrated that direct harvesting of biological structures, rather than biochemical modification of protein sequences, is a viable route to create unique, optically active materials.

The Mechanisms of Natural Engineering and Future Research

Where cell culture is concerned, it is enough to know that cells *do* make optical nanostructures, which can be farmed appropriately. However, in the future an alternative may be to emulate the natural engineering processes ourselves, through reacting the same concentrations of chemicals under the same environmental conditions and possibly substituting analogous nano- or macro-machinery.

To date, the process best studied is the silica cell wall formation in diatoms. The valves are formed by the controlled precipitation of silica within a specialized membrane vesicle called the silica deposition vesicle (SDV). Once inside the SDV, silicic acid is converted into silica particles, each measuring approximately 50 nm in diameter. These then aggregate to form larger blocks of material. Silica deposition is molded into a pattern by the presence of organelles such as mitochondria spaced at regular intervals along the cytoplasmic side of the SDV [26]. These organelles are

Biomimetics of Optical Nanostructures, Fig. 6 (continued) in the presence of nickel sulfate (**e**). Note the enlargement of pores, and hence change in optical properties, in (**e**). (**f**) "Photonic crystal" of the weevil *Metapocyrtus* sp., section through a scale, transmission electron micrograph; scale bar = 1 μm (see Parker [33]). (**g**) A comparatively enlarged diagrammatic example of cell membrane architecture: tubular cristae in mitochondria from the chloride cell of sardine larvae (From Threadgold [37]). Evidence suggests that preexisting internal cell structures play a role in the manufacture of natural nanostructures; if these can be altered then so will the nanostructure made by the cell

thought to physically restrict the targeting of silica from the cytoplasm, to ensure laying down of a correctly patterned structure. This process is very fast, presumably due to optimal reaction conditions for the synthesis of amorphous solid silica. Tight structural control results in the final species-specific, intricate exoskeleton morphology.

The mechanism whereby diatoms use intracellular components to dictate the final pattern of the frustule may provide a route for directed evolution. Alterations in the cytoplasmic morphology of *Skeletonema costatum* have been observed in cells grown in sublethal concentrations of mercury and zinc [27], resulting in swollen organelles, dilated membranes, and vacuolated cytoplasm. Frustule abnormalities have also been reported in *Nitzschia liebethrutti* grown in the presence of mercury and tin [28]. Both metals resulted in a reduction in the length to width ratios of the diatoms and fused pores and a reduction in the number of pores per frustule. These abnormalities were thought to arise from enzyme disruption either at the silica deposition site or at the nuclear level. We grew *Coscinodiscus wailesii* in sublethal concentrations of nickel and observed an increase in the size of the pores (Fig. 6d, e) and a change in the phospholuminescent properties of the frustule. Here, the diatom can be "made to measure" for distinct applications such as stimuli-specific sensors.

Further, *trans*-Golgi-derived vesicles are known to manufacture the coccolithophore 3D "photonic crystals" [29]. So the organelles within the cell appear to have exact control of (photonic) crystal growth ($CaCO_2$ in the coccolithophores) and packing (SiO_3 in the diatoms) [30, 31]. Indeed, Ghiradella [9] suggested that the employment of preexisting, intracellular structures lies behind the development of some butterfly scales, and Overton [32] reported the action of microtubules and microfibrils during butterfly scale morphogenesis. Further evidence has been found to suggest that these mechanisms, involving the use of molds and nanomachinery (e.g., Fig. 6f, g), reoccur with unrelated species, indicating that the basic "eukaryote" (containing a nucleus) cell can make complex photonic nanostructures with minimal genetic mutation [33]. The ultimate goal in the field of optical biomimetics, therefore, could be to replicate such machinery and provide conditions under which, if the correct ingredients are supplied, the optical nanostructures will self-assemble with precision.

For further information on the evolution of optical devices in nature, including those found in fossils, or when they first appeared on earth, see references [34, 35], and for information on cell culture in a *Morpho* butterfly, see reference [36].

Cross-References

- Bio-Optics
- Bio-Photonics
- Exoskeleton
- Nanomaterials
- Self-Assembly of Nanostructures
- Structural Colors

References

1. Parker, A.R.: 515 Million years of structural colour. J. Opt. A. **2**, R15–R28 (2000)
2. Parker, A.R., McPhedran, R.C., McKenzie, D.R., Botten, L.C., Nicorovici, N.-A.P.: Aphrodite's iridescence. Nature **409**, 36–37 (2001)
3. Miller, W.H., Moller, A.R., Bernhard, C.G.: The corneal nipple array. In: Bernhard, C.G. (ed.) The Functional Organisation of the Compound Eye, pp. 21–33. Pergamon Press, Oxford (1966)
4. Yoshida, A., Motoyama, M., Kosaku, A., Miyamoto, K.: Antireflective nanoprotuberance array in the transparent wing of a hawkmoth *Cephonodes hylas*. Zool. Sci. **14**, 737–741 (1997)
5. Gale, M.: Diffraction, beauty and commerce. Phys. World **2**, 24–28 (1989)
6. Boden, S.A., Bagnall, D.M.: Biomimetic subwavelength surfaces for near-zero reflection sunrise to sunset. In: Proceedings of the 4th World Conference on Photovoltaic Energy, Conversion, Hawaii (2006)
7. Parker, A.R., Hegedus, Z., Watts, R.A.: Solar-absorber type antireflector on the eye of an Eocene fly (45 Ma). Proc. R. Soc. Lond. B **265**, 811–815 (1998)
8. Beale, B.: Fly eye on the prize. The Bulletin. 46–48 (1999)
9. Yablonovitch, E.: Liquid versus photonic crystals. Nature **401**, 539–541 (1999)
10. Cohen, R.E., Zhai, L., Nolte, A., Rubner, M.F.: pH gated porosity transitions of polyelectrolyte

multilayers in confined geometries and their applications as tunable Bragg reflectors. Macromolecules **37**, 6113 (2004)
11. Ghiradella, H.: Structure and development of iridescent butterfly scales: lattices and laminae. J. Morph. **202**, 69–88 (1989)
12. Berthier, S.: Les coulers des papillons ou l'imperative beauté. Proprietes optiques des ailes de papillons, p. 142. Springer, Paris (2005)
13. DeSilva, L., Hodgkinson, I., Murray, P., Wu, Q., Arnold, M., Leader, J., Mcnaughton, A.: Natural and nanoengineered chiral reflectors: structural colour of manuka beetles and titania coatings. Electromagnetics **25**, 391–408 (2005)
14. Kinoshita, S., Yoshioka, S., Fujii, Y., Okamoto, N.: Photophysics of structural color in the *Morpho* butterflies. Forma **17**, 103 (2002)
15. Watanabe, K., Hoshino, T., Kanda, K., Haruyama, Y., Matsui, S.: Brilliant blue observation from a *Morpho*-butterfly-scale quasi-structure. Jap. J. Appl. Phys. **44**, L48–L50 (2005)
16. Zhang, W., Zhang, D., Fan, T., Ding, J., Gu, J., Guo, Q., Ogawa, H.: Bio-mimetic zinc oxide replica with structural color using butterfly (*Ideopsis similis*) wings as templates. Bioinspir. Biomim. **1**, 89–95 (2006)
17. Fuhrmann, T., Lanwehr, S., El Rharbi-Kucki, M., Sumper, M.: Diatoms as living photonic crystals. Appl. Phys. B **78**, 257–260 (2004)
18. Lehmann, V.: On the origin of electrochemical oscillations at silicon electrodes. J. Electrochem. Soc. **143**, 1313 (1993)
19. Rosi, N.L., Thaxton, C.S., Mirkin, C.A.: Control of nanoparticle assembly by using DNA-modified diatom templates. Angew. Chem. Int. Ed. **43**, 5500–5503 (2004)
20. Cullis, A.G., Canham, L.T., Calcott, P.D.J.: The structural and luminescence properties of porous silicon. J. Appl. Phys. **82**, 909–965 (1997)
21. De Stefano, L., Rendina, I., De Stefano, M., Bismuto, A., Maddalena, P.: Marine diatoms as optical chemical sensors. Appl. Phys. Let. **87**, 233902 (2005)
22. Sandhage, K.H., Dickerson, M.B., Huseman, P.M., Caranna, M.A., Clifton, J.D., Bull, T.A., Heibel, T.J., Overton, W.R., Schoenwaelder, M.E.A.: Novel, bioclastic route to self-assembled, 3D, chemically tailored meso/nanostructures: shape-preserving reactive conversion of biosilica (diatom) microshells. Adv. Mater. **14**, 429–433 (2002)
23. Rorrer, G.L., Chang, C.H., Liu, S.H., Jeffryes, C., Jiao, J., Hedberg, J.A.: Biosynthesis of silicon-germanium oxide nanocomposites by the marine diatom *Nitzschia* frustulum. J. Nanosci. Nanotechnol. **5**, 41–49 (2004)
24. Hildebrand, M., Palenik, B.: Grant report 'Investigation into the Optical Properties of Nanostructured Silica from Diatoms' *Storming Media. Pentagon report no. A266314M* 1–7 (2003)
25. Radloff, C., Vaia, R.A., Brunton, J., Bouwer, G.T., Ward, V.K.: Metal nanoshell assembly on a virus bioscaffold. Nano Lett. **5**, 1187–1191 (2005)
26. Schmid, A.M.M.: Aspects of morphogenesis and function of diatom cell walls with implications for taxonomy. Protoplasma **181**, 43–60 (1994)
27. Smith, M.A.: The effect of heavy metals on the cytoplasmic fine structure of *Skeletonema costatum* (Bacillariophyta). Protoplasma **116**, 14–23 (1983)
28. Saboski, E.: Effects of mercury and tin on frustular ultrastructure of the marine diatom *Nitzschia liebethrutti*. Water Air Soil Pollut. **8**, 461–466 (1977)
29. Corstjens, P.L.A.M., Gonzales, E.L.: Effects of nitrogen and phosphorus availability on the expression of the coccolith-vesicle v-ATPase (subunit C) of *Pleurochrysis* (Haptophyta). J. Phycol. **40**, 82–87 (2004)
30. Klaveness, D., Paasche E.: Physiology of coccolithophorids. In: Biochemistry and Physiology of Protozoa, vol. 1, 2nd edn. Academic Press, London (1979)
31. Klaveness, D., Guillard, R.R.L.: The requirement for silicon in *Synura petersenii* (Chrysophyceae). J. Phycol. **11**, 349–355 (1975)
32. Overton, J.: Microtubules and microfibrils in morphogenesis of the scale cells of *Ephestia kuhniella*. J. Cell Biol. **29**, 293–305 (1966)
33. Parker, A.R.: Conservative photonic crystals imply indirect transcription from genotype to phenotype. Rec. Res. Develop. Entomol. **5**, 1–10 (2006)
34. Parker, A.R.: In the Blink of an Eye, p. 316. Simon & Schuster/Perseus Press, London/Cambridge, MA (2003)
35. Parker, A.R.: A geological history of reflecting optics. J. R. Soc. Lond Interface **2**, 1–17 (2005)
36. Parker, A.R., Townley, H.E.: Making photonic structures via cell culture: *Morpho* butterfly scales. Bioinspl. Biomim. Bionano. **3**, 1–5 (2014)
37. Threadgold, L.T.: The Ultrastructure of the Animal Cell, p. 313. Pergamon press, Oxford (1967)

Biomimicked Microneedles

► Biomimetic Mosquito-Like Microneedles

Biomimicry

► Biomimetics

Biomolecular Mechanics

► Mechanical Properties of Hierarchical Protein Materials

Biomolecules

▶ DNA Manipulation Based on Nanotweezers

Bionic Ear

▶ Bioinspired CMOS Cochlea

Bionic Eye

▶ Artificial Retina: Focus on Clinical and Fabrication Considerations

Bionic Microneedles

▶ Biomimetic Mosquito-Like Microneedles

Bionics

▶ Biomimetics

Bio-optics

▶ Structural Color in Animals

Biopatterning

▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Bio-photonics

▶ Biomimetics of Optical Nanostructures
▶ Structural Color in Animals

Biophysics

▶ DNA Manipulation Based on Nanotweezers

Bioprobes

▶ Biosensors

Biornametics: Architecture Defined by Natural Patterns

Ille C. Gebeshuber[1], Petra Gruber[2] and Barbara Imhof[3]
[1]Institute of Applied Physics, Vienna University of Technology, Wien, Austria
[2]Transarch – Biomimetics and Transdisciplinary Architecture, Vienna, Austria
[3]LIQUIFER Systems Group, Wien, Austria

Synonyms

Arts-based research; Biomimetics; Lateral thinking; Transdisciplinary approaches

Definition

Biornametics – Architecture Defined by Natural Patterns – is an emerging contemporary design practice that explores a new methodology to interconnect scientific evidence with creative design in the field of architecture. The word biornametics is generated from "ornament," referring to the famous Austrian architect Adolf Loos, and "biomimetics," the abstraction of good design from living nature (see "▶ Biomimetics").

Overview

In biornametics research, role models from nature, static and dynamic patterns (e.g., nanostructured surfaces or materials with functional hierarchy

from the nano- to the macroscale), are investigated, and the findings are applied to design strategies. The emergence of patterns in nature at all scales of existence of organisms as one of the most important signs of life – order – is not arbitrary but highly interconnected with boundary conditions, functional requirements, systems requirements, material, and structure. The three main areas of investigation for role models in biornametics are, firstly, **surface patterns, nanotextured surfaces, and nanostructured materials**; secondly, **shape, growth, and deployable structures;** and, thirdly, **adaptation and reorganization**. Biological building strategies rely basically on repetition, variation, and self-similarity. Often simple building blocks are arranged with molecular precision and thus achieve diverse and highly specialized material properties.

Biornametics research aims at understanding the functionality of these natural patterns by extracting the principles found in current nanotechnology research and transferring these principles to an architectural interpretation. Colors are just one very important example. In contrast to pigment colors, structural colors that are found on some butterfly wings and beetles and even on plants [1–3] are primarily determined by the geometry of the underlying material. Interesting is also the generation of these surfaces and materials, as well as multifunctional properties such as durability, degradation, or self-repair. Further examples include plant-environment interactions such as the pitcher plant that lures animals onto a super-sliding surface or the lotus leaf's self-cleaning properties and nanostructured composite materials with high toughness such as the abalone shell. The patterns in these role models from living nature fulfill their purpose and are surprisingly elegant and appeal to the aesthetic dimension of the human perception (Fig. 1). The transfer of surface patterning to architectural elements may deliver added or integrated functionality or reinterpret specific functions on another scale.

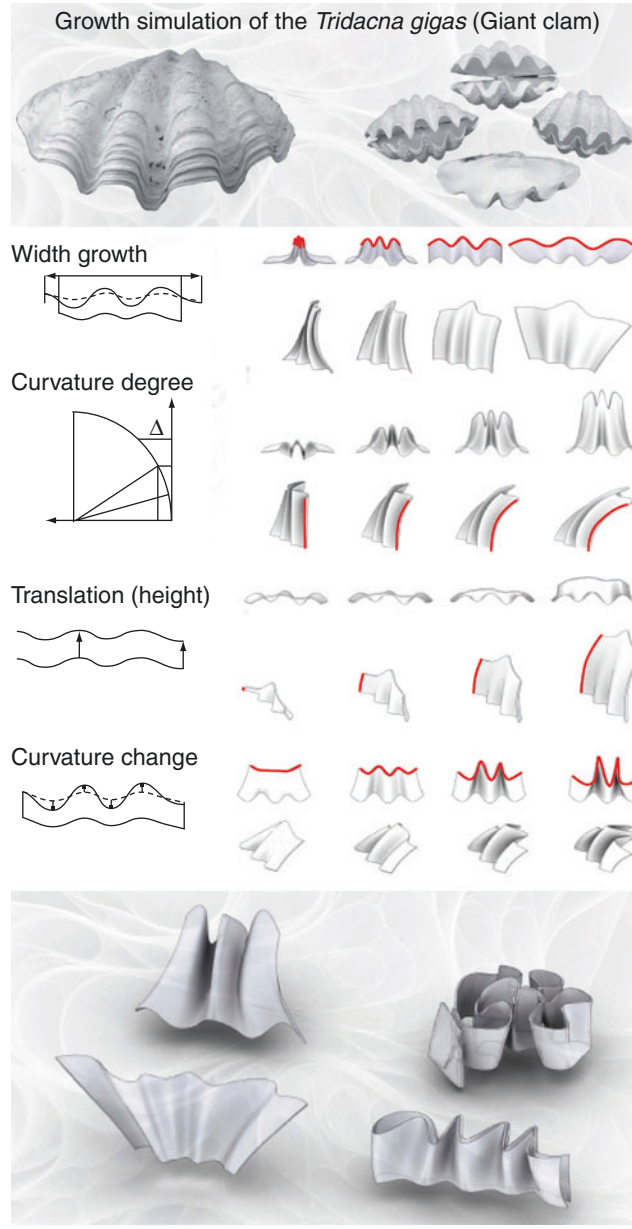
Summary

Biornametics – Architecture Defined by Natural Patterns – is a transdisciplinary arts-based design practice that explores different approaches to a new design methodology to interconnect scientific evidence from the life sciences with creative design in the field of architecture. The starting point of *biornametics* takes on the history of one of the composed parts of the word "ornament" referring to Adolf Loos [4] and extends into another "biomimetics [bionik in German]" [5–7].

Biornametics especially refers to the more recent terminology where the "New Ornament" as an emerging contemporary design practice is based on digital techniques. It is less concerned with serial rationality but with algorithmic, digital operations and connecting the processes of planning and production. The second part of the artificial word biornametics stems from biomimetics and means the strategic search for nature's solutions. The hypothesis underlying this strategy is that living nature has evolved in a process of continuing adaptation to become a complex changing environment and that the exploitation of highly optimized solutions is likely to deliver innovations that provide more intelligence and higher efficiency than our standard methods.

Biornametics research and the simulation of role models from nature, static and dynamic patterns (e.g., growth principles, movement patterns, adaptation and differentiation as key for emergence of patterns, etc.), provide the basics for subsequent applications in art. The emergence of patterns in nature at all scales of existence of organisms as one of the most important signs of life – order – is not arbitrary but highly interconnected with boundary conditions, functional and systems requirements, materials, and structure. The findings from role models and their simulation are applied to design strategies and translated into prototypical architectural concepts.

As a result of the first ever biornametics research cycle, performed between 2010 and 2013, a physical installation was set up temporarily in the Silver Gallery of the University of Applied Arts in Vienna, Austria, demonstrated one of the unlimited possibilities of interpretation of a single principle, and was thus closely related to the perception of the executing team. The goal was to build a lightweight and cost-efficient



Biornametics: Architecture Defined by Natural Patterns, Fig. 1 Ornaments for architecture inspired by a growth simulation of the giant clam, starting from the initial nanocrystalline composite, followed by the formation of a two-dimensional disk protrusion and the formation of crystalline curls

spatial structure with an inherent potential to move, or to be moved, by stimuli from the external environment.

A paramount step in biornametics research is panel discussions with panelists from different backgrounds (such as architecture, mathematics, engineering, physics, and biomimetics) to produce more aspects of methodology and scientific input with feedback from the interested general public. The aim of such explorative debates is to present different approaches to science-based design and to reason about the value of incorporating findings from the life sciences into architectural design methodologies. Springer Scientific Publishing published results from the first biornametics research cycles in a book that takes this approach further and creates the context of the current discussion of arts-based research [8].

Future continuous research can be deducted from the potential that was discovered and lies in the theme of growth and how one can make use of the functions the growing element offers during different phases and the topic of nanotechnology in building and creating in a yet rather unknown scale.

Background

The first biornametics research cycle started in 2010. The Austrian Science Fund installed in the year 2009 a funding program for arts-based research, called PEEK. In this program any person engaged in arts-based research who has the necessary qualifications can apply for funding for up to 36 months. PEEK goals are to support high-quality and innovative arts-based research in which artistic practice is integral to the inquiry. In this way, research capacity, quality, and international standing of arts-based researchers in Austria shall be increased. Further benefits are an increase of both public awareness and awareness within the academic and the arts communities of arts-based research and its potential applications.

The team around architect and artist Barbara Imhof was one of the first to secure PEEK funding: for the first ever cycle of biornametics research, architecture defined by natural patterns. Biornametics research combines scientists, architects, and artists in a highly creative, innovative team.

Host research institution of this first cycle of biornametics research was the University of Applied Arts, Vienna, Austria; key people were Barbara Imhof and Petra Gruber. Collaborating institutions were located in Austria, Germany, Italy, the United Kingdom, and Malaysia. The duration of this first cycle of biornametics research was from May 2010 until March 2013.

Aims of Biornametics Research

The intellectual aim of biornametics arts-based research is the exploration of an aesthetic and functional interpretation together with utilization of new manufacturing technologies and elaboration of biomimetic design methods and the New Ornament.

1:1 implementation is used to communicate the approach and its theme as test bed experimenting the physical translation process and experience the implications arising from unexpected solutions or developments.

Aims of biornametics arts-based research:

- Research, analysis, and simulation of dynamic patterns from nature.
- Patterns from nature are investigated and deep principles behind morphogenesis defined. Analogies in architectural design are identified to clarify application possibilities.
- Successful transfer of found principles to architectural application shall deliver intelligent, efficient, and ecological designs which also will create new forms and new aesthetics.
- Deliver visualization and prototypes of the architectural interpretation.
- Application to an architectural scenario is crucial for the success of transfer. The development of a real project manifests the validity of the approach.
- Further development and refinement of the biornametics methodology to transfer nature's patterns to architectural application.
- The development of a methodology for the transfer of information into the architectural discipline is important as a pathway for follow-up studies. The biomimetic method goes beyond superficial interpretations of findings from nature.
- Serve as a role model for interdisciplinary architectural design work.

Integration of findings from other disciplines such as the life sciences is not an easy task. The new design approach needs a formal framework that allows dynamic exchange of information. Preconditions for this exchange are the willingness to cross the borders between disciplines and to deal with other "languages" (terminologies) and cultures [9]. By involving different researchers and consultants with various backgrounds such as nanotechnology, biomimetics,

informatics, mathematics, structural engineering, architecture, and design, a multidisciplinary team is created that investigates and develops novel design strategies for architecture with an exemplary application structure.

Limitations

The integration of new production technologies to the research approach can only be achieved in follow-up attempts (e.g., subsequent projects together with industry). The main reason for this is to keep the simulations and computational results as broad and systemic as possible, to allow for potential application developments in a wide variety of fields. Collaboration with industry will supply the necessary production workspaces and machines.

Effects of the Design Practice Outside the Core Field of Research

Biornametics research is cross-disciplinary work based on a scientific approach to architectural design; it is a new teaching method with high potential to strengthen innovation in architecture. A continuously growing database of potential role models and principles provides the basis for further more specific follow-up arts-based research approaches that explore the applicability in the field of architecture. One more general aspect that shows great potential to be developed further is the integration of new production technologies into the process.

Research Areas

The research areas for biornametics research refer to a main issue in biomimetics: the relation between structure and function. Three initial main areas of investigation for role models from nature were defined in the first research cycle:

1. Surface patterns, nanotextured surfaces, and nanostructured materials
2. Shape, growth, and deployable structures
3. Adaptation and reorganization

These three areas reflect the specific scientific expertise in the scientific team of the first biornametics research cycle: expert consultant George Jeronimidis is versatile in biomimetics and fiber structures, and therefore the role models focus rather on plants than on animals, and Ille C. Gebeshuber is an expert in nanotextured surfaces and structural colors.

In biornametics research the sample role models from nature to be investigated need to be chosen with regard to:

- Availability of reliable information
- Tangibility of scales and processes
- Exclusion of biochemistry and metabolic processes
- Expected innovative potential for spatial and architectural applications

The methodology starts with a selection of role models from nature ("scientific input"). At the same time a primary investigation in successful transfers, architectural applications, and outcome scenarios delivers exemplary knowledge and guarantees novelty of the selected approach.

The base of biornametics research, patterns and their scientific exploration, partly finds a feedback in the architectural theory of the ornament. After the first phase of data collection starts the computation of models, followed by concept development and design.

The first transfer generally takes place when the selected role model patterns are analyzed regarding their potential application and their abstracted principles. Digital simulations are used to investigate the principles. In the second transfer analogy is established by exploring applications where the abstracted principles could be applied.

The final step is the proto-architectural implementation of showcase applications to be exhibited and communicated to the scientific and architectural community by means of publications and scientific papers in journals and conference proceedings.

Surface Patterns, Nanotextured Surfaces, and Nanostructured Materials

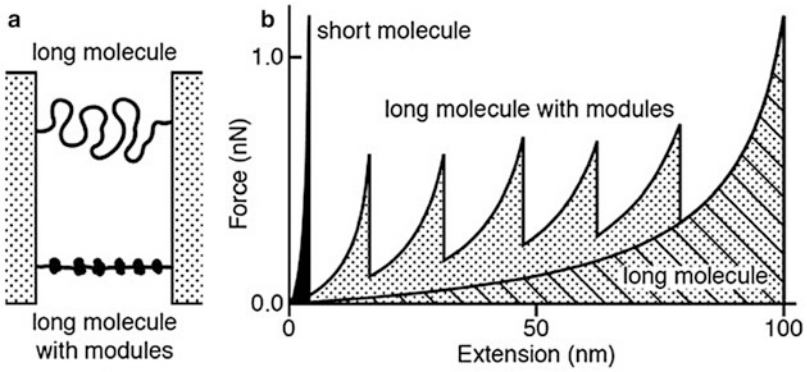
Biological systems exhibit a wealth of functional units highly optimized for a range of parameters, also on the nanoscale, by only combining a handful of starting materials. Biological building strategies rely basically on repetition, variation, and self-similarity. Often simple building blocks are arranged with molecular precision and thus achieve diverse and highly specialized material properties.

Colors are just one very important example. In contrast to pigment colors, physical colors that are found on some butterfly wings and beetles are primarily determined by the geometry of the underlying material. The process of generation of these surfaces and materials is also interesting, as well as their properties such as durability, degradation, or self-repair.

Further examples include plant-environment interaction, for example, the pitcher plant that lures animals onto a super-sliding surface or the well-known self-cleaning principle that was discovered in the lotus leaf. Patterning on the nanoscale also produces materials that have unequaled properties, for example, the abalone shell ([10]; Fig. 2).

The patterns found do not only fulfill their purpose but are surprisingly elegant and appeal to the aesthetic dimension of the human perception. The transfer of surface patterning to architectural elements may deliver added or integrated functionality or reinterpret specific functions on another scale.

Biornametics research role models with relevance to surface patterns, nanotextured surfaces, and nanostructured materials are photonic crystals in the jewel beetle (Fig. 3); super-sliding nanostructure-based slippery surfaces of the yellow pitcher plant (Fig. 4); multifunctional nanostructures on the *Morpho* butterfly (Fig. 5), providing coloration, tuned wettability, unidirectional water runoff, and further nanoscale-based properties; superhydrophobic surfaces of the giant Amazon water lily (Fig. 6); complex nanocomposite of the abalone shell (Fig. 7), with high fracture resistance and iridescent colorations (abalone parts are used in jewelry, for their color play and durability); opal-like photonic crystals of the peacock feather (with exquisite control of the color-generating nanostructures for the various colors, Fig. 8); the octopus (Fig. 9), who can quickly and reversibly change color; the *Strelitzia* plant (Fig. 10) with impressive superhydrophobic properties; and, last but not least, insect wings (Fig. 11) and more generally surfaces of insects with various functionalities based on nanoscale organization of the material.



Biornametics: Architecture Defined by Natural Patterns, Fig. 2 Concept for an adhesive [10]. (**a**) Two ways to attach two particles: with a long molecule or with a long molecule with nodules. (**b**) When stretched, a short molecule can only be extended a little and would then break. A long molecule would be stretched much more and finally break. However, a long molecule with nodules, with sacrificial bonds that break before the backbone of the molecule breaks, increases the toughness of the adhesive. Such a strategy is applied in the abalone shell and also in diatom adhesives (Reprinted by permission from Macmillan Publishers Ltd: Nature, Smith et al. [10], copyright (1999)

Shape, Growth, and Deployable Structures

The topic of morphogenesis in nature is about the development of shapes in general. In the context of biornametics research, the interest lies in the dynamics of shapes and shape change. The topic focuses on the ontogenetic development of three-dimensional complex shapes (in contrast to an evolutionary perspective) and on other phenomena related to shape change in organisms.



Biornametics: Architecture Defined by Natural Patterns, Fig. 3 Jewel beetles are living jewels. Their cuticle is organized similar to precious opal, a highly prized iridescent gemstone. The brilliant coloration originates from photonic crystals, composed of spheres of a couple of hundreds of nanometers in diameter, arranged in hexagonal densest packing, acting as a Bragg reflector for visible light

The development of organisms is based on cell division, the basis for the generation of tissues and organs. The principles of cell growth as investigated in microbiology can be included, but the focus of the topic lies on the research findings in developmental biology.

Interesting issues are the principles of growth in organisms and the differentiation of tissues and materials [11]. The time-based rules of growth and the spatial geometric definition of growth principles (growth patterns) are of specific importance. Specific topics, for example, branching, should be included.

Especially in the plant realm, the relation between growth and deployment is of interest. On a shorter timescale, fast deployment promises to yield significant biornametics research results or, more general, fast shape change in organisms, which aims at, for example, defense or attraction purposes.

The research of mechanisms leading to shapes and shape change in nature shall go beyond the generation and control of complex geometries by establishing a strong link to efficiencies and functionality. This is also a focus in current architectural research and development.

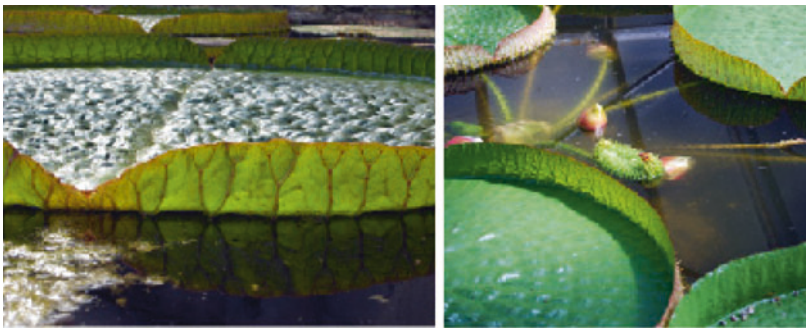
Biornametics research role models with relevance to shape, growth, and deployable structures are the *Morpho* butterfly (Fig. 5) with



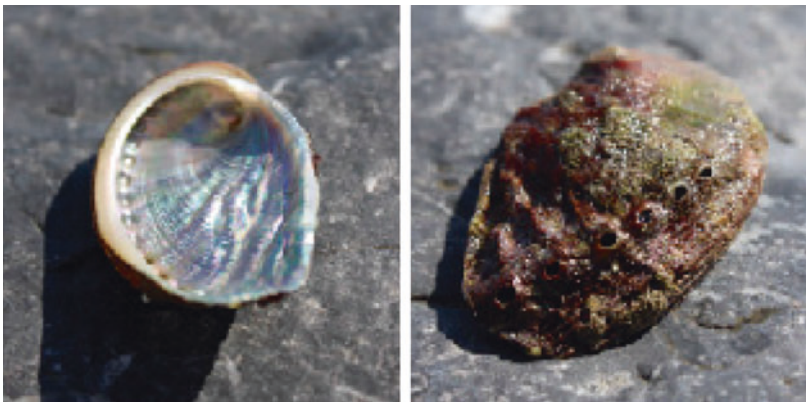
Biornametics: Architecture Defined by Natural Patterns, Fig. 4 Nanoscale structures are the physical basis for the super-slippery inner surfaces of pitcher plants



Biornametics: Architecture Defined by Natural Patterns, Fig. 5 Multifunctional nanostructures on the *Morpho* butterfly and on various other butterflies provide brilliant coloration, temperature regulation, water runoff paths (unidirectional), and much more



Biornametics: Architecture Defined by Natural Patterns, Fig. 6 The giant Amazon water lily grows up to 2.5 m in diameter. Hierarchical micro- and nanostructures provide exquisite self-cleaning properties

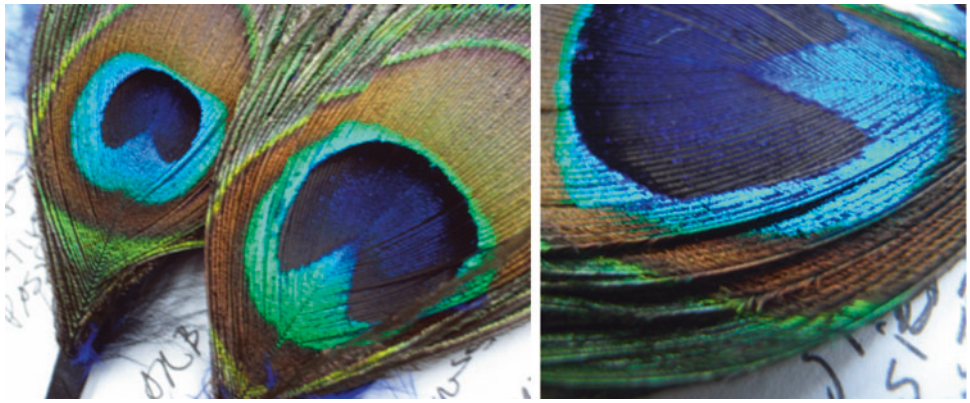


Biornametics: Architecture Defined by Natural Patterns, Fig. 7 The abalone shell is a fascinating natural nanocomposite made from calcite and some % of adhesive protein matrix. The fracture toughness of the composite is 1000 times larger than the one of a single calcite crystal

interesting nanoscale-based biomechanics; the giant Amazon water lily (Fig. 6) which can grow up to 2.5 m in diameter, supported by functionalities based on nanostructures, nanomechanics, and nanochemistry; the abalone shell (Fig. 7), in which the self-assembly of its toughness-providing complex nanocomposite of protein matrix and calcite bricks can inspire in situ growth of building materials, using, for example, CO_2 from the air and water, similar to various further biomineralizing organisms [12]; and spider silk (Fig. 12), forming strong, tough, large, mechanically stable lightweight constructions.

Adaptation and Reorganization

The topic Adaptation and Reorganization treats the stabilization capacity in dynamic and living systems. The adaptive capacities include



Biornametics: Architecture Defined by Natural Patterns, Fig. 8 The brilliant colors of the peacock originate from nanostructures, not from pigments. In the *blue*, *green*, and *light blue* regions, the size of the color-generating nanostructures only differs by some nanometers – an exquisite example of control on the nanoscale by animals



Biornametics: Architecture Defined by Natural Patterns, Fig. 9 The color change abilities of the octopus are legendary. Just recently researchers found the nanoscale reason behind the fast and reversible color change: photonic crystals with tunable size



Biornametics: Architecture Defined by Natural Patterns, Fig. 10 The beautiful shape of *Strelitzia* flowers has inspired the shape of quite a few houses. And also on the nanoscale, *Strelitzia* is inspiring: unidirectional water runoff properties on the plant, as if painted by the brush of an artist

structural change that might affect physical properties such as strength, stiffness, or mass/surface ratio. Also deformation relates to the topic Adaptation and Reorganization; it is about form change due to disturbances. Adaptation can also refer to organizational aspects, for example, the implementation of failure-tolerant systems or mass management tools in case of circulation.

Adaptation and reorganization in biological systems are triggered by a change in environment that the organism or system is either subjected to or actively looking for. Therefore methods of adaptation comprise active, explorative strategies and passive, responsive methods to environmental changes. Examples for both explorative strategies and response can be found in climbing plants, lianas, and vines that actively look for a suitable environment and react to change in structural support with reorganization of tissue. Hierarchical vine-tree-like carbon nanotube architecture inspired by the natural role model shows interesting properties of self-assembly into scaffolds with extraordinary material properties that can yield completely new building materials, for architecture and artistic applications, as well as devices that have to withstand specific conditions

(parameters in tunable materials are responsive to environmental conditions).

In these cases, adaptation is often achieved by local decision-making and local differentiation. The topic focuses on long-term processes that aim at creating or reestablishing equilibrium in a dynamic system that can deliver valuable strategies for the creation of built environment.

Biornametics research role models with relevance to adaptation and reorganization are spider silk, which can mechanically adapt to changing mechanical conditions by unfolding extra units (such as the need when an insect is trapped, preventing rupture of the net, Fig. 12); the octopus (Fig. 9), with amazing adaptation and reorganization properties on all length scales, of high importance for adaptive buildings and reorganizing structures depending on external needs (e.g., deployable guest rooms); the *Strelitzia* flower (the physics of unfolding flowers is complex and has only recently been treated in detail, Fig. 10); and insect wings (Fig. 11), with their amazing transparent, tough, and strong lightweight deployable structures that serve for a lifetime.



Biornametics: Architecture Defined by Natural Patterns, Fig. 11 Insects can deploy their wings without anybody from the outside pulling. This makes them interesting for space applications. And on the nanoscale, insect eyes and wings exhibit various structures responsible for functionalities such as antireflectivity, increased reflectivity, and self-cleaning

Role Model Research

Role model research is an integral part of the biornametics approach. Being together in a botanic park (or, even better, in a virgin rainforest) with high species diversity, with scientists, architects, biologists, artists, and people from further fields, inspires interdisciplinary and lateral thinking, transdisciplinary discussions, and cross-fertilization of ideas. The first round of biornametics research established 37 role models, 10 of them with functionalities based on nanoscale properties (see Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12). This collection shall be extended with each cycle, resulting in a "treasure chest" for





Biornametics: Architecture Defined by Natural Patterns, Fig. 12 Spider silk is a strong and lightweight material. Its functionality is based on properties emerging from smart arrangement of perfect nanocrystals within a soft matrix and further levels of hierarchy at various length scales

biornametics researchers and further interesting audience aiming at inspiration from living nature.

Research in the areas of surface patterns, nanotextured surfaces, and nanostructured materials; shape, growth, and deployable structures; as well as adaptation and reorganization yields:

- Relevant new and established literature that is new to the team
- Finding of new interesting phenomena fitting to the three topics
- Good photographs/videos of phenomena (for use in websites, publications, talks, outreach work, etc.)
- Establishing and outfitting of the website with first information

Role Model Search

Starting with the three main research areas, the team applies a solution-based methodology to extract information from living nature and investigates the three areas and the role models through scientific papers. In this bottom-up process, biology delivers the input, with basic topics and specific examples. From these inputs, deep principles are extracted, in an abstraction step, which are subsequently realized in an artificial material, and implemented to architecture, taking into consideration respective boundary conditions and target aims.

Role Model Data Sheets

The role models need to be investigated further regarding their value for the specific arts-based research approach. A datasheet needs to be developed for each role model. These serve as working draft of the observed role model phenomena and present the underlying scientific information for the arts-based research approach.

The datasheets are organized in two sheets per role model containing an image, the Latin name, the area of defined research, an abstract with a selection of the most interesting features, references, first application ideas, and a functional diagram of the feature to be explored in the further process.

There are ten role models from the first cycle of biornametics research that are based on nanoscale effects; they are introduced in more detail below.

Role Model 01: Jewel Beetle, Coleoptera

Jewel beetles (Fig. 3) are tropical beetles. Their brilliant, iridescent, metallic-looking coloration is highly durable – million-year-old fossil jewel beetles still show magnificent coloration. The color is generated by periodic nanostructures that interact with light. For the coloration, the size of the structures is important, and far less important is the material of the structures (at this size, all is transparent) – which makes this role model very interesting for transfer to technological applications. In some cases, these nanostructures are reactive, i.e., they swell or shrink with certain stimuli, simultaneously changing color. The principle of structural coloration is an example of the deep principle *structure rather than material* and is often applied in living systems.

Role Model 02: *Sarracenia flava*, Pitcher Plant

The yellow pitcher plant (Fig. 4) is a carnivorous plant. Its insides are covered with nanostructures that result in a highly slippery surface for prey insects, yet some symbionts such as pitcher plant spiders can perfectly manage the slippery surfaces and walk on it. Tunable surfaces, which are slippery for certain materials and sticky for others, promise high application potential in architecture.

Role Model 03: *Morpho peleides*, Blue *Morpho* Butterfly

The *Morpho* butterfly (Fig. 5) has such strong, brilliant colors that it can be seen from kilometers away – flashes of blue light in the Costa Rica rainforest. The scales of certain butterflies are covered with multifunctional periodic nanostructures that control wettability, cause coloration, tune water runoff directions, control temperature, etc. Transfer potential to architecture is high, especially for walls and skins of buildings.

Role Model 04: *Victoria amazonica*, Giant Water Lily

The giant Amazon water lily (Fig. 6) can have up to 2.5 m in diameter. Similar to the lotus leaf, its leaves show self-cleaning abilities that are not based on ultrasmooth surfaces but on hierarchical structures, bridging from the nano- via the micro- and meso- to the macroscale. Self-cleaning

surfaces are of high interest in human applications (buildings, cars, dishes, clothes, etc.), and also this effect is based on *structure rather than material* and can easily be transferred to art and architecture applications, by using different materials. The size of the structures is crucial though, in combination with the surface tension of the fluids that will come in contact with the surface.

Role Model 05: *Haliotis*, Red Abalone Shell

The abalone shell (Figs. 2 and 7) mainly consists of calcium carbonate, a rather brittle material. Nevertheless its fracture toughness is 1000 times higher than a crystal of the pure material. The secret is in the few % of protein in the material and in the structure of the arrangement. Little bricks of calcium carbonate, a couple of hundreds of nanometers in size, are glued together with proteins as mortar. The protein is soft, and an attempt to break the abalone shell results in some little brick breaking, but most of the energy being dissipated in the soft protein matrix. Furthermore, the glue between the calcium carbonate bricks shows highly interesting nanomechanical properties (Fig. 2) that majorly contributes to the fracture toughness of the material: it is a self-healing adhesive. The transfer of properties of the complex nanocomposite of the abalone shell can yield the development of fracture-resistant building materials, which is interesting for earthquake-prone regions, and strong, tough, self-repairing adhesives that are needed in various architecture applications.

Role Model 06: *Pavo cristatus*, Peacock

The peacock (Fig. 8) is a beautiful bird. Its feathers have already been mentioned by Sir Isaac Newton in his classic book *Opticks* as an example for structural colors. When watched with high-resolution microscopes, the physical basis of the brilliant colors of the peacock becomes obvious: feather material is arranged as photonic crystals, with little spheres of similar size in hexagonal densest packing. Regions with different colors on the peacock feather refer to different sizes of the spheres in the natural opal. These size differences are on the order of just some tens of nanometers. The order of this arrangement is amazing.

Role Model 23: Spider Silk

Spider silk (Fig. 11) is one of the toughest materials in living nature. Cloth woven from spider silk is stronger than similar materials woven from Kevlar fibers. On the nanoscale, spider silk is a complex nanocomposite, with beta sheet nanocrystals embedded in semi-amorphous protein phase.

Role Model 34: *Octopus vulgaris*, Common Octopus

In octopus (Fig. 9), part of the impressive and rapid color change is based on nanostructures and their interaction with light. The size of the photonic crystal changes, by smart rearrangement of the matrix between the single optically relevant particles, and this causes the color change – this effect is physically similar to the color change in soap bubbles when they become thinner with time.

Role Model 36: *Strelitzia reginae*, Crane Flower or Bird of Paradise

There are various houses with shapes inspired by the extraordinary form of *Strelitzia* flowers (Fig. 10). *Strelitzia* leaves exhibit unidirectional water runoff characteristics and superhydrophobic properties that are based on minuscule surface structures. Transfer of these properties to surfaces of buildings can result in unique, beautiful patterns for water runoff, in the case of rain, for example, and even the appearance of messages and other text on windows, invisible on sunny days. They can also inspire bird-protective coatings, if their reflective properties are in the UV range, a wavelength portion of light that cannot be seen by people but can very well be detected by birds, who can see in the UV range. In this way, the huge number of birds that die every year from collisions with buildings (this number is in the billions!) can be drastically reduced, without compromising the visual appearance of the building to the human eye. A further property worthwhile of being transferred to architecture and art, and technology in general, is self-cleaning abilities based not on toxic chemical coatings but on structural coatings, which can in principle be obtained by structuring the most outer

material of the building, without adding another layer (biomimetic principle: *integration instead of additive construction*).

Role Model 37: Insect Wings

Insect wings (Fig. 11), such as the self-deployable wings of ladybirds, have been inspiring self-deployable structures for space applications and further uses in territories that are unfriendly to people for various decades. With the arrival of the age of nanotechnology, further amazing inspiring properties of insect wings and more generally surfaces of insects have been identified. The nanostructures in these organisms are multifunctional and serve various purposes depending on the environment, current situation, and further parameters. Self-cleaning, beautiful non-bleaching coloration, temperature regulation, control of wetting behavior, self-repair, and various optical properties (e.g., reflective, antireflective) have been identified and are starting to be used in applications such as antireflective surfaces of optical lenses and glasses (inspired by the moth eye).

Role Model Categories Diagram

In the following step of the biornametics research methodology, the role models are categorized into the three main established groups. Due to the multifunctionality of biological materials, there will be many overlaps, so some role models fit into several groups, and the aspects that are interesting to investigate are ambiguous.

For example, spider silk (Fig. 11) fits into the category *shape, growth, and deployable structures* and into the category *reorganization and adaptation*.

> Some spiders use glues while others deploy strands of fine filaments for fixing flies. Recent work has provided new insights into the mechanical properties of these nanoscale ropes.
>
> A spider's web, typically in combination with a powerful poison, is its principal tool in the struggle for survival. It is a highly tuned, lightweight net-structure that relies heavily on skillful engineering and the fitting deployment of a most versatile material: the spider's famous silk. A typical web spider makes up to seven different types of silk, some of which can be tuned to suit environmental circumstances by adjustment of the spider's spinning physiology. Individual silk threads, each around a micrometer in diameter, are actively drawn from microscopic spinning spigots perched on motile spinneret mounds, resembling miniature batteries of battle-ship gun-turrets. . . . The resulting filamentous tangle clings to, and often totally covers, a pair of much thicker support fibers issuing from spigots on the main spinnerets. This multiple-fiber rope can be further reinforced by crimped, spring-like fibers from yet another set of spigots, which pull up tight the by now rather complex, composite fiber. ([13], text slightly adapted)

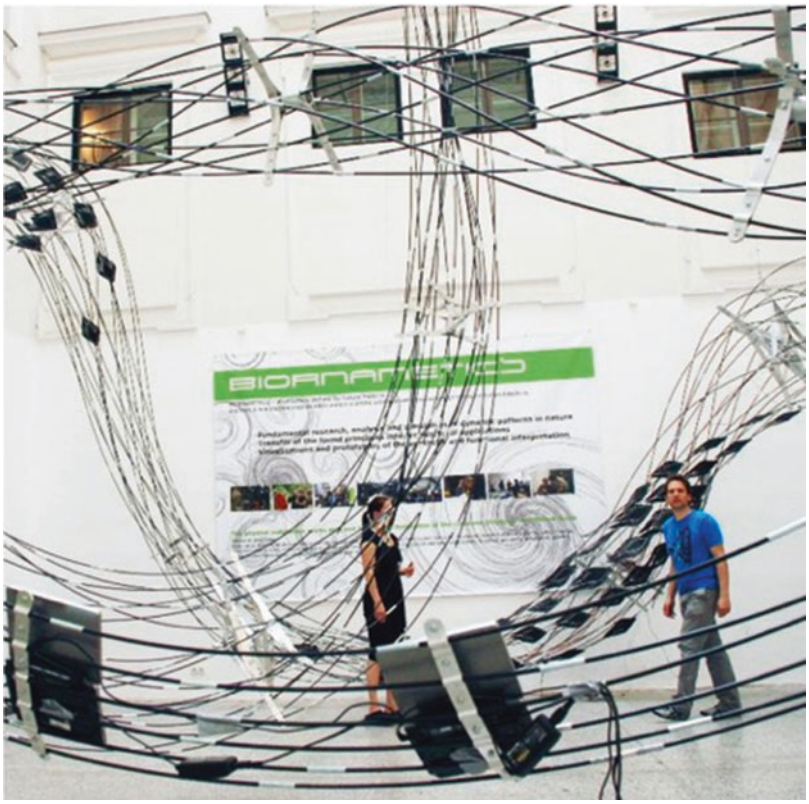
Thus the spider silk follows interesting growth principles during its "spiderfacturing." In this way it also adapts to the neighboring environment so well that it can provide the framework of catching food for its producer.

Art Installation

A physical installation shall be the result of the translation of specific principles encountered, analyzed, and discussed during the research work of the biornametics research approach. The installation shall demonstrate one of the unlimited possibilities of interpretation of a single principle and thus be closely related to the perception of the executing team. The goal in the first round of biornametics research was to build a lightweight and cost-efficient spatial structure with an inherent potential to move, or to be moved, by stimuli from the external environment (Fig. 13).

The primary structural system of the installation reflects a specific organization of fibers arranged in concentric layers twisted clockwise and counterclockwise around a virtual axis. This organization of fibers can be found as a structural principle in several role models investigated during the research, displaying a high capacity to resist and adapt at the same time. This capacity depends on the geometry of the arrangement and the material properties of the fibers. Possible deformations can only happen within certain limits relative to the stiffness of the material and structure. The adaptive capacities of the structure exceeded by far those of conventional structural building elements.

Glass fiber tent poles were chosen for the fibers because of their structural properties, the



Biornametics: Architecture Defined by Natural Patterns, Fig. 13 Art installation in the Silver Gallery at the Vienna University of Applied Arts, Austria. The single parts of the installation were developed in the biornametics research approach

modularity of the system, and the easy availability at appropriate costs. A few reversible and numerous irreversible joints had to be defined. The so-called spacers kept the elements at specified distances and positions in order to maintain the spatial arrangement of the overall system.

The joints between the spacers and the fibers were designed to produce as little friction as possible and so allow easy movement by gliding along the fiber's axis. On the other hand, fixing of the connection points was necessary to ensure that the system would lock temporarily in a chosen position.

The use of a test setup with sensors and movement control would have exceeded the given budget and time restrictions, so the installation was fixed in space, but change in shape was enabled by a controlled and synchronized setup process, specifically by altering unlocked joints in specified places. The whole construction was suspended from the ceiling, and its performance demonstrated a resistance sufficient to maintain its shape.

Adaptable stiffness of the structure (not necessarily of the material involved) could be a point of departure for future development of more flexible and even reversible building systems (e.g., structures exerting increasing resistance under increasing external forces). On a second level of hierarchy, the installation demonstrated another adaptive process often found in nature: speeds of transformation varying in reaction to specific environmental changes. The system consisted of reactive leaflike structures that were partially fixed to the primary fiber structure together with sensory, control, and actuation equipment. The elements were added in a cantilevering mode and therefore reached a higher level of flexibility and of visible change.

Future Directions of the Field

In the fourth round of the PEEK funding scheme, *GrAB*, *Growing as Building*, the sequel arts-based

research approach to biornametics, was funded by the Austrian Science Fund FWF in the frame of PEEK 2011. *GrAB – Growing As Building* – runs from 2013 until the end of 2015. GrAB takes growth patterns and dynamics from nature and applies them to architecture with the goal of creating a new living architecture. The aim of GrAB is to develop architectural concepts for growing structures. Three main directions are investigated: transfer of abstracted growth principles from nature to architecture, integration of biology into material systems, and intervention of biological organisms and concepts with existing architecture. Key issues of investigation are mechanisms of genetically controlled and environmentally informed, self-organized growth in organisms and the differentiation of tissues and materials.

Research parameters include, for example, size, height, speed, and properties such as the stiffness or flexibility of structures, which are equally important in living systems and in architecture [14].

The methodology of biomimetic information transfer used is based on refined methods from biornametics research. Computer simulations remodeling relevant principles and physical models are used to understand selected natural phenomena and inform the translation process. Artistic and architectural tools and methods are jointly used with practices from the natural sciences. A Biolab provides hands-on experience with growth of organisms.

The architectural interest lies in the development of structures in a specific environment following an interaction of elements derived from natural pattern formation. To support the transfer of dynamic, growing structures, recent advancements in processing technologies such as additive manufacturing systems are investigated and possibly integrated into design concepts and both analogue and digital models. Scales tackled can include materials (also nanoscale), built elements and structures, as well as urban systems.

Along with architects and artists, scientists from the field of biology, biomimetics, and engineering including students collaborate to develop this arts-based research approach in a holistic way and to supply all the necessary expertise.

The work on GrAB commenced in June 2013 and will run for 2.5 years until the end of November 2015. The host research institution is the University of Applied Arts, Vienna; key people are Barbara Imhof and Petra Gruber. Collaborating institutions are located in Austria, the United Kingdom, Germany, the Netherlands, and Ethiopia.

Nanoscience and nanotechnologies are important for various aspects of GrAB, with the language of life being spoken and programmed on the nanoscale, with evident effects on the micro- and mesoscale, and finally on the macroscale of the built environment and urban systems.

Cross-References

- ▶ Biomimetics
- ▶ Nanostructures for Coloration (Organisms Other Than Animals)
- ▶ Structural Color in Animals

References

1. Diah, S.Z.M., Karman, S.B., Gebeshuber I.C.: Nanostructural colouration in Malaysian plants: lessons for biomimetics and biomaterials. J. Nanomat. 2015, Art.# 878409(15p.) (2014)
2. Kinoshita, S.: Structural Colors in the Realm of Nature. World Scientific Publishing Company, Singapore (2008)
3. Lee, D.: Nature's Palette: The Science of Plant Color. University of Chicago Press, Chicago (2007)
4. Loos, A.: Ornament and Crime: Selected Essays. Studies in Austrian Literature, Culture, and Thought. Translation Series. Ariadne Press, Riverside (1997) - (written in German in 1908)
5. Bar-Cohen, Y.: Biomimetics: Biologically Inspired Technologies. CRC Press, Boca Raton (2005)
6. Bhushan, B.: Biomimetics: Bioinspired Hierarchical-Structured Surfaces for Green Science and Technology. Springer, Heidelberg/New York/Dortrecht/London (2012)
7. Gruber, P., et al.: Biomimetics – Materials, Structures and Processes: Examples, Ideas and Case Studies. Springer, Heidelberg/Dortrecht/London/New York (2011)
8. Imhof, B., Gruber, P.: What is the Architect Doing in the Jungle? Biornametics. Springer, Wien/New York (2013)
9. Gebeshuber, I.C.: We have to establish a common language. In: Imhof, B., Gruber, P. (eds.) What is the

Architect Doing in the Jungle? Biornametics, pp. 68–69. Springer, Wien/New York (2013)
10. Smith, B.L., et al.: Molecular mechanistic origin of the toughness of natural adhesives, fibres and composites. Nature **399**, 761–763 (1999)
11. Gordon, R.: The Hierarchical Genome and Differentiation Waves: Novel Unification of Development, Genetics and Evolution. World Scientific, Hackensack (1999)
12. Gebeshuber, I.C.: Biomineralization in marine organisms. In: Kim, S.-K. (ed.) Springer Handbook of Marine Biotechnology, pp. 1279–1300. Springer, Dortrecht/Heidelberg/London/New York (2015)
13. Vollrath, F.: Spider silk: thousands of nano-filaments and dollops of sticky glue. Curr. Biol. **16**(21), R925–R927 (2006)
14. Gruber, P.: Biomimetics in Architecture: Architecture of Life and Buildings. Springer, Wien/New York (2010)

Biosensing

▶ Biosensors

Biosensors

Henry O. Fatoyinbo and Michael P. Hughes
Centre for Biomedical Engineering, University of Surrey, Surrey, UK

Synonyms

Bioprobes; Biosensing; Nanobiosensors

Definition

A biosensor is a system or device which incorporates a biologically active material in intimate contact with an appropriate transduction element for the purpose of detecting the concentration or activity of chemical species in any type of sample.

Overview

The important components of a biosensor are (1) a bioreceptor (e.g., enzymes, antibody, microorganism, or cells); (2) a transducer of the physicochemical signal, and (3) a signal processor to interpret the information that has been converted. Depending on the type of physicochemical signal produced, i.e., electrochemical, mass, thermal, or optical, the transduction element can come in various forms, such as electrodes for amperometric and potentiometric detection, field effect transistors (FET), piezoelectric quartz (TSM, QCM), thermistors, and optical fibers/optoelectronic systems. Biosensors have found many applications in a range of industries including defense, medicine, food, pharmaceutical, and healthcare owing to their rapid response time in detection, coupled with the emergence of complementary technologies. As fabrication technologies become more sophisticated, research and development has continually reduced the size of biosensing devices, giving rise to submicron detection systems known collectively as *nanobiosensors*.

One of the first descriptions of a man-made biosensor was the e*nzyme electrode*, coined by Clark and Lyons in 1962 [1]. In this system an oxidoreductase enzyme was held next to a platinum electrode in a membrane sandwich. Holding the platinum anode to a positive potential (0.6 V) relative to an Ag/AgCl reference electrode, an enzymatic catalyzed oxidation reaction of glucose produced hydrogen peroxide detected by the platinum electrode. It was not until 1967 that Updike and Hicks gave the necessary details required to build a functional enzyme electrode for glucose monitoring [2], paving the way for the first commercial biosensor for measurement of glucose in whole blood to come on to the market in 1974. The biological entity (i.e., enzyme) catalyzing this reaction produced a physicochemical signal, which was converted to an electrical signal, similar in process to the different sensory systems found in the human body.

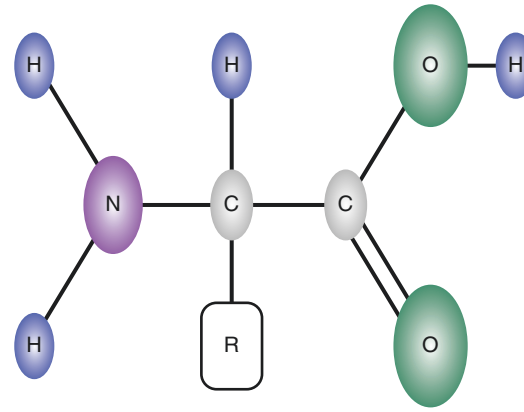
Biosensors which incorporate an electrochemical (EC) transduction element monitor an electroactive species which is produced or consumed by the action of the biological element. These include amperometric, potentiometric, conductimetric, and field effect transistor (FET)-based detection [3, 4]. Enzyme-based electrodes are the most

commonly used biosensors associated with electrochemical transduction elements (i.e., amperometric and potentiometric). The basic principle behind amperometric detection (e.g., glucose electrode) works by holding the voltage potential between a working and reference electrode constant, while measuring the variation in current in relation to the concentration of the active species. In contrast, potentiometric detection (e.g., pH electrode) is based on holding the current constant or at zero and measuring the variation in voltage as a logarithmic proportion of the reactive species. Oxidoreductases and hydrolases are commonly used enzymes in these systems, and can be immobilized on to the electrode in a variety of ways. To improve these sensor characteristics, high porosity membrane structures (e.g., polycarbonate) allow a more controlled substrate flux to the immobilized enzyme layer, situated just above the detection electrode. Both potentiometric and amperometric biosensing systems receive considerable amounts of attention as they have the advantages of being low cost, single-use devices with a high degree of reproducibility. The associated instrumentation is also very inexpensive and compact, lending these biosensors to on-site measurements in fields as varied as environmental monitoring to bioprocess monitoring.

Classification of Bioreceptors

Enzymes and Antibodies

The vast majority of biochemical reactions taking place in a cell are permitted through the catalytic ability of enzymes. Enzymes are globular proteins; their primary structure consists of sequences of L-amino acids linked via condensation reactions to form polypeptide chains. The basic generic structure of an amino acid can be seen in Fig. 1, where the letter R denotes the side chain of the peptide. These peptides can be classified as polar neutral (cysteine and tyrosine being the most hydrophilic amino acids), nonpolar neutral (hydrophobic), basic (negatively charged), or acidic (positively charged) depending on the attached groups. There are 20 different types of L-amino acids, and with variations in size from



Biosensors, Fig. 1 Generic structure of an amino acid residue. The R-side group denotes the amino group

62 to 2,500 amino acid residues an enormous combination of polypeptides can be created. The primary structured polypeptide is the starting point for the final globular protein. As the amino acids begin to interact with each other through ionic interactions, van der Waals forces, hydrogen bonding, and hydrophobic bonding, complex folding arrangements begin to occur. This secondary structure, commonly termed α-helix or β-pleated sheet, may contain regions of regular folding patterns, stabilized by weak interactions of neighboring carboxylic (–COOH) and amine ($–NH_2$) side groups. Neighboring interactions of the R-side groups determine the 3D conformational shape of the protein, referred to as the tertiary structure. Quaternary structures are formed when separate macromolecules (e.g., cofactors) get linked in a specific order forming enzymes of high stereospecificity, regiospecificity, and chemoselectivity.

Enzymes are classified into six major groups: (1) oxidoreductases, (2) transferases, (3) hydrolases, (4) lyases, (5) isomerases, and (6) ligases (Table 1). They are able to retain their catalytic activity within a limited range of parameters (i.e., pH, temperature, fluid forces, and chemical agents), but are well suited for amperometric and potentiometric biosensors due to their high turnover numbers, defined by the net number of substrate molecules reacted per catalyst active site per unit time. The dependence of the speed of reaction to the concentration of substrate can be described

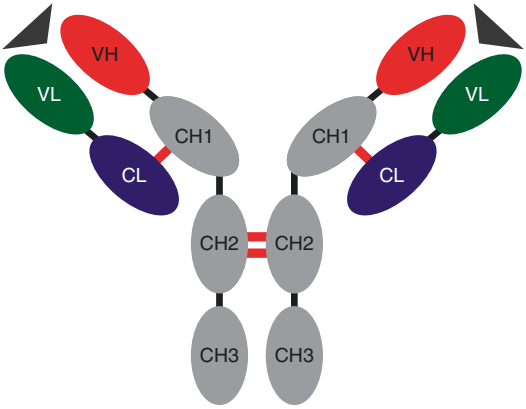
Biosensors, Table 1 Class of enzymes and examples within each class

Enzyme definition	Examples
Oxidoreductases: catalyze oxidation and reduction reactions	*Glucose oxidase*, *Alcohol dehydrogenase, Lactate dehydrogenase, Diacetyl reductase*
Transferases: transfer functional groups	*Thiaminase, DNA methyltransferase, Choline acetyltransferase*
Hydrolases: catalyze the hydrolysis of bonds	*Phospholipase, Lipoprotein lipase, Phosphatase, Dexyribonuclease, RNase, Lysozyme*
Lyases: cleaves bonds, not via hydrolysis or oxidation	*Carbonic anhydrase, Tryptophan synthase, Adenylate cyclise*
Isomerases: catalyzes isomerization change of molecules	*Protein disulfide isomerise, Topoisomerase, Phosphoglucomutase*
Ligases: links two molecules via covalent bonds	*DNA ligase, CTP synthase, Acetyl-CoA carboxylase*

using the equation derived by Michaelis–Menten in Eq. 1, where V is the reaction rate, [S] is the substrate concentration, and K_m is the Michaelis–Menten constant at which the substrate concentration gives the reaction rate which is half of V_{max} (maximum rate at substrate saturation). Although enzymes are most widely used in biosensing systems, they have a high cost factor in terms of production and purification as they have to be derived from biological sources:

$$V = \frac{V_{\max}[S]}{K_m + [S]} \quad (1)$$

Antibodies, also known as immunoglobulins (Ig), are a class of globular proteins (~150 kDa) which bind to specific antigenic structures forming an antibody-antigen complex, and they have widespread applications in clinical diagnostics and biotechnology. Their basic structure comprises of four polypeptide chains arranged in a Y configuration (Fig. 2). There are two heavy chains and two light chains connected by disulfide bonds. The base of the Y-structure is known as the



Biosensors, Fig. 2 Representative structure of an IgG antibody consisting of four polypeptide chains, interlinked with disulfide bridges (CH2–CH2; CH1–CL). CH1–CH3 represents the constant heavy chains; CL represents the constant light chains; VL and VH are the variable light and heavy molecules, respectively, which form the paratope regions for antigenic binding (*triangle*)

Fc (fragment crystallizable) region and is found in the constant domain containing the heavy chains which define the class or isotype of antibody in mammals (i.e., IgA, IgM, IgG, IgD, and IgE). The Fab (fragment-antigen binding) region is found in the variable domain and contains the light chains and heavy chains which make up the amino terminal–shaped antigen binding site, known as the paratope, the most important region for antigen binding on the antibody. The two binding sites on the arms of the antibody can be extremely variable, allowing millions of different antibodies with slightly varying paratopes to exist. This lends antibodies to be widely applied in immunoassays for molecule detection across cells, large proteins, and even amino acids.

Nucleic Acids

Nucleic acids are a group of linear polymers made up of nucleotides. Nucleotides comprise a purine or pyrimidine nucleobase, a pentose sugar, and a phosphate group. The five known nucleotides or bases are (1) Adenine (A), (2) Guanine (G), (3) Cytosine (C), (4) Thymine (T), and (5) Uracil (U). The base pairing (b.p) of these nucleotides in the biopolymer molecule DNA takes the form of A-T and G-C. Deoxy-ribonucleic acid (DNA) and ribonucleic acid (RNA) are double-stranded helix

and single-stranded polynucleotides, respectively, though this is more of a generalization rather than a rule. DNA and RNA have very similar chemical structures, apart from two distinct differences: (1) DNA contains 2-deoxyribose, a sugar lacking one oxygen atom compared to the ribose sugar of RNA; (2) RNA uses the nucleobase Uracil as a base, while DNA uses Thymine. The main role of DNA is the long-term storage of the genetic code of living organisms, containing instructions for the construction of macromolecules needed in cells such as proteins and RNA. The genetic information is coded into groups of three adjacent nucleotides termed codons, which represent the 20 different amino acids, the start codon, and the stop codon encoded by RNA. RNA can be considered as the workhorse for protein synthesis, copying the genetic information through transcription (messenger RNA) and translation (transfer RNA and ribosomal RNA) providing information to synthesize all the proteins needed by cells in the body. Development of microarrays has led to a new generation of biochips capable of multiplexing for rapid drug discovery and environmental analysis as will be discussed later.

Cells and Microorganisms

A vast range of microbes (i.e., yeast, algae, bacteria) have been utilized in biosensor systems, having the advantage of being massively produced through cell culturing. They are classified into two major groups based on the sensing technique employed: electrochemical microbial biosensors and optical microbial biosensors [5]. The basic principle is that certain strains of microorganisms preferentially metabolize certain compounds, with the metabolic waste being excreted detected by a transduction system. Although microbial biosensors are less sensitive than enzymatic biosensors, and the overall detection period can be longer, they are well suited to environmental analysis (e.g., Biological Oxygen Demand (BOD)) where the global effect of some toxin is detected rather than a specific toxin or compound. An interesting and growing area of research which uses this principle is that of *synthetic biology*, where a commonly used microbial chassis (e.g., *E. coli*, *S. cerevisiae*) is transfected with a gene that is turned on in the presence of a specific substrate. This has rapid sensing applications in healthcare monitoring of viruses and toxins.

Artificial Bioreceptors

The application of artificial bioreceptors is based on the naturally occurring structures of catalytic antibodies (abzymes), where these antibodies have been produced artificially by directing them against an analogue compound of the transition state for a given reaction. These abzymes have the same specific recognition of natural antibodies and the catalytic activity of enzymes reacting to substrates (antigen) to form an end product with comparable kinetics.

Molecular-imprinted polymers (MIPS) are another form of artificial bioreceptors. They are created by mixing functionalized monomers with an analogue or target molecule which binds to the monomers through hydrogen bonds, hydrophobic bonds, or ionic interactions, thus creating a template around which the polymer will form [6]. Through polymerization reactions (i.e., UV irradiation) in the presence of a cross-linking agent a 3D rigid structure is created with complementary sites for the target molecule. Removal of the template target leaves a molecularly imprinted polymer designed for specific recognition of target molecules. Although this method for creating receptors is useful for hard to obtain natural bioreceptors, there are issues regarding the polymerization stage which can often lead to wrong structure formations.

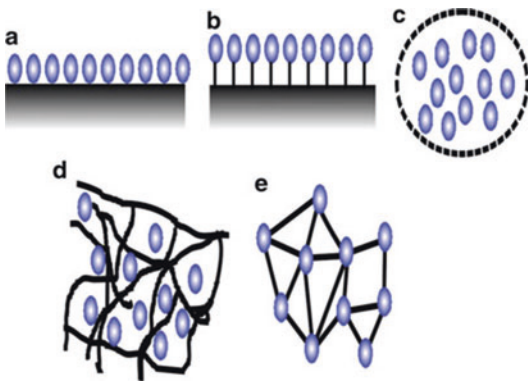
Unlike enzymes, antibodies and MIPS form ligand-receptor complexes which do not produce physicochemical signal, so quantitative or qualitative detection must be made via indirect or direct methods. Indirect methods use a secondary biological system label (bioconjugate) which emits a physicochemical signal, such as a light-emitting probe (e.g., Horseradish peroxidase) or a radioactive element, commonly used in immunosensor applications. Direct methods do not require labeling, and involve measurement changes in mass, electrical properties, or optical properties as the ligand-receptor complex is formed. The non-catalytic activity of these receptors means that once the complex has been formed reuse of

the receptor in the biosensor can only be achieved through device washing to dissociate the complex.

Immobilization Techniques

Attachment of bioreceptors to the surface or in very close proximity of a signal transduction element is usually a prerequisite for biosensors to operate efficiently. A range of immobilization methods have been developed for immobilizing cells and nanoscale bioparticles (enzymes, antibodies, etc.) on and within a variety of substrates including microscope glass slides, micro- and nano-electrodes (i.e., gold, metal oxides, platinum, silver, etc.), SiO_2, polymers, and hydrogels. The technique employed is very much dependent on the bioreceptor and the substrate to which it has to be immobilized for subsequent signal transduction. Originally developed for enzyme immobilization, techniques can be classified into five broad areas which are adsorption, inclusion, covalent coupling on activated surfaces, cross-linking, and confinement (Fig. 3).

Inclusion uses polymerization reactions to embed or encapsulate the bioreceptor within a cross-linked system such as a hydrogel. Mainly used to immobilize whole cells or subcellular fractions, it provides a homogeneous distribution of bioreceptor within the gel but the reagents used for cross-linking (i.e., photo-initiators) can produce free radicals on UV-activated polymerization which can be detrimental to the biological structures present. A well-developed technology, widely used in fiber-optic biosensors is that of the *sol-gel* technology. Initially centered on metalloproteins it has been extended to encapsulate proteins, enzymes, antibodies, and whole cells with the biomolecules retaining their structural integrity and biological function [7].



Biosensors, Fig. 3 Representation of techniques employed for bioreceptor (*sphere*) immobilization on and within substrates for biosensor. (**a**) Adsorption; (**b**) covalent attachment to functionalized surfaces; (**c**) encapsulation in a semipermeable membrane; (**d**) inclusion in gels; (**e**) cross-linking using cross-linking agents, i.e., glutaraldehyde

Adsorption processes occur due to simple interactions of the bioreceptor and the substrate. The substrate can be either charged, leading to ionic bonds being created, or uncharged leading to weaker bonds (hydrogen, hydrophilic, and van der Waals effects). Due to varying environmental factors which may occur at the interface, desorption of the bioreceptors is common when change in pH or ionic strength is present, making this technique less widely used for immobilization.

Entrapment involves the bioreceptor remaining in solution, compartmentalized by a semipermeable membrane, allowing small molecules through. This technique is particularly useful in hydrogel biosensing applications where control is of huge importance such as in drug delivery systems which rely upon the slow release of drugs into a system based on the target analyte concentration.

Cross-linking is a technique whereby a bifunctional agent such as glutaraldehyde encourages attachment of polypeptide structures to each other or on to substrate surfaces to make the complex more stable. This technique is particularly useful in the processes involving adsorption or inclusion.

Covalent binding of proteins on activated surfaces can be achieved through the covalent bonds developed between functional groups not required in the biorecognition process. The surface activation process prior to immobilization can be achieved either via a direct method whereby amine ($–NH_2$), carboxylic (–COOH), hydroxyl (–OH), or sulfhydryl (–SH) groups on the substrate surface are functionalized through a range of reagents before covalent binding can occur

with the functional group of the bioreceptor. Another common method is silanization of glass surfaces with an organosilane containing an organic functional group at one end and an alkoxysilyl group at the other end of the molecule.

Carboxylic groups can be activated using the acyl nitride method or carbodiimide method. The acyl nitride method transforms the carboxylic acid group to hydrazides through esterification by methanol in acidic solution, then treatment with hydrazine. A final treatment step with sodium nitrite gives acyl nitride which forms an amide bond with the amine groups of the bioreceptor. The carbodiimide method reacts in acidic solution with the carboxylic group forming a key intermediate, *O*-acylisourea a carboxylic ester, subsequently forming an amide bond with the amine group of the bioreceptor. Hydroxyl groups are activated using cyanogen bromide forming an intermediate which reacts with the amine group of the bioreceptor. Amine groups can be activated using diazonium salts, while sulfhydryl groups are activated through the formation of disulfide bridges between thiol groups of the bioreceptor and the substrate, after surface treatment with 2,2′-dipyridyldisulfide. Although most of these techniques are straightforward to perform, ready to use activated surfaces are commercially available (e.g., Sigma Aldrich). As nanostructures, nanoparticles (NP), and thin films become increasingly incorporated into biosensors, variations to these immobilization techniques for biomolecule attachment have been applied such as layer by layer (LbL) and Langmuir–Blodgett (LB) films for better molecular architectures and ultimately enhanced performances in biosensing systems [8].

Detection Technologies for Nanobiosensors

A wide range of biosensor technologies have been developed since the electrode enzyme of 1962. In recent decades nanomaterials have made significant headway in to biosensor systems for improving the limit of detection (LOD) in a host of complex sample solutions. These new nanobiosensors (or bionanosensors) typically use traditional detection techniques with nanotechnology to develop a host of enhanced bioelectronic devices for a variety of biomedical research and industrial applications. Along with optical, electrochemical, and mass transduction systems, biosensors have also utilized thermistors and magnetic beads as transduction elements. The state of art for current nanotechnologies involved in biosensing is extremely large and intertwined between detection technologies; hence for simplification the following section serves as an illustrative description of novel, and by no means exhaustive, approaches in the most common biosensing techniques.

Electrochemical-Based Detection

Carbon Nanotube Biosensors

An emerging building block of novel nanostructures and devices used in electrochemical biosensing applications are *carbon nanotubes* (CNT). Synthesized in various forms (e.g., single walled (SWCNTs), multi-walled (MWCNTs)), CNTs possess a large surface per unit mass and excellent mechanical and electrical properties making them extremely powerful in electronic detection of biomolecules, viruses, and cells. Functionalization of CNTs with different chemical groups has been achieved using both covalent and non-covalent methods [9]. At low potentials, various analytes can be oxidized with little surface fouling, making them ideal for high selectivity and reusability. Commonly used is the –COOH group for covalent bonding on CNTs via oxidizing treatment, which may also create other functional groups (e.g., hydroxyl, carbonyl, etc.) at defect sites of the graphene sheet. These other functional groups can be easily converted to carboxylic acids through perchloric acid ($HClO_4$) treatment or potassium permanganate ($KMnO_4$), which sets the surface for immobilization of bioreceptors through the carbodiimide method. CNT electrochemical (CNT-EC) biosensing has been reported for detecting glucose and neurotransmitters and a range of other biological entities, including nucleic acids and viruses. Immobilization of single-stranded DNA

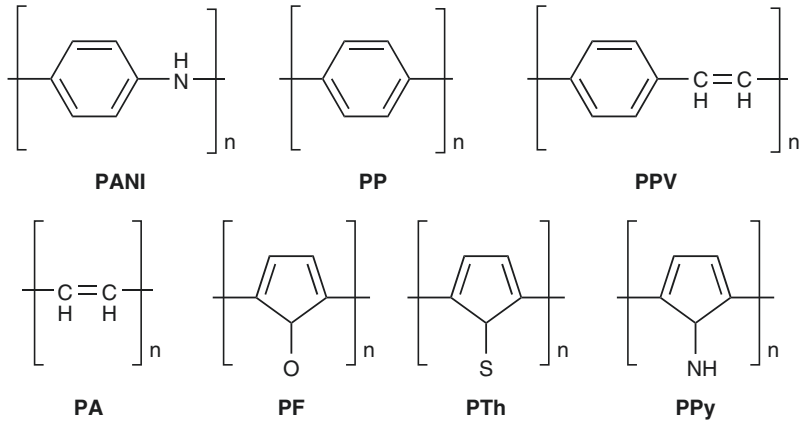
(ssDNA) molecules, mainly through covalent attachment, on CNTs has led to the development of electrochemical DNA biosensors which have been used to sense nucleic acids, genes, and DNA molecules. In addition, the enhanced amplification in signal transduction has seen their use in detecting DNA hybridization reactions. There are a host of doped and nanocomposite CNT electrodes developed for a variety of DNA detection applications, with sensing techniques that include impedance spectroscopy, surface conductance, and differential pulse voltammetry. Studies into the characteristics of immobilized ssDNA and hybridization of DNA have also been conducted using techniques such as cyclic voltammetry (CV) and electrochemical impedance spectroscopy (EIS).

Since about 85 % of the biosensor market, worth approximately US$ 8.5 billion, belongs to glucose sensors, a large amount of novel biosensing ideas are based around blood glucose monitoring [9]. Strategies based on the presence or absence of enzymes as the bioreceptor element has been applied to CNTs for glucose monitoring. The enzyme glucose oxidase (GOD), which catalyzes the oxidation of glucose, has been immobilized on an Ag-NP/CNT/Chitosan composite giving a rapid and steady response to the substrate. This was also demonstrated on MWCNTs/ZnO-NPs, where GOD was electrostatically bound to ZnO-NPs based on isoelectric point differences. Incorporation of GOD into a colloidal Au-CNT composite showed that the apparent Michaelis–Menten constant showed a high affinity of the enzyme to the substrate under the microenvironment of the AuNPs. MWCNTs modified with ferrocene electrodes have been immobilized with GOD, and gave an improved response to glucose oxidation as the ferrocene acted as an electron transfer mediator while the MWCNT acted as the conductor. Though doped-CNT electrodes can promote electron transfer in redox reactions interference could result as endogenous O_2, unless extracted, may react with GOD producing H_2O_2, thus underestimating the glucose level. CNT-based biosensors not employing enzymes for glucose monitoring tend to be low in cost and can be achieved through, for instance, the direct growth of CuNPs on CNTs through self-assembly or via the use of immobilized microbes on CNT-modified carbon paste electrodes [10].

CNT-EC biosensors largely tend to be based upon amperometric sensing, with the CNT either surface modified with an oxide or chemically doped with a variety of NPs for label-free detection of biomolecules such as immunoglobulins and insulin. A wide range of CNT-bioreceptor bioconjugate-based biosensors have also recently been developed for real-time detection with detection limits down to a zeptomole (10^{-21}) based on double-step amplification. For example, immunoaffinity reactions (i.e., formation of antibody-antigen complexes) have been detected on CNTs as a function of mechanical and electrical properties of the CNT; CNT-FET (field effect transistors) have been used to detect mismatches of dsDNA sequences through decreases in conductance; viral antibodies have been immobilized on to CNTs through the use of polymers to detect viruses; and metal-NPs have been used for DNA attachment to CNTs for enhanced detection of sequence-specific target DNA and human hepatitis B virus ssDNA with LOD as low as 0.035 pM [10].

Conducting Polymer Biosensors

Another class of nanomaterial which has recently received widespread interest for biosensor applications are a class of functional polymers known as *conducting polymers* (CP). Their attractiveness as electrochemical transducers in sensing stems from their highly conjugated polymer chain composed of alternating single and double carbon–carbon bonds, which can be doped or de-doped in a controllable manner. When used as nanostructures, these unique properties are further enhanced with large surface area, nanoscale size, and the exhibition of quantum effects making response times significantly faster than bulk polymers. The primary principle behind CPs as sensing devices is based on the "before and after" sensing of a target molecule on the surface on the CP, meaning that the sensitivity of the CP is highly dependent on its surface area where target molecules have the ability to increase or decrease



Biosensors, Fig. 4 Chemical structures of common conductive polymer units: polyaniline *PANI*), poly(para-phenylene) *PP*, poly (phenylenevinylene) *PPV*, polyacetylene *PA*, polyfuran *PF*, polythiophene *PTh*, polypyrrole *PPy*

mobile charge carriers, thus affecting the conductivity of the CP. Figure 4 shows the typical molecular structures of CPs for biosensing applications. Techniques used to synthesize CP nanostructures include hard- and soft-template methods and electrospinning. In the hard-template method, which is most commonly used, a template membrane is required to guide the growth of the nanost ructures within the pores leading to a completely controlled nano-morphology dictated by pores and channels, which can give CP nanowires of 3 nm in diameter. As the template needs removing post-process, the technique has some limitations in that the nanostructures can be destroyed or deformed. Various preparation protocols using the hard-template method have been described using either chemical or electrochemical polymerization, but the important variable in these processes is the actual template which can take the form of porous alumina membranes (Al_2O_3), polycarbonate membranes, or nanofibers [11]. In contrast, the soft-template method sometimes referred to as a self-assembled method requires no template. It is a cheap and fairly simple technique based on non-covalent interactions driving the self-assembly (e.g., π–π stacking interactions, hydrogen bonding, and dispersive forces), and can use oligomers, DNA, surfactants, colloidal particles, or colloids to synthesize CP nanostructures, also via both electrochemical and chemical methods [11].

Conducting polymers are extremely versatile when used in biosensing devices and have been used as immobilization matrices, artificial bioreceptors, and transducers for electrical charge from redox reactions [12]. Based on conductometric detection, a change of electrical conductivity of the CP nanostructure upon interaction with target analytes can be measured. The basis of potentiometric detection is the detection of changes to the chemical potential of the CP due to internal shifts in the anion equilibrium. Using these two detection techniques, CP-based biosensors have been developed based on nanothin films and nanostructures. Nanothin CP films, produced by electrochemical polymerization, spin coating, or layer by layer techniques, tend to have bio-active molecules embedded allowing target analytes adsorbing on to the film diffuse into the film for interaction with the receptor. As examples, an ultra thin film of PPy-GOD composite was fabricated for potentiometric sensing of glucose, showing a high sensitivity to glucose at low current. An organic thin film transistor (OTFT) for glucose monitoring had a channel of immobilized GOD on poly (3,4-ethylenedioxythiophene)-polystyrene sulfonate (PEDOT-PSS) CP film, giving rapid response times (∼20 s) and high selectivity between drain current and glucose concentration in test conditions. Using layer by layer self-assembly, an ion-sensitive field effect transistor (ISFET) was fabricated with polyaniline (PANI) as the thin film channel giving glucose sensitivity of <0.8 μA/mM, further indicating the potential of nanothin film CP as low cost biosensor materials. Other defining factors which influence the performance of these nanothin films include the thickness of the film, the roughness of

the film, and the nanoporosity of the film. CP nanostructures include CP-nanoparticles (CP-NP) which tend to be dispersed on electrode surfaces to increase area/volume ratio for favorable adsorption of biomolecules; nanoarrays of nanotubes/nanowires in highly ordered alignment have shown increased sensitivity in hybridization, mass transport, and kinetic studies; and individually addressable 1D nanostructures (nanowires/nanotubes) which have shown extremely powerful transduction capabilities due to the high contact area of the analyte on the nanostructure which produces polarons and bipolarons along the backbone of the CP [11]. This effectively affects the electrical conductivity of the CP nanotube or nanowire which subsequently results in an enhanced and amplified signal based on nanosurface interactions.

Nanobiosensing has evolved even further through the development of ultrasensitive, electrical biosensors known as *nanowire biosensors* (inorganic or semiconductor), which tend to be comparable in size of the biological or chemical species being sensed. This enables excellent primary transduction of signals to interfacing macroscopic instrumentation, through direct integration of nanodevices (e.g., nano-FET) with miniaturized systems.

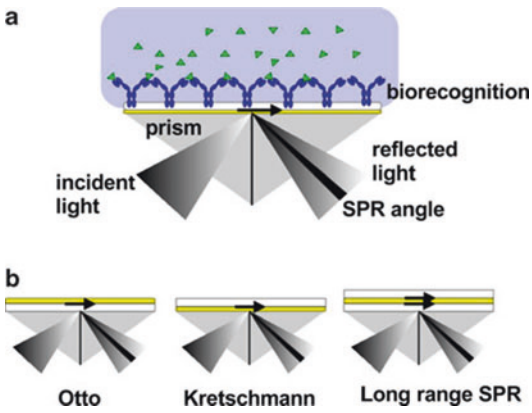
Optical-Based Biosensors

Variations in optical phenomena based on the resultant interactions of target analyte and bioreceptor form the basis of optical transducer technologies. Fiber-optic technology has been a key driving force in the development of many of these optical techniques for biosensor applications. As such the term *optrodes* has come to define devices which incorporate optic fibers as transduction elements in sensors. A fiber-optic biosensor is a device in which a reaction phase is associated with a transduction system made from optical fibers or bundles of optical fibers directly coupled to some form of light detector. The reaction phase (e.g., sol-gel film) contains immobilized bioreceptors which generates an optical signal upon biorecognition with target analytes, transmitting the signal through the optic fiber to the detector.

Covering different regions of the electromagnetic spectrum (UV, visible, IR, NIR), optical sensors are based on various optical principles (absorbance, fluorescence, luminescence, reflectance), which allow the measurement of a range of properties including light intensity, refractive index, scattering, diffraction, and polarization. Nanotechnology has played an important part in enhancing optical sensing techniques as will be highlighted.

Surface-Plasmon Resonance (SPR)

The most common technology used in optical biosensing is surface-plasmon resonance (SPR). Belonging to the class of refractometric sensing it exploits *surface-plasmon polaritons* (electromagnetic waves) to probe interactions of analytes in solution forming complexes with bioreceptors immobilized on the SPR sensor surface through adsorption. This highly sensitive method is based on the collective oscillatory excitement of electrons propagating between a metal film and dielectric interface, leading to a total absorption of light at a particular angle of incidence (Fig. 5). Surface plasmons are optically excited if the projection of the incident light wave vector along the x-axis matches the propagation constant of the surface plasmon (k_{sp}), described as:



Biosensors, Fig. 5 (**a**) Typical setup of a SPR biosensor, with the *black arrow* showing the direction of the SPP wave at the interface of the metallic surface and the dielectric (*white*); (**b**) typical configurations used in SPR biosensors

$$k_{sp} = \frac{\omega}{c}\left[\frac{\varepsilon_m \varepsilon_d}{\varepsilon_m + \varepsilon_d}\right]^{0.5} \quad (2)$$

where ε_m and ε_d are the dielectric constants of metal and dielectric, respectively, ω is the angular frequency of light, and c is the speed of light in vacuum.

A change to k_{sp} resulting from a change to the refractive index of the sensed medium, i.e., the dielectric, subsequently alters the resonance conditions of the surface plasmons interacting with the optical wave. Three relevant characteristics are critical to sensor design and performance, which are:

1. *Enhancement of electric field*: at the resonance angle, the intensity of the electric field at metal-dielectric interface is significantly enhanced due to the smaller complex permittivity in the dielectric compared to the metal.
2. *Propagation length*: the propagation constant, *ksp*, contains complex permittivity quantities for both metal and dielectric (Eq. 2) resulting in a complex wave vector. The imaginary component of the propagation constant (k_{sp}") represents the attenuation by metal absorption and radiative losses, implying the field intensity decays by a characteristic length of $1/2k_{sp}$". The surface plasmon decay length (δ_{sp}) is thus defined in Eq. 3, where λ_0 is the optical wavelength:

$$\delta_{sp} = \lambda_0 \frac{\left(\varepsilon'_m\right)^2}{2\pi\varepsilon''_m}\sqrt{\left(\frac{\varepsilon'_m + \varepsilon_d}{\varepsilon'_m \varepsilon_d}\right)^3} \quad (3)$$

3. *Penetration depth*: the other component of the parallel wave vector is the perpendicular wave vector to the interface which is an imaginary amplitude wave vector that decays in the z-direction. The penetration depth (δi), a couple of hundred nanometers in water, is defined as the distance in the perpendicular direction to the interface where the electromagnetic field inside the medium (i) falls to *1/e* (~37 %) and can be expressed by Eq. 4:

$$\delta_i = \frac{\lambda_0}{2\pi}\sqrt{\left|\frac{\varepsilon'_m + \varepsilon_d}{\varepsilon_i^2}\right|} \quad (4)$$

Optical excitation methods for surface plasmons include prism, grating, and waveguide coupling. The characteristic measurement of the resonance has lead to SPR sensors to be categorized into angular (monochromatic light), wavelength (polychromatic light), intensity, or phase modulation [13].

SPR-based biosensors work on the basis of an alteration to the refractive index near the surface of the functionalized sensor when a biomolecular interaction occurs. The performance of the SPR-based biosensor is highly dependent on the intrinsic optical performance of SPR sensor (sensitivity and resolution) and the characteristics of the functionalized surface. Advances in SPR sensor performance has led to increased sensitivity of systems containing immobilized biorecognition elements, with commercial systems having refractive index resolution (LOD equivalent) as low as 1×10^{-7} RIU (Refractive Index Unit). Researchers worldwide are continuing to construct novel coupling approaches, mainly based on Kretschmann's configuration (as depicted in Fig. 5b), for greater sensitivity such as the multilayer structure long range surface plasmons (LRSPs) which offer greater optical field enhancement, lower propagation loss, and longer field penetration (for virus and bacteria detection) over conventional surface plasmons. Optical fiber-based sensors and integrated optical waveguide SPR sensors offer the development of miniaturized systems for in vivo applications and multichannel sensing on a single chip [13].

Functionalization of the metallic surface (typically Au film) is commonly achieved through covalent attachment as this creates a stable and strong binding to the surface and allows regeneration of the sensor interface after removing analytes from the surface but not the attached bioreceptor. The functional self-assembled monolayers (SAMs) of thiolated organic compounds is a routine approach for covalent attachment of thiol groups of the bioreceptors, while polymeric layers such as carboxymethylated dextran (CM-dextran) thin films offer advantages of reduced binding interference, reduction in nonspecific binding by acting as a buffer, and increased attachment points over SAMs. Although gold thin films are commonly used

substrates for SPR biosensors because of their stability in aqueous environments and easy functionalization through alkanethiol SAMs, they have been found not to be the best candidates for high sensitivity SPR sensing. Indium tin oxide (ITO) has been reported to be a better candidate, though excitation and detection would occur in the infrared range. For conventional visible range excitation and detection, silver/gold layers have been used and lamellar overcoat structures, where a thin layer of a dielectric (oxide based) is deposited over silver metal thin film providing efficient protection from the aqueous solution which could damage the silver film [14]. There has also been a large focus on polymeric thin-film electrodeposition of oligomeric and conducting polymers on to gold surfaces for SPR biosensing. They have been used as they possess properties (hydrophobicity, pH, surface charge) which reduce nonspecific adsorption, which in turn increases the sensitivity and specificity of a biosensor. The conductive polymer polypyrrole has been intensively used in biosensors due to its biocompatibility, stability, and conductivity. Entrapment of bioreceptors in polypyrrole has been shown to be less attractive as the hydrophobicity of the substrate makes target molecule diffusion difficult. A more promising approach is the functionalization of the conducting polymer through covalent binding of the bioreceptor on the carboxyl groups. This method of functionalization has become a routine technique, based on electrospotting, for DNA array SPR chip fabrication.

The most common bioreceptor used on SPR-based biosensors are antibodies and commercial SPR systems (BIAcore^{TM}, real-time biospecific interaction analysis) have been developed for pharmaceutical drug discovery, antibody characterization, proteomics, immunogenicity, and much more in life science research. Essentially, SPR-based biosensing is a label-free detection method, but for signal amplification and enhanced surface chemistry new approaches which involve nanomaterials (e.g., CNTs, NP, nanowires) have been developed [13]. The effect of nanostructures on surface plasmons has led to a growing area of photonic studies known as nanoplasmonics. Metal-NPs are incorporated into SPR-based biosensors due to the lack of small molecule detection with conventional systems. As low molecular weight compounds produce a negligible change in the refractive index of the sensing surface, metal-NPs overcome this due to their localized SPR (LSPR) frequency (refer to ▶ Optical Properties of Metal Nanoparticles). This occurs as a result of the incident light interacting with particles much smaller than the incident wavelength, and the presence of metal-NPs is characterized by intense adsorption and light scattering. The enhanced electric field has shown to give an order of magnitude increase in sensitivity (5×10^{-8} RIU) and has enhanced other spectroscopic methods such as surface-enhanced Raman spectroscopy (SERS). For instance, SPR signal output has been increased through mass or enzymatic processes at the bioreceptor-ligand complex. Immobilized peptide nucleic acids (PNA) for point mutation detection were hybridized with DNA-modified nanoparticles. The LOD for this ultrasensitive mismatch was as low as 1 fM using this technique [15]. A similar approach was applied to Poly-T modified nanoparticles, which bind to surface-immobilized Poly-A tails forming a ternary surface complex, with LOD for 19–23 mer as low as 10 fM.

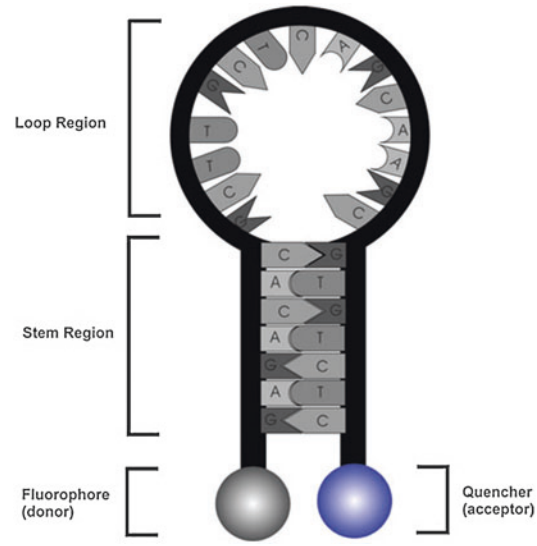
Experimental measurements based on labeling may be problematic due to metal-NP influences on binding kinetics and biomolecular interactions. A different approach toward incorporating nanotechnology into SPR-based systems focuses on the sensing matrices, which alleviate problems associated with biomolecular labeling. SPR-based immunosensors had immobilized AuNPs encapsulated in dendrimers on SAM-modified gold surfaces providing an enhanced surface density for analyzing insulin, while the sensitivity of SPR-based hormone assays were significantly enhanced using AuNPs [14]. A sensing medium which had gold embedded in a hydrogel and functionalized for small molecule detection swelled upon biorecognition, which increased the distance between AuNPs in the hydrogel. A shift in the SPR curve to higher angles was a result of the intercorrelation of both SPR and LSPR.

Molecular Beacons and Quantum Dots

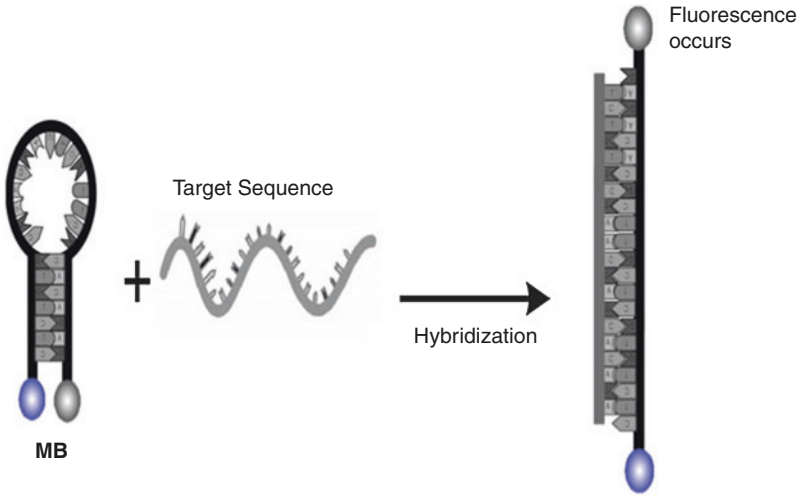
Luminescence is the emission of electromagnetic radiation in the ultraviolet, visible, and infrared spectra from atoms or molecules as a result of the transition of electronically excited states to a lower energy state, usually the ground state. There are various forms of luminescence (e.g., phosphorescence, bioluminescence, chemiluminescence, and fluorescence) which have been applied to biosensors and fiber-optic biosensing for a diverse range of activities ranging from glucose monitoring and genosensors to environmental monitoring and biopharmaceutical process monitoring.

Fluorescence is a widely used optical technique in biosensing, and there are three types of sensing employed. First is direct sensing when a specific target molecule is detected before and after a reaction or change takes place. The second form involves an indirect detection by the addition of a dye (e.g., FITC, GFP) which will tag the presence of the target molecule. The final form is known as fluorescence resonance energy transfer (FRET), which generates a unique fluorescence signal. Developed by Förster and Weber, FRET requires two probes: a donor (fluorophore) and an acceptor (chromophore) [10, 16]. Irradiating the donor with light of an appropriate wavelength produces an oscillating dipole which resonates with the dipole of the acceptor in close proximity. This dipole–dipole interaction is a radiationless transfer of energy between the donor and acceptor. As such if the acceptor is another fluorophore with an excitation wavelength that overlaps with emission wavelength of the donor, the transfer of energy from the donor to the acceptor will stimulate fluorescence of the acceptor. The two major strategies used to develop FRET biosensors are: (1) two chain probes in which the fluorophores are on two different molecules resulting in intermolecular FRET when two molecules come into close proximity to each other, or (2) FRET due to intramolecular conformational change based on single-chain probes with fluorophores attached to different regions of a single molecule.

Optical DNA nanobiosensors have been developed through the use of molecular beacons (MB) and quantum dots (QD) in optical nano-imaging systems. MBs are single-stranded nucleic acid probes consisting of a stem, a loop, and a fluorophore/quencher pair (Fig. 6). The close proximity (~10 nm) of the donor fluorophore and acceptor quencher in the hairpin, made possible by the "locking" mechanism of the stem region (~4–7 nucleobase pairs), effectively turns off the fluorescence via FRET. They are widely used in fluorometric analysis of nucleic acids. A conformational change of the MB occurs upon hybridization with the complementary target (ssDNA or RNA) which increases the physical separation distance between the fluorophore and quencher, thus providing a fluorescence signal (Fig. 7). Different color fluorophores can be used simultaneously for detection of multiple targets in the same solution. This label-free detection tool is sensitive and its importance in biosensor technology offers high selectivity with real-time monitoring of hybridization for analysis of gene polymorphism in genetically modified organisms, polymerase chain reaction (PCR), and ligase reactions. They have been used to detect bacteria, viruses, and other microorganisms in a variety of conditions; studies of genotyping, genetic disorders associated with point mutations, and drug discovery have been conducted with various



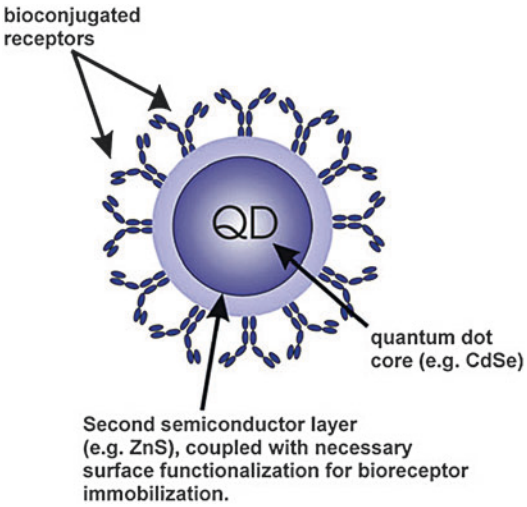
Biosensors, Fig. 6 A molecular beacon consisting of a loop region, stem region, and a fluorophore/quencher pair separated by ~10 nm, inhibiting fluorescence



Biosensors, Fig. 7 Hybridization reaction of a molecular beacon with a target sequence oligomer. The increased separation distance between the fluorophore and the quencher allows fluorescence to occur

aspects of MBs. Other innovative approaches for using MBs as nanobiosensors have included visualizing the transport and distribution of mRNA; incorporating MBs in flow cytometry studies for oxidative stress analysis; detecting mutations in the *rpoB* of *Mycobacterium tuberculosis* (TB) via MB-qPCR as a screening tool for multidrug-resistant tuberculosis; real-time detection of dsDNA; and linking MBs to the fuse silica surface for drug delivery. These applications highlight the importance of MBs as a powerful tool in intracellular monitoring of nucleic acids and in the ongoing development of highly selective optical biosensors.

QDs are nanoscale semiconductor structures grown by epitaxial techniques in nanocrystals by chemical methods. They consist of a colloidal core (e.g., cadmium selenide, cadmium sulfide, and indium arsenide) covered with surface coatings which decrease leaching of the metal from the core (Fig. 8). Based on the reduced size of QDs, they possess discrete electronic energy which gives rise to their unique optical properties and their use as a fluorescent probe for biomolecular imaging or even as a donor in molecular beacons [17]. In comparison to organic dyes, and fluorescent proteins, QDs are controllable in many ways including their size-oriented optical and electronic properties, tunable light emission, resistance against photobleaching, and simultaneous excitation of multiple fluorescence



Biosensors, Fig. 8 Quantum dot for nanobiosensing applications conjugated with antibodies on the outer surface

colors [10]. More recently, development of heavy metal-free QDs with similar optical properties has overcome issues regarding cytotoxicity presented by heavy metals such as cadmium. Their photo-stable fluorescence and water solubility upon coating with amphiphilic polymers has seen QDs being used in biodetection systems, biosensing and high-throughput bio-analytical systems, quantitative cellular investigative studies, and nanobiomedicine.

Surface immobilization of antibodies, DNA, proteins, biopolymers, peptides, and nucleic acid on QDs has significantly advanced bio-imaging and biosensing over the past decade. Bioreceptor attachment via cross-link molecules with diverse reactive groups, e.g., –SH, NH_2, or –COOH, is attained through covalent or electrostatic interactions. A broad absorption spectra with high extinction coefficients and a full width at half maximum (FWHM) of 20–30 nm is afforded by QDs. Emissions of QDs can be quenched by gold nanoparticles (AuNPs), and this has been exploited in fluorescent assays for DNA detection where QDs and AuNPs were used as a FRET donor-acceptor couple based on cadmium selenide QDs bioconjugated with a short DNA strand and hybridized with a AuNP-DNA strand. The development of a single-QD-based nanobiosensor for detection of multiple DNA of HIV-1 and HIV-2 has been developed [10] using the nanobiosensor as a local nanoconcentrator, capable of significantly amplifying both fluorescence signal and FRET signals, with potential use for point of care diagnostics and high-throughput screening. QDs have also been applied in chip-based DNA microarrays for analyzing single nucleotide polymorphism and multi-marker discovery, with anticipated LODs as low as 2 nM concentrations of DNA-NP probes. Interestingly, the use of QDs has been discovered in gene technology, where QD-FRET has been utilized in labeling and recognizing target sequences of DNA with red, blue, and green QDs in different combinations. Discrimination and detection of a single base pair mismatch has been reported using a QD-DNA nanobiosensor, potentially leading to very advanced applications in clinical diagnostics of DNA mutants with drug resistance. Chemiluminescence and bioluminescence resonant energy transfer (CRET and BRET, respectively) generate an excited state donor through chemical and biochemical reactions and are believed to be more advantageous than FRET due to lower background from optical excitations [17].

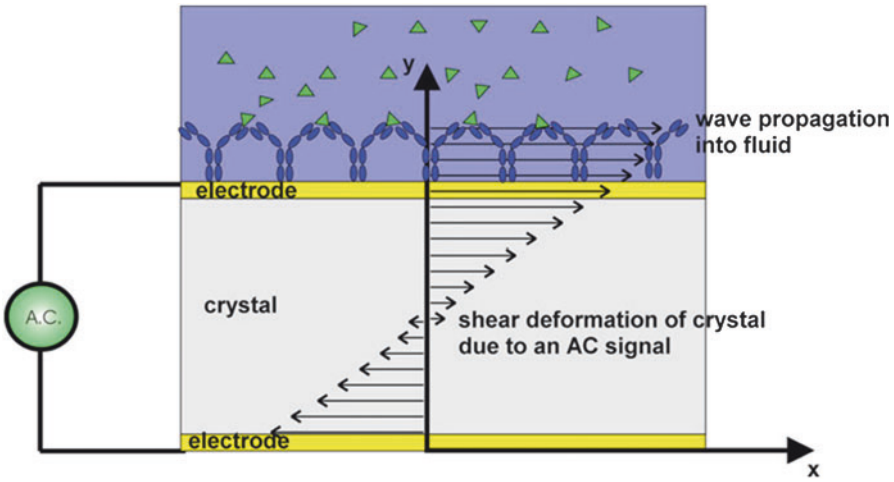
Mass Sensitive Biosensors

Piezoelectric crystals, more commonly referred to as quartz crystal microbalances (QCM) are mass sensitive devices which transform a mass loading at the gold electrode sensor surface, due to adsorbing analytes, into a change in the intrinsic resonance frequency of the QCM. When a low voltage AC signal is applied to thin layer electrodes sandwiching a quartz crystal, the piezoelectric nature of the crystal produces a shear (tangential) deformation causing both surfaces to move in parallel, but in opposite directions, thereby producing acoustic waves which propagate through the bulk of the crystal perpendicular to the surface (Fig. 9). The resonance frequency of the crystal depends on the shear acoustic wave velocity and the thickness of the crystal substrate. The fundamental resonance frequency and odd overtones are the only harmonics which can be excited as predicted using the shear acoustic wave (ν) propagation equation. Also, as ν is very dependent on the properties of the propagating material, i.e., Young's Modulus and density, thickness shear mode (TSM) resonators are well suited for determination of material-specific parameters, which are difficult to obtain using other biosensing techniques such as SPR.

Able to operate in three modes (i.e., flexural, extensional, and thickness shear), thickness shear mode is the most sensitive with the resonant frequency change being a measure of the thin layer mass attached to the crystal surface. It is also the only mode which can be electrically excited with a low voltage, as the other modes have a coupling factor of zero. In a vacuum, the relation between mass and resonant frequency is considered to be proportional to the added thin layer solid mass on the device surface as described by the Sauerbrey equation (Eq. 5), where f_0 is the resonance frequency of the unperturbed quartz resonator, F_q is the frequency constant of the crystal, ρ_q is the quartz density, and A_{el} is the electrode area:

$$\Delta f_m = -\frac{f_0^2}{F_q \rho_q} \frac{\Delta m_s}{A_{el}} \tag{5}$$

To take into consideration a QCM in contact with fluids, Kanazawa and Gordon proposed a theoretical relationship which factors in the material properties of both quartz and a Newtonian



Biosensors, Fig. 9 Illustration of a quartz crystal microbalance biosensor depicting the shear acoustic wave propagating in the bulk of the crystal and into the fluid in contact with one surface. Typical decay lengths of the wave into the fluid are of the order of nanometers, making ligand-complex interactions with surface-immobilized nanoscale bioreceptors and thin film deposition of materials on the surface detectable due to a change in resonance frequency

fluid as defined in Eq. 6, where μ_q is the shear module in the x-direction, ρ_l is the density of the liquid, and η is the liquid viscosity:

$$\Delta f_1 = -\sqrt{f_0^3 \frac{\eta_l \rho_l}{\pi \mu_q \rho_q}} \tag{6}$$

Suitable oscillator circuits have been designed for the application of these sensors to biological detection in liquid samples, combined with modified surface chemistries of the sensing surface for the immobilization of bioreceptors. QCMs typically operate at frequencies between 1 MHz and 10 MHz, with a frequency-determination limit of 0.1 Hz and a LOD of mass bound to the electrode surface of 10^{-10}–10^{-11} g, subject to calibration of the instrumentation. QCM-based biosensors have been used extensively by researchers for detection of micron-sized bacteria such as *Escherichia coli* and to study the mechanisms involved in surface adhesion and formation of microbial biofilms [18]. The sensitivity of QCMs has been shown to increase with the incorporation of NPs.

It is appropriate to mention here that DNA biosensors, generically known as genosensors, have also been applied to QCMs. These genosensors monitor and detect in real-time DNA hybridization events, through immobilized macromolecules. An AuNP-amplified QCM DNA biosensor for *E. coli* detection had AuNPs immobilized on the surface of the QCM, increasing the surface area of the sensor and allowing more ssDNA probes to be absorbed on to the surface which then captured the biotinylated *E. coli* DNA. Amplification has been enhanced further using outer-avidin-coated AuNPs giving an LOD as low as 2×10^3 CFU/mL. A similar technique has recently been applied for the specific and rapid detection of avian influenza virus H5N1. A magnetic nanobead amplification-based QCM immunosensor was fabricated, whereby polyclonal antibodies against the AI H5N1 virus surface antigen hemagglutinin were immobilized on the gold surface through SAMs of 16-mercaptohexadecanoic acid. Capture of the virus on the immobilized antibodies resulted in a decrease in resonant frequency of the QCM, which was further amplified by the magnetic nanobeads coated with anti-H5 antibodies bound between antibody and antigen.

The real-time detection of lysozymes on a QCM nanosensor has been achieved using immobilized molecular-imprinted polymers (MIPs) giving a detection limit of 1.2 ng/mL and

a high selectivity and sensitivity in a concentration range of 0.2–1,500 μg/mL. They include the use of QCMs in bacteria detection, hemostasis studies; immunoassay studies; carbohydrate binding; oligonucleotide interactions; protein, enzyme kinetic, and cellular studies; and small molecule interactions via MIP and small molecule surface immobilization [18]. The vast majority of these applications had the gold surface of the QCM chemically functionalized through SAMs of both natural and artificial bioreceptor elements which aided in the detection of specific entities, thereby increasing the limit of detection and quantification significantly in comparison to some optical techniques.

Microarray Biosensors

Microarrays, also known as biochips, are 2D or 3D molecular receptor arrays which allow for the simultaneous detection of a large number of substances through the interaction of target analytes with substrate-bound bioreceptors. There are several categories of microarrays (e.g., DNA, proteins, aptamers, whole-cell) which typically comprise several hundred to several thousand "spots" or "zones" on which the bioreceptor (s) can be immobilized. The applications of these microarray biosensors are far-reaching and have been used in gene expression, protein-expression profiling, point mutation analysis, immunodiagnostics, drug screening, cell analysis, and DNA sequencing to name a few. A brief description of the most common microarray technologies to date is provided.

DNA microarrays consist of immobilized cDNA (<5,000 b.p) or oligonucleotides (<80 b.p) which interrogate target analytes (e.g., DNA) in sample solutions. In principle, the assays developed largely depend on the specific molecular recognition between complementary sequences of nucleobases. Depending on the method of detection or "microarray reading," bioreceptors have usually been immobilized on glass/silicon substrates, metal-coated (i.e., ITO, gold) glass substrates, or polymeric substrates. Companies (i.e., Affymetrix) have developed precision equipment for high-density "spotting" of bioreceptors on functionalized substrates for rapid screening studies, along with appropriate microarray reading software. The most widely used detection method for DNA microarrays is fluorescence detection. Target molecules are tagged with a fluorophore and upon hybridization with the bioprobe, the spot or zone will fluoresce. Other forms of detection for DNA microarrays include chemiluminescence/electrochemiluminescence. DNA microarray technologies have been used to map the genome sequence of multiple organisms including humans and yeast.

Protein microarrays can generally be classed as either analytical/diagnostic or functional microarrays. They were developed upon completion of genome mapping, leading to the prediction of open reading frames (ORF) from genomes making it possible to express the full or partial proteome with large-scale cloning and gene expression methods. An analytical form of these microarrays is the antibody microarray which forms complexes with the target protein or antigen, and offers tremendous potential in clinical research and diagnosis. The functional form of protein microarrays which possess high-density proteins or peptides can represent complete or partial proteome of an organism. They are mainly used in studies involving protein–protein, protein–DNA, and protein–small molecule interactions, also offering wide-ranging applications in clinical cancer studies, viral infections, and autoimmune disease studies [19].

A modified analogue to DNA is peptide nucleic acids (PNA), which possesses a neutral peptide backbone (polyamide) instead of a negatively charged phosphate-sugar backbone. In comparison to DNA, PNA exhibits high chemical stability, resistance to enzymatic degradation, the ability to recognize specific nucleic acid sequences, strand invasion capabilities, high thermal stability, and ionic strength and the ability to hybridize with nucleic acids. Since their discovery in 1991, researchers have shown PNA-based biosensors to be extremely versatile in a range of systems including electrochemical, mass, and SPR-based on its unique biophysical properties. For example, μ-patterning of DNA on to microarrays has been achieved using electrohydrodynamics (EHD). DNA was found to be

affected by thermal and electrical properties upon immobilization, but the use of PNA in the patterning process did not affect its function or structural integrity.

Future Trends

Biosensors have undergone a rapid transformation over the past half century. It continually remains in a state of flux with innovative applications of nanotechnology for even faster, cheaper, and smaller biosensing systems for areas as diverse as personalized healthcare, point of care (POC) diagnostics, and rapid biodetection of pathogens in bioterrorism. These are being realized as microfabrication and miniaturization of electronic and transduction elements become incorporated into systems known collectively as lab-on-a-chip (LOC) systems or micro-total analysis systems (μ-TAS). These systems rely on the complete processing and transportation of biological samples on a single microchip through micro- and nanofluidic channels, from sample preparation to sample analysis through advanced biosensing techniques. Novel material properties, the development and integration of particle manipulation techniques, and increasing processing power will drive the next generation of biosensing nanosystems.

Graphene, a carbon allotrope isolated from graphite, has been reported as a new alternative to CNTs and metal alloy NPs ever since its discovery in 2004 [20]. Advances in graphene-based biosensors show significant advantages over CNT and metal-NPs. Purported problems associated with CNTs and metal-NPs, such as inconsistent signal amplification and metal impurities interfering with redox reaction with the biomolecules are thought to be avoided with the use of graphene in electrochemical nanobiosensor research. There is still a need of bioconjugation protocol development for bioreceptor immobilization on graphene before the uptake of electrochemical graphene-based biosensors becomes more commonplace.

Particle manipulation through AC electrokinetic phenomena, including electrowetting deposition (EWOD), *dielectrophoresis* (DEP), and *AC electro-osmosis* (ACEO) have shown tremendous potential in bionanotechnology. They have shown simplicity in integration with various biosensing transducers, and coupled with high specificity and sensitivity for particle manipulation in a range of media solutions, the rate of detection in biosensor systems is significantly greater than conventional methods (i.e., adsorption and electrostatic attraction) which rely heavily on bringing the target analyte in contact to the bioreceptor for high-quality sensing.

Cross-References

- ▶ AC Electrokinetics of Nanoparticles
- ▶ Arthropod Strain Sensors
- ▶ Carbon Nanotubes
- ▶ Dielectrophoresis
- ▶ Electrospinning
- ▶ Graphene
- ▶ Nanoparticles
- ▶ Nanostructured Functionalized Surfaces
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Quantum Dot Toxicity
- ▶ Self-Assembly of Nanostructures
- ▶ Sol–Gel Method
- ▶ Surface Plasmon-Polariton-Based Detectors
- ▶ Synthetic Biology

References

1. Clark Jr., L.C., Lyons, C.: Electrode systems for continuous monitoring in cardio vascular surgery. Ann. N. Y. Acad. Sci. **102**, 29–45 (1962)
2. Updike, S.J., Hicks, G.P.: The enzyme electrode. Nature **214**, 986–988 (1967)
3. Blum, L., Marquette, C.: Biosensors. From the glucose electrode to the biochip. In: Boisseau, P., Houdy, P., Lahmani, M. (eds.) NanoScience – Nanobiotechnology and Nanobiology, pp. 871–909. Springer, New York (2010)
4. Kress-Rogers, E.: Handbook of Biosensors and Electronic Noses. Taylor and Francis/CRC Press, Boca Raton (1996)
5. Su, L., Jia, W., Hou, C., Lei, Y.: Microbial biosensors: a review. Biosens. Bioelectron. **26**, 1788–1799 (2011)
6. Haupt, K.: Molecularly imprinted polymers in analytical chemistry. Analyst **126**, 747–756 (2001)
7. Jerónimo, P.C.A., Araújo, A.N., Montenegro, M.C. B.S.M.: Optical sensors and biosensors based on sol–-gel films. Talanta **72**, 13–27 (2007)

8. Siqueira Jr., J.R., Caseli, L., Crespilho, F.N., Zucolotto, V., Oliveira Jr., O.N.: Immobilization of biomolecules on nanostructured films for biosensing. Biosens. Bioelectron. **25**, 1254–1263 (2010)
9. Vashist, S.K., Zheng, D., Al-Rubeaan, K., Luong, J.H. T., Sheu, F.-S.: Advances in carbon nanotube based electrochemical sensors for bioanalytical applications. Biotechnol. Adv. **29**, 169–188 (2011)
10. Dolatabadi, J.E.N., Maschinchian, O., Ayoubi, B., Jamali, A.A., Mobed, A., Losic, D., Omidi, Y., Guardia, M.: Optical and electrochemical DNA nanobiosensors. Trends Anal. Chem. **30**(3), 459–472 (2011)
11. Xia, L., Wei, Z., Wan, M.: Conducting polymer nanostructures and their application in biosensors. J. Colloid Interface Sci. **341**, 1–11 (2010)
12. Teles, F.R.R., Fonseca, L.P.: Applications of polymers for biomolecule immobilization in electrochemical biosensors. Mater. Sci. Eng. C-Biomim. Supramol. Syst. **28**, 1530–1543 (2008)
13. Abbas, A., Linman, M.J., Cheng, Q.: New trends in instrumental design for surface plasmon resonance-based biosensors. Biosens. Bioelectron. **26**, 1815–1824 (2011)
14. Wijaya, E., Lenaerts, C., Maricot, S., Hastanin, J., Habraken, S., Vilcot, J.-P., Boukherroub, R., Szunerits, S.: Surface plasmon resonance-based biosensors: from the development of SPR structures to novel surface functionalization strategies. Curr. Opin. Solid State Mater. Sci. **15**(5), 208–224 (2011)
15. Scarano, S., Mascini, M., Turner, A.P.F., Minunni, M.: Surface plasmon resonance imaging for affinity based biosensors. Biosens. Bioelectron. **25**, 957–966 (2010)
16. Ibraheem, A., Campbell, R.E.: Designs and applications of fluorescent protein based biosensors. Curr. Opin. Chem. Biol. **14**, 30–36 (2010)
17. Algar, W.R., Tavares, A.J., Krull, U.J.: Beyond labels: a review of the applications of quantum dots as integrated components of assays, bioprobes, and biosensors utilizing optical transduction. Anal. Chim. Acta **673**, 1–25 (2010)
18. Becker, B., Cooper, M.A.: A survey of the 2006–2009 quartz crystal microbalance biosensor literature. J. Mol. Recognit. **24**(5), 754–787 (2011)
19. Chen, H., Jiang, C.J., Yu, C., Zhang, S., Liu, B., Kong, J.: Protein chips and nanomaterials for application in tumor marker immunoassays. Biosens. Bioelectron. **24**, 3399–3411 (2009)
20. Kuila, T., Bose, S., Khanra, P., Mishra, A.K., Kim, N. H., Lee, J.H.: Recent advances in graphene-based biosensors. Biosens. Bioelectron. **26**(12), 4637–4648 (2011)

Bond-Order Potential

► Reactive Empirical Bond-Order Potentials

Bone Remodeling

J. A. Sanz-Herrera and E. Reina-Remo
School of Engineering, University of Seville, Seville, Spain

Synonyms

Bone turnover

Definition

Bone remodeling is a dynamic physiological process in which the combined effect of bone formation and bone resorption occurs at a specific location of the bone architecture to enable bones to adapt to mechanical stresses, to repair its microstructure and thus maintain the mechanical integrity of the skeleton, and to maintain mineral homeostasis. Bone remodeling consists of two main subprocesses, bone resorption by osteclasts followed by bone formation by osteoblasts without causing large changes in bone quantity, geometry, or size. An imbalance in the regulation of bone remodeling's two subprocesses results in many metabolic bone diseases, such as osteoporosis.

Overview

Bone is a metabolically active tissue capable of adapting its structure to mechanical stimuli and repairing structural damage. It has the capability of forming new osseous tissue at locations that are damaged or missing, such as in fracture healing, distraction osteogenesis or bone implants interface, or removing bone matrix in case of disuse. (Distraction osteogenesis: surgical process in which new bone formation is induced through gradual traction in an osteotomy site to reconstruct skeletal deformities or lengthen bones.) This calcified, vascular, connective tissue has many functions, including support, protection, mineral storage, blood cell production,

hematopoiesis, and locomotion through the attachment of muscles.

There are three basic mechanisms involved in the turnover and development of bone: the longitudinal growth, modeling, and remodeling. The former, the longitudinal growth, ceases with closure of the epiphyseal plates after the growing period. The second, modeling, serves to alter the bone that is present and to determine its geometry and size [1]. Bone modeling at any surface involves an uncoupled activity of osteoclast activation (A) followed either by formation (A-F) or resorption (A-R). (Osteoclast: specialized cells capable of bone resorption.) The latter, bone remodeling, starts with fetal osteogenesis and continues throughout life. It involves the coupled action of the osteoclastic and osteoblastic activities at localized areas on the bone surface (A-R-F) and differs from bone modeling, in which these processes occur independently at distinct anatomical locations (Osteoblast: bone forming cells). By the process of remodeling, bone is continuously resorbed and rebuilt at about one to two million microscopic sites per adult skeleton by teams of cells or remodeling units. These remodeling units are structures which persist for a variable but unknown period of time, traveling through the bone structure to carry out the functions of bone turnover in an ordered and predictable manner. Osteoclast–osteoblast interaction is needed for coupling bone resorption and formation both temporarily and spatially [1, 2]. Temporarily, since in each remodeling cycle, there occurs a succession of cellular events in the same place but at different times. Spatially, since in each remodeling unit, the same events are occurring in different places but at the same time.

Biological Principles

Bone Structure

At the structural level, mature bone can be classified into compact/cortical bone and cancellous/trabecular bone. *Cancellous bone* has large, open spaces surrounded by thin, interconnected plates of bone. The large spaces contain red bone marrow where hematopoiesis occurs and the plates of bone are trabeculae composed of several layers of lamellae. By contrast, *compact bone* is much denser than cancellous bone, with spaces reduced in size and lamellar organization more precise and thicker. The typical range of apparent density of cancellous bone is 1.0–1.4 g/cm^3 compared to about 1.8–2.0 g/cm^3 for compact bone [3]. (Apparent density: mass per unit volume of a region of bulk bone, which includes Haversian canals, marrow spaces, and other soft tissue spaces.) Cancellous bone can be found in the epiphysis or heads of long bones while cortical bone is found in the diaphysis or shaft of the bone and the outer layer of the trabecular bone. In the entire skeleton around 80 % of the total skeletal mass is composed of cortical bone and only 20 % is cancellous bone and the volume fraction (including internal pores like lacunae and canaliculi) is 0.8 mm^3/mm^3 and 0.2 mm^3/mm^3, respectively.

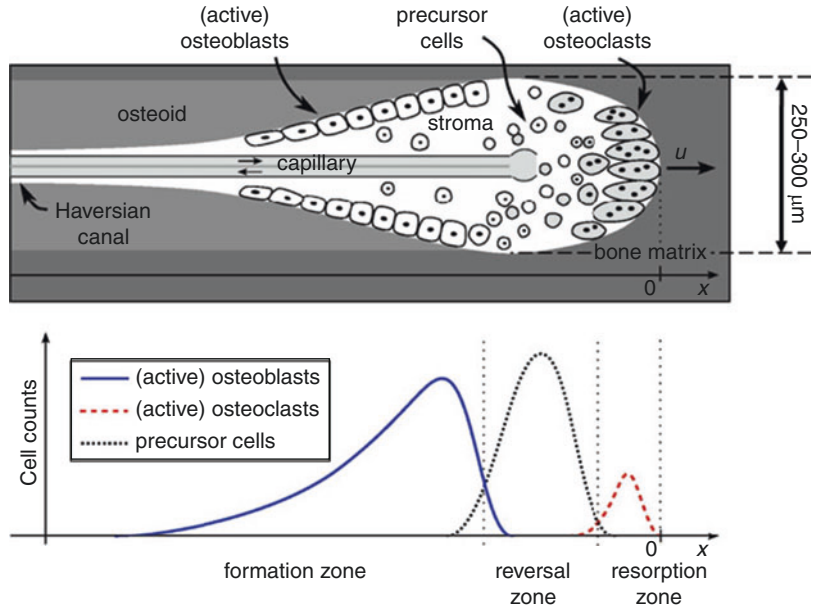
At the tissue level, bone may be divided into two wide categories:

- *Woven bone/primary bone*: This type of bone is laid down rapidly as a disorganized arrangement of collagen fibers and osteocytes. In adult skeleton, this young regenerated bone can be found in a fracture callus, during distraction osteogenesis and in areas undergoing active endochondral ossification.
- *Lamellar bone/secondary bone*. This mature bone can exist in cortical and trabecular bone. Unlike woven bone, this tissue is well organized and regular.

The Remodeling Cells

The remodeling process requires the synchronized activities of molecular mechanisms and multiple cellular participants, such as osteoblasts, osteoclasts and osteocytes to ensure that bone formation and bone resorption occur sequentially.

- *Osteoclasts*: Are multinucleated cells derived from hematopoietic cells formed by the fusion of mononuclear progenitors of the monocyte/macrophage family. They are the principal resorptive cells of bone by secretion of acid and proteases.



Bone Remodeling, Fig. 1 Cross-sectional diagram of a BMU in cortical bone. Osteoclasts resorb the bone matrix at the front while osteoblasts lay down osteoid toward the back to refill the cavity. The central capillary provides a supply of precursor cells, as well as various nutrients [5]

- *Osteoblasts*: Are mononuclear bone cells of mesenchymal origin that have several important roles in bone remodeling: expression of osteoclastogenic factors, production of bone matrix proteins (osteoid) and bone mineralization. It represents a heterogeneous family of cells, including immature osteoblast lineage cells, and differentiating and mature matrix-producing osteoblasts.
- *Osteocytes*: Terminally differentiated osteoblasts embedded in the mineralized collagen matrix that are likely to determine which bone surface osteoclasts will resorb. These cells are the most abundant cells in bone, accounting for 90–95 % of all bone cells. They may act as sensors for the chemical and mechanical signals, such as microcracks and loss of mechanical loading. Osteocytes communicate with each other and with cells on the bone surface, via the dendritic processes encased in canaliculi (Canaliculi: microscopic canals between the various lacunae of ossified bone).

The Remodeling Unit

The remodeling process is not performed individually by each cell but by groups of cells acting as organized units. Clusters of osteoclasts and osteoblasts are arranged within temporary anatomical structures known as "basic multicellular units" (BMUs). A BMU consists of around 20 osteoclasts and several hundred osteoblasts. Active BMUs consists of a leading front or "head" of bone resorbing osteoclasts and a tail portion occupied by osteoblasts (Fig. 1). The "head" contains a capillary bud to supply nutrients and probably to supply the progenitor cells for osteoclasts and osteoblasts. BMUs are not permanent structures since they form in response to a specific signal, perform their function and die, leaving a few residual lining cells and osteocytes. The number of BMUs working in a specified volume of tissue depends on the "activation frequency" and the "sigma period" [4]. The former refers to the frequency with which new remodeling cycles are created per unit cross-sectional area of arbitrary thickness and thus is a measure of the rate of bone turnover. The latter refers to the speed with which the BMU travels through the tissue space. It quantifies the time of completion of one remodeling cycle at a given location in the bone. Sigma periods are frequently subdivided into resorption and formation periods. Formation is much slower than resorption. In humans, the total remodeling cycle is about 4 months, about 3 weeks for resorption, and about 3 months for rebuilding the lost bone.

The Remodeling Cycle

There are three principal stages in the BMU lifetime: activation of osteoclasts, resorption of the old bone, and formation of new bone matrix (A-R-F) (Fig. 1). The spatial and temporal arrangements of the cells within the BMUs ensure a coordination between the different phases of the remodeling process. The A-R-F process may be further divided into six separated phases: activation, resorption, reversal, formation, mineralization, and quiescence (see Fig. 1):

- *Activation phase*: Activation takes place when an activation signal or stimulus, including chemical, mechanical, and electrical signals, converts a previously inactive bone surface into a remodeling one. It is important to distinguish between origination and activation. The former occurs only once for each BMU at a quiescent surface while the latter is a continuous process that occurs at the cutting edge of the BMU. As the lifespan of the individual cells in a BMU is much shorter than that of the BMU, activation involves gathering of the cells that will form the BMU. In both cortical and cancellous bones, osteoclasts participate in activation or recruitment. In addition, bone lining cells of osteoblastic nature are available on the cancellous surfaces. Systemic hormones, growth factors, and interleukins play a role at this step, helping to recruit new osteoclasts by enlarging the precursor pool.
- *Resorption phase*: Bone resorption is a multistep process initiated by the proliferation of immature osteoclast precursors, the commitment of these cells to the osteoclast phenotype, and finally the degradation of the organic and inorganic phases of bone by the mature resorptive cells. It involves the following steps:
  - Attachment of the osteoclasts to the target matrix: Newly created osteoclasts attach themselves to a bone surface and form a ruffled membrane.
  - Acidification of resorption lacunae. Osteoclasts acidify the extracellular microenvironment. For instance, they secrete both hydrochloric acid, to resorb hydroxyapatite, and proteases like cathepsin K to degrade collagen and other bone matrix proteins.
  - Bone demineralization. The acidic milieu mobilizes the mineral phase of bone and provides an optimal environment for organic matrix degradation.
  - Dissolution of the bone tissue forming a resorption lacuna.
  - The products of bone degradation are endocytosed by the osteoclast, transported, and finally released to the extracellular space.
  - Detachment of osteoclasts and movement to a new site of bone degradation.
- *Reversal phase*: The transition from osteoclastic to osteoblastic activity takes several days (around 9 days) and results in a cylindrical space lying between the resorptive region and the refilling region [1]. The length of this region may vary considerably, depending on the lag between the resorptive and formative phases.
- *Formation phase*: In this step, the osteoblasts initially synthesize the organic matrix called osteoid to restore the previously resorbed bone. The apposition rate of matrix and of mineral is more rapid initially, with an initial rate of 1–2 μm/day which gradually decreases to zero as the cavity is filled. As an individual ages, the amount of bone deposited in the resorption lacunae by osteoblasts is less than that previously removed by osteoclasts. This negative balance of osteoblastic relative to osteoclastic activity is responsible for the loss of bone that occurs with age and that when pronounced leads to osteoporosis.
- *Mineralization phase*: Following deposition of the unmineralized bone matrix, called osteoid, mineral is deposited within and between the collagen fibers. This process is delayed by a period of time known as the mineralization lag time, which is normally about 15 days. Once begun, approximately 60 % of the mineralization of the osteoid occurs during the first few days [4]. This is called primary mineralization. The remainder of the mineral is added at a decreasing rate for around 6 months during the secondary mineralization.

- *Quiescence*: Once the resorption and formation processes are completed, the osteoclasts disappear, and the osteoblasts become osteocytes or Haversian canal lining cells, or they disappear.

Cortical Versus Trabecular Remodeling

Bone remodeling has been well described in cortical [1] and cancellous bone [6]. Current evidence indicates that the differences in the remodeling activities of these two types of bone reside in the different mechanical environments of the bone cells in cortical or cancellous bone.

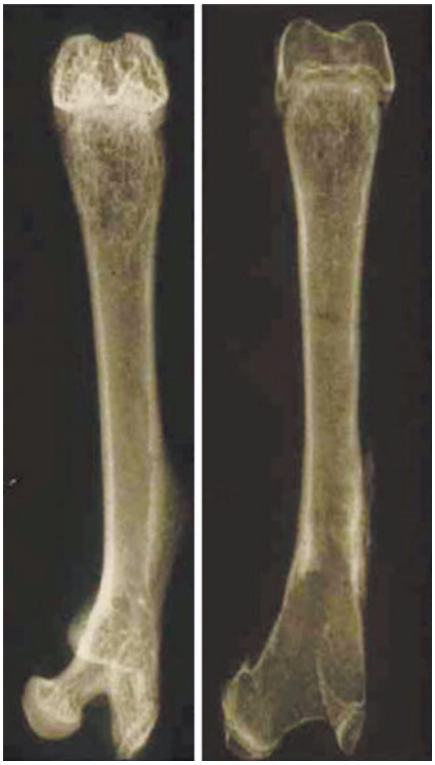
For instance, in *cortical remodeling*, since BMUs become isolated within the cortex and osteoclasts and osteoblasts need nourishment, the tunnel cannot be entirely refilled during the formation phase. Each BMU leaves a central passageway in the bone to support the metabolism of the BMU and bone matrix osteocytes, and to carry calcium and phosphorous to and from the bone when necessary. This passageway is called Haversian canal and is typically 40–50 μm in diameter in humans and contains two capillaries: a "supply" and a "return" vessel. In human adults, osteonal BMUs replace about 5 % of compact bone each year [3].

In *trabecular and endocortical surfaces* since BMUs work on trabeculae surfaces, the cells do not dig and refill tunnels but trenches (half-tunnels). Because a BMU is about the same diameter as a trabecula, the remodeling unit in cancellous bone is equivalent to half of a cortical remodeling unit. BMUs replace human adult trabeculae bone at a higher rate than cortical bone, about 25 % each year [3] since the surface/volume relationships are much greater in the cancellous bone. Another difference is that cancellous bone BMUs proceed along a surface lined with lining cells, which could participate in remodeling, whereas there are no lining bone cells within the cortex. Cancellous bone has also more ready access to marrow cells than cortical bone.

Key Research Findings

Bone remodeling is the final phenomenological consequence, observable at the bone microarchitectural level, of a cascade of biological events which take place at different spatial scales ranging from molecular to cell/tissue levels. The nature of the phenomena involved in the process of bone remodeling is diverse including molecular mechanisms due to genetic factors or interaction between cells, where mechanics plays an important role. The importance of the former (genetics) versus the latter (mechanics) has been widely discussed during the last decades. Specifically, a number of cytokines, receptors, and proteins have been reported to have an important role in bone remodeling. On the other hand, it has been evidenced that mechanical factors such as the interstitial fluid flow, dynamic loading, or the presence of microcracks influence bone remodeling. Both mechanics and genetics are coupled effects, such that mechanics or genetics themselves cannot explain the phenomenon of bone remodeling [7]. In fact, it may be assumed that mechanical forces acting at the tissue level induce a cascade of molecular mechanisms activated at a lower observation (protein) scale. Then, molecular mechanisms provide a cue transcripted at the cell level which develops a specific function. This hypothesis is termed *mechanotransduction*.

The effect of genetics in contrast to mechanics was highlighted in an animal experiment presented in Ref. [8]. The experiment analyzed bone growth rather than bone remodeling and proceeded as follows. The cartilaginous primordia of the femur (the cartilage anlage where femur develops from) of a developing mouse was isolated to analyze the effect of the mechanical loading during growth. It was carried out transplanting this tissue into the mouse's spleen site, where mechanical loads are absent, during growth. As a result, the transplanted femur in the spleen, where no mechanical loading was developed, was compared after growth to a normal physiological one, which was subjected to a normal mechanical environment. Figure 2 shows the appearance of both femurs with and without a normal loading environment during growth. It can be observed the lack of bone tissue of the unloaded femur in terms of lower bone mass distribution and organization, thinner cortical wall, undeveloped femoral



Bone Remodeling, Fig. 2 Two mouse femurs showing the isolated effect of genetics (*right*) and the combined effect of genetics and mechanics (*left*) during growth (From Chalmers and Ray [8])

neck, and curvature of the femoral diaphysis [9]. Even though bone growth during development is a different process than bone remodeling, this experiment illustrates the importance of both genetics and mechanics, and genetics itself, in bone tissue function and form.

In the following sections, some key research findings related to the cellular and molecular mechanisms as well as the mechanical factors involved in bone remodeling are discussed.

Cellular and Molecular Mechanisms Involved in Bone Remodeling

Bone remodeling is the coordinated action of BMUs, as previously introduced. First, a signal, probably coming from a mechanical cue or hormone, activates bone cells in the so-called activation phase. It is thought that osteocytes are the cells involved in the transduction of the above-mentioned mechanical stimulus into a biological signal. In this phase, PTH, a calciotropic hormone secreted by the parathyroid glands, binds to the seven-transmembrane G-protein-coupled receptor (called the PTH receptor) of osteoblastic cells. It activates protein kinase A, protein kinase C, and calcium intracellular signaling pathways of osteoblastic cells, which induces a wave of transcriptional response that produces a secretion of molecules that recruit osteoclast precursors, induce osteoclast differentiation and activation, and establish *bone resorption* [10]. Then, in response to PTH, osteoblasts produce the chemokine MCP-1 (monocyte chemoattractant protein-1), which is a chemoattractant for osteoblasts precursors [10]. In addition, during the resorption phase, cytokines CSF-1, RANKL, and OPG are modulated in response to PTH to activate and maintain osteoclasts activity. Specifically, these cytokines promote osteoclasts proliferation, spreading, motility, cytoskeletal organization, and coordinate differentiation from osteoclasts precursors to multinucleated osteoclasts. Then, bone organic matrix is degraded and dissolved by matrix metalloproteinases and a collection of collagenolytic enzymes with a low pH [10]. At the *reversal phase*, a cell from the osteoblast lineage removes collagen remnants from the bone surface and prepares it for subsequent osteoblast-mediated bone formation. In the *formation phase*, mechanical strain on bone and PTH signaling, via PTH receptors on osteocytes, activate a signaling pathway which determines the early stage of bone deposition. Moreover, mesenchymal stem cells or early osteoblast progenitors differentiate and secrete molecules that ultimately form replacement bone [10]. The molecules include collagen I proteins and non-collagenous ones, such as proteoclycans, glycosylated proteins, and lipids. During the *mineralization phase*, once an equal quantity of resorbed bone has been formed, an activation signal informs the remodeling cycle to cease.

Many of the above summarized processes which take place during different phases of bone remodeling are under research and some of them are still poorly understood. However, it has been briefly evidenced that many times a mechanical

signal is involved in the related molecular and cellular mechanisms present in bone remodeling. In the next section, a discussion of the effects of mechanical signals on bone remodeling is introduced from a cell/tissue perspective.

Mechanics in Bone Remodeling

The important role that different mechanical signals have in the process of bone remodeling has been evidenced in recent years. It is thought that mechanical forces, transferred by muscle and bones, interact at the protein level allowing a transcription into biochemical signals which influences cell function. Some of the mechanisms of the transcription of a mechanical cue into a biochemical signal involved at the molecular level in bone remodeling have been exposed in the previous section. This process, termed mechanotransduction, remains to be elucidated.

A phenomenological consideration of the effect of mechanics in bone remodeling aims to establish a cause-effect statement of the problem. In this sense, "cause" would be a certain mechanical factor and "effect" would be its specific consequence on the continuous formation/resorption of bone, i.e., bone remodeling, observable at the macroscopic level of the bone architecture. In bone remodeling, osteocytes have been traditionally considered as a candidate for bone mechanosensing [7]. Cowin and coworkers have reported that osteocytes sense a range of strain levels exerted at different frequencies at the bone tissue, by means of the fluid flow pressure in the lacunar-canalicular porosity system [11]. Specifically, it has been suggested that strain, shear stress, or fluid flow give rise to changes in the ion channel conductance in the osteocytic membrane, which releases intracellular Ca^{2+} that regulates the opening and closing of membrane ion channels in the communicating junctions. Therefore, the gap junctions modulate the intracellular potential and current that passes through the network of interconnected osteocytes and surface osteoblasts which induce their function [11].

As discussed above, both mechanical load (strain), frequency and fluid flow within the bone pores are essential stimulus for bone adaptation. On one hand, it has been shown that application of load frequency stimulates bone formation when load is applied over 0.5 Hz. Specifically, Hsieh and Turner [12] reported more bone formation in a rat forelimb when application load was 10 cycle/s versus 1 cycle/s. Moreover, small loads were applied to the sheep hindlimbs at 30 cycle/s [13] showing encouraging results. As a conclusion, experimental results suggest that bone formation is proportional to the product of frequency times strain amplitude [11, 14]. It has been proved that a certain strain level at a given frequency may lay down within the physiological regime which stimulates osteocytes' mechanosensory system [11]. On the other hand, interstitial fluid flow circulates through the bone pores, canaliculi, and canals associated to the different porosity levels of bone. Interstitial fluid flow has been postulated as another important stimulus for bone formation due to the transport of nutrients to individual cells buried within the matrix. Moreover, the indirect shear stress and hydrostatic pressure produced by the fluid circulation may induce the necessary stimulus to activate the process of bone remodeling. Different sources to induce fluid movement are gravity forces, capillary forces (i.e., vasculature), hydrostatic pressure-induced flow of blood through vasculature, or forces exerted by muscles and bones and transmitted to the bone fluid due to the poroelastic architecture of the bone tissue.

A consequence of the effect of mechanics on bone formation and remodeling is the loss of bone mass in the absence of activity or bone disuse. It has been widely reported that astronauts, under the effect of microgravity and hence low mechanical stimulation of bones, undergo a rate of bone mass loss of the skeleton approximately of 1 % per month. In this context, LeBlanc et al. [15] reported the measurements made from 1990 to 1995 on 18 cosmonauts involving 12 space missions measured before and after missions lasting from 4 to 14.4 months. Eighteen crew members received a lumbar spine and hip scan before and after flight; 17 crew members received an additional whole body scan. Results, shown in Table 1, were expressed as percent change from baseline per month of flight in order to account for the different flight times.

Bone Remodeling, Table 1 Bone mineral density and body composition changes ($p < 0.01$) significantly different from baseline, after 4–14.4 months of space flight

Variable	N	%/Month change	SD
Lumbar spine	18	−1.06	0.63
Femoral neck	18	−1.15	0.84
Trochanter	18	−1.56	0.99
Total body	17	−0.35	0.25
Pelvis	17	−1.35	0.54
Arm	17	−0.04	0.88
Leg	16	−0.34	0.33

Adapted from [15]
N number of samples, *SD* standard deviation

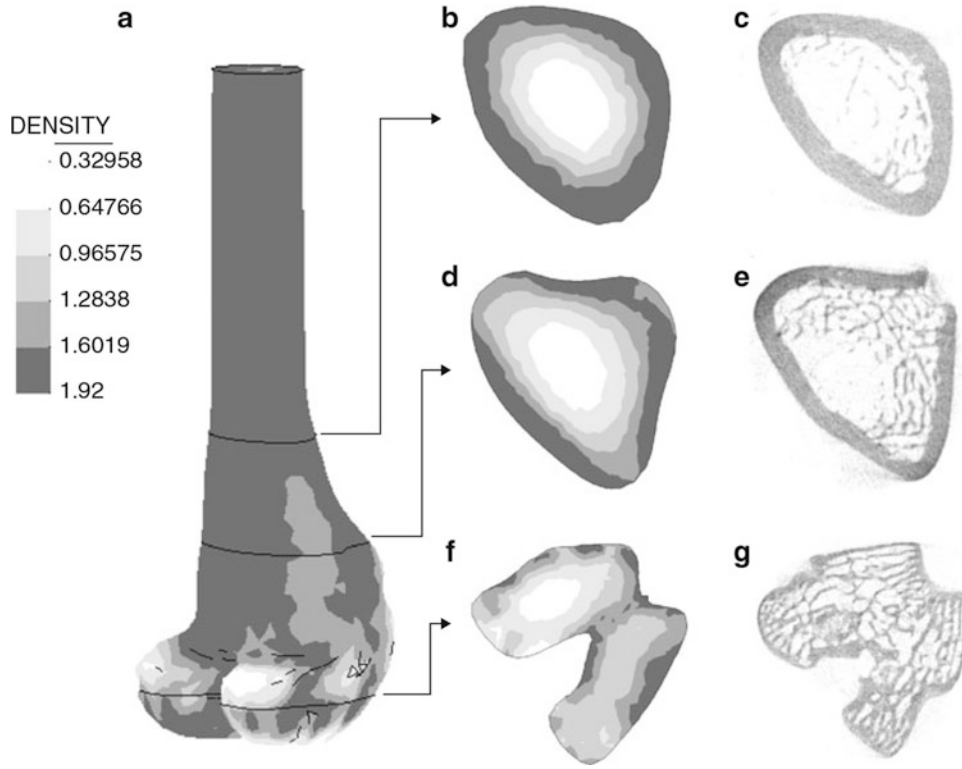
Moreover, another evidence of the importance of the mechanical forces in bone remodeling is the fact that the skull may actually gain bone due to the increased fluid pressure owing to the caudal shift of fluids in microgravity [7].

Another interesting mechanical cue for bone remodeling is the presence of microcracks along the bone tissue. Microcracks appear as a consequence of accumulated microdamage due to repetitive loading. Microdamage is even produced at physiological levels of strain and has been reported to be a mechanical signal to initiate the process of tissue repair during bone remodeling [16]. On the other hand, under severe or extreme amplitude of the loading, for example athletes during competition, microcracks may eventually produce failure or breaking of the bone organ.

Future Directions

Understanding bone remodeling as well as the implicated mechanisms in such a process may provide insight in bone function and related bone diseases. Specifically, once the mechanical and molecular mechanisms involved are elucidated, one may control bone formation or resorption by the design of drugs, biomaterials or actuators which interact with the bone tissue. To this end, how forces are transferred from muscles and skeleton to the protein level, and how these forces induce a cascade of biochemical signals, i.e., mechanotransduction, is a keystone which remains to be investigated.

A proper knowledge of bone remodeling mechanisms may add some light for the explanation of the process, and hopefully may provide a clinical solution for diseases involving abnormalities in bone remodeling such as osteoporosis and osteosclerosis. In one hand, osteoporosis, the most common bone disease, is a remodeling disease in which patients have a low bone mass condition with a high risk of fracture. This disease reflects an imbalance in the skeletal turnover so that bone resorption exceeds bone formation. On the other hand, patients with osteosclerosis have an increased bone mass due to stimulated osteoblast activity. These bone diseases and their consequences are major public health problems. Their study may benefit advances in the field of bone remodeling such as bone healing, distraction osteogenesis, or osteointegration processes around implants, among many other orthopedic protocols and techniques. Mathematical models have been developed for the analysis of bone remodeling. Currently, this is an intense area of research carried out by engineers, mathematicians, and physicists. The main objective of mathematical modeling is to predict information for a number of applications with a clinical or research interest to reduce animal experimentation, reduce cost and time, improve protocols, and aid the clinician to take decisions related to diseases in bone remodeling. Specially, many models have focused on the analysis of mechanics in bone remodeling from a continuum approach. The input of these models is normally a so-called mechanical stimulus, namely, fluid flow velocity, shear stress, pressure, or a certain combination of the components of the stress or strain tensors. Normally, models assume a higher amount of deposited bone in places where the mechanical stimulus is higher. Continuum models may be classified into phenomenological and mechanistic ones. In one hand, phenomenological models aim at predicting how macroscopic forces induce bone function and form during bone remodeling [17–21]. Many phenomenological parameters have to be calibrated in this kind of models whereas results should be analyzed in qualitative terms. Figure 3 shows a numerical simulation of bone remodeling and how it compares to actual



Bone Remodeling, Fig. 3 Simulation in silico of bone remodeling of a white New Zealand rabbit femur: macroscopic bone density [g/cc] distribution. (**a**, **b**, **d**, **f**) Simulated until convergence. (**c**, **e**, **g**) Actual tomographies for different sections (From [22])

clinical results [22]. Mechanistic models establish a link between mechanical forces and the implicated biological functions to explain certain features of bone remodeling [23–25]. Mechanistic approaches are usually more complex since they analyze the coupled effect of mechanics and biology in bone remodeling. In the recent years, the term in silico approach has been coined to classify mathematical models into the different scientific methodologies present in the clinical practice such as bone remodeling.

Recently, the use of biomaterials to enhance or improve the function of a damaged tissue, or to promote tissue growth, has been proposed. This new discipline is named *Tissue Engineering* and offers many applications in the case of bone tissue. Specifically, new biomaterials (ceramic-based, bioactive glasses, polymers) are used as a scaffold which interacts with the bone tissue, with the purpose that cells colonize the scaffold biomaterial, attach, and develop their specific function finally leading to the growth of new bone tissue. Several requirements must be fulfilled by the scaffold biomaterial to develop its proper function. First, it must be a biocompatible biomaterial which interacts with the bone tissue. Moreover, it must be strong enough to support physiological loads at its specific location allowing cell attachment and fluid circulation in the interior. Mechanics is also an important factor in bone tissue engineering, as the signaling pathways are similar to the bone tissue but here the extracellular matrix is the biomaterial itself. The gold standard in bone tissue engineering is to promote the growth of functional bone tissue in circumstances where nature has lost its ability to develop new tissue under a disease or age-related complications.

References

1. Frost, H.M.: Tetracycline-based histological analysis of bone remodeling. Calcif. Tissue Res. **3**, 211–237 (1969)
2. Parfitt, A.M.: The actions of parathyroid hormone on bone: relation to bone remodeling and turnover, calcium homeostasis, and metabolic bone disease. Part I of IV parts: mechanisms of calcium transfer between blood and bone and their cellular basis: morphological and kinetic approaches to bone turnover. Metabol. Clin. Exp. **25**, 809–844 (1976)
3. Martin, R.B., Burr, D.B., Sharkey, N.A.: Skeletal Tissue Mechanics. Springer, New York (1998)
4. Parfitt, A.M.: The physiologic and clinical significance of bone histomorphometric data. In: Recker, R.R. (ed.) Bone Histomorphometry: Techniques and Interpretations. CRC Press, Boca Raton (1983)
5. Buenzli, P.R., Pivonka, P., Smith, D.W.: Spatio-temporal structure of cell distribution in cortical bone multicellular units: a mathematical model. Bone **48**, 918–926 (2011)
6. Eriksen, E.F., Gundersen, H.J., Melsen, F., Mosekilde, L.: Reconstruction of the formative site in iliac trabecular bone in 20 normal individuals employing a kinetic model for matrix and mineral apposition. Metab. Bone Dis. Relat. Res. **5**, 243–252 (1984)
7. Cowin, S.C.: Tissue growth and remodeling. Annu. Rev. Biomed. Eng. **6**, 77–107 (2004)
8. Chalmers, J., Ray, R.D.: The growth of transplanted foetal bones in different immunological environments. J. Bone Joint Surg. **44B**, 149–164 (1962)
9. Goodship, A.E., Cunningham, J.L.: Pathophysiology of functional adaptation of bone in remodeling and repair in-vivo. In: Cowin, S.C. (ed.) Bone Mechanics Handbook, pp. 26-1–26-31. CRC Press, Boca Raton (2001)
10. Raggatt, L.J., Partridge, N.C.: Cellular and molecular mechanisms of bone remodeling. J. Biol. Chem. **286**, 25103–25108 (2010)
11. Weinbaum, S., Cowin, S.C., Zeng, Y.: A model for the excitation of osteocytes by mechanical loading-induced bone fluid shear stresses. J. Biomech. **27**, 339–360 (1994)
12. Hsieh, Y.F., Turner, C.H.: Effects of loading frequency on mechanically induced bone formation. J. Bone Miner. Res. **16**, 918–924 (2001)
13. Rubin, C., Turner, A.S., Bain, S., Mallinckrodt, C., McLeod, K.: Anabolism. Low mechanical signals strengthen long bones. Nature **412**, 603–604 (2001)
14. Robling, A.G., Castillo, A.B.., Turner, C.H.: Biomechanical and molecular regulation of bone remodeling. Annu. Rev. Biomed. Eng. **8**, 455–498 (2006)
15. LeBlanc, A., Schneider, V., Shackelford, L., West, S., Oganov, V., Bakulin, A., Voronin, L.: Bone mineral and lean tissue loss after long duration space flight. J. Musculoskelet. Neuronal Interact. **1**, 157–160 (2000)
16. Bentolila, V., Boyce, T.M., Fyhrie, D.P., Drumb, R., Skerry, T.M., Schaffler, M.B.: Intracortical remodeling in adult rat long bones after fatigue loading. Bone **23**, 271–281 (1998)
17. Pauwels, F.: Eine neue theorie über den einflub mechanischer reize auf die differenzierung der stützgewebe. Z. Anat. Entwicklungsgesch. **121**, 478–515 (1960)
18. Cowin, S.C., Hegedus, D.H.: Bone remodeling I: a theory of adaptive elasticity. J. Elast. **6**, 313–326 (1976)
19. Huiskes, R., Weinans, H., Grootenboer, H.J., Dalstra, M., Fudala, B., Sloof, T.J.: Adaptive bone-remodeling theory applied to prosthetic-design analysis. J. Biomech. **20**, 1135–1150 (1987)
20. Beaupré, G.S., Orr, T.E., Carter, D.R.: An approach for time-dependent bone modeling and remodeling: theoretical development. J. Orthop. Res. **8**, 651–661 (1990)
21. Doblaré, M., García, J.M.: Application of an anisotropic bone-remodeling model based on a damage-repair theory to the analysis of the proximal femur before and after total hip replacement. J. Biomech. **34**, 1157–1170 (2001)
22. Sanz-Herrera, J.A., García-Aznar, J.M., Doblaré, M.: Micro–macro numerical modeling of bone regeneration in tissue engineering. Comput. Methods Appl. Mech. Eng. **197**, 3092–3107 (2008)
23. Hernandez, C.J., Beaupré, G.S., Carter, D.R.: A model of mechanobiologic and metabolic influences on bone adaptation. J. Rehabil. Res. Dev. **37**, 235–244 (2000)
24. Huiskes, R., Ruimerman, R., van Lenthe, G.H., Janssen, J.D.: Effects of mechanical forces on maintenance and adaptation of form in trabecular bone. Nature **405**, 704–706 (2000)
25. Hazelwood, S.J., Martin, R.B., Rashid, M.M., Rodrigo, J.J.: A mechanistic model for internal bone remodeling exhibits different dynamic responses in disuse and overload. J. Biomech. **34**, 299–308 (2001)

Bone Turnover

▶ Bone Remodeling

Boron- and/or Nitrogen-Doped Carbon Nanotubes

▶ Light-Element Nanotubes and Related Structures

Boron Nitride Nanotubes (BNNTs)

▶ Chemical Vapor Deposition (CVD)
▶ Physical Vapor Deposition

Bottom-Up Fabrication of Plasmonic Materials

▶ Plasmonics in Self-Organized Media

Bottom-Up Nanofabrication

▶ Electric Field–Directed Assembly of Bioderivatized Nanoparticles

Bottom-Up Processing

▶ Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly

Boundary Lubrication

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Greases; Perfluoropolyethers; Petroleum lubricants; Synthetic lubricants

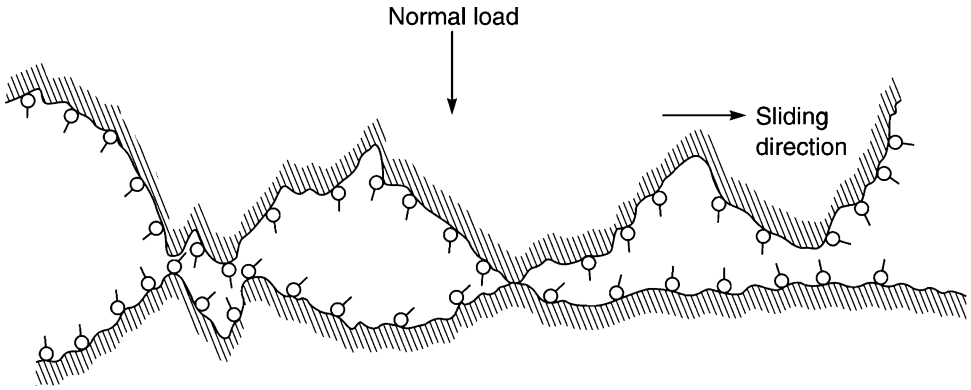
Definition

Lubricants are commonly used for reduction in friction and wear of interfaces. In some applications, the solid surfaces are so close together that some asperities come in contact and others are mitigated by a thin film of lubricant, Fig. 1. Under these conditions, the lubricant viscosity is relatively unimportant and the physical and chemical interactions of the lubricant with the solid bodies controls friction and wear. Even a monolayer of adsorbed molecules may provide some protection against wear. Lubrication in some situations can be achieved by the use of multimolecular lubricant films. Monolayer lubrication is referred to as boundary lubrication and multimolecular lubrication is referred to as mixed lubrication. Boundary lubrication usually occurs under high-load and low-speed conditions in machine components such as bearings, gears, cam and tappet interfaces, and piston ring and liner interfaces. Boundary lubrication forms a last line of defense. In many cases, it is the regime which controls the component life [1–12].

Various lubricants and greases are used for lubrication of machine components operating in various lubrication regimes [3–7, 12–21]. Additives are commonly used to provide the desirable properties and interaction with the interface.

Overview

For the case of two contacting bodies coated with a continuous solid monolayer of lubricant with a load too small to cause plastic deformation, the interface is in equilibrium under load for some time since the films prevent contact between the substrates. The films do not allow any state of lower free surface energy than the initial state. The films will thus lubricate over a considerable sliding distance if the bodies are subjected to low-speed sliding, although they will eventually be worn away. Alternatively, if the temperature is raised slightly above the melting point of the films, the admolecules will acquire some mobility and no state involving more than one complete monolayer can be stable. Activation energy for migrating away from the loaded region is provided thermally – and mechanically also, if a low-speed sliding is imposed. The two layers initially present thus penetrate each other and



Boundary Lubrication, Fig. 1 Schematic of two surfaces separated by a boundary layer of lubricant

adsorbed molecules tend to move away from the loaded interface. An equilibrium between a partial monolayer at the loaded interface and a surrounding vapor, if it exists, is generally metastable since the state for which the bodies are in direct contact usually has lower energy than any lubricated one.

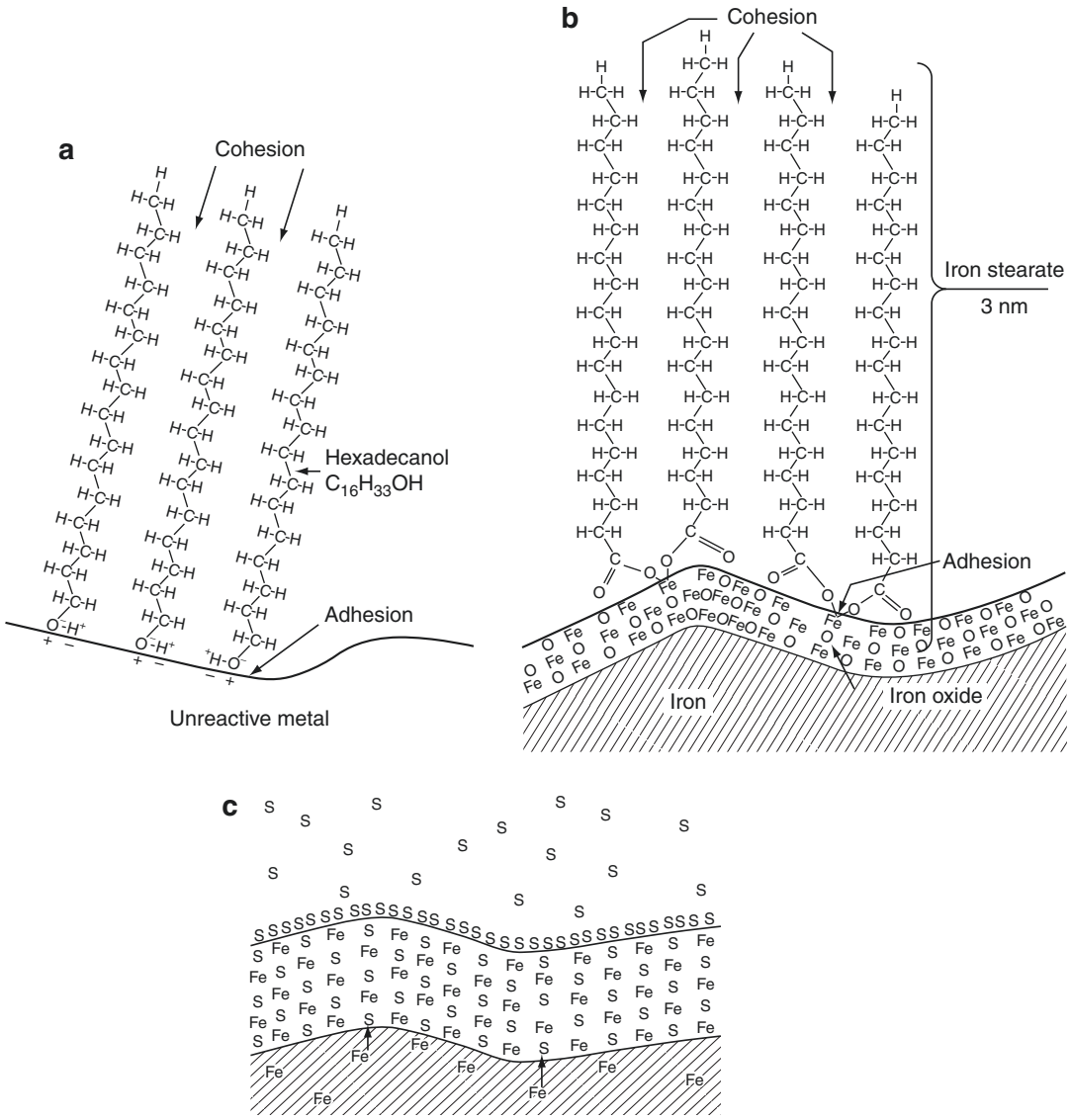
The boundary films are formed by physical adsorption, chemical adsorption, and chemical reaction; for typical examples, see Fig. 2. The physisorbed film can be of either monomolecular (typically <3 nm) or polymolecular thickness. The chemisorbed films are monomolecular, but stoichiometric films formed by chemical reaction can have a large film thickness. In general, the stability and durability of surface films decrease in the following order: chemical reaction films, chemisorbed films, and physisorbed films.

A good boundary lubricant should have a high degree of interaction between its molecules and the sliding surface. As a general rule, liquids are good lubricants when they are polar and thus able to grip solid surfaces (or be adsorbed) [3–6, 14]. Polar lubricants contain reactive functional groups with low ionization potential or groups having high polarizability. Boundary-lubrication properties of lubricants are also dependent upon the molecular conformation and lubricant spreading. Examples of nonpolar and polar molecules are shown in Fig. 3. In the case of Z-DOL, a hydrogen atom, covalently bonded with oxygen atom in the O–H bond, exposes a bare proton on the end of the bond. This proton can be easily attracted to the negative charge of other molecules because the proton is not shielded by electrons, and this is responsible for the polarity of the O–H ends. Likewise, the lone pairs of electrons in the oxygen and fluorine atoms in both molecules are unshielded, and can be attracted to positive charges of other molecules, and thus exhibit electronegativity which is responsible for some polarity. The CF_3 end in Z-15 is symmetric, and its polarity is low.

In addition to the polarity of liquids, the shape of their molecules governs the effectiveness, which determines whether they can form a dense, thick layer on the solid surface. Ring molecules or branch chain molecules tend to be poorer than straight chain molecules because there is no way in which they can achieve a high packing density. Straight chain molecules with one polar end, such as alcohols and soaps of fatty acids, are highly desirable because they enable a thick film to be formed with the polar end tightly held on the surface and the rest of the molecule normal to the surface. If the sliding surface has to operate under humid conditions, the lubricant should be hydrophobic (i.e., it should not absorb water or be displaced by water).

The most readily observed cause of breakdown of thin-film lubrication is the melting of a solid film and degradation of liquid films, but some degree of lubrication may persist to a higher temperature. Sliding speed and load influence the performance of multilayers.

The properties of the solid surface that are desirable for good lubrication are discussed.



Boundary Lubrication, Fig. 2 (**a**) Schematic diagram representing the physisorption with preferred orientation of three polar molecules of hexadecanol to a metal surface; (**b**) schematic diagram representing the chemisorption of stearic acid on an iron surface to form a monolayer of iron stearate, a soap; (**c**) schematic representation of an inorganic film formed by chemical reaction of sulfur with iron to form iron sulfide [10]

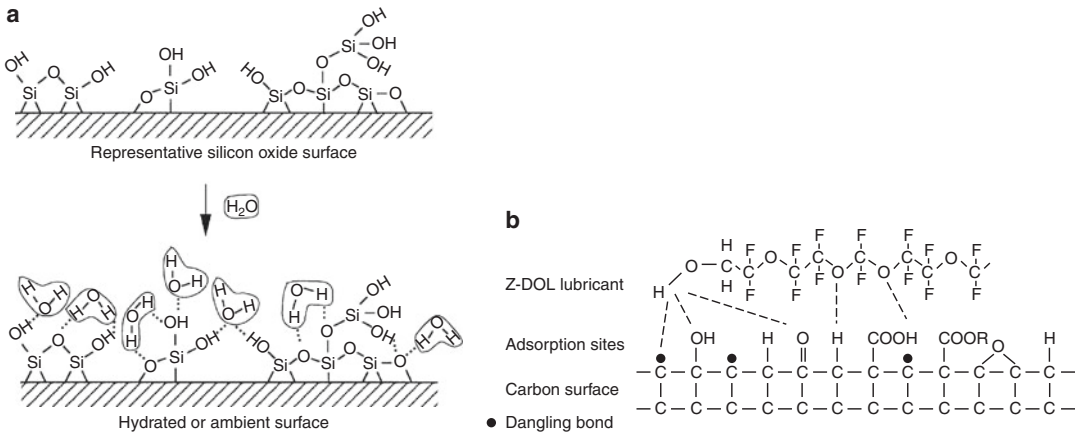
Nonpolar molecule

Fomblin Z-15 $CF_3-O-(CF_2-CF_2-O)_m-(CF_2-O)_n-CF_3$ ($m/n \sim 2/3$)

Polar molecule

Fomblin Z-DOL $HO-CH_2-CF_2-O-(CF_2-CF_2-O)_m-(CF_2-O)_n-CF_2-CH_2-OH$

Boundary Lubrication, Fig. 3 Structures of nonpolar and polar (−OH) organic lubricant molecules



Boundary Lubrication, Fig. 4 Schematic illustrations of (**a**) a hydrophilic silicon-oxide surface before and after adsorption of water molecules; hydrogen bonding occurs between the solid surface and water molecules. (**b**) A hydrogenated diamond-like carbon surface with adsorbed polar perfluoropolyether (Z-DOL) lubricant molecules; • represents dangling bonds

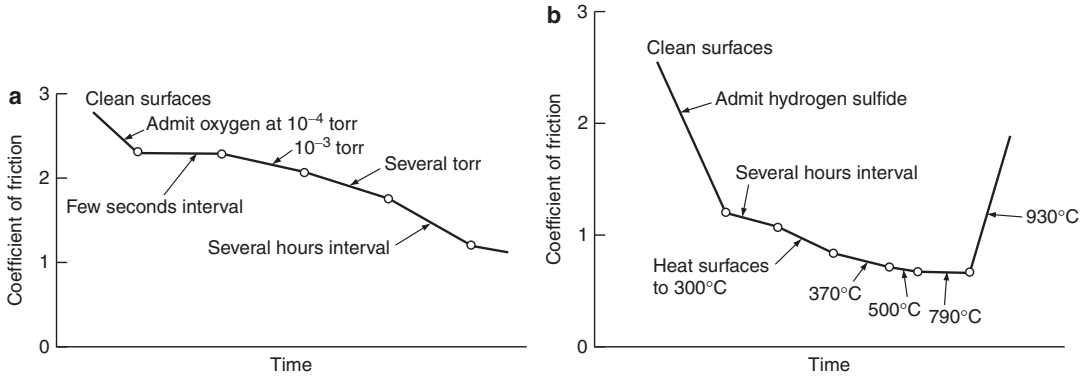
A solid should have a high surface energy so that there will be a strong tendency for molecules to adsorb on the surface. Consequently, metals tend to be the easiest surfaces to be lubricated. The solid surface should have a high wetting (or low contact angle) so that the liquid lubricant wets the solid easily. For better lubrication, the surface should be reactive to the lubricant under test conditions so that durable, chemically reacted films can form. Another property of solid surfaces is hydrophobicity. The surfaces should be highly functional with polar groups and dangling bonds (unpaired electrons) so that they can react with lubricant molecules and adsorb them. Examples of hydrophilic silicon oxide surface and its reactivity to ambient water and amorphous carbon surface with polar groups and dangling bonds which promote adsorption of perfluoropolyether molecules are shown in Fig. 4 [14]. Additives to the lubricants can also enhance the formation of chemically reacted films.

Data showing the effect of environment and types of lubricants and their interaction with solid surfaces on boundary lubrication behavior are now presented.

Effect of Adsorbed Gases

Boundary films occur on almost all surfaces because they reduce the surface energy and are thus thermodynamically favored. Normally, air covers any surface with an oxidized film plus adsorbed moisture and organic material. Inadvertent lubrication by air is the most common boundary lubrication. Figure 5a shows the reduction in the coefficient of friction that is obtained by the adsorption and/or chemical reaction of oxygen on clean iron surfaces outgassed in a vacuum (roughly 10^{-6} torr or mmHg). The coefficient of friction is markedly reduced by admission of oxygen gas, though the oxygen pressure is very low (roughly 10^{-4} torr). As oxygen pressure is allowed to increase, the friction is reduced still more. Finally, if the surfaces are allowed to stand for some period of time, the adsorbed oxygen film becomes more complete and the friction drops still further. Note that the seizure of clean metals is prevented by even a trace of oxygen, as obtained at 10^{-4} torr [22].

Figure 5b shows the effect of the addition of hydrogen sulfide on the coefficient of friction of outgassed (clean) iron surfaces; the friction reduced abruptly and appreciably. It is necessary to heat the surface over 790 °C before the decomposition of the film takes place and friction rises. It is probable that hydrogen sulfide reacts with the clean iron surfaces to form an iron sulfide (FeS) film.

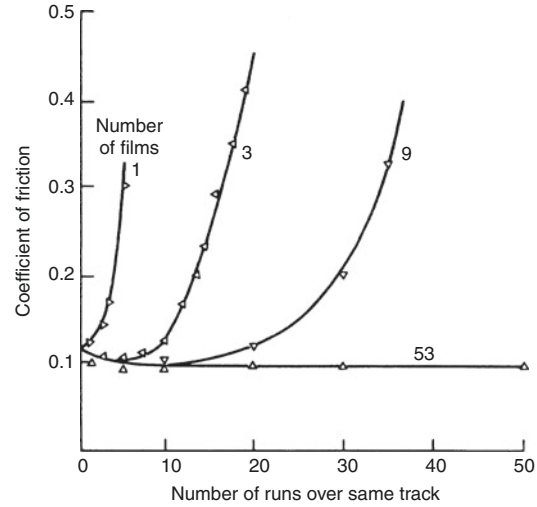


Boundary Lubrication, Fig. 5 (**a**) Effect of oxygen on the coefficient of friction of outgassed iron surfaces and (**b**) effect of hydrogen sulfide on the coefficient of friction of outgassed iron surfaces [22]

Effect of Monolayers and Multilayers

It is possible to show by use of monomolecular layers and multimolecular layers that a very thin film of lubricant at the surface can be effective in reducing friction. In studying the effects of monolayers or multilayers, it is convenient to use the well-known Langmuir–Blodgett (L–B) technique. This technique involves floating an insoluble monolayer on the surface of water and then transferring it from the surface of the water to the surface of the solid (by successive dippings) to which the monolayer or multilayer is to be applied. This technique is convenient for deposition of films of known and controllable thicknesses.

Bowden and Tabor [23] deposited the films of a long-chain fatty acid (stearic acid) on a stainless-steel surface. The lubricated surface was slid against an unlubricated surface at 10 mm/s, and the coefficient of friction was recorded from the beginning of sliding. Data shown in Fig. 6 for a monolayer and with multilayers of 3, 9, or 53 films show that the greater the number of films, the longer it takes to wear off or displace this protective film and, consequently, the longer the time in which the film is an effective boundary lubricant. The films deposited by the L–B technique are not entirely equivalent to the type of protective film developed from lubricants in practice with respect to either molecular packing or composition. The stearic acid films, however, were close packed and regularly oriented with the polar group in the water surface.



Boundary Lubrication, Fig. 6 Wear behavior of a number of stearic acid films deposited on stainless-steel sliding against unlubricated stainless-steel surface [23]

Israelachvili et al. [24] have elegantly measured the frictional force or shear stress (frictional force divided by the apparent area of contact) required to sustain sliding (shearing) of two molecularly smooth mica surfaces with various molecular layers(s) of a liquid film in between. The two liquids used were octamethylcyclotetrasiloxane (OMCTS) and cyclohexane, and have mean molecular diameters of 0.85 and 0.5 nm, respectively. Measurements were made at sliding velocities ranging from 0.25 to 2 μm/s after steady-state sliding was attained. They found that the shear stress depended on the

number of boundary liquid layers, Table 1; in cyclohexane, for example, shear stress fell by about an order of magnitude per additional layer. In other words, the friction is quantized depending on the number of molecular layers separating the surfaces. By extrapolation, one may infer that when seven to ten layers are present, the shear stress of the liquid film would have fallen to the value expected for bulk continuum Newtonian flow. It is noteworthy that this is about the same number of layers as when the forces across a thin film and the whole concept of viscosity begin to be described by continuum theories [25].

Boundary Lubrication, Table 1 Shear stress as a function of number of boundary layers trapped between two mica surfaces for octamethylcyclotetrasiloxane (OMCTS)[a] and cyclohexane[b]

	Shear stress (MPa)	
Number of layers	OMCTS	Cyclohexane
1	8.0 ± 0.5	$2.3 \pm 0.6 \times 10$
2	6.0 ± 1.0	1.0 ± 0.2
3	3.0 ± 1.0	$4.3 \pm 1.5 \times 10^{-1}$
4	Not measured	$2.0 \pm 1.0 \times 10^{-2}$

[a]Molecular diameter ~ 0.85 nm
[b]Molecular diameter ~ 0.5 nm

Effect of Chemical Films

The addition of a small trace of a fatty acid (polar lubricant) to a nonpolar mineral oil or to a pure hydrocarbon can bring about a considerable reduction in the friction and wear of chemically reactive surfaces. Typical results taken from [23] are given in Table 2. In these experiments, friction was measured using identical materials sliding against each other and lubricated with a (nonpolar) paraffin oil or with 1 % (polar) lauric acid added to the paraffinic oil. They found that with unreactive metals (such as Pt, Ag, Ni, and Cr) and glass-the fatty acid is no more effective than a paraffin oil. In contrast, the results for lubrication of reactive metals (such as Zn, Cd, Cu, Mg, and Fe) show that very effective lubrication can be obtained with a 1 % solution of lauric acid in paraffin oil. Bowden and Tabor [23] have further shown that a 0.01 % solution of lauric acid in paraffin oil reduces the coefficient of friction of chemically reactive cadium surfaces from 0.45 to 0.10. Even smaller concentrations (0.001 %) of lauric acid can reduce friction slowly with time (after a few hours). In the case of the less-reactive metals, such as iron, which cannot be lubricated by a 1 % solution of fatty acid, they are well lubricated by a more concentrated solution. These results indicate the strong effects of films formed by chemical reaction on friction and wear.

Boundary Lubrication, Table 2 Efficiency of lubrication by paraffin oils or 1 % lauric acid in paraffin oil compared with reactivity of metal to lauric acid [23]

	Coefficient of friction at 20 °C			Transition temperature (°C)	%Acid reactive[a]
Metal	Clean	Paraffin oil	1 % Lauric acid in paraffin oil		
Unreactive					
Platinum	1.2	0.28	0.25	20	None
Silver	1.4	0.8	0.7	20	None
Nickel	0.7	0.3	0.28	20	None
Chromium	0.4	0.3	0.3	20	None
Glass	0.9	0.4	0.4	20	None
Reactive					
Zinc	0.6	0.2	0.04	94	10
Cadmium	0.5	0.45	0.05	103	9.3
Copper	1.4	0.3	0.08	97	4.6
Magnesium	0.6	0.5	0.08	80	Trace
Iron	1.0	0.3	0.2		Trace

[a]Estimated amount of acid involved in the reaction assuming formation of a normal salt

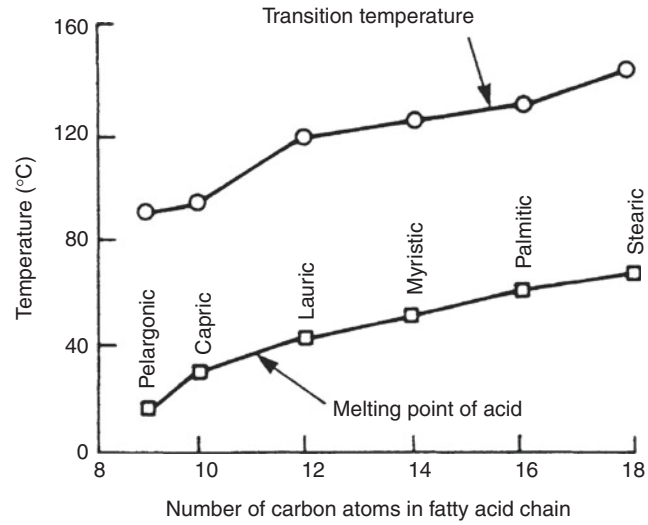
Friction and wear measurements on monolayers of various substances show that whereas a single monolayer of a given polar compound on a reactive surface provides low friction and wear, on unreactive surfaces it may be completely ineffective, and as many as ten or more layers may be needed. Bowden and Tabor [23] deposited multilayers of stearic acid and metal soaps (believed to be responsible for good lubrication behavior when stearic acid is used as a lubricant on a metal surface) on unreactive and reactive metal surfaces and made the friction measurements. Table 3 shows the number of layers required for effective lubrication. It is found that a larger number of films are needed for unreactive surfaces (see also Ling et al. [11]).

Boundary Lubrication, Table 3 Lubrication of various metal surfaces by layers of stearic acid and metal stearates deposited by the Langmuir–Blodgett technique [23]

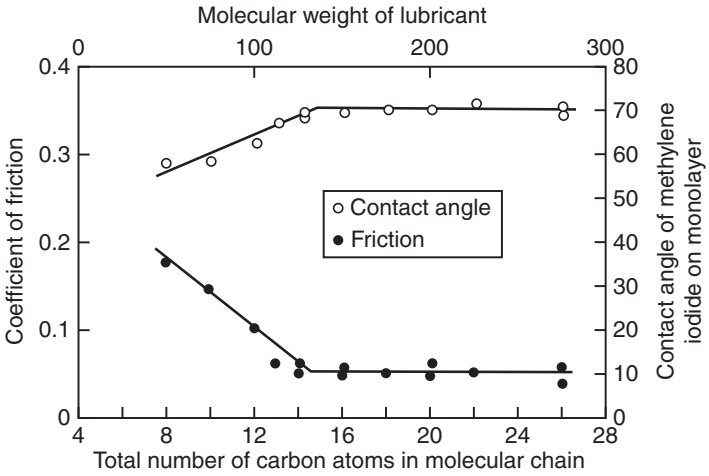
	Number of layers for effective lubrication	
Metal	Stearic acid	Metal soap (Cu or Ag stearates)
Unreactive		
Platinum	>10	7–9
Silver	7	3
Nickel	3	3
Reactive		
Copper	3	3
Stainless steel	3	1

Parraffins, alcohols, ketones, and amides become ineffective lubricants at the bulk melting point of the lubricant. When the melting occurs, the adhesion between the molecules in the boundary film is diminished and breakdown of the film takes place. The increased metallic contact through the lubricant film leads to increased friction and wear. With saturated fatty acids on reactive metals, however, the breakdown does not occur at their melting points but at considerably higher temperatures. This is shown in Fig. 7 for a series of fatty acids on steel surfaces, and it is seen that breakdown (transition temperature) occurs at 50–70 °C above the melting point. The transition temperature corresponds approximately to the melting point of the metallic soaps formed by chemical reaction [23]. The actual value of the breakdown temperature depends on the nature of the metals and on the load and speed of sliding. The esters of saturated fatty acids also behave like acids, except that the difference between the transition and melting temperatures (T_t–T_m) decreases with increasing ester group length, approaching zero at 26 carbon atoms.

Thus, lubrication is affected not by the fatty acid itself but by the metallic soap formed as a result of the chemical reaction between the metal and the fatty acid. Also, T_t–T_m is a measure of the



Boundary Lubrication, Fig. 7 Breakdown or transition temperature of fatty acids on steel surfaces and their melting points as a function of chain length [23]



Boundary Lubrication, Fig. 8 Effect of chain length (or molecular weight) on coefficient of friction (of stainless-steel sliding on glass lubricated with a monolayer of fatty acid) and contact angle (of methyl iodide on condensed monolayers of fatty acids on glass) [26]

strength of adsorption that is due to a dipole–metal interaction.

Effect of Chain Length (or Molecular Weight)

The effect of chain length of the carbon atoms of paraffins (nonpolar), alcohols (nonpolar), and fatty acids (polar) on the coefficient of friction was studied by Bowden and Tabor [23] and Zisman [26]. Figure 8 shows the coefficient of friction of a stainless-steel surface sliding against a glass surface lubricated with a monolayer of fatty acid. It is seen that there is a steady decrease in friction with increasing chain length. At a sufficiently long chain length, the coefficient of friction reaches a lower limit of above 0.07. Similar trends were found for other lubricant films (paraffins and alcohols). Zisman [26] also found that the contact angle of methylene iodide on a monolayer rises to a high constant value for polar, long-chain molecules of paraffins, alcohols, and fatty acids, indicating an increase in packing to an optimum condition with an increase in the number of carbon atoms in the chain. Owens [27] has confirmed that in polymer lubrication a complete surface coverage occurs with fatty acids having long chain lengths.

Zisman [26] and Owens [27] have shown that the durability of the lubricant film also increases with an increase in the film chain length. These results suggest that monolayers having a chain length below 12 carbon atoms behave as liquids (poor durability), those with chain lengths of 12–15 carbon atoms behave like a plastic solid (medium durability), whereas those with chain lengths above 15 carbon atoms behave like a crystalline solid (high durability).

Principal Classes of Lubricants

Liquid lubricants (oils) include natural organics consisting of animal fat, vegetable oils, mineral (or petroleum) fractions, synthetic organics, and mixtures of two or more of these materials. Various additives are used to improve the specific properties [3–7, 12, 15, 17–20]. A partial list of the lubricant types is shown in Table 4.

Closure

Boundary lubrication is accomplished by mono- or multimolecular films. The films are so thin that their behavior is controlled by interaction with the substrates. Molecular structure of the lubricants and functionality of the substrate affect the type and degree of bonding of the lubricant to the substrate.

Liquid lubricants include mineral (or petroleum) and synthetic organics. Various additives are used to improve specific properties. Mineral oils are excellent boundary lubricants and by far the most used lubricants. Synthetic lubricants can be used at greater extremes of

Boundary Lubrication, Table 4 Types of liquid lubricants (oils)

Natural organics	Synthetic organics
Animal fat	Synthetic hydrocarbons (polybutene)
Shark oil	Chlorinated hydrocarbons
Whale oil	Chlorofluorocarbons
Vegetable oils	Esters
Mineral (petroleum) oils	Organic acid
Paraffinic	Fatty acid
Naphthenic	Dibasic acid (di)
Aromatic	Neopentyl polyol
	Polyglycol ethers
	Fluoro
	Phosphate
	Silicate
	Disiloxane
	Silicones
	Dimethyl
	Phenyl methyl
	Chlorophenyl methyl
	Alkyl methyl
	Fluoro
	Silanes
	Polyphenyl ethers
	Perfluoropolyethers

environment, including temperature, humidity, and vapor pressure. Mineral oils are typically used up to a maximum temperature of about 130 °C and some synthetic oils up to about 370 °C. However, synthetic lubricants are more expensive than mineral oils. There are several properties of lubricants which are important for lubrication; their relative importance depends upon the industrial application. Additives are commonly used to modify friction and wear of lubricants and greases. These are classified as friction modifier, antiwear, and extreme pressure.

Greases are used where circulating liquid lubricant cannot be contained because of space and cost and where cooling by the oil is not required or the application of a liquid lubricant is not feasible.

Cross-References

- ▶ Nanotechnology
- ▶ Nanotribology
- ▶ Reliability of Nanostructures

References

1. Anon: Limits of lubrication. Special issue of Tribol. Lett. **3**, 1 (1997)
2. Beerbower, A.: Boundary Lubrication, Report No. AD-747336, U.S. Department of Commerce, Office of the Chief of Research and Development. Department of the Army, Washington, DC (1972)
3. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
4. Bhushan, B.: Modern Tribology Handbook, vol. 1 – Principles of Tribology; vol. 2 – Materials, Coatings, and Industrial Applications. CRC Press, Boca Raton (2001)
5. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
6. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
7. Booser, E.R.: CRC Handbook of Lubrication. Theory and Design, vol. 2. CRC Press, Boca Raton (1984)
8. Godfrey, D.: Boundary lubrication. In: Ku, P.M. (ed.) Interdisciplinary Approach to Friction and Wear, vol. SP-181, pp. 335–384. NASA, Washington, DC (1968)
9. Iliuc, I.: Tribology of Thin Films. Elsevier, New York (1980)
10. Ku, P.M.: Interdisciplinary Approach to the Lubrication of Concentrated Contacts, vol. SP-237. NASA, Washington, DC (1970)
11. Ling, F.F., Klaus, E.E., Fein, R.S.: Boundary Lubrication – An Appraisal of World Literature. ASME, New York (1969)
12. Loomis, W.R.: New Directions in Lubrication, Materials, Wear, and Surface Interactions – Tribology in the 80's. Noyes, Park Ridge (1985)
13. Bhushan, B.: Tribology and Mechanics of Magnetic Storage Devices, 2nd edn. Springer, New York (1996)
14. Bhushan, B., Zhao, Z.: Macro- and microscale tribological studies of molecularly-thick boundary layers of perfluoropolyether lubricants for magnetic thin-film rigid disks. J. Info. Storage Proc. Syst. **1**, 1–21 (1999)
15. Bisson, E.E., Anderson, W.J.: Advanced Bearing Technology, SP-38. NASA, Washington, DC (1964)
16. Boner, C.J.: Modern Lubricating Greases. Scientific Publications, Broseley (1976)
17. Braithwaite, E.R.: Lubrication and Lubricants. Elsevier, Amsterdam (1967)
18. Evans, G.G., Galvin, V.M., Robertson, W.S., Walker, W.F.: Lubrication in Practice. Macmillan, Basingstoke (1972)
19. Gunderson, R.C., Hart, A.W.: Synthetic Lubricants. Reinhold, New York (1962)
20. Gunther, R.C.: Lubrication. Bailey Brothers and Swinfen, Folkestone (1971)
21. McConnell, B.D.: Assessment of Lubricant Technology. ASME, New York (1972)
22. Bowden, F.P.: The influence of surface films on the friction, adhesion and surface damage of solid, the

fundamental aspects of lubrication. Ann. NY Acad. Sci. **53**, Art 4, 753–994 (1951)
23. Bowden, F.P., Tabor, D.: Friction and Lubrication of Solids, vol. I. Clarendon, Oxford (1950)
24. Israelachvili, J.N., McGuiggan, P.M., Homola, A.M.: Dynamic properties of molecularly thin liquid films. Science **240**, 189–191 (1988)
25. Israelachvili, J.N.: Intermolecular and Surface Forces, 2nd edn. Academic, San Diego (1992)
26. Zisman, W.A.: Durability and wettability properties of monomolecular films on solids. In: Davies, R. (ed.) Friction and Wear, pp. 110–148. Elsevier, Amsterdam (1959)
27. Owens, D.K.: Friction of polymers I. Lubrication. J. Appl. Poly. Sci. **8**, 1465–1475 (1964)

Brain Implants

▶ MEMS Neural Probes

Brushes

▶ Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

Buckminsterfullerene

▶ Fullerenes for Drug Delivery

Bulk Acoustic Wave MEMS Resonators

▶ Laterally Vibrating Piezoelectric Resonators

Bundles

▶ Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

C

C_{60}

► Fullerenes for Drug Delivery

Cancer

► Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites

Cancer Modeling

► Models for Tumor Growth

Capacitive MEMS Switches

Dimitrios Peroulis
School of Electrical and Computer Engineering, Birck Nanotechnology Center, Purdue University, West Lafayette, IN, USA

Synonyms

Electrostatic RF MEMS switches; Micromechanical switches; RF MEMS switches

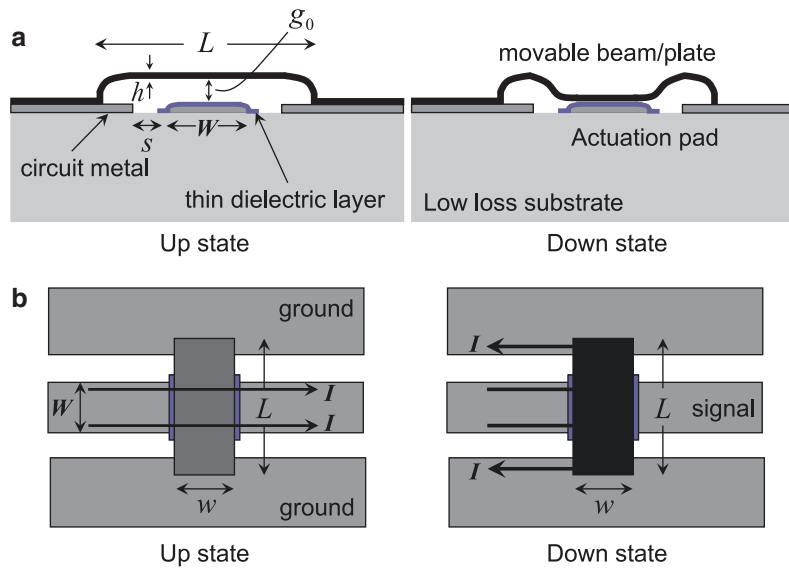
Definition

Capacitive micro-electro-mechanical systems (MEMS) switches are a special type of micromachined switches that control radio frequency (RF) signal paths in microwave and millimeter-wave circuits through mechanical motion and contact.

Overview

Capacitive and direct current (dc)-contact MEMS switches are among the most important micromachined devices for high-frequency applications due to their near-ideal RF performance. Dc-contact switches function similarly to conventional relays: micromachined beams or plates move under the influence of an appropriately applied force (e.g., electrostatic force) to open or close a metal-to-metal contact. While micromachined beams or plates are also utilized in capacitive switches, these switches rely on metal-to-dielectric contacts to implement their on and off states. Capacitive switches are particularly attractive for demanding high-frequency communications, electronic warfare, and radar systems due to their ultralow loss (<0.1–0.2 dB up to 40 GHz), high isolation (>20–50 dB for frequencies beyond 10 GHz), very high linearity (>66 dBm third-order intercept point), and near-zero power consumption (∼tens of nJ per

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1



Capacitive MEMS Switches, Fig. 1 (**a**) Side-view and (**b**) top-view schematics of a typical shunt capacitive MEMS switch. Both the *up* and *down* states are shown

switching cycle and zero quiescent power for electrostatically actuated switches). When compared to solid-state switches, capacitive switches are relatively slow devices with speeds ranging in the tens to hundreds of microseconds range. This speed is primarily limited by switch inertia and squeeze film damping. Their relatively large lateral dimensions of tens or hundreds of μm allow capacitive switches to handle several 100 mW of RF power. Long-term operation, however, can only be achieved if they are hermetically sealed in order to avoid contamination- and humidity-induced failure. Hermetically sealed capacitive switches have successfully switched over 100 billion cycles at room temperature and under low RF power conditions (20 dBm). Despite the aforementioned RF advantages, capacitive switches are currently not available commercially and are not widely utilized in defense or communication systems. This is primarily due to the facts that (1) high-yield manufacturing processes are not widely available yet and (2) their main failure modes such as dielectric charging, dc/RF gas discharge and metal creep and the physics behind them have not been adequately understood and addressed today.

Switch Structure and Actuation Mechanisms

Figure 1 shows a typical capacitive MEMS switch [1]. This is a shunt switch configuration, and is the dominant capacitive switch configuration in the literature today. The signal travels down the center conductor, and if the switch closes, will return along the outside conductors. It is, however, possible to design geometries for series configurations. Their characteristics, nevertheless, are very similar to the ones found in shunt switches. Consequently, this entry focuses primarily on shunt capacitive switches.

The switch in Fig. 1 consists of a movable beam or plate that is anchored to the switch substrate. While typically the term "plate" may better characterize the geometry of Fig. 1 for typical lateral switch dimensions, it is common in the RF MEMS literature to refer to this geometry as a fixed-fixed beam. Hence the term "beam" is adopted here to describe such capacitive switch geometries. Cantilever beams are also possible particularly for series switch configurations. The fixed-fixed beam anchors are typically metallic and are connected to the RF line. For example, they can be connected to the ground planes of a

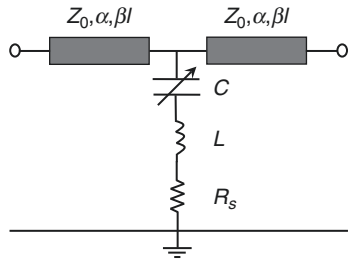
coplanar waveguide line as shown in Fig. 1b. The movable beam is typically composed of a thin-film (thickness h of 0.5–2 μm) metal such as gold, aluminum, nickel, or molybdenum. It is also possible that the beam is comprised of multiple layers including thin-film dielectrics such as silicon nitride or silicon dioxide and metals. One or more metallic pads are placed underneath the beam. In the simplest case, a single metallic pad is placed under the beam as shown in Fig. 1b. In this case, this pad is the center conductor of the coplanar waveguide line. This pad is usually covered by a thin (0.1–0.3 μm) dielectric layer such as silicon nitride, silicon dioxide, or more recently amorphous [2] or ultrananocrystalline [3] diamond.

The beam's length and width are determined by the required down-state switch capacitance as explained in the following section. This results in length and width in the tens to hundreds of μm for 5–40 GHz capacitive switches. The thickness of the dielectric layer also impacts the down-state capacitance. While high capacitance is in general required, thicknesses lower than 0.1 μm are hard to achieve in practice due to the need to handle high electric fields (dc and/or RF) across this dielectric. The gap (g_0) between the bottom surface of the movable beam and the top surface of the dielectric layer is determined by the need to minimize the up-state capacitance. A low up-state capacitance is necessary for low insertion loss. Capacitive switches in the 5–40 GHz range typically have gaps in the 2–5 μm range.

The switch has two states of operation. In the up state, the RF signal goes through the signal line almost unaffected by the movable beam. The up state is also called "zero biased" or "on state". In the down state the RF signal does not go through the RF line because it is reflected (see following section). The down state is also called the "biased" or "off state." An actuation mechanism is required to move the switch beam between these two states. Several possible actuation schemes exist, including electrostatic, electrothermal, magnetostatic, and piezoelectric. Table 1 summarizes the main advantages and drawbacks of each actuation scheme. The vast majority of reported capacitive switches are electrostatically actuated. They use the same actuation principle as the one originally proposed when the first capacitive switch was invented and reduced to practice in 1994 [4, 5]. In this original scheme a dc actuation voltage is applied between the movable beam and the actuation pad underneath it (Fig. 1). The beam is attracted due to the generated electrostatic field and collapses on the switch dielectric layer. Despite the need to generate a high dc voltage (30–100 V), which can be readily accomplished using a dc-dc converter, electrostatic switches exhibit the most desirable electromechanical characteristics, including the fastest possible response,

Capacitive MEMS Switches, Table 1 Main advantages and drawbacks of common actuation schemes for capacitive RF MEMS switches

Actuation mechanism	Advantages	Drawbacks
Electrostatic	Zero quiescent power consumption, easy biasing circuit, fast transient response (tens of microseconds)	Need to generate high voltage, high voltage may lead to charging and breakdown issues
Magnetostatic	Low voltage, high contact pressure, potentially low power with latching mechanism	Low quiescent power consumption requires latching, slower than electrostatic due to increased switch size
Electrothermal	Low voltage, high contact pressure, size comparable to electrostatic schemes	High quiescent power consumption (mW), slow response time (tens to hundreds of milliseconds)
Piezoelectric	Same as electrostatic by with low voltage	Difficult to achieve high-quality piezoelectric layer and integrate it with RF circuit at low temperatures



Switch Impedance

$$Z_s = R_s + j\omega L + \frac{1}{j\omega C},\ C = \begin{cases} C_{UP} \\ C_{DOWN} \end{cases}$$

$$f_0 = \frac{1}{2\pi\sqrt{LC}}$$

Typical Values		
	X-band	K-band
C_{UP}/C_{DOWN}	0.1/6 pF	0.04/3 pF
L	4–80 pH	6–50 pH
R_s	0.1–0.3 Ω	0.1–0.3 Ω

$$Z_s = \begin{cases} 1/j\omega C & for \quad f << f_0 \\ R_s & for \quad f = f_0 \\ j\omega L & for \quad f >> f_0 \end{cases}$$

Capacitive MEMS Switches, Fig. 2 Lumped-element equivalent circuit of the capacitive switch shown in Fig. 1. Typical values are provided for the lumped components of this circuit

zero quiescent power consumption, and the easiest possible biasing circuits. The resulting high fields though may lead to (gas and solid) dielectric charging and their associated reliability issues. More detailed discussion can be found in the last section.

RF Performance

Figure 2 shows a simple but physically meaningful and accurate lumped-element equivalent circuit of the switch shown in Fig. 1. It also includes typical values for all the equivalent circuit components [1]. The parameters α and β in this figure represent the attenuation constant and propagation constant of the transmission line respectively. The up- and down-state capacitance values are the most critical ones in this equivalent circuit. The up-state capacitance can often be accurately calculated by a typical quasi-static expression

$$C_{UP} = C_{PP} + C_{ff} = \frac{\varepsilon_0 A}{g_0 + \frac{t_d}{\varepsilon_r}} + C_{ff}$$

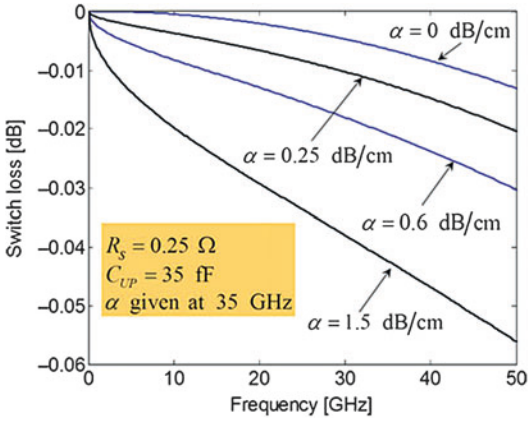
where C_{PP} and C_{ff} are the parallel-plate and fringing-field capacitances, respectively, A is the RF area of the switch ($A = Ww$ in Fig. 1), g_0 is the initial switch height, t_d is the dielectric layer thickness, and ε_r is this layer's dielectric constant. For typical switch geometries, the fringing-field capacitance could reach 25–50 % of the parallel plate capacitance. If improved accuracy is needed, a full-wave simulation is performed to estimate the switch up-state capacitance. The up-state capacitance must be sufficiently small to minimize the up-state insertion loss. Assuming a well-designed switch where the contributions from L and R_s can be ignored in the up state, the switch up-state reflection coefficient can be calculated from

$$S_{11} = \frac{-j\omega C_{UP} Z_0}{2 + j\omega C_{UP} Z_0}$$

where Z_0 is the characteristic impedance of the transmission line (typically 50 Ω). For example, an up-state capacitance of 70 fF results in $S_{11} < -10$ dB up to approximately 30 GHz. The up-state switch ohmic loss is the other critical up-state characteristic. The total ohmic loss of the switch can be calculated as

$$Loss = 1 - |S_{11}|^2 - |S_{21}|^2$$

This depends on (a) the attenuation α (dB/cm) of the transmission line underneath the movable beam and (b) on the switch series resistance R_s.



Capacitive MEMS Switches, Fig. 3 Simulated ohmic loss for a typical shunt capacitive switch. The attenuation *a* (dB/cm) depends on the transmission line characteristics

Well-designed capacitive switches can exhibit a total loss of less than 0.1 dB up to 40 GHz. Figure 3 shows numerical values for this loss for typical switch characteristics.

The switch down-state capacitance is more complicated to calculate because the switch beam may not be perfectly flat against the dielectric layer. Even in good designs this may not be possible due to the roughness of the layers involved. A model that is often used to capture the nonideal down-state switch capacitance is [1]

$$C_{DN} = \frac{\varepsilon_0 A}{2}\left(\frac{1}{r + \frac{t_d}{\varepsilon_r}} + \frac{\varepsilon_r}{t_d}\right)$$

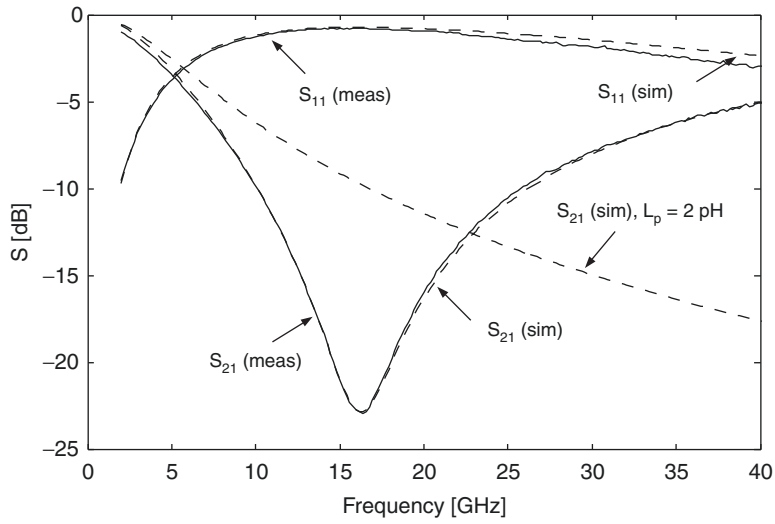
where r is the roughness amplitude. The down-state fringing-field capacitance is not included in this equation because it is typically not significant (<5 % of the parallel-plate capacitance) due to the small dielectric layer thickness. As shown by the equation above, the experimentally achieved down-state capacitance can vary greatly depending on the true contact area and the dielectric layer characteristics including its roughness. In practice, it is difficult to avoid a 30–50 % degradation of the down-state capacitance compared to the theoretical parallel-plate value. A high down-state capacitance (2–5 pF) is typically required in order to achieve an acceptable isolation (>20–50 dB) level at the desired frequency. One way to achieve this is to decrease the dielectric layer thickness. However, this dielectric layer needs to sustain very high electric fields (50–150 V/μm) across its thickness. Given typical fabrication process limitations that prohibit high-temperature growth processes for the dielectric layer (please see fabrication section), dielectric layers thinner than 0.1–0.2 μm become impractical. A second way to increase the down-state capacitance is to increase the dielectric constant. For instance, barium-strontium-titanate (BST) or strontium-titanate-oxide (STO) films with dielectric constants up to 400 can be employed. Besides additional fabrication complexities, these films have not been thoroughly studied in MEMS switches and may exhibit unacceptably high dielectric charging. Most switches typically employ some form of silicon dioxide or silicon nitride dielectric with dielectric constant in the 3.9–7.5 range. For switches with very low inductances (L < 1–2 pH), the switch isolation can be approximately calculated as

$$S_{21} = \frac{2}{2 + j\omega C_{DN} Z_0}$$

A more accurate calculation reveals that the switch inductance and series resistance also determine the total down-state isolation. In particular, its inductance cancels the down-state capacitance at the switch resonant frequency

$$f_0 = \frac{1}{2\pi\sqrt{LC_{DN}}}$$

This frequency can be adjusted by controlling the switch physical geometries. Switch inductances in the range of 1–100 pH can be readily achieved [1]. However, higher switch inductance typically results in a higher switch series resistance. Typical switch resistance values range in the 0.1–2 Ω range. The series resistance is the primary limiting factor of the switch isolation at its resonant frequency. At that frequency the switch isolation can be calculated as



Capacitive MEMS Switches, Fig. 4 Simulated and measured down-state scattering parameters of a capacitive MEMS switch (After Ref. [6] with permission)

$$S_{21} = \frac{2R_S}{2R_S + Z_0} \approx \frac{2R_S}{Z_0} \text{ at } f = f_0$$

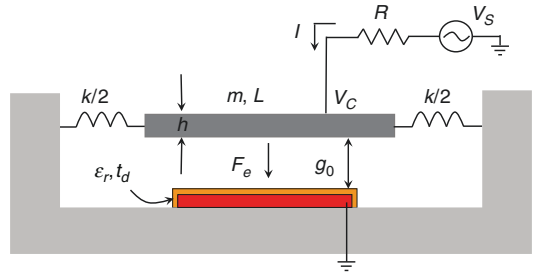
Figure 4 shows measured and simulated results of a typical capacitive switch [6] with $C_{DN} = 1.1\,\text{pF}$, $L = 87\,\text{pH}$, $R_s = 1.95\,\Omega$. This figure also shows the expected performance when the inductance is reduced to $L = 2\,\text{pH}$.

Electromechanical Considerations: Static Behavior

Figure 5 shows a simple but physically meaningful one-dimensional electromechanical model of the switch geometry of Fig. 1. The beam is modeled as a spring-mass system with a spring constant k. This spring constant depends on (a) the beam geometry, (b) the electrostatic force distribution on the beam, and (c) the residual stress of its structural film. The residual stress σ (MPa) is due to the fabrication process and depends on the exact deposition conditions. Typically a tensile stress (σ(MPa)) is needed in order to avoid buckling. The spring constant can be expressed as

$$k = k_1 + k_2$$

where k_1 depends on the first two factors and k_2 depends on the residual stress. The exact values



Capacitive MEMS Switches, Fig. 5 One-dimensional electromechanical model of a capacitive RF MEMS switch

can be calculated based on the specific switch design. For example, for the fixed-fixed beam of Fig. 1 and assuming that $W = L/3$ and that the electrostatic attractive force is uniformly distributed along the beam section directly above the coplanar waveguide center conductor, the spring constant can be calculated as [1]

$$\begin{aligned} k &= k_1 + k_2 \\ &= 32Ew\left(\tfrac{h}{L}\right)^3\left(\frac{27}{49}\right) + 8\sigma(1-\nu)w\left(\frac{h}{L}\right)\left(\frac{3}{5}\right) \end{aligned}$$

where ν is the beam's Poisson's ratio. For usual beam geometries and fabrication processes with residual stress in the order of 10–50 MPa, the second term dominates the spring constant. Typical spring constant values range from 10 to 50 N/m in order to provide sufficiently high restoring

force and avoid stiction issues. While low spring-constant designs have been successfully demonstrated [7], special care needs to be taken in avoiding stiction and self-actuation due to high RF power [8]. It is also important to mention that the above spring constant calculations are based on small-deflection theory. A nonlinear spring constant may need to be derived if this condition is not satisfied.

The electrostatic force F_e on the switch beam can be calculated as

$$F_e = \frac{\partial W_e}{\partial g} = \frac{1}{2} V^2 \frac{\partial C(g)}{\partial g} \approx \frac{1}{2} \frac{\varepsilon_0 A V^2}{g^2}$$

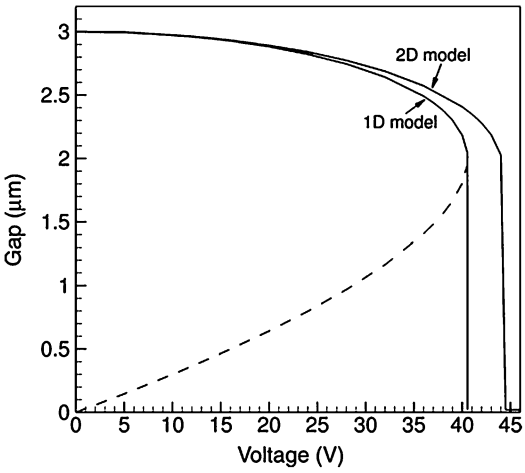
where W_e is the stored electrostatic energy, g is the switch gap between the beam and the actuation pad, and V is the applied electrostatic voltage. The last approximation is based on assuming a parallel-plate capacitance approximation and by ignoring the dielectric contribution. The static switch gap can be calculated by taking into account the static equilibrium of the forces applied on the beam

$$\frac{1}{2} \frac{\varepsilon_0 A V^2}{g^2} k(g_0 - g)$$

The above equation can be solved for the applied voltage as

$$V = \sqrt{\frac{2k}{\varepsilon_0 A} g^2 (g_0 - g)}$$

This equation is plotted in Fig. 6 for typical switch parameters. As Fig. 6 shows, there are two possible gaps for any given actuation voltage. This is not observed in practice and is a result of the unstable behavior of the beam. In particular, for small actuation voltages, the electrostatic force is increased proportionally to $\frac{1}{g^2}$. However, the restoring force is only increased proportionally to g. Hence, there is a critical gap beyond which the restoring force cannot hold the beam and the beam collapses on the dielectric surface. This critical gap can be found from the previous equation by taking its derivative and setting it up to zero



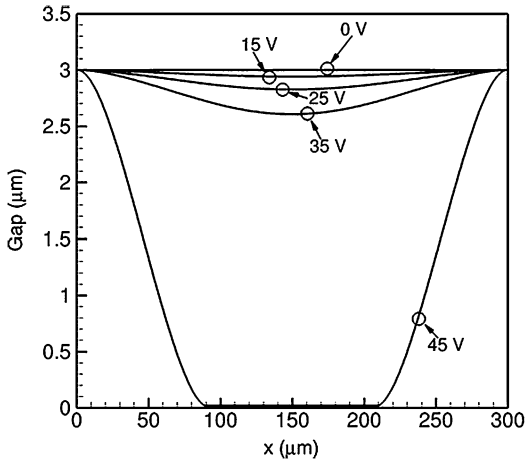
Capacitive MEMS Switches, Fig. 6 Simulated gap-voltage relationship for a capacitive MEMS switch with the following characteristics: $L = 300$ μm, $w = 100$ μm, $W = 120$ μm, $h = 1$ μm, $g_0 = 3$ μm, (Young's modulus) = 79 GPa, $\sigma = 10$ MPa. These results have been o btained with the PRISM center online simulation tool in MEMShub [13]

$$\frac{\partial V}{\partial g} = 0 \overset{g_c = \frac{2}{3} g_0}{\rightarrow}$$

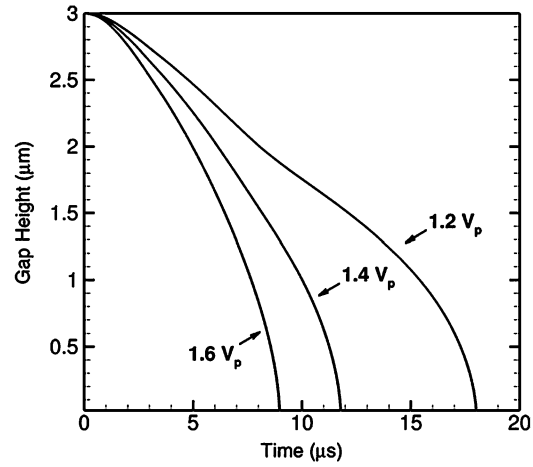
The voltage, therefore, required for actuating the beam, called the pull-in or pull-down voltage V_p, is given by

$$V_p = V(g_c) = \sqrt{\frac{8 k g_0^3}{27 \varepsilon_0 A}}$$

Figure 6 also shows the voltage-gap relationship as obtained by a two-dimensional beam model [13]. The gap plotted is between the center of the beam and the actuation pad. This curve is slightly different because the beam is not deformed as a perfectly flat object as assumed by the one-dimensional model. Figure 7 shows the actual deformed shape of the movable beam for voltages up to the pull-down voltage. Actuation voltages in the range of 30–100 V are typical in RF MEMS switches. Notice, however, that once the beam is actuated, the voltage required to hold the beam down (hold-down voltage V_h) is much lower because the gap between the beam and the actuation pad is much lower than g_0. It is hard to



Capacitive MEMS Switches, Fig. 7 Simulated shape of the movable beam of a capacitive MEMS switch for different bias voltages. The characteristics of the switch geometry are the same as in Fig. 6. These results have been obtained with the PRISM center online simulation tool in MEMShub [13]



Capacitive MEMS Switches, Fig. 8 Simulated switching time (pull-down) of a capacitive MEMS switch for different bias voltages. The characteristics of the switch geometry are the same as in Fig. 6. These results have been obtained with the PRISM center online simulation tool in MEMShub [13]

analytically calculate the hold-down voltage because it depends on many fabrication-dependent conditions such as the adhesion force between the beam and the dielectric layer. Typical hold-down voltages are in the range of 5–15 V [1]. Consequently, the gap-voltage relationship is strongly hysteretic. It is also worth mentioning that nonideal conditions such as an initial beam curvature and nonlinear bending are not included in this model. These effects can be captured by more complicated nonlinear beam models such as the ones presented in [9, 10].

Electromechanical Considerations: Dynamic Behavior

The dynamic behavior of the capacitive MEMS switch of Fig. 1 can be approximately captured by a one-dimensional model as described by the following equation

$$mg''(t) + bg(t) + kg(t) = F$$

where m is the switch mass, b is the damping coefficient, and f is the externally applied force. RF MEMS switches are typically packaged in an environment of 1 atm in order to avoid excessive ringing due to an underdamped response. As a result, the switch damping is dominated by squeeze-film damping as the gas under the beam is displaced during the switch motion. Due to the small gap g_0, it is in general difficult to accurately calculate the damping coefficient particularly in the near-contact region. This is further complicated by the possible existence of holes in the beam that aid its fabrication and substantially improve its switching speed. While there are several published approximations that can yield a reasonable approximation to the damping coefficient [1], accurate macro-models based primarily on rarefied gas dynamics only recently started becoming available [11, 12]. An equivalent way to characterize the switch damping is by its mechanical quality factor defined as

$$Q = \frac{k}{\omega_0 b}$$

where $\omega_0 = \sqrt{\frac{k}{m}}$ is the switch mechanical frequency. Typical switches show mechanical frequencies in the 20–100 kHz range and quality factors in the 0.5–2 range. Figures 8 (switch closure) and 9 (switch opening) illustrate dynamic

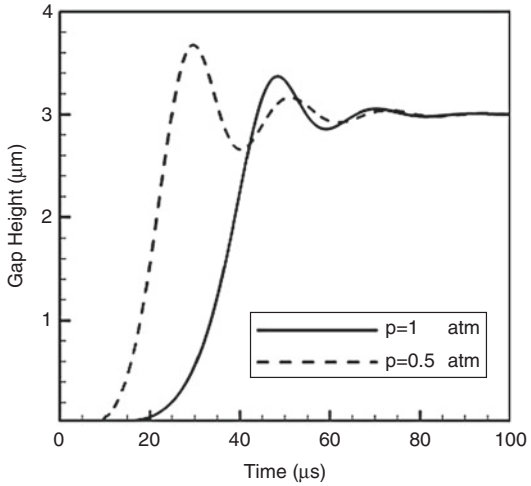
responses as calculated by two-dimensional beam models that accurately capture squeeze-film damping [13]. Notice that the displacement at the center of the beam is plotted in these graphs.

The switching speed can also be estimated based on the simple one-dimensional model. While it is difficult, in general, to derive an exact analytical solution, this model can provide reasonable approximations for common cases. For example, for $Q > 2$, the closing time can be estimated by [1]

$$t_s \approx 3.67 \frac{V_p}{V_s \omega_0}$$

where V_s is the applied voltage. Other limiting cases can be found in [1]. In general, closing times in the 5–50 μs range can be achieved. A similar range is typically possible for the release times.

The one-dimensional model can also be used to estimate the velocity and acceleration of the switch. Switching velocity in the 1–10 m/s range can be observed in the near-contact region. Due to its small mass, the switch acceleration can exceed 10^6 m/s^2 in the same region.

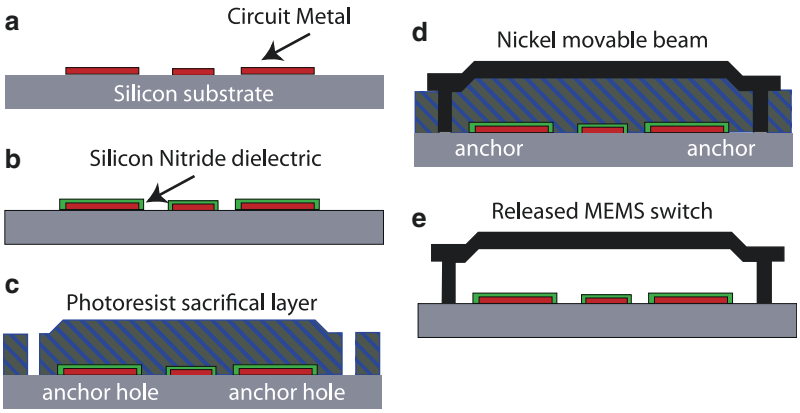


Capacitive MEMS Switches, Fig. 9 Simulated switching time (release) of a capacitive MEMS switch for different pressure levels. The characteristics of the switch geometry are the same as in Fig. 6. These results have been obtained with the PRISM center online simulation tool in MEMShub [13]

Fabrication Methods

Capacitive switches can be fabricated with conventional micromachining processes and require a small number of masks. Figure 10 illustrates the masks of a typical fabrication process.

- Step (a): The first mask defines the circuit metal of Fig. 1 after this metal layer is deposited on the substrate through evaporation, sputtering, or electroplating. Gold, aluminum, and copper are common metal choices. A smooth metal surface is particularly important directly underneath the switch beam in order to minimize local electric field enhancement.
- Step (b): The second mask defines the dielectric layer to cover the portion of the metal that



Capacitive MEMS Switches, Fig. 10 Simplified typical fabrication process for a capacitive MEMS switch

will be under the beam. Unless a high melting temperature metal has been deposited (e.g., tungsten), a relatively low-temperature process is required for the deposition and patterning of the dielectric layer in order to avoid damaging the circuit metal. Plasma-enhanced chemical vapor deposition (PECVD) is the most common process for depositing silicon nitride/oxide films. This is usually followed by a reactive ion etching (RIE) step that helps etching the unwanted dielectric layer parts.

- Step (c): The third mask is used to define the sacrificial layer of the switch. The sacrificial layer is the layer upon which the beam will be deposited, and this material needs to be removed at the end of the process to release the beam. The beam anchor points are defined in this step by selectively etching the sacrificial layer. This can be a dry- or a wet-etch step. More complicated processes may involve an additional planarization step before the next mask. The choice of the sacrificial layer material is critical as it controls many important parameters of the switch design, including the residual stress of the beam. Common material choices include photoresists, other photoconductive polymers, and polyimides. There is no sufficient understanding in the open literature of the exact processes that are involved in controlling the beam layer residual stress in the presence of a sacrificial layer. Important parameters though include, among others, the atomic structures of each film and the deposition temperatures of each film.
- Step (d): The fourth mask defines the actual beam layer. A variety of processes can be utilized including evaporation, sputtering, and electroplating of the beam layer(s). This step is also critical in determining the final residual stress of the beam.
- Step (e): The beam is finally released by etching (wet or dry) the sacrificial layer and by drying (if needed) the switch wafer. Etching of the sacrificial layer can also influence the beam residual stress particularly if a high-temperature process is necessary. If wafer drying is needed, this needs to be done carefully in order to avoid damage to the beam or causing stiction to the substrate. Stiction may occur if drying involves removing a liquid with high surface tension (e.g., water) underneath the beam. Such a liquid will pull the beam down as it evaporates. Special drying processes and equipment based on supercritical carbon dioxide have been successfully developed [14] and followed by many MEMS researchers.

The last step in the fabrication process is packaging, which is discussed in the following section.

Packaging

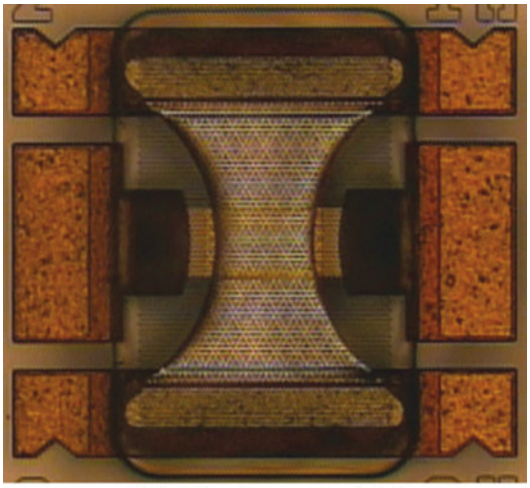
Hermetic packaging is required for capacitive RF MEMS switches to avoid any contamination- or humidity-induced early failure. While conventional hermetic packages exist, they are not well suited for capacitive RF MEMS switches or circuits. First, if switches need to be inserted in conventional hermetic packages, these switches will have to be diced first since several thousands of them can be simultaneously fabricated on a wafer. Dicing released switches is particularly dangerous for the switches and may considerably reduce the process yield. Second, conventional hermetic packages are expensive (∼tens of dollars/package) and not well-suited for cost-driven consumer applications. Third, they typically exhibit a relatively high insertion loss, which is often much higher than the switch itself (e.g., a DC-40 GHz package could exhibit 0.6 dB at 20 GHz [1]).

As a result, it is important to follow a cost-effective on-wafer hermetic packaging scheme. In this case, a wafer-scale package is first completed and then dicing follows. A wide variety of approaches have been developed so far to accomplish this. These approaches can be divided into three main categories:

- Two-wafer hermetic packages completed by fusion, glass-frit, thermocompression, eutectic, or anodic bonding. These packages are created by bonding two wafers together using one of the aforementioned approaches. The main advantage of these techniques is that they

result in excellent hermetic bonds. Their main drawback is that they may require high temperatures (300–1,000 °C depending on the technique) with the exception of low-temperature eutectic bonds (e.g., indium–gold bonds). In addition, these techniques tend to be relatively expensive since packaging cost usually accounts for 60–80 % of the total cost.

- Two-wafer quasi-hermetic packages completed by low-temperature polymer or solder-bump bonding. This technique is similar to the previous one, except that sealing is achieved by low-temperature bonding (room temperature −150 °C). A wide variety of polymers can be used for this. While low temperatures can be achieved, packages fabricated with such bonding typically exhibit very low but nonzero leak rates [1].
- Hermetic packages fabricated by on-wafer microencapsulation using micromachining techniques [15]. These techniques do not require a second wafer cap. Instead, every switch or switching circuit is encapsulated in a tiny package on its own wafer by a microfabricated technique. Typical temperatures in this process are in the 200–250 °C range. These techniques are ideally suited for the small size and high RF bandwidth of MEMS devices and typically result in low-cost fabrication. Particular attention needs to be paid though to ensure compatibility between the switch and package fabrication processes. Figure 11 shows a switch packaged with this technique.



Capacitive MEMS Switches, Fig. 11 The packaged memtronics switch (After Ref. [15] with permission)

Circuits and Applications

Capacitive MEMS switches may be employed in a number of circuits mostly for communication, radar, and electronic warfare systems [1]. Variable capacitors and impedance tuners for cell phones and other radios in mobile form factors constitute the most important applications in the commercial sector. MEMS variable capacitors (varactors) can be formed by connecting in parallel several capacitive MEMS switches and selectively activating them. Several companies including WiSpry and Cavendish Kinetics are pursuing this particularly due to the large cell phone market size.

The most important of the high-frequency circuit applications (>10 GHz) are high-isolation switching packets, true-time delay networks and phase shifters, reconfigurable impedance tuners for amplifiers and antennas, high-quality-factor reconfigurable filters, and tunable oscillators. Many of these circuits exploit the near-ideal RF performance of capacitive switches. Consequently, optimal performance can be usually achieved by employing a circuit- or sub-system-level package instead of a device-level package. Examples of several of these circuits can be found in [1].

Failure Mechanisms and Reliability

Capacitive MEMS switches suffer from high electric dc and/or RF fields through narrow gaps and dielectric layers. These fields can readily reach 5–50 V/μm in the up state and may increase further during actuation. Such fields may cause field emission and ionize the gas in the switch gap [16]. The long-term effects of field emission and

gas discharge are not known at this point. In addition, when such fields are applied across a thin-film solid dielectric, charges may get trapped in the dielectric layer leading to dielectric charging. A number of studies have been completed (see for example [17, 18]) focusing mostly on charges trapped in the bulk of the dielectric. However, surface charging of the solid dielectric that is also influenced by gas ionization is potentially more detrimental to the switch performance and is not well understood today. Charging phenomena are the leading cause of failure in capacitive switches today. Long-term drift of actuation voltage, stiction, and breakdown can be observed as a result of these charging issues.

Besides solid and gas dielectric charging, metal creep is another potential failure mechanism. Creep may be developed in the movable beam material if a switch is subjected to a constant stress. For example, if a switch is left in its down state for a long time (typically tens to thousands of hours), the beam material may creep resulting in a temporary or permanent change of the switch spring constant. Several recent papers show it is potentially an area of concern for capacitive MEMS switches [19, 20]. Creep at high temperature may be even a more significant area of concern.

Other possible failure modes include

- Beam buckling due to high temperatures. This may be caused during release process, normal high-temperature operation, or due to high RF currents through the movable beam under high RF power conditions.
- Self-actuation of the switch movable beam due to high RF power [8]. A high RF voltage may result in self-actuation because the attractive electrostatic force is proportional to the square of the switch voltage. This limits the switch power handling.
- Hot switching failure. When a capacitive switch needs to interrupt high RF currents or sustain high transient RF voltages, abrupt chemical changes may occur at its surfaces leading to premature wear and failure. This is related to the dielectric charging phenomena.
- Shock-induced failure. High shocks (>30,000–100,000 g) may result in beam fracture particularly if contact is achieved. Such events are rare in most applications.

Failures related to cycling-induced fatigue, crack generation, and fracture are not typically observed under normal operating conditions. However, they may become important at extreme temperatures particularly for movable beams based on thin-film metals. Despite the aforementioned failure modes, the best switches today have achieved over 100 billion cycles under typical laboratory conditions when driven by 30 kHz bipolar bias waveforms with approximately 35 V peak amplitude. These devices were hot-switched at a power level of 20 dBm at 35 GHz [21]. However, these cannot be considered typical results. Early failures are found in several wafer samples. Additional research is required to increase the observed reliability and limit early failures that are commonly due to poor fabrication process control.

Cross-References

- ▶ Basic MEMS Actuators
- ▶ NEMS Piezoelectric Switches
- ▶ Piezoelectric MEMS Switch

References

1. Rebeiz, G.M.: RF MEMS Theory, Design, and Technology. Wiley, Hoboken (2003)
2. Webster, J.R., Dyck, C.W., Sullivan, J.P., Friedmann, T.A., Carton, A.J.: Performance of amorphous diamond RF MEMS capacitive switch. Electron. Lett. **40**(1), 43 (2004)
3. Goldsmith, C., Sumant, A., Auciello, O., Carlisle, J., Zeng, H., Hwang, J.C.M., Palego, C., Wang, W., Carpick, R., Adiga, V.P., Datta, A., Gudeman, C., O'Brien, S., Sampath, S.: Charging characteristics of ultra-nano-crystalline diamond in RF MEMS capacitive switches. In: Proceedings of the IEEE MTT-S International Microwave Digest, Anaheim, pp. 1246–1249, May 2010
4. Goldsmith, C.L., Kanack, B.M., Lin, T., Norvell, B.R., Pang, L.Y., Powers, B., Rhoads, C., Seymour, D.: Micromechanical microwave switching. US Patent 5,619,061, 31 Oct 1994

5. Goldsmith, C.L., Yao, Z., Eshelman, S., Denniston, D.: Performance of low-loss RF MEMS capacitive switches. IEEE Microwave Wireless Compon. Lett. **8**(8), 269–271 (1998)
6. Peroulis, D.: RF MEMS devices for multifunctional integrated circuits and antennas. PhD Dissertation, The University of Michigan, Ann Arbor (2003)
7. Peroulis, D., Pacheco, S.P., Sarabandi, K., Katehi, L.P.B.: Electromechanical considerations in developing low-voltage RF MEMS switches. IEEE T. Microw. Theory **51**(1), 259–270 (2003)
8. Peroulis, D., Pacheco, S.P., Katehi, L.P.B.: RF MEMS switches with enhanced power-handling capabilities. IEEE T. Microw. Theory **52**(1), 59–68 (2004)
9. Snow, M.: Comprehensive modeling of electrostatically actuated MEMS beams including uncertainty quantification. MSc thesis, Purdue University, West Lafayette (2010)
10. Younis, M.I., Abdel-Rahman, E.M., Nayfeh, A.: A - reduced-order model for electrically actuated microbeam-based MEMS. J. Microelectromech. Syst. **12**(5), 672–680 (2003)
11. Guo, X., Alexeenko, A.: Compact model of squeeze-film damping based on rarefied flow simulations. J. Micromech. Microeng. **19**(4), 045026 (2009)
12. Parkos, D., Raghunathan, N., Venkattraman, A., Alexeenko, A., Peroulis, D.: Near-contact damping model and dynamic response of micro-beams under high-g loads. In: Proceedings of the IEEE International Conference on Micro Electro Mechanical Systems (MEMS), Cancun, pp. 465–468, Jan 2011
13. Ayyaswamy, V., Alexeenko, A.: Coarse-grained model for RF MEMS device. MEMShub.org, http://memshub.org/resources/prismcg. Accessed Mar 2011
14. Tousimis: http://www.tousimis.com/. Accessed Mar 2011
15. Memtronics: http://www.memtronics.com/. Accessed Mar 2011
16. Garg, A., Ayyaswamy, V., Kovacs, A., Alexeenko, A., Peroulis, D.: Direct measurement of field emission current in E-static MEMS structures. In: Proceedings of the 24th IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2011), pp. 412–415, Jan 2011
17. Peng, Z., Yuan, X., Hwang, J.C.M., Forehand, D.I., Goldsmith, C.L.: Superposition model for dielectric charging of RF MEMS capacitive switches under bipolar control-voltage waveforms. IEEE T. Microw. Theory **55**(12), 2911–2918 (2007)
18. Papaioannou, G., Exarchos, M.-N., Theonas, V., Wang, G., Papapolymerou, J.: Temperature study of the dielectric polarization effects of capacitive RF MEMS switches. IEEE T. Microw. Theory **53**(11), 3467–3473 (2005)
19. Hsu, H.-H., Peroulis, D.: A viscoelastic-aware experimentally-derived model for analog RF MEMS varactors. In: Proceedings of the 23rd IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2010), Wanchai, pp. 783–786, Jan 2010
20. McLean, M., Brown, W.L., Vinci, R.P.: Temperature-dependent viscoelasticity in thin Au films and consequences for MEMS devices. IEEE/ASME J. Microelectromech. Syst. **19**(6), 1299–1308 (2010)
21. Goldsmith, C., Maciel, J., McKillop, J.: Demonstrating reliability. IEEE Microw. Mag. **8**(6), 56–60 (2007)

Capillarity Induced Folding

▶ Capillary Origami

Capillary Flow

Prashant R. Waghmare and Sushanta K. Mitra
Micro and Nano-scale Transport Laboratory, Department of Mechanical Engineering, University of Alberta, Edmonton, AB, Canada

Synonyms

Passive pumping; Surface tension–driven flow

Definition

The fluid flow in an enclosed conduit due to simultaneous changes in the inherent surface energies of fluid and solid surface of the conduit.

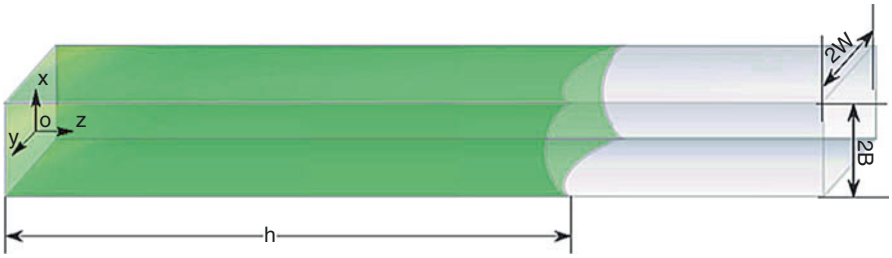
Introduction

Everything in the universe has its own state of energy, which is represented by possible combinations of 132 elements of periodic table. The simplest form of each element is an atom and each atom has three different components: proton, electron, and neutron. Further, each element has a fixed number of electrons, which arrange themselves in different shells, and the number of electrons in each shell can be determined by Bohr's theory. Several individual elements or atoms do not have sufficient number of electrons

in their outer shell and this makes it unstable and further the element tries to become stable by searching for required electrons. The finding of sufficient electrons at the outer shell results in the formation of a molecule or in bulk cluster of molecules. The surface or interface formation with this cluster of molecules creates an imbalance in the arrangement and orientation of the molecules in the cluster, mainly across the interface. This imbalance represents the energy of system. Every system in the universe tries to attain the minimum state of energy and, therefore, in the case of liquid in the air, it has been observed that the raindrops always attain the spherical shape. Liquid molecules have capability to orient themselves. Hence, they can form the shape of minimum energy but in the case of solid molecules, they try to minimize the state of energy by covering up the liquid if it comes in contact. In this case, the solid surfaces are not in equilibrium with the saturated vapor, i.e., they are not in a state of minimum energy. The moment at which conduit or channel with high energy level surfaces comes in contact with lower energy level liquid interface, the solid surface tries to envelope itself with the liquid to attain the minimum energy. In the case of the higher energy liquid in comparison with the solid surfaces, the liquid interface tries to minimize its surface. The prior case can be illustrated with the water (72 dyn/cm) in glass channel (~200–300 dyn/cm) and the mercury (~700–735 dyn/cm) with the same glass channel is an example of the latter case.

The recent developments in the microfabrication technologies allowed to fabricate features of sizes from micro- to nanoscales. The surface to volume ratio of the feature increases as the scale of the feature decreases which in turn makes surface forces dominant over other forces. Because of high surface forces, very high pressure is required to pump the fluid in microchannels. Hence, researchers have developed different nonmechanical pumping mechanism with the help of electrokinetic and/or megnetohydrodynamic approaches. Generation and actuation of electric and/or magnetic field are an additional burden on the system which increases both the fabrication process and cost of the device. Therefore, attempts are being made to develop a flow without any external means. The fluid flow can be achieved by controlling surface chemistry, fluid properties like surface tension, or by changing the geometries. Such transport of fluid is called autonomous flow or autonomous pumping which is an ideal transport mechanism for microfluidic applications. The dominance of surface forces at microscale plays a significant role in deciding the pumping approach in microfluidic devices. Hence, nowadays surface tension–driven flow in microfluidic devices has widely attracted the attention of researchers. As explained earlier, the capillary action is the interplay between the surface energies or surface tension between the fluid and solid surface in contact. Moreover, at microscale, due to very high surface to volume ratio, the possibility of available surface area is very high. One can easily pump the fluid with capillarity provided the fluid has lower surface energy than the solid surfaces. For optimum design and function of any microfluidic device which works on capillary flow principle, it is essential to predict its behavior in advance and it is therefore necessary to perform the theoretical analysis of capillary flow within the microchannels. The prediction of the temporal variations in the flow front position along capillary length is the ultimate goal of the theoretical analysis. Therefore, generally the analysis is performed to predict the flow front position, i.e., penetration depth for given working and operating conditions.

Over the last century, the capillary phenomenon has become a topic of interest due to its importance in several areas. First time in literature, Washburn proposed a closed form solution for the penetration depth in a channel of millimeter dimension. The closed form solution is derived by balancing the surface tension force to viscous force and it is observed that the penetration depth is proportional to the square root of the time. As explained earlier, capillary phenomenon is the change in the surface energy process and to encompass the concept of change in the surface energy, one can use the thermodynamic approach for analysis. The surface energy of solid is the



Capillary Flow, Fig. 1 Schematic of the microchannel of width 2B, depth 2 W considered for theoretical modeling [6]

topic of ongoing debate. Moreover, there are several effects like dynamic contact angle, inlet effects, reservoir effects, suspension flow, etc., which require tedious and cumbersome analysis. Attempts are also being made to present analysis with microscopic energy balance approach where different forces are accounted in terms of different forms of energy. On the other hand, hydrodynamic models, based on the conventional fluid mechanics principle, are easy to implement. Such models are mainly developed by two distinct approaches in the literature, namely, differential and integral approach. The moving fluid-air interface with the differential approach becomes computationally costly, whereas the integral approach with moving control volume provides a simple form of ordinary differential equation. In such modeling or analysis, the governing equation for flow front transport is obtained by balancing different forces like viscous, inertial, gravity, pressure forces, etc. The velocity-dependent terms like inertial and viscous terms of the governing equation are determined with the velocity profile across the channel. In the literature, the steady state assumption is applied from the very entrance of the channel neglecting the entrance length effect. This assumption has been widely adopted till date as done by Washburn where the time and length scale used for the validation of the theoretical model was big. Hence, the assumption of a steady state velocity profile holds true. Whereas, in the case of microfluidic channels the length and timescale is very small; hence, the Washburn prediction does not follow the observation as demonstrated by Saha and Mitra [1, 2]. The later part of this entry is dedicated to emphasize the importance of such microscale effects in the analysis. Main emphasis is given to the integral approach based on modeling due to the ease of its adaptability. Therefore, in the next section, the overview of the modeling of the capillary transport is discussed in brief.

Mathematical Modeling

Figure 1 shows the microchannel of width 2B and depth of 2 W is considered for the theoretical modeling. The momentum equation in integral form for homogeneous, incompressible, and Newtonian fluid can be written as [3]:

$$\sum F_z = \frac{\partial}{\partial t}\int_0^h \int_{-W}^{W} \int_{-B}^{B} \rho v_z dx dy dz + \int_{-W}^{W} \times \int_{-B}^{B} v_z(-\rho v_z) dx dy \tag{1}$$

where $\sum F_z$ refers to all forces present during the development of the fluid-air interface, ρ is the fluid density. Generally, forces present during the capillary transport are viscous (F_v), gravity $\left(F_g = 4\rho g h B W\right)$, pressure forces at the flow front (F_{pf}), and at the inlet (F_{pi}).

$$\sum F_z = \underbrace{F_v}_{\text{velocity dependent}} + \underbrace{F_g + F_{pf} + F_{pi}}_{\text{velovity independent}} \tag{2}$$

The Eq. 1 contains three velocity-dependent terms, namely, transient, convective, and viscous force which is generally determined as velocity profile, v_z, across the channel. The fully developed

flow assumption, i.e., Poiseuille flow assumption, is widely used in the literature neglecting the transience in the velocity profile [4, 5]. The consequence of such assumptions particularly at microscales will be discussed in detail in the later part.

The steady state velocity profile can be used to determine the velocity-dependent terms of the momentum equation. Moreover, remaining pressure force terms are calculated with available expressions (from the literature) for pressure fields at respective locations. The pressure force at the fluid-air interface can be determined by well-known Young-Laplace equation with fluid surface tension (σ) and equilibrium contact angle (θe). The approximated pressure field expression at the entrance of the microchannel is used widely to determine the pressure force at the entrance of the microchannel. Levin et al. [7], for the first time in the literature, claimed that atmospheric pressure cannot be used as entrance pressure at the inlet of the microchannel. The pressure field expression for circular capillary is derived by assuming a separate hemispherical control volume as fluid source other than the control volume considered within the microchannel. Further several researchers have used same expression for rectangular microchannels with an assumption of equivalent radius. In such analysis, with an equivalent radius assumption, the hemispherical control volume of equivalent radius of projected area of rectangular microchannel entrance is presented in Eq. 3.

$$p(o,t) = p_{atm} - \left\{ 1.11\rho\sqrt{BW}\frac{d^2h}{dt^2} + 1.58\rho\left(\frac{dh}{dt}\right)^2 + \frac{1.772\mu\, dh}{\sqrt{BW}\, dt} \right\} \tag{3}$$

The importance of an appropriate entrance pressure field expression for rectangular microchannel is also discussed in detail in later part of the study. Finally, determining all terms of Eq. 1 and rearranging as per order of differential operator, one can obtain the dimension form of the ordinary differential equation which governs the capillary transport in the microchannels. Further, nondimensional governing equation as shown in Eq. 4 can be obtained by performing nondimensional analysis with characteristic time, $t_0 = \frac{\rho(2B)^2}{12\mu}$ and characteristic length $h_0 = 2B$.

Capillary Flow, Table 1 Constants of the generalized nondimensional governing equation for a capillary flow in a microchannel with fully developed velocity profile [3]

Constants	Expressions
C_1	$\frac{0.55}{\sqrt{\gamma}}$
C_2	0.958
C_3	1
C_4	$0.295\sqrt{\gamma}$
C_5	$\frac{B_o}{144Oh^2}$
C_6	$\frac{\gamma - cos\,\theta_e}{72Oh^2}$

$$(h^* + C_1)\frac{d^2h^*}{dt^{*2}} + C_2\left(\frac{dh^*}{dt^*}\right)^2 + (C_3 + C_4h^*)\frac{dh^*}{dt^*} + C_5h^* + C_6 = 0 \tag{4}$$

The coefficients of Eq. 4 are tabulated in Table 1. Two nondimensional numbers are obtained, i.e., Bond number (Bo) and Ohnesorge number (Oh). The constants of this equation are functions of different nondimensional groups like Ohnesorge number (Oh), Bond number (Bo), and aspect ratio (γ). The Ohnesorge number represents the ratio of viscous to surface tension force, i.e., $Oh = \frac{\mu}{\sqrt{2B_{\rho\sigma}}}$, the Bond number dictates the ratio of gravity to surface tension force, i.e., $\left(B_O = \frac{\rho g(2B^2)}{\sigma}\right)$.

The solution of Eq. 4 predicts the penetration depth with capillary flow. The numerical [8] and analytical [3] solutions of Eq. 4 are available in the literature.

Revisiting the Assumptions for Microscale Applications

As mentioned earlier, the velocity-dependent terms of momentum equation are determined with an assumption of a steady state. It is assumed that at the very entrance of the microchannel the flow is fully developed. However, in reality, three different flow regimes can be observed in the capillary flow: entry regime, Poiseuille regime,

where the flow is fully developed, and the regime behind the fluid-air interface, i.e., surface tension regime. The steady state assumption, i.e., parabolic velocity profile assumption, is valid for steady state flow, whereas capillary flow is inherently a transient phenomenon. The parabolic velocity profile assumption is a reasonably good assumption for macroscale capillaries as shown by several researches [5, 7, 9]. Moreover, such assumptions are only valid in the case of a very high viscous fluid or very low Reynolds number flow which may not be true in every case [10]. Hence, it is important to consider and analyze such transience in the analysis at microscale. This can be tackled by considering the transient or developing velocity profile instead of the steady state velocity profile as explained in the following section.

The transient momentum equation in the direction of the flow for pressure-driven flow is:

$$\rho \frac{\partial v_z}{\partial t} = \mu \frac{\partial^2 v_z}{\partial x^2} + \frac{dp}{dz} \tag{5}$$

The velocity in Eq. 5 is a combination of steady and transient part of velocity as depicted in Eq. 6 [11]:

$$v_z(x,t) = v_{z\infty}(x) + v_{zt}(x,t) \tag{6}$$

where $v_{z\infty}(x)$ is the fully developed or steady state velocity, i.e.,

$$v_{z\infty}(x) = \frac{B^2}{2\mu}\frac{dp}{dz}\left[1 - \left(\frac{x}{B}\right)^2\right] \tag{7}$$

The transient part of the velocity can be obtained by separation of the variable method which can be given as [6]:

$$v_{zt}(x,t) = 2\sum_{n=1}^{\infty}(-1)^n\left[\frac{1}{B\mu\lambda_n^3}\frac{dp}{dz}\right]cos\,(\lambda_n x)exp \times \left(-v\lambda_n^2 t\right) \tag{8}$$

where v is the kinematic viscosity of the fluid and $\lambda_n = \frac{(2n-1)\pi}{2B}$. By combining Eqs. 5 and 7 the transient velocity profile $v_z(x, t)$ can be obtained by:

$$v_z(x,t) = \left\{\sum_{n=1}^{\infty}(-1)^n\frac{1}{B\mu}\left(\frac{2}{\lambda_n^3}\right)cos\,(\lambda_n x)exp \times \left(-v\lambda_n^2 t\right) + \frac{1}{2\mu}\left(B^2 - x^2\right)\right\}\frac{dp}{dz} \tag{9}$$

Further, the average velocity across the channel can be represented as:

$$v_z(t)_{avg} = \frac{B^2}{3\mu}\left[1 - \sum_{n=1}^{\infty}\frac{96}{(2n-1)^4\pi^2}exp \times \left(-\frac{(2n-1)^2\pi^2 vt}{4B^2}\right)\right]\frac{dp}{dz} \tag{10}$$

Finally, the velocity profile in terms of the penetration with transient velocity is:

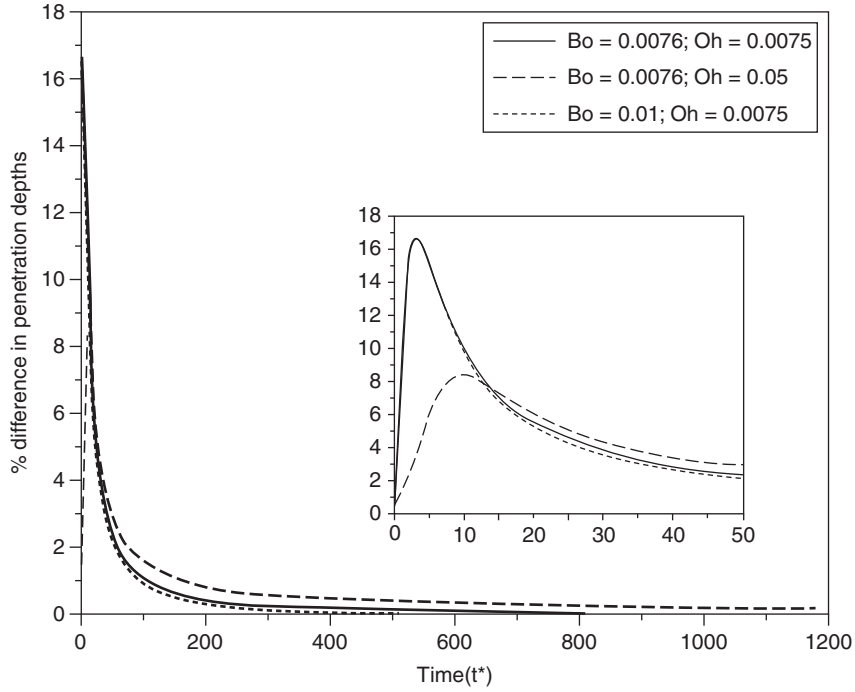
$$v_z(x,t) = \frac{B^2}{2\mu}\alpha_1\left\{\sum_{n=1}^{\infty}(-1)^n\left[-\frac{4}{(B\lambda_n)^3}\right]cos\,(\lambda_n x)exp\left(-v\lambda_n^2 t\right) + \left(1 - \frac{x^2}{B^2}\right) \times \left\{\frac{1}{\alpha_1\left[1 - \sum_{n=1}^{\infty}\beta_1 exp\left(-\lambda_n^2 vt\right)\right]}\right\}\frac{dh}{dt} \tag{11}$$

where

$$\alpha_1 = \frac{\left[(\phi)^4 - 4exp - \frac{\phi^2 t}{3}\right]}{\left[(\phi)^4 - 6exp - \frac{\phi^2 t}{3}\right]}$$

and $\phi = \lambda_n B$.

A similar approach can be followed as explained in the previous section and further the governing equation for capillary transport can be derived with the transient velocity profile provided in Eq. 11. Moreover, the difference in the penetration depth with both approaches, i.e., with the steady state and transient velocity profile, under different operating conditions can be compared. Figure 2 shows the difference in the penetration depth in such cases where the difference in the penetration depth is more at the beginning of

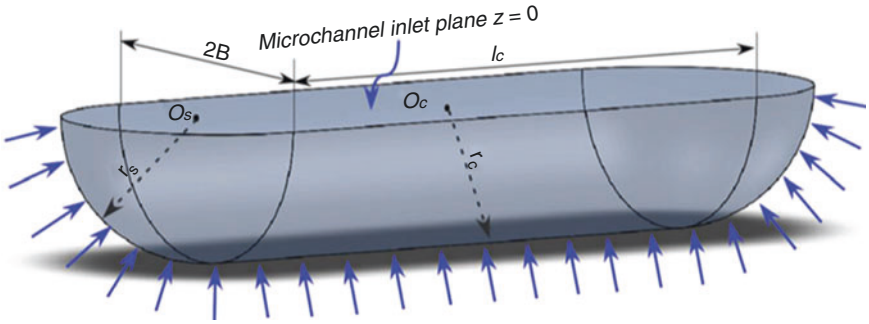


Capillary Flow, Fig. 2 Transient response in the difference in the penetration depths with the fully developed (steady state) and developing (unsteady) velocity profile under different conditions [6]

the filling process, as shown in the inset of Fig. 2. This difference in the penetration depth decreases as the flow progresses along the microchannel where the flow becomes a fully developed flow. This can also be explained with the help of boundary layer theory which is the effect of fluid viscosity. The boundary layer thickness increases as the viscosity of fluid increases because of the retardation of flow due to increase in the viscosity, whereas in the case of the fluid density, the effect is opposite to viscosity. Therefore, the difference in penetration depth with the high density fluid (Bo $= 0.01$) is higher than the difference with the high viscous fluid. It is evident from the analysis that the transience effect in the analysis has a significant impact on the filling process prediction, particularly at the beginning of the filling process. At microscale, such difference needs to be accounted prior to the design.

As discussed earlier, the pressure force at the entrance of the microchannel is determined with the help of the pressure field at the microchannel entrance. Several researchers [3–8] have adopted the pressure field expression with an equivalent radius assumption. Levin et al. [6] developed an entrance pressure field expression for circular capillary, assuming a hemispherical control volume as a separate control volume at the entrance which is responsible for a sink flow at the entrance of capillary and the pressure field. Moreover, a similar expression for rectangular capillaries is extended with an equivalent radius assumption. In such cases, the radius of circular capillary is replaced by the equivalent radius of projected area at the entrance of the channel. This is not a realistic representation for noncircular capillaries particularly for high aspect ratio microchannels where it is not appropriate to consider the hemispherical control volume for the sink flow or pressure field at the microchannel entrance. In the case of such geometries, the control volume needs to be considered as a combination of semicylinder and hemisphere as shown in Fig. 3.

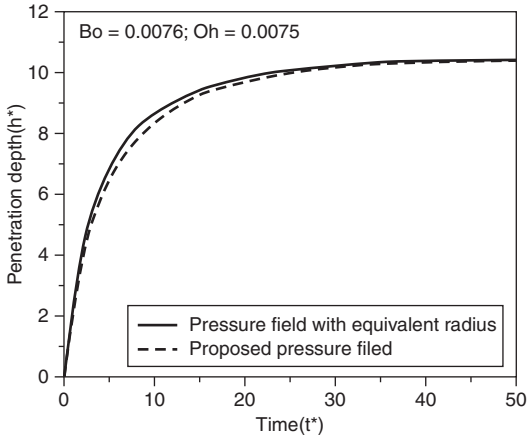
The detailed derivation of the pressure field expression with this control volume can be seen in [6] which is:



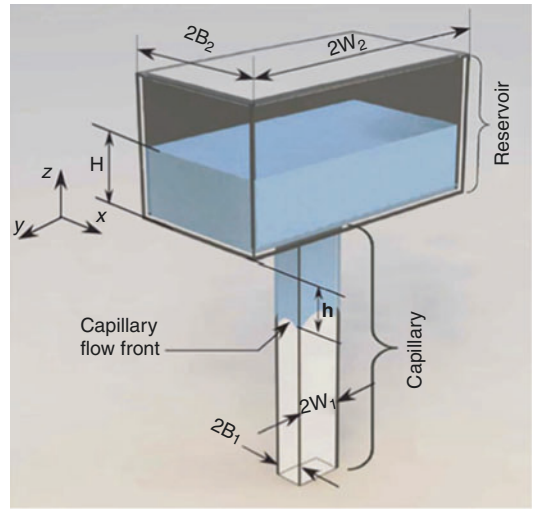
Capillary Flow, Fig. 3 The fluid volume from infinite reservoir considered as control volume for pressure field expression analysis in the case of rectangular microchannel. The *arrow* shows the direction of the fluid flow from the reservoir into the microchannel [6]

$$p(0,t) = p_{atm} - \rho B\left\{\left[\frac{4\gamma + 3(1-\gamma)}{24}\right]\left\{\pi\left[\frac{1}{2\pi} + \frac{2}{\pi^2} + \frac{6}{10}\right]\right\} + \left[1 - \frac{2}{\pi} ln\frac{R_\infty}{B}\right]\right\}\frac{d^2h}{dt^2}$$
$$+\rho\left\{\left[\frac{1(1-\gamma)}{\pi^2} - \frac{6}{5}\right] - \left[\frac{4\gamma + 3(1-\gamma)}{6}\right] \times \left[\frac{(2-\gamma)}{2\pi} - \frac{(1-\gamma)}{\pi^2}\right]\right\}\left(\frac{dh}{dt}\right)^2 - \frac{4\mu}{B} \times \left[(2-\gamma) + \frac{(1-\gamma)}{\pi}\right]\frac{dh}{dt} \quad (12)$$

where R_∞ represents the radial distance far away from the control volume in the reservoir, where the sink action, i.e., entrance pressure force, disappears. One can re-derive the governing Eq. 4, using pressure field expression presented in the Eq. 12, and determine the effect of such a pressure field on the analysis. Figure 4 shows the comparison of variations in the penetration depth with recently proposed pressure and with equivalent radius field expressions. The approximated pressure field overpredicts the penetration depth. The difference in the penetration depth with the proposed pressure field is significant, which shows that it is important to consider the proposed pressure field for a rectangular microchannel rather than an approximated pressure field. The transport with a capillary action is the balance among surface, viscous, and other body forces which retard the flow as it progresses. Hence, the capillary flow always attains a steady state which is generally termed as an equilibrium penetration depth in the literature. If the length of the channel is longer than that of the equilibrium penetration depth, then flow front cannot reach the outlet and, therefore, the assistance to the capillary flow is attempted in such cases. Passive or nonmechanical pumping approaches combined with the capillary flow serve this enhancement. The scaling analysis suggests that the gravity force is less dominant at microscale [12], but several researchers have demonstrated that gravity can be used as an assistance to the capillary flow [13–15]. Generally, the capillary flow analysis is performed with an assumption of infinite reservoir. Hence, the reservoir effect and the gravitational force from the reservoir are generally neglected in the analysis. To accommodate the entrance effect of finite size reservoir at the inlet of the microchannel in the theoretical modeling, the entrance pressure field is developed for the arrangements shown in Fig. 5. The rectangular microchannel with rectangular reservoir on the top of the microchannel is considered and the pressure field with the gravity and reservoir effect is developed in the flow [16]. Moreover, this pressure field is used to obtain the governing equation for capillary flow under the influence of gravity head from the reservoir.



Capillary Flow, Fig. 4 The comparison of variations in the penetration depth with equivalent radius and recently proposed pressure field expressions. Figure 4 shows the comparison of penetration depth for $\gamma = 0.9$ with the corresponding difference in the penetration depth [6]

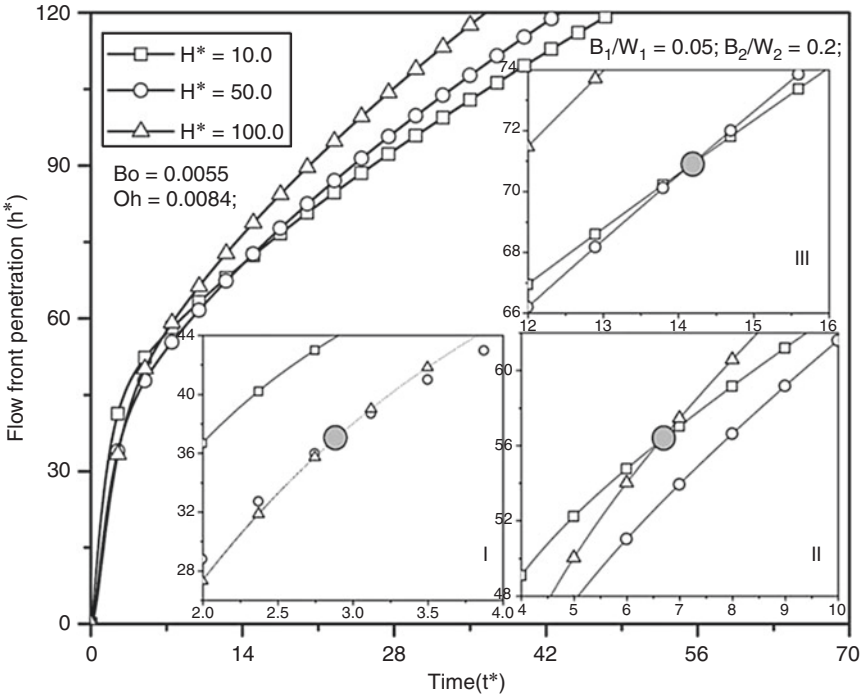


Capillary Flow, Fig. 5 Schematic of a gravity-assisted capillary flow in a vertically oriented capillary of width $2B_1$ and depth $2\ W_1$. The additional gravitational head from the fluid in a finite reservoir of size ($2B_2 \times 2\ W_2$) is assisting the capillary flow [16]

The reservoir with three different levels of fluid in the reservoirs (H^*), namely, 10, 50, and 100, is considered for the analysis. Figure 6 shows the variations in the penetration depth (h^*) with different operating conditions. This analysis represents the interplay between the surface tension force and gravity head from the reservoir. The capillary flow takes place in the channel which remains same for all three cases, whereas the level of the fluid from the reservoir increases from case I to case III. Thus, for three cases, the capillary effect is the same but the gravity head is different. At the beginning of the transport, the fluid from the reservoir offers less inertia to the fluid transport and the capillary force dominates over the gravity from the reservoir. Therefore, at the beginning of the transport, the penetration depth with a lower reservoir fluid level ($H^* = 10.0$) is higher than the other two penetration depths as shown in the inset I. Similarly, the penetration depth with the highest reservoir fluid level ($H^* = 100.0$) has the lowest penetration depth as compared to others.

Moreover, as the fluid progresses in the microchannel, the momentum from the reservoir fluid assists the capillary flow and the gravitational force, due to which the fluid from the reservoir becomes dominant over the capillary force within the microchannel. This results in transcendence among the penetration depths with a different gravity head. The penetration depth with the highest gravity head ($H^* = 100$) surpasses the penetration depth with gravity head $H^* = 50$ and $H = 10^*$ in inset I and II two, respectively. This can be attributed to as an interplay between the surface tension force, i.e., the capillarity and gravitational force from reservoir. In the case of microfluidic applications, the sizes of reservoir and microchannel are comparable to each other. Hence one cannot neglect the effect of the reservoir in such cases, particularly if it is surface tension–driven pumping. Further, one can assist the capillary flow with an appropriate arrangement of reservoir.

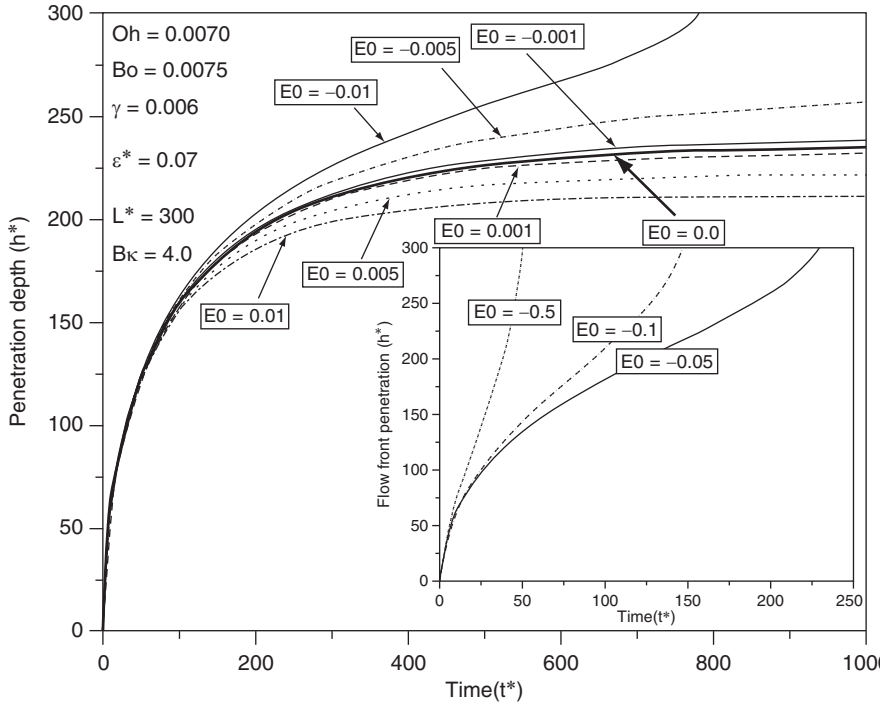
There are always certain limitations to the autonomous pumping which make them inadequate in long microchannels. Hence, it is important to enhance the pumping ability by other means. Further enhancement in the capillary flow can be achieved by coupling the capillary flow with the electroosmotic flow which is one of the electrokinetic pumping mechanisms. In most of the cases, the inner wall of a microchannel always has surface charges due to different



Capillary Flow, Fig. 6 Transient response of a flow front transport for different gravitational heads in the reservoir with Bo = 0.0055, Oh = 0.0084, $B_1/W_1 = 0.05$ $B_2/W_2 = 0.2$. (**I**) Flow front penetration rate for $H^* = 100$ surpasses the penetration rate for $H^* = 50$. (**II**) Flow front penetration rate for $H^* = 100$ surpasses the penetration rate for $H^* = 10$. (**III**) Flow front penetration rate for H = 50 surpasses the penetration rate for $H^* = 10$ [16]

mechanisms like ionization, dissociation of ions, isomorphic substitution, etc., [17]. These surface charges distribute ions of the electrolytes in a specific pattern when brought into contact with an electrolyte which is generally termed as the formation of electrical double layer (EDL). After applying the electric field across the channel, the movement of the ions takes place, which results in the movement of the fluid due to an electric field [18]. The electrolyte solution is transported with the capillary action; one can further assist the capillary flow. An additional body force due to electroosmotism is added to Eq. 2, which accommodates the additional effect of electroosmosis. Further one can analyze the interplay between the capillarity and electroosmotisms as presented in the recent studies [19]. Figure 7 shows the variation in the penetration depth of the capillary flow under the influence of electroosmotism. Through a nondimensional analysis, a new nondimensional number is proposed, i.e., Eo which represents the ratio between the surface tension force and electroosmotic force. The direction of the electroosmotic flow can be reversed by changing the electric field direction. Hence, negative and positive Eo numbers are observed in the analysis. The negative Eo numbers represent the change in the direction of the electric field as compared to positive Eo numbers. The pure capillary flow can be seen as Eo = 0.

The variation in the penetration depth under three different operating conditions is shown in Fig. 7. As observed in the pure capillary case (Eo = 0), the penetration depth attains the equilibrium penetration depth, whereas in the case of − Eo numbers, the equilibrium penetration depth increases with increment in the magnitude of − Eo numbers. This represents that in the case of electroosmotic flow with − Eo number, the capillary flow is assisted by electroosmotism. In this analysis, the nondimensional length of the



Capillary Flow, Fig. 7 Variation in the penetration depth for vertically oriented channel with water as electrolyte, where B = 100 μm, W = 400 μm, L = 75 mm, $\zeta = -75$ mV and constant contact angle is 27°. The inset shows the variation in the electric field within the electrolyte as the flow front progresses under different applied voltages [19]

microchannel (L^*) is considered as 300 and with –Eo = −0.01 the entire filling of the microchannel is observed. In the case of positive Eo numbers, it is observed that the electroosmotism acts in a opposite direction of the capillary flow. Hence, it retards the flow and this can be observed by the decrement in the equilibrium penetration depth with the increment in the positive Eo numbers.

For the enhancement in the capillary flow, a transport with additional gravity head and electroosmotic forces are considered. A generalized theoretical modeling for a gravity-assisted capillary flow with reservoir effects and electroosmotically assisted capillary flow is reported in brief. It is observed that even though the scaling among forces suggests that the gravitation force is negligible at microscale, the reported analysis infers that with a finite reservoir, an added advantage due to gravity can be a useful tool to transport the fluid at microscale. This added force for the capillary transport can be utilized without any additional burden in the design of the LOC device. The electroosmotically assisted capillary flow model suggests that in a combined flow the electrokinetic parameters have an important influence on the capillary flow. Such electrokinetic flow approaches can be coupled to enhance the capillary flow transport in the microchannel.

The wetting properties of the fluid decide the capability of pumping with a capillary flow. Therefore, it is important to know the precise magnitude of wetting properties like contact angle and surface tension of the working fluid. The microfluidics has become a promising option for biomedical application and inclusion of biomolecules is an unavoidable part in such applications. In most cases, the biomolecules are attached with the microbeads and transported to the desired locations. It is evident from the experimental analysis that the inclusion of microbeads changes the wetting behavior drastically [20]. Therefore, it is

necessary to consider the effect of microbeads in the fluid for the analysis. This can be done by considering the following expressions for surface tension and contact angle: density and viscosity which are functions of volume fraction of microbeads. Such correlations of the surface tension and contact angle are provided in [20]. Such expressions for the variation in the contact angle and surface tension with the volume fraction can be readily used in modeling transport processes of microbead suspensions in micro-capillaries, used in the microfluidic devices.

In passive pumping, particularly with the capillary flow, different aspects due to microscale effects like aspect ratio–dependent velocity profile, contact angle at four walls, fluid-air interface dynamics in the case of suspension flow, etc., need to be investigated in detail. Theoretically the concept of electroosmotically assisted capillary flow has been presented but the experimental demonstration of such phenomena is also an interesting area of research. Moreover, wetting of biomolecule suspensions under transient effects instead of the steady state is also needed to be studied. The experimental study of the flow behind the front and at the entrance of the microchannels is also an interesting study to perform.

Cross-References

- ▶ AC Electroosmosis: Basics and Lab-on-a-Chip Applications
- ▶ Electrowetting
- ▶ Micro/Nano Flow Characterization Techniques
- ▶ Micropumps
- ▶ Surface Tension Effects of Nanostructures
- ▶ Wetting Transitions

References

1. Saha, A., Mitra, S.K.: Numerical study of capillary flow in microchannels with alternate hydrophilic-hydrophobic bottom wall. J. Fluid Eng. Trans. ASME **131**, 061202 (2009)
2. Saha, A., Mitra, S.: Effect of dynamic contact angle in a volume of fluid (VOF) model for a microfluidic capillary flow. J. Colloid Interface Sci. **339**, 461–480 (2009)
3. Xiao, Y., Yang, F., Pitchumani, R.: A generalized flow analysis of capillary flows in channels. J. Colloid Interface Sci. **298**, 880–888 (2006)
4. Washburn, E.: The dynamics of capillary flow. Phys. Rev. **17**, 273 (1921)
5. Chakraborty, S.: Electroosmotically driven capillary transport of typical non-Newtonian biofluid in rectangular microchannels. Anal. Chim. Acta **605**, 175–184 (2007)
6. Waghmare, P.R., Mitra, S.K.: A comprehensive theoretical model of capillary transport in rectangular microchannels. Microfluid. Nanofluid. (2011). doi:10.1007/s10404-011-0848-8
7. Levin, S., Reed, P., Watson, J.: A theory of the rate of rise a liquid in a capillary. In: Kerker, M. (ed.) Colloid and Interface Science, p. 403. Academic, New York (1976)
8. Marwadi, A., Xiao, Y., Pitchumani, R.: Theoretical analysis of capillary-driven nanoparticulate slurry flow during a micromold filling process. Int. J. Multiph. Flow **34**, 227 (2008)
9. Dreyer, M., Delgado, A., Rath, H.: Fluid motion in capillary vanes under reduced gravity. Microgravity Sci. Technol. **4**, 203 (1993)
10. Bhattacharya, S., Gurung, D.: Derivation of governing equation describing time-dependent penetration length in channel flows driven by non-mechanical forces. Anal. Chim. Acta **666**, 51–54 (2010)
11. Keh, H., Tseng, H.: Transient electrokinetic flow in fine capillaries. J. Colloid Interface Sci. **242**, 450 (2001)
12. Nguyen, N., Werely, S.: Fundamentals and Applications of Microfluidics. Artech House, New York (2003)
13. Yamada, H., Yoshida, Y., Terada, N., Hagihara, T., Teasawa, A.: Fabrication of gravity-driven microfluidic device. Rev. Sci. Instrum. **79**, 124301 (2008)
14. Jong, W.R., Kuo, T.H., Ho, S.W., Chiu, H.H., Peng, S. H.: Flows in rectangular microchannels driven by capillary force and gravity. Int. Commun. Heat Mass Transf. **34**, 186–196 (2007)
15. Kung, C., Chui, C., Chen, C., Chang, C., Chu, C.: Blood flow driven by surface tension in a microchannel. Microfluid. Nanofluid. **6**, 693 (2009)
16. Waghmare, P.R., Mitra, S.K.: Finite reservoir effect on capillary flow of microbead suspension in rectangular microchannels. J. Colloid Interface Sci. **351**(2), 561–569 (2010)
17. Hunter, R.: Zeat Potential in Colloid Science, Principle and Applications, Principle and Applications. Academic, London (1981)
18. Israelachvili, J.N.: Intermolecular and Surface Forces. Academic, London (1998)
19. Waghmare, P.R., Mitra, S.K.: Modeling of combined electroosmotic and capillary flow in microchannels. Anal. Chim. Acta **663**, 117–126 (2010)
20. Waghmare, P.R., Mitra, S.K.: Contact angle hysteresis of microbead suspensions. Langmuir **26**, 17082–17089 (2010)

Capillary Origami

Supone Manakasettharn[1], J. Ashley Taylor[2] and Tom N. Krupenkin[2]
[1]National Nanotechnology Center (NANOTEC), National Science and Technology Development Agency (NSTDA), Pathum Thani, Thailand
[2]Department of Mechanical Engineering, The University of Wisconsin-Madison, Madison, WI, USA

Synonyms

Capillarity induced folding; Elasto-capillary folding; Surface tension–powered self-assembly

Definition

Capillary origami is the folding of an elastic planar structure into a three-dimensional (3D) structure by capillary action between a liquid droplet/bubble and a structure surface.

Why Capillary Origami?

The fabrication of 3D structures is one of the major challenges for micro- and nano-fabrication. Folding of an elastic planar structure after patterning and release is one technique to fabricate a 3D structure using self-assembly. The term *origami* is taken from the Japanese art of paper folding; while the actuation of the folding is accomplished by using capillary forces of a fluid droplet, hence, the technique has been termed *capillary origami*. The combination of the folding process with capillary forces has resulted in a new technique for micro- and nano-fabrication.

History

The term *capillary origami* was first introduced in 2007 by Charlotte Py et al. to describe the folding of a polydimethylsiloxane (PDMS) sheet into a 3D structure by using capillary forces created by a water droplet [1]. As early as 1993, Syms and Yeatman demonstrated that 3D structures could be fabricated by folding surfaces using capillary forces produced by molten solder [2]. Later Richard R. A. Syms introduced the term *surface tension-powered self-assembly* to describe the technique [3, 4]. Both of these techniques are quite similar in that 3D structures can be produced by folding elastic thin films. Both use capillary forces for self-assembly. In the first example, various liquids such as water are used, while for the second study, molten metals such as solder were used, which then solidified fixing the 3D microstructures.

Principles

At the macroscale level, the influence of capillary forces is negligible compared to other forces such as gravity, electrostatic, or magnetic. Because capillary forces scale linearly with the characteristic size of the system, at submillimeter dimensions, capillary forces begin to dominate since the majority of other forces decrease much more rapidly than the first power of the length. For example, a human cannot walk on water because capillary forces produced at the water surface are much smaller than the gravitational force acting on a human, which scales as the cube of the length. On the other hand, the much smaller water strider can easily walk on water because capillary forces are large enough to balance the gravitational force produced by the water strider. For capillary origami, capillary forces need to be large enough to counteract the weight of the liquid droplet and the structural forces of the planar layer.

In terms of energy, for capillary origami, one needs to consider the interplay of three different energies: capillary energy, bending energy, and gravitational potential energy. For a two-dimensional (2D) model, the capillary energy per unit length of the interface (2D analog of the surface energy) is defined as $E_c = L'\gamma$, where L' is the length of the interfacial surface of the fluid and γ is the surface tension [5]. The bending energy per unit length is approximately $E_b = \frac{LB}{2R^2}$, where

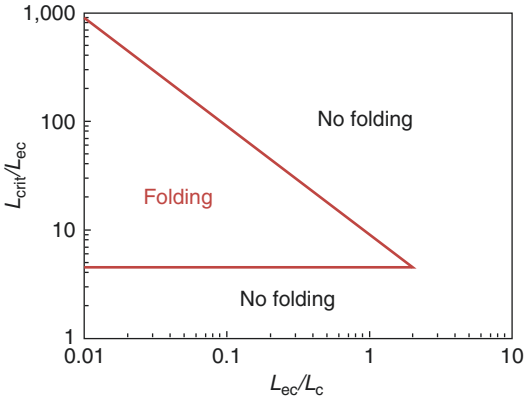
L is the length of the structure and R is the radius of curvature [6]. $B = \frac{Eh^3}{12(1-\nu^2)}$ is the bending rigidity of the structure, where E is Young's modulus, h is thickness of the layer, and ν is Poisson's ratio. If one only considers the mass of the fluid, assuming that it is much larger than the mass of the structure, then the gravitational potential energy per unit length is $E_g = \rho S g z$, where ρ is the density, S is the surface area, g is the constant of gravity, and z is the height of the center of mass. By neglecting the effect of gravity and assuming complete circular folding, [7] derived simplified criteria for folding considering the interplay between capillary and bending energies, which are expressed as

$$\sqrt{2\pi} < \frac{L_{\text{crit}}}{L_{\text{ec}}} < 2\sqrt{2\pi}\frac{L_{\text{c}}}{L_{\text{ec}}} \qquad (1)$$

where L_{crit} is the critical length of a structure for folding to occur, $L_{\text{c}} = \sqrt{\frac{\gamma}{\rho g}}$ is the capillary length [5], and $L_{\text{ec}} = \sqrt{\frac{B}{\gamma}}$ is the elasto-capillary length [8].

The simplified criteria for folding derived from Eq. 1 can be plotted as shown in Fig. 1. To fold a structure requires $\frac{L_{\text{ec}}}{L_{\text{c}}} < 2$ or $L_{\text{c}} > \frac{L_{\text{ec}}}{2}$ or $\gamma > \frac{\sqrt{B\rho g}}{2}$ indicating that the capillary length must be larger than half of the elasto-capillary length so that the capillary effect can overcome bending rigidity of the structure. The other requirement for folding is $\frac{L_{\text{crit}}}{L_{\text{ec}}} > \sqrt{2\pi} \cong 4.44$ or $L_{\text{crit}} > 4.44 L_{\text{ec}}$ confirming that the length of the structure should also be long enough for a liquid droplet to wet the surface to produce sufficient capillary forces to fold the structure. For the 3D structures, the folding criteria become more complex. In particular in 3D, the critical length also depends on the shape of the initial template such that $L_{\text{crit}} \cong 7L_{\text{ec}}$ for squares and $L_{\text{crit}} \cong 12L_{\text{ec}}$ for triangles [9]. Figure 2 shows examples of capillary origami structures of a pyramid, a cube, and a quasi-sphere obtained from folding triangle-, cross-, and flower-shaped PDMS sheets, respectively [10].

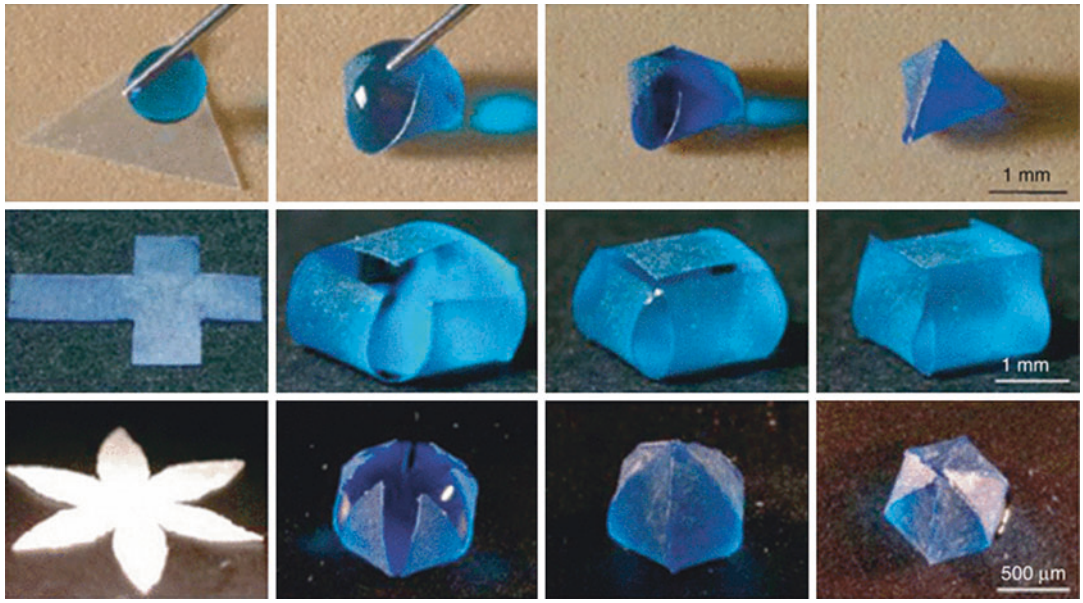
The concept of capillary origami has also been modeled at the nanoscale. Molecular dynamics simulations show the folding of planar graphene nanostructures by water nanodroplets [11]. Figure 3 summarizes the folding by presenting a phase-like diagram by plotting droplet radius (R_d) versus the ribbon width (w). From the graph, four different areas or phases can be identified corresponding to four different bending modes: nonfolding, sliding, zipping, and rolling. Figure 3 (left) shows the nonfolding phase where the nanoribbon end does not fold around the nanodroplet. The nonfolding phase is adjacent to the sliding phase where the nanoribbon end folds around the nanodroplet and then slides on the nanoribbon surface. The right top of the figure shows the zipping phase where the width of ribbon is several times greater than the droplet radius allowing the ribbon to fold around the droplet in the orthogonal direction. In the rolling phase ($w \leq \frac{1}{2}R_d$), the ribbon folds around the droplet and then continues wrapping in a scroll-like manner.



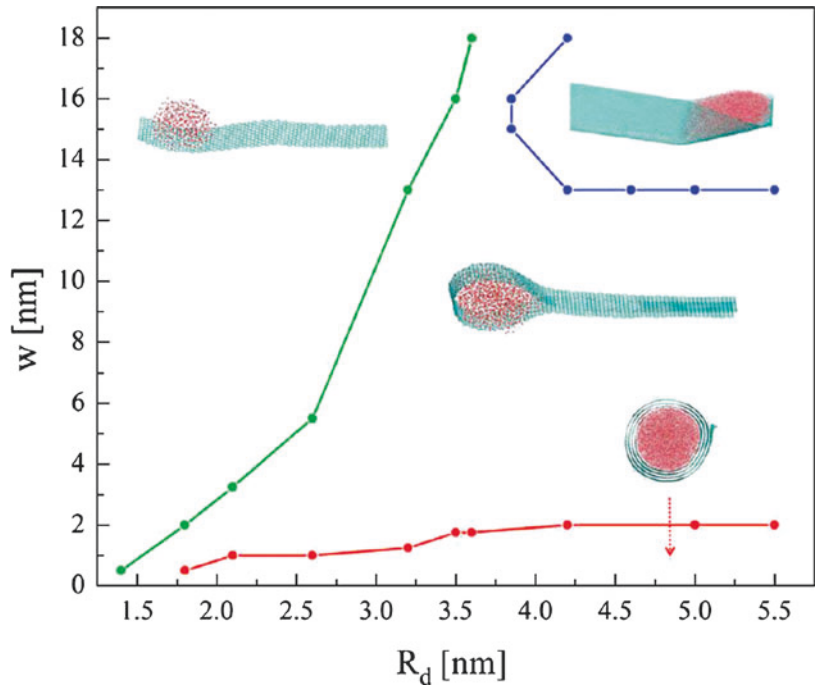
Capillary Origami, Fig. 1 Folding criteria plotted from Eq. 1 assuming complete circular folding and neglecting the effect of gravity

Applications

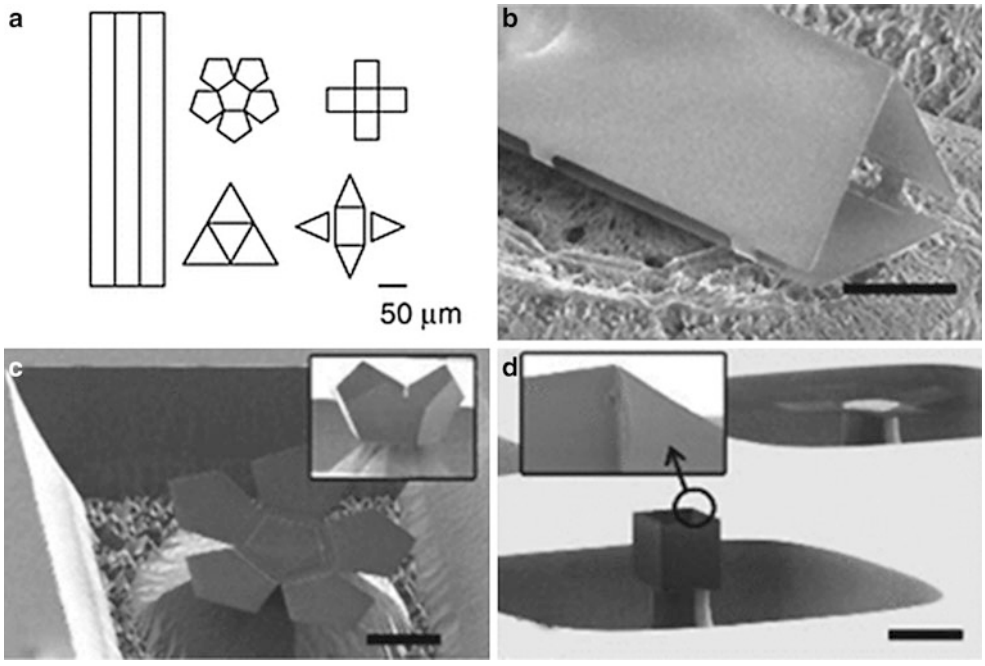
Capillary origami has been used to fabricate a number of 3D microstructures. Figure 4 illustrates the self-assembly of structures with various geometries. The initial planar templates are shown in Fig. 4a, and the folded final 3D microstructures are shown in Fig. 4b–d. The initial planar templates with lengths ranging from 50 to 100 μm and



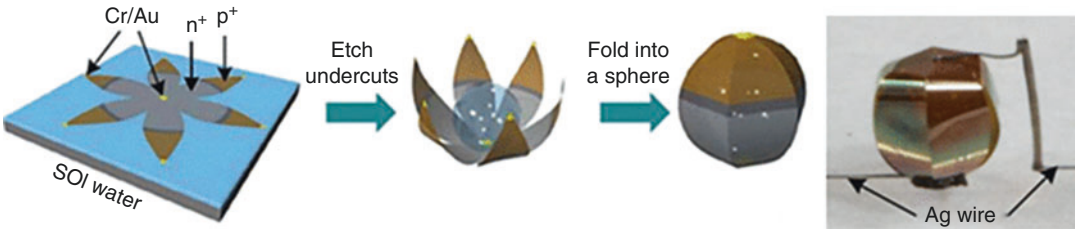
Capillary Origami, Fig. 2 Capillary origami 3D structures of a pyramid, a cube, and a quasi-sphere obtained by folding triangle-, cross-, and flower-shaped PDMS sheets, respectively, actuated with a water droplet (Reprinted with permission from Ref. [10]. Copyright 2007, American Institute of Physics)



Capillary Origami, Fig. 3 The phase diagram of a nanodroplet and graphene nanoribbon showing four different folding dynamics (Reprinted with permission from Ref. [11])



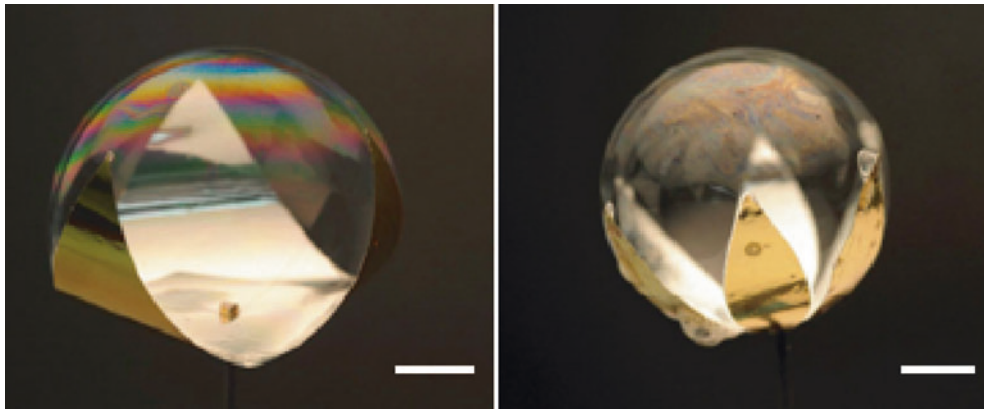
Capillary Origami, Fig. 4 (**a**) Schematics of initial templates. (**b**)–(**d**) SEM images of 3D microstructures after folding (scale bar: 50 µm) (Reprinted with permission from Ref. [12]. Copyright 2010, American Institute of Physics)



Capillary Origami, Fig. 5 Three schematics from left to right showing steps to fabricate a spherical-shaped silicon solar cell. The image at the far right shows the final spherical-shaped silicon solar cell (Images reprinted from Ref. [13] with permission)

a thickness of 1 µm were fabricated from silicon nitride thin films deposited and patterned by using standard micromachining processing typically used for integrated circuit and MEMS fabrication. Water droplets then were deposited on the templates to fold 3D microstructures [12]. Figure 5 shows another example of microfabrication of a quasi-spherical silicon solar cell based on capillary origami. After fabrication by conventional micromachining processing, the initial flower-shaped silicon template was folded into a sphere using a water droplet. Unlike conventional flat solar cells, this spherical solar cell enhanced light trapping and served as a passive tracking optical device, absorbing light from a wide range of incident angles [13].

Besides using liquid droplets, capillary origami structures can be constructed by using soap



Capillary Origami, Fig. 6 Capillary origami 3D structures formed from triangle- and flower-shaped templates using soap bubbles (scale bar: 2 cm) (Images reprinted from Ref. [14] with permission)



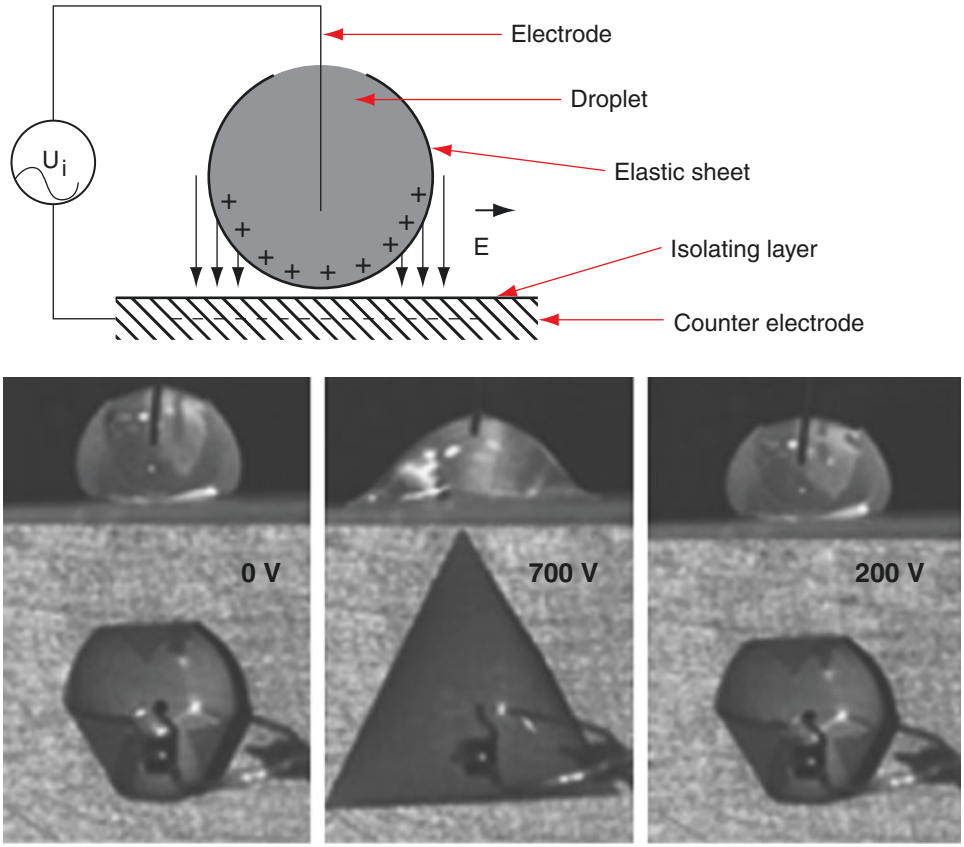
Capillary Origami, Fig. 7 The folding of an artificial flower when submerged in water (Reprinted with permission from Ref. [15]. Copyright 2009, American Institute of Physics)

bubbles as shown in Fig. 6. The weight of a soap bubble is much less than that of a liquid droplet especially for large droplets capable of covering centimeter-size structures when gravitational forces become significant. A soap bubble was shown to fold a centimeter-size elastic structure, which cannot be accomplished using a liquid droplet [14].

Petals of a flower also can be folded into a structure similar to capillary origami when submerged in water as shown in Fig. 7. The folding of the flower in water is accomplished by the interplay of elastic, capillary, and hydrostatic forces. During submersion, hydrostatic pressure pushes against the back of petals, and surface tension prevents water from penetrating through the spacing between petals resulting in trapping an air bubble inside a flower. The inside of the folded flower remains dry protected by the air bubble [15].

Structures formed by capillary origami also can be actuated by using electrostatic fields to reversibly fold and unfold then. For this application, we need to take into account the interplay of capillary, elastic, and electrostatic forces. As shown in Fig. 8, an electric field was applied between the droplet and the substrate. When the voltage was increased, the electrostatic force increased eventually overcoming capillary forces resulting in unfolding of the PDMS sheet. When the voltage was decreased below a certain threshold, the electrostatic force was no longer strong enough to prevent capillary forces from again folding the elastic sheet [16].

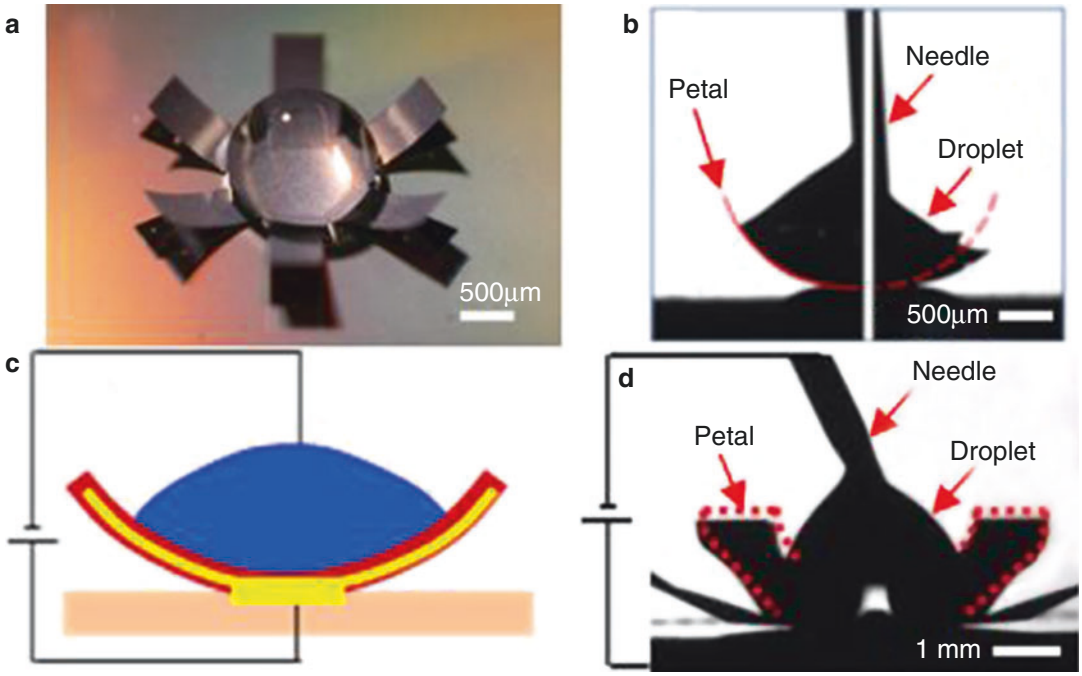
For flower-shaped polycrystalline silicon microstructures (called microflowers), the degree



Capillary Origami, Fig. 8 Capillary origami controlled by an electric field. The schematic of the experimental setup is shown at the far left, and three images to the right show the results of increasing voltage from 0 V to 700 V and decreasing voltage from 700 V to 200 V (Images from Ref. [16] – reproduced by permission of The Royal Society of Chemistry)

of folding was dynamically controlled by two different methods as shown in Fig. 9 [17]. The change in the bending angle was achieved by changing the volume of a liquid droplet by using an automated syringe pump. The petals reversibly bent and relaxed as the liquid was added and withdrawn. The electrowetting process (Fig. 9c) was also used to control bending. Capillary forces exerted by the liquid droplet on the petals can be changed by changing its contact angle by applying a voltage to the conducting liquid droplet placed on the petals. A thin dielectric layer of SiO_2 was thermally grown on the conductive Si petals to fabricate the electrowetting device. Both methods were used to reversibly actuate the petals without completely wrapping around the droplet, leaving part of the petal surface free to reflect incident light. The direction of the incoming light can then be dynamically controlled by the bending of the petal demonstrating an example of an optofluidic device.

Capillary origami is a simple and inexpensive method to fabricate 3D structures at the submillimeter scale. By using capillary forces, intricate and delicate 3D thin-film structures can easily be fabricated, which would be difficult to obtain by other methods. More applications exploiting the advantages of capillary origami itself or in combination with electric fields can readily be envisioned. Ultimately one expects to see more commercial products based on this versatile technique.



Capillary Origami, Fig. 9 (**a**) A microflower with a captured microdroplet. (**b**) Dependence of petal angle on liquid volume (only half of a frame is shown to facilitate comparison). As the amount of liquid captured by a microflower decreases, the petal angle also decreases. (**c**) Schematics of a process showing a droplet on the dielectric layer surrounding the conductive polycrystalline silicon petal of the microflower. (**d**) Electrowetting actuation of a microflower. The image shows petal position without applied voltage, and the red dash lines indicate petal positions with the voltage applied (Reprinted with permission from Ref. [17]. Copyright 2011, American Institute of Physics)

Cross-References

- ▶ Electrowetting
- ▶ Self-Assembly for Heterogeneous Integration of Microsystems
- ▶ Surface Tension Effects of Nanostructures

References

1. Py, C., Reverdy, P., Doppler, L., Bico, J., Roman, B., Baroud, C.N.: Capillary origami: spontaneous wrapping of a droplet with an elastic sheet. Phys. Rev. Lett. **98**, 156103 (2007)
2. Syms, R.R.A., Yeatman, E.M.: Self-assembly of three-dimensional microstructures using rotation by surface tension forces. Electron. Lett. **29**, 662–664 (1993)
3. Syms, R.R.A.: Surface tension powered self-assembly of 3-D micro-optomechanical structures. J. Microelectromech. Syst. **8**, 448–455 (1999)
4. Syms, R.R.A., Yeatman, E.M., Bright, V.M., Whitesides, G.M.: Surface tension-powered self-assembly of microstructures – the state-of-the-art. J. Microelectromech. Syst. **12**, 387–417 (2003)
5. Berthier, J.: Microdrops and Digital Microfluids. William Andrew Pub, New York (2008)
6. Timoshenko, S., Woinowsky-Krieger, S.: Theory of Plates and Shells, 2nd edn. McGraw-Hill, New York (1959)
7. de Langre, E., Baroud, C.N., Reverdy, P.: Energy criteria for elasto-capillary wrapping. J. Fluids Struct. **26**, 205–217 (2010)
8. Bico, J., Roman, B., Moulin, L., Boudaoud, A.: Adhesion: elastocapillary coalescence in wet hair. Nature **432**, 690 (2004)
9. Py, C., Reverdy, P., Doppler, L., Bico, J., Roman, B., Baroud, C.N.: Capillarity induced folding of elastic sheets. Eur. Phys. J. Spec. Top. **166**, 67–71 (2009)
10. Py, C., Reverdy, P., Doppler, L., Bico, J., Roman, B., Baroud, C.: Capillary origami. Phys. Fluids **19**, 091104 (2007)
11. Patra, N., Wang, B., Kral, P.: Nanodroplet activated and guided folding of graphene nanostructures. Nano Lett. **9**, 3766–3771 (2009)
12. van Honschoten, J.W., Berenschot, J.W., Ondarcuhu, T., Sanders, R.G.P., Sundaram, J., Elwenspoek, M., Tas, N.R.: Elastocapillary fabrication of

three-dimensional microstructures. Appl. Phys. Lett. **97**, 014103 (2010)
13. Guo, X., Li, H., Yeop Ahn, B., Duoss, E.B., Jimmy Hsia, K., Lewis, J.A., Nuzzo, R.G.: Two- and three-dimensional folding of thin film single-crystalline silicon for photovoltaic power applications. Proc. Natl. Acad. Sci. U. S. A. **106**, 20149–20154 (2009)
14. Roman, J., Bico, J.: Elasto-capillarity: deforming an elastic structure with a liquid droplet. J. Phys. Condens. Matter **22**, 493101 (2010)
15. Jung, S., Reis, P.M., James, J., Clanet, C., Bush, J.W. M.: Capillary origami in nature. Phys. Fluids **21**, 091110 (2009)
16. Pineirua, M., Bico, J., Roman, B.: Capillary origami controlled by an electric field. Soft Matter **6**, 4491–4496 (2010)
17. Manakasettharn, S., Taylor, J.A., Krupenkin, T.N.: Bio-inspired artificial iridophores based on capillary origami: fabrication and device characterization. Appl. Phys. Lett. **99**, 144102 (2011)

Carbon MEMS

Rodrigo Martinez-Duarte, Monsur Islam and Rucha Natu
Multiscale Manufacturing Laboratory, Department of Mechanical Engineering, Clemson University, Clemson, SC, USA

Synonyms

C-MEMS; Carbon microelectromechanical systems

Definition

A process used to fabricate miniaturized glass-like carbon structures through the carbonization of previously shaped organic precursors.

Overview

Carbon MEMS combines different patterning techniques with carbonization to derive miniaturized glass-like carbon features from an organic precursor. Since glass-like carbon is a brittle material, the carbonization of an already patterned precursor compares advantageously against subtractive manufacturing of glass-like carbon bulk. To date, the most common precursor patterning technique used in C-MEMS has been conventional photolithography, but many other techniques could be used. These cover a wide range: from relatively inexpensive like electrospinning, stamping, molding, polymer machining, and casting to next-generation lithography (NGL) techniques such as nanoimprint lithography (NIL) and electron beam lithography. The choice of precursor patterning technique is dictated by the quality, complexity, and final dimensions of the desired carbon part. Shrinkage of the precursor structure occurs during carbonization and must be accounted for. However, the existence of commercial high-quality precursors, like the negative photoresist SU-8, and availability of standardized fabrication tools make the fabrication process and the dimensional control highly reproducible.

Historical Background

Carbon can be found as a number of allotropes including diamond, lonsdaleite, buckminsterfullerenes, graphene, carbyne, graphite, carbon nanofoams, diamond-like carbon, amorphous carbon, and those carbons derived from the pyrolysis of organic materials, better known as glass-like carbons. Different precursors to obtain glass-like carbon may be used, including phenolic resins, polyfurfuryl alcohols, cellulose, polyvinyl chloride, and polyimides.

The first documented modern derivation of glass-like carbon from an organic resin (phenolic in this case) appears to be from 1915, when Weintraub and Miller in Massachusetts, USA, derived disks of a "very bright, shiny looking carbon with hardness equal or greater than 6 on the Mohs mineral scale." These disks were then used for better microphones in telephone transmitters. Their fabrication protocol featured slow heating of a hardened resin to a temperature close to 700 °C in about 1 week, followed by firing at temperatures from 800 to 1100 °C in just a few hours. The advantage of patterning an organic precursor, rather than carbon bulk, was realized

since then [1]. A sustained flow of publications on glass-like carbon did not begin until 1962, when Davidson, at the General Electric Co. in Kent, England, derived glass-like carbon from cellulose [2], and Yamada and Sato, at the Tokai Electrode Manufacturing Co. in Nagoya, Japan, published preliminary characterization results of carbon derived from organic polymers [3]. They named such carbon "glassy carbon," a term that was later registered as a trademark. Around the same time, in 1963, Lewis, Redfern, and Cowlard at the Plessey Company in the UK postulated the use of glass-like carbon, named "vitreous carbon" by the authors and later registered as trademark as well, as an ideal crucible material for semiconductors. Later that year, Redfern and Greens disclosed several production processes to derive this "vitreous carbon" in a patent [4]. In 1965, the advantages of glass-like carbon electrodes for voltammetry and analytical chemistry were characterized by Zittel and Miller, from Oak Ridge National Laboratory in the USA, using "glassy carbon" from the Tokai Electrode Manufacturing Co. [5]. In 1967, Cowlard and Lewis published a detailed description of the properties of "vitreous carbon," the fabrication process and its potential applications [6]. The decade of 1970 brought a significant interest on the use of glassy carbon as a material for different implants and biomedical instrumentation and also witnessed an explosion of the interest on glassy carbon by the analytical and electrochemistry communities, which still remains strong. In 1971, a structural model for glass-like carbon was postulated by Jenkins and Kawamura [7]. This model is up to this date the only one capable of explaining most of the experimental results obtained with glass-like carbon.

Carbon derived from organic polymers by pyrolysis in inert atmosphere has been historically known by three different names: "vitreous carbon," "glassy carbon," and glass-like carbon. Although highly referenced in implant-related publications during the 1970s, the term "vitreous carbon" started to fall in disuse by the end of that decade. "Vitreous carbon" is now better identified with reticulated vitreous carbon (RVC), a material introduced in the late 1970s by Chemotronics International Inc. from Ann Arbor, Michigan. The commercialization of Tokai's "glassy carbon" electrodes targeting the electrochemistry market made "glassy carbon" the term of preference for the electroanalytical chemistry community to refer to glass-like carbon. In 1995, the IUPAC (International Union of Pure and Applied Chemistry) defined glass-like carbon as the material derived by the pyrolysis of organic polymers and recommended that the terms "glassy carbon" and "vitreous carbon," which had been introduced as trademarks, should not be used as synonymous for glass-like carbon.

From the microfabrication standpoint, in 1983 Lyons et al. at AT&T Bell Laboratories published their work on the use of photodefined novolac resist patterns as precursors for carbon microstructures [8]. The drive behind this effort was finding an alternative to carbon films deposited by chemical vapor deposition. Miniaturized glass-like carbon 3D structures were not reported until the late 1990s by Schueller and coworkers at Harvard University. In their process, polydimethylsiloxane (PDMS) molds were fabricated using soft lithography and then used to pattern furfuryl alcohol-modified phenolic resins and phenol-formaldehyde resins, which were subsequently carbonized [9]. By 2000, Kostecki and colleagues were obtaining further results on the fabrication of planar carbon microelectrodes from positive photoresists to study the influence of their geometry in their electrochemical response [10]. Such work ignited the use of pyrolyzed photoresist films, or PPF, in fields such as electrochemistry. In 2002, the derivation of carbon from negative photoresists was reported by Singh et al. [11]. The obtained carbon showed higher electrical resistivity and vertical shrinkage than the one synthesized from positive resists. In 2004, structures with aspect ratios higher than 10 were reported by Wang et al. at the University of California, Irvine (UCI) [12, 13]. These authors coined the term Carbon MEMS. Since then, most of the work identified as Carbon MEMS has used SU-8 as carbon precursors and photolithography to shape 3D structures. Other precursors include resorcinol–formaldehyde gels [14] and cellulose. Starting in 2009, electrospinning of SU-8 and PAN (polyacrylonitrile) has been developed toward obtaining carbon fibers in the nanoscale [15, 16].

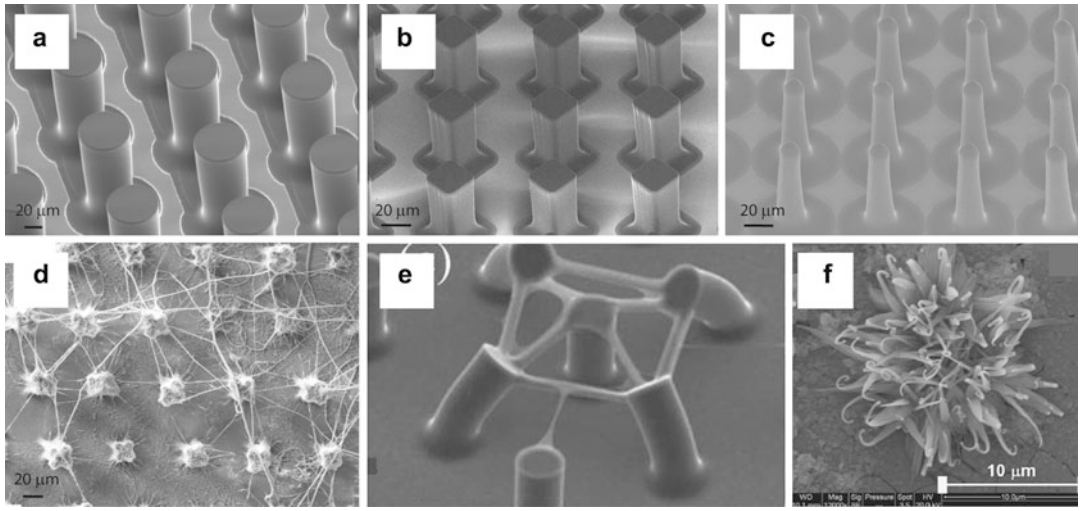
Material Properties of Carbon MEMS

The carbon obtained in a Carbon MEMS process resembles glass-like carbon. Although the choice of precursor and patterning technique has an impact on the properties of the resultant carbon, the following properties can serve as a starting reference. Although a consensus on the crystalline structure of glass-like carbon has not been reached to this date, the most widely accepted model [7] is the one that considers this type of carbon as made up of tangled and wrinkled aromatic ribbon molecules that are randomly cross-linked by carbon–carbon covalent bonds. The ribbon molecules form a networked structure, the unit of which is a stack of high strained aromatic ribbon molecules. Such structure of crystallites reflects the features of thermosetting resins structure which are commonly used as precursors for glass-like carbons. Glass-like carbon is widely considered as impermeable to gases and extremely inert, with a remarkable resistance to chemical attack from strong acids and other corrosive agents such as bromine. The material will react with oxygen at temperatures above few hundred degrees centigrade, making oxygen plasma a favored etching process. Glass-like carbon has a hardness of 6–7 on Mohs scale, a value comparable to that of quartz. Its density ranges from 1.4 to about 1.5 g cm^{-3}, compared to 2.3 g cm^{-3} for graphite. X-ray diffraction studies have shown that glass-like carbon presents an extremely small pore size, ~50 Å, of a closed nature, and that has an amorphous structure [31–38]. Glass-like carbon features a coefficient of thermal expansion of 2.2–3.2 × 10^{-6} K^{-1} which is similar to some borosilicate glasses. Its Young's modulus can vary between 10 and 40 GPa. An indicative electrical resistivity of glass-like carbon is 1 × 10^{-4} Ω m when the carbonization temperature is as high as 900 °C. The electrical resistivity drops further as the temperature increases over this value or the dwell time at temperatures below it increase [17, 18]. One of the reasons glass-like carbon is preferred by electrochemists is the fact that it has a wider electrochemical stability window than platinum and gold.

Photolithography as Toolbox for Carbon MEMS

The embracing of photolithography to fabricate the precursor photoresist structure has brought significant advantages and enabled rapid and notable developments. Photolithography refers to patterning with light, since the chemical composition of the photoresist changes upon being exposed to light of specific wavelengths, usually 365 nm for SU-8. In SU-8, light starts a cross-linking reaction in the matrix that makes the exposed section less soluble in a developer. The photolithography mask, a patterned stencil that enables the selective pass of light, must then be designed accordingly to the desired final topography: transparent areas in the mask must correspond to the topography to be fabricated. Traditional photolithography of a negative-tone photoresist involves a set of basic processing steps: spin coating a photoresist on a substrate, soft bake to evaporate the casting solvent, exposure through the mask to initiate cross-linking, postexposure bake to finalize cross-linking, and developing to obtain the exposed topography. Optimization of specific steps in traditional photolithography protocols has enabled the fabrication of a number of structures detailed below. The use of electrospinning has also emerged as an alternative to spin coating to deposit nanoscaled polymer fibers on a substrate, either as bulk or as single fibers (Fig. 1).

The fabrication of high-aspect ratio carbon structures, like those shown in Fig. 1a, b, was the first milestone of Carbon MEMS technology. SU-8 was an ideal choice of precursor to achieve this, given its low absorbance of light at wavelengths above 360 nm and the possibility to implement very thick layers in a single spin-coating step. Multistep photolithography has been used to fabricate dense arrays of carbon electrodes of different shapes to be used for capacitors [20, 21], heaters [22], glucose sensors [23], batteries [24–27], bioparticle manipulation using electrokinetics [28–36], and cell scaffolding [37]. By varying the exposure setup and dose, wires and suspended plates can also



Carbon MEMS, Fig. 1 (**a**) High-aspect ratio carbon electrodes on planar connecting leads. SU-8 precursor structure was fabricated using two-layer photolithography. (**b**) Further example of high-aspect ratio carbon structures. (**c**) Tapered carbon structures fabricated using backexposure during photolithography. (**d**) Example of SU-8 wires suspended between carbon posts. The fibers were made with electrospinning and later carbonized (Reprinted from [16] with permission from Elsevier). (**e**) Wire arrays made using electron beam lithography. Reprinted from [19] with permission from Elsevier. (**f**) Example of a fractal structure after carbonizing a sol–gel polycondensation of resorcinol and formaldehyde (Reprinted from [14], with permission from Elsevier)

be manufactured. The implementation of a photolithography process on a releasable substrate or on a sacrificial layer leads to free-standing SU-8 structures which were used as micromolds after carbonization [38]. Only the carbonization of free-standing SU-8 structures leads to a true isometric shrinkage. Carbonization of precursor structures anchored to a substrate yields carbon structures with a slight distortion on their base. UV backexposure of an SU-8 layer through a transparent substrate, fused silica, for example, leads to the development of tapered structures resembling cones (Fig. 1). Exposure dose can be tailored to specify the taper of the angle [39]. SU-8 wires suspended between two posts were made in 2006 using electron beam lithography. After carbonization, a number of carbon wire arrays were obtained (Fig. 1e) [19]. Suspended carbon wires like those shown in Fig. 1d have also been made using electrospinning in recent years [16, 40–42]. Electrospinning was also used to deposit a fiber mat on a substrate which was then patterned using photolithography. The fibrous structures were then carbonized to obtain highly porous carbon structures [15].

Other methods to shape the precursor exist. For example, in 2009, carbon fractals and microspheres were obtained by pyrolyzing a sol–gel polycondensation of resorcinol with formaldehyde [14]. The amount of surfactant during sol–gel synthesis determined the size of the spheres and the eventual switch to fabricating fractal structures (Fig. 1f). Micromolding of parylene has also been used to fabricate high-aspect ratio carbon posts [43].

Shrinkage

Carbonization is the process by which solid residues with a high content of carbon are obtained from organic materials, usually by pyrolyzing them in an inert atmosphere. The degree of shrinkage and carbon yield, the ratio of the weight of carbon to the weight of the original polymer sample, varies depending on the choice

of carbon precursor, its degree of cross-linkage, and its shape. For example, SU-8 features a carbon yield of 35–60 % [39]. The variability seems to depend on the shape of the precursor micro- or nanostructure.

In the case of Carbon MEMS, carbonization usually takes place in a furnace under a nitrogen, vacuum of forming gas atmosphere under a flow around 2000 ml/min. The carbonization protocol features three stages: (1) a temperature ramp from room temperature to 200–300 °C at 10 °C/min, followed by a 30 min dwell (this step is to completely eliminate the solvent and allow for any residual oxygen to be evacuated from the chamber and prevent combustion of the polymer as the temperature is raised further), (2) a temperature ramp to 900–1000 °C at 10 °C/min with a 1–4 h dwell, and (3) a natural cooldown to room temperature by turning the furnace off. Carbonization is a complex process with many reactions taking place concurrently, including dehydrogenation, condensation, hydrogen transfer, and isomerization. The pyrolysis process of organic compounds can be divided into three major steps: (1) pre-carbonization, (2) carbonization, and (3) annealing. During pre-carbonization ($T < 300$ °C), molecules of solvent and unreacted monomer are eliminated from the polymeric precursor. The next step, carbonization, can be further divided into two substages: (a) from 300 to 500 °C, when heteroatoms such as oxygen and halogens are eliminated causing a rapid loss of mass while a network of conjugated carbon systems is formed and hydrogen atoms start being eliminated, and (b) from 500 to 1200 °C, where hydrogen, oxygen, and nitrogen atoms are almost completely eliminated and the aromatic network is forced to become interconnected. At this point, permeability decreases and density, hardness, Young's modulus, and electrical conductivity increase. The carbon content of the structures carbonized at 900 °C is expected to exceed a mass fraction of 90 % in weight. At $T \sim 1300$ °C, more than 99 % carbon can be found. Annealing is usually carried out at temperatures above 1200 °C to allow for the gradual elimination of any structural defects and evolution of any further impurities.

Applications

Several applications have been demonstrated using Carbon MEMS. Some examples include the development of 3D architectures for lithium-ion batteries [24, 25], carbon plates in PEM fuel cells derived from machined polyimide [44, 45], post arrays which were electrochemically activated or decorated with carbon nanotubes to achieve supercapacitors [20, 21], and the development of carbon microspheres and fractal-like structures for sensors and batteries [14]. Other sensor applications include post arrays for rapid quantification of glucose in low concentrations [23] and interdigitated arrays with very narrow gaps to implement redox-based detection of dopamine with an amplification factor up to 25 [46]. C-MEMS devices have also been used as substratum for cell growth [47] and as carbon scaffolds to induce stem cell differentiation [37]. Extended arrays of carbon posts have been integrated in flow-through microfluidic devices to implement high-throughput cell separation and manipulation using electric fields in a technique now known as carbon-electrode dielectrophoresis [28, 29, 48]. This technique has been used to purify viable bacteria from an antibiotic-treated sample [31], to implement a sample preparation step that increases the sensitivity of traditional PCR protocols [34], and to extract DNA from a sample [32]. The combination between centrifugal microfluidics and carbon-electrode DEP marked an important step toward a sample-to-answer diagnostic platform [30]. Particle transport using electric fields has also been implemented [36], as well as high-throughput electrical lysis of different cells [33]. Other applications include the use of Carbon MEMS techniques to fabricate robust, inexpensive carbon shapes for the micromolding of bulk metallic glasses [38, 49].

Future Directions of the Field

Although most of the Carbon MEMS work has been done using SU-8 photoresist and photolithography, there is a strong interest on using different precursors, including biopolymers [50].

Development of techniques for shaping these precursors in the micro- and nanoscale will be needed, i.e., extrusion-based additive manufacturing. Electrospinning has been gaining importance as a technique to fabricate precursor nanostructures and is expected to be developed further. A broader use of composites is also expected: the patterning of SU-8-CNTs composites has been published [21] and preliminary work has also been disclosed using SU-8- silver and SU-8-silica composites. The use of catalysts to lower the energy required for carbonization is also of high interest, as well as understanding the shrinking process.

References

1. Weintraub, E., Miller, L.B.: Microphone. US Patent 1,156,509 (1915)
2. Davidson, H.W.: The properties of G.E.C. Impermeable carbon. Nucl. Eng. **7**, 159–161 (1962)
3. Yamada, S., Sato, H.: Some physical properties of glassy carbon. Nature **193**, 261–262 (1962)
4. Redfern, B., Greens, N.: Bodies and shapes of carbonaceous materials and processes for their production. US Patent 3,109,712 (1963)
5. Zittel, H.E., Miller, F.J.: A glassy-carbon electrode for voltammetry. Anal. Chem. **37**, 200–203 (1965)
6. Cowlard, F.C., Lewis, J.C.: Vitreous carbon – a new form of carbon. J. Mater. Sci. **2**, 507–512 (1967)
7. Jenkins, G., Kawamura, K.: Structure of glassy carbon. Nature **231**, 175–176 (1971)
8. Lyons, A., Wilkins, C., Robbins, M.: Thin pinhole-free carbon films. Thin Solid Films **103**, 333–341 (1983)
9. Schueller, O.J.A., Brittain, S.T., Whitesides, G.M.: Fabrication of glassy carbon microstructures by pyrolysis of microfabricated polymeric precursors. Adv. Mater. **9**, 477–480 (1997)
10. Kostecki, R., Song, X.Y., Kinoshita, K.: Influence of geometry on the electrochemical response of carbon interdigitated microelectrodes. J. Electrochem. Soc. **147**, 1878–1881 (2000)
11. Singh, A., Jayaram, J., Madou, M., Akbar, S.: Pyrolysis of negative photoresists to fabricate carbon structures for microelectromechanical systems and electrochemical applications. J. Electrochem. Soc. **149**, E78–E83 (2002)
12. Wang, C., Taherabadi, L., Jia, G., Madou, M., Yeh, Y., Dunn, B.: C-MEMS for the manufacture of 3D microbatteries. Electrochem. Solid-State Lett. **7**, A435–A438 (2004)
13. Wang, C., Jia, G., Taherabadi, L.H., Madou, M.J.: A novel method for the fabrication of high-aspect ratio C-MEMS structures. J. Microelectromechan. Syst. **14**, 348–358 (2005)
14. Sharma, C.S., Kulkarni, M.M., Sharma, A., Madou, M.: Synthesis of carbon xerogel particles and fractal-like structures. Chem. Eng. Sci. **64**, 1536–1543 (2009)
15. Sharma, C.S., Sharma, A., Madou, M.: Multiscale carbon structures fabricated by direct micropatterning of electrospun mats of SU-8 photoresist nanofibers. Langmuir **26**, 2218–2222 (2010)
16. Sharma, C.S., Katepalli, H., Sharma, A., Madou, M.: Fabrication and electrical conductivity of suspended carbon nanofiber arrays. Carbon N. Y. **49**, 1727–1732 (2011)
17. Park, B.Y., Taherabadi, L., Wang, C., Zoval, J., Madou, M.J.: Electrical properties and shrinkage of carbonized photoresist films and the implications for carbon microelectromechanical systems devices in conductive media. J. Electrochem. Soc. **152**, J136–J143 (2005)
18. Mardegan, A., Kamath, R., Sharma, S., Scopece, P., Ugo, P., Madou, M.: Optimization of carbon electrodes derived from epoxy-based photoresist. J. Electrochem. Soc **160**, B132–B137 (2013)
19. Malladi, K., Wang, C., Madou, M.: Fabrication of suspended carbon microstructures by e-beam writer and pyrolysis. Carbon N. Y. **44**, 2602–2607 (2006)
20. Beidaghi, M., Chen, W., Wang, C.: Electrochemically activated carbon micro-electrode arrays for electrochemical micro-capacitors. J. Power Sources **196**, 2403–2409 (2011)
21. Chen, W., Beidaghi, M., Penmatsa, V., Bechtold, K., Kumari, L., Li, W.Z., Wang, C.: Integration of carbon nanotubes to C-MEMS for on-chip supercapacitors. IEEE Trans. Nanotechnol. **9**, 734–740 (2010)
22. Jeong, O.C., Konishi, S.: Three-dimensionally combined carbonized polymer sensor and heater. Micromechanics Sect. Sensors Actuators (SAMM), based Contrib. Revis. from Tech. Dig. IEEE 20th Int. Conf. Micro Electro Mech. Syst. (MEMS 2007) - MEMS 2007, IEEE 20th Int. Conf. M. 143, 97–105 (2008).
23. Xu, H., Malladi, K., Wang, C., Kulinsky, L., Song, M., Madou, M.: Carbon post-microarrays for glucose sensors. Biosens. Bioelectron. **23**, 1637–1644 (2008)
24. Min, H.-S., Park, B.Y., Taherabadi, L., Wang, C., Yeh, Y., Zaouk, R., Madou, M.J., Dunn, B.: Fabrication and properties of a carbon/polypyrrole three-dimensional microbattery. J. Power Sources **178**, 795–800 (2008)
25. Teixidor, G.T., Zaouk, R.B., Park, B.Y., Madou, M.J.: Fabrication and characterization of three-dimensional carbon electrodes for lithium-ion batteries. J. Power Sources **183**, 730–740 (2008)
26. Wang, C., Madou, M.: From MEMS to NEMS with carbon. Biosens. Bioelectron. **20**, 2181–2187 (2005)
27. Wang, C., Taherabadi, L., Jia, G., Kassegne, S., Zoval, J., Madou, M.: Carbon-MEMS architectures for 3D microbatteries. Proc. SPIE **5455**, 295–302 (2004)

28. Martinez-Duarte, R., Renaud, P., Madou, M.: A novel approach to dielectrophoresis using carbon electrodes. Electrophoresis **32**, 2385–2392 (2011)
29. Jaramillo, M.D.C., Torrents, E., Martinez-Duarte, R., Madou, M.J., Juarez, A.: On-line separation of bacterial cells by carbon-electrode dielectrophoresis. Electrophoresis **31**, 2921–2928 (2010)
30. Martinez-Duarte, R., Gorkin, R.A., Abi-Samra, K., Madou, M.J.: The integration of 3D carbon-electrode dielectrophoresis on a CD-like centrifugal microfluidic platform. Lab Chip **10**, 1030–1043 (2010)
31. Elitas, M., Martinez-Duarte, R., Dhar, N., McKinney, J.D., Renaud, P.: Dielectrophoresis-based purification of antibiotic-treated bacterial subpopulations. Lab Chip **14**, 1850–1857 (2014)
32. Martinez-Duarte, R., Camacho-Alanis, F., Renaud, P., Ros, A.: Dielectrophoresis of lambda-DNA using 3D carbon electrodes. Electrophoresis **34**, 1113–1122 (2013)
33. Mernier, G., Martinez-Duarte, R., Lehal, R., Radtke, F., Renaud, P.: Very high throughput electrical cell lysis and extraction of intracellular compounds using 3D carbon electrodes in lab-on-a-chip devices. Micromachines **3**, 574–581 (2012)
34. Jaramillo, M.D.C., Martínez-Duarte, R., Hüttener, M., Renaud, P., Torrents, E., Juárez, A.: Increasing PCR sensitivity by removal of polymerase inhibitors in environmental samples by using dielectrophoresis. Biosens. Bioelectron. **43**, 297–303 (2013)
35. Martinez-Duarte, R.: Carbon-electrode dielectrophoresis for bioparticle manipulation. ECS Trans. **61**, 11–22 (2014)
36. Rouabah, H.A., Park, B.Y., Zaouk, R.B., Morgan, H., Madou, M.J., Green, N.G.: Design and fabrication of an ac-electro-osmosis micropump with 3D high-aspect-ratio electrodes using only SU-8. J. Micromech. Microeng. **21**, 035018 (2011)
37. Amato, L., Heiskanen, A., Caviglia, C., Shah, F., Zór, K., Skolimowski, M., Madou, M., Gammelgaard, L., Hansen, R., Seiz, E.G., Ramos, M., Moreno, T.R., Martínez-Serrano, A., Keller, S.S., Emnéus, J.: Pyrolysed 3D-carbon scaffolds induce spontaneous differentiation of human neural stem cells and facilitate real-time dopamine detection. Adv. Funct. Mater. **24**, 7042–7052 (2014)
38. Schroers, J., Kumar, G., Madou, M., Martinez-Duarte, R.: Carbon molds for use in the fabrication of bulk metallic glass parts and molds. US 2012/0125071 A1 (2012)
39. Martinez-Duarte, R.: SU-8 photolithography as a toolbox for carbon MEMS. Micromachines **5**, 766–782 (2014)
40. Maitra, T., Sharma, S., Srivastava, A., Cho, Y.-K., Madou, M., Sharma, A.: Improved graphitization and electrical conductivity of suspended carbon nanofibers derived from carbon nanotube/polyacrylonitrile composites by directed electrospinning. Carbon N. Y. **50**, 1753–1761 (2012)
41. Sharma, S., Sharma, A., Cho, Y.-K., Madou, M.: Increased graphitization in electrospun single suspended carbon nanowires integrated with carbon-MEMS and carbon-NEMS platforms. ACS Appl. Mater. Interfaces **4**, 34–39 (2012)
42. Canton, G., Do, T., Kulinsky, L., Madou, M.: Improved conductivity of suspended carbon fibers through integration of C-MEMS and electromechanical spinning technologies. Carbon N. Y. **71**, 338–342 (2014)
43. Naka, K., Konishi, S.: Micro and nano structures of carbonised polymer through pyrolytic transformation from polymer structures. Micro Nano Lett. **1**, 79 (2006)
44. Park, B.Y., Madou, M.J.: Design, fabrication, and initial testing of a miniature PEM fuel cell with micro-scale pyrolyzed carbon fluidic plates. J. Power Sources **162**, 369–379 (2006)
45. Lin, P.-C., Park, B.Y., Madou, M.J.: Development and characterization of a miniature PEM fuel cell stack with carbon bipolar plates. J. Power Sources **176**, 207–214 (2008)
46. Heo, J.I., Shim, D.S., Teixidor, G.T., Oh, S., Madou, M.J., Shin, H.: Carbon interdigitated array nanoelectrodes for electrochemical applications. J. Electrochem. Soc. **158**, J76 (2011)
47. Teixidor, G.T., Gorkin, R.A., Tripathi, P.P., Bisht, G. S., Kulkarni, M., Maiti, T.K., Battacharyya, T.K., Subramaniam, J.R., Sharma, A., Park, B.Y., Madou, M.: Carbon microelectromechanical systems as a substratum for cell growth. Biomed. Mater. **3**, 034116 (2008)
48. Martinez-Duarte, R.: Carbon-electrode dielectrophoresis for bioparticle manipulation. ECS Trans. **61**, 11–22 (2014)
49. Martinez-Duarte, R.: Fabrication of Carbon Micro Molds. University of California, Irvine (2009)
50. Islam, M., Martinez-Duarte, R.: Additive manufacturing of carbides using renewable resources. In: Proceedings of the ASME 2015 IMECE. In press, Houston (2015)

Carbon Microelectromechanical Systems

► Carbon MEMS

Carbon Nanotube (CNT) Arrays

► Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

Carbon Nanotube Materials

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Carbon Nanotube NEMS

Max Zenghui Wang
Department of Electrical Engineering and Computer Science, Case School of Engineering, Case Western Reserve University, Cleveland, OH, USA

Synonyms

CNT NEMS; CNT resonator; Nanotube adsorption sensor; Nanotube force sensor; Nanotube mass sensor; Nanotube resonant sensor; Nanotube resonator

Definition

Carbon nanotube nanoelectromechanical systems (NEMS) are devices that combine the electrical and mechanical degrees of freedom in carbon nanotubes, the one-dimensional (1D) form of crystalline nanocarbon. It holds the record of the smallest NEMS made to date, with the part in motion being a single molecule – an individual carbon nanotube. The most common type of carbon nanotube NEMS device is carbon nanotube resonator, which generally assumes the geometry of a suspended carbon nanotube field-effect transistor (FET), with source and drain electrodes connecting both ends of the nanotube and a third nearby gate electrode forming a capacitor with the nanotube. The suspended segment of the nanotube can be actuated electrostatically by applying a bias voltage (DC, AC, or both) between the nanotube and the gate electrode, and the mechanical motion of the device can in turn be detected electrically by monitoring the electronic transport through the nanotube FET. Combining the outstanding electrical and mechanical properties of its building material, carbon nanotube NEMS devices exhibit, in several aspects, superior performance currently unavailable in NEMS based on other nanomaterials. In particular, leveraging the ultrasmall mass in motion, NEMS resonators based on carbon nanotubes have demonstrated extreme sensitivity to external stimuli (e.g., mass, force). In addition, the atomically perfect crystalline surface of carbon nanotube makes it a unique platform for studying surface physics processes. Together with the ultrahigh sensitivity to adsorbed mass, exotic phenomena have been observed using carbon nanotube NEMS, such as phase transitions in the pseudo-1D system formed by the adsorbed gas atomic layer on the surface of an individual carbon nanotube.

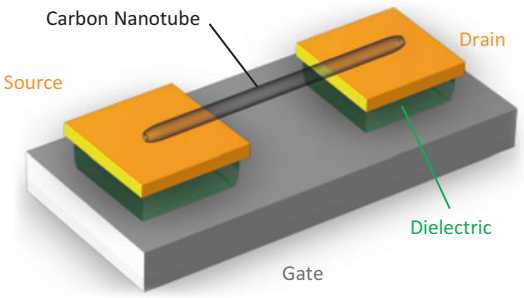
Introduction

Since its discovery, carbon nanotube (CNT) has been extensively studied for its outstanding physical properties. Structurally, CNT can be categorized into single-walled carbon nanotubes (SWCNT) and multi-walled carbon nanotubes (MWCNT). A SWCNT is a hollow cylinder whose entire surface (except the ends) is made of carbon atoms arranged in the graphitic hexagonal honeycomb lattice. It can be conceptually constructed by rolling a piece of graphene (a single layer of graphite) along one of the directions that allow the carbon atoms to seamlessly stitch to their counterparts on the opposite edge of the graphene sheet. Typical SWCNTs have diameter on the order of 1 nm. A MWCNT consists of multiple coaxial layers, with each layer being an individual SWCNT. Consequently, MWCNT generally has greater diameter and larger mass than SWCNT. Most carbon nanotube NEMS devices are constructed based on SWCNTs to take advantage of their small mass and diameter.

Nanoelectromechanical systems are devices that have both electronic and mechanical degrees of freedom, with one or more dimensions of the motional part being on the scale of 100 nm or less. The scaling down of the device dimensions can give rise to new physical phenomena and enhanced

functionalities. NEMS devices are often used for sensing, signal processing, and actuation at the nanoscale.

By having an individual carbon nanotube as the part in motion, CNT NEMS devices leverage the ultrasmall diameter/ultrahigh aspect ratio and ultralow mass of nanotubes, as well as their electronic properties (high current density, quantized and ballistic electronic transport, etc.) and mechanical (high flexibility, high elastic modulus, etc.) properties. While several NEMS based on multi-walled nanotubes utilize the interlayer motion to build linear [1] and rotational [2] bearings, the great majority of carbon nanotube NEMS devices are resonators [3–20] with the resonator body made entirely by a nanotube, with the first one [3] demonstrated in 2004. The rest of the article thus focuses on CNT NEMS resonators.



Carbon Nanotube NEMS, Fig. 1 Schematic of a carbon nanotube NEMS resonator/suspended-channel FET

Geometries/Structures

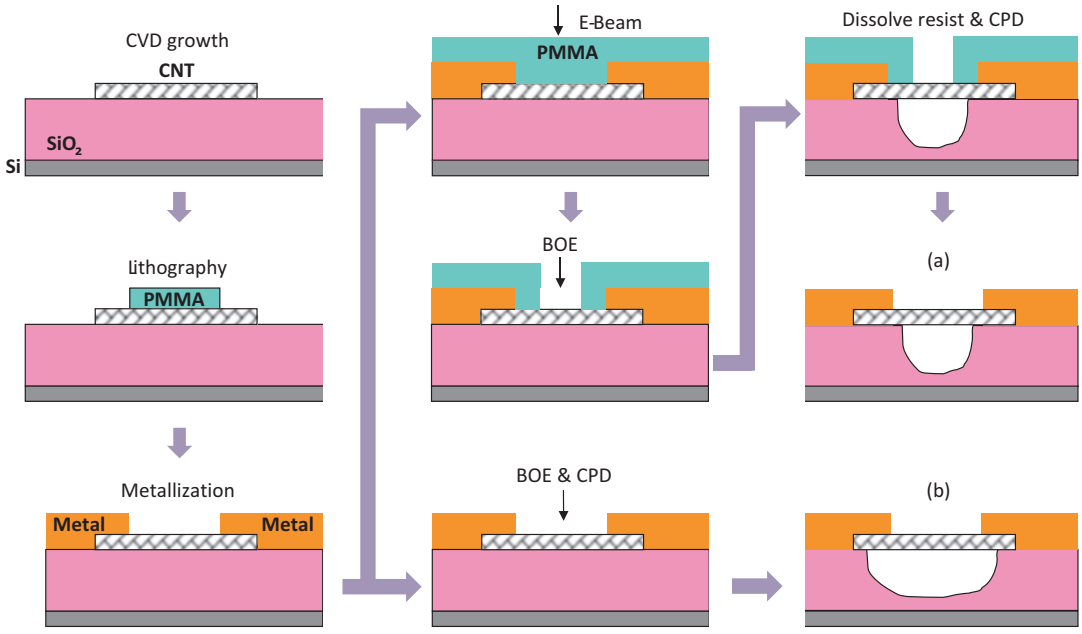
Most carbon nanotube NEMS resonators assume the geometries of a doubly clamped beam/string [3–18] and function as suspended-channel FETs (Fig. 1). The carbon nanotube is suspended over a microtrench and is mechanically anchored at both ends and electrically connected to the source and drain electrodes. In some cases, the electrodes themselves serve as the mechanical anchors for the nanotube (i.e., the nanotube is directly suspended between the two electrodes). The gate electrode is typically located underneath the nanotube (toward the bottom of the trench): a common configuration is to use the heavily doped Si back gate that extends the entire device die; alternatively, local gate at the bottom of the trench (through aligned deposition of metal layers) can also be used. Specially designed gate (e.g., side gate) has also been demonstrated in carbon nanotube NEMS [4].

A less common type of carbon nanotube NEMS takes the form of a singly clamped beam [19, 20]. To date, such structure has been realized using multi-walled carbon nanotubes (with the thinnest being double walled). Electrically, the nanotube is directly connected to one electrode (the clamp) and faces an opposing electrode which is not in contact with the nanotube. Additional gate electrodes along the length of the nanotube can also be used, even though they generally do not provide the canonical gating effect as in the FET structure (as the singly clamped structure does not facilitate electronic transport measurements). Compared with the doubly anchored suspended FET structure, the singly clamped beam structure presents some unique opportunities such as utilization of the field-emission properties of carbon nanotube, at the cost of forfeiting the access to carbon nanotube's electronic transport properties and the capability of electrostatically tuning the mechanical motion through gating, which are detailed in following sections.

Making Carbon Nanotube NEMS

There are a number of ways to fabricate doubly clamped carbon nanotube NEMS, which can be categorized into two main schemes: (i) grow/deposit carbon nanotube on the substrate first and then fabricate the suspended structure and (ii) fabricate the device structure first without carbon nanotube and then grow/transfer the nanotube onto the prefabricated structure.

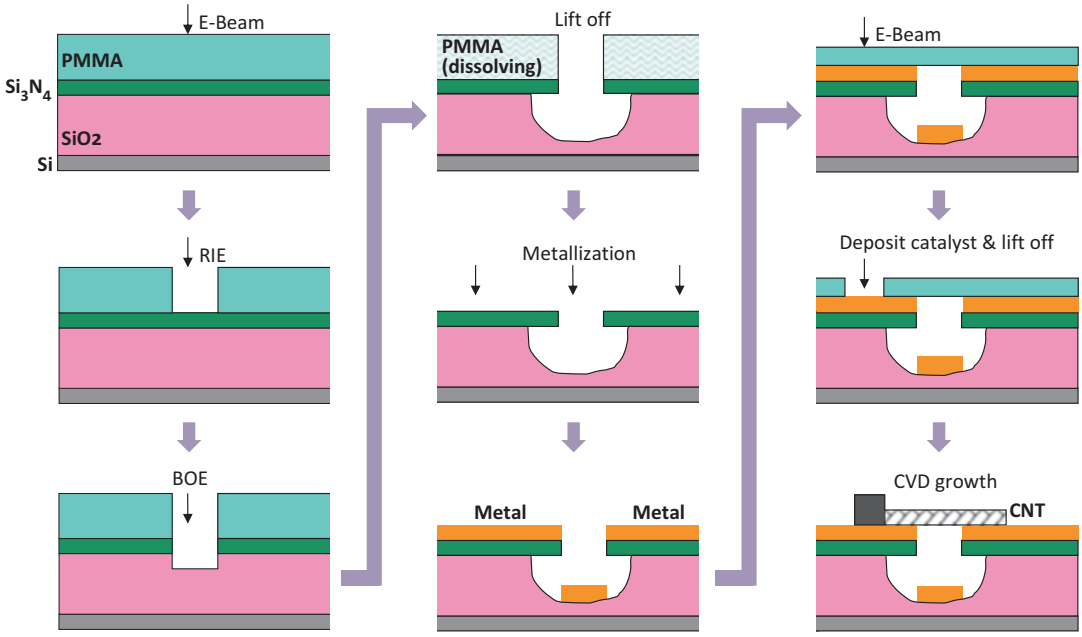
Typical process flow for making doubly clamped carbon nanotube NEMS resonator using scheme (i) is illustrated in Fig. 2. First, carbon nanotube is grown on a SiO_2-covered Si wafer using chemical vapor deposition (CVD). Metal electrodes are then patterned on the carbon nanotube (either aligned to a selective nanotube or



Carbon Nanotube NEMS, Fig. 2 Schematic of fabricating doubly clamped carbon nanotube NEMS resonator using scheme (i), with two major categories of resulting device geometry (**a**) and (**b**) shown

patterned randomly but in a large array) through photolithography, electron-beam lithography, or contact (stencil) mask followed by electron-beam or thermal evaporation. Non-suspended carbon nanotube FET is obtained upon completion of the above steps. To suspend the electrically contacted nanotubes, often a second layer of aligned lithography is performed to open etch windows at locations where suspended segment of carbon nanotube is desired. Such etch window may or may not enclose part of the electrodes (i.e., the inner edges of the electrodes may or may not be exposed in the etch window). In some cases, the metal contacts themselves are used as etch masks, and the second lithography step (for open etch windows) is omitted. Once the etch window is defined, an etch step is performed to remove the underlying oxide in the exposed area to suspend the carbon nanotube. Buffered oxide etch (BOE) is most commonly used, often followed by critical point drying (CPD) to prevent the nanotube from collapsing and sticking to the bottom of the etching-formed microtrench (due to the capillary force during the drying process), while vapor HF etch is a practical alternative.

Depending on the size and geometry of the etch window, two major categories of device geometry can result from this fabrication scheme: (a) the nanotube segment between the metal electrodes is partially suspended in the middle section and clamped to the SiO_2 substrate on both sides and (b) the entire nanotube segment between the metal electrodes is suspended, and the electrodes themselves, while also serving as mechanical clamps for the suspended nanotube, are also partially suspended. Each geometry has its own unique attributes. In the former case (a), the mechanical clamp through van der Waals interaction with the SiO_2 is often sufficiently strong and provides good mechanical clamping, but the electronic properties measured between the electrodes contain information from both suspended and unsuspended segments. In the latter case (b), the entire electronic signal comes from suspended nanotube, but the suspended metal layer can be floppy and exhibit its own resonances during measurement. Note that due to the nondirectional nature of the etch processes (BOE or vapor HF), it is not practical to obtain devices where the boundary of the etched microtrench exactly aligns



Carbon Nanotube NEMS, Fig. 3 Schematic of fabricating doubly clamped carbon nanotube NEMS resonator using scheme (ii)

with the edge of metal electrodes, a third geometry which presumably possesses the advantages of both geometries (a and b) mentioned above.

Typical fabrication process for building carbon nanotube NEMS using scheme (ii) is shown in Fig. 3. The first part is to fabricate the device structure without the nanotube. It can be generally considered as the procedure as in scheme (i) without the nanotube growth at the beginning. Nevertheless, due to the absence of carbon nanotube, a number of things can be done differently. First, directional etch (such as RIE) of SiO_2 can be used instead of BOE or vapor HF, which can result in vertical sidewalls of the microtrench which can be aligned to the edges of the metal electrodes (e.g., using the metal as an etch mask). Second, the choice of supporting material for the carbon nanotube is expanded. For example, a Si_3N_4 can be added on top of the SiO_2 to facilitate undercutting of SiO_2 (which can result in substantially high aspect ratio structures) while providing strong mechanical support to the otherwise floppy metal electrodes [5]. These features are desirable for creating suspended carbon nanotube NEMS in the following step. For example, the undercut will prevent the nanotube from sticking to the SiO_2 microtrench sidewall during growth, yielding a higher suspension rate; and the nitride support will prevent metal from collapsing during the high-temperature growth process while also serving as a stiff clamp once the nanotube is suspended between electrodes. Once the device structure is prepared, a final step is performed to deposit carbon nanotube across the microtrench (and between the electrodes), often through CVD growth [5, 6]. Other techniques, such as stamp transfer, can also be used [7].

Both fabrication schemes have their advantages and limitations. In scheme (i), individual nanotubes can be identified, even characterized (such as using electronic transport and AFM), before additional effort is invested to turn it into a NEMS device by suspending the nanotube. In addition, with proper procedures (such as CPD), the success rate for achieving individual suspended devices is reasonably high, while the effort in making each device is substantial (as devices are fabricated on individual bases). In contrast, in scheme (ii), the device structures can be prefabricated at wafer scale, but for CVD growth, it requires an additional lithography

step to pattern the catalyst (otherwise, the randomly grown nanotubes could electrically short the electrodes), and the yield of the growth step (chances of obtaining an individual suspended carbon nanotube across a microtrench and bridging a pair of electrodes) is far from unity. Nevertheless, the large-scale processes (patterning and growth) largely make up for the total device yield. While the CVD growth may impose limitations on the material choice for metal electrodes (to sustain during the high-temperature growth), the nanotube transfer technique [7] offers a versatile alternative, as long as the nanotube-metal contacts are properly treated (such as annealing or use freshly deposited metal) to ensure good electronic performance.

Besides the abovementioned differences in device structures and materials between the two fabrication schemes, another important contrast is the cleanness of the carbon nanotube. In scheme (i), the nanotube undergoes a number of wet processes and is directly exposed to various chemicals (resist, polymer, etchant, solvent, etc.), which could lead to contaminants and defects on the nanotube surface. However, this does not necessarily preclude the resulting carbon nanotube NEMS device to exhibit certain desired performances, such as ultrahigh sensitivity to admass. Consequently, atomic-level mass sensing has been demonstrated on carbon nanotube NEMS resonators fabricated using this scheme [8–10]. In contrast, in scheme (ii), the carbon nanotube does not undergo any additional processing and thus can maintain its as-grown atomically perfect surface. Therefore, devices fabricated using this scheme are used in applications where pristine nanotube surface is required, such as studying transport in ultraclean suspended nanotubes or using nanotube as a substrate for investigating monolayer gas adsorption on low-dimensional surfaces [5, 11].

Singly clamped nanotube NEMS structures are often made by mounting the carbon nanotube onto a predefined structure, which often involves steps that are not compatible with standard nanofabrication processes, such as attachment of nanotube inside SEM/TEM. The fabrication is often performed at the level of individual devices and is thus highly labor intensive.

Actuation of Device Motion

To date, most nanotube NEMS resonators are operated in vacuum (mTorr or below), with a few measured in low pressure (a few Torr). The most common actuation scheme for carbon nanotube NEMS is electrostatic actuation through a gate electrode, which prevails in the doubly clamped device structure. In this configuration, the nanotube and its gate form the two electrodes of a capacitor. When a voltage is applied across the capacitor (with 0 net electric charge), an attractive electrostatic force is generated between the two electrodes, with its magnitude equal to

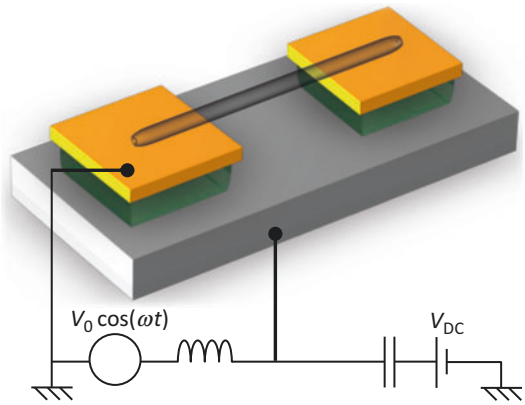
$$F = \frac{1}{2} C' V^2, \tag{1}$$

where C' is a derivative of the capacitance with respect to the distance between its two electrodes and V is applied voltage. Accordingly, the nanotube is attracted toward the gate when a finite gate voltage exists and is driven into oscillation when an AC voltage is applied. For carbon nanotube NEMS resonators, when the driving signal is at the frequency of the device's natural frequency, mechanical resonance with enhanced amplitude ensues.

It is worth noting that in most cases, a DC voltage is applied in addition to an AC voltage (Fig. 4), and there are two main reasons. First, Eq. 1 shows that the driving force is proportional to V^2. If a pure AC voltage $V = V_0\cos(\omega t)$ is applied, $F \propto V^2 \propto \cos^2(\omega t) = (1 + \cos(2\omega t))/2$, which has no $\cos(\omega t)$ term. This suggests that the device will be driven at 2ω instead; thus, the response will be more complicated than that of a simple harmonic resonator. Instead, when a DC component V_{DC} is included, $V = V_0\cos(\omega t) + V_{\mathrm{DC}}$, then

$$\begin{aligned} F \propto V^2 &= (V_0 \cos(\omega t) + V_{DC})^2 \\ &= V_{DC}^2 + 2V_{DC}V_0 \cos(\omega t) \\ &\quad + V_0^2 \cos^2(\omega t), \end{aligned} \tag{2}$$

which, in the case of $V_{\mathrm{DC}} \gg V_0$ (which is often the case), the driving force is mostly at 1ω. Second, Eq. 2 shows that there will be a static term ${V_{\mathrm{DC}}}^2$ in the force applied. Due to the ultra-flexibility of



Carbon Nanotube NEMS, Fig. 4 Electrostatic actuation of doubly clamped carbon nanotube NEMS resonators

carbon nanotube (in comparison with thicker devices with similar geometry, such as lithography-defined nanowires), this static force is capable of generating sufficient tension in the nanotube and significantly tuning its resonance frequency. It is not atypical for a resonance to be tuned by more than 200 % with just a few volts of DC gate voltage.

For singly clamped devices, a nearby electrode (similar to the gate electrode in the doubly clamped case) can be used to capacitively actuate the vibratory motion [19]. Resonant motion driven using electromagnetic wave has also been demonstrated [20]. Additionally, due to the ultrasmall mass of carbon nanotubes, it is possible to directly visualize their completely undriven thermomechanical resonances in electron microscopes as a position-dependent blurring of the nanotube image. This has been observed for both singly and doubly clamped devices.

Readout of Device Motion

Electronic readout prevails in doubly clamped carbon nanotube NEMS resonators. This readout scheme utilizes the fact that motion-induced on-tube charge fluctuation modulates the conductance of the CNT FET and thus generates an oscillation in the amplitude of the electronic current through the device. Specifically, the device displacement δz induces a change in the nanotube-gate capacitance $\delta C = C'\delta z$, which leads to an additional on-tube electric charge δQ, which further leads to a modulation in the device conductance δG. Specifically, the charge variation can be expressed as

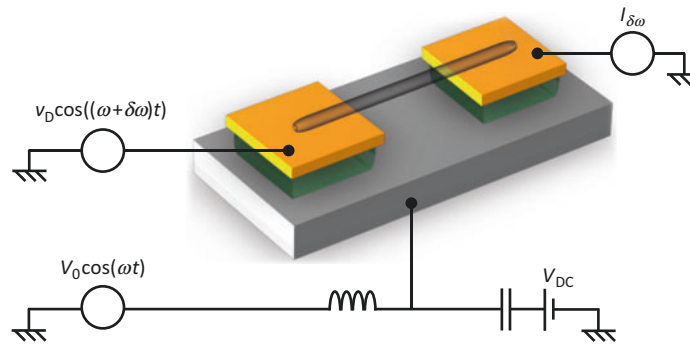
$$\delta Q = V_g \delta C + C \delta V_g. \quad (3)$$

The first term in Eq. 3 accounts for the motion-induced change in capacitance. When the device is mechanically on resonance, the motion amplitude increases substantially from the off-resonance background. Consequently, the first term surges at the device's resonance frequency. The second term represents the conventional gating effect in CNT FET devices and is present even in non-suspended devices. When the device is on resonance, the motion amplitude can be as large as several nm, while the vacuum gap to the gate is typically 100 s of nm. The change in gate capacitance (δC) is therefore only $2 \sim 3$ orders of magnitude smaller than the static gate capacitance of the nanotube, which is sometimes on the same order of magnitude as the $\delta V_g / V_g$ ratio. Consequently, the contribution of both terms in Eq. 3 ($V_g \delta C$ and $C \delta V_g$) can be on the same order of magnitude when the device is on resonance, and the mechanical resonance is often manifested as a sharp feature on top of a slowly varying background.

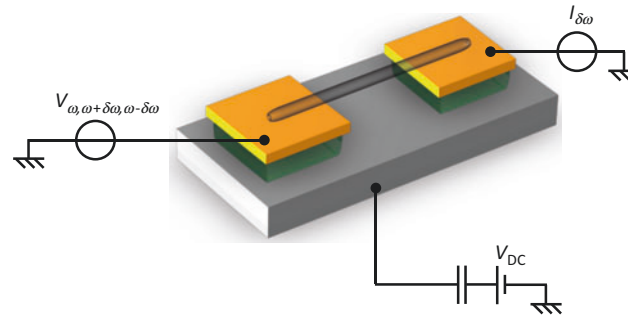
However, such current fluctuation – which is at the same frequency as the device motion – is often masked by the capacitive current generated by the electrostatic driving signal on the gate, which is also at the same frequency, but at a much larger amplitude, due to the large parasitic capacitance from the electrodes (including the wire bonding pads), which is typically several orders of magnitude larger than the nanotube's capacitance.

To overcome this challenge, frequency downmixing technique is often used to extract the motion-induced signal from the large capacitive background. There are two types of mixing techniques: one uses amplitude modulated (AM) signal and the other one uses frequency modulation (FM).

The basic operation principle of AM downmixing roots from a two-source mixing technique [3], which is illustrated in Fig. 5. In this setup, two



Carbon Nanotube NEMS, Fig. 5 Schematic of the two-source mixing setup for resonant motion readout



Carbon Nanotube NEMS, Fig. 6 Schematic of the one-source (*AM*) mixing setup for resonant motion readout

radio-frequency (RF) signals offset by a small frequency difference $\delta\omega$ (typically in the kHz range) are applied to the gate and drain of the CNT FET, respectively. The RF signal at frequency ω generates a nanotube motion at the same frequency and thus the modulation of device conductance: $\delta G = \delta G_0 \cos(\omega t)$. The drain bias applied is at a slightly different frequency: $V_{\text{Drain}} = V_{\text{D}} \cos((\omega + \delta\omega)t)$. The modulation in the current through the nanotube FET is therefore

$$\begin{aligned}\delta I &= V_{\text{Drain}} \times \delta G \\ &= \delta G_0 V_{\text{D}} \cos\left((\omega + \delta\omega)t\right) \cos\left(\omega t\right) \\ &= \delta G_0 V_{\text{D}} (\cos\left((2\omega + \delta\omega)t\right) + \cos\left(\delta\omega t\right))/2\,,\end{aligned} \tag{4}$$

which has a term at the difference frequency $\delta\omega$ and still carries the information of the device motion (through the term δG). In contrast, the capacitive current, though still much larger in amplitude, remains purely at frequency ω. Using a lock-in amplifier referenced at the difference frequency $\delta\omega$, it is possible to reject the electronic signal from all other frequencies and detect the device motion (at ω) by examining the down-mixed FET current at $\delta\omega$.

The AM mixing technique simplifies the two-source scheme by requiring only one RF signal source (Fig. 6). The output of the RF source is AM modulated at the lock-in reference frequency $\delta\omega$ and is applied to the drain electrode

$$V_D = (1 + A\cos\left(\delta\omega t\right))V_0 \cos\left(\omega t\right), \tag{5}$$

which has components at the frequencies ω, $\omega + \delta\omega$, and ω-$\delta\omega$. A is the modulation depth and ranges between 0 and 100 % (values close to 100 % are often used). The gate voltage only carries a DC component in this case. The AC component on the drain effectively creates an alternating bias voltage between the nanotube and the gate and drives the vibration of the nanotube. When considering the motion of the nanotube resonator, this is equivalent to applying these three AC components at the gate. Thus, the AM

scheme is equivalent to the two-source setup, with both the conductance modulation and the drain bias having components in ω, $\omega + \delta\omega$, and ω-$\delta\omega$. Consequently, the mixing current has low frequency components at DC, $\delta\omega$, and $2\delta\omega$. Typically, the $\delta\omega$ component is being measured by the lock-in amplifier.

In the AM readout scheme, even though the capacitive current from the contacts are removed through frequency down-mixing, the non-motional current (the second term in Eq. 3) has a finite term at $\delta\omega$, which presents a frequency-dependent background, and can sometimes hinder the observation of mechanical resonance. The FM readout scheme [12], a later-developed technique, removes this non-motional mixing current in addition to the capacitive current, allowing more efficient detection of mechanical resonance.

For singly clamped devices, direct electronic readout is unavailable as the nanotube itself does not typically form a closed path in the circuit. One can force the open circuit to close by using the nanotube as a field emitter by applying a high voltage across the tube and an opposing cathode and monitor the field-emission current as the nanotube vibrates. Such readout scheme can be destructive, and irreversible shortening of the nanotube due to the large current has been observed [20]. Alternatively, the vibrational amplitude can be estimated directly through imaging in electron microscope. While in such scheme quantitative measurement remains very challenging, imaging offers the opportunity of directly observing the mode shape of the individual resonances, similar to AFM measurements [13].

Frequency Tuning

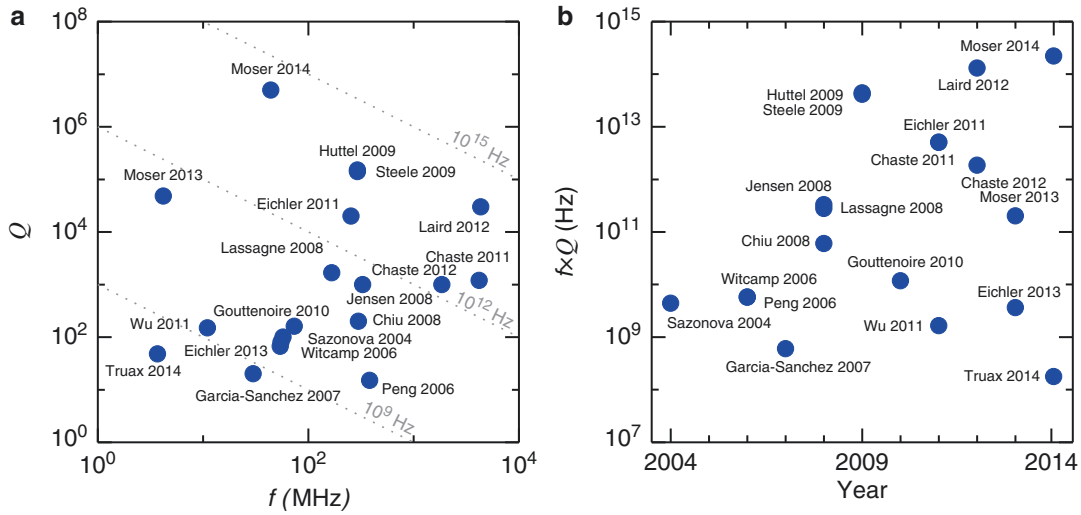
The resonance frequency of carbon nanotube NEMS resonator can often be effectively tuned via adjusting the DC gate voltage. There are two main effects involved: electrostatic tensioning, which leads to an increase in resonance frequency as V_g is increased, and capacitive softening, which has the opposite effect.

For doubly clamped devices, the attractive force resulting from the application of gate bias induces tension in the nanotube and thus increases its resonance frequency. The magnitude of the gate voltage affects the elastic regime in which the nanotube NEMS resonator operates [3]. At small gate bias, the flexural rigidity dominates, and the nanotube vibrates as a bending beam. As V_g increases, the flexural rigidity is overcome by the electrostatic force from the gate, and the nanotube operates as a hanging chain that cannot be elongated (i.e., the chain has negligible flexural rigidity but infinite extensional rigidity). As the gate voltage further increases, the extensional modulus becomes important, and the nanotube behaves like an elastic string. In these different regimes, the frequency has different dependences on the gate voltage.

Due to the large aspect ratio of carbon nanotube, it is common for the suspended segment in a doubly clamped device to exhibit finite slack (i.e., the length of the nanotube is longer than the distance between its two clamping points). It is worth noting that the amount of initial slack affects how the frequency increases with V_g in the different regimes, and such effect further depends on individual resonant modes [3, 14].

Capacitive softening roots from the fact that the nanotube is vibrating in a nonuniform electric field, and the finite field gradient reduces the restoring force toward the equilibrium position (a negative effective spring constant). With finite slack, the nanotube can thus have both in-plane and out-of-plane resonant motion, which exhibit different responses to the capacitive softening effect [4], as the observation of such effect requires the device motion be in the direction of the electric field gradient.

Unlike in doubly clamped devices, tension cannot be easily introduced in singly clamped nanotube NEMS resonators. Thus, such device typically operates in the bending regime, and the use of multi-walled nanotubes (with greater bending rigidity) can help achieve higher resonance frequencies. Further, the gate tuning effect, though it exists, is much smaller than in doubly clamped devices in which tension can be effectively introduced.



Carbon Nanotube NEMS, Fig. 7 Partial summary of (**a**) f_{res}, Q and (**b**) $f_{res} \times Q$ for carbon nanotube NEMS resonators studied between 2004 and 2014. *Dashed lines* in (**a**) represent contours of constant $f_{res} \times Q$

Device Performances

For NEMS resonators, two important device parameters are the resonance frequency f_{res} and quality factor Q. While f_{res} can be effectively turned by gate voltage, within the commonly used V_g range (on the order of ± 10 V), the f_{res} variation is generally within one order of magnitude. In contrast, other device parameters, such as dimension and initial tension, have much larger effects on f_{res}. Specifically, shorter devices have higher f_{res} (e.g., $f_{res} = 4.2$GHz for an 110-nm-long nanotube resonator [15], while $f_{res} = 4.2$ MHz for a 4-μm-long device made by the same group of researchers [16]), and devices with lower tension can have very low f_{res} (e.g., $f_{res} = 3.7$ MHz for a 1.89-μm-long nanotube suspended over a MEMS structure which can introduce slack/reduce tension in the nanotube in a controlled way [17]).

The quality factor of carbon nanotube resonators typically increases as temperature decreases: Q ranges between $10^1 \sim 10^3$ at room temperature, and increases to $10^1 \sim 10^6$ at cryogenic temperatures, due to reduction of phonon-phonon scattering and suppression of electron motion (in the Coulomb blockade regime), two mechanisms that dissipate energy during the nanotube motion. Advancement in transduction techniques has further enabled electronic readout of the completely undriven thermomechanical motion of carbon nanotube NEMS resonators, and by removing external driving and electrostatic noise from the gate electrode, researchers have observed mechanical resonance with intrinsic Q up to 5×10^6 [18]. Figure 7a provides a partial list of f_{res} and Q for some of the nanotube resonators demonstrated between 2004 and 2014, and Fig. 7b shows how $f_{res} \times Q$, one of the most important figures of merit for NEMS devices, evolves throughout the decade.

Applications

Carbon nanotube NEMS resonators are highly sensitive to external stimuli due to the small mass and high flexibility of carbon nanotubes and are thus used in a number of sensing applications.

The resonance frequency of a NEMS resonator is sensitive to any adsorption on its surface, which increases the mass in motion (and thus decreases f_{res}). On the first order (not considering the mode shape of individual resonance modes and the detailed location of adsorbed mass), the frequency

shift δf due to adsorption of additional mass δm for a resonator with mass m is

$$\delta f = \frac{\delta m}{2m} f_{\text{res}}. \tag{6}$$

Given the same frequency resolution, a smaller resonator mass m leads to a finer adsorption mass resolution δm. Given the ultrasmall mass of an individual carbon nanotube (especially single-walled carbon nanotube), it is natural that CNT NEMS resonators are used for ultimate mass sensing. Initial works have achieved mass sensitivity of $10^{-21} \sim 10^{-22}$g [8, 9, 19], sufficient for resolving individual Argon atoms. From Eq. 6, it is also clear that a higher f_{res} leads to a higher frequency-to-mass responsivity (f_{res} shift per adsorbed mass). Consequently, years after the initial experiments, with improvement in device fabrication and transduction scheme, and very importantly, by using a carbon nanotube resonator with a much higher resonance frequency (f_{res} ~2GHz, one order of magnitude higher than in previous works), mass sensing at single proton level (10^{-24}g) has been demonstrated [10].

In addition to mass sensing, carbon nanotube NEMS resonators have also been used for ultrasensitive force detection. In contrast to mass sensing, where high f_{res} (which means short nanotube with high tension) is desirable, to achieve more sensitive force detection, the opposite is preferred. This is because long nanotubes with little initial tension have ultralow spring constant, and thus, its frequency can be highly sensitive to external forces which modify the spring constant of the resonator system. Using a 4-μm-long nanotube resonator with f_{res} as low as 4.2 MHz, researchers have demonstrated force sensitivity of $10^{-20}\text{N/Hz}^{-1/2}$ at cryogenic temperatures [16].

Compared with other NEMS structures fabricated using top-down techniques (such as lithography and etching), carbon nanotube resonator is unique in that the surface of the motional part (the nanotube) can be atomically well defined, especially for CNT NEMS fabricated using the growth-on-prefabricated-structure scheme. This offers the opportunity of exploring surface phenomena on a highly curved and atomically perfect surface, such as phase transition in the adsorbed monolayer on the nanotube surface. Historically, graphite surface has been the canonical platform for studying surface adsorption and phase behavior in two-dimensional adsorption systems, thanks to its simple and well-defined surface structure (the honeycomb pattern formed by carbon atoms). The surface of a single-walled carbon nanotube inherits the advantages of the graphite surface and offers a few additional properties unavailable in conventional two-dimensional (2D) graphitic surfaces: first, it is highly curved and thus imposes an additional boundary condition for the adsorbed layer (along the circumferential direction), and due to this strong confinement, the resulting adsorption system approaches the 1D limit; second, the significant reduction of carbon atoms (compared with those on and under conventional 2D graphitic surface) leads to decreased adsorption potential energy and thus can shift the equilibria of phase transitions in the adsorption system. When a phase transition takes place in the adsorbed atomic layer (which is accompanied by a sudden change in its density, as in 3D phase transitions), the mass of the resonator undergoes an abrupt change, which is reflected in a sudden shift in resonance frequency. Experimentally, phase transitions have been observed for Ar, Kr, and ^{4}He, with several low-dimensional phases, including commensurate phase – a phase specific to adsorption systems, in which the arrangement of adsorbate atoms conforms to the underlying substrate atoms – being identified [5, 11].

Besides sensing applications, carbon nanotube NEMS resonators have also been demonstrated as active components in RF circuits. Both singly clamped [20] and doubly clamped [12] structures have been used. In one example, researchers utilize the field-emission current from a singly clamped nanotube resonator to function the NEMS device as the antenna, tuner, amplifier, and demodulator in a radio. In another case, using the FM mixing technique, researchers demonstrate digital demodulation function with a doubly clamped carbon nanotube NEMS resonator, with data transfer rate meeting the GSM specifications.

Cross-References

- Carbon Nanotube NEMS
- Carbon-Nanotubes
- Graphene NEMS
- Nanomechanical Properties of Nanostructures
- Nanomechanical Resonant Sensors and Fluid Interactions
- NEMS Mass Sensors
- NEMS Resonant Chemical Sensors
- NEMS Resonant Mass Sensors

References

1. Cumings, J., Zettl, A.: Low-friction nanoscale linear bearing realized from multiwall carbon nanotubes. Science **289**, 602–604 (2000)
2. Fennimore, A.M., Yuzvinsky, T.D., Han, W.Q., Fuhrer, M.S., Cumings, J., Zettl, A.: Rotational actuators based on carbon nanotubes. Nature **424**, 408–410 (2003)
3. Sazonova, V., Yaish, Y., Ustunel, H., Roundy, D., Arias, T.A., McEuen, P.L.: A tunable carbon nanotube electromechanical oscillator. Nature **431**, 284–287 (2004)
4. Wu, C.C., Zhong, Z.: Capacitive spring softening in single-walled carbon nanotube nanoelectromechanical resonators. Nano Lett. **11**, 1448–1451 (2011)
5. Wang, Z., Wei, J., Morse, P., Dash, J.G., Vilches, O.E., Cobden, D.H.: Phase transitions of adsorbed atoms on the surface of a carbon nanotube. Science **327**, 552–555 (2010)
6. Peng, H.B., Chang, C.W., Aloni, S., Yuzvinsky, T.D., Zettl, A.: Ultrahigh frequency nanotube resonators. Phys. Rev. Lett. **97**, 87203 (2006)
7. Wu, C.C., Liu, C.H., Zhong, Z.: One-step direct transfer of pristine single-walled carbon nanotubes for functional nanoelectronics. Nano Lett. **10**, 1032–1036 (2010)
8. Chiu, H.Y., Hung, P., Postma, H.W.C., Bockrath, M.: Atomic-scale mass sensing using carbon nanotube resonators. Nano Lett. **8**, 4342–4346 (2008)
9. Lassagne, B., Garcia-Sanchez, D., Aguasca, A., Bachtold, A.: Ultrasensitive mass sensing with a nanotube electromechanical resonator. Nano Lett. **8**, 3735–3738 (2008)
10. Chaste, J., Eichler, A., Moser, J., Ceballos, G., Rurali, R., Bachtold, A.: A nanomechanical mass sensor with yoctogram resolution. Nat. Nanotechnol. **7**, 301–304 (2012)
11. Lee, H.-C., Vilches, O.E., Wang, Z., Fredrickson, E., Morse, P., Roy, R., Dzyubenko, B., Cobden, D.H.: Kr and ^{4}He adsorption on individual suspended single-walled carbon nanotubes. J. Low Temp. Phys. **169**, 338–349 (2012)
12. Gouttenoire, V., Barois, T., Perisanu, S., Leclercq, J.-L., Purcell, S.T., Vincent, P., Ayari, A.: Digital and FM demodulation of a doubly clamped single-walled carbon-nanotube oscillator: towards a nanotube cell phone. Small **6**, 1060–1065 (2010)
13. Garcia-Sanchez, D., Paulo, A.S., Esplandiu, M.J., Perez-Murano, F., Forro, L., Aguasca, A., Bachtold, A.: Mechanical detection of carbon nanotube resonator vibrations. Phys. Rev. Lett. **99**, 85501 (2007)
14. Eichler, A., Moser, J., Chaste, J., Zdrojek, M., Wilson-Rae, I., Bachtold, A.: Nonlinear damping in mechanical resonators made from carbon nanotubes and graphene. Nat. Nanotechnol. **6**, 339–342 (2011)
15. Chaste, J., Sledzinska, M., Zdrojek, M., Moser, J., Bachtold, A.: High-frequency nanotube mechanical resonators. Appl. Phys. Lett. **99**, 213502 (2011)
16. Moser, J., Güttinger, J., Eichler, A., Esplandiu, M.J., Liu, D.E., Dykman, M.I., Bachtold, A.: Ultrasensitive force detection with a nanotube mechanical resonator. Nat. Nanotechnol. **8**, 493–496 (2013)
17. Truax, S., Lee, S.-W., Muoth, M., Hierold, C.: Axially tunable carbon nanotube resonators using co-integrated microactuators. Nano Lett. **14**(11), pp 6092–6096 (2014), doi:10.1021/nl501853w
18. Moser, J., Eichler, A., Güttinger, J., Dykman, M.I., Bachtold, A.: Nanotube mechanical resonators with quality factors of up to 5 million. Nat. Nanotechnol. **9**, 1007–1011 (2014), doi:10.1038/nnano.2014.234, advance online publication
19. Jensen, K., Kim, K., Zettl, A.: An atomic-resolution nanomechanical mass sensor. Nat. Nanotechnol. **3**, 533–537 (2008)
20. Jensen, K., Weldon, J., Garcia, H., Zettl, A.: Nanotube radio. Nano Lett. **7**, 3508–3511 (2007)

Carbon Nanotube-Metal Contact

Wenguang Zhu
Department of Physics and Astronomy, The University of Tennessee, Knoxville, TN, USA

Synonyms

Carbon nanotube-metal interface

Definition

Carbon nanotube-metal contacts are widely present in many carbon nanotube-based nanodevices, and their electronic structures may significantly

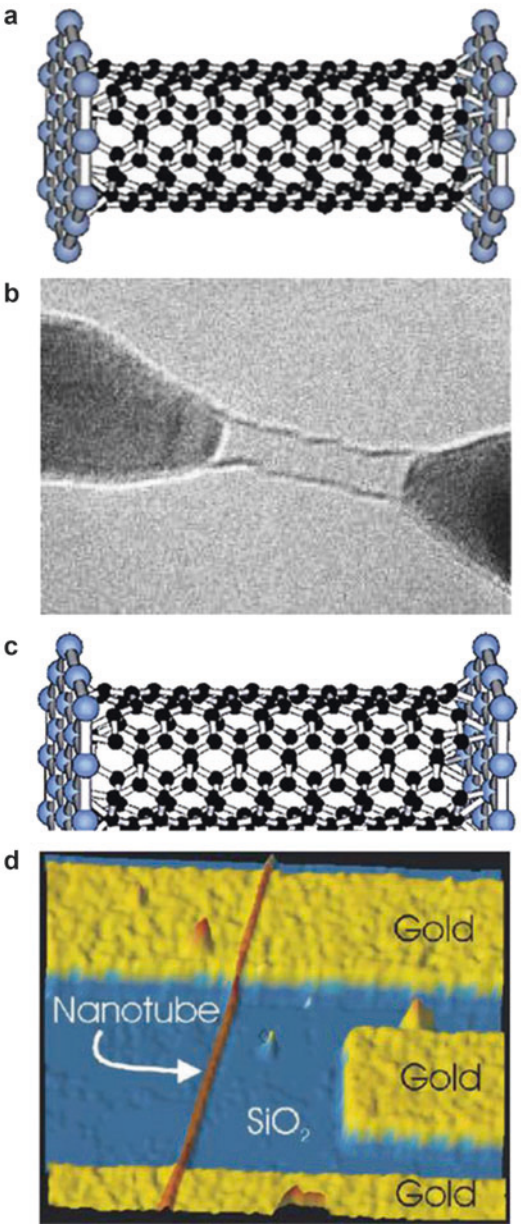
influence the operation and performance of carbon nanotube-based nanodevices.

Overview

Carbon nanotubes (CNTs) are quasi-one-dimensional materials with remarkable mechanical and electronic properties promising a wide range of applications from field-effect transistors (FETs) and chemical sensors to photodetectors and electroluminescent light emitters. In most of these CNT-based nanodevices, metals are present as electrodes in contact with the CNTs. Many factors including the CNT-metal contact geometry, microscopic atomic details at the interface, and the resulting electronic structure can play a significant role in determining the functionality and performance of the devices. For instance, it has been demonstrated that an individual semiconducting CNT can operate either as a conventional FET or an unconventional Schottky barrier transistor, depending on the properties of the metal-CNT contact. In general, the electrical transport characteristics of the CNT-metal systems are sensitive to the choice of metal element as the electrode.

CNT-Metal Contact Geometry

There are two types of interface geometries of CNT-metal contacts, i.e., end contact and side contact [1, 2]. The end-contact geometry refers to the cases where metals are merely in contact with the open ends of one-dimensional CNTs, as illustrated in Fig. 1a. This contact geometry can be naturally achieved in the catalytic CVD growth of CNTs, where CNTs sprout from catalytic metal particles with the CNT axis normal to the metal surface. Figure 1b shows a sample experimental image of an end contact between a single-wall CNT and two Co tips in an in situ electron microscopy setup. The side-contact geometry refers to the cases where metals are in contact with the sidewall of CNTs, as illustrated in Fig. 1c. This contact geometry occurs when a CNT lays on the surface of a flat metal substrate. In most of CNT-based nanodevices, such as CNT FETs, metal strips are deposited from above to cover the CNTs laying on the surface as to build



Carbon Nanotube-Metal Contact, Fig. 1 A schematic illustration of (**a**) an end contact and (**c**) a side contact between a CNT and a metal (From Ref. [3], Fig. 1). (**b**) A CNT forming end contacts with Co tips (From Ref. [4], Fig. 1). (**d**) A CNT forming side contacts on gold electrodes (From Ref. [2], Fig. 24)

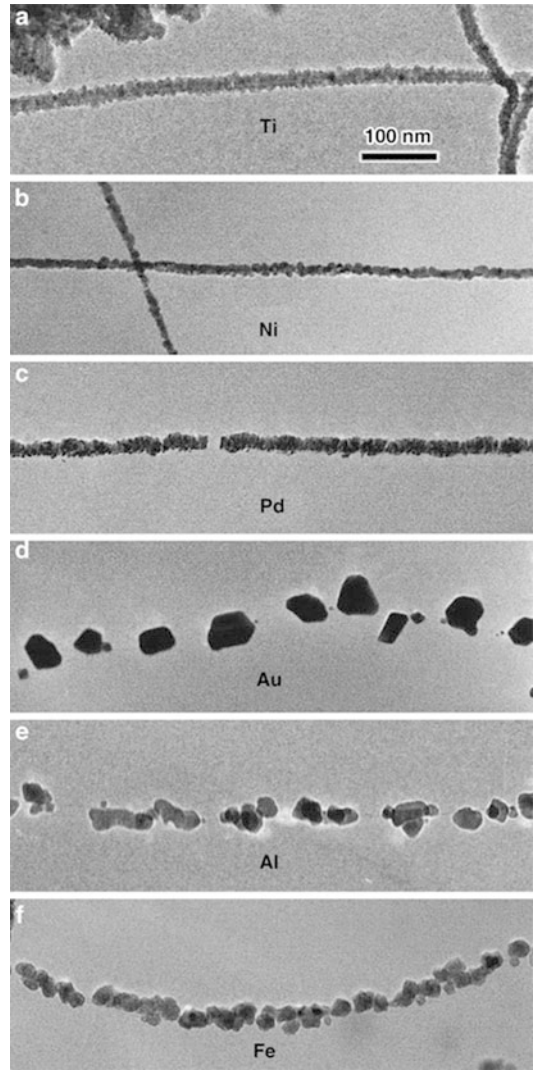
electrodes, fully covering sections of the CNTs, as shown in Fig. 1d. Among these two contact geometries, the side-contact geometry is more technologically relevant to CNT-based nanodevices.

Bonding and Wetting Properties of Metals on CNTs

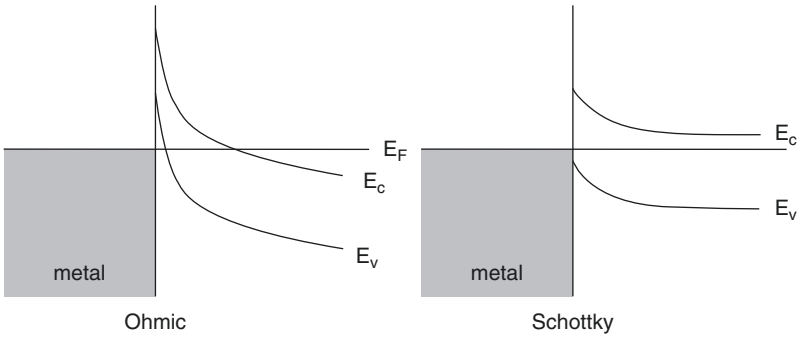
In the end-contact geometry, metals form strong covalent bonds with carbon atoms at the open ends of CNTs [5]. The bonding energy can be as high as, e.g., 7.6 eV for a single bond at a CNT–Co contact, according to density functional calculations. Due to the strong covalent nature of the bonding, large mismatch-induced strains or high tensile strength can be built up at the interface.

In the side-contact geometry, metals and CNTs form much weaker bonds due to the nearly chemically inert side walls of CNTs [5]. Single-wall CNTs are built up of a cylindrically closed sheet of graphene, in which carbon atoms arranged in a honeycomb structure form very stable *sp2*-hybridized covalent bonds with the *pz*-orbitals of carbon extending normal to the sidewalls. The interaction between metals and CNTs in the side-contact geometry is determined by the hybridization between the carbon *pz*-orbitals and the unbonded orbitals of the metals. Alkali and simple metals have binding energy around 1.5 eV per atom. Some transition metal atoms with unpaired *d* electrons, such as Sc, Ti, Co, Ni, Pd, Pt, form strong bonds with a binding energy around 2.0 eV per atom, whereas the transition metals with fully occupied *d* orbitals such as Cu, Au, Ag, and Zn have relatively weak binding with a binding energy less than 1.0 eV per atom. On the other hand, the binding energy of metal on CNTs also depends on the radius of CNTs. In general, the larger is the radius, the weaker the binding energy.

The wettability of metals on CNTs is critical to the electrical transport properties at CNT-metal contacts. In addition, CNTs can be used as templates to produce metallic nanowires with controllable radius by continuously coating the sidewalls of CNTs with metals. Experiments using different techniques such as electron beam evaporation, sputtering, and electrochemical approaches have achieved continuous coating of Ti and quasi-continuous coating of Ni and Pd on CNTs [5, 6]. Such metallic nanowires are ideal to be used as conducting interconnects in nanodevices. Metals such as Au, Al, Fe, Pb form isolated discrete clusters rather than a uniform coating layer on the surface of CNTs. Figure 2 shows sample TEM images of Ti, Ni, Pd, Au, and Fe coatings on CNTs [6]. The correlation between the wettability



Carbon Nanotube-Metal Contact, Fig. 2 TEM images of (**a**) Ti, (**b**) Ni, (**c**) Pd, (**d**) Au, (**e**) Al, and (**f**) Fe coatings on carbon nanotubes



Carbon Nanotube-Metal Contact, Fig. 3 A schematic illustration of the energy levels for an ohmic and a Schottky contact between a metal and a semiconducting CNT

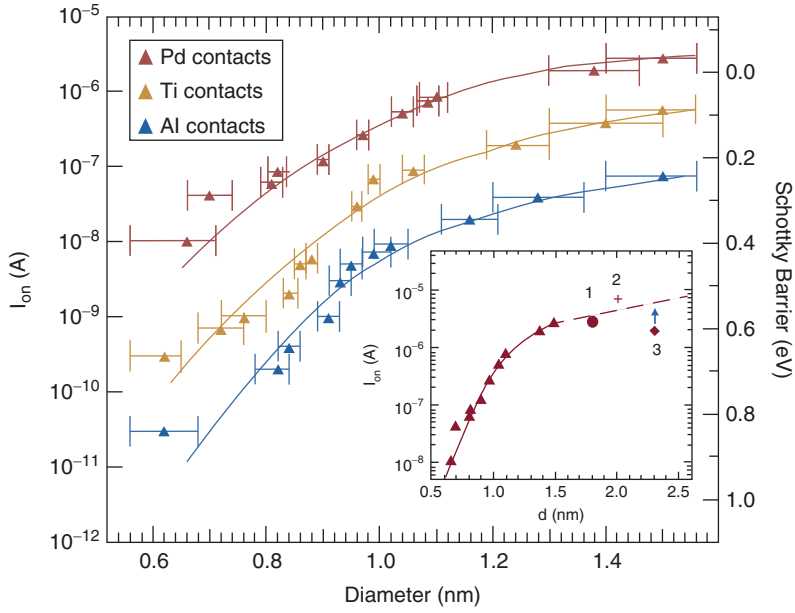
of these metals and their binding energies on CNTs is clear, i.e., metals with relatively strong binding energies with CNTs tend to form uniform coatings.

Electronic Structures of CNT-Metal Contacts

The electronic structure of CNT-metal contacts has a significant impact on the operation and performance of CNT-based nanodevices [2]. Due to the one-dimensional nature of CNTs and their special contact geometries, CNT-metal contacts exhibit some unusual features when compared to traditional planar contacts.

For metallic CNTs, as in contact with metals, ohmic contacts are normally formed at the interface, where no interface potential barrier exists, and the contact resistance is primarily determined by the wettability of the metal and the local atomic bonding and orbital hybridization at the interface [2]. Palladium is found to be optimal as electrodes to make ohmic contacts with metallic CNTs.

Semiconducting CNTs form either ohmic contacts or Schottky barriers at the interface with metals [2]. Figure 3 schematically illustrates the energy levels for an ohmic and a Schottky contact between a metal and a semiconducting CNT. A distinctive feature of CNT-metal contacts from traditional planar metal/semiconductor interfaces is that the height of the Schottky barrier formed at CNT-metal contacts strongly depends on the work function of the metal for a given semiconducting CNT [7, 8]. In general, at traditional planar metal/semiconductor interfaces, the Schottky barrier height shows very weak dependence on the metal work function due to the so-called Fermi-level pinning effect [9]. The strong dependence of the Schottky barrier on the metal work function in CNT-metal contacts is attributed to the reduced dimensionality of CNTs, which entirely changes the scaling of charge screening at the interface, making the depletion region decay rapidly in a direction normal to the interface and thus significantly weakening the Fermi-level pinning effect [7]. Experimental and theoretical work has shown that the interface Schottky barrier regions are much thinner in one dimension than those in three dimensions. In this case, charge carrier tunneling through the Schottky barriers becomes important. Because of the involvement of tunneling and thermionic emission in the carrier transport at the interfaces, the dependence of the on-current of CNT transistors on the Schottky barrier becomes very strong. Figure 4 shows experimental CNT-FET on-current and Schottky barrier height as a function of the CNT diameter for three different metal electrodes, Pd, Ti, and Al [8]. When the metal work functions are in the valence or conduction band of the semiconducting CNTs, ohmic contacts will likely be formed at the interface. Experimental measurements have shown that certain metals with high work functions, such as Pd, can produce nearly ohmic contacts with semiconducting CNTs. Ohmic contacts are more desirable in devices where contact resistance needs to be minimized. In addition to the metal work function, other factors, such as the



Carbon Nanotube-Metal Contact,
Fig. 4 Experimental CNT-FET on-current (*left axis*) and computed Schottky barrier height (*right axis*) as a function of the CNT diameter for three different metal electrodes, Pd, Ti, and Al

contact geometry and the chemical bonding at the interface also play important roles in the transport properties of CNT-metal contacts.

Cross-References

- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Carbon-Nanotubes
- ▶ CMOS-CNT Integration

References

1. Banhart, F.: Interactions between metals and carbon nanotubes: at the interface between old and new materials. Nanoscale **1**, 201–213 (2009)
2. Anantram, M.P., Léonard, F.: Physics of carbon nanotube electronic devices. Rep. Prog. Phys. **69**, 507–561 (2006)
3. Palacios, J.J., Pérez-Jiménez, A.J., Louis, E., SanFabián, E., Vergés, J.A.: Phys. Rev. Lett. **90**, 106801 (2003)
4. Rodríguez-Manzo, J.A., et al.: Small **5**, 2710–2715 (2009)
5. Ciraci, S., Dag, S., Yildirim, T., Gülseren, O., Senger, R. T.: Functionalized carbon nanotubes and device applications. J. Phys. Condens. Matter **16**, R901–R960 (2004)
6. Zhang, Y., Franklin, N.W., Chen, R.J., Dai, H.J.: Metal coating on suspended carbon nanotubes and its implication to metal-tube interaction. Chem. Phys. Lett. **331**, 35–41 (2000)
7. Léonard, F., Tersoff, J.: Role of fermi-level pinning in nanotube Schottky diodes. Phys. Rev. Lett. **84**, 4693–4696 (2000)
8. Chen, Z.H., Appenzeller, J., Knoch, J., Lin, Y.-M., Avouris, P.: The role of metal – nanotube contact in the performance of carbon nanotube field-effect transistors. Nano Lett. **5**, 1497–1502 (2005)
9. Tung, R.T.: Recent advances in Schottky barrier concepts. Mater. Sci. Eng. R **35**, 1–138 (2001)

Carbon Nanotube-Metal Interface

- ▶ Carbon Nanotube-Metal Contact

Carbon Nanotubes

- ▶ Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae

Carbon Nanotubes (CNTs)

- ▶ Chemical Vapor Deposition (CVD)
- ▶ Physical Vapor Deposition

Carbon Nanotubes for Chip Interconnections

Gilbert Daniel Nessim
Chemistry Department, Bar-Ilan Institute of Nanotechnology and Advanced Materials (BINA), Bar-Ilan University, Ramat Gan, Israel

Synonyms

Carbon nanotubes for interconnects in integrated circuits; Carbon nanotubes for interconnects in microprocessors

Definition

Chip interconnections electrically connect various devices in a microprocessor. Today's established technology for interconnects is based on copper. However, it may be technically challenging to extend copper use to future interconnects in microprocessors with smaller lithographic dimensions due to materials properties limitations. Carbon nanotubes are currently investigated as a potential replacement for future integrated circuits (microprocessors). Although carbon nanotubes are a clear winner against copper in terms of materials properties, multiple fabrication challenges need to be overcome for carbon nanotubes to enter the semiconductor fab and replace copper for chip interconnections.

Motivation

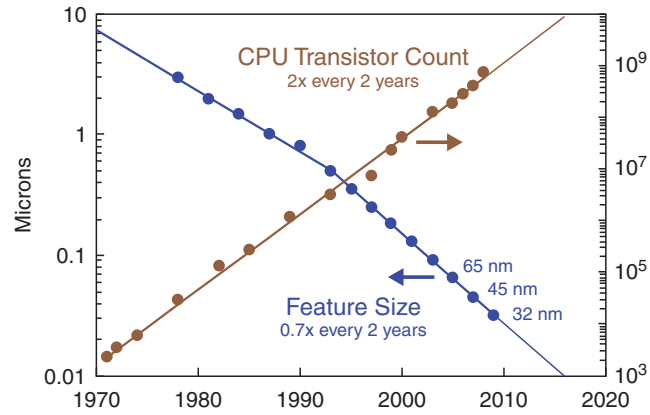
Following over 40 years of successful fulfillment of Moore's law, stating that the number of transistors in a chip doubles every 2 years, we have already moved from microelectronics to nanoelectronics [1]. Although the *"end of scaling"* has been predicted many times in the past, enormous technical challenges, especially quantum mechanical issues and billion-dollar lithography investments, are a serious threat to further miniaturization (Fig. 1).

Today's latest processors are manufactured using the 32-nm technology. To move toward the 22-nm node and beyond, issues such as lithographic limitations, leaking currents in ultra-thin dielectrics (only a few monolayers thick), insufficient power and thermal dissipation, and interconnect reliability must be resolved [1]. At the transistor level, the performance is negatively affected by increased off-state currents due to short channel effects, increased gate leakage due to tunneling through nanometer-thin dielectric layers, and increased overall gate capacitance due to decreasing gate pitch.

Although quantum mechanical tunneling and leakage currents may eventually stop further scaling, efficient heat removal from a chip is currently the biggest obstacle. In this respect, the many kilometers of copper interconnects present in today's chips are the main culprit for heat generation. For instance, in 2004, Magen et al. [2] showed that for a microprocessor fabricated with the 0.13-μm-node technology consisting of 77 million transistors, interconnects consumed more than 50 % of the total dynamic power. Given the increased length of interconnects, their reduced cross section, and the increased current densities circulating into the interconnects of our latest chips, the problem has been further exacerbated.

Additionally, copper interconnects are a major contributor to the total resistance-capacitance (RC) delay of the chip, can fail by electromigration, and need a liner to avoid diffusion into the silicon. Bottom line, the interconnect issue is so serious that the International Semiconductor Roadmap [3] (ITRS, an expert team assessing the semiconductor industry's future technology requirements for the next 15 years) indicates copper interconnects as a possible dealbreaker to further miniaturization for IC nodes beyond 22 nm.

Many technology options are currently under investigation to replace copper for interconnects. Among them, we can mention other metals (mainly silicides), wireless, plasmonics, and optical interconnects. Most notably, there has been an intense research effort on new nanotechnology materials such as carbon nanotubes, which, at



Carbon Nanotubes for Chip Interconnections, Fig. 1 Moore's Law: Transistor count has doubled while feature size has decreased by 0.7X every 2 years (Figure reprinted with permission from Kuhn [1])

the theoretical level, could solve all the above technical issues suffered by copper.

The plan of this section is to first introduce the reader to copper interconnects' fabrication and limitations. Next, we will compare copper to carbon nanotubes (CNTs) and detail possible models for implementation. An important paragraph will focus on the state-of-the-art of CNT fabrication, prior to concluding on the outstanding issues and outlook for future CNT-based chip interconnections.

Background on Copper Interconnects and Dual-Damascene Process

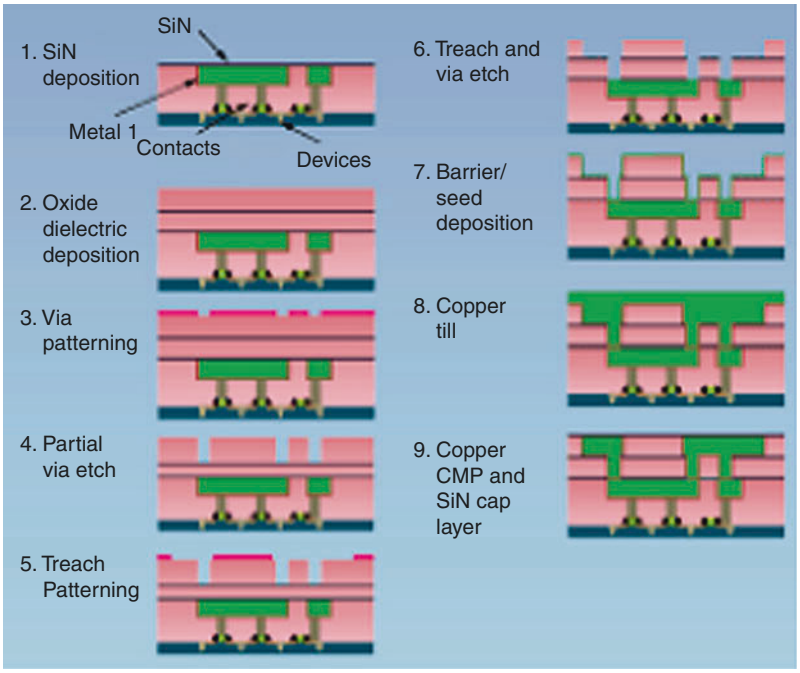
In 1997, IBM introduced the revolutionary *"dual-damascene"* process to fabricate copper interconnects and to replace aluminum interconnects, the industry standard at the time. Compared to aluminum, copper presents two major advantages: (1) 50 % lower resistivity (Cu ≈ 1.75 μm cm vs. Al ≈ 3.3 μm cm) and (2) higher current densities before failure by electromigration (up to 5×10^6 A/cm^2) [4, 5]. Although, as a material, copper was a clear winner against aluminum, fabrication challenges delayed its introduction. Historically, we may be at a similar juncture with carbon nanotubes compared to copper as we were with copper compared to aluminum in 1997: in spite of their superior materials properties, mainly fabrication issues are now preventing the introduction of nanotubes in the semiconductor industry to replace copper interconnects.

Copper diffuses into silicon, generating mid-gap states that significantly lower the minority carrier lifetime and which lead to leakage in diodes and bipolar transistors. Copper also diffuses through SiO_2 and low-k dielectrics, and therefore requires complete encapsulation in diffusion barriers. Since no dry etches were known for copper, IBM's bold innovation of polishing using chemical mechanical polishing (CMP), after electroplating the copper, was significantly at odds with the technological processes at that time in semiconductor fabrication.

The copper dual-damascene process consists of the following steps:

- Develop a pattern for wires or vias by patterned etching of the dielectric.
- Deposit a barrier layer (usually Ta) to prevent copper diffusion into silicon.
- Deposit a copper seed layer.
- Fill the vias with copper using electrodeposition.
- Remove excess copper using CMP.
- Repeat the process to lay the alternating layers of wires and vias which will form the complete wiring system of the chip (Fig. 2).

Typical microprocessor design follows a *"reverse scaling"* metallization scheme with multiple layers of interconnects labeled as local, intermediate, and global interconnects, with increasing width. Very thin local interconnects locally connect gates and transistors within a functional block and are usually found in the lower two



Carbon Nanotubes for Chip Interconnections, Fig. 2 Dual-damascene process of copper filling an interconnect via (Figure reprinted with permission from Jackson et al. [21])

metal layers. The wider and taller intermediate interconnects have lower resistance and provide clock and signal within a functional block up to 4 μm. Global interconnects are found at the top metal layers and provide power to all functions in addition to connecting functional blocks through clock and signal. They are usually longer than 4 μm (up to half of the chip perimeter) and exhibit very low resistance to minimize RC delay and voltage drop. Below are a typical cross section of an I.C. chip and a possible implementation using CNTs (Figs. 3 and 4).

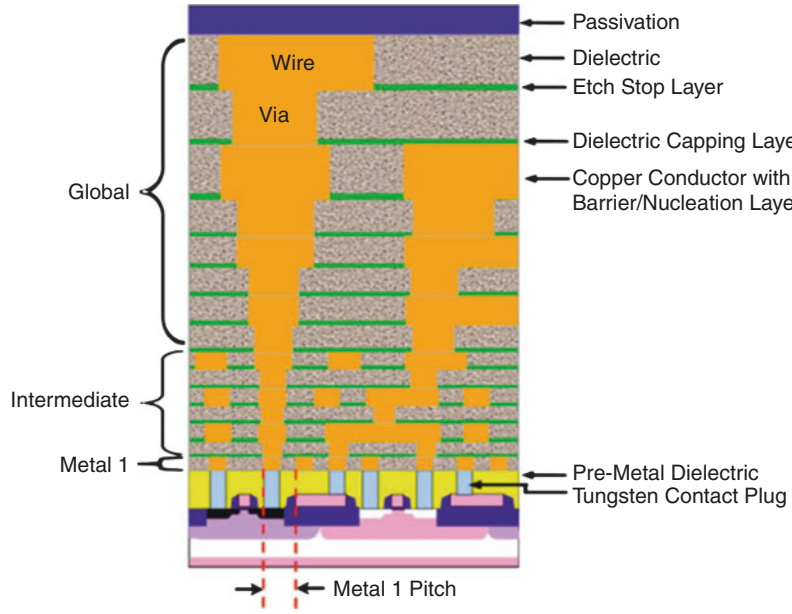
Limitations of Copper Interconnections

Copper interconnects have efficiently scaled down to the current 32-nm-node microprocessors, although this has required many technological advances to allow ever-shrinking copper cross sections to carry increasing currents without failure. However, we may be very close to smashing against a technical wall because of materials failure and related fabrication issues.

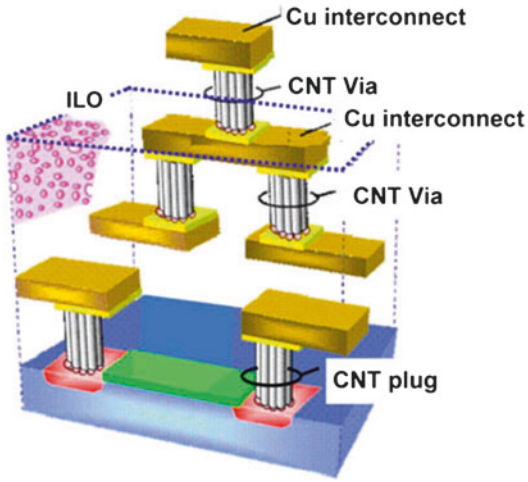
Alternative materials or technologies would require many changes in semiconductor fabrication and massive investments; thus the large semiconductor companies are doing the impossible to extend copper application to future nodes. It is clear that only when up against an insurmountable technical wall will the semiconductor industry switch to a new technology.

Electrical resistance is a major issue now that copper interconnect cross sections are comparable to the mean free path of electrons in copper (~40 nm in Cu at room temperature). Grain boundary and surface scattering are significant contributors to the increased resistance, especially now that we have reached nanoscale dimensions. At the microstructural level, the grain boundaries play an important role, hence, among other fabrication concerns, controlling the copper grain size during electrodeposition has allowed to limit the grain boundary scattering impact thus far.

The steep rise in interconnect resistance for smaller IC nodes is a major source of RC delays and directly affects the chip reliability by increasing the risk of electromigration failure, a major issue for further downscaling. Electromigration is the transport of material caused by the gradual movement of the copper ions due to the momentum transfer between conducting electrons and



Carbon Nanotubes for Chip Interconnections, Fig. 3 Typical cross sections of hierarchical scaling in current microprocessor (Figure reprinted with permission from the Semiconductor Industry Association [22])



Carbon Nanotubes for Chip Interconnections, Fig. 4 Schematic view of possible implementation of carbon nanotube via interconnects in lieu of copper (Figure reprinted with permission from Awano et al. [7])

diffusing metal atoms, which occurs for high current densities, which can create voids leading to open circuits. Given that downscaling leads to a reduction of the interconnects' cross section, the problem is amplified at subsequently smaller nodes. To compound the issue, the need for a resistive diffusion barrier layer, also called a liner (usually Ta), to avoid copper diffusion into silicon, further reduces the available conductive copper cross section, thus increasing the risk of electromigration failure, especially as the operating temperature rises.

In addition to the increased resistance and the electromigration failure risk, many other aspects of the dual-damascene process are becoming potential sources of failure as the node shrinks. Among the many integration concerns, we can mention materials issues such as interface adhesion between the different materials (copper, low-k dielectrics, etc.), liner effectiveness, metal voids, CMP interface defects, etc. Concurrently, there is a long list of process-related issues such as the need for etch/strip/clean processes (to avoid damage to low-k dielectric materials), atomic layer deposition (ALD) processes to deposit liners, copper plating and CMP techniques, etc.

A few interesting numerical estimates taken from the 2009 projections from ITRS, [3] provide the reader with the magnitude of the technical challenge to extend copper interconnect technology (Table 1).

As already mentioned, many alternative technologies are currently being investigated for replacement of copper as interconnect material that would require significant chip redesigns and

Carbon Nanotubes for Chip Interconnections, Table 1 Selected critical parameters for copper use as interconnects in future IC nodes (Data from the Semiconductor Industry Association [22])

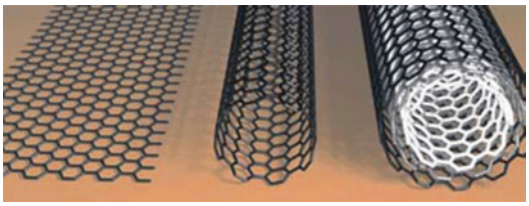
Year of production (estimated)	2010	2015	2020
MPU/ASIC metal 1½ pitch (nm) (contacted)	45	25	14
Total interconnect length (m/cm^2) – metal 1 and 5 intermediate levels, active wiring only	2,222	4,000	7,143
Barrier/cladding thickness (for Cu intermediate wiring) (nm)	3.3	1.9	1.1
Interconnect RC delay [ps] for a 1-mm Cu intermediate wire, assumes no scattering and an effective ρ of 2.2 μΩ-cm	1,132	3,128	9,206

new fabrication technologies. Some examples include optical interconnects, radio frequency (RF) interconnects, plasmonics, and 3-D interconnects (probably still copper). The interested reader can find more details on these alternative technologies in the review paper from Havemann et al. [6] (now a little dated) or in the latest ITRS report on interconnects [3].

The Case for Carbon Nanotube Interconnects

An interesting solution, which has been the subject of intense research in recent years, is to replace copper with carbon nanotubes. If a reliable and repeatable fabrication process consistent with Complementary Metal Oxide Semiconductor (CMOS) technology requirements could be developed, integration into existing chip architectures may not require significant process redesign (Fig. 5).

Carbon nanotubes, which can be visualized as rolled sheets of graphene, have been widely investigated as a promising new material for many electrical device applications [5, 7] (e.g., transistor (CNT-FET), interconnects) as they exhibit exceptional electrical, thermal, and mechanical



Carbon Nanotubes for Chip Interconnections, Fig. 5 Graphical representations of ideal graphene sheet, SWCNT, MWCNT (Figure reprinted with permission from Graham et al. [22])

properties [8]. When comparing materials properties, CNTs are a clear winner against copper. Studies show that CNTs are stable for current densities up to 10^9 A/cm^2, two orders of magnitude higher than copper. CNTs can exhibit multichannel ballistic conduction over distances of microns. Because of their higher chemical stability relative to copper, diffusion barriers (liners) are not needed for CNTs, thus allowing a larger conductive cross section compared to copper for the same technology node. Additionally, their mechanical tensile strength (100 times that of steel) and their high thermal conductivity (comparable to diamond) give CNTs an edge compared to copper. Finally, growing CNTs in high aspect ratio vias could allow the design of chips with higher interlayer spacing to reduce overall RC losses and to decrease chip-layer energy dissipation [9].

Before examining possible models of CNT-based interconnect architectures, it is important to clearly understand CNTs' electrical properties, which represent the most critical material limitation to resolve with respect to copper. The electronic band structures of single-wall CNTs (SWCNTs) and of graphene are very similar. For graphene and metallic SWCNTs, the valence band and the conduction band touch at specific points in the reciprocal space. For semiconducting SWCNTs, the conduction band and the valence band do not touch. Semiconducting SWCNTs have been extensively studied as channels in transistor devices while metallic SWCNTs have been considered for applications such as IC interconnects and field emission.

The resistance of a CNT contacted at both ends is the sum of three resistances [5, 10]:

$$R_{CNT} = R_Q + R_L + R_{CONTACT}$$

where R_Q is the quantum resistance, R_L is the scattering resistance, and $R_{CONTACT}$ is the contact resistance. We will now discuss these three resistances.

An ideal (defect-free) metallic SWCNT electrically contacted at both ends, in the absence of scattering or contact resistance, exhibits a resistance $R = 2R_Q \approx 13\,\text{k}\Omega$ as a SWCNT has two conduction channels. The quantum resistance $R_Q = 6.5\,\text{k}\Omega$ is due to the mismatch between the number of conduction channels in the nanotube and the macroscopic metallic contacts. The one-dimensional confinement of electrons, combined with the requirement for energy and momentum conservation, leads to ballistic conduction over distances in the order of a micron.

The scattering resistance is due to impurities or nanotube defects that reduce the electron mean free path, and depends on the length l of the SWCNT:

$$R_L = \frac{1}{2R_Q}\left(\frac{l}{l_0}\right)$$

For defect-free SWCNT lengths below a micron, we can neglect the scattering resistance.

The contact resistance, which results from connecting the SWCNT to a contact (usually metallic), depends strongly on the material in contact with the nanotube, and on the difference between their work functions. The work functions of multiwall CNTs (MWCNTs) and SWCNTs have been estimated to be 4.95 and 5.10 eV, respectively [10]. Palladium has been found to be one of the materials minimizing the contact resistance, better than titanium or platinum contacts (which exhibited nonohmic behavior when in contact with CNTs) [10].

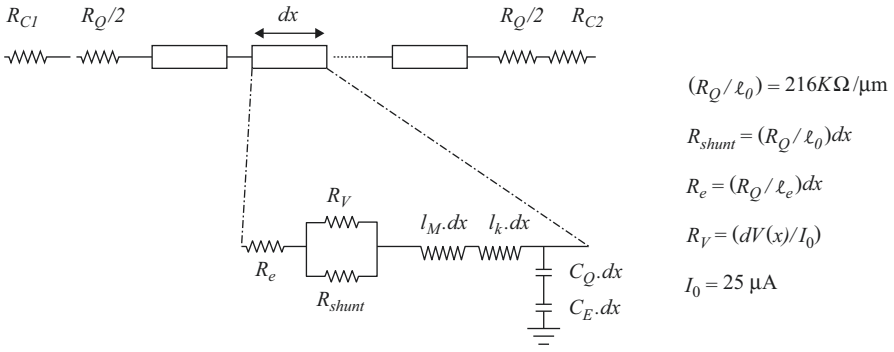
For interconnect applications, most often bundles of SWCNTs are considered. It is important to note that the coupling between adjacent SWCNTs is negligible since, for defect-free SWCNTs, the electrons would rather travel along the SWCNT axis (ballistic path) than across SWCNTs because of the large inter-CNT tunneling resistance (2–140 MΩ) [10]. Thus, the resistance of a bundle of SWCNTs can be viewed as a parallel circuit of the resistances of the individual SWCNTs. If we have n SWCNTs, the resistance of the bundle will be:

$$R_{SCWNT\ bundle} = \frac{R_{SCWNT}}{n}$$

The above overview related to SWCNTs. The electrical properties of MWCNTs have not been as extensively studied because of the additional complexities arising from their structure, as every shell has different electronic characteristics and chirality, in addition to interactions between the shells [11]. Geometrically, the interwall distance in a MWCNT is 0.34 nm, the same as the spacing between graphene sheets in graphite. What still has to be clarified is how the conductivity of a MWCNT varies with the number of walls.

Initially, it was thought that the conductance of a MWCNT occurred only through the most external wall, which seems to be the case at low bias and temperatures, where electronic transport is dominated by outer-shell conduction. However, theoretical models and experimental results indicate that shell-to-shell interactions can significantly lower the resistance of MWCNTs with many walls [5, 10].

One view is that the conductivity of a MWCNT with n walls is simply n times the conductivity of a SWCNT. Li et al. [12] experimentally measured an electrical resistance of only 34.4 Ω for a large MWCNT with outer diameter of 100 nm and inner diameter of 50 nm (= > 74 walls). This value was much lower than the one that could be calculated assuming all walls participated separately in the electrical conduction (i.e., calculated as the parallel of the resistances for each wall) showing that interwall coupling contributes to additional channels of conductance. Naeemi et al. [13] also assumed intercoupling between CNT walls in their models to increase the channels of conduction with increasing number of walls. However, their conductance was lower compared to that measured experimentally by Li's team.



Carbon Nanotubes for Chip Interconnections, Fig. 6 Equivalent circuit model of metallic SWCNT used in HSPICE simulations (Figure reprinted with permission from Naeemi and Meindl [14])

Bottom line: The conductivity of MWCNTs increases with the number of walls but the exact relationship has not yet been exactly clarified.

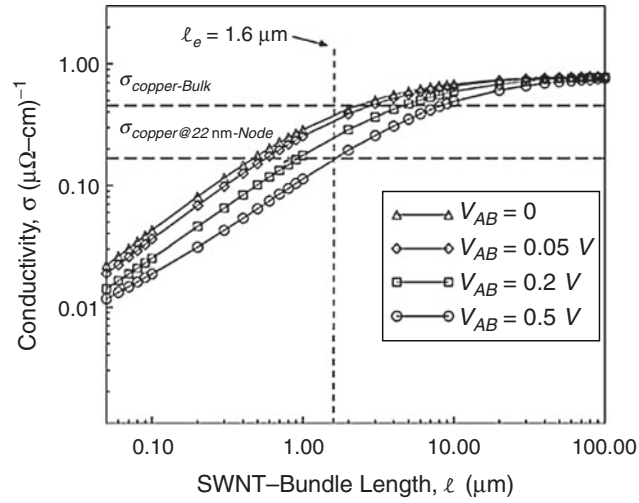
Models of CNTs as Interconnects

Various studies investigated replacing copper interconnects with bundles of CNTs (SWCNTs or MWCNTs) or with one large MWCNT. A first approach consists of using densely packed SWCNTs. Modeling SWCNTs as equivalent electrical circuits and using SPICE simulations, Naeemi et al. [14] showed that a target density of SWCNTs of at least 3.3×10^{13} CNTs/cm^2 was required. Currently achieved maximum densities for CNTs in vias barely reach 10^{12} CNTs/cm^2, which is still an order of magnitude smaller than required. Furthermore, since statistically only one-third of the SWCNTs grown are metallic (the other two-third are semiconducting), the conduction of the bundle will only occur in the metallic SWCNTs (Fig. 6).

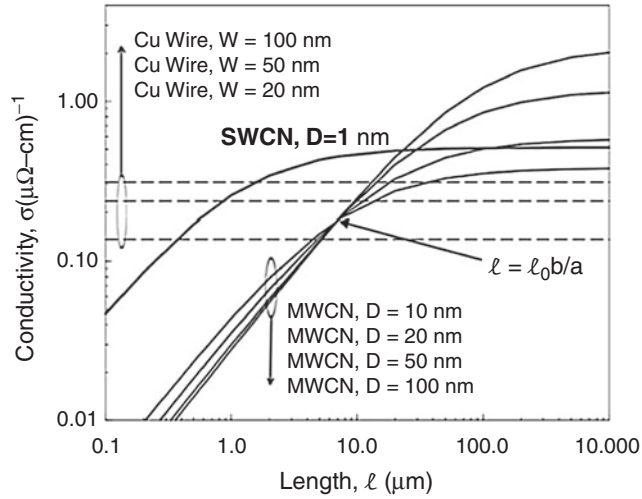
Naeemi et al. [14] also compared SWCNT bundles to copper as local, intermediate, and global interconnects. They showed that, in SWCNT bundles, resistance and kinetic inductance decreased linearly with the number of nanotubes in the bundle, while magnetic inductance changed very slowly. The resistance of a bundle of SWCNTs with sufficient metallic nanotubes was smaller than the resistance of copper wires, while capacitance was comparable. SWCNT bundles also fared better compared to copper in reducing power dissipation, delay, and crosstalk. For local interconnects, they quantified the improvements as 50 % reduction in capacitance, 48 % reduction of capacitance coupling between adjacent lines, and 20 % reduction in delay. For intermediate interconnects, the improvements were more marked, especially in terms of improved conductivities. For global interconnects, dense SWCNT bundles proved critical to improve bandwidth density (Fig. 7).

Using MWCNTs, which are all electrically conductive as they exhibit multiple channels of conduction (compared to only one-third metallic SWCNTs), could lower the resistivity of the bundle, although fewer of them can be packed in the same space because they usually have larger diameters (but also require a lower packing density compared to SWCNTs). In a different modeling study, Naeemi et al. [13] explored the suitability of MWCNTs as replacement for copper interconnects. They concluded that for long lengths (over 100 μm), MWCNTs have conductivities many times that of copper and even of SWCNT bundles. However, for short lengths (less than 10 μm), dense SWCNT bundles can exhibit a conductivity that is twice that of MWCNT bundles. Thus, for via applications, they recommended using dense bundles of SWCNTs or, alternatively, bundles of MWCNTs with small diameter (i.e., with few walls) (Fig. 8).

Using an individual MWCNT with large diameter could offer high conductivity due to the participation of multiple walls to significantly



Carbon Nanotubes for Chip Interconnections, Fig. 7 Conductivity of densely packed SWCNT bundles versus length for various bias voltages (Figure reproduced with permission from Naeemi and Meindl [14])



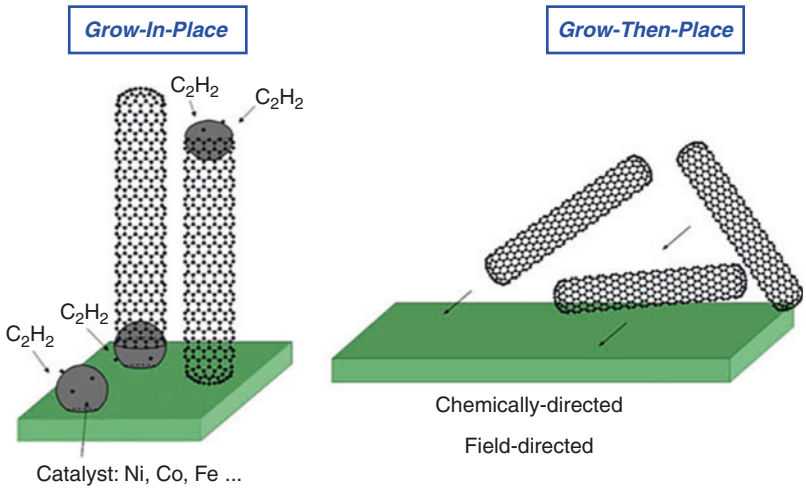
Carbon Nanotubes for Chip Interconnections, Fig. 8 Conductivity of MWCNTs with various diameters compared to Cu wires and dense bundles of SWCNTs (Figure reproduced with permission from Naeemi and Meindl [13])

increase the channels for conduction. However, as previously mentioned, the exact relationship between the number of walls and the conductance has yet to be clarified.

In conclusion, the choice of nanotubes may differ depending on the type of interconnect. For instance, dense bundles of SWCNTs or MWCNTs with few walls may be more suitable for small-section vertical vias, while dense bundles of larger MWCNTs may be more appropriate for long-range interconnects. The option of using a large MWCNT which fills all the space available needs to be further investigated. It is also plausible that hybrid systems of copper/SWCNTs/MWCNTs may be the best solution; for instance, small-section vertical vias may be replaced by dense SWCNT bundles, while larger long-range horizontal interconnects may still use copper, dense bundles of MWCNTs, or even metal-CNT composites.

Practical Implementation: Fabrication State of the Art and Outstanding Issues

Reliable and repeatable high-yield CNT fabrication compatible with CMOS standards is the main bottleneck in replacing copper in chip interconnections. Although hundreds of research teams have focused their efforts on nanotube growth and thousands of papers detailing growth recipes have been



Carbon Nanotubes for Chip Interconnections, Fig. 9 Pictorials comparing the "grow-in-place" and "grow-then-place" techniques (Reproduced with permission from Professor Carl V. Thompson [18])

published, surprisingly, very few have focused on the growth on conductive substrates at CMOS-compatible processing temperature [7, 10, 15, 16]. It is still a challenge to reliably and consistently synthesize CNTs on conductive layers at temperatures below 400–450 °C, the maximum temperature allowed in CMOS fabrication to avoid disrupting previous diffusion patterns. Furthermore, it is still difficult to precisely control CNT diameter and height, although chemical vapor deposition (CVD) from thin films of controlled thickness [17] or from nanoparticles [16] of controlled size has shown encouraging results.

To utilize carbon nanotubes in industrial applications, two main approaches have been considered: *"grow-in-place"* and *"grow-then-place"* [18] (Fig. 9).

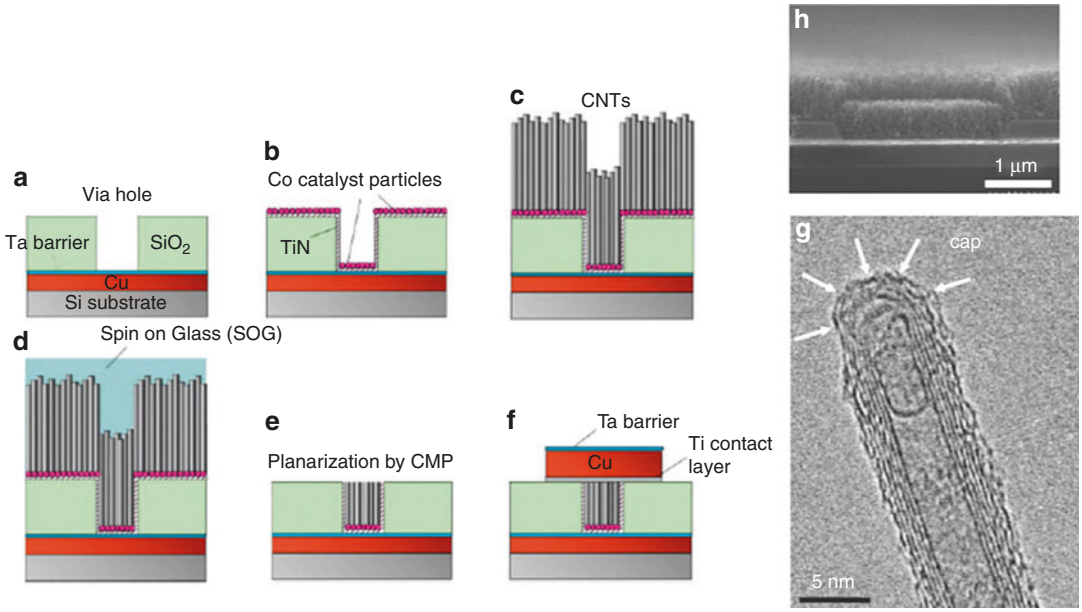
Grow-then-place: This technique consists of first preparing nanotubes and subsequently transferring them to a substrate. Arc discharge and laser ablation are the main techniques used to synthesize free-standing nanotubes. The nanotubes may be subsequently selected (e.g., separating SWCNTs or metallic SWCNTs) and purified prior to use. To transfer them to another substrate, CNTs are usually functionalized in a way that they will attach to pre-patterned areas of the substrate which will attract functionalized CNTs. An interesting technique for interconnect vias is based on using electrophoresis to push CNTs dispersed in a liquid solution into a matrix with pits (e.g., porous alumina matrix).

The advantages of this method are that it places no restrictions on the process or temperature used for CNT synthesis and allows to pretreat the CNTs (e.g., select, purify, functionalize). The major drawback is that no successful and repeatable technique to transfer the CNTs to the substrate has been developed to date. The challenge of resolving this issue appears too high to make this technique a candidate for the CMOS industry. However, free-standing, purified, CNTs are manufactured by many companies and sold for other applications (e.g., CNT-polymer composites).

Grow-in-place: this technique usually consists of preparing the sample with a catalyst present in the locations where the nanotubes will be synthesized. For instance, a thin catalyst film can be deposited using e-beam evaporation or sputtering; alternatively, nanoparticles can be deposited on a substrate. Synthesis is usually performed using thermal or assisted (e.g., plasma) CVD.

This method has several advantages: (1) good control of nanotube position (CNTs will grow where there are catalysts), (2) proven recipes to obtain crystalline CNTs (at least on insulating substrates), (3) proven capabilities to obtain carpets of vertically aligned CNTs, (4) physical contact with the substrate, (5) electrical contact with the substrate, and (6) CVD techniques are commonplace in the CMOS industry.

The major drawbacks are that (1) the processing temperature should be below 400–450 °C (CMOS-compatibility), thus putting



Carbon Nanotubes for Chip Interconnections, Fig. 10 Process to synthesize CNTs into pits, SEM cross section, and TEM showing crystalline MWCNT (Reprinted with permission from Yokoyama et al. [23])

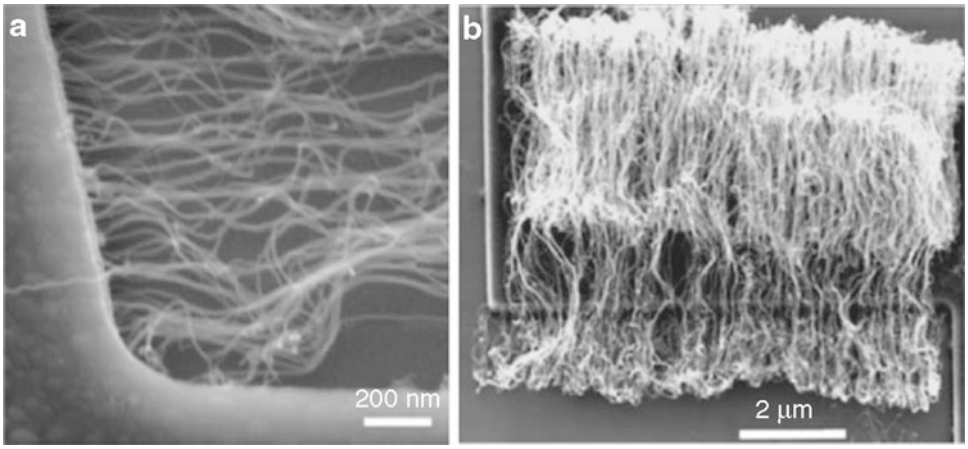
serious limits on the synthesis method and (2) the CNTs should be directly synthesized on the substrate of choice, usually a metallic layer to provide electrical contact. Although growing dense carpets of crystalline CNTs on insulating substrates such as alumina or silicon oxide has been achieved by many, CNT growth on metallic layers still remains a serious challenge. Interactions between the catalyst and the metallic substrate (e.g., alloying) are the major impediments for the successful growth of dense carpets of CNTs on metallic layers.

In addition to interesting results obtained from university research, good progress on the growth and characterization of carbon nanotubes for interconnects has been achieved by industrial laboratories, initially by Infineon and now by the Fujitsu laboratories. In 2002, Kreupl et al. [19] of Infineon showed that bundles of CNTs could be grown in pits of defined geometry.

More recently, Awano et al. [7] of Fujitsu grew bundles of MWCNTs in a 160 nm via at 450 °C and measured an electrical resistance of 34Ω (the CNT density observed was 3×10^{11} CNTs/cm^2). This follows a previous result obtained earlier by the same team where they grew MWCNTs into a 2 μm via at temperatures close to 400 °C with the lowest resistance measured of 0.6Ω after CMP and annealing in a hydrogen atmosphere [5] (Fig. 10).

Kreupl et al. [19] succeeded in growing a single MWCNT into a 25-nm hole and measured a high resistance of 20–30 kΩ. Given the difficulty of growing a single large MWCNT and the difficult task of making a precise electrical measurement, there may be room for further improvement if we could grow an individual, crystalline (defect-free) MWCNT with the maximum number of walls for a given external diameter, thus maximizing the number of channels of conduction.

Most of the effort on CNT synthesis to replace copper interconnects has focused on vertical growth of dense carpets of CNTs, which can be achieved by a high density of active catalyst dots. In contrast there have been fewer successful reports of horizontal growth, with less spectacular results. Many techniques have been used to achieve horizontal alignment among which we can mention high gas flow rates, electric fields, and epitaxial techniques to guide horizontal alignment of the nanotubes [10]. In 2010, Yan et al. [20] obtained an interesting horizontal



Carbon Nanotubes for Chip Interconnections, Fig. 11 Scanning electron microscope images of dense carpets of horizontally aligned CNTs grown using CVD (Figure reproduced from Yan et al. [20])

growth of bundled CNTs with a density of 5×10^{10} CNTs/cm^2, which is approaching what has been achieved for vertical CNT growth (although still over an order of magnitude lower compared to the best result for vertical CNT growth) (Fig. 11).

Although the experimental results obtained are encouraging, there are still numerous challenges that need to be resolved for CNTs to enter the semiconductor fab:

1. Increase the CNT areal density by one or two orders of magnitude. For SWCNTs, assuming all of them are metallic, a packing density of 10^{13}–10^{14} CNTs/cm^2 is required to compete with copper in terms of resistance, while for MWCNTs, the required packing density is lower and depends on the number of channels of conduction (i.e., number of walls). This will require, among other considerations, adequate catalyst and underlayer materials choice and deposition, possible surface pretreatment (e.g., plasma, reduction, and etching), maximum nucleation of active catalyst dots, and optimizing the CNT growth process. Although CNT areal density is an important issue, it may not be the dealbreaker.
2. Minimize the contact resistance between the CNTs and the substrate. To achieve maximum conductivity, the choice of the appropriate underlayer is critical; specific metals (e.g., Pd) and possibly silicides are good candidates. For MWCNTs, it is also important to ensure electrical contact with all the walls.
3. When using SWCNTs, synthesize only metallic SWCNTs (on average one-third of the SWCNTs grown) which are the ones participating in the electrical conduction. This is closely linked to the issue of chirality control, for which no solution has been proposed to date. Selective catalyst choice may provide an alternative avenue to synthesize a higher fraction of metallic SWCNTs.
4. Control growth direction of CNTs. This is especially challenging for horizontal interconnects where the directionality and the packing density achieved are still lagging compared to vertical growth of CNTs, despite some interesting progress in this area [10, 20].
5. Synthesize crystalline, defect-free CNTs to ensure maximum electrical conductivity in the nanotube. This is challenging, especially when combined with the requirement of growing CNTs at low temperature to achieve CMOS compatibility.
6. Synthesize CNTs at temperatures below 400–450 °C to ensure CMOS compatibility.
7. Repeatably yield the same CNT structures when the same process conditions are applied. This is a major issue since important variations in the structure and shape of the CNTs grown have been experimentally observed.

Conclusions and Outlook for CNTs as Chip Interconnections

In the past decade, the synthesis of CNTs and the understanding of their growth mechanisms have massively improved. However, for CNTs to enter the CMOS fab and replace copper, significant challenges still need to be resolved. In my opinion, the most significant challenge to overcome is developing a reliable and repeatable fabrication process consistent with CMOS conditions. To achieve that tall order, we need to improve our understanding of the CNT growth mechanisms. Although many simulation models and many experimentally-based insights have been achieved [10], there are still many questions related to CNT growth mechanisms that have not been fully answered. For instance:

- Which precursor gases favor CNT growth and which gases hinder CNT growth? What is the role of the gases in the resulting level of crystallinity of the CNTs grown? Could we pretreat the gases to improve the CNT yield or structure?
- What is the exact role of the catalyst? How does its materials properties and its lattice structure influence the resulting CNTs grown (in shape and structure)?
- What is the role of the underlayer (layer below the catalyst) and its interactions with the catalyst? Why is it so challenging to grow CNTs on metallic layers?

In addition to improving our mechanistic understanding of CNT growth, I believe that a parallel effort focused on developing better reactors is needed. Most researchers use standard CVD-based systems that were designed for a general purpose. A customized reactor, where the same growth conditions can be repeatably achieved with very small variations could provide the repeatability in results (CNT structures) that has eluded us so far.

Carbon nanotubes are already becoming a manufacturing reality in mechanical engineering applications (e.g., CNT-based composites) and many interesting results have been obtained to develop novel CNT structures for electrical applications. Although the jury is still out, if process repeatability could be achieved, we could hope, not only that CNTs will enter the CMOS fabs and replace copper for chip interconnections, but also that they will lead to innovative ventures requiring lower investments to develop integrated circuits with radically new architectural designs using carbon nanotubes as new building blocks.

Cross-References

- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes
- ▶ Chemical Vapor Deposition (CVD)
- ▶ CMOS-CNT Integration
- ▶ Nanotechnology
- ▶ Physical Vapor Deposition
- ▶ Synthesis of Carbon Nanotubes

References

1. Kuhn, K.J.: Moore's Law Past 32 nm: Future Challenges in Device Scaling. Intel Publication, Hillsboro (2009)
2. Magen, N., Kolodny, A., Weiser, U.: Interconnect-power dissipation in a microprocessor. In: Proceedings of the 2004 International Workshop, Paris, 1 Jan 2004
3. ITRS. International Technology Roadmap for Semiconductors – Interconnect 2009, International Sematech, Austin
4. Goel, A.K.: High-Speed VLSI Interconnections, 2nd edn. Wiley/IEEE, Hoboken (2007)
5. Nessim, G.D.: Carbon Nanotube Synthesis for Integrated Circuit Interconnects. Massachusetts Institute of Technology, Cambridge, MA (2009)
6. Havemann, R.H., Hutchby, J.A.: High-performance interconnects: an integration overview. Proc. IEEE **89**(5), 586–601 (2001)
7. Awano, Y., Sato, S., Nihei, M., Sakai, T., Ohno, Y., Mizutani, T.: Carbon nanotubes for VLSI: interconnect and transistor applications. Proc. IEEE **98**(12), 2015–2031 (2010)
8. Dresselhaus, M.S., Dresselhaus, G., Avouris, P. (eds.): Carbon Nanotubes: Synthesis, Structure, Properties, and Applications. Springer, Berlin (2001)
9. Chen, F., Joshi, A., Stojanović, V., Chandrakasan, A.: Scaling and evaluation of carbon nanotube interconnects for VLSI applications. In: Nanonets Symposium 07, Catania, 24–26 Sept 2007
10. Nessim, G.D.: Properties, synthesis, and growth mechanisms of carbon nanotubes with special focus on thermal chemical vapor deposition. Nanoscale **2**(8), 1306–1323 (2010)

11. Collins, P.G., Avouris, P.: Multishell conduction in multiwalled carbon nanotubes. Appl. Phys. **74**(3), 329–332 (2002)
12. Li, H.J., Lu, W.G., Li, J.J., Bai, X.D., Gu, C.Z.: Multichannel ballistic transport in multiwall carbon nanotubes. Phys. Rev. Lett. **95**(8), 086601 (2005)
13. Naeemi, A., Meindl, J.D.: Compact physical models for multiwall carbon-nanotube interconnects. IEEE Electron. Device Lett. **27**(5), 338–340 (2006)
14. Naeemi, A., Meindl, J.D.: Design and performance modeling for single-walled carbon nanotubes as local, semiglobal, and global interconnects in gigascale integrated systems. IEEE Trans. Electron Devices **54**(1), 26–37 (2007)
15. Nessim, G.D., Seita, M., O'Brien, K.P., Hart, A.J., Bonaparte, R.K., Mitchell, R.R., Thompson, C.V.: Low temperature synthesis of vertically aligned carbon nanotubes with ohmic contact to metallic substrates enabled by thermal decomposition of the carbon feedstock. Nano Lett. **9**(10), 3398–3405 (2009)
16. Awano, Y., Sato, S., Kondo, D., Ohfuti, M., Kawabata, A., Nihei, M., Yokoyama, N.: Carbon nanotube via interconnect technologies: size-classified catalyst nanoparticles and low-resistance ohmic contact formation. Phys. Status Solidi **203**(14), 3611–3616 (2006)
17. Nessim, G.D., Hart, A.J., Kim, J.S., Acquaviva, D., Oh, J.H., Morgan, C.D., Seita, M., Leib, J.S., Thompson, C.V.: Tuning of vertically-aligned carbon nanotube diameter and areal density through catalyst pre-treatment. Nano Lett. **8**(11), 3587–3593 (2008)
18. Thompson, C.V.: Carbon nanotubes as interconnects: emerging technology and potential reliability issues. In: 46th International Reliability Symposium; 2008: IEEE CFP08RPS-PRT, p. 368, 2008
19. Kreupl, F., Graham, A.P., Duesberg, G.S., Steinhogl, W., Liebau, M., Unger, E., Honlein, W.: Carbon nanotubes in interconnect applications. Microelectron. Eng. **64**(1–4), 399–408 (2002)
20. Yan, F., Zhang, C., Cott, D., Zhong, G., Robertson, J.: High-density growth of horizontally aligned carbon nanotubes for interconnects. Phys Status Solidi. **247** (11–12), 2669–2672 (2010)
21. Jackson, R.L., Broadbent, E., Cacouris, T., Harrus, A., Biberger, M., Patton, E., Walsh, T.: Processing and integration of copper interconnects. In: Solid State Technology. Novellus Systems, San Jose (1998)
22. Graham, A.P., Duesberg, G.S., Hoenlein, W., Kreupl, F., Liebau, M., Martin, R., Rajasekharan, B., Pamler, W., Seidel, R., Steinhoegl, W., Unger, E.: How do carbon nanotubes fit into the semiconductor roadmap? Appl. Phys. **80**, 1141–1151 (2005). Copyright 2005, Springer Berlin/Heidelberg
23. Yokoyama, D., Iwasaki, T., Yoshida, T., Kawarada, H., Sato, S., Hyakushima, T., Nihei, M., Awano, Y.: Low temperature grown carbon nanotube interconnects using inner shells by chemical mechanical polishing. Appl. Phys. Lett. **91**, 263101 (2007). Copyright 2007, American Institute of Physics

Carbon Nanotubes for Interconnects in Integrated Circuits

▶ Carbon Nanotubes for Chip Interconnections

Carbon Nanotubes for Interconnects in Microprocessors

▶ Carbon Nanotubes for Chip Interconnections

Carbon Nanowalls

▶ Chemical Vapor Deposition (CVD)

Carbon-Nanotubes

▶ Robot-Based Automation on the Nanoscale

Car–Parrinello Molecular Dynamics

Mauro Boero[1] and Atsushi Oshiyama[2]
[1]Institut de Physique et Chimie des Materiaux de Strasbourg (IPCMS), University of Strasbourg and CNRS, UMR 7504, Strasbourg, France
[2]Department of Applied Physics, The University of Tokyo, Tokyo, Japan

Synonyms

Ab initio molecular dynamics; CPMD; DFT-based molecular dynamics; First principles molecular dynamics

Definition

The Car–Parrinello molecular dynamics (CPMD) is an extension of the Lagrangian formalism of

classical molecular dynamics in which the model potential describing the interaction among atoms is replaced by the total energy functional of the system as provided by the Density Functional Theory (DFT). The electronic wavefunctions are explicitly introduced as new dynamical variables. The simultaneous Euler-Lagrange equations of motion for both sets of dynamical variables, atomic coordinates and electronic wavefunctions, avoid the explicit minimization of the DFT total energy at each step of the dynamics. Instead, they introduce a fictitious dynamics of the wavefunctions representing an adiabatic updating on-the-fly of the electronic structure along the atomic dynamics.

Introduction

The main target in atomic-scale simulations is to reproduce in a realistic way physical and chemical events occurring in materials. Specifically, the scope of First Principles Molecular Dynamics (FPMD) is to study a system of *interacting* nuclei and electrons by recreating it on a computer in a way as close as possible to nature and by simulating its dynamics over a *physical length of time* relevant to the properties of interest. The inherent complexity of the simulated systems, from solids to biological macromolecules, calls for methods able to go beyond the simple calculation of the electronic structure of a given set of coordinates $\mathbf{R}_I$ representing the positions of atoms. This is exactly the idea that started the entire field of Molecular Dynamics (MD).

From an historical point of view, the MD approach was introduced by Alder and Wainwright [1] in the late 1950s to study the interactions of hard spheres. Many important insights concerning the behavior of simple liquids emerged from their studies, but due to the limitations of the computational facilities and the pioneering stage of the MD, it was only in 1964 that the first dynamical simulation could be done. That milestone case focused on liquid Ar with the interatomic interaction modeled by a truncated Lennard-Jones potential [2]. In a nutshell, any MD method is an iterative numerical scheme for solving some equations of motion (EOM), representing the physical evolution of the system under study. Modeling the interaction of atoms with an analytic potential $V(\mathbf{R}_I)$, especially when chemical bonds evolve in time and they are broken or formed is a hard task not yet solved apart from a very limited class of chemical species. On the other hand, the electronic structure for a general many-body system can be determined with a computationally reasonable workload by means of the density functional theory (DFT), originally proposed in the early 1960s by Kohn, Hohenberg, and Sham [3, 4]. Its importance in the advancement of computational quantum chemistry and related fields was acknowledged by the Nobel Prize in Chemistry in 1998 awarded jointly to Walter Kohn and John A. Pople. Joining the two fields, MD and DFT, is exactly what the Car–Parrinello method is about, extending the range of both concepts [5, 6].

A Brief Overview of Density Functional Theory: The CPMD Potential

The DFT is a formulation of the many-body quantum mechanics in terms of an electron density distribution, $\rho(\mathbf{x})$, which describes the ground state of a system composed of interacting electrons and point-like nuclei having positions $\{\mathbf{R}_I\}$ [7]. All along the text, atomic units will be used for simplicity. In practice, single-particle wavefunctions $\psi_i(\mathbf{x})$ are used to express the many-body mathematical function ρ $(\mathbf{x})$. The dramatic simplification, is the fact that not even the specific analytic form of the complex function $\psi_i(\mathbf{x})$ matters, but only its square modulus, so that the electron density reads

$$\rho(\mathbf{x}) = \sum_{i=1}^{N^{occ}} f_i |\psi_i(\mathbf{x})|^2 \qquad (1)$$

This expression is clearly a single Slater determinant constructed from wavefunctions representing all the N^{occ} occupied orbitals. The coefficients f_i are the (integer) occupation numbers, and they are equal to 1 in the case in which the spin is explicitly

considered (spin-unrestricted) or equal to 2 if the spin is neglected and energy levels are considered as doubly-occupied (spin-restricted). Furthermore, the wavefunctions $\psi_i(\mathbf{x})$ are subject to the orthonormality constraint

$$\int \psi_i^*(\mathrm{x})\psi_j(\mathrm{x})\, d^3x = \delta_{ij} \tag{2}$$

as in any quantum mechanics approach. The Kohn-Sham (KS) DFT total energy of the system in its ground state is then written as

$$E^{KS}[\{\psi_i\}] = E_k[\{\psi_i\}] + E_H[\rho] + E_{xc}[\rho] + E_{eI}[\rho] + E_{II} \tag{3}$$

In Eq. 3, the first three terms on the right-hand side (E_k, E_H, E_{xc}) describe all the electron–electron interactions, the fourth term (E_{eI}) refers to the electron–nucleus interaction, and the fifth one (E_{II}) corresponds to the nucleus–nucleus interaction. More explicitly, E_k is the Schrödinger-like kinetic energy expressed in terms of the single-particle wavefunctions $\psi_i(\mathbf{x})$ as

$$E_k[\{\psi_i\}] = \sum_{i=1}^{N^{occ}} f_i \int \psi_i^*(\mathbf{x}) \left(-\frac{1}{2}\nabla^2\right) \psi_i(\mathbf{x})\, d^3x \tag{4}$$

It must be remarked that this expression for the kinetic energy does not depend on the density $\rho(\mathbf{x})$ but directly on the wavefunctions. The second term, E_H, is the Hartree energy, i.e., the Coulomb electrostatic interaction between two charge distributions

$$E_H[\rho] = \frac{1}{2}\int\int \frac{\rho(\mathbf{x})\rho(\mathbf{y})}{|\mathbf{x}-\mathbf{y}|} d^3x\, d^3y \tag{5}$$

The exchange interaction and the electron correlations due to many-body effects are represented by the term $E_{xc}[\rho]$, whose exact analytical expression is unknown. There are good approximations derived from the homogeneous electron gas limit for the exchange interaction [7], the so-called local density approximation (LDA), whose name comes from the fact that a homogeneous distribution of interacting electrons is assumed, in which $\rho(\mathbf{x})$ depends just on the local point $\mathbf{x}$. Similarly, in the LDA version of the correlation energy [7], the explicit analytic form of the functional comes from a parameterization of the results of quantum Monte Carlo calculations. Due to the insufficiency of a simple LDA approximation for many real systems, nonlocal approximations, including the gradient of the density, are often adopted and the exchange-correlation functional becomes $E_{xc}[\rho,\ \nabla\rho]$. In practical applications, however, the gradient enters only with its modulus, thus adding only a modest computational cost. The electrostatic interaction between electrons and nuclei, is then

$$E_{eI}[\rho] = -\int \sum_{I=1}^{M} \frac{Z_I \rho(\mathbf{x})}{|\mathbf{x}\text{-}\mathbf{R}_I|} d^3x \tag{6}$$

where Z_I is the charge of the Ith nucleus. However, in practice, this expression "as is" is computationally expensive. In fact, two different length scales come into play: a small one for the core electrons, characterized by rapidly varying wavefunctions, especially in the region very close to the nucleus, and a longer one for the valence electrons that form chemical bonds and vary more smoothly. Clearly, the first one would dominate and add a computational workload that would make impractical simulations of large systems. To overcome this problem, one can observe that core electrons are generally inert and do not participate to chemical bonds. This crucial observation led to the use of pseudopotentials [6]. Namely, core electrons are eliminated and a potential describing the core–valence interaction is built by fitting to the all-electron solutions of the Schrödinger or Dirac equation for the single atom of the chemical species considered. In a pseudopotential (PP) approach, the electron–nucleus interaction is rewritten as

$$E_{eI}[\rho] = \int d^3x V_{ps}(\mathbf{x} - \mathrm{R}_I) \cdot \rho(\mathbf{x}) \tag{7}$$

Finally, the fifth and last term in right-hand side of Eq. 3 is simply the Coulomb interaction between two classical nuclei I and J and is written as

$$E_{II} = \sum_{I<J}^{M} \frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} \tag{8}$$

where Z_I and Z_J are the net valence charge in a PP approach. The total energy E^{KS} of the ground state of such a system of interacting electrons and nuclei is obtained by minimizing the KS functional with respect to the single-particle orbitals $\psi_i(\mathbf{x})$, which, in practice, means solving the KS Schrödinger-like equations

$$\frac{\delta E^{KS}}{\delta \psi_i^*} \equiv H^{KS}\psi_i(\mathbf{x}) = \varepsilon_i \psi_i(\mathbf{x}) \tag{9}$$

The Basis Set Issue

A somehow arbitrary issue is the proper definition of $\psi_i(\mathbf{x})$. The answer is the selection of a proper basis set on which orbitals can be expanded. One possible choice is a localized basis set expressing the one-electron wavefunctions as

$$\psi_i(\mathbf{x}) = \sum_{k=1}^{M} c_i^k \phi_k(\mathbf{x}; \{\mathbf{R}_I\}) \tag{10}$$

and the number of analytic functions used, M, is also an indicator of the computational cost of the quantum calculation, in the obvious sense that the larger the basis set, the higher the computational workload. One of the most popular basis sets is represented by Gaussian-type orbitals (GTO)

$$\phi_k(\mathbf{x} - \mathbf{R}_I) = \phi_k(\mathbf{r}) = N_k \cdot r_x^{k_x} r_y^{k_y} r_z^{k_z} \cdot \exp\left(-\alpha_k \cdot r^2\right) \tag{11}$$

where $\mathbf{r} = \mathbf{x} - \mathbf{R}_I$. When such a basis set is used, the constants $N_\mathbf{k}$, and $\alpha_\mathbf{k}$ are kept fixed during the electronic structure calculation, whereas the coefficients c_i^k are allowed to vary until they are fully optimized [8]. It must be remarked that orbitals expanded in a localized basis set depend on the atomic positions $\mathbf{R}_I$. As a consequence, in any calculation in which the forces acting on the ions are required, the explicit derivatives of these wavefunctions with respect to $\mathbf{R}_I$ must be computed, leading to non-Hellmann-Feynman force components known in the literature as Pulay forces [6, 8]. An alternative basis set rather popular in physics is represented by plane waves (PW)

$$\psi_i(\mathbf{x}) = \sum_{\mathrm{G}=0}^{\mathrm{G}^{\max}} c_i(\mathbf{G})\, e^{i\mathbf{G}\cdot\mathbf{x}} \tag{12}$$

where the sum is truncated at a suitable cut-off $\mathbf{G}^{\max}$. In this case, no dependence on the atomic coordinates and no arbitrariness in the increase in the number of basis functions exist.

First Principles Molecular Dynamics

Until the early 1980s, few applications of DFT went beyond the static calculations of the electronic structure. Nonetheless, finite temperature and entropy effects are two of the dominant features in matter and their role is often far from negligible. In this respect, the FPMD has represented a huge step forward. In this particular combination of DFT and MD, the interactions among atoms, instead of being described by an analytical function $V(\{\mathbf{R}_I\})$ of the atomic coordinates $\mathbf{R}_I$, is directly computed from the DFT total energy E^{KS}, which is simultaneously a function of the electron wavefunctions and the atomic coordinates. The Born–Oppenheimer (BO) approximation [9] allows to disentangle the motions of the electrons and the nuclei, and each time the nuclei $\mathbf{R}_I(t)$ are displaced from given positions at time t to new positions $\mathbf{R}_I(t + \Delta t)$ at a subsequent time $t + \Delta t$, an optimization of the electronic structure has to be performed. Then the forces acting on the nuclei are estimated from the gradient of E^{KS} with respect to the ionic position and the variables $\mathbf{R}_I(t)$ are updated to $\mathbf{R}_I(t + \Delta t)$ by solving via finite differences the Newton-like EOM

$$M_I \ddot{\mathbf{R}}_I = -\nabla_{\mathbf{R}_I} \min_{\{\psi_i\}} E^{KS}[\{\psi_i\}, \{\mathbf{R}_I\}] \tag{13}$$

This iterative procedure assumes that the electronic structure is recomputed and the full diagonalization of the Hamiltonian is performed at each time step t along the discrete trajectory $\{\mathbf{R}_I(t)\}$.

Car–Parrinello Molecular Dynamics

An alternative to this scheme, which has represented a real breakthrough in first principles dynamical simulations, was proposed in 1985 by R. Car and M. Parrinello [5]. The scope (and driving force) was to overcome the two major efforts arising in FPMD: On one hand one has to integrate the equations of motion for the nuclear positions as in Eq. 13, which represent the long timescale part to the problem. On the other hand, one has to propagate dynamically the smooth time-evolving (ground state) electronic subsystem. The Car–Parrinello molecular dynamics (CPMD) is able to satisfy this second requirement in a numerically stable way and makes an acceptable compromise for the time step length of the nuclear motion. The formulation is an extension of a classical molecular dynamics Lagrangian in which the electronic degrees of freedom (wavefunctions) are added to the system, along with any other dynamical variable $q_\alpha(t)$, i.e., thermostats, barostats, reaction coordinates, etc.

$$\mathcal{L}^{CP} = \frac{1}{2}\sum_I M_I \dot{\mathbf{R}}_I^{\,2} + \sum_i \mu \int \left|\dot{\psi}_i(\mathbf{x})\right|^2 d^3x + \frac{1}{2}\sum_\alpha \eta_\alpha \dot{q}_\alpha^2 - E^{KS}[\rho, \{\mathbf{R}_I\}, q_\alpha] + \sum_{ij}\lambda_{ij}\left(\int d^3x \psi_i^*(\mathrm{x})\psi_j(\mathrm{x}) - \delta_{ij}\right) \quad (14)$$

The first term on the right-hand side of Eq. 14 is the kinetic energy of the nuclei, the second one the fictitious kinetic energy of the electrons representing the update of the wavefunctions during the dynamics, the third one the kinetic term of any further dynamical variable used in the sense specified above, the fourth one is the DFT total energy, and the last addendum is the orthonormality constraint for the wavefunctions. The kinetic energy for the electronic degrees of freedom is the main novelty of the CPMD approach: A strategy to update on-the-fly the wavefunctions when ions undergo a dynamical displacement, avoiding expensive iterative diagonalization required by the BO approach at each time step. The Euler–Lagrange EOM are as follows:

$$\mu\ddot{\psi}_i(\mathbf{x}) = -\frac{\delta E^{KS}}{\delta \psi_i^*} + \sum_j \lambda_{ij}\psi_j(\mathbf{x}) \quad (15)$$

$$M_I\ddot{\mathbf{R}}_I = -\nabla_{\mathbf{R}_I} E^{KS} \quad (16)$$

$$\eta_\alpha \ddot{q}_\alpha = -\frac{\partial E^{KS}}{\partial q_\alpha} \quad (17)$$

The fictitious mass μ assigned to the orbitals $\psi_i(\mathbf{x})$ is the parameter that controls the speed of the updating of the wavefunctions with respect to the nuclear positions. For this reason, it determines the degree of adiabaticity of the two subsystems, electrons and nuclei.

It is straightforward to give a Hamiltonian, instead of a Lagrangian, formulation of the CPMD method, via a simple Legendre transform after defining the momenta

$$\pi_i(\mathbf{x}) = \frac{\delta \mathcal{L}^{CP}}{\delta \dot{\psi}_i}(\mathbf{x}) = \mu\dot{\psi}_i(\mathbf{x})$$

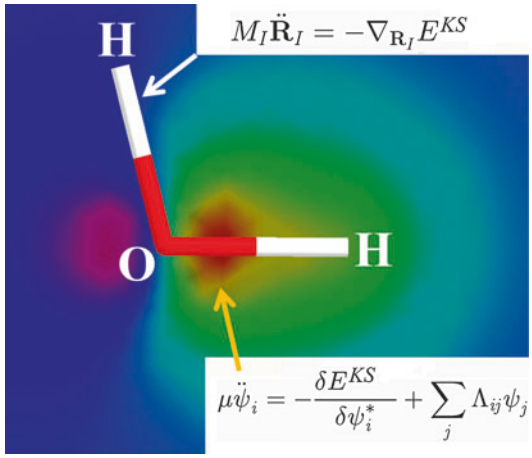
$$\pi_i^*(\mathbf{x}) = \frac{\delta \mathcal{L}^{CP}}{\delta \dot{\psi}_i(\mathbf{x})} = \mu\dot{\psi}_i^*(\mathbf{x}) \quad (18)$$

$$\mathbf{p}_i = \nabla_{\dot{\mathbf{R}}_I}\mathcal{L}^{CP} = M_I\dot{\mathbf{R}}_I \quad (19)$$

$$\dot{\zeta}_\alpha = \frac{\partial \mathcal{L}^{CP}}{\partial \dot{q}_\alpha} = \eta_\alpha \dot{q}_a \quad (20)$$

so that the Hamiltonian reads

$$H^{CP} = \sum_I \frac{\dot{\mathbf{p}}_I^{\,2}}{2M_I} + \sum_i \int \frac{\pi_i^*(\mathbf{x})\,\pi_i(\mathbf{x})}{\mu} d^3x + \sum_\alpha \frac{\dot{\zeta}_\alpha^2}{2\eta_\alpha} - E^{KS}[\rho, \{\mathbf{R}_I\}, q_\alpha] - \sum_{ij}\lambda_{ij}\left(\int \psi_i^*(\mathbf{x})\psi_{\mathrm{j}}(\mathbf{x})\, d^3x - \delta_{ij}\right) \quad (21)$$



Car–Parrinello Molecular Dynamics, Fig. 1 Atomic structure (sticks; *red* = O, *white* = H) of a water molecule and electronic wavefunction of an O–Hs-bond in terms of density map (*blue*: $|\psi_i(\mathbf{x})|^2 = 0$, *red*: $|\psi_i(\mathbf{x})|^2 =$ maximum). The two sets of dynamical variable evolve in time according to the equations of motions indicated, coupled via E^{KS}

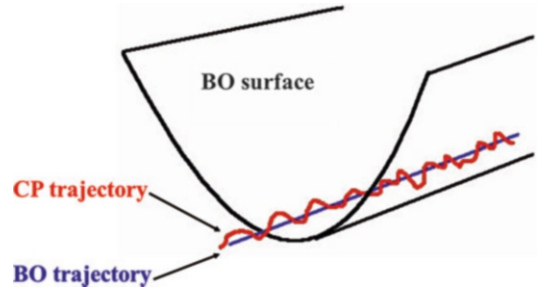
and the CPMD equations of motion (Fig. 1) will be given by the corresponding Hamilton EOMs.

A rigorous mathematical proof of this scheme has been given by Bornemann and Schütte [10], showing that the CPMD trajectory $\{\mathbf{R}^{\mathrm{CP}}(t)\}$ stays close to the BO one $\{\mathbf{R}^{\mathrm{BO}}(t)\}$ and the upper bound is proportional to the square root of the fictitious mass μ

$$\left|\mathrm{R}^{CP}(t) - \mathrm{R}^{BO}(t)\right| < C \cdot \sqrt{\mu} \qquad (22)$$

where C is a positive constant (Fig. 2).

The fact that the CPMD was a milestone step forward in realistic simulations of materials at various thermodynamics conditions can be easily seen by the number of publications in first principles molecular dynamics (FPMD) before and after 1985. Indeed, the original Car–Parrinello publication has more than 7,000 citations in 2014 (source: ISI Web of Science), and, to acknowledge the importance of the method, the international Physics and Astronomy Classification Scheme (PACS) introduced in 1996 a new identification number, 71.15.Pd, to classify Car–Parrinello related publications. Since then, the method has been applied to a wide variety of materials, ranging from solids, to liquids, and to biological systems [11, 12].



Car–Parrinello Molecular Dynamics, Fig. 2 Schematic representation of a Car–Parrinello trajectory (*red line*) with respect to a Born–Oppenheimer dynamics (*blue line*) one a given DFT-based potential energy surface

Numerical Details

Although it is not restriction neither of DFT [13] nor of CPMD [14], PWs are often used as a convenient basis set for the coding of CPMD, since they have several good properties: (i) Accuracy can be systematically improved in a fully variational way, (ii) PWs are independent from atomic positions (i.e., no Pulay forces [6]), (iii) PWs can be easily distributed in parallel processing. However, it must be observed that the fact that PWs are not localized can lead to inefficiencies for small clusters or surfaces placed in a large simulation cell. The equations of motion are discretized by finite differences, via a Verlet or velocity-Verlet algorithm [15]. The ionic variables $\mathbf{R}_I$(t) are updated at a rate Δt, while the electronic degrees of freedom are updated at a rate $\Delta t/\mu^{1/2}$, i.e.,

$$\begin{aligned}\frac{\mu}{\Delta t^2}\left[c_i(\mathbf{G}, t+\Delta t) + (\mathbf{G}, t-\Delta t) - (\mathbf{G}, t)\right] \\ = -\sum_{G'} \left\langle \mathbf{G}\middle| H^{CP} \middle|\mathbf{G}'\right\rangle c_i(\mathbf{G}') \\ + \sum_j \Lambda_{ij} c_j(\mathbf{G})\end{aligned} \qquad (23)$$

For most of the applications, Δt and μ fall in the range 3–5 au and 300–600 au, respectively. Of course, the (quantum) time scale of electrons is dominating in this kind of approaches and simulations times are of the order of few tens or, at very best, hundreds of ps. As far as the system size is

concerned, with N electrons and $\mathbf{G}^{\max}$ PWs, $\mathbf{G}^{\max}$ being integer, the scaling of the various parts composing the CPMD algorithm is $O(N\ \mathbf{G}^{\max})$ for the kinetic term, $O(N\ \mathbf{G}^{\max}\ \log\ \mathbf{G}^{\max})$ for the local potential and $O(N^2\ \mathbf{G}^{\max})$ for both the nonlocal term and wavefunctions orthogonalization procedure.

Second-Generation Car–Parrinello Molecular Dynamics

The inherent high computational cost associated to the electronic structure calculations has limited the affordable timescales for several years. Only the most advanced high performing computer platforms recently available have allowed to increase the system size to about a thousand of atoms and simulation times towards hundreds of picosecond. Yet, many phenomena still call for a substantial boost. These are, for instance, diffusion in solids or, in the case of glasses generation from the melt, a less rapid cooling rate suitable to avoid numerically induced structural problems. While linear scaling methods can be a viable way to access larger system sizes, they still have to face the problem of the simulation timescale. Moreover, the crossover point at which linear scaling methods become advantageous has remained fairly large, especially if high accuracy is needed. An interesting attempt at overcoming these limitations has been proposed in 2007 [16]. The basic idea is to join the advantages of both the BO approach and the CPMD; in a "nutshell"

	CPMD	BO
Conservation of constants of motion	**Good**	Convergence dependent
Electronic optimization	**Not needed**	Needed
Hamiltonian diagonalization	**Not needed**	Needed
Integration step Δt	Small	**Large**
Minimum of the BO surface	Approximate	**Exact**

These two approaches have nearly complementary features as sketched above. Following the CPMD formulation, it can be remarked that the Lagrangian formulation for the propagation of the wavefunctions is stable by construction, thus providing a reliable integration. This *stability* feature must then be preserved. Concerning the *efficiency*, large integration steps Δt are desirable and possibly a small, or better zero, deviation from the BO surface should be kept all along the dynamics to get a high *accuracy*. The mathematical result of this list of requirement resumes into a modified ionic EOM

$$M_I\ddot{\mathbf{R}}_I = -\frac{\partial \mathrm{E}_{NSC}}{\partial \mathbf{R}_I} + \sum_{i,j}\Lambda_{ij}\frac{\partial}{\partial \mathbf{R}_I}\left\langle \psi_i \middle| \psi_j \right\rangle - 2\frac{\partial \langle \psi_i|}{\partial \mathbf{R}_I}\left(\frac{\delta \mathrm{E}_{NSC}}{\delta \langle \psi_i|} - \sum_j \Lambda_{ij}|\psi_j\rangle\right) \tag{24}$$

While the first term in the right hand of the equation is clear, the rest seems a bit puzzling at a first glance. Indeed, in the original formulation [16], the selected basis set is not PW, but a localized basis set as in Eq. 10. Hence, the electronic wavefunctions depend also on the atomic coordinates $\mathbf{R}_I$ and the request of orthogonality at each step is released to save time, meaning that the scalar product $<\psi_i|\psi_j>$ is no longer vanishing. Analogously, the total energy E^{DFT} is not reoptimized as in full self-consistent BO procedures and for this reason is indicated as non-self-consistent energy E_{NSC}. Wavefunctions are propagated according to an algorithm which resembles the original CPMD formulation in the sense that second order EOMs are used, but a damping term (first-order derivative) is present which reminds a sort of steepest-descent algorithm typical of the BO dynamics. The net result is the electron dynamics,

$$\mu\frac{d^2}{dt^2}|\psi_i\rangle + \gamma\frac{d}{dt}|\psi_i\rangle = -\frac{\delta E_{NSC}}{\delta\langle\psi_i|} + \sum_j \Lambda_{ij}|\psi_i\rangle \tag{25}$$

which is then solved via a predictor–corrector scheme. With no pretention of completeness, the procedure can be summarized as follows. On a

first instance, in a localized basis set $\{|q>\}$, the electronic wavefunctions are expanded as

$$|\psi_i\rangle = \sum_{q=1}^{M} C_{iq}|q\rangle \tag{26}$$

on the M functions composing the localized basis. Then the NxM matrix of the expansion coefficient is written as

$$\mathbf{C} = [C_{iq}] = \begin{pmatrix} C_{11} & . & . & C_{1M} \\ . & . & & . \\ . & & . & . \\ C_{N1} & . & . & C_{NM} \end{pmatrix} \tag{27}$$

and the density matrix becomes $\mathbf{P} = \mathbf{C}\mathbf{C}^T = \mathbf{PSP}$. The MxM matrix indicated as $\mathbf{S}$ is just given by the expansion coefficient and its matrix elements have the usual form

$$S_{qq'} = \sum_{i=1}^{N} C_{iq}^{*} C_{iq} \tag{28}$$

Hence, the (DFT) total energy can be rewritten as $E^{\text{tot}}[\mathbf{C},\mathbf{R}_I]$ which can be used in a straightforward way to write a BO dynamics

$$M\ddot{\mathbf{R}}_I = -\nabla_{\mathbf{R}_I} \min\{E^{\text{tot}}[\mathbf{C},\mathbf{R}_I]\} \quad \forall \mathbf{C}^T \cdot \mathbf{S} \cdot \mathbf{C} = \mathbf{1} \tag{29}$$

under the given constraint on C which resumes in an implicit orthogonality condition. However it must be kept into account that: (i) Diagonalization and minimization of E^{tot} are required in BO; (ii) Hellman-Feynman forces are just one component since Pulay forces due to the local basis set are present. Residual force components appear due to non-self-consistency (NSC) of the approach. To take into account all the points above, the basic strategy can be summarized in four major points:

1. Propagate the electronic variables in time according to the CP original idea of updating on-the-fly to avoid expensive full diagonalization operations
2. Use a good propagation algorithm $\mathbf{C}(t_n) = f(\mathbf{C}(t_{n\text{-}1}),\ldots,\mathbf{C}(t_{n\text{-}m}))$ depending on previous time steps $m \in [1,K]$ time steps
3. Select the appropriate number of steps K to keep $\mathbf{C}(t_n)$ as close as possible to the (electronic) ground state
4. Enforce convergence on the BO surface, correct this propagation CPMD-like afterwards

Point 3 corresponds to the first move in the numerical integration procedure and it can be identified as the "predictor" part directly deriving from a standard numerical integration of the CPMD type equations of motion. Point 5, instead, is the "corrector" needed afterwards to better converge the wavefunctions and to restore the neglected self-consistent loop. The use of not necessarily fully converged wavefunctions at the predictor propagation stage allows for large integration steps, thus resulting in a remarkable boost in the dynamics.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Computer Modeling and Simulation of Materials
- ▶ Electronic Structure Calculations
- ▶ First Principles Calculations
- ▶ Molecular Dynamics Method
- ▶ Molecular Dynamics Simulations of Interactions Between Biological Molecules and Nanomaterials
- ▶ Molecular Dynamics Simulations of Nanobiomaterials
- ▶ Surface Electronic Structure

References

1. Alder, B.J., Wainwright, T.E.J.: Phase transition for a hard sphere system. Chem. Phys. **27**, 1208 (1957)
2. Rahman, A.: Correlation in the motion of atoms in liquid argon. Phys. Rev. **136**, A405 (1964)
3. Hohenberg, P., Kohn, W.: Inhomogeneous electron gas. Phys. Rev. **136**, B864 (1964)
4. Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Phys. Rev. **140**, A1133 (1965)
5. Car, R., Parrinello, M.: Unified approach for molecular dynamics and Density-Functional theory. Phys. Rev. Lett. **55**, 2471 (1985)
6. Marx, D., Hutter, J.: Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods. Cambridge

University Press, New York (2009). ISBN 978-0521898638
7. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. Oxford University Press, New York (1989). ISBN 0-19-504279
8. Hehre, W.J., Radom, L., Schleyer, P.V.R., Pople, J.A.: Ab Initio Molecular Orbital Theory. Wiley, New York (1986). ISBN 978-0471812418
9. Born, M., Oppenheimer, J.R.: Zur Quantentheorie der Molekeln. Ann. Phys. **84**, 457 (1927)
10. Bornemann, F.A., Schütte, C.: A mathematical investigation of the Car-Parrinello method. Numer. Mat. **78**, 359 (1998)
11. Boero, M., Tateno, M.: Quantum theoretical approaches to proteins and nucleic acids. In: Oxford Handbook of Nanoscience and Technology, Vol. 1: Basic Aspects, pp. 549–598. Oxford University Press, New York (2010). ISBN 978-0199533046
12. Boero, M.: Reactive simulations for biochemical processes. In: Atomic-Scale Modeling of Nanosystems and Nanostructured Materials. Lecture Notes in Physics, vol. 795, pp. 81–98. Springer, Berlin/Heidelberg (2010). ISBN 978-3-642-04650-6
13. Oshiyama, A., Iwata, J.: Large-scale electronic-structure calculations for nanomaterials in density functional theory. J. Phys. Conf. Ser. **302**, 012030 (2011)
14. Car, R., Parrinello, M.: The unified approach to density functional and molecular dynamics in real space. Solid State Commun **62**, 403 (1987)
15. Pang, T.: An Introduction to Computational Physics. Cambridge University Press, New York (2000). ISBN 978-1840858839
16. Kühne, T.D., Krack, M., Mohamed, F.R., Parrinello, M.: Efficient and accurate Car-Parrinello-like approach to Born-Oppenheimer molecular dynamics. Phys. Rev. Lett. **98**, 066401 (2007)

Carrier-Free Electrophoresis

▶ Micro Free-Flow Electrophoresis (μFFE)

Catalyst

▶ Chemical Vapor Deposition (CVD)
▶ Physical Vapor Deposition

Catalytic Bimetallic Nanorods

▶ Molecular Modeling on Artificial Molecular Motors

Catalytic Chemical Vapor Deposition (CCVD)

▶ Chemical Vapor Deposition (CVD)

Catalytic Janus Particle

▶ Molecular Modeling on Artificial Molecular Motors

Cathodic Arc Deposition

▶ Physical Vapor Deposition

Cavity Optomechanics

▶ Nano-optomechanical Systems (NOMS)

Cell Adhesion

▶ Bioadhesion
▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Cell Adhesion and Detachment

▶ Biological Breadboard Platform for Studies of Cellular Dynamics

Cell Micro-patterning

▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Cell Migration

▶ Biological Breadboard Platform for Studies of Cellular Dynamics

Cell Patterning

► Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Cellular and Molecular Toxicity of Nanoparticles

► Cellular Mechanisms of Nanoparticle Toxicity

Cellular Dynamics

► Biological Breadboard Platform for Studies of Cellular Dynamics

Cellular Electronic Energy Transfer

► Micro/Nano Transport in Microbial Energy Harvesting

Cellular Imaging

► Electrical Impedance Tomography for Single-Cell Imaging

Cellular Mechanisms of Nanoparticle Toxicity

Francelyne Marano, Fernando Rodrigues-Lima, Jean-Marie Dupret, Armelle Baeza-Squiban and Sonja Boland
Unit of Functional and Adaptive Biology (BFA), Laboratory of Molecular and Cellular Responses to Xenobiotics, UMR CNRS 8251, Univ Paris Diderot, (Sorbonne Paris Cité), Paris cedex 13, France

Synonyms

Cellular and molecular toxicity of nanoparticles

Definition

The interaction between nanoparticles and cell triggers a cascade of molecular events which could induce a toxicity and cell death. They are associated with the uptake of nanoparticles, their persistence at cellular level, and their ability to release free radicals and to induce an oxidative stress. The resulting activation of molecular pathways and transcription factors could lead to a pro-inflammatory response or, depending on the level of free radicals, apoptosis.

Background

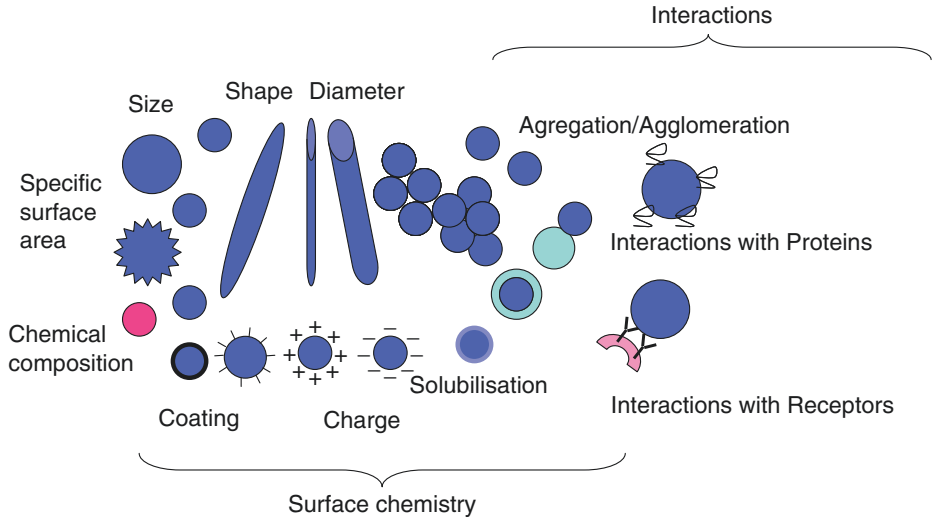
The last 5 years have shown an increasing number of papers on the mechanisms of nanoparticle (NP) cytotoxicity. What are the reasons? It is likely that the specific useful properties which appear at nanoscale can also lead to adverse effects. This hypothesis is strongly supported by in vivo and in vitro studies to compare the toxicity of NPs with their fine counterparts of the same chemical composition. These results have clearly demonstrated a higher toxicity of particles at nanoscale than at microscale. Moreover, it appears from experimental studies that solid nano-sized particles could be translocated beyond the respiratory tract and could induce a systemic response. The interstitial translocation of a same mass of particles is higher for ultrafine than fine particles after intratracheal instillation in rats [1]. Surface area, which is strongly increased for NPs compared to micro-particles of same chemical composition, and surface reactivity are considered as the principal indicators of NP reactivity. It was shown that a toxic response could be observed even to apparently nontoxic substances when the exposure occurred in the nanometer size range. All these observations have led to the development of a new field of toxicology, nanotoxicology [2]. However, the toxicological mechanisms which sustained the biological response are not yet clear and a matter of debates.

The concerns about the toxicity of engineered nanoparticles, which are increasingly used for industrial and medical applications, came also

from the knowledge on the toxicity of non-intentional atmospheric particles. Short-term epidemiological studies in Europe and North America have showed an association between cardiorespiratory morbidity and mortality and an increased concentration of atmospheric fine particles [3]. Moreover, the long-term epidemiological studies have also demonstrated an association between exposure to atmospheric particles (particulate matter or PM10 and 2.5) and increased cancer risk [3]. In parallel, in vitro and in vivo studies on fine and ultrafine airborne particles such as diesel exhaust particles, PM2.5, gave causal explanations to these adverse health effects (reviewed in Ref. [1]). They allow defining the molecular events induced by these particles in lung cells. The major event is a pro-inflammatory response which is characterized by the release of various cytokines (pro-inflammatory mediators), associated with the activation of transcription factors and signaling pathways. This was especially demonstrated for diesel exhaust particles (DEP), a major component of urban PM in Europe. These events are mostly induced by organic components of the DEP and are probably mediated by the generation of reactive oxygen species (ROS) during the metabolism of organic compounds (for a review, see Ref. [4]). These findings were used as a background for the researches on biological mechanisms induced by NPs considering that fine and ultrafine atmospheric particles have great similarities with NPs, especially diesel exhaust particles which are of nano size and aggregate after their release in the atmosphere.

It became rapidly obvious that the understanding of the cellular and molecular mechanisms leading to the biological effects of NPs was essential for the development of safe materials and accurate assays for risk assessment of engineered NPs [2], and several recent reviews were focused on demonstrated or hypothetic cellular mechanisms of these responses [4–6].

The first event, when NPs enter in contact with the human body by inhalation, oral and dermal exposure, or intravenous application, is their interaction in the biological fluids and the cellular microenvironment with biological molecules such as proteins thus forming a protein corona [7]. Consequently, NPs do not directly interact with the cell membrane but through the protein and/or lipids of the corona. NP-bound proteins may recognize and interact with membrane receptors or could bind nonspecifically to cellular membranes. Whatever these interactions, they seem to play a central role which could determine further biological responses. In particular, these interactions may drive the uptake of NPs by the first target cells at the level of the biological barriers such as immune cells (macrophages, dendritic cells, and neutrophils) or epithelial and endothelial cells. This uptake seems to be general for many NPs which are able to bind proteins at their surface, and the paradigm of "Trojan horse" was developed to explain this uptake and the further biological responses. One of the first responses is the direct or indirect production of ROS which is associated with the size, the chemical composition, and the surface reactivity of the NPs. This common response occurs for a large number of NPs even with different chemical patterns and different abilities to form agglomerates, and thus the paradigm of the central role of oxidative stress was developed [5]. These authors suggested that "although not all materials have electronic configurations or surface properties to allow spontaneous ROS generation, particle interactions with cellular components are capable of generating oxidative stress." Further activation of nuclear factors and specific genetic programs are associated with the level of ROS production leading to cell death by necrosis and apoptosis or adaptive responses such as pro-inflammatory responses, antioxidant enzyme activation, repair processes, effects on cell cycle control, and proliferation. Over the last years, numerous in vitro studies have confirmed this hypothesis leading to the development of assays using the detection of ROS or oxidative stress for the screening of NPs. However, new data during the last year have pointed out other specific effects of NPs which are not related to oxidative stress. For example, NPs can interact with membrane receptors, induce their aggregation, and mimic sustained physiological responses through specific signaling pathways in the target cells. This type of mechanism may contribute to the development of diseases but

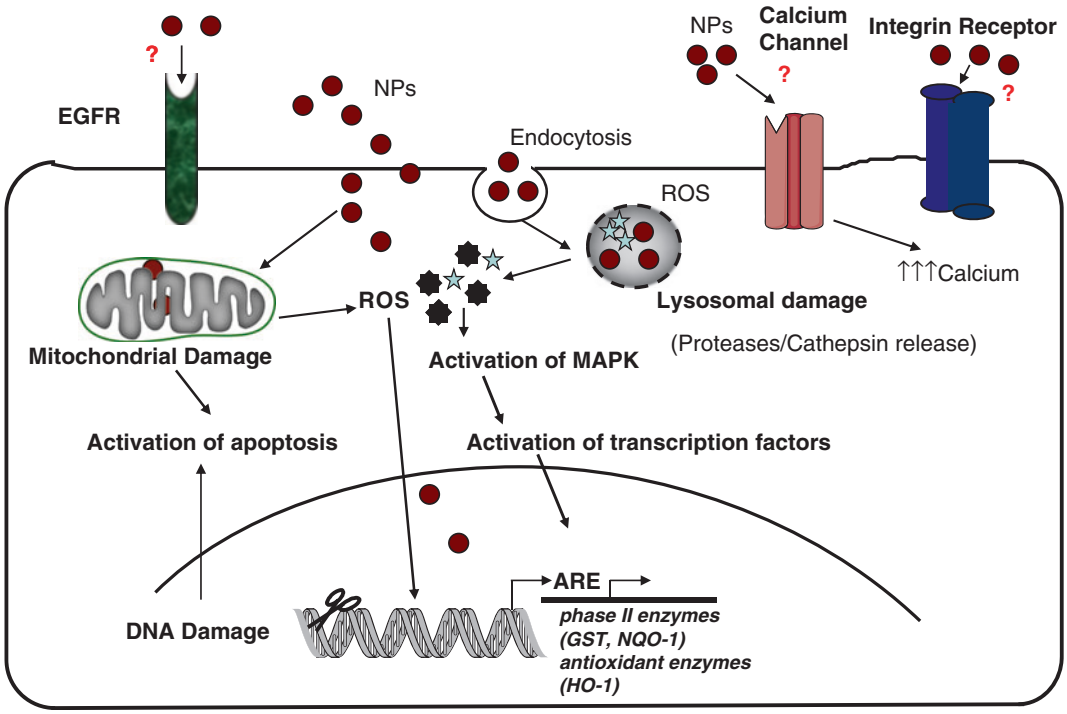


Cellular Mechanisms of Nanoparticle Toxicity, Fig. 1 Different physicochemical characteristics of the nanomaterials involved in their biological activity: size, surface area, shape, bulk chemical composition, surface chemistry including solubility as well as surface charge or coatings, and interactions between particles leading to agglomeration and aggregation as well as with proteins leading to "corona" or with receptors in the cell membrane

could also be of use to develop therapeutic strategies whereby NPs activate or block specific receptors.

Cellular Uptake of Nanoparticles and Their Fate at Cellular Level

The uptake of particles by specialized immune cells in human is a normal process which leads to their removal and contributes to the integrity of the body. However, depending on the level of the uptake, this process could induce an increasing release of inflammatory mediators and disturbance of the normal functions of phagocytes such as the clearance and the destruction of pathogens. One of the knowledge of the last fifty years on the effects of a sustained exposure to airborne particles, especially at occupational level, is the concept of overloading. If the mechanisms of clearance are not sufficient to eliminate the particles and if they are persistent, the particles could accumulate in the tissues, leading to a sustained inflammation and chronic pathologies. This was demonstrated not only for exposure to quartz, asbestos, coal, and mineral dusts but also for longtime exposure to heavy PM-polluted atmospheres such as in Mexico City [1]. These questions of uptake and persistence are fundamental for risk assessment evaluation of NPs. This may explain the number of papers published recently that analyze the mechanisms of uptake, the behavior, and the translocation of various NPs. So far, it appears that the response depends on several different parameters: the NP surface and its specific chemical composition resulting from the engineering processes; the capacity of NPs to form aggregates (strongly bonded or fused NPs) or agglomerates (collection of weakly bound NPs); and the methods used for dispersion and experimental preparation, which determine the ability of NPs to adsorb or not specific biological compounds such as proteins, to form the "corona" and to interact with biological membranes [7]. The amount and the structural/functional properties of the adsorbed proteins drive the interactions of these nanomaterials with the membranes and their uptake (Fig. 1). Recent studies have clearly identified a number of serum proteins such as albumin, IgG, IgM, IgA, apolipoprotein E, cytokines, or transferrin that bind to carbon black, titanium dioxide, acrylamide,



Cellular Mechanisms of Nanoparticle Toxicity, Fig. 2 A schematic representation of NP-triggered cellular pathways through membrane receptors, ROS production, and implication of oxidative stress in these responses. NPs could induce activation of EGF or integrin receptors can lead to apoptosis, inflammation, or proliferation. ROS produced by NPs in immediate cellular environment or inside the cells lead to activation of redox-dependent signaling pathways like MAPK and the activation of transcription factors, e.g., AP-1, NF-kB, or Nrf2. They migrate to the nucleus and modify gene expression of cytokines, phase 2 enzymes (glutathione S transferase or GST, quinone oxidoreductase 1 or NQO-1), and antioxidant enzymes (heme oxygenase 1 or HO-1). Oxidative stress could also result in the damage of different organelles like the mitochondria, lysosomes, and nucleus resulting to apoptosis. Accumulation of high intracellular calcium levels through a direct effect on calcium channel might also act as an alternative mechanism for the induction of these mechanisms (Adapted from Marano et al. [11])

or polystyrene NPs [8]. Among the identified proteins, several are ligands for cellular receptors and may contribute to the biological effects of NPs. For example, receptor aggregation induced by NPs could lead to cell signaling: coated gold NPs were able to bind and cross-link IgE-Fc epsilon receptors leading to degranulation and consequent release of chemical mediators [9].

On another hand, integrins such as $\alpha_5\beta_3$ are known to play a key role in cell signaling, and their activation by extracellular ligands can modulate biological processes such as matrix remodeling, angiogenesis, tissue differentiation, and cell migration. These receptors were recently demonstrated as important membrane targets for carbon NPs, and their activation induced lung epithelial cell proliferation which was due at least in part to β1-integrin activation [6].

As far as uptake process is concerned, it is likely that different cell types might have different uptake mechanisms, even for the same NPs. The possible pathways of cellular uptake were previously described by several authors (see Ref. [10]). It could occur through phagocytosis, macropinocytosis, clathrin-mediated endocytosis, non-clathrin- and non-caveolae-mediated endocytosis, caveolae-mediated endocytosis, or diffusion (Fig. 2). These mechanisms have been described for different NPs and may occur for the same NP depending on the cell type, the medium, and the level of aggregation. Therefore, uptake processes are considered as very complex and not

easy to measure. Dawson et al. [12] have postulated that the uptake depends mostly on the size: NPs less than 100 nm can enter the cells and less than 40 nm in the nucleus. It was also suggested that the size of the NPs determine caveolin- versus clathrin-dependent uptakes [13]. However, these oversimplified scenarios are refuted by obvious discrepancies in the recent literature about the optimal size, shape, and mechanisms of internalization of NPs.

The surface charge of the NPs could be an important factor for uptake since the negatively charged surface membrane could favor the positively charged NPs for higher internalization. However, negatively charged NPs were also shown to have enhanced uptake as compared to unfunctionalized NPs, perhaps by their possible interactions with proteins. Endocytosis of small NPs is energy dependent and associated with lipid rafts, dynamin, and F-actin mechanisms. Phagocytosis and macropinocytosis are mostly involved in the endocytosis of large particles (more than 500 nm) and also in the uptake of the aggregates or agglomerates of NPs which could be promoted by their opsonization in the biological fluids. Macropinocytosis (which is one kind of pinocytosis) is also an important mechanism for positively charged NPs and TiO_2 or carbon black aggregate internalization [14].

The behavior of the NPs after their uptake is another important question, but, surprisingly, as far as now, little is known about the intracellular fate of NPs. Most of the transmission electron microscopy (TEM) observations have shown the NPs in cytoplasmic vesicles limited by membranes. These vesicles could further be transported in the cytoplasm through the microtubule network. The bio-persistence of nanomaterials which are resistant to degradation in the endosomal compartment could be one of the factors of further toxicity and accumulation. However, several metal oxide NPs are toxic after dissolution in the cell. Indeed, the uptake of ZnO NPs into the lysosomal acidic medium accelerates their dissolution and the release of Zn^{2+} ions in the cytoplasm. Their excess could induce cytokine production and cytotoxicity and the initiation of acute inflammation at the level of the target organ such as the lung.

NPs such as TiO_2 or carbon black NPs were also observed free in the cytoplasm of cells [14]. Two explanations may be put forward. The first one is that NPs could directly enter by diffusion through the lipid bilayer. It has been shown that cationic NPs could pass though cell membranes by generating transient holes without membrane disruption [15]. Another possible explanation could be the release of NPs after rupture of endosomal compartment. It was described that cationic NPs, after binding to lipid groups on the cell surface membrane, could be endocytosed in vesicles and accumulated in the lysosomal compartment. Within, they are able to sequestrate protons which could lead to the activation of proton pumps and further rupture of the ion homeostasis and lysosomal accumulation of water. The subsequent lysosomal swelling and membrane rupture lead to the cytoplasmic release of NPs [16]. In proliferating cells, these cytoplasmic NPs, associated or not with microtubules, could enter in the nucleus during the mitosis, which could explain that nonsoluble NPs were observed in the nucleus [14]. More rarely, NPs were also observed within the mitochondrial matrix but, so far, no explanation was given to explain this organelle localization.

The Cellular Stress Induced by Nanoparticles and Its Biological Consequences

The last 10 years of research conducted on the mechanisms of toxicity of non-intentional as well as engineered NPs has led to the establishment of a consensus within the scientific community of toxicologists to consider the central role of oxidative stress in cellular responses to NPs leading to inflammation or apoptosis [5, 17] (Fig. 2). The concept of oxidative stress was developed for many years to explain dysfunctions leading to pathologies. Oxidative stress could occur when reactive oxygen species (ROS) are overproduced leading to an imbalance between ROS production and antioxidant defense capacity. Oxidative stress could also occur when the organism shows a deficiency in antioxidant systems and, especially,

in antioxidant enzymatic systems (superoxide dismutase, catalase, glutathione peroxidase). An increased concentration of ROS, exceeding the antioxidant capacity of the cells, can lead to oxidative damage at molecular or cellular level.

ROS have important cellular roles either by acting as second messengers for the activation of specific pathways and gene expressions or by causing cell death. In the hierarchical oxidative stress model in response to NPs, Nel et al. [5] propose that a minor level of oxidative stress leads to the activation of the antioxidant protection, whereas, at a higher level, cell membrane and organelle injuries could lead to cell death by apoptosis or necrosis, but specific signaling pathways and gene expression are involved at each step. The induction of oxidative stress by several NPs is due to their ability to produce ROS (e.g., TiO_2) or to lead to their production. The surface properties of NPs modulate the production of ROS, and the smaller they are, the higher is their surface area and their ability to react with biological components and to produce ROS. However, if this cellular induction appears to be general, all the NPs are not able to produce ROS, and the cellular increase of the latter could be an indirect effect of the uptake.

ROS interact nonspecifically with biological compounds, yet some macromolecules are more sensitive such as the unsaturated lipids, amino acids with a sulfhydril group (SH), and guanine sites in nucleic acids. When lipid bilayer is attacked by ROS, cascade peroxidation occurs leading to the disorganization of the membranes and of their functions (exchange, barriers, information). The most sensitive proteins contain methionine or cysteine residues, especially in their active site, and their oxidation could lead to modifications of their activity and even to their inactivation.

The adaptive cellular responses to NPs are associated with the modulation of different redox-sensitive cellular pathways. Tyrosine kinases and serine/threonine kinase such as mitogen-activated protein kinases or MAP kinases were especially studied (ERK, p38, and JNK) in association with several transcription factors such as NFκB. The free radical can degrade the NFκB inhibitor IκB by the activation of the cascades leading to its proteolysis. The activation of NFκB induces its translocation within the nucleus and its link to consensus sequences in the promoter of numerous genes leading to their transcription. This is also the case for other transcriptions factors such as AP1 and NrF2. The latter plays an essential role in the antioxidant response element (ARE)-mediated expression of phase 2 enzymes such as NQO1 (NADPH quinone oxidoreductase-1) and antioxidant enzymes such as heme oxygenase-1 (HO-1). Indeed, HO-1 was found to be activated by CeO_2 NP exposure of human bronchial cells via the p38-Nrf-2 signaling pathway. The ability of NPs to interact with these signaling pathways could partially explain their cytotoxicity. Recently, TiO_2 and SiO_2 NPs were demonstrated in vitro and in vivo to induce the release of IL1β and IL1α, two potent mediators of innate immunity, via the activation of the inflammasome, a large multiprotein complex containing caspase 1 which cleaves pro IL1-β in its active form. These results lead to consider that these NPs could induce a potent inflammatory response. However, the mechanisms leading to this activation are not yet clear.

Another important target of ROS produced by NPs is DNA. Oxidative damage of DNA could generate intrachain adducts and strand breakage. The bond between the base and deoxyribose could also be attacked leading to an abasic site, and the attack on the sugar could create a single-strand break. The genotoxicity of NPs begins to be studied and recent reviews pointed out the possible genotoxic mechanisms.

However, oxidative stress appears now not sufficient to explain all the biological effects of NPs. The role of epidermal growth factor receptor (EGFR) was investigated by the group of K. Unfried with the demonstration that carbon black NPs induce apoptosis and proliferation via specific signaling pathways both using EGFR [18]. Carbon black NPs could also impair phagosome transport and cause cytoskeletal dysfunctions with a transient increase of intracellular calcium not associated with the induction of ROS since antioxidants did not suppress the response,

which could be due to a direct effect on ion channels that control the calcium homeostasis in the cell [19]. Even if all the mechanisms are not completely demonstrated, it appears now that transmembrane receptors are implicated in NP-induced cell signaling and could lead to specific biological responses to NPs.

Nanoparticles and Cell Death

NPs have also been shown to induce either apoptotic or necrotic cell death in a variety of in vitro systems depending on the concentration and duration of exposure. This induction of cell death mechanism by NPs might act as a basis of different pathologies, and consequently it is important to understand NP-induced apoptosis pathways. Cells are able to undergo apoptosis through two major pathways, the extrinsic pathway with the activation of death receptors and the intrinsic pathway with the central role of mitochondria, its permeabilization, and the release of cytochrome *c* leading to the activation of apoptosome. Recently, the permeabilization of lysosomal membrane was also shown to initiate apoptosis with the release of catepsins and other hydrolases from the lysosomal lumen. The molecular pathways of apoptosis induction by carbon black and titanium dioxide NPs in human bronchial epithelial cells were recently studied. It was shown that the initial phase of apoptosis induction depends upon the chemical nature of the NPs. Carbon black NPs triggered the mitochondrial pathway, with the decrease of mitochondrial potential, the activation of bax (a proapoptotic protein of the Bcl2 family), and the release of cytochrome *c*, and the production of ROS is implicated in the downstream mitochondrial events. TiO_2 NPs induced lysosomal pathway with lipid peroxidation, lysosomal membrane destabilization, and cathepsin B release [20]. Lysosomal permeabilization has also been shown to be important in silica NP-induced apoptosis. These results point out the necessity of a careful characterization of the molecular mechanisms involved by NPs and not just describing at the final outcome.

Future Directions of Research

The interactions between nanomaterials and their biological target are essential to explain their biological effect, and the interest of the recent researches on the cellular mechanisms induced by NPs is to take into account the specificity of the cells and their microenvironment. The first step is the formation of the corona in biological fluids whose composition and affinity kinetics strongly depend on the characteristics of NPs and, especially, their size and surface reactivity. This coating of proteins influences the aggregation, the final size, and, finally, the uptake of NPs via the interaction with the membranes, their specific receptors, or lipid rafts. This could determine if the nanomaterial is bioavailable and if NPs induce or not adverse interactions. The central mechanism proposed to explain the biological response is the oxidative stress. However, this paradigm is debated because very similar oxidative stress effects observed in cellular models and induced by different particles could lead in vivo to different pathological effects. It is now obvious that oxidative stress is a common and nonspecific mechanism in toxicology and that the responses at the level of the cell depend on the perturbation of the redox balance with a few number of induced signaling pathways. The different biological responses could depend on the tissue specificity which could lead to different diseases observed after occupational or environmental exposure to well-known particles or fibers.

Recent studies have also shown that NPs could develop a response without a direct contact with the cells but after an induction of secreted factors, which is the "bystander effect." Small molecules such as purines could be increased at cytoplasmic level in response to NPs, transferred through the gap junctions within a tissue to activate specific receptors [10]. Moreover, NP-induced apoptosis was also demonstrated to be propagated through hydrogen peroxide-mediated bystander killing in an in vitro model of human intestinal epithelium. These specific responses could explain the differences observed in vivo. Finally, the

interactions of NPs with proteins, enzymes, cytokines, and growth factors, outside or inside the cell, lead to modifications of the functions of these proteins with a possible indirect pathological effect.

The large variety of engineered NPs on the market and under development makes these studies very complex. However, the development of safe nanomaterials depends on better knowledge of these specific interactions.

Cross-References

- ▶ Genotoxicity of Nanoparticles
- ▶ In Vivo Toxicity of Carbon Nanotubes
- ▶ Quantum Dot Toxicity
- ▶ Toxicology: Plants and Nanoparticles

References

1. Donaldson, K., Borm, P.: Particle Toxicology, p. 434. CRC Press, Boca Raton, Florida, USA (2007)
2. Oberdorster, G., Oberdorster, E., Oberdorster, J.: Nanotoxicology: an emerging discipline evolving from studies of ultrafine particles. Environ. Health Perspect. **113**, 823–839 (2005)
3. Brunekreef, B., Holgate, S.T.: Air pollution and health. Lancet **360**, 1233–1242 (2002)
4. Marano, F., Boland, S., Baeza-Squiban, A.: Particle-associated organics and proinflammatory signaling. In: Donaldson, K., Borm, P. (eds.) Particle Toxicology, pp. 211–226. CRC Press, Boca Raton, Florida, USA (2007)
5. Nel, A., Xia, T., Madler, L., Li, N.: Toxic potential of materials at the nanolevel. Science **311**, 622–627 (2006)
6. Unfried, K., Albrecht, C., Klotz, L.O., Mikecz, A.V., Grether-Beck, S., Schins, R.P.F.: Cellular responses to nanoparticles: target structures and mechanisms. Nanotoxicology **1**, 52–71 (2007)
7. Nel, A.E., Madler, L., Velegol, D., Xia, T., Hoek, E.M., Somasundaran, P., Klaessig, F., Castranova, V., Thompson, M.: Understanding biophysicochemical interactions at the nano-bio interface. Nat. Mater. **8**, 543–557 (2009)
8. Lynch, I., Salvati, A., Dawson, K.A.: Protein-nanoparticle interactions: what does the cell see? Nat. Nanotechnol. **4**, 546–547 (2009)
9. Huang, Y.F., Liu, H., Xiong, X., Chen, Y., Tan, W.: Nanoparticle-mediated IgE-receptor aggregation and signaling in RBL mast cells. J. Am. Chem. Soc. **131**, 17328–17334 (2009)
10. Bhabra, G., Sood, A., Fisher, B., Cartwright, L., Saunders, M., Evans, W.H., Surprenant, A., Lopez-Castejon, G., Mann, S., Davis, S.A., Hails, L.A., Ingham, E., Verkade, P., Lane, J., Heesom, K., Newson, R., Case, C.P.: Nanoparticles can cause DNA damage across a cellular barrier. Nat. Nanotechnol. **4**, 876–883 (2009)
11. Marano, F., Hussain, S., Rodriges-Lima, F., Baeza-Squiban, A., Boland, S.: Nanoparticles: molecular target and cell signaling. Arch. Toxicol. 85(7):733-41 (2011)
12. Dawson, K.A., Salvati, A., Lynch, I.: Nanotoxicology: nanoparticles reconstruct lipids. Nat. Nanotechnol. **4**, 84–85 (2009)
13. Rejman, J., Oberle, V., Zuhorn, I.S., Hoekstra, D.: Size-dependent internalization of particles via the pathways of clathrin- and caveolae-mediated endocytosis. Biochem. J. **377**, 159–169 (2004)
14. Hussain, S., Boland, S., Baeza-Squiban, A., Hamel, R., Thomassen, L.C., Martens, J.A., Billon-Galland, M.A., Fleury-Feith, J., Moisan, F., Pairon, J.C., Marano, F.: Oxidative stress and proinflammatory effects of carbon black and titanium dioxide nanoparticles: role of particle surface area and internalized amount. Toxicology **260**, 142–149 (2009)
15. Gratton, S.E., Ropp, P.A., Pohlhaus, P.D., Luft, J.C., Madden, V.J., Napier, M.E., Desimone, J.M.: The effect of particle design on cellular internalization pathways. Proc. Natl. Acad. Sci. U. S. A. **105**, 11613–11618 (2008)
16. Xia, T., Kovochich, M., Liong, M., Zink, J.I., Nel, A. E.: Cationic polystyrene nanosphere toxicity depends on cell-specific endocytic and mitochondrial injury pathways. ACS Nano **2**, 85–96 (2008)
17. Ayres, J.G., Borm, P., Cassee, F.R., Castranova, V., Donaldson, K., Ghio, A., Harrison, R.M., Hider, R., Kelly, F., Kooter, I.M., Marano, F., Maynard, R.L., Mudway, I., Nel, A., Sioutas, C., Smith, S., Baeza-Squiban, A., Cho, A., Duggan, S., Froines, J.: Evaluating the toxicity of airborne particulate matter and nanoparticles by measuring oxidative stress potential – a workshop report and consensus statement. Inhal. Toxicol. **20**, 75–99 (2008)
18. Sydlik, U., Bierhals, K., Soufi, M., Abel, J., Schins, R. P., Unfried, K.: Ultrafine carbon particles induce apoptosis and proliferation in rat lung epithelial cells via specific signaling pathways both using EGF-R. Am. J. Physiol. Lung Cell. Mol. Physiol. **291**, L725–L733 (2006)
19. Moller, W., Brown, D.M., Kreyling, W.G., Stone, V.: Ultrafine particles cause cytoskeletal dysfunctions in macrophages: role of intracellular calcium. Part. Fibre Toxicol. **2**, 7 (2005)
20. Hussain, S., Thomassen, L.C., Feracatu, I., Borot, M. C., Andreau, K., Fleury, J., Baeza-Squiban, A., Marano, F., Boland, S.: Carbon black and titanium oxide nanoparticles elicit distinct apoptosic pathways in bronchial epithelial cells. Part. Fibre.Toxicol. **7**(10), 1–17 (2010). Online 16 Apr

Cellular Toxicity

▶ Nanoparticle Cytotoxicity

Chaos

▶ Nonlinear and Parametric NEMS Resonators

Characterizations of Zinc Oxide Nanowires for Nanoelectronic Applications

▶ Fundamental Properties of Zinc Oxide Nanowires

Charge Transfer

▶ Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Charge Transfer on Self-Assembled Monolayer Molecules

▶ Charge Transport in Self-Assembled Monolayers

Charge Transport in Carbon-Based Nanoscaled Materials

▶ Electronic Transport in Carbon Nanomaterials

Charge Transport in Self-Assembled Monolayers

Jeong Young Park
Graduate School of EEWS (WCU), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

Synonyms

Charge transfer on self-assembled monolayer molecules

Definition

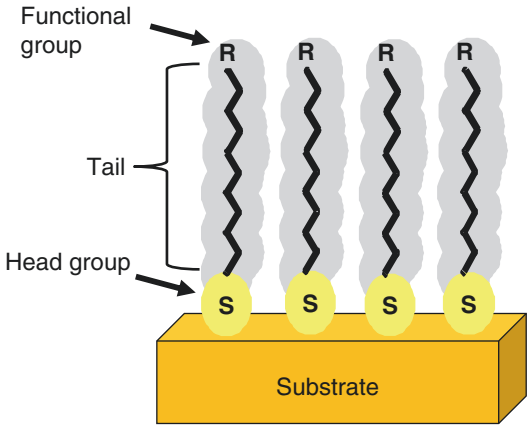
Charge transport in self-assembled monolayers (SAMs) is the transport of an electron or a hole through an organized molecule layer which is bound to a substrate.

Overview

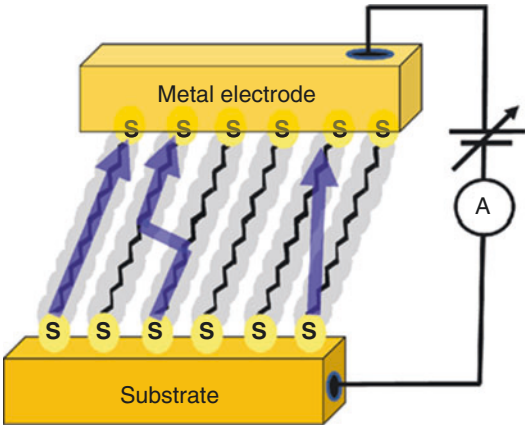
Charge Transport Through Organic Molecules

Significant studies on charge transport properties through organic molecules have been carried out in the general area of molecule-based and molecule-controlled electronic devices, often termed "molecular electronics" [1, 2]. Self-assembled monolayers (SAMs) are composed of an organized layer of amphiphilic molecules in which one end of the molecule, the "head group," shows a special affinity for a substrate [3]. SAMs also consist of a tail with a functional group at the terminal end, as seen in Fig. 1.

Charge transport of organic molecules is usually limited by hopping processes and is therefore dominated by surface ordering. Self-assembled monolayers are a good model system of molecular electronics due to the ordered surface structure. In order to measure charge transport in a self-assembled monolayer, the substrate surface should be metallic. For example, a gold surface



Charge Transport in Self-Assembled Monolayers, Fig. 1 Schematic of a self-assembled monolayer (*SAMs*) showing the head group that is bound to the substrate. SAMs consist of a tail with a functional group at the terminal



Charge Transport in Self-Assembled Monolayers, Fig. 2 Scheme of charge transport mechanisms through self-assembled monolayers. The dominant charge transport mechanism in a molecular junction involves "through-bond" (*TB*) tunneling, and "through space" (*TS*) as illustrated in the left and right transport channel, respectively

exhibits strong bonds with alkanethiol through S–H bonds. The other electrode should also be metallic for charge transport through the self-assembled monolayer. The measurement scheme of charge transport through a self-assembled monolayer that represents a conductor-molecule-conductor junction is shown in Fig. 2.

Charge Transport Mechanism

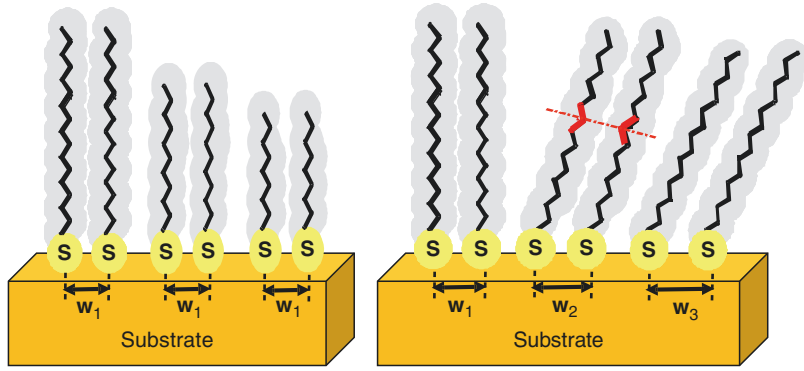
For insulating molecules, such as alkane chains, electron transport occurs via tunneling mechanisms. When such molecules are placed between electrodes, the junction resistance changes exponentially: $R = R_0 \exp(\beta s)$, with electrode separation s, where R_0 is the contact resistance and β a decay parameter. In most experiments, the separation s is the length of the alkane chain. However, length is not the only important parameter. Conformation and molecular orientation relative to the electrodes are also important. Other factors need to be considered as well, including energy positions of the highest occupied and lowest unoccupied molecular orbitals (HOMO, LUMO), electrode work function, and nature of the bonding to the electrodes.

Charge transport mechanisms through self-assembled monolayers consist mainly of three processes [4]. The dominant charge transport mechanism in a molecular junction is "through-bond" (TB) tunneling, where the current follows the bond overlaps along the molecules (as illustrated in the left transport channel of Fig. 2). Another contribution involves the charge transport from electrode to electrode, in which the molecule plays the role of a dielectric medium that is called "through space" (TS), as illustrated in the right transport channel of Fig. 2. The last contribution of charge transport pathway involves a chain-to-chain coupling as illustrated in the middle of Fig. 2. As the molecular chains tilt, the decrease of the electron tunneling distance leads to a lateral hop between the neighboring molecular chains.

Two Pathway Models

If electron transport was determined purely by tunneling through the alkane chains, one would expect the value of β to equal zero, since the tunneling distance is the same for all tilt angles. The nonzero value of β indicates the existence of either intermolecular charge transfer or variations in the S-Au bonding as a function of tilt that affect the conductivity in an exponential way with angle.

Slowinski et al. [4] proposed a two-pathway conductance model involving "through-bond" tunneling, and the "chain-to-chain" coupling.



Charge Transport in Self-Assembled Monolayers, Fig. 3 Scheme of the measurement of the junction resistance for two different situations: (*1*) decreasing of the alkane chain (*left* part), and (*2*) the tilting of the alkane chain while maintaining the same number of carbon atoms (*right* part), which will yield the resistance (*1*) per unit length of molecule or (*2*) tilting angle of the molecules, respectively

Assuming no effects due to changes in S-Au bonding, the first pathway is independent of tilt, while the second depends on the tilt angle. The tunneling current, thus, is given by

$$I_t = I_0 \exp(-\beta_{TB} d) + I_0 n_s \exp[-\beta_{TB}(d - d_{CC} \tan \Theta)] \times \exp(-\beta_{TS} d_{CC})$$

where I_t is the current at a specific tilt angle Θ, d is the length of the molecule, n_s is a statistical factor accounting for the number of pathways containing a single lateral hop as compared to those containing only through-bond hops, d is the diameter of the molecule chains, β_{TB} and β_{TS} are respectively through-bond and through-space decay constants. For example, in case of C16 alkanethiol molecule chains, $d_{CC} = 4.3$ Å, $d = 24$ Å, and $n_s = 16$, i.e., the number of carbon atoms in the molecule.

Decay Constant upon Shortening and Tilting of Molecules

The junction resistance is dependent on electrode spacing for two different situations: (1) shortening of the alkane chain [5] and maintaining the same width (w) between chains and (2) tilting of the alkane chain but changing w [6, 7]. These measurements will yield the resistance per unit length of molecule or tilting angle of the molecules, respectively. The conductance decay constant β has already been measured using SAMs with different chain lengths when the separation between electrodes decreases as a function of the alkane chain length (the left image of Fig. 3). The decay constant, β, upon tilting of molecules can be measured using deformation with an AFM tip and simultaneous measurement of current (the right image of Fig. 3). This methodology will be described in the next section.

Basic Methodology

Preparation of Self-Assembled Monolayer

The organic molecular films on various types of substrates (conducting, semiconducting, or insulating substrates) have been prepared using techniques such as the Langmuir-Blodgett technique, dipping the substrates into solution with molecules, drop casting, or spin-coating [8].

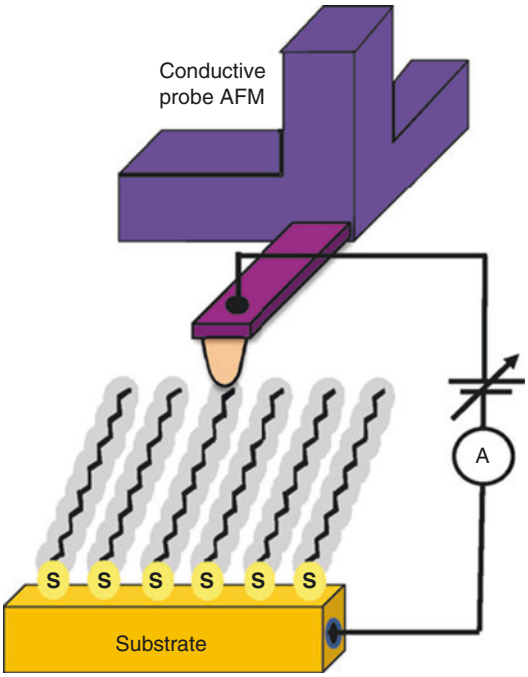
As one example, details on the preparation of an alkanethiol SAM will be described below. Gold substrates (200–300 nm of gold coating over 1–4 nm of chromium layer n glass) are prepared by butane flame annealing in air after cleaning in acetone, chloroform, methanol, and a piranha solution (1:3; H_2O_2:H_2SO_4). The resulting surface consisted of large grains with flat terraces of (111) orientation (sizes up to 400 nm) separated by monatomic steps. Flatness and cleanness were tested by the quality of the lattice-resolved images of the gold substrate.

Two types of hexadecanethiol (C16) self-assembled monolayer can be formed on Au (111): complete monolayers of the molecules and islands of molecules covering only a fraction of the substrate. In the first case, the film was produced by immersing the substrate in 1 mM ethanolic solution of C16 for about 24 h, followed by rinsing with absolute ethanol and drying in a stream of nitrogen to remove weakly bound molecules. Incomplete monolayers in the form of islands were prepared by immersing the substrate in a 5 μM ethanolic solution of C16 for approximately 60 s, followed by rinsing. Samples consisting of islands facilitate the determination of the thickness of the molecular film relative to the surrounding exposed gold substrate. The molecular order of the islands improves with storage time at ambient conditions.

Techniques to Measure Charge Transport in Self-Assembled Monolayers

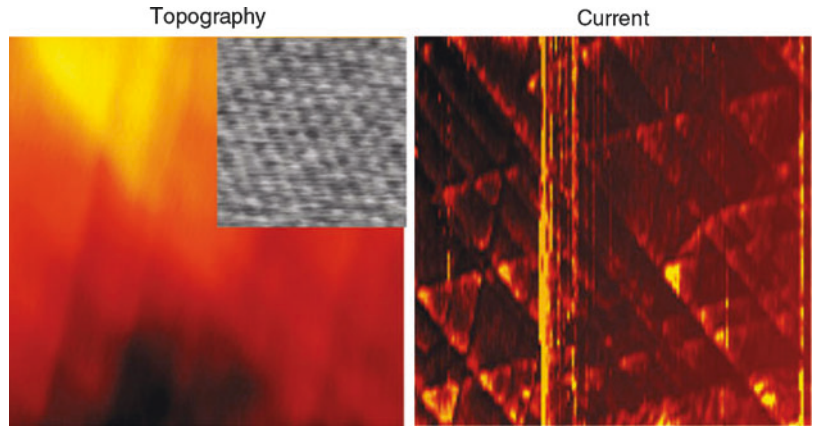
The current through a thiol SAM on a hanging Hg drop electrode can be measured in an electrochemical solution. The current was measured as a function of the monolayer thickness that can be tuned by two methods: by changing the number of carbons in the alkane chain and therefore its length; or, expansion of the Hg drop such that the monolayer surface coverage was reduced and the molecules increased their tilt angle with respect to the surface. Slowinski et al. determined the decay constants $\beta_{TB} = 0.91/Å$ and $\beta_{TS} = 1.31/Å$ by both a fit to their experimental data and by independent ab initio calculations. Mercury drop expansion experiments by Slowinsky et al. have shown a dependence of the current through the alkanethiol monolayers on surface concentration, prompting the authors to suggest the existence of additional pathways for charge transfer, like chain-to-chain tunneling.

Scanning tunneling microscopy and scanning tunneling spectroscopy have been used to reveal the atomic scale surface structure and charge transport properties of SAM layers [9, 10]. STM has been used to reveal various phases of surface structure and atomic scale defects, which could play a crucial role in the electrical transport.



Charge Transport in Self-Assembled Monolayers, Fig. 4 Scheme of conductance measurements of SAM with a conductive-probe atomic force microscopy (*CP-AFM*) system

Conductance measurements were performed with a conductive-probe atomic force microscopy (CP-AFM) system. The use of AFM with conducting tips provides the ability to vary the load on the nanocontact and also opens the way for exploring electron transfer as a function of molecular deformation. A junction is fabricated by placing a conducting AFM tip in contact with a metal-supported molecular film, such as a self-assembled monolayer (SAM) on Au, as shown in Fig. 4. The normal force feedback circuit of the AFM controls the mechanical load on the nanocontact while the current–voltage (*I–V*) characteristics are recorded. The possibility to control the load on the contact is an unusual characteristic of this kind of junction and provides the opportunity to establish a correlation between the mechanical deformation and electronic properties of organic molecules. The normal force exerted by the cantilever was kept constant during AFM imaging, while the current between tip and sample was recorded. It is crucial to carry out the experiment in



Charge Transport in Self-Assembled Monolayers, Fig. 5 AFM images (200 × 200 nm) of topography, and current images obtained simultaneously for a full monolayer of C16 on Au (111) surface. Lattice-resolved images of the film (*inset* in the *left* figure) reveal a lattice image of SAM (size: 2 × 2 nm)

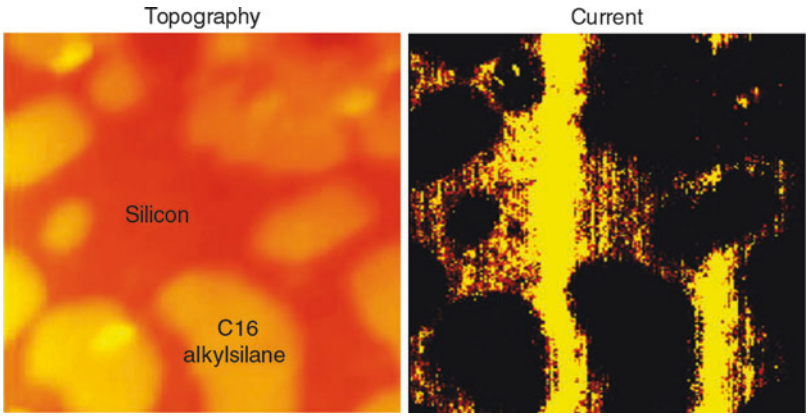
the low load regime so that there is no damage to the surface. This can be confirmed by inspection of the images with Ångstrom depth sensitivity as well as by the reproducibility of the current and adhesion measurements. If the measured conductance did not change at constant load and did not show time-dependent behavior in the elastic regime, the tip experiences minimal changes during subsequent contact measurements.

Key Research Findings

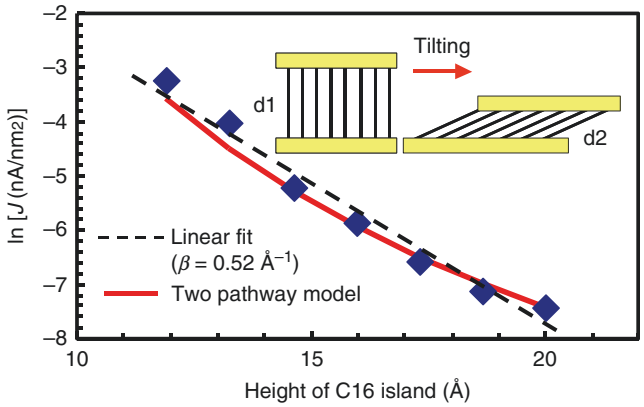
The molecular tilt induced by the pressure applied by the tip is one major factor that leads to increased film conductivity. By measuring the current between the conductive AFM tip and SAM as a function of the height of the molecules, the decay parameter (β) can be obtained [11]. Wold et al. studied the junction resistance as a function of load using AFM. The resistance was found to decrease with increasing load within two distinct power law scaling regimes [12]. Song et al. examined the dependence of the tunneling current through Au-alkanethiol-Au junctions on the tip-loading force [13]. It is found that the two-pathway model proposed by Slowinsky et al. can reasonably fit with the results, leading the authors to conclude that the tilt configuration of alkanethiol SAMs enhances the intermolecular charge transfer.

Figure 5 shows topography and current images obtained simultaneously for a full monolayer of C16 on an Au (111) surface. The topographic image reveals the commonly found structure of the gold film substrate, composed of triangular-shaped terraces separated by atomic steps. Lattice-resolved images of the film (inset in the left figure) reveal a ($\sqrt{3} \times \sqrt{3}$)-R30° periodicity of the molecules relative to the gold substrate. Qi et al. measured current–voltage (I–V) characteristics on the C16 alkanethiol sample for loads varying between −20 and 120 nN, and found that the current changes in a stepwise manner and the plateaus are associated with the discrete tilt angle of the molecules. A stepwise response of the SAM film to pressure has been observed previously in other properties such as film height and friction of alkanesilanes on mica and alkanethiols on gold.

In order to measure the thickness of the self-assembled monolayer upon molecular deformation, the SAM islands that partially cover the substrate can be used. The heights of the islands can be obtained from topographical AFM images, while charge transport properties of alkanesilane SAMs on silicon surface are measured using AFM with a conducting tip. In this manner, the load applied to the tip-sample contact can be varied while simultaneously measuring electric conductance. Figure 6 shows the topographic and current images, respectively, that were acquired simultaneously on hexadecylsilane islands on a silicon surface. The image size is 500 × 500 nm. The hexadecylsilane islands are 100–200 nm in diameter and have a height of 1.6 nm at the applied load of 0 nN (or effective total load of 20 nN). It is also



Charge Transport in Self-Assembled Monolayers, Fig. 6 AFM images (500 × 500 nm) of topographic, and current images, respectively, that were acquired simultaneously on hexadecylsilane SAM islands on silicon surface



Charge Transport in Self-Assembled Monolayers, Fig. 7 Semilog plot of current density (nA/nm^2) as a function of the height of the hexadecylsilane SAM islands on a silicon surface. A decay constant (β) = 0.52 ± 0.04 Å^{-1} was found for the current passing through the film as a function of tip-substrate separation

clear that the current measured on the alkanesilane island is much smaller than that measured on the silicon surface.

These changes were shown to correspond to the molecules adopting specific values of tilt angle relative to the surface, and explained as the result of methylene groups interlocking with neighboring alkane chains. In the case of complete monolayers of alkanethiol SAM, the junction resistance (R) was measured as a function of the applied load [6]. These data were converted to current versus electrode separation by assigning each step in the current to a specific molecular tilt angle, following the sequence established in previous experiments. It was found that ln(R) increases approximately linearly with tip-surface separation, with an average slope β = 0.57 (±0.03) Å^{-1}. Similar measurement of the decay parameter upon the molecular tilts was carried out with a scanning tunneling microscope and simultaneous sensing of forces. By measuring the current as a function of applied load, a tunneling decay constant β = 0.53 (±0.02) Å^{-1} was obtained [14].

In the case of hexadecylsilane molecules, the local conductance of hexadecylsilane SAM islands on a silicon surface was measured with conductive-probe AFM. A semilog plot of current density (nA/nm^2) was obtained as a function of the height of the hexadecylsilane SAM islands on a silicon surface, as shown in Fig. 7. A decay constant (β) = 0.52 ± 0.04 Å^{-1} was found for the current passing through the film as a function of tip-substrate separation [7]. Figure 7 shows the best fit of the two-pathway model with the experimental current measurement as a function of the heights of molecule islands by using the fitting parameters of β_{TB} and β_{TS} that are 0.9 and 1.1 Å^{-1}, respectively. The good fit indicates that

the two-path tunneling model is a valid model to describe this observation.

While saturated hydrocarbon chains mainly interact with each other via weak van der Waals forces, much stronger intermolecular π-π interactions can be present in organic films comprised of conjugated/hybrid molecules. This influences charge transport significantly [5, 15]. In a conductance AFM study of two SAM systems, Fang et al. revealed the role of π-π stacking on charge transport and nanotribological properties of SAM consisting of aromatic molecules [16]. The two model molecules chosen in this study are (4-mercaptophenyl) anthrylacetylene (MPAA) and (4-mercaptophenyl)-phenylacetylene (MPPA). In MPPA, the end group is a single benzene ring, while in MPAA it is changed to a three fused benzene ring structure. This structural difference induces different degrees of lattice ordering in these two molecular SAM systems. Lattice resolution is readily achieved in the MPAA SAM, but it is not possible for the MPPA SAM under the same imaging conditions, indicating the MPAA is lacking long-range order. However, it is important to note that even without long-range order, the stronger intermolecular π-π stacking in the MPAA SAM greatly facilitates charge transport, resulting in approximately one order of magnitude higher conductivity than in the MPPA SAM.

Future Directions for Research

In this contribution, the basic concept of and recent progress on charge transport studies of organic SAM films formed by saturated hydrocarbon molecules and conjugated molecules has been outlined. Several techniques, including AFM, STM, and hanging Hg drop electrode, are used to elucidate the charge transport properties of SAM layers. A number of molecular scale factors such as packing density, lattice ordering, molecular deformation, grain boundaries, annealing induced morphological evolution, and phase separation play important roles in determining charge transport through SAM films. High resolution offered by scanning probe microscopy (SPM) is a key element in identifying and studying microstructures (e.g., molecular tilt, lattice ordering, defects, vacancies, grain boundaries) in organic films and their effects on electronic properties. Other advanced surface characterization techniques, such as SAM with nano-electrodes, in combination with conductive-probe atomic force microscopy, and spectroscopic techniques such as ultraviolet photoemission spectroscopy (UPS) and inverse photoemission spectroscopy (IPES), could be promising venues to explore the correlation between microstructures and electronic properties of organic films.

Cross-References

- Atomic Force Microscopy
- Conduction Mechanisms in Organic Semiconductors
- Electrode–Organic Interface Physics
- Scanning Tunneling Microscopy
- Self-Assembly

References

1. Aviram, A., Ratner, M.A.: Molecular Electronics: Science and Technology. New York Academy of Sciences, New York (1998)
2. Reed, M.A., Zhou, C., Muller, C.J., Burgin, T.P., Tour, J.M.: Conductance of a molecular junction. Science **278**, 252–254 (1997)
3. Ulman, A.: An Introduction to Ultrathin Organic Films from Langmuir-Blodgett to Self-Assembly. Academic, Boston (1991)
4. Slowinski, K., Chamberlain, R.V., Miller, C.J., Majda, M.: Through-bond and chain-to-chain coupling. Two pathways in electron tunneling through liquid alkanethiol monolayers on mercury electrodes. J. Am. Chem. Soc. **119**, 11910–11919 (1997)
5. Salomon, A., et al.: Comparison of electronic transport measurements on organic molecules. Adv. Mater. **15**, 1881–1890 (2003). doi:10.1002/adma.200306091
6. Qi, Y.B., et al.: Mechanical and charge transport properties of alkanethiol self-assembled monolayers on a Au(111) surface: the role of molecular tilt. Langmuir **24**, 2219–2223 (2008). doi:10.1021/la703147q
7. Park, J.Y., Qi, Y.B., Ashby, P.D., Hendriksen, B.L.M., Salmeron, M.: Electrical transport and mechanical properties of alkylsilane self-assembled monolayers on silicon surfaces probed by atomic force microscopy. J. Chem. Phys. **130**, 114705 (2009)

8. Barrena, E., Ocal, C., Salmeron, M.: Molecular packing changes of alkanethiols monolayers on Au(111) under applied pressure. J. Chem. Phys. **113**, 2413–2418 (2000)
9. Bumm, L.A., Arnold, J.J., Dunbar, T.D., Allara, D.L., Weiss, P.S.: Electron transfer through organic molecules. J. Phys. Chem. B **103**, 8122–8127 (1999)
10. Xu, B.Q., Tao, N.J.J.: Measurement of single-molecule resistance by repeated formation of molecular junctions. Science **301**, 1221–1223 (2003)
11. Wang, W.Y., Lee, T., Reed, M.A.: Electron tunnelling in self-assembled monolayers. Rep. Prog. Phys. **68**, 523–544 (2005)
12. Wold, D.J., Haag, R., Rampi, M.A., Frisbie, C.D.: Distance dependence of electron tunneling through self-assembled monolayers measured by conducting probe atomic force microscopy: unsaturated versus saturated molecular junctions. J. Phys. Chem. B **106**, 2813–2816 (2002). doi:10.1021/jp013476t
13. Song, H., Lee, H., Lee, T.: Intermolecular chain-to-chain tunneling in metal-alkanethiol-metal junctions. J. Am. Chem. Soc. **129**, 3806 (2007)
14. Park, J.Y., Qi, Y.B., Ratera, I., Salmeron, M.: Noncontact to contact tunneling microscopy in self-assembled monolayers of alkylthiols on gold. J. Chem. Phys. **128**, 234701 (2008). doi:10.1063/1.2938085
15. Yamamoto, S.I., Ogawa, K.: The electrical conduction of conjugated molecular CAMs studied by a conductive atomic force microscopy. Surf. Sci. **600**, 4294–4300 (2006)
16. Fang, L., Park, J.Y., Ma, H., Jen, A.K.Y., Salmeron, M.: Atomic force microscopy study of the mechanical and electrical properties of monolayer films of molecules with aromatic end groups. Langmuir **23**, 11522–11525 (2007)

Chem-FET

► Nanostructure Field Effect Transistor Biosensors

Chemical Beam Epitaxial (CBE)

► Physical Vapor Deposition

Chemical Blankening

► Chemical Milling and Photochemical Milling

Chemical Dry Etching

► Dry Etching Processes

C

Chemical Milling and Photochemical Milling

Seajin Oh and Marc Madou
Department of Mechanical and Aerospace Engineering and Biomedical Engineering, University of California at Irvine, Irvine, CA, USA

Synonyms

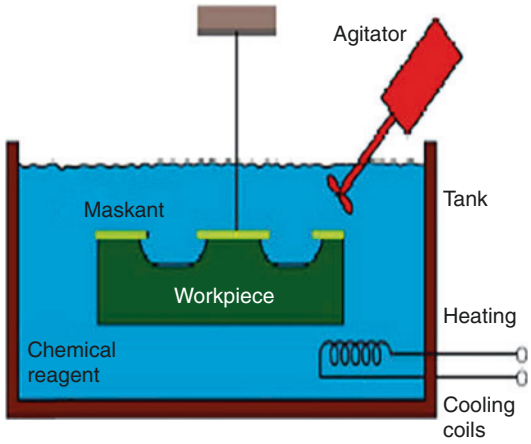
Chemical blankening; Photoetching; Photofabrication; Photomilling

Definition

Photochemical milling (PCM), also known as photochemical machining, is the process of fabricating high precision metal workpieces using photographically produced masks and etchants to corrosively remove unwanted parts. This process is called wet etching in MEMS fabrication techniques and can be also applied to nonmetal materials. Wet etching, when combined with nanolithography, is a useful process to fabricate detailed nanostructures by extremely controlled removal (Fig. 1).

Overview

Photochemical machining (PCM) produces three-dimensional features by wet chemical etching (Fig. 2). PCM yields burr-free and stress-free metal products and allows for the machining of a wide range of materials which would not be suitable for traditional metal working techniques. PCM is also known as photoetching,



Chemical Milling and Photochemical Milling, Fig. 1 (**a**) Schematic illustration of photochemical milling process

photomilling, photofabrication, or chemical blankening [1]. There is a special type of photochemical milling that uses light for initiating or accelerating the wet etching process in metal or semiconductor materials.

The combination of photoresists and wet etching enables the fabrication of very detailed structures with complex geometry or large arrays of variable etching profiles in thin (<2 mm) flat metal sheets. Photoresists are made of synthetic polymers having consistent properties. Liquid photoresist coats a thin film by dipping or spin casting which enables the production of detailed patterns, but often creates pinholes in the thin layer. Thick dry photoresist films applied by hot lamination have advantages of process simplicity and reliability although the materials are expensive.

The process of wet etching is based on the redox chemistry of etchant reduction and metal oxidation, which results in the formation of soluble metal-containing ions that diffuse away from the reaction metal surface. Many metals commonly used in manufacturing industry are etched readily in aqueous solutions comprising etchant (e.g., ferric chloride). Metal oxides and virtually all materials can be etched with a proper selection of etchant regardless of different etching rates.

Wet etching in the PCM process is isotropic where the etchant attacks both downward into the material and sideways under the edge of the resist layer and the ratio of the depth to the undercut is termed the etch factor (Fig. 3). In MEMS device fabrication, the undercut plays a key role in fabricating free-standing microstructure patterns (e.g., beams and cantilevers) that are necessary where microstructures have to be flexible, thermally isolated, small mass, double-sided contacts with gases or liquid surroundings. In most cases, the free-standing material is deposited as a film on a substrate surface, termed a sacrificial material. The etchant, possessing a lateral etching component, must be sufficiently selective not to attach the free-standing material. Further, novel nanolithography techniques enable to create submicrometer-scaled features.

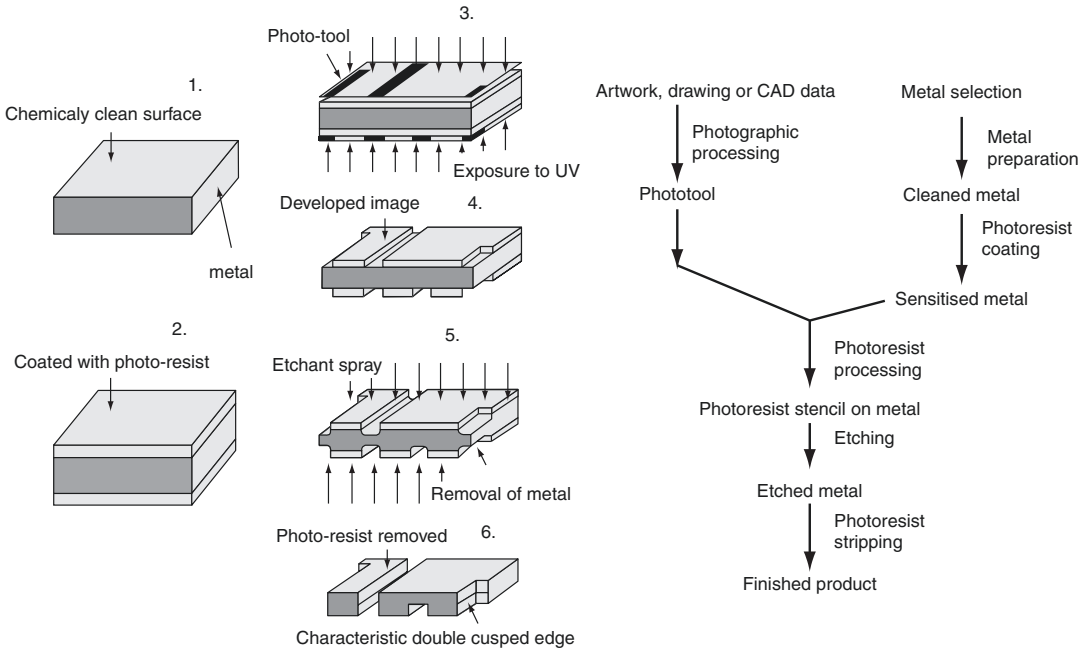
Basic Methodology

Photolithography

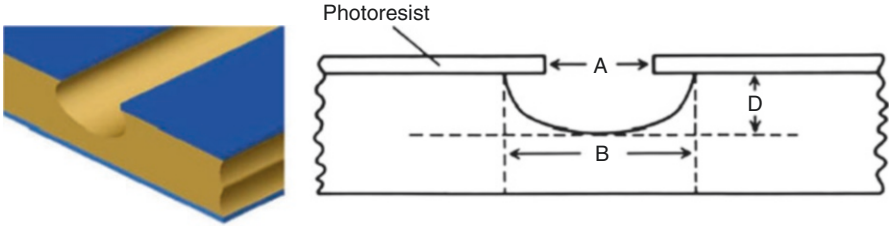
An image of the profile of the flat feature is generated by computer-aided design (CAD) and electronically transferred onto a photographic film to produce a phototool, a photolithographic mask as known in MEMS. The photoresist is exposed through the phototool from ultraviolet source. There are two types of photoresist – negative and positive. UV lights soften the positive film, and the exposed area is released in the developing solution. The negative photoresist film has reversed pattern developing characteristics. Wet resist is applied to a metal sheet by a dipping process while spin coating is commonly used in micromachining (Fig. 4).

Wet Etching

In wet chemical etching, the components of the solid are changed into soluble chemical components which are transported by diffusion or convection away from the surface into the bulk of the solution. The solvent molecules form a shell around the dissolved solid particles that are mobile in the liquid phase. The specific interactions of the components of the liquid with the solid determine the reaction rate, which is attributed to a greater etching selectivity of the solid



Chemical Milling and Photochemical Milling, Fig. 2 PCM process in metal sheet etching. *1*. Chemically clean the metal surface. *2*. Coat both sides of the plate with photoresist that adheres to the metal when exposed to UV light. *3*. Expose plate and phototool to ensure image transfer. *4*. Develop and create photomask image. *5*. Spray metal with etchant or dip it in hot acidic solution to etch all material other than part covered with photoresist. *6*. Rinse the plate to ensure photoresist and etchant removal Ref. [2]



Chemical Milling and Photochemical Milling, Fig. 3 Schematic illustration of etch detail through line openings in the patterned photoresist (*blue*). Etch factor = Depth of etch (*D*)/Undercut [½ (B − A)] Ref. [2]

than dry etching methods. Water is used as the solvent in most wet etching processes [3].

1. In a metal or a semiconductor, the dissolution of metal or semiconductor is accompanied by an electron transfer and obeys the laws of electrochemistry. The oxidation reaction of metal or semiconductor M releases metal ions into solution and produce electrons.

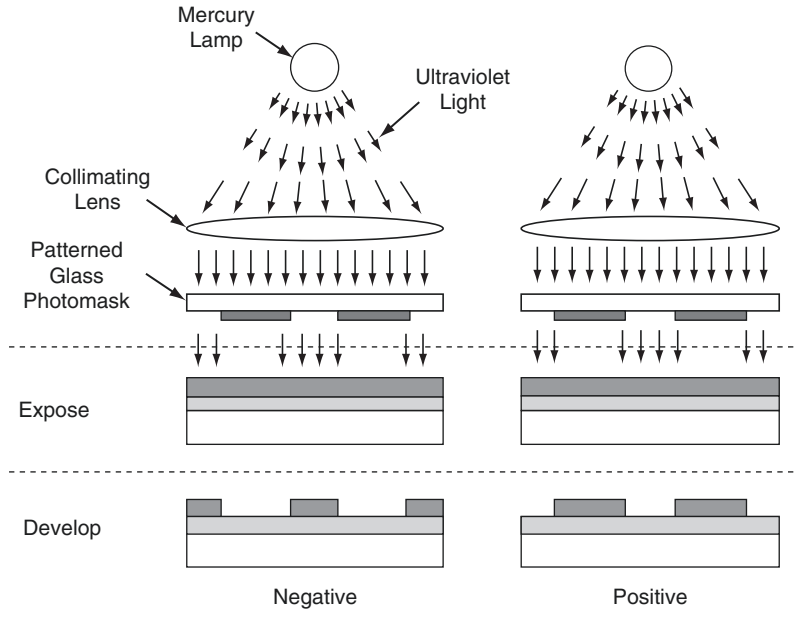
$$M = M^{n+} + ne^{-} \quad (\text{anodic partial process})$$

A secondary chemical redox process takes place to transfer the liberated electrons to an oxidizing agent OM in the solution.

$$OM + ne^{-} = OM^{n-} \quad (\text{cathodic partial process})$$

The anodic and cathodic partial processes results in etching metal.

$$M_{\text{in solid}} + OM_{\text{in solution}} = M^{n+} + OM^{n-}$$



Chemical Milling and Photochemical Milling, Fig. 4 Positive and negative photoresist Ref.[2]

The etch rate corresponds to the number of metal ions produced at the solid surface per time unit, which is proportional to the interchanged anodic partial current I over the surface A, I/A. Metals and semiconductors often dissolve as complexes (e.g., $[MY_x]^+$) where smaller molecules or ions (ligands) form a chemically bound primary shell around the central atom. The etch rate can be changed by varying the concentration of reactants, temperature, viscosity, and convection of the solution.

2. In dielectric materials, acid–base reactions take place if the material to be etched reacts with hydrogen or hydroxyl ions. The cations are solvated by water as a strong polar solvent and they can diffuse rapidly into the bulk of the solution. Etching processes are applicable with oxides and hydroxides of metals and semiconductors at low pH-values,

$$M_xO_y + 2yH^+ = xM^{2y/x+} + yH_2O$$
$$M(OH)_x + xH^+ = M^{x+} + xH_2O$$

M = metal or semiconductor (e.g., Cu, Si)

Similarly, at very high pH-values some metals and semiconductors form stable water dissolvable complexes that are easily solvated by water.

Acidic anions, such as chloride and fluoride, or neutral molecules react as ligands, a typical example of which is the etching of SiO_2 in HF-containing etchants.

$$SiO_2 + 6HF = H_2SiF_6 + H_2O$$

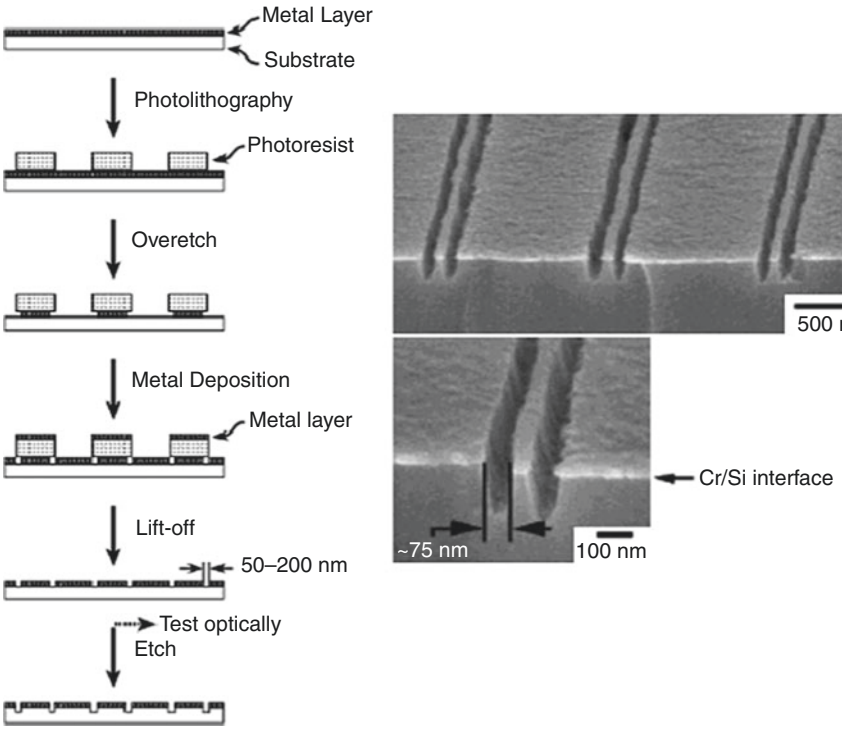
Salt-like film is also dissolved by complexing agents. For example, copper chloride is etchable in neutral KCl solutions.

$$CuCl + 2Cl- = [CuCl_3]^{2-}$$

Key Research Findings for Nanotechnology

Edge Lithography

Undercutting by isotropic wet etching is applied to transfer the edges of a photoresist pattern into a feature of the final pattern, as illustrated in Fig. 5. The process generates 50 nm scale trenches by controlled undercutting. Currently, patterning features <100 nm are possible by advanced lithography techniques – deep ultraviolet, electron beam writing, extreme ultraviolet, and x-ray photolithography – but are prohibitively



Chemical Milling and Photochemical Milling, Fig. 5 (*Left*) Schematic illustration of the process generating nanometer-scale lines by controlled undercutting. A pattern is produced in the photoresist by photolithography or soft lithography. An isotropic etch is applied to the substrate beneath the photoresist. Shallow undercutting of the base layer, followed by evaporation into the exposed areas, and lift-off generates ~50 ± 200 nm gaps in the thin film at the edges of the photoresist pattern. The pattern is then used as an optical filter. (*Right*) SEM image of a cross section of linear trenches at the edges of a 100 nm line transferred into a Si <100 > substrate. The trenches are 250 nm deep Ref. [4]

expensive. In contrast, edge lithography is a convenient, inexpensive technique for patterning features with nanometer-scale dimensions.

Wet Etching for Maskless Patterning

In the photochemical etching process, light exposure can increase the charge carrier density in near surface areas which enhances the anodical as well as the cathodical partial processes in wet etching. For example, defect electrons in the lower band left by light exposure support the release of cations and at the same time the released electrons in the conduction band are readily accepted by an oxidizing agent in the solution. In the same process, the intensive exposure with a focused beam enables direct pattern generation without a lithographic mask. This method is specified preferentially for patterning semiconductors deposited on a nonconducting substrate (e.g., GaN on sapphire) [5].

Future Directions for Research

Wet chemical etching is a simple, inexpensive, and well-understood process. The process plays a key role in the field of MEMS and nanotechnology. Continuous effort is being made to provide controllability, repeatability, and, most importantly, detail in fabricating microstructures. New etchant compositions have been developed to apply wet chemical etching to the materials constituting new devices. A representative example is selective removal of metal nitride films on sapphire for a

new version of a light-emitting diode. At the same time, great efforts have been made to develop new lithography techniques for nanoscale patterns such as proximal probe lithography, very thin to monolayer lithography, and soft lithography. PCM equipped with the emerging lithography techniques can cause the paradigm shift in the creation of nanoscaled features [6].

Cross-References

- ▶ Dry Etching Processes
- ▶ DUV Photolithography and Materials
- ▶ Electron Beam Lithography (EBL)
- ▶ EUV Lithography
- ▶ Nanoimprint Lithography
- ▶ Nanotechnology
- ▶ Stereolithography

References

1. Abate, K.: Photochemical etching of metals. Met. Finish. **100**(6A), 448–451 (2002)
2. Allen, D.M.: Photochemical machining: from manufacturing's best kept secret to a $6 billion rapid manufacturing process. CIRP J. Manuf. Syst. **53**, 559–572 (2005)
3. Köhler, J.M.: Wet chemical etching method. In: Etching in Microsystem Technology. Wiley, Weinheim (1999)
4. Love, J.C., Paul, K.E., Whitesides, G.M.: Fabrication of nanometer-scale features by controlled isotropic wet chemical etching. Adv. Mater. **13**(8), 604–607 (2001)
5. Bardwell, J.A., Webb, J.B., Tang, H., Fraser, J., Moisa, S.: Ultraviolet photoenhanced wet etching of GaN in K2S2O8 solution. J. Appl. Phys. **89**(7), 4142–4149 (2001)
6. Madou, M.: Fundamentals of Microfabrication, 2nd edn. CRC Press, Boca Raton (2002)

Chemical Modification

▶ Nanostructures for Surface Functionalization and Surface Properties

Chemical Solution Deposition

▶ Sol–Gel Method

Chemical Vapor Deposition (CVD)

Yoke Khin Yap and Dongyan Zhang
Department of Physics, Michigan Technological University, Houghton, MI, USA

Synonyms

Aerosol-assisted chemical vapor deposition (AACVD); Atmospheric pressure chemical vapor deposition (APCVD); Atomic layer chemical vapor deposition (ALCVD); Atomic layer deposition (ALD); Atomic layer epitaxial (ALE); Boron nitride nanotubes (BNNTs); Carbon nanotubes (CNTs); Carbon nanowalls; Catalyst; Catalytic chemical vapor deposition (CCVD); Chemical vapor deposition (CVD); Cold-wall thermal chemical vapor deposition; Dissociated adsorption; Double-walled carbon nanotubes (DWCNTs); Graphene; High-pressure carbon monoxide (HiPCO); Hot filament chemical vapor deposition (HFCVD); Hot-wall thermal chemical vapor deposition; Inductively coupled-plasma chemical vapor deposition (ICP-CVD); Low-pressure chemical vapor deposition (LPCVD); Metalorganic chemical vapor deposition (MOCVD); Multiwalled carbon nanotubes (MWCNTs); Nanobelts; Nanocombs; Nanoparticles; Nanotubes; Nanowires; Plasma-enhanced chemical vapor deposition (PECVD); Single-walled carbon nanotubes (SWCNTs); Thermal chemical vapor deposition; Ultrahigh vacuum chemical vapor deposition (UHVCVD); Vertically-aligned carbon nanotubes

Definition

Chemical vapor deposition (CVD) is referred to deposition process of thin films and nanostructures through chemical reactions of vapor-phase precursors. Since CVD can be conducted using high purity precursors, it likely leads to thin film and nanostructures with high purity. The use of vapor-phase precursors also enables better control on the composition and

doping of thin films and nanostructures. For example, Si thin films can be deposited by decomposition of silane gas (SiH_4) by plasma or heat as follows: SiH_4 (g) → Si (s) + $2H_2$ (g). Doping of Si films with boron will lead to p-type Si films and can be achieved by the addition of B_2H_6 gas.

Classification

Classification by Operating Pressures

Chemical reactions involved in a CVD technique can be initiated by many ways leading to the classification of various types of CVD approaches. For example, CVD can be classified according to the operating pressures as follows:

- *Atmospheric pressure CVD* (APCVD) is referred to CVD processes that are conducted at atmospheric pressure. The advantage of APCVD is a simple experimental setup without the need of a vacuum system. The potential drawback will be the undesired contamination.
- *Low-pressure CVD* (LPCVD) is referred to CVD processes at pressures below and close to atmospheric pressure. One of the purposes of reducing the operation pressure is to avoid undesired reactions between precursors. LPCVD can also improve film uniformity.
- *Ultrahigh vacuum CVD* (UHVCVD) is referred to LPCVD processes at a very low pressure, typically below ~10^{-8} torr.

There are *CVD* processes that operate at high pressures. See details in section "Classification by Excitation Techniques."

Classification by Excitation Techniques

In addition, CVD can be classified by the excitation techniques that initiate the chemical reactions. For example:

- *Plasma-enhanced CVD* (PECVD) is referred to CVD processes that employed plasmas to initiate the needed chemical reactions. In general, PECVD could reduce the growth temperatures as the chemical reactions in PECVD are not ignited by heat. There are many subclassification of PECVD depending on the type of AC potential used for plasma generation. For example, RF-PECVD and MW-PECVD employed radio frequency (RF, typically at 13.56 MHz) and microwave (MW, 2.45 GHz) potential to dissociate gases and produce the needed chemical reactions. Some PECVD are classified by the configuration of plasma generation. For instance, inductively coupled-plasma CVD (ICP-CVD) is actually RF-PECVD that uses an induction coil as the RF electrode outside the vacuum chamber and is sometimes called remote plasma CVD [1]. On the other hand, two RF plasmas can be used in a CVD system. For example, a dual RF-PECVD approach was demonstrated for the growth of vertically aligned carbon nanofoils/nanowalls and carbon nanotubes (CNTs) [2, 3].
- *Thermal CVD* is referred to CVD processes that employed heat to initiate the needed chemical reactions. The most common thermal CVD technique employed external furnace to control the growth temperatures of the entire reaction zone (e.g., vacuum quarts/glass chambers). This is sometimes called *hot-wall thermal CVD*. The advantages of this approach are including the potential of large-scale synthesis. In contract, there are several approaches to achieve the so-called *cold-wall thermal CVD*. For example, hot filaments (HF) are used to heat up the temperatures of the adjacent substrates while the needed chemical reactions are ignited with higher temperatures on the filaments. This approach is called *hot filament CVD* (HFCVD). Other heating approaches can also be used including IR lamps, laser beams, and passing current flows through a suspended Si chip [4].

Classification by the Precursor Type and Feeding Procedure

CVD can also be classified by the type of precursors or the procedures where precursors are introduced into the reaction chamber. Some of the examples are described as follows:

- *Aerosol-assisted CVD (AACVD)* involves the use of carrier gases (usually inert gases such as

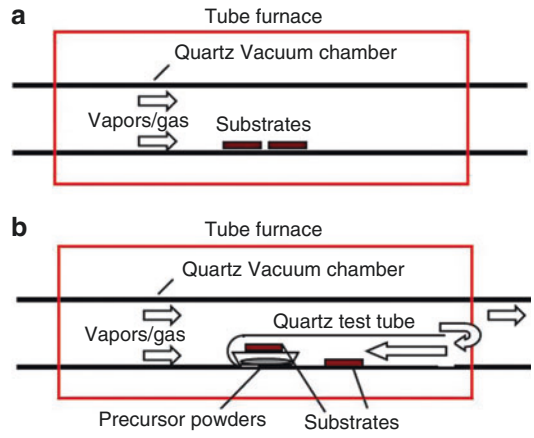
Ar, He, etc.) to transfer vapors of liquid-phase precursor into the reaction chamber in the form of aerosol. This approach enables the use of liquid precursors for the CVD process. The most well-known AACVD is *metalorganic CVD* (*MOCVD*) where metal organic solids are dissolved in organic solvents. For example, trimethylgallium [TMGa, $Ga(CH_3)_3$] is often used as the source of Ga to form GaN when it reacts with ammonia (NH_3) at high temperatures. Thus *AACVD* can be viewed as a subclassification of *thermal CVD*.

- *Atomic layer CVD* (*ALCVD*) is better known as *atomic layer deposition* (*ALD*) or *atomic layer epitaxial* (*ALE*) [5]. ALD allows conformal deposition of monolayer of binary (or ternary) compounds by introducing the two (or three) reacting precursors in a pulsed mode one after another. For example, monolayers of zinc sulfide (ZnS) can be deposited by first exposing the growth surface with $ZnCl_2$. Once chemisorption of $ZnCl_2$ is completed, the reaction chamber will be purged with an inert gas. Thereafter, hydrogen sulfide (H_2S) will be introduced to the chamber so that the following reaction will take place on the growth surface: $ZnCl_2$ (g) + H_2S (g) → ZnS (s) + HCl (g). Since the chemical process is self-limiting, only one monolayer (or less) of ZnS will be formed in each cycle. The advantage of *ALD* is conformal coating on the full surface of the sample and the precision of film thickness control. The major drawback is the extreme low deposition rate. *ALD* is usually conducted at low temperatures (200–400 °C) and may be viewed as a subclassification of *thermal CVD*. However, plasma and metal organic precursors are sometimes used in *ALD*.

Examples of CVD Approaches for Nanotechnology

Catalytic Chemical Vapor Deposition (CCVD)

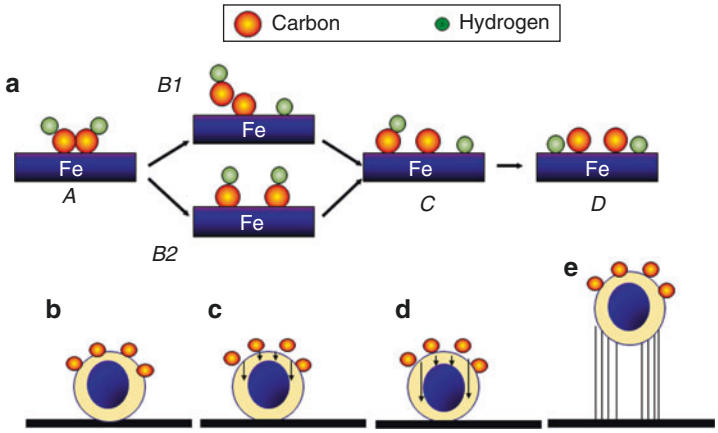
Catalytic chemical vapor deposition (*CCVD*) is the simplest and most popular technique for the synthesis of carbon nanotubes (CNTs), graphene, and nanowires (NWs). *CCVD* is simply a *thermal CVD* approach with the use of catalyst that induces the formation of nanomaterials. A typical experimental layout for *CCVD* is shown in Fig. 1. As shown, the *CCVD* system consists of a tube furnace and a quartz tube chamber where chemical reactions are taking place.



Chemical Vapor Deposition (CVD), Fig. 1 (**a**) Typical setup for CCVD. (**b**) Modified double-tube configuration

Synthesis of Vertically Aligned Carbon Nanotubes by CCVD

A typical experimental setup of *CCVD* for the synthesis of carbon nanotubes (CNTs) is shown in Fig. 1a. In this case, the substrates (usually oxidized Si) are first deposited with catalyst films such as Fe, Ni, or Co by electron beam evaporation, sputtering, pulsed laser deposition, etc. [6–8]. Then, these samples were then annealed at 600 °C in Ar, N_2, or H_2 for about 30 min. During the annealing process, these catalyst films will be converted into nanoparticles that will serve as the growth sites for CNTs. After the annealing, the growth temperatures (~600–800 °C) and growth ambient (usually Ar, H_2, or their mixtures) will be set prior to the introduction of hydrocarbon source gas (methane, CH_4; ethylene C_2H_4, acetylene, C_2H_2). These hydrocarbon gases will be decomposed on the surface of the catalyst nanoparticles through the chemical process of dissociative adsorption [6].

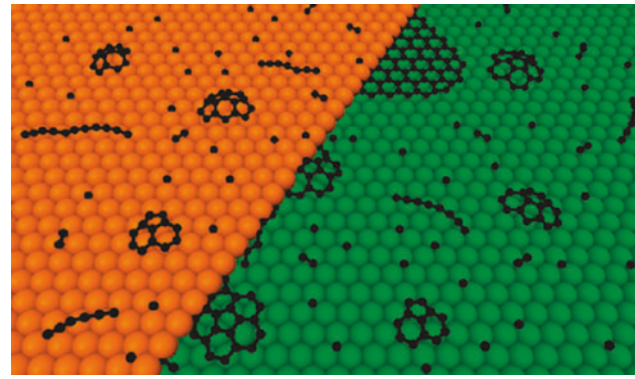


Chemical Vapor Deposition (CVD), Fig. 2 (**a**) Sequences of dissociative adsorption of C_2H_2 on Fe surface (see text for detailed description). The (**b**) decomposed carbon atoms (**c**) diffused into the subsurface of the solid-core Fe nanoparticle until (**d**) supersaturation and then (**e**) segregate as nanotubes

For example, dissociated adsorption of C_2H_2 is summarized in Fig. 2 [6]. Figure 2a shows the adsorption of a C_2H_2 molecule (step *A*) on the surface of the Fe nanoparticle. This will lead to either the breaking of C–H bond (step *B1*) to form C_2H and H fragments or the breaking of C=C bond (step *B2*) to form two C–H fragments. The catalytic function of the Fe nanoparticle is to reduce the energy required for decomposition by a charge transfer from hydrocarbon molecules to Fe. According to a first principle calculation, the dissociation energy of the first hydrogen atom from an isolated C_2H_2 (step *A* to *B1*) in vacuum can be reduced from 5.58 to 0.96 eV. On the other hand, the energy barrier between *A* and *B2* is 1.25 eV. The C–H bond breaking (step *B1*) is followed by C–C bond breaking (step *C*) with a potential barrier of 1.02 eV, whereas C=C bond breaking (step *B2*) is followed by C–H bond breaking (step *C*) with an energy barrier of 0.61 eV. Both modes (*A* to *B1* to *C or A* to *B2* to *C*) are possible and give one C–H fragment, one C, and one H. The decomposition of C_2H_2 is completed after the breaking of the last C–H bond (step *D*) with the need of a potential energy of 0.61 eV.

The decomposed carbon atoms (Fig. 2b) will then diffuse into the subsurface of the Fe nanoparticle (Fig. 1c). At typical growth temperature (650–800 °C), these nanoparticles are not melted even after considering the eutectic point of Fe-C phase. Since dissociative adsorption is an exothermic process, the near surface temperatures of the catalytic nanoparticles will be higher than the growth temperatures. Since the melting of nanoparticles starts from their surfaces, it is possible that the near surface region of the particles is melted. This will form the gas-liquid interface between carbon and Fe solid-core nanoparticles. Due to the high diffusion, rate of carbon in Fe melts; a Fe-C alloy will start to form. When these nanoparticles become supersaturated with carbon (Fig. 2d) to a value critical for growth at the solid-liquid interface, the excess carbon will segregate as carbon nanotubes (Fig. 2e). A tip-growth mode is illustrated in this figure where the nanoparticles remained at the tips of CNTs. A base-growth mode is also possible where the nanoparticles remained at the bases of CNTs. Thus, the diameters of CNTs depend to certain extent on the diameters of the nanoparticles. By controlling the thickness of the catalyst films and other parameters, *CCVD* enables the growth of vertically aligned (VA) single-, double-, and multi-walled CNTs [8].

In fact, one of the most impressive achievements in the growth of vertically aligned single-walled CNTs (VA-SWCNTs) is the so-called super growth [9]. This water-assisted approach was based on *CCVD* with the addition of 20–500 ppm of water vapors. These water vapors were introduced into the growth chamber by flowing carrier gas (Ar, or He with 40 % H_2) through a water bubbler. Ethylene was used as the carbon source along with various catalysts



Chemical Vapor Deposition (CVD), Fig. 3 Nucleation of graphene on Ni (111) surface (Reprinted with permission from J. Gao et al., J. Am. Chem. Soc. C **133**, 5009 (2011). Copyright (2011) American Chemical Society)

[Fe (1 nm), Al (10 nm)/Fe (1 nm), Al_2O_3 (10 nm)/Fe (1 nm), to Al_2O_3 (10 nm)/Co (1 nm)]. VA-SWCNTs with the length of several mm and diameter of 1–3 nm were reported. This approach also enabled the growth of double- and multi-walled CNTs [10] (Fig. 3).

The abovementioned *CCVD* approaches are achievable at relatively low temperatures and lead to VA-SWCNTs. Historically SWCNTs were grown by *CCVD* at higher temperatures in a powder form. In 1996, Dai et al. demonstrated the growth of SWCNTs at 1200 °C by using carbon monoxide (CO) as the carbon source gas and supported MoO_x powder as the catalyst [11]. In this case, powder form of SWCNTs was grown on the catalyst that was loaded on a quartz boat. Later, the growth temperature was being able to reduce to 1000 °C by using CH_4 as the source gas, resulting in the growth of randomly distributed SWCNTs on powders of supported metal-oxide catalysts [12]. The diameters of these SWCNTs are 1–6 nm. In addition, *CCVD* could also be modified into a high-pressure mode for large-scale synthesis of SWCNTs [13]. In this approach, liquid iron pentacarbonyl, $Fe(CO)_5$, was used as the catalyst and was introduced into the CVD chamber by CO gas. At operational pressures of 1–10 atm and temperatures of 800–1200 °C, $Fe(CO)_5$ will thermally be decomposed into iron clusters in gas phase and react with CO gas to produce SWCNTs. The yield of SWCNTs was found to increase with temperature and pressure. The average diameter of SWCNTs was decreased from ~1.0 nm at 1 atm to 0.8 nm at 10 atm. This approach is now known as high-pressure carbon monoxide (HiPCO) technique, one of the major techniques for large-scale synthesis of SWCNTs with small diameters. Finally, alcohol was also used as the carbon source for the synthesis of SWCNTs by *CCVD* [14]. In this case, ethanol vapors (5 torr) were supplied to the reaction chamber that contained Fe/Co catalyst supported with zeolite at 700–800 °C.

Synthesis of Graphene by CCVD

CCVD is also applicable for the synthesis of two-dimensional (2D) materials. For example, the growth of graphene by CCVD has been demonstrated using Ni [15, 16] and Cu plates [17]. Theory suggests that the decomposition of hydrocarbon gases on Ni or Cu are similar to the dissociative adsorption discussed for the growth of CNTs discussed earlier with differences on the atomic configuration [18]. The major differences are (1) carbon atoms dissolve into the bulk / sub-surface of the Ni plate and led to the formation of few layered graphene (FLG); (2) contrarily, the growth process is limited on the surface in the Cu case and result in the formation of mono layered graphene (MLG) [17]. Theory also suggests that the growth of graphene on Cu surface by methane (CH4) gas could happened with hydrocarbon radical instead of free carbon atoms [18].

The nucleation of 2D materials such as graphene is also different from that in 1D CNTs. for the case of CNTs, nucleation is confined on

the surface and subsurface regions of the zero dimensional metal particles. The nucleation of graphene involves a lot more surface diffusion processes [19]. As shown in Fig. 3, the process involves the transformation of 1D carbon chain to 2d sp2 carbon clusters/networks. Furthermore, it was suggested that graphene nucleation near a step edge is more preference to that on a terrace.

Synthesis of ZnO Nanostructures by CCVD

On the other hand, *CCVD* can be modified into a *double-tube* configuration as shown in Fig. 1b. This approach will enable the use of solid precursors to generate the needed growth vapors. For example, various ZnO nanostructures can be grown by such a setup by using ZnO and graphite powders as the precursors [20, 21]. As shown, these powders were loaded on a ceramic combustion boat which is contained at the end of a closed-end quartz tube. The substrates can be placed a distance away from the boat. The following reactions will occur at 1100 °C when oxygen gas (O_2) is introduced:

$$\mathrm{ZnO\,(s)\ +\ C\,(s)\ \rightarrow\ Zn\,(g)\ +\ CO\,(g)}$$

$$\mathrm{2Zn\,(g)\ +\ O_2(g)\ \rightarrow\ 2ZnO\,(s)}$$

Based on this approach, various ZnO nanostructures can be grown with and without the use of catalyst (gold films, Au), including nanotubes [20], nanowires, nanobelts, nanocombs [21], and nanosquids [22].

Synthesis of Boron Nitride Nanotubes by CCVD

The double-tube *CCVD* configuration in Fig. 1b was also used for the growth of boron nitride nanotubes (BNNTs) [23–25]. In this case, boron, magnesium oxide, and iron oxide were used as the precursors and were loaded on the ceramic boat. The possible chemical reaction at 1100–1200 °C is 4B (s) + MgO (s) + FeO (s) → $2B_2O_x$ (g) + MgO_y (s) + FeO_z (s), where x, y, and z are yet to be determined. When ammonia gas (NH_3) is introduced into the chamber, the generated boron oxide vapors (B_2O_x) will react with NH_3 to form BNNTs. The partially reduced MgO_y and FeO_z are the possible catalysts for the formation of BNNTs [23]. It is noted that the synthesis of BNNTs by these chemical processes usually requires temperatures above 1350 °C. The key success for the low-temperature growth discussed here is due to the so-called "growth vapor trapping" approach obtained by placing the bare substrates directly on the ceramic boat. These substrates trapped the growth vapors from the precursors and enhanced the nucleation rate of BNNTs at low temperatures. In recent experiments, Fe, Ni, or MgO films were coated on the substrates used as the catalysts [24, 25]. Such an approach leads to patterned growth of BNNTs. The possible chemical reactions are

$$\mathrm{B_2O_2(g)\ +\ MgO\,(s)\ +\ 2NH_3(g)\ \rightarrow\ 2BN\,(s,\ BNNTs)\ +\ MgO(s)\ +\ 2H_2O(g)\ +\ H_2(g),\ or}$$

$$\mathrm{B_2O_2(g)\ +\ Ni\ (s)\ or\ Fe\ (s)\ +\ 2NH_3(g)\ \rightarrow\ 2BN\ (s,\ BNNTs)\ +\ Ni\ (s)\ or\ Fe\ (s)\ +\ 2H_2O\ (g)\ +\ H_2(g)}$$

Acknowledgments Yoke Khin Yap acknowledges the support from the National Science Foundation (Award number DMR-1261910).

Cross-References

- ▶ Atomic Layer Deposition
- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Carbon-Nanotubes
- ▶ Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)
- ▶ Synthesis of Carbon Nanotubes

References

1. Tsu, D.V., Lucovsky, G., Dvidson, B.N.: Effects of the nearest neighbors and the alloy matrix on SiH stretching vibrations in the amorphous SiOr:H (0<r<2) alloy system. Phys. Rev. B. **40**, 1795–1805 (1989)
2. Menda, J., et al.: A Dual-RF-Plasma Approach for Controlling the Graphitic Order and Diameters of Vertically-Aligned Multiwall Carbon Nanotubes. Appl. Phys. Lett. **87**, 173106 (2005) (3 pp)

3. Hirao, T., et al.: Formation of vertically aligned carbon nanotubes by dual-RF-plasma chemical vapor deposition. Jpn. J. Appl. Phys. **40**, L631–L634 (2001)
4. van Laake, L., Hart, A.J., Slocum, A.H.: Suspended heated silicon platform for rapid thermal control of surface reactions with application to carbon nanotube synthesis. Rev. Sci. Instrum. **78**, 083901 (2007) (9 pp)
5. Leskelä, M., Ritala, M.: Atomic layer deposition chemistry: Recent developments and future challenges. Angew. Chem. Int. Ed. **42**, 5548–5554 (2003)
6. Kayastha, V.K., et al.: Controlling dissociative adsorption for effective growth of carbon nanotubes. Appl. Phys. Lett. **85**, 3265–3267 (2004)
7. Kayastha, V.K., et al.: High-density vertically aligned multiwalled carbon nanotubes with tubular structures. Appl. Phys. Lett. **86**, 253105 (2005) (3 pp)
8. Kayastha, V.K., et al.: Synthesis of vertically aligned single- and double- walled carbon nanotubes without etching agents. J. Phys. Chem. C **111**, 10158–10161 (2007)
9. Hata, K., et al.: Water-assisted highly efficient synthesis of impurity-free single-walled carbon nanotubes. Science **306**, 1362–1364 (2004)
10. Yamada, T., et al.: Size-selective growth of double-walled carbon nanotube forests from engineered iron catalysts. Nat. Nanotechnol. **1**, 131–136 (2006)
11. Dai, H., et al.: Single-wall nanotubes produced by metal-catalyzed disproportionation of carbon monoxide. Chem. Phys. Lett. **260**, 471–475 (1996)
12. Kong, J., Cassell, A.M., Dai, H.: Chemical vapor deposition of methane for single-walled carbon nanotubes. Chem. Phys. Lett. **292**, 567–574 (1998)
13. Nikolaev, P., et al.: Gas-phase catalytic growth of single-walled carbon nanotubes from carbon monoxide. Chem. Phys. Lett. **313**, 91–97 (1999)
14. Maruyama, S., et al.: Low-temperature synthesis of high-purity single-walled1 carbon nanotubes from alcohol. Chem. Phys. Lett. **360**, 229–234 (2002)
15. Obraztsov, A.N., Obraztsova, E.A., Tyurnina, A.V., Zolotukhin, A.A.: Chemical vapor deposition if thin graphite films of nanometer thickness. Carbon **45**, 2017–2021 (2007)
16. Kim, K.S. et al.: Large-scale pattern growth of graphene films for stretchable trasparent electrodes. Nature **457**, 706–710 (2009)
17. Li, X.S. et al.: Large-area synthesis of high-quality and uniform graphene films on copper foils. Science **324**, 1312–1314 (2009)
18. Zhang. W., Wu, P., Li, Z., Yang, J.: First-principles thermodynamics of graphene growth on Cu surfaces. J. Phys. Chem. C **115**, 17782–17787 (2011)
19. Gao, J., Yop, J., Zhao, J., Yakobson, B.I., Ding, F.: Graphene nucleation on transition metal surface: Structure transformation and role of the metal step edge. J. Am. Chem. Soc. **133**, 5009–5015 (2011)
20. Mensah, S.L., et al.: Formation of single crystalline ZnO nanotubes without catalysts and templates. Appl. Phys. Lett. **90**, 113108 (2007)
21. Mensah, S.L., et al.: Selective growth of pure and long ZnO nanowires by controlled vapor concentration gradients. J. Phys. Chem. C **111**, 16092–16095 (2007)
22. Mensah, S.L., et al.: ZnO nnosquids: banching nnowires from nnotubes and nnorods. J. Nanosci. Nanotechnol. **8**, 233–236 (2008)
23. Lee, C.H., et al.: Effective growth of boron nitride nanotubes by thermal chemical vapor deposition. Nanotechnology **19**, 455605 (2008)
24. Lee, C.H., et al.: Patterned growth of boron nitride nanotubes by catalytic chemical vapor deposition. Chem. Mater. **22**, 1782–1787 (2010)
25. Wang, J., Lee, C.H., Yap, Y.K.: Recent advancements in boron nitride nanotubes. Nanoscale. **2**, 2028–2034 (2010)

Chemical Vapor Machining (CVM)

▶ Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Chemical-Assisted Ion Beam Etching (CAIBE)

▶ Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Chemistry of Carbon Nanotubes

▶ Functionalization of Carbon Nanotubes

Chemotherapeutic

▶ Use of Nanotechnology in Pregnancy

Chitosan

▶ Chitosan Nanoparticles

Chitosan Nanoparticles

Burcu Aslan[1], Hee Dong Han[2,4], Gabriel Lopez-Berestein[1,3,4,5] and Anil K. Sood[2,3,4,5]
[1]Department of Experimental Therapeutics, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA
[2]Gynecologic Oncology, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA
[3]Cancer Biology, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA
[4]Center for RNA Interference and Non–coding RNA, M.D. Anderson Cancer Center, The University of Texas, Houston, TX, USA
[5]The Department of Nanomedicine and Bioengineering, UTHealth, Houston, TX, USA

Synonyms

Chitosan; Drug delivery system; Nanoparticles

Definition

Chitosan nanoparticles are biodegradable, nontoxic carriers for nucleotides and drugs with the potential for broad applications in human disease.

Overview

Characteristics of Chitosan

Chitosan is a natural cationic polysaccharide composed of randomly distributed *N*-acetyl-D-glucosamine and β-(1,4)-linked D-glucosamine. Chitosan can be chemically synthesized via alkaline deacetylation from chitin, which is the principal component of the protective cuticles of crustaceans [1]. Chitosan is biodegradable in vivo by enzymes such as lysozyme, which is endogenous and nontoxic [2]. In addition, biodegradation of chitosan is highly associated with the degree of deacetylation. These properties render chitosan particularly attractive for clinical and biological applications as a highly biocompatible material with low toxicity and immunogenicity.

Several techniques have been designed to assemble chitosan nanoparticles (CH-NPs) as a drug delivery system including emulsions, ionotropic gelation, micelles, and spray drying. A variety of therapeutic agents can be loaded into CH-NPs with high efficiency, which can then be injected intravenously, intraperitoneally, or intrathecally.

Chemical Modification of Chitosan

The abundant amine and hydroxyl groups present in chitosan offer a unique opportunity to attach targeting ligands or imaging agents. Numerous derivatives of chitosan have been designed and tailored to improve the physicochemical and adhesive properties of nanoparticles such as size, shape, charge, density, and solubility. Quaternized chitosan, *N*,*N*,*N*-trimethyl chitosan, thiolated chitosan, carboxyalkyl chitosan, sugar-bearing chitosan, bile acid-modified chitosan, and cyclodextrin-linked chitosan are among the modifications frequently utilized in chitosan-based drug delivery systems. Each modification offers unique properties and characteristics. For instance, trimethyl chitosan is soluble over a wide pH range and enhances the condensation capacity of plasmid DNA at neutral pH due to fixed positive charges on its backbone. Thiolation of chitosan provides free sulphydryl groups on its side chains and forms disulfide bonds with cysteine-rich subdomains of mucoglycoproteins on cell membranes and increases cellular uptake. In addition, both modifications have been used as nonviral carrier systems to combine the advantages of trimethyl chitosan and thiolated chitosan while minimizing their shortcomings [3].

Hydrophobic moieties can also be attached to chitosan to facilitate the incorporation of insoluble drugs, i.e., hydrophobic glycol chitosan nanoparticles. Chemical conjugation of hydrophobic 5β-cholanic acid to the hydrophilic glycol chitosan backbone allows for the incorporation of the water-insoluble drug camptothecin [4].

Key Research Findings

Preparation of Chitosan Nanoparticles

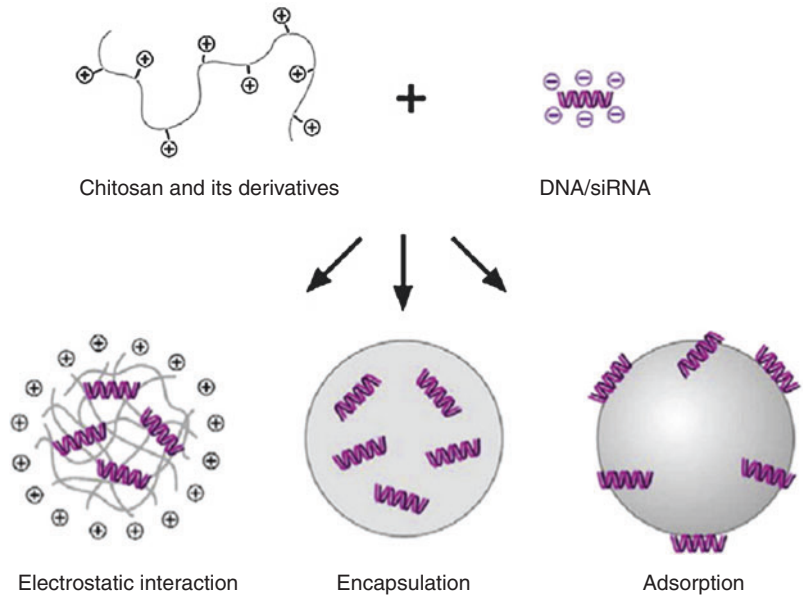
The amino groups of chitosan backbone can interact with anionic molecules such as tripolyphosphate (TPP). The ionic cross-linking of chitosan is advantageous since the method is easy and mostly performed under mild conditions without using organic solvents. Ionotropic gelation of chitosan using TPP for the incorporation of low molecular drugs, proteins, DNA/siRNA have been demonstrated [5]. CH-NPs are rapidly formed through ionic interactions between the negatively charged phosphates of TPP and positively charged amino groups of chitosan.

Drug Delivery

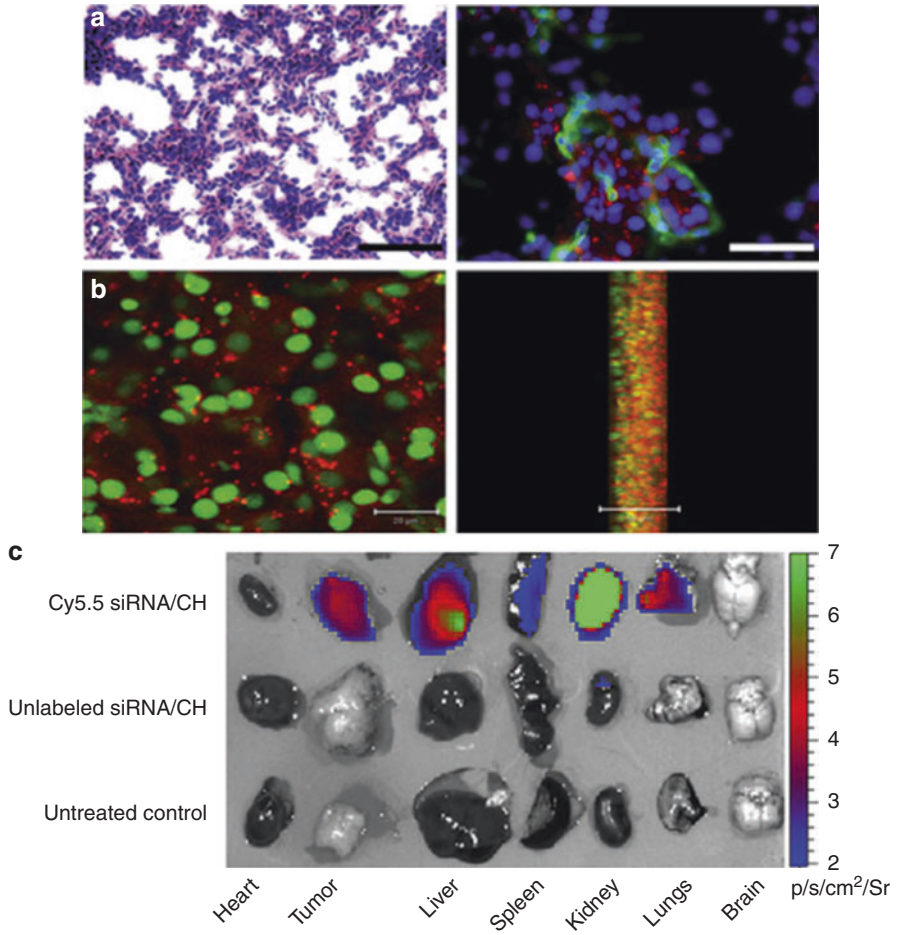
Ringsdorf first reported the concept of polymer–drug conjugates for delivering small molecule drugs [6]. The concept of polymer–drug conjugates allows chemical conjugation of a drug using a biodegradable spacer. The spacer is usually stable in the bloodstream, but cleaved at the target site by hydrolysis or enzymatic degradation. Based on this concept, several polymer–drug conjugates have been developed such as glycol chitosan conjugated with doxorubicin, which forms self-assembled nanoparticles in an aqueous condition. A paclitaxel-chitosan conjugate that can be cleaved at physiological conditions was developed for oral delivery of paclitaxel. *N*-succinyl-chitosan-mitomycin conjugates demonstrated high antitumor efficacy against a variety of murine tumor models of leukemia, melanoma, and primary and metastatic liver tumors.

Gene Delivery

Recently, plasmid DNA, siRNA, and oligonucleotide-loaded CH-NPs have been used for targeted gene silencing. Positively charged chitosan can easily form polyelectrolyte complexes with negatively charged nucleotides based on electrostatic interaction, incorporation, or adsorption, as illustrated in Fig. 1 [7]. The positive charge on the surface of CH-NPs is desirable to prevent aggregation due to electrostatic repulsion and increase binding efficiency with the negatively charged cell membrane by enhancing electrostatic interactions. However, other cellular uptake mechanisms such as clathrin-mediated endocytosis, caveolae-mediated endocytosis, and macropinocytosis may also be involved [8]. Moreover, the configuration of chitosan is modified under acidic pH by triggering the opening of tight junctions. Acidic pH results in an increase in the number



Chitosan Nanoparticles, Fig. 1 Preparation of chitosan-based DNA/siRNA nanoparticles based on different mechanisms [7]



Chitosan Nanoparticles, Fig. 2 (**a**) Fluorescent siRNA distribution in tumor tissue. Hematoxylin and eosin, original magnification 2003 (*left*); tumor tissues were stained with anti-CD31 (*green*) antibody to detect endothelial cells (*right*). The scale bar represents 50 mm. (**b**) Fluorescent siRNA distribution in tumor tissue. Sections (8 mm thick) were stained with Sytox green and examined with confocal microscopy (scale bar represents 20 mm) (*left*); lateral view (*right*). Photographs taken every 1 mm were stacked and examined from the lateral view. Nuclei were labeled with Sytox green and fluorescent siRNA (*red*) was seen throughout the section. At all time points, punctated emissions of the siRNA were noted in the perinuclear regions of individual cells, and siRNA was seen in >80 % of fields examined. (**c**) Shows fluorescence intensity overlaid on white light images of different mouse organs and tumor

of protonated amines on the chitosan leading to a further disruption on membrane organization. It has also been reported that chitosan can swirl across the membrane lipid bilayer and facilitate the cellular uptake of the polyplex due to increase in mole fraction of chitosan, leading to reduction in the polymeric chains which results in decreased molecular weight [7]. In addition, Lu et al. have recently reported the biodistribution of intravenously administered siRNA-CH-NPs in tumor-bearing mice. They demonstrated that CH-NPs allowed for a higher localization of siRNA in tumor tissues compared to other organs (Fig. 2) [9].

Electrostatic interactions between protonated amines of chitosan and negative charge of DNA or siRNA leads to spontaneous formation of highly compact encapsulation of either DNA or siRNA into CH-NPs [10]. Gel electrophoresis has been used to assess hydrogen bonding and hydrophobic interactions between chitosan and DNA [11]. However, a major limitation of siRNA

delivery is its rapid degradation in plasma and cytoplasm. It has been reported that stable chitosan/siRNA complexes can protect siRNA degradation in circulation of CH-NPs in bloodstream to overcome extracellular and intracellular barriers. On the other hand, disassembly is also needed to allow release of siRNA. This emphasizes the importance of an appropriate balance between protection and release of siRNA for biological functionality [12].

Positively charged CH-NPs can bind to negatively charged cell surfaces with high affinity. CH-NPs are known to overcome endosomal escape via its "proton sponge effect." Once these nanoparticles penetrate into an acidifying lysosomal compartment, the unsaturated amino groups of chitosan distrain protons that are delivered by proton pumps (vATPase) which is called the "proton sponge mechanism." Subsequent lysosomal swelling and rupture leads to endolysosomal escape of nanoparticles [13].

Targeted Delivery

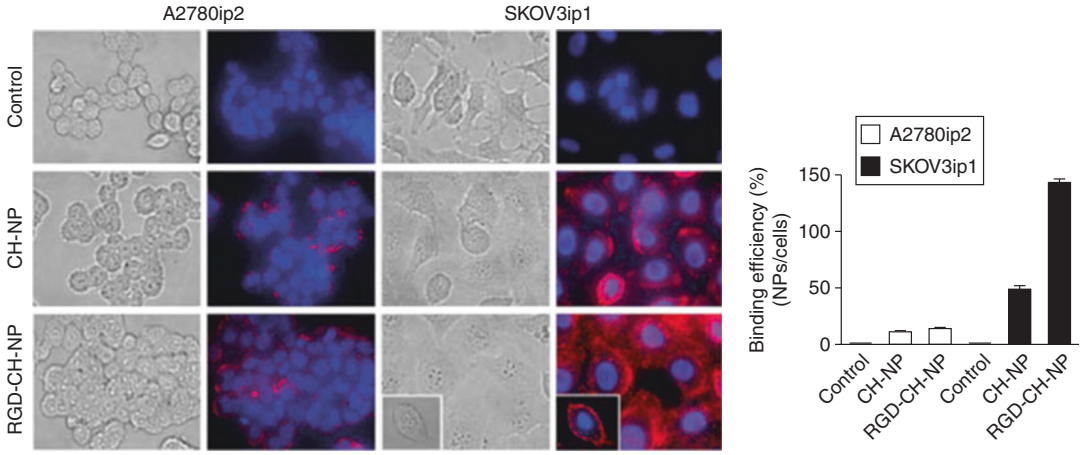
An ideal delivery system should lead to enhanced concentrations of therapeutic payloads at disease sites, and minimize potential non-desirable off-target effects, and ultimately raise the therapeutic index. Differences between tumor and normal tissue microenvironment and architecture, such as vascularization, overexpressed receptors, pH, temperature, ionic strength, and metabolites, can be exploited for selective targeting.

Targeted delivery systems have been designed to increase and/or facilitate uptake into target cells, and to protect therapeutic payloads. Recent work comparing non-targeted and targeted nanoparticles have shown that the primary role of the targeting ligands is to enhance selective binding efficiency to receptor in cell surface and cellular uptake into target cells, and to minimize accumulation in normal tissues. The addition of targeting ligands that provide specific ligand-receptor binding on nanoparticle-cell surface interactions can play a vital role in the ultimate location of nanoparticles. For example, nanoparticles decorated with specific moieties such as peptides, proteins, or antibodies can be targeted to cancer cells via cell-surface receptor proteins such as transferrin or folate receptors (known to be increased on a wide range of cancer cells). These targeting ligands enable nanoparticles to bind on to cell-surface receptors and penetrate cells by receptor-mediated endocytosis.

Target selective ligand-labeled CH-NPs can enhance receptor-mediated endocytosis. Various receptors on the tumor cell surface have been established as a target-binding site to achieve selective delivery. The overexpression of transferrin and folate in certain tumors has been exploited to deliver CH-NPs conjugated with these receptor's ligands [14, 15]. Another example is the $\alpha v\beta 3$ integrin, which is overexpressed in a wide range of tumors, and is largely absent in normal tissues. Han et al. [16] have recently reported that the administration of RGD peptide-labeled CH-NPs led to increased tumor delivery of siRNA-CH-NP and enhanced antitumor activity in ovarian carcinoma models (Figs. 3 and 4).

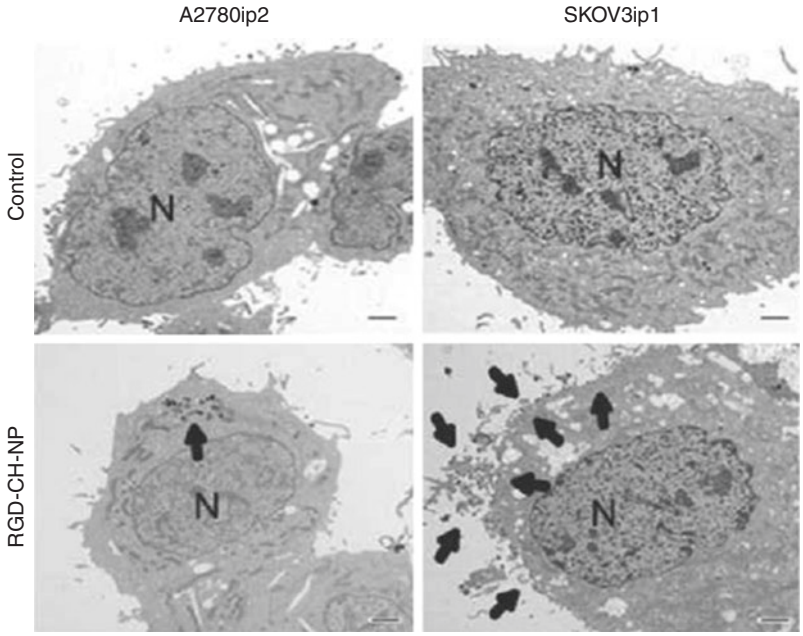
Chitosan-Based Environmental-Responsive Particles

Physiological alterations such as pH, temperature, ionic strength, and metabolites in the microenvironment of tumor have gained increased interest in terms of targeted therapy. Potential differences in these parameters between tumor and normal tissue can be used for enhanced targeting. A novel, pH responsive NIPAAm/CH-NPs containing camptothecin and paclitaxel was successfully used to enhance tumor uptake and antitumor activity [17]. On the other hand, thermosensitive CH-g-poly(*N*-vinylcaprolactam) composite has been developed by an ionic cross-linking method and incorporated with 5-FU (5-fluorouracil). This study showed that 40 % of drug is released from the particles when the temperature is above a lower critical solution temperature of 38 °C while only 5 % of drug is released below 38 °C, which confirms the drug release mechanism of this polymeric carrier system is based on temperature. These nanoparticles were more toxic to cancer cells while devoid of toxicity to normal cells, leading to enhance antitumor efficacy [18].



Chitosan Nanoparticles, Fig. 3 Binding of Alexa555 siRNA/RGD-CH-NPs and Alexa555 siRNA-CH-NP in SKOV3ip1 or A2780ip2 cells by fluorescence microscopy

Chitosan Nanoparticles, Fig. 4 Binding of siRNA/RGD-CH-NPs in SKOV3ip1 or A2780ip2 cells by transmission electron microscopy against ovarian cancer cells in vitro



Chitosan-Based Magnetic Nanoparticles

Chitosan-based magnetic nanoparticles have been developed for magnetic resonance (MR) imaging via passive and tissue-specific targeting. The low-oxidizing ferromagnetic materials are the most commonly used compounds for nanoparticle formulations and provide a stable magnetic response. The accumulation of magnetic nanoparticles when injected intravenously can be induced when the tissue or organ is subjected to an external high-gradient magnetic field [19]. Drugs, DNA plasmids, or bioactive molecules are released into target tissues and effectively taken up by tumor cells after accumulation of these carriers. On the other hand, magnetic nanoparticles such as iron oxide nanoparticles

(Fe_3O_4) are applied to oscillating magnetic fields, which results in the generation of heat and holds potential for rapid heating of tumor tissue.

When magnetic nanoparticles are administered systemically, they are rapidly coated with plasma proteins. These particles are taken up by the reticuloendothelial system, leading to decreased circulation time. Magnetic nanoparticles coated with hydrophilic materials such as chitosan could provide longer circulation time. In addition, chitosan-coated magnetic nanoparticles, which are used in magnetic resonance imaging, can also be functionalized and used as theranostic carriers due to amino and hydroxyl groups of chitosan.

Application of Chitosan Nanoparticles for SiRNA Delivery

Recently, siRNA targeted to EZH2 (a critical component of the polycomb repressive complex 2 [PRC2]), loaded CH-NPs were shown to enhance the delivery of siRNA to tumors, leading to downregulation of the target protein and subsequently enhanced antitumor activity. Lu et al. demonstrated target gene silencing using EZH2 siRNA loaded CH-NPs, leading to increased antitumor efficacy in animal tumor models. Moreover, Zhang et al. reported the use of intrathecal administration of siRNA targeted against specific muscarinic receptor subtypes loaded in CH-NPs in a rat model of pain. CH-NP/siRNA was distributed to the spinal cord and the dorsal root ganglion [20]. The administration of Chitosan-M_2-siRNA caused a large reduction in the inhibitory effect of muscarine on the rat paw withdrawal threshold from a heat stimulus. These studies support the use of CH-NPs as a delivery system for siRNA into neuronal tissues in vivo.

Future Directions for Research

Chitosan nanoparticles offer a unique potential for clinical and biological applications due to low immunogenicity, low toxicity, and high biocompatibility. In addition to its advantages such as protonated amine groups, chitosan can increase binding efficiency with cells because of electrostatic interactions. Therefore, CH-NPs may be used for broad applications in human disease. Moreover, chitosan allow modifications that will exploit the inherent physicochemical properties by conjugation of selective ligands. The highly desirable specific targeting of drugs has been elusive to date; however, CH-NPs can bring us closer to this goal.

Acknowledgments Portions of this work were supported by the NIH (CA 110793, 109298, P50 CA083639, P50 CA098258, CA128797, RC2GM092599, U54 CA151668), the Ovarian Cancer Research Fund, Inc. (Program Project Development Grant), the DOD (OC073399, W81XWH-10-1-0158, BC085265), the Zarrow Foundation, the Marcus Foundation, the Kim Medlin Fund, the Laura and John Arnold Foundation, the Estate of C. G. Johnson, Jr., the RGK Foundation, and the Betty Anne Asche Murray Distinguished Professorship.

Cross-References

- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Nanomedicine
- ▶ Nanoparticle Cytotoxicity
- ▶ Nanoparticles

References

1. Muzzarelli, R.A.A.: Chitin, pp. 1–37. Pergamon, Elmsford (1977)
2. Khor, E.: Chitin: Fulfilling a Biomaterials Promise. Elsevier, Oxford (2001)
3. Zhao, X., Yin, L., Ding, J., Tang, C., Gu, S., Yin, C., Mao, Y.: Thiolated trimethyl chitosan nanocomplexes as gene carriers with high in vitro and in vivo transfection efficiency. J. Control. Release **144**, 46–54 (2010)
4. Min, K.H., Park, K., Kim, Y., Bae, S.M., Lee, S., Jo, H. G., Park, R.W., In-San Kim, I.S., Jeong, S.Y., Kim, K., Kwon, I.C.: Hydrophobically modified glycol chitosan nanoparticles-encapsulated camptothecin enhance the drug stability and tumor targeting in cancer therapy. J. Control. Release **127**, 208–218 (2008)
5. Park, J.H., Saravanakumar, G., Kim, K., Kwon, I.C.: Targeted delivery of low molecular drugs using chitosan and its derivatives. Adv. Drug Deliv. Rev. **62**, 28–41 (2010)
6. Ringsdorf, H.: Structure and properties of pharmacologically active polymers. J. Polym. Sci. Polym. Symp. **51**, 135–153 (1975)

7. Lai, W.F., Lin, M.C.M.: Nucleic acid delivery with chitosan and its derivatives. J. Control. Release **134**, 158–168 (2009)
8. Nam, H.Y., Kwon, S.M., Chung, H., Lee, S.Y., Kwon, S.H., Jeon, H., Kim, Y., Park, J.H., Kim, J., Her, S., Oh, Y.K., Kwon, I.C., Kim, K., Jeong, S.Y.: Cellular uptake mechanism and intracellular fate of hydrophobically modified glycol chitosan nanoparticles. J. Control. Release **135**, 259–267 (2010)
9. Lu, C., Han, H.D., Mangala, L.S., Ali-Fehmi, R., Newton, C.S., Ozbun, L., Armaiz-Pena, G.N., Hu, W., Stone, R.L., Munkarah, A., Ravoori, M.K., Shahzad, M.M.K., Lee, J.W., Mora, E., Langley, R. R., Carroll, A.R., Matsuo, K., Spannuth, W.A., Schmandt, R., Jennings, N.J., Goodman, B.W., Jaffe, R.B., Nick, A.M., Kim, H.S., Guven, E.O., Chen, Y. H., Li, L.Y., Hsu, M.C., Coleman, R.L., Calin, G.A., Denkbas, E.B., Lim, J.Y., Lee, J.S., Kundra, V., Birrer, M.J., Hung, M.C., Lopez-Berestein, G., Sood, A.K.: Regulation of tumor angiogenesis by EZH2. Cancer Cell **18**, 185–197 (2010)
10. Mao, S., Sun, W., Kissel, T.: Chitosan-based formulations for delivery of DNA and siRNA. Adv. Drug Deliv. Rev. **62**, 12–27 (2010)
11. Messai, I., Lamalle, D., Munier, S., Verrier, B., Ataman-Onal, Y., Delair, T.: Poly(D, Llactic acid) and chitosan complexes: interactions with plasmid DNA. Colloids Surf. A Physicochem. Eng. Asp. **255**, 65–72 (2005)
12. Liu, X., Howard, K.A., Dong, M., Andersen, M.O., Rahbek, U.L., Johnsen, M.G., Hansen, O.C., Besenbacher, F., Kjems, J.: The influence of polymeric properties on chitosan/siRNA nanoparticle formulation and gene silencing. Biomaterials **28**, 1280–1288 (2007)
13. Nel, A.E., Mädler, L., Velegol, D., Xia, T., Hoek, E.M. V., Somasundaran, P., Klaessig, F., Castranova, C., Thompson, M.: Understanding biophysicochemical interactions at the nano–bio interface. Nat. Mater. **8**, 543–557 (2009)
14. Mao, H.Q., Roy, K., Troung-Le, V.L., Janes, K.A., Lin, K.Y., Wang, Y., August, J.T., Leong, K.W.: Chitosan–DNA nanoparticles as gene carriers: synthesis, characterization and transfection efficiency. J. Control. Release **70**(3), 399–421 (2001)
15. Fernandes, J.C., Wang, H., Jreyssaty, C., Benderdour, M., Lavigne, P., Qiu, X., Winnik, F.M., Zhang, X., Dai, K., Shi, Q.: Bone-protective effects of nonviral gene therapy with Folate–Chitosan DNA nanoparticle containing Interleukin-1 receptor antagonist gene in rats with adjuvant-induced arthritis. Mol. Ther. **16**(7), 1243–1251 (2008)
16. Han, H.D., Mangala, L.S., Lee, J.W., Shahzad, M.M. K., Kim, H.S., Shen, D., Nam, E.J., Mora, E.M., Stone, R.L., Lu, C., Lee, S.J., Roh, J.W., Nick, A.M., Lopez-Berestein, G., Sood, A.K.: Targeted gene silencing using RGD-Labeled chitosan nanoparticles. Clin. Cancer Res. **16**(15), 3910–3922 (2010)
17. Li, F., Wu, H., Zhang, H., Gu, C.H., Yang, Q.: Antitumor drug paclitaxel loaded pH-sensitive nanoparticles targeting tumor extracellular pH. Carbohydr. Polym. **77**(4), 773–778 (2009)
18. Rejinold, N.S., Chennazhi, K.P., Nair, S.V., Tamura, H., Jayakumar, R.: Carbohydr. Polym. (2010). doi:10.1016/j.carbpol.2010.08.052
19. Veiseh, O., Gunn, J.W., Zhang, M.: Design and fabrication of magnetic nanoparticles for targeted drug delivery and imaging. Adv. Drug Deliv. Rev. **62**, 284–304 (2010)
20. Zhang, H.M., Chen, S.R., Cai, Y.Q., Richardson, T.E., Driver, L.C., Lopez-Berestein, G., Pan, H.L.: Signaling mechanisms mediating muscarinic enhancement of GABAergic synaptic transmission in the spinal cord. Neuroscience **158**, 1577–1588 (2009)

Clamping Loss

▶ Anchor Loss in MEMS/NEMS

Clinical Adhesives

▶ Bioadhesives

Cluster

▶ Synthesis of Subnanometric Metal Nanoparticles

C-MEMS

▶ Carbon MEMS

CMOS (Complementary Metal-Oxide-Semiconductor)

▶ CMOS MEMS Fabrication Technologies

CMOS MEMS Biosensors

Michael S.-C. Lu
Department of Electrical Engineering, Institute of Electronics Engineering, and Institute of NanoEngineering and MicroSystems, National Tsing Hua University, Taiwan, Republic of China

Synonyms

Integrated biosensors

Definition

CMOS MEMS biosensors are miniaturized biosensors fabricated on CMOS (complementary metal-oxide semiconductor) chips by the MEMS (microelectromechanical systems) technology.

Overview

A biosensor is a device designed to detect a biochemical molecule such as a particular deoxyribonucleic acid (DNA) sequence or particular protein. Many biosensors are affinity-based, meaning that they use an immobilized capture probe that selectively binds to the target molecule being sensed. Most biosensors require a label attached to the target, and the amount of detected label is assumed to correspond to the number of bound targets. Labels can be fluorophores, magnetic beads, gold nanoparticles, enzymes, or anything else allowing convenient binding and detection; however, labeling a biomolecule can change the associated binding properties, especially for protein targets.

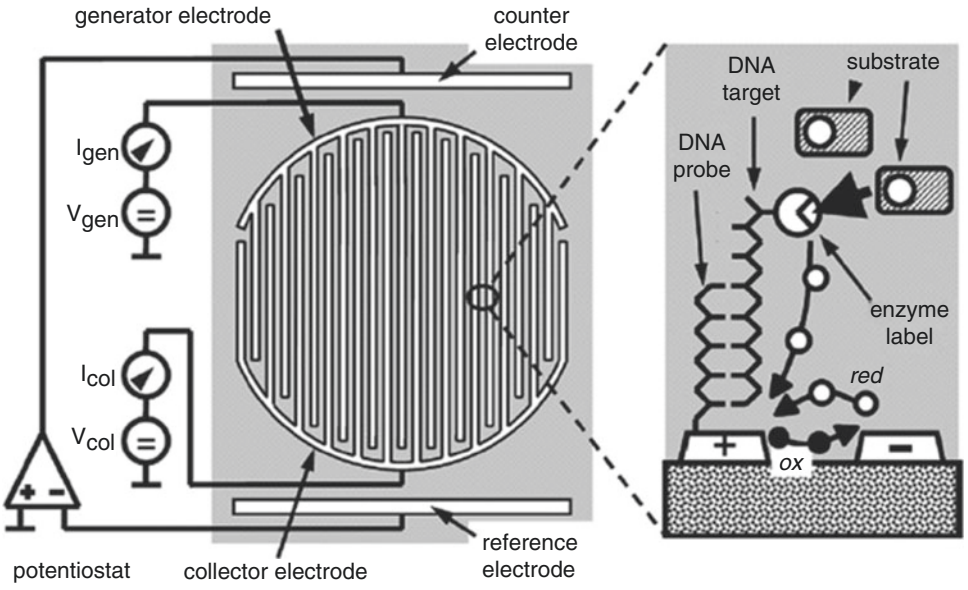
An electrical biosensor is capable of detecting a binding event by producing an electrical current and/or a voltage. Conventional optical detection methods require external instruments that are expensive and not amenable to miniaturization. Electrical bioassays hold great promise for numerous decentralized clinical applications ranging from emergency-room screening to point-of-care diagnostics due to their low cost, high sensitivity, specificity, speed, and portability. Miniaturization of an electrical biosensor can be achieved through microfabrication – widely known as the MEMS technology; in addition, the sensing circuits can be embedded on the same chip through integrated-circuit (IC) processes, among which the CMOS technology is the most popular choice for implementing various analog and digital circuits. A fully integrated CMOS MEMS biosensor array is capable of providing real-time high-throughput detection of multiple samples. CMOS biosensors can be implemented based on the electrochemical, impedimetric, ion-sensitive, magnetic, optical, and micromechanical approaches, which require different MEMS processes to construct the sensing interfaces. Some of the methods require labeling and some are label-free for bio-signal transduction. More details are provided in the following sections.

Sensing Principles and Key Research Findings

Electrochemical Biosensors

Electrochemical biosensors have been the subject of basic as well as applied research for many years. In 1970, Dr. Leland C. Clark demonstrated that glucose could be measured in whole blood with the presence of the glucose oxidase enzyme. Commercial glucose sensors based on electrochemical detection have been developed since then. Electrochemical biosensors typically depend on the presence of a suitable enzyme in the biorecognition layer to catalyze reaction of electroactive substances. Affinity-based electrochemical sensors use enzymes as labels that bind to antibodies, antigens, or oligonucleotides with a specific sequence.

Electrochemical biosensors can employ potentiometric, amperometric, and impedimetric sensing principles to convert the chemical information into a measurable electrical signal. In potentiometric devices, ion selective electrodes (ISE) are commonly used to transduce a biorecognition event into a potential signal between the working and the reference electrodes whose value, depending on the concentration of the analyte,



CMOS MEMS Biosensors, Fig. 1 Schematic of the CMOS interdigitated microelectrode for electrochemical DNA detection. On the *right*: schematic illustration of the redox-cycling process (Schienle et al. © 2004 IEEE)

can be predicted by the Nernst equation. The current flowing through the electrode is equal to or near zero.

Amperometric biosensors operate by applying a constant potential and monitoring the current associated with the reduction or oxidation of an electroactive species. The sensors can work in three- or four-electrode configurations. The former case consists of a reference, a working, and a counter electrode. The four-electrode setup has an additional working electrode such that oxidation and reduction take place at anode and cathode simultaneously. Working electrodes are normally made of noble metals which are critical to the operation of a sensor. Since these metals are not CMOS compatible, the electrodes must be formed after fabrication of CMOS circuits. Silver/silver chloride (Ag/AgCl) is commonly used as the reference electrode in many electrochemical biosensors. Since not all particles oxidized at the anode reach the cathode, a potentiostat, whose input and output are connected to a reference and to a counter electrode, respectively, is required to provide the difference current to the electrolyte and regulate the potential of the electrolyte to a constant value.

Interdigitated microelectrodes have been adopted in many electrochemical biosensors for quantitative analysis. The width and spacing of microelectrodes are reduced by microfabrication. The main advantage is the enhanced redox current due to fast redox recycling, as the chemical products produced at one side of the electrodes are readily collected at the other side of adjacent electrodes and regenerated to the original states. The relationship between the produced redox current and the analyte concentration has been derived by Aoki et al. [1].

Electrochemical DNA hybridization biosensors rely on the conversion of the DNA base-pair recognition event into an electrical signal. Schienle et al. [2] reported a CMOS electrochemical DNA sensor array with each sensing element consisting of interdigitated gold electrodes separated by 1 μm. As depicted in Fig. 1, probe molecules were immobilized on the gold surface through thiol coupling and the target molecules were tagged by an enzyme label (alkaline phosphatase). A chemical substrate (para-aminophenyl phosphate) was applied to the chip after hybridization. The enzyme label on the matched DNA strands cleaved the phosphate group and

generated an electroactive compound (para-aminophenol), which was subsequently oxidized and reduced as the indicator of successful DNA hybridization. Levine et al. [3] also reported a CMOS electrochemical DNA sensor array which was operated based on conventional cyclic voltammetry (CV). During the measurement the electrode potential was scanned up and down in order to produce the redox reactions associated with the ferrocene labels attached to DNA strands. In addition to DNA detection, CMOS electrochemical sensors have been realized to allow high-throughput detection of dopamine and catecholamine release from adrenal chromaffin cells [4], since the release of neurotransmitters from secretory vesicles of biological cells is closely related to the function of a nervous system.

Impedimetric Biosensors

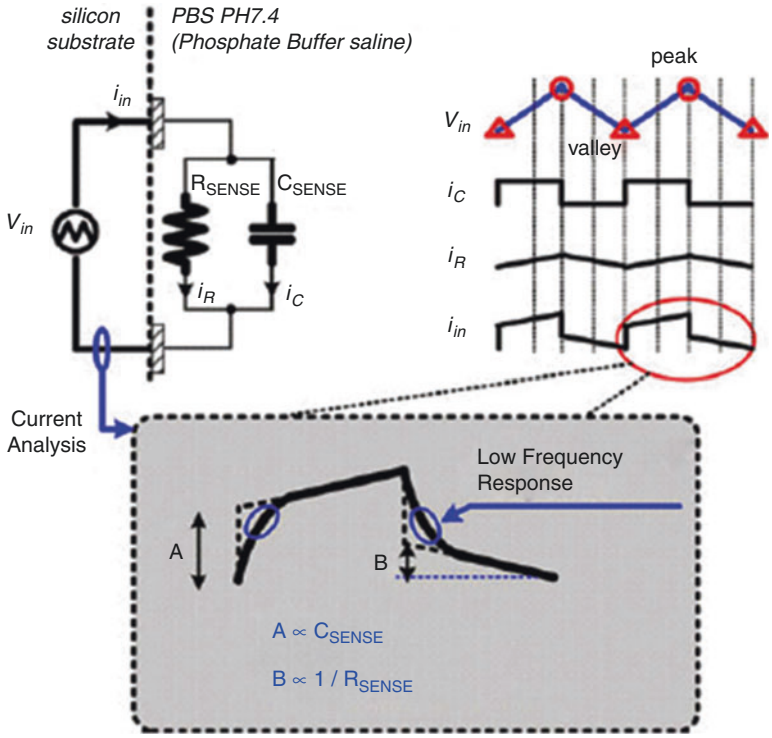
Changes in the electrical properties of a sensing interface (e.g., capacitance, resistance) can occur when a target biomolecule interacts with a probe-functionalized surface. A conventional impedimetric biosensor measures the electrical impedance of an electrode-solution interface in a.c. steady state with constant d.c. bias conditions. This approach, known as electrochemical impedance spectroscopy (EIS), is accomplished by imposing a small sinusoidal voltage over a range of frequencies and measuring the resulting current. The current–voltage ratio gives the impedance, which consists of both energy dissipation (resistor) and energy storage (capacitor) elements. Results obtained by EIS are often graphically represented by a Bode plot or a Nyquist plot. EIS reveals information about the reaction mechanism of an electrochemical process since different reaction steps can dominate at certain frequencies. Impedimetric biosensors can detect a variety of target analytes by simply varying the probe used. Changes in the impedance can be correlated to DNA hybridization, antigen–antibody reaction, or be used to detect biological cells.

The interface impedance is commonly represented by an equivalent circuit model for analysis. E. Warburg [5] first proposed that the interface impedance can be represented by a polarization resistance in series with a polarization capacitor. The interface capacitance possesses a frequency dependency and is commonly represented by the Gouy-Chapman-Stern model as the series combination of the double-layer capacitance (Helmholtz capacitance) and the diffuse layer capacitance (Gouy-Chapman capacitance).

In addition to the frequency-domain measuring method, interface impedance changes can be measured by the potentiostatic step method where small potential steps are applied to the working electrode and the transient current responses, as determined by the time constant of the interface resistance and capacitance, are measured accordingly. Lee et al. [6] reported a fully integrated CMOS impedimetric sensor array for label-free detection of DNA hybridization. The changes in the reactive capacitance and the charge-transfer resistance on the gold sensing electrodes were extracted by applying a triangular voltage waveform and monitoring the produced currents. As illustrated in Fig. 2, the current flowing through the capacitor is associated with the slope of the applied triangular wave, while the current flowing through the resistor is in proportion to the magnitude of the triangular wave. The reported detection limit was 10 nanomolar (nM).

It is important to distinguish the differences between faradaic and non-faradaic biosensors for impedance detection. In electrochemical terminology, a faradaic process involves charge transfer across an interface, while a transient current can flow without charge transfer in a non-faradaic process by charging a capacitor. The faradaic EIS requires the addition of a redox species which is alternately oxidized and reduced by the transfer of charges to and from the metal electrode. In contrast, no additional reagent is required for non-faradaic impedance spectroscopy. The associated impedance change is predominantly capacitive with the charge transfer resistance being omitted.

Miniaturized CMOS capacitive sensors have been developed for numerous biosensing applications. As the sensor size is small, monolithic integration provides the benefit of enhancing the signal-to-noise ratio by reducing the parasitic capacitance observed at the sensing node, which



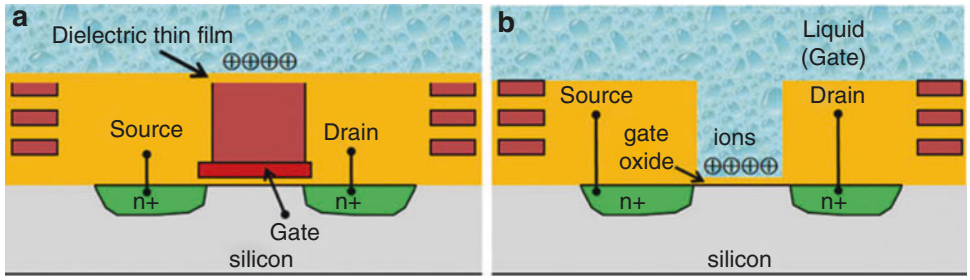
CMOS MEMS Biosensors, Fig. 2 Schematic of the impedance extraction method (Adopted by Lee et al. © 2010 Elsevier)

would otherwise negatively impact the detection limit during direct capacitance measurement. Stagni et al. [7] reported an 8 × 16 CMOS DNA sensor array where the bindings of complementary DNA strands on gold microelectrodes reduced the dielectric constant of the electrode-analyte impedance and the associated change was used to modify the charging and discharging transients of the detection circuit. Capacitive sensitivity can be enhanced by use of interdigitated microelectrodes with the minimum gap defined by the adopted CMOS process. Lu et al. [8] reported very sensitive detection of the neurotransmitter dopamine in the sub-femtomolar (fM) range. CMOS capacitive sensors have also been used for monitoring cellular activity since the morphological and physiological states of biological cells have correlations with their electrical properties [9].

The impedance change of a biosensing event can be made purely resistive with additional chemical modifications. Li et al. [10] reported a DNA array detection method in which the binding of immobilized DNA probes functionalized with gold nanoparticles produced a conductivity change between adjacent metal electrodes, which were made of gold metal in order to withstand the chemical in the clean processing. The detected signal can be further enhanced by additional silver deposition. Silicon dioxide was used as the underlying material under the gap for immobilization of the DNA probes. Detection limit of the target DNA concentration was 1 picomolar (pM).

Ion-Sensitive Field Effect Transistors (ISFET)

ISFETs were first developed in the early 1970s and have been utilized in various biosensing applications which depend on the type of receptor used for analyte recognition or how a signal is generated. Immunologically modified FETs and DNA-modified FETs detect the change of surface charges by monitoring the associated current–voltage relationship. Cell-based FETs detect the potential changes produced by biological cells due to the flow of ions across the cell membranes upon stimulations. For applications where it is required to build an ISFET array to provide detection of multiple samples at different locations,



CMOS MEMS Biosensors, Fig. 3 (**a**) Schematic of the floating-gate ISFET structure. (**b**) Schematic of the open-gate ISFET structure

monolithic integration is then preferred in order to reduce wiring complexity and noise interference.

CMOS ISFET sensors can be built by using a floating-gate or an open-gate structure as depicted in Fig. 3. The floating-gate structure is easier to fabricate as it requires minimal post-processing after completion of a conventional CMOS process. Silicon dioxide or silicon nitride in the CMOS passivation thin films can be used as the material for surface functionalization. The produced signal due to charges on sensor surface is capacitively coupled to the floating gate through a relatively thick dielectric layer which reduces the signal-coupling efficiency. The gate in an open-gate ISFET is removed and replaced by the aqueous solution whose potential is commonly set via a reference electrode. The exposed gate oxide is used for surface functionalization. Since the gate oxide thickness is only tens of angstroms in a conventional sub-μm CMOS process, sensitivity is thus significantly enhanced as compared to a floating-gate ISFET.

Accumulated charges on the ISFET surface modulate the threshold voltage of a MOS transistor, leading to a channel current change under fixed voltage biases of the drain, source, and gate terminals. The threshold voltage of an open-gate ISFET is expressed as:

$$V_{th} = E_{ref} - \Psi + \chi_{sol} - \frac{\phi_{si}}{q} - \frac{Q_{ox} + Q_{ss} + Q_B}{C_{ox}} + 2\phi_F$$

where E_{ref} is the reference electrode potential, Ψ is a chemical input parameter as a function of solution pH value, χ_{sol} is the surface dipole potential of the solvent, ϕsi is the work function of silicon, q is the electron charge, C_{ox} is the gate oxide capacitance per unit area, Q_{ox} and Q_{ss} are the charges in the oxide and at the oxide-silicon interface, Q_B is the depletion charges in silicon and ϕ_F is the difference between the Fermi potential of the substrate and intrinsic silicon.

Kim et al. [11] reported the use of p-type ISFET fabricated in a standard CMOS process for detection of DNA immobilization and hybridization. Gold was deposited by post-processing as the gate material for immobilizing DNA due to its chemical affinity with thiols. The channel current increased during hybridization due to the negative charges present in the phosphate groups of DNA strands. Li et al. [12] presented a post-CMOS fabrication method to make open-gate ISFETs and demonstrated ultrasensitive dopamine detection in the fM range.

Magnetic Biosensors

Some of the magnetic biosensors require the use of magnetic beads as labels attached to the samples in order to induce a measurable electrical signal when specific binding on sensor surface occurs. Magnetic beads are made of small ferromagnetic or ferrimagnetic nanoparticles that exhibit a unique quality referred to as superparamagnetism in the presence of an externally applied magnetic field. This phenomenon, as discovered by Louis Néel (Nobel Physics Prize winner in 1970), has been used in numerous applications such as magnetic data storage and magnetic resonance imaging (MRI).

Several methods can be used to electronically detect the existence of magnetic beads, such as the GMR (giant magnetoresistance) effect discovered by the 2007 Nobel Physics Prize winners Albert Fert and Peter Grünberg. The effect appears in thin film structures composed of alternating ferromagnetic and nonmagnetic layers. A significant change in the electrical resistance is observed depending on whether the magnetization of adjacent ferromagnetic layers is in a parallel alignment (the low-resistance state) due to the applied magnetic field or an anti-parallel alignment (the high-resistance state) in the absence of the magnetic field.

Han et al. [13] reported a CMOS DNA sensor array that adopted the spin valve structure to observe the GMR effect. Magnetic thin films of nanometers in thickness were deposited and patterned after the conventional CMOS process. The biotinylated analyte DNA was captured by complementary probes immobilized on the sensor surface. Then streptavidin-coated magnetic labels were added and produced specific binding to the hybridized DNA. The stray magnetic field of the magnetic labels was detected as a resistance change in the sensor. In general, signal modulation is required in the sensing scheme in order to separate the true bio-signals from the false ones caused by drifts or ionic solution interference. Other than the GMR principle, detection of magnetic beads can be achieved based on the Hall effect of a CMOS sensor. As discovered by Edwin Hall in 1879, the effect produces an electric field perpendicular to the magnetic induction vector and the original current direction. Detection of a single magnetic bead has been demonstrated [14].

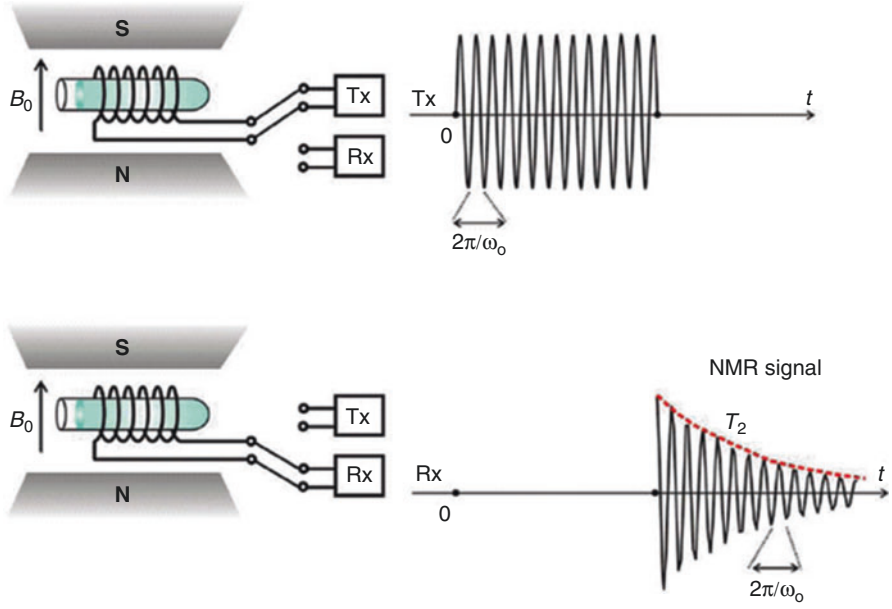
To eliminate the needs of externally applied magnetic fields and post-CMOS fabrication for a magnetoresistive biosensor, Wang et al. [15] presented an inductive approach that can detect existence of a single magnetic bead on a CMOS chip. The sensing scheme used a highly stable integrated oscillator with an on-chip LC resonator. An a.c. electrical current through the on-chip inductor produced a magnetic field that polarized the magnetic particles present in its vicinity, leading to an increased effective inductance and therefore a reduced oscillation frequency. As the frequency shift due to a single micron-size magnetic bead is typically a few parts per million (ppm) of the resonant frequency, the sensing oscillator needs to have small phase noises at small offset frequencies to achieve a stable frequency behavior.

A miniaturized CMOS NMR (nuclear magnetic resonance) system has been reported by Sun et al. [16] for applications in biomolecular sensing. NMR was first discovered by Isidor Rabi who was awarded Nobel Prize in Physics in 1944. Magnetic nuclei, like ^{1}H and ^{31}P, could absorb radio frequency (RF) energy when placed in a magnetic field of a strength specific to the identity of the nuclei. The nucleus is described as being in resonance when the absorption occurs. Different atomic nuclei within a molecule exhibit different resonant frequencies for the same magnetic field strength. Essential chemical and structural information about a molecule can thus be studied by observing such magnetic resonant frequencies.

The reported CMOS NMR system consists of a magnet (static magnetic field), an RF coil surrounding a sample, and an RF transceiver linked to the coil as shown in Fig. 4. The RF magnetic field produced through the coil at the right frequency ω_o can excite nuclei spins within the sample. Once the RF excitation is stopped and the receiver is connected to the coil, the detected NMR signal displays an exponential relaxation in the precession of the net magnetic moment. Both the resonant frequency and the relaxation's characteristic time are specific for the sample to be studied.

Optical Biosensors

Optical detection methods such as luminometry and fluorometry can be utilized to make CMOS-based biosensors. Luminometric methods, such as luciferase-based assays, involve the detection and quantification of light emission as a result of a chemical reaction. Luciferase is a generic term for the class of oxidative enzymes used in bioluminescence. Such methods have been used to detect pathogens and proteins, perform gene expression and regulation studies, and sequence DNA. Fluorometric methods require an excitation source to stimulate photoemission of the fluorescent-tagged species and optical filters to separate the generated photoemission from the



CMOS MEMS Biosensors, Fig. 4 Operation of the CMOS NMR system (Sun et al. © 2009 IEEE)

high background interference. Luminometry is more amenable to miniaturization and integration than the fluorometric methods since no filters or excitation sources are required.

Eltoukhy et al. [17] reported a CMOS photodetector array which was directly integrated with a fiber-optic faceplate with immobilized luminescent reporters/probes for DNA synthesis by pyrosequencing. Pyrosequencing is a "sequencing by synthesis" technique which involves taking a single strand of the DNA to be sequenced and then synthesizing its complementary strand enzymatically. Synthesis of the complementary strand is achieved with one base pair at a time by monitoring the activity of the DNA synthesizing enzyme with another chemiluminescent enzyme. The challenge of such a detection system is to achieve high sensitivity despite the presence of relatively high dark currents. The reported sensor array was able to detect emission rates below 10^{-6} lux over a long integration time (30 s).

Huang et al. [18] reported a CMOS microarray chip that leveraged low-cost integration of solid-state circuits for fluorescence-based diagnostics. The featured time-gated fluorescence detection was able to significantly reduce interferences from the external excitation source, eliminating the need for additional optical filters. Direct immobilization of DNA capture strands on the CMOS chip allows placement of optical detectors in close vicinity to the fluorescent labels, leading to improved collection efficiency and providing imaging resolution limited by pixel dimensions rather than diffraction optics.

In addition to the aforementioned approaches, a method based on detection of the incident light intensity using a normal light source has been developed [19]. The technique uses gold nanoparticles with silver enhancement to induce opacity on top of CMOS photodetectors when specific bindings occur. It does not require an external optical scanner and a specific light source as needed in the fluorescence-based method.

Micromechanical Cantilevers

Microcantilever-based biosensors have attracted considerable attention as a means of label-free detection of biomolecules. Intermolecular forces arising from adsorption of small molecules are known to induce surface stress. The specific binding between ligands and receptors on the surface of a microcantilever beam thus produces physical

bending of the beam. Optical detection of the beam deflection due to hybridization of complementary oligonucleotides has been demonstrated [20]; however, the method requires external instruments that are not amenable to monolithic integration.

Piezoresistive detection is a viable alternative for CMOS integration because it is compatible with aqueous media. The piezoresistive effect is associated with the resistivity change of a semiconductor subject to a mechanical strain. By placing a MOS transistor into the base of a cantilever, modulation of the channel current underneath the gate region can be measured as a result of adsorption-induced surface stress [21]. Sensing resolution of the static beam deflection is limited by the flicker noise – the dominant source of noise in MOS transistors at low frequencies.

Summary

A miniaturized biosensing platform can be achieved through monolithic integration of sensing devices and detection circuits by the CMOS MEMS technology. Sensing devices that can be directly fabricated on a CMOS chip are introduced, including those based on electrochemical, impedimetric, ion-sensitive, magnetic, optical, and micromechanical approaches. Some of the methods require labeling (e.g., magnetic beads, nanoparticles) and some are label-free for bio-signal transduction. Some approaches use the devices (e.g., transistors) or the materials (e.g., metal electrodes) in a CMOS process for sensing, such that sensor performances can be enhanced in accordance with the scaling of CMOS technologies. Arrays of sensors can be conveniently fabricated on a CMOS chip such that sensing resolution and accuracy can be enhanced through statistical analysis of the collected data.

Cross-References

- ▶ Biosensors
- ▶ Nanogap Biosensors

References

1. Aoki, K., Morita, M., Niwa, O., Tabei, H.: Quantitative analysis of reversible diffusion-controlled currents of redox soluable species at interdigitated array electrodes under steady-state conditions. J. Electroanal. Chem. **256**, 269–282 (1988)
2. Schienle, M., Paulus, C., Frey, A., Hofmann, F., Holzapfl, B., Schindler-Bauer, P., Thewes, R.: A fully electronic DNA sensor with 128 positions and in-pixel A/D conversion. IEEE J. Solid State Circuits **39**, 2438–2445 (2004)
3. Levine, P.M., Gong, P., Levicky, R., Shepard, K.L.: Active CMOS sensor array for electrochemical biomolecular detection. IEEE J. Solid State Circuits **43**, 1859–1871 (2008)
4. Ayers, S., Berberian, K., Gillis, K.D., Lindau, M., Minch, B.A.: Post-CMOS fabrication of working electrodes for on-chip recordings of transmitter release. IEEE Trans. Biomed. Circuits Syst. **4**, 86–92 (2010)
5. Warburg, E.: Ueber das Verhalten sogenannter unpolarisbarer Elektroden gegen Wechselstrom. Ann. Phys. Chem. **67**, 493–499 (1899)
6. Lee, K., Lee, J., Sohn, M., Lee, B., Choi, S., Kim, S. K., Yoon, J., Cho, G.: One-chip electronic detection of DNA hybridization using precision impedance-based CMOS array sensor. Biosens. Bioelectron. **26**, 1373–1379 (2010)
7. Stagni, C., Guiducci, C., Benini, L., Riccò, B., Carrara, S., Samorí, B., Paulus, C., Schienle, M., Augustyniak, M., Thewes, R.: CMOS DNA sensor array with integrated A/D conversion based on label-free capacitance measurement. IEEE J. Solid State Circuits **41**, 2956–2963 (2006)
8. Lu, M.S.-C., Chen, Y.C., Huang, P.C.: 5 × 5 CMOS capacitive sensor array for detection of the neurotransmitter dopamine. Biosens. Bioelectron. **26**, 1093–1097 (2010)
9. Ghafar-Zadeh, E., Sawan, M., Chodavarapu, V.P., Hosseini-Nia, T.: Bacteria growth monitoring through a differential CMOS capacitive sensor. IEEE Trans. Biomed. Circuits Syst. **4**, 232–238 (2010)
10. Li, J., Xue, M., Lu, Z., Zhang, Z., Feng, C., Chan, M.: A high-density conduction-based micro-DNA identification array fabricated with a CMOS compatible process. IEEE Trans. Electron Devices **50**, 2165–2170 (2003)
11. Kim, D.S., Jeong, Y.T., Park, H.J., Shin, J.K., Choi, P., Lee, J.H., Lim, G.: An FET-type charge sensor for highly sensitive detection of DNA sequence. Biosens. Bioelectron. **20**, 69–74 (2004)
12. Li, D.C., Yang, P.H., Lu, M.S.-C.: CMOS open-gate ion-sensitive field-effect transistors for ultrasensitive dopamine detection. IEEE Trans. Electron Devices **57**, 2761–2767 (2010)
13. Han, S., Yu, H., Murmann, B., Pourmand, N., Wang, S.X.: A high-density magnetoresistive biosensor array with drift-compensation mechanism. In: IEEE International Solid-State Circuits Conference (ISSCC)

Digest of Technical Papers, pp. 168–169, San Francisco, 11–15 Feb 2007
14. Besse, P., Boero, G., Demierre, M., Pott, V., Popovic, R.: Detection of a single magnetic microbead using a miniaturized silicon Hall sensor. Appl. Phys. Lett. **80**, 4199–4201 (2002)
15. Wang, H., Chen, Y., Hassibi, A., Scherer, A., Hajimiri, A.: A frequency-shift CMOS magnetic biosensor array with single-bead sensitivity and no external magnet. In: IEEE International Solid-State Circuits Conference (ISSCC) Digest of Technical Papers, pp. 438–439 (2009)
16. Sun, N., Liu, Y., Lee, H., Weissleder, R., Ham, D.: CMOS RF biosensor utilizing nuclear magnetic resonance. IEEE J. Solid State Circuits **44**, 1629–1643 (2009)
17. Eltoukhy, H., Salama, K., El Gamal, A.: A 0.18-μm CMOS bioluminescence detection lab-on-chip. IEEE J. Solid State Circuits **41**, 651–662 (2006)
18. Huang, T.D., Sorgenfrei, S., Gong, P., Levicky, R., Shepard, K.L.: A 0.18-μm CMOS array sensor for integrated time-resolved fluorescence detection. IEEE J. Solid State Circuits **44**, 1644–1654 (2009)
19. Xu, C., Li, J., Wang, Y., Cheng, L., Lu, Z., Chan, M.: A CMOS-compatible DNA microarray using optical detection together with a highly sensitive nanometallic particle protocol. IEEE Electron Device Lett. **26**, 240–242 (2005)
20. Fritz, J., Baller, M.K., Lang, H.P., Rothuizen, H., Vettiger, P., Meyer, E., Güntherodt, H.-J., Gerber, C., Gimzewski, J.K.: Translating biomolecular recognition into nanomechanics. Science **288**, 316–318 (2000)
21. Shekhawat, G., Tark, S.H., Dravid, V.P.: MOSFET-embedded microcantilevers for measuring deflection in biomolecular sensors. Science **311**, 1592–1595 (2006)

CMOS MEMS Fabrication Technologies

Hongwei Qu[1] and Huikai Xie[2]
[1]Department of Electrical and Computer Engineering, Oakland University, Rochester, MI, USA
[2]Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

Synonyms

CMOS (complementary metal-oxide-semiconductor); CMOS-MEMS; Integration; MEMS (micro-electro-mechanical systems)

Definition

CMOS-MEMS are micromachined systems in which MEMS devices are integrated with CMOS circuitry on a single chip to enable miniaturization and performance improvement. CMOS-MEMS also refers to microfabrication technologies that are compatible with CMOS fabrication processes.

Overview

Microelectromechanical systems (MEMS) leverage semiconductor fabrication technologies to manufacture various miniature sensors and actuators. Due to their low cost and small size as well as their much improved reliability, MEMS devices have been widely used even in our daily life, e.g., MEMS accelerometers for automobiles' airbags, MEMS gyroscopes for electronic stability program (ESP) in automobile braking systems, MEMS tire pressure sensors; digital micromirror device (DMD)-enabled portable projectors, MEMS inkjet printers, MEMS resonators as frequency references, etc. Moreover, smart cell phones are now equipped with MEMS gyroscopes and accelerometers for motion actuated functions. They are also installed with surface-mounted MEMS microphones for even smaller size. The worldwide MEMS market reached 6.5 billion US dollars in 2010 [1].

Continuous miniaturization, expanded functionalities, lower cost, and improved performance are the ultimate goals of MEMS. The nature of MEMS strongly suggests direct integration of mechanical structures with electronics whose fabrication is dominated by CMOS technologies. In the last couple of decades, great efforts have been made in the integration of MEMS structures with ICs on a single CMOS substrate. The pioneering work for CMOS-MEMS transducers was done by H. Baltes and his coworkers at the Swiss Federal Institute of Technology Zurich (ETH) [2]. They employed both wet bulk silicon micromachining and surface micromachining techniques in the fabrication of integrated CMOS-MEMS devices. With the great advances in IC and MEMS

technologies, the current focus of CMOS-MEMS integration technology is on the modification and standardization of CMOS technology to accommodate MEMS technology. One of the best-known commercial CMOS-MEMS devices is the digital micromirror device (DMD) manufactured by Texas Instruments. Recently, some CMOS foundries, such as TSMC, X-Fab and Global Foundries, have begun to offer CMOS-MEMS services for research and product developments.

This entry summarizes a variety of CMOS-MEMS technologies and devices that have been developed. Particular materials needed in associated CMOS-MEMS will also be introduced. Typical MEMS devices, including inertial sensors, resonators, and actuators, are exemplified in featuring the respective technologies.

Classification of CMOS-MEMS Technologies

MEMS can be integrated with CMOS electronics in many different ways. One common way to categorize CMOS-MEMS technologies is from the perspective of manufacturing processes. Based on the process sequence, CMOS-MEMS technologies can be classified into three categories: Pre-CMOS, Intra-CMOS, and Post-CMOS [3]. Due to its popularity and accessibility, post-CMOS will be described in more detail in this entry.

Pre-CMOS

It is widely accepted that pre-CMOS technologies are represented by the modular integration process originally developed at Sandia National Laboratories. As suggested by the name, in pre-CMOS technology, MEMS structures are pre-defined and embedded in a recess in silicon wafer; and the recess is then filled with oxide or other dielectrics. The wafer is then planarized prior to the following process steps for CMOS electronics [4]. In this "MEMS first" process, although MEMS structures are pre-defined, a wet etch after the completion of the standard CMOS processes is required to release the pre-defined MEMS structures. Due to the involvement of photolithography process needed for patterning the MEMS in the recess, the thickness of the MEMS structures is constrained by the lithographical limit.

Other methods for the formation of MEMS structures in pre-CMOS technologies, including wafer bonding and thinning for epitaxial and SOI wafers in which MEMS are prefabricated, have also been reported in the fabrication of a variety of MEMS devices.

Inter-CMOS

In early 1990s, Analog Devices, Inc (ADI) specifically developed a MEMS technology based on its BiCMOS process. This "iMEMS" technology, originally dedicated to manufacturing CMOS-MEMS accelerometers and gyroscopes, is an intermediate-CMOS-MEMS, or Inter-CMOS-MEMS technology in which the CMOS process steps are mixed with additional polysilicon thin-film deposition and micromachining steps to form the sensor structures [5]. Infinion's pressure sensors are also fabricated using this kind of Inter-CMOS-MEMS technology. To reduce the residual stress in structural polysilicon, high temperature annealing is normally required in the Inter-CMOS-MEMS, which could pose a potential risk to CMOS interconnect and active layers. Thus, the thermal budget should be carefully designed. Moreover, it is almost impractical to perform intermediate CMOS and MEMS processed in separate foundries due to possible contaminations in the wafer transfer and processes. Therefore, a dedicated foundry used for both CMOS and MEMS is necessary for Inter-CMOS-MEMS technology, which may not take full advantages of mainstream technologies in either area.

Limitations of Pre- and Inter-CMOS-MEMS

Since surface micromachining and polysilicon are typically used, most of the Pre- and

Inter-CMOS-MEMS technologies suffer from the limitations of thin-film structural materials. (1) Structural curling and cost associated with stress compensation: Due to the residual stress in the deposited thin films in the device, polysilicon structures often exhibit curling after release, resulting in reduced sensitivity, lower mechanical robustness, and increased temperature dependence. Although stress compensation can be realized via multiple controlled process steps, the associated cost is quite high. (2) Small size and/or mass: The curling of thin-film structures in turn limits the size of the overall microstructure. The mass is further reduced due to the small thickness of thin-film polysilicon. For inertial sensors, the smaller the structure mass, the lower performance of the device. (3) Parasitics: In a surface micromachined polysilicon accelerometer, depending on polysilicon wiring path, the parasitic impedance may considerably lower the static and dynamic performance of the device. (4) Cost and suboptimal processes of the dedicated foundry needed: The dedicated foundry combining CMOS and MEMS fabrications needed for pre- and inter-CMOS-MEMS is normally expensive and suboptimal for either fabrication. It is against the modern trends in which flexible accessibility of optimal and cost-effective processes are preferable.

Post-CMOS

Post-CMOS-MEMS refer to the CMOS-MEMS processes in which all MEMS process steps are performed after the completion of the CMOS fabrication. The advantages of post-CMOS-MEMS over Pre- and Inter-CMOS-MEMS include process flexibility and accessibility and low cost. In contrast to both Pre-CMOS-MEMS and Inter-CMOS-MEMS, for Post-CMOS-MEMS technology, the fabrications of CMOS circuitry and MEMS structures are performed independently. The flexibility of foundry access makes it possible to take advantages of both advanced CMOS technologies and optimal MEMS fabrication. This is particularly attractive to research community in exploration of state-of-the-art in MEMS. Some design rules may need to be changed to accommodate MEMS structure design in the CMOS design stage. Meanwhile, post-CMOS microfabrication should be carefully designed, particularly considering the thermal budget, so as not to affect the on-chip CMOS electronics.

According to how MEMS structures are formed, post-CMOS-MEMS technologies fall into two categories: additive and subtractive. In additive post-CMOS-MEMS, structural materials are deposited on a CMOS substrate. In subtractive post-CMOS-MEMS, MEMS structures are created by selectively etching CMOS layers. Apparently, additive post-CMOS-MEMS methods require more stringent material compatibility with the CMOS technologies used. Thus, they are less utilized than subtractive post-CMOS-MEMS. The following introduction will focus more on subtractive post-CMOS-MEMS.

Additive MEMS Structures on CMOS Substrate

In additive post-CMOS-MEMS, metals, dielectrics, or polymers are deposited and patterned to form MEMS structures normally on top of the CMOS layers. Some commercial MEMS products are fabricated using additive post-CMOS-MEMS approaches. In this category, the best-known product is probably the digital mirror device (DMD), the core of the digital light processing (DLP) technology developed by Texas Instruments. In a DMD, tilting mirror plates and their driving electrodes are fabricated directly on top of CMOS circuits. Three sputtered aluminum layers are used to form the top mirror plate and the two parallel-plate electrodes for electrostatic actuation, respectively. The driving electrodes are addressed via CMOS memory cell. To release the mirror plate and top electrodes in the post-CMOS-MEMS fabrication of the mirrors, deep-UV hardened photoresist is used as the sacrificial layer.

In some circumstances where CMOS protection is well designed, electroplating can also be used to grow microstructures on top of CMOS electronics. Other structural and sacrificial materials, such as polycrystalline SiGe and Ge, have

CMOS MEMS Fabrication Technologies, Table 1 Representative thin-film deposition additive post-CMOS-MEMS technologies

Authors and references	Institute	Structural material	Sacrificial material	Interconnect material	Year
Hornbeck [7]	Texas instruments	Al	Photoresist	Al	1989 (invented in 1987)
Yun et al. [8]	UC-Berkeley	Polysilicon	SiO_2	W/TiN	1992
Franke et al. [9]	UC-Berkeley	Poly-SiGe	Ge or SiO_2	Al	1999
Sedky, Van Hoof et al. [10]	IMEC	Poly-SiGe	Ge	Al	1998

been used to create CMOS-MEMS as well [6]. Additive post-CMOS-MEMS processes, along with their respective materials, are summarized in the following table (Table 1).

In addition to the approaches of forming MEMS structures on top of the CMOS substrate by thin-film deposition, wafer bonding provides another method to directly integrate MEMS structures on CMOS substrate [11]. For instance, a prefabricated polysilicon capacitive acceleration sensor wafer is bonded to a CMOS wafer with read-out electronics. In a wafer-bonded piezoresistive accelerometer, the micromachined bulk silicon proof mass was sandwiched by a bottom glass cap and a top CMOS chip on which the conditioning circuit was integrated. SOI-CMOS-MEMS has also been attempted for monolithic integration of electronics with bulk MEMS structures. With 3-dimensional packaging enabled by technological breakthroughs such as through-silicon vias (TSVs), this integration method promises to be further developed in manufacturing complex microsystems. MEMS suppliers, including STMicroelectronics and InvenSense, have adopted wafer-to-wafer or chip-to-wafer bonding CMOS-MEMS integration.

Subtractive Post-CMOS-MEMS

In these devices, MEMS structures are formed from built-in CMOS thin-film stacks including metals and SiO_2, or from the silicon substrate. These materials are patterned and removed partially by wet or dry etching to release the MEMS structures. This section describes the thin-film and bulk CMOS-MEMS formed by such subtractive processes.

Subtractive CMOS-MEMS by Wet Etching

The first generation of CMOS-MEMS sensors was fabricated using a post-CMOS subtractive process in which silicon substrate was completely or partially removed using a wet etching method, leaving behind thin-film or bulk MEMS structures [2]. For thermal sensors in which beams or membranes consisting of dielectric layers, the substrate silicon is normally etched away completely to obtain thermally isolated structures. The silicon dioxide membrane can act as an intrinsic etch stop layer in backside silicon anisotropic wet etch using KOH, ethylene diamine-pyrocatechol (EDP), or Tetramethylammonium hydroxide (TMAH). A high-Q RF MEMS filter with an inter-metal dielectric layer as structural material was reported by IBM. A medical tactile sensor array was also reported in which the aluminum sacrificial layer was etched from the backside of the wafer after the CMOS substrate was etched through [12].

The silicon substrate can also be included in the MEMS structures using a wet etch process. The first method is to perform a time-controlled backside etch with a well-calibrated etching rate. A uniform single crystal silicon membrane with a desired thickness can be created. This method has been widely used in industry for fabrication of large volume products such as integrated pressure sensors. In cases where the silicon membrane thickness is not critical, even mechanical processing such as grinding can be used to create the backside cavity.

The second method involves the utilization of an automatic etch stop technique to create silicon membranes or MEMS structures. In this case, an anisotropic etch stops at the electrochemically

CMOS MEMS Fabrication Technologies, Table 2 CMOS-MEMS devices enabled by subtractive process wet etching

Authors and references	Institutions	Year	Device	Device structure	Etching method
Wise et al. [16]	U. of Michigan	1979	Pressure sensor	Silicon diaphragm	Backside EDP etching
Wise et al. [14]	U. of Michigan	1985	Neuron probe array	CMOS thin films and Si substrate	EDP etching, p++ etching stop
Yoon and Wise [17]	U. of Michigan	1990	Mass flow sensor	CMOS thin films	Backside, SiO_2 etching stop
Baltes et al. [2]	ETH Zurich	1996	Thermal capacitor	CMOS thin films	Front side etching
Haberli et al. [18]	ETH Zurich	1996	Pressure sensor	CMOS thin films	Front side etching of aluminum as sacrificial layer
Schneider et al. [19]	ETH Zurich	1997	Thermal sensor	CMOS thin films and suspended substrate	PN junction electrochemical etch stop
Akiyama et al. [20]	U. of Neuchatel, ETH Zurich	2000	AFM probe	CMOS thin films and silicon substrate	N-well electrochemical etch stop
Schaufelbuhl et al. [21]	ETH Zurich	2001	Infrared imager	CMOS thin films	Backside KOH
Verd et al. [22]	U. of Barcelona	2006	RF MEMS	CMOS thin films	Front side SiO_2 etching

biased p-n junction formed between the n-well and p-type substrate in CMOS [13]. Although the electrochemical electrode design and implementation are complicated, this process can be specifically used in the fabrication of highly sensitive pressure/force and thermal sensors. The anisotropic etch stop can also occur at highly doped ***p*** regions in the substrate. This method has been used in fabrication of many suspended structures including neural probes [14]. Note that the p++ doping process may not be available in a standard CMOS process. In the case where only a small portion of the silicon substrate needs to be removed to reduce the circuit-substrate coupling, a wet silicon etch can be performed from the front side. In wet silicon etching, either silicon nitride or additional polymers or both can be used to protect the front CMOS and pads.

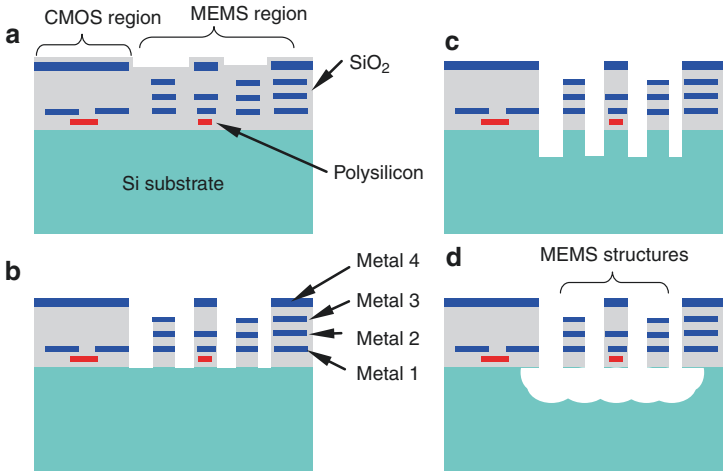
Polymers sensitive to analytes can be coated on finished CMOS-MEMS structures for chemical and biological sensing. For example, the first CMOS-MEMS electronic nose was demonstrated by forming polymer-coated CMOS thin-film cantilevers on a CMOS chip [15].

Table 2 summarizes some representative devices that were fabricated using wet etching when this technology was dominant in post-CMOS micromachining. Bibliographies of these efforts can be found in the above citations in this section.

Subtractive Post-CMOS-MEMS by Dry Etching

Dry etching processes have quickly become popular in microfabrication for both MEMS research and industry. Particularly, the deep reactive ion etching (DRIE) technology, or Bosch process, has revolutionized subtractive post-CMOS microfabrication [23]. This section describes thin-film and bulk CMOS-MEMS devices fabricated using dry etching processes.

Most dry etching processes are based on plasma processes, such as reactive ion etch (RIE) and DRIE, while etchants in the vapor phase can also be used for dry etching. For example, vapor XeF_2 provides good isotropic etching of silicon, which has been used for releasing CMOS thin-film MEMS structures [24]. The combination of RIE and DRIE, performed from the front or back side, or both sides, has been used to fabricate a large variety of CMOS-MEMS devices. Depending on the structural materials and etching methods employed, subtractive post-CMOS can be divided into two types: thin-film processes and bulk processes.

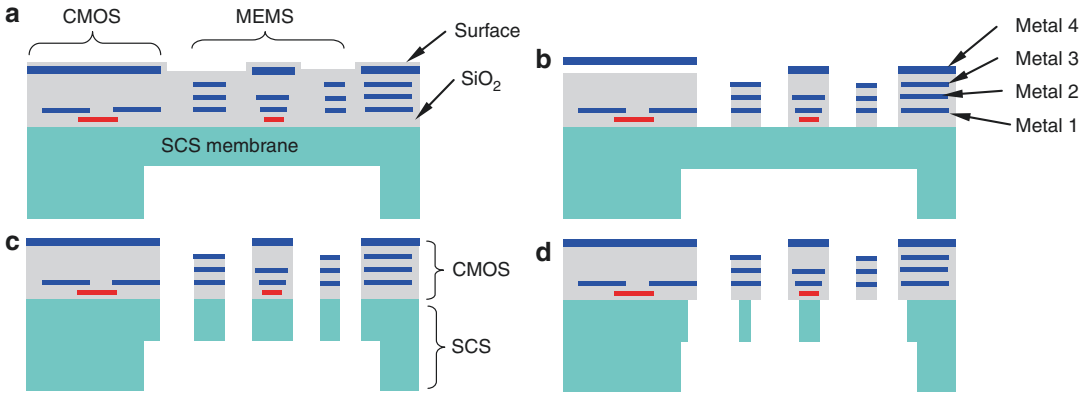


CMOS MEMS Fabrication Technologies,
Fig. 1 CMU post-CMOS fabrication process for MEMS structures made of CMOS thin films

Thin-Film Post-CMOS-MEMS Dry Processes In thin-film processes, structural materials are composed of CMOS thin films. Figure 1 depicts the process flow of a thin-film post-CMOS-MEMS process, which was originally developed at Carnegie Mellon University [25]. A sequenced process consisting of an isotropic SiO_2 etching, a silicon DRIE and an isotropic Si RIE releases the MEMS structure. In these process steps, the top metal layer acts as a mask to form the MEMS structures and to protect the CMOS circuitry, as seen in Fig. 1a, b. Anisotropic and isotropic silicon etching complete the process flow, as seen in Fig. 1c, d. Various inertial sensors have been fabricated using this thin-film technology. In all these inertial sensors, mechanical springs and proof masses are formed by the multiple-layer CMOS stacks consisting of SiO_2 and metals. The sensing capacitance is formed from sidewall capacitance between comb fingers. The multiple CMOS metal layers inside the comb fingers and other mechanical structures allow very flexible electrical wiring, facilitating different sensing schemes including vertical comb-drive sensing. Akustica, Inc. has commercialized digital microphones using a modified version of this process. Other sensors have also been demonstrated using similar thin-film technology, such as humidity sensors and chemical sensors.

All these thin-film post-CMOS dry etching processes have excellent CMOS compatibility and accessibility as well as design flexibility. However, a major issue is the large vertical curling and lateral buckling of suspended MEMS structures, which is caused by the residual stress in the stacked thin-film CMOS layers. Although structural curling can be tolerated for some small devices such as RF MEMS, for devices such as inertial sensors that need relatively large size, the impact of structural curling can be severe.

Bulk CMOS-MEMS Dry Process In order to overcome the structural curling and to increase the mass, flatness, and robustness of MEMS structures, single crystal silicon (SCS) may be included underneath the CMOS thin-film stacks. The SCS silicon structures are formed directly from the silicon substrate using DRIE. Figure 2 illustrates the process flow [26]. The process starts with the backside silicon DRIE to define the MEMS structure thickness by leaving a 10–100 μm-thick SCS membrane (Fig. 2a). Next, the same anisotropic SiO_2 etch as in the thin-film process is performed on the front side of wafer (chip) to expose the SCS to be removed (Fig. 2b). The following step differs from the thin-film process in that an anisotropic DRIE, instead of isotropic etch, finalizes the structure release by etching through the remaining SCS diaphragm, as shown in Fig. 2c. With the SCS underneath the CMOS interconnect layers included, large and flat MEMS microstructures can be obtained. If necessary, an optional time-controlled isotropic silicon etch can be added. This step will undercut the SCS



CMOS MEMS Fabrication Technologies, Fig. 2 DRIE bulk CMOS-MEMS process flow

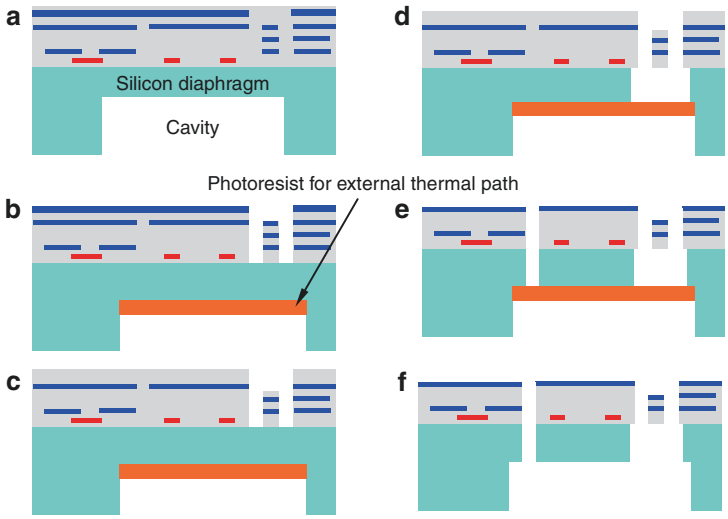
underneath the designed narrow CMOS stacks to create thin-film structures (Fig. 2d). This step is particularly useful for fabricating capacitive inertial sensors. It can be used to form the electrical isolation structures between sensing fingers and silicon substrate.

The DRIE CMOS-MEMS technology has shown great advantages in the fabrication of relatively large MEMS devices such as micromirrors. Large flat mirror can be obtained by including portion of silicon substrate underneath the aluminum mirror surface. A CMOS-MEMS gyroscope with a low noise floor of 0.02°/s/sqrtHZ has also been demonstrated using this technology [27].

By attaching SCS underneath the CMOS stack comb fingers, the sensing capacitance can be considerably increased for larger signal-to-noise ratio (SNR). Although CMOS thin films are still used in some microstructures for electrical isolation, the length of the thin-film portion is minimal to reduce the temperature effect. Compared to the thin-film dry CMOS-MEMS process, a backside silicon DRIE step is added. This requires an additional backside lithography step to define the region for MEMS structures. The maximum thickness of the MEMS structures is limited by the aspect ratio that the silicon DRIE can achieve.

An Improved Bulk CMOS-MEMS Process The bulk CMOS-MEMS process depicted in Fig. 2 is useful in fabrication of many devices where SCS structures are desired to improve both mechanical and electrical performance of the devices. However, for some devices, very fine structures are formed in step (c) in Fig. 2; so the damage caused by the step (d) to these fine structures may be severe. This is particularly true for the fabrication of capacitive inertial sensors where narrow-gap sensing comb fingers are needed. For instance, in performing the isotropic silicon undercut to form the narrow CMOS beams for electrical isolation, the SCS in the comb fingers is also undercut. The sensing gap increases due to the undesired undercut greatly reduce sensitivity and signal-to-noise ratio (SNR). If the undercut occurs in mechanical structures such as suspension springs, the characteristics of the device will also be affected. Another issue is related to the thermal effect in the plasma etch for the SCS undercut. Upon completion of the silicon undercut, the greatly reduced thermal conductance from the isolated structure to the substrate can cause a temperature rise on the released structures. Slight over-etch is often necessary to accommodate process variations, but this will generate a large temperature rise on the suspended structures which in turn dramatically increases the SCS etching rate, resulting in uncontrollable and damaging results [28].

A modified dry bulk CMOS-MEMS process has been demonstrated to effectively address the issues caused by the undesired SCS



CMOS MEMS Fabrication Technologies, Fig. 3 The modified bulk CMOS-MEMS process for separate etching of CMOS beams and SCS microstructures. Backside photoresist coating effectively reduces temperature in the device release, reducing deleterious non-uniform etching

undercut [28]. In the refined process illustrated in Fig. 3, the etching of the CMOS connection beams is performed separately from the etching of the microstructures where SCS is needed. The top metal layer is specifically used to define the connection beams. After their formation, the top metal layer is removed using a plasma or a wet etch. Then other microstructures are exposed after a SiO_2 etch. The direct etch-through of the remaining silicon on the microstructures will complete the release process. To reduce the thermal effect described above, a thick photoresist layer is patterned on the backside of the cavity. In the release step, the applied photoresist provides a thermal path that reduces the temperature rise on the etched-through structures. The removal of the photoresist using O_2 plasma etching completes the entire microfabrication process. Owing to the monolithic integration and large proof mass enabled by the inclusion of SCS, bulk CMOS-MEMS inertial sensors have demonstrated better performance than their thin-film counterparts [29]. The photoresist coating can also be replaced by sputtering a layer of metal such as aluminum.

Combined Wet/Dry Processes

In addition to the integration methods described above, efforts have been continuously made to integrate CMOS with MEMS using the combination of different microfabrication technologies. By combing silicon anisotropic wet etch with DRIE, some sophisticated surface and bulk MEMS structures such as bridges and cantilever arrays can be created. A multi-sensor system was demonstrated using a combined etch process [30]. In the accelerometers reported in [31], isotropic wet etching is used to remove metal layers in CMOS thin stacks to create parallel-plate-like vertical capacitors for gap-closing sensing. A silicon RIE follows to release the MEMS devices and break the coupling between the sensing thin films and the substrate. Sensitivities are largely increased with the gap-closing sensing compared to comb-finger sensing.

Summary

CMOS-MEMS technologies have been placed in pre-CMOS, intra-CMOS, and post-CMOS categories. Both pre-CMOS and intra-CMOS have issues such as dedicated foundries with suboptimal and less cost-effective processes. So it is normally impractical for academic research community to access these dedicated facilities. Post-CMOS provides excellent CMOS compatibility, foundry accessibility, and design flexibility, and the cost is also relatively low. While the

process standardization and industrialization of CMOS-MEMS technologies are in continuous progress, innovative processing technologies have opened up new pathways for integration. Wafer bonding–based integration has blurred the boundary between pre- and post-CMOS-MEMS integrations. SOI-CMOS-MEMS have also been aggressively explored. The technologies involved in this new exploration have emerged as enabling means for three-dimensional and systems-in-package integrations. More recently, technologies to co-fabricate many subsystems including nano-systems are being pursued enthusiastically.

CMOS-MEMS integration will continually evolve with the emergence of new fabrication technologies and new materials. While some companies have demonstrated promising CMOS-MEMS products, more joint efforts from research community and industries are needed for new process transfer and standardization to allow large volume fabrication of new products.

Cross-References

- ▶ Integration
- ▶ MEMS
- ▶ Nanofabrication
- ▶ Sensors

References

1. Johnson, R.C.: MEMS market projected to hit double-digit growth, again. www.eetimes.com. Accessed 14 July 2011
2. Baltes, H., Paul, O., et al.: IC MEMS microtransducers. In: Proceedings of International Electronic Device Meeting IEDM '96, San Francisco, pp. 521–524 (1996)
3. Baltes, H., Brand, O., Fedder, G.K., Hierold, C., Korvink, J.G., Tabata, O.: CMOS-MEMS: Advanced Micro and Nanosystems, 1st edn. Wiley, Weinheim (2005)
4. Smith, J.H., Montague, S., Sniegowski, J.J., Murray, J. R., McWhorter, P.J., Smith J.H.: Embedded micromechanical devices for the monolithic integration of MEMS with CMOS. In: Proceedings of International Electronic Device Meeting, IEDM '95, Washington, DC, pp. 609–612 (1995)
5. Kuehnel, W., Sherman, S.: A surface micromachined silicon accelerometer with on-chip detection circuitry. Sens. Actu. A Phys. **45**, 7–16 (1994)
6. Franke, A.E., Heck, J.M., King, T.J., Howe, R.T.: Polycrystalline silicon-germanium films for integrated microsystems. J. Micromech. Sys. **12**, 160–171 (2003)
7. Hornbeck, L.: Deformable-mirror spatial light modulators and applications. SPIE Crit. Rev. **1150**, 86–102 (1989)
8. Yun, W., Howe, R.T., Gray, P.R.: Surface micromachined, digitally force-balanced accelerometer with integrated CMOS detection circuitry. Technical Digest of Solid State Sensors and Actuators Workshop, Hilton Head Island, pp. 126–131 (1992)
9. Franke, A.E., Bilic, D., Chang, D.T., Jones, P.T., King, T.J., Howe, R.T., Johnson, G.C.: Post-CMOS integration of germanium microstructures. In: The 12th IEEE International Conference on Micro Electro Mechanical Systems, Orlando, pp. 630–637 (1999)
10. Sedky, S., Fiorini, P., Caymax, M., Loreti, S., Baert, K., Hermans, L., Mertens, R.: Structural and mechanical properties of polycrystalline silicon germanium for micromachining applications. J. MEMS **7**, 365–372 (1998)
11. Fedder, G.K., Howe, R.T., Liu, T.J., Quevy, E.P.: Technologies for cofabricating MEMS and electronics. Proc. IEEE **96**, 306–322 (2008)
12. Salo, T., Vancura, T., Brand, O., Baltes, H.: CMOS-based sealed membranes for medical tactile sensor arrays. In: Proceedings of International Conference on Micro Electro Mechanical Systems, Kyoto, pp. 590–593 (2003)
13. Muller, T., Brandl, M., Brand, O., Baltes, H.T.: An industrial CMOS process family adapted for the fabrication of smart silicon sensors. Sen. Actu. A Phys. **84**, 126–133 (2000)
14. Najafi, K., Wise, K.D., Mochizuki, T.K.: A high-yield IC-compatible multichannel recording array. IEEE T. Electron. Dev. **32**, 1206–1211 (1985)
15. Baltes, H., Koll, A., Lange, D.H.: The CMOS MEMS nose-fact or fiction? In: Proceedings of IEEE International Symposium on Industrial Electronics ISIE '97, Guimaraes, vol. 1, pp. SS152–SS157 (1997)
16. Borky, J.M., Wise, K.D.: Integrated signal conditioning for silicon pressure sensors. IEEE T. Electron. Dev. **ED-27**, 927–930 (1979)
17. Yoon, E., Wise, K.D.: A multi-element monolithic mass flowmeter with on-chip CMOS readout electronics. Technical Digest of Solid State Sensors and Actuators Workshop, Hilton Head, pp. 161–164 (1990)
18. Haberli, A., Paul, O., Malcovati, P., Faccio, M., Maloberti, F., Baltes, H.: CMOS integration of a thermal pressure sensor system. In: IEEE International Symposium on Circuits and Systems, ISCAS '96, Atlanta, vol. 1, pp. 377–380 (1996)
19. Schneider, M., Muller, T., Haberli, A., Hornung, M., Baltes, H.: Integrated micromachined decoupled CMOS chip on chip. In: Proceedings of 10th IEEE

International Workshop on MEMS, Nagoya, pp. 512–517 (1997)
20. Akiyama, T., Akiyama, T., Staufer, U., de Rooij, N.F., Lange, D., Hagleitner, C., Brand, O., Baltes, H., Tonin, A., Hidber, H.R.: Integrated atomic force microscopy array probe with metal-oxide-semiconductor field effect transistor stress sensor, thermal bimorph actuator, and on-chip complementary metal-oxide-semiconductor electronics. J. Vac. Sci. Technol. B **18**, 2669–2675 (2000)
21. Schaufelbuhl, A., Schneeberger, N., Munch, U., Waelti, M., Paul, O., Brand, O., Baltes, H., Menolfi, C., Huang, Q., Doering, E., Loepfe, M.: Uncooled low-cost thermal imager based on micromachined CMOS integrated sensor array. J. MEMS **10**, 503–510 (2001)
22. Verd, J., Uranga, A., Teva, J., Lopez, J.L., Torres, F., Esteve, J., Abadal, G., Perez-Murano, F., Barniol, N.: Integrated CMOS-MEMS with on-chip readout electronics for high-frequency applications. IEEE Electron. Dev. Lett. **27**, 495–497 (2006)
23. Laermer, F., Schilp, A.: Method of anisotropically etching silicon. US Patent 5,501,893, Robert Bosch Gmbh (1992)
24. Kruglick, E.J.J., Warneke, B.A., Pister, K.S.: CMOS 3-axis accelerometers with integrated amplifier. In: Proceedings of International Conference on Micro Electro Mechanical System, MEMS-98, pp. 631–636. Heidelberg. (1998)
25. Fedder, G.K., Santhanam, S., Reed, M.L., Eagle, S.C., Guillou, D.F., Lu, M.S.C., Carley, L.R.: Laminated high-aspect-ratio microstructures in a conventional CMOS process. Proceedings of International Conference on Micro Electro Mechanical Systems, MEMS-96, San Diego, pp. 13–18 (1996)
26. Xie, H., Erdmann, L., Zhu, X., Gabriel, K.J., Fedder, G. K.: Post-CMOS processing for high-aspect-ratio integrated silicon microstructures. J. Microelectromech. Syst. **11**, 93–101 (2002)
27. Xie, H., Fedder, G.K.: Fabrication, characterization, and analysis of a DRIE CMOS-MEMS gyroscope. IEEE Sens. J. **3**, 622–631 (2003)
28. Qu, H., Xie, H.: Process development for CMOS-MEMS sensors with robust electrically isolated bulk silicon microstructures. J. Microelectromech. Syst. **16**, 1152–1161 (2007)
29. Qu, H., Fang, D., Xie, H.: A monolithic CMOS-MEMS 3-axis accelerometer with a low-noise, low-power dual-chopper amplifier. IEEE Sens. J. **8**, 1511–1518 (2008)
30. Hagleitner, C., Lange, D., Hierlemann, A., Brand, O., Baltes, H.: CMOS single-chip gas detection system comprising capacitive, calorimetric and mass-sensitive microsensors. IEEE J. Solid-St. Circc **37**, 1867–1878 (2002)
31. Tsai, M.H., Sun, C.M., Liu, Y.C., Wang, C.W., Fang, W.L.: Design and application of a metal wet-etching post-process for the improvement of CMOS-MEMS capacitive sensors. J. Micromech. Microeng. **19**, 105017 (2009)

CMOS-CNT Integration

Huikai Xie and Ying Zhou
Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, USA

C

Synonyms

Monolithic integration of carbon nanotubes

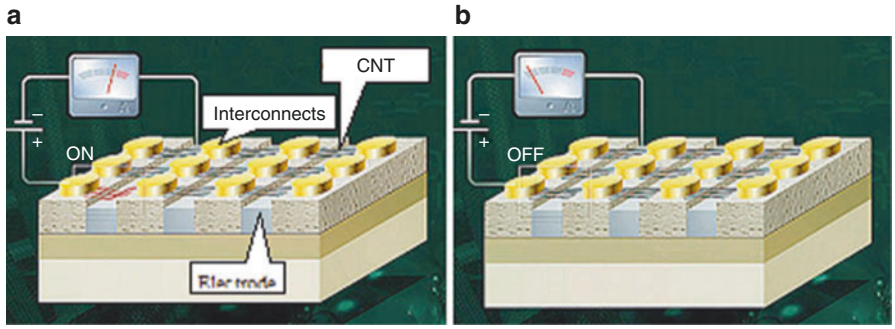
Definition

Carbon nanotubes can be directly grown on CMOS substrate without degrading the performance of CMOS electronics.

Introduction

With numerous outstanding electrical, mechanical, and chemical properties, carbon nanotubes (CNTs) have been explored for various applications with great success. As CMOS circuits possess powerful interfacing, signal amplification, conditioning, and processing capabilities, it is also highly desired to integrate CNTs with CMOS. CNTs may be either used as part of CMOS electronics or as sensing elements to form functioning nano-electromechanical systems (NEMS). Figure 1 shows a nanotube random-access memory (NRAM) which uses CNT ribbons as switches [1].

Many sensors based on CMOS-CNT hybrid systems have also been demonstrated, including mechanical, thermal, and chemical sensors [3, 4]. The integration of CMOS circuits with CNT sensors can increase signal-to-noise ratio and dynamic range, lower power consumption, and provide various controls and automations. Other efforts have been made to use multiwall carbon nanotubes (MWNTs) as CMOS interconnect for high frequency applications [5], or to apply CNT-based nano-electromechanical



CMOS-CNT Integration, Fig. 1 The structure of NRAM at (**a**) on and (**b**) off states [2]

switches for leakage reduction in CMOS logic and memory circuits [6].

However, monolithic integration of CMOS and CNTs is still very challenging. Most CMOS-CNT systems have been realized either by a two-chip solution or low-throughput CNT manipulations. In this entry, CMOS-CNT integration approaches are reviewed, with a particular focus on a localized heating CNT synthesis method that can grow CNTs on foundry CMOS.

CNT Synthesis

There are three main methods for carbon nanotube synthesis: arc-discharge [7], laser ablation [8], and chemical vapor deposition (CVD) [9]. The first two methods involve evaporation of solid-state carbon precursors and condensation of carbon atoms to form nanotubes, where high annealing temperature, typically over 1000 °C, is required to remove defects and thus produce high-quality nanotubes. However, they tend to produce a mixture of nanotubes and other by-products such as catalytic metals, so the nanotubes must be selectively separated from the by-products. This requires post-growth purification and manipulation.

In contrast, the CVD method employs a hydrocarbon gas as the carbon source and involves heating metal catalysts in a tube furnace to synthesize nanotubes. Nanotubes can grow either on the top (tip growth) or from the bottom (base growth). The diameters and locations of the grown CNTs can be controlled via catalyst size and catalyst patterning, and the orientation can be guided by an external electric field. Suitable catalysts that have been reported include Fe, Co, Mo, and Ni [10]. Compared to the arc-discharge and laser ablation methods, CVD uses much lower growth temperature, but it is still too high for direct CNT growth on CMOS substrates.

In addition, during or after CNT growth, electrical contacts need to be formed for functional CNT-based devices. It is reported that Mo provides good ohmic contacts with nanotubes and shows excellent conductivity after growth, with resistance ranging from 20 kΩ to 1 MΩ per tube [11]. Several other metals, such as palladium, gold, titanium, tantalum, and tungsten, have also been investigated as possible electrode materials.

CMOS-CNT Integration

To integrate CNT on CMOS, there are several factors that must be taken into account: temperature budget, material compatibility, CNT type, CNT quality, and contamination. Depending on when CNTs are made, CMOS-CNT integration technology can be categorized as follows:

- Pre-CMOS: CMOS processes will be performed after CNTs are synthesized in place
- Intra-CMOS: CNT growth steps are inserted into CMOS fabrication steps
- Post-CMOS: CNTs are introduced after all CMOS processes have been done

For pre-CMOS, CNTs must go through standard CMOS process steps. This is very difficult to realize. There are temperature constraints, material compatibility, and contamination issues. There is no report about pre-CMOS CNT integration yet. For intra-CMOS, CNTs will be introduced at a later stage in the CMOS fabrication sequence, so it is easier to protect CNTs than in the pre-CMOS case. But temperature and contamination issues still must be considered. Post-CMOS, on the hand, completely eliminates CMOS contamination issues. It has potential to achieve mass production and low cost, but the temperature remains a limiting factor.

Intra-CMOS CNT Integration

Intra-CMOS (High Temperature CNTs)

Thermal CVD has been used to grow CNTs directly on CMOS substrate. For example, Tseng et al. demonstrated, for the first time, a process that monolithically integrates SWNTs with n-channel metal oxide semiconductor (NMOS) FET in a CVD furnace at 875 °C [12]. However, the high synthesis temperature (typically 800–1000 °C for SWNT growth) may damage the aluminum metallization layers and change the characteristics of the on-chip transistors as well. Ghavanini et al. assessed the deterioration level of CMOS transistors with certain CNT CVD synthesis conditions applied, and they reported that one PMOS transistor lost its functions after the thermal CVD treatment (610 °C, 22 min) [13]. As a result, the integrated circuits in Tseng's thermal CVD CNT synthesis can only consist of NMOS and use n+ polysilicon and molybdenum as interconnects, which make it incompatible with foundry CMOS processes.

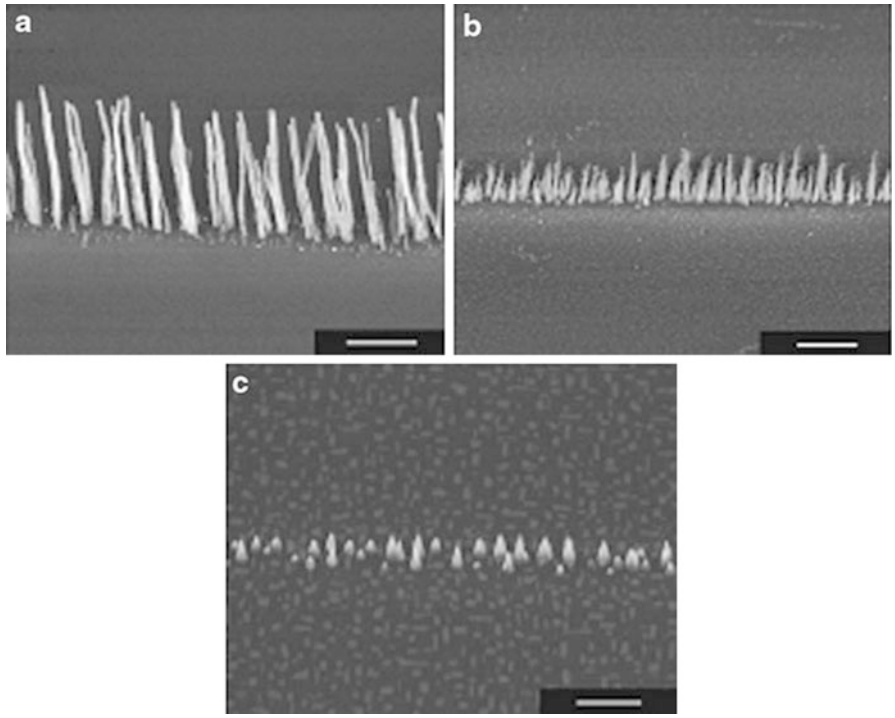
Intra-CMOS (Low Temperature CNTs)

Some other attempts have been made to develop low temperature growth using various CVD methods. Hofmann et al. reported vertically aligned carbon nanotubes grown at temperature as low as 120 °C by plasma-enhanced chemical vapor deposition (PECVD) [14]. However, the decrease in growth temperature jeopardizes both the quality and yield of the CNTs, as shown in Fig. 2. The synthesized products are actually defect-rich carbon nanofibers rather than MWNTs or SWNTs.

Intra-CMOS (Localized Heating)

Localized heating: To accommodate both the high temperature requirement (800–1000 °C) for high-quality SWNT synthesis and the temperature limitation of CMOS processing (<450 °C), CNT synthesis based on localized heating has drawn great interest recently. Englander et al. demonstrated, for the first time, the localized synthesis of silicon nanowires and carbon nanotubes based on resistive heating [15]. The fabrication processes are shown in Fig. 3. Operated inside a room temperature chamber, the suspended micro-electromechanical system (MEMS) structures serve as resistive heaters to provide high temperature at predefined regions for optimal nanotube growth, leaving the rest of the chip area at low temperature. Using the localized heating concept, direct integration of nanotubes at specific areas can be potentially achieved in a CMOS compatible manner, and there is no need for additional assembly steps. However, the devices typically have large sizes and their fabrication processes are not fully compatible with the standard foundry CMOS processes. Although this concept has solved the temperature incompatibility problem between CNT synthesis and CMOS circuit protection, the fabrication processes of microheater structures still have to be well designed to fit into standard CMOS foundry processes and the resistor materials must be selected to meet the CMOS compatibility criteria.

Using the localized heating technique described above, on-chip growth using CMOS micro-hotplates was demonstrated by Haque et al. [16]. As shown in Fig. 4, tungsten was used for both the micro-hotplates (as the heating source) and interdigitated electrodes for nanotubes contacts. MWNTs have been successfully synthesized on the membrane, and simultaneously connected to CMOS circuits through tungsten metallization. Although tungsten can survive the high temperature growth process, and has high connectivity and conductivity, Franklin et al. reported that no SWMTs were found to grow from catalyst particles on the



CMOS-CNT Integration, Fig. 2 SEM images of vertically aligned CNFs grown by PECVD deposition at (**a**) 500 °C, (**b**) 270 °C, and (**c**) 120 °C (scale bars: (**a**) and (**b**) 1 μm and (**c**) 500 nm) [14]

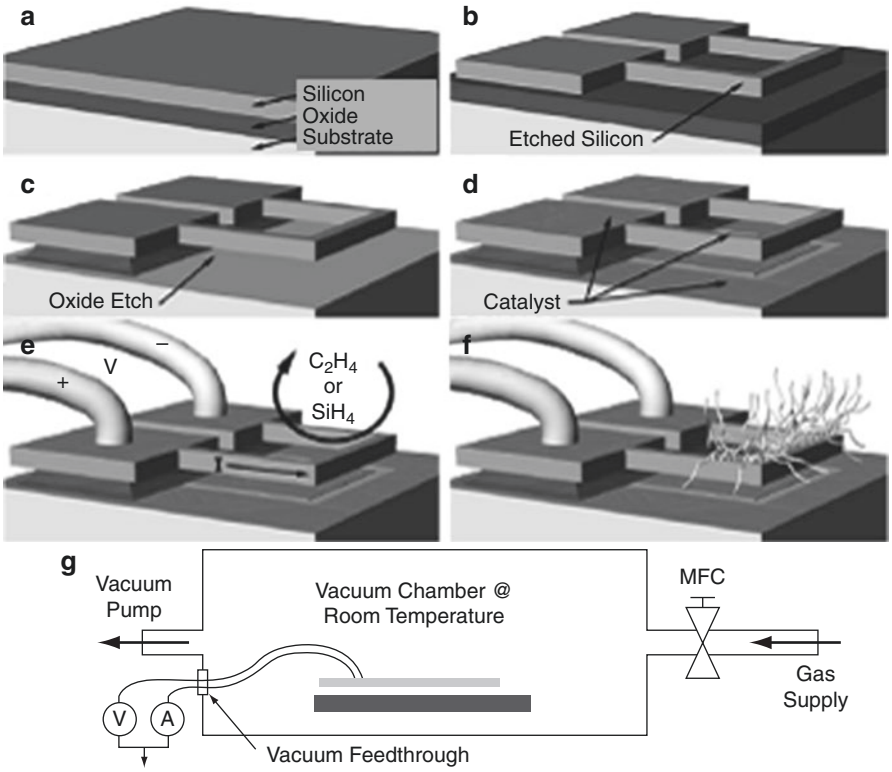
tungsten electrodes, presumably due to the high catalytic activity of tungsten toward hydrocarbons [11]. Further, although the monolithic integration has been achieved, the utilization of tungsten, a refractory metal, as interconnect metal is limited in foundry CMOS, especially for mixed-signal CMOS processes. Moreover, this approach requires a backside bulk micromachining process and is limited to SOI CMOS substrates.

Post-CMOS CNT Integration

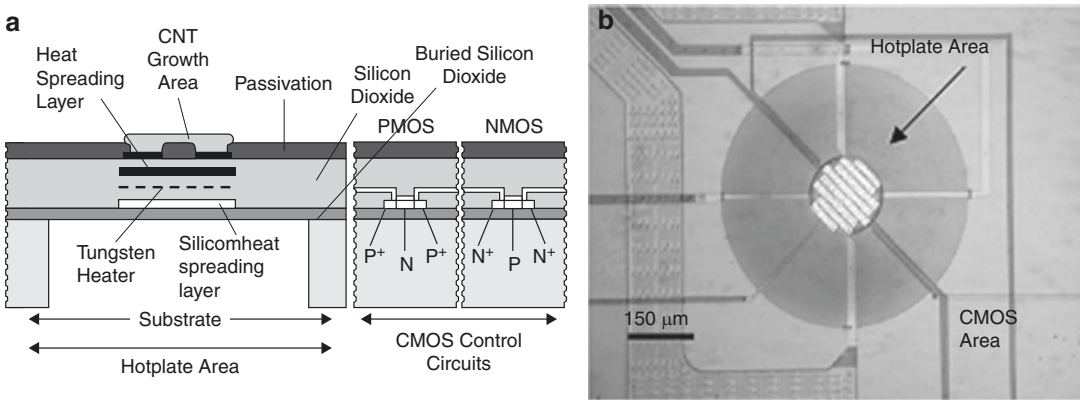
Post-CMOS (CNT Transfer and Assembly)

To overcome the temperature limitation, one possible solution is to grow nanotubes at high temperature first and then transfer them to the desired locations on CMOS substrates at low temperature. However, handling, maneuvering, and integrating these nanostructures with CMOS chips/wafers to form a complete system are very challenging. In the early stage, an atomic force microscope (AFM) tip was used to manipulate and position nanotubes into a predetermined location under the guide of scanning electron microscope (SEM) imaging [17]. Although this nanorobotic manipulation realized precise control over both the type and location of CNTs, its low throughput makes large scale assembly prohibitive.

Other post-growth CNT assembly methods include surface functionalization [18], liquid-crystalline processing [19], dielectrophoresis (DEP) [20], and large scale transfer of aligned nanotubes grown on quartz [21]. A 1 GHz CMOS circuit with CNT interconnects has been demonstrated using a DEP-assisted assembly technique [5]. The fabrication process flow and the assembled MWNT interconnect are shown in Fig. 5. The DEP process provides the capability of precisely positioning the nanotubes in a noncontact manner, which minimizes the parasitic capacitances and allows the circuits to operate at more than 1 GHz. However, to immobilize the DEP-trapped CNTs in place and to improve the electrical contact between CNTs and the



CMOS-CNT Integration, Fig. 3 Fabrication process and localized heating concept [15]

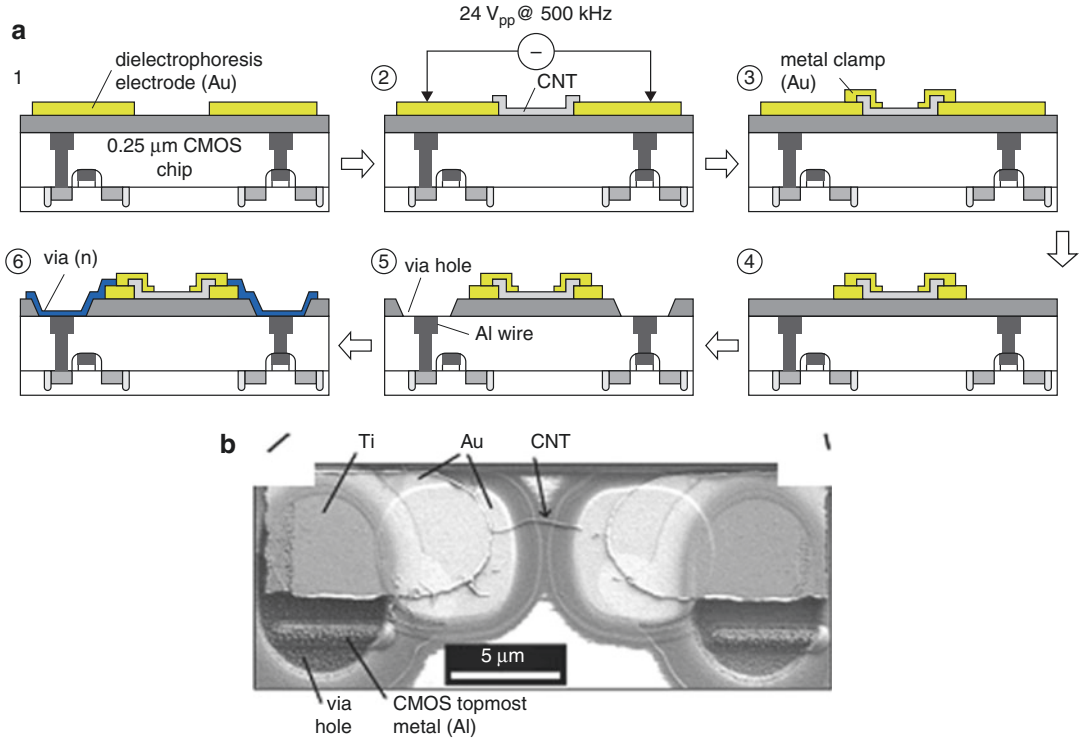


CMOS-CNT Integration, Fig. 4 (**a**) Schematic of the cross-sectional layout of the chip. (**b**) Optical image of the device top view showing the tungsten interdigitated electrodes on top of the membranes, heater radius = 75 µm, membrane radius = 280 µm [16]

electrodes, metal clamps must be selectively deposited at both ends of the CNTs (Fig. 5a, step 3). The process complexity and low yield (~8 %, due to the MWNT DEP assembly limitation) are still the major concerns.

Post-CMOS (Localized Heating)

Monolithic CMOS-CNT integration is desirable to fully utilize the potentials of nanotubes for emerging nanotechnology applications, but the approaches introduced above still cannot

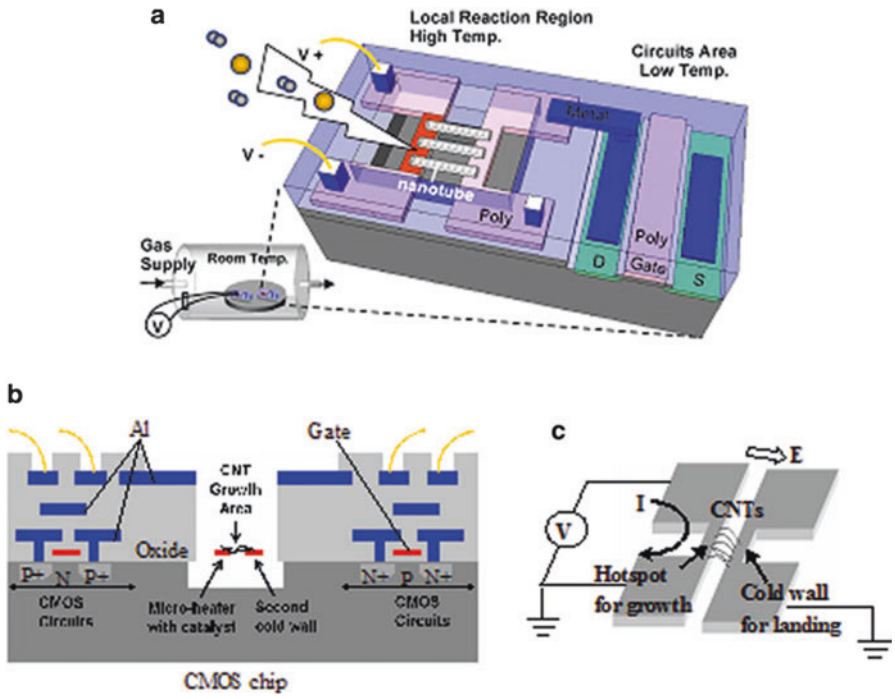


CMOS-CNT Integration, Fig. 5 (**a**) Process flow to integrate MWNT interconnects on CMOS substrate. (**b**) SEM image of one MWNT interconnect (wire and via) [5]

meet all the requirements and realize complete compatibility with CMOS processes. To solve the problem, a simple and scalable monolithic CMOS-CNT integration technique using a novel maskless post-CMOS surface micromachining processing has been proposed. This approach is fully compatible with commercial foundry CMOS processes and has no specific requirements on the type of metallization layers and substrates.

As illustrated in Fig. 6, the basic idea of the monolithic integration approach is to use maskless post-CMOS MEMS processing to form micro-cavities for thermal isolation and use the gate polysilicon to form resistors for localized heating as well as the nanotube-to-CMOS interconnect. The microheaters, made of the gate polysilicon, are deposited and patterned along with the gates of the transistors in the standard CMOS foundry processes. One of the top metal layers (i.e., the metal-3 layer as shown in Fig. 6b) is also patterned during the CMOS fabrication. It is used as an etching mask in the following post-CMOS microfabrication process for creating the micro-cavities. Finally, the polysilicon microheaters are exposed and suspended in a micro-cavity on a CMOS substrate. The circuits are covered under the metallization and passivation layers, as illustrated in Fig. 6b. Unlike the traditional thermal CVD synthesis in which the whole chamber is heated to above 800 °C, the CVD chamber is kept at room temperature all the time, with only the microheaters activated to provide the local high temperature for CNT growth (Fig. 6a, the red part represents the hot microheater).

The top view of a microheater design is shown in Fig. 6c. There are two polysilicon bridges: one as the microheater for generating high temperature to initiate CNT growth and the other for CNT landing. With the cold wall grounded, an E-field perpendicular to the surface of the two bridges will be induced during CNT growth. Activated by localized heating, the nanotubes will start to grow from the hotspot



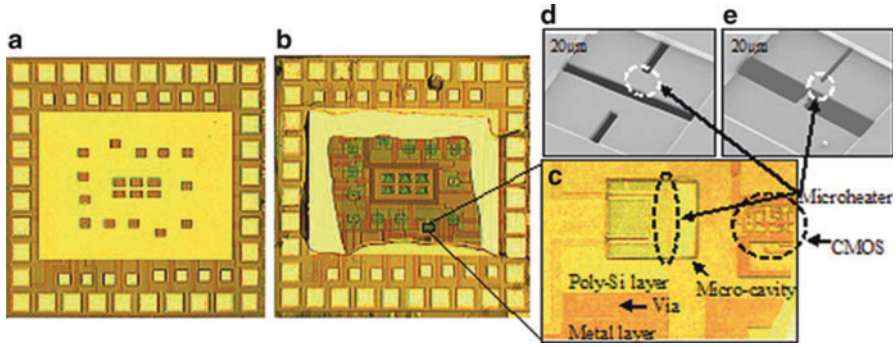
CMOS-CNT Integration, Fig. 6 (**a**) The 3D schematic showing the concept of the CMOS-integrated CNTs. The CVD chamber is kept at room temperature all the time. The *red* part represents the hot microheater that has been activated for high temperature nanotube synthesis. (**b**) Cross-sectional view of the device. (**c**) The schematic 3D microheater showing the local synthesis from the hotspot and self-assembly on the cold landing wall under the local electric field

(i.e., the center of the microheater) and will eventually reach the secondary cold bridge under the guide of the local E-field. Since both the microheater bridge and the landing bridge are made of the gate polysilicon layer and have been interconnected with the metal layers in CMOS foundry process, the as-grown CNTs can be electrically connected to the CMOS circuitry on the same chip without any post-growth clamping or connection steps.

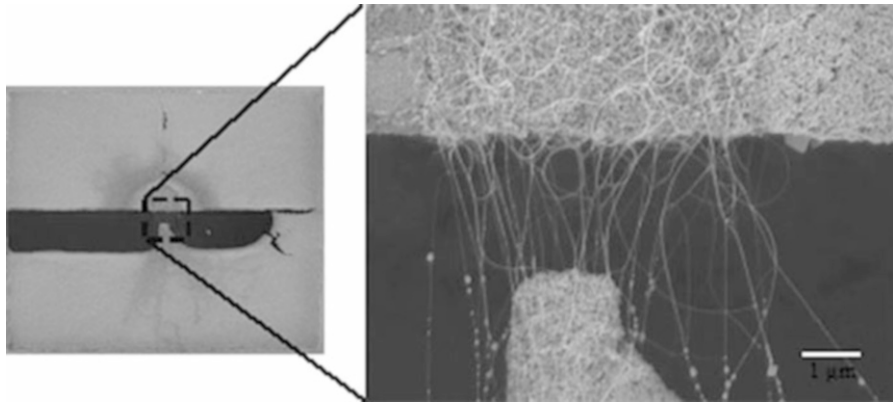
This technology has been verified at the chip level. The CMOS chips were fabricated in the AMI 0.5 μm 3-metal CMOS process. Optical microscope images of a CMOS chip before and after MEMs fabrication are shown in Fig. 7a, b. The total chip area is 1.5 × 1.5 mm^2, including test circuits and 13 embedded microheaters. SEMs of two microheaters are shown in Fig. 7d, e, with resistances of 97 and 117 Ω, respectively. At about 2.5 V, red glowing was observed for the design in Fig. 7e. This voltage was also used for the CNT growth.

Figure 8 shows one device with successful CNT growth, where individual suspended carbon nanotubes were grown from the 3 × 3 μm microheater shown in Fig. 7e and landed on the near polysilicon tip. The overall resistance of the CNTs is measured between the microheater and the cold polysilicon wall at room temperature. The typical resistances of in situ synthesized CNTs range from 5 to15 MΩ. The resistance variation from device to device is mainly due to the variation of the CNT quantity grown on each microheater. Junction effects of Schottky contacts were observed for self-assembled polysilicon/CNTs/polysilicon heterojunctions.

After successful synthesis of carbon nanotubes, the influence of the localized heating on nearby CMOS circuits was evaluated. Simple circuits, such as inverters, were tested and proved working properly. There was no change to the rising and falling time after the CNT growth. The dc electrical characteristics of individual



CMOS-CNT Integration, Fig. 7 (**a**) The CMOS chip photograph (1.5 × 1.5 mm^2) after foundry process; (**b**) The CMOS chip photograph after post-CMOS process (before final DRIE step); (**c**) Close-up optical image of one microheater and nearby circuit. CMOS circuit area, although visible, is protected under silicon dioxide layer. Only the microheater and secondary cold wall within the micro-cavity are exposed to synthesis gases. Polysilicon heater and metal wire are connected by via. (**d**) and (**e**) Closed-up SEM images of two microheaters



CMOS-CNT Integration, Fig. 8 Localized synthesis of carbon nanotubes grown from the 3 × 3 μm microheater, suspended across the trench and landed on the secondary polysilicon tip

transistors had no considerable change after CNT growth, demonstrating the CMOS compatibility of this integration approach.

Summary

CMOS-CNT integration has been demonstrated by using both intra- and post- CMOS processes. Several methods have been developed to overcome the temperature conflict between CNT growth and CMOS, including using high temperature refractory metals for interconnect, low temperature CVD, transferring/assembling CNTs prepared off site, and localized heating. Among these techniques, localized heating is very promising. Truly monolithic CNT-CMOS integration has been demonstrated on foundry CMOS substrate by employing MEMS and localized heating. This post-CMOS microfabrication is maskless, and the CNT growth does not affect the characteristics of the transistors on the same chip.

Cross-References

- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes
- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Synthesis of Carbon Nanotubes

References

1. Zhang, W., Jha, M., Shang, L.: NATURE: A hybrid nanotube/CMOS dynamically reconfigurable architecture. Design Automation Conference, 2006 43 rd ACM/IEEE, pp. 711–716 (2006)
2. Nantero, I.: "NRAM®," in http://www.nantero.com/mission.html, 2000–2009
3. Agarwal, V., Chen, C.-L., Dokmeci, M. R., Sonkusale, S.: A CMOS integrated thermal sensor based on single-walled carbon nanotubes. IEEE Sensors 2008 Conference, pp. 748–751 (2008)
4. Cho, T.S., Lee, K.-J., Kong, J., Chandrakasan, A.P.: A 32-uW 1.83-kS/s carbon nanotube chemical sensor system. IEEE J. Soild-State Circuits **44**, 659–669 (2009)
5. Close, G.F., Yasuda, S., Paul, B., Fujita, S., Wong, H.-S.P.: A 1 GHz integrated circuit with carbon nanotube interconnects and silicon transistors. Nano Lett. **8**, 706–709 (2008)
6. Chakraborty, R.S., Narasimhan, S., Bhunia, S.: Hybridization of CMOS with CNT-based nano-electromechanical switch for low leakage and Robust circuit design. IEEE Trans. Circuits Syst. **54**, 2480–2488 (2007)
7. Journet, C., Maser, W.K., Bernier, P., Loiseau, A., LamydelaChapelle, M., Lefrant, S., Deniard, P., Lee, R., Fischerk, J.E.: Large-scale production of single-walled carbon nanotubes by the electric-arc technique. Nature **388**, 756–758 (1997)
8. Guo, T., Nikolaev, P., Thess, A., Colbert, D.T., Smalley, R.E.: Catalytic growth of single-walled nanotubes by laser vaporization. Chem. Phys. Lett. **243**, 49–54 (1995)
9. Cassell, A.M., Raymakers, J.A., Kong, J., Dai, H.: Large scale CVD synthesis of single-walled carbon nanotubes. J. Phys. Chem. B **103**, 6484–6492 (1999)
10. Meyyappan, M.: Carbon Nanotubes: Science and Applications. CRC Press, New York (2005)
11. Franklin, N.R., Wang, Q., Tombler, T.W., Javey, A., Shim, M., Dai, H.: Integration of suspended carbon nanotube arrays into electronic devices and electromechanical systems. Appl. Phys. Lett. **81**, 913–915 (2002)
12. Tseng, Y.-C., Xuan, P., Javey, A., Malloy, R., Wang, Q., Bokor, J., Dai, H.: Monolithic integration of carbon nanotube devices with silicon MOS technology. Nano Lett. **4**, 123–127 (2004)
13. Ghavanini, F.A., Poche, H.L., Berg, J., Saleem, A.M., Kabir, M.S., Lundgren, P., Enoksson, P.: Compatibility assessment of CVD growth of carbon nanofibers on bulk CMOS devices. Nano Lett. **8**, 2437–2441 (2008)
14. Hofmann, S., Ducati, C., Robertson, J., Kleinsorge, B.: Low-temperature growth of carbon nanotubes by plasma-enhanced chemical vapor deposition. Appl. Phys. Lett. **83**, 135–137 (2003)
15. Englander, O., Christensen, D., Lin, L.: Local synthesis of silicon nanowires and carbon nanotubes on microbridges. Appl. Phys. Lett. **82**, 4797–4799 (2003)
16. Haque, M.S., Teo, K.B.K., Rupensinghe, N.L., Ali, S. Z., Haneef, I., Maeng, S., Park, J., Udrea, F., Milne, W. I.: On-chip deposition of carbon nanotubes using CMOS microhotplates. Nanotechnology **19**, 025607 (2008)
17. Huang, X.M.H., Caldwell, R., Huang, L., Jun, S.C., Huang, M., Sfeir, M.Y., O'Brien, S.P., Hone, J.: Controlled placement of individual carbon nanotubes. Nano Lett. **5**, 1515–1518 (2005)
18. Liu, J., Casavant, M.J., Cox, M., Walters, D.A., Boul, P., Lu, W., Rimberg, A.J., Smith, K.A., Colbert, D.T., Smalley, R.E.: Controlled deposition of individual single-walled carbon nanotubes on chemically functionalized templates. Chem. Phys. Lett. **303**, 125–129 (1999)
19. Ko, H., Tsukruk, V.V.: Liquid-crystalline processing of highly oriented carbon nanotube arrays for thin-film transistors. Nano Lett. **6**, 1443–1448 (2006)
20. Schwamb, T., Schirmer, N.C., Burg, B.R., Poulikakos, D.: Fountain-pen controlled dielectrophoresis for carbon nanotube-integration in device assembly. Appl. Phys. Lett. **93**, 193104 (2008)
21. Ryu, K., Badmaev, A., Wang, C., Lin, A., Patil, N., Gomez, L., Kumar, A., Mitra, S., Wong, H.-S.P., Zhou, C.: CMOS-analogous wafer-scale nanotube-on-insulator approach for submicrometer devices and integrated circuits using aligned nanotubes. Nano Lett. **9**, 189–197 (2009)

CMOS-MEMS

► CMOS MEMS Fabrication Technologies

CMOS-MEMS Resonators

Sheng-Shian Li
Institute of NanoEngineering and MicroSystems, National Tsing Hua University, Hsinchu, Taiwan

Synonyms

MEMS resonators integrated with circuits; MEMS/IC integrated resonator circuits; Micromechanical resonators fabricated using CMOS-MEMS technologies

Definition

Based upon a strict definition, a CMOS-MEMS resonator is fabricated using a CMOS foundry orientated process to realize MEMS/IC integration for resonator applications, such as oscillators and filters. The integrated CMOS-MEMS fabrication platform is a key technology to reduce the form factor, enhance the performance, increase the functionality, and facilitate circuit integration for portable sensing devices and Internet of Things (IoTs). The integrated CMOS-MEMS circuits greatly reduce the parasitic stray capacitances at the MEMS-circuit interface so that it not only improves the system responses at high frequencies but also reduces the power consumption of the electronic circuits. This platform benefits the future portable/wearable electronic products for smaller size and longer standby time.

Principle of Operation

CMOS-MEMS resonators are mainly formed by the CMOS back-end-of-line (BEOL) materials, including interlayer dielectrics (ILD), e.g., SiO_2, and conductive metals, e.g., AlCu or Cu [1]. The target of this technology is to facilitate MEMS/IC monolithic integration without the need of additionally complex processes, such as heterogeneous integration using wafer bonding or 3D IC. Figure 1 presents two configurations of CMOS technology where the BEOL is highlighted. As can be seen, the BEOL materials will be used as MEMS structures of resonators, while the transistor circuits are naturally integrated with the resonators, thus leading to a much smaller form factor and less parasitics as compared to the current technologies (e.g., two-chip or heterogeneous integration solution). Since the capacitive transducers inherently exist, they are used as the major driving and sensing mechanism for CMOS-MEMS resonators [1–4] although the piezoresistive readout [5] can also be implemented once the polysilicon material (cf. Fig. 1) serves as the piezoresistors for motion detection. In this entry, only the pure capacitively transduced CMOS-MEMS resonator technology, i.e., the mainstream transduction in CMOS-MEMS, is introduced.

Resonator Structure

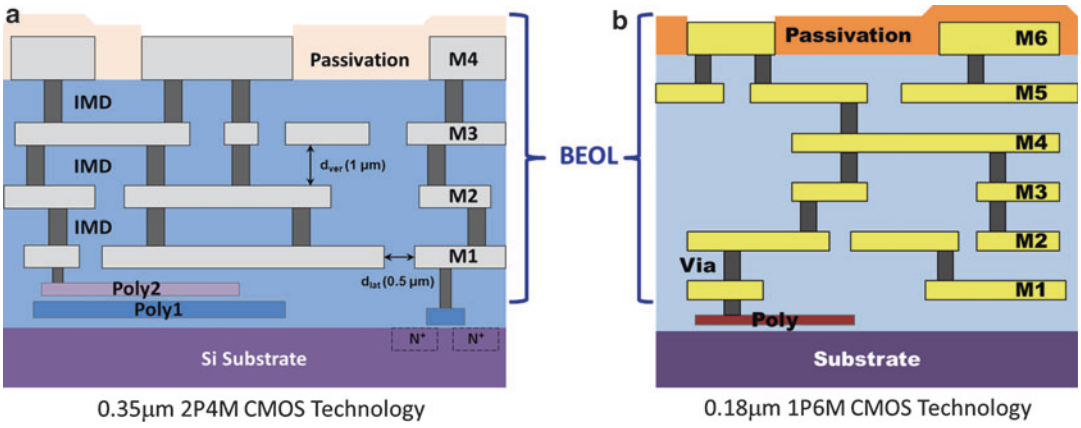
Based on various maskless post-CMOS release processes (i.e., no additional lithography step is required), the CMOS-MEMS resonator structures can be categorized into (a) mere-metal, (b) metal-rich composite, (c) and oxide-rich composite configurations, as shown in Fig. 2. The implemented processes resulting in structures in Fig. 2 will be discussed in the next section. With these structures, the CMOS-MEMS resonators exhibit different performance in terms of quality factor (Q) and thermal stability (i.e., temperature coefficient of frequency, TC_f).

Resonance Frequency

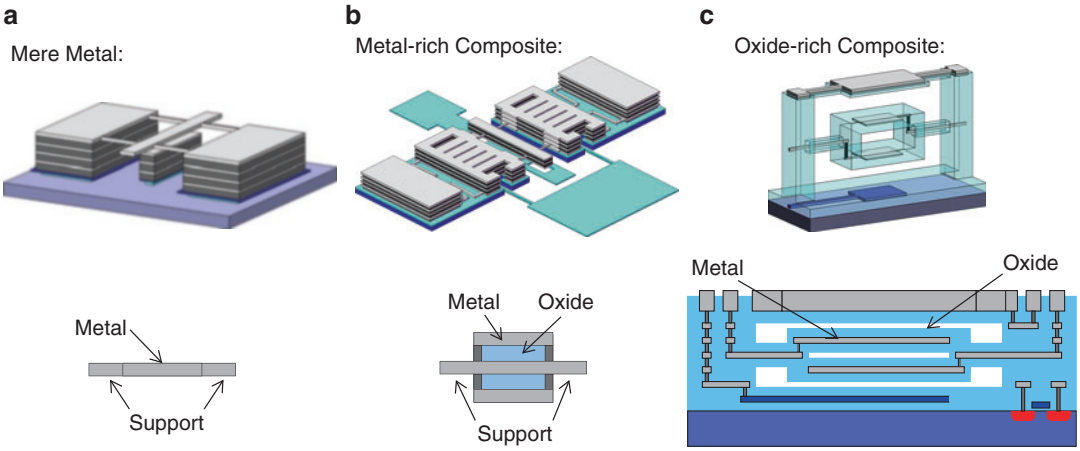
To determine the resonance frequency of the CMOS-MEMS composite beam resonators indicated in Fig. 2b and c, the Euler-Bernoulli approach can be used to attain a first-order approximation of the resonance frequency (f_{nom}) using a formula given by

$$f_{nom} = \frac{1}{2\pi}\sqrt{\frac{k_m}{m_{re}}} = \frac{1}{2\pi}(\beta_1 L_r)^2\sqrt{\frac{\sum(E_i I_i)}{\sum(\rho_i A_i)}}\frac{1}{L_r^2} \quad (1)$$

where k_m and m_{re} are the mechanical stiffness and effective mass, respectively, of the beam resonator, L_r is the length of the beam, β_1 represents the frequency parameter of a fundamental mode of the beam, i represents the corresponding CMOS structural materials (e.g., metal and SiO_2), and E_i, ρ_i, A_i, and I_i are the Young's modulus, density, cross-sectional area, and moment of inertia of each structural layer, respectively. If the mere-metal resonator is designed, Eq. 1 can be greatly simplified by using only one uniform structural material. Precise resonance frequencies can be simulated by the finite-element analysis once more complicated or accurate structures of the CMOS-MEMS resonators are considered, as shown in Fig. 3.



CMOS-MEMS Resonators, Fig. 1 CMOS configurations. (**a**) 0.35 µm 2-poly-4-metal (2P4M) CMOS technology node. (**b**) 0.18 µm 1-poly-6-metal (1P6M) CMOS technology node



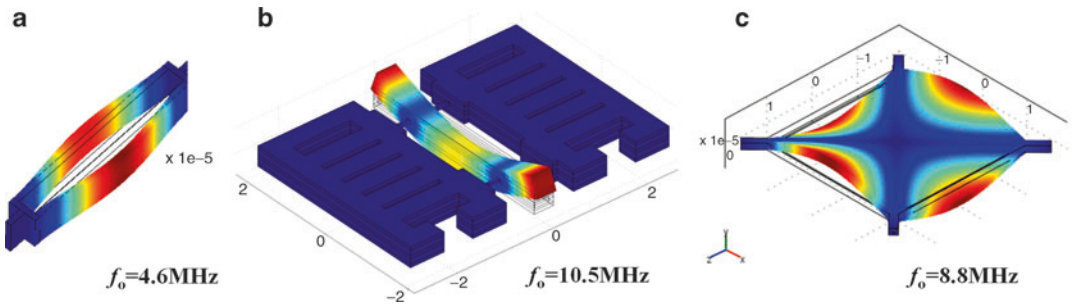
CMOS-MEMS Resonators, Fig. 2 Typical structures of CMOS-MEMS resonators in their perspective and cross-sectional views. (**a**) Mere-metal resonator. (**b**) Metal-rich composite resonator where metal layers enclose oxide. (**c**) Oxide-rich composite resonator where oxide layers enclose metal

Transduction and Overall Transfer Function

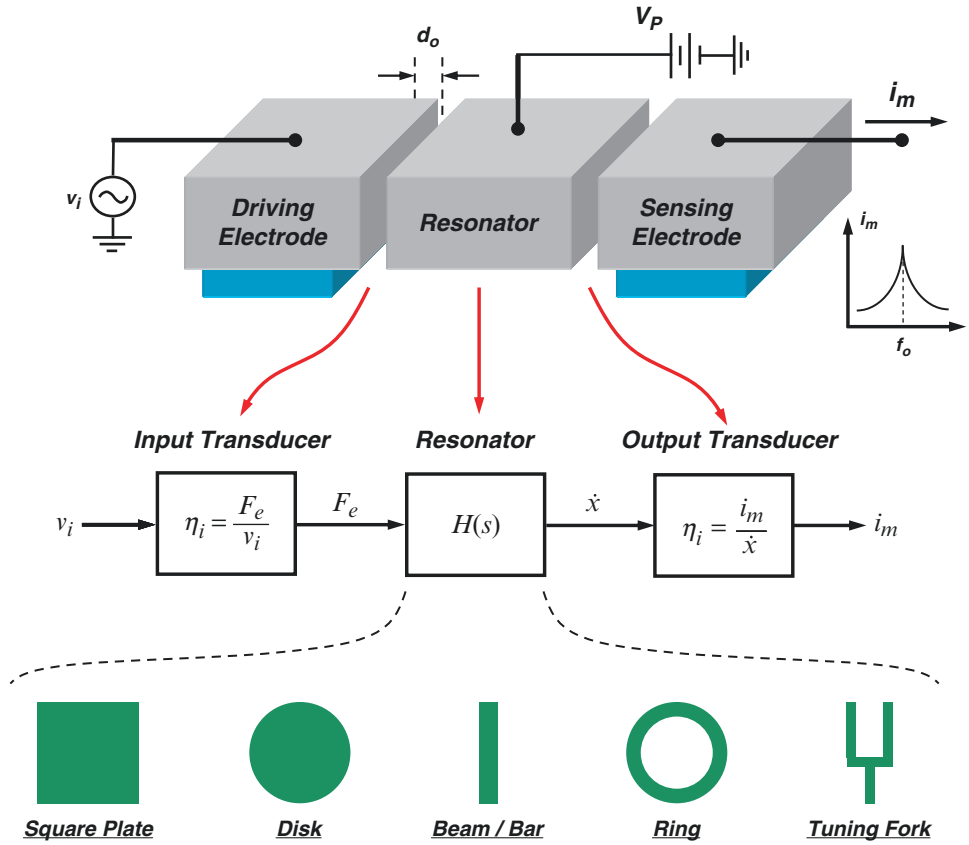
After the resonance frequency is set through Eq. 1 based on the geometrical dimensions and material properties of a given resonator, one has to determine the capacitive transduction efficiency to obtain the most important parameter, the motional impedance R_m, and overall transmission of the CMOS-MEMS resonator. A block diagram shown in Fig. 4 can be used to facilitate the understanding of the CMOS-MEMS resonator which is modeled in an electrical domain by linking the mechanical transfer function $H(s)$ with the input/output electromechanical coupling coefficients, η_i and η_o. The overall admittance transfer function $Y = i_m/v_i$ of the CMOS-MEMS resonator can be expressed by

$$Y(s) = \frac{i_m(s)}{v_i(s)} = \left(\frac{F_e}{v_i}\right) \times \left(\frac{\dot{x}}{F_e}\right) \times \left(\frac{i_m}{\dot{x}}\right) = \eta_i \times H(s) \times \eta_o \quad (2)$$

where v_i and i_m represent the small-signal input driving voltage and output motional current of the resonator, F_e is the electrostatic force generated by v_i to drive the resonator into resonance, and $\dot{x}$

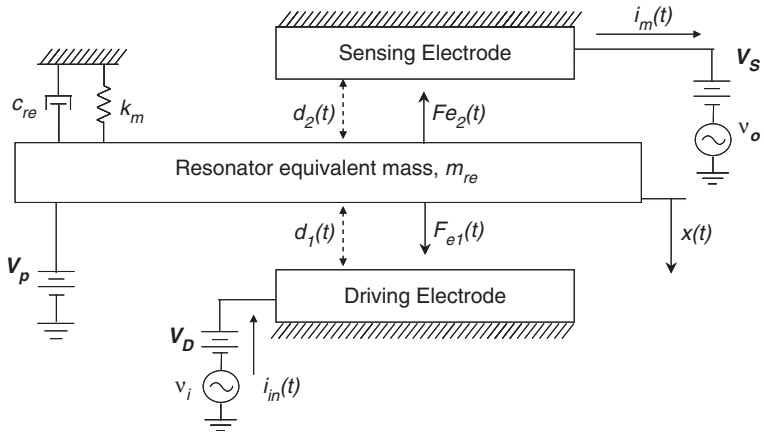


CMOS-MEMS Resonators, Fig. 3 Finite-element simulated modal frequencies and corresponding mode shapes of CMOS-MEMS resonators. (**a**) Tuning fork. (**b**) Free-free beam. (**c**) Flexural mode plate



CMOS-MEMS Resonators, Fig. 4 Conceptual schematic and block-diagram representation of a generalized capacitive MEMS resonator

represents the velocity of the resonator due to vibration. This is why η_i and η_o are treated as the electromechanical coupling coefficients where η_i converts the "electrical signal v_i" into the "mechanical force F_e" while η_i transforms the "mechanical velocity $\dot{x}$" into the "electrical current i_m". The mechanical transfer function $H(s)$ in Fig. 4 represents a general high-Q biquad response (cf. inset of Fig. 4) of the vibrating resonators, some of which are illustrated at the



CMOS-MEMS Resonators, Fig. 5 Equivalent lumped mechanical model for a capacitive MEMS resonator

bottom. Each type of resonators in Fig. 4 can be modeled by a simple mass-spring-damper system, as shown in Fig. 5, where the capacitive transducer gaps and voltage differences between the electrodes and the resonator are assumed to be different to preserve generality.

To derive the equivalent mechanical transfer function, $H(s)$, the equation of motion for the second-order mechanical system is used, given by

$$m_{re}\ddot{x} + c_{re}\dot{x} + k_m x = F_e \tag{3}$$

where x is the position of the mass away from its equilibrium position, m_{re} is the equivalent mass, c_{re} is the equivalent damping, k_m is the equivalent linear spring constant of the resonator, and F_e is the external driving force, respectively. Each element in Eq. 3 is assumed to be linear.

The force-to-velocity transfer function $H(s)$ can then be derived as

$$H(s) = \frac{\dot{x}}{F_e} = \frac{sX(s)}{F_e(s)} = \frac{1/m_{re}}{s^2 + \frac{c_{re}}{m_{re}}s + \frac{k_m}{m_{re}}}. \tag{4}$$

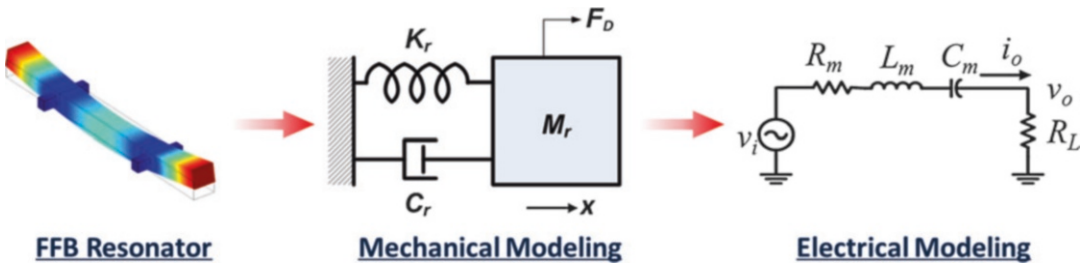
In Fig. 5, V_D, V_S, and V_P are the DC bias voltages applied into the input electrode, output electrode, and resonator structure, respectively. Based on the derivation of the capacitive transducers, the first-order electrostatic force and the output motional current of the resonator are given by

$$\begin{aligned} F_e &= -(V_P - V_D)\frac{\partial C_1}{\partial x}v_i \\ &= (V_P - V_D)\frac{\varepsilon_o A_1}{d_1^2}v_i \text{ and } \eta_i \equiv \frac{F_e}{v_i} \\ &= -(V_P - V_D)\frac{\partial C_1}{\partial x} \end{aligned} \tag{5}$$

$$\begin{aligned} i_m &= -(V_P - V_S)\frac{\partial C_2}{\partial x}\dot{x} \\ &= (V_P - V_S)\frac{\varepsilon_o A_2}{d_2^2}\dot{x} \text{ and } \eta_o \equiv \frac{i_m}{\dot{x}} \\ &= -(V_P - V_S)\frac{\partial C_2}{\partial x} \end{aligned} \tag{6}$$

where ε_o is the air permittivity; C_1 and C_2 are capacitors of the drive/sense ports, respectively; A_1 and A_2 are capacitive transducer areas of the drive/sense ports, respectively; and d_1 and d_2 are the capacitive transducer gaps of the drive/sense ports, respectively.

To simply describe the modeling procedure, Fig. 6 shows the CMOS-MEMS resonator, such as a free-free beam resonator, first modeled in mechanical domain as a mass-spring-damper (*MKC*) system. Then the *MKC* system is further transformed into a real *RLC* equivalent circuit in electrical domain through the electromechanical analogy. Based on Eqs. 2, 3, 4, 5, and 6, an equivalent series *RLC* model of an electromechanical resonator shown in Fig. 6 can be obtained with the motional elements expressed as



CMOS-MEMS Resonators, Fig. 6 Equivalent lumped electrical model for a capacitive MEMS resonator

$$L_m = \frac{m_{re}}{\eta_i \times \eta_o}, C_m = \frac{\eta_i \times \eta_o}{k_m}, R_m = \frac{c_{re}}{\eta_i \times \eta_o}. \tag{7}$$

For most capacitively transduced micromechanical resonators, their larger-than-conventional impedance (i.e., motional impedance R_m) is the main issue for system integration. Assuming the gap size d_o and overlap area A of the input and output transducers are symmetric, the motional impedance R_m in Eq. 7 can be further simplified as

$$R_m = \frac{c_{re}}{\eta_e^2} \cong \frac{\sqrt{k_m m_{re}} d_o^4}{Q \varepsilon_o^2 A^2 V_P^2} \tag{8}$$

where $\eta_e = \eta_i = \eta_o$, $d_o = d_1 = d_2$, $A = A_1 = A_2$, and Q is the quality factor of the resonator.

It is evident that to reduce R_m of a capacitive resonator, the stiffness, mass, gap size, and overlap area must be carefully designed under a V_P-constrained condition (high bias V_P is not desirable). Several design approaches can be adopted to reduce R_m, including (i) the gap size d_o should be as small as possible, (ii) the transducer overlap area A should be as large as possible, and (iii) the Q factor of the resonator should be as high as possible.

Thermal Stability

The nominal mechanical resonance frequency of the composite beam resonator (cf. Fig. 2b) composed of three structural materials, including metal (i.e., aluminum alloy), tungsten, and oxide, can be expressed as a combination of the resonance frequency of each portion for a given composite beam by an algebraic manipulation of Eq. 9 and given by [6]

$$f_{nom}{}^2 = \frac{m_{\text{metal}}}{m_r} f_{\text{metal}}{}^2 + \frac{m_{\text{tungsten}}}{m_r} f_{\text{tungsten}}{}^2 + \frac{m_{\text{oxide}}}{m_r} f_{\text{oxide}}{}^2 \tag{9}$$

where m_{metal}, m_{tungsten}, and m_{oxide} are the portion of mass of the composite beam for metal, tungsten, and oxide constituents, respectively, and f_{metal}, f_{tungsten}, and f_{oxide} are the mechanical resonance frequencies of beams for metal, tungsten, and oxide constituents, respectively. As a result, the linear temperature coefficient of frequency (TC_f) of the composite free-free beam resonator can be derived by taking the derivative of Eq. 9 with respect to temperature and then simplified as

$$TC_{f1} = \frac{TC_{f1,\ \text{metal}} + A \cdot TC_{f1,\ \text{tungsten}} + B \cdot TC_{f1,\ \text{oxide}}}{1 + A + B} \tag{10}$$

where $A = \frac{m_{\text{tungsten}} \cdot f^2_{\text{tungsten}}}{m_{\text{metal}} \cdot f^2_{\text{metal}}} = \frac{E_{\text{tungsten}} \cdot I_{\text{tungsten}}}{E_{\text{metal}} \cdot I_{\text{metal}}}$ and $B = \frac{m_{\text{oxide}} \cdot f^2_{\text{oxide}}}{m_{\text{metal}} \cdot f^2_{\text{metal}}} = \frac{E_{\text{oxide}} \cdot I_{\text{oxide}}}{E_{\text{metal}} \cdot I_{\text{metal}}}$ and where $TC_{f1,\text{metal}}$, $TC_{f1,\text{tungsten}}$, and $TC_{f1,\text{oxide}}$ are the linear TC_fs of metal, tungsten, and oxide, respectively; E_{metal}, E_{tungsten}, and E_{oxide} are Young's moduli of metal, tungsten, and oxide, respectively; and I_{metal}, I_{tungsten}, and I_{oxide} are the moments of inertia of metal, tungsten, and oxide, respectively. In addition, the linear TC_f of each material in Eq. 10 can be expressed as [6]

$$TC_{f1,\ \text{mat}} = \frac{TC_{E1,\ \text{mat}} + \alpha_{1,\ \text{mat}}}{2} \tag{11}$$

where $TC_{E1,\,mat}$ and $\alpha_{1,\,mat}$ are the linear TC_E and the coefficient of thermal expansion (CTE), respectively, of each structural material. In Eq. 11, $TC_{E1,\,mat}$ is the major parameter to induce the resonance frequency drift as temperature changes (i.e., $TC_{f1,\,mat}$), indicating the structural oxide with positive TC_E is effective for passive temperature compensation of CMOS-MEMS resonators. To approach the ultimate goal of zero TC_f, thermal stability of the CMOS-MEMS composite resonators can be further improved by manipulating the ratio A and ratio B in Eq. 10, respectively. As an extreme case when both A and B are close to zero, it represents that TC_f of such a resonator is exactly the same with that of metal-type resonators, while structures with A of zero and infinite B belong to oxide-type resonators exhibiting positive TC_f.

Methods of Fabrication

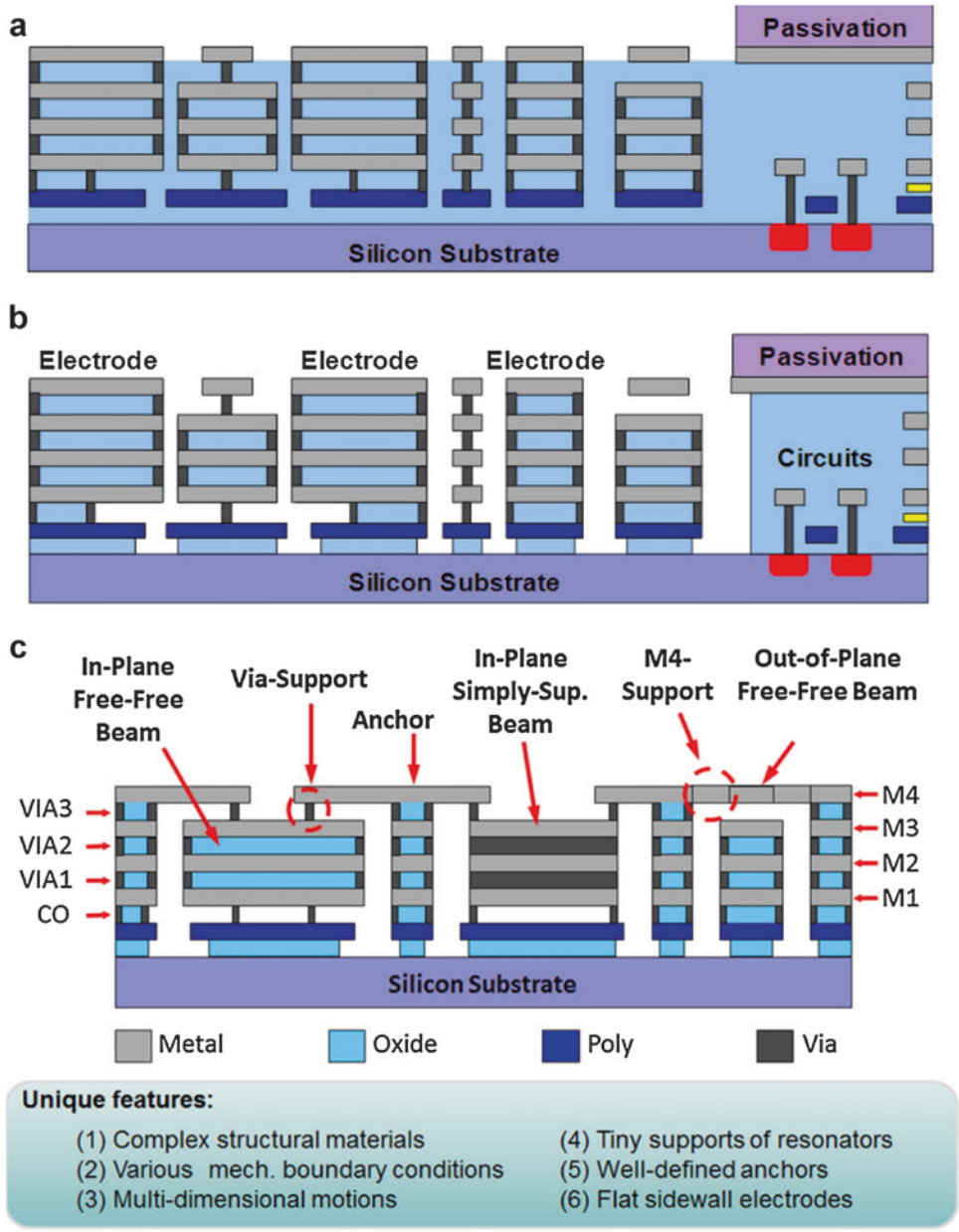
Several CMOS-MEMS resonator platforms [1–4] co-fabricating mechanical resonators and their amplifier circuits for MEMS/IC integration have been developed in the past decades. These platforms can be mainly categorized into two different release methods – oxide removal and metal removal maskless post-CMOS processes, implemented in both 0.35 μm 2-poly-4-metal (2P4M) and 0.18 μm 1-poly-6-metal (1P6M) CMOS technologies. In the following contents, their pros and cons will be discussed.

Oxide Removal Release Process in a 0.35 μm CMOS

To fabricate resonators using the platform illustrated in Fig. 7, chips are manufactured utilizing a standard 0.35 μm 2-poly-4-metal CMOS service from the Taiwan Semiconductor Manufacturing Company (TSMC) with a cross-sectional view shown in Fig. 7a, including two polysilicon layers underneath four metal layers [1]. CMOS circuit areas are masked by the passivation layer mostly comprised of silicon nitride while the MEMS regions (i.e., openings) expose sacrificial oxide to the release etchant. A commercial SiO_2 etchant, silox vapox III, with very high selectivity to metal layers, vias (i.e., tungsten), and contacts (i.e., tungsten) is utilized to release the resonator structures as depicted in Fig. 7b without the help of critical point dryers while the transistor circuits are still protected by the passivation layer. Figure 7c indicates that CMOS-MEMS resonators fabricated using this platform specifically possess several unique features including (i) complex structural materials which can be made of metal-oxide composite, metal composite, and mere-metal structures; (ii) various mechanical boundary conditions for resonators, such as fixed, pinned-pinned boundary, and free-free boundary designs; (iii) multidimensional displacements of resonators capable of in-plane and out-of-plane motions with respect to the substrate surface; (iv) standard CMOS vias (VIA) and contacts (CO) acting as tiny supports of resonators to effectively isolate the vibrating energy from resonators to their anchors; (v) well-defined anchors without undercut issue which is often seen in conventional CMOS-MEMS [7–11] or SOI process [12]; and (vi) better transducer efficiency attained by utilizing via-connected walls to form a flat sidewall electrodes, all of which offer a variety of flexible options suited for sensor and RF applications. Figure 8 presents the fabricated resonators, such as beam and comb-drive resonators. Note that the minimal electrode-to-resonator gap spacing is limited by the feature size of the 0.35 μm CMOS technology node such that the motional impedance of the fabricated resonators is in several MΩs due to their low electromechanical coupling coefficient.

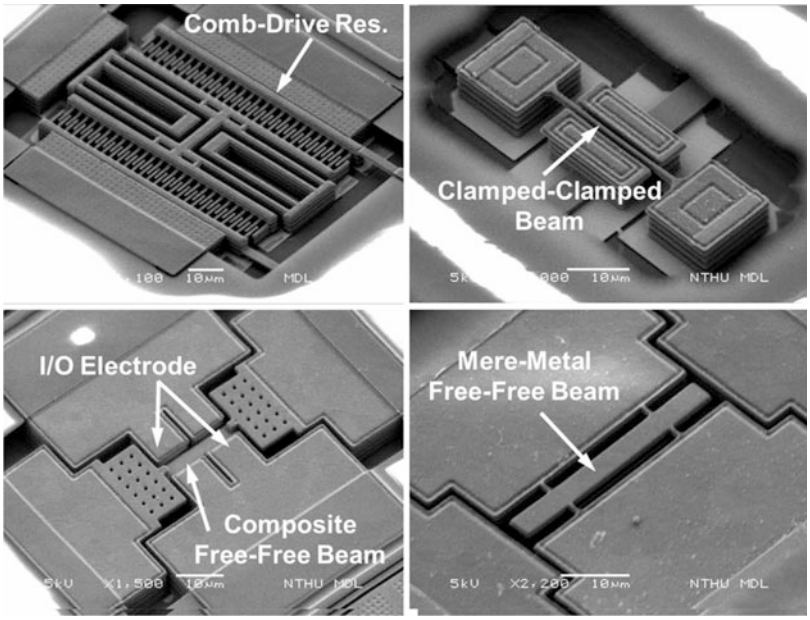
Oxide Removal Release Process in a 0.18 μm CMOS

To address the issue of high motional impedance in the previous platform, one can transfer this oxide removal process of section "Oxide Removal Release Process in a 0.35 μm CMOS" into a 0.18 μm CMOS as indicated in Fig. 9a [2]. The cross-sectional view of CMOS 0.18 μm BEOL and SEM of the fabricated resonator are shown in Fig. 9b, c, respectively. The electrode-to-resonator gap spacing is reduced



CMOS-MEMS Resonators, Fig. 7 Fabrication process flow for a CMOS-MEMS resonator in a 0.35 μm 2P4M technology under the oxide removal post process. (**a**) Chip prepared by the CMOS foundry. (**b**) Device is released by a buffer HF solution. (**c**) Various configurations of CMOS-MEMS resonators are realized in the platform

through the minimal feature size of the 0.18 μm CMOS (1.8X smaller than the 0.35 μm CMOS), while the transducer area can be increased by the six-metal stacking (four-metal stacking in the 0.35 μm CMOS). The combined merit of the gap and area leads to much lower motional impedance as compared to the 0.35 μm CMOS-MEMS resonators. The use of the 0.18 μm 1P6M platform implies the advanced technology nodes would lower the motional impedance, reduce the DC bias voltage, and attain higher resonance frequencies.



CMOS-MEMS Resonators, Fig. 8 SEM photos of the fabricated resonators using the oxide removal release process in a 0.35 μm CMOS technology

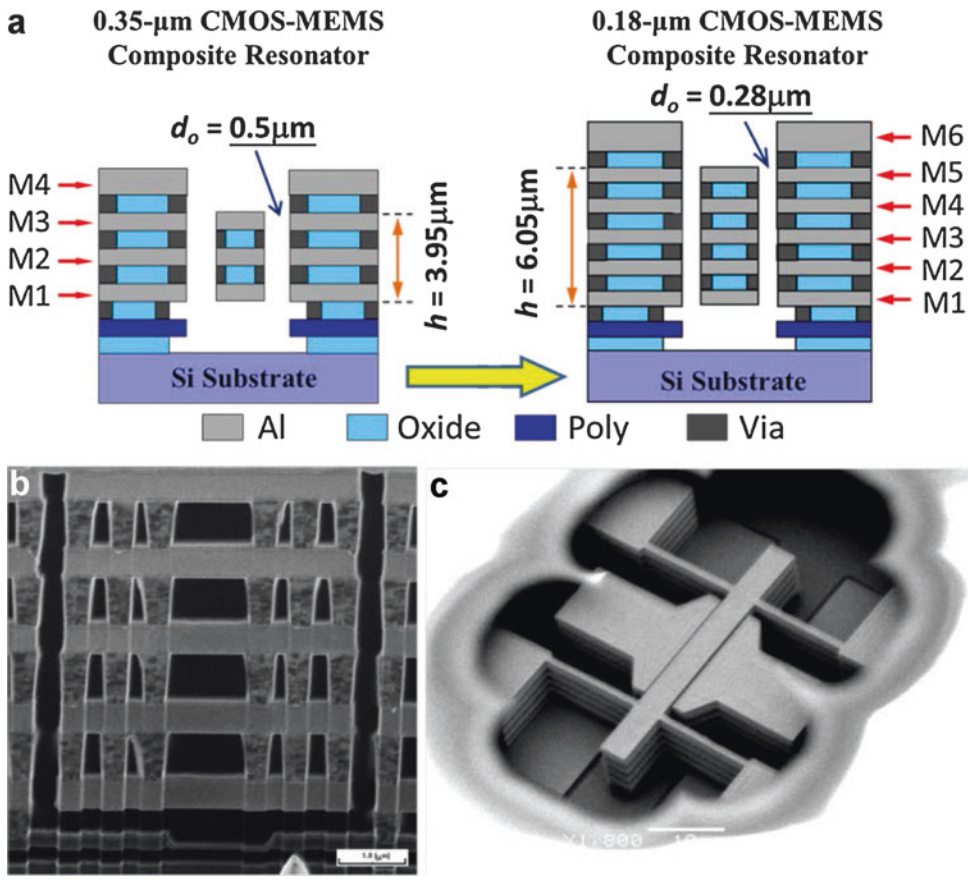
Metal Removal Release Process in a 0.35 μm CMOS

Although the oxide removal process only requires one maskless wet etching step, the metal-rich feature of the aforementioned resonators places a bottleneck on quality factor Q since the structural metal (AlCu) is often treated as a lossy acoustic material. Q of metal-rich resonators is often lower than 2,000, which is not sufficient for practical oscillator and filter implementations [3]. To address the low-Q issue in the oxide removal process, a metal removal post process has been developed as an alternative to form low-loss oxide-rich resonator structures to enable high Q.

As indicated in Fig. 10a, the device is first prepared by the CMOS foundry using a standard 0.35 μm 2P4M CMOS process. Then the etching solution containing H_2SO_4 and H_2O_2 is utilized to remove metal layers. This maskless etching technique provides excellent selectivity between the metal and dielectric layers; therefore, the issues from time-based etching in sections "Oxide Removal Release Process in a 0.35 μm CMOS" and "Oxide Removal Release Process in a 0.18 μm CMOS" can be greatly alleviated. Finally, reactive ion etching (RIE) is used to remove the passivation layer for the later wire bonding. Figure 10b presents a resonator SEM photo and its cross-sectional view, showing the metal layers are embedded inside the main SiO_2 structure. Q of the oxide-rich resonator is greater than 6,000, which is 6X higher than the metal-rich counterparts in sections "Oxide Removal Release Process in a 0.35 μm CMOS" and "Oxide Removal Release Process in a 0.18 μm CMOS."

Metal Removal Release Process in a 0.18 μm CMOS

The release approach in section "Metal Removal Release Process in a 0.35 μm CMOS" has also been transferred into a 0.18 μm CMOS node as shown in Fig. 11 to gain the smaller feature size of the advanced CMOS for low motional impedance R_m [4]. To fabricate the resonators, the chips are manufactured using a standard TSMC 0.18 μm 1P6M CMOS foundry process. Then a metal wet etchant, comprising H_2SO_4 and H_2O_2, is utilized to remove the sacrificial metals, hence providing lateral air-gap spacing d_o of 0.38 μm. Finally, the RIE is utilized to open bond pads for later wire bonding. As can be seen in the cross-sectional view of the resonator in Fig. 11b, the resonator structure consists of more than 97 % SiO_2. As shown in Fig. 11c, such oxide-rich resonators



CMOS-MEMS Resonators, Fig. 9 CMOS-MEMS resonator in a 0.18 μm 1P6M CMOS-MEMS platform. (**a**) Technology transformation. (**b**) Cross-sectional view of 0.18 μm BEOL. (**c**) Free-free beam resonator realized in a 0.18 μm CMOS-MEMS platform

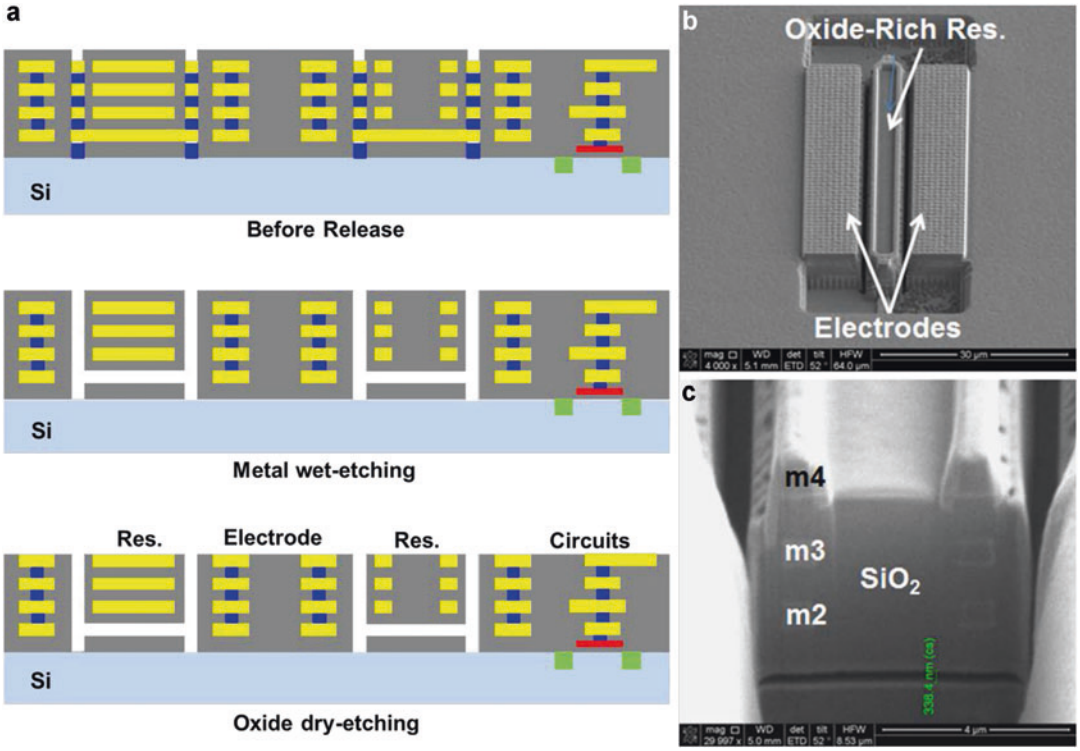
possess Q more than 10,000, which is comparable to commercial single-crystal silicon (SCS) or polysilicon-based resonators. Even tested in air, Q of the resonators is still greater than 6,000, an evidence of less squeeze film damping effect and suitable for resonant sensor applications.

Main Research Accomplishment

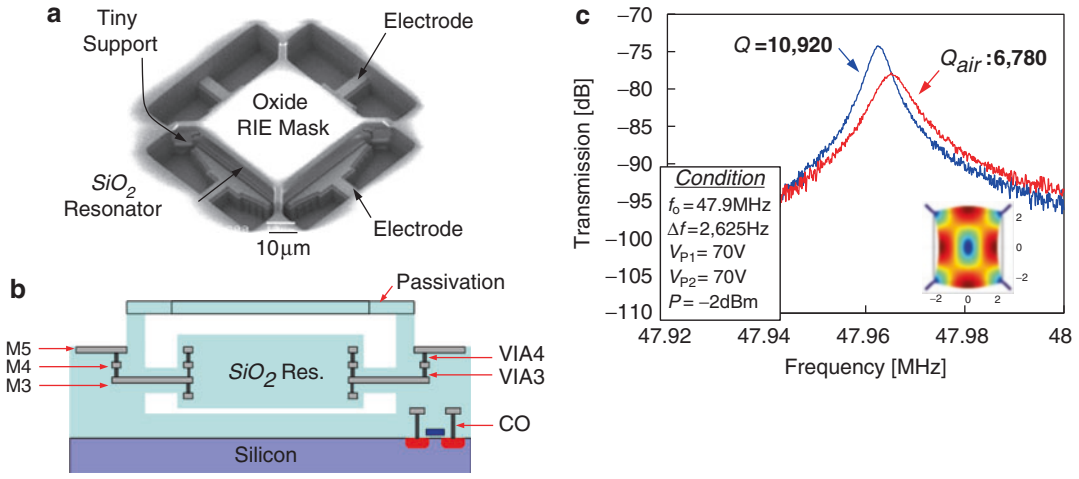
Motional Impedance R_m Reduction

The motional impedance governed by Eq. 8 is strongly dependent on the electrode-to-resonator gap spacing d_o. Figure 12 presents the performance comparison of similar free-free beam designs implemented in the 0.35 μm 2P4M and 0.18 μm 1P6M platforms, showing the advanced technology node (0.18 μm) would significantly lower the motional impedance R_m due to its smaller gap spacing (1.8X) and larger transduction area (1.6X) as compared to the 0.35 μm 2P4M technology.

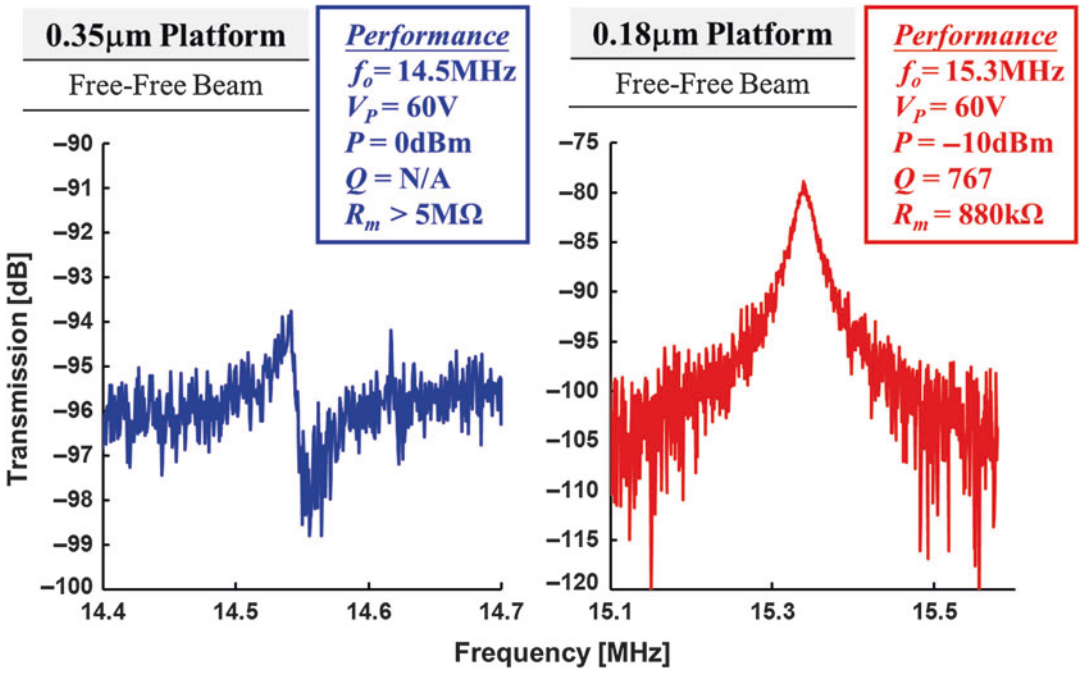
In addition to the use of the advanced technology node, a mechanically coupled array design [13] offers an efficient approach to reduce the motional impedance R_m which is inversely proportional to the number of the constituent resonators in an array device. The more resonators mechanically coupled as an array, the smaller the motional impedance. Since the coupled array



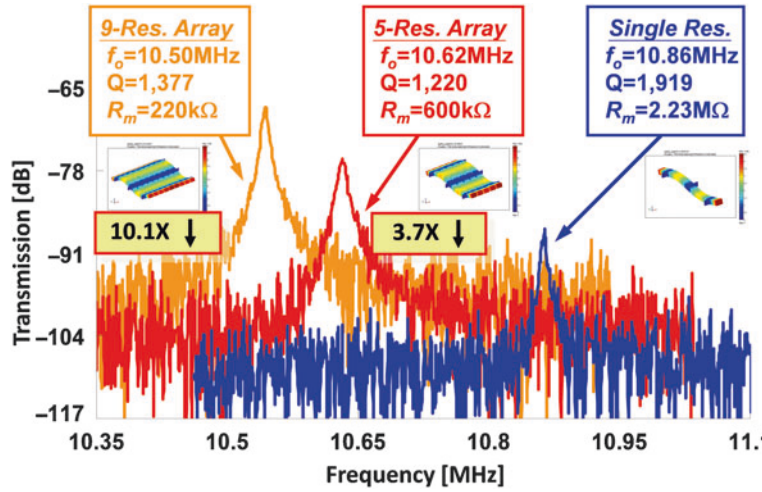
CMOS-MEMS Resonators, Fig. 10 CMOS-MEMS resonator using the metal removal post process in a 0.35 μm 2P4M CMOS-MEMS platform



CMOS-MEMS Resonators, Fig. 11 (**a**) SEM, (**b**) cross-sectional configuration, and (**c**) frequency responses of a CMOS-MEMS resonator fabricated in a 0.18 μm 1P6M technology under the metal removal post process

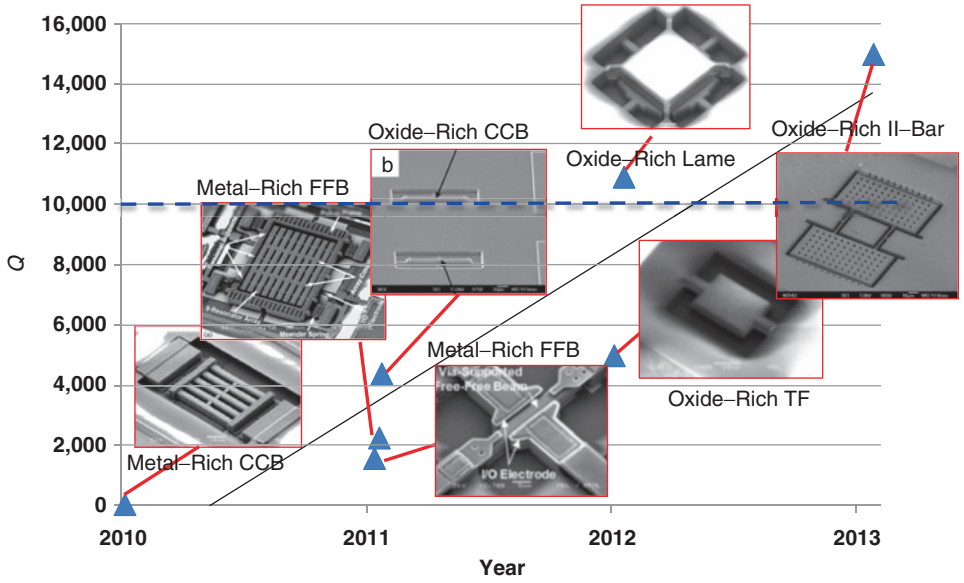


CMOS-MEMS Resonators, Fig. 12 Measured frequency characteristics of free-free beam resonators using the (**a**) 0.35 µm 2P4M and (**b**) 0.18 µm 1P6M CMOS-MEMS platforms



CMOS-MEMS Resonators, Fig. 13 Comparison of a single resonator and a nine-resonator array under the same DC bias and input power

becomes a multi-degree-of-freedom mechanical system, a high-velocity coupling scheme is used to accentuate the desired mode shape and reject the spurious modes. As shown in Fig. 13, the proposed nine-resonator and five-resonator arrays [13] have been experimentally characterized to have 10.1X and 3.7X smaller motional impedances, respectively, as compared to a single resonator. In addition, the resonator array benefits from the large transduction area to effectively reduce the required DC bias voltage while maintaining reasonable motional impedance.



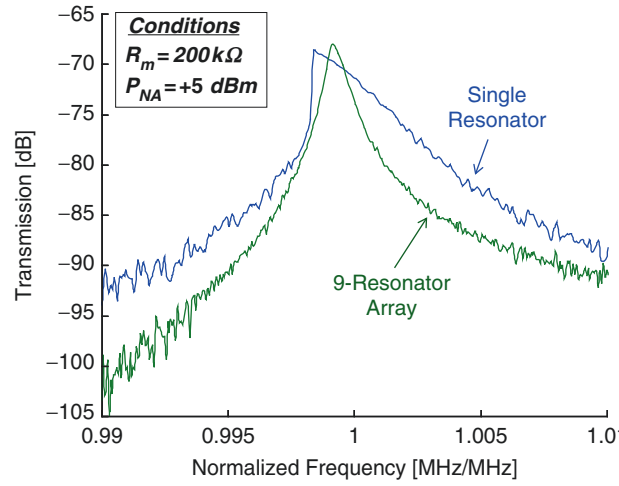
CMOS-MEMS Resonators, Fig. 14 Progress on quality factor of the CMOS-MEMS resonators developed in our group

Quality Factor *Q* Improvement

As compared to the single-crystal or polysilicon-based MEMS resonators, Q of the CMOS-MEMS resonators becomes a main issue to be used in timing reference and spectral processing applications. Due to the lossy nature of the metal materials in the CMOS back-end-of-line (BEOL), CMOS-MEMS metal-rich composite resonators often suffer Q value limited to 1,000. Several approaches are proposed to resolve this bottleneck, including (i) adopting proper mechanical boundary condition and implement nodal support design (i.e., support at the motionless locations of the resonator to avoid acoustic loss), such as the use of a free-free beam resonator with nodal supports instead of a clamped-clamped beam version; (ii) utilizing high-Q structural material in BEOL, such as SiO_2-rich composite structure in Fig. 10, to attain high Q; and (iii) taking advantage of high-Q bulk mode instead of low-Q flexural mode, as illustrated in Fig. 11. Figure 14 finally presents the progress on quality factor of our developed CMOS-MEMS resonators in the past 5 years. As can be seen, Q greater than 10,000 has been realized by the combination of the nodal support design, oxide-rich structural material, and bulk-mode operation, thus greatly benefiting the future oscillator and filter performance.

Power Handling P_o Enhancement

Upon high-input power levels, capacitive resonators often suffer the Duffing nonlinearity which resulted from both capacitive and mechanical nonlinearities, hence degrading output power and phase noise performance of oscillators implemented using those resonators. The mechanically coupled array design [13] described in section "Motional Impedance R_m Reduction" not only reduces their motional impedance R_m but greatly enhances the power handling capability P_o since the effective stiffness of the N-resonator array ideally is N times higher than that of a single resonator, thus contributing N times larger power handling ability. Figure 15 presents the frequency characteristics of single resonator and nine-resonator array designs under the same input power and motional impedance, where the former falls into the nonlinear regime, while the latter still preserves the linear vibration with a symmetrical frequency response and high Q.



CMOS-MEMS Resonators, Fig. 15 Comparison of the single resonator and nine-resonator array under the same motional impedance and input power

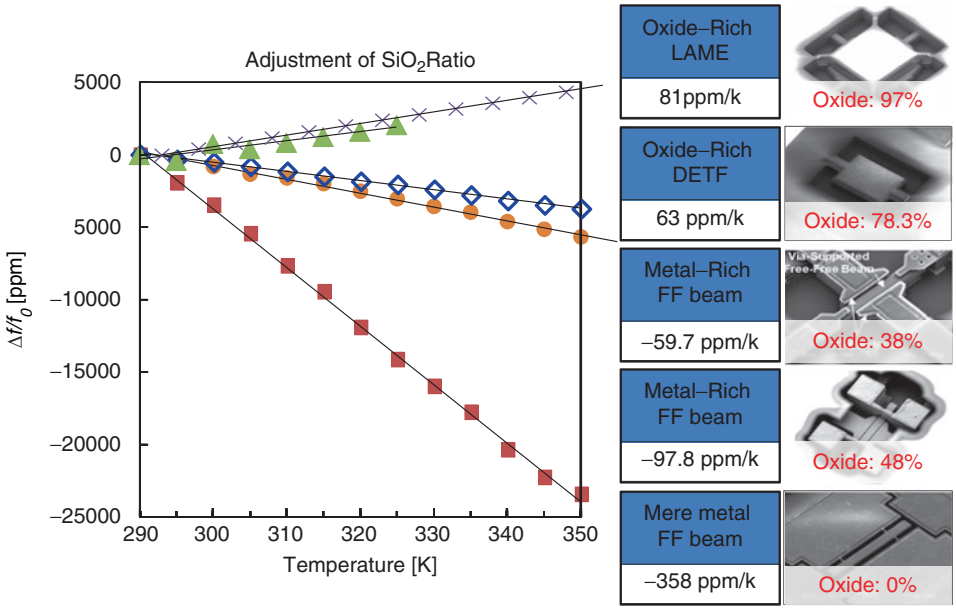
Temperature Compensation

The CMOS-MEMS resonators mostly composed of metals are sensitive to temperature variation due to the very negative temperature coefficients of Young's modulus (TC_Es) of their constituent metal materials. To improve the thermal stability of those CMOS-MEMS resonators, the silicon dioxide (SiO_2) of the CMOS BEOL possessing positive TC_E offers a simple and effective temperature compensation scheme where metal-oxide composite structures can be utilized to build resonators capable of improving their thermal stability [1]. For example, the oxide constituent ratio of the CMOS-MEMS resonators can be adjusted as indicated in Fig. 2 where the mere-metal, metal-rich, and oxide-rich structures would lead to very negative, slightly negative, and very positive temperature coefficients of frequency (TC_fs). Figure 16 presents the comparison of fractional frequency change versus temperature measurements for mere-metal, metal-rich composite, and oxide-rich composite beam resonators, indicating the near-zero TC_f is plausible once the ratio between SiO_2 and metal is optimized, such as manipulation of A and B in Eq. 10 to enable $TC_f \sim 0$. Thus, the use of SiO_2 as part of the composite structures in CMOS-MEMS resonators brings an easy and effective temperature compensation scheme [14].

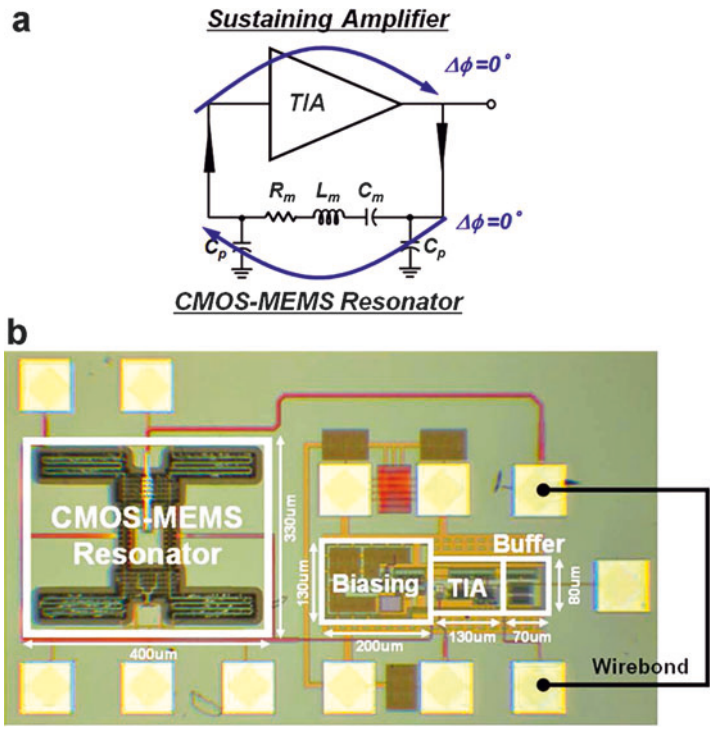
Examples of Application

Oscillator Implementation

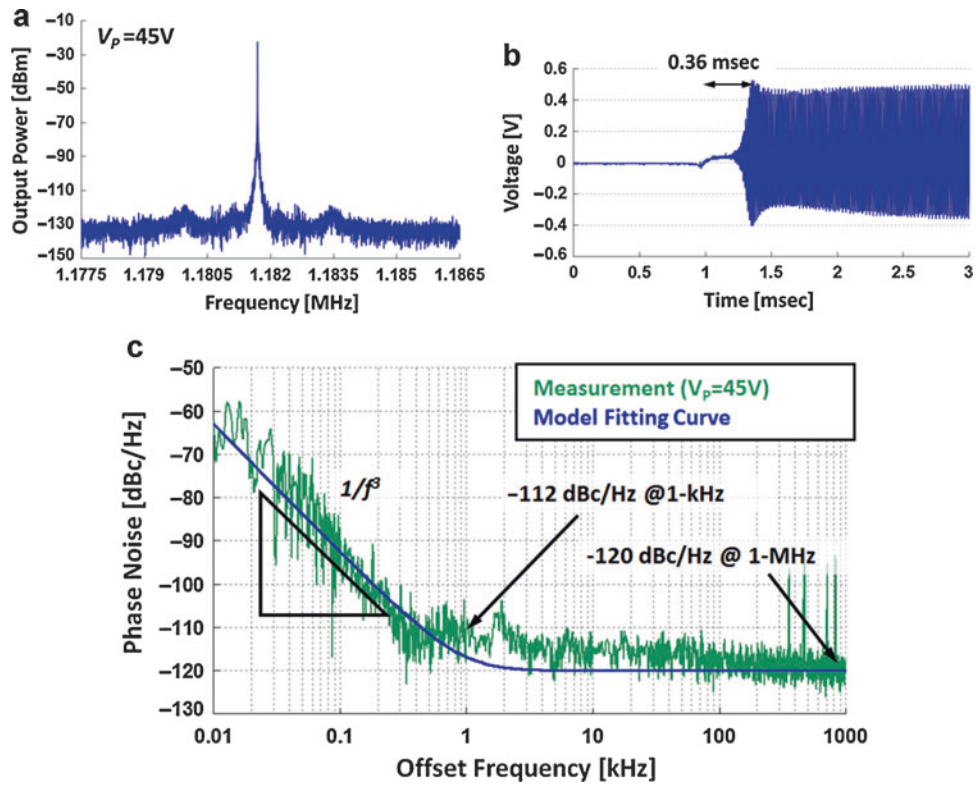
After addressing the major bottlenecks and their corresponding solutions to enhance the overall performance of the CMOS-MEMS resonators, the oscillator implementation is of great importance for timing, frequency synthesizing, and sensing applications. As depicted in Fig. 17a, a micromechanical resonator and its sustaining amplifier can form a closed loop to enable oscillation once the Barkhausen criteria (i.e., loop gain >1 and loop phase $= 0°$) are satisfied. Using a physical implementation shown in Fig. 17b as an example, a single-chip CMOS-MEMS oscillator has been successfully demonstrated in vacuum with the measured frequency-domain output spectrum and time-domain waveform shown in Fig. 18a, b, respectively. The measured phase noise performance in Fig. 18c is comparable to the silicon-based oscillators. Even in air (i.e., with



CMOS-MEMS Resonators, Fig. 16 Temperature coefficients of frequency can be adjusted by the constituent ratio of SiO_2 in CMOS-MEMS resonators



CMOS-MEMS Resonators, Fig. 17 (**a**) Top-level schematic and (**b**) optical view of the monolithic CMOS-MEMS oscillator

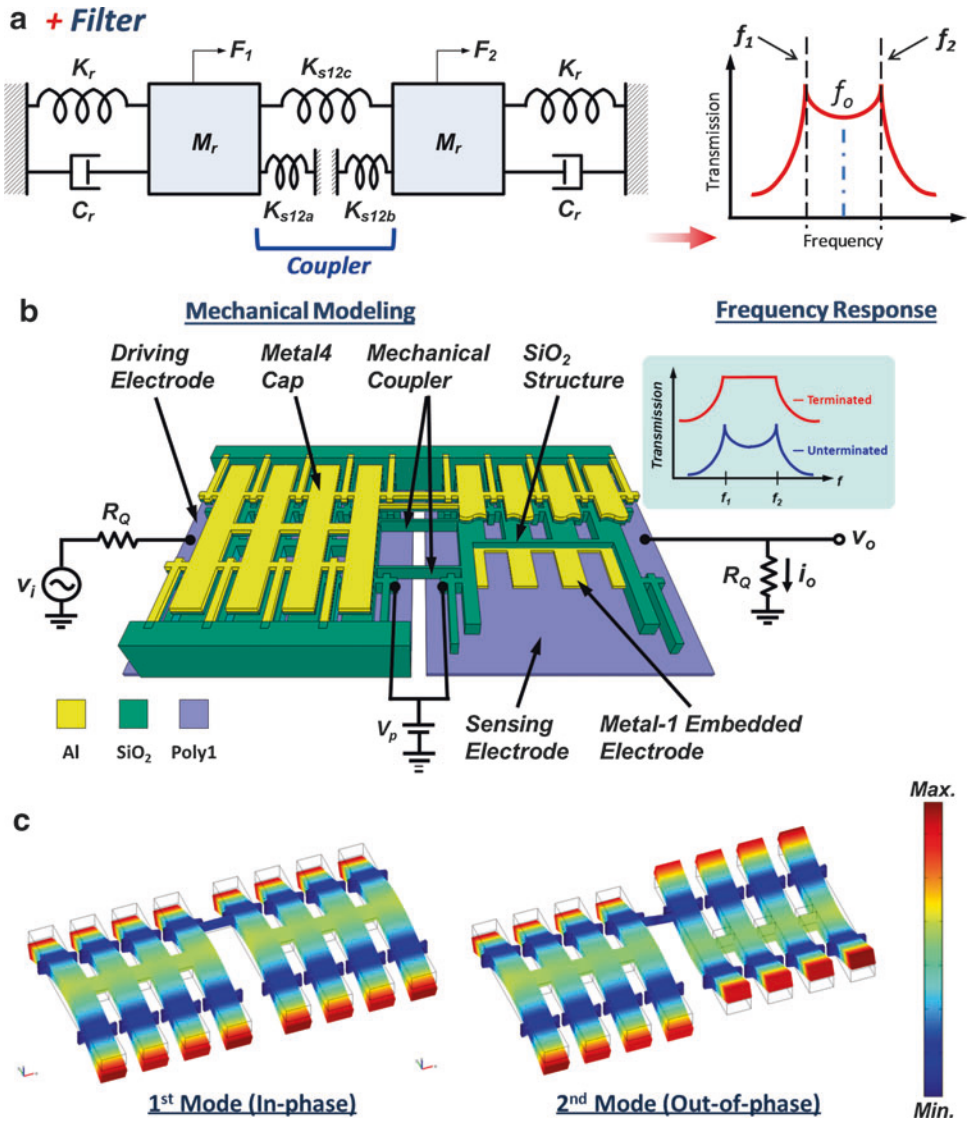


CMOS-MEMS Resonators, Fig. 18 (**a**) Frequency-domain response, (**b**) time-domain waveform, and (**c**) phase noise performance of the CMOS-MEMS oscillator

large squeeze film damping), the oscillator is still functional, indicating great potential to be used in gas, chemical, and mass sensing applications based on the frequency-shift mechanism. The abovementioned approaches to improve quality factor and power handling of the CMOS-MEMS resonators would eventually benefit the close-to-carrier and far-from-carrier phase noise performance, respectively, of the implemented oscillators. In addition, the reduction of the motional impedance described in the previous section would also lead to the overall phase noise reduction for the CMOS-MEMS integrated oscillators.

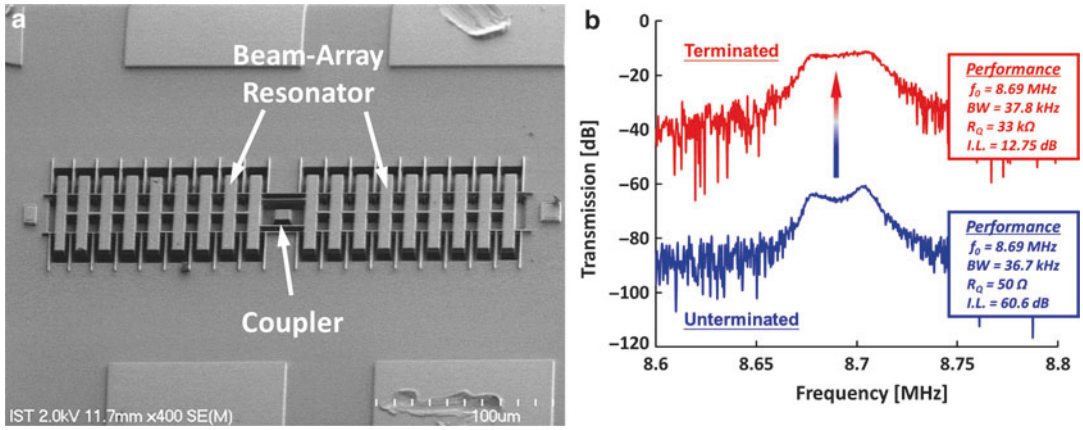
Filter Implementation

Another significant building block for frequency control and wireless communication is the frequency selection element. The developed CMOS-MEMS resonators can serve as the fundamental units for a band-pass filter. Since the resonance frequency of the CMOS-MEMS resonators is limited, it is best to be used in sensor front end with much lower operation frequencies. Figure 19a illustrates the concept of a fourth-order band-pass filter realized by two mechanical resonators, each of which is modeled by a mass-spring-damper system, coupled by a mechanical spring system to form a desired passband of the filter. The center frequency of the filter is determined by the resonance frequency of the constituent resonators, while the bandwidth is mainly set by the spring constant of the mechanical coupler. To simplify the filter design, the constituent resonators should be identical and the coupling element is designed in its quarter-wavelength configuration, allowing the center frequency of the filter to be set at the resonance frequency of



CMOS-MEMS Resonators, Fig. 19 (**a**) Equivalent mechanical circuit model, (**b**) perspective-view schematic, and (**c**) simulated filter modes of a mechanically coupled CMOS-MEMS band-pass filter

the resonators (f_o) while the frequencies (f_1 and f_2) of the two filter modes possess equal distance (i.e., separation) to the filter center frequency (f_o). Figure 19b presents the physical implementation of a CMOS-MEMS filter comprising two CMOS-MEMS resonator arrays mechanically coupled through a quarter-wavelength mechanical link to enable a narrow-bandwidth band-pass filter for channel selection [15]. The in-phase and out-of-phase filter modes shown in Fig. 19c form a passband of the filter. The fabrication process is similar to that of the CMOS-MEMS resonators in the previous section using the metal removal release process in a 0.35 μm CMOS technology [3], and the SEM view of the fabricated filter is shown in Fig. 20a. After filter termination, a flat passband and real insertion loss can be obtained as shown in Fig. 20b.



CMOS-MEMS Resonators, Fig. 20 (**a**) SEM view and (**b**) un-terminated and terminated frequency responses of a mechanically coupled CMOS-MEMS filter

Cross-references

► CMOS-MEMS

References

1. Chen, W.-C., Fang, W., Li, S.-S.: A generalized CMOS-MEMS platform for micromechanical resonators monolithically integrated with circuits. J. Micromech. Microeng. **21**(6), 065012 (2011)
2. Li, C.-S., Hou, L.-J., Li, S.-S.: Advanced CMOS-MEMS resonator platform. IEEE Electron Device Lett. **33**(2), 272–274 (2012)
3. Liu, Y.-C., Tsai, M.-H., Chen, W.-C., Li, M.-H., Li, S.-S., Fang, W.: Temperature-compensated CMOS-MEMS oxide resonators. IEEE/ASME J. Microelectromech. Syst. **22**(5), 1054–1065 (2013)
4. Chen, W.-C., Li, M.-H., Liu, Y.-C., Fang, W., Li, S.-S.: A fully-differential CMOS-MEMS DETF oxide resonator with Q > 4,800 and positive TCF. IEEE Electron Device Lett. **33**(5), 721–723 (2012)
5. Li, C.-S., Li, M.-H., Chin, C.-H., Li, S.-S.: Differentially piezoresistive sensing for CMOS-MEMS resonators. IEEE/ASME J. Microelectromech. Syst. **22**(6), 1361–1372 (2013)
6. Melamud, R., Chandorkar, S.A., Kim, B., Lee, H.K., Salvia, J.C., Bahl, G., Hopcroft, M.A., Kenny, T.W.: Temperature-insensitive composite micromechanical resonators. IEEE/ASME J. Microelectromech. Syst. **18**(6), 1409–1419 (2009)
7. Fedder, G.K., Santhanam, S., Reed, M.L., Eagle, S.C., Guillou, D.F., Lu, M.S.-C., Carley, L.R.: Laminated high-aspect-ratio microstructures in a conventional CMOS process. Sensors Actuators A **57**, 103–110 (1996)
8. Lo, C.-C., Chen, F., Fedder, G.K.: Integrated HF CMOS-MEMS square-frame resonators with on-chip electronics and electrothermal narrow gap mechanism. In: Technical Digest, Transducers'05, pp. 2074–2077. Seoul (2005)
9. Uranga, A., Teva, J., Verd, J., Lopez, J.L., Torres, F., Esteve, J., Abadal, G., Perez-Murano, F., Barniol, N.: Fully CMOS integrated low voltage 100 MHz MEMS resonator. IEEE Electron. Lett. **41**(24), 1327–1328 (2005)
10. Verd, J., Uranga, A., Teva, J., Lopez, J.L., Torres, F., Esteve, J., Abadal, G., Perez-Murano, F., Barniol, N.: Integrated CMOS-MEMS with on chip read-out electronics for high frequency applications. IEEE Electron Device Lett. **27**(6), 495–497 (2006)
11. Teva, J., Abadal, G., Uranga, A., Verd, J., Torres, F., Lopez, J.L., Esteve, J., Pérez-Murano, F., Barniol, N.: From VHF to UHF CMOS-MEMS monolithically integrated resonators. In: Technical Digest, 21st IEEE International Conference on Micro Electro Mechanical Systems (MEMS'08), pp. 82–85. Tucson. (2008)
12. Pourkamali, S., Hao, Z., Ayazi, F.: VHF single crystal silicon elliptic bulk-mode capacitive disk resonators, part II: implementation and characterization. IEEE/ASME J. Microelectromech. Syst. **13**(6), 1054–1062 (2004)
13. Li, M.-H., Chen, W.-C., Li, S.-S.: Mechanically-coupled CMOS-MEMS free-free beam resonator arrays with enhanced power handling capability. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **59**(3), 346–357 (2012)
14. Chen, W.-C., Fang, W., Li, S.-S.: High-Q integrated CMOS-MEMS resonators with deep-submicron gaps and quasi-linear frequency tuning. IEEE/ASME J. Microelectromech. Syst. **21**(3), 688–701 (2012)
15. Chen, C.-Y., Li, M.-H., Li, C.-S., Li, S.-S.: Design and characterization of mechanically-coupled CMOS-MEMS filters for channel-select applications. Sensors Actuators A Phys. **216**, 394–404 (2014)

CNT assembly

▶ CNT Handling and Integration

CNT Biosensor

▶ Nanostructure Field Effect Transistor Biosensors

CNT Handling and Integration

Zhan Yang
Jiangsu Key Laboratory for Advanced Robotics Technologies and Collaborative Innovation Center of Suzhou, Nano Science and Technology, Soochow University, 215021, Suzhou, China

Synonyms

CNT assembly; Nano Manipulation; Nanorobotics

Definition

Manipulation matter at nanoscale is one of the main purposes of nanotechnology. In with the publication of the pioneering work, "Engines of Creation" in 1986, atom-by-atom assembly became less a dream and more a search for a practical technology to implement it. Researchers were getting more adept at using Atomic Force and Scanning Tunneling Microscopes (STMs), culminating with the now famous writing of the IBM logo in 1989 with 35 xenon atoms. The team, led by Andreas Heinrich of IBM Research-Almaden (California), has successfully stored one magnetic bit of data with just 12 atoms of iron and a full byte of data in 96 atoms. This represents a storage density that is at least 100 times denser than the largest hard drive platters or flash memory chips in 2012.

In the recent two decades, nanorobotic manipulation was developing rapidly. Nanorobotic manipulation was applied to atomic force microscope, scanning electron microscope, transmission electron microscope towards to nanomanufacture nanoassembly and nanofabrication.

The AFM consists of a cantilever with a sharp tip (probe) at its end that is used to scan the specimen surface. The cantilever is typically silicon or silicon nitride with a tip radius of curvature on the order of nanometers. When the tip is brought into proximity of a sample surface, forces between the tip and the sample lead to a deflection of the cantilever according to Hooke's law. Depending on the situation, forces that are measured in AFM include mechanical contact force, van der Waals forces, capillary forces, chemical bonding, electrostatic forces, magnetic forces (see magnetic force microscope, MFM), Casimir forces, solvation forces, etc. Along with the force, additional quantities may simultaneously be measured through the use of specialized types of probes (see scanning thermal microscopy, scanning Joule expansion microscopy, photothermal microspectroscopy, etc.). Typically, the deflection is measured using a laser spot reflected from the top surface of the cantilever into an array of photodiodes. Other methods that are used include optical interferometry, capacitive sensing, or piezoresistive AFM cantilevers. These cantilevers are fabricated with piezoresistive elements that act as a strain gauge. Using a Wheatstone bridge, strain in the AFM cantilever due to deflection can be measured, but this method is not as sensitive as laser deflection or interferometry. Prof. N. Xi's team developed a real-time, force feedback nanomanipulation system.

Scanning electron microscope is a real-time nanoscale observation equipment. With the benefit of big specimen chamber, a complex nanomanipulation system with multiple-DOFs could be constructed. Recently, the environmental scanning electron microscope (ESEM) with nanorobotic manipulation was employed to characterize the biological property of cells.

Prof. Fukuda's team is a leader of nanorobotic manipulation in FESEM from the beginning of

this century. The carbon nanotube was picked-up, assembled, measured, and fabricated by multiple DOFs nanomanipulator.

Dr. M. Rizuan and Dr. Y. Shen constructed a nanomanipulation system in order to characterize the biological property of cells with low pressure and different humidity. Two types of experiments have been conducted, i.e., mechanical properties of individual yeast cells and forces on cell-substrate area. Different kinds of spring constants (0.02 N/m, 0.09 N/m, and 0.7 N/m) and tip's shapes (sharp, flat, and needle-like) of the cantilevers have been used during the experiments. From the analysis, the compressed forces needed to penetrate the cell wall of the yeast cells using sharp and flat tips (0.02 N/m for both) which are in the range of 87–278 nN and 57–207 nN, respectively. Locality mechanical measurement has been performed using needle-like tip (0.09 N/m) on single-cell and mother-daughter cells where the elastic properties of the cells are in the range of 1.32–3.95 MPa.

Nanorobotic manipulation is a rapid developing technology on nanofabrication and nanoassembly. It is powerful with nanomaterial measurement and manipulation. With the control theory, the automation is employed to the nanorobotic manipulation system. A high efficacy nanorobotic manipulation center will be applied to nanomaterial, nanosensor, and biological samples.

Introduction

This entry is focused on the study of CNT interaction with environment and the application of CNT for pH and temperature sensing.

CNT Handling

Carbon nanotubes (CNTs) [1] have been widely used as nanobuilding blocks to assemble nanostructures and nanodevices. Therefore, the interaction between a CNT and an environment is important. In the present study, the interaction forces, which consisted mainly of van der Waals [2] forces, between a single multiwalled carbon nanotube tip and a gold surface were evaluated under a scanning electron microscope by means of nanorobotic manipulation. The influence of electron beam irradiation was also investigated. It was found that electron beam irradiation can increase the interaction forces by producing contamination at the contact area. The van der Waals forces were calculated theoretically, and the theoretical values were close to the experimental values. In the following, the measurement of van der Waals force between a CNT tip and a gold surface is introduced first, and then a method of manipulating CNTs in 3D is presented.

Measurement of van der Waals Force Between a CNT Tip and Gold Surface

At the nanoscale, van der Waals force is one of the predominant surface forces that also include the electrostatic force, the Casimir force, etc. The measurement of nanoscale surface forces can help characterize the properties of nanomaterials and also be useful for the assembly of nanostructures and nanodevices. Carbon nanotubes (CNTs) as typical nanomaterials have been used for atomic force microscope (AFM) probes, nanotweezers, nanoposition sensors, and three-dimensional nanostructures. The interaction between a CNT and a substrate, including the van der Waals forces, is important for the assembly of these mechatronic applications. The deformation of a CNT placed on a substrate by surface van der Waals forces will significantly modify the idealized geometry of the free nanotube. The interaction has been investigated by AFM and molecular-mechanics simulations. The joint interaction forces with two CNT tips by chemical bonds were also investigated. Cumings and Zettl calculated the attractive van der Waals forces between the inner and outer layer. They demonstrated a lowfrictional nanoscale linear bearing realized from multiwalled carbon nanotubes (MWNTs). Zheng et al. showed through theoretical calculations that the extruded core of a MWNT oscillating on the outer shell caused the excess van der Waals interaction energy and

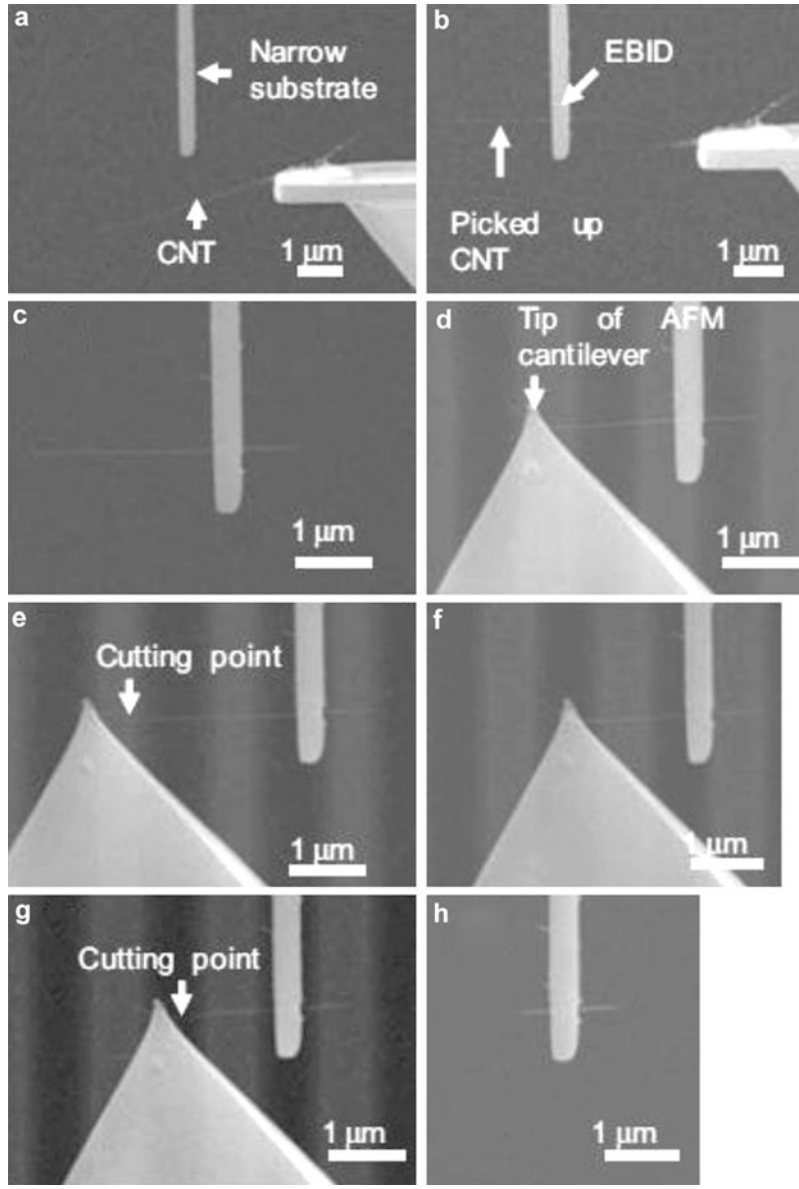
found that the oscillation frequency can be significantly higher than 1 GHz [3]. The effects of surface forces in a three-terminal nanorelay were investigated by Jonsson et al. [4]. They showed that van der Waals forces have a significant impact on the characteristics of the relay. For the calculation or analysis of van der Waals forces, a parameter called the Hamaker constant is an important value.

In previous studies, the Hamaker constant between a CNT and a surface was obtained by approximate calculation [5] or by original experiments [6]. Akita et al. obtained the Hamaker constant of 60×10^{-20} J for the sidewall of a CNT adhered on a silicon material surface [6]. For a CNT and a gold surface, Walkeajärvi et al. just used a typical value of 10×10^{-20} J for their calculation. For two CNTs, a previous report used the value of graphite to replace it. The van der Waals forces between a CNT tip and a gold surface are not clearly understood experimentally and need to be investigated. In particular, the interaction force is considered to be changed by irradiation of the electron beam during the observation in an electron microscope, because electron beam irradiation produces contamination on the observation area.

From an engineering viewpoint, it is important to control the interaction forces between CNT probes and metallic surfaces for nanodevice fabrication and assembly. This entry reports on the measurement of the van der Waals forces between a CNT tip and a gold surface under a scanning electron microscope (SEM) based on nanorobotic manipulation. The influence of electron beam irradiation was also investigated. The van der Waals forces were calculated theoretically and compared with the experimental values.

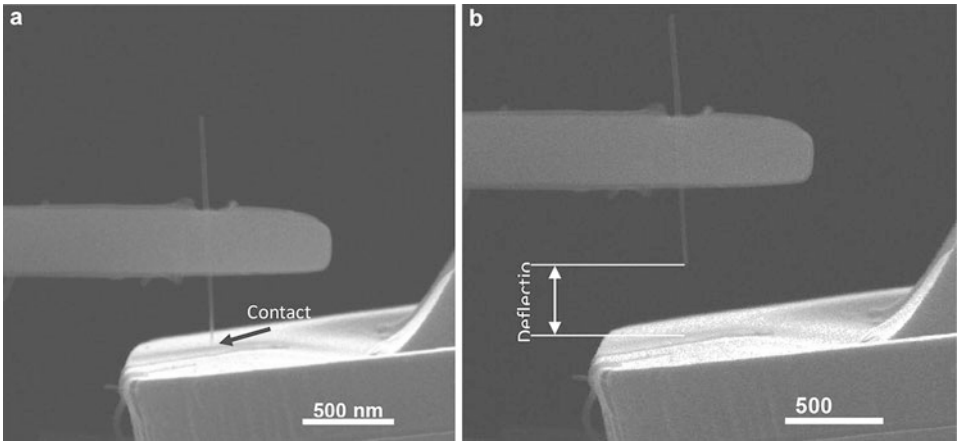
For the measurement of the interaction forces between the CNT tip and the surface of a gold substrate, a single MWNT probe was first prepared as shown in Fig. 1. The MWNT, typically 20–50 nm in diameter, was synthesized by the standard arc-discharge method. Figure 1a shows a single MWNT with diameter of 30 nm that was picked up from the CNT bundle by a nanorobotic manipulator under a SEM (JEOL, JSM-6500 F). The MWNT was fixed at the narrow substrate with electron beam–induced deposition (EBID) as shown in Fig. 1b. The deposition is caused by the dissociation of molecules adsorbed to a surface by high-energy electrons. In the present experiments, tungsten hexacarbonyl was used as precursor by filling into a glass tube and introducing it into the vicinity of the sample. The magnified image of the narrow substrate with MWNT is shown in Fig. 1c. It was etched to a width of 350 nm by a focused ion beam (FIB) process. The MWNT probe should be prepared with no impurities and straight with short length, because the bending phenomenon occurs easily and makes it difficult to measure the vertical interaction forces. Some methods such as electron beam–induced fabrication assisted with oxygen gas or current-induced fabrication can be used to adjust the CNT length. Here, the current-induced fabrication process was used. One end of the CNT was touched at the tip of the AFM cantilever as shown in Fig. 1d. The part of the CNT near the cantilever tip evaporated by joule heating when the electric current applied was 7–20 mA. The length of the MWNT probe was adjusted by using the same process as shown in Fig. 1d, e by the first process. Figure 1f, g show the second process. Finally, the MWNT probe is fabricated as shown in Fig. 1h.

In the present work, the interaction force measurement between the CNT probe and the gold surface of an AFM cantilever was performed experimentally. The AFM cantilever was coated by a gold layer with thickness of 40 nm. The interaction forces were measured from the deflection of the AFM cantilever. The CNT probe and AFM cantilever were connected to the ground to avoid the occurrence of electrostatic forces. The SEM was kept at low magnification (less than 10,000 times) during the operation to avoid contamination. Other type forces caused by chemical bonding can be negligible, because the CNT probe and the surface were kept clean under vacuum conditions. Therefore, the interaction force between the CNT tip and the surface of the AFM cantilever is considered to consist mainly of van der Waals forces. In the experiment, the AFM cantilever tip was moved to the

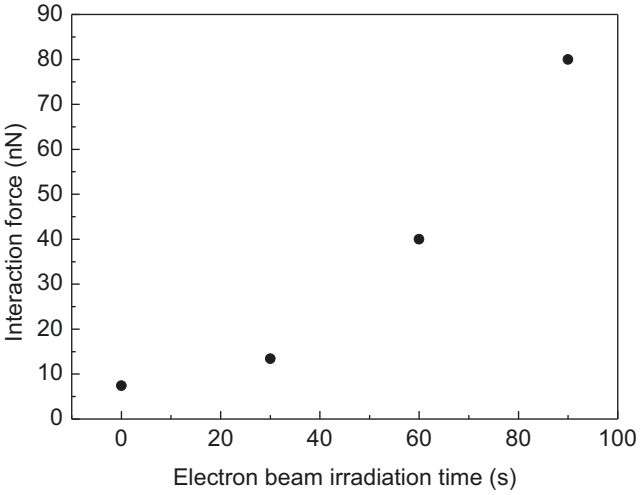


CNT Handling and Integration, Fig. 1 A multiwalled carbon nanotube was placed on an etched substrate. The length of the CNT was adjusted by the current-induced fabrication process

CNT tip by a nanorobotic manipulator, until they touched each other (Fig. 2a). Then the CNT probe was moved in the inverted direction until release of the adhesion by van der Waals forces. The position of the AFM cantilever was controlled in step motions, each step of moving was about 2 nm. When the CNT probe was released from the AFM cantilever surface, the AFM cantilever was stopped in a certain distance. The AFM cantilever was released where the attraction is less than the tension caused by the deformation of the cantilever. The deflection of the AFM cantilever was 370 nm in the present experiment (Fig. 2b). From this, the

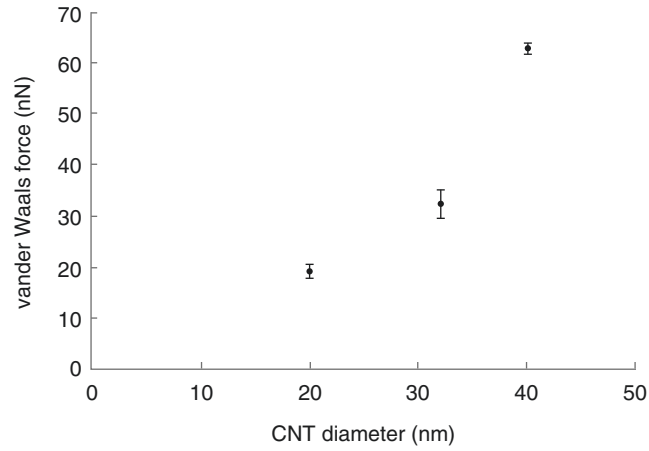


CNT Handling and Integration, Fig. 2 The tip of a single CNT was adhered to an AFM cantilever surface. After that, the AFM cantilever was pulled back by a nanomanipulator. The interaction force between the carbon nanotube tip and the surface can be obtained from the deflection of cantilever



CNT Handling and Integration, Fig. 3 The relationship between the interaction force and electron beam irradiation time

interaction forces from the deflection of the AFM cantilever were calculated according to $F = \delta d$, where d is the deflection and is the elastic constant of the AFM cantilever, is 0.02 N/m (Olympus Corp., OMCL-TR400). Therefore, the interaction force between the CNT tip and the gold surface was calculated as 7.4 nN. The interaction force measurement was also performed under electron beam irradiation to affect the influence of interaction forces. The interaction forces were measured after irradiation for different times. After the above experiment to measure the van der Waals interaction forces, the contact area, where the CNT tip was adhered on the gold surface, was magnified to 30,000 times. The acceleration voltage of the SEM was 15 kV and the beam current was 30 pA. The irradiation periods were set at 30, 60, and 90 s. The interaction forces were obtained as 13.4, 40, and 80 nN, respectively, as shown in Fig. 3. The results show that electron beam



CNT Handling and Integration, Fig. 4 Three teams experiment results with error bar of van der Waals force

irradiation can increase the interaction forces between the CNT tip and the surface. The increment of interaction force is almost proportional to the electron beam irradiation time period.

The increase of interaction force induced by electron beam irradiation is considered to result from the contamination produced at the contact area. In the experiment, the precursors come from the vacuum pump oils in the sample chamber; hence the contamination should be mainly amorphous carbon. This suggests that electron beam irradiation can be used to control the interaction forces in the nanonewton range for the assembly of CNT nanobuilding blocks.

We assume that the interaction between CNT tip and the surface is nonretarded and additive. The interaction energy can be described by Lennard-Jones's pair potential as the following equation

$$W = \frac{r^2\pi^2 C\rho_1\rho_2}{12D^2} = \frac{r^2 A}{12D^2} \tag{1}$$

where r is the radius of CNT, and C is the coefficient in the atom-atom pair potential. ρ_1 and ρ_2 are the numbers of atoms per unit volume in the two materials, and D the distance between the CNT tip and the surface, as shown in Fig. 5. A is the Hamaker constant.

Then the van der Waals forces are obtained by taking the derivative of the pair potential with respect to the distance,

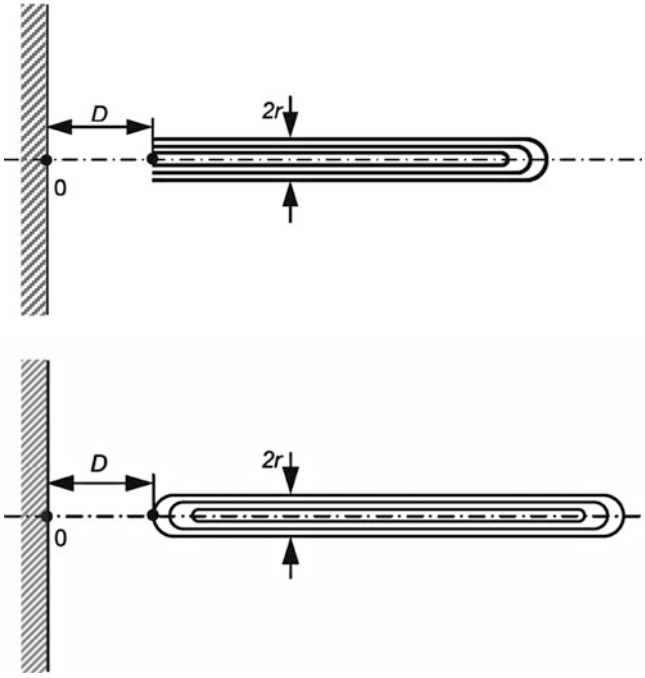
$$F = W' = \frac{r^2 A}{6D^3} \tag{2}$$

To calculate F, the Hamarker constant $A_{\mathrm{CNT}-Au}$ between the CNT and the gold surface interacting across vacuum is needed. The unknown $A_{\mathrm{CNT}-Au}$ can be obtained approximately by using the combining relation as the nonretarded Hamaker constant for carbon nanotube and gold surface [3], which is given by

$$A_{\mathrm{CNT}-Au} = \sqrt{A_{\mathrm{CNT}} A_{Au}}\,, \tag{3}$$

where A_{Au} and A_{CNT} are $5.47 \times 10^{-19} J$ [2] and $0.284 \times 10^{-19} J$ [5] from previous report, respectively. Thus, the $A_{\mathrm{CNT}-Au}$ is obtained as $1.246 \times 10^{-19} J$ from Eq. 3.

If the CNT is considered as a column, the van der Waals forces between the CNT tip and the gold surface are calculated as 77, 197, and 308 nN at the CNT diameters 20, 32, and 40 nm as shown in Fig. 4. It is simply assumed that the minimum distance D is about 0.3 nm, and the r is used as the radius of the CNT in Fig. 5 The calculated result is so much larger than the value of experimental result. Therefore, this mode which considers the CNT as a column cannot precisely get the van der Waals forces.



CNT Handling and Integration, Fig. 5 Schematic drawing of a CNT adhered to a surface with a tiny separation D. The CNT is $2r$ in diameter with open cap and closed cap

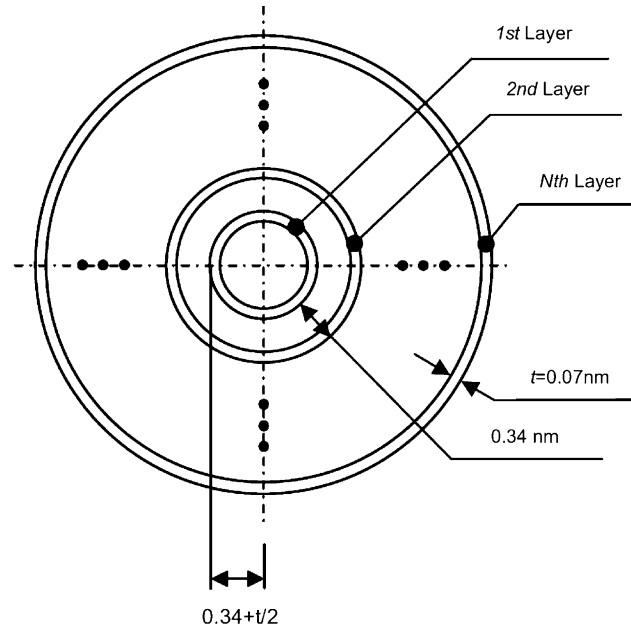
We suggest that the model of CNT can be considered as a multilayered cylinder as shown in Fig. 6. The cross-sectional area of the CNT is less than that calculated with the column model. Each layered cross-sectional area is described by following equations,

$$\begin{aligned} A_1 &= \pi(0.34 + t/2)^2 - \pi(0.34 - t/2)^2 \\ A_2 &= \pi(0.34 + t/2 + 0.34)^2 - \pi(0.34 - t/2 + 0.34)^2 \\ &\vdots \\ A_N &= \pi[0.34 + t/2 + 0.34(N-1)]^2 - \pi[0.34 - t/2 + 0.34(N-1)]^2 \end{aligned} \tag{4}$$

And the total cross-sectional area is,

$$A = \sum_{n=0}^{n=N-1} \left[\pi(0.34 + t/2 + 0.34n)^2 - \pi(0.34 - t/2 + 0.34n)^2\right] \tag{5}$$

Here, t is 0.07 nm, the diameter of carbon molecular. Then the total area is calculated as 148 nm^2. The distance between two layers is 0.34 nm that has clearly observed by a TEM [7]. N is the number of layers. The relationship between the number of layers and the radius of CNT can be approximated by $0.34N \approx r$ in our case. We obtained that the layer's number of the CNT is about 29, 47, and 59 at the CNT diameters at 20, 32, and 40 nm. Finally, the van der Waals forces were obtained theoretically as 15.9, 41.3, and 64.8 nN at the CNT diameters at 20, 32, and 40 nm from the Eq. 2. This result is very close to the experimental data. It demonstrates that the cylinder model is good or accurate enough for calculation of van der Waals force between CNTs and the surface. For more precise calculation, the CNT should be considered as a unit of carbon atoms. In that case, the sum of the forces between all individual atoms and the substrate needs to be calculated. Therefore, the shapes of CNT tip should be also taken into more careful consideration. Those are our future works.



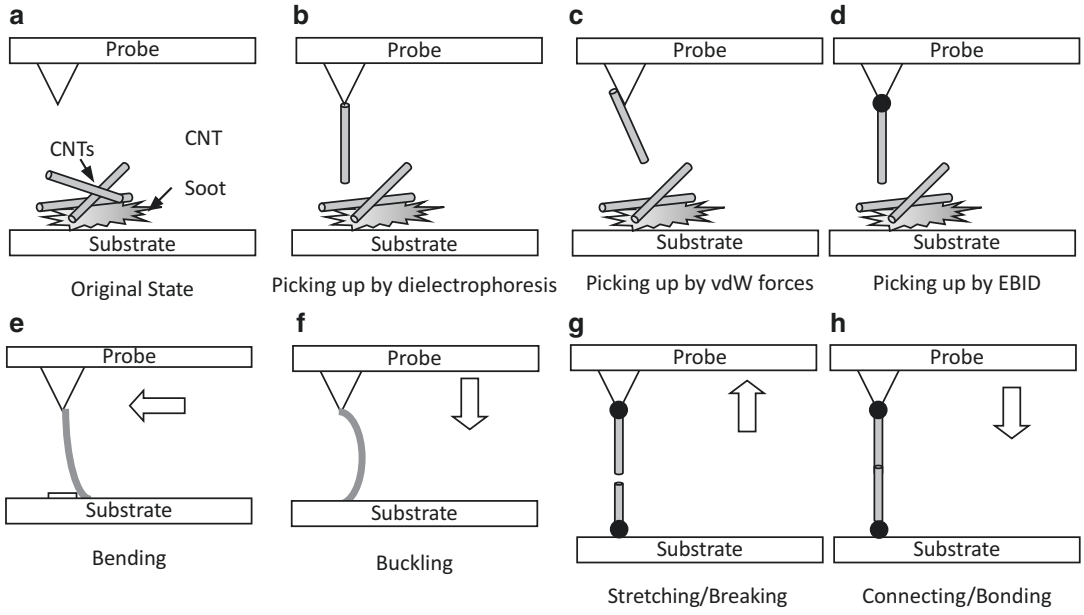
CNT Handling and Integration, Fig. 6 Schematic drawing of a CNT described by multilayered cylinder model. The distance between layers of CNT is 0.34 nm. The t is the diameter of carbon molecule with 0.07 nm

In summary, a single MWNT was actually prepared with the nanorobotic manipulation, electron beam–induced deposition, and current-induced fabrication. The interaction forces between the MWNT tip and the gold surface were investigated and measured with an AFM cantilever, which were mainly consisted of van der Waals forces. The method proposed is easy to operate and have a large measurement range. It can be used for evaluation of connection strength during assembling of nanodevices. We also found that the electron beam irradiation can increase the interaction forces. The increase of interaction force induced by electron beam irradiation might be produced by the contamination. We suggest that the electron beam irradiation should be used to adjust the interaction forces in nanonewton scale during the assembly of CNT.

For calculating the van der Waals forces, the CNT was described by a multilayered cylinder model. The result of the experiment is very close to theoretical calculation. That model can be used to estimate the nanometer order van der Waals forces.

3D CNT Handling

Manipulations of CNTs in 3D space are very important techniques for assembling them into structures and devices. The basic step for this is to pick up a single tube from nanotube soot (Fig. 7a). This has been shown first by using dielectrophoresis through nanorobotic manipulations (Fig. 7b). The interaction between a tube and the atomic flat surface of AFM cantilever tip has been shown to be strong enough for picking up a tube onto the tip (Fig. 7c). By using electron beam–induced deposition (EBID), it is possible to pick up a nanotube onto a probe (Fig. 7d). For handling a tube, weak connection between the tube and the probe is desired. Bending and buckling a CNT as shown in Fig. 7e, f are important for the in-situ property characterizations of a nanotube, which is a simple way to get the Young's modulus of a nanotube without damaging the tube if performed in elastic range and hence can be used for the selection of a tube with desired properties. Plastic bending or buckling can generate intramolecular kink junctions of CNTs. Combined bending and buckling with shape fixing with EBID can be used for the shape modifications of a nanotube. Stretching a nanotube between two probes or a probe and a substrate has brought out several interesting results. The first demonstration of 3D nanomanipulations of nanotubes took this one as an example to show



CNT Handling and Integration, Fig. 7 3D manipulations of CNTs. Basic technique for such manipulations is to pick up an individual tube from CNT soot (**a**) or oriented array. (**b**) shows a free-standing nanotube is picked up by dielectrophoresis generated by nonuniform electric field between the probe and substrate, (**c**) and (**d**) show the same manipulation by contacting a tube with the probe surface or fixing (e.g., with EBID) a tube to the tip. Vertical manipulations of nanotubes also include bending (**e**), buckling (**f**), stretching/breaking (**g**), and connecting/bonding (**h**). This family is open for new strategies

the breaking mechanism of a MWNT and the tensile strength of CNTs. By breaking a MWNT in a controlled manner, some interesting nanodevices have been fabricated. This technique – destructive fabrication – has been presented to get sharpened and layered structures of nanotubes and to improve the length control of a nanotube. Bearing motion has also been observed in an incompletely broken MWNT. The reverse process, namely the connection of broken tubes, has been demonstrated recently, and the mechanism is revealed as rebonding of unclosed dangling bonds at the ends of a broken tube. Based on this interesting phenomenon, mechanochemical nanorobotic manipulations have been presented.

3D nanorobotic manipulations have opened a new route for the assembly of nanotubes into nanodevices.

In the following, we will focus on our recent advances in this field.

CNT Integration

A CNT integration method by mechanical engineering is nanorobotic manipulation [8–10] which has advantages such as real-time controllability, nanometer lever positioning resolution, and three-dimensional fabrications. The automation nanofabrication has a potential for auto nanostructure assembly fabrication and measurement. Nanomanipulation allows for the detection and manipulation of tiny entities such as single molecules, nanotubes, cells, viruses, proteins, and DNA molecules.

Fabrication of a pH Sensor Nanoprobe with Tungsten Oxide and Platinum Nanowires Based on Nanorobotic Manipulation

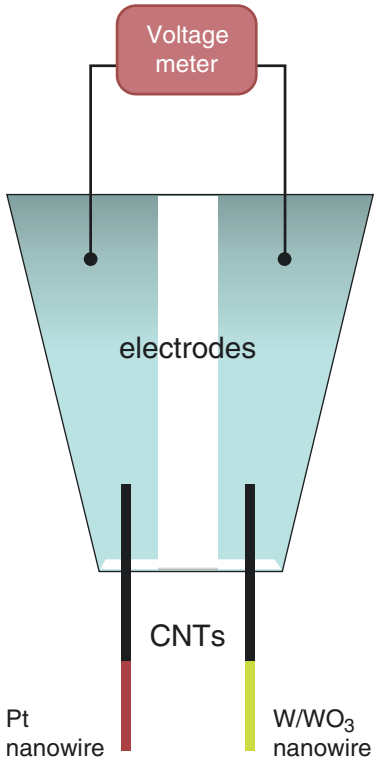
A novel solid-state pH sensor nanoprobe based on a tungsten oxide nanowire work electrode and a platinum nanowire counter electrode is reported. A cantilever is used for the electrodes

of the device and etched by a focused ion beam (FIB). Two multiple-walled carbon nanotubes (MWCNTs) are assembled to the separated electrodes of the cantilever. A tungsten probe is etched by an FIB to 300 nm in diameter and 25.4 μm in length. Then the probe is coated by parylene and the tip cut to expose the tungsten. A tungsten nanowire is used for the work electrode, and a platinum nanowire is used for the counter electrode. The nanowires are fabricated via the field emission method from carbon nanotube tips by introducing hexacarbonyltungsten ($W(CO)_6$) and trimethylcyclopentadienylplatinum ($CpPtMe_3$), respectively, inside a field emission electron microscope. The nanoprobe pH sensor is tested in acidic (pH 4.0, pH 6.1, and pH 6.a9) and alkaline (pH 8.0) buffer solutions and different responses are found. The sensitivity of the pH sensor nanoprobe was measured as −35.04 mv/pH.

Chemical detection of single cells and their local environment has attracted much interest recently. The pH value is a promising parameter to describe the living condition of single cell. For examples in animal cells, pH controls the activity of key enzymes, protein synthesis, DNA and RNA synthesis, cell cycle, and ion conductance by membranes. The extracellular fluid pH is usually $7.35 \sim 7.45$, and in vascular smooth muscle cell (VSMC) the pH value is usually $7.0 \sim 7.2$. In some pathological conditions, such as hypoxia, acidosis, or alkalosis, pH can be fluctuated between 6.5 and 8.1, of VSMC pH corresponding generate phase when large fluctuations. And the oxygen changes primer pH to vascular effects of very complex, it may be by changes in intracellular Ca ion release and re-uptake, the activity of calcium channels on the plasma membrane or to adjust the sensitivity of the calcium ions in the contraction protein vasoconstriction state. The pH sensors towards biological cell require a resolution of 1 pH. The measurement of the pH value of a single cell is usually performed by visualizing the color change of indicators, chemical probes, H^+ sensitive microelectrodes, and nuclear magnetic resonance (NMR) spectrometry.

Traditionally, the measuring methods for pH values fall roughly into four categories: indicator reagents, pH test strips, metal electrode methods (hydrogen electrode, quinhydrone electrode, and antimony electrode method), and glass electrode methods. In recent decades, new techniques of pH detection have been developed such as optical-fiber-based pH sensors [11], mass-sensitive pH sensors [12], metal oxide pH sensors [13], conducting polymer pH sensors [14], nanoconstructed cantilever-based pH sensors [15], ISFET-based pH sensors [16], and pH-image sensors [17].

To detect the pH distribution for cytoplasm and understand the relation between pH and enzymes and protein in detail within a single cell, which has a scale of several to several tens of micrometers, many nanoscale pH sensors have been developed. Zhang et al. reported a solid-state pH sensor based on WO_3-modified vertically aligned MWCNTs, which use a WO_3-CNT array for the pH detection electrode. This sensor has sensitivity close to −40 mV/pH [18]. Fenster et al. reported a single tungsten nanowire pH- sensitive electrode, which has a sensitivity of −61.1 mV/pH [19]. The previous two pH sensors contain nanoscale tungsten pH sensitive electrodes and macroscale stranded hydrogen electrodes which are impossible to apply to a single cell. We have designed double nanometal electrodes which can avoid the size problem in cell pH detection (Fig. 8). We aimed to fabricate the pH sensor by using electrodes by which sensing of the proton concentration becomes possible if either the electrochemical equilibrium involves protons in the reaction or if a separation of the electrolyte is achieved by a proton-specific ionophore. The probe was fabricated by the platinum and tungsten oxide nanowire electrodes grown via field emission technique. Because of the scale of the nanowires, nanoscale resolution positioning and manipulation was required. To this end a nanorobotic manipulation system was employed. Figure 8 shows the concept of the proposed device. In this research, a platinum and tungsten dual nanowire electrode pH sensor has been fabricated. A nanoprobe pH sensor evaluation



CNT Handling and Integration, Fig. 8 Concept draws of nanowire probe pH sensor

system has been constructed, and the pH response of this sensor has been tested in acidic buffer solution (pH = 4, 6.1, 6.9) and alkaline buffer solution (pH = 8).

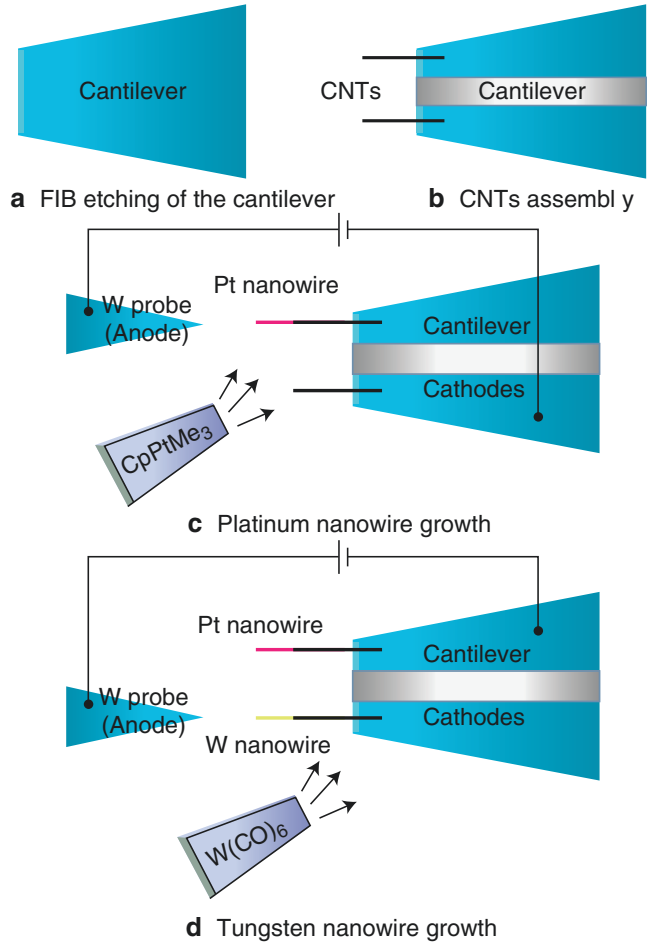
During the field emission nanowire growth, an MWCNT is employed for the cathode and a tungsten probe for the anode. Unit 1 and unit 4 were used to assemble the MWCNT emitters and set the tungsten probes as the anode.

The tungsten probe is etched by an FIB and then coated with parylene. Then the tip of the tungsten probe is opened using the FIB. The tungsten was set to manipulator 4 for anode. The carbon nanotubes were picked up by an AFM cantilever. An AFM cantilever (Olympus, OMCL-TR400PB-1) is set on the manipulator as the end-effector to pick up the MWCNT from the carbon bulk of MWCNTs which were synthesized by the arc-discharge method. The MWCNT is set up as the emitter (cathode) which has a low field emission voltage.

The fabrication procedure is shown in Fig. 9. A PZT cantilever was used as the substrate. The cantilever is etched by an FIB for separated electrodes (a). An FESEM was introduced for real-time observation and two units of manipulators inside the specimen chamber of FESEM were used to fabricate the device. An AFM cantilever was assembled to the manipulator to pick up CNTs from the carbon bulk. The CNTs were cut to 1.7 μm length by current, and fabricated to PZT cantilever (b). A tungsten probe was used for the anode in field emission. The probe was etched by an FIB and then coated with parylene. The FIB was then used to cut the tip of the probe to expose the conductive material to concentrate the electro field. The etched tungsten probe was positioned to the counter of the CNTs. For field emission of CNTs, the emission characteristics should be measured before growth. The trimethylcyclopentadienylplatinum precursor was inserted into the specimen chamber of the FESEM and a constant current applied to CNT 1 and etched tungsten probe to grow the platinum nanowire (c). Using the precursor hexacarbonyl tungsten and applying a constant current to the etched tungsten probe the CNT 2 resulted in the growth of the tungsten nanowire (d).

A tungsten nanoprobe (MicroManipulator Inc, MODEL 7B) was used as the anode for nanowire growth via field emission. To decrease the branching of the nanowire, the applied field needs to be concentrated. The cathode for the field emission is an MWCNT which is tens of micrometeres in diameter. The tip radius of the tungsten probe is nearly 1 μm as shown in Fig. 10a and needs to be reduced. The FIB technique is used to etch the tungsten tip to a 300 nm thick plane which is shown in Fig. 10b. The probe is then turned clockwise through 90°, and the plane etched to form a probe 300 × 300 nm in cross section (Fig. 10c). A parylene coating unit (Dix-coating, DACX-Lab) was employed to coat a 300–400 nm thickness of the polymer (Fig. 10d). The tip of the probe was etched by an FIB to remove the polymer as shown in Fig. 10e.

A dynamic force microscope (DFM) cantilever (KEYENCE OP-75041 DFM/SS) was used for the substrate CNTs. The cantilever comprises

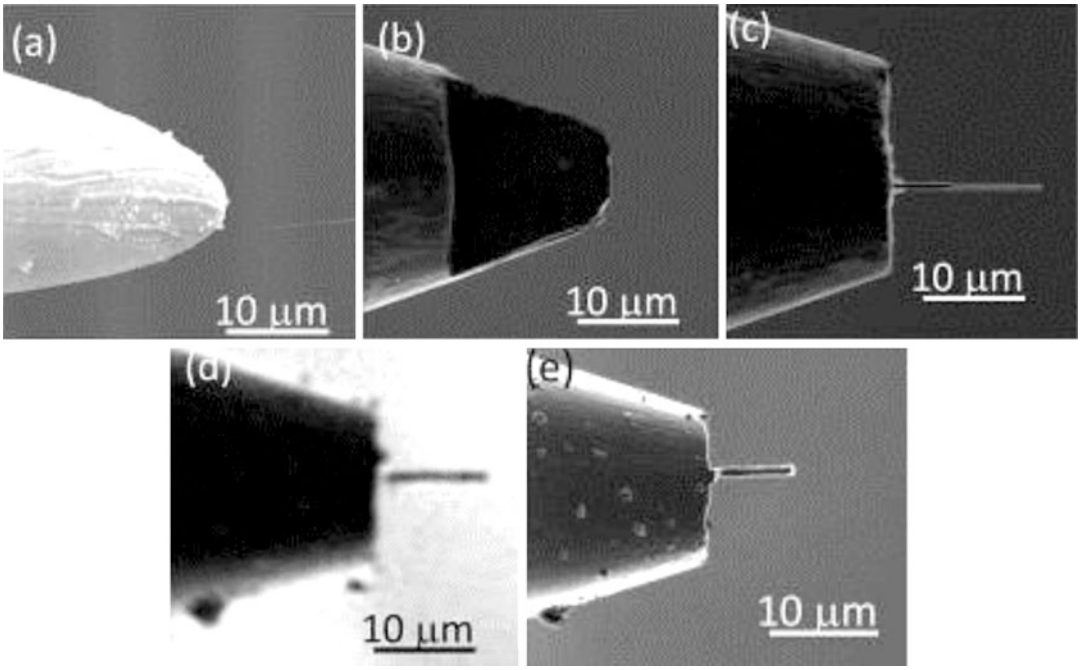


CNT Handling and Integration, Fig. 9 Concept draws of nanowire probe pH sensor

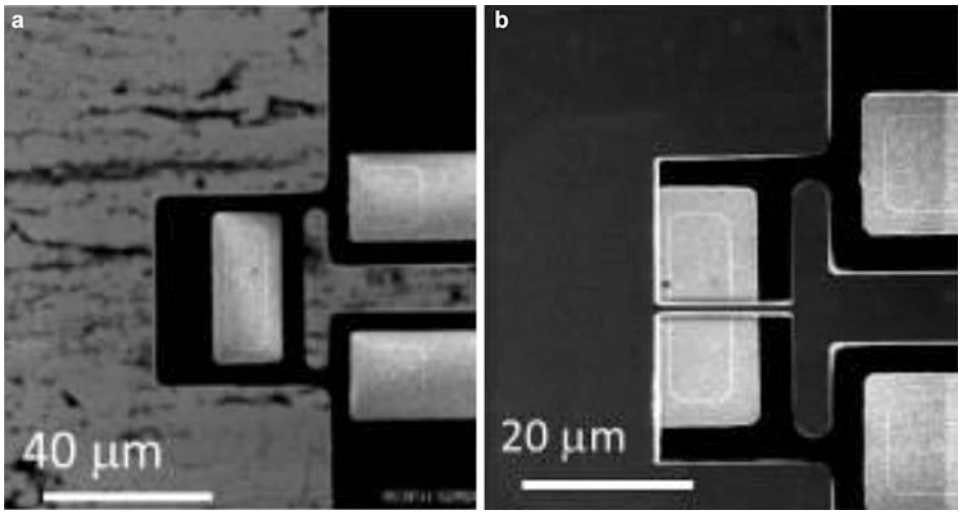
three layers. The bottom layer and the upper layer are insulators and the middle layer is alumina. We cut the edge of the cantilever and then separated the cantilever with FIB etching. The gap of the electrodes is 700 nm. Figure 11a, b, respectively, shows the cantilever before and after etching.

The CNTs were picked up from the carbon bulk by the AFM cantilever controlled by the nanorobotic manipulator. As shown in Fig. 12, the CNTs were assembled to electrodes of the PZT cantilever individually. The CNT 1 is 1.75 μm in length by 20 nm in diameter, and the CNT 2 is 1.78 μm in length by 31 nm in diameter as shown in Fig. 13a. Figure 13a, b shows the platinum nanowire growth on the tip of CNT 1. The platinum nanowire was grown from the MWCNT emitter by using the CpPtMe3 (STREM CHEMICALS purity 99 %) precursor. The pressure of the FESEM specimen chamber was about 5×10^{-3} Pa. A constant field emission current 1 μm was applied for 300 s with the gap at 0.62 μm. The gap of CNT 1 and the etched tungsten probe is set using the nanorobotic manipulator (a). The platinum nanowire is 209 nm in length (b). Figure 13c, d shows the tungsten nanowire growth on the tip of CNT 2. The tungsten nanowire grew from the MWCNT emitter by intruding $W(CO)_6$ (ALDRICH, 97 %) precursor. The pressure of the FESEM specimen chamber was again about 5×10^{-3} Pa. A constant field emission current of 500 nA was applied for 60 s at a gap of 0.70 μm. The gap of CNT 2 and etched tungsten probe is set by the nanorobotic manipulator (c). The platinum nanowire is 907 nm in length (d).

We constructed a pH sensor evaluation system as shown in Fig. 14. The buffer solutions were

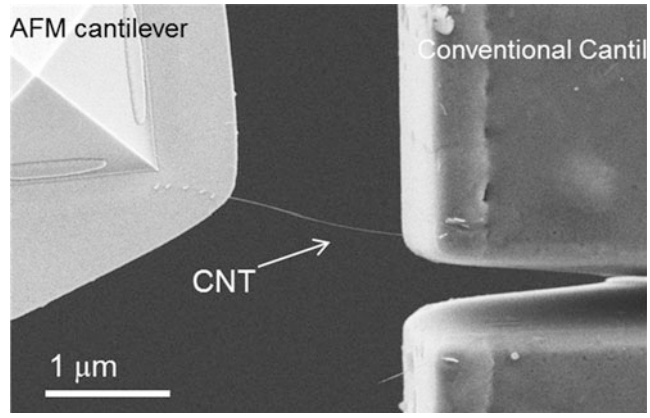


CNT Handling and Integration, Fig. 10 Fabrication of anode by FIB etching. (**a**) a tungsten probe tip (**b**) FIB etching to a plate (**c**) turned 90° and FIB etched to a needle. (**d**) etched tungsten needle coved with a parylene layer. (**e**) FIB etching the tip of the needle in order to cut the parylene layer at the tip area



CNT Handling and Integration, Fig. 11 Fabrication of the cantilever for output electrodes by FIB

introduced into the flow system. Next, a nanoprobe pH sensor was inserted to detect the electropotential together with a reference pH sensor for pH detection. A voltmeter which was constructed using a pA meter and a large resistance was employed to measure the electro potential. The characteristics were evaluated as the electro potential and pH values. Because of liquid



CNT Handling and Integration, Fig. 12 Fabrication of the cantilever for output electrodes by FIB

surface tension, the nanoprobe pH sensor has a risk of bending and being damaged. However, several buffer solutions are required for testing which increases the risk of damage to the nanoprobe pH sensor. It was necessary to provide a solution to this problem, and for this purpose we used continuous flow measurements. A container with an inlet and outlet was fabricated. Initially, buffer solutions of pH = 4 were flowed into the container and then the pH sensor nanoprobe was inserted into the solution. A reference pH sensor was placed close to the pH sensor nanoprobe in the container to monitor the real-time pH value. The electromotive force of the tungsten oxide and platinum nanowires was measured using a voltmeter. For the second step, a buffer solution of pH = 6.9 was flowed into the container, mixing with the previously introduced solution. The pH value changed to 6.1 as measured by the reference pH sensor. The electromotive force of the nanowires was measured. We continued to flow in the buffer solution of pH value 6.9 until the pH value of the container changed to 6.9. The electromotive force of nanowires was measure in this solution. A buffer solution of pH value of 9 was used to flow into the container, and when the pH value changed to 8, we again measured the electromotive force of the nanowires.

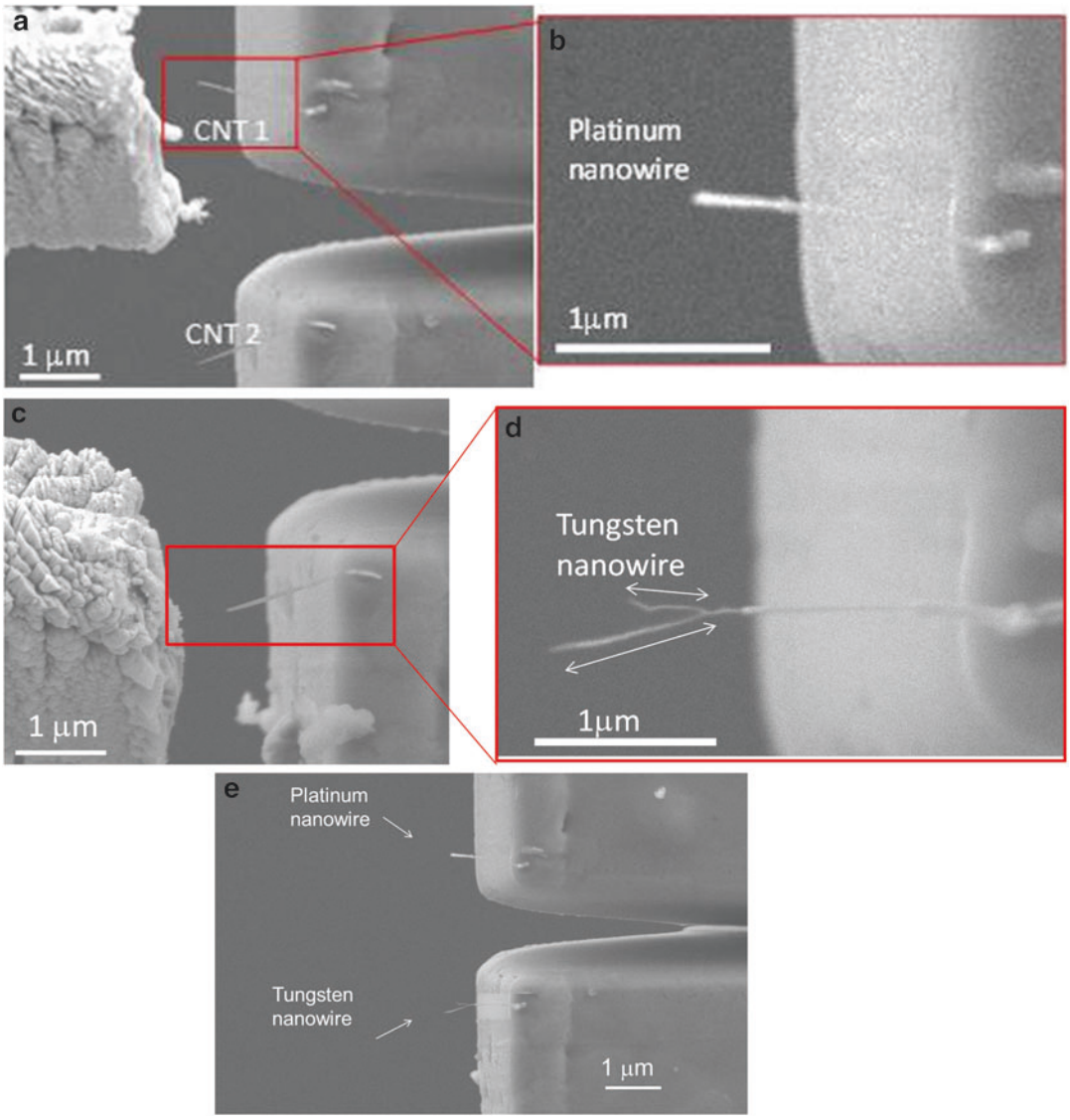
In order to prove that the nanoprobe pH sensor could be used to detect hydrogen ions based on the tungsten and tungsten oxide nanowire and the platinum nanowire and to discover the pH response of the CNTs, we fabricated double CNT electrodes and tested their pH response. Two CNTs were picked up from CNT bulks and assembled to FIB etched separated electrodes of a cantilever. The double CNTs device was tested with the buffer solution flow system using solutions with pH values of 4, 6.4, and 8. The electromotive force of the CNTs was a constant value of 0.018 mV.

The term pH value was coined to describe the solution pressure $\mathrm{pH} = -\lg\left(\mathrm{cH/mol\ dm^{-3}}\right)$, in 1909 by Søren P. L. Sørensen pondus Hydrogenii or potentia Hydrogenii, of hydrogen ions in aqueous solutions. International Union of Pure and Applied Chemistry (IUPAC) defines the quantity pH in terms of the molality basis activity αH of hydrogen ions in solution:

$$pH = -\lg[a(H^+)] = -\lg\left[\frac{m(H^+)\gamma_m(H^+)}{m^0}\right] \quad (6)$$

Where $\alpha(H^+)$ is the activity of hydrogen ion H^+ (hydrogen 1+) in aqueous solution, $m(H^+)$, $\gamma_m(H^+)$ is the activity coefficient of H^+ (molality basis) at molality $m(H^+)$, and $m^0 = 1\ \mathrm{mol/kg}$ is the standard molality.

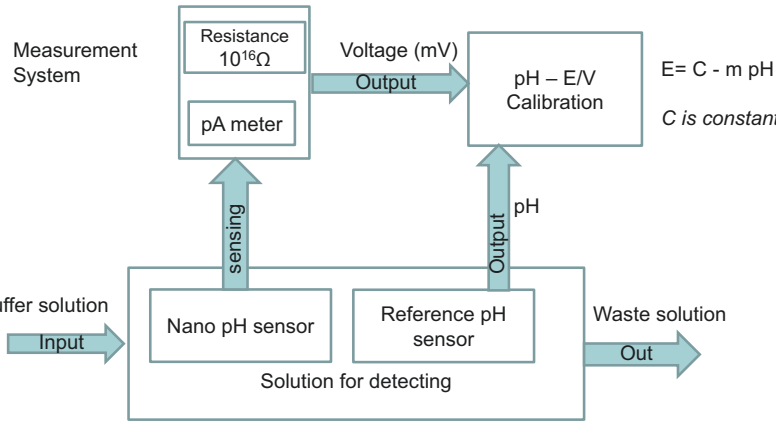
The nanoprobe pH sensor was evaluated using acidic and alkaline buffer solutions with pH 4 and 9 respectively, based on the solution flow system for a series of five times. The results reveal that the nanoprobe pH sensor can detect the pH for buffer solutions of acid (pH 4. 01, pH 6.1, and pH 6.9) and alkali (pH 8) (Fig. 15). Sensor 1 was used to test the buffer solutions at pH 4 and pH 9 to confirm that the sensor has a pH response in acidic and alkaline media. The open circuit potentials



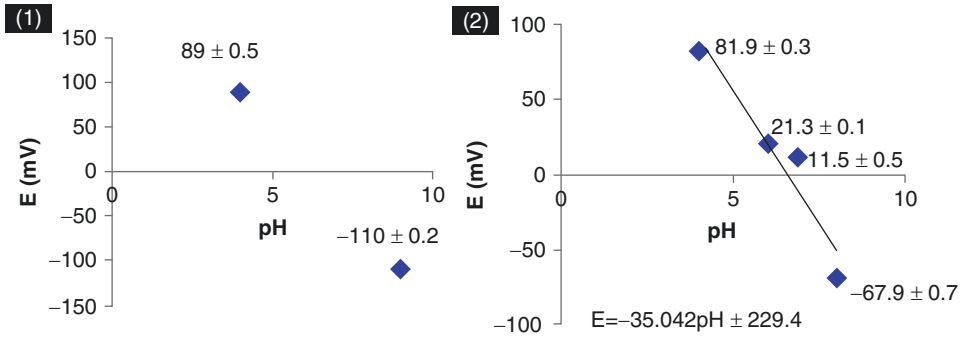
CNT Handling and Integration, Fig. 13 (**a**, **b**) show typical growth of the platinum nanowire and (**c**, **d**) show the tungsten nanowire by field emission which was used for the work and the counter electrode. (**e**) shows the assembled pH probe sensor

(OCP) were 89.1 ± 0.5 and −110.0 ± 0.2 mV, respectively. The results of sensor 1 show that the pH sensor has different electromotive forces under acidic and alkaline conditions. Sensor 2 was used to test the pH sensitivity of the tungsten oxide and platinum nanowire pH sensor. The OCP values in buffer solutions of pH 4.0, 6.1, 6.9, and 8.0 were 81.9 ± 0.3, 21.3 ± 0.1, 11.5 ± 0.5, and −67.9 ± 0.7 mV, respectively. By calculation, the sensitivity of the nanowire pH sensor was −35.04 mV pH^{-1}. The error of the sensor is 0.1 mV level. The resolution of open circuit potentials are 0.01 mV and significant figure should be 0.1 mV. The experimental pH accuracy is 0.8 pH.

The performance of pH sensors is usually characterized by measuring the OCP of the electrodes in solutions with various pH values.



CNT Handling and Integration, Fig. 14 Block diagram for the nano pH sensor evaluation



CNT Handling and Integration, Fig. 15 pH sensors evaluation in the buffer solutions

A single-phase interaction electrode may similarly be envisaged, though no example of a pH electrode involving oxygen-deficit phases has been provided [18, 19]. By omitting hydration, the surface mechanism can be expressed as follows:

$$WO_3 + 2nH^+ + 2ne^- \Leftrightarrow WO_{3-n} + nH_2O \quad (7)$$

According to the Nernst equation for the equilibrium case, the electrode potential can be stated:

$$\begin{aligned} E &= E^0 - \frac{R \cdot T}{z \cdot F} \ln \frac{\alpha_{(WO_{3-n})}}{\alpha_{(WO_3)} \cdot \alpha_{(H^+)}^{2n}} \\ &= E^0 - \frac{R \cdot T}{z \cdot F} \ln \frac{\alpha_{(WO_{3-n})}}{\alpha_{(WO_3)}} - \frac{R \cdot T}{F} \ln \frac{1}{\alpha_{(H^+)}} \\ &= E^0 - \frac{2.303 \cdot R \cdot T}{F} pH \\ &= E^0 + mpH \end{aligned} \quad (8)$$

Where E is the measured potential, E^0 is the standard potential which is zero, R is the gas constant (8.314 J K^{-1} mol^{-1}), T is the absolute temperature (K), z is the signed ionic charge (mol), and F is the Faraday constant (96487.3415 C mol^{-1}). At room temperature $(T = 298\ K)$, during the reaction the $\frac{\alpha_{(WO_{3-n})}}{\alpha_{(WO_3)}}$ approximate to 1 [20], the slope m should be

$$\begin{aligned} m &= -2.303 \cdot R \cdot T/F = -0.0591V/pH \\ &= -59.1mV/pH \end{aligned} \quad (9)$$

At the counter electrode, the platinum nanowire surface reaction can be described as follow [20]:

$$2H_2O \Leftrightarrow O_2 + 4H^+ + 4e^- \quad (10)$$

Here, platinum plays a role as catalyst, so the performance of this pH sensor is

−59.1 mV/pH. The sensitivity of the pH sensor nanoprobe which we fabricated was −35.04 mV/pH. The reason that this sensitivity had a bias with respect to that from the Nernst equation is due to the fact that the tungsten nanowire is not of a constant thickness. And the method of tungsten growth is field emission. The growth tungsten plays a role as field emission tip which is high temperature up to 1000 K. The high temperature leads to the decrease of sensitive of tungsten/tungsten trioxide nanowire hydrogen ion sensitive. The tungsten nanowire grown via field emission has a polycrystalline structure and the lattice direction was random. Various tungsten oxides covered the surface of the tungsten nanowire. The thickness of the tungsten trioxide is not constant, which leads to the observed bias in the value of the sensitivity.

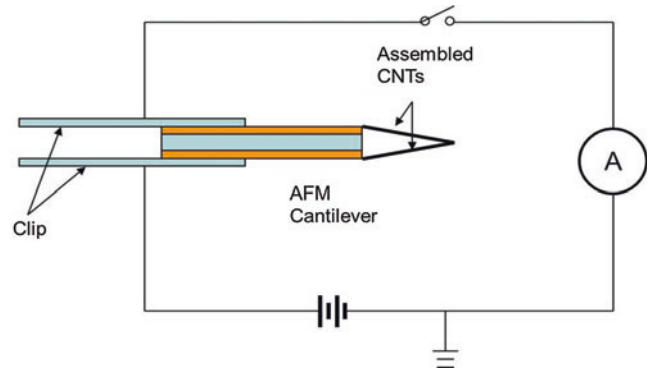
We aim to detect the pH value and pH distribution inside a cell. Common cell has a size of 5–50 μm. We design an electrochemistry pH sensor with a potential to inject to the cell and measure the pH distribution. The issue is how to make an integrated probe type pH sensor. The previous research had designed the pH sensor with a sensitive nanowire work electrode, however the counter electrode is several centimeter in diameter. It is impossible to insert into a cell and measure the pH distribution inside the cell. This is the "size problem." Here, we integrated the nanowork electrode and counter electrode into a single device with a gap $1 \sim 2\mu m$ in between. The proposed pH sensor has to insert cell and detection the pH.

The advantages of the pH sensor are as follows. First, the design of this pH sensor is an integrated probe type sensor; to the author's knowledge this is the first design of a nanoscale integrated probe type pH sensor. Second, in the fabrication the nanowires were used for work electrode and counter electrode and fabricated on one probe. This sensor has a potential to detect pH value inside the cell. Second, the sensitivity of pH sensor is near to the theory calculation. And with the calibration and calculation the resolution of this pH sensor is 0.8 pH.

CNT Probe Thermal Sensor Based on Nanorobotic Manipulation

This part reports a thermal sensor assembled by two multiwalled carbon nanotubes (MWCNTs) inside a field emission scanning electron microscope (FESEM) which constructed a probe tip in order to detect the local environment of single cell. An atomic force microscope (AFM) cantilever was used for substrate which is composed by Si_3N_4 and double side was covered by gold layer. Two MWCNTs were assembled to both side of AFM cantilever individually by employing a nanorobotic manipulation. A second AFM cantilever was used as an end-effector to manipulate the MWCNTs to touch each other, and used electron beam induced deposition (EBID) to bonding the two MWCNTs. The MWCNTs probe thermal sensor was evaluated inside a thermostated container at 25–60 °C. The experimental results show the positive characteristics of the temperature coefficient of resistance (TCR).

Nanomechanical and nanoelectromechanical devices have been widely investigated in recent years. The precise and controllable nanofabrication methods are important to develop the novel type of nanodevices. One-dimensional nanostructure such as nanowires and nanotubes are basic building blocks for NEMS. Since carbon nanotubes (CNT) have been discovered many applications have been developed such as mass sensor, nanoradio, and nanomotor. Nanowire is a solid one-dimensional nanostructure which includes metallic nanowires, semi conductive nanowires, and nonmetal nanowires which have applications as nanoneedles, nanohotwire transistor, and nanogenerator. One challenge is to assemble and apply a nanosensor to detect the local environment of single cell. The single cell is the basic element of life and it is the converging point of meter, energy and information. Because of the size of the cell, nanodevice is needed for measurement and manipulation. A thermometer is a device that measures temperature or temperature gradient using a variety of different principles of mechanical, thermojunctive, thermoresistive, and radiative (infrared and optical pyrometers). To use as a sensor, the extremely small size of the CNT can



CNT Handling and Integration, Fig. 16 Concept drawing of the MWCNTs probe thermal sensor

provide accurate measurement at nanoscale size and reduces the possibility of disturbing the local environment. In addition, the small size sensor implies very low power consumption. To use as a temperature sensor, CNT can have the smallest time constant and provides the extremely rapid time response with the measured object's temperature.

C. K. M. Fung. et al. reported a CNT-based thermal sensor. With two Au electrodes, the bundled CNTs were assembled in between on the substrate. C. Y. Kuo et.al. fabricated lateral growth of CNTs between two electrodes thermal sensor. The CNTs grow on the substrate between electrodes. F. Arai, et al. reported a single CNT thermal sensor with a CNT assembled on four electrodes. The former sensors have advantages of smaller size, lower energy cost, and suitable to measure the specimens within 2D surface. Due to the restriction of the architecture design, these sensors can only measure the specimens with the size larger than 1 mm scale.

We proposed a thermal sensor assembled by two multiwalled carbon nanotubes (MWCNTs) inside a field emission scanning electron microscope (FESEM). The advantage of the MWCNTs probe sensor is to detect the local environment of single cell with high accuracy as shown in Fig. 16. The basic idea is to detect temperatures by the thermal coefficient of resistance. The biological cells are commonly alive and cultured at temperature for 35–37 °C. When the temperature goes up to 43 °C most cells will be exterminated. We aim to evaluate the thermal sensor in the range of 30–60 °C.

A nanowire probe is designed by temperature coefficient of resistance (TCR) method. The electrical resistance of a conductor is dependent upon collisional processes within the wire; the resistance could be expected to increase with temperature since there will be more collisions. An intuitive approach to temperature dependence leads one to expect a fractional change in resistance which is proportional to the temperature change:

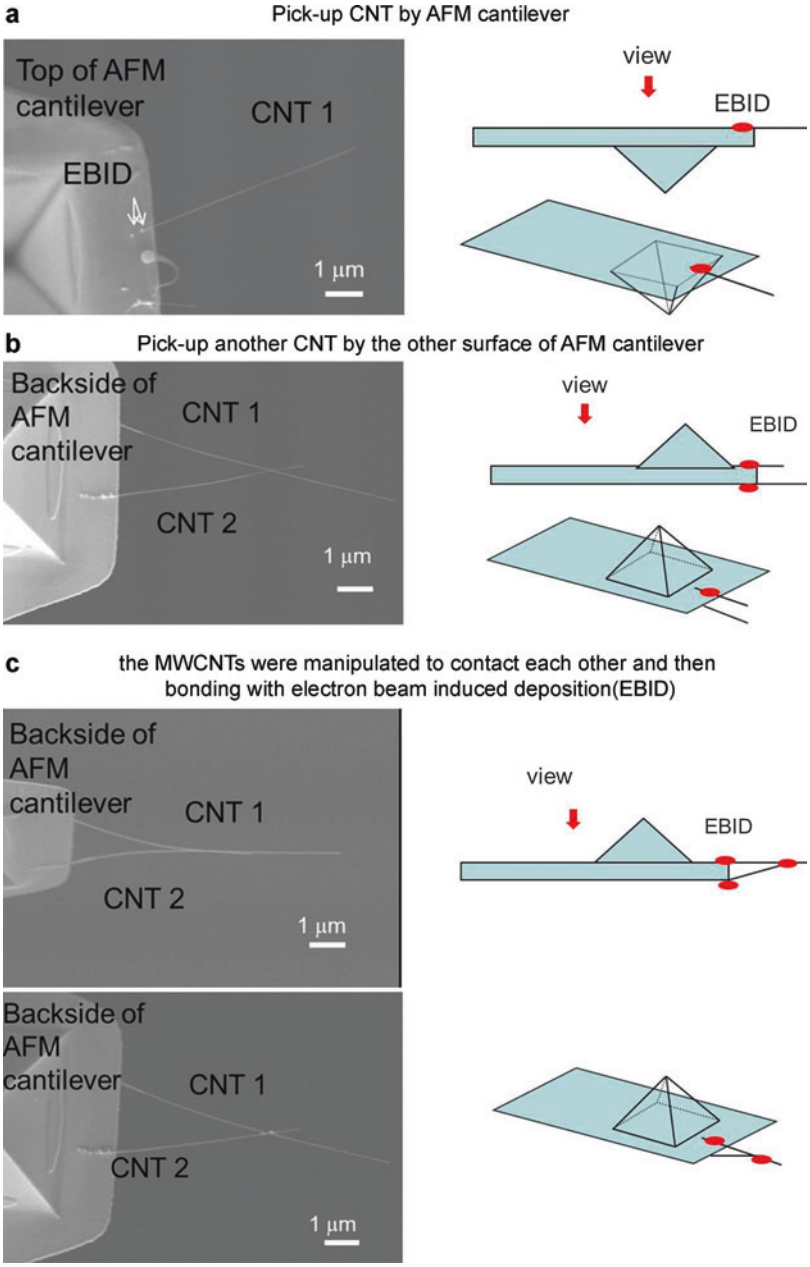
$$\frac{\Delta R}{R_0} = \alpha \Delta T \tag{11}$$

Here ΔR is the resistance change. R_0 is the initial resistance at 0 K. α is the temperature coefficient of resistance. ΔT is the temperature change.

Based on temperature coefficient of resistance, we designed a probe type thermal sensor by CNT. The basic design is the resistance of CNT change since temperature change. An AFM cantilever was used for substrate which is constructed by Si_3N_4 and coated by Au layer both sides. CNTs were assembled to the surface of AFM cantilever. Then use another cantilever to force the CNTs to touch each other.

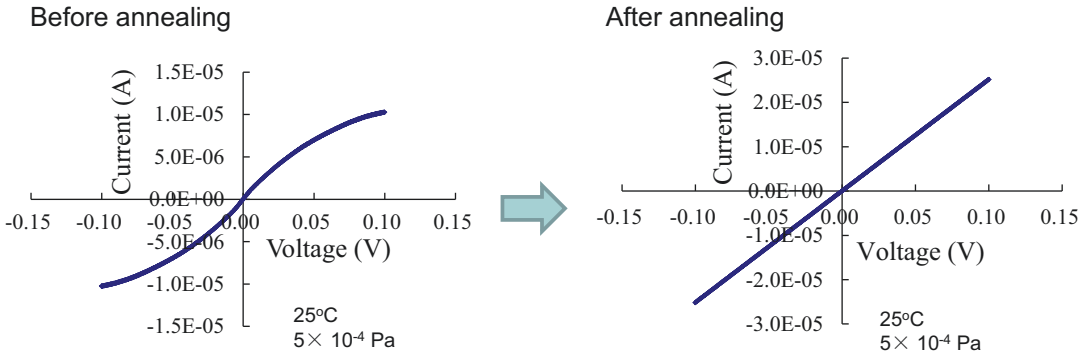
An AFM cantilever (Olympus, OMCL-TR400PB-1) is used as an end-effector to pick up the MWCNT. An AFM cantilever is constructed by Si_3N_4 and coated by Au layer on two sides. We use AFM cantilever as robot hand for MWCNT pick up and assembly. The MWCNT probe is fabricated from its bulks which are synthesized by arc-discharge method.

First, the AFM cantilever was manipulated to pick up a CNT from CNT bulk, which was

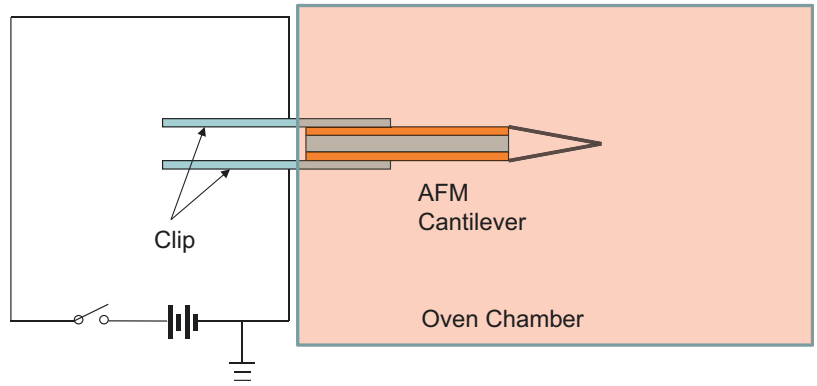


CNT Handling and Integration, Fig. 17 Assembly of MWCNTs probe thermal sensor. (**a**). Pick up CNT by AFM cantilever. (**b**) Pick up another CNT by the other surface of AFM cantilever. (**c**). The MWCNTs were manipulated to contact each other and then bonding with electron beam induced deposition (*EBID*)

synthesized by arc-charge method. The top surface of cantilever was used to pick up CNT 1. The cantilever was moved to below the CNT and went up in order to use Au surface to touch and bonding the CNT with van der Waals force. Electron beam induced deposition points were used to reinforce the bonding of CNT 1 and Au substrate as shown in Fig. 17a. Second step, we turned the cantilever over. And we assembled CNT 2 to the back surface of the Au surface as shown in Fig. 17b. Third step, we used another cantilever to manipulate CNT 1 in order to touch CNT 2. Then, an EBID was deposited on the CNT's contact point to improve the bonding as shown in Fig. 17c. In total four MWCNTs probe thermal sensors were fabricated. The MWCNTs were synthesized by



CNT Handling and Integration, Fig. 18 Annealing of CNT



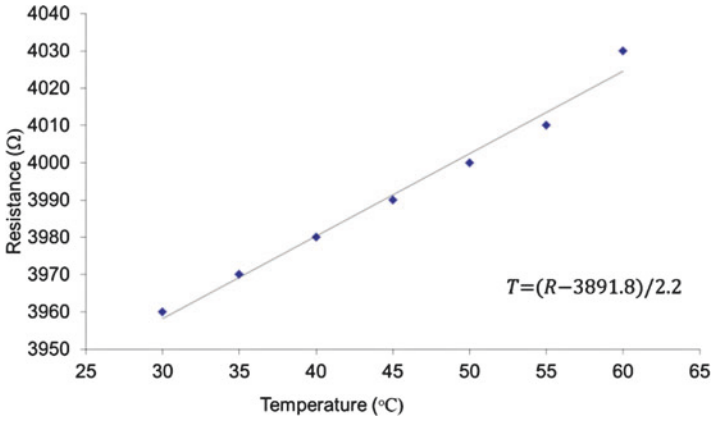
CNT Handling and Integration, Fig. 19 Schematic of evaluation of thermal sensor

arc-charge method. Defects of MWCNTs are issues of mechanical and electrical applications. Annealing is a method to solve this problem. We applied sweep voltage −1–1 V to anneal the CNT. The Fig. 18 shows the anneal I-V curve. Figure 19 shows the schematic of thermal sensor evaluation. A probe thermal sensor was placed to the holder which is colored yellow. The two sides of AFM cantilever were connected to positive and negative electrodes individually. The probe sensor was set into an oven and tested for I-V curve from 30 °C to 60 °C for every five degrees.

The oven was set to 30 °C and maintained at this temperature for 2 h. Then we tested the I-V response of the MWCNT probe thermal sensor. The resistance was 3960 Ω with the standard error of 1.5 Ω. For next step, the oven temperature was increased from 5°C to 35 °C in 1 h and maintained 2 h. The resistance was 3970 Ω with the standard error of 7.4 Ω. The test points at 40 °C, 45 °C, 50 °C, 55 °C and 60 °C was set-up. The tested resistance was 3980 Ω, 3990 Ω, 4000 Ω, 4010 Ω, and 4030 Ω with standard error 1.0 Ω, 1.9 Ω, 1.3 Ω 11.1 Ω, and 1.0 Ω.

The result of CNT annealing was shown in the Fig. 20. The I-V curve had been returned to linear. The defects of CNTs were repaired. The mechanism can be described like this, the CNT were cover some of amorphous carbon. When sweep current is applied to the CNT the remote joule heating [21] investigated to the amorphous carbons and rebuilt the defects of bonding at CNTs. When the charge carriers flow pass the CNT, the vibration of lattices of CNT will absorb the charge carrier energy and create a phonon.

The experimental results show that temperature coefficient of resistance of CNT was linear in Fig. 20. The relationship is temper = resistance −3891.8/2.2. And the standard error was at 10 Ω level. The metallic MWCNT usually has a



CNT Handling and Integration, Fig. 20 Evaluation result

resistance of several k Ω. Here the increase in temperature was set to 5 °C because of the oven's limitations. According to the Eqs. 5 and 6, the temperature coefficient of resistance is calculated as 0.05 %.

The total resistance thermal sensor

$$R_{\text{total}} = R_{\text{CNT}} + R_{c1} + R_{c2} + R_{c3} \tag{12}$$

Where, R_{total} is the total resistance of nanothermal sensor. R_{c1} is contact resistance of CNT 1 and Au substrate. R_{c2} is the contact resistance of CNT 2 and Au substrate. R_{c3} is the contact resistance of two CNTs. And the resistance of CNT can be calculated by the following equation.

$$R_{\text{CNT}} = \rho_{NT} \times \frac{L}{\pi\left(\frac{d}{2}\right)^2} \tag{13}$$

Where, R_{CNT} is the resistance of CNT. ρ_{NT} is the resistivity of the CNT. L is the length of the CNT. d is the diameter of the CNT. In this experiment, CNT 1 is approximately 3 μm in length and 37 nm in diameter. The CNT 2 is approximately 3 μm in length and 42 nm in diameter. The ρ_{NT} is $3.8 \times 10^{-5}\Omega\text{cm}$ [22]. By the calculation the $R_{\text{CNT}} = 1.8\,\text{k}\Omega$.

Chun Lan et al. discussed the metal carbon contact resistance and resistance of CNT. The MWCNT resistivity is 0.33 $k\Omega/\mu m$ and specific contact résistance of CNT and Au surface is 4.4 $k\Omega\mu m$.

$$R_{\text{CNT}} = 33\frac{k\Omega}{\mu m} \times 6\mu m \approx 1.98k\Omega \tag{14}$$

$$R_{c1} + R_{c2} = \frac{4.4k\Omega\mu m}{2.5\mu m} = 1.76\ k\Omega \tag{15}$$

So

$$R_{c3} \approx 210\ \Omega\ \sim 390\Omega \tag{16}$$

The calculation results of CNT resistance meet the reference data and the theoretical calculation. The MWCNT can be considered as a graphene cylinder. The single crystal graphite shows the linear thermal response when the temperature is higher than 200 K. The evaluation result of nanothermal sensor is linear.

In summary, this section presented the CNT integration and applications and fabrication of a functional pH sensor with a double metallic nanowire. A cantilever was used for the electrodes of the device and etched by an FIB. Two CNTs were assembled to the separated electrodes of the cantilever. A tungsten probe was etched by an FIB to 300 nm in diameter and 25.4 μm in length. Then the probe was coated by parylene and the tip cut to expose the tungsten. A tungsten nanowire with 907 nm length and a platinum nanowire with 209 nm length grew from the tip of the CNTs via field emission by introducing hexacarbonyltungsten and trimethylcyclopentadienylplatinum, respectively, inside a field

emission electron microscope. The acid and alkali response of the pH sensor was tested; the sensor was able to detect the pH value, and the performance of the pH sensor was −35.04 mv/pH. We hope that these findings will be rapidly adopted in the field of biochemical detection.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Nanorobotic Assembly
- ▶ Nanorobotics
- ▶ Scanning Electron Microscopy

References

1. Iijima, S.: Helical microtubules of graphitic carbon. Nature **354**(6348), 56–58 (1991)
2. Visser, J.: Van der Waals and other cohesive forces affecting powder fluidization. Powder Technology. **58**, 1–10 (1989)
3. Zheng, Q.S., Liu, J.Z., Jiang, Q.: Excess van der Waals interaction energy of a multiwalled carbon nanotube with an extruded core and the induced core oscillation. Phys. Rev. B **65** (2002)
4. Jonsson, L.M., Nord, T., Kinaret, J.M., Viefers, S.: Effects of surface forces and phonon dissipation in a three-terminal nanorelay. J. Appl. Phys. **96** , 629–635 (2004)
5. Dong, L.X., Arai, F., Fukuda, T.: Nanoassembly of carbon nanotubes through mechanochemical nanorobotic manipulations. Jpn. J. Appl. Phys. Part 1 **42**, 295–298 (2003)
6. Akita, S., Nishijima, H., Nakayama, Y.: Influence of stiffness of carbon-nanotube probes in atomic force microscopy. J. Phys. D-Appl. Phys. **33** , 2673–2677 (2000)
7. Xia, Y.N., Whitesides, G.M.: Soft lithography. Annu. Rev. Mater. Sci. **28** , 153–184 (1998)
8. Lixin, D., Arai, F., Fukuda, T.: Destructive constructions of nanostructures with carbon nanotubes through nanorobotic manipulation. IEEE/ASME Trans. Mechatron. **9**, 350–357 (2004)
9. Nakajima, M., Arai, F., Fukuda, T.: In situ measurement of young's modulus of carbon nanotubes inside a TEM through a hybrid nanorobotic manipulation system. IEEE Trans. Nanotechnol. **5**, 243–248 (2006)
10. Heping, C., Ning, X., Guangyong, L., Jiangbo, Z., Prokos, M.: Planning and control for automated nanorobotic assembly. In: Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on, pp. 169–174. (2005)
11. Deboux, B.J.C., Lewis, E., Scully, P.J., Edwards, R.: A novel technique for optical fiber pH sensing based on methylene blue adsorption. J. Lightwave Technol. **13**, 1407–1414 (1995)
12. Ruan, C., Zeng, K., Grimes, C.A.: A mass-sensitive pH sensor based on a stimuli-responsive polymer. Anal. Chim. Acta **497**, 123–131 (2003)
13. Yao, M.W.S., Madou, M.: A pH electrode based on melt-oxidized iridium oxide. J. Electrochem. Soc. **148**, 29–36 (2001)
14. Talaie, A., Lee, J.Y., Lee, Y.K., Jang, J., Romagnoli, J. A., Taguchi, T., et al.: Dynamic sensing using intelligent composite: an investigation to development of new pH sensors and electrochromic devices. Thin Solid Films **363**, 163–166 (2000)
15. Bashir, R., Hilt, J.Z., Elibol, O., Gupta, A., Peppas, N. A.: Micromechanical cantilever as an ultrasensitive pH microsensor. Appl. Phys. Lett. **81**, 3091–3093 (2002)
16. Bergveld, P.: Thirty years of ISFETOLOGY: what happened in the past 30 years and what may happen in the next 30 years. Sensors Actuators B Chem. **88**, 1–20 (2003)
17. Nomura, Y.K.S., Hiraishi, N., Nakajima, M., Nikaido, T.: Application of semiconductor pH-imaging sensor to the imaging of extracted human carious dentin. Anal. Sci. **17**, 539–541 (2001)
18. Fenster, C., Smith, A.J., Abts, A., Milenkovic, S., Hassel, A.W.: Single tungsten nanowires as pH sensitive electrodes. Electrochem. Commun. **10**, 1125–1128 (2008)
19. Zhang, W.-D., Xu, B.: A solid-state pH sensor based on WO_3-modified vertically aligned multiwalled carbon nanotubes. Electrochem. Commun. **11**, 1038–1041 (2009)
20. Kurzweil, P.: Metal oxides and ion-exchanging surfaces as ph sensors in liquids: state-of-the-art and outlook. Sensors **9** , 4955–4985 (2009)
21. Baloch, K.H., Voskanian, N., Bronsgeest, M., Cumings, J.: Remote Joule heating by a carbon nanotube. Nat. Nanotechnol. **7**, 316–319 (2012)
22. Primak, W., Fuchs, L.H.: Electrical conductivities of natural graphite crystals. Phys. Rev. **95**, 22–30 (1954)

CNT NEMS

- ▶ Carbon Nanotube NEMS

CNT Resonator

- ▶ Carbon Nanotube NEMS

CNT-FET

▶ Nanostructure Field Effect Transistor Biosensors

Coarse-Grained and Hybrid Simulations of Nanostructures

Richard Gowers and Paola Carbone
School of Chemical Engineering and Analytical Science, The University of Manchester, Manchester, UK

Synonyms

Mesoscopic simulations; Multiscale simulations

Definition

In computational chemistry coarse-grained (CG) models are defined as molecular models where some details (i.e., degrees of freedom) of the original chemical structure have been removed. The resulting models are a coarser description of the chemical systems compared with the original ones and can then be used to perform either molecular dynamics or Monte Carlo simulations [1]. The reduction of the models' degrees of freedom enables the simulation of systems whose size is comparable with that of the experimental ones and the timescale spanned by these simulations can reach microseconds.

Overview

Computer modeling is a powerful technique to gain molecular level details of chemical systems under different physical conditions and enables to relate macroscopic observations with changes in the chemical and physical state of the system. However, all modeling techniques rely on computer hardware, and therefore their use is limited by the available computer power. State-of-art simulations can nowadays reach the size of few millions of atoms, but if standard high-performance computers are used, the system size usually does not exceed few hundred thousand particles. Indeed, during a molecular simulation, the number of interatomic interactions that must be computed every iteration is proportional to N^2, where N is the total number of system particles [1]. This heavy use of the CPUs limits not only the size of molecular models but also the timescale the system can be simulated for.

One way to circumvent this problem is to reduce the number of interacting particles (N) in the systems, simplifying the models and modifying the original interacting parameters to include in an implicit way the neglecting details. The simplest example of such coarse graining is the development of united-atom (UA) force fields [2] where the hydrogen atoms and the aliphatic carbon to which they are covalently bonded are modeled as a single entity. The assumption underlying the development of such force fields is that the physics of the model is not affected by neglecting the explicit interactions involving the aliphatic hydrogen atoms. A similar decision on how many and which atomistic details can be neglected in a molecular model (procedure known as mapping scheme) is the key decision that must be carefully made every time a new coarse-grained model is developed if some of the system chemical and physical features have to be maintained. It has been indeed shown that mapping schemes which retain different features of the original molecule perform differently depending on the property analyzed [3].

The model simplification done in the UA force fields can therefore be carried on at much larger scale: large group of atoms can be lumped up in single super-atoms (or beads) and even entire colloidal particles can be modeled as single rigid bodies.

A very broad distinction can be made between those CG models aiming at preserving some chemical details of the original system and those which instead preserve only the shape of large

molecular aggregates. The former can reproduce with a certain degree of accuracy the system enthalpy, while the latter reproduce the only entropic effects, and it is typically employed in modeling colloidal systems [3–6].

Procedures

Since different features of the atomistic model can be used as target properties, several procedures to develop the effective interactions between the beads have been proposed. Some of the most popular methods used in materials science are briefly introduced below.

Structural Based Model

The target property to reproduce is the structure of the atomistic model. The CG non-bonded part of the force field is therefore refined until it reproduces structural correlation functions such as the radial distribution function (RDF) of the atomistic system. The quality of the agreement is quantified using a merit function of the form

$$\chi^2 = \sum \left(y_{\text{target}} - y_{CG} \right)^2 \qquad (1)$$

where y_{target} and y_{CG} are the correlation functions calculated from the atomistic and CG model, respectively. The sum is over all coordinates and simulations. The minimization of χ is a nonlinear inverse problem for which no analytical expression is available and where linearization is unsuccessful. An iterative procedure is therefore run and the CG force field parameters are consequently adjusted. Two iterative procedures are usually employed: one uses Monte Carlo simulations [7] and the other one molecular dynamics [8]. In both cases, the new CG model is built upon the atomistic one and the existence of a one-to-one correspondence between the two models enables an easy interchange between the two model resolutions [3].

Thermodynamic Models

The aim of these types of models is to reproduce some thermodynamic properties of the atomistic system. The CG potential is usually modeled using a Mie (generalized Lennard-Jones) function

$$V(r) = C_\varepsilon \left\{ \left(\frac{\sigma}{r} \right)^n - \left(\frac{\sigma}{r} \right)^m \right\} \qquad (2)$$

where n and m are the exponents controlling the softness of the potential, σ is the particle diameter, ε is the potential well depth, and C is chosen in such a way that the minimum of the potential corresponds to $-\varepsilon$.

The key parameters to control and iteratively optimize are ε and σ. In some procedure the exponents n and m can also be varied. The target properties for the CG model include partition coefficients [9], density, and interfacial tension [10]. Another more systematic approach to obtain the sought force field parameters involves the solution of a molecular-based equation of state (EoS) which takes advantage of the fact that for some molecular fluids the SAFT EoS can be solved [11].

Force Matching

In this type of method [12, 13] the CG interactions are again parameterized using the underlying atomistic interactions, but in this case the difference to be minimized is that between the atomistic and CG forces and the minimization is solved using a variational approach

$$\chi^2 = \frac{1}{3N} \left\langle \sum_{I=1}^{N} \left(F_{\text{target}} - F_{CG} \right)^2 \right\rangle \qquad (3)$$

where the angular brackets denote average over the trajectory, F_{target} is the net force on the atomistic site I, F_{CG} is the force on the same CG site, and N is the number of sites (beads) in the CG model. Within this type of CG model a one-to-one correspondence between the atomistic and CG resolution is again established, and reintroduction of atoms in the mesoscopic model is possible.

Excess Entropy Model

In this case the function which needs to be minimized is the relative entropy (S_{rel}) which is defined as the difference between the distributions of configurations generated by the atomistic (P_{at})

and CG (P_{CG}) models. Using the Kullback–Leibler divergence formalist, one can define S_{rel} as

$$S_{\text{rel}}(U) = \int P_{\text{at}} \frac{P_{\text{at}}}{P_{CG}(U)} dR \tag{4}$$

where U is the CG potential and the average is evaluated over the CG configurational space, but weighted according to the atomistic probability distribution, P_{at}. S_{rel} provides a variational framework for determining the approximate CG potential (U) that reproduces target atomistic distributions [14, 15].

Dissipative Particle Dynamics

The techniques presented above despite simplifying the molecular model still retain a quite high degree of chemical specificity. If however much larger systems must be simulated, a mesoscopic approach such as dissipative particle dynamics (DPD) can be employed. Such method can sample large conformational space in relatively short time, and therefore systems of the order of hundreds of nanometers can be simulated [16, 17]. In this method, each bead represents a much larger amount of material than the previously discussed methods, such as an entire molecule or a few molecules of solvent. Because each DPD particle represents such large portion of the chemical system, the standard analytical functions used to model the particle-particle non-bonded interactions (Eq. 2) cannot be used. Therefore, in DPD the forces are expressed as the sum of three contributions

$$F_{ij} = F_{ij}^{C}(r_{ij}) + F_{ij}^{D}(v_{ij}) + F_{ij}^{R}(\theta) \tag{5}$$

where F_{ij}^{C} represents a conservative repulsive force between particles and it is notably weaker than the standard short distance forces (Eq. 2) allowing particles to pass through each other (the so-called soft potential). Such weak short-range interaction is a necessary consequence of the much larger mapping scheme used in this modeling technique. F_{ij}^{D} is a dispersive force between particles which is a function of their relative velocity and represents the effect of fluid viscosity. Finally, F_{ij}^{R} is a random force, a function of a Gaussian random number θ, which replicates the thermal and vibrational energy of the system.

Because all of these forces, including the random force, are applied between pairs of particles, momentum is conserved throughout the system. DPD simulations are tuned to replicate hydrodynamic properties by adjusting the strength of each of these three forces. DPD has been used to model problems which are otherwise out of reach using particle-based methods, including modeling self-assembly and phase diagram of complex fluids.

Mean Field Theory

A radically different approach to coarse graining is that which simplifies the molecular systems at such level that its discrete nature (the fact that is made by individual atoms) is replaced by a continuum mean field. In this context the molecular system is modeled using a grid of points on which an effective field, representing the averaged interaction caused by the presence of all the system particles within a cutoff distance, which is identified with the mesh of the grid, acts. In its basic form, to obtain the field (H_{eff}), the free energy of the system is minimized with respect to the distribution of all possible configurations of the system, P. Since both P and H_{eff} are not known, an iterative process is used until self- consistency is reached [18]:

$$P = \frac{\exp\left(\frac{-W}{k_B T}\right)}{\sum_N \exp\left(\frac{-W}{k_B T}\right)} \tag{6}$$

$$W = \frac{\partial H_{\text{eff}}}{\partial P} \tag{7}$$

where k_B and T are the Boltzmann constant and the temperature, respectively, W is the intermolecular potential, and the summation is run over all the N possible states of the system.

Applications in Materials Science

The procedure briefly presented above can be used in a variety of contexts in both materials science and biology. All have been indeed used

to model synthetic and biopolymers, complex and simple liquids, surfactants, and mixtures of them [3–6]. Below, three examples of such applications in materials science are presented.

Predicting Self-Assembly Properties for Amphiphilic Copolymers

Amphiphilic copolymers are polymer chain formed by more than one type of monomer each with a different polarity. When dissolved in solvents, they self-assemble in a variety of morphologies which are function of the relative chemical affinity of the solvent and monomers and the chain microstructure (the monomer sequence within the chain). Such self-assembled nanometric structure can then be used in a variety of applications from drug and gene delivery agents to emulsion stabilizers [19]. The a priori prediction of their phase diagram is very difficult to achieve, and modeling can help in avoiding the synthesis of many different materials. Using an atomistic model for such simulations is however not possible; therefore CG models have become a popular choice. Mainly using DPD [20] and thermodynamics model [21], it has been recently possible to follow at the self-assembly process and build entire phase diagrams for such systems [20].

Predicting the Long-Time Dynamics of Ionic Liquids

Ionic liquids (ILs) are high molecular weight ion pairs formed by organic cations and bulky counterions. They are liquid at room temperature and below so that they can be exploited in a considerable number of extraction and other chemical processes. Knowing the dynamics of ILs at low temperature can therefore improve the design of the ideal solvent for a specific extraction process; however such experiments are very difficult to carry out since the IL's viscosity increases very rapidly lowering the temperature [22]. CG modeling can help in gaining such data as the dynamics of the model is sped up compared with the experimental one by a factor which depends on the CG scheme used. Using CG models it has been possible to show that certain type of ILs with long aliphatic chains can self-assemble [23] and that at low temperatures they present an increased dynamic heterogeneity which is due to an increasing number of slow ions, while the amount of fast ions stays almost constant with temperature [24].

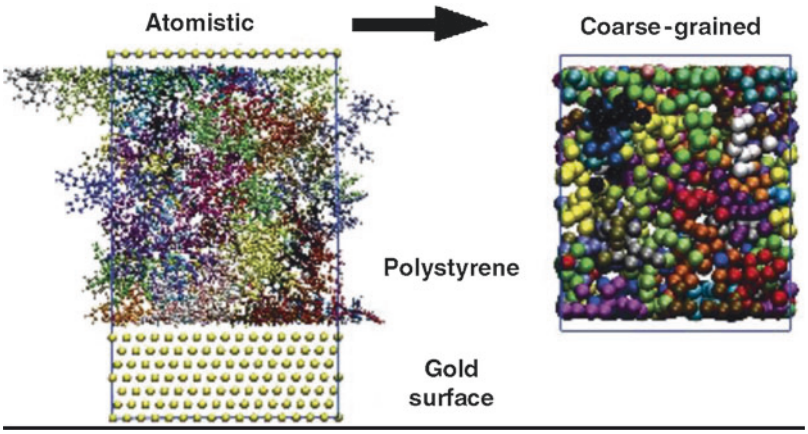
Calculating Interphase Thickness of Polymer Films on Solid Surfaces

When in contact with a solid surface, polymers exhibit complex behavior and understanding and characterizing this behavior is important in the design and development of nanostructures. Such systems are difficult to model as interactions at both small and long length scales need to be included.

A common approach to studying these systems therefore is to propagate information from the smaller length scales into the larger-scale models, and such an approach has been used for modeling polystyrene in contact with a gold surface [25]. Using data from atomistic-scale simulations of small amounts of the nanostructure, a coarser model was constructed (see Fig. 1). This model had reduced the number of particles in the polymer by means of coarse graining and also modeled the gold surface implicitly. This meant that each particle in the polymer interacted with the surface through a single function of normal distance instead of being a sum of interactions with many particles in the surface. This mesoscale model enabled the simulations of much larger volumes and led to the identification of an interphase thickness in which the polymer presents structural properties different than those in bulk.

Future Directions

Coarse-graining techniques have become powerful and well-established tools to access large length and timescales. The use of such techniques however is not without any drawbacks. In certain cases the degrees of freedom coarse-grained away in the mesoscale models represent crucial details without which the behavior of the models no longer faithfully recreate the phenomena that are to be studied. This is, for example, the case in molecular systems where short-range directional non-bonded interactions such as hydrogen bonds are present [3].



Coarse-Grained and Hybrid Simulations of Nanostructures, Fig. 1 An example of the reduction in detail possible when using coarse graining.

To circumvent this limitation typical of traditional coarse-graining techniques, the possibility of mixing both coarser and finer levels of detail in a single simulation in so-called hybrid-scale simulations has been recently explored. These two levels of detail are present simultaneously in the model allowing specific parts of the system to be modeled at a higher level of detail while allowing other parts to use a computationally cheaper model [26].

Two main approaches can be identified for such hybrid-scale simulations. Indeed one can mix either CG and atomistic force fields or particle-based (CG or atomistic) and field models. Both approaches are briefly explained below.

Hybrid Atomistic/Coarse-Grained Models

In models where atoms and CG beads are simultaneously present, in order to calculate the interactions between pairs of atoms or beads, preexisting standard force fields can be used [27]. However, interactions between atoms and beads, the so-called cross terms, must also be evaluated. These cross term interactions could be parameterized using one of the previously described coarse-graining methods; however this is an undesirable option because it would greatly aggravate the work required to derive the model and increase the number of parameters within the model. Instead, a commonly used solution is the use of virtual sites which are geometrical points identified in the finer (atomistic) resolved regions of the model which collect the interactions coming from the model coarser parts. If the CG model is derived from the atomistic one, as it is in the case of structural based or force matching models, the CG mapping scheme is built on the chain atomic structure and each bead center of mass corresponds either to a specific atom or to the center of mass of all the atoms lumped up in it. Thus, it is easy to identify in the atomistic resolved part of the model the positions of the virtual beads and use them as pinning position to collect any mixed interactions. The VS can therefore be seen as coarse-grained representation of a group of atoms, placed at their center of mass, which facilitate their interaction with coarse-grained beads. When a coarse-grained particle interacts with the atoms of a certain group, it is with the virtual site representation of that group. To ensure that all particles can interact with each other, all atoms must be mapped to such a virtual site.

Once all such pairwise interactions have been evaluated, the net force on each virtual site is transferred onto the constituent atoms. This is done on a mass-weighted basis, so that heavier atoms get a larger share of the force from the virtual site. This process ensures that Newton's third law is obeyed and, when used in molecular dynamics simulations, that linear momentum is conserved:

$$F_{\text{atom}} = F_{vs} \cdot \frac{m_{\text{atom}}}{m_{vs}} \tag{8}$$

where $\boldsymbol{F}$ refers to the net force and m the mass of the particle. Using this method, non-bonded

interactions no longer cross between the different resolutions and the coarse-grained force field can be used to describe and parameterize these interactions. This means that once an atomistic and coarse-grained force field is chosen, it is also possible to start creating hybrid-scale models.

Using virtual sites, models which mix both atoms and coarse-grained beads can be constructed. In these models selected chemical motifs are kept at a fine level of detail, while all other sections are coarse-grained. The choice of where to leave fine levels of detail is left to intuition, typically dipoles, including hydrogen bonding sites, or areas where steric factors are of great importance are kept at fine resolution. This approach allows a detailed description of some parts of the model to be combined with a less computationally intensive description of other parts of the model [27, 28]. The Hamiltonian for the simulation system can be defined as

$$H = K_{\text{atoms}} + K_{CG} + V_{\text{bonds}} + V_{nb(cg-cg)} + V_{nb(aa-aa)} + V_{nb(vs-cg)} \quad (9)$$

where K refers to the contribution from kinetic energy and V refers to potential energy between particles.

Adaptive Resolution Models

If what needs to be modeled at high resolution is a particular recurring chemical motif in the molecular system, then the model resolution can be static, which means that during the construction of the model, one decides which chemical moieties are modeled with the different resolutions and this model choice is kept for the entire simulation. However, in some cases it is more suited to define a geometrical region within the simulation box which is modeled with a different resolution. In this case, the total number of degrees of freedom of the model will change over the time depending on how many chemical species enter or leave that region (or regions) [29]. Usually systems are constructed to have a region of interest containing molecules modeled at atomic details, surrounded by a region of coarse-grained molecules. As simulations of such systems progress, the molecules are free to move between these two regions of different resolution, and a method to allow the transition between the two model resolutions is therefore required. To handle such transition, a buffer region (known as healing region) between the two levels of detail is used which allows the molecules to gradually change from one level of detail to another. In this healing region, the models switch from using one force field to the other as a function of their position within the healing region

$$V = \lambda V_{AA} + (1 - \lambda) V_{CG} \quad (10)$$

where λ refers to a switching function between the two force fields. Great care must be taken to ensure that there is no discontinuity in the force acting on the particles, as this would create mass transfer between different regions of the simulation box. Transitioning from the coarse region into the fine detail region is particularly difficult as it involves reintroducing the degrees of freedom that have previously been discarded.

Hybrid Particle Field Models

As the computational cost of both molecular dynamics and Monte Carlo simulation methods scales as $O(N^2)$, increasingly larger simulations become too demanding to produce results in a timely manner. An alternative to this particle-based approach is hybrid models where the particles surrounding a specific atoms or CG bead is replaced by a continuum field and the interparticle non-bonded interactions are calculated with respect to that external field. This method scales as $O(N)$, making the simulation of extremely large systems possible.

This approach is becoming popular in modeling polymer blends or solutions where the polymer chains self-assemble creating specific morphologies. A commonly used field method for simulating such polymer nanostructures is the single-chain mean field theory (SCMF) [30]. In this, single polymer chains or large molecules are separated and each determines its non-bonded interaction with respect to an external

field rather than having to interact with every other particle in every other molecule. To construct the mean field, self-consistent field (SCF) theory is used, where the particle density of a given chemical species (ϕ_A) is measured at all points on a discrete grid. The force caused by the mean field for a particle of species A at a given position is then given as

$$F_A(r) = -k_B T \sum_{A'} \chi_{AA'} \frac{\partial \phi_{A'}(r)}{\partial r} - \frac{1}{\kappa}\left(\sum_A \frac{\partial \phi_A(r)}{\partial r} - 1\right) \quad (11)$$

where k_B and T are the Boltzmann constant and temperature. The term $\chi_{AA'}$ is a Flory-Huggins mixing parameter between two species and must be defined between all combinations of species and κ represents the compressibility of the system. The first term defines how favorably different species can mix, while the second term tries to balance the particle density. This field varies both spatially and temporally and must be updated as the simulation progresses. However the variation of the field is slow with respect to the intramolecular forces, meaning that several particle moves can be done before the field must be updated. Within this approach while the intermolecular interactions are calculated via the SCMF theory, the intramolecular interactions (i.e., bonds, angles, and torsions) are handled at the same level of precision as conventional particle-based simulations. Initial work on this method used Monte Carlo to perform the particle moves [31] as the use of Monte Carlo allows the possible conformational space of the molecules to be sampled effectively. More recently molecular dynamics has also been used [32]. Since the calculation of the interactions with a discrete grid is computationally quick to perform, with each particle only interacting with eight different points rather than all their neighbors, this method is also particularly well suited to the increasing parallel nature of the hardware available, with some software able to achieve hyper-parallelism (achieving a speedup greater than n using n processors) [32].

Cross-References

- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Mesoscopic Modeling

C

References

1. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Oxford University Press, Oxford (2002)
2. Ponder, J.W., Case, D.A.: Force fields for protein simulations. Adv. Protein. Chem. **66**, 27–85 (2003)
3. Karimi-Varzaneh, H., van der Vegt, N.F.A., Müller-Plathe, F., Carbone, P.: How good are coarse-grained polymer models? A comparison for atactic polystyrene. ChemPhysChem **13**, 3428 (2012)
4. Voth, G.A.: Coarse-Graining of Condensed Phase and Biomolecular Systems, CRC edn. Taylor and Francis, Boca Raton (2008)
5. Carbone, P., Avendaño, C.: Coarse-grained methods for polymeric materials: Enthalpy and entropy driven models. Wiley Interdiscip. Rev.: Comput. Mol. Sci. **4**(1), 62–70 (2014)
6. Noid, W.G.: Perspective: Coarse-grained models for biomolecular systems. J. Chem. Phys. **7**, 139 (2013)
7. Lyubartsev, A.P., Laaksonen, A.: Calculation of the effective interaction potentials from radial distribution functions – a reverse Monte-Carlo approach. Phys. Rev. E **52**(4), 3730–3737 (1995)
8. Baschnagel, J., Binder, K., Doruker, P., Gusev, A.A., Hahn, O., Kremer, K., Mattice, W.L., Müller-Plathe, F., Murat, M., Paul, W., et al.: Bridging the gap between atomistic and coarse-grained models of polymers: Status and perspectives. Adv. Polym. Sci. **152**, 41–156 (2000)
9. Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., de Vries, A.H.: The MARTINI force field: Coarse grained model for biomolecular simulations. J. Phys. Chem. B **111**(27), 7812–7824 (2007)
10. Shinoda, W., Devane, R., Klein, M.L.: Multi-property fitting and parameterization of a coarse grained model for aqueous surfactants. Mol. Simul. **33**(1–2), 27–36 (2007)
11. Avendaño, C., Lafitte, T., Galindo, A., Adjiman, C.S., Jackson, G., Müller, E.A.: SAFT-gamma force field for the simulation of molecular fluids. 1. A single-site coarse grained model of Carbon Dioxide. J. Phys. Chem. B **115**(38), 11154–11169 (2011)
12. Ercolessi, F., Adams, J.B.: Interatomic potentials from 1st-principles calculations – the force-matching method. Europhys. Lett. **26**(8), 583–588 (1994)
13. Noid, W.G., Chu, J.W., Ayton, G.S., Krishna, V., Izvekov, S., Voth, G.A., Das, A., Andersen, H.C.: The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. J. Chem. Phys. **128**(24), 244114 (2008)

14. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**, 79 (1951)
15. Shell, M.S.: The relative entropy is fundamental to multiscale and inverse thermodynamic problems. J. Chem. Phys. **129**, 144108 (2008)
16. Groot, R.D., Warren, P.B.: Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. J. Chem. Phys. **107**, 4423–4435 (1997)
17. Espanol, P.: Dissipative particle dynamics. In: Handbook of Materials Modeling, pp. 2503–2512. Springer, Dordrecht (2005)
18. Altland, A., Simons, B.D.: Condensed Matter Field Theory, 2nd edn. Cambridge University Press, Leiden (2010)
19. Alexandridis, P., Lindman, B.: Amphiphilic Block Copolymers Self-Assembly and Applications. Elsevier, Amsterdam (2000). ISBN 978-0-444-82441-7
20. Ortiz, V., Nielsen, S.O., Discher, D.E., Klein, M.L., Lipowsky, R., Shillcock, J.: Dissipative particle dynamics simulations of polymersomes. J. Phys. Chem. B **109**, 17708–17714 (2005)
21. Nawaz, S., Carbone, P.: Coarse-graining poly(ethylene oxide)-poly(propylene oxide)-poly(ethylene oxide) (PEO-PPO-PEO) block copolymers using the MARTINI force field. J. Phys. Chem. B **118**, 1648–1659 (2014)
22. Rogers, R.D., Seddon, K.R.: Ionic Liquids, Science **302**, 792–793 (2003)
23. Wang, Y.T., Voth, G.A.: Unique spatial heterogeneity in ionic liquids. J. Am. Chem. Soc. **127**, 12192–12193 (2005)
24. Karimi-Varzaneh, H., Muller-Plathe, F., Balasubramanian, S., Carbone, P.: Studying long-time dynamics of imidazolium-based ionic liquids with a systematically coarse-grained model. Phys. Chem. Chem. Phys. **12**, 4714–4724 (2010)
25. Johnston, K., Harmandaris, V.: Hierarchical multiscale modeling of polymer–solid interfaces: atomistic to coarse-grained description and structural and conformational properties of polystyrene–gold systems. Macromolecules **46**, 5741 (2013)
26. Abrams, F.C., Delle Site, L., Kremer, K.: Dual-resolution coarse-grained simulation of the bisphenol-A-polycarbonate/nickel interface. Phys. Rev. E **67**, 21807 (2003)
27. Rzepiela, A.J., Louhivuori, M., Peter, C., Marrink, S. J.: Hybrid simulations: combining atomistic and coarse-grained force fields using virtual sites. Phys. Chem. Chem. Phys. **13**, 10437–10448 (2011)
28. Di Pasquale, N., Marchisio, D., Carbone, P.: Mixing atoms and coarse-grained beads in modelling polymer melts. J. Chem. Phys. **137**, 164111 (2012)
29. Praprotnik, M., Delle Site, L., Kremer, K.: Multiscale simulation of soft matter: from scale bridging to adaptive resolution. Annu. Rev. Phys. Chem. **59**, 545–571 (2008)
30. Müller, M., de Pablo, J.J.: Computational approaches for the dynamics of structure formation in self-assembling polymeric materials. Annu. Rev. Mater. Res. **43**, 1–34 (2013)
31. Daoulas, K.C., Müller, M., de Pablo, J.J., Nealey, P.F., Smith, G.D.: Morphology of multi-component polymer systems: single chain in mean field simulation studies. Soft Matter **2**, 573–583 (2006)
32. Milano, G., Kawakatsu, T.: Hybrid particle-field molecular dynamics simulations for dense polymer systems. J. Chem. Phys. **130**, 214106 (2009)

Coarse-Grained Molecular Dynamics

► Dissipative Particle Dynamics, Overview

Cochlea Implant

► Bioinspired CMOS Cochlea

Cold Field Electron Emission from Nanostructured Materials

► Field Electron Emission from Nanomaterials

Cold-Wall Thermal Chemical Vapor Deposition

► Chemical Vapor Deposition (CVD)

Compliant Mechanisms

Larry L. Howell
Department of Mechanical Engineering, Brigham Young University, Provo, UT, USA

Synonyms

Compliant systems; Flexures; Flexure mechanisms

Definition

Compliant mechanisms gain their motion from the deflection of elastic members.

Main Text

Compliant mechanisms offer an opportunity to achieve complex motions within the limitations of micro- and nano-fabrication. Because compliant mechanisms gain their motion from the constrained bending of flexible parts, they can achieve complex motion from simple topologies. Traditional mechanisms use rigid parts connected at articulating joints (such as hinges, axles, or bearings), which usually requires assembly of components and results in friction at the connecting surfaces [1–3]. Because traditional bearings are not practical and lubrication is problematic, friction and wear present major difficulties.

Nature provides an example of how to effectively address problems with motion at small scales. Most moving components in nature are flexible instead of stiff, and the motion comes from bending the flexible parts instead of rigid parts connected with hinges (for example, consider hearts, elephant trunks, and bee wings). The smaller the specimen, the more likely it is to use the deflection of flexible components to obtain its motion. And so it is with man-made systems as well, the smaller the device, the greater the advantages for using compliance [1].

Advantages of Compliant Mechanisms

Some of the advantages of compliant mechanisms at the micro- and nanoscales include the following:

Can be made from one layer of material. Compliant mechanisms can be fabricated from a single layer. This makes them compatible with many common microelectromechanical system (MEMS) fabrication methods, such as surface micromachining, bulk micromachining, and LIGA. For example, consider the folded beam suspension shown in Fig. 1. This device is often used as a suspension element in MEMS systems. It offers a simple approach for constrained linear motion, and also integrates a return spring function. The device can achieve large deflections with reasonable off-axis stiffness. The compliant mechanism makes it possible to do these functions with a single layer of material.

No assembly required. Compliant mechanisms that gain all of their motion from the deflection of flexible components are "fully compliant mechanisms," where devices that combine both traditional and compliant elements are called "partially compliant mechanisms." Fully compliant mechanism can usually be fabricated without assembly of different components.

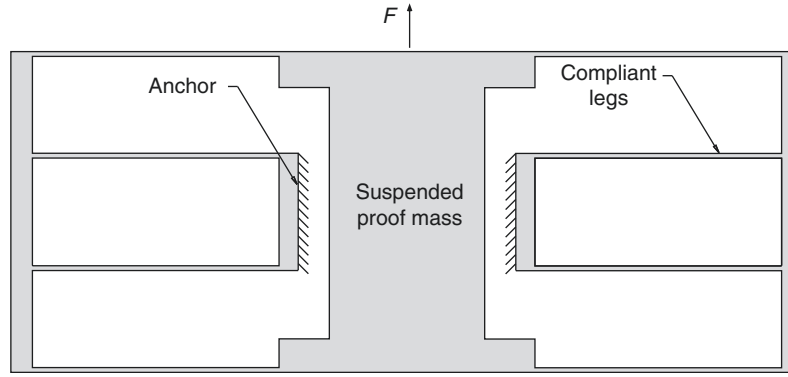
Small footprint. Some compliant mechanisms can also be designed to have a small footprint on the substrate on which they are built. Various strategies can be used to decrease the size of a mechanism. Figure 2 shows a thermal actuator that uses multiple layers to achieve a small footprint.

Friction-free motion. Because compliant mechanisms gain their motion from deflection of flexible members rather than from traditional articulating joints, it is possible to reduce or eliminate the friction associated with rubbing surfaces. This results in reduced wear and eliminates the need for lubrication, as described next.

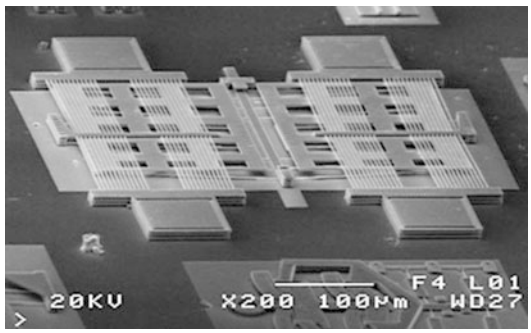
Wear-free motion. Wear can be particularly problematic at small scales, and the elimination of friction can result in the elimination of wear at the connecting surfaces of joints. For devices that are intended to undergo many cycles of motion, eliminating friction can dramatically increase the life of the system.

No need for lubrication. Another consequence of eliminating friction is that lubricants are not needed for the motion. This is particularly important at small scales where lubrication can be problematic.

High precision. Flexures have long been used in high precision instruments because of the repeatability of their motion. Some reasons for compliant mechanisms' precision are the



Compliant Mechanisms, Fig. 1 A folded-beam suspension is an example of a widely used compliant mechanism in microelectromechanical systems (MEMS) applications

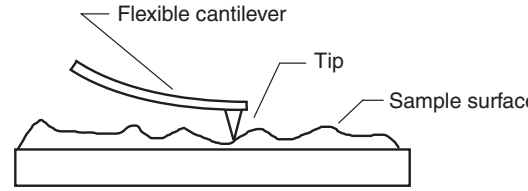


Compliant Mechanisms, Fig. 2 This scanning electron micrograph shows a thermal actuator that uses multiple layers of compliant elements to achieve large amplification with a small footprint

backlash-free motion inherent in compliant mechanisms and the wear-free and friction-free motion described above. The cantilever associated with an atomic force microscope (Fig. 3) is an example application.

Integrated functions. Like similar systems in nature, compliant mechanisms have the ability to integrate multiple functions into few components. For example, compliant mechanisms often provide both the motion function and a return-spring function. Thermal actuators are another example of integration of functions, as described later.

High reliability. The combination of highly constrained motion of compliant mechanisms, the relative purity of materials used in micro/nanofabrication, and wear-free motion result in high reliability of compliant mechanisms at the micro/nanoscale.



Compliant Mechanisms, Fig. 3 The cantilever of an atomic force microscope (AFM) is an example of compliance employed in high-precision instruments

Challenges of Compliant Mechanisms

Compliant mechanisms have many advantages, but they also have some significant challenges. A few of these are discussed below [1].

Limited rotation. One clear drawback of compliant mechanisms is the general inability to undergo continuous rotation. Also, if a fully compliant mechanism is constructed from a single layer of material, then special care has to be taken to ensure that moving segments of the compliant mechanism do not collide with other segments of the same mechanism.

Dependence on material properties. The performance of compliant mechanisms is highly dependent on the material properties, which are not always well known.

Nonlinear motions. The deflections experienced by compliant mechanisms often extend beyond the range of linearized beam equations. This can make their analysis and design more complicated.

Fatigue analysis. Because most compliant mechanisms undergo repeated loading, it is

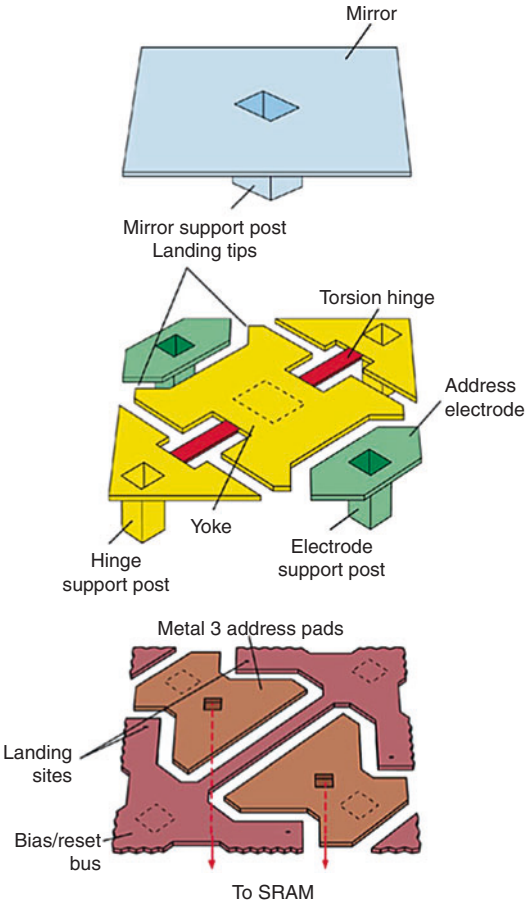
important to consider the fatigue life of the device. Interestingly, because of the types of materials used and their purity, many MEMS compliant mechanisms will either fail on their first loading cycle or will have infinite fatigue life. Because of the low inertia of MEMS devices, it is often easy to quickly test a MEMS device to many millions of cycles. Factors such as stress concentrations, the operating temperature, and other environment conditions can affect the fatigue life.

Difficult design. Integration of functions into fewer components, nonlinear displacements, dependence on material properties, the need to avoid self-collisions during motion, and designing for appropriate fatigue life, all combine to make the design of compliant mechanisms nontrivial and often difficult.

Example Applications of Compliant Mechanisms

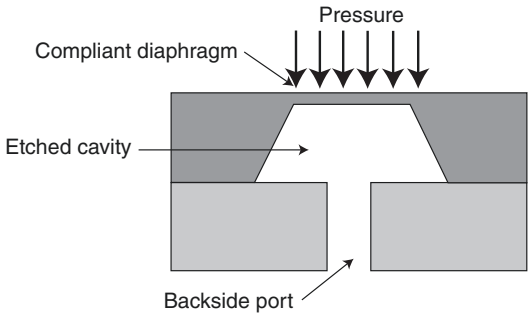
Examples of MEMS compliant mechanisms are shown here to further illustrate their properties and to demonstrate a few applications.

Digital Micromirrors. One of the most visible commercially available microelectromechanical systems is Texas Instruments' Digital Micromirror Device (DMD™), which is used in applications such as portable projectors. The DMD is a rectangular array of moving micromirrors that is combined with a light source, optics, and electronics to project high-quality color images. Figure 4 shows the architecture of a single DMD pixel. A 16-μm-square aluminum mirror is rigidly attached to a platform (the "yoke"). Flexible torsion hinges are used to connect the yoke to rigid posts. An applied voltage creates an electrostatic force that causes the mirror to rotate about the torsion hinges. When tilted in the on position, the mirror directs light from the light source to the projection optics and the pixel appears bright. When the mirror is tilted in the off position, the light is directed away from the projection optics and the pixel appears dark. The micromirrors can be combined in an array on a chip, and each micromirror is associated with the pixel of a projected image. The torsion hinges use compliance to obtain motion while avoiding rubbing parts that cause friction and wear. The hinges can be deflected thousands of times per second and infinite fatigue life is essential.

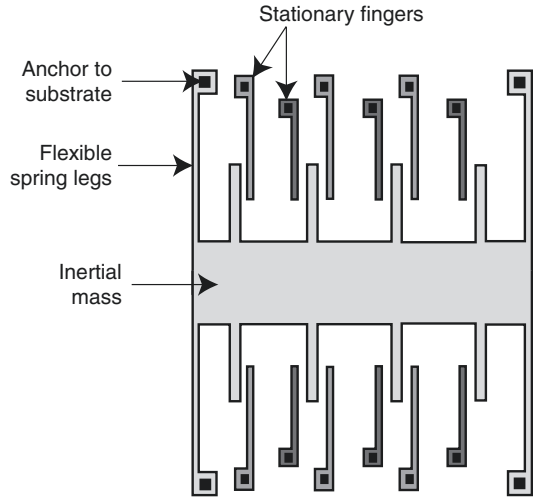


Compliant Mechanisms, Fig. 4 Texas Instrument's Digital Micromirror Device (DMD™) uses compliant torsion hinges to facilitate mirror motion (Illustration courtesy of Texas Instruments)

Piezoresistive pressure sensors. A sensor is a device that responds to a physical input (such as motion, radiation, heat, pressure, magnetic field), and transmits a resulting signal that is usually used for detection, measurement, or control. Advantages of MEMS sensors are their size and their ability to be more closely integrated with their associated electronics. Piezoresistive sensing methods are among the most commonly employed sensing methods in MEMS. Piezoresistance is the change in resistivity caused



Compliant Mechanisms, Fig. 5 The strain on a compliant diaphragm of a piezoresistive pressure sensors results in a detectable change in resistance, which is correlated with the pressure



Compliant Mechanisms, Fig. 6 This accelerometer makes use of compliant legs that deflect under inertial loads. The deflection results in a detectable change in capacitance and is correlated with the corresponding acceleration

by mechanical stresses applied to a material. Bulk micromachined pressure sensors have been commercially available since the 1970s. A typical design is illustrated in Fig. 5. A cavity is etched to create a compliant diaphragm that deflects under pressure. Piezoresistive elements on the diaphragm change resistance as the pressure increases; this change in resistance is measured and is correlated with the corresponding pressure.

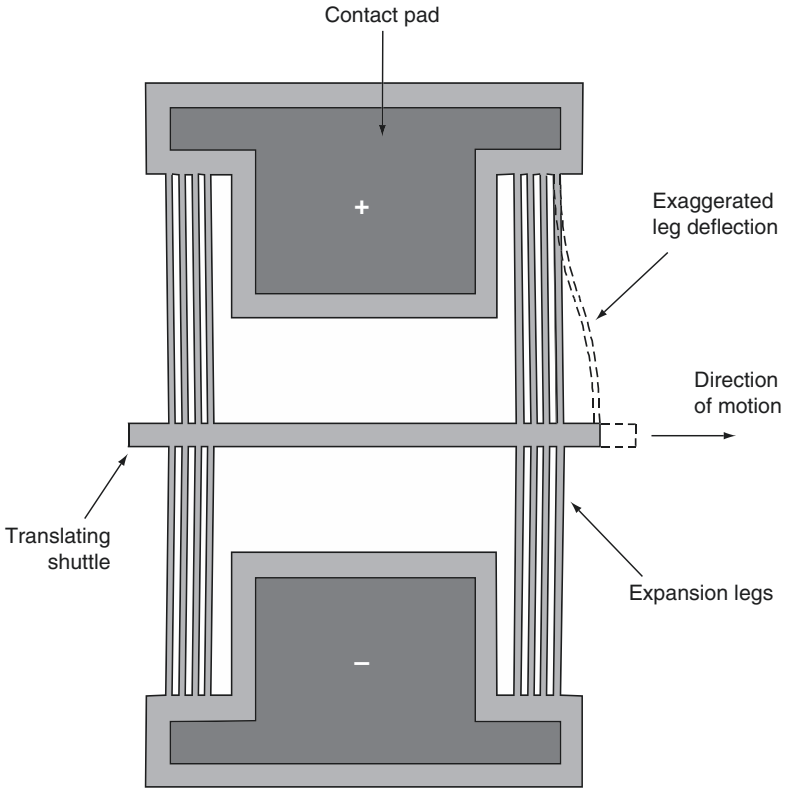
Capacitive acceleration sensors. Accelerometers are another example of commercially successful MEMS sensors. Applications include automotive airbag safety systems, mobile electronics, hard drive protection, gaming, and others. Figure 6 illustrates an example of a surface micromachined capacitive accelerometer. Acceleration causes a displacement of the inertial mass connected to the compliant suspension, and the capacitance change between the comb fingers is detected.

Thermal actuators. A change in temperature causes an object to undergo a change in length, where the change is proportional to the material's coefficient of thermal expansion [4]. This length change is usually too small to be useful in most actuation purposes. Therefore, compliant mechanisms can be used to amplify the displacement of thermal actuators. Figure 7 illustrates an example of using compliant mechanisms to amplify thermal expansion in microactuators. Figure 8 shows a scanning electron micrograph of a thermomechanical in-plane microactuator (TIM) illustrated in Fig. 7. It consists of thin legs connecting both sides of a center shuttle. The leg ends not connected to the shuttle are anchored to bond pads on the substrate and are fabricated at a slight angle to bias motion in the desired direction. As voltage is applied across the bond pads, electric current flows through the thin legs. The legs have a small cross-sectional area and thus have a high electrical resistance, which causes the legs to heat up as the current passes through them. The shuttle moves forward to accommodate the resulting thermal expansion. Advantages of this device include its ability to obtain high deflections and large forces, as well as its ability to provide a wide range of output forces by changing the number of legs in the design.

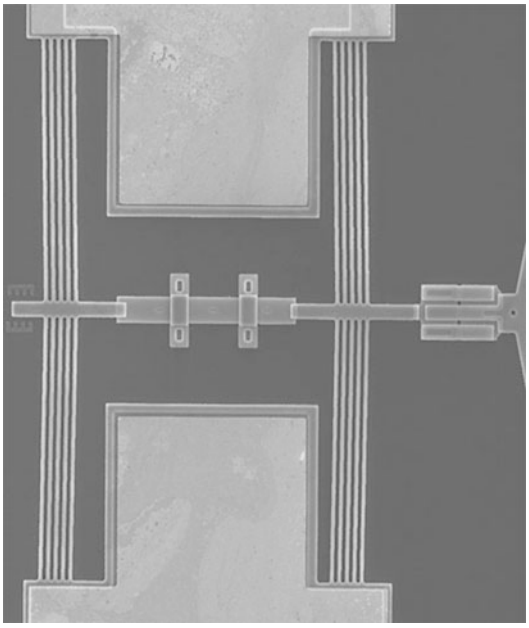
Analysis and Design of Compliant Mechanisms

Multiple approaches are available for the analysis and design of compliant mechanisms. Three of the most developed approaches are described below.

Finite element analysis. Finite element methods are the most powerful and general methods available to analyze compliant mechanisms.



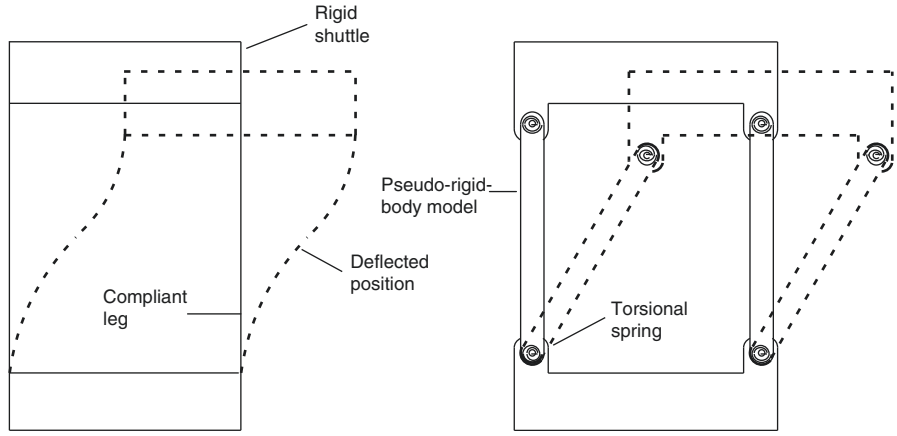
Compliant Mechanisms, Fig. 7 A schematic of a thermomechanical in-plane microactuor (TIM) that uses compliant expansion legs to amplify the motion caused by thermal expansion



Compliant Mechanisms, Fig. 8 A scanning electron micrograph of a thermal actuator illustrated in Fig. 7

Commercial software is currently available that has the capability of analyzing the large, nonlinear deflections often associated with compliant mechanisms. The general nature of the method makes it applicable for a wide range of geometries, materials, and applications. Increasingly powerful computational hardware has made it possible to analyze even very complex compliant mechanisms. It is also possible to use finite element methods in the design of compliant mechanisms, particularly once a preliminary design has been determined. But in the early phases of design, other methods (or hybrid methods) are often preferred so that many design iterations can be quickly analyzed.

Pseudo-rigid-body model. The pseudo-rigid-body model is used to model compliant mechanisms as traditional rigid-body mechanisms, which opens up the possibility of using the design and analysis methods developed for rigid-body mechanisms in the design of compliant mechanisms [1]. With the



Compliant Mechanisms, Fig. 9 The pseudo-rigid-body model of the compliant parallel-guiding mechanism consists of appropriately located pin joints and torsional springs (this device is a building block of other devices, such as the folded-beam suspension)

pseudo-rigid-body model approach, flexible parts are modeled as rigid links connected at appropriately placed pins, with springs to represent the compliant mechanism's resistance to motion. Extensive work has been done to develop pseudo-rigid-body models for a wide range of geometries and loading conditions. Consider a simple example. The micromechanism shown in Fig. 9 has a rigid shuttle that is guided by two flexible legs. (Note that the folded-beam suspension in Fig. 1 has four of these devices connected in series and parallel.) The pseudo-rigid-body model of the mechanism models the flexible legs as rigid links connected at pin joints with torsional springs. Using appropriately located joints and appropriately sized springs, this model is very accurate well into the nonlinear range. For example, if the flexible legs are single-walled carbon nanotubes, comparisons to molecular simulations have shown the pseudo-rigid-body model to provide accurate results [5]. The advantages of the pseudo-rigid-body model are realized during the early phases of design where many design iterations can be quickly evaluated, traditional mechanism design approaches can be employed, and motions can be easily visualized.

Topology optimization. Suppose that all that is known about a design is the desired performance and design domain. Topology optimization shows promise for designing compliant mechanisms under such conditions. The advantage is that very little prior knowledge about the resulting compliant mechanism is needed, and any biases of the designer are eliminated [6]. Topology optimization is often integrated with finite element methods to consider many possible ways of distributing material with the design domain. This has the potential to find designs that would not otherwise be discovered by other methods. Infinite possible topologies are possible and finite element methods can be employed to evaluate the different possibilities. The resolution of the design domain mesh can be a limiting factor, but once a desirable topology is identified, it can be further refined using other approaches.

Conclusion

Compliant mechanisms provide significant benefits for micro- and nano-motion applications. They can be compatible with many fabrication methods, do not require assembly, have friction-free and wear-free motion, provide high precision and high reliability, and they can integrate multiple functions into fewer components. The major challenges associated with compliant mechanisms come from the difficulty associated with their design, limited rotation, and the need to ensure adequate fatigue life. It is likely that compliant mechanisms will see increasing use in micro- and

nano-mechanical systems as more people understand their advantages and have tools available for their development.

Cross-References

- ▶ AFM
- ▶ Basic MEMS Actuators
- ▶ Biomimetics
- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Insect Flight and Micro Air Vehicles (MAVs)
- ▶ MEMS on Flexible Substrates
- ▶ Nanogrippers
- ▶ Piezoresistivity
- ▶ Thermal Actuators

References

1. Howell, L.L.: Compliant Mechanisms. Wiley, New York (2001)
2. Lobontiu, N.: Compliant Mechanisms: Design of Flexure Hinges. CRC Press, Boca Raton (2003)
3. Smith, S.T.: Flexures: Elements of Elastic Mechanisms. Taylor & Francis, London (2000)
4. Howell, L.L., McLain, T.W., Baker, M.S., Lott, C.D.: Techniques in the design of thermomechanical microactuators. In: Leondes, C.T. (ed.) MEMS/NEMS Handbook, Techniques and Applications, pp. 187–200. Springer, New York (2006)
5. Howell, L.L., DiBiasio, C.M., Cullinan, M.A., Panas, R., Culpepper, M.L.: A pseudo-rigid-body model for large deflections of fixed-clamped carbon nanotubes. J. Mech. Robot. **2**, 034501 (2010)
6. Frecker, M.I., Ananthasuresh, G.K., Nishiwaki, S., Kikuchi, N., Kota, S.: Topological synthesis of compliant mechanisms using multi-criteria optimization. J. Mech. Des. **119**, 238–245 (1997)

Compliant Systems

- ▶ Compliant Mechanisms

Composite Materials

- ▶ Theory of Optical Metamaterials

Computational Chemistry for Drug Discovery

Giulia Palermo[1,2] and Marco De Vivo[1]
[1]Department of Drug Discovery and Development - CompuNet, Istituto Italiano di Tecnologia, Genoa, Italy
[2]Laboratory of Computational Chemistry and Biochemistry, Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Synonyms

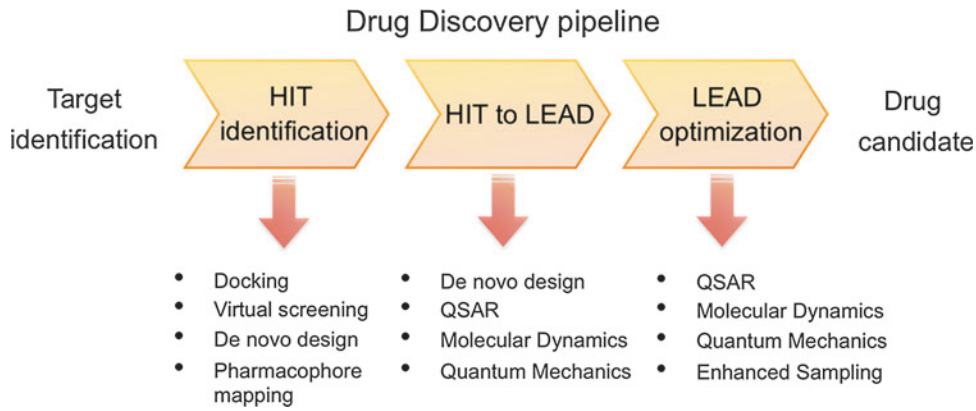
Drug design; Molecular modeling

Definition

Computational chemistry uses physics-based algorithms and computers to simulate chemical events and calculate chemical properties of atoms and molecules. In drug design and discovery, diverse computational chemistry approaches are used to calculate and predict events, such as the drug binding to its target and the chemical properties for designing potential new drugs.

Overview

Computational methods are nowadays routinely used to accelerate the long and costly drug discovery process. Typically, once the drug discovery target is selected, drug discovery activities are divided into those for (1) the *hit identification* phase, in which the aim is the identification of chemical compounds with a promising activity toward the target; (2) the *lead generation* phase, in which hit compounds are improved in potency against the target; and, finally, (3) the *lead optimization* phase, in which lead compounds are optimized, generating drug-like molecules ultimately able to exert their beneficial pharmacological effect in patients (Fig. 1). Computations can help in all these drug discovery activities, from drug target identification, commonly a receptor



Computational Chemistry for Drug Discovery, Fig. 1 Drug discovery pipeline, typically formed by three phases, namely, "hit identification," "hit to lead," and "lead optimization." Different computational approaches can be applied to each phase of the drug discovery pipeline

or an enzyme, to the design and optimization of a new drug-like compound. While computational methods for target identification rely mainly on computational sciences such as bioinformatics and computational genomics, different computational approaches are used once the target has been identified and the search for small molecule inhibitors has commenced, starting from the hit identification phase, moving toward the lead generation and optimization phases (Fig. 1). At that point, routinely applied computational chemistry approaches include methods for structure-based drug design (SBDD), when structural data of the target protein are available, and ligand-based drug design (LBDD), when structural information of the target is missing or not fully reliable. Overall, these methods facilitate the identification of promising chemical scaffolds that interfere favorably with the target's function, producing a positive pharmacological effect. Experimental biochemical and pharmacological data on the new compounds, such as their in vitro inhibitory potency and in vivo efficacy, can be used to check the computational predictions while also forming the basis upon which better models can be constructed, leading to the design of superior compounds [1].

The impact of computational chemistry on drug discovery has been intensified in the last few decades by the rapid development of faster architectures and better algorithms for time-affordable high-level computations. Several theoretical methods that were once prohibitive for effective drug discovery are now increasingly used for hit identification and lead generation. Recently, for example, CPU-intensive and GPU-based free energy perturbation (FEP) calculations have been applied to accurately estimate the binding free energy of closely related chemical analogs, generating very promising results for lead generation and optimization [2]. Also, classical molecular dynamics (MD) is currently proposed as a practical computational tool for studying the energetics and kinetics of a ligand binding to a target protein. This is relevant for lead optimization, representing a new frontier in computationally driven drug discovery. Recently, in fact, compounds have been evaluated not only for their ability to bind tightly to the target but also for their capacity to remain bound to the target for a long time (i.e., considering k_{on} and k_{off} of binding), increasing the chances of efficacy in vivo. Finally, in the last decade, quantum mechanics (QM) has become ever more accessible for performing SBDD for lead generation and optimization. For example, QM and hybrid QM/MM methods are increasingly used to study the interaction of covalent inhibitors with a drug discovery target.

It is challenging to tailor a perfect fit between a new compound and its target in order to generate potent inhibitory effects (i.e., high affinity), which is the goal during the lead generation phase.

However, there are several other challenges during the lead optimization phase. A potent inhibitor is not a drug. There are other physicochemical variables that dictate the pharmacokinetics (PK) of each compound, affecting their drug-likeness and, ultimately, their efficacy and safety in vivo. Absorption, distribution, metabolism, excretion, and toxicity (ADMET) are key parameters that need to be optimized to generate a drug candidate with good chances of success in clinical trials. ADMET prediction at early stages of the drug discovery process is key to preventing, or at least limiting, later failures in costly clinical trials. In this respect, computational methods for chemometrics and quantitative structure–activity relationship (QSAR) approaches play a prominent role in creating predictive models for selecting and prioritizing compounds, typically during the lead optimization phase.

Thus, each computational chemistry method can impact and accelerate a given phase of the drug discovery process, from docking and MD for hit identification and lead generation to QSAR for ADMET optimization (see Fig. 1). Detailed methodological descriptions of each method can be found in many excellent review articles and books that focus on the theoretical background of computational chemistry [3–5]. This essay aims instead to comprehensively outline the applicability of the computational chemistry methods and approaches used nowadays to accelerate drug design and discovery, with particular emphasis on SBDD. The point is to show how each computational approach suits better a certain phase of the drug discovery pipeline, from SBDD for the hit identification and lead generation phases to QSAR methods for lead optimization, where drug-like properties are tuned to generate a promising drug candidate. The everyday use of once-prohibitive computational methods, such as MD- and QM-based methods for SBDD, will also be highlighted.

Computational Methods for SBDD

Computational approaches to structure-based drug design (SBDD) rely on knowledge of the target protein structure, which is usually provided by high-resolution X-ray crystallography or NMR data. Through a detailed analysis of the interaction between the target structure and the (new) ligands, SBDD approaches allow informed decisions to be made in designing more potent (i.e., with high affinity for the target) and selective compounds. Therefore, these methods are mostly used for hit identification and during the hit-to-lead phase (Fig. 1). A wide range of computational chemistry approaches can be applied to SBDD, including force field-based methods such as molecular docking calculations, classical MD- or Monte Carlo (MC)-based simulations, and more sophisticated QM-based methods [1, 6]. Ultimately, the new rationally designed compounds must be experimentally evaluated to verify whether they are potent inhibitors. The experimental test of each compound demonstrates the interdisciplinary nature of effective drug discovery. It is also essential for establishing and evaluating the predictive power of computational approaches in identifying and generating promising drug candidates.

Force Field-Based Approaches for SBDD

Computational approaches based on molecular mechanics (MM) allow the energy and several properties of molecular systems to be computed [3]. The basic functional form of a force field describes the potential energy of the system, which is determined as the sum of different contributions that are parameterized to reproduce experimental or ab initio data. The interactions are divided into bonded interactions and nonbonded interactions. The typical force field equation (Chart 1) contains the terms for the bonded (highlighted in red) and nonbonded (highlighted in blue) interactions. In detail, the bonded terms describe the chemical bonds between two neighboring atoms, bond angles between three atoms, and dihedral angles between four atoms. Improper dihedral–angle terms can be additionally applied to maintain planar or tetrahedral conformations. The functional form of the bond and angle terms is

Computational Chemistry for Drug Discovery, Chart 1 Typical force field equation

$$V = \sum_{\text{bonds}} K_R(\mathrm{R} - \mathrm{R}_{eq})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{eq})^2 + \sum_{\text{dihedrals}} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j}\left[\frac{A_{ij}}{\mathrm{R}_{ij}^{12}} - \frac{B_{ij}}{\mathrm{R}_{ij}^{6}} + \frac{q_i q_j}{\varepsilon \mathrm{R}_{ij}}\right]$$

bonded interactions

non-bonded interactions

Lennard-Jones term (van der Waals interactions)

Coulomb term (electrostatic interactions)

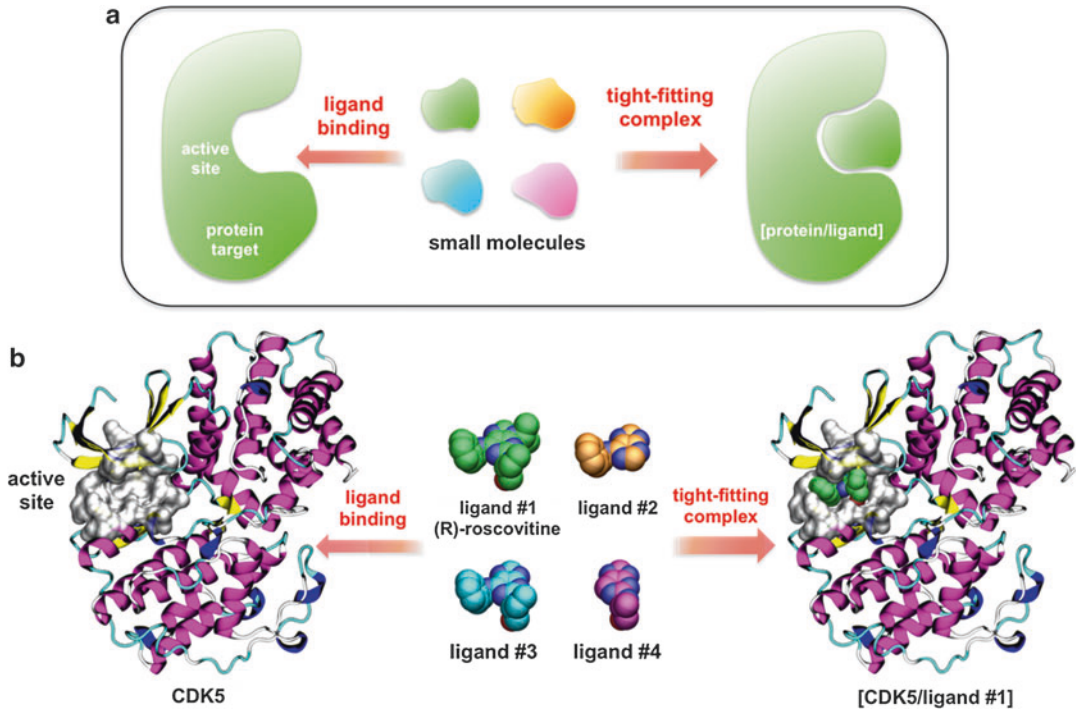
$$V_{vdW} = \sum_{i<j}\left[\frac{A_{ij}}{\mathrm{R}_{ij}^{12}} - \frac{B_{ij}}{\mathrm{R}_{ij}^{6}}\right] \quad + \quad V_{el} = \sum_{i<j}\left[\frac{q_i q_j}{\varepsilon \mathrm{R}_{ij}}\right]$$

quadratic, while the dihedral term uses a trigonometric function. In the typical force field equation, **R** is the distance between two atoms i and j that are bound together via a covalent bond, θ is the bond angle, $\mathbf{R}_{\text{eq}}$ and θ_{eq} refer to equilibrium bond lengths and angles, and K_R and K_θ are the vibrational constants. V_n is the torsional barrier corresponding to the nth barrier of a given torsional angle with phase γ. The last term of the typical force field equation refers to the nonbonded interactions, which are composed of a Lennard-Jones term for the van der Waals interactions and a Coulomb term for the electrostatic interactions between atoms i and j.

Force fields like AMBER, OPLS, GROMOS, and CHARMM are extensively used to study standard biomolecular systems such as protein, DNA, and RNA. Corrections and extensions are recurrently developed and reported for treatment of carbohydrates, lipids, ions, and other molecules of relevant biological interest, including drugs. Particularly relevant for drug design are the generalized AMBER force field (GAFF) and CHARMM general force field (CgFF), which provide suitable parameters for computations of drug-like organic molecules. The main advantage of using force field-based approaches is their relatively low computational cost compared to QM-based methods. For this reason, force field-based methods are by far the most applied in computational drug discovery. These include molecular docking and virtual screening of millions of novel compounds to identify new inhibitors, as well as de novo design and classical molecular dynamics simulations (see below).

Virtual Screening and Molecular Docking for Drug Discovery

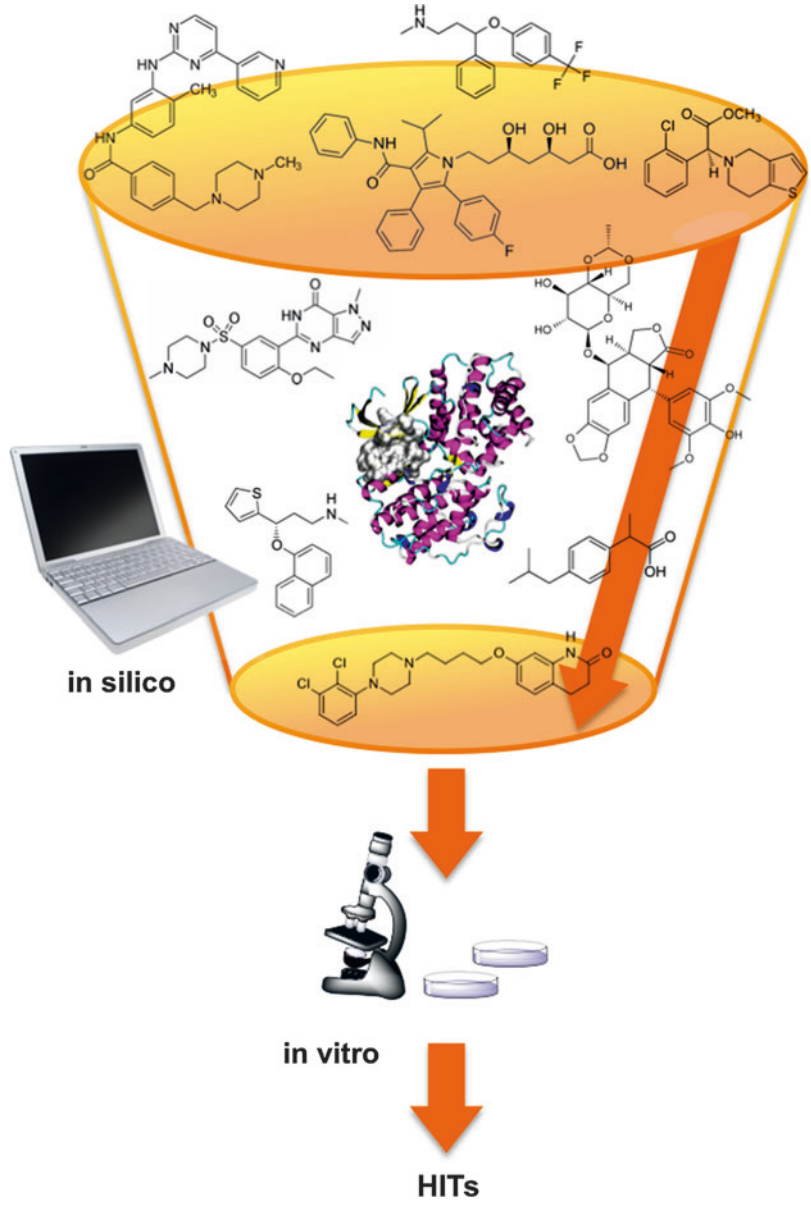
Virtual screening (VS) and molecular docking are broadly applied approaches in computational SBDD [1]. The ultimate goal of these methods is to identify compounds that dock well inside a targeted pocket (Fig. 2). These computational approaches are thus mostly used to identify promising hits and generate new leads. A compound able to fit into a target pocket would likely interfere with the target function, representing a promising chemical scaffold for the design of a new drug. GLIDE, GOLD, AUTODOCK, DOCK, and ICM-Dock are some of the more popular docking software. They are all used to increase the chances of identifying novel bioactive molecules from among the many small molecules that can be (virtually) screened. Starting from large compound collections, which can contain millions of different small molecules, VS protocols filter those collections and reduce them into smaller sets of active compounds with promising predicted activity against the target (Fig. 3). Usually, a library of hundreds of thousands or millions of virtual compounds is screened against a selected target. Then, a more sophisticated docking protocol, for example, a protocol that includes some receptor flexibility, can be used to rescore the best few thousand scored compounds. It is usually advisable to visually inspect the highly scored compounds, looking at each pose. This step reduces the risk of false-positive results, often induced by highly scored compounds that bind by adopting unlikely conformations or bind far from the pocket region of interest. After visual inspection, a few compounds are selected for



Computational Chemistry for Drug Discovery, Fig. 2 (**a**) Scheme of the ligand binding of a small molecule to the target proteins. (**b**) Binding of the tight-fitting ligand #1 [(R)-roscovitine] to the cyclin-dependent kinase-5 (CDK5) [23]. Other ligands (#2 to #4) are also shown. CDK5 is an attractive pharmacological target for drug discovery. CDK5 deregulation is implicated in many neurodegenerative processes such as Alzheimer's and Parkinson's diseases and amyotrophic lateral sclerosis. CDK5 is depicted with ribbons. The protein active site is represented in molecular surface, while a series of small ligand analogs with different chemical features – ligands #1–4 – are shown in space-filling representation. Different colors highlight the different characteristics of the four ligands

in vitro testing. VS is thus typically used for hit identification during drug discovery, where the goal is to identify new chemical scaffolds with promising, i.e., weak, inhibitory activity against the target. Hits are then used as a starting point to develop leads, i.e., more potent inhibitors (Fig. 1). Usually, hit compounds are experimentally identified as inhibitors characterized by IC_{50} (i.e., the concentration of inhibitor required to block 50 % of the target activity) around ~10–100 or higher microM activity, while lead compounds have IC_{50} in the low nanoM range.

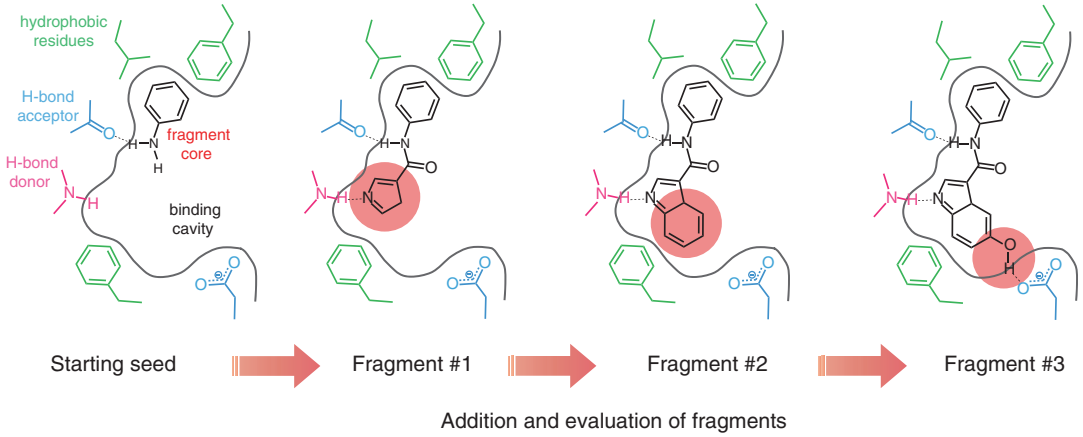
To predict and evaluate ligand–target interactions, molecular docking must correctly pose the ligand into the catalytic site and associate an energy score with each pose. There are two main steps during each docking calculation: the posing of the ligand into the pocket and the scoring of the final pose. To identify the possible binding poses, a conformational sampling of the ligand in the target active site is performed for each docking run. Each ligand pose is generated by exhaustive conformational searches performed by systematic or deterministic searches of rotatable bonds or by mapping the target/ligand geometric complementarities. More sophisticated search methods employ stochastic searches such as MC sampling procedures or evolutionary algorithms. Then, each final pose is scored, according to the chosen scoring function, and in VS, compounds are ranked according to their relative final energy score. Thus, a good docking pose is characterized by a favorable score, which should reflect and match experimental data, if available. The scoring step is thus as crucial as the posing step, where the ligand is accommodated into the binding site.



Computational Chemistry for Drug Discovery, Fig. 3 A typical virtual screening protocol starts with a large compound collection, which is virtually screened against a given target. Then, the most promising ligands are tested in vitro, leading to the identification of hit compounds (characterized by a concentration for 50 % target inhibition (IC_{50}) of ~100–10 μM). Ultimately, hit compounds will be transformed into lead molecules, which are more potent inhibitors that will be also tested in vivo for efficacy and safety

Although posing and scoring are often considered as independent problems, they are actually interconnected. This becomes clear when we consider that the primary criterion for the conformational selection of the bound ligand is its energy score, which is used to rank and select ligand poses among all the possible binding modes. Nowadays, there are several functions for scoring ligands, from force field-based scoring functions to knowledge-based potentials. Each scoring function has its own peculiarities and drawbacks. Care is therefore required in their selection and initial benchmarking, before embarking in a VS campaign. At times, multiple scoring functions are used to generate a consensus score, as the average or linear combination of single scoring function.

The challenges in achieving a correct score in VS and docking calculations include a proper treatment of electrostatics, as well as an exact



Computational Chemistry for Drug Discovery, Fig. 4 The de novo design of ligands is performed by building each compound within the active site. On the basis of an X-ray structure of the target protein, compounds are built within the binding cavity, using a number of fragments that can be joined together to create a new drug-like compound. Selected interaction centers are shown in green (hydrophobic residues), blue (H-bond acceptors), and magenta (H-bond donors). A starting fragment core (seed) is then grown into a bigger compound that can favorably interact with the target

evaluation of the desolvation energy of the ligand and binding pockets, and the explicit inclusion of entropic effects upon ligand binding. Recently, there has been intense investigation of the role of buried waters in the protein's pockets, which could act as a bridge for better ligand binding. A further challenge is that these methods consider a quasi-static complementarity between the ligand and its target. The restriction of target flexibility allows a limited induced fit upon binding but can potentially cause the miss of an important fraction of good hits/inhibitors during VS. To overcome this, recent docking algorithms and protocols allow multiple conformations of a few active-site residues that can be affected upon ligand binding, while other approaches try to solve the same issue via MD simulations or MC search [7]. Although rigid docking calculations are still preferred for VS of large compound libraries, flexible docking protocols are increasingly used nowadays, showing better predictivity than more static approaches.

Computational De Novo Design of Drug-Like Molecules

Computational de novo drug design means designing novel compounds from scratch, building them directly inside the binding site of the target macromolecule, piece by piece. As with docking and VS, the final goal is to explore chemical diversity and identify hit compounds that can fit well into the cavity of interest and likely interfere with the target function. De novo design programs include BOMB, BUILDER, GenStar, CONCERT, and several others, each with slightly different features and specifications [8]. The design process builds the new ligand directly within the binding pocket from an initial seed that can be as small as a hydrogen atom (Fig. 4). Then, ligand-growing and ligand-linking approaches are mainly used to transform one, or more, small fragments into a single drug-like compound with good predicted affinity for the target. With the ligand-growing method, functional groups are added to chemical handles of the core fragment, which is placed into the active site. Functional groups are usually available from an exhaustive list containing all sorts of aromatic and nonaromatic rings, aliphatic chains of varying length, etc. The user can therefore cherry pick the most interesting functional groups from the list and add those to the chemical handles of the core, generating new compounds. In contrast, ligand-linking methods use functional groups to

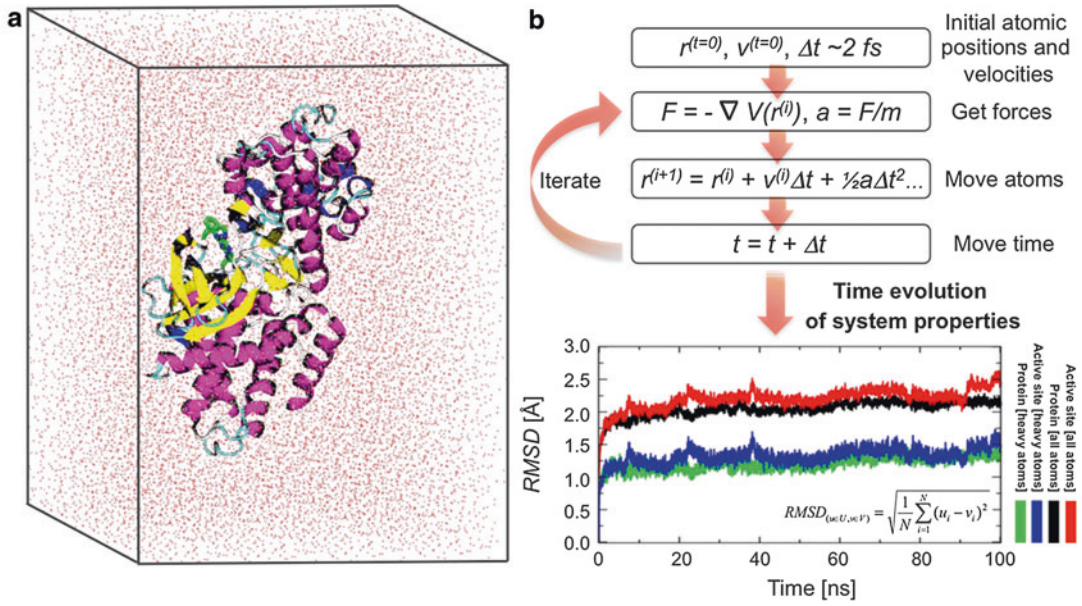
connect two or more fragments found or placed in the pocket, forming a single drug-like compound. In both procedures, the ligand construction is carried out by adding typical drug fragments that, once linked to the starting chemical core, are evaluated by considering several favorable conformers, usually followed by a geometry optimization of the newly formed compound within the target structure. Then, the obtained target/ligand complexes are scored in the same way and with the same limitations as scoring functions for VS and molecular docking.

The main advantage of de novo approaches, compared to docking and VS, is the possibility of exploring a virtually infinite search space and expanding chemical diversity. Drug-like molecules have been estimated to be in the order of 10^{60}–10^{100}. In this extremely large chemical space, the de novo design method allows the user to focus on chemical structures of interest, expanding on what is found in commercial libraries and converging on promising scaffolds that can be optimized ad hoc. The main risk is combinatorial explosion, due to the many combinations of functional groups and new compounds that could be constructed. In addition, the chemical space is limited by the synthesizability of the newly designed compounds, which might easily become too complex to be made in the lab. To overcome this, synthesizability constraints are often included in protocols for de novo computational design. Close collaboration between computational and synthetic chemists is also recommended. Ultimately, de novo molecular design methods help the hit identification phase by widening the chemical horizon for novel chemical scaffolds [2].

Molecular Dynamics for SBDD

Classical molecular dynamics (MD) simulations compute the time evolution of a system by applying Newtonian mechanics [3]. The method relies on force fields, which have been appropriately parameterized to describe biomolecular systems, such as the AMBER, OPLS-AA, and GROMOS force fields (see section above). These classical force fields provide a good description of the energetics of macromolecules such as proteins and nucleic acids. Also, in classical MD simulations of biological systems, there can be an explicit description of the water environment, with realistic molecular systems that can easily be constituted by a few hundred thousand atoms or more. In this respect, widely applied force fields for water molecules in biomolecular systems are the TIPnP models (Fig. 5).

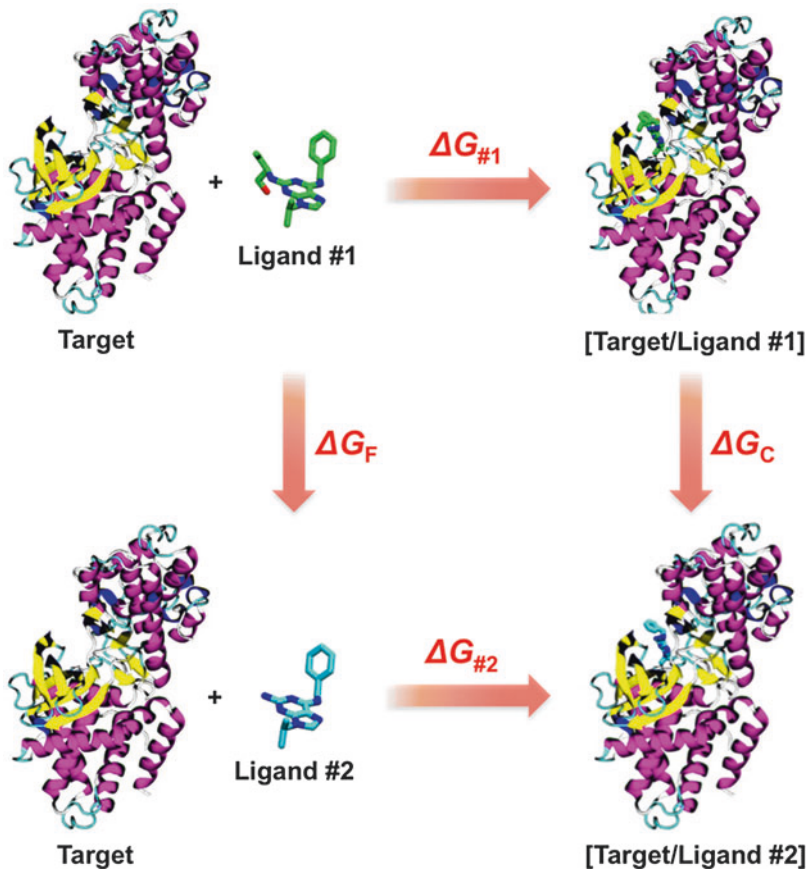
In drug design, classical MD simulations are mainly used to accurately calculate the binding free energy of the ligand to its target, explicitly taking into account the flexibility of the system, temperature, and entropic effects. MD is, however, much more computationally expensive than VS and docking. Therefore, MD is mostly recommended for, and so far applied to, lead optimization, where a promising compound is improved through small chemical modifications. MD simulations can suggest favorable chemical modifications through a detailed, although computationally demanding, understanding of the ligand–target interactions. MD simulations of realistic models of pharmaceutically relevant targets include the fatty acid amide hydrolase (FAAH) embedded into an explicit membrane environment [9, 10], the muscarinic acetylcholine receptors (mAChRs) and other GPCRs model systems bound to drugs, voltage-gated ion channels, and protein kinases, among others [11]. These types of studies elucidate the dynamical behavior of the ligand/target complex, highlighting key interactions that are likely responsible for drug binding and, ultimately, drug efficacy. In addition, the ability to simulate drug binding and unbinding events permits researchers to evaluate the residence time of the ligand into the binding pocket, which directly relates to the k_{on} and k_{off} of binding. The correct balance of these parameters is quite important in assuring a stable binding, which is at the basis of drug efficacy. Finally, the use of short MD runs as a post-processing and rescoring tool for docking results has been shown to improve the original docking score [7]. In this case, the docking binding pose relaxes, after introducing full flexibility into the system through MD, generating an optimized ligand/target complex that



Computational Chemistry for Drug Discovery, Fig. 5 (**a**) Typical MD model system, which shows a protein in explicit solvent. In this example, CDK5 is in complex with (R)-roscovitine, embedded in explicit water molecules [23]. (**b**) Scheme of MD simulation steps, which can provide information on the time evolution of certain selected system properties – e.g., the root mean square deviation (RMSD) expressed in Å of the overall protein and its active site, compared to the initial X-ray structure

leads to more accurate energy scores. Indeed, MD has lately been proposed, and preliminarily used, as a tool for screening medium-sized compound libraries (a few hundred or so compounds), with a protocol for dynamic docking that is expected to replace static VS in the near future [11].

Many biochemical events, such as diffusion events in the cell, can happen at long timescales spanning from micro- or milliseconds to seconds or even much longer. Those events are usually termed "rare" events. Often, these rare events require the overcoming of energetics barriers, which (in the case of ligand binding) are related to large protein conformational changes or ligand diffusion. That is, the biological phenomenon of ligand binding often consists of simulating a rare event. These simulations, and the accurate calculation of the associated free energy, require enhanced sampling methods, which can efficiently explore the free energy landscape of the rare biochemical event under investigation. These methods include umbrella sampling, steered MD, accelerated MD, thermodynamic integration, metadynamics, and others. They are often used in computational biophysics to study a ligand's binding/unbinding mechanism and other rare events [3]. A conventional method for the characterization of chemical events in computational biophysics is thermodynamics integration used to sample rare events by defining a path between the initial and final state. Then, thermodynamic parameters characterizing the system are changed gradually so that, at each stage, the system is in "constrained equilibrium." More simply, in the case of ligand binding, a suitable reaction coordinate can be chosen to discriminate between the bound and unbound states. Then, forces acting on this coordinate are monitored. The integration of the averaged forces along the binding path provides a good estimate of the free energy of the binding process. This method has been widely applied to numerous issues in biochemistry and biophysics,



Computational Chemistry for Drug Discovery, Fig. 6 Schematic representation of a thermodynamic cycle for the evaluation of the relative free binding energies of ligand #1 and ligand #2 to CDK5 [23]

including enzymatic chemical reactions [12–15], because it offers a rigorous evaluation of the free energy of the investigated process. With thermodynamic integration, one is essentially limited to compute the free energy landscape of one or two collective variables. In a more complex case, a more advanced method, e.g., metadynamics, is needed. Metadynamics allows the free energy landscape to be mapped by adding an external potential to the simulation, overcoming energetics barriers, and returning a thorough description of the explored free energy space. To restrict the virtually immense searchable configuration space, which is dictated by the degrees of freedom of the model systems, one must specify a restricted number of collective variables (i.e., the coordinates of interest) that properly describe the chemical event under study. In this way, for example, the path of the ligand, from the solvent to the target active site, can be simulated and examined, providing useful information on the dynamics and energetics of binding [16].

The free energy perturbation (FEP) method is a final example of free energy calculation methods that have successfully impacted SBDD. FEP calculations have demonstrated predictive power in calculating relative free binding energies of close analogs of active chemical scaffolds [2]. In this case, perturbations of each compound, which involve force field parameters, convert the initial ligand #1 into ligand #2, through a thermodynamic cycle (Fig. 6). The binding free energy difference between the ligands *#1* and *#2* is calculated as $\Delta\Delta G_{\text{bind}} = \Delta G_{\#1} - \Delta G_{\#2} = \Delta G_F - \Delta G_C$, where ΔG_F and ΔG_C are the energies associated with the unbound ligands in solution and with the target/ligand complexes, respectively. This allows the energetics of binding

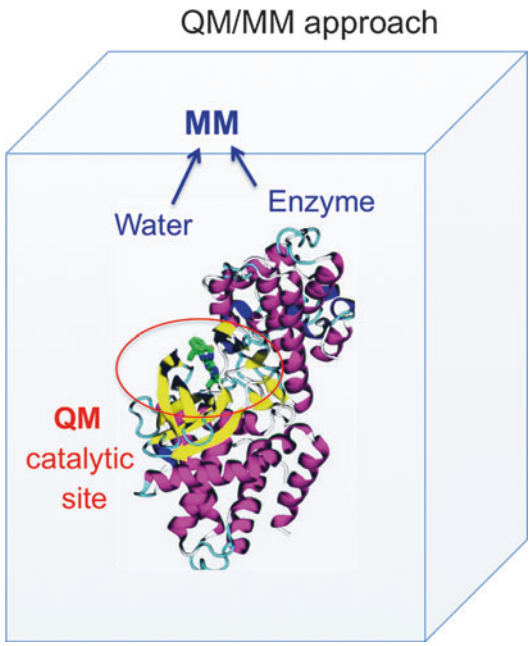
for the ligand *#1* and *#2* to be calculated, taking into full account the cost of the ligand's hydration. FEP theory has been broadly used for computational drug design purposes in combination with a Monte Carlo (MC) conformational search [17]. In this approach, a stochastic algorithm is used to randomly change the ligand within the target's active site. These random changes can also involve translational or rotational degrees of freedom of the ligand and of the residue side chains of the target. Whether a step is accepted or rejected is decided based on the Metropolis criterion, which generally accepts steps that lower the overall energy and occasionally accepts steps that increase energy. This prevents the system from getting stuck in local energy minima, allowing hopping over the energy barriers. As a drawback, an MC search lacks information about the timescale of the binding process.

All these methods and approaches can provide detailed mechanistic insights for discriminating between a few selected compounds that have successfully passed previous screening tests. Importantly, the continual development of more powerful processors and algorithms is allowing systems of increasing size and complexity to be studied. Protein folding, drug binding, membrane transport, and large conformational changes of proteins can nowadays be observed by performing long-timescale simulations, up to tens of microseconds, and more. For example, access to longer timescales has encouraged the study of binding kinetics. Calculations of the association and dissociation constants for the target/ligand complex – i.e., K_{on} and K_{off}, respectively – provide important information on the (ir)reversibility of the drug binding process, helping in the lead optimization phase. Overall, *state-of-the-art* MD simulations are nowadays a powerful SBDD tool.

Quantum Mechanics for SBDD

Quantum mechanics (QM) methods are the most advanced and computationally demanding methodologies, which are used to characterize the structure, dynamics, reactivity, and energetics of molecules. QM methods can be mainly classified as ab initio and density functional theory (DFT) methods. Ab initio QM methods aim to solve the Schrödinger equation, dealing with the wave function of the system. A variety of ab initio schemes have been proposed, from the simplest Hartree–Fock (HF) theory to the more sophisticated Møller–Plesset (MP) perturbation theory that includes correlation effects and applies a perturbation to the HF solution. In DFT, the central idea, proposed by the theorem of Hohenberg and Kohn in 1964, is that the ground-state energy of a system of interacting electrons is a unique functional of its electron density. Nowadays, calculations at the DFT level of theory are the method of choice for studying the electronic structure of realistic biochemical model systems, due to a rather balanced equilibrium between accuracy and computational affordability. DFT-based approaches in drug design are used to look at system properties that strongly depend on electronic structure, such as electrostatic potential maps used to characterize the binding site of the target or QM calculations to define the most probable protonation state of key residues located in the binding site. QM-derived scoring functions for better affinity evaluations in docking calculations is another field of application of QM for drug design. QM is also crucial for developing MM force fields to describe nonstandard residues, such as drug-like molecules [18].

The most common SBDD use of QM is the study of the catalytic and inhibitory mechanisms of pharmaceutically relevant enzymes. An atomic-level understanding of the inhibitory mechanism and its energetics can provide important information for improving existing drugs and designing new and more potent inhibitors. Indeed, QM-based investigation of enzymatic reaction mechanisms are linked to drug discovery through the design of inhibitors that resemble the structure and physicochemical properties of the enzymatic transition state (TS). This is because QM methods allow the study of bond-forming/breaking reactions, which characterize enzymatic catalysis and covalent inhibition. During new bond-forming events, polarization and charge transfer effects



Computational Chemistry for Drug Discovery, Fig. 7 QM/MM approach. A small part of the enzyme is treated at the QM level, while the remaining part of the system, i.e., water and protein, is treated at the MM level of theory

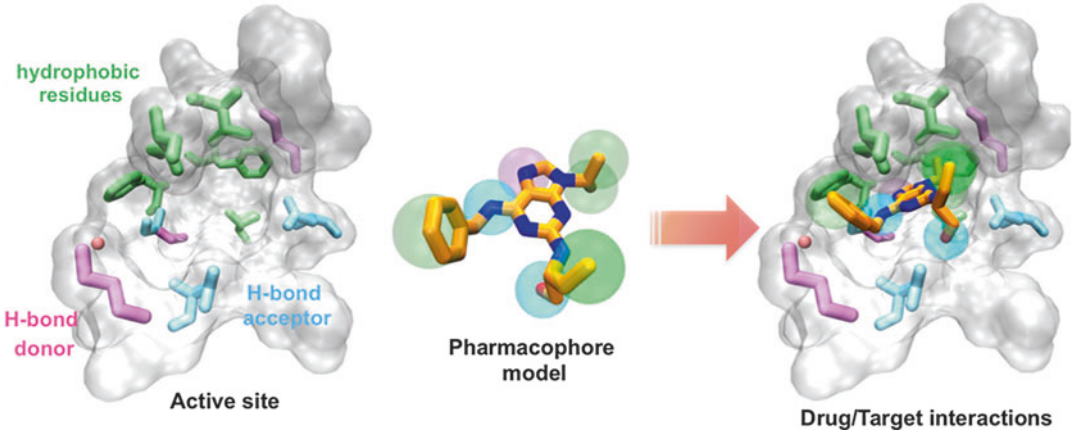
between the ligand and the active-site residues can be captured fully only via an explicit description of the electronic structure. Thus, QM-based methodologies allow the catalytic and inhibitory mechanisms to be deciphered in great detail. This information can ultimately facilitate and guide the design of better inhibitors, such as potent TS analogs [18].

Nevertheless, the applicability of most QM methods is limited to studying relatively small molecular systems in the gas phase, while biological systems in a realistic solution environment usually comprise thousands of atoms (~20,000 to ~100,000 atoms c.a.). This picture becomes even more complicated when dealing with membrane proteins, where the lipid bilayer can exert a role for drug binding. In this case, the model system can consist of ~250,000 atoms or more, making the use of QM prohibitive [15, 19]. To overcome this, QM methods are usually paired with classical MM approaches in the hybrid QM/MM scheme (Fig. 7). This allows the crucial part of the system in which the reaction takes place (the protein active site and the ligand) to be treated at the QM level, while the rest of the system is described with a classical MM force field. Since its first appearance in 1976, the QM/MM approach has become increasingly popular in studying enzymatic reaction mechanisms [20]. It has been of paramount importance in modern computational chemistry for allowing the multiscale modeling of complex chemical systems. The scientific impact of the QM/MM approach was acknowledged with the 2013 Nobel Prize in Chemistry.

For studying biological systems, the QM/MM scheme has been successfully coupled to ab initio MD methods, such as Car–Parrinello and Born–Oppenheimer MD. In this way, model systems of a few hundred atoms (i.e., the protein active site and the ligand) are simulated at their actual temperature, thus including entropic effects, during picosecond timescale trajectories. Of several different schemes for first-principles-based dynamics, the Car–Parrinello method has been widely used to study numerous pharmaceutically relevant targets, including phosphatases, hydrolases, and lactamases [6, 10]. Ab initio MD permits an atomic-level understanding of catalytic and inhibitory mechanisms in biological systems, providing results that can be directly compared with experiments. The practical impact of these high-level computations seems more applicable to the lead optimization phase, while a much broader applicability of QM-based methods is foreseen in SBDD over the next decade [18, 20].

Ligand-Based Drug Design Approaches

Ligand-based drug design (LBDD) approaches are applied when the three-dimensional structure of the target is unknown or when it is not fully reliable (i.e., homology models of the target are based on poor sequence identity) [1]. In this case, the experimental activity data of active compounds are generally used to construct pharmacophore models. These models include an ensemble of steric and electronic features



Computational Chemistry for Drug Discovery, Fig. 8 An example of a pharmacophore model of (R)-roscovitine, which tightly binds CDK5 [23]. Crucial interacting residues of CDK5 are shown in *green* (hydrophobic residues), *blue* (H-bond acceptors), and *magenta* (H-bond donors). The regions of the ligand that match those types of interaction are indicated with the same color

(usually indicated as descriptors) that are likely to be those mainly responsible for activity (Fig. 8). Typical descriptors are physicochemical features such as molecular weight, geometry, surface accessible area, aromaticity index, electronegativity, polarizability, and solvation properties. In general, a proper set of descriptors should cover a broad chemical diversity space. In this way, the more relevant descriptors for building predictive pharmacophores are likely to be recognized. These are then used to help identify new active compounds, for example, in a virtual library of drug-like molecules. The same principles are valid for molecular similarity methods, where similitudes in geometrical or physicochemical properties among active compounds are used to identify new active compounds from among the many that can be virtually screened. A similarity coefficient, such as the Tanimoto or Euclidean coefficients, is then used to identify whether a set of new compounds is likely to interact with a given target [5].

Geometrical and electronic descriptors are used to build a quantitative structure–activity relationship (QSAR), which results in a mathematical model able to predict the biological activity of new compounds [5]. For example, activity data of active molecules can be used to extract mathematical models for early prediction of activity or, more often, metabolic properties that could generate toxicity problems. Conventional 3D QSAR models consider the ensemble of conformations, orientations, tautomers, stereoisomers, and protonation states of the initial ligand set. More exotic multidimensional 4D or 5D QSAR models have also been developed, which take into account all energy contributions of ligand binding, including solvation energy and conformational entropy. Usually, linear regression and principal component analysis can be combined for a partial least square analysis that directly relates the reduced set of descriptors to the biological activity. However, a nonlinear relation is often present between the biological activity and the descriptors used. Therefore, nonlinear regression models using machine-learning algorithms have recently been introduced in QSAR. Of these, artificial neural networks algorithms are used to discover the relationship between descriptors and biological activity via an iterative process [21].

Absorption, distribution, metabolism, excretion, and toxicity (ADMET) are the key parameters that can be optimized with QSAR approaches to get a drug candidate with a proper drug-like pharmacokinetic (PK) profile. ADMET and PK

properties are critical to the success of any drug discovery program. A poor drug-like PK profile can result in the rapid metabolism and elimination of the compound from the body. A drug-like compound should be characterized by certain physicochemical properties, which are loosely recapitulated in Lipinski's rule of five. This says that an optimal orally bioavailable drug should have a molecular weight less than 500, less than 5 H-bond donor sites, and less than 10 H-bond acceptor sites, while the log of the octanol/water partition coefficient (logP), a measure of hydrophobicity, should be below 5. Despite its practical utility for quickly evaluating drug-likeness, it is nowadays clear that this rule can only estimate the compound's probable success by highlighting potentially problematic aspects of its physicochemical and structural profile [22]. In fact, many approved drugs, such as large macrolide antibiotics (MW $\sim$ 1000), differ in one or more descriptors from Lipinski's rule. Other important features can heavily affect the PK profile. For ideal absorption and distribution, an aqueous solubility above 10^{-6} M is advised (logS > -6). Transport properties, such as membrane permeability and brain/blood partitioning (log BB), for allowing blood–brain barrier penetration must also be taken into account in certain drug discovery projects. Additionally, interactions with influx/efflux transporter proteins or metabolic enzymes can greatly affect the final therapeutic effect and should be considered in the early phases of a drug discovery program. That is, for a given molecular structure, properties such as solubility, membrane permeability, partition coefficients, blood–brain barrier penetration, plasma protein binding, and metabolite formation are computed and compared against a database of known drugs. Computational tools and methods for predicting ADMET properties and producing predictive statistical models are thus valuable. These statistical models are trained on experimental ADMET data obtained from several tested compounds. These computational tools are therefore crucial for selecting, prioritizing, and optimizing promising lead compounds.

Conclusions and Perspective

This essay describes some key aspects of the most effective computational chemistry methods for accelerating drug discovery, showing how different challenges can be tackled with these methods. Each computational method is suited to a particular drug discovery phase (see Fig. 1). Indeed, over the last two decades, it has become increasingly evident that computational methods can accelerate the identification, design, and optimization of small molecules as promising drug candidates. SBDD approaches suit better the hit identification and lead generation phase, where new active small molecules are designed, identified, and transformed into potent inhibitors. QSAR methods are usually better for lead optimization, where drug-like properties are tuned to increase efficacy and lower attrition in drug development. In addition to conventional methods (such as virtual screening and QSAR), methods that were once prohibitive for effective drug design, such as molecular dynamics simulations and quantum-mechanics-based methods, are nowadays frequently used in all phases of the drug discovery pipeline, broadening drug discovery's computational arsenal. In fact, the advent of faster algorithms and better computer hardware will allow more extensive use of molecular dynamics and first-principles-based simulations. We therefore envisage that computational chemistry methods will have an even greater impact on drug discovery in the future years.

References

1. Jorgensen, W.L.: The many roles of computation in drug discovery. Science **303**, 1813–1818 (2004)
2. Jorgensen, W.L.: Efficient drug lead discovery and optimization. Acc. Chem. Res. **42**, 724–733 (2009)
3. Frenkel, D., Smit, B.: Understanding Molecular Simulation. Academic, San Diego (2002)
4. Jensen, F.: An Introduction to Computational Chemistry. Wiley, Chichester, England (1998)
5. Leach, A.R.: Molecular Modelling: Principles and Applications, 2nd edn. Prentice Hall, New York (2001)
6. Carloni, P., Rothlisberger, U., Parrinello, M.: The role and perspective of ab initio molecular dynamics in the

study of biological systems. Acc. Chem. Res. **35**, 455–464 (2002)
7. Nichols, S.E., Baron, R., Ivetac, A., McCammon, J.A.: Predictive power of molecular dynamics receptor structures in virtual screening. J. Chem. Inf. Model. **51**, 1439–1446 (2011)
8. Schneider, G., Fechner, U.: Computer-based de novo design of drug-like molecules. Nat. Rev. Drug Discov. **4**, 649–663 (2005)
9. Palermo, G., Branduardi, D., Masetti, M., Lodola, A., Mor, M., Piomelli, D., Cavalli, A., De Vivo, M.: Covalent inhibitors of fatty acid amide hydrolase: a rationale for the activity of piperidine and piperazine aryl ureas. J. Med. Chem. **54**, 6612–6623 (2011)
10. Palermo, G., Rothlisberger, U., Cavalli, A., Vivo, M. D.: Computational insights into function and inhibition of fatty acid amide hydrolase. Eur. J. Med. Chem. **91**, 15–26 (2015)
11. Borhani, D.W., Shaw, D.E.: The future of molecular dynamics simulations in drug discovery. J. Comput. Aided Mol. Des. **26**, 15–26 (2012)
12. De Vivo, M., Dal Peraro, M., Klein, M.L.: Phosphodiester cleavage in ribonuclease H occurs via an associative two-metal-aided catalytic mechanism. J. Am. Chem. Soc. **130**, 10955–10962 (2008)
13. De Vivo, M., Ensing, B., Dal Peraro, M., Gomez, G. A., Christianson, D.W., Klein, M.L.: Proton shuttles and phosphatase activity in soluble epoxide hydrolase. J. Am. Chem. Soc. **129**, 387–394 (2007)
14. De Vivo, M., Ensing, B., Klein, M.L.: Computational study of phosphatase activity in soluble epoxide hydrolase: high efficiency through a water bridge mediated proton shuttle. J. Am. Chem. Soc. **127**, 11226–11227 (2005)
15. Palermo, G., Campomanes, P., Cavalli, A., Rothlisberger, U., De Vivo, M.: Anandamide hydrolysis in FAAH reveals a dual strategy for efficient enzyme-assisted amide bond cleavage via nitrogen inversion. J. Phys. Chem. B.**119**(3), 789–801 (2015)
16. Ensing, B., De Vivo, M., Liu, Z.W., Moore, P., Klein, M.L.: Metadynamics as a tool for exploring free energy landscapes of chemical reactions. Acc. Chem. Res. **39**, 73–81 (2006)
17. Jorgensen, W.L., Thomas, L.L.: Perspective on free-energy perturbation calculations for chemical equilibria. J. Chem. Theory Comput. **4**, 869–876 (2008)
18. De Vivo, M.: Bridging quantum mechanics and structure-based drug design. Front. Biosci. Landmrk. **16**, 1619–1633 (2011)
19. Palermo, G., Stenta, M., Cavalli, A., Dal Peraro, M., De Vivo, M.: Molecular simulations highlight the role of metals in catalysis and inhibition of type II Topoisomerase. J. Chem. Theory Comput. **9**, 857–862 (2013)
20. Lodola, A., De Vivo, M.: The increasing role of QM/MM in drug discovery. Adv. Protein Chem. Struct. Biol. **87**, 337–662 (2012)
21. Nantasenamat, C., Isarankura-Na-Ayudhya, C., Prachayasittikul, V.: Advances in computational methods to predict the biological activity of compounds. Expert Opin. Drug Discov. **5**, 633–654 (2010)
22. Lipinski, C.: Chris Lipinski. Interview by Peter Kirkpatrick. Nat. Rev. Drug Discov. **11**, 900–901 (2012)
23. Mapelli, M., Massimiliano, L., Crovace, C., Seeliger, M.A., Tsai, L.H., Meijer, L., Musacchio, A.: Mechanism of CDK5/p25 binding by CDK inhibitors. J. Med. Chem. **48**, 671–679 (2005)

Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering

A. T. Conlisk
Department of Mechanical Engineering, The Ohio State University, Columbus, OH, USA

Synonyms

Microscale fluid mechanics; Nanoscale fluid mechanics

Definition

Because of the small scale of the fluid conduits, electric fields must often be used to transport fluids especially at the nanoscale. This means that the fluids must be electrically conducting, and so microfluidics and nanofluidics require the user to be knowledgeable in fluid mechanics, heat and mass transfer, electrostatics, electrokinetics, electrochemistry, and, if biomolecules are involved, molecular biology.

Introduction

The term microfluidics refers generally to internal flow in a tube or channel whose smallest dimension is under 100 μm. Nanofluidics refers to the

same phenomenon in a conduit whose smallest dimension is less than 100 nm.

Microchannels and nanochannels have large surface-to-volume ratio, so that surface properties become enormously important. In fully developed channel flow, the pressure drop $\Delta p \sim \frac{1}{h^3}$, where h is the small dimension, and so the pressure drop is prohibitively large for a nanoscale channel. Thus a solvent fluid such as water, proteins and other biomolecules, and other colloidal particles are most often transported electrokinetically. This means that the art of designing micro- and nanodevices requires a significant amount of knowledge of fluid flow and mass transfer (biofluids are usually multicomponent mixtures) and often heat transfer, electrostatics, electrokinetics, electrochemistry, and molecular biology. Details of the character of these fields of study can be found in the book by the author [1].

The study of micro-/nanofluidics requires knowledge of all of the abovementioned fields and so has a unifying effect. Moreover, nanofluidics opens the door to the discovery of the structure and conformation of biomaterials such as proteins and polysaccharides through molecular simulation.

In dealing with devices with small-scale features, microscale and below, there are three activities that normally comprise the design process; these are the following:

- Modeling: computational and theoretical
- Fabrication
- Experimental methods

Modeling is often done prior to the fabrication process as a guide as to what can be done. Experimental methods are usually used to assess the performance of a device, among other purposes.

Surface-to-Volume Ratio

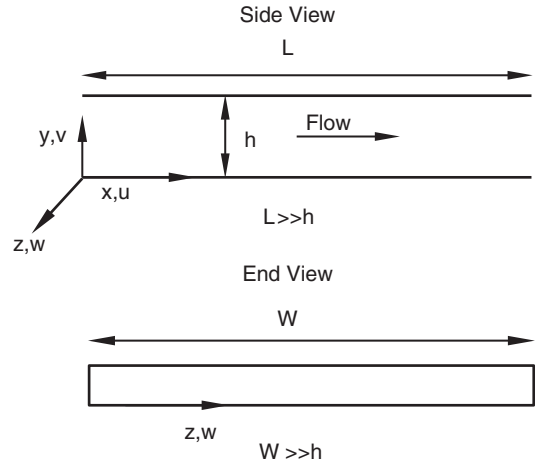
Consider a channel of rectangular cross section having dimensions in the (x, y, z) coordinate system of (L, h, W) with the primary direction of fluid motion being in the x direction. Then the surface-to-volume ratio is given

$$\frac{S}{V} = 2\left(\frac{1}{L} + \frac{1}{h} + \frac{1}{W}\right) = 6m^{-1} \tag{1}$$

for a channel having all three dimensions $L = h = W = 1m$. On the other hand, for a channel having dimensions (10 μm, 10^{-2} μm, 10 μm),

$$\frac{S}{V} \sim 2 \times 10^{8} m^{-1} \tag{2}$$

This means that surface properties become very important at the microscale and nanoscale and surfaces are routinely engineered to achieve a desired objective. In most devices the nanoscale features interface directly with microscale features. A typical channel geometry is depicted in Fig. 1.



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 1 Geometry of a typical channel. In applications $h \ll W, L$, where W is the width of the channel and L is the length in the primary flow direction. u, v, and w are the fluid velocities in the x, y, and z directions

Fluid Mechanics

Micro- and nanofluidics generally involve the flow of electrically conducting fluids, *electrolyte solutions*, that are assumed to be incompressible, having a constant density. Generally, the flows are internal, bounded on each side by walls, and are assumed to be fully developed. In this case, referring to Fig. 1, the governing equation for the velocity u in a channel is given by

$$\mu \frac{\partial^2 u}{\partial y^2} = \frac{\partial p}{\partial x} - B_x \tag{3}$$

where p is the pressure and B_x is a body force. The no-slip condition is applied at each wall: $u = 0$ at $y = 0, h$.

Mass Transfer

The molar flux of species A for a dilute electrically conducting mixture is

$$\vec{N}_A = -D_{AB}\nabla_{CA} + m_A z_A c_A \vec{E} + c_A \vec{V} \tag{4}$$

Here D_{AB} is the diffusion coefficient, R is the universal gas constant, T is the temperature, $z_A m_A$ is called the ionic mobility with $m_A = \frac{FD_{AB}}{RT}$, z_A is the valence, $F = 96500 \frac{Coul}{mole}$ is Faraday's constant, and $\vec{E}$ is the electric field. Equation 4 is called the Nernst-Planck equation, and the electric field term in the flux equation is called *electrical migration*. The boundary condition of interest here is that the solid walls in Fig. 1 are impermeable to species A, or $N_{A_y} = 0$ at $y = 0, h$.

In the absence of a velocity field $\vec{V}$ and in one dimension, Eq. 4 can be integrated to give

$$c_A = c_{A0} e^{-zA\phi} \tag{5}$$

and this is termed the *Boltzmann distribution* for the concentration of species A.

The Electric Field

An electric field is set up around any charged body and is defined as the force per charge on a surface. Electrical charges are either positive or negative, and like charges repel and opposite charges attract. For two bodies of charge q and q', the *electric field* is defined by

$$E = \frac{F}{q'} = \frac{q}{4\pi\epsilon_e r^2} \quad \frac{N}{C} \tag{6}$$

and is directed outward from the body of charge q and toward the body having a charge q' if $q > 0$, and the electric field is in the opposite direction if $q < 0$. In general, the electric field is a vector. This formula is called *Coulomb's Law*, and ϵ_e is called the *electrical permittivity.* The electrical permittivity is a transport property like the viscosity and thermal conductivity of a fluid.

The electric field due to a flat wall having a surface charge density σ in $\frac{Coulomb}{m^2}$ on one side is directed normal to the surface and has magnitude

$$E = \frac{\sigma}{2\epsilon_e} \tag{7}$$

A wire is characterized as having a line charge density, and if charges are distributed over a volume, a volume charge density is defined and called ρ_e in $\frac{Coulomb}{m^3}$.

The *electrical potential* is defined as the work done in moving a unit of charge, and mathematically

$$\phi = -\int_a^b \vec{E} \bullet d\vec{s} \tag{8}$$

The units of the electric potential are $\frac{Nm}{C} = 1\,Volt = 1\,V$. This formula is similar to the formula for mechanical work given by

$$W = -\int_a^b \vec{F} \bullet \vec{ds} \tag{9}$$

In differential form the electrical potential is given by

$$\vec{E} = -\nabla\phi \tag{10}$$

For a single charge *Gauss's Law* is given by

$$\int\int_S \epsilon_e \vec{E} \bullet d\vec{A} = q \tag{11}$$

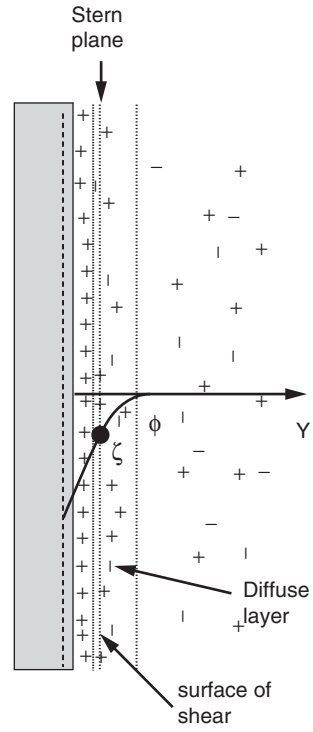
For a volume that contains a continuous distribution of charge, ρ_e, summing over all the charges using the definition of the integral, Gauss's Law becomes

$$\int\int_S \epsilon_e \vec{E} \bullet d\vec{A} = \int\int\int_{\mathcal{V}} \rho_e d\mathcal{V} \tag{12}$$

Using Eq. 12 and the differential form of the definition of the electrical potential, it follows that

$$\nabla^2\phi = -\frac{\rho_e}{\epsilon_e} \tag{13}$$

This is a *Poisson equation* for the potential given the volume charge density. The combination of Eqs. 4 and 13 is called the *Poisson-Nernst-Planck* system of equations.



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 2 The electric double layer (EDL) consists of a layer of counterions pinned to the wall, the Stern layer, and a diffuse layer of mobile ions outside that layer. The wall is shown as being negatively charged and the ζ potential is defined as the electrical potential at the Stern plane

Electrochemistry

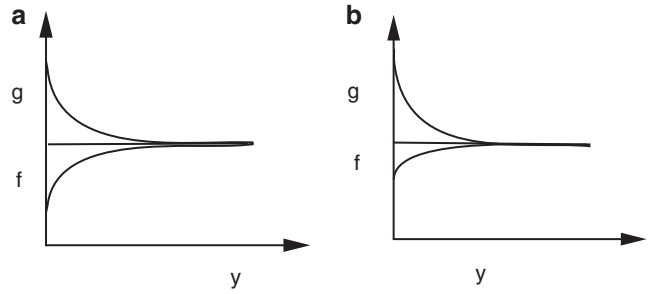
Electrochemistry may be broadly defined as the study of the electrical properties of chemical and biological material [2]. In particular much of electrochemistry pertinent to micro- and nanofluidics involves the study of the behavior of *ionic solutions* and the *electrical double layer* (*EDL*). Electrochemistry of electrodes is important to understand the operation of a *battery.*

An ionic or electrolyte solution is a mixture of ions, or charged species immersed in a solvent, often water. It is the charged nature of ionic solutions that allows the fluid to move under the action of an electric field, provided by electrodes placed upstream and downstream of a channel in a *nanopore membrane*. The term *membrane* is used to mean a thin sheet of porous material that allows fluid to flow in channels that make up the porous part of the membrane. Those channels are often like those channels depicted in Fig. 1.

Because the surface-to-volume ratio is so large in a nanoscale channel, the properties of the surface are extremely important. Fluid can be moved by an electric field if the surfaces of a channel are charged. If the surface is negatively charged, a surplus of positive ions will arrange themselves near the wall. This is shown in Fig. 2. It is because of this excess charge that allows fluid to be transported by an externally applied electric field.

The nominal length scale associated with the EDL is the *Debye length* defined by

$$\lambda = \frac{\sqrt{\epsilon_e RT}}{FI^{1/2}} \tag{14}$$



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 3 (**a**) Debye-Hückel [5] picture of the electric double layer. Here *g* denotes the cation mole fraction and *f* denotes the anion mole fraction. The Debye-Hückel model assumes the cation and anion wall mole fractions are symmetric about a mean value which occurs for low surface charge densities. (**b**) Gouy-Chapman model [3, 4] of the EDL allows many more counterions than coions to collect near the charged surface and is valid at higher surface charge densities. From [1]

where F is Faraday's constant, ϵ_e is the electrical permittivity of the medium, I is the *ionic strength*, $I = \sum_i z_i^2 c_i$, c_i is the concentrations of the electrolyte constituents at some reference location, R is the universal gas constant, z_i is the valence of species i, and T is the temperature.

The ion distribution within the EDL can be described by using the number density, concentration, or mole fraction. Engineers usually prefer the dimensionless mole fraction, whereas chemists usually use concentration or number density.

There are two views of the ion distribution within the electrical double layer that are generally thought to be valid and have been verified by numerical solutions of the governing equations (see the section on electrokinetic phenomena below). The Gouy-Chapman [3, 4] model of the electric double layer allows counterions to collect near the surface in much greater numbers than coions. This model as numerical solutions suggest [1] occurs at higher surface charge densities. The Debye-Hückel picture assumes that coions and counterions collect near the surface in roughly equal amounts, above and below a mean value. These pictures are depicted in Fig. 3.

Molecular Biology

The nanoscale is the scale of biology since many proteins and other biomolecules have nanoscale dimensions. Many of the applications of nanofluidics such as rapid molecular analysis, drug delivery, and biochemical sensing have a biological entity as an integral part of their operation. Moreover, using nanofluidic tools, DNA sequencing is now possible. The book by Alberts [6] is a useful tool for learning molecular biology.

Nucleic acids are polymers consisting of nucleotides. Those based on a sugar called *ribose* are called ribonucleic acids (RNA), and those based on *deoxyribose* are called deoxyribonucleic acids (DNA). RNA is single stranded, while DNA is usually double stranded although single-stranded DNA (ss-DNA) does exist. Nucleotides contain five-carbon sugars attached to one or more phosphate groups (a phosphorus central atom surrounded by four oxygens) and a base which can be either adenine (A), cytosine (C), guanine (G), or thymine (T). Two nucleotides connected by a hydrogen bond are called a base pair (bp). *Protein synthesis* begins at a gene on a particular strand of a DNA molecule in a cell [6].

There are seven basic types of proteins classified according to their function, although different authors use different terms to describe each class; see, for example, Alberts [6], panel 5-1. *Enzymes* are catalysts in biological reactions within the cell. For example, the immune system responds to foreign bacteria and viruses by producing *antibodies* that destroy or bind to the antigen, the foreign agent. The antigen is the catalyst, or

Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering,

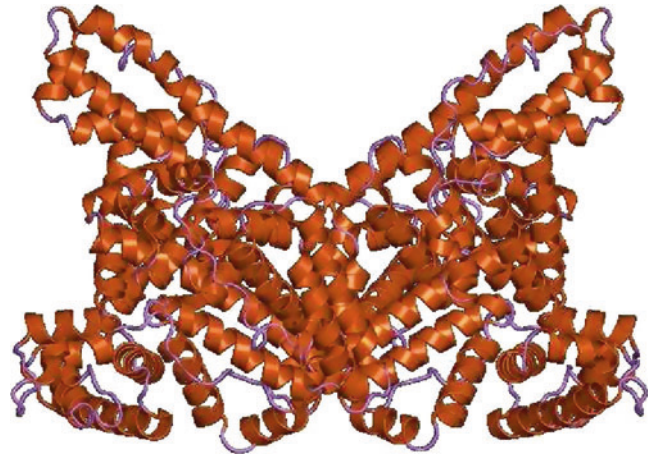


Fig. 4 Ribbon view of the protein albumin depicting its folding pattern, the *secondary structure* of a protein. From the European Bioinformatics Institute, public domain www.ebi.ac.uk

reaction enhancer, for inducing the immune response: the production of the antibodies. Proteins are responsible for many of the essential functions of the body, including moving material into and out of cells, regulating metabolism, managing temperature and *pH*, and muscle operation, among other functions.

Proteins are large and complex molecules, polymers made up of a total of 20 amino acids and held together by peptide bonds. The 20 amino acids have *side chains* that can be *basic*, *acidic*, *polar*, or *nonpolar*. Because they are so large, they cannot be described easily in a single chemical formula or picture. Thus molecular biologists depict proteins and other macromolecules in distinct levels of structure. The *primary structure* is the amino acid sequence, the order in which the 20 amino acids appear. The *secondary structure* depicts the folding properties of a protein as depicted in Fig. 4. Proteins are further described by more complex folding of the secondary structure (*tertiary structure*) and a *quaternary structure* if the protein has more than one backbone.

Proteins are usually negatively charged, and thus nanopore membranes for rapid molecular analysis can be used to separate different types of proteins and other biomolecules based on different values of size and charge. Biomolecules are what is termed soft material in that they are porous and deform under stress. Indeed, recent measurements of the conformation of albumin show that it may take the shape of a wedge, looking like a piece of pie. With the explosive growth of computer capability, conformations of biomolecules are actually being computed using molecular simulation tools like molecular dynamics and Monte Carlo schemes.

Ion channels are natural conical nanopores whose walls are made of proteins that play a crucial role in the transport of biofluids to and from cells. The basic units of all living organisms are cells. In order to keep the cells functioning properly, there needs to be a continuous flux of ions in and out of the cell and the cell components. The cell and many of its components are surrounded by a plasma membrane which provides selective transfer of ions. The membrane is made up of a double layer of lipid molecules (lipid bilayer) in which proteins are embedded. Ion channels are of two categories: carrier and pore. The *carrier* protein channel is based on the binding of the transport ion to a larger macromolecule, which brings it through the channel. A *pore* ion channel is a narrow, water-filled tunnel, permeable to the few ions and molecules small enough to fit through the tunnel (approximately 10 A in diameter).

Electrokinetic Phenomena

As the scale of the channels in a nanopore membrane becomes smaller, pressure, the normal

means for driving fluids through pipes and channels at macroscale (Fig. 1), becomes very difficult [1] since the pressure drop required scales as h^{-3} where h is the (nanoscale) channel height. Since in many applications the fluids used are electrically conducting, electric fields can be used to effectively pump fluid. Moreover electrically charged particles can move relative to the bulk fluid motion, and thus species of particles can be separated.

These *electrokinetic phenomena* are generally grouped into four classes [1]:

1. Electroosmosis (electroosmotic flow): the bulk motion of a fluid caused by an electric field
2. Electrophoresis: the motion of a charged particle in an otherwise motionless fluid or the motion of a charged particle relative to a bulk motion
3. Streaming potential or streaming current: the potential induced by a pressure gradient at zero current flow of an electrolyte mixture
4. Sedimentation potential: the electric field induced when charged particles move relative to a liquid under a gravitational or centrifugal or other force field

By far the two most important of these phenomena are electroosmosis and electrophoresis, and for the purposes of the theme of this entry, electroosmosis is discussed exclusively.

The dimensionless form of the streamwise momentum equation in the fully developed flow region in the absence of a pressure gradient is

$$\epsilon^2 \frac{\partial^2 u}{\partial y^2} = -\beta \sum_i z_i X_i \tag{15}$$

and the Poisson equation for the potential in dimensionless form is

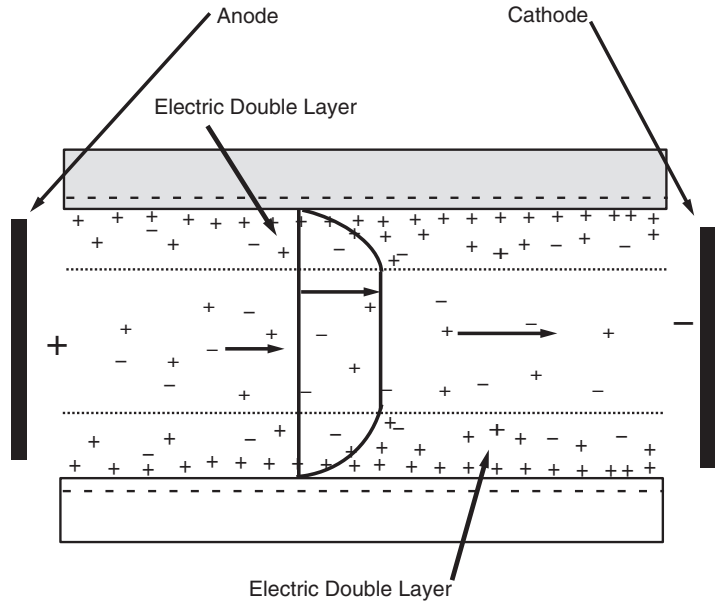
$$\epsilon^2 \frac{\partial^2 \phi}{\partial y^2} = -\beta \sum_i z_i X_i \tag{16}$$

where the partial derivatives in this one-dimensional fully developed analysis are really total derivatives, $\epsilon = \frac{\lambda}{h}$ and $\beta = \frac{c}{I}$ where c is the total concentration including the solvent and I is the ionic strength. Here X_i is the mole fraction, but if the electrolyte concentrations are scaled on the ionic strength, $I = \sum_i z_i c_i$, $\beta = 1$. It is seen from Eq. 15 that the combination of the electrodes that create an electric field and the excess charge in the electrical double layers produces the electrical force that balances the viscous force causing the electrolyte to move, and this is depicted in Fig. 5.

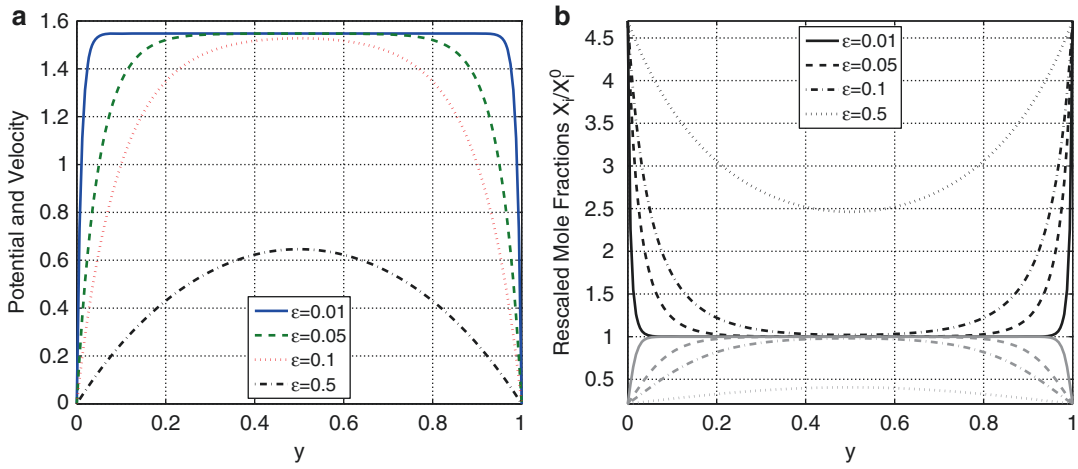
The fluid velocity satisfies the no-slip condition at the wall, and the electric potential satisfies $\phi(0) = \phi(1) = 0$. Then both the equations and the boundary conditions are identical and on a dimensionless basis $u(y) = \phi(y)$. In reality, the potential does not vanish at the wall, but if a Dirichlet boundary condition holds and the potential satisfies $\phi = \zeta$ at $y = 0{,}1$ and thus $u = \phi - \zeta$, where ζ is the *dimensionless* ζ-potential at the wall. Results for the potential and velocity and the concentrations scaled on the ionic strength are presented in Fig. 6 [1]. Note that for $\epsilon << 1$, the fluid velocity is constant away from the walls of the channel, unlike the Poiseuille flow of pressure-driven flow.

Dimensional Analysis

The equivalence of the dimensionless velocity and the dimensionless potential is important because the measurements of the velocity are equivalent to the measurements of the electric potential. Moreover, it is noted that a Debye length of $\lambda = 1$ nm in a $h = 100$ nm channel gives the same value of $\epsilon = 0.01$ as a $\lambda = 100$ nm Debye length in a $h = 10$ μm channel so that microscale measurements can be validated by nanoscale computations and, conversely, nanoscale computations can be validated by miscoscale experiments. Note that this similarity analysis does not apply for unsteady flow since there is no time derivative in the potential equation.



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 5 The combination of electrodes in the regions upstream and downstream of a charged channel or membrane, usually fluid reservoirs, causes electroosmotic flow



Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering, Fig. 6 (**a**) Potential, velocity, and rescaled mole fractions for a 1:1 electrolyte for various values of ϵ. Here the dimensional potential on both walls is $\zeta^* = -40mV$. In (**b**) the mole fractions are rescaled based on the upstream reservoir mole fractions as $\frac{X_i}{X_i^0}$. The cations are plotted in black lines and the anions are plotted in gray lines. From [1]

Closure

Physics of fluids at the nanoscale is dominated by the large surface-to-volume ratio inherent at this length scale, and thus the surfaces of a channel or membrane become extremely important. Indeed, the pressure drop across a rectangular channel in a nanopore membrane, $\Delta p \sim \frac{1}{h^3}$, is prohibitively large for the efficient operation of a nanofluidic device if $h = O(nm)$. Thus fluid, charged biomaterials such as proteins, and colloidal particles are most often

transported electrokinetically at the nanoscale. The art of designing micro- and nanofluidic devices therefore requires a significant amount of knowledge of fluid flow, mass transfer and sometimes heat transfer, electrostatics, electrokinetics, electrochemistry, and molecular biology.

The common thread is micro-/nanofluidics, which plays the role of unifying and integrating these fields. In particular, nanofluidics opens the door to reveal the structure and behavior of flows around nanoparticles and the conformation of proteins and other biomolecules using molecular simulations.

Cross-References

- ▶ Applications of Nanofluidics
- ▶ Electrokinetic Fluid Flow in Nanostructures
- ▶ Nanochannels for Nanofluidics: Fabrication Aspects
- ▶ Rapid Electrokinetic Patterning and Its Applications

References

1. Conlisk, A.T.: Essentials of Micro- and Nanofluidics with Application to the Biological and Chemical Sciences. Cambridge University Press, Cambridge, UK (2013)
2. Bockris, J.O.M., Reddy, A.K.N.: Modern Electrochemistry. Ionics, vol. 1, 2nd edn. Plenum Press, New York (1998)
3. Gouy, G.: About the electric charge on the surface of an electrolyte. J. Phys. A **9**, 457–468 (1910)
4. Chapman, D.L.: A contribution to the theory of electrocapillarity. Phil. Mag. **25**, 475–481 (1913)
5. Debye, P., Huckel, E.: The interionic attraction theory of deviations from ideal behavior in solution. Z. Phys. **24**, 185 (1923)
6. Alberts, B., Bray, D., Hopkin, K., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Essential Cell Biology. Garland Publishing, New York (1998)

Computational Studies

- ▶ Surface Modeling of Ceramic Biomaterials

Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Leonid V. Zhigilei[1], Alexey N. Volkov[2] and Avinash M. Dongare[3]
[1]Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA, USA
[2]Department of Mechanical Engineering, University of Alabama, Tuscaloosa, AL, USA
[3]Department of Materials Science and Engineering, and Institute of Materials Science, University of Connecticut, Storrs, CT, USA

Synonyms

Carbon nanotube materials; Computer modeling and simulation of materials; Dislocation dynamics; Mesoscopic modeling; Molecular dynamics method; Multiscale modeling; Nanocrystalline materials; Nanofibrous materials and composites

Definitions

Nanomaterials (or nanostructured materials, nanocomposites) are materials with characteristic size of structural elements on the order or less than several hundreds of nanometers at least in one dimension. Examples of nanomaterials include nanocrystalline materials, nanofiber, nanotube, and nanoparticle-reinforced nanocomposites, as well as multilayered systems with submicron thickness of the layers.

Atomistic modeling is based on atoms as elementary units in the models, thus providing the atomic-level resolution in the computational studies of material structure and properties. The main classical atomistic methods in materials research are (1) molecular dynamics technique that yields "atomic movies" of the dynamic material behavior through the integration of the equations of motion of atoms and molecules, (2) Metropolis Monte Carlo method that enables evaluation of

the equilibrium properties through the ensemble averaging over a sequence of random atomic configurations generated according to the desired statistical-mechanics distribution, and (3) kinetic Monte Carlo method that provides a computationally efficient way to study systems where the structural evolution is defined by a finite number of thermally activated elementary processes.

Mesoscopic modeling is a relatively new area of computational materials science that considers material behavior at time and length scales intermediate between the atomistic and continuum levels. Mesoscopic models are system/phenomenon specific and adopt coarse-grained representations of the material structure, with elementary units in the models designed to provide a computationally efficient representation of individual crystal defects or other elements of micro/nanostructure. Examples of the mesoscopic models are coarse-grained models for molecular systems, discrete dislocation dynamics model for crystal plasticity, mesoscopic models for nanofibrous materials, cellular automata, and kinetic Monte Carlo Potts models for simulation of microstructural evolution in polycrystalline materials.

Computer Modeling of Nanomaterials

Rapid advances in synthesis of nanostructured materials combined with reports of their enhanced or unique properties have created, over the last decades, a new active area of materials research. Due to the nanoscopic size of the structural elements in nanomaterials, the interfacial regions, which represent an insignificant volume fraction in traditional materials with coarse microstructures, start to play the dominant role in defining the physical and mechanical properties of nanostructured materials. This implies that the behavior of nanomaterials cannot be understood and predicted by simply applying scaling arguments from the structure–property relationships developed for conventional polycrystalline, multiphase, and composite materials. New models and constitutive relations, therefore, are needed for an adequate description of the behavior and properties of nanomaterials.

Computer modeling is playing a prominent role in the development of the theoretical understanding of the connections between the atomic-level structure and the effective (macroscopic) properties of nanomaterials. Atomistic modeling has been at the forefront of computational investigation of nanomaterials and has revealed a wealth of information on structure and properties of individual structural elements (various nanolayers, nanoparticles, nanofibers, nanowires, and nanotubes) as well as the characteristics of the interfacial regions and modification of the material properties at the nanoscale. Due to the limitations on the time and length scales, inherent to atomistic models, it is often difficult to perform simulations for systems that include a number of structural elements that is sufficiently large to provide a reliable description of the macroscopic properties of the nanostructured materials. An emerging key component of the computer modeling of nanomaterials is, therefore, the development of novel mesoscopic simulation techniques capable of describing the collective behavior of large groups of the elements of the nanostructures and providing the missing link between the atomistic and continuum (macroscopic) descriptions. The capabilities and limitations of the atomistic and mesoscopic computational models used in investigations of the behavior and properties of nanomaterials are briefly discussed and illustrated by examples of recent applications below.

Atomistic Modeling

In atomistic models [1, 2], the individual atoms are considered as elementary units, thus providing the atomic-level resolution in the description of the material behavior and properties. In classical atomistic models, the electrons are not present explicitly but are introduced through the interatomic potential, $U\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right)$, that describes the dependence of the potential energy of a system of N atoms on the positions $\vec{r}_i$ of the atoms. It is assumed that the electrons adjust to changes in atomic positions much faster than the atomic nuclei move (Born-Oppenheimer approximation) and the potential energy of a system of

interacting atoms is uniquely defined by the atomic positions.

The interatomic potentials are commonly described by analytic functions designed and parameterized by fitting to available experimental data (e.g., equilibrium geometry of stable phases, density, cohesive energy, elastic moduli, vibrational frequencies, characteristics of the phase transitions, etc.). The interatomic potentials can also be evaluated through direct quantum mechanics–based electronic structure calculations in so-called first-principle (*ab initio*) simulation techniques. The *ab initio* simulations, however, are computationally expensive and are largely limited to relatively small systems consisting of tens to thousands of atoms. The availability of reliable and easy-to-compute interatomic potential functions is one of the main conditions for the expansion of the area of applicability of atomistic techniques to realistic quantitative analysis of the behavior and properties of nanostructured materials.

The three atomistic computational techniques commonly used in materials research are

1. Metropolis Monte Carlo method – the equilibrium properties of a system are obtained via ensemble averaging over a sequence of random atomic configurations, sampled with probability distribution characteristic for a given statistical mechanics ensemble. This is accomplished by setting up a random walk through the configurational space with specially designed choice of probabilities of going from one state to another. In the area of nanomaterials, the application of the method is largely limited to investigations of the equilibrium shapes of individual elements of nanostructure (e.g., nanoparticles) and surface structure/composition (e.g., surface reconstruction and compositional segregation [3]).
2. Kinetic Monte Carlo method – the evolution of a nanostructure can be obtained by performing atomic rearrangements governed by predefined transition rates between the states, with time increments formulated so that they relate to the microscopic kinetics of the system. Kinetic Monte Carlo is effective when the structural and/or compositional changes in a nanostructure are defined by a relatively small number of thermally activated elementary processes, e.g., when surface diffusion is responsible for the evolution of shapes of small crystallites [4] or growth of two-dimensional fractal-dendritic islands [5].
3. Molecular dynamics method – provides the complete information on the time evolution of a system of interacting atoms through the numerical integration of the equations of motion for all atoms in the system. This method is widely used in computational investigations of nanomaterials and is discussed in more detail below.

Molecular Dynamics Technique

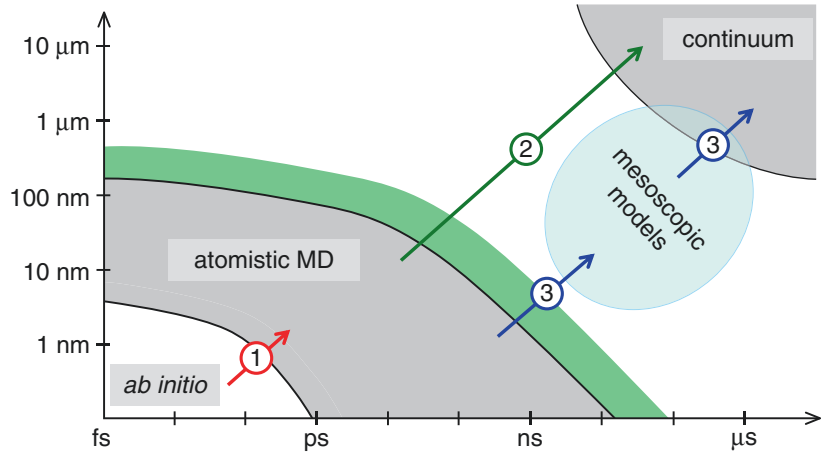
Molecular dynamics (MD) is a computer simulation technique that allows one to follow the evolution of a system of N particles (atoms in the case of atomistic modeling) in time by solving classical equations of motion for all particles in the system,

$$m_i \frac{d^2 \vec{r}_i}{dt^2} = \vec{F}_i, \quad i = 1, 2, \ldots, N \qquad (1)$$

where m_i and $\vec{r}_i$ are the mass and position of a particle i and $\vec{F}_i$ is the force acting on this particle due to the interaction with other particles in the system. The force acting on the ith particle at a given time is defined by the gradient of the interparticle interaction potential $U\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right)$ that, in general, is a function of the positions of all the particles:

$$\vec{F}_i = -\vec{\nabla}_i U\left(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N\right) \qquad (2)$$

Once the initial conditions (initial positions and velocities of all particles in the system) and the interaction potential are defined, the equations of motion, Eq. 1, can be solved numerically. The result of the solution is the trajectories (positions and velocities) of all the particles as a function of time, $\vec{r}_i(t)$, $\vec{v}_i(t)$, which is the only direct output of a MD simulation. From the trajectories of all



Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling, Fig. 1 Schematic representation of the time- and length-scale domains of fist-principle (*ab initio*) electronic structure calculations, classical atomistic MD, and continuum modeling of materials. The domain of continuum modeling can be different for different materials and corresponds to the time and length scales at which the effect of the micro/nanostructure can be averaged over to yield the effective material properties. The *arrows* show the connections between the computational methods used in multiscale modeling of materials: The *red arrow* #1 corresponds to the use of quantum mechanics–based electronic structure calculations to design interatomic potentials for classical MD simulations or to verify/correct the predictions of the classical atomistic simulations; the *green arrow* #2 corresponds to the direct use of the predictions of large-scale atomistic simulations of nanostructured materials for the design of continuum-level constitutive relations describing the material behavior and properties; and the two *blue arrows* #3 show a two-step path from atomistic to continuum material description through an intermediate mesoscopic modeling

particles in the system, however, one can easily calculate (sometimes by sampling the ensemble of initial conditions) the spatial and time evolution of structural and thermodynamic parameters of the system. For example, a detailed atomic-level analysis of the development of the defect structures or phase transformations can be performed and related to changes in temperature and pressure in the system (see examples below).

The main strength of the MD method is that only details of the interatomic interactions need to be specified and no assumptions are made about the character of the processes under study. This is an important advantage that makes MD to be capable of discovering new physical phenomena or processes in the course of "computer experiments." Moreover, unlike in real experiments, the analysis of fast nonequilibrium processes in MD simulations can be performed with unlimited atomic-level resolution, providing complete information on the phenomena of interest.

The predictive power of the MD method, however, comes at a price of a high computational cost of the simulations, leading to severe limitations on time and length scales accessible for MD simulations, as shown schematically in Fig. 1. Although the record length-scale MD simulations have been demonstrated for systems containing more than 10^{12} atoms (corresponds to cubic samples on the order of 10 µm in size) with the use of hundreds of thousands of processors on one of the world's largest supercomputers [6], most of the systems studied in large-scale MD simulations do not exceed hundreds of nanometers even in simulations performed with computationally efficient parallel algorithms (shown by a green area extending the scales accessible for MD simulations in Fig. 1). Similarly, although the record long time scales of up to hundreds of microseconds have been reported for simulations of protein folding performed through distributed computing [7], the duration of most of the simulations in the area of materials research does not exceed tens of nanoseconds.

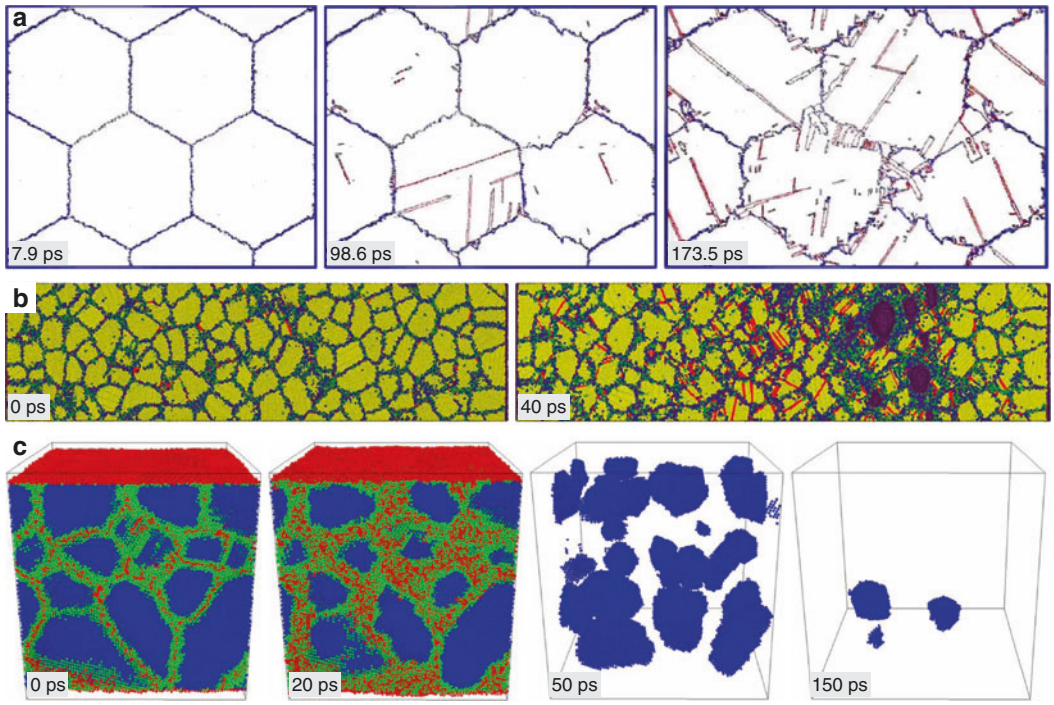
Molecular Dynamics Simulations of Nanomaterials

Both the advantages and limitations of the MD method, briefly discussed above, have important implications for simulations of nanomaterials. The transition to the nanoscale size of the structural features can drastically change the material response to the external thermal, mechanical, or electromagnetic stimuli, making it necessary to develop new structure-property relationships based on new mechanisms operating at the nanoscale. The MD method is in a unique position to provide a complete microscopic description of the atomic dynamics under various conditions without making any *a priori* assumptions on the mechanisms and processes defining the material behavior and properties.

On the other hand, the limitations on the time and length scales accessible to MD simulations make it difficult to directly predict the macroscopic material properties that are essentially the result of a homogenization of the processes occurring at the scale of the elements of the nanostructure. Most of the MD simulations have been aimed at investigation of the behavior of individual structural elements (nanofibers, nanoparticles, interfacial regions in multiphase systems, grain boundaries, etc.). The results of these simulations, while important for the mechanistic understanding of the elementary processes at the nanoscale, are often insufficient for making a direct connection to the macroscopic behavior and properties of nanomaterials.

With the fast growth of the available computing resources, however, there have been an increasing number of reports on MD simulations of systems that include multiple elements of nanostructures. A notable class of nanomaterials actively investigated in MD simulations is nanocrystalline materials – a new generation of advanced polycrystalline materials with sub-micron size of the grains. With a number of atoms on the order of several hundred thousands and more, it is possible to simulate a system consisting of tens of nanograins and to investigate the effective properties of the material (i.e., to make a direct link between the atomistic and continuum descriptions, as shown schematically by the green arrow #2 in Fig. 1). MD simulations of nanocrystalline materials addressing the mechanical [8, 9] and thermal transport [10] properties as well as the kinetics and mechanisms of phase transformations [11, 12] have been reported, with several examples illustrated in Fig. 2. In the first example, Fig. 2a, atomic-level analysis of the dislocation activity and grain-boundary processes occurring during mechanical deformation of an aluminum nanocrystalline system consisting of columnar grains is performed and the important role of mechanical twinning in the deformation behavior of the nanocrystalline material revealed [8]. In the second example, Fig. 2b, the processes of void nucleation, growth, and coalescence in the ductile failure of nanocrystalline copper subjected to an impact loading are investigated, providing important pieces of information necessary for the development of a predictive analytical model of the dynamic failure of nanocrystalline materials [9]. The third example, Fig. 2c, illustrates the effect of nanocrystalline structure on the mechanisms and kinetics of short-pulse laser melting of thin gold films. It is shown that the initiation of melting at grain boundaries can steer the melting process along the path where the melting continues below the equilibrium melting temperature and the crystalline regions shrink and disappear under conditions of substantial undercooling [12].

The brute-force approach to the atomistic modeling of nanocrystalline materials (increase in the number of atoms in the system) has its limits in addressing the complex collective processes that involve many grains and may occur at a micrometer length scale and above. Further progress in this area may come through the development of concurrent multiscale approaches based on the use of different resolutions in the description of the intragranular and grain boundary regions in a well-integrated computational model. An example of a multiscale approach is provided in Ref. [13], where scale-dependent constitutive equations are designed for a generalized finite element method (FEM) so that the atomistic MD equations of motion are reproduced in the regions where the FEM mesh is refined down to atomic level. This and other multiscale approaches can help to focus computational efforts on the important regions of the system



Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling, Fig. 2 Snapshots from atomistic MD simulations of nanocrystalline materials: (**a**) mechanical deformation of nanocrystalline Al (only atoms in the twin boundaries left behind by partial dislocations and atoms in disordered regions are shown by *red* and *blue* colors, respectively) [8]; (**b**) spallation of nanocrystalline Cu due to the reflection of a shock wave from a surface of the sample (atoms that have local fcc, hcp, and disordered structure are shown by *yellow*, *red*, and *green/blue* colors, respectively) [9]; and (**c**) laser melting of a nanocrystalline Au film irradiated with a 200 fs laser pulse at a fluence close to the melting threshold (atoms that have local fcc surroundings are colored *blue*, atoms in the liquid regions are *red* and *green*, in the snapshots for 50 and 150 ps the liquid regions are blanked to expose the remaining crystalline regions) [12]

where the critical atomic-scale processes take place. The practical applications of the multiscale methodology so far, however, have been largely limited to investigations of individual elements of material microstructure (crack tips, interfaces, and dislocation reactions), with the regions represented with coarse-grained resolution serving the purpose of adaptive boundary conditions. The perspective of the concurrent multiscale modeling of nanocrystalline materials remains unclear due to the close coupling between the intragranular and grain boundary processes. To enable the multiscale modeling of dynamic processes in nanocrystalline materials, the design of advanced computational descriptions of the coarse-grained parts of the model is needed so that the plastic deformation and thermal dissipation could be adequately described without switching to fully atomistic modeling.

Mesoscopic Modeling

A principal challenge in computer modeling of nanomaterials is presented by the gap between the atomistic description of individual structural elements and the macroscopic properties defined by the collective behavior of large groups of the structural elements. Apart from a small number of exceptions (e.g., simulations of nanocrystalline materials briefly discussed above), the direct analysis of the effective properties of nanostructured

materials is still out of reach for atomistic simulations. Moreover, it is often difficult to translate the large amounts of data typically generated in atomistic simulations into key physical parameters that define the macroscopic material behavior. This difficulty can be approached through the development of mesoscopic computational models capable of representing the material behavior at time and length scales intermediate between the atomistic and continuum levels (prefix *meso* comes from the Greek word μέσος, which means middle or intermediate).

The mesoscopic models provide a "stepping stone" for bridging the gap between the atomistic and continuum descriptions of the material structure, as schematically shown by the blue arrows #3 in Fig. 1. Mesoscopic models are typically designed and parameterized based on the results of atomistic simulations or experimental measurements that provide information on the internal properties and interactions between the characteristic structural elements in the material of interest. The mesoscopic simulations can be performed for systems that include multiple elements of micro/nanostructure, thus enabling a reliable homogenization of the structural features to yield the effective macroscopic material properties. The general strategy in the development of a coarse-grained mesoscopic description of the material dynamics and properties includes the following steps.

1. Identifying the collective degrees of freedom *relevant for the phenomenon under study* (the focus on different properties of the same material may affect the choice of the structural elements of the model);
2. Designing, based on the results of atomic-level simulations and/or experimental data, a set of rules (or a mesoscopic force field) that governs the dynamics of the collective degrees of freedom;
3. Adding a set of rules describing the changes in the properties of the dynamic elements in response to the local mechanical stresses and thermodynamic conditions.

While the atomistic and continuum simulation techniques are well established and extensively used, the mesoscopic modeling is still in the early development stage. The central problem of the mesoscopic modeling is the design of the mesoscopic force field or, in general, the "equations of motion" that describe the dynamics of the collective degrees of freedom in the mesoscopic system. In principle, the equations for the collective degrees of freedom can be derived from the equations of motion of atoms and formulated as the so-called generalized Langevin equation by applying the Mori-Zwanzig projection operators [14, 15]. While formally exact, this equation is difficult to apply to real systems, since a number of parameters of this equation should be evaluated by considering the so-called projection dynamics that is different from the real microscopic dynamics of the corresponding atomistic system [16]. Moreover, the generalized Langevin equation is a stochastic integrodifferential equation with memory effects, which is difficult to apply to large-scale mesoscopic simulations. Therefore, apart from polymer systems, where the Mori-Zwanzig formalism is shown to be a practical tool for constructing mesoscopic computational models [16–18], the mesoscopic force fields for various nanomaterials are usually designed based on semiempirical approaches. There is no universal mesoscopic technique or methodology in the development of such semiempirical force fields, and the current state of the art in mesoscopic simulations is characterized by the development of system-/phenomenon-specific mesoscopic models.

The mesoscopic models used in materials modeling can be roughly divided into two general categories: (1) the models based on lumping together groups of atoms into larger dynamic units or particles and (2) the models that represent the material microstructure and its evolution due to thermodynamic driving forces or mechanical loading at the level of individual crystal defects. The basic ideas underlying these two general classes of mesoscopic models are briefly discussed below.

The models where groups of atoms are combined into coarse-grained computational particles are practical for materials with well-defined structural hierarchy (that allows for a natural choice of

the coarse-grained particles) and a relatively weak coupling between the *internal* atomic motions inside the coarse-grained particles and the *collective* motions of the particles. In contrast to atomic-level models, the atomic structure of the structural elements represented by the coarse-grained particles is not explicitly represented in this type of mesoscopic models. On the other hand, in contrast to continuum models, the coarse-grained particles allow one to explicitly reproduce the nanostructure of the material. Notable examples of mesoscopic models of this type are coarse-grained models for molecular systems including polymers [19–21] and mesoscopic models for carbon nanotubes and nanofibrous materials [22–26]. The individual molecules (or mers in polymer molecules) and nanotube/nanofiber segments are chosen as the dynamic units in these models. The collective dynamic degrees of freedom that correspond to the motion of the "mesoparticles" are explicitly accounted for in mesoscopic models, while the internal degrees of freedom are either neglected or described by a small number of internal state variables. The description of the internal states of the mesoparticles and the energy exchange between the dynamic degrees of freedom and the internal state variables becomes important for simulations of nonequilibrium phenomena that involve fast energy deposition from an external source, heat transfer, or dissipation of mechanical energy.

Another group of mesoscopic models is aimed at a computationally efficient description of the evolution of the defect structures in crystalline materials. The mesoscopic models from this group include the discrete dislocation dynamics model for simulation of crystal plasticity [27–29] and a broad class of methods designed for simulation of grain growth, recrystallization, and associated microstructural evolution (e.g., phase field models, cellular automata, and kinetic Monte Carlo Potts models) [27, 28, 30]. Despite the apparent diversity of the physical principles and computational algorithms adopted in different models listed above, the common characteristic of these models is the focus on a realistic description of the behavior and properties of individual crystal defects (grain boundaries and dislocations), their interactions with each other, and the collective evolution of the totality of crystal defects responsible for the changes in the microstructure.

Two examples of mesoscopic models (one for each of the two types of the models discussed above) and their relevance to the investigation of nanomaterials are considered in more detail next.

Discrete Dislocation Dynamics

The purpose of the discrete dislocation dynamics (DD) is to describe the plastic deformation in crystalline materials, which is largely defined by the motions, interactions, and multiplication of dislocations. Dislocations are linear crystal defects that generate long-range elastic strain fields in the surrounding elastic solid. The elastic strain field is accounting for ~90 % of the dislocation energy and is responsible for the interactions of dislocations among themselves and with other crystal defects. The collective behavior of dislocations in the course of plastic deformation is defined by these long-range interactions as well as by a large number of local reactions (annihilation, formation of glissile junctions or sessile dislocation segments such as Lomer or Hirth locks) occurring when the anelastic core regions of the dislocation lines come into contact with each other. The basic idea of the DD model is to solve the dynamics of the dislocation lines in elastic continuum and to include information about the local reactions. The elementary unit in the discrete dislocation dynamics method is, therefore, a segment of a dislocation.

The continuous dislocation lines are discretized into segments, and the total force acting on each segment in the dislocation slip plane is calculated. The total force includes the contributions from the external force, the internal force due to the interaction with other dislocations and crystal defects that generate elastic fields, the "self force" that can be represented by a "line tension" force for small curvature of the dislocation, the Peierls force that acts like a friction resisting the

dislocation motion, and the "image" force related to the stress relaxation in the vicinity of external or internal surfaces. Once the total forces and the associated resolved shear stresses, τ^*, acting on the dislocation segments are calculated, the segments can be displaced in a finite difference time integration algorithm applied to the equations connecting the dislocation velocity, v, and the resolved shear stress, e.g., [28]

$$v = A\left(\frac{\tau^*}{\tau_0}\right)^m \exp\left(-\frac{\Delta U}{kT}\right), \tag{3}$$

when the displacement of a dislocation segment is controlled by thermally activated events (ΔU is the activation energy for dislocation motion, m is the stress exponent, and τ_0 is the stress normalization constant) or

$$v = \tau^* b/B, \tag{4}$$

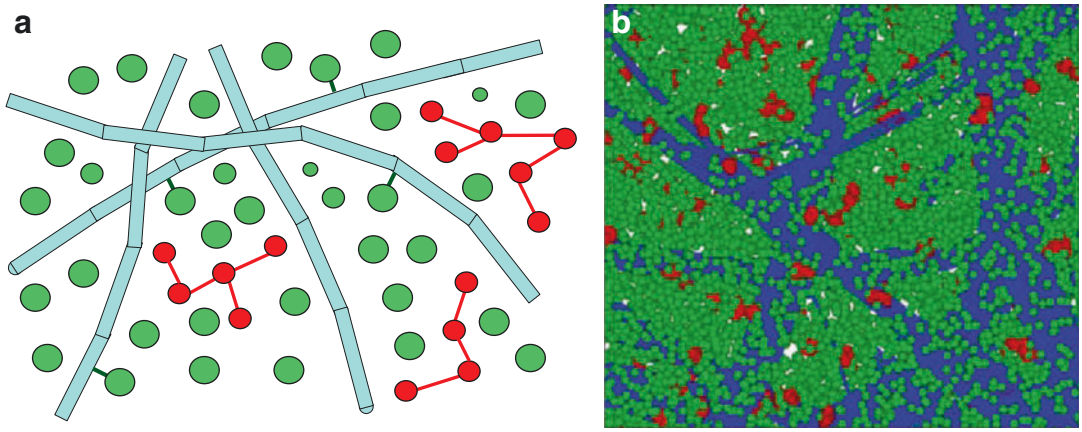
that corresponds to the Newtonian motion equation accounting for the atomic and electron drag force during the dislocation "free flight" between the obstacles (B is the effective drag coefficient and b is the magnitude of the Burgers vector of the dislocation).

Most of the applications of the DD model have been aimed at the investigation of the plastic deformation and hardening of single crystals (increase in dislocation density as a result of multiplication of dislocations present in the initial system). The extension of the DD modeling to nanomaterials is a challenging task as it requires an enhancement of the technique with a realistic description of the interactions between the dislocations and grain boundaries and/or interfaces as well as an incorporation of other mechanisms of plasticity (e.g., grain boundary sliding and twinning in nanocrystalline materials). There have only been several initial studies reporting the results of DD simulations of nanoscale metallic multilayered composites, e.g., [31]. Due to the complexity of the plastic deformation mechanisms and the importance of anelastic short-range interactions among the crystal defects in nanomaterials, the development of novel hybrid computational methods combining the DD technique with other mesoscopic methods is likely to be required for realistic modeling of plastic deformation in this class of materials.

Mesoscopic Model for Nanofibrous Materials

The design of new nanofibrous materials and composites is an area of materials research that is currently experiencing a rapid growth. The interest in this class of materials is fueled by a broad range of potential applications, ranging from fabrication of flexible/stretchable electronic and acoustic devices to the design of advanced nanocomposite materials with improved mechanical properties and thermal stability. The behavior and properties of nanofibrous materials are defined by the collective dynamics of the nanofibers and, in the case of nanocomposites, their interactions with the matrix. Depending on the structure of the material and the phenomenon of interest, the number of nanofibers that has to be included in the simulation in order to ensure a reliable prediction of the effective macroscopic properties can range from several hundreds to millions. The direct atomic-level simulation of systems consisting of large groups of nanofibers (the path shown by the green arrow #2 in Fig. 1) is beyond the capabilities of modern computing facilities. Thus, an alternative two-step path from atomistic investigation of individual structural elements and interfacial properties to the continuum material description through an intermediate mesoscopic modeling (blue arrows #3 in Fig. 1) appears to be the most viable approach to modeling of nanofibrous materials. An example of a mesoscopic computational model recently designed and parameterized for carbon nanotube (CNT)-based materials is briefly discussed below.

The mesoscopic model for fibrous materials and organic matrix nanocomposites adopts a coarse-grained description of the nanocomposite constituents (nanofibers and matrix molecules), as schematically illustrated in Fig. 3. The individual CNTs are represented as chains of stretchable cylindrical segments [22], and the organic matrix is modeled by a combination of the



Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling, Fig. 3 Schematic representation of the basic components of the dynamic mesoscopic model of a CNT-based nanocomposite material (**a**) and a corresponding molecular-level view of a part of the system where a network of CNT bundles (*blue* color) is embedded into an organic matrix (*green* and *red* color) (**b**)

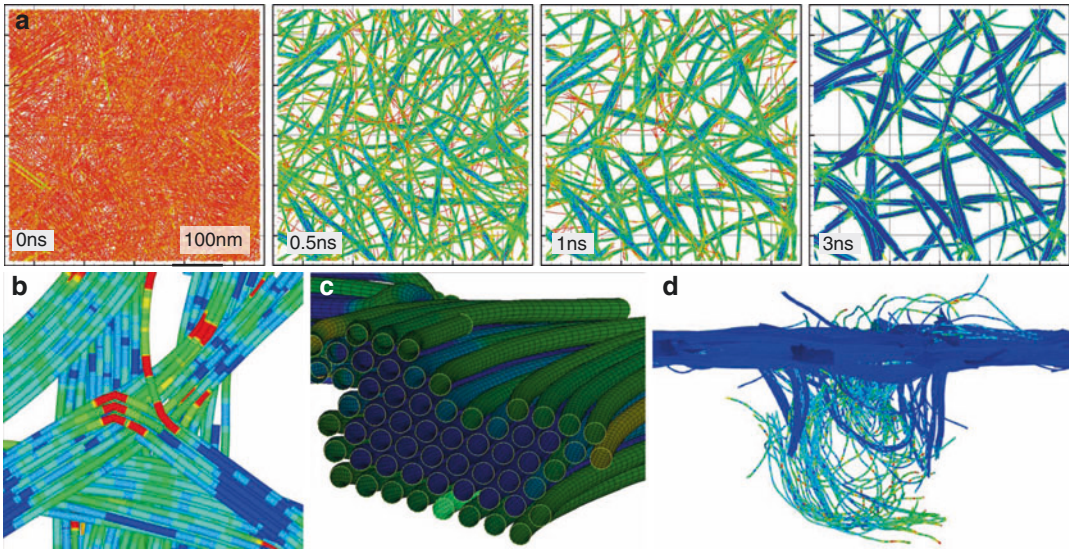
conventional "bead-and-spring" model commonly used in polymer modeling [19, 20] and the "breathing sphere" model developed for simulation of simple molecular solids [21] and polymer solutions [32].

The degrees of freedom, for which equations of motion are solved in dynamic simulations or Metropolis Monte Carlo moves are performed in simulations aimed at finding the equilibrium structures, are the positions of nodes defining the segments, the positions of the molecular units, and the radii of the spherical particles in the breathing sphere molecules. The potential energy of the system can be written as

$$U = U_{T(\text{int})} + U_{T-T} + U_{M-M} + U_{M(\text{int})} + U_{M-T} \quad (5)$$

where $U_{T(\text{int})}$ is the potential that describes the internal strain energy associated with stretching and bending of individual CNTs, U_{T-T} is the energy of intertube interactions, U_{M-M} is the energy of chemical and nonbonding interactions in the molecular matrix, $U_{M(\text{int})}$ is the internal breathing potential for the matrix units, and U_{M-T} is the energy of matrix–CNT interaction that can include both nonbonding van der Waals interactions and chemical bonding. The internal CNT potential $U_{T(\text{int})}$ is parameterized based on the results of atomistic simulations [22] and accounts for the transition to the anharmonic regime of stretching (nonlinear stress–strain dependence), fracture of nanotubes under tension, and bending buckling [33]. The intertube interaction term U_{T-T} is calculated based on the tubular potential method that allows for a computationally efficient and accurate representation of van der Waals interactions between CNT segments of arbitrary lengths and orientation [23]. The general procedure used in the formulation of the tubular potential is not limited to CNTs or graphitic structures. The tubular potential (and the mesoscopic model in general) can be parameterized for a diverse range of systems consisting of various types of nano- and microtubular elements, such as nanotubes, nanorodes, and microfibers.

Simulations performed with the mesoscopic model demonstrate that the model is capable of simulating the dynamic behavior of systems consisting of thousands of CNTs on a timescale extending up to tens of nanoseconds. In particular, simulations performed for systems composed of randomly distributed and oriented CNTs predict spontaneous self-assembly of CNTs into continuous networks of bundles with partial hexagonal ordering of CNTs in the bundles, Fig. 4a–c [23, 33]. The bending buckling of CNTs (e.g., see



Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling, Fig. 4 Snapshots from mesoscopic simulations of systems consisting of (10,10) single-walled carbon nanotubes: (**a**) spontaneous self-organization of CNTs into a continuous network of CNT bundles (CNT segments are colored according to the local intertube interaction energy) [23]; (**b**) an enlarged view of a structural element of the CNT network (CNT segments colored according to the local radii of curvature and the *red* color marks the segments adjacent to buckling kinks) [33]; (**c**) a cross section of a typical bundle showing a hexagonal arrangement of CNTs in the bundle [23]; (d) snapshot from a simulation of a high-velocity impact of a spherical projectile on a free-standing thin CNT film [34]

Fig. 4b) is found to be an important factor responsible for the stability of the network structures formed by defect-free CNTs [33]. The structures produced in the simulations are similar to the structures of CNT films and buckypaper observed in experiments. Note that an atomic-level simulation of a system similar to the one shown in the left panel of Fig. 4 would require $\sim 2.5 \times 10^9$ atoms, making such simulation unfeasible.

Beyond the structural analysis of CNT materials, the development of the mesoscopic model opens up opportunities for investigation of a broad range of important phenomena. In particular, the dynamic nature of the model makes it possible to perform simulations of the processes occurring under conditions of fast mechanical loading (blast/impact resistance, response to the shock loading, etc.), as illustrated by a snapshot from a simulation of a high-velocity impact of a spherical projectile on a free-standing thin CNT film shown in Fig. 4d [34]. With a proper parameterization, the mesoscopic model can also be adopted for calculation of electrical and thermal transport properties of complex nanofibrous materials [35, 36].

Future Research Directions

The examples of application of the atomistic and mesoscopic computational techniques, briefly discussed above, demonstrate the ability of computer modeling to provide insights into the complex processes that define the behavior and properties of nanostructured materials. The fast advancement of experimental methods capable of probing nanostructured materials with high spatial and temporal resolution is an important factor that allows for verification of computational predictions and stimulates the improvement of the computational models. With further innovative development of computational methodology and the steady growth of the available computing resources, one can expect that both atomistic and

mesoscopic modeling will continue to play an increasingly important role in nanomaterials research.

In the area of atomistic simulations, the development of new improved interatomic potentials (often with the help of *ab initio* electronic structure calculations, red arrow #1 in Fig. 1) makes material-specific computational predictions more accurate and enables simulations of complex multicomponent and multiphase systems. Further progress can be expected in two directions that are already actively pursued: (1) large-scale MD simulations of the fast dynamic phenomena in nanocrystalline materials (high strain rate mechanical deformation, shock loading, impact resistance, response to fast heating, etc.) and (2) detailed investigation of the atomic structure and properties of individual structural elements in various nanomaterials (grain boundaries and interfaces, nanotubes, nanowires, and nanoparticles of various shapes). The information obtained in large-scale atomistic simulations of nanocrystalline materials can be used to formulate theoretical models translating the atomic-level picture of material behavior to the constitutive relations describing the dependence of the mechanical and thermal properties of these materials on the grain size distribution and characteristics of nanotexture (green arrow #2 in Fig. 1).

The results of the detailed analysis of the structural elements of the nanocomposite materials can be used in the design and parameterization of mesoscopic models, where the elementary units treated in the models correspond to building blocks of the nanostructure (elements of grain boundaries, segments of dislocations, etc.) or groups of atoms that have some distinct properties (belong to a molecule, a mer unit of a polymer chain, a nanotube, a nanoparticle in nanocomposite material, etc.). The design of novel system-specific mesoscopic models capable of bridging the gap between the atomistic modeling of structural elements of nanostructured materials and the continuum models (blue arrows #3 in Fig. 1) is likely to become an important trend in the computational investigation of nanomaterials. To achieve a realistic description of complex processes occurring in nanomaterials, the description of the elementary units of the mesoscopic models should become more flexible and sophisticated. In particular, an adequate description of the energy dissipation in nanomaterials can only be achieved if the energy exchange between the atomic degrees of freedom, excluded in the mesoscopic models, and the coarse-grained dynamic degrees of freedom is accounted for [34]. A realistic representation of the dependence of the properties of the mesoscopic units of the models on local thermodynamic conditions can also be critical in modeling of a broad range of phenomena.

In general, the optimum strategy in investigation of nanomaterials is to use a well-integrated multiscale computational approach combining the *ab initio* and atomistic analysis of the constituents of nanostructure with mesoscopic modeling of the collective dynamics and kinetics of the structural evolution of the material and leading to the improved theoretical understanding of the factors controlling the effective material properties. It is the improved understanding of the connections between the processes occurring at different time and length scales that is likely to be the key factor defining the pace of progress in the area of computational design of new nanocomposite materials.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Active Carbon Nanotube-Polymer Composites
- ▶ Carbon-Nanotubes
- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Mechanical Properties of Nanocrystalline Metals
- ▶ Modeling Thermal Properties of Carbon Nanostructure Composites
- ▶ Molecular Modeling
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Plasticity Theory at Small Scales
- ▶ Reactive Empirical Bond-Order Potentials
- ▶ Self-Assembly of Nanostructures

References

1. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Clarendon, Oxford (1987)
2. Frenkel, D., Smit, B.: Understanding Molecular Simulation: From Algorithms to Applications. Academic, San Diego (1996)
3. Kelires, P.C., Tersoff, J.: Equilibrium alloy properties by direct simulation: oscillatory segregation at the Si-Ge(100) 2×1 surface. Phys. Rev. Lett. **63**, 1164–1167 (1989)
4. Combe, N., Jensen, P., Pimpinelli, A.: Changing shapes in the nanoworld. Phys. Rev. Lett. **85**, 110–113 (2000)
5. Liu, H., Lin, Z., Zhigilei, L.V., Reinke, P.: Fractal structures in fullerene layers: simulation of the growth process. J. Phys. Chem. C **112**, 4687–4695 (2008)
6. Germann, T.C., Kadau, K.: Trillion-atom molecular dynamics becomes a reality. Int. J. Mod. Phys. C **19**, 1315–1319 (2008)
7. http://folding.stanford.edu/
8. Yamakov, V., Wolf, D., Phillpot, S.R., Mukherjee, A. K., Gleiter, H.: Dislocation processes in the deformation of nanocrystalline aluminium by molecular-dynamics simulation. Nat. Mater. **1**, 45–49 (2002)
9. Dongare, A.M., Rajendran, A.M., LaMattina, B., Zikry, M.A., Brenner, D.W.: Atomic scale studies of spall behavior in nanocrystalline Cu. J. Appl. Phys. **108**, 113518 (2010)
10. Ju, S., Liang, X.: Investigation of argon nanocrystalline thermal conductivity by molecular dynamics simulation. J. Appl. Phys. **108**, 104307 (2010)
11. Xiao, S., Hu, W., Yang, J.: Melting behaviors of nanocrystalline Ag. J. Phys. Chem. B **109**, 20339–20342 (2005)
12. Lin, Z., Bringa, E.M., Leveugle, E., Zhigilei, L.V.: Molecular dynamics simulation of laser melting of nanocrystalline Au. J. Phys. Chem. C **114**, 5686–5699 (2010)
13. Rudd, R.E., Broughton, J.Q.: Coarse-grained molecular dynamics and the atomic limit of finite elements. Phys. Rev. B **58**, R5893–R5896 (1998)
14. Zwanzig, R.: Memory effects in irreversible thermodynamics. Phys. Rev. **124**, 983–992 (1961)
15. Mori, H.: Transport, collective motion, and Brownian motion. Prog. Theor. Phys. **33**, 423–455 (1965)
16. Hijón, C., Español, P., Vanden-Eijnden, E., Delgado-Buscalioni, R.: Mori-Zwanzig formalism as a practical computational tool. Faraday Discuss. **144**, 301–322 (2010)
17. Akkermans, R.L.C., Briels, W.J.: Coarse-grained dynamics of one chain in a polymer melt. J. Chem. Phys. **113**, 6409–6422 (2000)
18. Li, Z., Bian, X., Caswell, B., Karniadakis, G.E.: Construction of dissipative particle dynamics models for complex fluids via the Mori–Zwanzig formulation. Soft Matter. **10**, 8659–8672 (2014)
19. Peter, C., Kremer, K.: Multiscale simulation of soft matter systems. Faraday Discuss. **144**, 9–24 (2010)
20. Colbourn, E.A. (ed.): Computer Simulation of Polymers. Longman Scientific and Technical, Harlow (1994)
21. Zhigilei, L.V., Leveugle, E., Garrison, B.J., Yingling, Y.G., Zeifman, M.I.: Computer simulations of laser ablation of molecular substrates. Chem. Rev. **103**, 321–348 (2003)
22. Zhigilei, L.V., Wei, C., Srivastava, D.: Mesoscopic model for dynamic simulations of carbon nanotubes. Phys. Rev. B **71**, 165417 (2005)
23. Volkov, A.N., Zhigilei, L.V.: Mesoscopic interaction potential for carbon nanotubes of arbitrary length and orientation. J. Phys. Chem. C **114**, 5513–5531 (2010)
24. Buehler, M.J.: Mesoscale modeling of mechanics of carbon nanotubes: self-assembly, self-folding, and fracture. J. Mater. Res. **21**, 2855–2869 (2006)
25. Ostanin, I., Ballarini, R., Potyondy, D., Dumitrică, T.: A distinct element method for large scale simulations of carbon nanotube assemblies. J. Mech. Phys. Solids **61**, 762–782 (2013)
26. Torabi, H., Radhakrishnan, H., Mesarovic, S.D.J.: Micromechanics of collective buckling in CNT turfs. J. Mech. Phys. Solids **72**, 144–160 (2014)
27. Raabe, D.: Computational Materials Science: The Simulation of Materials Microstructures and Properties. Wiley-VCH, Weinheim/New York (1998)
28. Kirchner, H.O., Kubin, L.P., Pontikis, V. (eds.): Computer simulation in materials science. Nano/meso/macroscopic space and time scales. Kluwer Academic, Dordrecht/Boston/London (1996)
29. Groh, S., Zbib, H.M.: Advances in discrete dislocations dynamics and multiscale modeling. J. Eng. Mater. Technol. **131**, 041209 (2009)
30. Holm, E.A., Battaile, C.C.: The computer simulation of microstructural evolution. J Miner. Met. Mater. Soc. **53**, 20–23 (2001)
31. Akasheh, F., Zbib, H.M., Hirth, J.P., Hoagland, R.G., Misra, A.: Dislocation dynamics analysis of dislocation intersections in nanoscale metallic multilayered composites. J. Appl. Phys. **101**, 084314 (2007)
32. Leveugle, E., Zhigilei, L.V.: Molecular dynamics simulation study of the ejection and transport of polymer molecules in matrix-assisted pulsed laser evaporation. J. Appl. Phys. **102**, 074914 (2007)
33. Volkov, A.N., Zhigilei, L.V.: Structural stability of carbon nanotube films: the role of bending buckling. ACS Nano **4**, 6187–6195 (2010)
34. Jacobs, W.M., Nicholson, D.A., Zemer, H., Volkov, A. N., Zhigilei, L.V.: Acoustic energy dissipation and thermalization in carbon nanotubes: atomistic modeling and mesoscopic description. Phys. Rev. B **86**, 165414 (2012)
35. Volkov, A.N., Zhigilei, L.V.: Scaling laws and mesoscopic modeling of thermal conductivity in carbon nanotube materials. Phys. Rev. Lett. **104**, 215902 (2010)
36. Volkov, A.N., Zhigilei, L.V.: Heat conduction in carbon nanotube materials: strong effect of intrinsic thermal conductivity of carbon nanotubes. Appl. Phys. Lett. **101**, 043113 (2012)

Computational Systems Bioinformatics for RNAi

Zheng Yin, Yubo Fan and Stephen TC Wong
Center for Bioengineering and Informatics, Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston, TX, USA

Synonyms

Automatic data analysis workflow for RNAi; Systems level data mining for RNAi

Definition

Computational systems bioinformatics for RNAi screening and therapeutics is defined as complete computational workflow applicable to the hypothesis generation from large-scale data from image-based RNAi screenings as well as the improvement of RNAi-based therapeutics; the workflow includes automatic image analysis compatible to large-scale cell image data, together with unbiased statistical analysis and gene function annotation.

Introduction

RNA interference (RNAi) defines the phenomenon of small RNA molecules binding to its complementary sequence in certain messenger RNA, recruiting a specific protein complex to dissect the whole mRNA and thus silencing the expression of the corresponding gene. It is a highly conserved system within living cells to quantitatively control the activity of genes. In 1998, Fire et al. first clarified the causality of this phenomenon and named it as RNAi [1], and the following decade saw RNAi evolving into a powerful tool for gene function study. In 2006, Fire and Mello were awarded Nobel Prize in Physiology or Medicine; and by 2007, scientific papers using high-throughput screening based on RNAi kept piling up while clinical trials of RNAi-based therapeutics on various diseases raised the expectation on a trend of soon-to-come "super drugs." Unfortunately, by late 2009 nearly all the first trend clinical trials have been terminated, meanwhile, researchers are struggling to effectively quantify the high-content information obtained from RNAi-based screening experiments.

Although facing obstacles, it is still believed that the combination of nanotechnology and systems biology would restore and amplify the glory of RNAi on both research and therapeutic areas. This entry will summarize the challenges facing RNAi-based therapeutics – especially the difficulty of delivery and some possible solution through chemoinformatics in nanometer scale; also difficulties facing RNAi-based high-content screening will be reviewed with suggestions on possible solutions.

RNAi-Based Therapeutics: How-to and What-to Deliver

Currently, various RNAi-based therapeutics are being tested in clinical trials but before the application becomes clinical, several problems have first to be solved. One of the critical challenges is how to specifically and effectively deliver the objects into the targeted cells. Serious side-effects in patients could be caused by off-target effects or immune response as reported in literatures [2, 3].

Apparently, the therapeutic goal is to achieve RNAi therapy, that is, systemically administered nucleic acids must survive in circulation long enough to reach their target tissue, enter the desired cells, "escape" their endosome or delivery packaging, and finally become incorporated into the RNA-induced silencing complex (RISC) – a towering task, and surprisingly, researchers have advanced a number of plausible solutions in recent years, including the use of specialized nanoparticle filled with an RNAi-based cancer therapy to target human cancer cells and silence the target gene. However, what making the task even more difficult for therapeutics is: everything now happens in vivo.

Chemoinformatics Solutions

Libraries have been built based on natural products and combinatorial chemistry with millions of compounds to date. The compound library is used to screen and locate small molecules to bind a particular protein, RNA or DNA. Virtual screening utilizing molecule fragments even can design unknown compounds with reasonable binding affinity to desired targets. However, these large-scale screening compounds tend to bind multiple targets (off-target effects) even after optimizations [4]. Also, it is virtually impossible to derive a solid algorithm or theory to screen and locate all bindings and inhibitions and their combinations every protein currently know. The lack of genome-scale coverage is a clear disadvantage of a compound based assay. Along with the lack of specificity comes the challenge of on-target potency. Even after a target has been identified the drug-like compounds have to be further engineered to increase efficiency and decrease off-target activities.

To a large extent, chemoinformatic approaches, where the target of a compound can be predicted by in silico alignment, modeling algorithms, and virtual screening, provide a more efficient way to screen compounds to the desired targets. A successful application has been reported for the inhibition of SARS protease [5]. The chemoinformatics approach excels at optimization around a small number of well defined targets, while RNAi approaches can more readily identify unknown pathways and phenotypes with no prior knowledge of the target.

Data Analysis for RNAi HCS: Challenge from Millions of Cells

Large volumes of datasets generated from RNAi HCS of RNAi prohibits manual or even semi-manual analysis; thus, automated data analysis is desperately needed [6, 7]. In the context of genome-wide RNAi HCS using cultured cells from Drosophila, a series of automated methods on cell image processing [8], online phenotype discovery [9, 10], cell classification, and gene function annotation [11] have been developed. All these methods are integrated into an automated data analysis pipeline, G-CellIQ (Genomic Cellular Image Quantitator), to support genome-wide RNAi HCS.

A lot of decisions need to be made en route to a genome-scale RNAi screens, including the selections of appropriate animal models, reagents for igniting RNAi, screening formats, and type of readouts (see [12] for a review). The focus here is arrayed high-content RNAi screening, a systematic screen with reagents spotted in 384- or 96-well plates and where each gene or gene group is knocked down individually. The readout of HCS is obtained through microscopy, which captures multiple phenotypic features simultaneously [13]. Compared with other screening approaches, RNAi HCS can offer broader insight into cellular physiology and provide informative and continuous phenotypic data generated by RNAi. However, the automated processing and analysis of the magnitude of image data presents great challenges to applying such findings at the genome level.

Image Processing: Cell Segmentation and Quantification

The scale of datasets has always been a huge obstacle. For example, in [6], approximately 17000 overlapping cells were segmented semi-manually across 10 months from an HCS targeting around 200 genes or gene combinations. However, such low-dimensional screens are not genome-wide, as they can only cover 1–2 % of the genome. For automatic data analysis, the following work needs to be accomplished.

1. Preprocessing: where empty images, incomplete cells, and other artifacts are identified and discarded
2. Cell segmentation: where dense or overlapping cells are segmented accurately
3. Feature extraction: where informative features are extracted to quantify cell morphology

Cell segmentation is the cornerstone for the whole data analysis workflow. While existing image analysis methods can handle the processing of standard images, they are limited in their scope

and capability to handle genome-wide RNAi HCS analysis. Thresholding methods basically set a cutoff on the intensity of pixels and classify them into background and foreground (cell), and they may fail due to uneven background and illumination levels. Rule-based correction on over- and under-segmentation starts from relatively simple segmentation methods (like watershed), and use a series of heuristic rules, like distance between neighborhood nuclei and properties of putative cell boundaries, such methods suffer from difficulties in devising rules to merge the cell cytoplasm.

Phenotype Identification, Validation, and Classification

Given the quantified features describing cell morphology, the following work is essential to address the biological function of morphological profiles.

1. Define biologically meaningful phenotypes to compose cell populations based on single cell morphology.
2. Model existing phenotypes and identify novel phenotypes online to continuously generate new data.
3. Assign cellular phenotypes into different sub-phenotypes to address morphological changes caused by RNAi treatment.

Statistical tests, artificial neural networks [6], Support Vector Machine-Recursive Feature Elimination (SVM-RFE) [14], genetic algorithms, and various other methods have been used to select or extract informative subsets of features and model certain phenotypes. However, phenotypes are usually defined a priori from pilot datasets. Human intervention is currently necessary for image-based datasets of genetic or chemical perturbations where the dynamic range of cellular phenotypes cannot be predicted before data collection. Failing to accurately measure phenotypic variations will cause concomitant classification errors and mislead functional analysis; and it is impossible to perform manual analysis during the screening process where millions of images are acquired. Thus, the ability of these screens to identify new phenotypes is greatly limited [9].

Statistical Analysis and Gene Function Annotation

RNAi HCS inherits various statistical questions from traditional high-throughput screening, and the properties of image readouts raise specific challenges.

1. Summarizing gene function scores from quantified morphological change
2. Data triage and normalization based on readout from positive and negative control wells
3. Repeatability test and consolidation of scores from biological replicates
4. Cluster analysis, visualization, and biological interpretation of results

A comprehensive review [15] summarizes different dataset generated by RNAi screens and traditional small molecule screens. It also reviews statistical analysis methods applicable to most of problems outlined above. Quantitative morphological signatures (QMS) [6] represent the efforts on interpreting RNAi HCS datasets: use the similarity score to a panel of existing phenotypes to explore the broad phenotypic space. However, problems remain open when the discriminative ability of such scores is confounded by multiple phenotypes following a single RNAi treatment, so repeatability tests can become more complicated [6]. The use of publicly available databases on drugs and disease-related biological processes to interpret RNAi HCS datasets also remains unresolved.

G-CellIQ: An Integrated Automated Data Analysis Tool for RNAi High-Content Screening

Computational Architecture of G-CellIQ

G-CellIQ (**G**enomic **CELL**ular **I**maging **Q**uantitator) is developed to process large volumes of digital images generated from large-scale HCS studies. The workflow for G-CellIQ can be simplified as "three modules handling three databases." Images generated from HCS are stored in a Raw image database; the Image processing/cell morphology quantification

module segments each image into single cells and creates a quantified cell database; the Phenotype modeling and cell classification module compares each cell's morphology to a panel of "reference" phenotypes (which can be defined both manually and automatically) and generates a morphology score; the annotation of gene function module then summarize single cell scores into scores for cell populations, images and wells, and the consolidated scores for involved genes form a single gene function profile database.

Image Processing and Cell Morphology Quantification

A segmentation method consisting of nuclear segmentation, cell body segmentation, and over-segmentation correction [8] is used in G-CELLIQ. Each cell body is described by 211 morphological features; automatic image quality control is applied to filter images where the signal from certain channels is extremely dark or bright or cells are located at the edge of an image.

Nuclear segmentation cell nuclei are first separated from background using a binarization method which implements an adaptive thresholding [8]. To segment clustered nuclei, the nuclei shape and intensity information are integrated into a combined image and processed with Gaussian filter and the nuclei centers are detected as local maxima in the gradient vector field (GVF); after that, nuclei are segmented using the marker-controlled seeded-watershed method.

Cell body segmentation Preliminary cell body segmentation is done using an adaptive thresholding algorithm. Due to the large size of HCS dataset, seeded-watershed method is used to segment the touching cell bodies. Results from nuclei segmentation are used as the seed information [16].

Over-segmentation correction is necessary when there are multiple nuclei existing within cells. For cell segments smaller than a given size threshold, its neighboring cell segments sharing the longest common boundary are determined and the intensity variation in a rectangular region across the common boundary of touching cell segments are calculated. If the intensity variation was smaller than a given threshold, the corresponding cell segments are merged.

Feature extraction The detailed shape and boundary information of nuclei and cell bodies is obtained through the proposed segmentation method. To capture the geometric and appearance properties, 211 morphology features belonging to five categories were extracted following [11].

Online Phenotype Discovery, Phenotype Modeling, and Cell Classification

Expert opinion is implemented to identify a panel of reference phenotypes while candidate groups of informative features are selected to describe typical phenotypes. SVM-RFE and GA-SVM method are used to select the feature sets with the best performance for cross validation. A series of SVM classifiers are trained to differentiate the reference phenotypes from all others, and a continuous value rather than binary class label is used as the output of SVM to indicate each cell's morphological similarity to typical phenotypes.

A novel method is designed to do online phenotype discovery [9, 10]; it is based on online phenotype modeling and iterative phenotype merging. For the modeling part, each existing phenotype is modeled through a Gaussian Mixture Model (GMM), and each model is continuously updated according to information of newly incorporated cells with a minimum classification error (MCE) method. For the merging part, the newly generated cell population is iteratively combined with each existing phenotype, one at a time. Then an improved gap statistics method is used to identify the number of possible phenotypes in the combination. Through cluster analysis, some of the cells in the new populations are assigned into the same cluster as samples from existing phenotypes, and those cells are merged by existing phenotypes to help update the phenotype models. On the other hand, some cells are never merged and remain as candidate novel phenotype for validation by statistical tests [9, 10].

Annotation of Gene Function

After assigning each cell a score vector corresponding to its similarity to reference

phenotypes, all cells in control condition are pooled to model a baseline for cell morphology. All morphology scores are normalized to the Z-score relative to this control baseline. The scores for the qualified cells are then averaged to form scores for each well. In order to select repeatable wells from those undergoing the same dsRNA treatment, a series of repeatability tests are applied to scores for different well. The weighted average of the scores is then calculated for the repeatable wells and generated scores for each treatment condition (TC). Similar procedure consolidates the score from biological replicate TCs to form a score vector for each gene. Hierarchical clustering is implemented group genes with similar function scores into the same phenocluster.

RNAi HCS Applying G-CellIQ

Since more than 75 % of human disease genes have *Drosophila* orthologs, the pursuit of genes involved in normal fly morphogenesis and migration is expected to reveal mechanisms conserved in humans [17]. An example of using G-CellIQ in the context of RNAi HCS using cultured *Drosophila* cell lines is presented next.

Regulatory of cell shape change. A genome-scale RNAi screening is carried out for regulators of *Drosophila* cell shape. Drosophila Kc187 cell lines are utilized in the screen and wild-type cells have hemocyte-like properties. Using dsRNA to target and inhibit the activity of specific genes/proteins, the role of individual genes in regulating morphology can be systematically determined.

Work in [9, 10] relies on part of the genome dataset to target the group of kinase-phosphatases in order to develop and validate online phenotype discovery methods. A panel of five existing phenotypes is set by expert labeling and online phenotype discovery. In order to address the level of penetrance in this dataset, the phenotypic scores for single cells are sorted according to similarity to wild-type cells, and if a certain well has significantly less (at least one standard-deviation) wild-type cells than control wells, wild-type cells from this well are removed to reveal the phenotypic change relating to RNAi treatment.

Roles of Rho family small GTPases in development and cancer. Three automated and *quantitative* genome-wide screens for dsRNAs that induce the loss of the Rho-induced cytoskeletal structures (lamellipodia, filopodia, and stress fibers) are performed to identify putative Rho protein effectors. Candidate effectors identified in such image-based screens are readily validated in the context of the whole organism using the large number of mutant fly lines coupled with the vast arrays of in vivo techniques available to fly biologists [18].

Following image segmentation, feature extraction, classification of single cells, and scoring of individual wells, an image descriptor is assigned to each gene. Hierarchical clustering can then be used to cluster the image descriptor and identify groups of genes when targeted by RNAi result in quantitatively similar morphologies. Previous studies reported in [6] demonstrated that clustering results identifies groups of functionally related genes that operate in similar signaling pathways. These groups of genes are termed as "Phenoclusters" [19]. To validate the hypothesis generation ability of G-CELLIQ, 32 dsRNAs/wells are randomly selected from a dataset where individual kinases and phosphatases were inhibited by RNAi. Each dsRNA/well was assigned an image descriptor, and hierarchical clustering was used to group genes/wells. dsRNAs in this analysis clustered into two broad groups. One group of 19 conditions included 10/10 control conditions, as well as dsRNAs targeting the Insulin receptor (InR). Strikingly, the other large cluster of 13 conditions included 3/3 dsRNAs previously identified in a genome-wide screen for regulators of MAPK/ERK activation downstream of the EGF/EGFR activity [20]. These results demonstrate that automated high-throughput imaging can discriminate distinct morphologies and be used to model functional relationships between signaling molecules.

Cross-References

▶ RNAi in Biomedicine and Drug Delivery

References

1. Fire, A., et al.: Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature **391**(6669), 806–811 (1998)
2. Hornung, V., et al.: Sequence-specific potent induction of IFN-[alpha] by short interfering RNA in plasmacytoid dendritic cells through TLR7. Nat. Med. **11**(3), 263–270 (2005)
3. Grimm, D., et al.: Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. Nature **441**(7092), 537–541 (2006)
4. Copeland, R.A., Pompliano, D.L., Meek, T.D.: Drug-target residence time and its implications for lead optimization. Nat. Rev. Drug Discov. **5**(9), 730–739 (2006)
5. Plewczynski, D., et al.: In silico prediction of SARS protease inhibitors by virtual high throughput screening. Chem. Biol. Drug Des. **69**(4), 269–279 (2007)
6. Bakal, C., et al.: Quantitative morphological signatures define local signaling networks regulating cell morphology. Science **316**, 1753–1756 (2007)
7. Zhou, X., Wong, S.T.C.: Computational systems bioinformatics and bioimaging for pathway analysis and drug screening. Proc. IEEE **96**(8), 1310–1331 (2008)
8. Li, F.H., et al.: High content image analysis for human H4 neuroglioma cells exposed to CuO nanoparticles. BMC Biotechnol. **7**, 66 (2007)
9. Yin, Z., et al.: Using iterative cluster merging with improved gap statistics to perform online phenotype discovery in the context of high-throughput RNAi screens. BMC Bioinform. **9**(1), 264 (2008)
10. Yin, Z., et al.: Online phenotype discovery based on minimum classification error model. Pattern Recogn. **42**(4), 509–522 (2009)
11. Wang, J., et al.: Cellular phenotype recognition for high-content RNA interference genome-wide screening. J. Mol. Screen. **13**(1), 29–39 (2008)
12. Perrimon, N., Mathey-Prevot, B.: Applications of high-throughput RNAi screens to problems in cell and developmental biology. Genetics **175**, 7–16 (2007)
13. Carpenter, A.E., Sabatini, D.M.: Systematic genome-wide screens of gene function. Nat. Rev. Genet. **5**(1), 11–22 (2004)
14. Loo, L., Wu, L., Altshuler, S.: Image based multivariate profiling of drug responses from single cells. Nat. Methods **4**(5), 445–453 (2007)
15. Birmingham, A., et al.: Statistical methods for analysis of high-throughput RNA interference screens. Nat. Methods **6**(8), 569–575 (2009)
16. Yan, P., et al.: Automatic segmentation of RNAi fluorescent cellular images with interaction model. IEEE Trans. Inf. Technol. Biomed. **12**(1), 109–117 (2008)
17. Reiter, L.T., et al.: A systematic analysis of human disease-associated gene sequences in *Drosophila melanogaster*. Genome Res. **11**, 1114–1125 (2001)
18. Bier, E.: Drosophila, the golden bug, emerges as a tool for human genetics. Nat. Rev. Genet. **6**(1), 9–23 (2005)
19. Piano, F., et al.: Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. Curr. Biol. **12**(22), 1959–1964 (2002)
20. Friedman, A., Perrimon, N.: Functional genomic RNAi screen for novel regulators of RTK/ERK signaling. Nature **444**, 230–234 (2006)

Computer Modeling and Simulation of Materials

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Computer Modeling of Electrochemical Systems

▶ Electrochemical Interfaces for Energy Storage and Conversion

Concentration Polarization

▶ Concentration Polarization at Micro-/Nanofluidic Interfaces

Concentration Polarization at Micro-/Nanofluidic Interfaces

Vishal V. R. Nandigana and N. R. Aluru
Department of Mechanical Science and Engineering, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana – Champaign, Urbana, IL, USA

Synonyms

Concentration polarization; Micro-/Nanofluidic devices; Nonlinear electrokinetic transport

Definition

Concentration polarization (CP) is a complex phenomenon observed at the interfaces of micro/nanofluidic devices due to the formation of significant concentration gradients in the electrolyte solution resulting in accumulation and depletion of ions near the interfaces.

This chapter provides an overview of the underlying theory and physics that is predominantly observed on the integration of microfluidic channels with nanofluidic devices. Nonlinear electrokinetic transport and concentration polarization phenomenon are discussed in detail along with recent advancements in utilizing this phenomenon for designing novel devices.

Electrical Double Layer (EDL) and Electroosmotic Flow

A solid in contact with an aqueous solution acquires a surface charge (σ_s) due to the dissociation of ionizable groups on the solid walls. The fixed surface charge on the solid surface in contact with the liquid develops a region of counterions (ions with charges opposite to the solid surface) in the liquid to maintain the electroneutrality at the solid–liquid interface. This screening region is denoted as the electrical double layer (EDL) or Debye length (DL). For instance, in the case of silica channel with *KCl* electrolyte solution, the dissociation of silanol groups would make the channel negatively charged and affect the distribution of K^+ counterions in the solution. The layer where the ions get strongly attracted toward the channel surface due to the electrostatic force is called the inner layer with a typical thickness of one ion diameter. The outer Helmholtz plane separating the liquid and the diffusive layer constitute the liquid side part of the EDL. The ionic species in the diffusive layer are influenced by the local electrostatic potential, and the species distribution at equilibrium can be described by the Boltzmann equation. The thickness of the diffusive layer spans between 1 and 100 nm. The EDL thickness (λ_D) is given by [1]:

$$\lambda_D = \left(\frac{\varepsilon_0 \varepsilon_r RT}{F^2 \sum_{i=1}^{m} z_i^2 c_0} \right) 1/2 \qquad (1)$$

where F is Faraday's constant, z_i is the valence of ionic species i, c_0 is the bulk concentration of the electrolyte solution, ε_0 is the permittivity of free space, ε_r is the relative permittivity of the medium, m is the total number of ionic species, R is the universal gas constant, and T is the absolute temperature.

The thickness of the EDL plays a significant role in the transport of miniaturized devices. In microfluidic channels, the fluid transport is often controlled by electric fields, as it eliminates the use of external mechanical devices [1]. The electric field aids better control when compared to using pressure-driven techniques. Furthermore, the electric fields overcome the high pressures needed to transport the fluid at such length scales as the pressure follows a power–law relation with respect to the height (h) of the channel. The electric field acts on the charged counterions present at the interface of solution and stationary charged wall (i.e., at the EDL regions), resulting in the motion of the fluid which is referred as electroosmotic flow (EOF) [1]. As the EDL thickness in these channels is much smaller compared to their height $\left(\frac{h}{\lambda_D} \gg 1\right)$, the fluid flow has a plug-like flow characteristic. However, recent advancements in the fabrication technology [2] have motivated researchers around the globe to investigate the transport phenomenon in channel sizes of the order of few hundreds of nanometers. Transport in these devices is referred to as "nanofluidics." The electrical double layer in these devices spans much of the diameter or channel height leading to many interesting transport phenomena compared to its microscopic counterpart. The electroosmotic velocity no longer follows a plug-like flow characteristic but follows a Poiseuille-like (parabolic) characteristic as the electrokinetic body force is not just confined to a thin layer adjacent to the channel surface. Along with the aforementioned difference, the micro and nanofluidic systems also exhibit a different ion transport characteristic which is discussed below. In nanofluidic systems, as the EDL thickness

becomes comparable to the channel height $\left(\frac{h}{\lambda_D} \approx 1\right)$, there is a predominant transport of the counterions inside the channel, thus enabling the channel to be ion-selective [1]. These features are not observed in the microfluidic channels, as the counter-ionic space charge is confined to a very thin layer adjacent to the surface and the region away from the surface is essentially quasi-electroneutral (i.e., both co-ions and counterions are present away from the surface). Along with the EDL, the surface charge also plays a prominent role in controlling the transport inside the nanofluidic systems [2]. Similar ion-selective phenomenon was also observed in the intraparticle and intraskeleton mesopores of particulate and in membrane science [3].

Owing to the differences in the electrokinetic transport phenomena between the micro- and nanofluidic devices, the integration of these two devices paves way to complex physics. The models and the underlying theory developed to understand the electrokinetic transport in such systems are elaborated in section "Electrokinetic Theory for Micro/Nanochannels." A detailed discussion on the concentration polarization phenomenon and its applications are presented in section "Concentration Polarization." Finally, a brief summary is presented in section "Summary."

Electrokinetic Theory for Micro/Nanochannels

A complete set of equations for modeling the electrokinetic transport and to account for the EDL effects in micro/nanofluidic channels are presented. To understand the electrokinetic transport, space charge model developed by Gross et al. [4] is used extensively in the literature. The model solves the classical Poisson–Nernst–Planck (PNP) equations, which describe the electrochemical transport and the incompressible Navier–Stokes along with the continuity equations are solved to describe the movement of the fluid flow. These coupled systems of equations are more intensive, mathematically complicated, and computationally expensive. Though many linearized approximations were proposed to this model to study the electrokinetic transport [2], the governing equations of the complete nonlinear space charge model is discussed in this chapter.

In electrokinetic flows, the total flux is contributed by three terms: a diffusive component resulting from the concentration gradient, an electrophoretic component arising due to the potential gradient, and a convective component originating from the fluid flow. The total flux of each species in the solution is given by

$$\boldsymbol{\Gamma}_i = -D_i \nabla c_i - \Omega_i z_i F c_i \nabla \phi + c_i \boldsymbol{u} \tag{2}$$

where $\boldsymbol{\Gamma}_i$ is the flux vector, D_i is the diffusion coefficient, Ω_i is the ionic mobility, c_i is the concentration of the i^{th} species, $\boldsymbol{u}$ is the velocity vector of the fluid flow, and ϕ is the electrical potential. Note that the ionic mobility is related to the diffusion coefficient by Einstein's relation, $\Omega_i = \frac{D_i}{RT}$ [5]. The electrical potential distribution is calculated by solving the Poisson equation,

$$\nabla \cdot (\in_r \nabla \phi) = -\frac{\rho_e}{\in_0} \tag{3}$$

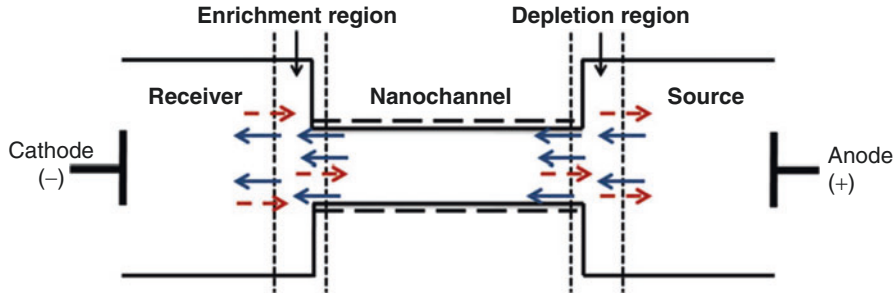
where ρ_e is the net space charge density of the ions defined as

$$\rho_e = F\left(\sum_{i=1}^{m} z_i c_i\right) \tag{4}$$

The mass transfer of each buffer species is given by the Nernst–Planck equation,

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot \Gamma_{\mathrm{i}} \tag{5}$$

Equations 3, 5, and 2 are the classical Poisson–Nernst–Planck (PNP) equations, which describe the electrochemical transport. The incompressible Navier–Stokes and the continuity equations are considered to describe the movement of the fluid flow through the channel, i.e.,



Concentration Polarization at Micro-/Nanofluidic Interfaces, Fig. 1 Schematic illustration of ion-enrichment and ion-depletion effect in cation-selective micro/nanofluidic channel. The *solid arrows* indicate the flux of cations and the *dotted arrows* indicate the flux of anions. At the nanochannel–anodic junction, both the anions and cations are depleted, while there is an enhancement of both the ions at the cathode–nanochannel junction

$$\rho\left(\frac{\partial \boldsymbol{u}}{\partial t}+\boldsymbol{u}\cdot\nabla\boldsymbol{u}\right)=-\nabla p+\mu\nabla^2\boldsymbol{u}+\rho_e\boldsymbol{E} \quad (6)$$

$$\nabla\cdot\boldsymbol{u}=0 \quad (7)$$

where $\boldsymbol{u}$ is the velocity vector, p is the pressure, ρ and μ are the density and the viscosity of the fluid, respectively, and $\boldsymbol{E}=-\nabla\phi$ is the electric field. $\rho_e\boldsymbol{E}$ is the electrostatic body force acting on the fluid due to the space charge density and the applied electric field. Elaborate details on other simplified models are discussed in the review article of Schoch et al. [2].

Concentration Polarization

Concentration polarization (CP) is a complex phenomenon observed at the interface regions of micro/nanofluidic devices due to the formation of significant concentration gradients in the electrolyte solution near the interfaces causing accumulation of ions on the cathodic side and depletion of ions on the anodic side for a negatively charged nanochannel surface. This phenomenon was also observed in the field of colloid science and in membrane science which was extensively studied for over 40 years, and the early works of CP phenomenon is comprehensively reviewed by Rubinstein et al. [6]. The pioneering works from Rubinstein and his coworkers had revealed electrokinetic instabilities [7] in the concentration polarization regions leading to the breakdown of limiting current and resulting in the overlimiting conductance regimes in the ion exchange membranes. Such complex phenomenon could not be postulated using the classical equilibrium model of EDL [6]. All the underlying CP physics observed near the interfaces of micro/nanochannels are summarized in the following subsections.

Enrichment/Depletion Effects

In micro/nanofluidic devices, Pu et al. [8] first experimentally observed the CP effects near the interfaces and provided a simple model to explain the accumulation and depletion physics which is summarized below. For a negatively charged nanochannel, the EDL would be positively charged. For an overlapped EDL, as discussed in section "Electrical Double Layer (EDL) and Electroosmotic Flow," the nanochannel becomes ion-selective, resulting in higher cation concentration than anions. Thus, the flux of cations is higher compared to the anions in the nanochannel. With the application of positive potential at the source microchannel or reservoir (see Fig. 1), the cations move from the source (anode) reservoir to the receiving (cathode) reservoir end, while the anions move in the opposite direction through the nanochannel. At the cathodic side, the anion flux from the ends of reservoir to the nanochannel junction is higher compared to the anion flux from the junctions to the nanochannel as the anions are repelled by the negatively charged nanochannel. This difference in fluxes causes an accumulation

of anions at the cathode–nanochannel junction. The cation flux from the nanochannel to the cathode junction is greater than from the cathode junction to the reservoir as the cations have to balance the anions present at this junction. This results in an accumulation of cations as well at the nanochannel–cathode junction. At the anodic side, the anion flux from the nanochannel to the anode junction cannot balance the anion flux from the anode–nanochannel junction to the reservoirs due to the limited anions passing through the nanochannel. This results in a depletion of anions at this junction. The cation flux from the reservoir to the anode–nanochannel junction is less than the cation flux entering the nanochannel as the cations are attracted by the positively charged nanochannel. This in turn leads to the depletion of cations at this junction. To summarize, for a negatively charged nanochannel, both the cations and anions accumulate at the cathodic interface and are depleted at the anodic interface. The phenomenon is reversed for a positively charged nanochannel surface. A schematic diagram highlighting the accumulation and the depletion physics for a negatively charged nanochannel is displayed in Fig. 1.

Nonlinear Electroosmosis

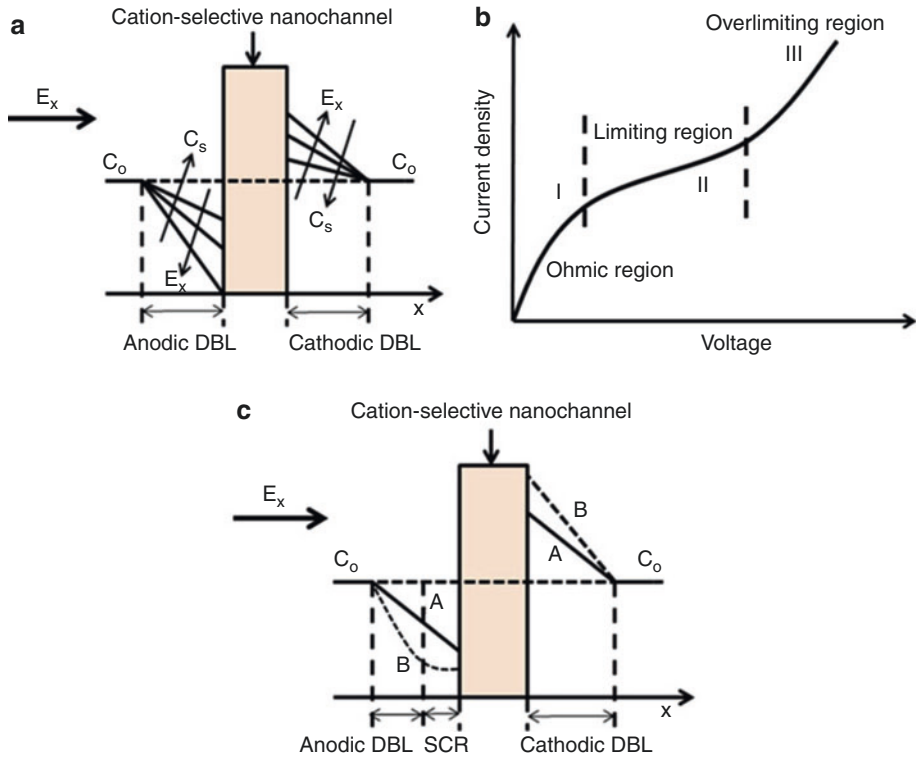
As discussed in the previous section, the integration of micro/nanofluidic devices leads to the accumulation and depletion of ions near the interfaces. Several numerical and experimental studies were carried out to gain a better understanding of the physics at these interfaces. The studies revealed complex and interesting physics at the depletion interface compared to the enrichment side. Rubinstein et al. [6] theoretically predicted the presence of space charges at the depletion region under large electric fields. The presence of the induced space charges near the depletion interface results in a nonequilibrium electrical double layer outside the nanochannel. The induced space charges under the action of the external electric field lead to nonlinear electroosmosis or otherwise known as electroosmosis of the second kind. The electroosmotic flow of the second kind was found to be directly proportional to the square of the applied electric field. Furthermore, the induced space charges also result in the generation of vortices at the depletion interface along with inducing large pressure and voltage gradients at this junction. Jin et al. [9] also reported similar physics from their extensive numerical study. In the case of a flat ion exchange membrane, Rubinstein et al. [7] derived a 2D nonequilibrium electroosmotic slip (u_s) for an applied voltage (V) using the linear stability analysis to impose strong vortex field near the membrane:

$$u_s = -\frac{1}{8}V^2 \frac{\overline{\frac{\partial^2 c}{\partial x \partial y}}}{\overline{\frac{\partial c}{\partial y}}} \tag{8}$$

where x and y are the axes parallel and perpendicular to the ion-exchange membrane, respectively. Experiments performed by Kim et al. [10] also reveal the nonequilibrium EOF near the micro/nanofluidic junctions. The application of electric field on the surface of particles also results in such induced space charges which spread over a larger region than the primary EDL resulting in highly chaotic flow patterns. A recent review by Höltzel and Tallarek [3] provide a detailed discussion on the polarization effects around membranes, packed beds, and glass monoliths. Recent advancements by Rubinstein and Zaltzman [11], however, revealed that the extended space charge region was not a part of the EDL, but develop from the counterion concentration minimum zone with the co-ions expelled under the action of the electric field. They further claim that the space charges would be present in the system even without equilibrium EDL. Their analysis included the study of the space charge dynamics in concentration polarization regions using 1D and three-layer models of EDL. From the understanding of these extended space charge layers, another important phenomenon, namely, the nonlinear current characteristics in micro/nanochannels is addressed below.

Nonlinear Current–Voltage Characteristics

There has been growing interest in the development of micro/nanofluidic devices as ionic filters



Concentration Polarization at Micro-/Nanofluidic Interfaces, Fig. 2 (**a**) Schematic distribution of ionic concentration in equilibrium concentration polarization under axial electric field (E_x). The local electroneutrality is maintained at both enrichment (cathodic interface) and depletion (anodic interface) diffusion boundary layers (DBL). The concentration gradients become steeper with the decrease in the ionic strengths (c_s) and at higher electric fields, (**b**) displays the nonlinear current–voltage characteristics for an ion-selective micro/nanochannel, and (**c**) shows the nonequilibrium concentration distribution due to the induced space charge region (SCR) (shown as *dotted lines* (*B*)) in the depleted region under very large electric fields

and nanofluidic batteries to control both ionic and molecular transport in aqueous solutions [2]. As discussed before, when the diameter/height of a charged channel scales comparable to the EDL, there is a predominant transport of the counterions inside the channel. Thus, the transport of electrical current inside the nanochannel is primarily due to the counterions. This feature enables the micro/nanochannel to be used as an ion exchange membrane. However, understanding the passage of ionic currents through such ion-selective solids is the most fundamental physical problem that has stimulated extensive research in this field for over a decade. Furthermore, the concentration polarization physics near the interfaces play a pivotal role in understanding the current–voltage characteristics. Figure 2a shows that the concentration gradients (CP regions) near the interfaces become steeper with the decrease in the ionic strength (c_s) and at higher electric fields ($\boldsymbol{E}_x$). At low electric fields, the current increases linearly with the applied voltage following the Ohm's law (region I in Fig. 2b).

However, at higher electric fields, the ion concentration in the depleted CP zone (i.e., near the anodic interface region) reaches toward zero and the classical Levich analysis [5] predicts a diffusion-limited current saturation according to which a saturation of current density occurs at a constant level described as the "limiting-current density" (region II in Fig. 2b).

The ionic current can be calculated considering the Fick's first law:

$$I = nFAD\frac{dc}{dx} \tag{9}$$

where n is the number of electrons transferred per molecule, D is the diffusion coefficient, and A is the electrode surface area. The concentration gradient is generally approximated by a linear variation (5),

$$I = nFAD\frac{c_0 - c(x=0)}{\delta} \tag{10}$$

where c_0 is the bulk electrolyte solution and δ is the diffusion boundary layer (DBL) thickness at the solid–liquid interface and $c(x=0)$ represents the concentration at the anodic (depletion) solid–liquid interface. δ typically ranges between 10 and 400 μm in ion exchange membranes [6], while it depends on the microchannel length in the case of micro/nanochannels [10].

From Eq. 10 it is clear that the current I reaches a maximum value when the concentration at $c(x=0)=0$, resulting in a limiting/saturation current as predicted by the classical Levich theory.

$$I_{lim} = \frac{nFADc_0}{\delta} \tag{11}$$

However, experimental studies in micro/nanofluidic devices (also in membrane science) revealed ionic currents larger than the limiting value and this regime was termed as the overlimiting current regime (region III in Fig. 2b). Further, the limiting current region is termed as the limiting resistance region (in micro/nanofluidic devices) due to the large but finite limiting differential resistance as the current does not saturate to a limiting value but has a slope which is smaller than the ohmic region. The nonlinear current characteristics in micro/nanochannels are a subject of intensive discussions in the literature [3]. Earlier studies indicated that water dissociation effects leading to the generation of H^+ and OH^- ions were responsible for such overlimiting currents. Later, Rubinstein et al. [7] postulated that some mechanism of mixing should be present that destroys the DBL as lower δ leads to higher currents (from Eq. 11). Maletzk et al. [12] coated the surface of cation-exchange membranes with a gel which does not allow mixing. From these experiments, they observed a saturation of current with no enhancements in the current, thereby confirming the earlier postulation of Rubinstein. Further, the fluid flow in the overlimiting regime revealed strong fluctuations indicating convection close to the surface. This convection was first attributed to the gravitational buoyant forces due to the concentration and temperature gradients. However, later theories have argued that the convection was not due to the gravitational instability in CP zones. Dukhin et al. [13] suggested the mechanism for mixing to be electroconvection. The type of electroconvection present in the overlimiting regime was revealed as the electroosmotic flow of the second kind. As discussed in the previous section, the induced space charges in the depletion zone (see Fig. 2c) under the action of the electric field results in the EOF of second kind and this convective instability tends to destroy the DBL leading to overlimiting currents as shown in Fig. 2b. Experiments by Kim et al. [10] also revealed the nonlinear currents due to the nonequilibrium EOF in the ion-selective micro/nanochannels. Similar physics was also observed in the permselective membranes and ion-selective particles [3]. In spite of all these postulations, the physics behind the extended space charge layer still remains largely unclear and there are a lot of potential research opportunities to fully understand this complex physics.

Propagation of Concentration Polarization

In this section, the conditions and the scenarios which can lead to the propagation of CP in micro/nanofluidic devices are highlighted. Zangle et al. [14] highlighted the phenomena of concentration polarization propagation using a simplified model of charged species transport and validated the same by conducting experiments and by comparing with other experimental results. The CP phenomenon was found to be governed by a

type of Dukhin number, relating the bulk and the surface conductance. The inverse Dukhin number, for a symmetric electrolyte was specified as:

$$\frac{G_{bulk}}{G_\sigma} = \frac{Fhzc_0}{\sigma} \tag{12}$$

where G_{bulk} is the bulk conductance, G_σ is the surface conductance, c_0 is the concentration outside the EDL, and σ is the wall surface charge density. Zangle et al. postulated that CP depends on Dukhin number and not on the ratio of channel height to the Debye length $\left(\frac{h}{\lambda D}\right)$. Using their simplified model, they showed that both the enhancement and the depletion regimes at the interfaces of micro/nanofluidic channels propagate as shock waves under the following condition:

$$c^*_{o,r} h^*_n < \max\left(\upsilon^*_2, 2\upsilon^*_2 - 1\right) \tag{13}$$

where $c^*_{o,r} h^*_n = \frac{(\upsilon_1 z_1 - \upsilon_2 z_2) F h_n c_{o,r}}{-2\upsilon_1 \sigma}$ is an inverse Dukhin number describing the ratio of bulk to surface conductance as mentioned before. $\upsilon^*_2 = \frac{\upsilon_2 z_2 F_\eta}{\zeta_n \in}$ is the mobility of the co-ion nondimensionalized by the electroosmotic mobility. $c_{o,r}$ is the reservoir electrolyte concentration, h_n is the nanochannel height, υ_1 and υ_2 are the mobilities, and z_1 and z_2 are the valences of the positive and negative ionic species, respectively. ζ_n is the nanochannel zeta potential, $\in$ is the permittivity and η is the viscosity. Elaborate details of the model can be referred in [14]. From this model, they proposed a thumb rule to avoid propagation of CP and it was found that $c^*_{o,r} h^*_n \gg 1$. This condition for CP propagation was compared with 56 sets of experimental literature values and was found to give a sufficient first-hand prediction with regard to the concentration polarization propagation.

Though the model considers the effects of surface charge and the electrolyte concentration, the finite *Pe* effects which also play a critical role in the concentration polarization were not considered. Further, the experiments of Kim et al. [10] and the numerical studies performed by Jin et al. [9] also revealed that the applied potential also plays a pivotal role in the concentration polarization generation and propagation apart from the inverse Dukhin number. The shortcoming of the model was also highlighted by Zangle et al. in their work. Thus, still a clear and complete understanding of the CP regimes is yet to be reached and continuous efforts are being made to understand the physics at the micro/nanofluidic junctions to design advanced and novel devices.

Applications

In this section, various applications that have been developed utilizing the concentration polarization phenomenon are addressed. The applications range from preconcentrating biomolecules to fluid pumping and mixing and also in water desalination. A brief discussion of the aforementioned applications is presented below.

Preconcentration

Wang et al. [15] used the depletion region observed at the micro/nanojunction to preconcentrate proteins. The energy barrier created at the depletion region (due to the large voltage drop induced at this junction) prevents the entry of charged molecules into the nanochannel. This results in an increase in the concentration of the molecules near the depletion region. In their experiments, Wang et al. used two anodic microchannels which were independently controlled so that the direction of EOF can be aligned perpendicular to the axis of the nanopore. An increase of about 10^6–10^8 fold in the concentration of the protein was reported in their study. Over the past couple of years, similar preconcentration devices utilizing the CP effects were experimentally fabricated [3, 14]. Wang et al. [16] also presented an experimental approach to improve the binding kinetics and the immunoassay detection sensitivity using concentration polarization in micro/nanofluidic devices. The antigens were preconcentrated at the depletion region due to the strong electric field gradients resulting in the enhancement in the binding rates with the antibody beads.

Seawater Desalination

The phenomenon of concentration polarization witnessed in ion-selective membranes was successfully implemented to address the freshwater shortage issue by providing energy-efficient solution to water desalination. A microfluidic device was fabricated which provides 99 % salt rejection at 50 % recovery rate at a power consumption of less than 3.5 Wh/L [17]. The CP depletion layer acts as a barrier for any charged species and these species were diverted away from the desalted water using suitable pressure and voltage fields. Their design also ensured salt ions and other debris to be driven away from the membrane thereby preventing any membrane fouling which is often observed in other desalination techniques.

Mixing, Pumping, and Other Applications

As discussed earlier, the induced space charges observed at the depletion region of micro/nanochannel along with the large electric field gradients at this region result in strong vortices. Kim et al. [18] enhanced the mixing efficiency of microfluidic devices using the vortices created at this interface. Further, as the electroosmotic flow of the second kind observed at the depletion region is directly proportional to the square of the applied electric field, Kim et al. [19] was able to pump fluids using the nonequilibrium EOF and observed a fivefold increase in the volumetric flow rates compared to similar devices utilizing equilibrium EOF. Yossifon et al. [20] used an asymmetric microchannel in conjunction with the nanochannels. The application of forward and reverse bias (at the overlimiting regime) voltage led to an asymmetric space charge polarization which resulted in the rectification of the current. Such membranes have potential applications in selective species separation.

Summary

The origin and the underlying physics that is present at the interfaces of micro/nanofluidic devices were discussed. The complex phenomenon of concentration polarization (CP) leading to the enrichment and depletion of ions near the micro–nano junctions and the concepts of induced space charges and nonlinear electrokinetic transport were briefly discussed in this chapter. Further, the various controversies surrounding the physical mechanism for nonlinear current characteristics have been highlighted. The criteria for concentration polarization propagation and the various applications that have been developed utilizing the concentration polarization phenomenon were discussed. Though, a lot of extensive work has been carried out to understand the CP physics, a clear and complete understanding of the CP regimes and the induced space charge dynamics is yet to be reached and efforts need to be directed in this area to understand the physics at the micro/nanofluidic junctions to design novel devices.

Cross-References

- ▶ Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering
- ▶ Electrokinetic Fluid Flow in Nanostructures
- ▶ Integration of Nanostructures within Microfluidic Devices
- ▶ Surface-Modified Microfluidics and Nanofluidics

References

1. Karniadakis, G.E., Beskok, A., Aluru, N.R.: Microflows and Nanoflows: Fundamentals and Simulation. Springer, New York (2005)
2. Schoch, R.B., Han, J., Renaud, P.: Transport phenomena in nanofluidics. Rev. Mod. Phys. **80**, 839–883 (2008)
3. Höltzel, A., Tallarek, U.: Ionic conductance of nanopores in microscale analysis systems: where microfluidics meets nanofluidics. J. Sep. Sci. **30**, 1398–1419 (2007)
4. Gross, R.J., Osterle, J.F.: Membrane transport characteristics of ultrafine capillaries. J. Chem. Phys. **49**, 228–234 (1968)
5. Probstein, R.F.: Physiochemical Hydrodynamics: An Introduction. Wiley, New York (1994)
6. Rubinstein, I.: Electrodiffusion of Ions. SIAM, Philadelphia (1990)

7. Rubinstein, I., Zaltzman, B.: Electro–osmotic slip of the second kind and instability in concentration polarization at electrodialysis membranes. Math. Models Methods Appl. Sci. **11**, 263–300 (2001)
8. Pu, Q., Yun, J., Temkin, H., Liu, S.: Ion–enrichment and ion–depletion effect of nanochannel structures. Nano Lett. **4**, 1099–1103 (2004)
9. Jin, X., Joseph, S., Gatimu, E.N., Bohn, P.W., Aluru, N.R.: Induced electrokinetic transport in micro – nanofluidic interconnect devices. Langmuir **23**, 13209–13222 (2007)
10. Kim, S.J., Wang, Y.-C., Lee, J.H., Jang, H., Han, J.: Concentration polarization and nonlinear electrokinetic flow near a nanofluidic channel. Phys. Rev. Lett. **99**, 044501 (2007)
11. Rubinstein, I., Zaltzman, B.: Dynamics of extended space charge in concentration polarization. Phys. Rev. E. **81**, 061502 (2010)
12. Maletzki, F., Rösler, H.-W., Staude, E.: Ion transfer across electrodialysis membranes in the overlimiting current range: stationary voltage current characteristics and current noise power spectra under different conditions of free convection. J. Membr. Sci. **71**, 105–116 (1992)
13. Dukhin, S.S.: Electrokinetic phenomena of the second kind and their applications. Adv. Colloid Interface Sci. **35**, 173–196 (1991)
14. Zangle, T.A., Mani, A., Santiago, J.G.: Theory and experiments of concentration polarization and ion focusing at microchannel and nanochannel interfaces. Chem. Soc. Rev. **39**, 1014–1035 (2010)
15. Wang, Y.-C., Stevens, A.L., Han, J.: Million–fold preconcentration of proteins and peptides by nanofluidic filter. Anal. Chem. **77**, 4293–4299 (2005)
16. Wang, Y.C., Han, J.: Pre-binding dynamic range and sensitivity enhancement for immuno-sensors using nanofluidic preconcentrator. Lab Chip **8**, 392–394 (2008)
17. Kim, S.J., Ko, S.H., Kang, K.H., Han, J.: Direct seawater desalination by ion concentration polarization. Nat. Nanotechnol. **5**, 297–301 (2010)
18. Kim, D., Raj, A., Zhu, L., Masel, R.I., Shannon, M.A.: Non-equilibrium electrokinetic micro/nano fluidic mixer. Lab Chip **8**, 625–628 (2008)
19. Kim, S.J., Li, L.D., Han, J.: Amplified electrokinetic response by concentration polarization near nanofluidic channel. Langmuir **25**, 7759–7765 (2009)
20. Yossifon, G., Chang, Y.-C., Chang, H.-C.: Rectification, gating voltage and interchannel communication of nanoslot arrays due to asymmetric entrance space charge polarization. Phys. Rev. Lett. **103**, 154502 (2009)

Conductance Injection

► Dynamic Clamp

Conduction Mechanisms in Organic Semiconductors

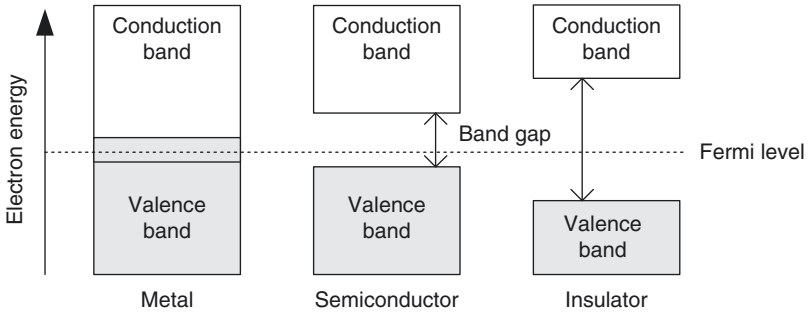
Weicong Li and Harry Kwok
Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

Definition

Conduction mechanisms in organic semiconductors refer to the means by which electronic charges move through organic semiconductors under external stress particularly under the influence of an electrical field.

Overview

In order to understand the conduction mechanisms in organic semiconductors, it is necessary to first introduce the concept of *band theory*, which is well established in solid-state physics. Solids in general are made of atoms, each of which is composed of a positively charged nucleus surrounded by negatively charged electrons. In quantum mechanical terms, these electrons effectively reside in discrete energy states in orbits. When a large number of atoms (of order 10^{20} or more) are brought together to form a solid, the discrete energy states are so close together that energy bands begin to form. At the same time, there will be gaps between the energy bands which are known as the *band gaps*. Because of the presence of these energy gaps, there will be some energy bands that are almost fully occupied (known as the *valence bands*) and energy bands that are almost unoccupied (known as the *conduction bands*). Based on the band theory, solids are typically divided into the following three categories: *metals*, *semiconductors*, and *insulators*. In metals, there is an overlap between the energy bands so that the energy bands are partly filled by electrons at any temperature (even $T = 0$ K), while both the semiconductors and the insulators have fully filled



Conduction Mechanisms in Organic Semiconductors, Fig. 1 Band structures of metal, semiconductor, and insulator

valence bands and empty conduction bands at $T = 0$ K. To further study the distribution of electrons occupying the energy states in solids, it is also necessary to introduce the concept of Fermi level into the band theory, which represents the maximum energy of states that electrons can occupy at $T = 0$ K. Accordingly, all the allowed energy states below the Fermi level are occupied by electrons, and all the energy states above it are empty. When temperature is above 0 K, the probability that electrons occupy the state with energy E under thermodynamic equilibrium condition is given by Fermi-Dirac distribution function:

$$f(E) = \frac{1}{1 + \exp[(E - E_F)/kT]} \qquad (1)$$

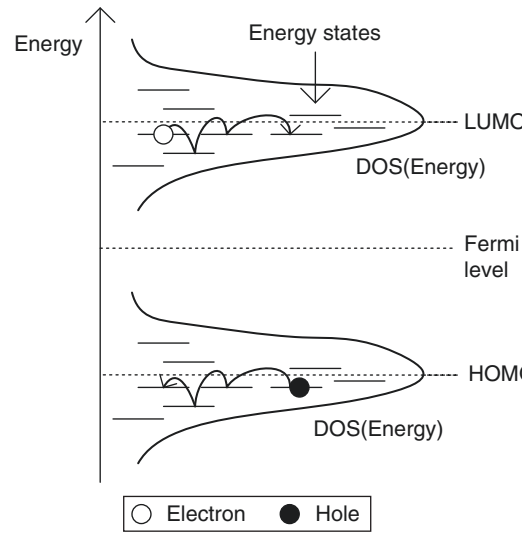
where E_F is the Fermi level, k is the Boltzmann constant, and T is the temperature in Kelvin. As mentioned earlier, in semiconductor and insulator, the valence band is fully occupied by electrons, and the conduction band is empty at $T = 0$ K. Therefore, one can infer that the Fermi level lies in the bandgap, between the valance and conduction bands. On the other hand, in metal, due to the fact that the energy bands are partly filled by electrons at any temperature, the Fermi level lies within the energy bands. The band structures of metal, semiconductor, and insulator, and the position of Fermi level in them are shown in Fig. 1. The distinction between the semiconductors and the insulators appears when temperature rises above 0 K. Because the band gap between conduction and valence bands in semiconductors is much narrower than that found in insulators, a fair amount of electrons can be thermally excited from the valence band to the conduction band in semiconductors at finite temperature, leading to measurable conductivity. This is not found in the insulators due to the larger band gaps even at room temperature, which lead to negligible probability of electrons occupying energy states in the conduction band, according to Eq. 1. The characteristic semiconductor band structure has allowed it to play an important role as the materials of choice in the prosperous electronic industry in the last few decades.

While inorganic semiconductors such as silicon dominated the electronic industry in the twentieth century, tremendous effort has been spent in the research and development of organic electronics in last decade due to the fact that organic semiconductors are usually easier and cheaper to form. Soluble organic materials, such as conjugated polymers, can be deposited in liquid phase (e.g., by printing and spin coating) onto large substrate areas at low processing temperature (below 100 °C). Due to this advantage, organic electronics are particularly attractive in the making of displays, sensors, light sources, photovoltaic panels, radio-frequency identification detectors (RFID), and in devices used in optical communications. As a consequence, research on the charge conduction mechanisms in organic semiconductors and devices is of significant importance.

In general, conduction mechanisms primarily describe how electronic charges (referred to as carriers) move inside the solids under the influence of an external electrical field. The process produces a current. At the macroscopic level, the current density J in solids produced by external electrical field is given by

$$J = env = en\mu F \tag{2}$$

where e is elementary charge of a single carrier, n the charge density, and v is the drift velocity. Furthermore, v can be expressed as the product of the charge mobility μ and the electrical field F. As can be seen in Eq. 2, a large current requires the presence of a substantial number of mobile charge carriers (electrons or holes). In organic semiconductors, mobile carriers are known to be produced from the distributed π-bonds, which are covalent chemical bonds resulting from the overlap of atomic orbitals. Thus, the limited current flow in many organic semiconductors are related to their irregular molecular structures which can result in low charge mobility in comparison to values found in silicon and other inorganic semiconductors. In addition, the more established conduction mechanisms based on band theory normally found in inorganic crystalline semiconductors are absent in the organic semiconductors. As mentioned earlier, band theory states that carriers could only exist and move in either the conduction bands or the valence bands because there are permitted energy states where carriers can reside and their movement between the energy states will produce a current. The use of "energy band" diagrams (see Fig. 1) to explain charge transport in organic semiconductors however is not possible. This is because of the presence of high densities of defects and trap states. Instead, charge transport in organic semiconductors is directly explained in terms of the energy (orbital) states which are termed either the *lowest unoccupied molecular orbital* (LUMO) or the *highest occupied molecular orbital* (HOMO). As such, LUMO and HOMO levels are not genuine energy bands and they are used merely to serve as references to demarcate ground state energy and the next activated state energy [1]. The distribution of these energy states known as the *density of states* (*DOS*) is usually considered to be Gaussian centered at the LUMO and HOMO (see Fig. 2). The width of the Gaussian energy states depends on both the regularity of the molecular structure and the impurities present in the organic semiconductor.

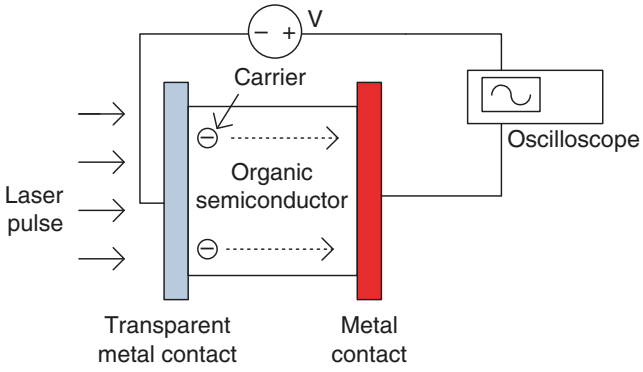


Conduction Mechanisms in Organic Semiconductors, Fig. 2 Density of (energy) states in an organic semiconductor

As expected, in most organic semiconductors both the carrier density n and the charge mobility μ are low and the value of the latter often depends on the strength of the electrical field F in contrast to what is observed in inorganic semiconductors. In some organic semiconductors, the molecular structures can be highly disordered and different conduction mechanisms are found to predominate depending on the associated manufacturing process.

Basic Methodology

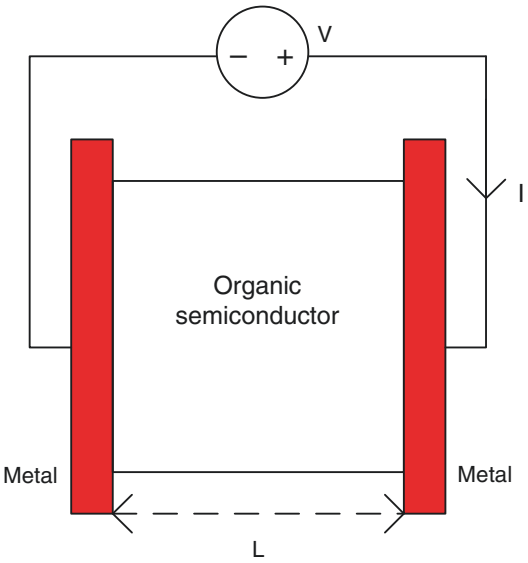
Many useful techniques have been proposed to study the conduction mechanisms in organic semiconductors including time-of-flight (TOF) experiment, space charge limited current (SCLC) measurement, and field-effect measurements using organic field-effect transistors (OFETs). These techniques, combined with the dependence on temperature, provide important information on mobility, trap concentration which are useful to assess the conduction mechanisms. Brief introductions to several different techniques commonly used are given here.



Conduction Mechanisms in Organic Semiconductors, Fig. 3 A schematic of the time-of-flight (TOF) experiment setup

Time-of-Flight (TOF) Experiment

As implied by its name, time-of-flight (TOF) experiment is the method of measuring the time it takes for one or a few carrier to travel a distance through the solid. When TOF experiment is used to study the conduction mechanisms in organic semiconductors (see Fig. 3), two metal electrodes (forming the contact) are deposited on the two ends of the organic semiconductor (one of the two is usually transparent). Initially, a few carriers are generated at one end near the transparent metal electrode using a short laser pulse with energy greater than the energy difference between HOMO and LUMO levels of organic semiconductor. The photo-generated carriers are then drifted toward the opposite end by an external electrical field generating a current pulse. By measuring the time delay of the current pulse, the velocity and the mobility of the charge carriers through the organic semiconductors can be computed.



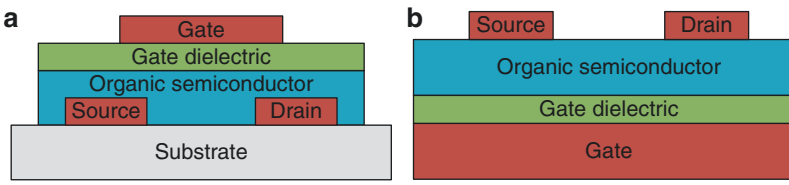
Conduction Mechanisms in Organic Semiconductors, Fig. 4 A simplified schematic of the setup used for space charge limited current (SCLC) measurement

Space Charge Limited Current (SCLC) Measurement

Due to the low mobility of carriers in organic semiconductors, the measured I–V characteristics usually deviate from Ohm's law (i.e., the linear relationship between current and voltage). This is illustrated in Fig. 4. In this case, if efficient charge injection from the metal electrode is achieved by choosing a suitable metal, the I–V characteristics will follow the space charge limited current equation as given by

$$I = \frac{9}{8}\theta\varepsilon_0\varepsilon_r\mu A\frac{V^2}{L^3} \qquad (3)$$

where θ is a parameter dependent on the traps present in the semiconductors, ε_0 the free space permittivity, ε_r the relative dielectric constant, μ drift mobility of injected charge carrier, A the cross section area of semiconductor, and L the distance between metal contacts. By measuring the I–V characteristics at different temperatures, one can determine the mobility and the trap density of traps. Thus, SCLC measurement is a very useful technique reflecting the conduction mechanisms in organic semiconductors.



Conduction Mechanisms in Organic Semiconductors, Fig. 5 Two device configurations for the OFETs: (**a**) top-gate, (**b**) bottom-gate

Measurement Based on the Organic Field-Effect Transistor (OFET)

Field-effect transistor (FET) is an electronic device widely used in active circuits. It consists of a semiconductor with a conducting channel, an isolated gate controlling charge flow in the channel, a gate dielectric between the semiconductor and the gate, as well as source and drain regions forming the output terminals. Organic field-effect transistor (OFET) is a field-effect transistor formed on an organic semiconductor. The device configuration can have a top-gate or a bottom-gate as shown in Fig. 5.

The basic operation principle of the OFETs is very simple. When a bias voltage is applied between gate and source electrodes, carriers are injected from the source into the organic semiconductor forming an extremely thin accumulation layer ($2 \sim 3$ nm) at the interface between organic semiconductor and the dielectric. The carriers conduct a current across the source and the drain regions, and the current depends on the gate voltage as well as the charge mobility which is also dependent on the gate voltage and the drain-to-source voltage. The operation of the OFETs therefore relies on carrier accumulation in the field-effect structure in contrast to the case of the inorganic FETs which rely on either charge depletion or inversion. Therefore, OFETs are efficient tools to investigate the interfacial conduction mechanisms, while TOF and SCLC measurements are mainly used to study the bulk conduction mechanisms in the organic semiconductors. For example, *I–V* characteristics of OFETs are usually analyzed to determine parameters such as the charge mobility and the threshold voltage both of which are closed related to the density of traps at the interface. In addition, spectroscopic techniques are sometimes used to probe the morphology of the organic semiconductor interface, to look for the potential relationship between regularity of molecular structure and conduction performance.

Key Research Findings

Band-Like Transport

For highly purified and ordered organic molecular crystals, it is possible that band-like charge transport similar to that of the inorganic semiconductors may occur. The main feature found in band-like charge transport is the fact that the temperature dependence of the charge mobility has the following form:

$$\mu(T) \propto T^{-n}, \text{ with } n = 1, 2, 3 \ldots \qquad (4)$$

In practice, n is usually positive which leads to increasing charge mobility when temperature decreases. In general, because the electrons are usually weakly delocalized even in the highly ordered organic crystals, the band widths of the HOMO and the LUMO are small compared to energy bands found in the inorganic semiconductors. As a result, room temperature charge mobilities observed in organic semiconductor crystals can only reach values in the range 1–20 cm^2/Vs [2]. Band-like charge transport has been observed in small-molecules and in single-crystal organic semiconductors (such as rubrene) formed by vapor deposition process. As a matter of fact, in the majority of organic semiconductors, traps and defects are formed during deposition which tends to destroy band-like properties.

Polaron Transport

A polaron is a quasiparticle composed of a charge carrier and its induced polarization field. In many organic materials, due to the low charge

mobilities, carriers tend to polarize their surrounding lattice. As a result, polarization fields are formed around the carriers, which can no longer be considered as "naked." Instead, the carriers will be localized in potential minima created by the so-called molecular deformations [1]. In other words, a charge is trapped by the deformation it induces. Such an entity is known as a "polaron." Polarons can move between molecules similar to the carriers except that they also carry the deformations along. In many disordered molecular organic semiconductors, deformations associated with the trapped charges can be considerable and conduction mechanisms characterizing polaron transport are under intensive research by many research groups across the world.

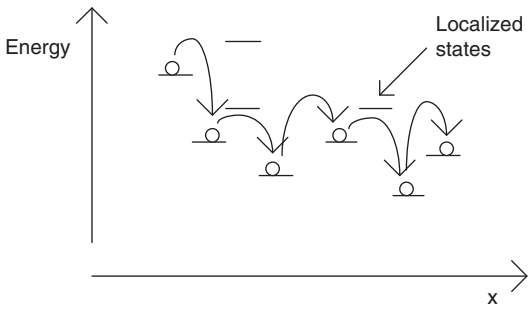
Variable Range Hopping (VRH) Transport

For most organic semiconductors such as polymers and oligomers, their molecular structures are highly disordered and have considerable densities of defects and traps. The energy band diagrams are no longer suitable to describe the densities of states as these energy states are now localized. Furthermore, band-like charge transport can no longer explain the observed low charge mobilities and the fact that their values increase with temperature (as opposed to what is observed in band-like charge transport). One of the widely accepted theories, known as the variable range hopping (VRH) transport, is proved to give a reasonable explanation by describing charge transport in terms of hopping of the charge carriers between localized states as shown in Fig. 6.

Hopping can be used to explain the lower mobility found in disordered organic semiconductors and instead of the power law dependence on temperature as in band-like charge transport, the temperature dependence in VRH charge transport exhibits temperature-dependent activation as well as dependence on the applied electric field as given by [3]

$$\mu(F,T) \propto \exp(-\Delta E/kT) * \exp\left(\beta\sqrt{F}/kT\right) \quad (5)$$

where μ is mobility, F is the electrical field, T is the temperature, ΔE is the activation energy, and β is a parameter related to disorder.



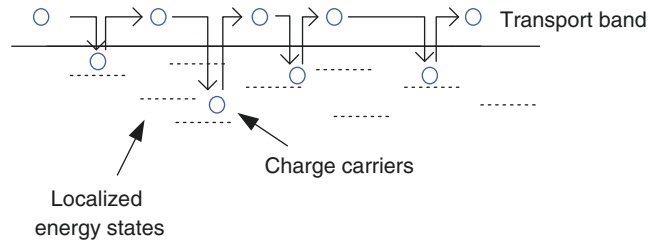
Conduction Mechanisms in Organic Semiconductors, Fig. 6 Hopping transport in organic semiconductors

Multiple Trap and Release (MTR) Transport

Charge transport in OFETs is affected by defects and impurities which exist in the intrinsic part of the organic semiconductors per se and can also be linked to an inferior semiconductor/dielectric interface. As a result, the performance of OFETs is sample-dependent, which is one of the major difficulties in characterizing the properties of OFETs. As mentioned earlier, VRH transport is more suitable to account for charge transport in highly disordered organic semiconductors and, in contrast, another well-established and widely accepted charge transport model known as the "multiple trap and release" (MTR) model is frequently applied to the relatively well-ordered organic semiconductors, such as small molecules and molecular crystals. The basic principle of MTR model includes two important components: (1) a transport band containing delocalized energy states whereby carriers can move freely and (2) the presence of a high density of localized energy states located in the vicinity of the edge of transport band acting as traps. During charge transport, carriers move freely in the transport band with a high probability of being trapped at the localized energy states and then subsequently thermally released into transport band again. The basic illustration of MTR transport process is shown in Fig. 7.

The effective mobility (μ_{eff}) in the MTR model is actually smaller than the "real" mobility (μ_0) in the transport band in the absence of localized energy states and is given by [4]

$$\mu_{\text{eff}} = \mu_0 \alpha \exp(-E_t/kT) \quad (6)$$



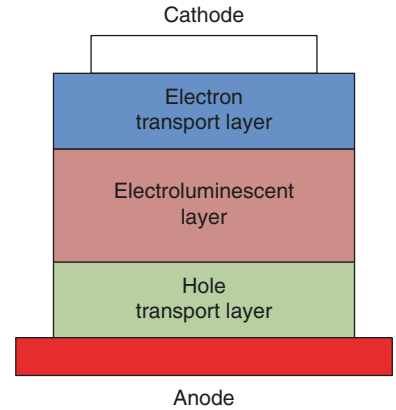
Conduction Mechanisms in Organic Semiconductors, Fig. 7 Multiple trap and release transport (*MTR*) model

where α is the ratio of the effective density of energy states at the edge of the transport band to the density of traps in the localized energy states, and E_t is the energetic distance between the edge of the transport band and the localized energy states. If the localized energy states are energetically dispersive, α and E_t must be recalculated according to the trap distribution.

In the study of OFETs, the MTR model is widely used to account for charge transport due to the fact that it offers a reasonable explanation on the gate voltage dependent mobility usually observed in OFETs. As mentioned above, unlike inorganic semiconductors, organic semiconductors usually have Gaussian density of states (DOS). When a bias is applied to the gate of an OFET, the Fermi level at the dielectric-semiconductor interface will be shifted toward the transport band so that a fair amount of localized energy states near the edge of the transport band will be filled when the Fermi level is located closer to the transport band. As a result, the mobility of the carriers in the MTR model is actually improved because of the reduced density of traps leading to a reduced value of E_t. Therefore, the gate-voltage dependent effective mobility of the carriers in the MTR transport model is given by [5]

$$\mu_{\text{eff}} = \mu_0 \frac{N_c}{N_{t0}} \left[\frac{C_i (V_G - V_T)}{q N_{t0}} \right]^{\frac{T_0}{T} - 1} \tag{7}$$

where μ_0 is the mobility of the carriers in the transport band, N_c the effective density of states at the edge of the transport band, N_{t0} the total density of traps, C_i the capacitance of the insulator per unit area, and T_0 is a characteristic temperature related to the distribution of the DOS.

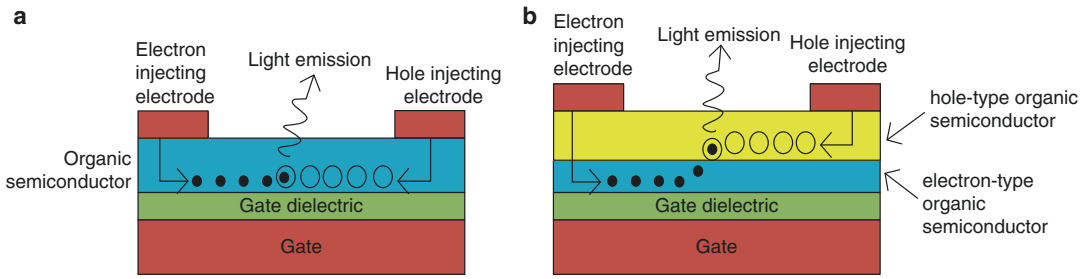


Conduction Mechanisms in Organic Semiconductors, Fig. 8 Schematic of a typical organic light-emitting diode (OLED)

Devices

Organic Light-Emitting Diode (OLED)

Organic light-emitting diode (OLED) is an electroluminescent diode composed of organic materials serving as the electroluminescent layer and charge transport layer. The typical OLED structure is shown in Fig. 8. During operation of an OLED, a bias voltage is applied between anode and cathode. Holes (electrons) are injected from the anode (cathode) into the electroluminescent layer through the hole (electron) transport layer. Because electrons and holes exist simultaneously in the same layer, there is a high probability that they recombine with each other due to electrostatic forces, leading to radiative emission. Therefore, efficient charge injection from both cathode and anode is requisite for the efficient operation of the OLED and current conduction is dominated by space charge limited current as introduced earlier.



Conduction Mechanisms in Organic Semiconductors, Fig. 9 Schematic illustrations of organic light-emitting field-effect transistors (OLEFETs) with light emission in: (**a**) the single-layer configuration and (**b**) the multilayer configuration

Organic Light-Emitting Field-Effect Transistor (OLEFET)

Organic light-emitting field-effect transistor (OLEFET) is a novel organic device combining the function of current conduction of an OFET with electroluminescence in an OLED. The operation of the OLEFET is actually the same as that of the OFET. However, if proper materials are chosen as the source and the drain to give efficient charge injection and under favorable voltage bias condition, electrons and holes can be injected and transported separately in the OFET channel(s). This type of charge transport is known as ambipolar charge transport, which is unique and only found in an organic field-effect transistor. Furthermore, in ambipolar charge transport if the electrons and holes are allowed to recombine radiatively in an emitter layer to give out light, this type of OFET with electroluminescence functionality is usually called OLEFET. Various device structures have been proposed to realize ambipolar charge transport and light emission in OLEFETs, and, in most cases, the proposed structures fall into two main categories as far as the charge layers are concerned. These are the single-layer OLEFET and multilayer OLEFET as shown in Fig. 9.

Cross-References

- ▶ Electrode–Organic Interface Physics
- ▶ Flexible Electronics
- ▶ Optical and Electronic Properties
- ▶ Surface Electronic Structure

References

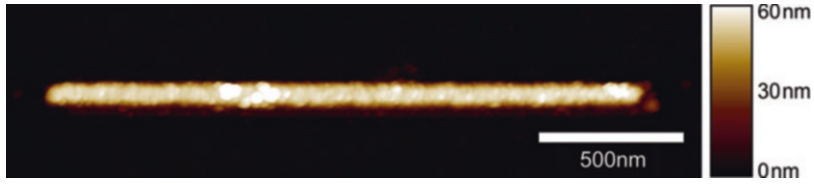
1. Kwok, H.L., Wu, Y.L., Sun, T.P.: Charge transport and optical effects in disordered organic semiconductors. In: Noginov, M.A., Dewar, G., McCall, M.W., Zheludev, N.I. (eds.) Tutorials in Complex Photonic Media, pp. 576–577. SPIE Press, Bellingham (2009)
2. Podzorov, V., Menard, E., Borissov, A., Kiryukhin, V., Rogers, J.A., Gershenson, M.E.: Intrinsic charge transport on the surface of organic semiconductors. Phys. Rev. Lett. **93**, 086602 (2004)
3. Brütting, W.: Physics of Organic Semiconductors. Wiley, Weinheim (2005)
4. Horowitz, G.: Organic field-effect transistors. Adv. Mater. **10**, 365–377 (1998)
5. Bao, Z., Locklin, J.: Organic Field-Effect Transistors. CRC Press, Boca Raton (2007)

Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy

J. Vogt[1], C. Huck[1], F. Neubrech[2] and A. Pucci[1]
[1]Kirchhoff Institute for Physics, Heidelberg University, Heidelberg, Germany
[2]4th Physics Institute, University of Stuttgart, Stuttgart, Germany

Synonyms

Electrical conductivity; Localized plasmon polaritons in the infrared; Nanoantennas



Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy, Fig. 1 AFM picture of a typical gold nanowire prepared by EBL on CaF_2. The measured average height of the wire is ca. 60 nm

Definition

Metal nanorods of a few micron length and a much smaller diameter feature strong plasmonic resonances in the infrared region and therefore act rather similar to radio antennas, but the spectral shape of such resonances is related to the conductivity of the nanoantenna material. In the scientific literature, analytic approaches exist which explain the relationship between the electronic conductivity and the resonance spectrum that is due to plasmon polaritons as mixed excitations from free electrons and photons.

Introduction

Metal nanowires are produced, for example, by electron-beam lithography (EBL), electrochemical, and wet chemical methods [1]. Usually they are inspected by scanning electron microscopy (SEM) or atomic force microscopy (AFM) in order to get geometric information; see Fig. 1. But, neither SEM nor AFM can deliver conductivity information of a nanoobject. Electrical measurements are not simple because of the necessary contacts and thus contact resistance comes into play. Nevertheless, because of the importance of the ohmic losses in electronics, many studies already investigated the change of resistivity with the size and the crystalline quality of interconnect structures [2]. The results demonstrate that the two main contributions to the resistivity increase with shrinking height and width of metal interconnects which are surface scattering and grain-boundary scattering. The Matthiessen rule according to which the several contributions add up to the total resistivity [3] was proven to be a valid approximation in most cases.

For plasmonic resonances of nanostructures, increased resistivity leads to higher damping and thus to less-efficient performance of antennas, sensors, and other devices. Vice versa, information on resistivity or conductivity, respectively, is included in the plasmonic resonance spectrum.

Infrared Conductivity of Metals

Below the onset of interband transitions, the infrared (IR) optical properties of metals are determined by the collective oscillations of free-charge carriers called plasmons. The circular frequency (ω)-dependent Drude dielectric function

$$\varepsilon(\omega) = \varepsilon_\infty - \frac{\omega_p^2}{\omega(\omega + i\omega_\tau)} \qquad (1)$$

is a good description of that behavior [4]. Effects from interband transitions on the background polarizability are included in ε_∞. This background permittivity can have values clearly above 1, which becomes important in the near infrared where the negative real part of the Drude term reaches the same order of magnitude. For gold and silver, ε_∞ becomes frequency dependent and complex in the visible, but for other metals like iron, chromium, and platinum, such behavior starts already in the mid-infrared, and the application of the models in this article is then restricted to the far IR below about 1000 cm^{-1} (in wave numbers). The Drude parameters ω_p and ω_τ describe the plasma frequency and the relaxation rate of the free charge carriers, respectively. Frequency-independent parameters (see Table 1) are in reasonable accord with the IR optical

Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy, Table 1 Drude parameters for selected metals at room temperature from the data collection by Ordal et al. [5]. The parameters are given as wave numbers $\varpi = \omega/(2\pi c)$ in cm^{-1} units. They were derived from fits to far and mid-IR spectra of polycrystalline bulk material (ε_∞ was set equal to one)

	ω_τ	ω_p
Gold	215	72,800
Silver	145	72,700
Aluminum	660	119,000

behavior of such metals where only s and p electrons contribute to the conductivity, for example, aluminum, silver, and gold (but not iron). With ω_p much higher than IR frequencies and $\omega_\tau << \omega_p$, the Drude dielectric function has a strong negative real part in the IR. For ω_τ spectrally located in the mid-IR (typical situation occurring for iron or defect-rich noble metals), the skin depth in absorbing region is

$$\delta \approx \frac{c}{\omega_p}\sqrt{\frac{2\omega_\tau}{\omega}}, \qquad (2)$$

but it is $\delta \approx c/\omega_p$ for $\omega_\tau << \omega$ in the reflecting region (c is the vacuum velocity of light). Typical values for skin depths of metals in the mid-IR are of the order of a few 10 nm and thus of the same order of magnitude as the diameters of typical plasmonic structures used in sensing applications. Any increase in resistivity may lead to a strong increase in the skin depth and thus significantly increases the ratio between absorption and reflection [4].

In the dc limit, the Drude parameters determine the dc resistivity $\rho_{dc} = \omega_\tau / \left(\varepsilon_v \omega_p^2\right)$ [3]; ε_v is the vacuum permittivity. Slight deviations from the IR optical parameters may occur due to an anisotropy of scattering [4] and the energy dependence of many body effects [6].

Electron Scattering

From a variety of studies, it is known that, as theoretically predicted, the contribution $\omega_{\tau s}$ of electronic surface scattering to resistivity increases directly proportional to the inverse wire diameter (with a proportionality factor depending on surface roughness and surface chemistry), e.g., [2] and references therein, and, according to Matthiessen's rule, [3] has to be added to the defect scattering rate $\omega_{\tau d}$. Therefore, the total electronic relaxation rate (inverse lifetime) is

$$\omega_\tau = \omega_{\tau e}(t,\omega) + \omega_{\tau p}(t) + \omega_{\tau s} + \omega_{\tau d} \qquad (3)$$

with $\omega_{\tau e}$ and $\omega_{\tau p}$ as the electron–electron and the electron–phonon scattering rate, respectively [2–4].

Above temperatures t of about 10 K, the temperature-dependent term in electron–electron scattering is small compared to electron–phonon scattering [3, 4]. The frequency-dependent term in electron–electron scattering is also negligible for metals with conduction carried by s and p electrons only [5, 6], like, for example, Au in the infrared where the dielectric function follows the simple Drude model with a constant ω_τ. The temperature dependence of ω_τ is almost due to the temperature-dependent number of phonons. With the Debye temperature θ_D, the Debye model for phonons leads to the Bloch–Grüneisen relation [7] for the scattering rate

$$\omega_{\tau p} \propto \left(\frac{t}{\theta_D}\right)^5 \int_0^{\theta_D/t} \frac{z^5}{(e^z - 1)(1 - e^{-z})} dz \qquad (4)$$

for a free electron gas in the dc limit. This relation predicts $\omega_\tau \propto t/\theta_D$ for $t >> \theta_D$, $\omega_\tau \approx \omega_{\tau d} + \omega_{\tau s}$ for $t << \theta_D$, and a transition region with $\omega_\tau \propto (t/\theta_D)^5$ (Bloch relation [3]), respectively. In the infrared, this relation is no more valid since additionally to the electron and phonon, also a photon with energy $\hbar\omega$ is involved, and the scattering rate

$$\omega_{\tau p} = \frac{1}{\tau_0}\left[\frac{2}{5} + 4\left(\frac{t}{\theta_D}\right)^5 \int_0^{\theta_D/t} \frac{z^4}{e^z - 1} dz\right] \qquad (5)$$

derived by Holstein is considered as a better approximation [8]. The rate $1/\tau_0$ depends on the kind of metal and could be approximated by

Holstein's bulk dc-scattering rate at sufficiently low temperature [4].

Granular Metal

Beyond the percolation threshold, the IR optical properties of conductive but granular metal films can be described by a Drude-type dielectric function ε^* with an effective plasma frequency $\omega_{p,eff}$. The analytic approach to the effective dielectric function ε_{eff} of inhomogeneous media that originally was developed by Bruggeman [9],

$$F\frac{\varepsilon-\varepsilon_{\text{eff}}}{\varepsilon+(D-1)\varepsilon_{\text{eff}}}+(1-F)\frac{\varepsilon_{\text{host}}-\varepsilon_{\text{eff}}}{\varepsilon_{\text{host}}+(D-1)\varepsilon_{\text{eff}}}=0, \tag{6}$$

with the spatial filling factor F, dimension D, and a host medium with ε_{host} can be used to estimate the effective Drude parameters. For a metal-island film $D = 2$ and in vacuum, without adsorbates or cover layers, $\varepsilon_{host} = 1$. With $1/\varepsilon \approx 0$ which is a valid approximation below a certain frequency, it follows $\varepsilon_{eff} = (2F - 1)\varepsilon$. Furthermore, with $\varepsilon_{eff}/F = \varepsilon^*$, a relation that accounts for equal absorption in both the models, the result is $\omega^2_{p,eff} = (2 - 1/F)\omega^2_p$. For $D = 3$ the same approximation leads to $\omega^2_{p,eff} = \omega^2_p(3F - 1)/(2F)$. For $D = 2$, the lowering of the squared-effective plasma frequency is more pronounced than for the $D = 3$ case. So, far-field spectra are especially sensitive to a morphology consisting of one layer of grains, which is mostly relevant for metal nanostructures [10]. The grainy structure of the wire in Fig. 1 could be clearly recognized. Upon annealing the morphology can be changed towards larger grains.

Infrared Plasmonic Resonances

Very similar to radio frequency (RF) antennas, resonantly excited plasmonic nanostructures act as optical antennas that concentrate energy of electromagnetic radiation to a confined volume of sub-wavelength scale. For example, metallic nanorods ("nanoantennas") with μm-sized lengths L show plasmonic resonances in the IR spectral range. Related to the not-negligible skin depth δ, the simple $\lambda/2$-dipole behavior known from RF antennas, where the relationship between L and the resonant wavelength λ_{res} is given by $2\ L = \lambda_{res}$, is not valid for nanoantennas at optical (including IR) frequencies [1]. The modified relation for cylindrical rods

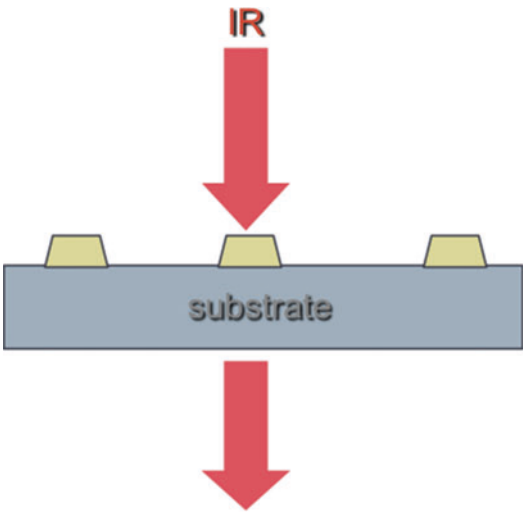
$$2L = \lambda_{\text{eff}} = n_2\left(\lambda_{\text{res}}/\lambda_{\text{p}}\right) + n_1 \tag{7}$$

(for the radius $R << L$) where λ_p denotes the plasma wavelength of the antenna's material better describes the optical antenna behavior in the case $\omega_\tau << \omega_{res}$ and makes clear that $\lambda_{res} \sim 1/\omega_p$ [11]. Thus granularity effects the resonance position; see the lowered effective plasma frequency above. The coefficients n_1 and n_2 depend on R and on the dielectric properties ε_∞ and ε_s, the dielectric constant of the nonabsorbing surrounding medium at the resonance frequency:

$$n_1 \approx R\left\{\left[2\pi\left(13.74 - 0.12\cdot\frac{\varepsilon_\infty + 141.04\varepsilon_s}{\varepsilon_s}\right)\right] - 4\right\},$$
$$n_2 \approx \frac{2\pi R}{\varepsilon_s}\cdot 0.12\cdot\sqrt{\varepsilon_\infty + 141.04\varepsilon_s}. \tag{8}$$

The metal was described as a free-electron gas according to the Drude model with negligible relaxation rate ω_τ compared to photon frequencies ω. Such conditions are fulfilled for perfect, crystalline gold (clearly also silver and copper) nanorods in the mid-IR, but not in case of strong defect scattering when δ becomes frequency dependent.

Plasmonic extinction in the IR (e.g., as measured in normal transmittance geometry, see Fig. 2, with polarization along the long-nanowire axis and the nanowire on a substrate that without nanowire is used for the reference spectrum [1]) contains an important contribution from absorption. This contribution becomes smaller (compared to scattering) for larger R (or larger height h and width w) of the antenna. As a rough approximation, these effects can be described within the radiation-corrected



Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy, Fig. 2 Scheme of the IR transmittance measurement at normal incidence of light. For studies of individual wires, the focal plane should be as small as possible but should include the full wire. The reference measurement has to be performed on the identical bare substrate with the same IR optics. The polarization of the IR light is important

quasistatic approximation that allows first-order estimates of resonances also for particles substantially larger than 1 % of wavelength if the incident light is uniform across the object, i.e., for IR antennas at normal incidence of light up to few 10 nm in diameter [12, 13]. In this approximation, for the fundamental resonance and $\omega_\tau << \omega$, the scattering cross section

$$\sigma_{\mathrm{sca}} = \omega_{\mathrm{p}}^2 \frac{Vr}{n_{\mathrm{s}} c} \times \frac{\omega^4 T_{\mathrm{L}}}{\left[\left(\omega_{\mathrm{res}}^2 - \omega^2\right)^2 + \omega^2 \left(\omega_\tau + \omega^2 T_{\mathrm{L}}\right)^2\right]} \tag{9}$$

and the absorption cross section

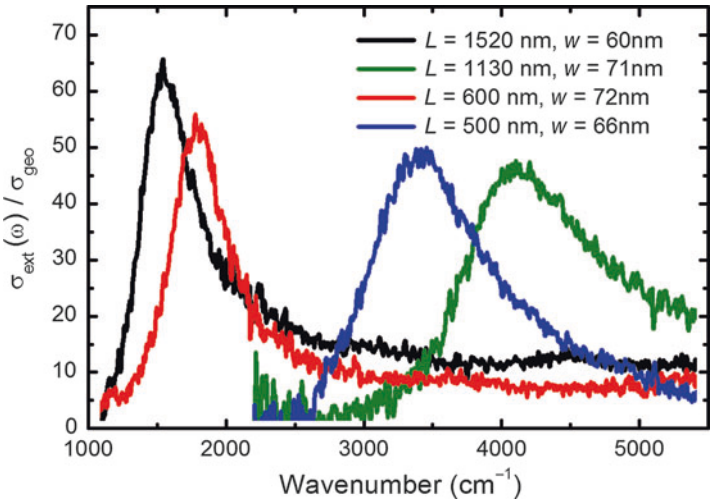
$$\sigma_{\mathrm{abs}} = \omega_{\mathrm{p}}^2 \frac{Vr}{n_{\mathrm{s}} c} \times \frac{\omega^2 \omega_\tau}{\left[\left(\omega_{\mathrm{res}}^2 - \omega^2\right)^2 + \omega^2 \left(\omega_\tau + \omega^2 T_{\mathrm{L}}\right)^2\right]} \tag{10}$$

differ in their asymmetric frequency-dependent spectral shape; notice the different exponents of ω in the two numerators. $T_{\mathrm{L}} = Vrn_{\mathrm{s}}\,\omega_{\mathrm{p}}{}^2/(6\pi c^3)$ is the Larmor-time parameter (with the vacuum velocity c of light) depending on the volume V (where the electrons are excited in the particle), the refractive index n_{s} of the surrounding medium, and the local-field ratio r [14]. For a spheroid in the quasistatic approximation, $r = \varepsilon_{\mathrm{s}}/[\varepsilon_{\mathrm{s}}+F(\varepsilon_\infty-\varepsilon_{\mathrm{s}})]$ with F as the depolarization factor [15]. For a sphere, for example, $r = 3\varepsilon_{\mathrm{s}}/(\varepsilon_\infty + 2\varepsilon_{\mathrm{s}})$ for an infinitely long needle $r = 1$ which is a reasonable approximation also for nanowires. For a plasmonic dipole of an infinitesimally small volume, T_{L} becomes also infinitesimally small and the damping of the oscillator gets equal to the electronic scattering rate ω_τ. However, related to the scattering of light in case of larger objects with finite V, the radiation damping rate $\omega_{\mathrm{rad}} = \omega^2 T_{\mathrm{L}}$ has to be included in the denominators of both the quasistatic expressions σ_{abs} and σ_{sca}. Typical values of ω_{rad} at $\omega = \omega_{\mathrm{res}}$ for IR nanoantennas from gold (with h and w similar to 50 nm, $n_{\mathrm{s}} \geq 1$) are of the order of the relaxation rate and thus are not negligible.

In experiments, resonances may appear much broader compared to the expectation based on the damping rates if the symmetry of the cylindrical shape is broken due to production errors. Then, for example, the even-order resonances (at 2 ω_{res}, usually not dipole active [1]) can be excited also by normal incidence of light. Further (Gaussian) broadening appears due to random parameter variations in a nanowire ensemble. In ensembles with interwire distances of the order of the resonance wavelength and below, nanowire interaction is an important issue in [16] not presented here.

Infrared Spectroscopy of Metallic Nanowires

With IR microspectroscopy (intense light sources are preferred), the far-field extinction related to a nanowire's plasmonic resonances can be detected well if a small aperture with the wire in its center is used [1, 16]. The IR spectra benefit from a carefully measured reference spectrum from the bare (identical) substrate. The extinction cross section $\sigma_{\mathrm{ex}} = \sigma_{\mathrm{abs}} + \sigma_{\mathrm{sca}}$ of a single nanoparticle



Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy, Fig. 3 Individual nanowire's extinction (normalized to the geometrical shadow) measured at room temperature in normal transmittance for an electric field parallel to the long axis of the wire. The transmittance is normalized to the IR spectrum of the ZnS substrate without any wire. The geometrical parameters (length L and width w) of the electron-beam lithographically produced gold nanostripes are given in the figure. The stripes' height was about 60 nm

(see examples in Fig. 3) on a thick transparent substrate (with refractive index $n_{\text{substrate}}$) can be estimated from the relative transmittance spectrum $T_{\text{rel}}(\omega)$ at normal incidence of light via the relation $\sigma_{\text{ext}} = A_0 \cdot (1 - T_{\text{rel}}) \cdot (n_{\text{substrate}} + 1)/2$ where A_0 is the focal area (with one particle in the center) [1, 16]. If the preconditions mentioned above are fulfilled, a fit based on the Eqs. 8 and 9 can deliver the parameters ω_τ and ω_{res} (and thus λ_{res} that via Eqs. 7 and 8 can be compared to the value expected from the ideal plasma frequency).

Conclusion

Bringing together well-established knowledge on metal optical properties, metallic resistivity, and optical antennas, the strong relationship between electrical conductivity parameters and the fundamental plasmonic resonance of nanowires in the infrared is clarified.

Cross-References

- ▶ AFM
- ▶ Gold Nanorods
- ▶ Effective Media
- ▶ Electron Beam Lithography (EBL)
- ▶ Local Surface Plasmon Resonance (LSPR)
- ▶ Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods
- ▶ Plasmonics
- ▶ Spectromicroscopy

References

1. Pucci, A., Neubrech, F., Aizpurua, J., Cornelius, T., de la Chapelle, M.L.: Electromagnetic nanowire resonances for field-enhanced spectroscopy. In: Wang, Z. (ed.) One-Dimensional Nanostructures, pp. 175–216. Springer, New York (2008)
2. Steinhögl, W., Schindler, G., Steinlesberger, G., Engelhardt, M.: Size-dependent resistivity of metallic wires in the mesoscopic range. Phys. Rev. B **66**, 075414 (2002)
3. Ashcroft, N.W., Mermin, N.D.: Solid State Physics. Sounders College Publishing, Orlando (1976)
4. Abeles, F.: Optical properties of metals. In: Abeles, F. (ed.) Optical Properties of Solids, pp. 93–162. North Holland, Amsterdam (1972)
5. Ordal, M.A., Bell, R.J., Alexander Jr., R.W., Long, L. L., Querry, M.R.: Optical properties of fourteen metals in the infrared and far infrared: Al, Co, Cu, Au, Fe, Pb, Mo, Ni, Pd, Pt, Ag, Ti, V, and W. Appl. Optics **24**, 4493–4499 (1985)

6. Young, C.-Y.: The frequency and temperature dependence of the optical effective mass of conduction electrons in simple metals. J. Phys. Chem. Solids **30**, 2765–2769 (1969)
7. Ziman, J.M.: Principles of the Theory of Solids. Cambridge University Press, Cambridge (1979)
8. McKay, J.A., Rayne, J.A.: Temperature dependence of the infrared absorptivity of the noble metals. Phys. Rev. B **13**, 673–685 (1976)
9. Bittar, A.: The Bruggeman Effective Medium Theory Applied to the Optical Properties of Inhomogeneous Materials. Physics and Engineering Laboratory, Lower Hutt (1984)
10. Zhang, X., Stroud, D.: Optical and electrical properties of thin films. Phys. Rev. B **52**, 2131–2137 (1995)
11. Novotny, L.: Effective wavelength scaling for optical antennas. Phys. Rev. Lett. **98**, 266802 (2007)
12. Sarid, D., Challener, W.A.: Modern Introduction to Surface Plasmons. Cambridge University Press, Cambridge (2010)
13. Kats, M.A., Yu, N., Genevet, P., Gaburro, Z., Capasso, F.: Effect of radiation damping on the spectral response of plasmonic components. Opt. Express **19**, 21748–21753 (2011)
14. Doyle, W.T.: Electrodynamic response of metal spheres. J. Opt. Soc. Am. A **2**, 1031–1034 (1985)
15. Pelton, M., Bryant, G.: Introduction to Metal-Nanoparticle Plasmonics. Wiley, Hoboken (2013)
16. Weber, D., Pucci, A.: Antenna interaction in the infrared. In: de la Chapelle, M.L., Pucci, A. (eds.) Nanoantenna: Plasmon -Enhanced Spectroscopies for Biotechnological Applications, pp. 175–194. Pan Stanford Publishing, Singapore (2013)

Confocal Laser Scanning Microscopy

Reinhold Wannemacher
Madrid Institute for Advanced Studies, IMDEA Nanociencia, Madrid, Spain

Synonyms

Confocal scanning optical microscopy (CSOM); Laser scanning confocal microscopy

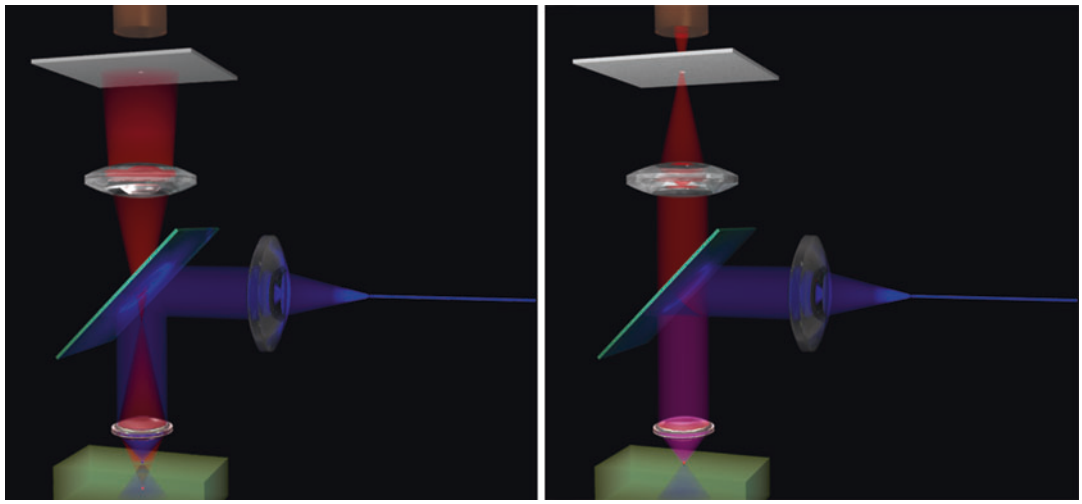
Definition

A confocal laser scanning microscope (CLSM) images a point light source used for excitation onto the sample via the objective lens and images the excited focal volume onto a point detector using reflected, transmitted, emitted, or scattered light. In contrast to conventional microscopes, this scheme permits strong rejection of out-of-focus light and optical sectioning of the sample. In order to obtain an image, the focal volume must be scanned relative to the sample. Scanning can be performed in the lateral as well as in the axial directions, and three-dimensional images of the sample can be generated in this way.

Operating Principle

Figure 1 illustrates the working principle of a fluorescence confocal microscope. A point light source, here the end of an optical fiber carrying the excitation light, is imaged onto the sample by the objective lens via a beam splitter. Emitted light from the focal spot is imaged through the beam splitter onto a pinhole via an auxiliary lens. The light is registered by the detector, because it passes through the pinhole (left-hand side of Fig. 1). Fluorescence from a fluorophore at an out-of-focus position in the excitation cone, on the other hand, arrives at the screen defocused. Therefore, only a small fraction of it passes through the pinhole and reaches the detector. The effect is understated by the schematic figure and actually much stronger in a real microscope because of the short focal length of the objective lens. The optical sectioning capability is the core of confocal microscopy and allows to render the object three-dimensionally under different angles by appropriate software once a stack of images at different depths has been acquired. In addition, the lateral resolution is slightly improved in confocal microscopy, compared to conventional microscopy, when the pinhole is small. On the other hand, an image can be acquired in this way only by serial scanning of the sample or of the excitation beam (for technical improvements in this respect see section "Confocal Microscopy Involving Modified Illumination").

Because a CLSM operates with light, it may be used to image many different physical quantities. These may be simply reflected or transmitted intensity (brightfield confocal microscopy) or the



Confocal Laser Scanning Microscopy, Fig. 1 Principle of confocal laser scanning microscopy, demonstrating the strong rejection of out-of-focus light

intensity of fluorescence excited in the sample (fluorescence confocal microscopy). Other options include polarization and phase of reflected or transmitted light, as well as the intensity, wavelength, lifetime, time correlation, or recovery after photobleaching of fluorescence from the sample or intensity, wavelength, and polarization of inelastically (Raman) scattered light. Moreover, a nonlinear response of the sample to the optical excitation near the laser focus, based, for example, on multiphoton excitation, second harmonic generation, or stimulated scattering processes may be used for confocal microscopy. Some of these options will be discussed in section "Variants of Confocal Laser Scanning Microscopy."

Optical sectioning and three-dimensional image acquisition being the essential feature of confocal optical microscopy, it is worth mentioning here that an alternative (diffraction-limited) brightfield optical microscopy technique with similar capability is digital holographic microscopy, although this technique is far less widely known and used. Here, the object is reconstructed from an intensity camera image and no scanning is necessary. Moreover, a phase image is obtained in addition to an amplitude image. Significant improvements in object reconstruction algorithms have been made in recent years, and lens-based versions with external reference beam as well as lensless versions have been demonstrated. Lens-based instruments with external reference beam are commercially available.

Basic Theory of the Confocal Microscope

The spatial resolution of modern high-quality conventional as well as confocal microscopes is limited by diffraction. This means that within the design spectral range of the objective lens aberrations, such as spherical aberration, astigmatism, coma, field curvature, distortion, and chromatic aberration, have a significantly smaller impact on the resolution of the microscope than diffraction. An important exception to this statement arises from aberrations due to refraction, when sample regions well inside refracting samples have to be imaged.

Most modern microscope objectives are now infinity corrected, that means they are corrected for forming an image at infinity. A tube lens is in this case required to form a real image at finite distance. The limitations by diffraction are, however, in any case dominated by the objective lens and not by the tube lens. This is due to the dependence of the diffraction limit on the opening angle of the rays contributing to the image, which is much smaller for the tube lens than for the objective lens.

Point Spread Function of the Confocal Microscope

The three-dimensional intensity distribution in the image space corresponding to a single-point object, demagnified by the magnification of the optical system, is called the (intensity) point spread function (PSF) of the lens. The PSF of a confocal microscope with an infinitesimally small pinhole is given by [1, 2]:

$$\mathrm{PSF}_{\mathrm{CF}}(x,y,z) = \mathrm{PSF}_{\mathrm{ill}}(x,y,z) \cdot \mathrm{PSF}_{\mathrm{det}}(x,y,z) \tag{1}$$

Here, $\mathrm{PSF}_{\mathrm{CF}}$, $\mathrm{PSF}_{\mathrm{ill}}$, and $\mathrm{PSF}_{\mathrm{det}}$ represent the point spread functions for the confocal imaging and the illumination and detection paths, respectively. Neglecting the contribution from the tube lens, as well as aberrations of the objective lens, the latter two functions are simply the point spread functions of a simple lens, which, in the paraxial and scalar approximation, can be calculated by means of the Huygens-Fresnel principle as [3]

$$\mathrm{PSF}(x,y,z) = |h(x,y,z)|^2 \tag{2}$$

with

$$h\left(\vec{r}\right) = \frac{\mathrm{C}'}{\lambda}\int\int_{A}\frac{\mathrm{e}^{iks}}{s}\mathrm{dA} \approx \frac{\mathrm{C}}{\lambda}\int\int_{\Omega} e^{-ik\vec{q}\cdot\vec{r}}\mathrm{d}\Omega. \tag{3}$$

Here, $h\left(\vec{r}\right)$ represents the scalar complex amplitude of the field in the image space at a position $\vec{r} = (x,y,z)$ relative to the location of the geometric focus; λ is the wavelength, $k = 2\pi/\lambda$: s is the distance between the point P at position $\vec{r}$ in the image space and a point Q at position $f\,\vec{q}$ in the pupil A of the lens of focal length f: $\vec{q}$ is a unit vector in the direction of Q: Ω is the solid angle subtended by the aperture of the lens as seen from the origin at the geometric focus, and C and C′ are constants. Equation 3 assumes homogeneous illumination of the lens. Inhomogeneous illumination can be taken into account by multiplying the integrand with a corresponding pupil function (compare section "Apodization"). For a confocal microscope operating in reflection or transmission modes $\mathrm{PSF}_{\mathrm{ill}}(\mathrm{x,\ y,\ z}) \approx \mathrm{PSF}_{\mathrm{det}}(\mathrm{x,\ y,\ z})$, and both functions are then identical to the one given in Eq. 2. This results in

$$\mathrm{PSF}_{\mathrm{CF}}(x,y,z) = |h(x,y,z)|^4. \tag{4}$$

Experimentally, this function would be observed when a point object is scanned through the focus of the instrument in both cases. It should be kept in mind here that the paraxial approximation is contrary to the actual typical situation in optical microscopy. The paraxial approximation, nevertheless, works surprisingly well even for N.A. ≈ 0.5 ($\theta_0 = 30°$ in the case of a dry lens, see below) and gives reasonable estimates of the diffraction limit even for higher numerical aperture objective lenses. Deficiencies of the scalar approximation will be discussed in section "Deficiencies of the Scalar and Paraxial Approximation: Effects of the Vector Character of Light."

Figure 2 displays the three-dimensional intensity point spread function of a conventional (a) and a confocal (b) microscope calculated according to Eqs. 1 and 3. The numerical aperture of the objective lens is N.A. = 0.5. z is the coordinate along the axis of the lens and r the coordinate perpendicular to the optical axis, measured in wavelengths. The drastic reduction in side lobes for a confocal microscope is immediately evident. This also leads to drastic reduction in laser speckle in the case of coherent illumination.

Single-Point Resolution in the Focal Plane

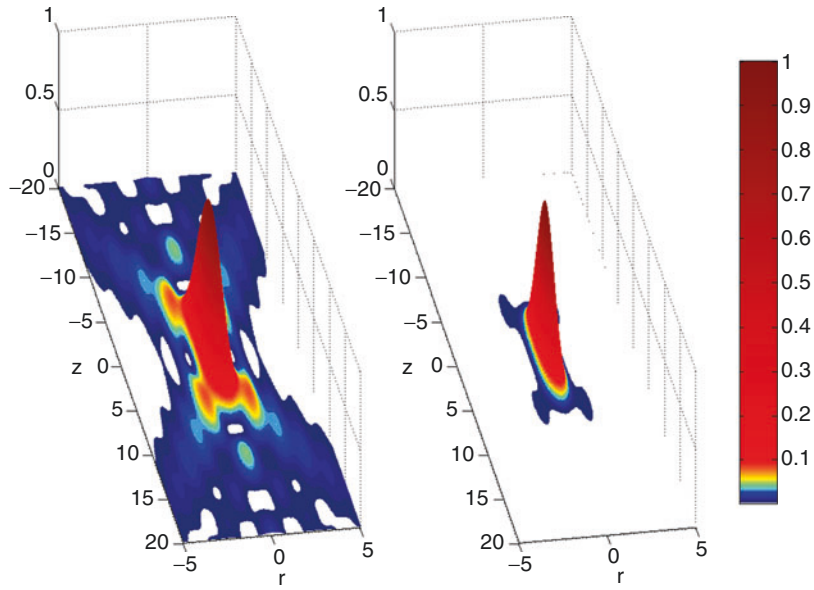
In the focal plane ($z = 0$), the PSF of the conventional microscope, as calculated from Eq. 3, coincides with the well-known Airy pattern

$$\mathrm{PSF} = \left(\frac{2J_1(\mathrm{v})}{\mathrm{v}}\right)^2 \tag{5}$$

with

$$\mathrm{v} = kr \cdot n\sin\theta_0 = kr \cdot \mathrm{N.A.} \tag{5a}$$

Here, r is the distance from the axis; θ_0 is the angle of a marginal ray passing through the aperture



Confocal Laser Scanning Microscopy,
Fig. 2 Intensity point spread function of a conventional (**a**) and a confocal (**b**) microscope in the scalar and paraxial approximations. The numerical aperture of the objective lens is N.A. = 0.5. z is the coordinate along the axis of the lens and r is the lateral dimension, both measured in wavelengths

toward the geometric focus, relative to the optical axis; and n is the refractive index in the image space. From this expression, the lateral full width at half maximum (FWHM) of the PSF of the conventional microscope

$$\mathrm{FWHM} = \frac{0.51\lambda}{\mathrm{N.A.}} \tag{5b}$$

is easily calculated. Equation 3 yields

$$\mathrm{PSF_{CF}} = \left(\frac{2\mathrm{J}_1(\mathrm{v})}{\mathrm{v}}\right)^4 \tag{5c}$$

for the PSF of the confocal microscope in reflection mode, and therefore

$$\mathrm{FWHM_{CF}} = \frac{0.37\lambda}{\mathrm{N.A.}}. \tag{5d}$$

Equations 5b and 5d demonstrate the enhancement in lateral (single-point) resolution for a confocal microscope relative to a conventional one, in the case when the pinhole is closed completely. As an example, for $\lambda = 488$ nm, N.A.=0.9, FWHM = 277 nm, and $\mathrm{FWHM_{CF}} = 201$ nm.

Single-Point Resolution on the Axis: Depth Response

Similarly, the PSF of the conventional microscope on the optical axis is calculated from Eq. 3 as

$$\mathrm{PSF}(u) = \left(\frac{\sin(u/4)}{u/4}\right)^2 \tag{6}$$

with

$$u = nkz\sin^2\theta_0. \tag{6a}$$

It turns out that the definition

$$u = 4kz\sin^2(\theta_0/2) = 2nkz(1 - \cos\theta_0) \tag{6b}$$

which is equivalent to Eq. 6a in the paraxial approximation is more appropriate at higher numerical apertures. The corresponding FWHM of the depth response is therefore

$$\mathrm{FWHM} = \frac{0.89\lambda}{n(1 - \cos\theta_0)} \tag{6c}$$

for the conventional microscope. Correspondingly, for the confocal microscope

$$\mathrm{PSF}_{\mathrm{CF}}(u) = \left(\frac{\sin(u/4)}{u/4}\right)^4 \tag{6d}$$

and

$$\mathrm{FWHM}_{\mathrm{CF}} = \frac{0.64\lambda}{n(1-\cos\theta_0)}. \tag{6e}$$

For the parameters used in the example above, $\lambda = 488$ nm, N.A. $= 0.9$, FWHM $= 770$ nm, and $\mathrm{FWHM}_{\mathrm{CF}} = 554$ nm. The FWHM of the point spread functions on the optical axis of the microscope is therefore about three times larger than the lateral FWHM for the conventional as well the confocal microscope, although the precise value depends on the numerical aperture and although the scalar theory used here is not applicable at large numerical aperture (see section "Deficiencies of the Scalar and Paraxial Approximation: Effects of the Vector Character of Light").

V(z)

A clearer demonstration of the different optical sectioning capabilities of the conventional and confocal microscopes is obtained with planar objects instead of point objects. The depth response of a confocal microscope operating in reflection is often characterized by axially scanning a mirror through the focus position and registering the light intensity behind the pinhole during the scan. The corresponding amplitude function is called $V(z)$, an expression coined originally for the acoustic microscope, which is also a confocal instrument and to which the same scalar theory is applicable. Because the image of the illuminating infinitesimal pinhole is moving by a distance of $2z$ when the mirror moves by a distance z, the intensity detected behind the pinhole is derived from Eq. 6b, by replacing u by $2u$:

$$\begin{aligned} I(z) &= |V(z)|^2 = \left|\frac{\sin(u/2)}{u/2}\right|^2 \\ &= \left|\frac{\sin(nkz(1-\cos\theta_0))}{nkz(1-\cos\theta_0)}\right|^2 \end{aligned}. \tag{7}$$

This equation predicts a central maximum of width

$$\mathrm{FWHM}_{\mathrm{CF}} = \frac{0.44\lambda}{n(1-\cos\theta_0)} \tag{8}$$

which, in paraxial approximation $\theta_0 << 1$ for a dry lens, becomes

$$\mathrm{FWHM}_{\mathrm{CF}} = \frac{0.89\lambda}{\mathrm{N.A.}^2}. \tag{9}$$

Equation 7 also implies symmetric side lobes for negative and positive defocus. As an example, for a dry lens of N.A. $= 0.8$ and an operating wavelength of 488 nm, Eq. 8 yields FWHM $= 537$ nm. In a conventional microscope, on the other hand, the signal received by a large area detector would be independent of the position of the mirror.

The simple theory presented so far predicts a symmetric $V(z)$ function. Whereas the width of the main peak of the $V(z)$ is typically very close to measurements performed with real lenses, the aberrations present in any real objective lens typically lead to deviations as far as the side maxima are concerned and in particular to asymmetry in $V(z)$ for positive and negative defocus. This may be used for quantitative characterization of, for example, the amount of spherical aberration present in the optical system. Interferometric versions of confocal microscopy, however, have been more traditionally used for this purpose.

Two-Point Resolution: Rayleigh and Sparrow Criteria

The single-point resolution of the confocal microscope, as given by the PSF discussed above, is in most cases not the relevant quantity to judge the resolution of the instrument, because what is really desired is the capability to resolve certain details of a microscopic object consisting of various parts. It is therefore important to quantify the two-point resolution of the instrument, which means the capability to resolve two-point objects close to each other. There is some arbitrariness in this definition, because it depends on the subjective judgment, under which conditions two-point objects are resolved in an image. Only the lateral two-point resolution will be discussed here.

The Rayleigh criterion defines two-point sources as resolved, if the image of the second

point lies at the first zero of the image of the first one or at a larger distance. This leads to a lateral resolution

$$d_{\mathrm{R}} = \frac{0.61\lambda}{\mathrm{N.A.}} \tag{10}$$

for the conventional microscope and

$$d_{\mathrm{R,CF}} = \frac{0.56\lambda}{\mathrm{N.A.}} \tag{11}$$

for the confocal microscope, just 8 % less than for the conventional microscope. In the case of coherent sources, d_{R} depends on the phase difference of the sources: whereas two out-of-phase coherent sources can be clearly resolved, because there will be a zero of intensity halfway between the images of the two sources, the sources cannot be resolved, if they are in phase with each other, because the maximum will lie in the middle between the images of the two individual sources.

There is another criterion for the two-point resolution, which is more generally applicable, because it does not refer to a zero of the response function. The Sparrow criterion states that two sources are considered to be resolved, when the intensity halfway between the two images is the same as the one at the individual image locations. For incoherent illumination, this results in

$$d_{\mathrm{S}} = \frac{0.51\lambda}{\mathrm{N.A.}} \tag{12}$$

for *both* the conventional and confocal microscopes.

Coherence in Brightfield and Fluorescence Microscopy

The imaging of extended objects differs significantly for the conventional and confocal microscopes, respectively. For a CLSM operating in brightfield (reflection or transmission) mode, the imaging is spatially coherent, because the illumination generates a spatially coherent field distribution in the focus of the objective lens, as given by the complex amplitude point spread function of the objective lens. For a CLSM in reflection mode, this field distribution has to be multiplied by the reflectance R of the sample, and this weighted field distribution is then imaged onto the pinhole implying convolution with the combined amplitude point spread function of the objective and pinhole relay lenses. Assuming imaging of the object onto the pinhole by the same objective lens that is used for excitation and neglecting contributions to the point spread function from the pinhole relay lens in the second imaging step, the intensity behind the infinitesimally small pinhole can therefore be written as the convolution

$$I = \left| \iiint h(x,y,z) R(x,y,z) h(-x,-y,-z) \mathrm{d}x\mathrm{d}y\mathrm{d}z \right|^2 . \tag{13}$$

Here, $h(x,y,z)$ is the amplitude PSF of the objective lens (compare Eq. 3). For the case of an even PSF, Eq. 13 is equivalent to

$$I = \left| h^2 * R \right|^2 \tag{14}$$

that means the signal is given by the absolute square of the convolution of the amplitude PSF of the confocal microscope, $h^2\left(\vec{r}\right)$, with the local amplitude reflectivity of the sample.

In the case of a conventional microscope, on the other hand, the illumination is approximately incoherent and therefore

$$I = |h|^2 * |R|^2 . \tag{15}$$

In reality, for the conventional microscope, imaging is partially coherent, because the emission from each emitting point on the illumination source is imaged, due to diffraction at the condenser aperture, into a finite spatial region, which is occupied by a coherent field due to that point emitter, and the regions in the image space corresponding to neighboring, incoherently emitting points on the source partially overlap on the sample. This is true for critical illumination, where the spatial region would be given by the PSF of the condenser, as well as for Köhler illumination, where the spatial region is the whole illuminated region of the sample.

Fluorescence Confocal Microscopy

Fluorescence imaging is incoherent. Assuming that the fluorescence intensity is proportional to the excitation intensity, the signal obtained in a confocal microscope with an infinitesimally small pinhole is

$$I = \left(|h(\lambda)|^2 |h(\beta\lambda)|^2\right) * f \qquad (16)$$

where it has again been assumed that, as in standard commercial confocal microscopes, the same objective lens is used for excitation and fluorescence imaging, respectively. Here, f represents the distribution of fluorescent centers in the sample, λ the excitation wavelength, $\beta\lambda$ the fluorescence wavelength, and β the ratio of both wavelengths, the Stokes ratio. In the case of several different types of emitters, f would have to be weighted according to the spectral contribution of each emitter to the detected signal, which depends on the filters employed in detection for rejection of the excitation and also on the wavelength-dependent sensitivity of the detector. For a point emitter placed at the focus, the convolution with a δ function just yields the first two terms on the right-hand side of Eq. 16. For a conventional microscope, on the other hand, because the whole sample is illuminated, the single-point resolution is only determined by the intensity PSF of the objective lens at the fluorescence wavelength, and the excitation wavelength is irrelevant.

Effects of Finite Pinhole Size

A finite size of the pinhole is obviously required in order to obtain a measurable signal. This will reduce the lateral resolution as well as the optical sectioning capability. It can be shown [2] that for a single-point object, the lateral resolution is almost unaffected by the size of the pinhole, if

$$v_P = \frac{2\pi r_p}{\lambda M} \sin\alpha \leq 0.5 \qquad (17)$$

where r_P is the radius of the pinhole and M the magnification of the lens. The maximum value allowed by Eq. 17 therefore sets a reasonable value for the pinhole size, if optimum lateral resolution and reasonable signal are desired. As an example, in the case of a 100×/0.8 N.A. objective lens and $\lambda = 514$ nm, a critical diameter of the pinhole of 10.2 µm is calculated from Eq. 17. The depth discrimination, as measured by moving a mirror through focus, on the other hand, is less affected and is essentially unaltered if

$$v_p \leq 2.5 \qquad (18)$$

It is obvious from these equations that the pinhole size must be adapted when the objective lens is changed.

Apodization

Equation 3 assumed rectangular apodization, that is, homogeneous illumination of the lens pupil and neglects reflection losses at the lens. For a given objective lens, rectangular apodization yields the smallest FWHM of the focus in the focal plane, at the expense of larger side maxima, compared to pupil functions falling off toward the edge of the objective lens. It can be achieved only approximately with Gaussian laser beams and is then equivalent to loss of a large fraction of the power of the excitation beam. Gaussian apodization, on the other hand, increases the FWHM, but reduces the side maxima.

Deconvolution

As described by Eq. 16, the image acquisition process in confocal fluorescence microscopy can be modeled as a convolution of the spatially dependent fluorescence of the sample with a point spread function (PSF) of the imaging system. This is true also for conventional non-confocal fluorescence microscopy. In addition, random noise is superimposed on the image. Deconvolution with the PSF would naturally seem the appropriate way to determine the true fluorescence distribution in both cases. Applied to confocal images, the resolution may be improved. In the case of conventional microscopy, optical sectioning and removal of

out-of-focus blur may be achieved by post-processing instead of employing hardware in the optical setup.

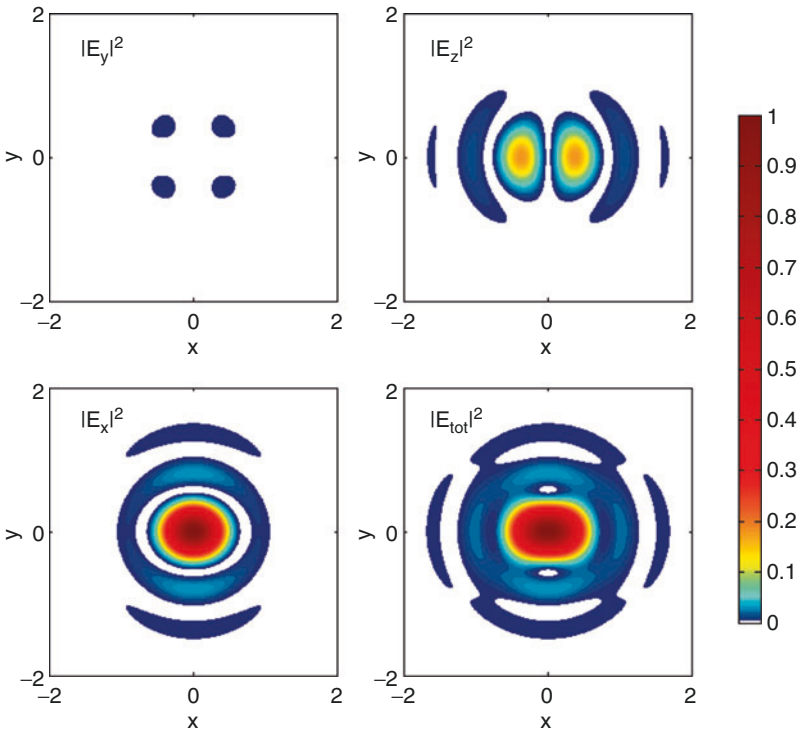
In general, the 3D PSF, necessary for this procedure, can be obtained experimentally or analytically. In the experimental methods, images of one or more point-like objects are collected. The problem with this technique lies in the poor signal-to-noise ratio obtainable with very small objects and the fact that the PSF may vary depending on the sample. In analytical calculations of the PSF, aberrations of the optical system are often partially taken into account, whereas, on the other hand, the scalar approximation is most often used and the effects of the vector character of light (compare section "Deficiencies of the Scalar and Paraxial Approximation: Effects of the Vector Character of Light") are neglected.

Many 3D deconvolution methods are currently employed and some are available in commercial and noncommercial software packages [4]. The simplest class are *neighboring* methods, in which out-of-focus blur is removed by subtraction of neighboring (filtered) images within a stack. This method does not sufficiently remove noise. In contrast to that, *linear* methods apply deconvolution to the whole stack of images at once. Examples are inverse filtering, Wiener filtering, the linear least squares, and the Tikhonov filtering techniques. The last three methods do not restore high-frequency object components beyond the bandwidth of the PSF, and inverse filtering suffers from noise amplification. All methods are very sensitive to error in the PSF. Therefore, constrained iterative nonlinear algorithms are often employed, in which, starting from a guess for the true object, an error is minimized under certain constraints (positiveness of the image, finite support of the sample, etc.). In cases of strong noise in the image, *statistical* iterative methods (like the maximum likelihood method) are favored. A computational alternative are *blind deconvolution* methods, in which the PSF of the optical system and the "true object" are simultaneously determined in a converging iteration. These methods are, however, computationally demanding and sensitive to noise, and solutions may be nonunique.

Deficiencies of the Scalar and Paraxial Approximation: Effects of the Vector Character of Light

In view of many more recent developments in confocal microscopy, it appears useful to shortly discuss deviations from the simple scalar theory. These deviations become increasingly important with increasing numerical aperture of the objective lens. In the case of linear polarization of the excitation beam, cylindrical symmetry is lost. The field in the focal plane and exactly on axis is then polarized in the direction of the excitation, but, away from the axis, it is elliptically polarized with a longitudinal component pointing in the direction of the optical axis. The intensity PSF becomes elongated and approximately elliptical, with the major axis of the ellipse in the direction of the excitation. Both effects increase with increasing numerical aperture. In the case of linearly polarized excitation, the two-point resolution of an ordinary optical microscope equipped with a well-corrected high numerical aperture objective lens (as well as that of a similar confocal microscope) therefore depends on the orientation of the line connecting the two points relative to the incoming polarization (as well as on the orientation of the dipolar point reflectors or absorbers/emitters). Figure 3 shows the PSF for numerical aperture N.A. = 0.95 and the absolute squares of all electric field components in the focal plane. In the scalar approximation, the lines of constant intensity would, of course, be circles, which is clearly not the case in the figure appearing in the lower-right corner of Fig. 3. Moreover, the maxima of the longitudinal field occur on both wings of the main maximum, along the direction of the incoming polarization (x direction), and the maximum absolute square of E_z is approximately 20 % of that of E_x.

The vector diffraction problem was first solved by Richards and Wolf [5]. A somewhat more physical treatment employs expansion of the field in the image space into vector multipoles centered at the focus [6]. This latter approach is also of interest for matching the focal field distribution to the fields of a dipole via the amplitude distribution and the polarization in the pupil plane. In this way, the coupling of the field to single atoms or molecules can be significantly enhanced, which is of



Confocal Laser Scanning Microscopy, Fig. 3 PSF of a well-corrected microscope objective of a high numerical aperture lens (N.A. = 0.95) for the case of linear polarization of the incoming beam, calculated using the Debye-Wolf integral. A constant pupil function has been assumed here and, correspondingly, reflection losses in the lens have been neglected. z is the coordinate along the axis of the lens and x, y are the lateral coordinates, all measured in wavelengths. The incoming polarization is along the x axis

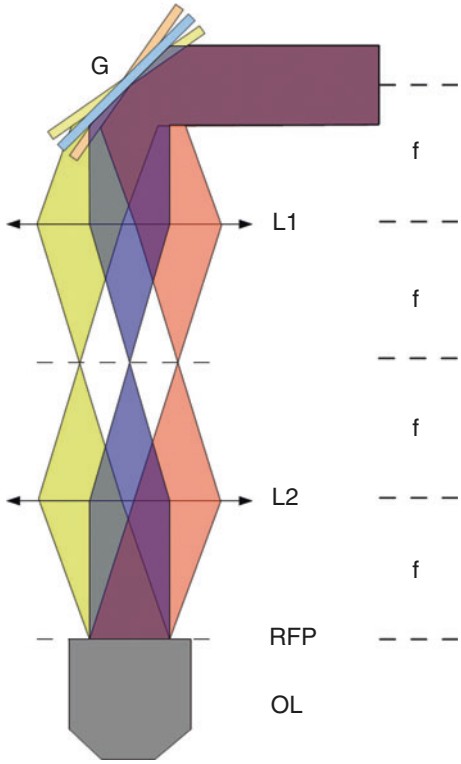
interest, for example, for quantum optical applications. The vectorial approach is in general also required to treat focusing of other distributions of intensity, phase, and polarization in the pupil plane of the lens. Examples relevant for applications include radially polarized excitation, producing a longitudinally polarized focus, azimuthally polarized excitation, which leads to a doughnut-shaped intensity distribution in the focal plane, or combinations of these distributions with scalar vortices, that means helical phase fronts. A longitudinal focus is essential for tip-enhanced Raman microscopy (TERS, compare section "Confocal Raman Microscopy," ▶ Scanning Near-Field Optical Microscopy), which combines a near-field technique with confocal imaging. Other distributions are relevant for optical tweezers.

Instrumental Details

Scanning Techniques

Whereas scanning the sample is an option, particularly in laboratory setups, many commercial confocal microscopes employ lateral scanning of the laser beam and sample scanning in the vertical direction. Beam scanning is typically achieved using mirrors mounted on galvanometer motors, which allow to vary the angle at which the beam passes the rear focal plane of the objective lens. An example for a telecentric 4f system that allows to vary this angle without displacing the beam in the rear focal plane is shown in Fig. 4.

Recently, resonant galvanometer-based beam scanning systems have become commercially available, which employ torsion spring-based sinusoidal oscillations of the scan mirror with frequencies in the kilohertz range with open-loop operation for the fast scan axis. This permits frame rates on the order of 30 frames per second and in this way allows to study fast processes, such as diffusion in biological cells or to avoid blurring of the image due to movement of organs in in vivo studies. Alternative fast-scanning confocal microscopes are discussed in section "Variants of Confocal Laser Scanning Microscopy."



Confocal Laser Scanning Microscopy, Fig. 4 Telecentric lens system minimizing beam walk off. *f* focal length of lenses L1 and L2, *G*: scan mirror mounted on galvanometer scanner, *RFP* rear focal plane of objective lens OL

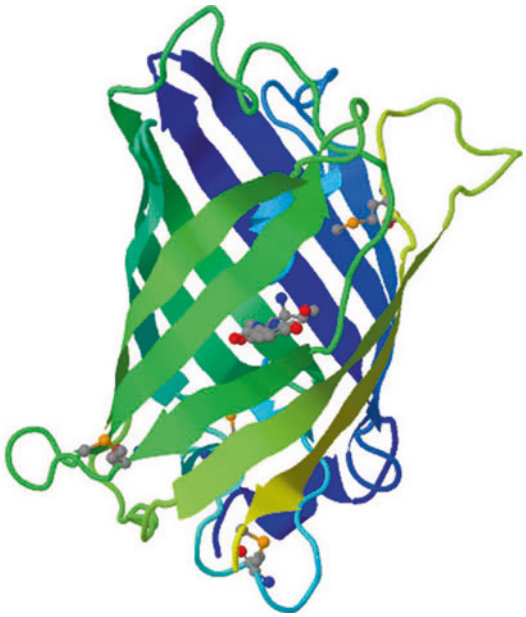
Excitation Sources and Beam Delivery

Ion lasers, HeNe lasers, diode-pumped solid-state lasers, and diode lasers have all been used as continuous wave light sources in confocal laser scanning microscopy. Argon ion lasers provide a choice of several excitation wavelengths and are therefore still popular in spite of their low efficiency. In order to combine several laser beams, dichroic beam splitters are employed. In many cases, a software-controlled acousto-optic tunable filter (AOTF) selects the excitation wavelength (s) of choice from this combined beam. The AOTF is based on the diffraction of light from ultrasonic waves generated in a birefringent crystal by an ultrasonic transducer. Incident and diffracted waves propagate as ordinary and extraordinary wave in the crystal, respectively, or vice versa and are therefore polarized perpendicular to each other. This allows convenient rejection of the incident light by a polarizer. Because of momentum conservation, the difference in the optical wave vectors of both beams must be equal to the acoustic wave vector. This means that the wavelength of the diffracted beam is controlled by the ultrasonic frequency. In a collinear AOTF, both optical beams and the acoustic wave propagate in the same direction, independent of the optical wavelength.

The optical output of the AOTF is typically fed into an optical fiber, which delivers the beam to the input optics of the confocal scan head of the microscope. In brightfield reflection confocal microscopy, the beam splitter which directs the excitation light toward the objective lens (compare Fig. 1) induces a considerable loss for the excitation as well as for the detected light. This loss is minimized by a 50/50 beam splitter. In fiber-based confocal microscopes, instead of beam splitters, 2 × 2 fiber couplers are typically employed. An improved version would make use of optical circulators, but these are presently not widely available for wavelengths in the visible range. In fluorescence confocal microscopy, the detected wavelength differs from the excitation wavelength and therefore dichroic dielectric beam splitters are used which are highly reflecting (transmitting) at the excitation (detection) wavelength and minimize losses in this way.

Detection

Standard detectors in confocal laser scanning microscopes are photomultipliers, which are in many cases operated in the analogue mode, which means by measuring the anode current and integrating over the pixel dwell time. Photon counting, on the other hand, is typically employed in fluorescence correlation and fluorescence lifetime microscopy (compare section "Variants of Confocal Laser Scanning Microscopy"), where usually avalanche photodiodes with high quantum efficiency and fast response times replace photomultipliers as detectors. Confocal



Confocal Laser Scanning Microscopy, Fig. 5 Tertiary structure of the green fluorescent protein [8]. The fluorescent chromophore, composed of three amino acids is located in the center of the beta-barrel protein cage, length about 4 nm, which prevents quenching of the fluorescence by water

microscopes which allow to spectrally disperse the light passing the confocal aperture often employ a charge-coupled camera (CCD) attached to a spectrograph to register the spectra. Back-illuminated Peltier or liquid nitrogen cooled CCDs provide high quantum efficiency (above 90 % over a wide spectral range) and low background noise, which is important for single molecule detection or when working with less photostable fluorescent probes.

Fluorescent Probes

Many samples are autofluorescent and therefore allow fluorescent imaging without having to introduce additional fluorescent probes. In biological samples, however, autofluorescence is typically weak, and therefore the sample frequently had to be stained with appropriate dyes. The latter, however, are often highly phototoxic in living cells. An important development in light microscopy, including confocal laser scanning microscopy, started in the year 1994 when the green fluorescent protein (GFP, see Fig. 5) from the jellyfish *Aequorea victoria* was genetically expressed in bacteria making them fluorescent at room temperature. In the same way, it is now generally possible to label proteins of interest in biological cells with fluorescent proteins by genetic manipulation, which can be achieved, for example, by injection of a virus vector. The number of fluorescent proteins used in the field has exploded by now and they are widely used in optical microscopy because of their considerably reduced phototoxicity, brightness, and photostability [7]. Genetically modified fluorescent proteins from *Aequorea victoria* now cover the spectral range from the deep blue to yellow and others derived from Anthozoa species (corals and anemones), as well as other sources, span the entire visible spectrum. The tertiary structure and size of these fluorescent proteins are very similar to those derived from *Aequorea victoria*, although the amino acid sequences are quite different. Red-emitting species with correspondingly longer excitation wavelengths are of particular interest because of reduced autofluorescence, deeper penetration, and better resistance toward high excitation density of biological tissue in this spectral region. Research is ongoing to improve brightness and photostability, reduce oligomerization and pH sensitivity, improve the appropriateness for fusion tagging, and reduce the time required for maturation of the protein in living organisms.

Variants of Confocal Laser Scanning Microscopy

Confocal Microscopy Involving Modified Illumination

Slit-Scanning Confocal Microscopes

Scanning a line focus, generated, for example, by a cylindrical lens, over the sample and imaging this line focus onto a slit aperture parallel to the image of the line focus still provides the optical

sectioning capability of the confocal microscope, because the slit rejects out-of-focus light. At the same time, the frame rate is significantly increased, because scanning is necessary only in one direction. A spatially sensitive detector must be used to resolve light passing through different positions along the exit slit. This may be achieved by imaging the exit slit onto a one-dimensional detector array, read out synchronously with the scan, or by scanning an image of the exit slit, synchronously with the scan, across a two-dimensional detector, such as a CCD camera, forming a confocal image in this way. Disadvantages over the single-point scanning technique include reduced lateral resolution in the direction of the line focus, enhanced out-of-focus background, and, in the case of coherent illumination, increased laser speckle.

Spinning Disk Confocal Microscopes

Instead of scanning a single-point focus across the sample, multiple focal spots may be simultaneously generated and imaged each onto a confocal aperture. A white light version of such a confocal microscope based on a spinning Nipkow disk was introduced by Petran and Hadravsky already in the 1960s and later improved by Xiao, Corle, and Kino. The disk contains pinholes arranged in a spiral pattern, which are slightly displaced such that the whole sample is illuminated after the disk has rotated by a certain angle. In more recent versions, light from each focal spot on the sample passes the same pinhole in the disk that was used for excitation. A two-dimensional detector, such as the eye of the observer or a CCD camera, is used to register a confocal image while the disk is spinning. Nipkow disk-based confocal microscopes are now commercially available from several manufacturers and provide fast confocal imaging, but at the cost of reduced flexibility, because the pinhole size cannot be varied and because the beams cannot be steered at will, as it is necessary, for example, in some experiments involving photobleaching. Moreover, cross-talk between the different focal spots may occur. Another version of confocal microscopy with multiple focal spots, *swept field confocal microscopy*, leaves the pinhole array stationary and sweeps the image of this array over the sample. By switching between different pinhole arrays, the pinhole size can be varied.

The light efficiency of Nipkow disk-based confocal microscopes may be significantly improved by adding a microlens array, mounted on a second disk, which is spinning on the same axis as the Nipkow disk and placed on top of the latter. Each microlens focuses incoming light onto one of the pinholes of the Nipkow disk. A dichroic beam splitter between both disks may be used to direct the detected fluorescence onto a camera. Another option is to use slit-shaped apertures on the Nipkow disk, which results in the same advantages and disadvantages as already described in section "Slit-Scanning Confocal Microscopes." For a review of applications of spinning disk microscopes in life science, see reference [9].

Chromatic Brightfield Confocal Microscopy

A chromatic confocal microscope operating in reflection deliberately introduces chromatic aberrations into the imaging system. Scanning in the vertical direction is then replaced by simultaneous detection of different spectral components which encode the depth information, because the depth of the focus depends on the wavelength. A complete stack of images can be acquired in this way in a single two-dimensional mechanical scan of the sample or of the excitation beam. Broadband excitation may be provided by a white light lamp or by a femtosecond laser-generated supercontinuum.

Structured Illumination Microscopy (SIM)

The SIM technique does not employ any pinhole and can be used with white light, but is related to confocal microscopy in its optical sectioning capability [10]. Optical sectioning is achieved by acquiring a sequence of images of the sample with structured illumination. The simplest case of structured illumination is thereby produced by placing a grid of fully transparent and fully opaque stripes of equal width (one half the period L of the grid) into the illumination path and projecting this grid onto the sample. Only sample

structures that are in focus will lead to significant variations of the image when the grid is displaced along the direction of periodicity, because only in focus the image of the grid within the sample is sharp. After acquiring three images with the grid displaced by 0, L/3, and 2 L/3, the optical section can be calculated (in the simplest version of the SIM algorithm) as the root mean square of the three differences between the three images. More sophisticated deconvolution algorithms are available and many other structures for illumination can be used. Movement of a grid illumination pattern across the sample may be replaced by the generation of arbitrary patterns by digital mirror devices (DMD) based on microelectromechanical systems (MEMS) or on spatial light intensity modulators (SLM), based on liquid crystals.

Whereas optical sectioning can be achieved more easily in this way than in a standard confocal system with a single-point focus, there are also some problems related to this approach. SIM works badly in strongly scattering samples, because small differences on a large background have to be determined. This is related to a significant loss in bit resolution and, hence, dynamic range in the final image. Moreover, the optical sectioning capability of SIM is slightly worse than for the standard single focus confocal microscope.

Another version of SIM employs either a grid pattern or random aperture arrays on a spinning disk, which both allow a large throughput of the light from the illumination source of the order of 50 %. Because of cross-talk between the light transmitted through neighboring apertures, the image acquired through the disk will contain a part that is not in focus. This part has to be subtracted from the image. A corresponding conventional image, to be subtracted from the partly confocal image, may be acquired by tilting the disk slightly, reflecting a second light source from the rear side of the disk toward the microscope objective and registering the corresponding image using a second camera. This procedure allows rapid optical sectioning and, hence, in vivo imaging of biological samples with very good signal-to-noise ratio with a comparatively simple instrument. A similar approach of subtracting the conventional image is based on a DMD, instead of a spinning disk, and was termed *programmable array microscope*.

Versions of structured illumination microscopy providing a moderate degree of super-resolution are based on Moiré patterns produced by projecting a high spatial frequency grid onto the sample (high-resolution SIM, HR-SIM). The Moiré pattern arises, because the observed signal is the product of the spatial distribution of the excitation with the concentration of the fluorophore and therefore contains spatial frequencies equal to differences between sample spatial frequencies and the one of the grid. The grid pattern must not only be shifted, but also be rotated, in order to be able to calculate the image. The method is able to increase the resolution by a factor of two beyond the diffraction limit.

Confocal Microscopy Beyond Brightfield and Standard Fluorescence

Confocal Raman Microscopy

In the Raman spectroscopy mode, the inelastically scattered light from the sample is detected, where the frequency shift toward lower (Stokes signal) or higher photon energy (anti-Stokes signal) coincides with an internal vibration of the sample. Raman scattering is typically very weak, and strong rejection of elastically scattered laser light is necessary. Historically, triple monochromators were often used for this purpose and still yield the highest spectral resolution, but in recent years dielectric long-pass filters with ultra-sharp transmission edges and high rejection factors have become available. This simplifies the instruments significantly, increases light efficiency, and in combination with slit-scanning techniques (see section "Slit-Scanning Confocal Microscopes") allows rapid multispectral confocal Raman imaging with acquisition times per pixel in the millisecond range.

Coherent anti-Stokes Raman scattering (CARS) and *stimulated Raman scattering (SRS)* microscopies are nonlinear variants of Raman

microscopy based on stimulated Raman scattering. Both techniques require two short-pulse lasers operating at different frequencies ν_1 and ν_2. The overlapping beams are focused by the microscope objective into the sample. When the difference in the optical frequencies is tuned to the frequency of a characteristic vibrational frequency ν_v of the sample ($\nu_1 - \nu_2 = \nu_v$), it will excite this vibration and at the same time generate anti-Stokes Raman scattered light at a frequency $2\nu_1 - \nu_2 = \nu_1 + \nu_v$ (CARS) or weakly deplete the pump and enhance the Stokes beams (SRS). Stimulated Raman scattering can be several orders of magnitude stronger than spontaneous Raman scattering and therefore allows rapid label-free imaging of a particular molecule in the sample. At least one of the two lasers necessary for CARS or SRS has to be tunable. This requirement maybe fulfilled, for example, by a Ti:sapphire laser or an optical parametric oscillator. Video rate in vivo SRS microscopy has been recently demonstrated and offers a number of advantages over CARS microscopy.

Another variant of confocal Raman microscopy, tip-enhanced Raman scattering (TERS) employs optical near fields of a sharp tip in order to increase the spatial resolution in Raman scattering over the diffraction limit. Confocal excitation and detection thereby reduce elastically scattered background in this setup.

Multiphoton Microscopy

Fluorescence excitation may in general be based on a linear process, in which a single photon excites the emitter into the excited state at energy E_1 from which it fluoresces, or on nonlinear processes, in which the emitter simultaneously absorbs n photons of energy E_1/n. Nonlinear processes are usually much less likely to occur than linear ones, with the probability strongly decreasing with the number of photons. Therefore, pulsed lasers are used for excitation, in which the optical power is concentrated in short pulses of widths in the femtosecond to picosecond range and the intensity during the pulse is very high. In addition, focusing the beam to a submicron spot leads, of course, to strong additional enhancement of the excitation probability.

Multiphoton microscopy has several important advantages over optical microscopy employing linear excitation. First, it is inherently confocal, in the sense that out-of-focus contributions to the detected signal are very small and therefore a confocal pinhole is not required (or the existing pinhole can be opened fully without losing the optical sectioning capability). This is advantageous, in particular for strongly scattering samples, because it yields a higher detection efficiency. Second, the photon energy used for excitation is only one half (for two-photon excitation) or one third (for three-photon excitation) of that used in the linear case, and the excitation wavelength therefore typically lies in the near infrared or even further in the infrared. Because the scattering in biological tissue and other inhomogeneous materials with inhomogeneities on a scale of the wavelength or below decreases strongly with the wavelength (proportional to λ^{-4} for very small inhomogeneities), the penetration depth is considerably larger for multiphoton excitation compared to single-photon excitation of the same fluorophore. This means that three-dimensional imaging deep into tissue becomes possible. Third, photo-induced damage to the sample and the fluorophore is reduced, also because of the lower photon energy.

Instead of making use of multiple photon excitation of a fluorescent chromophore, it is also possible in some samples to detect light due to second harmonic generation (SHG) and to use that for label-free imaging. Other label-free and, in addition, chemically specific, variants of multiphoton microscopy are CARS microscopy and SRS microscopy, as described in section "Confocal Raman Microscopy."

Fluorescence Lifetime Imaging (FLIM)

The lifetime of a fluorophore may vary depending on local pH, oxygen, or ion concentrations, or on intermolecular interactions, for example, fluorescence resonance energy transfer (FRET, see section "Fluorescence Resonance Energy Transfer

(FRET)"). On the other hand, within some limits, it does not respond to the intensity of the excitation light, the fluorophore concentration, or photobleaching. The fluorescence lifetime is therefore a useful physical quantity that can be used for imaging and quantitative analysis of local pH, ion concentrations, or intermolecular interactions. Fluorescence lifetimes are typically in the range of a few picoseconds, when dominated by non-radiative processes, to several tens of nanoseconds, when limited by the radiative transition rate. In some cases it is useful to employ fluorophores with very long lifetimes, in particular when strong autofluorescence is present. Performing a lifetime measurement at each position of the excitation laser focus within the sample and representing the corresponding values as a gray scale or color value from a lookup table yield a confocal lifetime image. At the same time, a standard fluorescence intensity image can be obtained.

A common method of measuring fluorescence lifetime is time-correlated single-photon counting (TCSPC). Here, a correlator, triggered by the short pulse of the exciting laser measures arrival times of fluorescence photons, typically detected by an avalanche photodiode with a short response time, and generates a histogram of delays. Fitting an exponential function to the histogram yields the fluorescence lifetime as the decay time of the exponential. Complications may arise when the decays are actually non-exponential. In the case of relatively long lifetimes in the nanosecond range, instead of TCSPC, a gated image intensifier may be used to measure the number of photons falling into a time window defined by the gate. Both techniques operate in the time domain. When the lifetime is comparatively long, it is also possible to modulate the laser pulses in the MHz range and detect the phase shift of the corresponding modulation in detected fluorescence intensity, which depends in a simple way on the fluorescence lifetime. TCSPC, however, is the most flexible way of measurement, because it is not restricted to long lifetimes and allows to analyze non-exponential decays as well.

Fluorescence Resonance Energy Transfer (FRET)

FRET [11] is a resonant non-radiative energy transfer from a donor to an acceptor fluorophore due to the dipolar interaction. As the dipolar interaction energy is proportional to the third power of the donor-acceptor distance R and because the probability for FRET to occur involves the square of an off-diagonal matrix element of the dipolar interaction, it falls off as R^6 and depends on the relative orientation between the molecules and the spectral overlap between the emission spectrum of the donor and the absorption spectrum of the acceptor. The probability is highest for parallel orientation of the donor and acceptor transition dipoles. Because of the steep fall, off FRET can only occur if R is sufficiently small, that means, if the donor and the acceptor are sufficiently close to each other. Because the critical distance is only 1–10 nm, typically 4–6 nm, in all practical cases, the FRET mechanism provides a *molecular ruler* for measuring the donor-acceptor distance and in this way provides an indirect mechanism for studying structure on the nanometer scale, which cannot directly be resolved by diffraction-limited optical microscopy (superresolution microscopy, compare section "Super-Resolution," might, however, in the future partly supersede FRET studies). This is of particular interest for protein-protein and intra-protein interactions. For this purpose, the proteins of interest have to be labeled by fluorophores with overlapping emission and absorption spectra. In many cases fluorescent proteins (compare section "Fluorescent Probes") are used for this purpose. A sensitive measure for FRET, which can be used for imaging, is the ratio of intensities of the donor and acceptor fluorescence peaks.

A problem with FRET confocal microscopy is that the signal-to-noise ratio is often very poor. Therefore, in many cases only the occurrence or absence of FRET is detected. The signal-to-noise ratio may be improved by measuring the donor lifetime, instead of the acceptor/donor fluorescence intensity ratio. This imaging option is usually known as FLIM-FRET (*fluorescence lifetime imaging-fluorescence resonance energy transfer*).

Fluorescence Recovery After Photobleaching (FRAP)

This microscopy technique, most often performed in a confocal setup, bleaches the fluorescence of a certain sample region, often an intracellular organelle, and registers the recovery of fluorescence due to diffusion of the fluorescently labeled molecules. It may therefore be used to investigate the mobility of the target molecule within the surrounding structures.

Fluorescence Correlation Spectroscopy (FCS)

In contrast to the previous techniques, FCS is usually used not as an imaging technique, but with a fixed position of the confocal volume within the sample. The laser (usually a continuous wave laser) thereby excites fluorescent particles within this volume, and particle movement in and out of the volume produces fluorescence intensity fluctuations. The autocorrelation function of these fluctuations provides information about the concentration, diffusion coefficient, and the mass of the particles. The diffusion coefficient depends on the viscosity of the medium via the Einstein-Smoluchowski relation. It may also depend on interactions of the particles with a microstructured environment.

A variant of FCS is fluorescence cross-correlation spectroscopy, in which the fluorescence intensity fluctuations of two fluorophores, labeling two different molecules, are measured simultaneously in two different channels. If the two molecules are bound in a dimer, the fluctuations will be highly correlated. Otherwise, no cross correlation is expected. The degree of cross correlation then is a measure of how many of the two different species of molecules are bound to each other.

Confocal Microscopy Sensitive to Phase

Interferometric versions of confocal laser scanning microscopy have been reported relatively early in the literature and were based on Mach-Zehnder or Michelson interferometers, for measurements in transmission or reflection, respectively [1]. In this way, it is possible to measure object topography or refractive index variations with interferometric precision. Interferometric confocal microscopes have the advantage over conventional interferometric microscopes that the shape of the wave front of the reference beam and are irrelevant because only phase and amplitude at the pinhole is important. This means that the requirement for matching optics, otherwise necessary in interferometric microscopy, is strongly relaxed.

Often two beam splitters and two detectors, each with its own pinhole, are employed in the detection beam path of interferometric confocal microscopes based on Michelson interferometers. This allows to separate the conventional confocal signal and the pure interference signal as the sum and difference of the outputs of the two detectors. This is based on the fact that two beams at the inputs of a symmetric lossless beam splitter are combined with a $+\pi/2$, $-\pi/2$ phase shift relative to each other at the two outputs of the beam splitter, respectively. A non-interferometric variant of phase-sensitive confocal microscopy, *differential phase-contrast confocal microscopy*, is obtained by omitting the mirror generating the reference beam in the Michelson interferometer and obscuring one half of each of the relay lenses in a complimentary manner [2].

The sensitivity of interferometric confocal microscopes may be enhanced by using a spectrally shifted reference beam (*heterodyne interferometric confocal microscopy*). Topographic resolution of about 0.01 nm using such a heterodyne interferometric confocal microscope has been reported. In addition, the optical sectioning capability of the confocal microscope and the corresponding dependence of the signal on defocus can be used to avoid phase unwrapping ambiguities.

Another variant of interferometric confocal microscopy is called *4π microscopy*. Here, two opposing objective lenses are used for excitation and/or detection. This increases the available numerical aperture and therefore enhances the resolution, in particular, in the axial direction. In addition, because of the coherent excitation, the

technique is inherently interferometric due to the interference of the two counterpropagating excitation beams in the focal region. This generates a standing wave interference pattern which modulates the main lobe of the focus in the axial direction and in this way allows to improve the axial resolution by a factor of about 4.5. The side lobes, due to interference, within the confocal main lobe may be effectively suppressed in the case of two-photon excitation. This suppression is due to the nonlinearity of the excitation process and can be additionally enhanced, if the sample fluorescence is also detected through both objective lenses in an interferometric setup. Because the position of maxima of the PSF for excitation and detection depends on the phase of both beams used for excitation and detection, precise control of the phase of both beams is required.

It should be mentioned here that a non-interferometric technique for phase measurement in optical microscopy is based on the so-called transport-of-intensity equation which can be derived from the paraxial time-dependent wave equation and relates the intensity and phase of a paraxial monochromatic wave to its longitudinal intensity derivative. The technique requires, however, the measurement of very small changes of intensity as a function of small defocus and therefore requires high bit resolution and long integration times.

Superresolution

Historically, the Abbe diffraction limit had been an insurmountable barrier in optical microscopy for many years. Many structures of interest, on the other hand, are significantly smaller than this limit and, particularly in biology, there has been a strong desire to significantly improve the resolution. In recent years, the diffraction limit has been broken in numerous ways, and this is based on some of the most exciting modern advancements in optics, which is still, although very old, a rapidly developing area of physics. The conceptually simplest way to achieve sub-Abbe resolution is by way of deconvolution (compare section "Deconvolution"). This does not, however, allow to achieve resolution enhancements of an order of magnitude, as urgently desired in many cases. One route to satisfy this demand employs non-propagating near fields. These near fields carry all optical information on length scales below the Abbe limit. The fact that these near fields are lost in far-field optical imaging can be viewed as the reason for the limited resolution of standard (including confocal) optical microscopes. Near-field related techniques (compare "▶ Scanning Near-Field Optical Microscopy") will not be discussed here (as an exception, compare TERS, section "Confocal Raman Microscopy"), although some of them may be combined with confocal imaging, for example, *solid immersion lens microscopy* (SIL) or *total internal reflection microscopy* (TIRF). Other routes, however, have opened the way to far-field optical nanoscopy in recent years. They are based on prior knowledge about the sample (PALM/STORM) or on optical nonlinearity (STED, SSIM) [12]. STED uses a specially designed optical excitation and is a scanning microscopy technique. PALM/STORM is non-confocal, because parallel detection using a CCD camera is used. These techniques have reached lateral resolution in the range of 20 nm, and ongoing research attempts to further push the resolution toward the molecular level (1–5 nm). Commercial instruments based on these techniques are increasingly becoming available [13]. The techniques will be shortly described, because of their potential as far-field microscopy techniques to partly supersede standard confocal fluorescence laser scanning microscopy.

Superresolution by Prior Knowledge (Profilometry, PALM/STORM)

In cases where it is known that one and only one sharp interface between two homogeneous materials or one and only one point emitter or reflector in the focus is present, the position of the interface in the z direction or the three-dimensional position of the point emitter/reflector can be determined

with a precision that is far beyond the Abbe limit and is limited only by the total amount of photons detected.

The simplest such technique consists in profiling a surface using a standard non-interferometric CLSM. By adjusting the z position to the wing of the $V(z)$ function, very slight changes in the topography of the sample surface can be measured, if the material and, hence, the reflectivity do not vary. Sub-nanometer depth resolution has been achieved in this way, averaged over the diffraction-limited lateral size of the confocal spot. As discussed in section "Confocal Microscopy Sensitive to Phase," the technique may be combined with interferometry to increase the depth resolution to about 0.01 nm.

In a principally similar way, *photoactivated localization* microscopy (PALM) and *stochastic optical reconstruction microscopy* (STORM) determine the position of single-point emitters with nanometric precision. Because it must be avoided to have more than one molecule in the focal volume and such a sparse distribution of fluorescent emitters would not allow sub-diffraction-limited resolution, it is necessary to separate the contributions from many individual molecules in some way. *Temporal* separation is key to current single molecule-based superresolution techniques. Other options, such as spectral separation, or more sophisticated schemes have also been discussed, however.

Current techniques use *photoactivable* molecules, which can be statistically turned on by another light source. Irreversible or reversible processes may be employed to turn off an activated subset. Determination of the positions of the activated molecules and multiple repetition then results in a software-generated image with nanometric resolution. Typically, thousands of images must be acquired and, correspondingly, imaging is slow, with a trade-off between spatial and temporal resolution. Effective frame rates of reconstructed images of about 1/(3 min) at a Nyquist-limited resolution of about 50 nm have been demonstrated for frames of 50 × 50 pixels. The obtainable resolution depends on the brightness of the fluorophores in the "on" state, as well as on the achievable contrast between "on" and "off" states. Fluorescent proteins as well as fluorescent dyes are currently in use [14].

PALM has hitherto mostly been used in total internal reflection configuration and is then restricted to essentially two-dimensional imaging. Lateral resolution down to ≈20 nm has been achieved. A three-dimensional STORM technique has achieved an image resolution of 20–30 nm in the lateral dimensions and 50–60 nm in the axial dimension in an illuminated sample volume of a few micrometers in thickness. An interferometric variant of PALM employing self-interference of fluorescent photons has improved the axial resolution to <20 nm for optically thin samples.

It may finally be worth noting that, whereas photoswitching is considered by some authors as a kind of optical nonlinearity, an optically nonlinear response is not essential to the breaking of the diffraction limit in the far field in the case of single molecule-based techniques. For example, sub-diffraction imaging due to binding and detachment of diffusing probes has been demonstrated, which obviously does not involve any optical nonlinearity. Moreover, photoswitching and temporal selection of molecules in general represent conceptually only one, although currently the standard and most successful option to ensure that only one molecule is detected within the PSF of the objective lens. It may therefore be said that in this case the breaking of the diffraction barrier is based on the knowledge, however obtained, to have only one molecule within the PSF and on the facts that the center-of-mass position of a diffraction-limited spot can be determined to a much higher precision than given by the size of this spot and that this position corresponds to the position of the fluorescent molecule.

Superresolution Due to Nonlinearity (SSIM, STED)

Some examples of confocal imaging methods employing optical nonlinearity have already

been discussed in sections "Confocal Raman Microscopy"(CARS, SRS), and "Multiphoton Microscopy." By reducing the size of the PSF for photons of the same wavelength, optical nonlinearity is in principle able to enhance the resolution. In multiphoton microscopy, however, because long wavelength photons are used for excitation of fluorophores, the resolution is actually worse than that of its linear counterpart.

Saturated structured illumination microscopy, SSIM, on the other hand, provides resolution enhancement without fundamental limit. This nonlinear variant of SIM (compare section "Structured Illumination Microscopy (SIM)") employs a nonlinear dependence of the fluorescence intensity on the excitation density, which is possible, for example, by saturation of the excited state. In this case, the effective illumination generally contains higher spatial frequency components leading to increased resolution. Optical resolution beyond 50 nm has been reported using this technique.

Stimulated emission depletion microscopy (STED) selectively de-excites fluorescing molecules within the PSF of the confocal microscope by stimulated emission, using a second laser. This laser beam passes a spiral phase mask, which generates a zero of intensity in the center of the beam cross section. Stimulated emission and corresponding deexcitation of fluorophores then occur everywhere within the PSF, except close to the axis of the beam, and the size of the region, from which fluorescence that can be collected depends on the intensity of the deexciting laser. Lateral resolution of about 20 nm has been reported using this setup. If STED is combined with a 4Pi setup, fluorescing spherical focal volumes of 40–45 nm diameter can be generated. STED is currently able to operate considerably faster than PALM. Video-rate STED microscopy with about 60 nm lateral resolution has been reported with a frame rate of 28 frames/s.

Cross-References

- ▶ Optical Techniques for Nanostructure Characterization
- ▶ Scanning Near-Field Optical Microscopy

C

References

1. Corle, R.C., Kino, G.S.: Confocal Scanning Optical Microscopy and Related Imaging Systems. Academic, San Diego (1996)
2. Wilson, T.: Confocal Microscopy. Academic, London (1990)
3. Born, M., Wolf, E.: Principles of Optics, 6th edn. Pergamon, Oxford (1980)
4. Sarder, P., Nehorai, A.: Deconvolution methods for 3-D fluorescence microscopy images. IEEE Signal Proc. Mag. **23**, 32–45 (2006)
5. Richards, B., Wolf, E.: Electromagnetic diffraction in optical systems. II. Structure of the image field in an aplanatic system. Proc. Roy. Soc. A **253**, 358–379 (1959)
6. Sheppard, C.J.R., Török, P.: Efficient calculation of electromagnetic diffraction in optical systems using a multipole expansion. J. Mod. Opt. **44**, 803–818 (1997)
7. Chudakov, D.M., Matz, M.V., Lukyanov, S., Lukyanov, K.A.: Fluorescent proteins and their applications in imaging living cells and tissues. Physiol. Rev. **90**, 1103–1163 (2010)
8. Ormo, M., Cubitt, A.B.., Kallio, K., Gross, L.A., Tsien, R.Y., Remington, S.J.: Crystal structure of the Aequorea victoria green fluorescent protein. Science **273**, 1392–1395 (1996). Image from the RCSB PDB (www.pdb.org) of PDB ID 1EMA (http://dx.doi.org/%2010.1021/ja1010652)
9. Gräf, R., Rietdorf, J., Zimmermann, T.: Live cell spinning disk microscopy. Adv. Biochem. Eng. Biotechnol. **95**, 57–75 (2005)
10. Langhorst, M.F., Schaffer, J., Goetze, B.: Structure brings clarity: structured illumination microscopy in cell biology. Biotechnol. J. **4**, 858–865 (2009)
11. Piston, D.W., Kremers, G.J.: Fluorescent protein FRET: the good, the bad and the ugly. Trends Biochem. Sci. **32**, 407–414 (2007)
12. Hell, S.W.: Far-field optical nanoscopy. Science **316**, 1153–1158 (2007)
13. Chi, K.R.: Ever-increasing resolution. Nature **462**, 675–678 (2009)
14. Heilemann, M., Dedecker, P., Hofkens, J., Sauer, M.: Photoswitches: key molecules for subdiffraction-resolution fluorescence imaging and molecular quantification. Laser Photon. Rev. **3**, 180–202 (2009)

Confocal Scanning Optical Microscopy (CSOM)

▶ Confocal Laser Scanning Microscopy

Conformal Electronics

▶ Flexible Electronics

Contact Angle

▶ Surface Engineering, Tailored Wettability, and Applications

Contour-Mode Resonators

▶ Laterally Vibrating Piezoelectric Resonators

Contraception

▶ Use of Nanotechnology in Pregnancy

Contrast Enhancement

▶ Magnetic Nanoparticles for Biomedical Applications

Cooling of Electronic Components

▶ Microchannel Flows as Heat Sinks

Core/Shell Nanostructures

▶ Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods

Coupled Mode Theory

▶ Harnessing Disorder at the Nanoscale

Coupling

▶ Nonlinear and Parametric NEMS Resonators

Coupling Clamp

▶ Dynamic Clamp

CPMD

▶ Car–Parrinello Molecular Dynamics

Creep

► Nanomechanical Properties of Nanostructures

Crystallite

► Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

Cuticle

► Arthropod Strain Sensors

Cylindrical Gold Nanoparticles

► Gold Nanorods

D

Decoherence and Dissipation in Nanoelectronic Devices

► Microscopic Modeling of Solid-State Quantum Devices

Decoration of Carbon Nanotubes

► Functionalization of Carbon Nanotubes

Defects

► Influence of Defects in Photocatalysis

Deicing

► Micro-/Nanostructured Icephobic Materials

Densification

► Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

Density

► Interfacial Investigation of Protein Films Using Acoustic Waves

Density-Matrix Theory

► Microscopic Modeling of Solid-State Quantum Devices

Deoxyribonucleic Acid

► DNA from First Principles

Deposited Metal Nanostructures

► Simulation of Supported Metal Clusters

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Dermal and Transdermal Delivery

Biana Godin[1] and Elka Touitou[2]
[1]Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA
[2]Institute of Drug Research, School of Pharmacy, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel

Synonyms

Dermal delivery – dermal absorption, skin penetration, cutaneous delivery, skin delivery

Transdermal delivery – percutaneous absorption, transdermal permeation, transcutaneous delivery

Definition

Dermal delivery refers to the process of mass transport of active ingredients applied on the skin to various skin strata.

Transdermal delivery refers to the entire process of mass transport of substances applied on the skin surface and includes their absorption by each layer of the skin, their uptake by microcirculation of the skin, and distribution in the systemic circulation.

Introduction

History of Therapeutic Systems Applied on the Skin

The history of skin application of therapeutic substances goes back to ancient times. Numerous carriers, including conventional semisolid bases (creams, gels, ointments), matrix systems (clays, polymers), and liquid systems (solutions, emulsions, suspensions), are being used for cutaneous application of therapeutics for centuries. Transdermal delivery, or delivery of active ingredients across the skin, was actually prototyped in ancient China and Egypt through the use of certain plasters and ointments. Several examples of medications topically applied on the skin to attain regional or systemic effect include the mustard plaster, used as a remedy for severe chest congestion, and Chinese balm or plaster containing menthol and ginger (rich in salicylates) for regional pain relief. In the case of mustard plaster, which later was introduced to the US Pharmacopeia, powdered mustard seeds (*Brassica nigra*) mixed with warm water were applied to the patient's chest with a cloth. Following the application, the change in the conditions to moist and warm activated some ingredients in mustard and particularly an enzyme myrosin. This enzyme is responsible for a hydrolysis of sinigrin and the release of active agent allyl isothiocyanate, which is transdermally absorbed being a low mw (~100 Da) mildly lipophilic substance. Other transdermal formulations recognized by the US Pharmacopeia in the early twentieth century included belladonna plaster (analgesic) and mercury ointment (for treatment of syphilis) [1].

Active development of the transdermal delivery systems started in the mid of the 1970s, exploring only the molecules which could intrinsically permeate the skin. The first transdermal delivery system, the scopolamine transdermal patch for treatment of motion sickness able to deliver the drug for 3 days, was granted with the FDA approval in 1979. The system was developed by the team of Alejandro Zaffaroni, who established ALZA Corporation. Currently, there are more than a dozen marketed transdermal delivery systems containing active ingredients such as nitroglycerin (cardiovascular conditions), nicotine (smoking cessation), various hormone combinations (contraceptives, hormone replacement therapy, etc.), lidocaine (local anesthesia), fentanyl (pain management), oxybutinin (urinary incontinence), methylphenidate (attention deficiency disorder), selegiline (depression), rotigodin (Parkinson's diseases), rivastigmine (dementia), granisetron (nausea and vomiting), and others [2].

Throughout the history, the vast majority of medicaments applied to the skin surface were intended to cure superficial conditions. It can be explained by the intrinsic barrier function of the skin, which efficiently inhibits penetration of

exogenous substances. To enable transdermal absorption or, on the contrary, enhanced accumulation of a therapeutic agent on the skin surface, it is important to thoroughly understand the skin physiology and barrier structure as well as and the physicochemical properties of drug molecules and carrier systems.

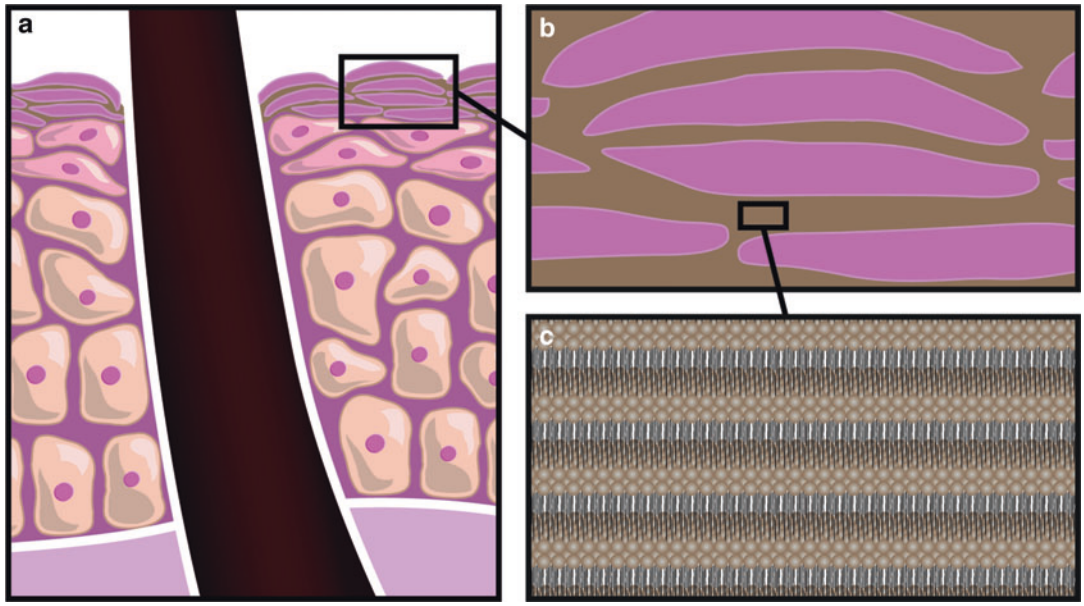
Skin Barrier

The skin, or integument, is the largest body organ, comprising 15 % of body weight and having a surface area of ~2 m^2. Among the several functions of the skin, the main one is being a physical barrier to the environment; controlling the inward and outward passage of water, electrolytes, and xenobiotics; and providing protection against pathogens, ultraviolet radiation, and mechanical insults. There are three structural layers to the skin: the epidermis, the dermis, and subcutis. Hair follicles and sebaceous, sweat, and apocrine glands are regarded as skin appendages. Epidermal layers of the skin have a dynamic turnover with cells of the outer strata continuously replenished by inner cells moving up to the surface from the basal layer. Epidermis and particularly its outermost layer stratum corneum (SC) represent the most important control elements in the transport of molecules into and across the skin [3, 4]. The actual investigation into the barrier properties of the skin started in the early 1920s. It was initially proposed by Rein that an intermediate layer of cell between SC and viable epidermis is responsible for the resistance to transcutaneous transport of molecules. This hypothesis was further modified by the groups of Blank et al. and Sheuplein et al. who have shown for the first time that SC is the limiting step in the process of passive diffusion across the skin. In 1944, Winsor and Burch have shown that removal of SC by sandpaper dramatically increased permeation of water and solutes into the skin. Nevertheless, in later years, it was demonstrated by Michaels and colleagues that some of the exogenous substances will readily bypass the skin as a function of their physicochemical structure. Great contribution to the understanding of skin barrier structure was done by Kligman and colleagues, who in a series of studies separated the SC membrane and has shown that it consists of well-defined polygonal corneocytes. The composition of intercorneocyte lipid domain [5] of the SC was first described by groups of Elias, Gray, and Yardley [3]. It was shown that the amount of the lipid in keratinocyte as well as the overall composition of the lipid layers in the skin strata is strongly affected by keratinocyte differentiation status. While undifferentiated keratinocytes in basal epidermis contain mainly sphingomyelin and phosphoglycerides, differentiated cells located in the stratum corneum (skin surface) contained mainly ceramides, cholesterol, and fatty acids in approximately equal molar ratios. The key enzymes responsible for the formation of the epidermal permeability barrier are serine palmitoyltransferase for the sphingolipids, β-hydroxymethylglutaryl-coenzyme A reductase (HMG-CoA reductase) for cholesterol, and acetyl-coenzyme A carboxylase [3, 6].

Well along, the widely known model of SC was proposed by Barry and Williams. This model describes SC as a "brick and mortar" structure where the bricks are corneocytes, terminally differentiated metabolically inactive cells of the skin epithelium filled with insoluble keratins, and mortar is the continuous intercellular lipid phase, composed mostly of ceramides, cholesterol, free fatty acids, and cholesteryl esters (Fig. 1). This unique chemical composition of SC and very low water content (∼15 %w/w) impart to this protective layer a highly lipophilic nature. Thus, only low molecular weight xenobiotics with intermediate lipophilicity can cross this intrinsic barrier and be absorbed to the deeper layers of the skin or, through the network of blood capillaries located in the dermis, to the systemic circulation [7, 8]. Efficient administration of substances with distinct physicochemical characteristics (ionic, highly lipophilic, large molecular weight, etc.), aimed to act in the deep viable skin strata, regionally in the immediately underlying tissues (muscles, ligaments, joints, etc.) or systemically, requires various techniques to enable their transport across the SC barrier.

Percutaneous Absorption

The term percutaneous or transdermal absorption refers to the entire process of mass transport of



Dermal and Transdermal Delivery, Fig. 1 (**a**) Schematic presentation of the skin structure. (**b**) Stratum corneum, the outermost layer of the skin and the main barrier for absorption, consists of terminally differentiated anucleated corneocytes filled with proteins and (**c**) intercellular stratum corneum lipids. The overall structure is frequently referred to as a "brick-and-mortar" wall

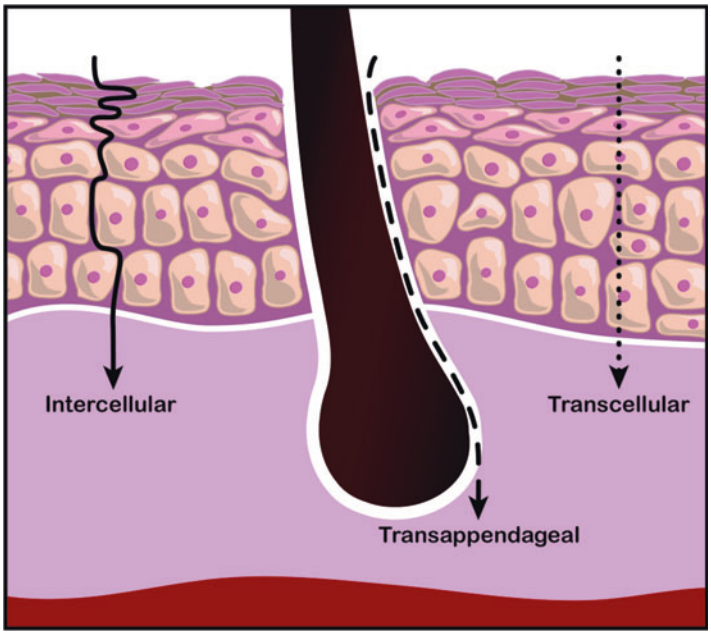
substances applied topically and includes their absorption by each layer of the skin and finally their uptake by microcirculation of the skin. It can be broken down into a series of sequentially occurring transport events: (i) deposition of a penetrant molecule onto the SC; (ii) diffusion through SC; (iii) diffusion through the viable epidermis; (iv) passage through the upper part of papillary dermis; and, finally (v), uptake into the microcirculation for subsequent systemic distribution. Since the viable tissue layers and the capillaries are quite permeable to solutes, diffusion across the SC is often regarded as the rate-limiting step.

The cells of the SC are anucleated and metabolically inactive; thus, passage through this layer is a passive process governed by classic diffusion laws, and a simple equation for steady-state flux can be used to describe it. When the cumulative mass of a diffusant, m, passing per unit area, A, through a membrane is plotted, at some point the graph approaches linearity, and its slope yields the steady flux, J (dm/dt) (Eq. 1) [7, 8]:

$$J = dm/dt = ADC_0K/h \qquad (1)$$

where C_0 represents the constant donor drug concentration; K, the partition coefficient of solute between membrane (*SC*) and the medium in which the molecule is applied; D, the diffusion coefficient; and h, the membrane thickness. From Eq. 1, the ideal properties needed for a molecule to penetrate stratum corneum are low molecular weight and high but balanced (optimal) partition coefficient and a low melting point, correlating with good solubility. Permeation modifiers can affect various parameters in the equation to tailor the absorption into or across the skin. Skin permeation enhancers, for example, can increase the diffusion coefficient by affecting the partition coefficient to increase the solubility of the substance in SC, decrease the thickness of the membrane, or temporarily reduce/diminish the membrane properties of SC.

A molecule applied on the intact skin surface has three possible routes to reach the viable tissue: intracellular, through SC lipids; intercellular across the corneocytes; and via skin appendages (Fig. 2). The majority of transdermally absorbed molecules are believed to permeate the skin



Dermal and Transdermal Delivery,
Fig. 2 Schematic presentation of the routes for permeation of the agent applied to skin surface to the viable layers of the skin and across the skin: intercellular, transappendageal, and transcellular routes

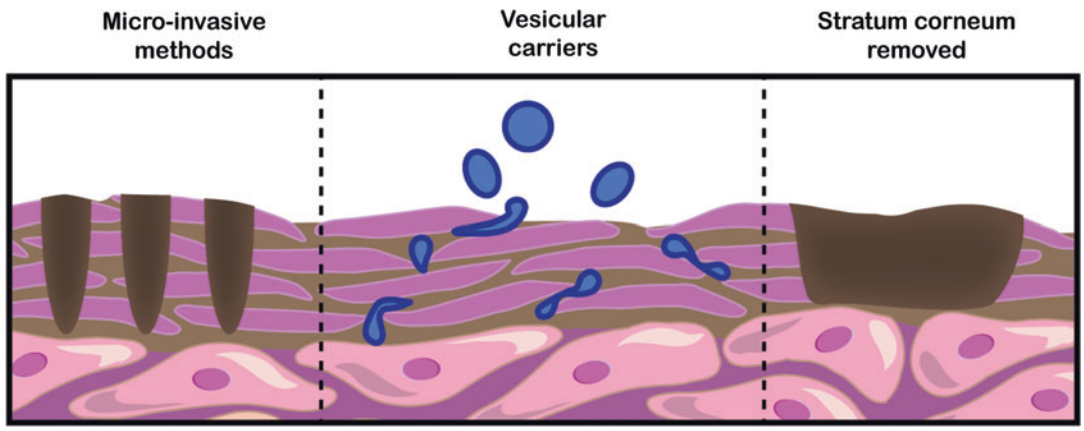
through the intercellular passage, and therefore, skin permeation enhancement methods disrupt or bypass the crystalline, semicrystalline, and liquid crystal domain formed by intercellular SC lipids. Because of the low fractional appendageal area (about 0.1 %), this pathway usually adds little to steady-state drug flux, with the exclusion being highly polar molecules or macromolecules that struggle to cross intact SC. However, appendages may function as shunts important at short times prior to steady-state diffusion. Some nanovectors can enhance accumulation of bioactive materials in skin appendages, thus benefiting conditions such as acne and alopecia.

Successful transdermal delivery aims to achieve the balance between attaining therapeutic levels of the drug in the required tissue and safety to the skin. Some of the transdermal delivery technologies assume that relatively large, micron-scale defects in the upper skin strata can be well tolerated if the viable strata of the skin (viable epidermis and dermis) are not majorly damaged [2]. In the case of microinvasive techniques, the permeation of the molecules occurs across the physically created channels in the skin barrier.

Nano- and Microtechnologies in Dermal and Transdermal Delivery

The main goal of formulating the active agents in a carrier system is to achieve delivery of therapeutically relevant levels at the target tissue lowering the incidence of systemic adverse reactions. The desired site of action can vary from the skin surface to various body organs/systems. With a very few exceptions, the ultimate destination of the molecule applied on the skin will depend on the vehicle or vector system.

In the past three decades, the research focus in the area of cutaneous and transcutaneous delivery largely shifted toward the application of micro- and nanotechnologies to affect the permeation behavior of active agents applied on the skin surface. These technologies range from systems which enhance the transport of bioactives across the skin barrier (e.g., specially designed vesicles, microinvasive methods) to those impeding skin penetration of small and mildly lipophilic molecules [9, 10]. There are several nano- and microtechnology-based mechanisms/carriers by which an enhanced dermal and transdermal systemic delivery can be achieved as schematically



Dermal and Transdermal Delivery, Fig. 3 Nano- and microtechnology-based methods to enhance skin absorption of active materials: microinvasive techniques (e.g., microfabricated needles, jet injectors), specially designed vesicular carriers (e.g., ethosomes, transfersomes), electrically assisted methods to temporarily remove SC barrier (e.g., radiofrequency ablation, electroporation)

presented in Fig. 3. These techniques are used when the systemic effect is desired or when the applied agent should affect viable strata of the skin or beneath tissues and can be broadly categorized to specially designed nanovectors that temporarily affect SC barrier properties and/or passively deliver drugs across the SC (e.g., ethosomes, transfersomes, surfactant-based elastic vesicles), minimally invasive techniques that actively birch SC and beneath strata (e.g., microneedles), and methods for removing portions of the SC (e.g., microporation, microabrasion) [10]. On the other hand, to maximize the effect of the drugs for alleviating superficial conditions and to minimize whole-body distribution and related side effects, increased concentration and contact time of an active ingredient with skin surface/upper epidermis, while minimizing its absorption, are required. In this case, nano- and microtechnology-based approaches can be useful for preventing transcutaneous absorption of low molecular weight mildly lipophilic agents which intrinsically cross the skin barrier (e.g., corticosteroids, sunscreens, retinoids). Large multilamellar liposomes, nonionic surfactant vesicles (niosomes), solid nano- and microparticles, dendrimers, and dendritic-core multi-shell nanotransporters are being investigated in this essence.

Further in this entry, main techniques for skin application of drugs aiming at different purposes are described.

Nano- and Micro-vectors

There is a great versatility of the nano- and microvectors suitable for topical, regional, and transdermal drug delivery of various active agents with distinct mechanisms of actions. These include phospholipid- and surfactant-based systems (liposomes, niosomes, ethosomes, transfersomes, elastic vesicles, micelles, etc.) and solid nano- and microparticles (solid lipid nanoparticles [SLN], polymeric nano- and microparticles, solgel particles, etc.) [10–15]. Several of these technologies were successfully tested in clinical trials and/or clinically used. Design considerations, structures, and modes of actions of the nano- and microvectors are further discussed with examples of therapeutics/nutraceutics that were shown to benefit from skin application by means of a particular carrier.

Vesicles

Most early reports on traditional *liposomes*, containing phospholipids, water, and cholesterol, were published by Mezei and colleagues in the beginning of 1980s. These and further studies have shown that when applied to the skin,

liposomes enable localized effect, depositing the encapsulated molecule in the SC. In vitro and in vivo studies in various animal models (from mice to rabbits) confirmed the local delivery behavior of liposomes [11, 12, 15, 16]. While corticosteroids and retinoids were the most investigated drugs for liposome-based delivery, reports on xanthines, local anesthetics, proteins, enzymes, tissue growth factors, immunomodulators, and other biologically active substances have also been published. Results from a clinical study on a corticosteroid betamethasone dipropionate encapsulated in liposomes have shown that an anti-inflammatory effect was more pronounced in patients with superficial skin inflammation (eczema) and not in patients with psoriasis, a condition involving deep skin strata [6, 17]. A few in vivo studies and clinical trials with liposomally encapsulated retinoids confirmed enhanced topical localization with improved tolerability [3, 10, 11]. The mode of action of traditional liposomes was investigated by a number of scientists in the area of dermal and transdermal drug delivery. Starting from Junginger and colleagues who provided valuable insights on the understanding of the interactions taking place on the liposome-stratum corneum interface using freeze-fracture electron microscopy, groups of Weiner, Touitou, Barry, Puglisi, Downing, Cevc, and others investigated mechanisms involved by a variety of microscopy techniques, radiolabeled vesicle distribution studies, and physicochemical analyses. It was shown that the ultrastructure of the intercellular lipid domains in SC is not affected by the liposomal application; no permeation of the liposome components to the deep layers of the skin is observed, and in general, liposomes do not penetrate SC and appear as stacks on the skin surface [3, 6, 11, 13, 16, 18, 19]. This can explain the formation of drug reservoir in the SC and lack of penetration into the deeper layers of the skin. In general, liposomes are recognized as a safe delivery system and approved for the last two decades by the regulatory agencies for skin application of pharmaceuticals (e.g., for acne and actinic keratosis) and cosmetic active ingredients as well as for other routes of drug administration (e.g., antifungal and anticancer intravenous medications) [11, 16, 17].

More recent efforts on the front of the *phospholipid-based vesicles* shifted toward systems that can enhance the transport of active ingredients across the SC barrier. Ethosomes and ultradeformable bilayer vesicles, also called transfersomes, represent the two systems most investigated in this endeavor.

Ethosomes, mainly composed of phospholipids, relatively high concentrations of ethanol, and water, were introduced by Touitou et al. [14, 20]. The action of this vesicular system is generated by the presence of phospholipid vesicles and ethanol in the system. The presence of ethanol in the vesicle's structure enables malleability required to cross the "mortar" domain of the SC. On the other hand, ethanol also temporarily fluidizes the lipid phase in the SC, thus providing a background for permeation of the soft vesicles. The mode of action, proposed by Touitou et al. in 2000, can, thus, be denominated into three main simultaneously occurring processes: (1) fluidizing effect of ethanol on SC lipid domain and on phospholipid membrane in the ethosome, (2) entry and passage of the soft ethosomes across the SC lipid "mortar" phase, and (3) fusion of the ethosomal vesicles with phospholipids in the viable skin strata and release of encapsulated components beneath the SC barrier. Moreover, the presence of alcohol enables encapsulation of molecules with a wide spectrum of physicochemical features. In vitro, in vivo, and clinical studies have shown that the composition of the system can tune their behavior in terms of delivery to various deep skin strata or transdermally. Numerous reports on ethosomal systems with various active ingredients were published by Touitou et al. and other research groups. These include antimicrobial and antiviral agents for deep skin infections, acne medications, anti-inflammatory agents, cannabinoids, anti-alopecia therapeutics, anti-parkinsonian trihexyphenidyl HCl, antianxiety drug buspirone HCl, antipyretic paracetamol and ibuprofen, large molecular weight compounds such as insulin, and many others [14, 18, 21, 22]. Ethosomes are safe carriers, and no skin adverse effects were reported with ethosomal products or in clinical studies. They are marketed for treatment of herpes simplex infection and used in cosmeceutical products.

Highly deformable bilayer vesicles, also called transfersomes, were introduced by Cevc as deformable liposomes mainly composed of phospholipids, an "edge activator," such as sodium cholate and water [23]. The inventors argue that due to the presence of a polar substance in the lipid phase of the aggregate, the resulting vesicle can squeeze when the osmotic gradient is applied and enters through the SC pores with less than one tenth of its diameter. In the proposed mechanism of action, two features are claimed to be important: (1) nonoccluded conditions and (2) presence of the hydration gradient from the anhydrous SC (10–15 % water) to the viable skin strata (~70 % water). When applied to the intact skin, the ultradeformable transfersome vesicles can cross a nanoporous barrier essentially intact despite the extreme bilayer vesicle deformation and appreciable intra-pore friction. This is enabled by the self-controlled local bilayer composition adjustment to the shape and overall curvature imposed by the pore on a vesicle. Later, Barry and colleagues evaluated in vitro skin transport of molecules from transfersomes vs. traditional vesicles. Both types raised maximum flux and skin deposition compared to saturated aqueous drug solution (maximum thermodynamic control) under a nonoccluded environment, but results were not as dramatic as detailed in earlier work. Transfersomes were proven in in vitro and in vivo studies transdermally to deliver steroids, NSAIDs, and large molecules (e.g., insulin, interferon alpha, serum albumin) across human skin [23, 24]. A transfersome-based therapeutic, Diractin® or ketoprofen in Transfersome® gel developed by Idea AG, was approved for clinical use. However, phase III clinical studies in Europe and the United States comparing Diractin® with the locally applied ketoprofen-free vesicles (used as an "epicutaneous placebo"), for the treatment of osteoarthritis of the knee, were discouraging and revealed no statistically significant difference between Diractin® and the placebo.

Vesicles based on *synthetic surfactants* were also suggested for skin application of active ingredients. In this category, niosomes and elastic vesicles are further briefly described as carriers for topical or systemic drug delivery, respectively.

Niosomes, proposed by Handjani-Vila et al., are nonionic surfactant vesicles comprised of bilayer assemblies formed from certain synthetic amphiphilic molecules in aqueous medium. Niosomes can contain various surfactants generally used in pharmaceutical preparations, including polyoxyethylene alkyl ethers, sorbitan esters, polysorbate-cholesterol mixtures, alkyl glycerol ethers, and others. Similarl to liposomes, these carriers initially developed in L'Oréal for cosmetic applications [25] enable enhanced retention of active agents within the superficial SC strata, as confirmed by various microscopy techniques. Various cosmetic products currently contain niosomes.

Synthetic surfactant-based *elastic vesicles*, introduced by Bouwstra's research group, were designed according to similar principles as phospholipid-based transfersomes [26]. The most frequently reported "edge activator" in the formulation of these vesicles is octaoxyethylene-laurate ester (PEG-8-L) [27]. The molar concentration of PEG-8-L directly affects the bilayer elasticity. Elastic vesicles were investigated for enhanced percutaneous transport of a number of drugs such as pergolide, rotigotine, estradiol, immune modulators, and ketorolac. As of today, no clinical studies with these systems have been reported.

Microemulsions

Microemulsions typically contain two phases (aqueous and organic) and surfactant/co-surfactant and, due to a very low surface tension between the components, form spontaneously [28, 29]. Cyclodextrins were also used as a component in microemulsions. In general, the type of emulsion (w/o vs. o/w emulsion), the droplet size, the emollient, the emulsifier, as well as the surfactant organization (micelles, lyotropic liquid crystals) in the emulsion may affect the cutaneous and percutaneous absorption. This system was used mainly for solubilization of lipophilic agents for skin application, thus increasing the concentration gradient and enabling skin deposition of higher concentrations of therapeutics. In some cases, this can allow for mild increase in percutaneous penetration as reported in several in vitro and in vivo studies.

Solid Nano- and Microparticles

This category of drug vectors includes such nanovectors as solid lipid nanoparticles (SLN), solgel particles, polymeric nano- and microparticles, metal- (gold, silver, iron) and metal oxide (titanium oxide, zinc oxide, silica)-based particles, and microsponges [30]. As it was mentioned above, efficient penetration of intact skin by nondeformable particles (organic or inorganic nanoparticles, conventional liposomes, micelles, etc.) is both elusive and unlikely. Best evidence for this comes from skin penetration studies with incompressible inorganic nanoparticles which lack any direct ability to affect the cutaneous barrier. Various sizes of inorganic and metallic particles were tested for skin delivery (30–200 nm), and all were found confined to the skin surface. Consistent with the notion, it is a general agreement in the field that solid particles with diameters of >30 nm accumulate on the skin surface and cannot transport across the skin, except through hair follicles. On the other hand, when the skin barrier is breached, as in the case of psoriasis, solid particulates can be of benefit, providing controlled release of substances in the skin [31].

Solid lipid nanoparticles (*SLN*), initially designed to enable extended release of intravenous therapeutics, are relatively large on the nanoscale (<500 nm) dispersions of melted and solidified lipids [32]. Among the common SLN constituents are triglycerides, partial glycerides, fatty acids, steroids, and waxes which are formulated into solid particles by methods such as high-pressure homogenization and microemulsion formation. Diffusion of the active ingredients through the solid matrix is a slow process that enables prolonged release of the molecule on the superficial skin layers. Due to the low concentrations of the free active molecules present on the surface at each time point, the diffusion gradient and thus the driving force for passive diffusion are considerably low than when the actives are applied in liquid formulation. This prevents transdermal absorption of an active ingredient making SLN, and other solid particles, attractive carriers for substances that have to act on the skin surface but possess intrinsic physicochemical characteristics to be absorbed to the viable strata and from there to systemic circulation [13]. Sunscreens, perfumes, and insect repellents belong to this category. Being solid and opaque, SLN also possess an inherent ability to act as a physical sunscreen.

Polymeric carriers are another category of the solid nano- and micro-vectors for topical delivery [33, 34]. There are a plenty of structures and building blocks that can be used for the design of polymeric vectors. The polymers currently investigated and used for preparations of particles in biomedical applications range from natural polymers, such as chitosan and tyrosine derivatives and biocompatible and biodegradable poly(lactic acid) (PLA), poly(glycolic acid) (PGA), and their copolymers (PLGA) to other representatives in the class, such as polymethyl methacrylate (PMMA), polycaprolactone, polyalkylcyanoacrylates, polyvinyl acetate (PVA), and polystyrene [9]. Skin application of polymeric nano- and microscale systems represents a less explored venue comparing to the knowledge of their physicochemical properties and design variables accumulated in the drug delivery field. The first patent on polymeric particles related to skin delivery by Nuwayser was approved soon after the appearance of the first transdermal systems. The inventor proposed to use polymeric particles as a controlled release component in the formulation of transdermal patches. Since then, several, mostly lipophilic drugs were incorporated into the polymeric particles applied on the skin, including steroids, minoxidil, and sunscreens such as octyl methoxycinnamate and benzophenone-3. Studies have shown that based on their shape and structure, polymeric particles can either accumulate in hair shafts enabling intrafollicular depot or retard skin permeation of active ingredients. A *microsponge delivery system* (MDS) [11] is based on porous nondegradable polymeric microspheres, typically 10–25 μm in diameter. MDS can entrap various active ingredients and then release them onto the skin over a time and in response to trigger (e.g., rubbing, temperature, pH change), while remaining on the skin surface. The technology is being used in cosmetics,

over-the-counter skin care, sunscreens, and prescription products. MDS was formulated with benzoyl peroxide (acne medication), hydroquinone (bleaching cosmetic agent), fluconazole (antifungal), and retinoids (acne).

Solgel systems, introduced by Avnir in 1984, are highly porous, reactive organosilicates formed based on the hydrolytic polycondensation of silicon alkoxides with a general formula $Si(OR)_4$or R_0 $nSi(OR)_4$Àn where R_0 contains vinyl, methacryl, or epoxy groups. A flexible network of inorganic oxide or hybrid nanocomposites can be produced through reacting R_0 with additional groups. As a result, the characteristics of solgel nanocomposite are halfway between those of polymers and silica-based glasses. Solgel-based systems were successfully clinically tested (phase II) for treatment of rosacea and acne and are used in sunscreen formulations [35].

Microinvasive Techniques for Enhanced Skin Absorption

Several microinvasive techniques are under investigation for transdermal delivery applications. Principles of ballistic high-velocity needle-free particles, microneedles, and skin microporation methods are described below.

Ballistic Needle-Free High-Velocity Particle System

Ballistic needle-free high-velocity particle system or PowderJect system was invented more than half a century ago for military use. Pulsed microjet injectors could be used to deliver drugs for local as well as systemic applications without using needles. The technology is based on transporting solid nanoparticles or liquid nanodroplets across the SC layer into the viable tissues, driven by a supersonic shock wave of helium. These systems are currently in use for delivery of proteins and vaccine. Jet-based noninvasive vaccination [36] involves shooting suitably coated gold particles into the skin. The injectors in clinical use contain compressed gas that can discharge the contents across the skin barrier. Among the claimed advantages for these delivery systems are freedom from pain and needle phobia, improved efficacy and bioavailability, controlled release, accurate dosing, and safety. However, the reported problems with bruising and particles bouncing off skin surfaces are the main reasons for the PowderJect uncommon use. Another disadvantage is the use of extremely small volumes, lowering the application to very potent therapeutics. Nanoliter-volume pulsed microjets comprised of 10–15 nL droplets were created to minimize the pain associated with this microinvasive transdermal drug delivery. To control the depth of jet penetration into the dermis, thus minimizing the pain, parameters such as velocity and volume of the microinjections can be tuned: a high velocity (>100 m/s) of microjets allows their entry into the skin, whereas the extremely small volumes (from 2 nL) limit the penetration depth. Due to the local adverse effects, the technology is generally proposed for acute usage, as in the case of skin immunization. Among currently used ballistic devices are Antares (insulin delivery), Mini-Jet, Biojector, Inject, and Penjet.

Microneedles (MN)

Microneedles are microinvasive tools to overcome the skin barrier and are currently one of the most vibrant domains of transdermal delivery research by active means [2, 26, 37, 38].

MN were proposed by Prausnitz and coworkers in 1998 for transdermal delivery of high molecular weight bioactives [2]. MN consist of an array of microprojections, ranging from 20 to 2,000 μm in dimensions, and smaller arrays are being developed. Application of MN to the SC surface creates transport pathways of micron dimensions that allow for transport of macromolecules and aggregates such as supramolecular complexes and even nanoparticles. A variety of techniques and designs of MN have been proposed. MN can be produced by microfabrication and laser-based micromolding; can be saw diced and/or etched; can be biodegradable (polymer or nanoporous semiconductors based) or non-dissolving (metallic, silica), solid, or hollow; and can be designed for sampling or drug/vaccine delivery. The design of MN can involve various geometries, which have a direct impact on the diameter and the depth of the pores produced in the skin. As an example, MN can be blunt or possess a very high aspect ratio, can incorporate

the active ingredient into the hollow structure, or can be coated with the substance to deliver. Another key parameter in the design of efficient MN-based delivery system is the application and driving pressures. Generally, due to the skin elasticity, the average diameter of the transepidermal pores left in the skin after microneedle extraction equals to roughly only half of the needle diameter at the base. Microneedles were proven to be efficient in transcutaneous immunization and delivery of opiates and proteins; the long-term safety still remains an open question. There have been a number of clinical studies, which involve mainly metallic MN. Safety of MN especially for chronic administration of therapeutics still holds a question mark. For the last decade, the technology has been under development by several industrial companies mainly for the purpose of vaccination, including Zosano Pharma (Macroflux), 3M Corporation (3 M hollow microstructured transdermal system, hMTS), Corium International Inc. (MicroCor), NanoPass (MicronJet), and BD Technologies (Soluvia™ prefillable microinjection system).

Radiofrequency or Electroporation

Generally, there is less data in the literature related to the methods of *upper skin layer removal* for enhanced percutaneous absorption. Only the main principles of the techniques are given below. *Radiofrequency or electroporation*-based removal of the superficial skin strata is based on microelectronic devices which use high-voltage currents [39]. Further, *lasers* with different wavelengths and types, including the erbium:yttrium-gallium-garnet laser, the neodymium-doped yttrium-aluminum-garnet laser, and the CO2 laser, are studied to increase skin transport of the drugs [2, 40]. This means to create a single or multiple (an array) microchannels across the stratum corneum deep into the viable epidermis. The efficacy of transport depends on the electrical parameters and the physicochemical properties of drugs. When applied in vivo, high-voltage pulses are reported to be well tolerated, inducing only muscle contractions; however, since the techniques may involve viable tissue, the safety for acute and chronic administration of actives yet remains the major question. A few studies reported in the literature show improved skin permeation of ionic molecules such as granisetron hydrochloride and diclofenac sodium, using the above techniques. Enhanced delivery of proteins has also been reported.

Summary

Extensive research during the past few decades has focused on various strategies utilizing nano- and microscaled delivery systems to enhance skin transport of dermal and systemic therapeutics or to retard cutaneous absorption of agents that should primarily act on the skin surface. This entry focused on suitable techniques for delivery of therapeutics either from the skin surface to the deep viable layers or to the systemic circulation. While currently available methods for administration of therapeutics onto the skin span over a wide range of micro- and nanotechnologies, new methods are being introduced in this dynamic research area. In many cases, the safety of chronic or acute use on the skin will determine the fate of the proposed method.

Cross-References

- ▶ Liposomes
- ▶ Nanoparticles
- ▶ Radiofrequency
- ▶ Solid Lipid Nanoparticles (SLN)

References

1. Scheindlin, S.: Transdermal drug delivery: PAST, PRESENT, FUTURE. Mol. Interv. **4**, 308–312 (2004)
2. Prausnitz, M.R., Langer, R.: Transdermal drug delivery. Nat. Biotechnol. **26**, 1261–1268 (2008)
3. Wertz, P.W.: Current understanding of skin biology pertinent to skin penetration: skin biochemistry. Skin Pharmacol. Physiol. **26**, 217–226 (2013)
4. Elias, P.M., Ferngold, K.R. (eds.): Skin Barrier. Taylor & Francis Group, New York (2006)
5. van Smeden, J., Janssens, M., Gooris, G.S., Bouwstra, J.A.: The important role of stratum corneum lipids for the cutaneous barrier function. Biochim. Biophys. Acta (BBA) Mol. Cell Biol. Lipids. **1841**, 295–313 (2014)

6. Lowther, J., Naismith, J.H., Dunn, T.M., Campopiano, D.J.: Structural, mechanistic and regulatory studies of serine palmitoyltransferase. Biochem. Soc. Trans. **40**, 547–554 (2012)
7. Hadgraft, J., Guy, R. H.: Feasibility assessment in topical and transdermal delivery: Mathematical models and in vitro studies. In: Guy, R. H., Hadgraft, J., (eds.) Transdermal Drug Delivery, CRC Press, Taylor and Francis Group an Informa business, Boca Raton, FL, USA. 1–23 (2002)
8. Godin, B., Touitou, E.: Transdermal skin delivery: predictions for humans from in vivo, ex vivo and animal models. Adv. Drug Deliv. Rev. **59**, 1152–1161 (2007)
9. Godin, B., Touitou, E.: Nanoparticles aiming at specific targets – dermal and transdermal delivery. In: Domb, A.J., Tabata, Y., Ravi Kumar, M.N.V., Farber, S. (eds.) Nanoparticles for Pharmaceutical Applications. American Scientific Publishers, Valencia (2007)
10. Touitou, E., Godin, B.: Vesicular carriers for enhanced delivery through the skin. In: Touitou, E., Barry, B.W. (eds.) Enhancement in Drug Delivery, pp. 255–278. CRC press, Boca Raton (2006)
11. Korting, H.C., and Schafer-Korting, M.: Carriers in the topical treatment of skin disease. Handb. Exp. Pharmacol. **197**, 435–468 (2010)
12. Godin, B., Touitou, E.: Nanoparticles aiming at specific targets – dermal and transdermal delivery. In: Domb, A.J., Tabata, Y., Ravi Kumar, M.N.V., Farber, S. (eds.) Nanoparticles for Pharmaceutical Applications, pp. 192–211. American Scientific Publishers, Valencia (2007)
13. Schafer-Korting, M., Mehnert, W., Korting, H.C.: Lipid nanoparticles for improved topical application of drugs for skin diseases. Adv. Drug Deliv. Rev. **59**, 427–443 (2007)
14. Touitou, E., Ainbinde, D.: Ethosomes – an innovative carrier for enhanced delivery into and across the skin: original research article: ethosomes – novel vesicular carriers for enhanced delivery: characterization skin penetration properties, 2000. J. Control. Release **190**, 44–46 (2014)
15. de Araujo, D.R., da Silva, D.C., Barbosa, R.M., Franz-Montan, M., Cereda, C.M., Padula, C., Santi, P., de Paula, E.: Strategies for delivering local anesthetics to the skin: focus on liposomes, solid lipid nanoparticles, hydrogels and patches. Expert Opin. Drug Deliv. **10**, 1551–1563 (2013)
16. Rahimpour, Y., Hamishehkar, H.: Liposomes in cosmeceutics. Expert Opin. Drug Deliv. **9**, 443–455 (2012)
17. El Maghraby, G.M., Barry, B.W., Williams, A.C.: Liposomes and skin: from drug delivery to model membranes. Eur. J. Pharm. Sci. **34**, 203–222 (2008)
18. Shumilov, M., Bercovich, R., Duchi, S., Ainbinder, D., Touitou, E.: Ibuprofen transdermal ethosomal gel: characterization and efficiency in animal models. J. Biomed. Nanotechnol. **6**, 569–576 (2010)
19. Honeywell-Nguyen, P.L., de Graaff, A.M., Groenink, H.W., Bouwstra, J.A.: The in vivo and in vitro interactions of elastic and rigid vesicles with human skin. Biochim. Biophys. Acta **1573**, 130–140 (2002)
20. Touitou, E., Dayan, N., Bergelson, L., Godin, B., Eliaz, M.: Ethosomes – novel vesicular carriers for enhanced delivery: characterization and skin penetration properties. J. Control. Release **65**, 403–418 (2000)
21. Ainbinder, D., Paolino, D., Fresta, M., Touitou, E.: Drug delivery applications with ethosomes. J. Biomed. Nanotechnol. **6**, 558–568 (2010)
22. Shumilov, M., Touitou, E.: Buspirone transdermal administration for menopausal syndromes, in vitro and in animal model studies. Int. J. Pharm. **387**, 26–33 (2010)
23. Cevc, G.: Rational design of new product candidates: the next generation of highly deformable bilayer vesicles for noninvasive, targeted therapy. J. Control. Release **160**, 135–146 (2012)
24. Cevc, G., Vierl, U.: Nanotechnology and the transdermal route: a state of the art review and critical appraisal. J. Control. Release **141**, 277–299 (2010)
25. Hougeir, F.G., Kircik, L.: A review of delivery systems in cosmetics. Dermatol. Ther. **25**, 234–237 (2012)
26. Windbergs, M., Hansen, S., Schroeter, A., Schaefer, U.F., Lehr, C.M., Bouwstra, J.: From the structure of the skin barrier and dermal formulations to in vitro transport models for skin absorption: skin research in the Netherlands and in Germany. Skin Pharmacol. Physiol. **26**, 317–330 (2013)
27. Loan Honeywell-Nguyen, P., Wouter Groenink, H.W., Bouwstra, J.A.: Elastic vesicles as a tool for dermal and transdermal delivery. J. Liposome. Res. **16**, 273–280 (2006)
28. Vadlamudi, H.C., Narendran, H., Nagaswaram, T., Yaga, G., Thanniru, J., Yalavarthi, P.R.: Microemulsions based transdermal drug delivery systems. Curr. Drug Discov. Technol. **11**, 169–180 (2014)
29. Lapteva, M., Kalia, Y.N.: Microstructured bicontinuous phase formulations: their characterization and application in dermal and transdermal drug delivery. Expert Opin. Drug Deliv. **10**, 1043–1059 (2013)
30. Prow, T.W., Grice, J.E., Lin, L.L., Faye, R., Butler, M., Becker, W., Wurm, E.M., Yoong, C., Robertson, T.A., Soyer, H.P., Roberts, M.S.: Nanoparticles and microparticles for skin drug delivery. Adv. Drug Deliv. Rev. **63**, 470–491 (2011)
31. Katare, O.P., Raza, K., Singh, B., Dogra, S.: Novel drug delivery systems in topical treatment of psoriasis: rigors and vigors. Indian. J. Dermatol. Venereol. Leprology. **76**, 612–621 (2010)
32. Pardeike, J., Hommoss, A., Muller, R.H.: Lipid nanoparticles (SLN, NLC) in cosmetic and pharmaceutical dermal products. Int. J. Pharm. **366**, 170–184 (2009)
33. Pohlmann, A.R., Fonseca, F.N., Paese, K., Detoni, C. B., Coradini, K., Beck, R.C., Guterres, S.S.: Poly (-caprolactone) microcapsules and nanocapsules in

drug delivery. Expert. Opin. Drug Deliv. **10**, 623–638 (2013)
34. Zhang, Z., Tsai, P.C., Ramezanli, T., Michniak-Kohn, B.B.: Polymeric nanoparticles-based topical delivery systems for the treatment of dermatological diseases. Wiley Interdiscip. Rev. Nanomedicine. Nanobiotechnology. **5**, 205–218 (2013)
35. Fireman, S., Toledano, O., Neimann, K., Loboda, N., Dayan, N.: A look at emerging delivery systems for topical drug products. Dermatol. Ther. **24**, 477–488 (2011)
36. Kis, E.E., Winter, G., Myschik, J.: Devices for intradermal vaccination. Vaccine **30**, 523–538 (2012)
37. van der Maaden, K., Jiskoot, W., Bouwstra, J.: Microneedle technologies for (trans)dermal drug and vaccine delivery. J. Control. Release **161**, 645–655 (2012)
38. Schoellhammer, C.M., Blankschtein, D., Langer, R.: Skin permeabilization for transdermal drug delivery: recent advances and future prospects. Expert Opin. Drug Deliv. **11**, 393–407 (2014)
39. Staples, M., Daniel, K., Cima, M.J., Langer, R.: Application of micro- and nano-electromechanical devices to drug delivery. Pharm. Res. **23**, 847–863 (2006)
40. Lin, C.H., Aljuffali, I.A., Fang, J.Y.: Lasers as an approach for promoting drug delivery via skin. Expert Opin. Drug Deliv. **11**, 599–614 (2014)

Detection of Nanoparticle Biomarkers in Blood

► Detection of Nanoparticulate Biomarkers in Blood

Detection of Nanoparticulate Biomarkers in Blood

Rajaram Krishnan[1] and Michael J. Heller[2,3]
[1]Biological Dynamics, Inc., University of California San Diego, San Diego, CA, USA
[2]Department of Nanoengineering, University of California San Diego, La Jolla, CA, USA
[3]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

Synonyms

Detection of nanoparticle biomarkers in blood

Definition

Detection of nanoparticulate biomarkers in blood is the ability to find nanoparticles in blood.

Overview

Challenges with Isolating and Detecting Nanoparticles in Blood

While the potential applications of nanotechnology in medicine are rapidly growing, a number of issues still need to be resolved before nanomedicine translates from the lab to the bedside [1]. Two important challenges will be the monitoring of drug delivery nanoparticles and the detection of cell-free-circulating (cfc) DNA nanoparticulate biomarkers [2, 3]. Presently, considerable research efforts are being carried out on the development of new drug delivery nanoparticle therapeutics [4]. For both research and clinical applications, it will be important to develop rapid, sensitive, and inexpensive monitoring techniques for determining levels of drug delivery nanoparticles directly into the patient's blood [2, 5, 6]. In another related area, the ability to detect cfc-DNA, cfc-RNA, and other cell-free-circulating nanoparticulate biomarkers directly into blood would represent a major advance for early cancer screening [7], residual disease detection [8], and chemotherapy monitoring [9].

Unfortunately, present methods for isolating nanoparticles and nanoparticulate biomarkers in the 10–500 nm range from blood are complex, expensive, and time consuming [5–9]. Adding even more challenge is that relatively large blood samples (1–10 ml) are needed when assaying for very low levels of nanoparticles or biomarkers. Thus, sample preparation is more often the weak link in monitoring and diagnostic assays than is the intrinsic sensitivity of the downstream detection technology. The sample preparation process for blood often involves centrifugation, filtration, washing, extractions, and a number of other steps before the analyte(s) can be identified [5–9]. The extended amount of time between blood drawing, cell separation, analyte extraction, and the final analysis also leads to

significant loss of sensitivity and selectivity and to the degradation of the analytes [7–9]. Thus, there is a critical need for a novel robust technology, which will allow a variety of nanoscale entities to be manipulated, isolated, and rapidly detected *directly* from whole blood and other biological samples. Meeting this challenge will be required before therapeutic monitoring can be carried out in cost-effective point-of-care (POC) settings.

Dielectrophoresis as an Isolation Method

Dielectrophoresis (DEP) is an intrinsically powerful technique for separating cells and nanoparticles [10, 11]. DEP is the induced motion of particles produced by the dielectric differences between the particles and media in an asymmetric alternating current (AC) electric field [12]. For spherical particles, the DEP force equation is given by:

$$F_{DEP} = 2\pi\varepsilon_m r^3 \mathrm{Re}[K(\omega)]\nabla E^2_{RMS} \qquad (1)$$

where

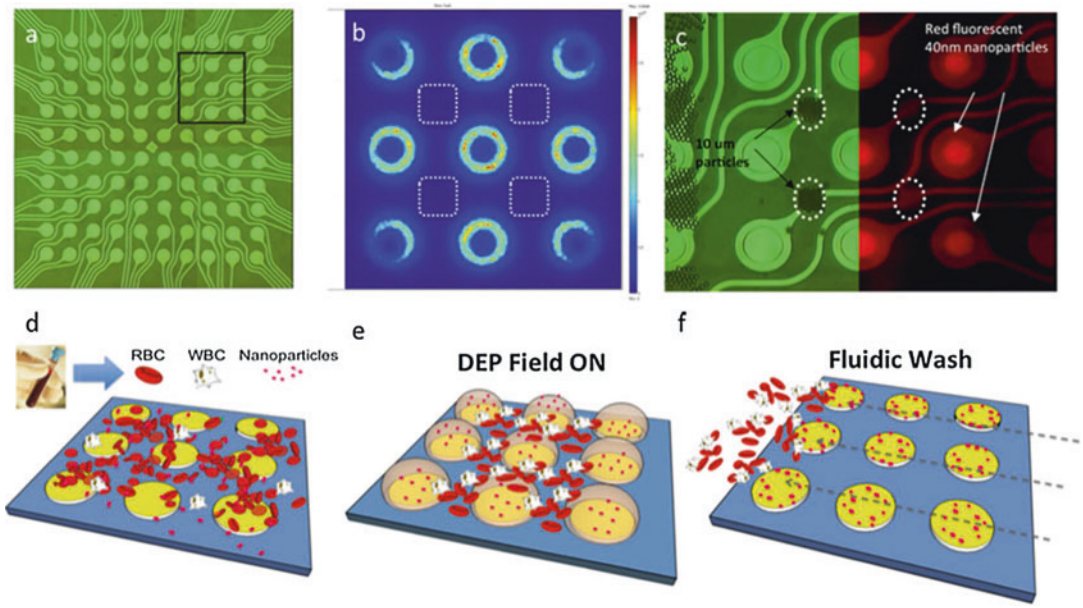
$$\mathrm{Re}[K(\omega)] = \mathrm{Re}\left(\frac{\varepsilon^*_p - \varepsilon^*_m}{\varepsilon^*_p + \varepsilon^*_m}\right) \qquad (2)$$

where ε^*_p and ε^*_m are the complex dielectric permittivities of the particle and medium respectively, defined by $\varepsilon^* = \varepsilon - j\sigma/\omega$, where $j^2 = -1$, ε is the dielectric constant and σ is the conductivity. If a particle has a positive Re[K(ω)], it will migrate to the high-field regions, and if it has a negative Re[K(ω)], it will migrate to the low-field regions. Unfortunately, the use of DEP for practical applications has been limited by the need to carry out the process under low conductance (ionic strength) conditions [13, 14]. Thus, blood or any other high ionic strength sample has to be significantly diluted before DEP separations can be carried out [14]. Recently, a high conductance (HC) DEP method that allows both nanoparticles and cfc-DNA nanoparticulates to be manipulated, isolated, and detected under high ionic strength conditions has been developed [15–17]. HC-DEP sets the stage for new "seamless" sample to answer systems, which will allow a variety of drug delivery nanoparticles and other nanoscale biomarkers to be rapidly isolated and analyzed from clinically relevant amounts of complex undiluted biological samples.

Basic Methodology

Microelectrode Array Device Used for Dielectrophoresis

In these studies, all HC-DEP experiments were carried out using an electronic microarray device with 100 circular platinum microelectrodes 80 μm in diameter (Fig. 1a). The microarray is overcoated with a 10-μm thick porous polyacrylamide hydrogel layer and enclosed in a microfluidic cartridge, which forms a 20 μL sample chamber. While only a 3 × 3 subset of microelectrodes was used for the DEP experiments, all 100 microelectrodes can be used if desired. AC electric fields were applied to the nine microelectrodes in a checkerboard-addressing pattern where each microelectrode has the opposite bias of its nearest neighbor. Figure 1b shows the asymmetric electric field distribution, which produces DEP positive (high-field) regions on the microelectrodes and the DEP negative (low-field) regions between the microelectrodes [15–17]. At AC frequencies in the 5,000–10,000 Hz range, nanoparticles, cfc-DNA, and other cellular nanoparticulates from about 10 to 500 nm in size will concentrate in the DEP high-field regions over the microelectrodes, while micron-size particles and cells will concentrate in the low-field regions. Figure 1c shows a composite half bright field/half fluorescence image for a prior DEP separation of 10 μm polystyrene particles and 40 nm red fluorescent nanoparticles carried out in 1× TBE buffer. The 10 μm particles are isolated into the low-field regions, and 40 nm red fluorescent nanoparticles are concentrated into the high-field regions. Detailed procedures for high conductance (HC) DEP separations were described earlier [15–17]. The general scheme for the HC-DEP separation of nanoparticles from whole blood (red and white cells) is shown in Fig. 1d–f. When the DEP field is applied, the nanoparticles (red dots) concentrate into the high-field regions where they



Detection of Nanoparticulate Biomarkers in Blood, Fig. 1 Microarray device and flow diagram for the separation of nanoparticles from blood. (**a**) Shows the electronic microarray device with 100 circular platinum microelectrodes 80 μm in diameter. (**b**) Alternating current (*AC*) electric fields were applied to a subset of nine microelectrodes in a checkerboard-addressing pattern. The electric field model for this geometry shows the DEP field maxima or high-field regions exist on the microelectrodes and the DEP field minima or low-field regions exist in the areas between the electrodes (*white dotted square*). (**c**) Shows a composite *half bright field*/*half red* fluorescence image for a prior DEP separation where 10 μm particles are isolated into the low-field regions and red fluorescent 40 nm nanoparticles are concentrated into the high-field regions (**d**) Shows the diagram of the microarray with *red* and *white* blood cells and *red* fluorescent nanoparticles before the DEP field is applied. (**e**) Shows the microarray after the DEP field is applied, with *red* fluorescent nanoparticles moving to the microelectrodes and the larger *red* and *white* blood cells moving between the microelectrodes. (**f**) Shows fluidic wash removing the blood cells, while the nanoparticles remain in the high-field regions

are held firmly on the microelectrodes, and blood cells move into the low-field regions between the microelectrodes where they are held less firmly (Fig. 1e). A simple fluidic wash easily removes the blood cells while the nanoparticles remain in the high-field regions (Fig. 1f). The highly concentrated fluorescent nanoparticles can now be easily detected.

Preparation of Buffy Coat Blood and Whole Blood Experiments

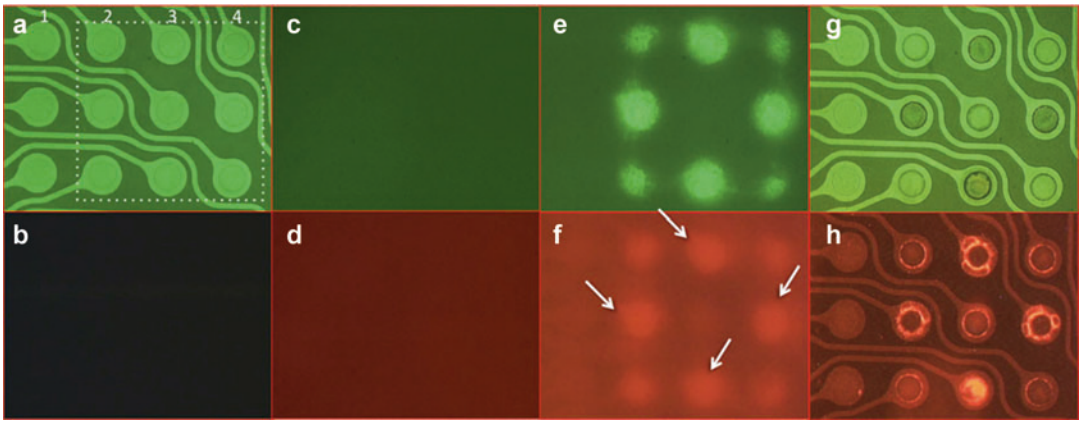
For the HC-DEP experiments, concentrated 5× Tris Borate EDTA (TBE) buffer solution was obtained from USB Corporation (Cleveland, Ohio, USA) and diluted to 1× concentration and Dulbecco's phosphate buffer saline (1× PBS) solution was obtained from Invitrogen (Carlsbad, CA, USA) and diluted to 0.5× concentration. Human buffy coat blood was obtained from San Diego Blood Bank (San Diego, CA). Buffy coat blood contains ~1/10 the number of red blood cells as whole blood. Whole blood samples were obtained from adult female Sprague Dawley rats using proper protocols. Conductivity measurements were made with an Accumet Research AR-50 Conductivity meter using 2 cell (range: 10–2,000 μS) and 4 cell (range: 1–200 mS) electrodes. Buffer conductivities were: 1× TBE: 1.09 mS/cm; 0.5× PBS: 7.6 mS/cm; buffy coat blood: 8.6 mS/cm, whole rat blood: 5.2 mS/cm (nanoparticle experiments). Fluorescent polystyrene nanoparticles (FluoSpheres) with NeutrAvidin were purchased from Invitrogen (Carlsbad, CA). The nanoparticles were 0.04 μm (40 nm) in diameter and red fluorescent (e × 585/em605). For human buffy coat blood

experiments and rat blood experiments, 10 μL of the 40 nm red fluorescent nanoparticles was added from the stock solution to 300 μL of either buffy coat blood or whole rat blood. For comparison purposes, the number of nanoparticles in an Abraxane dosage [18] was calculated by using the recommended dosage of 260 mg/m^2 combined with an average estimated body surface area of 1.92 m^2 leading to a value of approximately 0.5 g of Abraxane per dosage. Since the mean diameter of an Abraxane nanoparticle is 130 nm, and estimating a density of ~1 g/mL, the number of circulating nanoparticles was calculated to be approximately 7.5×10^{10} nanoparticles per ml of blood (assuming 6 L of blood). The red fluorescent nanoparticles were thus serially diluted to the following amounts: 9.5×10^9, 9.5×10^{10}, and 9.5×10^{11} per mL in human buffy coat blood. The microarrays were controlled using a custom-made switching system that allows individual control over the voltage applied to each microelectrode. The microelectrodes were set to proper AC frequency and voltages using an Agilent 33120A Arbitrary Function Generator. AC frequencies ranged from 1,000 to 10,000 Hz, at 20 V peak to peak (pk-pk). The DEP field was generally applied for about 10–20 min. The DEP separations were visualized using a JenaLumar epifluorescent microscope (red fluorescence nanoparticles E × 585 nm, Em 605 nm). Both bright field and red fluorescent images were captured using an Optronics 24-bit RGB CCD camera. The image data was processed using a Canopus ADVC-55 video capture card connected to a laptop computer with Adobe Premiere Pro and Windows Movie Maker. The final fluorescence intensity data images were created by inputting fluorescent image frames of the video into MATLAB.

Key Research Findings

The initial HC-DEP experiment demonstrates the separation and detection of 40 nm red fluorescent nanoparticles in human buffy coat blood (0.86 S/m). Buffy coat blood was used in this experiment because it allows the complete DEP separation process to be observed. The DEP experiment was carried out at 10,000 Hz and 10 V pk-pk. Figure 2a, b shows the microarray in bright field and red fluorescence before the buffy coat blood sample was applied. Twelve microelectrodes are seen in the bright field image (Fig. 2a), and the red fluorescence image is dark (Fig. 2b). *All images show 12 microelectrodes, with the three control microelectrodes in column 1 not activated, and nine microelectrodes in columns 2, 3, and 4, which are activated*. Figure 2c, d shows images of the microarray in bright field and red fluorescence after 20 μL of buffy coat blood containing 300 ng/μL of 40 nm red fluorescent nanoparticles was added. The microelectrodes are not visible in the bright field image because of the high cell density (Fig. 2c), but a red fluorescent background is now seen in the fluorescent image (Fig. 2d). Figure 2e, f shows the microarray after the DEP field has been applied for 12 min to the nine microelectrodes. The bright field image shows the microelectrodes becoming more visible as the blood cells move into the low-field regions (Fig. 2e), and the fluorescent image shows the red fluorescent nanoparticles concentrating in the high-field regions around the microelectrodes (Fig. 2f). Finally, Fig. 2g, h shows the microarray after washing with 0.5 × PBS buffer to remove the blood cells. The bright field image shows the microarray is clear of cells (Fig. 2g), while the red fluorescent image shows intense fluorescence from the concentrated nanoparticles (Fig. 2h). The overall time from sample application to detection was less than 30 min.

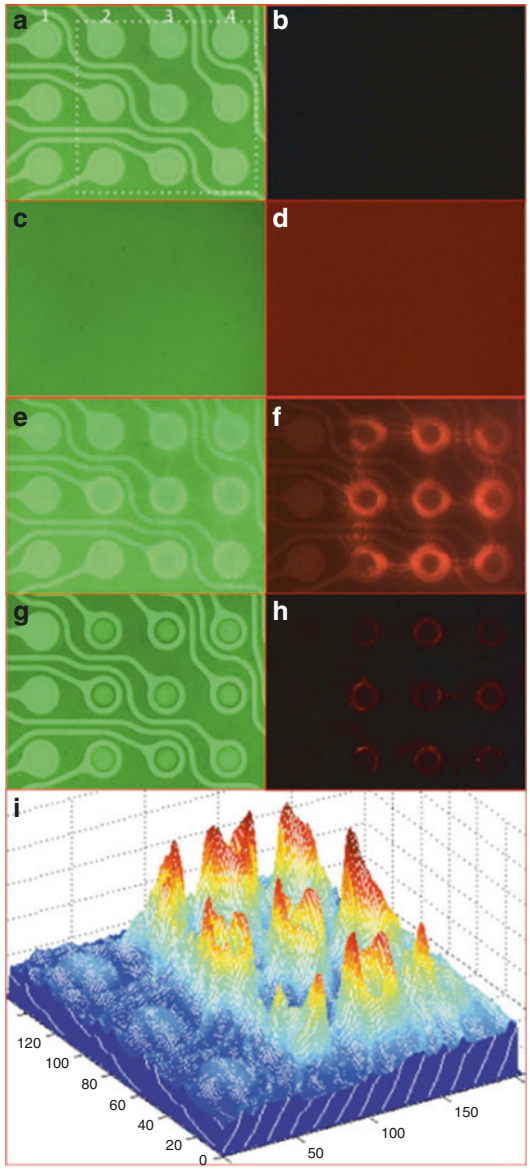
The second HC-DEP experiment demonstrates the separation and detection of 40 nm red fluorescent nanoparticles in undiluted whole rat blood (0.52 S/m). This experiment was carried out at 10,000 Hz and 10 V pk-pk. Figure 3a, b shows the microarray before the whole blood sample was added. Figure 3c, d shows images of the microarray after 20 μL of whole blood containing 300 ng/μL of 40 nm red fluorescent nanoparticles was added. The DEP field was then applied for 11 min. Because of the very high cell density of



Detection of Nanoparticulate Biomarkers in Blood, Fig. 2 Separation and detection of 40 nm *red* fluorescent nanoparticles in buffy coat blood. (**a**) Shows the microarray in bright field and *red* fluorescence before the buffy coat blood sample was applied to the microarray. (**b**) Same as (**a**) in *red* fluorescence. (**c**) Shows the microarray in bright field after 20 μL of buffy coat blood containing 300 ng/μl of 40 nm *red* fluorescent nanoparticles were added to the microarray. (**d**) Same as (**c**) in *red* fluorescence. (**e**) Shows the microarray after the DEP field has been applied to a set of nine electrodes (columns 2, 3, and 4), while the three electrodes in the column 1 remained un-activated. The blood cells move away from the microelectrodes. (**f**) Same as (**e**) in *red* fluorescence where the *red* fluorescent nanoparticles can be seen concentrating on the microelectrodes. (**g**) Bright field image shows the microarray after it was washed with to remove the buffy coat blood cells (**h**) *Red* fluorescent image now shows the fluorescent nanoparticles concentrated on the microelectrodes

whole blood, the movement of cells into the low-field regions and concentration of fluorescent nanoparticles into the high-field regions was not observable (while blood cells are present). After one fluidic wash with 0.5× PBS, the bright field image shows the microelectrodes with only a few cells present (Fig. 3e), and the red fluorescent image shows intense fluorescence from the nanoparticles concentrated around the microelectrodes (Fig. 3f). The microarray was then washed five times with 0.5× PBS, and the final images show no cells present (Fig. 3g), with significant red fluorescence remaining on the microelectrodes (Fig. 3h). Figure 3i is a 3D fluorescent intensity image (after the first wash) showing the relative levels of fluorescence on the activated microelectrodes versus the un-activated microelectrodes. The overall time from sample application to detection was less than 30 min.

HC-DEP experiments were now carried out in order to determine if nanoparticles could be detected in blood at the basic dosage range used for present drug delivery nanoparticles. By way of example, the dosage for Abraxane drug delivery nanoparticles is approximately 7.5×10^{10} particles/mL blood [18]. Red fluorescent 40 nm nanoparticles at 9.5×10^{9}, 9.5×10^{10}, and 9.5×10^{11} particles/mL were made up in buffy coat blood, and HC-DEP was carried out at 10,000 Hz and 10 V pk-pk for 20 min. Figure 4a–c shows the bright field image for the three different concentrations of nanoparticles, where the blood cells can be seen moving away from the activated microelectrodes into the low-field regions. Figure 4d–f now shows the red fluorescent images of the three different concentrations of nanoparticles after the microarray was washed with 0.5× PBS. Figure 4d shows the nanoparticles at the lowest concentration (9.5×10^{9} particles/mL), with the enlargement of the microelectrode clearly showing fluorescence from the concentrated nanoparticles. Figure 4g–i shows the 3D fluorescent intensity images showing the relative levels of fluorescence on the activated microelectrodes versus the un-activated microelectrodes. These results demonstrate the intrinsic ability of HC-DEP to detect nanoparticles in blood at dosage levels now used for drug delivery nanoparticles.



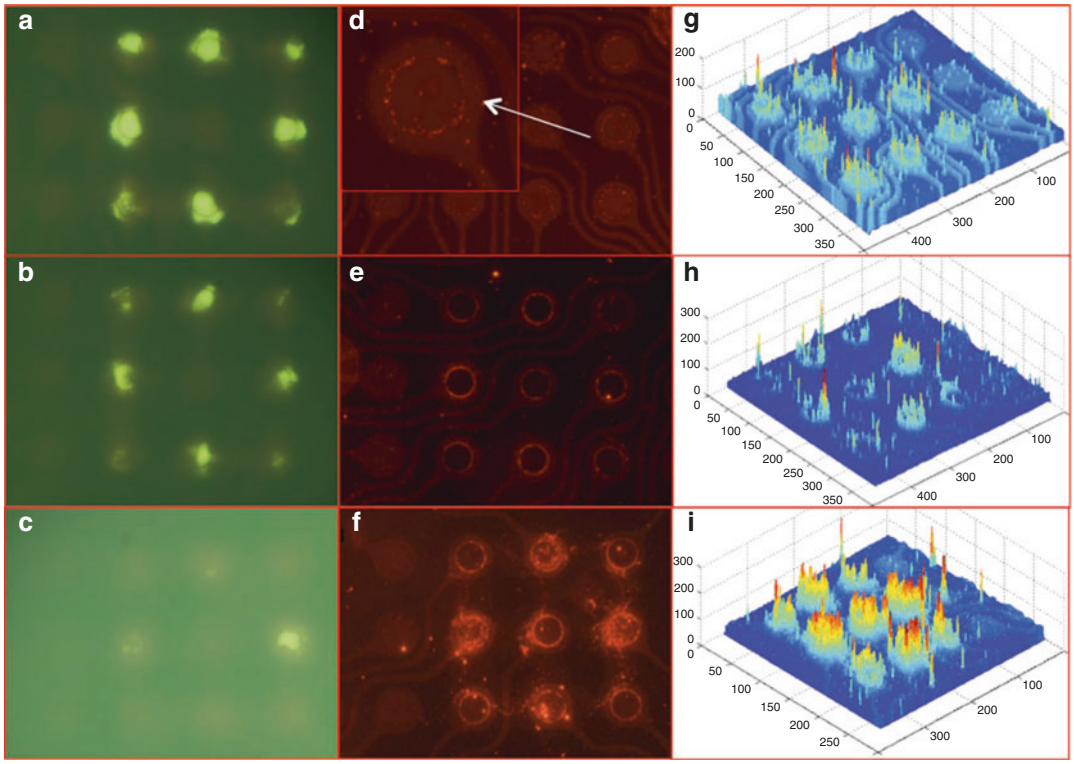
Detection of Nanoparticulate Biomarkers in Blood, Fig. 3 Separation and detection of 40 nm *red* fluorescent nanoparticles in whole blood. (**a**) Bright field image before the whole blood sample was added. (**b**) Same as (**a**) in *red* fluorescence. (**c**) Bright field image after 20 μL of whole blood containing 300 ng/μL of 40 nm *red* fluorescent nanoparticles was added. (**d**) Same as (**c**) in *red* fluorescence. (**e**) Bright field image after DEP field was applied to a set of nine microelectrodes (columns 2, 3, and 4), and the blood was removed. The bright field image shows the underlying microelectrodes with some blood cells scattered across the surface. (**f**) Same as (**e**) where *red* fluorescent image shows the fluorescent nanoparticles concentrated over the microelectrode structures (columns 2, 3,

Further experiments were carried out to determine the detection levels for the red fluorescent 40 nm nanoparticles in 1× TBE using dilution series of nanoparticles, which ranged from 2.8×10^8 to 2.8×10^{10} particles/mL. HC-DEP was carried out at 10,000 Hz and 10 V pk-pk for 15 min. The results are presented in Fig. 5a (graph) and show that the nanoparticles can be detected down to the 2.8×10^8 particles/mL level. Because only a 1 μL volume of solution is actually being affected by the DEP field (on nine microelectrodes), the true number of nanoparticles being detected on each microelectrode is $<3 \times 10^5$ particles per microelectrode. A final experiment was now carried out to verify by scanning electron microscopy (SEM) that the 40 nm nanoparticles are truly being deposited into the high-field regions on the microelectrodes. For these experiments, a microelectrode array without a hydrogel layer was used. Figure 5d shows a red fluorescent image of one of the microelectrodes before the DEP field is applied. Figure 5e shows an image of the microelectrode after the DEP field was applied at 10,000 Hz and 10 V pk-pk for 4 min. Red fluorescence can be clearly seen concentrated around and on the microelectrode. Finally, Fig. 5f shows the SEM image of the same microelectrode with clusters of the 40 nm nanoparticles clearly visible. In previous work, similar SEM results for HC-DEP using 200 nm nanoparticles were shown [17].

Future Directions for Research

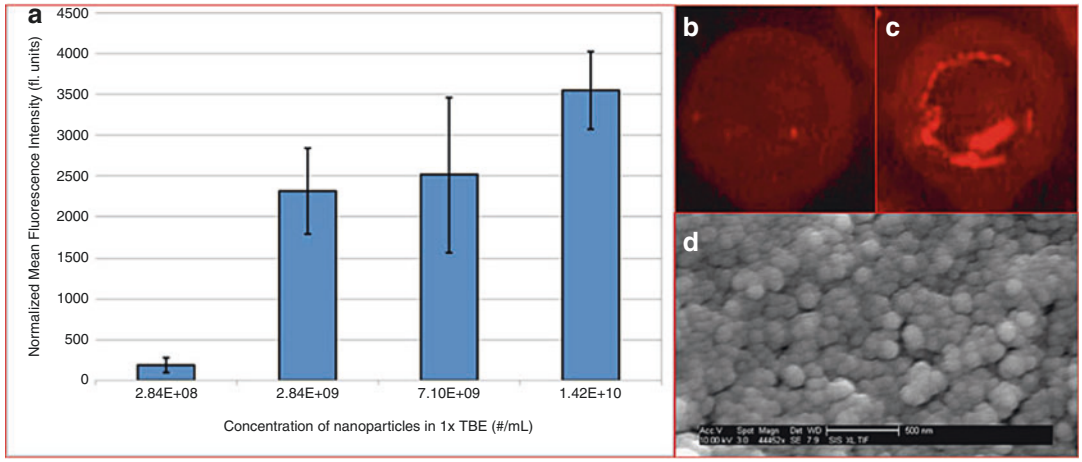
The ability to rapidly detect and monitor the concentration of drug delivery nanoparticles will be important for future nanomedicine applications. HC-DEP has now been shown to be able to rapidly

Detection of Nanoparticulate Biomarkers in Blood, Fig. 3 (continued) and 4). (**g**) Bright field image after washing five times with 0.5× PBS – no cells remain. (**h**) Same as (**g**) – Red fluorescence from nanoparticles still remains on the microelectrodes. (**i**) 3D fluorescent intensity image after the first 0.5× PBS buffer wash showing the relative levels of fluorescence on all nine microelectrodes (produced using MATLAB)



Detection of Nanoparticulate Biomarkers in Blood, Fig. 4 HC-DEP was carried out on samples of red fluorescent 40 nm nanoparticles at 9.5×10^9, 9.5×10^{10}, and 9.5×10^{11} particles/mL inserted in buffy coat blood. Blood cells are seen moving away from the activated microelectrodes. Figures (**a**) Bright field image for 9.5×10^9 nanoparticles/mL concentration, (**b**) Bright field image for 9.5×10^{10} nanoparticles/mL concentration, (**c**) Bright field image for 9.5×10^9 nanoparticles/mL concentration. (**d**) *Red* Fluorescent Image of (**a**) after the microarray was washed with 0.5× PBS. (**e**) *Red* Fluorescent Image of (**b**) after the microarray was washed with 0.5× PBS. (**f**) *Red* Fluorescent Image of (**c**) after the microarray was washed with 0.5× PBS. (**g**) 3D images showing the relative fluorescence intensity levels on the activated microelectrodes versus the un-activated microelectrodes for (**d**). (**h**) 3D images showing the relative fluorescence intensity levels for (**e**). (**i**) 3D images showing the relative fluorescence intensity levels for (**f**)

isolate and detect 40 nm nanoparticles from whole blood and buffy coat blood at clinically relevant levels, $\sim 9 \times 10^9$ particles/mL. Further studies in buffer show that even lower detection limits can easily be achieved, 2.8×10^8 particles/mL. The fact that the isolation of nanoparticles can be carried out directly in whole blood, with no sample processing, means that HC-DEP holds considerable promise for being a viable and cost-effective method for point-of-care applications. At the AC frequencies being used (5,000–10,000 kHz), HC-DEP provides a powerful nanoscopic tool that will allow a variety of important nanoscale entities in the 10–500 nm range to be manipulated and isolated relative to larger cells (red and white blood cells) and numerous smaller proteins and metabolic biomolecules. In addition to having high conductance (0.5–0.9 S/m) [19], whole blood is also one of the most complex of biological samples. Other important nanoscale entities that could be isolated using HC-DEP include: nanoparticle imaging agents, virus, prions, chylomicrons, large antibody and immunoglobulin complexes, and other cellular nanoparticulate biomarkers (nuclei, mitochondria, ribosomes, lysosomes, and various storage vacuoles). It should also be kept in mind that at higher AC frequencies (>10,000 Hz), DEP has the intrinsic ability for the higher resolution separation of different nanoparticles based on their AC crossover frequency and dielectric



Detection of Nanoparticulate Biomarkers in Blood, Fig. 5 Detection levels for the *red* fluorescent 40 nm nanoparticles in 1× TBE buffer. (**a**) Graph shows results after HC-DEP for a dilution series of nanoparticles, ranging from 2.8×10^8 to 2.8×10^{10} particles/mL. Nanoparticles could be detected down to the 2.8×10^8 particles/mL level. (b) *Red* fluorescent image before DEP field is applied. (**c**) *Red* Fluorescent image after the DEP field is applied. (**d**) SEM image of the microelectrode with clusters of the 40 nm nanoparticles clearly visible

properties [20]. A final advantage for HC-DEP is the fact that the entities being isolated do not have to be labeled. The ability to post-label a specific biomarker or analyte will prove to be a significant advantage for future drug monitoring and diagnostic applications. Thus, numerous molecular biological detection techniques including PCR, immunochemistry, specific fluorescent dyes, and in situ hybridization can be easily carried out in the same sample chamber. Overall, the results of this study have shown that HC-DEP has enormous potential for a number of new "seamless" sample to answer nanomedicine applications.

Cross-References

- ▶ Dielectrophoresis
- ▶ Dielectrophoresis of Nucleic Acids

References

1. Editorial: Healthy challenges. Nat. Nanotechnol. **2**, 451 (2007)
2. Nishiyama, N.: Nanomedicine: nanocarriers shape up for long life. Nat. Nanotechnol. **2**, 203–204 (2007)
3. Ferrari, M.: Cancer nanotechnology: opportunities and challenges. Nat. Rev. Cancer **5**, 161–171 (2005)
4. Mu, L., Feng, S.S.: A novel controlled release formulation for the anticancer drug paclitaxel (Taxol(R)): PLGA nanoparticles containing vitamin E TPGS. J. Control. Release **86**, 33–48 (2003)
5. Duncan, R.: The dawning era of polymer therapeutics. Nat. Rev. Drug Discov. **2**, 347–360 (2003)
6. Nishiyama, N., Kataoka, K.: Current state, achievements, and future prospects of polymeric micelles as nanocarriers for drug and gene delivery. Pharmacol. Ther. **112**, 630–648 (2006)
7. Sozzi, G., et al.: Quantification of free circulating DNA as a diagnostic marker in lung cancer. J. Clin. Oncol. **21**, 3902–3908 (2003)
8. Board, R.E., et al.: DNA methylation in circulating tumor DNA as a biomarker for cancer. Biomarker Insights **2**, 307–319 (2007)
9. Gautschi, O., et al.: Circulating deoxyribonucleic acid as prognostic marker in non-small-cell lung cancer patients undergoing chemotherapy. J. Clin. Oncol. **22**, 4157–4164 (2004)
10. Albrecht, D.R., Underhill, G.H., Wassermann, T.B., Sah, R.L., Bhatia, S.N.: Probing the role of multicellular organization in three-dimensional microenvironments. Nat. Methods **3**, 369–375 (2006)
11. Morgan, H., Hughes, M.P., Green, N.G.: Separation of submicron bioparticles by dielectrophoresis. Biophys. J. **77**, 516–525 (1999)
12. Ramos, A., Morgan, H., Green, N.G., Castellanos, A.: Ac electrokinetics: a review of forces in microelectrode structures. J. Phys. D Appl. Phys. **31**, 2338–2353 (1998)
13. Hughes, M.P.: Nanoparticle manipulation by electrostatic forces. In: Goodard, W.A., Brenner, D.W., Lyshevski, S.E. (eds.) Handbook of Nanoscience,

Engineering, and Technology, 2nd edn, pp. 16–32. CRC Press/Taylor and Francis Group, Boca Raton (2007)
14. Cheng, J., et al.: Preparation and hybridization analysis of DNA/RNA from E. coli on microfabricated bioelectronic chips. Nat. Biotechnol. **16**, 541–546 (1998)
15. Krishnan, R., Sullivan, B.D., Mifflin, R.L., Esener, S.C., Heller, M.J.: Alternating current electrokinetic separation and detection of DNA nanoparticles in high-conductance solutions. Electrophoresis **29**, 1765–1774 (2008)
16. Krishnan, R., Heller, M.J.: An AC electrokinetic method for enhanced detection of DNA nanoparticles. J. Biophotonics **2**, 253–261 (2009)
17. Krishnan, R., Dehlinger, D.A., Gemmen, G.J., Mifflin, R.L., Esener, S.C., Heller, M.J.: Interaction of nanoparticles at the DEP microelectrode interface under high conductance conditions. Electrochem. Commun. **11**, 1661–1666 (2009)
18. Green, M.R., et al.: Abraxane, a novel Cremophor-free, albumin-bound particle form of paclitaxel for the treatment of advanced non-small-cell lung cancer. Ann. Oncol. **17**, 1263–1268 (2006)
19. Hirsch, F.G., et al.: The electrical conductivity of blood: I. relationship to erythrocyte conductivity. Blood **5**, 1017–1035 (1950)
20. Green, N.G., Ramos, A., Morgan, H.: Ac electrokinetics: a survey of sub-micrometre particle dynamics. J. Phys. D Appl. Phys. **33**, 632–641 (2000)

Device Modeling

► Microscopic Modeling of Solid-State Quantum Devices

DFT-Based Molecular Dynamics

► Car–Parrinello Molecular Dynamics

DFT-Based Simulations of Electrified Interfaces

► Electrochemical Interfaces for Energy Storage and Conversion

Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering

Eberhard Burkel[1] and F. Zhang[2]
[1]Institute of Physics, University of Rostock, Rostock, Germany
[2]School of Materials Science and Engineering, Southeast University, Nanjing, China

Synonyms

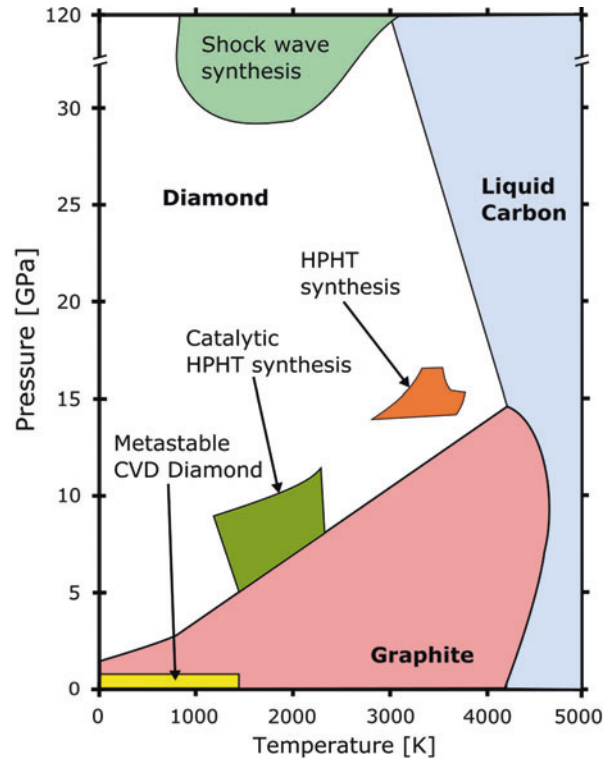
Diamond synthesis, Phase transformation of carbon to diamond, SPS of carbon allotropes

Definition

Nanodiamonds have excellent mechanical and optical properties, large surface areas and tunable surface structures. The nanodiamonds have very broad applications and, since they are nontoxic, also biomedical ones. So far, these artificial diamonds are mainly created by energy-consuming high-pressure and high-temperature methods. Recently, the spark plasma sintering technology proved to be an energy-saving method to generate nanometer and micrometer-sized diamond grains in short processing times.

Introduction

About six decades ago, the first artificial diamonds were synthesized. Since then, diamond particles and diamond films have been successfully obtained by various methods including high pressure and high temperature (HPHT) under static and shock pressure, combustion flames and chemical vapor deposition with radio frequency plasma and microwave plasma. For the synthesis of diamonds, the HPHT is still the most popular commercial method and it is applied to graphite, fullerenes, and carbon nanotubes. The achieved



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering,
Fig. 1 Pressure-temperature carbon phase diagram (Redrawn from [1] with permission of John Wiley and Sons ©)

materials show properties close to the ones of natural diamonds. Excellent surface coatings are achieved by ultrafine diamond grains with sizes of 5 nm–10 nm synthesized by explosives. Figure 1 shows the pressure-temperature phase diagram of carbon [1]. Generally, the HPHT synthesis of diamonds is applied at pressures above 5.0 GPa and at temperatures above 1300 °C.

However, C60 fullerenes are also compressed under non-hydrostatic pressures up to 30 GPa at room temperatures and transformed to small diamond crystallites at 20 GPa and 2000 °C using a multi-anvil cell achieving nanocrystalline cubic diamond with crystal sizes of 5 nm–10 nm [2]. In addition, the incorporation of solvent catalysts such as Ni, Co, Fe, and other transition metals and their alloys is a crucial point as aid for the diamond synthesis in the HPHT method.

About one decade ago, it was observed that, during spark plasma sintering (SPS) already at pressures of 80 MPa and temperatures of 1500 °C, carbon nanotubes became unstable and transformed to diamonds without any catalysts being involved [3]. Spark plasma sintering, also known as field-assisted sintering (FAST) or pulsed electric current sintering (PECS), is an electric field-assisted sintering procedure utilizing pulsed direct currents [4]. During this process, the necessary energy required in the diamond transition is delivered. The studies indicate that the SPS has a potential to be used as an alternative method for the diamond generation.

The Innovative Spark Plasma Sintering Technology

The science and technology of new materials is in the focus of ongoing research and development efforts worldwide. One main focus is directed toward the synthesis, characterization and applications of nanostructured and composite materials. The multifunctionality of such materials is caused by their internal structures of micro- and nanometer-sized units. However, the physical, chemical, and mechanical properties of

nanomaterials change not only due to the pure size reduction but also due to the dominant influence of boundary effects and the appearance of quantum mechanical phenomena. The size and surface/interface effects specific to nanostructured solids are nowadays best exploited in the development of new magnetic and lightweight materials. Significant progress has been achieved for hard and soft magnetic materials and exchange-spring composite magnets as well as for nanostructured alloys with improved mechanical properties. Nanomaterials can be metals, ceramics, polymeric materials or their composites. In the last years, innovative processing technologies like the spark plasma sintering have emerged. This high-speed powder consolidation technology has a broad application potential. During SPS, pulsed electric currents and mechanical pressure are applied on precursor powders in a graphite die. Pulsed electrical direct currents through the die and through the powder compacts, directly, in case they are conductive, lead to rapid heating while uniaxial pressure is applied on the powder. The Joule's heating effects can be strongly enhanced. Due to small gaps and small contact points of neighboring particles, plasma discharges and rapid local heating can occur and will lead to surface activation diffusion processes [5–7]. During the SPS synthesis, uniaxial pressure and pulsed currents allow to separate the densification of powders from the grain growth and, thus, to influence the internal structure of the materials. The essential advantages of SPS over conventional sintering methods like pressureless sintering, gas-pressure sintering, hot-pressing or hot isostatic pressure (HIP) are lower sintering temperatures, shorter exposure to elevated temperatures and no need for binder or additives as well as important improvements in the properties of the sintered materials. Processing under extreme pressure and temperature conditions leads to important changes of the density, electron configuration and free energy of the constituent phases. The resulting chemical equilibrium changes allow the access to a wide range of new compounds. External pressure can further induce reactions between chemical components at much lower temperatures, thus facilitating the synthesis of metastable phases with valuable modified properties. SPS has already demonstrated its potential as an energy-saving sintering method, for nanotechnology, for rapid manufacturing of monophasic materials and composites, for biomedical applications and for energy-saving and energy-harvesting materials [8]. This core technology has recently attracted high interest in the fields of material manufacturing, energy, health care, automotive and aeronautics.

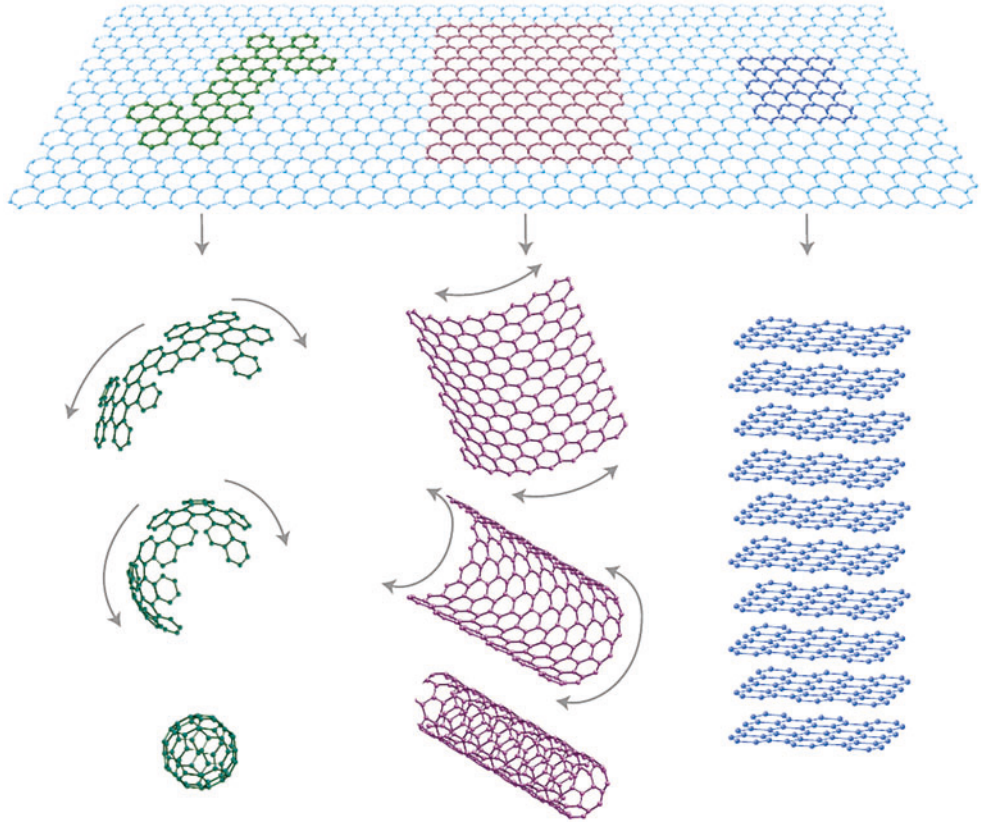
D

Nanocarbons

The nanocarbon materials include fullerenes like C60, carbon nanotubes (CNTs), graphene and nanodiamonds. These allotropes arise since carbon atoms with their original 6 electrons in the states $1s^2 2s^2 p^2$ are able to form various hybrid orbitals due to the energies of the 2 s and 2p states having almost similar energies allowing not only the simple states but also linear combinations of them to be solutions of the Schrödinger equation. Four equal bonding orbitals are formed in the sp^3 state, two in the sp^2 state and one in the sp state, forming localized and directed σ bonds. The other electrons are delocalized, forming the so-called π bonds.

The fullerene C60 is a carbon allotrope that consists of 60 carbon atoms, arranged as 12 pentagons and 20 hexagons. The shape resembles the one of a soccer ball. The discovery of C60 has stimulated strong activities in chemistry and opened with the fullerene chemistry a branch studying a new family of molecules based on fullerenes. The 1996 Nobel Prize for Chemistry has been awarded to H.W. Kroto, R.F. Curl, and R.E. Smalley for their discovery of C60 in 1985 as a new allotrope of carbon.

Carbon nanotubes have attracted many scientists worldwide. The small dimensions, strength and remarkable physical properties of these structures make them a unique material with a whole range of promising applications. Carbon nanotubes were discovered accidentally by Sumio Iijima 1991 while studying the surfaces of graphite electrodes used in an electric discharge. Carbon nanotubes with an incredible



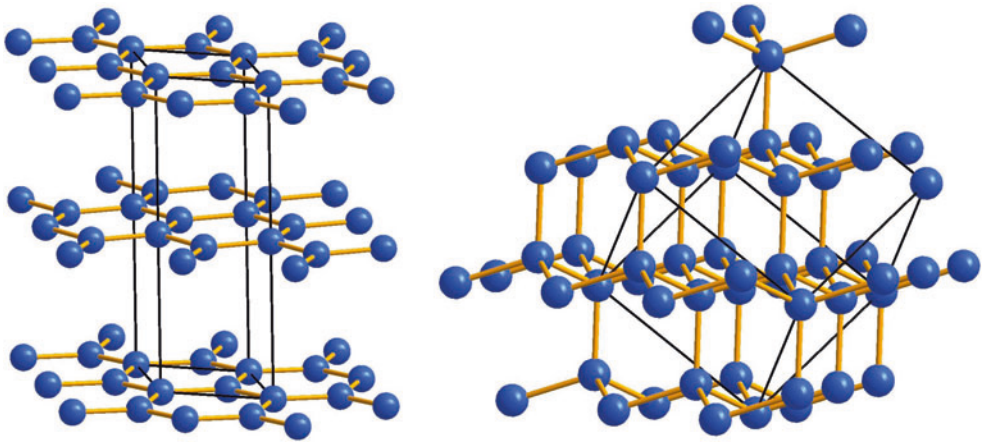
Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 2 Schematic of carbon allotropes (Reprinted from [9] with permission of Nature Publishing Group ©)

strength and fascinating electronic properties appear to be ready to overtake fullerenes in the race to the technological marketplace.

Graphene is the name given to a monolayer of carbon atoms tightly packed into a two-dimensional (2D) honeycomb lattice which is the basic building block for carbon materials of other dimensions, as shown in Fig. 2. It is bound by σ bonds that are sp^2 hybridized. Each carbon has one π orbital which contributes to a delocalized network of electrons. It can be wrapped up into 0D fullerenes, rolled into 1D nanotubes, or stacked into 3D graphite (Figs. 2 and 3). Recent years have witnessed many breakthroughs in research on graphene and significant advance in the mass production of this material. This one-atom-thick net of carbons uniquely combines extreme mechanical strength, exceptionally high electronic and thermal conductivities, impermeability to gases and many other supreme properties all of which make it highly attractive for numerous applications [9].

Nanoscale diamond particles were first produced by detonation in the USSR in the 1960s, but they remained essentially unknown to the rest of the world until the end of the 1980s. Then a number of important breakthroughs led to wider interest in these particles which are known as nanodiamonds. Nanodiamonds have excellent mechanical and optical properties, large surface areas and tunable surface structures. The nanodiamonds have very broad applications and, since they are nontoxic, also biomedical ones [10].

In diamond, the carbon atoms are bonded in a tetrahedral structure with sp^3 hybridization. Since there is no π orbital, there is no delocalized network of electrons which makes diamond an



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 3 Structure of graphite (*left*) and diamond (*right*) with the indicated bonds and unit cells

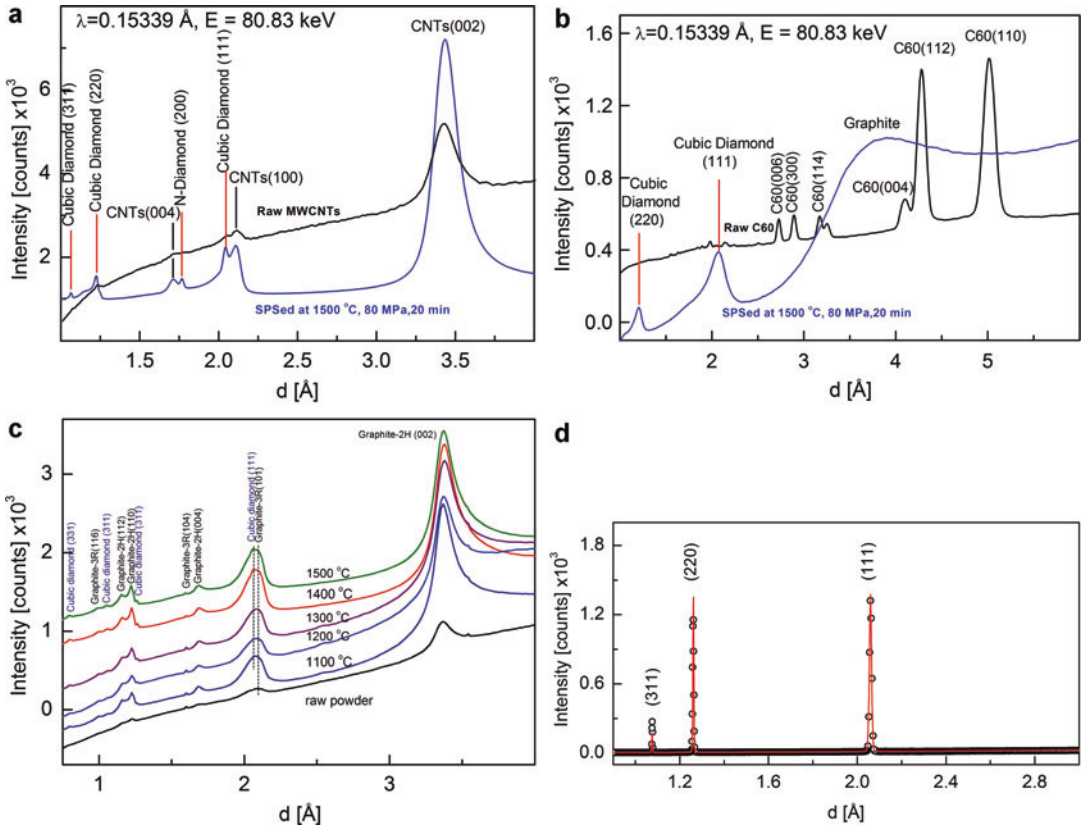
insulator. The diamond structure can be seen as two interpenetrating face-centered cubic structures displaced by one fourth of the cube diagonal. Figure 3 shows the structure of graphite (left) and the structure of diamond with indicated bonds and unit cells. The diamond lattice is rotated to make the similar staple of carbon planes visible.

Stability of Nanocarbon Under SPS

As already discussed, the widespread method to prepare diamonds is the HPHT method with pressures in the GPa region. Therefore, it is important to analyze the behavior of nanocarbons under the low pressures of a few MPa during the SPS treatment and to confirm the suitability of this method for a successful diamond synthesis. A variety of carbon allotropes, such as carbon nanotubes, C60, graphene nanoplatelets and graphite were investigated with respect to their thermal stabilities under SPS conditions [11, 12].

In these experiments, multiwall carbon nanotubes (MWCNT, 10–20 nm), graphite powders (about 10 μm) and graphene nanoplatelets (C-750) were used. The nanoplatelets consisted of short stacks of graphene sheets having platelet shapes with an average thickness of 2 nm and a lateral diameter of less than 2 μm. For the SPS treatment, the carbon powders are usually pressed into a graphite die to form a disk-shaped sample of 20 mm diameter and 5 mm thickness. A heating rate of 100 K min^{-1} was adopted and the sintering process lasted typically 20 min. The sintering temperatures were varied up to 1500 °C. The axial pressure has been 80 MPa and the sintering process was performed under vacuum condition. The applied pulsed direct current for SPS was 1000 A–2000 A. The sintered carbon samples were etched in a solution of concentrated H_2SO_4 (90 vol.%) and HNO_3 (10 vol.%). The etched samples were washed using deionized water repeatedly and dried in an oven.

The phase identification of the raw nanocarbon material and of the SPS sintered samples can be performed by diffraction experiments using high energetic synchrotron radiation X-rays. Figure 4 shows the exemplary spectra already as functions of the lattice spacings. The raw C60 shows diffraction peaks corresponding to spacings of C60 as given in the Fig. 4a. After SPS sintering, diffraction peaks of cubic diamond at spacings 2.06 Å and 1.23 Å appear, as well as a broad graphite peak. Since the C60 diffraction peaks disappeared, this indicates that the C60 was transformed into diamond and graphite during the SPS process. The raw MWCNTs (Fig. 4b) show a main diffraction peak corresponding to the CNT (002) spacing and weaker ones, corresponding to the (100) and (004) spacings. After the SPS process, the MWCNT diffraction peaks are still present with increased intensities

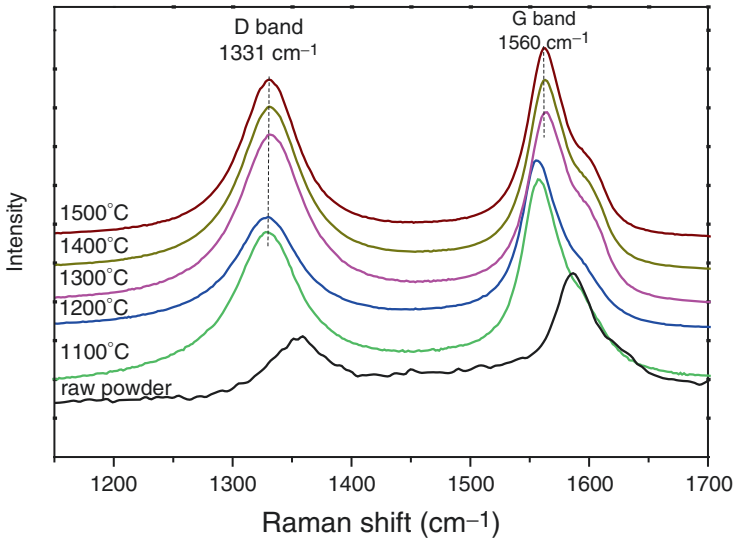


Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 4 X-ray diffraction patterns of (**a**) the raw C60 and the spark plasma sintered C60 at 1500 °C and 80 MPa for 20 min, of (**b**) the raw MWCNT and sintered MWCNT at 1500 °C and 80 MPa, and (**c**) the raw and sintered graphene nanoplatelets at 1100–1500 °C and 80 MPa and (**d**) of the purified diamond crystals

and reduced halfwidths indicating that the crystallinity improved. However, more important, new peaks corresponding to cubic diamond (111), (220) and nanodiamond (200) appear demonstrating the successful synthesis of diamond.

The diffraction peaks of the raw graphene nanoplatelets (Fig. 4c) are very weak and broad indicating their low crystallinity and their nanostructures. The raw graphene nanoplatelets have characteristic peaks corresponding to the graphite-2H (002) and graphite-3R (101) plane spacing. The intensities of the peaks increase continuously with the sinter temperature and the graphite-3R (104), (116) and graphite-2H (004), (110), (112) peaks are visible at temperatures from 1100 °C to 1500 °C. This shows that the crystallization and graphitization of the graphene nanoplatelets depend on the SPS temperatures. In addition, peaks corresponding to cubic diamond (111), (220), (311) and (331) planes can be identified in this temperature range. Figure 4d shows the diffraction pattern of extracted micrometer-sized grains of diamond after the SPS synthesis and purification with the typical diamond reflections.

Raman spectroscopy is another valuable tool to study the phase changes by analysis of typical dynamical features of the materials. Figure 5 shows the Raman spectra of the raw graphene nanoplatelets and the sintered samples at 1100 °C up to 1500 °C and at 80 MPa with typical modes. Here, the so-called graphite-like G-band peak gives information on the inplane stretching of the sp^2-bonded carbon in a two-dimensional



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering,
Fig. 5 Raman spectra of the raw and sintered graphene nanoplatelets at 1100–1500 °C and 80 MPa

graphitic hexagonal lattice and the disorder D-band peak gives information on the internal crystalline defect densities within the structure. The spectrum of the raw graphene nanoplatelets shows lower intensity in the D-band at 1355 cm^{-1} than in the G-band at 1585 cm^{-1} which indicates defects in the initial nanoplatelets. After SPS processing, the D-band shifted to 1331 cm^{-1} for the samples sintered at 1100–1500 °C. This value is typical for cubic diamond and its sp^3 hybridization, thus, proving a structural change to diamond introduced by the SPS process. The main peak of the G-band has shifted to 1556 cm^{-1} at temperatures of 1100 °C and 1200 °C and to 1560 cm^{-1} at 1300 °C–1500 °C with roughly unchanged contributions at 1585 cm^{-1} of the raw nanoplatelets. The shift of the G-band indicates a major change in the inplane C-C bond stretching proving the formation of sp^2-bonded carbon vibrations in graphite. It is because of the densification and graphitization of the graphene nanoplatelets during the SPS. Some graphene nanoplatelets are bonded together when the stackings match and are formed into nanographite stacks at higher temperatures. The Raman spectra of SPS sintered fullerenes and carbon nanotubes also confirmed the formation of nanodiamonds.

Scanning electron microscopy (SEM) images are shown in Fig. 6a, b for the raw and the sintered C60, in Fig. 6c, d for the raw and sintered MWCNT and in Fig. 6e, f for the raw and sintered graphene nanoplatelets, all sintered at 1500 °C and 80 MPa. The raw C60 powders are nanoparticle agglomerates and show bundles of C60 (Fig. 6a). Some diamond crystals with sizes from 2 μm to 8 μm can be observed (Fig. 6b). The fibrous structures of the raw MWCNTs are very obvious in Fig. 6c. After sintering, the fibrous structures disappear and some diamonds are found. Figure 6d shows one diamond crystal with a particle size of around 35 μm. In the background, there are no longer any MWCNTs visible. The raw powder of graphene shows many agglomerates with platelet shapes and particle sizes from nanometers to several micrometers (Fig. 6e). Some bright diamond particles ranging from 2 μm to 4 μm are found in the 1500 °C sintered samples (Fig. 6f). A larger diamond particle with size about 20 μm is found in the 1400 °C sintered sample (Fig. 6c). The diamond particle sizes are ranging from 2 μm to 4 μm in the 1500 °C sintered sample (Fig. 6d).

In addition, transmission electron microscopy (TEM) can give further information on the structure. The result of a C60 sample that was spark plasma sintered at 1200 °C is presented in Fig. 7 showing a diamond crystal and some C60 particles. The particle sizes of the C60 in this sintered sample lie in the range of 40 to 100 nm which is

Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering,

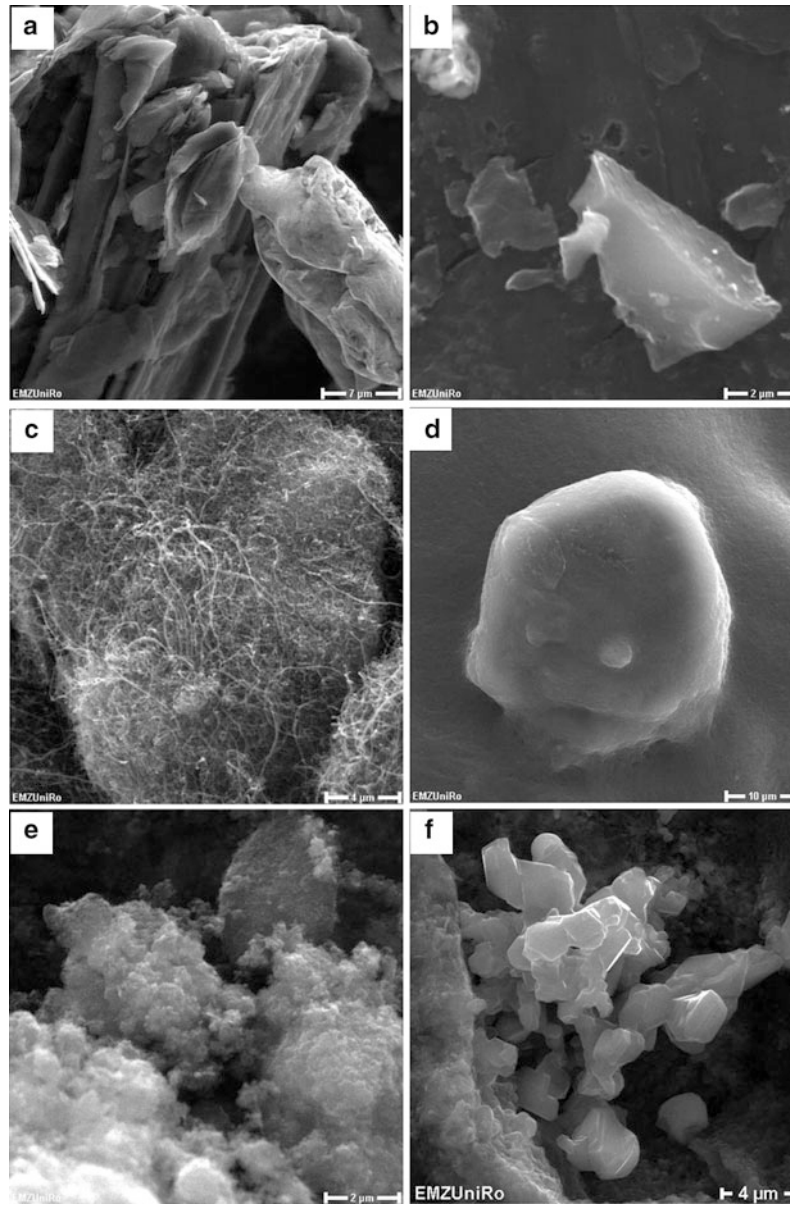
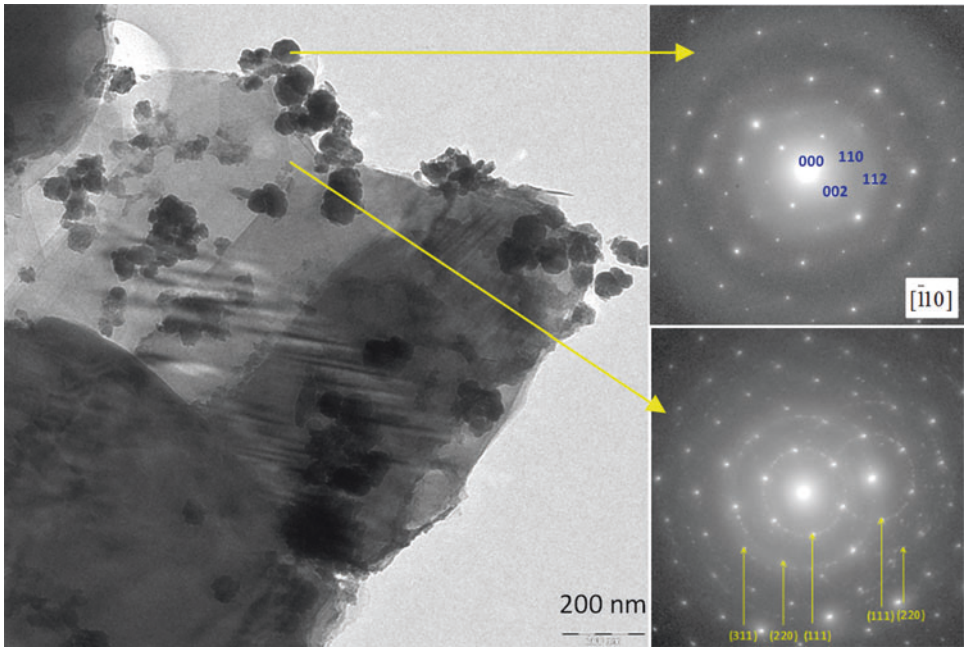


Fig. 6 SEM images of the raw and the spark plasma sintered C60, of (**a**, **b**) the raw and sintered MWCNT and of (**c**, **d**) the raw and sintered graphene nanoplatelets (**e**, **f**), all sintered at 1500 °C and 80 MPa

identical to the sizes of the raw C60. The C60 particles are present with the diamond crystal together. This clearly indicates that the diamond is transformed from these C60 particles, directly. The selected area diffraction patterns show the diffraction rings for C60 in the top one and for diamond with cubic polycrystals in the bottom one. No formation of an intermediate phase can be detected. The transformation from C60 to diamond is seen as a direct carbon transformation process with structural reconstruction without intermediate phase [13].

The particle size of the diamond crystals did not increase continuously but decreased again at the temperature of 1500 °C. Nevertheless, the degree of diamond crystallinity at 1500 °C is higher than the one at the lower temperatures. In a pressure-temperature diagram, the best



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 7 TEM image of the C60 to diamond with selected area diffraction patterns of C60 (*top*) and diamond (*bottom*)

transformation of graphene nanoplatelets to diamond occurs in the temperature range of 1300 °C–1400 °C at the pressure of 80 MPa.

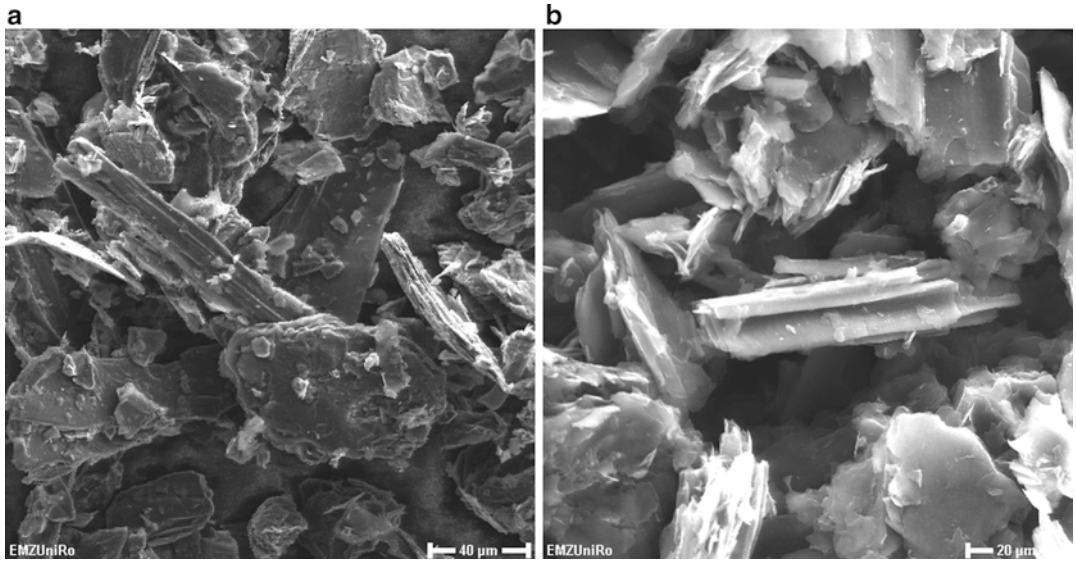
Stability of Graphite Under SPS

Figure 8 shows the SEM images of a raw and a spark plasma sintered graphite sample, prepared at 1500 °C and 80 MPa for 20 min. The raw graphite shows the typical layered structure (Fig. 8a). After the SPS treatment at similar conditions as for C60, MWCNTs and graphene, no diamond is present in the sample. Still, the typical structure of the raw graphite is visible in Fig. 8b. X-ray investigations and Raman spectroscopy confirm that there is no diamond formation if a pure graphite sample is processed by SPS at these mild conditions. The graphite is the most stable crystalline modification of carbon among the MWCNTs, graphene, C60 and graphite allotropes under the SPS.

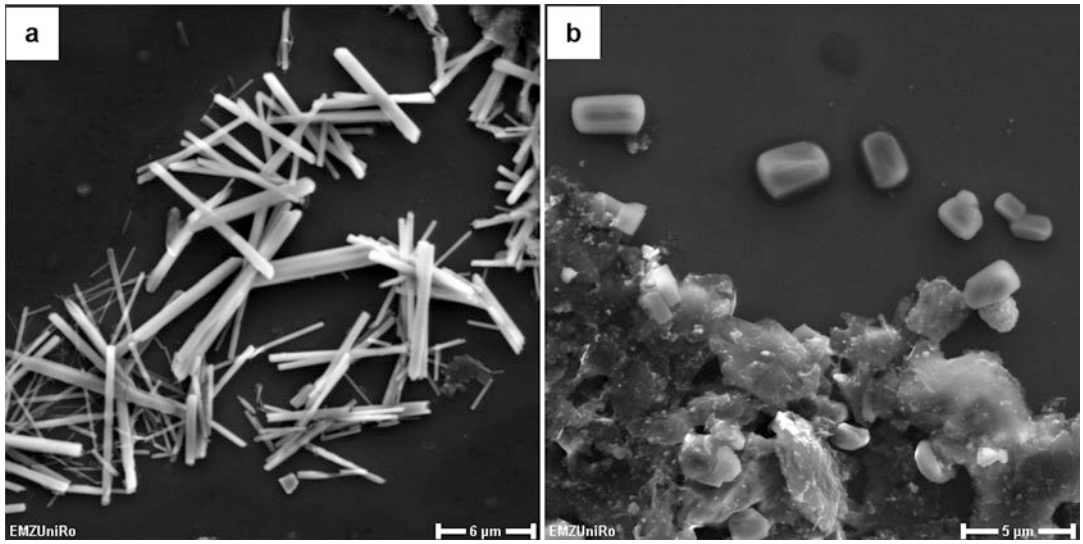
Diamond Formation from Graphite with Various Catalysts Under SPS

Pure graphite is the main carbon source for the diamond synthesis with the HPHT method. As mentioned before, the discussed conditions of modest temperatures and pressures in the SPS technology are not sufficient to transform pure graphite into diamond.

However, adding catalytic materials like Ni or quasicrystalline Al-Cu-Fe powders to pure graphite powders allows the transformation of graphite into diamond during a SPS treatment at 1300 °C and 50 MPa for 20 min [14]. Figure 9 shows the SEM images of the diamonds obtained after etching the sintered samples. The Ni catalyst supported the growth of nano- and microrods from graphite (Fig. 9a). The diameter of the diamond rods ranges from 80 nm to 2 μm. Therefore, the SPS method is suited for the synthesis of diamond nano- and microrods. Quasicrystalline Al-Cu-Fe powder used as catalyst allows the

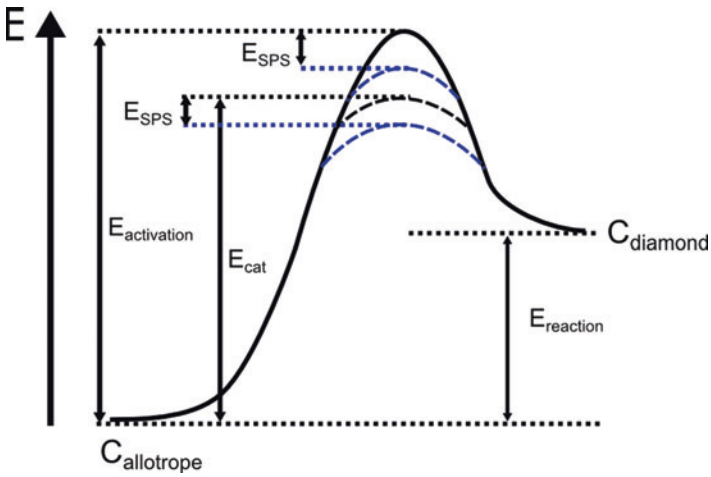


Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 8 SEM images of the raw graphite (**a**) and the spark plasma sintered graphite (**b**) at 1500 °C and 80 MPa for 20 min



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 9 SEM images of the diamonds gained from the spark plasma sintered graphite samples at 1300 °C and 80 MPa mixed with Ni (**a**) and Al-Cu-Fe quasicrystal (**b**) as catalytic materials

conversion of graphite to well-shaped diamond grains from 1 μm to 3 μm (Fig. 9b). Figure 10 demonstrates schematically the energy barrier between diamond and various carbon allotropes. In the case of graphite, for instance, the high-energy barrier of the activation energy, $E_{activation}$, compared with the real reaction energy, $E_{reaction}$, can easily be crossed with HPHT. With the



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering,
Fig. 10 Schematic view of the energy barrier for the transformation of various carbon allotropes into diamond

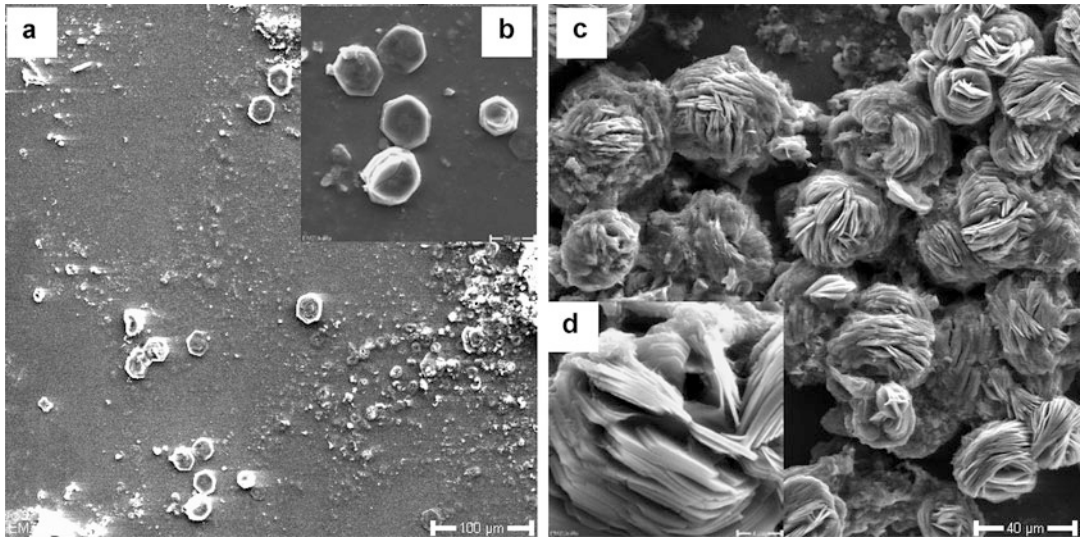
application of a catalyst, the energy barrier is reduced (marked by E_{cat}) and the SPS treatment delivers the necessary energy, E_{SPS}, to cross the barrier, even for graphite. For the other allotropes, the SPS treatment alone can deliver the necessary energy for the barrier crossing. However, the application of catalysts can promote the transition, further.

Diamond Synthesis from Carbon Nanotubes with Solvent Catalysts

In the diamond synthesis with the SPS method, the atmospheres (vacuum or argon gas), the pulsed direct current intervals, and the use of catalysts have a direct influence on the diamond sizes and the transition rates that can be achieved [15, 16]. For a demonstration, MWCNTs/NiFe and graphite/NiFe mixtures were used as raw materials. The microstructures and diamond phases were analyzed by X-ray diffraction, Raman spectroscopy and scanning electron microscopy. It was proven that diamond can be synthesized by SPS in an argon gas atmosphere with a minimum pressure of 9.5 MPa from both mixtures, MWCNTs/NiFe and graphite/NiFe. High-quality diamond crystals with hexahedron structures can be found in the sample (Fig. 11a). A high-magnification SEM micrograph of such diamond crystals obtained by SPS is shown in Fig. 11b. The flowerlike structured carbons visible in the sample (Fig. 11c, d) are seen as a proof that sparking plasmas occurred leading to such carbon flowers. The diamond shapes and sizes increase by the use of Ar gas. This indicates that the Ar gas increases the plasma effects in the SPS. The variations of the pulsed current intervals lead also to changes in the structures and the sizes of the synthesized diamond grains. This demonstrates that the localized temperatures can be varied by the pulse intervals thus influencing the growth of the diamond nuclei.

The Role of SPS in the Transformation from Carbon to Diamond

For further understanding of the nucleation of the diamond grains, in situ diffraction experiments using synchrotron radiation were performed at a pressure of 80 MPa up to a temperature of 1500 °C using standard steady current heating [12]. The diffraction patterns indicated clearly that the graphene nanoplatelets are stable under these conditions which can be attributed to the lack of current pulses being present in the SPS treatment. In these experiments, not the powder but only the sample container was heated by Joule's heating. It was demonstrated that similar pressure and temperature conditions and only steady current heating did not allow the nucleation of diamond grains. Therefore, it is concluded that the pulsed current heating has an important influence. In the

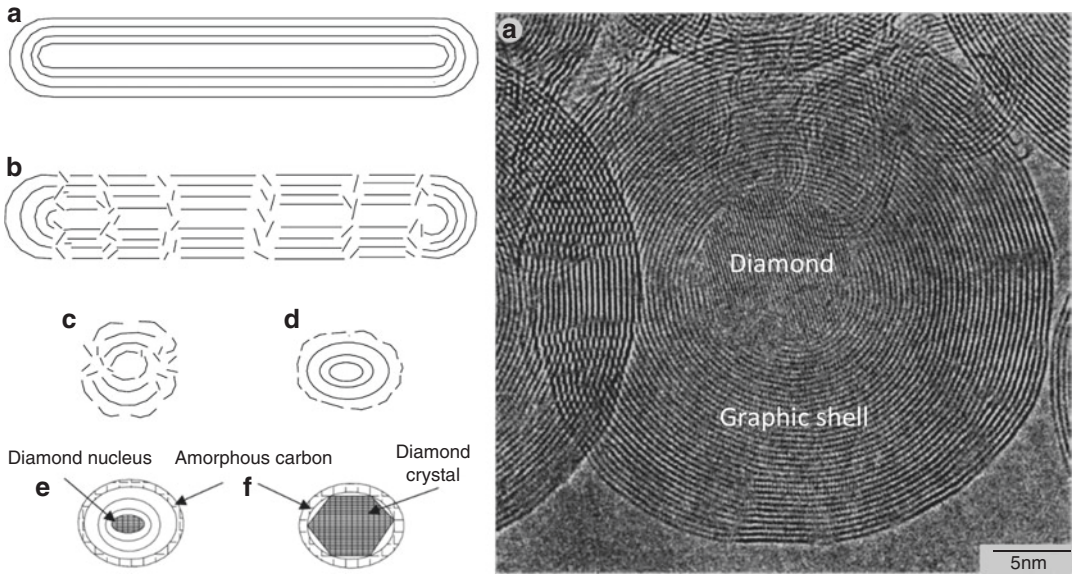


Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 11 SEM micrographs of the spark plasma sintered MWCNTs/FeNi at 1200 °C and 9.55 MPa in Ar gas atmospheres with various magnifications showing the diamond crystals (**a**, **b**) and carbon flowers (**c**, **d**)

SPS method, the electric direct current employs on-off pulses with high current values up to 6000 A at low voltages up to approximately 5 V which can generate highly localized Joules heating up to a few thousand degrees Celsius between particles, within minutes. The current density in the SPS is typically on the order of 10^3 A/cm^2 and it is highly concentrated at the intergranular contacts or interfaces [17], thus leading to localized high temperatures. Momentary plasmas can be generated at the intergranular contacts or interfaces of the powder particulates by the pulsed direct current. There exists a certain activation barrier from the graphene (sp^2) to diamond (sp^3). Bonds have to be broken and rearranged with a different hybridization. The energy of the occurring plasmas allows to pass the activation energy barrier from graphene to diamond, as sketched in Fig.10. The plasmas generated by the pulsed direct current increase the entropy of the whole carbon system thus promoting the diamond formation. The generation of plasmas during the SPS is the main thermodynamic reason for the diamond nucleation. However, the graphite itself cannot be transformed into diamond at the same SPS conditions. This means that graphene nanoplatelets, fullerenes and MWCNTs behave really different than graphite. The graphene nanoplatelets are nanometer-thick graphite with an average particle thickness of 2 nm consisting of less than 10 sheet stacks of graphene. The properties of the graphene nanoplatelets are similar to the properties of the single-layer graphene but not to the ones of graphite.

The model of CNT transformation into diamond in SPS was proposed in [18]: CNT fracture to form an intermediate phase of bulky onions and the diamond nucleates at the cores of onions as sketched in Fig. 12 (left) inspired by the TEM observations in Fig. 12 (right) [13].

From this model, it is predicted that the diameter of CNTs determines the transformation rates to diamond crystals. It was found that the sizes of the achieved diamond particles can be increased if the sizes of the used CNTs are reduced. These results match with the theoretical assumptions. The transformation of C60 to diamond is a direct solid-state phase transformation without intermediate phase (see Fig. 7). It is a SPS-induced phase transformation with a new arrangement of



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 12 The model of CNT transformation into diamond in SPS [18] and TEM observation (Reprinted from [19] with permission of Nature Publishing Group ©)

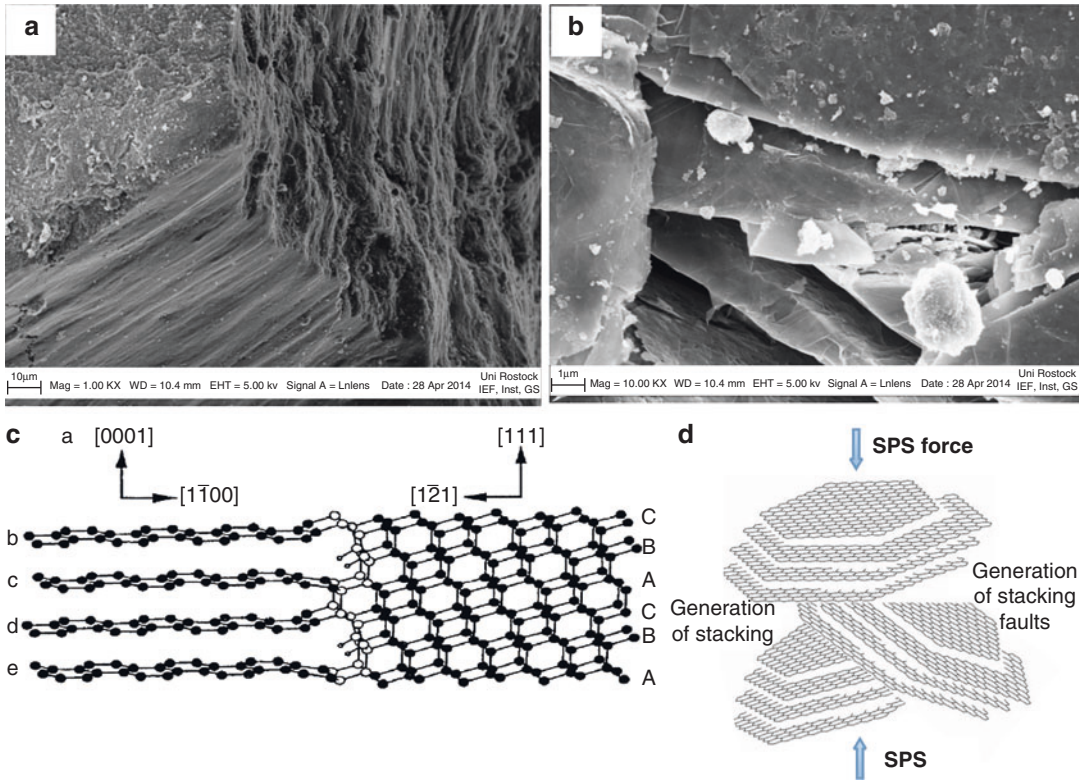
carbon atoms in this extreme physical field during the SPS process.

As demonstrated, diamond can be directly synthesized by SPS from graphene nanoplatelets at a very low pressure of 80 MPa without any catalysts being involved. The diamond is observed in all the graphene samples treated at 1100 °C up to 1500 °C with SPS. The particle sizes of the diamonds (up to 20 μm) obtained from the graphene nanoplatelets are much smaller than those from carbon nanotubes (up to 150 μm) and from fullerenes (up to 250 μm) at the same pressure and temperature range. Whereas graphene is a pure sp^2-hybridized material, carbon nanotubes and fullerenes have fractions of sp^3 hybridization due to angular strains.

Figure 13 shows a SEM micrograph of a bulk sample of graphene nanoplatelets treated by SPS at 1400 °C with a fracture surface in the right part. The fractured surface and the sample wall (Fig. 13a) show a clear layer structure perpendicular to the direction of the uniaxial pressure of the SPS almost similar to graphite [12]. The layers are stacks parallel to the top and bottom surfaces of the compacted and sintered graphene body. Etching the fracture surface reveals additional insight into the fine structure. During the SPS, the layered graphene nanoplatelets were packed randomly at the uniaxial pressure of 80 MPa. The fracture surface in Fig. 13b clearly shows that the graphene nanoplatelets are not only pressed parallel to each other but that they are folded and bent causing also a large amount of stacking faults in such SPS-treated bulk graphene solid. Therefore, during the SPS treatment, planar stacks of graphene nanoplatelets are pressed together to form graphite stacks. Hereby, the bent and tilted graphene nanoplatelets create also a considerable amount of stacking faults that are able to trigger the diamond nucleation (visible in the center of Fig. 13b) confirming the findings in Fig. 4a.

High-resolution TEM analysis of the SPS-treated nanoplatelets clearly shows a diamond nucleus with a lattice spacing of 2.06 Å ± 0.02 Å within the compacted, bent and tilted nanoplatelets (Fig. 14).

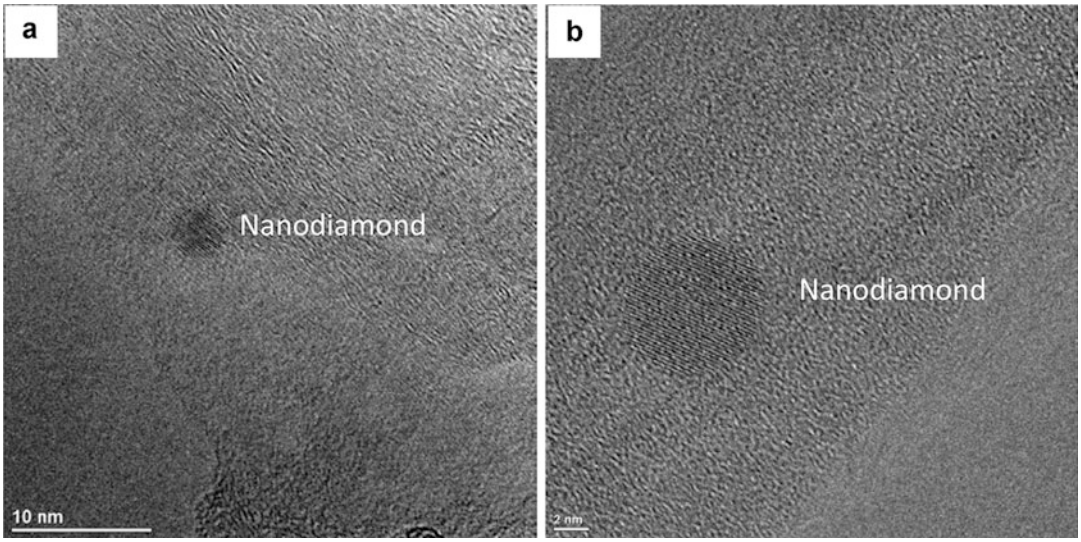
Already, Lambrecht et al. [20] found that the diamond formation will just become possible through the formation of stacking faults in graphite (Fig. 13c). Based on this, a schematic



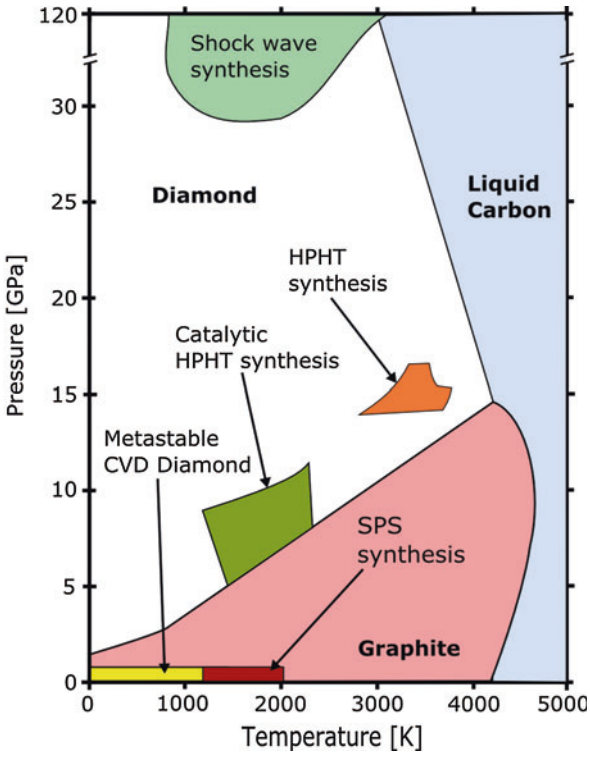
Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 13 SEM micrographs of a fracture surface of the spark plasma sintered graphene nanoplatelet bulk sample before etching showing the stacks (**a**) and a fracture surface after etching with folded nanoplatelets and a diamond grain in the center part (**b**), a schematic with the proposed nucleation mechanism of the diamond nuclei formation at graphite layers (**c**) (Reprinted from [20] with permission of Nature Publishing Group ©) and an illustration indicating the generation of stacking faults during SPS of the layered nanoplatelets (**d**)

illustration of the mechanism for the transformation from graphene nanoplatelets to diamond during the SPS is sketched in Fig. 13d. During the SPS consolidation, the contact areas of tilted and bent graphene layers give rise to a considerable amount of stacking faults and locally high dislocation densities. Due to the applied pressure, diffusion will occur and new interlayer covalent bonds among the graphene nanoplatelets are created. The interlayer covalent bonds will lead to the formation of regions of sp^3-hybridized carbon atoms that are then serving as diamond nucleation sites. In pure graphite, however, there are not enough stacking faults created at the low pressure of 80 MPa applied in the SPS which inhibits the diamond formation as described before.

In conclusion, graphene nanoplatelets can be transformed to diamond without any catalysts being involved at the low pressure of 80 MPa and at temperatures from 1100 °C to 1500 °C in a SPS process. Due to the pulsed current, the locally extremely high temperatures along with possible plasma formation support the creation of new bonds in the regions of high densities of stacking faults serving as nucleation centers for the diamond growth. Thus, the pressure-temperature carbon phase diagram (Fig. 1) can be expanded to the SPS method leading to Fig. 15.



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 14 High-resolution TEM analysis of the SPS-treated nanoplatelets (**a**, **b**) with a nanodiamond, a nucleus with a lattice spacing of 2.06 Å in the compacted, bent, and tilted nanoplatelets (Images by J. Barthel, RWTH Aachen University)



Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering, Fig. 15 Pressure-temperature carbon phase diagram with the SPS method included (Modified from [1] with permission of John Wiley and Sons ©)

References

1. Bundy, F.P.: The P.T phase and reaction diagram for elemental Carbon. J. Geophys. Res. **85**(B12), 6930 (1980) and http://www.ch.ic.ac.uk/rzepa/mim/century/gifs/cphased.gif
2. Nunez, R.M., Monceau, P., Hodeau, J.L.: Crushing C60 to diamond at room temperature. Nature **355**, 237–239 (1992)
3. Zhang, F., Shen, J., Sun, J., Zhu, Y.Q., Wang, G., McCartney, G.: Conversion of carbon nanotubes to diamond by a spark plasma sintering. Carbon **43**, 1254–1258 (2005)
4. Zhang, F., Burkel, E.: Novel titanium manganese alloys and their macroporous foams for biomedical applications prepared by field assisted sintering. In: Laskovski, A.N. (ed.) Biomedical Engineering, Trends, Researches and Technologies, pp. 203–224. InTech, Rijeka (2011)
5. Zhan, G.D., Kuntz, J., Wan, J., Garay, J., Mukherjee, A. K.: A novel processing route to develop a dense nanocrystalline alumina matrix (<100 nm) nanocomposite material. J. Am. Ceram. Soc. **86**(1), 200–202 (2003)
6. Omori, M.: Sintering, consolidation, reaction and crystal growth by the spark plasma system (SPS). Mater. Sci. Eng. A **287**, 183–188 (2000)
7. Chen, W., Anselmi-Tamburini, U., Garay, J.E., Groza, J.R., Munir, Z.A.: Fundamental investigations on the spark plasma sintering/synthesis process: I. Effect of dc pulsing on reactivity. Mater. Sci. Eng. A **394**, 132–138 (2005)
8. Orrù, R., Licheri, R., Locci, A.M., Cincotti, A., Cao, G.: Consolidation/synthesis of materials by electric activated/assisted sintering. Mater. Sci. Eng. R **63**, 127–287 (2009)
9. Geim, A.K., Novoselov, K.S.: The rise of graphene. Nat. Mater. **6**, 183–191 (2007)
10. Mochalin, V.N., Shenderova, O., Ho, D., Gogotsi, Y.: The properties and applications of nanodiamonds. Nat. Nanotechnol. **7**(1), 11–23 (2012). 2007
11. Zhang, F., Mihoc, C., Ahmed, F., Latte, C., Burkel, E.: Thermal stability of carbon nanotubes, fullerene and graphite under spark plasma sintering. Chem. Phys. Lett. **510**, 109–114 (2011)
12. Zhang, F., Eisenhut, F., Burkel, E.: Pulsed direct current field induced phase transformation in graphene nanoplatelets. Appl. Phys. Lett. **104**, 253108 (2014)
13. Zhang, F., Ahmed, F., Holzhuter, G., Burkel, E.: Growth of diamond from fullerene C60 by spark plasma sintering. J. Cryst. Growth **340**, 1–5 (2012)
14. Zhang, F., Burkel, E.: Synthesis of diamond using spark plasma sintering. In: Lakshmanan, A. (ed.) Sintering of Ceramics-New Emerging Techniques, pp. 27–58. InTech, Rijeka (2012)
15. Zhang, F., Ahmed, F., Bednarcik, J., Burkel, E.: Diamond synthesis through the generation of plasma during spark plasma sintering. Phys. Status Solidi A **11**, 2241–2246 (2012)
16. Zhang, F., Adam, M., Otterstein, E., Burkel, E.: Pulsed electric field induced diamond synthesis from carbon nanotubes with solvent catalysts. Diamond Relat. Mater. **20**, 853–858 (2011)
17. Yang, K., He, J., Su, Z., Reppert, J.B., Skove, M.J., Tritt, T.M., Rao, A.M.: Inter-tube bonding, graphene formation and anisotropic transport properties in spark plasma sintered multi-wall carbon nanotube arrays. Carbon **48**(3), 756–762 (2010)
18. Shen, J., Zhang, F., Sun, J., Zhu, Y.Q., McCartney, G.: Spark plasma sintering assisted diamond formation from carbon nanotubes at very low pressure. Nanotechnology **17**, 2187–2191 (2006)
19. Banhart, F., Ajayan, P.M.: Carbon onions as nanoscopic pressure cells for diamond formation. Nature **382**, 433–435 (1996)
20. Lambrecht, W.R.L., Lee, C.H., Segall, B., Angus, J.C., Li, Z., Sunkara, M.: Diamond nucleation by hydrogenation of the edges of graphitic precursors. Nature **364**, 607–610 (1993)

Diamond Synthesis

▶ Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering

Dielectric Force

▶ Dielectrophoresis

Dielectrophoresis

Nicolas G. Green[2,3] and Hossein Nili[1,2,3]
[1]Nano Research Group, University of Southampton, Highfield, Southampton, UK
[2]School of Electronics and Computer Science, University of Southampton, Highfield, Southampton, UK
[3]Nano Research Group, Faculty of Physical and Applied Sciences, University of Southampton, Southampton, UK

Synonyms

AC electrokinetics; Dielectric force

Definition

Dielectrophoresis (DEP) is defined as the motion of polarizable particles under the influence of an applied *nonuniform* electric field, with the force arising from the interaction of the field and the dipole moment induced in the particle.

Introduction

Dielectrophoresis is a second-order effect arising from the net effect of the dielectric force acting on interfaces between different dielectrics. It is second order in that the field generates a charge on the interface due to the different polarizability of the dielectrics and then acts on that charge to produce a force. The second-order nature also means that the time average of the force in an AC field is nonzero. Where the interface is a closed surface surrounding a volume such as a particle or a droplet which is free to move, the net force results in motion where the field is nonuniform. This dielectrophoretic motion is due to the fact that the charge on the interface forms a dipole across the particle and the force on the two poles of the dipole is different since the electric field is nonuniform.

The phenomenon of dielectrophoresis was first recorded in the late nineteenth century as an effect arising in the behavior of soil samples. The term dielectrophoresis was coined by Herbert Pohl in 1951 to describe the motion of biological particles under the influence of nonuniform electric fields [1]. Pohl went on to study the effect in detail and in 1978 published the seminal work on the subject [2]. With the application of semiconductor fabrication methods allowing researchers to easily build field-generating electrodes on the scale of biological cells, dielectrophoresis gained traction in the late twentieth century. Applications followed for manipulation, characterization, and separation of micrometer-sized particles such as cells and bacteria down to nanometer-scale particles such as viruses and proteins.

Interfacial Polarization of Dielectric Materials

The forces which give rise to dielectrophoresis are general in that a force is generated on any surface between materials or any internal discontinuity in a material. Understanding of the force starts with an understanding of dielectric materials: materials which possess bound charges which *polarize* under the influence of an applied electric field. The bound charges can only move minute distances under the influence of an externally applied electric field.

Polarization

In electrostatics, the force acting between charges or the force on a charge in an electric field can be written as $\mathbf{F}_c = q\mathbf{E}$, where q is the charge and $\mathbf{E}$ is the electric field. When positive and negative charges are placed in an electric field, they are therefore displaced in opposite directions. If those charges are bound to atoms or molecules, they will move a short distance and then be held in displaced positions, producing a dipole moment in the atom/molecule.

The dipole moment is a fundamental object in electrostatics consisting of two equal and opposite charges, $+q$ and $-q$ for example, displaced by a vector distance $\mathbf{d}$. The dipole moment in this case is given by $\mathbf{p} = q\mathbf{d}$. Many molecules possess fixed dipole moments, and in the case of dipoles arising from an applied electric field, the molecule then has an *induced* dipole moment. The process of this happening in a dielectric material, with all the molecules developing dipole moments, is referred to as *polarization*. In a dielectric, consisting of N molecules per cubic meter, the electric polarization $\mathbf{P}$, or the average vector dipole moment per molecule, is given by $\mathbf{P} = N\mathbf{p}$.

There are three basic polarization processes. Under an applied electric field, the center of charge of the electron cloud in a molecule moves slightly with respect to the center of charge of the nucleus; this is referred to as electronic polarization. In a solid such as crystalline potassium chloride, the ions of different signs move in different directions when subjected to an electric field; this

is referred to as atomic polarization. If the molecules possess a permanent dipole moment, the material is described as polar, and the molecules both align with and become further polarized in an applied electric field; this is referred to as orientational polarization.

Charge Movement in Materials

Polarization causes charges to gather, either at points within the dielectric or at the surface [3, 4]. Differences in polarization on either side of an interface mean that there is a net charge at that interface. This charge can be generated by bound charges from polarization as described above or in real rather than ideal dielectrics by free charge associated with electrical conduction. In water-based biological systems where dielectrophoresis is typically applied, there is generally a fairly high conductivity arising from the ions and salts suspended in the solutions.

The important relationships can be derived simply by looking at charge densities and the relationship with polarization and the electric field. In a material, the bound volume charge density is related to the polarization by

$$\rho_b = -\nabla \cdot \mathbf{P} \tag{1}$$

and the total charge, which is the sum of the free charge density ρ_f and the bound charge density, is related to the electric field by Gauss's law:

$$\nabla \cdot \mathbf{E} = \frac{\rho_f + \rho_b}{\varepsilon_o} \tag{2}$$

where ε_o is the permittivity of free space. Substituting Eq. 1 into Eq. 2 gives

$$\nabla \cdot (\varepsilon_o \mathbf{E} + \mathbf{P}) = \rho_f \tag{3}$$

This is generally written as

$$\nabla \cdot \mathbf{D} = \rho_f \tag{4}$$

where $\mathbf{D} = \varepsilon_o \mathbf{E} + \mathbf{P}$ is termed the electric flux density. For a linear isotropic dielectric, $\mathbf{P}$ is proportional to $\mathbf{E}$ such that $\mathbf{P} = \varepsilon_o \chi \mathbf{E}$ where χ is the electric susceptibility of the dielectric. Therefore, for a linear and isotropic dielectric, from Eq. 3:

$$\mathbf{D} = \varepsilon_o \mathbf{E} + \mathbf{P} = \varepsilon_o (1 + \chi)\mathbf{E} = \varepsilon_o \varepsilon_r \mathbf{E} \tag{5}$$

where $\varepsilon_r = (1 + \chi)$ is the relative permittivity of the dielectric.

Frequency-Dependent Charge Processes

The extension of the simple electrostatic model to incorporate polarization effects and processes which depend on frequency can be illustrated using some simple models [3]. The most basic aspect of charge processes which involve frequency is that of conduction from the movement of free charges. This can be illustrated using a model parallel plate capacitor of area A and plate separation d, connected to an alternating potential ϕ of angular frequency ω, as shown in Fig. 1a. With an ideal dielectric filling the capacitor, the impedance is

$$Z = \frac{1}{i\omega C} \tag{6}$$

where $i^2 = -1$ and the capacitance C is given by

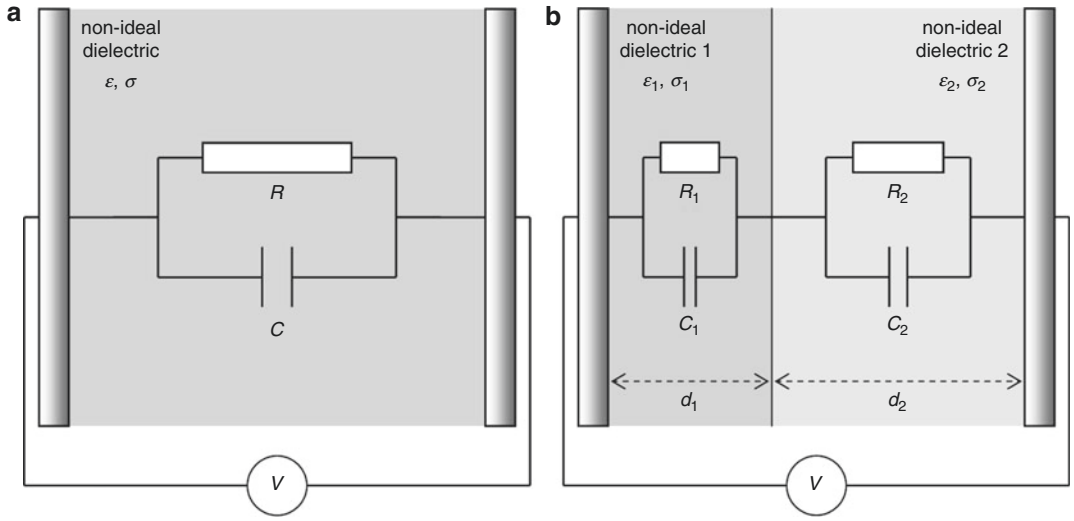
$$C = \varepsilon_o \varepsilon_r \frac{A}{d} \tag{7}$$

Most real dielectrics have a nonzero conductivity, and the permittivity is dependent on the frequency of the applied electric field. With the ideal dielectric replaced with a real dielectric of conductivity σ, the current in the circuit is the same as if the circuit were replaced with a capacitor (capacitance C from Eq. 8) representing the ideal dielectric and resistance in parallel given by

$$R = \frac{d}{\sigma A} \tag{8}$$

The total impedance of the circuit is then

$$Z = \frac{1}{1/R + i\omega C} = \frac{1}{\sigma \frac{A}{d} + i\omega \varepsilon_o \varepsilon_r \frac{A}{d}} = \frac{1}{i\omega \left(\varepsilon_o \varepsilon_r - i\frac{\sigma}{\omega}\right)\frac{A}{d}} \tag{9}$$



Dielectrophoresis, Fig. 1 (**a**) Parallel plate capacitor of plate separation d and area A containing a dielectric of permittivity $\varepsilon = \varepsilon_r \varepsilon_o$ and conductivity σ has the same impedance as the parallel circuit shown. (**b**) The same parallel plate capacitor completely filled by two dielectrics, with permittivities ε_1, ε_2, conductivities σ_1, σ_2, and widths d_1 and d_2 has the equivalent circuit shown, comprising that of the individual dielectrics in series

The real dielectric therefore behaves as a capacitor:

$$C = \tilde{\varepsilon}\,\frac{A}{d} \tag{10}$$

with a *complex polarizability* given by

$$\tilde{\varepsilon} = \varepsilon_o \varepsilon_r - i\frac{\sigma}{\omega} \tag{11}$$

This gives the frequency-dependent behavior of a basic homogeneous dielectric material. In general, the behavior of a dielectric is more complicated. Polarization mechanisms have a frequency-dependent nature, and the material has additional frequency dependencies due to relaxation mechanisms.

Since polarization mechanisms involve the movement of charge to create dipoles and the rate of this movement is finite, each mechanism therefore has an associated characteristic time constant which is the time necessary for the polarization to reach its maximum value. As the frequency of the applied alternating field increases, the period over which the dipole moment is able to "relax" and then follow the field decreases. Therefore, there is a frequency at which these two periods are the same and beyond which the polarization no longer reaches its maximum. At much higher frequencies, no polarization occurs and a relaxation or dispersion has occurred in the polarizability.

Of the three basic polarization mechanisms outlined previously, orientational polarization has the longest relaxation time. Atomic and electronic polarization will align with the field up to frequencies of the order of 10^{14} Hz and from the perspective of dielectrophoresis, which typically operates only up to 10^8 Hz, can be considered constant, i.e., $\mathbf{P}_a = \varepsilon_o \chi_a \mathbf{E}$. The orientational polarization is given by a complex expression which accounts for the frequency variation:

$$\mathbf{P}_{or} = \frac{\varepsilon_o \chi_{or} \mathbf{E}}{1 + i\omega\tau_{or}} \tag{12}$$

where τ_{or} is the *relaxation time* and χ_{or} is the low-frequency static limit for the orientational polarization mechanism. The total frequency-dependent polarization is then

$$\mathbf{P} = \varepsilon_o \left(\chi_a + \frac{\chi_{or}}{1 + i\omega\tau_{or}} \right) \mathbf{E} \tag{13}$$

At the low-frequency limit, $\chi = \chi_a + \chi_{or} = \varepsilon_s - 1$, where ε_s is the relative permittivity measured in a static electric field. At the high-frequency limit, $\chi = \chi_a = \varepsilon_\infty - 1$, where ε_∞ is the relative permittivity at sufficiently high frequency that there is no orientational polarization. χ_{or} is therefore given by $\varepsilon_s - \varepsilon_\infty$ and the total polarization given by

$$\mathbf{P} = \varepsilon_o(\varepsilon^* - 1)\mathbf{E} \tag{14}$$

where

$$\varepsilon^* = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1 + i\omega\tau_{or}} \tag{15}$$

This description gives the frequency-dependent polarization of a dielectric material and is normally written in separate real and imaginary components as the Debye formulations:

$$\varepsilon^* = \varepsilon' - i\varepsilon'' \text{ where } \varepsilon' = \varepsilon_\infty + \frac{\varepsilon_s - \varepsilon_\infty}{1 + \omega^2\tau_{or}^2} \text{ and } \varepsilon'' = \frac{\omega\tau_{or}(\varepsilon_s - \varepsilon_\infty)}{1 + \omega^2\tau_{or}^2} \tag{16}$$

Interfacial Polarization

As stated previously, experimental systems are further complicated by the presence of many dielectrics with differing properties, and the difference in polarizability causes charges to build up at the discontinuities or interfaces between the dielectrics [3]. As the polarizabilities of each dielectric are frequency dependent, so the surface charge density at the interface is also frequency dependent. As a result, the permittivity and conductivity of the whole system exhibit additional dispersions due solely to the polarization of the interfaces. The study of such heterogeneous dielectrics was first discussed by Maxwell and extended by Wagner, and the dispersion in permittivity arising because of an interface is referred to as Maxwell-Wagner interfacial polarization. This can be illustrated by reference to a simple parallel plate capacitor as shown in Fig. 1b where there are two dielectrics with different properties. The two dielectrics have permittivities ε_1 and ε_2 and conductivities σ_1 and σ_2, respectively, and they can be represented as two resistor/capacitor parallel circuits in series. The impedance of the equivalent circuit is given by

$$Z = Z_1 + Z_2 = \frac{R_1}{1 + i\omega R_1 C_1} + \frac{R_1}{1 + i\omega R_2 C_2} \tag{17}$$

This can be reworked as Eq. 10, with complex permittivity given by

$$\tilde{\varepsilon} = \varepsilon_o\varepsilon' - i\varepsilon_o\varepsilon'' - i\frac{\sigma}{\omega} \tag{18}$$

ε' and ε'' are given by the Debye formulations (Eq. 16) with

$$\varepsilon_\infty = \frac{d\varepsilon_1\varepsilon_2}{d_1\varepsilon_2 + d_2\varepsilon_1}$$
$$\varepsilon_s = \frac{d\left(d_1\varepsilon_1\sigma_2^2 + d_2\varepsilon_2\sigma_1^2\right)}{\left(d_1\sigma_2 + d_2\sigma_1\right)^2} \tag{19}$$
$$\tau = \varepsilon_o\frac{d_1\varepsilon_2 + d_2\varepsilon_1}{d_1\sigma_2 + d_2\sigma_1}$$

σ is the conductivity of the whole system, given by

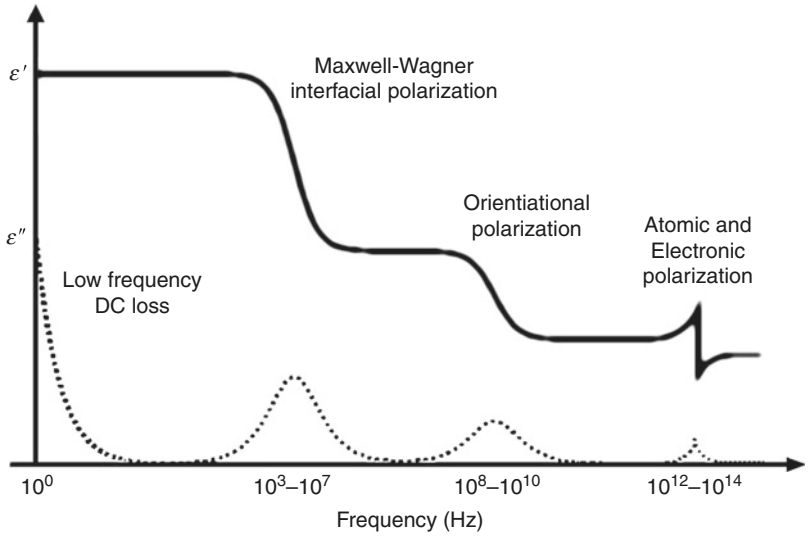
$$\sigma = \frac{d\sigma_1\sigma_2}{d_1\sigma_2 + d_2\sigma_1} \tag{20}$$

The resulting permittivity of the whole material in the capacitor has a Debye-type relaxation solely due to interfacial polarization. The frequencies for interfacial polarization occur at lower frequencies than for orientational polarization, and a plot of the complex permittivity of the system as a whole would look something like Fig. 2.

Dielectrophoresis

Interfacial Polarization of Particles

A particle suspended in a liquid is a special case of a multi-dielectric system with a closed surface surrounding a volume of a dielectric material (the particle) [3]. Some particles have a more complicated internal structure, which introduce



Dielectrophoresis, Fig. 2 The frequency variation of the complex permittivity of a dielectric, taking into account the typical relaxation mechanisms

additional interfaces to the model. Under the influence of an applied electric field the particle/liquid interface is charged by Maxwell-Wagner interfacial polarization.

A schematic diagram of the resulting effect on the particle is shown in Fig. 3. The difference in polarizability results in a surface charge density around the interface at the surface of the particle. This charge has an opposite sign on either side of the particle in line with the electric field, therefore producing an induced dipole in the particle. The direction of the dipole is either with the electric field if the particle polarizability is greater than the fluid or against the field if the particle polarizability is smaller. The dipole magnitude and direction in general depend, by inference, not only on the particle internal properties but also the properties of the suspending fluid medium.

Determining an expression for the dipole moment of a spherical particle is an exercise in solving Poisson's equation using spherical harmonics which can be found in references [5]. The solution of the electrical potential in spherical polar coordinates is

$$\phi_m = Er\cos\theta\left[\frac{a^3}{r^3}\left(\frac{\tilde{\varepsilon}_p - \tilde{\varepsilon}_m}{\tilde{\varepsilon}_p + 2\tilde{\varepsilon}_m}\right) - 1\right] \qquad (21)$$

in the fluid medium (indicated by the subscript m) and

$$\phi_p = -\frac{3\tilde{\varepsilon}_m}{\tilde{\varepsilon}_p + 2\tilde{\varepsilon}_m} Er\cos\theta \qquad (22)$$

inside the particle (indicated by subscript p). Here, a is the radius of the particle. Comparison of Eq. 21 with the potential of a dipole:

$$\phi_{\text{dipole}} = \frac{qd\cos\theta}{4\pi\varepsilon_m r^2} \qquad (23)$$

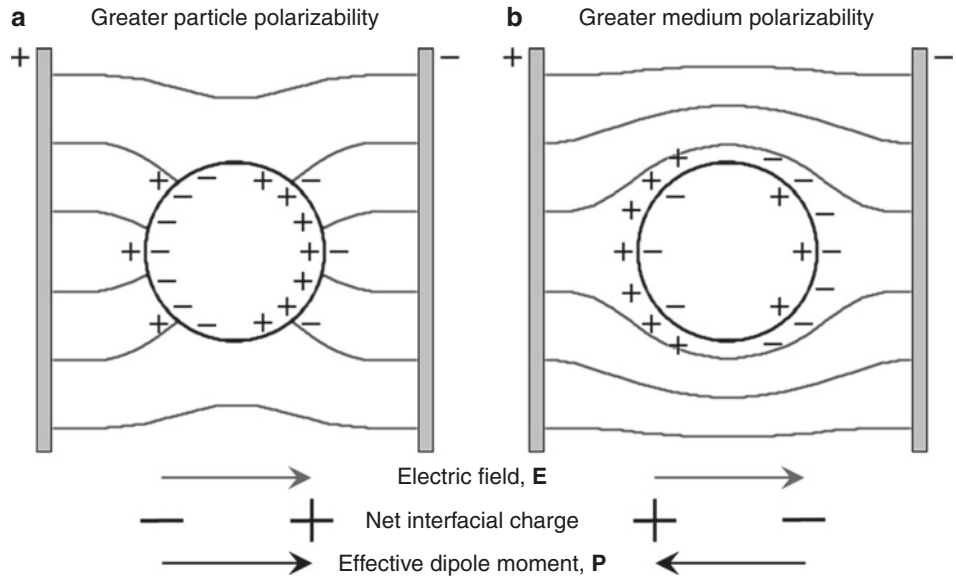
shows that the electrical potential in the medium is the sum of the original uniform field and a dipole moment $\mathbf{p}$ where

$$\mathbf{p} = 4\pi\varepsilon_m\left(\frac{\widetilde{\varepsilon}_p - \widetilde{\varepsilon}_m}{\widetilde{\varepsilon}_p + 2\widetilde{\varepsilon}_m}\right)a^3\mathbf{E} \qquad (24)$$

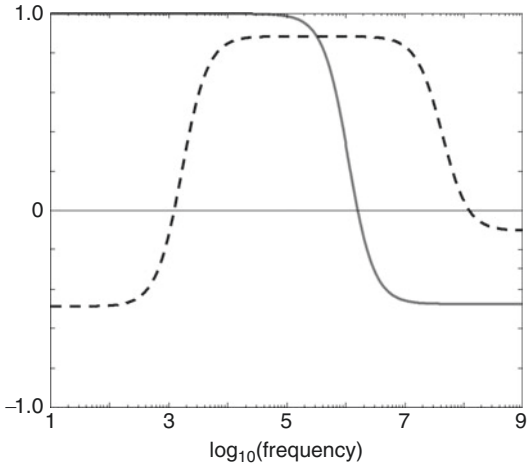
This is the effective dipole moment of the sphere, with a size that is frequency dependent and undergoes a dispersion because of Maxwell-Wagner interfacial polarization. The effective polarizability of a spherical dielectric particle is

$$\tilde{\alpha} = 3\varepsilon_m\left(\frac{\widetilde{\varepsilon}_p - \widetilde{\varepsilon}_m}{\widetilde{\varepsilon}_p + 2\widetilde{\varepsilon}_m}\right) \qquad (25)$$

and the frequency dependency of the dipole and the effective polarizability is described by the Clausius-Mossotti factor:



Dielectrophoresis, Fig. 3 Electric field lines, charges (distribution and net) at dielectric interfaces, electric fields, and effective dipole moment vectors for when a dielectric particle in suspension is subjected to a uniform electric field. The effective dipole moment will be aligned (**a**) with or (**b**) against the applied electric field, depending on whether the particle is more polarizable or less polarizable than the suspending medium



Dielectrophoresis, Fig. 4 Variations with frequency of the Clausius-Mossotti factor for two different particle types. Dispersions are due to Maxwell-Wagner interfacial polarization

$$\tilde{f}_{CM} = \frac{\tilde{\varepsilon}_p - \tilde{\varepsilon}_m}{\tilde{\varepsilon}_p + 2\tilde{\varepsilon}_m} \tag{26}$$

The variation with frequency of the Clausius-Mossotti factor for two different particle types is shown in Fig. 4.

Translational Force on a Dipole in a Nonuniform Field

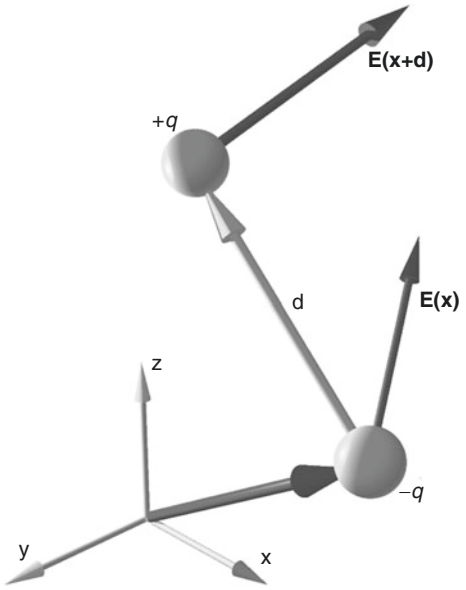
The simplest way to assign a force to a particle in a field is to consider the force exerted on a dipole by a nonuniform electric field [3, 4]. The dipole, shown in Fig. 5 (where the two equal and opposite charges $+q$ and $-q$ are separated by the vector $\mathbf{d}$), experiences a net force due to the nonuniform electric field $\mathbf{E}$ due to the fact that the field at and therefore the force on the two poles of the dipole are not equal. The force can be represented as a sum of the forces on the two charges:

$$\mathbf{F} = q\mathbf{E}(\mathbf{x} + \mathbf{d}) - q\mathbf{E}(\mathbf{x}) \tag{27}$$

where $\mathbf{x}$ can be taken to be the position vector of $-q$ without loss of generalization. This equation can be simplified if $|\mathbf{d}|$ is much smaller than $|\mathbf{x}|$ by expanding the electric field using a Taylor series about $\mathbf{x}$ to give

$$\mathbf{E}(\mathbf{x} + \mathbf{d}) = \mathbf{E}(\mathbf{x}) + d \cdot \nabla\mathbf{E}(\mathbf{x}) + \ldots \tag{28}$$

where all terms of order d^2 or higher have been neglected. If the dipole moment $\mathbf{p} = q\mathbf{d}$ remains



Dielectrophoresis, Fig. 5 The charges of the dipole $\mathbf{p} = Q\mathbf{d}$ in a nonuniform field $\mathbf{E}$ experience a different value of the electric field and as a result a different Coulomb force

finite in the limit $|\mathbf{d}| \rightarrow 0$, then substituting Eq. 28 into 27 gives the first order force on an infinitesimal dipole as

$$\mathbf{F} = \mathbf{p} \cdot \nabla \mathbf{E} \tag{29}$$

Only in a nonuniform field does this force have a nonzero net value.

Dielectrophoretic Force

The resulting movement of particles is referred to as dielectrophoresis (DEP) and the force referred to as the dielectrophoretic force [3, 4]. Equation 29 is restricted to cases where the gradient of the field across the particle can be assumed to be constant, often referred to as the dipole approximation. In cases where the field is very nonuniform across a distance comparable in size to the particle, the higher order terms can no longer be ignored.

In the dipole approximation, the dielectrophoretic force on a particle is found by substituting the effective dipole moment, Eq. 24 into Eq. 29 to give

$$\mathbf{F}_{DEP} = \upsilon\alpha \mathbf{E} \cdot \nabla \mathbf{E} \tag{30}$$

where υ is the volume of the particle and α is the effective polarizability (in this initial case, a static polarizability). This can be rewritten using a vector identity as

$$\mathbf{F}_{DEP} = -\frac{1}{2}\upsilon\alpha\nabla(\mathbf{E} \cdot \mathbf{E}) = -\frac{1}{2}\upsilon\alpha\nabla|\mathbf{E}|^2 \tag{31}$$

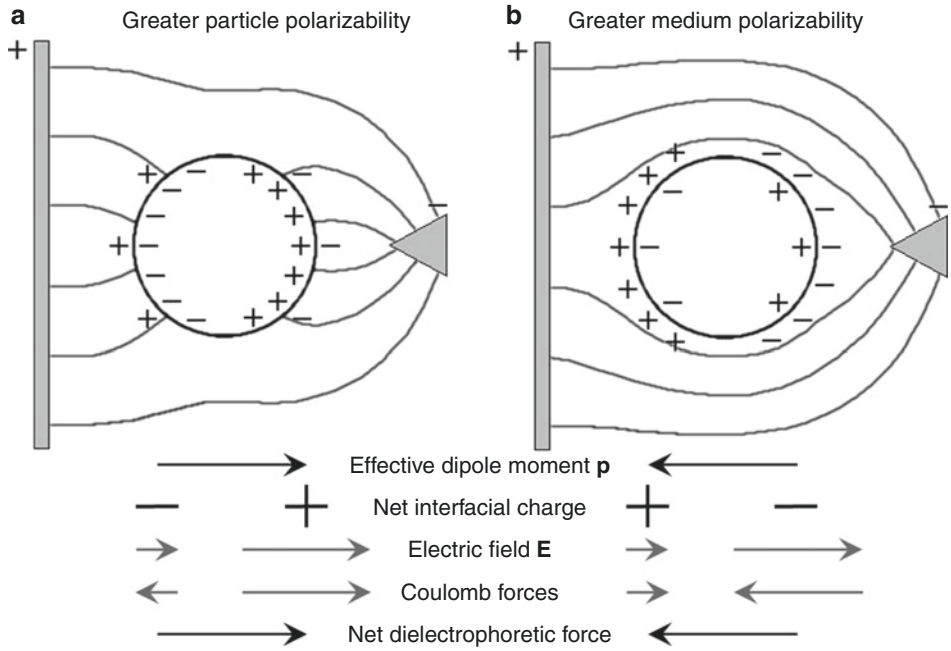
Since the effective polarizability of the particle can be either positive or negative, the dielectrophoretic force can act either in the direction of increasing field strength or in the opposite direction. Where particles are attracted toward regions of high field strength, the movement is referred to as *positive dielectrophoresis*. The opposite situation, where particles are repelled from high field regions, is called *negative dielectrophoresis*. The difference in the physical mechanism is illustrated in Fig. 6, where Coulomb forces on the two poles of the induced dipole are different owing to the nonuniform electric field. The force on the pole in the strongest electric field is the strongest but is toward high fields when the particle polarizability is greater (positive DEP) and away from high fields when the medium polarizability is greater (negative DEP).

Dielectrophoresis in an AC Field

Equation 31 is derived for a static electric field [3]. Deriving the force for an AC field requires an extension to account for the frequency and phase aspects associated with the time varying charging. This extension also exposes additional force terms that produce related movement in particles.

Assuming an applied potential of a single frequency ω, then the time-dependent values in the system can be represented using phasors. An arbitrary, harmonic potential can be defined as $\phi(\mathbf{x}, t) = \mathrm{Re}\left[\tilde{\phi}(\mathbf{x})e^{i\omega t}\right]$ where the tilde indicates the complex phasor $\left(\tilde{\phi} = \phi_R + i\phi_I\right)$ and Re[...] indicates the real part of. The electric field is then given by $\mathbf{E}(\mathbf{x}, t) = \mathrm{Re}\left[\tilde{\mathbf{E}}(\mathbf{x})e^{i\omega t}\right]$ where the vector $\tilde{\mathbf{E}} = -\nabla\tilde{\phi} = -(\nabla\phi_R + i\nabla\phi_I)$ is the corresponding phasor. The dipole moment of the particle is therefore $\tilde{\mathbf{p}} = \upsilon\tilde{\alpha}\mathbf{E}e^{i\omega t}$ and the time-averaged force on the particle is

$$\langle \mathbf{F}_{DEP} \rangle = \frac{1}{2}\mathrm{Re}\left[(\tilde{\mathbf{p}} \cdot \nabla)\mathbf{E}^*\right] \tag{32}$$



Dielectrophoresis, Fig. 6 Electric field lines, charges (distribution and net) at dielectric interfaces, directions of the applied electric field, effective dipole moment, Coulomb forces, and net dielectrophoretic force for when a dielectric particle in suspension is subjected to a nonuniform electric field. (**a**) When the particle is more polarizable than the suspending medium, the effective dipole moment is aligned with the applied electric field and the DEP force acts to move the particle toward the region of highest electric field intensity. (**b**) When the particle is less polarizable than the suspending medium, the effective dipole moment is aligned against the applied electric field and the DEP force acts to move the particle away from the region of highest electric field strength

where * indicates complex conjugate.

If the phase is constant across the system, the field phasor can be assumed to be real without loss of generality (i.e., $\tilde{\mathbf{E}} = \mathbf{E} = -\nabla\phi_R$). The time-averaged force is then

$$\langle \mathbf{F}_{DEP} \rangle = \frac{1}{2} \upsilon \mathrm{Re}[\tilde{\alpha}](\mathbf{E} \cdot \nabla)\mathbf{E} \tag{33}$$

Using vector identities and the fact that the field is irrotational gives

$$\langle \mathbf{F}_{DEP} \rangle = \frac{1}{4} \upsilon \mathrm{Re}[\tilde{\alpha}] \nabla(\mathbf{E} \cdot \mathbf{E}) \tag{34}$$

This is the dielectrophoretic force, which is commonly written as

$$\langle \mathbf{F}_{DEP} \rangle = \frac{1}{4} \upsilon \mathrm{Re}[\tilde{\alpha}] \nabla |\mathbf{E}|^2 \tag{35}$$

Often the field is quoted using *root-mean-square* (*rms*) values so that Eq. 35 becomes

$$\langle \mathbf{F}_{DEP} \rangle = \frac{1}{2} \upsilon \mathrm{Re}[\tilde{\alpha}] \nabla |\mathbf{E}_{rms}|^2 \tag{36}$$

This is also the expression for the force in a DC field. The field term in Eq. 35 is the magnitude of the phasor and the difference between the two equations is the factor of a half. When using AC fields, it is therefore extremely important to understand whether you are using amplitude or RMS values for the electric field. Inspection of these equations shows that the dielectrophoretic force depends on the volume of the particle and the gradient of the field magnitude squared. The force also depends on the real part of the effective polarizability.

This can be quantified for a spherical particle, where again the variation in the magnitude of the

force with frequency is given by the real part of the Clausius-Mossotti factor. By substitution of the effective polarizability of a sphere (Eq. 25), the full expression for the time-averaged DEP force on a spherical particle is

$$\langle \mathbf{F}_{DEP} \rangle = \pi \varepsilon_m a^3 \mathrm{Re}\left[\frac{\widetilde{\varepsilon}_p - \widetilde{\varepsilon}_m}{\widetilde{\varepsilon}_p + 2\widetilde{\varepsilon}_m}\right] \nabla |\mathbf{E}|^2 \qquad (37)$$

Examination of this equation shows that the magnitude of the force depends on the particle volume (a^3), the permittivity of the suspending medium, and the gradient of the field strength squared. The real part of the Clausius-Mossotti factor defines the frequency dependence and direction of the force. Expanding on the previous description, positive DEP occurs if the polarizability of the particle is greater than the suspending medium ($\mathrm{Re}\left[\tilde{f}_{CM}\right] > 1$) and the particle moves toward regions of high electric field strength. Negative DEP occurs if the polarizability of the particle is less than the suspending medium ($\mathrm{Re}\left[\tilde{f}_{CM}\right] < 1$) and the particles are repelled from regions of high field strength.

Dielectrophoresis in a Field with a Spatially Dependent Phase

For a general AC field, such as that generated by the application of multiple potentials of different phase, the derivation of the dielectrophoretic force is more involved [3]. The electric field in this case is $\mathbf{E}(\mathbf{x}, t) = \mathrm{Re}\left[\tilde{\mathbf{E}}(\mathbf{x}) e^{i\omega t}\right]$, where the vector $\tilde{\mathbf{E}} = -\nabla\tilde{\phi} = -(\nabla\phi_R + i\nabla\phi_I)$ is the corresponding *complex* phasor. The expression for the time-averaged force on the particle can then be derived from Eq. 32 with the vectors now consisting of complex components. The equation for the force is

$$\begin{aligned}\langle \mathbf{F}_{DEP} \rangle &= \frac{1}{2}\mathrm{Re}\left[(\tilde{\mathbf{p}} \cdot \nabla)\tilde{\mathbf{E}}^*\right] \\ &= \frac{1}{2}\upsilon \mathrm{Re}\left[\tilde{\alpha}(\tilde{\mathbf{E}} \cdot \nabla)\tilde{\mathbf{E}}^*\right]\end{aligned} \qquad (38)$$

Again, using vector identities and the facts that the electric field is irrotational and has zero divergence (Gauss's law), the force expression becomes

$$\begin{aligned}\langle \mathbf{F}_{DEP} \rangle = &\frac{1}{4}\upsilon \mathrm{Re}[\tilde{\alpha}]\nabla\left(\tilde{\mathbf{E}} \cdot \tilde{\mathbf{E}}^*\right) \\ &- \frac{1}{2}\upsilon \mathrm{Im}[\tilde{\alpha}]\left(\nabla \times \left(\tilde{\mathbf{E}} \times \tilde{\mathbf{E}}^*\right)\right)\end{aligned} \qquad (39)$$

where Im [. . .] is the imaginary part of the function. This can be rewritten as

$$\begin{aligned}\langle \mathbf{F}_{DEP} \rangle = &\frac{1}{4}\upsilon \mathrm{Re}[\tilde{\alpha}]\nabla\left|\tilde{\mathbf{E}}\right|^2 \\ &- \frac{1}{2}\upsilon \mathrm{Im}[\tilde{\alpha}]\left(\nabla \times \left(\mathrm{Re}\left[\tilde{\mathbf{E}}\right] \times \mathrm{Im}\left[\tilde{\mathbf{E}}\right]\right)\right)\end{aligned} \qquad (40)$$

where $\left|\tilde{\mathbf{E}}\right|^2 = \left|\mathrm{Re}\left[\tilde{\mathbf{E}}\right]\right|^2 + \left|\mathrm{Im}\left[\tilde{\mathbf{E}}\right]\right|^2$. If there is no spatially varying phase, the phasor of the electric field can be taken to be real (i.e., $\mathrm{Im}\left[\tilde{\mathbf{E}}\right] = 0$), the second term on the right-hand side of equation is then zero, and the expression for the force becomes that of Eq. 35. However, if there is a spatially varying phase, as in the case of traveling wave dielectrophoresis, the complete force expression must be used. The first term in the force expression depends on the frequency in the same manner as Eq. 35. The second term in the force equation depends, however, on the imaginary part of the effective polarizability or rather the imaginary part of the Clausius-Mossotti factor. This force is zero at high and low frequencies, rising to a maximum value at the Maxwell-Wagner interfacial relaxation frequency.

A general statement would then be that the *general* dielectrophoretic force is divided into two components, one of which is the DEP force and another that depends on the phase of the field:

$$\begin{aligned}\langle \mathbf{F}_{DEP} \rangle = &\underbrace{\frac{1}{4}\upsilon \mathrm{Re}[\tilde{\alpha}]\nabla\left|\tilde{\mathbf{E}}\right|^2}_{\text{Dielectrophoresis}} \\ &\underbrace{- \frac{1}{2}\upsilon \mathrm{Im}[\tilde{\alpha}]\left(\nabla \times \left(\mathrm{Re}\left[\tilde{\mathbf{E}}\right] \times \mathrm{Im}\left[\tilde{\mathbf{E}}\right]\right)\right)}_{\text{Travelling Wave Dielectrophoresis}}\end{aligned} \qquad (41)$$

Generally, for traveling wave dielectrophoresis to be effective, a frequency must be chosen where both the real part is substantially high **and** the imaginary part, which only has a value close to the relaxation frequency, is nonzero.

Applications of Dielectrophoresis

The critical feature of dielectrophoresis is that it requires field nonuniformities in order to create movement. The rapid development of dielectrophoresis as a method was made capable by the production of microelectrodes using semiconductor fabrication methods [3, 4].

In addition to this, as most biological particles are dielectrics, dielectrophoresis is particularly fitted to lab-on-a-chip applications where multiple processes involving characterization, manipulation, and separation of bioparticles are integrated onto a single chip. Dielectrophoresis is advantageous over other means of inducing particle motion in its noninvasive nature and easy integration onto micro-devices, not relying on any moving parts. Early biomedical applications of dielectrophoresis were limited to studying the response of cells to electric fields for characterization purposes. Nowadays, the technique has found widespread use for a variety of biotechnological and diagnostic applications. Examples include separation of human breast cancer cells from blood and trapping and manipulation of DNA – of importance to the development of point-of-care devices for detection and identification of pathogenic microorganisms, aggregation of hemispherical cells as a first step toward the creation of artificial stem cell microniches in vitro, and creation of engineered skin with artificial placodes of different sizes and shapes in different spatial patterns [6].

Dielectrophoresis has established its status as a versatile technique for separation of dielectric – biological and otherwise – particles in suspension. Through the frequency-dependent characteristics of dielectric particles, DEP has access to a wide range of particle properties and can be modified by the choice of suspending medium. Factors such as double layer and electro-convective effects can complicate DEP separation of dielectric particles [7]. Yet it has been shown that once these phenomena are correctly identified and incorporated in the design process, they can take their part in the toolbox of electrokinetic techniques for separation applications [8, 9].

Advances in fabrication technology have broadened the applicability of dielectrophoresis through realization of micro- and nanoelectrode geometries that are capable of generating electric fields strong enough to move nanoparticles such as viruses and chromosomes [10]. At the nanoscale, effects such as thermal and hydrodynamic forces find increased significance. Yet it has been shown that the "nuisances" are controllable and not significant enough to hinder nanoparticle motion due to dielectrophoresis [11].

Applications of dielectrophoresis are not limited to the realm of diagnostics involving biological particles. As early as 1924, Hatfield achieved separation of minerals using DEP; interestingly, the work has been pursued by other research groups up until very recently. Other examples where dielectrophoresis has been used for industrial applications include construction of a current-limiting fuse using DEP collection of conductive particles between two electrodes, depositing a patterned coating of a nanostructured material onto a substrate using positive DEP, and the fabrication of nanoscale devices, composed of movable components brought together in a fluid medium by exertion of the dielectrophoretic force – among other interactions [4].

With further developments in the design of suitable electrode geometry and in the modeling of phenomena that occur alongside dielectrophoretic motion of particles, DEP applications are expected to advance further, enabling manipulation and characterization of a yet wider range of particles.

Cross-References

- ▶ AC Electrokinetics of Nanoparticles
- ▶ Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly
- ▶ Dielectrophoresis of Nucleic Acids
- ▶ Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes
- ▶ Electric Field–Directed Assembly of Bioderivatized Nanoparticles

References

1. Pohl, H.A.: The motion and precipitation of suspensoids in divergent electric fields. J. Appl. Phys. **22**, 869–871 (1951)
2. Pohl, H.A.: Dielectrophoresis: The Behavior of Neutral Matter in Nonuniform Electric Fields. Cambridge University Press, Cambridge (1978)
3. Morgan, H., Green, N.G.: AC Electrokinetics of Colloids and Nanoparticles. Research Studies Press, Baldock (2003)
4. Jones, T.B.: Electromechanics of Particles. Cambridge University Press, Cambridge (1995)
5. Lorrain, P., Corson, D.R., Lorrain, F.: Electromagnetic Fields and Waves. W.H. Freeman, New York (1988)
6. Pethig, R.: Review article – dielectrophoresis: status of the theory, technology, and applications. Biomicrofluidics **4**, 022811 (2010)
7. Ramos, A., Morgan, H., Green, N.G., Castellanos, A.: AC electric-field-induced fluid flow in microelectrodes. J. Colloid Interface Sci. **217**, 420–422 (1999)
8. Gascoyne, P.R.C., Vykoukal, J.: Particle separation by dielectrophoresis. Electrophoresis **23**, 1973–1983 (2002)
9. Morgan, H., Hughes, M.P., Green, N.G.: Separation of submicron bioparticles by dielectrophoresis. Biophys. J. **77**, 516–525 (1999)
10. Hughes, M.P.: AC electrokinetics: applications for nanotechnology. Nanotechnology **11**, 124–132 (2000)
11. Ramos, A., Morgan, H., Green, N.G., Castellanos, A.: AC electrokinetics: a review of forces in microelectrode structures. J. Phys. D Appl. Phys. **31**, 2338–2353 (1998)

Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly

Shengqin Wang[1] and Yingxi Zhu[2]
[1]Institute of Materials Research and Engineering, A*STAR (Agency for Science, Technology and Research), Singapore, Singapore
[2]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, USA

Synonyms

AC electrokinetic manipulation; Dielectrophoresis; Nanoassembly

Definition

Dielectrophoresis (DEP)-directed assembly is an electrokinetic manipulation process, in which a force exerted on a particle subjected to a nonuniform electric field is employed to localize the particle in space with micro- or nanoscale precision. DEP force originated from the interparticle interaction due to the induced dipoles on the particles subject to the spatial gradient of the electric field. DEP is applicable for any particles, charged or uncharged, which makes it advantageous over electrophoresis that requires a net charge on the particle for Coulombic interaction to drive the particle in an applied electric field.

Background

DEP was first adopted as a term to describe the motion of particles under nonuniform electric fields in the 1950s by Pohl, who defined this effect as "the motion of suspended particles relative to that of the solvent resulting from polarization forces produced by an inhomogeneous electric field" [1, 2]. In 1966, Pohl and Hawk made the first demonstration of the use of DEP in the collection of living cells [3]. Despite the initial breakthrough, at that time the DEP technique was limited within the manipulation of cells and other microorganisms of the size in the micrometer scale due to the limitations in achieving high field gradients. For submicrometer-sized particles, a strong field gradient of the applied electric field is required in order to overcome the randomization of particles due to Brownian motion (or to provide sufficient energy to overcome the counteracting influence of diffusion). Thus, it was until the 1970s after the advancement of photo and electron beam lithographic techniques for fabricating microdevices [4] the DEP manipulation of submicrometer-sized particles was employed in sufficiently large field gradients by using microelectrodes. The use of microelectrodes also possesses two additional advantages: (1) low voltages can be effectively used to generate high electric field strengths across a wide range

of frequencies; and (2) due to the high surface-to-volume ratio of nanocolloids, Joule heating system can be easily and quickly dissipated.

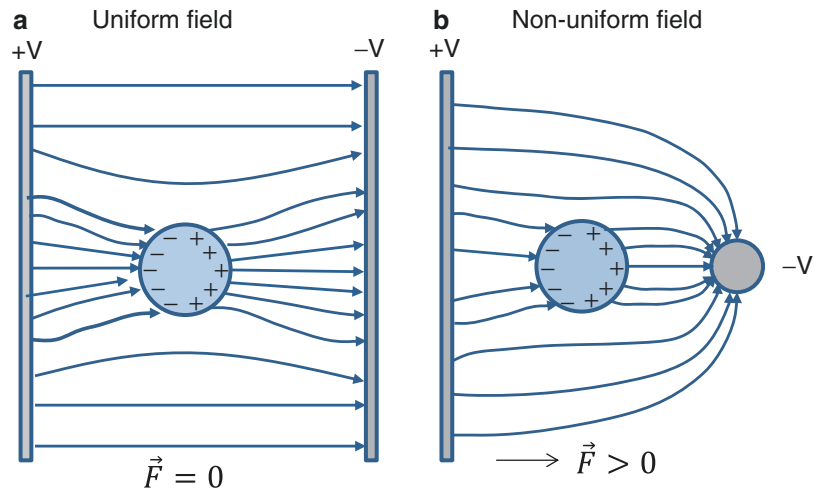
The early studies using microelectrode-based DEP were still largely concentrated on biotechnology and biomedical applications, including the responses of cells to DEP forces, cell sorting, bacterial manipulation, tissue engineering, and biosensors [4, 5]. Despite the early interest, there was little further attempt to work with nanomaterials until Washizu and co-workers demonstrated the manipulation and collection of biopolymers such as DNA and protein near planar microelectrodes in the 1990s [6, 7]. Since then, a wide range of nanocolloids such as viruses and submicrometer latex spheres have successfully been collected in microfluidic devices combined with DEP [8–13].

In the past decade, the application of DEP has been widely extended to a broad range of fields from biotechnology to materials science. Of particular interest, various materials in the nanometer scale, including polymer and lipid vesicles, carbon nanotubes, graphene, metal and metal oxide nanoparticles, and semiconducting nanowires and nanorods, have been manipulated and assembled into well-defined structures using the DEP technique [14–23]. DEP has been proven to be an efficient and versatile method for the directed assembly of various functional nanomaterials, thus providing a bridge connecting the top-down and bottom-up approaches that are the two independent basic methods used in the nanotechnology. The versatility, precision, and speed of DEP-directed assembly over the traditional bottom-up assembly approaches also impose great potentials to fabricate various nanodevices, such as microminiature biomedical device for diagnosis and electronic transistors.

Theory

DEP is a phenomenon in which a force is exerted on a polarizable particle in the nonuniform electric field. In an electric field, the particle is polarized to bear a dipole on its surface, thereby yielding a transient Coulombic attraction between the polarized particle and the electric field. As illustrated in Fig. 1, in a uniform field, the forces experienced by the positive and negative poles are balanced so that the time-averaged net force is zero. However, in a nonuniform field, the forces become imbalanced across the particle, leading to the time-averaged nonzero net force executed on the particle.

As particles are generally suspended in a medium, the polarization and resulting DEP response of the surrounding medium need to be taken into account for the DEP-driven motion of the particle in suspension. When the suspended particle has the higher polarizability than that of the medium, the induced dipole of the particle lies in the same direction as the electric field so that the resulting DEP force drives the particle toward the



Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly, Fig. 1 Schematic illustration of dielectrophoresis (*DEP*)

region of higher field strength, which is the so-called *positive* DEP (pDEP). On the contrary, if the polarizability of the particle is lower than that of the medium, the particle is driven toward the region of weaker electric field, which is the so-called *negative* DEP (nDEP). Between these two situations, the DEP force executed on the particle is zero if the particle has the same polarizability as the medium.

In an alternating-current (AC) electric field, the dipole moment of a particle can be written as

$$\tilde{\mathrm{p}} = v\tilde{\alpha}Ee^{i\omega t}, \tag{1}$$

where v is the particle volume, $\tilde{\alpha}$ is the polarizability, and E and ω are the magnitude and angular frequency of the electric field, respectively. As the time-averaged DEP force on the dipole is described as

$$\mathrm{F}_{DEP} = \frac{1}{2}Re[\,(\tilde{p}\cdot\nabla)E], \tag{2}$$

the DEP force is obtained as

$$\mathrm{F}_{DEP} = \frac{1}{4}vRe[(\tilde{\alpha})]\nabla|E|^2. \tag{3}$$

For a spherical particle, the effective polarizability in the medium is given by

$$\tilde{\alpha} = 3\varepsilon_m\tilde{f}_{CM}, \tag{4}$$

where

$$\tilde{f}_{CM} = \left(\frac{\widetilde{\varepsilon}_p - \widetilde{\varepsilon}_m}{\widetilde{\varepsilon}_p + 2\widetilde{\varepsilon}_m}\right) \tag{5}$$

is the Clausius-Mossotti (CM) factor and determined by the difference between the complex permittivity of the particle, $\widetilde{\varepsilon}_p$, and the medium, $\widetilde{\varepsilon}_m$. Substituting $\tilde{\alpha}$ for the effective polarizability of a sphere into Eq. 3, the full expression for the time-averaged DEP force is

$$\mathrm{F}_{DEP} = 2\pi\varepsilon_m a^3 Re\left[\left(\tilde{f}_{CM}\right)\right]\nabla|E|^2. \tag{6}$$

Apparently, the DEP force directly depends on the CM factor and the electric field gradient. As the complex permittivity of both the particle and the medium is frequency dependent as

$$\widetilde{\varepsilon} = \varepsilon - i\left(\frac{\sigma}{\omega}\right), \tag{7}$$

the real part of the CM factor reaches a low-frequency limiting value of $\frac{\sigma_p-\sigma_m}{\sigma_p+2\sigma_m}$, which relies solely on the conductivity of the particle, σ_p, and the medium, σ_m. Conversely, the high-frequency limiting value of the CM factor is $\frac{\varepsilon_p-\varepsilon_m}{\varepsilon_p+2\varepsilon_m}$, which is dominated by the permittivity of the particle, ε_p, and the medium, ε_m. Therefore, the CM factor and a spherical particle can vary between the low-frequency and high-frequency limiting values of 1 and -0.5, respectively. Apparently, pDEP occurs at $Re\left[\left(\tilde{f}_{CM}\right)\right] > 0$, while nDEP occurs at $Re\left[\left(\tilde{f}_{CM}\right)\right] < 0$.

In addition to the DEP force, a particle suspended in the medium also undergoes Brownian motion that can be treated as an effective randomizing force, which can be approximately described as

$$\mathrm{F}_{\mathrm{thermal}} = \frac{k_B T}{v^{\frac{1}{3}}}, \tag{8}$$

where k_B is the Boltzmann constant, T is the temperature, and v is the particle volume [4]. In order to precisely control the DEP-directed motion and particle assembly, the imposed DEP force must overcome the random thermal fluctuation. As F_{DEP} is proportional to v while F_{thermal} is inversely proportional to $v^{1/3}$, the thermal motion is especially more prominent for smaller particles so that a larger field gradient is required to direct their assembly. In addition to the thermal fluctuation, there are several other electrokinetic effects on the particle, which resulted from the motion of the solvent molecules under an electric field, such as electrothermal and electroosmotic forces [24]. These forces are also frequency dependent and occasionally need to be taken into account in the DEP assembling of particles.

In some practical situations, many particles of interest may not be simply in the symmetric spherical geometry and could also carry charges on

their surfaces. For particles of nonspherical shapes, several methods have been proposed to modify the derivation of the polarizability as well as the resulting DEP characteristics [25–28]. For charged particles in aqueous media, the DEP behavior can be greatly influenced by the relaxation and conduction of counterions in the electric double layer at the particle-water interaction [29]. The effect of the polarized double layer could become prominent for nanoparticles (NPs), where the thickness of the double layer can be comparable to or even larger the particle size. Consequently, Eq. 5 needs to be modified to take the charging of the electric double layer into account and describe the polarizability of the charged particle in aqueous media.

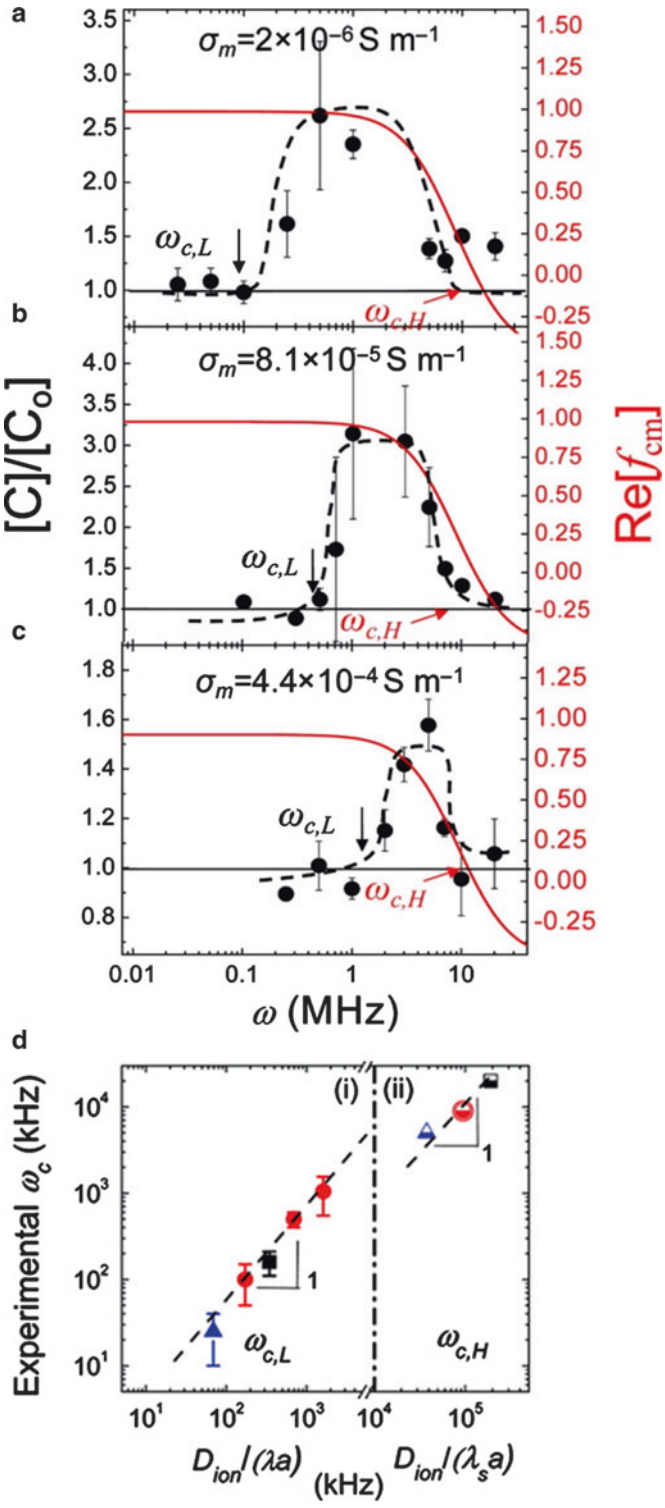
With the advancement of microfluidic technology and DEP theory, DEP has been recently employed to directly assemble various nanocolloids in addition to the traditional targets of cells, viruses, biopolymers, or microsphere particles. Below several representative examples with key findings in the progress of DEP-directed assembly are described.

Examples and Key Findings

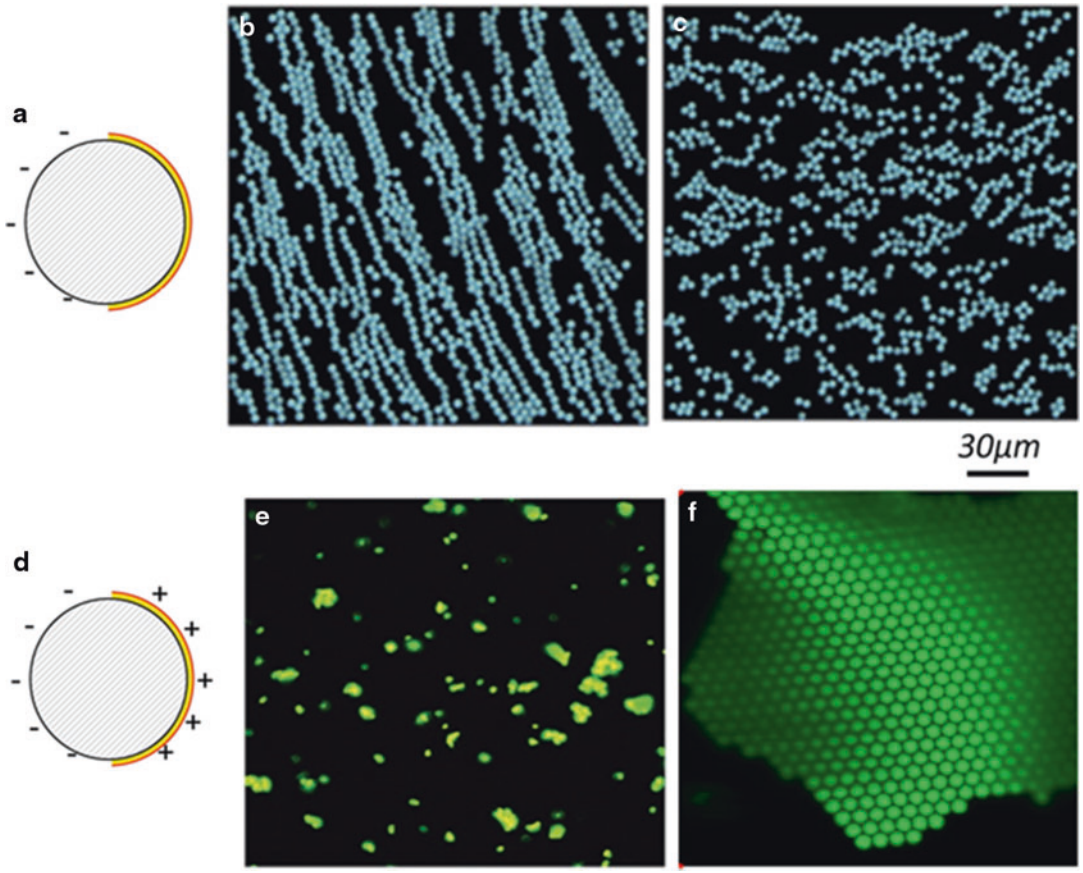
DEP Manipulation and Assembly of Nanoparticles and Janus Nanoparticles

Classical theories have predicted that due to the severely weakened F_{DEP} with reduced size, the deterministic DEP force on nanoparticles could hardly overcome Brownian motion; as a result, DEP becomes nearly inapplicable for colloidal manipulation and directed assembly at nanoscales [5]. However, recent experimental studies have demonstrated the assembly of functional nanowires and many other hierarchical nanostructures from nanoparticles in suspensions by DEP [22, 30–32]. Recent theoretical studies have endeavored to elucidate the induced dipole and intriguing DEP behaviors of submicrometer colloids of 50–500 nm in diameter by accounting the electric double layer effect for the interfacial polarization mechanism [29, 33]. It was suggested that submicrometer particles experience a DEP force stronger than that predicted by the classical Maxwell-Wagner (MW) theory, because AC field-induced tangential ionic currents in the Stern and diffuse layer result in high surface conductivity of dielectric colloidal particles. Moreover, when the particle radius, a, is further decreased to become comparable to the Debye length, λ (i.e., $a \approx \lambda$), the normal conductive field due to counterion migration in the diffuse layer cannot be negligible any more. As a consequence, nanoparticles should appear different DEP behaviors from micrometer or submicrometer particles.

To enable the in situ characterization of the nanoparticle assembly under nonuniform AC electric fields, fluorescence correlation spectroscopy (FCS) is employed, thanks to its high sensitivity and temporal resolution ($\sim 10^{-6}$ s). The typical concentration of fluorescence-labeled species is 10^{-12}–10^{-9} M in molar concentration or $\sim 10^{-3}$–1 particle or molecule over the ~ 1 μm [3] focal volume, which is 2–3 orders of magnitude lower than the common concentration limit required for microscopic visualization [34]. By employing FCS to detect fluorescence-labeled nanocolloids between microelectrodes, the DEP characteristics of nanoparticles of $a = 5$–25 nm were investigated, surprisingly showing a strong dependence on frequency of the applied AC electric field [35]. As shown in Fig. 2a–c, the nanoparticles surprisingly exhibit two distinct crossover frequencies in response to applied nonuniform AC electric fields between microelectrodes, one corresponding to the transition from zero or nDEP to pDEP and the other one corresponding to the transition from pDEP to nDEP, in contrast to one single characteristic DEP crossover frequency of micrometer-sized particles in suspensions. The results imply two distinct dipole relaxation time scales of a polarized nanoparticle under nonuniform AC fields. By examining the dependence of the two crossover frequencies with particle size and medium conductivity in Fig. 2d, it is found that the two distinct crossover frequencies resulted from the normal space charging in the diffuse layer and the counterion adsorption in the Stern layer, respectively. Because the normal conductivity and electric current in the diffuse layer are negligible for particles



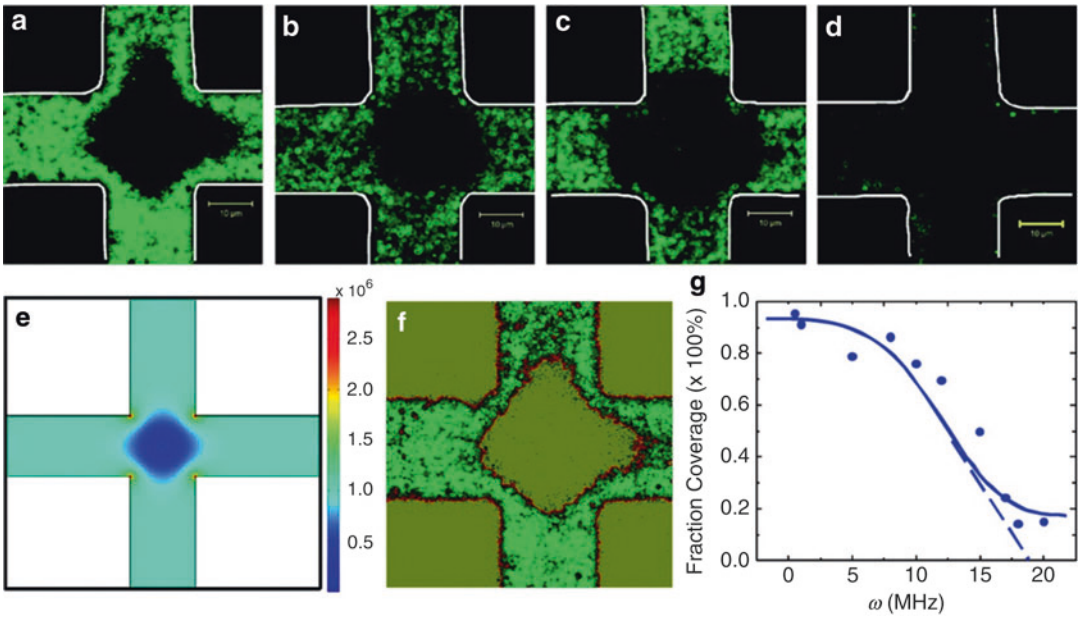
Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly, Fig. 2 The normalized concentration profiles, C/C_0, plotted on the left coordinate axis, against the frequency of the applied AC electric fields for PS latex nanoparticles of diameter of 20 nm suspended in aqueous media of varied medium conductivity, $\sigma_m = 2 \times 10^{-6}$ (**a**), 8.1×10^{-5} (**b**), and 4.4×10^{-4} (**c**) S/m. The theoretically estimated real part of the CM factor, $f_{CM}(\omega)$, is plotted in the red solid lines on the right coordinate axis. The dashed line is a guideline only and provides comparison with the theoretical $Re[f_{CM}(\omega)]$. (**d**) The collapse of experimental measured lower crossover frequency, $\omega_{c,L}$, with varied particle diameter, a and σ_m, suggesting the scaling of $\omega_{c,L}$ with $D_{ion}/(\lambda a)$, shown in the *left panel* (i) in solid symbols, and higher crossover frequency, $\omega_{c,H}$, with varied a, suggesting the scaling of $\omega_{c,H}$ with $D_{ion}/(\lambda_s a)$, shown in the right panel (ii) in semifilled symbols, for a = 10 nm (*black squares*), 20 nm (*semifilled red circles*), and 50 nm (*semifilled blue triangles*). In (**d**), D_{ion} is the ion diffusion constant, λ is the Debye length of the solution, and λ_s is the thickness of the Stern layer (Reproduced from Froude et al. [35], with permission from American Chemical Society)



Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly, Fig. 3 (**a**) Schematic illustration of half side-charged JPs. Micrograph of the assembly of the half side-charged JPs by DEP in aqueous suspensions of $\sigma_m = 0.0007$ (**b**) and 0.06 (**c**) S/m under applied AC field of $\omega = 1$ MHz and $V_{pp} = 10$ V between a pair of microelectrodes of planar configuration separated by 50 μm. (**d**) Schematic illustration of zwitterionic JPs. (**e**) Micrograph of zwitterionic JPs in deionized water in the absence of applied AC field, exhibiting small aggregates formed in the bulk suspension. (**f**) Micrographs demonstrating that zwitterionic JPs with opposite surface charges on each hemisphere can rapidly form irreversible crystalline assembly in deionized water under the applied nonuniform AC fields (Reproduced from Zhang and Zhu [37], with permission from American Chemical Society)

of $a \gg \lambda$, only the tangential conductive current in the double layer is often accounted, leading to one single crossover frequency for micrometer particles. It was also demonstrated that the DEP force imposed on nanoparticles could be stronger than that predicted by classic theory. The assembly of nanoparticles using DEP is therefore practical and has great potential applications for the fabrication of nanoparticle clusters or arrays.

The advanced DEP theory, which has considered the double layer polarization effect, can be also applied to understand and control the DEP behavior of anisotropic nanocolloids in suspensions. One of the simple model anisotropic particles is the Janus particles (JPs), which are biphasic colloids with two hemispheres of distinct chemical and physical properties. It is recently demonstrated that Janus latex particles, the half hemisphere of which is coated with a thin gold layer, experience positive DEP at all the frequency ranges from 25 kHz to 20 MHz and can form pearl chains of JPs by imposed DEP forces in nonuniform AC electric fields [36]. It is also found that the pearl chaining of JPs strongly depends on the medium conductivity [37]. As shown in Fig. 3a–c, shorter pearl chains of JPs



Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly, Fig. 4 Fluorescence micrographs of the DEP-induced assembly of liposome under applied AC field of constant amplitude, $V_{pp} = 5$ V, and varied frequency of (**a**) 500 kHz, (**b**) 3 MHz, (**c**) 15 MHz, and (**d**) 18 MHz. Panels (**a**–**d**) indicate the decreased liposome coverage between microelectrodes with increasing AC frequency, where the transition from positive to negative DEP takes place; (**e**) simulated AC electric field gradient between the quadruple microelectrodes by using Gauss's law with COMSOL; (**f**) the illustration of image analysis corresponding to the original fluorescence micrograph shown in panel (**a**) to determine the liposome coverage between microelectrodes based on the total areas of fluorescence intensity; (**g**) liposome surface coverage against the applied AC frequency for a plain liposome. A crossover frequency, ω_c, is thereby defined as the frequency extrapolated from the liposome coverage frequency plot where liposome coverage decreases to reach the saturation of typically less than 20 %. The determined ω_c herein for a plain liposome is ~18 MHz (Reproduced from Fronde and Zhu [39], with permission from American Chemical Society)

are assembled from the suspension at low medium conductivity of 0.015 S/m, in sharp contrast to random aggregates at higher medium conductivity of 0.06 S/m. It is interesting to observe that at high medium conductivity, JPs can undergo a transition from nDEP to pDEP, which is opposite to the pDEP to nDEP transition typically observed for the homogeneous particles. This unique DEP response is resulted from induced charge electroosmosis (ICEO) flow due to the imbalanced polarizability of two hemispheres in the JPs.

In addition to form one-dimensional chain-like assembly of JPs, zwitterionic JPs that bear opposite charges on two hemispheres can be driven to assemble into closely random packed crystals in large area by DEP as shown in Fig. 4d–f; in contrast, only small aggregates exist in the suspension in the absence of AC fields. The DEP-directed assembly of JPs with tunable surface chemistry can be used to guide the rapid synthesis of highly hierarchical structured materials from anisotropic building blocks.

DEP Manipulation and Assembly of Surfactant Micelles and Lipid Vesicles

DEP is applicable to manipulate complex nanocolloids, such as lipid and surfactant assemblies in the vesicle or micelle forms. In this session, the application of DEP on two model demonstrative systems, lipid unilamellar vesicle (liposome) and surfactant micelle, is discussed.

Liposome is a synthetic spherical vesicle composed of a unilamellar lipid bilayer, which has been widely used to administrate nutrients and

pharmaceutical drugs by utilizing their hollow, aqueous interior core capable of encapsulating various molecules [38]. The DEP assembly of liposomes of different surface treatments for suspension stability was recently investigated [39]. To prevent liposomes from fusion, two strategies, using charged nanoparticles and calcium phosphate coating, were employed, giving rise to distinct surface chemistry and topology.

Theoretically, for the complex structure of liposomes, the CM factor needs to be modified by considering the polarization across different regions of distinct permittivity and conductivity in each layer. Using the multi-shell model [25], the dielectric property of liposome is described as

$$\widetilde{\varepsilon}_{lip} = \widetilde{\varepsilon}_{LB} \frac{{\gamma_{12}}^3 + 2\left(\frac{\widetilde{\varepsilon}_{IM} - \widetilde{\varepsilon}_{LB}}{\widetilde{\varepsilon}_{IM} + 2\widetilde{\varepsilon}_{LB}}\right)}{{\gamma_{12}}^3 - \left(\frac{\widetilde{\varepsilon}_{IM} - \widetilde{\varepsilon}_{LB}}{\widetilde{\varepsilon}_{IM} + 2\widetilde{\varepsilon}_{LB}}\right)}, \tag{9}$$

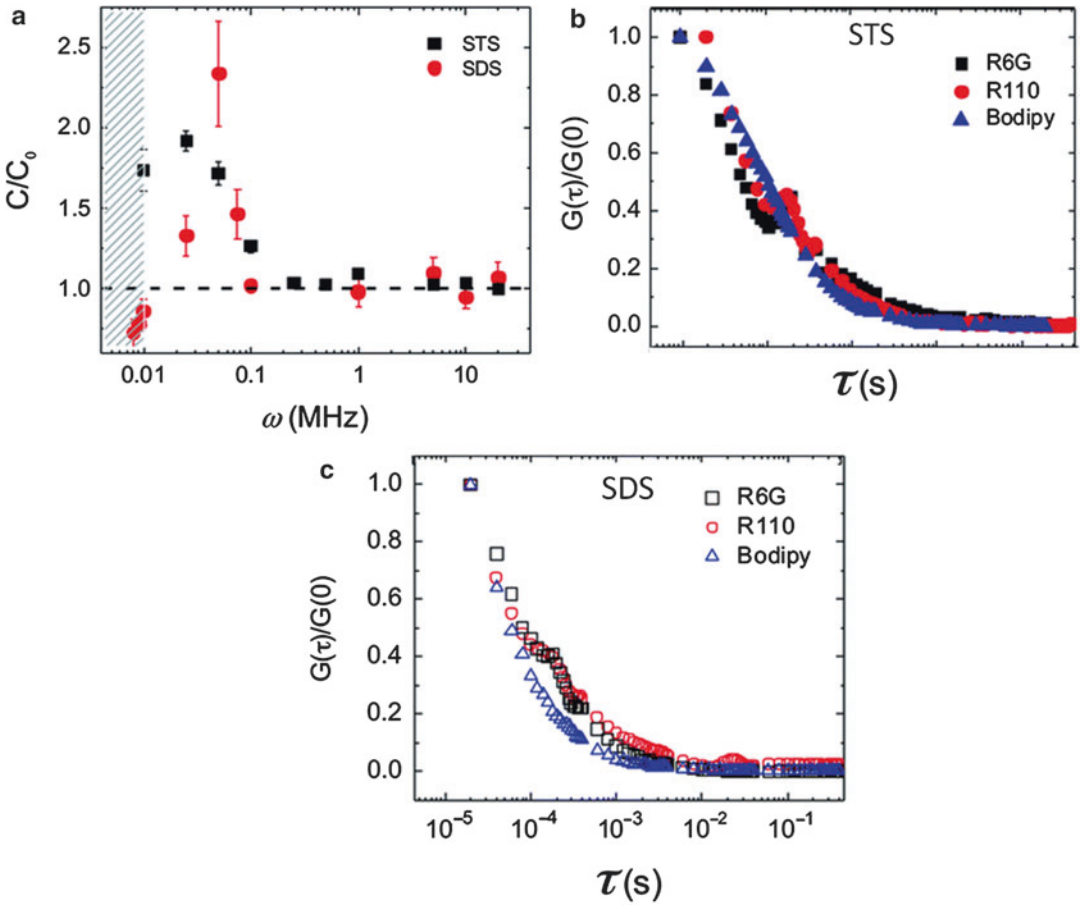
where $\widetilde{\varepsilon}_{LB}$ is the permittivity of the lipid bilayer, γ_{12} is a ratio of the outer to inner radius of the spherical shell, and $\widetilde{\varepsilon}_{IM}$ and $\widetilde{\varepsilon}_{LB}$ are the permittivity of the interior medium (IM) and lipid bilayer (LB), respectively. Using the classical MW theory, the CM factor for liposome could be deduced as

$$\tilde{f}_{CM} = \left(\frac{\widetilde{\varepsilon}_{lip} - \widetilde{\varepsilon}_{m}}{\widetilde{\varepsilon}_{lip} + 2\widetilde{\varepsilon}_{m}}\right). \tag{10}$$

It appears that in this shell model each interface separating the dielectric regions would contribute to the ultimate DEP behavior of the liposome in the suspension.

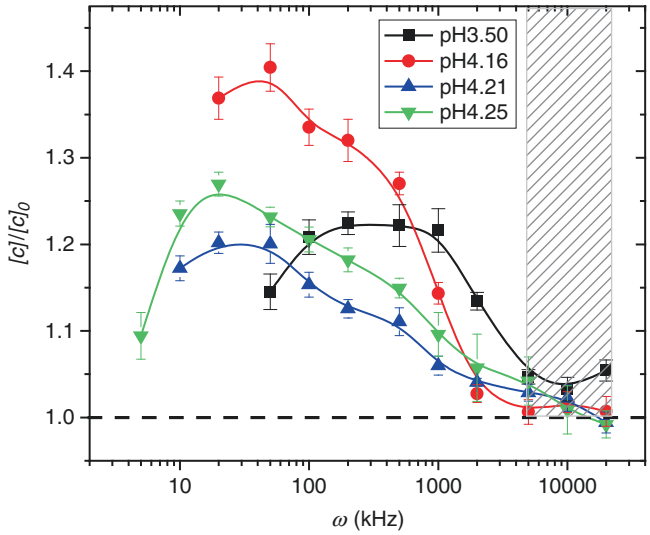
The key DEP behaviors of liposomes in aqueous buffer suspension between microelectrodes in a quadruple configuration are summarized in Fig. 4a–d. As the applied AC frequency is increased from 100 kHz to 20 MHz, collected liposomes at the microelectrode interfaces repel away from the microelectrode and diffuse into the aqueous suspension as indicated by the reduced fluorescence intensity. The consistence of fluorescence intensity with the simulated AC electric field gradient between the quadruple microelectrodes shown in Fig. 4e confirms the pDEP response of liposomes at low AC frequency but zero DEP at high frequency as a hallmark of a shell structure. The crossover frequency for the transition from pDEP to zero DEP is found to be ~18 MHz. The good agreement between the measured crossover frequency and the theoretical prediction using Eqs. 9 and 10 indicates that the electrokinetic behavior of the liposome can be well described by the multi-shell model. In addition, based on the study of the role of surface coating chemistry and structure on the DEP response of liposomes, it is found that the liposome surface construct, rather than surface conductivity, plays a critical role in the DEP characteristics of a shelled nanocolloid, which is much different from that of a solid dielectric particle. It is noted that when the liposome particle has a highly conductive interior, surface conductivity changes exhibit little effect on the DEP behavior of the liposomes even in the conductivity-controlled low-frequency region. Changes in the permittivity of multilayer coatings, on the other hand, can be used to effectively tune the polarizability of the shelled particle to a much greater extent, significantly altering the high-frequency DEP response of liposomes in suspension.

Surfactant-assembled micelles, which have been used in many chemical, cosmetic, and pharmaceutical products, represent another complex nanocolloidal construct. With micelles composed of sodium tetradecyl sulfate (STS) or sodium dodecyl sulfate (SDS) and fluorescence probe encapsulated in the interior hydrophobic core, the structural dynamics and assembly of micelles by DEP in nonuniform AC electric field were recently reported [40]. In response to applied electrical fields of frequency of 10–500 kHz, both micelles exhibit a pDEP behavior with increased micelle concentration in the strong field regions, as indicated in Fig. 5a. The estimated DEP crossover frequency from pDEP to nDEP is found approximately in the range of 200–500 kHz and agrees well with the advanced DEP theory in which the Stern layer conductance is accounted, indicating that the micelles have similar DEP behaviors as the charged nanoparticles. Most surprisingly, at a further



Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly, Fig. 5 (**a**) The normalized concentration profiles, C/C_0, plotted on the left coordinate axis, against AC frequency for STS-R6G (*black squares*) and SDS-R6G (*red circles*) micelles in salt-free aqueous suspension under applied AC field of constant amplitude, $V_{pp} = 10$ V, across microelectrodes in a quadruple configuration. Normalized autocorrelation curves, $G(\tau)/G(\tau = 0)$, for STS micelles (**b**) and SDS (**c**) micelles encapsulated with R6G (*black squares*), R110 (*red circles*), and Bodipy (*blue triangles*), both in salt-free aqueous suspensions under AC fields of $V_{pp} = 10$ V and $\omega = 5$ kHz. The autocorrelation curves obtained with micelles encapsulated with R6G or R110 exhibit the similar unusual "double plateau," indicating the existence of the free encapsulated molecules in solution but not with the ones encapsulated with Bodipy (Reproduced from Goodrich et al. [40], with permission from Royal Chemical Society)

low-frequency range of 5–10 kHz, a unique DEP response behavior of the micelle was observed: the correlation curves recorded by FCS manifest the existence of two distinct species of fluorescent entities in the suspension as indicated in Fig. 5b and c, suggesting the release of encapsulates from micelles induced by AC fields; in addition, such field-driven release of encapsulates exhibits a strong dependence on the chemistry of its encapsulated molecules. When charged molecules, for instance, Rhodamine 6G or Rhodamine 110, are encapsulated in the micelles, the anomalous release of the probe encapsulated in the micelle is observed but no release is found with neutral probes, such as Bodipy, encapsulated in micelles. Considering the difference between the charged and neutral molecular encapsulates, it is speculated that the conductivity of the interior regime of the micelle is the determining factor for the field-controlled release due to electromigration of charged probes. A future study with quantification, which also demands the advancement of



Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly, Fig. 6 The normalized concentration profiles, C/C_0, of a synthetic polyelectrolyte, poly(2-vinyl pyridine) (P2VP), against AC frequency in salt-free aqueous solutions of varied pH under applied nonuniform AC fields of $V_{pp} = 20$ V across a pair of planar microelectrodes separated by 30 μm. The dashed zone highlights the DEP behaviors of P2VP at AC frequencies higher than the crossover frequency (Reproduced from Wang and Zhu [41], with permission from American Institute of Physics)

molecular AC polarization theory, is desired to fully understand this intriguing AC electrokinetic characteristic in the complex micelle system.

DEP Manipulation of Synthetic Polyelectrolytes

The last demonstrative case involves synthetic polyelectrolytes, which are polymers bearing charged groups in their repeating units. By employing FCS, the DEP behavior of a synthetic and positively charged polyelectrolyte, poly (2-vinyl pyridine) (P2VP), was investigated at varied frequency of applied AC electric fields [41]. As P2VP is an annealed polyelectrolyte whose charge density is not constant but tunable by pH or applied electric fields, the local proton concentration in the vicinity of a PE chain can be significantly different from that in the bulk due to counterion condensation and thereby can be modified by external electric fields. In response to a homogeneous AC field, an induced swollen coil-to-collapsed globule conformational transition is observed and reversible upon AC frequency sweep when applied field strength exceeds a critical value, yet a hysteresis in AC frequency-dependent conformation is also observed [42]. However, no conformational change of P2VP is observed by imposed DEP force in response to spatially nonuniform AC fields of similar electric field and frequency, and instead a net flow and an accumulation of P2VP toward high field regions near two microelectrodes are observed, in sharp contrast to the response of P2VP under uniform AC fields that cause isotropic depletion of counterions and subsequent polarization normal to a P2VP chain. As shown in Fig. 6, in nonuniform AC fields within the frequency window of 100 kHz to 5–10 MHz, P2VP exhibits the pDEP response and becomes concentrated at the microelectrode interface. It is thus postulated that due to the field-driven net motion of both P2VP chains and localized counterions, the depletion of localized counterions away from a P2VP chain is significantly weakened or suppressed, resulting in negligible counterion concentration-induced dipolar energy on a P2VP chain, leading to no conformational transition. Nevertheless, the DEP behavior of P2VP at high frequencies above 10 MHz shows a strong dependence on the pH of the solution: at lower

pH, a zero to positive DEP transition appears as increasing the frequency across 10 MHz; however, no clear transition is observed at higher solution pH. The abnormal pDEP behavior for the synthetic polyelectrolyte at high frequencies is unpredicted by the classical MW theory. However, this abnormal pDEP of the synthetic polyelectrolyte is found similar to the reported DEP response for DNAs with short chains [43]. It is argued that the flexibility of the polymer chain could be an important factor for its electrokinetic properties in solutions. The distinct conformational and transport behaviors of a single annealed PE chain under varied microelectrode layouts can be further exploited in lab-on-chip device design to effectively manipulate and transport polyelectrolytes including DNA and protein under the application of varied AC fields.

Perspectives

Over the last decades, rapid and continuous development of lithography techniques has made enormous progress in the fabrication of metal and/or insulator structures. Nowadays, the fabrication of complex and custom designed microelectrodes to generate strong electric field gradient as well as tunable field distributions becomes feasible and nearly a routine job in the lab. On the other hand, nanomaterials, with their vast diversity in chemical composition, structure, and functionality, have been developed incredibly fast and reshaped the landscape of materials today. It turns out that the assembly of nanomaterials into hierarchical structure becomes a critical step to transfer their fascinating properties to advanced devices or products for broad applications. As a bridge linking between the top-down and bottom-up approaches, DEP has been demonstrated to be a robust method with the potential for the large-scaled integration of nanomaterials into nanoscaled functional devices. Future research could be considered along the following directions. First, DEP-directed assembly of various nanomaterials to fabricate functional devices is still of great interest, especially for the anisotropic and heterogeneous nanomaterials. To achieve the high controllability and efficiency of DEP-directed nanomaterial assembly, particular efforts, including the innovative design of microelectrode arrays and optimal operational parameters, need to be taken care of. Second, theoretical modeling on the electrokinetics of nanomaterials is further advanced. The complexity arising from the polarization of condensed counterions, asymmetry and heterogeneity in surface chemical composition, structure, and geometry of nanomaterials, needs to be addressed explicitly to predict the DEP characteristic of structured nanomaterials. Finally, the scaled-up nano-manufacturing based on DEP and other electrokinetic effects could be promising. The integration of DEP with the current micro- and nano-fluidic device could lead to further accelerating the nanomaterial assembly process toward practical industrial applications.

Cross-References

- ▶ AC Electrokinetics
- ▶ AC Electrokinetics of Colloidal Particles
- ▶ AC Electrokinetics of Nanoparticles
- ▶ Dielectrophoresis
- ▶ Dielectrophoretic Assembly
- ▶ Electric Double Layer Capacitor
- ▶ Induced Charge Electroosmosis
- ▶ Nanofluidics
- ▶ Nanoparticle Self-Assembly

References

1. Pohl, H.A.: The motion and precipitation of suspensoids in divergent electric fields. J. Appl. Phys. **22**, 869–871 (1951)
2. Pohl, H.A.: Some effects of nonuniform fields on dielectrics. J. Appl. Phys. **29**, 1182–1188 (1958)
3. Pohl, H.A., Hawk, I.: Separation of living and dead cells by dielectrophoresis. Science **152**, 647–649 (1966)
4. Pohl, H.A.: Dielectrophoresis. Cambridge University Press, Cambridge (1978)
5. Morgan, H., Green, N.G.: AC Electrokinetics: colloids and nanoparticles. Research Studies Press, Baldock/Hertfordshire/England (2002)
6. Washizu, M., Kurosawa, O.: Electrostatic manipulation of DNA in microfabricated structure. IEEE Ind. Appl. Trans. **26**, 1165–1172 (1990)

7. Washizu, M., Suzuki, S., Kurosawa, O., Nishizaka, T., Shinohara, T.: Molecular dielectrophoresis of biopolymer. IEEE Ind. Appl. Trans. **30**, 835–843 (1994)
8. Chang, H.-C., Yeo, L.Y.: Electrokinetically-driven microfluidics and nanofluidics. Cambridge University Press, New York (2009)
9. Zhang, C., Khoshmanesh, K., Mitchell, A., Kalantarzadeh, K.: Dielectrophoresis for manipulation of micro/nano particles in microfluidic systems. Anal. Bioanal. Chem. **396**, 401–420 (2010)
10. Hoffman, P.D., Zhu, Y.X.: Double-layer effects on low-frequency dielectrophoresis-induced colloidal assembly. Appl. Phys. Lett. **92**, 224103 (2008)
11. Akin, D., Li, H., Bashir, R.: Real-time virus trapping and fluorescent imaging in microfluidic devices. Nano Lett. **4**, 257–259 (2003)
12. Grom, F., Kentsch, J., Müller, T., Schnelle, T., Stelzle, M.: Accumulation and trapping of hepatitis A virus particles by electrohydrodynamic flow and dielectrophoresis. Electrophoresis **27**, 1386–1393 (2006)
13. Hoffman, P.D., Sarangapani, P.S., Zhu, Y.: Dielectrophoresis and AC-induced assembly in binary colloidal suspensions. Langmuir **24**, 12164–12171 (2008)
14. Shekhar, S., Stokes, P., Khondaker, S.I.: Ultrahigh density alignment of carbon nanotube arrays by dielectrophoresis. ACS Nano **5**, 1739–1746 (2011)
15. Pethig, R.: Review article-dielectrophoresis: status of the theory, technology, and applications. Biomicrofluidics **4**, 022811 (2010)
16. Nocke, A., Wolf, M., Budzier, H., Arndt, K.-F., Gerlach, G.: Dielectrophoretic alignment of polymer compounds for thermal sensing. Sens. Actuators A **156**, 164–170 (2009)
17. Daeha, J., Chunder, A., Lei, Z., Saiful, I.K.: High yield fabrication of chemically reduced graphene oxide field effect transistors by dielectrophoresis. Nanotechnology **21**, 165202 (2010)
18. Li, X., Chin, E., Sun, H., Kurup, P., Gu, Z.: Fabrication and integration of metal oxide nanowire sensors using dielectrophoretic assembly and improved post-assembly processing. Sens. Actuators B **148**, 404–412 (2010)
19. Freer, E.M., Grachev, O., Duan, X., Martin, S., Stumbo, D.P.: High-yield self-limiting single-nanowire assembly with dielectrophoresis. Nat. Nanotechnol. **5**, 525–530 (2010)
20. Raychaudhuri, S., Dayeh, S.A., Wang, D., Yu, E.T.: Precise semiconductor nanowire placement through dielectrophoresis. Nano Lett. **9**, 2260–2266 (2009)
21. Barsotti, R.J., Vahey, M.D., Wartena, R., Chiang, Y. M., Voldman, J., Stellacci, F.: Assembly of metal nanoparticles into nanogaps. Small **3**, 488–499 (2007)
22. Gierhart, B.C., Howitt, D.G., Chen, S.J., Smith, R.L., Collins, S.D.: Frequency dependence of gold nanoparticle superassembly by dielectrophoresis. Langmuir **23**, 12450–12456 (2007)
23. Krupke, R., Hennrich, F., von Lohneysen, H., Kappes, M.M.: Frequency dependence of gold nanoparticle superassembly by dielectrophoresis. Science **301**, 344–347 (2003)
24. Castellanos, A., Ramos, A., Gonzalez, A., Green, N. G., Morgan, H.: Electrohydrodynamics and dielectrophoresis in microsystems: scaling laws. J. Phys. D-Appl. Phys. **36**, 2584–2597 (2003)
25. Irimajiri, A., Hanai, T., Inouye, A.: A dielectric theory of "multi-stratified shell" model with its application to a lymphoma cell. J. Theor. Biol. **78**, 251–269 (1979)
26. Green, N.G., Jones, T.B.: Numerical determination of the effective moments of non-spherical particles. J. Phys. D-Appl. Phys. **40**, 78–85 (2007)
27. Winter, W.T., Welland, M.E.: Dielectrophoresis of non-spherical particles. J. Phys. D-Appl. Phys. **42**, 045501 (2009)
28. Yang, C.Y., Lei, U.: Dielectrophoretic force and torque on an ellipsoid in an arbitrary time varying electric field. Appl. Phys. Lett. **90**, 153901 (2007)
29. Basuray, S., Chang, H.C.: Induced dipoles and dielectrophoresis of nanocolloids in electrolytes. Phys. Rev. E **75**, 060501 (2007)
30. Hermanson, K.D., Lumsdon, S.O., Williams, J.P., Kaler, E.W., Velev, O.D.: Dielectrophoretic assembly of electrically functional microwires from nanoparticle suspensions. Science **294**, 1082–1086 (2001)
31. Kumar, S., Seo, Y.K., Kim, G.H.: Manipulation and trapping of semiconducting ZnO nanoparticles into nanogap electrodes by dielectrophoresis technique. Appl. Phys. Lett. **94**, 153104 (2009)
32. Zheng, L.F., Li, S.D., Brody, J.P., Burke, P.J.: Manipulating nanoparticles in solution with electrically contacted nanotubes using dielectrophoresis. Langmuir **20**, 8612–8619 (2004)
33. Gagnon, Z., Chang, H.C.: Aligning fast alternating current electroosmotic flow fields and characteristic frequencies with dielectrophoretic traps to achieve rapid bacteria detection. Electrophoresis **26**, 3725–3737 (2005)
34. Elson, E.L., Magde, D.: Fluorescence correlation spectroscopy. I. Conceptual basis and theory. Biopolymers **13**, 1–27 (1974)
35. Froude, V.E., Godfroy, J.I., Wang, S.Q., Dombek, H., Zhu, Y.X.: Anomalous dielectrophoresis of nanoparticles: A sensitive and rapid characterization by single-particle laser spectroscopy. J. Phys. Chem. C **114**, 18880–18885 (2010)
36. Zhang, L., Zhu, Y.X.: Dielectrophoresis of Janus particles under high frequency AC-electric fields. Appl. Phys. Lett. **96**, 141902 (2010)
37. Zhang, L., Zhu, Y.X.: Directed assembly of janus particles under high frequency ac-electric fields: effects of medium conductivity and colloidal surface chemistry. Langmuir **28**, 13201–13207 (2012)
38. Felnerova, D., Viret, J.F., Gluck, R., Moser, C.: Liposomes and virosomes as delivery systems for antigens, nucleic acids and drugs. Curr. Opin. Biotechnol. **15**, 518–529 (2004)
39. Fronde, V.E., Zhu, Y.X.: Dielectrophoresis of lipid unilamellar vesicles (liposomes) of contrasting surface constructs. J. Phys. Chem. B **113**, 1552–1558 (2009)

40. Goodrich, V.E., Connor, E., Wang, S.Q., Yang, J.F., Zhao, J., Zhu, Y.X.: AC-electrokinetic manipulation and controlled encapsulate release of surfactant based micelles. Soft Matter **9**, 5052–5060 (2013)
41. Wang, S., Zhu, Y.: Manipulating single annealed polyelectrolyte chain under ac-electric fields: collapse versus accumulation. Biomicrofluidics **6**, 024116 (2012)
42. Wang, S.Q., Chang, H.C., Zhu, Y.X.: Hysteretic conformational transition of single flexible polyelectrolyte under resonance ac-electric polarization. Macromolecules **43**, 7402–7405 (2010)
43. Henning, A., Bier, F.F., Holzel, R.: Dielectrophoresis of DNA: Quantification by impedance measurements. Biomicrofluidics **4**, 022803 (2010)

Dielectrophoresis of Nucleic Acids

David J. Bakewell
Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, UK

Synonyms

Translational movement of polynucleotides in a nonuniform electric field

Definitions

1. **Dielectrophoresis** (DEP) is the translational motion of an electrically polarizable body by the action of an externally applied, spatially nonuniform, electric field. DEP can occur with an electrically neutral body in a spatially nonuniform field and depends on the *dielectric* properties of the body relative to the surrounding medium. The word *dielectrophoresis* derives from the Greek word "*phorein*" where a particle "is *carried* as a result of its *dielectric* properties" [1]. DEP is distinguished from **electrophoresis** that requires a net charge on the body for Coulombic movement to occur in an externally applied, spatially uniform, electric field. DEP is also distinguished from **electrorotation** (EROT) that refers to *rotational* motion of a body in a spatially inhomogeneous electric field.
2. **Nucleic acids** (NAs) are responsible for information storage, distribution, translation, and control throughout a living system, such as a cell. There are two types of NAs central to life: deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). DNA, discovered more than a century ago by Friedrich Miescher, is called a nucleic acid because it was extracted from the nuclei of cells (originally called "nuclein") and showed acidic chemical properties [2, 3]. Nowadays, it is known that DNA can also be extracted from organelles in a cell, e.g., mitochondria, or other life forms, e.g., viruses; nonetheless, the term "nucleic" remains. Both DNA and RNA are comprised of nucleotides that consist of a nitrogenous base, a pentose sugar, and a phosphate link. A third type of NA is not naturally occurring; rather it is a synthetically made NA, e.g., peptide nucleic acid (PNA).

Introduction

The first main exposition of dielectrophoresis (DEP) as an important method for moving an electrically neutral body was by Herbert Pohl [4], although the phenomenon was known as far back as 600 BC. DEP of a body suspended in a medium is usually achieved by applying an electric potential between electrodes immersed in low conductivity electrolyte medium. Application of an electric potential can yield an electric field that induces dipoles in the body. If the field is sufficiently nonuniform, the interaction of the dipole charges yields a net force on the body and causes it to move. Dependence on the dielectric polarizability of the body, rather than reliance on net electrical charge, is a key reason for considerable interest in using DEP for noncontact movement of micro- and nanosized bodies. A corollary of the motion based on dielectric properties is that radio-frequency (RF) electric fields can be used to move bodies by DEP and hence avoid problematic hydrolysis at the electrode-medium interface. Developments in microfabrication methods that allow the generation of high electric field nonuniformities using RFs and low voltages have

enabled DEP to become a popular method for moving bodies in microenvironments, e.g., lab-on-chip (LOC) and micro-total analysis system (μTAS).

There is considerable interest in using DEP for moving nucleic acids (NAs) which are large molecules that play a central role in life. Both types of naturally occurring NAs, DNA and RNA, possess net negative electrical charge, and electrophoresis is a commonly used tool in molecular biology laboratories today. DEP is less well known in current biology laboratories so that the focus of this entry lies with the DEP motion of these NAs in their native suspension at room temperature (25 °C), i.e., water. Research into DEP in general, and in particular DEP of DNA, has dramatically increased in recent years, forming about 10 % of the several hundreds of journal articles published per annum on DEP. A database search (Web of Science™ v.5.16, 2015, Thomson Reuters), for example, shows that in the last decade there are more than thirty journal publications per annum involving DEP of DNA. The topic is of interest to nanotechnologists due to the central role DNA plays in life sciences and its properties as a nano-construction material. Progress in this field has been discussed in highly informative reviews [1, 5, 6].

Conversely, there are comparatively little published data on DEP of RNA and no journal reports solely on DEP of PNA. This is perhaps because this RNA is more complicated to prepare and handle than DNA; for instance, RNA is easily degraded (e.g., by an enzyme, RNase) without extra laboratory precautions and facilities. This is somewhat unfortunate because RNA, for many scientists and technologists, is more important than DNA (see also "RNA world" hypothesis [3]). The following sections discuss fundamental structural, chemical, and physical properties of NAs for DEP. This is followed by a DEP of DNA application example and a summary of broader contributions including a reference to the use of PNA in experiments. A brief summary of DEP of RNA follows. Due to a need for concise referencing, many of the names of individual authors and groups are referred to as listed in reviews and books, rather than individual citations.

Fundamentals

This section describes the fundamentals of NAs, their structural and electrical properties in water, their inducible polarizability, and the basic principles of DEP. The focus is on DNA since there are far fewer DEP studies on RNA and PNA and it is likely similar electrokinetic principles apply.

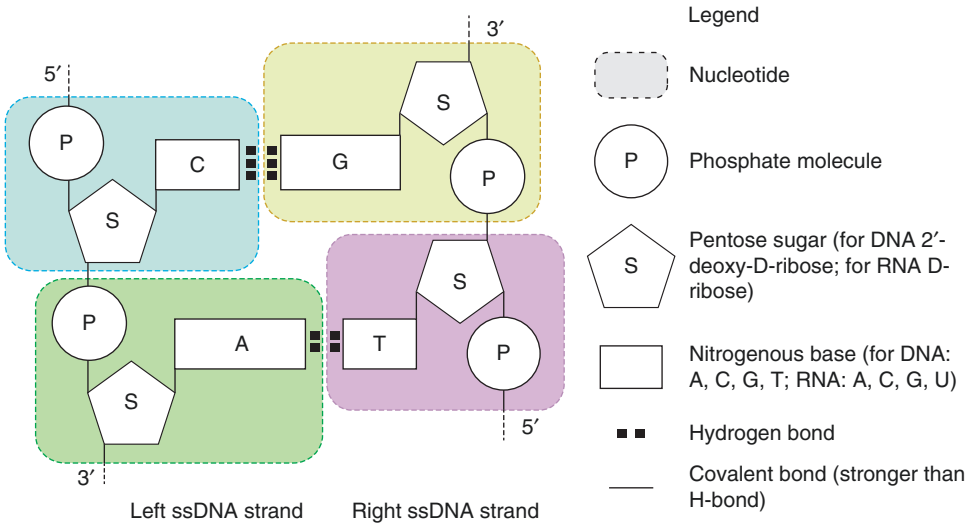
Nucleic Acids

The basic chemical building blocks of natural NAs are nucleotides, outlined in Fig. 1, and consist of three characteristic components:

- A nitrogen-containing base, indicated by each rectangle in Fig. 1, that is either:
  - Purine with two possibilities: adenine (A) or guanine (G)
  - Pyrimidine with three possibilities: cytosine (C), thymine (T), or uracil (U)
- A pentose sugar ring (denoted by S within each pentagon in Fig. 1):
  - For DNA the sugar is 2′-deoxy-D-ribose (with "′" read as "prime").
  - For RNA the sugar ring is D-ribose.
- A phosphate (denoted by P within each circle in Fig. 1)

In DNA and RNA, the nucleotides are covalently linked together with the phosphates acting as bridges between the pentose sugars. The 5′ phosphate group of one nucleotide is linked to the 3′ hydroxyl group of the next nucleotide. The linkages between the nucleotides, on each of the left and right strands in Fig. 1, are often referred to as the "sugar-phosphate backbone." DNA is principally distinguished from RNA by the pentose sugar rings; in DNA the hydroxyl (OH) group on the 2′ carbon atom in the D-ribose ring is replaced by hydrogen (H) [3] so that the oxygen atom can be said to be "removed." The "removal" of the oxygen atom is encapsulated by the prefix "deoxy"; so the prefix followed by "ribonucleic acid" can be helpful for remembering the long name for the acronym "DNA."

The schematic arrangement, i.e., two dimensional (2D) of the three constituents for each of the four nucleotides that comprise DNA, is shown in



Dielectrophoresis of Nucleic Acids, Fig. 1 Ladder sketch of dsDNA nucleic acid showing the constituents of each nucleotide containing a base (*rectangle*), sugar ("*S*" inside a *pentagon*), and phosphate ("*P*" inside a *circle*) and pairing of *left* and *right* strands. Not to scale; for twist and writhe angles between molecules in 3D, see text for references

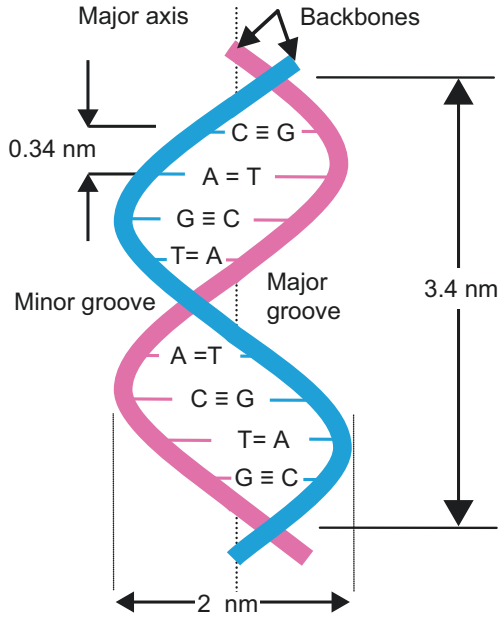
Fig. 1. A key property of NAs is that a single strand of linked nucleotides can form *complementary* pairs with another strand. That is, each base of a nucleotide on one strand can form hydrogen (H) bonds with the base of a nucleotide on the opposite strand according to rules given below. This occurs for all bases on each strand with purines pairing with pyrimidines according to the following rules:

- DNA contains A, C, G, and T and they pair as A with T and C with G.
- RNA contains A, C, G, and U and they pair as A with U and C with G.

DNA and RNA are distinguished not only by their pentose sugars; they differ by one of the bases (T rather than U). Only very rarely do exceptions to these rules occur. There are a number of different structures of DNA; two of the most common are single-stranded DNA (ssDNA) and double-stranded DNA (dsDNA). Figure 1 shows complementary base pairing where A and C on the left ssDNA strand form H-bonds with T and G, respectively, on the right strand. There are different forms of dsDNA, namely, A, B, and Z [7]. The standard double-helical DNA is usually referred to as (Crick and Watson) B-form.

Figure 2 shows a very short length of dsDNA as it occurs in aqueous (or water) solution, that is, in the energetically favorable state as a three-dimensional (3D) double helix. The sugar-phosphate double-helical backbones are shown as thick lines. Structural "backbone" is somewhat of a misnomer as it is the stacking of the nitrogen bases that gives the molecule its mechanical stiffness [7]. The sugar-phosphate backbones are hydrophilic and lie on the exterior of the macromolecule. The hydrophobic nucleotide base pairs (bp) lie within the macromolecule, away from the polar water molecules. They are represented symbolically (A – T, C – G pairs) with, respectively, double and triple H-bonds. The displacement along the major helical axis between base pairs is shown as 0.34 nm, so that the length between repetitive positions of the double helix is about 3.4 nm (10 bp). The diameter is about 2 nm.

Importantly, the H-bonds shown in Figs. 1 and 2 are weaker than the covalent bonds. Consequently, by moderate heating or by using a different medium, each dsDNA macromolecule can be reversibly separated into two complementary ssDNA strands. These ssDNA strands, in a solution



Dielectrophoresis of Nucleic Acids, Fig. 2 A short fragment of dsDNA in 3D double-helical B-form showing basic dimensions (2 nm diameter and 0.34 nm per base pair length) and exhibiting major and minor grooves of the "backbone." Sketch only, not to scale; see text for references and details

with available nucleotides bases, can act as templates, thus enabling copying, or *replication*. ssDNA also enables *detection* whereby a particular sequence of nucleotides will pair up with another strand with the complementary sequence. In a situation where a particular base does not match (e.g., A – A) and the others do is called a *mismatch*. These fundamental properties underpin the emergence of new technologies, such as nanospheres and microarrays, where applications of DEP are being explored for enhancing concentration and other developments – as described in section "Dielectrophoresis of DNA: General."

RNA is mainly involved with reading of DNA, called *transcription*, and its *translation* into proteins and other macromolecules associated with cellular information. There are a number of different types of RNA that perform different functions inside a cell:

– Messenger RNA (mRNA) is concerned with the transfer of genetic "blueprint" information that codes for proteins. There are 20 different amino acids (AAs) found in nature [3]. Fragments of DNA that code for a sequence of AAs that constitutes a protein, using the A, C, G, and T base alphabet, are called *genes*. A triplet of bases is required to sufficiently code for 20 AAs. That is, four bases to choose independently for first (base) position, four bases for the second, and so on yield a total of $4^3 = 64$ possible AAs. This is more than enough to code for the actual 20 AAs needed to make proteins so the genetic code is said to be "*degenerate*" [3]. The triplet of bases that codes for each AA (and also DNA reading instructions) is called a *codon*.
– Transfer RNA (tRNA) reads or decodes each mRNA codon and transfers an appropriate AA onto a polypeptide chain that is being synthesized during the production of a particular protein. tRNAs are small RNAs (about 80 nucleotides) with a characteristic clover leaf shape and act as adapters.
– Ribosomal RNA (rRNA) constitutes part of the ribosome. These are located in the cytoplasm of a cell and are responsible for catalyzing the synthesis of AA peptides and hence protein, by decoding mRNA.
– Noncoding RNA (ncRNA) is an RNA that does not code for protein (i.e., not mRNA) and includes rRNA, tRNA, and other forms of RNA involved in modification, splicing, and regulation [3]. For example, RNA interference (RNAi) is a process involving gene expression silencing, i.e., preventing part of the genetic code being read. ncRNAs include small interfering RNA (siRNA) and microRNA (miRNA) that have been involved with developmental timing [3]. There are many applications of ncRNA and the variety of functions is continuing to be discovered.

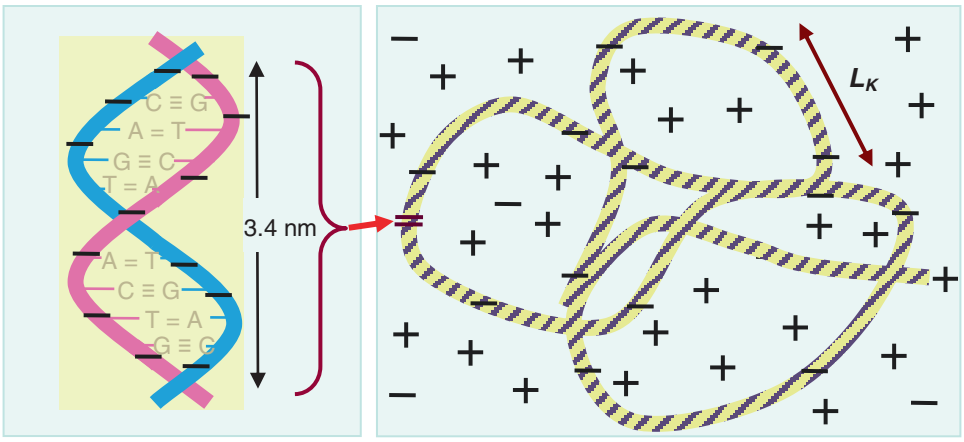
If RNA is extracted, for example, from fish liver cells, the dominant amount of RNA is rRNA with the other amounts of RNA being much smaller. Consequently, mRNA needed for understanding protein production, e.g., for each gene, has to be carefully separated from the total amount of RNA. DNA and RNA are both

biological *polymers*, or biopolymers, with each nucleotide being the smallest subunit (or "mer"). For example, a short strand of 20 nucleotide ssDNA can be referred to as a 20-*mer*. A short strand of up to 50 nucleotides of ssDNA, or ssRNA, can be generally referred to as an *oligonucleotide, oligos* being Greek for "few" [3], and they are often made synthetically. Strands of NAs greater than 50 nucleotides are referred to as *polynucleotides*.

Electrical Properties of Nucleic Acids

DNA is classified as a biological polyelectrolyte. These are biomolecules with a large number of charged chemical groups. An electrolyte or aqueous environment with salts acts as a supply for dissociated ions, for example, Na^+, Mg^{2+}, OH^-, and Cl^-. DNA and RNA possess a net *negative* charge so they attract ions of the opposite charge, or *counterions*, by Coulombic forces to restore charge neutrality. The positive counterions, or cations, form a cloud around the NA and electrically screen the negative charge. Further away from the macromolecule, the counterions are more diffuse and the screening occurs in such a way as to attract other negative ions, or anions, with the same charge as the NA, and they are called *coions*.

The double-helical structure for DNA is shown to be straight in Fig. 2 but this only applies for relatively short lengths. The mechanical properties of DNA are partly determined by its chemical composition and the composition of the solvent or suspending *medium*. In an aqueous medium with low counterion concentration, there is little electrostatic screening so the polyelectrolyte repels itself and it tends to be straight. Conversely, in solvents with high counterion concentrations, there is considerable electrostatic screening so the DNA repels itself less and it is more flexible. A measure of straightness is known as the persistence length, L_p, as sketched in Fig. 3. Typically for standard biological conditions for dsDNA, $L_p = 50$ nm [6]. The behavior of dsDNA in water can be modeled as a wormlike chain with each straight link equal to the Kuhn length that is twice the persistence length, $L_K = 2L_p$ [8]. In this chain, due to thermal motion, the next link is independently and randomly oriented from the previous link. Each link is about 100 nm or about 300 bp. The persistence lengths of ssDNA are much shorter than dsDNA, with a minimum of about 1 nm. The following discussion on dielectric polarization mechanisms includes both of the NAs but focuses on DNA since the literature, reviewed in [6], is much more extensive.



Dielectrophoresis of Nucleic Acids, Fig. 3 dsDNA (or DNA) as a negatively charged polyelectrolyte in solution attracts counterions (mostly cations). These form around the polyelectrolyte to electrically screen the charge of the DNA. Wormlike chain model of dsDNA, with close-up inset, in aqueous suspension, showing relatively straight segments of Kuhn length L_K. Not to scale

In addition, dielectric spectroscopic measurements of the electrical properties of rRNA are often confounded by the presence of ribosomal proteins (see the work by Bonincontro et al. (1997), cited in [9], and references to other work therein).

Dielectric Polarization Parameters

Polarization is a term that describes how charges, within a dielectric, respond to an externally applied electric field. Charges that are free to move over a long range reveal their movement on a macroscopic scale as conduction. If the movement of the charges is restricted, they are said to be "polarized" [4]. In this respect, polarization is the "capability" of the charges to move in response to an applied electric field. The polarization can be expressed, in terms of an equivalent circuit with permittivity parameters, as the real (in-phase) permittivity ε' or *dielectric constant* [10] and the free movement of charges as the imaginary (out-of-phase) permittivity ε'' or *dielectric loss*. The permittivity parameters are frequency dependent, $\varepsilon \equiv \varepsilon(\omega)$. In addition the conductivity (or low-frequency Ohmic loss) of the electrolyte solution σ is included [11]. The complex permittivity ε is the combination of these:

$$\varepsilon(\omega) = \varepsilon'(\omega) - j[\varepsilon''(\omega) + \sigma/\omega] \tag{1}$$

where $j = \sqrt{-1}$; the angular frequency is $\omega = 2\pi f$ (rad s^{-1}) with frequency f (Hz). Often $\varepsilon'' << \sigma/\omega$, the superscript $'$ notation is omitted in the literature and the complex permittivity is written:

$$\underline{\varepsilon}^* = \varepsilon - j\sigma/\omega \tag{2}$$

where the underscore denotes a complex quantity, "*" denotes complex conjugate, and ε is the permittivity (real part of $\underline{\varepsilon}^*$) that is assumed constant for the frequencies of interest.

A quantitative measure of the responsiveness of a body to an applied external electric field is the polarizability, α. A highly polarizable body, for example, features many charges that are responsive to an electric field and their movement is in some way restricted. An important parameter for experimentally determining the value of the polarizability of a NA, such as DNA, is the dielectric increment, or decrement [10–13]:

$$\Delta\varepsilon' = \varepsilon'_{rl} - \varepsilon'_{rh} \tag{3}$$

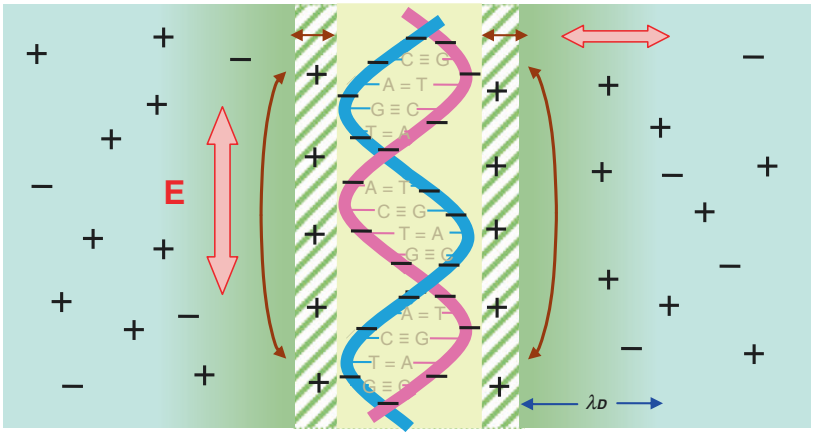
where ε_{rl} and ε_{rh} are the low- and high-frequency relative permittivities, or limiting dielectric constants, respectively. The predicted polarizability, α_m, for a macromolecule determined using the experimentally measured dielectric decrement, $\Delta\varepsilon'$, is given by

$$\alpha_m = \frac{3\varepsilon_o \Delta\varepsilon'}{C_m} \tag{4}$$

where ε_0 is the permittivity of free space, C_m is the number density (m^{-3}) of macromolecules, and the factor "3" accounts for random macromolecular orientation. Using dielectric mixture theory, the polarizability of each DNA, or RNA, macromolecule can be determined from the suspension by calculating the volume fraction of the NA. Another parameter is the relaxation time constant, τ_R(s) that is the time duration for charges to redistribute or "relax" after being perturbed by an electric field. The charges resonate at a dispersion frequency, f_R, or angular frequency ω_R:

$$\tau_R = \frac{1}{f_R} = \frac{2\pi}{\omega_R} \tag{5}$$

Dispersions arise when the oscillating cloud of charges can no longer follow the alternating electric field, and this effect is observed in both real and imaginary parts of the complex permittivity. NAs can have more than one dielectric dispersion characterized by a decrement and associated time constant. Commercial time-domain dielectric spectrometers can be used to estimate the dispersion decrement $\Delta\varepsilon'$ and time constant τ_R using a sample, e.g., 150 μl, of macromolecules suspended in a medium, such as water. Characteristics of dielectric permittivities and losses for calf-thymus DNA, a macromolecule used in many dielectric spectroscopy studies, versus frequency



Dielectrophoresis of Nucleic Acids, Fig. 4 Counterion polarization on a short segment of DNA: a negatively charged sugar-phosphate double helix attracts mainly counterions. Components of an applied AC electric field (*red filled bidirectional arrows*) cause counterion movement in the longitudinal direction – along the DNA backbone – and transverse direction (*brown bidirectional line arrows*). The counterion movement tends to be within the "cylindrical" condensed layer (*green shaded* ////). The diffuse layer has characteristic length, λ_D, as shown and is referred to as the Debye screening length – see text for details

are compiled and shown in [6]. Although there are variations in the experimentally measured permittivity, ε', they all tend to decrease with an increase in frequency and the polarizability follows the same trend.

The types of polarization include electronic, atomic, molecular, interfacial (or space-charge), and counterion polarization. The first three are attributed to the displacement, or orientation, of bound charges; the latter two concern movement on a larger scale. Of the first three types of polarization, only molecular dipole polarization tends to feature in the literature concerning dielectric properties of NAs. The asymmetric distribution of electrons in molecules gives rise to *permanent* dipoles. The interaction of such a dipole with an externally applied electric field causes a torque that attempts to orient the molecule in the field direction. The polarization is appropriately called *orientation*, or *dipole*, polarization. Water is an example of a molecule with a permanent dipole and manifests permittivity values, via ε and σ, that are practically frequency independent up to 17 GHz [11].

The last two kinds of polarization, interfacial and counterion, involve large-scale charge movement. Currently, there is no universal consensus in the literature on the high-frequency polarization mechanisms for macromolecules, such as DNA or RNA. At present, the emphasis tends to favor Maxwell-Wagner interfacial polarization for molecular structures with well-defined interfaces with water molecules and counterion fluctuation polarization, along the longitudinal axis, for DNA and similarly for RNA.

Counterion Fluctuation Polarization for DNA

It is the response of the counterions to an externally applied AC electric field that results in counterion polarization. Since there is Coulombic attraction between the charged body and the counterion layer, the counterions attempt to pull the charged body along with them as they follow the electric field. The counterion polarization mechanism is due to solution counterions (such as Na^+, Mg^{2+}, etc.) interacting with negatively charged phosphate groups along the DNA or RNA polyion backbone. The counterions move freely along the lengths of the macromolecule in response to the component of the external electric field parallel to the major axis or "backbone" (Fig. 4). Their migration results in an induced dipole moment and, hence, an identifiable polarizability.

A popular model of counterion polarization is the "Mandel-Manning-Oosawa" model developed in various stages by these authors cited in [6, 13, 14]. In this model, the counterions move

freely along macromolecular "subunit lengths" and are permitted to cross from one subunit to a neighboring subunit only by overcoming "potential barriers." The *subunit length*, L_s, is described as the length along the average macromolecular conformation between "breaks," or "potential barriers," resulting from perturbations in the equipotentials due to conformational processes, folding, etc. Counterions moving along these *subunit* lengths, under the influence of an external electric field, manifest *high*-frequency dispersion that is molecular weight *independent*. This polarization contrasts with *molecular-weight-dependent* counterion movement along the entire macromolecular contour length that results in *low*-frequency dispersion and may account for the static permittivity. A proportion of the counterions are so strongly attracted to the polyelectrolyte that they are said to "condense" onto the polyion backbone. Essentially there are three distinct phases:

(i) *Condensed* counterions – these are sufficiently, but non-locally, bound (or "delocalized") to the phosphate groups of the DNA and thereby neutralize a fraction of the DNA charge. In many respects the condensed counterion layer is similar to the Stern layer of the double-layer model for colloids.
(ii) *Diffuse* counterions which are responsible for neutralizing the remainder of the DNA charge, with a density which decreases exponentially with distance from the axis.
(iii) *Bulk* ions or "added salt" ordinary aqueous solution ions.

In terms of their contribution to polarizability, the condensed counterion phase is the most important. A feature of the condensed state is that the local concentration of counterions around the DNA does not tend to zero when the suspending medium or bulk electrolyte concentration does. Condensation occurs when the condition $\xi \geq 1/|z|$ is satisfied where z is the valence of each counterion and the charge density parameter, ξ, is given by

$$\xi = q^2/(4\pi\varepsilon_m k_B T b) \tag{6}$$

where q is the elementary charge, $\varepsilon_m = \varepsilon_0\varepsilon_{rm}$ is the permittivity of the bulk electrolyte medium, ε_{rm} is the relative permittivity, k_BT is the Boltzmann temperature, and b is the average distance between charged sites; for B-DNA double helix, $b = 1.73$ Å. The "Manning-Mandel-Oosawa" model yields a generalized expression for scalar longitudinal polarizability, α_s, per subunit length:

$$\alpha_s = \frac{z^2 q^2 {L_s}^2 n_{cc} A_{st}}{12 k_B T} \tag{7}$$

where A_{st} is the stability factor of the ionic phase and includes mutual repulsion between fixed charges on the backbone and the effect of Debye screening,

$$A_{st} = [1 - 2(|z|\xi - 1)\ln(\kappa_s b)]^{-1}. \tag{8}$$

Referring to Eq. 8, the reciprocal of the Debye screening length, shown in Fig. 4, $\lambda_D = \kappa_s^{-1}$ (m) is given by

$$\kappa_s = \left[\left(\frac{N_{Av} q^2}{\varepsilon_m k_B T}\right)\left(\sum_i C_i z^2 + \frac{C_p}{\xi}\right)\right]^{0.5} \tag{9}$$

where N_{Av} is Avogadro's number and C_i and C_p are the respective molar concentrations of ions in the bulk and diffuse phase and phosphate groups. In the simplified case where $C_p = 0$, Eq. 9 reduces to the standard formula for the Debye reciprocal length of a balanced electrolyte [4, 12, 13]. In Eq. 7, n_{cc} is the number of condensed counterions that can be predicted theoretically:

$$n_{cc} = \phi_c L_s/|z|b \tag{10}$$

where the fraction of condensed counterions is $\phi_c = 1 - |z^{-1}|\xi^{-1}$. Equations 6, 7, 8, 9, and 10 are equated with Eq. 4 that determines the polarizability from a dielectric spectroscopy experiment. The relation, $\alpha_m C_m = \alpha_s C_s$, where the number density of subunits is $C_s = N_{Av} C_p b/L_s$, is useful to bridge the theory and experiment. An expression for L_s based on the measured dielectric decrement $\Delta\varepsilon'$ is thus derived (see Bakewell et al. (2000) cited in [13, 14]):

$$L_s = \sqrt{\frac{9\Delta\varepsilon'}{\pi\varepsilon_{rm}(|z|\xi - 1)A_{st}N_{Av}C_p b}}. \quad (11)$$

Assuming, for example, $T = 298$ K (25.0 °C), $\varepsilon_{rm} = 78.4$, and $z = 1.00$ for monovalent cations and $C_p = 2.72$ (mol/m^3) and $C_i = 1.10$ (mol/m^3), then $\xi = 4.13$, $\kappa_s = 9.754 \times 10^7$ (m^{-1}), $A_{st} = 3.76 \times 10^{-2}$, and $L_s = 3.31 \times 10^{-8}\sqrt{\Delta\varepsilon'}$ (m). An expression for the relaxation time in terms of the subunit length can be also derived:

$$\tau = \frac{{L_s}^2 q}{\pi^2 \mu k_B T}. \quad (12)$$

Assuming the mobility value $\mu = 8.00 \times 10^{-9}$ (m^2 V^{-1} s^{-1}) at 25 °C, $L_s = 1.42 \times 10^{-4}\sqrt{\tau}$ (m). A sample of dsDNA with measured values, $\Delta\varepsilon' = 9.0$ and $\tau = 500$ ns, has predicted subunit lengths from Eqs. 11 and 12 that approximately concur with each other, $L_s \cong 100$ nm. Physically, this value is close to the wormlike chain Kuhn length for dsDNA. In the framework of the counterion model, the result would support the notion that equipotentials arise from natural curvature of the dsDNA suspended in solution.

Maxwell-Wagner Polarization for DNA

Interfacial, or *space-charge*, polarization arises from electrical charges being restricted in their movement at the interfaces between layers of different dielectric materials. Interfacial polarization results in dielectric dispersions when the aggregate of dissimilar materials is exposed to AC electric fields [4, 10–13]. Maxwell-Wagner (MW) polarization models for DNA suspensions have been undertaken by a number of researchers (see Grosse (1989) and Saif et al. (1991) and their respective coworkers cited in [13, 14]).

MW interfacial polarization models the DNA macromolecule with the amino acid sugar-phosphate double helix as a 2 nm diameter, long cylindrical insulating core with low conductivity and permittivity. The insulator core is surrounded by a highly conducting sheath that represents the bound counterions on the negatively charged sugar-phosphate backbone. The conductive sheath interfaces with an electrolyte medium with low conductivity. The dielectric decrement for MW dispersion for DNA is predicted to be

$$\Delta\varepsilon' \cong 8\varepsilon_{rm}v_f/3 \quad (13)$$

where v_f is the volume fraction. Similarly, the high-frequency relaxation is approximated in terms of the DNA *suspension* conductivity, σ_s (or conductivity of the electrolyte medium, σ_m) and fraction of condensed counterions, ϕ_c,

$$f_R = \sigma_m\phi_c/\left(4\pi\varepsilon_o\varepsilon_{rm}v_f\right) \quad (14)$$

where $\sigma_m = \left(1 - v_f\right)\sigma_s/(1 - \phi_c)$. For example, using $\sigma_s = 8.5 \times 10^{-3}$ (S/m) at 25 °C, $v_f = 0.20\,\%$, and the fraction of condensed counterions, $\phi_c = 0.25$, Eqs. 13 and 14 yield $\Delta\varepsilon' = 0.42$ and $f_R \cong 160$ MHz.

An alternative approach that mimics the rod-shaped DNA macromolecule with conducting sheath by a randomly oriented shelled, prolate ellipsoid with a very short minor axis and an extremely long major axis, predicted a very high relaxation frequency of hundreds of MHz (see Bone and Small (1995) cited in [6, 14] and other reports by their group). The MW interfacial polarization model tends to predict high relaxation frequencies in the order of 100s of MHz and it is a less successful model for explaining the dielectric behavior of DNA than the counterion fluctuation model. This is perhaps because relatively dilute suspensions of DNA exist as wormlike entanglements with a poorly defined interface compared with the well-defined spherical, or rod-like, boundaries of colloids or some viruses.

Other Polarization Mechanisms

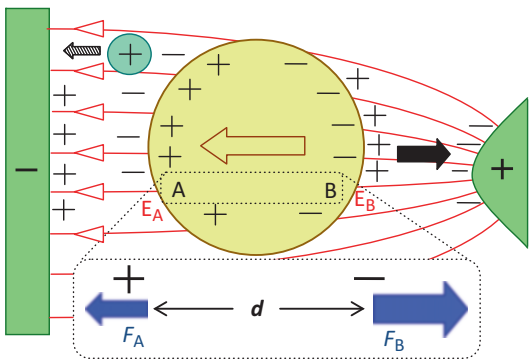
A variety of other polarization mechanisms include counterion polarization transverse to the DNA axis, rotation of water-bound molecules along and across the grooves of the DNA, and relaxation of DNA polar groups. These polarization mechanisms tend to predict very high-frequency dispersions up to hundreds of MHz. Also the relaxation times have temperature dependence that predicts much higher activation

energies than measured and tend to have received less significance in the literature (see works by De Xammara Oro, Mashimo, and Takashima referenced in Bone and Small (1995) and others cited in [6, 14]).

Dielectrophoresis

The principles of DEP are often described for an electrically neutral body although in practice the body, such as DNA, is often charged. Application of an electric field external to a neutral body induces an uneven spatial distribution of electrical charge, or *dipoles*, as shown in Fig. 5.

The external electric field also acts on the displaced charges themselves, so that the electric field can be said to act *twice* on the body. If the electric field is uniform, the field flux acting on electrical charges within each localized region yields Coulombic forces that cancel each other and the net force is zero. On the other hand, in a *nonuniform* electric field, Fig. 5, the density of the electric flux varies spatially between regions within the body, so that there is an imbalance of Coulombic forces and they sum to a nonzero net force. If the body is *more* polarizable than the surrounding medium, it moves *toward* the region of highest field nonuniformity. This is the case in Fig. 5; the body moves to the right, as shown, and the movement is called *positive* DEP (pDEP). Conversely, if the body is *less* polarizable than the surrounding medium, it moves *away from* the region of highest field nonuniformity (i.e., to the left, not shown) and it is called *negative* DEP (nDEP).



Dielectrophoresis of Nucleic Acids, Fig. 5 Dielectrophoresis: a neutral body (*yellow circle*) in a nonuniform electric field (*red unfilled arrows*) experiences a net force depending on the spatial distribution of electric field strength. In this example, where the polarizability of the body is greater than the surrounding medium, it moves to the right (*black arrow*). The field lines near the body are simply sketched; see the literature, e.g., [13], for a more accurate description. A positive test charge (\\\\\\\ *arrow*) moves to the left according to the direction of the electric field flux. **Inset**: charges at A and B displaced d apart interact with the electric field, E_A and E_B, generating Coulombic forces, F_A and F_B, that act in opposite directions (*blue-filled white-dot patterned arrows*), as shown. The force at B is greater than at A, $F_B > F_A$, so the neutral body moves to the right (*black-filled arrow*)

The imbalance of forces on the neutral body can be modeled by considering two elementary charges $+q$ and $-q$ at A and B, distance d apart, shown in Fig. 5 inset. The electric field at B is stronger than at A, $E_B > E_A$. Assuming the forces are acting in the horizontal direction, the sum of the Coulombic forces approximates with a Taylor series:

$$\begin{aligned} F_{total} &= F_A + F_B = qE_B - qE_A \\ &= q\,[E(x+d) - E(x)] \cong q\,d\frac{dE}{dx} \\ &\cong p\frac{dE}{dx} \end{aligned} \tag{15}$$

where $p = qd$ (C m) is the induced (or effective) dipole moment. This can be written as

$$p = \alpha v E \tag{16}$$

where α is the induced polarizability, or effective dipole moment, per unit volume, v, in unit electric field and has units Farad per meter (Fm^{-1}). It is assumed, for the force to be proportional to the derivative of E in Eq. 15, that the dimension of the dipole p is small compared with the characteristic length of the electric field nonuniformity. Any spatial electric field phase variation is considered to be negligible and so that consideration is confined solely to the in-phase component of the DEP force, and the polarizability is real. Combining Eqs. 15 and 16,

$$F_{DEP}|_{DC} = \alpha v E\frac{dE}{dx} = \frac{1}{2}\alpha v\frac{dE^2}{dx} \tag{17}$$

where the subscript signifies that this equation is applicable for a constant (DC) signal applied to electrodes and is positive (i.e., acting toward the right). Expressions for the net force acting on a neutral bioparticle in a nonuniform electric field have also been developed using the energy variation principle and Maxwell stress tensor approach (see discussion and reference to Wang et al. (1997) cited in Chap. 4 of [13]). The effective moment method is the most straightforward and can accommodate situations where there is dielectric loss.

Since the electric field acts twice, *inducing* and *acting on* dipoles within the body, the DEP force predicted by Eq. 15 is proportional to the *square* of the magnitude of the electric field. The body moves in accordance with the strength and inhomogeneity of the electric field, the volume and dielectric polarizability of the body, and also the dielectric properties of the suspending medium. Importantly, the net movement of the body does *not* depend on the direction of the electric field or on the electrode polarity. Consequently, time-varying alternating current (AC) electrode voltages can be used to move the body rather than constant time direct current (DC) potentials. This feature, in turn, avoids problems of hydrolysis reactions occurring between electrode surfaces and bulk medium. AC radio frequencies above a few kHz applied to electrodes in moderately salty electrolytes, for example, do not incur hydrolysis. This enables DEP to be implemented using microelectrodes with micron to submicron features for miniaturized applications.

Since AC signals are more useful than DC, the response observable of macromolecules moving in liquids implies that the small time-averaged DEP force is considered. It is evaluated as half the DC value. Generalizing to three dimensions,

$$\left\langle \vec{F}_{DEP}(\underline{x},t)\right\rangle_t = \vec{F}_{DEP}(\underline{x}) = \frac{1}{4}\alpha(f)v \vec{\nabla}\left|\vec{E}(\underline{x})\right|^2 = \frac{1}{2}\alpha(f)v \vec{\nabla}\left|\vec{E}_{rms}(\underline{x})\right|^2 \quad (18)$$

where $\vec{\nabla}$ is the gradient operator and "→," $\langle\rangle_t$, and "*rms*" denote vector quantity, "small time" average over a RF electric field oscillation period, and root mean square, respectively. The frequency dependence of the polarizability is made explicit and the DEP force is understood to be "almost instantaneous." In terms of understanding and computing the DEP force, the electric field gradient for realistic geometric electrode designs can be determined analytically for simple cases or by electromagnetic simulation software. The frequency-dependent polarizability, α, is inferred from dielectric measurements using Eq. 4 or predicted from Eq. 7, and the volume is usually known or can be approximated.

Note that for a sphere with radius r with MW polarization, $\alpha = 3\varepsilon_m \mathrm{Re}\{f_{CM}\}$ where f_{CM} is the frequency-dependent Clausius-Mossotti factor and "Re" denotes real part. The time-averaged force given by Eq. 18 becomes

$$\vec{F}_{DEP}(\underline{x}) = 2\pi\varepsilon_m r^3 \mathrm{Re}\{f_{CM}\}\,\vec{\nabla}\left|\vec{E}_{rms}(\underline{x})\right|^2 \quad (19)$$

which is the starting formula in many scientific publications, e.g., [1, 4–6]. In terms of scaling relations used in nanotechnology, a simple and effective way of considering the DEP force is to recall that E is determined by Poisson's or Laplace's equation, so

$$F_{DEP} \propto \frac{r^3}{d^3}V^2. \quad (20)$$

In Eq. 20, d represents the characteristic length of the structure, or feature, size generating the nonuniform electric field and V is the voltage across the electrodes. This means that if r and d decrease by the same amount, e.g., by 1/1000 from μm to nm, in body and feature size, the DEP force remains the same. This principle of invariance to scale enables DEP to be an attractive means of moving biomolecules on the nanoscale even if earlier work was on the micron scale or larger. It also remains proportional to the square of the voltage across the electrodes.

Dielectrophoretic Transport

NAs have little inertial force compared with viscous drag force so that the former can be neglected. The latter can be approximated in the first instance if a charged macromolecule and associated counterions and coions, moving in a medium (e.g., DNA in water), behave hydrodynamically as if it is a semirigid body, such as a sphere or an ellipsoid. As discussed in section "Electrical Properties of Nucleic Acids," the rationale for this is that DNA behaves as a wormlike chain retaining a roughly spherical (or ellipsoidal) shape due to ions and water molecules being attracted and dragged along with it. The complex of DNA, ions, and water can be made equivalent to a body with a hydrodynamic radius, r_h, and an effective drag force is approximated.

As a consequence, the concentration flux of a population of single, noninteracting NAs or an ensemble of experiments of single NAs, arising from a DEP force, can be expressed in a very simple way as

$$\vec{J}_{DEP}(\underline{x},t) = c(\underline{x},t)\vec{F}_{DEP}(\underline{x})/\zeta \qquad (21)$$

where the drag coefficient for a sphere is $\zeta = 6\pi\eta r_h$ where η is the dynamic viscosity of the medium. The incessant collisions of water molecules on the NA give rise to Brownian motion and hence to a thermally driven diffusion flux:

$$\vec{J}_{Diff}(\underline{x},t) = -D\ \vec{\nabla}\, c(\underline{x},t) \qquad (22)$$

where the Einstein diffusion coefficient is $D = k_B T/\zeta$ and all symbols have been defined. The contributions from DEP and diffusion can be combined into a continuity equation, also known as the modified diffusion equation (MDE) or Fokker-Planck equation:

$$\frac{\partial c(\underline{x},t)}{\partial t} = -\vec{\nabla}\cdot\left(\vec{J}_{DEP}(\underline{x},t) + \vec{J}_{Diff}(\underline{x})\right). \qquad (23)$$

The evolution of Eq. 23 is a starting point for very simple analytical and numerical simulations of DEP-driven transport of NAs.

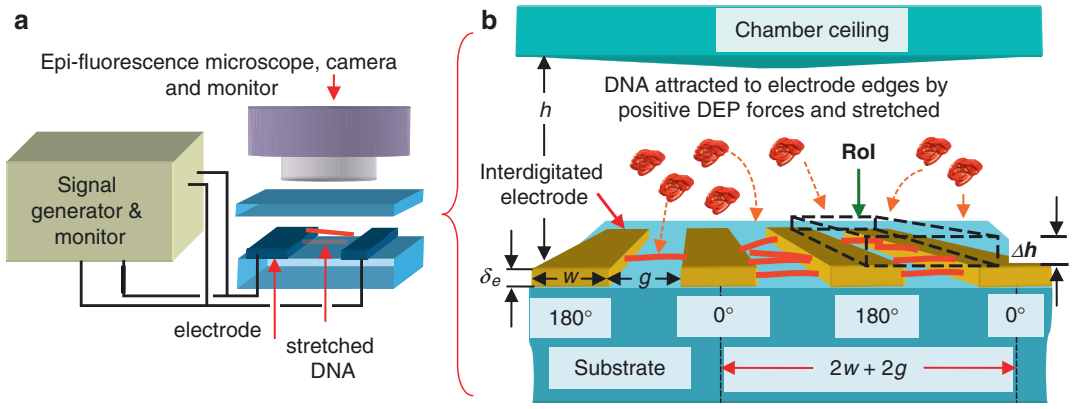
Applications

There is much interest in controlling the movement of DNA in microfabricated environments for two general application areas: first, DNA base sequence and gene expression analyses in LOC and μTAS microdevices and related biotechnological and biomedical applications and, second, using DNA as nano-scaffolds or constructing material, in molecular electronics. A key motive for this application is that "top-down" fabrication approaches are limited on the nanoscale so there is strong technological interest in alternative "bottom-up" approaches. Although there is considerable research on DEP of DNA, there is very little work reported so far on DEP of RNA or PNA.

This section starts with an introductory example of DEP of DNA. Investigations indicate DNA macromolecules tend to exhibit pDEP, i.e., their induced polarizability is greater than the surrounding medium. As discussed in the previous section, the pDEP could be due to the dilute suspensions of DNA existing as wormlike entanglements involving the aqueous medium so that there is a poorly defined polyelectrolyte/medium interface. The behavior contrasts with model colloids, e.g., polystyrene nanospheres, or viruses that can exhibit both pDEP and nDEP – depending on the applied RF, medium conductivity, and other experimental factors. The section discussion continues on to address wider contributions from the scientific community to DEP of NA.

Positive Dielectrophoresis of DNA: An Introductory Example

DEP of DNA is typically accomplished with a standard RF signal generator, signal monitoring oscilloscope, and measurement apparatus. A wider range of RFs and conductivities can be used but a benchmark RF is 1 MHz with reverse osmosis (RO) water or buffer with conductivity less than 1 millisiemen per meter (mS/m). The measurement process often involves imaging using fluorescence microscopy with modifications that enable recording and data processing. An experimental arrangement for observing DNA transport is shown in Fig. 6a. Tracking of DNA



Dielectrophoresis of Nucleic Acids, Fig. 6 Scheme of DEP experimental apparatus: (**a**) AC signal generator, monitoring oscilloscope, epi-fluorescence microscope, camera, monitor, DEP of DNA collection experiment. (**b**) Detailed inset of DEP collection showing interdigitated electrodes (fabricated with width w, interelectrode gap g, and thickness δ_e – not to scale) with electrical potential phases 180° apart and housed in a chamber with height h. DNA molecules (*red*) are attracted to the electrode edges and across the interelectrode gaps by positive DEP and stretched, as shown. An example of a region of interest (*RoI*) volume (height Δh), for DEP collection quantification, is shown on the *right* by the *box outlined* by *black dashes*

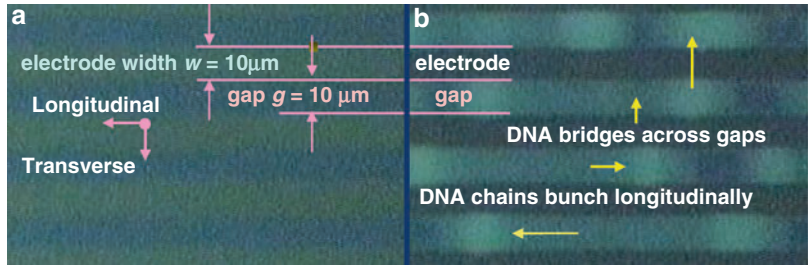
macromolecules as they move in real time is achieved by labeling them with a fluorescent dye and observing their micron-scale movement with an epi-fluorescence microscope with typically 200 × or above magnification, as shown. Often an anti-strand-breaking agent is used with the fluorescent dye and a camera is used to record and display DNA movement on a PC/TV screen.

A perspective close-up view of the microelectrode interdigitated "finger" electrodes that are responsible for generating electric fields that drive DEP macromolecule transport is shown in Fig. 6b. The thickness of the electrodes, δ_e, has been exaggerated for illustration and is typically much less than the transverse width, w, and interelectrode gap, g, $\delta_e << w, g$. Typical standard microfabrication methods are used: UV patterned resist on glass or other suitable substrates, evaporated with metal, developed, and lifted off. The AC potentials to the electrodes are 180^0 apart so the potential pattern is repeated in the transverse direction every $2w + 2g$ so from Eq. 18 the DEP force repeats and is symmetric about each center of the electrode. The regions of the planar array with the highest nonuniformity are the electrode edges so switching on the DEP attracts the DNA to that region. The DNA is shown in Fig. 6b to be coiled in the suspension; plasmid DNA, for example, has a shape similar to a rubber band that has been, cut, twisted, and rejoined again. If the DEP force is sufficiently strong, the coiled DNA collected by the pDEP forces becomes stretched by the electric field and builds up to form bridges, parallel to the field lines, that span between adjacent electrodes, as shown.

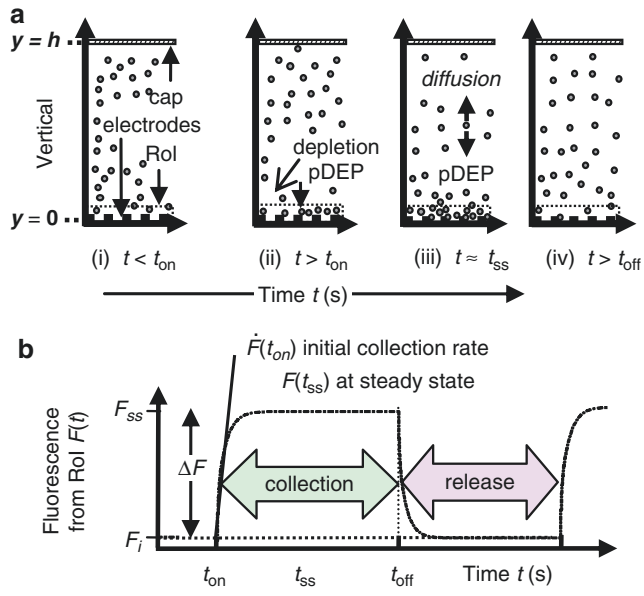
Quantification of pDEP collections of plasmids can be achieved using fluorescence spectroscopy, in a similar manner to nanoparticles, by designating a three-dimensional (3D) volumetric region of interest (RoI) (see [1, 14] and references therein to Bakewell and Morgan (2004)). As illustrated on the right of Fig. 6b, the 3D RoI involves assignment of a 2D planar area parallel to the substrate and encompassing the electrode edges, with a small height, Δh, e.g., that is within the micron-scale depth of focus of the microscope objective. Within the RoI volume (V_{RoI}), for sufficiently low concentrations, the fluorescence is proportional to the plasmid concentration:

$$F(t) \propto \int_{V_{RoI}} c(\underline{x}, t)\, d\underline{x}. \tag{24}$$

Consequently, the fluorescence can be used as a measure of the amount of DNA within an RoI.



Dielectrophoresis of Nucleic Acids, Fig. 7 DAPI-labeled DNA plasmid suspension: half-frame width images: (**a**) before the onset of DEP and (**b**) 0.13 s after the onset of DEP with $V_o = 4.5$ V, $f = 100$ kHz. See text for further details



Dielectrophoresis of Nucleic Acids, Fig. 8 Cartoon showing DNA collection under the action of positive DEP force and release after the DEP force is switched off. (**a**) Cartoon showing DNA concentration within the chamber with RoI (*side view*, not to scale) shown by the *dotted box outline* near the array at the *bottom* of the chamber; (**b**) fluorescence quantified within RoI as a function of time, $F(t)$, showing DNA collection and release

An example of a part video-frame microscope image of 12 kilo-bp (kbp) plasmid DNA, stained with fluorescent dye, 4′,6-diamidino-2-phenylindole (DAPI), taken from a similar study to [14], is shown in Fig. 7a before DEP was applied. The electrodes comprised 10 nm titanium onto glass substrate, 10 nm palladium, overlaid with 100 nm gold, and with dimensions $w = g = 10$ μm. The negatively supercoiled plasmid DNA strands, suspended in low conductivity medium, cannot be seen individually but they were known to be about 4 μm in contour length or 2 μm looped end to end. Figure 7b shows the DNA has collected into the interelectrode gap 0.13 s after pDEP was applied. One plasmid alone cannot bridge the interelectrode gap, so it is thought that the strands elongate and link to each other, similar to a pearl chain, as they position between the electrodes. The medium has started to evaporate so the transversally stretched DNA strands have tended to bundle longitudinally in places, as shown. After the pDEP force is switched off, Brownian motion is responsible for the DNA strands being released (image not shown) and moving away from the planar array into the bulk solution.

Quantification of pDEP collection and release of plasmid DNA, [1, 14], were achieved by exploiting the geometric properties of the electrodes. The longitudinal length, l, of the electrodes, $l >> w + g$, enabled symmetric and periodic patterns of the DEP to be efficiently image processed that resulted in collection and release time profiles. A simplified side-view cartoon of pDEP of plasmid DNA, suspended in an aqueous medium within a chamber, is shown in Fig. 8a. Accompanying collection and release fluorescence profiles are shown in Fig. 8b, as the

pDEP is switched on and off, or time modulated. Near the electrodes at the bottom of the chamber, the RoI is indicated by the box-dotted outline shown in Fig. 8a stages (i) to (iv). The DNA plasmids are initially uniformly suspended, Fig. 8a(i), with a concentration that, within the RoI volume, yields the *initial* fluorescence, $F_i = F(t_{on})$, as shown in the corresponding plot in Fig. 8b. After pDEP is switched on at $t = t_{on}$, the plasmids are attracted toward a planar interdigitated electrode array by pDEP, particularly near electrode edges where the DEP force is strong. As the plasmids rapidly collect near the electrode edges, the concentration increases so that the fluorescence of the plasmids within the RoI region above the array is shown to increase for a short time afterward, $t > t_{on}$, as shown in Fig. 8b. The concentration of the plasmids above the RoI becomes depleted, and over time, the depletion zone progresses toward the microdevice cap, Fig. 8a(ii). Eventually the pDEP flux becomes balanced by thermally driven diffusion and the concentration reaches steady state (SS) at $t = t_{ss}$, shown in Fig. 8b(iii), with fluorescence, $F_{ss} = F(t_{ss})$, shown in Fig. 8b.

Switching off pDEP at $t = t_{off}$ releases the plasmid DNA so the fluorescence of the plasmids, within the RoI, is shown to decrease in Fig. 8b. The plasmids continue to diffuse from the array and return to the bulk solution, Fig. 8a(iv). Two quantification summary parameters, shown in Fig. 8b, are (i) the initial collection rate and (ii) the change from initial to SS concentration. The former assumes, initially, diffusion is negligible in Eq. 23. Consequently, from Eq. 24, to a first approximation, the rate of change of fluorescence is proportional to the DEP force, $\dot{F}(t_{on}) \propto F_{DEP}$ where the overhead "$\cdot$" denotes time derivative. The relation for the initial collection rate enables a comparison of DEP forces for different experimental parameter values, e.g., voltage and frequency. The latter parameter, initial to SS fluorescence change, $\Delta F = F(t_{ss}) - F(t_{on})$, can also be compared for different experimental parameters.

The results in [14] indicated that the strength of collections also decreased with increasing frequency from 100 kHz to 5 MHz, and these trends concurred with dielectric polarizability measurements for the same DNA plasmid. The results also showed that for these electrodes, the predicted pDEP force is lower than that expected from theory. A number of reasons were presented including the possible effects of AC electroosmosis (EO) (also called electroosmotic flow (EOF)), polarization saturation, and distortive effect of the DNA on the electric field itself.

D

Dielectrophoresis of DNA: General

There have been many variations and advancements on the basic arrangement in the previous section. The literature indicates that Washizu's group in Japan pioneered most of the early and novel experiments on DEP of DNA, starting in the 1990s with their fluid-integrated circuit (FIC). The details of their work, e.g., Washizu and coworkers (1990–2007), can be found in several very comprehensive reviews [1, 5, 6], along with many important and recent contributions from newly emergent groups around the world. The state of the art and advancement for nanotechnology can be summarized into a couple of general categories:

1. Generation of DEP force, microfabricated environment, and DNA preparation:
   (a) The length of DNA that can be moved by DEP depends on many factors including the smallest feature size of the electrodes, position of the DNA macromolecules within the nonuniform electric field, and ranges from 20 bp to tens tens of kbp and upward; it is limited by the practicalities of sample preparation, e.g., a 1000 kbp or 1 Mega-bp (Mbp) of DNA has a contour length of 340 μm, long enough to be sheared in a micropipette tip during preparation.
      - Many experiments have used 48.5 kbp (contour length 16.5 μm) λ-phage dsDNA and more recently short sequences (e.g., 20-mer) of ssDNA that are commercially available.
      - Other types of DNA and including chemical modifications and attachments are discussed later in this section.

(b) Methods for measuring DEP movement:
- Fluorescence microscopy, as described in the introductory example, has been the standard method for visualizing and measuring DEP of DNA. Suitable dyes for visualization with fluorescence microscopy include DAPI, YOYO, TOTO, acridine orange, and PicoGreen.
- Methods that do not require attachment of a fluorescence dye, i.e., are label-free; these include electrical impedance spectroscopy and special optical techniques, e.g., surface-enhanced Raman spectroscopy (SERS). In the former, the presence of the DNA alters the capacitance between the measurement electrodes. An ingenious arrangement is to use the same electrodes for DEP as for impedance measurement and quantification, as reported by Bier, Hölzel, and coworkers in Germany [15].

(c) Voltage and frequency range:
- Most DEP experiments use standard RF signal generators that range from 0 to 20 MHz and up to 30 V peak to peak. Investigators have tended to use a low-frequency range from 30 to 1 kHz or high-frequency, e.g., 20 kHz to 20 MHz. In principle there are no restrictions except to avoid hydrolysis at low RFs and this, in turn, depends on experimental factors, e.g., solution conductivity. The high-frequency response is influenced by impedance characteristics of the electrodes and supply buses.

(d) Direct DEP force:
- "Traditional" DEP uses metal electrodes that connect directly with buses to the AC supply. The electrodes have often been microfabricated using standard photolithographic methods as described for Figs. 6b and 7. Consequently, planar metal electrodes are typically up to a few hundred nanometers in thickness and are essentially quasi-two dimensional (~2D) with designs that are castellated, corrugated, triangular, sinusoidal, polynomial, etc., to optimize DEP response. The metals used are inert or, at least, biocompatible and can be single- or multilayered. They include gold, aluminum, titanium, platinum, palladium, etc. There are variations on the use of traditional metals; examples are:
  - Metal oxides, such as indium tin oxide (ITO), have become popular to use as material for electrodes in more recent years due to high electrical conductivity combined with optical transparency (e.g., see early work by Fiedler et al. (1998) and Suehiro and Pethig (1998) cited in [1]).
  - To avoid chemical reactions that are likely to occur if low frequencies and high medium conductivities are used, electrode surfaces exposed to the medium are coated with thin layers that act to electrically insulate, e.g., porous hydrogel. An example is given later in this section.
- Carbon nanotube (CNT) end features with diameters as small as 1 nm can achieve high gradients of the square of the electric field magnitude, and their nanosized scale is similar to the diameter of DNA. Consequently, single- or multi-walled carbon nanotubes (SWCNT or MWCNT) are attractive candidate electrodes since their nanoscale features are capable of yielding high DEP forces (see Tuukkanen et al. (2005–2007) cited in [6, 16]).

(e) Other methods for generating DEP forces do so *indirectly*, i.e., use metal electrodes connected to the ac signal supply to generate a potential difference across the space between them. The *supply* electrodes are located sufficiently far away from the nonuniformity, or RoI, so any electrochemical reactions that may occur at the metal/solution medium interface do so without interfering with DEP transport.
- Floating-potential electrodes (FPEs) are microfabricated metal electrodes

positioned between the supply electrodes but *not* directly connected to them. Instead, submerged in electrolyte, they assume the value of the electrical potential according to their position between the supply electrodes and subsequently act to distort the electric field near and around each of them (see authors Asbury and van den Engh (1998), Washizu et al. (1995), and Washizu (2005), cited in [5, 6]).

- Electrodeless DEP (eDEP) or insulating DEP (iDEP), mainly developed during the last decade, uses *insulating boundaries* between the supply electrodes to distort the electric field. Without needing metal microfabricated electrodes to create electric field nonuniformities, metallic degradation near the RoIs is of little concern, so that iDEP enables much lower frequencies, down to 0 Hz (DC), to be applied. Consequently, higher inducible polarizabilities, and hence, higher DEP forces, can be used to move DNA. Typically, device boundaries are made from 3D microfabricated insulating posts made from quartz, polydimethylsiloxane (PDMS), or polymethyl methacrylate (PMMA) (see Chou et al. (2002, 2003), Tegenfeldt et al. (2004), and Swami et al. (2009), cited in [1, 5, 6], [5, 17], and [1, 15], respectively). The relative advantages of iDEP, including the use of plastics that enables injection molding or hot embossing, and opportunities for mass fabrication are described in a very informative review by Regtmeier et al. (2011) cited in [17]. A summary of materials, underlying principles of operation, and applications is also reviewed. Variations of iDEP include:
  - "Liquid electrodes" whereby the electric potential difference across the supply electrodes is taken up by the electrolyte medium, and cross-sectional variations between the insulating walls, housing the medium, govern the shape of the electric field nonuniformity. This method includes involving insulating membranes that capacitively couple adjacent liquid electrodes, thus circumventing joule heating problems that can occur with iDEP. See Regtmeier et al. (2011) cited in [17] that refers to the work by Demierre et al. (2007, 2008) and Shafiee et al. (2009, 2010) and associated groups that introduced the term "contactless" DEP (cDEP), i.e., DEP motion that avoids direct contact of biological specimens with the electrodes.
  - Nanoscale pipettes, or nanopipettes, are used for material transport and nanolithography. Ying et al. (2004) and others of Klenerman's group in the UK, cited in [1, 5, 6], describe, in an elegant series of experiments, DEP combined with EP and EO that enables high-precision pulsatile, in-and outflow of 40-mer ssDNA and dsDNA from the end of a borosilicate glass nanopipette. Other NAs (1 kb ssDNA and a single-nucleotide triphosphate) were also studied. See further, recent, work by the Klenerman group referred to by Regtmeier et al. (2011) cited in [17] for their important contributions to miniaturized biological analysis.
- Optically induced DEP (ODEP) or optoelectronic tweezers (OETs) – as distinguished from "optical laser tweezers" – have been developed where an optical pattern, e.g., from a commercial photo-projector, illuminated onto a photoconductive layer of amorphous silicon, gives rise to a localized DEP force. The optical image has the benefit of acting as an electrode and being able to change position and shape in real time; consequently, it can be titled "virtual electrode." Chiou et al. (2005) initially presented this

important work; it is cited in [18] that describes developments of the method and includes other relevant optical manipulation techniques.

2. DEP of DNA purpose or outcome:
   (a) Electrostatic stretching and trapping of DNA to electrodes using DEP:
      - Temporary positioning: since DNA is flexible on the micro scale, most reported DEP experiments, one way or another, involve trapping and positioning DNA for the duration DEP force is switched on.
      - Semipermanent positioning or immobilization: use of chemical modification to bind the DNA to the electrode or substrate so that it remains tethered after the DEP force is switched off, e.g., avidin binding (by complementary biotin-linked DNA) and thiol labeling of gold electrodes (complementary thiol labeling of DNA strands) (see works by Germishuizen et al. (2003, 2005) and coworkers of Burns in the USA (2002, 2006) that modified DNA stretching with polymer enhanced media, cited in [1, 5, 6]).
      - Applications of stretching and positioning of DNA, typically along electric field lines, include:
        – Measuring the lengths of stretched and oriented dsDNA prepared using restriction enzymes and simple cutting of stretched DNA using a UV laser (see works by Washizu et al. (1990, 1995) cited in [5]).
        – Measuring the conductivity of DNA for molecular electronics (see the work by Burke and coworkers (2003, 2004) cited in [5]).
        – Probing and modeling the mechanical, elastic properties of DNA (see Dalir et al. (2009) cited in [1]).
        – Molecular visualization and elucidation: elegant and important work by Kabata et al. (1993), cited in [6, 13, 15], used pDEP to provide *direct* experimental evidence that RNA polymerase *slides* along DNA. Previously, only indirect evidence had been available and the putative sliding process contrasted with other possible mechanisms, e.g., intersegment transfer, association, and dissociation. Pulsed pDEP (1 MHz, 1 s duration) was used to attract and stretch strands of DNA, with biotin attached at both ends, between 80 and 100 μm apart parallel edges of aluminum electrodes that had been vacuum deposited onto a glass substrate. After the pDEP was switched off, stretched (λ or T7) DNA strands remained anchored onto the glass substrate at both ends by the biotin linkages that had coupled with avidin that had been adsorbed onto the glass surface. The strong avidin-biotin anchoring allowed *E. coli* RNA polymerase to move along the DNA strands and be imaged using a microscope/camera and subsequently analyzed.
        – Molecular surgery: optical tweezer control of a nanosphere coupled to an enzyme, e.g., DNase I, with cutting the bases of electrostatic stretched DNA was demonstrated by Yamamoto et al. (2000) in Washizu's group as cited in [1, 5, 6]. These elegant experimental examples highlight that DNA base information can be used for sub-nanometer position-dependent function in conjunction with DEP.
        – DNA scaffolds: there is a large corpus of literature on DEP of CNTs, metals, their oxides, and fabrication of field-effect transistors (FETs) (see the review by Zhang et al. [16]). In addition to the current protocols on attachment of metal colloids to DNA, e.g., by Mirkin and coworkers or in situ growth on substrates, DEP

has been proposed for a solution-based method of metallized DNA.
  - DNA origami, which often entails folded long ssDNA strands being paired with oligonucleotide ssDNA, has been structured into 2D sheets that can be used as "nanobreadboard" with 6 nm resolution. Examples of 2D sheets, rectangular and "smiley," have been manipulated by DEP and optimal experimental parameters comprehensively reported (see Kuzyk et al. (2008) cited in [6, 16] followed by a review, Kuzyk (2011), cited in [18]).

(b) Separation and purification of DNA:
  - DEP chromatography has been used to separate large (e.g., 48.5 kbp λ-phage dsDNA) and small (22 base ssDNA oligonucleotide) and also medium-large (2 kbp DNA) (see Kawabata and Washizu (2001) cited in [1, 5, 6]). Described simply and referring to the introductory example in Fig. 6b and descriptions [13, 16], the chromatograph is a type of field-flow fractionation (FFF) where the DNA solution, externally driven by a pump, flows with the fluid lateral to the array (in the direction parallel the array-substrate surface) and interacts with the vertically DEP-driven force generated by the electrode array. Large DNA macromolecules with strong polarizability and pDEP force are pulled down onto the array, whereas small DNA fragments with weak polarizability and pDEP force continue to flow though the chamber, hence, separating the mixture.
  - The separation of comparably sized DNA by pDEP is difficult; however, it has been made possible with nanosphere technology. Called "DEP enhancement," Kawabata and Washizu (2001), cited in [1, 5, 6], described the mixing and subsequent DEP-driven separation of 35 kbp T7-DNA with 48.5 kbp λ-DNA (contour length 16.5 μm) that had been linked to 2 μm diameter polystyrene microspheres. Using low conductivity medium (less than 20 mS/m) and energizing frequency 2.5 MHz, the T7-DNA was moved by pDEP, whereas the λ-DNA-bead complexes underwent nDEP, thus separating the mixture. The reason for the λ-DNA-bead complexes being moved by nDEP was attributed to the bead dominating the movement; on their own, microspheres at these experimental conditions undergo nDEP. In contrast, both types of DNA, on their own, are attracted by pDEP so as to span the shortest portions of the interelectrode, corrugated gaps – in a similar manner to the plasmid DNA described in Fig. 6b and 7b.
  - Recently, continuous-flow separation of linear 2.686 and 6.0 kbp DNA, circular topological variants, e.g., parental (6.766 kbp), *mini-circle* 2.247 (kbp) DNA, and DNA complexes (cancer drug and protein) on a microfluidic chip has been reported by Anselmetti's group in Germany [17]. Building on previous work by this group and others, the results are important for advancing the development of gene vaccines and other investigations where DNA macromolecular shape and topology play a significant biological role.
  - Circulating cell-free (ccf) DNA, an important biomarker for cancer and other diseases, has been recently separated from blood and plasma by Heller's group in the USA [19]. In advancing earlier work cited in [19], problems with using high conductivity medium that arises from using blood were circumvented by using planar platinum electrodes that had been coated with a 20 μm layer of porous

hydrogel. The DNA captured by DEP was reported to be compatible to genomic DNA isolated using current clinical "gold standard" methods – thus demonstrating the practical utility of DEP of DNA that could save resources in healthcare.

(c) DEP of DNA has been combined with other biological, chemical, and electrokinetic methods, e.g., EROT, EP, EO, or EOF, and electrostatic liquid actuation. In addition to the work by Klenerman's group mentioned above, three recent and notable examples include:
- Development of DNA concentrators (see Yokokawa et al. (2010) cited in [9, 15]). In that work, the DEP of 48.5 kbp λ-DNA was compared with low-level DC EP (i.e., electrode voltage was sufficiently low, less than a few volts, so that electrolysis did not occur). It was found that low-level DC EP concentrated the λ-DNA better than DEP but the medium conductivity was low (2 mS/m) and possible degradation to the electrodes by using a DC voltage over time is unclear.
- iDEP, achieved by using insulating constrictions, has resulted in a ten-fold enhancement of DNA hybridization kinetics down to 10 p-molar concentration sensitivity using relatively high ionic buffer strengths (see Swami et al. (2009) cited in [1, 15]).
- DEP has also been combined with nanosphere technology to significantly improve the hybridization kinetics of DNA with savings in the use of reagents and with lower ionic strength and simpler protocols than current microarray and Southern blot methods (see Cheng et al. (2010) cited in [1, 9]). This offers a molecular-scale DEP enhancement and offers promising new advances in detection sensitivity, speed, etc., for genomics research and diagnostics.

(d) DEP of DNA in "packages" or as "linked":
- Viruses contain DNA or RNA encapsulated with a protein coat. Viruses have been intensively studied for DEP response, particularly at RFs where the DEP changes from being positive to negative, thus optimizing virus separation (see works by separate groups, e.g., Morgan in the UK and Führ in Germany, and their respective coworkers reviewed in [1, 5, 6]).
- Chromatin that constitutes chromosomes and contains DNA wrapped around histones and other protein has been lysed from bacteria (*E. coli*), transported by DC EP, and trapped by iDEP at 500 Hz, as described by Prinz et al. (2002) cited in [1, 5, 6].
- DNA ligated with PNA and nanospheres showed that DNA could be maneuvered to a unidirectional orientation by DEP and moreover imaged using a fluorescence microscope without the need for intercalating dyes (see Hölzel et al. (2003) cited in [1, 5, 6]).

Dielectrophoresis of RNA

DEP has been used to rapidly collect RNA from the nucleus of a living cancer cell using atomic force microscope (AFM) [20]. The AFM cone-shaped probe tip had been structurally modified to form nanometer-scale concentric electrodes that enabled mRNA to be attracted and subsequently extracted from the cell nucleus. This novel and promising application of DEP may pave the way for enhanced analysis of single-cell extracts with many applications, particularly, as mentioned in section "Nucleic Acids," the mRNA forms only a small proportion of the total RNA in a cell.

Time-dependent DEP collections on of 16S and 23S subunit rRNA, extracted from *E. coli*, onto planar interdigitated electrodes over a wide RF range, 3 kHz – 50 MHz, have been reported by the Pethig group in the UK [9]. The in-depth study used total internal reflection fluorescence (TIRF) microscopy with an aqueous suspension of 13 mS/m conductivity. The study revealed a positive to negative DEP transition above 9 MHz and confirmed a quadratic dependence of pDEP force with respect to voltage for low values ($3V_{rms}$) – as

predicted by Eq. 19. Being the first demonstration of fast capture and release of rRNA, DEP shows considerable promise for application in rRNA-based biosensing devices.

Other reports at present have involved DEP of RNA in context, i.e., not as the main focus of study but in conjunction with DEP of DNA, or of cells, e.g., the lysing of cell contents (Cheng et al. (1998) cited in [6, 14]) where cells undergo DEP positioning, or investigations on the effect of DEP on gene expression in cells, etc.

Outlook and Concluding Remarks

There have been many achievements and DEP of NAs, particularly DNA, continues to be an active and promising research field with a scope that has advanced with new materials, e.g., PDMS, PMMA, ITO, CNTs, etc., and technologies, e.g., microfluidics, nanospheres, microarrays, electronic nanodevices, and scanning probe microscopes. Detection methods have also advanced, be it either electrical, e.g., impedance spectroscopy, or optical, e.g., SERS and TIRF. There is substantial opportunity for advancing theoretical understanding of polarization mechanisms that underpin DEP-induced motion of NAs; at present, current models favor counterion fluctuation. There is also considerable opportunity for advancement of nanoscale DEP applications, whether it is directed toward nano-construction or advancing DNA hybridization kinetics that is needed in genomics and sensor research. DEP of RNA appears to be research field open for investigation, particularly with the worldwide drive for single-cell analysis and quantitative systems biology. There is also much to be accomplished in terms of integrating DEP-based methods with other electrokinetic, LOC, and μTAS miniaturization technologies and establishing reliable and quantifiable measures of functionality and performance.

Cross-References

- ▶ AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation
- ▶ AC Electrokinetics of Nanoparticles
- ▶ Applications of Nanofluidics
- ▶ Automatic Data Analysis Workflow for RNAi
- ▶ Biornametics: Architecture Defined by Natural Patterns
- ▶ Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering
- ▶ Computational Systems Bioinformatics for RNAi
- ▶ Concentration Polarization at Micro-/Nanofluidic Interfaces
- ▶ Dielectrophoresis
- ▶ Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly
- ▶ Dielectrophoretic Assembly
- ▶ Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes
- ▶ DNA Computing
- ▶ DNA FET
- ▶ DNA from First Principles
- ▶ DNA Manipulation
- ▶ DNA Manipulation Based on Nanotweezers
- ▶ DNA Origami as Programmable Nanofabrication Tools
- ▶ Electrokinetic Fluid Flow in Nanostructures
- ▶ Electrokinetic Manipulation
- ▶ Implementing Alternating Nanolaminates in Trenched Energy Storage Systems
- ▶ Micro- and Nanofluidic Devices for Medical Diagnostics
- ▶ Micro-/Nanofluidic Devices
- ▶ Monolithic Total Internal Reflection Biosensor
- ▶ Nanochannels for Nanofluidics: Fabrication Aspects
- ▶ Nanofluidic Channels
- ▶ Oligonucleotide Amplification
- ▶ RNA Interference (RNAi)
- ▶ RNAi in Biomedicine and Drug Delivery
- ▶ Short-Interfering RNA (siRNA)
- ▶ siRNA Delivery
- ▶ Surface-Modified Microfluidics and Nanofluidics
- ▶ Systems Level Data Mining for RNAi
- ▶ Total Internal Reflection (Fluorescence) Velocimetry
- ▶ Uptake/Internalization/Sequestration/Biodistribution

References

1. Pethig, R.: Review article-dielectrophoresis: status of the theory, technology, and applications. Biomicrofluidics **4**(2), 35 (2010)
2. Dahm, R.: Discovering DNA: friedrich miescher and the early years of nucleic acid research. Hum. Genet. **122**(6), 565–581 (2008)
3. Nelson, D.L., Cox, M.M.: Lehninger Principles of Biochemistry, 6th edn. WH Freeman, New York (2013)
4. Pohl, H.A.: Dielectrophoresis. Cambridge University Press, Cambridge (1978)
5. Lapizco-Encinas, B.H., Rito-Palomares, M.: Dielectrophoresis for the manipulation of nanobioparticles. Electrophoresis **28**(24), 4521–4538 (2007)
6. Hölzel, R.: Dielectric and dielectrophoretic properties of DNA. IET Nanobiotechnol. **3**(2), 28–45 (2009)
7. Calladine, C.R., et al.: Understanding DNA – The Molecule and How it Works, 3rd edn. Academic, London (2004)
8. Bloomfield, V.A., Crothers, D.M., Tinoco Jr., I.: Nucleic acids: Structures, Properties, and Functions. University Science, Sausalito (2000)
9. Giraud, G., et al.: Dielectrophoretic manipulation of ribosomal RNA. Biomicrofluidics **5**, 24116 (2011)
10. Takashima, S.: Electrical Properties of Biopolymers and Membranes. Adam Hilger, Philadelphia (1989)
11. Grant, E.H., Sheppard, R.J., South, G.P.: Dielectric behaviour of biological molecules in solution. In: Harrington, W.F., Peacocke, A.R. (eds.) Monographs on Physical Biochemistry. OUP, Clarendon Press, Oxford (1978)
12. Pethig, R.: Dielectric and Electronic Properties of Biological Materials. Wiley, New York (1979)
13. Morgan, H., Green, N.G.: AC Electrokinetics: Colloids and Nanoparticles. Research Studies Press/IoP Publishing, Baldock/Philadelphia (2003)
14. Bakewell, D.J., Morgan, H.: Dielectrophoresis of DNA: time- and frequency-dependent collections on microelectrodes (vol 5, pg 1, 2006). IEEE Trans. Nanobioscience **5**(2), 139–146 (2006)
15. Henning, A., Bier, F.F., Hölzel, R.: Dielectrophoresis of DNA: quantification by impedance measurements. Biomicrofluidics **4**(2), 9 (2010)
16. Zhang, C., et al.: Dielectrophoresis for manipulation of micro/nano particles in microfluidic systems. Anal. Bioanal. Chem. **396**(1), 401–420 (2010)
17. Viefhues, M., Regtmeier, J., Anselmetti, D.: Fast and continuous-flow separation of DNA-complexes and topological DNA variants in microfluidic chip format. Analyst **138**(1), 186–196 (2013)
18. Kayani, A.A., et al.: Optofluidics incorporating actively controlled micro- and nano-particles. Biomicrofluidics **6**(3), 32 (2012)
19. Sonnenberg, A., et al.: Dielectrophoretic isolation and detection of cancer-related circulating cell-free DNA biomarkers from blood and plasma. Electrophoresis **35**(12–13), 1828–1836 (2014)
20. Nawarathna, D., et al.: Targeted messenger RNA profiling of transfected breast cancer gene in a living cell. Anal. Biochem. **408**(2), 342–344 (2011)

Dielectrophoretic Assembly

▶ Electric-Field-Assisted Deterministic Nanowire Assembly

Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes

Didi Xu[1], Arunkumar Subramanian[2], Lixin Dong[3] and Bradley J. Nelson[1]
[1]Institute of Robotics and Intelligent Systems, ETH Zurich, Zurich, Switzerland
[2]Department of Mechanical and Nuclear Engineering, Virginia Commonwealth University, Richmond, VA, USA
[3]Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA

Synonyms

Dielectrophoresis; Electrokinetic manipulation; Nanoassembly; Nanoelectrodes

Definition

Dielectrophoreis is an electrokinetic manipulation process where the interactions of a nonuniform electric field with a polarizable particle are employed to localize the particle in space with nanoscale precision. The nonuniform field strength around the particle exerts a dielectrophoretic (DEP) force on the dipole induced within the polarizable particle, which drives the particle toward the electric field maximum or minimum. In addition, the resulting torque aligns the particle along the direction of the applied electric field.

Dielectrophoretic assembly has been widely used for high precision assembly of carbon nanotubes (CNTs) in order to overcome disadvantages such as the need for a complex setup and poor controllability, which are associated with other techniques to integrate CNTs onto functional nanostructures. Since this technique is performed at room-temperature, in a noncorrosive fluidic environment, and at low voltage, the electric field-based method is compatible with further postintegration nanostructuring steps such as those used in the Integrated Circuit (IC) industry.

Overview

Functional nanostructures, which incorporate an individual multiwalled carbon nanotube (MWNT) as the fundamental electromechanical transduction element, can be used to create tubular switches, memories, sensors, and resonators/oscillators with improved performance. This entry highlights two key unit processes that enable the fabrication of such constructs and provides an overview of potential applications using recently demonstrated nanomechanical switches. The first unit process relates to dielectrophoretic nanoassembly of single MWNTs onto arrays of nanoelectrode pairs fabricated on silicon chips [1, 2]. Appropriately optimized electrode designs insure that the assembled MWNT bridges the nanoelectrodes at its distal ends while remaining fully suspended and flat in the region between them (Fig. 1).

The second unit process involves the modification of the nested-shell, native MWNT geometry using Joule-heating-induced vaporization of carbon atoms at specific locations on the nanotube [2–4]. By engineering heat sinks along the NT length to modulate thermal dissipation and Joule-heating-induced temperatures, the vaporization of carbon atoms within the NT has been controlled with atomic-to-nanoscale [5]. This technique to selectively modify the NT geometry with control over location, length, and shells has provided us with the capability to create structures where some shells can axially translate or rotate with respect to others. Nanoelectromechanical systems (NEMS) devices based on such intershell displacement mechanisms have been demonstrated to be ultrafast, fatigue-free constructs with very low friction.

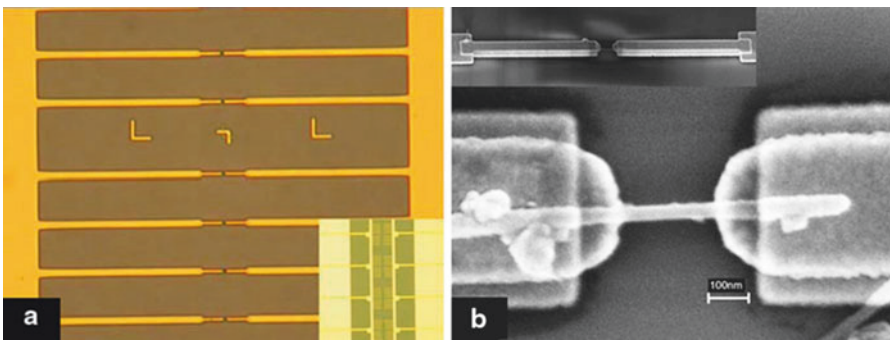
Basic Methodology

DEP Assembly Technique

Based on Pohl's theory [6], dielectrophoretic forces F_{DEP} and torques T_{DEP} can be calculated as

$$F_{DEP} = (p \cdot \nabla)E \tag{1}$$

$$T_{DEP} = p \times E \tag{2}$$



Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes, Fig. 1 Dielectrophoretic assembly of MWNTs. (**a**) Nanoelectrode array design with a lower magnification image in inset. (**b**) SEM images of an assembled MWNT nanostructure at high and low magnification (Reproduced with permission from Ref. [2], Copyright 2007 IOP)

where p is the induced dipole moment of the particle and E is the electric field.

Because both DEP forces and torques are dependent on multiple parameters, such as particle and suspending medium properties and the applied electric field, the nanotube (NT) is modeled as a prolate ellipsoid with homogeneous dielectric properties. The large aspect ratio of NTs results in a needle-shaped particle with $l \gg r$, so the governing equations in dielectrophoresis could be simplified as [7, 8]

$$F_{DEP} = \frac{1}{2}\pi r^2 l \varepsilon \mathrm{Re}\left(\frac{\widetilde{\varepsilon}_p - \widetilde{\varepsilon}_m}{\widetilde{\varepsilon}_m}\right)\nabla|E|^2 \qquad (3)$$

$$\langle T_3 \rangle = \frac{1}{2}\pi r^2 l \varepsilon_m E^2 \sin\theta\cos\theta \mathrm{Re}\left(\frac{\left(\widetilde{\varepsilon}_p - \widetilde{\varepsilon}_m\right)^2}{\widetilde{\varepsilon}_m\left(\widetilde{\varepsilon}_m + \widetilde{\varepsilon}_p\right)}\right) \qquad (4)$$

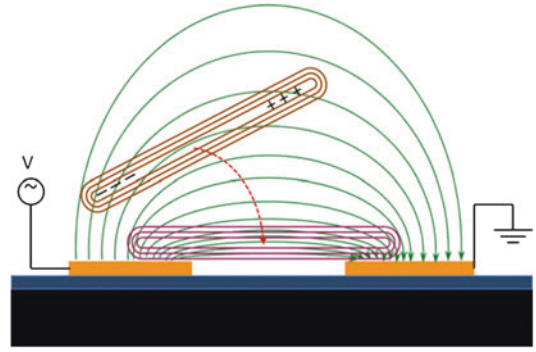
where $\widetilde{\varepsilon}_p$ and $\widetilde{\varepsilon}_m$ are complex dielectric permittivities of the particle and the suspending medium, r and l are the radius and length of the CNT, and the angle between the electric field and the longest axis 1 is θ.

Another important aspect influencing the motion of CNTs is the viscous drag force, which is proportional to the relative velocity, retards the CNT in all directions. If gravity, Brownian motion, electro-osmosis, etc., are all negligible during DEP process, the instantaneous translation velocity is proportional to the instantaneous dielectrophoretic force [3, 4]. Similarly, the rotational friction factor of the CNT rotating about its center in x–y plane can also be calculated by hydrodynamic equations [9]. The schematic is shown in Fig. 2.

$$v_T \approx \frac{\boldsymbol{F}_{DEP}}{f} \qquad (5)$$

The constant f is the translation friction factor and depends on a range of parameters such as size, shape, and fluid viscosity.

The standard operation procedure of DEP assembly is as follows. First, the MWNTs are suspended and sonicated in ethanol (at a concentration of 5 μg/ml) for 10 h to insure homogeneity.



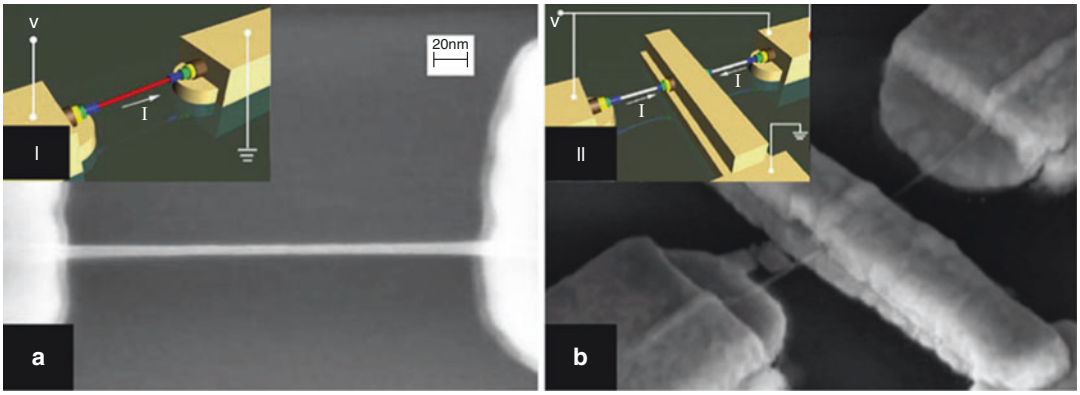
Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes, Fig. 2 Dielectrophoresis concept

The chip is then immersed in a reservoir containing this suspension and a nonuniform electric field is applied with a high frequency function generator. After deposition, the chip is removed from the reservoir, rinsed in clean ethanol, and brown-dried with nitrogen. The assembly parameters related to the target nanomaterial and nanoelectrode structure such as the AC/DC voltage components, polarizing frequency, and deposition time are modulated to optimize device deposition yields.

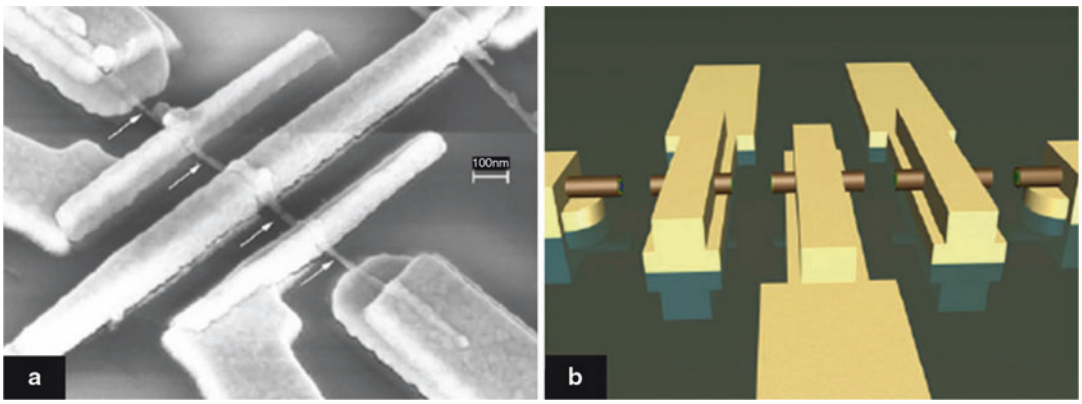
Current-Induced Shell Engineering of MWNT

Current driven shell engineering is useful in cases to employ Joule heating to vaporize carbon atoms and thin the NT between device contacts in order to tune its electrical resistance, mechanical stiffness, and clamping configuration [2]. A typical DEP assembled MWNT is sandwiched between two layers of metal at each end to improve electrical contact. The metallic contacts serve as heat sinks, and shell removal due to thermal stress occurs in the suspended segment, starting from the outermost and proceeding towards the inner ones. During current-induced vaporization, NT temperatures are estimated to be as high as 2,000 K. It is proven that at these elevated temperatures, defects such as kinks and holes are initiated at MWNT mid-lengths and result in individual shell removal in the region between contacting electrodes [3, 4]. An SEM image of a shell-structured nanotube is shown in Fig. 3a.

Modulation of Joule heat dissipation along NT lengths using additional heat sinks can be



Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes, Fig. 3 Formation of nanobearing with different architectures. (**a**) Bearings created by partial electric breakdown between two metallic contacts. (**b**) Piecewise shell engineering of single NTs between metallic contacts and floating contacts (Reproduced with permission from Ref. [2], Copyright 2007 IOP)



Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes, Fig. 4 High-density NT bearings. (**a**) Telescoping segments formed with a 220 nm pitch and separated by 6–10 nm gaps. The *arrows* point to the inter-segment gaps. (**b**) Schematic illustration of the core shell mechanisms formed with the inter-segment gaps exaggerated to reveal the shell structure (Reproduced with permission from Ref. [2], Copyright 2007 IOP)

employed to achieve atomic-to-nanoscale control over shell etching locations on a MWNT [5]. It is found that the electric breakdown position can be controlled in MWNTs by introducing additional metallic contacts located at different segments of a NT. A nanotube assembled onto three contacts is shown in Fig. 3b, and a schematic is shown in its inset. By applying a bias to the electrodes at distal ends and grounding the central metal contact (Fig. 3b), currents are driven through both suspended segments of the NT simultaneously and remove its outer shells in the respective segments. Detailed information on this technique can be found in references [2] and [5].

Key Research Findings

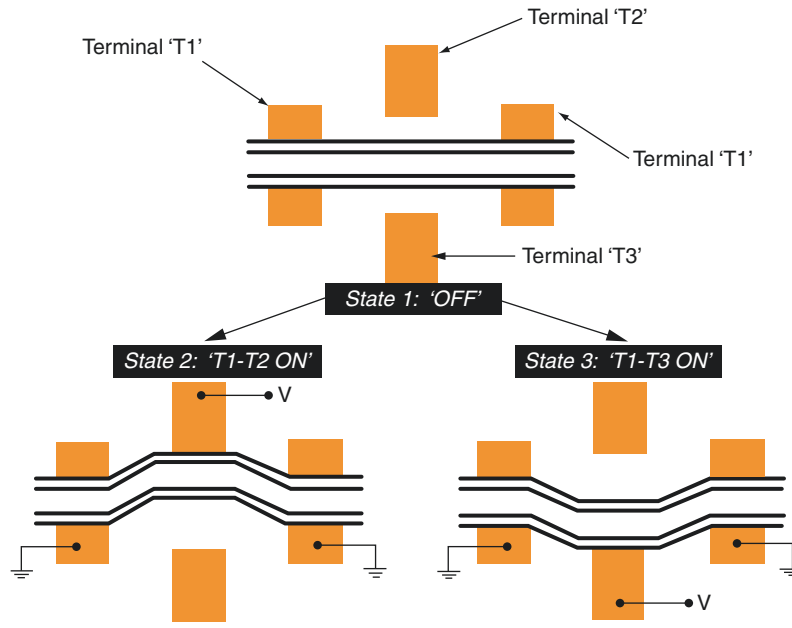
Nanotube Bearings

High bearing densities can be achieved through Joule-heat-driven vaporization of different segments of a NT, that is, a NT is assembled onto multiple metallic contacts using floating electrode DEP [2, 10]. Figure 4a shows independent,

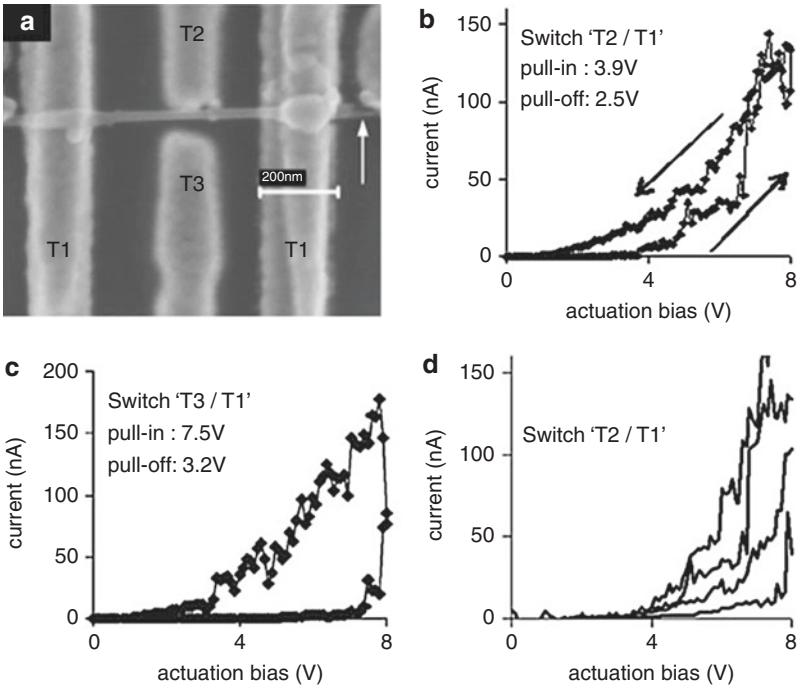
bidirectional nanobearings engineered at different locations along the length of a single MWNT bridging five electrodes. With a 220 nm device length, these constructs represent the smallest bearings reported to date and are expected to exhibit lower friction forces and faster response times during telescoping core movements. Each of the three NT segments in the central part of the fabricated device is anchored to the metal at only its outer shells, with inner shells capable of sliding inside the outer housing in both directions. An illustration of this shell structure is shown in Fig. 4b. This structure has a number of potential applications. One possible application involves resonators based on shell displacements, where neighboring nanotube segments electrically excite oscillations. The configurational stability and frequency spectrum of these constructs have been investigated. The most significant conclusion that emerges from the analysis of these open-capped structures is that, for a given NT (in terms of the number of shells and outer diameter), there exist only a finite number of configurations that are energetically stable. For an 11 shell 8 nm outer diameter CNT, resonant frequencies from 1.4 to 4.9 GHz for various stable inner shells exist [11]. Using these constructs, ultralow power electromechanical switches can be realized by electrostatically controlling the movement of NT cores and thereby, modulating the intersegment conductivity [10].

Nanotube Switch

Another energy-efficient, three-state NEMS switch which has been realized using the DEP assembled nanostructures is shown in Fig. 5 [12]. In this device, the contacts at the NT's distal ends are held at a common, ground potential and form the first terminal, T1, of the NEMS switch. The gate electrodes, which are located on either side of the NT in close proximity, represent terminals "T2" and "T3," respectively. In the device "off-state," the NT is not in contact with the gate electrodes on either side. When a voltage is applied to one of the gate electrodes, it induces an attractive electrostatic force on the nanotube and deforms it elastostatically. Beyond a threshold bias, which is characteristic of the NT Young's modulus, diameter, number of shells, length and device actuation gap, the NT comes in the contact with the gate electrode and the switch is turned "ON." Thus, the circuit between terminals T1_T2 and T1_T3 can be turned ON by applying the pull-in voltage at the respective actuation electrodes.



Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes, Fig. 5 Three states of the device. The "OFF state" in the absence of external bias is shown on *top*. The "T1-T2 ON" state when a voltage is applied to terminal "T2" is shown in *bottom left*. The *bottom-right* image illustrates the "T1-T3 ON" state when a bias is applied to terminal "T3" (Reproduced with permission from Ref. [11], Copyright 2009 ACS)



Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes, Fig. 6 Device characterization. (**a**) SEM image of a device with terminal labeling shown. The arrow points to a current-induced break in the NT to decouple it from the array. (**b**) I-V characteristics of the switch between terminals "T1-T2." The *arrows* point to the increasing and decreasing voltage parts of the curve. (**c**) I-V characteristics of the switch between terminals "T3-T1." (**d**) Device operation over four cycles shown to demonstrate the repeatability in device behaviors (Reproduced with permission from Ref. [11], Copyright 2009 ACS)

An SEM image of a representative device is shown in Fig. 6a, and its operational characteristics are shown in Fig. 6b–d. The *I-V* curves during the switching transition initiated by a bias applied to terminal T2 is shown in Fig. 6b. One important characteristic that emerges from these curves is the hysteresis in device operation. This is due to the impact of van der Waals forces between the NT and gate electrode, which is onset at pull-in and opposes the subsequent pull-off event. The switching characteristics of the device constructed between terminals T3 and T1 can be seen in Fig. 6c. The pull-in voltage for this terminal was found to be higher as compared to that at terminal "T2" due to the larger NT-electrode gap at this terminal. The repeatability of these devices is highlighted in Fig. 6d using results from four cycles of switching induced by terminal T2. It can be observed from this plot that the turn-on bias remains repeatable.

Future Directions for Research

Results highlighted in this entry provide interesting pointers toward future nanodevices based on single MWNTs with novel performance regimes. Though significant progress has been made, future avenues for research and development exist in each of the following thrust areas: nanofabrication, device engineering, and characterization. For instance, though DEP-based assembly has been demonstrated to be a robust method to assemble diverse materials with nanoscale precision, further improvements in assembly yields are required before these could be used within industrial processes. This is an important challenge that requires the identification of optimum deposition parameters within the multi-variable process.

In the area of device development, a number of different engineering embodiments can be

realized using nanofabrication processes outlined in this report. These include microwave resonators based on elastostatic deformations and inter-shell displacements within MWNTs. These devices have the potential for use in near-term industrial applications such as mass, force, and resonant-pressure sensors. Another possible area of future research relates to developing integrated MEMS/NEMS platforms for probing the electromechanical properties of nanomechanical elements. For instance, appropriately designed constructs realized using a combination of DEP assembly and shell engineering can be used to understand the electromechanical coupling between individual shells within a MWNT and to provide insights into its suitability for use as piezoresistors in nanotransducers.

Cross-References

- ▶ AC Electrokinetics
- ▶ AC Electroosmosis: Basics and Lab-on-a-Chip Applications
- ▶ Dielectrophoresis
- ▶ Self-Assembly

References

1. Xu, D., Subramanian, A., Dong, L.X., Nelson, B.J.: Shaping nanoelectrodes for ultrahigh precision dielectrophoretic assembly of carbon nanotubes. IEEE Trans. Nanotechnol. **8**(4), 449–456 (2009)
2. Subramanian, A., Dong, L.X., Tharian, J., Sennhauser, U., Nelson, B.J.: Batch fabrication of carbon nanotube bearings. Nanotechnology **18**(7), 075703 (2007)
3. Huang, J.Y., Chen, S., Jo, S.H., Wang, Z., Han, D.X., Chen, G., Dresselhaus, M.S., Ren, Z.F.: Atomic-scale imaging of wall-by-wall breakdown and concurrent transport measurements in multiwall carbon nanotubes. Phys. Rev. Lett. **94**(23), 236802 (2005)
4. Molhave, K., Gudnason, S.B., Pedersen, A.T., Clausen, C.H., Horsewell, A., Boggild, P.: Transmission electron microscopy study of individual carbon nanotube breakdown caused by Joule heating in air. Nano Lett. **6**(8), 1663–1668 (2006)
5. Subramanian, A., Choi, T.-Y., Dong, L.X., Tharian, J., Sennhauser, U., Poulikakos, D., Nelson, B.J.: Local control of elctric current driven shell etching of multiwalled carbon nanotubes. Appl. Phys. Mater. Sci. Process. **89**(1), 133–139 (2007)
6. Pohl, H.A.: Dielectrophoresis the Behavior of Neutral Matter in Nonuniform Electric Fields. Cambridge University Press, Cambridge (1978)
7. Jones, T.B.: Electromechanics of Particles. Cambridge University Press, Cambridge (1995)
8. Morgan, H., Green, N.G.: AC Electrokinetics: Colloids and Nanoparticles. Research Studies, Hertfordshire (2003)
9. Koenig, S.H.: Brownian motion of an ellipsoid. A correction to Perrin's results. Biopolymers **14**, 2421–2423 (1975)
10. Subramanian, A., Dong, L.X., Nelson, B.J., Ferriera, A.: Supermolecular switches based on multiwalled carbon nanotubes. Appl. Phys. Lett. **96**(7), 073116 (2010)
11. Subramanian, A., Dong, L.X., Nelson, B.J.: Stability and analysis of configuration-tunable, bi-directional MWNT bearings. Nanotechnology **20**(49), 495704 (2009)
12. Subramanian, A., Alt, A.R., Dong, L.X., Kratochvil, B.E., Bolognesi, C.R., Nelson, B.J.: Electrostatic actuation and electromechanical switching behavior of one-dimensional nanostructures. ACS Nano **3**(10), 2953–2964 (2009)

Dielectrophoretic Phenomena

▶ AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation

Differential Scanning Calorimetry

▶ Nanocalorimetry

Digital Microfluidics

▶ Electrowetting

Dipolar Particle Chaining

▶ AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation

Dipole–Dipole Particle Interaction

► AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation

Dip-Pen Nanolithography

Yi Zhang[1], Rüdiger Berger[2] and Hans-Jürgen Butt[2]
[1]Shanghai Institute of Applied Physics, Chinese Academy of Sciences, Shanghai, China
[2]Max Planck Institute for Polymer Research, Mainz, Germany

Synonyms

Hard-tip soft-spring lithography; Polymer pen lithography; Scanning-probe lithography

Definition

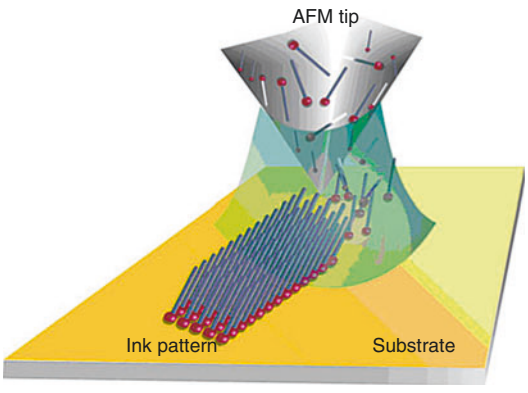
Dip-pen nanolithography (DPN) is an atomic force microscope (AFM) [1]) based patterning technique in which molecules are transported from a sharp tip to a surface [2, 3]. With DPN, the molecules to be deposited are first loaded onto the tip by dipping the tip into a melt or solution of the substance to be deposited. Similar to a pen which people use to write notes on a paper, the atomically sharp tip is scanned in contact with a sample surface. The ink molecules are transported to the surface driven by capillary forces normally generated by the condensed water or melted ink when tip contacts surface. Typically, DPN allows surface patterning with a feature size on scales of less than 100 nm [4]. DPN is a quite unique method for patterning because it depends on the delivery of molecules to surfaces rather than energy (in the form of light, force, heat, or electricity) to make patterns. The DPN principle can be called a positive lithography technique in the sense that material is added rather than removed.

The History of DPN

The research on the deposition of organic materials by the AFM tip to a solid surface can be dated back to 1995 [2]. However, the first use of the term "dip-pen nanolithography" and the fast development of the DPN technique started in 1999 [3] in Prof. Chad Mirkin's lab in Northwestern University. Since then the excellent attributes of the DPN, such as direct-write capability, ultrahigh resolution, molecular ink and substrate general, and programmable abilities, have attracted researchers all over the world. Currently, several tens of groups are employing DPN technique for their researches and contributing on the development of the DPN technique at all aspects. In the last 5 years, a constant value of around 60–70 articles per year was published under the term "DPN." This value characterizes DPN to be a method of scientific importance and not to be a scientific niche. Meanwhile, commercialized instruments for parallel patterning of ink materials have been developed in early 2000s.

Description of the Method

The pens. Conventional DPN utilizes an AFM probe to transport ink materials from an ink-coated tip to a surface when the tip is engaged (Fig. 1). The size of AFM tip is on the scale of tens of



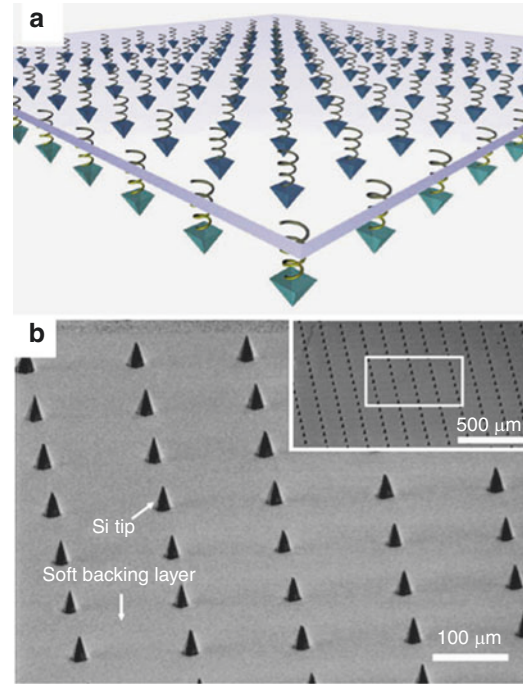
Dip-Pen Nanolithography, Fig. 1 Schematic drawing indicating ink transporting from an AFM tip onto a solid substrate (Image was reprinted from [5] with permission)

nanometers to several micrometers. Accordingly, the feature size fabricated by the tip can be controlled in the range from about 10 nm to several micrometers by selecting suitable tips and depositing parameters. The motion of the AFM tip on the surface can be precisely controlled by the AFM system, which makes it convenient to fabricate programmed patterns. The same AFM tip also serves as an imaging tool for characterizing written nanoscopic features on a surface.

In order to control the deposition of molecules, specialized AFM tips were fabricated. For example, AFM cantilevers terminated with poly(methyl methacrylate) (PMMA) colloids have been used to pattern molecules by allowing the polymer on the tip to swell under different humidity conditions [6]. Even hollow tips connected with a reservoir were used for controlling the droplet size for liquid nano dispensing applications [7]. New procedures based on novel probes boost the potential of DPN into different research areas.

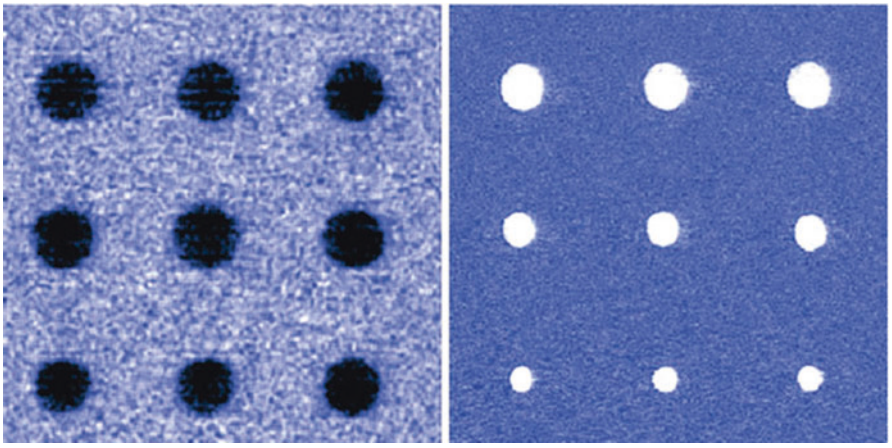
As a patterning technology, throughput is the biggest limitation of the single-probe DPN. Thus, parallelization capabilities must be developed in order to be competitive with standard lithographic techniques [8]. A possible solution to DPN parallelization is to develop, to fabricate, and to use two-dimensional (2D) tip arrays [9]. Linear and 2D arrays carrying more than 10,000 tips have been developed [8] and are commercially available. These tools allow one to fabricate molecule-based structures over long distances (centimeters) or large areas (square centimeters) with high efficiency. Moreover, the pens in a tip array can be independently actuated and used to make a different nanostructure, which provides more flexible control in DPN. The material that makes up the parallel tip arrays can be soft polymers [10]. The most recent development in this field indicated that a hard silicon tip array on a soft baking layer can overcome the throughput problems of conventional DPN and the resolution limits imposed by the use of soft polymer tips (Fig. 2) [11].

Ink materials. With DPN, "ink" substances are patterned to endow surfaces with predesignated functions at specific locations. Small molecules such as alkanethiols were first used as ink materials to be patterned on gold surface. Alkanethiol molecules deposited on gold surfaces normally undergo a self-assembly process and form a closely packed monolayer. Then DPN has been developed to pattern a variety of "ink–substrate" combinations. Inks studied to date include many types of small molecules, metal salts, nanoparticles, polymers, oligonucleotides, peptides, and proteins, while the substrates used include gold, glass, quartz, silicon, mica, germanium, and so on. In addition, researchers also have developed DPN techniques to pattern different chemical [12] or biological ink molecules on one substrate, so that multiple chemical or biochemical functionalities are integrated [4].

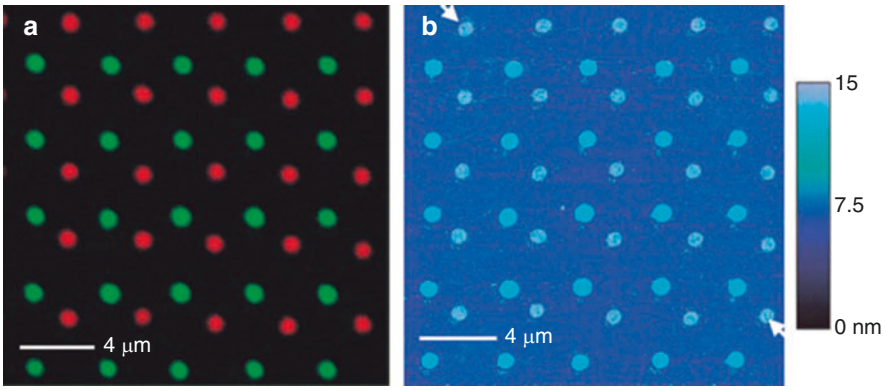


Dip-Pen Nanolithography, Fig. 2 (**a**) Schematic illustration of a hard tip array on a soft baking layer (the springs). (**b**) An SEM image of a real silicon tip array on a SiO_2/PDMS/glass substrate. The *inset* shows a large area of the array (Image was reprinted from [11] with permission)

Alkanethiols were the first type of chemicals [1, 12] that were directly patterned on gold surface with DPN (Fig. 3). Alkanethiol molecules are important because they can be used to conveniently tailor the local chemical properties of the substrate by selecting appropriate tail group of the molecule. In addition, alkanethiols are easy to



Dip-Pen Nanolithography, Fig. 3 Lateral force microscopy images of two dot arrays made by two alkanethiols, 1-octadecanethiol (*left*) and 16-mercaptohexadecanoic acid (*right*), respectively, on thermally evaporated gold films on silicon oxidized substrates



Dip-Pen Nanolithography, Fig. 4 Direct patterning of multiple-DNA inks by DPN. (**a**) Fluorescence image of two different DNA arrays. The arrays were made in two steps, each with an AFM tip that was coated with one DNA ink. After the first array was completed, the AFM tip was retracted and replaced with the second one coated with another DNA ink. The second tip was then relocated to the region of the first DNA array and fabricated the second array in-between the dots of the first array. (**b**) AFM image of the two DNA arrays shown in (**a**) decorated with gold nanoparticles. The DPN arrays were treated with a solution containing two different particles that were modified with DNA complementary to the first and second arrays, respectively. The two particles selectively assembled on the correct arrays (Image was reprinted from [14] with permission)

handle during tip coating and can be repeatable patterned on gold surface. Since then, alkanethiols have been widely explored as a model system for DPN to investigate the ink transport through AFM tip, molecular diffusion on the substrates, and phase separation of mixed ink molecules.

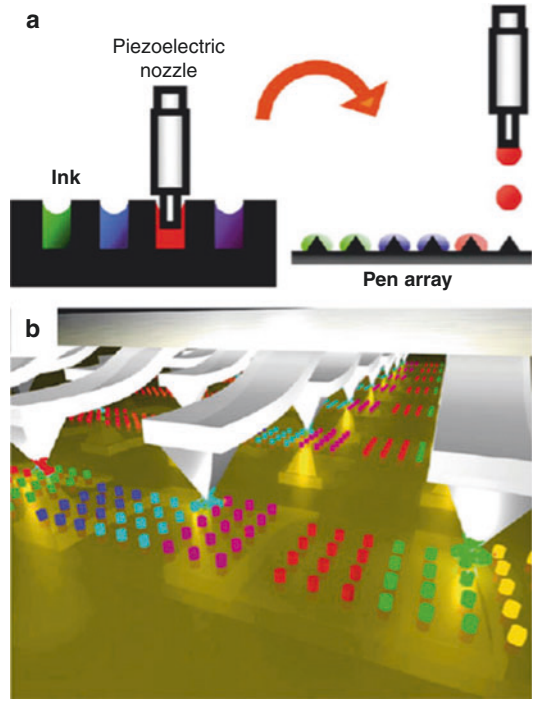
Patterning of biomolecules on various surfaces with DPN is particularly attractive [13]. The biomolecules can be patterned in a direct-write manner in either ambient or inert environments without risking cross-contamination. Since DPN has a very high spatial resolution, it is capable of fabricating high-density micro- or nano-arrays (Fig. 4) for biomolecular assays, including DNA and protein detection. The high spatial resolution that DPN brings to this field would allow an assay to screen for a correspondingly larger number of targets, or allow a fixed number of targets to be screened with a correspondingly smaller sample volume in a shorter period of time.

Tip coating. Tip coating is an important step of a DPN process. Researchers have developed a series of methods to coat AFM tips with molecular inks. For soluble ink materials, one can dip the tips into a dilute solution containing the ink molecules. When the tip is taken out from the solution, the solvent should be removed, normally with a gas flow, prior to deposition. Coating of the tip may be conducted by placing the tips in the ink vapor and by cooling the tip. For liquid inks, tip coating is relatively simple, and dip-coating is normally adopted.

To facilitate tip inking and to increase the quantity of adsorbed ink, researchers have developed novel tip modifications methods. For example, AFM tips made of or coated with poly (dimethylsiloxane) (PDMS) can load more ink than bare AFM tips. Nowadays, researchers have developed microscopic inkwells as well as inkjet printing (Fig. 5) techniques that could be used to ink the different pens in a tip array for simultaneous DPN patterning of multiple inks [15].

Ink transport. Ink molecules transport from coated tip to solid substrate driven by a capillary force. It is a complicated process and is influenced by a lot of parameters. These include the chemical makeup of the ink, the substrate, and the tip, which affect the distribution and mobility of the ink on the tip and substrate, and the environment conditions under which the experiment is carried out. Since the ink-coated tip contact the substrate during DPN, a water meniscus would normally condense between a hydrophilic tip and a hydrophilic surface under room temperature and ambient conditions. It has been found that relative humidity of the environment affects ink transport [16], and today most DPN is conducted in a humidity-controlled chamber that helps to regulate ink transport. In a water-free or solvent-free environment, molecule transfer is hampered but not impossible. It is evident that different molecules are transported to the surface by fundamentally different mechanisms, which may be attributed to the different size, shape, solubility, and diffusion constant of the ink molecules.

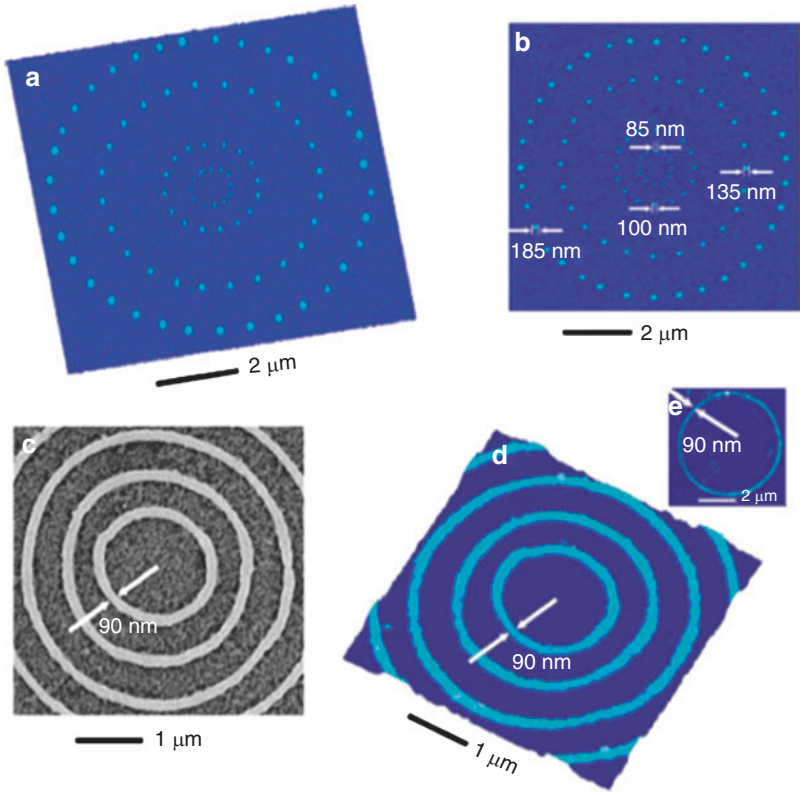
To a better control over the ink transport, researchers have developed approaches that use ink chaperones to overcome the solubility and surface interactions that previously inhibited uniform pattern formation by DPN. One approach uses poly(ethylene glycol) (PEG) as a matrix to facilitate the transport of inks with different sizes and solubilities to an underlying substrate. Some others use viscous liquids such as glycerol, dimethylformamide, and so on, to assist ink transport. In some cases, the use of a liquid will also benefit biomolecule inks to keep their biological functions during DPN process.



Dip-Pen Nanolithography, Fig. 5 Schematic drawings for (**a**) inking of pen arrays by inkjet printing and (**b**) multiplexed dip-pen nanolithography. After the nozzle of the ink jet printer was inked, the arrayed pens were independently addressed with different inks using the inkjet printer. The inked pens then make patterns containing multiplexed compositions on the surface (Image was reprinted from [15] with permission)

Applications of DPN

By depositing ink molecules to surfaces, DPN has the ability to create architectures with controlled chemical composition and physical structure. With significant advances in the parallel printing



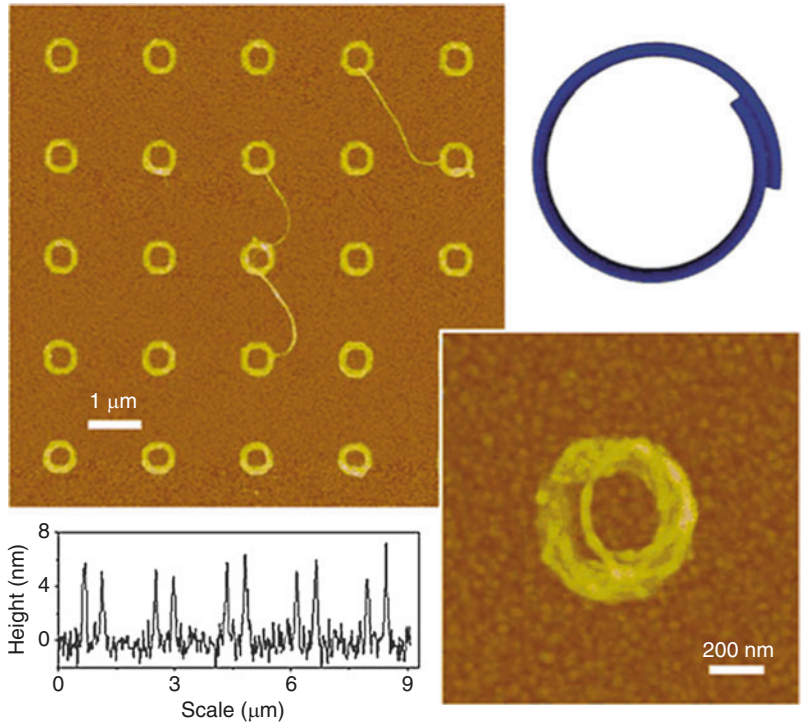
Dip-Pen Nanolithography, Fig. 6 AFM images (**a**, **b**, **d**, **e**) and SEM image (**c**) of Au nanostructures on a silicon substrate (Image was reprinted from [18] with permission)

capabilities in the context of DPN with arrayed tips, it has shown great potential to increase the throughput of DPN. Moreover, state-of-the-art DPN allows the individual control of the movement of single pens within a pen array. Therefore, further development of parallel DPN would dramatically increase the complexity and sophistication of the structures one can fabricate with DPN. These capabilities make DPN a unique and highly desirable tool for depositing ink materials on a variety of surfaces, which is important for fundamental studies and direct applications ranging from molecular electronics to materials assembly, and for investigating biological recognition [13].

Generating hard nanostructures. One approach is to write inorganic ink materials that are dispersed in a solvent by DPN. Then the solvent will be removed from the patterned features by post-DPN treatments such as heating or light curing, resulting in well-defined hard inorganic nanostructures [13]. The DPN fabricated hard nanostructures can serve as catalysts, nanoscale conductive circuits, and so on. For example, carbon nanotubes were synthesized on DPN-patterned nanoparticle catalysts [17].

Another approach is the combination of DPN with wet chemical etching. Using such an approach both metal and semiconductor nanostructures were fabricated (Fig. 6). DPN-generated alkanethiol monolayers can be employed as resists for creating solid-state structures by standard wet-etching techniques. By using this method, nanometer-sized complex features can be generated.

Generating templates for hierarchical assembly of materials. DPN-generated chemical templates of various sizes, shapes, and compositions have been used for the directed assembly of a lot of nanoscale building blocks such as carbon nanotubes (Fig. 7) and nanoparticles. In addition, DPN was found useful in bridging the nano-gaps between nanoscale building blocks, which should



Dip-Pen Nanolithography, Fig. 7 (*Upper left*) AFM image and (*lower left*) height profiles of single-walled carbon nanotube (SWNT) rings assembled on a DPN-generated nanoarray. (*Upper right*) A model and (*lower right*) a zoom-in AFM image of a coiled SWNT ring. To achieve the assembly of SWNT, a gold substrate was first patterned with rings of 16-mercaptohexadecanonic acid (MHA). The bare gold regions of the substrate were then passivated with 1-octadecanethiol (ODT). The substrate was treated with a drop of 1,2-dichlorobenzene containing SWNTs. Because 1,2-dichlorobenzene wets the MHA rings but not the ODT passivated regions, SWNTs were selectively assembled on the MHA rings of the substrate (Image was reprinted from [19] with permission)

find broad applications, such as integration of nanocircuits, nanosoldering, and repair of failed nanoelectronics.

Fabricating biological nanoarrays. One can either directly fabricate nanoarrays of biological molecules on surfaces via DPN, or generate biological nanoarrays on a template fabricated with DPN methods. A variety of biological molecules, including DNA, peptides, proteins, viruses, and bacteria, have been patterned using direct-write or indirect adsorption approaches. With biological nanoarrays, a larger number of targets can be screened rapidly in an individual experiment with small amount of sample. For example, DPN-generated nanoarrays of the antibody have been used to screen for the HIV-1 virus antigen in serum samples [20].

Variants of DPN

Researchers constantly extend the basic capabilities of DPN beyond the range of patterning via a coated AFM tips (reviewed by Salaita et al. [13]). For example, researchers have employed heatable AFM tips to achieve melting and afterward patterning of some ink materials that are solids at room temperature and do not have appreciable solubility in a carrier solvent. For another example, by combining DPN with electrochemistry, a surface can also be electrochemically modified and then rapidly coupled to ink material transported from the tip. To facilitate tip inking and ink transporting, several research groups have developed so-called nanofountain or nanopipet as AFM tips to deliver reagents directly to a

substrate. These innovative variations of DPN have significantly expanded its capabilities.

Future Directions

Comparing with other lithographic techniques, DPN has unique properties and advantages. DPN is a positive lithography method. Therefore DPN is complementary to other scanning-probe-based, negative patterning methods [9]. DPN is superior over printing techniques in terms of patterning resolution, while still holds the capability of massive parallelization. The development on the independently actuated pen-array techniques, including software and hardware, is still underway. Further development in this direction would not only provide more flexible control and higher parallelization in DPN process, but also benefit other scanning-probe-based lithography techniques. Then DPN should also find broader applications especially in industry of biotechnology.

In addition, DPN is very attractive for basic research. DPN is an appealing method to deposit a variety of organic molecules on surfaces. The zoo of organic molecules covers a variety of special functionalities, e.g., electrical conductivity, optical properties, and wetting properties. Thus, the behavior of small entities of molecules in respect to these properties can be studied on a nanometer lengths scale with ultrasmall amounts of material.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Micro- and Nanomanipulation for Nanomanufacturing
- ▶ Nanoscale Printing
- ▶ Scanning Probe Microscopy
- ▶ Self-Assembled Monolayers

References

1. Binnig, G., Quate, C.F., Gerber, C.: Atomic force microscope. Phys. Rev. Lett. **56**, 930–933 (1986)
2. Jaschke, M., Butt, H.-J.: Deposition of organic material by the tip of a scanning force microscope. Langmuir **11**, 1061–1064 (1995)
3. Piner, R.D., Zhu, J., Xu, F., Hong, S.H., Mirkin, C.A.: "Dip-pen" nanolithography. Science **283**, 661–663 (1999)
4. Ginger, D.S., Zhang, H., Mirkin, C.A.: The evolution of dip-pen nanolithography. Angew. Chem. Int. Ed. **43**, 30–45 (2004)
5. Smith, J.C., Lee, K.B., Wang, Q., Finn, M.G., Johnson, J.E., Mrksich, M., Mirkin, C.A.: Nanopatterning the chemospecific immobilization of cowpea mosaic virus capsid. Nano Lett. **3**, 883–886 (2003)
6. Kramer, M.A., Jaganathan, H., Ivanisevic, A.: Serial and parallel dip-pen nano lithography using a colloidal probe tip. J. Am. Chem. Soc. **132**, 4532–4533 (2010)
7. Meister, A., Krishnamoorthy, S., Hinderling, C., Pugin, R., Heinzelmann, H.: Local modification of micellar layers using nanoscale dispensing. Microelectron. Eng. **83**, 1509–1512 (2006)
8. Mirkin, C.A.: The power of the pen: development of massively parallel dip-pen nanolithography. ACS Nano **1**, 79–83 (2007)
9. Vettiger, P., Despont, M., Drechsler, U., Durig, U., Haberle, W., Lutwyche, M.I., Rothuizen, H.E., Stutz, R., Widmer, R., Binnig, G.K.: The "Millipede" – more than one thousand tips for future AFM data storage. IBM J. Res. Develop. **44**, 323–340 (2000)
10. Huo, F.W., Zheng, Z.J., Zheng, G.F., Giam, L.R., Zhang, H., Mirkin, C.A.: Polymer pen lithography. Science **321**, 1658–1660 (2008)
11. Shim, W., Braunschweig, A.B.., Liao, X., Chai, J., Lim, J.K., Zheng, G., Mirkin, C.A.: Hard-tip, soft-spring lithography. Nature **469**, 516–520 (2011)
12. Hong, S.H., Zhu, J., Mirkin, C.A.: Multiple ink nanolithography: toward a multiple-pen nano-plotter. Science **286**, 523–525 (1999)
13. Salaita, K., Wang, Y.H., Mirkin, C.A.: Applications of dip-pen nanolithography. Nat. Nanotechnol. **2**, 145–155 (2007)
14. Demers, L.M., Ginger, D.S., Park, S.J., Li, Z., Chung, S.W., Mirkin, C.A.: Direct patterning of modified oligonucleotides on metals and insulators by dip-pen nanolithography. Science **296**, 1836–1838 (2002)
15. Wang, C.: Self-correcting inking strategy for cantilever arrays addressed by an inkjet printer and used for dip-pen nanolithography. Small **2**, 1666–1670 (2008)
16. Rozhok, S., Piner, R., Mirkin, C.A.: Dip-pen nanolithography: what controls ink transport? J. Phys. Chem. B **107**, 751–757 (2003)
17. Kang, S.W., Banerjee, D., Kaul, A.B.., Megerian, K. G.: Nanopatterning of catalyst by dip pen nanolithography (DPN) for synthesis of carbon nanotubes (CNT). Scanning **32**, 42–48 (2010)
18. Zhang, H., Mirkin, C.A.: DPN-generated nanostructures made of gold, silver, and palladium. Chem. Mater. **16**, 1480–1484 (2004)
19. Wang, Y.: Controlling the shape, orientation, and linkage of carbon nanotube features with nano affinity

templates. Proc. Natl. Acad. Sci. U. S. A. **103**, 2026–2031 (2006)
20. Lee, K.B., Kim, E.Y., Mirkin, C.A., Wolinsky, S.M.: The use of nanoarrays for highly sensitive and selective detection of human immunodeficiency virus type 1 in plasma. Nano Lett. **4**, 1869–1872 (2004)

Direct Laser Writing

▶ Laser Beam Machining

Directed Assembly

▶ Electric-Field-Assisted Deterministic Nanowire Assembly

Directed Self-Assembly

▶ Electric Field–Directed Assembly of Bioderivatized Nanoparticles

Disjoining Pressure and Capillary Adhesion

Seong H. Kim
Department of Chemical Engineering, Pennsylvania State University, University Park, PA, USA

Synonyms

Adsorbate adhesion; Liquid condensation; Liquid film; Liquid meniscus; Meniscus adhesion

Definition

Disjoining pressure, introduced by Derjaguin in 1936, may be defined as the difference between the thermodynamic equilibrium state pressure applied to surfaces separated by a thin film and the pressure in the bulk phase with which the film is in equilibrium [1]. Capillary adhesion refers to the adhesion due to a liquid meniscus bridging two solid surfaces which are wetted partially or completely by the liquid film. A curvature of the liquid meniscus induces a pressure difference across the meniscus, which depends on the liquid surface tension and the shape of the meniscus. The pressure difference as well as the surface tension of the liquid result in a net attractive force if the liquid wets the solid surface.

Occurrence

If you have been at a beach, you would have experienced capillary adhesion whether you recognized it or not. Walking on wet sand is easier than walking on dry sand. The wet sand can support your body weight, while the dry sand cannot and your feet sink into the sand. To build a sand castle the sand has to be wet. Wet sand can be shaped because particles adhere to each other. The strong adhesion is caused by liquid menisci, which form around the contact areas of two neighboring particles [2]. The force caused by such a liquid meniscus is called "capillary force" also termed meniscus force. Capillary adhesion phenomena are found and involved in various technical and biological systems. Capillary forces must be taken into account in studies of the adhesion between particles or particles to surfaces which are involved in processing powders, soils, and granular materials [3] and friction and wear of two contacting solid surfaces. They are also important at technological interfaces such as heads on magnetic hard disc memory and microelectromechanical system (MEMS) devices where they cause stiction [4]. Capillary forces play important roles in adhesion of insects, spiders, frogs, and geckos [5].

A liquid meniscus at the solid contact surface is readily formed if the solid surface can be wet with liquid. For example, if the solid surface is covered with a thin liquid film such as nonvolatile lubricants used for magnetic hard discs, the meniscus can be formed by flowing the liquid from the thin

film into the contact region when an opposing surface comes into contact. Even if the solid surface is not covered with preloaded liquid films, the meniscus can be formed around the solid contacting point if the vapor in the surrounding medium is readily adsorbed on the solid surfaces. This phenomenon is called capillary condensation and it is quantified by Kelvin's equation (see below).

Key Research Findings

Disjoining Pressure

The disjoining pressure, Π, is the difference between the pressure of the interlayer on the surfaces confining it and the pressure in the bulk phase, the interlayer being part of this phase and/or in equilibrium with it [1]. Π can be attractive or repulsive. For example, in the case of two plates immersed in a liquid, Π is the pressure that must be applied to the liquid between the plates, in excess of the external pressure to the liquid, to maintain a given separation. The concept can be applied to soap films. In the case of the adsorbed layer at the solid–vapor interface, Π can be thought of as a mechanical pressure that must be applied to a bulk substance to bring it into equilibrium with a given film of the same substance.

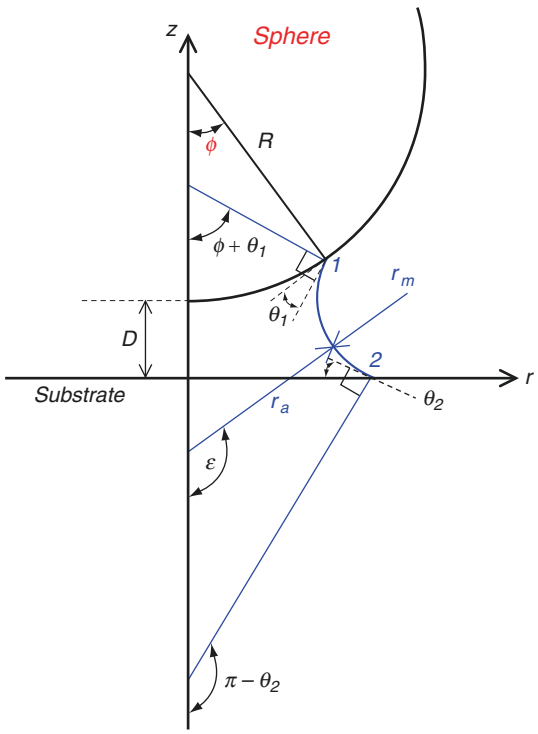
Various methods exist to measure the disjoining pressure of liquid films. These include thin film balance [6], contact angle studies [7], and atomic force microscopy (AFM) [8]. The disjoining pressure can be regarded as a net of several components including dispersion interactions, overlapping of diffuse double layers in the case of charged surfaces, overlapping of adsorbed layers of neutral molecules, and alteration of solvent structure in the discrete region (such as salvation or hydration) [1]. The dispersion (van der Waals) interaction contribution can be estimated using the Derjaguin approximation, $\approx A/6\pi h^3$, where A is the Hamaker constant (typically ~10^{-20} J) and h is the film thickness. Any deviation from this approximation must be due to specific chemical interactions such as ionic and covalent bonding of liquid molecules or structuring of liquid layers in the vicinity of solid surfaces.

Young–Laplace Equation

In the continuum model, the governing equation for the surface curvature of a liquid meniscus bridging two solid surfaces is the Young–Laplace equation [1, 3]. A curved liquid–vapor interface supports a pressure difference called Laplace pressure:

$$\Delta P_L = \gamma\left(\frac{1}{r_m} + \frac{1}{r_a}\right) = \frac{\gamma}{r_e} \qquad (1)$$

Here, γ is the surface tension of the liquid, r_m and r_a are the meridional and azimuthal radii of curvature of the surface, respectively, as shown in Fig. 1, and re is the effective radius of curvature. The curvature defined by r_m is in the plane of the paper and that defined by r_a is normal to the plane. Since the radius center of r_m is outside the meniscus, it has a negative value. The liquid surface is isobaric, meaning that r_e is constant along the meniscus surface, while r_m and r_a vary along



Disjoining Pressure and Capillary Adhesion, Fig. 1 Geometrical presentation of a liquid bridge between a sphere and a flat substrate [13]

the surface. Figure 1 also shows the contact angles of liquid with the sphere and the substrate surfaces (θ_1 and θ_2), which arise from the solid–liquid equilibrium (Young equation). The exact solution of the differential Young–Laplace equation for this geometry can be found in references [3] and [9].

Liquid Meniscus Formation Due to Vapor Adsorption

Thermodynamic equilibrium requires the chemical potential of the liquid meniscus to be equal to that of the vapor. Assuming the molar volume (V_m) of the liquid substance to be constant and applying the Young–Laplace equation, (1), the Gibbs free energy of the liquid meniscus at a constant temperature is expressed as $\Delta G_{liq} = \int V_m \mathrm{dP} = V_m \Delta P_L = V_m \gamma / r_e$. The Gibbs free energy of a vapor with respect to its saturation state is $\Delta G_{vap} = R_g \mathrm{T} \cdot \ln(p/p_{sat})$. Setting ΔG_{liq} equal to ΔG_{vap} and replacing r_e with r_K gives the Kelvin equation [1, 3]:

$$V_m \Delta P_L = V_m \frac{\gamma}{r_k} = R_g T \mathrm{1n}(p/p_{sat}) \qquad (2)$$

$$r_k = \frac{\gamma V_m}{R_g T} = \frac{1}{\mathrm{1n}(p/p_{sat})} = \frac{\lambda_k}{\mathrm{1n}(p/p_{sat})} \qquad (3)$$

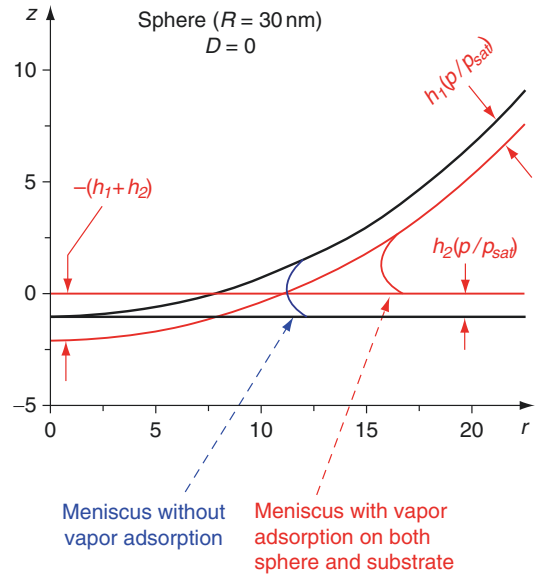
Consider that the solid surface is covered with the adsorbate film which is in equilibrium with the vapor phase (Fig. 2). Then, the isothermal Gibbs energy relationship between the adsorbate and the vapor is as follows [10]:

$$V_m \Pi(h) = R_g T \mathrm{1n}(p/p_{sat}) \qquad (4)$$

where $\Pi(h)$ is the disjoining pressure of the adsorbed layer with a thickness h. The $V_m \Pi(h)$ term represents the molecular interactions of the adsorbate molecules with the solid surface [11]. From (2) and (4), it is obvious that the meniscus should be in equilibrium with the adsorbate film.

Capillary Force Calculation

The capillary force from the liquid meniscus depends on two components – the axial



Disjoining Pressure and Capillary Adhesion, Fig. 2 Meniscus profiles for a sphere with a radius of 30 nm without (*blue line*) and with (*red line*) considering the equilibrium thickness of the adsorbate layer on the sphere, $h_1(p/p_{sat})$, and the thickness on the flat substrate, $h_2(p/p_{sat})$. The meniscus contact angle is set to $\theta_1 = \theta_2 = 0°$ [13]

components (along the z-axis) of the surface tension force F_{st} and the Laplace pressure force F_{pl} at the point where the meniscus and the sphere meet (at the upper boundary point 1). For the geometry shown in 1, the two components can be calculated with the following equations:

$$F_{st} = 2\pi R \sin(\phi) \times \gamma \sin(\theta_1 + \phi) \qquad (5)$$

$$\begin{aligned} F_{pl} &= -\pi R^2 \sin^2(\phi) \times \frac{\gamma}{r_k} \\ &= \pi R^2 \sin^2(\phi) \times \frac{R_g T}{V_m} \mathrm{1n}(p/p_{sat}) \end{aligned} \qquad (6)$$

Here the positive sign is chosen for the attractive force. The term ϕ is often called a filling angle. The total capillary force is then the sum of these two components:

$$\begin{aligned} F_{cap} &= F_{st} + F_{pl} = 2\pi\gamma R \\ &\times \left[\sin(\phi) \sin(\theta_1 + \phi) - R \sin^2(\phi)/(2r_k) \right] \end{aligned} \qquad (7)$$

Since the meniscus is in equilibrium, the force at the boundary point 1 is the same as the force at any point on the meniscus.

Simple Approximation

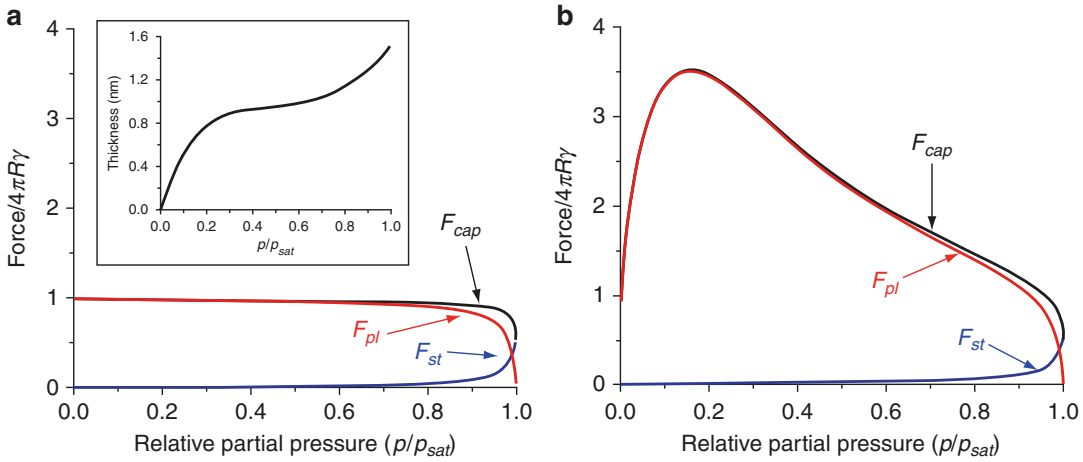
One of the widely used equations to describe the capillary force is $F_{cap} = 4\pi R\gamma \cos\theta$ [10]. This equation is derived from the Young–Laplace equation which contains two unknowns (principal radii of the meniscus) through several assumptions. One assumption is that the meniscus shape can be approximated to be circular. It also assumes that the radius of the tip is much larger than the cross-sectional radius of the liquid meniscus which, in turn, is much larger than the meridional curvature (external curvature) of the meniscus surface. This approximate form describes many macroscopic capillary phenomena; however, it does not work well as the contacting solid system gets smaller.

More Accurate Calculations

Figure 3 shows the capillary force (F_{cap}) and its two components (F_{pl} and F_{st}) for R = 100 nm calculated using the exact solution of the Young–Laplace equation, (1) [12, 13]. The force scale is normalized with $4\pi R\gamma$. The calculation shows a drastic difference in the p/p_{sat} dependence of F_{cap} due to the vapor adsorption. When the solid surface is assumed to be free of adsorbates (Fig. 3a), the capillary force decreases monotonically but insignificantly with p/p_{sat} until p/p_{sat} approaches the saturation point [14]. In contrast, the calculation including a type-II adsorption isotherm (inset in Fig. 2a) shows a large partial pressure dependence (Fig. 3b) [12, 13]. *Fcap* reaches a maximum value at p/p_{sat} ~0.15 and then decreases upon further increase of p/p_{sat}. Note that the maximum p/p_{sat} range coincides with the knee point of the type-II adsorption isotherm which corresponds to the completion of the monolayer thickness. The result calculated with the adsorption isotherm is in good agreement with experimental measurement data [15]. The capillary force exerted by the meniscus formed by vapor adsorption and condensation is strongly affected by the structure of the adsorbate layer [16] as well as the geometry of the contacting solid surfaces [17].

Meniscus Stretch and Liquid Film on Solid

When the vapor adsorption from the environment is negligible (e.g., water does not adsorb readily on hydrophobic surfaces), the Laplace pressure of the meniscus ($P(h)$) under equilibrium conditions is equal to the disjoining pressure ($\Pi(h)$) of the



Disjoining Pressure and Capillary Adhesion, Fig. 3 Total (*black*), Laplace pressure component (*red*), and surface tension component (*blue*) of the capillary force for a sphere with R = 100 nm on a flat surface (**a**) without and (**b**) with taking into account the adsorption isotherm of ethanol. The forces are normalized with $4\pi R\gamma$, where γ = surface tension of liquid [13]. The inset in (**a**) shows the model adsorption isotherm used in this calculation. The isotherm curve is produced by fitting the experimental data reported in reference [15]

liquid film deposited on the solid surface. Mate et al. demonstrated the use of AFM in calculating $r_{eff}(h)$ for a liquid lubricant film [18]. By slowly retracting an AFM probe in contact with a liquid film, the meniscus is stretched in a quasi-equilibrium state (the stretch rate is slow enough to allow for the change in the meridional and azimuthal radii in a manner that $r_{eff}(h)$ remains constant throughout the meniscus stretching). Then, $r_e(h)$ is obtained by fitting the force-distance curve data to the following equation:

$$\frac{F(d)}{4\pi R\gamma} = \left(1 + \frac{d}{2r_e(h)}\right) \tag{8}$$

where $F(d)$ is the force measured by AFM, R is the radius of the AFM probe tip, and d is the tip-sample separation distance. The $r_e(h)$ is extracted from the slope of the linear portion of the retraction curve where the meniscus is stretched from the AFM tip.

Bowles and White later proposed a numerical solution to calculate the shape of the retraction curve based on the shape of the meniscus with a given $r_e(h)$ [19]. The White method analyzes this retraction curve based on a given $r_e(h)$ and the cantilever spring constant k. The White method fits the experimental data with the following equations:

$$\frac{F(d)}{4\pi R\gamma} = \frac{\lambda}{4\alpha^2} \tag{9}$$

where $\lambda = \frac{r_e(h)}{R}$ and $\alpha = \frac{r_e(h)}{r_0(d)}$ are two parameters for the numerical solution of the shape of the meniscus on the AFM probe and $r_0(d)$ is a moving boundary where the lower edge of the meniscus meets the unperturbed liquid film of thickness h. Once $r_e(h)$ is determined, the disjoining pressure of the liquid film and the Laplace pressure of the liquid meniscus can be calculated using the surface tension of the liquid film.

Cross-References

▶ Nanotribology

References

1. Adamson, A.W.: Physical Chemistry of Surfaces, 5th edn. Wiley, New York (1990)
2. Bocquet, L., Charlaix, E., Ciliberto, S., Crassous, J.: Moisture-induced ageing in granular media and the kinetics of capillary condensation. Nature **396**, 735–737 (1998)
3. Kralchevsky, P.A., Nagayama, K.: Particles at Fluid Interfaces and Membranes. Elsevier, Amsterdam (2001)
4. Kim, S.H., Asay, D.B., Dugger, M.T.: Nanotribology and MEMS. NanoToday **2**, 22–29 (2007)
5. Autumn, K., Peattie, A.M.: Mechanisms of adhesion in Geckos. Integr. Comp. Biol. **42**, 1081–1090 (2002)
6. Claesson, P.M., Ederth, T., Bergeron, V., Rutland, M. W.: Techniques for measuring surface forces. Adv. Colloid Interface Sci. **67**, 119–183 (1996)
7. Leger, L., Joanny, J.F.: Liquid spreading. Rep. Prog. Phys. **55**, 431–486 (1992)
8. Saramago, B.: Thin liquid wetting films. Curr. Opin. Colloid Interface Sci. **15**, 330–340 (2010)
9. Orr, F.M., Scriven, L.E., Rivas, A.P.: Pendular rings between solids: meniscus properties and capillary force. J. Fluid Mech. **67**, 723–742 (1975)
10. Israelachvili, J.N.: Intermolecular and Surface Forces, 2nd edn. Academic, San Diego (1992)
11. Mate, C.M.: Application of disjoining and capillary pressure to liquid lubricant films in magnetic recording. J. Appl. Phys. **72**, 3084–3090 (1992)
12. Asay, D.B., de Boer, M.P., Kim, S.H.: Equilibrium vapor adsorption and capillary force: exact Laplace-Young equation solution and circular approximation approaches. J. Adheshion Sci. Technol. **24**, 2363–2382 (2010)
13. Hsiao, E., Marino, M.J., Kim, S.H.: Effects of gas adsorption isotherm and liquid contact angle on capillary force for sphere-on-flat and cone-on-flat geometries. J. Colloid Interface Sci. **352**, 549–557 (2010)
14. Butt, H.J., Kappl, M.: Normal capillary forces. Adv. Colloid Interface Sci. **146**, 48–60 (2009)
15. Asay, D.B., Kim, S.H.: Molar volume and adsorption isotherm dependence of capillary forces in nanoasperity contacts. Langmuir **23**, 12174–12178 (2007)
16. Asay, D.B., Kim, S.H.: Effects of adsorbed water layer structure on adhesion force of silicon oxide nanoasperity contact in humid ambient. J. Chem. Phys. **124**, 174712 (2006)
17. Köber, M., Sahagún, E., García-Mochales, P., Briones, F., Luna, M., José Sáenz, J.: Nanogeometry matters: unexpected decrease of capillary adhesion forces with increasing relative humidity. Small **6**, 2725–2730 (2010)
18. Mate, C.M., Lorenz, M.R., Novotny, V.J.: Atomic force microscopy of polymeric liquid films. J. Chem. Phys. **90**, 7550–7555 (1989)
19. Bowles, A.P., Hsia, Y.T., Jones, P.M., Schneider, J.W., White, L.R.: Quasi- equilibrium AFM measurement of disjoining pressure in lubricant nano-films I: Fomblin Z03 on silica. Langmuir **22**, 11436–11446 (2006)

Dislocation Dynamics

► Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Disordered Plasmonics

► Harnessing Disorder at the Nanoscale

Dispersion

► Fate of Manufactured Nanoparticles in Aqueous Environment

Dissipative Cavity

► Harnessing Disorder at the Nanoscale

Dissipative Particle Dynamics, Overview

Xuejin Li, Zhen Li, Xin Bian, Mingge Deng, Changho Kim, Yu-Hang Tang, Alireza Yazdani and George Em Karniadakis
Division of Applied Mathematics, Brown University, Providence, RI, USA

Synonyms

Coarse-grained molecular dynamics; Fluctuating hydrodynamics; Mesoscopic simulation; Rheology; Soft matter

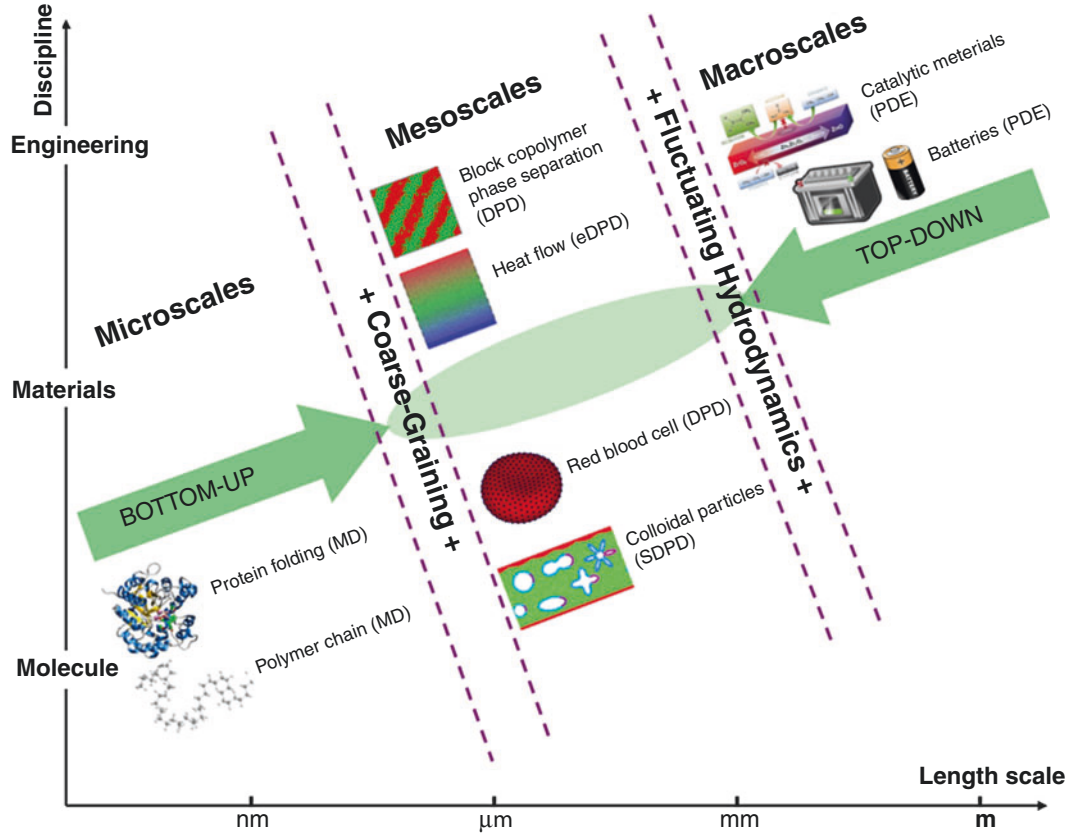
Definition

Dissipative particle dynamics (DPD) is a stochastic mesoscopic simulation technique that describes *clusters* of molecules moving together in a Lagrangian fashion subject to simplified pairwise conservative, dissipative and random forces.

Introduction

Natural systems can be described at different scales based on both spatial and temporal size. In general, there are three different scales, i.e., micro-, meso-, and macroscales. A microscopic event occurs at nanometers in length and nanoseconds in time or, even less, governed by quantum mechanics or classical laws. Macroscale describes physical objects or phenomena that are measurable and visible directly with the naked eye, and thus, the mean free path of molecules is far smaller than the characteristic length of the geometry. A macroscopic event is usually described by continuum partial differential equations (PDEs), such as Navier-Stokes equations for fluid dynamics. Mesoscale is somewhat in between, which is typically in order of micrometers and microseconds. Examples of such mesoscopic systems include colloidal suspension, polymer solution, and biological membranes. Since the behaviors of such systems can be much more complex than that of a simple fluid, they are usually called complex fluids. There exist broad applications of complex fluid systems in industrial technologies, especially in the field of biological and biomedical sciences such as biochemical lab-on-a-chip systems, separation devices, and drug delivery systems.

Computer simulations have become an essential part of mathematical modeling of many natural systems in modern science and technology, serving as a bridge between theories and experiments (see Fig. 1). Molecular dynamics (MD) has emerged as one of the first and yet widely used simulation methods in many areas of physics and chemistry starting from the 1950s. However, the classical MD is practical only for simulations up to microscales. It is difficult to simulate larger and more complex systems because of prohibitively expensive computations. Meanwhile, sophisticated computational and numerical discretization techniques have been developed over the past decades for the solution of PDEs at macroscale,



Dissipative Particle Dynamics, Overview, Fig. 1 Illustration of the length scales in "soft" materials. MD, DPD, and PDE stand for molecular dynamics, dissipative particle dynamics, and partial differential equations, respectively, and eDPD and SDPD indicate energy-conserving DPD and smoothed DPD, respectively

such as various computational fluid dynamics (CFD) methods. However, these PDEs are inadequate to capture mesoscopic events in complex fluids, since these PDEs are based on simple constitutive equations, while the dynamics of a complex fluid is fundamentally affected by its microscopic structure. Furthermore, the standard PDE formulations often ignore the stochastic nature of a system observed at mesoscale, which is a characteristic feature for many physical phenomena.

Mesoscopic simulation methods have been developed to overcome the aforementioned problems, aiming at modeling complex fluids with efficient computational costs. Dissipative particle dynamics (DPD), which describes *clusters* of molecules moving together in a Lagrangian fashion, is a typical mesoscopic simulation method for the dynamic and rheological properties of simple and complex fluids [1]. DPD combines Lagrangian features from MD and coarser spatial-temporal scales from lattice-gas automata (LGA), and, therefore, it is faster than MD and more flexible than LGA. The first form of DPD was reformulated by Espanöl and Warren such that it produces a correct thermal equilibrium state [2]. This is considered as the standard form of DPD. Several improved DPD models [3–6], which are capable of representing complicated fluid properties more accurately, as well as efficient algorithms have also been developed. Rigorous foundations of DPD methodology have also been investigated by both top-down (from macroscopic description to mesoscopic description) and bottom-up (from microscopic to mesoscopic) approaches.

Theoretical Formulation

In a DPD simulation, a particle represents a cluster of molecules, and the position and momentum of the particle are updated in a continuous phase space at discrete time steps. The equation of motion and pairwise interacting forces of particles read

$$\dot{\mathbf{r}} = \mathbf{v}_i,\ m_i\dot{\mathbf{v}}_i = \sum_{j\neq i}\left(\mathbf{F}_{ij}^C + \mathbf{F}_{ij}^D + \mathbf{F}_{ij}^R\right), \quad (1)$$

where $\mathbf{r}_i$, $\mathbf{v}_i$, and m_i are position, velocity, and mass of particle i, respectively. $\mathbf{F}_{ij}^C$ is referred to as conservative force obtained from a prescribed potential between particles i and j. It is repulsive and leads particles to be evenly distributed in space. Roughly speaking, it can be understood as a pressure force; $\mathbf{F}_{ij}^D$ has a negative sign and is proportional to the velocity difference of two particles. Therefore, it is dissipative and resists velocity difference of any interacting pair of particles. Hence, $\mathbf{F}_{ij}^D$ dissipates the kinetic energy of the system. This amount of energy must be put back into the system, to keep the system at a constant temperature. $\mathbf{F}_{ij}^R$ is exactly for this purpose. It is a random force and injects kinetic energy back into the system. Effectively, $\mathbf{F}_{ij}^D$ and $\mathbf{F}_{ij}^R$ implement a thermostat so that thermal equilibrium is achieved. The magnitudes and functional forms of the two forces are related by the so-called *fluctuation-dissipation theorem*. The typical forms of the three forces are as follows [2, 7]:

$$\mathbf{F}_{ij}^C = a_{ij}\omega^C\left(r_{ij}\right)\mathbf{e}_{ij}, \quad (2a)$$

$$\mathbf{F}_{ij}^D = -\gamma\omega^D\left(r_{ij}\right)\left(\mathbf{e}_{ij}\cdot\mathbf{v}_{ij}\right)\mathbf{e}_{ij}, \quad (2b)$$

$$\mathbf{F}_{ij}^R = \sigma\omega^R\left(r_{ij}\right)\theta_{ij}\delta t^{-1/2}\mathbf{e}_{ij}, \quad (2c)$$

where a_{ij}, γ, and σ reflect, respectively, the strength of conservative, dissipative, and random forces. Here, θ is a *Gaussian white noise* ($\theta_{ij} = \theta_{ji}$), and $\mathbf{e}_{ij}$ is the unit vector pointing from particle j to i; also, ω^C, ω^D, and ω^R are unspecific weighting functions of relative distance r_{ij}. A common choice of the weighting functions is

$$\omega^C\left(r_{ij}\right) = \begin{cases}\left(1 - r_{ij}/r_c\right)^s, & r_{ij} < r_c,\\ 0, & r_{ij} \geq r_c,\end{cases} \quad (3)$$

$$\omega^R\left(r_{ij}\right) = \begin{cases}1 - r_{ij}/r_c, & r_{ij} < r_c,\\ 0, & r_{ij} \geq r_c,\end{cases} \quad (4)$$

where $s = 1$ is the most widely adopted for the classical DPD method. However, other choices (e.g., $s = 0.25$) for the envelopes have also been used. Also, r_c is the cutoff radius, which defines the extent of the interaction range.

To satisfy the fluctuation-dissipation theorem, two conditions must be further enforced [2]:

$$\omega^D\left(r_{ij}\right) = \left[\omega^R\left(r_{ij}\right)\right]^2, \quad \sigma^2 = 2\gamma k_{\mathrm{B}}T, \quad (5)$$

where k_{B} is the Boltzmann's constant and T is the absolute temperature.

Two important implications of the DPD forces in Eq. 2 must be explicitly noted: firstly, DPD is considered as a reduction model of the underlying microscopic dynamics. By construction, it focuses on the coarse-grained properties and intentionally ignores irrelevant degrees of freedom on the microscopic level. As a result, interparticle potential (the derivative of which is $-\mathbf{F}_{ij}^C$) is much softer than that of MD method; hence, it can potentially access longer time and length scales than are possible using conventional MD simulations. Furthermore, $\mathbf{F}_{ij}^D$ and $\mathbf{F}_{ij}^R$ together account for the lost microscopic information. Secondly, by design, all three forces act along the line of particle centers $\mathbf{e}_{ij}$ and are symmetric by interchanging particle indices. Therefore, the momentum is locally conserved. The fulfillment of the conservation laws guarantees that a DPD system approaches the PDEs of fluid dynamics, such as Navier-Stokes equations, at the macroscopic scale.

For the purpose of modeling simple and complex fluids, α, γ (or σ), and s are free parameters to calibrate so that the desired properties of a target system can be achieved. The time evolution of velocities and positions of particles is determined by Newton's second law of motion similar to the MD method, which is usually integrated using a modified velocity-Verlet algorithm [7].

Extensions of DPD

The classic DPD method is a minimal working version for mesoscopic simulations of fluids. Due to the lack of necessary degrees of freedom, it cannot model some specific problems, e.g., heat flow in non-isothermal systems and liquid-vapor interface in two-phase fluid systems. To this end, extra degrees of freedom, such as an internal energy variable and freedom to alter the equation of state, have been introduced to DPD system for particular applications [8]. In this section, several extensions of the DPD method will be described briefly.

Energy-Conserving DPD

The classic DPD method is limited to isothermal systems and is not valid for modeling non-isothermal processes because the energy is not conserved in particle interactions. To extend the standard formulation of DPD with application to heat flow in non-isothermal fluid systems, the conservation of energy was applied to DPD by introducing the particle internal energy ε_i as an additional variable [3]. Along this new variable, a local entropy function $s(\varepsilon_i)$ and the temperature $T_i = (\partial s_i/\partial \varepsilon_i)^{-1}$ are defined for each particle. The energy-conserving DPD model is known in the literature as eDPD or DPDE, where for interacting particles, the dissipation induced by the dissipative forces is utilized in raising the internal energy of the particles through viscous heating. Moreover, the temperature differences between particles produce variations in the internal energies through heat conduction.

The key aspect of eDPD is that the temperature T_i is defined on each particle, and the fluctuation-dissipation theorem is applied locally based on the particle temperature T_i rather than the thermodynamic temperature of the system. As a result, eDPD allows temperature gradients and can be used in non-isothermal problems, where thermal transport plays a critical role.

Many-Body DPD

The conservative force $\mathbf{F}_{ij}^{C}$ is responsible for the thermodynamic behavior of the DPD system and has the form shown in Eq. 2. For a single-component DPD fluid, the equation of state (EOS) of the system can be approximated as $p = \rho k_B T + \alpha a_{ij} \rho^2$, which is a quadratic function of the particle density ρ. This monotonic function of EOS does not contain a van der Waals loop and excludes the possibility of simulating phenomena involving vapor-liquid coexistence or free surface.

To model vapor-liquid coexistence, the many-body DPD (MDPD) was developed by modifying the expression of the conservative force [5]. A direct way to produce higher order terms in EOS is making the conservative force in MDPD dependent not only on the interparticle distance but also on the instantaneous local particle density. Therefore, the conservative force in MDPD is effectively a many-body force and it generally includes both attractive and repulsive components. With higher order terms in EOS, MDPD has been successfully applied to simulations of droplet- and wetting-related phenomena.

Fluid Particle Model

The fluid particle model (FPM) is a generalization of the classic DPD method [4]. FPM considers both linear and angular momenta of the particles and includes both central and noncentral forces between particles, while the classic DPD method considers only linear momentum and includes only central forces. By introducing torques and angular velocities of the particles, both linear and angular momenta of FPM are conserved. Also, the dissipative and random forces in FPM satisfy the fluctuation-dissipation theorem, and they together act as a thermostat to maintain the temperature of the FPM system at a constant.

Since the rotational motion of individual particle is included, FPM shows high computational efficiency in studying the dynamical and rheological properties of colloidal suspensions in simple fluid solvents. The reason is that each colloidal particle can be represented by a single FPM particle rather than a cluster of particles. FPM has been used to investigate a range of rheological behaviors and volume fractions of the colloidal suspensions and shows good agreement with experimental and theoretical results.

Smoothed DPD

The classic DPD model is a bottom-up approach and it is usually considered as a coarse-grained MD model, in which particles are governed by a coarse-grained force field. Similarly to the MD method, the inputs of DPD are the particle interactions, while the macroscale quantities, such as viscosity, diffusivity, compressibility, and EOS, are output properties. As a result, DPD has no restrictions on constitutive equations, which provides flexibility in modeling complex materials and multicomponent systems. However, a shortcoming of DPD is that it employs empirical expressions for the particle interactions, and its parameters are not related in a straightforward way to usual macroscale quantities. Therefore, in practice, the DPD units have to be mapped to physical units based on output properties.

Alternatively, *smoothed* DPD (SDPD) is a top-down approach [6]. It was developed from smoothed particle hydrodynamics (SPH). In fact, SDPD is a Lagrangian discrete model for simulating Navier-Stokes hydrodynamics that includes thermal fluctuations consistently. Unlike the classic DPD method, the parameters in the governing equations of SDPD have clear physical meaning. The inputs are the viscosity, equation of state, temperature, and other parameters required by the fluctuating Navier-Stokes equations. However, since the constitutive equations and macroscale properties of the system must be given as inputs, SDPD works well with simple fluids but loses its effectiveness in modeling complex materials.

Applications in Different Fields

In the past decade, the DPD method has been successfully extended to many applications including simple fluid hydrodynamics, polymer solutions and melts, biological membranes, colloidal suspensions, and blood flow [9, 10]. A few examples are presented in the following:

Polymer Solution

A polymer is a large molecule composed of many repeated subunits bonded together. In DPD, a polymer can be represented by linking collections of particles into chains with appropriate spring forces expressed as $\mathbf{F}_{ij}^{S} = k_{\text{bond}}\left(1 - r_{ij}/r_0\right)\mathbf{e}_{ij}$ [9]. Numerous simulations have testified that the DPD model is capable of capturing many essential physical phenomena of the polymer systems. For example, the DPD model has been applied to polymers in solution to study effects of solvent quality on the conformation and relaxation of polymer chains [11]. The dynamics of polymer chains in simple shear flow such as Couette and Poiseuille flow have also been simulated. Simple shear flows are commonly generated by imposing a constant driving body force (Poiseuille flow), or a driving velocity on the boundary shear planes (Couette flow). The static and dynamic properties of polymer chain under shear, such as radius of gyration, diffusivity, and viscosity, were computed and compared to experimental results [12]. The conformational changes and translocation dynamics of polymer and DNA molecules through microfluidic channels have also been investigated by DPD [13].

Amphiphilic Systems

Amphiphilic molecules, which contain hydrophilic head and hydrophobic tails, can be synthesized with various structures of the block type. In a selective solvent of one of the blocks, they can self-assemble into a wide variety of morphologies. A typical model of amphiphilic molecules is the two-letter ("black-and-white") H-T model, composed of a single head particle (denoted by H) and a linear chain of tail particles (denoted by T). The amphiphilic molecules are then immersed in solvent bath (denoted by W).

To model the amphiphilic nature of amphiphilic molecules, the hydrophobic and hydrophilic interactions emerge from the relative interaction strengths. In theory, the functional dependence of the Flory-Huggins parameter χ [14] on DPD repulsion parameter a_{ij} places the simulation results in the proper context. The relationship between a_{ij} and χ has been established by Groot and Warren [15]: $a_{ij} = a_{ii} + 3.27\chi_{ij}$ at particle density $\rho = 3$ and $a_{ij} = a_{ii} + 1.45\chi_{ij}$ at $\rho = 5$. Typically, for simplicity, the repulsive parameter related to the interaction between two alike DPD particles (H-H, T-T, W-W, and H-W) is

treated the same, while the parameter related to the interaction between the hydrophilic and hydrophobic particles (H-T and T-W) is made larger than the repulsion parameter between two similar particles, which ensures that the hydrophobic tails of amphiphilic molecules are sufficiently shielded from the hydrophilic head and solvent. Various morphologies, such as spherical and cylindrical micelles, toroids, bilayers, and vesicles, have been observed depending on the selectivity of solvents, amphiphilic concentration, and the structure of the amphiphilic molecules [16]. A selection of vesicle shapes obtained from the DPD simulations is displayed in Fig. 2.

By adding a bond-bending potential to the hydrophobic tails, a rod-coil amphiphilic molecule can be modeled. Based on the functionality of the rigid-rod blocks, there is growing interest in amphiphilic systems with rod-coil structures, and a number of intriguing phases, including wavy lamellae, zigzags, and hollow spherical micelles, have been observed through self-assembly from DPD simulations [17]. These simulations have shown the applicability of DPD method to study the self-assembly process of amphiphilic systems.

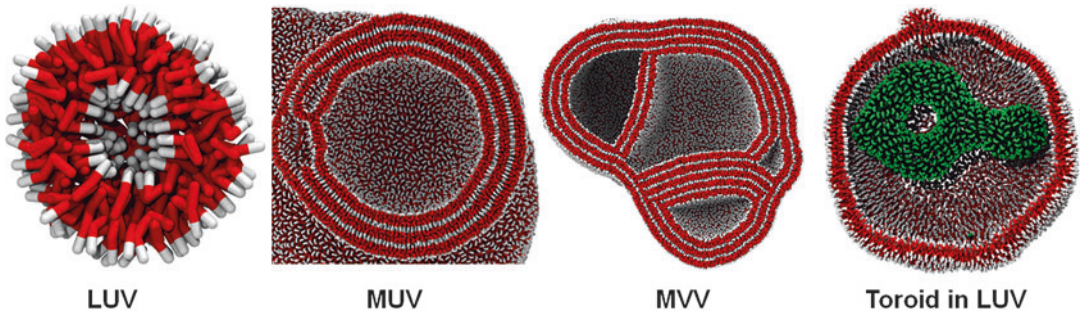
Blood Flow

Blood is primarily composed of microscopic cellular particles such as red blood cells (RBCs), leukocytes, and platelets. The most abundant cells in vertebrate blood are RBCs. A human RBC is a nucleus-free cell with a biconcave shape when not subject to any external stress and is approximately 8.0 μm in diameter and 2.0 μm in thickness. The RBC membrane consists of a lipid bilayer supported by an attached spectrin-based cytoskeleton, and its extreme deformability allows the RBC to squeeze without any damage when passing through narrow capillaries in microcirculation. A multiscale RBC (MS-RBC) model [18], which is based on spectrin-level RBC model [19], has been developed for blood simulations.

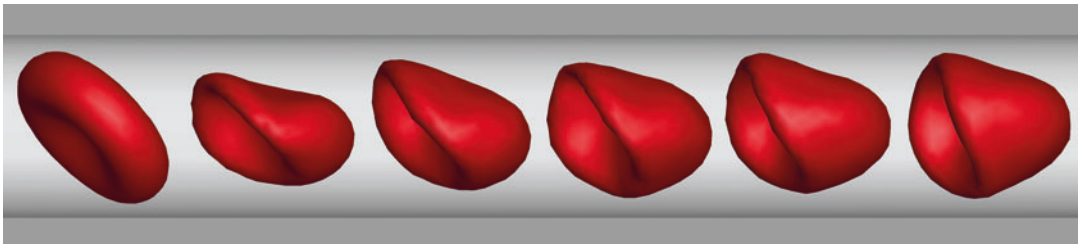
In the MS-RBC model, the membrane of RBC is represented by a two-dimensional triangulated network with a collection of DPD particles, while constraints on the area and volume conservation of RBC are imposed to mimic the area-preserving lipid bilayer and the incompressible interior fluid. The RBC membrane interacts with the fluid particles through DPD forces, and the temperature of the system is controlled through the DPD thermostat. The internal and external fluids are modeled by collections of free DPD particles, and their separation is enforced by bounce-back reflections of these particles at the RBC membrane surface. The MS-RBC has been successfully applied in RBC simulations, such as RBC dynamics in Poiseuille flow (see Fig. 3), RBC thermal fluctuations, and RBCs in diseases like malaria and sickle cell anemia [20].

Outlook

As a coarse-grained molecular dynamics method, dissipative particle dynamics has been



Dissipative Particle Dynamics, Overview, Fig. 2 Self-assembled complex microstructures of amphiphilic molecules. These microstructures included: large unilamellar vesicle (*LUV*), multilamellar vesicle (*MLV*), multivesicular vesicle (*MVV*), and a toroid in LUV. The hydrophilic and hydrophobic particles are rendered in *white* and *red/green*, respectively. Slices of these shapes are shown for clarity



Dissipative Particle Dynamics, Overview, Fig. 3 Sequence of RBC shapes from biconcave to parachute in Poiseuille flow in tube

demonstrated to be a promising mesoscopic method for simulating the dynamic and rheological properties of simple and complex fluids. The classical DPD method and its extended models have already been applied successfully to a wide variety of phenomena occurring at the mesoscales. However, there remain many open questions both with regard to the foundations of the method and in various applications. For example, the issue of memory effects and the scaling up of DPD in even bigger spatiotemporal scales are two important open questions.

Cross-References

- ▶ Computer Modeling and Simulation of Materials
- ▶ Mesoscopic Modeling
- ▶ Multiscale Modeling

References

1. Hoogerbrugge, P.J., Koelman, J.M.V.A.: Simulating microscopic hydrodynamic phenomena with dissipative particle dynamics. Europhys. Lett. **19**, 155–160 (1992)
2. Español, P., Warren, P.: Statistical mechanics of dissipative particle dynamics. Europhys. Lett. **30**, 191–196 (1995)
3. Español, P.: Dissipative particle dynamics with energy conservation. Europhys. Lett. **40**, 631–636 (1997)
4. Español, P.: Fluid particle model. Phys. Rev. E **57**, 2930–2948 (1998)
5. Warren, P.B.: Vapor-liquid coexistence in many-body dissipative particle dynamics. Phys. Rev. E **68**, 066702 (2003)
6. Español, P., Revenga, M.: Smoothed dissipative particle dynamics. Phys. Rev. E **67**, 026705 (2003)
7. Groot, R.D., Warren, P.B.: Dissipative particle dynamics: Bridging the gap between atomistic and mesoscopic simulation. J. Chem. Phys. **107**, 4423–4435 (1997)
8. Pivkin, I.V., Caswell, B., Karniadakis, G.E.: Dissipative particle dynamics. In: Lipkowitz, K.B. (ed.) Reviews in Computational Chemistry, pp. 85–110. Wiley, New Jersey (2010)
9. Groot, R.D.: Applications of dissipative particle dynamics. Lect. Notes Phys. **640**, 5–38 (2004)
10. Lu, Z.-Y., Wang, Y.-M.: An introduction to dissipative particle dynamics. In: Monticelli, L., Salonen, E. (eds.) Biomolecular Simulations: Methods and Protocols, Methods in Molecular Biology, pp. 617–633. Springer, New York (2013)
11. Kong, Y., Manke, C.W., Madden, W.G., Schlijper, A. G.: Effect of solvent quality on the conformation and relaxation of polymers via dissipative particle dynamics. J. Chem. Phys. **107**, 592–602 (1997)
12. Fan, X.J., Phan-Thien, N., Ng, T.Y., Wu, X.H., Xu, D.: Microchannel flow of a macromolecular suspension. Phys. Fluids **15**, 11–21 (2003)
13. Fan, X.J., Phan-Thien, N., Chen, S., Wu, X.H., Ng, T.Y.: Simulating flow of DNA suspension using dissipative particle dynamics. Phys. Fluids **18**, 063102 (2006)
14. Doi, M., Edwards, S.F.: The Theory of Polymer Dynamics. Oxford University Press, London (1986)
15. Groot, R.D., Madden, T.J.: Dynamic simulation of diblock copolymer microphase separation. J. Chem. Phys. **108**, 8713–8724 (1998)
16. Yamamoto, S., Maruyama, Y., Hyodo, S.: Dissipative particle dynamics study of spontaneous vesicle formation of amphiphilic molecules. J. Chem. Phys. **116**, 5842–5849 (2002)
17. Huang, J.-H., Ma, Z.-X., Luo, M.-B.: Self-assembly of rod-coil diblock copolymers within a rod-selective slit: A dissipative particle dynamics simulation study. Langmuir **30**, 6267–6273 (2014)
18. Fedosov, D.A., Caswell, B., Karniadakis, G.E.: Dissipative particle dynamics modeling of red blood cells. In: Pozrikidis, C. (ed.) Computational Hydrodynamics of Capsules and Biological Cells, pp. 183–218. CRC Press, Boca Raton (2010)

19. Li, J., Dao, M., Lim, C.T., Suresh, S.: Spectrin-level modeling of the cytoskeleton and optical tweezers stretching of the erythrocyte. Biophys. J. **88**, 3707–3719 (2005)
20. Fedosov, D.A., Pivkin, I.V., Pan, W.X., Dao, M., Caswell, B., Karniadakis, G.E.: Multiscale modeling of hematologic disorders. In: Ambrosi, D., Quarteroni, A., Rozza, G. (eds.) Modelling of Physiological Flows, pp. 289–331. Springer, New York (2012)

Dissociated Adsorption

▶ Chemical Vapor Deposition (CVD)

DNA Computing

▶ Molecular Computing

DNA FET

▶ Nanostructure Field Effect Transistor Biosensors

DNA from First Principles

Rosa Di Felice[1,2], Agostino Migliore[3] and Daniele Varsano[2]
[1]Department of Physics and Astronomy, University of Southern California, Los Angeles, CA, USA
[2]Center S3, CNR Institute of Nanoscience, Modena, Italy
[3]Department of Chemistry, Duke University, Durham, NC, USA

Synonyms

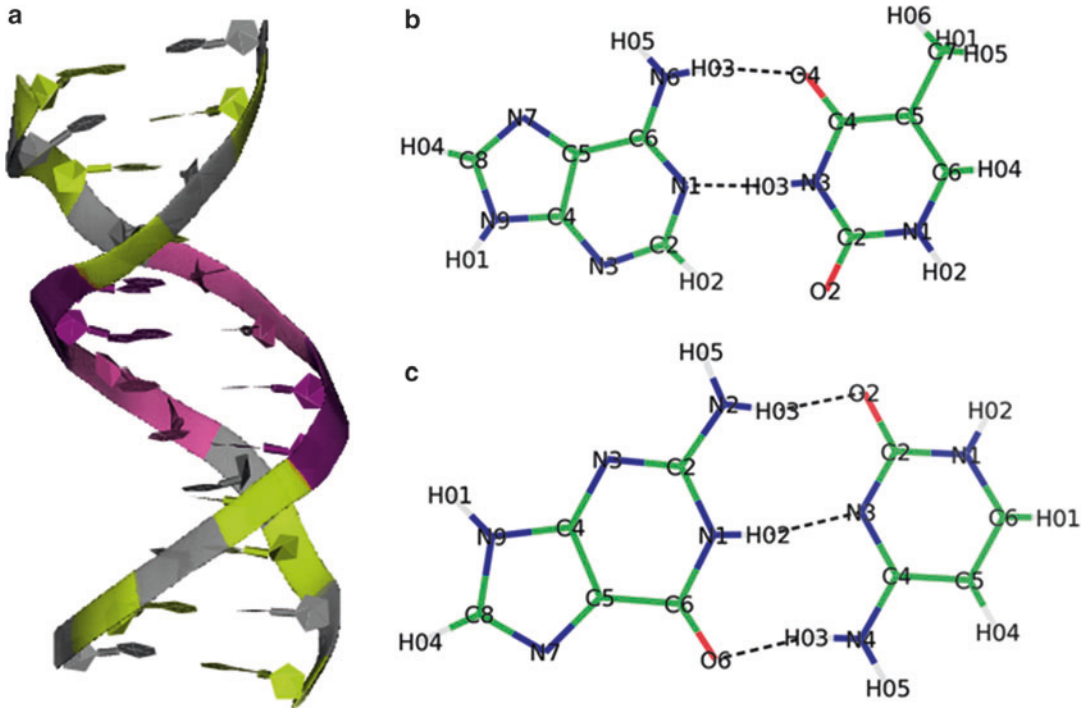
Deoxyribonucleic acid; Electronic structure of nucleic acids; First-principle calculations; Structural fluctuations

Definition

DNA is the acronym of deoxyribonucleic acid, also known as the molecule of life. It is the basic constituent of the genetic code. Although genetic studies date back to the activities of Gregor Johann Mendel in the 1860s, the three-dimensional (3D) structure of the genetic molecule was disclosed by James Watson and Francis Crick only about a century later in 1953, after the pioneering X-ray crystallization experiments of Rosalind Franklin and Maurice Wilkins. A canonical DNA molecule is a right-handed double helix with a pitch of 3.4 nm and a diameter of 2.1 nm, made of two strands running in opposite directions. Each strand is a sequence of four different nucleotides (guanine = G, cytosine = C, adenine = A, and thymine = T) separated from each other along the strand by 0.34 nm and connected through a deoxyribose backbone. The two strands are connected by Watson-Crick hydrogen bonding between the bases: guanine pairs specifically to cytosine and adenine pairs specifically to thymine. Guanine and adenine are purines, formed by two coplanar heterocycles (one penta-ring and one hexa-ring); cytosine and thymine are pyrimidines, formed by one heterocycle (hexa-ring). The structure of the double helix and of the base pairs (bp's) is illustrated in Fig. 1 [1].

The electronic structure of materials is at the core of their response to the environment and external stimuli (e.g., light) and of their exploitation for practical purposes: DNA is a biological material that does not violate this rule. In DNA, the response to the environment, as well as the inherent properties, is a crucial factor for its biological role, including normal function and disease; the exploitation for practical purposes includes nanomechanical machines, molecular electronics, and molecular computing devices.

Computational approaches to DNA target both the structure and the electronic properties (the latter include optical excitations and mobile charges). They can be either "parameterized" or "from first principles": the former quality means that the methodology contains external information from experiments or theories, and the latter



DNA from First Principles, Fig. 1 Structure of canonical double-stranded DNA (dsDNA). (**a**) 3D structure in a ribbon representation of a canonical B-DNA molecule, the Drew-Dickerson dodecamer with palindromic sequence 5'-CGCGAATTCGCG-3' (pdb code 1BNA). Different colors identify different bases: G = yellow; C = silver; A = purple; T = pink. (**b**, **c**) 3D structure in a stick representation of H-bonded AT (**b**) and GC (**c**) pairs. Atomic labels are shown at the atomic sites. *White*, *green*, *blue*, and *red* sticks represent hydrogen, carbon, nitrogen, and oxygen atoms. Hydrogen bonds are shown with *black dashed lines*

quality means that the methodology does not contain any information external to the investigated system, but only the chemical nature of the constituent atoms and a viable initial atomic structure. First-principle calculations are feasible only for small fragments because they are too cumbersome for entire DNA molecules: in fact, since they do not require external parameters, they must solve the quantum states of the electrons in a self-consistent manner. Yet, knowledge of the double-helical structure of entire oligomers is required as a viable starting point of first-principle computational procedures: this can be either obtained from the nucleic acid database (a structure is identified by a 4-letter code in a file with extension "pdb"; http://www.rcsb.org/pdb/home/home.do) or from empirical simulations (e.g., classical molecular dynamics or Monte Carlo). At odds with inorganic materials, the electronic properties of a soft DNA molecule are intertwined with the conformational fluctuations. Therefore, first-principle calculations of the electronic structure should address a variety of conformations in order to have a relation with reality. Electronic structure calculations of DNA bases, base pairs, and small fragments of stacked base pairs can be performed with wavefunction-based methods and density-based methods. This entry briefly defines the methodologies, presents milestones achieved with representative methodologies, and discusses the intimate relation between atomic and electronic structure.

First-Principle Methods that Have Been Used to Study DNA

The title "DNA from first principles" refers to computational studies of the physicochemical properties of DNA fragments and oligomers.

Such properties are primarily the 3D atomic structure, the ground-state electronic energy and density, and the response of the electrons to light and external voltage. More complicated spectroscopic features may also be determined but are omitted from this entry. First-principle studies can be conceptually organized in two categories: (fp1) determination of electronic structure for one (or several) frozen atomic configuration(s) and (fp2) determination of the structure and structural motions.

The first problem is tackled with a variety of computational methods that solve for the properties of the electrons, either in the ground state or under the influence of optical and electrical excitations, at constrained atomic positions. The second problem is tackled with ab initio molecular dynamics (MD) or molecular mechanics (MM) simulations, which evolve the atomic geometry in the effective potential of the on-the-fly computed electron density. Molecular dynamics simulations solve second-order equations of motion and give the evolution of the system in time. Molecular mechanics, in the most popular nuance, consists of solving first-order equations of motions for the search of equilibrium configurations, namely, a number of local minima that are closest to trial structures.

For a system of N electrons and K nuclei with exclusively Coulombic interactions, the total Hamiltonian is:

$$H = -\sum_{i=1}^{N} \frac{1}{2}\nabla_i^2 - \sum_{I=1}^{K} \frac{1}{2M_I}\nabla_I^2 - \sum_{i=1}^{N}\sum_{I=1}^{K} \frac{Z_I}{r_{iI}} + \sum_{i=1}^{N}\sum_{j>1}^{N} \frac{1}{r_{ij}} + \sum_{i=1}^{K}\sum_{j>1}^{K} \frac{Z_I Z_J}{R_{IJ}}. \quad (1)$$

In Eq. 1, atomic units are used, for which Planck's constant is 2π, the mass of the electron is 1, and the charge of the electron is 1; M_I is the ratio of the mass of nucleus I to the mass of the electron; Z_I is the atomic number of nucleus I. Electrons and nuclei are labeled by lower- and upper-case subscripts, respectively. The additive terms in Eq. 1 from left to right are the kinetic energy of electrons, the nuclear kinetic energy, and the electron-nucleus, electron-electron, and nucleus-nucleus Coulomb interactions. Within the Born-Oppenheimer approximation, based on the fact that nuclei are much heavier and therefore much slower than electrons, the many-body wavefunction is factorized into a wavefunction for the electrons that depends parametrically on the nuclear coordinates and a nuclear wavefunction. In such conditions the many-body Schrödinger equation can be separated in the electronic and nuclear problems. In the solution of the Schrödinger equation for electrons, the nuclear kinetic energy is zero and the nuclear Coulomb interaction enters as an external potential: this is category (fp1) in the introduction above, namely, the search for the electronic structure of frozen systems. On the other hand, the ground-state electronic Hamiltonian can be used to determine the Hellmann-Feynman forces and evolve the nuclear coordinates either in time or toward the local minimum: this is category (fp2) above. The nuclei are treated as classical particles because of the heavy mass, relative to the electron mass. In many first-principle approaches, a nucleus is replaced by an ion core, which is the nucleus plus the core electrons, which are not explicitly included in the solution of the electronic Schrödinger equation. Electrons are inherently quantum particles, namely, their behavior cannot be explained or determined with the laws of classical physics. Therefore, first-principle methods are based on a quantum mechanical (QM) formulation.

The electronic problem is solved either by wavefunction-based methods, in which the Schrödinger equation for the many-electron wavefunction is solved, or density-based methods, in which the total energy is written as a functional of the electron density whose minimization produces the ground state. In all approaches summarized in the next three subsections, the single-electron wavefunction (a molecular orbital) needs to be expanded in a basis set. Popular basis sets are Gaussians (Pople and Dunning sets) and

plane waves. Gaussians are localized on atoms, while plane waves are delocalized and offer a convergence criterion. Localized basis sets are mostly used for molecules; plane waves are more appropriate for periodic crystals. Localized basis sets may cause a basis set superposition error (BSSE) for systems composed of multiple components, which must be corrected to predict meaningful formation energies [2].

Wavefunction-Based Methods

The many-electron wavefunction is antisymmetric for the exchange of any two electrons. Its most general form is an expansion of Slater determinants. In any computer implementation, this expansion is truncated to one (single-determinantal approaches) or more (multi-determinantal approaches) Slater determinants. In the first case, the many-body problem is reduced to many one-electron problems in which the potential energy in a single-electron Hamiltonian is the sum of the electron-nucleus Coulomb interaction and the Hartree and exchange terms of the electron-electron Coulomb interaction. The Hartree term is the energy of electrostatic interaction of an electron with all electrons. The exchange term removes the unphysical interaction of each electron with itself (i.e., the so-called electronic self-interaction energy) that is inherent in the Hartree term, by an exchange (or Fermi) hole surrounding each electron. This is a manifestation of the Pauli exclusion principle that forbids two electrons with the same spin to have the same spatial localization (namely, to occupy the same molecular orbital). This is the Hartree-Fock (HF) method, which neglects all (dynamical) correlations among electrons apart from the static correlation established by the Pauli principle. In fact, the correlation energy is defined, in computational chemistry and physics, as the difference between the exact energy and the HF ground-state energy of a system. The HF method is solved by a self-consistent field (SCF) procedure starting from guess single-particle orbitals and changing them iteratively under the orthonormalization condition until the energy change between two iteration steps is below a prefixed threshold, and it is assumed that the minimum system energy is thus achieved.

Electron correlation is included in the configuration interaction (CI) scheme by keeping more than one Slater determinant in the many-body wavefunction expansion. This procedure is computationally very cumbersome, because many Slater determinants should be included to respect physical conditions, such as the extensive nature of the energy: all approximate forms of CI, except for full CI (which is exact in the limit of complete basis set), do not have this property. Methods beyond HF (or post-HF methods) that recover this property are coupled cluster (CC) approaches and Møller-Plesset (MP) perturbation theory. The most affordable of such methodologies, which have given satisfactory results for DNA bases and base complexes, are CC with single and double excitations (CCSD) and MP to the second order (MP2). Other successful post-HF methods are the complete active space self-consistent field (CASSCF) approach and its second-order perturbation formulation (CAS-PT2): these methods yield reliable results in complex chemical situations.

All these methods, along with semiempirical treatments (INDO, ZINDO, and similar), are efficiently described in standard quantum-chemistry textbooks [2].

Density-Based Methods

Wavefunction-based methods reach a great accuracy and therefore predictive power. However, their use is limited to few-atom systems due to their big computational cost and bad scalability with the size of the system. A viable way to solve, even if in an approximate way, the many-electron problem relies on density-based methods, where the electronic density instead of the wavefunction is used as the fundamental variable. Density functional theory (DFT) is based on the Hohenberg-Kohn (HK) theorems (1964) [3] and permits, via the solution of the Kohn-Sham (KS) ansatz (1965), to afford the calculation of the ground-state properties of systems of many electrons with reasonable accuracy and predictive power.

The first HK theorem proves that the ground-state properties of a many-electron system, subject to an external potential $V_{ext}(r)$, are uniquely determined by its electronic charge density. In the second HK theorem, an energy functional of the density is defined:

$$E[\rho] = F[\rho] + \int V_{ext}(r)\rho(r)dr, \quad \text{with} \quad F[\rho] = T[\rho] + V_{ee}[\rho],$$

where T is the electron kinetic energy and V_{ee} is the electron-electron interaction. $F[\rho]$ is a universal functional, in the sense that it is the same in all physical systems. The theorem proves that the exact ground-state energy is a global minimum of this functional and the density that minimizes the functional is the exact ground-state density ρ_0.

The HK theorems reduce the complexity of the many-electron problem from 3N spatial degrees of freedom of the wavefunction (N number of electrons) to just three spatial degrees of freedom of the density. This is an exact reformulation of the many-body Schrödinger equation, but its effective exactness is limited because the form of the universal functional $F[\rho]$ is unknown.

Practical applications of DFT are doable in the KS ansatz, whose basic idea is to replace the interacting many-electron system with a fictitious system (the KS system) made of noninteracting electrons that generate the same ground-state density ρ_0 of the real system. For the KS system, the energy functional is

$$E[\rho] = Ts[\rho] + \frac{1}{2}\int \frac{\rho(r)\rho(r')}{|r-r'|} drdr' + \int V_{ext}(r)\rho(r)dr + E_{xc}[\rho], \quad (2)$$

where $T_s[\rho]$ is the kinetic energy of the noninteracting system. The second term is the classical electron-electron repulsion (Hartree term), and $E_{xc}[\rho]$ is the "exchange-correlation" functional, which contains all the many-body effects of the system. The wavefunctions of the KS system can be then expressed as a single Slater determinant, as the system is composed of independent particles. Minimizing Eq. 2 with respect to the density brings to the KS equations, a set of N coupled single-particle Schrödinger-like equations:

$$\left(-\frac{1}{2}\nabla^2 + \int \frac{\rho(r)}{|r-r'|}dr' + V_{ext}(r) + \frac{\partial E_{xc}[\rho]}{\partial \rho}\right)\psi_i^{KS} = \varepsilon_i^{KS}\psi_i^{KS}, \quad (3)$$

in which the subscript i labels each of the N electrons.

The total density can then be written as a sum over all electrons of the single-particle electron densities in terms of the single-particle wavefunctions:

$$\rho(r) = \sum_{i=1}^{N} f_i \left|\psi_i^{KS}\right|^2 \quad (4)$$

where f_i is the occupation number and $\frac{\partial E_{xc}[\rho]}{\partial \rho}$ is the exchange-correlation potential, $V_{xc}[\rho]$. It is clear that the complexity of the problem is strongly reduced, but the dependence of the E_{xc} functional on the density is not known, and therefore, the solution of Eq. 3 requires approximations. Various approximations have been proposed since the early days of DFT, and development of exchange-correlation functionals is still an active field of research.

The simplest family of E_{xc} functionals derives from the local density approximation (LDA), according to which the functional depends on the electron density only at each point. The LDA was successfully used for several years for periodic crystals and is reliable in cases in which the electron density resembles that of a homogeneous electron gas: this condition poorly describes the situation in molecules, and this is why traditionally the electronic structure of molecules was computed with wavefunction-based methods. The first level of improvement relative to the LDA is constituted by the GGA (general gradient approximation) functionals, where an explicit dependence on the density gradient is also considered: popular examples are PBE and BLYP.

GGA functionals gave the earliest acceptable results of DFT on molecular systems. More recently, hybrid functionals were developed: these are not pure functionals of the density, but they also depend on the KS wavefunctions, including fractions of the exact exchange term from the Hartree-Fock (HF) theory and exchange-correlation terms from other approximations, for instance, GGA. Several hybrid functionals were developed, differing from each other essentially for the amount of exact exchange. Hybrid functionals yield a better description of molecular properties with respect to "pure" density functionals. It is important to note that these functionals have a dependence on empirical parameters, usually assigned by fitting procedures on atomic and molecular experimental data. The most popular and successful hybrid functionals that have been considered for studying DNA from first principles are B3LYP (20 % of HF exchange), PBE0 (25 %), BHH (50 %), M05-2X (57 %), and the M06 family with different percentages of exact exchange. BHH and the M06 family are now considered the most accurate for DNA, at least for spectroscopy and electron transfer predictions. Another important set of functionals for the study of DNA is the VdW-DF family, designed to include in DFT the description of Van der Waals interactions, which are entirely missed from "pure" density functionals.

Kohn-Sham energies and wavefunctions computed by DFT can be used as a starting point for Green function-based many-body perturbation theory (MBPT) methods, as the GW approximation for the computation of quasiparticle energies and the Bethe-Salpeter equation for absorption spectra. MBPT methods correct DFT eigenvalues for estimating quasiparticle energy levels as measured, e.g., by photoemission experiments.

The basic idea of the DFT, i.e. the possibility to consider the electronic density as the fundamental variable to solve the many-body problem, has been extended in 1984 to time-dependent phenomena (TDDFT) [4]. TDDFT is based on the Runge-Gross theorem, and the time-dependent KS (TDKS) equations allow calculations of the dynamical properties of many-electron systems in the presence of time-dependent potentials. TDDFT, as DFT, is an exact theory and in principle can be used for any time-dependent situation; it is mostly used for the computation of photoabsorption cross-section spectra and excitation energies of molecular systems. Exchange-correlation potentials in TDKS equations are much more complex than in static DFT, because at each time, they depend on the value of the density at all previous times. Most TDDFT calculations are done with potentials derived in the adiabatic approximation, which consists of considering the ground-state exchange-correlation potential evaluated at the instantaneous time-dependent density, i.e. discarding all the memory effects:

$$V_{xc}^{A}(r,t) = V_{xc}^{gs}[\rho_0](r)\big|_{\rho_0(r)=\rho(r,t)}. \qquad (5)$$

Despite its crudeness, the approximation in Eq. 5 gave reliable results for the optical absorption and circular dichroism (CD) spectra of DNA bases and fragments of stacked and H-bonded pairs. CD spectroscopy is particularly powerful to discriminate between various helix motifs, and the possibility of computing this property opens the way, by comparison to experimental data, to resolve DNA structures that cannot be resolved by conventional methods such as X-ray and nuclear magnetic resonance (NMR).

The quality of TDDFT results is strongly dependent on the choice of the exchange-correlation functional; in particular, charge-transfer excitations are poorly described by local or semilocal functionals.

Ab Initio Molecular Dynamics

Molecular dynamics (MD) simulations solve the equations of motion of each atom in a target system in three dimensions:

$$\vec{F}_i = m_i \vec{a}_i, \qquad (6)$$

where i is a running index over all atoms in the system, m_i its mass, $\vec{a}_i$ its acceleration, and $\vec{F}_i$ the force acting on it at a given instant of time. In turn, the force is expressed as the first derivative of the potential with respect to the atomic coordinate and the acceleration as the second derivative of the

atomic coordinate with respect to time, which results in the second-order differential equations:

$$\frac{m_i d^2 \vec{r}_i}{dt^2} = -\frac{dV}{d\vec{r}_i}. \quad (7)$$

The numerical solution of Eq. 7 yields the ith atomic coordinate at time $t + \delta t$ given its value at times t and $t - \delta t$. When continued in time, this procedure gives the time evolution of the structure. Necessary ingredient to implement this procedure is the knowledge of the interatomic potential $V\left(\vec{r}_i\right)$.

The most widely used implementation of MD is a classical version in which information about the electrons in the system is not necessary, because the potential is parametrized from experimental or theoretical data: this is not a first-principle approach; we mention it here to highlight, by comparison, the crucial aspect of ab initio MD. With the available computer power, classical MD simulations of 10–20 bp DNA molecules can routinely cover a time span of the order of 1 microsecond. The timescale of 1 millisecond is presently attainable only with a special hardware, the computer Anton, donated to the University of Pittsburgh by David E. Shaw and available to the research community by proposal submission and evaluation.

Ab initio MD does not use empirical parameters to express $V\left(\vec{r}_i\right)$. Each atom is composed of a nucleus and a certain number of electrons. In the spirit of the Born-Oppenheimer approximation, the motion of nuclei and electrons occurs on very different timescales, because their masses differ by three orders of magnitude. Consequently, the solution of the many-body problem nuclei + electrons can be separated: (i) at frozen atomic coordinates, the electronic ground state of the system can be determined with the electronic structure self-consistent methods summarized above; (ii) the ground-state energy of the electronic system, along with the ion-ion interaction, defines the instantaneous potential for that atomic configuration and can be used to solve Eq. 7; (iii) the new instantaneous coordinates are found. Then the cycle goes on giving the time evolution of the system. This is also known as Born-Oppenheimer molecular dynamics. This approach is rather cumbersome because it requires at each atomic configuration the attainment of the exact ground state of the electrons, which may require many self-consistent cycles. It was unfeasible 30 years ago, when in 1985 Roberto Car and Michele Parrinello invented the first practical solution of ab initio MD, now known as the Car-Parrinello (CP) method [5]. CP MD does not search for the instantaneous electronic ground state at each atomic configuration and rather evolves nuclei and electrons simultaneously in a Lagrangian formulation, keeping the electrons very close to the ground state by a suitable choice of the integration time and effective mass. The time step for solving Eq. 7 with classical MD is of the order of 1 femtosecond. The time step for solving Eq. 7 with quantum MD (either Born-Oppenheimer or CP) is of the order of attosecond, because of the resolution necessary to probe the electron dynamics, which is about three orders of magnitude faster than the nucleus dynamics. Consequently, the early ab initio MD simulations were limited to few picoseconds and nowadays can reach the order of nanoseconds.

Another approach to solve the quantum MD problem is based on the density functional tight-binding (DFTB) MD, in which the electronic problem is formulated as a tight-binding approach with parameters derived from DFT: this approach is not fully first principle.

Examples of applications of CP and DFTB MD to DNA oligomers are included later.

Electronic Structure Calculations of DNA Fragments

The field of "DNA from first principles" evolved from addressing individual components to larger fragments and then entire oligomers, as the methods became more efficient and the computers more powerful. Particularly, crucial achievements were from the hardware point of view, the attainment of efficient parallel scaling of electronic structure and molecular dynamics codes, and from the methodological point of view, the

development of gradient-corrected and hybrid density functionals in DFT.

Structure, Electronic Structure, and Optical Properties of Bases and Base Pairs

Quantum mechanical studies of DNA bases and base pairs were initially driven by the need of knowing the structure of nucleic acids to a high degree of accuracy, for implications in biology and medicine.

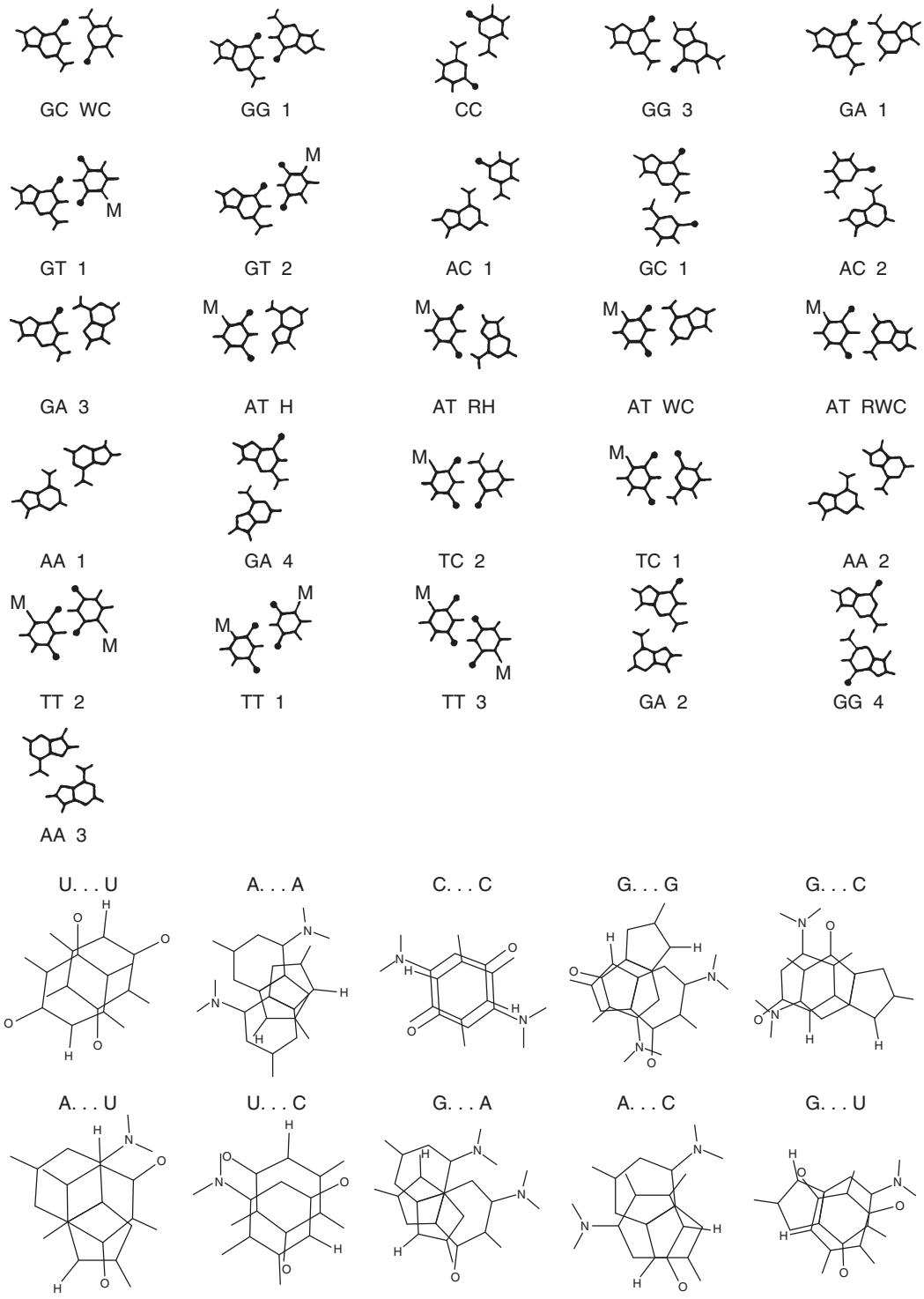
Ab initio methods were applied since the late 1980s to investigate equilibrium structures of isolated, stacked, and H-bonded bases, as well as the pairing energies. These studies had the twofold objective of studying the structure and determining interaction energy parameters to enable classical simulations. The early calculations were performed without inclusion of electronic correlations, while calculations done in the mid-1990s started to include a consistent treatment of electronic correlations, which substantially enhanced the accuracy and predictive power.

An early work that still remains a landmark was done in the 1990s by Hobza, Leszczynski, Sponer, and collaborators [6]. The individual bases were investigated at the MP2 level with medium-sized polarized basis sets of atomic orbitals, to disclose the energetics of tautomers, finding that for each base the most favorable tautomer is the one that occurs in nature. The results also disclosed the internal structure of each base, which was later an important input knowledge for studies of base pairs. Other properties of interest in the studies of single bases were the protonization energies, the electrostatic potential, the polarizability, the gas-phase electron affinities and ionization potentials, and the nonplanarity. In the past decade, individual bases were also the object of coupled cluster, multiconfigurational perturbation, and DFT + GW methods, with comparable results.

The first reliable data on the geometry and binding energy of all possible H-bonded base pairs (Fig. 2) was obtained by relaxing the structure at the HF level and determining for the optimized bases the stabilization energies at the MP2 level, using a minimal basis set with diffuse polarization functions and including the BSSE. The results revealed that the Watson-Crick GC pair is the most energetically favorable, while the most unfavorable is an AA pair. It was found that the number of H-bonds is not the only factor that controls the energetics of H-bonding; the dipole interaction is another major determinant. Several pairs were found to be not exactly planar, especially due to rotations of the NH_2 group because of partial sp^3 N hybridization. Later on, canonical and noncanonical H-bonded base pairs were computed with higher-order wavefunction-based methods and DFT, with particular attention to noncanonical pairs with a biological relevance (e.g., methylated cytosine).

The computation of stacked base pairs (called dinucleotides in genomic contexts, Fig. 2) presents more difficulties than that of H-bonded pairs, because stacking implicitly involves dispersion interactions, which are accounted for in the correlation part of the Coulomb electron-electron interaction, which is absent in HF and difficult to grasp in post-HF methods with limited basis sets and in DFT. Thus, the early studies were done without structural relaxation, because the optimized structure could not be trusted. Alternatively, researchers performed single-point (frozen-structure) calculations of several structures for each dinucleotide, varying the internucleotide distance, angle, and inclination and determining the minimum energy geometry for each dinucleotide. For the most favorable pair, the stabilization energy could then be computed at the MP2 level. In a second stage, MP2 was applied for full gradient optimizations of the stacked pairs. Results indicated that the stacking energies for the various pairs span approximately the same range as H-bonding energies. More refined calculations including a better description of the correlation (coupled cluster method in the complete basis set limit) started in 2002, and nowadays, available results from wavefunction-based calculations for the stacking energy of DNA bases are very accurate, indicating energies in the range of 0–15 kcal/mol depending on the context, with an accuracy of around 3 %.

The application of DFT methods to DNA base pairs became credible only with the advent of hybrid functionals that include a part of exact



DNA from First Principles, Fig. 2 H-bonded and stacked base pairs found from first principles. H-bonded (*top*) and stacked (*bottom*) base pairs that have been computed by various first-principle methods in the 1990s (Adapted with permission from Hobza and Šponer [6]; copyright 1999 American Chemical Society)

exchange and after 2008 the VdW-DF family of functionals. However, not all functionals perform well. Specialists in the field can now find in the literature an assessment of which functionals are reliable and which are not, so that basic tests are less mandatory than in the past, though it is always a good idea to carry out tests of the functional and of the basis set before undertaking any prediction of structural, optical, and electronic properties (including electron transfer parameters) of DNA bases and their pairs. DFT methods have become increasingly important with increasing the size of the target system, because correlated wavefunction-based approaches become unfeasible beyond a small size.

Computation of the optical properties of nucleobases and base pairs is an important tool for understanding the interaction of DNA with sunlight. Excited states of nucleic acids are the first step of chains of events responsible for UV photodamage, leading to photocarcinogenesis; they also play a role in the mechanisms of ultrafast non-radiative decay pathways and DNA photoprotection. Both wavefunction-based methods and TDDFT approaches have been used to study optical properties of DNA bases [7]. There is a good agreement between different computational methods in the calculation of the bright $\pi\pi^*$ transitions (from a bonding orbital π to an antibonding orbital π^*), also relative to experimental findings. Such a good agreement is not found for the "dark" $n\pi^*$ transitions (from a nonbonding state to an antibonding orbital π^*). The discrepancies are partially due to the lack of reliable experimental data for these transitions, because of the very weak oscillator strength. Despite the difficulty of predicting $n\pi^*$ transitions from first principles, they are intensely studied because of their biological relevance. In fact, detailed knowledge of the relative energies of $n\pi^*$ and $\pi\pi^*$ excitations is of paramount importance in understanding the processes of dissipation of the electronic energy by internal conversion and prevention of photochemical reactions.

Gas-phase calculations carried out at the CIS, CAS-PT2, and TDDFT levels agree in assigning the energy order between $\pi\pi^*$ and $n\pi^*$ transitions: the lowest-energy excitation is $\pi\pi^*$ in guanine and cytosine and $n\pi^*$ in thymine and uracil. The situation is more controversial for adenine: wavefunction-based methods predict the lowest-energy excitation to be $\pi\pi^*$, while TDDFT(B3LYP) results indicate the lowest-energy excitation to be $n\pi^*$, very close in energy to a $\pi\pi^*$ excitation.

Some excited-state calculations have been done with the inclusion of solvent effects either implicitly with dielectric models or explicitly (mostly water), allowing comparison with solution-phase experiments. It has been found that the presence of the solvent destabilizes the $n\pi^*$ transitions (blueshift), while the $\pi\pi^*$ transitions are almost unaffected. The different sensitivity of the $n\pi^*$ and $\pi\pi^*$ excited states to the solvent environment leads to a reversal of the ordering of these transitions in thymine, uracil, and adenine, as computed by TDDFT.

Besides the interpretation of spectroscopy experiments, a thorough knowledge of the excited-state properties of DNA bases is of crucial importance for the rationalization of the observed non-radiative decay of the excited states to the ground state. Several computational studies have been dedicated to this issue by looking at the excited-state potential energy surfaces along a reaction coordinate, and different mechanisms have been proposed, involving nonadiabatic coupling between low-lying excited states and conical intersections between excited- and ground-state energy surfaces.

Calculations of the optical excitations of Watson-Crick base pairs revealed that both $\pi\pi^*$ and $n\pi^*$ transitions could be identified as excitations of the monomeric units, leading to an absorption spectrum very similar to the sum of the monomeric spectra. While the $\pi\pi^*$ excitations are only slightly affected by H-bonding, the $n\pi^*$ excitations reveal a consistent blueshift. Different calculations also showed the appearance of excited states of charge-transfer (CT) character (excitations formed by a transition from an electron in a moiety to a virtual state in the complementary moiety). These excitations are particularly interesting for implications in proton-transfer reactions between bases upon light excitation [8]. They have higher energies than the first $\pi\pi^*$ and $n\pi^*$, except for TDDFT

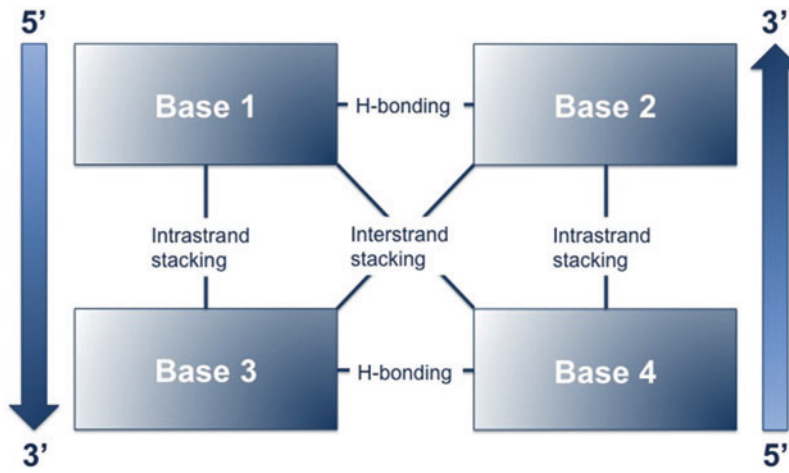
calculations with local or semilocal functionals, which poorly predict CT excitations.

The most evident effect of stacking interactions in the optical properties of nucleobases is the hypochromism, namely, the reduction of the absorption intensity upon stacking: the intensity of a peak for stacked bases is lower than the sum of the intensities for individual bases. This effect is routinely used in experiments to monitor denaturation of the DNA double helix. TDDFT calculations on stacked bases correctly predict hypochromicity, as well as some measured sequence-dependent shifts. Combining computational affordability and reliability, TDDFT is now considered a reference framework to study the optical absorption properties of nucleic acid fragments, except for CT excitations that rigorously still need higher first-principle methods (or calibrated exchange-correlation functionals). Despite this limitation, in the last years, TDDFT has also been applied to elucidate decay mechanisms of stacked base pairs, trying to explain the observed long lifetime in terms of possible decay channels that involve CT states. Long-lived excited states are potentially responsible for DNA photodamage, which is a relevant target in biomedical research.

Structure, Electronic Structure, and Optical Properties of Stacked Base-Pair Dimers

Calculations on stacked base-pair dimers (Fig. 3) constitute a step toward more realistic systems in which both H-bonding and stacking interactions are present and influence each other in a sequence-dependent manner. The results of quantum chemical calculations based on CCSD(T) theory produced an energy ranking for ten possible dinucleotides that appear in base-pair dimers with idealized structures. Individual intrastrand stacking energies are in the range of −10.8 to −1.6 kcal/mol, and individual interstrand stacking energies are between −4.8 and 3.1 kcal/mol, with a total stacking energy for the stacked dimers in the range between −17.3 (CG step) and −11.2 (GG step) kcal/mol. Cooperativity between stacking and H-bonding was revealed in vertical ionization energies in a dimer of two AT pairs with the purines on the same strand.

It was also recognized that the magnitude of the intrinsic gas-phase stacking energy could not be determined unambiguously using a single geometry for each sequence, because of the strong dependence of the interaction energy on the structural configuration. Sequential multistep procedures, based on pruning base-pair dimers from different snapshots from a molecular dynamics simulation and computing the electronic structure of such dimers, have been used to have quantitative estimations of relative energies between different structures. These calculations showed that the sequence dependence of the thermodynamic stability of DNA cannot be explained in terms of intrinsic interactions between stacked base pairs only.



DNA from First Principles, Fig. 3 Scheme of bp dimers. Cartoon of a base-pair dimer with indication of H-bonding and interstrand and intrastrand stacking interactions. Each of Bases 1 and 3 can be G, C, A, or T; Bases 2 and 4 will be the complementary base

Sequence-dependent conformations of DNA base-pair dimers have been also the subject of study using DFT methods. These studies have provided additional information on the sequence-dependent DNA stability, for instance, highlighting the role of the binary interaction between adjacent sets of base pairs in the twist variability observed in high-resolution experiments and identifying the role of the methyl group of thymine in the stabilization of the base-pair dimers.

Calculations of optical properties and excited-state energy profiles on stacked base-pair dimers appeared in the first decade of the twenty-first century and were performed by TDDFT. Using the adiabatic local density approximation, it has been shown that the absorption spectra of stacked bp dimers reveal the fingerprints of both H-bonding (spectral shifts) and stacking (hypochromicity) [9]. TDDFT calculations performed using the M05-2X functional and including solvent effects via a continuum model on stacked AT dimers identified dark CT states that involve both interstrand and intrastrand effects. By relaxing the lowest bright and dark excited-state geometries, Vincenzo Barone and collaborators were able to disclose the specific role of H-bonding and stacking in the excited-state decay and in emission processes in stacked bp dimers. These calculations showed that the optical absorption in DNA fragments is a concerted activity of all bases, while emission occurs essentially from individual monomers, mainly thymine. It was shown that the ultrafast excited-state relaxation channel exists in the $(A:T)_2$ dimer and that CT states between stacked adenines are responsible for the long living states observed experimentally in A-rich DNA. The dynamics of excited states was related to femtosecond pump-probe measurements in single-strand poly-A and double-strand poly(dA)-poly(dT) fragments by Bern Kohler and coworkers.

Electronic Coupling Parameters Between Stacked Bases and Base Pairs

Charge (electron or hole) transfer in DNA, which can involve either electrons or holes, is an important property that has been studied for both biomedical research (oxidative damage and repair, biosensors) and nanotechnology (electrical materials and devices).

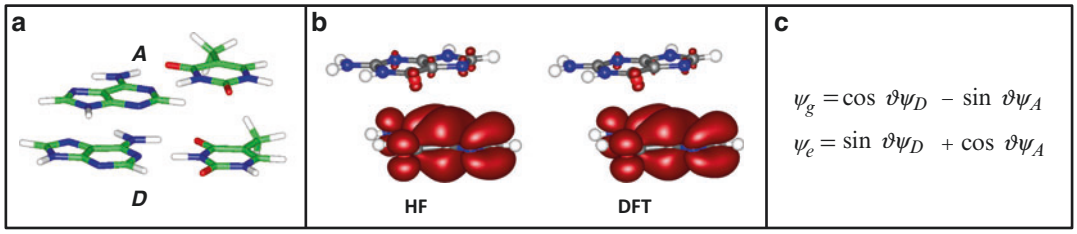
All information required to investigate a charge-transfer system can be derived from the wavefunctions for its ground state and for the excited states that correspond to the charge localizations before and after the transfer process. Such wavefunctions can be obtained by DFT, HF, or post-HF methods.

Diverse models have been proposed to describe hole transfer through DNA depending on the chain length: ballistic transport, superexchange, hopping, and phonon-assisted polaronic hopping. Whatever the mechanism for charge propagation through the DNA chain, the electronic coupling between adjacent nucleobase pairs is a critical determinant of charge-transfer efficiency.

Since hole transfer through the π-stacked nucleobases dominates the charge propagation, the electronic couplings can be computed for the fundamental DNA unit shown in Fig. 4, where D (A) is the electron donor (acceptor) group. The inter-pair electronic coupling V_{DA} may be obtained from the adiabatic ground (ψ_g) and first-excited (ψ_e) wavefunctions. The diabatic (localized) electronic wavefunctions with the excess charge (an electron hole) on $A(\psi_A)$ and $D(\psi_D)$ are linear combinations of ψ_g and ψ_e. A master approach to computing the coupling from the adiabatic wavefunctions is the generalized Mulliken-Hush (GMH) method, where

$$V_{DA} = \frac{\mu_{ge}\Delta E_{ge}}{\sqrt{\left(\mu_g - \mu_e\right)^2 + 4\mu_{ge}^2}}. \qquad (8)$$

In Eq. 8, μ_{ge} is the adiabatic dipole transition moment, i.e., the matrix element of the dipole moment operator between ψ_g and ψ_e; $\mu_g(\mu_e)$ is the mean dipole moment in state $\psi_g(\psi_e)$; and $\Delta E_{ge} = \langle\psi|H|\psi_e\rangle - \left\langle\psi_g|H|\psi_g\right\rangle$ is the adiabatic vertical excitation energy (H is the electronic Hamiltonian). This method has been mostly implemented by using HF one-electron wavefunctions. In the HF approach, ψ_D and ψ_A may be obtained as symmetry-broken diabatic wavefunctions. Symmetry breaking means that



DNA from First Principles, Fig. 4 Charge transfer through DNA. (**a**) Adenine-thymine dimer. A base pair is the donor (*D*) group and the other is the acceptor (*A*) group. (**b**) A reduced model for hole transfer between the two adenines. In a one-electron picture, the HOMO of the base couple represents the ground-state charge density. HF and DFT calculations produce very similar HOMO distributions for this system. (**c**) The adiabatic and diabatic sets of electronic states are related by a rotation of an angle ϑ in the system state space. $\psi_D(\psi_A)$ is not an electronic wavefunction for the $D(A)$ group, but a wavefunction for the total *D*-*A* system with the hole essentially localized on $D(A)$

the wavefunction has a lower symmetry than the nuclear framework [2]. This error in the HF calculation of a wavefunction is exploited to obtain generally nonorthogonal wavefunctions whose asymmetry implies charge localization in *A* or *D*.

Related to the above approach is the fragment charge difference method, which writes the electronic coupling as

$$V_{DA} = \frac{\Delta q_{ge} \Delta E_{ge}}{\sqrt{\left(\Delta q_g - \Delta q_e\right)^2 + \left(\Delta q_{ge}\right)^2}}. \qquad (9)$$

Here, $\Delta q_g(\Delta q_e)$ is the difference between the excess charges on the *D* and *A* fragments in state $\psi_g(\psi_e)$; $\Delta q_{ge} = \sin 2\vartheta$, where ϑ is the rotation in the system state vector space that brings from the adiabatic to the diabatic states under the condition of maximum *D*-to-*A* CT (see Fig. 4). The fragment charges are well defined in a DNA dimer or in other systems with well-separated *D* and *A* groups, for which Eq. 9 is then isomorphic to Eq. 8. Equation 9 has been evaluated using the HF Koopmans approximation, which allows estimating ΔE_{ge} as the energy difference between the highest occupied molecular orbital (HOMO) and the HOMO–1 for the neutral DNA system. These electronic orbitals become the lowest unoccupied molecular orbital (LUMO) and the HOMO in the cationic CT system, and changes in the orbitals are neglected.

Applications of this method to DNA dimers with all possible combinations of nucleobases showed secondary effects of the backbone and high sensitivity of the inter-pair coupling to DNA step parameters, thus corroborating that thermal fluctuations can produce significant changes in the electronic couplings, hence failure of the Condon approximation (which neglects the coupling dependence on the nuclear coordinates *Q*). This conclusion was confirmed by coupling calculations on molecular dynamics (MD) snapshots via an INDO/S implementation of the GMH method and via a method that estimates the coupling from the overlap between the nonorthogonal HOMOs of individual bases.

Indeed, in most real situations, the initially produced charge localization and its transfer to a neighboring site amount to initial and final states with non-negligible spatial overlap, and the effective electronic coupling or (generalized) transfer integral or ET matrix element (or, also, electronic coupling matrix element) that enters the CT rate constant is [10]

$$V_{DA}(Q) \equiv \frac{1}{1 - S_{DA}^2(Q)} \left| H_{DA}(Q) - S_{DA}(Q) \frac{H_{DD}(Q) + H_{AA}(Q)}{2} \right|, \qquad (10)$$

where $H_{IF} = \langle \psi_I | H | \psi_F \rangle$ $(I, F = D \text{ or } A)$, $S_{DA} = \langle \psi_D | \psi_A \rangle$, and the Q dependence is explicitly shown. Given physical localized electronic states, a good orthogonal diabatic set needs to be related to the nonorthogonal one by the Löwdin transformation [10], which imposes a constraint on the rotation angle ϑ.

The fragment charge difference method, using HF wavefunctions, neglects the dynamic electronic correlation, which is instead included, albeit with approximations, in DFT. The use of DFT faces two technical issues: the inaccuracy in the description of the adiabatic excited state and unphysical electronic self-interaction [2]. The first problem was solved by a recent approach that avoids using ψ_e [10] by writing the effective electronic coupling in Eq. 10 as

$$V_{DA} = \left| \frac{\alpha\beta}{\alpha^2 - \beta^2} \Delta E_{DA} \left(1 - \frac{\alpha^2 + \beta^2}{2\alpha\beta} S_{DA} \right) \frac{1}{1 - S_{DA}^2} \right| \tag{11}$$

where $\alpha \equiv \left\langle \psi_D | \psi_g \right\rangle, \beta \equiv \left\langle \psi_A | \psi_g \right\rangle$, and $\Delta E_{DA} = \langle \psi_D | H | \psi_D \rangle - \langle \psi_A | H | \psi_A \rangle$. This formula can be used in any computational setup, but has been mainly implemented with DFT and hybrid-DFT approaches, defining the many-body electronic states as Slater determinants of KS orbitals. Despite a long-standing debate on the use of these orbitals, several studies have shown that the "the KS molecular orbital model is physically well-founded and has certain advantages over the HF model" [11] because it includes electronic correlations.

Equation 11 can be used for any Q and S_{DA} values, except for the transition-state coordinate Q_t (where $\Delta E_{DA} = 0$; note that ΔE_{DA} can be used as a reaction coordinate [12]). At $Q = Q_t$, Eq. 11 has an eliminable discontinuity. The condition $Q = Q_t$ defines a zero-measure subset of a molecule conformational space, and Eq. 11 can be used with high accuracy also in the close proximity of Q_t (as measured by $\Delta E_{DA}/V_{DA} \ll 1$), where the Condon approximation is valid.

Hybrid DFT implementations of Eq. 11 using the BHH hybrid exchange-correlation functional and the Minnesota hybrid meta-functionals M06-Hf and M06-2X led to descriptions of DNA electronic properties (including V_{DA} and ΔE_{gs}) in agreement with post-HF results from CASSCF and CAS-PT2. Knowledge of the ground-state wavefunction of the CT system, in addition to the diabatic states where the hole is localized on the donor or on the acceptor, offers the best description of a CT system insofar as this state is properly described. In this regard, functionals such as B3LYP, PBE, and B97 do not allow for suitable implementations of Eq. 11, because they poorly describe electron density tails. For example, even a slight delocalization of the excess charge in the ground state can lead to a significant overestimation of the electronic coupling for weakly coupled off-resonance D and A groups. The diabatic CT states can be obtained from fragment approaches or from constrained DFT (CDFT), which provides the most "natural" approach to charge localization within DFT [10].

The fragment-orbital approach is adopted (i.e., CT between the HOMOs of neighboring DNA bases is considered) in DFTB methods, which are based on the expansion of the DFT energy with respect to charge density fluctuations. DFTB uses simplifications that enable relatively fast electronic coupling estimates in entire DNA oligomers.

Larger Fragments

Electronic properties of DNA fragments larger than dimers of base pairs have been studied using high-accuracy first-principle methods. Particular attention has been devoted to CT through trimers of bases and base pairs.

In the case of superexchange, the electronic coupling involved in the CT rate constant is

$$V_{\text{superexchange}}(D, A) = \frac{V(D, \mathrm{B}_1) V(\mathrm{B}_n, A)}{\Delta E(D, \mathrm{B}_1)} \prod_{j=1}^{N-1} \frac{V(\mathrm{B}_j, \mathrm{B}_{j+1})}{\Delta E(D, \mathrm{B}_{j+1})}, \tag{12}$$

where N (for a trimer, $N = 1$) is the number of nucleobases bridging the D and A base pairs (or bases, when full localization of the transferring

charge on bases is assumed to reduce the computational cost); $V(D, B_1)$, $V\left(B_j, B_{j+1}\right)$, and $V(B_n, A)$ are the couplings between adjacent base pairs; and $\Delta E(D, B_j)$ are the respective energy gaps. Equation 12 represents a tight-binding bridge model, where electronic couplings between nonadjacent sites are neglected, and assumes that the localization of the transferring excess charge on a bridge site would imply much higher system energy than its localization in *D* or *A*. Hence, the bridge levels correspond to virtual states, that is, superexchange is bridge-mediated *D*-*A* tunneling. Equation 12 is derived using the Green's function formalism [13] and reduces to Larsson's formula for a single bridge.

The hole transfer matrix elements involved in Eq. 12 have been computed by using molecular orbitals from semiempirical (CNDO/S) computations, HF wavefunctions, DFT wavefunctions constructed from the HOMOs of individual bases, CASSCF on base trimers, and full-electron hybrid-CDFT computations on base-pair trimers.

Tunneling through the coupling in Eq. 12 shows an approximately exponential decay with the *D*-*A* distance, and charge hopping (i.e., transient real occupation of bridging bases by the transferring excess charge) takes place at sufficiently large *D*-*A* distances. Each hop amounts to a superexchange process. The hopping mechanism is expected to produce an approximately inverse relation between the number of hopping sites (namely, *D*-*A* distance) and *D*-to-*A* arrival rate. A different inverse-power law is observed in highly nonlinear DNA chains, such as DNA hairpins. The transition from coherent tunneling to incoherent hopping has been predicted theoretically and experimentally. First-principle evaluation of this mechanism transition is nowadays enabled by supercomputers that allow DFT computations on hundreds of atoms.

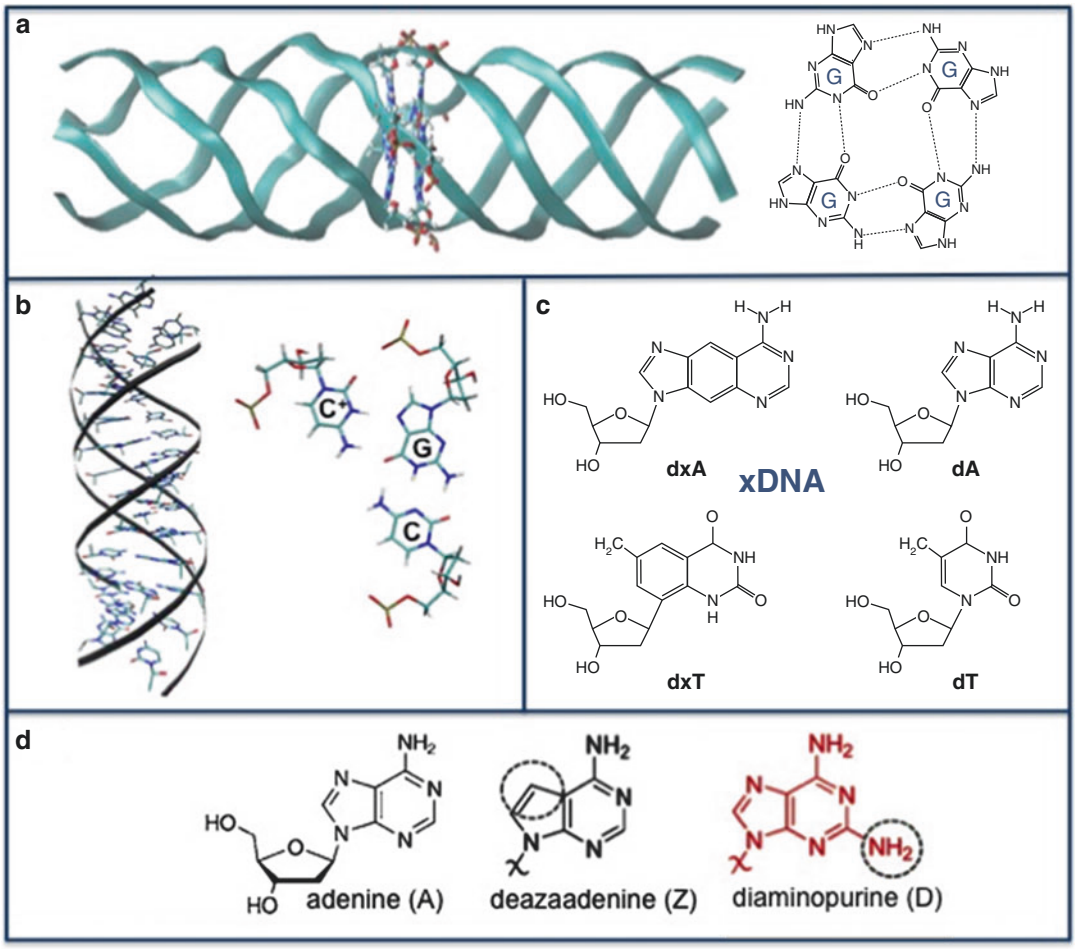
An alternative interpretation of the experiments that showed the coherent-to-incoherent transition of the hole transport mechanism at sufficiently long *D*-*A* distances is based on the phonon-assisted polaron hopping mechanism. Thermal nuclear motions (phonons) support the propagation (hopping) of the geometry distortion and environmental charge polarization that accompanies and promotes the hole localization over some bases (which is a "delocalization" compared to the hole localization on a single base). Hence, the charge transport proceeds through hopping of the quasiparticle made of the electron hole and surrounding polarization field (polaron). The delocalization of the polaron is limited by the environmental charge polarization, the internal reorganization of the base pairs, and the same continuous nuclear fluctuations.

The delocalization of the hole on more base pairs can lead to an intermediate mechanism, distinct from both superexchange and hopping, where coherence is not completely lost during the hole motion. Experimental evidence reported by Barton's group at Caltech points to transport via DNA domains that form transiently, irrespective of the charge and therefore distinct from the DNA segments implied in a polaron. This interpretation might be related to recent theory on resonant enhancement of charge transport efficiency: fluctuations to transient geometries that bring multiple nucleobases into electronic energy degeneracy may support band-like transport during the resonance lifetimes.

Exotic Helical Conformations and Modified Bases

Nucleic acid conformations are not limited to the canonical Watson-Crick double helix. Topological and chemical modifications are viable, either naturally or through (bio)chemical synthesis. The most relevant modifications in biology and medicine are the G-quadruplex (topological) and methylated cytosine (chemical). Other nucleic acid motifs that have been studied from first principles include triplex DNA (a triple helix mediated by protonated cytosine), xDNA (in which each base can be expanded by an aromatic cycle), and duplex DNA where natural bases are modified with the insertion or removal of functional groups or oxidation. These motifs are illustrated in Fig. 5.

The G-quadruplex (Fig. 5a) is a quadruple helix constituted of only the guanine base. The stacking unit is the G-tetrad, composed of



DNA from First Principles, Fig. 5 Modified DNAs that have been studied from first principles. Structure of selected hexotic nucleic acids. (**a**) The G-quadruplex, with formula of the stacking unit on the right. (**b**) The C^+GC triplex, with atomic structure of the stacking unit on the right. (**c**) Size-expanded bases dxA and dxT, compared to dA and dT. (**d**) Chemical modifications of adenine that have been used in charge-transfer experiments

four guanines fortified by eight H-bonds. G-quadruplexes can form from genomic sequences (only 2–4 G-tetrads, ~1–2 nm long including edge loops) or be produced by enzymatic synthesis (several tetrads, up to 1 μm long): in both cases, the folding results in several different polymorphs. The rigidity of this nucleic acid motif, the high density of stacking sites that give $\pi-\pi$ electronic coupling, and the possibility of controlling the length and the folding process make it appealing for nanotechnologies. Indeed, a special G-quadruplex constituted of four parallel strands connected by biotin-avidin recognition is the only form of nucleic acids that has so far supported significant electrical currents over lengths of up to 100 nm when deposited on an inorganic substrate. First-principle DFT calculations of the electronic coupling parameter between two adjacent G-tetrads in a frozen configuration, extracted from a MD trajectory of a parallel G-quadruplex at room temperature, revealed a much higher value than all published values for duplex DNA and for other G-quadruplex polymorphs. Furthermore, the ab

initio electronic coupling was used to validate a multi-hopping model that successfully fits the measured current-voltage characteristics. Earlier work on the ground-state electronic structure of G-quadruplex was done on two stacked tetrads and on a periodic wire: for the latter, channels in which electrons can move with high mobility were found. Dynamical approaches to the G-quadruplex are mentioned in the next section.

CT parameters have been computed from first principles for frozen xDNA base-pair dimers with different structures, revealing that this variety can be more conducive to charge transfer than natural dsDNA. However, a strong dependence on the structural details was also revealed, to an extent that should be analyzed in a dynamical fashion in order to have predictive power for real measurements.

M-DNA is another possible modification that is not illustrated in Fig. 5: it is obtained by doping duplex DNA with transition metal cations. There is no three-dimensional atomic structural resolution for M-DNA. Hypothetical models, which are supported by high-level DFT calculations by Di Felice and coworkers of the ground-state electronic structure and optimized geometries, consist of a metal ion that breaks the hydrogen bonds within a natural base pair and forms partially chemical bonds with the bases. M-DNA mimics have been obtained experimentally and simulated from first principles also using hydroxypyridone instead of the natural bases.

Other DNA modifications illustrated in Fig. 5d are described below with a combination of classical MD and DFT.

Relation Between Atomic and Electronic Structure

It was already noted in the preceding sections that the atomic structure and electronic structure are intertwined. In particular, excited-state electronic properties, including CT, are very sensitive to structural details. Because in reality the conformation of nucleic acids is dynamic and not static, it is important to compute the electronic and optical properties not for just one conformation but for several of them. The average of the electronic properties for several viable dynamical conformations, or the weighted average over few representative structures that embody the whole trajectory, can be taken as a good estimate of the real situation. The dynamics should be computed for stable oligomers or wires in a solution environment. This computational program has become feasible only in the twenty-first century and more effectively in the past few years. Two possible strategies of doing such calculations are summarized, along with salient results, in the next two subsections.

We note that the backbone has been neglected in most of the QM calculations on frozen atomic configurations, guided by the fact that charge-transfer and optical excitations involve electronic orbitals of the bases. However, dynamical calculations, which target the determination of the structure and structural fluctuations, cannot omit the backbone because this is an integral part of the structure. Hence, in ab initio MD, the backbone is directly included in the quantum dynamics, while in a combined multistep approach, it must be included in the classical MD step and can be included or excluded in the QM step. The environment must also be included in molecular dynamics, and usually, it has been included as explicit solvent molecules and counterions. QM calculations can be done in the gas phase or solution phase: most solution-phase QM calculations on DNA have been done with implicit solvent models.

Ab Initio Molecular Dynamics Simulations

The straightforward approach is to perform MD simulations in which the potential energy to evolve the motion of the nuclei, instead of being parametrized, is given by the ground-state electronic energy at that instantaneous configuration. The nuclei are evolved with Newton's classical equations of motions, but with forces that encompass the full quantum information for the electrons.

The group of Michele Parrinello published in 2002 in *Physical Review Letters* a report on the

electronic structure and optical conductivity of wet Z-DNA, a periodic d(GpC) obtained from a quenched CP MD simulation. In 2005, the same group published in the same journal a dynamical simulation of the same periodic polymer at both room temperature and low temperature, on the timescale of 1 picosecond. In both letters, the BLYP functional was used: we have pointed out above that BLYP is not capable of catching subtle correlation effects; yet, the ground-state electronic structure is qualitatively correct. These works excluded a strict relationship between helix distortions and charge localization and suggested a proton-coupled charge-transfer mechanism.

Marcus Elstner, Gianaurelio Cuniberti, and coworkers simulate charge transfer through DNA with DFTB simulations: the methodology has been published in the *Journal of Physical Chemistry* in 2008. This approach can directly compute the electronic couplings, in a structurally sensitive manner. Only the relevant part of the system, where one wants to evaluate CT, is treated with a quantum method, while the rest of the system is classical (QM/MM). The structure influences the charge-transfer parameters simultaneously in two ways: (i) by computing the transfer integrals for many instantaneous configurations and (ii) through a coupling term between the QM and MM parts. The method is very efficient and enables simulations of realistic oligomers. However, it is based on a simplified description of both the ground-state electronic structure and electronic couplings. It has been applied to a variety of duplex DNA sequences and a G-quadruplex, taking into account chemical modifications of the bases. For the G-quadruplex, this methodology revealed superior charge-transfer efficiency than in duplex DNA.

Intrinsic limitations of these methodologies are the short time over which the dynamics can be sampled, which is generally of the order of picoseconds, and the restricted precision on the electronic structure.

Combined Multistep Procedure

The multistep procedure consists of these four steps: (I) performing classical molecular simulations in explicit solvent of stable oligomers over timescales that can reach 1 ms and are routinely hundreds of nanoseconds; (II) extracting bp dimers or G-tetrads or C^+GC triads (Fig. 5), including or not the backbone, from several dynamical snapshots at regular time intervals or from representative structures obtained by clustering the trajectories or by physicochemical principles; (III) computing the ground-state electronic structure and/or electronic couplings and/or optical properties for such small fragments of stacked units; and (IV) evaluating the desired property as a time average or weighted average.

Uzi Landman, Gary Schuster, and coworkers published in *Science* in 2001 quantum mechanical calculations for native and oxidized configurations of the 4-bp B-DNA duplex d(5'-$G_1A_2G_3G_4$) with an intact sugar-phosphate backbone including Na^+ ions and a hydration shell. The quantum calculations were performed, in the framework of DFT with a gradient-corrected functional, for several configurations extracted from a classical MD simulation of the same oligomer at room temperature. This work proposed an ion-gated mechanism for charge transport through DNA, based on the fact that the dynamical distribution of ions and water molecules controls the charge delocalization through the stack. In this early work, the electronic coupling parameters were not explicitly computed.

Troisi and Orlandi in 2002 published a similar approach but including the explicit calculation of electronic couplings as a function of time. The work included classical MD simulations of some duplex DNA oligomers over 40 ps, followed by quantum chemical calculations between all the pairs of adjacent bases extracted from the dynamical snapshots at regular time intervals of 10 fs. Overall, couplings between adjacent bases showed an oscillatory behavior and a large variance.

The hybrid multistep approach was applied by Di Felice and collaborators to duplex and triplex DNA, with publications appearing in 2012–2013. In all cases, the classical MD simulations were of the order of 20–50 ns. For duplex DNA and

G-quadruplex, the target property was charge transfer, while it was optics for triplex DNA. Overall, these studies proved that the methodology is powerful and sufficiently manageable to be applicable to many different structures.

Alexander Voytiuk, in collaboration with the groups of Anh Tuan Phan and Maria-Elisabeth Michel-Beyerle, applied the procedure to compute transfer integrals for many different G-quadruplex motifs, for identifying which topology yields high inter-tetrad electronic coupling and which geometrical parameter of the helix mostly affects the electronic coupling. In this study, published in 2013, the MD simulations were conducted for 200 ns each, for various quadruplex topologies. The electronic structure was simplified at the semiempirical INDO/S level, and the fragment charge difference method was used for the computation of the electronic coupling parameter. The study pointed out a remarkable dependence of CT rates on the G-quadruplex structure, while it could not identify a single helix parameter that by itself correlates strongly to the inter-tetrad electronic coupling. This negative conclusion is a hint to the cooperativity of the entire quadruplex structure to determine the electronic coupling. Rigorously, this study is not from first principles, because INDO/S is a semiempirical treatment of the QM problem, but we mention it here because it is very similar in spirit to combined MD + QM studies in which the QM part is from first principles.

The groups of George Schatz and Jiry Sponer applied a similar methodology to investigate the optical excitations and the relative stability of DNA duplexes with various sequences, respectively. A combination of MD and quantum chemistry was also applied, in a somewhat different spirit, to compute the optical conductivity of wet DNA and physical approaches to DNA sequences: for the optical conductivity, the full time dependence was not studied; for DNA sequencing, the study was actually not from first principles because a parametric tight-binding model was employed for the electronic structure.

Conclusion

The investigation of DNA from first principles has undergone a terrific progress since the first applications in the mid-1980s. The early studies could only be done for individual bases in frozen configurations and at a relatively low level of theory (missing or poor description of electronic correlations). Nowadays, it is possible to compute entire oligomers at high levels of theory (good description of electronic correlations) including structural fluctuations. Results of first-principle calculations of DNA bases, base pairs, base-pair dimers, larger fragments, and entire oligomers have contributed to our understanding of the structure, electronic and optical properties, charge-transfer efficiency, and structure-electronic relationships in a variety of nucleic acid motifs, including duplex and triplex DNA and the G-quadruplex. This knowledge is fundamental to control oxidative damage and repair of the genetic material and to manipulate DNA derivatives for nanotechnology applications. Yet, we still need to develop the link from fundamental properties to principles of operation, as well as the rules for controlling the structure to obtain a desired function. We expect that in the near future, this gap will be filled if the enthusiasm in DNA from first principles will be sustained: the available information will be largely expanded and will be much more accurate, because of the present possibility of computing real systems with high accuracy, enabled by theoretical, software, and hardware developments. Ultimately, studies of DNA from first principles will have a huge impact not only on fundamental science and technology but also on society through health-care and commercial products.

Cross-References

- ▶ Computer Modeling and Simulation of Materials
- ▶ DNA FET
- ▶ First Principles Calculations
- ▶ Nanomaterials

References

1. Calladine, C.R., Drew, H., Luisi, B., Travers, A.: Understanding DNA – The Molecule and How It Works. Elsevier Academic Press, San Diego (2004)
2. Jensen, F.: Introduction to Computational Chemistry, 2nd edn. Wiley, Chichester (2007)
3. Fiolhais, C., Nogueira, F., Marques, M.A.L. (eds.): A primer in Density Functional Theory. Lecture Notes in Physics, vol. 620. Springer, Berling/Heidelberg (2003)
4. Marques, M.A.L., Ullrich, C., Nogueira, F., Rubio, A., Burke, K., Gross, E.K.U. (eds.): Time-Dependent Density Functional Theory. Lecture Notes in Physics, vol. 706. Springer, Berlin/Heidelberg (2006)
5. Car, R., Parrinello, M.: Unified approach for molecular dynamics and density-functional theory. Phys. Rev. Lett. **55**, 2471–2474 (1985)
6. Hobza, P., Šponer, J.: Structure, energetics, and dynamics of the nucleic acid base pairs: nonempirical *Ab initio* calculations. Chem. Rev. **99**, 3247 (1999)
7. Crespo-Hernández, C.E., Cohen, B., Hare, P.M., Kohler, B.: Ultrafast excited-state dynamics in nucleic acids. Chem. Rev. **104**, 1977–2019 (2004)
8. Šponer, J., Lankas, F., Leszczynski, J. (eds.): Computational studies of RNA and DNA, Challenges and Advances in Computational Chemistry and Physics. Springer, Heidelberg (2006)
9. Varsano, D., Di Felice, R., Marques, M.A.L., Rubio, A.: A TDDFT study of excited states of DNA bases and base assemblies. J. Phys. Chem. B **110**, 7129 (2006)
10. Migliore, A.: Nonorthogonality problem and effective electronic coupling calculation: application to charge transfer in π-stacks relevant to biochemistry and molecular electronics. J. Chem. Theory. Comput. **7**, 1712–1725 (2011)
11. Baerends, E.J.: Perspective on "Self-consistent equations including exchange and correlation effects". Theor. Chem. Acc. **103**, 265–269 (2000)
12. Warshel, A.: Dynamics of reactions in polar solvents. Semiclassical trajectory studies of electron-transfer and proton-transfer reactions. J. Phys. Chem. **86**, 2218–2224 (1982)
13. Nitzan, A.: Chemical Dynamics in Condensed Phases. Oxford University Press, Oxford (2007)

DNA Manipulation

▶ Robot-Based Automation on the Nanoscale

DNA Manipulation Based on Nanotweezers

Nicolas Lafitte[1], Yassine Haddab[2], Yann Le Gorrec[2], Momoko Kumemura[1], Laurent Jalabert[1], Christophe Yamahata[3], Nicolas Chaillet[2], Dominique Collard[1] and Hiroyuki Fujita[4]
[1]LIMMS/CNRS-IIS (UMI 2820), Institute of Industrial Science, The University of Tokyo, Meguro-ku, Tokyo, Japan
[2]FEMTO-ST/UFC-ENSMM-UTBM-CNRS, Besançon, France
[3]Microsystem Lab., Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[4]Center for International Research on MicroMechatronics (CIRMM), Institute of Industrial Science, The University of Tokyo, Meguro-ku, Tokyo, Japan

Synonyms

Biocharacterization; Biomolecules; Biophysics; MEMS; Microsystems

Definition

DNA manipulation based on nanotweezers is the handling of DNA molecules using very small-size tweezers for characterization purpose.

Introduction

Since the 1990s, studies on single-molecule biophysics have brought a large amount of information on DNA properties. The development of manipulation techniques and especially the emergence of new tools have allowed direct measurement of physical and chemical properties of DNA molecules. These properties turn out important for cell machinery understanding as structural modifications in the molecule induce changes in the interaction properties with proteins, and conversely.

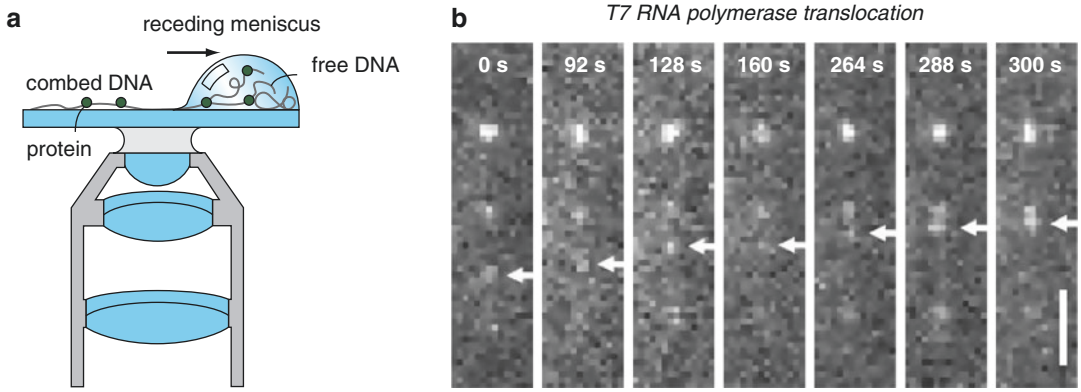
For example, topoisomerase enzymes unwind and wind DNA in order to facilitate DNA replication. A helicase protein moves along the unwound DNA, separating the two annealed nucleic acid strands of the double-stranded DNA. Polymerase enzymes reproduce new strands against the single-stranded DNA templates. During all the replication process, the physical properties and the biological functions of the DNA change according to the interactions with the proteins allowing (or proscribing) the sequence of events [1–4].

Large advances in fluorescence microscopy have been achieved enabling single-molecule visualization. Scientists developed powerful manipulation techniques where the DNA is attached to a substrate and observed by fluorescence [5, 6]. For instance, polymerase enzymes were fluorescently labeled to monitor their activity on a single molecule of DNA previously combed onto a surface (Fig. 1) [7]. Meanwhile, new approaches were developed to directly interact with the molecules, e.g., optical tweezers, magnetic tweezers (Fig. 2), atomic force cantilevers (known as AFM), and microfibers [8–10]. The response of single DNA molecules to a stretching or a twisting stress gives direct access to the physical properties of the molecule. Remarkable mechanical properties of DNA were discovered, and protein interactions with DNA were measured by the changes in the mechanical properties of the strands [11, 12].

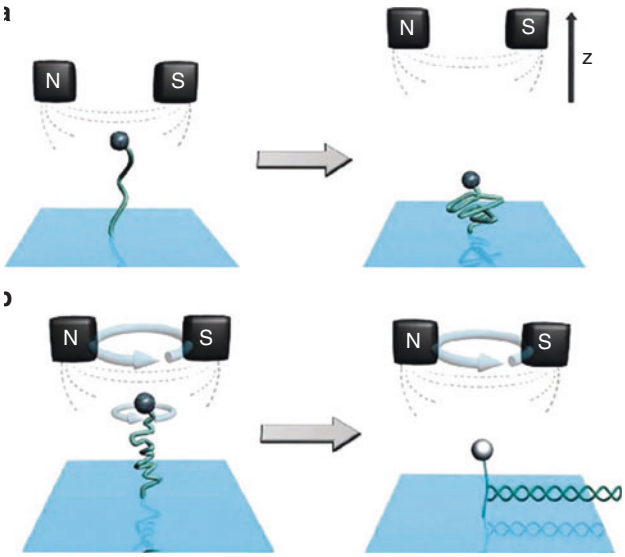
However, all these new techniques, providing remarkable access to the biophysical properties of the molecules, still have a low throughput since the setup preparation and the experimentations are quite long and done one at a time. To move toward systematic biological or medical analysis, micro- and nano-electromechanical systems (known as MEMS or NEMS for Micro/Nano-Electro-Mechanical Systems) are more appropriate tools. They can be easily integrated into biological manipulation systems. Furthermore they are produced at low cost by an efficient microfabrication process. In [13], DNA molecule bundles are trapped, stretched, and characterized in solution by silicon nanotweezers (Fig. 3).

Working Principle

Figure 3 shows a three-dimensional illustration of the nanotweezers. Two sharp tips are used as electrodes for either dielectrophoresis trapping (cf. "Trapping of DNA molecules" section) or conductivity measurements on the molecules. One tip is fixed when the other one can be moved through an electrostatic actuation. The gap between the electrodes can be sensed through



DNA Manipulation Based on Nanotweezers, Fig. 1 Attachment of single DNA molecules on a glass surface. (**a**) DNA is stretched and immobilized using hydrophobic glass substrate and a receding air-water interface. After rehydratation of the sample the DNA stays firmly attached to the glass slide [5]. (**b**) Real-time visualization of the motion of RNA polymerase along combed DNA strands. The directional movement of the RNA polymerase along a DNA molecule is observed using the incorporation of fluorescent into RNA strand (Scale bar = 2.5 μm) (Image from Ref. [7])



DNA Manipulation Based on Nanotweezers, Fig. 2 Schematic sketch of the magnetic tweezer technique. (**a**) The setup consists of a DNA molecule to a magnetic bead and to a glass surface, for instance. A pair of magnets above the bead produces a magnetic field gradient (*dashed lines*) along the axial direction, which results in a force on the bead directed up toward the magnets. The magnets can be lowered (raised) to increase (decrease) the stretching force acting on the DNA. Magnet north and south poles are labeled N and S, respectively. (**b**) Thanks to the magnetic properties of the bead, torque forces can also be applied to the sample. Starting from a relaxed configuration, the DNA molecule becomes more and more tangled as the magnets rotate, and eventually loops of helices (plectonemes) are formed. The formation of plectonemes induces a progressive linear decrease in the DNA extension, which is analogous to the decrease in extension of a twisted and tangled telephone cord (Image from Ref. [20])

a differential capacitor that measures the relative displacement of the moving electrode. Thus, the device consists of three parts:

- Two sharp tips (Fig. 3c)
- A series of electrostatic comb-drive actuators (Fig. 3a)
- A differential capacitive sensor (Fig. 3b)

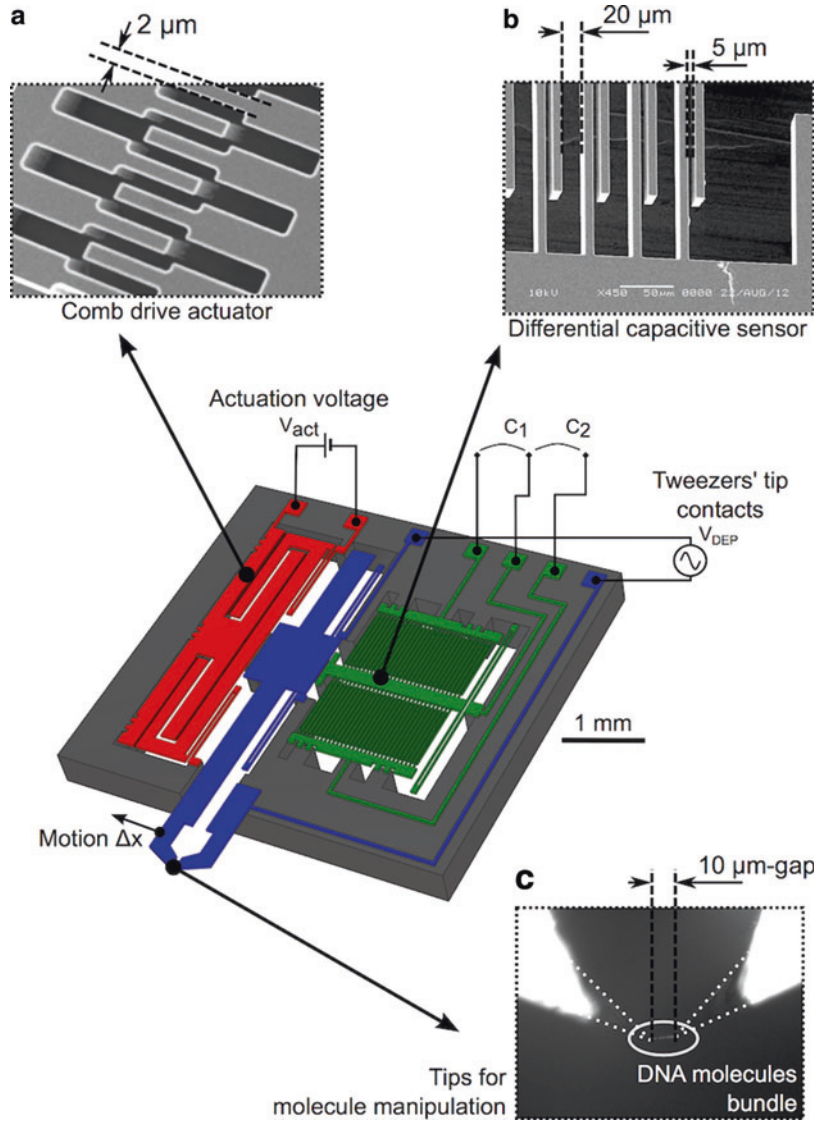
Since [14], laterally driven electrostatic comb drive architectures are commonly used in MEMS devices with many industrial applications (e.g., microresonators, switches, micromirrors, accelerometers). When a voltage is applied between the two electrodes, the mobile parts move along the electric field paths tending to reduce the gap between the electrodes.

A differential capacitance sensor measures the tip displacement. This sensor is well adapted for bulk micromachining and compact integration [15]. The sensor consists of a series of central plates that move relatively to fixed external plates, forming two variable capacitances C_1 and C_2. The values of the capacitances change inversely with the gap between the electrodes. However, the measurement is linear for small displacements (<100 nm) compared to the capacitance gap (5 μm). Hereafter, the capacitance variations are measured with an electronic read-out allowing static or dynamic displacement measurements.

The micromachining of the device and of the sharp tips is explained in the next section, and the role of the sharp tips is discussed in the DNA trapping section.

Microfabrication of the Nanotweezers

The silicon nanotweezer with sharp tips is fabricated using standard micromachining from a



DNA Manipulation Based on Nanotweezers, Fig. 3 3D schematic of the silicon nanotweezers for DNA manipulation and characterization. Overall dimensions are 4.5 × 4.5 × 0.5 mm. The tweezers consist of an electrostatic actuator in red, a capacitive sensor in *green,* and sharp tips in *blue*. Applying a voltage V_{act} moves the mobile tip, and the displacement changes the sensor capacitance values C_1 and C_2, which are measured with an adequate electronic read-out. On the other hand the tips are also electrodes that allow applying an electric field in between the tips for molecule trapping. (**a**), (**b**), and (**c**) are respectively close-up views of the actuator comb, the sensor capacitances, and the sharp tips with a bundle of DNA molecules

silicon-on-insulator (SOI) wafer [16]. The SOI wafer is based on a standard 400 µm-thick handle silicon wafer, 2 µm-thick buried oxide (BOX) layer, and 30 µm-thin silicon device layer. The crystalline orientation of the top silicon layer is <100>. The BOX is used both as a stopper for the deep reactive ion etching (DRIE) of the top and the backside silicon and as an electrical insulation layer between the comb drive actuator, the tips, and the capacitive sensor. Controlling the device thickness using SOI wafers and the in-plane dimensions using highly anisotropic etching (i.e., DRIE) allow getting uniform tweezers' characteristics and performances.

The sharp tips are fabricated by wet anisotropic etching along the <111> crystallographic planes. The tips engineering needs a sacrificial silicon nitride patterning, a local oxidation of silicon (LOCOS), and wet etching in TMAH.

Finally the process only requires three lithographic masks: one for defining the area of the sharp tips with silicon nitride (mask #1), one for microstructuring the top silicon layer (mask #2),

and one for the backside etching of the handling substrate (mask #3).

Here is the process flowchart:

1. Deposition of silicon nitride (SiN, 30 nm) by Low-Pressure Chemical Vapor Deposition (LPCVD)
2. Lithography along the <100> plane (mask #1) and Reactive Ion Etching (RIE) of SiN
3. Lithography of the top silicon structures (mask #2)
4. DRIE of top silicon, stopping on the BOX layer
5. Cleaning in "piranha"solution (called SPM) and slightly in buffered HF (10 s)
6. LOCOS of 300 nm
7. RIE of SiN 30 nm (sharp tips area)
8. TMAH wet anisotropic etching of silicon in order to obtain {111} facets (the other structures are protected by the 300 nm-thick oxide)
9. Cleaning in SPM and in buffered HF (until removing the 300 nm oxide layer)
10. Protection oxide growth (100 nm-thick)
11. Lithography on the backside of the wafer (double side alignment, mask #3)
12. DRIE of the backside 400 μm-thick silicon, stopping on the BOX layer.
13. Releasing of the movable structures in vapored HF
14. Deposition by evaporation of aluminum (Al, 100 nm) on the top side for enhancing DNA trapping.

Trapping of DNA Molecules

Figure 4c shows a scanning electron microscope (SEM) image of a DNA bundle anchored between the tips of nanotweezers. The trapping of molecule bundle with silicon tweezers is achieved when an AC electric field causes positive dielectrophoresis (called in the literature DEP): the DNA molecules elongate along the electric field lines and move toward the regions where the field is higher [17]. Furthermore, it is known that DNA molecules adhere to several electrochemically active metals such as Al or Pt. Sharp aluminum-coated electrodes create high electric gradients that cause the ends of the DNA strands to be attracted to the edges of the electrodes and become attached, resulting in the permanent formation of a bridging structure. Indeed, the adhesion of DNA molecules is strong on the metal as the molecules can be grafted (covalent bounds) on the metal and the metal oxide by its biphosphate function, present on the phosphate-deoxyribose backbone of the DNA, creating P-O-Metal link by hydrolysis reaction.

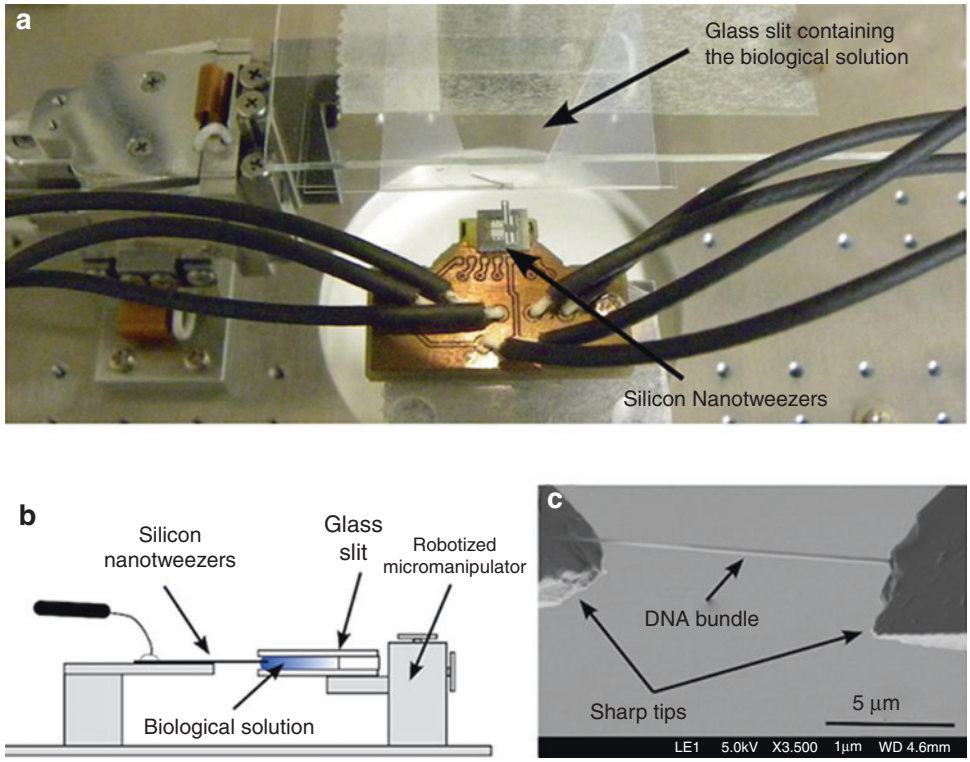
The method can be used to trap single DNA molecules by tuning the parameters appropriately. Two approaches for single-molecule studies are currently investigated:

- A double-stranded λ-DNA molecule can be isolated by DC electrophoresis and consecutively trapped by AC DEP in a microfluidic device.
- Using pulsed DEP, silicon tweezers can be used to trap a single DNA molecule in an aqueous solution.

For all the biological experiments, a solution of double-stranded λ-DNA (48,502 base pairs, 16 μm-long) was used. Double-stranded λ-DNA is the DNA of a bacteriophage known as Enterobacteria phage λ or λ-phage. After dilution in deionized water, a small droplet of the solution was deposited on a glass slide. The tweezers were mounted on an optical microscope and brought in contact with the surface of the droplet with a 3D precision micromanipulator (Fig. 4). Then, a high AC electric field (2 MV/m or 2 V/μm at 1 MHz frequency) was applied for few seconds in order to capture DNA molecules. After the experiments, the bundle could be broken and removed by blowing air and rinsing the tips with water. The DEP experiment could be repeated as long as the aluminum coating remained on the silicon tips.

Biocharacterization on DNA Molecules

The characterization of biological samples is done through the mechanical response of the system {tweezers + DNA bundle}. The frequency response of the tweezers being well identified,

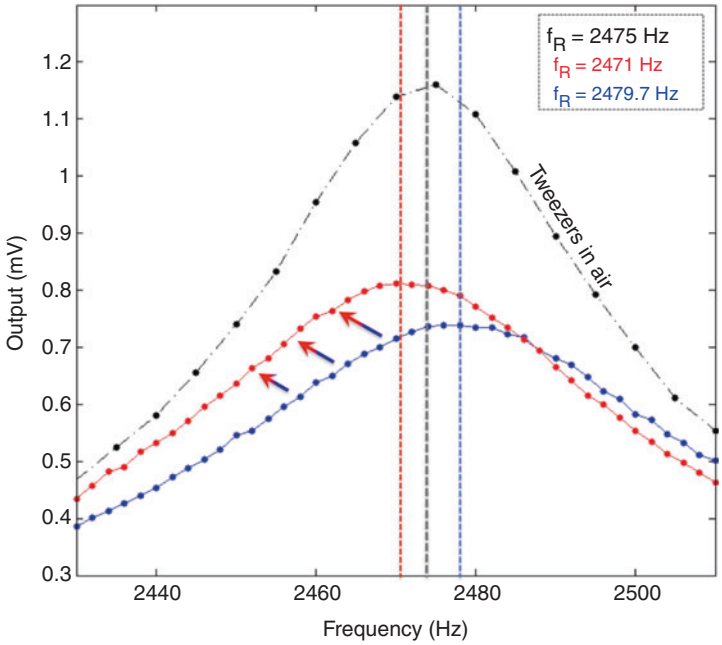


DNA Manipulation Based on Nanotweezers, Fig. 4 Trapping of DNA molecules with silicon nanotweezers by dielectrophoresis (*DEP*). (**a**) The experimental setup consists of the nanotweezers on its PCB with the electrical connections and a glass slit containing the biological solution. (**b**) The schematic shows the importance of the micromanipulators, on which are mounted the nanotweezers and the glass slit, for an accurate insertion of the tips in the DNA solution. (**c**) is a scanning electron microscope (*SEM*) image of a trapped molecule bundle. An AC voltage (V_{DEP}) applied between the two tips of the tweezers allows the attraction and the stretching of dielectric molecules as DNA

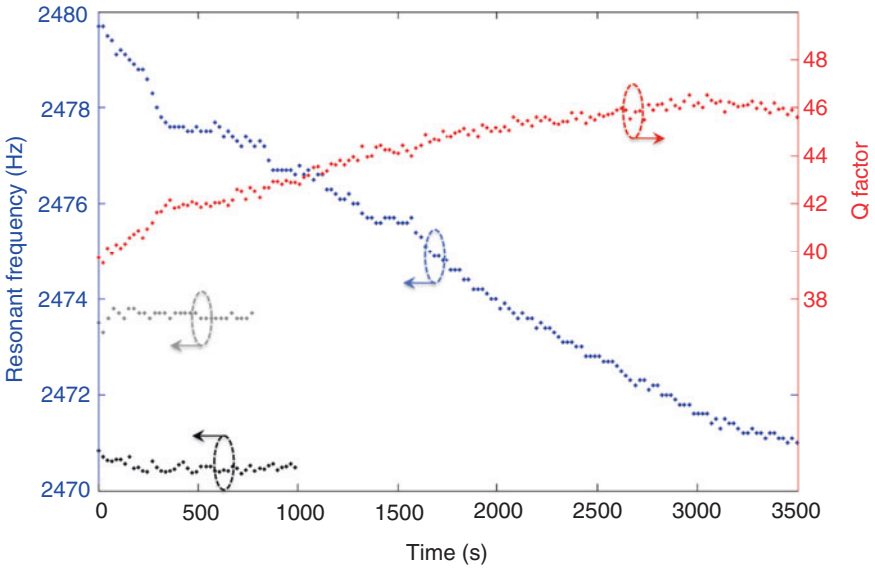
the changes in the frequency response are attributed to mechanical properties of the molecules. Recently, the real-time monitoring of the number of trapped DNA molecules during the dielectrophoresis was proven. During the trapping, the resonant frequency of the system increases resulting from the addition of the bundle stiffness to the sole tweezers' one. Simultaneously, the quality factor tends to decrease as the viscous loss increases with the bundle formation. In Fig. 5, a clear difference can be observed between frequency responses of the nanotweezers with and without trapped DNA molecules.

The DNA physical properties change with protein interactions, inducing changes of the whole system frequency response. Consequently nanotweezers emerge as a relevant tool even for characterization of DNA-protein interactions. In a recent study, the digestion of the bundle by restriction enzymes was monitored in real time through the frequency response of the nanotweezers. HindIII proteins are known as restriction enzymes for DNA. Indeed HindIII recognizes a specific sequence in the DNA code where to cleave the molecule [18].

As for the trapping of molecules, the resonance of the system {Tweezers + DNA bundle} is monitored. Conversely, the resonant frequency decreases till the initial resonant frequency of the sole tweezers when the digestion of the bundle is finished. The quality factor increases (Fig. 6). The evolution of the number of the molecules in the bundle, deduced from the single molecule stiffness ($k_{\lambda\text{-DNA}} = 30$ µN/m), shows an exponential decay



DNA Manipulation Based on Nanotweezers, Fig. 5 Tweezer resonance evolution during digestion of the DNA bundle. *Blue dots* are the frequency response of the tweezers with a bundle (in solution). *Red dots* are the response of the tweezers after digestion of the bundle (without any bundle in solution). For comparison, the resonance of the sole tweezers in air is plotted in *black* [19]



DNA Manipulation Based on Nanotweezers, Fig. 6 Time evolution of the resonance properties of the system {Tweezers + DNA bundle}. Resonant frequency (in *blue dots*) decreases with the digestion of the bundle by the restriction enzymes, when the quality factor (in *red dots*) increases. For comparison, resonant frequency of the sole tweezers is shown in deionized water (in *black dots*) and in a buffer without restriction enzymes (in *gray dots*) [19]

with time, the extracted time constant depending inversely on the enzyme concentration [19].

The minimum number of molecules that can be sensed using such silicon nanotweezers (about 20 molecules) depends on the minimum frequency variation that can be experimentally measured. In recent works dynamic control strategies have been developed to improve the overall

system sensitivity to the variation of the number of DNA molecules. The objective of such control strategy is the detection of a variation on a single molecule of DNA.

Outlook

Since the 1990s, remarkable techniques and tools were developed in order to perform single-molecule manipulation. Although these recent technological advances had a significant impact on the understanding of the interactions between DNA and the molecular machinery they present some drawbacks that open new research issues.

Indeed, in most of the biological experiments on DNA, manipulations rely on complex systems as optical or magnetic tweezers. These techniques have a low throughput since their implementation is quite long and difficult. Furthermore the molecule preparation is done one at a time. Here the direct sensing of biomolecular reactions is demonstrated using silicon nanotweezers. The experimental results obtained with the integrated MEMS nanotweezers showed the efficiency of such devices for DNA characterization. Nevertheless the main drawbacks of MEMS stem from the large size and high stiffness of the system relatively to the biological molecule properties. Control strategies that are currently under development aim at reducing the tweezer stiffness using feedback to improve the sensitivity of the whole process toward single DNA molecule variation.

The improvement of resolution in DNA sensing brought new perspectives and questions to address. The complexity of the DNA machinery is still far from being elucidated. More complex techniques or technologies are under development with, for instance, combination of different tools with fluorescent visualization, in order to follow multiprotein multienzyme complexes. Lastly, new developments on magnetic tweezers have recently achieved first results inside the nucleus of living cells. Performing such in vivo experiments inside cells will provide inestimable knowledge of molecule functions in their inherent environment.

Cross-References

- ▶ 3D Micro-/Nanomanipulation with Force Spectroscopy
- ▶ Basic MEMS Actuators
- ▶ Dielectrophoresis of Nucleic Acids
- ▶ Nanogrippers
- ▶ Optical Tweezers

References

1. Bruce, A., Alexander, J., Julian, L., Martin, R., Keith, R., Peter, W.: The Molecular Biology of the Cell. Garland Science, New York (2002)
2. Neuman, K.C.: Single-molecule measurements of DNA topology and topoisomerases. J. Biol. Chem. **285**(25), 18967–18971 (2010). doi:10.1074/jbc.R109.092437
3. Dumont, S., Cheng, W., Serebrov, V., Beran, R.K., Tinoco, I., Pyle, A.M., Bustamante, C.: RNA translocation and unwinding mechanism of HCV NS3 helicase and its coordination by ATP. Nature **439**(7072), 105–108 (2006). doi:10.1038/nature04331
4. Hamdan, S.M., Johnson, D.E., Tanner, N.A., Lee, J.-B., Qimron, U., Tabor, S., van Oijen, A.M., Richardson, C.C.: Dynamic DNA helicase-DNA polymerase interactions assure processive replication fork movement. Mol. Cell **27**(4), 539–549 (2007). doi:10.1016/j.molcel.2007.06.020
5. Van Mameren, J., Peterman, E., Wuite, G.: See me, feel me: methods to concurrently visualize and manipulate single DNA molecules and associated proteins. Nucleic Acids Res. **36**, 4381–4389 (2008)
6. Haustein, E., Schwille, P.: Single-molecule spectroscopic methods. Curr. Opin. Struct. Biol. **14**(5), 531–540 (2004). doi:10.1016/j.sbi.2004.09.004
7. Kim, J.H., Larson, R.G.: Single-molecule analysis of 1D diffusion and transcription elongation of T7 RNA polymerase along individual stretched DNA molecules. Nucleic Acids Res. **35**(11), 3848–3858 (2007). doi:10.1093/nar/gkm332
8. Neuman, K.C., Nagy, A.: Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. Nat. Methods **5**(6), 491–505 (2008). doi:10.1038/nmeth.1218
9. Moffitt, J., Chemla, Y., Smith, S., Bustamante, C.: Recent advances in optical tweezers. Annu Rev Biochem **77**, 205–228 (2008)
10. Cluzel, P., Lebrun, A., Heller, C., Lavery, R., Viovy, J. L., Chatenay, D., Caron, F.: DNA: an extensible molecule. Science **271**(5250), 792–794 (1996)
11. Bustamante, C., Bryant, Z., Smith, S.B.: Ten years of tension: single-molecule DNA mechanics. Nature **421**(6921), 423–427 (2003). doi:10.1038/nature01405

12. Strick, T., Allemand, J., Croquette, V., Bensimon, D.: Twisting and stretching single DNA molecules. Prog. Biophys. Mol. Biol. **74**(1–2), 115–140 (2000)
13. Yamahata, C., Collard, D., Legrand, B., Takekawa, T., Kumemura, M., Hashiguchi, G., Fujita, H.: Silicon nanotweezers with subnanometer resolution for the micromanipulation of biomolecules. J. Microelectromech. Syst. **17**(3), 623–631 (2008)
14. Tang, W., Nguyen, T.-C.H., Howe, R.: Laterally driven resonant microstructures. Sens. Actuators **20**, 25–32 (1989)
15. Sun, Y., Fry, S.N., Potasek, D.P., Bell, D.J., Nelson, B. J.: Characterizing fruit fly flight behavior using a microforce sensor with a new comb-drive configuration. J Micromech. Syst. **14**, 4–11 (2005)
16. Senturia, S.: Microsystem Design. Kluwer, Boston (2000)
17. Washizu, M., Kurosawa, O.: Electrostatic manipulation of DNA in microfabricated structures. Ind. Appl. IEEE Trans. **26**(6), 1165–1172 (1990)
18. Roberts, R.J.: How restriction enzymes became the workhorses of molecular biology. Proc. Natl. Acad. Sci. U. S. A. **102**(17), 5905 (2005)
19. Kumemura, M., Collard, D., Yoshizawa, S., Fourmy, D., Lafitte, N., Jalabert, L., Takeuchi, S., Fujii, T., Fujita, H.: Direct bio-mechanical sensing of enzymatic reaction On DNA by silicon nanotweezers. In: Micro Electro Mechanical Systems (MEMS), 2010 I.E. 23rd International Conference on, Hong-Kong, pp. 915–918 (2010)
20. Salerno, D., Brogioli, D., Cassina, V., Turchi, D., Beretta, G.L., Seruggia, D., Ziano, R., Zunino, F., Mantegazza, F.: Magnetic tweezers measurements of the nanomechanical properties of DNA in the presence of drugs. Nucleic Acids Res. **38**(20), 7089–7099 (2010). doi:10.1093/nar/gkq597

DNA Origami as Programmable Nanofabrication Tools

Haorong Chen, Feiran Li, Jing Pan, Jungwook Choi and Jong Hyun Choi
School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Synonyms

Structural DNA nanotechnology

Definition

DNA origami is a self-assembly technique, where a long single-stranded DNA (scaffold), usually from the genome of a bacteriophage, is folded into a designed path with the help of a set of synthetic oligonucleotides (staples). This bottom-up nanofabrication approach is capable of generating large numbers of identical DNA structures with designed geometry. The structures can be further functionalized or decorated to produce versatile nanostructures.

Introduction

Apart from being the genetic material for various forms of life, DNA has also emerged as a promising engineering material for nanotechnology. Based on its excellent ability to recognize its complementary sequence, the binding of DNA can be programmed by sequence design. This idea gave birth to the field of structural DNA nanotechnology. Carefully designed oligonucleotides have been demonstrated to assemble into polyhedrons [1], 2D lattice [2], and 3D crystals that facilitate X-ray crystallography [3]. The DNA origami method, first introduced in 2006 [4], marked another important development in structural DNA nanotechnology. This method significantly increased the available complexity of self-assembled DNA structures because of the increased number of synthetic oligonucleotides involved. Importantly, the method also relaxed the rather stringent requirement for stoichiometry and oligonucleotide purification [4], and thus, it simplified the procedure and made custom DNA structures more accessible to the science and engineering community.

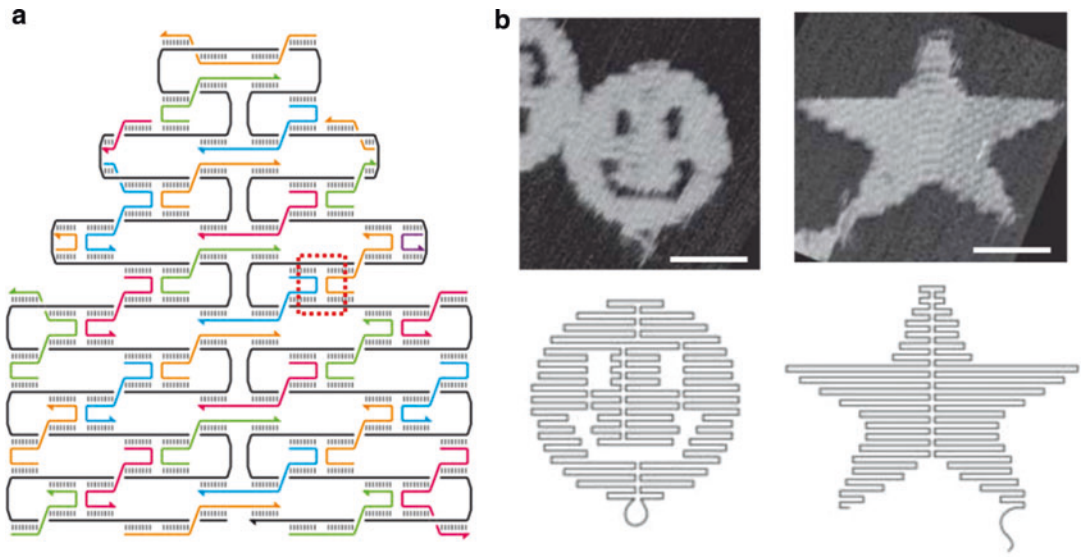
DNA origami technique uses double-stranded B-form DNA [4] as building blocks to construct a programmed structure through self-assembly. A rather complex geometry of hundreds of double-stranded structural units can be programmed by designing the sequence of hundreds of staples. An eight-base-long

double-stranded basic unit is only about 2 nm in diameter and 3 nm long. Such small size means that features as small as a few nanometer can be generated, which is not easy even for advanced lithographic techniques. The ability of DNA origami technique to controllably generate complex structures with nanometer accuracy has made it a valuable nanofabrication tool for research purposes. The parallel, bottom-up nature of the method also makes large-scale production possible. This essay focuses on the fabrication of various nanostructures using DNA origami. The fabrication strategies are described along with the unique properties and applications of the resulting structures.

Design Strategies for DNA Origami of Versatile Geometric Shapes

The synthesis/fabrication of DNA origami typically involves mixing the scaffold with the appropriate excess of a designed staple set and thermal annealing of the mixture in one pot. During annealing, the mixture is briefly heated to 75–95 °C to disrupt any secondary structure and then slowly cooled down to let complementary domains hybridize in a thermal cycler [4]. It is also possible to substitute the thermal annealing process with chemical annealing by controlling the concentration of denaturant such as formamide [5]. Since the synthesis procedure is rather established, the development of new DNA origami structure is usually centered on designing the folding path and staple sequences. As with all other design strategies, one should first choose the scaffold to be used because the length of the scaffold will determine the maximum size (area) of the final structure. The single-stranded genome of m13mp18 phage is the most widely used scaffold, while other DNAs of known sequence (such as PhiX174 [6] and lambda DNA fragment [7]) can be used as well. Two established design strategies for 2D and 3D structures will be discussed below.

DNA origami of arbitrary 2D shapes can be designed with an established procedure. As illustrated in Fig. 1a, the folding path of the long scaffold (colored in black) is first laid out in a raster fill pattern to approximate the target geometry. Then, the staples are added to the design diagram to connect adjacent rows of the scaffold by forming crossovers (circled in dotted red).



DNA Origami as Programmable Nanofabrication Tools, Fig. 1 Design schematics of DNA origami. (**a**) Scaffold and staple layout for a 2D DNA origami design based on raster fill. A *red* dotted box circles one crossover that interconnects two double helices. (**b**) AFM images of 2D DNA origami with complex geometry, along with the corresponding folding paths [4]. Scale bar: 50 nm

The positioning of the staples should follow certain rules to minimize the strain of the resulting structure by allowing each DNA double helix to make integers of full turns between crossovers. A common practice is to place staples into a staggered periodic array where the crossovers in the same row are 32 bases apart from each other [8, 9]. Finally, the sequence of each staple is determined based on the complementary scaffold domain that it is designed to bind. A software package has been offered to generate a sequence design from a bitmap graph of the target shape [10]. 2D DNA origami has demonstrated its usefulness in 2D self-assembly and nanolithography as will be discussed in later sections.

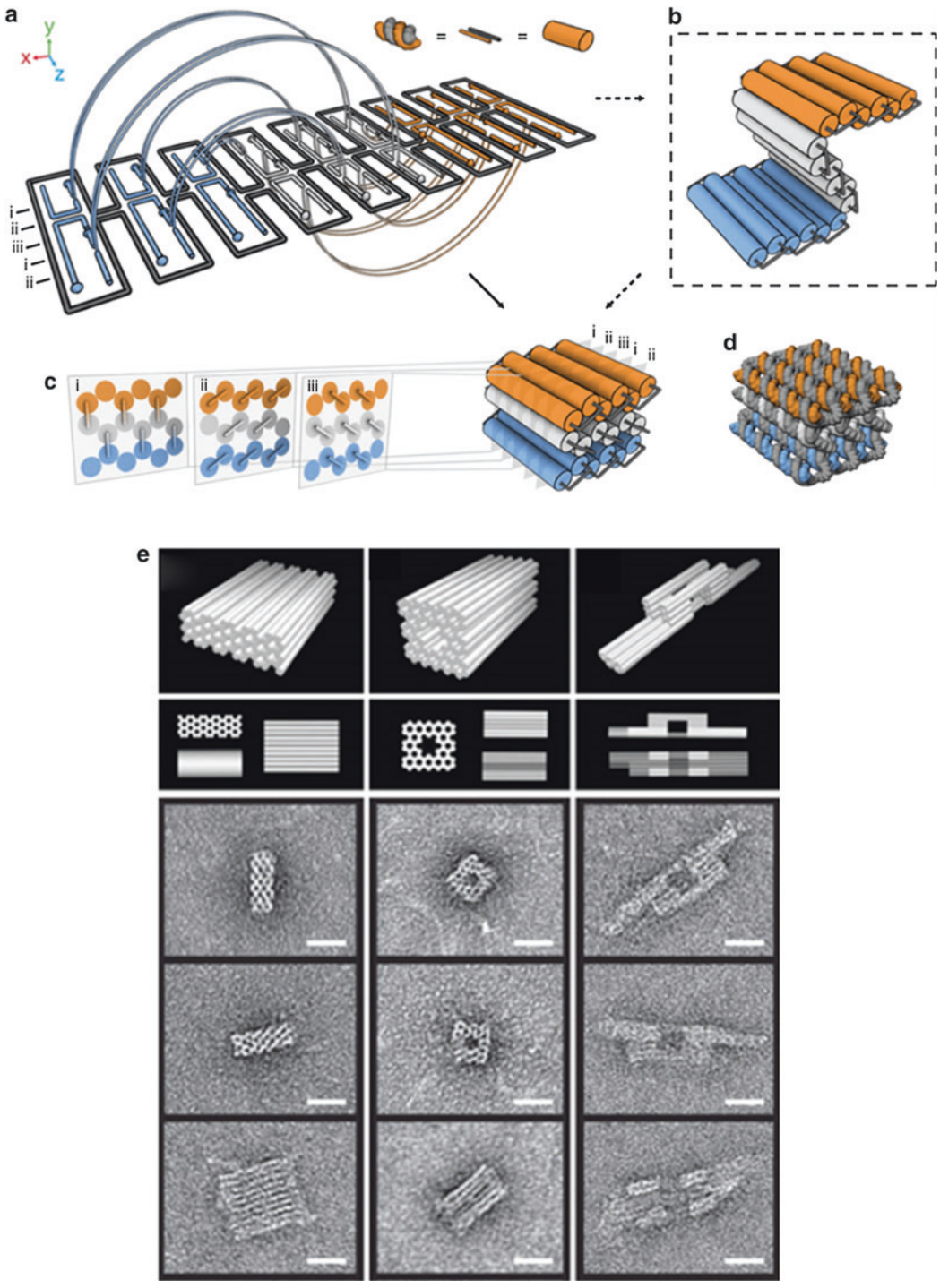
By extending the folding path of a scaffold out of plane, 3D DNA origami can be designed. Figure 2a–d illustrates a standardized strategy for folding a scaffold in 3D. Similar to the 2D origami design strategy, the scaffold only travels along one direction (Fig. 2a). Therefore, when the scaffold is hybridized with staples, the resulting DNA double helices will all be parallel to each other. The staples are placed to connect the double helices into a hexagonal array as shown in Fig. 2c. The staples form three sets of crossovers that connect every helix with three adjacent neighbors and stabilize the hexagonal array [11]. Based on this framework, various designs such as shown in Fig. 2e can be generated by specifying the length and starting position of each double helix in the array. For efficiently implementing this design strategy and preventing errors, an open-source computer-aided design tool, caDNAno [11], is available.

Besides the standardized design strategies, there are also specialized strategies for designing special structures such as wire frame-like tensegrity [12] or gridiron [13] structures. New structures may also be created by modifying the standard designs. For example, curvature and twist can be generated by deliberately introducing strain in a 3D structure. By controlled inserting and deleting of bases between crossovers in a minimally strained standard design, the new structure will be strained and change to a new conformation in response [14]. In summary, the design of DNA origami is rather flexible. While standard procedures lead to fast and low-error designs, there is also an ample room for creative twists and novel structures. The ever increasing library of available geometry illustrates the versatility of the DNA origami method.

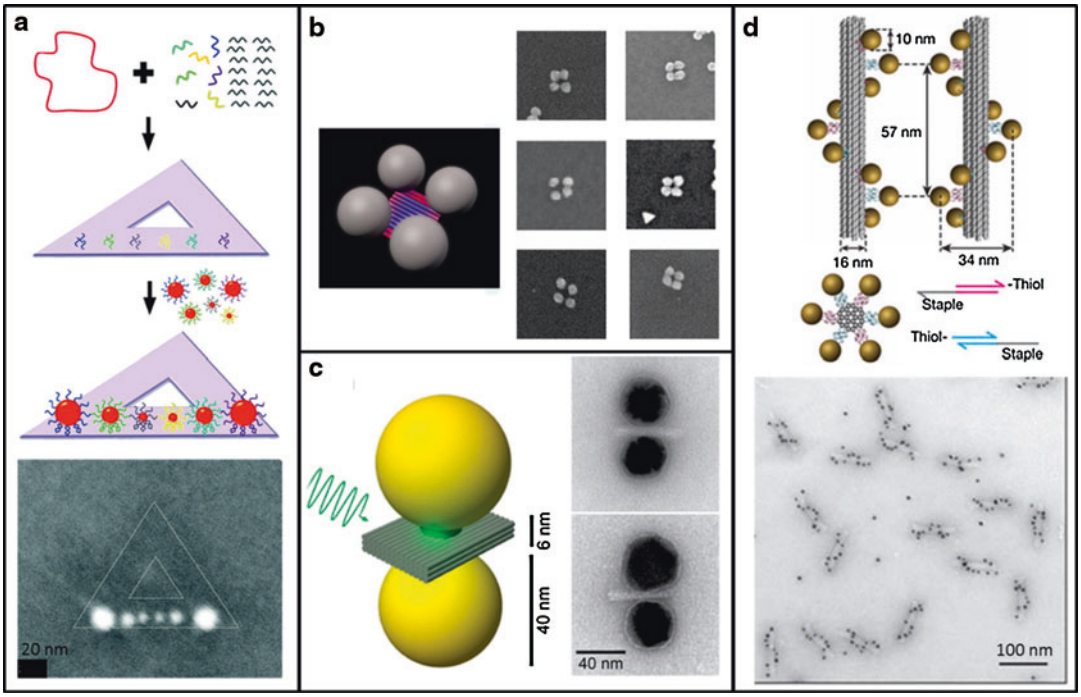
DNA Origami as Nanobreadboards

One of the most powerful abilities of DNA origami is to enable the organization of nanocomponents with a few nanometer accuracy [4]. As governed by sequence complementarity, each staple has a unique position in the assembled DNA origami structure. By functionalizing selected staples, modifications and decorations can be specifically added to predefined positions. The DNA origami template functions as a nanobreadboard where each staple is addressable as a decoration site for attaching nanocomponents including inorganic nanocrystals and functional biomolecules [4]. This approach is useful in inducing novel properties by controlling distance-dependent interactions between nanocomponents.

The accurate spatial control of DNA origami has been demonstrated to facilitate efficient construction of plasmonic structures. By controllably patterning gold or silver components in close proximity, the collective motion of the free electrons can yield novel optical effects such as large enhancement of electrical field in the near field [15] or even negative refractive index [16]. DNA origami has been used as an effective template in organizing gold nanoparticles (AuNPs) into plasmonic structures. The general assembly strategy is illustrated in Fig. 3a [17]. Selected staples are extended with sticky ends (unhybridized single-stranded domains), while the AuNP surface is conjugated with thiolated single-stranded DNA complementary to the sticky ends. The DNA origami structure formed after annealing the scaffold-staple mixture displays sticky ends on its surface, while excess free staples are removed via filtration [4, 18] or gel electrophoresis [11]. Finally, the DNA origami and AuNPs are incubated in an appropriate ratio to allow the sticky ends to hybridize with the DNA on AuNP surface, hence anchoring the AuNPs to the origami.



DNA Origami as Programmable Nanofabrication Tools, Fig. 2 (**a**) Folding path of the hexagonal double-helix array. For clarity, the double helices are simplified as two parallel strands or one rod. (**b**) Rod presentation showing how the structure is folded in 3D. (**c**) Cross-section views illustrating three sets of crossovers that connect the double helices. (**d**) Molecular model of the hexagonal double-helix array. (**e**) Schematics and TEM images of 3D DNA origami designs. Scale bar: 20 nm [11]



DNA Origami as Programmable Nanofabrication Tools, Fig. 3 Design schematic and TEM images of AuNP assemblies templated by DNA origami. (**a**) Heterogeneous AuNP chain on a triangular origami [17]. (**b**) SERS antenna consists of 4 AuNPs on four corners of a DNA origami tile [19]. (**c**) SERS antenna consists of 2 AuNPs on opposite side of a three-layered DNA origami [20]. (**d**) AuNP helices templated by columnar DNA origami [21]

To make the anchorage robust, three or more neighboring staples are usually extended with the same sticky ends so that the AuNPs can be anchored by multiple strands (i.e., multivalency). It is worth noting that one DNA origami structure can express multiple groups of distinct sticky ends. Correspondingly, multiple types of AuNPs (or different kinds of nanoparticles) may be respectively conjugated with corresponding thiolated DNA strands. The heterogeneous AuNP chain shown in the TEM image of Fig. 3a demonstrates the versatility of DNA origami-based AuNP assemblies.

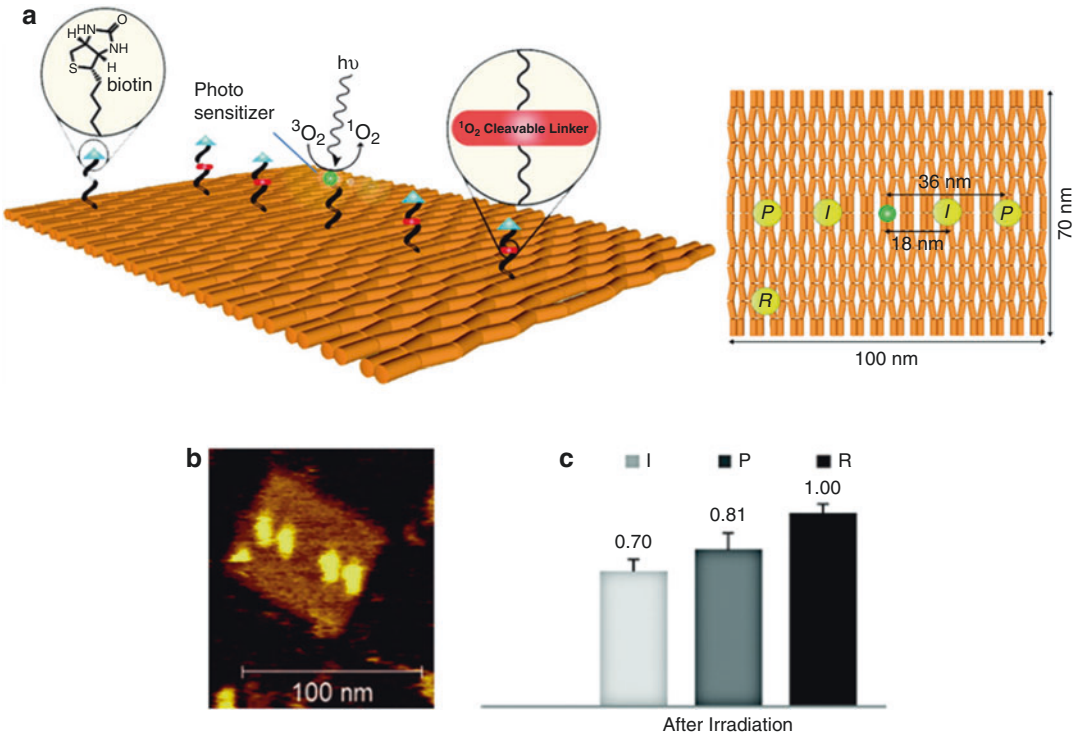
Figure 3b shows a plasmonic antenna fabricated by placing AuNPs on the four corners of a rectangular DNA origami [19]. The nanoparticles are then enlarged through electroless plating of silver. As the growing particles squeeze down the gaps between them, hotspots are created for surface-enhanced Raman scattering (SERS) measurements. The Raman fingerprint signals produced by four-particle assemblies are reported to be 100 times stronger than those of single particles. Figure 3c presents a different plasmonic antenna construction, where two gold nanoparticles are placed on opposite sides of a multilayered DNA origami [20]. The separation was accurately controlled by the thickness of the DNA template and can be tuned by changing the number of layers. When DNA-binding analytes were introduced into the hotspot, a SERS enhancement factor of 3.1×10^5 was observed. Figure 3d shows the arrangement of 9 AuNPs on each DNA origami rod into a helix, demonstrating DNA origami's capability in 3D nanofabrication. The helical arrangement generates chiral optical response near the plasmonic resonance wavelength in the visible range [21]. This plasmonic property can be drastically magnified by enlarging the particles through electroless plating. The resulting solution

produces strong optical rotary dispersion that can be detected by even naked eyes.

Semiconductor quantum dots (QDs) are another class of important optical nanocomponents that have been used to decorate DNA origami platforms. Given their excellent photo-properties such as high fluorescence quantum yield and long-term photostability [22], the ability of organizing QDs should increase their utility in various engineering applications. Liddle et al. have reported the assembly of QD and AuNP on the same DNA origami [23]. By controlling the distance between the particles and their relative number ratio, the fluorescence lifetime of the QD was engineered. The QD assembly onto DNA origami was first achieved via the high-affinity interaction between biotin and streptavidin [23]. To facilitate the assembly, a set of selected staples were replaced with biotinylated strands. After one-pot annealing, the resulting DNA origami had biotins exposed in designed positions, which conjugated with streptavidin coated on the QD surface [24], thereby anchoring the QDs on DNA origami. As with AuNPs, multivalent binding with several neighboring biotinylated staple strands makes the anchorage robust and increases the assembly yield. When compared to DNA base pairing on origami, however, the biotin-streptavidin interaction is not information rich and facilitates the assembly of homogeneous QDs only. Thus, DNA hybridization is a more powerful approach to graft various kinds of QDs on origami, although it requires special attention and extra efforts to conjugate DNA on QDs, because QDs are usually synthesized with small organic ligands, and the transfer to aqueous solution for DNA conjugation requires multistep ligand exchange and/or encapsulation processes. Such procedure often introduces unwanted sample heterogeneity and surface defects. To overcome these limitations, Choi et al. proposed an alternative approach of using DNA for the dual purposes of particle capping and targeting recognition [24]. This strategy has been further explored and is now generalized by several groups where phosphorothiolated, single-stranded chimera DNA is employed for dual purposes: the phosphorothiolated part is used as the capping domain, while the rest of the sequence binds to its complementary strand as the recognition domain [25, 26]. This scheme allows one to directly synthesize various kinds of QDs (e.g., CdS and PbS) that express distinct sequences (with size tunability and material composition variability) and assemble them on a DNA origami in a highly controlled manner.

Besides various nanoparticles, the conjugation with small molecules brings a variety of functionalities to DNA origami structures. By incorporating modified phosphoramidite building blocks during oligonucleotide synthesis, various chemical groups including biotin, thiol group, and fluorescent dyes may be incorporated to a staple oligonucleotide. The well-defined distance between the staples with functional moieties in a DNA origami structure can be exploited for sub-diffraction imaging calibration standard [27] or as research platforms for investigating distance-dependent molecular interactions. Figure 4a illustrates one of the latter examples where one photosensitizer and five biotinylated extensions are placed precisely on a 2D DNA origami [28]. The biotin moieties are conjugated with streptavidin to produce a height contrast for AFM imaging, and four of the biotin extensions have singlet-oxygen cleavable linkages. Upon irradiation, the photosensitizer at the center of the origami tile starts to generate singlet oxygen molecules which diffuse and cleave the linkage. With this setup, the cleavage efficiency was probed as a function of distance, and information about the diffusion length and lifetime of singlet oxygen was obtained.

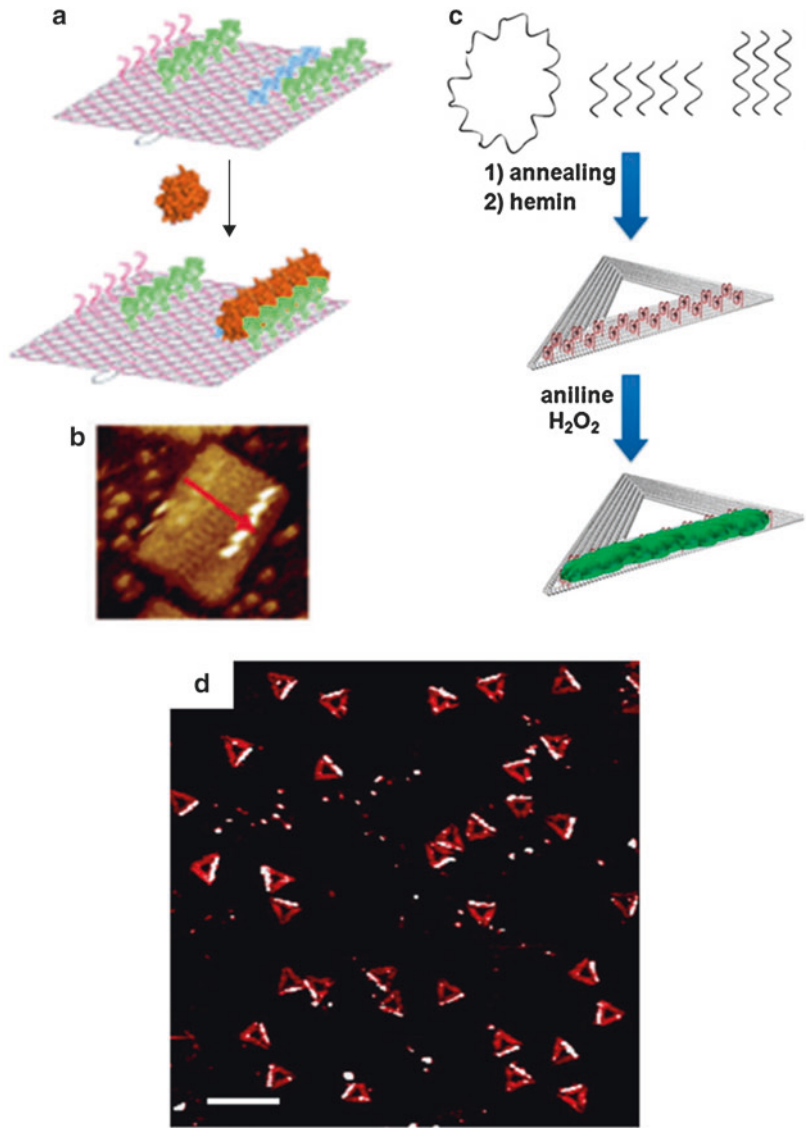
DNA origami may be used as a template to organize functional DNA strands such as aptamers, which are identified from in vitro selection for high binding affinity toward given target molecules (typically, small proteins and organic compounds) [29]. By expressing aptamer sequences on the surface, DNA origami can mimic the function of antibodies. This is shown to be a useful mechanism for molecular diagnostics and for organizing nanomaterials. Figure 5a illustrates a DNA origami with extensions of thrombin-binding aptamers [30] which capture thrombin, a serine protease involved in the blood clotting process. The capture event can be clearly



DNA Origami as Programmable Nanofabrication Tools, Fig. 4 DNA origami used for studying distance dependence of singlet oxygen cleavage. (**a**) Schematics that illustrate the placement of photosensitizers, cleavable linkers, and biotinylated staples on a 2D rectangular DNA origami. (**b**) AFM image of the DNA origami after conjugating streptavidin to biotinylated staples. (**c**) Percentage of uncleaved linkages at different distances [28]

observed with AFM (Fig. 5b). This example demonstrates that combined with AFM, DNA origami can serve as a platform for immunoassay at the single-molecule level. In Fig. 5c, triangular DNA origami tiles are extended with a 17-nucleotide hemin-binding aptamer, which contains a guanine (G)-rich sequence [31]. This aptamer forms a G-quadruplex secondary structure in the presence of potassium ions which facilitates recognition of hemin [32]. The hemin-containing extension was shown to behave as a peroxidase-mimicking enzyme and subsequently catalyze the polymerization of aniline through facilitating its oxidation into aniline radical by H_2O_2. Because the catalysis is limited in a short range, the polymer forms only in the vicinity of an aptamer extension. Therefore, by designing the spatial placement of aptamer extensions, polyaniline can be controllably formed in specified locations on DNA origami (Fig. 5d). Such spatial control of conjugated polymers holds a great promise for organic nanoelectronics.

Finally, the ability to arrange proteins on DNA origami implicates an access to enormously rich biochemical functions. With the maturation of the conjugation strategies, endless possibilities could unfold. For example, the assembly of proteins may result in more effective biochemical catalysis. Figure 6a illustrates the spatially controlled enzyme cascade reaction on a DNA origami. In this example, glucose oxidase and horseradish peroxidase were conjugated to thiolated oligonucleotides via *N*-[(ε-maleimidocaproyloxy) sulfosuccinimide ester] (sulfo-EMCS) bifunctional linkers. They were subsequently anchored onto the origami structure via DNA hybridization [8]. By assembling the two enzymes close to each other (Fig. 6b), the intermediate product was more effectively transferred between them, resulting in a significantly increased overall



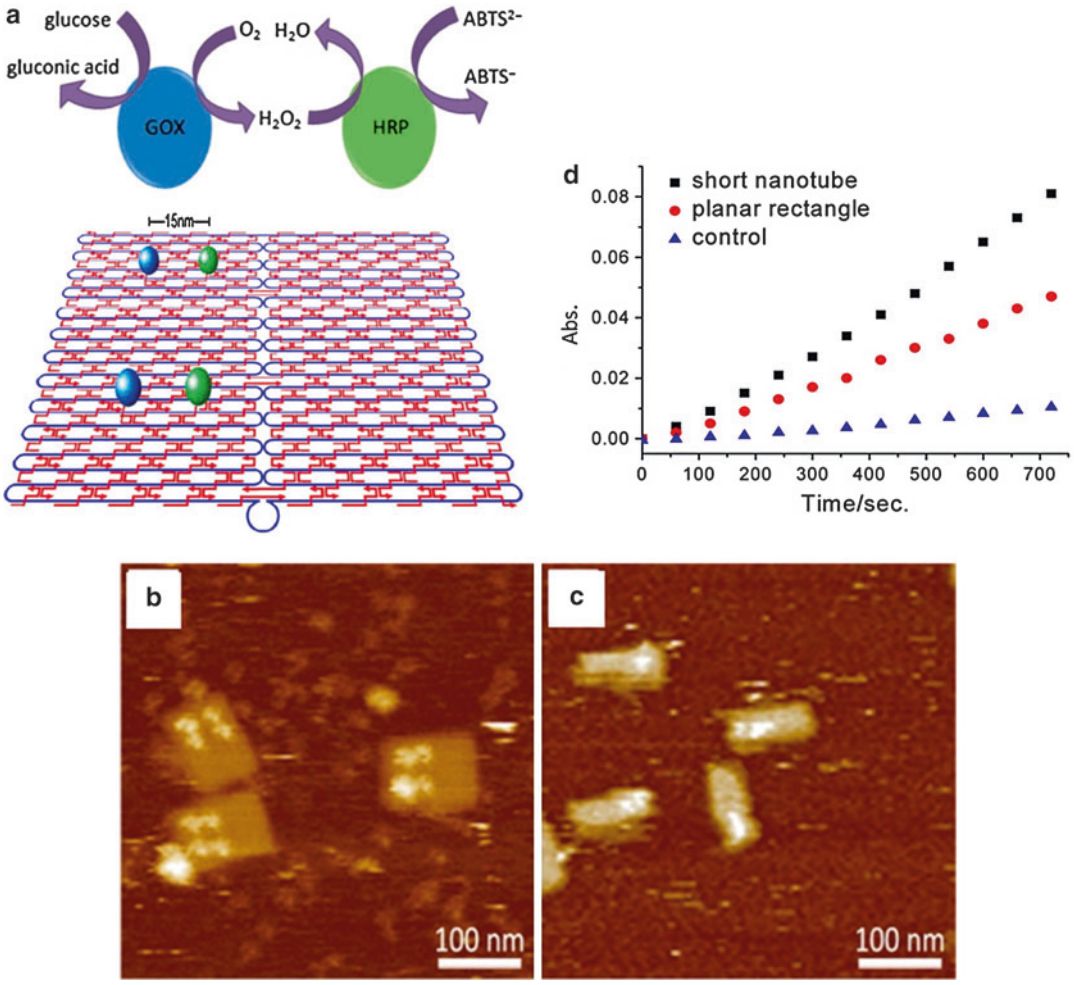
DNA Origami as Programmable Nanofabrication Tools, Fig. 5 (**a**) Schematics of 2D DNA origami structures functionalized with aptamer extensions. (**b**) AFM image of a row of thrombin captured on a DNA origami surface [30]. (**c**) Schematic of hemin-binding aptamers on DNA origami that recognize hemin and catalyze the polymerization of aniline at desired locations. (**d**) AFM image of triangular DNA origami with polyaniline deposited on one arm [31]. Scale bar: 200 nm

reaction rate. DNA origami was demonstrated to further increase the catalytic efficiency by forming a semi-closed tube around the enzyme cascade (Fig. 6c), which reduces dissipation of an intermediate product and creates a higher local concentration.

DNA Origami as Nanostencils for Fabricating Inorganic Materials

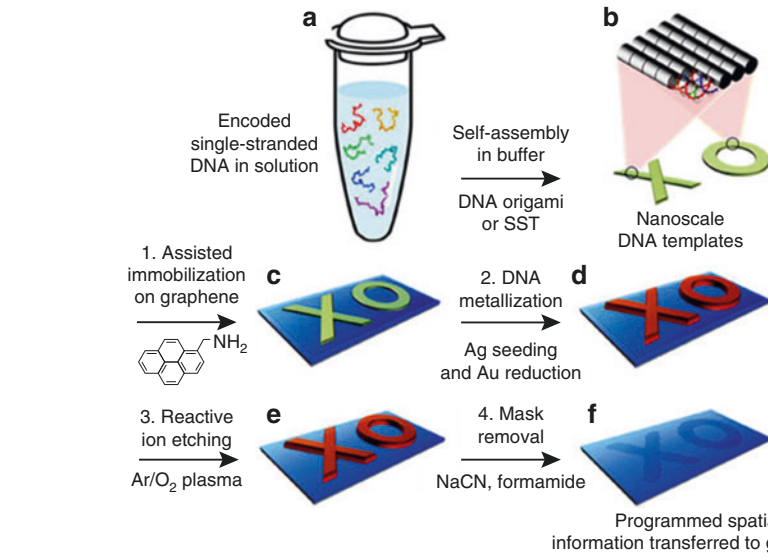
The ability of DNA origami to produce arbitrary 2D geometry with nanometer accuracy makes it a potential nanolithography tool. Compared with optical lithography, the small feature size of DNA origami enables it to achieve much better resolution. Given the nature of its synthesis and deposition, DNA origami is a scalable and parallel process that is faster in fabricating large quantities of nanostructures, which is advantageous over other high-resolution methods such as electron beam lithography (EBL) and focused ion beam (FIB) milling. Several successful demonstrations of nanopatterning with DNA origami have been reported [18, 33, 34].

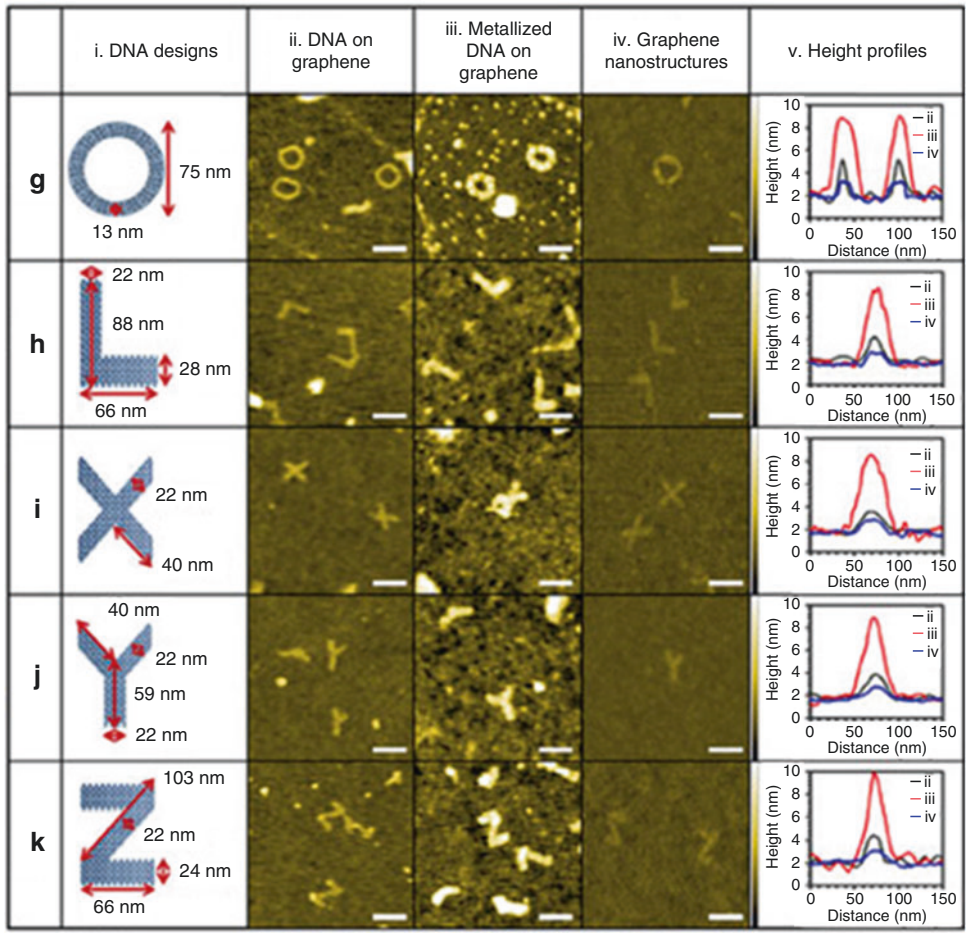
Figure 7 illustrates the process of transferring programmed DNA origami geometry to



DNA Origami as Programmable Nanofabrication Tools, Fig. 6 Conjugation of proteins on a DNA origami tile. (**a**) Schematic of glucose oxidase and horseradish peroxidase assembled into enzyme cascade on DNA origami. (**b**, **c**) AFM images of the enzyme cascade on flat DNA origami (**b**) and semi-closed DNA origami tubes (**c**). (**d**) Catalytic kinetics of the enzyme cascades compared to control [8]

underneath graphene [18]. Briefly, DNA origami templates are incubated with glutaraldehyde to adsorb the molecule onto the surface. Then the templates are immobilized onto graphene surface with the assistance of 1-pyrenemethylamine, which adsorbs onto graphene by π-π stacking and aids the adhesion of DNA origami. After deposition, the DNA template is metalized by a sequential treatment with silver nitrate and gold enhancement solution. The adsorbed glutaraldehyde reduces silver ions into metal seeds on DNA skeleton. Gold then deposits around the silver seeds, resulting in a fully metalized mask, which protects graphene during the subsequent Ar/O_2 etching. Finally, the metal mask and DNA residue are removed by $NaCN/H_2O$ and deionized formaldehyde, leaving only the patterned graphene. This example illustrates two key challenges for pattern transfer from DNA template: the effective attachment of the template and the strengthening of that template. Atomically thin materials such as graphene are probably the best suited for DNA origami-based molecular lithography because they need only a limited exposure to plasma etching and do not require thick masks which generally have a lower resolution.





DNA Origami as Programmable Nanofabrication Tools, Fig. 7 (**a**–**f**) Step-by-step illustration of the pattern transfer process. (**g**–**k**) Schematics, AFM images, and height profiles at different processing stages. Scale bars: 100 nm [18]

Oxides from chemical vapor deposition (CVD) can be a good alternative to metal in preparing the etching masks [33]. Compared with metalized masks, which show significant grains and extra particles (Fig. 7g–k), CVD oxides are potentially smoother with less uncontrolled nucleation. For CVD-based oxide growth, a substrate with deposited DNA origami is placed in a glass desiccator with two open vials of precursor and ammonia. Tetraethyl orthosilicate (TEOS) and titanium isopropoxide are the respective precursors that hydrolyze to yield SiO_2 and TiO_2. The hydrolysis reaction is catalyzed by water. Therefore, the different amount of adsorbed water results in different rates of oxide growth. Overall, CVD can convert a DNA template into a hard mask from vapor phase and circumvent several possible issues related with liquid-phase methods, such as fouling and DNA desorption.

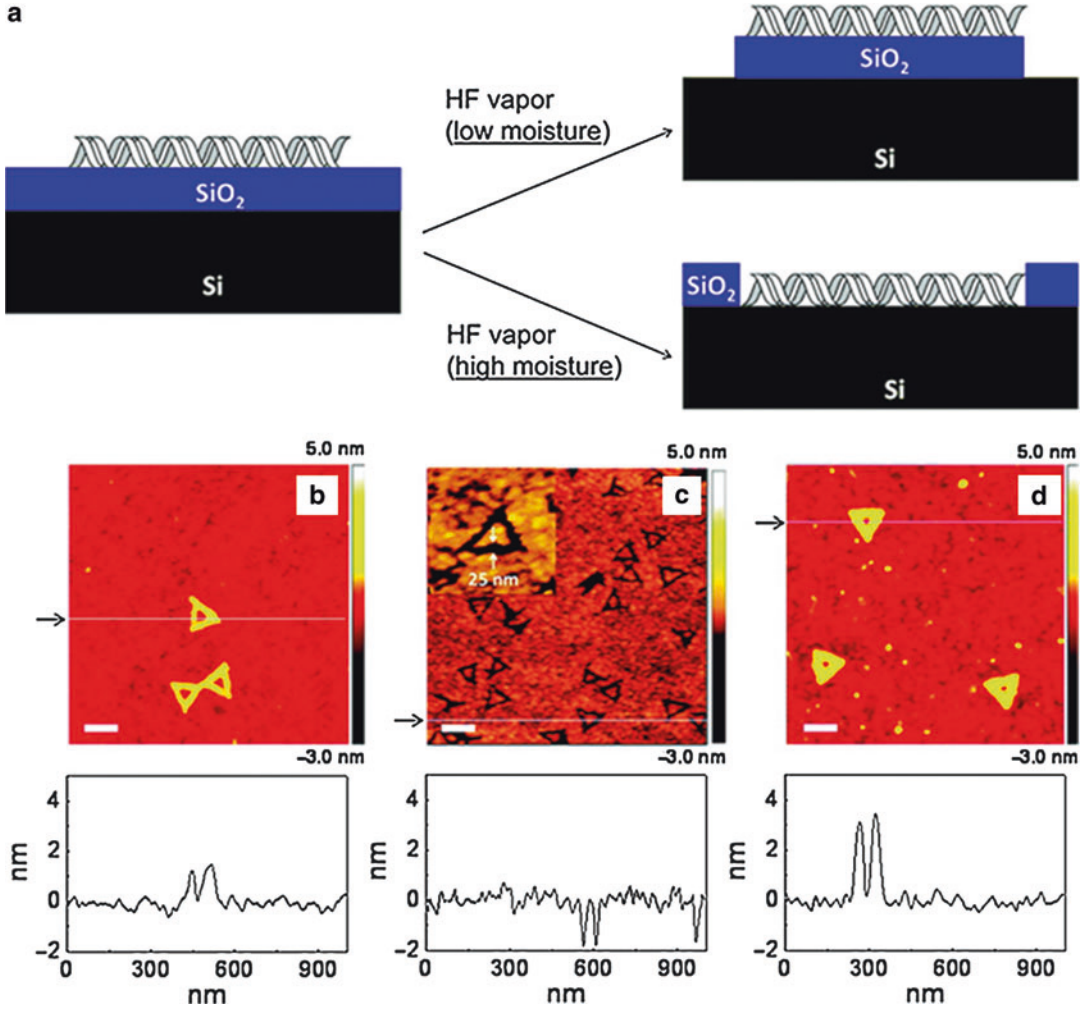
DNA origami also has been reported to directly template the etching of SiO_2 substrates without using any hard masks [34]. Similar to the oxide precursors, etching by vapor phase HF needs water as a catalyst. Therefore, the differential adsorption of water in various areas produces a contrast mechanism (Fig. 8a). Compared to DNA, SiO_2 surface has higher water adsorption at low relative humidity, but lower water adsorption at high relative humidity. Therefore, etch rate in DNA region is slower at low humidity and faster at high humidity. Figure 8a illustrates that both positive and negative tone patterns can be produced, depending on the relative humidity. In either scenario, a height contrast of about 2 nm has been demonstrated (Fig. 8c, d). This strategy may be applied for etching thin oxides, such as those grown by atomic layer deposition (ALD).

The fragile nature of DNA generally does not allow it to directly function as an etch mask. Novel contrast mechanisms such as presented above are needed for effectively transferring programmed DNA geometry to other materials. DNA origami has been demonstrated to template the deposition of metals and oxides. However, more diverse chemical strategies should be established to enable DNA-templated growth of a wider range of materials. Then, DNA origami structures may either provide functionality on their own or serve as hard masks for etching other functional materials. The graininess and fidelity of the processes also need to be optimized in order to fully exploit the high spatial resolution of DNA origami.

Scaling Up DNA Origami Structures

Most DNA origami designs use the 7 kb m13 genome and about 200 staples. As a result, a 2D structure typically covers an area less than 100 × 100 nm [35]. While DNA origami represents a breakthrough in the complexity and size for molecular self-assembly, further scaling up is sought after for organizing nanocomponents into more sophisticated systems and better integration with top-down fabrication methods. One possible route for scaling up is through the use of larger scaffolds. However, the increased number of staples for folding such a scaffold also means increased similarities between binding domains and increased misfold rates. In addition, very long single strands are fragile and may result in more fragmented partial structures. An alternative strategy is to hierarchically assemble preformed DNA origami into larger and more complex structures, as discussed below.

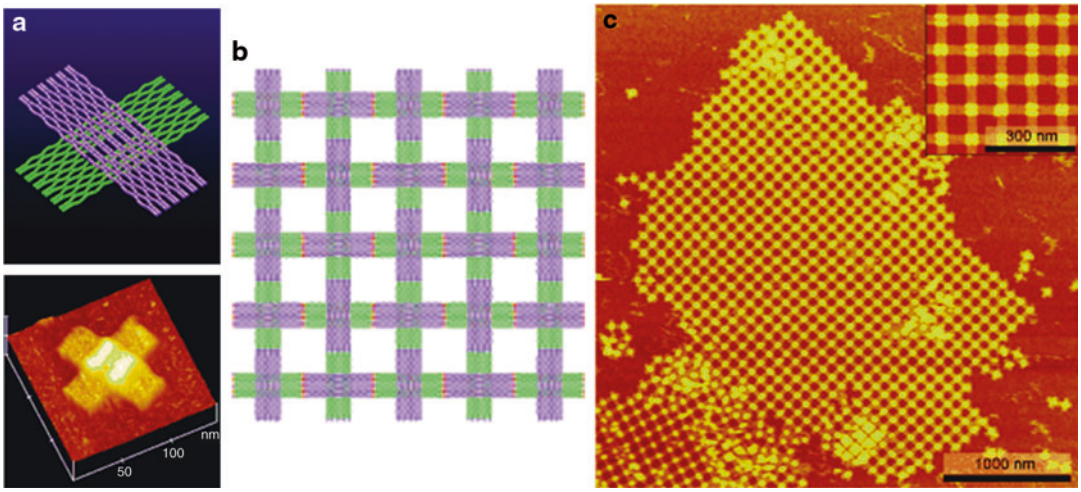
A hierarchical assembly strategy has been shown to create microscale structures from preformed DNA origami monomers. Figure 9 shows a 2D array of interconnected, cross-shaped DNA origami. Because DNA origami are observed to connect more effectively in the direction parallel to DNA helices [36], the cross-shaped DNA origami was designed in order to let the 2D array propagate effectively in both directions. As illustrated in Fig. 9a, the cross consists of two rectangular subunits connected perpendicularly, and all of the four edges are perpendicular to the DNA strands to facilitate effective linking. Eight sticky ends extend from each edge. Their programmed hybridization to the sticky ends on another edge will connect two DNA origami structures in an exactly predesigned fashion. Figure 9c presents the resulting crystalline 2D array which exceeds 2 × 2 μm. Such size adequately matches the feature lengths of



DNA Origami as Programmable Nanofabrication Tools, Fig. 8 (**a**) Schematic of the two etching regimes that result in positive and negative tones. (**b**–**d**) AFM images of DNA origami and the corresponding etched patterns of negative/positive tone. Depending on the moisture condition, the presence of DNA either promotes etching of adjacent silicon oxide (**c**) or hinders the etching (**d**). Height profiles are presented below the corresponding AFM images. Scale bar: 100 nm [34]

photolithography. However, the assembly mechanism does not actively limit the expansion of the 2D array. As a result, its overall size and shape is not well controlled. Moreover, DNA origami units in this periodical array cannot be distinguished from each other; hence, they cannot be addressed individually.

An "origami of origami" strategy has been proposed to produce well-defined DNA origami assemblies, where each origami unit can be addressed. Figure 10a illustrates this strategy: two kinds of scaffolds are utilized, one for constructing the normal DNA origami units and another for constructing a loose framework [6]. The normal origami tiles have specially designed sticky ends for directing them to a predetermined location in the framework. Collectively, the framework and the unit tiles form a much larger structure of a predesigned geometry. It is worth noting that each tile in the assembly is unique because of their unique set of sticky ends. This allows every point in the assembly to be individually addressed;

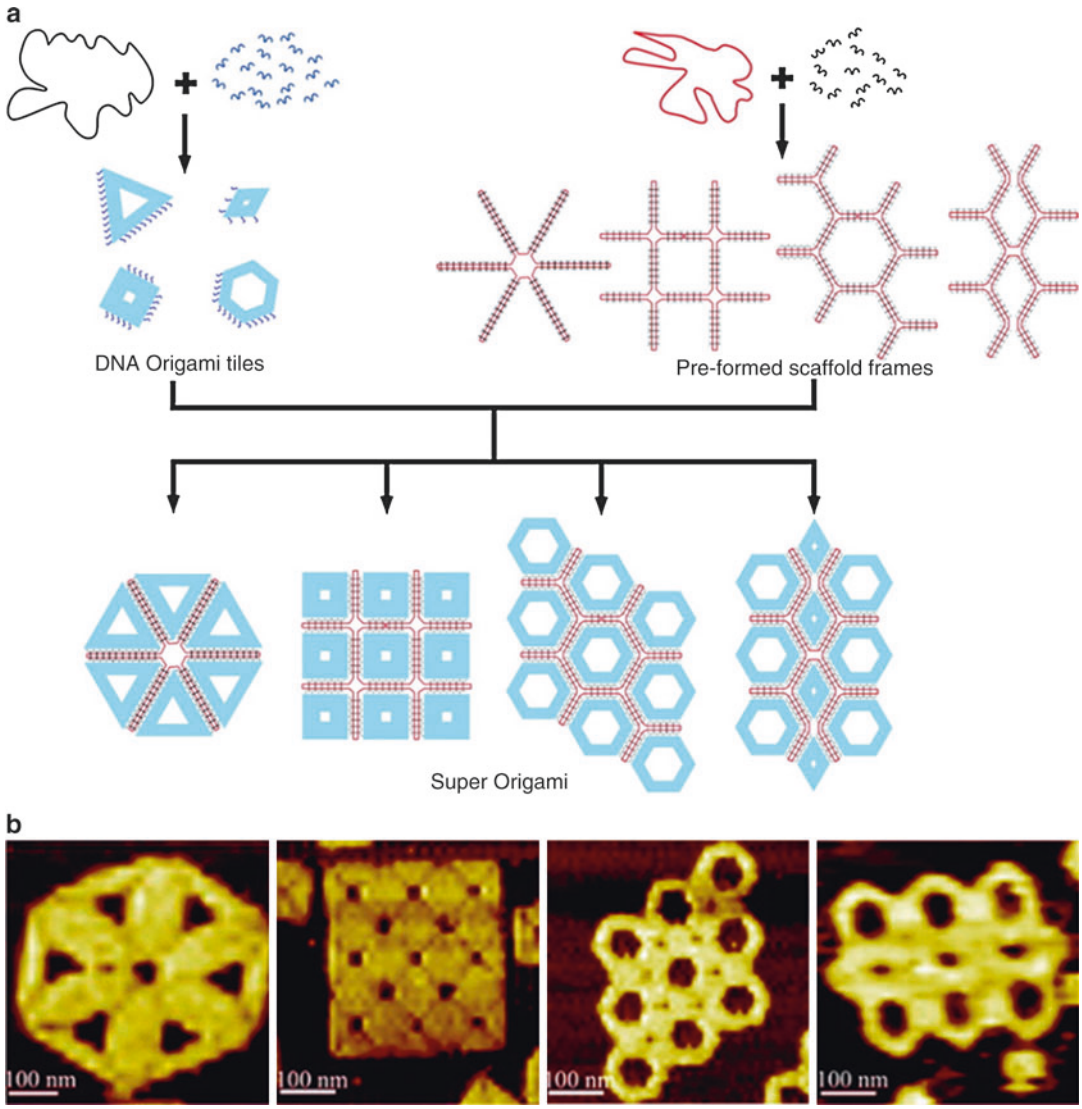


DNA Origami as Programmable Nanofabrication Tools, Fig. 9 (**a**) Schematic and AFM image of individual cross-shaped DNA origami. (**b**) Schematic for connecting individual cross origami into a crystalline 2D array. (**c**) AFM image of a 2D array about 2 × 3 μm in size [36]

to decorate a point, one only needs to functionalize the point on the corresponding DNA origami tile. After assembling the tiles, the functionalization will be automatically carried to the predetermined position in the superstructure. With this strategy, about ten DNA origami tiles can be grouped together, and therefore, an approximately ten-time scale-up has been achieved without compromising addressability.

In the hierarchical assembly strategy, reversible linkage between DNA origami units provides the larger systems with dynamic properties. Figure 11 presents such an example that uses azobenzene, a chemical moiety that undergoes reversible photo-isomerization in response to illumination of different wavelength [37]. Azobenzene can be incorporated into DNA to facilitate photo-controlled hybridization: planar *trans* isomer stacks with the nucleobases and stabilizes the double helix, whereas out-of-plane *cis* isomer destabilizes the duplex due to steric hindrance (Fig. 11a) [37]. By switching azobenzene moiety from *trans* to *cis* isomers, base pairing can be disrupted, while switching from *cis* to *trans* form promotes the hybridization (Fig. 11b). Therefore, by extending DNA origami with azobenzene-functionalized sticky ends, the linkage between DNA origami can be controlled by light of different wavelengths [38]. Figure 11d demonstrates the dynamic switch of DNA origami between monomer and dimer forms within the time frame of a few minutes. Such remote control capability should prove a valuable tool for regulating various processes at the nanoscale.

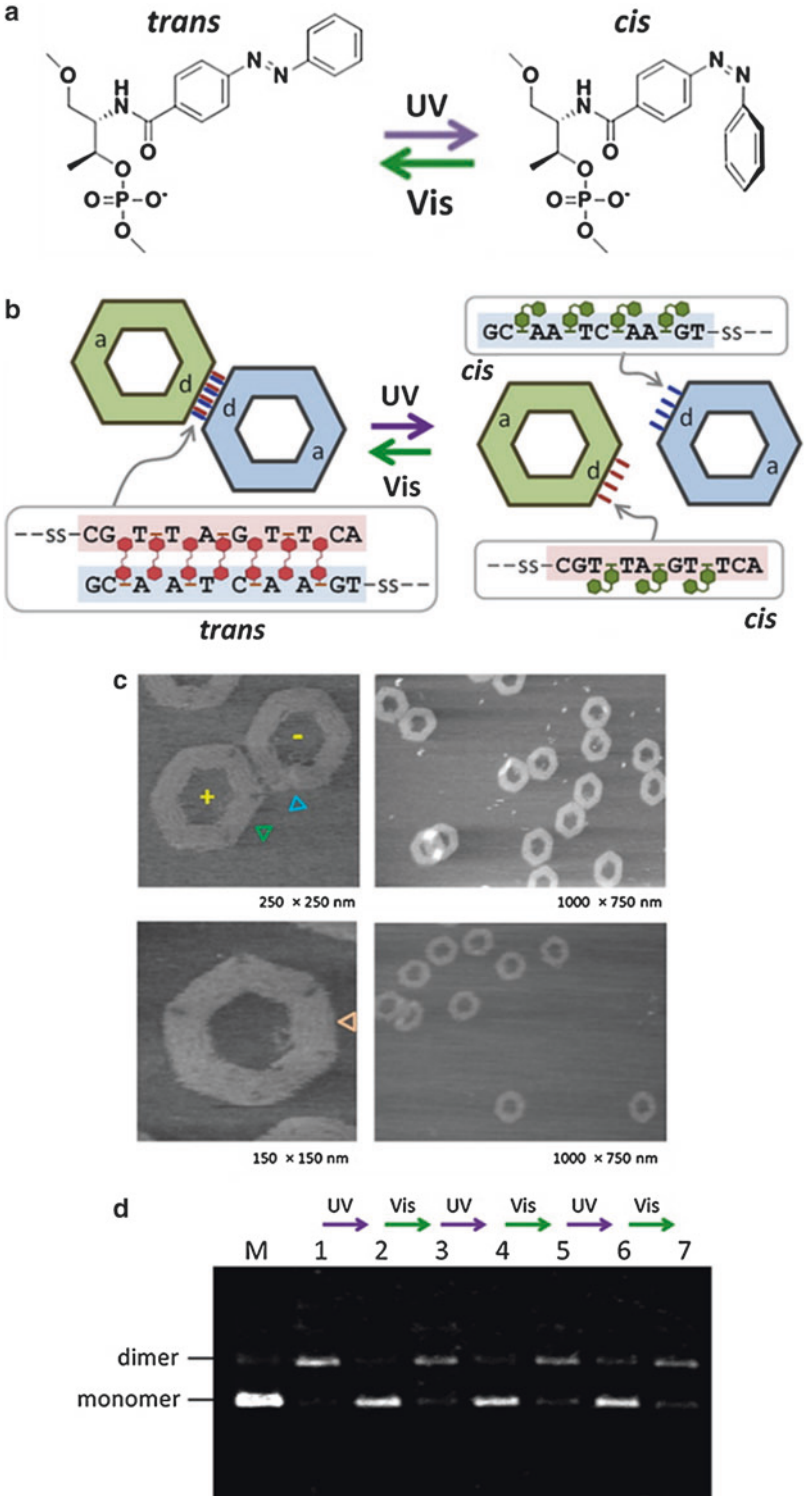
Reversible linkage during the hierarchical assembly can also be realized by DNA hybridization. Compared with azobenzene that has only two available conformations, DNA is information rich and allows the assembled structure to switch between multiple linking configurations. Figure 12 shows such a versatile strategy. Short cylindrical DNA origami monomers are designed to have unhybridized binding domains on both upper and lower edges (Fig. 12a). The linkers that simultaneously hybridize to binding sites on both edges stack the short cylinders into an elongated tube. Different linker sets that pair up the binding sites differently cause a variation in the relative rotation between two connected DNA origami units and thus result in tubules of chosen chiral angles [9]. Moreover, the linkers may also be extended with unhybridized domains (the so-called toeholds) in order to allow disassembly via toehold-mediated strand displacement [39]. For disassembly, a set of single strands (i.e., release



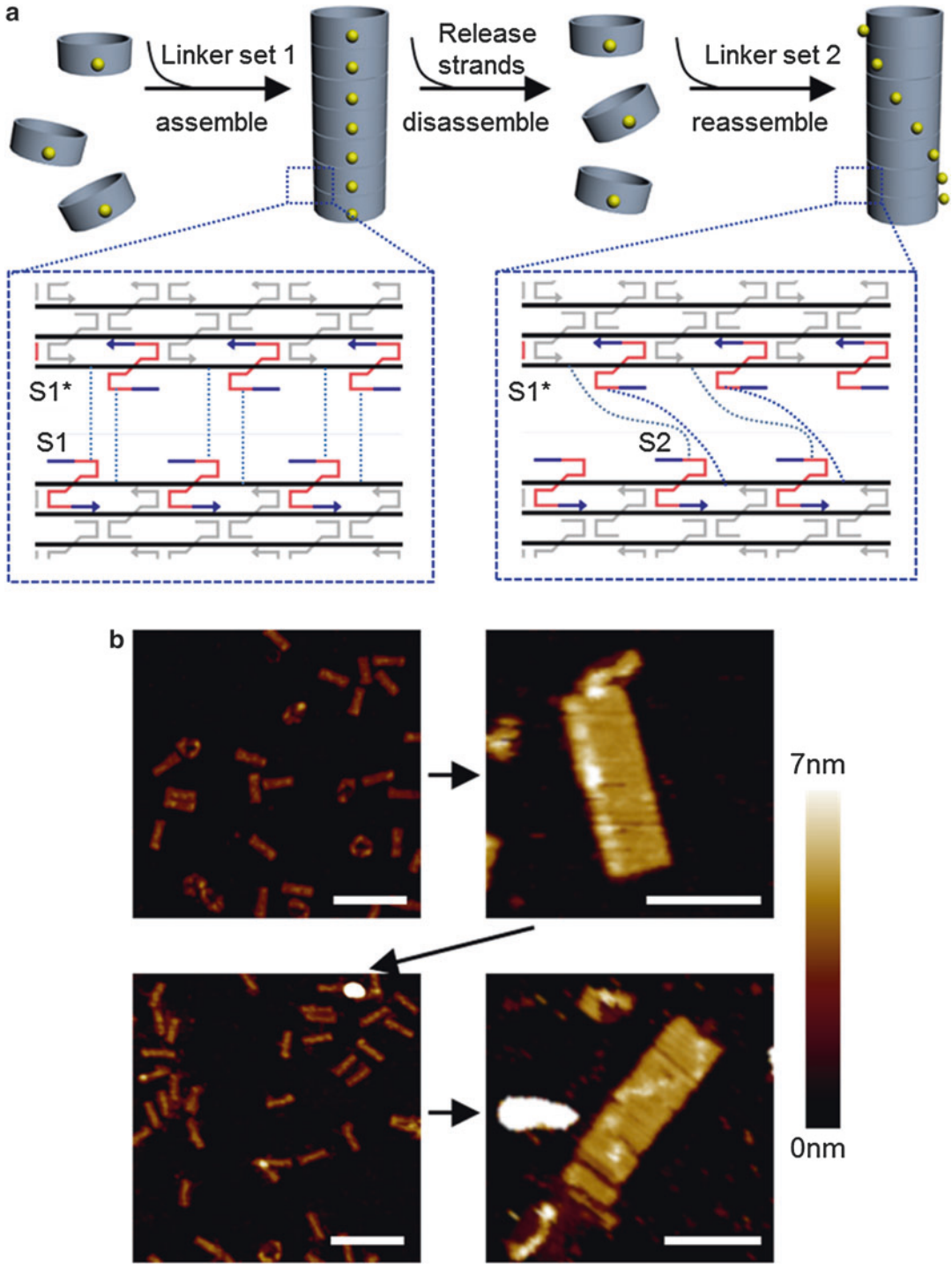
DNA Origami as Programmable Nanofabrication Tools, Fig. 10 (**a**) Schematic of the process for creating "origami of origami" structures. (**b**) AFM images of the assembled super origami structures corresponding to respective designs [6]

strands) are introduced to the solution. The release strands are exactly complementary to the linkers. They first hybridize to the toehold domain of the linkers and then bind more and more bases of the linkers via random branch migration [39]. Because of the increased number of base pairs, the release strands eventually hybridize with the linkers completely and disengage them from linking the two neighboring origami units together, thereby releasing the linkages and causing disassembly. The disengaged linkers leave unhybridized binding sites intact which are available for another round of assembly. In Fig. 12b, the hollow ring-shaped DNA origami units are assembled into a nonchiral tubule, disassembled, and then reassembled into a chiral tubule (the chosen helical angle of ~45°), demonstrating the dynamic nature of the assembly strategy.

The hierarchical assembly strategy has been demonstrated to produce structures that exceed



DNA Origami as Programmable Nanofabrication Tools, Fig. 11 (**a**) Schematic of the photoisomerization of azobenzene moiety. (**b**) Scheme of photo-regulated DNA origami dimerization. (**c**) AFM images of DNA origami dimers and monomers. (**d**) Agarose gel electrophoresis image demonstrating reversible photo-induced assembly and disassembly [38]



DNA Origami as Programmable Nanofabrication Tools, Fig. 12 (**a**) Illustration of the mechanism for reversibly linking DNA origami monomers into tubules of designed chirality. *Black* and *gray*: scaffold and staples. *Red* and *blue*: linking domains and toehold domains of the linkers. (**b**) AFM images demonstrating the reversible assembly and disassembly of DNA origami tubules. *White* dot patterns created by hairpin staples demonstrate the chirality of the tubules. Scale bar: 200 nm [39]

the size limit of a single DNA origami while retaining the spatial accuracy. The increased size and complexity combined with reversible and photo-responsive assembly mechanisms should open up new possibilities of sophisticated nanosystems.

DNA Origami for Constructing Dynamic and Tunable Structures

DNA is much more compliant compared to many other materials commonly used in nanotechnology such as silicon and gold – this is especially the case for single-stranded DNA. As a result, DNA-based structures can be designed to significantly change their configuration in response to environmental cues, similar to many conformation-changing proteins. This unique property opens up broad possibilities for dynamic structures and novel functions.

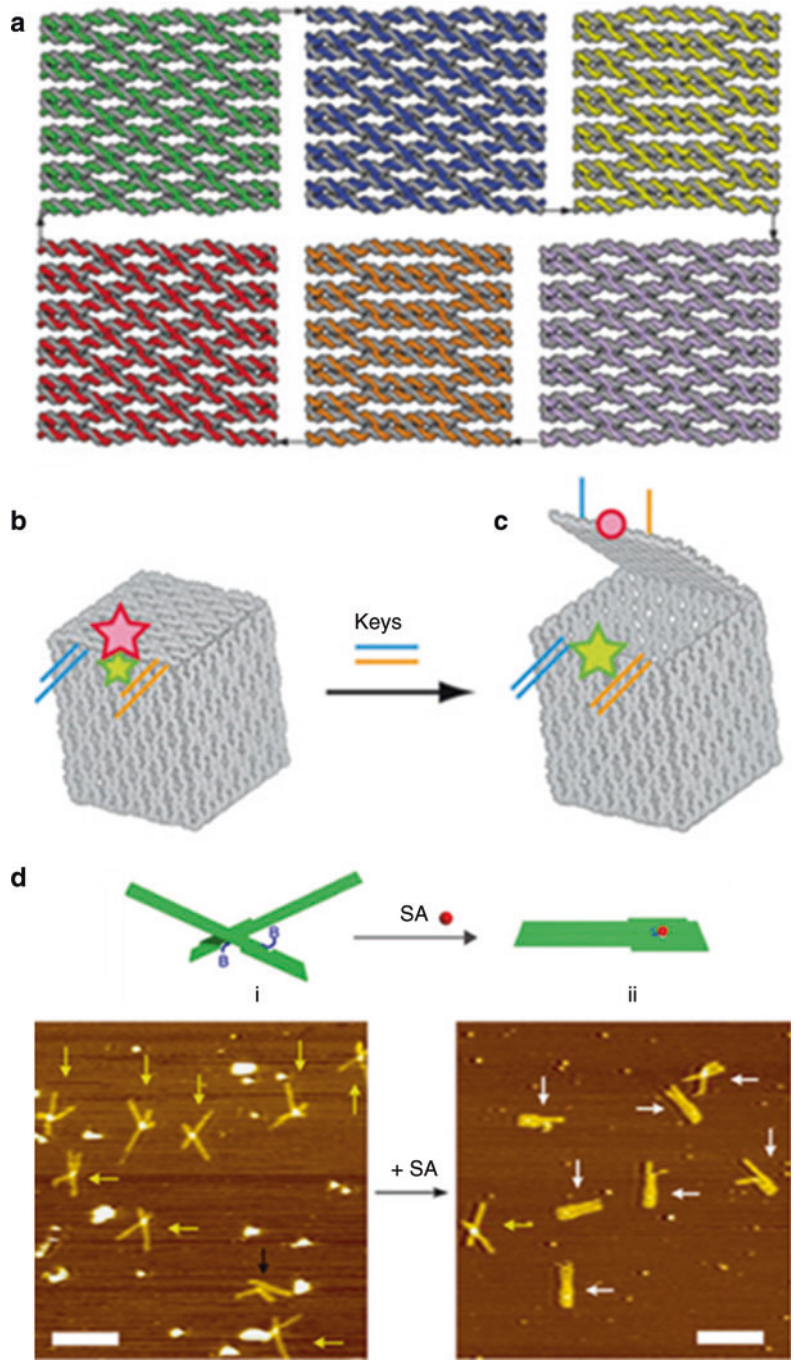
Highly compliant single-stranded DNA has been used as free hinges for creating dynamic structures. Figure 13a, b illustrates a DNA origami box that uses such mechanism [40]. By design, the scaffold folds into a box formed by six interconnected rectangles. One top side of the box (the lid) is hinged by several staples with a four-base-long single-stranded flexible domain. The lid is locked closed by two pairs of partially hybridized extensions (Fig. 13b). A "key" DNA can disrupt the hybridization through toehold-mediated strand displacement. When the two pairs of locking extensions are both disrupted, the lid is free to open (Fig. 13c). Douglas et al. have further demonstrated that a similar structure can open in response to other chemical analytes such as platelet-derived growth factor (PDGF) if the corresponding aptamer sequences are used as locking strands [41]. As an application, drugs may be loaded inside the structure, which are accessible from the outside only when it is opened by target analytes. Such dynamic structure has been shown to sense environmental signals and expose payloads accordingly. This responsive behavior makes dynamic DNA structures promising drug delivery carriers.

Figure 13d illustrates a molecular forceps that pinches its target analyte [42]. This DNA origami structure consists of two single strands connecting two rigid arms. Both arms contain a chemical group for recognizing a molecular target. The forceps normally assumes the most relaxed open conformation. Upon addition of a target, the recognition groups on both arms bind to the target and cause the structure to pinch close. This conformation change can be clearly visualized with AFM and offers a target-sensing mechanism with high sensitivity and selectivity.

While single-stranded DNA behaves much like free hinges, double-stranded DNA can be used to construct springlike connections that are able to generate considerable restoring forces. The structure shown in Fig. 14a has a thin (one corrugated layer) double-stranded joint [43], and another single-stranded connection is placed away from the joint. The natural tendency for single-stranded DNA to coil up produces the tensile force that bends the joint. The tensile force may be tuned by changing the length of the single strand. Figure 14b shows that the bending angle responds to the changing force, clearly demonstrating the elastic nature of double-stranded DNA.

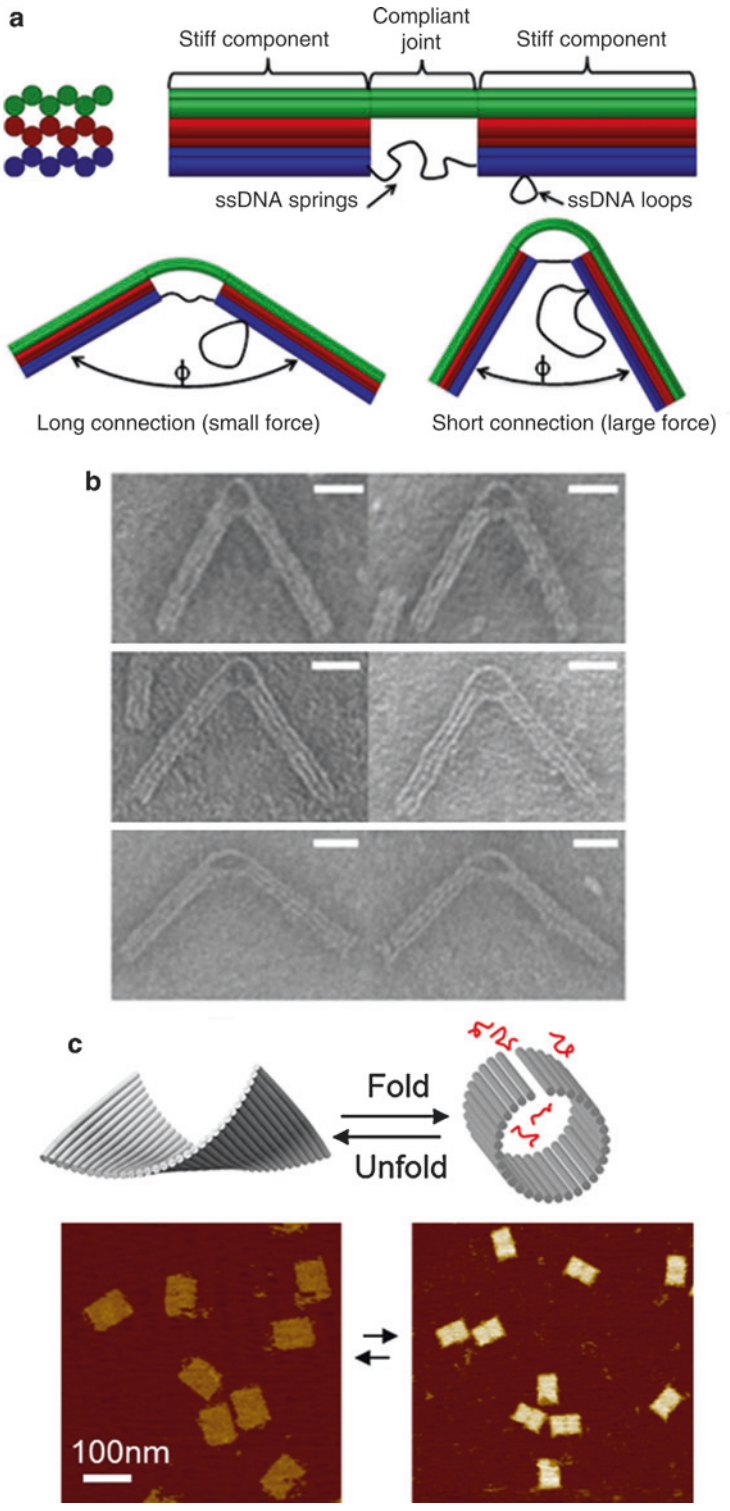
A similar elastic behavior has been demonstrated for 2D DNA origami in solution. As illustrated in Fig. 14c, a 2D rectangular DNA origami has considerable flexibility, which causes the structure to frequently bend over and close the two opposite edges at 40 °C or higher [44]. After introducing linkers that connect the two edges, the majority of the DNA origami structures fold into short tubules as they are frozen in the bent-over conformation. By investigating the kinetics of such a folding process, Chen et al. calculated the bending flexibility of inter-helix crossovers in DNA origami and explained the elastic behavior of the origami structure [44]. They were also able to accelerate folding kinetics by up to 20 times through tuning the structural elasticity by introducing structural defects (i.e., intentionally excluding several staple strands in the origami structure).

The examples discussed here have introduced reconfigurable structures based on the compliance of single-stranded DNA and the elasticity of double-stranded DNA. In analogy, many proteins undergo constant conformational fluctuations [45]



DNA Origami as Programmable Nanofabrication Tools, Fig. 13 (**a**) Folding path of the origami box. (**b**) Schematic of the origami box locked by two pairs of hybridizing extensions. The two pentagrams represent two fluorophores on opposite edges that form efficient FRET (fluorescence resonance energy transfer) pair. (**c**) The box is free to open when the hybridizations are disrupted by strand displacement. The FRET pair is disrupted as the lid opens and can be used to monitor the origami configuration [40]. (**d**) Schematic and AFM images of DNA origami forceps. Streptavidin is used as a model molecular target to be captured by biotin. Scale bar: 300 nm [42]

and experience drastic structural changes in response to environmental stimuli such as light irradiation and chemical binding. A better understanding of the mechanical properties of DNA will enable more dynamic structures with tailored conformations and well-controlled kinetic properties and may even result in transformable structures that emulate proteins in functionality.



DNA Origami as Programmable Nanofabrication Tools, Fig. 14 (**a**) Schematic of a mechanically tunable structure consisting of two stiff bodies, one compliant joint and one ssDNA spring. (**b**) TEM images of the compliant structure tuned into different bending angles. Scale bar: 20 nm [43]. (**c**) Schematic and AFM images of 2D rectangular DNA origami in equilibrium and the folded conformation [44]

Outlooks

While novel properties and promises for applications have been demonstrated, DNA origami technique is still in the phase of rapid development, growing as powerful, programmable nanofabrication tools. Novel geometries generated by innovative design schemes are likely to bring new applications. The development of robust linking schemes will facilitate more versatile assemblies of nanocomponents with higher yields and enable novel functionalities. New chemical strategies will allow the transfer of versatile DNA origami geometries into wider varieties of materials with increased fidelity. Further breakthroughs in scaling up are needed to match the size of DNA origami structures to the feature length scale of lithographic techniques. When the DNA origami is developed as a viable, bottom-up manufacturing technology and integrated with traditional top-down methods, there would be little limit on the complexity and accuracy of the achievable nanostructures and could revolutionize a wide range of industries. Finally, the ability to fabricate compliant dynamic structures is a rather special feature of the DNA origami method. With increased understanding of the fundamental properties of DNA, combined with advances in chemistry, highly functional DNA origami that emulates proteins may be possible, and DNA nanotechnology may join force with synthetic biology.

Cross-References

- ▶ Nanoparticles
- ▶ Nanotechnology
- ▶ Self-Assembly of Nanostructures

References

1. He, Y., Ye, T., Su, M., Zhang, C., Ribbe, A.E., Jiang, W., Mao, C.: Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. Nature **452**(7184), 198–201 (2008)
2. Yan, H., Park, S.H., Finkelstein, G., Reif, J.H., LaBean, T.H.: DNA-templated self-assembly of protein arrays and highly conductive nanowires. Science **301**(5641), 1882–1884 (2003)
3. Zheng, J., Birktoft, J.J., Chen, Y., Wang, T., Sha, R., Constantinou, P.E., Ginell, S.L., Mao, C., Seeman, N. C.: From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal. Nature **461**(7260), 74–77 (2009)
4. Rothemund, P.W.: Folding DNA to create nanoscale shapes and patterns. Nature **440**(7082), 297–302 (2006)
5. Jungmann, R., Liedl, T., Sobey, T.L., Shih, W., Simmel, F.C.: Isothermal assembly of DNA origami structures using denaturing agents. J. Am. Chem. Soc. **130**(31), 10062–10063 (2008)
6. Zhao, Z., Liu, Y., Yan, H.: Organizing DNA origami tiles into larger structures using preformed scaffold frames. Nano Lett. **11**(7), 2997–3002 (2011)
7. BR, H., Liedl, T., Shih, W.M.: Folding DNA origami from a double-stranded source of scaffold. J. Am. Chem. Soc. **131**(26), 9154–9155 (2009)
8. Fu, Y., Zeng, D., Chao, J., Jin, Y., Zhang, Z., Liu, H., Li, D., Ma, H., Huang, Q., Gothelf, K.V.: Single-step rapid assembly of DNA origami nanostructures for addressable nanoscale bioreactors. J. Am. Chem. Soc. **135**(2), 696–702 (2012)
9. Chen, H., Cha, T.-G., Pan, J., Choi, J.H.: Hierarchically assembled DNA origami tubules with reconfigurable chirality. Nanotechnology **24**(43), 435601 (2013)
10. Andersen, E.S., Dong, M., Nielsen, M.M., Jahn, K., Lind-Thomsen, A., Mamdouh, W., Gothelf, K.V., Besenbacher, F., Kjems, J.: DNA origami design of dolphin-shaped structures with flexible tails. ACS Nano **2**(6), 1213–1218 (2008)
11. Douglas, S.M., Dietz, H., Liedl, T., Högberg, B., Graf, F., Shih, W.M.: Self-assembly of DNA into nanoscale three-dimensional shapes. Nature **459**(7245), 414–418 (2009)
12. Liedl, T., Högberg, B., Tytell, J., Ingber, D.E., Shih, W.M.: Self-assembly of three-dimensional prestressed tensegrity structures from DNA. Nat. Nanotechnol. **5**(7), 520–524 (2010)
13. Han, D., Pal, S., Yang, Y., Jiang, S., Nangreave, J., Liu, Y., Yan, H.: DNA gridiron nanostructures based on four-arm junctions. Science **339**(6126), 1412–1415 (2013)
14. Dietz, H., Douglas, S.M., Shih, W.M.: Folding DNA into twisted and curved nanoscale shapes. Science **325**(5941), 725–730 (2009)
15. Hao, E., Schatz, G.C.: Electromagnetic fields around silver nanoparticles and dimers. J. Chem. Phys. **120**(1), 357–366 (2004)
16. Shalaev, V.M.: Optical negative-index metamaterials. Nat. Photonics **1**(1), 41–48 (2007)
17. Ding, B., Deng, Z., Yan, H., Cabrini, S., Zuckermann, R.N., Bokor, J.: Gold nanoparticle self-similar chain structure organized by DNA origami. J. Am. Chem. Soc. **132**(10), 3248–3249 (2010)
18. Jin, Z., Sun, W., Ke, Y., Shih, C.-J., Paulus, G.L., Wang, Q.H., Mu, B., Yin, P., Strano, M.S.: Metallized

DNA nanolithography for encoding and transferring spatial information for graphene patterning. Nat. Commun. **4**, 1663 (2013)
19. Pilo-Pais, M., Watson, A., Demers, S., LaBean, T.H., Finkelstein, G.: Surface-enhanced Raman scattering plasmonic enhancement using DNA origami-based complex metallic nanostructures. Nano Lett. **14**(4), 2099–2104 (2014)
20. Kühler, P., Roller, E.-M., Schreiber, R., Liedl, T., Lohmüller, T., Feldmann, J.: Plasmonic DNA-origami nanoantennas for surface enhanced Raman spectroscopy. Nano Lett. **14**(5), 2914–2919 (2014)
21. Kuzyk, A., Schreiber, R., Fan, Z., Pardatscher, G., Roller, E.-M., Högele, A., Simmel, F.C., Govorov, A. O., Liedl, T.: DNA-based self-assembly of chiral plasmonic nanostructures with tailored optical response. Nature **483**(7389), 311–314 (2012)
22. Medintz, I.L., Uyeda, H.T., Goldman, E.R., Mattoussi, H.: Quantum dot bioconjugates for imaging, labelling and sensing. Nat. Mater. **4**(6), 435–446 (2005)
23. Ko, S.H., Du, K., Liddle, J.A.: Quantum-dot fluorescence lifetime engineering with DNA origami constructs. Angew. Chem. Int. Ed. **52**(4), 1193–1197 (2013)
24. Choi, J.H., Chen, K.H., Strano, M.S.: Aptamer-capped nanocrystal quantum dots: a new method for label-free protein detection. J. Am. Chem. Soc. **128**(49), 15584–15585 (2006)
25. Deng, Z., Samanta, A., Nangreave, J., Yan, H., Liu, Y.: Robust DNA-functionalized core/shell quantum dots with fluorescent emission spanning from UV–vis to near-IR and compatible with DNA-directed self-assembly. J. Am. Chem. Soc. **134**(42), 17424–17427 (2012)
26. Ma, N., Sargent, E.H., Kelley, S.O.: One-step DNA-programmed growth of luminescent and biofunctionalized nanocrystals. Nat. Nanotechnol. **4**(2), 121–125 (2008)
27. Steinhauer, C., Jungmann, R., Sobey, T.L., Simmel, F. C., Tinnefeld, P.: DNA origami as a nanoscopic ruler for super-resolution microscopy. Angew. Chem. Int. Ed. **48**(47), 8870–8873 (2009)
28. Helmig, S., Rotaru, A., Arian, D., Kovbasyuk, L., Arnbjerg, J., Ogilby, P.R., Kjems, J., Mokhir, A., Besenbacher, F., Gothelf, K.V.: Single molecule atomic force microscopy studies of photosensitized singlet oxygen behavior on a DNA origami template. ACS Nano **4**(12), 7475–7480 (2010)
29. Ellington, A.D., Szostak, J.W.: Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures. Nature **355**(6363), 850–852 (1992)
30. Tintoré, M., Gállego, I., Manning, B., Eritja, R., Fàbrega, C.: DNA origami as a DNA repair nanosensor at the single-molecule level. Angew. Chem. Int. Ed. **52**(30), 7747–7750 (2013)
31. Wang, Z.-G., Liu, Q., Ding, B.: Shape-controlled nanofabrication of conducting polymer on planar DNA templates. Chem. Mater. **26**(11), 3364–3367 (2014)
32. Travascio, P., Li, Y., Sen, D.: DNA-enhanced peroxidase activity of a DNA aptamer-hemin complex. Chem. Biol. **5**(9), 505–517 (1998)
33. Surwade, S.P., Zhou, F., Wei, B., Sun, W., Powell, A., O'Donnell, C., Yin, P., Liu, H.: Nanoscale growth and patterning of inorganic oxides using DNA nanostructure templates. J. Am. Chem. Soc. **135**(18), 6778–6781 (2013)
34. Surwade, S.P., Zhao, S., Liu, H.: Molecular lithography through DNA-mediated etching and masking of SiO_2. J. Am. Chem. Soc. **133**(31), 11868–11871 (2011)
35. Pinheiro, A.V., Han, D., Shih, W.M., Yan, H.: Challenges and opportunities for structural DNA nanotechnology. Nat. Nanotechnol. **6**(12), 763–772 (2011)
36. Liu, W., Zhong, H., Wang, R., Seeman, N.C.: Crystalline two-dimensional DNA-origami arrays. Angew. Chem. **123**(1), 278–281 (2011)
37. Asanuma, H., Liang, X., Nishioka, H., Matsunaga, D., Liu, M., Komiyama, M.: Synthesis of azobenzene-tethered DNA for reversible photo-regulation of DNA functions: hybridization and transcription. Nat. Protoc. **2**(1), 203–212 (2007)
38. Yang, Y., Endo, M., Hidaka, K., Sugiyama, H.: Photocontrollable DNA origami nanostructures assembling into predesigned multiorientational patterns. J. Am. Chem. Soc. **134**(51), 20645–20653 (2012)
39. Zhang, D.Y., Winfree, E.: Control of DNA strand displacement kinetics using toehold exchange. J. Am. Chem. Soc. **131**(47), 17303–17314 (2009)
40. Andersen, E.S., Dong, M., Nielsen, M.M., Jahn, K., Subramani, R., Mamdouh, W., Golas, M.M., Sander, B., Stark, H., Oliveira, C.L.: Self-assembly of a nanoscale DNA box with a controllable lid. Nature **459**(7243), 73–76 (2009)
41. Douglas, S.M., Bachelet, I., Church, G.M.: A logic-gated nanorobot for targeted transport of molecular payloads. Science **335**(6070), 831–834 (2012)
42. Kuzuya, A., Sakai, Y., Yamazaki, T., Xu, Y., Komiyama, M.: Nanomechanical DNA origami 'single-molecule beacons' directly imaged by atomic force microscopy. Nat. Commun. **2**, 449 (2011)
43. Zhou, L., Marras, A.E., Su, H.-J., Castro, C.E.: DNA origami compliant nanostructures with tunable mechanical properties. ACS Nano **8**(1), 27–34 (2013)
44. Chen, H., Weng, T.-W., Riccitelli, M., Cui, Y., Irudayaraj, J., Choi, J.H.: Understanding the mechanical properties of DNA origami tiles and controlling the kinetics of their folding and unfolding reconfiguration. J. Am. Chem. Soc. **136**, 6995–7005 (2014)
45. Yang, H., Luo, G., Karnchanaphanurach, P., Louie, T.-M., Rech, I., Cova, S., Xun, L., Xie, X.S.: Protein conformational dynamics probed by single-molecule electron transfer. Science **302**(5643), 262–266 (2003)

DOLLOP (Dynamically Ordered Liquid-Like Oxyanion Polymer)

▶ Prenucleation Clusters

Doping in Organic Semiconductors

Yabing Qi
Energy Materials and Surface Sciences (EMSS) Unit, Okinawa Institute of Science and Technology, Onna-son, Kunigami-gun, Okinawa, Japan

Synonyms

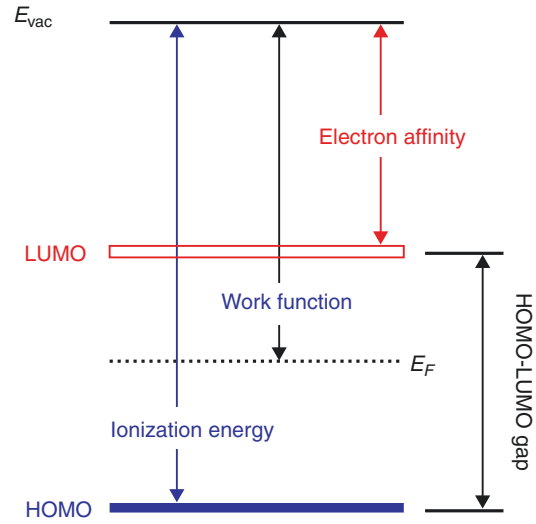
Doping of organic semiconductors

Definition

There are two major ways to dope an organic semiconductor: chemical doping and electrochemical doping. Because of the additional complexity due to the presence of an electrolyte solution in the case of electrochemical doping, the focus is placed on chemical doping. In chemical doping, dopant molecules (typically a few weight percent) are introduced into an organic semiconducting host material, either by co-evaporation in vacuum or spin casting from solution.

Operation Principles

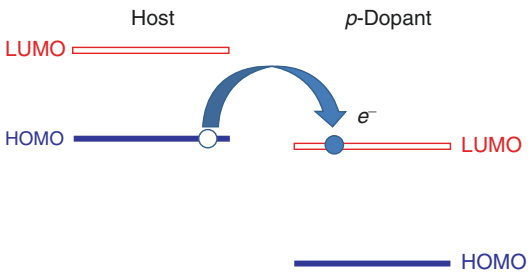
Doping in organic semiconductors can be rationalized based on energetics, i.e., charge carriers (electrons or holes) "flow" from the higher energy levels of host molecules to the lower energy levels of dopant molecules to minimize the total energy. It is known that the energy levels in atoms are discrete and narrow in width. On the other hand, as a large number of atoms are brought to proximity to form solids with an ordered crystal structure, the energy levels evolve to continuous bands.



Doping in Organic Semiconductors, Fig. 1 Schematic drawing of energy levels in molecules

Energy levels in molecules are somewhere between these two extremes, not as discrete as those in atoms, but also not as continuous as those in crystals. Figure 1 illustrates the energy levels in molecules. The highest energy levels that are occupied by electrons are the so-called highest occupied molecular orbitals (HOMO). Any energy level below HOMO is occupied by electrons. The lowest energy levels that are not occupied are the so-called lowest unoccupied orbitals (LUMO). Any energy level above LUMO is unoccupied. The energy required to excite an electron in HOMO of a molecule to the vacuum level is defined as the ionization energy (IE). The energy gain for an electron falling from the vacuum level to fill LUMO is electron affinity (EA). The energy difference between the Fermi level and the vacuum level is work function (WF).

In order to efficiently *p*-dope a host material, the EA of the dopant molecules needs be comparable to or even greater than the IE of the host molecules (see Fig. 2). Vice versa, an efficient n-type dopant is expected when the dopant molecules have an IE comparable to even lower than the EA of the host material. As an example, α-NPD (*N*,*N*′-diphenyl-*N*,*N*′-bis (1-naphthyl)-1,1′-biphenyl-4,4′-diamine) is a widely used hole transport material in organic light emitting diodes



Doping in Organic Semiconductors, Fig. 2 Schematic drawing of charge transfer for p-type doping

(OLEDs). The ionization energy and electron affinity of α-NPD are 5.4 and 1.4 eV, respectively [3, 6]. An efficient p-type dopant for α-NPD should have an EA comparable to 5.4 eV. F_4-TCNQ (2,3,5,6-tetrafluoro-7,7,8,8-tetracyanoquinodimethane) has an EA of 5.24 eV, which has been shown to efficiently *p*-dope α-NPD [2]. It is worthwhile to note that in Fig. 2 an organic molecule dopant is assumed, although materials other than organic molecules have been used as dopants. For example, alkali metals have been used as p-type dopants and halogens have been used as n-type dopants.

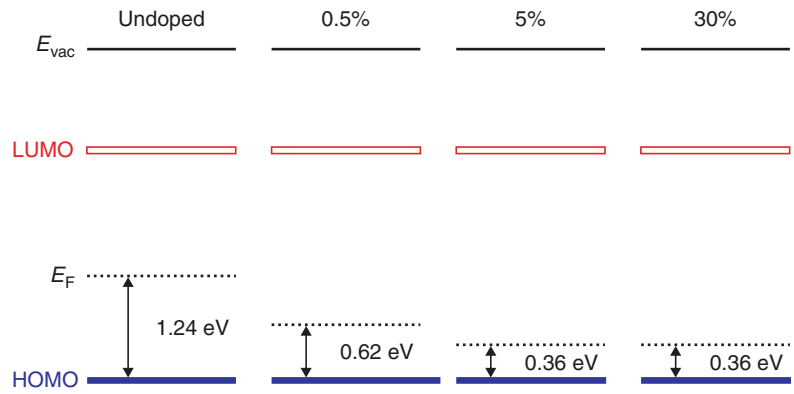
Basic Methodology and Key Research Findings

Experimental techniques that have been commonly used in doping studies of organic semiconductors include ultraviolet photoemission spectroscopy (UPS), inverse photoemission spectroscopy (IPES), x-ray photoelectron spectroscopy (XPS), Rutherford backscattering (RBS), charge transport measurements, and scanning tunneling microscopy (STM).

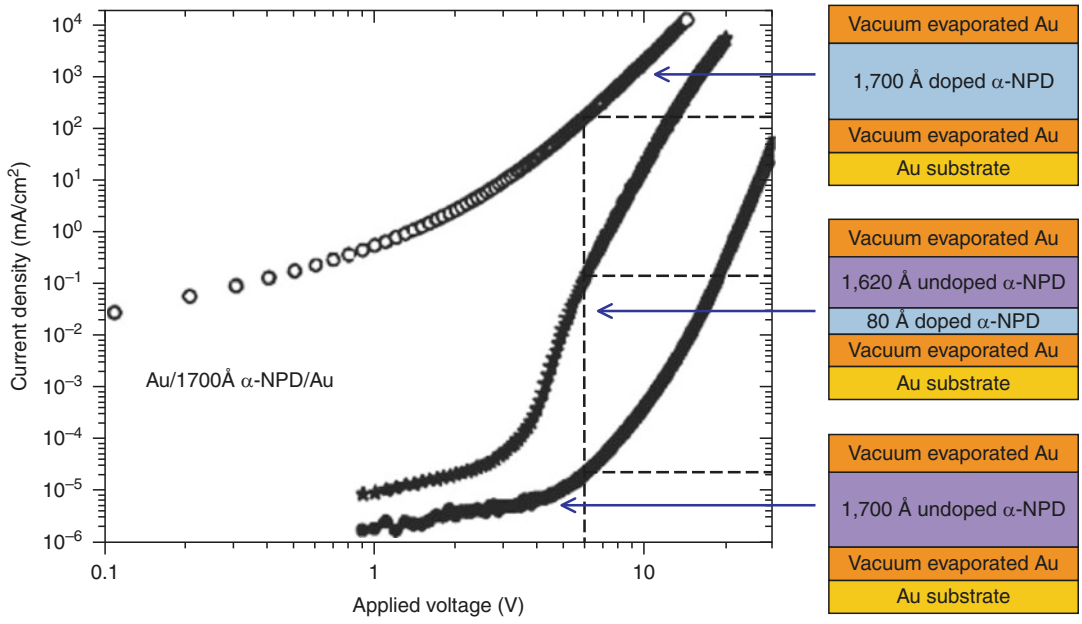
The first set of measurements for a typical doping study are usually UPS and IPES measurements on neat films of host and dopant molecules, which provide fundamental information such as IE and EA. With such information for both materials, a crude judgment can be made whether doping can take place between these two materials. While electrochemical measurements combined with optical absorption measurements are also often cited in literature to yield values of IE and EA, it should be noted that (1) electrochemical measurements are usually performed in solution where intermolecular interaction is rather weak and (2) optical gap values need to be corrected for exciton binding energy in order to be compared with the HOMO–LUMO gap (= IE − EA) measured by UPS and IPES.

In order to accurately assess effectiveness of doping, more direct measurements are required, one of which is to study the Fermi level position change by UPS as a function of doping concentration. When organic materials are not intentionally doped (so-called intrinsic), the Fermi level is often close to the mid-gap position. A Fermi level shift from the near mid-gap position to HOMO is expected for the p-type doping. For the n-type doping, on the other hand, the Fermi level shifts toward LUMO. Figure 3 schematically depicts the Fermi level shift toward HOMO as the concentration of F_4-TCNQ in α-NPD increases from 0 % to 0.5 %, and then to 5 % [2]. It is worthwhile noting that as the doping concentration increases further to 30 %, the Fermi level remains at approximately the same position, and is pinned at ∼ 0.4 eV above HOMO. This is a result of the tail of density of states extended into the HOMO–LUMO gap. It has been proposed that various types of static and/or dynamic disorder in organic films are responsible for these states.

0.019w?>The second direct way to visualize the doping effect is to measure charge transport properties of undoped and doped films. Bulk conductivity enhancement and charge injection improvement at organic–electrode interfaces are two major doping-induced advantages, both of which can be verified via charge transport measurements. These two features are demonstrated in Fig. 4, where the current–voltage (I-V) characteristics are compared for undoped α-NPD, 0.5 % F_4-TCNQ interface–doped α-NPD, and 0.5 % F_4-TCNQ bulk–doped α-NPD samples [2]. Detailed structures of the three types of samples are schematically shown on the right in Fig. 4. Due to the amorphous nature of α-NPD films, charge carrier hopping is believed to be the main conduction mechanism in α-NPD. The first sample is an undoped sample (represented by solid circles in Fig. 4). For this



Doping in Organic Semiconductors, Fig. 3 Position of the Fermi level in the HOMO–LUMO gap of ***α***-NPD as a function of doping concentration of F_4-TCNQ. The values are taken from Gao and Kahn [2]. Reprinted with permission. Copyright 2003, American Institute of Physics



Doping in Organic Semiconductors, Fig. 4 I-V curves of undoped α-NPD, 0.5 % F_4-TCNQ interface–doped α-NPD, and 0.5 % F_4-TCNQ bulk–doped α-NPD samples. Detailed structures of the three types of samples are schematically shown on the right. Reprinted with permission from Gao and Kahn [2]. Copyright 2003, American Institute of Physics

sample, the I-V curve basically follows a straight line with the slope approximately equal to 1 at low applied voltages (between 0.9 and 4 V). The electrical current in this region is dominated by ohmic conduction due to free charge carriers. As the applied bias increases to 6 V, the current begins to climb up with a much steeper slope. This rapid increase is called injection onset, caused by charge carrier injection at the organic/electrode interfaces. The steep slope is most likely caused by the filling-up of trap states in α-NPD, which in turn facilitate charge carrier hopping in α-NPD. Doping can be used to improve injection, which is demonstrated by the interface-doped α-NPD sample (represented by stars in Fig. 4). I-V characteristics at low applied voltages (between 0.9 and 2 V) share the same features as those of the undoped sample, i.e., linear on log-log scale with a slope of 1.

The current in this regime is within a factor of 5 of that of the undoped one, which is expected from the fact that only a small portion of the film is doped. On the other hand, injection properties are significantly improved, and the injection onset occurs at a much lower bias (~3 V). For the bulk-doped sample (represented by open circles in Fig. 4), the low bias conductivity (i.e., in the ohmic conduction region) is 5 orders of magnitude higher than that of the undoped sample. Along the same line, doping has been implemented in organic field-effect transistors (OFETs) to tackle the contact resistance issue, which is one of the common problems. Studies have shown that contact resistance in pentacene OFETs has been reduced by sevenfold by inserting a 10-nm-thick doped interface layer [4]. It is important to note that the stability of the devices was not affected significantly by doping of the contacts [4].

In some cases, as dopant molecules donate (or accept) electrons from the host material, oxidation states of some specific atoms in the dopant molecules change. This change can be used as an indicator for doping. Oxidation states can be determined by XPS. For example, such a change is revealed in an XPS study on cobaltocene and cobaltocene-doped tris(thieno)hexaazatriphenylene derivative [1].

Future Directions for Research

In the past decade or so, applications involving organic semiconductor (especially OLED-based products such as OLED lighting, OLED TV, OLED cell phone, etc.) have been commercialized successfully. Meanwhile, there is a rapidly growing interest in using doping to improve performance of these organic based devices [5]. Despite increased interest and research efforts along this direction to make doping a process readily applicable in organic based devices, several key issues need to be addressed appropriately. First of all, air stability of dopants and doped films are crucial. This is particularly true for n-type dopants, because with a high-lying HOMO n-dopants are prone to oxidation when exposed to air. Secondly, to cut down fabrication cost, it is highly desirable to be able to introduce dopants via ambient pressure, low-temperature, and high-throughput techniques such as spin coating, spary pyrolysis, inkjet printing, etc. Thirdly, to ensure reliable operation of organic-based devices, dopants need to be spatially confined within the targeted region. Therefore, diffusional stability of dopants in the host matrix is of great value. Last but not least, systematic research is necessary to search for new dopants that are of low cost and high performance.

Cross-References

- ▶ Charge Transport in Self-assembled Monolayers
- ▶ Conduction Mechanisms in Organic Semiconductors
- ▶ Electrode–Organic Interface Physics
- ▶ Flexible Electronics
- ▶ Hybrid Solar Cells
- ▶ Maxwell–Wagner Effect
- ▶ Organic Actuators
- ▶ Organic Bioelectronics
- ▶ Organic Photovoltaics: Basic Concepts and Device Physics
- ▶ Organic Sensors
- ▶ Surface Electronic Structure

References

1. Chan, C.K., Amy, F., Zhang, Q., Barlow, S., Marder, S. R., Kahn, A.: Chem. Phys. Lett. **431**, 67 (2006)
2. Gao, W., Kahn, A.: J. Appl. Phys. **94**, 359 (2003)
3. Hill, I., Kahn, A.: J. Appl. Phys. **84**, 5583 (1998)
4. Tiwaria, S.P., Potscavage Jr., W.J., Sajoto, T., Barlow, S., Marder, S.R., Kippelen, B.: Org. Electron. **11**, 860 (2010)
5. Walzer, K., Maennig, B., Pfeiffer, M., Leo, K.: Chem. Rev. **107**, 1233 (2007)
6. Wan, A., Hwang, J., Amy, F., Kahn, A.: Org. Electron. **6**, 47 (2005)

Doping of Organic Semiconductors

- ▶ Doping in Organic Semiconductors

Double-Walled Carbon Nanotubes (DWCNTs)

▶ Chemical Vapor Deposition (CVD)

Droplet Microfluidics

▶ Electrowetting
▶ Lattice Boltzmann Methods for Nanofluidics

Drug Delivery

▶ Use of Nanotechnology in Pregnancy

Drug Delivery and Encapsulation

▶ Acoustic Nanoparticle Synthesis for Applications in Nanomedicine

Drug Delivery System

▶ Chitosan Nanoparticles

Drug Design

▶ Computational Chemistry for Drug Discovery

Drug Targeting

▶ Magnetic Nanoparticles for Biomedical Applications

Dry Adhesion

▶ Gecko Adhesion

Dry Etching Processes

Salman B. Inayat[1], Avinash P. Nayak[1,2], V. J. Logeeswaran[1,2] and M. Saif Islam[1,2]
[1]University of California, Davis, CA, USA
[2]Electrical and Computer Engineering, University of California – Davis Integrated Nanodevices and Nanosystems Lab, Davis, CA, USA

Synonyms

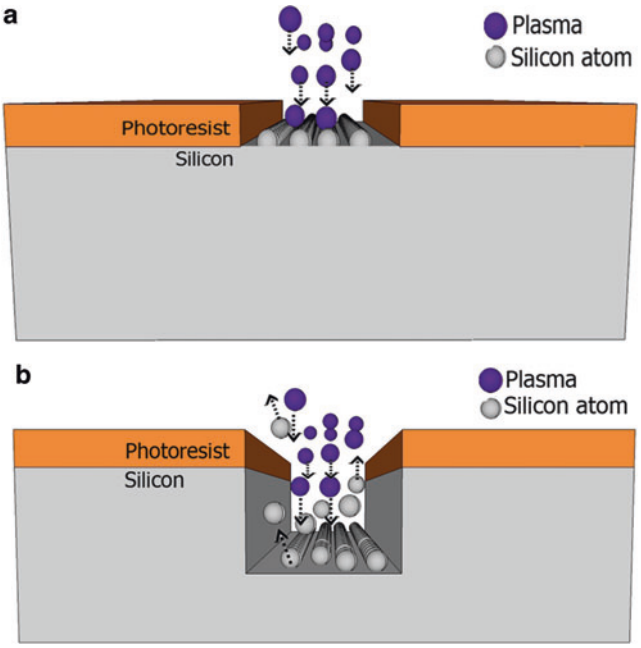
Chemical dry etching; Gas etching; Physical-chemical etching; Physical dry etching; Plasma etching

Definition

In dry etching, substrate material is removed by impinging radicals in plasma or by etchant gases. The reaction that takes place can be done utilizing high kinetic energy of particle beams, chemical reaction, or a combination of both.

Physical Dry Etching

Physical dry etching requires high energy kinetic energy (ion, electron, or photon) beams to etch off the substrate atoms. When the high-energy particles knock out the atoms from the substrate surface, the material evaporates after leaving the substrate. There is no chemical reaction taking place and therefore only the material that is unmasked will be removed. The physical reaction taking place is illustrated in Fig. 1.



Dry Etching Processes, Fig. 1 The plasma hits the silicon wafer with high energy to knock off the Si atoms on the surface. (**a**) The plasma atoms hitting the surface. (**b**) The silicon atoms being evaporated off from the surface

Chemical Dry Etching

Chemical dry etching (also called vapor phase etching) involves a chemical reaction between etchant gases to attack the silicon surface or the substrate. The process does not use liquid chemicals or etchants. The chemical dry etching process is usually isotropic and exhibits high selectively. Anisotropic dry etching has the ability to etch with finer resolution and higher aspect ratio than isotropic etching. Due to the directional nature of dry etching, undercutting can be avoided. Figure 2 shows a rendition of the reaction that takes place in chemical dry etching. Some of the ions that are used in chemical dry etching is tetrafluoromethane (CH_4), sulfur hexafluoride (SF_6), nitrogen trifluoride (NF_3), chlorine gas (Cl_2), or fluorine (F_2).

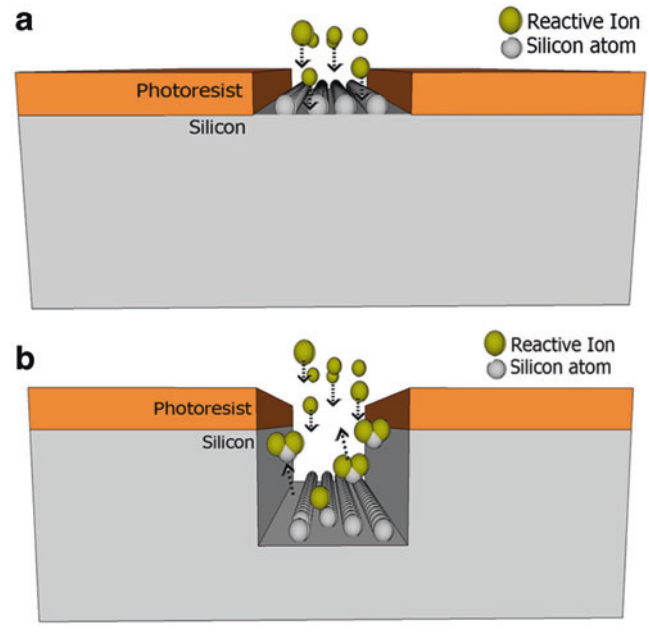
Reactive Ion Etching

Reactive ion etching (RIE) provides high level of resolution using both physical and chemical etching mechanisms. RIE is one of the most diverse and most widely used processes in semiconductor industry and research settings. Since the process combines both physical and chemical interactions, the process is much faster. The high-energy collision from the ionization helps to dissociate the etchant molecules into more reactive species.

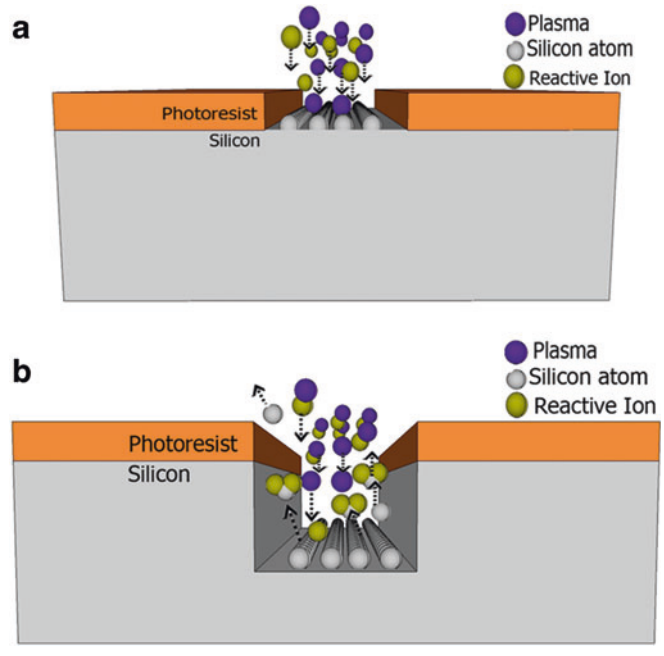
In the RIE process, cations are produced from reactive gases which are accelerated with high energy to the substrate and chemically react with the silicon. The typical RIE gases for Si are CF_4, SF_6, and $BCl_2 + Cl_2$. As seen in Fig. 3, both physical and chemical reactions are taking place. Table 1 shows some commonly used gases for etching semiconductor and dielectric thin films.

Deep Reactive Ion Etching (DRIE)

Major challenges toward fabricating high-aspect-ratio structures with dry etching include (a) low etch rates, (b) inability to maintain high-aspect ratios for etching depths of 10 μ or more, and (c) weakness of photoresist masks to survive extended etching runs [1]. Deep reactive ion etching (DRIE) provides a viable solution to these problems and successfully etches structures as deep as the thickness of a silicon wafer with extremely high-aspect ratios of 20–30:1 [1]. The DRIE process is



Dry Etching Processes, Fig. 2 Process of a reactive ion interacting with the silicon surface. (**a**) The interaction between the reactive ion and the silicon atom. (**b**) A bond between the reactive ion and the silicon atom then chemically remove the silicon atoms from the surface



Dry Etching Processes, Fig. 3 RIE process. This process involves both physical and chemical reactions to etch off silicon

comprised of two steps. In the polymerization or passivation step, a thin 10 mm layer is deposited on the side walls and bottom of a trench etched in the previous etching step. The (plasma) etching step following the passivation step removes the polymer from the bottom of the trench and proceeds with etching of silicon while the side walls remain protected by the polymer. In this way high anisotropy is achieved through repeated succession of etch and passivation sequence.

DRIE works on the mechanism of plasma etching in the presence of process gases (SF6 for

Dry Etching Processes, Table 1 Commonly used gases for etching semiconductor and dielectric thin films

Material	Reactive gas	Etching rate (Å/min)	Etch mask	Selectivity
Si (a-Si)	CF_4, SF_6, $BCI_2 + Cl_2$	~500	Resist meal (cr, Ni, Al)	~20:1
				~40:1
SiO_2	$CHF_3 + O_2$,, $CF_4 + H_2$	~200	Resist metal (Cr, Ni, Al)	~10:1
				~30:1
Si_3N_4	$CF_4 + O_2$ (H_2), CHF_3	~100	Resist metal (Cr, Ni, Al)	~10:1
				~20:1
GaAs	Cl2, $Cl_2 + BCl_3$	~200	Sl_3N_4 metal (Cr, Ni)	~10:1
				~20:1
InP	CH_4/H_2	~200	Sl_3N_4 metal (Cr, Ni, Al)	~40:1
Photoresist, any polymer	O_2	~500	Sl_3N_4 metal (Cr, Ni)	~50:1

etching and C4F8 for polymerization) in the etching chamber. Like any plasma etching, DRIE requires the presence of sustained plasma through the generation of secondary elections after the initial plasma-generating electrons have been consumed. These secondary electrons are emitted by the cathode when struck by the ions inside the plasma. Collection of positive ions around the substrate (cathode) forms a sheath or a dark space. This sheath ensures further accumulation of positive ions responsible for etching. When the substrate placed on cathode is a conductor, the sheath is maintained with its positive polarity and the bombardment of substrate is sustained.

In case of an insulating substrate, however, the substrate accumulates the positive charges nullifying the field in chamber (between anode and cathode). The plasma ceases and etching process stops. In DRIE process the process chambers are operated with high frequency RF source. The rapidly alternating polarity of electrodes prevents the positive charges from accumulating on the substrate which are neutralized in the second half cycle by the impinging electrons (substrate acts as anode in the second half cycle).

Process Issues

Bull's eye effect: The phenomenon by which edges of wafer etch faster than the central regions [2].

Microloading: The phenomenon by which isolated features etch faster than those located in a dense array.

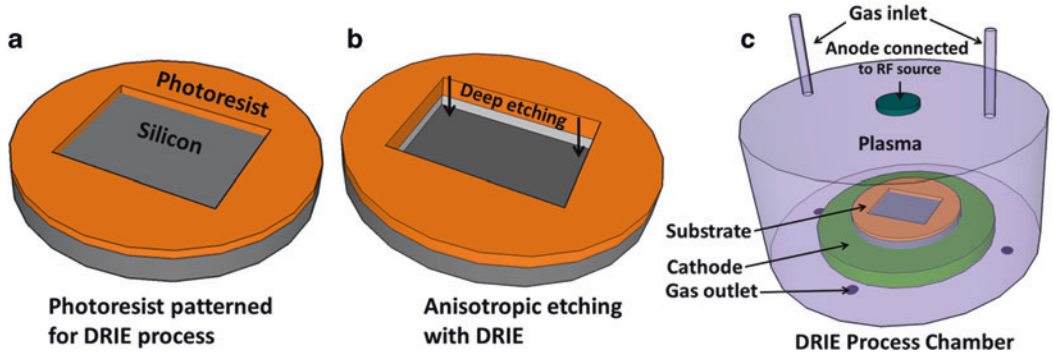
Aspect ratio dependent etch (ARDE): The occurrence of smaller features etching slower than those with larger dimensions.

Inductively Coupled Plasma: Reactive Ion Etching (ICP-RIE)

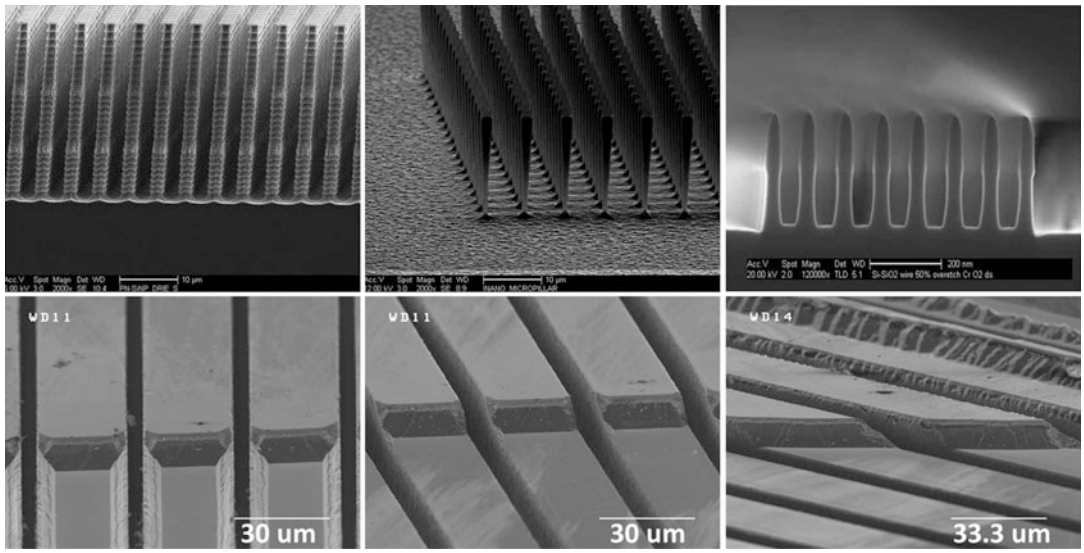
Although plasma etching provides high etch rates at low chamber pressure for increased directionality, the plasma density is low due to low chamber pressure. Low plasma densities not only degrades selectivity but also renders inability to maintain high-aspect ratios [1]. High-density plasma etchers, namely, ICP-RIE, provide a solution to abovementioned problems by producing high-density plasma at low pressure. ICP-RIE comprises of magnetic coils that create a magnetic field to control the ion-producing electrons. The magnetic field ensures electrons are prevented from colliding with electrodes or chamber walls. This increases the probability of electrons for successful ion creating collisions, thus increasing the plasma density at low pressure.

Applications

Dry etching is used to etch some selective areas of a semiconductor surface in order to form recesses



Dry Etching Processes, Fig. 4 (**a** and **b**) Wafer preparation for deep reactive ion etching (**c**) Inside the DRIE process chamber



Dry Etching Processes, Fig. 5 Silicon nanopillars [3], nanowalls [4], and nanoridges fabricated using deep, ICP, and regular reactive ion etching

in material, such as mesa, contact holes, or via holes, to otherwise remove portions of semiconductor or dielectric layers where predominantly vertical sides are desired. Along with semiconductor manufacturing, micromachining, and display production, the removal of organic residues by oxygen plasmas is also described as a dry etch process. Dry etching is particularly useful for materials and semiconductors, which are chemically resistant and could not be wet etched, such as silicon carbide or diamond.

Logeeswaran et al. [3] fabricated high-aspect-ratio vertically oriented silicon micro-pillars using deep reactive ion etching (DRIE) to develop highly efficient portable and flexible electronic devices for future applications. As shown in Fig. 4, a highly doped p-type Si (100) substrate with a doping concentration of $\sim 10^{19}$ cm^{-3} was patterned with a 2-μm dot etch mask. The substrate was kept at 10 °C with SF_6 and C_4F_8 flows of 300 and 150 sccm, respectively, a source RF power at 1,800 W and a substrate power at 20 W for a total etching time of ~6 min. These high-aspect-ratio structures were etched using a special DRIE recipe called Bosch process. The Bosch process is a pulsed-multiplexed etching technique which

alternates between two modes to achieve extremely long and vertical micro-scaled nanostructures (Fig. 5).

Chen et al. [5] carried out an extensive study to determine the dependence of the etching performance. Factors such as applied coil or electrode power, reactant gas flow rates, duty cycles, and chamber pressures were considered. These process parameters were modeled and experimentally determined to see the effect on the surface morphology and the mechanical performance of the silicon structures.

References

1. Marc, J.: Madou, Fundamentals of Microfabrication and Nanotechnology. CRC Press, Boca Raton (2012)
2. http://nanolab.berkeley.edu/labmanual/chap7/7.22sts2.pdf
3. Logeeswaran, V.J., et al.: Harvesting and transferring vertical pillar arrays of single-crystal semiconductor devices to arbitrary substrates. Electron Devices IEEE Trans. on. **57**, 1856–1864 (2010); and Logeeswaran, V. J., Jinyong, O.H., Avinash, P., Nayak, A.M., Katzenmeyer, S.G., Kristin H.G., Alec T.A., Nibir, K., Dhar, N.P., Kobayashi, S-Y.W., Saif, I.M.: A perspective on nanowire photodetectors: current status, future challenges and opportunities. IEEE J. Sel. Top. Quantum Electron. **17**(4), 1002 (2011)
4. Hakan, K., Logeeswaran, V.J., Saif I.M.: Fabrication of 3D-silicon micropillars walls decorated with aluminum-ZnO/ZnO nanowires for optoelectric devices. Phys. Status Solidi A. **210**, 1–4 (2013)
5. Chen, et al.: Effect of Process Parameters on the Surface Morphology and Mechanical Performance of Silicon Structures After Deep Reactive Ion Etching (DRIE), vol. 11. ETATS-UNIS: Institute of Electrical and Electronics Engineers, New York (2002)

Dual-Beam Piezoelectric Switches

▶ NEMS Piezoelectric Switches

Duffing

▶ Nonlinear and Parametric NEMS Resonators

DUV Lithography

▶ DUV Photolithography and Materials

DUV Photolithography and Materials

Garry J. Bordonaro
Cornell NanoScale Science and Technology Facility, Cornell University, Ithaca, NY, USA

Synonyms

193-nm lithography; 248-nm lithography; DUV lithography; Photolithography

Definition

Deep ultraviolet (DUV) photolithography is the process of defining a pattern in a thin photosensitive polymer layer (photoresist) using controlled 254–193-nm light such that the resulting polymer pattern can be transferred into or onto the underlying substrate by etching, deposition, or implantation. The exposing light is passed through a chrome-on-quartz photomask, whose opaque areas act as a stencil of the desired pattern. The exposed polymer is then subjected to a chemical development process where the unwanted areas of polymer are removed, leaving the target areas unprotected from subsequent processing.

In semiconductor manufacturing, the light sources used are typically excimer lasers and produce either 248-nm or 193-nm light. The light is usually passed through the photomask and then through a reduction lens which reduces the pattern size by a factor of 5 or 4 times.

Introduction

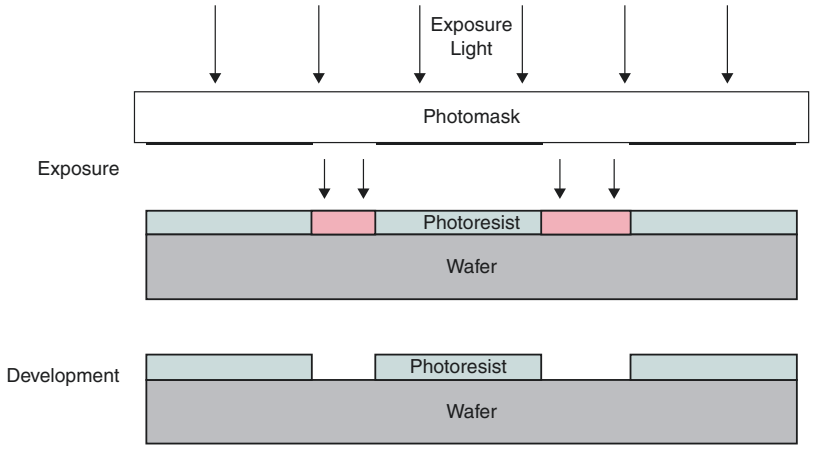
The process of photolithography has been performed since at least the 1950s in the

manufacturing of circuit boards and discrete electronic devices, initially using cyclized rubber-based photoresists as the patterning medium in the process of contact lithography, where the photomask is brought into contact or near-contact with the substrate for direct 1:1 patterning, using mask-aligning exposure tools. Modern DNQ-novolac (diazonaphthoquinone) photoresists were developed in the early 1970s allowing features smaller than 2 μm to be printed and opened the door to the age of mass-produced silicon-based integrated circuits. The resulting economies of scale eventually led to the development of the personal computer for the mass market. The growth of this market, in turn, drove the development of advanced photoresist materials and exposure tools to satisfy the demands of the market for faster, smaller, and cheaper technology products.

The process of UV (ultraviolet, 436–365 nm) photolithography as performed for the last 30 years remains basically unchanged in that time. A design pattern is generated using computer-aided design (CAD) software. The pattern data is converted into a standardized file format, usually GDSII or, more recently, OASIS. This data is then converted by specialized software into the format needed for the specific mask-writing tool to be used and then sent to the mask writer. The mask writer exposes the pattern on a thin resist film (100–500 nm) coated on top of a thin chrome layer (1,000 Å typ.), which has been deposited on top of the photomask substrate, which is most often fused silica or quartz, but can be other types of glass. The exposed pattern is then chemically developed, revealing the chrome layer underneath in the shape of the desired pattern. Next, the uncovered chrome is etched away with an acid, removing the designed pattern from the chrome layer. The end result is an opaque stencil made from thin chrome on a transparent plate, which is suitable for use in either a mask aligner for contact photolithography or a projection photolithography tool. The photomask can then be used many times, particularly in projection photolithography, where the pattern area of the mask never contacts the wafer or the tool.

The device substrate, usually a silicon wafer, is spin coated with a thin photoresist layer and baked at a temperature near the glass-transition temperature of the resist to remove most of the solvent. Exposure of the mask pattern into the resist is then performed, either by bringing the wafer and mask together before exposure (contact or proximity photolithography, 1:1) or by mounting the mask in front of a (usually reduction) lens and projecting the image onto the wafer (projection photolithography, 4:1 typ.). Production tools currently are virtually all reduction projection printers, and they are what the remainder of this discussion will focus on. The DNQ-novolac chemistries used in UV photolithography today are substantially the same as when they were first formulated, with some tweaking of wavelength absorption properties and sensitivities. The exposure light causes a chemical reaction within the photoresist, forming acid (for DNQ resist, indene carboxylic acid) in the exposed areas. An aqueous base solution (developer) removes these acidic areas of the resist, creating a three-dimensional replica of the pattern seen on the photomask. Positive-tone photoresists such as this are dominant in usage, but negative-tone versions, where the exposed areas *remain* after development, also are available. The high-pressure mercury (Hg) arc lamp light sources used still possess the same output spectrum characteristics as before, and the exposure tools using them, while more sophisticated and automated than before, are essentially the same (Fig. 1).

The emergence of modern-day DUV (deep ultraviolet, 248–193 nm) exposure tools and chemistry has profoundly changed this established process. What was once a comparatively forgiving series of process steps has now become a critically controlled time and temperature ballet with incredibly complex machinery and computerized systems. The development of DUV photolithography has caused engineers to push the boundaries of our understanding of the behavior of light, as well as our concepts regarding mechanical and chemical systems, beyond



DUV Photolithography and Materials, Fig. 1 Basic photolithography process

what we once considered the limits of our abilities. Most remarkably, it is done on a daily basis in high-volume manufacturing (HVM) and with great success.

History

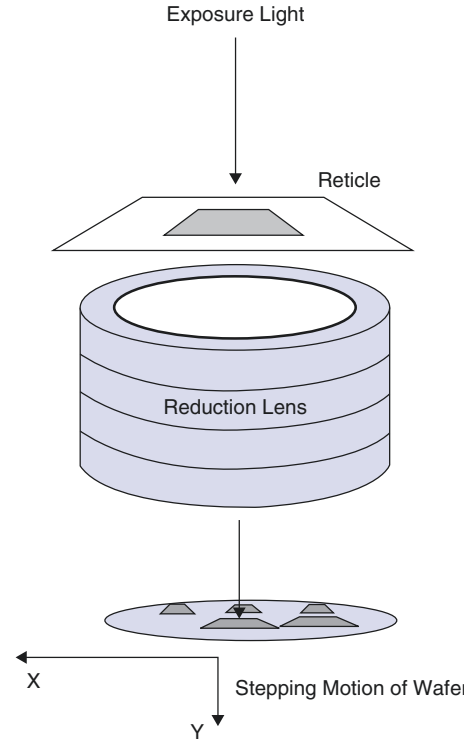
Modern DUV lithography had its beginning in the early 1980s at IBM with Frechet, Willson, and Ito and their development of chemically amplified resists. The drive toward ever-smaller features in the structures that make up the electronic devices in the marketplace required the use of ever-shorter-wavelength exposure light, which in turn required new photoresists with appropriate sensitivity to these new wavelengths. Unfortunately, the existing photoresist materials were too absorptive and insensitive in the DUV wavelength regions to be practical. New polymer systems were developed with sufficient transparency at these wavelengths, but their sensitivity to exposure was still far too low. Chemically amplified (CA) resists overcame this by using a complex heat-activated de-protection method that multiplied the acid formed by exposure, which increased the material's DUV sensitivity by some orders of magnitude, leading to today's superior performers with resolution capabilities far better than DNQ-sensitized materials could ever achieve.

However, this was not an overnight success story. Although these resists began to appear in the early 1980s, and IBM had a small production line in Burlington, Vermont, in the late 1980s, HVM using 248-nm photolithography did not begin until the late 1990s. There were two primary reasons for this: a lack of adequate photoresists and delayed tool development. The CA resists described previously would take another 10 years to reach the stage of development where they were robust enough to survive the rigors of mass production with reasonable cost. Problems with airborne contamination and etch resistance also took years to overcome, but in the end DUV resist performance surpassed expectations, with etch resistance nearly equal to DNQ materials and with much better sensitivity and resolution. The exposure tooling also encountered many roadblocks during development, including difficulty in identifying suitable fused silica material for lenses and masks, developing suitable 248-nm light sources, and, perhaps most importantly, finding funds to research all of these issues. The SEMATECH consortium played a vital role in this by helping to organize and direct efforts and funding in these matters. Industry-university collaborations also played key roles in areas like materials and chemical engineering. They reprised these roles during the similar progression in the development of 193-nm materials and tools in the 1990s. Thankfully, lessons learned in resist

development and material qualification during the run-up to 248-nm HVM shortened the 193-nm development timeline by about 50 %. This learning was also important in later investigations into the viability of 157-nm (F_2) photolithography, a technology that was ultimately rejected.

DUV Processing

DUV photolithography is at its core the same as UV photolithography. The two methods share many of the same basic processing steps: photoresist is spin coated onto wafers, baked, exposed, post-baked, and developed. Similar tools are used for both technologies in HVM: spin coaters, hot plates, step-and-repeat systems (steppers), and aqueous developers. However, there are important differences in these processes besides the minimum value of the resulting pattern's critical dimensions (CD). The devil is, as usual, in the details (Fig. 2).



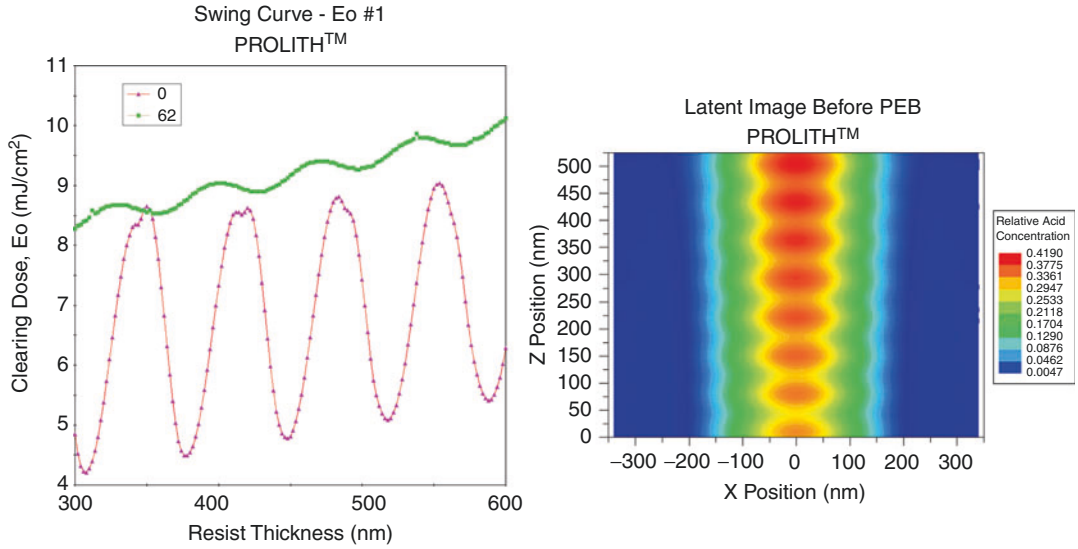
DUV Photolithography and Materials, Fig. 2 Wafer stepper schematic

Wafer Preparation

Photolithographic processing always begins with cleaning of the wafer surface. Depending on the state of the wafer (new or in process), the cleaning process can take many forms and be either dry or wet. Final results depend heavily on the cleanliness of the wafer surface, so the importance of this step should not be minimalized. Typical new wafer cleaning may include the "standard cleans," SC1 and SC2, which derive from the well-known "RCA clean" originally used in vacuum tube manufacturing. SC1 is a 5:1:1 solution of water, ammonium hydroxide, and hydrogen peroxide heated to ~75 °C. SC2 is a 6:1:1 solution of water, hydrochloric acid, and hydrogen peroxide heated to ~75 °C. Both baths are usually performed for ~10 min each, and an additional step consisting of a dip in dilute hydrofluoric acid is sometimes included as well. Cleaning of wafers which are "in process" (partially completed) can take many forms, each dependent on the previous processing steps performed on the wafer, as specific residue will need to be addressed.

Once a clean surface is achieved, the paths of the two processes differ: UV wafers typically undergo a "priming" step, where the wafer surface is dehydrated using heat and sometimes vacuum and then coated with a chemical primer, hexamethyldisilazane (HMDS), to maintain the dehydrated state and optimize adhesion of the photoresist to the wafer surface [1]. Dehydration is necessary as photoresist will preferentially attach to water on the surface, causing serious degradation of resist adhesion. HMDS application will contain any remaining water and greatly increase the wetting angle of the surface, rendering it hydrophobic. DUV wafers, on the other hand, typically do not undergo this priming step.

In DUV lithography, a bottom antireflective coating (BARC) is usually applied to the surface of the wafer by spin coating. During exposure, light penetrates progressively deeper into the photoresist until complete exposure is achieved. At that point, the light reaches the bottom of the resist and is reflected by the wafer surface. BARC is a thin polymer layer which is designed, when applied at the proper thickness, to absorb the



DUV Photolithography and Materials, Fig. 3 Standing wave dose swing of 63 % reduced to 1 % by BARC; 248 nm on Si

light which would otherwise be reflected back into the resist by the wafer. BARC is also used in some UV processing where topography of the wafer surface can cause "notching" or unwanted exposure due to reflected light, particularly in the case of metal lines on the wafer. Light can be reflected randomly by irregularities in the surface, especially by features that have been etched into the wafer, or by deposited material on top of it, creating a background dose which can affect feature sizes. In some cases, the reflected light can be focused onto an area of resist, exposing it and creating unexpected patterns. BARC is necessary in most DUV processing due to the high reflectivity of silicon (or almost any material) at DUV wavelengths. This reflectivity, in combination with additional internal reflections caused by the refractive index mismatch at the air-resist interface, forms a standing wave within the bulk of the resist. The resulting constructive or destructive total effect of this wave on the exposure dose makes CD control increasingly difficult as feature sizes shrink, and the impact on the dose becomes more significant. Therefore, BARC is used in most 248-nm processing and in all 193-nm processing. In fact, in advanced 193-nm processing, multilayer BARCs are common (Fig. 3).

Antireflective coatings are comprised of various polymer materials, usually linked to dyes to give them the desired wavelength absorbance. The polymers can be acrylic, vinyl, polyether, polyurethane, polysaccharide, or polystyrene, among others. The polymer is dissolved in a photoresist bowl-compatible solvent to allow easy integration into the processing line. There must be no intermixing with the photoresist applied after bake and no chemical interactions that might cause footing or undercut at the base of the resist feature. Some BARCs can be developed away with the resist (D-BARC) to simplify processing or for applications such as metal lift-off.

There are two important benefits of this step: the internal reflective interference is greatly reduced or eliminated, allowing better process control, and the adhesion of the photoresist to the surface (now BARC) is greatly improved, negating the need for any priming [2]. The inevitable trade-off is that the process requires an additional spin-and-bake step at elevated temperature (∼180–220 °C), and sometimes an

additional dry etch step after development to open the ARC at the bottom of the resist before the actual process step is performed. The etch chemistry can be O_2 plasma, or the same process that will be used in the following wafer etch step, but characterization of the etch rate is necessary. The thickness of the BARC must be adjusted to optimize performance with the existing film stack on the wafer at the time of exposure due to the interaction of the light with each of the layers and must be re-optimized for any changes in the films during each step in the process. Photolithographic computer modeling can be of great help here, narrowing the empirical testing parameters to acceptable size.

Another type of antireflective coating, top-layer ARC, or TARC, is also used in some situations. This layer works in much the same manner as BARC but is applied to the top of the resist layer. This layer can serve two purposes: it can act as an absorber for light reflected back from the wafer surface, reducing the standing wave intensity, and it can serve as a protective barrier layer for the photoresist, keeping any harmful chemicals from reaching it. In many immersion lithography applications (see DUV Exposure), it is used for this second function, preventing the immersion fluid, usually water, from leaching chemicals from the surface of the resist and degrading the resulting image.

Wafer Flatness

One important difference between UV and DUV processing is the flatness requirement of the wafer substrate. Silicon wafers start as slices from a long single-crystal ingot, called a boule. These slices are polished to the required thickness and flatness on either one side or both sides. The surface polish is usually required to be <1-nm RMS, with <0.5 nm being more desirable. Wafers polished on both sides (double polished) typically have the flattest surfaces and least thickness variation. As will be shown, flatness is important for successful DUV lithography.

The most stringent requirement for wafer flatness in UV lithography is 2-μm TTV (total thickness variation). The *least* stringent requirement for projection DUV lithography wafer flatness is 2-μm TTV and is 0.5-μm TTV for many critical processes. This more demanding requirement is due to the properties of projection optics outlined below.

The resolution of a reduction lens is calculated based on the Rayleigh criterion:

$$R = \frac{k_1 \lambda}{NA}$$

where

R = resolution of the system (1/2 of the smallest printable pitch)
λ = wavelength of exposure light
NA = numerical aperture of the lens
k_1 = process-dependent factor ($\geq$0.8 typ. DNQ, $\leq$0.6 typ. DUV)

However, the depth of focus of the same system is similarly calculated:

$$DOF = \frac{k_2 \lambda}{NA^2}$$

where

DOF = depth of focus (usable focus range on axis)
λ = wavelength of exposure light
NA = numerical aperture of the lens
k_2 = process-dependent factor ($\leq$1 typ.)

The first equation shows that smaller-wavelength light sources result in smaller features being printed (λ is directly proportional to R). It also shows that larger NA lenses result in smaller features being printed (NA is *inversely* proportional to R).

The second equation shows that smaller-wavelength light sources result in a smaller depth of focus (λ is directly proportional to DOF). It also shows that larger NA lenses result in a *much* smaller depth of focus (NA is *inversely exponentially* proportional to DOF) [3].

From the information above it is clear that the same mechanisms that produce smaller features also decrease the depth of focus, which is typically the most important limiter of process latitude. The

DUV Photolithography and Materials, Table 1 Various exposure wavelengths and their production capabilities at the time of the next wavelength technology introduction to HVM

λ (nm)	NA	Resolution (μm)	Total depth of focus (μm)	Overlay (nm)
436	0.38	1.000	3.020	±0.120
365	0.60	0.350	1.010	±0.050
248	0.75	0.130	0.340	±0.025
193	0.93	0.065	0.220	±0.012
193i[a]	1.35	0.038	0.106	±0.005

[a]Immersion, currently the leading-edge technology

DOF is the tolerance of the system to vertical variations in the target surface, which can include resist nonuniformity, chuck non-flatness or contamination, focus system noise, and physical vibrations, as well as wafer thickness excursions. In other words, the *difficulty* of printing features increases *exponentially* as the minimum pitch decreases linearly. If the parameters for the first generation of HVM 248-nm exposure tools ($\lambda = 248$ nm, NA $= 0.7$, $k_1 = 0.6$, $k_2 = 1$) are substituted into these equations, the following values result: resolution $= 213$ nm and depth of focus $= 506$ nm. Since the DOF figure represents the total available focus latitude, this might also be stated as ±253 nm. At the time of these tools' introduction, the typical *minimum* photoresist thickness was ~600 nm, which already exceeded the focus capabilities of the new DUV systems.

This loss of latitude in processing caused a paradigm shift in the lithographic process flow: from then on, photoresists would have to be increasingly thinner, and all surfaces requiring DUV exposure would have to be very flat regardless of their current state of processing. The industry would come to depend on chemical mechanical polishing (CMP) to provide the flat surfaces required for successful patterning of critical layers. CMP is a wet wafer polishing technique, where a slurry containing etch chemistries specific to the wafers to be polished is introduced to large rotating polishing pads. The polishing is performed using controlled pressure, continuous slurry flow, and constant motion of the wafer and pads to yield a uniformly flat surface across the wafer. Contamination is a constant issue in CMP processing, as any material left behind, such as copper, can contaminate an entire processing line. Post-CMP cleaning is, therefore, a critical step in the process flow. One example of innovative processing fostered by this flatness requirement is damascene metallization, where insulating material is deposited, patterned, and etched; then metal is deposited to overfill the etched features and finally CMP polished back down to the insulator, leaving a flat surface with metal in the previously etched areas. Many other variations exist, but the goal is that every process performed must result in a flat surface for the next lithography step.

Another important observation to be made from the above calculation is the powerful resolving ability of the photoresists used in DUV processing. The factor k_1 is an empirically derived number based on actual processing results. It can be influenced by many elements of wafer processing, such as uniformity, resist contrast, and development. The k_1 factor has a direct linear relationship to the system resolution. Note that the k_1 value is much smaller for DUV resists than for typical UV resists. This is due to the chemical amplification process described earlier. The above k_1 value is higher than typical values found in HVM today, which are often less than 0.4. This remarkable property of DUV photoresist enabled the printing of features *smaller* than the wavelength of the light used for exposure. In fact, 193-nm exposure tools were first introduced to HVM at around the 110-nm node, almost *one half* the exposure wavelength, a feat deemed impossible only a few years earlier (Table 1).

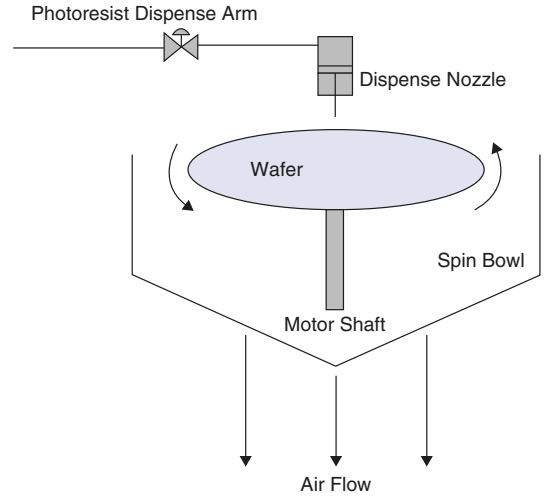
Photoresist Coating and Development

Spin coating of photoresist is universally performed due to the efficiency and uniformity achieved with this process. Other methods exist,

such as spray coating or meniscus, but they are reserved for special applications and do not typically appear in DUV processing. Wafer coating is performed by large, complex machines using precise control of dispense volume, spin speed, acceleration, and temperature parameters, to name only a few. While the tools used in UV processing are similar, the DUV versions are more tightly controlled in all parameters, have airborne chemical filtration (CA resists are sensitive to airborne contamination, e.g., ammonia and NMP), and require higher uniformity of hot-plate surface temperatures.

DUV photoresists are comprised of various materials including polyhydroxystyrene (PHS), acrylic polymer, and phenolic resin. The three general types of 248-nm photoresists are t-BOC PHS, ESCAP (PHS t-butyl acrylate), and acetal resin [4]. 193-nm resists can be polyacrylate, polymethylmethacrylate, or silsesquioxane resin, to name a few. The solvents used are primarily propylene glycol monomethyl ether acetate (PGMEA) and ethyl lactate, along with anisole and butyl acetate. Other materials are added to act as acid generators, acid inhibitors, and base quenchers, as well as surfactants to eliminate streaking during spin coating.

A typical resist coating process for DUV lithography has two primary steps: BARC coating and resist coating. Beginning with a properly cleaned surface, the wafer is spin coated with BARC, usually using a dynamic (wafer spinning) dispense process: a nozzle is moved out over the rotating wafer, and resist is released onto the wafer surface at a precise position, sometimes with continuous motion of the nozzle across the wafer, with a carefully controlled rate and amount. The wafer's rotation is then ramped up at a high rate of acceleration (5–30-k rpm/s) to the final casting speed of 2–6-k rpm, although usually the maximum is 3-k rpm for 200-mm or larger wafers. The volume of exhaust air from the spin bowl is adjusted to obtain the maximum film thickness uniformity across the wafer and to simultaneously obtain the fewest number of defects. The wafer is then baked on a hot plate at 170–210 °C for 60 s to remove solvent from the film and is then moved to a chill plate for 20–30 s (Fig. 4).



DUV Photolithography and Materials, Fig. 4 Resist coating schematic

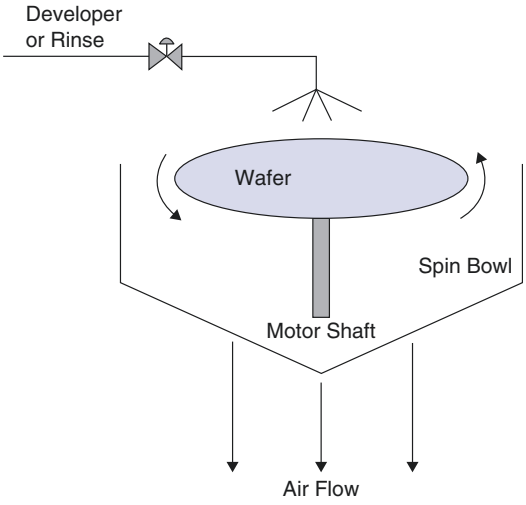
A hot-plate bake is analogous to baking in a convection oven but has some important benefits. The hot plate can bring the wafer into intimate contact for uniform thermal transfer across the wafer, often using vacuum for better contact between the wafer and the plate, and the process drives the solvent out of the film from the inside at a high rate. Oven baking causes the resist to form an outer skin, which thickens as the drying process continues at a decreasing rate due to the increasing skin thickness. A typical hot-plate bake time is 60 s, while an equivalent oven bake process may take 30 min or more and cannot maintain similar uniformity across all wafers in a batch.

The BARC thickness can be anywhere from 40 to 120 nm depending on the application. After cooling, the wafer is coated with photoresist using the same methods as before, but this time it is baked at 110–140 °C for 45–90 s to remove most of the casting solvent from the film, and then moved to a chill plate for 20–30 s to stop the bake process in a controlled way. All wafers must be baked for exactly the same length of time at exactly the same temperature for wafer-to-wafer repeatability to be maintained. The final thickness of the resist can be anywhere from <100 nm to >1 μm depending on the application. All of these process steps are performed by

automated modules utilizing tight control of dispense volume and bake time and temperature, as well as the temperatures of the chemicals and the individual parts of the tool.

After exposure, DUV wafers must be postexposure baked (PEB) to activate the acid amplification process. The PEB is a critical step, as the time and temperature both have a strong impact on the final CD and sidewall profile of the resist. Exposure of the resist by the tool creates a small amount of acid in the film but not enough to cause well-developed features to form. The PEB activates the acid amplification process, which greatly multiplies the volume of acid in the exposed areas, making the resist soluble in the developer solution. The acid profile in the resist is sensitive to both the time and temperature of the PEB, as they determine the final acid diffusion length as well as the final acid concentration in the resist. Once the wafer cools it is placed on a spin chuck, and a developer solution, usually tetramethylammonium hydroxide (TMAH) or recently tetrabutylammonium hydroxide (TBAH), is applied to the spinning surface by stream, spray, or other methods to form a puddle on the entire (now stationary) wafer. The developer remains in place for 30–60 s and is then rinsed off with deionized water as the wafer ramps up to speed and continues to spin until dry. There have been recent reports of using alcohols as the developer to produce a negative-tone image in the resist, which is very useful in 193-nm processing as there are no negative-tone 193-nm resists currently available. The amount of water and the length of the rinse time are crucial to minimizing the creation of defects on the wafer from this process. Uniformity and repeatability are also key components of a successful development process (Fig. 5).

Because of the remarkable sensitivity of CA photoresists, any variation in temperature at any position on the wafer can cause a change in the final CD at that location. It is not uncommon for a DUV resist to have a CD change of 1 nm/°C. That may not seem like much, but it can make up more than 2 % of current device CDs, which sometimes have a total error budget of 5 % or less. Such CD variations can wreak havoc with the performance of multi-gigahertz devices and significantly



DUV Photolithography and Materials, Fig. 5 Development schematic

reduce yields in the fab (fabrication facility or semiconductor factory). The uniformity of DUV hot plates is usually specified to be $\leq$0.5 % variation across the surface, which can be larger than a 300-mm square. DNQ resists used for UV processing are not nearly as susceptible to these variations as engineers typically only perform PEB to reduce standing wave profiles in the resist sidewall.

DUV resist processing spurred the introduction of proximity hot-plate baking, a method where instead of pulling the wafer into close contact with the surface using vacuum, the wafer is suspended a few tens of microns above the hot-plate surface using, for example, ceramic beads at the wafer edge. The purpose of this spacing is to eliminate across-chip line-width variations (ACLV) due to local changes in the bake temperature of the wafer that might be caused by contamination or other nonuniformity of the hot-plate surface. Longer bake times are required for this, but the resulting increase in across-chip and across-wafer CD control makes it worthwhile. Any variations in the CD from one area of a chip to another can decrease the maximum working speed of the device by skewing clock timings or even cause the device to fail completely.

The uniformities of the spin-coated film thicknesses are very carefully controlled. With 193-nm

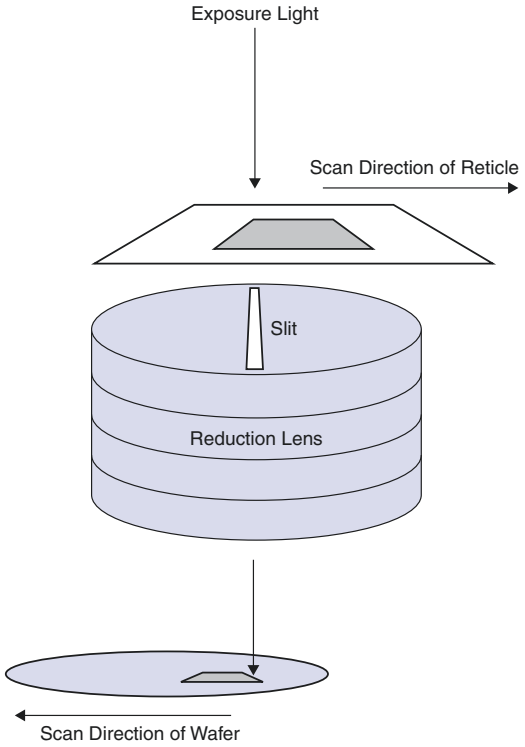
resist film thickness on the order of 100 nm and feature sizes reaching below 30 nm, there is no room for error. Any variation in the resist thickness (or the BARC thickness) will cause a measureable change in the CD at that location. For this reason, thickness variation across a 300-mm wafer is targeted to be <5 nm. Perhaps an even more difficult control issue is the level of defectivity. Defects are irregularities in the resist film which act to distort the exposed image. Some examples are particles on top of, underneath, or imbedded into the resist film. If defects are in the same location as device features, they can cause incorrect imaging of the pattern, leading to poor device performance or failure. They can be caused by many things, including the resist itself. Defect control is important in all photolithography, but because the CDs are so small in DUV processing, the problem is exacerbated. The allowable defect level for ≥30-nm defects on a 300-mm wafer is about $0.01/cm^2$. The introduction of immersion lithography (see DUV Exposure), where water is used between the exposure lens and the wafer surface, has created entirely new types of defects. Some of these have been solved by hardware engineering, while others have required material and chemical solutions. Some solutions include modification of the wafer coating system and the use of protective top coating layers.

Considering that many wafer fabs start more than 10,000 new wafers each week and that each wafer will undergo as many as 30 or more photolithography steps during the approximately 45-day process, it should be apparent that complete control of the coating process is essential to keeping CDs in spec and production profitable.

DUV Exposure

The tools used for DUV exposure are by and large the same as those used for g-line (436-nm) and i-line (365-nm) UV exposure. Economical HVM is only made possible by using the ubiquitous wafer stepper. This high-throughput, high-yield, high-uptime platform enables the incredible economies of scale that today supply the marketplace with billions of transistors at a cost to the end user of < $0.0000002 each. These aforementioned qualities are indeed required as the tools can cost anywhere from $7 M to more than $40 M each. Steppers use computer-controlled automated systems to load and pre-align wafers, compensate for environmental conditions, load and align reticles (photomasks used in projection exposure tools), step the wafers to each programmed location for alignment and focus measurements, expose the wafers, and unload the wafers to start the process again. The pattern image on the reticle is projected through the reduction lens and focused onto the wafer to print each image, which may contain multiple devices or die. Each die will eventually be cut from the wafer and packaged into a chip for use in an electronic system. The placement accuracy of each die is measured by a laser interferometry system which monitors the wafer stage position. The wafer is leveled at the start to maximize the available focus range across the surface. Each die location on the wafer is measured for focus before each exposure using grazing-angle light beams, and the individual die locations can be measured for accurate overlay of layer-to-layer patterns using through-the-lens grating comparison or off-axis diffraction systems. The location of each stepper is scrutinized with regard to environmental factors such as electrical noise and vibration, as these can have a detrimental impact on the printed images. All of these systems must work together at high speeds and with high accuracy while remaining operational >95 % of the time. This is, of course, a simplification of what have been described as the most complex machines ever built.

DUV steppers utilize additional systems to level the wafer for *each exposure location* to maximize available DOF, correct focus error to a few tens of nanometers, and correct individual exposure lens element positions to reduce distortion within the field. Air filtration systems remove airborne contaminants from the atmosphere within the tool to prevent degradation of the resist profile and damage to the optical elements. Motors are used to generate motion opposite to the motions of the tool caused by the stage jumping from die to die. Alignment tolerances for both the wafer and the reticle are tighter due



DUV Photolithography and Materials, Fig. 6 Wafer stepper-scanner schematic

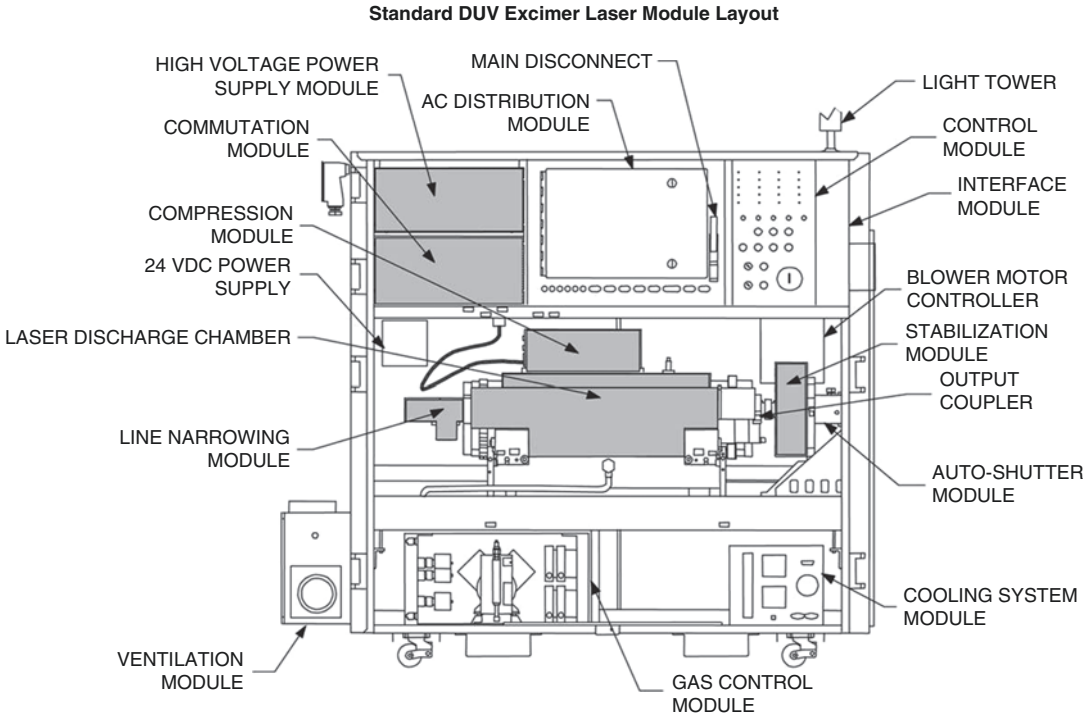
to shrinking feature sizes. At the same time, stage speeds and field sizes are increased to enhance productivity (Fig. 6).

The ever-increasing demands on DUV lithography for better alignment and lower distortion led to the introduction of the stepper-scanner systems now widely used in HVM. Scanners use a limited slit of the reduction lens for exposure, in order to optimize the best parts of the lens assembly. The reticle is scanned across this slit from end to end, while the wafer die location is scanned in the opposite direction under the lens. The result is a larger die than would be possible with a static system, with better distortion control and better overlay performance. The stage steps to the next die position and the scanning repeats in the reverse direction. Wafer stages can move at up to 700 mm/s, requiring counterbalances to move opposite the wafer and reticle stages to offset recoil, reduce stress on the factory floor, and dampen vibrations. Scanner throughput has reached 250 aligned and exposed 300-mm wafers per hour using two wafer stages in tandem: one performs wafer focus and alignment mapping, while the other steps and scans the wafer under the exposure lens.

Light Sources

Illumination systems for DUV exposure no longer use mercury arc lamps as the source, as the output intensity at DUV wavelengths is too low. Instead, excimer lasers provide high-power illumination with up to 90 W at 6-KHz pulse frequency and sub-half picometer bandwidth. Excimer lasers have proven to be very reliable, and constant improvements have made them increasingly powerful and cost effective. The laser can be located remotely from the tool, as far away as 20 m. The beam delivery systems utilize automated aiming mirrors in the nitrogen-purged beam line to maintain optimum positioning at the tool. A complex system of lenses, expanders, and homogenizers shape the beam and ensure uniformity across the reticle and from exposure to exposure. The laser uses a fluorine gas-filled chamber (KrF for 248 nm, ArF for 193 nm) with high-voltage electrodes to create the illumination source. The chamber gas depletes over time, and new gas can be injected to replenish the mix, until the chamber must eventually be flushed and refilled. These laser gas injections and refills are monitored and performed automatically by the stepper software to keep laser power and bandwidth within specifications. Laser maintenance is usually one of the less troublesome aspects of DUV lithography (Fig. 7).

The ramp-up of DUV photolithography also marked the acceptance of shaped illumination into HVM. Though proven for some time, this resolution enhancement technique (RET) was not often implemented in production environments. With the advent of sub-wavelength printing, shaped sources became a cost-effective enhancement to the wafer stepper. By using annular (off-axis) illumination, smaller features can be printed with greater depth of focus, albeit at the expense of (sometimes) throughput and printability of certain other feature sizes. Even greater enhancement is possible using quadrupole or, in the extreme, dipole illumination. One-



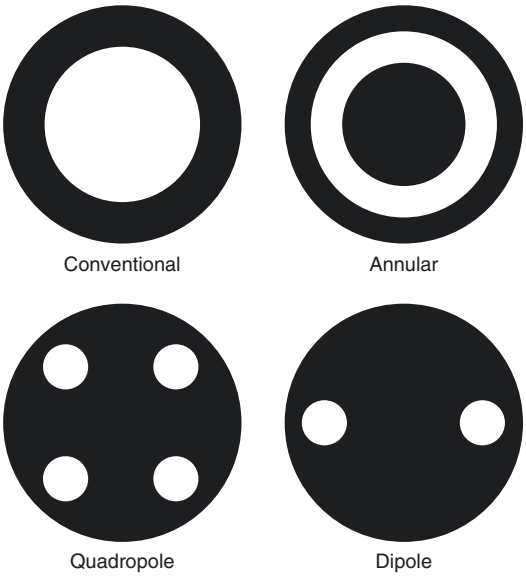
DUV Photolithography and Materials, Fig. 7 Two hundred and forty-eight-nanometer DUV excimer laser diagram

dimensional gratings can be printed at the diffraction limit of the system using binary masks, simply by inserting a shaped source. Programmability also was introduced, allowing any size annulus to be formed from keyboard commands and source shapes to be called up automatically. Today these mechanisms have been developed to the point of "free-form" source-shape capability. Arrays of thousands of mirrors are used to create any random shape calculated to optimize the process with any particular mask. The forms of both the source and the mask are calculated together in a technique called source mask optimization (SMO). Using free-form sources, the engineer is allowed almost limitless flexibility in process optimization (Fig. 8).

Polarization of the illumination is now one of the variables the operator can manipulate to optimize tool performance with regard to particular pattern orientations. Because of the high angles of incidence of the exposure light to the resist surface, light is transmitted or reflected by the photoresist depending on its polarization. Polarization can be set in orthogonal directions or in circular ones. Controlling the polarization of the exposure light can assist the engineer in controlling the printability of the resulting pattern.

Immersion Systems

The NA of exposure lenses has become so high, 0.93 in air, that the angles of light exiting the lens are as high as 70° from normal, leading to reflection of the light at the resist surface. The two polarizations of the light (TE and TM) will behave differently with one polarization reflecting away more than the other, and the penetrating light polarizations focusing at different depths within the resist. To overcome this, current leading-edge system vendors have implemented immersion systems at 193 nm, where purified water is used as the transport medium between the exposure lens and the resist. The water fills the space between the



DUV Photolithography and Materials, Fig. 8 Illumination source-shape examples

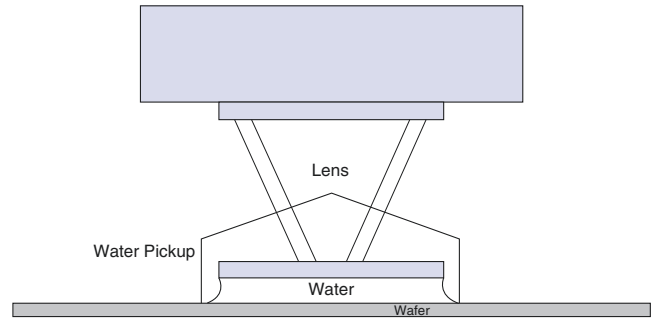
lens and the wafer surface, creating a higher index medium for the light to pass through. The higher index of the water is a better match to the indexes of both the lens glass and the photoresist, reducing both the exit angle of the light leaving the lens and the angle of incidence of the light entering the resist, thus reducing the reflectivity and polarization losses. The water must be injected into the space without wetting any other parts of the system and must be kept only in the lens-wafer space throughout the exposure process. At the end of the exposures, it must be removed from the wafer without leaving any residue. To prevent defects the water must be bubble-free and leave no droplets behind on the surface, even though the wafer moves at very high speeds. Elaborate systems consisting of de-gassers, showerheads, and pickup rings are used to inject and remove the water from the moving wafer surface (Fig. 9).

This additional complexity, introduced by immersion and polarization, drives the cost of tooling up but also allows features to be printed on a half pitch as small as 38 nm in a single exposure. No smaller pitch can be resolved using these wavelengths unless a significantly higher refractive index medium can be found to replace the water currently in use, a doubtful prospect.

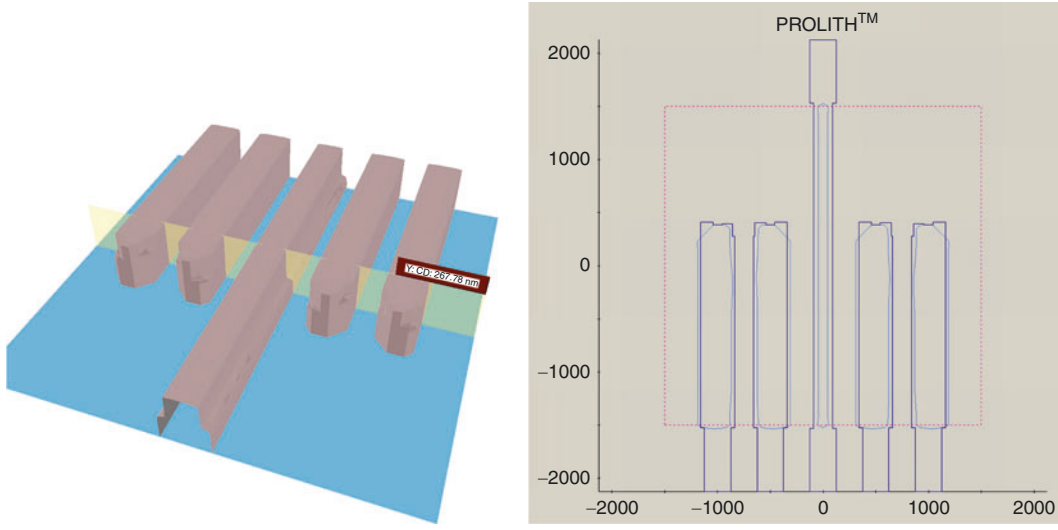
Mask Making

During the 1980s, mask making for the newly introduced reduction UV steppers initially relaxed the requirements for mask making compared to contact lithography, leading to what has been called the "mask makers' holiday," which lasted for some years. With features five times the size of the wafer CD, and only one field to make instead of the entire wafer, making acceptable masks presented a relatively simple task. With the introduction of DUV exposure at 4 × reduction, this holiday was abruptly ended. The industry struggled to meet increasingly demanding specifications for feature size and placement, line-edge roughness, and defects. Shortly thereafter, even greater pressure was brought to bear by the inclusion of optical proximity correction (OPC) features, which are assist features on the reticle sized below the resolution of the system, and phase-shift masks (PSM), which have some clear areas of the mask etched to a depth that causes the light passing through them to exit the mask 180° out of phase with the rest of the light. OPC may include assist features such as scattering bars or serifs, which are sub-resolution shapes that act to force the diffracted light into desired shapes on the wafer. Phase-shifting areas can be used either to define features or to improve the fidelity of the printed pattern. By placing out-of-phase light at the edge of a feature, the location in the resist can be precisely defined. Critical level masks can today have features nearly as small as the wafer CDs, making masks very difficult and expensive to manufacture.

The addition of OPC and PSM to a mask can cause the volume of the data necessary to write the masks to explode, making file size and data rates problematic. The tools used to write the masks, laser and e-beam writers, can cost $10 M or more and can take many hours to complete each mask. Inspection and repair add more time and effort to the task. Using these RETs in manufacturing, one single mask can cost $60 K or more. Given that a mask set for a device may require 30 or more masks, with perhaps 10 critical levels, mask costs can surpass the $1 M per set mark. Frequent revisions in production, sometimes known as "re-spins," can



DUV Photolithography and Materials, Fig. 9 Immersion scanner schematic



DUV Photolithography and Materials, Fig. 10 Computer model output of resist profile and mask OPC

obviously get very expensive. This in turn has driven the use of modeling in manufacturing.

Computer Modeling

Simulation of photolithography has been ongoing for around 30 years, initially in an effort to better understand and explain the process and then later as a way to optimize processing parameters. It has today evolved to the point where models can quickly and accurately predict not only the resulting resist pattern but also the optimum illumination shape and mask pattern corrections. By modeling the process, expensive hours of DUV tool time are saved by greatly reducing the on-wafer testing required to explore the process parameter space. Dose, focus, and illumination settings can all be determined before even making the mask. Simulation of the process over the entire chip can be done to locate problems and optimize yield before "tape out" of the mask design (submission of final data to the mask vendor), saving enormous time and effort and thus cost. Current applications of computer power leverage the enormous data volumes generated in manufacturing to model systems and extract error contributions from individual system components such as the reticle, the lens, and even the wafer chuck. Once these values are measured, corrections can be applied to reduce or eliminate them (Fig. 10).

The resolution of DUV exposure systems has now been pushed so far into previously unknown,

nonlinear regimes that the accepted concepts regarding refractive imaging systems cannot predict their behavior. The patterns that need to be exposed on the wafer cannot be defined in traditional ways due to previously unseen phenomena. Modeling and simulation have now allowed lithographers to reverse engineer the optical image, creating a new approach called inverse lithography technology (ILT). A mask made using ILT may look nothing like the final printed pattern, but it is a collection of shapes which will, when used in conjunction with the complementary illumination source shape, cause a diffracted pattern that will ultimately form the desired image in the photoresist. The SMO approach combines ILT with variable shaping of the light source, in some cases with single pixel addressing. These techniques have allowed engineers to push the resolution of exposure systems to the limits of diffraction with acceptable control.

One of the major mask-making issues that came to light in DUV lithography has to do with an unexpected effect that occurs when printing feature sizes below the wavelength of the exposure light. During exposure processes with low k_1 values, where the process becomes nonlinear, defects on the mask are printed on the wafer at dimensions larger than would normally be expected. This mask error enhancement factor (MEEF) can in some cases exceed 4 × or more, magnifying any error into a potential "killer" defect. Modeling these potential errors has led to better mask inspection methods focusing on the conditions particular to creating MEEF in order to prevent unexpected pattern failures. Simulations are also used to compute full-chip OPC to maximize the process window and increase yield. Without this computational modeling power, the advancement necessary to maintain "Moore's law" would have lost momentum long ago.

Inspection

One of the most difficult aspects of DUV processing is the inspection process. Trying to determine exactly what is on the wafer after lithography is much more difficult than in typical UV patterning, due to the much smaller feature sizes and overlay requirements. DUV lithography has driven the rapid advancement of optical inspection tools utilizing actinic (248-nm, 193-nm) illumination for both wafers and masks. Scanning electron microscopes (SEM) are routinely used in HVM to check in-line wafers, which can then feedback and feed forward data to other line processing tools. The information is then used to adjust processing parameters to compensate for measured results and improve yield. Atomic force microscopy (AFM) is also useful for determining surface characteristics and resist profiles. In recent years, a new method, called scatterometry, has gained popularity for measurement of dimensions which are difficult to resolve using standard tools. This method has the advantage of not requiring the formation of an image of the target to calculate its dimensions, something which can be impossible to do with light and is increasingly challenging with any tool at the scale of leading-edge devices. SEM inspection also has an unfortunate side effect on DUV resist: shrinkage. The area under inspection by the electron beam will actually change size through the process of inspection, making any resulting measurements questionable.

Another unique issue has also appeared in 193-nm DUV: the formation of a haze on the chrome side of reticles is causing new inspection techniques to be investigated. This haze can, over time, reduce transmission of the reticle affecting CD and even causing placement errors. New cleaning techniques are under investigation in order to counter this problem. Reticle inspection in general has undergone rapid development as 248- and 193-nm tools have advanced, in the quest to keep pace with shrinking CDs and exploding data sets.

Future Issues

One recurring theme in the descent to molecular-scale devices has been the battle against line-edge roughness (LER). LER may become the ultimate limiter of photolithography as the actual feature sizes approach the magnitude of the roughness. It

manifests as small variations in sidewall linearity or smoothness, either vertically or horizontally. These variations affect the line width of device structures after pattern transfer, making control of the performance of the device difficult or impossible. When the uniformity of the printed line width is beyond the control of the lithographer, working devices become unreliable or unattainable. The operational speeds and the timing of events within the devices depend on the uniform size of its structures. Once this uniformity is compromised, delicate timings are changed and device operations may fail.

The root cause of these variations in the sidewall plane has not definitively been explained by modeling, and ongoing work includes both investigations into the causes of LER and the development of new materials and techniques to reduce or eliminate it. Materials such as molecular glass are intended to reduce the size of the polymer molecules in an effort to reduce the width variations. New methods and materials for development of the resist after exposure are also possible solutions. LER can be reduced by decreasing the sensitivity of the resist, thus increasing the required dose, but this would result in an unacceptable decrease in throughput.

High-index immersion materials have been investigated as a way to extend the usefulness of 193-nm tools. Replacement of the water used in immersion lithography systems with higher index fluid could extend 193-nm exposure resolution for another node or two at most. Possible candidates have usually included fluorinated liquids, but most have been dismissed as not having high enough index to be worth the investment to make them viable.

Current work is focused on double-patterning scenarios, where two mask patterns are exposed in the same layer. The use of two separate patterns solves the problem of shrinking pitch values but requires either a resist that "forgets" the unwanted effects of the first exposure by "freezing" the image or that an etch step be performed before the second mask is exposed. The latter method places demanding overlay requirements on the scanner as well as a heavy burden on the mask makers. This level of precision is unprecedented, with error tolerances in the range of 2–3 nm for layer-to-layer registration. Another effort is concentrating on "spacer" patterning, also known as pitch splitting. This is a method of self-aligned patterning where the wafer is exposed and developed, a new layer of material is deposited, and then the material is etched back to form the new structures. This produces a pattern with about ½ the original pitch. This technique has been used by memory chip manufacturers for some time and has now evolved to triple and quadruple versions as the industry waits for EUV to finally reach readiness for HVM.

Conclusion

DUV photolithography is a dynamic and rapidly advancing technology. The race between manufacturers to make the smallest, fastest, and cheapest devices is seemingly endless. However, they may be approaching the end of their abilities to make things smaller. EUV lithography (13.5 nm) is preparing to enter the production phase by the 10-nm node, but some uncertainty still exists. Even if it does come into the fab, how long will it be useful? Current predictions are that it could last until the 5-nm node, if it is ever reached. At the same time, multiple e-beam and imprint technologies, as well as directed self-assembly, are still in the running. If they do work, will anyone be able to afford these technologies?

One thing is certain: as soon as this entry goes to press, it will become outdated. The latest advancements will change the direction of the industry, probably in unforeseen ways. In order to ascertain the current implementations of DUV lithography, the reader should investigate the SPIE Advanced Lithography Conference proceedings published since this entry was updated (2014), as well as search the publications of SEMATECH, ITRS, IEEE, BACUS, and IEDM. DUV photolithography will likely remain in favor longer than many would imagine, but the leading edge will eventually pass it by.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Immersion Lithography Materials

References

1. Dammel, R.: Diazonaphthoquinone-Based Resists. SPIE, Bellingham (1993)
2. Mack, C.A.: Field Guide to Optical Lithography. SPIE, Bellingham (2006)
3. Sheats, J.R., Smith, B.W.: Microlithography Science and Technology. Marcel Dekker, New York (1998)
4. Helbert, J.N.: Handbook of VLSI Microlithography. Noyes, Park Ridge (2001)

Dye Sensitized Solar Cells

- ▶ Nanomaterials for Excitonic Solar Cells

Dye-Doped Nanoparticles in Biomedical Diagnostics

Vladimir Gubala
Biomedical Diagnostics Institute, Dublin City University, Glasnevin, Dublin, Ireland

Synonyms

Dye-doped nanospheres in biomedical diagnostics

Definition

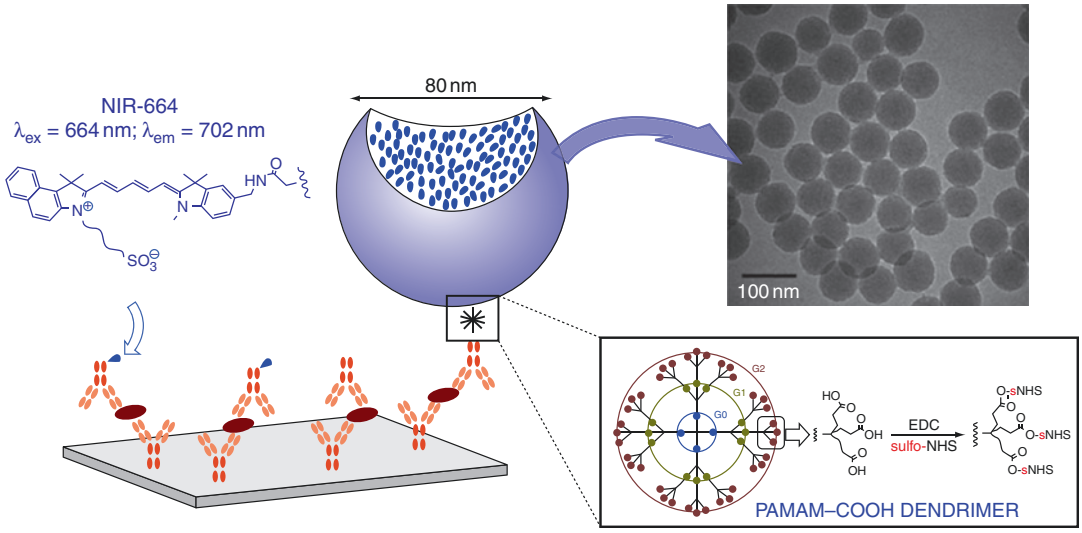
Dye-doped nanoparticles are new class of fluorescent labels. They are typically made of silica or polystyrene with tens of thousands of fluorescent dyes entrapped in the pores of its polymer matrix.

Overview and General Concept

When designing devices for biomedical diagnostics, it is essential to maximize the signal to noise ratio to achieve clinically relevant sensitivity and limits of detection. Simple assay designs involve first capturing an antigen onto a surface using one antibody, then measuring the surface concentration by visualizing the captured antigen through its reaction with a second, labeled antibody. There is increasing interest in fluorescence-based array sensors (biochips) for biomedical applications. While fluorescence detection offers high sensitivity, there is generally a low level of fluorescence signal from the biochip due to a monolayer of fluorescent labels, hence the importance of enhancing the fluorescence. In antibody-based assays, the measured signal can be amplified through the replacement of molecular luminophore with doped nanoparticles (NP) (Fig. 1).

Fluorescent labels are used for a range of applications including immunosorbent assays [1], immunocytochemistry [2], flow cytometry [3], and DNA/protein microarray analysis [4]. Current research in biomedical diagnostics is moving to inexpensive devices, using biochips that often have to measure biomolecule concentrations at the picomolar level in blood sample volumes on the microliter scale, without the possibility of any user manipulation of the sample. This is exceptionally challenging to achieve in a miniature bioassay device. First, the active biomolecules (antibody or nucleic acid) must be immobilized at the appropriate surface density while maintaining its reactivity. Second, the measured signal should be maximized relative to noise and background contributions. In antibody-based assays, nonspecific binding (NSB) controls the background response and often results in a detection limit that is much higher than that defined by the equilibrium constants for the binding events.

To meet these new demands, fluorescent labels with improved physical and chemical properties are required: for example, high photostability and fluorescence intensity, with a reproducible signal under a variety of chemical and biological conditions. Excellent photostability and significant



Dye-Doped Nanoparticles in Biomedical Diagnostics, Fig. 1 A cartoon illustrating the concept of fluorescence-linked immunosorbent assay using the detection antibody labeled with molecular fluorophore (NIR-664 in this particular case) and 80 nm nanoparticles, doped with tens of thousands of molecular dye molecules, thus significantly increasing the brightness of the label. The TEM image shows the high monodispersity and size-uniformity of 80-nm silica nanoparticles doped with NIR-664, prepared by reverse microemulsion method. Inset with PAMAM dendrimer highlights a unique modification of the NP surface that ensures high colloidal stability and efficient bioconjugation to biomolecules of interest

reductions in photobleaching is required for applications where high intensity of prolonged excitations are applied, such as intracellular optical imaging. Dye-doped nanoparticles (typically silica or polystyrene) stand out as excellent candidates as it is possible to dope silica NPs with a large number of fluorophores, increasing the total fluorescence of the label significantly. Despite the effect of fluorescence quenching phenomena within an NP with a large amount of dye incorporated in a small volume, the goal of producing a particle with brighter fluorophores protected inside an organic matrix, thereby increasing photostability and quantum efficiency is largely successful. Owing to their intense signal, fluorescent nanoparticles (NP) are useful as labels since they can be measured directly, without the need for any amplification step.

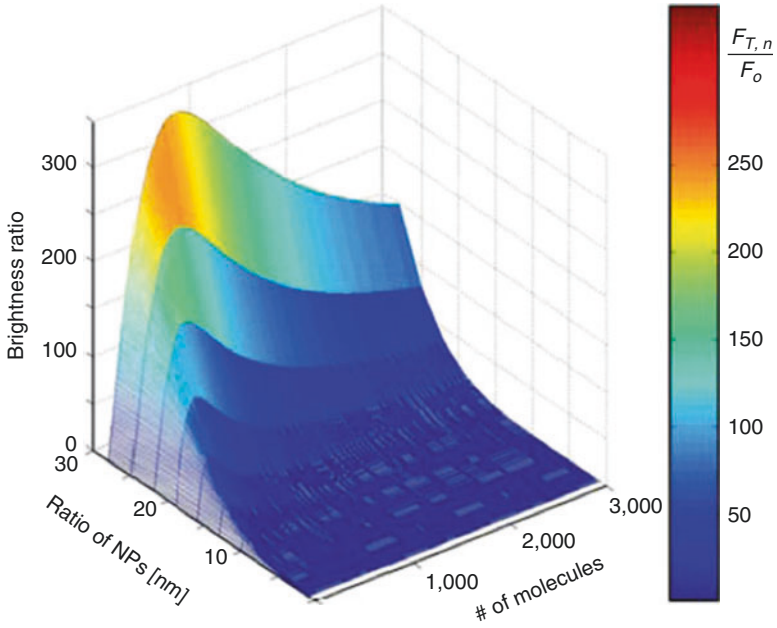
How to Select Appropriate Dye-Doped NPs for Your Application?

Dye-doped NPs are currently offered in a wide range of sizes, with entrapped fluorophores covering most of the useful spectrum of UV/V radiation and with a surface modification to contain chemically reactive groups, usually –COOH and $–NH_2$. Most of the scientists using dye-doped NPs will make their decision based on the characteristics of the instrument (e.g., excitations source and emission filters) and the type of application.

Typically, at least three parameters should be defined when selecting the appropriate dye-doped NP:

(a) What is the desired fluorescence property of the NP for your application?

The choice of the luminophore, particularly its excitation and emission wavelengths, obviously depends on the capacity of the instrument that is used to excite the fluorophores and collect the emitted light. However, it also depends on the application. At near infrared wavelengths there is low background interference from the fluorescence of biological molecules, solvent, and substrates. For example, whole blood has a



Dye-Doped Nanoparticles in Biomedical Diagnostics, Fig. 2 A computer model showing the correlation between the brightness ratio and the radius and number of dye molecules entrapped inside NP

weak absorption in the NIR region, thus reducing the need for whole-blood filtering for assays using whole blood. NIR light can also penetrate skin and tissue to several millimeters and this can enable fluorescence detection in dermatological or in vivo diagnostic devices. When preparing dye-doped NPs, it is also important to remember that not every readily available dye can be doped into the NP matrix. This will be further discussed later in this entry.

(b) What is the desired size of the NP for your application?

In case of bioassays, the answer to this question is specific to the design of the assay. While it is tempting to think that the brightness will increase with increased size of the NP, one must also consider the increased surface area and different monolayer packing arrangement of larger NP when compared to smaller ones.

It is not necessarily true that by increasing the diameter of NP say by an order of 2, the fluorescence will also increase by the same order (Fig. 2). In general, dye-doped NPs in the range of 20–200 nm were successfully used in bioassays in the wider research community. As for optical imaging, the rate of uptake of nanoparticles by mammal cells depends on the particle size and shape. A number of articles emerged recently reporting that the optimum size of nanoparticles to enter the cells is around 50 nm. The cellular uptake of $\sim$ 50 nm NP is typically the highest and the NPs were shown to be trapped inside vesicles in the cytoplasm; however they did not enter the cell nucleus. Shape also plays an important role, with spherical nanoparticles being taken in larger numbers than rod-shaped nanoparticles. This is presumably due to the difference in curvature affecting the contact area with the cell membrane receptors.

(c) What is the desired surface modification for efficient conjugation with biomolecules?

It is relatively easy to functionalize silica or polystyrene NPs with reactive groups that enable bioconjugation. However, each step in the conjugation process modifies the zeta potential of the NP, thus affecting colloidal stability. Proteins are usually immobilized on the surface of polystyrene beads by physical

adsorption. It is a relatively fast method for formation of reasonably homogeneous protein layer. For better results, it is often desirable to block the non-active surface area with bovine serum albumin or similar blocking agents. In case of silica particles, the colloidal stability and the density of the active molecules on the surface can be fine-tuned by carefully executed surface chemistry. Some aspects of the surface modification strategies of dye-doped silica NP will be discussed later in this entry.

Basic Methodology

Synthesis of Dye-Doped Silica Nanoparticles

One obvious advantage of using silica particles as opposed to polystyrene is the significantly higher density of fused silica (2.2–2.6 g/cm^3), which facilitates the manipulation and purification of samples (centrifugation at 15,000 rpm for 1–2 min is usually sufficient to separate the beads from supernatant). Also, silica NPs are relatively nontoxic and chemically inert. They can be prepared in a range of sizes and the post-synthetic surface modification can be achieved with good reproducibility.

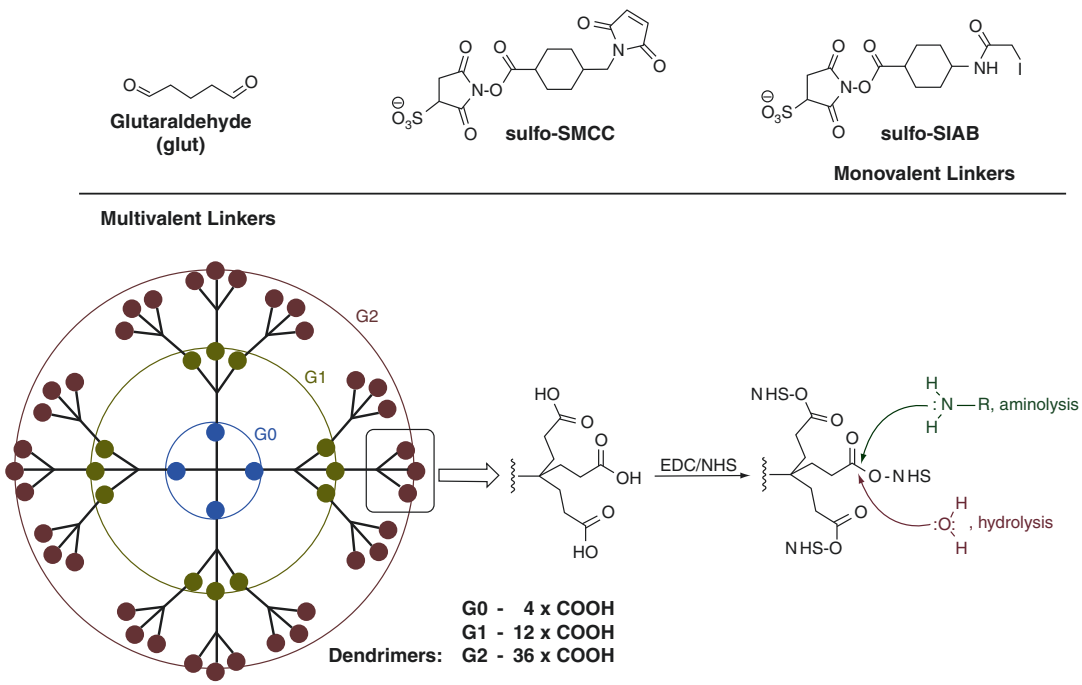
There are two main methods for synthesis of dye-doped silica particles.

1. Sol–gel technique via Stober method, which consists of the hydrolysis of a silica alkoxide precursor (such as tetraethylorthosilicate, TEOS) in an ethanol and aqueous ammonium hydroxide mixture.
2. *Reverse microemulsion* method based on a water-in-oil reverse microemulsion system.

Both methods have certain advantages and disadvantages. For example, the former method is comparatively simple and both organic and inorganic dyes can be incorporated in the silica matrix. A modified Stober method has been used to produce highly fluorescent dye-doped core–shell silica NPs with narrow size distribution by the Wiesner group at Cornell University (NY, USA) [5]. The latter technique is using stabilized water nanodroplets formed in the oil solution. Silane hydrolysis and the formation of NPs with the dye trapped inside occur inside such small microreactors. The size of the NP is determined by the nature of surfactant, the hydrolysis agent, and other parameters, such as the reaction time, oil/water ratio, etc. The resulting NPs usually show a high degree of uniformity and water dispersity. One of the limitations of this method is that, in most cases, it works well with inorganic dyes, some of which have lower quantum yields compared to organic fluorophores. By both techniques, the dye molecule can be either physically adsorbed inside the pores of NP or covalently linked to its matrix. The second option is preferred, as the non-covalently attached fluorophores can eventually diffuse or "leach out" into a solution over time, thus reducing the overall brightness of the NP.

Surface Functionalization of Dye-Doped NPs

In general, the benefits of using dye-doped NPs can only be realized if they are efficiently coated with biorecognition elements (in most cases antibody, oligonucleotides, etc.), have good colloidal stability, and the ratio of specific to nonspecific binding (NSB) is sufficiently great. The colloidal stability of NPs can be modulated by carefully executed surface modification. Sufficient separation of NPs can be achieved by increasing either steric or electrostatic repulsions between two or more adjacent NPs. Current approaches to prepare colloidal solutions with low polydispersity index are mostly based on electrostatic interactions. In polystyrene particles, the introduction of charge on the surface can be done by copolymerization reaction of polystyrene monomer and its charged analogue. And even then, a longer-term colloidal stability is guaranteed only after addition of polar surfactants into the sample. Quite often, this surfactant must be removed prior the immobilization of the antibody or DNA. On the other hand, the flexible silica chemistry provides versatile routes for surface modification. Different types of functional groups can be easily introduced onto the NP surface enabling conjugation with biomolecules of interest. The post-synthetic modification of the particle surface with charged species (e.g.,



Dye-Doped Nanoparticles in Biomedical Diagnostics, Fig. 3 Structures of the most popular monovalent linkers and a cartoon illustrating the multivalent feature of a Newkome-type dendrimer, activated by a dehydrating agent (This figure was reprinted with permission of Elsevier)

phosphates, carboxylic acid, etc.) is not only aimed to reduce particle aggregation but also to provide some sort of reactive handle, where the reaction with antibodies takes place. Most of the biomolecules used as molecular recognition elements are charged species (proteins, DNA, etc.) and contribute positively toward the decrease of the polydispersity index of the sample.

One of the most elegant and efficient methods to achieve good colloidal stability of silica NPs is to introduce a mixture of charged precursors on the particle surface. The first one is usually chemically inert with low pKa value, assuring that the charge is present when working at physiologically relevant pH (7.0–7.4). The second one, typically at much lower quantity provides reactive group for the reaction with biomolecules or other cross-linking agents. In 2006, Tan and colleagues [6] proposed a way of improving the colloidal stability of silica NPs through the addition of a negatively charged, nonreactive alkoxysilane with phosphate group along with its counterpart carrying a reactive amine. Other very common groups that can be used as reactive "handles" include carboxylic acids, isothiocyanates, and epoxy group. Normally, the ratio between the nonreactive and reactive groups is relatively high (10:1) in order to maintain the suspension stability. As a consequence, the number of attachment sites on the NP surface is therefore limited. For this reason, it might be quite difficult to achieve high protein coupling ratios. Gubala et al. [7] overcame this disadvantage by using multivalent linkers such as dendrimers as antibody scaffolds (Fig. 3).

Bioconjugation Strategies

As mentioned previously in this entry, the most straightforward method to immobilize large biomolecules like antibodies is through physical adsorption [8]. This method is very popular with polystyrene beads; however, it does not offer direct possibility of modulating the density of the captured protein on the NP surface. The

control over the active surface area, which is the area of the NP surface covered with antibodies, can be achieved by using cross-linking reagent.

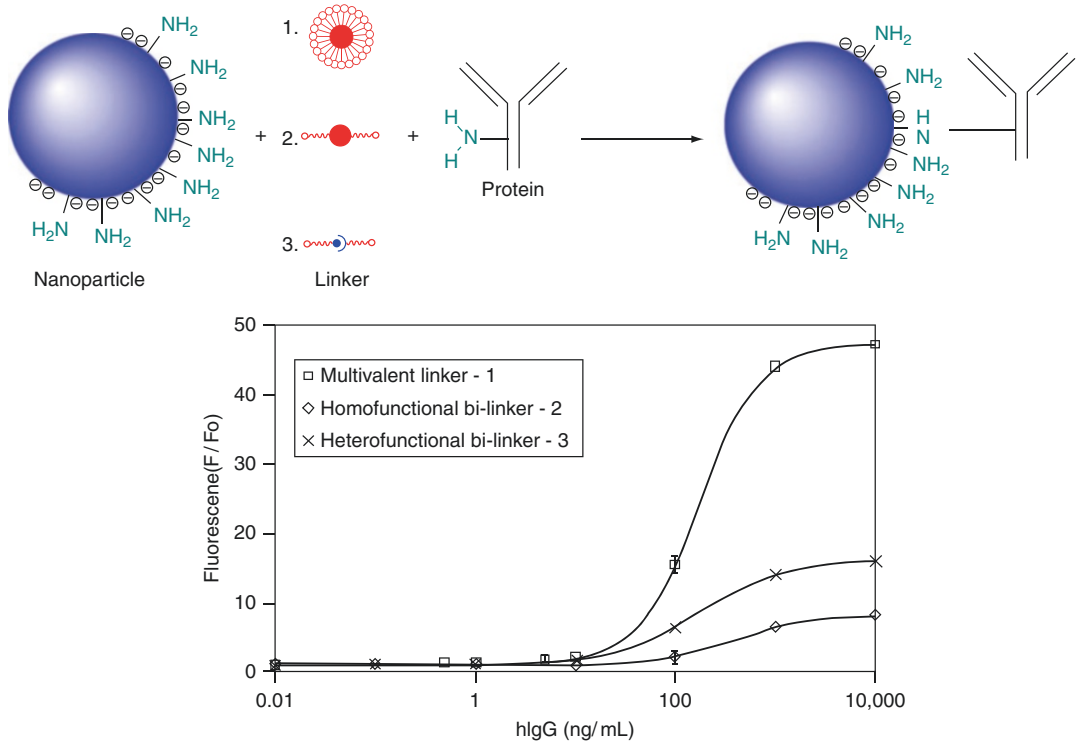
There is a great choice of commercial hetero- or homo-functional linkers available; however, their effect on NP stability, aggregation, solubility, and efficiency of bioconjugation was until recently poorly documented. This was striking, particularly when considering that this is the actual layer "sensed" by the molecule to be labeled.

In a sensitive and functional bioassay, one obvious issue in the use of NPs as labels is the fraction of the coupled antibody that is in fact active or available for reaction with antigen. This fraction can be rather small, which in turn can lead to diminished sensitivity and increased nonspecific binding. The minimization of NSB is essential for sensitive detection in an assay. Thus, it is clear that the strategy used to attach an antibody to NP surface is a key element that affects the activity of the bound antibody, the nonspecific binding, and the surface binding of particles.

Improved colloidal stability of silica NPs has been previously achieved by the addition of negatively charged nonreactive organosilanes in addition to organosilanes with functional groups available for bio-immobilization. The number of attachment sites on the nanoparticle surface is limited in order to maintain the suspension stability of the NP system. For this reason it is difficult to achieve high protein coupling ratios. Gubala [7] and colleagues at Biomedical Diagnostics Institute have recently demonstrated some advantages of using multivalent molecules such as dendrimers as antibody coupling scaffolds as illustrated in Fig. 3. Dendrimers are monodisperse, nano-sized, hyperbranched "starburst polymers," with growing numbers of terminal functional groups with increasing generation number that have been extensively investigated for a variety of biomedical applications. By changing the generation of dendrimers, the authors were very efficiently able to control the surface area populated by biomolecules. The authors demonstrated that antibody-sensitized NPs prepared by using a multivalent linker showed a significantly lower limit of detection and higher sensitivity than with the homo- and hetero-functional cross-linkers (Fig. 4). The difference in performance was remarkable, particularly when compared with glutaraldehyde, one of the most common cross-linker still employed by many research laboratories. They reasoned that the multivalency of the dendrimers is one of the most significant factors behind the increase in the detection sensitivity. Dendrimers had a positive effect on NP stability and aggregation, were more reactive with biological materials, and were capable of immobilizing antibody at the appropriate surface density while maintaining its activity to the analyte of interest. Moreover, the multivalency of the dendrimer is a significant factor responsible for the improvement in the reaction yields even at lower protein concentration. This is very important, particularly when considering the cost of the bioorganic material that is used in bioconjugation reactions.

Dye-Doped Nanoparticles and the Implications on Assay Kinetics

A surface capture immunoassay can be considered in two ways: either incubation of antibody-sensitized particles with antigen followed by capture onto an antibody-sensitized surface of those particles that have bound antigen; or capture of antigen onto an antibody-sensitized surface followed by capture of antibody-sensitized particles onto the surface-bound antigen. In either case, the kinetics of reaction of a sensitized particle with a sensitized surface is a key element of the description of the process and hence of the assay design. The size of the particles and the fraction of the particle surface that is active for the antibody–antigen reaction are the important parameters. A simple kinetic model was developed for particle capture onto an antigen-loaded assay surface by means of an antibody–antigen reaction. This is similar to kinetic formulations treating single molecules, but also accounts for the fact that each particle has a multiplicity of antibodies present on its surface. The NPs-surface-bound antibodies can be considered as active binding sites. Depending on the coverage of such active sites on the particle, a collision of the particle with an antigen may or may not lead



Dye-Doped Nanoparticles in Biomedical Diagnostics, Fig. 4 Bioconjugation strategies that rely on the use of cross-linking agent. The content of $–NH_2$ (or $–NH_3^+$) groups on the surface of the NP is typically limited to ~ 10 %. Higher amount of amines can cause aggregation. The lower density of reactive amines is offset by the multivalency of dendrimers. Their effect is threefold: (**a**) the use of dendrimers positively affects the colloidal stability of the NPs, (**b**) the reactivity of each amine on the surface is amplified by the number of dendrimers' surface groups, hence significantly improving reaction efficiency with usually expensive biomolecules, (**c**) dendrimers, due to their size, not only serve as linkers but also as spacers between the surface of the NP and the captured recognition elements, thus maintaining the activity of antibodies, oligonucleotides, etc. The net result is more sensitive assays when compared to the monovalent linkers and also improved assay kinetics

to reaction, with a probability proportional to the surface coverage of active sites on the particle and the surface coverage of antigen in the reaction well.

The capture reaction, of antibody-functionalized NPs onto the assay surface, can be written as:

$$P + S \rightleftarrows PS \qquad (1)$$

with forward rate constant k_{on} and reverse rate constant k_{off}. Here, S denotes an active antigen site on the assay capture surface. If the NP is bound to the surface by a single antibody–antigen interaction, then k_{off} will simply be the dissociation rate constant for the antigen–antibody complex. The association rate, k_{on}, will be determined by the probability of a reactive collision between an antibody-functionalized NP and an unoccupied reactive antigen site on the assay surface. A reactive collision, leading to coupling of the NP to the assay surface, would occur when a reactive part of the NP surface (an antibody, oriented with the binding site exposed) collides with a reactive part of the assay surface (an antigen, oriented with the epitope exposed). NPs are in a continuous state of collision with the assay surface, at a rate determined by the diffusion coefficient of the NPs (hence by their radius and the viscosity of the

reaction medium) and by their concentration. As the NP concentration on the assay surface builds up, particles not only occupy sites but also physically block the assay surface area, on account of their size. The probability of obtaining a successful reactive collision between particle and capture surface also depends on the fraction of the capture surface that is active for reaction. Therefore, the probability that the collision will be with an active site on the assay capture surface will be a product of the following:

- The fraction of the total assay capture surface that is covered by accessible antigen with the binding epitope correctly exposed to facilitate antibody binding
- The fraction of the total assay surface that is unblocked by bound NPs
- A coverage-dependent factor that expresses the requirement that a NP requires a space whose smallest dimension is at least as large as the NP diameter, in order that the NP can fit into the space

The considerations of reaction probability are embedded in the capture rate constant, k_{on}, which depends on both the fraction of the surface of the antigen-loaded capture plate that is in fact active for the capture reaction (i.e., it depends on the surface state of the adsorbed antigen) and on the fraction of the surface of the antibody-sensitized NP that is active for reaction. It can be written as:

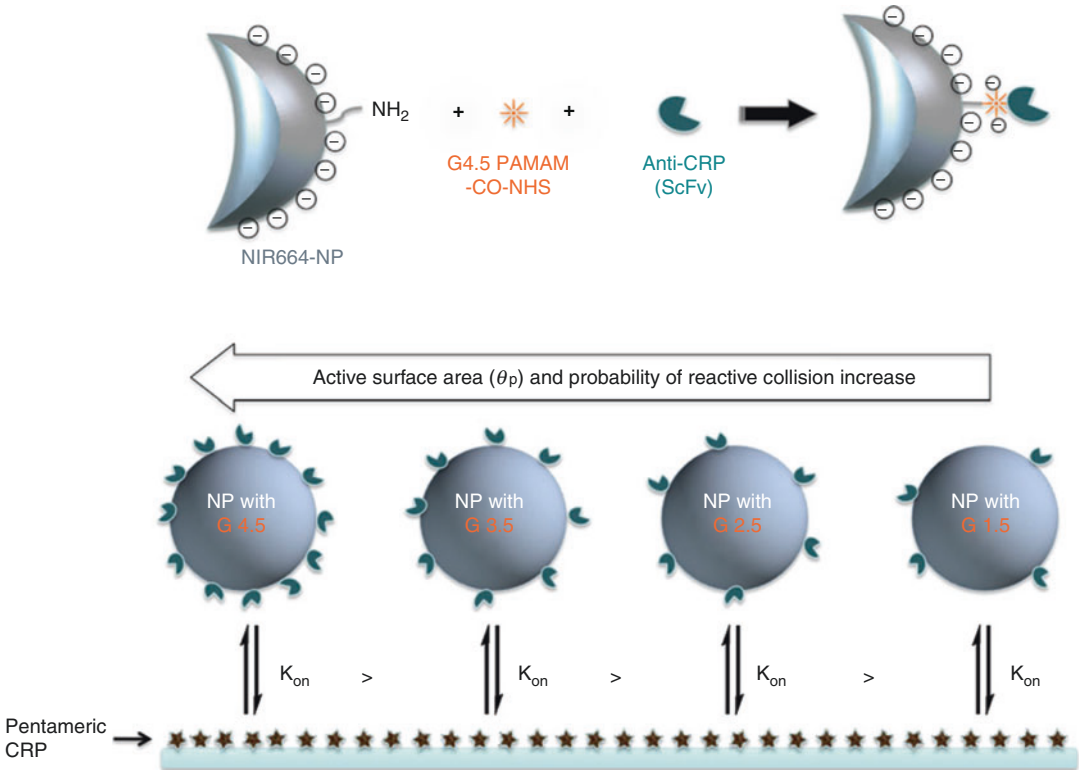
$$k_{on} = k_D \varepsilon \theta s \theta p \tag{2}$$

where k_D is the diffusion-limit rate constant for NP consumption by the antigen-surface, dependent on the NP radius and the viscosity of the medium; ε is the reaction efficiency for a reactive collision; θ_S is the fraction of the capture surface area (here the antigen-loaded capture plate) that is active for the capture reaction, which in this case is dependent on the surface coverage of active adsorbed antigen on the prepared capture surface; and θ_P is the active fraction of the NP surface, in this case dependent on the surface coverage of active antibody attached to the NP. Therefore, the central objective of surface functionalization with is to increase θ_P, which can be achieved by using dendrimer as a linker (Fig. 5).

Gubala et al. have shown how the use of dendrimers as multivalent linkers effectively controls and maximizes the active fraction of the particle surface. Dendrimers have been presented as efficient coupling agents to conjugate antibodies to the surface of dye-doped silica nanoparticles. The protein-binding capacity of the dendrimer-sensitized NPs increased by the same factor as the number of surface carboxylate groups on the dendrimer used. The highest generation of the dendrimers (G4.5) used in their study showed the highest surface binding rate and the highest signal in the direct-binding FLISA. The effect of increased active antibody loading significantly outweighed any effect of a decreased NP diffusion coefficient compared to that of a molecular dye-labeled antibody. The practically important parameter, the ratio of the equilibrium fluorescence (offset corrected) to the nonspecific offset, or signal to background ratio, increased by a factor of ~ 4 for the G4.5 dendrimer-conjugated NPs compared to the molecular dye-labeled antibody.

Future Directions

Successful detection devices for biomedical diagnostics frequently require high sensitivity and low LOD. Moreover, a device for point-of-care diagnostics must be both inexpensive and reliable under a variety of experimental conditions. In this entry the advantages of employing dielectric, dye-doped NPs have been highlighted for the purpose of enhancing the performance of fluorescence-based assays. The objective of this entry was to discuss the versatility of silica (or eventually polystyrene) as a host material for fluorescent dyes for many applications, specifically in the fields of nanobiotechnology and the life sciences. Silica as a host material represents an ideal scaffold for encapsulation of organic dyes in a silica matrix thus enhancing their stability and performance.



Dye-Doped Nanoparticles in Biomedical Diagnostics, Fig. 5 *Top* – A cartoon illustrating a bioconjugation reaction between single-chain fragments of C-reactive protein antibody (ScFv) and dye-doped silica NP. By varying the dendrimer generation, it is possible to vary the density of ScFv fragments on the surface of NP, hence modulate the active surface area of NPs. The net outcome is increased probability of reactive collision and hence association K_{on} constant. The effect of increased active antibody loading significantly outweighed any effect of a decreased NP diffusion coefficient compared to that of a molecular dye-labeled antibody

It is obvious from the growing literature on this topic that this area is still in a growth phase and that significant new developments are likely to emerge in the coming decades. Although co-doping of NPs with multiple dyes of distinguishable spectral properties has been demonstrated, the application of such barcoding NPs in multiplexed assays is expected to grow significantly. Similarly, the use of high-brightness NPs as labels in flow cytometry applications is attracting considerable interest. The versatility of silica as a host material and the feasibility to modulate the core–shell architecture provide a unique environment for innovative research toward creating integrated nanomaterials with highly fine-tuned functionality in the field of nanotechnology and beyond.

Cross-References

- ▶ Biosensors
- ▶ Nanoparticle Cytotoxicity
- ▶ Nanoparticles

References

1. McDonagh, C., Stranik, O., Nooney, R., MacCraith, B.D.: Nanoparticle strategies for enhancing the sensitivity of fluorescence-based biochips. Nanomedicine **4**, 645–656 (2009)
2. Oertel, J., Huhn, D.: Immunocytochemical methods in haematology and oncology. J. Cancer Res. Clin. Oncol. **126**, 425–440 (2000)
3. Brelje, T.C., Wessendorf, M.W., Sorenson, R.L.: Multicolor Laser Scanning Confocal Immunofluorescence Microscopy: Practical Application and

Limitations, Cell Biological Applications of Confocal Microscopy, 2nd edn, p. 165. Academic, San Diego (2002)
4. Schaferling, M., Nagl, S.: Optical technologies for the read out and quality control of DNA and protein microarrays. Anal. Bioanal. Chem. **385**, 500–517 (2006)
5. Burns, A., Ow, H., Wiesner, U.: Fluorescent core-shell silica nanoparticles: towards "Lab on a Particle" architectures for nanobiotechnology. Chem. Soc. Rev. **35**, 1028–1042 (2006)
6. Yan, J., Estévez, M.C., Smith, J.E., Wang, K., He, X., Wang, L., Tan, W.: Dye-doped nanoparticles for bioanalysis. Nano Today **2**, 44–50 (2007)
7. Gubala, V., Le Guevel, X., Nooney, R., Williams, D.E., MacCraith, B.: A comparison of mono and multivalent linkers and their effect on the colloidal stability of nanoparticle and immunoassays performance. Talanta **81**, 1833–1839 (2010)
8. Hermanson, G.T.: Bioconjugate Techniques, 2nd edn. Academic/Elsevier, London (2008)

Dye-Doped Nanospheres in Biomedical Diagnostics

▶ Dye-Doped Nanoparticles in Biomedical Diagnostics

Dynamic Clamp

Thomas Nowotny[1] and Pablo Varona[2]
[1]Centre for Computational Neuroscience and Robotics, School of Engineering and Informatics, University of Sussex, Brighton, UK
[2]Dpto. de Ingenieria Informatica, Universidad Autónoma de Madrid, Madrid, Spain

Synonyms

Artificial synapse; Conductance injection; Coupling clamp; Electronic expression; Electronic pharmacology; Hybrid network method; Neural activity-dependent closed loop; Reactive current clamp; Synthesized conductance injection; Synthesized ionic conductance; Synthesized synaptic conductance

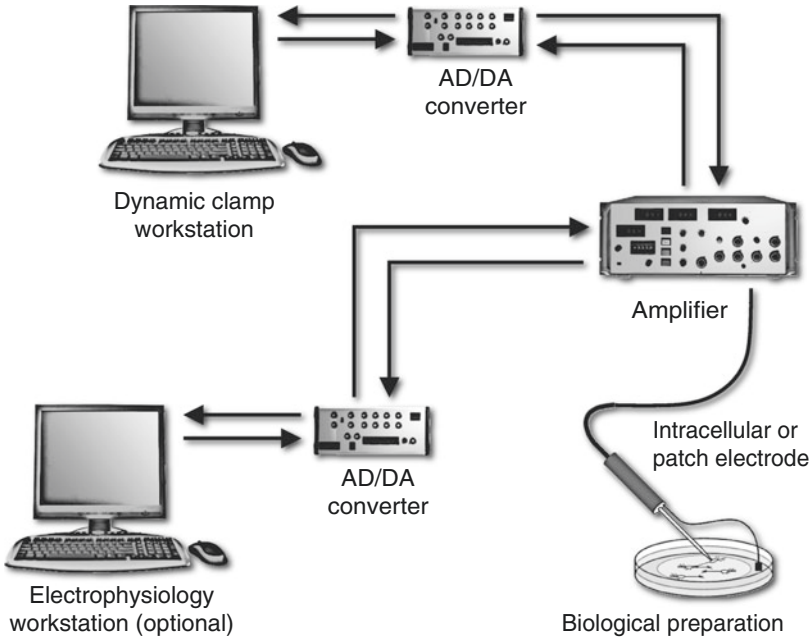
Definition

Dynamic clamp is a method of introducing simulated electrical components into electrically active biological cells using a fast real-time feedback loop between the cells and a computer.

The Dynamic Clamp Concept

Traditionally, neurophysiologists have used current and voltage clamp protocols to assess the electrical properties of neurons. In the current clamp technique, a current (typically a pulse) is injected into the neuron while the membrane potential is being recorded. In voltage clamp, the membrane potential is kept at a controlled value while the transmembrane current is being recorded. Both techniques have contributed to the understanding of the biophysical properties of excitable cells and allowed the design of conductance-based models of neurons.

The dynamic clamp technique operates in a cyclic way (see Fig. 1). An electrode is inserted into a neuron, and the membrane potential is recorded, amplified, and digitalized into a computer that calculates a current to inject into a postsynaptic neuron. This postsynaptic cell can be the same or a different one from which the membrane potential is being recorded. The injected current is calculated after solving a model that describes ionic or synaptic conductances or simulates the dynamics of individual neurons or neural networks. This process is repeated indefinitely with a fixed frequency. The time between two consecutive membrane potential acquisitions is known as the update rate. The correct selection of the update rate is critical for the proper function of the application. The maximum update rate is usually determined by the data acquisition board. A slow update rate can prevent the correct simulation of conductances and a realistic stimulation. On the other hand, a realistic update rate may require a fast computer, depending on the computational load of the models which can consist of several differential equations that need to be solved online. Depending on the goals of the experiment,



Dynamic Clamp, Fig. 1 The general dynamic clamp setup. The dynamic clamp system forms a closed observation-stimulus loop in which the measured and amplified membrane potential is input to the dynamic clamp hardware/software (dynamic clamp workstation), which calculates a corresponding transmembrane current according to a model specified by the user. The software then issues an appropriate current command, which is converted into a physical current injection by the amplifier. A second, independent, analog to digital (A/D) converter and PC with standard electrophysiology recording software monitors and saves all aspects of the experiment (electrophysiology workstation). The measurement-injection loop is repeated at 10–20 kHz, making the interaction essentially instantaneous for the target neuron

hardware or real-time software implementation may be required to achieve a realistic closed loop.

Through the closed-loop interaction, the dynamic clamp technique can simulate the addition of new ionic and synaptic conductances into a living neuron. Modelers and experimentalists have used dynamic clamp to design new experiments on excitable cell properties which were impossible or difficult to carry out with classical techniques. For example, dynamic clamp has been used to simulate the effects of introducing or removing conductances, simulating drug effects, increasing or decreasing neural activity, simulating in vivo conditions, and modifying or building neural dynamics with artificial synapses and artificial neurons in hybrid circuits (see below).

The dynamic clamp allows for a large variety of activity-dependent experiments to address complex information processing and learning in neural systems [1, 2]. The closed-loop interaction can reveal dynamics hidden under traditional electrophysiological protocols. Dynamic clamp can also be used to control neural dynamics under normal or pathological conditions and to induce specific types of neural activity in circuits whose targets are under investigation.

History of Dynamic Clamp

Dynamic clamp was independently invented by Sharp et al. [3] at Brandeis University and Robinson and Kawai [4] at Cambridge. The term "dynamic clamp" was coined by the former group, while Robinson and Kawai originally referred to the technique as "synthesized synaptic conductance transients." The two systems also differed in their technical implementation.

Sharp et al. used a generic data acquisition board and developed the Microsoft DOS-based

Dynamic Clamp, Table 1 Common dynamic clamp implementations

Name	Technology	URL/publication	Comments
RTXI (Real-Time eXperiment Interface)	Real-Time Linux, Comedi	http://www.rtxi.org/ [13]	Based on RTLDC, MRCI, and RTLAB
StdpC	Windows based, custom DigiData driver, NI-DAQmx	http://sourceforge.net/projects/stdpc/ [11]	Partially based on Dynclamp 2/4
G-clamp	LabVIEW RT, NI-DAQmx, embedded RT hardware subsystem	http://www.hornlab.neurobio.pitt.edu/ [15]	
SM-1, SM-2 amplifiers by Cambridge Conductance1	Custom analog hardware and DSP board	[4,17]	
Dynamic Clamp in CED signal	Uses DSP capabilities of CED DAQs	http://www.ced.co.uk/	Offers a subset of models in StdpC; released in Fall 2010
QuB	Windows based, NI-DAQmx	http://www.qub.buffalo.edu/wiki/index.php/Dynamic_Clamp [14]	
RTBiomanager	Real-Time Linux, Comedi	http://www.ii.uam.es/~gnb [16]	

dclamp software to emulate artificial ionic and synaptic conductances.

Robinson and Kawai followed the "hardware route" and built analog circuitry (the SM-1 amplifier) to inject dynamic clamp conductances into neurons.

The MS DOS-based dclamp was later replaced by other solutions, the main developments being real-time Linux-based (RTLDC, RTLAB, MRCI) solutions [5–8] and the Windows-based Dynclamp 2/4 [9].

Dynclamp 2/4 was further developed by Nowotny et al. since 2003 and under the name StdpC [10, 11], including for the first time plastic synapses, in particular spike-timing-dependent plasticity (STDP). Other additions included experimental automation through a scripting mechanism, a modern GUI based on QT (Trolltech, now Nokia), and most recently active electrode compensation [12].

The real-time Linux-based projects RTLDC, RTLAB, and MRCI merged in 2004 into the common RTXI (real-time experimentation interface) project [13]. A major step forward for the usability of this system was the recent release on live CDs.

Besides these longer-term projects, independent other implementations of dynamic clamp have been developed over the years. These include QuB [14], g-Clamp [15], and the dynamic clamp subsystem of the commercial CED Signal software.

One of the more recent systems (RTBiomanager; [16]) now promises another step change in dynamic clamp research in systematically extending the concept of closed-loop real-time experimentation beyond the realm of pure electrophysiology.

The currently available dynamic clamp systems and their characteristic features are summarized in Table 1. They can roughly be categorized into hardware- (SM-1, SM-2) and software-based solutions. The latter category roughly divides into Windows-based (StdpC, QuB, g-Clamp) and real-time Linux-based (RTXI, RTBiomanager) solutions. The hardware-based and real-time Linux-based systems guarantee hard real-time constraints, whereas the Windows-based solutions work in a soft real-time framework.

Dynamic Clamp Applications/ Configurations

Since its inception, dynamic clamp has seen several different extensions and modifications. The most common types of experiments are the following.

Simulated Chemical Synapses

The simulation of chemical synapses reaches back to the original publications of Sharp et al. [3] and Robinson and Kawai [4]. It involves measuring the membrane potential of two neurons and injecting current into one of them – the designated postsynaptic cell – in order to simulate the action of a nonexisting synapse or increasing the conductance of existing ones. Some researchers have also compensated existing synapses by using a negative conductance value. Further extensions of this protocol include:

1. Short-term depression of the simulated synapse
2. Spike-timing-dependent plasticity of the simulated synapse
3. Using a simulated model neuron as the presynaptic and/or postsynaptic neuron
4. Using an electronic circuit for the presynaptic neuron

A typical description of a chemical synapse is given by

$$
\begin{aligned}
I_{\text{syn}} &= g_{\text{syn}} S\left(V_{\text{post}} - V_{\text{rev}}\right) \\
\frac{dS}{dt} &= \alpha\left(V_{\text{pre}}\right)(1 - s) - \beta_0 s \\
\alpha\left(V_{\text{pre}}\right) &= \left\{ \begin{array}{ll} \alpha_0 & 0 \le t - t_{\text{spike}} \le t_{\text{release}} \\ 0 & otherwise \end{array} \right\}
\end{aligned}
$$

where I_{syn} denotes the postsynaptic current, g_{syn} the maximal synaptic conductance, S the fraction of active transmitter, V_{rev} the reversal potential of the synapse, and α_0 and β_0 are rate constants of synaptic transmitter activation and removal, t_{release} the duration of transmitter release in response to a presynaptic spike, and t_{spike} the time of the last occurrence of such a spike.

Simulating Electrotonic Connections (Gap Junctions)

A gap junction is essentially an Ohmic conductance between two cells through which current can flow in two directions. In dynamic clamp, a simulated gap junction is achieved by monitoring the membrane potential of two cells and injecting the appropriate currents into both cells according to Ohm's law and the observed membrane potential difference.

$$
\begin{aligned}
I_1 &= g_{\text{gap}}(V_2 - V_1) \\
I_2 &= g_{\text{gap}}(V_1 - V_2)
\end{aligned}
$$

where I_1 and I_2 are the currents injected into the two neurons, V_1 and V_2 are their respective membrane potentials, and g_{gap} the constant conductance of the gap junction.

A common extension are asymmetric or rectifying gap junctions, in which current can preferentially or exclusively flow in one direction.

Simulating Voltage-Dependent Ion Channels (Hodgkin–Huxley Conductances)

Voltage-dependent ion channels allow current flows into and out of a cell depending on its own membrane potential. For a simulated voltage-gated ion channel, the membrane potential of a cell is monitored, and the current of the simulated ionic conductance is injected into the same cell. Commonly, the conductance is described in the Hodgkin–Huxley formalism, for example, the sodium current is given by

$$
\begin{aligned}
I_{\text{Na}} &= g_{\text{Na}} m^3 h (V - V_{\text{Na}}) \\
\frac{dm}{dt} &= \alpha_m(1 - m) - \beta_m m \\
\alpha_m &= \frac{(-35 - V)/10}{\exp((-35 - V)/10) - 1} \\
\beta_m &= 4\exp((-60 - V)/18)
\end{aligned}
\qquad
\begin{aligned}
\frac{dh}{dt} &= \alpha_h(1 - h) - \beta_h h \\
\alpha_h &= 0.07\exp(-60 - V)/20 \\
\beta_h &= \frac{1}{\exp((-30 - V)/10) + 1}
\end{aligned}
$$

where I_{Na} denotes the sodium current, m and h are the so-called activation and inactivation variables, V denotes the membrane potential of the neuron, and V_{Na} is the sodium reversal potential.

Pattern Clamp

Pattern clamp is the simple combination of simulating a presynaptic neuron and a strong unidirectional gap junction. It nevertheless deserves a separate discussion because if the gap junction conductance is large, the electrical activity of the target neuron will be almost identical to the simulated presynaptic activity pattern, allowing the experimenter to virtually control: "clamp" the neuron to a particular activity pattern (hence the name "pattern clamp"). While quite simple from a theoretical point of view, this mode of dynamic clamp is very powerful as it allows the experimenter to interact with the rest of the neuronal system through the clamped neuron's natural synapses. This can include the release of neuromodulators and synaptic plasticity of the neuron's synapses.

Hybrid Circuits of Real and Artificial Neurons

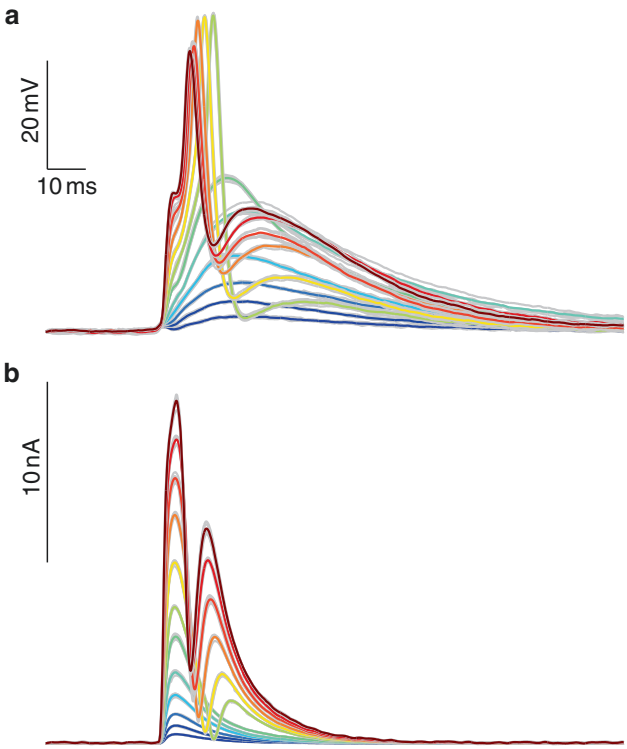
Hybrid circuits of multiple living and artificial neurons implemented with dynamic clamp protocols provide a unique way to test model neuron and network dynamics within a realistic cellular and synaptic environment [18]. A hybrid circuit can be used to bridge between different description levels and allows precise changes in single cellular or synaptic parameters in order to assess their direct participation in the collective output. In this way, hybrid methods can result in more physiological ways of stimulating networks, creating direct links between experiments and modeling, and providing new tools and validation methods for neural network analysis.

Examples of Dynamic Clamp Results

Typical results from standard dynamic clamp applications are illustrated in Figs. 2, 3, and 4. They were obtained using the StdpC dynamic clamp software on Windows XP and using a National Instruments M series PCI-6229 data acquisition board (Figs. 2 and 3) and the real-time Linux system RTLDC on Fedora Core using a National Instruments PCI-MIO-16E-4 data acquisition board (Fig. 4).

Figure 2 illustrates results from simulating an excitatory chemical synapse between two cells of the mollusk *Lymnaea stagnalis* in an in vitro whole brain preparation. The cerebral giant cell (CGC) and the motoneuron B1 were impaled with sharp microelectrodes. A second sharp electrode was inserted into the B1 neuron for the purpose of current injection. The dynamic clamp was configured to use the tonically firing CGC as the presynaptic cell and the quiescent B1 neuron as the postsynaptic cell. The synapse was configured with reversal potential 0 mV, activation threshold −20 mV, activation slope of 25 mV, and time scale of 10 ms. The maximal synaptic conductance was varied, taking values 10, 20, 30, 50, 80, 120, 170, 230, 300, 390, 490, and 600 nS (encoded by line color in Fig. 2). Note, how the simulation of a chemical synapse in dynamic clamp is different from injecting a predefined EPSP, in particular if a postsynaptic spike is evoked (current reversal during the spike).

Figure 3 illustrates the results of performing pattern clamp on the CGC neuron in the mollusk *Lymnaea stagnalis*. An accelerating-decelerating burst pattern was defined using the StdpC spike generator facility. The pattern was imposed onto the CGC neuron using a high conductance-simulated gap junction. Figure 3 shows the results for gap junction conductances of 300 nS (B) and 2,000 nS (C). For the lesser conductance, pattern clamp is only partially successful, and some spikes of the pattern are displaced or missing. At 2,000 nS, the pattern clamp is almost flawless, including exact timing of all spikes in the burst pattern. Interestingly, even though the gap junction conductance is almost seven times higher in C than in B, the injected currents are only two to three times stronger due to the better control achieved. It is also important to note that the spikes in the CGC are of higher amplitude than the spikes of the target activity pattern (Fig. 3d, e). This indicates that the CGC spikes are genuine and will travel down the axon rather than being merely strong deflections of the membrane potential at the soma.

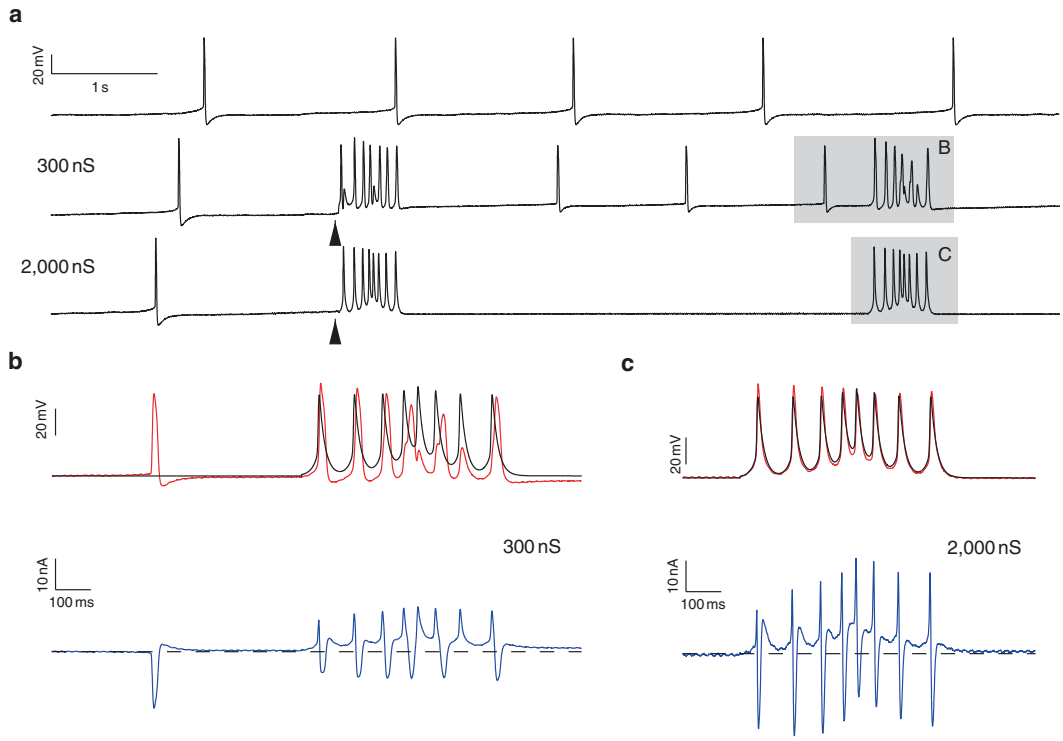


Dynamic Clamp, Fig. 2 Excitatory postsynaptic potentials (EPSPs) and excitatory postsynaptic currents (EPSCs) of a simulated, excitatory, chemical conductance introduced into a molluskan cell. (**a**) EPSPs for simulated maximal synaptic conductance of 10, 20, 30, 50, 80, 120, 170, 230, 300, 390, 490, and 600 nS aligned to the occurrence of the presynaptic spike. For each maximal conductance, five individual EPSPs (*thin gray lines*) and their average thick lines (*colored online*) are shown. (**b**) Corresponding EPSCs. The *colors* of **a** and **b** are matched to mark the corresponding data sets. Note how the introduction of a simulated synaptic conductance is different from injecting a fixed current waveform: When a postsynaptic spike is elicited (*yellow to red traces (color online)*), the current through the simulated ion channel shows a marked dip and almost reverses (Figure modified from Kemenes et al. [19])

A somewhat more unusual example of dynamic clamp is shown in Fig. 4. Here, dynamic clamp is used to synchronize a model neuron and a real cell in order to fit the properties of the cell online and in real time. Within the dynamic clamp software, two copies of a neuron model with slightly different parameters are simulated together with a simulated coupling to a real cell impaled with two electrodes. Through the simulated coupling, the cell and the model cells synchronize (partially) which allows to assess how much the dynamics of each of the model cells differs from the real cells'. After a fixed testing period, the model cell with the more appropriate dynamics is retained and further compared to a new model. Through this iterative process, the model cells become increasingly similar to the real cell allowing indirect deduction of cell properties.

Limitations

Like any other electrophysiological technique, dynamic clamp has practical and principal limitations. The main principal limitation is that dynamic clamp simulates the electrical effect of simulated channels in the cell membrane, not the flow of specific ions. Where the presence of the ions is important as, for example, with calcium ions that act as second messengers, simulating an ion channel will have a different effect than the actual channel would have. A second principal limitation is related to the position of the electrode(s) used.

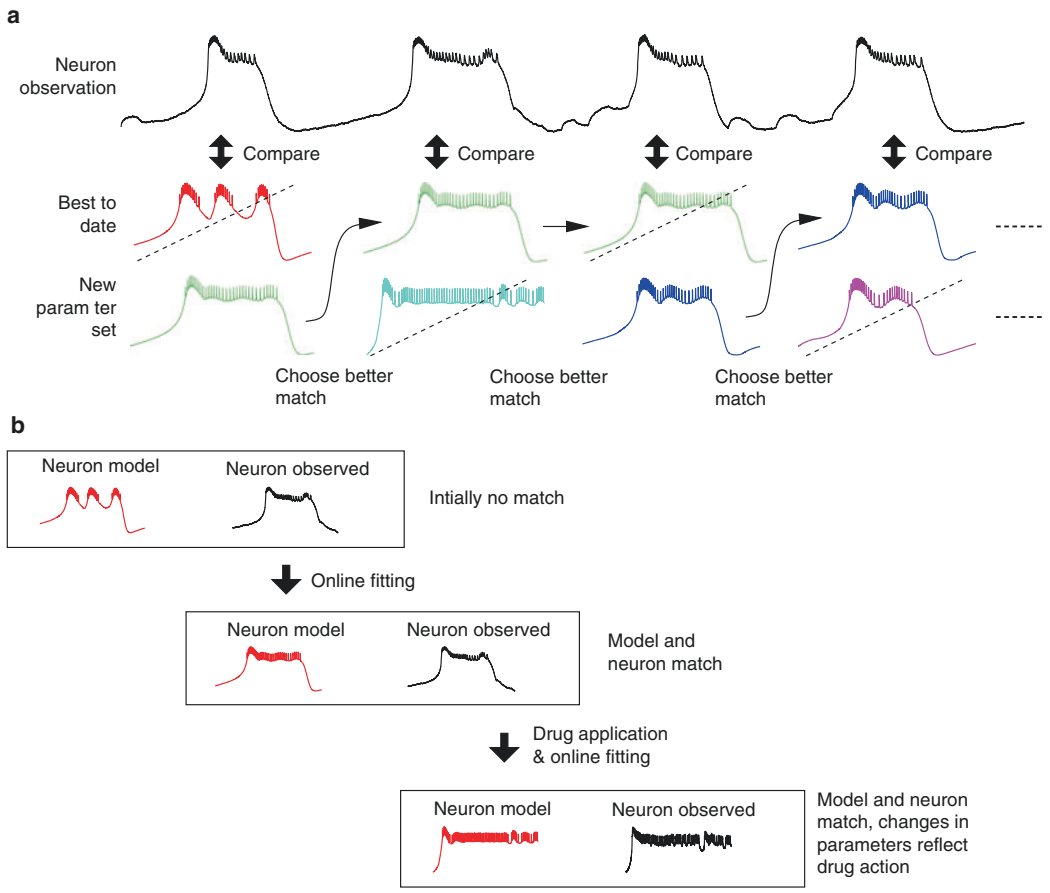


Dynamic Clamp, Fig. 3 Results of a pattern clamp experiment. The target cell spikes tonically when the pattern clamp is off (*panel a, top*). When the gap junction coupling in pattern clamp is not strong enough (gap junction conductance of 300 nS), the target neuron will only be clamped partially (*panel a, middle*), but sufficient conductance (gap junction conductance of 2,000 nS) leads to a complete clamp (*panel a, bottom*). Pattern clamp is switched on at the *black arrowheads*. Panels b and c are detailed views of the episodes shaded in *gray* in panel a. The *gray* (*red online*) trace was recorded from the target neuron; the *black* trace shows the desired pattern. Currents injected are shown in the *bottom panels of b and c.* With 300 nS conductance, the pattern clamp is not able to prevent intrinsic spiking (*panel a, middle and panel b*). This is readily achieved at 2 μS maximal conductance, even though the currents injected for this >6 times larger conductance are only two to three times larger owing to the better match achieved (From Kemenes et al. [19])

If electrodes are far from the locus in the cell where real channels would be located, the effect of the injected current can be quite different from the current that flows through the channels directly. If, for example, the current injecting electrode is inserted into the soma, a Na current is simulated that normally would be located at the axon hillock, and the axon hillock is electrically far removed from the soma; the dynamic clamp experiment may be difficult to interpret.

Practical limitations depend on the dynamic clamp implementation and the used hardware.

Dynamic clamp protocols rely critically on the update rate used for the closed-loop interaction. In many applications, this update rate must be strictly accomplished, requiring precise timing and no jitter. Specific hardware, fast DSP systems, or real-time software technology can overcome these problems. Software implementations are more suitable than hardware approaches as they are easier to program, generally inexpensive, and more flexible to modify and customize to a particular experiment. A common problem in electrophysiology, and more so in dynamic clamp, is the series resistance and capacitance of electrodes. Common techniques such as bridge series resistance compensation and capacitance compensation are generally not suitable for dynamic clamp due to the rapidly changing current injections. Most dynamic clamp experiments have, therefore, been conducted with two separate electrodes in the same cell, one for membrane potential



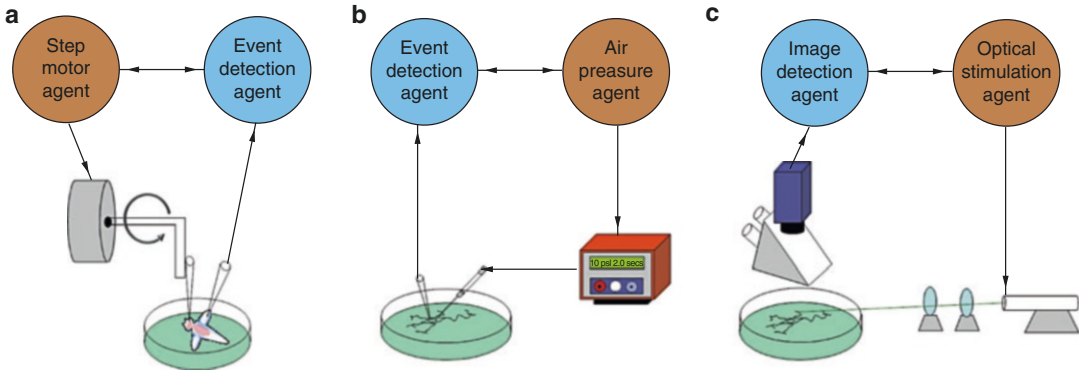
Dynamic Clamp, Fig. 4 Dynamic clamp used for online fitting. (**a**) Two model neurons with slightly differing parameters are simulated alongside a real cell (a lateral pyloric (LP) cell of the lobster) to which they are coupled through a simulated electrical coupling. After a test period, the better matching model cell is retained and compared to a model with a new, slightly different, parameter set. (**b**) Initially, the model cells' dynamics can be quite different from the dynamics of the real cell. After some time, the models fit the cell increasingly better and eventually should be able to track slow changes within the cell. The data shown in this figure are illustrative excerpts from experiments with a prototype system programmed in RTLDC and used with a lobster LP cell (Levi and Nowotny 2005, unpublished)

measurements and one for current injection. Others have used low-resistance patch clamp electrodes to avoid the problem.

A recent alternative solution to the electrode artifact problem was introduced by Brette et al. [12], who used a sophisticated calibration method to estimate series resistance and capacitance artifacts with a digital model of the electrode and subtract them digitally and in real time. Using this new active electrode compensation (AEC) allows experimenters to conduct dynamic clamp experiments with a single high-resistance electrode.

Future Directions

The dynamic clamp concept has an enormous potential to precisely control the spatiotemporal aspects of a stimulus and to build activity-dependent stimulus–response loops to interact with neural systems beyond the realm of electrophysiology. Establishing these loops can be an essential step toward understanding the dynamics of many neural processes and can bridge between traditionally disparate levels of analysis (e.g., subcellular, network, and system dynamics).



Dynamic Clamp, Fig. 5 Generalization of the dynamic clamp concept for different stimulation and recording techniques. Computer algorithms implemented as intelligent agents monitor neural activity to exert goal-driven closed-loop stimulation. Panel **a**: activity-dependent mechanical stimulation. Neural activity is measured with intra- or extracellular recordings. Specific events detected in this activity are used to control a step motor for mechanical stimulation of sensory systems. Closed-loop interaction can be used to automatically find receptive fields of cells that react to the sensory stimulation throughout the nervous system [15]. Panel **b**: activity-dependent drug microinjection. A microinjector delivers a timely and precise amount of neurotransmitter or neuromodulator to achieve a desired type and level of neural activity. Panel **c**: closed-loop optical recording with real-time event detection used to drive laser stimulation which can go down to nanoscale precision

The same principles used in the dynamic clamp technology can be generalized to develop new techniques of activity-dependent stimulation by enhancing or going beyond the electrophysiological realm in a broad spectrum of research in nervous systems (see Fig. 5).

For example, a generalization of the dynamic clamp concept has been used to find receptive fields throughout the sensorimotor transformation by implementing activity-dependent mechanostimulation [20]. In this protocol, a step motor is controlled to exert mechanic stimulation as a function of neural events recorded with microelectrodes throughout this transformation. The protocol can explore different motor angles and speed until neural activity correlated with the stimuli is found. Another dynamic clamp extension has implemented activity-dependent real-time drug microinjection to stimulate neural circuits [21] as a function of specific patterns recorded in the electrical activity. This protocol is used to constrain neural activity within specified limits and to produce robust temporal patterns. Other protocols can include closed-loop optical recording and stimulation, for example, in two-photon microscope setups which can lead to highly realistic experimental conditions that overcome the dynamic clamp limitations discussed above. Further extensions of the closed-loop interaction with neural systems will lead to novel brain–machine interfaces and prosthetic devices.

Cross-References

- ▶ Biomimetics
- ▶ Biosensors
- ▶ Computational Systems Bioinformatics for RNAi
- ▶ Electrode–Organic Interface Physics
- ▶ Lab-on-a-Chip for Studies in *C. elegans*
- ▶ Nanorobotics for Bioengineering
- ▶ Synthetic Biology

References

1. Prinz, A.A., Abbott, L.F., Marder, E.: The dynamic clamp comes of age. Trends Neurosci. **27**(4), 218–224 (2004)
2. Destexhe, A., Bal, T. (eds.): Dynamic Clamp: From Principles to Applications. Springer Series in Computational Neuroscience. Springer, New York (2009)
3. Sharp, A.A., O'Neil, M.B., Abbott, L.F., Marder, E.: The dynamic clamp: artificial conductances in

biological neurons. Trends Neurosci. **16**(10), 389–394 (1993)
4. Robinson, H.P., Kawai, N.: Injection of digitally synthesized synaptic conductance transients to measure the integrative properties of neurons. J. Neurosci. Methods **49**(3), 157–165 (1993)
5. Dorval, A.D., Christini, D.J., White, J.A.: Real-time linux dynamic clamp: a fast and flexible way to construct virtual ion channels in living cells. Ann. Biomed. Eng. **29**, 897–907 (2001)
6. Culianu, C.A., Christini, D.J.: Real-time Linux experiment interface system: RTLab. In: Proceedings of the IEEE 29th Annual Bioengineering Conference, Newark, pp. 51–52 (2003)
7. Raikov, I., Preyer, A.J., Butera, R.: MRCI: a flexible real-time dynamic clamp system for electrophysiology experiments. J. Neurosci. Methods **132**, 109–123 (2004)
8. Butera Jr., R.J., Wilson, C.G., DelNegro, C.A., Smith, J.C.: A methodology for achieving high-speed rates for artificial conductance injection in electrically excitable biological cells. IEEE Trans. Biomed. Eng. **48**(12), 1460–1470 (2001)
9. Pinto, R.D., Elson, R.C., Szücs, A., Rabinovich, M.I., Selverston, A.I., Abarbanel, H.D.I.: Extended dynamic clamp: controlling up to four neurons using a single desktop computer interface. J. Neurosci. Methods **108**, 39–48 (2001)
10. Nowotny, T., Zhigulin, V.P., Selverston, A.I., Abarbanel, H.D., Rabinovich, M.I.: Enhancement of synchronization in a hybrid neural circuit by spike-timing dependant plasticity. J. Neurosci. **23**(30), 9776–9785 (2003)
11. Nowotny, T., Szücs, A., Pinto, R.D., Selverston, A.I.: Stdpc: a modern dynamic clamp. J. Neurosci. Methods **158**(2), 287–299 (2006)
12. Brette, R., et al.: High-resolution intracellular recordings using a real-time computational model of the electrode. Neuron **59**(3), 379–391 (2008)
13. Lin, R.J., Bettencourt, J., White, J., Christini, D.J., Butera, R.J.: Real-time experiment Interface for biological control applications. Conf. Proc. IEEE Eng. Med. Biol. Soc. **1**, 4160–4163 (2010)
14. Milescu, L.S., Yamanishi, T., Ptak, K., Mogri, M.Z., Smith, J.C.: Real-time kinetic modelling of voltage-gated ion channels using dynamic clamp. Biophys. J. **95**, 66–87 (2008)
15. Kullmann, P.H.M., Wheeler, D.W., Beacom, J., Horn, J.P.: Implementation of a fast 16-bit dynamic clamp using LabVIEW-RT. J. Neurophysiol. **91**, 542–554 (2004)
16. Muñiz, C., Rodriguez, F.B., Varona, P.: RTBiomanager: a software platform to expand the applications of real-time technology in neuroscience. BMC Neurosci. **10**, 49 (2009)
17. Robinson, H.P.C.: A scriptable DSP-based system for dynamic conductance injection. J. Neurosci. Methods **169**(2), 271–281 (2008)
18. Le Masson, G., Renaud-Le Masson, S., Debay, D., Bal, T.: Feedback inhibition controls spike transfer in hybrid thalamic circuits. Nature **417**, 854–858 (2002)
19. Kemenes, I., Marra, V., Crossley, M., Samu, D., Staras, K., Kemenes, G., Nowotny, T.: Dynamic clamp with StdpC software. Nat. Protoc. **6**(3), 405–417 (2011)
20. Muñiz, C., Levi, R., Benkrid, M., Rodríguez, F.B., Varona, P.: Real-time control of stepper motors for mechanosensory stimulation. J. Neurosci. Methods **172**(1), 105–111 (2008)
21. Chamorro, P., Levi, R., Rodriguez, F.B., Pinto, R.D., Varona, P.: Real-time activity-dependent drug microinjection. BMC Neurosci. **10**, 296 (2009)

Dynamic Force Measurements on Modified Surfaces with AFM

Yan Wu[1], Kaushik K. Rangharajan[2] and Shaurya Prakash[2]
[1]Department of Engineering Physics, University of Wisconsin – Platteville, Platteville, WI, USA
[2]Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, OH, USA

Synonyms

Amplitude-phase-distance (APD) curves; Dynamic force spectroscopy

Definition

Dynamic force measurement is a variation of the force spectroscopy mode of an Atomic Force Microscope (AFM), where the amplitude and phase response of an oscillating AFM probe are used to detect the interaction forces between the probe and the sample or substrate of interest. In contrast to quasi-static force measurements where the micro-cantilever's deflection signal can be directly related to the interaction force between the AFM tip and sample surface using Hooke's law, the interpretation of amplitude and phase data in dynamic force measurement often requires detailed analysis of a micro-cantilever's vibration

using dynamic models. When applied at solid/water interfaces, dynamic force measurement is sensitive to solvation forces, van der Waals forces, electric double-layer forces, and viscous damping forces. Dynamic force measurements therefore lead to a useful technique for characterizing chemically modified surface in electrolyte solutions.

Introduction to Force Spectroscopy

Shortly after the invention of AFM by Binnig and Quate [1], surface scientists began to use AFM as a tool not only for high-resolution topography imaging but also for quantitative measurements of interaction forces between the AFM tip and a sample surface [2, 3]. The tip-sample interaction forces provide valuable information on local material properties such as elasticity, hardness, Hamaker constant, adhesion, and surface charge densities [4]. The application of the AFM in measuring the tip-sample interaction force is called force spectroscopy. Force spectroscopy has wide applications in surface science, materials engineering, and biology.

Most force spectroscopy is done with a quasi-static force measurement. In quasi-static force measurements, the cantilever-based AFM probe is brought into contact with a solid surface at a constant speed. The deflection at the free end of the micro-cantilever beam due to tip-sample interaction force is detected using the optical lever technique, where a laser light reflects off the end of the cantilever and strikes a detector. For small deflection (with respect to the length of the beam), the cantilever can be modeled as a linear spring with spring constant k_c. Assuming static equilibrium during the approaching and/or retracting process of the cantilever, the balance between the tip-sample interaction force and the spring force gives the expression of the measured interaction force via Hooke's law:

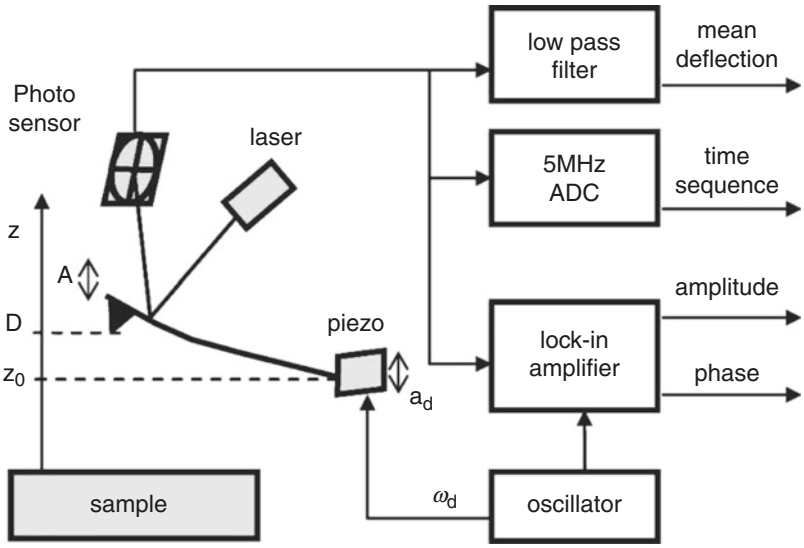
$$F_i = k_c \cdot h \tag{1}$$

where h is the deflection of the cantilever beam. Typically, the deflection and the vertical position of the cantilever are monitored during the entire process of approaching and retracting the AFM tip to and from the sample surface. As the result, a curve of measured force versus tip-sample distance can be constructed.

Dynamic Force Measurement

The interest in dynamic force measurements started when tapping-mode AFM became a dominant imaging mode for AFM applications [5, 6]. During a dynamic force measurement, the micro-cantilever is driven by an excitation signal at the resonance frequency of the micro-cantilever. The amplitude of the oscillation and the phase lag of the cantilever response with respect to the excitation signal are measured as a function of tip-sample distance. Analogous to quasi-static force-distance curves, the data from dynamic force measurements are called amplitude-phase-distance (APD) curves. An example of a dynamic force measurement setup is shown in Fig. 1. In Fig. 1, the excitation of the cantilever is done acoustically, i.e., via exciting a piezo-ceramic at the base of the micro-cantilever. There are other means to drive the cantilever, including magnetic [7] and photothermal excitations [8].

Most of the early work in dynamic force measurements was done for better understanding of the cantilever dynamics in tapping-mode AFM, especially with the goal of providing the interpretation of the phase signal in tapping-mode AFM [9–11]. Consequently, early work on dynamic measurements focused on large-amplitude oscillation (on the order of a few tens of nanometers) to match the typical operation condition of tapping-mode imaging. These types of APD curves are useful to correlate experimental observations such as attractive versus repulsive imaging regions in tapping mode with theories on cantilever dynamics [11]. Since the tip goes through almost the entire force field from attractive to repulsive region in every cycle of oscillation, large-amplitude dynamic force measurements fall short in capturing the detailed dependence of tip-sample interaction forces as function of



Dynamic Force Measurements on Modified Surfaces with AFM,
Fig. 1 Schematic diagram showing the experimental setup dynamic force microscopy. The tip oscillates with an average distance, *D*, from the sample surface with amplitude, *A*. The piezo provides the force needed to oscillate the micro-cantilever

tip-sample distance as in quasi-static force measurement.

In the late 1990s, researchers started to use small-amplitude dynamic force measurements in detecting the solid/liquid interfacial forces [12, 13]. Previously, there had been a lot of work on using quasi-static force curves to study solid/liquid interfacial forces [4, 14, 15]. When using either dynamic or quasi-static force measurements to study solid/liquid interfacial force, the micro-cantilever and the AFM tip are typically immersed in liquid to avoid the influence of forces from liquid/air interface. Compared with quasi-static force measurements, there are several advantages of using small-amplitude dynamic force measurement in liquid environments: first, the cantilever deflection signal tends to pick up greater thermal drift noise in liquid environment than in air, which allows much better error minimization in dynamic force measurements in a liquid environment since the force is not directly calculated from the cantilever deflection signal. The noise due to thermal drift can be rejected by modulating the deflection signal with an oscillating cantilever and selecting only the signal at the modulating frequency. Second, the viscous damping force is much stronger in a liquid than in air. It is hard to separate the conservative forces such as van der Waals force and electric double-layer force from dissipative viscous damping force in quasi-static force measurements. In dynamic force measurements, however, conservative forces and dissipative forces produce in-phase response and out-of-phase responses in the APD curves, respectively, making it possible to decouple the two kinds of forces. Dynamic force measurements can be done on the same location where tapping-mode scan is performed without losing the spot because there is no need to change the AFM probe. This is convenient when the nature of application demands tapping-mode imaging, for example, soft biological membranes or surfaces with nano-bubbles.

It is important to keep the oscillation amplitude small when conducting dynamic force measurements in a liquid. In addition to the fact that the large-amplitude oscillation is not sensitive to small changes in force fields, large-amplitude oscillation excites higher modes of vibration of the cantilever beam simultaneously in liquids [16]. The energy transfer from fundamental vibration mode to higher vibration mode complicates the analysis of cantilever dynamic response in order to obtain the tip-sample interaction force. For small-amplitude oscillation, the complicated cantilever beam dynamics can be simplified using point mass model with the dominant vibration mode being the fundamental mode of the cantilever.

Dynamic Model for Small-Amplitude Dynamic Force Measurement in Liquids

The equation of motion of an oscillating cantilever in a liquid, shown in Fig. 1, can be approximated in low Reynolds number, Re, flow to the first order by a point mass versus beam system [17]:

$$m\ddot{z} = -b\dot{z} - k_c(z - z_0) + k_c a_d \cos(\omega_d t) + F_i \tag{2}$$

where z is the position of the tip at any time and z_0 is the position at zero cantilever deflection. The coefficient, m, is the effective mass of the cantilever, including the additional fluid mass being accelerated, and k_c is the spring stiffness of the cantilever. The tip-sample interaction force, F_i, includes the electrical double-layer force and van der Waals force. The hydrodynamic damping, b, and the tip-sample interaction force, F_i, are all tip-sample distance dependent. Including the higher harmonic modes, the steady-state solution of Eq. 2 for the oscillation of the cantilever tip can be approximated as

$$z = D + \sum_{n\geq 1} A_n \cos(n\omega_d t - \phi_n) \tag{3}$$

where D is the mean value of the tip-sample distance, and A_n and ϕ_n are the amplitude and phase shift of the oscillation at the n^{th} harmonics, respectively.

In typical dynamic measurements, Re ranges from 10^{-6} (at the tip radius) to 10^{-3} (at the widest beam width). For Re $<<$ 1, the flow around the tip and the beam of the cantilever is classified creeping flow, with standard solutions existing for many different conditions. If the oscillation amplitude is small compared to the range of the conservative interfacial forces including the van der Waals force and the electric double-layer force, we can assume the total tip-sample interaction force, F_i, is a single value function of tip position z and can be approximated by

$$\begin{aligned} F_i(z) &\approx F_i(D) + \frac{\partial F_i}{\partial z}(z - D) \\ &= \langle F_i \rangle - k_i(z - D) \end{aligned} \tag{4}$$

where $< F_i >$ is the averaged interaction force over one cycle and k_i is the force gradient or the effective stiffness of the tip-sample interaction. For low Reynolds number flow, the forces arising from interactions between the AFM tip, liquid, and surfaces can be superimposed, and the virial theorem [18] can be used to estimate the time averaged forces. Note that in Eqs. 2 and 4, F_i represents the summation of all conservative interfacial forces in aqueous medium including van der Waals force, electric double-layer force, hydration force, and others. At steady state, the virial theorem for the point mass model is $(1/2)m\langle \dot{z}^2 \rangle = -(1/2)\langle F_i z \rangle$ and the power balance at steady state, $\langle k_c a_d cos(\omega_d t)\dot{z} \rangle + \langle -b\dot{z}^2 \rangle + \langle F_i \dot{z} \rangle = 0$, giving

$$\frac{a_d \cos\phi_1}{A_1} = \left(\frac{k_i}{k_c} + 1\right)\frac{1}{{A_1}^2}\sum_{n\geq 1} {A_n}^2 - \left(\frac{\omega_d}{\omega_r}\right)^2 \frac{1}{{A_1}^2}\sum_{n\geq 1} n^2 {A_n}^2 \tag{5a}$$

and

$$\frac{a_d \sin\phi_1}{A_1} = \frac{b\omega_d}{k_c {A_1}^2}\sum_{n\geq 1} n^2 {A_n}^2 \tag{5b}$$

where ω_r is the cantilever's resonance frequency, and it equals $(k_c/m)^{1/2}$. Equation 5a shows that the in-phase component of the APD curve, $\cos\varphi_1/A_1$, is related to the conservative interaction (the gradient of the single valued interaction force, k_i), and Equation 5b shows the out-of-phase component, $\sin\varphi_1/A_1$, is related to the dissipative interaction (hydrodynamic damping, b).

If the amplitudes at higher harmonics are negligible, Eq. 5 can be simplified to

$$k_i = \left(\frac{\cos\phi_1}{A_1} - \frac{\cos\phi_\infty}{A_\infty}\right) a_d k_c \tag{6a}$$

and

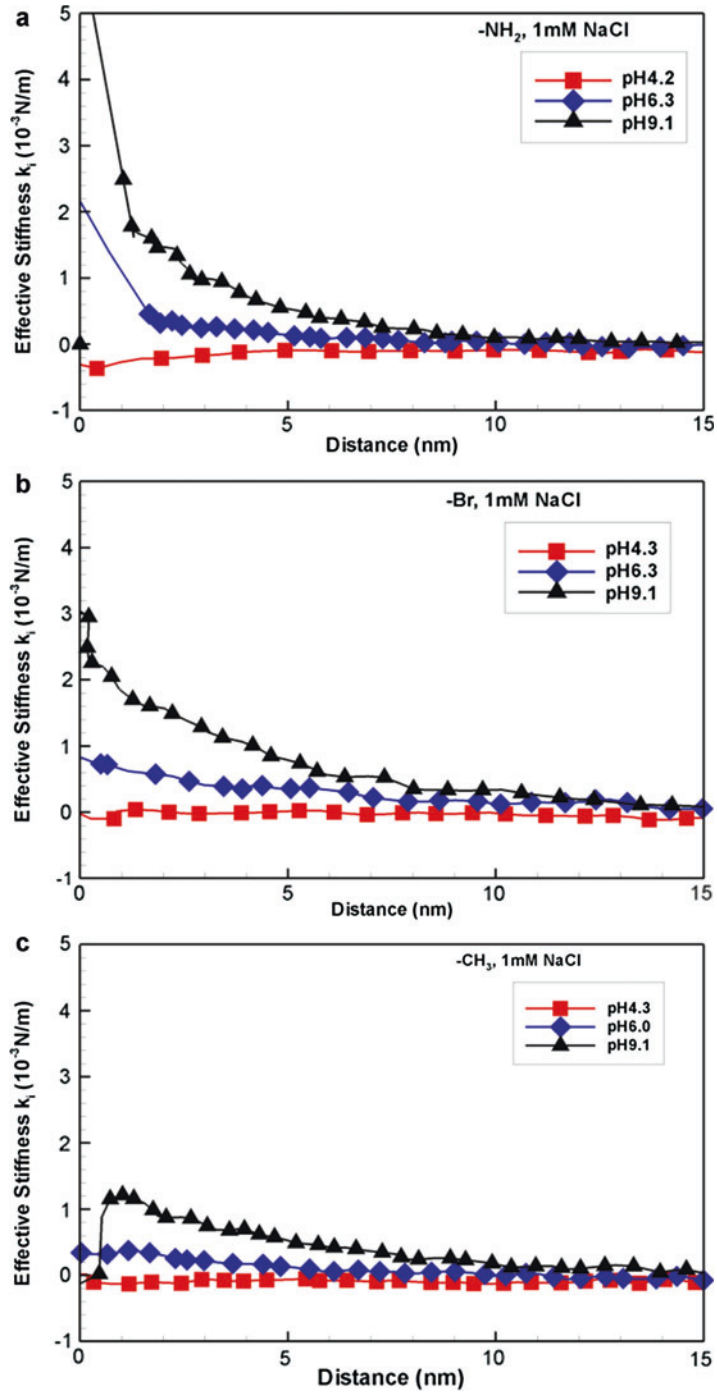
$$b_s = \left(\frac{b}{b_c} - 1\right) b_c = \left(\frac{A_\infty \sin\phi_1}{A_1 \sin\phi_\infty} - 1\right) b_c \quad (6b)$$

where A_∞ and ϕ_∞ are fundamental mode amplitude and phase when the AFM probe is far away from the solid/water interface. In dynamic force curve measurements, A_∞ and ϕ_∞ are obtained experimentally by measuring the amplitude and phase signal when the oscillating tip is at least a few microns away from the sample surface. The hydrodynamic damping coefficient, b, sensed in the dynamic measurements consists of a term associated with the damping of the cantilever, b_c, that stays relatively unchanged when the tip-sample distance is in the tens of nanometer range and the cantilever beam is at least several microns away from the sample surface and a distance-dependent term, b_s, which arises due to the squeeze film damping between the tip and sample surface [19]. Equation 6 translates the measured data in dynamic force measurements, A_1 and ϕ_1, to physical quantities, k_i and b_s, at the solid/liquid interface. The rest of parameters in Equation 6 can be measured independently from APD curves. Both k_c and b_c can be determined by the thermal noise spectrum of the cantilever when the tip is far away from the surface [20, 21]. The amplitude of acoustic excitation, a_d, can be approximated to be A_∞/Q_∞, where Q_∞ is the quality factor of the cantilever resonance curve measured at a distance far away from the surface.

Equation 6 shows that the dynamic force curve can be used to calculate the effective stiffness of the electric double-layer interaction and hydrodynamic damping at solid/water interfaces if the contribution of the higher harmonics to cantilever oscillation is negligible. In general, the following two conditions are necessary if we can safely neglect the higher harmonics without losing much insight to the physics of the interaction. First, the mechanical stiffness of the cantilever needs to be much greater than the effective stiffness of the tip-sample interaction. Second, the oscillation amplitude needs to be small compared to the characteristic range of the interaction, which is the thickness of the electric double layer and is usually taken to be equal to the Debye length of the electrolyte solution [17], but can vary to 3–4 times the Debye length [22, 23].

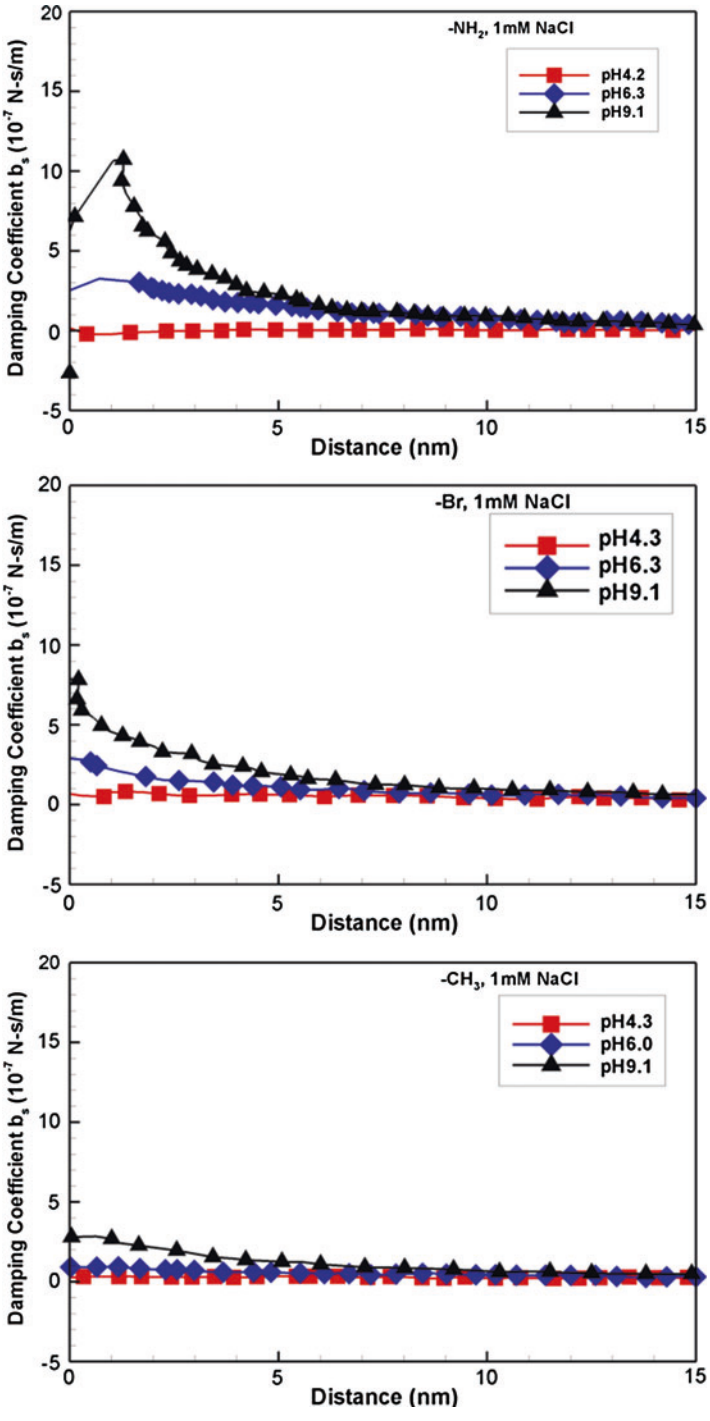
Applications of Dynamic Force Measurement on Modified Surfaces in Liquid

The understanding of solid/liquid interfacial forces is fundamental to a variety of fields, including tribology, colloidal systems, electrochemical cells, biomolecular interactions, and electrokinetic transport through micro- and nanofluidic devices [22, 24]. Dynamic force measurements have shown to be sensitive to conservative interactions such as the electrostatic double-layer force and dissipative interactions from hydrodynamic damping in aqueous solutions. An application of particular interest is to characterize modified surfaces or engineered surfaces in micro- and nanofluidic devices. Due to the high surface-area-to-volume (SA/V) ratio in micro- and nanofluidic devices, the material properties of channel walls such as surface charge density, surface roughness, and surface hydrophobicity become dominating factors in determining the fluidic transport inside the channel [24, 25]. For example, macroscale fluid mechanics assumes a no-slip condition at the wall, which implies that the velocity of the fluid layer in contact with the walls is the same as that of the walls. In micro- and nanofluidic channels, however, finite slip of fluid flow can be achieved by using hydrophobic material for channels or modify the surface of channel wall to be hydrophobic [26]. The unique surface-mediated flow in micro- and nanofluidic channels opens the possibility to control fluidic transport characteristics by systematically engineering the surface properties of channel walls [22, 24]. Dynamic force measurement has proven to be a useful tool in characterizing the modified surfaces for micro- and nanofluidic devices because it provides information not only on the electric double-layer interaction but also on hydrodynamic damping coefficients at



Dynamic Force Measurements on Modified Surfaces with AFM, Fig. 2 Plots of effective stiffness versus tip-sample distance for functionalized silica surfaces immersed in 1 mM NaCl solutions titrated at three different pH values

solid/liquid interfaces. Hydrodynamic damping coefficients are important in micro- and nanofluidic devices as they can provide insights into the physical implications of surface-fluid interactions. One particular surface-fluid interaction of interest is the drainage velocity of the liquid thin film confined between two surfaces approaching one another. The drainage is related



Dynamic Force Measurements on Modified Surfaces with AFM, Fig. 3 Plots of hydrodynamic damping versus tip-sample distance over functionalized silica surfaces immersed in 1 mM NaCl solutions titrated at three different pH values

to the surface properties of the confining surfaces. For example, the drainage rate (or fluid drainage velocity) is often related to the equilibrium surface forces or can be used to determine interfacial properties for both Newtonian and non-Newtonian fluids. Fluid drainage between surfaces with different combinations of hydrophilicity and hydrophobicity [27] has been used as an

indirect way to characterize the boundary slip effect, with the underlying mechanisms for slip being the subject of continued debate among the scientific community.

In the following, an illustrative example for the application of dynamic force measurements on modified surfaces is described [28]. In this example, silica surfaces were chemically modified yielding surfaces with –Br, -NH_2, and –CH_3 functional group terminations, respectively, using previously reported methods comprising self assembly and "click" chemistry [29]. The dynamic force measurement data were obtained using gold-coated silicon nitride probes (Biolever, Olympus) with a rectangular cantilever (nominal spring constant of 0.03 N/m). The cantilever oscillation amplitude was less than 2 nm when the tip was within 15 nm distance from sample surface. Once the APD curves were measured, the effective interaction stiffness and hydrodynamic damping can be calculated based on Eq. 6. Figure 2 shows the effective interaction stiffness between the AFM tip and the chemically modified surfaces with –Br, –NH_2, and –CH_3 functional groups in 1 mM NaCl solutions at different pH values, respectively. Assuming electric double-layer interaction is dominating in the tip-sample distance range of interest (1–15 nm) in this study, the effective interaction stiffness, k_i, can be approximated to be the derivative of electric double-layer force over distance in this case, as expressed in

$$k_i \approx g\sigma_s\sigma_T \exp\left(-\frac{D}{L_D}\right) \tag{7}$$

where σ_S and σ_T are the surface charge densities of the sample and the AFM tip, respectively, L_D is the Debye length of the electrolyte solution, and g can be considered as a geometry constant when z is of the same order of magnitude of the tip radius. Equation 7 assumes small surface potentials (smaller than or of the order of kT/e, i.e., around 25 mV) and that D is large enough compared to L_D. A higher absolute value of effective stiffness indicates a stronger double-layer interaction force and stronger surface charge density at the sample surface. One can compare charge densities between different modified surfaces, evaluate the effect of changing pH value in the solution on surface charges, and find out the electric double-layer interaction ranges from the results in Fig. 2.

The distance-dependent hydrodynamic damping at the modified surfaces sensed by the micro-cantilever is shown in Fig. 3. Figure 3 shows that there are distinct hydrodynamic damping responses among the silica surfaces with dissimilar functional groups. For example, there is a significant increase in the damping factor of the –Br surface as the distance between the probe and sample decreases, but by contrast there is almost no change for the –CH_3-terminated surface under similar test conditions. With each kind of surface, the solution pH value also has an impact on the damping curves, most likely due to the fact that pH acts as a control variable to modulate the effective surface charge density, thereby changing the interaction forces between the tip and the surface.

The example shows dynamic force measurement is a powerful tool to probe solid/liquid interfacial forces. When the oscillation amplitude of a micro-cantilever is much less than electrical double-layer interaction range, there are distinctive dynamic responses to chemically modified surfaces in aqueous electrolyte solutions within a few Debye lengths from the surfaces. The effective interaction stiffness is highly dependent on the surface charge. The hydrodynamic damping also shows strong dependence on the surface hydrophobicity.

Cross-References

- ▶ AFM in Liquids
- ▶ AFM, Tapping Mode
- ▶ Nanostructures for Surface Functionalization and Surface Properties

References

1. Binnig, G., Quate, C.F., Gerber, C.: Atomic force microscope. Phys. Rev. Lett. **56**(9), 930–933 (1986)
2. Weisenhorn, A.L., et al.: Forces in atomic force microscopy in air and water. Appl. Phys. Lett. **54**(26), 2651–2653 (1989)
3. Martin, Y., Wickramasinghe, H.K.: Magnetic imaging by "force microscopy" with 1000 Å resolution. Appl. Phys. Lett. **50**(20), 1455–1457 (1987)
4. Butt, H.-J., Cappella, B., Kappl, M.: Force measurements with the atomic force microscope: technique, interpretation and applications. Surf. Sci. Rep. **59** (1–6), 1–152 (2005)
5. García, R., Pérez, R.: Dynamic atomic force microscopy methods. Surf. Sci. Rep. **47**(6–8), 197–301 (2002)
6. Anczykowski, B., Krüger, D., Fuchs, H.: Cantilever dynamics in quasinoncontact force microscopy: spectroscopic aspects. Phys. Rev. B **53**(23), 15485–15488 (1996)
7. Xu, X., Raman, A.: Comparative dynamics of magnetically, acoustically, and Brownian motion driven microcantilevers in liquids. J. Appl. Phys. **102**(3), 034303 (2007)
8. Ratcliff, G.C., Erie, D.A., Superfine, R.: Photothermal modulation for oscillating mode atomic force microscopy in solution. Appl. Phys. Lett. **72**(15), 1911–1913 (1998)
9. Paulo, Á.S., García, R.: Unifying theory of tapping-mode atomic-force microscopy. Phy. Rev. B **66**(4), 041406 (2002)
10. Hölscher, H.: Quantitative measurement of tip-sample interactions in amplitude modulation atomic force microscopy. Appl. Phys. Lett. **89**(12), 123109 (2006)
11. Tamayo, J., García, R.: Relationship between phase shift and energy dissipation in tapping-mode scanning force microscopy. Appl. Phys. Lett. **73**(20), 2926–2928 (1998)
12. Lantz, M., et al.: Dynamic force microscopy in fluid. Surf. Interface Anal. **27**(5–6), 354–360 (1999)
13. O'Shea, S.J., Lantz, M.A., Tokumoto, H.: Damping near solid – liquid interfaces measured with atomic force microscopy. Langmuir **15**(4), 922–925 (1999)
14. Ducker, W.A., Senden, T.J., Pashley, R.M.: Measurement of forces in liquids using a force microscope. Langmuir **8**(7), 1831–1836 (1992)
15. Butt, H.-J., Jaschke, M., Ducker, W.: Measuring surface forces in aqueous electrolyte solution with the atomic force microscope. Bioelectrochem. Bioenerg. **38**(1), 191–201 (1995)
16. Melcher, J., et al.: Origins of phase contrast in the atomic force microscope in liquids. Proc. Natl. Acad. Sci. **106**(33), 13655–13660 (2009)
17. Wu, Y., Gupta, C., Shannon, M.A.: Effect of solution concentration, surface bias and protonation on the dynamic response of amplitude-modulated atomic force microscopy in water. Langmuir **24**(19), 10817–10824 (2008)
18. Goldstein, H., Poole Jr., C., Safko, J.: Classical Mechanics. Addson-Wesley, Reading (1981)
19. O'Shea, S.J., Welland, M.E.: Atomic force microscopy at solid – liquid interfaces. Langmuir **14**(15), 4186–4197 (1998)
20. Hutter, J.L., Bechhoefer, J.: Calibration of atomic-force microscope tips. Rev. Sci. Instrum **64**(7), 1868–1873 (1993)
21. Nnebe, I., Schneider, J.W.: Characterization of distance-dependent damping in tapping-mode atomic force microscopy force measurements in liquid. Langmuir **20**(8), 3195–3201 (2004)
22. Prakash, S., Yeom, J.: Nanofluidics and Microfluidics: Systems and Applications. Elsevier, Waltham (2014)
23. Conlisk, A.T.: Essentials of Micro- and Nanofluidics: With Applications to the Biological and Chemical Sciences. Cambridge University Press, Cambridge (2012)
24. Prakash, S., Karacor, M.B., Banerjee, S.: Surface modification in microsystems and nanosystems. Surf. Sci. Rep. **64**(7), 233–254 (2009)
25. Prakash, S., Karacor, M.B.: Characterizing stability of "click" modified glass surfaces to common microfabrication conditions and aqueous electrolyte solutions. Nanoscale **3**(8), 3309–3315 (2011)
26. Bhushan, B., Wang, Y., Maali, A.: Boundary slip study on hydrophilic, hydrophobic, and superhydrophobic surfaces with dynamic atomic force microscopy. Langmuir **25**(14), 8117–8121 (2009)
27. Lauga, E., Brenner, M., Stone, H.: Microfluidics: the no-slip boundary condition. In: Handbook of Experimental Fluid Dynamics. Springer, New York (2005)
28. Wu, Y., et al.: Dynamic response of AFM cantilevers to dissimilar functionalized silica surfaces in aqueous electrolyte solutions. Langmuir **26**(22), 16963–16972 (2010)
29. Prakash, S., et al.: "Click" modification of silica surfaces and glass microfluidic channels. Anal. Chem. **79**(4), 1661–1667 (2006)

Dynamic Force Spectroscopy

▶ Dynamic Force Measurements on Modified Surfaces with AFM

E

E-Beam Lithography (EBL)

► Electron Beam Lithography (EBL)

ECM

► Electrochemical Machining (ECM)

Ecotoxicity

► Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae
► Toxicology: Plants and Nanoparticles

Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates

Mélanie Auffan[2,3], Catherine Santaella[3,5], Alain Thiéry[3,6], Christine Paillès[2,3], Jérôme Rose[8,9], Wafa Achouak[3,5], Antoine Thill[3,7], Armand Masion[1,3], Mark Wiesner[3,4] and Jean-Yves Bottero[1]
[1]Aix-Marseille Université, CNRS, IRD, CEREGE UM34, Aix–en–Provence, France
[2]CEREGE, UMR 6635 CNRS/Aix-Marseille Université, Aix–en–Provence, France
[3]iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France
[4]CEINT, Center for the Environmental Implications of NanoTechnology, Duke University, Durham, NC, USA
[5]Laboratoire d'Ecologie Microbienne de la Rhizosphère et d'Environnements Extrême, UMR 6191 CNRS–CEA–Aix–Marseille Université de la Méditerranée, CEA Cadarache, St Paul lez Durance, France
[6]IMEP, UMR 6116 CNRS/IRD, Aix-Marseille Université, Marseille, France
[7]Laboratoire Interdisciplinaire sur l'Organisation Nanométrique et Supramoléculaire, UMR 3299, CEA/CNRS SIS2M, Gif–sur–Yvette, France
[8]CEREGE UMR 6635–CNRS–Université Paul Cézanne Aix–Marseille III, Aix–Marseille Université, Aix–en–Provence, France
[9]GDRI ICEINT: International Center for the Environmental Implications of Nanotechnology, CNRS–CEA, Aix–en–Provence, France

Synonyms

Environmental toxicology; Nano-ecotoxicology; Toxicity of metal and metal oxide nanoparticles on Prokaryotes (bacteria), unicellular and invertebrate Eukaryotes

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Definition

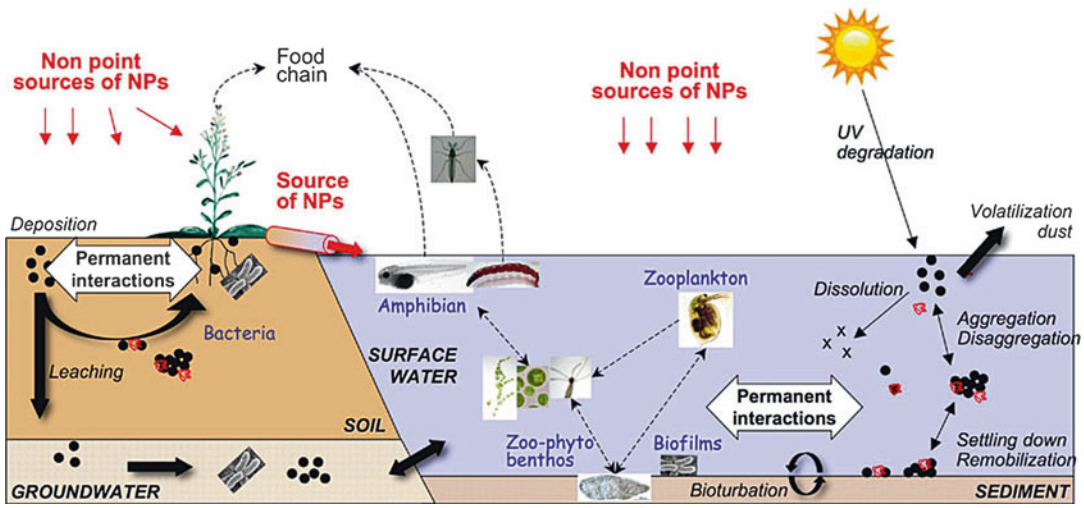
A toxic product is a chemical compound that harms the environment by affecting the biological organisms. Due to their novel properties, nanoparticles (NPs) cannot be considered as other organic and inorganic xenobiotics in the environment, e.g., pesticides, HAPs and PCBs dissolved metals, or also medicines. NPs have a mass, a charge, and above all a surface area. They are subject to phenomena of classical and quantum physics. Their reactivity means that their surface atoms are labile, easily change their redox state, and highly reactive with respect to compounds in the environment. Considering the huge range of applications using NPs, it seems reasonable to expect their dissemination in the environment at each step in their life cycle, from design through production to use and disposal of finished products. To date, the data available show that NPs can cross biological membranes and distribute themselves within different compartments of living organisms, or can also induce a remote toxicity. Consequently, there is a need to elucidate how the NPs can lead to adverse effects on organisms in their natural environment, considering not only their effects on target organs and life traits, but also their fate and transfers within the food webs (Fig. 1) (Box 1).

Box 1: Ecosystem and Food Chains

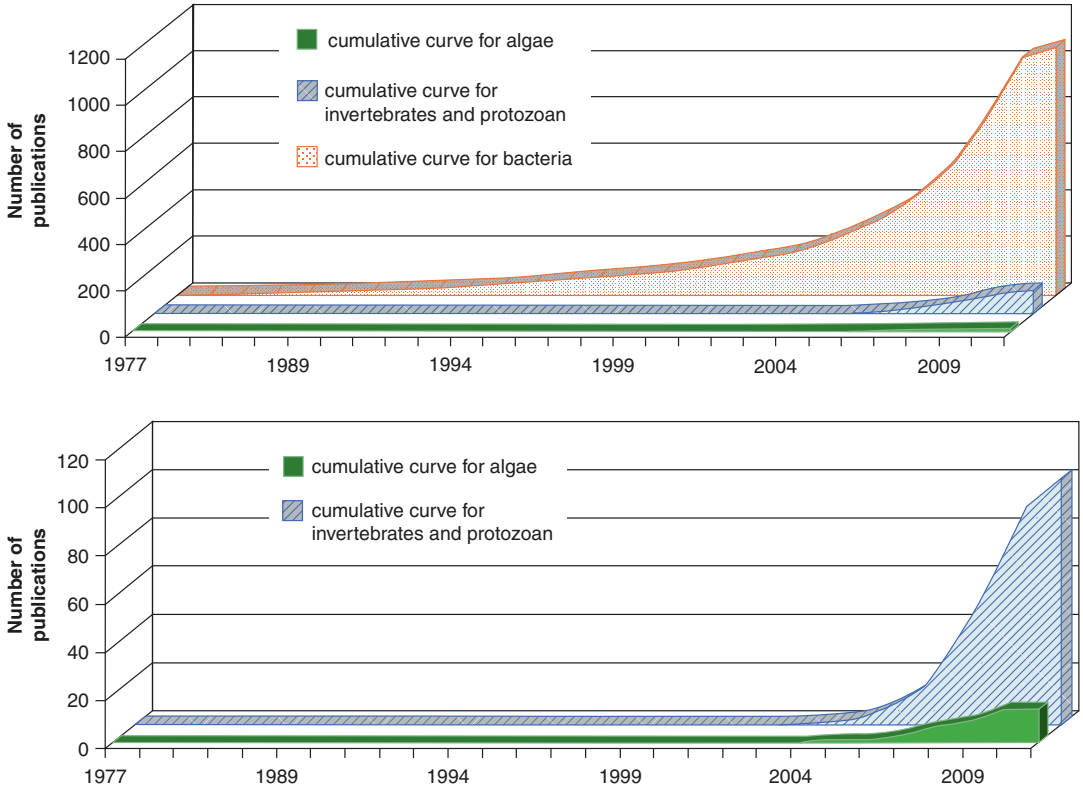
An ecosystem is the basic functional unit in ecology, as it includes both organisms and their abiotic environments (Fig. 1). It represents the highest level of ecological integration based on energy and biomass transfers. It is defined as a specific unit of all the organisms occupying a given area that interacts with the physical environment producing distinct trophic structure, biotic diversity, and material cycling.

Overview

The number of available publications (Fig. 2) dealing with the environmental impacts of NPs (called nano-ecotoxicology [1]) is presented in Fig. 2. Studies regarding algae, protozoa, and invertebrates are more recent (since 2006) than the ones regarding bacteria (which started in the 1980s) (*ISI Web of Knowledge source*) (Box 2). As a consequence, researches on bactericidal effects are more advanced (already at the scale of biofilms and the microbial community



Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates, Fig. 1 Fate and transport of nanomaterials within aquatic and terrestrial ecosystems and the relationships with the living organisms



Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates, Fig. 2 Evolution of the nano-ecotoxicology studies focused on the algae, invertebrates and protozoan, and bacteria since the 1980s (based on the ISI Web of Knowledge database)

structures), while for the other models nano-ecotoxicology is still in its infancy (at the individual scale).

About 30 species of invertebrates are studied regarding the three main ecosystemic domains: freshwater, marine, and terrestrial. For each domain, the Crustacean Cladocera Daphniidae (freshwater), the Mollusks Bivalvia Mytilidae, and Annelids Polychaetes (marine) are the most studied invertebrates, while the terrestrial (earthworms and isopods) examples are scarce.

Considering the three-domain system that divides cellular life into archaea, bacteria, and eukaryote domains (based on differences in 16 S rRNA genes), the studied models are mostly Eukaryotes belonging to Opisthokonts Animals, Alveolata (Ciliata), Plants Chlorophyta (green algae), or Prokaryotes with planktonic bacteria. Following the Whittaker's model distinguishing the differences in nutrition, most of the studied invertebrates belong to the ingestion nutrition group (active suspensory feeders (Daphniidae), suspension feeders (Mytilidae), or carnivorous (Cnidaria, Polychaeta)), others are detritivorous (*Blaberus*), deposit feeders (*Eisenia*), or fluid suckers (*Caenorhabditis elegans*), and most of the autotrophic organisms are the eukaryotic photosynthetic green algae (Chlorophyta). Protozoa are unicellular heterotrophic organisms, belonging to the Ciliophora group, feeding by phagotroph.

Box 2: Bacteria as a Good Model for Ecotoxicology

Life on earth evolved from bacteria and life on earth depends on bacteria, as they make the essential elements oxygen, carbon,

(continued)

nitrogen, and sulfur available for other life forms on our planet. They play important roles in the global ecosystem. They are key players of the breakdown of dead organisms, and the release of nutrients back into the environment. Each human body hosts ten microorganisms for every human cell. Microbes inhabiting humans have been recently identified and shown to play a beneficial role in health by contributing to human metabolism while in turn they benefit from the nutrient-rich niche in the intestine. Since bacteria constitute the largest biological surface, nanoparticles (NPs) may encounter in different ecosystems from air, soil, water to insect and animal bodies. Their contribution to NPs fate in different ecosystems has to be taken into account, as they may be involved in NPs transformation, sequestration, solubilization, or aggregation.

The biology and functional ecology of bacteria in different ecosystems make them good models for NPs toxicity studies. NPs may show deleterious effect on microbial physiology and survival, altering hence ecosystems functioning. It is hence worth investigating NPs effects of exposure and subsequent bioaccumulation to determine their toxicity on microbial communities and model bacteria.

Key Principles

Not intended to be an exhaustive list of results, this article discusses the relationships between the potential ecotoxicity of NPs with respect to their physicochemical properties, the organism functions, and the relationships between species. It has been recently shown that the surface reactivity and the chemical destabilization of metallic and metal oxide NPs were involved in their potential toxicity. NPs reactivity can be spontaneous (e.g., adsorption, oxidation, or reduction) or induced by an external event (e.g., the photocatalytic activity of TiO_2) [2]. Chemical destabilization can lead to redox changes within the lattice structure, the generation of reactive oxygen species (ROS), and to the release of ions in solution (e.g., Zn ions). The dissolution can be either reductive (e.g., the reduction of $Ce^{4+}O_2$ NPs into Ce^{3+}) or oxidative (e.g., Ag^0 NPs oxidation into Ag^+, Cu^0 NPs oxidation into Cu^{2+}) [3]. All these modifications can be critical events if the protective mechanisms of unicellular or pluricellular organisms are overwhelmed by a strong affinity with NPs. To highlight the relationship between surface reactivity and the ecotoxicological effects, this article is focused on metallic and metal oxide NPs.

Key Research Findings

Photocatalytic NPs: The Example of TiO_2

TiO_2 NPs (anatase and rutile) are the most used photocatalytic NPs as self-cleaning, opacifying, antimicrobial, antifouling agents, or UV-absorber in cosmetics, paints, etc., (see [4] and references therein). Recent studies have shown that they can be disseminated in the environment. For instance, the leaching out of TiO_2 NPs from exterior facades releases few $\mu g.L^{-1}$ of TiO_2 into stream water [5].

Bacteria

The ecotoxicity of TiO_2 NPs has been studied on rod-shaped *Escherichia coli*, *Pseudomonas fluorescens*, *Vibrio fischeri*, *Salmonella typhimurium*, bacillus-shaped *Nitrosomonas europaea* Gram-negative bacteria, and *Bacillus subtilis* Gram-positive bacteria [6, 7].

The mechanisms of ecotoxicity of TiO_2 mainly rely on a photocatalytic process at the surface of NPs generating ROS. The toxicity requires a contact between bacteria and TiO_2 NPs leading to the loss of membrane integrity and an increase in outer membrane permeability. The concentration of TiO_2 affecting bacterial survival in photocatalytic conditions varies between 10^{-3} and 5×10^3 $mg.L^{-1}$ [8]. For *E. coli*, the antimicrobial activity depends on biological parameters such as physiological state, generation, and initial concentration of bacteria [9] as well as the

chemical structure and the shape of the TiO_2 NPs. Surprisingly, membrane damages were also observed in the absence of light, e.g., [10].

The uptake of well-dispersed TiO_2 NPs was observed by transmission electron microscopy in *S. typhimurium* [7] treated with 8 and 80 ng.mL^{-1} of TiO_2 for 60 min. This induced a weak mutagenic potential in *S. typhimurium* strains (TA98, TA1537). Mutagenic effects were also observed toward *E. coli* (WP2uvrA) [7, 11].

Size-dependent toxicity of TiO_2 NPs is controversial. Studies have shown that particle size did not affect antibacterial activity [12, 13]. However, Ivask et al. have shown that whereas nano and bulk TiO_2 both decrease viability of bacteria at high concentrations (>4,000 mg.L^{-1}), the TiO_2 NPs were the only ones inducing bioluminescence of the superoxide anion sensing bacteria from 100 mg.L^{-1} [14]. Another size effect regards the size-dependent point of zero charge (PZC) of TiO_2, with a PZC of 4.8 for the smallest particles (∅ 3.6 nm) and higher PZC values (6.2, 6.5, 6.8) for the largest ones (∅ 8, 25, 200 nm) [6]. Since many experiments are performed at pH 5.5–6.8, exposure to bacteria and the adsorption onto their walls will be sensitive to this change in PZC [13]. Moreover, initially agglomerated TiO_2 NPs can be dispersed by bacteria as *P. fluorescens*. Such phenomenon can enhance the transport of NPs in the environment [15].

Algae

Since 2006, less than 20 papers illustrate the effects of NPs on marine and freshwater primary producers (algae). Overall, these studies demonstrate that primary producers (algae) are among the more sensitive forms of aquatic life to metallic NPs. The most studied freshwater green algae are *Pseudokirchneriella subcapitata*, but the results are controversial. Warheit et al. [16] observed acute effects with EC_{50} values of 21 mg.L^{-1} for TiO_2 NPs (∅ 140 nm), and 72 h EC_{50} of 5.83 mg.L^{-1} for TiO_2 NPs (∅ 25–70 nm). However, no toxicity was found at concentrations up to 100 mg.L^{-1} for TiO_2 NPs (∅ <40 nm) [17]. The absence of ecotoxicity is attributed to the rapid aggregation/coagulation [17], whereas the toxicity is related to entrapping of algal cells by aggregated TiO_2 NPs [18]. Such controversy highlights that many studies address the question of ecotoxicity but very few concern the mechanisms of toxicity or internalization.

The effects of particle size and UV photoactivation were assessed using TiO_2 (anatase, ∅ 25 and 100 nm) on freshwater green algae *Desmodesmus subspicatus* [19]. The EC_{50} value was 44 mg L^{-1} for the 25 nm NPs and greater than 50 mg L^{-1} for the 100 nm-TiO_2. UV photoactivation caused no additional effects. In some cases, the presence of TiO_2 NPs can also be beneficial. It increases the dry weight, chlorophyll synthesis and metabolism in photosynthetic organisms.

Finally, in a recent study [20], the toxicity of TiO_2 NPs was assessed for the first time on marine diatoms, chlorophyceae, and prymnesiophyceae. Ninety six hours NOEC were observed on all species in presence of TiO_2 NPs (anatase, ∅ 15–20 nm, 10–1,000 µg.L^{-1}).

Invertebrates

Terrestrial Invertebrates Ecotoxicological studies on terrestrial invertebrates are limited. They mostly regard the crustacean isopod *Porcellio scaber* [21, 22], the model *C. elegans*, and the earthworm *Eisenia fetida* [23, 24]. TiO_2 NPs (∅ from 15 to <50 nm) spread over the surface of dry leaf (1 mg TiO_2.g^{-1} of food for 14 days) did not induce effect on mortality, weight change, or feeding behavior toward *Porcellio*. However, *E. fetida* exposed to artificial soils containing more than 1 g.kg^{-1} of TiO_2 NPs for 7 days bioaccumulates and exhibit significant mitochondria damage. Moreover, a strong inhibition of the cellulase activity of *E. fetida* occurred that directly impacted the digestion. However, it has not been clarified whether cellulase originates from the earthworms or the symbionts.

Freshwater Invertebrates The adverse effects of TiO_2 NPs on aquatic invertebrates have been more studied than other inorganic NPs. In freshwater crustaceans, only branchiopods have been tested: the cladocerans *Ceriodaphnia dubia*, *Chydorus sphaericus*, and the anostracans

Thamnocephalus platyurus, as well as the brine shrimp *Artemia*.

Daphniid *Daphnia magna* are the most tested animals. TiO_2 NPs were mostly detected in their gut, but no clear conclusions can be drawn on the size-dependent toxicity of NPs. However, the aggregation state is important to consider [16]. Variable 48 h EC_{50} and LC_{50} ranging from 5.5 to 20,000 $mg.L^{-1}$ have been reviewed [1]. Others studies regard *Hydra attenuata*, the Gastropod *Physa acuta*, and midge larvae *Chironomus riparius*. For instance, sublethal doses of alumina coated-TiO_2 NPs reduced embryo growth rate and hatchability, and induced developmental deformities of the embryos on the freshwater snail *Physa acuta* [25]. Embryonic growth and hatchability tests are useful endpoints to assess chronic toxicity and toxic thresholds of NPs in sediment environment. Although no snail mortality was observed significant change in antioxidant levels were observed at 0.05 or 0.5 $g.kg^{-1}$. In this study, the exposure duration appeared as a significant factor. Consequently, long-term exposure has to be taken into account since alterations of ecological populations and ecological functionality are possible.

Marine Invertebrates A significant decrease in casting rate, increase in cellular and DNA damage in coelomocytes was measured toward the sediment dwelling marine Polychaetes *Arenicola marina* [26]. Using Raman Scattering Microscopy, TiO_2 aggregates of diameter >200 nm were located in the lumen of the gut and attached to the outer epithelium. On the bivalve *Mytilus galloprovincialis*, 24 h in vivo tests highlight that TiO_2 NPs (∅ 22 nm, 51 $m^2.g^{-1}$) induced lysosomal membrane destabilization in the hemocytes and digestive gland [27]. TiO_2 also increased the catalase and glutathione transferase activity in the digestive glands (not in the gills) which implies an oxidative stress mechanism.

Chemically Unstable NPs

Bacteria

Metallic Ag NPs have antimicrobial properties, and are the most commonly used nanomaterials in consumer products (see [28, 29] and references therein). The widespread use of Ag NPs will increase the load up of silver into wastewater treatment plants. Predicted environmental concentration (PEC) for Ag NPs are 0.5–2 ngL^{-1} in surface waters, 32–111 ngL^{-1} in sewage treatment plant effluents, and 1.3–4.4 $mg.kg^{-1}$ in sewage sludge. Recently, silver sulfide NPs have been evidenced in the reducing and sulfur-rich environment of sewage sludge.

A wide range of Ag NPs sizes, shapes, crystal structure, and surface chemistry were exposed to bacteria at different pH and ionic strength. The toxicity proceeds through release of free silver ions associated with specific effects of Ag NPs. Ag ions inhibit the oxidation of endogenous substrates by *E. coli* and cause efflux of accumulated phosphates and metabolites. Ag^+ interferes with the respiratory chain, induces the collapse of the proton motive force, the de-energization of the cell, the disruption of DNA replication, and disturbs phosphate uptake. Silver binds to proteins or enzymes and also competes for cell surface binding sites of bacterial cells with metal ions, such as copper. Ag NPs attach to the surface of the Gram-negative membranes, can penetrate inside the bacteria, and drastically disturb the cell permeability and respiration, dissipate the proton motive force, collapse the plasma membrane potential, and deplete intracellular ATP levels. The mode of action of Ag NPs (∅ 9.3 nm) 0.4 nM is similar to that of Ag^+ ions 6 μM. Free silver and Ag NPs generate ROS and cell membrane damage. Using superoxide anion sensor *E. coli* [14] showed that generation of ROS by Ag NPs contributed to the toxicity. Interestingly, extracellular polymeric substances produced by bacteria, neutralized free silver ions, and decreased the toxicity of Ag NPs [30]. River humic acids also mitigated the bactericidal action on *P. fluorescens*.

Size-, crystallinity-, and shape-dependent effects were observed in terms of antibacterial activity and ROS production. Direct interaction with the bacteria preferentially occurred for NPs ranging within 1–10 nm. Among tested Ag NPs sizes (10, 30–40, and 100 nm), only the 10 nm NPs inhibited methicillin-resistant *S. aureus* (the main cause of nosocomial infections worldwide) [31].

Quantum dots (QDs) are colloidal semiconductors with composition- and size-dependent absorption and emission. They are used as biological probes cellular imaging, tumor targeting, and diagnostics. The minimum inhibitory concentrations of ZnO NPs against *E. coli*, *Pseudomonas aeruginosa*, and *Staphylococcus aureus* were found between 125 and 500 $\mu g.mL^{-1}$ [32]. ZnO NPs induce toxicity via the dissolution and release of Zn^{2+} ions [6, 33]. The toxicity depends on the media composition (pH, ionic strength) and the illumination conditions. Both Gram-negative and Gram-positive membranes are disorganized after interactions with ZnO NPs [6, 34]. The increase in membrane permeability leads to the internalization of NPs [34]. A study using mutant strains has shown that CdSe and CdSe/ZnS QDs (∅ <5 nm) can penetrate into *B. subtilis* and *E. coli* upon purine-processing mechanisms [35]. Extracellular polymeric substances excreted by bacteria prevent NPs internalization and decrease the toxicity observed [36]. Interestingly, ZnO NPs inside the cell are less crystallized than the initial NPs [34]. This raises the question of whether there is a direct internalization of the NPs, or a dissolution/precipitation within the cells.

CdTe QDs also induce toxicity toward Gram-positive methicillin-resistant *Shewanella aureus* and *B. subtilis*, with EC_{50} concentrations tenfold higher than toward Gram-negative strains *E. coli* and *P. aeruginosa* [37]. CdTe NPs bind to bacteria and impair the antioxidant system by photogeneration of ROS [37]. In that case, the release of Cd^{2+} has little or no role in toxicity of CdTe NPs [37], but the formation of TeO_2 and CdO by surface oxidation may cause cytotoxicity [38].

CeO_2 NPs are used as diesel fuel additives, polishing surface agents, as well as in coating and microelectronics. CeO_2 are nonstoichiometric NPs with cerium atoms in +4 and +3 oxidation states, and oxygen vacancies [39]. EC_{50} was near 5 mg L^{-1} for *E. coli* and ranged from 0.27 to 6.3 $mg.L^{-1}$ for *Anabaena* in pure water [40]. Under exposure to CeO_2 NPs, *N. europaea* cells have shown larger sedimentation coefficient than the control [6]. The toxicity of CeO_2 NPs is exerted by direct contact with cells [40], membrane damages [6], cell disruption [40], and release of free Ce^{3+} [41]. Even if free Ce^{3+} is toxic for *Anabaena* in pure water, it could not explain by itself the cytotoxicity observed. However, the reduction of the Ce^{4+} into Ce^{3+} at the surface of the CeO_2 NPs correlates with the cytotoxicity toward *E. coli*. No oxidative stress response was detected with *E. coli* or *B. subtilis*, but CeO_2 NPs and $CeCl_3$ alter the electron flow, and the respiration of the bacteria [42]. Pelettier et al. [42] also observed the disturbance of genes involved in sulfur metabolism, and an increase in the levels of cytochrome terminal oxidase (*cydAB*) transcripts known to be induced by iron limitation.

Algae

The short-term toxicity of Ag NPs (∅ 25 nm) and $AgNO_3$ was tested toward the freshwater algae *Chlamydomonas reinhardtii* [43]. The toxicity measured by EC_{50} for Ag^+ was 18 times larger than for Ag NPs and much higher than those determined to inhibit algal growth in other studies. The toxic effects observed here reflect toxicity to photosynthesis. TEM showed no aggregation of Ag NPs. However, the measured Ag^+ concentration in the suspension of Ag NPs by itself cannot account for the observed toxicity since only 1 % of the Ag had been oxidized [43]. Interestingly, once normalized by the concentration of dissolved Ag^+, the toxicity of Ag NPs was more important than that of $AgNO_3$. This suggests the importance of interactions between the surfaces of Ag NPs and algae cells.

In contrast, Ag^+ released from the oxidative dissolution of Ag NPs (∅ 60–70 nm) severely suppressed cell growth, photosynthesis, and chlorophyll production in marine diatom *T. weissflogii* [44]. In both studies [43, 44] Ag NPs were not found to be toxic to either freshwater or marine unicellular algae at concentrations that could be expected in contaminated water bodies.

Regarding ZnO NPs, their toxicity toward the freshwater algae *P. subcapitata* was solely attributed to rapid dissolution at pH 7.6 with a saturation solubility in the range of 1 mg L [45]. Toxicity of ZnO NPS (∅ 20–30 nm) was assessed on marine phytoplankton (diatoms, chlorophytes, and prymnesiophytes) [20]. Even in saline water,

the toxicity was related to dissolution, release, and uptake of free Zn ions.

Protozoa

Studies on adverse effects of NPs against unicellular eukaryotes are scarce. The toxicity of Ag NPs (Ø 30–40 nm) on the Ciliate Protozoa *Paramecium caudatum* was reported by Kvitek et al. [46]. No toxicity was observed below 39 $mg.L^{-1}$, whereas ionic silver induced toxicity above 0.4 $mg.L^{-1}$. The authors noticed that the surfactant/polymer modification strongly affected the toxic behavior of Ag NPs.

The toxicity of CuO NPs (30) was investigated on Protozoans *Tetrahymena thermophila* [47]. A clear accumulation of the CuO NPs was observed in their food vacuoles. In all cases toxic effects of Ag and CuO NPs to protozoa were caused by their solubilized fraction. Interestingly, experiment performed in natural water versus artificial water remarkably (up to 140-fold) decreased the toxicity of CuO NPs depending mainly on the concentration of dissolved organic carbon.

Invertebrates

Terrestrial Invertebrates The effects of Cu NPs (∅ 20–40 nm, and <100 nm) and Cu ions were assessed on *Eisenia* in artificial soil [48]. Both Cu ions and NPs are bioaccumulated suggesting a route of entry into terrestrial food chains. No subchronic toxicity of Cu NPs was observed at concentrations up to 65 $mgCu.kg^{-1}$ soils. Increases in metallothionein expression occurred from 20 mg kg^{-1} of Cu NPs to 10 mg kg^{-1} of $CuSO_4$. An oxidized and amorphous layer at the surface of the Cu NPs appears immediately in the soil and depends on particle size (XANES-based experiment). This highlights the necessity to consider the surface chemistry of Cu NPs during nano-ecotoxicology studies since the toxicant is likely to be more reflective of CuO than metallic Cu. *Eisenia* was also exposed to $AgNO_3$ and Ag NPs (PVP-coated) [49]. *Eisenia* consistently avoid soils contaminated with Ag NPs or $AgNO_3$ with faster avoidance for $AgNO_3$. In natural soil, *Eisenia* accumulated significantly higher concentrations of Ag than in artificial soil. Ag NPs incubated in the soil were 10–17 % Ag(I), suggesting that Ag ions may be responsible for the effects observed with Ag NPs (EXAFS-based experiment). The earthworm *Lumbricus terrestris* was exposed to Ag NPs. Increased apoptotic activity was detected in a range of tissues both at acute and sublethal concentrations (down to 4 $mg.kg^{-1}$ soil). Comparing exposure in water and soil showed reduced bioavailability in soil reflected in the apoptotic response [50]. Toxicity of Ag NPs was also observed on *Drosophila melanogaster* in terms of egg development and pupae hatching. Ahamed et al. [51] show that coated-Ag NPs induce heat shock stress, oxidative stress, DNA damage, and apoptosis toward *D. melanogaster*.

C. elegans is a good model to assess the toxicity of chemically instable NPs at low doses. For Ag NPs, growth inhibition occurs at concentrations in the low $mg.L^{-1}$ levels [52], and the metallothionein-deficient strain exhibited greater sensitivity than wild-strain. Both Ag and ZnO NPs release Zn and Ag ions that are responsible for most of the toxicity, but the NPs themselves did also cause toxicity on the nematode [52, 53]. CeO_2 NPs (∅ 8.5 nm) also affect *C. elegans* at low doses inducing a decrease in the lifespan (−12 % at 1 nM), ROS accumulation, and oxidative damage [54]. The best developed paradigm to explain these toxic effects are an oxidative stress and the ROS generation when the CeO_2 NPs reversibly alternate between $Ce_{(3+)}$ and $Ce_{(4+)}$ [54]. To date, further studies are needed to demonstrate whether the surface reactivity of CeO_2 NPs could trigger the generation of ROS or not.

Freshwater Invertebrates Heinlaan et al. [55] compared the toxicity of Cu and ZnO NPs between two branchiopods, *Daphnia magna* and an anostraca *Thamnocephalus platyurus*. The L $(E)C_{50}$ values of Cu NPs (90–224 $mg.Cu.L^{-1}$) are greatly higher than ZnO NPs (1.1–16 $mg.L^{-1}$). The toxicity of Cu NPs was also compared to Ag NPs [56] on *Ceriodaphnia dubia*. Whatever the water matrix tested, Ag and CuO NPs were highly toxic to the water flea. The lower toxicity of Ag NPs is explained by the silver complexation with dissolved organic compounds naturally present in the tested waters [57].

A TEM histological study on *Daphnia magna* observed the dispersion of Cu NPs in the midgut [58]. Questions raised on the respective localization of NPs between the peritrophic membrane and the midgut epithelium microvilli. This was completed by a study on the biokinetic uptake and efflux [59] showing that more than 70 % of Ag NPs accumulate in the daphnids through ingestion. This highlights the importance of NPs transport along the aquatic food chain.

One study focused on *Hydra attenuata* has shown that CdSe/Cd QDs induced non-synchronous tentacle retraction. The tentacle-writhing behavior is related to the position of metal core inside the QDs [60]. The organic coated QDs were able to enter into the head after 1 h, and the entire body within 24 h. The freshwater mussels, *Elliptio complanata*, are also disturbed by CdTe QDs: decrease in viability, lipid peroxidation of the gills and gut, and fall down of the immune activity of hemocytes [61]. The toxic effects are linked to the toxicity of their metal core, the coating, and dissolution rate. However, Velzeboer et al. [17] do not observe clear effects at concentrations up to 100 $mg.L^{-1}$ toward the cladoceran *Chydorus sphaericus* exposed to CdSe/ZnS QDs. This lack of toxicity is explained by the aggregation of the NPs that decreases the bioavailability.

Ag NPs are efficient larvicides for mosquitoes involved in malaria fever. Sap-Iam et al. [62] found that 90 % of *Aedes aegypti* larvae died after 3 h contact with 5 $mg.L^{-1}$ of Ag NPS (∅ 10 nm, PMA-coated). The involved mechanism is the penetration of Ag NPs through the larval membrane. Ag NPs in the intracellular space can bind to sulfur-containing proteins or phosphorous-containing compounds like DNA, leading to the denaturation of organelles and enzymes. However, Ag NPs had no effects on the hatchability of mosquito eggs due to the chitinized serosal cuticle that prevents the penetration within the eggs.

Marine Invertebrates A comparison between CuO NPs and the soluble form of Cu was done on *Scrobicularia plana* and *Nereis diversicolor* [63]. The annelid and bivalve organisms showed different behaviors and biochemical responses, but oxidative stress was observed for both species (increase in catalase and glutathione transferase activities).

Toxicity of QDs was also assessed on a coastal marine and salt lake rotifer, *Brachionus manjavacas* [64]. Population exposed to 0.3 $\mu g.mL^{-1}$ of QDs (∅ 37 nm) exhibited 50 % of decrease in the growth rate while larger particles (∅ up to 3,000 nm) caused no change. These larger particles remained confined in the gut, indicating NPs sizes as a critical factor in the internalization by tissue. Moreover, transfer of the F_1 offspring from NPs-exposed maternal females into NPs-free media demonstrated that NPs are cleared from the animals without residual adverse effects [64].

Chemically Stable NPs: The Example of Gold

Metallic Au NPs are used as contrast agents in electron microscopy, optical sensors, catalysts, and for therapeutic uses. Au metal is extremely resistant to oxidation, and essentially insoluble under ambient conditions. Although Au is resistant to oxidative dissolution, excess Au ions can remain in the suspension at the end of the synthesis process. Therefore, the possibility that Au associated with organisms was taken up as Au ions cannot be discounted without a determination of the speciation of Au in the exposure media [65].

Bacteria

Toxicity of Au NPs has been studied on *S. typhimurium*, *Salmonella enteritidis*, *Streptococcus mutans*, *E. coli*, and *Listeria monocytogenes*. Ammonium functionalized Au NPs (∅ 2 nm) showed 3.1 ± 0.6 μM LC_{50} on *E. coli*. The toxicity is related to the strong electrostatic attraction between the negatively charged bilayer and the positively charged NPs [66]. This affinity for the surface was also observed with *S. typhimurium* strain TA 102 [67]. No toxic or mutagenic effects were observed. However, Au NPs demonstrated a Minimum Inhibitory Concentration (MIC) at 97 $\mu g.mL^{-1}$ on *S. mutans*, which is a possible infectious etiology in dental caries. As a comparison, silver showed an average MIC of 4.86 $\mu g.mL^{-1}$.

E

Algae

One study focused on the impacts of Au NPs (∅ 10 nm, amine-coated) toward the marine algae *Scenedesmus subspicatus*. LD_{50} value was reached at 1.6 10^5 Au NPs $cells^{-1}$ (∼2.4 mg. L^{-1}). Au NPs were strongly adsorbed onto the cell wall leading to progressive intracellular and wall disturbances [68].

Invertebrates

Terrestrial Invertebrates Few in vivo studies document the interactions between Au NPs and terrestrial organisms. One recent paper evidences that Au NPs are bioavailable from soil to the model detritivore *Eisenia fetida* and have the potential to enter terrestrial food webs [65]. They are distributed among tissues, and affect the earthworm reproduction. Using several sizes of Au NPs, it was found that primary particle size (∅ 20 or 55 nm) did not influence accumulated concentrations on a mass basis, while on a particle number basis the 20 nm NPs were more bioavailable. Differences in bioavailability can be related to the aggregation behavior in porous media [65].

The impacts of Au NPs (Ø 50 nm) on the central nervous system (CNS) of male cockroach *Blaberus discoidalis* was assessed [69]. Negatively charged Au NPs interacted with the roach, transferred inside the nerve cord within 17 days, and encapsulated by the proteins of the CNS. No major impact on the life expectancy was observed after 2 months, whereas the locomotion was affected.

The effects toward *Drosophila melanogaster* wild-type (Oregon R+) were also studied using Au NPs (citrate-capped) [70]. Ingested Au NPs induced a strong reduction of the lifespan and fertility, DNA fragmentation, and overexpression of the stress proteins HSP_{70}. Au NPs were homogeneously distributed through the deep layers of the enteric tissue and in the reproductive organs. Most of them were in the endosomes near the lamellar structures of the rough endoplasmic reticulum. In the germinal tissues, Au NPs were homogeneously distributed in the spermatocytes and flagella, but only in the cytoplasm of the ovarioles. No data are available on the motility of spermatozoids [70].

Freshwater Invertebrates The impacts of Au NPs (amine-coated, ∅ 10 nm) was investigated on the freshwater benthic bivalve *Corbicula fluminea* [68]. Contamination via trophic exposure (7 days with *Scenedesmus subspicatus*) induced internalization through the bronchial and digestive barriers. The histological penetration of Au NPs in the stomach cells (cytoplasm and nucleus), or in epithelial cells of digestive gland (lysosomal vesicles) depended on the speed and intensity of the digestion. Strong bioconcentration factors were obtained in gills and visceral mass, respectively. Interestingly, dose-dependent genetic expressions were observed on the gills of *Corbicula* but not in the visceral mass [68]. At low doses [1.6 10^2 Au NPs.$cells^{-1}$], there was an overexpression of genes associated to oxidative stress, while for [1.6 10^4 Au NP.$cells^{-1}$], there was also decrease in the mitochondrial activity and respiratory rate. For the highest concentration [1.6 10^5 Au NPs.$cells^{-1}$], biochemical mechanisms of detoxification were activated (changes in the glutathione transferase expression and metallothionein concentrations).

Marine Invertebrates Relatively little is known about how Au NPs nanoparticles interact with marine organisms. Au NPs (∅ 15 nm) incubated with *Mytilus edulis* were detected in the digestive gland due to impaired feeding [71]. This suggests that Au NPs cause a modest level of oxidative stress sufficient to oxidize thiols in glutathione and proteins but without causing lipid peroxidation or induction of thioredoxin reductase activity. However, the mechanisms by which Au NPs might cause oxidative stress are still poorly understood.

Interspecies Relationships

Effects on Biofilms and Bacterial Communities

Effects of NPs were assessed toward biofilms and communities. In these complex systems, the microbial responses do not always correlate with results from studies performed on monocultures in vitro. For instance, TiO_2 NPs from wastewater and seawater sludge have low toxicity on *V. fischeri* (see [4] and references therein).

However, low concentration (5.3 $mg.L^{-1}$) of TiO_2 NPs added to a stream microcosm under ambient UV radiation [72] damaged the cell walls of planktonic microorganisms due to intercellular action of ROS and accumulated in the benthic biofilms. Cell membrane damages were more pronounced in free-living cells than in biofilm cells, indicating the protective role of cell encapsulation against TiO_2 NPs. In biofilms, the surface properties of NPs govern their mobility, diffusion, and distribution, and structural parameters such as roughness and cell distribution of biofilm can be affected. Ag NPs also induced sloughing of *P. putida* biofilms. The presence of fulvic acid (which could play the role of a natural surfactant) reduced sloughing but increased uptake and bioaccumulation of the NPs [28]. Au NPs were also aggregated within flagella or biofilm network made from *Salmonella enteritidis* and *Listeria monocytogenes*. While bacterial cell walls and cytoplasm membrane were disintegrated, Au NPs did not penetrate into the bacteria [73].

Nitrifying bacteria are main players in wastewater treatment plants and in soil. Ag NPs concentrations lower than 1 $mg.L^{-1}$ inhibit nitrifying organisms taken from a nitrifying reactor. This decrease correlates with the fraction of Ag NPs less than 5 nm, the intracellular ROS concentrations, but not with the photocatalytic ROS fractions (see [29] and references therein).

Soil bacterial communities were also studied after exposure to ZnO NPs (0.05, 0.1, and 0.5 $mg.g^{-1}$ soil) in microcosms over 60 days. An alteration of the community structure and a decrease in the microbial biomass were observed [74]. The effects were higher than that of TiO_2 NPs, as reflected by lower DNA and stronger shifts in bacterial community composition, and modification of the soil enzyme activities [75]. However, while Ag NPs are bactericidal agents, bacterial community of estuarine sediments exposed to Ag NPs (for 20 days, 1 $mg.L^{-1}$, ∅ 58 nm) did not exhibit significant changes. This is related to the chloride amount in estuary water that alters the physicochemical behavior of Ag NPs [76] and the speciation of ionic silver.

Moreover, engineered NPs can be transferred from prey to predator. One study focused on bare CdSe QD accumulated in the *P. aeruginosa* bacteria, and transferred to and biomagnified in the *Tetrahymena thermophila* protozoa that prey on the bacteria. Cadmium concentrations in the protozoa predator were five times higher than their bacterial prey. However, QDs-treated bacteria affect the protozoa without lysing them. Consequently, intact QD remain available to higher trophic levels [77].

Effects of Invertebrate Assemblages

Land application of biosolids from wastewater treatment will be a major pathway for the introduction of manufactured NPs to the environment. The model organisms *Nicotiana tabacum* L. cv *xanthi* and *Manduca sexta* (tobacco hornworm) was used to investigate plant uptake and the potential for trophic transfer of Au NPs (∅ 5, 10, 15 nm) [78]. This study presents the first evidence of trophic transfer of NPs from a terrestrial primary producer to a primary consumer as well as the first evidence of biomagnification of NPs within a terrestrial food web by mean factors of 6.2, 11.6, and 9.6 for the 5, 10, and 15 nm treatments, respectively.

Recent studies focused on freshwater simplified food webs. The first one regards carboxylated and biotinylated QDs transferred to rotifers *Brachionus calyciflorus* through dietary uptake of ciliated protozoans *Tetrahymena* interacting with *E. coli* [79]. QDs accumulation was limited in the ciliates and no enrichment was observed in the rotifers. The contaminated *E. coli* formed aggregates that were not eaten by *Tetrahymena*. No toxicity was observed in single cells *Tetrahymena* nor in *Brachionus calyciflorus*. Another study regards the trophic transfer of TiO_2 NPs in the food chain including *Daphnia magna* and its predator, the zebra fish *Danio rerio* [80]. Eight-day-old *D. magna* were exposed to TiO_2 (∅ 21 nm, 0.1 and 1 $mg.L^{-1}$) and transferred to the 3-month-old *D. rerio* tanks. No biomagnification of TiO_2 NPs from *Daphnia* to zebra fish was found, but the dietary intake was found to be one of the main NPs exposure route for higher trophic level aquatic organisms.

However these studies regarding short trophic links do not take into account important

parameters such as the colloidal destabilization of the NPs, the interaction between the surface of the NPs and (in)organic molecules naturally occurring or bio-excreted, or the flux between compartments of the ecosystems (aqueous phase, sediments, biota). To work under more realistic scenarios of exposure, few studies are performed on mesocosms. Ferry et al. [81] were the first to study the fate of positively charged Au NPs (65 × 15 nm) in estuarine mesocosms containing sediments, biofilms, primary producers, filter feeders, grazers, and omnivores. While water and sediments represented 99 % of the total mass of the system, on a per mass basis the filter feeders and biofilms were the most effective sink of Au NPs. The estimated mass of biofilms was less than 0.5 % but they recovered 60 % of NPs. This high affinity is related to the negative surface charge of the biofilm interacting with the positively charged Au NPs and to the large surface area developed by the biofilms. These results are interesting since biofilms offer (1) a route into the food web through grazing by detritivores and (2) a route for mineralization through biofilm calcification.

Conclusion: Future Directions

Past experience with chemicals (e.g., methyl mercury, DDT, and PCBs) revealed dietary uptake at lower trophic levels and accumulation up the *food chain* to be an important route of contaminant exposure. All the observations that NPs can biomagnify highlight the importance of considering *dietary uptake* as a pathway for NPs exposure and raises questions about potential *ecoreceptor and human exposure* to NPs from long-term land and chronical exposure.

Ecotoxicity studies carried out at concentrations much higher than would ever be expected in the environment provide necessary data regarding the mechanisms of NPs toxicity. However complementary studies on the impact of NPs at concentrations comparable to those encountered in an environmental situation are required. To date, one challenge is to combine data from bioindicator organisms reflecting environmental health, and suitable biomarkers enlightening physiological processes [82].

Acknowledgements The authors would like to thank the CNRS and CEA for funding the International Consortium for the Environmental Implications of NanoTechnology and also the NSF and the US-EPA for funding the Center for the Environmental Implications of NanoTechnology. They also acknowledge financial support from the French National Agency (ANR) in the frame of the PNANO/NANAN, and P2N/MESONNET projects. We also apologize to the authors whose work was not cited. Due to the limitation of the number of references, we favored the citation of reviews and the most recent articles.

Glossary

EC_{50} The half maximal effective concentration referring to the concentration that induces a response halfway between the baseline and maximum after some specified exposure time.

EXAFS Extended X-ray absorption fine structure.

IC_{25} The inhibition concentration 25 % is the statistical analysis used in chronic whole effluent toxicity tests to estimate the sublethal effects of the effluent sample. IC_{25} estimates of the concentration causing 25 % reduction in growth or reproduction of test organisms.

LC_{50} The lethal concentration 50 % is the dose required to kill half the members of a tested population for a test duration. It is a general indicator of a substance's acute toxicity.

LD_{50} The median lethal dose 50 %.

MIC Minimum inhibitory concentration.

NOEC No observed effect concentration, also called a no observable effect level (NOEL).

PEC Predicted environmental concentration.

ROS Reactive oxygen species.

TEM Transmission electron microscope.

XANES X-ray absorption near edge structure.

Cross-References

▶ Effect of Surface Modification on Toxicity of Nanoparticles
▶ Fate of Manufactured Nanoparticles in Aqueous Environment
▶ Nanoparticle Cytotoxicity

▶ Nanostructures for Surface Functionalization and Surface Properties
▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity

References

1. Kahru, A., Dubourguier, H.-C.: From ecotoxicology to nanoecotoxicology. Toxicology **269**(2–3), 105–119 (2010)
2. Auffan, M., Rose, J., Bottero, J.Y., Lowry, G., Jolivet, J.P., Wiesner, M.R.: Towards a definition of inorganic nanoparticles from an environmental, health, and safety perspective. Nat. Nanotechnol. **4**, 634–641 (2009)
3. Auffan, M., Rose, J., Wiesner, M.R., Bottero, J.Y.: Chemical stability of metallic nanoparticles: a parameter controlling their potential toxicity in vitro. Environ. Pollut. **157**, 1127–1133 (2009)
4. Sharma, V.K.: Aggregation and toxicity of titanium dioxide nanoparticles in aquatic environment – a review. J. Environ. Sci. Health A Tox. Hazard. Subst. Environ. Eng. **44**(14), 1485–1495 (2009)
5. Kaegi, R., Sinnet, B., Zuleeg, S., Hagendorfer, H., Mueller, E., Vonbank, R., Boller, M., Burkhardt, M.: Release of silver nanoparticles from outdoor facades. Environ. Pollut. **158**(9), 2900–2905 (2010)
6. Fang, X.H., Yu, R., Li, B.Q., Somasundaran, P., Chandran, K.: Stresses exerted by ZnO, CeO_2 and anatase TiO_2 nanoparticles on the Nitrosomonas europaea. J. Colloid Interface Sci. **348**(2), 329–334 (2010)
7. Kumar, A., Pandey, A.K., Singh, S.S., Shanker, R., Dhawan, A.: Cellular uptake and mutagenic potential of metal oxide nanoparticles in bacterial cells. Chemosphere (2011). doi:10.1002/cyto.a.21085
8. Matsunaga, T., Okochi, M.: TiO_2-mediated photochemical disinfection of *Escherichia coli* using optical fibers. Environ. Sci. Technol. **29**, 501–505 (1995)
9. Rincón, A.-G., Pulgarin, C.: Bactericidal action of illuminated TiO_2 on pure *Escherichia coli* and natural bacterial consortia: post-irradiation events in the dark and assessment of the effective disinfection time. Appl. Catal. B Environ. **49**, 99–112 (2004)
10. Adams, L.K., Lyon, D.Y., Alvarez, P.J.J.: Comparative eco-toxicity of nanoscale TiO_2, SiO_2, and ZnO water suspensions. Water Res. **40**(19), 3527–3532 (2006)
11. Pan, X., Redding, J.E., Wiley, P.A., Wen, L., McConnell, J.S., Zhang, B.: Mutagenicity evaluation of metal oxide nanoparticles by the bacterial reverse mutation assay. Chemosphere **79**(1), 113–116 (2010)
12. Paoli, G.D.: A rapid GC-MS determination of gamma-hydroxybutyrate in saliva. J. Anal. Toxicol. **32**, 298–302 (2007)
13. Jiang, W., Mashayekhi, H., Xing, B.S.: Bacterial toxicity comparison between nano- and micro-scaled oxide particles. Environ. Pollut. **157**(5), 1619–1625 (2009)
14. Ivask, A., Bondarenko, O., Jepihhina, N., Kahru, A.: Profiling of the reactive oxygen species-related ecotoxicity of CuO, ZnO, TiO_2, silver and fullerene nanoparticles using a set of recombinant luminescent *Escherichia coli* strains. Anal. Bioanal. Chem. **398**(2), 701–716 (2010)
15. Horst, A.M., Neal, A.C., Mielke, R.E., Sislian, P.R., Suh, W.H., Madler, L., Stucky, G.D., Holden, P.A.: Dispersion of TiO nanoparticle agglomerates by *Pseudomonas aeruginosa*. Appl. Environ. Microbiol. **76**(21), 7292–7298 (2010)
16. Warheit, D.B., Hoke, R.A., Finlay, C., Donner, E.M., Reed, K.L., Sayes, C.M.: Development of a base set of toxicity tests using ultrafine TiO_2 particles as a component of nanoparticle risk management. Toxicol. Lett. **171**, 99–110 (2007)
17. Velzeboer, I., Hendriks, A.J., Ragas, A.M., van de Meent, D.: Aquatic ecotoxicity tests of some nanomaterials. Environ. Toxicol. Chem. **27**(9), 1942–1947 (2008)
18. Aruoja, V., Dubourguier, H.C., Kasemets, K., Kahru, A.: Toxicity of nanoparticles of CuO, ZnO and TiO_2 to microalgae *Pseudokirchneriella subcapitata*. Sci. Total Environ. **407**(4), 141–146 (2009)
19. Hund-Rinke, K., Simon, M.: Ecotoxic effect of photocatalytic active nanoparticles (TiO_2) on algae and daphnids. Environ. Sci. Pollut. Res. **13**, 1–8 (2006)
20. Miller, R.J., Lenihan, H.S., Muller, E.B., Tseng, N., Hanna, S.K., Keller, A.A.: Impacts of metal oxide nanoparticles on marine phytoplankton. Environ. Sci. Technol. **44**, 7329–7334 (2010)
21. Jemec, A., Tisler, T., Drobne, D., Sepcic, K., Jamnik, P., Ros, M.: Biochemical biomarkers in chronically metal-stressed daphnids. Comp. Biochem. Physiol. **147C**, 61–68 (2008)
22. Drobne, D., Jemec, A., Tkalec, Z.P.: In vivo screening to determine hazards of nanoparticles: nanosized TiO_2. Environ. Pollut. **157**, 1157–1164 (2009)
23. Hu, C.W., Li, M., Cui, Y.B., Li, D.S., Chen, J., Yang, L.Y.: Toxicological effects of TiO_2 and ZnO nanoparticles in soil on eartworm *Eisenia fetida*. Soil Biol. Biochem. **42**, 586–591 (2010)
24. Heckman, L.-H., Hovgaard, M.B., Sutherland, D.S., Autrup, H., Besenbacher, F., Scott-Fordsmand, J.J.: Limit-test toxicity screening of selected inorganic nanoparticles to the earthworm *Eisenia fetida*. Ecotoxicology **20**, 226–233 (2011)
25. Musee, N., Oberholster, P.J., Sikhwivhilu, L., Botha, A.-M.: The effects of engineered nanoparticles on survival, reproduction, and behavior of freshwater snail, Physa acuta (Draparnaud, 1805). Chemosphere **81**, 1196–1203 (2010)
26. Galloway, T., Lewis, C., Dolciotti, I., Johnston, B.D., Moger, J., Regoli, F.: Sublethal toxicity of nano-titanium dioxide and carbon nanotubes in a sediment dwelling marine polychaete. Environ. Pollut. **158**, 1748–1755 (2010)

27. Canesi, L., Fabbri, R., Gallo, G., Vallotto, D., Marcomini, A., Pojana, G.: Biomarkers in *Mytilus galloprovincialis* exposed to suspensions of selected nanoparticles (Nano carbon black, C60 fullerene, Nano-TiO_2, Nano-SiO_2). Aquat. Toxicol. **100**, 168–177 (2010)
28. Fabrega, J., Fawcett, S.R., Renshaw, J.C., Lead, J.R.: Silver nanoparticle impact on bacterial growth: effect of pH, concentration, and organic matter. Environ. Sci. Technol. **43**(19), 7285–7290 (2009)
29. Marambio-Jones, C., Hoek, E.: A review of the antibacterial effects of silver nanomaterials and potential implications for human health and the environment. J. Nanopart. Res. **12**(5), 1531–1551 (2010)
30. Dimkpa, C.O., Calder, A., Gajjar, P., Merugu, S., Huang, W., Britt, D.B., McLean, J.E., Johnson, W.P., Anderson, A.J.: Interaction of silver nanoparticles with an environmentally beneficial bacterium, *Pseudomonas chlororaphis*. J. Hazard. Mater. **188**, 428–435 (2011)
31. Ayala-Núñez, N., Lara Villegas, H., del Carmen Ixtepan Turrent, L., Rodríguez Padilla, C.: Silver nanoparticles toxicity and bactericidal effect against methicillin-resistant *Staphylococcus aureus*: nanoscale does matter. Nanobiotechnology **5**(1), 2–9 (2009)
32. Premanathan, M., Karthikeyan, K., Jeyasubramanian, K., Manivannan, G.: Selective toxicity of ZnO nanoparticles toward gram-positive bacteria and cancer cells by apoptosis through lipid peroxidation. Nanomed. Nanotechnol. Biol. Med. **7**(2), 184–192 (2011)
33. Li, M., Zhu, L.Z., Lin, D.H.: Toxicity of ZnO nanoparticles to *Escherichia coil*: mechanism and the influence of medium components. Environ. Sci. Technol. **45**(5), 1977–1983 (2011)
34. Huang, Z., Zheng, X., Yan, D., Yin, G., Liao, X., Kang, Y., Yao, Y., Huang, D., Hao, B.: Toxicological effect of ZnO nanoparticles based on bacteria. Langmuir **24**, 4140–4144 (2008)
35. Kloepfer, J.A., Mielke, R.E., Nadeau, J.L.: Uptake of CdSe and CdSe/ZnS quantum dots into bacteria via purine-dependent mechanisms. Appl. Environ. Microbiol. **71**(5), 2548–2557 (2005)
36. Brayner, R., Dahoumane, S.A., Yepremian, C., Djediat, C., Meyer, M., Coute, A., Fievet, F.: ZnO nanoparticles: synthesis, characterization, and ecotoxicological studies. Langmuir **26**(9), 6522–6528 (2010)
37. Dumas, E., Gao, C., Suffern, D., Bradforth, S.E., Dimitrijevic, N.M., Nadeau, J.L.: Interfacial charge transfer between CdTe quantum dots and gram negative vs gram positive bacteria. Environ. Sci. Technol. **44**(4), 1464–1470 (2010)
38. Schneider, R., Wolpert, C., Guilloteau, H., Balan, L., Lambert, J., Merlin, C.: The exposure of bacteria to CdTe-core quantum dots: the importance of surface chemistry on cytotoxicity. Nanotechnology **20**(22), 225101 (2009)
39. Schubert, D., Dargusch, R., Raitano, J., Chan, S.-W.: Cerium and yttrium oxide nanoparticles are neuroprotective. Biochem. Biophys. Res. Commun. **342**(1), 86–91 (2006)
40. Rodea-Palomares, I., Boltes, K., Fernandez-Pias, F., Leganas, F., Garcia-Calvo, E., Santiago, J., Rosal, R.: Physicochemical characterization and ecotoxicological assessment of CeO_2 nanoparticles using two aquatic microorganisms. Toxicol. Sci. **119**(1), 135–145 (2011)
41. Zeyons, O., Thill, A., Chauvat, F., Menguy, N., Cassier-Chauvat, C., Orear, C., Daraspe, J., Auffan, M., Rose, J., Spalla, O.: Direct and indirect CeO_2 nanoparticles toxicity for *E. coli* and *Synechocystis*. Nanotoxicology **3**(4), 284–295 (2009)
42. Pelletier, D.A., Suresh, A.K., Holton, G.A., McKeown, C.K., Wang, W., Gu, B.H., Mortensen, N.P., Allison, D.P., Joy, D.C., Allison, M.R., Brown, S.D., Phelps, T.J., Doktycz, M.J.: Effects of engineered cerium oxide nanoparticles on bacterial growth and viability. Appl. Environ. Microbiol. **76**(24), 7981–7989 (2010)
43. Navarro, E., Piccapietra, F., Wagner, B., Marconi, F., Kaegi, R., Odzak, N., Sigg, L., Behra, R.: Toxicity of silver nanoparticles to *Chlamydomonas reinhardtii*. Environ. Sci. Technol. **42**(23), 8959–8964 (2008)
44. Miao, A.-J., Schwehr, K.A., Xu, C., Zhang, S.-J., Luo, Z., Quigg, A., Santschi, P.H.: The algal toxicity of silver engineered nanoparticles and detoxification by exopolymeric substances. Environ. Pollut. **157**, 3034–3041 (2009)
45. Franklin, N.M., Rogers, N.J., Apte, S.C., Batley, G.E., Gadd, G.E., Casey, P.S.: Comparative toxicity of nanoparticulate ZnO, bulk ZnO, and $ZnCl_2$ to a freshwater microalga (*Pseudokirchneriella subcapitata*): the importance of particle solubility. Environ. Sci. Technol. **41**(24), 8484–8490 (2007)
46. Kvitek, L., Vanickova, M., Panacek, A., Soukupova, J., Dittrich, M., Valentova, E., Prucek, R., Bancirova, M., Milde, D., Zboril, R.: Initial study on the toxicity of silver nanoparticles (NPs) against *Paramecium caudatum*. J. Phys. Chem. **C113**, 4296–4300 (2009)
47. Blinova, I., Ivask, A., Heinlaan, M., Mortimer, M., Kahru, A.: Ecotoxicity of nanoparticles of CuO and ZnO in natural water. Environ. Pollut. **158**, 41–47 (2010)
48. Unrine, J.M., Tsyusko, O.V., Hunyadi, S.E., Judy, J.D., Bertsch, P.M.: Effects of particle size on chemical speciation and bioavailability of copper to earthworms exposed to copper nanoparticles. J. Environ. Qual. **39**(6), 1942–1953 (2010)
49. Shoults-Wilson, W.A., Reinsch, B.C., Tsyusko, O.V., Bertsch, P.M., Lowry, G.V., Unrine, J.M.: Role of particle size and soil type in toxicity of silver nanoparticles to earthworms. Soil Sci. Soc. Am. J. **75**(2), 365–377 (2011)
50. Lapied, E., Moudilou, E., Exbrayat, J.-M., Oughton, D.H., Joner, E.J.: Silver nanoparticle exposure causes apoptotic response in the earthworm *Lumbricus*

terrestris (Oligochaeta). Nanomedicine **5**(6), 975–984 (2010)
51. Ahamed, M., Posgai, R., Gorey, T.J., Nielsen, M., Hussain, S.M., Rowe, J.J.: Silver nanoparticles induced heat shock protein 70, oxidative stress and apoptosis in *Drosophila melanogaster*. Toxicol. Appl. Pharmacol. **242**, 263–269 (2010)
52. Meyer, J., Lord, C.A., Yang, X.Y., Turner, E.A., Badireddy, A.R., Marinakos, S.M., Chilkoti, A., Wiesner, M., Auffan, M.: Intracellular uptake and associated toxicity of silver nanoparticles in *Caenorhabditis elegans*. Aquat. Toxicol. **100**, 140–150 (2010)
53. Wang, H., Wick, R.L., Xing, B.: Toxicity of nanoparticulate and bulk ZnO, Al2O3 and TiO_2 to the nematode *Caenorhabditis elegans*. Environ. Pollut. **157**, 1171–1177 (2009)
54. Zhang, H., He, X., Zhang, Z., Zhang, P., Li, Y., Ma, Y., Kuang, Y., Zhao, Y., Chai, Z.: Nano-CeO_2 exhibits adverse effects at environmental relevant concentrations. Environ. Sci. Technol. **45**, 3725–3730 (2011)
55. Heinlaan, M., Ivask, A., Blinova, I., Dubourguier, H.C., Kahru, A.: Toxicity of nanosized and bulk ZnO, CuO and TiO_2 to bacteria *Vibrio fischeri* and crustaceans *Daphnia magna* and *Thamnocephalus platyurus*. Chemosphere **71**(7), 1308–1316 (2008)
56. Gao, J., Youn, S., Hovsepyan, A., Llaneza, V., Wang, Y., Bitton, G.: Dispersion and toxicity of selected manufactured nanomaterials in natural river water samples: effects of water chemical composition. Environ. Sci. Technol. **43**, 3322–3328 (2009)
57. Rodgers, J.H., Deaver, E., Rogers, P.L.: Partitioning and effects of silver in amended freshwater sediments. Ecotoxicol. Environ. Saf. **37**(1), 1–9 (1997)
58. Heinlaan, M., Kahru, A., Kasemets, K., Arbeille, B., Prensier, G., Dubourguier, H.-C.: Changes in the *Daphnia magna* midgut upon ingestion of copper oxide nanoparticles: a transmission electron microscopy study. Water Res. **45**, 179–190 (2011)
59. Zhao, C.-M., Wang, W.-X.: Biokinetic uptake and efflux of silver nanoparticles in *Daphnia magna*. Environ. Sci. Technol. **44**, 7699–7704 (2010)
60. Malvindi, M.A., Carbone, L., Quarta, A., Tino, A., Manna, L., Pellegrino, T.: Rod-shaped nanocrystals elicit neuronal activity in vivo. Small **4**(10), 1747 (2008)
61. Gagné, F., Auclair, J., Turcotte, P., Fournier, M., Gagnon, C., Sauvé, S., Blaise, C.: Ecotoxicity of CdTe quantum dots to freshwater mussels: impacts on immune system, oxidative stress and genotoxicity. Aquat. Toxicol. **86**(3), 333–340 (2008)
62. Sap-Iam, N., Homklinchan, C., Larpudomlert, R., Warisnoicharoen, W., Sereemaspun, A., Dubas, S.T.: UV irradiation-induced silver nanoparticles as mosquito larvicides. J. Appl. Sci. **10**(23), 3132–3136 (2010)
63. Griffitt, R.J., Weil, R., Hyndman, K.A., Denslow, N. D., Powers, K., Taylor, D., Barber, D.S.: Exposure to copper nanoparticles causes gill injury and acute lethality in zebrafish (*Danio rerio*). Environ. Sci. Technol. **41**(23), 8178–8186 (2007)
64. Snell, T.W., Hicks, D.G.: Assessing toxicity of nanoparticles using *Brachionus manjavacas* (Rotifera). Environ. Toxicol. **26**(2), 146–152 (2011)
65. Unrine, J.M., Hunyadi, S.E., Tsyusko, O.V., Rao, W., Shoults-Wilson, W.A., Bertsch, P.M.: Evidence for bioavailability of Au nanoparticles from soil and biodistribution within earthworms (*Eisenia fetida*). Environ. Sci. Technol. **44**(21), 8308–8313 (2010)
66. Goodman, C.M., McCusker, C.D., Yilmaz, T., Rotello, V.M.: Toxicity of gold nanoparticles functionalized with cationic and anionic side chains. Bioconjug. Chem. **15**, 897–900 (2004)
67. Wang, S., Lawson, R., Ray, P.C., Yu, H.: Toxic effects of gold nanoparticles on *Salmonella typhimurium* bacteria. Toxicol. Ind. Health **27**(6), 547–554 (2011)
68. Renault, S., Baudrimont, M., Mesmer-Dudons, N., Gonzalez, P., Mornet, S., Brisson, A.: Impacts of gold nanoparticle exposure on two freshwater species: a phytoplanktonic alga (*Scenedesmus subspicatus*) and a benthic bivalve (*Corbicula fluminea*). Gold Bull. **41**(2), 116–126 (2008)
69. Rocha, A., Zhou, Y., Kundu, S., Gonzalez, J.M., Vinson, S.B., Liang, H.: In vivo observation of gold nanoparticles in the central nervous system of *Blaberus discoidalis*. J. Nanobiotechnol. **9**, 1–9 (2011)
70. Pompa, P., Vecchio, G., Galeone, A., Brunetti, V., Sabella, S., Maiorano, G., Falqui, A., Bertoni, G., Cingolani, R.: In vivo toxicity assessment of gold nanoparticles in *Drosophila melanogaster*. Nano Res. **4**, 405–413 (2011)
71. Tedesco, S., Doyle, H., Blasco, J., Redmond, G., Sheehan, D.: Exposure of the blue mussel, *Mytilus edulis*, to gold nanoparticles and the pro-oxidant menadione. Comp. Biochem. Physiol. C Toxicol. Pharmacol. **151**(2), 167–174 (2010)
72. Battin, T.J., Fvd, K., Weilhartner, A., Ottofuelling, S., Hofmann, T.: Nanostructured TiO_2: transport behavior and effects on aquatic microbial communities under environmental conditions. Environ. Sci. Technol. **43**(21), 8098–8104 (2009)
73. Sawosz, E., Chwalibog, A., Szeliga, J., et al.: Visualization of gold and platinum nanoparticles interacting with *Salmonella enteritidis* and *Listeria monocytogenes*. Int. J. Nanomedicine **5**, 631–637 (2010)
74. Ge, Y., Schimel, J.P., Holden, P.A.: Evidence for negative effects of TiO_2 and ZnO nanoparticles on soil bacterial communities. Environ. Sci. Technol. **45**(4), 1659–1664 (2011)
75. Kim, S., Kim, J., Lee, I.: Effects of Zn and ZnO nanoparticles and Zn^{2+} on soil enzyme activity and bioaccumulation of Zn in *Cucumis sativus*. Chem. Ecol. **27**(1), 49–55 (2011)
76. Bradford, A., Handy, R.D., Readman, J.W., Atfield, A., Muhling, M.: Impact of silver nanoparticle contamination on the genetic diversity of natural bacterial

assemblages in estuarine sediments. Environ. Sci. Technol. **43**(12), 4530–4536 (2009)
77. Werlin, R., Priester, J.H., Mielke, R.E., Krämer, S., Jackson, S., Stoimenov, P.K., Stucky, G.D., Cherr, G. N., Orias, E., Holden, P.A.: Biomagnification of cadmium selenide quantum dots in a simple experimental microbial food chain. Nat. Nanotechnol. **6**, 65–71 (2011)
78. Judy, J.D., Unrine, J.M., Bertsch, P.M.: Evidence for biomagnification of gold nanoparticles within a terrestrial food chain. Environ. Sci. Technol. **45**(2), 776–781 (2010)
79. Holbrook, R.D., Murphy, K.E., Morrow, J.B., Cole, K. D.: Trophic transfer of nanoparticles in a simplified invertebrate food web. Nat. Nanotechnol. **3**, 352–355 (2008)
80. Zhu, X., Wang, J., Zhang, X., Chang, Y., Chen, Y.: Trophic transfer of TiO_2 nanoparticles from *Daphnia* to zebrafish in a simplified freshwater food chain. Chemosphere **79**, 928–933 (2010)
81. Ferry, J.L., Craig, P., Hexel, C., Sisco, P., Frey, R., Pennington, P.L., Fulton, M.H., Scott, I.G., Decho, A. W., Kashiwada, S., Murphy, C.J., Shaw, T.J.: Transfer of gold nanoparticles from the water column to the estuarine food web. Nat. Nanotechnol. **4**(7), 441–444 (2009)
82. Saez, G., Moreau, X., De Jong, L., Thiéry, A., Dolain, C., Bestel, I., Di Giorgio, C., De Meo, M., Bartélémy, P.: Development of new nano-tools: towards an integrative approach to address the societal question of nanotechnology? Nano Today **5**, 251–253 (2010)

Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment

Mana Man Na Yung[1], Catherine Mouneyrac[2] and Kenneth Mei Yee Leung[1]
[1]The Swire Institute of Marine Sciences and School of Biological Sciences, The University of Hong Kong, Pokfulam, Hong Kong, China
[2]LUNAM Université, Université Catholique de l'Ouest, Angers, France

Synonyms

Marine organisms; Nano-ecotoxicology; Toxic mechanisms; Zinc oxide nanoparticles (ZnO-NPs)

Definition

Ecotoxicity is generally defined as the subject in the field of ecotoxicology which studies the potential for biological, chemical, or physical stressors to affect ecosystems. The study of ecotoxicity of nanoparticles focuses on the toxic effects of nanoparticles on wildlife and their ecosystems and uncovers the natural uptake and toxic mechanisms of nanoparticles in organisms. It also investigates the influences of environmental factors on bioavailability and toxicity of the nanoparticles to the organisms. Nanoparticles are defined as particles having one or more dimensions between 1 and 100 nm that show distinct properties from their bulk counterparts of the same materials [1]. The use of nanoparticles in manufactured products has grown dramatically in the last decade. The continuously growing market for zinc oxide nanoparticles (ZnO-NPs), driven mainly by increasing demand for cosmetics and sunscreens, has grown to over 30,000 t annually [2]. ZnO-NPs are transparent ultraviolet (UV) absorbers providing a long-term and broad-band protection from UVA (320–400 nm) and UVB (290–320 nm) radiation. ZnO-NPs are also being used for their antimicrobial, anticorrosion, and catalytic properties in industries such as paints, coatings, textile, and electronics. Due to the widespread use of ZnO-NPs, coastal waters and the ocean would become an ultimate sink of these environmental contaminants. It is, therefore, essential to understand and investigate the fate and behavior of ZnO-NPs in the marine environment and examine their biological effects on saltwater organisms.

This review aims to discuss the behavior and transport of ZnO-NPs, describe the latest methods for characterizing their physicochemical properties, and highlight the environmental impacts of ZnO-NPs on various marine organisms and their potential toxic mechanisms. Influences of various environmental factors on the bioavailability and toxicity of ZnO-NPs to marine organisms are also examined. Based on these findings and knowledge gaps, we recommend some future research directions for better understanding the ecotoxicity of ZnO-NPs in marine ecosystems.

Fate and Behavior of ZnO-NPs in the Marine Environment

Since only very few studies had quantified the concentration of ZnO-NPs in the marine environment, the environmental concentrations were usually calculated based on mathematical modeling methods [3, 4]. Owing to the lack of analytical instruments able to characterize nanoparticles in complex matrices (e.g., surface water, wastewater treatment plant effluents, biosolids, sediments, and soil), the expected environmental concentrations have to be modeled based on the market penetrations of nano-products and the known usage of these products [5] or a life-cycle model to quantify the release of ZnO-NPs into the environment [4, 6, 7]. The modeled concentrations of ZnO-NPs in different environmental compartments are summarized in Table 1. The estimations vary greatly among different modeling methods, indicating a high uncertainty in the estimated concentration.

Nevertheless, the estimated concentrations of ZnO-NPs in the environment can be used as preliminary guidelines for developing analytical methods and forming the basis for risk assessment of ZnO-NPs [3].

The physicochemical characteristics of nanoparticle, together with the conditions of the surrounding environment (e.g., temperature, ionic strength, oxygen level, and natural organic matter), determine the behavior and transport of nanoparticles in the environmental system of concern [5]. The physicochemical properties of nanoparticles can be characterized by various microscopic and spectroscopic techniques (Table 2). The particle size and shape, aggregation/agglomeration, surface chemistry (e.g., functionalization), chemical composition, and dissolution are among the key parameters of physicochemical characteristics of ZnO-NPs. Several state-of-the-art analytical techniques, namely, transmission electron microscopy (TEM), atomic force microscopy (AFM), and dynamic light scattering (DLS), are widely applied to characterize nanoparticles. The advanced microscopic methods (e.g., TEM, AFM, etc.) can generally provide accurate dimensions of nanoparticles that are free from artifacts, but these techniques suffer from low sample numbers and high operating costs [8]. Since the scattering varies with particle radius, aggregates of the particles and the contaminants (e.g., dust) in the sample might mask the signal of the nanoparticles when using DLS techniques [8].

It is important to understand the persistence and distribution of nanoparticles in the environment in order to examine their exposure pathways and estimate their exposure concentrations. Once being released into the marine environment, ZnO-NPs are expected to have different behaviors. In general, the nanoparticles can stay in suspensions as individual particles, dissolve in the seawater, aggregate and form larger particles and subsequent deposit on sediment, adsorb onto the constituents in marine waters (e.g., dissolved organic matter, DOM), and transform chemically based on reduction-oxidation (redox) reactions or transform biologically in the presence of biota (e.g., microorganisms) in the marine environment [9] (Fig. 1).

Persistence

Elemental zinc is able to accumulate in water, sediment, and organisms in two different forms,

Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment, Table 1 Modeled environmental concentrations of ZnO-NPs in different compartments (Adapted from Ref. [3])

Surface water ($ng\ L^{-1}$)	Wastewater ($\mu g\ L^{-1}$)	Sludge ($mg\ kg^{-1}$)	Soil ($\mu g\ kg^{-1}$)	Sludge-treated soil ($\mu g\ kg^{-1}$)	Sediment ($\mu g\ kg^{-1}$)	Remarks	References
76,000		2,172		3,194		10 % market penetration	[5]
1–60	0.22–1.42	13.6–64.7	0.026–0.66	1.6–23.1	0.49–56		[6, 7]
0.1–100	1	10^1–10^2	1–10^2	10^1–10^2	10^2		[4]

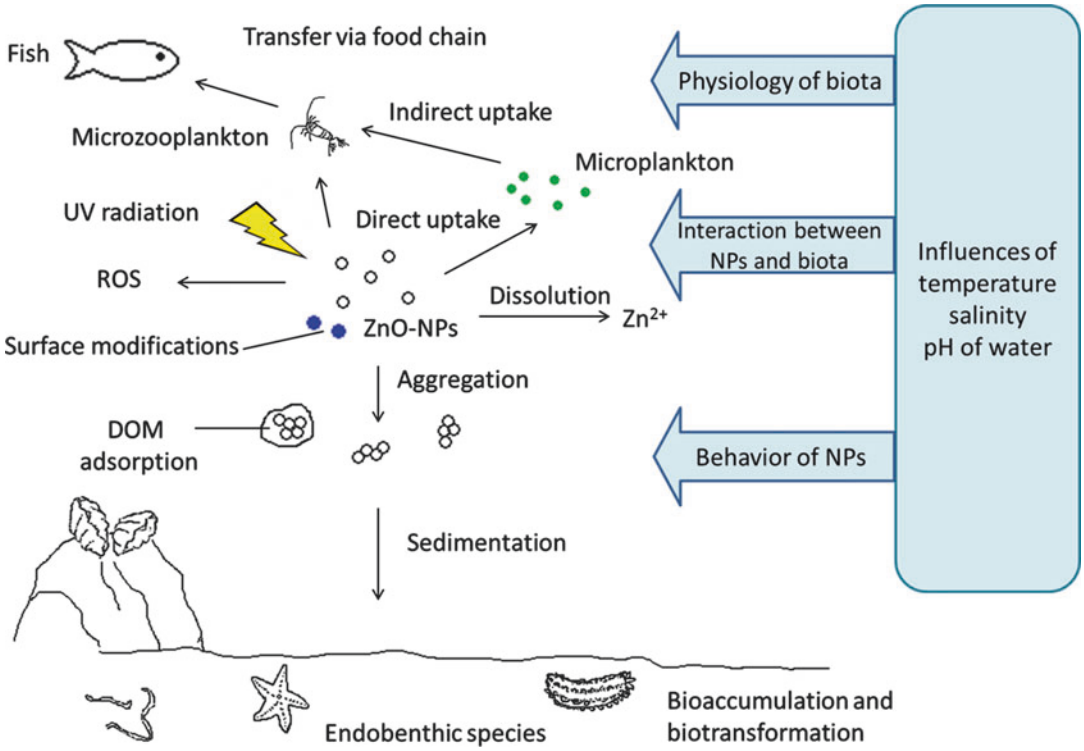
Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment, Table 2 Analytical techniques for the physicochemical characterization of nanoparticles (Modified from Ref. [9])

Parameters	Techniques	Detection range
Size and shape	Atomic force microscopy (AFM)	>0.1 nm
	Scanning electron microscopy (SEM)	1 nm
	Transmission electron microscopy (TEM)	>0.1 nm
Crystal structure	X-ray diffraction (XRD)	
Particle aggregate size	Atomic force microscopy (AFM)	>0.1 nm
	Dynamic light scattering (DLS)	0.2 nm – 10 μm
	Scanning electron microscopy (SEM)	>1 nm
	Transmission electron microscopy (TEM)	>0.1 nm
Surface area	Brunauer-Emmett-Teller (BET)	m^2/g -> 2,000 m^2/g
Surface charge	Zeta potential	
Chemical composition	Electron paramagnetic resonance (EPR)	mM range
	Fourier transform infrared spectroscopy (FT-IR)	ppm range
	Inductively coupled plasma mass spectroscopy (ICP-MS)	1 ppt–0.1 ppb
	Nuclear magnetic resonance (NMR)	mM range
	Raman spectroscopy	
	Scanning electron microscopy (SEM) and energy-dispersive X-ray spectroscopy (EDS)	1 nm
	Transmission electron microscopy (TEM) and energy-dispersive X-ray spectroscopy (EDS)	>0.1 nm
	Ultraviolet/visible spectroscopy (UV/Vis)	mM range
	X-ray diffraction (XRD)	
Dissolution	Dialysis	<5 nm
	Ultrafiltration	1 nm–1 μm
	Diffusive gradients in thin films (DGT)	Labile forms

namely, zinc ions and ZnO-NP. However, they have completely different properties in terms of their persistence in water. Zinc ion itself is intrinsically persistent, although it can be converted into other species (i.e., speciation) and can form zinc complexes, such as zinc-chloro complexes or zinc hydroxide with chemicals present in natural seawater. In contrast, ZnO-NPs are not necessarily persistent. ZnO-NPs can dissolve or aggregate, losing the properties of the single nanoparticle. Various researchers have been studying the factors affecting the dissolution or aggregation ability of ZnO-NPs in the marine environment. Fairbairn et al. [10] demonstrated that uncoated ZnO-NPs were soluble in seawater and the release of zinc ions from ZnO-NPs was dependent on time and concentration of ZnO-NPs. Specifically, at low initial concentration of 0.1 mg L^{-1} ZnO-NPs, the dissolution reached 100 %; whereas at high initial concentration of 10 mg L^{-1} ZnO-NPs, the dissolution decreased to 60 % after 6 h. Furthermore, Bian et al. [11] also showed that decreasing pH or in the presence of humic acid at pH 9–11 enhanced the release of dissolved zinc ions in water. Also, it has been shown that small ZnO-NPs (4 nm) dissolved more readily than larger ZnO-NPs (15 and 240 nm) at 1,000 mg L^{-1}, and such size effects on dissolution of ZnO-NPs can be explained by the Noyes-Whitney equations and Ostwald-Freundlich equations [11]. As some types of ZnO-NPs in commercial products have surface coatings (e.g., silica and polymer) to ensure their dispersion in aqueous media, these coated ZnO-NPs generally have a greater persistence in the marine environment than uncoated ZnO-NPs [12].

Particle Aggregation

Once the nanoparticles enter the water, they form homoaggregates according to the combined effect



Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment, Fig. 1 Schematic illustration of the behavior and transport of ZnO-NPs in the marine environment

of van der Waals and double layer force (DLVO) theory, which describes the forces between charged surfaces of particles interacting in a liquid medium. The combined effects of van der Waals attraction and the electrostatic repulsion due to an electric double layer of counterions that surround ZnO-NPs in an aqueous medium determine the stability of the ZnO-NP suspension. ZnO-NPs carry an electrical charge, which produces a force of mutual electrostatic repulsion between adjacent particles. If the charge is high enough to overcome the van der Waals attraction force, the ZnO-NPs will remain separate and are stabilized in suspension. However, the marine environment with a high ionic strength causes compression of the electric double layer, reduces the electrostatic repulsion forces between the nanoparticles, and consequently promotes the nanoparticles to aggregate and settle out of the suspension [11]. The degree of aggregation (homoaggregation and heteroaggregation) depends on the properties of the nanoparticles (e.g., particle morphology, size, surface area, and surface charge) and characteristics of the environmental system [8], as well as exposure protocols.

The translocation of nanoparticles depends largely on their size. Aggregates of nanoparticles usually have greater likelihood to deposit on sediments, and thus they are usually less mobile than individual nanoparticles. Nonetheless, there are other processes, such as dissociation or resuspension of nanoparticles due to changes in aquatic chemistry of the surrounding environment and the activity of aquatic organisms, that could complicate the deposition action of aggregated nanoparticles. The elimination of aggregates from the water column through deposition or sedimentation decreases the concentration of bioavailable ZnO-NPs for pelagic species. Although these ZnO-NP aggregates become less mobile, uptake by sediment-dwelling organisms or filter feeders is still possible.

Adsorption

Because of their high surface-to-volume ratio and their high surface reactivity, ZnO-NPs can adsorb onto pollutants or DOMs in the marine environment. DOMs such as humic acids and fulvic acids derived from algae or higher plants are ubiquitous in the environment. These plant materials are typically highly functionalized with carboxylic acid groups and/or phenolic groups; they are negatively charged at neutral pH and may also contain chromophores that absorb sunlight [13]. The absorption of light by chromophoric groups can lead to formation of short-lived reactive oxygen species (ROS), for example, singlet oxygen, $^1O_2$, which may increase the toxicity of ZnO-NPs or chemically transform them. The adsorption of DOMs onto ZnO-NPs can alter their surface charge and inhibit aggregation/agglomeration, thereby enhancing their mobility in the environment. For example, adsorption of humic acids onto ZnO-NPs has been shown to decrease the particle zeta potential, suggesting humic acid-coated ZnO-NPs could be more easily dispersed and suspended than uncoated ones because of their enhanced electrostatic repulsion [14]. Some studies have also shown that ZnO-NPs with humic acid coatings enhanced sorption of organic pollutants such as polycyclic aromatic hydrocarbons on the nanoparticles' surface [15].

Transformation

Transformation of ZnO-NPs in the environment will significantly influence their size, shape, and surface chemistry, and this process will affect their ultimate distribution, persistence, and toxicity in the environment. Redox reactions occurring on the surface of ZnO-NPs can lead to the production of ROS, which are in turn able to oxidize organic compounds in the environment. For instance, ROS released from nanoparticles can induce oxidative stress in vivo and in vitro on cell cultures of bacteria [16].

The effects of ZnO-NPs on microbial populations in the natural environment were well studied. Conversely, further examinations are needed to fully understand the contribution of bacteria and fungi in water or sediment to the degradation of engineered ZnO-NPs. Uptake of ZnO-NPs might also happen in higher organisms, and the biological fate of ZnO-NPs depends on the balance of four processes: absorption, distribution, metabolism, and excretion. Up to now, the process of biotransformation of ZnO-NPs by organisms in the natural environment and how such transformation affects the toxicity of ZnO-NPs are still unclear.

In summary, the environmental fate of ZnO-NPs depends on physical and chemical properties of ZnO-NPs and can be further influenced by environmental conditions. Modification will greatly influence the behavior and transport of ZnO-NPs and their interaction with organisms in the marine environment.

Toxic Effects of Zinc Oxide Nanoparticles on Marine Organisms

Over the past decade, there have been a growing number of studies to reveal the relationship between physicochemical characteristics and toxicity of ZnO-NPs to saltwater organisms in seawater. These studies not only examined the effects of ZnO-NPs to marine organisms at various levels of biological organization (e.g., molecular, cellular, developmental, and behavioral responses) but also investigated the potential uptake route, fate, and transport of ZnO-NPs in the marine ecosystem. This section provides an up-to-date summary and critical analysis on the documented information related to the exposure effects of ZnO-NPs on marine organisms from different taxonomic groups, including microalgae, molluscs, crustaceans, and fish (Table 3). Microalgae and crustaceans, which are the common test taxa, are more sensitive than molluscs. Fishes are comparatively less sensitive to ZnO-NP exposure. Growth and mortality were the general toxicity endpoint, while the common test duration was 96 h (Table 3).

Microalgae

Microalgae, primary producers in the marine environment, are particularly sensitive to a wide

Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment, Table 3 Toxicity of zinc oxide nanoparticles (ZnO-NPs) to marine organisms: an overview of available literature data

Primary particle size	Test species	Duration	Endpoint(s)	Effect concentration	References
	Algae				
6.3–15.7 nm (sphere) 242–862 nm (rod)	*Thalassiosira pseudonana*	100 h	Growth, uptake	EC100 = 10 mg L^{-1}	[18]
	Chaetoceros gracilis			EC100 = 10 mg L^{-1}	
	Phaeodactylum tricornutum			EC100 = 10 mg L^{-1}	
20 nm	*Thalassiosira pseudonana*	48 h	Growth rate, Fv/Fm, Chl *a*	IC50 = 0.39 μM (growth); 3.83 μM (Fv/Fm); 6.94 μM (Chl *a*)	[19]
20 nm	*Skeletonema costatum*	96 h	Growth inhibition	IC50 = 2.36 mg L^{-1}	[20]
	Thalassiosira pseudonana	96 h	Growth inhibition	IC50 = 4.56 mg L^{-1}	
100 nm	*Dunaliella tertiolecta*	96 h	Growth inhibition	EC50 = 2.42 mg L^{-1}	[21]
20–30 nm	*Thalassiosira pseudonana*	96 h	Growth rate	LOEC = 0.5 mg L^{-1}	[22]
	Skeletonema marinoi			LOEC = 1 mg L^{-1}	
	Dunaliella tertiolecta			LOEC = 1 mg L^{-1}	
	Isochrysis galbana			LOEC = 1 mg L^{-1}	
	Crustaceans				
20 nm	*Tigriopus japonicus*	96 h	Mortality	LC50 = 0.85 mg L^{-1}	[20]
	Elasmopus rapax			LC50 = 1.19 mg L^{-1}	
10–30 nm, 200 nm	*Artemia salina*	24, 96 h	Uptake, mortality	LC50 > 100 mg L^{-1}	[28]
Not reported	*Acartia tonsa*	7 days	Reproduction, survival (dietary exposure)	EC20 = 143 μg L^{-1} (reproduction); 112 μg L^{-1} (survival)	[29]
35 ± 10 nm	*Corophium volutator*	100 days	Growth	LOEC = 0.5 mg L^{-1}	[30]
10–30 nm	*Corophium volutator*	10 days	Uptake	1000 ng g^{-1} Zn (sediment)	[31]
	Mollusca				
31.7 ± 13.0 nm	*Crassostrea gigas*	6, 24, 48, 96 h	Uptake, mortality	LC50 (96 h) = 37.2 mg L^{-1}	[24]
21–34 nm	*Scrobicularia plana*	16 days	Uptake, biochemical and behavioral change	Increase in CAT; disturbed burrowing and feeding rates at 3 mg kg^{-1} (sediment)	[25]
24 ± 3 nm	*Mytilus galloprovincialis*	4 days	Uptake	LOEC = 2.5 mg L^{-1}	[26]

(*continued*)

E

Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment, Table 3 (continued)

Primary particle size	Test species	Duration	Endpoint(s)	Effect concentration	References
20–30 nm	*Mytilus galloprovincialis*	12 weeks	Respiration, uptake, growth, survival	LOEC = 2 mg L^{-1}	[27]
	Worm				
21–34 nm	*Hediste diversicolor*	16 days	Uptake, biochemical and behavioral change	Increase in GST (glutathione-S-transferase) activity at 3 mg kg^{-1} (sediment)	[25]
	Other invertebrates				
18.8 nm	*Lytechinus pictus*	96 h	Larval morphology	EC50 = 99.5 μg L^{-1}	[9]
	Fish				
20 nm	*Oryzias melastigma*	96 h	Oxidative stress responses	Upregulation of HSP70 at 4 and 40 mg L^{-1}	[20]

range of pollutants [17]. They are, therefore, often used as model organisms in standard toxicity tests for environmental risk assessment of chemical substances. The majority of studies on microalgae focused on the development of concentration-response relationships using algal growth inhibition or photosynthesis inhibition as the toxicity endpoint [17].

In general, toxic effects of ZnO-NPs varied significantly among different algal species (i.e., species dependent). A less silicified marine diatom *Phaeodactylum tricornutum* was shown to be less sensitive to ZnO-NPs than a more silicified diatom *Thalassiosira pseudonana* [18]. A significant decrease in cell-division rate was observed in *T. pseudonana*, which was suggested to be caused by the disruption of silicon uptake and frustule formation in the presence of zinc ions dissolved from the ZnO-NPs [18].

The toxic mechanisms of ZnO-NPs have been suggested to be dependent on physicochemical properties of the particles, such as size, shape [18], amount of dissolved free zinc ions [18, 19], and production of ROS [20, 21].

The median inhibition concentrations (IC50; concentrations caused 50 % growth inhibition when compared with the control group) of ZnO-NPs to *T. pseudonana* and *Skeletonema costatum* to ZnO-NPs were found to be 4.56 and 2.36 mg L^{-1}, respectively [20]. An exposure to bulk ZnO resulted in significantly higher IC50s at 6.65 and 2.97 mg L^{-1} for *T. pseudonana* and *S. costatum*, respectively, indicating that ZnO-NPs were more toxic to the test diatoms [20]. Studies suggested that dissolved zinc ions released from ZnO-NPs, rather than the nanoparticles themselves, have been demonstrated as the sole cause of toxicity in studies conducted on *T. pseudonana* [19]. Comparable IC50 values were obtained upon exposure to zinc ions or to ultrafiltrate of the ZnO-NPs smaller than <0.22 μm [19]. On the contrary, Peng and colleagues [18] argued that not all the observable toxicity in an exposure of *P. tricornutum* to ZnO-NPs of varying sizes and shapes could be attributed to bioavailable zinc ions, but morphology and aggregates of the nanoparticles played a part as well. They found that one-dimensional rod-shaped ZnO-NPs were more toxic than the spherical ones, whereas the size of nanoparticles was less influential to their toxicities to the microalgae [18]. The presence of ZnO-NP aggregates on the cell surface may also have inhibited algal growth through the release of zinc ions or mechanical damage by the nanoparticles. Based on the above studies, both ZnO-NP aggregates and soluble zinc ions can be attributable to their toxic effects to marine microalgae.

Miller et al. [22] explained that toxic effects of free zinc ions on marine algae *T. pseudonana*, *S. marinoi*, *Dunaliella tertiolecta*, and *Isochrysis galbana* were coupled with the nanoparticle-associated effects. For instance, adsorption of particles onto cell surfaces may facilitate damage because of both the production of ROS and dissolution of free zinc ions and, thus, disrupt cellular metabolic activities. Unless these activities occurring on the cell surface and the concentrations of the free zinc ions can be quantified throughout the experiments, the release of free zinc ions as the sole cause of ZnO-NP toxicities cannot be unequivocally judged.

Several studies [20–22] suggested that the generation of ROS can be a toxic pathway of ZnO-NPs to the marine organisms; however, none of those studies measured the ROS released by ZnO-NPs or quantified the intracellular ROS produced by the diatoms. The photocatalytic activity of ZnO-NPs in seawater is still unknown as is the associated phototoxicity to the marine algae. Further studies are required to confirm the ROS production and photocatalysis as toxic mechanisms of ZnO-NPs in the marine environment.

Molluscs

ZnO-NPs form aggregates rapidly in seawater and deposit on the sediment surface over the time [23]; therefore, benthic invertebrates, especially filter feeders and suspension feeders, have a higher risk of exposure to ZnO-NPs. Several studies had been conducted to study the bioaccumulation of ZnO-NPs in bivalve molluscs including clams, oysters, and mussels.

Exposure of ZnO-NPs (about 30 nm) to the oyster *Crassostrea gigas* caused an accumulation of zinc in gills and later in their digestive gland, which resulted in mitochondrial damage and oxidative stress [24]. The zinc burden could occur by both endocytosis of ZnO-NPs and uptake of soluble zinc. However, the exposure concentration used for the experiment (4 mg L^{-1}) [24] was much higher than the predicted environment concentration of ZnO-NPs in natural surface water (1–10 ng L^{-1}) and in effluents from sewage treatment plants (300–432 ng L^{-1}) [6]. At an environmental realistic concentration of 3 mg kg^{-1} ZnO-NPs in sediments, the marine intrasedimentary clam *Scrobicularia plana* was exposed to isotopically labeled ZnO-NPs (21–34 nm) for 16 days and consequentially showed behavioral disturbance and an increase in catalase activity in their tissues [25]. However, the authors could not distinguish whether the zinc uptake in the tissues of *S. plana* was resulted from nanoparticles or the soluble zinc or a combination of both.

A coastal marine mussel *Mytilus galloprovincialis* was found to have accumulated zinc when it was exposed to 1–10 mg L^{-1} ZnO-NPs for 4 days. Instead of excreting the ZnO-NPs in form of nanoparticles, they actually excreted them as pseudofeces [26]. Once settled on the seafloor, zinc-containing pseudofeces may be consumed by other marine benthic suspension feeders (e.g., clams), deposit feeders (e.g., polychaete worms, amphipods), and grazers (e.g., sea urchins); in this way, ZnO-NPs could be biotransformed in the marine ecosystem and significantly affect their fate and transport [26].

A chronic exposure of *M. galloprovincialis* to 2 mg L^{-1} ZnO-NPs (20–30 nm) for 12 weeks resulted in zinc accumulation in their somatic tissues and gonadal tissue, an increase in respiration rate, a reduced shell growth rate, and an increased mortality [27]. Moreover, the same study also demonstrated that ZnO-NPs exerted greater effect on mortality of small mussels. The removal of small individuals which serve as prey for various predators in the marine ecosystem could cause a cascade impact on the food web and energy flow and might lead to a change in the marine benthic community composition and structure in the long run [27].

Crustaceans

Small crustaceans, such as copepod, amphipods, and brine shrimp, are filter feeders and thus are especially susceptible to many pollutants compared with other pelagic marine animals

[17]. They are sensitive to metal stresses and are, in turn, used routinely as biomonitors for monitoring pollution in aquatic systems and as test species in standard toxicity tests. Exposure of these zooplanktons to ZnO-NPs has been associated with a number of harmful effects which are often linked to the physicochemical nature of the nanoparticles, such as size, dissolution, and aggregation of the nanoparticles. For example, ZnO-NPs of 10–30 nm were shown to be significantly more toxic to the brine shrimp *Artemia salina*, causing 26 % mortality at a concentration of 100 mg L^{-1}, when compared with ZnO-NPs of 200 nm which caused 18 % mortality at the same test concentration and duration (96 h) [28]. In other words, ZnO-NPs having smaller primary particle sizes can cause a higher toxicity than those with a large particle size, even though they will aggregate. The authors suggested that main causes of the toxic effects were due to the release of free zinc ions and the accumulation of ZnO-NPs in the gut of the brine shrimps [28]. They observed that the *Artemia* were unable to excrete the ingested particles that formed massive aggregates inside the guts after accumulation and such gut congestions could significantly affect the food uptake and assimilation leading to weight loss and illness. Wong et al. [20] also found that the mortality of marine copepod *Tigriopus japonicus* might be caused by mechanical injury due to the attachment of the ZnO-NP aggregates to their exoskeleton. In addition, these authors also found that the toxic effect was also dependent on the exposure concentration of ZnO-NPs and correlated well with the concentration of dissolved zinc ions in the exposure medium [20].

Limited work has been carried out on the effects of the dietary transfer of nanoparticles along the food chain. However, an exposure of the copepod *Acartia tonsa* to ZnO-NP-dosed diatom *T. weissflogii* at the highest concentration of 263 μg L^{-1} Zn showed a significant reduction in survival and reproduction by 52 % and 50 %, respectively, after 7 days of exposure [29]. These authors confirmed the transfer of ZnO-NPs to *A. tonsa* from the dosed algae, illustrating their potential for trophic transfer [29].

Toxic effects of ZnO-NPs were also tested with sediment-dwelling zooplanktons. Growth and reproduction of the amphipod *Corophium volutator* decreased after exposed to ZnO-NPs (35 nm) via both the water and sediment exposure routes [30, 31]. Zinc uptake by the amphipod was proved to be via the aqueous phase and ingestion of the zinc-adsorbed sediments.

In summary, effect assessments for ZnO-NPs on marine crustaceans are species dependent. Copepods and amphipods seem to be more sensitive to ZnO-NPs than brine shrimps. Toxic effects of ZnO-NPs to these small crustacean species are associated with the primary particle size, the concentration of free zinc ions, and the aggregates of the ZnO-NPs. More importantly, some studies reported that toxic effects of ZnO-NPs may also be occurring at environmentally realistic concentrations [29, 31].

Fish

Potential uptake routes of ZnO-NPs in fish include ingestion from water or sediment for demersal species, via the gill or the gut epithelial cells or through the chorion pore channel by diffusion for embryos [32]. Studies of toxicities of ZnO-NPs on marine fishes are scarce. The marine medaka fish *Oryzias melastigma* larvae, after exposure to ZnO-NPs of 4 and 40 mg L^{-1} ZnO-NPs for 96 h, showed an upregulation of a heat shock protein (HSP70) only. The *O. melastigma* larvae exposed to $ZnSO_4$ at the same zinc ion concentrations exhibited a significant increase in the protein expression of both HSP70 and superoxide dismutase (SOD). The results implied that the test medaka larvae exposed to ZnO-NPs were under non-oxidative stress, and the toxic effect of ZnO-NPs was mainly due to ionic zinc released from the nanoparticles [20].

In contrast to marine fishes, toxicity of ZnO-NPs to freshwater fish has been comparatively well documented, in particular for the zebra fish *Danio rerio*. Different adverse effects, including reduction in hatching rate of embryos, oxidative stress via intracellular ROS production, malformation of larvae [33], and inhibition of

SOD and catalase (CAT) activities in liver and gill tissues of zebra fish adults [34], have been observed in the literature. Across the studies, toxicity of ZnO-NPs to freshwater fishes was mainly sublethal and caused by oxidative stress. However, more relevant information on marine fish species is required before a robust conclusion can be drawn.

In summary, the recent exposure studies of marine organisms have demonstrated that ZnO-NPs can induce a wide range of toxic effects (including both acute lethal or chronic sublethal effects) in marine algae and invertebrates, but limited toxicity data were found in either embryonic or adult marine fish, possibly because fishes are generally more tolerant to ZnO-NP exposure [35]. A number of possible causes of ZnO-NP toxicities in aquatic organisms have been proposed by previous studies, including the ionic zinc dissolved from ZnO-NP into the exposure media, the bioaccumulation of ZnO-NPs, the interaction between the nanoparticles' aggregates and the target cells of the organisms, and oxidative stress associated to the production of extracellular and intracellular ROS.

Toxic Mechanisms of Zinc Oxide Nanoparticles

The mode of toxic action of ZnO-NPs involves at least three major mechanisms [36]. First is the dissociation of ZnO-NPs into toxic and bioavailable zinc labile (free Zn ions or weakly complexed forms with inorganic and organic ligands) forms in the exposure media. Second is the direct interaction between the nanoparticles and the cell or organisms that may cause mechanical damage, for instance, adhesion of the nanoparticles or their aggregates onto the cell may cause mechanical abrasion and damage, as well as onsite release of ROS on the cell surface, resulting in membrane lesions. Accumulation of the nanoparticles in the cell or body tissues may also trigger toxic effects when the concentration of nanoparticles reaches an effect threshold or when the amount of accumulated nanoparticles is greater enough to cause interruption or impairment of cellular and physiological processes. Third, ZnO-NPs may generate more ROS under ultraviolet (UV) radiation and disrupt the biological activities of the organisms in the vicinity.

Dissolved Ionic Zinc

Toxic effects of dissolved ionic zinc released from ZnO-NPs have been extensively proven in toxicity studies across a wide range of test species and summarized in a number of recent review articles [35, 36]. Most of the studies reported that ionic zinc dissociated from ZnO-NPs contributes solely or at least partially to the observed toxicity of ZnO-NPs [36]. Demonstration of ionic zinc as a cause of ZnO-NP toxicity is usually based on the findings that ZnO-NPs show comparable toxicity to the zinc salts ($ZnCl_2$ or $ZnSO_4$) or the filtrate portion of ZnO-NPs, together with a direct measurement of the concentration of dissolved Zn ions in the experimental medium.

Similar to ZnO-NPs, toxic effects of zinc ions also varied greatly among different marine species as zinc is an essential element to most marine organisms and the requirement of zinc varies among species [37]. Soluble Zn ions can inhibit growth and reduce photosynthesis ability in the marine diatom *T. pseudonana* after 48 h exposure [20], cause abnormal morphology in embryos of the sea urchin *Lytechinus pictus* embryos after 96 h of exposure at low μg L^{-1} range [10] and hinder growth, delay sexual maturation, and reduce reproduction in the amphipod *C. volutator* at concentrations ranging between 0.2 and 1 mg L^{-1} in a chronic exposure for 100 days [30].

Despite the wide recognition of the toxicity of the dissolved zinc ions, it is still uncertain about the extent of contribution of ionic zinc to ZnO-NP toxicity and its underlying mechanisms. As a typical aquatic exposure system contains both nanoparticles and zinc ions released from particle dissolution, the observed toxicity is usually a combination of both of the components [22, 36].

A recent study showed that ZnO-NP toxicity from nanoparticle-dependent effects and from dissolved zinc ions has distinct modes of toxic action as revealed by differential gene expression profiles in the freshwater water flea *Daphnia*

magna after exposure to ZnO-NPs or $ZnSO_4$ [38]. Therefore, expression of specific suits of biomarker genes can be useful to identify environmental exposure of ZnO-NPs and differentiate the exposure and toxic effects between the nanoparticle and ionic component [38]. Similar molecular or omics' approaches can be readily applied to reveal toxic mechanisms of ZnO-NPs and the dissolved zinc ions in marine organisms.

Direct Cellular Interaction and Bioaccumulation in Organisms

Particle dissolution and the release of zinc ions cannot always fully explain the toxicity of ZnO-NPs [22], while the particle-induced effect is suggested to be another important toxic mechanism of ZnO-NPs [18, 21, 30]. When the marine diatoms *P. tricornutum* were exposed to 40 mg L^{-1} of ZnO-NPs, there were obvious attachments of ZnO-NP aggregates onto their walls (as shown by TEM images), suggesting that growth inhibition was likely associated with the physicochemical interaction between the aggregates and the algal cell [18]. Similarly, exposure of nauplii of the marine copepod *T. japonicus* to ZnO-NPs showed physical injury when the ZnO-NPs attached onto the exoskeleton of the nauplii and hence hampered their movement [20]. Larvae of the brine shrimp *Artemia salina* were unable to excrete the ingested nanoparticles under waterborne exposure to ZnO-NPs [28]. After the marine oysters *C. gigas* were exposed to ZnO-NPs, they accumulated the nanoparticles inside vesicles in their gills and digestive gland tissues [24]. The adsorbed or the ingested ZnO-NPs could continuously release zinc ions and/or ROS, causing oxidative stresses, or biotransformation of the nanoparticles inside the organisms [24]. Although there is no systematic study on such a mode of toxic action up to now, these particle-induced mechanisms cannot be ignored.

The Generation of ROS

In addition, toxicity of ZnO-NPs to marine organisms may also be attributed to their photocatalytic activity under UV radiation, resulting in the generation of ROS in aquatic media [36]. Photochemistry behavior of ZnO-NPs is governed by the band gap which is the energy difference between the valence band (highest occupied molecular orbitals, HOMO) and the conduction band (lowest energy unoccupied molecular orbitals, LUMO) [39]. Light absorbed at wavelength less than or equal to the bandgap energy of ZnO will result in the excitation of an electron from the valence band to the conduction band; consequently, a hole or vacancy in the HOMO state of the molecular orbitals is created. Both the excited electrons and holes can migrate to the surface of ZnO and subsequently react in aqueous solution to form ROS such as superoxide (O_2^-) and hydroxyl radical (·OH). The superoxide (O_2^-) in aqueous solution can form oxygen and hydrogen peroxide by self-reaction (i.e., dismutation)
($O_2^- + O_2^- + 2H^+ \rightarrow H_2O_2 + O_2$). This process often occurs in living cells in the presence of superoxide dismutase (SOD) [39]. ZnO-NP, with band gap energy 3.37 eV (equivalent to 368 nm), can generate ROS under UV irradiation with energy at wavelength 368 nm or below, which comprises approximately 6 % of the sunlight energy reaching the earth's surface [36]. This toxicity pathway for ZnO-NPs is seldom studied because the majority of toxicity studies were conducted under laboratory fluorescence lighting which contains negligible or no UV radiation.

Photoinduced toxicity was suggested to be caused by natural sunlight but not laboratory lightings. For example, the 2-h LC50 of the nematode *Caenorhabditis elegans* was found to be 25 mg L^{-1} under the exposure of ZnO-NPs (60–100 nm) and natural sunlight, whereas same concentrations of ZnO-NPs induced no adverse effects under laboratory lighting or dark conditions [40]. In contrast, no significant growth inhibition was induced in the green algae *Pseudokirchneriella subcapitata* when exposed to UV pre-irradiated ZnO-NPs (<100 nm) under visible, UVA and UVB irradiation conditions at a concentration range between 0.05 and 0.3 mg ZnO-NP L^{-1} [41]. Most documented studies on photoinduced toxicity of ZnO-NPs were conducted with bacteria and freshwater species [34, 40]; the photocatalytic activity of ZnO-NPs

is less known in marine species. Thus, future research studies are needed to uncover the photocatalytic activity of ZnO-NPs in the seawater environment and its effects on ZnO-NP toxicity to marine organisms.

Influences of Environmental Factors on Toxicity of ZnO-NPs in the Marine Environment

ZnO-NPs under different environmental conditions can exhibit different toxicities to aquatic organisms [36]. Changes in temperature, salinity, and pH of seawater may change the physicochemical properties of ZnO-NPs, the physiology of the marine organisms, and the interactions between the nanoparticles and the organisms. Hence, the exposure pathway and the toxic effects of ZnO-NPs on organisms would be modified by the environmental condition accordingly. Since temperature, salinity, and pH are the key environmental variables and they vary significantly in marine ecosystems, this section specifically explores their effects on the toxicity of ZnO-NPs to marine organisms.

Temperature

Most aquatic organisms are ectotherms (animals whose regulation of body temperature depends on an external source). As a consequence, temperature is an important environmental factor governing their physiological processes. It is a well-known fact that temperature can influence all kinds of physiological processes. At the same time, temperature can alter properties of a chemical and hence its bioavailability and toxicity to marine organisms. In other words, temperature can influence the metabolic rate, movement, and feeding activity of an organism, causing changes in the toxicokinetics of contaminants (i.e., the rate of chemicals enter the body of an organism and the behaviour of the chemicals in the body), and thus the uptake, elimination and detoxification rates of the contaminants can be affected. In general, organisms living near to their thermal tolerance limits (in particular upper ones) exposed to a chemical toxicant would exhibit a greater combined toxicity than did organisms living under optimal conditions [42]. Up to now, there is only one comprehensive study to investigate the interacting effect of temperature and ZnO-NPs to marine organisms [43]. The results of this study indicate that zinc ion dissolution from ZnO-NPs increases with decreasing temperature and the interacting effect of temperature and ZnO-NPs varies among different species. First, Wong and Leung [43] found an increase in growth inhibition of the marine diatom *S. costatum* at both ends of its thermal tolerance limits at 15 °C and 28 °C based on IC10 values (IC10 = inhibitory concentration of 10 %) when exposed to ZnO-NPs for 96 h. Second, the amphipod *Melita longidactyla* mortality was positively correlated with temperature from 10 °C to 30 °C because they entered dormancy at low temperatures [43]. Third, the growth rate of larvae of the marine medaka fish *O. melastigma* was only reduced by temperature but not by the ZnO-NP exposure at 10 mg L^{-1}, probably because zinc is an essential trace element for fish [43]. Nonetheless, combined effects of temperature and ZnO-NPs on other marine organisms such as filter and deposit feeders like bivalve molluscs and polychaetes are still virtually unknown.

Salinity

Salinity fluctuations are common in estuaries and marine pollution often occurs mainly in coastal areas, where organisms would experience the interacting effect of salinity fluctuation and exposure to chemical pollutants. This interaction is complicated because salinity affects both the physiological processes of the organisms and the physicochemical characteristics of nanoparticles. An increase in salinity generally increases the degree of metal complexation due to the presence of chloride, sulfate, and organic matters in natural seawater. Salinity increase can reduce the ion dissolution from metal-oxide-based nanoparticles and thus lower the toxicity of the nanoparticles associated with the release of free metal ions [44].

However, studies on salinity effect on toxicity of ZnO-NPs are relatively scarce. A study on the salinity-dependent toxicity of ZnO-NPs showed that increasing in salinity from 5 to 35 PSU (practical salinity unit) decreased the dissolved

zinc ion concentration due to precipitation of the metal ions and consequently reduced the toxicity to the marine copepod *T. japonicus* [45].

pH

Changes in ion dissolution and aggregation size of ZnO-NPs can occur following from a change in pH. Miao et al. [19] compared the toxicity of ZnO-NPs on the marine diatom *T. pseudonana* at different pH levels (i.e., pH 7–9) and attributed the elevated toxicity of ZnO-NPs at low pH to an increase in the release of zinc ions from ZnO-NPs in the exposure medium. Similar findings were also reported by Bian et al. [11] that dissolution of ZnO-NPs increased with decreasing pH from pH 11 to pH 1. Moreover, maximum aggregate sizes of ZnO-NPs were observed at pH 9, where the pH level was close to the zero point of charge for ZnO (i.e., the repulsion interactions between nanoparticles decreased to a minimal point), and therefore there was a rapid sedimentation of ZnO-NPs.

Conclusions and Future Research Directions

The toxicity of ZnO-NPs can occur in certain species (e.g., the clam *S. plana*) under the exposure at environmentally realistic concentrations, suggesting that some marine organisms can be at risk in the natural environment with elevated levels of these nanoparticles. The major possible toxic mechanisms of ZnO-NPs, a the environmental factors affecting the behavior of ZnO-NPs and their toxicity to the organisms in the marine environment have been summarized and highlighted above. Based on our comprehensive review of relevant literature, we have identified several knowledge gaps which are needed to be filled for improving our current understanding on the ecotoxicity of ZnO-NPs for environmental risk assessment and management for this group of highly popular, commercialized nanoparticles.

1. Molecular toxic mechanisms to marine organisms should be thoroughly examined using advanced transcriptomic, proteomic, and metabolomic approaches in order to differentiate different modes of toxic action between the nanoparticles of ZnO-NPs and their associated dissolved zinc ions and between waterborne and dietary exposure.
2. Given that ZnO-NPs are highly photoactive, research focus should be laid on understanding the photoinduced toxicity under environmentally relevant UV radiation to an array of different marine species from different taxonomic groups. Tools and techniques should be developed to quantify the ROS released from ZnO-NPs and the intracellularly induced ROS in the environmental compartments and within the test marine organisms.
3. There are insufficient data on chronic effects of ZnO-NPs to marine organisms, and thus more chronic toxicity studies such as life-cycle studies are required to improve our understanding on the long-term and low-dose effect of these nanoparticles on selected marine species. Benthic organisms such as bivalves and polychaetes should be placed on a high priority as a subject of study, since ZnO-NPs have a high potential to aggregate and settle on the sediment in marine environment and bottom filter feeders and deposit feeders have a great potential to consume the nanoparticles associated with organic matter.
4. More toxicity data should be generated from marine fish species so as to reveal the sublethal toxic effects and associated toxic mechanisms of ZnO-NPs in this important group of higher-level organisms in the marine ecosystem.
5. Further ecotoxicological study of ZnO-NPs should also concentrate on bioaccumulation, cellular localization or tissue distribution, biotransformation, and trophic transfer of the nanoparticles in selected, typical marine food chains.
6. Combined effects of environmental stressors such as temperature, salinity, pH, UV radiation, and presence of DOM and combined toxic effects of ZnO-NPs and other pollutants to marine organisms should be investigated to better understand the behavior and toxicity of ZnO-NPs under environmentally realistic scenarios.

By filling up these knowledge gaps, our understanding on the ecotoxicology of ZnO-NPs in the marine environment could be significantly improved, while the information could also be used to derive environmental quality benchmarks such as water and sediment quality guidelines for regulating the use and release of ZnO-NPs, and hence offering better protection to marine ecosystems [46].

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity

References

1. Auffan, M., Rose, J., Bottero, J.Y., Lowry, G.V., Jolivet, J.P., Wiesner, M.R.: Towards a definition of inorganic nanoparticles from an environmental, health and safety perspective. Nat. Nanotechnol. **4**, 634–641 (2009)
2. Future Markets: The Global Market for Zinc Oxide Nanoparticles. Future Markets, London (2014)
3. Nowack, B.: The occurrence, behavior, and effects of engineered nanomaterials in the environment. In: Kim, J. (ed.) Advances in Nanotechnology and the Environment, pp. 197–207. Pan Stanford, Singapore (2012)
4. Boxall, A.B..A., Chaudhry, Q., Sinclair, C., Jones, A.D., Aitken, R., Jefferson, B., Watts, C.: Current and Future Predicted Environmental Exposure to Engineered Nanoparticles. Central Science Laboratory, Sand Hutton (2007)
5. Gottschalk, F., Sun, T.Y., Nowack, B.: Environmental concentrations of engineered nanomaterials: review of modeling and analytical studies. Environ. Pollut. **181**, 287–300 (2013)
6. Gottschalk, F., Sonderer, T., Scholz, R.W., Nowack, B.: Modeled environmental concentrations of engineered nanomaterials (TiO_2, ZnO, Ag, CNT, fullerenes) for different regions. Environ. Sci. Technol. **43**, 9216–9222 (2009)
7. Gottschalk, F., Sonderer, T., Scholz, R.W., Nowack, B.: Possibility and limitations of modeling environmental exposure to engineered nanomaterials by probabilistic material flow analysis. Environ. Toxicol. Chem. **29**, 1036–1048 (2010)
8. Domingos, R.F., Baalousha, M.A., Yon, J.N., Reid, M. M., Tufenkji, N., Lead, J.R., Leppard, G.G., Wilkinson, K.J.: Characterizing manufactured nanoparticles in the environment: multimethod determination of particle sizes. Environ. Sci. Technol. **43**, 7277–7284 (2009)
9. Tiede, K., Hassellöv, M., Breitbarth, E., Chaudhry, Q., Boxall, A.B..A.: Considerations for environmental fate and ecotoxicity testing to support environmental risk assessments for engineered nanoparticles. J. Chromatogr. A **1216**, 503–509 (2009)
10. Fairbairn, E.A., Keller, A.A., Mädler, L., Zhou, D., Pokhrel, S.: Metal oxide nanoparticles in seawater: linking physiochemical characteristics with biological response in sea urchin development. J. Hazard. Mater. **192**, 1565–1571 (2011)
11. Bian, S.W., Mudunkotuwa, I.A., Rupasinghe, T., Grassian, V.H.: Aggregation and dissolution of 4 nm ZnO nanoparticles in aqueous environments: influence of pH, ionic strength, size, and adsorption of humic acid. Langmuir **27**, 6059–6068 (2011)
12. Tang, E., Cheng, G., Ma, X., Pang, X., Zhao, Q.: Surface modification of zinc oxide nanoparticle by PMAA and its dispersion in aqueous system. Appl. Surf. Sci. **252**, 5227–5232 (2006)
13. Lowry, G.V., Wiesner, M.R.: Environmental considerations: occurrences, fate, and characterization of nanoparticles in the environment. In: Monteiro-Riviere, N.A., Lang Tran, C. (eds.) Nanotoxicology: Characterization, Dosing and Health Effects, pp. 369–389. Informa Healthcare, New York (2007)
14. Yang, K., Lin, D., Xing, B.: Interactions of humic acid with nanosized inorganic oxides. Langmuir **25**, 3571–3576 (2009)
15. Wang, X., Lu, J., Xu, M., Xing, B.: Sorption of pyrene by regular and nanoscaled metal oxide particles: influence of adsorbed organic matter. Environ. Sci. Technol. **42**, 7267–7272 (2008)
16. Lipovsky, A., Tzitrinovich, Z., Friedmann, H., Applerot, G., Gedanken, A., Lubart, R.: EPR study of visible light-induced ROS generation by nanoparticles of ZnO. J. Phys. Chem. C **113**, 15997 (2009)
17. Scown, T.M., van Aerle, R., Tyler, C.R.: Review: do engineered nanoparticles pose a significant threat to the aquatic environment? Crit. Rev. Toxicol. **40**(7), 653–670 (2010)
18. Peng, X., Palma, S., Fisher, N.S., Wong, S.S.: Effect of morphology of ZnO nanostructures on their toxicity to marine algae. Aquat. Toxicol. **102**, 186–196 (2011)
19. Miao, A.J., Zhang, X.Y., Luo, Z., Chen, C.S., Chin, W. C., Santschi, P.H., Quigg, A.: Zinc oxide-engineered nanoparticles: dissolution and toxicity to marine phytoplankton. Environ. Toxicol. Chem. **29**(12), 2814–2822 (2010)
20. Wong, S.W.Y., Leung, P.T.Y., Djurišić, A.B.., Leung, K.M.Y.: Toxicities of nano zinc oxide to five marine organisms: influences of aggregate size and ion solubility. Anal. Bioanal. Chem. **396**, 609–618 (2010)

21. Manzo, S., Miglietta, M.L., Rametta, G., Buono, S., Di Francia, G.: Toxic effects of ZnO nanoparticles towards marine algae *Dunaliella tertiolecta*. Sci. Total Environ. **445–446**, 371–376 (2013)
22. Miller, R.J., Lenihan, H.S., Muller, E.B., Tseng, N., Hanna, S.K., Keller, A.A.: Impact of metal oxide nanoparticles on marine phytoplankton. Environ. Sci. Technol. **44**(19), 7329–7334 (2010)
23. Keller, A.A., Wang, H., Zhou, D., Lenihan, H.S., Cherr, G., Cardinale, B.J., Miller, R., Ji, Z.: Stability and aggregation of metal oxide nanoparticles in natural aqueous matrices. Environ. Sci. Technol. **44**, 1962–1967 (2010)
24. Trevisan, R., Delapedra, G., Mello, D.F., Arl, M., Schmidt, E.C., Meder, F., Monopoli, M., Cargnin-Ferreira, E., Bouzon, Z.L., Fisher, A.S., Sheehan, D., Dafre, A.: Gills are an initial target of zinc oxide nanoparticles in oysters *Crassostrea gigas*, leading to mitochondrial disruption and oxidative stress. Aquat. Toxicol. **153**, 27–38 (2014)
25. Buffet, P.E., Amiard-Triquet, C., Dybowska, A., Faverney, C.R., Guibbolini, M., Valsami-Jones, E., Mouneyrac, C.: Fate of isotopically labeled zinc oxide nanoparticles in sediment and effects on two endobenthic species, the clam *Scrobicularia plana* and the ragworm *Hediste diversicolor*. Ecotoxicol. Environ. Saf. **84**, 191–198 (2012)
26. Montes, M.O., Hanna, S.K., Lenihan, H.S., Keller, A. A.: Uptake, accumulation, and biotransformation of metal oxide nanoparticles by a marine suspension-feeder. J. Hazard. Mater. **225–226**, 139–145 (2012)
27. Hanna, S.K., Miller, R.J., Muller, E.B., Nisbet, R.M., Lenihan, H.S.: Impact of engineered zinc oxide nanoparticles on the individual performance of *Mytilus galloprovincialis*. PLoS One **8**(4), e61800 (2013)
28. Ates, M., Daniels, J., Arslan, Z., Farah, I.O., Rivera, H.F.: Comparative evaluation of impact of Zn and ZnO nanoparticles on brine shrimp (*Artemia salina*) larvae: effects of particle size and solubility on toxicity. Environ. Sci. Process. Impacts. **15**, 225–233 (2013)
29. Jarvis, T.A., Miller, R.J., Lenihan, H.S., Bielmyer, G. K.: Toxicity of ZnO nanoparticles to the copepod *Acartia tonsa*, exposed through a phytoplankton diet. Environ. Toxicol. Chem. **32**(6), 1264–1269 (2013)
30. Fabrega, J., Tantra, R., Amer, A., Stolpe, B., Tomkins, J., Fry, T., Lead, J.R., Tyler, C., Galloway, T.S.: Sequestration of zinc from zinc oxide nanoparticles and life cycle effects in the sediment dweller amphipod *Corophium volutator*. Environ. Sci. Technol. **46**, 1128–1135 (2012)
31. Larner, F., Dogra, Y., Dybowska, A., Fabrega, J., Stolpe, B., Bridgestock, L.J., Goodhead, R., Weiss, D.J., Moger, J., Lead, J.R., Valsami-Jones, E., Tyler, C.R., Galloway, T.S., Rehkämper, M.: Tracing bioavailability of ZnO nanoparticles using stable isotope labeling. Environ. Sci. Technol. **46**, 12137–12145 (2012)
32. Handy, R.D., von der Kammer, F., Lead, J.R., Hassellöv, M., Owen, R., Crane, M.: The ecotoxicology and chemistry of manufactured nanoparticles. Ecotoxicology **17**, 287–314 (2008)
33. Zhu, X., Wang, J., Zhang, X., Chang, Y., Chen, Y.: The impact of ZnO nanoparticle aggregates on the embryonic development of zebrafish (*Danio rerio*). Nanotechnology **20**, 195103 (2009)
34. Xiong, D., Fang, T., Yu, L., Sima, X., Zhu, W.: Effects of nano-scale TiO_2, ZnO and their bulk counterparts on zebrafish: acute toxicity, oxidative stress and oxidative damage. Sci. Total Environ. **409**, 1444–1452 (2011)
35. Wong, S.W.Y., Leung, K.M.Y., Djurišić, A.B..: A comprehensive review on the aquatic toxicity of engineered nanomaterials. Rev. Nanosci. Nanotechnol. **2**(2), 79–105 (2013)
36. Ma, H., Williams, P.L., Diamond, S.A.: Ecotoxicity of manufactured ZnO nanoparticles – a review. Environ. Pollut. **172**, 76–85 (2013)
37. Bodansky, M.: Biochemical studies on marine organisms: II. The occurrence of zinc. J. Biol. Chem. **44**, 399–407 (1920)
38. Poynton, H.C., Lazorchak, J.M., Impellitteri, C.A., Smith, M.E., Rogers, K., Patra, M., Hammer, K.A., Allen, H.J., Vulpe, C.D.: Differential gene expression in *Daphnia magna* suggests distinct modes of action and bioavailability for ZnO nanoparticles and Zn ions. Environ. Sci. Technol. **45**, 762–768 (2011)
39. Hoffmann, M., Hotze, E.M., Wiesner, M.R.: Reactive oxygen species generation on nanoparticulate material. In: Wiesner, M.R., Bottero, J.Y. (eds.) Environmental Nanotechnology: Applications and Impacts of Nanomaterials, pp. 155–171. McGraw-Hill, New York (2007)
40. Ma, H., Kabengi, N.J., Bertsch, P.M., Unrine, J.M., Glenn, T.C., Williams, P.L.: Comparative phototoxicity of nanoparticulate and bulk ZnO to a free-living nematode *Caenorhabditis elegans*: the importance of illumination mode and primary particle size. Environ. Pollut. **159**(6), 1473–1480 (2011)
41. Lee, W.M., An, Y.J.: Effects of zinc oxide and titanium dioxide nanoparticles on green algae under visible, UVA, and UVB irradiations: no evidence of enhanced algal toxicity under UV pre-irradiation. Chemosphere **91**, 536–544 (2013)
42. Cairns, J., Health, A.G., Parker, B.C.: The effects of temperature upon the toxicity of chemicals to aquatic organisms. Hydrobiologia **47**, 135–171 (1975)
43. Wong, S.W.Y., Leung, K.M.Y.: Temperature-dependent toxicities of nano zinc oxide to marine diatom, amphipod and fish in relation to its aggregation size and ion dissolution. Nanotoxicology **8**(51), 24–35 (2014)
44. Nugegoda, D., Rainbow, P.S.: Salinity, osmolality, and zinc uptake in *Palaemon elegans* (Crustacea: Decapoda). Mar. Ecol. Prog. Ser. **55**, 149–157 (1989)
45. Park, J., Kim, S., Yoo, J., Lee, J.S., Park, J.W., Jung, J.: Effect of salinity on acute copper and zinc toxicity to

Tigriopus japonicus: the difference between metal ions and nanoparticles. Mar. Pollut. Bull. (2014). doi:10.1016/j.marpolbul.2014.04.038
46. Leung, K.M.Y., Merrington, G., Warne, M.St.J., Wenning, R.J.: Scientific derivation of environmental quality benchmarks for the protection of aquatic ecosystems: challenges and opportunities. Environ. Sci. Pollut. Res. **21**, 1–5 (2014)

Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae

Florence Mouchet[1,2], Périne Landois[3], Floriane Bourdiol[1,2,3], Isabelle Fourquaux[4], Pascal Puech[5], Emmanuel Flahaut[3,6] and Laury Gauthier[1,2]
[1]EcoLab – Laboratoire d'Écologie Fonctionnelle et Environnement, Université de Toulouse, INP, UPS, Castanet Tolosan, France
[2]CNRS UMR 5245, EcoLab, Castanet Tolosan, France
[3]Institut Carnot Cirimat, Université de Toulouse, UPS, INP, Toulouse, France
[4]CMEAB, Centre de Microscopie Electronique Appliquée à la Biologie, Université Paul Sabatier, Faculté de Médecine Rangueil, Toulouse, France
[5]CEMES, Toulouse, France
[6]CNRS, Institut Carnot Cirimat, Toulouse, France

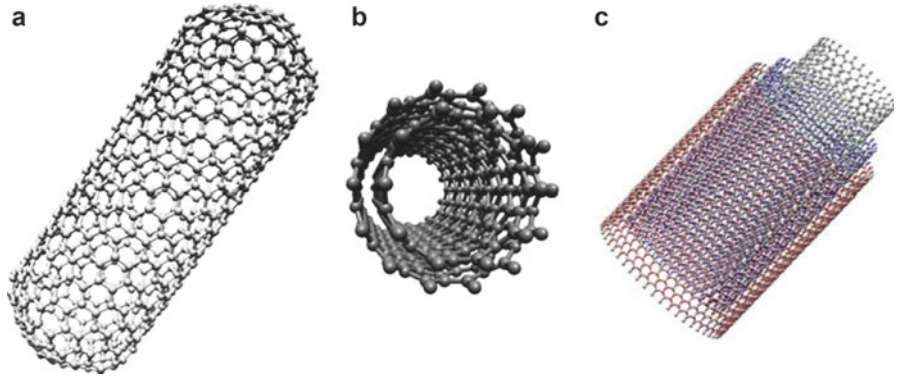
Synonyms

Amphibian larvae; Carbon nanotubes; Ecotoxicity

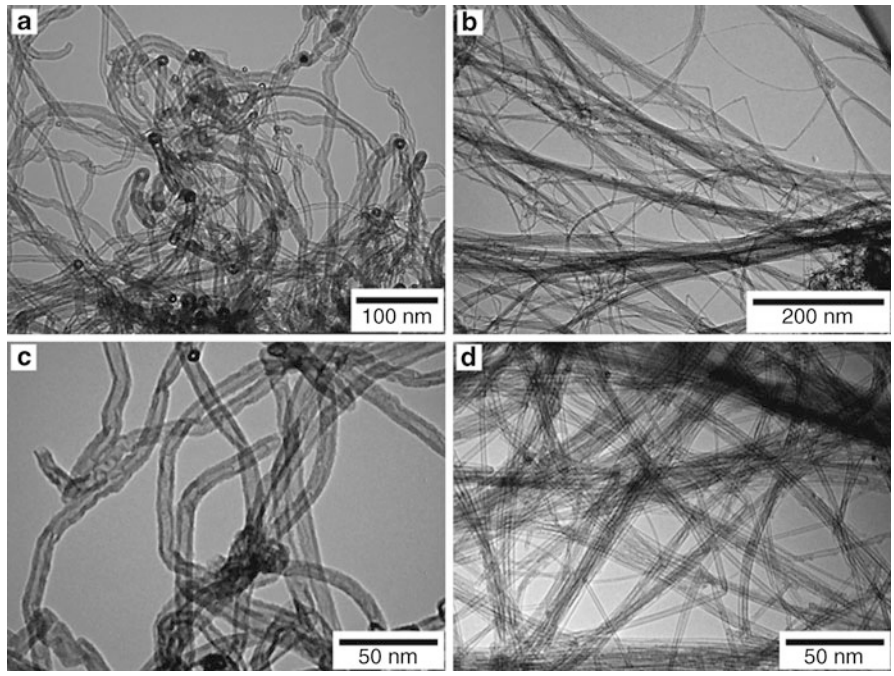
Definition

According to Truhaut [1], ecotoxicology is defined as the branch of toxicology concerned with the study of toxic effects, caused by natural or synthetic pollutants, to the constituents of ecosystems, animals (including human), vegetables, and microbial, in an integral context. Carbon NanoTubes (CNT), a man-made allotrope of carbon, are considered as emergent potential contaminants into the environment. The aim of the ecotoxicology is then to evaluate the potential toxicity of CNT toward living organisms. Amphibian's larvae are used as biological models to study the potential toxicity of CNT.

Carbon NanoTubes (CNT) are a man-made allotrope of carbon (1D). They were discovered in 1991 by Iijima [2] and are intensively studied since 20 years by the scientific community. Their structure can be described as a graphene sheet rolled up to form a hollow cylinder. CNT properties depend on their structure. Two main types of CNT are distinguished [3, 4]: Single-Walled CNT (SWNT) consisting of a single sheet and Multi-Walled CNT (MWNT) consisting of several concentric layers, having a larger diameter (Fig. 1). CNT may be closed by half-fullerene-like molecules. Among the MWNT, Double-Walled CNT (DWNT) are at the frontier between SWNT and MWNT, with morphology and properties very close to SWNT. CNT have a diameter from 1 to 100 nm and a length from less than one µm up to tens of µm or more (Fig. 2). Due to their exceptional physical (mechanic, electronic, thermal) and chemical properties particularly because of their nano-size, interest in CNT has grown rapidly and application areas are very wide: nanocomposites, electron transmitters, electromagnetic shielding, batteries, supercapacitors, gas storage (including hydrogen), and nano-medicine [5]. Their current applications in the daily life of developed society are numerous as prototypes (flat-screens, tyres, sport equipments, etc.) and some others are emerging (paints, technical clothes, pharmaceutical products, etc.). Global CNT production capacity was estimated in 2002 over 2.5 t per day, and the correspondent market was estimated at $12 millions for 2002 and was expected to grow up to $700 millions by 2005 [6]. The recent analysis of John Oliver [7] reports that the global market for CNT grades based on committed production reached $103 millions in 2009. This market is projected to reach $167.2 millions in 2010 and $1 billion in 2014 at a compound annual growth rate of 58.9 %. This same analyst states that MWNT grade market was nearly $103 million in 2009 and projected to reach $161 million in 2010 and $865.5 million in 2014, for a 5-year compound annual growth rate (CAGR) of 53.1 %.

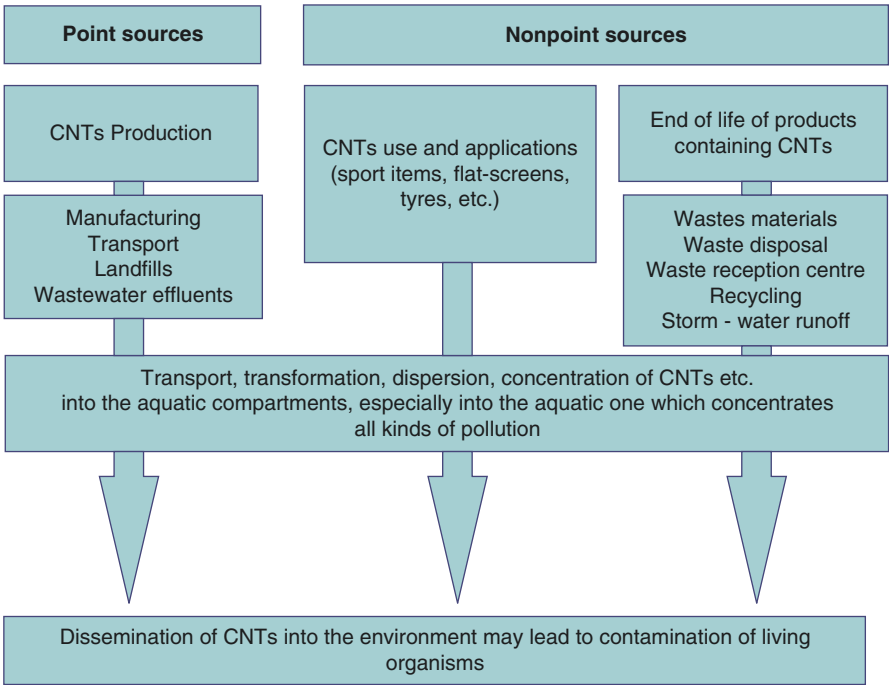


Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 1 Schematic representation of (**a**) Single-Walled Carbon Nanotube (SWNT), closed ends, (**b**) Double-Walled Carbon Nanotube, open ends, and (**c**) Multi-Walled Carbon Nanotube (MWNT), open ends (not to scale)



Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 2 TEM images of MWNT (**a**, **c**) and DWNT (**b**, **d**)

In respect to this economical context and the exponential production of CNT, it is likely that some of them will get into the environment by numerous ways (Fig. 3), during each step of their life cycle, from their production until the end of their life cycle, especially in the aquatic compartment which concentrates most pollutions. Indeed, CNT releases into the environment might come from different sources. Point sources may be the release of CNT at the production site, during manufacturing (via accident of release, landfills or wastewater effluents, etc.) or during their transport from the production site to industrial users. Nonpoint sources may be the dissemination of

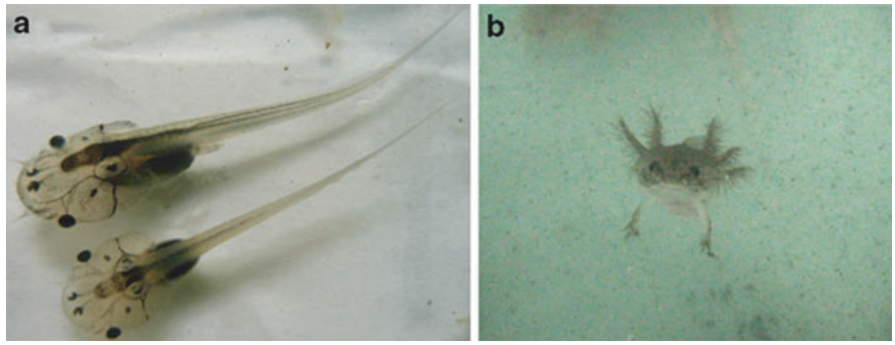


Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 3 Environmental context of the CNT contamination toward the environment

CNT during the use of products containing CNT, via numerous ways such as wear of products, until the end of their life (wet deposition from the atmosphere, storm-water runoff, surface water leakage with the waste disposal in dumps, etc.) Contamination by CNT, as for the case of all chemical substances, may involve biogeochemical cycles, i.e., transport, dispersion, transformation, and/or concentration into compartments of the environment, but also may involve biological systems due to the potential ingestion, uptake, and excretion mechanisms of living organisms.

In this context, CNT could be classified as emergent contaminants and must receive considerable attention as new, unknown, and potentially hazardous materials. To date, several main questions are established by the scientific community about the quantities of CNT disposed into the environment, their distribution into the environmental compartments, and their preferential areas of accumulation in these compartments, and finally their persistence or potential degradation in relation with some particular environmental conditions. On top of these geochemical concerns, the biological effects at molecular, sub-cellular, and cellular level from organism to ecosystem are of great concern and constitute the ecotoxicity area.

One strategy is to evaluate the potential ecotoxicological risk of CNT in the same way that risk assessment method is carried out during the assessment of the potential risk of a contaminant (e.g., chemical substances as pesticides or heavy metals) into the environment. This approach uses standardized methods in laboratory studies to determine biological effects using biological models, for example, amphibian larvae. Indeed, amphibians are well-known environmental health warning organisms due to their biphasic life cycle, permeable and sensitive eggs, skin, and gills [8]. Their specific physiology makes them particularly sensitive to the presence of contaminants in the water, which influences their behavior [9, 10]. In the same way, the huge quantity of DNA shared in a small number of chromosomes makes them good models for cytogenetic studies.

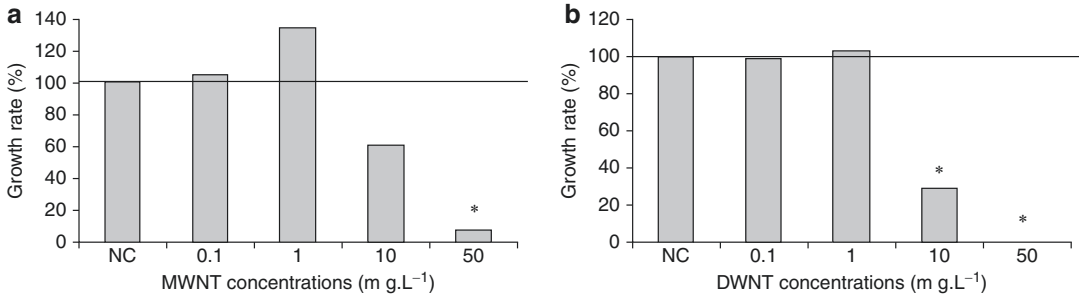


Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 4 *Xenopus laevis* (**a**) and *Ambystoma mexicanum* (**b**) larvae

Among amphibian species, *Xenopus* and axolotl models (Fig. 4) were used to evaluate the potential ecotoxicity of CNT in laboratory conditions [11–14] consist in exposing amphibian larvae to different concentrations of CNT in water using standardized methods or adapted ones if necessary. The nature of CNT depends on the synthesis method used (Catalytic Chemical Vapor Deposition (CCVD), Electric–arc discharge, and Laser ablation are the most commonly used), and on their physical–chemical state (e.g., functionalization, dispersion). Three biological endpoints were investigated in amphibian larvae: the acute toxicity (mortality), the chronic one according to the growth of larvae, and the genetic toxicity according to the induction of micronuclei in erythrocytes in the running blood as the expression of the clastogenic and/or aneugenic effects. Among toxic effects, genotoxicity may durably affect the aquatic ecosystems. The interaction between genotoxic compounds and DNA initially may cause structural changes in the DNA molecule. Unrepaired damage can generate other cell lesions and thus lead to tumor formation [15]. The evaluation of genotoxicity is particularly relevant because its assessment constitutes a warning system at short term to measure long-term effects at infra-lethal concentrations. Genotoxicity may be a useful biomarker as an intermediate endpoint in carcinogenesis, for example. This genetic toxicity represents the hidden risk because of its expression at long and very long term in the environment, contrary to the acute toxicity, for instance, which constitutes the visible risk. In the case of genetic toxicity, measured criteria will be DNA damage whereas in the case of acute toxicity, criteria will be mortality. In amphibian larvae, genome mutations may result in the formation of micronuclei, which are a consequence of chromosome fragmentation or malfunction of the mitotic apparatus.

Acute, chronic, and genetic toxicities are the endpoints which were being investigated in amphibian larvae according to the French national [16] and international [17] standardized methods of the MicroNucleus Test (MNT). The MNT has been widely used with many amphibian species: *Pleurodeles waltl*, *Ambystoma mexicanum*, and *Xenopus laevis* [8, 9, 18–20]. The sensitivity and reliability of the MNT to detect chromosomal and/or genomic mutations make it a good method to analyze the potential cytogenetic damage caused by pure substances [21–23]. The use of MNT may provide an important tool for the prediction of the potential long-term effects on amphibians in the environment. One of the key functions of such biomarkers (micronucleus) is providing an "early warning" signal of significant biological effects (changes at the genetic/molecular level) with sub-organism (molecular, biochemical, and physiological) responses preceding those occurring at higher levels of biological organization such as cellular, tissue, organ, whole-body levels and *in fine* at population level.

In this objective, two amphibian species were exposed to different kinds of CNT synthesized by



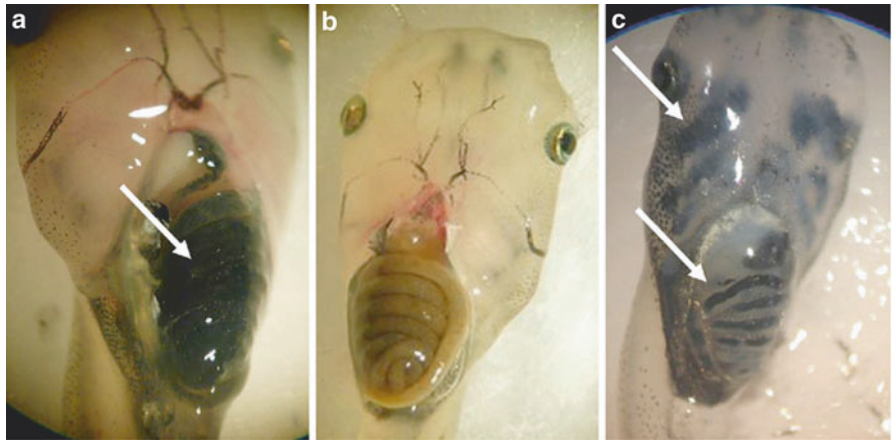
Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 5 Growth rate of *Xenopus larvae* exposed to 0.1, 1, 10 and 50 mg/L of raw CNT. * corresponds to a significant different size of larvae compared to the negative control group (mean value). *NC* Negative control

different CCVD synthesis processes: DWNT [24] and MWNT. The three endpoints, described before, were investigated and analyzed. Moreover, the presence of CNT was investigated in the larvae using traditional microscopy methods, but also transmission electron microscopy (TEM) and Raman spectroscopy. This theme around CNT localization in biological matrix is a delicate point in this kind of work. Raman spectroscopy is an optical technique. Monochromatic light (laser) is focused on the sample. The scattered light, composed of the elastic scattering (Rayleigh) and the inelastic scattering due to an interaction with some quanta of energy like phonons (Raman), is analyzed through a spectrometer. With SWNT or DWNT, one-dimensional resonance phenomenon increases strongly the Raman signal, while with large MWNT, the Raman signal is broadened and weaker.

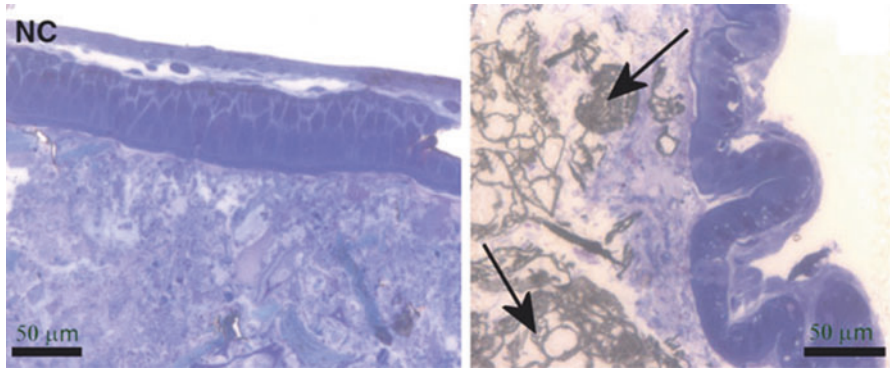
In axolotl, Mouchet et al. [11] exposed the amphibian larvae under adapted standardized conditions to raw DWNT. The results obtained demonstrated that raw DWNT are neither acutely toxic nor genotoxic to larvae whatever their concentration in the water (ranging from 1 to 1,000 mg. L^{-1}). Nevertheless, visual observation under binocular showed that inside the gut there were large black masses of ingested CNT even at the lowest concentration tested.

Mouchet et al. [12–14] exposed *Xenopus* larvae under standardized conditions to raw DWNT or MWNT in water (ranging from 1 to 50 mg. L^{-1}). The results showed no genotoxicity in erythrocytes of larvae. In contrast, results showed chronic toxicity at 50 mg L^{-1} of MWNT (Fig. 5a) and from 10 mg L^{-1} of DWNT (Fig. 5b). A visual inspection of the larvae under binocular after exposure showed that internal gills were black-colored and seemed to be blocked up in the case of DWNT (Fig. 6c) but not in the case of MWNT (Fig. 6a), compared to larvae exposed without CNT (Fig. 6b). In the same way, the visual inspection of the digestive tract in larvae exposed to DWNT or MWNT revealed the presence of black masses into the intestine lumen suggesting that the larvae may have absorbed CNT present in the exposure medium (Fig. 6). In amphibian larvae, xenobiotics could enter in the body of larvae by two principal ways: the dermal route and the digestive route leading to potentially high ingestion rates of particles, especially particles of food susceptible to vehicle agglutinated CNT bundles and water containing suspended particles.

Using microscopic observations of intestine cross-sections in control and exposed larvae (Fig. 7), black masses were clearly identified within food intake in the intestine lumen in exposed larvae, but black masses were not observed in epithelial or in chorionic tissues. It is not possible to identify CNT in the black masses of the intestine lumen by optical microscopy. Using TEM, the intestinal lumen, villi, and intestinal cells can be observed in detail. Figure 8 shows fibers in the intestinal lumen of larvae exposed to DWNT, near the villi but some of them are also observed in the control (not showed). These observations did not lead to the



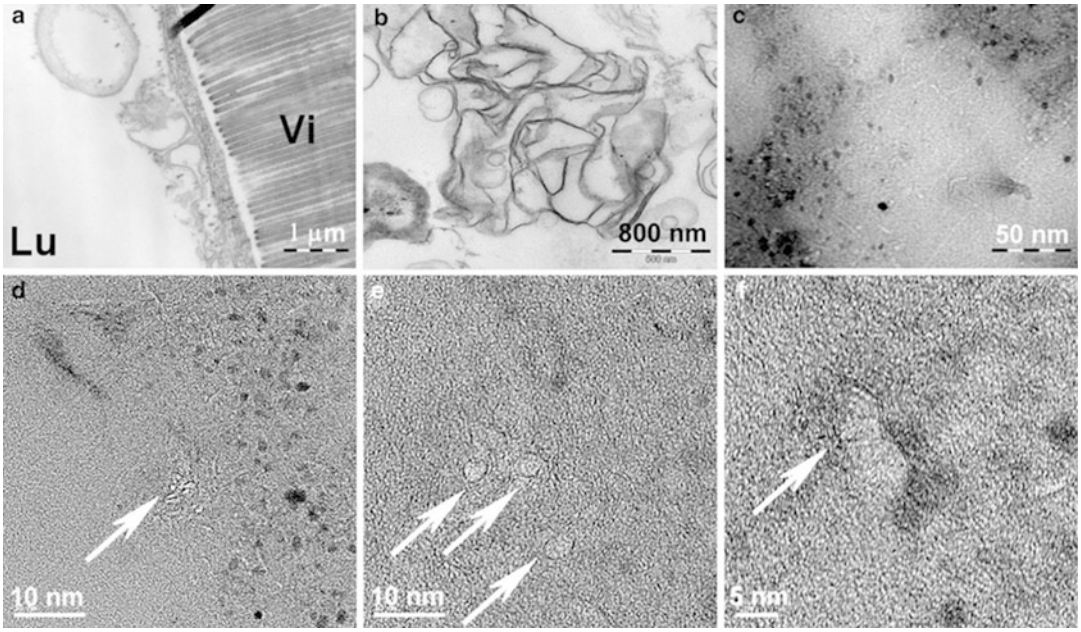
Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 6 Localization of CNT in larvae exposed to 10 mg L^{-1} of MWNT (**a**) or DWNT (**c**), in comparison to a control (exposure without CNT) larva (**b**). White arrows indicate the presence of CNT in gills in the case of DWNT exposure (**c**) and in the intestine in the case of both CNT (**a**) and (**c**)



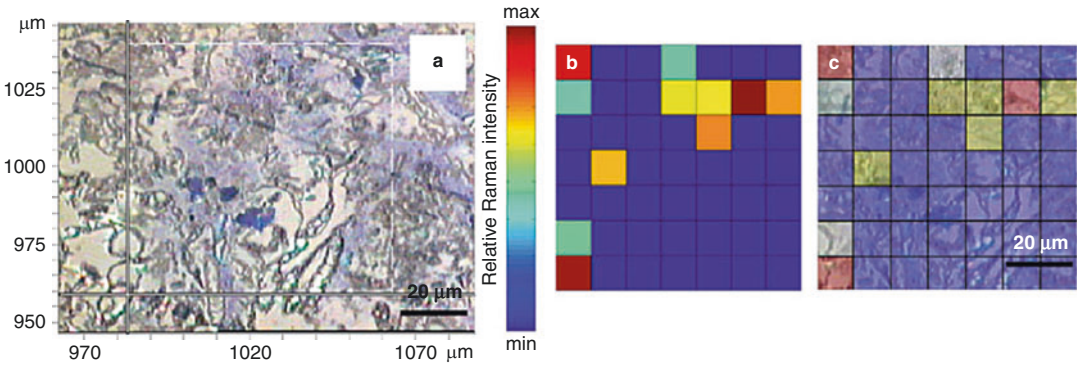
Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 7 Optical microscopy photographs of histological thins of intestine of *Xenopus* larvae exposed to 10 mg/L of DWNT (on the *right*) compared to the control group (on the *left*). *NC* Negative control. Black arrows indicate the presence of black fibrous structures in the intestine lumen of CNT-exposed larvae. Sections showed were stained with methylene blue

confirmation of the presence of CNT in the digestive tissue at cellular level. Fibers were then observed at higher magnification (Fig. 8b, c). Discrimination of these fibers was realized under High-Resolution TEM. Figure 8d shows isolated CNT. At this magnification (Fig. 8e, f), it was possible to show the presence of CNT in exposed larvae. Nevertheless, TEM is a local technique, also leading to possible ambiguity in terms of identification between CNT and biological compounds (ribosomes, organelles, etc.). Another method such as Raman spectroscopy is thus necessary to discriminate between them.

Two kinds of Raman analysis were performed on *Xenopus* intestine. The G or G'2D band can be used to identify and localize CNT in a biological matrix [12]. (1) Raman mapping is presented in Fig. 9 showing the optical microscopic image (slide) of an intestine sample of animal exposed to DWNT (Fig. 9a) and the corresponding Raman image from the same area (Fig. 9b). The superposition (Fig. 9c) clearly shows that, on the one hand, some of the black structures visible in optical microscopy do not contain any DWNT, but that, on the other hand, CNT are present in some other places although optical microscopy does not

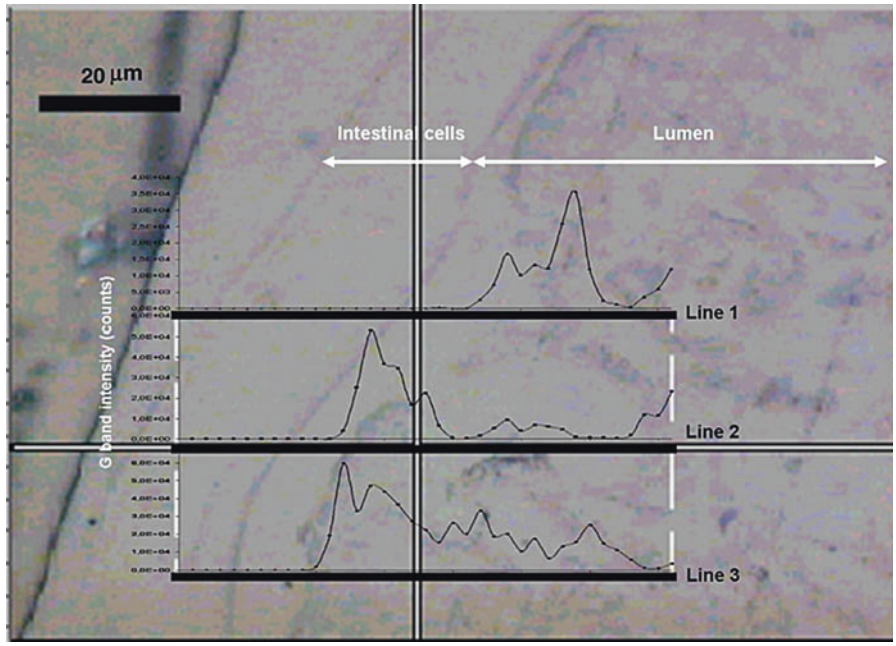


Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 8 Observation of area containing black fibrous structures in the intestine lumen of CNT-exposed larvae using Transmission Electron Microscopy (TEM) and 50 nm High-Resolution Transmission Electron Microscopy Electron micrographs (HRTEM) of *Xenopus* larvae intestine in animals exposed to DWNT. (**a**) TEM image of the intestine of a control larva, showing the lumen (Lu) from where nutrients are absorbed, as well as the intestinal wall (Vi: villi); (**b-f**) successive magnifications of (**a**) showing metal particles associated with CNT (**c**, **d**) and structures which could correspond to CNT (*white arrows*)



Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 9 Optical microscopy image of the studied intestine area (**a**); Raman intensity map (**b**) corresponding to the selected observation area of (**a**); Superposition of optical image and Raman intensity map (**c**). Red corresponds to maximum concentration of DWNT, yellow around half, and gray around 25 %. Blue corresponds to minimum concentration

show any clear evidence. (2) Figure 10 presents Raman line analysis in intestine, evidencing the presence of both kinds of CNT (DWNT and MWNT) in the intestine lumen, whereas no CNT were localized in the intestinal cells. Indeed, a sharp drop in intensity of the specific CNT band was observed. Raman spectroscopy analysis of different tissues of the larvae did not evidence



Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae, Fig. 10 Location of DWNT by Raman spectroscopy. The G band intensity extracted from Raman spectra is characteristic of the CNT. By acquiring spectroscopic lines, we follow the DWNT profile. Disappearance of the DWNT signal at the intestinal cell barrier is observed

the presence of CNT neither in blood (results not shown), nor in liver of amphibian suggesting that CNT do not cross the intestinal barrier.

To summarize, results showed different responses between the two amphibian species, *Xenopus* (anouran) and axolotl (urodeles) in relation with their different biology, larval growth, behavior, etc. A difference in sensitivity was generally observed in response to chemical pollutants, and *Xenopus* larvae are often found more sensitive than urodelian species to environmental substances [8]. Results highlight the potential risk of the studied CNT since acute and chronic toxicities were observed in larvae exposed to raw CNT. In *Xenopus*, exposure to CNT leads to toxicity usually above 10 mg. L^{-1}, which corresponds to a high concentration. This level of contamination seems to be probably unrealistic in the global environmental context, except in the case of an accident at the production site, for example. Since CNT are ingested by larvae, one cannot exclude the possibility that CNT may be found later in the food chain, once released into the environment. In contrast, no genotoxicity was observed in larvae exposed to raw CNT in these conditions. Acute toxicity observed in *Xenopus* larvae may be mediated by several biological mechanisms: (1) branchial obstruction and/or gas exchange perturbation and/or anoxia, (2) and/or intestinal obstruction, (3) and/or competition between CNT and food. The presence of CNT was evidenced in the larvae using both electron microscopy and Raman analysis. The Raman imaging technique is very sensitive and thus allows assessing without doubt that the intestine of CNT exposed larvae contains CNT, as already suggested by the simple visual inspection of the larvae and the TEM observation. However, the image superposition analysis using Raman spectroscopy (Figs. 9 and 10) clearly sharpened the first interpretation obtained by microscopy. Indeed, some of the black spots observed by optical microscopy do not contain any CNT, whereas CNT are clearly identified in some other places, although this technique does not show any clear evidence. Moreover, the abrupt vanishing of the

intensity of the specific CNT bands in Raman spectroscopy when crossing the intestinal cell wall strongly suggests that raw CNT do not cross the intestinal barrier of amphibians.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ Genotoxicity of Nanoparticles
- ▶ In Vivo Toxicity of Carbon Nanotubes
- ▶ Synthesis of Carbon Nanotubes
- ▶ Transmission Electron Microscopy

References

1. Truhaut, R.: Eco-toxicology-objectives, principles and perspectives. Ecotoxicol. Environ. Saf. **1**(2), 151–173 (1977)
2. Iijima, S.: Helical microtubules of graphitic carbon. Nature **354**(6348), 5658 (1991)
3. Bethune, D.S., Kiang, C.H., Devries, M.S., Gorman, G., Savoy, R., Vazquez, J.: Cobalt-catalyzed growth of carbon nanotubes with single-atomic-layerwalls. Nature **363**(6430), 605–607 (1993)
4. Iijima, S., Ichihashi, T.: Single-shell carbon nanotubes of 1-nm diameter. Nature **363**(6430), 603–605 (1993)
5. Eklund, P., Ajayan, P., Blackmon, R., Hart, A.J., Kong, J., Pradhan, B., Rao, A., Rinzler, A.: International Assessment of Research and Development of Carbon Nanotube Manufacturing and Applications. World Technology Evaluation Center, Baltimore, p. 138. (2007)
6. Fuji-Keizai USA: Inc. CNT Wordwide. Carbon Nanotubes – Wordwide Status and Outlook: Applications, Applied Industries, Production, R&D and Commercial Implications. p. 153 (2002)
7. NANO24D. BBC Research. Nanotechnologies. Carbon Nanotubes: Technologies and Global Markets. Published Jan 2000. Analyst: John Oliver. http://www.bccresearch.com/report/NAN024D.html (2010). Accessed 15 May 2010
8. Gauthier, L.: The amphibian micronucleus test, a model for in vivo monitoring of genotoxic aquatic pollution. Review paper. Int. J. Batrachol. **14**(2), 53–84 (1996)
9. Gauthier, L., Tardy, E., Mouchet, F., Marty, J.: Biomonitoring of the genotoxic potential (micronucleus assay) and detoxifying activity (EROD induction) in the river Dadou (France), using the amphibian *Xenopus laevis*. Sci. Total Environ. **323**, 47–61 (2004)
10. Bridges, C.M., Dwyer, F.J., Hardesty, D.K., Whites, D.W.: Comparative contaminant toxicity: are amphibian larvae more sensitive than fish? Bull. Env. Contam. Toxicol. **69**, 562–569 (2002)
11. Mouchet, F., Landois, P., Flahaut, E., Pinelli, E., Gauthier, L.: Assessment of the potential in vivo ecotoxicity of Double-Walled Carbon Nanotubes (DWNT) in water, using the amphibian *Ambystoma mexicanum*. Nanotoxicology **1**(2), 149–156 (2007)
12. Mouchet, F., Landois, P., Sarremejean, E., Bernard, G., Puech, P., Pinelli, E., Flahaut, E., Gauthier, L.: Characterisation and in vivo ecotoxicity evaluation of double-wall carbon nanotubes in larvae of the amphibian *Xenopus laevis*. Aquat. Toxicol. **87**, 127–137 (2008)
13. Mouchet, F., Landois, P., Puech, P., Pinelli, E., Flahaut, E., Gauthier, L.: CNT ecotoxicity in amphibians: assessement of Multi Walled Carbon Nanotubes (MWNT) and Comparison with Double Walled Carbon Nanotubes (DWNT). Special focus environmental toxicity of nanoparticles. Nanomedecine **5**(6), 963–974 (2010)
14. Mouchet, F., Landois, P., Datsyuk, V., Puech, P., Pinelli, E., Flahaut, E., Gauthier, L.: Use of the international amphibian micronucleus standardized procedure (ISO 21427-1) for in vivo evaluation of double-walled toxicity and genotoxicity in water. Environ. Toxicol. **26**(2), 136–145 (2011)
15. Vuillaume, M.: Reduced oxygen species, mutation, induction and cancer initiation. Mutat. Res. **186**(1), 43–72 (1987)
16. AFNOR: Association française de normalisation (the French National Organization for quality regulations). Norme NFT 90–325. Qualité de l'Eau. Evaluation de la génotoxicité au moyen de larves d'amphibien (*Xenopus laevis*, *Pleurodeles waltl*). AFNOR – Paris, p. 17. (2000)
17. ISO: ISO International Standard. Water quality – Evaluation of genotoxicity by measurement of the induction of micronuclei – Part 1: Evaluation of genotoxicity using amphibian larvae. ISO 21427–1, ICS: 13.060.70, GENOVA – CH (2006)
18. Ferrier, V., Gauthier, L., Zoll-Moreux, C., L'Haridon, J.: Genotoxicity tests in amphibians – a review. In: Microscale Testing in Aquatic Toxicology: Advances, Techniques and Practice, vol. 35, pp. 507–519. CRC, Boca Raton (1998)
19. Mouchet, F., Gauthier, L., Mailhes, C., Ferrier, V., Devaux, A.: Comparative study of the comet assay and the micronucleus test in amphibian larvae (*Xenopus laevis*) using benzo(*a*)pyrene, ethyl methanesulfonate, and methyl methanesulfonate: establishment of a positive control in the amphibian comet assay. Environ. Toxicol. **20**(1), 74–84 (2005)
20. Mouchet, F., Gauthier, L., Mailhes, C., Ferrier, V., Devaux, A.: Comparative evaluation of the genotoxicity of captan in amphibian larvae (*Xenopus laevis* and *Pleurodeles waltl*) using the comet assay and the micronucleus test. Environ. Toxicol. **21**(3), 264–277 (2006)

21. Mouchet, F., Cren, S., Cunienq, C., Deydier, E., Guilet, R., Gauthier, L.: Assessment of lead ecotoxicity in water using the amphibian larvae (*Xenopus laevis*) and preliminary study of its immobilization in meat and bone meal combustion residues. Biometals **20**(2), 113–127 (2007)
22. Jaylet, A., Gauthier, L., Zoll, C.: Micronucleus test using peripheral red blood cells of amphibian larvae for detection of genotoxic agents in freshwater pollution. In: Sandhu, S.S., et al. (eds.) *In Situ* Evaluations of Biological Hazards of Environmental Pollutants, pp. 71–80. Plenum Press, New York (1990)
23. Gauthier, L., Van Der Gaag, M.A., L'Haridon, J., Ferrier, V., Fernandez, M.: In vivo detection of waste water and industrial effluent genotoxicity: use of the newt micronucleus test (Jalyet test). Sci. Total Environ. **138**, 249–269 (1993)
24. Flahaut, E., Bacsa, R., Peigney, A., Laurent, C.: Gram-scale CCVD synthesis of double-walled carbon nanotubes. Chem. Commun. **12**, 1442–1443 (2003)

Ecotribology

► Green Tribology and Nanoscience

EC-STM

► Electrochemical Scanning Tunneling Microscopy

Effect of Surface Modification on Toxicity of Nanoparticles

Malgorzata J. Rybak-Smith
Department of Engineering Science, University of Oxford, Begbroke, UK

Synonyms

Influence of surface engineering on toxicity of nanoparticles; Influence of surface functionalization on toxicity of nanoparticles

Definitions

Nanoparticles – particles having at least one dimension between 1 and 100 nm.

Toxicity of nanoparticles – undesirable consequence of interactions of nanoparticles with human or mammalian cells (cytotoxicity), organs, such as lungs (pulmonary toxicity), spleen (splenic toxicity), liver (hepatotoxicity), kidneys (nephrotoxicity), immune system (immunocytotoxicity) and the nervous system (neurotoxicity).

Surface modification of nanoparticles – changing of physicochemical or biological characteristics of surface of nanoparticles by introducing chemical functional groups to the surface via creation of covalent bonds (covalent modification) or adsorption of biologically active molecules, such as proteins, surfactants, enzymes, antibodies, nucleic acids (non-covalent modification). Surface modification of nanoparticles can change their physicochemical properties, charge, size, hydrophobicity and make them more biocompatible for biomedical applications.

Nanoparticles are interesting novel materials which have a potential to be effectively used in nanomedicine and biotechnology. Their nano-sized character and exceptional physicochemical properties attract a lot of scientific attention. However, the size of nanoparticles allows them to penetrate various biological barriers within the body and cause toxicity. The potential toxicity of nanoparticles may also arise from interactions of various nanoparticles with biological systems and is highly undesirable. Toxicity of nanoparticles is a major obstacle for their use as potential drug or gene carriers, vaccines, or as biomedical devices. Medina et al. overviewed different classes of nanoparticles, including liposomes, carbon nanotubes (CNTs), gold shell nanoparticles, polymers, metallic nanoparticles, quantum dots (QDs), and carbon nanomaterials [1]. Nanoparticles can be used as biomarkers in molecular diagnosis or in drug delivery [1]. In the same entry, the pharmacological and diagnostic applications and potential toxicity of different classes of nanoparticles have been presented [1].

To minimize the toxicity of nanoparticles, surface engineering and functionalization can be

applied. Surface of nanoparticles, also surface charge, can be modified in numerous ways. Non-covalent modification of the surface of nanoparticles can be achieved by coating or wrapping with biological molecules which may diminish the toxicity of nanoparticles and therefore make them more biocompatible. Coating with proteins, peptides, surfactants, polymers or wrapping with nucleic acids are commonly used approaches. Another approach involves creation of chemical bonds between functional groups present on the surface of nanoparticles and other biological molecules attached to the surface, and is called covalent modification. In this case, the covalent attachment of polyethylene glycol, PEG (PEGylation), carbohydrate or peptide conjugation, and acetylation may be applied. Functionalization of surface of nanoparticles can significantly improve their utility in biomedical applications as long as the functionalization is not toxic for cells and tissues by itself.

The unique physicochemical properties of CNTs make them of great interest for many biomedical applications. However, the presence of structural defects on their surface, metal impurities, length, hydrophobicity, and surface roughness can significantly affect CNT utility in biomedicine. Moreover, CNTs have a large surface area to mass ratio. The greater surface allows the greater interaction of CNTs with human plasma blood proteins, cellular membranes, and adsorption of biological molecules [2]. Smart et al. reviewed biocompatibility aspects of CNTs interactions with fibroblasts, osteoblasts, neuronal cells, ion channels, cellular membranes, antibodies, and the immune system [3].

Yu et al. presented that chemical modifications of surface of nanoparticles may improve their biocompatibility and can change their human plasma protein binding patterns, cytotoxicity, influence Reactive Oxygen Species (ROS) generation and immune responses [4]. Numerous modified MWCNTs showed lower immune responses than their precursor-carboxylated MWCNTs (MWCNT-COOH) [4]. Magrez et al. presented in vitro toxicity studies of carbon-based materials, like multiwalled carbon nanotubes (MWCNTs), carbon nanofibers and carbon nanoparticles on lung tumor cells [5]. All of the tested nanomaterials were toxic. Furthermore, enhanced cytotoxicity was observed after functionalization of the nanoparticles after an acid treatment [5]. The effect of various chemical functionalizations of CNTs on their biocompatibility has been reviewed by Foldvari et al. [6]. Covalent functionalization using 1,3-dipolar cycloaddition is commonly used approach and results in formation of CNT intermediates with amine groups present on the surface which can be further used to attach bioactive molecules such as streptavidin, G-family proteins, or various peptides [6]. Moreover, the CNT intermediates can be used to attach small molecules, like amphotericin B (AmB), an effective antibiotic for treatment of chronic fungal infections [6]. Diminished toxicity and increased antibiotic activity of CNT-AmB conjugate was observed in Jurkat cells [6]. PEGylation is another commonly used approach which can improve the solubility and biocompatibility of CNTs [6]. Some bioactive molecules, for example, Cys-DNA or short interfering RNA (siRNA), can be non-covalently bound to single-walled carbon nanotubes (SWCNTs) [6]. The effect of functionalization of carbon nanoparticles such as SWCNTs, MWCNTs, carbon nanofibers, carbon graphite, carbon black, on cell toxicity was also discussed in the same review [6]. In general, the toxic effect of CNTs depends on their concentration, dose, purity, and surface properties of the CNTs (surface area, topology, presence of functional groups, presence of residues and impurities). SWCNTs and other carbon-based materials (e.g., carbon black, active carbon, carbon nanofibers) seem to be more toxic than MWCNTs [6]. The functionalization of CNTs can increase their dispersibility in biological media and therefore diminish their toxicity, compared with non-functionalized ("pristine") CNTs. "Soluble" functionalized CNTs may either enhance or decrease the toxicity compared with "pristine" CNTs. Possible enhanced toxicity of functionalized DWCNTs (oxidized) has been recently reported by de Gabory et al. [7]. SWCNTS and MWCNTs can cause cellular toxicity (apoptosis and necrosis were observed) [6]. The toxic effect depends on concentration of

CNTs and time of incubation. It was concluded that the smaller the diameter of CNTs and the greater the aspect ratio, the greater toxic effect is observed [6]. In vivo oral, dermal, and pulmonary toxicity of CNTs were also described in this review [6].

Currently, little is known about interactions of nanoparticles with human lungs [8]. Engineered nanoparticles can cause pulmonary toxicity in humans, experimental animals and cell culture [8]. The increased ratio of surface area to mass of nanoparticles means that more potentially reactive groups present on their surface may interact with the lung cells and therefore impact their pulmonary toxicity [8]. Clearance of inhaled nanoparticles from the lungs depends mostly on particle size and surface charge [9]. Charged nanoparticles bind to lung tissues more effectively than non-charged ones [9]. Various surface coatings (albumin, lecithin, polysorbate 80, peptide attachments, PEG) were also found to influence nanoparticle cellular uptake and fate [9]. Nanoparticles associate with proteins or biopolymers present in human plasma or other biological fluids and form a dynamic protein corona [10]. A long-lived protein corona formed from human plasma was studied on 50 and 100 nm sized, unmodified, carboxyl- and amine-modified polystyrene nanoparticles (PS) [10]. Surprisingly, both size and surface properties of the nanoparticles had a very considerable role in determining the nanoparticle coronas on the different particles of the same materials [10]. The size and surface modification of nanoparticles can impact the reactivity of biologically active proteins in the corona and likely have biological consequences [10].

Toxicity of various engineered nanoparticles, including CNTs, fullerenes, dendrimers, QDs, gold nanoparticles, silica, in vivo and in vitro, was summarized by De Jong et al. [11]. Some of the nanoparticles can cause lung inflammation, progression of plaque formation in vivo and oxidative stress in vitro and in vivo. Studies in vitro revealed that some of nanoparticles can cause platelet aggregation, inhibition of macrophage phagocytosis, and affect function of mitochondria [11]. Nevertheless, surface modification of QDs with *N*-acetylcysteine resulted in reduction of toxicity, compared to non-modified quantum dots, which can cause oxidative stress to plasma membranes and cellular compartments [11]. The effect of PEGylation on gold nanorod cytotoxicity was also presented in the same review [11]. Composition of nanoparticle formulation may influence interactions of the nanoparticles with tissues or cells and their potential toxicity [11]. Functionalization of nanoparticles may make them more specific for drug delivery purposes, diminish their toxicity to non-targeted organs, change their therapeutic effect, and make them safer and more biocompatible [11]. Coating of nanoparticles may prevent their agglomeration. For example, the use of polymers, like PEG, poly(vinylpyrrolidone) (PVP), chitosan, dextran, or surfactants as coatings, can keep nanoparticles separated in colloidal suspensions [11]. Crossing of the blood–brain barrier (BBB) by nanoparticles is especially challenging in drug delivery. Nanoparticle surface charges can affect the BBB crossing and must be considered for the brain toxicity. Neutral nanoparticles were found to be not toxic for the BBB unlike cationic nanoparticles [11]. The toxicity of anionic nanoparticles was concentration dependent [11].

In vitro and in vivo toxicity studies on engineered gold nanoparticles (GNPs), including atomic clusters, colloidal particles, nanorods, gold nanoshells, and nanowires, were reviewed in great detail by Khlebtsov and Dykman [12]. Surface functionalization of GNPs can significantly influence their biodistribution in blood, liver, kidney, lung, spleen, muscle, brain, and bone [12]. Some of tested GNPs were coated with PEG, albumin, maltrodextrin, tumor necrosis factor (TNF), PEG-SH, IgG, maltose among many, and biodistribution studies on various animal models were summarized [12]. Interestingly, the adsorption of apolipoprotein on GNP surface may make possible for them to cross the BBB [12]. Also, cytotoxicity studies in vitro with cell cultures like human leukemia (K562), human breast carcinoma (SK-BR-3), human hepatocarcinoma (HepG2) and many more were presented in the same review [12]. The influence of a variety of surface coatings, such as DNA/transferrin,

triphenylphosphine, cyclodextrin, cysteine, folic acid polyvinylpyrrolidone, glycolipid on GNP cytotoxicity was summarized [12].

Chemical modification of nanoparticles can impact their cellular uptake and cellular localization. The functionalization can change the binding of biological molecules, antibodies, etc., to the surface of nanoparticles and may have impact on drug delivery. Polymer-based nanoscale sized materials such as polymer micelles, nanogels, liposomes, polymer-DNA complexes can be of use in pharmacology. Batrakova and Kabanov broadly reviewed the biological activity of Pluronic block copolymers consisting of PEO and poly(propylene oxide) (PPO) blocks [13]. The amphiphilic character of these copolymers allows the interactions with hydrophobic surfaces and biological membranes [13]. These copolymers can self-assemble into pluronic micelles in aqueous solutions [13]. The core of these micelles is made up of hydrophobic PPO blocks while the shell consists of hydrophilic PEO chains [13]. The PEO shell can not only provide increased solubility, but also prevent the unwanted interactions with cells or plasma components. Pluronics have an ability to incorporate into plasma membranes followed by translocation into the cells which has an effect on various cellular functions [13]. Interestingly, the composition of the block copolymers can affect their biological functions. Block copolymers with intermediate lengths of PPO block and quite hydrophobic structure (such as Pluronic P85 or Pluronic L61), were the most efficiently transported into the cells [13]. On the other hand, hydrophilic block copolymers did not incorporate into the lipid bilayers and were not transported into the cells [13].

Fluorescent nanocrystals, also called quantum dots (QDs), can be utilized in drug delivery or cellular imaging [14]. The core of the QDs is usually a semiconductor like cadmium-selenium (CdSe), cadmium-tellurium (CdTe), indium-phosphate (InP), or indium-arsenate (InAs). Zinc sulfide (ZnS) is commonly used as the shell to stabilize the core of the QDs [14]. The structure of the QDs allows of the coating of the QDs and conjugation of biologically active molecules, such as antibodies, peptide ligands, which may reduce their toxicity and increase bioactivity [14]. Toxicity of cadmium and core material was widely discussed in this review [14]. Encapsulation of QD core with ZnS can not only stabilize the core of QD, but also improve their optical properties and bioavailability [14]. Toxicity studies of various QDs in vitro and in vivo were summarized by Ghaderi et al. [14]. In most of studies, CdSe and CdFe were the cores and ZnS was the shell [14]. The majority of tested QDs were not coated and were found cytotoxic for some of cell lines [14]. The size of QD plays an essential role in the observed cytotoxicity. In general, the smaller size the higher toxicity [14]. In one of the studies, preincubation of cells with *N*-acetylcysteine and bovine serum albumine (BSA) considerably reduced the cell death caused by QDs [14]. In some studies, generation of ROS caused the QDs toxicity [14]. Results of QD toxicity studies on mice and rats were also presented in the review [14]. CdSe/ZnS were coated with PEG and BSA and the latter were shown to be distributed to the liver, spleen, and other tissues [14]. Little is currently known about crossing of QDs through the BBB [14].

Jain et al. reviewed dendrimers toxicity in detail [15]. Positively charged dendrimers can interact with negatively charged biological membranes and cause the membrane destruction via nanohole formation and further cell lysis. Cationic dendrimers were shown to be less cytotoxic than anionic or PEGylated dendrimers [15]. Jain et al. summarized the cytotoxicity of positively charged dendrimers, polypropyleneimine (PPI), and polyamido amine (PAMAM) against several cancer cell lines, including B16F10, CCRF, HepG2, COS-7, and Caco-2 [15]. In the same review, results of cytotoxicity and membrane disruption studies were presented for PPI dendrimers, PEG conjugated PPI dendrimers and PPI dendrimers with neutral acetamide groups on cultured Human Umbilical Vein Endothelial Cells (HUVEC) [15]. Moreover, hemolytic and hematological toxicity of dendrimers, including PPI and PAMAM, were presented [15]. The effect of various types of dendrimers, including glycine-, lactose-, mannose-, phenylalanine-coated PPI dendrimers on different blood parameters like

white and red blood cells, hemoglobin, hematocrit, were summarized in the review [15]. It was concluded that the cationic dendrimers might show significant destruction of the hematological parameters [15]. However, immunogenicity has not been yet confirmed for dendrimers [15]. Results of biocompatible and surface engineered dendrimers, such as polyether imine dendrimers, SN-38 complexed G4-PAMAM dendrimers, glycine-coated PPI dendrimers, phenylalanine-coated PPI dendrimers, lactose-coated PPI dendrimers, PPI dendrimers with peripheral neutral acetamide groups, carboxylic acid-terminated PAMAM dendrimers, anionic PAMAM dendrimers, lauroyl- and PEGylated PAMAM dendrimers, and PEGylated PPI dendrimers, on various cell lines were summarized [15]. It was concluded that modification of the surface of dendrimers can diminish their toxicity [15]. Coating of surface of dendrimers with glycine, lactose, mannose, and galactose, PEG can also improve their hemolytic toxicity profile [15]. Surface engineering makes dendrimers more biocompatible for clinical applications. It allows them to mask the cationic charge on their surface which can be achieved by PEGylation, acetylation, peptide conjugation, and make them neutral. Also, introducing of negative charge on the surface of dendrimers makes them more biologically compatible [15]. Chemical modification of the surface of dendrimers can also allow them to be used as drug carriers or DNA and gene conjugates [15].

So far little is known about the interactions of nanoparticles with the central nervous system [16]. Hu et al. summarized results of recent in vitro studies on various cell lines, including cultured neuronal phenotype (PC12), brain cultures of immortalized mouse microglia (BV2), rat dopaminergic neurons (NR7), primary cultures of embryonic rat (Sprague–Dawley) striatum, and mouse neural stem cells (NSCs) [16]. Zinc oxide (ZnO) and anionic magnetic nanoparticles, for example, were neurotoxic to NSCs and PC12 cells, respectively [16]. Results of in vivo studies showed that the presence of charge on nanoparticles can have an influence on their neurotoxicity [16]. Changing the physicochemical properties of nanoparticles by conjugation of, for example, wheat germ agglutinin or coating with polysorbate surfactants, can impact the BBB transport mechanism [16]. Coating of nanoparticles with surfactants (Tween) may influence brain toxicity [16]. Cationic nanoparticles (gold, polystyrene) were shown to cause hemolysis and blood clotting whilst anionic particles were usually non-toxic [16].

Nanomaterials when placed in contact with human or mammalian body fluids and tissues are recognized by the immune system. Modification of nanoparticle surface with hydrophilic agents can influence their potential toxicity in vivo and in vitro. Coating or wrapping of nanoparticles can modify their interaction with blood proteins and cells. Dobrovolskaia et al. reviewed on how surface modification of various nanoparticles can change plasma protein binding and influence the nanoparticles interactions with red blood cells and platelets [17].

Functionalized nanoparticles can be unrecognizable by the immune system or can enhance the immune responses. Zolnik et al. presented how manipulation of physicochemical properties of nanoparticles can change their interactions with the components of the immune system [18]. The interactions may be undesirable and cause immunostimulation (interaction of nanoparticles with the complement system, cytokine secretion, immunogenicity) or immunosuppression. Nanoparticles can be recognized as foreign material entering the body. This may lead to toxicity. If nanoparticles (like carbon nanotubes) are too long they cannot be digested by macrophages and can cause granuloma formation. Modification of the surface properties of nanoparticles by changing their hydrophobicity and/or surface charge, may have a positive influence on compatibility of nanoparticles with the immune system. Immunosuppression may lower the body's defense against infection or may improve the therapeutic benefits of treatments for allergies and autoimmune diseases [18].

Nanoparticles can bind to coagulation proteins, blood plasma components, and cell-surface proteins. Binding of immune system complement proteins activates the complement system and

results in strong binding of several complement proteins to the CNTs [2]. Such complement activation may influence subsequent interaction of the CNTs with cells and tissues and induce potential toxicity. Protein binding occurs when CNTs are placed in contact with human blood or lung fluids and can have biological consequences [2]. Interactions of various CNTs with human plasma proteins and pulmonary surfactant proteins were reported [2]. Pre-coating of CNTs with bovine serum albumin (BSA), human serum albumin (HSA), human fibrinogen (FBG) or the detergent Tween 20 did not prevent the complement activation [2]. Frequently the surface of CNTs is not completely covered by the coating molecules, so some of potential binding sites on the surface of CNTs stay unobstructed and available for binding of various complement proteins [2].

The major limitation of nanoparticles in drug delivery is their rapid uptake by the mononuclear phagocyte system (MPS). However, the presence of PEG chains on the surface of nanoparticles can significantly change their interactions with blood proteins. Prolonged circulation of nanoparticles in the bloodstream is crucial for drug delivery purposes (e.g., particles should not be cleared by the liver and spleen before reaching their target). Moghimi et al. presented protein-binding processes to poly(ethylene glycol)–phospholipid (PEG-PL) engineered nanoparticles and the influence of surface PEGylation on complement activation and the fate of these engineered nanoparticles [19]. The effect of PEGylated liposome composition and size on complement activation in human serum was presented [19]. Anionic liposomes were shown to activate complement more than the methylated conjugates without charge [19].

Stable dispersion of nanoparticles in biological media is a necessary step toward their biomedical applications. It can be achieved by functionalization of CNT's hydrophobic surface with various chemical groups or adsorption of biologically active molecules. Although coating of CNTs with PEG_{5000}-phospholipid conjugates improves their stability in a biological buffer, the PEG-PL-coated CNTs can still activate the human complement system via the lectin pathway [19]. Complement activation by SWNTs in human serum with addition to *N*-acetylglucosamine, D-galactose, non-specific antibody, and Anti-MASP2 antibody in C1q-depleted serum was investigated by Moghimi et al. [19]. Interestingly, PEG-phospholipid conjugates which form stable micellar structures with the size about 30 nm did not activate complement [19]. PEGylated liposomes with the size from 100 to 120 nm activated complement via classical and alternative pathway [19]. The lectin pathway was shown to be activated by 2–5 nm diameter CNTs with a length of 250 nm [19].

Effect of surface PEGylation of various liposomes on complement activation was presented by Moghimi et al. [20]. Anionic liposomes were shown to activate complement in vitro. However, incorporation of phosphatidylethanolamine-methoxypoly(ethylene glycol) (PE-mPEG) into the liposomal bilayer abolished the complement activation and was shown to be liposome-concentration dependent [20]. The drug Doxil®, commonly used in cancer therapy, is a negatively charged PEGylated liposome with doxorubicin enclosed. The drug is a strong activator of the human complement system and even its surface modification with PEG does not suppress complement activation [20]. However, the length of PEG can be an important factor in this as shown in some in vivo studies. Longer PEG chains were much more effective in preventing of complement activation than the shorter chains [20]. Other nanoparticles, very similar in structure to Doxil®, like hydrogenated soy phosphatidylcholine (HSPC) and cholesterol with the size of 100 nm, were shown not to activate the complement system in vitro. However, the presence of PEG_{2000}-conjugated distearoyl phosphatidylethanolamine, can cause significant complement activation in vitro [20]. Also, introduction of hydrogenated soy phosphatidylglycerol causes significant complement activation in vitro measured by an increase of SC5b-9 levels in human serum [20]. The negative charge plays a crucial role in the activation and the surface density of methoxypoly(ethylene glycol), mPEG, cannot prevent complement activation by liposomes as they cannot mask the negatively charged phosphodiester moieties [20].

Assessment of nanoparticles toxicity in vivo is a necessary step if the nanomaterials are to be used in biomedicine as efficient drugs or gene carriers. Functionalization of their surface by changing the surface charge and chemical composition can have a considerable impact on their toxic effect in vivo. Some nanomaterials can initiate oxidative stress in cells and release of ROS. This can initiate inflammation processes and cellular damage. Moreover, organs like the liver, spleen, kidney, and lungs can be affected by the oxidative stress [21]. Nanomaterials can also interact with mitochondria and nucleus, and such interactions are thought to be the main source of their toxicity [21]. Primary causes of various nanomaterials toxicity in vivo were summarized by Aillon et al. [21]. The authors discussed the influence of physicochemical properties of dendrimers, CNTs, QDs, and gold nanoparticles on their toxicity [21]. Examination of nanoparticle toxicity in vivo is still very limited. For most nanoparticles, their nano-size can cause their toxicity. Additionally, the surface charge on dendrimers, the length of CNTs, and the shape of gold nanoparticles can enhance their toxicity in vivo. Functionalization of the surface of the nanomaterials can diminish their toxicity. Chemical modification of the dendrimers surface can mask their charge, CNT hydrophobic surface can become more hydrophilic, and they can also become shorter by cutting using sonication. Hydrophobicity of "pristine" CNTs is a main obstacle in their biomedical applications. Moreover, they have a tendency to aggregate which can cause slow clearance from the body and oxidative stress in organs like liver, spleen, or lungs. Their length causes inefficient phagocytosis and damage of macrophages [21].

Studies on mice revealed that PEG-polyester dendritic hybrids did not cause any pathological changes in the liver, kidney, heart, lungs, and intestine of animals [21]. Similarly, PEGylated melanine dendrimer administration into mice did not cause liver and kidney toxicity [21].

Cadmium and selenium, heavy metals mostly used in QD cores, can be toxic in vivo [21]. Cadmium not only can accumulate in human tissues but also cross the BBB. However, electrostatic charge modification, adsorption, covalent modification, or chelation can have a considerable impact on the QD biocompatibility in vivo. As mentioned earlier, ZnS is commonly used as stabilizer of the QD surface, but also secondary coating of the QD can be applied to diminish its toxicity [21]. However, there is still a debate between researchers if the secondary coating of QD can prevent the cadmium leakage from their core [21]. Stability of the QD complex which consists of core, shell, and secondary coating is a very important factor in their toxicity [21].

Various covalent and non-covalent modifications of the surface of nanoparticles are essential approaches that allow to overcome the toxicity problems associated with nanomaterials. Surface functionalization gives an excellent opportunity to design novel nanoparticles with desirable physicochemical properties and biological activity. Nevertheless, it is essential that the functionalization is not damaging for cells or tissues by itself. Ideally, the functionalization should diminish nanoparticles toxicity and make them more biocompatible. This can significantly enhance their chances to be used in pharmacology and biomedical fields. The modification of surface properties of the nanoparticles can change their reactivity and have considerable impact on their cytotoxicity, ROS generation, immunocytotoxicity, pulmonary toxicity, and neurotoxicity. It can also influence their interactions with human and mammalian blood proteins and their adsorption on the surface. However, as the interactions with biological environments are quite complex, achieving the desirable biological or therapeutical responses of nanoparticles may be time consuming and not always work out. Sometimes quite opposite outcomes can be observed and the functionalized nanoparticles can be even more toxic either in vivo or in vitro than non-functionalized ones [2, 5, 19, 20]. Nevertheless, functionalization of nanoparticles surface can be considered as a useful tool for researchers and may reduce nanoparticles toxicity by making them more suitable for various biomedical applications.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Functionalization of Carbon Nanotubes
- ▶ Gold Nanorods
- ▶ In Vivo Toxicity of Carbon Nanotubes
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Nanoparticle Cytotoxicity
- ▶ Nanoparticles
- ▶ Quantum Dot Toxicity

References

1. Medina, C., Santos-Martinez, M.J., Radomski, A., Corrigan, O.I.: Nanoparticles: pharmacological and toxicological significance. Br. J. Pharmacol. **150**, 552–558 (2007)
2. Rybak-Smith, M.J., Pondman, K., Salvador-Morales, C., Flahaut, E., Sim, R.B.: Recognition of carbon nanotubes by the human innate immune system. In: Klingeler, R., Sim, R.B. (eds.) Multi-Functional Carbon Nanotubes for Biomedical Applications, pp. 183–210. Springer, Heidelberg (2011)
3. Smart, S.K., Cassady, A.I., Lu, G.Q., Martin, D.J.: The biocompatibility of carbon nanotubes. Carbon **44**, 1034–1047 (2006)
4. Yu, Y., Zhang, Q., Mu, Q., Zhang, B., Yan, B.: Exploring the immunotoxicity of carbon nanotubes. Nanoscale Res. Lett. **3**, 271–277 (2008)
5. Magrez, A., Kasas, S., Salicio, V., Pasquier, N., Seo, J. W., Celio, M., Catsicas, S., Schwaller, B., Forró, L.: Cellular toxicity of carbon-based nanomaterials. Nano Lett. **6**, 1121–1125 (2006)
6. Foldvari, M., Bagonluri, M.: Carbon nanotubes as functional excipients for nanomedicines: II. Drug delivery and biocompatibility issues. Nanomed. Nanotechnol. **4**, 183–200 (2008)
7. de Gabory, L., Bareille, R., Daculsi, R., L'Azou, B., Flahaut, E., Bordenave, L.: Carbon nanotubes have a deleterious effect on the nose: the first in vitro data. Rhinology **49**(4), 445–452 (2011)
8. Card, J.W., Zeldin, D.C., Bonner, J.C., Nestmann, E. R.: Pulmonary applications and toxicity of engineered nanoparticles. Am. J. Physiol. Lung Cell. Mol. Physiol. **295**, L400–L411 (2008)
9. Yang, W., Peters, J.I., Williams III, R.O.: Inhaled nanoparticles – a current review. Int. J. Pharm. **356**, 239–247 (2008)
10. Lundqvist, M., Stigler, J., Elia, G., Lynch, I., Cedervall, T., Dawson, K.A.: Nanoparticle size and surface properties determine the protein corona with possible implications for biological impacts. Proc. Natl. Acad. Sci. U. S. A. **105**(18), 14265–14270 (2008)
11. De Jong, W.H., Borm, P.J.A.: Drug delivery and nanoparticles: applications and hazards. Int. J. Nanomed. **3**, 133–149 (2008)
12. Khlebtsov, N., Dykman, L.: Biodistribution and toxicity of engineered gold nanoparticles: a review of in vitro and in vivo studies. Chem. Soc. Rev. (2010). doi:10.1039/c0cs00018c
13. Batrakova, E., Kabanov, A.V.: Pluronic block copolymers: evolution of drug delivery concept from inert nanocarriers to biological response modifiers. J. Control. Release **130**, 98–106 (2008)
14. Ghaderi, S., Ramesh, B., Seifalian, A.M.: Fluorescence nanoparticles "quantum dots" as drug delivery system and their toxicity: a review. J. Drug Target. (2010). doi:10.3109/1061186X.2010.526227
15. Jain, K., Kesharwani, P., Gupta, U., Jain, N.K.: Dendrimer toxicity: let's meet the challenge. Int. J. Pharm. **394**, 122–142 (2010)
16. Hu, Y.-L., Gao, J.-Q.: Potential neurotoxicity of nanoparticles. Int. J. Pharm. **394**, 115–121 (2010)
17. Dobrovolskaia, M.A., McNeil, S.E.: Immunological properties of engineered nanomaterials. Nat. Nanotechnol. **2**, 469–478 (2007)
18. Zolnik, B.S., Gonzàlez-Fernàndez, Á., Sadrieh, N., Dobrovolskaia, M.A.: Minireview: nanoparticles and the immune system. Endocrinology **151**(2), 458–465 (2010)
19. Moghimi, S.M., Andersen, A.J., Hashemi, S.H., Lettiero, B., Ahmadvand, D., Hunter, A.C., Andersen, T.L., Hamad, I., Szebeni, J.: Complement activation cascade triggered by PEG-PL engineered nanomedicines and carbon nanotubes: The challenges ahead. J. Control. Release **146**, 175–181 (2010)
20. Moghimi, S.M., Szebeni, J.: Stealth liposomes and long circulating nanoparticles: critical issues in pharmacokinetics, opsonization and protein-binding properties. Prog. Lipid Res. **42**, 463–478 (2003)
21. Aillon, K.L., Xie, Y., El-Gendy, N., Berklan, C.J., Forrest, M.D.: Effects of nanomaterial physicochemical properties on in vivo toxicity. Adv. Drug. Deliver. Rev. **61**, 457–466 (2009)

Effective Media

- ▶ Theory of Optical Metamaterials

Elastic and Plastic Behavior of Materials with Nanometric Dimensions

- ▶ Mechanical Properties of Nanostructures

Elastic Modulus Tester

► Nanoindentation

Elasto-capillary Folding

► Capillary Origami

Electric Cell-Substrate Impedance Sensing (ECIS)

► Impedance Sensing Techniques for Biomedical Microdevices

Electric Double Layer Capacitor

► Nanomaterials for Electrical Energy Storage Devices

Electric Field–Directed Assembly of Bioderivatized Nanoparticles

Youngjun Song[1] and Michael J. Heller[2,3]
[1]Department of Electrical and Computer Engineering, University of California San Diego, San Diego, CA, USA
[2]Department of Nanoengineering, University of California San Diego, La Jolla, CA, USA
[3]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA

Synonyms

Bottom-up nanofabrication; Directed self-assembly; Nanoparticle self-assembly

Definition

Use of DC electric field microarray devices and electronic techniques to carry out the directed self-assembly of bioderivatized nanoparticles into high order structures.

Overview

Electronic microarray devices which produce reconfigurable electric field patterns can be used to control the self-assembly of nanocomponents into higher order structures. These devices allow the hierarchical assembly of nanocomponents to be carried within microscale geometries of electronic circuitry. Such nanofabrication devices and technologies have the potential for producing highly integrated nano/micro/macrostructures including nanophotonic-based large-scale arrays and displays; high-efficiency multiple band gap photovoltaics; fuel cells with nanostructured intermetallic catalysts and integrated fluidic channeling; smart-morphing nanocomposite materials; in vivo biosensor and mother-ship drug delivery devices. Electronic microarray devices that were originally developed for molecular diagnostic applications allow DNA, proteins, and almost any charged nanostructure to be rapidly transported and specifically bound at any site on the array surface. Microarray devices that have onboard integrated CMOS circuitry also allow rapid switching and precise control of the currents/voltages sourced to the microarray. Such microarray devices have been used to carry out the rapid highly parallel assisted self-assembly of biotin- and streptavidin-derivatized fluorescent nanoparticles and DNA-derivatized nanoparticles into multilayer structures. Electric field–directed assembly has major advantages over more classical passive layer-by-layer (LBL) self-assembly processes, which are generally one order of magnitude slower, nonparallel, and not reconfigurable in real time. The use of electronic microarray devices for assisted self-assembly represents a unique synergy of combining the best aspects of "top-down" and "bottom-up" technologies into a potentially viable nanomanufacturing process.

It also enables a more bio-inspired logic to be followed for the hierarchal self-assembly of nanostructures into higher order structures, materials, and devices.

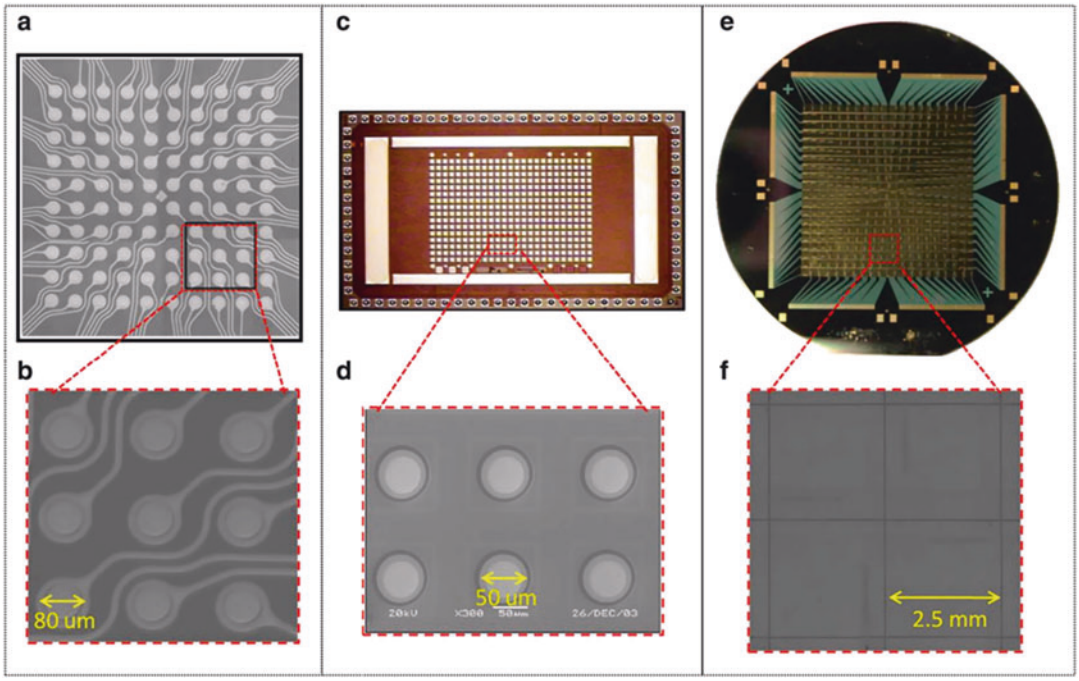
Nanotechnology Top-Down and Bottom-Up Processes

Nanotechnology includes a wide range of new ideas and concepts that are likely to enable novel applications in electronics, photonics, materials, energy conversion processes, biosensors, and biomedical devices. Some of the many challenges and opportunities in nanotechnology were discussed in the early National Nanotechnology Initiative meeting [1, 2]. One of the more important challenges is the area of nanofabrication, in particular the development of technologies which will lead to cost-effective nanomanufacturing processes. Enormous efforts are now being carried out on refining the more classical "*top-down*" photolithography processes to produce silicon-integrated (CMOS) electronic devices with nanometer scale features. While this goal has been achieved, the process requires billion dollar fabrication facilities and it does appear to be reaching some fundamental limits. So-called *bottom-up* self-assembly processes are also being studied and developed as possible new ways for producing nanodevices and nanomaterials. While there are now numerous examples of promising nanocomponents such as organic electron transfer molecules, quantum dots, carbon nanotubes, and nanowires, their assembly into higher order structures has been limited [3, 4]. Thus, the development of a viable cost-effective bottom-up self-assembly nanofabrication process that allows billions of nanocomponents to be assembled into a higher order structure still remains a considerable challenge.

Living System and DNA Self-assembly – Living systems represent the best example of self-assembly processes that can be used for developing strategies for bottom-up nanofabrication. Living systems contain many types of biomolecules which have high fidelity recognition properties including deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. Proteins serve as structural elements (collagen, keratin, silk, etc.), as binding recognition entities (antibodies) and as efficient catalytic macromolecules (enzymes). The nucleic acids DNA and RNA are involved in the storage and transfer of genetic information. All these biomolecules are able to interact and organize into higher order nanostructures which translate genetic information and perform biomolecular syntheses and energy conversion metabolic processes. Ultimately, all these biomolecules and bionanostructures are integrated and organized within membrane structures to form the entities called cells. Cells in turn can then replicate and differentiate (via these nanoscale processes) to form and maintain living organisms. Thus, biological systems have developed the ultimate "bottom-up" nanofabrication processes that allow component biomolecules and nanostructures with intrinsic self-assembly and catalytic properties to be organized into highly intricate living organisms. Of all the different biomolecules that could be useful for nanofabrication, the nucleic acids (DNA and RNA) with their high fidelity recognition and intrinsic self-assembly properties represent a most promising material for creating nanoelectronic, nanophotonic as well as many other types of organized nanostructures [5–7]. DNA, RNA, and other synthetic DNA analogues are programmable molecules which through their base sequence have intrinsic molecular recognition and self-assembly properties. Short DNA sequences, called oligonucleotides, are readily synthesized by automated techniques and can be modified with a variety of functional groups such as amines, biotin moieties, fluorescent or chromophore groups, and with charge transfer molecules. Additionally, synthetic DNA molecules can be attached to quantum dots, metallic nanoparticles, carbon nanotubes; as well as to surfaces like glass, silicon, gold, and semiconductor materials. Synthetic DNA molecules (oligonucleotides) represent an ideal type of "Molecular Legos" for the self-assembly of nanocomponents into more complex two- and three-dimensional higher order structures. At a first level, DNA sequences can be used as a kind of template for assembly on solid surfaces. The technique involves taking complementary DNA sequences and using them as a kind of selective glue to bind other DNA-modified macromolecules

or nanostructures together. The base pairing property of DNA allows one single strand of DNA with a unique base sequence to recognize and bind together with its complementary DNA strand to form a stable double-stranded DNA structures. While high fidelity recognition molecules like DNA allows one to "self-assemble" higher order structures, the process has some significant limitations. First, for in vitro applications (i.e., in a test tube), DNA and other high fidelity recognition molecules like antibodies, streptavidins, and lectins work most efficiently when the complexity of the system is relatively low. That is as the complexity of the system increases (more unrelated DNA sequences, proteins, and other biomolecules), the high fidelity recognition properties of DNA molecules are overcome by nonspecific binding and other entropy-related factors, and the specific hybridization efficiency for the DNA molecules is considerably reduced. Under in vivo conditions (inside living cells), the binding interactions of high fidelity recognition molecules like DNA are much more controlled and compartmentalized, and the DNA hybridization process is assisted by structural protein elements and active dynamic enzyme molecules. Thus, new bottom-up nanofabrication processes based on self-assembly using high fidelity recognition molecules like DNA should also incorporate strategies for directing and controlling the overall process.

Electronic Microarray DNA Hybridization Technology – A variety of electronic microarrays have been designed and fabricated for DNA genotyping diagnostic applications [8–11]. An early 100 test-site electronic microarray which has been commercialized (Nanogen) has an inner set of 80-μ diameter platinum microelectrodes (Fig. 1a, b). Each microelectrode has an individual wire interconnect through which current and voltage are applied and regulated. The microarrays are fabricated from silicon wafers, with insulating layers of silicon dioxide, platinum microelectrodes, and gold connecting wires. Silicon dioxide/silicon nitride is used to cover and insulate the conducting wires, but not the surface of the platinum microelectrodes. The whole surface of the microarray is covered with several microns of hydrogel (agarose or polyacrylamide) which forms a permeation layer. The permeation layer is impregnated with streptavidin which allows attachment of biotinylated DNA probes or other entities. The ability to use silicon and microlithography for fabrication of the DNA microarray allows a wide variety of devices to be designed. More sophisticated higher density electronic microarrays with 400 test-sites have on-chip CMOS control elements for regulating the current and voltages to the microelectrode at each test-site (Fig. 1c, d). These control elements are located in the underlying silicon structure and are not exposed to the aqueous samples that are applied to the chip surface when carrying out the DNA hybridization reactions. DNA hybridization reactions are carried out on these electronic microarrays by first applying a positive DC bias of ~2–4 V to address or spot specific DNA molecules to the selected test-sites on the array. These DNA molecules are usually oligonucleotide "capture" probes or target DNA sequences that have been functionalized with biotin molecules. These biotinylated DNA molecules become strongly bound to streptavidin molecules which are cross-linked to the hydrogel layer covering the underlying microelectrode. In the next step, a positive DC electric field is applied to specific microelectrodes directing the hybridization of the other DNA molecules to the DNA sequences attached to the selected test-sites. (During the addressing and hybridization processes, other microelectrodes on the microarray are biased negative.) Electronic addressing also allows DNA probes to be spotted onto the array in a highly reproducible manner. Electronic microarrays can be formatted in a variety of ways that include reverse dot blot format (capture/identity sequences bound to test-sites), sandwich format (capture sequences bound to test-sites), and dot blot format (target sequences bound to test-sites). DNA hybridization assays involve the use of fluorescent DNA reporter probes designed to hybridize to specific target DNA sequences. The reporter groups are usually organic fluorophores that have been either attached to oligonucleotide probes or to the target/sample DNA/RNA sequences. After electronic addressing and hybridization are carried out, the microarray is analyzed using the



Electric Field–Directed Assembly of Bioderivatized Nanoparticles, Fig. 1 *Three different array devices for electric field–directed nanoparticle assembly.* (**a**) View of a 100 test-site microarray device with 80-μ diameter platinum microelectrodes, the device is approximately 5 × 5 mm in size. (**b**) A close-up of the 80 μ microelectrodes. (**c**) A 400 test-site CMOS electronic microarray device with 50-μ diameter platinum microelectrodes and four perimeter counterelectrodes, the device is approximately 5 × 7 mm in size. (**d**) A close-up of the 50-μ diameter microelectrodes. (**e**) A 4″ silicon wafer size electric field–directed assembler device. The total active deposition area is 5 cm^2 with 200 platinum electrodes which are 2.5 × 2.5 mm with 8 μm gaps. (**f**) Shows a close-up of four of the electrodes

fluorescent detection system. Reference [12] provides a good example of single nucleotide polymorphism (SNP) genotyping analysis that has been carried out on using electronic microarrays.

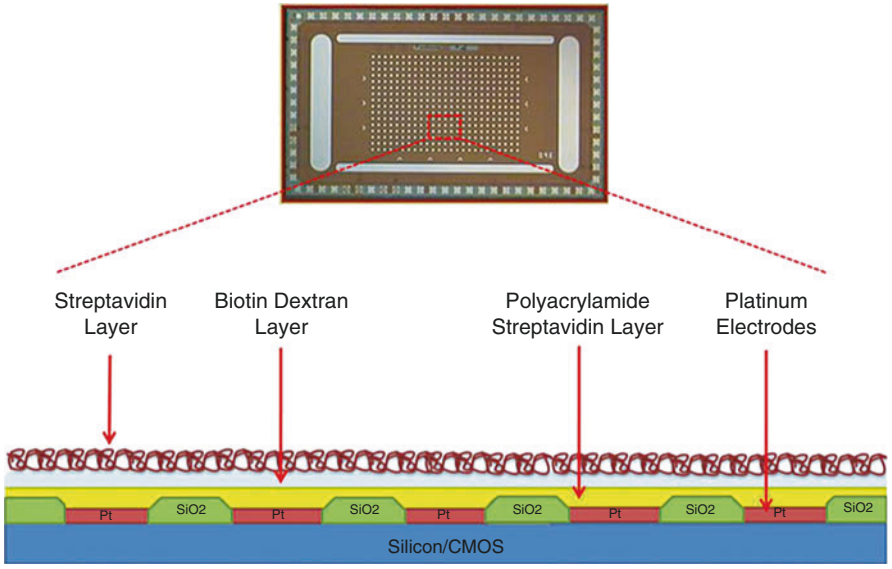
Electric Field–Directed Nanofabrication

While electronic microarray devices were originally developed for DNA genotyping diagnostic applications, the resulting electric fields are able to transport any type of charged molecule or structure including proteins, antibodies, enzymes, nanostructures, cells, or micron-scale devices to or from any of the sites on the array surface. In principle, these active devices serve as a "motherboard" or "host-board" for the assisted assembly of DNA molecules or other entities into higher order or more complex structures. Since the DNA molecules have intrinsic programmable self-assembly properties, and can be derivatized with electronic or photonic groups, or attached to larger nanostructures (quantum dots, metallic nanoparticles, nanotubes), or microstructures, this provides the basis for a unique bottom-up nanofabrication process. Thus, electronic microarray devices can serve as motherboards that allow one to carry out a highly parallel electric field "pick and place" process for the heterogeneous integration of molecular, nanoscale, and micron-scale components into complex three-dimensional structures. Electric field–directed self-assembly technology is based on three key physical principles: (1) the use of functionalized DNA or other high fidelity recognition components as "Molecular Lego" blocks for nanofabrication; (2) the use of DNA or other high fidelity recognition components as a

"selective glue" that provides intrinsic self-assembly properties to other molecular, nanoscale, or micron-scale components (metallic nanoparticles, quantum dots, carbon nanotubes, organic molecular electronic switches, micron and submicron silicon lift-off devices and components); and (3) the use of an active microelectronic array devices to provide electric field assistance or control to the intrinsic self-assembly of any modified electronic/photonic components and structures [13, 14].

Electric Field–Directed Self-assembly of Bioderivatized Nanoparticles – In addition to the more classical top-down processes such as photolithography, so-called bottom-up processes are being developed for carrying out self-assembly of nanostructures into higher order structures, materials, and devices [15]. To this end, considerable efforts have been carried out on both passive and active types of layer-by-layer (LBL) self-assembly processes as a way to make three-dimensional layered structures, which can have macroscopic x-y dimensions [16]. In cases where patterned structures are desired, the substrate material is generally pre-patterned using masking and a photolithographic process [17]. Other approaches to patterning include the use of optically patterned ITO films and active deposition of the nanoparticles. Nevertheless, limitations of passive LBL as well as active assembly processes provide considerable incentive to continue the development of better paradigms for nanofabrication and heterogeneous integration. Of particular importance to self-assembly processes involving biomolecular-derivatized nanoparticles is the prevention of nonspecific binding. Nonspecific binding due to cooperative hydrophobic, hydrogen bonding and ionic interactions between nanoparticles and the assembly substrate may often be as strong as the specific ligand binding interactions. Thus, once a nanoparticle has been bound nonspecifically, it is often difficult to remove it without perturbing or damaging the more specific nanoparticle interactions and structures. Nonspecific binding is often a concentration-dependent problem, whose probability increases when large amounts of derivatized nanoparticles interact with a physically and/or chemically complex substrate. Alternatively, trying to carry out stringent self-assembly at very low concentrations of biomolecular-derivatized nanoparticles is a slow process, and may still result in nonspecific binding to unprotected areas of the substrate. Therefore, there is a clear need for a rapid, directed assembly process which utilizes low concentrations of nanocomponents and reduces nonspecific interactions throughout the repeated multistep processes that will be required for fabricating complex higher order structures.

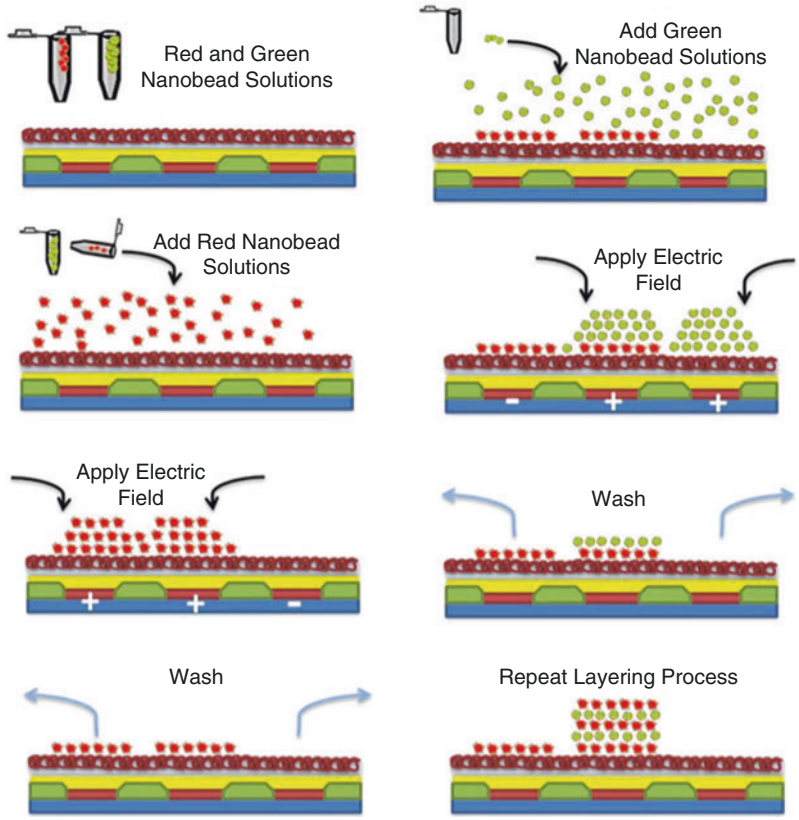
Electronic Microarrays and Techniques – Electronic microarrays and electric field techniques have now been used to carry out the highly parallel directed self-assembly of biomolecular-derivatized nanoparticles into multilayered higher order structures. The rapid and highly parallel assembly of biotin-, streptavidin-, and DNA-derivatized nanoparticles into structures with over 40 layers has been achieved [18–20]. The optimized electric field process allows nanoparticles at very low concentration in the bulk solution to be rapidly concentrated and specifically bound to the activated sites with minimal nonspecific binding. The electric field device used to obtain such results was an ACV 400 CMOS electronic microarray (Nanogen). This CMOS-controlled electronic microarray contains 400 50-μm diameter platinum microelectrodes in a 16 × 25 grid arrangement (Fig. 1c). The device has four additional large rectangular counterelectrodes on its perimeter. Each of the 400 microelectrodes is capable of independently producing from 0 to 5 V at up to 1 μA per microelectrode. The 400 microelectrodes on the array can be biased positive, negative, or left neutral in basically any combination; however, all four of the perimeter electrodes must be biased at the same polarity (positive or negative) or left neutral. The electronic microarray device can be set to source either current or voltage, and the activation time for each microelectrode can be controlled independently. The microarray device uses platinum electrodes as they present an inert interface to the solution. Additionally, the platinum electrodes do not degrade or disintegrate from the electrolysis reactions which occur at the voltages (>1.2 V) necessary to produce rapid electrophoretic



Electric Field–Directed Assembly of Bioderivatized Nanoparticles, Fig. 2 *The 400 test-site CMOS electronic microarray device*. The 400 test-site device, with a cross-sectional diagram showing the silicon base, the CMOS control circuitry, the platinum microelectrodes, the silicon dioxide insulating structures between the microelectrodes, the overlaying polyacrylamide/streptavidin permeation gel layer, the biotin–dextran layer, and the final layer of streptavidin (the drawings are not to scale)

transport of nanoparticles in the bulk solution. The ACV 400 site microarray comes coated with a 10-μm thick (hydrated) porous permeation layer composed of polyacrylamide impregnated with streptavidin (Fig. 2). This permeation layer is designed to prevent analyte molecules (DNA, proteins, etc.) and nanoparticles from contacting the microelectrodes, and to ameliorate the adverse effects of the electrolysis products on the binding reactions. In order to improve nanoparticle binding density, the microarray surface was overlayered with a biotin–dextran layer and a high density streptavidin layer (Fig. 2). The basic scheme for the electric field–directed self-assembly of biotin and streptavidin nanoparticles is shown in Fig. 3, and a full description of the actual procedure is given in Ref. [19].

Procedure for Deposition of First Nanoparticle Layers – The microarray is first washed with a 100 mM solution of L-histidine to remove any excess streptavidin. The bulk of the liquid on the chip is then removed, and 20 μL of a 4 nM solution of biotinylated 40 nm red fluorescent nanoparticles in 100 mM L-histidine buffer is deposited onto the microarray surface. Because most of the derivatized nanoparticles used in the experiments have a net negative charge, microelectrodes used for deposition are biased positive, with the outer perimeter counterelectrodes being biased negative. A number of microelectrodes on the array are often left unbiased (not activated) and served as overall controls for nonspecific binding. Immediately after addition of the nanoparticle containing solution, the directed electrophoretic transport and deposition of nanoparticles to the specific array sites is carried out. In initial experiments, the microarray can be set up in a combinatorial fashion to test a variety of currents, voltages, and activation times to determine optimal deposition conditions. Generally, a current ramp from 0 to 0.4 μA in 0.025 μA increments is programmed along rows of microelectrodes on the array. Activation times from 0 to 20 s in 5 s intervals are programmed across groups of columns on the microarray. Several columns of microelectrodes are also left inactivated to serve as controls. Following directed electrophoretic transport of the biotinylated 40 nm nanoparticles, the microarray is washed six times with 100 mM L-histidine, and then all liquid is removed from the



Electric Field–Directed Assembly of Bioderivatized Nanoparticles, Fig. 3 *The electric field–directed nanoparticle layering process*. The initial step involves adding a solution containing the biotinylated 40 nm *red* fluorescent nanoparticles onto the microarray device. The next step involves applying a positive DC biased to the desired test-sites on the microarray which attracts and concentrates the negatively charged biotinylated nanoparticles to the surface. The biotinylated nanoparticles become bound to the streptavidin layer, and the unattached nanoparticles are then washed away. In the next step, a solution containing streptavidin 40 nm *yellow-green* fluorescent nanoparticles is placed onto the microarray; the positively biased test-sites on the microarray attract and concentrate the negatively charged streptavidin nanoparticles which now become bound to the biotinylated nanoparticle monolayer. The unattached nanoparticles are then washed away. The alternate addressing, binding, and washing of biotin and streptavidin nanoparticles are repeated to form the multilayer nanoparticles structures

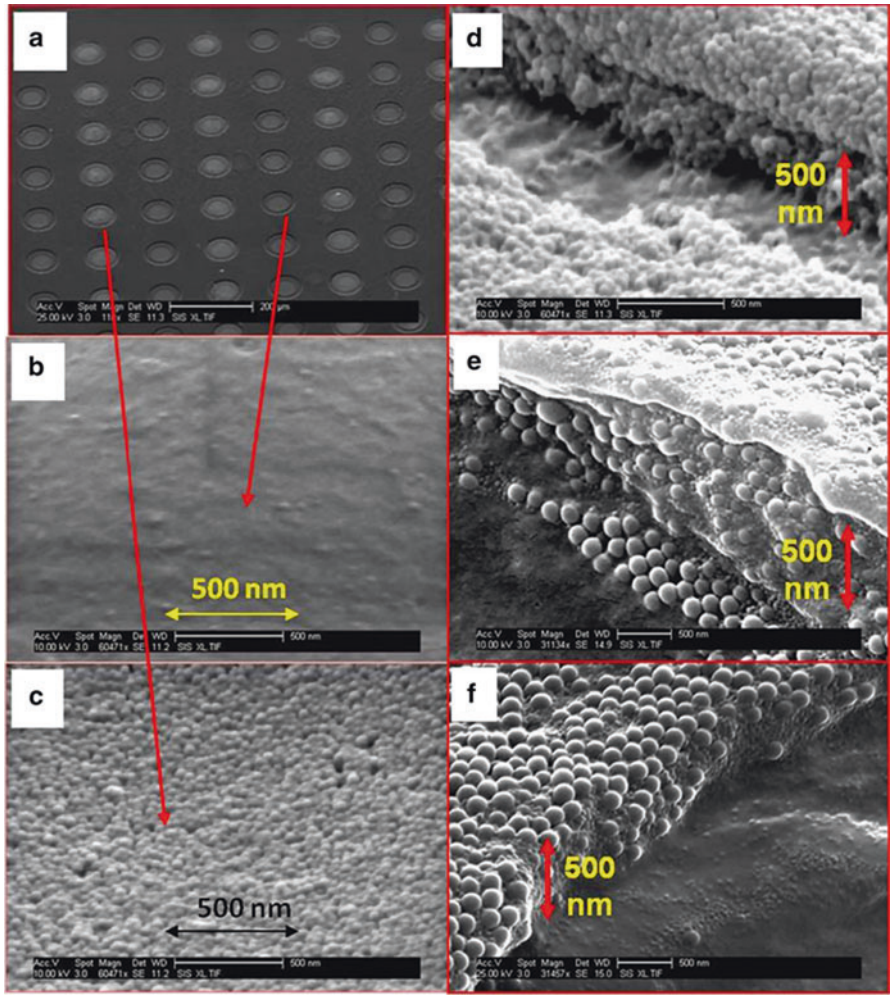
microarray to prevent crystallization of the histidine upon drying [19].

Procedure Multilayer Nanoparticle Depositions – The basic combinatorial directed deposition procedure is repeated in order to determine optimal conditions (current, voltage, and activation time) for producing best quality alternating layers of biotin and streptavidin nanoparticles. Template experiments can be used for determining nanoparticle layering conditions (which could be run on the same microarray in highly parallel manner). The first template experiment involves alternating the deposition of biotin 40 nm red fluorescent nanoparticles and streptavidin 40 nm yellow-green fluorescent nanoparticles in a grid type format. In this format, groups of columns of microelectrodes are set to different activation times (typically between 0 and 30 s), and the rows of microelectrodes are set to different current levels (0.025–0.4 μA in 0.0.025 μA increments). Several repeats of the combinatorial pattern are done across the whole microarray, and intervening columns of microelectrodes are left inactivated to serve as controls. This combinatorial approach allows for a relatively quick determination of the optimal time and current conditions necessary to achieve the best

deposition and layering. A second type of template experiment is used to determine the level of self-adhesion (nonspecific binding) between similar nanoparticles (biotin to biotin and streptavidin to streptavidin). This is carried out by having regions on the microarray only activated on alternate deposition cycles, when the same type of nanoparticle solution is exposed to the microarray. This experiment tests if the same type of nanoparticle would accumulate without the presence of its specific ligand binding partner. In these experiments, a single column per template is activated during every deposition to serve as a positive control for biotin-to-biotin nanoparticle binding and streptavidin-to-streptavidin nanoparticle binding. Additional columns of microelectrodes are not activated throughout the experiment, and serve as overall controls for more general nonspecific binding to the microarray. All of these experiments are monitored in real time and imaged by an epifluorescent microscope set at 540 nm emission for the yellow-green fluorescent nanoparticles and 610 nm emission for the red fluorescent nanoparticles. Imaging was carried out using a Hamamatsu Orca-ER camera [19].

Multilayer Biotin–Streptavidin Nanoparticle Structures – One of the main advantages of the electric field–directed self-assembly fabrication process is the ability to direct the streptavidin–biotin ligand binding to only occur on the specific microelectrodes which have been positively biased. Figure 4a shows a section of the 400 site CMOS microarray device where alternate rows of microelectrodes were used to carry out biotin–streptavidin nanoparticle layering. Rows with bright circular spots had 40 layers of alternating 40 nm red fluorescent biotin nanoparticles and 40 nm yellow-green fluorescent streptavidin nanoparticles deposited by applying a DC electric field. Rows with lighter circular spots had no electric field applied. Figure 4b shows a scanning electron microscope (SEM) image of a microelectrode which had no DC field applied. It can clearly be seen that no biotin nanoparticles have bound to this surface, despite the fact that it contains a streptavidin binding layer and has been exposed to the biotin nanoparticles 20 times. Figure 4c shows an SEM image of a microelectrode to which the DC electric field was applied. Nanoparticles have clearly been deposited on this surface. Multilayer nanostructures were verified by using SEM to image fractured cross sections of the 40 layer structures of alternating 40 nm red fluorescent biotin nanoparticles and 40 nm green fluorescent streptavidin nanoparticles (Fig. 4d). In other work, multilayer layers structures containing alternating 40 nm biotin nanoparticles and 200 nm streptavidin nanoparticles were fabricated. Figure 4e, f show the SEM image of cross sections of 40 layer structure of alternating 40 nm biotin nanoparticles and 200 nm streptavidin nanoparticles.

Multilayer DNA Nanoparticle Structures – In other work, electric field–directed hybridization was used to produce 20 layer nanostructures composed of DNA-derivatized nanoparticles [20]. Using the same CMOS electronic microarray device, DNA nanoparticles could be directed and concentrated such that rapid and specific hybridization occurs only on the activated sites. Nanoparticles layers were formed within 30 s of activation, and 20 layer structures were completed in under an hour. The results of this work carry significant implications for the future use of DNA nanoparticles for directed self-assembly nanofabrication of higher order structures. Advantages for using electric field directed fabrication of DNA nanoparticles include the following: First, the electric field process for the directed hybridization of DNA nanoparticles was significantly faster than the passive process; second, the electric field–directed process requires the use of only minimal concentrations of DNA nanoparticles, which would not be viable for passive hybridization process which requires much higher concentration of nanoparticles; third, the integrity of the biomolecular binding reactions (hybridization) and the ionic flux was maintained throughout the large number of process steps (20 depositions). Thus, high fidelity DNA hybridization and a large number/variety of unique DNA binding sequences can be used for nanofabrication; fourth, the higher order layered DNA nanoparticle structures can ultimately be removed from the microarray by a relatively simple lift-off procedure. Thus, an electric field array device can serve as a manufacturing platform for

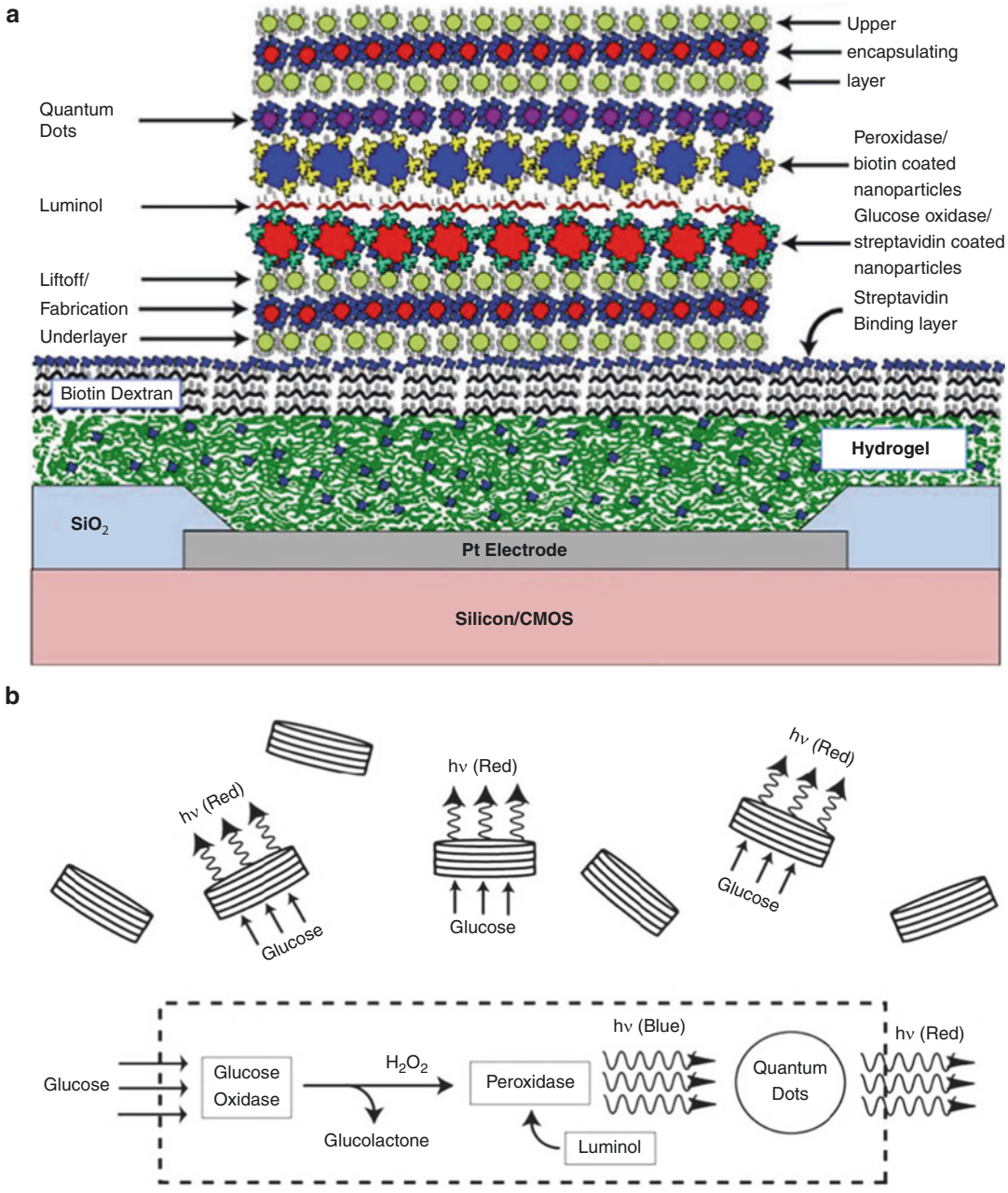


Electric Field–Directed Assembly of Bioderivatized Nanoparticles, Fig. 4 *Results of electric field–directed nanoparticle assembly/layering on a microarray device*. (**a**) Shows a section of the 400 site CMOS microarray device where alternate rows of microelectrodes were used to carry out nanoparticle layering. Rows with *bright circular spots* had 40 layers of alternating 40 nm *red* fluorescent biotin nanoparticles and 40 nm *green* fluorescent streptavidin nanoparticles deposited by applying a DC electric field. Rows with *lighter circular spots* had no electric field applied. (**b**) Shows an SEM image of a microelectrode which had no DC field applied. No nanoparticles have been deposited. (**c**) Shows a SEM image of a microelectrode to which the DC electric field was applied. Nanoparticles have been deposited on this surface. (**e**) Shows an SEM image of a cross section of a 40 layer structure of alternating 40 nm *red* fluorescent biotin nanoparticles and 40 nm *green* fluorescent streptavidin nanoparticles. (**f**) Shows an SEM image of a cross section of a 40 layer structure of alternating 40 nm biotin nanoparticles and 200 nm streptavidin nanoparticles. (**g**) Shows another SEM image of a cross section of a 40 layer structure of alternating 40 nm biotin nanoparticles and 200 nm streptavidin nanoparticles

making higher order 3D structures; and fifth, the overall process demonstrated the nanofabrication of higher order structures from a relatively heterogeneous group of materials, that is, biotin–dextran polymers, streptavidin, and two different sets of DNA nanoparticles.

Future Work

Some important potential future applications for the electric field–directed nanofabrication process include miniaturized chemical and biosensor devices, "micron-sized" dispersible chemo/



Electric Field–Directed Assembly of Bioderivatized Nanoparticles, Fig. 5 *Future glucose micro-biosensor device fabricated by electric field–directed assembly.* (**a**) The basic glucose micro-biosensor structure would include a bottom (lift-off) layer of biotin–streptavidin nanoparticles; a layer of glucose oxidase–streptavidin nanoparticles; a layer of luminol–dextran–biotin; a layer of streptavidin, a layer of peroxidase–biotin nanoparticles, a layer of *red* fluorescent streptavidin quantum dots, and a final upper capping layer of biotin–streptavidin nanoparticles. (**b**) When placed in a sample containing glucose, the glucose would diffuse into the micro-biosensor device and react with the glucose oxidase to produce hydrogen peroxide. The hydrogen peroxide and luminol diffusing to the peroxidase layer would catalyze the chemiluminescence (*blue light*). The *blue light* via fluorescent energy transfer to *red* fluorescent quantum dots would now produce detectable *red* fluorescence signal

biosensors for environmental and bioagent detection, lab-on a-chip devices, and in vivo diagnostic/drug delivery systems. Figure 5 shows the scheme for a future glucose micro-biosensor device fabricated using electric field–directed nanoparticle assembly. The basic glucose micro-biosensor structure would include a bottom (lift-off) layer of biotin–streptavidin nanoparticles; a layer of glucose oxidase–streptavidin nanoparticles; a layer of luminol–dextran–biotin; a layer of streptavidin, a layer of peroxidase–biotin nanoparticles, a layer of red fluorescent streptavidin quantum dots, and a final upper capping layer of biotin–streptavidin nanoparticles. When placed in a sample containing glucose, the glucose would diffuse into the micro-biosensor device and react with the glucose oxidase to produce hydrogen peroxide. The hydrogen peroxide and luminol diffusing to the peroxidase layer would catalyze the chemiluminescence (blue light). The blue light via fluorescent energy transfer to red fluorescent quantum dots would now produce detectable red fluorescence signal. This demonstrates that different fluorescent color responses could be created. Electronic array devices have other advantages including that they can be designed in a wide variety of shapes and sizes. To this end, larger "macroarray" devices have now been fabricated which can produce layered nanostructures with macroscopic dimensions. Figure 1e, f show a 4 in. silicon wafer size electric field array device. The total active deposition area for this device is 5 cm^2 with 200 platinum electrodes which are 2.5 × 2.5 mm with 8 μm gaps. Such "macroarray" devices will have applications for producing advanced fuel cells and photovoltaic materials. Overall electric field–directed nanofabrication using electronic array devices represents a unique synergy of combining the best aspects of "top-down" and "bottom-up" technologies into a viable nanofabrication process. It also represents a bio-inspired logic for self-assembly and hierarchal scaling of nanocomponents into integrated microscopic structures.

Cross-References

- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Biomimetic Synthesis of Nanomaterials
- ▶ Electric-Field-Assisted Deterministic Nanowire Assembly
- ▶ Quantum Dot Nanophotonic Integrated Circuits
- ▶ Self-Assembly
- ▶ Self-Assembly for Heterogeneous Integration of Microsystems
- ▶ Self-Assembly of Nanostructures
- ▶ Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components

References

1. National Research Council: Small Wonders, Endless Frontiers: Review of the National Nanotechnology Initiative. National Academy Press, Washington, DC (2002)
2. National Science and Technology Council: The National Nanotechnology Initiative – Strategic Plan. National Science and Technology Council, Washington, DC (2004)
3. Hughes, M.P. (ed.): Nanoelectromechanics in Engineering and Biology. CRC Press, Boca Raton (2003)
4. Goddard, W.A., Brenner, D.W., Lyashevski, S.E., Lafrate, G.J. (eds.): Handbook of Nanoscience. Engineering and Technology. CRC Press, Boca Raton (2003)
5. Bashir, R.: Biological mediated assembly of artificial nanostructures and microstructures. In: Goddard, W. A., Brenner, D.W., Lyashevski, S.E., Lafrate, G.J. (eds.) Handbook of Nanoscience, Engineering and Technology, pp. 15-1–15-31. CRC Press, Boca Raton (2003)
6. Heller, M.J., Tullis, R.H.: Self-organizing molecular photonic structures based on functionalized synthetic DNA polymers. Nanotechnology **2**, 165–171 (1991)
7. Hartmann, D.M., Schwartz, D., Tu, G., Heller, M.J., Esener, S.C.: Selective DNA attachment of particles to substrates. J. Mater. Res. **17**(2), 473–478 (2002)
8. Sosnowski, R.G., Tu, E., Butler, W.F., O'Connell, J.P., Heller, M.J.: Rapid determination of single base mismatch in DNA hybrids by direct electric field control. Proc. Natl. Acad. Sci. U. S. A. **94**, 1119–1123 (1997)
9. Heller, M.J., Tu, E., Holmsen, A., Sosnowski, R.G., O'Connell, J.P.: Active microelectronic arrays for DNA hybridization analysis. In: Schena, M. (ed.) DNA Microarrays: A Practical Approach,

pp. 167–185. Oxford University Press, New York (1999)
10. Heller, M.J.: DNA microarray technology: devices, systems and applications. Annu. Rev. Biomed. Eng. **4**, 129–153 (2002)
11. Heller, M.J., Tu, E., Martinsons, R., Anderson, R.R., Gurtner, C., Forster, A., Sosnowski, R.: Active microelectronic array systems for DNA hybridization, genotyping, pharmacogenomics and nanofabrication applications. In: Dekker, M. (ed.) Integrated Microfabricated Devices, pp. 223–270. Heller and Guttman, New York (2002)
12. Gilles, P.N., Wu, D.J., Foster, C.B., Dillion, P.J., Channock, S.J.: Single nucleotide polymorphic discrimination by an electronic dot blot assay on semiconductor microchips'. Nat. Biotechnol. **17**(4), 365–370 (1999)
13. Edman, C.F., Swint, R.B., Gurthner, C., Formosa, R. E., Roh, S.D., Lee, K.E., Swanson, P.D., Ackley, D.E., Colman, J.J., Heller, M.J.: Electric field directed assembly of an InGaAs LED onto silicon circuitry. IEEE Photon. Technol. Lett. **12**(9), 1198–1200 (2000)
14. US # 6,569,382, Methods and apparatus for the electronic homogeneous assembly and fabrication of devices, issued 27 May 2003
15. Mirkin, C.A., Letsinger, R.L., Mucic, R.C., Storhoff, J.J.: A DNA-based method for rationally assembling nanoparticles into macroscopic materials. Nature **382**, 607–609 (1996)
16. Mardilovich, P., Kornilovitch, P.: Electrochemical fabrication of nanodimensional multilayer films. Nano Lett. **5**(10), 1899–1904 (2005)
17. Hua, F., Shi, J., Lvov, Y., Cui, T.: Patterning of layer-by-layer self-assembled multiple types of nanoparticle thin films by lithographic technique. Nano Lett. **2**(11), 1219–1222 (2002)
18. Dehlinger, D.A., Sullivan, B., Esener, S., Hodko, D., Swanson, P., Heller, M.J.: Automated combinatorial process for nanofabrication of structures using bioderivatized nanoparticles. J. Assoc. Lab. Autom. **12**(5), 267–276 (2007)
19. Dehlinger, D.A., Sullivan, B.D., Esener, S., Heller, M. J.: Electric field directed assembly of biomolecular derivatized nanoparticles into higher order structures. Small **3**(7), 1237–1244 (2007)
20. Dehlinger, D.A., Sullivan, B., Esener, S., Heller, M.J.: Directed hybridization of DNA nanoparticles into higher order structures. Nano Lett. **8**, 4053–4060 (2008)

Electrical Conductivity

► Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy

Electrical Impedance Cytometry

David Holmes[1] and Benjamin L. J. Webb[2]
[1]London Centre for Nanotechnology, University College London, London, UK
[2]Division of Infection and Immunity, University College London, London, UK

Synonyms

Impedance cytometry; Single cell impedance spectroscopy

Definition

Electrical impedance cytometry is a technique whereby the dielectric or impedance properties of biological cells are measured. An externally applied electric field is used to probe the cell or sample of cells. This can be achieved either through the application of one or more discrete excitation frequencies or via broadband frequency measurement techniques. Typically, a potential is applied between a pair of electrodes and the resulting current flowing through the system is measured. The impedance of the system is the ratio of the voltage to the current passing through the system. The dielectric properties of the cells can be derived from this measurement through the use of appropriate models. The development of microfluidics and lab-on-a-chip type devices has allowed single cell impedance measurements to be performed with high sensitivity and high throughput.

Overview

The electrical measurements of biological samples has a long history dating back over a century to the pioneering work in the 1910s by Höber, who measured the low and high frequency conductivity of suspensions of erythrocytes [1]. A decade later Fricke described the electrical

properties of dispersive systems building on the work of Maxwell. Fricke was the first to measure the capacitive properties (8.1 mF/m^2) and thickness (3.3 nm) of the cell membrane – measurements which have subsequently been shown to be remarkably accurate [2]. The first single cell measurements were performed in the 1930s by Curtis and Cole [3].

Schwan [4] pioneered the field of biological impedance analysis; and was responsible for identifying the three main dielectric dispersions (α, β, γ) for suspensions of cells. These dispersions occur across a range of frequencies. The α-dispersion occurs at low-frequencies (~1 kHz) and arises from the polarization of the double layer around the particle. The β-dispersion occurs at intermediate frequencies (MHz) and arises from interfacial charging of the cell membrane. The γ-dispersion arises from the dipolar relaxation of water and occurs in the GHz range.

Experimental instrumentation and theoretical understanding have progressed in tandem, with highly sensitive techniques being developed, which allow the measurement of the dielectric properties of biological particles in fluid suspension. The majority of these techniques are however limited by the fact that they only provide an average value for the dielectric properties of a collection of particles. More recently, with the advent of micro- and nanofabrication techniques and the lab-on-a-chip, it has been possible to perform dielectric spectroscopic experiments on single biological particles suspended in physiological media.

This entry presents theory and models, which can be used to calculate the dielectric properties of cellular structures from impedance measurement. A practical realization of a single cell impedance measurement system is also presented along with data describing the electrical properties of single cells flowing in fluid suspension. Electrical measurement of cells growing on substrates and other more traditional bulk dielectric spectroscopy techniques (which measure the averaged properties of 100,000 s of cells) is not covered; much theoretical and experimental data is available for these techniques. Such techniques are robust and well characterized with a number of commercial systems available.

Theoretical Background

Maxwell Mixture Theory

The single-shelled spherical model can be used to describe a cell in suspension. Figure 1a shows the spherical model of a cell comprising a single shell; the dielectric and geometrical parameters are shown.

The impedance of a dilute suspension of particles in a uniform electric field can be calculated in terms of individual particle dielectric properties by Maxwell's mixture theory; where the equivalent complex permittivity $\tilde{\varepsilon}_{mix}$ of a mixture of particles and suspending medium is given by:

$$\tilde{\varepsilon}_{mix} = \tilde{\varepsilon}_m \frac{1 + 2\Phi\tilde{f}_{CM}}{1 - \Phi\tilde{f}_{CM}} \tag{1}$$

where

$$\tilde{f}_{CM} = \frac{\tilde{\varepsilon}_p - \tilde{\varepsilon}_m}{\tilde{\varepsilon}_p + 2\tilde{\varepsilon}_m} \tag{2}$$

and $\tilde{\varepsilon} = \varepsilon - j\sigma/\omega$ is the complex permittivity, the subscripts m and p refer to the medium and particle respectively, $i^2 = -1$, ω the angular frequency, $\tilde{f}_{CM}$ is the Clausius-Mossotti and Φ is the volume fraction.

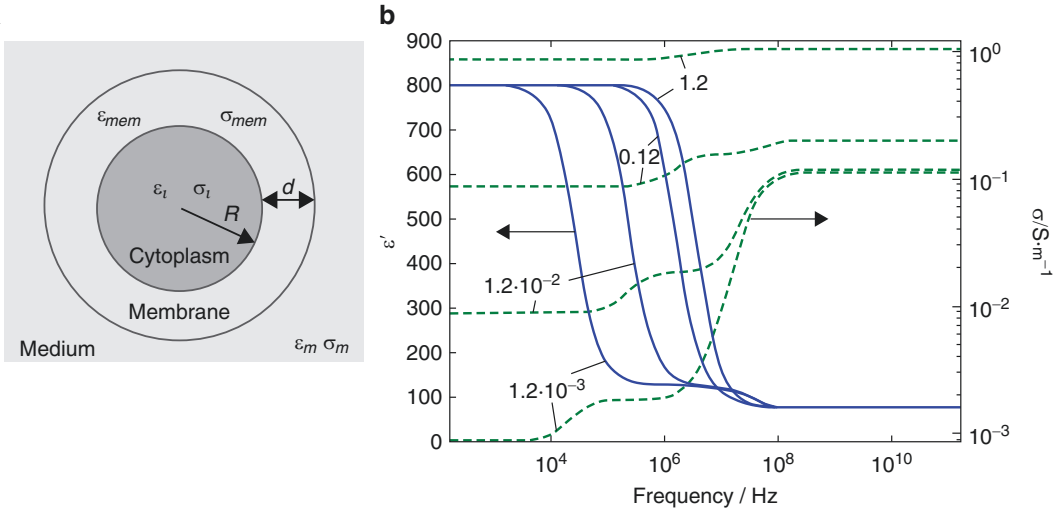
The complex permittivity of the cell, $\tilde{\varepsilon}_p$ is a function of the dielectric properties of membrane and cytoplasm, cell membrane $\tilde{\varepsilon}_{mem}$ and internal properties $\tilde{\varepsilon}_{cyto}$, and cell geometry (where R is the inner radius and d the membrane thickness) given by:

$$\tilde{\varepsilon}_p = \tilde{\varepsilon}_{mem} \frac{\gamma^3 + 2\left(\frac{\tilde{\varepsilon}_{cyto} - \tilde{\varepsilon}_{mem}}{\tilde{\varepsilon}_{cyto} + 2\tilde{\varepsilon}_{mem}}\right)}{\gamma^3 - \left(\frac{\tilde{\varepsilon}_{cyto} - \tilde{\varepsilon}_{mem}}{\tilde{\varepsilon}_{cyto} + 2\tilde{\varepsilon}_{mem}}\right)} \tag{3}$$

where

$$\gamma = \frac{R + d}{R} \tag{4}$$

This is only valid for dilute suspensions ($\Phi < 10$ %). Interaction between the particles



Electrical Impedance Cytometry, Fig. 1 (**a**) Single-shell electrical model of the spherical cell in suspension. (**b**) Plot showing the real and imaginary parts of the CM. The following parameters were used: $\varepsilon_o = 8.854 \times 10^{-12}$ F.m^{-2}, $R = 3 \times 10^{-6}$ m, $d = 5 \times 10^{-9}$ m, $\varepsilon_m = 80 \times \varepsilon_o$, $\varepsilon_{mem} = 5 \times \varepsilon_o$, $\sigma_{mem} = 10^{-8}$ S.m^{-1}, $\varepsilon_\iota = 60 \times \varepsilon_o$, $\sigma_\iota = 0.4$ S.m^{-1} and the medium conductivity varied between $\sigma_m = 1.2 \times 10^{-3}$ S.m^{-1} and 1.2 S.m^{-1}

must be taken into account for higher particle concentrations. The mixture theory can be extended to high volume fractions:

$$1 - \Phi = \left(\frac{\tilde{\varepsilon}_{mix} - \tilde{\varepsilon}_p}{\tilde{\varepsilon}_m - \tilde{\varepsilon}_p}\right)\left(\frac{\tilde{\varepsilon}_m}{\tilde{\varepsilon}_{mix}}\right)^{1/3} \quad (5)$$

The impedance of the system is related to the complex permittivity as follows:

$$\tilde{Z}_{mix} = \frac{1}{j\omega\tilde{\varepsilon}_{mix}} G_f \quad (6)$$

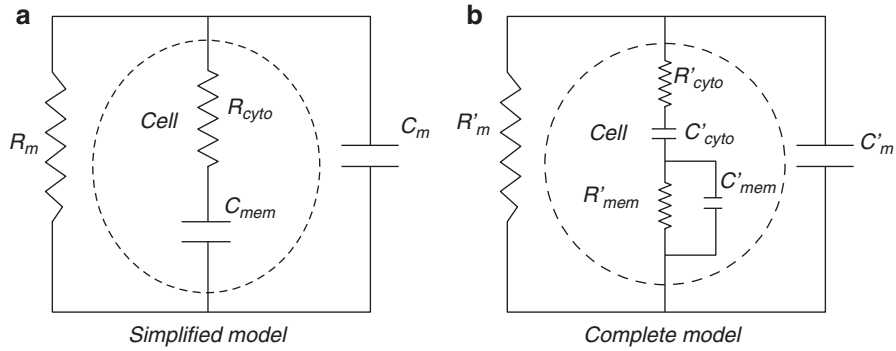
where G_f is a geometric constant, which for an ideal parallel plate electrode system is simply the ratio of electrode area to gap A/g (m). Combining Eq. 6 with Eq. 1 allows the complex permittivity of a suspension of cells to be determined from the frequency-dependent impedance measurements. Typically, a cell suspension is measured using macroscopic planar or cylindrical electrode geometries where G_f can be easily defined. This is not the case for single cell impedance analysis, where the cell is located between a pair of microelectrodes of comparable size to that of the cell (ca. 10–20 μm). In this case the electric field is not uniform and the effect of the divergent field (fringing field) must be considered to correctly model the impedance. This involves detailed analysis of the field geometry and a modification to G_f [5, 6]. It is possible to use more complex electrode geometries, incorporating guard electrodes to confine the electric field and simplify the analysis; however, this requires careful electrode design and complicates the fabrication process.

Equivalent Circuit Model

For simplified analysis of the system, an electrical circuit analog is often used. Such an approach was first developed by Foster and Schwan [7]. The cell can be described in terms of an equivalent circuit model, as shown in Fig. 2.

Simplified Circuit Model

To a first level of approximation, the cell can be modeled as a resistor that describes the cytoplasm in series with a capacitor for the membrane as shown in Fig. 2a. The cell membrane resistance is generally much greater than the reactance of the membrane and can therefore be ignored. Likewise, the capacitance of the cell cytoplasm can be ignored when its reactance is small compared to the cell cytoplasm resistance. The values of the



Electrical Impedance Cytometry, Fig. 2 Foster and Schwan's simplified circuit model for a single cell in suspension. (**a**) The cell is modeled as a resistor R_{cyto} (cytoplasm) and a capacitor C_{mem} (membrane) in series, with the suspending medium modeled as a resistor R_m and capacitor C_m. (**b**) The complete circuit model for a single-shelled particle in suspension. The particle is modeled as a resistor $R'c_{yt}$o and a capacitor $C'c_{yto}$ in series (cytoplasm) in combination with a resistor R'_{mem} and a capacitor C'_{mem} in parallel (membrane) (Adapted with permission from Sun et al. 2012, © 2010, Springer) [12]

electrical components in the simple model circuit are given below. The suspending medium is described by:

$$R_m = \frac{1}{\sigma_m(1 - 3\Phi/2)G_f} \tag{7}$$

$$C_m = \varepsilon_\infty G_f \tag{8}$$

The simplified cell components are given by:

$$C_{mem} = \frac{9\Phi R C_{mem,0}}{4} G_f \tag{9}$$

$$R_{cyto} = \frac{4\left(\frac{1}{2\sigma_m} + \frac{1}{\sigma_{cyto}}\right)}{9\Phi G_f} \tag{10}$$

with specific membrane capacitance (per unit area)$C_{mem,0} = \varepsilon_{men}/d$. The limiting high frequency permittivity of the suspension is related to the suspending medium permittivity according to:

$$\varepsilon_\infty \simeq \varepsilon_m \left[1 - 3\Phi \frac{\varepsilon_m - \varepsilon_{cyto}}{2\varepsilon_m + \varepsilon_{cyto}}\right] \tag{11}$$

This simplified circuit model has been used to interpret single cell impedance measurements providing good agreement with experiments (e.g., see Fig. 4).

Complete Circuit Model

For cases where the cell membrane conductance and cytoplasm capacitance values vary widely and cannot be ignored (e.g., during electroporation or cell lysis), a complete equivalent circuit model is required. This includes the resistance of the membrane and the capacitance of the cytoplasm. The complete circuit model is shown in Fig. 2b and the equations are given below. The suspending medium is described by:

$$R'_m = \frac{1}{\sigma_0 G_f} \tag{12}$$

$$C'_m = \varepsilon'_\infty G_f \tag{13}$$

The complete model cell components are described by:

$$R'_{mem} = \frac{1}{G_f}\left[\frac{\tau_1 + \tau_2}{\Delta\varepsilon_1 + \Delta\varepsilon_2} - \frac{1}{k_2 + k_3} - \frac{\tau_1\tau_1(k_2 + k_3)}{(\Delta\varepsilon_1 + \Delta\varepsilon_2)^2}\right] \tag{14}$$

$$C'_{mem} = \frac{\tau_1\tau_2(k_2 + k_3)}{(\Delta\varepsilon_1 + \Delta\varepsilon_2)R'_{mem}} \tag{15}$$

$$R'_{cyto} = \frac{1}{(k_2 + k_3)G_f} \tag{16}$$

$$C'_{cyto} = (\Delta\varepsilon_1 + \Delta\varepsilon_2)G_f \tag{17}$$

where R'_{mem} and C'_{mem} are the resistance and capacitance of the cell membrane, respectively. R'_{cyto} and C'_{cyto} are the resistance and capacitance of the cell cytoplasm, respectively. Full details of the above equations can be found in [8, 9].

Coulter Counter

Resistive pulse sensors are a specific case of the impedance cytometer. More commonly known as the Coulter counters, the technique measures the DC resistance between two electrically isolated fluid-filled chambers and the change in this current as a particle passes through a small connecting orifice [10]. Two large electrodes are positioned at either side of the orifice. As a particle passes through the orifice it displaces the conductive fluid and alters the resistance, this change results in a current pulse, the magnitude of which is proportional to the cell size. A simplified analysis of the Coulter counter was presented by Deblois and Bean [11]. The resistance of a tube, diameter D_t, length L_t, filled with electrolyte of resistivity ρm is:

$$R_t = \frac{4\rho_m L_t}{\pi D_t^2} \tag{18}$$

Sun et al. [12] showed that Maxwell's approximation can be used to evaluate the resistance of the tube containing a spherical particle (diameter d_p), giving an expression for the resistance, R_{mix} of a particle suspension:

$$R_{mix} = \frac{4\rho_{mix} L_t}{\pi D_t^2} = \frac{4\rho_m L_t}{\pi D_t^2}\left(1 + \frac{d_p^3}{D_t^2 L_t}\right) \tag{19}$$

The resistance change ΔR due to the particle is given by:

$$\Delta R = R_{mix} - R_t = \frac{4\rho_m d_p^3}{\pi D_t^4} \tag{20}$$

This expression is based on the assumption that the particle diameter is small compared with the tube (i.e., $\Phi << 1$), but ignores other issues such as access resistance to the tube. A solution with applicability over a broader range was given by Deblois and Bean [11]:

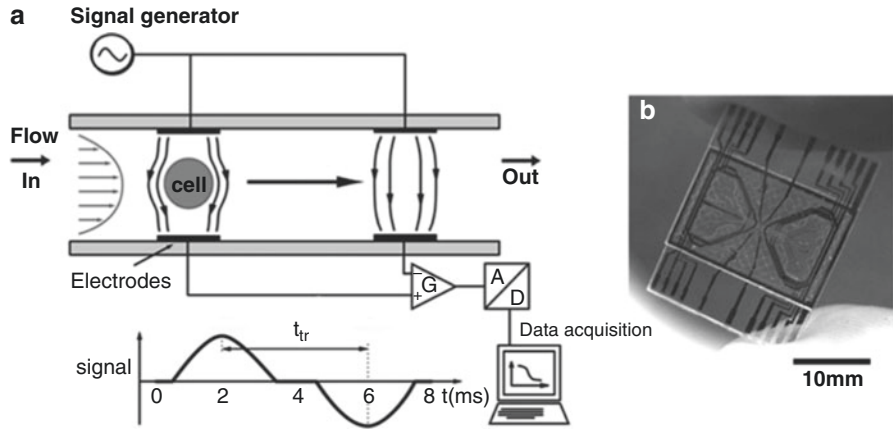
$$\Delta R_{(D_t/L_t<<1)} = \left(\frac{4\rho_m d_p^3}{\pi D_t^4}\right) \cdot F\left(\frac{d_p^3}{D_t^3}\right) \tag{21}$$

where $F(d_p{}^3/D_t{}^3)$ is a correction term that accounts for the nonuniformity of the current density within the tube.

Basic Methodology

The most successful implementation of a microfluidic impedance cytometry chip, in terms of measurement sensitivity and accuracy, is that of the facing electrode geometry as proposed by Gawad [13]. Other geometries have been implemented such as planar electrode geometries [14] and the so-called liquid electrodes [15]. The facing electrode implementation of a single cell impedance measurement system is shown in Fig. 3a. Electrodes on the upper and lower walls of the microfluidic channel are connected to an AC voltage source, which provides excitation signals at one or more frequencies. To minimize signal noise and drift a differential measurement is performed. The current flowing in the system is measured as a cell passes from one arm of the sensing circuit (two electrode pairs) to the other. The transit time between peaks, *ttr* allows the particle velocity to be monitored, giving information about particle position within the channel and sample flow rates. An example of a microfluidic impedance cytometry chip is shown in Fig. 3b; the chip is constructed from glass substrates with platinum microelectrodes and channel sidewalls formed in polymer. The upper and lower channel walls (glass) allow for microscope observation of cells flowing though the device.

A simplified schematic of the impedance detection system is shown in Fig. 4. A differential detection circuit is used. The electrode pair on the left of Fig. 4a illustrates the equivalent circuit model for the detection region with no cell between the electrodes, the right electrode pair



Electrical Impedance Cytometry, Fig. 3 (**a**) Schematic diagram showing the implementation of a single cell impedance measurement system. Electrodes on the upper and lower walls of the microfluidic channel are connected to an AC voltage source (providing excitation at one or more frequencies). The differential current flowing in the system is measured as a cell passes from one arm of the sensing circuit to the other. The transit time between peaks, t_{tr} allows the particle velocity to be monitored. (**b**) Photograph of a microfluidic impedance chip

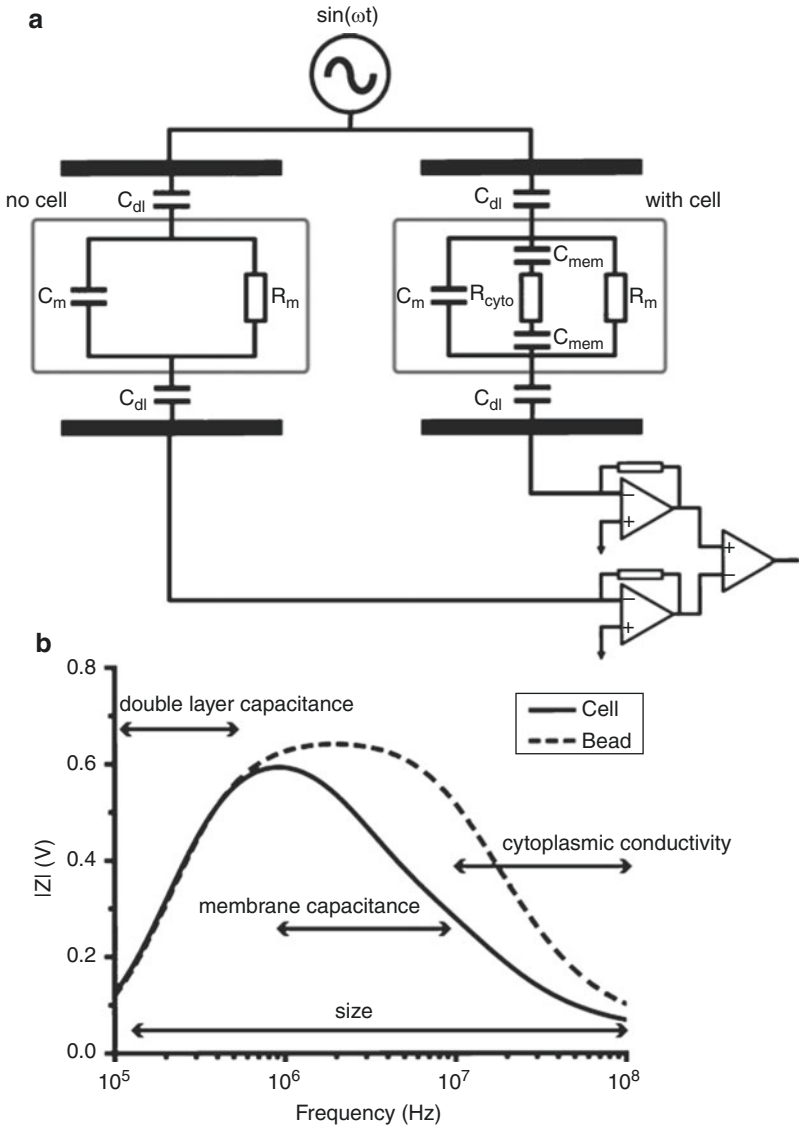
has the cell between the electrodes. As described above, R_m and C_m are the equivalent resistance and capacitance of the medium, respectively, C_{mem} is the equivalent capacitance of the cell membrane, and R_{cyto} is equivalent resistance of the cell cytoplasm. C_{dl} represents the electrical double layer capacitance at the electrode–liquid interface. The impedance magnitude from a polymer bead (i.e., an insulating particle) and a cell of the same diameter are shown in Fig. 4b including a graphical indication of the frequency ranges over which the different structures of the cell have an influence on the impedance signal [16].

Key Research Findings

Single Cell Impedance Cytometer for Blood Analysis

One of the most commonly ordered tests in the doctor's office is complete blood count (CBC). Although simple and relatively inexpensive, this test is a very effective screening tool for a wide range of conditions. For example, a lower number of red blood cells could indicate anemia. A higher number of neutrophils could indicate an infection in progress, while changes in the number of lymphocytes could reveal diseases. Current standards for CBC follow stringent requirements for quality control and other regulations and rely on automated machines located either in hospitals or centralized laboratories. Impedance-based counting of white blood cells and differentiating the major subtypes has been demonstrated and has the potential to replace the bulky hematology analyzers, currently only available in hospitals and centralized research labs.

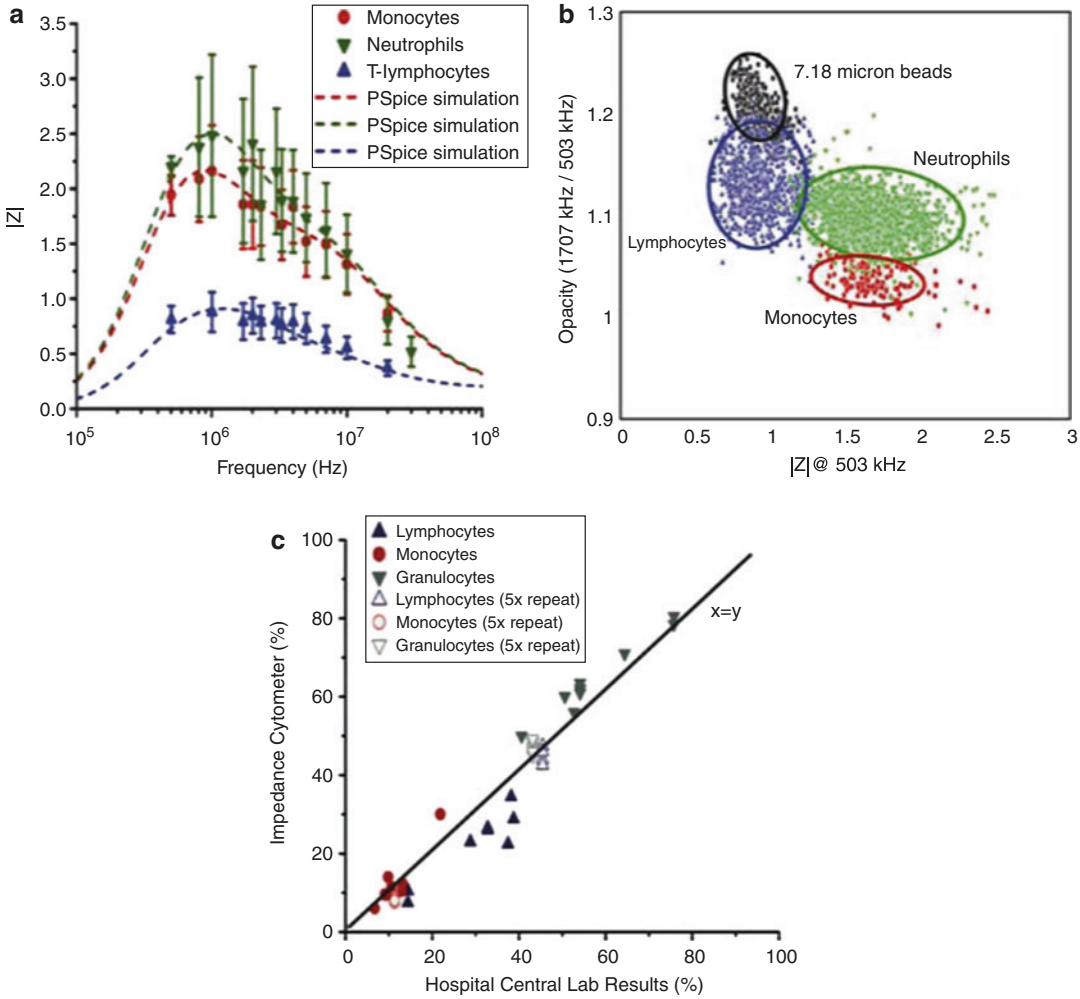
Figure 5a shows the impedance magnitude versus frequency for purified leukocyte subpopulations (T-lymphocytes, monocytes, and neutrophils). The mean and standard deviation of ~1,500 cells is plotted at each frequency point. The dashed lines in the plot show the best-fit circuit simulation (using PSPICE) incorporating the equivalent circuit model, electrical double layer, and differential amplification electronics of the system. Figure 5b shows the impedance-based analysis of clinical samples; the lymphocyte, neutrophil, and monocyte populations are clearly visible. Tests with patient samples are presented in Fig. 5c and show 95 % correlation against commercial (optical/Coulter) blood analysis equipment, demonstrating the potential clinical utility of the impedance microcytometer for a point-of-care blood analysis system [16].



Electrical Impedance Cytometry, Fig. 4 (**a**) Simplified schematic of the impedance detection system shows the differential detection system. The electrode pair on the *left* illustrates the equivalent circuit model with no cell between the electrodes; the *right* electrode pair has the cell between the electrodes. R_m and C_m are the equivalent resistance and capacitance of the medium, respectively, C_{mem} is the equivalent capacitance of the cell membrane, R_{cyto} is equivalent resistance of the cell cytoplasm. C_{dl} represents the electrical double layer capacitance at the electrode–liquid interface. (**b**) The frequency-dependent impedance magnitude from a polymer bead (i.e., an insulating particle) and a cell of the same diameter (Adapted with permission from Holmes et al. 2009, copyright © 2009, RSC) [16]

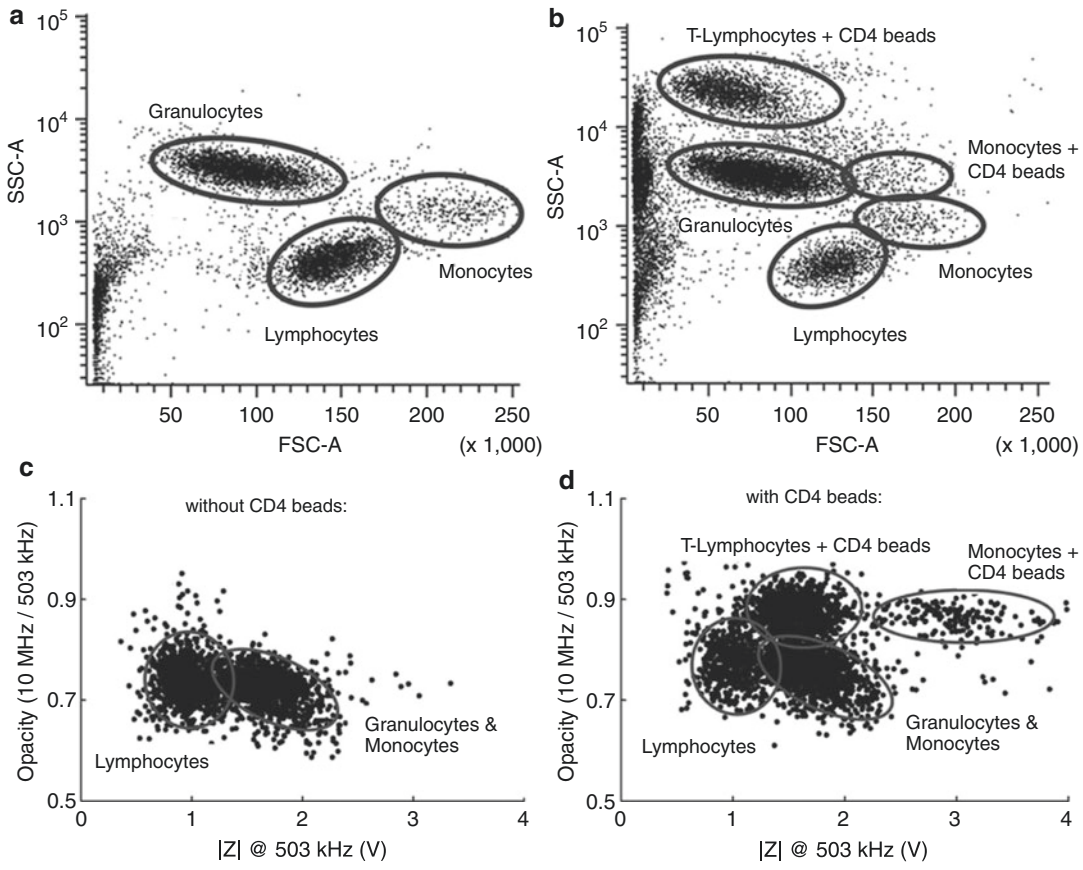
Impedance Labeling for Impedance Discrimination Based on Antigenic Expression

Electrical impedance cannot detect different antigen expressing cell subtypes with similar morphology; however, a method analogous to fluorescent antibody labeling was recently demonstrated which allows the discrimination of antigenically specific white blood cells [17].



Electrical Impedance Cytometry, Fig. 5 (**a**) Data showing the impedance magnitude versus frequency for purified leukocyte subpopulations (T-lymphocytes, monocytes, and neutrophils). The mean and standard deviation of ~1,500 cells is plotted at each frequency point. The *dashed lines* show the best fit circuit simulation incorporating the equivalent circuit model, electrical double layer, and differential amplification electronics. (**b**) Data showing discrimination of leukocyte subpopulations from clinical blood sample. (**b**) Comparison of impedance-based blood analysis with hospital central lab analysis using a traditional hematology analyzer (Adapted with permission from Holmes et al. 2009, copyright © 2009, RSC) [16]

This was demonstrated through the impedance identification of lymphocyte subset expressing the CD4 surface protein. A suspension of blood cells was mixed with micron sized polymer beads coated with an antibody that specifically recognizes the CD4 antigen. The beads bind to the CD4 expressing cells (T-lymphocytes and monocytes), and in doing so change the impedance properties of these cells. Figure 6 shows the same blood sample with and without the addition of CD4 impedance labels. The bead complexed cells can be clearly seen as separate population on the scatter plot. The T-lymphocyte and monocyte populations are easily discriminated from one and other due to the larger size of the monocytes. The use of labels to selectively alter the impedance properties of certain cell subpopulations brings the molecular specificity of antibody labeling techniques to electrical cell discrimination.



Electrical Impedance Cytometry, Fig. 6 (**a**) FACS data for saponin-lysed whole blood showing the three dominant leukocyte subpopulations. Some debris from the *red* cells is also visible, but the granulocyte and lymphocyte populations can be clearly identified, together with the smaller monocyte population (note axis on log scale). (**b**) The same blood sample complexed with CD4 beads, showing two new populations, the CD4+ T-lymphocytes and the subpopulation of monocytes that bind the beads. (**c**, **d**) show scatter plots from the impedance cytometer. The x-axis shows impedance measured at 503 kHz, which is a measure of cell size, and the y-axis shows the opacity (ratio of impedance at 10 MHz and 503 kHz). The CD4+ cell subpopulations are clearly visible in the impedance plot shown in (**d**) and are comparable with (**b**) (Adapted with permission from Holmes et al. 2010, copyright © 2010, ACS) [17]

Future Directions of Research

Impedance measurements represent a powerful noncontact method of analysis, with the potential to extract a wealth of information relating to the physiological and morphological properties of cells and other microscopic particles. The fact that the technique can probe cellular structures and extract information without the need for labeling makes it appealing for a range of diagnostic and other metrology applications where labeling can be expensive and may modify the cells under investigation (e.g., cell labeling can lead to activation).

Extension of the frequency range of the current impedance cytometry techniques to both higher and lower frequencies will allow probing of surface and internal properties of cells. Low-frequency (~1 kHz) measurement of the α-relaxation of cells is an important area of research, which has been poorly investigated to date due to the lack of experimental tools. High frequency (<50 MHz) measurements will allow probing of the intracellular compartment; current

methods rely upon suspending cells in buffer of low conductivity to enhance the signal response and shift the membrane relaxation to lower frequencies (i.e., within the frequency response of the experimental apparatus). Impedance measurement of the cytoplasmic compartment will allow label-free monitoring of cell cycle, intracellular parasite detection (e.g., malaria), and a host of other biophysical measurements. Broadband frequency measurements have been implemented allowing full impedance spectra to be obtained rapidly and accurately [18].

Scaling down of impedance detection systems to allow measurements to be performed on viruses and other nanoparticles may be possible. A number of nonmammalian cell types have been measured using impedance cytometry (e.g., bacteria [19], marine algae [20]), and this combined with the integration of powerful sample preparation techniques and particle focusing will lead to the development of a range of powerful, robust, and portable devices, which will have relevance to a broad range of environmental, clinical, and life science applications.

Cross-References

- ▶ Electrical Impedance Tomography for Single-Cell Imaging
- ▶ Micro- and Nanofluidic Devices for Medical Diagnostics
- ▶ Microfluidic Whole-Cell Biosensor
- ▶ On-Chip Flow Cytometry
- ▶ Single Cell Impedance Spectroscopy

References

1. Höber, R.: Eine Methode die elektrische Leitfaehigkeit im Innern von Zellenzu messen. Arch. Ges. Physiol. **133**, 237–259 (1910)
2. Fricke, H.: The electric capacity of suspensions of a red corpuscles of a dog. Phys. Rev. **26**, 682–687 (1925)
3. Curtis, H.J., Cole, K.S.: Transverse electric impedance of Nitella. J. Gen. Physiol. **21**, 189–201 (1937)
4. Schwan, H.P.: Electrical properties of tissue and cell suspensions. Adv. Biol. Med. Phys. **5**, 147–209 (1957)
5. Sun, T., Green, N.G., Gawad, S., Morgan, H.: Analytical electric field and sensitivity analysis for two microfluidic impedance cytometer designs. IET Nanobiotechnol. **1**, 69–79 (2007)
6. Linderholm, P., Seger, U., Renaud, P.: Analytical expression for electric field between two facing strip electrodes in microchannel. Electron. Lett. IEE **42**, 145–147 (2006)
7. Foster, K.R., Schwan, H.P.: Dielectric properties of tissues and biological materials: a critical review. Crit. Rev. Biomed. Eng. **17**, 25–104 (1989)
8. Sun, T., Gawad, S., Green, N.G., Morgan, H.: Dielectric spectroscopy of single cells: time domain analysis using Maxwell's mixture equation. J. Phys. D Appl. Phys. **40**, 1–8 (2007)
9. Sun, T., Bernabini, C., Morgan, H.: Single-colloidal particle impedance spectroscopy: complete equivalent circuit analysis of polyelectrolyte microcapsules. Langmuir **26**(6), 3821–3828 (2009)
10. Bayley, H., Martin, C.R.: Resistive-pulse sensing: from microbes to molecules. Chem. Rev. **100**, 2575–2594 (2000)
11. Deblois, R.W., Bean, C.P.: Counting and sizing of submicron particles by the resistive pulse technique. Rev. Sci. Instrum. **41**, 909–916 (1970)
12. Sun, T., Morgan, H.: Single-cell microfluidic impedance cytometry: a review. Microfluid. Nanofluid. **8**, 423–443 (2010)
13. Gawad, S., Cheung, K., Seger, U., Bertsch, A., Renaud, P.: Dielectric spectroscopy in a micromachined flow cytometer: theoretical and practical considerations. Lab Chip **4**, 241–251 (2004)
14. Gawad, S., Schild, L., Renaud, P.: Micromachined impedance spectroscopy flow cytometer for cell analysis and particle sizing. Lab Chip **1**, 76–82 (2001)
15. Demierre, N., Braschler, T., Linderholm, P., Seger, U., van Lintel, H., Renaud, P.: Characterization and optimization of liquid electrodes for lateral dielectrophoresis. Lab Chip **7**, 355–365 (2007)
16. Holmes, D., Pettigrew, D., Reccius, C.H., Gwyer, J.D., Berkel, C.V., Holloway, J., Davie, D.E., Morgan, H.: Leukocyte analysis and differentiation using high speed microfluidic single cell impedance cytometry. Lab Chip **9**, 2881–2889 (2009)
17. Holmes, D., Morgan, H.: Single cell impedance cytometry for identification and counting of CD4 T-cells in human blood using impedance labels. Anal. Chem. **82**, 1455–1461 (2010)
18. Sun, T., Holmes, D., Gawad, S., Green, N.G., Morgan, H.: High speed multi-frequency impedance analysis of single particles in a microfluidic cytometer using maximum length sequences. Lab Chip **7**, 1034–1040 (2007)
19. Bernabini, C., Holmes, D., Morgan, H.: Micro-impedance cytometry for detection and analysis of micron-sized particles and bacteria. Lab Chip **11**(3), 407–412 (2011)
20. Benazzi, G., Holmes, D., Sun, T., Mowlem, M.C., Morgan, H.: Discrimination and analysis of phytoplankton using a microfabricated flow-cytometer. IET Nanobiotechnol. **1**(6), 94–101 (2007)

Electrical Impedance Spectroscopy

▶ Electrical Impedance Tomography for Single-Cell Imaging

Electrical Impedance Tomography for Single-Cell Imaging

Tao Sun
Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA

Synonyms

Cellular imaging; Electrical impedance spectroscopy; Lab-on-a-chip; Rolled-up nanotechnology; Single cell analysis

Definition

Electrical impedance tomography (EIT) is an imaging technology that spatially characterizes the electrical properties of an object. It is an inverse problem: Given an object with unknown electrical properties, reconstruction of an image requires determination of the internal conductivity (or permittivity) profile of the object from a finite number of boundary measurements (current/voltage).

Introduction

The integration of mechanical elements, sensors, actuators, and electronics on a common substrate is known as Micro-Electro-Mechanical-Systems (MEMS). BioMEMS is a multidisciplinary subject across science and engineering and includes mathematics, biophysics, biochemistry, electrochemistry, microfluidics, micro/nano fabrication, integration, materials and surfaces, synthesis and detection technologies, accepted as a new concept – "Lab-On-a-Chip." For quantitatively understanding the functioning of the tissue and organisms, working at the cellular level is of prime interest for biomedical and clinical applications. Cellular analysis requires a combination of biophysical and biochemical approaches. Traditional analysis is performed with bulk techniques. However, measurements on large populations of cells can only provide average information. Individual cells, which look identical in appearance, generally have heterogeneous behavior. Moreover, the limitations of bulk techniques for studying the cell-to-cell and cell-to-surface interactions are apparent. Therefore, considerable attention is now paid to single cell analysis using microfluidic technologies [1].

Electrical impedance spectroscopy [2, 3] is a high-speed, noninvasive technique that is used to characterize the dielectric properties of biological cells. In the simplest sense, a biological cell consists of a conducting cytoplasm (with nucleus) covered by an ultrathin insulating cell membrane. When exposed to an ac electric field, at low frequencies the cell behaves as an insulating object. The cell membrane is equivalent to a capacitor and prevents the electric field lines from penetrating the cell. Therefore, information concerning cell size (and shape) can be obtained by measuring the electrical properties of the cell at low frequencies. At higher frequencies, the cell membrane electrically becomes short-circuited. Thus, the internal properties of the cell (i.e., cytoplasmic resistance) can be probed. For adherent cells, the kinetics during culture, such as cell growth, differentiation, and the effect of drugs can be monitored using electrical impedance. Cells are grown on the surface of an array of electrodes and the impedance is monitored as a function of time. This technique is termed electric cell-substrate impedance sensing [4].

In electrical impedance tomography (EIT) [5], measurements of a sample are made between multiple electrodes. The technique generates images of the conductivity distribution inside an object. Although, the spatial resolution of EIT cannot be compared with magnetic resonance imaging, the technique is harmless, less costly, and has good temporal resolution.

Electrical Impedance Tomography

EIT is an inverse problem: Given an object with unknown electrical properties, reconstruction of an image requires determination of the internal conductivity (or permittivity) profile of the object from a finite number of boundary measurements (current/voltage). The theory of EIT is composed of two parts: the forward problem and the inverse problem. The forward problem corresponds to the calculation of the potentials on the electrodes for a known injected current and conductivity distribution. Solving the forward problem involves constructing the global Jacobian (sensitivity) matrix of the system, which is required for image reconstruction and the solution of the inverse problem. For EIT, the electric field distribution in considered quasi-electrostatic. If only low frequency ac is used, the analysis is restricted to mapping the conductivity distribution of the object. Therefore, the electrical potentialϕ, satisfies Laplace's equation:

$$\nabla \bullet \sigma \nabla \phi = 0 \tag{1}$$

Equation 1 gives a unique solution for the electric potential distribution, if the boundary conditions are defined. The forward problem is well-posed in terms of existence, uniqueness and stability. However, in the case of EIT, since the potential distribution is dependent on the conductivity distribution, it cannot be solved analytically for arbitrary values of σ. Therefore, a general approach is to set up a numerical model that can simulate the potential values on the electrodes as close as possible to the measured voltage data in an experiment. The numerical model discretizes the domain into small elements so that an approximate solution can be obtained. Using the finite element method (FEM) with the defined boundary conditions, the governing equation of the system is:

$$\mathbf{A}\begin{bmatrix}\phi_p \\ V_l\end{bmatrix} - \begin{bmatrix}0 \\ 1\end{bmatrix} \text{ with } \mathbf{A} = \begin{bmatrix}\mathbf{A}_C & \mathbf{A}_E \\ \mathbf{A}_E^T & \mathbf{A}_D\end{bmatrix} \tag{2}$$

where $\mathbf{A}$ is the global admittance matrix of the system. The expressions of the compartments $\mathbf{A}C$, $\mathbf{A}E$, and $\mathbf{A}D$ can be found in [6]. The potential on every node ϕp in the finite element model and the voltage on every electrode Vl can be obtained by solving Eq. 2.

Image reconstruction requires the knowledge of the conductivity distribution. In EIT, the changes in conductivity σq within the domain, and the changes in the potential distribution ϕp are linked by the Jacobian (sensitivity) matrix $\mathbf{J}p,q$:

$$\mathbf{J}_{p,q} = \frac{\partial \phi_p}{\partial \sigma_q} \tag{3}$$

The inverse problem corresponds to calculation of the conductivity distribution when the injected current is known and the voltages are measured. It is "ill-posed" since the solutions of the inverse problem are not unique. The process of solving the inverse problem requires finding a stable value of σ, so that the difference between the numerical simulations $\phi(\sigma)$ and the measured voltage V is minimized.

$$f(\sigma) = \frac{1}{2}(\phi(\sigma) - V)^*(\phi(\sigma) - V) = \frac{1}{2}||\phi(\sigma) - V||^2 \tag{4}$$

To find the minima of the objective function $f(\sigma)$, a Taylor series is used to expand $f(\sigma)$, neglecting higher-order terms in Eq. 6:

$$f(\sigma + \Delta\sigma) = f(\sigma) + f'(\sigma)\Delta\sigma + \frac{1}{2}f''(\sigma)\Delta\sigma^2 \tag{5}$$

Using the Newton–Raphson method to search for a step $\Delta\sigma$, such that $f(\sigma + \Delta\sigma) \cong 0$, the iterative solution of Eq. 5 is:

$$\sigma_{n+1} = \sigma_n + \Delta\sigma \quad n > 0 \tag{6}$$

$$\Delta\sigma = -[Hf(\sigma_n)]^{-1}\nabla f(\sigma_n) \tag{7}$$

where $Hf(\sigma n)$ is the Hessian matrix and $\nabla f(\sigma n)$ is the gradient, given by:

$$\Delta f(\sigma) = f'(\sigma) = \phi'(\sigma)^*(\phi(\sigma) - V) \tag{8}$$

$$Hf(\sigma) = \phi'(\sigma)^*\phi'(\sigma) + \phi''(\sigma)(\phi(\sigma) - V) \tag{9}$$

For simplicity, the second-order derivative term in Eq. 9 can be neglected. Therefore, the expression for the adjusting step $\Delta\sigma$ becomes Eq. 6:

$$\begin{aligned}\Delta\sigma &= -[\phi'(\sigma)^*\phi'(\sigma)]^{-1}\phi'(\sigma)^*(\phi(\sigma)-V)\\ &= \phi'(\sigma)^{\dagger}(V-\phi(\sigma))\end{aligned} \tag{10}$$

where $\phi'(\sigma)^{\dagger}$ is the Moore–Penrose pseudo-inverse of $\phi'(\sigma)$, which is the inverse of the Jacobian matrix (Eq. 3).

The solutions given by Eqs. 6 and 10 are based on the Gauss–Newton algorithm. Other algorithms can be used to minimize Eq. 4. For example, the linear generalized Tikhonov regularization can be encapsulated into the Gauss–Newton algorithm, which gives calculation of $\Delta\sigma$ as:

$$\Delta\sigma = \left[\phi'(\sigma)^*\phi'(\sigma)+\lambda^2R^*R\right]^{-1}\left(\phi'(\sigma)^*V+\lambda^2R^*R\sigma\right) \tag{11}$$

where R is either the identity matrix or other regularization matrix including certain *image prior* assumptions about σ for reconstruction. λ is a positive scalar regularization parameter.

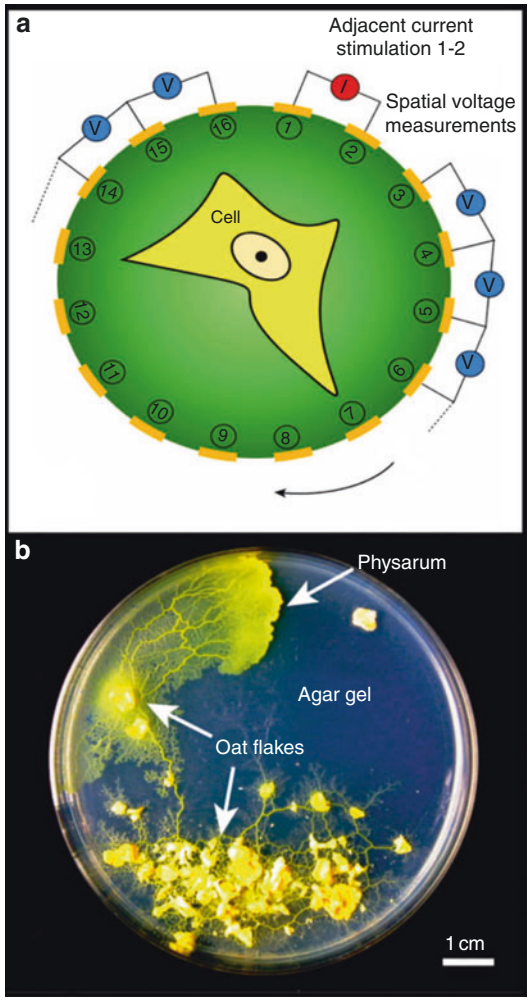
Over the past 20 years, a large number of reconstruction algorithms have been developed to improve the computational efficiency, reconstruction sensitivity, and image fidelity. Examples include the modified Newton–Raphson method, back-projection algorithm, variational method, maximum a posteriori approach, Monte Carlo sampling method, and direct reconstruction algorithms. Most biomedical and clinical uses of EIT are in vivo and used to detect disease such as breast cancer [7] or to monitor brain function [8]. Interest has centered on in vitro applications, e.g., Davalos et al. [9] proposed using EIT for detecting and imaging electroporation of cells in tissue in real time. Recently, there has been an interest in miniaturizing the technology. Hou et al. [10] patterned 32 electrodes in a rectangular configuration to provide a two-dimensional map of the conductivity of carbon nanotube thin films. Chai et al. [11] fabricated a set of 4 × 4 complementary metal oxide semiconductor microelectrode array to image a cell culture environment. Linderholm et al. [12] used an array of interdigitated electrodes to monitor the migration, stratification, and influence of detergent on cells in culture.

EIT for Single Cell Imaging

Individual cells can be identified from differences in size and dielectric properties using electrical techniques. Recently, an on-chip two-dimensional cell imaging technology using EIT was developed [13]. The cell can be cultured at any position within the circular array for imaging (Fig. 1). To perform EIT, a stimulating signal is injected through one pair of electrodes and the voltage across all sequential electrode-pair combinations measured. The electrical stimulations and measurements are noncontact to the cell. The perturbation in the conductivity distribution in the system is characterized by the voltage data, and images the cell are reconstructed using software termed "Electrical Impedance Tomography and Diffuse Optical Tomography Reconstruction Software," EIDORS [14]. The first step toward single cell impedance tomography was demonstrated by imaging a large multinuclear single-cell organism – the amoeboid plasmodium of the slime mold *Physarum polycephalum*, Fig. 1b.

EIT Chip Fabrication

The EIT chip contains 16 electrodes positioned in a circular pattern, and made from a printed circuit board (PCB). The electrode length and spacing are identical, the electrodes are 35 μm thick and the area of an individual electrode is 2.06 mm^2. A thin gold layer was galvanostatically plated onto the surface of the copper electrodes using a three-electrode potential station (PGSTAT128N, Autolab electrochemical instruments, the Netherlands) with a platinum-coated metal mesh counter electrode and Ag/AgCl electrode (REF321, Radiometer Analytical S. A., France) reference electrode. The thickness of the gold layers was 500 nm and controlled by adjusting the deposition time. To culture the plasmodium, a



Electrical Impedance Tomography for Single-Cell Imaging, Fig. 1 (**a**) Diagram showing the principle of electrical impedance tomography (EIT) for imaging single cell in culture in a circular electrode pattern. (**b**) Photograph showing *Physarum polycephalum* cultured (fed with oat flakes) on the surface of agar gel

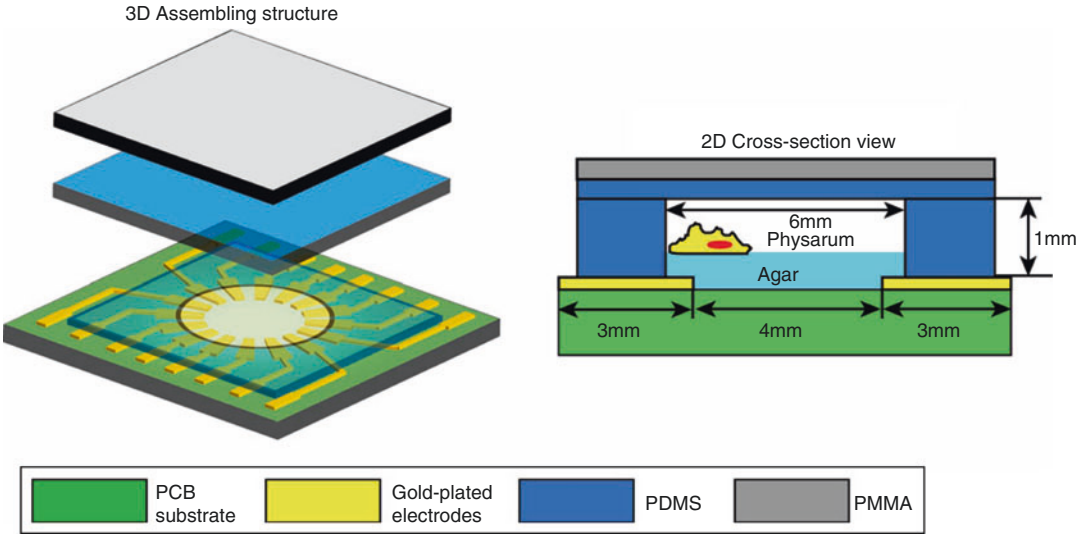
polydimethylsiloxane (PDMS) sheet (1 mm thick) with a 6 mm diameter hole was clamped over the EIT chip. A thin layer of agar gel was deposited over the surface of the electrodes, within the PDMS chamber, over which, the *Physarum polycephalum* was cultured. This avoids direct contact between the electrodes and the *Physarum polycephalum*. To eliminate evaporation, a second PDMS sheet and a polymethylmethacrylate (PMMA) plexiglass lid was used, as shown in Fig. 2.

System Instrumentation

Signal stimulations and measurements were performed using an Alpha-A impedance analyzer (Novocontrol Technologies, Germany) operating in four-electrode impedance measurement mode. A custom-designed electronic platform was built for multiplexed measurements, switching the pattern of stimulating signal and measuring voltages on different combinations of electrode pairs. Analogue chips, consisting of 16×8 crosspoint switches (Intersil CD22M3494) were used as multiplexers. The impedance analyzer and the multiplexing board were both controlled by a Linux PC through a GPIB-PCI interface (NI PCI-GPIB, National Instruments, USA) and a USB controller (FT2232D, FTDI Ltd. UK), respectively. Custom software was used for automated real-time measurements. The software sets the amplitude and frequency of the stimulating signal, the digital signals to the address lines of the multiplexers and controls the data acquisition. The highest temporal resolution of individual voltage measurement is 100 ms, which is limited by the data rate of the impedance analyzer.

EIT Measurement Scheme

The adjacent method was used for signal stimulation and voltage measurements. This method is more sensitive to variations in the impedance of objects near the electrodes than in the center of the chip. The stimulating signal is first applied through electrodes 1 and 2 (Fig. 1a) and the voltage measured from successive electrode pairs: 3–4, 4–5, . . ., 15–16, giving 13 voltage measurements. Note that voltages across electrode pair 16–1, 1–2, and 2–3 cannot be measured since the stimulating signal is injected from electrode pair 1–2. Then a second set of measurements is made with the current applied through electrodes 2 and 3 giving a further 13 sets of voltage data. This process is repeated, generating a total of $16 \times 13 = 208$ voltage measurements, which requires approximately 20 s for a full scan. Owing to reciprocity, only half of the measurements are linearly independent, all the data are used in the image reconstruction.



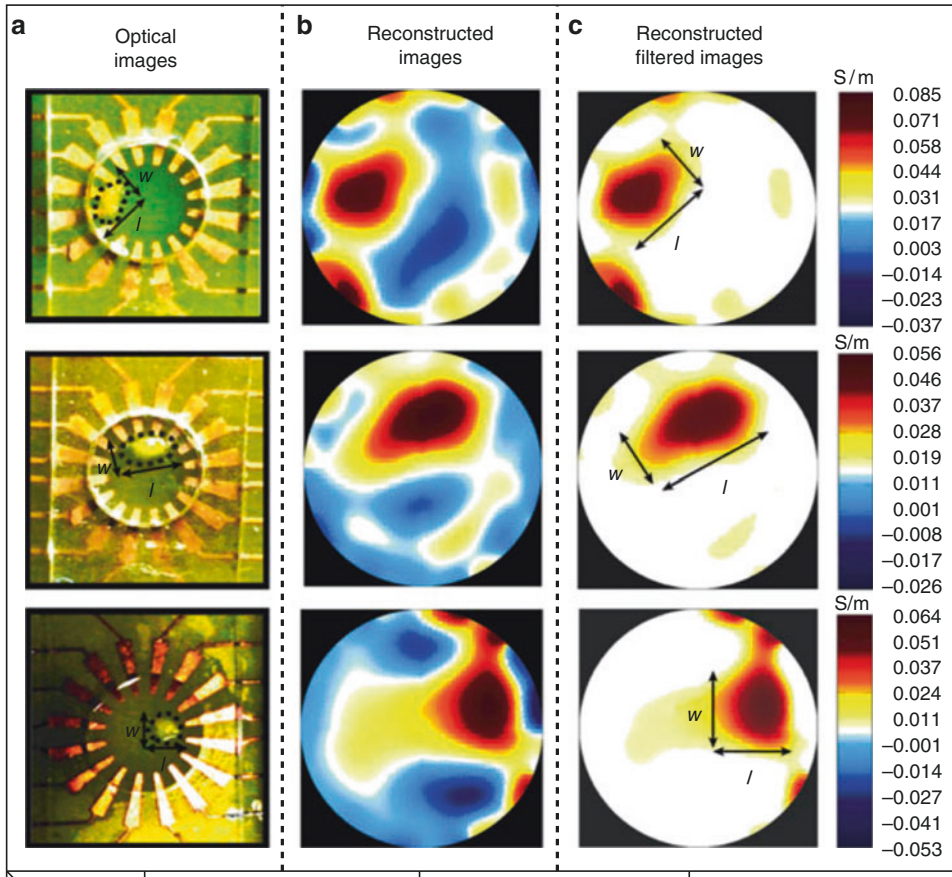
Electrical Impedance Tomography for Single-Cell Imaging, Fig. 2 Diagram showing the structure and cross section of the EIT chip used to image *Physarum* in culture

Physarum polycephalum in Culture

The plasmodium of the slime mold *Physarum polycephalum* is an amoeboid multinucleated single cellular organism. It consists of fanlike advancing fronts connected by tubular structures found in the posterior region (Fig. 1b). This organism has been long used as a model organism for the study of mitosis, membrane potential, and protoplasmic streaming. A *Physarum* plasmodium can extend to over several centimeters in diameter; nevertheless, it still maintains itself as a single cell by protoplasmic flow. Being a multinucleated cell, even if a cell is split into pieces, each piece can keep alive and act as single cell. It can be regarded as an autonomous distributed system and recently, its oscillatory behavior has been utilized directly for robotic control applications.

Results and Discussions

Different shapes of *Physarum polycephalum* were grown on the surface of the agar in different positions within the chip as shown in Fig. 3a. Two sets of voltage data were measured using the adjacent method (stimulating signal = 100 mV @ 200 kHz). One set of data was taken for pure agar gel in a closed EIT chip, providing a homogeneous background. The second set of data was measured after *Physarum* was grown on the agar. Images were reconstructed using 5,184 elements in the forward model. The initial value of *hyper-parameter* was set to 0.1 and adjusted in the range between 0.05 and 0.2 for smooth boundaries between *Physarum polycephalum* and the background. Figure 3b shows the corresponding reconstructed images from the voltage data. Brighter (*red*) corresponds to higher conductivity and darker (*blue*) to lower conductivity. The brighter region where the *Physarum* cell locates (positive values in the reconstructed differential conductivity) can be clearly seen. The mirror ghost images are equivalent to regions of lower conductivity (negative values in the reconstructed differential conductivity), and produce variations in the reconstructed conductivity values. Because the ghost images are low conductivity regions, and the difference between *Physarum* and the background is positive, the ghost images were eliminated by setting a threshold in the reconstructed conductivity, as shown in Fig. 3c. There are other artifacts, exhibiting both high and low conductivity regions, distributed along the edges of the



	Top physarum sample			Middle physarum sample			Bottom physarum sample		
	Optical	Electrical	Relative difference	Optical	Electrical	Relative difference	Optical	Electrical	Relative difference
Width (*w*)	1.41 mm	1.45 mm	2.8 %	1.61 mm	1.64 mm	1.8%	1.76 mm	1.82 mm	3.4%
Length (*l*)	2.47 mm	2.54 mm	2.8 %	3.19 mm	3.27 mm	2.5%	1.76 mm	1.82 mm	3.4%

Electrical Impedance Tomography for Single-Cell Imaging, Fig. 3 (**a**) Optical images of *Physarum polycephalum* on the agar gel in the EIT chips, (**b**) Corresponding reconstructed images. The *Physarum* cell is more conductive than the agar gel, exhibiting as bright (*red*) region in the reconstructed images. (**c**) Filtered images, the ghost images are eliminated by filtering the low conductivity regions

reconstructed images, which may come from the incompleteness of the one-dimensional segment model of the electrodes in the forward model.

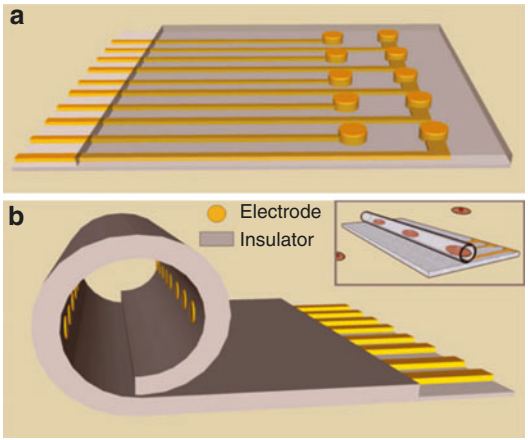
To analyze the resolution of the reconstructed images, the dimensions (the width and the length) of the *Physarum* cell from the optical and electrical images were compared. Dimensions of three samples (from top to bottom) are listed below the images in Fig. 3. The absolute difference in the dimensions between the optical and the reconstructed images ranges from 30 to 80 μm. The relative difference, calculated as the ratio of the absolute difference to the dimension in the optical image, shows that the minimum is 1.8 % and the maximum is 3.4 %. The average size of a single *Physarum* cell is approximately 2 mm in one dimension.

Conclusion and Outlook

The developed system would be useful in diagnostic and biomedical applications, such as cancer and cancer treatment monitoring for characterizing the spatial electrical properties of tumor cell colonies. Understanding the tumor growth is essential for screening programs, clinical trials, and epidemiological investigations. Skourou et al. [15] demonstrated the feasibility of using electrical impedance technique to detect small tumors (<3 mm) and tumor-associated changes. The tumor growth varies but is generally from a few mm to tens of mm [16], which falls into the imaging scale and resolution of the developed system. It can be expected that electrically mapping the electrical properties of tumor spatially will benefit the study in the dynamics in tumor cells spreading and proliferation.

Although impedance-based microfluidic cytometry [2, 3] has been widely used for cell manipulation and characterization using various approaches, the microfabrication of the devices leads to the difficulties in bonding the top and bottom of the microchannels, aligning the microfeatures (i.e., microelectrodes) exactly where they are required, fabricating the microelectrodes on the sidewalls of the microchannels, etc. A neat and smart technology is required to effectively address these problems in order to develop novel applications and push microfluidic cytometry into commercial markets, advocating point-of-care and on-side analysis.

A promising route is to use rolled-up microfabrication technology [17] to compensate the conventional microfabrication process. By using this technology, microchannels can be fabricated with a smooth inner/outer surface and a near to perfect circular cross section. Also, the microelectrodes can be located where they are needed with good control in position and alignment, thus offering the opportunities to address the above problems. Furthermore, the rolled-up technology can be applied to many materials and material combinations and thus it is easy to fabricate microchannels with customized surface chemical properties provided by the appropriate rolled-up materials.



Electrical Impedance Tomography for Single-Cell Imaging, Fig. 4 Schematic illustration of the proposed structure of rolled-up microspectroscope, integrated with pairs of microelectrodes for single cell imaging

However, up to now, only few reports can be found to apply this technology toward bio-applications. Very recently, researchers fabricated rolled-up transparent microtubes as confined culture scaffolds for individual yeast cells [18]. Good biocompatibility of rolled-up microtubes has been demonstrated by monitoring the growth of yeast cells inside the microtubes. It would be interesting to explore the rolled-up microtubes from microfluidic approaches in single cell analysis applications, where different patterns of microelectrodes will be integrated into the inner surface of the rolled-up microtubes for single cell manipulation and characterization under laminar microflow. Cells will be flown and trapped inside the microtube. Rolled-up multiple layers of circular electrode-arrays allows performing 3D single cell imaging [19], as shown in Fig. 4. The dimensions (diameter and length) of the microspectroscopes can be easily controlled by lithography, etching, and deposition techniques. Therefore, the tubular microspectroscopes can be designed for analyzing different sizes of biological cells to obtain the optimal signal. The microelectrodes can be fabricated exactly at the desired positions and also desired shapes; hence, by using conventional photolithography, different cell analyzing functions can be fabricated on a single chip that integrates

multifunctions for cell manipulation, identification, and imaging. Particularly, compared to the conventional microfluidic cytometry, the microelectrodes can be fabricated in a circular pattern inside the tubular microchannels. This allows many cell analysis applications to be extended from 2D to 3D, such as 3D single cell imaging using EIT by fabricating multiple sets of electrode-arrays inside the microtubes. Very recently, Caselli et al. [20] reported the simulation work of single cell imaging in a microfluidic cytometry using EIT. The authors reported detailed finite element modeling and numerical simulation results to demonstrate the feasibility of the device and method to extract information on cell morphology and to discriminate normal biconcave erythrocytes, spherocytes, and echynocytes. However, no device fabrication and experimental result has been reported yet. In summary, high-throughput and high-resolution microfluidic single cell impedance tomography is an important area that will continue to develop. It would be expected that 3D single cell imaging can lead to integrated point-of-care systems for biomedical and diagnostic applications.

Cross-References

- ▶ Electrical Impedance Cytometry
- ▶ Microfluidic Whole-Cell Biosensor

References

1. Whitesides, G.M.: The origins and the future of microfluidics. Nature **442**, 368–373 (2006)
2. Sun, T., Holmes, D., Gawad, S., Green, N.G., Morgan, H.: High speed multi-frequency impedance analysis for single particles in a microfluidic cytometer using maximum length sequences. Lab Chip **7**, 1034–1040 (2007)
3. Sun, T., Morgan, H.: Single-cell microfluidic impedance cytometry: a review. Microfluid. Nanofluid. **8**, 423–443 (2010)
4. Keese, C.R., Wegener, J., Walker, S.R., Giaever, I.: Electrical wound-healing assay for cells in vitro. Proc. Natl. Acad. Sci. U. S. A. **101**, 1554–1559 (2004)
5. Bayford, R.H.: Bioimpedance tomography (electrical impedance tomography). Ann. Rev. Biomed. Eng. **8**, 63–91 (2006)
6. Polydorides, N., Lionheart, W.R.B.: A matlab toolkit for three-dimensional electrical impedance tomography: a contribution to the electrical impedance and diffuse optical reconstruction software project. Meas. Sci. Technol. **13**, 1871–1883 (2002)
7. Cherepenin, V., Karpov, A., Korjenevsky, A., Kornienko, V., Mazaletskaya, A., Mazourov, D., Meister, D.: A 3D electrical impedance tomography (EIT) system for breast cancer detection. Physiol. Meas. **22**, 9–18 (2001)
8. Bagshaw, A.P., Liston, A.D., Bayford, R.H., Tizzard, A., Gibson, A.P., Tidswell, A.T., Sparkes, M.K., Dehghani, H., Binnie, C.D., Holder, D.S.: Validation of reconstruction algorithms for electrical impedance tomography of human brain function. Neuroimage **20**, 752–764 (2003)
9. Davalos, R., Otten, D.M., Mir, L.M., Runbinsky, B.: Electrical impedance tomography for imaging tissue electroporation. IEEE Trans. Biomed. Eng. **51**, 761–767 (2004)
10. Hou, T.-C., Loh, K.J., Lynch, J.P.: Spatial conductivity mapping of carbon nanotube composite thin films by electrical impedance tomography for sensing applications. Nanotechnology **18**, 315501 (2007)
11. Chai, K.T.C., Davies, J.H., Cumming, D.R.S.: Electrical impedance tomography for sensing with integrated microelectrodes on a CMOS microchip. Sens. Actuators B **127**, 97–101 (2007)
12. Linderholm, P., Marescot, L., Loke, M.H., Renaud, P.: Cell culture imaging using microimpedance tomography. IEEE Trans. Biomed. Eng. **55**, 138–146 (2008)
13. Sun, T., Tsuda, S., Zauner, K.-P., Morgan, H.: On-chip electrical impedance tomography for imaging biological cells. Biosens. Bioelectron. **25**, 1109–1115 (2010)
14. Adler, A., Lionheart, W.R.B.: Uses and abuses of EIDORS: an extensible software base for EIT. Physiol. Meas. **27**, S25–S42 (2006)
15. Skourou, C., Hoopes, P.J., Strawbridge, R.R., Paulsen, K.: Feasibility studies of electrical impedance spectroscopy of early tumor detection in rats. Physiol. Meas. **25**, 335–346 (2004)
16. Weedon-Fekjær, H., Lindqvist, B.H., Vatten, L.J., Aalen, O.O., Tretli, S.: Breast cancer tumor growth estimated through mammography screening data. Breast Cancer Res. **10**, R41 (2008)
17. Schmidt, O.G., Eberl, K.: Thin solid films roll up into nanotubes. Nature **410**, 168 (2001)
18. Huang, G.S., Mei, Y.F., Thurmer, D.J., Coric, E., Schmidt, O.G.: Rolled-up transparent microtubes as two-dimensionally confined culture scaffolds of individual yeast cells. Lab Chip **9**, 263–268 (2009)
19. Sun, T.: Microfluidic-based rolled-up 3D electrical impedance spectroscope for single cell analysis. Research Grant Proposal for Alexander von Humboldt Fellowship (unpublished material, 2009)
20. Caselli, F., Bisegna, P., Maceri, F.: EIT-inspired microfluidic cytometer for single-cell dielectric spectroscopy. J. Microelectromech. Syst. **19**, 1029–1040 (2010)

Electric-Field-Assisted Deterministic Nanowire Assembly

Theresa S. Mayer[1], Jeffrey S. Mayer[2] and Christine D. Keating[3]
[1]Department of Electrical Engineering and Materials Science and Engineering, The Pennsylvania State University, University Park, PA, USA
[2]Department of Electrical Engineering, Penn State University, University Park, PA, USA
[3]Department of Chemistry, Penn State University, University Park, PA, USA

Synonyms

Dielectrophoretic assembly; Directed assembly

Definition

Electric-field-assisted deterministic assembly uses the force induced on individual nanowires by a nonuniform electric field to position each nanowire with submicron registration accuracy relative to previously defined nanodevices on a substrate.

Overview

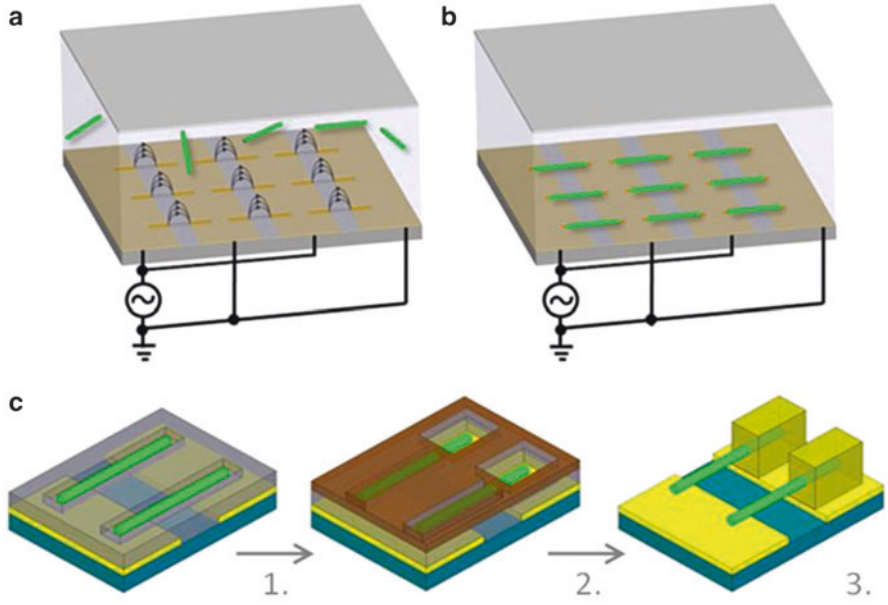
The electric-field force induced on solution-suspended, polarizable nanowires by a nonuniform alternating electric field can be used to attract, align, and position individual nanowires at predefined locations on a patterned substrate (e.g., Si integrated circuit) [1–4]. The magnitude and direction of the spatially varying electric-field force, which together control the final position and orientation of the nanowires, are determined by electrode structures on the substrate, bias voltages applied to the electrodes, the geometry and material properties of the nanowire, and material properties of the solution (Fig. 1a). The force field can be tailored to provide submicron registration (overlay) accuracy between each assembled nanowire and a specific nanoscale feature on the substrate at integration densities $>10^6$ nanowires/cm^2 and with yields >95 % [5]. Additionally, different nanowire populations (e.g., materials, coatings, etc.) can be directed to particular regions of the substrate by synchronizing the sequential delivery of each population with a unique electric-field force pattern [6]. Once assembled, the nanowires can be converted into one of many different device types – nanomechanical, electronic, optoelectronic – by adding mechanical and/or electrical contacts using conventional lithographic, deposition, and etching processes (Fig. 1b) [5, 6].

Basic Methodology

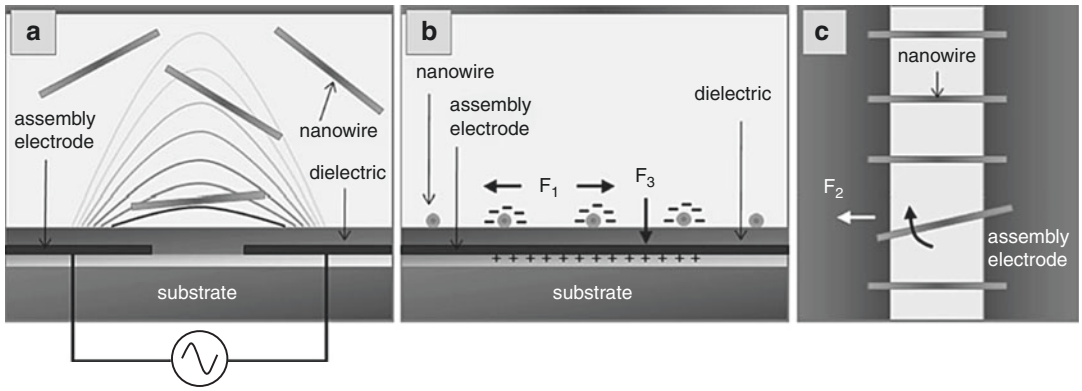
The electric-field-assisted deterministic nanowire assembly process can be divided into two phases: (1) attraction of the solution-suspended nanowires to a predefined area on the substrate and (2) alignment and positioning of the nanowires relative to other assembled wires or lithographic features on the substrate. In the first phase, the nanowires are suspended in the fluid far above the dielectric-coated substrate in a region where the gradient of the alternating electric field produced by the biased electrodes is relatively small but sufficient to polarize the individual nanowires. Each of the polarized nanowires experiences a force that is well modeled using the dielectrophoretic (DEP) force equation [7]:

$$\vec{F}_{DEP} = \left(\vec{p}.\nabla\right)\vec{E}\ ,$$

where $\vec{p}$ is the effective dipole moment of the particle resulting from its polarization by the electric field $\vec{E}$. More specifically, numerical solutions of the force using the Maxwell stress tensor (MST) method show that it is proportional to $\nabla\left|\vec{E}\right|^2$ [8], which is the same field dependence as $\vec{F}_{DEP}$ for a spherical dielectric particle analyzed by Pohl [7]. In general, the force is significantly larger than the other hydrodynamic, gravitational, and Brownian forces acting on the nanowires, and causes the wires to



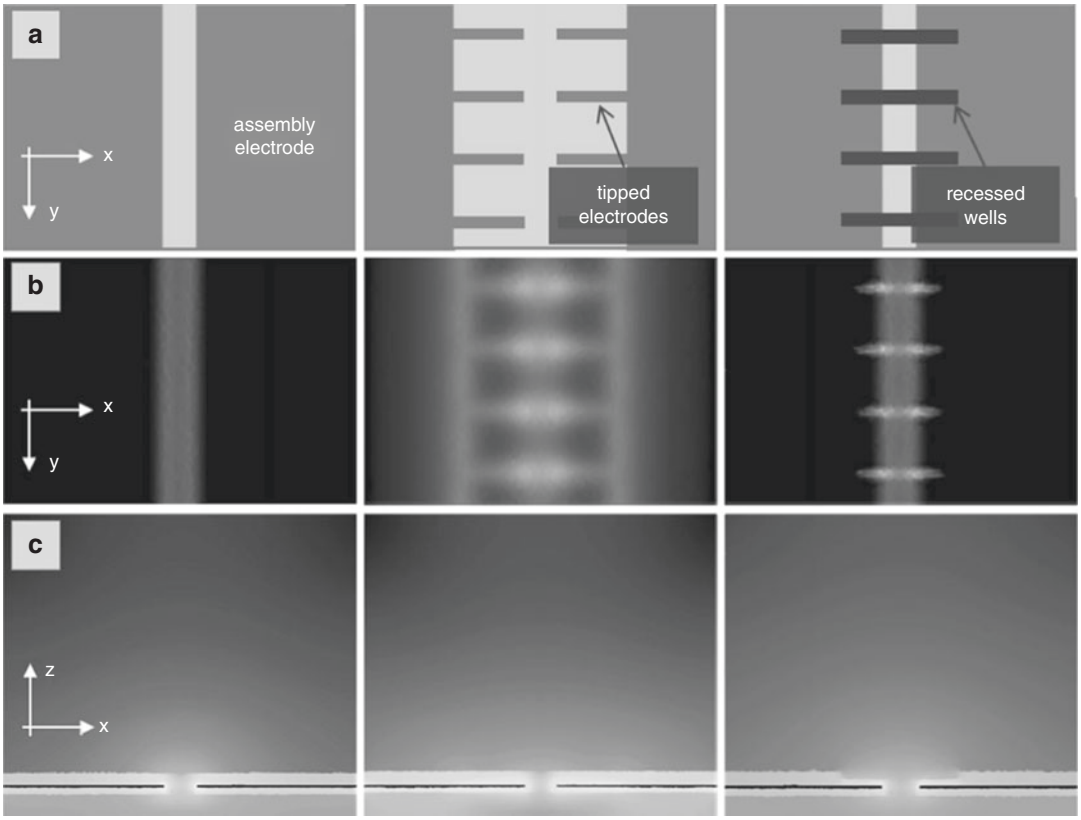
Electric-Field-Assisted Deterministic Nanowire Assembly, Fig. 1 *Deterministic nanowire assembly.* (**a**) Nonuniform electric fields originating from electrically biased electrode structures on the substrate extend into the fluid medium. (**b**) Electric-field forces acting on the solution-suspended nanowires are used to position individual nanowires within lithographic wells on the substrate. (**c**) Nanowires assembled in the wells are converted into cantilever devices by patterning windows to contact one end of each wire (*step 1*), electroplating metal clamps (*step 2*), and dissolving the sacrificial photoresist (*step 3*) (After Ref. [5])



Electric-Field-Assisted Deterministic Nanowire Assembly, Fig. 2 *Deterministic nanowire assembly.* (**a**) Cross section of interdigitated electrode structure showing electric field lines. *Dark gray* designates the highest and light gray the lowest electric field magnitude. The nanowires align tangent to the field lines and are attracted to the high field region in the gap between the electrodes. (**b**, **c**) Electric-field forces due to like charge on the wire tips (*F1*), opposite charge on the wires and on the biased metal electrode (*F2*, *F3*). Together these forces form uniformly spaced, aligned nanowire arrays

align tangent to the electric field lines and to move in the direction of the highest field gradient. For a simple interdigitated electrode structure, the nanowires are attracted from solution and are aligned such that they span the gap separating the biased electrodes (Fig. 2a).

In the second phase of the assembly process, the nanowires are on the substrate, and the



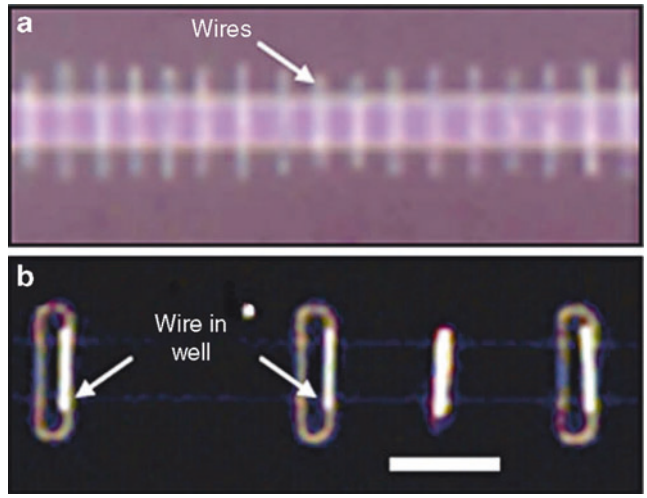
Electric-Field-Assisted Deterministic Nanowire Assembly, Fig. 3 *Electrode structures*. (**a**) *Top and cross-sectional views* of the three different electrode structures all having a gap width of 5 μm. The sharp tipped electrodes and recessed wells are 1 μm wide and 10 μm long. (**b**) *Top view* of the electric field magnitude measured at the dielectric coating/solution interface. (**c**) *Cross-sectional view* of the electric field magnitude (plotted on a log scale) measured along the *center* of the tipped electrodes and recessed wells. *Dark gray* designates the highest and *light gray* the lowest electric field magnitude

induced charge distribution and electric field are more complex, resulting in forces between neighboring wires as well as between these same wires and the buried electrodes (Fig. 2b, c). In the first case, the like charge on the nanowire tips gives rise to a net repulsive force between neighboring wires. This results in the formation of uniformly spaced arrays at high wire concentration. In the second case, if an aligned nanowire is offset from the center of the gap, the opposite charge on the wire and on the metal electrodes produces a lateral restoring force and torque. This causes the wires to align, centered across the gap, with end-to-end registration. It also fixes the position of the wires on the substrate when the solution is removed. As described in the next section "Key Research Findings," by tailoring the electrode structure design, the electric-field profile and forces the can be optimized to position individual wires with submicron registration accuracy relative to previously defined nanodevices on the substrate.

Key Research Findings

Assembly Electrodes and Electric-Field Profiles

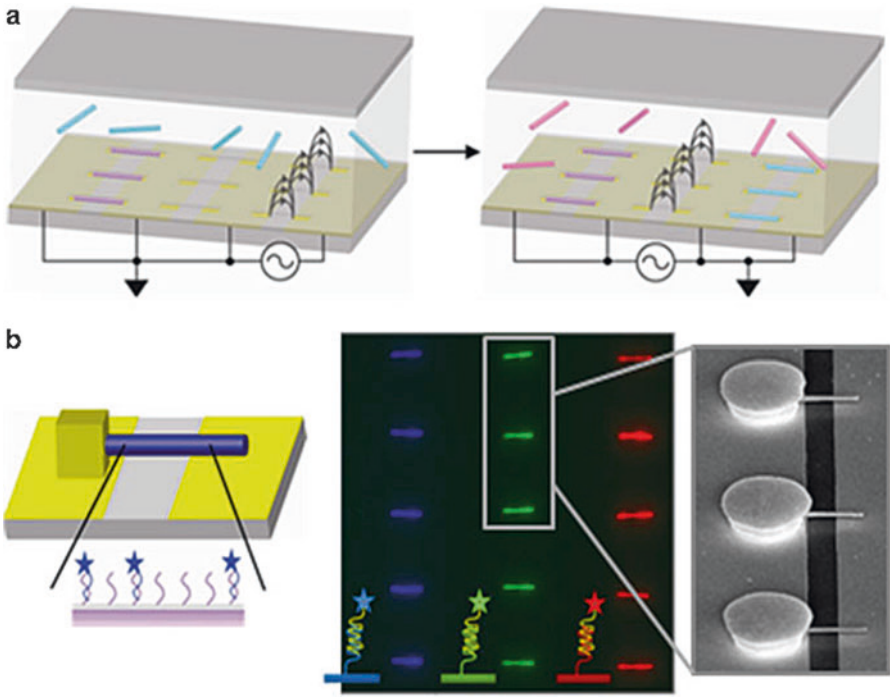
The electric-field forces used for nanowire assembly are defined by the nonuniform field originating from the electrically biased electrode structure. Three different structures are presented in Fig. 3 to illustrate the range of field profiles that



Electric-Field-Assisted Deterministic Nanowire Assembly, Fig. 4 *Nanowire assembly.* Dark field optical microscope images of 200 nm diameter, ~7 μm long Rh nanowires assembled on electrodes with (**a**) a uniform-width gap, and (**b**) recessed wells that span a uniform-width gap. The wires were suspended in isopropanol for assembly, and a 10 kHz, 7 V_{rms} sinusoidal voltage was applied across the interdigitated assembly electrodes. Uniformly spaced, well-aligned nanowire arrays were assembled using the uniform-gap electrode structure. Deterministic positioning of individual nanowires was achieved using recessed wells. Scale bar = 7 μm

can be realized: (1) electrodes with a uniform-width gap and a planar dielectric coating, (2) electrodes with sharp tips and a planar dielectric coating, and (3) electrodes with a uniform-width gap and a dielectric coating patterned with wells that span the gap. Figure 3 also shows simulations of the electric field magnitude (COMSOL Multiphysics) measured at the interface between the dielectric coating and the solution (x–y plane, Fig. 3b) and through the cross section of the structure (x–z plane, Fig. 3c). For the first structure, the field strength is highest at the edge of the electrodes, and it is uniform along the length of the gap. This gives field lines that radiate outward from the gap, with a field gradient that is the largest near the substrate surface. Optical microscope images taken after electric-field-assisted assembly of template-synthesized Rh nanowires (200-nm diameter, 7-μm long wires; Fig. 4a) demonstrate that this field profile leads to the formation of parallel wire arrays along each gap. The spacing between adjacent nanowires is uniform, but the absolute position of each wire in the array is not well controlled. This limits the registration accuracy of the assembled wires relative to existing device features on the substrate.

The next two electrode structures overcome this limitation by producing spatially confined regions of high field strength using lithographic features. For the sharp-tipped electrode structure, the electric field magnitude is the strongest in the region surrounding the tips, with a field gradient that is the largest in the gap between the tips. This results in individual nanowires preferentially assembling across the gap, with additional wires aligning in the fringing field at an angle relative to the gap. A similar field profile can be obtained by defining an array of recessed rectangular wells in the dielectric coating that separates the metal electrodes from the solution [5]. This structure has the advantage that it suppresses the lateral field between adjacent high field regions, which ensures that individual wires align to span the gap within the well (Fig. 4b). It is also possible to minimize the assembly of multiple wires on the sharp-tipped electrodes by adjusting the fluid flow to control the hydrodynamic forces acting on the wires during assembly [4]. Importantly, submicron registration accuracy between the assembled nanowire and features on the substrate can be achieved with both of these electrode structures.



Electric-Field-Assisted Deterministic Nanowire Assembly, Fig. 5 *Programmed nanowire assembly.* (**a**) Scheme illustrating programmed deterministic assembly of different populations of bioprobe-coated nanowire. Each population is directed to the recessed wells in the electrically biased regions of the substrate. (**b**) Following programmed assembly of three populations of bioprobe-coated wires, nanomechanical cantilever devices (*left*) were fabricated as described in Fig. 1. Fluorescence images (*center*) show that the fluorescently labeled targets hybridized selectively to their complementary probes. The field emission scanning electron microscope image (*right*) shows the fabricated nanocantilever devices (After Ref. [6])

Programmed Assembly of Different Nanowire Populations

Electric-field-assisted assembly can be used to deterministically position more than one type of nanowire on the same substrate. As illustrated in Fig. 5a, this is accomplished by synchronizing the sequential delivery of each nanowire population with a unique electric-field profile obtained by biasing a specific set of assembly electrodes. As a proof of concept, this programmed assembly approach was applied to create nanowire arrays containing three populations of wires, each functionalized with different nucleic acid probe molecules and positioned in different columns in the array [6]. Following assembly, the three types of wires were converted into nanomechanical cantilevers using conventional lithography and metal deposition processes. The probe-coated nanocantilever devices were incubated with a mixture of all three fluorescently labeled nucleic acid targets. Fluorescence microscopy images of the three column array (Fig. 5b) confirmed that the target molecules selectively hybridized to their complementary targets and that the assembly process positioned each nanowire population in the correct array column.

Examples of Applications

Adding heterogeneous populations of nanowire devices onto Si integrated circuits (ICs) will enable new and diverse functions ranging from energy harvesting to chemical and biological sensing [9]. Deterministic assembly offers a flexible and customizable alternative to traditional top-down nanofabrication because it completely decouples the nanowire synthesis process from

the IC manufacturing process. This allows the use of emerging template, solution, and catalyst-based protocols for inexpensive batch synthesis of diverse materials such as metals, semiconductors, metal oxides, conducting polymers, and molecule-coated wires [10]. Off-chip wire synthesis also removes the constraints of thermal budget and chemical compatibility imposed by top-down manufacturing methods. This permits the integration of nanowires with properties optimized for a specific function.

Future Directions for Research

Advancing the electric-field-assisted deterministic assembly technique for use in practical applications will require ongoing research to develop new and scalable nanomanufacturing strategies that can provide high-yield integration of nanowire devices with the required registration accuracy over large substrate areas. For Si-based systems-in-a-package (SiP), the nanowires should be assembled at the back end of the Si IC process. This involves a highly integrated, interdisciplinary research effort. For example, the availability of batch-synthesized nanowires with uniform properties (i.e., dimensions, microstructure, etc.) is essential to the success of this approach. Similarly, scalable methods should be developed to deliver aggregate-free, solution-suspended nanowires to large-area substrates with controlled and reproducible concentration. Additional research aimed at developing a comprehensive theoretical and experimental understanding of the electric-field forces will enable the design of new and more effective assembly electrode structures with patterns customized for different circuit applications. Finally, innovative nanofabrication methods should be applied to convert the assembled nanowires into high-performance electronic, optical, and sensing devices that can be monolithically integrated with the Si circuitry.

Cross-References

► Dielectrophoresis

References

1. Smith, P., Nordquist, C., Jackson, T.N., Mayer, T.S., Martin, B., Mbindyo, J., Mallouk, T.E.: Electric-field assisted assembly and alignment of metallic nanowires. Appl Phys Lett **77**, 1399–1401 (2000)
2. Duan, X., Huang, Y., Cui, Y., Wang, J., Lieber, C.M.: Indium phosphide nanowires as building blocks for nanoscale electronic and optoelectronic devices. Nature **409**, 66–69 (2001)
3. Raychaudhuri, S., Dayeh, S., Wang, D., Yu, E.: Precise semiconductor nanowire placement through dielectrophoresis. Nano Lett **9**, 2260–2266 (2009)
4. Freer, E., Grachev, O., Duan, X., Martin, S., Stumbo, D.: High-yield self-limiting single-nanowire assembly with dielectrophoresis. Nat Nanotechnol **5**, 525–530 (2010)
5. Li, M., Bhiladvala, R., Morrow, T., Sioss, J., Lew, K. K., Redwing, J.M., Keating, C.D., Mayer, T.S.: Bottom-up assembly of large-area nanowire resonator arrays. Nat Nanotechnol **3**(2), 88–92 (2008)
6. Morrow, T., Li, M., Kim, J., Mayer, T.S., Keating, C. D.: Programmed assembly of DNA-coated nanowire devices. Science **323**(5912), 352 (2009)
7. Pohl, H.: Dielectrophoresis: The Behavior of Neutral Matter in Nonuniform Electric Fields. Cambridge University Press, Cambridge/New York (1978)
8. Liu, Y., Chung, J., Liu, W., Ruoff, R.: Dielectrophoretic assembly of nanowires. J Phys Chem B **110**, 14098–14106 (2006)
9. International Technology Roadmap for Semiconductors: Emerging research devices. In: International Technology Roadmap for Semiconductors, 2009 edn. International Technology Roadmap for Semiconductors (2009)
10. Xia, Y., Yang, P., Sun, Y., Wu, Y., Mayers, B., Gates, B., Yin, Y., Kim, F., Yan, H.: One-dimensional nanostructures: synthesis, characterization, and applications. Adv Mater **15**(5), 353–389 (2003)

Electrocapillarity

► Electrowetting

Electrochemical Energy Storage and Conversion

► Electrochemical Interfaces for Energy Storage and Conversion

Electrochemical Interfaces for Energy Storage and Conversion

Clotilde S. Cucinotta[1] and Monica Kosa[2]
[1]School of Physics, Trinity College, Dublin, Ireland
[2]Department of Chemistry, Faculty of Exact Sciences, Bar Ilan University, Ramat Gan, Israel

Synonyms

Computer modeling of electrochemical systems; DFT-based simulations of electrified interfaces; Electrochemical energy storage and conversion; Electrodics from first principles; Modeling of batteries and fuel cells; Parameter-free study of electrochemical interfaces

Definition

Under equilibrium conditions, a bulk electrolyte is an electroneutral, isotropic, and homogeneous solution, and there are no electric fields along any preferential direction. However, at the frontier with another material with mobile free charges (ions or electrons) and a different chemical potential – e.g., the electrode – some anisotropy in the forces experienced by the particles arises, and charge redistribution occurs. Although the interfacial region remains overall neutral, either side of the boundary becomes charged to an equal and opposite extent, forming the so-called electrical double layer or electrified interface. This gives rise to a potential difference across the interface. This potential difference is usually not large (~1 V); however, the dimension of the interphase region is very small ($\sim 10^{-7}$ cm) with the consequence that the field gradient is enormous ($\sim 10^{7}$ V/cm). The potential difference across the electrified interface and the structure of the double layer are intimately related. The fundamental problem in electrochemical studies is to unravel this functional relationship and the effect of the field at the electrode–electrolyte interface.

Introduction

Electrochemistry represents the general framework for a wide variety of processes and phenomena, playing a crucial role in biology, materials science, and technology. The attention it receives from the scientific community is rapidly increasing because of its crucial role in a variety of energy generation and storage devices such as fuel cells, batteries, solar cells, and supercapacitors.

A vast amount of empirical data on electrochemical (EC) systems have been accumulated during the more than 150 years since Faraday's groundbreaking work. Not surprisingly, electrochemistry and its general applications in electron transfer reactions have given origin to a number of fundamental theoretical and computational models, such as the Debye–Huckel and Marcus theory.

The unifying motif of almost all (EC) systems – such as fuel cells, batteries, solar cells, and supercapacitors – is the coupled flow of matter and charge across the electrified interfaces (EI) between electrode and electrolyte. EC reactions, which usually produce or are driven by an electric current, ultimately dictate the behavior of most EC devices. Microscopic processes of this kind occur, for instance, in electrolytic cells, where water can be split into hydrogen and oxygen due to an electrical energy supply, or in batteries and fuel cells, where an electrical energy is derived from chemical reactions taking place within the cell. Understanding the microscopic details of these interfaces and of the processes taking place in their proximity, therefore, is a key to their optimization, enabling them to fulfill their promises for cleaner, cheaper, and safer energy technologies.

Experimental methods such as STM (scanning tunneling microscope), AFM (atomic force microscopy), SFG (sum frequency generation), and NF (near-field) spectroscopy sometimes achieve atomistic resolution. Most experiments, however, can only observe the macroscopic manifestation of EC conversion processes while the detail of their microscopic behavior remains vague. This implies that computer simulations,

which have the power to resolve processes atom by atom, are required to magnify our view of EC systems and should be used to validate experimental results, to design new experiments, and to predict exciting new effects of potential technological interest.

However, in EC systems, there is still a wide gap between theoretical understanding of microscopic phenomena and the macroscopic outcome of experiments.

Most of the current computational models target a single aspect of EC systems or phenomena, such as the electrostatic double layer (DL), the electronic structure at the electrode–electrolyte interface, the electron or ion transfer processes taking place at the electrode, the ion diffusion in electrolyte, etc. A unifying approach, able to cover the statistical mechanics and thermodynamics as well as material science or electronic structure of these complex systems, is still missing.

Such a comprehensive description can in principle be provided by ab initio simulations, relying on density functional theory (DFT) in the Kohn–Sham (KS) formulation [1] and able to simultaneously account for thermal motion, along with the changes in the electronic structure of the system. The full exploitation of this approach in electrochemistry, however, is still hampered by substantial difficulties [2].

Open issues and problems in electrochemistry target different levels of theory and applications, requiring the modeling of:

1. Inhomogeneous systems, with ionic species in solution and a variety of chemical bonds, from metallic or covalent to hydrogen bonding and nonbonded interactions going from long-range Coulomb to short-range but still important dispersion forces. The correct reproduction of the quantum chemistry of energy alignment at EI for this variety of interactions may require the introduction of correction terms to local or semilocal DFT potentials or the adoption of schemes beyond standard DFT. Moreover, in current materials used in already commercialized lithium-ion batteries and emerging photovoltaic cell technologies, for example, the charge conduction is governed by polarons, which are difficult to describe with standard DFT methods.
2. Complex adsorption, catalytic, and charge and mass transfer phenomena, occurring at the electrode–electrolyte interface. Liquid electrolytes require evaluation of the surface free energy rather than simple internal energy. Dynamical techniques, with their associated computational cost, are thus needed.
3. Double-layer formation/dynamics and the effect of varying the electrode potentials, affecting structure and phenomena occurring at EI. These include effective mass and charge dynamical reorganization, work function modifications, and energy flow at the electrode surfaces. Moreover, phase transition and volume changes are often induced by the EC process. In general, when complex forms of organization are the result of nonlinear interactions between their basic building blocks, the development of multiscale approaches, combining atomistic modeling with a continuum description, is advantageous.

All these aspects together imply large simulation samples and long equilibration times for finite temperature simulations and, in the case of charged electrodes, may require rather artificial assumptions on the distribution of charge in the system.

Theoretical Surface Science Approaches

There is a long history of theoretical, DFT-based surface science [3] – often in combination with experiments – resulting in a deep insight into the structure and reactions on metal and oxide surfaces as well as comprehensive studies on electronic, structural, and dynamical properties in energy materials [4].

An established and very popular model to predict reactivity and catalytic performance of different electrodes by means of simple descriptors was proposed by Hammer and Nørskov in 1995 [5]. In this model, adsorption energies are related to *d*-band position in the metal and to barriers in catalytic reactions. For instance, oxygen reduction

activity – which can be calculated in terms of activation energies, using the concept of Sabatier analysis – was related in a volcano plot to oxygen binding energy on different metal surfaces. The position of the center of *d*-electrons density with respect to metal Fermi level in some cases correlates linearly with oxygen atom chemisorption energy. However, this is only true if the metals under consideration are very similar. Even with this caveat, this approach has proven to be extremely useful, especially when oxygen binding energy is used as a parameter in describing the catalytic performance. However, it has to be said that using this approach, the complicated processes governing EC conversion are oversimplified and electrolyte presence is not considered explicitly.

Unraveling the atomistic structure and dynamics of metal–water interfaces is of crucial interest from both a fundamental and an applied point of view, in phenomena and technological processes such as corrosion, lubrication, heterogeneous catalysis, and electrochemistry. The established picture for water at this interface is of an icelike structure with enhanced density close to the electrode. These metal–water interfaces have been traditionally modeled statically, using the water bilayer model, with every second water molecule oriented parallel to the surface and the other molecules pointing one H atom either toward (H-down configuration) or away from the surface (H-up configuration). More recent studies [6] have shown that a remarkably rich variety of equilibrium structures, including dissociated or not hexagonal water rings, can form at the interface with even simple, flat surfaces, depending on the strain acting on the water layer and, in general, on the nature of the substrate. To address the open issues in electrode/electrolyte interface modeling, the theoretical surface science community has followed a stepwise approach based on first principles by considering systems with increasing complexity [7]. For instance, chemical trends in adsorbed halides- and water-induced work function changes have been correlated with water–metal interaction strength, charge redistribution at the interface, and metal electronegativity.

While an important progress has been achieved due to these static approaches, the equilibrium structure of metal–electrolyte interfaces in ambient condition is yet to be unraveled.

To take into account the effect of temperature in the modeling of the interface between electrode and water, numerous molecular dynamics (MD) studies [7, 8] used empirical and classical force fields models. The main limitation of most of these studies is that they do not explicitly include charge polarization effects as well as do not describe breaking and formation of chemical bonds and charge transfer at interfaces and in bulk electrolytes.

The only approaches suitable to describe the subtle interplay of electronic, ionic, and thermal effects are those based on first principles, especially when coupled with MD. Among those, DFT represents a very good compromise between accuracy and numerical efficiency. Recently, the increased availability of computational resources allowed a few first-principles MD studies to treat explicit water [8].

More recently, a method was devised to investigate coupled proton–electron transfer reactions [9]. This approach is based on ab initio MD but relies on foundations that might be seen as intermediate between simulation and theory. The half-reactions consisting of the addition/removal of protons/electrons are simulated in a homogeneous solution environment. Quantitative contact with electrochemistry is established by measuring free energies with respect to a common reference point, corresponding to the free energy of an aqueous hydronium ion H_3O^+. Once the thermochemical corrections are applied, free energy differences obtained by this method do correspond to the ΔG_0 measured with respect to the normal hydrogen electrode (NHE). This method is sophisticated and powerful in describing the thermodynamic bulk properties, but it lacks a comprehensive description of the metal/electrolyte interface that is central to electrochemistry. Perhaps more importantly, it does not provide a description of the overpotential related to the transfer of species to, from, or across the interface, which generally could be useful in determining the efficiency of the electrode and catalyst.

Finally, the method cannot be extended easily to real-time, real-space formulations of time-dependent DFT, which will be eventually required to describe the full kinetics of the system.

First-principles methods and MD are crucial for describing fundamental aspects of EC conversion, where chemical and electronic structure transformations occur. However, the complexity of these approaches and their high computational cost still restrict most of the existing studies to short simulation times and relatively small simulation cells, reducing the statistical validity of their conclusions.

A primary example of the currently limited ability to describe atomic and electronic scale processes at EI is our limited knowledge of the structure of the equilibrium DL and the effect of an external bias applied to the electrolyte through the electrodes even for the most basic metal–electrolyte interfaces. Secondly, the computed activation barriers for charge transfer at interfaces are usually not accurate (see however Ref. [4]) due to neglecting the effect of the environment.

In the following, we will present in some detail the methodological open issue of modeling the effect of the application of an external potential on the electrode and some of the proposed solutions.

In the final part of this entry, we will speak about modeling EC devices such as batteries and fuel cells, presenting some recent examples of the *state of the art* first-principles modeling of these devices.

Modeling the Effect of the Variation of the Electrode Potential: Conceptual Overview and Applications

The interest and the challenge in modern simulation of EC processes consist in representing the effect of variation of the electrostatic potential and thus the strong surface charge density redistribution and the fluctuating electric field at the electrode–electrolyte interface.

In principle, to describe an EC process a non-equilibrium approach should be used (such as the non-equilibrium Green's function formalism used in molecular electronics) and the open boundary conditions for describing the electrode, electrolyte, and chemical transformations occurring therein. Moreover, different timescales, ranging from that of slow solvent rearrangements to that of electron transfer up to fast solvent polarization effects, should be treated at the same level. These difficulties are the reason for the delay in advance of theoretical material science in this area, and only recently, the growth of computer power enabled treating such systems using DFT. The extensive computational power when using such approach, however, is still not available for the wide scientific community to be used in ordinary calculations.

The presence of an electrode potential distinguishes electrochemistry from catalysis and affects EC reactions by:

- Determining the chemical potentials of the electrons involved in half-reactions at the electrode surface
- Inducing the formation of a DL at the EI and a very strong electric field affecting the phenomena occurring therein

The simplest models of half-cell reactive processes use atomic clusters in gas phase and external parameters to simulate experimental equilibrium potential. To take into account the effect of an electric field on adsorption behavior at electrodes, a homogeneous electric field is sometimes applied perpendicularly to the cluster surface. The electrode potential is tuned by varying the external applied field; yet the relation of the EC potential to a known reference remains unclear.

Most of the *standard* simulation methods resort to periodic boundary conditions (PBC) to speed up the computation and minimize the effect of finite sample sizes, thus mimicking an infinite perfect crystal. Adopting periodic slabs provides a good description of the electronic and geometric structures, crucial for describing adsorption processes and relevant EC quantities at EI. However, the use of PBC makes it difficult to apply an electric field across the system, as required to

simulate a system under bias. The supercell geometry, coupled with current methods to account for long-range Coulomb interactions (Ewald and variations thereof), implies, in turn, that the system must be globally neutral. Previous models described the charge on the electrode by including additional occupied KS orbitals and/or resorting to fractional occupation numbers. In other cases, neutrality is enforced by adding a judicious distribution of fixed countercharges [7]. These could take the form of a homogeneous background charge uniformly distributed over the simulation cell [10], charged planes [11], a dipole layer into the vacuum region [12], or hydronium ions [13] placed at a fixed distance from the electrode surface in order to charge the electrode.

The conceptual evolution of these approaches will be now critically highlighted in some detail.

In 2004, Norskov demonstrated [13] that even for globally neutral supercells, electron transfer between electrodes can be effectively included in half-cell reaction free energies by adding a term which depends on the applied potential U as follows:

$$\Delta G = G_{\text{products}} - G_{\text{reactants}} - neU,$$

where e is the charge of an electron and n is the number of exchanged electrons between reactants and products. This method was applied for calculating reaction energies for O_2 reduction at fuel cell cathodes as a function of the applied EC potential, by establishing a computational reference to the hydrogen electrode at standard conditions, when hydrogen evolution is at equilibrium ($H_2 \leftrightarrow 2(H^+ + e^-)$). In this approach, similar to the thermodynamic approaches used in electrocatalysis, the electrode potential only affects Fermi level electrons, while the EC environment is considered as the adsorbate reservoir. The charge polarization at the electrode due to the induced electric field and the explicit effect of the EC environment on the interfacial processes is disregarded. Despite this severe limitation, the method is computationally viable and allowed establishing trends in electro-catalysis.

Subsequently, a dipole layer was introduced into the vacuum region to induce electrode polarization – thus accounting for the effect of an external field. The main challenge in this case is to know the strength of the electric field induced by the applied potential. The authors evaluated this contribution assuming plate capacitor behavior, $|E| = U/d$, and guesstimating the Helmholtz layer thickness.

Charging up the slabs leads effectively to a variation in the electrode potential. Since in periodic calculations the unit cell has to be neutral, a countercharge is needed to maintain cell neutrality. We will illustrate this issue in more detail in the following. For the moment, we note that once the slab is allowed to take on a net charge, it is possible to set the electrode charge also implicitly, by fixing the chemical potential of the electrons associated with the slab. Along this line, Lozovoi [11] presented a grand-canonical formulation of charged slabs, where a chemical potential, μ, is defined and kept fixed relative to a reference electrode at finite distance from the slab. However, restraining μ is often less efficient than varying the charge in the slab until the value corresponding to a desired μ is found.

The most important contribution along this line was provided in 2006 by Filhol [10]. In Filhol's approach, a net surface charge q is introduced onto the electrode via the variation of the total number of electrons present in the slab. Cell neutrality is enforced adding a constant countercharge background $-q/W$, where W is the cell volume (this countercharge is by default present in standard DFT codes in case of charged calculations in order to avoid electrostatic divergence). The potential drop at the EI induced by the presence of the charged electrode is quantified by comparing the Fermi level of the system to a reference potential at a defined distance from the metal/water interface.

Two potential references are defined, a vacuum and an aqueous state (double reference method). The first reference is the vacuum level for the uncharged system, giving the zero of the energy scale:

$$\phi_{0\text{bulk}} = \phi_{\text{bulk}} = \phi'_{\text{bulk}} - \phi'_{\text{vacuum}}.$$

The metal/electrolyte potential (no vacuum present) is aligned to this vacuum level by

aligning the bulk potential levels of the systems with and without vacuum. The second reference is the water potential, a fixed region far from the electrode at its position in the calculation at zero charge. Energy levels of the charged and relaxed system are aligned to this fixed water level:

$$\phi_{\text{charged}} = \phi'_{\text{charged}} - \phi'_{\text{water-charged}} + \phi'_{\text{water-uncharged}}.$$

The absolute reference is given by the NHE:

$$NHE = -4.85\text{eV} - \phi_{\text{charged}}.$$

The presence in this model of a constant countercharge background, mimicking the effect of the counterions, introduces a parabolic contribution to the potential outside the electrode. Interestingly, the artifacts arising from this distribution are reduced at the presence of polarizable water layers. However, since the counterions are modeled as uniformly spread in the electrolyte, rather than localized in the vicinity of the electrode surface, the electric field just outside the surface is too small.

An alternative solution was developed in 2007 by Rossmeisl and Norskov [13], who proposed a model where counterions are explicitly considered and the electrode potential is varied by varying the number of protons/electrons in the DL. Excess H atoms are thus included at fixed positions in the water layer, which should result in formation of solvated protons and transfer of electrons to the metal.

Free energies of solid/liquid interfaces with N metal atoms and n protons in the water layer are calculated as

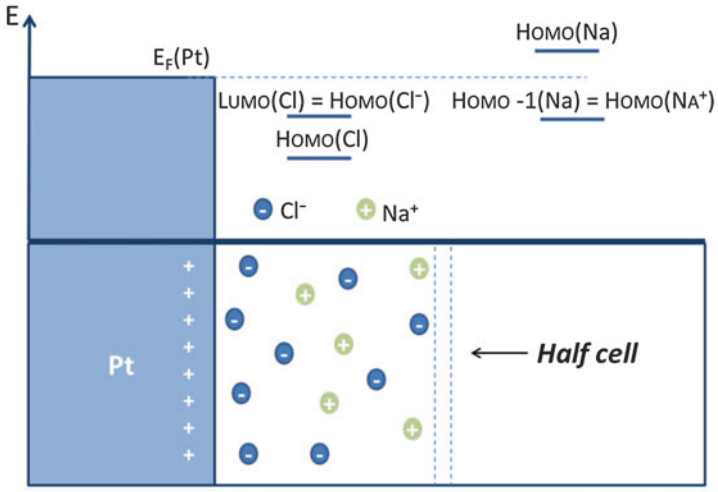
$$G_{\text{int}} = \left(G(N,n) - G(N,0) - n\mu_{H_2}/2\right)/N$$

where μ_{H_2} is the chemical potential of hydrogen. $G_{\text{int}} = 1/2C(U-U_{pzc})^2$ can be plotted as a function of the electrode potential U, calculated from the work function, extracting in this way the interface capacitance C. The potential of zero charge, U_{pzc}, is found by choosing the chemical potential μ_{H_2} in such a way that the G_{int} has its minimum for the configuration with no protons in the water layer. By choosing the μ_{H_2} equal to the free energy of H_2 at standard conditions, the minimum of G_{int} is at another value, representing the NHE. In this way, the work function is directly related to the potential scale of the NHE.

A different approach, proposed by Otani and Sugino [14], relies on performing nonperiodic slab calculations with effective screening medium. Green's function technique is used to solve the Poisson equation with these boundary conditions. The slab is sandwiched by a continuum medium characterized by dielectric constant ε. For instance, to model the structure of a Pt/water interface in acidic conditions, vacuum ($\varepsilon = 1$) is used for the Pt side of the slab and a perfect conductor ($\varepsilon = \infty$) beyond the water side. Changing the number of electrons produces a charge on the slab surface and induces a countercharge in the perfect conductor. A vacuum region is present between water and the perfect conductor. In this model, the countercharge is placed in a layer beyond the water, 15–20 Å far from the electrode. This again represents a fictitious condition because in a real interface, the countercharge is very close to the electrode surface.

A clever usage and tuning of the methodologies highlighted above [4] led to different applications, simulating half-reactions in electrocatalysis – sometimes including the modeling of kinetic effects [2, 4, 7]. For instance, the thermodynamic approach of Ref. [13] has been used to study equilibrium coverage of metal halides as a function of the electrode potential. H_2 dissociation barrier at water/metal interfaces at room temperature as well as structure and stability of water layers as a function of electrode coverage due to hydrogen and other anions has also been recently addressed. In these cases, the effect of the electrode potential has been included using the approach described in Ref. [7, 8].

Other times, different levels of theory have been combined, for instance, by possible inclusion of the presence of the electrolyte using a polarizable medium [15].



Electrochemical Interfaces for Energy Storage and Conversion, Fig. 1 The standard frozen-levels picture of the single-electron energy levels. A prototypic Pt/electrolyte interface is schematically represented. The electrolyte is water with Na^+ and Cl^- ions in solution. The separation between the filled and empty electronic states is decided by the Pt Fermi energy E_F. Filled molecular levels of water lay well below E_F, while the corresponding empty states are well above E_F. As a result, water is expected to be neutral. The HOMO state of the cation, Na^+ in our case, is above E_F, and thus, this species is expected to be fully ionized. Cl^-, of course, is the complementary case, since its LUMO is below E_F, and it becomes filled in the system under consideration

Single Energy Level Alignment and DFT

The validity of the highlighted theoretical approaches, which charge the electrode to simulate the effect of the application of an external potential, relies on a correct description of the single-electron energies. Charging the electrode surface will occur only provided the KS density of states reflects the ordering shown in Fig. 1. The DFT approach determines self-consistently the charge of each species, filling KS electron states according to their relative energy with respect to the system Fermi energy, determined in this case by the Fermi level of the electrode.

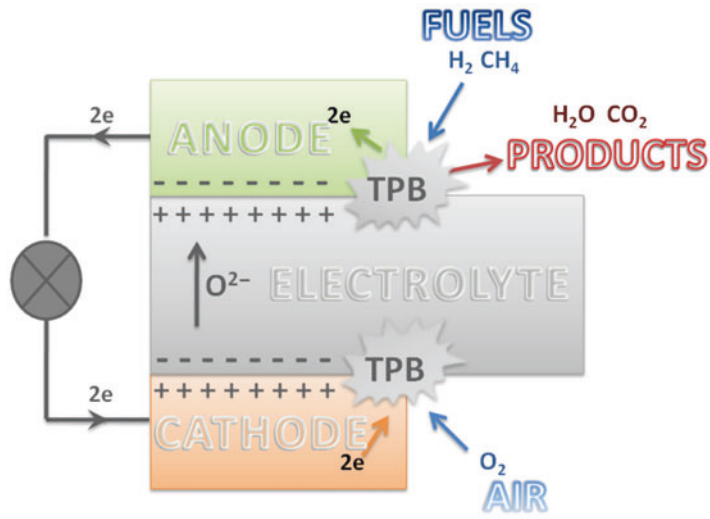
This points out an important methodological open issue to be addressed in view of a realistic description of EC interfaces: the well-known inadequacy of standard DFT in predicting single-electron energies for ions in solution, and for materials currently used as electrodes – such as metals and metal oxides.

In current DFT approximations, the energy of eigenstates depends on the occupation [9], and, as a result, the highest occupied level of cations may fall below the global Fermi energy upon discharging, and the lowest unoccupied level of anions might rise above E_F when accepting the extra negative charge. As a result, the DFT charge on ions in electrolytic solutions might turn out to be a fraction of the electron charge $|e|$, instead of reaching the integer value $\pm n$ as occurs in reality.

To prevent the artificial charge delocalization and improve the description of energy levels, the adoption of exact exchange schemes is required [9]. Such schemes are computationally extremely demanding. The system size and the simulation time become again an obstacle. Alternative, computationally viable solutions would be highly beneficial.

Electrochemical Systems: Batteries and Fuel Cells

One of the greatest scientific challenges of our time is to provide the solution to the dramatic increase in energy demand and costs. Such a solution should be of minimal environmental impact. Energy policies based on capture and sequestration of CO_2 as well as those providing the setting



Electrochemical Interfaces for Energy Storage and Conversion, Fig. 2 Operation of a solid oxide fuel cell (SOFC): At the cathode triple phase boundary (TPB), the oxygen is reduced with charge transfer from the cathode to the electrolyte. The O^{2-} ions diffuse through the electrolyte toward the anode. At the anodic TPB oxidation of the fuel and charge transfer from electrolyte to anode occur

for hydrogen production, storage, and distribution require massive infrastructures and are, at present, beyond the market capacity. In the short to intermediate future, EC devices such as batteries and fuel cells have a great potential for solving key aspects of the energy challenge.

Fuel cells (FCs) are EC devices in which a hydrogen-rich fuel reacts with an oxidizing agent, normally O_2. Fuel and oxygen are split at the opposite sides of an electrolyte, through which, an ion diffuses to combine with an ion with opposing charge, at the opposite side. This process allows the flow of electrons, i.e., electric power (see Fig. 2, which shows how a solid oxide fuel cells works). Like a combustion engine, FCs use a chemical fuel to generate power, but like in Batteries (Bs), the chemical energy is converted directly to electricity by means of an EC process, excluding the combustion step. As a consequence, FCs operate at much higher efficiency and generate fewer pollutants than traditional combustion methods. Bs convert chemical energy directly into electrical power, exploiting the difference in potential between electrodes. Nowadays, Bs power billions of portable electronic devices and are likely to represent the storage technology of choice in the near term for several other applications. The leading Li-based Bs technology represents a $50 billion dollar industry in the USA [16]. Energy storage devices can be utilized together with renewable but intermittent energy sources (like solar or wind) and/or conversion devices. Bs fulfills the need for readily available energy and could provide the means for the vehicle electrification, leading to a reduction of the carbon fingerprint.

Among the various types of FCs, solid oxide fuel cells (SOFCs) are of great interest, because of their ability to utilize, except from hydrogen, a wide variety of commonly available fuels at high efficiency, including waste [17]. Thus, SOFCs do not rely on the development of hydrogen infrastructure and represent a midterm solution to lower both the energy demand and the CO_2 emissions. Typically SOFCs find a natural application as combined heat and power devices for residential applications. They are convenient because they work with the present energy distribution system and potentially could drive the transition toward hydrogen-based economy.

Major factors responsible for the growth of fuel cell market include the ability to integrate fuel cells in stationary, portable, and transportation applications to function as an off-grid power source, with zero emissions. The key concerns in the industry pertain to the high cost of interconnection (mainly in SOFCs, due to their high operating temperature), catalyst for both the cathode and anode compartments, their commercialization, and establishment of fuel cell infrastructure [17].

A broad scientific effort is devoted by both the academic and industrial communities for

improving this energy-related technology; however, the gap separating the actual state of the art of the energy storage and conversion technology from that required for widespread commercialization is still large. On the one hand, increasing the specific energy capacity, the power, the safety, and the life cycle and lowering the cost of various devices remain a challenge, especially when one wishes to power hybrid electric vehicles. Improved electrode materials based on new ideas and advanced electrolytes with higher ionic conductivities and increased voltage stability must be developed. On the other hand, performances, durability, and cost of FCs must be improved. In this area, the major challenges consist of developing new electrolytes, electrodes, and catalyst materials/structures. For instance, while precious catalysts received a great attention and their performance reach the commercialization limit, the non-precious catalysts perform worse, and realization of their full potential is subject to extensive exploratory work.

Applications: Modeling Fuel Oxidation in Commercial Solid Oxide Fuel Cells

Unraveling microscopic processes at the electrode surface remains challenging for both experimental and computational communities [18]. As an example, fuel oxidation at the triple phase boundary (TPB) between Ni electrode, yttria-stabilized zirconia (YSZ) electrolyte, and fuel gas phase in commercial SOFCs was subject to numerous experimental and computational studies. However, in spite of these studies, it was still not clear where water was formed. Impedance spectroscopy measurements identified the rate-limiting step (RLS) with a charge-transfer reaction with an activation barrier in the range 1.04–1.68 eV [19].

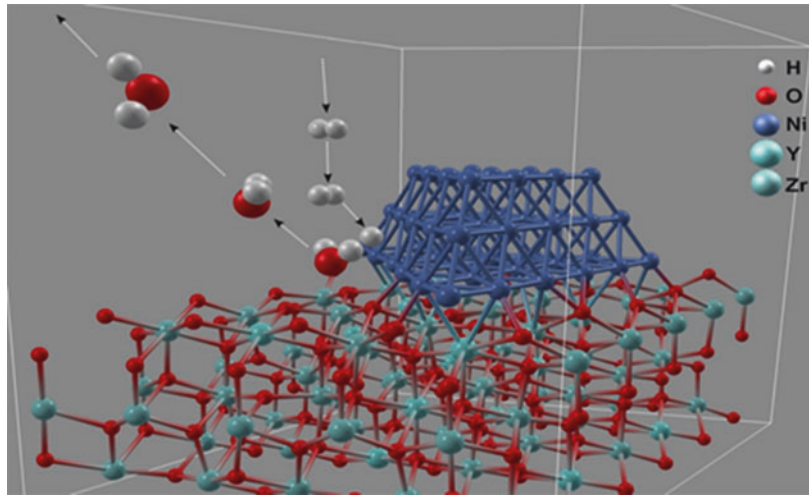
Kinetic modeling of impedance spectra had proposed three different assignments of the nature of the RLS, identified either with H^+, OH^-, or O^{2-} spillover.

In H^+ spillover ($H_{Ni} \rightarrow H_{YSZ} + e_{Ni}$), H atoms adsorbed on Ni hop either to a hydroxyl ion or an oxygen ion on the YSZ surface and here combine with oxygen to produce water. In O^{2-} spillover ($O^{2-}{}_{YSZ} \rightarrow O_{Ni} + 2e_{Ni}$), oxygen ions hop from the YSZ surface to the Ni surface, undergoing two charge-transfer reactions that take place before, after, or during the hop. In OH^- spillover, the OH^- formed on YSZ surface after H^+ spillover jumps back on the Ni anode. Following O^{2-} and OH^- spillover, water formation occurs on Ni surface. It has been also shown that the polarization resistance is reduced by a high water partial pressure.

In this respect, ab initio simulations provided crucial insight on the nature of the RLS and the role of adsorbed water. A recent perspective on DFT modeling of three-phase boundary SOFC electrode has been provided by Shishkin and Ziegler [20]. A realistic modeling of these processes implies that electrode dimension and chemical activity allow charge transfer from O^{2-} to Ni, as it results in the potential drop at the DL, during SOFCs operation.

A state of the art model for the Ni/YSZ interface [21] is shown in Fig. 3. This realistic model allowed to study explicitly the different processes which lead to water desorption from the TPB and their dependence on the local defective structure of the YSZ surface moiety. In this study, it has been proven that the most favorable oxidation pathway involves hydrogen spillover from Ni to YSZ surface, which represents the RLS of the process. This study has shown that water formation and desorption take place on the YSZ surface and are mediated by the diffusion of a hydroxyl group on a Zr site close to the Ni/YSZ interface. The process continues with formation of water by the direct transfer from Ni cluster of a second H atom and its subsequent release. The smaller polarization resistance experimentally observed when water partial pressure increases in the anodic chamber is also explained in terms of enlargement of the active triple phase boundary. In the article, it is shown how this is made possible by the presence of water in the environment, in the form of adsorbed hydroxyl groups on the YSZ surface.

Such a detailed mechanistic insight was made possible due to the combination of a variety of DFT-based techniques such as MD (to explore the



Electrochemical Interfaces for Energy Storage and Conversion, Fig. 3 Supercell used to model the triple phase boundary between electrode, electrolyte, and gas phase in the anodic chamber of Ni–YSZ based SOFCs. Graphics by Mario Valle, Swiss National Supercomputer Centre, Lugano, Switzerland

configurational space), metadynamics (often used as an exploratory tool to guess possible reaction pathways), and nudged elastic band approach (to quantify potential energy profiles) for the optimization of those reaction trajectories. It is worth mentioning that this study required extensive computational time. Therefore, the limitation of the computational facilities can play a critical role in mechanistic studies at electrode surfaces.

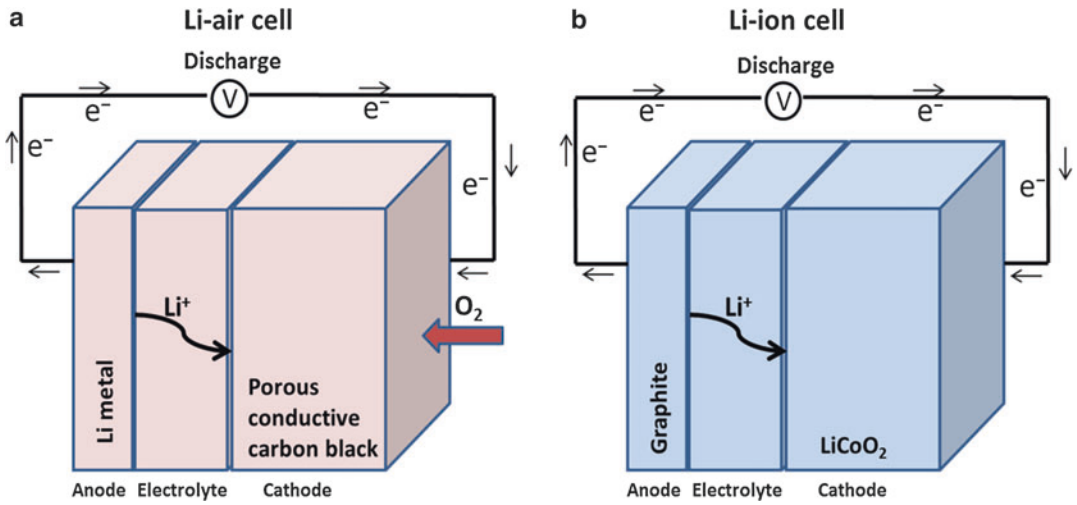
Cluster Models

An alternative approach for studying surface reactivity related to EC interphases is to use a cluster model of the active catalysts site, instead of periodic cell calculations. This approach has been particularly useful in modeling non-precious oxygen reduction reaction (ORR) catalysts, which were demonstrated to have comparable catalytic activity to that of Pt-based catalysts [22]. In such studies, the reaction paths and barriers are found by locating the all relevant minimum energy structures which represent reactants, products, or stable intermediates, as well as confirmed transition state species (i.e., structures having one imaginary frequency) on the potential energy surface. Usually such computational schemes are not prohibitively computationally expensive for using hybrid DFT potentials as well as post-Hartree–Fock electron correlation schemes, as opposed to slab periodic models. The favorable agreement between the computational data and experimental outcome may be partially attributed to the well-tuned DFT functionals, accounting for various types of interactions, typically found in metallo-organic catalysis [23].

Applications: Modeling Electrode Surfaces in Batteries

Understanding of the EI at the battery electrodes is particularly important due to the formation of the solid electrolyte interphase SEI [24, 25] which represents the actual EC surface during the stable operation of the battery. This intermediate layer could be formed due to the reactivity of the surface of the electrode with the electrolyte in the presence of chemically active species formed during battery operation. Therefore, the modeling of surface reactivity should take into account the non-perfect electrode covered with this additional layer. Only in recent years, DFT studies appeared tackling the formation of this layer or its possible precursors.

The importance of modeling the SEI or its precursors cannot be underestimated in metal–air batteries. Metal–air batteries have the highest energy density because most of the cell volume is occupied by the anode while the cathode active material, i.e., the O_2, is not stored in the battery,



Electrochemical Interfaces for Energy Storage and Conversion, Fig. 4 Schematic representation of (**a**) lithium–air cell and (**b**) lithium-ion cell components and the flow of the positive and negative charge on discharge of the cell

rather adsorbed constantly from air into the porous conductive carbon black (Fig. 4a). If Li metal is used as an anode, the specific capacity can be as high as 3,842 mA h g^{-1}. During the discharge of the cell, an oxidation reaction occurs at the anode, $Li \rightarrow Li^+ + e^-$.

Li^+ ion is moving across the electrolyte while the electrons flow in the external cycle to the cathode where a reduction reaction occurs, $2Li^+ + 2e^- + O_2 \rightarrow Li_2O_2$ (or $Li^+ + e^- + O_2 \rightarrow LiO_2$).

The EC performance of this type of cells strongly depends on the electrolyte. By an example, initially, organic carbonates were used; however, it was found that they are unstable toward the nucleophilic attack by the ORR products. Therefore, the understanding and optimization of the EC cathode electrolyte interface are crucial for battery performance.

Very recent studies tackled the reactivity of the electrode surface by modeling each possible reaction of the surface with electrolyte molecules. As an example, the reactivity of propylene carbonate with basic species derived from the Li_2O_2 molecule in solution (Li_2O_2, LiO_2^-, and O^{2-}) and the (100) surface of Li_2O_2 in the presence of the propylene carbonate solvent was recently studied [26]. A wealth of information is provided on the chemical reactivity of the ORR products with the electrolyte. As much as 15 individual reaction paths were examined, combining classical and DFT potentials together with free and potential energy methods.

The complexity of such studies is increased when taking into account particular electronic features of the bulk and surfaces. Again, for Li–air batteries, it was demonstrated that the stable surfaces of Li_2O_2 could be oxygen rich and have a half-metallic character, while the bulk is an insulator [27].

In rechargeable lithium-ion batteries (LIB), in which both the anode and the cathode are intercalation electrodes, the conductivity depends on the positive and negative charge conduction rate both in the bulk of the electrode and on their surface. Schematic representation of a Li-ion cell, which is typically used in portable devices, is shown in Fig. 4b. Interestingly, $LiFePO_4$ was successfully commercialized as a cathode in rechargeable LIBs. However, its surface characterization and the impact of surface composition on Li transport properties have been almost ignored. Recently, this trend has changed. For instance, NEB calculations on a vacuum slab model have shown that the diffusion of Li on the surface has higher activation barrier (0.52–0.77 eV) than for diffusion barrier in the bulk (0.19–0.29 eV) [28].

Summary and Future Perspectives

Understanding the basic principles of operation of novel materials and devices for energy storage and conversion opens new horizons within theoretical material science. It allows designing new strategies – going beyond a mere trial and error procedure – for improving actual technology, developing new materials, and establishing basic relationships in the electrochemistry of interfaces and electrolytes.

The constant development of computational power and new theoretical methods enabled the modeling of EC phenomena purely on the basis of microscopic information of the atomistic structure and our knowledge of electronic phenomena. In particular, parameter-free approaches such as DFT are unique in providing a comprehensive description of the subtle interplay between thermal, electronic, and ionic effects at EI.

The most important progresses in the simulation of realistic interfaces for energy storage and conversion was achieved by the theoretical surface science community. The studies summarized above demonstrated how DFT simulations can provide crucial insight on the nature of key mechanisms at EC bulk and interfaces in batteries and fuel cells. However, many important methodological issues are still open in electrochemistry. For instance, the disclosure of the functional relationship between DL structure, potential drop at EI, and effect of an applied potential is only in its infancy, especially for complex EI.

In this entry, we have highlighted the recent conceptual evolution leading to the development of the most popular methodologies to describe EC phenomena and processes. We have shown how standard simulation methods – which use periodic boundary conditions and supercells – resort to charging the electrode to simulate the effect of an applied potential to an electrolytic cell and how this approach allowed achieving relevant progress in the last 10–20 years. We illustrated how all these models reach global charge neutrality and how most of them reproduce the localization of the electric field and potential energy drop within a microscopic distance from the metal surface.

However, none of these methods currently provide a fully realistic and self-consistent view of the electrostatic DL characterizing an EI, due to the small size of the samples, the artificial assumptions on charge distribution, the lack of description of the open nature of EC systems, and how their dynamical behavior is affected by an applied external potential. Importantly, the kinetics of redox processes at the electrode fully including the effect of the solvent environment is still mostly disregarded.

More realistic models for EIs are needed, possibly including explicit dynamical description of electrolyte solution at the interface with the electrode and its effects on the complex activated phenomena occurring within the DL.

A combination of classical electrochemistry, electronic structure, and classical surface science knowledge will be essential for addressing the still numerous open issues in this intrinsically multidisciplinary research area. Crucially, a substantial progress will be enabled by the increasing availability of computational power.

Finally, the understanding of atomistic and molecular scale processes that occur in EC devices and their interplay can only be achieved by addressing the relevant questions using theoretical models, experimental techniques, computational methodology, and associated simulations as complementary tools.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Electrochemical Scanning Tunneling Microscopy
- ▶ Electrode–Organic Interface Physics
- ▶ Nanomaterials for Electrical Energy Storage Devices

References

1. Parr, R.G., Yang, W.: Density Functional Theory of Atoms and Molecules. Oxford University Press, Oxford, UK (1989)
2. Mueller, J.E., Fantauzzi, D., Jacob, T.: Multiscale Modeling of Electrochemical Systems

Jonathan – From Electrocatalysis: Theoretical Foundations and Model Experiments, 1st edn. In: Richard C. Alkire, Ludwig A. Kibler, Dieter M. Kolb, Jacek Lipkowski (eds.) Wiley-VCH Verlag GmbH & Co. KGaA (2013)
3. Axel Groß, A.: The virtual chemistry lab for reactions at surfaces: is it possible? Will it be useful? Surf. Sci. **500**, 347–367 (2002)
4. Keith, J.A., Anton, J., Kaghazchi, P., Jacob, T.: Modelling Catalytic Reactions on Surfaces with Density Functional Theory, pp. 1–38. Wiley-VCH, Verlag GmbH (2011)
5. Hammer, B., Nørskov, J.K.: Why gold is the noblest of all the metals. Nature **376**, 238–240 (2002)
6. Carrasco, J., Hodgson, A., Michaelides, A.: A molecular perspective of water at metal interfaces. Nat. Mater. **11**, 667–674 (2012)
7. Groß, A. Institut für Theoretische Chemie, Universität Ulm, Albert-Einstein-Allee 11, 89069 Ulm, Germany: Theory of solid/electrolyte interfaces, Ψk Scientific highlight of the month, October 2014
8. Spohr, E.: Some recent trends in computer simulations of aqueous DLs. Electrochim. Acta **49**, 23–27 (2003); Roman, T., Groß, A.: Structure of water layers on hydrogen-covered Pt electrodes. Catal. Today **202**, 183–190 (2013)
9. Cheng, J., Sprik, M.: Alignment of electronic energy levels at electrochemical interfaces. Phys. Chem. Chem. Phys. **14**, 11245–11267 (2012)
10. Taylor, C.D., Wasileski, S.A., Filhol, J.S., Neurok, M.: First principles reaction modelling of the electrochemical interface: consideration and calculation of a tunable surface potential from atomic and electronic structure. Phys. Rev. B **73**, 165402 (2006)
11. Lozovoi, A.Y., Alavi, A., Kohanoff, J., Lynden-Bell, R.M.: Ab initio simulation of charged slabs at constant chemical potential. J. Chem. Phys. **115**, 1661–1669 (2001)
12. Rossmeisl, J., Greeley, J., Karlberg, G.: Electrocatalysis and catalyst screening from density functional theory calculations. In: Marc Koper (ed.) from Fuel Cell Catalysis: A Surface Science Approach. Andrzej Wieckowski, Wiley (2009), ISBN: 978-0-470-13116-9
13. Norskov, J.K., Rossmeisl, J., Logadottir, A., Lindqvist, L., Kitchin, J.R., Blingaard, T., Jònsson, H.: Origin of the overpotential for oxygen reduction at a fuel-cell cathode. J. Phys. Chem. B **108**, 17886–17892 (2004)
14. Otani, M., Sugino, O.: First-principles calculations of charged surfaces and interfaces: a plane-wave nonrepeated slab approach. Phys. Rev. B **73**, 115407 (2006)
15. Andreussi, O., Dabo, I., Marzari, N.: Revised self-consistent continuum solvation in electronic-structure calculations. J. Chem. Phys. **136**, 064102 (2012)
16. http://batteryuniversity.com/learn/article/global_battery_markets
17. Gasteiger, H.A., Markovic, N.M.: Just a dream – or future reality? Science **324**, 48–49 (2009)
18. Adler, S.B., Bessler, W.G.: Elementary kinetic modelling of SOFC electrode reactions. In: Viel-stich, W. (ed.) Handbook of Fuel Cells – Fundamentals, Technology and Applications, vol. 5. Wiley, New York (2009)
19. Tuckerman, M.E.: Statistical Mechanics: Theory and Molecular Simulation. Oxford University Press, Oxford (2010)
20. Shishkin, M., Ziegler, T.: Direct modelling of the electrochemistry in the three-phase boundary of solid oxide fuel cell anodes by density functional theory: a critical overview. Phys. Chem. Chem. Phys. **16**, 1798–1808 (2014)
21. Cucinotta, C.S., Bernasconi, M., Parrinello, M.: Hydrogen oxidation reaction at the Ni/YSZ anode of solid oxide fuel cells from first principles. Phys. Rev. Lett. **107**, 206103(1–5) (2011)
22. Lefevre, M., Proietti, E., Jaouen, F., Dedelet, J.-P.: Iron-based catalysts with improved oxygen reduction activity in polymer electrolyte fuel cells. Science **324**, 71–74 (2009)
23. Zhao, Y., Truhlar, D.G.: Density functionals with broad applicability in chemistry. Acc. Chem. Res. **41**, 157–167 (2008)
24. Balaish, M., Kraytsberg, A., Ein-Eli, Y.: A critical review on lithium-air battery electrolytes. Phys. Chem. Chem. Phys. **16**, 2801–2822 (2014)
25. Peled, E.: The Electrochemical Behavior of Alkali and Alkaline Earth Metals in Nonaqueous Battery Systems -The Solid Electrolyte Interphase Model, J. Electrochem. Soc. **126**, 2047–2051 (1979)
26. Laino, T., Curioni, A.: A new piece in the puzzle of lithium/air batteries: computational study on the chemical stability of propylene carbonate in the presence of lithium peroxide. Chem. Eur. J. **18**, 3510–3520 (2012)
27. Radin, M.D., Rodrigues, J.F., Tian, F., Siegel, D.J.: Lithium peroxide surfaces are metallic, while lithium oxide surfaces are not. J. Am. Chem. Soc. **134**, 1093–1103 (2012)
28. Dathar, G.K.P., Sheppard, D., Stevenson, K.J., Henkelman, G.: Calculations of Li-ion diffusion in olivine phosphates. Chem. Mater. **23**, 4032–4037 (2011)

Electrochemical Machining (ECM)

Amitabha Ghosh
Bengal Engineering and Science University, Shibpur, Howrah, India

Synonyms

ECM

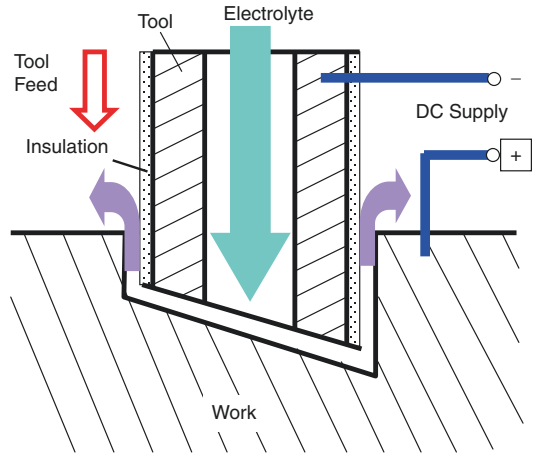
Definition

Electrochemical machining is a metal removal–based manufacturing process that depends upon electrochemical dissolution for metal removal mechanism. It is one of the advanced manufacturing processes for producing parts with complex shapes using metals and alloys with low machinability.

Overview

Though electrochemical machining process is of relatively recent origin, the reverse process – electroplating – has been in use since a long time. The basic process is indicated in Fig. 1.

A metallic tool is fed at a constant rate toward the work, and the gap between the tool face and the work surface (in the machining area) is supplied with copious flow of an electrolyte (often brine) at a reasonably high rate. A large direct current at a low voltage (typically 10–15 V) is supplied to the work–tool system, the positive terminal being the work. Electrochemical dissolution causes material to be removed from the work piece and the electrolyte–material combination being such that no deposition takes place on the tool electrode (and the dissolved material and the ohmic heat generated in the tool–work gap are carried away by the electrolyte). The shape of the machined portion of the work conforms to the shape of the tool face as indicated in Fig. 1. Once the desired shape is achieved the tool is withdrawn.



Electrochemical Machining (ECM), Fig. 1 Basic scheme of electrochemical machining

Process Mechanism

According to Faraday's laws, the amount of material dissolved during an electrochemical process is proportional to the quantity of electricity passed; besides, the amounts of different substances dissolved by the same amount of electricity are proportional to their chemical equivalent weights. Thus,

$$m \propto \in It \tag{1}$$

where m is the weight in gram of a substance dissolved, I is the current in amperes, t is the time in seconds and $\in$ is the gram equivalent weight of the substance (i.e., the work material). Equation 1 can be written in the form of the following equation:

$$m = \frac{\in It}{F} \tag{2}$$

Where F represents the constant of proportionality and is equal to 96,500 C. For ferrous chloride solution (brine) as the electrolyte, the following electrochemical reactions occur at the anode (work) and the cathode (tool); respectively:

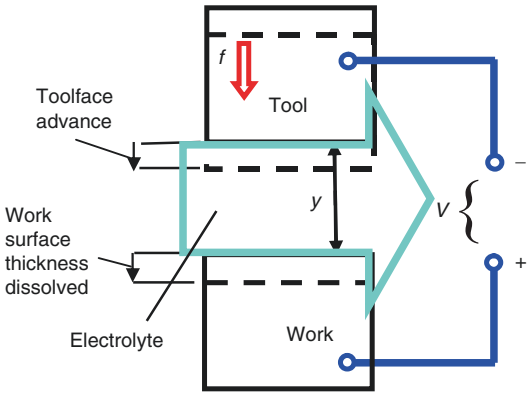
$$Fe \rightarrow Fe^{++} + 2e^{-}$$
$$2H_2O + 2e^{-} \rightarrow H_2 \uparrow + 2(OH)^{-}$$

Thus, Fe dissolves at the anode and H_2 is released at the cathode. The reaction in the electrolyte is as shown below:

$$Fe^{++} + 2(OH)^{-} \rightarrow Fe(OH)_2 \downarrow$$

The insoluble $Fe(OH)_2$ is removed with the electrolyte as a precipitate and no deposition at the cathode occurs.

The gram equivalent weight of the work metal, $\in$ is given by A/Z, where Z is the valence



Electrochemical Machining (ECM), Fig. 2 Mechanics of ECM

of the produced ions and A is the atomic weight. If ρ be the density of the work material in gm/cm^3, Eq. 2 yields the following relation for the volume rate of material removal from the work piece:

$$Q = \frac{AI}{96,500\rho Z}\text{cm}^3/\text{s} \tag{3}$$

The magnitude of the current I is given by

$$I = \frac{V - \Delta V}{R} \tag{4}$$

where V is the applied voltage in volts, ΔV represents the combined over voltage due to various effects and should be subtracted from the impressed voltage before Ohm's law can be applied, and R is the overall resistance (but most of it due to the electrolyte layer in the interelectrode gap). Typical electrolytic conductivity is 0.1–1.0 $\Omega^{-1}\text{cm}^{-1}$ whereas that of iron is about $10^5\Omega^{-1}\text{cm}^{-1}$.

Figure 2 shows symbolically the plane-parallel work–tool electrolyte combination with tool being advanced at a constant rate of f toward the work.

If the applied voltage is V and specific conductivity of the electrolyte is k, Eq. 2 yields the following relation for the current density at the work surface:

$$J = \frac{k(V - \Delta V)}{y} \tag{5}$$

Where, y is the gap thickness at the instant under consideration. So, after a short time interval Δt the tool face advances by $f \times \Delta t$. The work surface also recedes because of electrochemical dissolution. If δ be the depth of recession then using Eq. 3

$$\delta = Q \times \Delta t = \frac{AJ\,\Delta t}{96,500\rho Z}$$

Replacing J in the above equation from Eq. 5

$$\delta = \frac{Ak(V - \Delta V)}{96,500\rho Z y} \times \Delta t$$

But when the gap is stable the rate of depth of dissolution ($\delta/\Delta t$) and the rate of tool advance (f) must be equal. Hence, under equilibrium condition (i.e., $y = y_e$)

$$\frac{\delta}{\Delta t} = f$$

or,

$$\frac{Ak(V - \Delta V)}{96,500\rho Z y_e} = f$$

or,

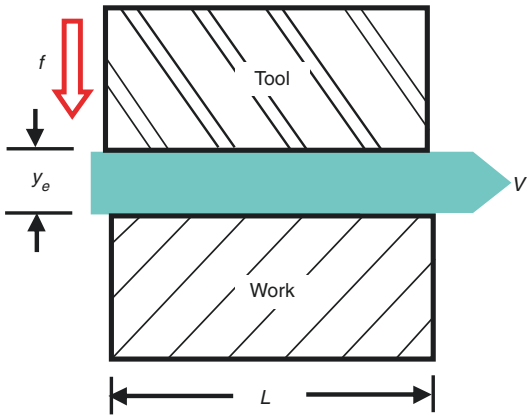
$$y_e = \frac{Ak(V - \Delta V)}{96,500\rho Z f} \tag{6}$$

It can be easily seen that this equilibrium condition is stable and for a given electrolyte–electrode combination, the equilibrium work–tool gap, y_e, depends upon the applied voltage and the feed rate.

Since an equilibrium gap is ultimately reached it is possible to arrive at a definite relationship between the tool and work geometries (i.e., a tool can be shaped to generate a desired work shape). This phenomenon makes the process suitable for machining objects with definite shapes.

Machining Force and Electrolyte Flow

At a first glance, it may appear that there is no machining force involved as there is no contact



Electrochemical Machining (ECM), Fig. 3 Electrolyte flow through work–tool gap

between the work and the tool. But, in reality, there exists large force acting on the electrodes due to a large pressure exerted by the electrolyte flowing through the work–tool gap.

Figure 3 indicates symbolically an ECM operation with an equilibrium work–tool gap of y_e. The flow velocity, v, is determined predominantly from the thermal conditions.

If the variation in conductivity of the electrolyte with temperature is ignored, then the following equation for the required flow velocity, v, of the electrolyte can be obtained [1]:

$$v = \frac{J^2 L}{k\Delta\theta\rho_e c_e} \tag{7}$$

where, J is the current density, L is the length of the flow path, k is the specific electrical conductivity of the electrolyte, $\Delta\theta$ is the allowed temperature rise in the electrolyte, ρ_e and c_e are the density and specific heat of the electrolyte. If the gap is y_e and the viscosity of the electrolyte is η, the total pressure required (assuming turbulent flow) to achieve a flow velocity of v is given by the relation [1].

$$p = \frac{0.3164\rho_e v^2 L}{4y_e (R_e)^{1/4}} + \frac{\rho_e v^2}{2} \tag{8}$$

where, R_e is the Reynolds's number. It can be easily demonstrated that to prevent boiling of the electrolyte the pressure for the electrolyte in a typical ECM operation is a few hundred kN/m^2. This can exert a reasonably large force on the tool face depending on the size.

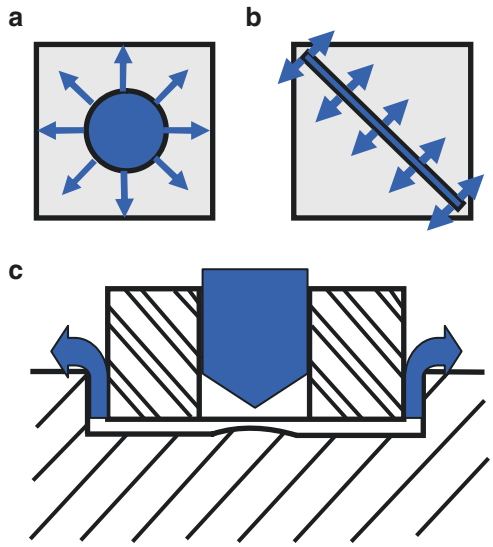
Tool Design

The three primary aspects associated with the tool design for ECM operation are (a) the determination of the geometric shape of the tool that generates the desired work shape, (b) electrolyte flow path design that ensures smooth and ample flow of electrolyte all over the working area, and (c) design from other considerations like strength, insulation of the areas of tool surface which should not take part in the electrochemical process and fixing arrangements.

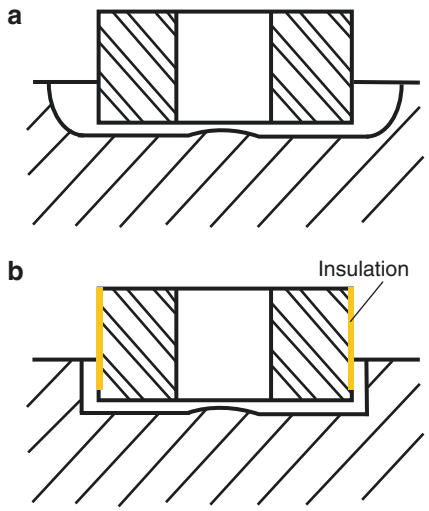
Since the equilibrium gap y_e can be accurately predicted it is possible to decide upon the required tool geometry that will generate the desired work shape. Design of electrolyte flow path is extremely important as the quality of the surface generated depends very significantly on the uniformity and smoothness of electrolyte flow. A large flow requires a large cross-sectional area of the orifice through which the flow emerges as indicated in Fig. 4a. This leads to a residual bulge on the machined surface of the work as shown in Fig. 4c. To avoid the residual bulge without reducing the flow path opening area, the passage can be in the form of a narrow but long slot as shown in Fig. 4b. Since the flow will be in a direction normal to the slot as shown, the slot geometry should be carefully decided so as to keep the whole machining area with ample and smooth flow of electrolyte.

To restrict machining only at the desired locations, it is essential to insulate some portions of the tool surface. Figure 5a, b indicates the result of ECM without and with insulation.

It is obvious that the tool material should be electrically conducting. As the desired shape has to be machined, tool material should have good machinability and reasonable mechanical strength.



Electrochemical Machining (ECM), Fig. 4 Tool face showing the flow exit holes (**a**) normal shape, (**b**) narrow and long passage, (**c**) residual bulge with a flow path as shown in (**a**)



Electrochemical Machining (ECM), Fig. 5 Machined shape (**a**) without insulation, (**b**) with insulation

Electrolytes

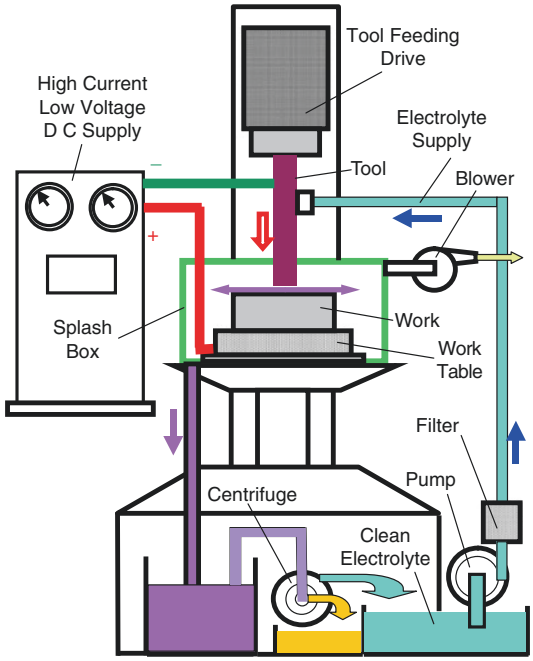
The functions accomplished by the electrolyte during ECM are: (a) completing the electrical path allowing passage of large current, (b) sustaining the electrochemical dissolution process, and (c) carrying away the heat and waste product from the machining area. So a good electrolyte should have large electrical specific conductivity. It should be also of such composition that deposition does not occur on the cathode (i.e., the tool) as otherwise the tool geometry can change. Besides, an electrolyte should be inexpensive, safe, and as noncorrosive as possible. Usually, an aqueous solution of inorganic compounds is used as shown below in the list:

Work material	Electrolyte
Ferrous	Chloride solution in water, usually NaCl
Nickel based	HCl or mixture of H_2SO_4 and brine
Titanium based	10 % HF + 10 % HCl + 10 % HNO_3
Co–Cr–W based	NaCl
WC based	Strong alkaline solutions

ECM Unit

The basic components of a typical ECM unit are shown in Fig. 6. It consists of a work table holding the job.

The tool is fed by a low-feed-rate tool feeding drive with a facility to quickly retrieve the tool at a fast rate. The machining zone is enclosed in a box and the collected electrolyte and the precipitate are accumulated in a tank. A centrifuge separates the precipitate as sludge and the clean electrolyte is stoned in another tank from where it is recirculated by a pump through a filter as shown. The H_2 generated by the process is removed using an exhaust blower to prevent the chance of any explosion. A low voltage high amperage DC supply provides the required large current at a low voltage as indicated. All the units must be made of corrosion-free material like stainless steel, polymeric materials, etc. The whole system is supported by a rigid and strong structure as very large electrolyte pressure can generate a large force between the tool and the electrode; and any deformation can jeopardize the operation. All these requirements result in high cost of such machines.



Electrochemical Machining (ECM), Fig. 6 Basic components of a typical ECM unit

References

1. Ghosh, A., Mallik, A.K.: Manufacturing Science, 2nd edn. Affiliated East West Press, New Delhi (2010)

Electrochemical Scanning Tunneling Microscopy

Ilya V. Pobelov, Chen Li and Thomas Wandlowski
Department of Chemistry and Biochemistry, University of Bern, Bern, Switzerland

Synonyms

EC-STM; In situ STM

Definition

Electrochemical scanning tunneling microscopy (EC-STM) is a conceptional and instrumental extension of scanning tunneling microscopy toward its application to charged solid–liquid interfaces (electrochemical systems).

Introduction

With the invention of scanning tunneling microscopy (STM) by Binnig and Rohrer, a unique new tool for high-resolution imaging of structures and process (dynamics) at conducting surfaces became available. Soon after its application in surface studies under ultrahigh vacuum (UHV) conditions, this method was also adopted to electrochemical systems under potentiostatic or galvanostatic control and "in situ" EC-STM was born. This technique advanced tremendously the knowledge on surface structure and reactivity in electrochemistry because imaging of electrode surfaces was now possible in situ, in real space, and with atomic-scale resolution [1–7].

Examples range from fundamental studies of reconstruction of metal single-crystal surfaces [5], ordered adsorption of anions [3, 6], and organic molecules [7] to the initial stages of metal deposition [8] and up to applications in electrocatalysis, plating, and corrosion [9]. During the last decade, in situ STM investigations were combined with in situ scanning tunneling spectroscopy (STS) to obtain electronic structure information [10] and to explore charge transport in single atom and molecular junctions [11] as well as in tunneling junctions [12]. Besides imaging with atomic-scale resolution, the STM tip has been also employed as a tool in nanostructuring of electrode surfaces with high spatial and chemical resolution [13].

The entry starts with a brief introduction into electrochemistry and EC-STM. Subsequently, the potential of STM in electrochemistry is demonstrated in selected structure studies with bare and adlayer (anions, molecules, metal atoms)-modified single-crystal surfaces and by monitoring simple reactions, such as metal deposition. Finally, single-molecule transport experiments and nanostructuring of surfaces by the STM tip will be addressed.

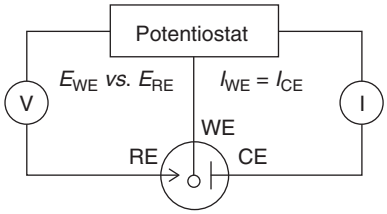
Overview and Key Principles

Fundamentals of Electrochemistry

Electrochemistry represents chemical reactions that are connected with the transfer of electric charges [14]. Many of these processes occur at electrified solid–liquid interfaces. Typically, the solid phase acts as an electron conductor (metal or semiconductor electrode) and the liquid phase as ionic conductor (e.g., an aqueous electrolyte solution). The interface is composed of electronic and ionic charge carriers and characterized by an electric field as high as 10^9 Vm^{-1}. This region is called "electrical double layer" (EDL). In the absence of electrochemical reactions, no charge transfer across the interface occurs. The latter then resembles a capacitor that can be charged or discharged upon application of an external voltage ("ideally polarizable interface"). In the presence of an electrochemical (Faraday) reaction, charge or, in other words, an electrochemical current passes the interface ("ideally nonpolarizable interface"). The classical route to the study of electrochemical reactions rests on current and voltage measurements. Both quantities can be determined with high precision and are easily accessible, although they lack direct chemical or structural specificity.

Electrochemical experiments are typically carried out under potentiostatic conditions (i.e., under potential control) in a three-electrode configuration composed of a working electrode (WE), the electrode of interest, a reference electrode (RE), and a counter electrode (CE), respectively (Fig. 1). For the discussion of a relation between the electrochemical and the absolute potential scales, we refer to Ref. [15]. All electrodes are immersed into the electrolyte solution. The setting up of the potential differences and the measurement of the current are controlled by a potentiostat (Fig. 1).

The potentiostat measures current and potential in the CE and RE circuits. Due to voltage drops and polarization effects, the potential difference between WE and RE can be measured precisely only if there is negligible current flowing through the RE. The potential difference between the WE and the CE is adjusted via a feedback loop



Electrochemical Scanning Tunneling Microscopy, Fig. 1 Schematics of a potentiostat and three-electrode electrochemical cell

in such a way that the desired value of the electrode potential E (E_{WE} vs. E_{RE}) is reached. As a consequence, the current flowing in the CE circuit (between WE and CE) is the current flowing at the WE at this electrode potential [14].

Basics of Scanning Tunneling Microscopy

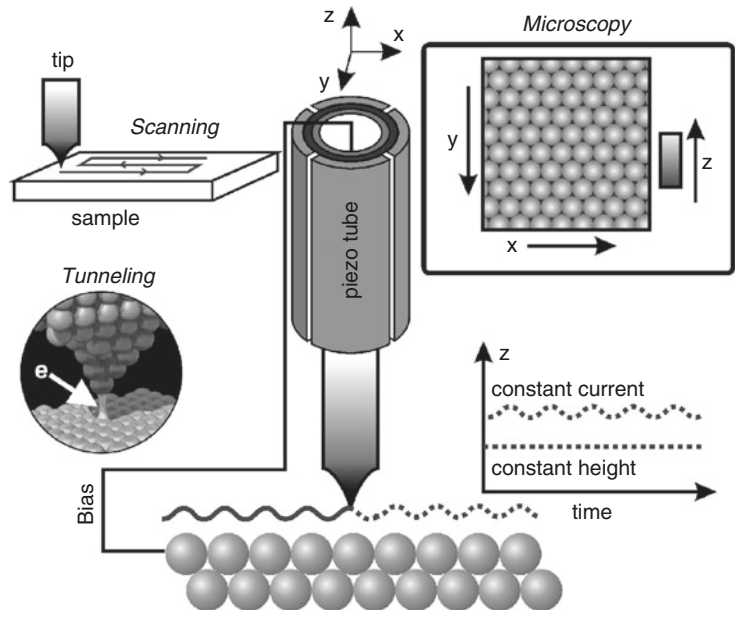
The STM technique is based on electron tunneling, which is well known from quantum mechanics. When a freely moving electron incidents an energy barrier of constant relative height Φ and width z, it may penetrate through it with the probability

$$P(z) = \exp\left(-\frac{4\pi}{h}\sqrt{2m_e\Phi}\cdot z\right) = \exp\left(-\kappa\sqrt{\Phi}\cdot z\right) \quad (1)$$

Here $h = 6.626 \times 10^{-34}$ J · s is Planck's constant, $m_e = 9.11 \times 10^{-31}$ kg is the electron mass, and $\kappa = 10.12\ \mathrm{eV}^{-1/2} \cdot \mathrm{nm}^{-1}$ is a combination of constants. In STM, a sharp conductive tip (typically metal) is brought up to few nanometers close to a conductive surface (Fig. 2). The potential difference E_b (bias voltage) applied between an atomically sharp tip and the sample causes electron tunneling via the insulating gap, which could be ultrahigh vacuum (UHV), gas, or liquid. In the low-bias approximation, the tunneling current is proportional to E_b and to the electron tunneling probability:

$$I_T(z) \propto E_b \exp\left(-\kappa\sqrt{\Phi}\cdot z\right) \quad (2)$$

The tunneling current I_T is measured by a highly sensitive preamplifier with a typical operational



Electrochemical Scanning Tunneling Microscopy,
Fig. 2 Schematics of an STM. *Upper left*: tip scanning on the sample; *lower left*: electron tunneling between tip and sample; *upper right*: obtained STM image; *lower right*: two operation modes

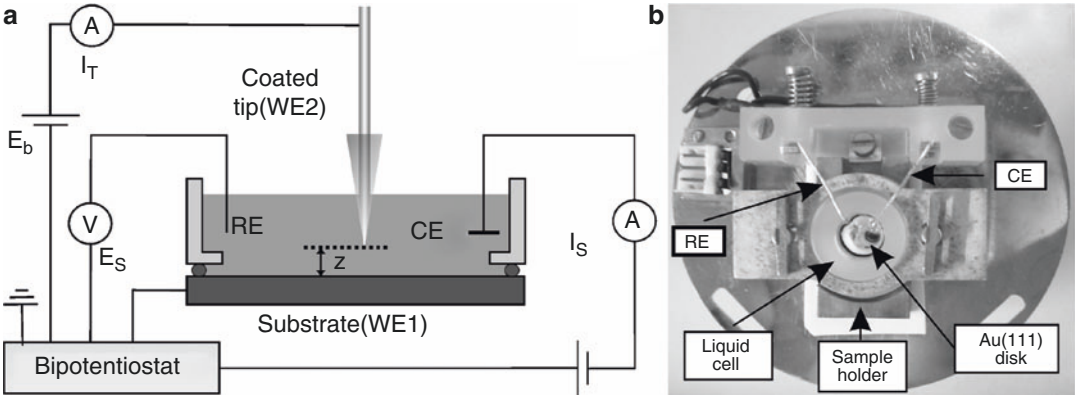
range of 1 pA to 100 nA. The *x,y,z*-movement of the tip in the nanometer scale is performed by a piezoelectric tube based on ceramics (often lead–zirconium–titanate) exhibiting a piezoelectric effect, that is, expanding or shrinking upon application of a voltage. One may distinguish between two operational modes of the STM (Fig. 2). In constant-current mode, a preset tunneling current I_T is kept constant by adjusting the z position of the tip at every measured point via a feedback circuit. The information on vertical displacement of the tip as a function of the lateral position is recorded as an array of (x, y, z) data and presented as a three-dimensional (3D) or two-dimensional (2D) image of the local density of states, which reflects, in the case of metals, the topography of the substrate surface directly.

In constant height mode the vertical position of the tip is kept constant during the lateral scanning while the value of I_T is recorded as a function of the tip position. The array of (x, y, I_T) data provides information on the surface topography and can be presented in a similar way as results of the constant-current mode STM.

In an STM experiment, one may achieve lateral *x,y* resolution of 0.1 nm and a vertical resolution up to 0.01 nm. This unprecedented capability of STM is due to the strong dependence of the tunneling current on the distance z between tip and sample (cf. Eq. 2). With a typical barrier height of one to a few eV, an increase of z by the size of a single metal atom (0.2–0.5 nm) decreases the tunnel current up to three orders of magnitude. Atomically sharp STM tips are required to avoid artifacts, and in this case, only the most protruding part of the tip acts as a probe. Finally, it is noted that STM per se is not providing "chemical information" on surface (and tip) composition!

Electrochemical (In Situ) Scanning Tunneling Microscopy

Tunneling in an electrochemical experiment takes place in an electrically conducting environment, the electrolyte (aqueous, organic, or ionic liquid as examples). Electrochemical reactions may occur at the tip as well as at the substrate. The resulting electrochemical current I_F is superimposed on the tunneling current I_T. To minimize I_F at the tip relative to I_T, the tip is coated with an inert and insulating layer made of glass, apiezon, polyethylene, or electrophoretic paint [1]. The remaining electrochemically active area of well-insulated tips amounts to 10^{-8}–10^{-7} cm^2 and yields residual electrochemical ("leakage") currents of typically less than 5 pA, which do not interfere with tunneling currents of 0.1–10 nA.



Electrochemical Scanning Tunneling Microscopy, Fig. 3 (**a**) Schematics of an EC-STM setup; (**b**) assembled cell

From an electrochemical point of view, substrate and tip represent two working electrodes (WE1 and WE2 in Fig. 3a), in which potentials E_S and E_T are independently controlled with respect to the RE, and the bias voltage is given by the potential difference $E_b = E_T - E_S$. This functionality is controlled by a bipotentiostat, which represents a potentiostat with two independently adjustable working electrodes (four-electrode potentiostat).

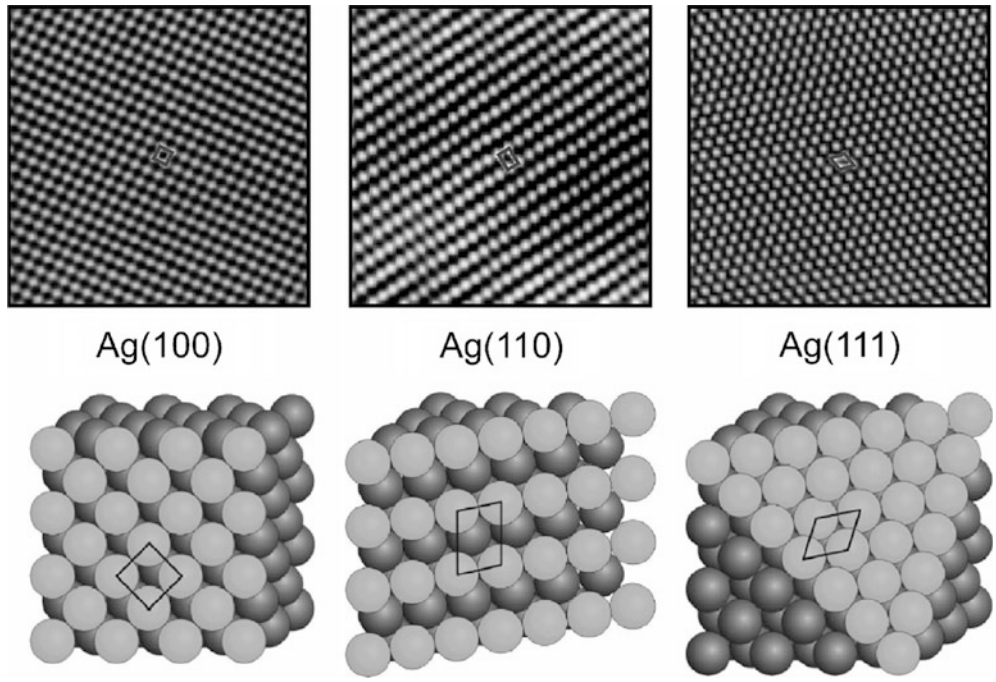
The type of substrate and its potential E_S are determined by the electrochemical problem to be studied. The choice of tip material and potential E_T are dictated by the requests of high mechanical and, in addition, chemical stability to minimize the faradic current I_F at the tip. Atomically sharp tips are typically prepared by electrochemical etching of thin metal wires, such as tungsten, platinum–iridium, or gold.

The choice of the RE depends on electrolyte and electrochemical problem under study. Examples are a Pt wire in acidic electrolytes, reference electrodes of the first kind (Me in Me^{z+}-containing electrolyte such as Cu/Cu^{2+} or Ag/Ag^{+}) or of the second kind (Ag/AgCl, Au/Au_xO_y, etc.). The materials of the CE are less critical. Often Pt or Au wires are chosen. Attempts to remove oxygen and to control the environment during an STM experiment led to the development of various custom-built environmental chambers. Figure 3b illustrates, as an example, a typical EC-STM cell.

Besides high-resolution atomic-scale imaging, even up to video acquisition rates [16], most EC-STM setups are also capable of performing scanning tunneling spectroscopy (STS) measurements. The latter are carried out at fixed lateral (x,y) position of the tip above the substrate. In *current–distance STS experiments*, the tunneling current is measured at constants E_S and E_T as a function of the vertical tip displacement Δz. The independent control of the potentials of both tip and substrate with respect to the RE allows various modes of *current–voltage spectroscopy* in an electrochemical environment. In *variable bias STS mode*, the potential of one electrode (typically of the substrate WE1, E_S) is fixed, while that of the other (WE2, E_T) is swept over a wide range of potentials. This regime is equivalent to an ex situ current–voltage STS experiment with two electrodes. The *constant bias mode of current–voltage STS* is unique for the electrochemical environment. Here, both E_S and E_T are swept simultaneously, so that their difference E_b is kept constant. Tunneling spectroscopy is an important technique for studies of electronic properties and in particular of electron transfer at electrochemical solid–liquid interfaces [10, 12].

Key Research Findings and Selected Applications

In the following section, the potential of EC-STM is demonstrated in structure studies on bare and adsorbate-modified single crystals and by



Electrochemical Scanning Tunneling Microscopy, Fig. 4 *Top row*: STM images of three low-index faces of silver single-crystal electrodes in 0.05 M H_2SO_4, 7 × 7 nm. *Bottom row*: models of the low-index faces of a FCC metal. The surface unit cells are indicated (Adapted from Ref. [17])

monitoring simple reactions such as metal deposition. Finally, single-molecule electron transport experiments and nanostructuring of surfaces are described.

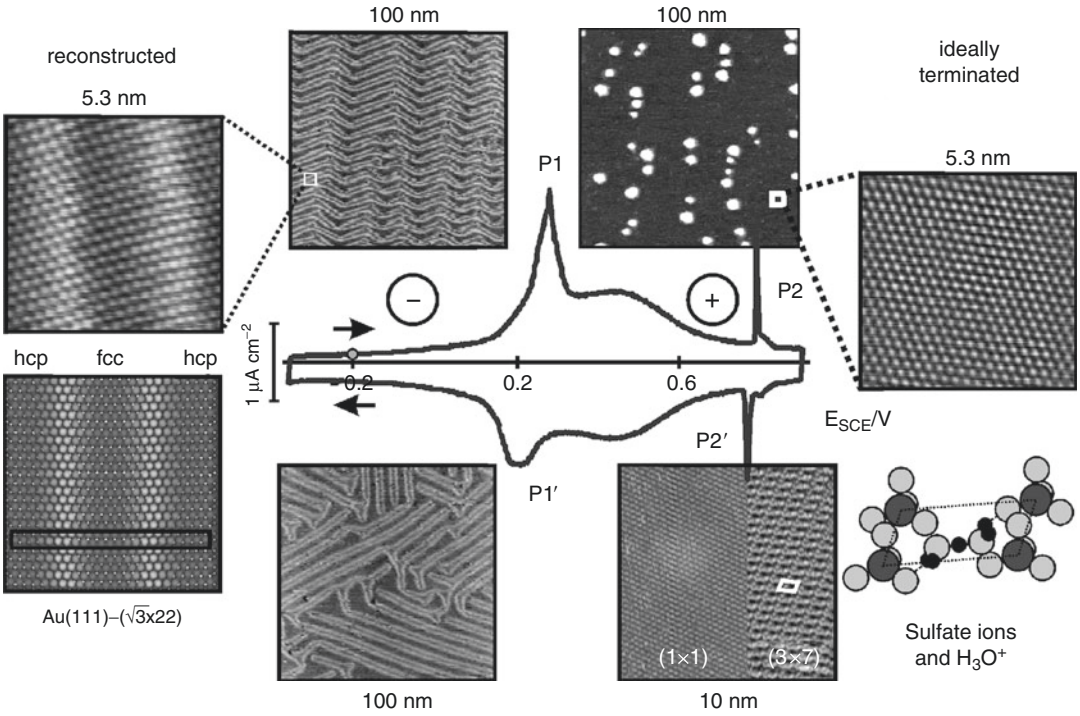
Structure of Single-Crystalline Surfaces

For a detailed understanding of surface processes, the use of single crystals with structurally well-defined surfaces is highly desirable. After establishing high-quality and lab-accessible preparation techniques, such as flame annealing, inductive annealing, and various chemical etching protocols, STM revealed unprecedented microscopic details of electrode surfaces under electrochemical conditions since the early 1990s up to the present [1–5]. The majority of atomic resolution studies was carried out with metals forming face-centered cubic (FCC) crystal lattices, such as Au, Ag, Cu, Ir, Ni, Pd, Pt, and Rh. Selected studies were also reported for Fe-type and semiconductor surfaces. As an example, Fig. 4 shows in situ STM images of the three low-index faces of silver single-crystal electrodes recorded in 0.05 M H_2SO_4 [17], together with schematic models of the respective surface structures.

The high spatial resolution of EC-STM provides access not only to steady-state surface structures, such as the lattice arrangement of surface atoms and defects (holes, steps, kinks, and/or islands) under potential control, but also to dynamic phenomena with a time resolution as high as several tens of milliseconds. Examples are step and island mobility [18] and surface reconstruction [5].

Surface Reconstruction

Experiments on the formation and lifting of surface reconstruction represent classical examples of surface processes driven by electrode potential and solution composition [5]. Surface reconstruction is the rearrangement of surface atoms into a different structure than the one dictated by the crystal plane, as driven by the minimization of the energy for surface atoms with an unsaturated coordination sphere. This phenomenon was first extensively studied in UHV. Conclusive evidence



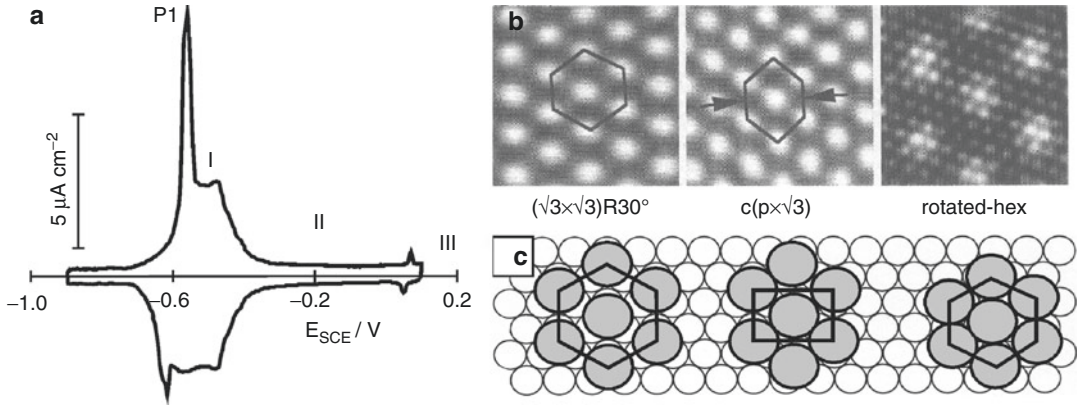
Electrochemical Scanning Tunneling Microscopy, Fig. 5 Cyclic voltammogram and STM images of Au(111) in 0.1 M H_2SO_4 (Adapted from Ref. [7])

about the existence of reconstructed surfaces under electrochemical conditions was obtained after electrochemists gained access to high-resolution structure-sensitive techniques, such as in situ STM. Examples are the low-index phases of Au (Au(111)-(p × √3), Au(100)-(hex), Au (110)-(1 × 2), and Au(110)-(1 × 3)) and Pt (Pt (100)-(hex), Pt(110)-(1 × 2)) as prepared by sputtering and annealing cycles in UHV or annealing in a Bunsen burner or hydrogen flame. In particular, in situ STM and surface X-ray scattering experiments revealed that the reconstructed surfaces of the three low-index Au single crystals are stable at negative surface charges and are lifted upon positive polarization. The process is reversible. The adsorption of anions was found to assist the lifting of the reconstruction [5].

Figure 5 shows large-scale and atomic-scale resolution images of the thermally induced Au (111)-(p × √3) reconstruction, as obtained after flame annealing in 0.1 M H_2SO_4 at $E_S = -0.20$ V, where the electrode is charged negatively. Changing the electrode potential toward positive charges causes the lifting of the reconstruction into the Au(111)-(1 × 1) structure. The extra amount of Au atoms from the ca. 4 % more densely packed (p × √3) structure forms monatomic high gold islands, which are rather mobile. At sufficiently high potentials, the adsorbed sulfate ions form a 2D-ordered (√3 × √7) adlayer on top of the Au(111)-(1 × 1) surface. Changing the potential back to negative charge densities causes the reconstruction to reappear. However, this so-called electrochemically induced reconstruction is less ordered as compared to the thermally induced reconstruction. For further details the reader is referred to Refs. [5, 7].

Adsorption of Anions

The study of adsorbates is of importance for electrochemistry, as many reactions proceed through intermediate adsorption stages [3, 6]. Anions exhibit a strong tendency to adsorb specifically at metal surfaces, that is, to establish a direct bond



Electrochemical Scanning Tunneling Microscopy, Fig. 6 Adlayers of iodide on Au(111)-(1 × 1). (**a**) Cyclic voltammogram of Au(111) in 50 mM NaI, scan rate 50 mV·s^{-1}. The potential ranges of stability for the different monolayers of I on Au(111) are indicated, (**b**) EC-STM images, and (**c**) models of the I monolayers on Au(111) (Adapted from Ref. [3])

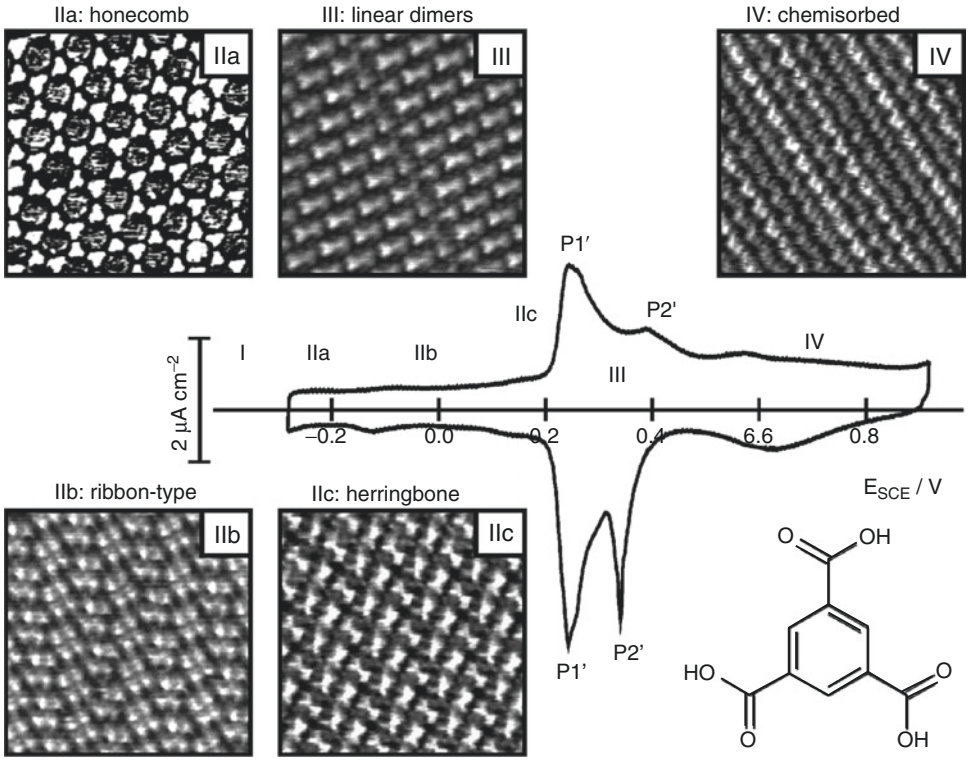
with the metal by partial loss of their solvation shell. The direct contact to the electrode reduces the ionic character markedly, which allows a much higher surface concentration than in the case of nonspecific (purely electrostatic) adsorption. Cyclic voltammograms often reveal sharp current spikes (Fig. 6a), indicative of phase transitions within the anionic adlayers and, hence, of the existence of 2D-ordered phases. Structural aspects of these adlayers have been studied in detail by EC-STM in combination with other structure-sensitive in situ techniques [3, 6].

As an example, Fig. 6 demonstrates interfacial structures formed by iodide adsorbed on Au(111) [3]. The cyclic voltammogram was recorded in 50 mM NaI aqueous electrolyte. The current peak P1 represents the lifting of the reconstruction Au(111)-(p × √3) → Au(111)-(1 × 1), while the potential regions I, II, and III, which are separated by characteristic current peaks, could be identified as the following 2D-ordered iodide adlayers: (√3 × √3)R30°, a compressed c (p × √3) phase, and a rotated hexagonal phase at the highest potential and iodide coverage, just before the onset of adlayer oxidation.

Adsorption and Self-Assembly of Organic Molecules

The adsorption and self-assembly of organic molecules at metal surfaces comprise a growing field of interest for fundamental research as well as technological applications. This is due to the immense importance of organic molecules in applied areas such as additives in plating baths (e.g., for superconformal growth in the Damascene process of Cu deposition [19]) or as corrosion inhibitors [9] and to the growing field of molecular nanotechnology for designing materials and functionalized surfaces with novel electronic, magnetic, photonic, mechanical, and sensing/diagnostic properties.

Employing defined substrate surfaces, such as metal single crystals, and tailored molecules as coded information, molecular recognition pattern, and self-organization as well as local functionalities have been studied by EC-STM under a wide range of conditions [2, 3, 7]. These investigations benefited from the application of concepts of supramolecular chemistry to create modified surfaces in the submonolayer, monolayer, and multilayer regime. Adlayer structure motifs are based, for instance, on hydrogen bonding and hydrophobic and electrostatic interactions, such as dipole, π-stacking, and metal ion–ligand coordination [7]. In the case of an electrochemical system, the substrate electrode WE1 acts as an electron source or sink and, eventually, as an additional coordination center. The electrode potential may be considered as a universal tuning source to fabricate and to address these molecular nanostructures.



Electrochemical Scanning Tunneling Microscopy, Fig. 7 Steady-state voltammogram for Au(111)-(1 × 1) in 0.05 M H_2SO_4 in the presence of 3 mM TMA, scan rate 10 mV·s^{-1}, together with typical EC-STM images of the TMA adlayers (size 10 × 10 nm): (*IIa*) hexagonal honeycomb phase, $E = -0.18$ V; (*IIb*) ribbon-type motif, $E = 0.03$ V; (*IIc*) herringbone motif, $E = 0.21$ V; (*III*) hydrogen-bonded linear dimers, $E = 0.34$ V; (*IV*) ordered chemisorbed striped phase of TMA at $E = 0.80$ V (Adapted from Ref. [7])

Structure motifs and processes of self-organization at surfaces can be investigated with atomic/molecular resolution employing EC-STM and STS.

A particular fascinating example of tailored molecular self-assembly at an electrified solid–liquid interface is benzene-1,3,5-tricarboxylic acid (trimesic acid (TMA); Fig. 7), a prototype material for engineering molecular nanostructures, on Au(111) in 0.05 M H_2SO_4 [7]. Depending on the assembly conditions and the applied electrode potential, five 2D-ordered molecular adlayers were observed by EC-STM (Fig. 7). At a negatively charged or uncharged Au (111) surface, the molecules are adsorbed flat and assemble with increasing coverage in four distinctly different hydrogen-bonded patterns: the low-coverage hexagonal honeycomb phase (IIa), sometimes even with guest molecules incorporated, a ribbon-type motif (IIb) and a herringbone motif (IIc) pattern, and rows of hydrogen-bonded linear dimers (III). At a positively charged surface, the TMA molecules change their orientation from flat to upright and one carboxyl group deprotonates. The resulting carboxylate group coordinates with the positively charged surface and rows of a new stacking structure are formed (IV). These structure transitions are reversible upon changing the electrode polarization. The detailed mechanisms were explored in a combined EC-STM and infrared spectroscopy study [7].

Concluding this paragraph one should emphasize that the interpretation of EC-STM images of molecular adlayers requires great caution because the observed contrast pattern provides information on the electronic structure of the entire

tunneling junction, and not a straightforward adlayer topography, as often naively assumed in the description of molecular adlayers.

Metal UPD and OPD

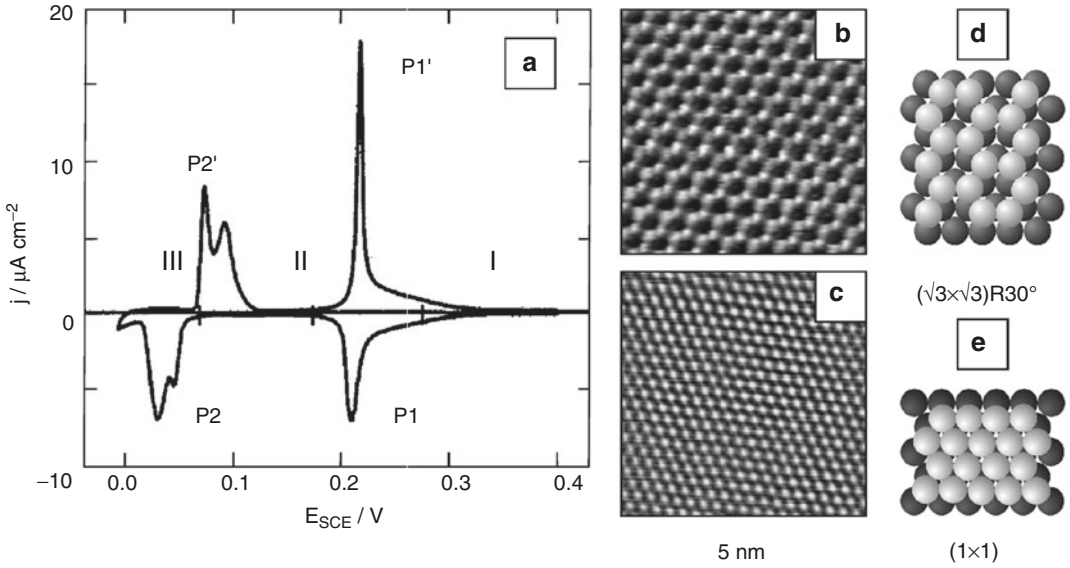
The deposition of metals from aqueous solutions has attracted the attention of electrochemists for more than a hundred years. This is related to the importance in many industrial-scale processes such as plating and metal refining. Contemporary frontiers in this field include the optimization of the electrochemical copper deposition in the fabrication of high-quality nanometer-wide integrated circuit interconnects (Damascene process) [19]. A primary characteristic of the metal deposition process is the equilibrium potential E_{eq}, at which the rates of metal deposition and dissolution are equal. The dependence of E_{eq} on the concentration of the metal ions Me^{n+} in a solution, $C_{Me}{}^{n+}$, is described by the Nernst equation:

$$E_{eq} = E^0 + \frac{RT}{nF}\ln(a_{Me^{n+}}) \cong E^0 + \frac{RT}{nF}\ln(C_{Me^{n+}}) \quad (3)$$

with E^0 being the standard potential of the redox pair Me^{n+}/Me and $(a_{Me^{n+}})$ as the thermodynamic activity of the metal ions in solution. When an external potential $E < E_{eq}$ is applied to the metal electrode, which is in contact with a metal ion Me^{n+}-containing solution, bulk deposition (or overpotential deposition (OPD)) of the metal Me takes place. At $E > E_{eq}$ the bulk dissolution of the metal phase is promoted. However, a (sub)monolayer of metal ions is deposited at $E > E_{eq}$ when the energy of the metal–substrate interaction is higher than that between the atoms of the deposited metal. This process is referred to as underpotential deposition (UPD) [8, 20]. Metal deposition onto a foreign substrate (e.g., Cu on Au) often starts with the formation of a monolayer at underpotentials, that is, at potentials positive to the Nernst potential for the respective bulk phase. UPD of metals is important as the first stage of many metal bulk deposition processes, to tailor electrocatalytic properties of metal surfaces as well as in fundamental electrochemical surface science. The structures and process involved were studied in detail by EC-STM, in combination with classical electrochemical and other structure-sensitive techniques for a wide range of systems and conditions [2, 3, 20].

As an example, Fig. 8 represents the archetype UPD system Cu on Au(111) in 0.05 M H_2SO_4 [20–22]. The cyclic voltammogram shows two well-defined pairs of current peaks P1/P1′ and P2/P2′ corresponding to energetically different adsorption/desorption processes in regions I to III. In the first step (P1), the transition between randomly adsorbed copper ions and (hydrogen) sulfate ions (region I) into an ordered layer of copper ions and co-adsorbed sulfate ions takes place. Among other structure-sensitive techniques, EC-STM revealed a $(\sqrt{3} \times \sqrt{3})R30^\circ$ pattern of bright features (region II) [22]. The correct interpretation of this contrast pattern is based on a surface X-ray scattering study of Toney et al. [23]. These authors demonstrated that Cu ions adsorb in hollow positions of the underlying Au(111) lattice and form a honeycomb-like hexagonal network with a 2/3 Cu coverage. Sulfate ions are positioned in the center of the hexagons with one S-O bond facing toward the electrolyte, which corresponds to a 1/3 sulfate coverage. The pattern resolved by EC-STM represents only the co-adsorbed anion structure. Further decrease of the potential leads to a second interfacial phase transition in this system, as indicated by the peak P2, and a pseudomorphic monolayer of Cu with the same structure as the underlying Au(111) surface is formed (region III). The potential range III is very narrow and limited by the onset of copper overpotential (bulk) deposition (OPD), which proceeds at $E < E_{eq}$ with $E_{eq} = 0.012$ V versus SCE.

OPD of Cu on Au(hkl) was among the first systems studied by EC-STM. The research on Cu-OPD and many other technologically important metals can be divided in three main directions addressing (1) the initial stages of bulk deposition, (2) the growth of the metal phase, and (3) the influence of additives on deposit morphology and deposition rate. Structure and dynamic studies



Electrochemical Scanning Tunneling Microscopy, Fig. 8 (**a**) Cyclic voltammogram of a Au(111)-(1 × 1) electrode in 0.1 M H_2SO_4 and 1 mM $CuSO_4$, scan rate 1 mV·s^{-1} (Adapted from Ref. [24]). (**b**, **c**) (√3 × √3) R30° and (1 × 1) Cu UPD adlayers on Au(111) (Adapted from Ref. [21]). The models of both phases are depicted in panels (**d**) and (**e**)

by EC-STM contributed substantially to the advancement of knowledge in all three target areas [2, 8, 20].

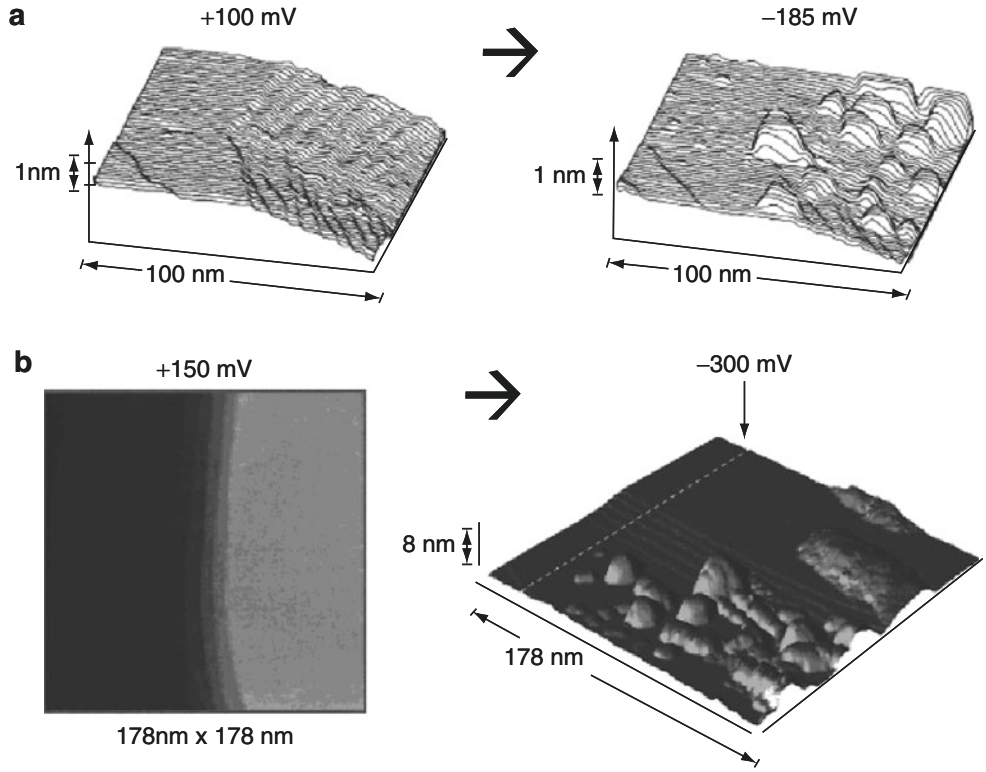
The morphology of the deposit during the initial stages of the OPD process depends on the nature and distribution of surface defects. For example, the deposition of Cu on Au(111) proceeds first by nucleation on surface defects, such as steps, kinks, and screw dislocation, and only at higher overpotentials also on terrace sites (Fig. 9). The nucleation-and-growth behavior leads to inhomogeneous film thicknesses and, hence, to rough metal overlayers. The presence of organic additives such as benzotriazole, thiourea, crystal violet, or polyethylene glycol, which act as so-called levelers and brighteners, modifies the deposition process in such a way that smooth and uniform overlayers emerge. As an example, the addition of crystal violet induces a quasi-two-dimensional growth, which causes the substrate to be quickly covered by a rather uniform Cu film on Au [22]. This particular study gives an illustration on the importance of additives for the optimization of deposition processes, such as that of Cu in the Damascene technology [19].

Current–Distance Spectroscopy: Measurements of the Tunneling Barrier

Current–distance STS represents a complementary approach for the characterization of the normal distribution of electric charges and adsorbates at electrochemical interfaces, as compared to STM imaging [21]. Particularly interesting insights on the electronic structure of adlayers were obtained by comparison of measured and simulated tunneling responses [10]. Figure 10a shows current–distance curves recorded on Au (111) in 0.1 M H_2SO_4 at different potentials. Unlike expected from a simple constant tunneling barrier model (cf. Eq. 1), the log (I)-Δz curves demonstrate pronounced deviations from linear behavior. They were also represented in terms of an effective barrier height (EBH), calculated as

$$\Phi_{\mathrm{eff}} = \kappa^{-2}\left(\frac{\partial \ln I}{\partial z}\right)^2 \qquad (4)$$

The profiles of the EBH obtained from curves in Fig. 10a are displayed in Fig. 10b. They exhibit minima and maxima with height and position depending on the electrode potential. The distance

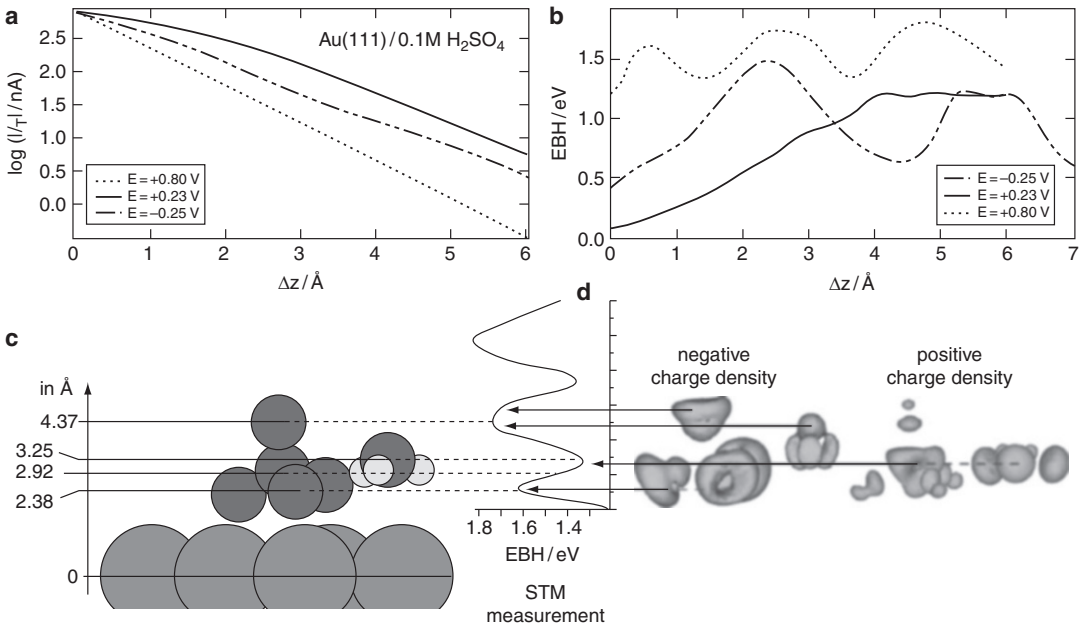


Electrochemical Scanning Tunneling Microscopy, Fig. 9 (**a**) STM images of Au(111)-(1 × 1) in 0.05 M H_2SO_4 + 5°10^{-5} M $CuSO_4$ at 0.1 V and −0.185 V versus Cu^{2+}/Cu; (**b**) Au(111)-(1 × 1) in 0.1 M H_2SO_4 + 10^{-3} M $CuSO_4$ at 0.15 V and −0.3 V versus Cu^{2+}/Cu (Adapted from Ref. [22])

scale in Fig. 10a, b is relative due to the experimental uncertainty of the absolute tip separation Δz in the STM experiment. To understand the origin of the variation of the EBH, structure and electron density distribution of the experimentally studied interfaces were modeled in DFT calculations [10]. As an example, Fig. 10c demonstrates the calculated optimal structure of the sulfate adlayer on Au(111) with the well-known (√3 × √7) R19.1° geometry (Fig. 5) and the corresponding distribution of positive and negative charge densities (Fig. 10d). Electron tunneling is facilitated by the positive local charge density and hindered by the negative ones. Therefore, the positions of the maxima and minima in the EBH profile were attributed to the maxima and minima of charge density (Fig. 10d). The distance scale of the experimental curves was shifted so that the features of the calculated charge distribution normal to the surface coincided with the features of the experimental curve. The report published in Ref. [10] demonstrates the power of complementary experimental and theoretical approaches to the study of charged interfaces by EC-STM/STS.

Single-Molecule Charge Transport and Switching

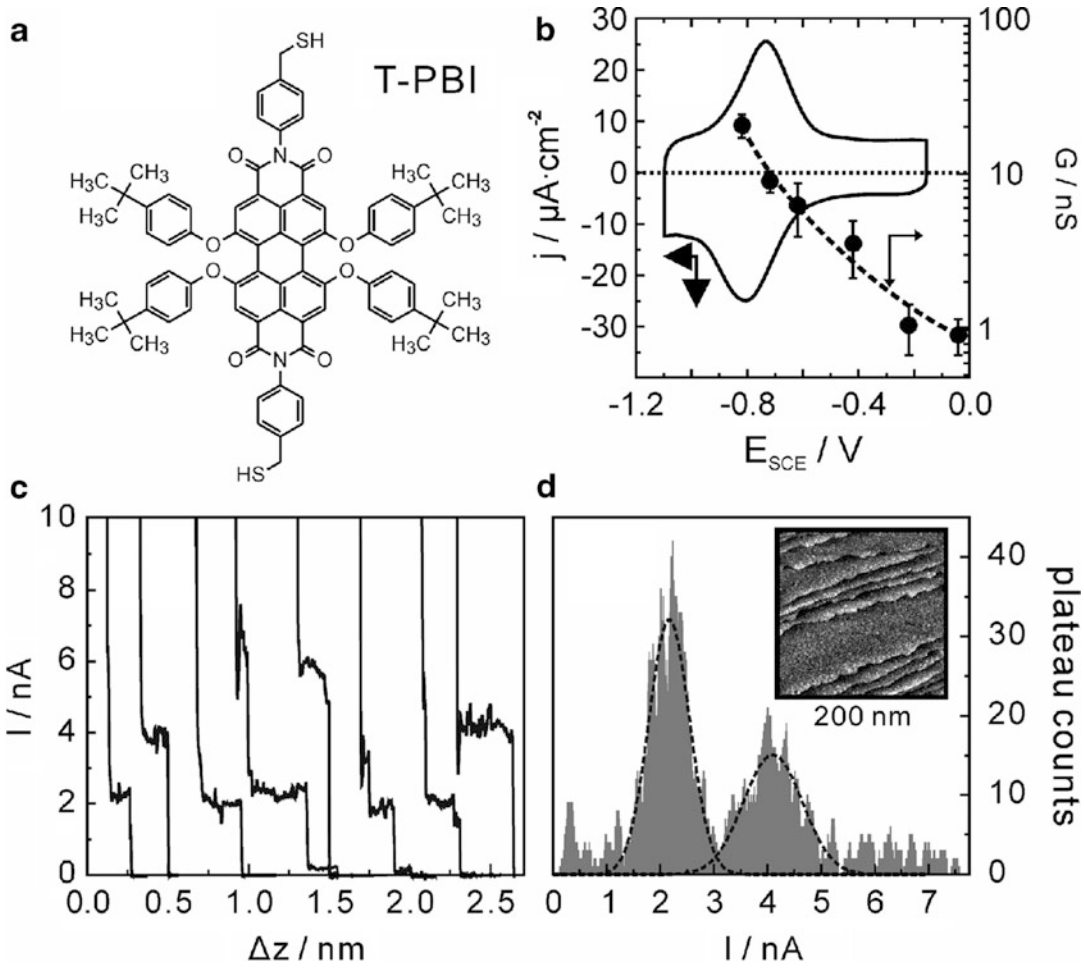
In this paragraph, an advanced application of EC-STM for studies of electrochemically controlled electron transport in single-molecule junctions is described [11]. The latter are formed by a molecule with two chemical anchoring groups trapped to two metal electrodes. The STM tip is used as an active tool for contacting one anchoring group of a molecule bound to the substrate electrode by the second anchoring group. The method is demonstrated for the redox-active molecule of thiol-terminated



Electrochemical Scanning Tunneling Microscopy, Fig. 10 (**a**) Experimental current–distance STS traces. (**b**) Corresponding effective barrier height profiles measured at the Au(111)/0.1 M H_2SO_4 interface at different potentials. (**c**) DFT-optimized structure of the sulfate adlayer adsorbed on Au(111), $E = 0.8$ V. (**d**) Comparison of calculated charge distribution for the structure shown in (**c**) and measured effective barrier height profiles (Adapted from Ref. [10])

perylene-3,4:9,10-tetra-carboxylic acid bisimide (T-PBI; Fig. 11a). The cyclic voltammogram of a T-PBI-modified Au(111) electrode (Fig. 11b) shows a broad current peak at $E^0 = 0.79$ V, which represents a reversible two-electron reduction/oxidation of the immobilized organic adlayer. A typical EC-STM image of a disordered densely packed T-PBI adlayer on an Au(111)-(1 × 1) surface is shown as inset in Fig. 11d.

Single Au |molecule| Au junction conductance measurements were carried out according to a "break junction" or "stretching" technique. A sharp Au STM tip, capable of imaging experiments with atomic resolution, was brought to a preset tunneling position. Subsequently, the STM feedback was switched off, and the tip approached the adsorbate-modified substrate surface at constant *x*-*y* position. The approach was stopped at a preset tunneling current of 5–20 μA, which corresponds to a strong tip–adlayer interaction. After a dwelling time of ~100 ms, sufficient to create molecular junctions between tip and substrate, the tip was retracted and the corresponding current–distance traces were recorded. Examples, acquired at $E_S = -0.82$ V and $E_b = 0.10$ V, are shown in Fig. 11c. The traces exhibit characteristic plateaus, which are separated by abrupt steps. These observations are attributed to the stochastic formation and breaking of individual respective multiple Au |T-PBI| Au junctions. Unlike in the case of a tunneling current, the current flowing through a molecular junction is rather constant during the separation of tip and sample electrodes and quickly drops by orders of magnitude when the junction is broken (Fig. 11c). The statistical analysis of several thousands of individual traces leads to conductance histograms, such as plotted in Fig. 11d. The histogram shows two pronounced peaks with positions close to 2 and 4 nA, with the second peak being broader and lower than the first one. The latter is a typical indication of molecular junctions corresponding to one and two molecules bound between electrodes.



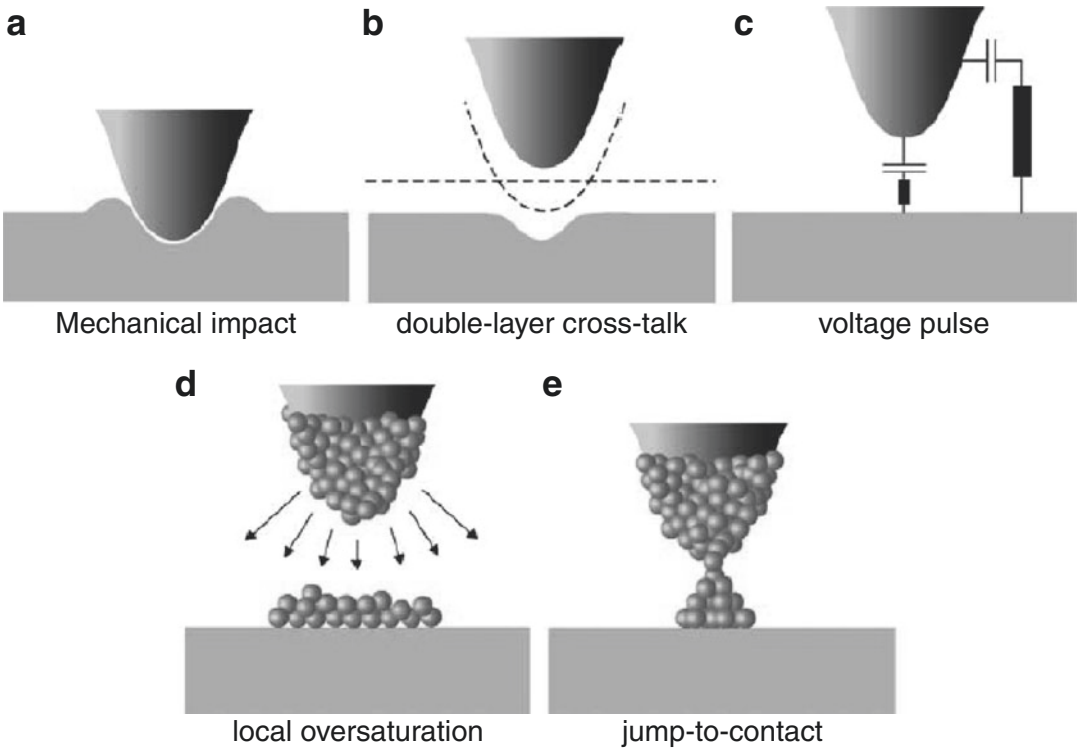
Electrochemical Scanning Tunneling Microscopy, Fig. 11 Single-molecule conductance experiments of T-PBI immobilized on Au(111)-(1 × 1) in 0.05 M $LiClO_4$, pH ≈ 12. (**a**) Chemical structure of T-PBI. (**b**) Cyclic voltammogram of a T-PBI monolayer, scan rate 0.2 $V \cdot s^{-1}$, and single-molecule conductance of T-PBI as a function of the substrate potential. (**c**) Typical current–distance retraction traces for T-PBI molecular junctions at $E_b = 0.10$ V, $E_S = -0.82$ V versus SCE. (**d**) Conductance histogram for T-PBI constructed from curves such as that presented in panel (**c**). The inset shows an STM image of a disordered T-PBI adlayer on Au(111) (200 × 200 nm, $I_T = 0.1$ nA; $E_S = -0.02$ V vs. SCE, $E_b = 0.1$ V) [11]

The single-junction conductance is estimated for this particular experiment at $E_S = -0.82$ V and $E_b = 0.10$ V as $G = 22 \pm 4$ nS. These experiments were repeated at different substrate potentials. Figure 11b illustrates the corresponding single-junction conductance as a function of substrate potential. The reduction of T-PBI upon decrease of the potential leads to an increase of the junction conductance by a factor of 20 over the whole measured potential range.

These experiments illustrate one of the frontier areas in electrochemical nanotechnology. They represent a direct approach toward exploring device-analog functions at electrified solid–liquid interfaces.

EC-STM-Based Nanostructuring

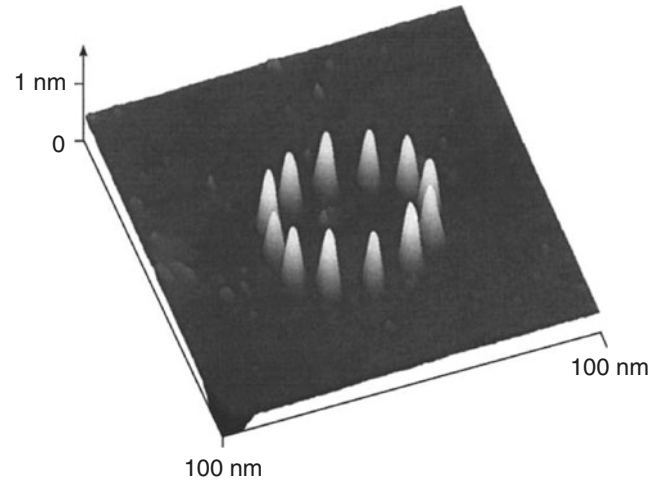
The STM tip can also be used as a tool to actively structure electrode surfaces [4, 13]. Several existing strategies are illustrated in Fig. 12: (a) formation



Electrochemical Scanning Tunneling Microscopy, Fig. 12 Various approaches to nanostructuring of metal electrodes by the EC-STM tip (Adapted from Ref. [13])

of active surface sites by mechanical impact of the tip, (b) local surface dissolution by double-layer cross talk upon application of a steady tip potential, (c) dissolution initiated by short voltage pulses, (d) deposition of metal upon the local oversaturation of solution with metal ions during a "burst-like" dissolution of the metal from the STM tip, and (e) transfer of metal clusters from the tip to the surface. While surface or monolayer defects as created by the mechanical impact of the tip are in general poorly defined and lead to non-reproducible surface features, the more controlled noncontact regimes allow fabrication of much more uniform, stable surface nanostructures with unique electric, magnetic, and catalytic properties. Fast dissolution of predeposited metal from the tip (Fig. 12d) can be used to create a local increase of the concentration of metal ions leading to metal deposition in the close vicinity of the tip (cf. Eq. 3). If the substrate potential E_S is only slightly more positive than E_{eq} as defined by the bulk concentration of metal ions, local metal deposition may occur. The generated metal clusters are metastable, as their dissolution at the substrate potential used is kinetically hindered [8].

In the "jump-to-contact" technique (Fig. 12e), metal predeposited on the tip is transferred to the substrate upon approaching the tip toward the surface without reaching mechanical contact [13]. Jump-to-contact occurs, and a metal bridge is formed between tip and the substrate. When the tip is retracted from the surface, the metal bridge breaks and leaves a small metal or alloy cluster on the surface (Fig. 13). Using this technique, small and uniform clusters of various metals can be generated on metal and semiconducting surfaces with high reproducibility and lateral precision having a wide range of unique properties. The tip potential is always held in the OPD range,



Electrochemical Scanning Tunneling Microscopy, Fig. 13 Cu clusters formed by the "jump-to-contact" technique on a Au(111) electrode in a solution of 0.05 M H_2SO_4 and 1 mM $CuSO_4$. The clusters are 0.8 nm high; the diameter of the circle is about 40 nm (Adapted from Ref. [4])

and provided that the metal deposition is sufficiently fast, the cluster transfer process can be repeated at kHz rates [13].

Cross-References

- ▶ AFM in Liquids
- ▶ Atomic Force Microscopy
- ▶ Charge Transport in Self-Assembled Monolayers
- ▶ Force Modulation in Atomic Force Microscopy
- ▶ Nanoindentation
- ▶ Nanoscale Properties of Solid–liquid Interfaces
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Plating
- ▶ Scanning Tunneling Microscopy
- ▶ Scanning Tunneling Spectroscopy
- ▶ Self-Assembly of Nanostructures

References

1. Siegenthaler, H.: STM in electrochemistry. In: Wiesendanger, R., Güntherodt, H.-J. (eds.) Scanning Tunneling Microscopy II, Surface Sciences, vol. 28, pp. 7–49. Springer, Berlin/Heidelberg (1995)
2. Gewirth, A.A., Niece, B.K.: Electrochemical applications of in situ scanning probe microscopy. Chem. Rev. **97**, 1129–1162 (1997)
3. Itaya, K.: In situ scanning tunneling microscopy in electrolyte solutions. Prog. Surf. Sci. **58**, 121–248 (1998)
4. Kolb, D.M., Simeone, F.C.: Characterization and modification of electrode surfaces by in situ STM. In: Bowker, M., Davies, P.R. (eds.) Scanning Tunneling Microscopy in Surface Science, Nanoscience and Catalysis, pp. 119–146. Wiley-VCH, Weinheim (2010)
5. Kolb, D.M.: Reconstruction phenomena at metal-electrolyte interfaces. Prog. Surf. Sci. **51**, 109–173 (1996)
6. Magnussen, O.M.: Ordered anion adlayers on metal electrode surfaces. Chem. Rev. **102**, 679–725 (2002)
7. Han, B., Li, Z., Li, C., Pobelov, I., Su, G., Aguilar-Sanchez, R., Wandlowski, T.: From self-assembly to charge transport with single molecules – an electrochemical approach. Top. Curr. Chem. **287**, 181–255 (2009)
8. Budevski, E., Staikov, G., Lorenz, W.J.: Electrochemical Phase Formation and Growth. Wiley-VCH, Weinheim (1996)
9. Oltra, R., Maurice, V., Akid, R., Marcus, P.: Local Probe Techniques for Corrosion Research. Woodhead Publishing, Cambridge (2007)
10. Simeone, F.C., Kolb, D.M., Venkatachalam, S., Jacob, T.: The Au(111)/Electrolyte interface: a tunnel-spectroscopic and DFT investigation. Angew. Chem. Int. Ed. **46**, 8903–8906 (2007)
11. Li, C., Mishchenko, A., Li, Z., Pobelov, I., Wandlowski, T., Li, X.Q., Würthner, F., Bagrets, A., Evers, F.: Electrochemical gate-controlled electron transport of redox-active single perylene bisimide molecular junctions. J. Phys. Condens. Matter **20**, 374122 (2008)
12. Zhang, J., Kuznetsov, A.M., Medvedev, I.G., Chi, Q., Albrecht, T., Jensen, P.S., Ulstrup, J.: Single-molecule electron transfer in electrochemical environments. Chem. Rev. **108**, 2737–2791 (2008)
13. Kolb, D.M., Simeone, F.C.: Electrochemical nanostructuring with an STM: a status report. Electrochim. Acta **50**, 2989–2996 (2005)

14. Bard, A.J., Faulkner, L.: Electrochemical Methods: Fundamentals and Applications. Wiley, New York (2001)
15. Trasatti, S.: The absolute electrode potential: an explanatory note. Pure Appl. Chem. **58**, 955–966 (1986)
16. Tansel, T., Magnussen, O.M.: Video STM studies of adsorbate diffusion at electrochemical interfaces. Phys. Rev. Lett. **96**, 026101 (2006)
17. Dietterle, M.: Untersuchungen zur elektrolytischen Cu-Abscheidung und zur Dynamik von Stufenkanten auf niedrigindizierten Ag-Elektroden: Eine in-situ STM Studie, Ph. D. thesis, Universität Ulm (1996)
18. Giesen, M.: Step and Island dynamics at solid/vacuum and solid/liquid interfaces. Prog. Surf. Sci. **68**, 1–153 (2001)
19. Vereecken, P.M., Binstead, R.A., Deligianni, H., Andricacos, P.C.: The chemistry of additives in damascene copper plating. IBM J. Res. Dev. **49**, 3–18 (2005)
20. Herrero, E., Buller, L.J., Abruña, H.D.: Underpotential deposition at single crystal surfaces of Au, Pt, Ag and other materials. Chem. Rev. **101**, 1897–1930 (2001)
21. Nagy, G., Wandlowski, T.: Double layer properties of Au(111)/H_2SO_4 (Cl) + Cu^{2+} from distance tunneling spectroscopy. Langmuir **19**, 10271–10280 (2003)
22. Schneeweiss, M., Kolb, D.: The initial stages of copper deposition on bare and chemically modified gold electrodes. Phys. Stat. Sol. A **173**, 51–71 (1999)
23. Toney, M.F., Howard, J.N., Richer, J.: Electrochemical deposition of copper on gold electrode in sulfuric acid: resolution of interfacial structure. Phys. Rev. Lett. **75**, 4472–4475 (1995)
24. Kibler, L.A.: Preparation and Characterization of Noble Metal Single Crystal Electrode Surfaces. ISE, Barcelona (2003)

Electrode–Organic Interface Physics

Michael G. Helander[2], Zhibin Wang[2] and Zheng-Hong Lu[1,2]
[1]Department of Physics, Yunnan University, Kunming, Yunnan, People's Republic of China
[2]Department of Materials Science and Engineering, University of Toronto, Toronto, ON, Canada

Synonyms

Band alignment; Energy-level alignment; Metal–organic interfaces

Definition

Electrode–organic interface physics describes the formation of an electric contact between an organic semiconductor material and an electrode, typically a metal. The alignment of the energy-levels of the organic material with the energy-levels of the electrode dictates the energetic barrier which must be overcome to move electrical charges between the electrode and organic material. Depending on the magnitude of the energetic barrier at the interface, the injection of charge across the interface will be different.

Introduction

Organic electronic devices are electrical devices, such as light emitting diodes, photovoltaics, or thin film transistors, in which the active materials consist entirely of organic materials with semiconducting properties. Since organic materials consist of discrete molecular units, there is no need for long range order between adjacent molecules or polymer chains. Organic semiconductors therefore have the advantage over traditional inorganic semiconductors, such as Si or GaAs, in that they do not require perfect single crystal films to operate in real devices. Complicated multilayer structures with nanometer-scale thicknesses can thus be easily fabricated from organic materials using low-cost roll-to-roll printing or vacuum coating techniques. However, the discrete nature of organic semiconductors also implies that they typically contain almost no *intrinsic* charge carriers (i.e., electrons or holes), and thus act as insulators until electrical charges are injected into them. In electrical device applications this means that all of the holes and electrons within a device must be injected from the anode and cathode respectively. As a result, device stability, performance, and lifetime are greatly influenced by the interface between the organic materials and the electrode contacts.

Energy-Level Alignment

Of particular interest for device applications is the energy-level alignment at electrode–organic

interfaces, or in other words the alignment of the molecular orbitals of the organic to the energy-levels of the electrode (typically the Fermi level). For traditional inorganic semiconductors the energy-level alignment at semiconductor–metal contacts is typically described by the Schottky barrier height, which is the potential barrier between the Fermi level of the metal and the conduction band or valence band edge of the semiconductor. Analogous to the Schottky barrier height at semiconductor–metal interfaces the energy-level alignment at electrode–organic interfaces is also typically represented by an injection barrier height ϕ_B. This injection barrier represents the potential barrier which must be overcome to move an electrical charge from the Fermi level of the electrode into the highest occupied molecule orbital (HOMO) or lowest unoccupied molecular orbital (LUMO) of the organic (or vice versa). Metal–organic interfaces were the first contacts to be extensively characterized in organic electronics, owing to their similarity to Schottky contacts in traditional semiconductors.

Work Function

The work function of a uniform surface of a conductor is defined as the minimum energy required to remove an electron from the interior of the conductor to just outside the surface, where "just outside" refers to a distance that is large enough that the image force is negligible, but small compared to the physical dimensions of the crystal (typically $\sim 10^{-4}$ cm). In other words, the work function is the difference between the electrochemical potential $\overline{\mu}$ of electrons in the bulk and the electrostatic potential energy $-e\Phi_{\text{vac}}$ of an electron in the vacuum just outside the surface [1]:

$$-e\phi_m = -e\Phi_{\text{vac}} - \overline{\mu}. \tag{1}$$

The energy-level corresponding to $-e\Phi_{\text{vac}}$ is referred to as the *local* vacuum level E_{vac} and is not to be confused with the vacuum level at infinity E_0, which represents an electron at rest at infinity (i.e., zero potential energy) [2]. Since the Fermi level E_F relative to the vacuum level at infinity E_0 is the electrochemical potential of electrons in the conductor [3],

$$E_0 - E_F = -\overline{\mu}, \tag{2}$$

the work function given by Eq. 1 is equivalent to the difference in potential energy of an electron between the local vacuum level E_{vac} and the Fermi level:

$$e\phi_m = E_{\text{vac}} - E_F. \tag{3}$$

From Eq. 3 it is clear that if the local vacuum level E_{vac} varies across a sample, as is the case for the different faces of a single crystal, the work function will also vary.

Schottky Contacts

The energy-level alignment at semiconductor–metal contacts has been extensively studied owing to their importance in microelectronics. In the Schottky–Mott limit the vacuum level of the semiconductor and metal align, forming a region of net space charge at the interface. No charge is transferred across the interface and hence the semiconductor bands are forced to bend to accommodate the potential difference. The Schottky barrier height for electrons is then given by the difference between the work function of the metal ϕ_m and the electron affinity of the semiconductor χ (i.e., the conduction band edge). However, the Schottky–Mott limit is rarely observed at most metal–semiconductor interfaces due to charge transfer across the interface (i.e., an interfacial dipole). As a result, the Fermi level cannot move freely in the bandgap of the organic and thus tends to be pinned by the dipole layer. The interface slope parameter is a convenient parameter that quantifies this phenomenon, and is commonly defined in terms of the Schottky barrier height for an electron or hole [4]:

$$S_\phi = \frac{d\phi_{\text{Bn}}}{d\phi_m} = -\frac{d\phi_{\text{Bp}}}{d\phi_m}, \tag{4}$$

where S_ϕ is the interface slope and ϕ_{Bn} and ϕ_{Bp} are the barrier heights for electrons and holes respectively. In the Schottky–Mott limit $S_\phi = 1$ (vacuum level alignment); and for Fermi level pinning $S_\phi = 0$ (interfacial dipole).

In the case of organic semiconductors many studies have experimentally demonstrated a strong correlation between the metal work function and the barrier height for holes or electrons. In analogy to inorganic semiconductors, in the Schottky–Mott limit the vacuum level of the organic and metal align, forming a region of net space charge at the interface. The barrier height for holes is then given by the offset between the HOMO of the organic E_{HOMO} and the work function of the metal:

$$e\phi_{\mathrm{Bp}} = E_{\mathrm{HOMO}} - e\phi_m. \tag{5}$$

However, since organic semiconductors contain almost no free charge carriers, band bending in the organic cannot fully accommodate the potential difference. Hence, the Schottky–Mott limit is rarely observed at metal–organic interfaces due to the formation of a strong interfacial dipole between the metal surface and organic molecules. Accounting for the effect of an interfacial dipole, the injection barrier for holes is then given by the following:

$$e\phi_{\mathrm{Bp}} = E_{\mathrm{HOMO}} - e(\phi_m - \Delta). \tag{6}$$

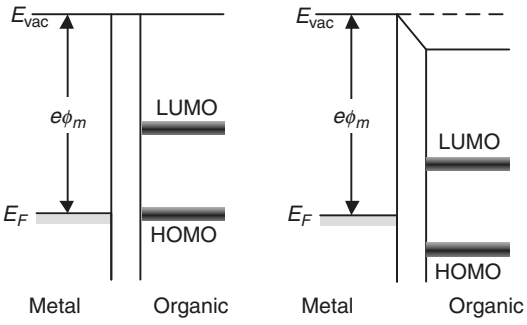
From the above equation it is clear that a negative dipole reduces the hole injection barrier, while a positive dipole enhances electron injection. From Eq. 6 an effective metal work function can be defined as the difference between the pristine metal work function and the dipole:

$$\phi_{m,\mathrm{eff}} = \phi_m - \Delta, \tag{7}$$

where $\phi_{m,eff}$ is the effective metal work function (Fig. 1).

Interface Dipole Theory

Deviation from the Schottky–Mott limit was first described by Bardeen [5] as the result of a large density of surface states; charge transfer between the metal and these surface states acts to pin the Fermi level. In fact surface states were first postulated by Tamm and Shockley prior to the Schottky model, but were largely ignored. Heine [6] later



Electrode–Organic Interface Physics, Fig. 1 Schematic diagram of energy-level alignment at the metal–organic interface in the Schottky–Mott limit (*left*) and with an interfacial dipole (*right*)

demonstrated that surface states do not exist in the gap of most metal–semiconductor interfaces, or at least not for modern devices produced by the high vacuum deposition of metal onto a clean semiconductor surface. In Bardeen's time semiconductor–metal contacts were formed by laminating a piece of metal and semiconductor, which inherently results in a "dirty" interface. Heine postulated that for a "clean" interface, gap states are induced in the semiconductor as a result of the rapidly decaying tail of the electronic wave function from the metal. These gap states would also tend to pin the Fermi level same as for Bardeen's surface states. Tersoff [7] later identified the gap states as metal-induced gap states (MIGs), which derive from the virtual gap states (VIGs) of the semiconductor complex band structure.

Regardless of their specific origin, interface states are independent of the energy-levels in the semiconductor, and hence can either be donor like or acceptor like when close to the valence or conduction bands respectively. A charge neutrality level ϕ_{CNL} is defined as the point at which the interface states are equally donor- and acceptor-like (i.e., the transition point from donor to acceptor states). Cowley and Sze [8] applied this notion to derive the dependence of the Schottky barrier height upon the metal work function. They argued that at semiconductor–metal interfaces the charge neutrality level of the semiconductor will tend to align with the Fermi level of the metal as a result of charge transfer between the metal and the interface states, forming an interfacial dipole. In other

words, the redistribution of charge at the interface (dipole) offsets the vacuum level of the metal and semiconductor, resulting in an effective metal work function that is different from the vacuum metal work function:

$$\phi_{m,\mathrm{eff}} = \phi_{\mathrm{CNL}} - S_\phi(\phi_m - \phi_{\mathrm{CNL}}). \quad (8)$$

The interfacial dipole is then the difference between $\phi_{m,eff}$ and ϕ_m:

$$\Delta = (1 - S_\phi)(\phi_m - \phi_{\mathrm{CNL}}). \quad (9)$$

Interface Slope Parameter

In their analysis, Cowley and Sze assumed a constant continuum of interface states across the semiconductor gap. Based on this assumption, the interface slope parameter, which in this context represents the dielectric screening strength of the semiconductor, is given by:

$$S_\phi = \frac{\varepsilon_i}{\varepsilon_i + e\delta D_s}, \quad (10)$$

where ε_i is the interface permittivity, δ is the characteristic thickness of the interface (atomic length), and D_s is the density of interface states. However, since these parameters can only be measured indirectly, accurate determination of the interface slope parameter using Eq. 10 is problematic at best. Luckily, Mönch [9] empirically found that the interface slope parameter (for weakly interacting interfaces) is dependent on the optical dielectric constant:

$$S_\phi = \frac{1}{1 + 0.1(\varepsilon_\infty - 1)^2}, \quad (11)$$

where ε_∞ is the optical dielectric constant (high frequency limit of the dielectric function). One important caveat to consider for Eq. 11 is that it only accurately describes the interface slope parameter for weakly interacting and laterally homogenous Schottky contacts. If interfacial chemical reactions occur, or if the interface is patchy, then S_ϕ often deviates from the ideal theoretical dielectric value.

Recently it was shown that organic semiconductors also follow the same empirical relationship between S_ϕ and ε_∞ described by Eq. 11, same as for Schottky contacts with solid Xenon. However, many organic molecules either interact fairly strongly with clean metal surfaces or do not wet the surface of metals uniformly. The experimentally determined value of S_ϕ for clean metal surfaces is, therefore, often different than the ideal theoretical value described by Eq. 11. For example, the experimental value of S_ϕ for C_{60} is nearly 0, while the theoretical value is 0.53 assuming $\varepsilon_\infty = 4$. In this case it is well known that C_{60} interacts strongly with most clean metal surfaces, which results in complete pinning of the Fermi level (i.e., $S_\phi = 0$). For terminated surfaces, such as an oxidized metal surface (e.g., Al/Al_2O_3), the interaction between the metal and organic is screened by the adsorbed surface termination layer. In this case, S_ϕ approaches the ideal theoretical value for a weakly interacting interface. Metal oxides in particular form ideal contacts with most organics since the first layer of adsorbed molecules tends to wet the oxide surface, forming a laterally homogeneous contact. One important consequence of this is that a clean metal and oxide-terminated metal will behave differently in contact with an organic semiconductor, even if the effective metal work function of the oxide-terminated metal is the *same* as the work function of the clean metal surface. In general, the clean metal will pin the Fermi level stronger than the oxide-terminated metal.

Since the interfacial dipole is indicative of charge transfer between the metal and semiconductor, electronegativity is often used in place of work function to describe the dependence of barrier height on the contacting metal. In this case a linear dependence of the injection barrier on the electronegativity *difference* of the metal and semiconductor is observed:

$$\phi_{m,\mathrm{eff}} = \phi_{\mathrm{CNL}} + S_x(X_m - X_s), \quad (12)$$

where X_m and X_s are the electronegativity of the metal and semiconductor respectively and S_x is

the electronegativity-based interface slope parameter given by:

$$S_x = \frac{d\phi_{\mathrm{Bn}}}{dX_m} = \frac{d\phi_{\mathrm{Bp}}}{dX_m}. \tag{13}$$

One important distinction between the two different interface slope parameters is that S_ϕ is a dimensionless quantity with an upper limit of 1, while S_x on the other hand is *not* dimensionless and has an arbitrary upper limit depending on the units of electronegativity used. There have been several attempts to link the two interface slope parameters via a constant of proportionality:

$$S_\phi = A_x S_x, \tag{14}$$

where the value of A_x varies depending on the electronegativity scale used. In the case of organic semiconductors, there is no clear definition of electronegativity and hence previous attempts to relate injection barriers at metal–organic interfaces to electronegativity values are in general not very useful.

Metal–Dielectric–Organic Interfaces

Although interface dipole theory can explain the energy-level alignment at metal–organic interfaces in most device applications, the metal contacts are often modified with various dielectric buffer layers to improve charge injection. Dielectric buffer layers, often referred to as electron or hole injection layers (EIL or HIL) since they dramatically enhance carrier injection, typically fall into one of three categories: (1) wide bandgap insulators such as LiF, (2) semiconducting transition metal oxides such as V_2O_5, or (3) organic molecules, such as fullerene (C_{60}). Despite the appreciable difference in material properties for these three different types of dielectrics, one important similarity has been consistently reported in the literature. For all three types of dielectrics, strong interfacial dipoles (i.e., a rapid change in vacuum level across the interface) have been reported at the interfaces formed with metals and organics.

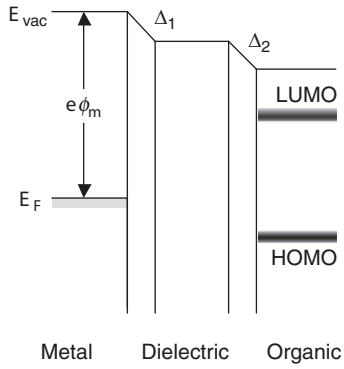
The energy-level alignment at real electrode–organic interfaces will therefore typically depend on the properties of the underlying metal as well as the dielectric buffer layer. A similar situation occurs at metal–oxide–semiconductor (MOS) interfaces used in advanced high-κ dielectric metal gate MOS field effect transistors (MOSFETs). It was discovered that the effective work function of the metal gate was strongly influenced by the properties of the oxide dielectric. Robertson [10] proposed that interface dipole theory could be applied to each of the interfaces (i.e., metal–oxide and oxide–semiconductor) to describe the trend in effective work function, or in other words that Eq. 12 could be applied to both interfaces to describe the net effective work function. A similar approach can thus be applied to metal–dielectric–organic interfaces to describe the energy-level alignment.

In order to calculate the net effect of the oxide layer on the injection barrier, the two interfaces can be considered in series. At the metal–dielectric interface the interaction between the dielectric and the metal creates an interfacial dipole layer that modifies the work function of the pristine metal surface. This results in the organic molecules deposited on top of the dielectric modified metal experiencing an effective metal work function as given by Eq. 12. Treating the metal–oxide as a layered anode structure, the effective metal work function can be plugged back into Eq. 12 to extract the effective work function of the metal–oxide after consideration of the effect of the organic layer:

$$\begin{aligned}\phi_{m,\mathrm{eff}} = &\left(1 - S_\phi^{\mathrm{org}}\right)\phi_{\mathrm{CNL}}^{\mathrm{org}} \\ &+ S_\phi^{\mathrm{org}}\left[\phi_{\mathrm{CNL}}^{\mathrm{ox}} + S_\phi^{\mathrm{ox}}(\phi_m - \phi_{\mathrm{CNL}}^{\mathrm{ox}}\right],\end{aligned} \tag{15}$$

where S_ϕ^{org}, S_ϕ^{ox}, $\phi_{\mathrm{CNL}}^{\mathrm{org}}$, and $\phi_{\mathrm{CNL}}^{\mathrm{ox}}$ are the interface slope parameters and charge neutrality levels of the organic and oxide respectively. The total combined dipole across the interface can then be calculated using Eq. 9. The barrier height for holes is then taken as the difference between this combined effective work function and the HOMO of the organic (Fig. 2).

Dielectrics typically have $S_\phi < 1$, which implies that the Fermi level is not completely



Electrode–Organic Interface Physics, Fig. 2 Schematic diagram of energy-level alignment at a metal–dielectric–organic interface

pinned at most metal–dielectric–organic interfaces. As a result, the barrier height at the interface is partially dependent on the work function of the underlying metal. However, in a few special cases S_ϕ can approach unity, implying that the Fermi level is pinned and hence the energy-level alignment at the interface is independent of the underlying metal. Buckminsterfullerene (C_{60}) is one example of such a dielectric layer. In contact with clean metal surfaces C_{60} can accept up to six electrons from the metal, becoming itself partially metallic. As a result, the optical dielectric constant (ε_∞) of the C_{60} approaches infinity, such that S_ϕ goes to 1 as given by Eq. 11. Therefore C_{60} can be used as a "universal" charge injection layer as it pins the Fermi level of any metal to its charge neutrality level.

Charge Injection and Transport

Up to this point only the barrier height at electrode–organic interfaces has been discussed. However, in real devices what is measured is the flow of current through the device as a function of applied bias. Therefore how the barrier height discussed above affects the injection and transport of charge in real devices is of interest. At traditional semiconductor–metal interfaces the process of charge injection is typically divided into two different regimes, space charge limited current (SCLC) and injected limited current (ILC).

Space Charge Limited Current

The first regime, which represents the maximum current that a semiconductor can sustain in the bulk (i.e., the amount of carriers in thermal equilibrium), is called the SCLC. A semiconductor–metal contact capable of injecting enough charges to sustain the SCLC is called an Ohmic contact. In most traditional semiconductors an Ohmic contact is assumed to represent a zero barrier height. One significant feature of SCLC is that the spatial distribution of electric field is described by $F(x) \propto x^{1/2}$, where x is the distance from the charge-injecting contact. The electric field at the interface of an Ohmic contact is therefore equal to zero. For unipolar transport (i.e., transport of only one type of charge carrier) in a perfect insulator (no intrinsic carriers) without traps the SCLC is given by the Mott-Gurney law [11]:

$$J = \frac{9}{8}\varepsilon_0\varepsilon\mu\frac{V^2}{d^3}, \tag{16}$$

where V is the applied voltage, d is the thickness of the film, and μ is the field-independent mobility.

With further consideration of an exponential distribution of trap states, the trap-charged limited current (TCLC) based on Eq. 16 follows:

$$J \propto \mu\frac{V^{l+1}}{d^{2l+1}}, \tag{17}$$

where l is a parameter derived from the trap distribution. Notice that the injection of charge depends only on the mobility, thickness of the semiconductor, and applied voltage as the barrier height is assumed to be zero.

In the case of organic semiconductors, however, the mobility typically depends on electric field. In disordered organic materials, it is believed that all electronic states are localized and participate in conduction through thermally activated hopping, which yields a Poole–Frenkel-like field dependence of the mobility:

$$\mu(F) = \mu_0\exp\left(\beta\sqrt{F}\right). \tag{18}$$

Under the assumption of a Poole–Frenkel dependence, an approximation to the SCLC for a field-dependent mobility is given by [12]:

$$J_{\mathrm{SCLC}} = \frac{9}{8}\varepsilon\varepsilon_0\mu_0 \exp\left(0.89\beta\sqrt{\frac{V}{d}}\right)\frac{V^2}{d^3}. \quad (19)$$

There has been significant experimental evidence that this model does indeed describe the SCLC in many organic semiconductors. One interesting caveat however due to the hopping-based transport in organic materials is that the barrier height for an Ohmic contact does not have to be zero in the typically operating electric field range of most device applications. Specific criteria for determining this critical threshold barrier height will be discussed in detail below.

Injection-Limited Current

The second regime of charge injection in semiconductors represents the situation where a significant barrier height exists at the interface, such that the transport of charge in the semiconductor is limited by the injection of charge at the semiconductor–metal contact. The ILC at Schottky contacts has traditionally been described by one of two processes. For high electric field strength and large barrier heights, Fowler–Nordheim (FN) tunneling, also known as field emission, dominates the injection current [4]:

$$J_{\mathrm{FN}} = \frac{q^3F^2}{8\pi h\phi_B}\exp\left[-\frac{8\pi\sqrt{2m^*}{\phi_B}^{3/2}}{3qhF}\right] \quad (20)$$

where F is the electric field strength, ϕB is the barrier height, m^* is the effective mass of an electron or hole in the semiconductor, q is the electronic charge, and h is the Planck constant. For thermally activated processes at lower electric field, Richardson–Schottky (RS) emission, also known as thermionic emission, dominates the injection current [4]:

$$J_{\mathrm{RS}} = A^*T^2\exp\left[-\frac{\left(\phi_B - \sqrt{qF/4\pi\varepsilon}\right)}{k_BT}\right] \quad (21)$$

where A^* is the modified Richardson constant, ε is the permittivity of the semiconductor, k_B is the Boltzmann constant, and T is the temperature.

In the case of organic semiconductors, the ILC was early on modeled as diffusion-limited thermionic emission. However, both FN tunneling and RS emission are derived for delocalized Bloch waves in the semiconductor, whereas the electronic states in most organic semiconductors are highly localized due to their discrete molecular nature. In addition, at room temperature FN tunneling typically requires electric field strengths several orders of magnitude greater than what are observed in real organic devices ($\sim 10^5$ V/cm).

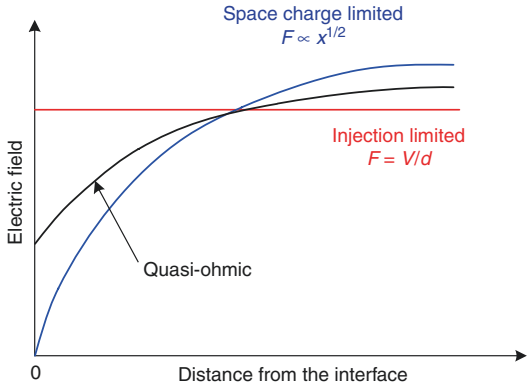
To overcome these deficiencies Scott and Malliaras [13] have proposed a modified thermionic emission model for the ILC in organic semiconductors based on the original solution to the drift-diffusion equation for injection into a wide bandgap intrinsic semiconductor solved by Emtage and O'Dwyer [14]. This model determines the equilibrium contributions to the current density for charge carriers recombining with their own image analogous to Langevin recombination of an electron–hole pair in the bulk:

$$J_{\mathrm{ILC}} = 4N_0\psi^2 e\mu F\exp\left(\frac{-e\phi_B}{k_BT}\right)\exp\left(f^{1/2}\right), \quad (22)$$

where N_0 is the density of chargeable sites in the organic film, ϕ_B is the barrier height, F is the electric field at the charge-injecting contact, μ is the electric field-dependent mobility, k_B is the Boltzmann constant, T is temperature, e is the electron charge, and ψ is a function of the reduced electric field ($f = e^3F/4\pi\varepsilon k_B^2T^2$):

$$\psi = f^{-1} + f^{-1/2} - f^{-1}\left(1 + 2f^{1/2}\right)^{1/2} \quad (23)$$

There has been significant experimental evidence that this model does indeed describe the ILC in many organic semiconductors. Also, under ILC conditions the spatial electric field distribution is assumed to be uniform such that $F(x) = V/d$; whereas for SCLC the value of V/d only gives the average value of the electric field.



Electrode–Organic Interface Physics, Fig. 3 Calculated electric field at the metal–organic interface as a function of the barrier height for a typical organic semiconductor with a mobility of 10^{-4}–10^{-3} $cm^2\ V^{-1}\ s^{-1}$ and a thickness of 1,000 nm

This suggests that SCLC can be distinguished from ILC by the electric field at the electrode–organic interface.

Quasi-Ohmic Contact

The two conditions discussed above, SCLC and ILC, correspond to the upper and lower limits respectively. What about the case in between SCLC and ILC, when the electric field at the electrode contact is $0 < F(x = 0) < V/d$ (see Fig. 3)? This indicates that there is in fact no clear boundary between SCLC and ILC, but rather an intermediate regime in between referred to as "quasi-Ohmic." Since the quasi-Ohmic regime exhibits characteristics of both SCLC and ILC, it is incorrect to apply the models for either case. Moreover, it is important to note that the boundaries of the quasi-Ohmic regime are dependent on the field-dependent carrier mobility μ, the barrier height ϕ_B, the applied voltage V, and the device thickness d. As a result the same electrode–organic contact may display characteristics of ILC, quasi-Ohmic, and SCLC in different ranges of applied bias. A new approach is thus required to properly describe the quasi-Ohmic regime.

As discussed above, the quasi-Ohmic regime in between SCLC and ILC cannot be described using traditional models. As a result a new approach that includes both the injection at the interfaces and the transport in the bulk of the organic is required to deal with this special case. One approach to deal with this situation is to use a time domain simulation under the transport models developed for inorganic semiconductors (i.e., the drift-diffusion and Poisson equations) to describe the distribution of electric field at steady state.

In the simulation, space x and time t are discrete. The injection current density described by Eqs. 22 and 23 serves as a boundary condition at $x = 0,\ t > 0$ (i.e., the charge-injecting contact). The time-dependent continuity equation follows:

$$\frac{\partial p(x,t)}{\partial t} = -\frac{1}{e}\frac{\partial J(x,t)}{\partial x}, \tag{24}$$

where p is the total density of holes and J is the conduction current density. The relation between the electric field and the charge density can be expressed by the Poisson equation as:

$$\frac{\varepsilon_r \varepsilon_0}{e}\frac{\partial F(x,t)}{\partial x} = p(x,t). \tag{25}$$

In Eq. 25, ε_r and ε_0 are the dielectric constants of the organic and the dielectric permittivity in vacuum respectively. The conduction current density can be calculated through the drift-diffusion equation:

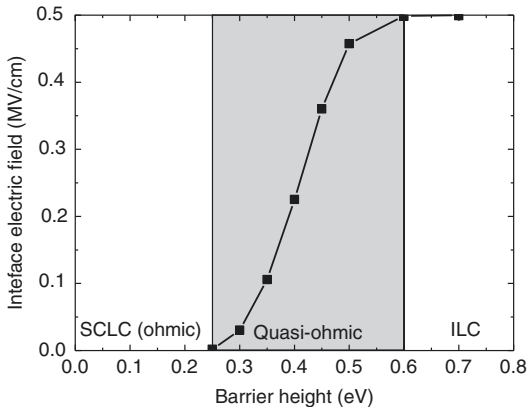
$$J(x,t) = ep(x,t)\mu(F)F - eD\frac{\partial p(x,t)}{\partial x}, \tag{26}$$

where D is the diffusion coefficient, which can be obtained from Einstein's relation (as a function of the field-dependent mobility):

$$D = \frac{\mu k_B T}{e}. \tag{27}$$

The other boundary conditions are:

$$\begin{cases} V = \int_0^d F dx \\ F(x, t = 0) = V/d, \\ p(x, t = 0) = 0 \\ J(x, t = 0) = 0 \end{cases} \tag{28}$$



Electrode–Organic Interface Physics, Fig. 4 Calculated electric field at the metal–organic interface as a function of the barrier height for a typical organic semiconductor with a mobility of 10^{-4}–10^{-3} $cm^2\ V^{-1}\ s^{-1}$ and a thickness of 1,000 nm

where d is the thickness of the film and V is the applied voltage. It is noted that the transient current density J_t is contributed by the displacement current and the response of the charge carrier density as:

$$J_t(x,t) = \varepsilon_r \varepsilon_0 \frac{\partial F(x,t)}{\partial t} + J(x,t). \qquad (29)$$

From Eq. 29 the total transient current at $x = d$ can be calculated until the steady state is reached. Also, the spatial distribution of electric field at steady state can be obtained from the simulation.

From these simulation results the boundaries of the quasi-Ohmic regime (i.e., the lower limit of SCLC and the upper limit of ILC) can then be defined as the convergence of the simulation results with the SCLC from Eq. 19 and the ILC from Eq. 22. Also, these boundaries represent the strict limits for SCLC and ILC in terms of the electric field at the charge-injecting contact. Figure 4 shows an example of the calculated electric field at the electrode–organic interface for a typical organic semiconductor with a mobility of 10^{-4}–10^{-3} $cm^2\ V^{-1}\ s^{-1}$. The calculation shows that the interfacial electric field depends strongly on the barrier height in the quasi-Ohmic regime. The interfacial electric field approaches zero and the average values (i.e., $F = V/d$) at low ($< \sim 0.25$ eV) and high ($> \sim 0.55$ eV) barrier height which respectively correspond to SCLC regime and ILC regime.

Future Directions

The energy-level alignment and charge injection at electrode–organic interfaces is based on the classical device physics developed for inorganic semiconductors. It can be applied to a wide variety of different materials and interfaces, even newly developed organic materials.

Cross-References

▶ Organic Photovoltaics: Basic Concepts and Device Physics

References

1. Herring, C., Nichols, M.H.: Thermionic emission. Rev. Mod. Phys. **21**, 185 (1949)
2. Hagstrum, H.D.: The determination of energy-level shifts which accompany chemisorption. Surf. Sci. **54**, 197 (1976)
3. Van Rysselberghe, P.: A note on work functions and chemical potentials. J. Chem. Phys. **21**, 1550 (1953)
4. Sze, S.M., Ng, K.K.: Physics of Semicondcutor Devices. Wiley, Hoboken (2007)
5. Bardeen, J.: Surface states and rectification at a metal semi-conductor contact. Phys. Rev. **71**, 717 (1947)
6. Heine, V.: Theory of surface states. Phys. Rev. **138**, A1689 (1965)
7. Tersoff, J.: Schottky barrier heights and the continuum of gap states. Phys. Rev. Lett. **52**, 465 (1984)
8. Cowley, A.M., Sze, S.M.: Surface states and barrier height of metal–semiconductor systems. J. Appl. Phys. **36**, 3212 (1965)
9. Mönch, W.: Role of virtual gap states and defects in metal–semiconductor contacts. Phys. Rev. Lett. **58**, 1260 (1987)
10. Robertson, J.: Band offsets of wide-band-gap oxides and implications for future electronic devices. J. Vac. Sci. Technol. B **18**, 1785 (2000)
11. Lampert, M.A., Mark, P.: Current Injection in Solids. Academic, New York (1970)

12. Murgatroyd, P.N.: Theory of space-charge-limited current enhanced by Frenkel effect. J. Phys. D Appl. Phys. **3**, 151 (1970)
13. Scott, J.C., Malliaras, G.G.: Charge injection and recombination at the metal-organic interface. Chem. Phys. Lett. **299**, 115 (1999)
14. Emtage, P.R., O'Dwyer, J.J.: Richardson–Schottky effect in insulators. Phys. Rev. Lett. **16**, 356 (1966)

Electrodics from First Principles

▶ Electrochemical Interfaces for Energy Storage and Conversion

Electrohydrodynamic Forming

▶ Electrospinning
▶ Nanostructures by Electrospinning

Electrokinetic Fluid Flow in Nanostructures

Francesca Carpino[1], Larry R. Gibson II[1], Dane A. Grismer[1] and Paul W. Bohn[1,2]
[1]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN, USA
[2]Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, IN, USA

Synonyms

Nanofluidics

Definition

Electrokinetic flow in nanostructures involves the study of the behavior, manipulation, and control of fluids that are confined to structures characterized by nanometer scale (typically $\leq$ 100 nm) dimensions. Fluids confined in these structures exhibit behaviors not observed on longer length scales, because characteristic physical scaling lengths, e.g., Debye length, in the fluid are commensurate with the dimensions of the nanostructure itself. A dominant factor in determining the electrokinetic (EK) flow behavior in physical nanostructures is the size of the electrical double layer (EDL) associated with the physical boundaries of the nanostructure and its relation to the channel dimensions.

Overview

Motivation. Although the engineering drivers for reducing the size of laboratory devices are clear (reduced materials costs, reduced power consumption, less waste, lower operating costs, etc.), the case for doing science at reduced dimensions needs to be made more carefully. Factors which justify working at the nanometer scale include: (1) new transport phenomena, (2) exploiting the enhanced surface-to-volume ratio, (3) using diffusion as a viable transport mechanism, and (4) integrating large molecules or molecular complexes with very small (1–10 nm) physical structures. In particular:

- The similarity of the Debye length, κ^{-1}, and the channel diameter, a, accesses a new electrokinetic flow regime ($\kappa a \sim 1$) that is not available at longer length scales, even in μm-scale capillaries. By changing κ^{-1} it is possible to move from a regime, $\kappa a >> 1$, where flow is dominated by electrophoresis to a regime, $\kappa a \sim 1$, where electroosmotic flow is the dominant transport mechanism. This is possible, because in the nanochannels there is a preponderance of counterions over co-ions; in fact it is easy to achieve conditions where every mobile counterion in the pore is of one polarity, i.e., there are no co-ions.
- The huge increase in surface-to-volume ratio dramatically increases the efficiency of physical partitioning and chemical reactions and reduces Joule heating effects. Comparing a 20 μm open tubular pore featuring a 10-nm thick coating to a 200-nm nanopore with the

same coating, the capacity factor, k', for molecular separations increases by a factor of 117.
- Due to their small size, adjacent nanostructures can be sufficiently close to one another to allow solution emerging from neighboring pores to rapidly mix through diffusion alone. Using the Stokes–Einstein infinite dilution diffusivity for a 1-nm particle in H_2O at 300 K and a structure supporting a pore density of $\sim 5 \times 10^8$ cm^{-2} yields a diffusion time of $\tau_D \sim 450$ μs to achieve full lateral mixing.

Early Studies. All investigations of nanoscale electrokinetic transport ultimately trace their origin to a seminal 1965 study of EK transport in nanometer scale cylindrical capillaries by Rice and Whitehead [1], in which the behavior of electroosmosis, streaming potential, current density distributions, and the electroviscous effect in cylindrical capillaries of nanometer dimensions were described within the Debye–Hückel approximation. Rice and Whitehead began with the Poisson–Boltzmann equation for a narrow cylindrical capillary at small, $\zeta < 25$ mV, zeta potential,

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{d\varphi}{dr}\right) = k^2\varphi \tag{1}$$

Where

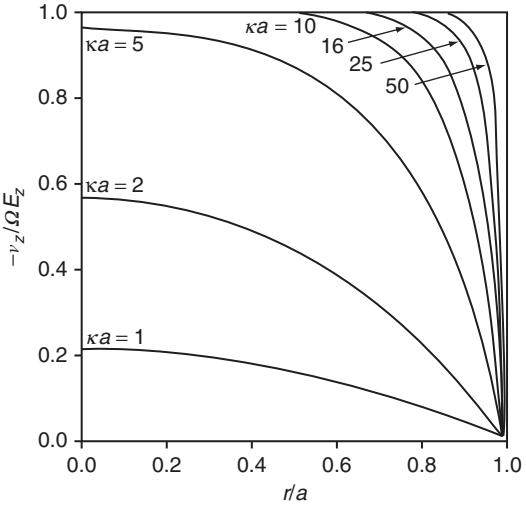
$$k = \sqrt{8\pi n e^2/\varepsilon kT} \tag{2}$$

is the inverse Debye length. The solution of (1) is

$$\varphi = \varphi_0 \frac{I_0(\kappa r)}{I_0(\kappa a)} \tag{3}$$

where I_0 is the modified Bessel function of the first kind, which allows the charge density to be recovered directly using the Poisson equation. For an infinite cylindrical tube, the equation of motion under a combination of electrical and pressure-driven flow can then be written as

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{dv_z}{dr}\right) = \frac{1}{\eta}\frac{dp}{dz} = \frac{F_z}{\eta} \tag{4}$$



Electrokinetic Fluid Flow in Nanostructures, Fig. 1 EOF velocity profiles in infinite cylindrical capillary (Adapted with permission from Ref. [1])

where the body force, F_z, is driven by the action of the applied field on the net charge density in the double layer, $\rho(r)$. The important result from this analysis is that two flow regimes are established. When $\kappa a >> 1$, the radial velocity profile, $v_z(r)$, reduces to the classical result, except very near the cylinder wall, i.e., the "plug flow" familiar from studies of capillary electrophoresis in μm-scale capillaries is recovered. When $\kappa a \sim 1$, $v_z(r)$ is proportional to $(a^2 - r^2)$, a behavior which is equivalent to the Poiseuille flow at nanometer dimensions, as shown in Fig. 1 and has been verified experimentally [2].

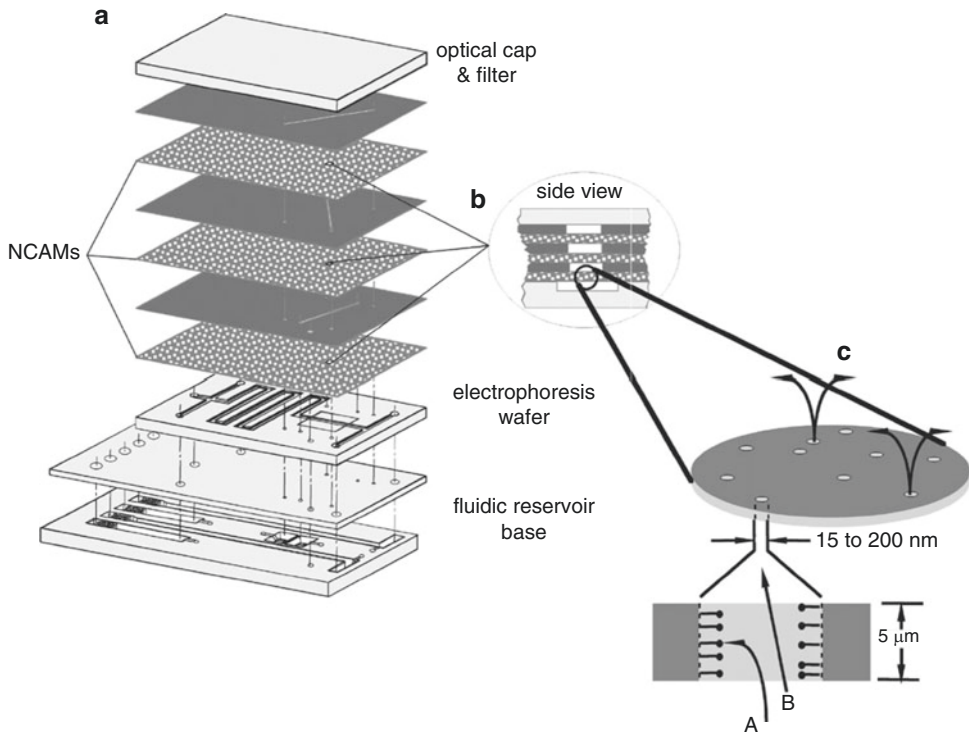
Fundamentals. When $\kappa a \sim 1$ there is significant interaction between the electrical double layers associated with opposing surfaces of the nanostructure. In general, transport across nanofluidic features at $\kappa a \sim 1$ depends on the sum of interactions among charge, electrophoretic mobility, molecular size, nanochannel cross section, ionic strength, and surface charge. A 100 nm × 1 μm nanochannel might exhibit a typical surface charge density ($\sigma \sim 2 \cdot 10^{-3}$ C m^{-2}), resulting in ~4,000 elementary charges on the interior wall of the nanochannel, requiring ≥1 mM counterion concentration in order to balance the surface charge. The ionic strength of the

solution can then be adjusted to control the relative populations of counter- and co-ions in the nanochannel. At low ionic strength, EDL overlap is significant, and counterion concentration and nanochannel conductance are determined purely by surface charge density. For common buffer concentrations (1–100 mM), κ^{-1} can range from 10 to 1 nm. Small molecule diffusion times across such nanochannels are of the order of 10^{-5}–10^{-4} s, which is advantageous if one wishes to implement heterogeneous reactions within nanochannels (*vide infra*). Adsorption (or desorption) of surface monolayers ($\Gamma = 10^{-10}$–10^{-11} mol/cm^2) within nanopores can cause the concentration to change ~1–5 mM, which has important implications for selective capture and subsequent release in purification and concentration applications.

Flow rate normalized to cross-sectional area is less in nanochannels than in microchannels and limits translocation velocities. In addition to applied field, molecules are affected by the electric field due to the electrical double layer (EDL). As a result, counterions with larger charge are more attracted to the surface of the nanochannel and their migration is impeded most. Co-ions are repelled from edges of the channel and are transported faster than counterions. Neutral species are not significantly affected by the EDL and exhibit intermediate velocities. This complex interplay between charge and molecular size can have interesting consequences for differential transport and molecular separations, especially of macromolecules. For example, molecules which occupy or sweep out a significant portion ($> a/10$) of the nanochannel can exhibit altered molecular transport characteristics. Smaller molecules, which approach the channel walls more closely, show slower translocation velocities than larger molecules, which present a larger cross section to the higher flow velocities near the center of the nanochannel. This phenomenon mediates Ogston sieving, which has been exploited to effect DNA separations in nanochannels [3]. In addition, macromolecules that are larger than the cross section of the nanochannel undergo entropically disfavored conformational changes upon entering the confined space represented by the nanochannel. Longer DNA strands surmount such entropic barriers more rapidly, because they have a larger contact area at the entrance and larger increase in entropy facilitates faster escape from the constriction, an observation which has been explored to fabricate nanoscale sieving media.

Applications

Electrokinetic Flow Switching. Clearly the special characteristics of nanoscale EK flow present rich possibilities for studying the fundamental physics of fluid transport at this scale. Beyond the interesting fluid physics, nanofluidic architectures present great opportunities for construction of integrated microfluidic architectures exhibiting digital fluidic transfer by exploiting the unique properties of nanocapillary array membranes (NCAMs) as the fluidic analogs of transistors in integrated electronic circuitry, (Fig. 2). NCAMs can be physically realized in a variety of forms, including spatially random distributions of nuclear track-etched nanopores in polycarbonate membranes or focused ion beam milled arrays in poly(methylmethacrylate). Fluid volume elements can be transferred from one microfluidic channel to another by manipulating the bias potential applied across an NCAM. Kuo et al. demonstrated that this fluid transfer is linear in driving potential up to a threshold value, V_{thresh}, beyond which quantitative mass transfer occurs, i.e., at $V_{appl} > V_{thresh}$ all of the analyte species which enter the cross section of the NCAM and the microfluidic channel are transferred quantitatively across the NCAM [4]. In addition, the pore size determines the mass selectivity of fluid transfer; surprisingly, hindered transport is observed for molecular sizes as little as $a/20$, where a is the pore diameter [5]. Integration of nanofluidic components also permits fluidic samples to be handled with discrimination based on molecular size and charge using valves, gates, and other fluidic components, similar to the way that integrated electronic components control information flow in VLSI circuits [6]. In addition, 3D integrated microfluidics constitute a powerful method to enhance fluid handling capabilities, because



Electrokinetic Fluid Flow in Nanostructures, Fig. 2 (**a**) Exploded schematic view of a multilevel vertically integrated μTAS system displaying a fluid input layer (*bottom*), a pre-separation (*electrophoresis*) layer, and three microfluidic layers separated by nanocapillary array membrane-switching layers. (**b**) Side view of the vertically stacked microfluidic channels separated by NCAMs. (**c**) Schematic diagram of a single NCAM with an exploded view of an individual nanopore with chemically derivatized interior. (**a**) and (**b**) denote two species that have been introduced at the mouth of the pore, with (**a**) being sequestered by the molecular recognition motifs lining the pore wall and (**b**) transferred across the pore (Adapted with permission from Ref. [16])

disparate fluidic manipulations such as preparation, concentration, tagging, separation, or affinity recognition may be accomplished in distinct physical locations.

Molecular Separations. Naturally occurring confined nanoscale geometries take advantage of several unique phenomena to produce ionic and molecular separations of remarkable power and utility. For example, transmembrane potassium channels routinely permit the passage of K^+, while excluding smaller, similarly charged Na^+ ions [7]. Such observations in nature have generated an interest in biomimetic nanostructures as tools to separate entities ranging from small molecules to proteins. Although synthetic replication of such complex ion channels has yet to be realized, significant progress in the field has been made.

The simplest form of a molecular separation utilizing nanostructures is filtration using a pressure gradient across a membrane. Provided both the surface charge density, σ, of the enclosing boundary and the ionic strength, I, of the mobile phase are negligible, size-based separation is feasible. Under these conditions, establishing either a pressure or concentration gradient will drive the extraction of appropriately sized analytes across the separation device. However, these options offer limited control, and the separations can be very slow.

Fortunately, EK transport is a powerful alternative. In fact, with the appropriate conditions to ensure $\kappa a \sim 1$ (*vide supra*), nanochannels, nanotubes, and nanopores may become completely perm-selective for counterions (relative to the zeta potential, ζ, of the structural

boundary). Garcia et al. successfully separated charged from uncharged dye molecules in nanochannels using EK transport [8]. In this experiment, the EDL was sufficiently large to localize the charged dye molecules (co-ions relative to ζ of the nanochannel walls) in the center of the nanochannel, while the uncharged molecules interacted with the channel walls. The induced electroosmotic velocity (maximal at the center of channel) eluted the charged dye molecules significantly faster than the uncharged species, essentially inducing radially varying capillary electrophoresis within the nanochannel.

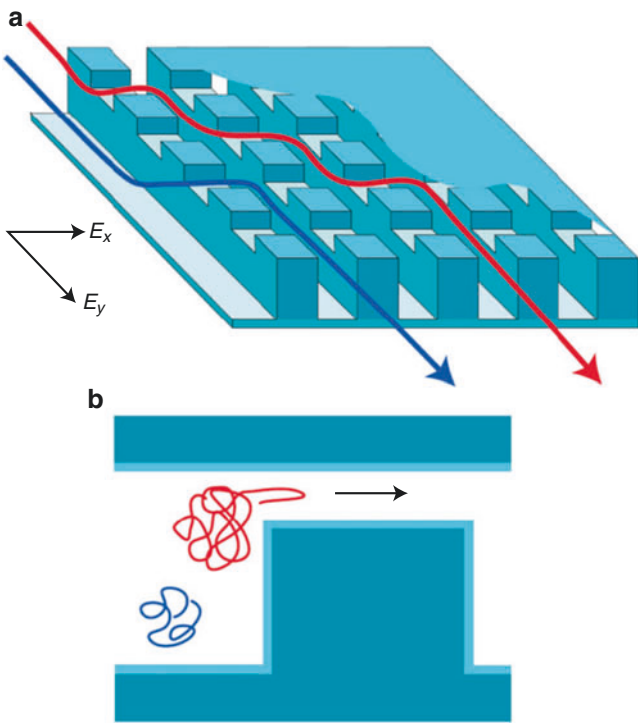
In addition to EK transport, molecular separations can be enhanced by modifying the nanostructure surface chemistry. For example, the perm-selectivity of nanopores (due to $\kappa a \sim 1$) in polycarbonate NCAMs can be enhanced by uniformly functionalizing pores in order to control both the pore size and surface charge density, thereby enabling size- and charged-based separations of small molecules as well as the resolution of mixtures of proteins of similar size but distinct isoelectric points (pI) [9, 10]. Similarly, Lokuge et al. realized actively tuned size-dependent separations of labeled dextrans, by modulating the temperature of NCAMs decorated with poly (N-isopropylacrylamide), thus causing the temperature-responsive polymer to cycle through its lower critical solution temperature, contracting and expanding to open and close the pores, respectively [11].

In order to extend the separation capabilities offered by nature, the screening quality of hydrophobic molecules has been exploited. Unfortunately, the conditions required for EK-enhanced separations typically do not favor hydrophobic molecules. In fact, high charge density regions within nanopores, for example, may act as barriers to entry. Fortunately, surfactants can be added at concentrations above the critical micelle concentration (CMC) to form micelles, which can be used to sequester hydrophobic molecules and transport them across the nanopores. This strategy was used to observe the micelle-enhanced transport of a sparingly soluble local anesthetic dissolved in nonionic surfactant across polymeric nanoporous membranes [12].

The fact that nanostructures can be fabricated to be commensurate with biomolecular scaling lengths also creates new modes for single-molecule manipulation and separation. For example, in bulk solution DNA molecules equilibrate to a spherical shape with a characteristic radius of gyration, R_0 [13]. To enter structures with any cross-sectional dimension less than $2R_0$, the molecule has to deform from its minimal energy state, a process that leads to entropic trapping. Figure 3 shows a periodic nanofilter array used to separate molecules with different mobilities, which are determined by the ease with which they deform from the preferred conformational state, a property that is inversely proportional to molecular size [14]. Additionally, confinement within nanochannels enables study of physical properties of DNA molecules in an uncoiled, linear form [15].

Whether applied independently or coupled, size-, charge-, and hydrophobicity-based molecular separations within nanopores, nanochannels, and nanotubes offer a range of options, especially when they are incorporated into larger microscale geometries and utilized as electrically switchable molecular gates [16]. Such devices have the potential to enhance or augment well-established separation processes, including microdialysis, capillary electrophoresis, and micellar electrokinetic chromatography positioned either upstream or downstream of the nanostructure.

Chemical Interactions and Reactions. Large surface-to-volume ratios and small cross-sectional areas make it possible for molecules traversing nanopores to have numerous interactions with the channel surface [15]. This phenomenon can be exploited to useful effect, for example, to selectively pattern the interior of nanochannels and nanopores in a process known as diffusion-limited patterning. The same phenomenon can be used to manipulate complex mass-limited samples, particularly in microfluidic/nanofluidic hybrid devices. Kim et al. exploited the molecular recognition abilities of antibodies to selectively capture, purify, and release target analytes from a mixture [6]. Gold-thiol chemistry was used to immobilize Fab' onto the electrolessly gold-plated interiors of NCAM nanopores. A mixture containing the target antigen, insulin, was transported through the

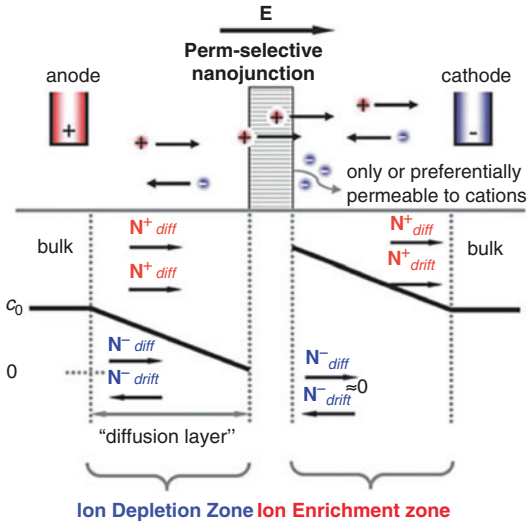


Electrokinetic Fluid Flow in Nanostructures, Fig. 3 (**a**) Orthogonal electric fields, E_x and E_y, serve as driving forces for the molecules and provide the energy needed to deform from their minimal energy states. (**b**) The longer molecule (*red*) has a greater rate of passage through confined regions due to a greater probability of deforming its shape than the shorter molecule (*blue*) (Reproduced with permission from Ref. [14])

NCAM by EK flow, after which, a releasing agent was applied to remove captured analytes. Mass spectra of the releasate showed strong specific retention of insulin – at 23-fold of the control intensity – but no presence of the other species, indicating the specific capture, concentration, and release of insulin. Thus, the high surface area within the NCAM and the increase in wall-collision frequency, Ω_W, render this immobilization strategy applicable to a broad range of analytes, especially those contained within mass-limited samples.

The effects of confinement in nanostructures are also apparent in altering the nature of fluidic transport. De Santo et al. observed sub-diffusive motion in nanochannels exhibiting high degrees of confinement [17]. Using fluorescence correlation spectroscopy (FCS), small, cationic rhodamine 6 G (Rh6G) molecules and charge-neutral labeled macromolecules, dextran and PEG, were observed under no-flow conditions in nanochannels of 30, 20, and 10 nm height. While Rh6G maintains its bulk diffusion coefficient in all cases, both dextran and PEG exhibit hindered diffusion – nearly an order of magnitude smaller at the highest degree of confinement. The decrease in diffusion for macromolecules is explained by anomalous diffusion, the degree of which increases with decreasing channel dimensions, until the theoretical limit for single-file diffusion is achieved. These findings emphasize the importance of surface interactions under confinement.

Concentration Polarization. Aside from direct effects on transport and reactivity inside nanofluidic channels, the presence of nanochannels can affect properties of liquid samples in adjacent microscale structures. Concentration polarization (CP), for example, occurs near ion-selective nanostructures. Surface-charged nanochannels/nanopores show selective permeability to ions due to EDL overlapping at $\kappa a \sim 1$ [18]. At low ionic strengths (<10 mM) the Debye length is comparable to the channel diameter, a, and the nanochannel is occupied entirely by counterions. Figure 4 illustrates CP at a cation perm-selective nanostructure [19]. When a dc bias is applied across the nanochannel, CP develops. On the anodic side, the anions move toward the anode migrating away from the anionic membrane interface where their concentration



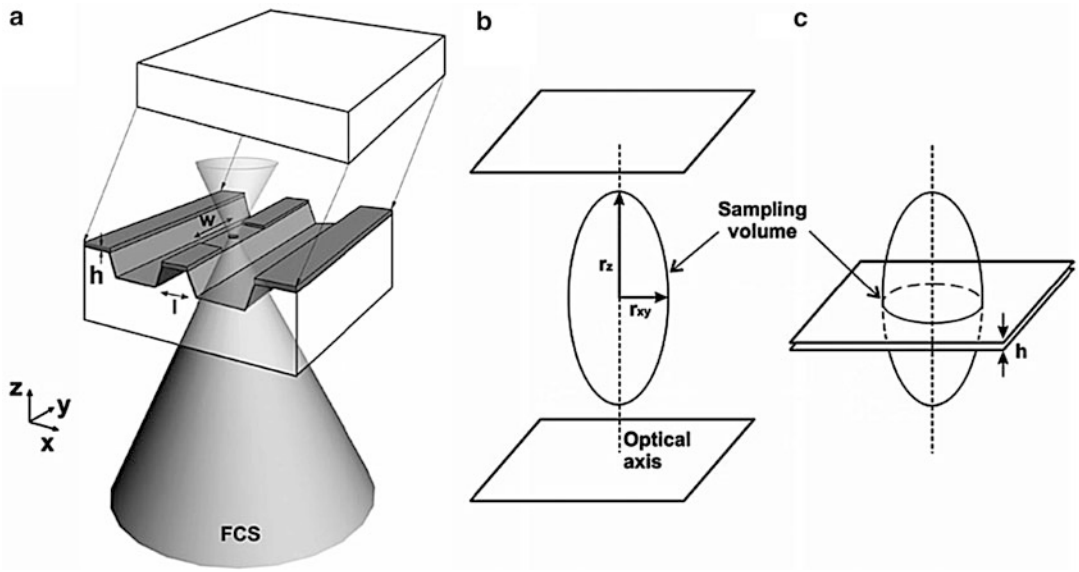
Electrokinetic Fluid Flow in Nanostructures, Fig. 4 Concentration polarization observed at the interfaces of a cation perm-selective nanostructure in the presence of current flow. Ion enrichment is observed at the cathodic side, while ions deplete at the anodic side (Reproduced with permission from Ref. [19])

decreases, while only cations can enter and pass through the perm-selective nanostructure. On the cathodic side, anions cannot enter the cation-selective nanostructure, so cations exit the membrane, causing an increase in local concentration. This phenomenon leads to concentration gradients established by the ion-enriched and ion-depleted zones thus created.

When an analyte bears the correct charge to be segregated in the ion enrichment zone, the local concentration can be increased, sometimes by orders of magnitude, thereby enhancing the sensitivity in chemical analysis applications. Han and coworkers observed strong nonlinear electrokinetic flow at the micro-nanostructure interface giving rise to strong CP [20]. The capability to preconcentrate analytes using CP prior to analysis has obvious applications in biosensing, where analytes may be present at concentrations far below available limits of detection (LOD). Wang et al. [21] utilized a nanofluidic preconcentrator and a bead-based immunoassay to enhance binding kinetics, sensitivity, and the dynamic range of detection in immuno-sensing of R-phycoerythrin, achieving a 500-fold sensitivity enhancement and lowering the LOD from 50 pM to <100 fm. Ramsey and coworkers [22] developed a comparable preconcentration system using a highly ion-conductive charge-selective polymer [poly-AMPS (2-acrylamido-2-methyl-1-propanesulfonic acid)]. Such sample preconcentration devices could be potentially integrated to a wide range of detectors, such as mass spectrometers, to improve the detection of low-level proteins.

Resistive Pulse Sensing. Because nanochannels are commensurate in size with macromolecular scaling lengths, they can be used as chemical and biochemical sensors. The translocation of large molecules through nanopores containing high ionic strength electrolyte can produce a transient ionic current simultaneous with the temporary pore blockage [23], which can be used to count and size macromolecules, using the same principles as in commercial Coulter counters. The resulting current transients are termed resistive pulses. The majority of effort has targeted the use of biological ion channels to support DNA or RNA translocation, with α-Hemolysin being the most popular choice due to its well-studied structure, stability, and ability to introduce specific structural modifications. α-Hemolysin is a 293-amino acid protein secreted by *Staphylococcus aureus*. Since the passage of each ssDNA or RNA molecule induces a current transient with a temporal duration proportional to its length, careful characterization of resistive pulse widths can be used to acquire information on molecular size. In favorable cases, these measurements can yield information capable of discriminating oligonucleotide composition [24].

Adding either chemical or geometric asymmetry to nanopores leads to structures with the ability to discriminate among chemical species. Conical nanopores can be produced from track-etched polymer membranes, which are etched in base, while neutralizing the incoming basic solution from the distal side. Conical nanopores display distinct advantages for resistive pulse sensing, including higher sensitivity due to ionic current being focused at the conical tip, larger ion currents at equivalent diameters, and being more resistant to fouling [25]. For example, protein biosensors have been fabricated from surface biofunctionalized nanotubes by electrolessly plating a layer of gold



Electrokinetic Fluid Flow in Nanostructures, Fig. 5 FCS used for observation of sample inside a nanochannel: (**a**) schematic of hybrid microfluidic/nanofluidic system and FCS probe (not to scale); (**b**) FCS ellipsoidal detection volume with height $2r_z$ and width of $2r_{xy}$ in the case of molecules diffusing freely in 3D; and (**c**) FCS detection volume of molecules confined between two walls diffusing in 2D (when $h << 2r_z$) (Reproduced with permission from Ref. [28])

onto the interior surfaces of conical nanopores, followed by self-assembly of thiolated biotin derivatives [26]. Streptavidin was subsequently detected, and blockage time was related to streptavidin concentration. This strategy highlights one of the strengths of metal-coated nanostructures – by exploiting self-assembly chemistry a wide range of sensor chemistries can be realized within the same physical structure.

Single-Molecule Manipulations. When coupled with traditional optically based single-molecule detection methods, nanofluidic features enable detection to be performed at more biologically relevant concentrations. The diffraction-limited spot size in a confocal fluorescence correlation spectroscopy (FCS) experiment is on the order of 0.2 fL, requiring nanomolar or lower concentrations for single-molecule analysis [27]. As seen in Fig. 5, nanoscale channels use only a fraction of this volume, decreasing the detection volume by up to more than an order of magnitude, while at the same time allowing a corresponding increase in the fluorophore concentration [28].

At the cost of flow-through experimental capabilities, zero mode waveguides (ZMWs) enable a further reduction in excitation volume beyond that of diffraction-limited techniques [15]. ZMWs are created by making a nanopore or nanopit in an opaque film such as Al or Au. The physical volume of these features confines the sample and excitation volumes to be on the order of 100 zL, (1 zL $= 10^{-21}$ L), more than three orders of magnitude smaller than the lower limit of the diffraction-limited focal volume in FCS experiments. Correspondingly, single-molecule observations can take place at physiologically relevant concentrations, 1 μM $< C <$ 1 mM, even for proteins with μM K_D values. The signal-to-noise is significantly increased, because the electromagnetic field is effectively confined within the ZMW, which enhances the signal relative to the reduced background by eliminating light propagation into the far field.

Conclusions

Realizing EK flow in nanostructures opens up a range of new phenomena that support important new applications. Especially important are those

applications that enable the space-time manipulations of molecular distributions in fluids, such as molecular separations, concentration polarization, and EK flow switching. In addition, new approaches to chemical sensing that exploit novel nanofluidic properties are realized through the scaling of mass transport for specific chemical interactions that in turn supports resistive pulse sensing and single-molecule manipulations.

Acknowledgments Work described in this entry that was carried out in the authors' laboratories was supported by the National Science Foundation, the Department of Energy and the Army Corps of Engineers.

Cross-References

► Applications of Nanofluidics
► Micro/Nano Flow Characterization Techniques

References

1. Rice, C.L., Whitehead, R.: Electrokinetic flow in a narrow cylindrical capillary. J. Phys. Chem. **69**, 4017–4024 (1965)
2. Ramsey, J.M., Alarie, J.P., Jacobson, S.C.: Molecular transport through nanometer confined channels. In Sixth International Conference on Miniaturized Chemical and Biochemical Analysis Systems, p. 314. Nara (2002)
3. Stein, D., van der Heyden, F.H.J., Koopmans, W.J.A., Dekker, C.: Pressure-driven transport of confined DNA polymers in fluidic channels. Proc. Natl. Acad. Sci. U. S. A. **103**, 15853–15858 (2006)
4. Kuo, T.C., Cannon, D.M., Chen, Y.N., Tulock, J.J., Shannon, M.A., Sweedler, J.V., Bohn, P.W.: Gateable nanofluidic interconnects for multilayered microfluidic separation systems. Anal. Chem. **75**, 1861–1867 (2003)
5. Zhang, Y., Timperman, A.T.: Integration of nanocapillary arrays into microfluidic devices for use as analyte concentrators. Analyst **128**, 537–542 (2003)
6. Kim, B.-Y., Swearingen, C.B., Ho, J.A., Romanova, E.V., Bohn, P.W., Sweedler, J.V.: Direct immobilization of Fab' in nanocapillaries for manipulating mass limited samples. J. Am. Chem. Soc. **129**, 7620–7626 (2007)
7. Doyle, D.A., Cabral, J.M., Pfuetzner, R.A., Kuo, A.L., Gulbis, J.M., Cohen, S.L., Chait, B.T., MacKinnon, R.: The structure of the potassium channel: molecular basis of K + conduction and selectivity. Science **280**, 69–77 (1998)
8. Garcia, A.L., Ista, L.K., Petsev, D.N., O'Brien, M.J., Bisong, P., Mammoli, A.A., Brueck, S.R.J., Lopez, G. P.: Electrokinetic molecular separation in nanoscale fluidic channels. Lab Chip **5**, 1271–1276 (2005)
9. Savariar, E.N., Krishnamoorthy, K., Thayumanavan, S.: Molecular discrimination inside polymer nanotubules. Nat. Nanotechnol. **3**, 112–117 (2008)
10. Osmanbeyoglu, H.U., Hur, T.B., Kim, H.K.: Thin alumina nanoporous membranes for similar size biomolecule separation. J. Membr. Sci. **343**, 1–6 (2009)
11. Lokuge, I., Wang, X., Bohn, P.W.: Temperature controlled flow switching in nanocapillary array membranes mediated by poly(*N*-isopropylacrylamide) polymer brushes grafted by atom transfer radical polymerization. Langmuir **23**, 305–311 (2007)
12. Bhown, A.S., Stroeve, P.: Micelle-mediated transport of a sparingly soluble drug through nanoporous membranes. Ind. Eng. Chem. Res. **46**, 6118–6125 (2007)
13. Schoch, R.B., Han, J.Y., Renaud, P.: Transport phenomena in nanofluidics. Rev. Mod. Phys. **80**, 839–883 (2008)
14. Austin, R.: Nanofluidics – a fork in the nano-road. Nat. Nanotechnol. **2**, 79–80 (2007)
15. Piruska, A., Gong, M., Sweedler, J.V., Bohn, P.W.: Nanofluidics in chemical analysis. Chem. Soc. Rev. **39**, 1060–1072 (2010)
16. Gatimu, E.N., King, T.L., Sweedler, J.V., Bohna, P.W.: Three-dimensional integrated microfluidic architectures enabled through electrically switchable nanocapillary array membranes. Biomicrofluidics **021502**, 1–11 (2007)
17. De Santo, I., Causa, F., Netti, P.A.: Subdiffusive molecular motion in nanochannels observed by fluorescence correlation spectroscopy. Anal. Chem. **82**, 997–1005 (2010)
18. Holtzel, A., Tallarek, U.: Ionic conductance of nanopores in microscale analysis systems: where microfluidics meets nanofluidics. J. Sep. Sci. **30**, 1398–1419 (2007)
19. Kim, S.J., Song, Y.-A., Han, J.: Nanofluidic concentration devices for biomolecules utilizing ion concentration polarization: theory, fabrication, and applications. Chem. Soc. Rev. **39**, 912–922 (2010)
20. Kim, S.J., Wang, Y.-C., Lee, J.H., Jang, H., Han, J.: Concentration polarization and nonlinear electrokinetic flow near a nanofluidic channel. Phys. Rev. Lett. **99**, 044501 (2007)
21. Wang, Y.-C., Han, J.: Pre-binding dynamic range and sensitivity enhacement for immuno-sensors using nanofluidic preconcentrator. Lab Chip **8**, 392–394 (2008)
22. Chun, H., Chung, T.D., Ramsey, J.M.: High yield sample preconcentration using a highly ion-conductive charge-selective polymer. Anal. Chem. **82**, 6287–6292 (2010)
23. Heng, J.B., Ho, C., Kim, T., Timp, R., Aksimentiev, A., Grinkova, Y.V., Sligar, S., Schulten, K., Timp, G.: Sizing DNA using a nanometer-diameter pore. Biophys. J. **87**, 2905–2911 (2004)

24. Rhee, M., Burns, M.A.: Nanopore sequencing technology: research trends and applications. Trends Biotechnol. **24**, 580–586 (2006)
25. Vlassiouk, I., Kozel, T.R., Siwy, Z.S.: Biosensing with nanofluidic diodes. J. Am. Chem. Soc. **131**, 8211–8220 (2009)
26. Siwy, Z., Trofin, L., Kohli, P., Baker, L.A., Trautmann, C., Martin, C.R.: Protein biosensors based on biofunctionalized conical gold nanotubes. J. Am. Chem. Soc. **127**, 5000–5001 (2005)
27. Levene, M.J., Korlach, J., Turner, S.W., Foquet, M., Craighead, H.G., Webb, W.W.: Zero-mode waveguides for single-molecule analysis at high concentrations. Science **299**, 682–686 (2003)
28. Durand, N.F.Y., Dellagiacoma, C., Goetschmann, R., Bertsch, A., Marki, I., Lasser, T., Renaud, P.: Direct observation of transitions between surface-dominated and bulk diffusion regimes in nanochannels. Anal. Chem. **81**, 5407–5412 (2009)

Electrokinetic Manipulation

▶ Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes

Electrokinetics

▶ Rapid Electrokinetic Patterning and Its Applications

Electron Beam Evaporation

▶ Physical Vapor Deposition

Electron Beam Lithography (EBL)

Nezih Pala and Mustafa Karabiyik
Department of Electrical and Computer Engineering, Florida International University, Miami, FL, USA

Synonyms

E-beam lithography (EBL)

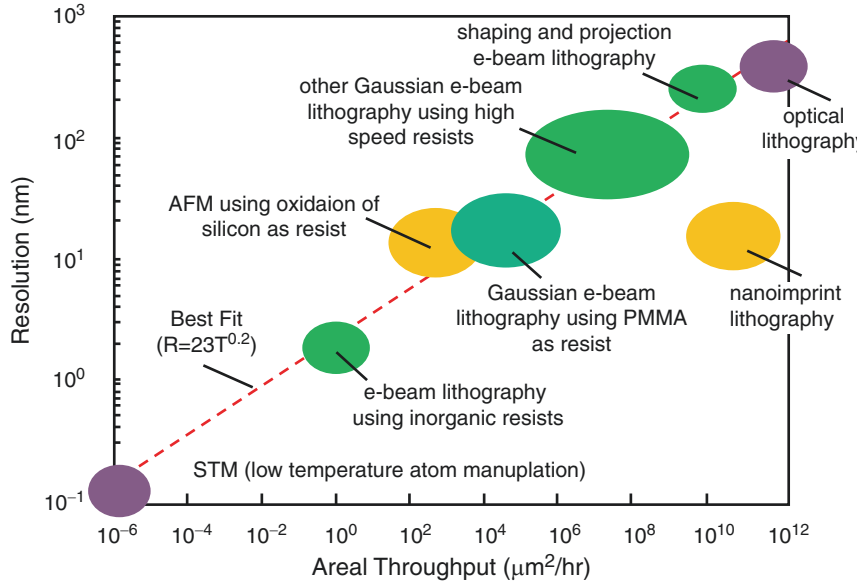
Definition

Electron beam lithography (often abbreviated as e-beam lithography or EBL) is the process of transferring a pattern onto the surface of a substrate by first scanning a thin layer of organic film (called resist) on the surface by a tightly focused and precisely controlled electron beam (exposure) and then selectively removing the exposed or nonexposed regions of the resist in a solvent (developing). The process allows patterning of very small features, often with the dimensions of submicrometer down to a few nanometers, either covering the selected areas of the surface by the resist or exposing otherwise resist-covered areas. The exposed areas could be further processed for etching or thin-film deposition while the covered parts are protected during these processes. The advantage of e-beam lithography stems from the shorter wavelength of accelerated electrons compared to the wavelength of ultraviolet (UV) light used in photolithography, which allows defining much smaller diffraction-limited features. On the other hand, direct writing of patterns by scanning electron beam is a slow process and results in low throughput. Therefore, EBL is used for preparing photomasks for photolithography or for direct writing of small-area, low-volume patterns for research purposes.

Principles of Electron Beam Lithography

There are two driving forces for the development of electron beam lithography and other patterning technologies (e.g., X-ray lithography) as potential alternatives to the conventional UV-based photolithography: higher resolution (smaller feature size) and cost.

The maximum resolution of an optical system is limited with the wavelength of the beam used. Conventional photolithography systems, which rely on UV light, can therefore achieve a minimum “diffraction-limited” feature size of several hundreds of nanometers. Using resolution enhancement techniques such as phase shifting and immersion lithography, it is possible to



Electron Beam Lithography (EBL), Fig. 1 Resolution of different lithographic techniques as a function of throughput

improve the resolution to some extent. However, the fundamental laws of optics will continue to be a limitation. Moreover, resolution enhancement techniques for photolithography caused both cost and cycle time of masks to steadily increase with the rising level of complexity for every technology node. The cost of an optical mask set for 90-nm technology has surpassed the $1 million mark and exceeds $2 million for the 65-nm node. As a result, prototyping is becoming extremely expensive and time consuming. Mask amortization for low- and medium-volume chip production is becoming all but unaffordable. For this reason, various maskless lithography approaches are being pursued in attempts to solve the cost and cycle time dilemmas. Especially in the framework of nanotechnology, variations of EBL emerge as valid alternatives for lithographic patterning (see Fig. 1).

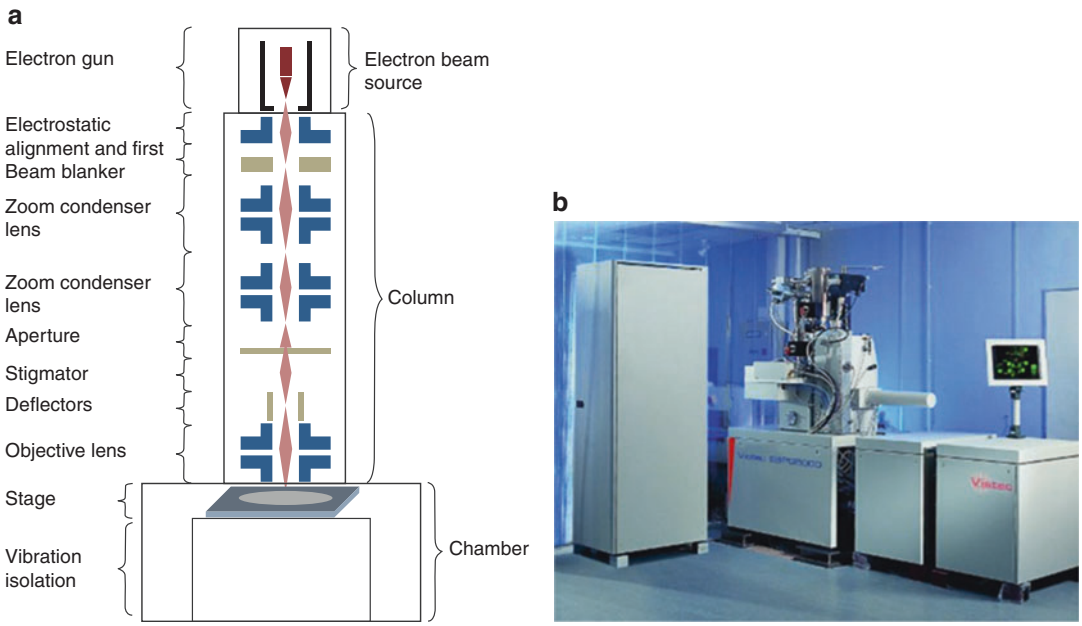
Electron Optics

The power of electron beam lithography to define nanometer-scale patterns comes from the short wavelength of electrons. In 1924, Louis de Broglie proposed that moving particles could show wave-like properties and therefore can be considered and modeled as waves in an appropriate framework. His theory was fully confirmed with the observation of electron diffraction in crystals in 1929. In his work, de Broglie developed a simple formula for the wavelength of a moving particle:

$$\lambda = \frac{h}{p} = \frac{h}{m \cdot v} \tag{1}$$

where h is Planck's constant (4.135×10^{-15} eV.s), p is the momentum, m is the mass of the electron (9.11×10^{-31} kg), and v is its speed. This neglects special relativity effects, because it is assumed that the electrons are traveling sufficiently below the speed of light. If the electron were traveling at the speed of light, c (2.9979×10^{8} m/s) which is its maximum theoretical speed, it would have a wavelength of 2.43 pm. An electron's velocity can be determined from its kinetic energy, by the formula

$$E_{kin} = \frac{1}{2} m v^2 \tag{2}$$



Electron Beam Lithography (EBL), Fig. 2 (**a**) Schematic of a typical electron beam lithography system. (**b**) A typical EBL system (Courtesy of Vistec Electron Beam Lithography Group)

Thus, the de Broglie wavelength for electrons with kinetic energy E_{kin} is:

$$\lambda = \frac{h}{\sqrt{2mE_{kin}}} = \frac{1.23}{\sqrt{E_{kin}}}[nm] \qquad (3)$$

where E_{kin} is expressed in eV. So if an electron were accelerated to an energy of 10 keV, then it would have $\lambda = 0.12$ A. This clearly shows that if lithography could be done with electrons, it would have a huge advantage over current optical lithography systems, which are limited by their wavelength. This is just a theoretical limit, however. EBL systems use projection and with most forms of EBL, no mask is used at all. So except in high-resolution systems (with very low throughput) where electron diffraction can come into play, the resolution of the system is limited by the beam column rather than the wavelength of the electrons. The resist also places limitations on the resolution of the beam, which can be greater than those effects, and in most cases, scattering in the resist places the limiting factor for the resolution.

Equipment

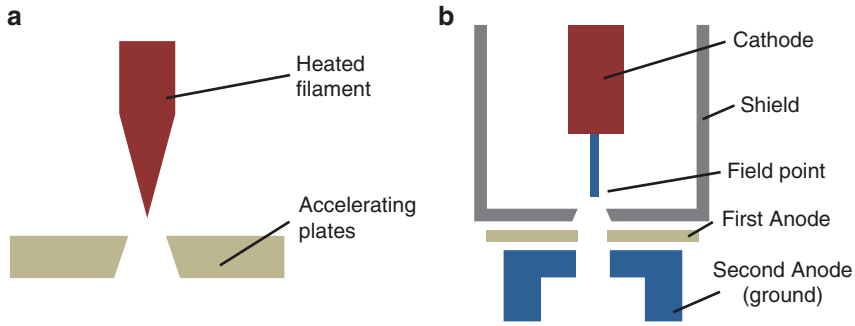
Although modern electron beam lithography systems are quite sophisticated to achieve very high resolution, precision, and repeatability, they all have the same fundamental parts in common. Figure 2 shows the schematic of a typical EBL system with major equipment parts.

Electron Sources

Electron source is the most essential part of any electron beam system. This is where the electrons are generated, which will eventually reach the resist on the sample and expose it. An ideal electron source should have a high intensity (brightness), high uniformity, small spot size, good stability, and long life. Brightness is measured in units of amperes per unit volume per steradian. Higher brightness is desired to have increased reaction rate with the resist and thereby shorter exposure time and higher throughput. High-resolution patterning, on the other hand, requires lower beam intensities. Virtual size of the source is also important since it determines the demagnification needed to be applied to the column. The smaller virtual size

Electron Beam Lithography (EBL), Table 1 Characteristics of different electron sources

Source type	Material	Brightness (A/cm^2rad)	Source size	Energy dispersion (eV)	Vacuum level (Torr)	Emitter temp. (K)
Tungsten Thermionic	W	~10^5	25 μm	2–3	10^{-6}	~3,000
LaB6 Thermionic	LaB_6	~10^6	10 μm	2–3	10^{-8}	~2,000–3,000
Thermal field aided emission	Zr/O/W	~10^8	20 nm	0.9	10^{-9}	~1,800
Cold field emission	W	~10^9	5 nm	0.22	10^{-10}	Ambience



Electron Beam Lithography (EBL), Fig. 3 Simplified cross-sectional schematics of (**a**) thermionic emission (**b**) field emission electron sources

provides smaller beam spot size on the wafer with minimum number of lenses. This makes the column less complicated, and allows a higher resolution to be achieved. Another very important feature of a good electron source is the low energy dispersion of the emitted electrons. An electron beam with a wide energy distribution is comparable to white light, while an energy beam with a narrow spectrum is comparable to a laser source. A narrow electron energy spread is desired as it reduces chromatic aberrations. Characteristics of different electron sources are summarized in Table 1 [1].

In an electron source, electrons are removed from cathode of the gun either by heating the cathode or by applying a large electric field. The first process is called thermionic emission (Fig. 3a) and the latter is called field emission (Fig. 3b). A combination of two, which is called thermal field–aided emission, is also sometimes employed.

For thermionic emission sources, tungsten or tungsten thoriated filaments were commonly used as emitters. In such sources, the filament is heated by passing current through it and electrons are emitted thermionically from a sharp tip. Tungsten filaments allow operation at pressures as high as 0.1 mtorr, but their low current density (~0.5 A/cm^2) results in very low brightness (~2 × 10^4 A/cm^2 sr). Thoriated tungsten cathodes have somewhat lower brightness at the same filament current and require higher vacuum (0.01 mtorr), but their maximum current density can be as large as 3 A/cm^2. Another disadvantage of tungsten sources is the high-energy spread due to the high operating temperature (2,700 K). Lanthanum hexaboride (LaB_6) is a newer thermionic source which can achieve very high current density over 20 A/cm^2 and good brightness of nearly 2 × 10^4 A/cm^2 sr at a lower temperature of 1,800 K. It has a slightly smaller source size than tungsten; however, it requires a better vacuum of 10^{-8} Torr, which can be easily achieved in present systems and must be well protected against sudden vacuum loss.

A major concern with all thermionic sources is the finite source size or crossover diameter. Because

a large volume of wire is heated, the thermionic sources produce broad beams. The energy distribution from such a source is also quite broad. This leads to focusing problems similar to those found in large optical partially incoherent sources. For typical LaB_6 sources, the crossover diameter is about 10 μm. Therefore, a demagnification of 100× is required to achieve 0.1-μm spot size. Even though the source has a large current density, the inherent source brightness is greatly reduced in order to achieve deep submicron resolution.

Field emission electron sources are based on application of an electric field high enough to extract the electrons through the surface potential barrier of the emitter (see Fig. 3b). In field emitters, cathode tip should be sharp, made of a material with reduced work function and be placed close to the extraction electrode to operate at a low voltage. Tungsten tips provide the extremely high fields necessary for electron extraction. An example of an electron source and the electrodes is shown in Fig. 3b. Two anodes are normally used. The first anode is the extraction electrode, and is used to extract the electrons from the cathode tip, and the second is used to accelerate the electrons to their full potential. The shield electrode is used to prevent thermally generated electrons from entering the beam. Thus, electrons are only emitted from the tungsten tip. Cold field emission sources have the disadvantage of high short-term noise, which is due to atoms adsorbing onto the tip. This causes fluctuations in the electron current in the short term as well as in the long term. Intensity fluctuations in the beam current caused by tip absorption can be reduced by ultrahigh vacuum environment.

Thermionic field emission sources are the best available electron sources. Thermal field emission sources combine the best of both field and thermionic emission sources. A tungsten needle (with approximately 1-μm tip) is used at only 1,800 K, thus increasing the lifetime of the source. The barrier to field emission of electrons is lowered because the needle is heated. Its brightness, source size, and energy spread are all better than the standard thermionic sources. Recently Zr/W/O tips have become available with crossover diameters as small as 200 Å for thermionic field emission. Such emitters can provide very high current densities up to 1,000 A/cm^2 and beam diameter as small as 100 Å without compromising the brightness. They, however, require a constant vacuum of at least 1×10^{-8} Torr for stable operation.

Once an electron beam is generated, it is focused and steered in the column using lenses, apertures, and deflectors. In the following sections, main components of a typical EBL column will be described.

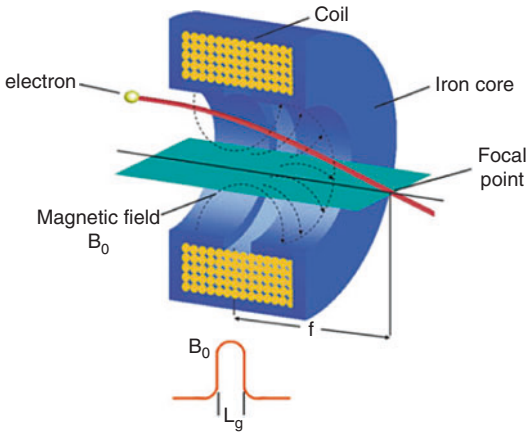
E

Electron Lenses

The physical principles of electromagnetic lenses can be described by using the basic laws of electromagnetism. Electron motion can be modeled as EM waves, which implies that it can be focused and manipulated analogous to the classical optics systems (geometric optics). At the same time, electrons maintain the characteristic properties of classical charged particles. Therefore, electrons are manipulated in two basic ways: through an external electric field or a magnetic field.

From the second law of Newton, $\mathbf{F} = m \cdot (d^2x/dt^2)$, the trajectory and speed of electrons can be controlled by external forces, more precisely, by electromagnetic forces. According to the Coulomb's law, an electric field will exert a force on a charged particle. An electron will experience a force of $\mathbf{F}_C = -e\mathbf{E}$, where e is the elementary charge of an electron, and $\mathbf{E}$ is the electric field. According to Lorentz's Law, an electron will experience a force when traveling perpendicular to a magnetic field. This force will be perpendicular to the magnetic field and the electron's velocity, and is given by $\mathbf{F}_L = -e.\boldsymbol{v} \times \mathbf{B}$. The contributions could be summed together; however, normally in EBL systems, both magnetic and electric fields are not used simultaneously.

Magnetic lenses are used primarily because the aberrations produced by magnetic lenses are less severe than those produced by electrostatic lenses. A cross section of a magnetic lens is shown in Fig. 4. The radial magnetic fields cause the electrons to experience rotation about the optical axis upon entering the lens. The electrons now have a



Electron Beam Lithography (EBL), Fig. 4 Cross-sectional schematic of a magnetic electron lens [2]

velocity component in the tangential direction about the optical axis. This component now interacts with the axial magnetic field, imparting a force on the electrons toward the optical axis, bringing them close to the axis.

Neglecting the actual aberrations that exist in the beam trajectories, electrons are focused to a certain distance, f, from the center of the lens, determined by the electron mass m, unit charge e, magnetic field B_0, the gap L_g, and V_0 the voltage representing the electron's velocity:

$$f \cong \frac{8mV_0}{eL_gB_0^2} \tag{4}$$

Electrostatic lens operation, to force electrons to converge in some point of the optical axis, is similar to magnetic lens one. The realization is accomplished by three plates provided with a central aperture. Central plate has a variable potential and the first and third plates are connected to the ground. In general, electrostatic lenses are used as condenser lenses of the electron source since the distortions inherent to these lenses are less critical here. The deflection unit is in charge of deviating the beam through the sample surface, within what is called the scan field. Ideally, the minimum degradation of beam is desired, i.e., precise deflection, constant beam size, and no hysteresis.

Beam Deflectors

Beam deflectors are used to scan the electron beam over the target, across an area called the scan field. To have a predictable deflection for precise patterning of ultrasmall features, this process must be very linear, with minimal degradation to the electron beam spot size and with no hysteresis. Like in electron lenses, deflection can be done by an electric field or by a magnetic field. Electrostatic beam deflection results in more distortions in the beam, so magnetic deflection is normally used. Beam deflectors are one of the most important components; however, any nonlinearities or problems with beam deflectors can easily be fixed in the software, which drives them.

Apertures

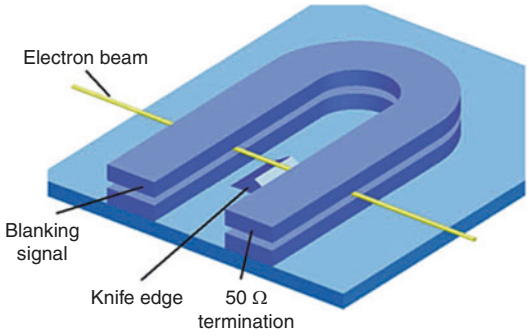
There are a few different types of apertures, blanking apertures, and beam-limiting apertures. Blanking apertures deflect the beam by deflecting the beam away from the aperture hole. Beam-limiting apertures set the beam convergence angle, α, (the angle between the beam trajectory and the normal line to the target) which controls the effect of lens aberrations and resolution but also limits the beam current.

Beam Blankers

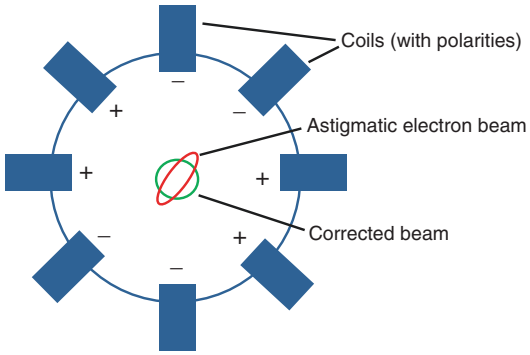
Beam blankers are used to turn off or "blank" the beam. This is necessary when doing vector scanning and the beam needs to be moved from one part of the wafer to another. The beam must be blanked in a time, which is very small compared to the time it takes to illuminate one pixel on the array, and the beam cannot move along the substrate while it is being blanked. A common blanker is shown in Fig. 5 and it works by applying a voltage to the upper plate, which deflects the electron beam away from the center of the column.

Stigmators

The imperfections in fabrication and assembling of the column cause astigmatism. Astigmatism results in different focusing conditions at different positions on the sample surface. This means that the ideal circular section of the beam becomes elliptic and, consequently distorts the image.



Electron Beam Lithography (EBL), Fig. 5 Schematics of a typical beam blanker



Electron Beam Lithography (EBL), Fig. 6 Schematic of a stigmator with eight poles. Alternate coils are connected in series but have different polarities. Astigmation of the electron beam can be corrected by changing the size of the current through the two sets of coils

Concerning to lithography, beam shape does not correspond to the model used for calculating the exposure dose; therefore, the pattern to be transferred is also distorted from the original design. The stigmator system is responsible for correcting beam shape to be circular again. It consists of four or eight poles that surround the optical axis. Adjustment is performed by the balance of electrical signal of the poles (Fig. 6).

Stage

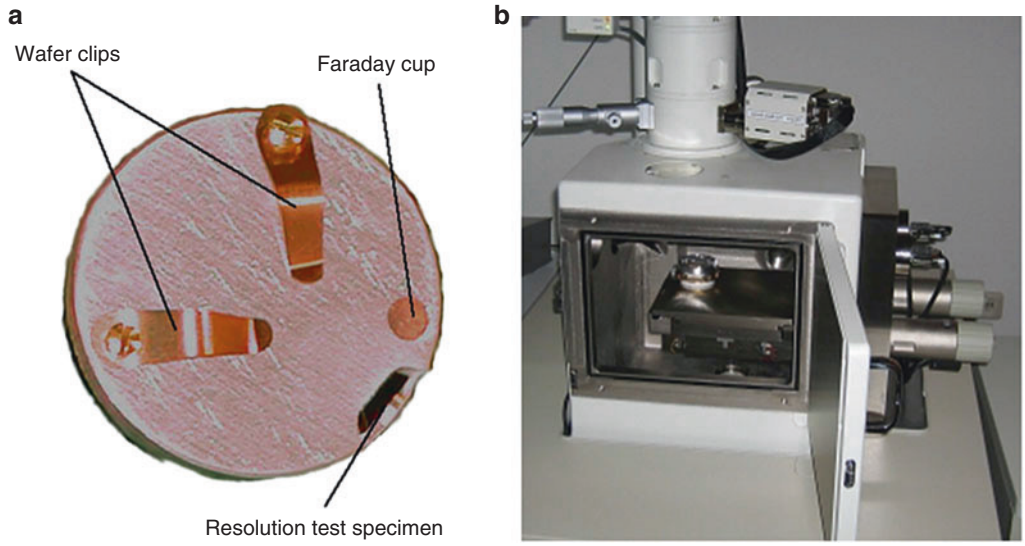
EBL patterning is performed in a high vacuum chamber. The sample is fixed on a sample holder by screws, spring-loaded mechanisms, double-sided tapes, or silver paint. Wafer holders usually contain a set of test standards for calibration focusing and position. Calibration subroutines can be performed immediately before the actual writing or even during the writing in between different fields as often as desired.

The dimensions and mobility of the stage determine the size and accessibility of the sample that can be patterned. The equipment used can lodge up to 300 mm wafers and movement is possible for XYZ, rotation and tilt (Fig. 7a). The capabilities of the system benefit from precise and motorized controlled displacements. The stage moved by piezoelectric motors and the position is controlled by laser interferometry system to achieve nanometer-scale precision. The stage is also placed on vibration isolation system to reduce the mechanical vibrations, which might affect the spatial precision. At the same time, computer monitors, transformers, and vacuum pumps are kept separate or controlled and shielding is used to avoid interferences. In addition, the chamber may also be equipped with a charge-coupled device (CCD) camera to visualize inside the chamber and to assess the control of sample positioning.

Limitations of the Equipment

Different from conventional optical lenses, electromagnetic lenses are only converging. In reference to aberrations, their quality is so poor that field size and convergence angle (numerical aperture) are limited. From electron source specifications, the beam diameter as a function of virtual source and column reduction is determined. The distortions caused by the column arise as spherical aberrations (astigmatism) and they are originated in both the lenses and the deflectors. Chromatic aberrations appear when electrons present a certain energetic spectra. This may be caused in the beam source itself, but also can be increased by the so-called Boersch and Loeffler effects, where energetic distribution increases as a result of electron collisions. Both phenomena contribute to decrease the precision of lenses and deflectors.

In general, the theoretical effective beam can be expressed as the quadratic sum of each contribution and optimal point is a compromise between all involved factors. For high resolution, the use of high magnification and beam energy is combined with low energetic dispersion and short



Electron Beam Lithography (EBL), Fig. 7 (**a**) A simple EBL sample holder with two wafer clips to hold the sample, a Faraday cup to measure the beam current, and gold on carbon resolution sample for resolution test. (**b**) Wafer chamber of an EBL system with the door open

focal distance. In consequence, write field is smaller and beam current should be reduced, which means that exposure process is slower, throughput is limited, and flexibility is constrained. The final spot diameter on the surface of the wafer is given by

$$d^2 = d_d^2 + d_o^2 + d_S^2 + d_C^2 \qquad (5)$$

where d_s is the spherical aberration, d_C is the chromatic aberration due to the nonzero energy distribution, and d_o is the perfect lens diameter, which is limited by the finite source size and space charge and d_d is the diffraction limit.

Another limitation is called stitching and stems from the limited deflection of the beam. Typically, for very small beam deflections electrostatic deflection "lenses" are used, larger beam deflections require electromagnetic scanning. Because of the inaccuracy and because of the finite number of steps in the exposure grid, the writing field is on the order of 100 μm–1 mm, larger patterns require stage movements. An accurate stage shift is critical for stitching (tiling writing fields exactly against each other) and pattern overlay (aligning a pattern to a previously made one). Large patterns spanning several write fields may suffer from *stitching errors* due to the finite inaccuracies in stage movement.

Resists for Electron Beam Lithography

Interactions between electromagnetic radiation (either UV or electron beam) and polymers, which are used as resists (either for optical or e-beam lithography) result in one of the following two chemical reactions: (1) *cross-linking* (2) *chain scission* (also known as fragmentation). The resists are usually long chains of carbon-based polymers. By interacting with incoming radiation, atoms in adjacent chains may be displaced and the carbon atoms will bond directly. This process is called cross-linking. Highly cross-linked molecules dissolve more slowly in an organic solvent (developer). A material in which cross-linking is the dominant reaction upon exposure is a *negative resist*. Radiation can also disrupt the polymer chains, breaking them up to smaller pieces. This reduces the molecular weight of the resist, which makes it more soluble in developer solution. A material in which chain

Electron Beam Lithography (EBL), Table 2 Commonly used EBL resists

Resist	Tone	Sensitivity @20 keV (μC/cm^2)	Resolution (nm)	Contrast	Etch resistance
PMMA	+	100	10	2.0	Poor
P(MMA-MAA)	+	5	100	1.0	Poor
PBS	+	1	250	2.0	Good
NEB-31	–	30	50	3.0	Good
EBR-9	+	10	200	3.0	Poor
ZEP	+	30	10	2.0	Good
UV-5	+	10	150	3.0	Good
SAL-606	–	8.4	100	2.0	Moderate
COP	–	0.3	1,000	0.8	Poor

scission is the dominant reaction upon exposure is a *positive resist*. The most important resist criteria are contrast and sensitivity for the exposure type and energy and etch resistance.

Tone determines the areas that will be removed in the development. For positive tone resists, patterned areas are removed with the developer, whereas for the negative ones, the result is the inverse of exposed features, resist is not dissolved where irradiated.

Sensitivity quantifies the minimum amount of delivered dose that is required to achieve the selective development, where dose is the number of electrons per unit area.

Resolution defines the minimum feature size or the smallest distance between two patterns that can be resolved.

Contrast describes how abrupt is the dependence of thickness on dose, which also has implications upon resolution. It is determined from the linear slope of the sensitivity curves.

Etch resistance determines the resist integrity under chemical (wet) and physical (dry) etching processes.

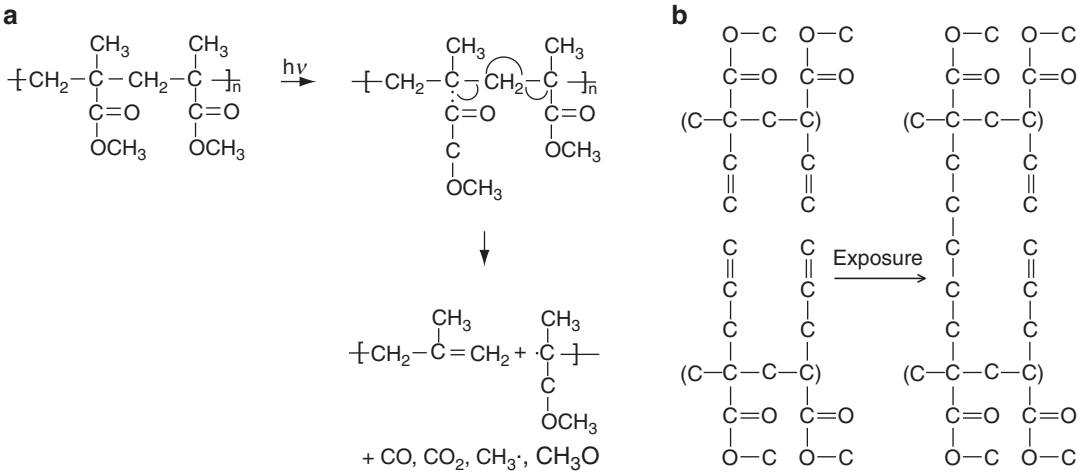
Other resist variables are adhesion to the different substrates, compatibility with conventional techniques, or shelf life. Table 2 summarizes the properties of commonly used EBL resists.

Positive Resists

Polymethyl methacrylate (PMMA) was one of the first resists developed for EBL and remains the most commonly used positive resist. PMMA has extremely high resolution, and its ultimate resolution has been demonstrated to be less than 10 nm. In PMMA, both cross-linking and fragmenting of the polymeric chains occur (see Fig. 8), but the rate of scission is much larger than that of cross-linking. If PMMA is exposed to ten times the normal critical dose, chain-linking will occur and it will behave like a negative resist; however, its resolution will be degraded to about 50 nm. One of the major disadvantages of PMMA is low sensitivity. PMMA also has poor resistance to dry etching, which makes it limited to mask making where only a thin Cr wet etch on a flat substrate is required. When it is used for direct writing, PMMA is typically used as the imaging resist for a liftoff process. PMMA is available in low-molecular-weight form and high-molecular-weight form, of which the former is more sensitive than the latter. Thus, if two layers of PMMA are used, with the low-molecular-weight form on the bottom, this will cause the resist to develop with an undercut, thus making it suitable for a liftoff process.

The copolymer, methyl methacrylate and methacrylic Acid (P(MMA-MAA)), provides a three- to fourfold improvement in sensitivity relative to PMMA and image thermal stability of 160 °C. Since P(MMA-MAA) can also be developed in MIBK:IPA solvent as used for PMMA, a single-step development of mixed layers of PMMA and P(MMA-MAA) is possible for enhancing the sensitivity and thermal stability of the resist.

A new resist called ZEP has an order of magnitude improvement in sensitivity versus PMMA and it has 2.5 times better etch resistance compared to PMMA although it is still not as resistant



Electron Beam Lithography (EBL), Fig. 8 (**a**) Mechanism of radiation-induced chain scission in PMMA (**b**) Cross-linking of PMMA structure where it is modified through the addition of a C=C side chain

as novolak-based resists. ZEP has about the same resolution as PMMA of about 10 nm and it has demonstrated an ability to produce lines of width 10 nm with a pitch of 50 nm. Other important positive resists include PBS (polybutene-1-sulfone) and EBR-9 (a copolymer of trifluoroethyl a-chloroacrylate and tetrafluropropyl a-chloroacrylate) which have high sensitivity. It is noted that the desired properties of a resist are high resolution and high sensitivity (high speed). Unfortunately, the resist that has higher sensitivity, including those mentioned here, usually has lower resolution, especially compared to PMMA.

Negative Resists

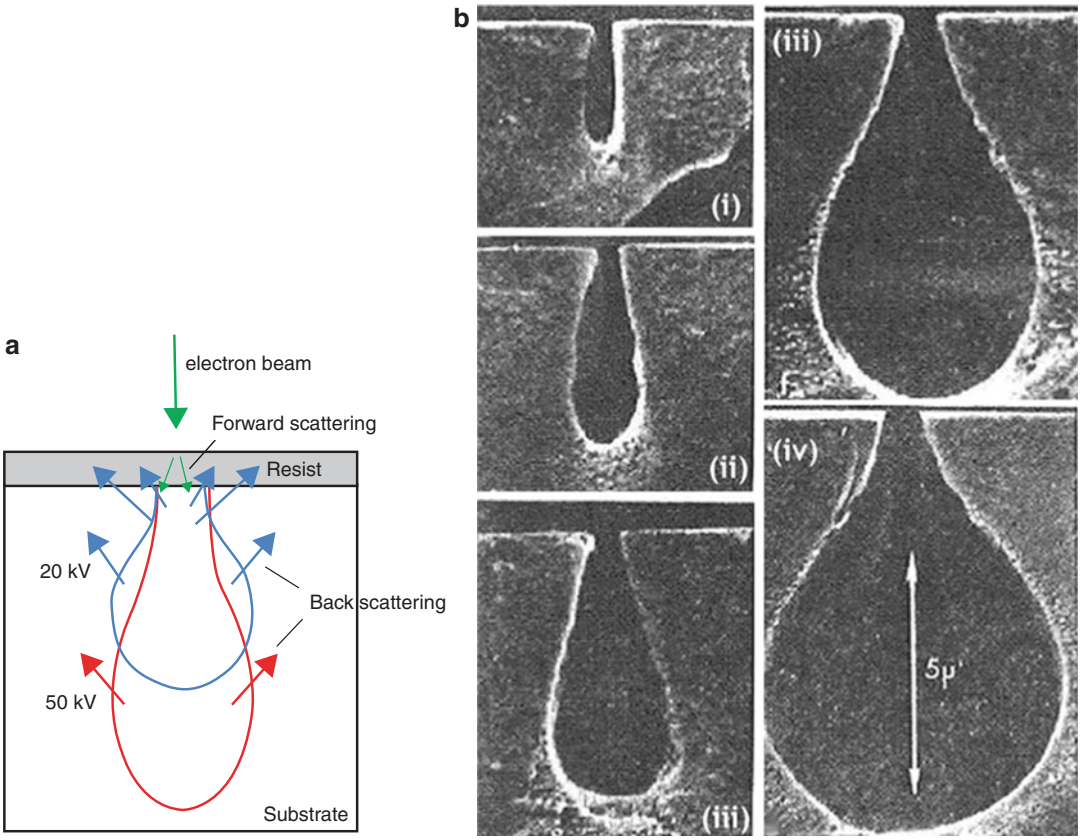
A number of negative EBL resists are also available. These resists have components on the polymer chain that enhance the cross-linking. Typical cross-linking components include chloromethyl styrene, epoxies, and vinyl groups. Negative resists tend to have less bias but they have problems with scum and swelling during development and bridging between features.

Commonly used negative e-beam resists consist of the Shipley advanced lithography (SAL) product line, an epoxy copolymer of glycidyl methacrylate and ethylacrylate (P(GMA-EA)), also known as COP, and a partially chloromethylated polystyrene (CMS). While COP has high sensitivity, CMS possesses modest resolution at modest sensitivity. The SAL offers many new deep ultraviolet (DUV) resists through the use of chemically amplified resist materials. Fast versions of CAP have demonstrated high-resolution capability with 100-nm lines at sensitivity higher than PMMA. Also, as it was mentioned before, PMMA can exhibit negative tone when exposed to a dose one order of magnitude higher [3].

Electron-Substrate Interactions

Electron beam generation and control technologies allow achieving beam diameters below the nanometer range. However, electrons being massive charged particles undergo interactions with the resist and the substrate and result in larger minimum resolution. Therefore, electron-solid interaction must be carefully taken into account to obtain small feature size by EBL.

Typical electron beam lithography machines use electron beams with 10–100 keV energy per electron. Therefore, the free path of an electron is 10 μm or more, which is at least an order of magnitude more than the resist thickness. Thus, the electrons can easily penetrate the resist layer and reach the substrate. When an electron beam hits the surface of a resist deposited on a substrate, it experiences elastic and inelastic collisions with



Electron Beam Lithography (EBL), Fig. 9 (**a**) Forward and backward scattering in resist and substrate. (**b**) Experimental evidence of pearl-shaped electron trajectories caused by electron-resist scattering. A 5-μm-thick layer of PMMA is developed with increasing times from (*i*) to (*iv*) [4]

the resist and substrate atoms and molecules. Elastic collisions result in backward scattering whereas inelastic collisions create small angle forward scattering (see Fig. 9). The energetic transfer in the resist mainly causes the resist exposure, whereas interaction with the substrate underneath is manifested as heating and creates more backscattered electrons. Usually, elastic collisions are expressed in terms of the unscreened single electron scattering of Rutherford. It evaluates the process as a probability of interaction that strongly depends on the solid atomic number and the electron energy. Inelastic collisions are the cause of the loss of electron energy and it is typically described by the Bethe's continuous slowing down model. As a result, electron trajectories are mostly dependent on electron energy and the substrate composition.

In the resist, forward scattering is the more probable process, i.e., main parameter is the energy, while the interaction with the substrate comes from backscattering, which is highly dependent on both charge energy and substrate nature. Thus, electron trajectories are typically distributed in a certain 3D interaction volume that reaches few microns in range of 10–20 keV for incoming electron energy. The usual pearl shape of electron distribution in the resists can be easily explained from the low atomic number and density of the resists (Fig. 7b). High-energy electrons initially are just slightly deviated from the incoming direction, since elastic contribution is weak ($\sim Z^2$), but once they slow down, lateral scattering dominates ($1/E^2$). Electron trajectories in resist and substrate are often modeled by Rutherford scattering and simulated using Monte Carlo

techniques. However, resultant increase in the beam diameter can also be estimated by an empirical formula [5]:

$$\Delta d = 0.9\left(\frac{t_R}{V_b}\right)^{1.5} \tag{6}$$

where Δd is the change in beam diameter after entering the resist, t_R is the resist thickness, and V_b is the accelerating voltage of the electron beam. Forward scattering is detrimental if very high resolution is required, and can be reduced by increasing the accelerating voltage. Forward scattering does have one advantage, however; it can be used to create a negative sidewall angle in the developed resist. This is advantageous when doing liftoff process with positive resist, because a slight negative sidewall angle is required. This eliminates the need for using two layers of different resists and simplifies the processing.

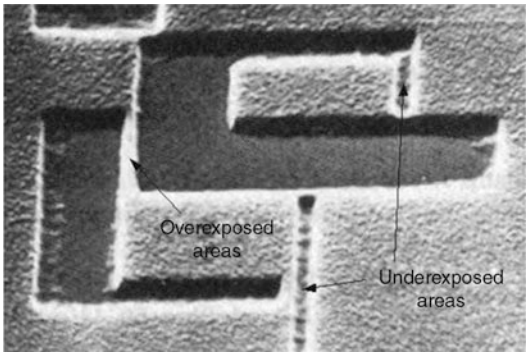
The electrons impinging on the resist and substrate have a large energy (typically 10–100 keV). Their forward scattering results in generation of *secondary electrons* with lower kinetic energies (typically 20–50 eV). The major part of the resist exposure is due to these electrons. Since they have low energies, their range is only a few nanometers. Therefore, they contribute little to the proximity effect. However, this phenomenon, together with the forward scattering, effectively causes a widening of the exposure region. This is one of the main limiting factors in resolution of e-beam lithography machines. The distance a typical electron travels before losing all its energy depends on both the energy of the primary electrons and the type of material it is traveling in. The fraction of electrons that are backscattered is roughly independent of beam energy. It does, however, have a strong relation to the substrate material. Substrates with a low atomic number give less backscattering than substrates with high atomic number.

Proximity Effects

As it is mentioned above, as the electrons penetrate the resist and the substrate, they experience many scattering events. In backscattering, an electron collides with the much heavier nucleus, which results in an elastic scattering event. The electron retains (most of) its energy, but changes its direction. The scattering angle may be large in this case. After large angle scattering events in the substrate, electrons may return back through the resist at a significant distance from the incident beam, thereby causing additional resist exposure outside the desired area. This phenomenon caused by electron backscattering is called *proximity effect*. It is important to correct for the proximity effects to avoid erroneous pattern developing.

Proximity effects may be observed in the final pattern in two different forms: intrashape and intershape proximity effects. The intrashape proximity effect causes smaller shapes to be underexposed relative to larger shapes written with the same dose. In this case, each pixel in a smaller shape gets less exposure from scattered electrons, because there are fewer neighboring exposed pixels. In a positive resist, intrashape proximity effects may result in underdeveloping of narrow lines or small features. Intershape proximity effects are the result of the fact that exposed shapes are also affected by other nearby shapes, with electrons scattering through the intervening resist. Densely packed shapes will strongly interact, providing much "cooperative exposure" between the shapes, and will thus require a lower exposure dose than sparse, isolated shapes, which get fewer scattered electrons and will thus be underexposed relative to their cousins with closer neighboring shapes. In a positive resist, intershape proximity effects may result in accidental exposure of narrow fields in between large exposed areas and thereby developing completely. These effects are illustrated in Fig. 10.

The first step of the proximity-effect correction is an accurate knowledge of the energy density profile deposited in the electron-resist layer due to a point or pixel exposure (often called point spread function). In general, this profile is a function of the system setup. An important property of these profiles is that the shape is independent of dose as well as position, assuming a planar and homogeneous substrate. This profile is often



Electron Beam Lithography (EBL), Fig. 10 SEM picture of a positive resist pattern affected by the intrashape and intershape proximity effects [5]

approximated by the sum of two Gaussian distributions:

$$f(r) = \frac{1}{\pi(1+\eta)}\left\{\frac{1}{\alpha^2}\exp\left[-\left(\frac{r}{\alpha^2}\right)^2\right] + \frac{\eta}{\beta^2}\exp\left[-\left(\frac{r}{\beta^2}\right)^2\right]\right\} \quad (7)$$

where r is lateral distance from incident beam; α is standard deviation of forward scattered electron distribution; β is standard deviation of backscattered electron distribution; η is ratio of energy deposited by forward scattered electrons to the energy deposited by backscattered electrons. The first term represents the contribution of forward scattered electrons, whereas the second term is related with the backward scattered electrons. Equation 7 is normalized so that

$$\int_0^\infty f(r)2\pi dr = 1 \quad (8)$$

for some substrates and resists, a better match between the experimental results and the model is achieved with the addition of a third contribution. As a matter of fact, deposited energy distribution is not uniquely caused by proximity effects in terms of electron scattering, so, due to this, more accuracy is obtained when an exponential decay modulates proximity function:

$$f(r) = \frac{1}{\pi(1+\eta+v)}\left\{\frac{1}{\alpha^2}\exp\left[-\left(\frac{r}{\alpha^2}\right)\right] + \frac{\eta}{\beta^2}\exp\left[-\left(\frac{r}{\beta^2}\right)^2\right] + \frac{v}{2\gamma^2}\exp\left[-\left(\frac{2}{\gamma^2}\right)\right]\right\} \quad (9)$$

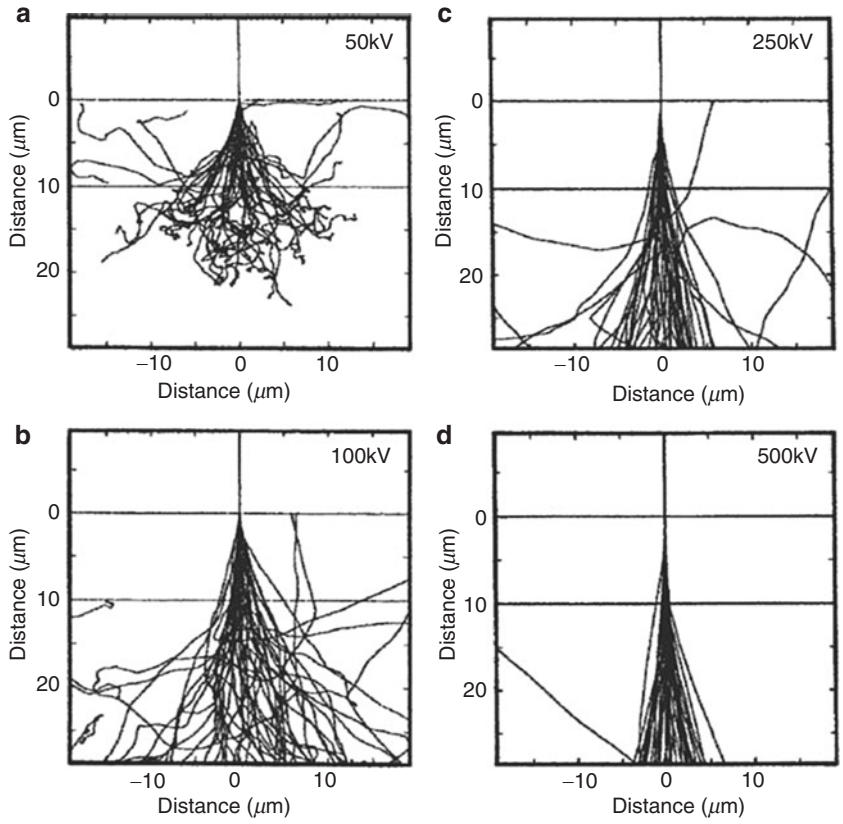
where v is efficiency of exponential function and γ is decay of exponential function.

The models discussed above are two-dimensional versions of an essential three-dimensional phenomenon. In general, the energy profile depends upon depth as well as radius. By averaging out the depth dependence, a two-dimensional profile can be obtained out of a three-dimensional profile. The simplified models allow great reduction in computation time for the exposure estimation and correction. For certain applications, it may be necessary to use a three-dimensional profile. In this case, a Monte Carlo simulation of electron scattering in the resist layer can be used. Electron scattering in resists and substrates can be modeled with reasonable accuracy by assuming that the electrons continuously slow down as described by the Bethe equation, while undergoing elastic scattering, as described by the screened Rutherford formula (Fig. 11).

Once the proximity function is correctly determined with the coefficients, proximity effects can be corrected for by several methods.

Dose modification and its variants are the most common correction. It is based on determination of the required dose for each pixel with a reasonable accuracy while being computationally practicable. Numerous variants of the "self-consistent dose correction" have been developed. In its simplest form, it is basically the reverse of the exposure estimation. Let Q_j be the dose applied to pixel j and let N be the total amount of pixels. The total energy on pixel i will be:

$$E_i = \sum_{j=1}^{N=} R_{ij}Q_j \text{ with } R_{ij} = \frac{\Delta V_f}{t}\left(\frac{1}{\pi\alpha^2}\exp\left(-\frac{r_{ij}^2}{\alpha^2}\right) + \frac{\eta}{\beta^2}\exp\left(-\frac{r_{ij}^2}{\beta^2}\right)\right\} \quad (10)$$



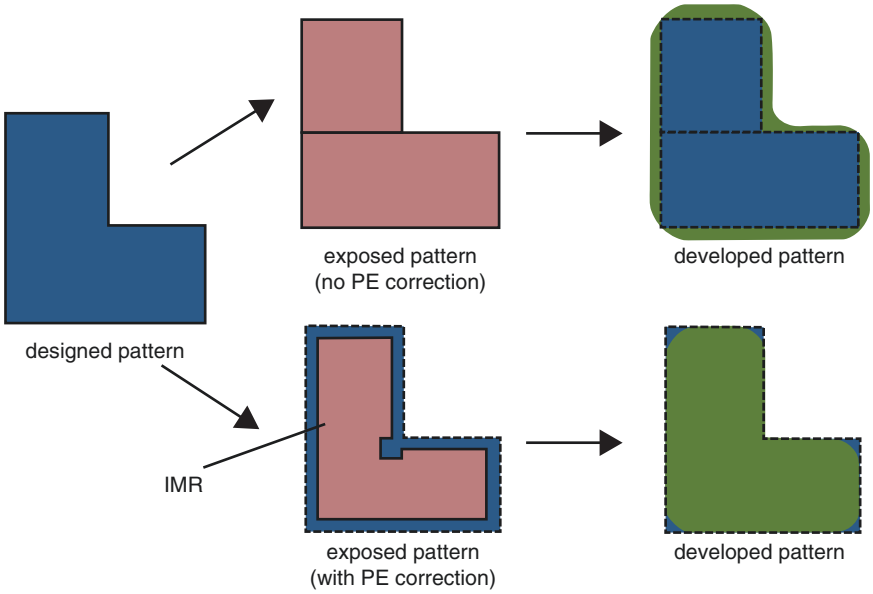
Electron Beam Lithography (EBL), Fig. 11 Electron trajectory simulations using Monte Carlo for 10 mm of resist on a silicon wafer for different electron energies [6]

where r_{ij} is the distance between pixel centers of i and j. This equation can be written in matrix notation for all i:

$$[E_i] = [R_{ij}][Q_j] \tag{11}$$

Solving this set of equations with matrix operations will provide a proximity-effect corrected pattern. However, this "self-consistent" scheme will not be a perfect correction since only the exposed pixels are considered in this equation. The problem can be simplified, thereby reducing the calculation time, by splitting the dose modification into a problem of forward scattering and back scattering correction. With dose modification, it is possible to achieve superior proximity-effect correction. The main disadvantage is that with very large circuits, it may require large computation times.

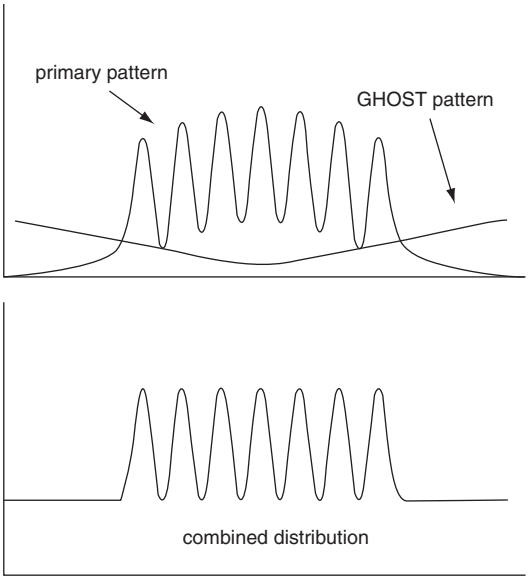
In another correction method, the shapes found in the pattern image are modified in such a way that the developed image will resemble the intended image as close as possible, while a single dose is used for the entire circuit. A good example of a shape modification method is the correction scheme in PYRAMID [7]. PYRAMID takes a pattern with rectangular circuit elements. The circuit is then passed to a correction hierarchy, which adjusts each element via pre-calculated rule tables. This rule table is created using exposure estimation as described earlier. The first step is to replace each rectangle with its inner maximal rectangle (IMR). The second step is to correct the effect of interaction among the different circuit elements. Each edge facing



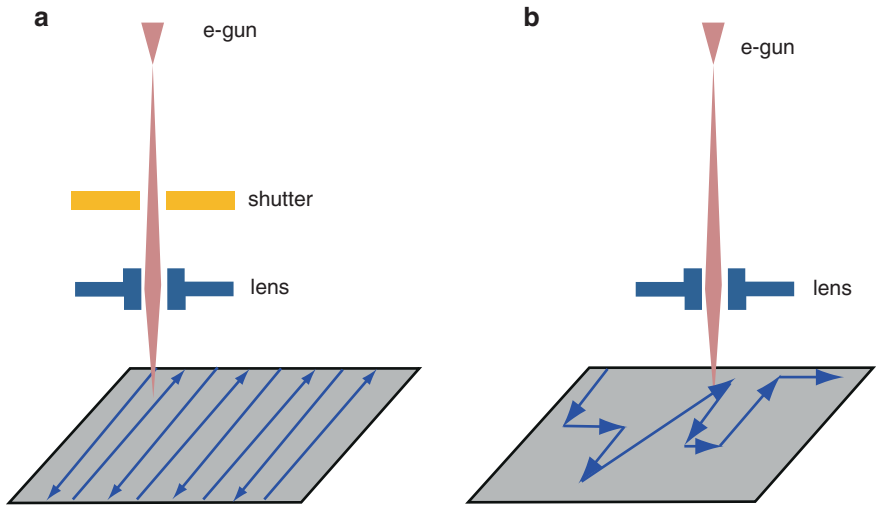
Electron Beam Lithography (EBL), Fig. 12 PE correction by shape modification at critical points

other circuit elements will be adjusted so that the midpoint of the edge will be equal to the experimentally determined development threshold. Even better results can be obtained by bending the edge when appropriate. The final step is to modify the shapes at critical points, that is, junctions between adjacent rectangles (Fig. 12). The major advantage of the shape modification method is that accurate results can be obtained without being computationally expensive and without losing throughput. However, it may not be as flexible as the other methods and experimental data are needed to obtain the necessary rule tables.

Background exposure correction, often referred to as GHOST, works by writing a second exposure, which is the inverse of the intended image (Fig. 13). This is done in such a way that the background dose is brought to a constant level. The main advantage of this method is that it is one of the easiest proximity-effect correction methods found in the literature and can be used with virtually all electron beam machines. However, there are several disadvantages with this method. One problem may be a contrast reduction (although this is true for many other correction types too). However, the main problem is that it only provides correction to the backscattering component, where the forward scattering remains uncorrected. There is also a loss of throughput because of the double exposure.



Electron Beam Lithography (EBL), Fig. 13 The figure on the top shows a pattern of lines, and the GHOST pattern superimposed. The figure on the bottom shows the summed contribution of the GHOST pattern and the line pattern [5]



Electron Beam Lithography (EBL), Fig. 14 Schematic representation of (**a**) raster scan (**b**) vector scan

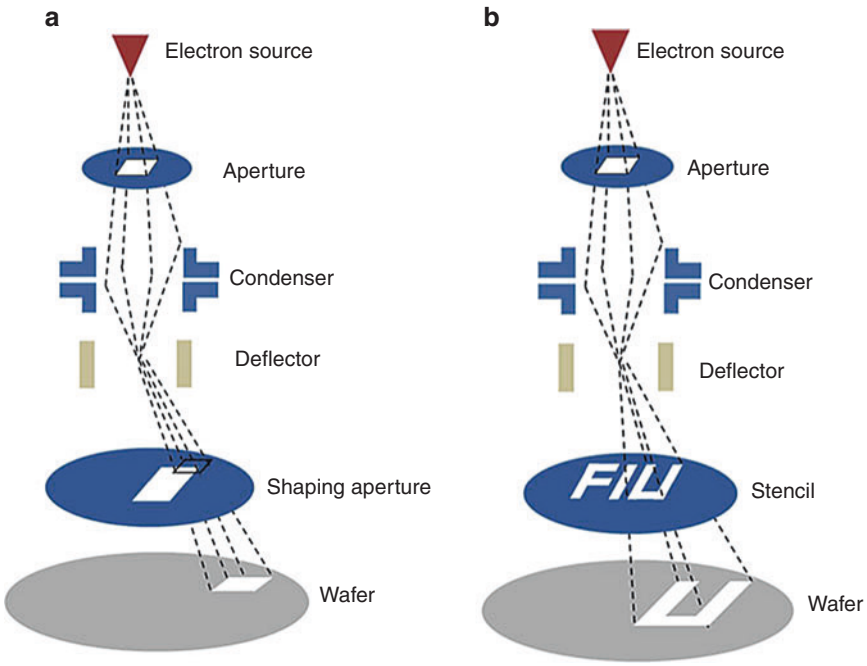
Electron Beam Lithography Systems

Although they all are based on the same basic principles, different types of electron beam lithography systems are available. These systems as well as some proposed future systems will be briefly discussed in this section.

Gaussian Beam EBL Systems

The most EBL systems today are based on scanning electron microscope (SEM) or scanning transmission electron microscope (STEM) systems since they share the same electron optics column. In fact, the first EBL systems were direct descendents of scanning electron microscopes and for research applications, it is very common to convert an electron microscope into an electron beam lithography system using a relatively low cost accessory (~ $100 k). This provides a huge saving especially for academic research facilities considering that dedicated e-beam writing systems are very expensive (~$4 M). SEMs normally have some lenses to focus the beam on the specimen, a beam-blanking mechanism, and some magnetic deflection coils used to scan the beam across the sample. Controlling these components by a computer allows transfer of any design in digital format to an exposure pattern. Digital-to-analog (DAC) converters are the key components for this transfer. The area to be scanned by the beam is divided into fields and subfields. This makes high speed (subfield) scanning achievable with DAC electronics when the address contains fewer bits (e.g., 12 bits). Then the large-field scanning is performed by tiling together subfields with a high-precision DAC with larger addressing (e.g., 20 bits or larger). For the given DAC sizes and 1-nm address grid, it can be estimated that the maximum subfield size is 2^{12} nm (~4 μm) and the maximum main field is 2^{20} nm (~1 mm) [8]. The scan trajectory is defined as the path, which the beam would be deflected along if every pixel were to be exposed. The actual pattern is written by opening and closing a shutter. The scan method in which the beam scans the entire field by following the bit map data organized serially is called *raster scan* (Fig. 14a). Since every pixel must be scanned serially in raster scan, exposure time is independent of the pattern and longer than the alternative method *vector scan*.

Vector scan method is developed to improve the throughput by directing the electron beam to only those parts of the chip to be exposed (Fig. 14b). The digital address of each area to expose is sent to DAC and the beam scans over only the pixels, which must be exposed. Some systems use two sets of DACs: The first DAC drives a deflector, which directs the beam onto



Electron Beam Lithography (EBL), Fig. 15 Simplified schematics of (**a**) variable shape aperture (**b**) cell projection approach with a stencil mask

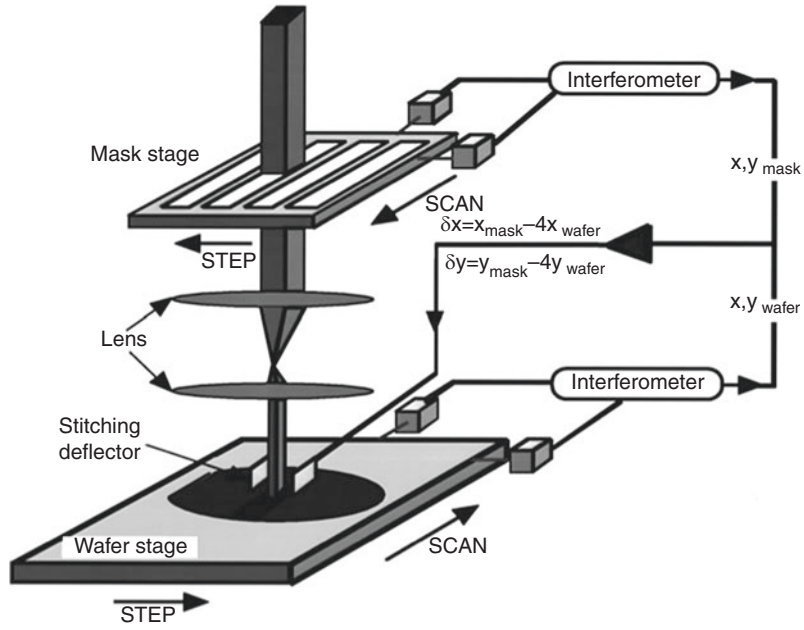
the corner of some primitive shape on the target wafer; the second DAC drives the same deflector or a different deflector in a raster pattern to fill in the shape. A moveable x – y substrate stage is often used, and is normally controlled using a laser interferometer to give accuracy down to the wavelength of the laser being used. Another advantage of the vector scan is the higher precision. Ignoring system aberrations and nonlinearities, the precision of the pattern address is determined by the width of the digital word. Therefore, pixels can be placed on an extremely fine grid if a high-speed, wide word DAC is used.

Variable-Shaped Beam and Mask Projection EBL Systems

As it was explained in detail before, direct write systems use a small electron beam spot that is moved with respect to the wafer to expose the pattern: one pixel at a time. This technique inherently suffers from low throughput to become a major process for economically manufacturing complex ICs. Inspired by the photolithography techniques, special EBL systems similarly have been developed to improve the throughput.

In one approach, the beam is partially intercepted by a square aperture. The electron beam is deflected and rotated electrostatically to adjust the width, length, and angle of the line exposed on the wafer. Since most images are made up of rectangles, hundreds or even thousands of pixels can be exposed simultaneously. Line width is somewhat more difficult to control in this method but for many applications, the increase in throughput is sufficient to compensate for the reduction in line width control (Fig. 15a).

In the cell projection approach, a mask with predetermined geometries similar to the photomasks in photolithography is placed between the wafer and the beam source (Fig. 15b). Since the short penetration length of electrons prohibits the use of a quartz mask, a stencil mask is used instead. Highly repetitive features such as DRAM cells can be transferred at a much higher throughput than conventional EBL. Cell projection has an advantage over shaped beam systems

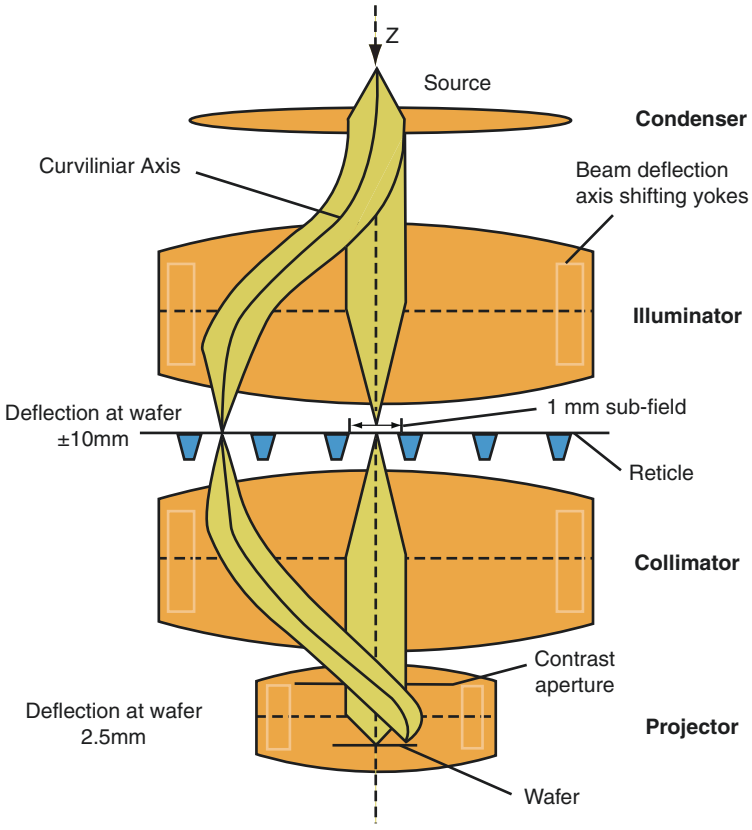


Electron Beam Lithography (EBL), Fig. 16 Schematic diagram of the SCALPEL principle showing the function of the scattering mask in forming image contrast [9]

in that complex shapes can be created on a stencil mask beyond rectangles and triangles, at a modest increase in throughput.

To be able to stop highly energetic electrons, the stencil masks in cell projection method have to be large even to pattern a small area. The energetic electrons absorbed by the mask also heat up the mask, which can significantly alter the size of the features. New projection EBL systems have been developed to solve these problems. SCattering with Angular Limitation Projection Electron-beam Lithography (SCALPEL) system developed by Bell Labs is a noteworthy approach. SCALPEL approach combines the high resolution and wide process latitude inherent in electron beam lithography with the throughput of a parallel projection system and avoids the problems of stencil masks. In the SCALPEL system, a mask consisting of a low atomic number membrane and a high atomic number pattern layer is uniformly illuminated with high-energy (100 keV) electrons (Fig. 16). The entire mask structure is essentially transparent to the electron beam and little of its energy is deposited there. The portions of the beam, which pass through the high atomic number pattern layer, are scattered through angles of a few milliradians. An aperture in the back focal plane of the electron projection imaging lenses stops the scattered electrons and produces a high contrast image at the plane of the semiconductor wafer. The SCALPEL approach separates the pattern formation and energy absorption to maintain pattern fidelity. The pattern is defined by the mask although the energy from the unwanted parts of the beam is deposited on a mechanical aperture [9]. Throughput on the SCALPEL system has been predicted to be approximately 45 wafers per hour. This is ten times higher than raster-scan EBL systems, and only 50 % slower than optical lithography machines. It is estimated by SCALPEL's designers that SCALPEL will be cheaper than optical lithography systems, which use RET and hence, could be used for the sub-100-nm ULSI nodes. However, further development and commercialization of SCALPEL was stopped in January 2001.

IBM has also developed a projection EBL system, which is dubbed projection reduction exposure with variable-axis immersion lenses (PREVAIL). The PREVAIL system increases the throughput by shifting the main optical axis of the electron beam through the use of variable-axis lens designed at IBM. Deflections within each subfield can be superimposed onto the magnetic



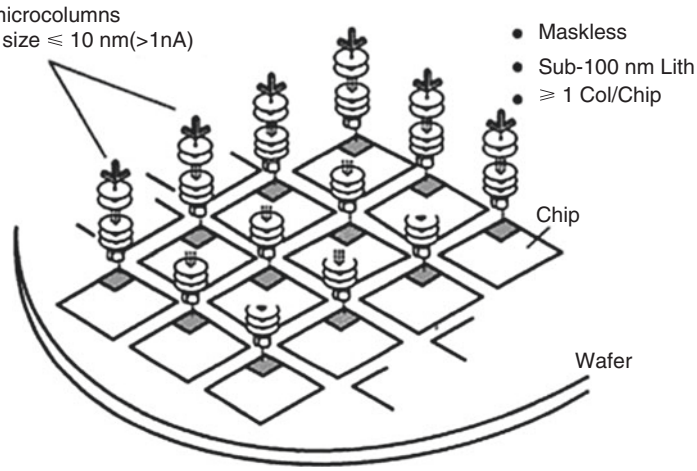
Electron Beam Lithography (EBL), Fig. 17 Core technology of the IBM's PREVAIL EBL system: Curvilinear variable-axis lenses (*CVAL*)

and electric fields used in the variable-axis lens. This variable-axis system allows the beam to be moved quicker from cell to cell on the mask, and reduces the dependence on slower mechanical components (Fig. 17).

Multi-beam EBL Systems

Throughput of EBL can also be increased by employing multiple beams. One way to obtain multiple beams is to use a series of electron sources rather than a single source. With the advancements of nanomanufacturing technologies, micromachined electron gun (MEG) devices have been proposed. A MEG device consists of monolithically integrated lenses feedback resistors and heaters. MEGs operate at much lower voltages (~200 V) than conventional electron guns and yield a spot size of 8.7 nm at a working distance of 20 μm. The lower voltage operation of these emitters mean that the electrons will have negligible backscatter and no proximity-effect corrections will be needed; however, exposure times may be longer due to the low narrow/low current beam. Since this technology is still at prototype stage, some details regarding the sample mounting and possible throughput are not clear. Very similar approach named sTM aligned field emission (SAFE) microcolumn arrays have also been proposed. SAFE technology uses a 2D array of miniaturized electron sources (Fig. 18). Although details of the manufacturing technology have not been fully disclosed, SAFE system is estimated to achieve a resolution of 25 nm and throughput of 60 count 200 mm diameter wafers per hour.

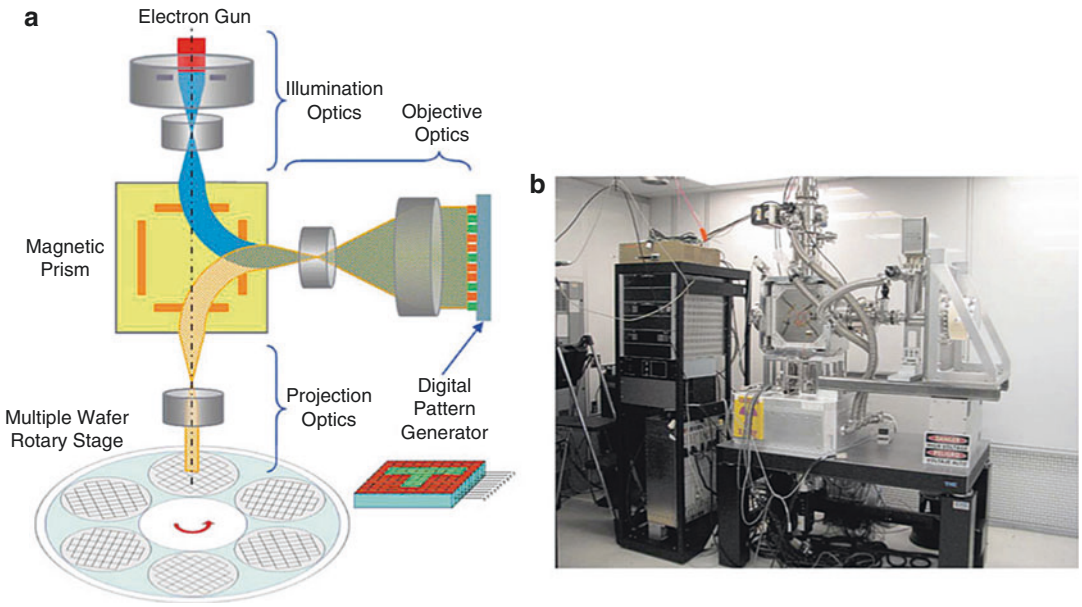
An alternative method proposed to obtain multiple beams is using one large electron beam and then dividing it into smaller beams by using miniature MEMS lenses and apertures. Smaller secondary beams are then focused on the target wafer [11]. This system has a 32×32 array of subelectron beams which cover a total of



Electron Beam Lithography (EBL), Fig. 18 Conceptual drawing of microcolumn array for parallel e-beam lithography Ref. [10]

3:6 × 3:6 mm area. It would require arrays of small lenses, blanking deflectors, blanking apertures, and beam deflectors, all implemented in a silicon substrate. The deflectors only need to deflect the beam in one direction, and a stage can move the sample in the other direction. No prototype of such system has been produced yet.

Recently, a novel maskless EBL system called the reflective electron beam lithography (REBL) has been proposed by KLA-Tencor Corporation [12]. REBL system incorporates six core technologies: reflective electron optics, the digital pattern generator, time domain integration, gray tone exposure, rotary stage, and optical wafer registration. Reflective electron optics enables the use of the digital pattern generator (DPG) chip to independently control 1×10^6 beams for massively parallel exposure. The electron optics ray diagram of the REBL nanowriter is shown in Fig. 19. A large area cathode of relatively low reduced brightness ($Br < 10^4$ Am^{-2} sr^{-1} V^{-1}) generates an illuminating electron beam. The cathode lens forms a crossover at the center of the condenser lens. The condenser lens forms an image of the source at the virtual center of the magnetic prism. The prism is set up to bend the optical axis exactly 90° and establish the focal lengths in the x and y planes to be equal, with equal magnifications in the x and y axes. When these three conditions are satisfied, the prism behaves as a 1:1 imaging lens, imaging the crossover from the center of the condenser lens to the center of the transfer lens. The transfer lens in turn images the source image at the virtual prism center onto the DPG through the DPG lens. The electrostatic DPG lens serves a dual purpose: It forms a virtual image of the crossover at infinity, and it decelerates the electrons to within a few volts of the cathode potential. When a mirror is turned "on" at the DPG, it will reflect electrons from the illuminating beam. These reflected electrons are reaccelerated back toward the prism as they travel through the accelerating field of the electrostatic DPG lens. A crossover is formed at the center of the transfer lens by the DPG lens. The transfer lens images the DPG at the virtual center of the prism but along the diagonal 90° clockwise from the first image formed on the illuminating arm. The reflected electron beam, going in the opposite direction, is therefore bent in the opposing direction. The prism thus bends the reflected beam away from the gun to project it onto the wafer. The prism also forms a crossover at the back focal plane of the demagnification lens, while the demagnification lens focuses the DPG image from the virtual center of the prism onto the wafer plane. The prototype REBL system has demagnification of 50 and DPG pixel pitch of 1.5 μm which produce a pixel pitch at the wafer plane of 30 nm. The system targets five to seven wafers levels per hour on average at the 45-nm node with extendibility to the commercial mainstream lithography market for the 32-nm node and below.



Electron Beam Lithography (EBL), Fig. 19 (**a**) Diagram of REBL nanowriter concept. (**b**) First REBL column on test stand for characterization [12]

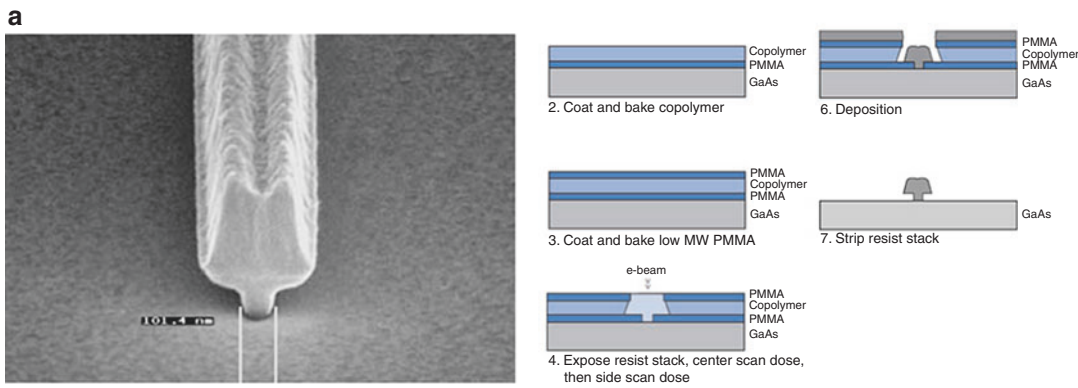
Examples of Application

EBL is arguably the most versatile technique in nanofabrication technologies. Ability of direct writing nanoscale features allows researchers to fabricate experimental proof-of-concept and prototype devices, circuits and systems in diverse fields including electronics, photonics, microelectromechanical systems, and material sciences. Although it is impossible to cover all these applications, selected ones will be presented in this section to give a glimpse of the capabilities of EBL.

Transistors for High-Frequency Applications

With their high breakdown voltage and large sheet carrier concentrations, nitride-based wide band gap materials and devices attracted great deal of attention in the last decade. Particularly AlGaN/GaN HEMTs have been researched for their potential applications in wireless communication and radar replacing the incumbent GaAs devices. Fabrication of AlGaN/GaN single devices and monolithically integrated microwave integrated circuits (MMICs) involves several nanofabrication steps. Among them, fabrication of transistor gate is particularly important as the gate length is the smallest dimension in the transistors and is the most important parameter that determines frequency response of the transistors. Since the gate resistance is an important parasitic element affecting key parameters like f_{max}, it should be minimized. To minimize the gate resistance for a given gate width, it is essential to have a large cross section. However, a large cross section would mean a long gate footprint affecting frequency response. This problem can be tackled by making a mushroom like (T-gate) structures as shown in Fig. 20. A T-gate is a structure that has a narrow footprint on the semiconductor and large head. The large head of the gate increases the cross-section area of the structure and thereby reduces the gate resistance. To fabricate T-gates one takes advantage of the difference in developing times for a given dose for the two commonly used e-beam resists, PMMA and copolymer (P(MAA-MMA)). This provides sufficient contrast to fabricate the three-dimensional shape required.

The process for fabrication of gates usually involves spinning three-layer stack, consisting of PMMA/copolymer/PMMA. Then, the resist is exposed with three passes: one strong centerline

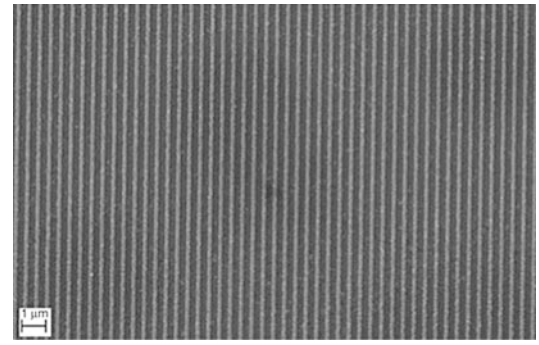


Electron Beam Lithography (EBL), Fig. 20 (**a**) SEM picture of a 100-nm "T-gate" of an AlGaN/GaN HEMT fabricated by 3-layer resist EBL. (**b**) Typical process flow for "T-gate" fabrication using tri-layer process (Courtesy of MicroChem Corp.)

dose, which helps in clearing the resist to define the foot of the gate first and two light side area doses to help develop a big head. CAD layout is designed accordingly. Patterns are written by an e-beam lithography tool using low current (~1 nA). After developing the resists, a metal stack (e.g., Ni/Au) with adequate thickness is deposited. Liftoff is performed usually in methylene chloride. Figure 20 shows the SEM picture of a T-gate with a footprint of 100 nm.

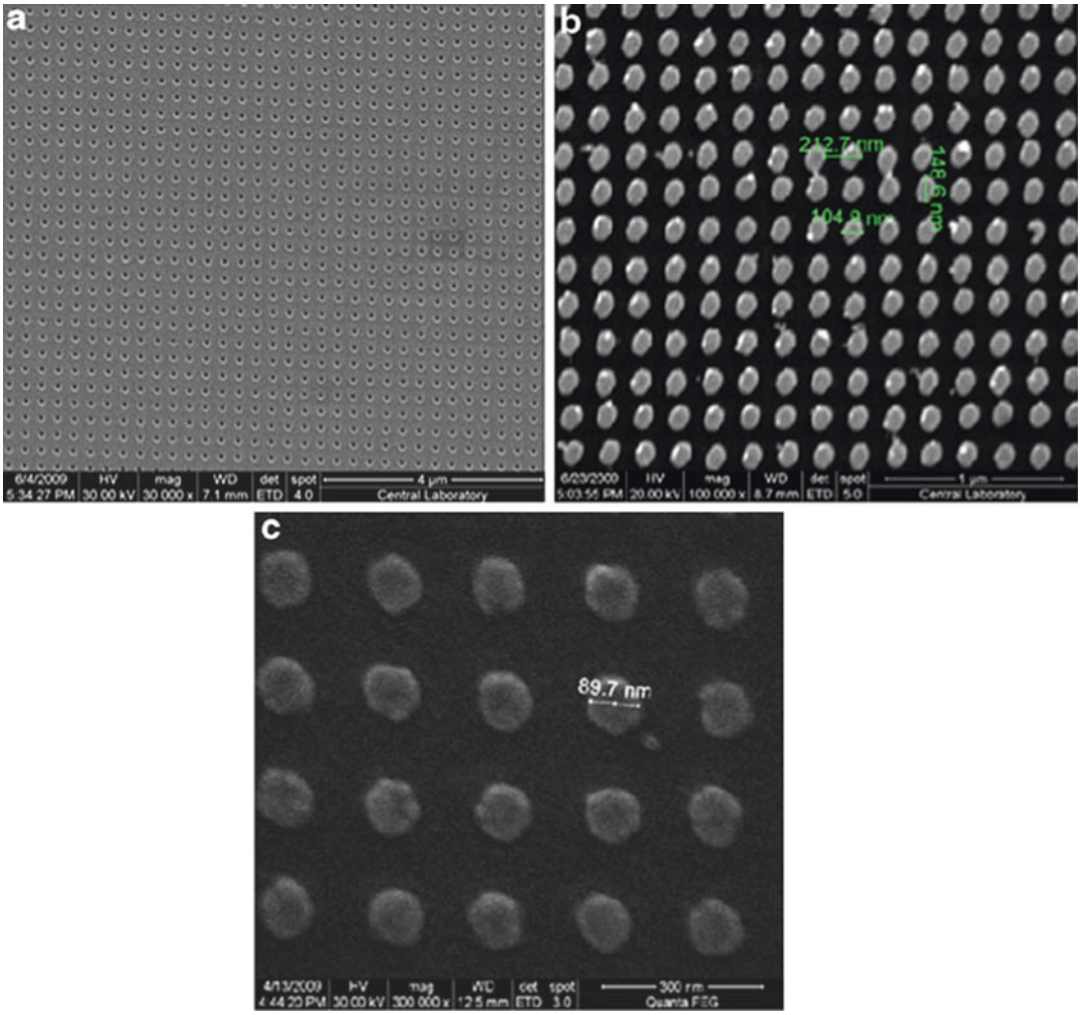
Surface Plasmonic Devices

Surface plasmon polaritons (SPPs) originate from the collective oscillations of conduction electrons coupled with photons propagating at metal-dielectric interfaces. Localization of SPPs can find wide range of applications in technology like SPP-based Bragg reflectors, emitters, and filters. Their massive localized electric fields are used to enhance weak Raman signals. Using surface plasmon enhancement even a single molecule can be detected on plasmonic surfaces. SPP cavity structures can be used as a template in surface-enhanced Raman spectroscopy (SERS) applications. Template supports a localized plasmonic mode that can be excited by the laser to be used in SERS. These templates are created by electron beam lithography. SPP-localized modes are adjusted by tuning the periodicity of the metallic grating. A sample image of metallic grating created by EBL is shown in Fig. 21.



Electron Beam Lithography (EBL), Fig. 21 *Top view* SEM image of a uniform grating fabricated by e-beam lithography with a period of 440 nm

PMMA is used as resist to create a uniform grating by EBL. The line width of the exposed regions depends on the following parameters: dose exposed per unit length, develop time, developer concentration, film thickness of PMAA, baking time and temperature, period of the grating (proximity affect), type of PMMA, silver thickness under PMMA, type of substrate. Correction of proximity effects is extremely important to obtain precise grating spacing, especially for small periods. In addition to using a PE correction software, an empirical "dose matrix" test method

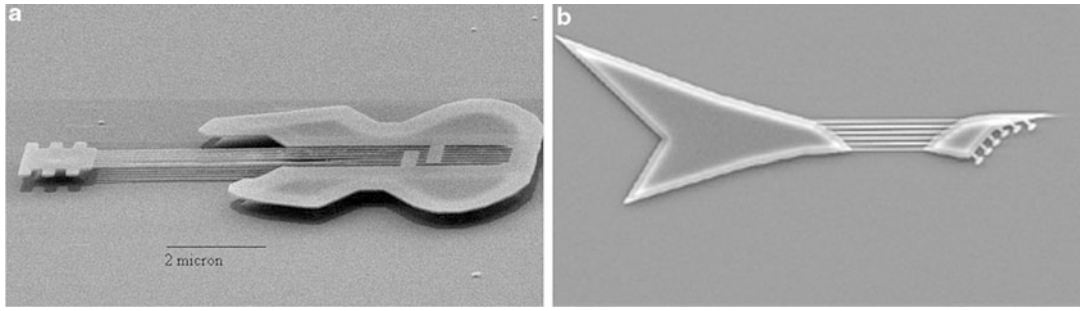


Electron Beam Lithography (EBL), Fig. 22 (**a**) Array of 100-nm diameter nanoholes with a period of 300 nm patterned on PMMA over an ITO-coated microscope slide. (**b**) SEM images of metal nanoparticles patterned over ITO-coated glass with different beam currents. (**c**) Enlarged image of the nanoparticles (Courtesy of Urcan Guler)

is also employed to determine the optimum exposure parameters. In a typical dose matrix test, the exposed dose on the resist is varied and other parameters are kept unchanged. Different fields are exposed with different doses. All fields are processed to the end. After checking the dose matrix with SEM, best exposure doses are selected to create gratings for measurements.

Light Trapping for Solar Cells

Light trapping with metal nanostructures on thin-film photovoltaic solar cells increases the solar cell efficiency. Light is coupled to metal nanoparticles and then coupled to the thin-film solar active layer over a wide range of angles. This procedure enhances the effective path length and absorption in the layer. So, metal nanoparticles can be used to increase efficiency of the devices based on the absorption of light. Enhancement effects of metallic nanoparticles are also demonstrated for dye-sensitized solar cells. Periodically structured metallic nanoparticles can trap the light onto the thin-film solar cell. Such structures can be created by EBL. First, an array of nanoholes is patterned on PMMA over an ITO-coated microscope slide Fig. 22a.



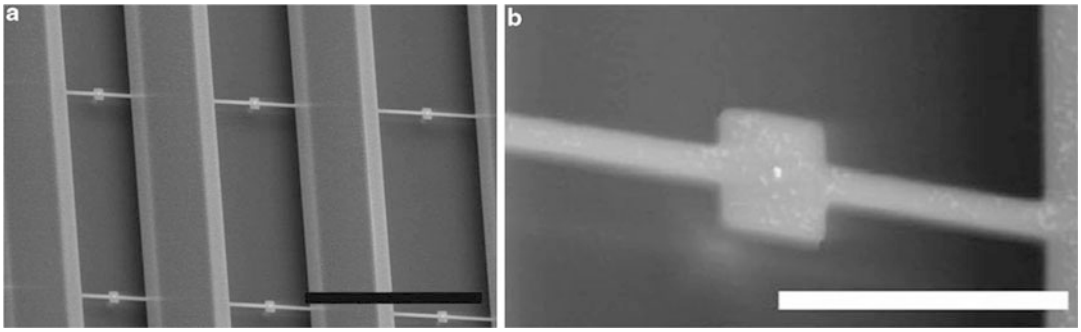
Electron Beam Lithography (EBL), Fig. 23 (**a**) World's smallest guitar, the "nano-guitar" fabricated on silicon using e-beam lithography [13] (**b**) Newer version of the "nano-guitar" which can actually be played by focused laser beams [14]

After standard EBL fabrication process, the desired nanoparticle-coated thin-film solar cell structure can be obtained (Fig. 21b, c). Period and size of the nanoparticles can be tuned for desired absorption spectrum of the incident light.

Micro and Nano Electromechanical Systems (MEMS and NEMS)

MEMS and NEMS were envisioned long time ago and put forward as a challenge by Richard Feynman in his seminal talk "There is plenty of room at the bottom" in 1959. Advancements in the nanofabrication technologies have allowed replication of many mechanical systems in nanometer scale and even creation of novel systems by taking advantage of physics at this scale. Electron beam lithography is one of the key technologies, which made possible the development of MEMS and NEMS. A striking example of the capabilities of MEMS fabrication by EBL is the *nano-guitar* In 1997, Cornell University researchers built the world's smallest guitar – about the size of a red blood cell – to demonstrate the possibility of manufacturing tiny mechanical devices using nanofabrication techniques [13]. It was 10 μm long, with six strings each about 50 nm, or 100 atoms, wide and made from crystalline silicon (Fig. 23a). Six years later the same group fabricated a new version of the nano-guitar, which is about five times larger than the original, but still so small that its shape can only be seen in a microscope [14]. Its strings are really silicon bars, 150 by 200 nm in cross section and ranging from 6 to 12 μm in length (Fig. 23b). The strings vibrate at frequencies 17 octaves higher than those of a real guitar, or about 130,000 times higher. What is unique about the second version is that it is possible to "play" it by sending focused laser beams. Since there is no practical microphone available for picking up the guitar sounds, however, the reflected laser light was computer processed to provide an equivalent acoustic trace at a much lower frequency. Both demonstrate the power of nanotechnology for a new generation of electromechanical devices.

One of the most widely explored application of MEMS and NEMS is sensitive detection of bound mass. With selective binding surfaces, the devices could be exquisitely sensitive to binding of selected chemical or biological species. In the selected case, EBL-supported nanofabrication capability was used to fabricate arrays of both polycrystalline silicon and silicon nitride resonators with evaporated gold contact pads. Circular gold pads 50–400 nm in diameter were deposited on the bridge resonators to serve as anchors for self-assembled thiol monolayer (SAM) for selective binding of target substances (Fig. 24). Detection of the resonant frequency shift allowed the determination of the mass of the adsorbed SAM. Nanoscale resonators presented



Electron Beam Lithography (EBL), Fig. 24 (**a**) SEM micrograph of arrays of bridge oscillators. (**b**) Oblique-angle SEM micrograph of oscillator. The diameter of the gold pad at the center is 50 nm. Scale bars correspond to 2 μm [15]

quality factor of Q = 8,500 and allowed detection of 2.7 attogram (=2.7 × 10^{-18} g) minimum resolvable mass [15].

Cross-References

► Scanning Electron Microscopy

References

1. Rius Sune, G: Electron lithography for nanofabrication. Ph.D. Dissertation, Universitat Autonoma de Barcelona (2008)
2. Herriott, D.R., Brewer, G.R.: Electron-beam lithography machines, Chapter 3. In: Brewer, G.R. (ed.) Electron-Beam Technology in Microelectronic Fabrication, pp. 141–216. Academic, New York (1980)
3. Tseng, A.A., Chen, K., Chen, C.D., Ma, K.J.: Electron beam lithography in nanoscale fabrication: recent development. IEEE Trans. Electron. Packag. Manuf. **26**(2), 141–149 (2003)
4. Goldstein, J., Newbmy, D., Joy, D., Lyman, C., Echlin, P., Lifshin, E., Sawyer, L., Michael, J.: Scanning Electron Microscopy and X-ray Microanalysis, 3rd edn. Springer, New York (2003). ISBN 0 306 47292 47299
5. Rai-Choudhury, P.: Handbook of Microlithography, Micromachining and Microfabrication, volume 1: Microlithography, vol. PM39. SPIE Press, Bellingham (1997)
6. Jones, G., Blythe, S., Ahmed, H.: Very high voltage (500 kV) electron beam lithography for thick resists and high resolution. J. Vac. Sci. Technol. B **5**(1), 120–123 (1987)
7. Lee, S., Cook, B.: PYRAMID – a hierarchical, rule-based approach toward proximity effect correction – part II: correction. IEEE Trans. Semicond. Manuf. **11**(1), 117–128 (1998)
8. Tennant, D.M., Bleir, A.R.: Electron beam lithography of nanostructures. In: Wiederrecht, G. (ed.) Handbook of Nanofabrication. Elsevier, Amsterdam (2009)
9. Harriott, L.R.: Scattering with angular limitation projection electron beam lithography for suboptical lithography. J. Vac. Sci. Technol. B **15**(6), 2130–2135 (1997)
10. Chang, T.H.P., Kern, D.P.: Arrayed miniature electron beam columns for high throughput sub-100 nm lithography. J. Vac. Sci. Technol. B **10**(6), 2743–2748 (1992)
11. Shimazu, N., Saito, K., Fujinami, M.: An approach to a high-throughput e-beam writer with a single-gun muliple-path system. Jpn. J. Appl. Phys. **35**(12B), 6689–6695 (1995)
12. Petric, P., Bevis, C., Carroll, A., Percy, H., Zywno, M., Standiford, K., Brodie, A., Bareket, N., Grella, L.: REBL: a novel approach to high speed maskless electron beam direct write lithography. J. Vac. Sci. Technol. B **B27**(1), 161 (2009)
13. Photo by D. Carr and H. Craighead, Cornell Press Release, July 1997.
14. Sekaric, L.: The high and low notes of the Universe, Physics News Update, Number 659 (2003)
15. Ilic, B., Craighead, H.G., Krylov, S., Senaratne, W., Ober, C., Neuzil, P.: Attogram detection using nanoelectro-mechanical oscillators. J. Appl. Phys. **95**, 3694 (2004)

Electron Beam Physical Vapor Deposition (EBPVD)

► Physical Vapor Deposition

Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells

Alexandra Portera and Eva McGuirea
Department of Materials, Imperial College London, London, UK

Synonyms

FIB-SEM; SEM; TEM

Definition

The application of ion beam and transmission electron microscopy techniques to image, analyze, and study the 3-D structure and chemistry of nanomaterials inside cells.

Overview

Direct imaging of the structure and chemical analysis of tissues, cells, and biomaterials at the nanoscale is critical for understanding a wide variety nanoscale processes in biological systems. Whether in comprehending the role of amyloid fibrils in Alzheimer's disease or the toxicity of nanomaterials, there is a significant need for such studies to develop the potential for new forms of treatment and to advance the design of bio- and nanomaterials [1–9]. Analytical electron microscopy is an indispensable technique for characterization of the material-biology interface because information can be obtained with very high spatial and energy resolution. However, this field remains wide open due to technical difficulties with sample preparation, particularly of hard-soft interfaces, and in achieving adequate spatial resolution and contrast between individual components within the cells and tissues while preserving them in their near native state and maintaining adequate stability of samples under the electron beam. Traditional preparation methods in electron microscopy use chemically intrusive stains and fixatives to preserve tissue and cell structure; however, these protocols will both distort the structure and modify the chemistry of the material. It is crucial to preserve the chemistry of the bio- or nanomaterial in the biological system as any slight modification in their chemistry or oxidation state caused during processing could make the results ambiguous and liable to misinterpretation which in turn could impact treatment strategies. Furthermore, the electron beam can damage the sample, altering the structure and changing the local chemistry. Methods to minimize this effect include the development of minimal dose methods as well as instruments that can operate at a range of accelerating voltages so that the experiments can be performed under optimum conditions. Once this has been achieved, the only option remaining is to cool the sample to reduce the damage rate. This entry aims to discuss development and application of analytical and 3-D microscopy techniques to analyze a range of bio-/nanomaterials interfaces at medium to high spatial resolution from ~1 μm to 1 nm using a range of techniques, namely dual-beam focused ion beam milling, energy-filtered TEM (EFTEM), high-angle annular dark-field scanning transmission electron microscopy (HAADF-STEM), and electron tomography. Traditional methods of sample preparation and using mass-thickness contrast for the imaging of biological samples will be discussed first as these still remain important techniques for both imaging and screening samples.

Sample Preparation

One of the greatest challenges for imaging nanoparticles within cells or biomaterials cell interfaces is in the sample preparation. When preparing samples for TEM, it is essential to induce minimal alteration to cell ultrastructure and also to the distribution and chemistry of nanoparticles within the cell. There are two main techniques available to prepare TEM samples at room temperature: fixation, dehydration, and embedding followed by ultramicrotomy; and focused ion beam milling (FIB). With ultramicrotomy, ultrathin resin-embedded sections containing the

cell–nanomaterial interface are mechanically sectioned using a diamond knife. Here the challenge lies in matching the hardness of the embedding resin with the hardness of the cell/nanomaterial to ensure that the interface does not damage during sectioning [1]. For this reason, choice of embedding resin is crucial. FIB is also an excellent technique for preparing and analyzing interfaces for electron microscopy [6, 9]. A clear advantage of FIB milling is that chemical processing and also mechanical damage can be obviated, thus reducing damage to the specimen prior to examination. The FIB process will be described later in this entry.

Staining

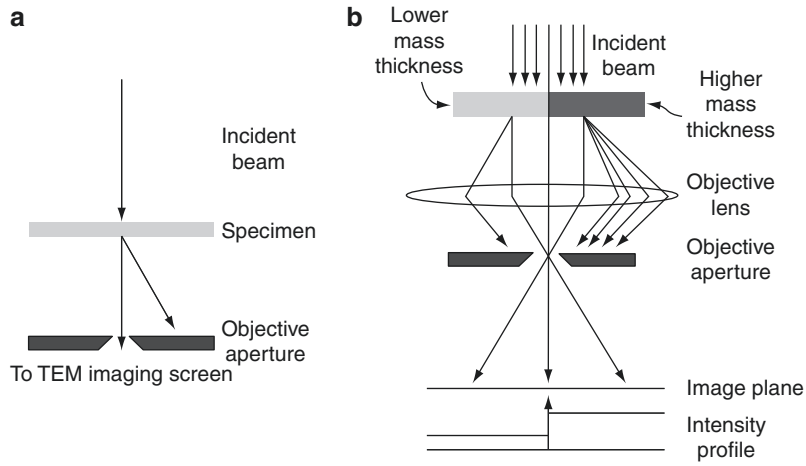
To increase contrast from cell organelles, heavy metal stains are often employed. Stains such as Os, Pb, and U leave heavy metals in specific regions of the structure. For example, osmium tetroxide is a strong oxidant that cross-links lipids; it therefore acts not only as a secondary fixative of membranes, but also increases contrast in TEM images because of the high atomic number of the osmium atoms. These stains can however, obscure visualization and characterization of low contrast nanostructures. To overcome this problem, stained and unstained sections have been compared to correlate distributions of nanoparticles within cell structure [4, 5].

Section Thickness

In ultramicrotomy, standard sample thicknesses for TEM sections can vary from 70 nm to several hundred nanometers. This results in the acquisition of a two-dimensional image containing three-dimensional information which can, in turn, result in the "blurring" of data or loss of information. For electron energy loss (EEL)-based techniques or high-resolution imaging, optimum contrast can be achieved by sectioning ultrathin sections of around 20 nm in thickness. By using these ultrathin sections, two-dimensional slices through cells are obtained, therefore minimizing the three-dimensional information and allowing for high-resolution imaging of nanoparticles within the cell. Contrast from cell organelles can also be significantly improved by sectioning at 20–40 nm or by leaving the section under the electron beam in the TEM to thin. The limitation with ultrathin sections, however, is durability of the section under the beam, thus, when energy or spatial resolution is not critical, thicker sample thickness (~70 nm) will provide sufficient contrast.

Rapid Freezing Techniques

Such as freeze substitution and freeze drying are also available to prepare cell–nanomaterial interfaces. These methods involve rapidly freezing the tissues or cells, typically in liquid propane or liquid nitrogen, either by drastically decreasing the temperature and/or by applying a high pressure. The cells are rapidly frozen to avoid ice crystal formation which would cause mechanical injury to the cells. After the cells/tissues have been frozen, these materials are slowly heated to allow the frozen water in the material to sublime from the solid phase to the gas phase. This process is called freeze drying. Sublimation of the material is required to avoid shrinkage, which would occur when rapidly heating the sample. If ice is not sublimed into the gas phase, the surface tension of the material will increase, and the material could collapse under this tension as the water leaves the liquid specimen. Freeze substitution is based on rapid freezing of tissues followed by solution ("substitution") of ice at temperatures well below 0 °C. For best morphological and histochemical preservation, substituting fluids should, in general, contain both chemical fixing agent and solvent for ice, e.g., 1 % solutions of osmium tetroxide in acetone or mercuric chloride in ethanol. Preservation of structure is generally poorer after substitution in solvent alone. Osmium tetroxide can also act to enhance the contrast acquired from the cell organelles. These techniques have many advantages over traditional room temperature preparation methods, particularly the preservation of structure, chemistry, and the accurate localization of many soluble and labile substances. Both these methods have been used to study the process of the formation of nanoparticles of amorphous calcium phosphate inside osteoblast and also neuronal cells. In these studies, cryo-fixation methods are critical to



Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells, Fig. 1 (**a**) The insertion of an objective aperture to select the direct beam removes most scattered electrons and may greatly increase image contrast. (**b**) The mechanism by which differences in mass and thickness create image contrast in a bright-field TEM image (Figure adapted from Ref. [11])

preserve both the chemistry of the labile calcium phosphate particles and the structure of the cell [10].

Bright-Field TEM Imaging and Mass-Thickness Contrast

Mass-thickness contrast is the contrast mechanism that is traditionally used for imaging biological samples. As electrons pass through the sample, they are scattered off-axis by elastic scattering events. Electrons that interact with the electron clouds of atoms in the sample are scattered through relatively small angles and may still contribute to the image, but electrons that are scattered due to interactions with an atomic nucleus are scattered through a higher angle and are less likely to contribute to the image generated. The cross section of scattering depends on the atomic number, Z, and on the density and thickness of the sample. In bright-field imaging, the direct beam is used to generate bright-field TEM images. The insertion of an objective aperture selects the angular range of electrons contributing to the image and may greatly improve image contrast as demonstrated in Fig. 1a. The mechanism by which contrast in a bright-field TEM image is formed from differences in mass and thickness is demonstrated in Fig. 1b. The areas of the sample that have higher mass thickness (ρt) produce areas of lower intensity in the resulting TEM image than areas of lower mass thickness. Mass-thickness contrast from cell organelles can be enhanced by staining the cells with heavy metals such as osmium tetroxide.

Analytical Transmission Electron Microscopy

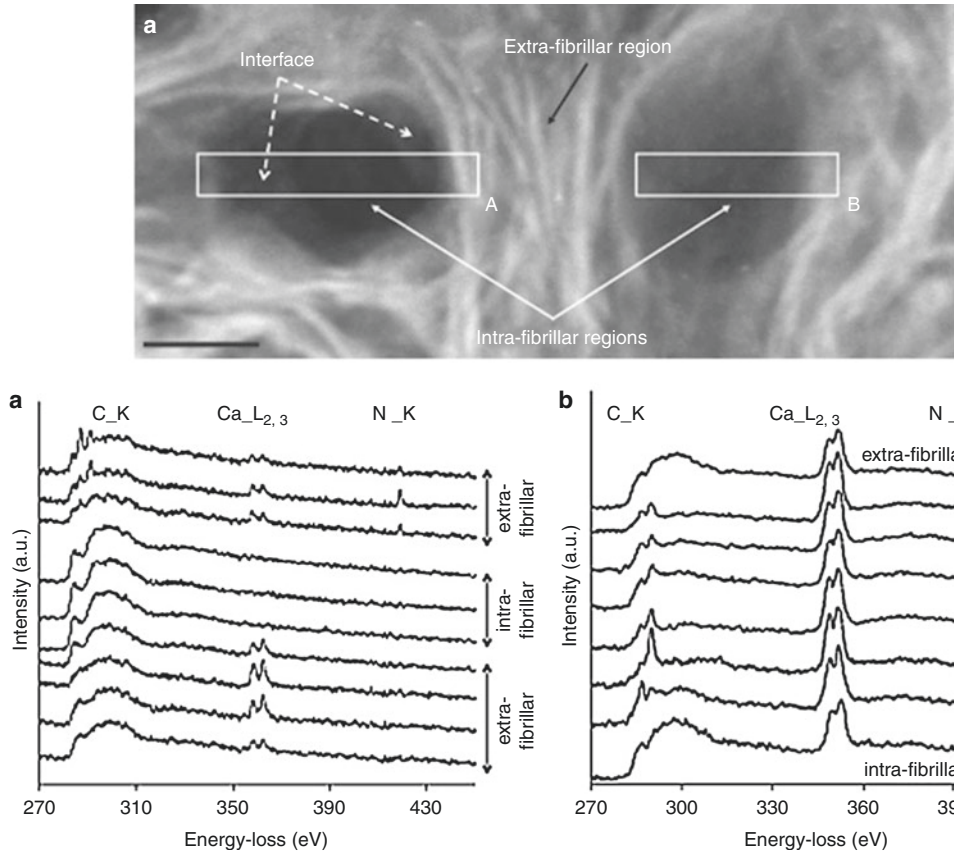
In a TEM, energy loss can either be measured in series, i.e., an image is acquired over a defined energy loss window, known as energy-filtered TEM (EFTEM); or in parallel, where an entire energy loss spectrum is acquired from a point in the specimen, known as electron energy loss spectroscopy (EELS). EELS is a technique where high-energy electrons incident on a sample scatter from, and therefore lose energy to, atoms in the sample. The energy lost in a particular scattering event is lost to an atom's electrons and only takes discrete values (quanta) which are different for every element. The electrons are transmitted through a sample, and then the energy that they lose is recorded in the form of a loss spectrum. The spectrum of each element is comprised of peaks or edges positioned at characteristic energy-losses. The shape and position of each edge also provides information about the bonding environment of the sample being analyzed. Therefore, EELS spectra provide detailed information about the elemental composition and bonding state within a sample with sub-nanometer spatial

resolution. EFTEM is a technique based on EELS which records images taken at different energies. The basis of this technique is that an energy-filter slit is used which only allows electrons through which have lost certain energies. By selecting different energies, elemental EFTEM maps can be produced. Using this technique, it is possible to acquire two-dimensional, elemental maps using electrons with an energy loss characteristic of a core level, interband transition or Plasmon resonance energy. With EFTEM, the EELS spectrum may be extracted by acquiring a series of energy-filtered images, using a narrow energy window (1–3 eV). Image processing software may then be employed to extract the EELS spectra over a given energy range.

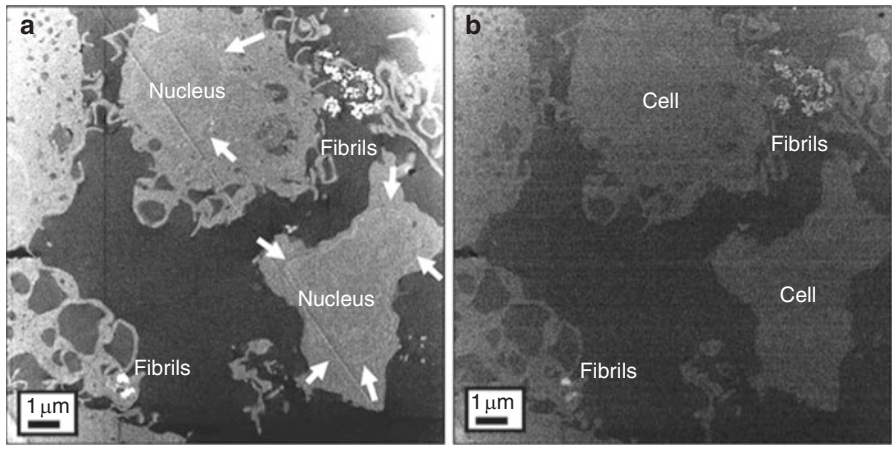
EFTEM is an excellent technique for nanobioscience applications because it can detect low atomic number elements (such as biological constituents) with high sensitivity. Using this technique, maps of the distributions of the elements in the sample can be resolved and, thus, the chemistry of the cell or presence of nanoparticles and their locations in the cells determined [3, 4]. For example, low-loss EFTEM enables clear differentiation between *unlabelled* C_{60}, SWNTs, and the cell cytoplasmic contents [3, 4]. TEM and EELS can be also applied to study intracellular free ion concentrations, the oxidation state of engineered nanoparticles in the cellular environment [12], mineral distributions in osteoporotic bone, or the formation of and the oxidation state of iron storage protein ferritin in the brains of patients suffering from Alzheimer's disease [13].

In scanning transmission electron microscopy (STEM), the incident electron beam is focused to a probe and scanned across the sample. Detectors placed at different post-specimen locations can collect different scattered or unscattered beams. The scattered (or unscattered) electrons are collected serially, and the image is displayed pixel by pixel. Combined with EELS, it can be used to map the chemistry or oxidation state of materials nanoscale resolution. In recent work, Jantou et al. have shown that STEM-EELS can provide new insights into the chemistry and bonding at the collagen-hydroxyapatite interface in ivory dentin and observed fine structure on the carbon K-edge spectra, suggesting that functional groups at the interface between the collagen and mineral phase can be identified [9]. This work potentially provides information on the modification of the surface amino acid groups by the inorganic phase [9]. Such approaches offer enormous potential for improved characterization of the chemistry of nanoparticles within tissues to advance knowledge of how these structures interact with cell organelles (Fig. 2).

HAADF-STEM is an imaging mode that is highly sensitive to the atomic number of the sample. HAADF-STEM relies on collecting electrons which have been scattered to high angles by Rutherford scattering. The HAADF detector has a large central aperture and collects only electrons that are scattered to relatively high angles (e.g., >30 mrad, the exact range of scattering angles is determined by the camera length and the geometry of the detector). The intensity of a HAADF-STEM image is therefore dependent on Z^n ($n \sim 2$). As the detector collection angles are reduced, the contribution of mass thickness (ρt) increases. Thus, it is possible for the user to control the kinds of electron/specimen interactions that contribute to an image (Fig. 3). HAADF-STEM has been combined with energy dispersive x-ray spectroscopy (EDX) to image Amyloid-beta aggregates inside cells using a selenium enhancement technique [7]. The rationale for this study is that a number of challenges arise when attempting to image the interactions between amyloid or other protein aggregates and cells or tissue. The most significant of these is a lack of contrast between the carbon-rich protein aggregate and the carbonaceous cellular environment, making the aggregates extremely difficult to distinguish and identify with confidence. A variety of strategies have previously been employed to overcome this, but these all involve the application of chemically intrusive labeling techniques such as immunogold labeling or tagging the proteins with fluorescent tags [7]. Instead, peptides were assembled from selenium-analogues of the sulfur-containing methionine peptides, and then used HAADF-STEM to detect the selenium-doped species selectively within the carbon-rich background of the cell. The presence of selenium was confirmed by the



Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells, Fig. 2 STEM image of a transverse section of ivory dentine. Fine structure on the carbon K-edge can be seen at the interface between the intra- and extra-fibrillar regions which can be attributed to the molecular amino acid signatures. P L2,3-, O K- (not shown), and Ca L2,3-edges can be used to investigate local variations in mineral chemistry [9]



Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells, Fig. 3 (**a**) HAADF-STEM image taken with inner and outer collection angles of 8.3 and 41.5 mrad, respectively. (**b**) HAADF-STEM image taken with inner and outer collection angles of 32.5 and 161.0 mrad, respectively. Bright contrast from the nulclear membrane is enhanced in (**a**). *Arrows* delineate the nuclear membrane

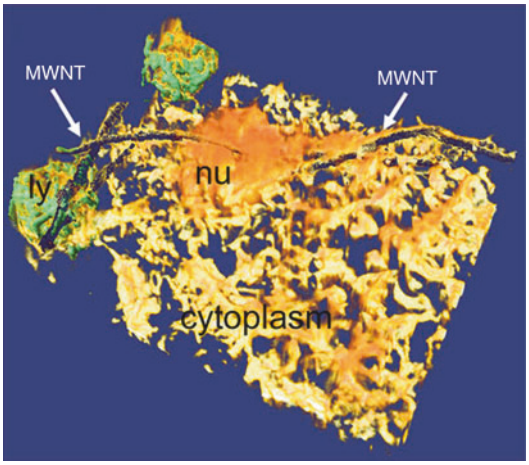
use of energy dispersive X-ray spectroscopy (EDX). The energies of the X-rays generated by the interactions between the electron beam and the atoms in the sample can be examined for peaks at energies characteristic of specific elements.

Electron Tomography

TEM and STEM imaging creates two-dimensional projections of a specimen. One disadvantage of this is that structural details from different depths within the specimen are superimposed on top of one another. Electron tomography (ET) involves collecting a series of images while tilting the specimen in the TEM around a single axis at regular intervals. It is impossible from this projection alone to determine the structure of the original object. For this reason, electron tomography is carried out – the process of taking a large number of two-dimensional electron micrographs of a specimen at different tilt angles and then reconstructing these to generate a volume that represents the original object.

The collection of a tomographic dataset produces a large number of two-dimensional projections of the region of interest which can be reconstructed into a three-dimensional volume that closely resembles the original object. The correlation between the three-dimensional reconstructed volume and the original object is limited by the data collected – the quality of the images and number of images collected – and the technique used to reconstruct the dataset.

Data collection for electron tomography involves tilting the specimen around an axis perpendicular to the electron beam and collecting images at successive angles. Specimen shift must be minimized during data acquisition, in particular in the direction parallel to the tilt axis, as only the area that is in every image can be used for the reconstruction of the data. Images are collected up to as high tilt angles as possible to minimize the size of the missing "wedge" in the dataset. It is impossible to take images of the specimen at tilt angles up to $\pm 90°$; therefore, there will always be some information missing from the dataset.

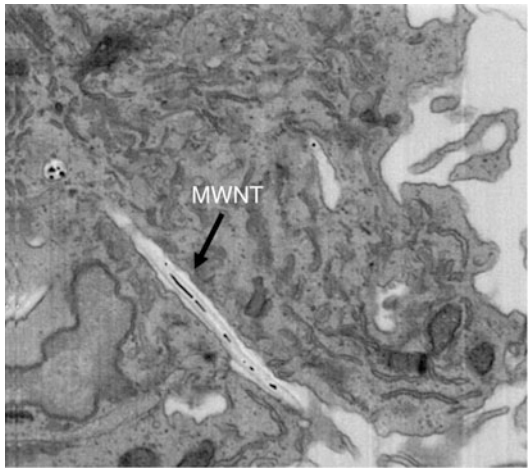


Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells, Fig. 4 HAADF-STEM tomogram of a necrotic macrophage cell illustrating MWNTs penetrating the nucleus (nu) [15]

A three-dimensional structure can then be reconstructed from the series of images collected to give high-resolution structural information about the material being analyzed. ET combined with EFTEM can also be used to yield three-dimensional compositional information [14]. Biological structures frequently need to be determined in all three dimensions because their chemical and physical properties are highly related to their topography. ET reconstructions are crucial to confirm that the nanostructures are localized within cell organelles and not just lying on the surface of the sample (Fig. 4).

Ion Beam Milling

A recent development in three-dimensional imaging is the "slice-and-view (S&V)" method in the dual-beam focused ion beam (DB-FIB). The DB-FIB combines an ultrahigh-resolution field emission gun scanning electron microscope (FEG-SEM) with a precise FIB system. In this system, the SEM uses electrons as a source for imaging and monitoring the milling process, whereas the FIB uses gallium positive ions as a source to ablate the surface of a specimen via the sputtering of substrate. This unique FIB-SEM combination allows the investigator to image the



Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells, Fig. 5 FEG-SEM image of a MWNT internalized inside a phagosome in a macrophage cell (Image taken from an osmicated cell in backscattered electron mode)

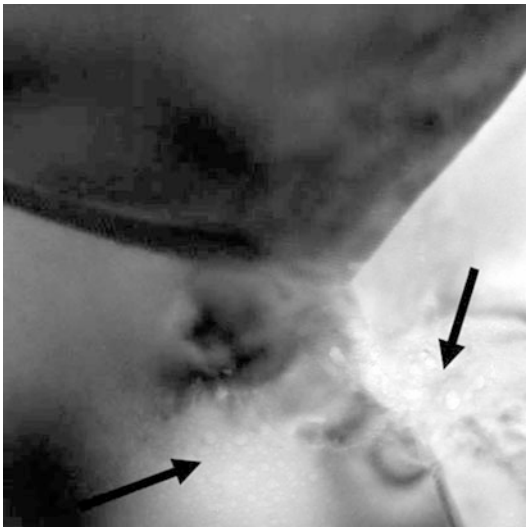
architecture of selected specimens by SEM and FIB in parallel. The focused ion beam can be used to mill away a precise volume to reveal the internal surfaces at a designated site in the specimen. The three-dimensional morphology of the sample can then be built up by milling and recording successive image slices (Fig. 5), the entire three-dimensional volume of the sample can then be reconstructed from the serial images to give information about the architecture of the sample. Milling can be performed with a precision of ~10 nm depth and up to 100 μm width allowing for analysis, in three dimensions, of the structure of materials from 0.1 mm to <100 nm. Approximately 500 serial images from a sample can be generated automatically overnight; therefore, both quantitative and qualitative information can be acquired. This technique has successfully been used to study the ultrastructure of mouse cortex tissue in three dimensions [16] and has great potential for mapping the distributions of nanoparticles inside tissues which can otherwise be analogous to the problem of "finding a needle in a haystack." The FEG-SEM method can also be used to prepare site specific lamella regions in the FIB for subsequent analysis in the TEM. More recently, this method has shown to be highly promising for preparing and studying frozen hydrated cells and tissue-material interfaces [17–19]. The advantage here is that this method involves no exposure to toxic or aggressive staining agents and the process induces little or no mechanical damage as would be experienced during cryo-ultramicrotomy.

Electron Beam Damage

Electron beam damage can be the limiting factor in the TEM study of biological materials. Many of the electron/specimen interactions that produce useful signals in the TEM result in damage of the specimen that can lead to structural and chemical changes in the sample [11]. This means that the region of the specimen under examination is no longer representative of the original sample and interpretation of the images and spectra is more difficult. Damage can result from both elastic and inelastic scattering of electrons. There are two main mechanisms of electron beam damage in the TEM: radiolysis and knock-on damage.

Radiolysis or ionization damage results in the breaking up of chemical bonds in the sample [11]. Organic solids, such as polymers and biological materials, are most prone to radiolysis [11]. Incident electrons can break down the main chain of a molecule, resulting in the loss of mass, or can break off side chains which may in turn react with other molecules in the sample and cross-link to form a new structure. Loss of mass results in a thinning of the sample or even the drilling of holes in the sample. Crystalline samples can lose their crystallinity due to ionization damage. The extent of damage due to radiolysis can be minimized by increasing the accelerating voltage as this decreases the interaction cross section. However, this leads to a reduction in image contrast, as less electron scattering takes place on interaction with the sample, and can result in an increase in knock-on damage.

Knock-on or atomic displacement damage occurs when electrons in the beam transfer enough energy to atoms in the sample to knock them out of their atomic sites leading to the creation of vacancies in the specimen [11].



Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells, Fig. 6 Void formation (indicated by *black arrows*) in an hydroxyaptite ceramic implant generated by knock-on damage to the sample under the electron beam

This damage mechanism will generate artifacts, such as amorphization of carbon nanotubes or void formation in calcium phosphate ceramics (Fig. 6). Electrons in the beam must possess enough kinetic energy to knock a specimen atom out of its site; this energy depends on the type of sample and is known as the threshold energy. This type of damage therefore increases with increasing accelerating voltage. Knock-on damage is not expected at an acceleration voltage of 100 kV (unless hydrogen atoms are present), but it is anticipated at higher voltages [20]. Cooling the specimen will not reduce the effects of knock-on damage. Keeping the energy of the electron beam low is the only way to avoid this type of damage. If energies above the threshold energy must be used, the total electron dose should be minimized by minimizing the time the electron beam is incident on the sample.

Heat is also a major cause of damage in the TEM. Much of the energy that is transferred to the sample when beam electrons collide with specimen electrons ends up as heat within the specimen [20]. Organic specimens are particularly susceptible to heat damage as they have low thermal conductivity, and so the region of interest cannot dissipate the thermal energy it gains under the beam. Heat damage can be minimized by using a higher accelerating voltage as this minimizes the interaction cross section.

Future Research Directions

Rapid developments are currently taking place in the field of electron microscopy, both to increase energy and spatial resolution, and also to increase automation of the equipment so that data can become significantly more quantitative and so it can be obtained with less work by the user, a feature that biological electron microscopy has traditionally suffered from. Exciting developments are occurring in the area of dual-beam FIB milling such that it will now be possible to correlate across length scales to link macroscale processes to structure and chemical changes occurring at the nanoscale. These methods will make major advancements in linking the physicochemical properties of nanoparticles to their effect on tissue structure and more generally will guide our understanding of tissue pathologies such as neurodegenerative diseases.

Acknowledgments The authors thank the contributions of Prof. David McComb, Imperial College, and Dr. Ben Lich at the FEI company, Eindhoven.

Cross-References

► Carbon Nanotubes
► Nanoparticle Cytotoxicity

References

1. Porter, A.E., Patel, N., Skepper, J.N., Best, S.M., Bonfield, W.: Comparison of in vivo dissolution processes in hydroxyapatite and silicon-substituted hydroxyapatite bioceramics. Biomaterials **24**, 4609–4620 (2003)
2. Porter, A.E., Gass, M.H., Muller, K., Skepper, J., Midgley, P., Welland, M.: Uptake of C_{60} by human monocyte macrophages, its localization and implications for toxicity: studied by high resolution electron microscopy and electron tomography. Acta Biomater. **2**(4), 409–419 (2006)

3. Porter, A.E., Gass, M.H., Muller, K., Skepper, J., Midgley, P., Welland, M.: Visualizing the uptake of C_{60} to the cytoplasm and nucleus of human monocyte-derived macrophage cells using energy-filtered transmission electron microscopy and electron tomography. Environ. Sci. Technol. **41**(8), 3012–3017 (2007)
4. Porter, A.E., Gass, M.H., Muller, K., Skepper, J., Midgley, P., Welland, M.: Direct imaging of single-walled carbon nanotubes in cells. Nat. Nanotechnol. **2**(11), 713–717 (2007)
5. Uchida, M., Willits, D.A., Muller, K., Willis, A.F., Jackiw, L., Jutila, M., Young, M.J., Porter, A.E., Douglas, T.: Intracellular distribution of macrophage targeting ferritin-iron oxide nanocomposite. Adv. Mater. **21**(4), 458–462 (2009)
6. Porter, A.E., Nalla, R.K., Minor, A., Jinshek, J.R., Kisielowski, C., Radmilovic, V., Kinney, J.H., Ritchie, R.O.: Changes in mineralization with aging-induced transparency in human root dentin. Biomaterials **26**(36), 7650–7660 (2005)
7. Porter, A.E., Knowles, T.P., Muller, K., Meehan, S., McGuire, E., Skepper, J., Welland, M.E., Dobson, C. M.: Imaging amyloid fibrils within cells using a Se-labelling strategy. J. Mol. Biol. **392**(4), 868–871 (2009)
8. Gentleman, E., Swain, R.J., Evans, N.D., Boonrungsiman, S., Jell, G., Ball, M.D., Shean, T.A., Oyen, M.L., Porter, A., Stevens, M.M.: Comparative materials differences revealed in engineered bone as a function of cell-specific differentiation. Nat. Mater. **8**, 763–770 (2009)
9. Jantou, V., Horton, M.H., McComb, D.W.: The nano-morphological relationships between apatite crystals and collagen fibrils in ivory dentine. Biomaterials **31**(19), 5275–5286 (2010)
10. Aronova, M.A., Kim, Y.C., Pivovarova, N.B., Andrews, S.B., Leapman, R.D.: Quantitative EFTEM mapping of near physiological calcium concentrations in biological specimens. Ultramicroscopy **109**, 201–212 (2009)
11. Williams, D.B., Carter, C.B.: Transmission Electron Microscopy: A Textbook for Materials Science. Springer, New York/London (2009)
12. Bhabra, G., Sood, A., Fisher, B., Cartwright, L., Saunders, M., Evans, W.H., Surprenant, A., Lopez-Castejon, G., Mann, S., Davis, S.A., Hails, L.A., Ingham, E., Verkade, P., Lane, J., Heesom, K., Newson, R., Case, C.P.: Nanoparticles can cause DNA damage across a cellular barrier. Nat. Nanotechnol. **4**, 876–883 (2009)
13. Collingwood, J.F., Chong, R.K.K., Kasama, T., Cervera-Gontard, L., Dunin-Borkowski, R.E., Perry, G., P'osfaig, M., Siedlake, S.L., Simpson, E.T., Smith, M.A., Dobson, J.: Three-dimensional tomographic imaging and characterization of iron compounds within Alzheimer's plaque core material. J. Alzheimers Dis. **14**, 235–245 (2008)
14. Gass, M.H., Koziol, K.K.K., Windle, A.H., Midgley, P.A.: Four-dimensional spectral tomography of carbonaceous nanocomposites. Nano Lett. **6**(3), 376–379 (2006)
15. Cheng, C., Muller, K., Koziol, K., Skepper, J.N., Midgley, P.A., Welland, M.E., Porter, A.E.: Toxicity and imaging of multi-walled carbon nanotubes in human macrophage cells. Biomaterials **30**(25), 4152–4160 (2009)
16. Knott, G., Marchman, H., Wall, D., Lich, B.: Serial section scanning electron microscopy of adult brain tissue using focussed ion beam milling. J. Neurosci. **28**(12), 2959–2964 (2008)
17. Marko, M., Hsieh, C., Scalek, R., Frank, J., Mannela, C.: Focused-ion-beam thinning of frozen hydrated biological specimens for cryo-electron microscopy. Nat. Methods **4**, 215–217 (2007)
18. Rigort, A., Bauerlien, F.J.B., Leis, A., Gruska, M., Hoffmann, C., Laugks, T., Bohm, U., Eibauer, M., Gnaegi, H., Baumeister, W.: Micromaching tools and correlative approaches for cellular cryo-electron tomography. J. Struct. Biol. **172**(2), 169–179 (2010)
19. Hayles, M.F., de Winter, D.A.M., Schneijdenberg, C. T.W.M., Meeldijk, J.D., Lucken, U., Persoon, H., de Water, J., de Jong, F., Humbel, B.M., Verkleij, A.J.: The making of frozen-hydrated, vitreous lamellas from cells for cryo-electron microscopy. J. Struct. Biol. **172**(2), 180–190 (2010)
20. Egerton, R.F., Li, P., Malac, M.: Radiation damage in the TEM and SEM. Micron **35**(6), 399–409 (2004)

Electron Transfer

► Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Electron Transport in Carbon Nanotubes and Graphene Nanoribbons

► Electronic Transport in Carbon Nanomaterials

Electron-Beam-Induced Chemical Vapor Deposition

► Electron-Beam-Induced Deposition

Electron-Beam-Induced Decomposition and Growth

► Electron-Beam-Induced Deposition

Electron-Beam-Induced Deposition

Guoqiang Xie
Institute for Materials Research, Tohoku University, Sendai, Japan

Synonyms

Electron-beam-induced chemical vapor deposition; Electron-beam-induced decomposition and growth

Definition

Electron-beam-induced deposition is a process for growing patterns with different sizes, shapes, and materials in the nanometer or submicron scale. In this process, an energetic electron beam in a vacuum chamber is focused on a substrate surface on which precursor molecules, containing the element to be deposited (e.g., organometallic compound or hydrocarbon), are adsorbed. As a result of complex beam-induced surface reactions, the precursor molecules adsorbed in and near to the irradiated area are dissociated into nonvolatile and volatile parts by the energetic electrons. The nonvolatile materials are deposited on the surface of the substrate to form a deposit, while the volatile parts are pumped away by the vacuum system. The advantage of electron-beam-induced deposition over conventional lithography methods is that two- and even three-dimensional (3D) structures are patterned and deposited simultaneously, making it a fast, one-step technique. Due to the controllability of the electron beam, the fabrication of the position-, size-, and morphology-controllable nanostructures can be easily achieved.

Overview

Originally, electron-beam-induced deposition was known as contamination growth in electron microscopy. Broers et al. [1] used contamination growth patterns as an etching mask to define 8-nm-wide metal lines. In the last decades, electron-beam-induced deposition for maskless nanofabrication has gained more importance as a tool, and it has been studied by many researchers because of its higher resolution due to the smaller beam spot. At an early stage, electron-beam-induced deposition has mainly been performed in scanning electron microscopes (SEMs) for various gases and conditions, and the technique was found to be very successful for a number of applications, making variously shaped nanostructures from metals such as W, Au, and Pt [2, 3]. However, the smallest structures fabricated in scanning electron microscopes have a typical width of 15–20 nm, no matter how small the primary electron beam is [4]. This is usually attributed to the fact that the distribution of secondary electrons, rather than primary electrons, determines the size of the deposits. The profile of the secondary electrons exhibits some spread, particularly for the energy range of scanning electron microscopes. Even for primary electrons of zero diameter, the profile of secondary electrons has a range of more than 15 nm [5]. Hence, it was believed that the resolution limit of electron-beam-induced deposition was at most 15 nm for the fabrication in scanning electron microscopes. Recently, several groups have tried to minimize the secondary electron effect by using transmission electron microscopes (TEMs) and scanning transmission electron microscopes (STEMs). Transmission electron microscopes utilize electrons of higher primary energy than scanning electron microscopes and can produce more focused beams with less secondary electron emission. Mitsuishi et al. [6] fabricated nanodots of 3.5 nm in diameter using a 200-keV field emission scanning transmission electron microscope. They found that the dot size does not depend on the substrate thickness; rather, the deposition period and speed are strongly dependent on the intensity of the electron beam. Tanaka et al. [7] have also fabricated nanodots containing W by electron-beam-induced

deposition in a 200-keV ultrahigh vacuum field emission transmission electron microscope (UHV-FE-TEM). The size of the dots can be controlled by changing the pressure in the chamber. The smallest particle size was about 1.5 nm in diameter.

Physics of Material Supply and Deposition

Aiming at a rate optimization in the electron-beam-induced deposition, one has to know how the deposition rate depends on the various relevant process parameters. Some theoretical analyses about the deposition process have been developed by Scheuer et al. [8, 9] for flood-beam systems and adapted for scanned-beam systems by Petzold and Heard [10]. Following the theoretical approach of Scheuer [8], it is assumed that only those molecules adsorbed at the substrate surface undergo dissociation induced by electron beam irradiation. Thus, only a monolayer of adsorbed molecules can contribute to the layer growth. Figure 1 illustrates the different processes on the surface [8]. The density (N) of adsorbed molecules on the substrate surface varied in time (t) is given by the following equation:

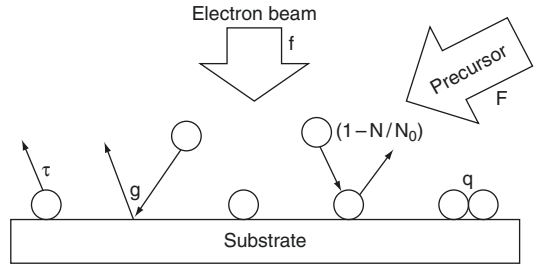
$$\frac{dN}{dt} = g \cdot F\left(1 - \frac{N}{N_0}\right) - \frac{N}{\tau} - q \cdot N \cdot f \quad (1)$$

where f is the electron flux density, q is the cross section for dissociation of the adsorbed molecules under electron beam irradiation, F is the molecular flux density arriving on the substrate, g is the sticking coefficient, N_0 is the molecule density in a monolayer, and τ is the mean lifetime of an adsorbed molecule.

If $dN/dt = 0$, equilibrium is obtained, resulting in an equilibrium density (N_E):

$$N_E = N_0 \cdot \frac{(g \cdot F/N_0)}{(g \cdot F/N_0) + 1/\tau + q \cdot f} \quad (2)$$

This means that $N_E \leq N_0$ for all supply and irradiation conditions, i.e., the maximum surface coverage at any time is one monolayer (ML). The growth rate (R) of the layer is given as follows:



Electron-Beam-Induced Deposition, Fig. 1 Illustration of surface interaction of the deposition process

$$R = v \cdot N \cdot q \cdot f \quad (3)$$

where v is the volume occupied by a dissociated molecule or its fractions. In the equilibrium case, $N = N_E$. From Eqs. 1, 2, 3, the steady-state growth equation is derived:

$$R = v \cdot N_0 \cdot \frac{(g \cdot F/N_0) \cdot q \cdot f}{(g \cdot F/N_0) + 1/\tau + q \cdot f} \quad (4)$$

In practice, the influence of the finite lifetime (τ) of the molecules can be made negligible, e.g., by cooling the sample. It is then worthwhile to note a few properties of the above equation. Obviously, the growth rate increases when the electron flux (f) and/or the molecular flux (F) is increased. In practice, these parameters are limited, e.g., by electron optical constraints, the vapor supply and differential pumping system, or simply the fact that only a limited vapor pressure can be tolerated above the sample before significant electron scattering starts. Under these conditions, the maximum rate obtainable for any given electron flux is

$$R_{\max} = v \cdot N_0 \cdot q \cdot f \quad (5)$$

and choosing

$$q \cdot f = g \cdot F/N_0 \quad (6)$$

yields 50 % of the maximum obtainable growth rate at supply F. This condition can be considered a reasonable compromise of parameters for efficient use of both electron and molecule supply. Increasing either flux to infinity while keeping the other one constant only increases the growth rate by a factor of 2, while decreasing either flux by

one order of magnitude reduces the growth rate by more than a factor of 5.

Key Research Findings

High-quality, intense, and focused electron beams have been used for the fabrication of nanomaterials with controlled position, size, and morphology. Electron-beam-induced deposition was used as an advanced method to fabricate nanostructures in situ by a maskless process. The reaction of precursor gas molecules with electrons near the material surface causes the formation of a solid deposit whose size and shape depend on the substrate materials and the profile of the electrons. Using the electron-beam-induced deposition technique, a variety of nanostructures, such as nanodots, nanowires, nanotubes, nanopatterns, two- or three-dimensional nano-objects, and so on, have been fabricated [2, 11–13].

When precursor gases containing tungsten (W), iron (Fe), and other metals are introduced into the microscope, noble-metal nanostructures can be fabricated in situ by electron-beam-induced deposition. The electron beam controls both the nucleation and growth processes. For the substrate materials, conventionally two kinds of materials, namely, conductive and insulator materials, are used. When the conductive materials, such as silicon, were used as a substrate, two- or three-dimensional nanostructures of W, Pt, Fe compounds, and so on, were fabricated by the electron-beam-induced deposition with the beam-scanning technique [2, 11]. This process is effective for the direct fabrication of nanosized structures by taking advantage of its fine probe of 1–10 nm in diameter. The effect of a vacuum on the size and growth process of the fabricated nanodots containing W was investigated using an ultrahigh vacuum field emission transmission electron microscope (UHV-FE-TEM). The size of the dots can be controlled by changing the dose of electrons and the partial pressure of the precursor. The smallest particle size obtained was about 1.5 nm in diameter at 1.5×10^{-6} Pa pressure, a probe size of 1 nm, and with 5 s irradiation [7]. Furthermore, nanodots, nanorods, and freestanding square frames were also fabricated by the electron-beam-induced deposition in an ultrahigh vacuum field emission scanning electron microscope (UHV-FE-SEM) using the precursor gas of iron carbonyl gas, Fe$(CO)_5$. The typical line width and length of the iron freestanding square frames were 30–50 nm and about 300 nm, respectively [14]. The iron freestanding square frame structures were magnetized by introducing the specimen into a magnetic field of about 1 T under the objective lens of the transmission electron microscope. After that, the objective lens was turned off, and the residual magnetic field was estimated to be less than 100 mT. The application of the magnetic field to iron three-dimensional nanostructures produced an electron hologram from the nanostructures. It was found that the magnetic field leaked from the nanostructure body, which appeared to act as a "nanomagnet" because of the square and ring shapes.

When insulator materials were used as a substrate, inhomogeneous deposition occurred because the charges accumulated on the insulator substrate upon electron irradiation during electron-beam-induced deposition. It has been reported that carbon dendrite-like or filament-like structures were grown on insulator substrates under electron beam irradiation in an electron microscope at a relatively poor vacuum condition [15]. The charging up of the surface of the insulator substrate due to electron beam irradiation and the high concentration of hydrocarbon molecules in the vacuum chamber are considered to be the formed reasons. By using insulator substrates such as Al_2O_3 [16] and SiO_2 [17], characteristic morphologies of nanostructures, such as arrays of nanowhiskers (or nanowires), arrays of nanodendrites, and fractal-like nanotree structures, were fabricated in transmission electron microscopes (TEMs) by the electron-beam-induced deposition. The typical size of the diameter of a nanowhisker, the tip of a nanodendrite, and the tip of a nanotree was about 3 nm and was almost completely independent of the size of the electron beam. The nanostructures with different morphologies can be obtained by controlling the intensity of the electron beam during electron-beam-induced deposition. The process of growth of the nanostructures is based on mechanisms in which the formation of the nanostructures is related to the nanoscale unevenness of the charge distribution on

the surface of the substrate, the movement of the charges to the convex surface of the substrate, and the accumulation of charges at the tip of the grown nanostructure. Furthermore, novel composite nanostructures of Pt nanoparticle/W nanodendrite or Au nanoparticle/W nanodendrite were also fabricated by the decoration of W nanodendrites with metallic elements [18]. The nanostructures have advanced features such as a large specific surface area, a freestanding structure on substrates, a typical size of several nanometers, and high purity. Because there are many possible combinations for the precursor and substrate, it is expected that various nanostructures may be fabricated on various substrates. Therefore, the nanostructures and their fabrication process may be applied in technology to realize various functional nanomaterials such as catalysts, sensor materials, and emitters.

Instead of using thin films of insulator materials, an anodic porous alumina membrane with the ordered nanohole arrays was used as a substrate; ordered array of nanostructures can be fabricated by the selective deposition of the precursor molecules into the ordered nanohole arrays using the electron-beam-induced deposition process. As an example, a porous alumina membrane on an aluminum substrate was prepared by a two-step anodization process, and then the aluminum was removed by etching. The diameter of the holes was about 40 nm, and the interval between the centers of the holes was about 100 nm. The total thickness of the membrane was about 1000 nm, but there were electron transparent regions at the bottoms of the holes (the barrier-layer thickness of the membrane was about 50 nm [19]). Tungsten hexacarbonyl, $W(CO)_6$, was used as a precursor gas source. The gas was supplied through a nozzle placed near the membrane on the side of the pores. The electron beam simultaneously irradiated the membrane from its reverse side. The deposition was carried out at room temperature. The diameter of electron beam during irradiation was about 1900 nm. The irradiation period was 1 min. It was found that the size of the formed pattern was almost the same as that of the electron-beam-irradiated area. Also, the deposition was only conducted in the pores of the electron-beam-irradiated area. Almost all of the pores in the electron-beam-irradiated area were filled with the deposits, while there was no deposit on the places other than the pores. The as-deposited nanoparticles contained many nanocrystallites. The nanocrystallites were identified to be the equilibrium phase of bcc structure metal W at room temperature [20].

Future Directions for Research

The further development of electron-beam-induced deposition requires the application of state-of-the-art electron microscopy technology such as precise positioning of a nanometer-scale electron beam, a high current for the achievement of small structures with a much higher fabrication speed, and an ultrahigh vacuum environment for eliminating contamination of the deposits.

There are many possible combinations for precursor and substrate; therefore, various functional nanostructures or compound nanostructures can be produced by electron-beam-induced deposition. However, the growth conditions for each combination of precursor and substrate should be different, and it should be further investigated. The fundamental processes involved in the interaction of an electron beam with substrates, the adsorption of precursor molecules on a substrate, and the decomposition of molecules by energetic electrons during electron-beam-induced deposition should also be further clarified.

Cross-References

- ▶ Atomic Layer Deposition
- ▶ Chemical Vapor Deposition (CVD)
- ▶ Electron Beam Lithography (EBL)
- ▶ Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)
- ▶ Self-Assembly
- ▶ Self-Assembly of Nanostructures

References

1. Broers, A.N., Molzen, W.W., Cuomo, J.J., Wittels, N.D.: Electron-beam fabrication of 80-Å metal structures. Appl. Phys. Lett. **29**, 596–598 (1976)
2. Koops, H.W.P., Kretz, J., Rudolph, M., Weber, M., Dahm, G., Lee, K.L.: Characterization and application

of materials grown by electron-beam-induced deposition. Jpn. J. Appl. Phys. (Part 1) **33**, 7099–7107 (1994)
3. Koops, H.W.P., Schössler, C., Kaya, A., Weber, M.: Conductive dots, wires, and supertips for field electron emitters produced by electron-beam induced deposition on samples having increased temperature. J. Vac. Sci. Technol. B **14**, 4105–4109 (1996)
4. Kohlmann-von Platen, K.T., Chlebek, J., Weiss, M., Reimer, K., Oertel, H., Brünger, W.H.: Resolution limits in electron-beam induced tungsten deposition. J. Vac. Sci. Technol. B **11**, 2219–2223 (1993)
5. Silvis-Cividjian, N., Hagen, C.W., Leunissen, L.H.A., Kruit, P.: The role of secondary electrons in electron-beam-induced-deposition spatial resolution. Microelectron. Eng. **61–62**, 693–699 (2002)
6. Mitsuishi, K., Shimojo, M., Han, M., Furuya, K.: Electron-beam-induced deposition using a subnanometer-sized probe of high-energy electrons. Appl. Phys. Lett. **83**, 2064–2066 (2003)
7. Tanaka, M., Shimojo, M., Mitsuishi, K., Furuya, K.: The size dependence of the nano-dots formed by electron-beam-induced deposition on the partial pressure of the precursor. Appl. Phys. A. **78**, 543–546 (2004)
8. Scheuer, V., Koops, H., Tschudi, T.: Electron beam decomposition of carbonyls on silicon. Microelectron. Eng. **5**, 423–430 (1986)
9. Koops, H.W.P., Weiel, R., Kern, D.P., Baum, T.H.: High-resolution electron-beam induced deposition. J. Vac. Sci. Technol. B **6**, 477–481 (1988)
10. Petzold, H.C., Heard, P.J.: Ion-induced deposition for x-ray mask repair: rate optimization using a time-dependent model. J. Vac. Sci. Technol. B **9**, 2664–2669 (1991)
11. Furuya, K.: Nanofabrication by advanced electron microscopy using intense and focused beam. Sci. Technol. Adv. Mater. **9**, 014110 (2008)
12. Song, M., Furuya, K.: Fabrication and characterization of nanostructures on insulator substrates by electron-beam-induced deposition. Sci. Technol. Adv. Mater. **9**, 023002 (2008)
13. Dong, L., Arai, F., Fukuda, T.: Electron-beam-induced deposition with carbon nanotube emitters. Appl. Phys. Lett. **81**, 1919–1921 (2002)
14. Takeguchi, M., Shimojo, M., Furuya, K.: Fabrication of alpha-iron and iron carbide nanostructures by electron-beam induced chemical vapor deposition and postdeposition heat treatment. Jpn. J. Appl. Phys. **44**, 5631–5634 (2005)
15. Banhart, F.: Laplacian growth of amorphous carbon filaments in a non-diffusion-limited experiment. Phys. Rev. E **52**, 5156–5160 (1995)
16. Song, M., Mitsuishi, K., Tanaka, M., Takeguchi, M., Shimojo, M., Furuya, K.: Fabrication of self-standing nanowires, nanodendrites, and nanofractal-like trees on insulator substrates with an electron-beam-induced deposition. Appl. Phys. A **80**, 1431–1436 (2005)
17. Xie, G.Q., Song, M., Mitsuishi, K., Furuya, K.: Growth of tungsten nanodendrites on SiO_2 substrate using electron-beam-induced deposition. J. Nanosci. Nanotech. **5**, 615–619 (2005)
18. Xie, G.Q., Song, M., Furuya, K., Louzguine, D.V., Inoue, A.: Compound nanostructures formed by metal nanoparticles dispersed on nanodendrites grown on insulator substrates. Appl. Phys. Lett. **88**, 263120 (2006)
19. Xie, G.Q., Song, M., Mitsuishi, K., Furuya, K.: Selective tungsten deposition into ordered nanohole arrays of anodic porous alumina by electron-beam-induced deposition. Appl. Phys. A **79**, 1843–1846 (2004)
20. Xie, G.Q., Song, M., Mitsuishi, K., Furuya, K.: Fabrication of ordered array of tungsten nanoparticles on anodic porous alumina by electron-beam-induced selective deposition. J. Vac. Sci. Technol. B **22**, 2589–2593 (2004)

E

Electronic and Optical Properties Calculations

▶ Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches

Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches

Letizia Chiodo[1], Giacomo Giorgi[2,3] and Maurizia Palummo[4]
[1]Center for Life Nano Science CLNS@Sapienza, Fondazione Istituto Italiano di Tecnologia and European Theoretical Spectroscopy Facility (ETSF), Rome, Italy
[2]Department of Chemical System Engineering, School of Engineering, The University of Tokyo, Tokyo, Japan
[3]CREST–JST, Tokyo, Japan
[4]Dipartimento di Fisica, Universitá di Roma Tor Vergata, European Theoretical Spectroscopy Facility (ETSF), Rome, Italy

Synonyms

Electronic and optical properties calculations; First-principles approaches; GW and BSE approaches; Self-energy and excitonic calculations in many-body systems

Definition

Ab initio electronic and optical properties of nanostructures.

Introduction

The interaction between oxides and light represents the fundamental process at the base of photocatalysis and solar-to-energy conversion. TiO_2 is one of the most studied oxides as well as testified by a large number of publications [1–4] both at fundamental and applicative levels of research.

Low-dimensional, nanostructured TiO_2-based materials, such as nanocrystallites, nanofibers, nanosheets, and nanotubes, offer the undeniable advantage of an enhanced available surface area that improves their photochemical and photophysical activities, compared with the high-dimensional counterparts.

Accordingly, they are receiving increasing attention as potential candidates for improving device performances. Particular attention has been paid to nanostructures of the anatase phase due to its higher energetic stability at the nanoscale and also to the envisaged superior performances of this phase with respect to rutile in photovoltaics (PV) and photocatalysis [5].

Both from experimental and theoretical points of view, deep research efforts are spent to investigate the role of nanostructuring and the influence of defects and of extrinsic doping to improve chemical and physical properties with respect to pure material.

Since most of the applications involve photoexcitation and the subsequent creation, transport, and dissociation of excitons, it is clear that a microscopic knowledge of their properties is a fundamental prerequisite to rationalize experimental data and improve device performances.

First-principles excited state methods, based on many-body perturbation theory [6] applied on top of density functional theory calculations [7, 8], provide a reliable tool to identify, explain, and predict the specific behavior related to the electronic and optical properties of nanostructures that may result hidden in the experimental processes.

In particular, here, the relationship among structure and electrooptical behavior will be highlighted. Strong excitonic effects, which allow a correct interpretation of the photoexcited processes, are present in all the considered nanostructures.

Theoretical Foundation

Density functional theory (DFT) [9] in the Kohn-Sham formalism [10] is a very powerful computational approach to predict a wide range of chemical and physical properties of materials.

However, it is well known that this mean-field theory does not correctly describe the electronic excited state properties of materials. At variance, within the Green's functions formalism, by means of the so-called Hedin's GW method [11] applied on top of DFT calculations, the calculated quasiparticle band structures result in good agreement with the experiments for a wide range of materials [12, 13]. Furthermore, within the same formalism, the solution of the two quasi-particles Bethe-Salpeter equation (BSE) [6, 14] allows to calculate spectroscopic properties (i.e., absorption, reflectivity, EELS, etc.), including many-body interactions, like local fields and excitonic effects, which result extremely important in systems of reduced dimensionality as those discussed here. While the theoretical framework described in the following is quite general and has been applied to a plethora of systems, inorganic and organic, finite and periodic ones, here the main focus will be on its application to titanium-dioxide low-dimensional structures.

The Kohn-Sham Formulation of the Density Functional Theory

Within density functional theory, the electron charge density $n\left(\vec{r}\right)$, rather than the electron wave function, plays a central role. DFT has been introduced by Hohenberg and Kohn (HK) in 1964 [9]. They demonstrated that the

ground state (GS) energy of a system of N-interacting electrons in an external potential V_{ext} (**r**) is a functional of the electronic density which is minimized at the correct GS density of the system. Kohn and Sham [10], a year later, reformulated the variational problem of the HK functional energy minimization in terms of the solution of a set of one-particle self-consistent equations (Kohn-Sham equations):

$$\left[-\frac{1}{2}\nabla^2 + V_{eff}^{KS}\left(\vec{r}\right)\right] \cdot \phi_i\left(\vec{r}\right) = \varepsilon_i \phi_i\left(\vec{r}\right) \quad (1)$$

where $n\left(\vec{r}\right) = \sum_i^N |\phi_i\left(\vec{r}\right)|^2$ is the electron charge density and $V_{eff}^{KS}\left(\vec{r}\right) = V_{ext}\left(\vec{r}\right) + V_{\mathrm{H}}\left(\vec{r}\right) + V_{xc}\left(\vec{r}\right)$, where the three parts of effective potential are the external, the Hartree, and the exchange-correlation potential. The last term, explicitly defined as the functional derivative with respect to the density of the exchange-correlation energy $E_{xc}[n]$, formally includes all the many-body effects of the system. Most of the drawbacks of DFT reside in the approximation used for the exchange-correlation potential. The oldest is the so-called local density approximation (LDA), where a local dependence on the density of the homogenous electron gas exchange-correlation energy is used [15, 16]. The inclusion of a dependence on the density gradient as in the generalized gradient approximation (GGA) [17] provides a better description of the nonlocality of the exchange-correlation hole [18]. New families of exchange-correlation functionals (PBE0, B3LYP and HSE06, meta-GGAs and hyper-GGAs functionals [8, 19]) have also been introduced in the last decades, giving more refined results for material properties than LDA and GGAs but at higher computational cost. The results shown in the last sections are mainly obtained at the GGA level of DFT, by using the PBE exchange-correlation functional.

Nevertheless, DFT is formally a "ground state" theory, and, as there is not a rigorous justification to interpret the KS eigenvalues ε_i as energy levels of the system, it is not surprising that the calculated electronic band structures, especially for semiconductors and insulators, do not compare very well with experiments. Therefore, in the next sections, a rigorous method to describe the charged and neutral electronic excited states of an interacting N-electron system will be presented.

Electronic Properties Within the Quasi-Particle Formulation of the Green's Functions Theory

The physical quantity which allows to describe the propagation of an electron added or extracted from an N-interacting electrons system is the time-ordered single-particle Green's function within the second quantization formalism [20, 21]:

$$G(1,2) = -i < \Phi_N | T\left(\hat{\psi}_H(1)\hat{\psi}_H^{\dagger}(2)\right) | \Phi_N >. \quad (2)$$

Here $1 = (x, t)$ (where t is a time coordinate, while $x = \left(\sigma, \vec{r}\right)$ indicate the spin and spatial coordinate and a similar definition stands for 2); $|\Phi_N >$ is the GS of the N-interacting particles system; $\hat{\psi}_H$, $\hat{\psi}_H^{\dagger}$ are the field operators in the Heisenberg representation; and T is the Wick time-ordering operator.

The ground state total energy, the expectation value of any single-particle operator, and the charged excitations spectra of the system can be calculated when G is known [20, 21]. Clearly being $|\Phi_N >$ an unknown quantity, Eq. 2 does not provide a practical way to calculate G in an arbitrary N-interacting system.

Nevertheless within the many-body perturbation theory (MBPT), it can be shown that G is connected to the Green's function of the noninteracting system G_0 (corresponding to the solution of the Hartree Hamiltonian) by the so-called Dyson equation:

$$G(12) = G_0(1,2) + \int G_0(1,3)\Sigma(3,4)G(4,2)d3d4. \quad (3)$$

where a nonlocal, time-dependent, non-Hermitian operator, the *self-energy* Σ, is introduced. This can

be thought as the effective potential that the electron or hole added to the system feels due to the change in the charge distribution induced by its own presence.

Using the spectral representation of G, and assuming that its dominant contribution comes from well-defined complex poles ε_i^{QP}, the electronic excitations of the many-body system can be obtained solving the so-called quasi-particle (QP) equation [6, 20, 21]:

$$\left(-\frac{1}{2}\nabla^2 + V_{ext} + V_H\right)\Phi_i^{QP}(x) + \int \Sigma\left(x, x', \varepsilon_i^{QP}\right)\Phi_i^{QP}(x')dx' = \varepsilon_i^{QP}\Phi_i^{QP}(x). \tag{4}$$

Then compared to KS electrons, which are completely independent particles, quasi-particles can be seen as weakly interacting dressed electrons characterized by a finite lifetime which is related to the imaginary part of the complex eigenvalues ε_i^{QP} [6, 22, 23]. It is worth to notice that when $\Sigma = 0$, Eq. 4 reduces to the Hartree approach, when $\Sigma = iGv$ (with v the unscreened Coulomb potential) to the Hartree-Fock, and when $\Sigma = V_{xc}$ to the Kohn-Sham scheme.

Now within the many-body perturbation theory, a set of five integral equations, known as the *Hedin* equations, can be derived, providing a formal scheme to find a suitable expression for Σ [11]. Starting from the equation of motion of the field operator and from the expression of the many-body Hamiltonian in the second quantization formalism, the equation of motion for G can be derived, and it results to depend on the two-particle Greens function. In a similar way, the two-particle Greens function equation of motion turns out to depend on the three-particle Greens function and so on, yielding an infinite hierarchy of equations [24]. It has been shown that this chain of equations can be broken by introducing in the Hamiltonian an external source that couples to the density and that at the end of the derivation is set to zero [11, 25]. In this way, the two-particle Greens function G_2 can be written in terms of a functional derivative of G with respect to this external source, and then G_2 is formally removed from the equation of motion of G [14]. Introducing other characteristic quantities of the system like the screened Coulomb potential $W(12)$, the time-ordered polarization $P(12)$, and the vertex function $\Gamma(123)$, the close set of five *Hedin* equations can be derived (see, i.e., Ref. [6] for their full expression). The *Hedin* equations offer the possibility to analyze the physical meaning of the quantities involved: the polarization P describes the response of the system to the additional particle or hole and the creation of electron-hole couples, and it is related to the dielectric function ε by the following expression $\varepsilon(1,\ 2) = \delta(1,\ 2) - \int\ v(13)P(32)d3$; W is the dynamical screened coulomb potential, and Γ is the vertex function which contains the information that the hole and the electron interact and is formally determined by the change in the potential upon excitation. Interestingly they lead directly to the possibility of a perturbative expansion in terms of the screened Coulomb interaction W. This is very important and physically sound; in fact a direct expansion in terms of the bare Coulomb interaction would be unsuitable due to the strength of this kind of interaction, whereas in many-electrons systems, the Coulomb interaction is usually screened via the polarization of the system, and the effective potential, by which the particles interact, is much weaker.

Since these equations are not exactly solvable, Hedin proposed an iterative approach to their solution. Starting with $\Sigma = 0$, at the end of the first iteration, the following expressions are obtained:

$$G = G_0; \quad \Gamma(123) = \delta(12)\delta(13);$$
$$P(12) = iG(12)G(21);$$
$$\Sigma(12) = iG(12)W(2,1).$$

This leads to an expression for the polarization function where the electron and hole propagators are independent (generally known as the random phase approximation (RPA) [26]) and defines the so-called GW approach for the self-energy calculation. Nowadays this is the most used approximation within the many-body approaches for the calculation of the QP excitation energies. In practical calculations, as the KS equations of the DFT are a better starting approximation than Hartree,

because a certain degree of exchange and correlation is included in V_{xc}, both G and W are calculated starting from the DFT-KS eigenvalues and eigenfunctions. Moreover, since Eq. 1 is formally similar to Eq. 4, the self-energy corrections are evaluated considering the term $\left(\Sigma - V_{xc}^{KS}\right)$ as a perturbation term of the KS equations. Due to the fact that in many systems the KS eigenfunctions are very similar to the GW ones, one may often neglect the off-diagonal terms in the many-body Hamiltonian and calculate the QP energies solving the following equation:

$$\varepsilon_i^{QP} = \varepsilon_i^{\mathrm{KS}} + \left\langle \phi_i^{\mathrm{KS}} \middle| \Sigma\left(\varepsilon_i^{QP}\right) - V_{xc}^{KS} \middle| \phi_i^{\mathrm{KS}} \right\rangle. \quad (5)$$

A further simplification consists in the inclusion of the frequency dependence of the self-energy by a first-order Taylor expansion around $\varepsilon_i^{\mathrm{KS}}$ [27, 28]:

$$\left\langle \Sigma\left(\varepsilon_i^{QP}\right) \right\rangle = \left\langle \Sigma\left(\varepsilon_i^{\mathrm{KS}}\right) \right\rangle + \left(\varepsilon_i^{QP} - \varepsilon_i^{\mathrm{KS}}\right) \left\langle \frac{\partial \Sigma(\omega)}{\partial \omega} \middle|_{\omega=\varepsilon_i^{KS}} \right\rangle + O\left(\left(\varepsilon_i^{QP} - \varepsilon_i^{\mathrm{KS}}\right)^2\right) \quad (6)$$

This approach is generally known as the perturbative G_0W_0 method, and it is the scheme used to calculate the QP self-energy corrections to the KS energies for TiO_2 clusters and sheets described in section "Applications."

Optical Properties By the Solution of the Bethe-Salpeter Equation

Starting from the GW approximation for the self-energy and multiplying the vertex by $-iGG$, a four-point integral equation for a generalized reducible ${}^4\overline{P}(1234)$ can be obtained at the second iterative step of the *Hedin* equations:

$$\begin{aligned} {}^4\overline{P}(1,1';2,2') &= {}^4P_{IQP}(1,1';2,2') \\ &+ {}^4P_{IQP}(1,1';3,3')K(3,3';4,4'){}^4\overline{P}(4,4';2,2') \end{aligned} \quad (7)$$

where the repeated arguments are integrated over and the kernel K contains two contributions [29]:

$$\begin{aligned} K(1,1',2,2') = &-\delta(1,2)\delta(1',2')W(1,1') \\ &+ \delta(1,1')\delta(2,2')\overline{v}(1,2). \end{aligned} \quad (8)$$

The term $\overline{v}$ is the unscreened coulomb potential where the long-range $\left(\vec{G} = 0,\ \vec{q} \to 0\right)$ Fourier transform component is set as zero. When the spin is not explicitly treated, this term gets a factor of two for singlet excitons. The second term, $-W$, is the screened Coulomb attraction between electron and hole. As mentioned before, while it is obtained from the functional derivative of the self-energy, the first term (here $\overline{v}$ to obtain the macroscopic dielectric function instead of the microscopic one) comes formally from the functional derivative of the Hartree potential [14] which enters in the equation for the reducible polarization function.

Using the basis of the KS wave functions ϕ_{nk} the Bethe-Salpeter Equation becomes

$$\begin{aligned} \overline{P} &= \left[1 - P_{IQP}K\right]^{-1}\overline{P}_{IQP} \\ &= \left[H^{exc} - I\omega\right]^{-1}_{(n_1,n_2),(n_3,n_4)}\left(f_{n_4} - f_{n_3}\right) \end{aligned} \quad (9)$$

where an effective excitonic Hamiltonian is introduced, defined as

$$H^{exc}_{(n_1,n_2),(n_3,n_4)} = \left(E_{n_2} - E_{n_1}\right)\delta_{n_1,n_3}\delta_{n_2,n_4} + \left(f_{n_2} - f_{n_1}\right)\times \int d\vec{r}_1\, d\vec{r}_1'\, d\vec{r}_2\, d\vec{r}_2\, \phi_{n_1}\left(\vec{r}_1\right)\phi^*_{n_2}\left(\vec{r}_1'\right) K\left(\vec{r}_1,\ \vec{r}_1',\ \vec{r}_2,\ \vec{r}_2'\right)\phi^*_{n_3}\left(\vec{r}_2\right)\phi_{n_4}\left(\vec{r}_2'\right)$$

Then, using the spectral representation [6] for the inverse of a matrix, the interacting polarization can be obtained solving an effective eigenvalue problem:

$$\sum_{(n_3,n_4)} H^{exc}_{(n_1,n_2),(n_3,n_4)} A_\lambda^{(n_3,n_4)} = E_\lambda A_\lambda^{(n_1,n_2)}. \quad (10)$$

In particular it can be shown that for the calculation of the absorption spectra, (i) one needs to

include only transitions with positive frequency [6] (i.e., (n_1, n_2) and (n_3, n_4) are pairs made of one valence and one conduction band, respectively) and (ii) only transitions with the same **k** for the initial and final state can be included. Then the macroscopic dielectric function reads as

$$\varepsilon_M(\omega) = 1 + \lim_{\vec{q}\to 0}\left(\vec{q}\right)\sum_{\lambda}\frac{\left|\sum_{v,c;\vec{k}}\left\langle v,\ \vec{k}-\vec{q}\ \left|e^{-i\vec{q}\vec{r}}\right|c,\ \vec{k}\right\rangle A_{\lambda}^{\left(v,c,\vec{k}\right)}\right|^2}{(E_{\lambda}-\omega)}. \tag{11}$$

The meaning of this equation is clear: instead of a simple sum of independent-particle optical matrix elements, they are weighted by the excitonic eigenstates A_λ. This occurs because the electron-hole Hamiltonian (Eq. 10) is not diagonal due to the presence of the nondiagonal excitonic kernel. This formulation allows to interpret the optical spectra (like absorption, reflectivity, transmission, EELS, etc.) in terms of mixing of single-particle transitions. Moreover the excitation energies in the denominator of Eq. 11 are related to the excitonic eigenvalues E_λ which, especially in low-dimensional systems as those considered here, are smaller with respect to the difference between the quasi-particle energy of occupied and unoccupied states, due to large exciton binding energies.

Applications

As applications of first-principles GS and excited state methods to the class of oxide nanostructures, two examples are discussed in the following: zero-dimensional atomic clusters [30, 31] and two-dimensional nanosheets [32] of titanium dioxide.

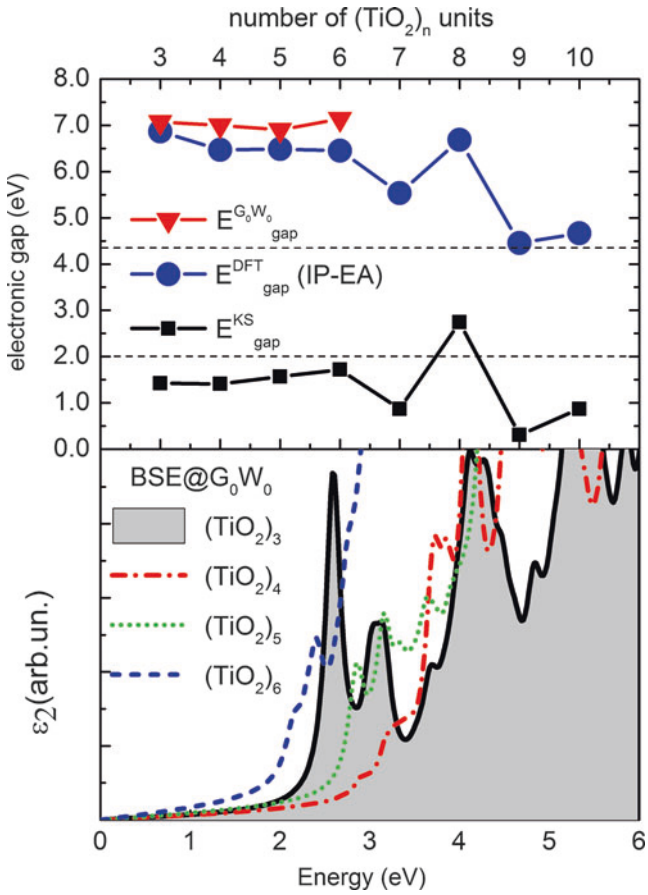
Numerical Details

The KS equations are solved here using a plane wave approach, as implemented in VASP [33, 34] and in the Quantum-ESPRESSO packages [35], while many-body calculations are performed using the plane wave Yambo code [36]. VASP code has been used to relax the structures using projected augmented wave (PAW) [37, 38] potentials for all the atomic species considered with a kinetic energy cutoff of 36.75 Ry and sampling of the Brillouin zone with different meshes according to the size of the system under investigation. Using the optimized geometrical structures, we then performed DFT calculations using Quantum-ESPRESSO package which is interfaced with the many-body code Yambo. Fictitious replica interactions occurring in the repeated cell approach are eliminated by using a vacuum thickness of more than 10 Å. Furthermore a cutoff in the coulomb potential is used both in the GW and BSE calculations, in the nonperiodic directions. Once the solution of the KS equations is obtained, it is possible to calculate several physical quantities of interest for the investigated systems; here the focus is on electronic band structures and optical spectra.

TiO_2-Based Atomic Clusters

Atomic anatase clusters of TiO_2 can be considered as the starting seeds for the synthesis of larger and more complex nanostructures for photocatalysis and photovoltaics. Due to the dimensions of the studied atomic clusters (up to 10 TiO_2 units, between 6 and 10 Å of size), giving finite-size quantization and shape-anisotropy, the electronic properties of these clusters strongly differ from the bulk.

In general, electronic and optical gap values present oscillations at growing size of the clusters, related to the symmetry of the electronic distribution (see Fig. 1, top panel). The largest KS gap (2.8 eV) is obtained for the $(TiO_2)_8$ system, which is the only cluster with a gap larger than the bulk anatase DFT value. The electronic gap, calculated as $E_{\text{gap}}^{\text{DFT}}$ (within ΔSCF, as IP-EA, with (IP) ionization potential and (EA) electron

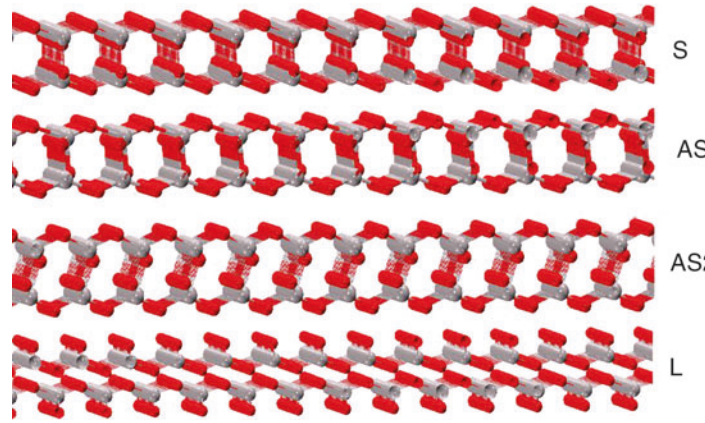


Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches, Fig. 1 *Top panel*: Electronic gap, evaluated at KS, DFT, and G_0W_0 level, for atomic clusters of increasing size. G_0W_0 converged results are reported only for smallest clusters ($(TiO_2)_n = 3,6$), due the computational cost of the approach. As reference values, *dashed lines* indicate the KS (1.8 eV) and G_0W_0 (3.8 eV) gap of the anatase bulk. *Bottom panel*: Imaginary part of the dielectric function (proportional to optical absorption) at BSE level, for the four smallest clusters $(TiO_2)_n = 3,6$

affinity) or via G_0W_0, has quite large values (in the range 4.5–7.0 eV), due to the strong quantum confinement of these atomic-size systems, and shows a quite good agreement with experimental data [30]. Because of the strong many-body effects, the electronic gaps of these inorganic molecules are also several eV (about 4 eV) larger of the Kohn-Sham [39] and optical gaps (Fig. 1). This is not surprising, as in localized systems the strong electron-hole interaction induces bound excitonic states, with e-h-binding energies much larger than in the bulk phase [30, 31, 40].

Indeed optical spectra of the smallest clusters, calculated at the many-body level, including the electron-hole interaction (i.e., by solving the BSE equation [6]), show a strong red shift from electronic to optical gap, with BSE gap larger than the KS-PBE gap and quite close to KS-PBE0 values [30].

Even more interesting information comes from the optical transitions analysis. In the quite asymmetric clusters, as $(TiO_2)_3$, the optical transitions described by BSE clearly show a charge transfer effect. With the hole localized on an O atom, the excitonic wave function is indeed localized on Ti atoms, denoting an internal charge transfer under photoexcitation. The first optical allowed transition involves the mixing of several states near the electronic gap and cannot be described simply as HOMO-LUMO transitions. In more symmetric clusters like $(TiO_2)_4$, any admixture of single-particle transitions is present in the lowest energy optical excitation. The first optically allowed, even if not very intense, transition, is at 2.9 eV, given by HOMO-2 → LUMO states. For $(TiO_2)_5$ and $(TiO_2)_6$ clusters, it has been found that an admixture of KS states contributes to optical transitions, but without net charge transfer effects.



Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches, Fig. 2 From *top* to *bottom: side view* of for symmetric anatase (*S*), asymmetric anatase (*AS1* and *AS2*), and lepidocrocite (*L*) isolated single-layer sheets

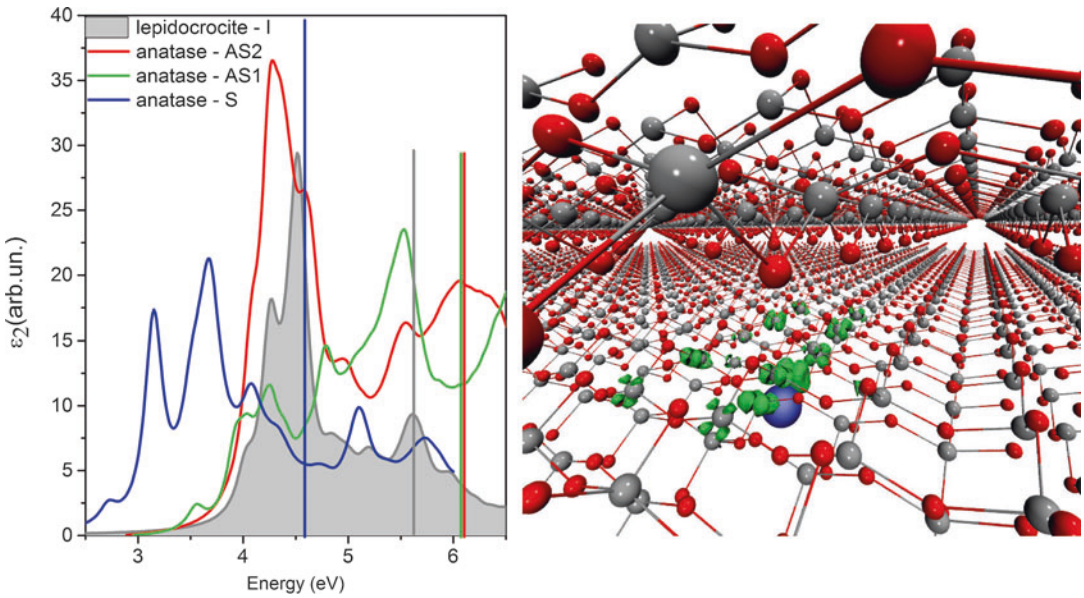
The different optical behavior is reflected on the different spectral shape (Fig. 1, bottom panel). The many-body approach thus offers a representation of optical properties which includes effects related to admixtures of KS states and introduces a refined description of charge transfer phenomena, which could affect excitonic distribution in TiO_2.

Overall, strong excitonic effects are observed, as the optical gaps of atomic clusters (defined from the first optically allowed transition) are in the range 2.0–3.0 eV, quite smaller than the electronic ones (about 7.0 eV). Well-defined optical peaks can be identified in the absorption spectra and may provide a useful tool for establishing the presence of well-determined structures in atomic clusters samples. The use of many-body methods becomes more and more important when the symmetry of the system is lowered, and the effects of point defects or other asymmetries affect the electronic properties of the system.

TiO_2-Based Nanosheets

Among all the possible morphological shapes of TiO_2-derived low-dimensional materials, the study of (001)-oriented nanosheets (NSs) is of broad interest for the enhanced optical activity when a high percentage of these facets are exposed [41]. A double nature characterizes such layered structures: one is the "bulk-cut" face obtained from anatase along [001], and the other is the so-called lepidocrocite [42, 43]. This latter phase can be thought as still derived from the former, but differently resulting from the gliding of the upper layer of the film over the lower along the Ti-O-Ti chain direction. A barrier-less path driven by the stress reduction characterizes the final bilayer. The importance of these NSs is also due to the fact that they are precursors of titania nanotubes (NTs) which in turn can be used for further assembling of nanostructured materials with different morphologies and applications [44]. In details, concerning the anatase (001)-oriented bulk-cut NSs, several structures (see Fig. 2) according to the nature of the Ti-O bond at the surface of the sheets are investigated: (1) a symmetric system (anatase-S) where the bonds Ti-O-Ti are identical (with the lattice cell parameters $a = b = 3.78$ Å) and (2) two asymmetric systems, i.e., with different Ti-O bonds at the surface, where the bulk-optimized lateral parameter is kept (anatase-AS1, still $a = b = 3.78$ Å) and also another structure where the total energy is minimized imposing the constraint of a square unit cell (anatase-AS2, $a = b = 3.51$ Å).

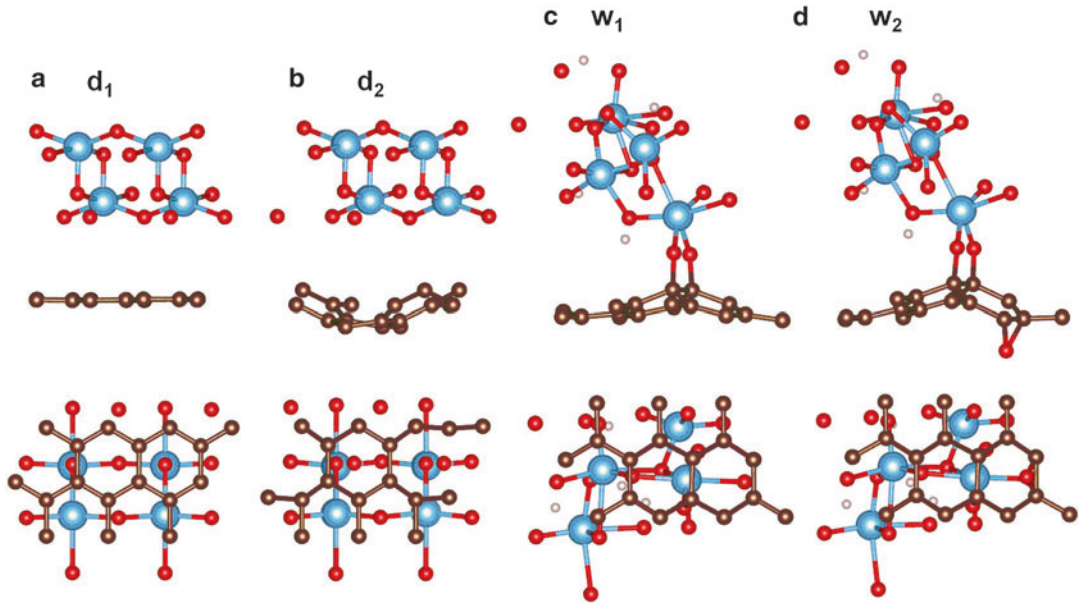
At variance, for the lepidocrocite structure, the system coordinates of Sato et al. [43] has been used as starting point and then accordingly reoptimized to find a structure (lepidocrocite-I, with $a = 3.76$ and $b = 3.03$ Å, respectively $a = b = 3.51$ Å; see Ref. [32] for more details). Ab initio calculations of the QP band structures reveal a change from indirect to direct bandgap when moving from anatase-S to lepidocrocite-I structure. This is an important aspect to be taken into account when evaluating the optical response of these nanostructures because the direct or indirect nature of the electronic and optical gap clearly affects the photoluminescence behavior. For all



Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches, Fig. 3 *Left panel*: imaginary part of the dielectric function calculated within the BSE approach, for anatase (S (*blue*), AS1 (*green*) and AS2 (*red*)) and lepidocrocite (*gray*)-isolated single-layer sheets. For each nanosheet, the vertical line indicates the corresponding minimum electronic gap calculated at the QP level of approximation within the G_0W_0 approach. *Right-panel*: plot of the excitonic wave function (*green isosurface*), fixing the hole near an oxygen bridge (*blue ball*) for an isolated double-layer of lepidocrocite

the computed isolated NSs structures, the electronic QP gaps range from 4.5 to 6 eV, while the corresponding optical gaps, calculated at the BSE level, are about 1.5 eV smaller. A systematic blue-shift of the optical gaps is observed going from the symmetric anatase (S) to the lepidocrocite (see the left panel of Fig. 3). The comparison with experimental absorption spectra (not shown here) results quite good in the case of lepidocrocite sheets both for single- and multiple-layer samples [32]. As for nanoclusters, the narrowing of the optical gap with respect to the corresponding electronic gap stems from the nanosheet-reduced dimensionality and from very low dielectric screening which induces a strong attractive electron-hole interaction. As a consequence, the optical spectra presented in Fig. 3 have a clear excitonic nature in the vis-UV region, presenting at the lower absorption edge more peaks than those obtained within the independent-particle treatment [32]. A sensitive oscillator strength redistribution is observed, indicating that, including the excitonic description, not only the energy associated to the transitions varies but also the nature itself of the states involved in the optical absorption changes due to the e-h interaction [32]. The left panel of Fig. 3 also shows a quite different optical response of the various layers. A direct comparison with experimental absorption data available from literature (not reported here) makes possible to establish an unambiguous relationship between atomic structures and optical properties of TiO_2-derived NSs [32]. An example of spatial exciton localization, described as the probability to find an electron when the hole is fixed in a given position, is represented in the right panel of Fig. 3 for the case of bilayer lepidocrocite sheet.

In the last years, the hybrid carbon-titania nanocomposites have been extensively investigated at experimental level for their promising applicability both in photocatalysis and photovoltaics [45–48]. Then the optical properties of graphene/TiO_2-based heterostructures represent a further very interesting field of analysis. The relevance of this class of materials in solar-to-energy

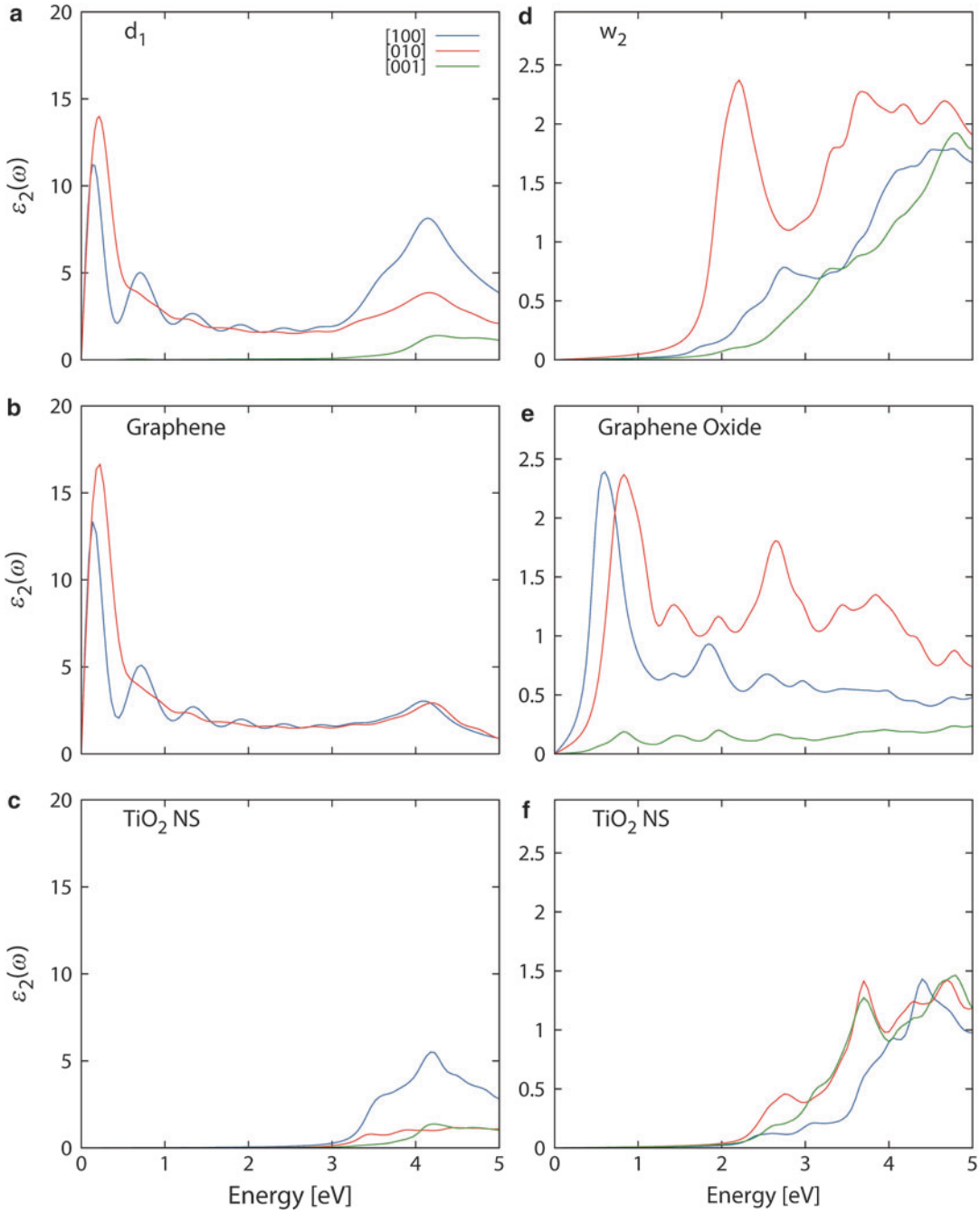


Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches, Fig. 4 *Side* and *top views* of graphene TiO_2 bilayer nanocomposite models. Structures (**a**) and (**b**) are the two dry models. (**c**, **d**) mimic the wet experimental environment (Reprinted with permission from Ref. [53]. Copyright 2014 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim)

conversion process stems from the combined nature of the heterostructure that embodies both the semiconducting properties of TiO_2 NS and the unique transport ones of graphene. Its large specific surface area, flexible structure, high transparency, and good electrical and thermal conductibility [49–52] make graphene the most promising materials in nanocomposite systems. It also exhibits excellent mobility of charge carriers, and it is expected to be able to slow the recombination of photogenerated electron-hole pairs contributing to the photocatalytic efficiency improvement of these hybrid materials. The almost complete lack of theoretical knowledge for such systems clearly associated with the different lattice parameter of the two subsystems has motivated to model nanohybrids formed by graphene- and TiO_2-based NSs. The simulations show that the presence of a direct bond (chemisorption) between the two-layered structures plays a crucial role in the electronic and optical properties of the final nanocomposite [53]. Figure 4 shows the atomic structure of the four models investigated in Ref. [53]. Two of them ($\mathbf{d_1}$ and $\mathbf{d_2}$) mimic the dry experimental assembling procedure [54], while the other two ($\mathbf{w_1}$ and $\mathbf{w_2}$) are modeled to reproduce the wet experimental environment. The presence of water makes the anatase NS hydroxylated at both surfaces. Furthermore the oxidized precursor of graphene, i.e., GO, is employed in assembling these wet heterostructures [55].

Structure optimization and calculation of electronic properties have been performed at the DFT level (see Ref. [53] for further details). The large size of the simulation cells did not allow the calculation of optical spectra by means of the GW/BSE approach. As demonstrated for the case of pure TiO_2-based nanosheets, the electronic bandgap opening ascribed to the GW correction is almost balanced by the inclusion of the excitonic effects at the BSE level of approximation. Then it can be fairly said that RPA and BSE optical spectra should be similar at a qualitative level.

Top panels of Fig. 5 report the imaginary part of the dielectric function of the physisorbed



Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches, Fig. 5 The imaginary part of dielectric function calculated as the independent level of approximation for the (**a**) $\mathbf{d_1}$ system, (**b**) graphene, and (**c**) TiO_2-NS. Same is reported for the (**d**) $\mathbf{w_2}$ hybrid, (**e**) graphene oxide, and (**f**) the hydrated TiO_2 NS. *Blue* and *red* are the two ordinary rays (x, y) and *green* is the extraordinary ray (z) (Reprinted with permission from Ref. [53]. Copyright 2014 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim)

heterointerface (model $\mathbf{d_1}$) and of one of chemisorbed systems (model $\mathbf{w_1}$). The imaginary part of the dielectric function of the two component layers of these two systems is also reported in the central and bottom panels. As visible from the figure, while the photo-responsivity in the visible region is not enhanced in the physisorbed heterostructure (panel a), in the chemisorbed one, a higher photo-responsivity in the VIS region is predicted (panel d). As discussed in detail in Ref. [53], this is due to the raise of the valence band maximum state which stems from the hybridization between the two-layered subunits.

Conclusions

The recently achieved impressive progress in the experimental investigation and synthesis of optoelectronic and solar energy device oriented oxide nanostructures demands for the use of reliable theoretical/computational simulations able to provide a microscopic understanding of the underlying chemistry and physics.

Not secondarily, theoretical predictions represent fundamental tools to guide designers towards new and more efficient architectures.

Parameter-free-excited state approaches based on the Green's function formalism applied on top of DFT calculations are nowadays successfully employed to investigate the electronic and optical properties of a large class of materials.

The absence of any external fitting parameters offers an accurate and unbiased microscopic description at quantum-mechanical level of the chemical-physical processes involved.

Many-body perturbation theory results obtained on top of mean-field density functional theory calculations have been discussed here with a particular focus on TiO_2-based nanostructures. The inclusion of many-body effects in the theoretical description is of primary importance to obtain a first firm comparison between theoretical and experimental data (when available). The excitonic nature of the main optical peaks clearly comes out from the analysis of the theoretical spectra of these low-dimensional systems.

Cross-References

- ▶ Computer Modeling and Simulation of Materials
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanostructured Materials for Sensing
- ▶ Optical and Electronic Properties
- ▶ Titanium Dioxide

References

1. De Angelis, F., Di Valentin, C., Fantacci, S., Vittadini, A., Selloni, A.: Theoretical studies on anatase and less common TiO_2 phases: bulk, surfaces, and nanomaterials. Chem. Rev. (2014). doi:10.1021/cr500055q, Article ASAP
2. Fujishima, A., Zhang, X., Tryk, D.A.: TiO_2 photocatalysis and related surface phenomena. Surf. Sci. Rep. **63**, 515–582 (2008)
3. Diebold, U.: The surface science of titanium dioxide. Surf. Sci. Rep. **48**, 53–229 (2003)
4. Chen, X., Mao, S. S.: Titanium dioxide nanomaterials: synthesis, properties, modifications Chem Rev. **107** (7), 2891–959 (2007)
5. Selloni, A.: Crystal growth: anatase shows its reactive side. Nat. Mater. **7**, 613–615 (2008)
6. Onida, G., Reining, L., Rubio, A.: Electronic excitations: density-functional versus many-body Green's-function approaches. Rev. Mod. Phys. **74**, 601–656 (2002), and references therein
7. Dreizler, R.M., Gross, E.K.U.: Density functional theory. Springer, Hedelberg (1990)
8. Burke, K.: Perspective on density functional theory. J. Chem. Phys. **136**, 150901 (2012), and references therein
9. Hohenberg, P., Kohn, W.: Inhomogeneous electron gas. Phys. Rev. **136**, B864–B871 (1964)
10. Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Phys. Rev. **140**, A1133–A1138 (1965)
11. Hedin, L.: New method for calculating the one-particle green's function with application to the electron-gas problem. Phys. Rev. **139**, A796–A823 (1965)
12. Aryasetiawan, F., Gunnarsson, O.: The GW method. Rep. Prog. Phys. **61**, 237–312 (1998)
13. Aulbur, W.G., Jonsson, L., Wilkins, J.W.: In: Ehrenreich, H., Spaepen, F. (eds.) Solid State Physics. Academic, New York (1999)
14. Strinati, G.: Application of the Green's functions method to the study of the optical properties of semiconductors. Riv. Nuovo Cimento **11**, 1–86 (1988)
15. Ceperley, D.M., Alder, B.J.: Ground state of the electron gas by a stochastic method. Phys. Rev. Lett. **45**, 566–569 (1980)
16. Perdew, J.P., Zunger, A.: Self-interaction correction to density-functional approximations for many-electron systems. Phys. Rev. B **23**, 5048–5079 (1981)

17. Perdew, J.P., Burke, K., Ernzerhof, M.: Generalized gradient approximation made simple. Phys. Rev. Lett. **77**, 3865–3868 (1996)
18. Gross, E.K.U., Dobson, F.J., Petersilka, M.: Density functional theory. Springer, New York (1996)
19. Cohen, A.J., Mori-Snchez, P., Yang, W.: Challenges for density functional theory. Chem. Rev. **112**(1), 289–320 (2012), and refs. therein
20. Fetter, L., Walecka, J.D.: Quantum theory of many body systems. McGraw-Hill, New York (1981)
21. Mattuck, R.D.: A guide to Feynman diagrams in the many body problem. McGraw-Hill, New York (1976)
22. Echenique, P.M., Pitarke, J.M., Chulkov, E.V., Rubio, A.: Theory of inelastic lifetimes of low-energy electrons in metals. Chem. Phys. **251**, 1–35 (2000)
23. Marini, A., et al.: Quasiparticle band-structure effects on the d hole lifetimes of copper within the GW approximation. Phys. Rev. B **66**, 161104(R) (2002)
24. Hedin, L., Lundquist, B.J.: Effects of electron–electron and electron–phonon interactions on the one-electron states of solids. In: Ehrereich, H., Seitz, F., Turnbull, D. (eds.) Solid State Physics, vol. 23, p. 1. Academic, New York (1969)
25. Schwinger, J.: On the Green's functions of quantized fields. I. Proc. Natl. Acad. Sci. **37**, 452–455 (1951)
26. Adler, S.L.: Quantum theory of the dielectric constant in real solids. Phys. Rev. **126**, 413 (1962); Wiser, N.: Phys. Rev. **129**, 62 (1963)
27. Hybertsen, M.S., Louie, S.G.: Electron correlation in semiconductors and insulators: band gaps and quasiparticle energies. Phys. Rev. B **34**, 5390–5413 (1986)
28. Godby, R.W., Schlüter, M., Sham, L.J.: Self-energy operators and exchange-correlation potentials in semiconductors. Phys. Rev. B **37**, 10159–10175 (1988)
29. Hanke, W.: Dielectric theory of elementary excitations in crystals. Adv. Phys. **27**, 287–341 (1978)
30. Chiodo, L., Salazar, M., Romero, A.H., Laricchia, S., Della Sala, F., Rubio, A.: Structure, electronic, and optical properties of TiO_2 atomic clusters: an ab initio study. J. Chem. Phys. **135**(24), 244704 (2011)
31. Chiodo, L., Iacomino, A., Palummo, M., Rubio, A.: Titania nanostructures electronic and optical response. In: Aliofkhazraei, M. (ed.) Handbook of Functional Nanomaterials, Nova Science Publishers, Ltd. (New York – USA) (2014)
32. Palummo, M., Giorgi, G., Chiodo, L., Rubio, A., Yamashita, K.: The nature of radiative transitions in TiO_2-based nanosheets. J. Phys. Chem. C **116**, 18495–18503 (2012)
33. Kresse, G., Furthmüller, J.: Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. Phys. Rev. B **54**, 11169–11186 (1996)
34. Kresse, G., Furthmüller, J.: Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Comput. Mater. Sci. **6**, 15–50 (1996)
35. Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Chiarotti, G.L., Cococcioni, M., Dabo, I., Dal Corso, A., Fabris, S., Fratesi, G., de Gironcoli, S., Gebauer, R., Gerstmann, U., Gougoussis, C., Kokalj, A., Lazzeri, M., Martin-Samos, L., Marzari, N., Mauri, F., Mazzarello, R., Paolini, S., Pasquarello, A., Paulatto, L., Sbraccia, C., Scandolo, S., Sclauzero, G., Seitsonen, A.P., Smogunov, A., Umari, P., Wentzcovitch, R.M.: QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. J. Phys. Condens. Matter **21**, 395502–395519 (2009)
36. Marini, A., Hogan, C., Gruning, M., Varsano, D.: yambo: an ab initio tool for excited state calculations. Comput. Phys. Commun. **180**, 1392–1403 (2009)
37. Blöchl, P.E.: Projector augmented-wave method. Phys. Rev. B **50**, 17953–17979 (1994)
38. Kresse, G., Joubert, D.: From ultrasoft pseudopotentials to the projector augmented-wave method. Phys. Rev. B **59**, 1758–1775 (1999)
39. Mowbray, D.J., Martinez, J.I., García-Lastra, J.M., Thygesen, K.S., Jacobsen, K.W.: Stability and electronic properties of TiO_2 nanostructures with and without B and N doping. J. Phys. Chem. C **113**, 12301 (2009)
40. Chiodo, L., Massaro, A., Laricchia, S., Della Sala, F., Cingolani, R., Salazar, M., Romero, A.H., Rubio, A.: Characterization of TiO_2 atomic crystals for nanocomposite materials oriented to optoelectronics. Opt. Quant. Electron. **44**(3–5), 291–296 (2012)
41. Mogilevsky, G., Chen, Q., Kulkarni, H., Kleinhammes, A., Mullins, W.M., Wu, Y.: Layered nanostructures of delaminated anatase:? Nanosheets and nanotubes. J. Phys. Chem. C **112**, 3239–3246 (2008)
42. Vittadini, A., Casarin, M.: Ab initio modeling of TiO_2 nanosheets. Theor. Chem. Acc. **120**, 551–556 (2008)
43. Sato, H., Ono, K., Sasaki, T., Yamagishi, A.: First-principles study of two-dimensional titanium dioxides. J. Phys. Chem. B **107**, 9824–9828 (2003)
44. Bavykin, D.V., Gordeev, S.N., Moskalenko, A.V., Lapkin, A.A., Walsh, F.C.: Apparent two-dimensional behavior of TiO_2 nanotubes revealed by light absorption and luminescence. J. Phys. Chem. B **109**, 8565–8569 (2005)
45. Wang, W.S., Wang, D.H., Qu, W.G., Lu, L.Q., Xu, A. W.: Large ultrathin anatase TiO_2 nanosheets with exposed {001} facets on graphene for enhanced visible light photocatalytic activity. J. Phys. Chem. C **116**, 19893 (2012)
46. Torres, S.M., Martinez, L.M.P., Figueiredo, J.L., Faria, J.L., Silva, A.M.T.: Design of graphene-based TiO_2 photocatalysts–a review. Environ. Sci. Pollut. Res. **19**, 3676 (2012)
47. Williams, G., Seger, B., Kamat, P.V.: TiO_2-graphene nanocomposites. UV-assisted photocatalytic reduction of graphene oxide. ACS Nano **2**, 1487 (2008)
48. Shen, J., Yan, B., Shi, M., Ma, H., Li, N., Ye, M.: One step of TiO_2-reduced oxide sheets. J. Mater. Chem. **21**, 3415 (2011)
49. Neto, A.H.C., Guinea, F., Peres, N.M.R., Novoselov, K. S., Geim, A.K.: The electronic properties of graphene. Rev. Mod. Phys. **81**, 109 (2009)

50. Geim, A.K.: Graphene: Status and Prospects. Science **324**, 530 (2009)
51. Novoselov, K.S., Jiang, Z., Zhang, Y., Morozov, S.V., Stormer, H.L., Zeitler, U., Maan, J.C., Boebinger, G. S., Kim, P., Geim, A.K.: Room-temperature quantum hall effect in graphene. Science **315**, 1379 (2007)
52. Nair, R.R., Blake, P., Grigorenko, A.N., Novoselov, K. S., Booth, T.J., Stauber, T., Peres, N.M.R., Geim, A. K.: Fine structure constant defines visual transparency of graphene. Science **320**, 1308 (2008)
53. Masuda, Y., Giorgi, G., Yamashita, K.: DFT study of anatase-derived TiO_2 nanosheets/graphene hybrid materials. Phys. Status Solidi B **251**(8), 1471–1479 (2014)
54. Meng, X., Geng, D., Liu, J., Li, R., Sun, X.: Controllable synthesis of graphene-based titanium dioxide nanocomposites by atomic layer deposition. Nanotechnology **22**, 165602 (2011)
55. Xiang, Q., Yu, J., Jaroniec, M.: Enhanced photocatalytic H_2-production activity of grahene-modified titania nanosheets. Nanoscale **3**, 3670 (2011)

Electronic Contact Testing Cards

▶ MEMS High-Density Probe Cards

Electronic Expression

▶ Dynamic Clamp

Electronic Pharmacology

▶ Dynamic Clamp

Electronic Properties

▶ Tight-Binding Simulations of Nanowires

Electronic Quantum Transport

▶ Microscopic Modeling of Solid-State Quantum Devices

Electronic Structure Calculations

▶ Ab Initio DFT Simulations of Nanostructures
▶ Tight-Binding Simulations of Nanowires

Electronic Structure of Nucleic Acids

▶ DNA from First Principles

Electronic Transport in Carbon Nanomaterials

Alejandro Lopez-Bezanilla[1], Stephan Roche[2,3], Eduardo Cruz-Silva[4], Bobby G. Sumpter[5] and Vincent Meunier[6]
[1]National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]CIN2 (ICN–CSIC), Catalan Institute of Nanotechnology, Universidad Autónoma de Barcelona, Barcelona, Spain
[3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[4]Department of Polymer Science and Engineering, University of Massachusetts Amherst, Amherst, MA, USA
[5]Computer Science and Mathematics Division and Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[6]Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, Troy, NY, USA

Synonyms

Charge transport in carbon-based nanoscaled materials; Electron transport in carbon nanotubes and graphene nanoribbons

Definition

Electronic transport is the transport of a charge carrier, either an electron or a hole, through a system that is coupled to two or more electrodes which serve as a source and sink for the charge carriers.

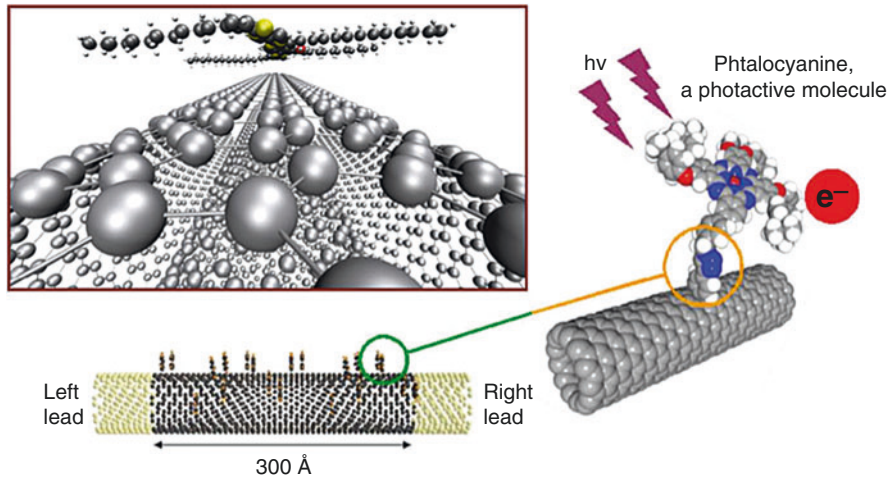
Single-walled carbon nanomaterials like carbon nanotubes and graphene nanoribbons are, respectively, tubular and planar arrangements of a one-atom-thick carbon sheet which possess a quasi one-dimensional (1D) geometry.

Introduction

As predicted by the miniaturization roadmap of the microelectronic industry, the gate length in metal-oxide-silicon-based field effect transistors (MOSFETs) has now been downscaled to a few tens of nanometers. This length reduction allows for the massive integration of billions of interconnected transistor devices, while preserving impressive current–voltage characteristics of individual components [1]. However, despite the sustained development of silicon-based nanoelectronics, the search for alternative materials for electronics has motivated the scientific community toward exploring the potential of molecular, organic, and carbon-based electronics such as carbon nanotubes or graphene [2]. The integration of carbon nanotubes (CNTs) and graphene nanoribbons (GNRs) as electron channels in electronic devices has become the focus of significant research and is the topic of this work. It should be stressed at the onset that, to date, carbon-based nanodevices have not been found to outperform silicon devices in mainstream microelectronics. Therefore, other types of applications need to be targeted. Among the wealth of possibilities, (bio)sensing at the nanoscale for medical, optical switches, photovoltaic cells, or nanoelectromechanical devices (NEMS) constitute some of the most promising alternatives [3, 4].

In spite of their promise, the use of such devices will remain elusive without a precise understanding of the intrinsic transport properties of carbon-based materials [5, 6]. Here, the advances made in the use of controlled chemical doping to fine-tune the transport properties of the pristine material as a particularly promising method of fundamental research are reviewed. Both experimental and theoretical studies have considered external agents like adsorbed atoms and molecules, substitutional impurities, and functional groups as possible candidates to introduce new functionalities in pristine carbon nanosystems. For instance, electroactive systems such as π-conjugated polymers display electronic and optical properties that can be tuned by appropriate molecular design. These systems also adopt several conformations with varying physical properties when molecular groups are subjected to oxidation or protonation through doping reactions. These polymers can also form molecular composites with other electroactive components of different chemical composition. Interesting properties emerge from the combination of CNTs and GNRs unique properties with those of the polymers themselves and open up many opportunities for potential applications, such as data storage media, photovoltaic cells, etc. - π-conjugated polymers are also known to strongly interact with a carbon nanotube surface through π-stacking interactions to form supramolecular adducts (Fig. 1a). This type of noncovalent functionalization allows for nanotube modification without the introduction of defects in the original geometry. At the same time, this procedure facilitates the molecular mixing of both composite components [7].

Conjugated π-networks decorate the surface of quasi-one-dimensional structures of CNTs and GNRs and confer to them either semiconducting or metallic properties, in addition to quasi-ballistic transport on micrometer length scales. Some external agents that do not necessarily modify the system geometry can alter the electronic transport features. For example, hole and electron doping chemistry does not significantly modify the position of individual carbon atoms but can create strongly localized states, electrostatic fields, and conduction gaps or localization phenomena. Conversely, the chemical modification of nanotubes or nanoribbons by covalent attachment of functional



Electronic Transport in Carbon Nanomaterials, Fig. 1 Physisorption (*top left*) and chemisorption (*bottom*) of organic molecules on carbon nanotubes. A phthalocyanine (*bottom*) is attached to a carbon nanotube by using a chain of phenyl groups that serve as linkers. This chemical attachment allows the nanotube to evacuate the electrons that result from photoactivation of the phthalocyanine. Modeling the degradation of electron transport abilities of the nanotube upon grafting of the phenyl groups is a fundamentally important task. In contrast, physisorbtion (*top left*) represents a noninvasive type of functionalization. The organic molecule(s) sit(s) in the proximity of the nanotube such that chemical bonds are not established and the electronic structure is only partially affected

groups causes simultaneous geometrical and electronic structure distortions. The alteration of the surface of the carbon nanostructures perturbs the crystal periodicity and modifies the electronic band structure and, in turn, the native quasi-ballistic nature of the electron transport properties [8, 9].

The experimental achievements in the field of nanotube functionalization have been spectacular. For instance, the "Click Chemistry" [10] approach succeeded in functionalizing and characterizing CNTs with zinc-phthalocyanine, with the aim of preparing a nanotube-based scaffold, which would make it easier to fabricate more complex functional structures. The attachment of phthalocyanines to nanotubes and fullerenes has emerged as a superb approach to carbon nanostructure phthalocyanine-based photovoltaic applications and other electronic devices [11]. Phthalocyanines are planar and electron-rich macrocycles characterized by remarkably high extinction coefficients in the red/near-infrared region (which is an important part of the solar radiation spectrum) and outstanding photostability and singular physical properties. These features make them exceptional building blocks for their incorporation in photovoltaic devices. Theoretical and computational studies support the understanding of how molecular states of phthalocyanines or other photoactive species mix with carbon nanotubes extended states (Fig. 1b).

Numerous questions regarding chemical functionalization still remain unsolved, especially since the goal is to develop novel applications. For instance, how does the interaction between nanotube and adsorbate affect their physical properties? What grafting process should be used to minimize the modification of the intrinsic and desirable electron transport capabilities of pristine nanotubes? In this regard, theoretical advances have allowed tremendous progress toward the understanding of some of the physical and chemical issues related to the modification of the electronic structure of carbon-based systems. However, the complexity of the calculations involving modified structures does not always permit a systematic comparison of theoretical modeling at size and timescales relevant to experiments. This complexity arises from the need to include a large number of atoms in the computer

simulation, as well as from the accuracy required to treat systems in a realistic manner. Notwithstanding, DFT-based approaches provide a set of powerful tools for the quantum mechanical investigations of the physical and chemical properties of nanosystems. The self-consistent calculation of the electronic structure of any material is critical whenever charge transfer between two interacting systems plays a major role in the interaction. Examples of this effect include chemical doping [12] or structural alterations introduced by topological defects that force local change in aromaticity [13]. Variations of the charge distribution leads to a number of effects, such as the appearance of local polarization and modification of the local geometry. For instance, when a functional group such as an aryl ring is positioned close to the surface of a CNT or a GNR, a covalent bond between two carbon atoms of both objects is established [14]. The covalent bond corresponds to the archetypical sharing of electrons between the two moieties and creates a dipolar moment that polarizes the charge distribution in the surrounding of the new carbon–carbon σ-bond between the individual molecules. The polarized charge distribution gives rise to an electrical field that perturbs the charge distribution that initially caused the polarization, and so forth in a self-consistent cycle.

The present entry primarily focuses on some recent works toward understanding the surface modification of carbon nanotubes and graphene nanoribbons by chemical doping such as adsorbed functional groups and substitutional doping. Depending on the chemical interaction, the transport properties of the carbon materials can be significantly altered. All the results shown here are based on DFT modeling, as implemented in the SIESTA [15, 16] and NWChem [17] codes. The localized orbital basis set, which constitutes one of the essential features of both codes, along with their reliable and fast calculations, are the main advantages of this approach. The tight-binding-like Hamiltonian obtained from the self-consistent calculation of an electronic structure presents the compactness required to study the transport property within the Green's function formalism.

In order to set the theoretical background to this entry, the principal concepts underlying the Landauer-Büttiker [18–20] theory of coherent electronic transport in low-dimensional systems are presented in section "The Landauer-Büttiker Quantum Transport Theory," along with a detailed description of the decimation technique in section "Green Function," which is particularly well suited for the calculation of long disordered systems. In section "Quantum Transport in Nanostructured Carbon Materials," a study of metallic and semiconducting functionalized carbon nanotubes by exohedral interactions with carbene and phenyl groups, endohedral amphoteric doping, and substitutional doping with N, P, B, S, and heteroatomics such as N–P is presented. In section "Summary and Conclusions," a summary and some conclusions are provided.

The Landauer-Büttiker Quantum Transport Theory

First, the basic ideas of the transport formalism are explored. The electronic transport calculations that will be presented are based on the Landauer-Büttikker formulation of conductance, which is particularly adequate for the study of electron motion along a device channel connected to two semi-infinite leads. The Landauer-Büttikker approach considers the transport properties of a phase coherent system formed by scattering region and leads, providing a simple and powerful framework for quantum transport in 1D mesoscopic systems. A device is defined as a material region where charge carriers spend a certain amount of time during their propagation. The electronic wave packets propagate quantum mechanically along a channel characterized by a cross-section and a longitudinal length. The channel is connected to left and right 1D semi-infinite leads that are in thermodynamical equilibrium with infinitely larger electron reservoirs. The Landauer-Büttikker approach is based on the classical concept of scattering: an incoming electron has a certain probability to be backscattered or transmitted when it crosses the device due to impurities. Material impurities are seen by

charges as attractive or repulsive potential barriers and the mutual interaction leads to various material conductance regimes. This transport model treats electron interferences from a quantum mechanical point of view corresponding to the coherent dynamics of the electron.

In 1957, Rolf Landauer introduced the revolutionary concept that contacts play an important role in transport phenomena, and should therefore be properly treated. In addition to the molecule or device resistance, the intrinsic resistance of the whole system is determined by the coupling of the channel with the electrodes. Landauer's formulation for the quantum mechanical problem of electronic transport consists in establishing a one-to-one correspondence between the conductance of a 1D conductor and the probability of an electron to transmit without reflection.

Landauer's model deals with strictly 1D systems, the conductance of which can be expressed by:

$$G = \frac{e^2}{h}\frac{T}{R},$$

where e is the electronic charge, h is the Planck constant, and T and R are the transmission and reflection probabilities, respectively. This equation has an apparent algebraic limitation: when T = 1 (i.e., perfect conductor) the conductance of the device becomes infinite. This result caused controversy at the time of Landauer's seminal publication since it is clear that no channel has an infinite conductance. Büttiker's work put the controversy to rest by considering the contribution of all the electronic channels of transmission. The resulting Landauer-Büttiker approach states that, in a two-probe system, the conductance per channel is

$$G_n = \frac{e^2}{h}T_n,$$

where T_n is the transmission coefficient for the *n*th channel. In this formula, several assumptions are made: there is no electron–electron interaction and each charge propagates coherently through the conductor with only possible elastic scattering events, since all the inelastic processes take place in the reservoirs. The final expression that takes into account all channel contributions, in addition to spin degeneracy therefore reads

$$G = G_0\sum_{n=1}^{N} T_n,$$

where $G_0 = \frac{2e^2}{h}$ is the quantum of conductance and the sum runs over all the different conducting channels at a given energy. This formula is usually known as the two-probe conductance formula. This description differs from Landauer's original formula in the conductance of a perfect 1D conductor is equal to G_0, which corresponds to maximum value of conductance. Reciprocally:

$$R = \frac{1}{G_0} = \frac{h}{2e^2} = 12.906\,\mathrm{k\Omega}$$

is the quantum of resistance. It is striking that this theory predicts that a perfect conductor has a minimum resistance larger than zero. The actual experimental verification of this theory devised in the 1950s was only achieved in the late 1980s by van Wees et al. [21] and Wharam et al. [22] who unambiguously demonstrated the finite minimum resistance of each individual channel of conductance.

Green Function

In this section, a primer on the use of Green functions in transport calculations is provided. In this work, electronic transport of single-particle problems are solved, the energetic of which is described by a density functional Hamiltonian, H, used in Schrödinger equation as:

$$H\Psi_i = \varepsilon_i\Psi_i \Rightarrow \left[-\frac{1}{2}\nabla^2 + V(r) - \varepsilon_i\right]\Psi_i(r) = 0,$$

where V(r) is a potential and ε_i are the eigenvalues associated to the eigenfunctions $\Psi_i(r)$. The wavefunctions can be expressed by an expansion in a complete set of basis functions $\{\varphi_\nu(r)\}$:

$$\Psi_i(r) = \sum_{v=1}^{N} c_{iv}\varphi_v(r).$$

Truncating the number of basis functions leads to an operational approximation. A sufficient large number of basis functions can reduce the imprecision. The Hamiltonian matrix elements of H in the $\{\varphi_v(r)\}$ basis are defined as:

$$H_{vv'} \equiv \int d3r\varphi_v^*(r)\left[-\frac{1}{2}\nabla^2 + V(r)\right]\varphi_{v'}(r)$$

and the overlap matrix element between two functions is defined as:

$$S_{vv'} \equiv \int d3r\varphi_v^*(r)\varphi_{v'}^*(r)$$

Inserting expression (2) into (1) a generalized eigenvalue problem is obtained:

$$\sum_{v'} H_{vv'}c_{iv'} = \varepsilon_i \sum_{v'} S_{vv'}c_{iv'} \quad \forall v.$$

From this expression, the Green's function is defined as:

$$\sum_{v''} (\varepsilon S_{vv''} - H_{vv'})G_{v''v'}(\varepsilon) \equiv \delta_{vv'}.$$

Within a tight-binding framework, the basis set can be formed by a set of atomic orbitals centered on each atom of the system that confers a compact form to the matrix. Inverting the matrix for a given value of ε is, in principle, the only calculation needed to determine the Green's function of the electronic system. In this work, the Hamiltonian and overlap matrices are provided by a DFT calculation of a system composed of hundred of atoms. Since the SIESTA and NWChem codes employ a basis set of atomic-like orbitals, the Hamiltonian matrix that result from a simulation meets the operational requirements of size and compactness to determine Green's function. Further, an appropriate choice of basis set size leads to a relatively compact set of matrices within the linear combination of atomic orbital model.

In a generalized consideration of the transport problem of a two-terminal geometry, the finite device is sandwiched between two semi-infinite electrodes, and the Hamiltonian of the entire system (i.e., an infinite size Hamiltonian) can be written as:

$$H = \begin{bmatrix} H_L & -\gamma_L & 0 \\ -\widetilde{\gamma}_L & H_0 & -\widetilde{\gamma}_R \\ 0 & -\gamma_R & H_R \end{bmatrix}$$

where H_0 is the device Hamiltonian, $H_{L,(R)}$ are the bulk left (right) electrode Hamiltonians, and $\gamma_{L,R}$ represent the tight-binding-like coupling between the electrodes and the device channel. The Green's function operator of the system can be expressed as:

$$[\varepsilon \cdot S - H]G(E) = \mathrm{I},$$

where I is the identity matrix and ε is the complex energy that imposes causality and is chosen to avoid the real axis poles of the Green's function by extending its domain to the complex plane. Depending on the positive or negative semi-infinite complex plane definition of ε, the Green's function can be labeled as $G_0^{r(a)}(\varepsilon)$,, that is, the retarded (advanced) Green functions of the device channel (fulfilling the relation $G_0^r(\varepsilon) = \tilde{G}_0^a(\varepsilon)$) where $\sim$ denotes the complex conjugated. Physically, these two Green functions represent an incoming and an outgoing wave formed upon the same excitation. It is worth noting that, while the contact function subspaces are semi-infinite, the device subspace has a finite representation. To obtain a finite matrix representation of the electrodes, it is useful to resort to the self-energy representation $\Sigma_{L,R}$ of both the left (L) and the right (R) electrodes:

$$\sum\nolimits_L = \widetilde{\gamma}_L g_L \gamma_L$$

$$\sum\nolimits_R = \widetilde{\gamma}_R g_R \gamma_R$$

Where

$$g_L = [E \cdot S - H_L]^{-1}$$

$$g_R = [\varepsilon \cdot S - H_R]^{-1}$$

are the (surface) Green functions of the left and right contact, respectively. The transformation of these two infinite matrices into two (*N, N*) matrices to obtain a finite expression is usually performed by taking advantage of the semi-infinite lattice symmetry of the two leads, either by analytical or iterative techniques. Once the self-energy terms are calculated, one can obtain the final expression of the Green's function matrix of the whole device as:

$$G_0 = \left[\varepsilon \cdot S - H_0 - \sum\nolimits_L - \sum\nolimits_R\right]^{-1}.$$

From this expression, the self-energies can be interpreted as additive potential terms of the device channel Hamiltonian whose final expression represents an effective Hamiltonian that accounts for the presence of the electrodes. The numerical representation within the chosen basis now corresponds to a matrix of finite rank. This is an undeniable advantage. Note that this standard procedure for the calculation of Green's function of a two-terminal system can be easily extended for multiterminal geometries, since each electrode can be included by a proper self-energy added to the expression of G_0.

Once the Green's function of the device channel is determined, the Caroli formula [23] is applied to calculate the transmission coefficients. In its compact form, the transmission between the right and left electrodes T(ε) is:

$$T(\varepsilon) = tr\left[\Gamma_R(\varepsilon) G_0^r(\varepsilon) \Gamma_L(\varepsilon) G_0^a(\varepsilon)\right],$$

where the contact terms, Γ_L and Γ_R for left (L) and right (R) electrodes are:

$$\Gamma_{R(L)}(\varepsilon) = i\left[\sum\nolimits_{R(L)}^{r}(\varepsilon) - \sum\nolimits_{R(L)}^{a}(\varepsilon)\right].$$

This expression for T(ε) is equivalent to that introduced in section "The Landauer-Büttiker Quantum Transport Theory" for the transmission coefficients.

Quantum Transport in Nanostructured Carbon Materials

Chemical Functionalization of Graphene Nanoribbons

Graphene is a zero-gap semiconductor and is therefore unsuitable for achieving efficient field-effect functionality with an acceptable ON/OFF states current ratio. One possibility to increase the zero-gap of two-dimensional (2D) graphene single layers is to shrink its lateral dimension. This dimensional confinement effectively opens an electronic bandgap that decreases in magnitude when the nanoribbon width increases. In spite of this, theoretical predictions and experimental observations have reported energy band-gaps far too small or very unstable in regard to edge reconstruction and defects, thus preventing envisioning these graphene-based devices from achieving improved performance relative to the ultimate CMOS-FETs.

To further introduce suitable band-gaps, one can instead recourse to larger width graphene nanoribbons (above 10 nm in lateral sizes) and compensate the loss due to the bandgap shrinking by triggering mobility gaps through doping or surface modification. Mobility gaps are direct consequences of a wide distribution of quasi-bound states over the valence band (for acceptor-type impurities) or the conduction band (for donor-type impurities) in the first conductance plateau when dopants are randomly distributed across the ribbon width because of the strong dependence of the scattering potential with the dopant position in respect to the ribbon edges. The electronic transport properties of GNRs modified by means of substitutional doping with atoms such as boron, nitrogen, or phosphorus and chemical attachment of functional groups have been simulated and has shown promising features [24].

Graphene is an aromatic material: its electronic ground state corresponds to completely delocalized π-electrons. It follows that all conjugated chemical bonds are equivalent. This symmetry is broken when a functional group is grafted onto graphene's surface since it removes a π-bond, thereby inducing an sp^2 to sp^3

rehybridization and a net magnetic moment [25]. However, this situation is not stable and as soon as a second atom is chemisorbed on graphene's basal plane, the unpaired electron participates in a second sp^3 hybridization and the magnetic moment vanishes. This formation of radicals associated with the removal of one p_z orbital in the conjugated graphene network also takes place in nanotube networks. Chemical functionalization of graphene-based materials is a promising strategy to reversibly tune the electronic properties without aggressively reducing the ribbon width. For instance, pairs of hydrogen (H–) and hydroxyl (OH–) functional groups attached on GNRs at opposite positions in a hexagonal graphene ring introduce the simultaneous appearance of two bonds with different charge densities in the system. The oxygen atom of the OH– group is a good acceptor of electrons while H– adopts a donor behavior when it is located in the vicinity of a C atom. The two equivalent bonds can also be treated by considering two phenyl rings covalently attached at opposite sides of a hexagonal ring.

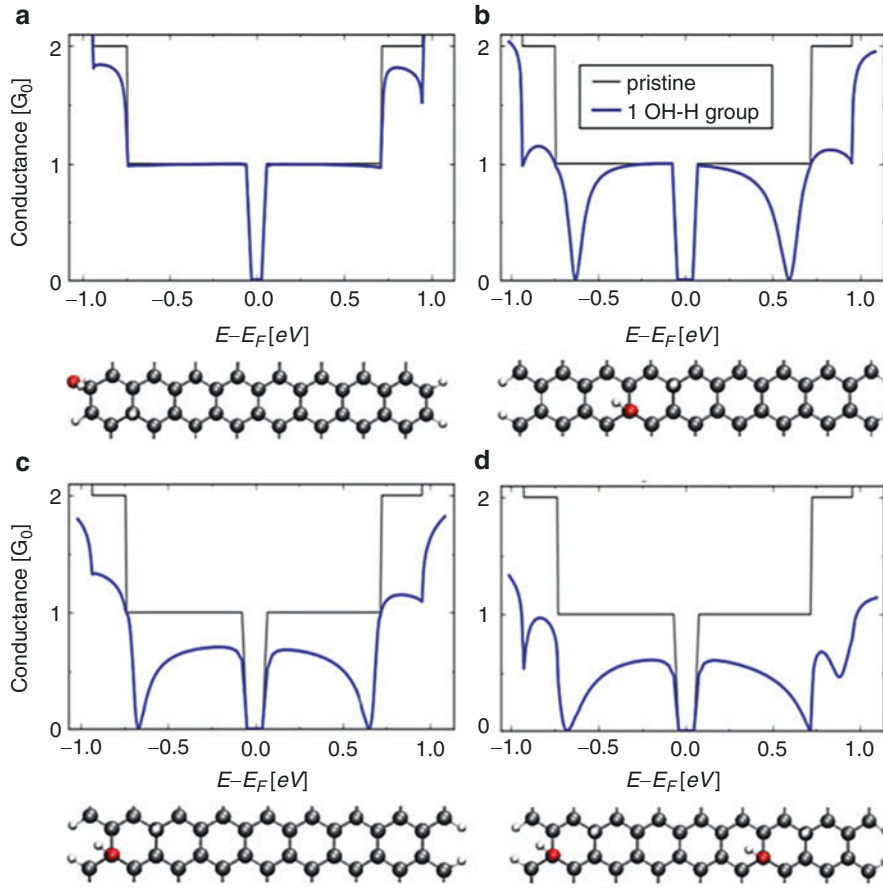
Several chemical groups can be simultaneously attached, which allows for a wide range of conductance and structural modulations depending on the relative positions of the addends. The particular configuration of a nanoribbon with functionalization at the edge is also of special interest. In this case, no localized states are introduced by the functional groups at energies close to the Fermi level and, therefore, the conductance ability of the ribbon is not significantly affected.

As explained in section "Green Function," a complete device can be set up once ab initio calculations are performed to obtain the Hamiltonian and overlap matrices associated with short functionalized nanoribbon sections. To compute the conductance of clean and defected GNRs, the multichannel Landauer-Büttiker technique described in sections "The Landauer-Büttiker Quantum Transport Theory" and "Green Function" is used. The channel is defined between two perfect GNR-based right (left) leads, which impose chemical potentials μ_1 (μ_2). This choice of leads also optimizes the coupling with the channel.

First the properties of an armchair-edged nanoribbon with 17 bonds along its length (hereafter called a17GNR) are discussed. The dependence of the a17GNR conductance as a function of the H– and OH– group position is shown in Fig. 2 for a set of different configurations. If the C–OH bond is located at the GNR edge, no backscattering is observed in the first conductance plateau (Fig. 2a), although some weak conductance decay is seen for higher energy sub-bands. Edge functionalization does not alter the electronic properties of the aGNR in the first sub-band since the impurity states are located far away from the Fermi level. In contrast, when the C–OH bond is shifted out of the edge line, the conductance is more significantly reduced as shown in Fig. 2b or in Fig. 2c. Figure 2b shows similarity with the situation for pair of grafted phenyls, whereas other locations (Fig. 2c) exhibit a larger suppression of conductance in the first plateau, although higher energy sub-bands remain largely unaffected. This dependence of the conductance on the location of the OH/H pair can be rationalized by analyzing the variation of electrostatic screening effects that dictate the effective scattering potential. The conductance decay is driven by backscattering on individual functional groups and by quantum interferences due to the proximity of the defects.

Zigzag-edged GNRs (zGNRs) display a number of different electronic features compared to aGNRs. As opposed to aGNRs, zGNRs present a magnetic ordering at the edges as a consequence of the strong localization of the unpaired π and π^* electrons in the ground state. These edge states are spin-degenerate but both ribbon edge states exhibit different spin orientations as a result of edge–edge interactions, and hence the difference between ferromagnetic and antiferromagnetic spin states disappears gradually for wide enough ribbons. First-principles calculations show that zGNRs are metallic in the ferromagnetic configuration whereas they exhibit semiconducting behavior in the antiferromagnetic state. zGNRs' magnetic properties have interesting implications for the development of electronic and spintronic devices.

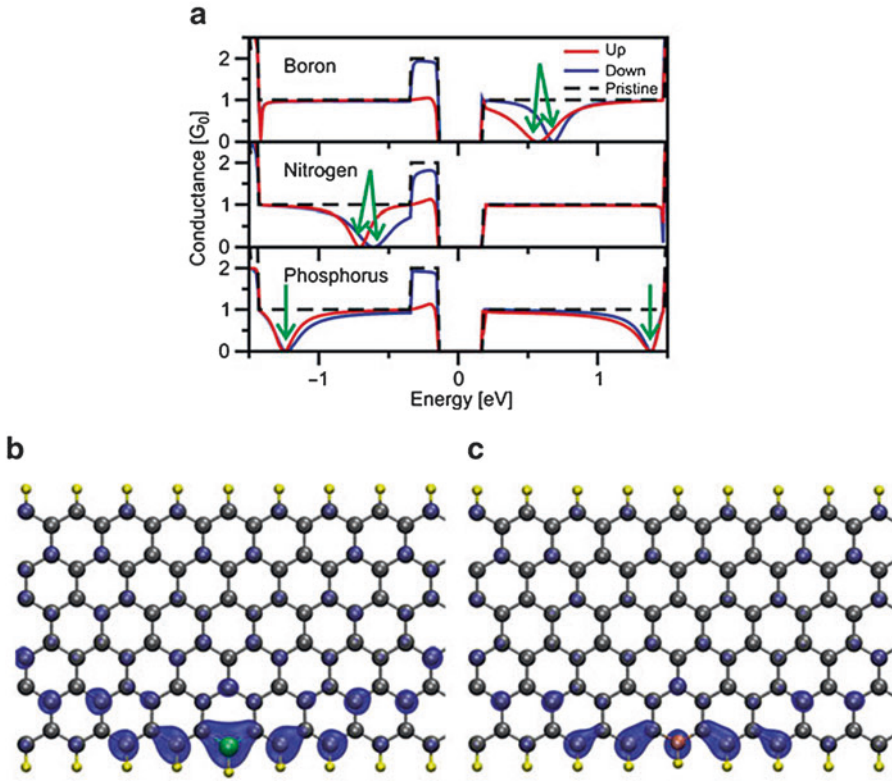
The case of GNR substitutionally doped with boron, nitrogen, and phosphorus atoms, which



Electronic Transport in Carbon Nanomaterials, Fig. 2 Conductance of narrow GNRs. Conductance of a17GNR as a function of defect position. The locations of the defects for (**a**–**d**) are illustrated at the *bottom*. The (**d**) case shows the conductance in the case of two pairs of functional groups

modify the electronic and transport properties of the ribbons, is now considered. The presence of a dopant typically induces localized states and a drop in conductance at some particular energies. The dependence of the conductance drop position in the transmission spectrum on the specific doping site was investigated within an ab initio approach. DFT is well suited to calculate the charge transfer between the doping atoms and the GNRs and thus to find the exact position in energy of the bound states. Nitrogen and phosphorus atoms preferentially substitute at the edges of zigzag and armchair nanoribbons, whereas boron atoms preferentially substitute at the edges of zGNRs but next to the edge of aGNRs. Spin-dependent conductance in the most energetically favorable boron-doped and nitrogen-doped zGNRs leads to the possibility of utilizing such structures in spintronic devices. The most favorable configuration of phosphorus atoms does not present a significant degree of spin-dependent behavior, but it displays both a donor-like and an acceptor-like state. The conductance of doped armchair nanoribbon exhibits spin degeneracy for all cases, but in phosphorus-doped zigzag ribbons, spin-dependent behavior can emerge as the doping is placed at particular sites across the ribbon width (Fig. 3).

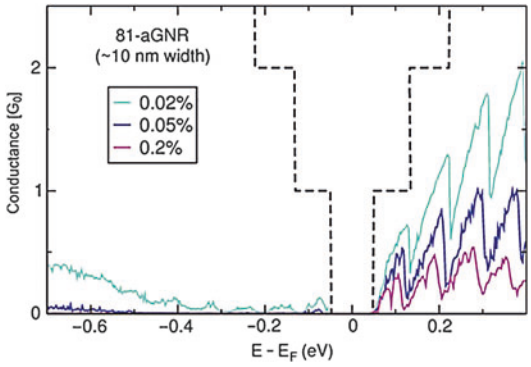
The large predicted variation of resonant energies with the dopant position with respect to the GNR edges [26] indicates that a random distribution of impurities along the ribbon surface should



Electronic Transport in Carbon Nanomaterials, Fig. 3 (**a**) Electronic quantum conductance for a zigzag graphene nanoribbon (12zGNR) doped with boron, nitrogen, or phosphorus. *Green arrows* indicate the conductance dips caused by the localized states associated to the doping atoms. In panels (**b**) and (**c**) illustrate the charge distribution of such localized states for boron (**b**) and nitrogen (**c**). All energies are relative to the Fermi energy of pristine nanotube

lead to a rather uniform reduction of the conductance over valence states on the first plateau of conductance. Turning to the quantum transport properties of boron doped GNRs up to 10 nm in width, first-principles calculations are combined with tight-binding models to investigate the effect of chemical doping on charge conduction for ribbons length up to 1 μm with randomly distributed substitutional boron impurities. Due to the large size of the corresponding unit cells, a statistical study of the mesoscopic transport using the fully ab initio Hamiltonian is not directly possible. Instead, a tight-binding model parameterized by accurate first-principles calculations is used. In a first step, the scattering potential induced by the dopant is extracted from those obtained with DFT. The analysis of the ab initio onsite and hopping terms around the doping atom obtained from self-consistent DFT calculations within an atomic-like basis, allows a simple nearest-neighbor tight-binding model to be built. The substitutional impurity modifies mainly the onsite terms, creating a potential well on a typical length scale of about 10 Å.

Figure 4 shows the conductance of a 10 nm wide armchair nanoribbon with a low dose for boron doping. For a doping density of about 0.2 %, the system presents a mobility gap of the order of 1 eV. When lowering the doping level to 0.05 %, the mobility gap reduces to about 0.5 eV and finally becomes less than 0.1 eV for lower density. The 0.2 % case is obtained for a fixed nanoribbon width and length, so that adjustments need to be performed for lateral or longitudinal size upscaling, but the method is straightforward once the transport length scales (mean free paths,



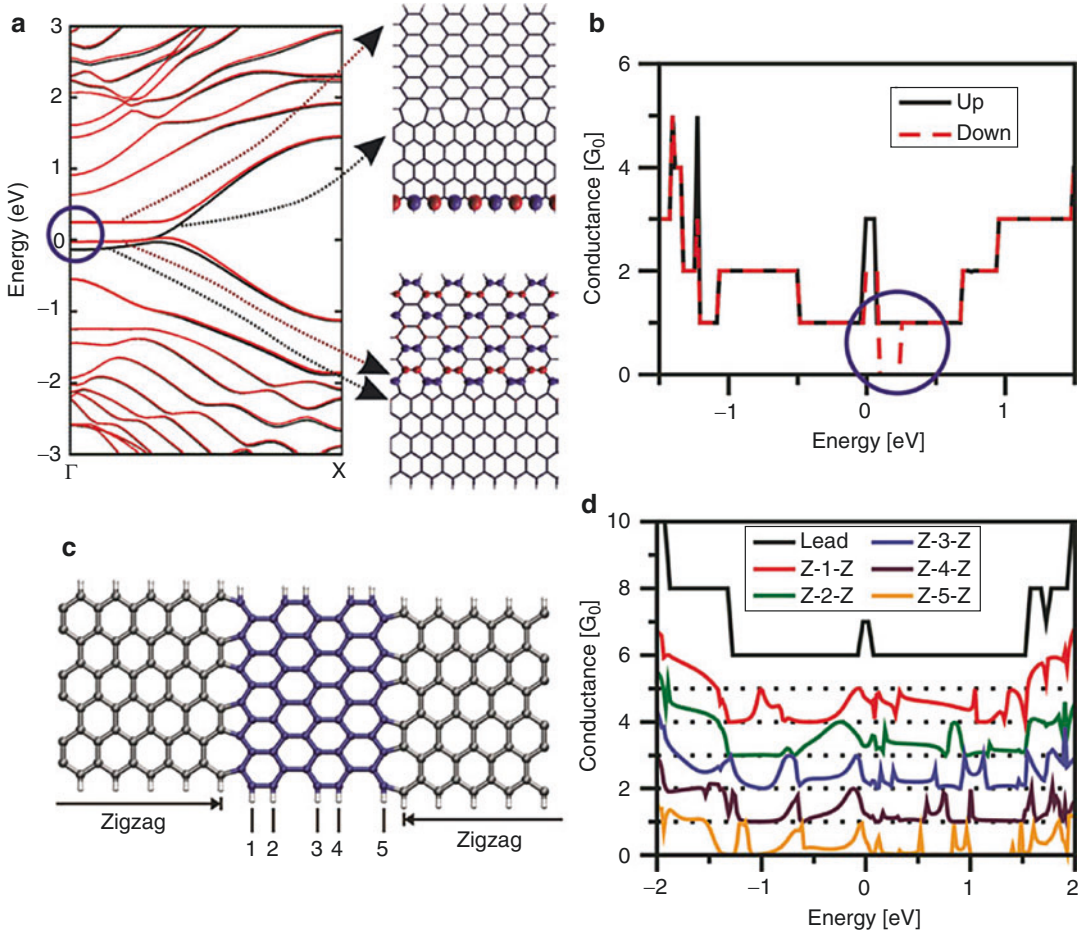
Electronic Transport in Carbon Nanomaterials, Fig. 4 Average conductance as a function of energy for a semiconducting 81-aGNR (*dashed line*) and three selected doping rates (0.02 %, 0.05 %, and 0.2 %, from *top to bottom*)

localization length) have been computed [24]. One notes however that the existence of mobility gaps (with conductance several orders of magnitude lower than the quantum conductance) can hardly allow a straightforward quantitative estimation of resulting ON/OFF current ratio, since this would require computing the charge flow in a self-consistent manner (using a Schrödinger-Poisson solver). This turns out to be critical since accumulated charges inside the ribbon channel screen the impurity potential in an unpredictable way. These types of calculations would need to be performed using nonequilibrium Green functions (NEGF) formalism and deserve further consideration [27].

To summarize the results presented thus far, it is clear that for any graphene-based application to be viable, it is critical to understand the effect of defects and disorder. In the following, junctions of nanoribbons based on an ordered array of structural defects [28] are treated. Such arrays of defects have been observed experimentally and could be visualized as grain boundaries in a graphene sheet [29]. In particular, an array of pentagons and heptagons can be used to join an armchair and a zigzag nanoribbon, as shown in Fig. 5. The spin resolved band structure for a (4, 4) hybrid nanoribbon is shown in Fig. 5a. Note the presence of an energy gap of ~0.2 eV in the minority spin states. For ribbons with an even width, there is a gap in the spin down states, that is, the systems exhibit half-metallicity and these hybrid systems behave as a spin polarized conductor (Fig. 5b). The isosurface plots of the wave functions near the Fermi level are indicated by the arrows on Fig. 5a. Some of these bands (top) are closely related to the zigzag nanoribbon edge states although, unlike zGNRs, there is only one edge state yielding a net magnetic moment. Another interesting case arises when an armchair segment is sandwiched in between two zGNRs electrodes, as shown in Fig. 5c. In this case, the two different lattices can be joined while keeping the threefold connectivity of each inner carbon atoms by introducing a set of pentagon-heptagon defects. The results presented in Fig. 5d indicate that the conductance near the Fermi level is reduced as the armchair section is increased, denoting a tunneling driven transport across the junctions.

Chemical Functionalization in Carbon Nanotubes

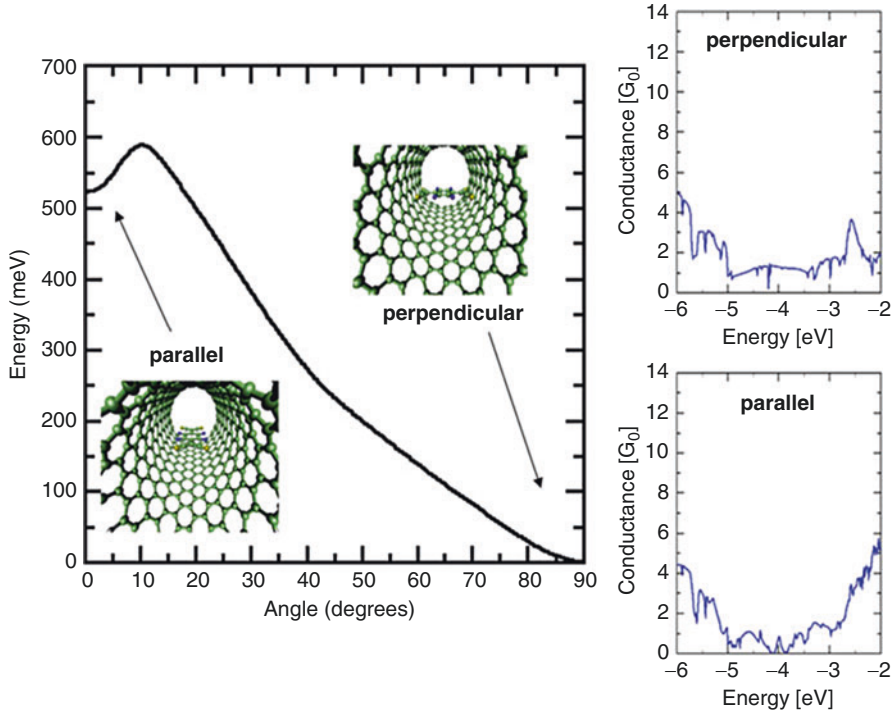
Amphoteric doping: It has been demonstrated that DFT-based transport calculations can be used at the fundamental level to devise a novel nanoelectronic device by exploiting the interplay between the intrinsic properties of pristine nanotubes and those of chemical dopants [30]. Here, a bistability is built in to form a nanoswitch whose conductance state depends on the relative position of a guest molecule inside of a carbon nanotube (see Fig. 6). Depending on the mode of operation, the proposed device is capable of acting as an electrical switch or a nonvolatile memory element. The switching mechanism consists of a significant change in the flow of electrons in a circuit due to different orientations of a molecule relative to the host nanotube. As shown in Fig. 6, one stable position of the molecule yields a high current across the device ("ON" conducting state) while a change in orientation is associated with an important, measurable decrease of the transmission property ("OFF" nonconducting state). The information is therefore stored in the form of the orientation of the molecule (parallel for "OFF" and perpendicular for "ON"), which retains its position even when the external source of energy is switched off



Electronic Transport in Carbon Nanomaterials, Fig. 5 (**a**) Computed spin resolved band structure for a (4,4) hybrid nanoribbon. The energy gap ($E_g = 0.27$ eV) on the minority spin (*red*) is indicated by a *blue circle*. The *arrows* show the wave function plot for the bands close to the Fermi level at the Γ-point. (**b**) Quantum conductance for a typical (4,4) hybrid graphene nanoribbon. The nanoribbon is spin degenerate in most of the energy spectrum, except for the region close to the Fermi level, where spin-polarized conductance is observed. (**c**) Example of a zigzag-armchair-zigzag device joined by 5–7 defects. (**d**) Quantum conductance for a set of these devices. It is observed that the conductance is reduced as the armchair section is increased, revealing a tunneling driven transport

(nonvolatility). In addition, due to the intrinsic mechanical properties of the nanotube, the molecule's orientation can be reliably switched back and forth mechanically. The information storage mechanism remains valid for any molecule–metallic nanotube combination as long as it displays the desired properties of significant charge transfer, and the presence of at least two metastable molecule positions with respect to the nanotube core. In addition, these two positions, in order to present distinct transport properties, must be characterized by clearly different types of interaction (such as different amounts of charge transfer for instance). These are necessary conditions but they are not always simple to guarantee a priori, and computational exploration is needed to properly select the correct combination of systems.

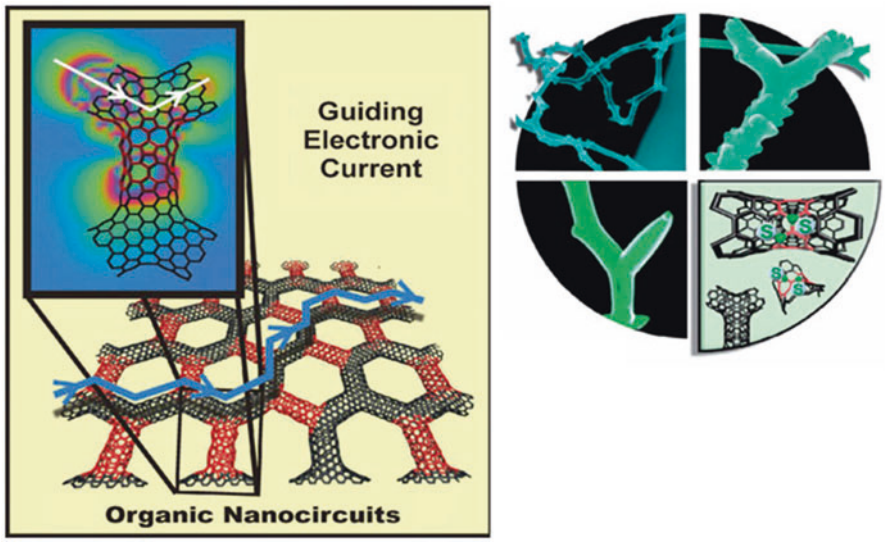
In this regard, it was shown that computational approaches paired with density functional theory studies can be routinely used to identify relevant combinations of molecules and wires/CNTs for realistic applications. In particular, a class of devices where a noncovalently bound molecule



Electronic Transport in Carbon Nanomaterials, Fig. 6 *Left panel*: total DFT energy of a "guest" F_4TCNQ molecule as a function of the angle between the main molecule axis and the "host" nanotube axis (inserts show the orientation). *Right panels*: conductance profile of the two different open systems where the active region is seamlessly connected to semi-infinite (10,10) carbon nanotubes

that acts as a passive gate can be extended to a large number of molecule-nanotube pairs, giving great versatility in the choice of nanotube diameter and doping molecule. Specifically, for a given molecule, it is possible not only to identify a nanotube that displays the characteristics of a memory element device but also to develop a detailed understanding of the criteria governing the functioning of this device.

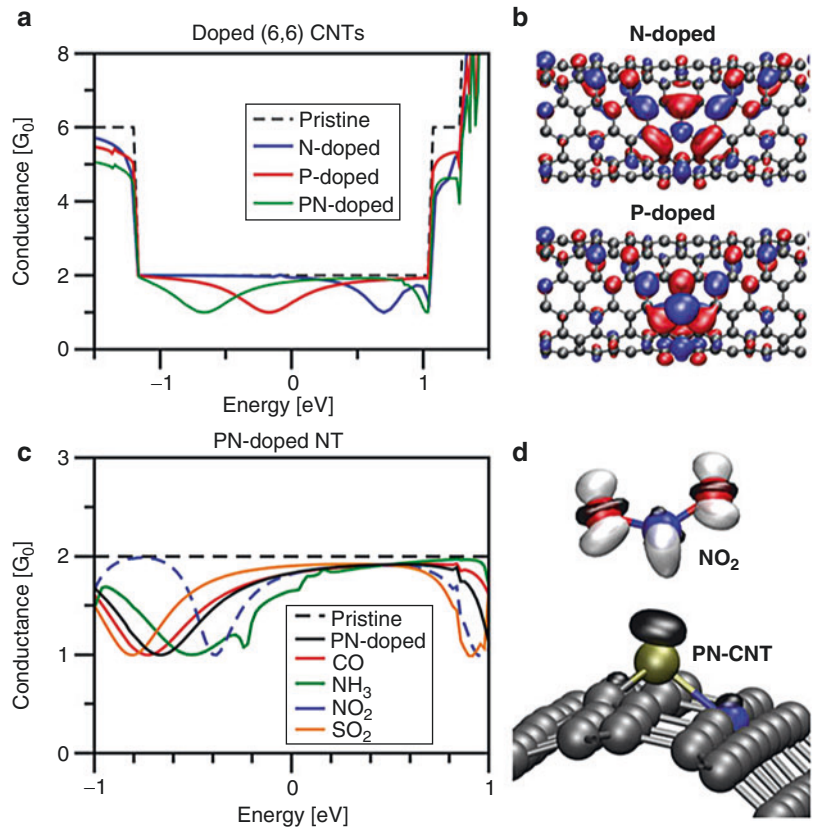
Nanonetworks: Carbon nanotubes can be combined to create 2D and 3D (three-dimensional) networks. These complex systems can be tailored from nanostructures as building blocks and are the foundations for constructing multifunctional nano- and microdevices. However, assembling nanostructures into ordered micronetworks remains a significant challenge in nanotechnology. The most suitable building blocks for assembling such networks are nanoparticle clusters, nanotubes, and nanowires. Unfortunately, little is known regarding the different ways networks can be created and their physicochemical properties as a function of their architecture. It is expected that, when 1D nanostructures are connected covalently [31], the resulting assemblies will possess mechanical, electronic, and porosity properties that are strikingly different from those of the isolated 1D blocks. In extensive theoretical studies, it has been shown that the properties of 2D and 3D networks built from 1D units are dictated by the specific architecture of these arrays. Specifically, the hierarchy concept as a practical way to design complex nanostructures from basic nano-building blocks was introduced. It was also demonstrated that nanotubes could be joined to make super networks with new properties emerging from those of the individual building blocks (i.e., the nanotubes). In addition to the unusual mechanical and electronic properties, the porosity of these systems makes them good candidates for catalysts, sensors, filters, or molecular and energy storage properties. The crystalline 2D and 3D



Electronic Transport in Carbon Nanomaterials, Fig. 7 *Left panel*: An ordered carbon nanotube network where defects introduced at the intersection of the nanotubes act as electron scattering centers (insert). *Right panel*: some of the different types of carbon nanotube architectures possible by introducing small amounts of sulfur into the precursors for the synthesis

networks are also expected to present unusual optical properties, in particular when the pore periodicity approaches the wavelength of different light sources, such as optical photonic crystals. The power of theoretical calculations in predicting novel materials with enhanced electronic and mechanical properties, using single atoms as building blocks, has been demonstrated repeatedly. These systems were studied by examining how atomic level modification of the networks can be made to tailor the properties of the current flow, a key ingredient for smart and addressable nanoarrays. The electrical current could be efficiently guided in 2D nanotube networks by introducing specific topological defects within the periodic framework. Using semiempirical transport calculations coupled with the Landauer-Büttiker formalism of quantum transport in multiterminal nanoscale systems, a detailed analysis of the processes governing the atomic-scale design of nanotube circuits was possible. When defects are introduced as patches in specific sites, they act as reflecting centers that re-inject electrons along specific paths, via a wave reflection process (see Fig. 7 left panel). This type of defect can be incorporated while preserving the threefold connectivity of each carbon atom embedded within the graphitic lattice. These findings introduce a new way to explore bottom-up design, at the nanometer scale, of complex nanotube circuits that could be extended to 3D nanosystems and applied in the fabrication of nanoelectronic devices.

Using knowledge acquired during the nanonetwork characterization and the behavior of nanostructures relative to local changes in chemistry, it was shown that the actual realization of networks can be obtained by inserting minute amounts of sulfur in precursors used for nanotube formation. This method was found to lead to the creation of a large range of structures, including crossbars and multiterminal junctions (Fig. 7 right panel) [32]. It also clarified the key role of sulfur in nanotube growth. Most notably, it was demonstrated that sulfur not only promotes the C-H separation in the precursor but also triggers the emergence of negative and positive local curvature, which are essential in the development of multiterminal, covalent systems. Here, the dopant's main role is to create nanostructures whose properties emerge from those of the individual components. For instance, sulfur triggers



Electronic Transport in Carbon Nanomaterials, Fig. 8 (**a**) Electronic quantum conductance of nitrogen, phosphorus, and phosphorus–nitrogen doped (6,6) SWCNTs, where the dips in conductance caused by scattering at the localized states can be observed. (**b**) Real space representation of the localized wavefunctions for nitrogen and phosphorus doped nanotubes. (**c**) Variations in the electronic conductance for a PN-doped nanotube after the physical/chemical absorption of different molecular species at the phosphorus localized state. (**d**) Variations of the charge density distribution around a NO_2 molecule absorbed at a PN site. *Black/white* clouds represent increase/decrease of the electronic charge density. All energies are relative to the Fermi energy of pristine nanotube

branching and initiates network creation but does not play a major beneficial role in the electronic transport properties of the resulting object.

Heteroatom Doping: Substitutional doping of carbon nanotubes and graphene clearly offers a practical path to tailor their physical and chemical properties by creating new states that modify their electronic structure [33]. The presence of these states, as discussed above, originates from the different electronic configurations of the doping atoms. These modifications in the electronic structure and surface reactivity can help to achieve new improved materials. Nitrogen and boron are among the most studied substitutional dopants used in nanotube research but phosphorus and phosphorus-nitrogen heteroatomic doping of carbon nanotubes have been successfully realized in experiments by using thermolysis of mixtures of ferrocene and triphenyl phosphine dissolved in benzylamine and ethanol. In order to fully understand the structure and properties of these new types of doped nanotubes, extensive electronic structure and transport calculations were used.

DFT calculations revealed an electronic band structure displaying the presence of localized (P) and semilocalized (PN) states around the doping atoms. In contrast to nitrogen, these states are normal to the nanotube surface and do not modify the intrinsic nanotube metallicity, and therefore semiconducting nanotubes remain so regardless of the doping. However, these electronic states behave as scattering centers for carriers with energies close to the localized state energy. Electronic transport calculations on pristine, nitrogen, phosphorus, and P–N doped nanotubes clarified the different effects of the dopants on their conductance. The dips observed in the quantum conductance (Fig. 8a) correspond to the localized and semilocalized states located around the doping atoms (Fig. 8b). The calculations of the quantum conductance on semiconducting zigzag nantubes showed that phosphorus doping do not modify their intrinsic

semiconducting behavior, as opposed to the effect observed for N-doped nanotubes. Phosphorus and PN doping in a (10,0) nanotube only creates bound and quasibound states around the phosphorus atom, which are dispersionless (flat bands in the band structure), and projected as sharp peaks in the density of states.

Due to the sp^3 hybridization, substitutional P atoms induce highly localized states which modify the chemical properties of the surface of carbon nanotubes. These sites have a strong affinity toward acceptor molecules. On the other hand, it was found that P–N co-dopants not only have a reduced affinity for acceptor molecules, but that the P–N bond can also accept charge, resulting in affinity toward donor molecules because of a partial positive charge at the phosphorus atom. For example, CO and NH_3 were found to adsorb only on PN-doped nanotubes, O_2 was adsorbed only on P-doped nanotubes, while NO_2 and SO_2 were adsorbed on both P- and PN-doped nanotubes. A natural bond orbital (NBO) analysis showed that the P atom has a very different chemical environment in P- and PN-doped nanotubes as a result of variations in charge and local bonding conditions.

The chemical changes at the localized electronic states affect its energy, resulting in different positions for dips in the conductance that are associated with them (Fig. 8c). SO_2 and CO cause only minor shifts in the conductance, while NH_3 causes the suppression of a conductance dip, and NO_2 and O_2 had similar effects in the conductance of P-doped carbon nanotubes. Since changes in the conductance at Fermi energy of about 10 % can be detected experimentally and, in particular, since modifications of the conductance slope near the Fermi energy can be an alternative for experimental measurements, the results of this study demonstrated the potential for a molecular sensor with identifiable selectivity that is based on P- and PN-doped carbon nanotubes.

Summary and Conclusions

At the nanoscale a number of new phenomena resulting from quantum effects and confinement, begin to dominate the processes governing the electron flow across the active device. Understanding those quantum effects is critical in the quest to develop new device concepts or novel mechanisms tuned to desired functionality and performance. Given the inherent difficulty in the integration of nanoscale systems like carbon nanotubes and graphene nanoribbons (e.g., controlled manipulation without agglomeration), into precise arrangements for devices, and the complexity in the assessment of the properties of these devices, a tremendous effort has been devoted in past years toward the development of theoretical methods and computational tools that make it possible to design and test new systems for desired and tailored characteristics. Realistic treatment of the fundamental properties of novel devices at the quantum mechanical level is now becoming a more routine task mainly due to accurate approaches such as density functional theory, nonequilibrium Green function formalism for electron transport in open-systems, and the advent of larger and more powerful computers. Extensive applications of these approaches are beginning to uncover new behavior in cabon-based materials that are uniquely suited for nanoelectronic applications.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Carbon-Nanotubes
- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Graphene
- ▶ Nanostructured Materials for Sensing

References

1. Narendra, S.G., Chandrakasan, A.P.: Leakage in Nanometer CMOS Technologies. Springer, New York (2006)
2. Avouris, P., Chen, Z., Perebeinos, V.: Carbon-based electronics. Nat. Nanotechnol. **2**(10), 605–615 (2007)
3. Allen, B.L., Kichambare, P.D., Star, A.: Carbon nanotube field-effect transistor-based biosensors. Adv. Mater. **19**(11), 1439–1451 (2007)

E

4. Vincent, D., Auvray, S., Borghetti, J., Chung, C.-L., Lefevre, R., Lopez-Bezanilla, A., Nguyen, K., Robert, G., Schmidt, G., Anghel, C., Chimot, N., Lyonnais, S., Streiff, S., Campidelli, S., Chenevier, P., Filoramo, A., Goffman, M.F., Goux-Capes, L., Latil, S., Blase, X., Triozon, F., Roche, S., Bourgoin, J.-P.: Carbon nanotube chemistry and assembly for electronic devices. C. R. Phys. **10**(4), 330–347 (2009)
5. Charlier, J.-C., Blase, X., Roche, S.: Electronic and transport properties of nanotubes. Rev. Mod. Phys. **79**(2), 677–732 (2007)
6. Castro Neto, A.H., Guinea, F., Peres, N.M.R., Novoselov, K.S., Geim, A.K.: The electronic properties of graphene. Rev. Mod. Phys. **81**(1), 109–162 (2009)
7. Sumpter, B.G., Meunier, V., Jiang, D.E.: New insight into carbon nanotube electronic structure selectivity. Small **4**, 2035 (2008); Linton, D., Driva, P., Sumpter, B., Inanov, I., Geohegan, D., Feigerle, C., Dadmun, M.D.: The importance of chain connectivity in the formation of non-covalent interactions between polymers and single-walled carbon nanotubes and its impact on dispersion. Soft Matter **6**, 2801 (2010)
8. Lopez-Bezanilla, A., Triozon, F., Latil, S., Blase, X., Roche, S.: Effect of the chemical functionalization on charge transport in carbon nanotubes at the mesoscopic scale. Nano Lett. **9**(3), 940–944 (2009)
9. Lopez-Bezanilla, A., Triozon, F., Roche, S.: Chemical functionalization effects on armchair graphene nanoribbon transport. Nano Lett. **9**(7), 2537–2541 (2009)
10. Kolb, H.C., Finn, M.G., Sharpless, K.B.: Click chemistry: diverse chemical function from a few good reactions. Angew. Chem. Int. Ed. **40**(11), 2004 (2001)
11. Campidelli, S., Ballesteros, B., Filoramo, A., Diaz, D., Torre, G., Torres, T., Aminur Rahman, G.M., Ehli, C., Kiessling, D., Werner, F., Sgobba, V., Guldi, D.M., Ciofi, C., Prato, M., Bourgoin, J.-P.: Facile decoration of functionalized single-wall carbon nanotubes with phthalocyanines via "click chemistry". J. Am. Chem. Soc. **130**(34), 11503–11509 (2008)
12. Biel, B., Blase, X., Triozon, F., Roche, S.: Anomalous doping effects on charge transport in graphene nanoribbons. Phys. Rev. Lett. **102**(9), 096803 (2009)
13. Dubois, S.M.M., Lopez-Bezanilla, A., Cresti, A., Triozon, F., Biel, B., Charlier, J.-C., Roche, S.: Quantum transport in graphene nanoribbons: effects of edge reconstruction and chemical reactivity. ACS Nano **4**(4), 1971–1976 (2010); Valiev, M., Bylaska, E.J., Govind, N., Kowalski, K., Straatsma, T.P., van Dam, H.J.J., Wang, D., Nieplocha, J., Apra, E., Windus, T.L., de Jong, W.A.: NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. Comput. Phys. Commun. **181**, 1477 (2010)
14. Jiang, D.E., Sumpter, B.G., Dai, S.: How do aryl groups attach to a graphene sheet? J. Phys. Chem. B **110**, 23628 (2006)
15. Ordejon, P., Artacho, E., Soler, J.M.: Self-consistent order-N density-functional calculations for very large systems. Phys. Rev. B **53**(16), 10441–10444 (1996)
16. Soler, J.M., Artacho, E., Gale, J.D., Garcia, A., Junquera, J., Ordejon, P., Sanchez-Portal, D.: The SIESTA method for ab initio order-N materials simulation. J. Phys. Condens. Matter **14**(11), 2745–2779 (2002)
17. Valiev, M., Bylaska, E.J., Govind, N., Kowalski, K., Straatsma, T.P., van Dam, H.J.J., Wang, D., Nieplocha, J., Apra, E., Windus, T.L., de Jong, W.A.: NWChem: a comprehensive and scalable open-source solution for large scale molecular simulations. Comput. Phys. Commun. **181**, 1477 (2010)
18. Landauer, R.: IBM J. Res. Dev. **32**, 306 (1988)
19. Büttiker, M.: IBM J. Res. Dev. **32**, 317 (1988)
20. Büttiker, M., Imry, Y., Landauer, R., Pinhas, S.: Generalized many channel conductance formula with application to small rings. Phys. Rev. B **31**(10), 6207–6215 (1985)
21. van Wees, B.J., et al.: Phys. Rev. Lett. **60**, 848–850 (1988)
22. Wharam, D.A., et al.: J. Phys. C: Solid State Phys. **21**, L209 (1988)
23. Caroli, C., Combescot, R., Nozieres, P., Saint-James, D.: J. Phys. C **4**, 916 (1971)
24. Biel, B., Triozon, F., Blase, X., Roche, S.: Nano Lett. **9**(7), 2725–2729 (2009)
25. Boukhvalov, D.W., Katsnelson, M.I.: Chemical functionalization of graphene with defects. Nano Lett. **8**(12), 4373 (2008)
26. Cruz-Silva, E., Barnett, Z.M., Sumpter, B.G., Meunier, V.: Structural, magnetic, and transport properties of substitutionally doped graphene nanoribbons from first principles. Phys. Rev. B **23**, 155445 (2011)
27. Brandbyge, M., Mozos, J.L., Ordejón, P., Taylor, J., Stokbro, K.: Density-functional method for non-equilibrium electron transport. Phys. Rev. B **65**, 165401 (2002)
28. Botello-Mendez, A.R., Cruz-Silva, E., Lopez-Urias, F., Sumpter, B.G., Meunier, V., Terrones, M., Terrones, H.: Spin polarized conductance in hybrid graphene nanoribbons using 5–7 defects. ACS Nano **3**, 3606–3612 (2009)
29. Simonis, P., Goffaux, C., Thiry, P.A., Biro, L.P., Lambin, P., Meunier, V.: STM study of a grain boundary in graphite. Surf. Sci. **511**, 319–322 (2002)
30. Wenchang, L., Meunier, V., Sumpter, B.G., Bernholc, J.: Density functional theory studies of quantum transport in molecular systems. Int. J. Quantum Chem. **106**, 3334 (2006); Meunier, V., Sumpter, B.G.: Amphoteric doping of carbon nanotubes by encapsulation of organic molecules: electronic properties and quantum conductance. J. Chem. Phys. **123**, 024705 (2005); Meunier, V., Sumpter, B.G.: Tuning the conductance of carbon nanotubes with encapsulated molecules. Nanotechnology **18**, 424032 (2007); Meunier, V., Kalinin, S.V., Sumpter, B.G.: Nonvolatile memory elements based on the intercalation of organic molecules inside carbon nanotubes. Phys. Rev. Lett. **98**, 056401 (2007)

31. Rodriguez-Manzo, J.A., Banhart, F., Terrones, M., Terrones, H., Gobert, N., Ajayan, P.M., Sumpter, B. G., Meunier, V.: Covalent metal-nanotube heterojunctions as ultimate nano-contacts. Proc. Natl. Acad. Sci. U. S. A. **106**, 4591 (2009)
32. Romo-Herrera, J.M., Sumpter, B.G., Cullen, D.A., Terrones, H., Cruz-Silva, E., Smith, D.J., Meunier, V., Terrones, M.: An atomistic branching mechanism for carbon nanotubes: sulfur as the triggering agent. Angew. Chem. Int. Ed. **47**, 2948 (2008); Romo-Herrera, J.M., Cullen, D.A., Cruz-Silva, E., Sumpter, B.G., Meunier, V., Terrones, H., Smith, D., Terrones, M.: The role of sulfur in the synthesis of novel carbon morphologies: from covalent Y-junctions to sea urchin-like structures. Adv. Funct. Mater. **19**, 1193 (2009); Sumpter, B.G., Huang, J., Meunier, V., Romo-Herrera, J.M., Cruz-Silva, E., Terrones, H., Terrones, M.: A theoretical and experimental study on manipulating the structure and properties of carbon nanotubes using substitutional dopants. Int. J. Quantum Chem. **109**, 97–118 (2009)
33. Cruz-Silva, E., López-Urías, F., Munoz-Sandoval, E., Sumpter, B.G., Terrones, H., Charlier, J.-C., Meunier, V., Terrones, M.: Electronic transport and mechanical properties of phosphorus and phosphorus-nitrogen doped carbon nanotubes. ACS Nano **3**, 1913–1921 (2009); Cruz-Silva, E., Cullen, D.A., Gu, L., Romo-Herrera, J.M., Munoz-Sandoval, E., Lopez-Urias, F., Sumpter, B.G., Meunier, V., Charlier, J.C., Smith, D. J., Terrones, H., Terrones, M.: Heterodoped nanotubes: theory, synthesis, and characterization of phosphorus-nitrogen doped multiwalled carbon nanotubes. ACS Nano **2**, 441 (2008); Sumpter, B.G., Meunier, V., Romo-Herrera, J.M., Cruz-Silva, E., Cullen, D.A., Terrones, H., Smith, D.J., Terrones, M.: Nitrogen-mediated carbon nanotube growth: diameter reduction, metallicity, bundle dispersibility, and bamboo-like structure formation. ACS Nano **1**, 369 (2007); Maciel, I.O., Campos-Delgado, J., Cruz-Silva, E., Pimenta, M.A., Sumpter, B.G., Meunier, V., Lopez-Urias, F., Muñoz-Sandoval, E., Terrones, M., Terrones, H., Jorio, A.: Synthesis, electronic structure and Raman scattering of phosphorous-doped single-wall carbon nanotubes. Nano Lett. **9**, 2267 (2009); Cruz-Silva, E., Lopez-Urias, F., Munoz-Sandoval, E., Sumpter, B.G., Terrones, H., Charlier, J.-C., Meunier, V., Terrones, M.: Phosphorus and phosphorus–-nitrogen doped carbon nanotubes for ultrasensitive and selective molecular detection. Nanoscale **3**(3), 1008–1013 (2011)

Electronic Visual Prosthesis

► Artificial Retina: Focus on Clinical and Fabrication Considerations

Electroplating

► Nanosized Nanocrystalline and Nanotwinned Metals

Electroresponsive Polymers

► Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP)

Electrospinning

Michael J. Laudenslager[1] and Wolfgang M. Sigmund[1,2]
[1]Department of Materials Science and Engineering, University of Florida, Gainesville, FL, USA
[2]Department of Energy Engineering, Hanyang University, Seoul, Republic of Korea

Synonyms

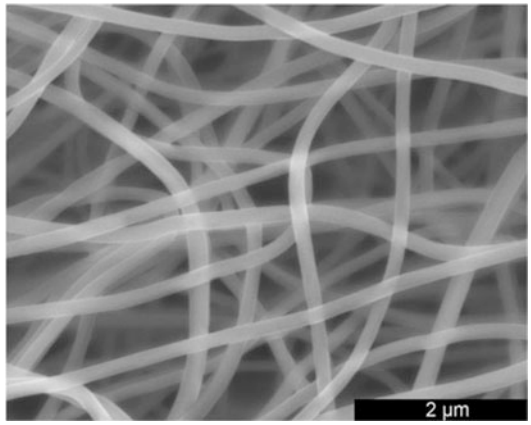
Electrohydrodynamic forming

Definition

Electrospinning is a process to obtain polymer, ceramic, metallic, and composite fibers from solutions, dispersions, or melts as a liquid jet accelerates through an electric field.

Introduction

Electrospinning produces long fibers with diameters that range from tens of nanometers to several microns. Despite the nanoscale diameters, the fiber lengths can reach several meters. Figure 1 shows a typical scanning electron microscope image of randomly oriented electrospun nanofibers.



Electrospinning, Fig. 1 Electrospun fibers spun from a titanium butoxide sol–gel and polyvinylpyrrolidone solution



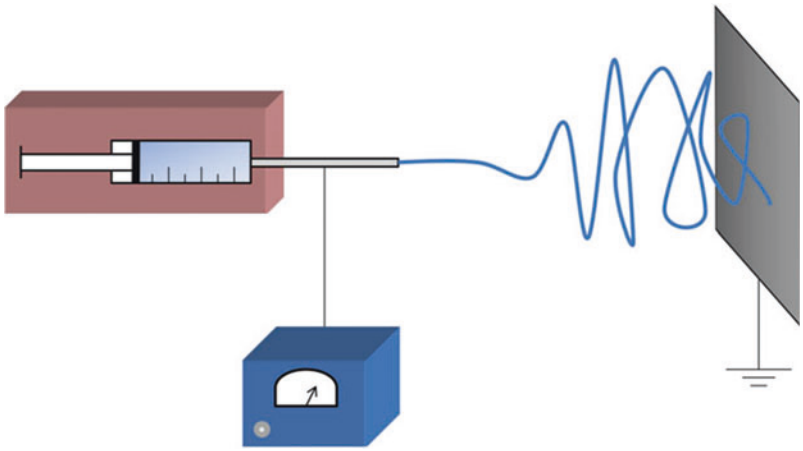
Electrospinning, Fig. 2 Three images showing droplet deformation and eventual fiber ejection as the electric field increases

While initial developments in the technique primarily dealt with polymers, further advances have demonstrated a diverse range of materials including ceramic, metallic, and composite systems. Additionally, this technique also provides control over the fiber structure and orientation.

The study of electrospinning begins with the field of electrohydrodynamics, which describes a range of phenomena that occur as liquids interact with electric fields. A single droplet when placed in an electric field deforms from a spherical droplet into an elongated Taylor cone [1]. With higher voltages, the droplet begins to eject fluid. Figure 2 shows three stages in droplet deformation as the electric field is increased. The droplet was formed from a highly viscous polyvinylpyrrolidone solution dissolved in water. Depending on the solution properties, two related phenomena occur: electrospraying and electrospinning.

Electrospraying forms discrete particles from nanometers to microns in diameter, whereas electrospinning forms long, continuous strands of fibers. However, a range of behaviors exists between these two extremes. In electrospinning, chain entanglements from high molecular weight polymers prevent the fluid from breaking apart. One of the great advantages of this process is that virtually any soluble polymer can be electrospun into nanofibers. As the strand accelerates through the field, it initially travels along a straight path. Due to instabilities in the electric field, its trajectory changes and the fiber whips around, spiraling toward nearby grounded surfaces. Although the study of electrohydrodynamics has existed for some time, electrospinning has only recently garnered much attention.

Electrospinning has several distinct advantages over other fiber processing techniques. Conventional fiber processing techniques are not readily capable of producing fibers with the nanoscale diameters that are characteristic of electrospun materials. Conversely, nanomaterial synthesis techniques can produce structures with similar diameters, but only at significantly shorter length scales. Furthermore, the electrospinning technique is applicable to an extensive range of material systems, while other nanomaterial growth methods are often highly specific to each material. The major drawback of the electrospinning technique is that the production rate of electrospun fibers is significantly slower than conventional techniques, typically on the order of milligrams per hour. Overall, the nanoscale dimensions and high aspect ratio of electrospun fibers creates many interesting avenues for research. The dimensions of these fibers



Electrospinning, Fig. 3 Typical laboratory scale electrospinning setup

greatly improve several of their properties such as porosity, surface area, and grain size (in the case of ceramic fibers). With these enhanced properties, electrospun fibers are applied to a number of disparate fields including filtration, cell growth, catalysis, energy storage and conversion devices, and hydrophobic surfaces.

Historical Developments in the Electrospinning Process

The first published use of the word electrospinning appeared in the literature in 1995. However, the study of charged fluids has existed for far longer. William Gilbert first published on the topic in 1600, where he describes the deformation of liquid droplets in an electric field [2]. Further insights into the field were reported by Lord Rayleigh who modeled the ability of charged fluids to eject liquid in 1882 [3]. The next major publications occurred in the 1964 by Sir Geoffrey Taylor. Taylor was the first to provide detailed descriptions of the deformation of liquid droplets [1]. To his credit, the distinctive shape of the deformed charged droplets has been given the name Taylor cones (or sometimes Gilbert-Taylor cones).

During this time, the industry also took notice of the field. Several important commercial developments in the field began in 1902. At this time, the US patent office recognized two methods for the dispersion of droplets via electrical means [4, 5]. Furthermore, in 1934, Anton Formhals patented the electrospinning process. Two years later, in 1936, the first commercial development of electrospun fibers began. I.V. Petryanov-Sokolov developed filters from electrospun fibers. Today, numerous companies currently manufacture electrospinning equipment and sell electrospun products.

Electrospinning Setup

Figure 3 shows a typical electrospinning setup. An advancement pump controls the flow rate of the polymer inside the syringe. Next, a power supply generates a charge on the liquid droplet. The electric field causes the droplet to deform and eject fluid. As the fluid travels, instabilities in the electric field cause the fiber to undergo a spiraling motion. This motion serves two purposes: It helps to thin the fiber, and it allows the solvent to evaporate prior to collecting on the substrate. The substrate is a grounded, conductive plate typically made from aluminum foil. Although this is the most commonly used setup, numerous modifications exist that can alter many of the properties of electrospun fibers.

Morphology of Electrospun Fibers

Porosity. As the fibers collect in randomly oriented meshes, the space left between fibers

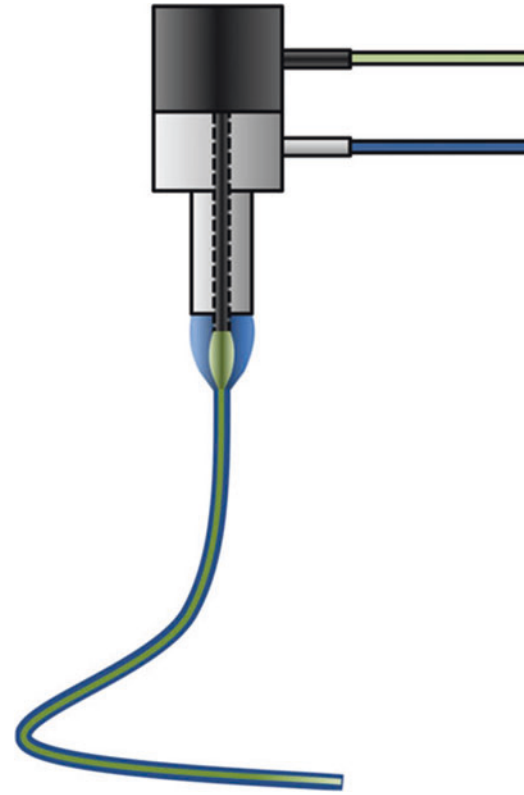
makes a highly porous structure. This porosity is dependent on the fiber diameter, and any alignment techniques that are used. The high surface area exhibited by electrospun mats is also due to the porosity. It is also possible to achieve porosity within individual fibers. Two methods are commonly used to produce porous fibers. One method exploits phase separation in a composite polymer system. Selectively removing one of the polymers after phase separation leaves behind a highly porous structure. Carefully adjusting the solvent, humidity, and molecular weight can also lead to highly porous surfaces with specific polymer systems [6]. In both cases, the fibers themselves are in the micron range, while the pores are on the nanoscale.

Core-shell and hollow fibers. Sun et al. used a novel electrospinning apparatus consisting of concentric capillaries to produce core-shell fibers [7]. Figure 4 demonstrates a typical core-shell electrospinning apparatus. Separate tubes feed material into each compartment of the device. In this case, two different polymers are used. The same device is also used to form hollow fibers. Pumping an immiscible oil through the inner capillary and subsequently removing the oil leaves behind hollow fibers [8].

Beaded fibers. Not all electrospun nanofibers form flat, uniform surfaces. Under certain conditions, large droplets called beads form on the electrospun strands. Several factors influence bead formation: low viscosity, low electric fields, and high surface tension. Therefore, adjusting the solution concentration, solvent, and ions present in the system can greatly influence the presence of beads [9].

Ribbons. These fibers do not have the circular cross section that is characteristic of nanofibers. Under certain conditions, a volatile solvent can induce the formation of a skin on electrospun fibers. As the material dries, the tubular structure collapses in on itself forming flattened ribbon structures [10].

Branched fibers. Short branches protruding from the sides of the nanofibers are reported due to perturbations in the cross-sectional areas of the fibers in polycaprolactone dissolved in acetone. These structures occur more frequently with large diameter fibers produced from highly viscous solutions in strong electrical fields [10].



Electrospinning, Fig. 4 Device for core-shell and hollow fiber production

Aligned fibers. Several different techniques produce aligned electrospun fibers [11]. The techniques either use a novel collection device or modify the electric or magnetic field to prevent the whipping motion. In one method, a high-speed mandrel replaces the standard collection plate. The rotating motion aligns the fibers around the mandrel. Another modification to the collector is to create a long gap in the collection substrate. The conductive sides of the gap attract the fibers, and residual charges within the fibers repel other fibers leading to parallel arrays collecting across the gap. A final technique involves using a magnetic field to prevent the whipping motion of fibers.

Methods to upscale production. The greatest limitation of the electrospinning process is the slow production rate. Increasing the fiber production output from a single needle is not possible.

However, it is possible to outscale the technique by assembling arrays of spinnerets in close proximity. This can result in numerous complications. Multiple spinnerets in close proximity cause electric field interference, which creates an uneven coating of fibers. Additionally, a single clogged needle in a large array could ruin the uniformity of the fiber mat. Several startup companies have begun developing other novel methods to increase fiber production. Some of these methods bypass the use of needles altogether. However, their efforts are beyond the scope of this entry.

Electrospun Materials

Polymers. The most frequently electrospun materials are polymers. Both synthetic and natural polymers are extensively reported for a variety of applications [12]. Due to the ability to use biocompatible materials and solvents, these materials are particularly interesting for biomedical and filtration applications. However, the electrospinning process is not limited to producing only polymeric fibers.

Ceramics. Several methods are able to produce ceramic fibers. The most common route to ceramic fibers is to incorporate a sol–gel material into a polymer. Heat treatments can burn out the polymer, leaving behind pure ceramic fibers. Similarly, incorporation of ceramic nanoparticles into electrospun polymer fibers followed by calcination is another pathway to ceramic fibers.

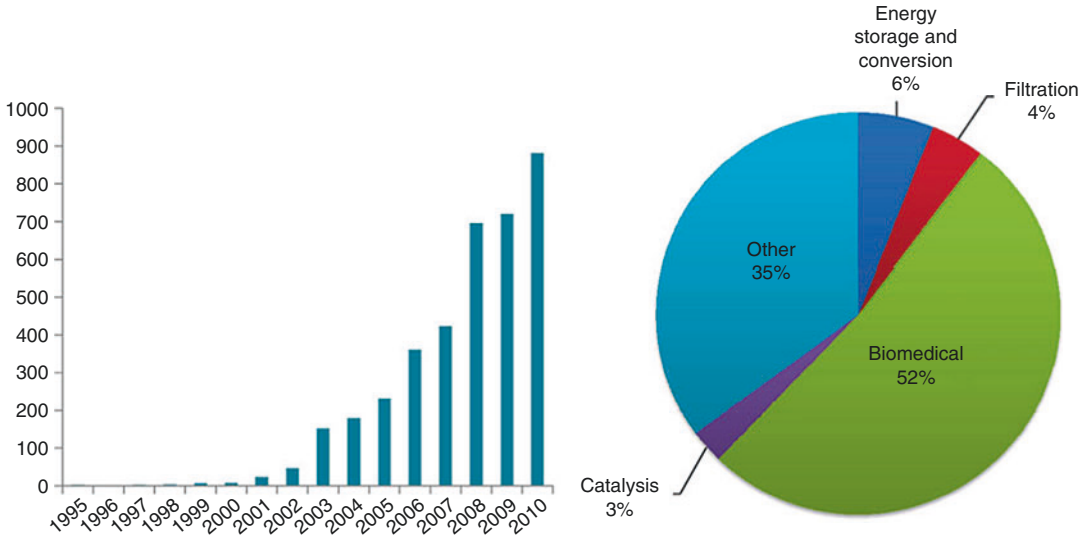
Metals. Only a few studies report purely metallic electrospun fibers. One method is to reduce ceramic nanofibers into metallic fibers. Another method uses electrospun fibers indirectly. Polymer nanofibers are coated in a thin metallic layer, and subsequently the polymer core is removed. This process leaves behind hollow metallic nanotubes. The potential advantages of metallic nanofibers for catalysis ensure that research into the area of metallic fibers will continue to expand.

Composites. As long as one of the materials is capable of being electrospun, the entire system will produce fibers. Due to this fact, a myriad of composite systems exist. These systems typically involve the incorporation of nanomaterials into the fibers. By carefully selecting the polymer and heat treatment conditions, the polymer can be turned into a carbon nanofiber support for the nanoparticles. Systems containing multiple polymers are also common. For the few polymers that prove intractable for electrospinning, a second polymer is often added creating a new composite system.

Analytical Models for Fiber Diameter

The unique properties of electrospun fibers stem from their nanoscale diameters. Therefore, it is of critical importance to understand the parameters that influence the fiber diameter. However, a complex relationship exists. The solution properties, Taylor cone shape and size, electric field, and whipping motion of the fiber all affect the fiber diameter. The solution properties (viscosity and surface tension) interact with gravity to form an equilibrium droplet shape. Appling an electric field deforms this shape and, at a sufficiently high voltage, ejects fluid from the droplet. Furthermore, adjusting the flow rate of the liquid through the capillary alters the size of the droplet. Finally, the radius of curvature as the jet whips toward the target further affects the final diameter.

Combining all of these terms, Fridrikh et al. proposed an analytical model to determine the final nanofiber diameter with an error of around 20 % for polymer fibers [13]. Equation 1 shows the relationship between the electrospinning parameters and the terminal fiber radius. The terminal radius depends on surface tension (γ), dielectric constant of the outside medium (ϵ), total current directed toward the lower electrode (I), radius of curvature divided by the diameter of the jet (χ), and flow rate (Q). This equation assumes that all solutions and voltages produce electrospun fibers, which is not the case. However, one can use the equation to determine which parameters have the greatest influence over the terminal fiber diameter. From the equation, it is apparent that flow rate and current are the most influential terms.



Electrospinning, Fig. 5 Electrospinning publications through 2010

$$r_{terminal} = \left(\gamma\varepsilon\frac{Q^2}{I^2}\frac{2}{\pi(2\ln\chi - 3)}\right)^{\frac{1}{3}} \quad (1)$$

Incorporation of ceramic precursors into polymer fibers greatly enhances the electrical conductivity by several orders of magnitude. For these highly conductive solutions, one must consider the conductivity due to the bulk charge and not limit the model to the surface charge alone. The high conductivity invalidates some of the critical assumptions in Eq. 1. Calcination further complicates the prediction of diameters in ceramic systems; fiber diameters shrink as calcination removes the polymeric material. However, the polymer loading and final fiber diameter is linearly related, and can be determined using a correction factor [14].

Properties of Electrospun Fibers

Mechanical properties. The mechanical properties of nanofibers can vary greatly depending on the materials involved and their processing conditions. The most common method to test fiber properties is to generate stress strain curves from large fiber mats. Measuring the properties of individual fibers is more challenging. To this end, several studies report various methods using atomic force microscopy measurements to determine the mechanical properties of individual fibers. The results of these studies suggest that the electrospinning process induces a molecular orientation within the samples, which increases the Young's modulus. The mechanical properties of ceramic fibers introduce special challenges. While polymer nanofibers are typically flexible, ceramic fibers are often extremely brittle. However, reports also exist of flexible ceramic fibers.

Applications and Publication Trends

Electrospinning applications span a range of diverse fields. Figure 5 shows the breakdown of publications according to their field from a title search on ISI web of knowledge. The first publications appeared in 1995, and the number of active researchers in the field has rapidly increased. The first publications in electrospun ceramic fibers appeared in 2003. Along with this development, publications in energy materials began to appear. The list of publications was categorized based on key terms relevant to each field. Papers that did not contain any of these terms are labeled "Other." This category includes research

into the fundamental concepts of electrospinning as well as modifications to the electrospinning setup to produce new structures and morphologies. The following sections outline how researchers utilize electrospun fibers in their respective fields.

Biomedical. Approximately half of electrospinning publications fall under the field of biomedical research. This is a broad area of research that comprises several subfields: scaffolds, wound healing, and drug delivery [15, 16]. Many biocompatible materials are electrospun for these applications. The porous nature of the fibers improves cell proliferation. Furthermore, by orienting the fibers, directed cell growth is possible. Drug delivery utilizes the high surface area to volume ratio of the fibers to increase the loading of drug molecules.

Energy materials. Recent developments in ceramic fiber processing have opened many new avenues in energy harvesting and conversion [17]. The one-dimensional structure of nanofibers confines grain growth, which can improve electrical conductivity. Researchers have assembled batteries, photovoltaics, hydrogen storage materials, capacitors, thermoelectric, and fuel cells from electrospun nanofibers. Some of these devices also make use of polymer fibers such as battery and fuel cell membranes.

Filtration. The highly porous network created by electrospun fibers makes them interesting filtration materials. As air flows through filters, there is a change in pressure. Higher drops in pressure require more energy to force air through the filtration devices. Assembling filters from thinner, uniform fibers, reduces the pressure drop, which lowers the energy requirements to force air through the filters. While most nanofilters are assembled from polymer fibers, ceramic fibers are also investigated due to their chemical and thermal stability [18].

Catalysis. The interconnected structure of electrospun fibers coupled with their high surface area makes electrospun fibers interesting catalyst materials. The connected fiber structure allows facile recovery of the catalytic material. Researchers have explored several methods to use nanofibers in catalysis. The most straightforward method is to embed catalytic nanoparticles in a polymer support material. Other studies have explored photocatalysis using ceramic fibers. Finally, a few researchers have heated polymer fibers embedded with metal salts to produce purely metallic nanofibers [19].

Super hydrophobic surfaces. Water repellent surfaces have many potential industrial applications, particularly due to the self-cleaning effect observed in super hydrophobic surfaces (surfaces with a water contact angle $>150^\circ$). A combination of high surface roughness and chemistry is necessary to achieve super hydrophobicity. By carefully controlling the morphology of electrospun fibers, a variety of super hydrophobic materials are reported [20]. These fibers often make use of the multiple scale of surface roughness created from beaded fibers.

Outlook

Electrospinning is an incredibly versatile nanomaterial processing technique. Using this technique, researchers can generate a wide range of materials and nanostructures from relatively simple, inexpensive equipment. These materials have the potential to purify air, heal wounds, deliver drugs, catalyze reactions, and produce and store clean energy. The full potential of electrospinning is just starting to be realized. As novel techniques are developed to enhance the output of electrospun fibers, industrial usage of the technique will increase. Further developments in this field will continue to unlock the potential of this technique to solve important scientific challenges.

Acknowledgments This work was supported by WCU (World Class University) program through the Korea Science and Engineering Foundation (R31-2008-000-10092).

Cross-References

- ▶ Lotus Effect
- ▶ Nanomaterials for Electrical Energy Storage Devices
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanomechanical Properties of Nanostructures

▶ Nanomedicine
▶ Nanostructures for Energy
▶ Scanning Electron Microscopy
▶ Sol–Gel Method
▶ TiO_2 Nanotube Arrays: Growth and Application

References

1. Taylor, G.: Disintegration of water drops in an electric field. Proc. R. Soc. Lond. A Math. Phys. Sci. **280**, 383–397 (1964). doi:10.1098/rspa.1964.0151
2. Gilbert, W., Mottelay, P.: De Magnete. Dover, New York (1958)
3. Rayleigh, L.: On the equilibrium of liquid conducting masses charged with electricity. Phil. Mag. **14**, 184 (1882)
4. Cooley, J.: Apparatus for electrically dispersing fluids. (1902)
5. Morton, W.: Method of dispersing fluids. Issued 29 July 1902
6. Casper, C.L., Stephens, J.S., Tassi, N.G., Chase, D.B., Rabolt, J.F.: Controlling surface morphology of electrospun polystyrene fibers: effect of humidity and molecular weight in the electrospinning process. Macromolecules **37**, 573–578 (2003). doi:10.1021/ma0351975
7. Sun, Z., Zussman, E., Yarin, A., Wendorff, J., Greiner, A.: Compound core-shell polymer nanofibers by co-electrospinning. Adv. Mater. **15**, 1929–1932 (2003)
8. Li, D., Xia, Y.: Direct fabrication of composite and ceramic hollow nanofibers by electrospinning. Nano Lett. **4**, 933–938 (2004). doi:10.1021/nl049590f
9. Fong, H., Chun, I., Reneker, D.H.: Beaded nanofibers formed during electrospinning. Polymer **40**, 4585–4592 (1999)
10. Reneker, D.H., Yarin, A.L.: Electrospinning jets and polymer nanofibers. Polymer **49**, 2387–2425 (2008)
11. Teo, W., Ramakrishna, S.: A review on electrospinning design and nanofibre assemblies. Nanotechnology **17**, 89 (2006)
12. Schiffman, J., Schauer, C.: A review: electrospinning of biopolymer nanofibers and their applications. Polym. Rev. **48**, 317–352 (2008)
13. Fridrikh, S., Yu, J., Brenner, M., Rutledge, G.: Controlling the fiber diameter during electrospinning. Phys. Rev. Lett. **90**, 144502 (2003)
14. Sigmund, W., et al.: Processing and structure relationships in electrospinning of ceramic fiber systems. J. Am. Ceram. Soc. **89**, 395–407 (2006)
15. Liang, D., Hsiao, B.S., Chu, B.: Functional electrospun nanofibrous scaffolds for biomedical applications. Adv. Drug Deliv. Rev. **59**, 1392–1412 (2007). doi:10.1016/j.addr.2007.04.021
16. Li, W.J., Mauck, R.L., Tuan, R.S.: Electrospun nanofibrous scaffolds: production, characterization, and applications for tissue engineering and drug delivery. J. Biomed. Nanotechnol. **1**, 259–275 (2005)
17. Laudenslager, M., Scheffler, R., Sigmund, W.: Electrospun materials for energy harvesting, conversion, and storage: a review. Pure Appl. Chem. **82**, 2137–2156 (2010). doi:10.1351/PAC-CON-09-11-49
18. Barhate, R., Ramakrishna, S.: Nanofibrous filtering media: filtration problems and solutions from tiny materials. J. Membr. Sci. **296**, 1–8 (2007)
19. Dersch, R., Steinhart, M., Boudriot, U., Greiner, A., Wendorff, J.: Nanoprocessing of polymers: applications in medicine, sensors, catalysis, photonics. Polym. Adv. Technol. **16**, 276–282 (2005)
20. Ma, M., Hill, R.M., Rutledge, G.C.: A review of recent results on superhydrophobic materials based on micro- and nanofibers. In: Superhydrophobic Surfaces, p. 241. Brill, Leiden (2009)

Electrostatic Actuation of Droplets

▶ Electrowetting

Electrostatic MEMS Microphones

Neal A. Hall
Electrical and Computer Engineering, University of Texas at Austin, Austin, TX, USA

Synonyms

MEMS capacitive microphone; MEMS condenser microphone; MEMS microphone; Silicon microphone

Definition

A microphone fabricated using surface and bulk silicon micromachining techniques and operating on the principle of a variable capacitance.

Applications and Background

Electrostatic MEMS microphones, like their larger counterparts, convert pressure waves in air into electrical signals. MEMS microphones were successfully commercialized in 2003 and began

competing with electret condenser microphones (ECMs) in consumer electronic device markets, most notably in cellular phones. Compared to ECMs which use heat-sensitive electret foils, MEMS microphones have a critical advantage in that they can withstand high temperatures encountered in standard lead-free solder reflow cycles used in automated pick and place manufacturing, and this in turn results in significant cost savings to system integrators [1]. Since their entry into the market, electrostatic MEMS microphones have become one of the highest growth areas for MEMS, growing from less than 300 million units shipped in 2007 to over one billion units shipped in 2011 [2]. Today many major semiconductor companies manufacture and sell electrostatic MEMS microphones including Knowles, Analog Devices, Infineon, and ST Microelectronics. The top four buyers of MEMS microphone in 2011 were Apple, Samsung, LG, and Motorola [3]. Apple's iPhone 4 product alone contains three electrostatic MEMS microphones – two in the body of the phone and a third in the mobile headset.

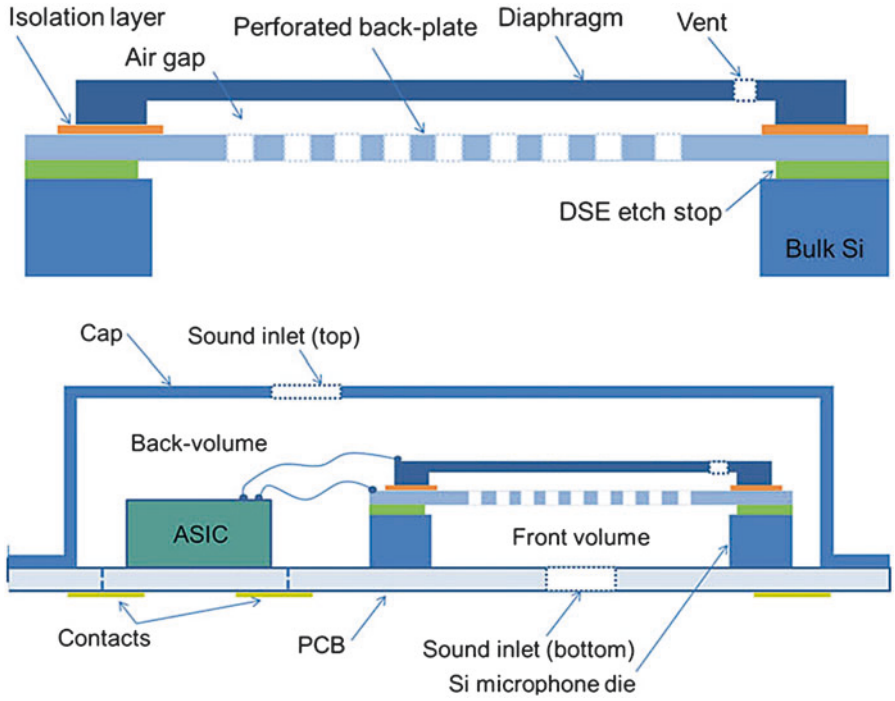
Operation and Construction

Field variables used to characterize sound include both pressure and particle vibrations. Although some new techniques attempt to measure particle velocity directly [4], the vast majority of microphones including electrostatic MEMS microphones are constructed as dynamic pressure sensors. Generally speaking, the range of sound pressures of interest in most applications is as low as 200 μPa root mean squared (rms) (20 dB) and as high as 10 Pa (114 dB). Electrostatic MEMS microphones must detect these small pressure signals in the presence of a large "DC" background atmospheric pressure, approximately 100 kPa. The frequency range typically of interest is the audio bandwidth, 20 Hz–20 kHz, although only recently have commercial models of MEMS microphones approached these limits. The corresponding wavelength range is 17 m–17 mm. For most of this range, MEMS microphones packages are significantly smaller than the wavelength of sound being measured.

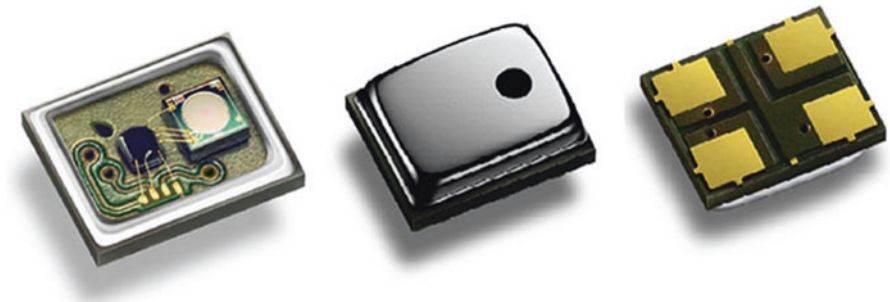
Figure 1 details important features of a silicon microphone die and also a complete surface mount electrostatic MEMS microphone package. Referring to the top image, a compliant, pressure-sensitive diaphragm is suspended over a rigid and perforated backplate to form a variable capacitor. When sound pressure is applied to the diaphragm, the diaphragm moves vertically and the resulting change in the parallel-plate capacitance is detected. The capacitive structure must be electrically biased and loaded with charge, and the high output impedance of the signal must be buffered with an amplifier. These features are implemented by a small application specific integrated circuit (ASIC) as shown in Fig. 1 bottom. To sense sound pressure effectively, one side of the diaphragm must be exposed to the incoming sound pressure while the opposing side must be sealed off from the environment. A common package to accomplish this is also shown in Fig. 1. The MEMS die and ASIC are both mounted on a common printed circuit board (PCB) which contains a small opening on the bottom surface to let sound pressure in. A sealed back volume is created by a cap mounted directly to the PCB. Electrical vias through the PCB lead to contacts on the bottom side to create a complete surface-mountable package.

As a variation to the system shown in Fig. 1 (bottom), the sound inlet is sometimes placed on the cap rather than the PCB to form a "top-inlet" configuration. A second common variation found on the MEMS die is a backplate fabricated above rather than below the diaphragm. Photographs and SEMs of a commercial electrostatic MEMS microphone that implements both of these variations are presented in Figs. 2 and 3, respectively. In the center image of Fig. 2, one can observe the sound inlet on the cap of the device. The SEMs in Fig. 3 highlight the perforated backplate suspended approximately 2 μm above the pressure sensitive membrane.

A final important feature to note in Fig. 1 is the presence of a small low-frequency vent, most commonly fabricated into the microphone diaphragm. The vent provides the device with a high flow resistance leakage path enabling sound pressure to bypass the diaphragm and reach the back volume. This feature creates a lower limiting frequency and prevents the diaphragm from



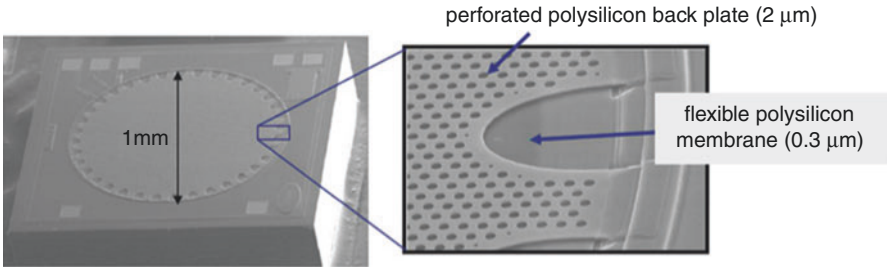
Electrostatic MEMS Microphones, Fig. 1 (*Top*) Schematic of a silicon MEMS die containing the sensing structure and (*bottom*) a schematic of a complete electrostatic MEMS microphone surface-mountable package. The package contains either a *top* or *bottom* sound inlet but typically not both



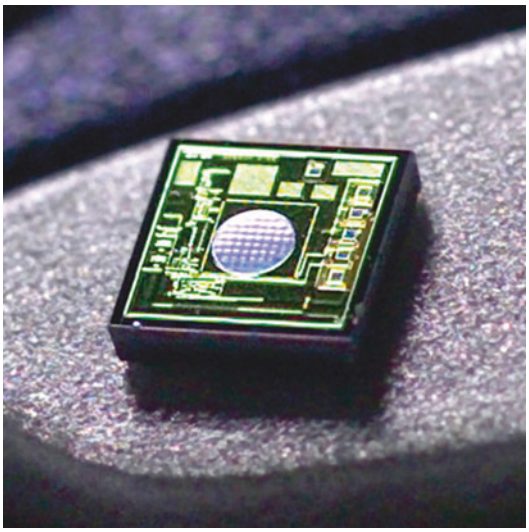
Electrostatic MEMS Microphones, Fig. 2 Photographs of a commercial MEMS microphone. This device has a footprint of 4.72 × 3.76 mm and a profile of 1.25 mm. The image at *left* is with the cap removed and shows details of the PCB, ASIC, and MEMS die. The figure at far *right* shows the electrical contacts on the *bottom* of the package that interface with a customer's system (Images borrowed courtesy of Infineon and first appearing in Ref. [5])

deforming and responding to DC atmospheric pressure and low-frequency barometric pressure fluctuations, which are orders of magnitude larger than the audible sound pressures that the device is designed to sense.

Although Fig. 1 presents the most common construction of MEMS microphones, single-chip architectures have also been demonstrated combing the MEMS capacitive sensing structure with the CMOS electronics on the same die. At least one commercial embodiment of this approach exists. Figure 4 below presents a photograph of a microphone made by Akustica (acquired by Bosch in 2009) in which one can observe the pressure-sensitive diaphragm structure along with electronics. The total die size is



Electrostatic MEMS Microphones, Fig. 3 Scanning electron micrographs (SEMs) of a silicon die containing an electrostatic MEMS sensing structure (Images borrowed courtesy of Infineon and first appearing in Ref. [5])

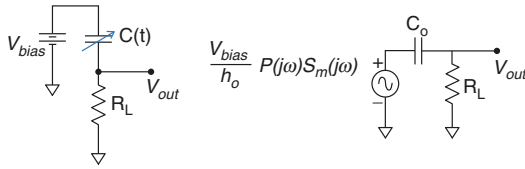


Electrostatic MEMS Microphones, Fig. 4 Micrograph of a single-chip electrostatic MEMS microphone made by Akustica. The sensing diaphragm can be seen in the center of the structure. This die measures approximately 1 × 1 mm (Image borrowed courtesy of Akustica, Inc. [6])

1 × 1 mm. An obvious advantage of this approach is size, with space being a premium in small consumer electronic devices (e.g., cell phones and laptop computers). A second advantage cited is potentially lower per unit cost. The MEMS structure is fabricated in layers already required in the CMOS process, so only few additional processing steps are required for the MEMS. The peripheral packaging requirements are also reduced. A third advantage cited is electronic design flexibility owing to the proximity between the sensing capacitance and the electronics, which reduces parasitic capacitance associated with bond pads and wire bonds that are otherwise required to route signals to off chip electronics as shown in the two-chip embodiment of Fig. 1. This proximity may also be advantageous for shielding against electromagnetic interference created by neighboring electronics in consumer products.

Transduction Principle

The variable capacitance transduction principle is presented in Fig. 5. The capacitive sensing structure is biased through a high impedance resistance, R_L, shown at left. Electrostatic forces pull the diaphragm in toward the backplate to create a standoff distance h_o typically between 1 and 2 μm and a nominal parallel-plate capacitance C_o. A phenomena known as electrostatic collapse of the diaphragm limits the bias voltage that can be applied [7], which is typically in the range of 2–11 V and depends on the design [1]. The instantaneous voltage across the parallel-plate capacitor is given by $v = q/C = qh/(\varepsilon_{oA})$, where C is the instantaneous capacitance, q is the instantaneous charge, h is the instantaneous backplate-diaphragm standoff distance, and A is the area of the structure. Small diaphragm vibrations create small changes in h, and the resulting small signal voltage generated about the bias point is therefore $dv = q_o/(\varepsilon_{oA}) \quad dh = (v_{bias}/h_o)dh$. The output impedance associated with the small signal voltage is the device capacitance itself, C_o, and the small signal equivalent circuit is that shown in Fig. 4 (right), with dh replaced by the product of the incoming acoustic pressure $P(j\omega)$ and the mechanical sensitivity of the device, $S_m(j\omega)$, defined below.



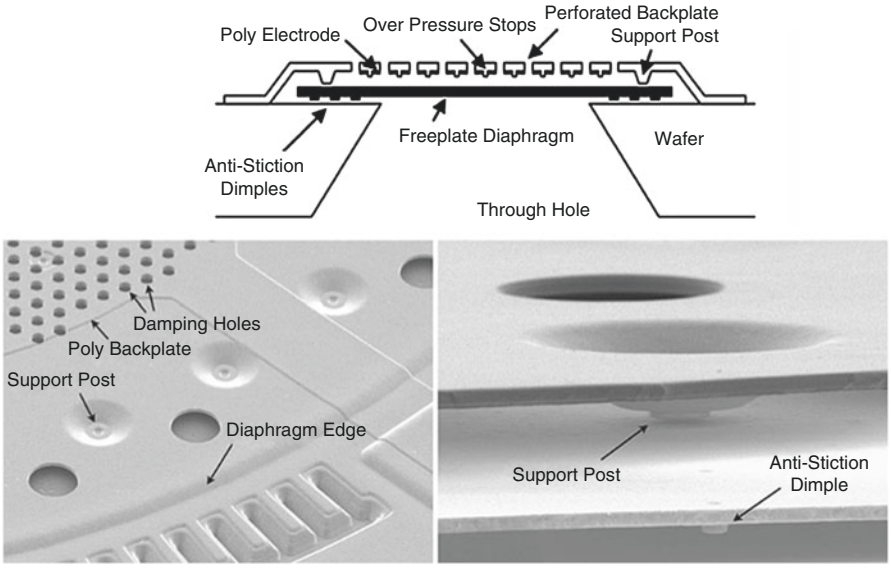
Electrostatic MEMS Microphones, Fig. 5 (*Left*) Circuit schematic for biasing an electrostatic MEMS microphone and measuring changes in capacitance, and (*right*) small signal equivalent circuit about the bias point

vb_{ias}/h_o is commonly referred to as an electrical sensitivity and has units of volts per meter of diaphragm displacement. It is also the electric field in the diaphragm-backplate gap created by application of the bias voltage. $S_m(j\omega)$ is a transfer function relating the displacement of the diaphragm to incoming sound pressure. As shown in Fig. 5 right, the small signal voltage generated is the product of the incoming pressure phasor and the electrical and mechanical sensitivities. The small capacitance of the device creates a high output impedance, and v_{out} in Fig. 4 is therefore followed by an amplifier with high input impedance such as a JFET [8]. Digital microphones are also common, in which case the buffered analog output is followed by a sigma-delta analog to digital converter.

Microfabrication of Sensing Structure

The structure shown in Fig. 1 (top) can be fabricated using many different fabrication process flows with many options for materials. Almost all electrostatic MEMS microphones use thin surface micromachined layers for the backplate and diaphragm, and a deep silicon etch (DSE) bulk micromachining step to form the front volume. The surface machined layers are typically between 100 nm and 2 μm thick. Figure 1 (top) suggests a simple fab flow option, in which the first step is to deposit or grow a material with high selectivity against a silicon deep reactive ion etch (DRIE) process, such as silicon dioxide. The first structural layer is then deposited and patterned to form the backplate. Polysilicon is a common material choice. A thin isolation layer (commonly silicon nitride) is deposited and etched to provide electrical isolation between the diaphragm and backplate, followed by deposition and etching of the sacrificial layer to form what will become the air gap. Low temperature oxide (LTO) and TEOS oxide are common choices for the sacrificial layer. After deposition and patterning of the diaphragm, which is commonly formed using polysilicon, the surface steps are completed and the device is etched through the silicon wafer from the opposing side up to the DSE etch stop. Both isotropic etches (e.g., DRIE) and anisotropic wet etches (e.g., KOH) are used in commercial versions of the technology [1]. The sacrificial layer(s) at the surface are then removed and the wafers are diced to realize completed die. Some commercial processes use a silicon on insulator (SOI) wafer as a starting point and use the epitaxial Si layer for formation of the microphone backplate [9].

As electrostatic MEMS microphones continue to mature into commercial device applications, several innovations at the microfabrication level continue to be driven by reliability and device-to-device repeatability criteria. For thin (1–2 μm thick) clamped diaphragms employed in MEMS microphones, even small residual film stress resulting from the deposition process can have a major influence on the compliance of the sensing diaphragm which in turn directly affects the total sensitivity of the device measured in volts/Pa. Controlling film stress in microfabrication processes to ensure device-to-device repeatability has practical limitations. Further, changing environmental conditions (e.g., temperature) can affect sensitivity output of a single device during the course of day-to-day operation. This has motivated the development of innovative structures designed for immunity of output sensitivity to these variations. Figure 6 presents a device by Knowles in which the sensing diaphragm is not clamped, but rather is freely floating until pulled into contact with the support posts by electrostatic forces upon biasing. By allowing the diaphragm to float freely rather than clamping and anchoring it to the substrate, zero diaphragm tension is achieved regardless of film stresses introduced in the deposition process. The free plate design is achieved by completely surrounding the plate with sacrificial layers rather than anchoring it to the substrate during the processing sequence.

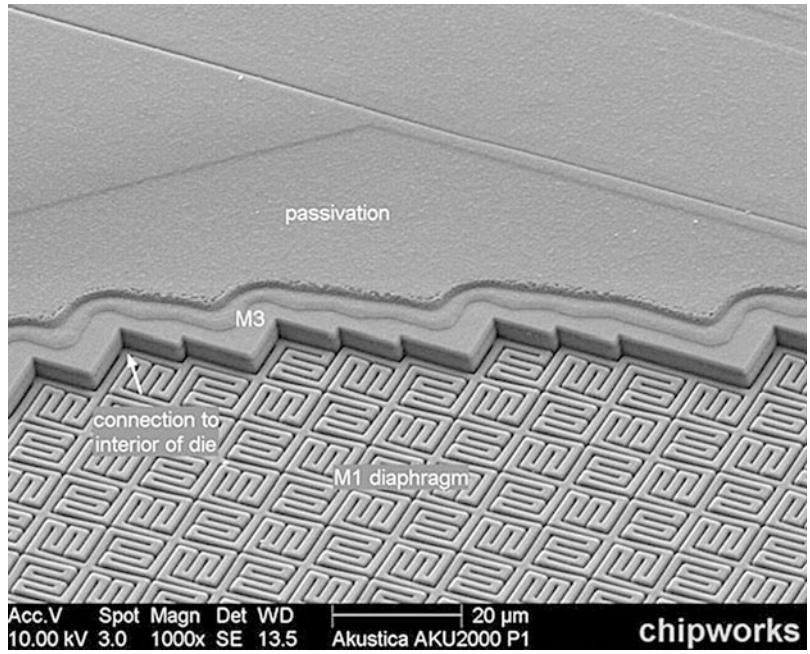


Electrostatic MEMS Microphones, Fig. 6 A schematic and SEMs highlighting a free diaphragm design by Knowles (Image borrowed courtesy of Knowles Acoustics and originally appearing in Ref. [1])

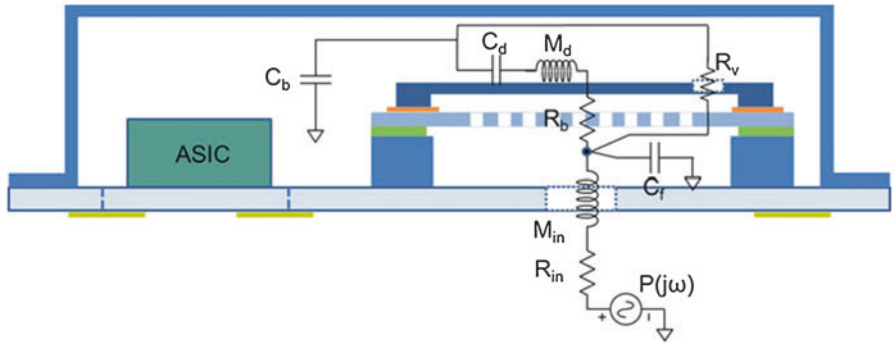
The sensing membrane in the single-chip microphone developed by Akustica is made using metal interconnect layers that are already part of standard CMOS processes [10]. Specifically, a metal layer is patterned to form a perforated screen which becomes buried in oxide layers during the subsequent CMOS processing steps. Upon completion of the CMOS process flow, the oxide on top of the wafer is etched down to the screen. The perforations in the screen then allow silicon etchant to reach the Si wafer beneath the screen to create a cavity behind the screen. This cavity is analogous and identical in function to the back volume labeled in Fig. 1. To create a sealed diaphragm, a conformal polymer is deposited to fill the screen perforations. Figure 7 shows an SEM of the metal screen layer forming the diaphragm. The interesting serpentine pattern of the screen in this design is aimed at removing bowing due to stresses inherent in the metal.

Dynamic Model of Package

More so than many other MEMS and Nanosystems, packaging has a profound effect on the overall dynamics of electrostatic MEMS microphones. The use of impedance-based network analogies such as the one presented in Fig. 8 are pervasive in the study of MEMS microphones [11]. Such models are based on the observation that acoustical and mechanical systems obey rules analogous to Kirchhoff's voltage and current laws. In acoustical impedance networks, pressure is a node variable and the sum of pressure drops across elements in a closed loop is equal to zero. Similarly, volumetric flow of air entering and leaving nodes in the network is conserved. In such models, acoustic pressure is a potential or effort variable and therefore analogous to voltage in electrical circuits. Acoustic volume velocity (with units of m^3/s) flows through various elements in the circuit and is therefore analogous to current. Cavities or volumes in acoustical systems have the ability to store potential energy by storing compressed air and are analogous to capacitors in electrical networks, while small inlets and outlets force air to accelerate upon passage through and therefore store energy by virtue of momentum, analogous to an inductor in an electrical network. Similarly, the mechanical compliance and mass of the diaphragm are represented by capacitors and inductors, respectively, with appropriately computed values. The backplate resistance R_b plays a dominant role in shaping the frequency response of the system and also in the noise analysis as



Electrostatic MEMS Microphones, Fig. 7 SEM showing the diaphragm layer and anchoring in Akustica's CMOS microphone (Image borrowed courtesy of MEMS Investor Journal, Inc. and Chipworks)

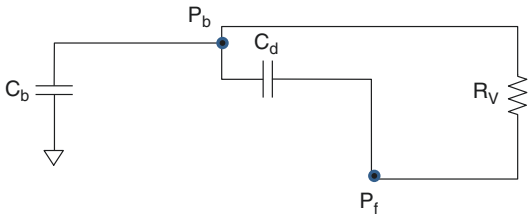


Electrostatic MEMS Microphones, Fig. 8 Network model superimposed on the physical package schematic to emphasize the important package dynamics in electrostatic MEMS microphones

discussed below. R_b is a flow resistance generated by the movement of air through the backplate perforations as the diaphragm vibrates relative the stationary backplate. Network models provide a way to combine mechanical, acoustical, and electrical features of the device into a unified model that can be simulated with common circuit simulator tools such as SPICE.

Figure 8 presents a common network model superimposed on the physical device structure. Subscripts *in*, *f*, *b, d* and *v* refer to elements associated with the inlet, front cavity, back cavity, diaphragm, and vent, respectively. The $S_m(j\omega)$ transfer function referenced in Fig. 5 relating an incoming pressure phasor to motion of the diaphragm may be modeled and computed based on the network model in Fig. 8.

Insight into the dynamics of the complete package arises from analysis of the network model in the low-frequency limit, the pass band, and near the fundamental resonance of the system. The model makes clear how the vent feature represented by R_v creates a lower limiting frequency by creating a low-frequency bypass around the diaphragm compliance C_d. To see this further, the section of the model relating the



Electrostatic MEMS Microphones, Fig. 9 Section of the low-frequency equivalent circuit of Fig. 8 relating the front volume pressure to the back volume pressure

front cavity pressure, P_f, to the drive pressure across the diaphragm ($P_f - P_b$) is shown in Fig. 9. A low-frequency approximation is shown and M_d and R_b are therefore not included.

For microphones packaged as shown in Fig. 8, C_b is often significantly larger than C_d. The low-frequency response of the microphone is then characterized by a simple first-order high-pass filter, with a zero at $\omega = 0$, and a pole or break frequency at $\omega = 1/(R_v C_b)$, above which the microphone operates in a flat band with $S(j\omega)$ dominated by the series compliance $C_b + C_d$. A fundamental resonance of the system characterized by two complex conjugate poles is common of electrostatic MEMS microphones. This LRC type resonance arises from the series combination of the inertial elements $M_d + M_{in}$, the compliance elements, $C_d + C_b$, and the loss elements $R_b + R_{in}$. This fundamental resonance is typically designed to be close to or above the desired upper limiting frequency of the microphone.

Dominant Noise Sources, dB(A) Weighting, SNR Definition

Signal-to-noise ratio is an important figure of merit for small acoustic transducers, defining the realm of applications in which a particular device technology is suited. Noise floors for electrostatic MEMS microphones are most meaningful when referred to an input pressure level. The input-pressure-referred-noise spectrum is commonly integrated using an A-weighted filter which takes into account the frequency response of a healthy human ear. The sound pressure level (SPL) of the A-weighted rms pressure level is computed and a dB(A) noise level is reported for the device. Common noise levels for MEMS microphones are around 34 dB(A). Often a signal-to-noise ratio (SNR) is reported, and in these cases the convention is to assume a 1 Pa rms, or 94 dB SPL, input. SNR is then defined by the following expression: SNR = 94 dB − A weighted noise. Additional performance-based figures of merit for MEMS electrostatic microphones include dynamic range, upper limiting frequency, and lower limiting frequency.

Scaling Challenges for Small-Scale Electrostatic Microphones

Electrostatic MEMS microphones face challenges in achieving low noise floors, and this has limited their ability to address applications such as hearing aids which require noise floors in the 20–25 dB (A) range. A fundamental noise limit in miniature microphones is thermal-mechanical noise due to the dissipative mechanisms in the acoustical and mechanical parts of the system [12]. In particular, the backplate resistance R_b labeled in Fig. 5 generates a thermal-mechanical pressure noise contribution that is seen at the input of the device across the pass band of operation and therefore impacts the input-pressure-referred noise directly. P_{tm}, the thermal-mechanical noise due to the mechanical flow resistance R_b, can be computed as:

$$P_{tm} = \frac{1}{A_d}\sqrt{4k_bTR_b} = \frac{1}{A_d}\sqrt{4k_bTR_{ba}A_d} \propto \sqrt{\frac{R_{ba}}{A_d}} \propto \frac{\sqrt{R_{ba}}}{a}$$

where k_b is Boltzmann's constant, T is the ambient temperature, A_d and a are the area and radius of the diaphragm and backplate (assumed to the same), and R_{ba} is the mechanical flow resistance per unit area. R_{ba} is a function of the perforation ratio and perforation geometry [13] but not the size of the backplate itself since R_{ba} is defined on a per unit area basis. For a given backplate design, decreasing the diaphragm radius by a factor of 2 results in a doubling (i.e., 6 dB) in the input-referred thermal noise limit. While backplates can be redesigned to achieve lower R_{ba}, this path faces practical limitations as higher perforation ratios reduce the total active capacitance of the device. Further, the lack of damping in the system can

result in an undesirably high resonance Q (i.e., small damping ratio) at the fundamental resonance of the system discussed above. In the pass band, the device behaves as a second-order system with a damping ratio:

$$\varsigma = \frac{R_{ab}A_d}{2M\omega_n} \propto \sqrt{\frac{R_{ba}}{\rho t \omega_n}}$$

where M is the combined inertia of the diaphragm and sound inlet, ω_n is the fundamental radial natural frequency of the system, and ρ and t are the density and thickness of the diaphragm. Within limits dictated by fabrication robustness, the latter two variables provide a means for achieving a desired damping ratio for an R_{ba} predetermined by noise considerations.

In addition to the above mechanical and acoustical scaling considerations, the lower cutoff frequency of the microphone may also be limited by the high-pass nature of the small signal electrical circuit shown in Fig. 5 (right). This electrical filter is independent of the lower limiting frequency inherent to $S_m(j\omega)$ discussed above due to the package dynamics. For a given bias resistor, the lower limiting frequency of the microphone will increase with reduced device capacitance. Although in principle GΩ bias resistors can be used, preventing leakage across such elements amid all required environmental operating conditions can prove challenging.

Other Techniques and Future Directions

Other types of MEMS microphones include piezoelectric and optical. The former have shown to be advantageous for special applications demanding high linearity and high dynamic range (e.g., aero-acoustic measurement arrays and sensors for wind tunnel testing) [14], while recent work with the later work aims to achieve significantly lower noise floors than that achievable with electrostatic MEMS microphones [15]. Active research is also aimed at demonstrating designs that are inherently directional. An innovation inspired by the directional hearing mechanism of a special parasitoid fly has recently been demonstrated [16]. Until recently, little attention was given to SNR as an important parameter for MEMS microphones in the consumer electronics commercial sector. Recently, however, manufacturers of electrostatic MEMS microphones have started introducing lower noise models, with the lowest noise models to date having input-referred noise levels equal to 29 dBA. This may be driven by a recent trend in smartphone applications: the use of multiple microphones in a single product for implementation of ambient noise cancelation algorithms. As a significant departure from traditional close-talking voice input, hands-free and headset-free application modalities may continue to drive performance increases and new innovations for electrostatic MEMS microphones.

Cross-References

- ▶ CMOS MEMS Fabrication Technologies
- ▶ Integrated Micro-acoustic Devices
- ▶ MEMS Packaging

References

1. Loeppert, P.V., Lee, S.B.: SiSonicTM – the first commercialized MEMS microphone. Presented at the Solid-state sensors, actuators, and microsystems workshop, Hilton Head Island (2006)
2. Tekedia.com: MEMS microphones break the billion unit barrier. http://tekedia.com/34372/mems-microphones-break-billion-unit-barrier/ (2012, January). Accessed 1 Feb 2012
3. Lowensohn, J.: Apple bests Samsung as buyer of tiny mics. http://news.cnet.com/8301-13579_3-57362174-37/apple-bests-samsung-as-buyer-of-tiny-mics/ (2012, January 19). Accessed 1 Feb 2012
4. De Bree, H.-E., Leussink, P., Korthorst, T., Jansen, H., Lammerink, T.S.J., Elwenspoek, M.: The μ-flown: a novel device for measuring acoustic flows. Sens. Actuators A Phys. **54**, 552–557 (1996)
5. Fuldner, M., Dehe, A.: Challenges of high SNR (signal to noise) silicon micromachined microphones. Presented at the 19th international congress on acoustics, Madrid (2007)
6. Akustica, Inc. New microphones boost acoustic performance in consumer electronics. http://www.akustica.com (2011). Accessed 1 Feb 2012
7. Osterberg, P., Yie, H., Cai, X., White, J., Senturia, S.: Self-consistent simulation and modelling of

electrostatically deformed diaphragms. In: Proceedings of the IEEE Micro Electro Mechanical Systems. An Investigation of Micro Structures, Sensors, Actuators, Machines and Robotic Systems, pp. 28–32, Oiso (1994)
8. Brauer, M., Dehe, A., Fuldner, M., Barzen, S., Laur, R.: Improved signal-to-noise ratio of silicon microphones by a high-impedance resistor. J. Micromech. Microeng. **14**, S86 (2004)
9. Weigold, J.W., Brosnihan, T.J., Bergeron, J., Zhang, X.: A MEMS condenser microphone for consumer applications. Presented at the 19th IEEE international conference on micro electro mechanical systems, Istanbul (2006)
10. Neumann, J.J.J., Kaighman, G.: CMOS-MEMS membrane for audio-frequency acoustic actuation. Sens. Actuators A Phys. **95**, 175–182 (2002)
11. Beranek, L.: Acoustics. McGraw-Hill, New York (1954)
12. Gabrielson, T.B.: Fundamental noise limits for miniature acoustic and vibration sensors. J. Vib. Acoust. **117**, 405–410 (1995)
13. Homentcovschi, D., Miles, R.N.: Modeling of viscous damping of perforated planar microstructures. Applications in acoustics. J. Acoust. Soc. Am. **116**, 2939–2947 (2004)
14. Horowitz, S., Nishida, T., Cattafesta, L., Sheplak, M.: Development of a micromachined piezoelectric microphone for aeroacoustics applications. J. Acoust. Soc. Am. **122**, 3428–3436 (2007)
15. Kuntzman, M., Garcia, C., Onaran, G., Avenson, B., Kirk, K., Hall, N.: Performance and modeling of a fully packaged micromachined optical microphone. J. Microelectromech. Syst. **20**, 828–833 (2011)
16. Miles, R.N., Su, Q., Cui, W., Shetye, M., Degertekin, F.L., Bicen, B., Garcia, C., Jones, S.A., Hall, N.A.: A low-noise differential microphone inspired by the ears of the parasitoid fly *Ormia ochracea*. J. Acoust. Soc. Am. **125**, 2013–2026 (2009)

Electrostatic RF MEMS Switches

► Capacitive MEMS Switches

Electrothermal Actuators

► Thermal Actuators

Electrothermomechanical Actuators

► Thermal Actuators

Electrowetting

C. J. Kim
Mechanical and Aerospace Engineering Department, University of California, Los Angeles (UCLA), Los Angeles, CA, USA

Synonyms

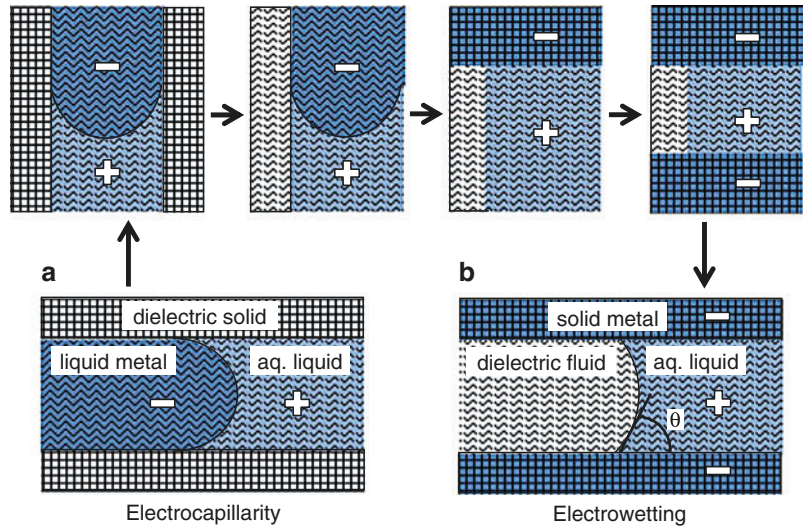
Digital microfluidics; Droplet microfluidics; Electrocapillarity; Electrostatic actuation of droplets; Electrowetting-on-dielectric (EWOD)

Definition

Making a surface more wetting to a liquid by applying voltages.

Introduction

When a material (typically solid) and a liquid are in contact, the application of an electric potential between them may cause the wettability of the material to increase, which is exhibited by a decrease of the observed contact angle. This phenomenon is called electrowetting – a term reminiscent of the more traditional electrocapillarity. In recent years, the development of various electrode and material configurations for technological applications – mostly in the rapidly expanding field of microfluidics – has given rise to additional terms such as electrowetting-on-dielectric (EWOD). Because technologies relying on the actuation of liquids by applied electrical potentials are still new, the terms used to describe such systems can be difficult to delineate. It will therefore be helpful to introduce a few terms related to electrowetting with a little bit of the history and theoretical background behind each of them. While the basic approaches to utilizing the electrowetting phenomenon to obtain physical results are presented here, applications of electrowetting to manipulate droplets or to develop devices



Electrowetting, Fig. 1 Comparison between electrocapillarity (**a**) and electrowetting (**b**), drawn to maintain the original configurations presented by Beni and Hackwood [2]. A *grid pattern* implies a solid, and a *wave pattern* implies a fluid. A *white* background implies a dielectric, and a *gray* (*blue* if seen in color) background implies a conductor

and systems are deferred to other related entries as well as other works referenced later in this entry.

Electrocapillarity and Electrowetting

A typical liquid (e.g., water) rises in a vertical glass capillary against gravity by the intermolecular attraction between the liquid and the inner wall of the capillary. A terminal height is reached when the gravity of the liquid column equals the attractive force or, more precisely, the vertical component of the liquid-air interfacial tension. The vertical component is larger if the liquid wets the glass surface better (e.g., the surface is clean). The degree of wetting is expressed by the contact angle θ of the liquid on the solid surface by the Young equation:

$$\gamma \cos\theta = \gamma_{sg} - \gamma_{sl} \tag{1}$$

where γ, γ_{sg}, and γ_{sl} are the interfacial energy (or interfacial tension, surface tension) between liquid and air (gas), solid and air, and solid and liquid, respectively. On the other hand, if the liquid does not wet the glass (e.g., mercury), i.e., the contact angle is greater than 90°, the liquid falls down in the capillary. A terminal falling depth against gravity is determined by the surface tension and the contact angle.

In 1875, Gabriel Lippmann showed that a mercury column in a glass capillary, which is dipped into an electrolyte bath, rises or falls when a voltage is applied between the mercury and the electrolyte [1]. This experiment, which is called the Lippmann electrometer, demonstrated that the interfacial tension between the mercury and the electrolyte is a function of the electric potential across the interface. Perhaps termed originally to depict the initial experimental configuration involving a capillary tube, *electrocapillarity* nevertheless describes a general phenomenon of interfacial energy changing by electric fields.

In 1981, Beni and Hackwood [2] introduced the term *electrowetting* to describe a manifestation of electrocapillarity that occurs when the configuration of materials and electrodes is somewhat different than that of the classical Lippmann electrometer. Specifically, in the traditional electrocapillarity measurement, voltage is applied between a metallic *liquid* (i.e., mercury) and an aqueous liquid (i.e., electrolyte), while both materials are surrounded by an insulating *solid* (i.e., glass capillary). For electrowetting, on the other hand, voltage is applied between a metallic *solid* and an aqueous liquid surrounded by an insulating *fluid* (e.g., air or oil). Figure 1 schematically illustrates these two phenomena, which operate under the same general principle but are distinguishable by how the materials are configured.

Readers may realize that, despite their apparent differences, electrocapillarity and electrowetting describe essentially the same principle. One can do a mental exercise that involves dictating phase changes of a given material. Rotate Fig. 1a 90° clockwise; remove the right dielectric solid and convert the left dielectric solid into a fluid (i.e., vapor or liquid); and convert the liquid metal into a flat solid metal and duplicate it at the bottom. Now one has Fig. 1b. So, the two terms describe one phenomenon depending on the users' focus of interest. If the interest is a liquid column moving in a solid capillary, electrocapillarity describes the phenomenon well. If the interest is how well a liquid wets a solid surface, e.g., what is the contact angle, electrowetting is a better description.

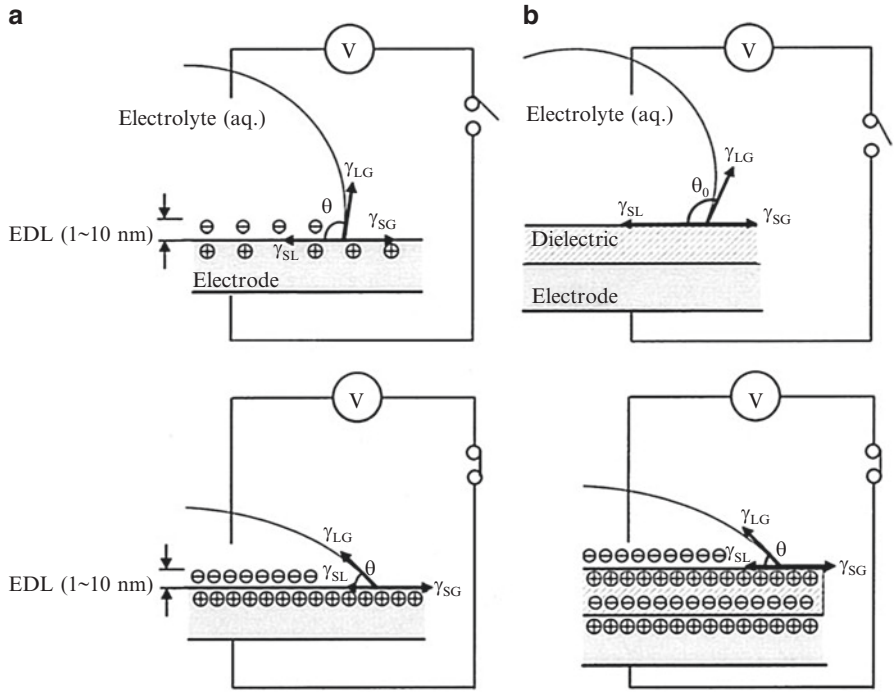
Some readers may raise another question, however. There exists a contact angle of the liquid electrolyte on the solid surface in the electrocapillarity of Fig. 1a, and there exists a column of the liquid electrolyte in a metallic capillary in the electrowetting of Fig. 1b. The confusion may be avoided if one notes that the main interest should be in the interface across which the electric potential is applied. In Fig. 1a, the liquid-liquid interface, across which the voltage is applied, travels in a capillary, exhibiting capillary action. In Fig. 1b, the solid–liquid interface, across which the electric potential is applied, forms a contact angle, exhibiting wetting.

Electrowetting and Electrowetting-On-Dielectric (EWOD)

Historically, for both the electrocapillarity and the electrowetting phenomena, the main interest was the interface between an aqueous liquid and a metal. Whether the metal was a liquid (for electrocapillarity) or a solid (for electrowetting), the electric double layer (EDL) at the interface played a critical role. Since the underlying mechanism for both phenomena was that the free interfacial energy can be modulated by increasing or decreasing the capacitive energy stored at the EDL via an electric potential across it, the mechanism would be valid only while the EDL functions as an insulating capacitor, i.e., below ~1 V range in practice. An applied voltage above such a limit would induce electrochemical reactions, such as electrolytic gas generation, and alter the surface. In the case of electrowetting, it was difficult to induce contact angle changes large enough to be useful without degrading the surface. Promising utilities were limited to the case of a liquid metal in an electrolyte-filled capillary, e.g., [3, 4].

Building on the findings of Minnema et al. [5], in 1993 Berge [6] reported that it is possible to perform electrowetting on a dielectric material when it is placed (as a solid film) between the liquid and the electrode. In a nutshell, the dielectric film functions as the primary energy-storing capacitor, thereby serving the same function as the EDL in conventional electrowetting. Since it required a very high voltage for electrowetting to be noticeable on an insulator, it is fitting that the first such observation was made during investigations of "water trees" (degradation structure grown in a polymer due to humidity and an electric field) in high-voltage cables [5]. Several kV were commonly used in the early reports. Since reducing the operating voltage range was simply a matter of making the dielectric layer thinner and of higher quality materials, the thin-film techniques of microelectromechanical systems (MEMS) played a timely role in helping experimenters reduce the dielectric thickness from the millimeter range [5, 6] to the micrometer range [7]. To distinguish it from the conventional practice of electrowetting on a metal surface and because it provided unprecedented viability for practical applications, Lee [8] and Moon et al. [7] named this new material configuration electrowetting-on-dielectric (EWOD). Figure 2 compares EWOD with conventional electrowetting.

Although a significantly higher voltage is necessary to charge the dielectric capacitor of EWOD than the EDL of electrowetting in order to obtain the same degree of wetting, much higher voltages can be applied to the dielectric for additional wetting. In other words, a larger contact angle change can be achieved with EWOD by allowing high voltages, whereas conventional electrowetting is severely limited by the small breakdown voltage of the EDL. Figure 3 explains the



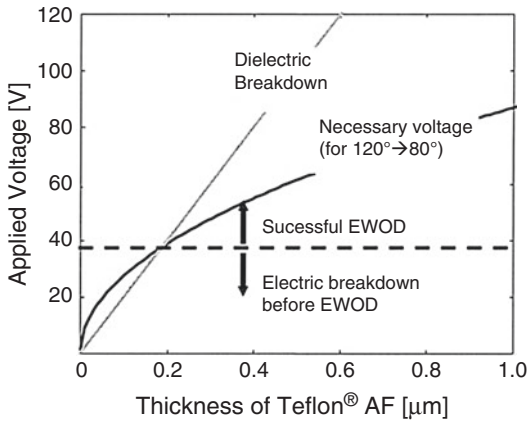
Electrowetting, Fig. 2 Electrowetting and EWOD [7]. (**a**) Illustration of conventional electrowetting. *Top*: With no external voltage applied, charges are distributed at the electrode-electrolyte interface, forming an EDL. *Bottom*: With an external voltage applied, charge density at EDL changes, and the contact angle decrease or increase. (**b**) Illustration of EWOD. *Top*: With no external voltage applied, there is little charge accumulation at the interface. *Bottom*: With an external voltage applied, charge accumulates at the interfaces, and the contact angle θ decreases

trade-off between dielectric thickness and operating voltage in EWOD, and why it is advantageous to use EWOD (compared to conventional electrowetting) despite the relatively high voltages that are needed. The thick solid curve in the figure shows that the voltage required to induce a specified increase in wetting (i.e., a reduction of contact angle $\Delta\theta$) is proportional to the square root of the thickness of the dielectric layer for a given dielectric material (Teflon AF is assumed in the figure), while the thin solid straight line shows that the breakdown voltage of the dielectric layer is proportional to the thickness. For successful operation, the electrowetting voltage should be smaller than the breakdown voltage. Although the required voltage continues to increase with the dielectric thickness, the margin of safety increases as well.

Despite the higher operation voltages, e.g., 15–80 V in air or 10–50 V immersed in oil in present experimental reports, EWOD is preferred over electrowetting on a conductor (which uses less than 2 V) in most cases. One may lower the voltage requirements of EWOD without sacrificing the electrowetting effect by coating the dielectric with a very thin layer of a very hydrophobic material. This strategy works by reducing the contact angle hysteresis, thereby enabling the contact lines (and thus droplets as well) to move more easily. A wide variety of droplet-medium combinations can be manipulated with EWOD, e.g., water in air, water in oil, oil-encapsulated water in air, oil in air, gas in water, etc. To learn more about the fundamentals of electrowetting and EWOD, readers are encouraged to consult [10].

Theories and Equations

The phenomenon of electrowetting can be interpreted (or understood) thermodynamically as well as electromechanically. In the



Electrowetting, Fig. 3 Electrowetting voltage and dielectric thickness [7]. The *thick solid line* represents the voltage required to obtain a certain degree of wetting (drawn for a contact angle reduction from 120° to 80° – enough reduction to slide a water droplet on many nonwetting solid surfaces in air) with Teflon AF as the sole dielectric, and the *thin solid line* represents the breakdown voltage for the same dielectric layer. A desired wettability increase is possible only where the electrowetting curve (the parabolic curve) stays safely below the *breakdown line* (the *straight line*). In other words, the dielectric should be thicker than where the two lines cross (~0.2 μ in the figure), as marked "Successful EWOD." If a smaller degree of wetting (e.g., moving a droplet on a solid surface coated with or immersed in oil) were acceptable, the corresponding electrowetting curve would be drawn lower, allowing lower voltages and thinner dielectrics. For conventional electrowetting, in contrast, the electrowetting curve would be drawn much lower than the EWOD curve in the figure due to the huge capacitance of the EDL (typically only several nanometers thick). Only ~1 V was used to induce enough wetting (70–40° to rough estimations) on gold to draw water in air environment [9]. However, that 1 V, which was the minimum needed to move the interface, was the maximum that could be applied before severe electrolysis. With no margin of safety, the actuation was not reversible

thermodynamic interpretation, assuming conventional electrowetting, the electrical energy accumulated at the solid–liquid interface, $cV^2/2$ (where c is the capacitance per area of the EDL and V the voltage across it), decreases the interfacial energy. For EWOD, the interpretation is slightly modified: the electrical energy accumulated in between the electrode and the liquid, $cV^2/2$, where c is the capacitance per area and V the voltage applied between them, makes the interfacial energy between the solid surface and the liquid decrease. The interfacial energy is exhibited as the contact angle of the liquid on the solid surface θ, which is easily measurable; as the solid–liquid interfacial energy decreases, so does the contact angle. By combining this energy equation with the Young equation and following the thermodynamic principle of energy minimization, the electrowetting equation (often called Lippmann-Young Equation) can be derived and written as follows:

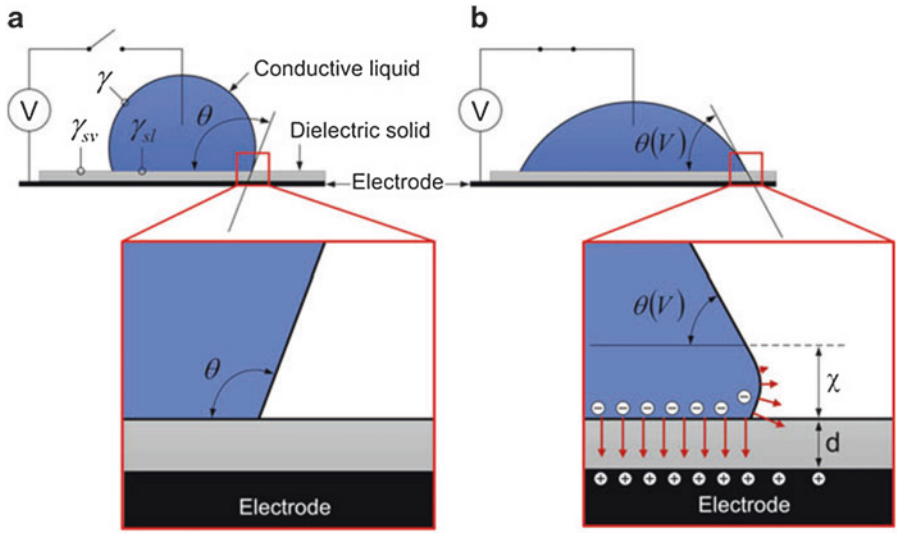
$$\cos\theta = \cos\theta_0 + \frac{cV^2}{2\gamma} \tag{2}$$

where θ_0 is the contact angle when there are no charges at the solid–liquid interface, i.e., no EDL exists, and γ is liquid-fluid (air being a fluid in Eq. 1) interfacial energy. The fluid may be a gas or a liquid that is immiscible with the liquid of interest, to which voltage is applied. For EWOD, where the specific capacitance can be obtained from the dielectric properties as $c = \varepsilon/d$ (ε is the permittivity, and d is the thickness of the dielectric), θ_0 coincides with the contact angle at no applied voltage, i.e., the Young angle, as the EDL has a negligible effect in this case. Considering the wide range of dielectric materials (i.e., representing a wide range of permitivities) and different liquid-fluid combinations used, it is convenient to define the second term on the right side of the equation as:

$$Ew = \frac{cV^2}{2\gamma} \tag{3}$$

The dimensionless Electrowetting Number Ew, defined as Eq. 3, expresses the importance of electrical energy relative to the free energy at the liquid-fluid interface. The liquid-fluid interface can be liquid–gas, liquid–vapor, or liquid-liquid if immiscible liquids. Although a liquid-air system is considered a standard in this essay, liquid-liquid systems of aqueous solution and oil are found in many electrowetting devices.

The importance of EWOD over conventional electrowetting and the growing view of the phenomenon as a physical force have led some researchers to the electromechanical interpretation of electrowetting, as summarized in Fig. 4. Considering the interfaces close to the solid–liquid-fluid

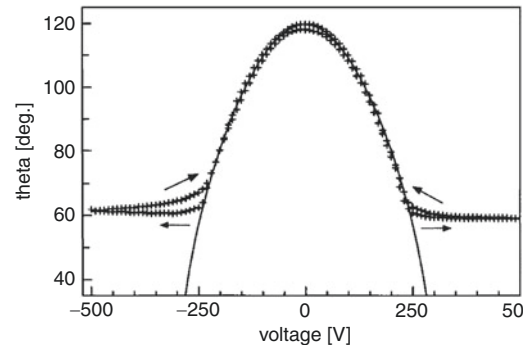


Electrowetting, Fig. 4 Electromechanical interpretation of electrowetting, illustrated for EWOD [10]. (**a**) With no voltage applied, the contact angle of a liquid is the Young angle. (**b**) With voltage V applied, the contact angle decreases by electrowetting. In the figure magnifying the area of solid–liquid-fluid triple point, χ is on the order of the dielectric thickness d, which is on the order of 1 μm in many EWOD devices

triple point as illustrated in Fig. 4b, one can see that the applied voltage produces normal electrostatic forces on the solid–liquid interface. Also, the fringe electric field on the liquid-fluid interface near the triple point exerts electrostatic traction normal to the liquid-fluid boundary right above the contact point (i.e., within χ in the figure), resulting in a net force parallel to the solid surface and causing droplet spreading. These electrostatic forces deform the liquid-fluid interface very close (sub-micron range in most practical cases) to the triple point, and this deformation reduces the slope of the interface away from the region. This reduced slope is the contact angle apparent to the observer. This pure electromechanical interpretation of electrowetting was made by theoretical derivations, e.g., [11, 12], and observed in experiments [13]. Under the electromechanical view, the electrowetting force is not infinitely concentrated at the contact point.

Electrowetting Curve

Since the electrowetting equation (Eq. 2), derived by both the thermodynamic and the electromechanical approaches, matches the experiments



Electrowetting, Fig. 5 A typical electrowetting curve [14] follows the electrowetting equation (Eq. 2) until the electrowetting effect saturates at high voltages (Reprinted with permission from American Chemical Society)

quite well in most cases, it plays an important role in understanding the phenomenon and performing research and development for applications. An example is shown in Fig. 5, which also reveals an important limitation: contact angle saturation. When the voltage is applied above a certain limit, a further increase of voltage does not decrease the contact angle anymore, e.g., not below 60° in Fig. 5. Instead, excessively high voltages only bring about detrimental effects,

such as the liquid-fluid interface becoming unstable and ejecting small droplets. Contact angle saturation is one of the most common sources of frustration for those engaged in research and development involving electrowetting. Despite its importance, however, the origin of saturation is still not well understood.

Summary

While explaining electrowetting, other closely related terms (electrocapillarity and EWOD) have also been described. Based on the same fundamental concept of electric potential affecting the interfacial energy, each portrays the electrowetting phenomenon in a somewhat different material configuration. Depending on the user's interest, one configuration may be more convenient than others, although EWOD is by far the most widely used configuration today. Basic theories of electrowetting have been presented through the conventional thermodynamic interpretation as well as the relatively new electromechanical interpretation, both of which result in the same electrowetting equation. The electrowetting equation is very useful as experimental data usually match the electrowetting curve. Today electrowetting is used in a wide range of applications including optical, biomedical, and even electronic.

Cross-References

► Dielectrophoresis

References

1. Lippmann, G.: Relations entre les phenomenes electriques et capillaries. Ann. Chim. Phys. **5**, 494–549 (1875)
2. Beni, G., Hackwood, S.: Electro-wetting displays. Appl. Phys. Lett. **38**, 207–209 (1981)
3. Prins, M.W.J., Welters, W.J.J., Weekamp, J.W.: Fluid control in multichannel structures by electrocapillary pressure. Science **291**, 277–280 (2001)
4. Lee, J., Kim, C.-J.: Surface-tension-driven microactuation based on continuous electrowetting. J. Microelectromech. Syst. **9**, 171–180 (2000)
5. Minnema, L., Barneveld, H.A., Rinkel, P.D.: An investigation into the mechanism of water treeing in polyethylene high-voltage cables. IEEE Trans. Electr. Insul. **EI-15**(6), 461–472 (1980)
6. Berge, B.: Electrocapillarity and wetting of insulator films by water. C. R. Acad. Sci. Ser. II **317**, 157–163 (1993)
7. Moon, M., Cho, S.K., Garrell, R.L., Kim, C.-J.: Low voltage electrowetting-on-dielectric. J. Appl. Phys. **92**, 4080–4087 (2002)
8. Lee, J.: Microactuation by continuous electrowetting and electrowetting: theory, fabrication, and demonstration. Ph.D. thesis, University of California, Los Angeles (2000)
9. Lee, J., Moon, H., Fowler, J., Schoellhammer, T., Kim, C.-J.: Electrowetting and electrowetting-on-dielectric for microscale liquid handling. Sens. Actuators **A95**, 259–268 (2002)
10. Nelson, W.C., Kim, C.-J.: Droplet actuation by electrowetting-on-dielectric (EWOD): a review. J. Adhes. Sci. Technol. (2012). (In print)
11. Jones, T.B.: On the relationship of dielectrophoresis and electrowetting. Langmuir **18**, 4437–4443 (2002)
12. Kang, K.H.: How electrostatic fields change contact angle in electrowetting. Langmuir **18**, 10318–10322 (2002)
13. Mugele, F., Buehrle, J.: Equilibrium drop surface profiles in electric fields. J. Phys. Condens. Matter **19**, 375112–375132 (2007)
14. Verheijen, H.J.J., Prins, M.W.J.: Reversible electrowetting and trapping of charge: model and experiments. Langmuir **15**, 6616–6620 (1999)

Electrowetting-on-Dielectric (EWOD)

► Electrowetting

Embryotoxic

► Use of Nanotechnology in Pregnancy

Emission Enhancement

► Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics

Energy-Level Alignment

► Electrode–Organic Interface Physics

Engineered Nanoparticles

▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity

Environmental Impact

▶ Fate of Manufactured Nanoparticles in Aqueous Environment

Environmental Impact Assessment

▶ Nano-TiO_2 Life-Cycle Assessment Perspective

Environmental Toxicology

▶ Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates

Epiretinal Implant

▶ Artificial Retina: Focus on Clinical and Fabrication Considerations
▶ Epiretinal Prosthesis

Epiretinal Prosthesis

Guoxing Wang[1] and Robert J. Greenberg[2]
[1]School of Microelectronics, Shanghai Jiao Tong University (SJTU), Minhang, Shanghai, People's Republic of China
[2]Second Sight Medical Products (SSMP), Sylmar, CA, USA

Synonyms

Epiretinal implant; Retinal prosthesis

Definition

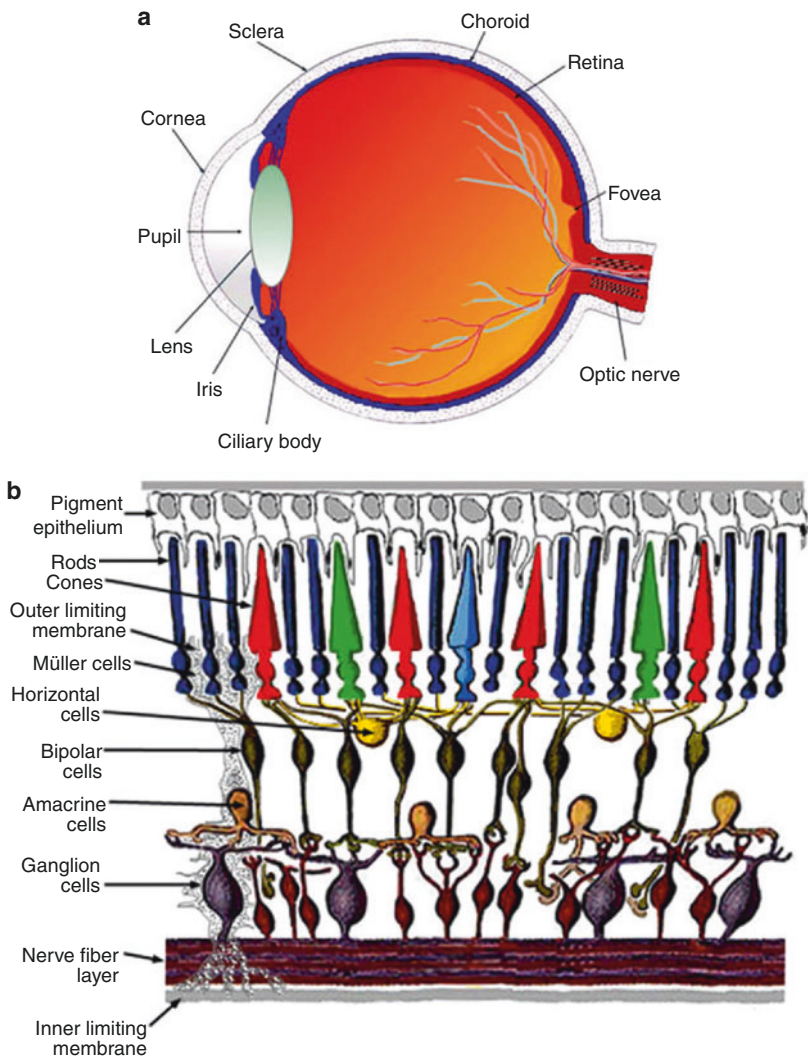
Epiretinal prosthesis refers to the device designed to restore partial vision for the blind people by stimulating the retina through electrodes on the retina surface.

Overview

Introduction to Retina

Vision is the main method for people obtaining information from the outside world. Unfortunately, it has been estimated that there are about 250 million blind patients in the world, and the number is increasing year by year. Researchers have tried many ways to address the causes of blindness. Electrical stimulation has been proposed as a means to restore some sight for blind patients. The idea behind this is similar to what has been done in cochlear implants, where electrical stimulation of hair cells in the inner ear has been used to successfully restore the hearing for the deaf people [1]. In retinal prosthesis, electrical pulses are being delivered to the retina, invoking a neural response that corresponds to the information of the image and creating a useful visual perception for the blind patients.

As is shown in Fig. 1, the retina is composed of several neuron cell layers, located in the back of the eye. The innermost layer in the retina that is closest to the lens and front of the eye is called the ganglion cell layer while the photosensors (the rods and cones) lie outermost in the retina against the pigment epithelium and choroid. The photosensors convert light into chemical/electrical signals which are conducted to the ganglion cells through intermediate layers of neurons. The ganglion cell axons form the optic nerve, carrying the image information to the brain. Interestingly, light has to travel through the thickness of the retina before striking and activating the rods and cones where the absorption of photons by the visual pigment of the photoreceptors is translated into a chemical/electrical message that can stimulate the succeeding neurons of the retina. The ganglion cells transmit all the information that has been processed by the retinal layers to the brain in terms of neuronal spikes.



Epiretinal Prosthesis, Fig. 1 Schematic of the human eye (**a**) and the retina-layered structure (**b**) (Adapted from Webvision, http://webvision.med.utah.edu/)

Main Causes of Blindness

Some retinal diseases, unfortunately, disturb the above process and can prevent the patients from getting useful vision. Age-related macular degeneration (AMD) is one of those and the leading cause of blindness in the developed world [2]. It results from abnormal aging of retina. As the cones of the fovea die, the central vision is affected first and thus patients lose the capability of seeing objects in details. Retinitis pigmentosa (RP) is another disease that causes blindness [2]. This is a genetically related disease and refers to a collective name for a number of genetic defects that result in photoreceptor loss. RP affects the rods (used in night vision) first and then the cones (used in ambient daylight levels). There are 100,000 people in the USA affected by retinitis pigmentosa, according to the National Eye Institute.

In both RP and AMD, the vision is impaired due to the damage to the photoreceptors that transform photons to neural signals. Postmortem evaluations of retina with RP or AMD have shown that a large number of cells remain relatively healthy in the inner retina compared to the outer retina [3]. This has led to the idea of building a device to electrically stimulate the

remaining viable cells in the retina to restore vision. In other words, the device can function as a replacement for the photoreceptors, converting the light information to electrical signals which can be interpreted by the remaining retinal neuron cells. The remaining ganglion cell neurons can pass the information to the brain, which perceives patterns of light and dark spots corresponding to the electrical signals generated by the device.

Types of Visual Prostheses

Retinal prostheses refer to the devices that mainly aim to fix the vision loss problems through electrically stimulating the retina. Electrodes are used to deliver electrical pulses to the retina. Depending on where the electrodes are placed, there are two leading approaches of retinal prostheses: epiretinal and subretinal. For epiretinal prosthesis, the electrodes are placed inside the vitreous, close to the inner layers of ganglion cells and nerve fibers. In contrast, in the sub-retinal prosthesis, the electrodes are placed under the retina where the rods and cones were. This entry will mainly explain the challenges and progresses associated with epiretinal prostheses. Other types of visual prostheses will be briefly discussed.

If the electrodes are implanted between the pigment epithelial layer and the outer layer of the retina, it is termed subretinal prosthesis. Compared with epiretinal approach, sub-retinal prosthesis has the potential advantage of closer proximity to the natural circuits of the retina and easier fixation of the electrodes since they are kind of sandwiched between the tissue layers. The disadvantages might include a more difficult surgery and more engineering challenges such as heat dissipation and power management. A German team led by the University Eye Hospital in Tübingen was formed in 1995 by Eberhart Zrenner to develop a subretinal prosthesis. Recently, they reported very promising results with acute human studies [4], but creating a reliable long-term subretinal device has proven more challenging. Also, currently the subretinal surgery requires filling the eye with toxic silicone oil. Clearly there are many issues to be resolved with the subretinal approach before it can be a viable practical therapy.

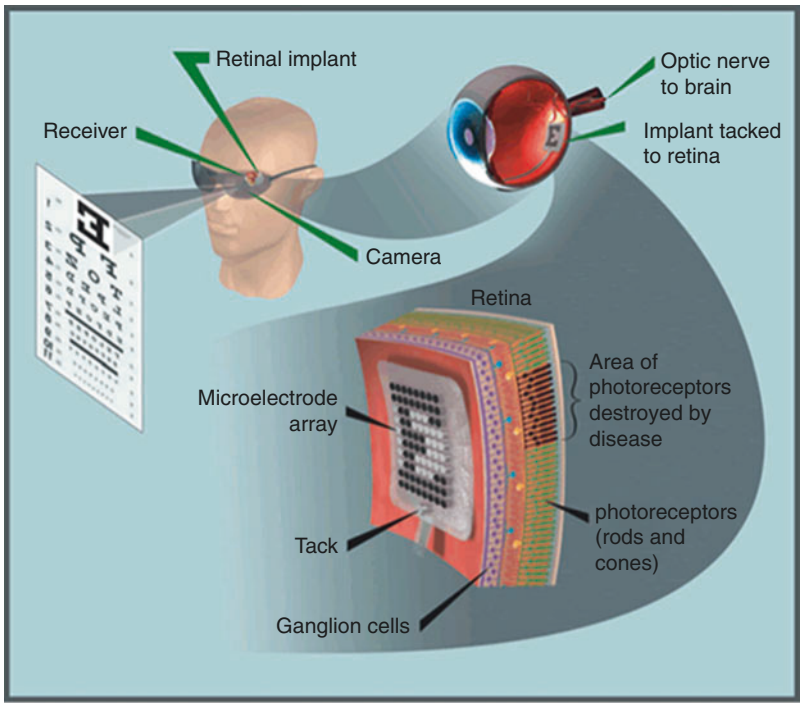
Another group led by Joseph Rizzo at the Massachusetts Eye and Ear Infirmary and John Wyatt at MIT also worked on retinal prosthesis and recently reported a subretinal stimulator with animal implantation results [5].

It is also worthwhile mentioning that similar to retinal stimulation, researchers worldwide have also proposed to electrical stimulation of optic nerve and visual cortex to invoke phosphenes to provide visional inputs to the patients. Among those are Dr. Claude Veraart from University of Catholique de Louvain at Belgium, Dr. Tohru YAGI from Tokyo Institute of Technology, Dr. Palankar from Stanford University, Dr. Nigel Lovell from University of New South Wales at Australia, and Dr. Mohammad Sawan from Ecole Polytechnique de Montreal at Canada. Despite all the challenges ahead, with the great efforts by the scientists, engineers, and entrepreneurs, there is no doubt people will move forward in curing blindness.

Principles of Epiretinal Prosthesis

Figure 2 shows the concept of epiretinal prosthesis. The device can be partitioned into two components, the external unit and the implant unit. The external unit includes the camera, signal processor, and power and data transmitter. The implant unit includes the power and data receiver, stimulator, and electrodes.

The whole device works as follows. An external video camera captures the image in the field of vision. This image is processed by a video signal processing unit that converts this image into stimulation data for the electrical stimulus. This data is then wirelessly transmitted to and received by an integrated circuit chip in the receiver inside the eye. The chip converts the signal to actual electrical stimulus currents which are delivered to the retina through the microelectrode array as shown in the figure. The electrode array is fixated on the retina by a retinal tack.



Epiretinal Prosthesis, Fig. 2 Scheme of epiretinal prosthesis components (Source: Artificial Retina Project, Department of Energy, http://artificialretina.energy.gov, January 2011)

Image Capture

A digital camera, most likely mounted on the patient's eyeglasses or head, captures the images in the field of vision. The raw data from the camera has to be processed to convert the video information to the format that can be understood by the implant chip to stimulate the retina. For example, the raw data needs to be downsampled, since most cameras today have orders of higher number of pixels than needed for a prosthetic device with only tens of electrodes. To date, Second Sight Medical Products, the leading company in retinal prosthesis, is carrying out a chronic clinical trial for a device with 60 electrodes (The Argus™ II). Devices with more electrodes are also being researched worldwide [6]. Besides downsampling, the video processing has to carry out complex procedures such as image enhancement.

Research is also being conducted to move the camera inside the eyeball. An intraocular camera used for image capturing may include an optical imaging system with an image sensor array, which may be enclosed in an implantable biocompatible housing together with other electronic components. The difficulty lies in the miniaturization of such an optical system in a way that is protected from the harsh corrosive environment of the body. Heat is also a challenge. Leaving the camera outside has the advantages of lower power consumption inside and thus less heat dissipation. In addition, it allows for convenient refinement and upgrades of signal processing algorithms and functionality in the exterior blocks.

Besides the imaging processing capability, the external unit should also provide options for personal customizations. The unit should have an easily accessible user interface that permits the patient to do some necessary image enhancement adjustments on his/her own. An error detection function should also be built in.

Wireless Telemetry

The implant needs power for the functions of electrical stimulation, image data recovery, and other circuits. The power required by the implant is provided by wireless telemetry. The overall power needed from telemetry is a function of the

stimulation parameters such as stimulation voltage and stimulation current amplitude, which in turn depend on many other parameters, such as the electrodes shape, distance to the retina, and even disease conditions of patients [3]. It has been estimated that a peak power of 250 mW may be needed for 1,000 electrodes stimulation [7]. During stimulation, the actual power needed by the implant may change with the stimulation patterns, which are a function of the image "seen" by the patients.

On the other hand, the maximum power that the implant can consume is governed by the safety issue that the temperature of the implant contacting the tissue should not exceed the normal temperature by a few degrees. Heating effects of RF power and chip operation have been studied experimentally and through modeling [8]. In sum, these studies suggest that wireless data and power transmission will not create a significant thermal effect, while the chip operation may generate a significant amount of heat. However, the liquid environment of the eye acts as a heat sink that is capable of dissipating a significant amount of power. An electronic chip positioned away from the retina can run at considerably higher power than a chip positioned on the retinal surface [9].

The required data rate for transmitting the image data from the external unit to the implant unit depends on the stimulus parameters and the refreshing rate of the stimulation. Assuming a total of 20 bits for one electrode stimulus current definition [10] and a refreshing rate of 50 frames per second, for 100 electrodes, the minimum data rate is $20 \times 50 \times 100 = 100$ kbps (kilobits per second). Obviously higher data rate will be needed if more electrodes are being stimulated.

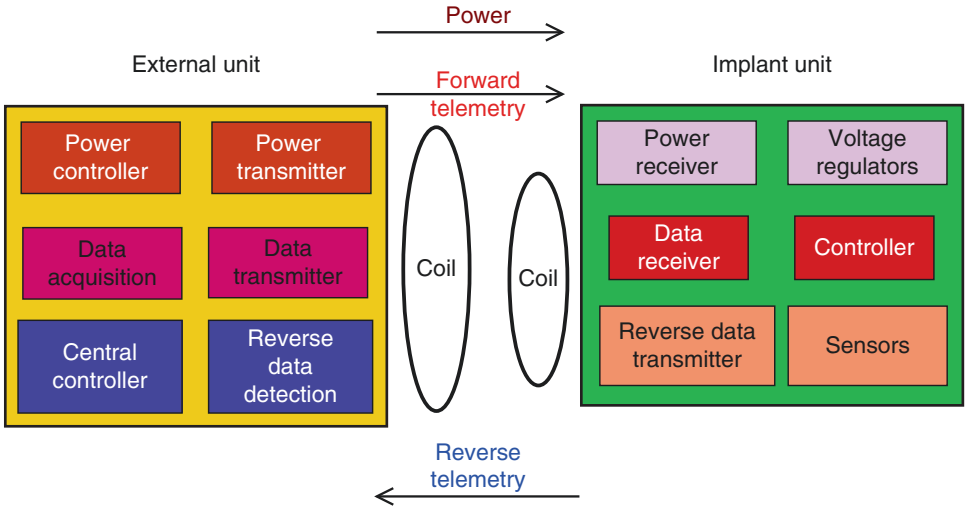
Besides relaying the image information from the external unit to the implant unit, it is also necessary to monitor the implant status for safety and also for optimal control purpose. For example, it is important to know the temperature of the implant and possibly shut down the device when the temperature exceeds a predetermined safety limit. Also, the change of pH value of the solution due to electrical stimulation could potentially be an indicator of tissue damage; thus it should be transmitted to the external unit for safety control [11]. The power level information inside the implant can also be detected and relayed back to the external unit to achieve an optimal power transfer system. The electrode impedance may be another parameter that needs to be transmitted back to the external unit for determining optimal stimulation patterns [12]. In sum, a communication channel is required to monitor the important implant parameters such as the temperature, pH value, power level, and electrode impedance, etc. These physiological signals generally change slowly and thus do not require a high data rate communication channel.

In summary, the telemetry system for retinal prosthesis needs to perform three functions, as shown in Fig. 3:

- Transmitting power to the implant
- Transmitting image data to the implant
- Transmitting data from implant to the external unit

To power the implant unit, the epiretinal prosthesis, like most other implantable applications, use inductive coupling as a means for wirelessly transmitting power. When an electrical current flows through a wire, a magnetic field is generated. A coil is formed when this conducting wire is wound into one or several loops. The magnetic field lines generated by each loop of coil combine with the lines generated by other loops to produce a concentrated field in the center of the coil. When another (receiving) coil is placed close to the transmitting coil, the magnetic field generated by the transmitting coil passes through the windings of the second coil and these two coils are magnetically (inductively) coupled. Further, if the current of the transmitting coil is alternating, by Faraday's law, the time-varying magnetic field generated by the alternating current induces an alternating voltage in the receiving coil, thus enabling energy transfer from the transmitting coil to the receiving coil.

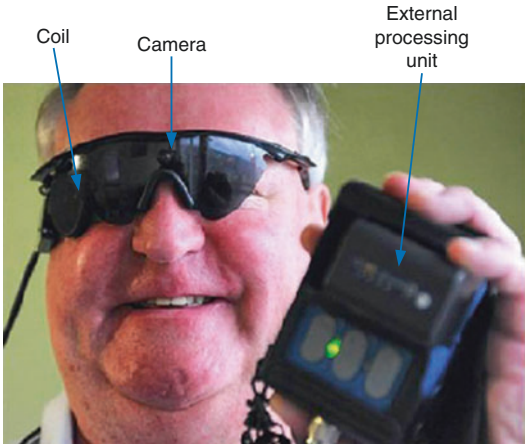
The whole system is powered by a portable battery outside the body with power transmitted to the implant through the coupled coil pair. The power amplifier at the primary side converts the battery DC power into radio frequency



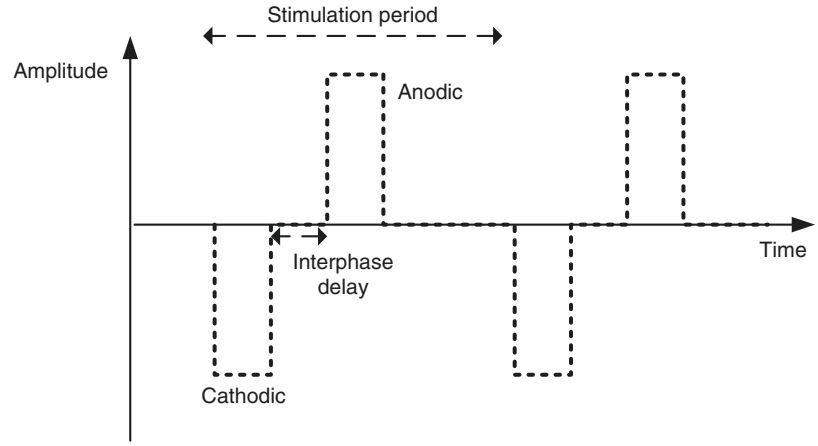
Epiretinal Prosthesis, Fig. 3 Illustration of an inductive power and data telemetry for epiretinal prosthesis

(RF) power to be radiated by the primary coil and thus to be received by the secondary coil. This power is then converted back to DC by the power recovery circuits to operate the implant electronics.

The size and shape of the coils are highly critical for an efficient power transfer and deserves special attention. Firstly, to facilitate the energy transfer, the two coils should be constructed with a shape that can maximize the coupling and be placed as close as possible. Secondly, the wire diameter of the coil should be chosen according to the frequency of the electrical current to minimize the losses associated with electrical resistance. For example, one way to reduce the resistance is to construct the coil with several wires (strands) twisted together (for example, litz wire) to increase the equivalent cross-section conducting area and reduce the resistance. However, the number of strands is normally limited by the maximum allowable size of the coils. Another important parameter of the coil is the number of the loops of the conductor (called number of turns). With the same current in the transmitting coil, more turns can generate higher magnetic field thus inducing a higher voltage across the implanted receiving coil. However, as the number of turns increases, the equivalent resistance of the coil increases too, which causes higher power dissipation on the coil. For a given size of coil, the number of strands trades off with the number of turns, and both should be optimized for a given application to achieve high power efficiency. The coil being implanted should be constructed as small as possible to minimize the patient discomfort and to ease surgery/injection procedure. The external coil, mounted on eyeglasses, however, is usually much larger for better coupling as shown in Fig. 4.



Epiretinal Prosthesis, Fig. 4 Photo of external unit developed by Second Sight Medical Products. Coil is on the side of the eyeglasses



Epiretinal Prosthesis, Fig. 5 Scheme of typical biphasic pulses for stimulating retina

The magnetically coupled coil pair can also be used to transmit data wirelessly from the external unit to the implant unit. By modulating the power carrier (used to transmit power), data can be sensed by the implant unit. Research is also being conducted to transmit data using a different carrier instead of the power carrier, mainly to increase the amount of data that can be transmitted without sacrificing the power transmission efficiency.

Stimulating the Retina

Electrical current is delivered from the electrodes to the retina, and presumably the retinal cells are stimulated, which generate neuronal spikes to be passed to brain. A high-density electrode array is composed of a substrate that provides mechanical support and the conducting material, for instance, a noble metal such as platinum, creating the electrical connection between the stimulator and the tissue. At the electrode/tissue interface, the electronic current is transformed into ionic current which sets up a voltage gradient that then activates the neurons.

It is desirable for the electrodes to be placed on to the macular region and close to the retina as this will create more efficient stimulation. The electrode array, as illustrated in Fig. 2, is fixated onto the retina by a tack. Optimizing this interface is an important area of research.

As shown in Fig. 5, a typical electrical stimulus is biphasic with the pulse of a cathodic phase and an anodic phase. The interphase interval (time between the two phase) separates the pulses slightly to make sure that the next pulse would not reverse the physiological effect of the previous one. The stimulus information contains the amplitude, width, interphase delay, and the repetition rate of the pulses. All this information highly correlates to the image information and the actual stimulation strategy may also be patient-specific. In other words, for the same picture being "seen" by the patient, the stimulation currents may be very different.

The integrated circuit chip is responsible for generating such current according to the commands it receives through data telemetry from the external unit. The stimulator is miniaturized to be accommodated within the constraints of the available space and designed to be controlled with the minimum power to cut down the possible heat dissipation inside the body.

Current Research Progress

In the early 1990s, Humayun and colleagues demonstrated that controlled electrical stimulation of the retina in individuals blind from RP and AMD results in visual percepts in the form of light dots. Experiments using pattern electrical stimulation of a human retina resulted in subjects perceiving

simple forms of vision with the percepts corresponding in time and location to the electrical stimulus. These were short-term tests, with the stimulating devices temporarily positioned on the retina.

In 1998, Second Sight Medical Products was founded to build a retinal prosthesis. A permanent implant, developed by Second Sight Medical Products, with 16 electrodes (4 × 4 array) was implanted in patients starting in 2002. The implanted subjects were able to use a camera to detect the presence or absence of ambient light, to detect motion, read a large letter, count objects, and recognize and differentiate objects such as cup from plate. The longest implant still in use by a patient today is over 7 years to date.

Starting in 2007, Second Sight went on to a clinical trial for its second-generation device, Argus II, which has 60 electrodes. So far there have been more than 30 patients implanted with centers in the USA, England, France, Geneva, Mexico, etc. The patients with such devices, for the first time, could see a series of letters, thus recognizing simple words and even sentences. Some patients could recognize numbers on cards held in their palm.

One natural assumption is that with more electrodes, the epiretinal prosthetic device could provide better vision inputs for the suffered blind patients and thus restore more meaningful vision to the patients, such as face recognition, unaided mobility.

Simulations of prosthetic vision suggest that 600–1,000 electrodes will be needed to restore visual function to a level that would allow blind to read, navigate a room unaided, and recognize faces [13, 14]. But, through the use of scanning, clearly far fewer can make a significant impact. As the number of stimulation outputs increases, the complexity of the stimulator electronics increases, requiring larger amounts of power to provide to the implant and a higher data rate, for transmitting the image information to the stimulating electronics. Protecting these high-channel-count devices from the body is also increasingly difficult as the numbers of electrodes increases. To date, only the Argus I and II from Second Sight have demonstrated long-term reliability in patients.

Future Direction for Research

With the advancement of technologies in the areas of microelectronics, mechanics, materials, there is no doubt that devices with better performance than current state of the art will emerge in the future. Needless to say, this involves a great deal of engineering challenges and trade-offs. Can 1,000 analog drivers be packed into a microchip that can be small enough to be implanted? Can enough data be transmitted into the device without consuming too much power to actually overheat the implant? Can a hermetic package be built to support 1,000 electrodes?

Another important area is to understand the neural response of the retina when stimulated by electrical current. Many experiments have been carried out to find out the relationship between stimulation current (such as the magnitude, waveform, and frequency) and the perception, but more work is needed.

Cross-References

- ▶ Bioinspired CMOS Cochlea
- ▶ Biomimetics
- ▶ Biosensors
- ▶ Nanostructured Functionalized Surfaces
- ▶ Organic Bioelectronics

References

1. Sanders, R.S., Lee, M.T.: Implantable pacemakers. Proc. IEEE **84**(3), 480–486 (1996)
2. Margalit, E., Sadda, S.R.: Retinal and optic nerve diseases. Artif. Organs **27**(11), 963–974 (2003)
3. Humayun, M.S., Prince, M., de Juan, E.J., Barron, Y., Moskowitz, M., Klock, I.B., Milam, A.H.: Morphometric analysis of the extramacular retina from postmortem eyes with retinitis pigmentosa. Invest. Ophthalmol. Vis. Sci. **40**(1), 143–148 (1999)
4. Zrenner, J., et al.: Subretinal electronic chips allow blind patients to read letters and combine them to words. Proc. R. Soc. B (2010). doi:10.1098/rspb.2010.1747
5. Shire, D.B., Kelly, S.K., Jinghua, C., Doyle, P., Gingerich, M.D., Cogan, S.F., Drohan, W.A., Mendoza, O., Theogarajan, L., Wyatt, J.L., Rizzo, J.F.: Development and implantation of a minimally invasive wireless subretinal neurostimulator. IEEE Trans. Biomed. Eng. **56**(10), 2502–2511 (2009)

6. Weiland, J.D., Humayun, M.S.: Visual prosthesis. Proc. IEEE **96**(7), 1076–1084 (2008)
7. Wang, G., Liu, W., Sivaprakasam, M., Kendir, G.A.: Design and analysis of an adaptive transcutaneous power telemetry for biomedical implants. IEEE Trans. Circ. Syst. I. **52**, 2109–2117 (2005)
8. Singh, V., Roy, A., Castro, R., McClure, K., Weiland, J., Humayun, M., Lazzi, G.: Specific absorption rate and current densities in the human eye and head induced by the telemetry link of a dual-unit epiretinal prosthesis. IEEE Trans. Antennas Propag. **57**(10), 3110–3118 (2009)
9. Piyathaisere, D.V., Margalit, E., Chen, S.J., Shyu, J.S., D'Anna, S.A., Weiland, J.D., Grebe, R.R., Grebe, L., Fujii, G., Kim, S.Y., Greenberg, R.J., De Juan Jr., E., Humayun, M.S.: Heat effects on the retina. Ophthalmic Surg. Lasers Imaging **34**(2), 114–120 (2003)
10. Sivaprakasam, M., Liu, W., Wang, G., Weiland, J.D., Humayun, M.S.: Architecture tradeoffs in high density microstimulators for retinal prosthesis. IEEE Trans. Circ. Syst. I Spec. Issue Biomed. Circ. Syst. **52**(12), 2629–2641 (2005)
11. Chu, A.P., Morris K., Greenberg R.J., Zhou D.M.: Stimulus induced pH changes in retinal implants. In: 26th Annual International Conference of the Engineering in Medicine and Biology Society, vol. 2, pp. 4160–4162. IEEE, San Francisco (2004)
12. Shah, S., Chu, A., Zhou, D., Greenberg, R., Guven, D., Humayun, M.S., Weiland, J.D.: Intraocular impedance as a function of the position in the eye, electrode material and electrode size. In: Conference Proceedings, 26th Annual International Conference of the Engineering in Medicine and Biology Society, vol. 2, pp. 4169–4171. IEEE, San Francisco (2004)
13. Cha, K., Horch, K.W., Normann, R.A.: Mobility performance with a pixelized vision system. Vis. Res. **32**(7), 1367–1372 (1992)
14. Hayes, J.S., Yin, J.T., Piyathaisere, D.V., Weiland, J., Humayun, M.S., Dagnelie, G.: Visually guided performance of simple tasks using simulated prosthetic vision. Artif. Organs **27**(11), 1016–1028 (2003)

EUV Lithography

Chimaobi Mbanaso and Gregory Denbeaux
College of Nanoscale Science and Engineering, University at Albany, Albany, NY, USA

Synonyms

Extreme ultraviolet lithography (EUVL); Soft x-ray lithography

Definition

EUV lithography is a technology that uses radiation near 13.5 nm wavelength to transfer patterns from a reflective mask to a substrate coated with a light-sensitive material called a photoresist. This patterning technique occurs under vacuum conditions in an all-reflective optics configuration and is used to fabricate integrated circuits with printed features that are smaller than 32 nm. The technology is considered to be the likely next-generation lithography method with the prospect of succeeding the current immersion lithography technique.

Overview

Semiconductor integrated circuits (ICs) which are embedded in various electronic devices are fabricated using a lithography process. As technology has advanced since the invention of the IC in 1958, lithography has played a vital role in patterning smaller features to produce faster and more powerful operating devices. Decreasing the exposure wavelength used for lithography is one of the ways in which the printing of ICs has steadily improved using optical lithography techniques. Extreme ultraviolet (EUV) lithography operating at 13.5 nm wavelength will possibly succeed the state-of-art lithography method as the next-generation technique of choice for continuing the shrinking of ICs [1–9]. Other next-generation lithography methods that have been considered include proximity x-ray lithography, electron projection lithography, ion projection lithography, and multi-e-beam direct-write [6, 10]. Before the adoption of the next lithography technology, important requirements such as throughput, cost of ownership, and image quality have to be considered [6].

In lithography, an optical projection printing system uses light to image a mask pattern onto a wafer substrate that has been coated with a light-sensitive material called a photoresist. The projection lens system typically reduces the mask pattern by a factor of four onto the photoresist [11, 12]. The lithography process defines the

critical dimensions of the printed features while other process steps such as etching and metallization contribute toward the overall fabrication of the device. In EUV lithography, 13.5 nm wavelength of light enables the transfer of patterns from a reflective mask to a photoresist-coated substrate. The short EUV wavelength is easily absorbed by most materials; hence, the lithography process must be operated under vacuum conditions and by using special reflective materials in the optical system in contrast to conventional lithography techniques that use refractive optics [3, 4].

One of the drivers for implementing EUV lithography is the capacity to print high-resolution features by the following fundamental relationships describing the lithographic system with resolution (R) and depth of focus (DOF). Thus,

$$R = k_1 \frac{\lambda}{NA} \tag{1}$$

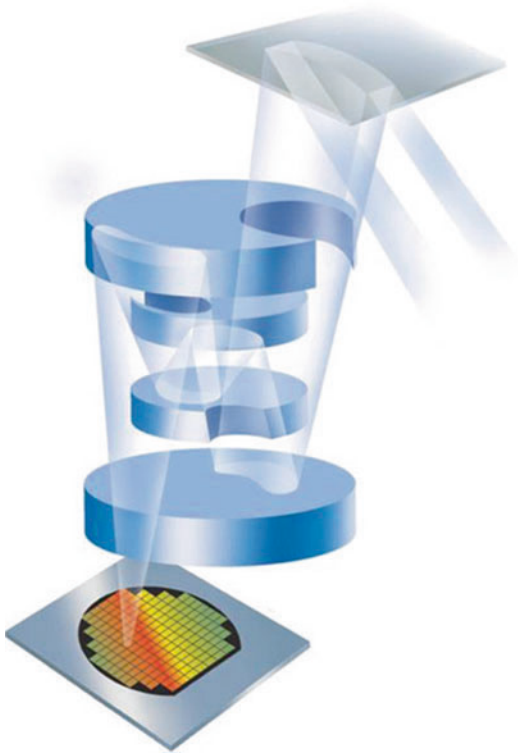
$$DOF = \pm k_2 \frac{\lambda}{NA^2} \tag{2}$$

where λ is the optical wavelength and NA is the numerical aperture of the projection system [3]. The parameters k_1 and k_2 are determined through empirical means and depend on the critical dimension (CD) tolerance and the size of an acceptable IC manufacturing process window [5]. The printing of smaller features in the semiconductor industry has followed a trend in which the number of transistors in an IC has doubled after approximately every 24 months (Moore's law) [13]. Strategies to improve the optical lithography technique to keep pace with Moore's law have led to modification of the lithography system according to Eq. 1; hence, decreasing the wavelength of light, fabricating high NA optics or decreasing k_1 have led to smaller patterned features [8]. The optical lithography technique that used the 436 nm g-line illumination wavelength during the early 1980s printed features with dimensions of over 2 μm by using a system with an NA of 0.15 and a k_1 of 0.8 [10]. At present, 193 nm lithography uses an immersion liquid along with various resolution enhancement techniques such as phase shifted masks, optical proximity correction, off-axis illumination, and double patterning lithography to print features below 32 nm [10]. Shortening the wavelength by using 13.5 nm is the next practical way to enable the printing of smaller features for future IC devices.

Initial concepts of EUV lithography were introduced in the 1980s from research efforts in Japan and the United States while using a broader radiation wavelength range of about 10–30 nm, referred to as soft x-rays [3, 4]. National laboratories such as the Lawrence Livermore National Laboratories (LLNL), Sandia National Laboratory (SNL), Lawrence Berkeley National Laboratory (LBNL), as well as the AT&T Bell Laboratories performed some of the early research in using EUV radiation in an all-reflective projection lithography system [4]. Since then, research efforts by industry consortia and semiconductor manufacturers have all been instrumental in leading developmental efforts in EUV lithography for high-volume manufacturing. Some of the early EUV development programs which commenced in the 1990s include Extreme Ultraviolet Limited Liability Company (EUV LLC), the European Extreme UV Concept Lithography Development System (EUCLIDES), and the Japanese Association of Super-Advanced Electronics Technologies (ASET). These programs and many others since then [4], have supported the development of light sources, optics, masks, photoresists, and vacuum equipments needed for EUV lithography. The sections that follow discuss these essential components of an EUV exposure tool and their development for this technology (Fig. 1).

EUV Light Sources

Historically, progressively shorter wavelengths in the lithography tools have been used as one of ways to reduce the feature sizes of IC devices. The visible g-line at 436 nm wavelength was used in the early 1980s for the lithography process [11]. Its wavelength successor was the ultraviolet i-line at 365 nm which was used in the late 1980s [11]. Both the g-line and i-line radiation wavelengths were produced with a mercury arc lamp.



EUV Lithography, Fig. 1 Schematic showing the EUV lithography projection optical system based on reflective mirrors (Courtesy of Carl Zeiss SMT GmbH)

With further decrease in feature sizes came the advent of deep ultraviolet (DUV) 248 nm wavelength lithography using a KrF (krypton fluoride) excimer laser in the late 1990s and then in 2001 [11], the ArF (argon fluoride) excimer laser was used for advanced lithography and has been modified to use an immersion liquid. ArF lithography uses light at 193 nm wavelength. The fluorine (F_2) excimer laser, which operates at 157 nm wavelength, was briefly considered to be the next lithography step toward shorter wavelengths but did not proceed to high-volume manufacturing [10, 11]. EUV lithography, operating at 13.5 nm, is considered the next step in shortening the wavelength of light used for lithography to continue the scaling of device features.

To generate high-power radiation at 13.5 nm for EUV lithography, high-temperature plasmas (approximately 30 eV) of a fuel material are used [3, 14]. For high-volume manufacturing applications, these hot plasmas of a target fuel such as xenon, tin, or lithium are generated from either an electrical discharge or by using an intense laser pulse. These two viable commercial solutions for generating EUV radiation are referred to as discharge produced plasma (DPP) and laser produced plasma (LPP). Other sources of EUV radiation such as synchrotrons were used for early demonstrations and pioneer studies of EUV lithography but requirements such as space restrictions limited their practical use for commercial applications [6, 7]. Tin is currently the leading target fuel for high-volume manufacturing EUV sources because of its higher conversion efficiency (CE) compared to xenon and lithium [3, 14]. CE, which is the ratio of the generated energy from the EUV source in the 2 % bandwidth of 13.5 nm wavelength to the energy input to the EUV source, is used to estimate the utility requirements and fuel choice [3].

In order to focus the light generated by the EUV source, collector optics that are either grazing incidence or normal incidence are used. The point where the light is focused is called the intermediate focus (IF) and connects the source-collector module to the illuminator and projection mirrors of the EUV lithography scanner. Only a portion of the EUV radiation emitted by sources is collected and transmitted to the IF due to factors such as the geometrical restrictions in the collector design, the efficiency of the collectors to reflect the EUV radiation, absorption of EUV by background gases and spectral purity filters (SPFs) along the propagation path to the IF [3]. SPFs are needed to suppress out-of-band radiation produced outside the 2 % bandwidth of the 13.5 nm wavelength by the plasma sources. The presence of these filters can further attenuate the in-band EUV power. Other filters include debris barriers to mitigate high-energy particles from the plasma source which can otherwise lead to degradation of reflective optics in close proximity to the source [3, 4].

EUV source specifications are determined based on consumer requirements which include throughput, cost of ownership, and image quality. For instance, for high-volume manufacturing, the EUV power generated must be sufficient to ensure

that a high wafer throughput, which is estimated to be about 100 wafers per hour, is achieved. This equates to a power of 115–180 W depending on the sensitivity of the photoresist at 5–10 mJ/cm^2, respectively. Hence, sources that meet the power requirements, while maintaining energy stability and reliability during operation, are needed. EUV sources can also be used for metrology purposes and the specifications for such applications are somewhat different from lithography exposure tools [6].

EUV Multilayer Optics

In contrast to prior optical lithography technologies that use transmission optics for propagating the light to the photoresist, EUV technology needs vacuum conditions and reflective optics to avoid absorption of the short wavelength at 13.5 nm. The optics system used for EUV lithography consists of reflective illumination optics used to effectively allow the mask to be illuminated and reflective projection optics to image the pattern on the photoresist. The reflective optics consist of repeated bilayer pairs to make a multilayer mirror material [3, 15]. The mathematical relationship between the bilayer spacing, d, incidence angle, θ, and wavelength, λ, is shown in the equation below,

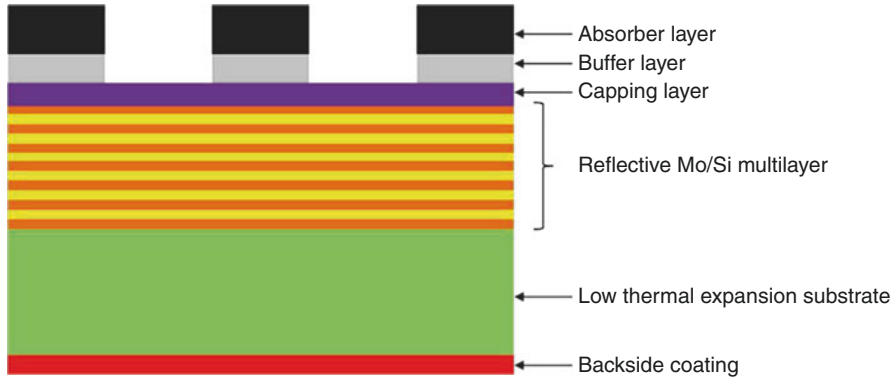
$$n\lambda = 2d\cos\theta \tag{3}$$

where n is an integer. The bilayer stack to reflect EUV radiation from a multilayer consists of a metal (high Z material) and a spacer (low Z material) which correspond to materials with high and low atomic numbers (Z). Historically, two of the common multilayers considered for EUV lithography were Molybdenum/Beryllium (Mo/Be) and Molybedenum/Silicon (Mo/Si) corresponding to best reflectivity performance at wavelengths in the 11.3–11.6 and 13.3–13.6 nm range, respectively [3]. The Mo/Si for 13.5 nm is widely used as the metal/spacer material of choice for EUV technology with a typical stack consisting of 40 Mo/Si layer pairs. These mirrors have shown a peak reflectivity of close to 70 % at 13.5 nm [15]. The quality of the multilayer coatings contributes to the reflectivity of the multilayer mirrors and is dependent on the deposition process. Magnetron sputtering [3, 15] is one of the common deposition techniques for EUV lithography mirrors because of the ability to coat large optics, the relatively fast sputtering rate and reproducibility from process to process. Some of the other deposition methods include ion-beam sputtering, electron-beam evaporation, and pulsed laser deposition [3].

Contaminants, which include water and hydrocarbons, in the presence of EUV photons can lead to oxidation or surface contamination resulting in a loss of reflectivity [3, 16]. The operational lifetime of the projection optics used in the EUV system must satisfy greater than 30,000 light-on hours [3, 5]. Thus, degradation at the surface of the optics must be kept at a minimum. Molecular contamination has been attributed to cracking of adsorbed molecules at the mirror surface by EUV-induced secondary electrons [16]. Typical species causing contamination are oxygen-containing molecules and carbon-containing molecules. A small increase in the layer of oxide (less than 1.5 nm) on the surface of these multilayer mirrors leads to a reflectance loss of greater than 1 %, and is unacceptable [3]. Such optics contamination will decrease the throughput of the exposure tool and printing uniformity. Strategies to reduce contamination of the optics include using protective capping layers, such as ruthenium, that are resistant to oxidation [3, 16]. Other strategies to extend the lifetime of the optics and reduce carbon growth include improvement of the vacuum conditions and adopting various cleaning techniques that would efficiently remove the carbon but not damage the underlying mirror. Possible removal methods include the use of molecular oxygen and atomic hydrogen [3].

EUV Masks

The EUV mask, which contains the pattern to be imaged onto the wafer, has to be reflective to avoid absorption of the short wavelength of light; hence, it is different from conventional



EUV Lithography, Fig. 2 Schematic showing a stacking structure of a reflective mask used for EUV lithography

transmissive optical masks. The EUV mask undergoes two main fabrication processes. These are the development of the mask blank and the patterning of the mask. For the mask blank fabrication, a flat and defect-free substrate with a low coefficient of thermal expansion material maintains the rigidity of the mask with minimum distortion [3, 6]. On this substrate, a reflective multilayer structure consisting of alternating layers of Mo/Si is deposited and covered by a capping layer to prevent oxidation. The multilayer on the EUV mask ensures high reflectivity at the 13.5 nm wavelength and consists of 40 pairs of Mo/Si bilayers as shown in Fig. 2. An absorber layer coating, antireflective coating on the topside of the mask as well as a conductive layer coating on the backside of the mask are all performed as part of the mask blank fabrication process. The absorber layer helps in the pattern-forming process by providing contrast to the reflective multilayer through absorption and minimum reflection, and materials such as tantalum nitride (TaN) are used for this function [3]. To minimize reflectivity during mask inspection using deep ultraviolet (DUV) light, an antireflective coating can be applied to the top of the absorber layer. A buffer layer material such as silicon dioxide (SiO_2) or ruthenium can be used to protect the multilayer structure during the etching or repair of the absorber layer [3]. The conductive backside coating is made to be compatible with the electrostatic chuck used to support and flatten the mask in the exposure tool as well as other process tools. Electrostatic chucking ensures that the clamping forces are evenly distributed over the mask area to maintain flatness. The next step in the fabrication of the EUV masks is the mask patterning by electron beam writing and dry etching. The writing process is similar to that used for conventional optical masks but with tighter restrictions on distortions and thermal specifications.

Mask defects can occur in the absorber layer, the multilayer reflective coating, or in the substrate below the multilayer coating. Substrate defects can propagate through the multilayer coating and cause a phase defect which is the perturbation of the reflected wave front [3]. Amplitude defects can be caused by particles near the top of the multilayer or if there is a flaw in the multilayer. Other defects can occur in the absorber layer where a portion of the absorber is missing from its desired location (clear defect) or if there is absorber material in an undesired location (opaque defect) [3]. Unlike optical lithography that uses a highly transparent membrane (pellicle) to be placed over the mask to provide protection from small particles that can fall and cause defects, the absorption of the short 13.5 nm wavelength makes it unusable in EUV lithography. Nevertheless, during shipping, storage, or handling, various protection schemes have been evaluated such as the dual pod concept [3]. The dual pod is a special mask carrier that gives double protection to the mask by using an inner and outer compartment to maintain a particle-free mask before use in an EUV exposure tool.

To avoid defects from being imaged on the wafer, the EUV mask must go through the manufacturing process defect-free and remain that way during use. Reliable and accurate EUV mask inspection is needed to capture and characterize printable mask defects at various levels of fabrication. Inspection tools that use EUV wavelength (actinic) and DUV wavelengths (non-actinic) have been used for this purpose [3]. However, to meet the requirements for high-volume manufacturing, actinic inspection and imaging tools will likely be needed due to the limits of DUV inspection techniques [3].

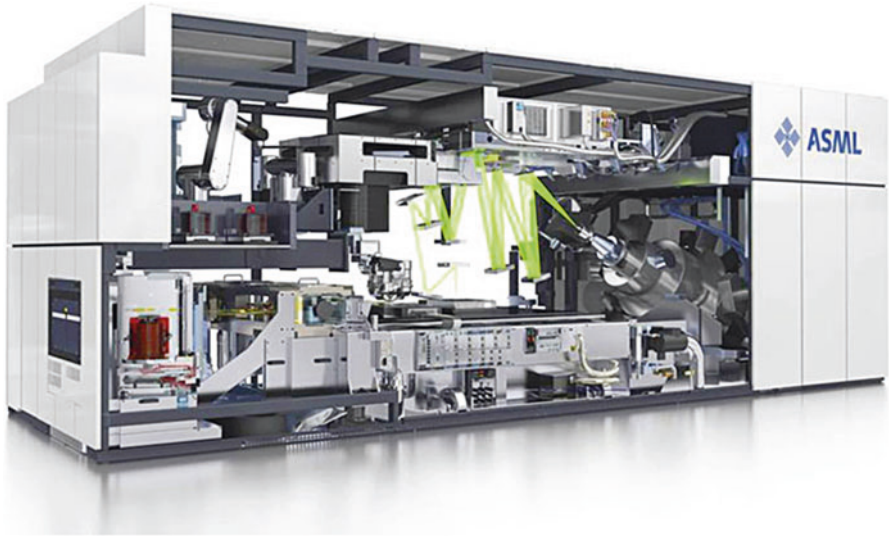
EUV Photoresists

Photoresists, which are the light-sensitive materials used in the lithography process, have undergone much progress from the Novolac® photoresists used in the i-line (365 nm) era to chemically amplified photoresists used for 248 and 193 nm lithography [12, 17]. The platform for EUV photoresists, also based on chemical amplification, is similar to those used for 248 and 193 nm optical lithography. These photoresists were first used in the mass production of semiconductor devices during the transition from the i-line of the mercury lamp to the KrF (248 nm) excimer laser [18]. These chemically amplified photoresists have been further employed from KrF (248 nm) to the current ArF (193 nm) immersion lithography and exhibit high sensitivity and resolution [6]. In chemically amplified photoresists, pattern formation occurs through deposition of energy in the photoresist upon exposure to the incident radiation leading to the generation of acids. The acids catalyze the pattern forming reaction at elevated temperatures referred to as the post exposure bake (PEB) step. The image in the photoresist is then developed in an aqueous base to reveal the pattern. The efficiency of these processes depends on the absorption efficiency of the incident radiation, the number of acids generated, and the efficiency of the catalytic chain reaction [12]. In EUV, the mechanism for interaction of radiation with the photoresist is different than exposures using 248 and 193 nm radiation technologies. The high energy of the EUV photon, 92 eV (13.5 nm), ionizes the components of the photoresist film to generate secondary electrons which contribute to the chemical reactions occurring during exposure [17, 18]. The properties of these photoresists depend on the photoacid generator (PAG) and the polymer used in the formulation and affects the specifications for resolution, line edge roughness (LER), and sensitivity [12]. The sensitivity of the photoresist required for EUV lithography is dependent on the power produced by the source. One of the concerns of using EUV photoresists in the vacuum environment with all the nearby reflective optics is the release of hydrocarbon contaminants during exposure, which could decrease the lifetime of the optics [19]. The outgassing of photoresist materials exposed to EUV wavelengths has been studied using various methods [19].

EUV Research Status

EUV technology has undergone rigorous research efforts which began over two decades ago in order to meet the requirements for high-volume manufacturing and cost of ownership. One of the critical areas that needs further development is the EUV source. Currently, the collectable power of EUV is limited and must be improved for EUV technology at high-volume manufacturing throughput levels. For EUV technology to be implemented, demonstrations of the reliability and stability for source power have to be achieved. Defect-free masks with the corresponding inspection infrastructure that can find all the printable EUV mask defects must be developed. SEMATECH (Semiconductor Manufacturing Technology, www.sematech.org) is contributing actively to develop the infrastructure for EUV blank and mask inspection tools. In addition, photoresists that meet the requirements of resolution, sensitivity, and line edge roughness (LER) need to be available as well.

Currently, ASML (ASM lithography, www.asml.com) has developed the semiconductor industry's first EUV lithography production platform (TWINSCAN NXE). The system platform is



EUV Lithography, Fig. 3 ASML NXE:3100 which is the pre-production TWINSCAN system EUV lithography scanners (Courtesy of Bill Pierson, ASML)

based on the learning and achievements from the EUV prototype tools (Alpha Demo Tools – ADT) located at IMEC (Interuniversity MicroElectronics Center, www.imec.be) in Belgium and CNSE (College of Nanoscale Science and Engineering, www.cnse.albany.edu) in Albany, New York [20]. The system architecture is based on a six-mirror lens, and the EUV radiation is generated from a tin-based plasma source [20]. The preproduction TWINSCAN NXE system is the NXE:3100 shown in Fig. 3, with *NA* as 0.25 and resolution of 27 nm, and six of them have been produced [20]. The NXE:3300B will follow the NXE:3100 to target volume production at a resolution of 22 nm and less at an *NA* of 0.32 in 2012 with throughput estimated to be 125 wafers per hour [20].

Cross-References

- ▶ DUV Photolithography and Materials
- ▶ Electron Beam Lithography (EBL)
- ▶ Immersion Lithography
- ▶ Immersion Lithography Materials
- ▶ Nanoimprint Lithography
- ▶ Nanoimprinting

References

1. Wu, B., Kumar, A.: Extreme Ultraviolet Lithography. McGraw-Hill, San Francisco (2009)
2. Wurm, S.: EUV lithography development and research challenges for the 22 nm half-pitch. J. Photopolym. Sci. Technol. **22**(1), 31–42 (2009)
3. Bakshi, V.: EUV Lithography. SPIE Press, Bellingham (2009)
4. Kemp, K., Wurm, S.: EUV lithography. C. R. Phys. **7**, 875–886 (2006)
5. Wurm, S.: Extreme ultraviolet lithography development in the United States. Jpn. J. Appl. Phys. **46** (B9), 6105–6112 (2007)
6. Wu, B., Kumar, A.: Extreme ultraviolet lithography: a review. J. Vac. Sci. Technol. B **25**(6), 1743–1761 (2007)
7. Gwyn, C.W., Stulen, R., Sweeney, D., Attwood, D.: Extreme ultraviolet lithography. J. Vac. Sci. Technol. B **16**(6), 3142–3149 (1998)
8. Chkhalo, N.I., Salashchenko, N.N.: Projection XEUV-nanolithography. Nucl. Instrum. Method Phys. Res. Sect. A **603**, 147–149 (2009)
9. Fay, B.: Advanced optical lithography development, from UV to EUV. Microelectron. Eng. **61/62**, 11–24 (2002)
10. Lin, B.J.: The ending of optical lithography and the prospects of its successors. Microelectron. Eng. **83**, 604–613 (2006)
11. Rothschild, M., Bloomstein, T.M., Fedynshyn, T.H., Kunz, R.R., Liberman, V., Switkes, M., Efremow, N. N., Palmacci, S.T., Sedlacek, J.H.C., Hardy, D.E., Grenville, A.: Recent trends in optical lithography. Lincoln Lab. J. **14**(2), 221–236 (2003)

12. Ronse, K.: Optical lithography – a historical perspective. C. R. Phys. **7**, 844–857 (2006)
13. Moore, G.E.: Cramming more components onto integrated circuits. Electronics **38**, 114–117 (1965)
14. Bakshi, V.: EUV Sources for Lithography. SPIE Press, Bellingham (2006)
15. Feigl, T., Yulin, S., Benoit, N., Kaiser, N.: EUV multilayer optics. Microelectron. Eng. **83**, 703–706 (2006)
16. Mertens, B., Weiss, M., Meiling, H., Klein, R., Louis, E., Kurt, R., Wedowski, M., Trenkler, H., Wolschrijn, B., Jansen, R., Runstraat, A., Moors, R., Spee, K., Ploger, S., Kruijs, R.: Progress in EUV optics lifetime expectations. Microelectron. Eng. **73–74**, 16–22 (2004)
17. Brainard, R.L., Barclay, G.G., Anderson, E.H., Ocola, L.E.: Photoresists for next generation lithography. Microelectron. Eng. **61/62**, 707–715 (2002)
18. Kozawa, T., Tagawa, S.: Radiation chemistry in chemically amplified photoresists. Jpn. J. Appl. Phys. **49**, 030001 (2010)
19. Dentinger, P.M.: Outgassing of photoresist materials at extreme ultraviolet wavelengths. J. Vac. Sci. Technol. B **18**(6), 3364–3370 (2000)
20. ASML's customer magazine: ASML Images. Summer Edition. Veldhoven (2010)

Evanescent-Wave Particle Visualization in Nano- and Microchannels

Minami Yoda
G. W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Synonyms

Nano-particle image velocimetry; Total internal reflection (fluorescence) velocimetry

Definition

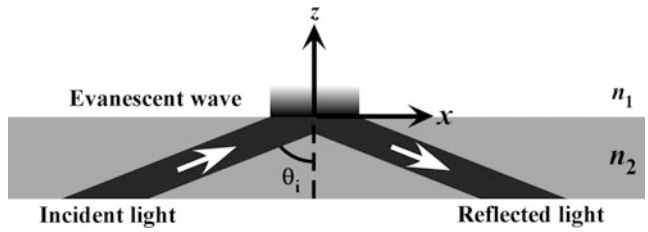
Evanescent-wave particle visualization is a technique for studying the dynamics of fluorescent colloidal particles, or spheres, of average radii a ranging from about 100 nm to a few μm. These particles, which are typically suspended in an aqueous buffer solution at volume fractions $\varphi < 1$ %, are illuminated by evanescent waves generated by the total internal reflection (TIR) of light at the refractive-index interface between a wall and the buffer, for example. In a flowing suspension, the technique can be used to measure near-wall velocities by tracking the motion of the particles over time. Moreover, the technique can also be used to measure the distribution of, and hence the forces acting upon, colloidal particles near the wall because the particle-wall separation h can be estimated from the brightness of the particle image.

Introduction

Surface forces are significant in the flows typical of micro- and nanofluidic devices with their relatively large surface areas and small volumes. Interfacial transport, specifically transport within the region less than 1 μm from the wall where surface forces are significant, is therefore of interest in the steady, creeping, and incompressible flows found in most micro- and nanofluidic devices.

From a fluid mechanics viewpoint, the most important parameter is flow velocity field. The energy required to drive an aqueous solution through nanochannels and microchannels, for example, is determined by the velocity gradient at the channel wall. Measuring this velocity field requires techniques with a spatial resolution at least an order of magnitude less than the overall dimension of the nanochannel or microchannel.

The most common velocimetry technique used in microfluidic devices, microparticle image velocimetry (μPIV), measures fluid velocity distributions by following the motion of suspended fluorescent polystyrene (PS) particles less than 3 μm in diameter with a density (1.05 g/cm^3) which is very close to that of most aqueous solutions, assuming that the particles follow the flow [1, 2]. Micro-PIV can measure the two in-plane components of the velocity without disturbing the flow by illuminating the entire flow and imaging a single two-dimensional plane of the flow. And, because it can use the same optical path to illuminate and image the flow, μPIV requires only the addition of a single optical window to the microfluidic device.



Evanescent-Wave Particle Visualization in Nano- and Microchannels, Fig. 1 Total internal reflection of light at the planar interface between two dielectric media. The reflected beam is slightly offset along the x-direction by the Goos-Hänchen shift

In the near-wall region, the velocity is mainly parallel to the wall because the fluid cannot penetrate the wall. Imaging the velocity field within 1 μm of the wall in a plane parallel to the wall with μPIV is difficult, however, because the thickness of the plane imaged by the technique, based on the "depth of correlation," is at least 2 μm, and optical interference (e.g., glare, reflections) greatly reduces the quality of the image.

Evanescent waves, however, can (as reviewed later) be used instead to illuminate only the flow within the first 0.5 μm next to the wall. They can therefore be used to visualize, and track, only the particles that are within the region of interest for studying surface forces.

As noted earlier, the particles are assumed to be passive tracers that follow the flow in particle velocimetry. This is not the case for colloidal particles, however, because these particles have significant Brownian diffusion. And near the wall, Brownian fluctuations will not necessarily have a mean of zero because the diffusion varies with distance from the wall.

Furthermore, the particle concentration is not uniform near the wall and varies along the wall-normal direction even in a suspension at rest due to particle-wall electric double layer and van der Waal interactions predicted by the classic Derjaguin-Landau-Verwey-Overbeek (DLVO) theory and additional forces associated with the pressure and voltage gradients often used to drive flows in micro- and nanofluidic devices. These phenomena (discussed in the next sections), which play a significant role for suspended particles within 1 μm of the wall, should be considered when determining the flow velocity field from these particle velocity data.

Key Concepts

Evanescent Waves

From basic optics, it is known that when light of wavelength λ is incident upon the planar interface between two homogeneous dielectric media of refractive indices n_1 and n_2 (where $n_2 > n_1$) at an angle of incidence θ_i exceeding the critical angle $\theta_c \equiv \sin^{-1}(n_1/n_2)$, TIR occurs, where all of the incident light is reflected, remaining in the more dense medium of refractive index n_2 (Fig. 1). The Fresnel conditions for TIR, however, give a nonzero amplitude transmission coefficient, and so there is a wave transmitted into the lower-refractive-index medium that cannot, on average, carry any energy across the interface [3].

This transmitted *evanescent wave*, which propagates parallel to the interface along the x-axis, has a purely imaginary wavenumber, and hence an intensity that decays exponentially as it propagates into the less dense medium of refractive index n_1. The intensity of the evanescent wave is then

$$I = I_o \exp\left\{-\frac{z}{z_p}\right\} \tag{1}$$

where I_o is the intensity of the evanescent wave at the interface, z is the distance normal to the interface measured from the interface, and z_p, the intensity-based penetration depth, is given by

$$z_p = \frac{\lambda}{4\pi}\left[n_2^2 \sin^2\theta_i - n_1^2\right]^{-1/2} \tag{2}$$

The illumination incident upon the interface should be linearly p-polarized to ensure that

Eq. 1 is valid [4]. The intensity of the evanescent wave is maximum at the interface, and this maximum value can be several (more than four) times the intensity of the light incident upon the interface [5]. Based on Eq. 1, the evanescent wave only illuminates the region within a few penetration depths of the wall. Evanescent waves, because they are confined to a region less than 1 μm from the wall, have been used in total internal reflection fluorescence (TIRF) microscopy to visualize cell-substrate contact and measure the binding kinetics of proteins to cell surface receptors in biophysics [5] and measure colloidal forces in surface science [6].

Particle-Wall Interactions in a Dilute Colloidal Suspension at Rest

Even in the absence of flow, it is well known in colloid science that colloidal particles near a wall suspended in a quiescent conducting medium (e.g., an aqueous electrolyte solution) will have a nonuniform distribution along the wall-normal direction, even for a dilute suspension where interparticle interactions are negligible, due to electrostatic interactions between the screening layers of mobile counterions that surround the particle and wall surfaces and van der Waals interactions due to dipole interactions. Gravitational interactions are for the most part negligible for these nearly neutrally buoyant particles.

When the solid surface (e.g., of the wall or the particle) is exposed to an aqueous solution, the surface will become charged, for example, by dissociation of surface groups or adsorption of ions. This charged surface will then in turn attract mobile counterions from the solution to form a charged screening layer, known as the electric double layer (EDL), whose thickness is characterized by the Debye length scale κ^{-1}, which is at most $O(10\text{ nm})$ for most aqueous solutions [7]. For a glass microchannel wall and PS particles, the wall and particle surfaces will become negatively charged for aqueous solutions at moderate pH values, and the particle-wall surfaces, will then have a repulsive interaction between their EDLs over distances of order κ^{-1} that depends upon the particle and wall zeta-potentials, which are also affected by the ionic strength and temperature of the solution.

The particle and wall surfaces will also experience attractive intermolecular van der Waals interactions. The combination of EDL repulsion mediated by van der Waals attraction, described by the DLVO theory [7], leads to a strongly nonuniform particle distribution along the wall-normal direction, with a particle depletion zone immediately next to the wall.

This steady-state nonuniform particle distribution should be a Boltzmann distribution, and so the potential energy of the particle-wall interactions $\phi(h)$ can be written in terms of the particle number density $c(h)$:

$$\frac{\phi(h)}{kT} = -\ln\left\{\frac{c(h)}{c_\infty}\right\} \tag{3}$$

where h is the particle-wall separation (so a particle touching the wall corresponds to a separation $h = 0$), k is Boltzmann's constant, T is the absolute temperature of the fluid, and the particle number density recovers to its nominal value c_∞ in the bulk of the flow far away from the wall, or for $h \to \infty$. The magnitude of the wall-normal "lift" force acting upon the particle is then the (magnitude of the) derivative of ϕ.

Hindered Brownian Diffusion

Colloidal particles with radii $a \leq 0.5$ μm suspended in an unbounded fluid are subject to significant Brownian diffusion due to Stokes' drag and thermal effects. The stochastic fluctuations due to Brownian diffusion are described by the Stokes-Einstein diffusion coefficient:

$$D_\infty = \frac{kT}{6\pi\mu a} \tag{4}$$

where k is the Boltzmann constant, T is the absolute temperature of the fluid, and μ is the viscosity of the fluid. In the near-wall region, however, the fluctuations in the particle position due to Brownian diffusion are reduced by hydrodynamic interactions with the wall. Moreover, the presence

of the wall hinders diffusion normal and parallel to the wall differently. For a particle whose center is at a wall-normal distance $z = O(a)$ from the wall, the correction factor for diffusion parallel to the wall is given by the Faxén relation [8]:

$$\beta_{\parallel} \equiv \frac{D_{\parallel}}{D_{\infty}} = 1 - \frac{9}{16}\frac{a}{z} + \frac{1}{8}\left(\frac{a}{z}\right)^3 - \frac{45}{256}\left(\frac{a}{z}\right)^4 - \frac{1}{16}\left(\frac{a}{z}\right)^5 \tag{5}$$

where $D_{\parallel}$ is a coefficient for hindered diffusion parallel to the wall. The correction factor for diffusion normal to the wall, which is an infinite series in z/a [9], can be approximated by regression of the series [10] as

$$\beta_{\perp} \equiv \frac{D_{\perp}}{D_{\infty}} = \frac{6h^2 + 2ah}{6h^2 + 9ah + 2a^2} \tag{6}$$

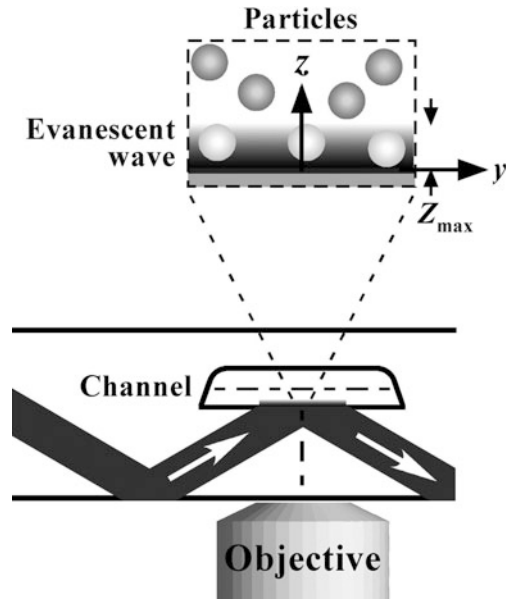
where $D_{\perp}$ is the coefficient for hindered diffusion normal to the wall and $h = z - a$ is the particle edge-wall separation. Near the wall, the diffusion coefficients become functions of the particle wall-normal position. Given that this position will change over time due to diffusion normal to the wall, the near-wall fluctuations of particles due to Brownian diffusion will not have a mean of zero, unlike the fluctuations described by Eq. 4.

The effects of hindered Brownian diffusion have been modeled using Fokker-Planck and Langevin approaches [11, 12], and the discrepancy between the particle and fluid velocities due to Brownian diffusion increases with a dimensionless group $\Pi \equiv D_{\infty}(\Delta t)/Z_{\max}^2$ that compares the displacements due to Brownian diffusion with the wall-normal dimension of the region illuminated by the evanescent wave for $\Pi > 1$. The effects of Brownian diffusion can therefore be reduced by minimizing Δt, the time interval between successive images.

The particles will also lag the flow if a large velocity gradient, or shear, causes the particles to rotate. The velocity of a sphere U near a solid planar wall convected by a simple shear flow rate of $\dot{\gamma}$ will be reduced by 3 % at normalized particle edge-wall separations $h/a = 2$ and by 8 % at $h/a = 0.5$ [13]. Particle rotation therefore only has a significant effect on particle dynamics very close to the wall, where there are relatively few particles due to repulsive particle-wall EDL interactions.

Evanescent-Wave Particle Visualization

Evanescent waves generated by the TIR of a laser beam at the interface between the glass wall of the channel and the flowing aqueous solution can be used to illuminate fluorescent PS particles of radius $a = 50$–500 nm suspended in the flow (Fig. 2). This illumination can be generated either by coupling the laser beam into the channel using a prism or by using a high–numerical aperture TIRF microscope objective. For blue light at $\lambda = 488$ where $n_1 = 1.34$ (i.e., the refractive index of water) and $n_2 = 1.46$ (i.e., the refractive index of fused silica), $\theta_c = 66.6^\circ$ and $z_p \approx 100$–200 nm by Eq. 2 for angles of incidence a few degrees greater than the critical angle. The fluorescence from the particles in the x-y plane parallel to the bottom wall is imaged



Evanescent-Wave Particle Visualization in Nano- and Microchannels, Fig. 2 Sketch of the cross section of flow through a glass microchannel containing suspended fluorescent particles. The particles are illuminated by evanescent waves from the TIR of a laser beam at the glass wall-fluid interface and imaged through the same (*lower*) wall of the channel. The flow and x direction point out of the page

from below through a microscope objective by an intensified or electron multiplying CCD (ICCD and EMCCD, respectively) camera.

The maximum z-distance from the wall Z_{max} where tracer particles are visible in the images is usually determined by the noise characteristics of the camera. For a 16-bit EMCCD camera image, for example, the background noise can be as great as 8 % of the maximum grayscale value. In this case, $Z_{max} = 3z_p$ based on Eq. 1, since the images of particles beyond that distance cannot be distinguished from the background noise in the images.

The x- and y-locations of the center of any particle image can be determined, for example, with a resolution of ~0.1 pixel by identifying the center of the correlation peak after cross-correlating the particle image with an axisymmetric Gaussian function. Moreover, the wall-normal, or z-, location of the particle can be determined from the grayscale value of the particle image. If the intensity I_p of the image of a fluorescent particle excited by evanescent waves also follows Eq. 1, the particle edge-wall separation is

$$h = z_p \ln\left\{\frac{I_p^0}{I_p}\right\} \tag{7}$$

where I_p^0 is the intensity of the image of a particle touching the wall (i.e., at $h = 0$). Note that particle-to-particle variations can introduce significant variations in I_p^0.

In most cases, the z-component of the velocity is negligible this close to the wall. The two other (x- and y-) components are typically obtained by estimating the displacements between a series of images separated by a time interval Δt using particle-tracking techniques [14] or by simply matching the particles in the first image to their nearest neighbor in the second image in cases where φ and Δt are small enough for the particle displacements to be less than the average interparticle distance. Each particle displacement is then assigned to its average wall-normal position over the two images in the pair.

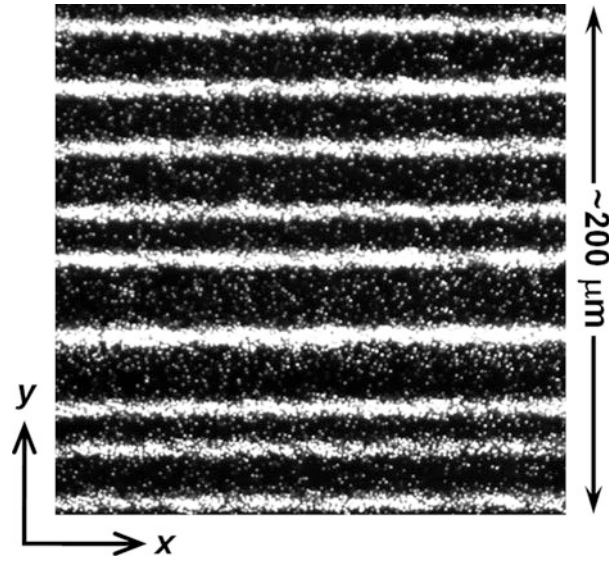
In multilayer nanoparticle tracking velocimetry (MnPTV) [15], one can obtain a velocity profile, typically consisting of three data points, all still at $z \leq Z_{max}$, which accounts for the nonuniform distribution of the particles. The velocity data from a sequence of images of a steady or periodic flow are divided, based upon their average wall-normal positions, into a few (usually three) layers and averaged over all the velocities in each layer. The wall-normal position of this average velocity is then also the average of the wall-normal positions of all the velocity samples in that layer. By effectively reducing the wall-normal dimension of the velocity data to that of a single layer (vs. Z_{max}), MnPTV should also be less affected by Brownian diffusion.

The particle edge-wall separations determined from Eq. 7 for a large ensemble ($O(10^4)$) of particles also can be used to estimate the steady-state distribution of the particles along the wall-normal direction. One of the more recent applications of evanescent-wave particle visualization in microfluidics entails directly measuring this distribution, quantified by the particle number density profile $c(h)$, and studying the differences in the near-wall distribution of colloidal particles in flowing, versus quiescent, dilute colloidal suspensions to determine the additional lift forces on near-wall particles due to the flow itself.

Examples

Over the past decade, a variety of phenomena have been studied using evanescent-wave particle visualization. A number of groups have studied hindered Brownian diffusion of particles ranging from $a = 100$ nm to 1.5 μm both parallel and normal to the wall [16–18] and for the most part reported results in good agreement with the predictions of Eqs. 5 and 6.

A variety of evanescent-wave particle velocimetry approaches using fluorescent PS tracers have been used to study slip of Newtonian liquids in Poiseuille flow through microchannels. Slip lengths as great as 96 nm have been observed in hydrophobic channels at $\dot{\gamma} = 1800\,s^{-1}$ [19], while other studies have reported smaller values for slip lengths, as great as 23 nm, in hydrophobic channels at $\dot{\gamma} = 2250\,s^{-1}$ [15]. Accounting for the



Evanescent-Wave Particle Visualization in Nano- and Microchannels, Fig. 3 Image of "bands" of $a = 245$ nm fluorescent PS particles suspended in an aqueous solution of 1 mM sodium tetraborate flowing through a ~30 μm deep fused-silica channel at $\dot{\gamma} = 1100\,s^{-1}$ and a dc electric field magnitude of 47 V/cm. The particles in these bands, which are aligned with the Poiseuille and EO flow directions, are illuminated by evanescent waves, and only particles at $h < 360$ nm are visible in this image

nonuniform near-wall distribution tracers significantly reduced the slip lengths estimated in the latter study, which noted that assuming uniformly distributed tracers would have easily tripled the slip length. Most recently, the technique has been used to study transport in nanochannels, with slip lengths as great as 189 nm reported in 410 nm deep hydrophilic channels [20]. Evanescent-wave particle velocimetry has also been used to measure velocity fields in electroosmotic (EO) flows driven by a voltage gradient or electric field [21] and estimate velocity profiles inside the EDL [22].

Other types of particle-like tracers have also been used to study near-wall flow dynamics. Near-wall flow velocities have been estimated from the displacements of both fluorescently labeled λDNA molecules [23] and ZnS-overcoated CdSe quantum dots [24] illuminated by evanescent waves, for example.

Recently, evanescent-wave particle visualization has also been used to study the near-wall dynamics of the particles themselves, versus the convection of these particles by the flow. It has been known for almost two decades that $a \approx 2$–5 μm neutrally buoyant particles in a dilute suspension are subject to so-called electrokinetic (or electroviscous) and hydrodynamic lift forces in shear (e.g., Poiseuille and Couette) flows that drive them away from the wall [25, 26]. More recently, both theoretical and experimental studies have shown that suspended particles are subject to "dielectrophoretic-like" forces due to a steady electric field parallel to the wall in electroosmotic flows [27, 28]. Evanescent-wave visualizations of dilute colloidal suspensions of $a \approx 100$–500 nm fluorescent PS and colloidal silica particles in both Poiseuille and EO flow through fused-silica microchannels have shown that these smaller particles are also subject to repulsive lift forces of O(1–10 fN) [29, 30] and that they can even be self-assembled into "bands" (Fig. 3) using a combination of Poiseuille and EO flow [31].

Cross-References

- ▶ Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering
- ▶ Electric Double Layer Capacitor

References

1. Adrian, R.J., Westerweel, J.: Particle Image Velocimetry. Cambridge University Press, Cambridge (2011)
2. Wereley, S.T., Meinhart, C.D.: Recent advances in micro-particle image velocimetry. Annu. Rev. Fluid Mech. **42**, 557–576 (2010)

3. Hecht, E.: Optics, 3rd edn. Addison-Wesley, Reading (1998)
4. Hertlein, C., Riefler, N., Eremina, E., Wriedt, T., Eremin, Y., Helden, L., Bechinger, C.: Experimental verification of an exact evanescent light scattering model for TIRM. Langmuir **24**, 1–4 (2008)
5. Axelrod, D., Burghardt, T.P., Thompson, N.L.: Total internal reflection fluorescence. Annu. Rev. Biophys. Bioeng. **13**, 247–268 (1984)
6. Prieve, D.C.: Measurement of colloidal forces with TIRM. Adv. Colloid Interf. Sci. **82**, 93–125 (1999)
7. Israelachvili, J.: Intermolecular and Surface Forces, 2nd edn. Academic, London (1992)
8. Faxén, H.: Der Widerstand gegen die Bewegung einer starren Kugel in einer zähen Flüssigkeit, die zwischen zwei parallelen Ebenen Wänden eingeschlossen ist. Ann. Phys. **4**, 89–119 (1922)
9. Brenner, H.: The slow motion of a sphere through a viscous fluid towards a plane surface. Chem. Eng. Sci. **16**, 242–251 (1961)
10. Bevan, M.A., Prieve, D.C.: Hindered diffusion of colloidal particles very near to a wall: revisited. J. Chem. Phys. **113**, 1228–1236 (2000)
11. Sadr, R., Hohenegger, C., Li, H., Mucha, P.J., Yoda, M.: Diffusion-induced bias in near-wall velocimetry. J. Fluid Mech. **577**, 443–456 (2007)
12. Huang, P., Guasto, J.S., Breuer, K.S.: The effects of hindered mobility and particle depletion of particles in near-wall shear flows and the implications for nanovelocimetry. J. Fluid Mech. **637**, 241–265 (2009)
13. Goldman, A.J., Cox, R.G., Brenner, H.: Slow viscous motion of a sphere parallel to a plane wall. Part II. Couette flow. Chem. Eng. Sci. **22**, 653–660 (1967)
14. Cierpka, C., Lütke, B., Kähler, C.J.: Higher order multi-frame particle tracking velocimetry. Exp. Fluids **54**, 1533 (2013)
15. Li, H.F., Yoda, M.: An experimental study of slip considering the effects of nonuniform colloidal tracer distributions. J. Fluid Mech. **662**, 269–287 (2010)
16. Choi, C.K., Margraves, C.H., Kihm, K.D.: Examination of near-wall hindered Brownian diffusion of nanoparticles: experimental comparison to theories by Brenner (1961) and Goldman et al. (1967). Phys. Fluids **19**, 103305 (2007)
17. Huang, P., Breuer, K.S.: Direct measurement of anisotropic near-wall hindered diffusion using total internal reflection velocimetry. Phys. Rev. E **76**, 046307 (2007)
18. Kazoe, Y., Yoda, M.: Measurements of the near-wall hindered diffusion of colloidal particles in the presence of an electric field. Appl. Phys. Lett. **99**, 124104 (2011)
19. Huang, P., Guasto, J.S., Breuer, K.S.: Direct measurement of slip velocities using three-dimensional total internal reflection velocimetry. J. Fluid Mech. **566**, 447–464 (2006)
20. Kazoe, Y., Iseki, K., Mawatari, K., Kitamori, T.: Evanescent wave-based particle tracking velocimetry for nanochannel flows. Anal. Chem. **85**, 10780–10786 (2013)
21. Sadr, R., Yoda, M., Zheng, Z., Conlisk, A.T.: An experimental study of electro-osmotic flow in rectangular microchannels. J. Fluid Mech. **506**, 357–367 (2004)
22. Sadr, R., Yoda, M., Gnanaprakasam, P., Conlisk, A.T.: Velocity measurements inside the diffuse electric double layer in electroosmotic flow. Appl. Phys. Lett. **89**, 044103 (2006)
23. Gai, H., Li, Y., Silber-Li, Z., Ma, Y., Lin, B.: Simultaneous measurements of the flow velocities in a microchannels by wide/evanescent field illuminations with particle/single molecules. Lab Chip **5**, 443–449 (2005)
24. Pouya, S., Koochesfahani, M., Snee, P., Bawendi, M., Nocera, D.: Single quantum-dot (QD) imaging of fluid flow near surfaces. Exp. Fluids **39**, 784–786 (2005)
25. Bike, S.G., Lazarro, L., Prieve, D.C.: Electrokinetic lift of a sphere moving in a slow shear flow parallel to a wall. I. Experiment. J. Colloid Interface Sci. **175**, 411–421 (1995)
26. Williams, P.S., Moon, M.H., Xu, Y., Giddings, J.C.: Effect of viscosity on retention time and hydrodynamic lift forces in sedimentation/steric field-flow fractionation. Chem. Eng. Sci. **51**, 4477–4488 (1996)
27. Yariv, E.: "Force-free" electrophoresis? Phys. Fluids **18**, 031702 (2006)
28. Liang, L., Ai, Y., Zhu, J., Qian, S., Xuan, S.J.: Wall-induced lateral migration in particle electrophoresis through a rectangular microchannel. J. Colloid Interface Sci. **347**, 142–146 (2010)
29. Kazoe, Y., Yoda, M.: An experimental study of the effect of external electric fields on interfacial dynamics of colloidal particles. Langmuir **27**, 11481–11488 (2011)
30. Cevheri, N., Yoda, M.: Lift forces on colloidal particles in combined electroosmotic and Poiseuille flow. Langmuir **30**, 13771–13780 (2014)
31. Cevheri, N., Yoda, M.: Electrokinetically driven reversible banding of colloidal particles near the wall. Lab Chip **14**, 1391–1394 (2014)

Excitonic Solar Cell

▶ Nanomaterials for Excitonic Solar Cells

Exoskeleton

▶ Arthropod Strain Sensors

Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms

Claire Coutris[1] and Erik J. Joner[2]
[1]Department of Plant and Environmental Sciences, Norwegian University of Life Sciences, Ås, Norway
[2]Bioforsk Soil and Environment, Ås, Norway

Synonyms

Bioaccessibility/bioavailability; Lumbricidae/oligochaeta/earthworms; Soil/terrestrial ecosystem/terrestrial compartment; Toxicity/ecotoxicity; Uptake/internalization/sequestration/biodistribution

Definition

Exposure and toxicity are the two variables that are required to assess the risk that a substance poses to an organism or the environment. Exposure determines to which extent an organism enters in contact with a substance, while toxicity measures to which extent an organism is impaired by the substance.

Background

The increasing use of engineered nanomaterials (ENMs) in consumer products is leading to a new type of environmental pollution for which the knowledge related to environmental hazards and risks is very limited. Toxicity and even ecotoxicity have been examined for a range of ENMs, and a majority of the reports conclude that several ENMs have toxic properties under certain conditions. Yet, exposure conditions seem to govern the outcome of such studies which range from instillation into mammalian airways to exposure of ENMs mixed into soil.

Soil is a particular matrix. It is recognized as a major recipient and sink for environmental pollution with ENMs [1], particularly through application of sewage sludge which contains most of the ENMs entering wastewater [2]. But soil is also an environmental compartment that is essential for food production and a number of important ecological and biogeochemical functions. Soil is an extremely complex matrix, with high specific surface area, a large amount of charged surfaces, abundant amorphous colloidal material, and a wide range of dissolved ions and dissolved organic matter, all of which eludes precise characterization. From one soil to another, there may be a large variation with respect to composition and properties due to differences in parent material and the extent and stage of biological and climate-dependent soil-forming processes. The complexity and diversity of soils is a challenge in terrestrial ecotoxicity studies, as soil properties will influence bioavailability and thereby exposure of organisms and ultimately the apparent toxicity to ENMs. Which properties are most important for ENM bioavailability probably depend on the ENM in question, but remains to be examined in detail.

Differences in soil composition translate into differences in soil properties that can be quantified by physical and chemical analyses. Though no systematic examination exist of which soil properties affect the bioavailability of different ENMs, certain soil properties that affect the bioavailability of other environmental pollutants are likely to be important. These are, therefore, reported in soil-based ecotoxicity studies on ENMs and include, or should include, texture (particle size distribution), organic matter content, pH, cation exchange capacity (CEC), and base saturation (BS; quantity of exchangeable Na, K, Mg, and Ca as a percentage of CEC). Even though other parameters may be equally important or provide more detailed information that may be used to explain interactions between a soil and ENMs, these are often too specialized to be available in most soil science labs (e.g., equipment for characterization of specific surface area, clay mineralogy, or soil organic matter quality).

When ENMs are characterized, specific surface area has become a useful parameter believed to predict both surface reactivity, capacity to

interact with dissolved ions and molecules, as well as propensity for dissolution. Small nanoparticles typically have specific surface areas (measured by BET; the Brunauer-Emmett-Teller method) in the order of 30–50 $m^2\ g^{-1}$ [3]. Though impressive it may seem, soil constituents like clay may have BET values an order of magnitude higher, ranging from 50 to 1,100 $m^2\ g^{-1}$. Also humic soil organic matter has BET values in the order of 500–800 $m^2\ g^{-1}$. As soils commonly contain at least 3–5 % organic matter and often at least 10 % clay, there are ample possibilities for interactions between ENMs and the soil matrix, which may affect exposure and toxicity to soil organisms.

Standard OECD and ISO test guidelines for soil organisms like earthworms commonly exclude the use of soil, and those standardized tests that permit the use of soil prescribe the use of soil types that are low in clay and organic matter. Artificial "soils" are frequently used in which weakly transformed peat constitute the organic fraction and sand and low swelling clay represent the mineral fraction. Not only are these "soils" poor as proxies for real soil with respect to surface area and surface charge, but they are also void of organisms which constitute an integral part of natural soils.

To test the impact of environmental toxicants on soil organisms, earthworms have frequently been used due to the ecological importance of their activity related to soil organic matter turnover and their burrowing that ensures soil porosity [4]. Earthworms are also convenient as test animals, as several stress-related and toxicological endpoints may be used in tests of limited duration to assess both acute and chronic effects at different levels (behavioral, physiological, reproductive, and genetic).

Eisenia fetida is a model earthworm used in standardized soil toxicity tests issued by the US-EPA and the OECD. This species has a short generation time and reproduces easily under lab conditions. Part of its genome is also sequenced for use in transcriptomic studies. Yet, the extensive use of this species is often criticized as it is epigeic (living in organic surface layer or compost), rather than endogeic or anecic (living and burrowing in soil, *stricto sensu*).

Exposure

The question of possible risks of ENMs does not depend solely on their toxicity, but also requires information on exposure (Risk = Hazard × Exposure). The latter aspect has, however, received far less attention than toxicity studies, probably due to methodological difficulties, as ENMs are difficult to track and quantify in organisms and their tissues, and even more so in complex media like water, sediments, soil, and waste. Exposure is somewhat differently defined in toxicology and ecotoxicology. In higher organisms in vivo, exposure is often concerned with toxicokinetic aspects of absorption, distribution, metabolism, and excretion (ADME), though the route of exposure is also given (oral, dermal, respiratory, etc.). In ecotoxicology, the perception of exposure is much closer to the colloquial understanding of the word and implies to which extent an organism gets in contact with a toxic compound, ingests it and excretes it. Two earthworm species living in the same soil but having different feeding habits (e.g., anecic vs. endogeic) may thus be exposed to ENMs in surface applied sewage sludge in very different ways. Also, a worm living in a clay-rich soil containing ENMs may be differently exposed compared to the same worm species living in a sandy soil containing the same type and amount of ENMs. This aspect of exposure is closely related to the concept of bioavailability, which is taking into account the interactions between a nutrient or a toxic compound and its surrounding matrix. Studies on earthworm exposure to ENMs have sought to vary both the way ENMs are provided (mixed into soil or feed) and matrix composition (soil vs. water, and soil composition). Also, pure measurements describing uptake and excretion have been made for some ENMs. The few earthworm exposure studies that exist are described under each category of ENM, together with data on ecotoxicity.

Experimental Evidence

Several types of metal and metal oxide nanomaterials have been studied for their

potential bioavailability and toxicity to earthworms, namely, Ag, Au, Co, Cu, Fe, Ni, Al_2O_3, SiO_2, TiO_2, ZnO, and ZrO_2. Most of these studies used the earthworm *E. fetida*, while others used *Lumbricus terrestris*, *L. rubellus*, or *E. veneta*. Also tests using the closely related pot worm *Enchytraeus albidus* have provided relevant data in similar tests, since this species also belongs to the subclass Oligochaeta. A variety of endpoints are reported, from survival, body burden, and reproduction to apoptosis and expression of genes involved in metal homeostasis and oxidative stress. State-of-the-art results are presented for each metal and metal oxide type below.

Ag

A variety of silver nanoparticles (Ag NPs) with a range of sizes (9–50 nm diam.) and coatings (citrate, oleic acid, polyvinylpyrrolidone, or none) have been exposed to earthworms through water, food, or different types of soil at concentrations ranging from 0.3 to 1,000 mg kg^{-1}. Bioaccumulation has generally been higher for $AgNO_3$ than for Ag NPs, and the various sizes and coatings tested have affected bioaccumulation to no or only limited extent. What has made a difference is the soil type used in these tests, so that higher bioaccumulation of Ag NPs has been observed in soil with lower organic matter and higher sand content. In artificial OECD soil, bioaccumulation factors (BAFs, being the concentration ratio of accumulated ENMs/soil ENMs) have been very low and seem to indicate that Ag NPs measured in earthworms merely result from incomplete gut clearance after exposure. In soil with lower organic matter and higher sand content, Ag NPs have been more bioaccessible and BAFs higher (BAF = 0.05–0.08 in sandy loam while <0.03 in OECD soil) [5, 6]. Survival and growth of worms were not affected by Ag NP concentrations as high as 1,000 mg kg^{-1}, contrary to $AgNO_3$, which reduced survival to approx. 3 % at the same concentration [7]. Reproduction measured as cocoon production has been affected at lower concentrations with significant reductions at ≥94 mg kg^{-1} for $AgNO_3$, and at ≥800 mg kg^{-1} Ag NPs in an experiment by Shoults-Wilson et al. [6]. In this experiment, no differences were found between particles of different size or with different coatings. A more sensitive endpoint yet is the measurement of avoidance, as demonstrated in an experiment by Shoults-Wilson et al. [8] where worms given the choice between uncontaminated soil and soil spiked with 7 mg kg^{-1} Ag NPs, avoided the latter when measured after 48 h exposure. A similar concentration of $AgNO_3$ induced the same response, but faster. This concentration is close to predicted environmental concentrations of Ag NPs in sludge from sewage treatment plants in Europe and the United States (1.29–6.24 mg kg^{-1}) [1]. Also enhanced frequency of apoptosis measured in the cuticle, the intestinal epithelium, and the chloragogenous tissue of earthworms (*L. terrestris*) exposed to concentrations as low as 4 mg kg^{-1} in soil indicate that earthworms may be negatively affected by environmentally relevant concentrations of Ag NPs [9]. The two latter reports do, however, not prove any lasting negative impact, but may simply be due to stress from which the worms can recover in time. The studies conducted so far have been based on freshly spiked media. It appears, however, that Ag NPs may become more bioavailable than $AgNO_3$ over time as Ag NPs can act as constant sources releasing ionic and bioavailable Ag. This is indicated, e.g., by the results of Shoults-Wilson et al. [10] who showed that 11–20 % of Ag in 30–50 nm Ag NPs was oxidized after 28 days in OECD soil, and supported by results from sequential chemical extractions of soils spiked with different types of Ag NPs (Coutris et al., unpublished results). Until more sensitive endpoints are examined, e.g., related to transcriptomic analyses, that can confirm or preclude proper toxicity effects of environmentally relevant AgNP concentrations, conclusions must be drawn from the physiological and reproductive endpoints available. Based on such data it seems like Ag NPs have a low toxicity and low bioavailability to earthworms.

Au

Gold NPs (4 and 18 nm diam., coated with citrate) have been found in the gut and the skin of earthworms exposed for 1 week to Au NPs in artificial OECD soil spiked with 3 mg Au kg^{-1}. A similar

body distribution was found in worms exposed 48 h on filter paper to a 10 mg L^{-1} Au NPs suspension [11]. In this study, Au uptake was particle size dependent, with uptake of Au ions being the highest followed by 4 and 18 nm Au NPs. In another experiment with Au NPs (20 and 55 nm diam., 5–50 mg kg^{-1} soil) Unrine et al. [12] confirmed that Au NPs were bioavailable to *E. fetida* through exposure in artificial OECD soil, and that Au ions were taken up to a larger extent than Au NPs. However, in this case, no difference in uptake was noticed between Au NPs on a mass basis (whereas on a particle number basis, more 20 nm Au NPs were taken up than 55 nm Au NPs). Although no effects on survival or growth were noticed, both ions and NPs had negative effects on earthworm reproduction. This response was not purely dose dependent, as e.g., 55 nm Au NPs had no negative effect on reproduction at 50 mg Au kg^{-1}, whereas at 20 mg kg^{-1} a negative effect was observed. Also here, ions were more toxic than NPs, but smaller NPs did not appear more toxic than larger ones, and the two size classes tested formed aggregates of similar size. Finally, none of the earthworm genes related to metal homeostasis and oxidative and cellular stress were expressed differently in control and Au NPs treatments, whereas $HAuCl_4$ at $\geq$2.5 mg kg^{-1} resulted in the upregulation of metallothionein [12].

Co

Cobalt NPs (Co core/Co_3O_4 shell, 4 nm diam.) have been found to be readily bioavailable to *E. fetida* exposed though their diet to 1 mg Co kg^{-1} (BAF = 0.69 ± 0.04 for Co NPs and 0.88 ± 0.05 for $CoCl_2$). Co from both particulate and ionic sources was only slowly excreted from the worms (biological half-life 200 days) and showed a rather similar biodistribution within internal organs of the worms [5]. The highest accumulation was found in the blood and the intestinal tract and the least in the reproductive organs. Still, at high concentrations (87 mg kg^{-1} food), Co from Co NPs have been found in reproductive organs and cocoons produced after the worms had been transferred to a non contaminated environment [13].

Cu

Copper NPs (Cu/CuO, 80 nm diam.) added to artificial OECD soil at 1,000 mg kg^{-1} had no acute effect on the survival of *E. fetida*, but induced a dramatic decrease in reproduction, whereas a similar concentration of $CuCl_2$ had both an acute and chronic effect [7]. The same particles have been tested at 450 and 750 mg kg^{-1} on *Enchytraeus albidus*, and indicated that both $CuCl_2$ and Cu NPs caused oxidative stress damage, yet in different ways and with more pronounced effects in the case of $CuCl_2$ [14]. An experiment by Unrine et al. [15], where *E. fetida* was exposed to artificial OECD soil spiked with Cu NPs (20–40 nm and 100 nm diam.), has showed that the expression of metallothionein genes was upregulated at $\geq$20 mg kg^{-1} Cu NPs and $\geq$10 mg kg^{-1} $CuSO_4$, while no effect on the regulation of other genes related to oxidative and general stress was detected. Synchrotron-based X-ray analyses indicated that both types of Cu NPs had been oxidized within 4 weeks so that they contained a mixture of zero-valent Cu, Cu_2O, and CuO. Smaller particles were also more oxidized than larger ones. The authors concluded that oxidized Cu NPs were able to enter the food chain from soil, but that only concentrations >65 mg kg^{-1} would induce adverse effects.

Fe

Ecotoxicity data for nano zero-valent iron (nZVI) on terrestrial macro- and mesofauna are only starting to emerge. Preliminary studies have shown that earthworms are sensitive to nZVI concentrations of $\geq$100 mg kg^{-1}. Both the compost worm *E. fetida*, used in standardized OECD tests, and the epigeic earthworm *Lumbricus rubellus* showed signs of adverse effects during long-term exposure to 100 mg nZVI kg^{-1}, whereas acute (mortality after 14 days exposure in sandy soil) toxicity was observed only at concentrations $\geq$500 mg nZVI kg^{-1} for both species (El-Temsah and Joner, unpublished results).

Ni

The only published study assessing the potential toxicity of Nickel NPs (20 nm diam.) to *E. fetida* did not reveal any acute effects on the survival or

reproduction. The test featured 28 days exposure in artificial OECD soil containing 1,000 mg kg^{-1} Ni NPs. On the other hand, a similar concentration of $NiCl_2$ induced a high mortality and impaired reproduction dramatically (98 % decrease in cocoon production and none of the cocoons hatched) [7].

Al_2O_3

Al_2O_3 NPs (12–14 nm diam.) added to artificial OECD soil at 1,000 mg kg^{-1} did not show any acute effects on the survival or reproduction of *E. fetida* [7]. Another study with similarly sized Al_2O_3 NPs (11 nm) found that cocoon production was reduced only at concentrations $\geq$3,000 mg kg^{-1}, and that avoidance was apparent when concentrations reached $\geq$5,000 mg kg^{-1} [16]. Although no predicted environmental concentrations are available for this type of nanomaterials, it seems unlikely that such concentrations will be found in the environment.

TiO_2

TiO_2 NPs (21 nm diam., 73 % anatase, 27 % rutile) added to artificial OECD soil at 1,000 mg kg^{-1} had no acute effect on the survival of *Eisenia fetida*, but induced a weak but significant adverse effect on reproduction [7]. TiO_2 NPs (10–20 nm diam., rutile) added to artificial OECD soil induced lipid peroxidation, DNA damage, and mitochondria damage only at very high concentrations (5,000 mg kg^{-1}) [17]. Another experiment with *Lumbricus terrestris* exposed to an aged TiO_2 composite (100–300 nm, 1–100 mg kg^{-1}) through water, food, and soil showed neither effect on survival nor bioaccumulation, but enhanced apoptotic frequency in the cuticle, the intestinal epithelium, and chloragogenous tissue of worms exposed to 100 mg kg^{-1} in water [18]. Similar effects were not observed when exposed in soil or feed.

ZnO

No acute toxic effects of Zn on survival were found in *E. veneta* exposed to $ZnCl_2$ or ZnO NPs ($\leq$100 nm diam.) at 250 and 750 mg ZnO kg^{-1} soil for 21 days [19]. Growth was observed in all worms irrespective of Zn source and dose, but was slightly reduced for $ZnCl_2$ at the highest concentration. Cocoon production was reduced by 50 % at 750 mg kg^{-1} in the case of ZnO NPs, whereas the same Zn concentration led to an almost complete inhibition in the case of $ZnCl_2$. No reduction in immune activity of coelomocytes was found in presence of ZnO NPs, even at the highest concentration tested, whereas a 20 % reduction was observed at 750 mg kg^{-1} for $ZnCl_2$. In another study with 7 days exposure of *E. fetida* to ZnO NPs (10–20 nm diam.) Hu et al. [17] observed increased lipid peroxidation at $\geq$100 mg kg^{-1}, reduced activity of superoxide dismutase and biochemical metabolism (measured as catalase activity) at $\geq$500 mg kg^{-1}, induced DNA damage at 1,000 mg kg^{-1}, and mitochondria damage at 5,000 mg kg^{-1}. This study made no comparisons to ionic Zn. A third study where *E. fetida* was exposed to ZnO NPs in agar for 4 days found close to 100 % mortality of ZnO NPs at 1,000 mg ZnO kg^{-1} [20]. The same authors also used a contact test where worms were exposed to ZnO NPs added as an aqueous suspension to filter paper, but this test showed no dose-dependent toxicity. Both the agar test and the filter paper test did, however, show that environmental conditions can influence toxicity, as lower negative effects were observed when ZnO NPs suspensions were prepared with water containing salts or dissolved humic material, as compared to exposure to ZnO NPs suspended in deionized water.

SiO_2 and ZrO_2

To date, the adverse effects of SiO_2 and ZrO_2 nanoparticles have only been examined one study, which did not reveal any detrimental effects on the survival or reproduction of *E. fetida* when exposed to 1,000 mg kg^{-1} SiO_2 NPs (5–15 nm diam.) or ZrO_2 NPs (20–30 nm diam.) in artificial OECD soil [7].

Concluding Remarks

Testing of metallic and oxide ENMs for adverse effects on soil invertebrates has only been attempted for a few years, and the limited amount

of data available, with respect to tested ENMs, earthworm species, ecotoxicological endpoints, and exposure conditions, calls for cautious conclusions. Still, it seems that ENMs made of elements with known antimicrobial properties, like Ag and Cu, are among those that should be treated with most concern as they have so far shown the lowest adverse effect concentrations. Earthworms may not only be affected by direct toxicity of these ENMs, but will also suffer if microorganisms on which they depend are affected. Earthworms strongly depend on microorganisms living both endosymbiotically in the coelomic cavity and in the digestive tract, and any adverse effects on these organisms are bound to affect the worm's physiology and health. Regarding the specific effects of ENM versus the same elements in a soluble form, there are several indications that the effect of ENMs stems, at least partly, from dissolving ions released by from the surface of ENMs. While free ions often react rapidly with other ions or more complex molecules to form insoluble metal species (e.g., AgCl), ENMs may continue to act in a toxic manner by surface interaction or continued release of ions. In this way, ions may appear more acutely toxic, while corresponding ENM may eventually appear to have more chronic effects.

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates
- ▶ Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ Fate of Manufactured Nanoparticles in Aqueous Environment
- ▶ Genotoxicity of Nanoparticles
- ▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Nanoparticle Cytotoxicity
- ▶ Nanoparticles
- ▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity
- ▶ Quantum Dot Toxicity
- ▶ Toxicology: Plants and Nanoparticles

References

1. Gottschalk, F., Sonderer, T., Scholz, R.W., Nowack, B.: Modeled environmental concentrations of engineered nanomaterials (TiO_2, ZnO, Ag, CNT, fullerenes) for different regions. Environ. Sci. Technol. **43**, 9216–9222 (2009)
2. Kiser, M.A., Westerhoff, P., Benn, T., Wang, Y., Perez-Rivera, J., Hristovski, K.: Titanium nanomaterial removal and release from wastewater treatment plants. Environ. Sci. Technol. **43**, 6757–6763 (2009)
3. Auffan, M., Rose, J., Bottero, J.-Y., Lowry, G.V., Jolivet, J.-P., Wiesner, M.R.: Towards a definition of inorganic nanoparticles from an environmental, health and safety perspective. Nat. Nanotechnol. **4**, 634–641 (2009)
4. Edwards, C.A., Bohlen, P.J.: Biology and Ecology of Earthworms. Chapman and Hall, London (1996)
5. Coutris, C., Hertel-Aas, T., Lapied, E., Joner, E.J., Oughton, D.H.: Bioavailability of cobalt and silver nanoparticles to the earthworm *Eisenia fetida*. Nanotoxicology (2012). doi:10.3109/17435390.2011.569094 (online Apr 2011)
6. Shoults-Wilson, W.A., Reinsch, B.C., Tsyusko, O.V., Bertsch, P.M., Lowry, G.V., Unrine, J.M.: Role of particle size and soil type in toxicity of silver nanoparticles to earthworms. Soil Sci. Soc. Am. J. **75**, 365–377 (2011)
7. Heckmann, L.H., Hovgaard, M.B., Sutherland, D.S., Autrup, H., Besenbacher, F., Scott-Fordsmand, J.J.: Limit-test toxicity screening of selected inorganic nanoparticles to the earthworm *Eisenia fetida*. Ecotoxicology **20**, 226–233 (2011)
8. Shoults-Wilson, W.A., Zhurbich, O.I., McNear, D.H., Tsyusko, O.V., Bertsch, P.M., Unrine, J.M.: Evidence for avoidance of Ag nanoparticles by earthworms (*Eisenia fetida*). Ecotoxicology **20**, 385–396 (2011)
9. Lapied, E., Moudilou, E., Exbrayat, J.M., Oughton, D.H., Joner, E.J.: Silver nanoparticle exposure causes apoptotic response in the earthworm *Lumbricus terrestris* (Oligochaeta). Nanomedicine **5**, 975–984 (2010)
10. Shoults-Wilson, W.A., Reinsch, B.C., Tsyusko, O.V., Bertsch, P.M., Lowry, G.V., Unrine, J.M.: Effect of silver nanoparticle surface coating on bioaccumulation and reproductive toxicity in earthworms (*Eisenia fetida*). Nanotoxicology (2011). doi:10.3109/17435390.2010.537382 (online Dec 2010)
11. Unrine, J., Bertsch, P., Hunyadi, S.E.: Bioavailability, trophic transfer, and toxicity of manufactured metal

and metal oxide nanoparticles in terrestrial environments. In: Grassian, V.H. (ed.) Nanoscience and Nanotechnology, pp. 343–364. Wiley, Hoboken (2008)
12. Unrine, J.M., Hunyadi, S.E., Tsyusko, O.V., Rao, W., Shoults-Wilson, W.A., Bertsch, P.M.: Evidence for bioavailability of Au nanoparticles from soil and biodistribution within earthworms (*Eisenia fetida*). Environ. Sci. Technol. **44**, 8308–8313 (2010)
13. Oughton, D.H., Hertel-Aas, T., Pellicer, E., Mendoza, E., Joner, E.J.: Neutron activation of engineered nanoparticles as a tool in studies on their environmental fate and uptake in organisms. Environ. Toxicol. Chem. **27**, 1883–1887 (2008)
14. Gomes, S.I.L., Novais, S.C., Gravato, C., et al.: Effect of Cu-nanoparticles versus one Cu-salt: analysis of stress biomarkers response in *Enchytraeus albidus* (Oligochaeta). Nanotoxicology (2012). doi:10.3109/17435390.2011.562327 (online Apr 2011)
15. Unrine, J.M., Tsyusko, O.V., Hunyadi, S.E., Judy, J.D., Bertsch, P.M.: Effects of particle size on chemical speciation and bioavailability of copper to earthworms (*Eisenia fetida*) exposed to copper nanoparticles. J. Environ. Qual. **39**, 1942–1953 (2010)
16. Coleman, J.G., Johnson, D.R., Stanley, J.K., et al.: Assessing the fate and effects of nano aluminum oxide in the terrestrial earthworm, *Eisenia fetida*. Environ. Toxicol. Chem. **29**, 1575–1580 (2010)
17. Hu, C.W., Li, M., Cui, Y.B., Li, D.S., Chen, J., Yang, L.Y.: Toxicological effects of TiO_2 and ZnO nanoparticles in soil on earthworm *Eisenia fetida*. Soil Biol. Biochem. **42**, 586–591 (2010)
18. Lapied, E., Nahmani, J.Y., Moudilou, E., et al.: Ecotoxicological effects of an aged TiO_2 nanocomposite measured as apoptosis in the anecic earthworm *Lumbricus terrestris* after exposure through water, food and soil. Environ. Int. **37**, 1105–1110 (2011)
19. Hooper, H.L., Jurkschat, K., Morgan, A.J., et al.: Comparative chronic toxicity of nanoparticulate and ionic zinc to the earthworm *Eisenia veneta* in a soil matrix. Environ. Int. **37**, 1111–1117 (2011)
20. Li, L.-Z., Zhou, D.-M., Peijnenburg, W.J.G.M., et al.: Toxicity of zinc oxide nanoparticles in the earthworm *Eisenia fetida* and subcellular fractionation of Zn. Environ. Int. **37**, 1098–1104 (2011)

Extreme Ultraviolet Lithography (EUVL)

► EUV Lithography

F

Fabrication of Graphene or Graphene Fabrication

► Synthesis of Graphene

Fate of Manufactured Nanoparticles in Aqueous Environment

Jérôme Labille[1], Natalia Pelinovskaya[1], Céline Botta[1], Jean-Yves Bottero[1] and Armand Masion[1,2]
[1]Aix-Marseille Université, CNRS, IRD, CEREGE UM34, Aix–en–Provence, France
[2]iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France

Synonyms

Aging; Alteration; Dispersion; Environmental impact; Life cycle; Titanium dioxide; Transfer

Definition

Manufactured nanoparticle. Manufactured nanoparticles are particles that have at least one dimension between 1 and 100 nm and are specifically engineered to have unique properties that do not exist in bulk materials with an identical composition.

Nanomaterial. A nanomaterial is a material that contains structured components with at least one dimension below 100 nm. These "nano-components" can be a solid (nanoparticle), liquid (nano-droplet), or gas (nanoporosity) phase, distributed in the material. A nanomaterial has no maximum size limit.

Nanoproduct. A nanoproduct is a commercial product that contains nanomaterial.

Colloid. A colloid is a material distributed in a dispersing medium, which is more submitted to Brownian random motion than to gravity attraction. Colloids typically have at least one dimension below 1 μm. They look homogeneous when observed by the naked eye.

Introduction

High-performance nanomaterials incorporating nanoparticles are making a massive intrusion into many industrial sectors from electronics to automotive, consumer goods, textiles, and pharmaceutics. More than 4,100 companies from 48 countries are involved in nanotechnology research, manufacturing, or applications [1]. In most of these applications, nanoparticles are surface coated with (in)organic compounds or doped with specific ions to obtain nanomaterials with specific characteristics. To date, more than 1,800 nanotechnology-based consumer products

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

introduced to the market have been identified and listed in the non-exhaustive database proposed by the Woodrow Wilson Institute [2]. Current trends all indicate that human exposure to nanomaterials will continue to increase in the coming years as more products are developed and put on the market. This has resulted in elevated anxiety concerning the potential risks that are associated with nanotechnology [3–5]. Accidents, normal aging during regular use, or inappropriate storage of the related wastes are possible ways for the dissemination of nanomaterials, or their related by-products, into the environment.

Titanium dioxide (TiO_2) nanoparticles represent the highest mass production of nanoparticles in the world. Its commercial production for the years 2006–2010 has been estimated at 5,000 t/year and more than 10,000 t/year for the years 2011–2014 [6]. Robichaud et al. [7] estimated its worldwide production level at approximately 2.5 million metric tons by 2025. Nano-TiO_2 are indeed widely used in industrial applications taking advantage of their UV blocking (e.g., sunscreen formulation) or photocatalytic properties (e.g., self-cleaning surfaces). A major part of the production is used for "personal care" products (sunscreens, cosmetics), while the rest is used in industrial applications such as paints, plastics, catalysts, and ceramics. Such a large mass logically raises serious concerns about the fate and impact of these related nanoproducts when released into the environment [8–10].

For instance, there are evidences for the release of TiO_2 nanoparticles (up to 300 nm) used in self-cleaning exterior facade into runoff water [11]. Further in the water cycle, in wastewater treatment plants, the presence of submicronic Ti-based particles has been measured in raw sewage (100–3,000 μg/L) and even in the effluents after treatment (5–15 μg/L) [12]. Cosmetics, and especially sunscreens, are the commercialized nanotechnology-based products that underwent the largest growth from 2007 to 2010 in Europe [13]. In sunscreens, TiO_2 is the most used inorganic filter, because of its effective UV reflection and absorption capacities and its adequate tolerance by human skin. Moreover, for transparency and efficacy reasons, nano-sized TiO_2 particles are widely used. To inhibit their photoactivity, TiO_2 nanoparticles used in sunscreens are surface modified using aluminum oxide/hydroxide and/or silicon dioxide. Moreover, lipophilic coatings are also grafted at the surface to facilitate the dispersion of the nanoparticles in the oil phase of the cream. The resulting material is commonly named TiO_2-based nanomaterial.

These functionalizations typically used on nano-TiO_2 within sunscreen will undoubtedly alter not only the behavior but also the toxic impact of the nano-objects upon their release into the environment. The hydrophobic character of the outermost coating indeed prevents the dispersion in water, while the mineral surface layer prevents phototoxicity. However, one can wonder about the lifetime of these coatings when the sunscreen undergoes regular weathering expected with its use [14].

Consequently, it is necessary for rigorous human and environmental risk assessment to consider exposure scenarios, not only related to bare nanoparticles but also to commercialized nanomaterials throughout their life cycle [15], taking into account the different stages of their surface alteration.

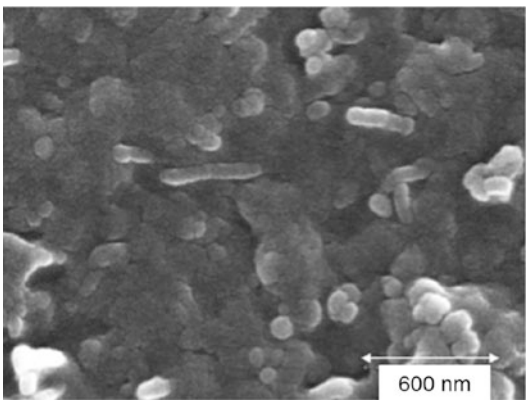
In this chapter, which focuses on the case of TiO_2 nanoparticles used in sunscreen as UV blockers, three related samples containing TiO_2 nanoparticles are compared in terms of behavior and fate in the aqueous environment, pure TiO_2 nanoparticles, TiO_2 nanomaterials used in sunscreen, and an actual sunscreen nanoproduct, so that gradual complexity in the chemical composition is explored from the bare mineral nanoparticles to the final commercialized cream. The effects of the corresponding surface properties can thus be evidenced as a function of their respective time evolution with aging.

In addition to these parameters intrinsic to the nano-object's composition, different environmental factors also play a role on the fate of nanoparticles in aquatic systems. pH, salinity, naturally occurring particulate or dissolved materials from both organic or mineral origin, and porosity of the system are as many factors that significantly control the physicochemical behavior of nanoparticles in water, in close relation with

their surface properties [16, 17]. The main physicochemical reactions involved in these phenomena are presented here.

Nano-TiO_2, from Bare Nanoparticles to Sunscreen Formulation

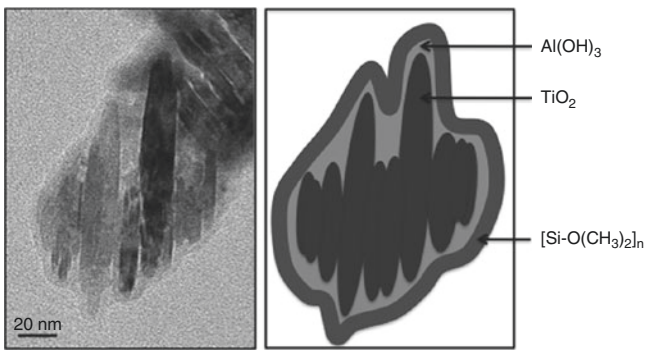
Pure TiO_2 nanoparticles used as bare nanoparticles were purchased by Alfa-Aesar Company under the trade name Titanium IV oxide, anatase, 32 nm, 39953. They consist of a powder of pure TiO_2 in the anatase form. As observed by scanning electron microscopy (Fig. 1), they are constituted of polydisperse agglomerates with sizes ranging from 100 to 600 nm. The fate of these pure TiO_2 nanoparticles in aqueous environment, and more particularly in a sandy porous medium saturated with water, was studied by Solovitch et al. [18].



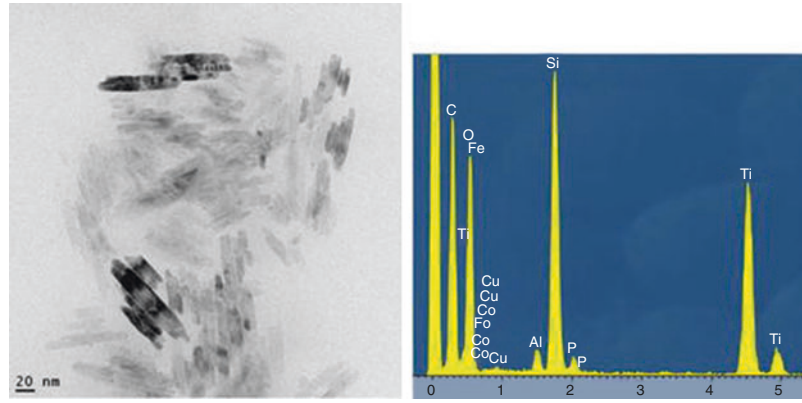
Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 1 Scanning electron microscope picture of the bare TiO_2 nanoparticles

The TiO_2 nanomaterial T-Lite™ SF (BASF, Germany) is widely used in sunscreens as mineral UV blocker. Its nano-TiO_2 core is functionalized in the aim to be dispersed in the oil phase of the sunscreen formulation. Its coating consists in a first layer of photocatalysis scavenger aluminum hydroxide [$Al(OH)_3$] and of an outer layer, the hydrophobic dispersing agent polydimethylsiloxane (PDMS) consisting in $[Si\text{-}O(CH_3)_2]_n$ repeat units (Fig. 2). The behavior and aging of the T-Lite nanomaterial in aqueous system was studied by Labille et al. [19] and Auffan et al. [20].

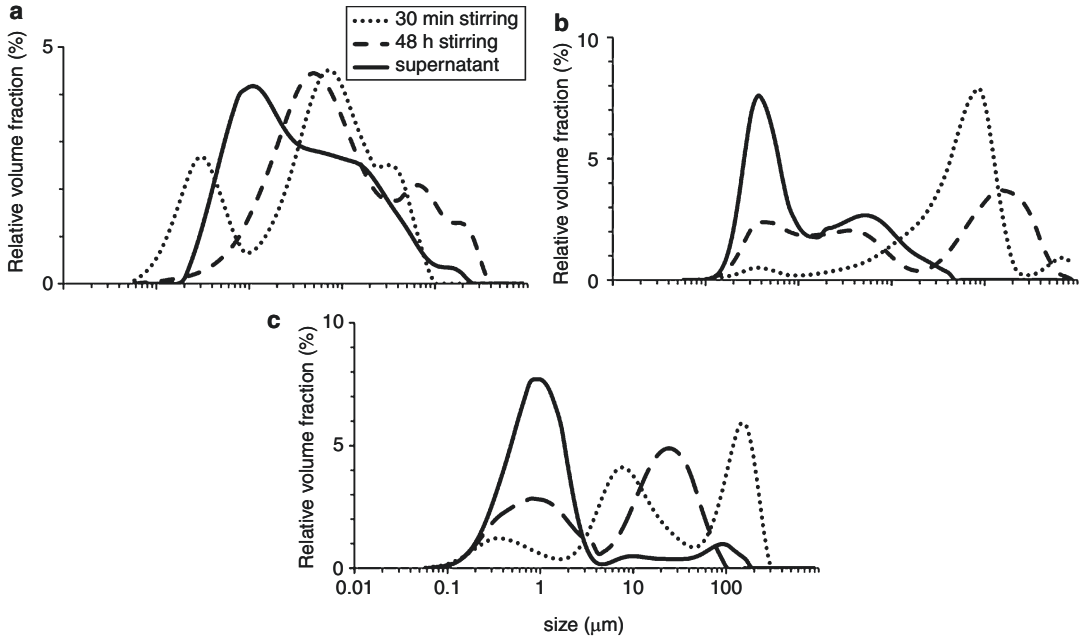
A commercial sunscreen corresponding to high-volume product was selected for comparison with bare TiO_2 nanoparticles and coated TiO_2 nanomaterials in terms of behavior in aqueous systems. This product displays a very high Sun Protection Factor (SPF $\geq$ 50), which corresponds to one in six of the sunscreen market. Its component list includes titanium dioxide and aluminum oxide. No information was given by the manufacturer on the crystal form, size, or coating of the TiO_2 particles used in this sunscreen formulation. However, the presence of Ti, Al, Si, and C elements concentrated in or on the nanoparticles (Fig. 3) suggests a typical coating involving aluminum oxide and PDMS similar to the TiO_2 nanomaterial T-Lite SF considered in this entry. The aging and behavior in water of this sunscreen product and of three other similar formulations was studied by Botta et al. [21]. It revealed that the TiO_2 concentration in the studied sunscreen is about 4 w/w%, the rest of the formulation consisting mainly of organic components.



Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 2 Transmission electron microscope picture of the TiO_2 nanomaterial (*left*) and scheme of the successive coatings (*right*)



Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 3 Shape (*left*, TEM observation) and chemical composition (*right*, EDX spectrum expressed in keV) of the nanoparticles identified in the investigated sunscreen



Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 4 Dispersion kinetics of the nanoparticulate samples in water. Size distributions after 30-min or 48-h stirring and after 48-h stirring +48-h settling; bare nano-TiO_2 (**a**), nanomaterial (**b**), and sunscreen (**c**)

Dispersion in Pure Water

The effect of the nanoparticle coating or formulation on their potential dispersion in aqueous system is presented below in terms of dispersion kinetics measured for the three different materials considered. Laser diffraction was used to measure the size distribution of the dispersed particles versus time. On Fig. 4 are reported the size distributions measured on the three samples after 30 min and 48-h stirring in pure water. The three sample dispersions were carried under the same hydrodynamic conditions with a magnetic stirrer rotating at 690 rpm in a 1 L vessel. The mixtures consisted of 100 mg powder (bare TiO_2 or TiO_2 nanomaterial) or 1.5 g of cream stirred in 250 mL of Milli-Q water. After 48-h stirring, in order to better distinguish the potential presence of

nanometric or micrometric particles remaining stable in suspension, the samples were let to settle out during 48 h, and the size distribution was measured in the supernatants.

Pure TiO_2

The time-resolved size distribution of the pure TiO_2 nanoparticles is shown on Fig. 4a. Considering the 32 nm crystallite size announced by the manufacturer, an unexpected larger and polydisperse size distribution is measured. At 30 min stirring, the smallest particle size class appears above 60 nm and centered around 250 nm, but large aggregates up to 100 μm in size are also measured. With stirring time, these smallest units tend to disappear, while larger aggregates tend to grow, reaching a maximum size of 300 μm at 48-h stirring. After 48-h settling, even if the size distribution measured in the supernatant tends to overexpress the finest particles while the larger ones have been settled and removed, the disappearance of the nanometric units from the dispersion is confirmed to the benefit of micrometric aggregates centered around 1–10 μm. This apparent aggregation with stirring time was certainly due to the pH of the mixture which was let natural at 5.5. This pH corresponds exactly to the isoelectric point of the TiO_2 nanoparticles, which favors self-aggregation (see pH effect section). However, even at more acidic pH, the crystallite size of 32 nm could not be measured as distinct units dispersed in water. Solovitch et al. [18], who submitted the TiO_2 nanoparticle to an acidic treatment to protonate their surface and favor dispersion, measured by dynamic light scattering a size distribution centered around 150 nm.

This tendency for nanometric crystallites to associate together and build larger agglomerates can be explained by their exacerbated surface energy. According to the Young-Laplace's equation, the surface tension of a particle is inversely proportional to its radius of curvature, i.e., its size. With very small nanoparticles, this leads to a surface pressure so high that they are thermodynamically unstable. In order to minimize such a pressure jump across the solid surface, nanoparticles may aggregate together into larger and dense clusters [22].

TiO_2 Nanomaterial

As anticipated for the water-repellent PDMS coating, the initial behavior of the nanomaterial stirred in water is typical for a hydrophobic material in aqueous media, so that the powder is initially retained at the air-water interface. Nevertheless, under stirring, the behavior of the nanomaterial evolves rapidly like a more hydrophilic material. As soon as 30 min of stirring, the solid phase starts to disperse in the aqueous phase (Fig. 4b). Some large agglomerates of ca. 200 μm are measured; they are probably formed by hydrophobic interactions arranging agglomerates in lumps to minimize the solid/liquid surface contact. However, a second and minor class of sizes also arises around 300 nm, corresponding to colloidal residues. Its relative contribution tends to increase rapidly with time and reaches values up to 35 vol% after 48-h stirring. This submicron size allows the corresponding colloidal residues to remain stable in suspension when stirring is stopped, as confirmed by their major contribution in the size distributions measured after settling period.

Such rapid dispersion in the aqueous phase of an originally hydrophobic material was explained by Auffan et al. and Labille et al. [19, 20] as being due to the desorption and alteration of the hydrophobic PDMS coating from the nanomaterial surface. Indeed, this was studied by measuring the release of Si from the nanomaterial to the solution during agitation in water by inductively-coupled plasma atomic emission spectroscopy (ICP-AES) [16] and by following the chemical structure of the PDMS by ^{29}Si CP-MAS NMR [20]. It revealed the desorption and alteration of most of the hydrophobic coating.

Concerning the $Al(OH)_3$ coating present in the initial nanomaterial, these works showed that it did not undergo significant alteration or dissolution in the alteration conditions used, thus remaining as a protective layer that prevents photocatalytic activity (see also section "Evaluation of the Nanoparticle Photoactivity").

F

Sunscreen Nanoproduct

Figure 4c shows the size distributions measured at the same stirring times for the sunscreen. Similar to the hydrophobic T-Lite nanomaterial, no size distribution could be measured at $t = 0$ due to the hydrophobic character which prevents immediate dispersion in water. Nevertheless, a relatively rapid dispersion in water was observed, with release of submicronic (<1 μm) particles as soon as 30 min of aging. Moreover, this seems to be a gradual process since the size distribution at 48-h stirring displays a higher amount of submicronic sizes.

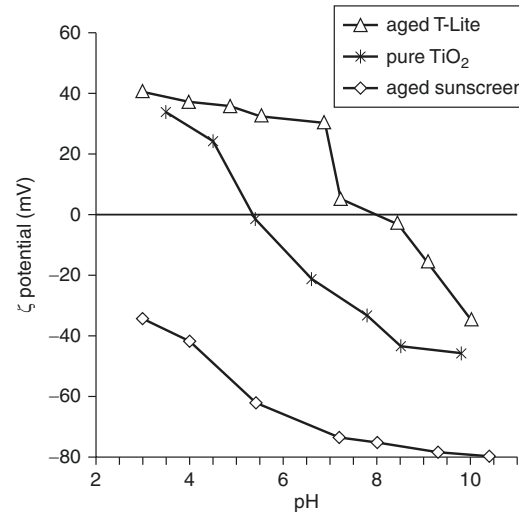
This tendency evidences the alteration of the initially hydrophobic material into residues displaying new hydrophilic surfaces, which enable their dispersion as submicronic entities. The size distributions of the residues constituting the stable suspension remaining after 48-h settling reveal a major fraction of this submicronic class. However, surprisingly, large agglomerates of several tens of micrometers are also measured in addition to the submicronic class. This larger class certainly corresponds to the agglomeration of organic matter into open structures with low density [21].

From the concentrations of the colloidal residues recovered after 48 h of stirring followed by 48 h of settling, a mass balance between the stable colloidal phase and the settled coarse phase can be estimated for the three nanoparticulate samples.

For the pure TiO_2 nanoparticles and TiO_2 nanomaterial, the ratio "dispersed/settled out" is about 50/50 and 25/75, respectively [19]. For the sunscreen, according to Botta et al. [21], the mass of the colloidal fraction dispersed in water represents between 20 % and 40 % of the initial sunscreen, while these colloids involve less than 30 w/w% of the total TiO_2 initially present.

Effect of pH on Nanoparticle Dispersion Stability

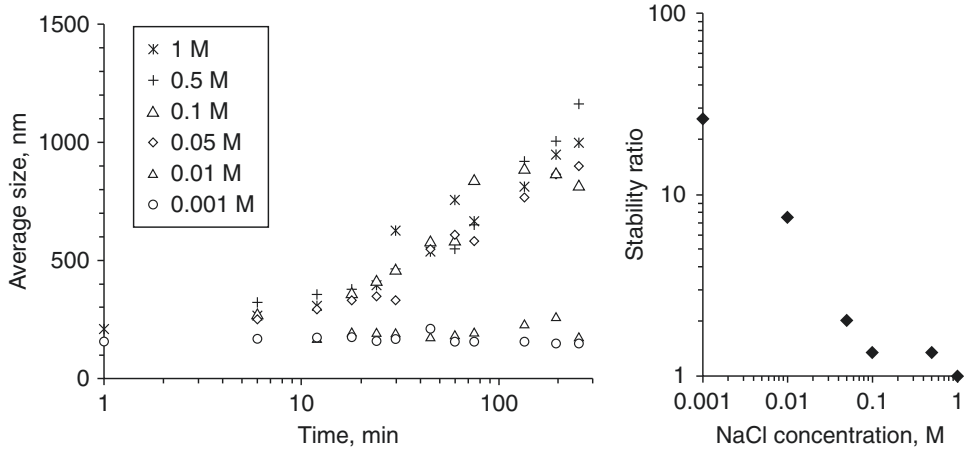
pH has a drastic effect on the dispersion stability. Adjusting the solution pH close to the isoelectric point of the dispersed particles will induce neutralization of the charged groups at their surface and thus the vanishing of the repulsive electrostatic interactions, at the benefit of the short-range and attractive van der Waals forces. This results in self-aggregation and sedimentation of the involved particles.



Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 5 Zeta potential evolution of nanoparticles versus pH

In order to better predict the dispersion stability of the nanoparticles as a function of pH, their electric surface potential can be estimated by electrophoretic mobility measurements. The values for the three samples of stable suspensions recovered after 48-h stirring +48-h settling were measured as a function of pH varying from 2 to 10. The zeta potential (ζ, mV) plots obtained are displayed in Fig. 5.

The titration curve obtained for the pure TiO_2 nanoparticles is typical for a metal oxide surface undergoing hydrolysis. The surface hydrolyzes in the presence of water to form hydroxide layers at the surface (≡M–OH). These polar hydroxyl groups confer to the oxide a hydrophilic and amphoteric character. They may become charged by reacting with H^+ or OH^- ions from the solution. At low pH, hydroxide surfaces adsorb protons and become positively charged (≡M–OH_2^+), whereas at high pH, proton desorption induces a negative surface charge (≡M–O^-). The overall surface charge of the oxide then results from the



Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 6 Aggregation kinetics of TiO_2 nanoparticles versus NaCl concentration (*left*); stability ratio of the TiO_2 suspension versus NaCl concentration (*right*)

respective expressions of all the hydroxyl sites according to the pH. The isoelectric point (IEP) of a surface is the pH of zero zeta potential. For TiO_2 surface, it is typically obtained in the range 4.8–6.9 depending mainly on synthesis specificities and relative development of the different lattice faces [17]. The TiO_2 nanoparticles considered in this entry have an isoelectric point at pH = 5.5 [18].

The zeta potential titration for the T-Lite nanomaterial revealed an isoelectric point for these colloidal residues at pH = 7.3. This IEP is characteristic for an aluminum hydroxide surface in agreement with the removal or alteration of the PDMS hydrophobic coating from the nanomaterial surface after 48-h stirring [19].

The stable suspension recovered from the commercial sunscreen displays a zeta potential remaining negative in all the pH range from 2 to 10. This does not fit with the surface charge expected from TiO_2 nor $Al(OH)_3$ used in the nanomaterials and identified respectively on the two other plots of Fig. 5. Nevertheless, Botta et al. [21] evidenced that some TiO_2 nanomaterials are well involved in the colloidal residues released from the sunscreen. In the panel of four sunscreens that they studied, TiO_2 constituted 0.6–9 wt% of the dry colloid mass. This is a relatively low proportion, which remains however in the same order as that in the initial sunscreen formulation (3–5 wt%). Thus, the organic components from the sunscreen formulation control the properties and behavior of the colloidal residues. Their negative surface charge is thus certainly due to some anionic organics such as disodium EDTA salt or oxidized siloxane polymers.

Effect of Salinity on Nanoparticles' Dispersion

Increasing salt concentration induces progressive screening of the repulsive electrostatic interactions between particles through compression of the electric double layer that surrounds each charged particle. The critical salt concentration above which attractive forces become predominant, termed the critical coagulation concentration (CCC), determines the limit for colloidal stability [17].

The aggregation kinetics of pure TiO_2 nanoparticles as a function of NaCl concentration are presented in Fig. 6 [18]. The stability ratio was calculated from the slope of the measured aggregation kinetics, as the ratio of the maximum aggregation rate obtained at the highest salt concentration to the aggregation rate measured at each salt concentration. The nanoparticles remained well dispersed at NaCl concentration below or equal to 10^{-3} mol/L, showing almost

Fate of Manufactured Nanoparticles in Aqueous Environment, Table 1 Critical coagulation concentrations measured with NaCl for the three nanoparticulate suspensions tested

	CCC_{NaCl}
Pure TiO_2	10^{-1} M
T-Lite	$2\ 10^{-2}$ M
Sunscreen	10^{-3} to 10^{-1} M

constant size over a period of some 250 min. A slow aggregation regime occurred at $10^{-2} < [NaCl] < 10^{-1}$ mol/L (stability ratio > 1). The critical coagulation concentration is defined as that enabling fast aggregation regime, i.e., a stability ratio ≈ 1. In the pure TiO_2 sample, CCC_{NaCl} falls around $[NaCl] = 10^{-1}$ mol/L.

Another approach to determine the CCC for a colloidal dispersion consists in measuring the turbidity in the supernatant after a settling period. The CCC corresponds to the minimum salt concentration at which the turbidity drops. On one hand, this approach is less time consuming, requiring only one punctual measurement on a series of samples. However, on the other hand, measuring only one point for each salt concentration deprives of any kinetics information and implies a result depending on the measurement time. In other words, lower CCC value could be typically measured at longer settling time, due to the slow coagulation process.

This second approach was followed with stable aqueous suspensions obtained from TiO_2 nanomaterial and sunscreen nanoproduct [19, 21]. The respective CCC_{NaCl} values obtained in these works are reported in Table 1.

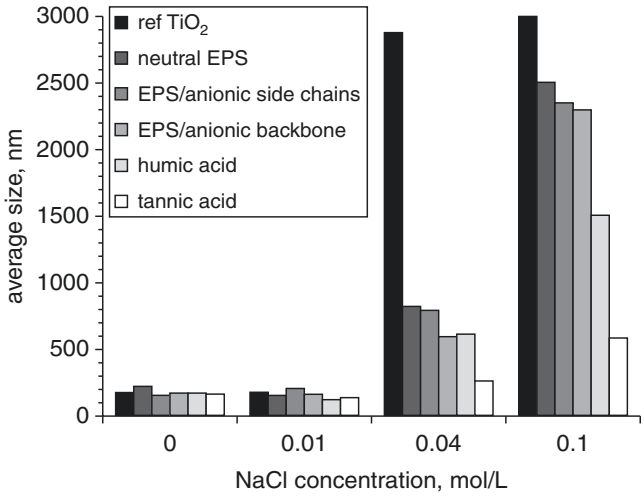
The typical salinity for seawater is around 5×10^{-1} mol/L. With the three nanoparticles' dispersions studied here having lower CCC_{NaCl} values, one could expect that they shall not remain stable in marine environment but rather undergo aggregation by coagulation mechanism. However, such simple consideration is mainly valid in synthetic system with very simple solution chemistry. In natural surface water, the presence of both suspended mineral particulate matter and natural organic matter must significantly alter the balance between dispersion and aggregation.

Effect of Natural Organic Matter on Nanoparticle Dispersion

In the two following examples of pure nano-TiO_2 and TiO_2 nanomaterial, different kinds of natural organic matter (NOM) were introduced into the nanoparticulate suspensions to evidence their effect on the nanoparticle dispersion stability. Three polysaccharides from rhizospheric bacterial origin (exopolymeric substances, EPS) and having high molecular weights (MW ≈ 2×10^6 Da): one neutral EPS and two branched anionic EPS with acidic residues located either in the backbone or in the side chains, a humic acid, and a tannic acid were tested. In all the presented experiment, the NOM concentration used were very diluted between 0.5 and 5 mg/L, so that no viscosity effect related to the high molecular weight was enabled.

Pure Nano-TiO_2

In Fig. 7 are reported the average sizes of the TiO_2 nanoparticles as a function of NaCl concentration and of the natural organic matter tested. In the absence of natural organic matter, the average size in the reference sample increases from 150 to 3,000 nm when NaCl concentration increases above 10^{-2} mol/L. This corresponds to the expected salt-induced aggregation observed in Fig. 6. The addition of natural organic matter contributed to favor nanoparticle dispersion despite the increasing salt concentration. This is typically due to the adsorption of the organic macromolecules onto the nanoparticle surface, inducing new repulsive interactions of steric and/or electrostatic nature. However, natural organic matters having different structural properties did not behave equally as a dispersing agent, depending strongly on the salt concentration. At $NaCl = 4 \times 10^{-2}$ mol/L, corresponding to the beginning of the fast aggregation regime (Fig. 6), the weak attraction forces occurring between the bare nanoparticles can be easily canceled by NOM addition, so that the average size obtained ranges six- to tenfold lower than the reference sample (250–800 nm). But at the highest salt



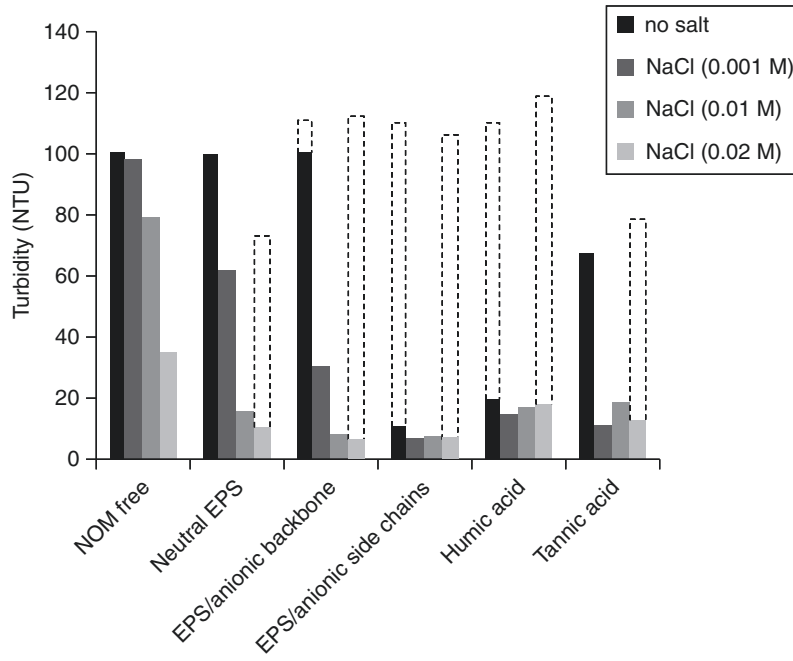
Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 7 Average size of TiO_2 nanoparticles as a function of NaCl concentration in the presence of different natural organic matters. pH = 8, TiO_2 = 50 mg/L, NOM = 1 mg/L

concentration 10^{-1} mol/L, the rapid aggregation measured with nanoparticles free of salt indicates the occurrence of stronger interparticle attraction forces. Such interactions could not be counterbalanced by all the NOM tested. EPS appeared as relatively inefficient dispersants. Humic acid enabled to decrease twofold the average aggregate size. Tannic acid appeared as the most efficient dispersant, limiting the average size at 500 nm despite the high salt concentration. The high charge density of humic and tannic acid molecules compared to EPS certainly favors nanoparticles' dispersion by creating new electrostatic repulsions between neighbor particles. However, when comparing neutral and anionic EPS, it appears that similar and weak dispersing effects were obtained with these three NOMs. The macromolecule anionicity is thus certainly not the driving parameter for dispersion. The molecular weight of the adsorbed macromolecules more likely appears determining in this example. Polysaccharides indeed display larger molecular weights, while tannic acid displays the lowest. With the experiment being conducted at a given NOM concentration, a lower molecular weight implies a higher number concentration, which favors surface saturation of the coated nanoparticles and thus their electrostatic stabilization. Zeta potential measurement of the nanoparticles coated with tannic acid molecules indeed reveals a significant negative surface charge thus supporting this mechanism.

TiO_2 Nanomaterial

The effect of NOM on the dispersion stability of the nanomaterial aqueous suspension is compared in Fig. 8 for two different concentration ratios: 2 and 20 wt% with respect to the mass of nanoparticles. Two opposite effects well known respectively as bridging flocculation and steric stabilization can be evidenced.

At 2 wt% of NOM, contrary to what was obtained with pure TiO_2 nanoparticles, the global effect observed was a destabilization of the dispersion. The supernatant turbidity was indeed significantly decreased in the presence of NOM, in varying extent depending on the type of NOM tested. This destabilizing mechanism is presumably bridging flocculation, in which organic macromolecules adsorb to more than one nanoparticle, creating bridges between neighboring units. The different NOM macromolecules evaluated were not all equal in terms of bridging ability. The EPS with anionic side chains and humic acid both demonstrated the highest ability to flocculate the nanoparticles, enabling an 80–90 % decrease in the turbidity in the absence of any salt. The other NOM sources required that salt be added to the solution to induce significant destabilization of the suspension. Salt addition indeed allows neighboring particles to come closer to each other, which may be a prerequisite condition for NOM-induced bridging



Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 8 Supernatant's turbidity of the T-Lite suspension after 48-h settling as a function of NaCl concentration and natural organic matter; NP = 25 mg/L; NOM/NP = 2 wt% (*solid histograms*) or 20 wt% (*dashed histograms*)

flocculation, depending on the macromolecule bridging efficiency. The less efficient macromolecule, the more salt addition required. Tannic acid and EPS with anionic backbone requiring salt concentration lower than the CCC_{NaCl} thus appear more efficient flocculent than neutral EPS which requires salt addition higher than the CCC_{NaCl}. The potential effects of molecular weight and spatial conformation cannot explain these discrepancies, since EPS with anionic side chains and humic acid are totally different in all these aspects, whereas both anionic EPS are more similar from this point of view. Nevertheless, the effects of the NOM anionicity and of the localization of this charge in the macromolecule likely appear as determining parameters. Indeed, the humic acid molecule is the most negatively charged, and the EPS with anionic side chains are likely to favor rapid attractive electrostatic interactions with the positively charged particles, which did not occur for the other types of NOM that were evaluated here. Indeed, the least-efficient flocculent has a neutral charge. The weak flocculent, EPS with anionic backbone, displayed acidic residues less exposed to the surrounding media. Finally, the tannic acid was a relatively good flocculent as soon as a minor salt addition was achieved. This condition is certainly due to its lower molecular weight compared to the humic acid, implying shorter interparticle bridges, which require prior approach of the neighboring particles via salt-induced partial screening of repulsive electrostatic interactions.

Conversely, at 20 wt% of NOM, further stabilization of the colloidal dispersion is observed, canceling salt-induced coagulation (Fig. 8). Similar to what was observed with pure TiO_2 nanoparticles, the mechanism involved here is again certainly steric or electrosteric stabilization. It is typically observed at high macromolecule concentration, when sorbed macromolecules saturate the surface of the particles, creating new repulsive interactions and thus dispersion stabilization.

Bridging flocculation mechanism is thus observed in the presence of a low amount of macromolecules sorbed to the colloid surface with regard to the total adsorption capacity of the colloid, while steric or electrostatic stabilization rather occurs at higher NOM concentration.

This has important consequences on the global fate of the nanoparticles in aqueous environments. Typically, in surface water where NOM concentrations range from 5 to 10 mg/L, the formation of stable nanoparticle dispersions may be favored. Conversely, in groundwaters where NOM concentration ranges from 0.1 to 2 mg/L, the flocculation of nanoparticles may be more likely. This effect has been observed elsewhere at the same weight proportions of NOM on clay colloids [23].

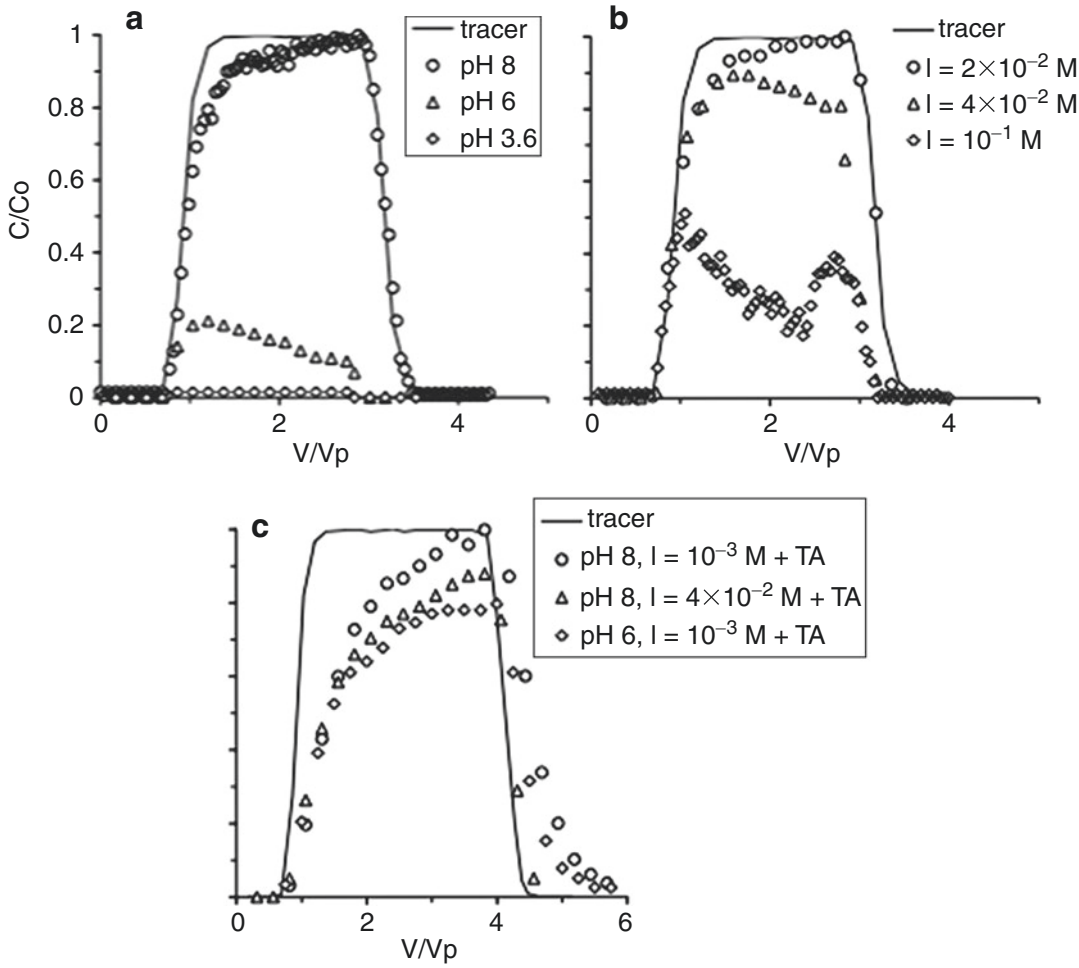
Since the ratio of nanoparticle to natural suspended matter plays an important role on the fate of both components, the nanoparticle concentration must be considered with caution in experimental fate studies. Predicted environmental concentration for Nano-TiO_2 in surface water ranges from 0.01 to 10 μg/L [24]. However, due to analytical detection limit, the aggregation kinetics presented here above and in most of the literature used nanoparticle concentrations higher than 10 mg/L. Such overestimation of the nanoparticle concentration completely alters the dispersion dynamic that may occur in real and complex environmental systems. In natural surface water, the probability that manufactured nanoparticles interact with each other is very low with regard to the likelihood of colliding with naturally occurring suspended matter present in significantly higher concentrations. It is thus reasonable to expect that the natural suspended matter will interact with the manufactured nanoparticles and potentially act as a carrier in the water column. This is the so-called heteroaggregation mechanism, which strongly affects the fate and transport potential of the manufactured nanoparticles. In the frame of ERA-NET SIINN Nanoheter Research Program [25], Labille and coworkers have studied this effect under realistic conditions, following both naturalist and mechanistic approaches. They showed that heteroaggregation of nano-TiO_2 indeed takes place in surface water with naturally occurring particulate or dissolved matter, depending on the sticking efficiency between the interacting components, i.e., on the chemical compositions of both these components and the dispersing solution [26].

Mobility Within Porous Media

The previous considerations of the main biophysicochemical reactions that may control the fate of particulate matter through the aqueous environment are most valid within the water column, which constitutes a relatively homogeneous dispersing phase. However, particulate matter undergoing, for instance, aggregation and sedimentation phenomena is expected to exit such monophasic and continuous system and to incorporate the more complex sediment medium. This latter consists of a solid, disordered, and porous phase saturated with water. The mobility or deposition of manufactured nanoparticles in this system is a real concern for bioaccessibility to benthic organisms. Similarly, bioaccessibility to terrestrial biota will follow the same issues as soil can be considered as another major porous receptacle for nanoparticles. *In fine*, via trophic chain or water transport through aquifers, this is exposure of humans to manufactured nanoparticles that may depend on their fate within porous media.

The diffusion column experiment is conventionally used to study the parameters determining the mobility or attachment of nanoparticles in a saturated porous medium. A mobile electrolyte circulates at a constant rate through a column filled with a stationary porous medium. At time 0, a volume of nanoparticles in suspension is injected into the circuit upstream of the column. The nanoparticle's concentration is measured in the mobile phase at the column outlet as a function of time and flux. The resulting evolution is called a breakthrough curve, typically box shaped. Many parameters can vary in the mobile phase (pH, ionic strength, hydrodynamic flow, water saturation) and in the stationary phase (porosity, mineralogy, grain size), and their effects on nanoparticle's attachment can be investigated.

The example presented in Fig. 9 deals with the pure TiO_2 nanoparticles injected within a sandy porous medium [18]. The breakthrough curves relative to different studied parameters (ionic strength, pH, presence of natural organic matter) are shown.



Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 9 Breakthrough curves of the TiO_2 nanoparticles within sand porous medium as a function of pH at I = 10^{-3} M (**a**), of salt concentration at pH = 8 (**b**), and in additional presence of tannic acid (**c**)

The influence of pH on the nanoparticle's retention in the column was studied at pH 3.6, 6, and 8 (Fig. 9a). The retention of nanoparticles in the column is nearly complete at pH = 3.6. This observation is consistent with the opposite surface charges of sand grains (−) and nanoparticles (+) at this pH, favoring electrostatic attractions between each surface. To the contrary, at pH = 8, particles and sand surfaces are both negatively charged, decreasing deposition down to 5 %. Finally, pH = 6 is very close to the pH of zero electric potential for the TiO_2 surface (IEP = 5.5; Fig. 5), which strongly favors self-aggregation of the nanoparticles, and alters their behavior in the porous medium. A high retention of TiO_2 is measured (78 %), certainly due to this aggregation effect. This is further confirmed when looking at the effect of salt concentration in the solution (Fig. 9b), since this parameter strongly controls the dispersion versus aggregation of the nanoparticle's suspension.

The ionic strength is expected to control the behavior of nanoparticles in the porous medium through electrostatic forces between sand grains and nanoparticles on one hand, but also between neighboring nanoparticles together on another hand. Both of these interactions are repulsive in nature at low ionic strength and may be screened

at high ionic strength, at the benefit of short-range attractive forces. A critical salt concentration can be defined beyond which deposition increases. This concentration is around 4×10^{-2} mol/L according to Fig. 9b. It is interesting to notice that this value coincides well with the critical coagulation concentration measured in a homogeneous and liquid medium (Fig. 6). The range 4×10^{-2} to 10^{-1} mol/L falls indeed exactly within the beginning of the fast aggregation regime. This strongly suggests a link between the aggregation of nanoparticles and their apparent deposition. Two reasonable mechanisms shall imply such a relation: (1) steric effects cause larger aggregates to deposit more likely than dispersed and smaller nanoparticles; and (2) drag forces through porous aggregates favor their deposition onto a collector, while individual nanoparticles rather follow streamline flow toward the column outlet [18].

Moreover, when the nanoparticle injection is stopped, their disappearance from the outlet solution is also much faster in conditions favoring their aggregation (pH 6 or NaCl $> 4 \times 10^{-2}$ M) than with well-dispersed nanoparticles which follow mainly the conservative tracer profile. This is due to size exclusion effect, which makes aggregates to visit a smaller portion of the available pore network within the column. Therefore, those aggregates that do not attach to the sand will pass the column more rapidly [18].

Furthermore, through its affinity for mobile or stationary inorganic surfaces, the presence of organic matter in the porous medium is likely to affect nanoparticle mobility. In a similar way to what is observed in dilute media, when organic molecules coat the nanoparticle or sand grain surface, the balance of the interactions between them will be modified by the new properties of each surface. Tannic acid was identified previously, among those tested (Fig. 7), as the natural organic matter displaying the highest dispersing effect on the TiO_2 nanoparticles even at high salt concentration. Adsorption isotherms of the tannic acid to sand or to TiO_2 nanoparticles revealed similar affinities for both surfaces.

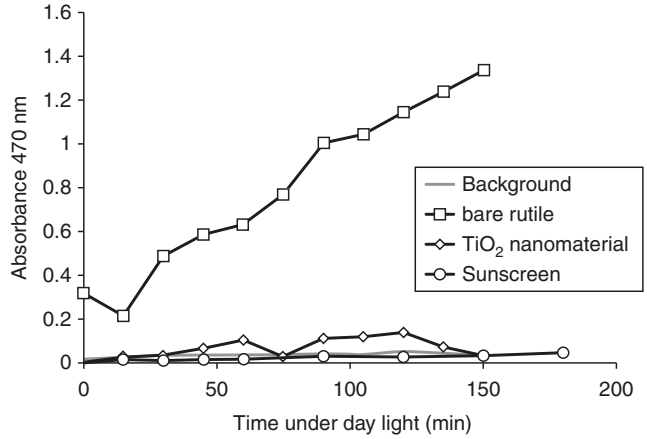
Figure 9c shows how tannic acid introduced into the nanoparticle suspension just before injection into the column strongly alters their breakthrough curve. At pH = 8, I = 10^{-3} M, where negligible deposition occurred in absence of NOM, due to both negatively charged surfaces, longer delay in the breakthrough curve and higher deposition rate are observed in the presence of the organic molecule. This is probably due to the affinity of the NOM molecule for both the sand and the nanoparticle surfaces, which may create bridges between both.

At pH = 8, I = 4×10^{-2} M, nanoparticles free of NOM undergo aggregation and display a typical breakthrough curve as previously described. In the presence of tannic acid, the coated nanoparticles remain well dispersed, and the breakthrough curve rather follows a delayed breakthrough similar to that obtained at I = 10^{-3} M, but with a slightly higher deposition.

At pH = 6, I = 10^{-3} M, while TiO_2 nanoparticles free of NOM are weakly surface charged and undergo significant aggregation and deposition, nanoparticles coated with tannic acid acquire a negative surface charge that favors dispersion and facilitates transfer through the porous medium.

Evaluation of the Nanoparticle Photoactivity

The potential photoactivity of the nanoparticles would imply under UV exposition the production of reactive oxygen species (ROS) that alter the neighboring organic matter. ROS are indeed responsible for oxidative stress which is a major cause of mortality for the living organisms exposed, due to Fenton reaction. The potential superoxide anion production by TiO_2 nanoparticles is presented here as a function of their coating or embedding. Bare TiO_2 is known to be highly photocatalytic and is commonly used as a positive reference. However, the colloidal residues released from the alteration of pure nanomaterial or of global sunscreen have been shown previously to remain respectively coated with an aluminum oxide layer or embedded in a large amount of organic components. The persistence of these surface layers certainly plays a significant role on the potential photoactivity of



Fate of Manufactured Nanoparticles in Aqueous Environment, Fig. 10 Assessment of ROS generation by the three nanoparticules' suspensions studied (pure nano-TiO_2, altered TiO_2 nanomaterial, and altered sunscreen nanoproduct). XTT (2,3-bis(2-methoxy-4-nitro- 5-sulfophenyl)-2H-tetrazolium-5-carboxanilide) was used to target and measure the superoxide anions. Measurements were performed at 470 nm

the colloids. The potential superoxide anion production by these samples was assessed under light excitation by the reduction of a tetrazolium dye, the XTT (2,3-bis(2-methoxy-4-nitro- 5-sulfophenyl)-2H-tetrazolium-5-carboxanilide). XTT is reduced by superoxide anion to produce water-soluble formazan molecule that can be photometrically monitored at the absorbance at 470 nm. In this experiment, the bare TiO_2 nanoparticles used as positive control were nano-rutile, in suspension at 50 mg/L. With the anatase TiO_2 form being naturally more photocatalytic than the rutile form used in sunscreens, the pure anatase TiO_2 nanoparticles studied in this entry would not have been a relevant reference. The absorbance data for the three samples are compared in Fig. 10. As expected, the rutile sample has an obvious photoactivity.

However, after the aging condition used in these works, the colloidal residues formed from nanomaterial or sunscreen did not generate superoxide. At the nanomaterial's surface, the Al-based layer remaining after alteration prevented the chemical interactions between the Ti atoms of the surface of the nano-TiO_2 core and the O_2 and/or H_2O molecules from the solution. This inhibited the promotion of electron-/hole + of the nano-TiO_2 core and the ROS generation. In the case of the sunscreen residues, not only the TiO_2 core of the nanomaterials used in the formulation was certainly still coated with a similar mineral ROS blocker, but the nanomaterials were also embedded within large aggregates of organic components that might not let the light pass. These two components both acted like barriers that prevented photo excitation of the TiO_2 core and assumed the absence of ROS production.

Conclusion

Among the wide concern of the risk assessment of nanotechnologies, this entry focused on the particular case of TiO_2 nanoparticles used in sunscreen as UV blockers. Three nanoparticulate samples, corresponding respectively to gradual steps of complexity in the chemical composition from the bare mineral nanoparticle to the final commercialized cream, were compared. Both the effects of the respective surface properties and of the environmental conditions on the fate of nanoparticles in aqueous system and their photoactivity could be evaluated.

In terms of surface properties, the dispersibility of nanoparticles in water highly depends on their affinity for the dispersing medium, i.e., their hydrophilic/hydrophobic character. While bare TiO_2 nanoparticles display a hydrophilic surface ready for easy aqueous dispersion, TiO_2-based

nanomaterials used in sunscreen formulation are more often functionalized with a hydrophobic coating. Moreover, final sunscreen formulation is also typically constituted of a water in oil emulsion, with hydrophobic character. However, along aging, the nanomaterial's coating was mainly altered and desorbed, and the sunscreen underwent significant textural reorganization, both tending to release in water submicronic residues with hydrophilic surface. These colloidal by-products involving TiO_2 nanoparticles then remained stable in aqueous suspension and thus ready for further transfer or interaction. The TiO_2 nanomaterial residues had a $Al(OH)_3$ surface newly exposed originating from their underlying functionalization. The sunscreen residues were constituted for more than 90 w/w% of organic components from the formulation, embedding the TiO_2 nanoparticles and driving the colloid's surface properties. In both altered samples, the non-exposure of bare TiO_2 mineral to the external environment prevented its well-known photoactivation, giving rather inactive colloids in terms of UV excitation and ROS production.

By controlling the occurrence and intensity of electrostatic interactions, both pH and salinity in water mainly drive not only the dispersion stability of nanoparticles in water but also their potential interaction and heteroaggregation with the surrounding elements (e.g., naturally occurring colloids, porous media, etc.). The respective surface charges of the involved elements with regard to that of nanoparticles are indeed a major parameter that determines the fate of both.

However, such easily predictable interactions are often screened in natural systems by the presence of natural organic matter. These molecules indeed likely interact with dispersed nano- or not nanoparticulate matter, strongly altering its surface properties, i.e., its dispersion stability. Two opposite effects can be typically observed and were evidenced again for nano-TiO_2 in this entry. Bridging flocculation inducing nanoparticle aggregation occurred at low NOM/nanoparticle ratio when adsorbed macromolecules were able to build bridges between neighbor nanoparticles. Steric or electrostatic repulsions, favoring nanoparticle's dispersion, were rather obtained at high NOM/nanoparticle's ratio when the nanoparticles' surface got saturated with adsorbed macromolecules.

Despite its focus on only one kind of nanoproduct, this entry clearly raises the major role played by the surface properties of the studied nanoparticles on their fate and impact when released into the aqueous environment. Surface properties are moreover expected to evolve continuously along the product life cycle, depending on the alteration and interaction enabled. This points out the high complexity in evaluating the global risk assessment of nanotechnologies and, to manage this, raises the key question of the relevance of the studied nanoparticles with regard to those potentially occurring in the environment.

F

Cross-References

- ▶ Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Nanostructured Functionalized Surfaces
- ▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity
- ▶ Small-Angle Scattering of Nanostructures and Nanomaterials
- ▶ Toxicology: Plants and Nanoparticles

References

1. Nanowerk database on nanomaterials, nanotechnology products and related companies and laboratories. http://nanowerk.com
2. Woodrow Wilson Institute: Inventory of manufacturer-identified nanotechnology-based consumer products introduced to the market. http://nanotechproject.org
3. Health and Consumer Protection Directorate General of the European Commission: Nanotechnologies, a preliminary risk analysis on the basis of a workshop, Brussels (2004)

4. U.S. Environmental Protection Agency: Nanotechnology white paper. Nanotechnology Workgroup, Science Policy Council, Washington, DC (2005)
5. SCENIHR: Opinion on the Appropriateness of Existing Methodologies to Assess the Potential Risks Associated with Engineered and Adventitious Products of Nanotechnologies, pp. 41–58. Health and Consumer Protection Directorate General of the European Commission, Brussels (2005)
6. United Nations Environment Programme: Chapter 7: Emerging challenges-nanotechnology and the environment. In: Geo Year Book 2007, United Nations Environment Programme Division of Early Warning and Assessment, Nairobi. http://unep.Org/geo/yearbook/yb2007/pdf/gyb2007_english_full.Pdf (2007)
7. Robichaud, C., Uyar, A., Darby, M., Zucker, L., Wiesner, M.: Estimates of upper bounds and trends in nano-TiO_2 production as a basis for exposure assessment. Environ. Sci. Technol. **43**, 4227–4233 (2009)
8. EWG: Sunscreen investigation. Section 4 nanotechnology and sunscreens. Environmental Working Group (2009), Washington, DC, http://www.ewg.org/2014sunscreen/
9. Senjen, R.: Nanomaterials – health and environmental concerns. In: Nanotechnologies in the 21st Century. EEB, Brussels (2009)
10. Wijnhoven, S.W.P., Dekkers, S., Hagens, W.I., De Jong, W.H.: Exposure to nanomaterials in consumer products. Letter report. RIVM, National Institute for Public Health and the Environment, The Netherlands (2009)
11. Kaegi, R., Ulrich, A., Sinnet, B., et al.: Synthetic TiO_2 nanoparticle emission from exterior facades into the aquatic environment. Environ. Pollut. **156**(2), 233–239 (2008)
12. Kiser, M.A., Westerhoff, P., Benn, T., Wang, Y., Pérez-Rivera, J., Hristovski, K.: Titanium nanomaterial removal and release from wastewater treatment plants. Environ. Sci. Technol. **43**(17), 6757–6763 (2009)
13. Reijnders, L.: The release of TiO_2 and SiO_2 nanoparticles from nanocomposites. Polym. Degrad. Stab. **94**(5), 873–876 (2009)
14. Wijnhoven, S.W.P., Dekkers, S., Kooi, M. et al.: RIVM report 2011 Nanomaterials in consumer products – Update of products on the European market in 2010, Report 340370003 (2010)
15. Som, C., Berges, M., Chaudhry, Q., et al.: The importance of life cycle concepts for the development of safe nanoproducts. Toxicology **269**(2–3), 160–169 (2010)
16. Brant, J., Labille, J., Wiesner, M., Rose, J., Bottero, J. Y.: Nanomaterials transport and fate in the environment. In: Wiesner, M., Bottero, J.-Y. (eds.) Environmental Nanotechnology Applications, and Impact of Nanomaterials. McGraw Hill, New York (2007)
17. Labille, J., Brant, J.: Stability of nanoparticles in water. Nanomedicine **5**(6), 985–998 (2010)
18. Solovitch-Vella, N., Labille, J., Rose, J., et al.: Concurrent aggregation and deposition of TiO_2 nanoparticles in a sandy porous media. Environ. Sci. Technol. **44**(13), 4897–4902 (2010)
19. Labille, J., Feng, J., Botta, C., et al.: Aging of TiO_2 nanocomposites used in sunscreen. Dispersion and fate of the degradation residues in aqueous environment. Environ. Pollut. **158**(12), 3482–3489 (2010)
20. Auffan, M., Pedeutour, M., Rose, J., et al.: Surface structural degradation of tio2-based nanomaterial used in cosmetics. Environ. Sci. Technol. **44**(7), 2689–2694 (2010)
21. Botta, C., Labille, J., Auffan, M., et al.: TiO_2-based nanoparticles released in water from commercialized sunscreens in a life cycle perspective: structures and quantities. Environ. Pollut. **159**, 1543–1550 (2011)
22. Jolivet, J.P., Froidefond, C., Pottier, A., et al.: Size tailoring of oxide nanoparticles by precipitation in aqueous medium. A semi-quantitative modelling. J. Mater. Chem. **14**(21), 3281–3288 (2004)
23. Labille, J., Thomas, F., Bihannic, I., Santaella, C.: Destabilization of montmorillonite suspensions by Ca^{2+} and succinoglycan. Clay Miner. **38**(2), 173–185 (2003)
24. Gottschalk, F., Sun, T., Nowack, B.: Environmental concentrations of engineered nanomaterials: review of modeling and analytical studies. Environ. Pollut. **181**, 287–300 (2013)
25. NANOHETER: SIINN Research Program (2013–2016) the ERA-NET on Nanosafety for a Safe Implementation of Innovative Nanoscience and Nanotechnology; Heteroaggregation of manufactured nanoparticles with naturally occurring colloids in surface water. http://nanoheter.cerege.fr (2015)
26. Praetorius, A., Labille, J., Scheringer, M., et al.: Heteroaggregation of titanium dioxide nanoparticles with model natural colloids under environmentally relevant conditions. Environ. Sci. Technol. **48**, 10690–10698 (2014)

Fatigue Strength

▶ Nanomechanical Properties of Nanostructures

FDTD

▶ Finite-Difference Time-Domain Technique
▶ Nanopolaritonics

Femtocalorimetry

▶ Nanocalorimetry

Fetus

► Use of Nanotechnology in Pregnancy

Fiber Reinforced Composite

► Arthropod Strain Sensors

Fibrillar Adhesion

► Gecko Adhesion

FIB-SEM

► Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells

Field Electron Emission from Nanomaterials

Mahendra A. More[1], Dilip S. Joag[1] and Richard G. Forbes[2]
[1]Centre for Advanced Studies in Materials Science and Condensed Matter Physics, Department of Physics, Savitribai Phule Pune University, Pune, Maharashtra, India
[2]Advanced Technology Institute and Department of Electrical and Electronic Engineering, Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey, UK

List of Abbreviations

CFE	Cold field electron emission
DC	Direct current
ECD	Emission current density
ES	Emission site
ESD	Emission site density
ET	Exactly triangular
FE	Field electron emission
FEF	Field enhancement factor
FEM	Field electron (emission) microscopy
FN	Fowler-Nordheim
ISQ	International System of Quantities
ITO	Indium tin oxide
LAFE	Large-area field emitter
LCL	Local conversion length
SN	Schottky-Nordheim
STFE	Single-tip field emitter
VCF	Voltage-to-barrier-field conversion factor

Synonyms

Cold field electron emission from nanostructured materials

Definition

Field electron emission is the quantum-mechanical tunneling of electrons from a material surface into vacuum, usually from electron states near the emitter Fermi level, under the influence of a high electric field typically of magnitude 1–10 V/nm.

Overview

Cold Field Electron Emission

In field electron emission (FE), applying a high negative electric field, of magnitude typically 1–10 V/nm, causes electron tunneling from a material surface into vacuum. If electrons are emitted mainly from states near the Fermi level, as usual in room-temperature emission, the process is called "cold field electron emission" (CFE). CFE from metals and, later, semiconductors and other materials has attracted experimental attention for more than 90 years. The first detailed account in English was given in 1922 by Lilienfeld [1], who wanted to build small X-ray devices for medical applications; the first satisfactory physical explanation was given by Fowler and Nordheim (FN) in 1928/1929 [2–4].

F

Note that the derivation in Ref. [2] lacks transparency and may not be mathematically valid; however, Eq. 18 in Ref. [2] is a correct asymptotic approximation for transmission coefficient, and the approximate formula Eq. 21 for emission current density is correctly derived (see Ref. [4]). However, Eq. 22 in Ref. [2] is not correct (see Ref. [3]). Reference [5] includes a brief overview of earlier history.

In the 1930s, Müller developed the clean, pointed-wire field electron emitter and with it the field (electron) emission microscope (FEM) [6]. By the 1960s, the surface-science technique of field emission microscopy was well developed [7]. Later, FE energy distributions were measured [8]. The theory of FE and FE energy distributions from metals and semiconductors, as it stood in the 1980s, is described in a textbook by Modinos [9].

The most useful surface-science application was the study of adsorbate behavior, notably surface diffusion on clean metal surfaces. Probe-hole techniques allow the current from a small surface area to be measured [10]. A field emitter exhibits faces of different crystallographic orientations; these can be accessed in a single experiment.

Field emitters have technological applications as electron sources. *Single-tip field emitter* (STFE) sources have high brightness and small optical size and are routinely used in electron microscopy and e-beam lithography [11, 12]. Two-dimensional emitter arrays can be fabricated by various techniques, giving *large-area field emitters* (LAFEs) suitable for flat panel displays and maybe electron tube devices [13, 14].

The original LAFE was the "Spindt cathode" [15], a lithographically created array of microfabricated metallic tips (commonly Mo). Until very recently, several factors limited its operating performance and lifetime. In particular, Joule heating at high emission currents, coupled with bombardment by ionized residual vacuum-system gases, caused tip erosion and blunting.

Spindt cathodes currently find use in ion sources and mass spectrometers. Huge efforts have been invested in developing displays based on Spindt cathodes, with some success [14]. There is an ongoing effort to find alternatives that are robust in industrial vacuum conditions and can deliver high uniform current density at low operating voltages. A key factor is to develop nanostructures with a high density (high number per unit area) of individual sharp emitters.

Emergence of Nanomaterials

As regards sharpness, the small size of nanomaterial emitters gives them inherent advantages. All metallic and semiconducting nanomaterials have the potential to deliver current at lower applied voltages than conventional field emitters. Recent years have seen sharp increases in studies of how to prepare nanomaterial films with high emission site density (ESD) and into the properties of the resulting LAFEs.

Many materials, including metals and ceramics, have been synthesized in nanometric forms. These are characterized by large surface-to-volume ratio, as compared with bulk materials. To be suitable for CFE, materials must have vacuum compatibility (low vapor pressure), mechanical robustness (ability to withstand the mechanical stress due to high electric fields), and resistance to ion bombardment. A desired target is an average or "macroscopic" current density J_M in the range 100–1000 A/m^2 (10–100 mA/cm^2), for applied macroscopic fields of around 1–20 V/μm. Also, it must be possible to deposit the nanomaterial onto a suitable flat substrate, with uniform ESD and relative ease.

The nanomaterials of interest to CFE can be broadly classified into: amorphous carbon, carbon nanotubes, and graphene; nanoforms of wide-bandgap semiconductors such as SiC, AlN, GaN, BN, Si, and ZnO; and nanoforms and nanostructures of established electron emitters such as LaB_6. Carbon nanotubes have small radii (up to ∼50 nm) and are attractive because of the high field enhancement offered. Wide-bandgap semiconductors have low work function and in some cases negative electron affinity. Silicon nanostructures can be easily integrated with microelectronic devices.

Reference [16] discusses early work on carbon-based LAFEs and other nanostructured heterogeneous materials and possible emission mechanisms. References [17] and [18] cover

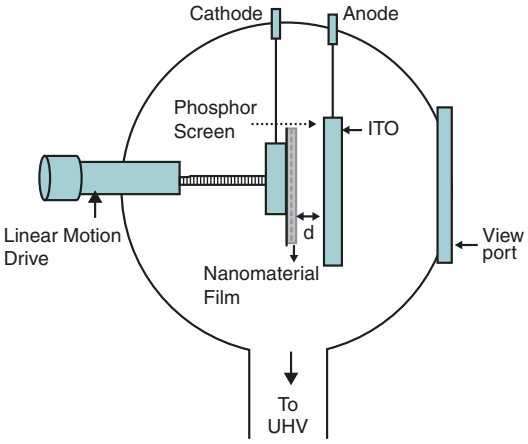
(ultimately unsuccessful) attempts to develop LAFEs based on amorphous carbon films. Reference [19] provides an overview of material issues related to cold cathodes. Reference [20] comprehensively reviews inorganic semiconductor nanostructures and their FE properties. References [21–23] comprehensively review the FE properties of carbon nanotubes. References [24] and [14] provide good general introductions to CFE and its applications in nanoelectronics. Many nanomaterials in principle have suitable basic properties, but only those that can be reliably deposited with a high ESD of sharp nanoforms are suitable for device applications.

Basic Production and Testing Methodologies

Nanostructured materials are usually deposited/coated onto a flat substrate (typically of area 1 cm^2), to form a LAFE. The substrate is an electrical conductor or semiconductor. The deposition method depends on the material properties. Methods used include chemical vapor deposition, arc plasma deposition, pulsed laser deposition, electrodeposition, electrochemical deposition, and spin coating. Si is the most common substrate, because it can be integrated with electronic device technology.

In the diode configuration often used for testing emitter properties, this coated substrate forms the cathode. The anode is normally an ITO-coated, conducting glass plate, with or without phosphor coating. The anode-cathode separation is typically in the range 100 μm to 5 mm, with the separation maintained by a quartz or alumina spacer.

Figure 1 shows a typical system. A holder carrying the anode and cathode is mounted in an ultrahigh-vacuum chamber that is subsequently evacuated to a pressure of 10^{-8}–10^{-7} mbar. A high voltage (typically up to 10 kV) from a stable DC supply is applied between anode and cathode. The emission current (typically 10^{-9}–10^{-3} A) is measured by an electrometer amplifier. Current-voltage characteristics are usually measured over a wide voltage range. Current stability is studied by setting the current initially at some appropriate level and then recording its value at regular time intervals. An FE image



Field Electron Emission from Nanomaterials, Fig. 1 Schematic of planar diode assembly mounted in a vacuum chamber

(showing the distribution of emission current density across the emitter) can be observed on the anode screen and recorded as appropriate.

Elements of Emission Theory

Introduction

In describing CFE theory for nanomaterials, three particular problems confront the writers of an encyclopedia entry. First, there are unfortunate clashes in terminology, notation, and practice between older theory papers and the newer experimental papers on nanomaterials. This work uses a compromise convention defined in "Appendix." Second, for interpreting CFE results, experimental nanomaterial papers often employ an oversimplified "elementary" equation originally used for undergraduate teaching about CFE from metal STFEs of large apex radius and, when describing LAFE emission, often omit an important correction factor (α_f below).

This elementary equation is not strictly applicable to any real material, but is often used as part of an "elementary" method of analyzing the slope of a Fowler-Nordheim plot (FN plot) (see below). The method aims to characterize emitter sharpness by extracting a value for its field enhancement factor (FEF) (see below). Unfortunately, the

method is not valid in many cases, which causes the third problem. Past inability to identify invalid cases has meant that many spuriously large FEF values have been published. A small survey [25] suggested that perhaps 40 % of the many hundreds of relevant papers in the literature report spuriously large FEF values.

Detailed CFE theory for nanomaterials is material specific, often exceptionally complicated and usually far from fully worked out. We give here a high-level overview of CFE theory and current-voltage data analysis that can in principle accommodate the equations conventionally used and, hopefully, more-accurate theory yet to be developed. Table 1 above gives definitions and values of universal constants used. As usual in FE contexts, fields, current densities, and related quantities are treated as positive, even though they would be negative in classical electromagnetism.

Basic Equations of Cold Field Electron Emission

In CFE theory, a central quantity is the *local emission current density* (local ECD) J_L at any arbitrary position on the emitter surface. The local surface electric field E_L at this position defines the barrier experienced by electrons tunneling there and is called here the (local) *barrier field*. J_L depends on E_L, and both vary with position across the surface. The emission current i_{es} from a single emission site ("ES") is found by integrating J_L over the area of the site; thus $i_{es} = \int_{ES} J_L dA$. If a characteristic surface position "C" within the site is considered (say, where J_L has a local maximum), and J_L has the value J_C there, then the result of integration can be written $i_{es} \equiv J_C A_{ns}$. This form of area definition was introduced in Ref. [3] and defines a *notional emission site area* A_{ns}. Similarly, if i_e denotes the *total emission current* from a LAFE, then the formula

$$i_e = A_n J_C \tag{1}$$

defines a *notional emission area* A_n for the LAFE, where J_C is the LAFE maximum value of J_L (or, alternatively, some other LAFE characteristic value of J_L).

CFE theory can be formulated at various levels of complexity. In *orthodox CFE theory*, which is a complexity level used by theoreticians to discuss experimental results (on the grounds that this is the simplest theory that has any chance of being nearly realistic), the tunneling barrier is assumed to be a Schottky-Nordheim (SN) barrier (see below), and J_C can be written in terms of the characteristic barrier field E_C by the linked equations:

$$J_C = \lambda_C J_{kC}. \tag{2a}$$

$$J_{kC} = a\phi^{-1}E_C^2 \exp\left[-bv_F\phi^{3/2}/E_C\right]. \tag{2b}$$

In Eq. 2b, a and b are, respectively, the first and second Fowler-Nordheim (FN) constants (see Table 1), ϕ is the local work function at point "C," and v_F is a particular value of a mathematical function best called the *principal*

Field Electron Emission from Nanomaterials, Table 1 Basic universal emission constants. Values are given in the units often used in field emission. e denotes the elementary positive charge, m_e the electron mass in free space, and h_P Planck's constant. In these units, e has the value 1 eV V^{-1}. (Based on 2014 values of the fundamental constants)

Name	Symbol	Derivation	Expression	Numerical value	Units
JWKB constant for electron	g_e	–	$4\pi(2m_e)^{1/2}/h_P$	10.246 33	$eV^{-1/2}\ nm^{-1}$
First Fowler-Nordheim constant	a	$z_S(e/g_e)^2$	$e^3/8\pi h_P$	1.541 434	$\mu A\ eV\ V^{-2}$
Second Fowler-Nordheim constant	b	$2g_e/3e$	$(8\pi/3)(2m_e)^{1/2}/eh_P$	6.830 890	$eV^{-3/2}\ V\ nm^{-1}$
Schottky constant	c	–	$(e^3/4\pi\varepsilon_0)^{1/2}$	1.199 985	$eV\ (V/nm)^{-1/2}$
(Constant in Eq. 20)	–	c^2	–	1.439 964	$eV^2\ (V/nm)^{-1}$
(Constant in Eq. 39)	–	ab^2	–	$7.192\ 494 \times 10^{-5}$	$A\ eV^{-2}\ nm^{-2}$
(Constant in Eq. 27)	–	bc^2	–	9.836 239	$eV^{1/2}$

Schottky-Nordheim (SN) barrier function. The value of v_F depends on the values of ϕ and E_C and typically is around 0.7.

The parameter J_{kC} defined by Eq. 2b is called the *characteristic kernel current density* (for the SN barrier), and the parameter λ_C is the *characteristic local pre-exponential correction factor.* The reason for introducing J_{kC} and splitting Eq. 2 as above is that – for given choices of ϕ_C, E_C, and barrier form (here the SN barrier) – J_{kC} can be calculated *exactly.* All theoretical uncertainty in predicting the ECD J_C appears in λ_C. Exact realistic calculation of λ_C would be intensely difficult, is in some cases beyond the current capabilities of quantum mechanics, and is unlikely to happen in the foreseeable future. For metals, λ_C is thought (in 2015) to lie in the range $0.005 < \lambda_C < 11$.

Equation 2 is the $J_C(E_C)$ form of the *orthodox Fowler-Nordheim-type (FN-type) equation*.

Inserting Eq. 2a in Eq. 1 yields the $i_e(E_C)$ form of this equation, namely,

$$i_e = A_n J_C = A_n \lambda_C J_{kC} \equiv A_f J_{kC}, \tag{3}$$

where the *formal emission area* A_f [$\equiv A_n\lambda_C$] is defined by Eq. 3.

Many experimental nanomaterial papers, however, use elementary FN theory as found in undergraduate textbooks. In elementary theory, the tunneling barrier is taken as exactly triangular, and the parameters v_F and λ_C are replaced by unity, yielding the *elementary FN-type equations*:

$$J_C^{el} = J_{kC}^{el} \equiv a\phi^{-1}E_C^2 \exp\left[-b\phi^{3/2}/E_C\right]. \tag{4}$$

$$i_e^{el} = A_f J_{kC}^{el}. \tag{5}$$

Current density and current values predicted by orthodox FN-type equations can be significantly different from those predicted by elementary FN-type equations.

Auxiliary Parameters

Area efficiency parameters. The parameters A_n and A_f are always much less than the geometrical (or *macroscopic*) *area* A_M of the LAFE (sometimes called its *footprint*). The *notional area efficiency* α_n (formerly called the "area efficiency of emission" and denoted by α_M) and the *formal area efficiency* α_f are defined via the relationships:

$$\alpha_n \equiv A_n/A_M, \quad \alpha_f \equiv A_f/A_M, \quad \alpha_f = \alpha_n\lambda_C. \tag{6}$$

These parameters are measures of how much of the device area is actually emitting. Values of α_f are not well known, but are thought to usually lie in the range $10^{-11} < \alpha_f < 10^{-3}$.

For a LAFE, the average (or *macroscopic*) *current density* J_M (averaged over the whole LAFE footprint) is defined by

$$J_M \equiv i/A_M. \tag{7}$$

This macroscopic current density is given in orthodox theory by

$$J_M = (A_n/A_M)J_C = \alpha_n J_C = \alpha_f J_{kC}, \tag{8}$$

and in elementary theory by

$$J_M^{el} = \alpha_f J_{kC}^{el} = \alpha_f a\phi^{-1}E_C^2 \exp\left[-b\phi^{3/2}/E_C\right]. \tag{9}$$

The values of all the area-like parameters may need to depend on whether they are being used in the orthodox equation or the elementary equation. In LAFE literature, α_f is often wrongly omitted from elementary FN-type equations where it ought to be present. The resulting defective equation massively overpredicts values of emission current density and emission current (which may be misleading for nonexperts) and probably also makes it difficult for experimentalists to see how to extract α_f-values from experimental results. It needs to be emphasized that omission of α_f is not the result of using an elementary model; rather it is a serious and widespread conceptual error.

The basic reasons for using "formal" area parameters, in addition to the existing "notional" area parameters, are twofold. (1) It is better to have only one parameter subject to theoretical uncertainty in an equation for i or J_M. (2) The "formal" parameters correspond more closely to

what experimentalists would deduce from FN plots in orthodox emission situations (see below), in contrast to the "notional" parameters, which correspond more closely to what an exact theory would calculate or to the emitting area that might be seen in an FE image.

Barrier field and related parameters. The *emission voltage* V_e (also called the "device voltage") is the voltage between the emitting region at an emitter tip (or emission site) and the counter-electrode. For simplicity, it is assumed here that V_e is the same for all emitter surface points, although this may not be true for some real emitters. The relation between E_L and V_e can be written in two equivalent ways:

$$E_L = \beta_V \; V_e \tag{10a}$$

$$E_L = V_e/\zeta_L \tag{10b}$$

where β_V is the *local voltage-to-barrier-field conversion factor* (VCF) and ζ_L [$\equiv 1/\beta_V$] is the *local conversion length* (LCL). Where measurement circumstances allow, a *characteristic VCF* $\beta_{V,C}$ and/or a *characteristic LCL* ζ_C can be found from experimental current-voltage data (see below). Experimental values of ζ_C typically lie in the range 20–1000 nm, with sharper emitters tending to have smaller ζ_C-values. Gomer's classic formula $E_a = V_e/k_a r_a \approx V_e/5r_a$ [7] (where "a" labels apex values) makes the LCL five times the emitter apex radius r_a, but this is not a good approximation for many LAFEs, because the value of ζ_C is determined by the whole system geometry.

Equation 10a has been the more commonly used (usually with the symbol β rather β_V). However, because the symbol β is widely used in modern CFE literature to denote field enhancement factors (FEFs) (see below), rather than VCFs, it is recommended that (in the interests of avoiding confusion Eq. 10b rather than Eq. 10a should be used in the future.

LAFEs are normally produced/tested in parallel-plane geometry, using a plate separation d_{sep} that is very much greater than the height h of any nanostructure on the cathode plate. In these conditions, a *(true electrostatic) macroscopic field* E_M can be defined by

$$E_M = V_e/d_{sep}, \tag{11}$$

and a (*true electrostatic) field enhancement factor* (FEF) β_F by

$$E_L = \beta_F E_M. \tag{12}$$

It follows that (in parallel-plane geometry) β_F, β_V, and ζ_L are related by

$$\beta_F = \beta_V d_{sep} = d_{sep}/\zeta_L. \tag{13}$$

Thus, for a LAFE, the characteristic FEF $\beta_{F,C}$ can be obtained from the characteristic VCF $\beta_{V,C}$ or (better) the characteristic LCL ζ_C.

Note that some authors call β_V a "field enhancement factor," despite the fact that it has different units and dimensions from β_F. This confusing convention adds to the case for replacing use of β_V by use of the local conversion length ζ_L.

For nonparallel-plate configurations or if h is comparable with d_{sep}, the term "macroscopic field" is sometimes given meanings or a definition different from that given above. These alternatives are not discussed here.

If the system geometry, including the detailed geometry of individual emitter/emission sites, is known, then FEF values and/or LCL values can be estimated by treating the objects involved as classical conductors and solving Laplace's equation numerically. However, this is not always valid; in particular, it will not be valid if field penetration and band bending occur [9, 26].

A situation often discussed is a classical conductor with the shape of a hemisphere on a cylindrical post of radius r_c, standing alone on one of two parallel planes, with the second plane (the counter-electrode) well separated from the first. There are no known exact analytical solutions to this electrostatic problem; the best approximate formulae are fits to numerical solutions of Laplace's equation. Various formulae for the FEF $\beta_{F,a}$ at the hemisphere apex have been given (see, e.g., [27–30]). The best simple formula,

good in the range $100 < (h/r_c) < 500$ and accurate to within 12 % [27], seems to be

$$\beta_{F,a} \approx 0.72\ h/r_c, \tag{14}$$

where h is the height of the post + hemisphere. An older version, with the factor 0.72 omitted, is often used, but is significantly less accurate.

For these formulae to be adequately valid, the counter-electrode must be separated from the post apex by a distance d_{gap} very large in comparison with h (ideally, $d_{gap} > 10h$). With an array of identical posts, the posts must be separated laterally by a distance at least h to $2h$; otherwise, the value of $\beta_{F,a}$ is significantly reduced; this effect is best called *mutual screening*.

Parameters of Merit for Large-Area Field Emitters

With a LAFE, the technological objective is to generate a macroscopic current density J_M that is both sufficiently uniform and as high as is needed, using an extraction voltage V_e that is as low as possible. Important engineering parameters are: (a) the characteristic field enhancement factor, $\beta_{F,C}$; (b) the characteristic local conversion length ζ_C; (c) the *onset voltage* $V_{e,on}$ at which J_M has some specified value $J_{M,on}$ (usually chosen as between 10^{-3} and 100 A/m^2); and (d) the *emission site density* (ESD). The formal area efficiency α_f would also be an important parameter if experimental practice were to measure this (where possible), but currently it is not.

The characteristic FEF $\beta_{F,C}$ assesses how "sharp" the most strongly emitting LAFE nanofeatures are; a primary requirement for a good emitter is sufficiently high $\beta_{F,C}$. The characteristic LCL ζ_C assesses the geometry of the whole engineered LAFE (which, in the finished product, would often have both an extracting electrode and an anode); ζ_C depends both on emitter sharpness and on how close the extracting electrode can be put to the emitter. The onset voltage $V_{e,on}$ also assesses these things, but additionally takes into account the emitter work function ϕ, since low ϕ leads to low $V_{e,on}$.

ESD values are best determined by measurements with an appropriate scanning machine (e.g., Ref. [31]), given some criterion for identifying emission sites. The formal area efficiency α_f would probably be linked in a qualitative or semi-quantitative way to the ESD, but few values are available (how to determine α_f reliably is a current research topic).

Another parameter often used is the *apparent macroscopic onset field* $E_{A,on}$ defined (in parallel-plane geometry), via measured onset voltage, as

$$E_{A,on} = V_{m,on}/d_{sep}. \tag{15}$$

$E_{A,on}$ is somewhat easier to determine than $\beta_{F,C}$ and ostensibly assesses the combined effect of emitter sharpness and low work function. It is related to $\beta_{F,C}$ by

$$E_{A,on} = (E_{C,on}/\beta_{F,C})(V_{m,on}/V_{e,on}) \tag{16}$$

where $E_{C,on}$ is the *characteristic local onset field*. The value of $E_{C,on}$ that corresponds to a given onset macroscopic current density $J_{M,on}$ varies somewhat, depending on LAFE details, but as an order-of-magnitude estimate, one can take $E_{C,on} \sim 1$ V/nm.

In practice, for non-lithographically produced emitters, the preparation of good emitters may involve compromise between adequate uniformity and high characteristic FEF. The advantage of lithographic techniques is that LAFEs can be made with high uniform ESD; the main disadvantages are the higher infrastructure and production costs involved.

Experimental CFE current-voltage characteristics are normally interpreted by using Fowler-Nordheim plots, as discussed below. To understand the associated theory, it is helpful to be aware of further aspects of basic CFE theory.

Conceptual-Level Theory of Local Emission Current Density (ECD)

For notational simplicity in the theory below, a modified notation is now used. The symbols E and J are used below to refer to *characteristic local values* of these parameters (and not to

macroscopic parameters); β and ζ refer to *characteristic* local values of FEF and LCL. The voltage conversion factor (VCF) represented above by β_V is not used (but its characteristic value would be given by $1/\zeta$). Note that this notation convention is different from that used in many experimental papers, which use E and J to denote macroscopic quantities.

The details of local ECD theory involve the electronic band structure of the emitting material. If emission comes mainly from a single electronic band or sub-band, the following approach is used.

An electron in a specific emitter state, approaching the surface from the inside, has an escape probability given by the *tunneling probability D* for the barrier it encounters. A reference state "R" is selected (usually the state with the highest D-value from among those that would be occupied at 0 K), and the corresponding tunneling probability D_R is evaluated. The contributions to the local ECD J_L from all occupied emitter states are then summed, and the result is written

$$J_L \equiv Z_R D_R. \tag{17}$$

The parameter Z_R is defined by Eq. 17 and is termed the *effective supply for state* R.

The mathematical form of Z_R can differ widely for different materials, but the form of D_R usually depends only on the nature of the barrier. This is much the same for all materials and is examined below. In more-general situations, emission may come from more than one band or sub-band and/or from well-defined individual electron states (including surface states); J_L may then need to be expressed as a sum of terms of the form of Eq. 17.

In practice, CFE theory is nearly always presented for the case of single-band emission where the reference state is that of a Fermi-level electron moving *forward* (i.e., at right angles to and toward the emitting surface). This electron sees a barrier equal to the local work function ϕ. Other parameters relevant to this reference state are subscripted "F." This theory, set out below and usually called *Fowler-Nordheim (FN) theory*, was originally derived for free-electron metals, but also applies (to some extent) to semiconductors if most of the emission comes out of surface states (which is thought to often be the case [9]) and if band-structure pinning occurs. FN theory is widely used as if it were an approximation applicable to all materials, although this is not in valid in all contexts.

Tunneling Probability Formulae for the CFE Regime

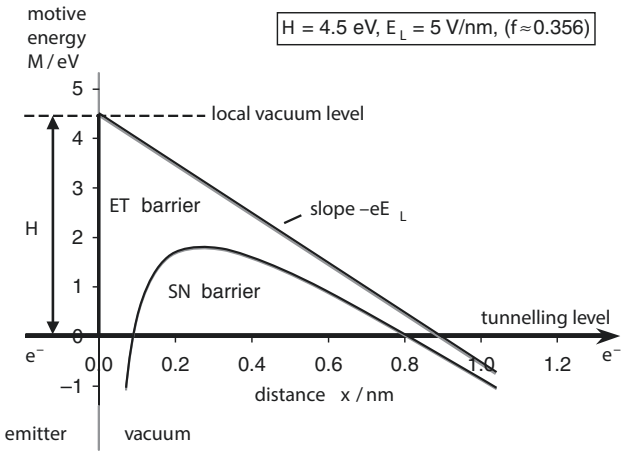
Deep tunneling is tunneling that occurs at energy levels well below the barrier maximum. In the CFE regime, by definition, nearly all escaping electrons (including any electron in state F) escape by deep tunneling. Consequently, a good approximation for D_F is the *Landau and Lifshitz tunneling probability formula* (see Ref. [32]):

$$D_F \approx P_F \exp[-G_F] \equiv P_F \exp\left[-v_F b \phi^{3/2}/E\right]. \tag{18}$$

For the electron in state F, v_F ("nu$_F$") is a correction factor related to the barrier's mathematical form; P_F is the tunneling pre-factor; and G_F is the barrier strength (see below) (also called the "Gamow exponent" and the "JWKB exponent"). As a rough rule, Eq. 12 is valid to better than 5 % if G_F and P_F are well known and $G_F > 3$; the accuracy of Eq. 12 improves significantly as G_F increases.

The *barrier form* (or "shape") is determined by what terms appear in the mathematical expression for the *electron motive energy* $M(x)$, where x is distance measured from the emitter's electrical surface. When the electron wave function separates in Cartesian coordinates (which is an underlying assumption usually made), $M(x)$ is identified by writing the Schrödinger-equation component for motion normal to the emitter surface in the form $\hat{K}\psi = -M\psi$, where $\hat{K}$ is the kinetic-energy operator. Applying this to state F gives $M_F(x)$. G_F is then defined by $G_F \equiv g_e \int \{M_F(x)\}^{1/2} dx$, where g_e is a universal constant defined in Table 1 and the integral is taken across the region where $M_F(x) \geq 0$; the barrier-form correction factor v_F is then defined via Eq. 18.

In principle, Eq. 18 and related definitions apply to a barrier of any well-behaved form.



Field Electron Emission from Nanomaterials, Fig. 2 Electron motive energy as a function of distance x from the emitter's electrical surface, for the exactly triangular (ET) barrier and for the Schottky-Nordheim (SN) barrier defined by Eq. 19. For both barriers depicted here, the zero-field height $H = 4.5$ eV and the barrier field $E_L = 5$ V/nm

In practice, CFE theory often uses the *Schottky-Nordheim (SN) barrier* $M^{SN}(x)$, which is the simplest realistic barrier (but is not physically correct for very sharp emitters). This barrier is illustrated in Fig. 2 and defined by

$$M^{SN}(x) = H - eEx - e^2/16\pi\varepsilon_0 x, \qquad (19)$$

where e is the elementary positive charge and ε_0 the electric constant. For the SN barrier, v_F is given by $v(f)$ ["vee(f)"], where $v(f)$ is a particular value of a mathematical function $v(l')$ called the *principal SN barrier function* [33, 34]. The parameter l' is a mathematical variable [34]. In the derivation of a formula for CFE current density through a SN barrier of zero-field height ϕ, l' is put equal to the *scaled barrier field* f for this barrier, where f is a dimensionless physical modeling parameter defined by

$$f \equiv E/E^* = c^2\phi^{-2}E. \qquad (20)$$

Here, E^* is the field necessary to reduce an SN barrier of zero-field height ϕ to zero, and c is the *Schottky constant* shown in Table 1. For notational simplicity, the parameter value $v(f)$ is sometimes denoted by the alternative symbol v_F, as earlier.

In older literature (e.g., [35, 36]), the function v is expressed in terms of a different mathematical variable y [$= +\sqrt{l'}$]. The change from y to l' is a mathematical simplification [34] that also helps physical understanding [33]. A good simple approximate formula for $v(l')$, and hence $v(f)$, now exists [37, 38]:

$$v_F \equiv v(f) \approx 1 - f + (1/6)f\ln f. \qquad (21)$$

This formula is valid to better than 0.35 % over the whole range $0 \le f \le 1$ and is adequate for nearly all technological purposes.

For the SN barrier, values of the tunneling pre-factor P_F have been calculated as a function of barrier field and zero-field barrier height by Mayer [39]. In practice, the pre-factor nearly always lies in the range 0.2–1.5 and can usually be neglected.

As already noted, many nanomaterial papers do not use the SN barrier but employ the elementary approximation that puts $P_F = 1$ and models tunneling as taking place through the exactly triangular (ET) barrier $M^{ET}(x) = H–eEx$. For this elementary model, $v_F = 1$.

Emission-Value-Based Fowler-Nordheim-Type Current-Voltage Equations

A feature of CFE literature is that FN-type equations (and related analyses) are formulated using several different pairs of independent and dependent variables, in particular $J_L(E_L)$, $J_M(E_M)$, $J_M(E_A)$, $i_e(V_e)$, and $i_m(V_m)$, where i_m and V_m are the measured current and voltage (which may be different from i_e and V_e) and E_A is the "apparent macroscopic field" discussed further below. Although $J_M(E_M)$- and $J_M(E_A)$-type formulations

are common in nanomaterial papers, $i_e(V_e)$- and $i_m(V_m)$-type formulations are theoretically more satisfactory and are used here. (It is shown later how to derive $J_M(E_M)$-type results from a current-voltage formulation, when circumstances permit.)

Equations above imply that the $i_e(V_e)$ form of the new standard FN-type equation can be written as the linked equations:

$$i_e = C_{iVe} V_e^2 \exp\left[-v_F b \phi^{3/2} \zeta_C / V_e\right] \tag{22a}$$

$$C_{iVe} \equiv A_f a \phi^{-1} \zeta_C^{-2} \tag{22b}$$

where C_{iVe} is defined by Eq. 22b. In so-called FN coordinates, Eqs. (22a) and (22b) together become

$$\begin{aligned} L_e\left(V_e^{-1}\right) &\equiv \ln\left\{i_e / V_e^2\right\} \\ &= \ln\{C_{iVe}\} - v_F b \phi^{3/2} \zeta_C / V_e, \end{aligned} \tag{23}$$

where $L_e\left(V_e^{-1}\right)$ is defined by Eq. 23. If any voltage dependence in C_{iVe}, ϕ, or ζ_C can be disregarded as "small," the tangent to this equation can be written [40] in the form

$$L_e = \ln\{r_{2012} C_{iVe}\} - s b \phi^{3/2} \zeta_C / V_e, \tag{24}$$

where s [33] and r_{2012} [40] are two of the SN barrier functions.

The *slope correction function* $s(f)$ and *2012 intercept correction function* $r_{2012}(f,\phi)$ are defined and (using Eq. 21) adequately approximated by

$$s(f) \equiv v_F - f \mathrm{d}v_F/\mathrm{d}f \approx 1 - f/6 \tag{25}$$

$$r_{2012}(f,\phi) \equiv \exp[-\eta \mathrm{d}v_F/\mathrm{d}f] \approx f^{-\eta/6} \mathrm{e}^{5\eta/6} \tag{26}$$

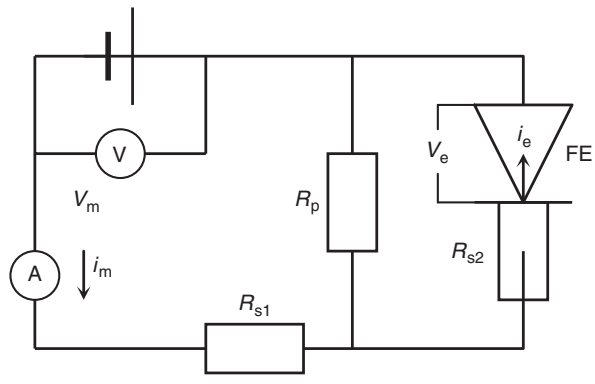
where the dimensionless work-function-related parameter η is given by ([33], see Eq. 6.1)

$$\eta \equiv b c^2 \phi^{-1/2}. \tag{27}$$

For illustration, $\eta = 4.64$ when $\phi = 4.50$ eV. Values of $s(f)$ are tabulated in [37] and [41], and values of $r_{2012}(f,\phi)$ in [40]; however, applying a calculator or a spreadsheet to Eqs. 25 and 26 provides values of s and r_{2012} that are adequate for most technological purposes. Illustrative values, for $\phi = 4.50$ eV and $f = 0.30$, are $s \approx 0.95$ and $r_{2012} \approx 120$. The suffix "2012" is being used to distinguish r_{2012} from an earlier function, now superceded, that was also called the intercept correction function, was denoted by r or r_N, and was defined differently.

Measured-Value-Based Fowler-Nordheim-Type Current-Voltage Equations

Figure 3 shows a schematic diagram of the circuits used to measure CFE current-voltage characteristics. A field electron emitter is a nonlinear device,



Field Electron Emission from Nanomaterials, Fig. 3 Illustrative measurement circuit for the current-voltage (i-V) characteristics of a field electron emitter, showing resistance in series and parallel with the emitting device FE. The emission parameters are denoted by $\{i_e, V_e\}$ and the measured parameters by $\{i_m, V_m\}$. Arrows denote the direction of *electron* flow

analogous to a pn-junction diode, and has an effective resistance R_e given by

$$R_e = V_e/i_e = \left(\phi\zeta_C^2/A_f a V_e\right)\cdot\exp\left[v_F b\phi^{3/2}\zeta_C/V_e\right]. \quad (28)$$

R_e is very large at low emission voltages, but becomes much smaller as V_e increases.

Careful system design can usually make the parallel resistance R_p very large (effectively infinite), but it is often impossible to eliminate the series resistances. When $R_p \approx \infty$, these combine to give a *total series resistance* R_s. Hence, conceptual distinction is needed between the *emission quantities* $\{i_e, V_e\}$ and the *measured quantities* $\{i_m, V_m\}$.

Usually one can take $i_e = i_m$, but V_e and V_m are related by

$$V_e = V_m - i_m R_s. \quad (29)$$

Defining a parameter Θ (the *voltage ratio*) by $V_e = \Theta V_m$ gives

$$\Theta \equiv V_e/V_m = R_e/(R_e + R_s). \quad (30)$$

Inserting this into Eq. 22a, and putting $i_m = i_e$, yields the $i_m(V_m)$ form of the new standard FN-type equation, which (as before) can be written as linked equations, namely,

$$i_m = C_{iVm} V_m^2 \exp\left[-v_F b\phi^{3/2}\zeta_C/\Theta V_m\right] \quad (31a)$$

$$C_{iVm} = A_f a \phi^{-1}\zeta_C^{-2}\Theta^2, \quad (31b)$$

where C_{iVm} is defined by Eq. 31b. In FN coordinates, Eq. 31 becomes:

$$L_m(V_{m^{-1}}) \equiv \ln\{i_m/V_{m^2}\} = \ln\{C_{iVm}] - v_F b\phi^{3/2}\zeta_C/\Theta V_m, \quad (32)$$

where $L_m(V_{m^{-1}})$ is defined by Eq. 32. The tangent to this curve can in principle be written in the mathematical form

$$L_m = \ln\{\rho C_{iVm}\} - \sigma b\phi^{3/2}\zeta_C/V_m, \quad (33)$$

where σ and ρ are generalized slope and intercept correction functions, but at the time of writing (in early 2015), it is – in most cases – impossible to reliably calculate (or even reliably estimate) values for σ and ρ. In most cases, this makes the usual methods of extracting reliable information from FN plots unworkable when $\Theta \neq 1$.

F

Interpretation of Experimental Current-Voltage Characteristics

Introduction

Existing data-analysis methods for measured current-voltage data are based either on the emission-value-based $[i_e(V_e)]$ form of the new standard FN-type equation given above (call these *orthodox methods*) or on the elementary equation derived from it (call this the *elementary method*). These methods were originally developed for firmly mounted metal emitters, of moderate to large apex radius, that were in good electrical contact with the high-voltage supply. Over the last few years, it has become increasingly clear that these traditional data-analysis methods do not work satisfactorily for many nanomaterial emitters, and that in principle it is important to use the measured-value-based $[i_m(V_m)]$ form of the relevant equation.

In practice, at present, this is usually only possible when we can approximate $i_m = i_e$ and $V_m = V_e$. However, for orthodox data-analysis methods to work adequately, *additional* theoretical constraints are necessary, and these give rise to the concept of the *orthodox emission situation* (specified below) – which is the situation in which orthodox methods do work adequately. If the actual emission situation is nonorthodox, then neither orthodox nor elementary data-analysis methods will work satisfactorily.

Thus, from the data-analysis point of view, what we need are a specification of what constitutes an orthodox emission situation, a reliable data-analysis method for orthodox situations, an experimental test of whether a given emission

situation is orthodox or not, and methodologies for carrying out data analysis in nonorthodox situations. The first three of these things are now in place; the fourth is the subject of active research, which is complicated because there are many possible physical causes for non-orthodoxy.

The Orthodox Emission Situation

The (theoretically perfect) orthodox emission situation complies with the following requirements [25]:

1. The emission voltage V_e is the same at all points on the emitter surface (i.e., there are no Fermi-level variations across the surface) and is equal to the measured voltage V_m (i.e., there is no voltage drop across a series resistance).
2. The measured current i_m is equal to the emission current i_e (i.e., there are no leakage currents).
3. The emission mechanism is deep tunneling through an SN barrier, and consequently the measured current is given by a version of the FN-type equation (Eq. 31).
4. In addition, in this equation (Eq. 31), (a) $\Theta = 1$; and (b) the only quantities that vary with V_m are V_m itself and the barrier-form correction factor v_F.
5. In particular, the local work function value ϕ is independent of both measured current and voltage and has the value assumed by the data analyst.

This set of theoretical assumptions (a) reduces Eq. 31 to Eq. 22, but with "e" everywhere replaced by "m," and (b) ensures that Eqs. 23, 24, 25, 26, and 27 are applicable. Theoretically, this is what existing orthodox analysis assumes (though in some cases further approximations are made). No real emitter or emission situation is "exactly orthodox," but many situations are "nearly orthodox," with orthodox data analysis adequately applicable.

However, with nanomaterial emitters, there are many real emission situations that are excluded by the orthodoxy requirements, particularly by requirement (4). These include situations where there is significant series resistance in the measurement circuit, situations where field-dependent changes in emitter geometry occur, and situations where the typical emission site apex radius is less than about 10–20 nm, to name but a few.

Orthodox Data Analysis Using Fowler-Nordheim Plots

A theoretical or experimental data plot made in FN coordinates is called a *Fowler-Nordheim plot (FN plot)*. Clearly, for an orthodox emission situation, the slope S_{iV}^{oes} and intercept $\ln\{I_{iV}^{oes}\}$ of the theoretical tangent described by Eq. 23 (but with "e" replaced by "m") are given via:

$$\begin{aligned} S_{iV}^{oes} &= -sb\phi^{3/2}/\zeta_C; \\ I_{iV}^{oes} &= r_{2012}\ \ C_{iVm} = r_{2012}A_f a\phi^{-1}\zeta_C{}^{-2}. \end{aligned} \tag{34}$$

The correction factors s and r_{2012} are both slowly varying functions of the measured voltage, because (in the orthodox emission situation) there is a well-defined linear relationship between V_m and the scaled barrier field f used above, namely,

$$V_m = (\zeta_C E^*)f. \tag{35}$$

Raw experimental data may also be plotted in this way, and the plot is often nearly linear. An exact straight line can be fitted to the experimental data, either manually or by linear regression, and its equation can be written

$$L_m = \ln\{I_{iV}^{fit}\} + S_{iV}^{fit}V_m{}^{-1}. \tag{36}$$

There are several slightly different ways of analyzing FN plots, but the most flexible is the *tangent method.* In this, one assumes that the experimental plot is equivalent to the theoretical tangent at some particular value $V_m = V_t$ that corresponds (via Eq. 35) to the value $f = f_t$. For this value the correction factors have the values s_t and r_t. One then takes the fitted slope and intercept in Eq. 36 to be good estimates of the theoretical slope and intercept in Eq. 34, which yields the formulae

$$\zeta_C = -S_{iV}^{fit}/s_t b\phi^{3/2} \tag{37}$$

$$A_f = I_{iV}^{fit} \times \left(\phi\zeta_C^2/r_t a\right) = \left(I_{iV}^{fit}\right)\left(S_{iV}^{fit}\right)^2/r_t s_t^2 ab^2\phi^2 \equiv \Lambda_t \left(I_{iV}^{fit}\right) \left(S_{iV}^{fit}\right)^2, \quad (38)$$

where Λ_t is the particular value $\Lambda(f_t,\phi)$ of the *area-data extraction parameter* $\Lambda(f,\phi)$ defined by

$$\Lambda(f,\phi) = 1/\left(r_{2012} s^2 ab^2\phi^2\right). \quad (39)$$

$\Lambda(f,\phi)$ is a new SN barrier function. Λ is broadly similar in purpose to the function Γ found in older literature (e.g., [42]); however, Λ is defined in a significantly different way and supercedes Γ.

The value chosen for f_t needs to correspond to a point near (or slightly to the left of) the middle of the range of $(1/V_m)$ values used to make the FN plot (or the part of it being analyzed). For an emitter with $\phi = 4.50$ eV, the value of f_t might typically be around 0.30 [43]. Corresponding values of other parameters in Eq. 38 are $s_t \approx 0.95$, $r_t \approx 120$, and $\Lambda_t \approx 6$ nm^2/A. This value of s_t can be used in a first application of Eq. 37.

Values derived as above for ζ_C are not particularly sensitive to the value chosen for f_t, but values of A_f can be. Better estimates of ζ_C and A_f can be obtained by using the extracted value of ζ_C and the observed midrange value of V_m^{-1} to obtain a revised estimate of f_t, via Eq. 35. Revised estimates of r_t, s_t, and Λ_t can then be obtained from relevant formulae given above, and revised estimates of ζ_C and A_f made by reapplying Eqs. 37 and 38.

If needed, the value of the characteristic FEF $\beta_{F,C}$ can be obtained from ζ_C via Eq. 13, and the value of the formal area efficiency α_f can be obtained from A_f via Eq. 6. Arguably, this is the best method of obtaining values of $\beta_{F,C}$ and α_f, since it is based on FN plots made using raw experimental data. Nevertheless, many experimental papers show FN plots ostensibly of the type $[\ln\{J_M/E_M{}^2\}$ vs $1/E_M]$ (often actually of the type $[\ln\{J_M/E_A{}^2\}$ vs $1/E_A]$). Obviously, this alternative approach requires pre-conversion of the raw experimental data. In orthodox emission situations (but not in nonorthodox situations), valid values of $\beta_{F,C}$ and α_F can be obtained from

$$\beta_{F,C} = -S_A^{fit}/s_t b\phi^{3/2} \quad (40)$$

$$\alpha_f = \Lambda_t \left(I_A^{fit}\right) \left(S_A^{fit}\right)^2, \quad (41)$$

where $S_A{}^{fit}$ and $I_A{}^{fit}$ are the slope and intercept of the straight line fitted to such a plot.

In the elementary data-analysis method, as applied to orthodox emission situations, the parameters s_t and r_r are replaced by unity. This gives rises to error of about 5 % in the estimation of ζ_C or $\beta_{F,C}$, but would give rise to error by a factor of order 100 in the estimation of A_f or α_f. In practice, due in part to the widespread habit in nanomaterial literature of writing down a defective elementary FN-type formula [for $J_M(E_M)$] that erroneously omits the factor α_f, estimates of emission area and formal area efficiency are not normally made in nanomaterial work. It would be useful and entirely possible to do so in orthodox emission situations, but – to avoid unnecessary error – it would be desirable to use the orthodox data-analysis formulae described above, rather than elementary data-analysis formulae.

A Quantitative Test for Lack of Orthodoxy

The formulae in the preceding Section give valid results only in emission situations that are "nearly orthodox." In nonorthodox situations the formulae generate spurious results. In particular, they generate spuriously large values of the characteristic FEF $\beta_{V,C}$ [25]. In [25], Forbes has recently described an experimental test for detecting lack of orthodoxy. This test can be applied to any type of FN plot (even when numerical values, but no units, are shown on the axes). The test involves extracting values of scaled barrier field f from an experimental FN plot, using the formula

$$f^{extr} = -s_t\eta/\left[S_X^{fit} \cdot \left(X^{-1}\right)^{expt}\right], \quad (42)$$

where X is the independent variable used in the plot, $(X^{-1})^{expt}$ is the experimental horizontal-axis value of interest, and S_X^{fit} is the fitted (negative) slope. One can take $s_t \approx 0.95$, as above. From this

formula, f-values can be found that correspond to the ends of the range of X-values measured (or to the ends of the range over which plot linearity is observed). This provides an apparent experimental range of f-values. The test assesses whether this extracted range is reasonable.

For any given value of the local work function, four indicative boundary f-values $\{f_{lb} < f_{low} < f_{up} < f_{ub}\}$ can be specified as described in [25]. All these values depend on the assumed value of local work function. For illustration, for $\phi = 4.50$ eV, the set of values is {0.10, 0.15, 0.45, 0.75}; f-value sets for other work function values are given in [25]. The test is an engineering triage test. (1) If the whole extracted range lies between f_{low} and f_{up}, then this range is "apparently reasonable." (2) If any part of the extracted range is below f_{lb} or above f_{ub}, then this range is "clearly unreasonable." (3) If part of the extracted range is outside the "apparently reasonable" range, but not in either of the "clearly unreasonable" ranges, then further investigation is required.

If the extracted range is "apparently reasonable," then the emission situation is orthodox or "nearly orthodox," and use of the formulae above is valid. If the extracted range is "clearly unreasonable," then it is almost certain that the formulae above will generate spurious results. If "further investigation is required," then there is a significant possibility the formulae above will generate unreliable results.

In work reported in [25], a small sample of 19 published FN plots, taken from nontraditional emitters of various kinds, were tested in this way. Approximately 40 % of these FN plots yielded "clearly unreasonable" f-value ranges, indicating that the related published FEF values are spuriously high. This suggests that there are a large number of unreliable FEF values in the literature and that often we do not know which are reliable and which are not.

If an FN plot exhibits two emission regimes with a break of slope between them, then this is a clear indication that the higher-current regime is nonorthodox. The lower-current regime may or may not be orthodox. Reference [25] also identified instances where a nearly linear (slightly curved) single-regime plot turned out to be nonorthodox.

It is arguable that the orthodoxy test should be applied to all published nanomaterial FN plots before any LAFE characterization parameter derived from such a plot is used for any other purpose, and that the orthodoxy test should be applied as part of all new FE investigations of nontraditional materials.

Clearly, a need exists to develop alternative methods of analyzing current-voltage data taken in nonorthodox FE situations. Methods currently being explored [44] include (a) development of an engineering ansatz that would enable a spurious FEF value to be modified toward the true FEF value; (b) attempts to derive V_e values from V_m values, by simulations that involve modeling the series resistance and comparing simulated $i_m(V_m)$-type FN plots with experimental plots; and (c) attempts to derive slope and intercept correction factors for nonplanar geometries where the emitter has small apex radius and to explore quantum-confinement effects in greater detail. There are many possible causes of non-orthodoxy and many conceivable methods of dealing with them. Investigation of how (where practicable) to deal with these problems is still in its early stages, and detailed comment here would be premature. For further information about FN-plot analysis, see Ref. [44].

Field Electron Emission from Planar Emitters

As indicated in the introduction, anisotropic one-dimensional (1D) and two-dimensional (2D) nanostructures have been recognized as potential field emitters for practical applications in various vacuum microelectronic devices, such as flat panel field emission displays (FEDs), X-ray tubes, and microwave tubes. Generally, for field emission research and development (R&D) studies, nanostructures are deposited onto a conducting substrate in a thin-film form. Such an assembly of nanostructures, called earlier a "large-area field emitter" (LAFE), is also known as a "multi-tip emitter." LAFEs are effective

emitters primarily because the high aspect ratios of the individual emitting structures lead to high field enhancement factors.

In principle, LAFEs can be deposited onto rounded surfaces, but it is more common to deposit them onto planar substrates, in which form they are also known as *planar emitters*. FE R&D studies of the 1D and 2D nanostructures are usually performed in a planar diode configuration (also known as close-proximity geometry). In this configuration, the emitter (cathode) and the anode are held parallel in close proximity, with the interelectrode separation ranging from a few tens of micrometers to a few millimeters. The anode can be a semitransparent cathodoluminescent phosphor screen or a metal disk, typically about 20 mm in diameter. As indicated earlier, use of a phosphor screen facilitates observation and recording of the FE image.

Differences in practical aspects of FE studies of a planar emitter and a traditional single-tip field emitter (STFE), i.e., a single microtip of well-defined shape and size, may be summarized as follows:

1. In the conventional configuration employed for STFE studies, the interelectrode separation is a few centimeters or more, whereas in the planar diode configuration, it ranges from a few tens of micrometers to a few millimeters.
2. In STFE-type geometry, the anode is often a partly hemispherical semitransparent cathodoluminescent phosphor screen (radius about 5–6 cm), whereas in the planar diode configuration, the anode is planar and can be a semitransparent cathodoluminescent screen or a metal disk 5–20 cm in diameter.
3. In STFE-type geometry, the electrostatic field E_a at the STFE apex is usually estimated via Gomer's formula, which (as noted earlier) can be written in the various forms

$$E_a = V_e/k_a r_a = V_e/\zeta_a = \beta_{V,a} V_e, \qquad (43)$$

where V_e is the emission voltage, r_a is the STFE apex radius, k_a is usually called a *shape factor* or a "field factor," ζ_a is the apex LCL, and $\beta_{V,a}$ is the apex VCF. In the planar diode configuration, as discussed earlier, two types of electrostatic field are needed, namely, the macroscopic field E_M as defined by Eq. 11 and a characteristic local barrier field E_C. If the characteristic point is chosen to be at the apex of one of the individual emitters, this local field can alternatively be denoted by E_a and is related to E_M by the apex version of Eq. 12, namely,

$$E_a = \beta_{F,a} E_M. \qquad (44)$$

A habit in LAFE literature in general, and in nanomaterial literature in particular, is to denote both β_F and β_V by the unadorned symbol β, which is potentially confusing. Further, though most authors call our β_F a "field enhancement factor" (FEF), some authors apply this name to our β_V. To resolve the confusion, statements have been made such as "the field enhancement factor, β, has the units of cm^{-1} (or nm^{-1}) in the conventional geometry, whereas it is dimensionless in the planar diode configuration." However, it is unsatisfactory to have two different physical quantities, with different physical dimensions, called by the same name and allocated the same symbol, and alternative approaches have been suggested.

One possibility is to distinguish the two symbols by a suffix and give a different name to β_V, which is the approach adopted in this entry, but this is not entirely satisfactory. Another approach was the attempt (e.g., [29]) to have β_F-type FEFs denoted by the symbol γ. This has had some limited success, but use of the symbol β to denote FEFs remains deeply entrenched. Yet another approach is the proposal made earlier that the use of VCFs (i.e., β_V-type parameters) be abandoned and be replaced by the use of local conversion lengths (LCLs) given by $\zeta_L = 1/\beta_V$.

4. In STFE-type geometry, one can easily achieve magnifications of around 10^5–10^6 and resolutions of around 3–4 nm (or sometimes better), whereas the planar configuration offers poor magnification and

F

resolution. In fact, for the simple planar configuration, precise magnification and resolution cannot be obtained.

5. In STFE-type geometry, the STFE apex field can easily be (approximately) estimated from Gomer's formula (Eq. 42 above). With a planar emitter (i.e., a LAFE), it is difficult to make reliable theoretical estimates of electrostatic field at the apex of any individual emitter, because the geometry is much more complicated.
6. With a planar emitter, a large number of individual emission sites contribute to the total emission current. When an emitter has a high surface density of emission sites, it may be particularly suitable as a high-current-density electron source for practical applications in vacuum microelectronic devices such as flat panel displays and X-ray tubes.
7. Furthermore, for planar nanostructured emitters fabricated from semiconductors with a narrow bandgap, noticeable enhancement in the emission current density has been observed as a result of illumination by visible light. Such effects may contribute toward the development of a new generation of nano-optoelectronic devices.
8. As regards emission current stability, planar cathodes are somewhat inferior to STFE-type cathodes, for three reasons. (i) The surface area available for capture of residual vacuum gases is much larger, and adsorbed molecules may migrate to become adsorbed at emission sites, subsequently becoming desorbed. Hence, there is more fluctuation in the emission current, resulting from more-frequent time-dependent changes in local work function induced by the adsorbed molecules. (ii) Extinction and generation of emission sites may result from bombardment by ions created by electron-impact ionization of residual gas molecules in space above the array, particularly under "industrial vacuum conditions." (iii) Provision for thorough "degasing" or "in situ cleaning" of the planar cathodes involves instrumental complexity, in contrast to the situation for STFE cathodes, which can easily be heated in situ by mounting them on a heating loop.
9. For STFEs, basic theory uses the well-defined concepts of emission current i_e and local emission current density J_L, and no confusion arises. With planar emitters, the concept of macroscopic current density J_M is also used. As noted earlier, J_M can be less than the LAFE characteristic local ECD J_C by many orders of magnitude. However, some nanomaterial papers fail to distinguish properly between the two types of current density and/or fail to identify which type of current density their symbol "J" represents. J_M is related in some way to the surface density of emission sites, but exact estimation of the number of emission sites (which involves identifying which nanostructures are emitting the electrons) is a difficult task.
10. For STFEs, experimental FN plots are always based directly on the raw experimental data and hence are of the type $[\ln\{i_m/V_m^2\}\ vs\ 1/V_m^2]$. With planar emitters, it has become conventional to pre-convert the raw experimental data and make plots *ostensibly* of the type $[\ln\{J_M/E_M^2\}\ vs\ 1/E_M]$ (though it is argued above that in this case too it would be better to use plots of type $[\ln\{i_m/V_m^2\}\ vs\ 1/V_m]$), based on the raw experimental data.

The term "ostensibly" is used above because the pre-conversion process of dividing the measured voltage V_m by the plate separation d_{sep} does not necessarily generate a value of true macroscopic field E_M (which is defined as V_e/d_{sep}). Rather it generates the mathematical parameter E_A that can be called an *apparent macroscopic field.* If voltage drop occurs in the measuring circuit, then E_A is not equal to the true macroscopic field E_M present in the LAFE. Further, in this case, the quantity extracted from the slope of an FN plot of type $[\ln\{J_M/E_A^2\}\ vs\ 1/E_A]$ is not a true FEF, but a *pseudo-FEF* that may differ significantly in value from the true macroscopic (electrostatic) FEF associated with the LAFE under analysis.

The mathematical analysis of FN plots of type $[\ln\{J_M/E_A^2\}\ \text{vs}\ 1/E_A]$ is complicated, confusing

and incompletely understood, and involves unsolved problems. Both conceptually and mathematically, it is easier to deal with plots of the type $[\ln\{i_m/V_m^2\}$ vs $1/V_m]$. Hence, if it is discovered (by applying the orthodoxy test to a plot of type $[\ln\{J_M/E_A^2\}$ vs $1/E_A]$) that an emission situation is nonorthodox, subsequent analysis (if possible at all) may be easier if the original raw data is replotted in the form $[\ln\{i_m/V_m^2\}$ vs $1/V_m]$.

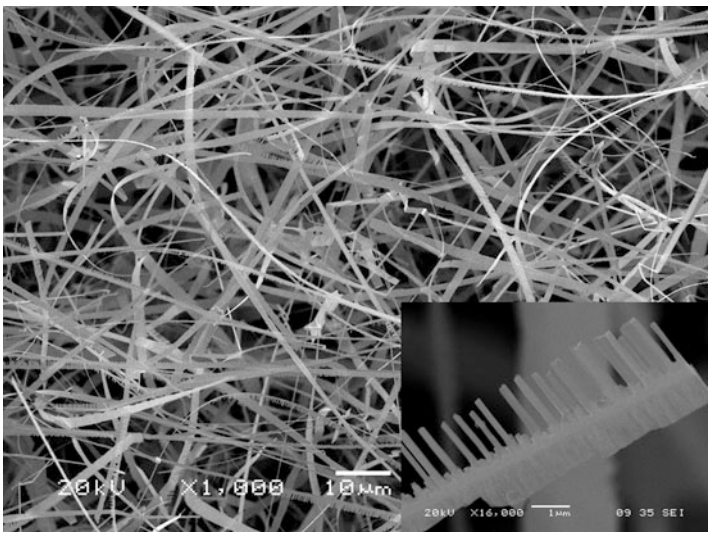
It is arguable that community use of (what are in reality) plots of type $[\ln\{J_M/E_A^2\}$ vs $1/E_A]$ has made it significantly more difficult for graduate students to recognize that arguments based on series resistance and Ohm's law may be the true explanation of puzzling experimental effects. This said, there is a good case for presenting LAFE results in the form of a direct plot of $J_M(E_A)$ since this makes it easier to compare the behavior of different nanostructures; similarly, there is a case for using plots of $J_M(V_m)$ to compare the performance of different engineered LAFEs. The best approach may be to use different forms of plot for different purposes.

Field Electron Emission from Cadmium Sulfide: A Case Study

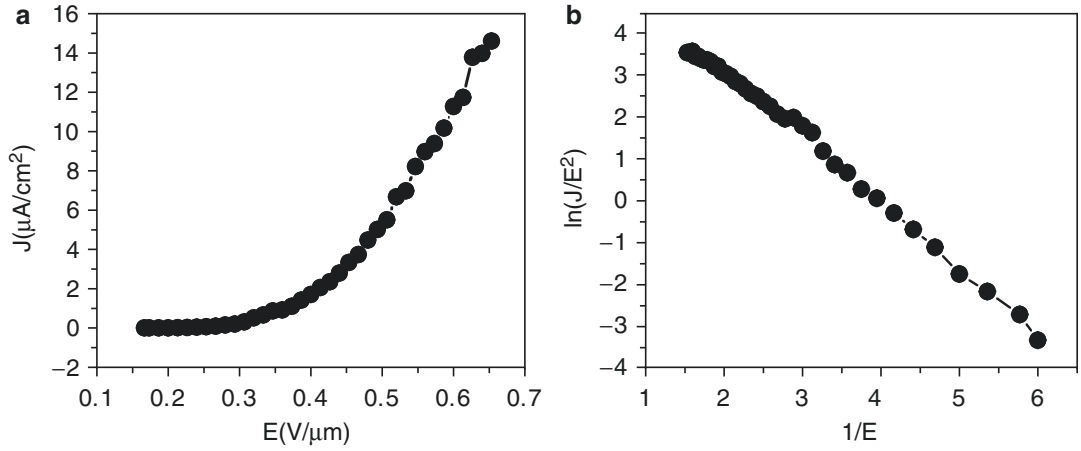
Recently, MAM and DSJ have investigated the field emission properties of various nanoforms of cadmium sulfide (CdS), a II–VI semiconductor that is also a photoconducting material with controllable morphology. Here we discuss the case of CdS nanocombs grown on a silicon substrate. The substrate is chosen so that the application could be integrated with silicon technology. Details of the synthesis of nanocombs, based on thermal evaporation of the CdS, are described elsewhere [45].

Figure 4 shows the SEM images of the as-synthesized CdS nanocombs grown on a silicon substrate. The structure of the nanocombs was analyzed using X-ray diffraction and was found to be the pure wurtzite CdS phase. The cathode as described above was mounted in a planar diode arrangement with a spacing of 1 mm between the anode and cathode. The whole assembly was placed in a vacuum chamber attaining a pressure near 10^{-8} mbar, and measurements were carried out at this pressure.

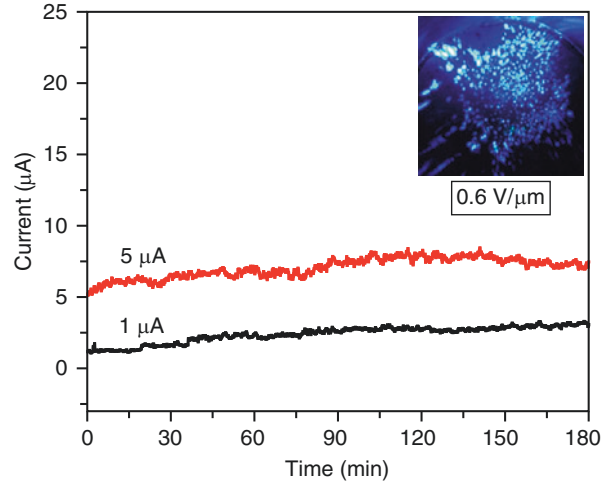
Current-voltage (i_m-V_m) and current-time (i_m-t) characteristics were measured, and the i_m-V_m data were converted to data relating macroscopic current density J_M to apparent macroscopic field E_A, using the parameter values $A_M = 1$ cm^2 and $d_{sep} = 1$ mm. Figure 5a depicts the $J_M(E_A)$ characteristics, and Fig. 5b the corresponding FN plot. (For simplicity, the symbols E and J are used in these figures to denote macroscopic quantities.) The FN plot is linear in the low-field region and



Field Electron Emission from Nanomaterials, Fig. 4 SEM image of as-synthesized CdS nanocombs grown on a Si substrate. The inset shows a closeup view of a nanocomb (Reproduced from Ref. [45]; copyright ©2011, Royal Society of Chemistry)



Field Electron Emission from Nanomaterials, Fig. 5 (**a**) Plot of macroscopic current density (here, J) against applied macroscopic field (here E) for the CdS nanocombs and (**b**) the corresponding Fowler-Nordheim (FN) plot (Reproduced from Ref. [45], copyright ©2011, Royal Society of Chemistry). In this plot, E (rather than E_M or E_A) denotes macroscopic field and J (rather than J_M) denotes macroscopic current density. In the FN plot, the units of E and J are, respectively, V/μm and μA/cm^2

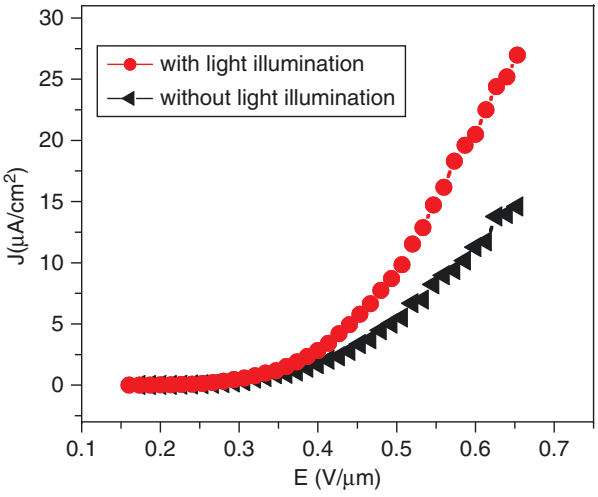


Field Electron Emission from Nanomaterials, Fig. 6 Plot of emission current versus time for the CdS nanocombs. The inset shows a FE image recorded at an apparent macroscopic field of 0.6 V/μm (Reproduced from Ref. [45]; copyright ©2011, Royal Society of Chemistry)

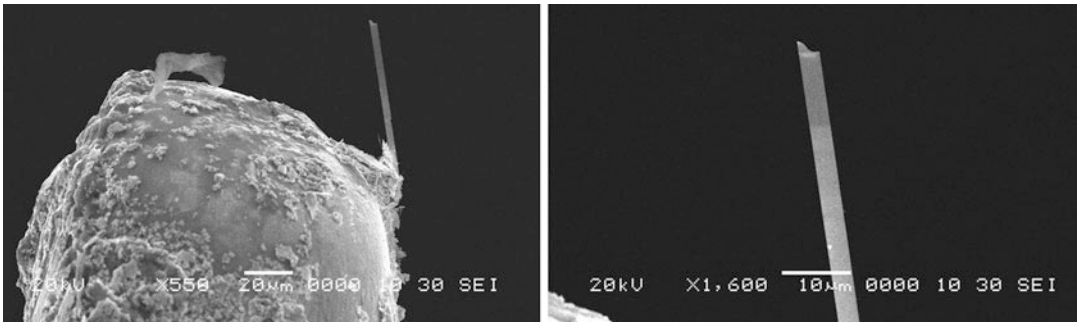
shows a slight saturation tendency in the high-field region. The apparent macroscopic onset field, defined as the field required to generate an emission current density of $\sim 10^{-3}$ A/m^2, is ~0.26 V/μm.

Figure 6 shows the i_m-t plots recorded to investigate the current stability at the base pressure of around 10^{-8} mbar. The average current at the preset levels of 1 μA and 5 μA is found to remain nearly constant. The inset of Fig. 6 shows the FEM image recorded at $E_A = 0.6$ V/μm. The FEM pattern helps when studying the emission uniformity over the emitting area, including emission spot flickering.

As noted above, CdS is a photoconducting semiconductor. Photosensitive FE studies were carried out by exposing the cathode to a constant-intensity visible light source. The cathode was illuminated with a 1300 W/m^2 halogen lamp, having intensity at the cathode measured as 80 W/m^2. Figure 7 shows the $J_M(E_A)$ plots with and without illumination. It can be seen clearly



Field Electron Emission from Nanomaterials, Fig. 7 Plot of macroscopic current density (here, J) against apparent macroscopic field (here, E) for the CdS nanocombs, with and without light illumination (Reproduced from Ref. [45]; copyright ©2011, Royal Society of Chemistry)

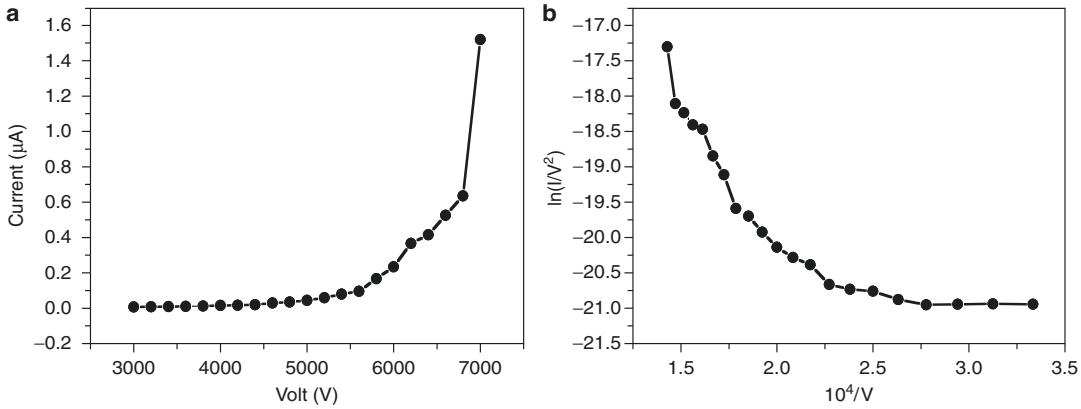


Field Electron Emission from Nanomaterials, Fig. 8 SEM images of the single nanowire attached to the tungsten needle

that the illumination enhances the CFE current density. The enhanced and unenhanced values (with and without illumination, respectively) of the apparent macroscopic onset fields (fields for drawing the current density 10^{-3} A/m^2) are 0.24 V/μm and 0.26 V/μm, respectively. The observed photosensitivity can be attributed to the creation, by photoexcitation, of electrons in excited states; in such states, electrons find it easier to propagate within the material and tunnel out of it.

An approximate value of the apparent characteristic field enhancement factor $\beta_{F,C}^{app}$ for the CdS LAFE can be calculated directly from the slope S_A^{fit} of the experimental FN plot (Fig. 5b), using Eq. 40, with $s_t = 1$ and the effective work function ϕ for field-emitting CdS taken here as 4.2 eV. This yields the value $\beta_{F,C}^{app} \approx 10^4$. This value looks implausibly large. Application of the orthodoxy test described above shows that Fig. 4b corresponds to the f-value range 0.49–1.4: this range is "clearly unreasonable," and we may conclude that the emission situation is nonorthodox and that the FEF value 10^4 is spuriously large. The exact reason for the non-orthodoxy of the CdS nanocomb is not currently known, but series resistance in the measuring circuit seems the most plausible explanation. Among the FN plots tested in Ref. [25] was one from a "CdS nanotip" [46]; this plot also was nonorthodox.

We, too, have studied the FE behavior of a single CdS emitter tip. A single pod from a CdS nanocomb was isolated and carefully attached it to a tungsten needle, using silver paste. Figure 8 shows the SEM image of a pod mounted on a tungsten tip. Figure 9a shows the $i_m(V_m)$



Field Electron Emission from Nanomaterials, Fig. 9 (**a**) Measured current-voltage characteristic for the single CdS nanowire shown in Fig. 8; (**b**) corresponding FN plot (Note the anomalous behavior of this FN plot)

characteristics for the emitter and Fig. 9b the corresponding FN plot. It is of interest that this single-emitter FN plot is markedly nonlinear, as compared with the LAFE behavior recorded in Fig. 4b. If the high-voltage and low-voltage parts of the FN plot are approximated by straight lines, then both these parts are found (via the test) to be nonorthodox. Detailed reasons for this anomalous behavior are not understood, but could, for example, relate to current-dependent resistance in the electron flow through the silver paint. The result illustrates the complications that can arise in nanomaterial studies.

Overall, the study of CFE from nanomaterials provides both exciting technological opportunities and interesting scientific challenges.

Appendix

Technical Conventions

Equation systems. The majority of nanomaterial papers now use the "modern" International System of Quantities and equations introduced in the 1970s [47]. This underlies the SI system of units and has recently been named the "International System of Quantities" (ISQ) [48]. However, some nanomaterial papers still use the old, hybrid, pre-SI system of equations and quantities used to describe field emission in the 1960s and earlier. In the old system the factor $4\pi\varepsilon_0$ does not appear, and the symbol e has context-dependent definitions and values. For clarity and in conformity with modern practice, this entry adheres strictly to the ISQ as regards quantities and equations.

Units and emission constants. However, where relevant, we use the ISQ-compatible system of atomic-level units that is customary in field emission. This system makes calculations simple when energies such as work functions are expressed in electron volts (eV) and local fields are in V/nm. Table 1 shows definitions and values of special universal constants used in this article, expressed in these units.

Nomenclature conventions. Newer nanomaterial literature often uses nomenclature conventions different from the older "mainstream" literature, but many theoretical papers use the older conventions. In particular, different meanings are often attached to the term "field" (and the symbol E or F) and to the term "current density" (and the symbol J). An indirect consequence, as noted in the main text, is that nanomaterial papers sometimes contain defective ECD equations in which the factor α_f (the "formal area efficiency") is missing. This difficulty can be avoided, as here, by careful definition or examination of what "current density" and "field" (and associated symbols) mean in a given paper and by the use of more specific terms and symbols where appropriate.

Although practice is not consistent in the literature, the symbols E and β, and the term β-value, tend to be used differently in mainstream and

Field Electron Emission from Nanomaterials, Table 2 Nomenclature conventions. In all cases the fields concerned are the negatives/magnitudes of the corresponding fields as defined in classical electrostatics. V_e denotes the emission voltage and V_m the measured voltage. "nwu" means "no well-defined usage"; "n/a" means "not applicable"

Quantity	Mainstream usage	Nanomaterial usage	Convention here
Electron energy	E	nwu	n/a
Local barrier field	F or F_L	nwu	E_L
Characteristic local barrier field	F or F_C	nwu	E_C
True macroscopic field ($\equiv V_e/d_{sep}$)	F_M	E	E_M
Apparent macroscopic field ($\equiv V_m/d_{sep}$)	nwu	E	E_A
Local conversion length ($\equiv V_e/E_L = 1/\beta_V$) short name:	n/a	n/a	ζ_L LCL
Voltage-to-barrier-field conversion factor ($\equiv E_L/V_e$) short name:	β β-value	nwu	β_V VCF
Field enhancement factor ($\equiv E_L/E_M$) short name:	γ nwu	β β-value	β_F FEF

nanomaterial papers. These usages, and the compromise convention adopted here, are shown in Table 2.

In principle, the symbol F is probably the better choice for the negative or magnitude of classical electrostatic field, as it allows E to be used for energy (or for classical electrostatic field); E is used here because this is the current nanomaterial community practice. The term "β value" is widely used in the literature, but is not used here because it is ambiguous. Rather, the acronym-based short name "FEF" or "VCF" is used, according to context. As indicated in the text, in order to avoid continuing confusion over the meaning of the symbol β, we suggest that in future the parameter "local conversion length" (ζ_L, $=1/\beta_V$) should be used instead of the parameter "voltage-to-barrier-field conversion factor" (β_V).

Cross-References

► Carbon Nanotubes

References

1. Lilienfeld, J.E.: The auto-electronic discharge and its application to the construction of a new form of X-ray tube. Am. J. Roentgenol **9**, 172 (1922)
2. Fowler, R.H., Nordheim, L.: Electron emission in intense electric fields. Proc. R. Soc. Lond. A **119**, 173 (1928)
3. Stern, T.E., Gossling, B.S., Fowler, R.H.: Further studies in the emission of electrons from cold metals. Proc. R. Soc. Lond. A **124**, 699 (1929)
4. Forbes, R.G., Deane, J.H.B.: Transmission coefficients for the exact triangular barrier: an exact general analytical theory that can replace Fowler and Nordheim's 1928 theory. Proc. R. Soc. Lond. A **467**, 2927 (2011)
5. Forbes, R.G.: Introduction to field electron emission and its theory. Tutorial lecture, 26th international vacuum nanoelectronics conference, Roanoke, July 2013. (Presentation available from American Vacuum Society, in .pdf form, via web search)
6. Müller, E.W.: Elektronenmikroscopische Beobactungen von FeldKazthoden. Z. Phys. **106**, 541 (1937)
7. Gomer, R.: Field Emission and Field Ionization. Harvard University Press, Cambridge, MA (1961); reprinted by AIP, New York (1992)
8. Gadzuk, J.W., Plummer, E.W.: Field emission energy distributions. Rev. Mod. Phys. **45**, 487 (1973)
9. Modinos, A.: Field, Thermionic and Secondary Electron Emission Spectroscopy. Plenum, New York (1984); reprinted by Springer, New York (2013)
10. Van Oostrom, A.G.J.: Validity of the Fowler-Nordheim model for field electron emission. Philips Res. Rept. Suppl. **1** (1966)
11. Swanson, L.W., Schwind, G.A.: Chapter 1, Review of the ZrO/W cathode. In: Orloff, J. (ed.) Handbook of Charged Particle Optics, 2nd edn. CRC Press, Boca Raton (2009)
12. Swanson, L.W., Schwind, G.A.: A review of the cold field electron cathode. Adv. Imag. Electron. Phys. **159**, 63 (2009)
13. Spindt, C., Holland, C.E., Rosengreen, A., Brodie, I.: Field-emitter arrays for vacuum microelectronics. IEEE Trans. Electron Devices **ED-38**, 2355 (1991)
14. Zhu, W. (ed.): Vacuum Microelectronics. Wiley, New York (2001)

15. Spindt, C.A., Brodie, I., Humphrey, L., Westerberg, E. R.: Physical properties of thin-field field emission cathodes with molybdenum cones. J. Appl. Phys. **47**, 5248 (1976)
16. Forbes, R.G.: Low-macroscopic-field electron emission from carbon films and other electrically nanostructured heterogeneous materials: hypotheses about emission mechanism. Solid State Electron. **45**, 779 (2001)
17. Silva, S.R.P., Carey, J.D., Khan, R.U.A., Gertner, E.G., Anguita, J.V.: Chapter 9, Amorphous carbon thin films. In: Halwa, H.S. (ed.) Handbook of Thin Film Materials. Semiconductor and Superconductor Thin Films, vol. 4. Academic, San Diego (2002)
18. Robertson, J.: Diamond-like amorphous carbon. J. Mater. Sci. Eng. R **37**, 129 (2002)
19. Xu, N.S., Huq, S.E.: Novel cold cathode materials and applications. Mater. Sci. Eng. R **48**, 47 (2005)
20. Fang, X., Bando, Y., Gautam, U.K., Ye, C., Golberg, D.: Inorganic semiconductor nanostructures and their field-emission applications. J. Mater. Chem. **18**, 509 (2008)
21. Milne, W.I. and 15 co-authors.: Carbon nanotubes. E nano-newsletter **13,** 5 (2008). http://www.phantomsnet.net/Foundation/Enano_newsletter13.php. Accessed 20 Feb 2015
22. Saito, Y. (ed.): Carbon Nanotubes and Related Field Emitters. Wiley-VCH, Weinheim (2010)
23. Bocharov, G.S., Eletskii, A.V.: Theory of carbon nanotube (CNT)-based electron field emitters. Nanomaterials **3**, 393 (2013)
24. Fursey, G.N.: Field Emission in Vacuum Microelectronics. Kluwer, New York (2005)
25. Forbes, R.G.: Development of a simple quantitative test for lack of field emission orthodoxy. Proc. R. Soc. Lond. A **469**, 20130271 (2013)
26. Zhang, X.-G., Pantelides, S.T.: Screening in nanowires and nanocontacts: field emission, adhesion force, and contact resistance. Nano Lett. **9**, 4306 (2009)
27. Edgcombe, C.J., Valdrè, U.: The enhancement factor and the characterization of amorphous carbon field emitters. Solid State Electron. **45**, 857 (2001)
28. Kokkorakis, G.C., Modinos, A., Xanthakis, J.J.: Local electric field at the emitting surface of a carbon nanotube. J. Appl. Phys. **91**, 4580 (2002)
29. Forbes, R.G., Edgcombe, C.J., Valdrè, U.: Some comments on models for field enhancement. Ultramicroscopy **95**, 57 (2003)
30. Djurabekova, F., Parviainen, S., Pohjonen, A., Norlund, K.: Atomistic modelling of metal surfaces under electric fields: direct coupling of electric fields to a molecular dynamics algorithm. Phys. Rev. E **83**, 026704 (2011)
31. Nilsson, L.-O., Groening, O., Groening, P., Kuettel, O., Schlapbach, L.: Characterization of thin film electron emitters by scanning anode field emission microscopy. J. Appl. Phys. **90**, 768 (2001)
32. Forbes, R.G.: On the need for a tunneling pre-factor in Fowler-Nordheim tunneling theory. J. Appl. Phys. **103**, 114911 (2008)
33. Forbes, R.G., Deane, J.H.B.: Reformulation of the standard theory of Fowler-Nordheim tunneling and cold field electron emission. Proc. R. Soc. Lond. A **463**, 2907 (2007)
34. Deane, J.H.B., Forbes, R.G.: The formal derivation of an exact series expansion for the principal Schottky-Nordheim barrier function v, using the Gauss hypergeometric differential equation. J. Phys. A: Math. Theor. **41**, 395301 (2008)
35. Burgess, R.E., Kroemer, H., Houston, J.M.: Corrected values of Fowler-Nordheim emission functions $v(y)$ and $s(y)$. Phys. Rev. **90**, 515 (1953)
36. Murphy, R.H., Good, E.L.: Thermionic emission, field emission, and the transition region. Phys. Rev. **102**, 1464 (1656)
37. Forbes, R.G., Deane, J.H.B.: Comparison of approximations for the principal Schottky-Nordheim barrier function $v(f)$, and comments on Fowler-Nordheim plots. J. Vac. Sci. Technol. B **28**, 2CA33 (2010)
38. Forbes, R.G.: Simple good approximations for the special elliptic functions in standard Fowler-Nordheim tunnelling theory for a Schottky-Nordheim barrier. Appl. Phys. Lett. **89**, 113122 (2006)
39. Mayer, A.: A comparative study of electron transmission through one-dimensional barriers relevant to field emission problems. J. Phys. Condens. Matter **22**, 1 (2010)
40. Forbes, R.G., Fischer, A., Mousa, M.S.: Improved approach to Fowler-Nordheim plot analysis. J. Vac. Sci. Technol. B **31**, 02B103 (2012)
41. Forbes, R.G.: Field electron and ion emission: basic formulae and constants. Chapter submitted for publication. In: Landolt-Börnstein Tables III/24 (Physics of Solid Surfaces). Springer, Berlin (2015). (updated edition to be published in 2016)
42. Forbes, R.G.: Field emission: New theory for the derivation of emission area from a Fowler-Nordheim plot. J. Vac. Sci. Technol. B **17**, 526 (1999)
43. Forbes, R.G.: Description of field emission current/voltage characteristics in terms of scaled barrier field values (f-values). J. Vac. Sci. Technol. B **26**, 209 (2008)
44. Forbes, R.G., Deane, J.H.B., Fischer, A, Mousa, M.S.: Fowler-Nordheim plot analysis: a progress report. Jordan J. Phys. **8**, 127 (2015). http://journals.yu.edu.jo/jjp/Vol8No3Contents2015.html
45. Chavan, P.G., Badadhe, S.S., Mulla, I.S., More, M.A., Joag, D.S.: Synthesis of single crystalline CdS nanocombs and their application in photo-sensitive field emission switches. Nanoscale **3**, 1078 (2011)
46. Zhai, T.Y., Fang, X.S., Bando, Y., Dierre, B., Liu, B.D., Zeng, H.B., Xu, X.J., Huang, Y., Yuan, X.L., Sekiguchi, T., Golberg, D.: Characterisation, cathodiluminescence, and field emission properties of morphology-tunable CdS micro/nanostructures. Adv. Funct. Mater. **19**, 2423 (2009)

47. Bureau International des Poids et Mesures (BIPM): The International System of Units, 8th Edn. BIPM, Sèvres, France (2006). http://www.bipm.org/utils/common/pdf/si_brochure_8_en.pdf. Accessed 20 Feb 2015
48. International Organization for Standardization (ISO): International Standard ISO/IEC 80000-1-2009: Quantities and units-Part 1: General. ISO, Geneva (2009)

Field-Driven Particle Assembly

▶ AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation

Filling of Carbon Nanotubes

▶ Functionalization of Carbon Nanotubes

Finite Element Methods for Computational Nano-optics

Sven Burger, Jan Pomplun and Frank Schmidt
Zuse Institute Berlin, Berlin, Germany

Synonyms

Finite-element analysis (FEA); Rigorous Maxwell solver

Definition

The finite-element method (FEM) is a numerical method for solving partial differential equations (PDEs). In the field of nano-optical devices, finite-element methods are mainly used for simulations of optical effects and optical device properties. The relevant models in this case are Maxwell's equations in various formulations. Main advantages of finite-element methods in this field are the exact representation of complex geometrical features present in real-world applications, the superior convergence properties which allow for very accurate results in relatively short computation times, and the stability and general applicability of the method to different simulation settings.

Mathematical Background

General

The finite element method offers a mathematically rigorous and elegant formalism for discretization of partial differential equations [6]. Instead of approximating the PDE itself, the key idea of FEM is the construction of an accurate finite-dimensional subspace of the original, infinite-dimensional solution space. The problem is then solved on this subspace, the finite-element space. A key concept for FEM is the so-called weak formulation of a PDE.

Weak Formulation of Maxwell's Equations

Physical laws are usually stated in a so-called strong form. For example Maxwell's equations in frequency domain read:

$$\begin{aligned} \mathbf{div}\,\mathbf{D} &= \rho, \\ \mathbf{curl}\,\mathbf{H} + i\omega\mathbf{D} &= \mathbf{J}, \\ \mathbf{curl}\,\mathbf{E} - i\omega\mathbf{B} &= 0, \\ \mathrm{div}\,\mathbf{B} &= 0. \end{aligned}$$

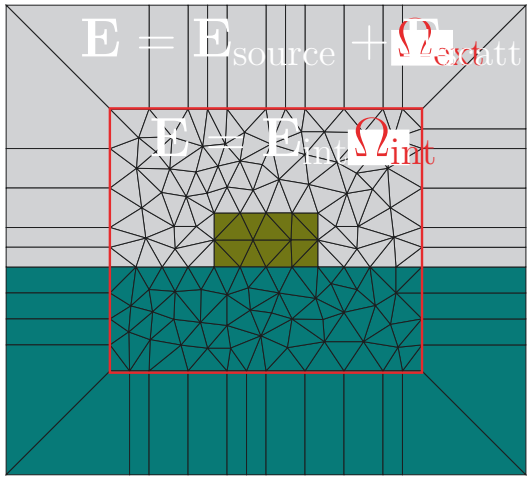
This set of equations is not closed. The electric and magnetic field $\mathbf{E}$ and $\mathbf{H}$ are connected to the electric displacement $\mathbf{D}$ and magnetic flux density $\mathbf{B}$ via constitutive relationships:

$$\begin{aligned} \mathbf{D} &= \varepsilon\mathbf{E} \\ \mathbf{B} &= \mu\mathbf{H}, \end{aligned}$$

where in the linear case ε and μ are the permittivity and permeability tensor. From above relationships a single equation for the electric field can be derived:

$$\mathbf{curl}\,\mu^{-1}\mathbf{curl}\,\mathbf{E} - \omega^2\varepsilon\mathbf{E} = i\omega\mathbf{J}. \qquad (1)$$

The electric current $\mathbf{J}$ is source of an electromagnetic field. In applications other types of sources



Finite Element Methods for Computational Nano-optics, Fig. 1 Triangulated geometry of a cross section of an optical waveguide (*green*) on an unbounded substrate (*blue*), surrounded by air (*gray*). The domain of interest is separated into an interior domain Ω_{int} and an unbounded exterior domain Ω_{ext}. In the interior the total field $\mathbf{E}_{\text{int}}$ is computed. In the exterior an incoming source field $\mathbf{E}_{\text{source}}$ can be predescribed and only the scattered field $\mathbf{E}_{\text{scatt}}$ is unknown. For waveguide mode and resonance problems $\mathbf{E}_{\text{source}}$ is zero, whereas for scattering problems $\mathbf{E}_{\text{source}}$ has to be a solution to Maxwell's equations in the exterior. For FEM, the unknown electric field is expanded into localized polynomial ansatz functions defined on the patches of the triangulation

like incident electromagnetic waves are also of great interest. In such a setup the nano-optical system is embedded into an infinite exterior. Numerically this problem is usually modeled by a total-field/scattered-field approach [9]. The domain of interest is separated into an infinite exterior Ω_{ext} which hosts the given incident field and the scattered field, and an interior Ω_{int} where the total field is computed. Figure 1 visualizes such a decomposition. A similar setup with zero incident field is used, when computing resonance or propagating mode problems on unbounded domains. For numerical simulations, the infinite exterior is treated using absorbing or transparent boundary conditions such as the perfectly matched layer (PML) method [1, 12] or the pole condition [8].

For application of FEM, the governing equation for the electric field (1) has to be transformed into weak formulation. First, it is tested with a vectorial function ϕ and integrated over $\mathbb{R}^3$:

$$\int_{\mathbb{R}^3} \phi\, \mathbf{curl}\, \mu^{-1} \mathbf{curl}\, \mathbf{E} - \omega^2 \phi \varepsilon \mathbf{E} = i\omega \int_{\mathbb{R}^3} \phi \mathbf{J}.$$

Integration by parts and the assumption that the electric field vanishes at infinity gives

$$\int_{\mathbb{R}^3} \mathbf{curl}\, \phi \mu^{-1} \mathbf{curl}\, \mathbf{E} - \omega^2 \phi \varepsilon \mathbf{E} = i\omega \int_{\mathbb{R}^3} \phi \mathbf{J}.$$

For a compact notation of the weak form of Maxwell's equations, the following bilinear and linear forms are introduced:

$$a(\phi, \mathbf{E}) = \int_{\mathbb{R}^3} \mathbf{curl}\, \phi\, \mu^{-1} \mathbf{curl}\, \mathbf{E} - \omega^2 \phi \varepsilon \mathbf{E},$$
$$f(\phi) = i\omega \int_{\mathbb{R}^3} \phi \mathbf{J}.$$

Furthermore, the function space H (**curl**) [6] is defined as the domain of a(ċ,ċ). The weak form of Maxwell's equations then reads as follows:

Find $\mathbf{E} \in \mathbf{H}(\text{curl})$ such that

$$a(\phi, \mathbf{E}) = f(\phi), \quad \forall \phi \in \mathbf{H}\,(\mathbf{curl}). \tag{2}$$

If a field $\mathbf{E}$ is solution to the strong form of Maxwell's equations (1) then it is also a solution to above infinite-dimensional variational formulation.

Finite-Element Discretization

A (conforming) finite element discretization of Maxwell's equations restricts the variational formulation (Eq. 2) to a finite-dimensional subspace V_h with $\dim V_h = N < \infty$.

Find $\mathbf{E}_h \in V_h$ such that

$$a(\phi_h, \mathbf{E}_h) = f(\phi_h), \quad \forall \phi_h \in V_h. \tag{3}$$

In order to derive an algebraic equation, a basis $\{\varphi_1, \ldots, \varphi N\}$ of V_h is constructed and the electric field is expanded according to $E_h = \sum_{i=1}^{N} e_i \varphi_i$.

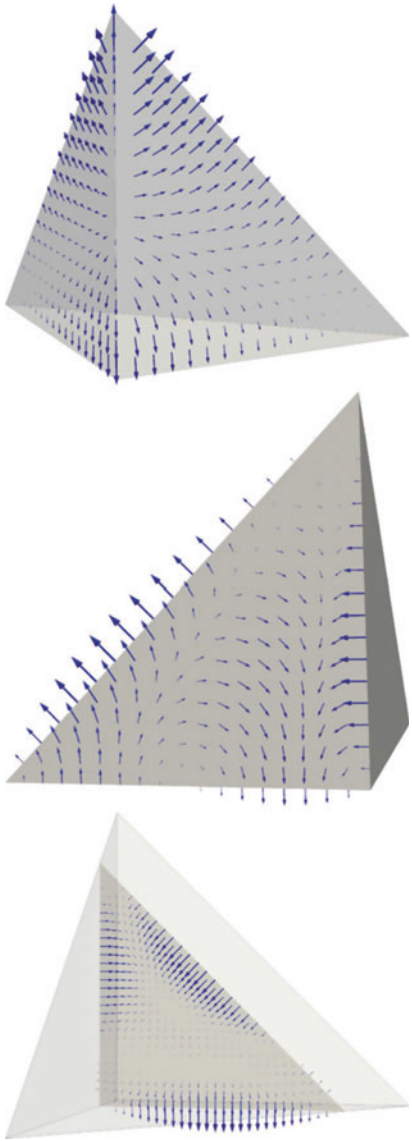
The variational problem is then tested with all elements of the basis which gives a linear system of equations:

$$\sum_{i=1}^{N} a\left(\varphi_j, \varphi_i\right) e_i = f\left(\varphi_j\right), \quad \forall j = 1, \ldots, N.$$

The matrix $A_{ji} = a(\varphi_j, \varphi_i)$ is sparse and can be decomposed with efficient sparse LU solvers to obtain the unknown expansion coefficients e_i of the electric field.

An important aspect of FEM is the construction of the basis functions φ_i. For their definition, the domain of interest is subdivided by a triangulator into a number of small patches. A major advantage of FEM is the possibility to decompose basically arbitrary geometries with unstructured meshes, using different patches like triangles and quadrilaterals in 2D and tetrahedrons, prisms, pyramids, and bricks in 3D. Even curved so-called isoparametric elements are allowed. Figures 1 and 3 show finite-element triangulations of a 2D waveguide cross section and of a 3D photonic crystal microresonator. On each of the patches a number of polynomial ansatz functions up to degree p are defined, which gives the order of the finite-element approximation. For Maxwell's equations their construction is a nontrivial task. To arrive at well-posed problems without artificial spurious solutions, the properties of the appearing differential operators **div** and **curl** in Maxwell's equations have to be respected [6, 7]. The degrees of freedom of such **H(curl)**-conforming finite elements are the tangential field components on the boundaries of the patches. This mathematical property respects the physical feature of the electric field being only tangentially continuous across interfaces, while the normal component can be discontinuous. Figure 2 shows a number of finite element ansatz functions on a tetrahedron. The possibility to use higher-order ansatz functions for the electric field is a great advantage of FEM in nano-optical simulations.

The accuracy of a finite-element solution depends on the choice of the finite-element space V_h in the variational formulation n. The better V_h approximates the continuous space **H** (**curl**), the more accurate the FEM solution gets. One possibility to increase accuracy is the subdivision of the mesh into smaller patches, which is referred to as h-refinement. Thereby, the quantity h denotes a



Finite Element Methods for Computational Nano-optics, Fig. 2 Higher-order H (**curl**)-conforming finite-element functions on a tetrahedron. The degrees of freedom of the ansatz functions are given by the tangential field components on a single edge (*top*), face (*middle*), and volume (*bottom*) of the patch. For the edge- and face-associated ansatz functions the tangential field component vanishes on all but one edge and face respectively. For the volume-associated ansatz function, the tangential component vanishes on all faces and edges

measure for the maximum size of the geometrical patches in the mesh. Self-adaptive mesh refinement can easily be implemented and leads to improved performance, especially in the case of multi-scale geometries. In contrast to other simulation techniques like finite-difference time-domain (FDTD), FEM furthermore allows for a high polynomial degree of the ansatz functions p to increase accuracy. Since often a dependence of the error proportional to h^p can be observed, this results in superior convergence properties and short computational times for FEM [6].

Application Example

Applications of FEM in the field of nano-optical devices include plasmonics, metamaterials, microstructured photonic fibers, microlithography, photonic crystal (PhC) materials and devices, optical semiconductor devices, micro-structured solar cells, microlithographic devices, optical metrology of nano-structures, and others. As a typical application example FEM simulations of high-Q resonances in PhC microcavities are reviewed [3].

Background: Photonic Microcavities

Photonic microcavities have been fabricated for various purposes [10]. One of the applications with potential high impact to technology is integrated optics: Integration of optical components with electronic circuits will provide increased functionality. Building blocks of integrated optical devices are optical waveguides, optical modulators, optical switches, and others. With their high quality factor (Q-factor) and small modal volume, one-dimensional (1D) PhC silicon-on-insulator microcavities are candidates for compact, integrated all-optical switching devices. High Q-factors have been reported for such devices, e.g., by Zain et al. [11]. Further designs of optical microcavities include 2D PhC resonators, micropillar cavities, microdisc, and microsphere resonators [10].

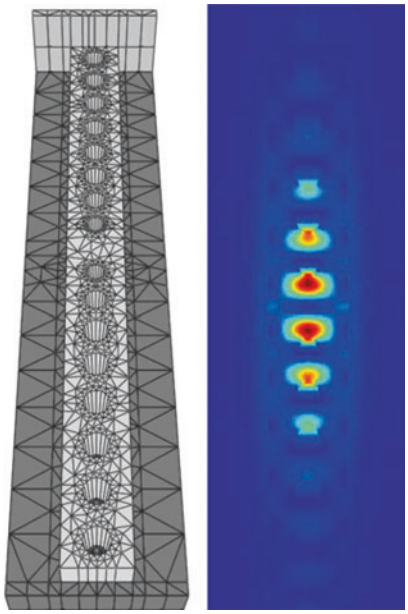
Numerical simulations of optical properties of microcavities are extremely helpful for an efficient design. Here, simulations of high-Q resonances in 1D PhC microcavities are reviewed. Numerical simulations of transmission spectra are performed using a time-harmonic FEM light scattering solver in combination with a propagating mode solver. Further an eigenmode solver is used for direct computation of resonance wavelengths and Q-factors. The used FEM solvers are parts of the program package *JCMsuite* [4].

Investigated Setup

The investigated microcavities consist of a silicon waveguide with square shape which is perforated with a finite number of cylindrical air pores and is supported by a silica substrate. The periodically arranged pores form a finite 1D PhC with a band gap in the relevant wavelength range. The cavity is realized by omitting one (central) pore. A specific setup with individual pore diameters and placements has been proposed and investigated experimentally by Zain et al. [11]. The tapered arrangement of the pore positions and diameters leads to high Q-factor and high transmission values. For FEM simulations, the geometry is discretized by prismatoidal elements [3]. Figure 3 (left) shows a part of a typical spatial mesh. A mirror symmetry allows to reduce the computational domain to one half of the cavity geometry. A typical computational domain has a size of some tens to some hundreds of cubic wavelengths. While these are moderate 3D computational domain sizes, the challenge of the simulation consists mainly in modeling the irregular geometry and in a precise resolution of the electromagnetic field. The very anisotropic intensity distribution on resonance can span many orders of magnitude. The model extends waveguide, substrate, and surrounding air to infinity by applying transparent boundary conditions to the corresponding faces.

Numerical Approach

For simulating transmission through the cavity at specific wavelength λ_0, first the fundamental propagation mode of the waveguide at λ_0 is computed. The obtained mode field is used as an incident field on the boundaries of the 3D computational domain, such that the mode propagates in direction of the center of the waveguide. Then the



Finite Element Methods for Computational Nano-optics, Fig. 3 *Left*: Geometry of a PhC microcavity (with less pores than the actual setup of Zain et al.), discretized with a prismatoidal mesh (*dark gray*: silica substrate, *light gray*: silicon waveguide, perforated by conical air pores). The structure is surrounded by air; substrate and waveguides extend to infinity. *Right*: Intensity distribution at a wavelength close to resonance, in an *xy* cross section at the center of the waveguide (*blue*/*dark*: low intensity, *red*/*bright*: high intensity)

scattered light field is computed using higher-order finite elements. In post-processes, energy fluxes through interfaces and field distributions in several cross sections of the 3D light field distribution are extracted from the near field solution. Transmission spectra are obtained by performing a series of simulations with (adaptively) varied wavelengths λ_0.

Alternatively to simulating the light scattering response of the device, an eigenmode solver can be used for directly simulating the resonance properties (resonance wavelength and Q-factor) [3].

Validation of Numerical Results

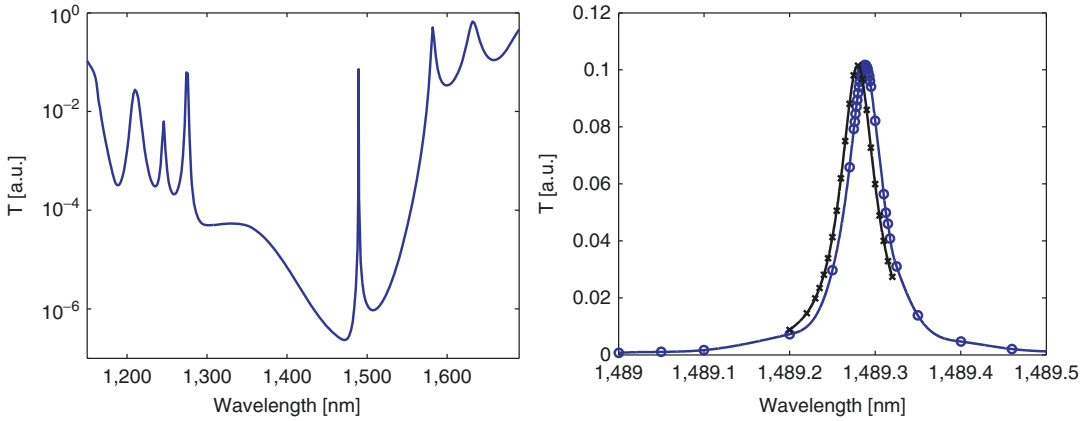
Typically, for a given simulation task each actual simulation will give a different quantitative result unless exactly the same method, the same implementation, the same hardware, and the same numerical discretization parameters are used. This is due to numerical discretization errors (which are present in any numerical method), approximations of the model, implementation errors, and other causes. Hence, it is essential to check numerical results using one or several validation strategies:

- Convergence of the results when numerical accuracy is increased: For FEM simulation this means that it has to be checked that the numerical results converge at the expected rate under mesh refinement and under increasing polynomial order of the finite-element ansatz functions [6].
- Convergence to analytical solutions for simple setups.
- Agreement with results obtained using different numerical methods, like in this case, e.g., FDTD [2] within the numerical error bounds of the respective methods. Agreement with results obtained with the same numerical method, but with different implementations [5].
- Comparison with experimental results: agreement with experimental results within experimental and numerical error bounds when experimental data is accessible.

For the reviewed example, simulation results have been validated by comparisons to independent numerical and experimental results. Also a convergence analysis has been performed, which shows that solutions with high numerical accuracy are reached. Further, it has been checked that resonance wavelength and cavity Q-factor computed from the complex eigenvalue of eigenmode simulations agree perfectly with results from light scattering simulations. Also, in a parameter study it has been shown that slight geometrical variations of the air pores sidewall angle and of feature placement on a nanometer scale can lead to significant changes of transmission wavelength and Q-factor [3].

Numerical Results

Figure 4 (left) shows the simulated transmission spectrum corresponding to the experiment of Zain



Finite Element Methods for Computational Nano-optics, Fig. 4 *Left*: Simulated transmission spectrum of a 1D PhC microcavity composed by an arrangement of 24 air pores with diameters between 131 and 182 nm in a rectangular silicon waveguide with a cross section of 500 × 260 nm (setup of Zain et al.). *Right*: Detail of the transmission spectrum: resonance transmission peak at ∼1,489 nm. *Blue circles*: time-harmonic simulation results for polynomial finite-element degree $p = 3$; *black crosses*: results for finer mesh; *lines*: spline interpolation. With increased numerical accuracy, resonance peak position shifts by about 5×10^{-6} (8×10^{-12} m), FWHM of the interpolation shifts by about 1 % (0.5×10^{-12} m)

et al. The 1D PhC arrangement leads to a bandgap between 1,200 and 1,600 nm in the transmission spectrum. At the cavity resonance wavelength of about 1,489.2 nm, a distinct transmission peak appears within the bandgap. Zooming into the wavelength range of 1,489–1,489.5 nm (Fig. 4, right) shows that the full width at half maximum of the resonance is only about 50 pm (picometer), corresponding to a resonance Q-factor of about 30,000. Figure 3 visualizes the computed electric field intensity at an incident wavelength close to the cavity's resonance wavelength. Clearly, the photonic intensity distribution is stored in a small modal volume of few cubic wavelengths. Figure 4 (right) also shows transmission peaks resulting from simulations with different numerical resolution. Increasing the resolution (i.e., in this case refining the mesh) shifts the resonance wavelength by 8 pm for the shown data and changes the full width by ∼0.5 pm. Together with convergence checks performed for similar setups [3], this indicates that a high accuracy of the numerical results has been reached. While the agreement with the experimentally observed resonance wavelength of 1,479 nm is very good, the experimentally observed Q-factor of about 147,000 is significantly higher than the simulation result. This can be explained by both, experimental error budgets, and differences between the experimentally realized device and the model geometry on a nanometer scale.

Conclusion

The reviewed example highlights several advantages of the use of finite-element methods for the simulation of nano-optical devices:

- Complex real-life geometries can be described accurately, with exact spatial resolution.
- Unstructured meshes allow for spatial adaptivity.
- Higher-order finite elements have very good convergence properties, ensuring results of very high accuracy.
- Isolated problems can be treated rigorously, even in the presence of structured exterior domains.
- The general finite-element setting allows to solve different partial differential equations, e.g., in this case equations modeling light scattering, mode propagation in waveguides, and resonance modes in cavities.

- The usage of direct, sparse matrix-solvers allows to obtain steady-state results at relatively low computational costs and typically several orders of magnitude faster compared to time-domain methods.

References

1. Bérenger, J.: A perfectly matched layer for the absorption of electromagnetic waves. J. Comput. Phys. **114**(2), 185–200 (1994)
2. Burger, S., Köhle, R., Zschiedrich, L., Gao, W., Schmidt, F., März, R., Nölscher, C.: Benchmark of FEM, waveguide and FDTD algorithms for rigorous mask simulation. In: Weed, J.T., Martin, P.M. (eds.) Photomask Technology, Proceedings of the SPIE, vol. 5992, pp. 378–389. (2005)
3. Burger, S., Schmidt, F., Zschiedrich, L.: Numerical investigation of photonic crystal microcavities in silicon-on-insulator waveguides. In: Photonic and Phononic Crystal Materials and Devices X, Proceedings of the SPIE, vol. 7609, p. 76091Q. (2010)
4. Burger, S., Zschiedrich, L., Pomplun, J., Schmidt, F.: JCMsuite: an adaptive FEM solver for precise simulations in nano-optics. In: Integrated Photonics and Nanophotonics Research and Applications, p. ITuE4. Optical Society of America, Boston (2008)
5. Hoffmann, J., Hafner, C., Leidenberger, P., Hesselbarth, J., Burger, S.: Comparison of electromagnetic field solvers for the 3d analysis of plasmonic nano antennas. In: Bosse, H., Bodermann, B. (eds.) Modeling Aspects in Optical Metrology, Proceedings of the SPIE, vol. 7390, p. 73900. (2009)
6. Monk, P.: Finite Element Methods for Maxwell's Equations. Oxford University Press, New York (2003)
7. Nedelec, J.: Mixed finite elements in R^3. Numer. Math. **35**, 315–341 (1980)
8. Schmidt F.: Solution of interior-exterior Helmholtz-type problems based on the pole condition concept: theory and algorithms. Habilitation thesis, Free University Berlin, Fachbereich Mathematik und Informatik (2002)
9. Taflove, A., Hagness, S.C.: The Finite-Difference Time-Domain Method, 3rd edn. Artech House, Boston (2005)
10. Vahala, K.J.: Optical microcavities. Nature **424**, 839–846 (2003)
11. Zain, A.R.M., Johnson, N.P., Sorel, M., De La Rue, R. M.: Ultra high quality factor one dimensional photonic crystal/photonic wire microcavities in silicon-on-insulator (SOI). Opt. Express **16**(16), 12084 (2008)
12. Zschiedrich, L., Klose, R., Schädle, A., Schmidt, F.: A new finite element realization of the perfectly matched layer method for Helmholtz scattering problems on polygonal domains in 2D. J. Comput. Appl. Math. **188**, 12–32 (2006)

Finite-Difference Frequency-Domain Technique

Georgios Veronis
Department of Electrical and Computer Engineering and Center for Computation and Technology, Louisiana State University, Baton Rouge, LA, USA

F

Definition

The finite-difference frequency-domain technique is a numerical technique for the solution of Maxwell's equations of electromagnetism in the frequency domain. It is based on approximating the derivatives in Maxwell's equations by finite differences, and results in a system of algebraic equations, which is solved using numerical linear algebra techniques.

Introduction to the Simulation of Plasmonic Devices

The finite-difference frequency-domain (FDFD) technique is a general-purpose numerical technique for the solution of Maxwell's equations of electromagnetism in the frequency domain. It can be applied to structures of any length scale and for sources of electromagnetic radiation of any frequency. Here, however, the focus is on the use of this technique in nano-optics, and in particular *plasmonics*.

Plasmonics is a rapidly evolving field of science and technology based on *surface plasmons*, which are electromagnetic waves that propagate along the interface of a metal and a dielectric. In surface plasmons, light interacts with the free electrons of the metal, which oscillate collectively in response to the applied field. Recently, nanometer scale metallic devices have shown the potential to manipulate light at the subwavelength scale using surface plasmons [1]. This could lead to photonic circuits of nanoscale dimensions.

Surface plasmons can be described by macroscopic electromagnetic theory, i.e., Maxwell's

equations, if the electron mean free path in the metal is much shorter than the plasmon wavelength [2]. This condition is usually fulfilled at optical frequencies [2]. In macroscopic electromagnetic theory, bulk material properties, such as dielectric constant, are used to describe objects irrespective of their size. Here, all materials are assumed to be nonmagnetic $(\mu = \mu_0)$ and are characterized by their bulk dielectric constant $\epsilon(\mathbf{r}, \omega)$, where $\mathbf{r} = (x, y, z) = x\hat{x} + y\hat{y} + z\hat{z}$, and ω is the angular frequency. However, for particles of nanometer dimensions a more fundamental description of their optical and electronic properties may be required [3].

Analytical methods, such as Mie theory [4], can only be applied to planar geometries or to objects of specific shapes (spheres, cylinders) and have therefore limited importance in the analysis of plasmonic devices and structures. Numerical simulation techniques are therefore very important for the analysis and design of plasmonic devices. Numerical modeling of plasmonic devices involves several challenges which need to be addressed. First, as mentioned above, plasmonic devices can have arbitrary geometries. Several techniques are specific for one type of geometrical configuration and are therefore not appropriate for modeling of arbitrary plasmonic devices.

Second, the dielectric constant of metals at optical wavelengths is complex, i.e., $\epsilon_r(\omega) = \epsilon_{\mathrm{Re}}(\omega) + i\epsilon_{\mathrm{Im}}(\omega)$ and is a complicated function of frequency [5]. Thus, several simulation techniques which are limited to lossless, nondispersive materials are not applicable to plasmonic devices. In addition, in time-domain methods the dispersion properties of metals have to be approximated by suitable analytical expressions [6]. In most cases, the *Drude model* is invoked to characterize the frequency dependence of the metallic dielectric function [7]

$$\epsilon_{\mathrm{r,\ Drude}}(\omega) = 1 - \frac{\omega_p^2}{\omega(\omega + i\gamma)}, \qquad (1)$$

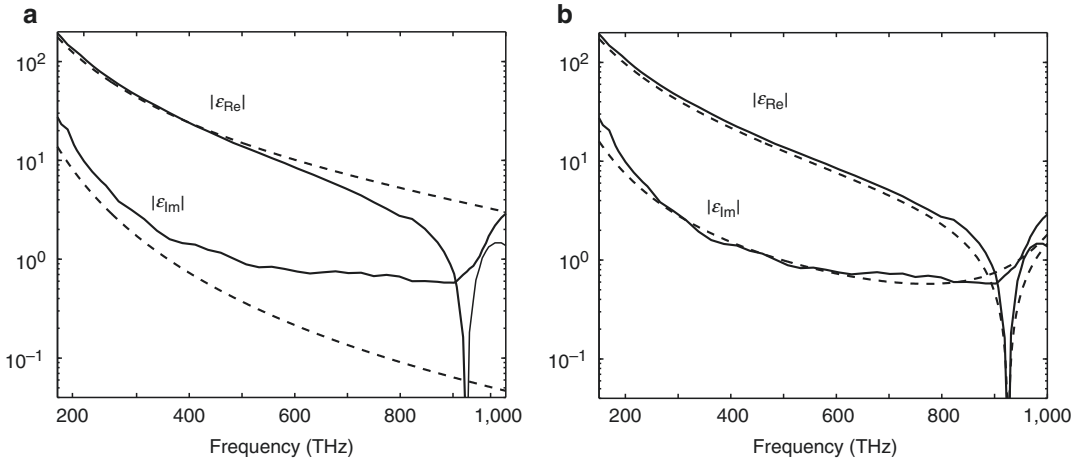
where ω_p and γ are frequency-independent parameters. However, the Drude model approximation is valid over a limited wavelength range [7]. The range of validity of the Drude model can be extended by adding Lorentzian terms to Eq. 1 to obtain the *Lorentz-Drude model* [7]

$$\epsilon_{r,\mathrm{LD}}(\omega) = \epsilon_{r,\mathrm{Drude}}(\omega) + \sum_{j=1}^{k} \frac{f_j \omega_j^2}{\left(\omega_j^2 - \omega^2\right) - i\omega\gamma_j}, \qquad (2)$$

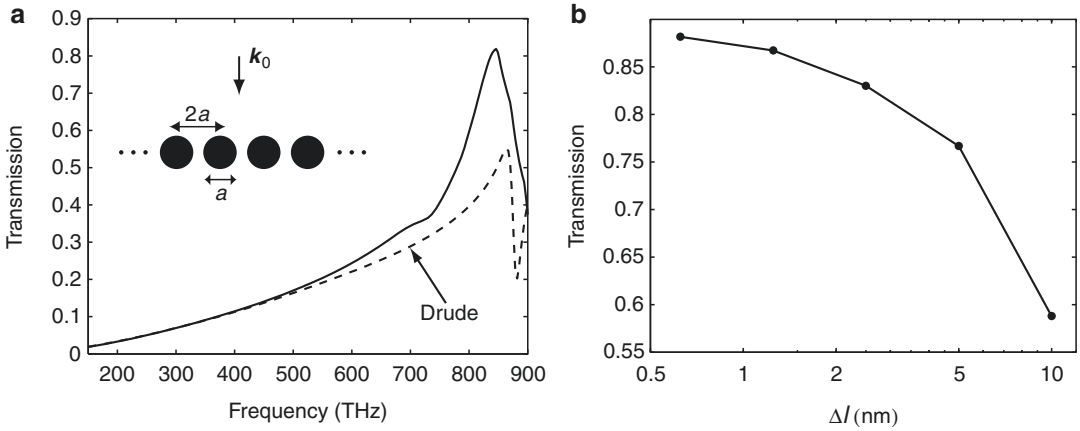
where ω_j and γ_j stand for the oscillator resonant frequencies and bandwidths, respectively, and f_j are weighting factors. Physically, the Drude and Lorentzian terms are related to intraband (free electron) and interband (bound electron) transitions, respectively [7]. Even though the Lorentz-Drude model extends the range of validity of analytical approximations to metallic dielectric constants, it is not suitable for description of sharp absorption edges observed in some metals, unless a very large number of terms are used [7]. In particular, the Lorenz-Drude model cannot approximate well the onset of interband absorption in noble metals (Ag, Au, and Cu), even if five Lorentzian terms are used [7]. In Fig. 1, the Drude and Lorentz-Drude models are compared with experimental data for silver. It is observed that even a five-term Lorentz-Drude model with optimal parameters results in a factor of two error at certain frequencies.

Third, in surface plasmons propagating along the interface of a metal and a dielectric, the field is concentrated at the interface and decays exponentially away from the interface in both the metal and dielectric regions [8]. Thus, for numerical methods based on discretization of the fields on a numerical grid, a very fine grid resolution is required at the metal-dielectric interface to adequately resolve the local fields. In addition, several plasmonic devices are based on components of subwavelength dimensions [8]. In fact, most of the potential applications of surface plasmons are related to subwavelength optics. The nanoscale feature sizes of plasmonic devices pose an extra challenge to numerical simulation techniques.

The challenges involved in modeling plasmonic devices will be illustrated here using a simple example: an infinite periodic array of silver



Finite-Difference Frequency-Domain Technique, Fig. 1 Real and imaginary part of the dielectric constant of silver at optical frequencies. The *solid lines* show experimental data [5]. The *dashed lines* show values calculated using (**a**) the Drude model and (**b**) the Lorentz-Drude model with five Lorentzian terms. The parameters of the models are optimal and obtained through an optimization procedure [7]



Finite-Difference Frequency-Domain Technique, Fig. 2 (**a**) Calculated transmission spectrum of an infinite array of silver cylinders (shown in the *inset*) for normal incidence and transverse magnetic (*TM*) polarization (involving only the E_x, E_y, and H_z vector field components). Results are shown for a diameter $a = 100$ nm. The *dashed line* shows the transmission spectrum calculated using the Drude model (Eq. 1) with parameters $\omega_p = 1.37 \times 10^{16} s^{-1}$ and $\gamma = 7.29 \times 10^{13} s^{-1}$. (**b**) Calculated transmission at 855 THz as a function of the spatial grid size Δl

cylinders illuminated by a plane wave at normal incidence (inset of Fig. 2a). The FDFD method, which is described in detail below, is used to calculate the transmission of the periodic array. This method allows one to directly use experimental data for the frequency-dependent dielectric constant of metals, including both the real and imaginary parts, with no approximation. The fields are discretized on a uniform two-dimensional (2-D) grid with grid size $\Delta x = \Delta y = \Delta l$. The calculated transmission as a function of frequency is shown in Fig. 2a. Also shown is the transmission of the structure calculated with the Drude model of Eq. 1. It is observed that the use of the Drude model results in substantial error. In general, the Drude model parameters

are chosen to minimize the error in the dielectric function in a given frequency range [9]. However, this approach gives accurate results in a limited wavelength range, as illustrated in this example. In general, the complicated dispersion properties of metals at optical frequencies pose a challenge in modeling of plasmonic devices not encountered in modeling of low- or high-index-contrast dielectric devices.

The calculated transmission at a specific wavelength as a function of the spatial grid size Δl is shown in Fig. 2b. It is observed that a grid size of $\Delta l \cong 1$ nm is required in this case to yield reasonably accurate results. The required grid size is directly related to the decay length of the fields at the metal-dielectric interface. In general, modeling of plasmonic devices requires much finer gird resolution than modeling of low- or high-index-contrast dielectric devices, due to the high localization of the field at metal-dielectric interfaces of plasmonic devices. The required grid size depends on the shape and feature size of the modeled plasmonic device, the metallic material used, and the operating frequency.

In the following sections, the FDFD technique is introduced, and it is also examined how FDFD addresses the challenges mentioned above.

Introduction to Finite-Difference Methods

FDFD is based on approximating the derivatives in Maxwell's equations by finite differences. A brief overview of the main features of finite-difference methods is therefore first provided here. In finite-difference methods, derivatives in differential equations are approximated by finite differences. To approximate the derivative $df/dx|_{x_0}$ Taylor series expansions $f(x)$ about the point x_0 at the point $x_0+\Delta x$ and $x_0-\Delta x$ are used to obtain [6]

$$\left.\frac{df}{dx}\right|_{x_0} = \frac{f(x_0+\Delta x)-f(x_0-\Delta x)}{2\Delta x} + O\left[(\Delta x)^2\right]. \tag{3}$$

Here, the notation $O[(\Delta x)^2]$ (to be read as " order $(\Delta x)^2$ ") denotes the remainder term and indicates its dependence on Δx, i.e., that it approaches zero as the second power of Δx. Thus, Eq. 3 shows that a *central-difference* approximation of the first derivative is *second-order accurate*, meaning that the remainder term in Eq. 3 approaches zero as the square of Δx.

In finite-difference methods, a continuous problem is approximated by a discrete one. Field quantities are defined on a discrete grid of nodes. The rectangular grid with node coordinates $\mathbf{r}_{ijk} = \left(x_i, y_j, z_k\right)$ is the simplest and most commonly used. A field quantity at nodal location $\mathbf{r}_{ijk}$ is denoted for convenience as $f|_{ijk} \equiv f\left(\mathbf{r}_{ijk}\right)$. Based on Eq. 3, the first derivative can be approximated by the following central-difference formula

$$\left.\frac{df}{dx}\right|_i \cong \frac{f|_{i+1} - f|_{i-1}}{2\Delta x}, \tag{4}$$

which is second-order accurate, based on the discussion above, if the rectangular grid is uniform, i.e., $x_i = i\Delta x$. For example, if the numerical resolution in the x direction is increased by a factor of 10 ($\Delta x' = \Delta x/10$), then the error introduced by the finite-difference formula in Eq. 4 reduces roughly by a factor of $10^2 = 100$. Similarly, the second derivative can be approximated by the formula

$$\left.\frac{d^2f}{dx^2}\right|_i \cong \frac{f|_{i+1} - 2f|_i + f|_{i-1}}{(\Delta x)^2}, \tag{5}$$

which is also second-order accurate on a uniform grid [6].

By replacing derivatives in differential equations with their finite-difference approximations, algebraic equations are obtained, which relate the value of the field at a specific node to the values at neighboring nodes. By applying the finite-difference approximation to all nodes of the grid, a system of linear equations of the form $\mathbf{Ax} = \mathbf{b}$ is obtained. Since the equation for the field at each point involves only the fields at a few (typically two, four, and six in one, two, and three dimensions, respectively) adjacent points, the resulting system matrix is extremely sparse. Such problems

can be solved efficiently if direct or iterative sparse matrix techniques are used.

The FDFD Method in One Dimension

The FDFD equations will now be derived by approximating the spatial derivatives in Maxwell's equations with finite differences. Assuming an exp($i\omega t$) harmonic time dependence for all field quantities, Maxwell's curl equations in the frequency domain take the form

$$\nabla \times \mathbf{E}(\mathbf{r}) = -i\omega\mu_0 \mathbf{H}(\mathbf{r}), \tag{6}$$

$$\nabla \times \mathbf{H}(\mathbf{r}) = \mathbf{J}(\mathbf{r}) + i\omega\epsilon(\mathbf{r})\mathbf{E}(\mathbf{r}). \tag{7}$$

It should be noted that the field vectors $\mathbf{E}(\mathbf{r}, t)$, $\mathbf{H}(\mathbf{r}, t)$ are real (measurable) quantities that can vary with time, whereas the vectors $\mathbf{E}(\mathbf{r}, \omega)$, $\mathbf{H}(\mathbf{r}, \omega)$ are complex phasors that do not vary with time. The former can be obtained from the latter by multiplying by exp($i\omega t$) and taking the real part. For example,

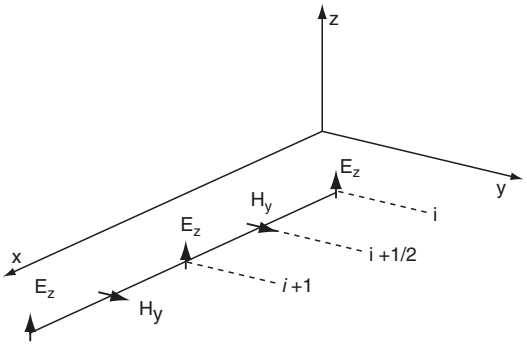
$$\mathbf{E}(\mathbf{r}, t) = \mathrm{Re}\{\mathbf{E}(\mathbf{r}, \omega)\exp(i\omega t)\}. \tag{8}$$

It should also be noted that in most cases for simplicity the frequency dependence of the complex phasors is not explicitly shown, e.g., $\mathbf{E}(\mathbf{r})$ is used instead of $\mathbf{E}(\mathbf{r}, \omega)$. Maxwell's equations can be simplified by considering electromagnetic fields and systems with no variations in two dimensions, namely, both y and z. By dropping all the y and z derivatives in Eqs. 6 and 7, and assuming a current source polarized in the z direction ($\mathbf{J} = J_z\hat{z}$), Maxwell's equations simplify to

$$\frac{\partial E_z}{\partial x} = i\omega\mu_0 H_y, \tag{9}$$

$$\frac{\partial H_y}{\partial x} = J_z + i\omega\epsilon E_z. \tag{10}$$

In the *Finite-Difference Time-Domain* (FDTD) technique a staggered spatial grid, known as the *Yee grid* [6], is used for interleaved placement of the electric and magnetic fields. The Yee grid enables the approximation of the continuous derivatives in space by second-order-accurate two-point centered finite differences. Since the spatial derivatives involved in Maxwell's equations in the frequency domain are exactly the same as those in the time domain, the Yee grid can also be used in FDFD. The placement of the electric and magnetic fields on a one-dimensional (1-D) Yee grid for FDFD is identical to the one for FDTD and is shown in Fig. 3.



Finite-Difference Frequency-Domain Technique, Fig. 3 The placement of electric and magnetic field vectors in FDFD for Maxwell's equations in one dimension

For simplicity a uniform rectangular grid with $x_i = i\Delta x$ is considered, and the derivatives in Eqs. 9 and 10 are replaced with their finite-difference approximations to obtain

$$\frac{E_z|_{i+1} - E_z|_i}{\Delta x} = i\omega\mu_0 H_y\big|_{i+1/2}, \tag{11}$$

$$\frac{H_y\big|_{i+1/2} - H_y\big|_{i-1/2}}{\Delta x} = J_z|_i + i\omega\epsilon|_i E_z|_i. \tag{12}$$

As mentioned above, by applying Eqs. 11 and 12 at all points in the Yee grid, a system of linear equations is obtained, which can be solved to find the electromagnetic fields. If the system obtained from Eqs. 11 and 12 was directly solved, all electric fields $\left(E_z|_1, E_z|_2, \ldots, E_z|_{N_x+1}\right)$ and all magnetic fields $\left(H_y\big|_{1+1/2}, H_y\big|_{2+1/2}, \ldots, H_y\big|_{N_x+1/2}\right)$ would have to be included in the vector of unknown fields. (A 1-D Yee grid terminated at

electric field positions was assumed here.) Thus, the system of linear equations would have a total of $2N_x + 1$ unknowns. However, in the system of equations obtained from the FDFD algorithm, it is straightforward to eliminate either all the electric or all the magnetic fields. To see this, Eq. 11 is applied at two adjacent grid points:

$$\frac{1}{i\omega\mu_0}\frac{E_z|_i - E_z|_{i-1}}{\Delta x} = H_y\big|_{i-1/2}, \tag{13}$$

$$\frac{1}{i\omega\mu_0}\frac{E_z|_{i+1} - E_z|_i}{\Delta x} = H_y\big|_{i+1/2}. \tag{14}$$

Equations 13 and 14 are now substituted into Eq. 12 to obtain

$$\frac{1}{\Delta x}\frac{1}{i\omega\mu_0}\left(\frac{E_z|_{i+1} - E_z|_i}{\Delta x} - \frac{E_z|_i - E_z|_{i-1}}{\Delta x}\right) = J_z|_i + i\omega\epsilon|_i E_z|_i, \tag{15}$$

which can also be written as

$$\frac{1}{(\Delta x)^2}E_z|_{i-1} + \left[-\frac{2}{(\Delta x)^2} + \omega^2\epsilon|_i\mu_0\right]E_z|_i + \frac{1}{(\Delta x)^2}E_z|_{i+1} = i\omega\mu_0 J_z|_i. \tag{16}$$

Thus, application of finite-difference approximations at the node location x_i results in a linear algebraic equation which relates the field $E_z|_i$ to the fields at the two adjacent nodes $E_z|_{i-1}$ and $E_z|_{i+1}$. By applying the finite-difference approximation to all internal nodes of the grid, a system of linear equations of the form $\mathbf{Ax} = \mathbf{b}$ is obtained, where $\mathbf{b}$ is determined by the source current $\mathbf{J}$:

$$\begin{aligned} a_{ii-1} &= a_{ii+1} = \frac{1}{(\Delta x)^2}, \\ a_{ii} &= -\frac{2}{(\Delta x)^2} + \omega^2\,\epsilon|_i\mu_0, \\ b_i &= i\omega\mu_0\, J_z|_i,\ x_i = E_z|_i. \end{aligned} \tag{17}$$

The FDFD equations for the boundary nodes depend on the boundary conditions at the boundary of the simulation domain. Typically, they involve only the two outermost nodes, e.g., $a_{11}x_1 + a_{12}x_2 = 0$. Thus, in the 1-D case the linear system matrix $\mathbf{A}$ is tridiagonal, and it is straightforward to eliminate the magnetic fields and obtain a system of equations which only involves the electric fields $\left(E_z|_1, E_z|_2, \ldots, E_z|_{N_x+1}\right)$ and has a total of $N_x + 1$ unknowns.

Finally, if Eqs. 9 and 10 are combined, the Helmholtz equation in 1-D is obtained:

$$\frac{\partial^2 E_z}{\partial x^2} + \omega^2\epsilon\mu_0 E_z = i\omega\mu_0 J_z. \tag{18}$$

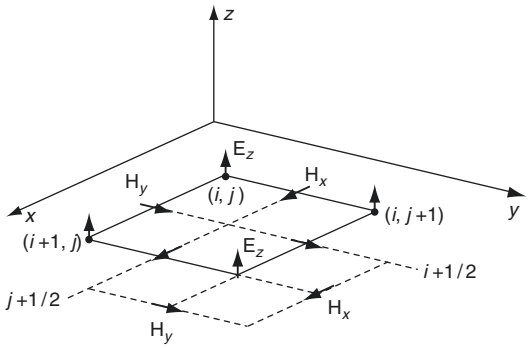
When the spatial derivatives in this equation are approximated by centered finite differences, Eq. 16 above is obtained. In other words, the FDFD equations can be directly obtained by discretizing the Helmholtz equation for the electric field.

The FDFD Method in Two and Three Dimensions

For the 2-D case, it is assumed that there are no variations of either the fields or the excitation in one of the directions, say the z direction. Thus, all derivatives with respect to z drop out from the two curl Eqs. 6 and 7. A current source polarized in the z direction ($\mathbf{J} = J_z\hat{z}$) is also assumed, so that only the transverse electric (TE) polarization (involving only the H_x, H_y, and E_z vector field components) is excited [6] and Maxwell's equations simplify to

$$\begin{aligned} \frac{\partial E_z}{\partial y} &= -i\omega\mu_0 H_x, \\ \frac{\partial E_z}{\partial x} &= i\omega\mu_0 H_y, \\ \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= J_z + i\omega\epsilon E_z. \end{aligned} \tag{19}$$

If Eq. 19 are combined, the 2-D Helmholtz equation is obtained



Finite-Difference Frequency-Domain Technique, Fig. 4 An FDFD unit cell for transverse electric (TE) waves. The small vectors with *thick arrows* are placed at the point in the grid at which they are defined. For example, E_z is defined at grid points (i, j), while H_y is defined at grid points $(i + 1/2, j)$

$$\frac{\partial^2 E_z}{\partial x^2} + \frac{\partial^2 E_z}{\partial y^2} + \omega^2 \epsilon \mu_0 E_z = i\omega\mu_0 J_z. \tag{20}$$

As mentioned in the previous section, the exact same FDTD Yee grid can be used in FDFD. A portion of the Yee grid for the TE case is depicted in Fig. 4.

The FDFD equations can be obtained by discretizing (Eq. 19). It is then straightforward to eliminate H_x and H_y to obtain the FDFD equations which only involve the E_z field. Here, however, the 2-D Helmholtz equation (20) is discretized to directly obtain the FDFD equations which only involve the E_z field. Once E_z is calculated by solving the FDFD equations, the H_x and H_y fields can be calculated by using the discretized versions of Eq. 19.

For simplicity a uniform rectangular grid with $x_i = i\Delta x$ and $y_j = j\Delta y$ is considered, and the derivatives in Eq. 20 are replaced with their finite-difference approximations to obtain

$$\begin{aligned} &\frac{E_z|_{i+1,j} - 2E_z|_{i,j} + E_z|_{i-1,j}}{(\Delta x)^2} \\ &+ \frac{E_z|_{i,j+1} - 2E_z|_{i,j} + E_z|_{i,j-1}}{(\Delta y)^2} \\ &+ \omega^2 \epsilon|_{i,j}\, \mu_0 E_z|_{i,j} = i\omega\mu_0 J_z|_{i,j}. \end{aligned} \tag{21}$$

Thus, application of finite-difference approximations at the node location $\mathbf{r}_{ij} = \left(x_i, y_j\right)$ results in a linear algebraic equation which relates the field $E_z|_{i,j}$ to the fields at the four adjacent nodes $E_z|_{i-1,j}, E_z|_{i+1,j}, E_z|_{i,j-1}, E_z|_{i,j+1}$. In the end, the finite-difference approximation is applied to all nodes of the grid to obtain one FDFD equation at each grid point (i, j). These equations form a system of linear equations $\mathbf{Ax} = \mathbf{b}$ with $(N_x + 1) \times \left(N_y + 1\right)$ equations and unknowns. The vector of unknown fields $\mathbf{x}$ will include the electric fields at all grid points (i, j), where $i = 1, 2, \ldots, N_x + 1$ and $j = 1, 2, \ldots, N_y + 1$. It should be noted that $\mathbf{x}$ is a 1-D vector. A one-to-one mapping between the 2-D electric field $E_z|_{i,j}, \quad i = 1, 2, \ldots, N_x + 1, \quad j = 1, 2, \ldots, N_y + 1$ and the 1-D vector of unknown fields $x_m, \quad m = 1, 2, \ldots, (N_x + 1) \times \left(N_y + 1\right)$ should therefore be constructed. Choosing the right mapping is in general important and can affect the convergence of the system of linear equations. One simple mapping function in the 2-D case is

$$m(i, j) = (i - 1)\left(N_y + 1\right) + j. \tag{22}$$

Using this one-to-one mapping function, the following system of linear equations is obtained:

$$\begin{aligned} a_{m(i,j)m(i-1,j)} &= a_{m(i,j)m(i+1,j)} = \frac{1}{(\Delta x)^2}, \\ a_{m(i,j)m(i,j-1)} &= a_{m(i,j)m(i,j+1)} = \frac{1}{(\Delta y)^2}, \\ a_{m(i,j)m(i,j)} &= -\frac{2}{(\Delta x)^2} - \frac{2}{(\Delta y)^2} + \omega^2 \epsilon|_{i,j}\, \mu_0, \\ b_{m(i,j)} &= i\omega\mu_0\, J_z|_{i,j}, \\ x_{m(i,j)} &= E_z|_{i,j}. \end{aligned} \tag{23}$$

The same approach can be applied to obtain the FDFD equations for the 2-D TM polarization case, as well as for the three-dimensional (3-D) case.

Numerical Dispersion of the FDFD Algorithm

A single plane wave propagating in a uniform medium has a simple *dispersion relation*

connecting the angular frequency ω and the wave number $k : \omega = ck$ [10], where c is the speed of light in the medium. Numerical techniques such as the FDFD technique introduce *numerical dispersion*. To derive the dispersion properties of the FDFD algorithm, a single wave or Fourier mode in space is considered. The dispersive properties of FDFD are assessed by obtaining the dispersion relation of the scheme, relating the frequency of a Fourier mode on the grid to a particular wavelength λ (or wave number k):

$$\omega = f_{\text{FDFD}}(k, \Delta x). \tag{24}$$

In the 1-D case, a plane wave propagating in a uniform medium is considered

$$E(x) = e^{\sqrt{-1}kx}, \tag{25}$$

and its discretized version is

$$E|_i = e^{\sqrt{-1}ki\Delta x}. \tag{26}$$

In the above equations, $\sqrt{-1}$ was used to avoid confusing it with the grid index i. In a uniform medium with no sources the FDFD equation (16) becomes

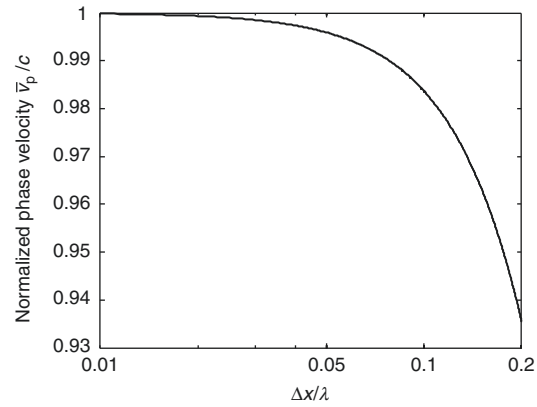
$$\frac{E_z|_{i+1} - 2E_z|_i + E_z|_{i-1}}{(\Delta x)^2} + \omega^2 \epsilon\mu_0 \, E_z|_i = 0. \tag{27}$$

Equation 26 is substituted into Eq. 27 to find

$$\frac{e^{\sqrt{-1}ki\Delta x}\left(e^{\sqrt{-1}k\Delta x} - 2 + e^{-\sqrt{-1}k\Delta x}\right)}{(\Delta x)^2} + \omega^2\epsilon\mu_0 e^{\sqrt{-1}ki\Delta x} = 0, \tag{28}$$

from which one obtains

$$\omega = ck\left(\frac{\sin\left(\frac{k\Delta x}{2}\right)}{\frac{k\Delta x}{2}}\right). \tag{29}$$



Finite-Difference Frequency-Domain Technique, Fig. 5 Dispersion of 1-D FDFD. Variation of numerical phase velocity with numerical resolution $\Delta x/\lambda$ in 1-D FDFD

Equation 29 is the dispersion relation of 1-D FDFD. It is observed that, in the limit $\Delta x \to 0$, Eq. 29 reduces to the exact dispersion relation $\omega = ck$. Using Eq. 29, the numerical phase velocity $\bar{v}_p \equiv \frac{\omega}{k}$ and group velocity $\bar{v}_g \equiv \frac{\partial\omega}{\partial k}$ of the FDFD scheme can be derived. The numerical phase velocity of 1-D FDFD as a function of the numerical resolution $\Delta x/\lambda$ is shown in Fig. 5. It is observed, for example, that for a spatial resolution of 20 grid points per wavelength $(\Delta x/\lambda = 0.05)$ the numerical error in the phase velocity is less than 1 %.

In the 2-D case, wave-like modes propagating in both x and y directions are considered

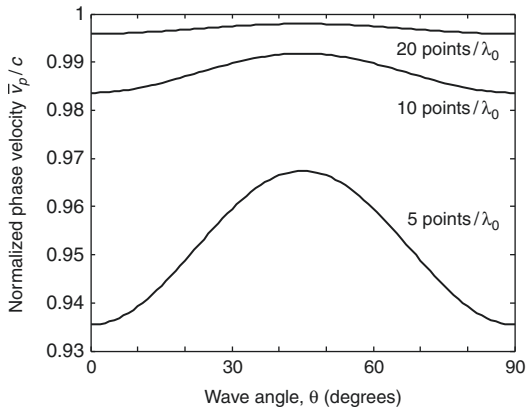
$$E(x, y) = e^{\sqrt{-1}\left(k_x x + k_y y\right)}, \tag{30}$$

and their discretized version is

$$E|_{i,j} = e^{\sqrt{-1}\left(k_x i\Delta x + k_y j\Delta y\right)}, \tag{31}$$

where k_x and k_y are the wave numbers in the x and y directions, respectively.

In a uniform medium with no sources, the 2-D FDFD equation (21) becomes



Finite-Difference Frequency-Domain Technique, Fig. 6 Dispersion of 2-D FDFD. Variation of numerical phase velocity with wave-propagation angle in a 2-D FDFD grid for three different cases of numerical resolution

$$\frac{E_z|_{i+1,j} - 2E_z|_{i,j}}{(\Delta x)^2} + \frac{E_z|_{i,j+1} - 2E_z|_{i,j} + E_z|_{i,j-1}}{(\Delta y)^2} + \omega^2 \epsilon \mu_0 \, E_z|_{i,j} = 0. \tag{32}$$

Equation 31 is substituted into Eq. 32 to find after some manipulation:

$$\omega^2 = c^2 \left[k_x^2 \left(\frac{\sin\left(\frac{k_x \Delta x}{2}\right)}{\frac{k_x \Delta x}{2}} \right)^2 + k_y^2 \left(\frac{\sin\left(\frac{k_y \Delta y}{2}\right)}{\frac{k_y \Delta y}{2}} \right)^2 \right]. \tag{33}$$

It is observed again that, in the limit $\Delta x \to 0$ and $\Delta y \to 0$, Eq. 33 reduces to the exact dispersion relation $\omega^2 = c^2\left(k_x^2 + k_y^2\right)$.

It can be seen from Eq. 33 that the numerical phase velocity $\bar{v}_p$ for the 2-D case is a function of angle of propagation through the FDFD grid. To see this, wave propagation at an angle θ with respect to the positive x axis is assumed, in which case $k_x = k\cos\theta$ and $k_y = k\sin\theta$, where $k = \sqrt{k_x^2 + k_y^2}$ is the wave number. Figure 6 shows a plot of normalized numerical phase velocity as a function of propagation angle θ. The dependence is, in general, relatively small (compared to dispersion errors due to the discretized grid), with $\bar{v}_p/c$ varying by only a few percent between $\theta = 45°$ and $\theta = 0°$, even for very coarse (e.g., $\Delta x = \Delta y = \lambda_0/5$) spatial grids. The dependence of $\bar{v}_p$ on propagation angle θ is known as *grid anisotropy* and is a source of additional numerical dispersion effects.

Finally, it is noted that, unlike FDTD, in FDFD only the spatial derivatives are approximated by finite differences. In fact, for time-harmonic sources and fields the FDTD equations reduce to the FDFD equations in the limit of $\Delta t \to 0$.

F

Waveguiding Structures

In the previous sections, it was described how Maxwell's equations are solved with the FDFD method when the fields are excited by a current source **J**. A slightly different approach is required if one is interested in determining the modes and the dispersion characteristics of waveguiding structures which are uniform in the z direction. If all field components have an $\exp(-\gamma z)$ dependence, it can be shown that Maxwell's equations reduce to [11]

$$-\epsilon_r k_0^2 h_x + \epsilon_r \frac{\partial}{\partial y}\left[\epsilon_r^{-1}\left(\frac{\partial h_y}{\partial x} - \frac{\partial h_x}{\partial y}\right)\right] - \frac{\partial}{\partial x}\left(\frac{\partial h_x}{\partial x} + \frac{\partial h_y}{\partial y}\right) = \gamma^2 h_x, \tag{34}$$

$$-\epsilon_r k_0^2 h_y - \epsilon_r \frac{\partial}{\partial x}\left[\epsilon_r^{-1}\left(\frac{\partial h_y}{\partial x} - \frac{\partial h_x}{\partial y}\right)\right] - \frac{\partial}{\partial y}\left(\frac{\partial h_x}{\partial x} + \frac{\partial h_y}{\partial y}\right) = \gamma^2 h_y, \tag{35}$$

where $\mathbf{H}(x, y, z) = \mathbf{h}(x, y)\exp(-\gamma z)$. Equations 34 and 35 also satisfy the transversality constraint for the magnetic field ($\nabla \cdot \mathbf{H} = 0$). They are discretized by applying the finite-difference method on the Yee grid. A sparse matrix eigenvalue problem of the form $\mathbf{Ah} = \gamma^2\mathbf{h}$ is then obtained, which can be solved using iterative sparse eigenvalue techniques [12]. An important feature of this formulation is the absence of spurious modes in the solution [12], which is therefore robust.

Quasistatic Approximation

Several plasmonic structures have dimensions much smaller than the wavelength λ of the incident light. Under these conditions, retardation effects are negligible and the quasistatic approximation $\mathbf{E}(\mathbf{r}) = -\nabla\varphi(\mathbf{r})$ can be used for the scattered electric field, where $\varphi(\mathbf{r})$ is the electrostatic potential. The field distribution problem then reduces to solving [13]

$$\nabla \cdot \left\{\epsilon(\mathbf{r})\left[-\nabla\varphi(\mathbf{r}) + \mathbf{E}^i(\mathbf{r})\right]\right\} = 0, \qquad (36)$$

which represents the current conservation law. Equation 36 is discretized using the finite-difference method, as described above. A sparse system of linear equations of the form $\mathbf{Ax} = \mathbf{b}$ is obtained, where $\mathbf{b}$ is determined by the incident field $\mathbf{E}^i(\mathbf{r})$.

Comparison to Other Numerical Techniques

FDFD is a frequency-domain technique and can thus treat arbitrary material dispersion. Nonuniform and/or nonorthogonal grids are required in FDFD for efficient treatment of curved surfaces and rapid field variations at material interfaces. In FDFD, as in all other methods which are based on discretization of the differential form of Maxwell's equations in a finite volume, absorbing boundary conditions (ABCs) are required, so that waves are not artificially reflected at the boundaries of the computational domain. Very efficient and accurate ABCs, such as the perfectly matched layer (PML) [6, 12], have been demonstrated for FDFD. As mentioned above, FDFD results in extremely sparse systems of linear equations. Such problems can be solved efficiently if direct or iterative sparse matrix techniques are used.

FDTD is also a finite-difference method, so its performance in modeling plasmonic devices is similar to the performance of FDFD. There are many similarities between the two techniques. Both can be used to model structures with arbitrary geometries. In addition, many of the methods used in combination with FDTD (total field/scattered field method, etc.) can also be used with FDFD. However, there are some major differences. First, as already mentioned above, in time-domain methods the dispersion properties of metals have to be approximated by suitable analytical expressions which introduce substantial error in broadband calculations. In addition, the implementation of methods for dispersive materials in FDTD requires additional computational cost and extra memory storage [6, 14]. On the other hand, in FDTD it is possible to obtain the entire frequency response with a single simulation by exciting a broadband pulse and calculating the Fourier transform of both the excitation and the response [6].

The *Finite-Element Frequency-Domain* (FEFD) method is a more powerful technique than FDFD, especially for problems with complex geometries. However, FDFD is conceptually simpler and easier to program. The main advantage of FEFD is that complex geometric structures can be discretized using a variety of elements of different shapes, while in FDFD a rectangular grid is typically used leading to staircase approximations of particle shapes [12, 15]. In addition, in FEFD fields within elements are approximated by shape functions, typically polynomials, while in FDFD a simpler piecewise constant approximation is used [15]. In short, FEFD is more complicated than FDFD but achieves better accuracy for a given computational cost [15].

Future Research Directions

The dispersion properties of metals at optical frequencies are a major challenge in modeling of plasmonic devices. Frequency-domain techniques can treat arbitrary material dispersion but require a large number of simulations to obtain the broadband response. One possible future research direction could be the development of suitable fast frequency-sweep methods [12]. Such methods allow the computation of the broadband response by simulating the device only at a limited number of frequencies. Another major challenge in

modeling of plasmonic devices is the very fine discretization required to adequately resolve the rapid field decay away from metal-dielectric interfaces. Although nonuniform and/or nonorthogonal grids increase the efficiency of standard numerical methods for such problems, it would be interesting to investigate whether numerical grids could be optimized specifically for treatment of plasmonic devices. Finally, the development of efficient numerical linear algebra techniques for the solution of the sparse systems of algebraic equations resulting from FDFD is also a major challenge. While there are general-purpose iterative techniques and preconditioners, it would be interesting to investigate whether such numerical linear algebra tools could be fine-tuned for frequency-domain simulation of plasmonic devices. In conclusion, there are many exciting research opportunities in developing more accurate and more efficient methods for plasmonic devices.

Cross-References

- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Finite-Difference Time-Domain Technique

References

1. Maier, S.A.: Plasmonics: Fundamentals and Applications. Springer, New York (2007)
2. Novotny, L., Hecht, B., Pohl, D.W.: Interference of locally excited surface plasmons. J. Appl. Phys. **81**, 1798–1806 (1997)
3. Prodan, E., Nordlander, P., Halas, N.J.: Effects of dielectric screening on the optical properties of metallic nanoshells. Chem. Phys. Lett. **368**, 94–101 (2003)
4. Bohren, C.F., Huffman, D.R.: Absorption and Scattering of Light by Small Particles. Wiley, New York (1983)
5. Palik, E.D. (ed.): Handbook of Optical Constants of Solids. Academic, New York (1985)
6. Taflove, A.: Computational Electrodynamics. Artech House, Boston (1995)
7. Rakic, A.D., Djurisic, A.B.., Elazar, J.M., Majewski, M.L.: Optical properties of metallic films for vertical cavity optoelectronic devices. Appl. Opt. **37**, 5271–5283 (1998)
8. Barnes, W.L., Dereux, A., Ebbesen, T.W.: Surface plasmon subwavelength optics. Nature **424**, 824–830 (2003)
9. Vial, A., Grimault, A.S., Macias, D., Barchiesi, D., De la Chapelle, M.L.: Improved analytical fit of gold dispersion: application to the modeling of extinction spectra with a finite-difference time-domain method. Phys. Rev. B **71**, 85416 (2005)
10. Jackson, J.D.: Classical Electrodynamics. Wiley, New York (1999)
11. Pereda, J.A., Vegas, A., Prieto, A.: An improved compact 2D fullwave FDFD method for general guided wave structures. Microw. Opt. Technol. Lett. **38**, 331–335 (2003)
12. Jin, J.: The Finite Element Method in Electromagnetics. Wiley, New York (2002)
13. Genov, D.A., Sarychev, A.K., Shalaev, V.M.: Plasmon localization and local field distribution in metal-dielectric films. Phys. Rev. E **67**, 56611 (2003)
14. Young, J.L., Nelson, R.O.: A summary and systematic analysis of FDTD algorithms for linearly dispersive media. IEEE Antennas Propag. Mag. **43**, 61–77 (2001)
15. Sadiku, M.N.O.: Numerical Techniques in Electromagnetics. CRC Press, Boca Raton (2001)

Finite-Difference Time-Domain Technique

Maxim Sukharev
Department of Applied Sciences and Mathematics, Arizona State University, Mesa, AZ, USA

Synonyms

FDTD

Definition

Finite-difference time-domain technique is one of the most popular numerical approaches in computational nano-optics. It is a numerical solver of macroscopic Maxwell's equations on a special grid in time domain utilizing finite differences.

Overview

The field of nano-optics has been tremendously impacted by both a fast growth of laser technology

and significant advances in nanofabrication techniques in recent years. Owing to such progress, researchers now have the ability to go far beyond the diffraction limit to explore optics at the nanoscale, opening up a wide variety of applications ranging from metamaterials, through plasmonic circuitry and efficient solar energy harvesting, to biomedical applications [1]. Among those applications is the long-standing question of controlling light at the subdiffraction scale aiming for optical nanodevices and nanoscale coherent sources operating in the visible region of spectrum [2]. From the fundamental physics point of view, it is also important to understand general behavior of electromagnetic (EM) near-fields and their interaction with atoms and molecules. From that perspective, information on EM field dynamics at the subwavelength scale and the ability to predict EM field's behavior in the presence of mesoscopic materials such as metal tips, metal nanoparticles (NP), and arrays thereof are needed.

The problem of EM field dynamics in the vicinity of nanostructural materials has been the topic of both experimental and theoretical works for several years [3]. It has been realized that materials composed of metal NPs are capable of strong EM field enhancements and coherent EM energy transport in space. Although modern optical measurements are capable of capturing some near-field phenomena, they obviously lack complete physical picture of EM dynamics. It is thus important to have reliable numerical approach, which may tackle various geometries of nanomaterials and yet is relatively easy to use.

Among many numerical methods used in nano-optics, the finite-difference time-domain (FDTD) technique is considered the most easily to utilize. It has recently become a widely used tool in nano-optics due to its straightforward code implementation and the capability to accurately model complex nanostructures.

Basic Concepts

With the assumption that EM radiation can be properly described by the set of Maxwell's equations, one has the following system to solve (MKS system of units is used unless specified otherwise) [4]

$$\begin{aligned} \mu_0 \frac{\partial \vec{H}}{\partial t} &= -\nabla \times \vec{E}, \\ \varepsilon_0 \frac{\partial \vec{E}}{\partial t} &= \nabla \times \vec{H} - \frac{\partial \vec{P}}{\partial t}, \end{aligned} \tag{1}$$

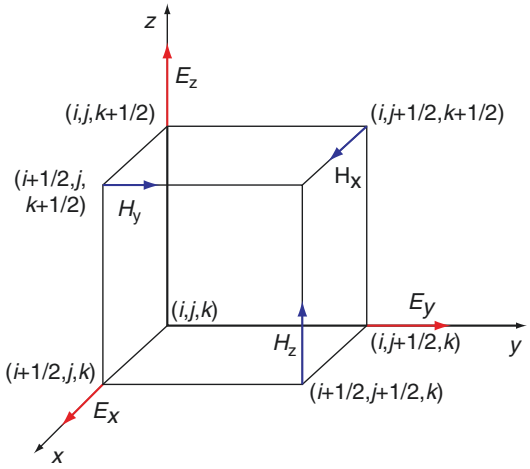
$\vec{E}$ and $\vec{H}$ are electric and magnetic fields, respectively, μ_0 and ε_0 are magnetic permeability and dielectric permittivity of the free space, and $\vec{P}$ is the macroscopic polarization of the nanomaterial. In spatial regions occupied by a metal nanostructure (such as metal NP, for instance), the second equation in (Eq. 1) is modified by replacing the polarization current by a current density $\vec{J}$. The latter is needed to take into account material dispersion as discussed below.

When one writes equations (Eq. 1) replacing spatial derivatives with finite difference and then propagates the resulting system in time, the problem becomes numerically unstable for several geometries important in nano-optics. Moreover, even a simple example of plane wave propagation in free space leads to unphysical results showing numerical dispersion. The careful examination of the source of such instabilities reveals the importance of curl nature of EM waves. This problem was successfully solved by Kane Yee in 1966. In his pioneer paper, Yee proposed a new approach, which then became a basis of FDTD technique [5]. The idea is to take into account the curl nature of EM fields using a special cube, where each component of EM radiation is displaced by a half of spatial step as shown in Fig. 1. One usually introduces discrete coordinates

$$\begin{aligned} x_i &= x_0 + \delta x(i-1), \quad i = 1, \ldots, N_x, \\ y_j &= y_0 + \delta_y(j-1), \quad j = 1, \ldots, N_y, \\ z_k &= z_0 + \delta z(k-1). \quad k = 1, \ldots, N_z. \end{aligned} \tag{2}$$

where δx, δ_y, and δz are spatial steps. A half of spatial step $\delta x/2$ is conventionally denoted in Fig. 1 as 1/2.

It is easy to show that with such arrangement of electric and magnetic field components Gauss's



Finite-Difference Time-Domain Technique, Fig. 1 The Yee cell. The size of each side of the cube is a half of spatial step

law for electric and magnetic fields is automatically satisfied in free space. Furthermore, the boundary conditions at each grid point are automatically maintained. The latter is critical for modeling of complex structures, where other methods are particularly hard to implement due to boundaries between materials. The Yee grid has several attractive technical features, including numerical stability and availability of the magnetic field. The latter is especially important if one considers structures with sharp corners at which the tangential components of $\vec{H}$ have singularities.

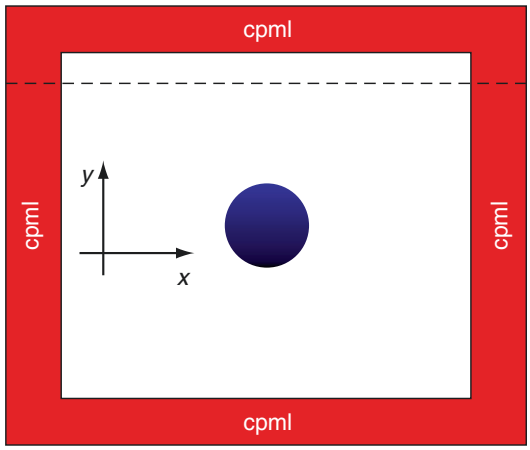
In nano-optics, one usually deals with initial value problems, where initial conditions for EM fields are given and one is interested in their subsequent dynamics, i.e., scattering and absorption of laser pulses by nanomaterials. It is thus important to develop reliable and numerically stable time propagation procedure. The most efficient and hence widely used technique is the leapfrog time-stepping algorithm. Here all electric field components are updated using known magnetic field components. Next magnetic field is updated and stored in memory using updated electric field. The process is then repeated until the desired propagation time is reached. It is important to note that such procedure is non-dissipative with no artificial decay of EM waves propagated on the grid. This technique has a second order accuracy with respect to time step.

The numerical convergence sensitively depends on spatial resolution δx and temporal step δt. While the choice of the latter is usually predetermined by the dimensionality of the problem (for three-dimensional systems the conventional time step is $\delta t < c = \left(\delta x\sqrt{3}\right)$), the former has to be determined by particular fraction of the characteristic wavelength, λ_c, of nanomaterial. For silver/gold nanostructures, this results in $\delta x \leq 1.5$ nm.

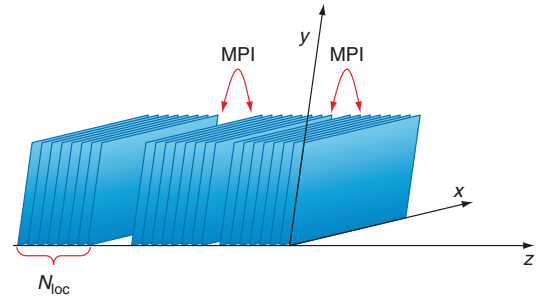
For simulations of open systems, one needs to impose artificial absorbing boundaries in order to avoid numerical reflection of outgoing EM waves back to the simulation domain. From among the various approaches that address this numerical issue, the perfectly matched layers (PML) technique is considered to be the most powerful [6]. It reduces the reflection coefficient of outgoing waves at the simulation region boundary to $\sim 10^{-8}$. Essentially, the PML approach surrounds the simulation domain by thin layers of nonphysical material that efficiently absorb outgoing waves incident at any angle. The most efficient and least memory intensive method is the convolution perfectly matched layers (CPML) absorbing boundaries. It should be noted that parameters of CPML have to be carefully chosen. The latter are usually problem dependent and have to be determined for each structure independently. However, the good starting point is to begin with the following set of CPML parameters and verify whether CPML introduces no reflections of outgoing waves back into the simulation domain

$$\kappa = 20,\ \sigma = 0.75,\ \alpha = 0.24. \qquad (3)$$

Optimal number of CPML layers varies from one simulation to another; however, 20 layers is usually enough to ensure efficient absorption at almost all incident angles. It is also important to note that for periodic structures supporting high Q resonances, CPML boundaries should be placed at least two characteristic wavelengths away from the system. For example, to simulate



Finite-Difference Time-Domain Technique, Fig. 2 Plane wave generation. A nanostructure is placed at the center of the simulation domain. The domain is surrounded by CPML layers on all sides in order to simulate an open system. Horizontal *dashed line* represents the incident plane wave generation



Finite-Difference Time-Domain Technique, Fig. 3 Parallel implementation. Schematic parallelization of the FDTD, where N_{loc} xy-slices are carried by each processor. Send-and-receive MPI subroutines are implemented at the boundaries between each processor

properly transmission and reflection of laser radiation by a silver film (with the thickness of 100 nm) with a single subwavelength slit (with a size of 70 nm), CPML boundaries must be placed 2 μm away from the film.

Depending on particular problem, various types of incident sources have to be implemented. The most common source in nano-optics is a plane wave. There are several ways to inject plane wave into the simulation domain. If one is interested in the normal incidence case, the most efficient and less memory-consuming technique is to include artificial current source of a given direction into second equation in (Eq. 1). It is illustrated in Fig. 2. Here dashed line represents the incident plane, at which every spatial point for x-component of electric field is driven by a current source of the form

$$E_{\text{inc}}(t) = E_0 f(t) \cos(\omega t + \varphi), \tag{4}$$

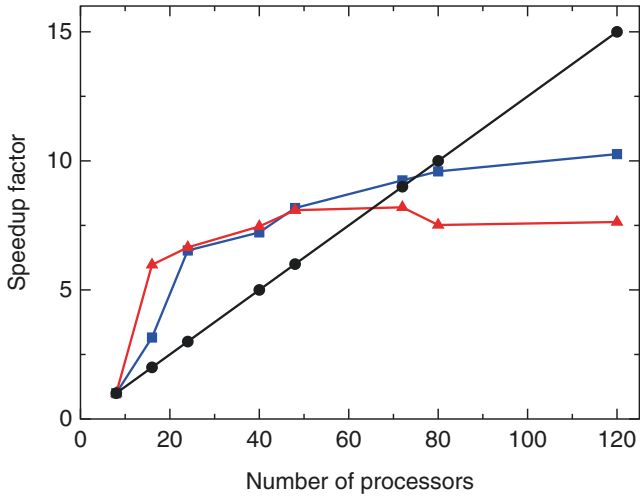
where $f(t)$ is the pulse envelope, ω and φ are incident frequency and the phase. It is important to use a smooth function of time in order to avoid generation of spurious waves. The widely used pulse envelope is

$$f(t) = \begin{cases} \sin^2\left(\pi \dfrac{t}{\tau}\right), & 0 \le t \le \tau, \\ 0, & t \ge \tau, \end{cases} \tag{5}$$

where τ is the pulse duration. Note that not only the envelope is 0 at the beginning and at the end of the pulse, but also its time derivative. It should be noted that the first term in the second equation in (Eq. 1) is also included – this ensures that the plane, at which the incident pulse is generated, is transparent for all waves that are scattered by a nanostructure. This source is referred to as soft source. Finally, both sides of the incident plane have to embedded into CPML regions as shown in Fig. 2.

One of the several advantages of FDTD is its simple finite difference layout, which allows a straightforward parallelization carried out on multiprocessor machines. As illustrated in Fig. 3, the FDTD scheme is partitioned onto a parallel grid by dividing the simulation space into M xy-slices, where M is the number of available processors. Point-to-point message passing interface (MPI) communication subroutines are implemented at the boundaries between slices. The number of xy planes in each slice denoted as N_{loc} usually varies in the range from 12 to 50. The optimal value of N_{loc} for particular problem is determined by scaling simulations – the same geometry is simulated using different number of processors.

The typical test for the parallelizability and portability of the parallelization scheme is shown in Fig. 4. The presented benchmarks of



Finite-Difference Time-Domain Technique, Fig. 4 Scaling example. Speedup factor defined in (Eq. 6) as a function of number of processors in double logarithmic scale. Ideal speedup is shown by *black circles, blue squares* represent the data obtained on ASU supercomputer Plasmon, and *red triangles* show simulations performed on NCSA cluster Abe

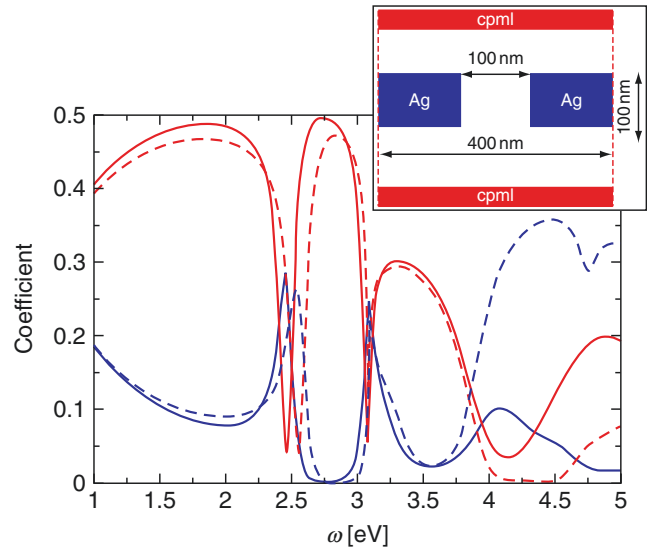
simulations are performed on both the Intel 64 Abe cluster at the NCSA and the Plasmon cluster – a 128-core AMD 64 Opteron–based machine that was assembled recently at ASU Polytechnic. The model tested corresponds to the scattering and absorption of an EM plane wave by a periodic array of silver NPs of 50 nm in diameter. In this problem, the two-dimensional FDTD grid has a size of 792 × 792 nm, with a step size of 1.1 nm. The number of time iterations within the leapfrog time-stepping technique is 650,000, resulting in total propagation time of 1.6 ps. An interesting observation in these studies is the vector eigenmodes of the interacting NPs. To test the efficiency of the parallelization procedure, the speedup factor is computed, defined as

$$S_n = \frac{T_8}{T_n}, \tag{6}$$

where T_8 is the execution time on eight processors and T_n is the execution time on n processors. The choice of the scaling is not arbitrary since Abe and Plasmon clusters have eight cores per node. According to Amdahl's law, the ideal speedup is $n/8$ [7].

As illustrated in Fig. 4, the measured speedup factor is nearly ideal. It degrades with increasing number of processors since in the model simulations, the grid size was fixed and hence the number of layers per processor decreased with increasing number of processors, resulting in more sending-and-receiving MPI operations per iteration with consequent latency of the network. Nonetheless, the speedup for the worst case of 120 processors, where number of layers per each processor is only 6, remains good: $S_{120} = 10.26$ as compared to the ideal speedup factor of 15.06 for 120 processors. Another interesting observation is the so-called nonlinear speedup regime for small number of processors, where measured *Sn* is exceeding the ideal limit. Similar results have been obtained for full three-dimensional FDTD codes for open systems and periodic structures demonstrating efficiency of the parallelization procedure. The nonlinear deviation of the simulated speedup from its ideal value can be explained as follows. By increasing the number of processors involved in the simulations while keeping the problem size constant, one effectively decreases the amount of swapping that the code performs. In other words, for those points in Fig. 4 at which the simulated speedup is higher than its ideal value, the memory occupied by the locally propagated matrices (consisting of the EM field components) is less than the cash size of each processor.

Using the computed speedup factor, it is informative to calculate another important performance measure, namely, the F-factor that specifies the portion of simulations that cannot be parallelized, i.e., the portion of an algorithm



Finite-Difference Time-Domain Technique, Fig. 5 Comparison between Drude-Lorentz (*solid lines*) and Drude (*dashed lines*) models for the frequency-dependent transmission (*blue dark lines*) and reflection (*red bright lines*) of EM radiation by a periodic slit array. The latter is schematically depicted in the inset. Here two vertical *red dashed lines* show periodic boundaries

that is sequential in nature. The F-factor is related to the speedup factor (Eq. 6) as

$$F = \frac{n - S_n}{S_n(n-8)}. \tag{7}$$

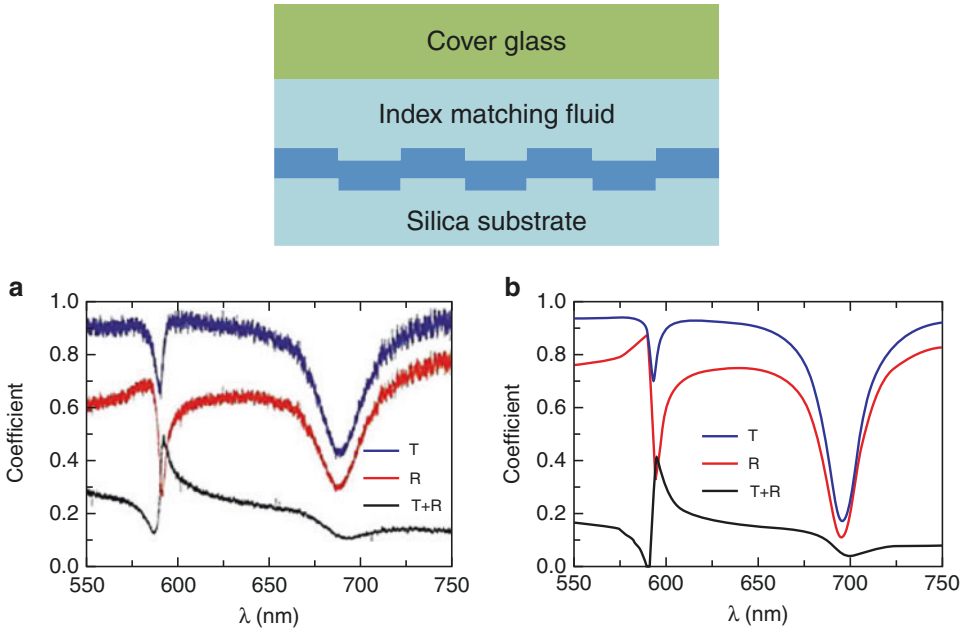
The calculated F-factor (Eq. 7) for both open and periodic codes shows outstanding parallelization. The largest F-factor is nearly 9 %, which indicates that only 9 % of the code is sequential.

To account for material dispersion in spatial regions of a nanostructure occupied by metal or dispersive dielectric, one employs the auxiliary differential equation method [4]. In general, the dielectric response of the metal is modeled using a standard Lorentz-Drude (LD) formulation [8]

$$\varepsilon(\omega) = \left[\varepsilon_r - \frac{\Omega_p^2}{\omega^2 + i\Gamma\Omega_p}\right] + \sum_n \frac{f_n}{\left(\omega_n^2 - \omega^2\right) + i\gamma_n\omega}. \tag{8}$$

The expression in brackets represents the standard Drude model that includes the bulk plasma frequency of the electron system, Ωp, and its width (damping), Γ. The material dispersion of metals can be accurately described within this model only in the limit of high frequencies. Sometimes it is desirable to model material dispersion over a wide frequency range. Hence, it is necessary to include a more general Lorentzian model (the second term in Eq. 8) that accurately covers the optical and near-infrared wavelength region. As seen in Eq. 8 the Drude model is a special case of the Lorentzian model describing an oscillator with the zero restoring force. Finally one needs to fit Eq. 8 to the experimentally observed dielectric response of, e.g., Ag and Au in the range 0.1–5 eV [9].

Figure 5 illustrates the applicability of the Drude model and compares it with more general description Eq. 8. The performed simulations calculate transmission and reflection properties of a periodic slit array perforated on a thin silver film (see insert in Fig. 5). In order to obtain transmission and reflection as functions of the incident frequency, the system is excited by an ultra-short pulse of the duration of 0.36 fs. In the frequency domain, such pulse is almost flat for all frequencies of interest. Once the system is excited, Maxwell's equations are propagated for about 2 ps and EM field components are recorded on both input and output sides of the structure as functions of time. Subsequent Fourier transformation of the detected fields and calculations of the Poynting vector result in transmission and reflection coefficients. It should be noted that the entire spectrum is obtained via a single FDTD run. Figure 5 shows direct comparison of the Drude model and the LD



Finite-Difference Time-Domain Technique, Fig. 6 Up-down diffraction grating. Experiment (*left frame*) and FDTD (*right frame*) data for the transmission (*black lines*, lowest curves in each frame), reflection (*red lines*, curves in the middle in each frame), and transmission + reflection (*blues lines*, upper curves in each frame) as functions of the incident wavelength in nanometers. The up-down grating setup is schematically depicted in the inset, where blue regions correspond to silver. Effective thickness of the grating is 35 nm and the period is 197 nm

model. Even though results obtained within the Drude model qualitatively agree with LD simulations for low-energy part of the spectrum, the Drude approach fails at frequencies above 4 eV.

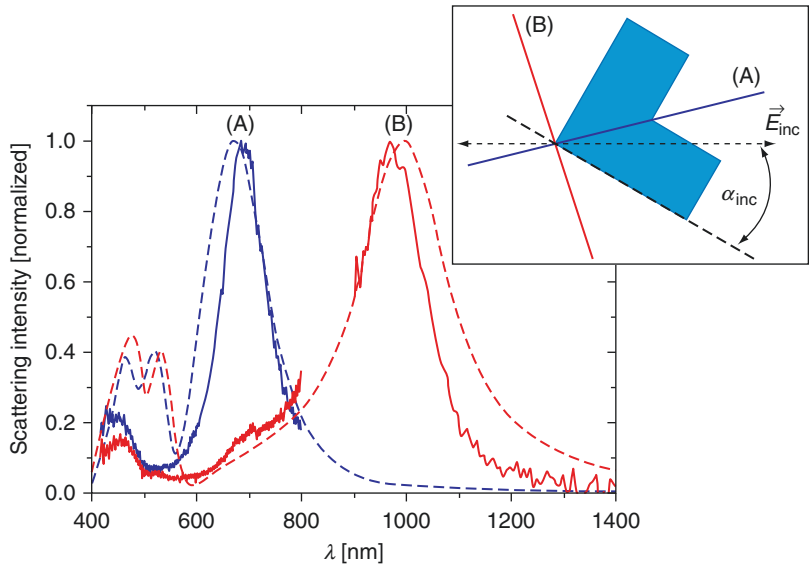
Applications

Up-Down Subwavelength Diffraction Gratings

It has been recently shown that up-down diffraction gratings produced on thin metal films may lead to high-quality optical resonances [10]. The latter have a wide variety of immediate applications ranging from surface-enhanced Raman spectroscopy to prototypes of nanoscale optical cavities for future nano-lasers. It is hence important to be able to predict optical response of such systems utilizing numerical simulations. Figure 6 demonstrates the efficiency of the FDTD technique and its ability to capture quantitatively general behavior of EM surface waves. Here experimental measurements are compared directly with FDTD data for up-down silver diffraction grating shown schematically in the inset. Due to symmetric dielectric environment, the grating exhibits two well-resolved resonant peaks corresponding to the symmetric and anti-symmetric surface modes. It should be noted that the resonance at high energy (short wavelength) becomes narrower when the effective thickness of the film decreases resulting in Q around 5,000 for thicknesses of 7 nm or smaller. The agreement between experiment and FDTD simulations is nearly ideal for the positions of these modes. Most importantly, widths and lines shapes obtained via FDTD are almost identical to those observed in the experiment.

Optical Properties of L-Shaped Nanoparticles

Several experimental groups have demonstrated that arrays of NPs with no center of inversion symmetry exhibit birefringence modifying the polarization of the incident radiation. For example, L-shaped NPs have two well-defined resonant



Finite-Difference Time-Domain Technique, Fig. 7 Scattering cross-section of a single silver L-shaped NP as a function of the incident wavelength in nanometers. FDTD simulations are shown as *dashed lines*, while experimental measurements are represented as *solid lines*. *Blue* (*dark*) color corresponds to the incident field polarized along the *blue* (*dark*) axis of symmetry (see the *inset*); *red* (*bright*) color shows the data for the polarization along the *red* (*bright*) axis. The *inset* shows the *top* view of the nanostructure

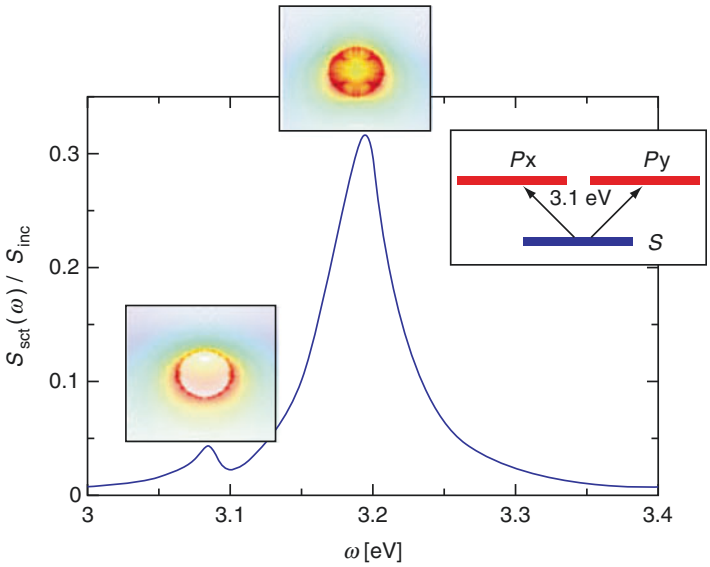
bands associated with the axes of symmetry (see the inset of Fig. 7). If the polarization of a pump pulse is not along one of these axes, the total EM field becomes elliptically polarized on the output side of the array [11]. Direct comparison of three-dimensional FDTD simulations with experimental measurements is presented in Fig. 7, again demonstrating a good agreement.

Classical Electrodynamics of Quantum Systems

All these and many other intriguing applications, however, mainly utilize linear optical response of nanomaterials. Most of the research has been centered at linear optics of metal nanostructures with several attempts to go beyond linear regime and consider nonlinear properties of materials including noble metal NPs aiming for a wide variety of nonlinear phenomena to appear. A newly emerging research field of nonlinear plasmonics follows two obvious routes, namely, combining linear metal nanosystems with highly nonlinear media or taking into account strongly inhomogeneous EM fields associated with EM surface waves that lead to spatial dependence of density of the conductive electrons in metals and hence result in nonlinear phenomena such as second harmonic generation [12]. The first approach deals with nonlinear media at the macroscopic level while even a single atom or molecule located near nanostructure at resonant conditions may be considered as a nonlinear optical system due to high EM near-fields. Thus, it is important to develop approach capable of capturing both size effects at the nanoscale and the time dynamics of EM fields. Generally speaking, such approach has to take into account Maxwell's equations and quantum dynamics of atoms or molecules in the vicinity of nanomaterials. The latter can be accomplished by coupling the Liouville equation that governs the evolution of the density matrix to Maxwell's equations via a polarization current term [13]. The resulting system of coupled equations thus contains the most accurate description of nonlinear phenomena at the nanoscale.

The goal is to couple Maxwell's equations (Eq. 1) to the equation that describes the dynamics of quantum system (such as a gas of cold atoms). In the presence of decoherence and with an assumption that the density of quantum gas is low enough to neglect Coulomb interaction between atoms, one can describe the quantum dynamics using the Liouville equation for a single atom

$$i\hbar \frac{d\hat{\rho}}{dt} = \left[\hat{H}, \hat{\rho}\right] - i\hbar\hat{\Gamma}\hat{\rho}, \tag{9}$$



Finite-Difference Time-Domain Technique, Fig. 8 Scattering cross-section of the atomic cluster as a function of the incident frequency in electron Volts. The *left inset* shows the energy level diagram. The *rightmost inset* depicts steady-state EM energy distribution associated with low-energy resonance (surface mode – analogue of the localized surface plasmon-polariton). The inset in the *middle* demonstrates the collective volume mode at 3.2 eV

where $\hat{\rho}$ is the density matrix of an atom and $\hat{H}$ is the Hamiltonian containing atom-EM field interaction. The relaxation processes are considered to be Markovian. The diagonal elements of the relaxation matrix $\hat{\Gamma}$ describe lifetime of the excited states, while nondiagonal elements take into account dephasing effects. Further assuming that EM field does not significantly change within a volume occupied by a single atom, the spatial dependence of the density matrix in (Eq. 9) is parametrical via EM field components.

To couple Eqs. 1 and 9, one needs to calculate a macroscopic polarization, $\vec{P}$, according to

$$\vec{P} = n_a \left\langle \vec{d} \right\rangle, \tag{10}$$

where n_a is the number density of atoms and $\left\langle \vec{d} \right\rangle = Tr\left(\hat{\rho}\vec{d}\right)$ is the expectation value of the atomic dipole moment, $\vec{d}$. The resulting system of coupled Maxwell-Liouville equations is then solved numerically on a computer. Note that the dipole–dipole interactions between atoms (without *nearest neighbor* approximation) are automatically included in this approach.

The numerical implementation of the proposed scheme is as follows:

1. In the regions occupied by atoms, Maxwell's equations are solved utilizing FDTD algorithm. First, magnetic field is updated according to Faraday's law. Next, using Ampere's law, electric field is updated with the macroscopic polarization current density (time derivative of Eq. 10), which uses density matrix at the previous time step.
2. With the knowledge of electric field components (stored in memory at two time steps), the density matrix is updated at each spatial point on the grid according to Eq. 9 using the fourth order Runge–Kutta scheme.
3. Finally, with the knowledge of electric field components and updated density matrix, one calculates the macroscopic polarization current at each grid point. Note that the time derivative of Eq. 10 in many cases can be written analytically, which simplifies calculations and improves numerical stability.

Figure 8 presents results of numerical integration of the coupled two-dimensional Maxwell-Liouville equations for the atomic cluster with a radius of 35 nm. The cluster is assumed to be comprised of two-level atoms with a number density corresponding to that of the ideal gas. It should be noted that in order to take into account all possible polarizations of EM field, excited states of each atom have to be doubly degenerate as depicted in the inset of Fig. 8.

The peak amplitude of the incident field is low such that the system exhibits only linear response. Maxwell-Liouville equations are propagated in time for as long as several picoseconds. The resulting EM fields are used to calculate scattering cross-section at a given incident frequency. Simulations are repeated for several frequencies. The data obtained show that the atomic cluster exhibits two resonances (unlike for an individual atom, where one would observe a single fluorescent spike) illustrating a significance of collective effects.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Finite-Difference Frequency-Domain Technique
- ▶ Optical Properties of Metal Nanoparticles

References

1. Hutter, E., Fendler, J.H.: Exploitation of localized surface plasmon resonance. Adv. Mater. **16**, 1685 (2004)
2. Murray, W.A., Barnes, W.L.: Plasmonic materials. Adv. Mater. **19**, 3771 (2007)
3. Kreibig, U., Vollmer, M.: Optical Properties of Metal Clusters. Springer, Berlin (1995)
4. Taflove, A., Hagness, S.C.: Computational Electrodynamics: The Finite-Difference Time-Domain Method, 3rd edn. Artech House, Boston (2005)
5. Yee, K.S.: Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. **14**, 302 (1966)
6. Berenger, J.P.: Perfectly matched layer (PML) for computational electromagnetics. In: Balanis, C.A. (ed.) Synthesis Lectures on Computational Electromagnetics. Morgan and Claypool, San Rafael (2007)
7. Pacheco, P.S.: Parallel Programming with MPI. Morgan Kaufmann, San Francisco (1993)
8. Pernice, W.H.P., Payne, F.P., Gallagher, D.F.G.: An FDTD method for the simulation of dispersive metallic nanostructures. Opt. Quant. Electron. **38**, 843 (2006)
9. Rakic, A.D., Djurisic, A.B.., Elazar, J.M., Majewski, M.L.: Optical properties of metallic films for vertical-cavity optoelectronic devices. Appl. Opt. **37**, 5271 (1998)
10. Sukharev, M., Sievert, P.R., Seideman, T., Ketterson, J.B.: Perfect coupling of light to surface plasmons with ultra-narrow linewidths. J. Chem. Phys. **131**, 034708 (2009)
11. Sung, J., Sukharev, M., Hicks, E.M., Van Duyne, R.P., Seideman, T., Spears, K.G.: Nanoparticle spectroscopy: birefringence in 2D arrays of L-shaped silver nanoparticles. J. Phys. Chem. C **112**, 3252 (2008)
12. Zeng, Y., Hoyer, W., Liu, J., Koch, S.W., Moloney, J. V.: Classical theory for second-harmonic generation from metallic nanoparticles. Phys. Rev. B **79**, 235109 (2009)
13. Mukamel, S.: Principles of Nonlinear Optical Spectroscopy. Oxford University Press, New York (1995)

Finite-Element Analysis (FEA)

▶ Finite Element Methods for Computational Nano-optics

First Principles Calculations

▶ Ab Initio DFT Simulations of Nanostructures

First Principles Molecular Dynamics

▶ Car–Parrinello Molecular Dynamics

First-Principle Calculations

▶ DNA from First Principles

First-Principles Approaches

▶ Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches

First-Principles Quantum Monte Carlo

Lucas K. Wagner
Department of Physics, University of Illinois, Urbana, IL, USA

Definition

A computational method that uses Monte Carlo techniques to solve many-body quantum problems. There are a number of different quantum Monte Carlo techniques, which vary from exact to approximate solution of the Schrödinger equation and can be applied to many different types of Hamiltonians.

Overview

There are many different types of quantum Monte Carlo techniques. Which one to use can be intimidating to a novice; however, there are actually only a few different concepts that are expressed in different languages for different problems.

Concept: Sampling

In a many-body quantum problem, the fundamental object is the many-body wave function $\Psi(\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \ldots)$ or, for finite temperature problems, the many-body density matrix $\rho(\mathbf{r'}_1, \mathbf{r'}_2, \ldots; \mathbf{r}_1, \mathbf{r}_2, \ldots)$. These objects are thus scalar functions on the space $\mathcal{R}^{3N}$ or $\mathcal{R}^{6N}$ in three dimensions, where N is the number of particles. Grid-based techniques require exponential numbers of points with respect to the number of particles to represent these functions, which is computationally infeasible.

The fundamental concept in QMC is that the wave function can be represented implicitly by the positions of a set of samples whose positions are chosen randomly. Observables are then measured on this set of samples, which are viewed stochastically. The central limit theorem is invoked to find the statistical error $\varepsilon = \sigma/\sqrt{N_s}$, where σ is the standard deviation of the quantity and N_s is the number of samples. The error is independent of the dimensionality of the problem. Stochastic sampling is a strategy that enables an algorithm to treat very high-dimensional spaces while potentially scaling weakly with the system size, assuming that the standard deviation σ is well controlled.

When evaluating a quantum Monte Carlo technique, one of the first objectives is to characterize what distribution the algorithm samples. This directly affects what quantities can be evaluated. The behavior of σ is also important. There are some problems for which Monte Carlo is not applicable because σ is not finite.

Concept: Monte Carlo Integration

Many high-dimensional problems can be written as an integration:

$$\int I(\mathbf{R}) d\mathbf{R}, \tag{1}$$

where $\mathbf{R}$ stands for the collective many-body coordinate $\mathbf{R} = \{r_1, r_2, \ldots\}$. If samples are generated with probability density $\rho(\mathbf{R})$, then the expectation value of any observable $A(\mathbf{R})$ on that sample set is

$$\langle A \rangle_\rho = \int A(\mathbf{R}) \rho(\mathbf{R}) d\mathbf{R}. \tag{2}$$

It is thus possible to write any integral as

$$\int I(\mathbf{R}) d\mathbf{R} = \int \frac{I(\mathbf{R})}{\rho(\mathbf{R})} \rho(\mathbf{R}) d\mathbf{R} = \left\langle \frac{I}{\rho} \right\rangle_\rho. \tag{3}$$

This equation suggests the Monte Carlo integration strategy: generate samples with some probability density ρ and accumulate the integral as the average of $\frac{I(\mathbf{R})}{\rho(\mathbf{R})}$.

There are several algorithms that can efficiently generate random samples given a density ρ. The most common is the algorithm of Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller [6], often called the Metropolis algorithm.

Concept: Importance Sampling

The stochastic uncertainty ε is proportional to the standard deviation σ of the quantity $\frac{I}{\rho}$ in Eq. 3. Since ε is inversely proportional to the square root of the number of independent samples, the cost of a Monte Carlo calculation to obtain a given uncertainty rises as the variance σ^2. Because of this, the ultimate efficiency of a Monte Carlo algorithm is strongly affected by the choice of the sampled function ρ. The concept of importance sampling is to choose ρ so that σ^2 is as small as possible. This technique can increase the efficiency of a Monte Carlo method by many orders of magnitude.

One particularly spectacular example of importance sampling is the case of evaluating the expectation value of the energy giving some wave function $\Psi(\mathbf{R})$. This integral is

$$\langle E \rangle = \frac{\int \Psi^*(\mathbf{R}) \hat{H} \Psi(\mathbf{R}) d\mathbf{R}}{\int |\Psi(\mathbf{R})|^2 d\mathbf{R}} \tag{4}$$

The typical choice of ρ is

$$\rho(\mathbf{R}) = \frac{|\Psi(\mathbf{R})|^2}{\int |\Psi(\mathbf{R})|^2 d\mathbf{R}}, \tag{5}$$

which results in the integral

$$\int \frac{\hat{H}\Psi(\mathbf{R})}{\Psi(\mathbf{R})} \rho(\mathbf{R}) d\mathbf{R} \tag{6}$$

The variance of the quantity $\frac{\hat{H}\Psi(\mathbf{R})}{\Psi(\mathbf{R})}$ has the special property that for an eigenfunction of the Hamiltonian, it is a constant. Therefore, as the wave function approaches an eigenfunction, the variance approaches zero, and the efficiency of the Monte Carlo process increases as well.

Flavors of Quantum Monte Carlo

For nanotechnology, there are at least two large classes of problems that may be served using quantum Monte Carlo techniques. The first is the ab initio or first-principles problem, in which it is required that a highly accurate prediction be made using only fundamental constants and atomic positions. The second is the model problem, in which there is an approximate low-energy problem in some usually discrete Hilbert space. While the algorithms themselves are often very similar to one another in both these cases, the considerations around the applications of the algorithms and the feasibility of sophisticated methods can result in quite different calculations. This entry will focus on the most common techniques used when the Hamiltonian to be solved is the first-principles one; that is,

$$\hat{H} = -\sum_i \frac{1}{2}\nabla_i^2 - \sum_{\alpha i} \frac{Z_\alpha}{r_{i\alpha}} + \sum_{\alpha\beta} \frac{Z_{\alpha\beta}}{r_{\alpha\beta}} + \sum_{ij} \frac{1}{r_{ij}}, \tag{7}$$

where atomic units ($\hbar = 1$, $m_e = 1$, $e = 1$) are used; i, j refer to electron indices; and α, β refer to nuclear indices. Notable techniques that are not covered in this entry include full configuration interaction quantum Monte Carlo [1] (FCI-QMC) and auxiliary field quantum Monte Carlo [7] (AF-QMC). More details are available in a review article [4].

Variational Monte Carlo (VMC)

VMC is a direct application of the variational method to obtain ground states. The wave function is expressed as a function of the many-body coordinate R and some parameters P. Then the expectation value of the energy is evaluated as

$$\langle E(P) \rangle = \frac{\int \Psi^*(\mathbf{R}, P) \hat{H} \Psi(\mathbf{R}, P) d\mathbf{R}}{\int |\Psi(\mathbf{R}, P)|^2 d\mathbf{R}} \tag{8}$$

using Monte Carlo techniques as explained above. The advantage of using Monte Carlo techniques is that the integral can be evaluated for any wave function for which the value and its derivatives can be evaluated efficiently.

A paradigmatic example of a wave function enabled by the VMC method is the Slater-Jastrow form:

$$\Psi(\mathbf{R}) = \mathrm{Det}\left[\phi_i(\mathbf{r}_j)\right] \exp\left(\sum_{\alpha, i\neq j} u(r_{i\alpha}, r_{j\alpha}, r_{ij})\right). \tag{9}$$

The first term in Eq. 9 is a Slater determinant of one-particle orbitals ϕ_i. Typically these orbitals are obtained from either a Hartree-Fock or density functional theory calculation. They may be optimized as part of the VMC procedure. The second term in Eq. 9 is called the Jastrow factor and explicitly includes the electron–electron distances. Typically, the function u is expanded as follows:

$$u(r_{i\alpha}, r_{j\alpha}, r_{ij}) = \sum_k c_k^{ei} a_k(r_{i\alpha}) + \sum_k c_m^{ee} b_k(r_{ij}) + \sum_{klm} c_{klm}^{eei}\left(a_k(r_{i\alpha})a_l(r_{j\alpha}) + a_k(r_{j\alpha})a_l(r_{i\alpha})\right)b_k(r_{ij}),$$

where the Greek letters stand for the index of the nuclei and a_k/b_k are one-dimensional functions. With the Jastrow factor, Eq. 8 becomes nonseparable, and Monte Carlo is needed to numerically evaluate the integral.

There are two main options to optimize the wave function. The first is direct energy optimization. This approach is surprisingly difficult to achieve and was only fully developed in the 2000s. Modern energy optimization techniques can handle thousands of parameters efficiently. Variance optimization is an older but still effective technique to optimize wave functions. It is directly related to the zero variance principle above. The variance optimization technique minimizes the variance of the local energy $E_L(\mathbf{R}) = \hat{H}\Psi(\mathbf{R})/\Psi(\mathbf{R})$ by varying the wave function parameters. While variance optimization is typically more efficient than energy optimization, it cannot efficiently optimize parameters that change the nodes of the wave function.

Common wave function forms: MSJ, backflow, pairing.

Convergence parameters: Number of parameters, accuracy of the optimization.

Main approximation: For a given Hamiltonian and boundary conditions, the primary approximation in VMC methods is the wave function *ansatz*. If this is close to the exact wave function, then the approximation will be small. If there is important physics not contained in the parameterization, the approximation will be larger. Simpler ground states may be artificially preferred over complex ground states because they are easier to parameterize.

Exact when: The wave function parameterization includes the exact wave function in its Hilbert space.

F

Projector Monte Carlo

Diffusion Monte Carlo (DMC) and reptation Monte Carlo (RMC) are based on the so-called imaginary time Schrödinger equation

$$-\frac{d\Psi(\mathbf{R}, \tau)}{d\tau} = (H - E_0)\Psi(\mathbf{R}, \tau), \tag{10}$$

which has a steady-state solution when Ψ is an eigenfunction and E_0 is adjusted to the eigenvalue. Non-steady-state solutions converge exponentially quickly to the eigenstate Φ_0 as τ goes to infinity.

In the language of Green's functions, the propagation of $\Psi(\mathbf{R}, \tau)$ is given by

$$\Psi(\mathbf{R}, \tau + \Delta\tau) = \int G(\mathbf{R}, \mathbf{R}', \Delta\tau)\Psi(\mathbf{R}', \tau)d\mathbf{R}'. \tag{11}$$

If the function $G(\mathbf{R}, \mathbf{R}', \Delta\tau) = \langle \mathbf{R}| \exp(-\Delta\tau\hat{H})|\mathbf{R}'\rangle$ were known, then one could use Monte Carlo integration to evaluate Eq. 11. It turns out that excellent approximations exist for G when $\Delta\tau$ is small. Then, since $\exp(-\Delta\tau\hat{H}) = \exp(-\Delta\tau\hat{H}/n)^n$, any large $\delta\tau$ can be expressed as multiple nested integrals; that is,

$$\Psi(\mathbf{R}, \tau + \Delta\tau) = \int G(\mathbf{R}, \mathbf{R}_1, \Delta\tau/n)\ldots G(\mathbf{R}_{n-1}, \mathbf{R}_n, \Delta\tau/n)\Psi(\mathbf{R}', \tau)d\mathbf{R}_1\ldots d\mathbf{R}_n. \tag{12}$$

This is a very high-dimensional integration over $3N_e n$ dimensions, where a typical value for n to ensure convergence might be 100–400.

The projector Monte Carlo techniques are thus based on performing the integral in Eq. 12 using Monte Carlo techniques. The reptation or path integral ground state methods perform this operation explicitly, by sampling a random variable in the space $\mathcal{R}^{3N_e n}$ with probability proportional to the integrand in Eq. 12. It is here that one first encounters the *fermion sign problem*: for fermion particles, the integrand is not necessarily positive and cannot be represented as a probability distribution. In fact, the sign of the integrand can fluctuate wildly, which ultimately causes the variance of any observable to increase exponentially as a function of the system size. Therefore, for fermions, an approximation is needed to make the simulations feasible.

Instead of using the bare projectors in Eq. 12, one typically uses the importance-sampled ones. In this technique, the starting distribution is proportional to $\Psi_T^2(\mathbf{R})$, where Ψ_T is the trial function (typically the Slater-Jastrow wave function). The projector is then modified to become the importance-sampled Green's function:

$$\tilde{G}(\mathbf{R}, \mathbf{R}', \Delta\tau) = \Psi_T(\mathbf{R}) G(\mathbf{R}, \mathbf{R}', \Delta\tau) / \Psi_T(\mathbf{R}'). \tag{13}$$

The final distribution sampled is the mixed distribution $\Psi_T(\mathbf{R})\Phi_{FN}(\mathbf{R})$, where $\Phi_{FN}(\mathbf{R}) = \Psi(\mathbf{R}, \infty)$, as one can see by performing the integral

$$\begin{aligned} &\int \tilde{G}(\mathbf{R}, \mathbf{R}', \Delta\tau)\Psi_T^2(\mathbf{R}')d\mathbf{R}'. \\ &= \Psi_T(\mathbf{R}) \int G(\mathbf{R}, \mathbf{R}', \Delta\tau)\Psi_T(\mathbf{R}')d\mathbf{R}' \\ &= \Psi_T(\mathbf{R})\Psi(\mathbf{R}, \Delta\tau). \end{aligned} \tag{14}$$

It can be shown that $\tilde{G}$ is much more easily approximated than G for finite timesteps, if the trial function Ψ_T is accurate [3]. The importance sampling transformation also makes the *fixed-node* approximation apparent; if the nodes (zeros) of Φ_{FN} are constrained to be the same as the nodes of Ψ_T, then the final distribution $\Psi_T(\mathbf{R})\Phi_{FN}(\mathbf{R})$ is guaranteed to be positive, and there is no sign problem at the cost of an approximation.

Convergence parameters: Time descretization.

Main approximation: For bosons, expectation values on the ground state can be found exactly. For small systems ($\lesssim 10$) of few fermions, the sign problem can be circumvented with sufficient computer time. For larger systems of fermions, fixed-node or similar approximations must be used.

Exact when: The ground state nodal surface is known.

Path Integral Monte Carlo

Path Integral Monte Carlo (PIMC) is closely related to the projector methods. In this case, the integrals to be evaluated are traces over the density matrix $\exp(-\beta\hat{H})$, where $\beta = 1/k_B T$, the inverse temperature. An expectation value of a local operator $\mathcal{O}(\mathbf{R})$ can be evaluated as

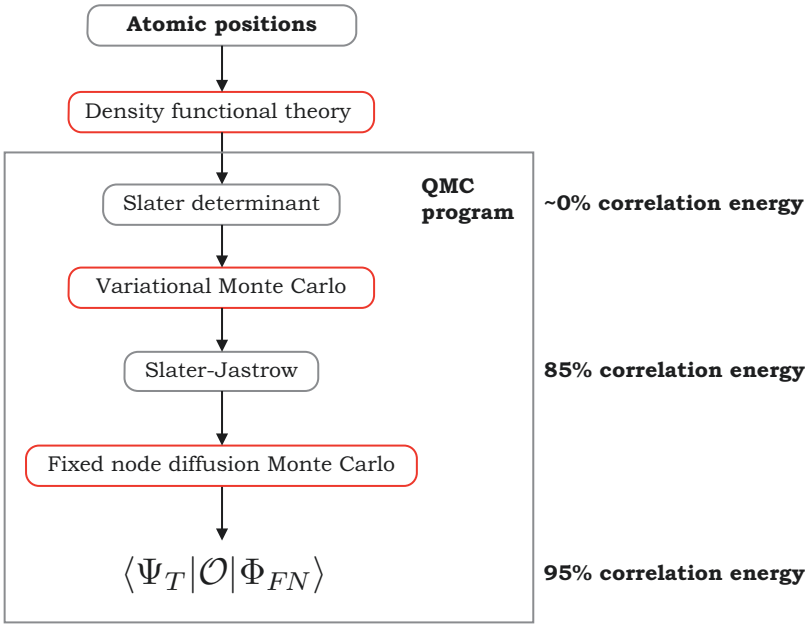
$$\langle \mathcal{O} \rangle = \int \langle \mathbf{R} | \exp(-\beta\hat{H}) | \mathbf{R} \rangle \, \mathcal{O}(\mathbf{R}) d\mathbf{R}. \tag{15}$$

This is very similar to Eq. 12 with $\Delta\tau = \beta$, except that the initial and final positions are the same, and there is no wave function.

Very similar considerations apply to the evaluation of Eq. 15, in terms of the sign problem and expanding the matrix elements as products of smaller steps. For lower temperature, β is larger, which corresponds to longer projection time. This then corresponds to a larger value of n, which are called "beads" in the language of PIMC.

Practical Calculations on First-Principles Models

The majority of quantum Monte Carlo calculations on first-principles models use a standard process, illustrated in Fig. 1. The most basic procedure is outlined here; there are many elaborations on this skeleton. First, a density functional theory (DFT) calculation is performed. The purpose of the DFT calculation is to generate a set of



First-Principles Quantum Monte Carlo, Fig. 1 Typical workflow for quantum Monte Carlo calculations of first-principles models

orbitals that can be used to create a Slater determinant starting wave function. The density functional itself is not relevant to the QMC calculation, other than the orbitals that it produces. Then, the Jastrow factor (Eq. 9) is added to the wave function and optimized using variational Monte Carlo techniques to form the Slater-Jastrow wave function. Finally, to improve the accuracy further, diffusion Monte Carlo is performed using the Slater-Jastrow wave function as the trial function.

Approximations in Diffusion Monte Carlo

The workflow outlined above has only a few well-defined approximations. They are summarized in Table 1. The controlled approximations can be removed in a polynomially scaling amount of time as listed. Once that is done, there are two main approximations: pseudopotentials and fixed node. Pseudopotentials are an approximation in the Hamiltonian, while fixed node is an approximation in the solution of the Hamiltonian. The relative importance of these two has been tested on small molecules by removing the fixed-node error. It seems that at least half of the error is due to fixed node, if Hartree-Fock pseudopotentials are used.

First-Principles Quantum Monte Carlo, Table 1 Approximations in the procedure outlined in Fig. 1

Approximation	Resolution
Controllable	
Timestep	Reduce the factor $\Delta\tau/n$ in Eq. 12 until the properties are converged
Finite size	In periodic models of infinite systems, increase the cell size used to simulate the system
Finite population	Use large numbers of walkers to perform the calculation
Uncontrolled	
Fixed node	Use a higher accuracy trial function
Pseudopotential	Test the pseudopotential carefully

What Can Be Calculated Using Quantum Monte Carlo Techniques

The fixed-node diffusion Monte Carlo can calculate expectation values of the following form:

$$\langle \Psi_T | \mathcal{O} | \Phi_{FN} \rangle, \tag{16}$$

where $\mathcal{O}$ is a quantum operator, Ψ_T is the trial function, and Φ_{FN} is the fixed-node wave function. This is called the mixed estimator. For the calculation of energies, this equates to the pure estimator

$$\langle \Psi_T | \hat{H} | \Phi_{FN} \rangle = \langle \Phi_{FN} | \hat{H} | \Phi_{FN} \rangle. \quad (17)$$

For other quantities, the mixed estimator is not exact, but can be corrected using the extrapolated estimator:

$$\langle \Phi_{FN} | \mathcal{O} | \Phi_{FN} \rangle \simeq 2 \langle \Psi_T | \mathcal{O} | \Phi_{FN} \rangle - \langle \Psi_T | \mathcal{O} | \Psi_T \rangle. \quad (18)$$

The mixed estimator can be written as follows:

$$\frac{\int \Psi_T^*(\mathbf{R}, P) \mathcal{O} \Phi_{FN}(\mathbf{R}, P) d\mathbf{R}}{\int \Psi_T^*(\mathbf{R}) \Phi_{FN}(\mathbf{R}) d\mathbf{R}}. \quad (19)$$

The final probability distribution of a diffusion Monte Carlo calculation is

$$\frac{\Psi_T^*(\mathbf{R}, P) \Phi_{FN}(\mathbf{R})}{\int \Psi_T^*(\mathbf{R}) \Phi_{FN}(\mathbf{R}) d\mathbf{R}}. \quad (20)$$

Performing some algebra, one can then obtain that

$$\langle \Psi_T | \mathcal{O} | \Phi_{FN} \rangle = \left\langle \frac{\mathcal{O} \Psi_T(\mathbf{R})}{\Psi_T(\mathbf{R})} \right\rangle_{\Phi_{FN} \Psi_T}. \quad (21)$$

Therefore, any static quantity can be calculated from the set of diffusion Monte Carlo walkers simply by calculating the operation of the quantum operator on the trial function and averaging over configurations.

It is a common practice to use an excited Slater determinant as a trial function to access excited states. If an appropriate symmetry can be identified that separates the ground and excited state, then this procedure results in a rigorous upper bound to the excited state energy [3]. Even in the cases where this does not occur, this method has been shown to obtain accurate estimates of the first excited state energies of solids and molecules [4, 5].

Implementations

There are a number of open-source or free-to-use packages that implement quantum Monte Carlo algorithms. Notable implementations are CASINO, QMCPACK, and QWalk, each of which has a user community.

Notable Applications

The applications listed here are chosen to give a sample of what these methods can be used to do. It is not exhaustive.

Homogeneous Electron Gas

One of the first applications of the diffusion Monte Carlo method was to compute the ground state energy of the interacting homogeneous electron gas as a function of density [2]. Because the sign problem in this system is relatively small, it was possible to obtain very accurate energies. These energies form the basis of all modern density functional theories.

High Accuracy Thermodynamics

When calculating thermodynamical quantities, one must evaluate the energy differences between different nuclear configurations. Typically these quantities are dominated by the change in internal energy, which is calculated by electronic structure methods. However, the error in these methods from the true ground state is often much larger than the change in internal energy. The methods rely on cancelation of errors, that is, that the error in the ground state of the atomic systems is approximately the same as the error in the condensed system. The projector methods, such as diffusion Monte Carlo, often obtain the lowest variational energies attainable for large systems. There is thus less error to cancel, and accuracy is much higher.

Strongly Correlated Systems

The Monte Carlo methods also occupy a unique niche which is ideal for strongly correlated condensed matter systems. They explicitly simulate the electron interactions, which can qualitatively change the physics in these materials and can treat periodic models of materials.

Cross-References

- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Computer Modeling and Simulation of Materials
- ▶ Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches
- ▶ First Principles Calculations

References

1. Booth, G.H., Thom, A.J.W., Alavi, A.: Fermion Monte Carlo without fixed nodes: a game of life, death, and annihilation in slater determinant space. J. Chem. Phys. **131**(5), 054106–054106–10 (2009)
2. Ceperley, D.M., Alder, B.J.: Ground state of the electron gas by a stochastic method. Phys. Rev. Lett. **45**(7), 566–569 (1980)
3. Foulkes, W.M.C., Hood, R.Q., Needs, R.J.: Symmetry constraints and variational principles in diffusion quantum Monte Carlo calculations of excited-state energies. Phys. Rev. B **60**(7), 4558–4570 (1999)
4. Foulkes, W.M.C., Mitas, L., Needs, R.J., Rajagopal, G.: Quantum Monte Carlo simulations of solids. Rev. Mod. Phys. **73**(1), 33–83 (2001)
5. Koloren, J., Mitas, L.: Applications of quantum Monte Carlo methods in condensed systems. Rep. Prog. Phys. **74**(2), 026502 (2011)
6. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. J. Chem. Phys. **21**(6), 1087–1092 (1953)
7. Zhang, S., Krakauer, H.: Quantum Monte Carlo method using phase-free random walks with slater determinants. Phys. Rev. Lett. **90**(13), 136401 (2003)

Flapping-Wing Flight

- ▶ Insect Flight and Micro Air Vehicles (MAVs)

Flash Evaporation Liquid Atomization

- ▶ Spray Technologies Inspired by Bombardier Beetle

Flexible Electronics

Muhammad A. Alam[1] and Satish Kumar[2]
[1]School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA
[2]G. W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

F

Synonyms

Bendable electronics; Conformal electronics

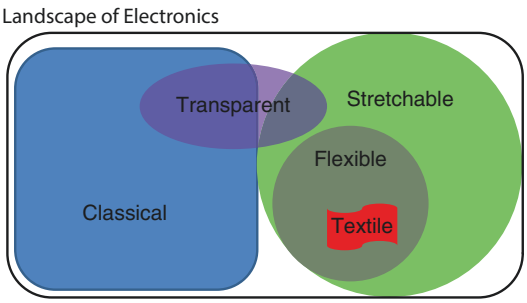
Definition of Flexible Electronics

Traditional electronics like integrated circuits or solar cells are generally built on thick inflexible substrates. Instead, flexible electronics – built on substrates like plastic or metallic foil – can be folded, wrapped, rolled, and twisted with negligible effect on its electronic function. Flexible electronics is typically composed of a bilayer of thin passive substrate (e.g., plastic, textile, etc.) topped with a second layer of active electronic components. In some applications, additional layers may be used for encapsulation and packaging. If the substrate is both flexible and stretchable for better conformal fit to highly curved surfaces, the system belongs to the broader class of "stretchable electronics" (Fig. 1). When flexible substrate is made of woven fibers like silk, some use a more descriptive term of "textile electronics" [1]. And finally, if both the electronics and substrate are made of optically transparent material – such as graphene or network of carbon nanotubes [2] on glass – it may be referred to as "transparent electronics" (Fig. 1).

History

Although the cables and connectors – thin metallic wires coated by plastic insulation – have always been flexible, it took the emergence of conductive polymers (doped organic

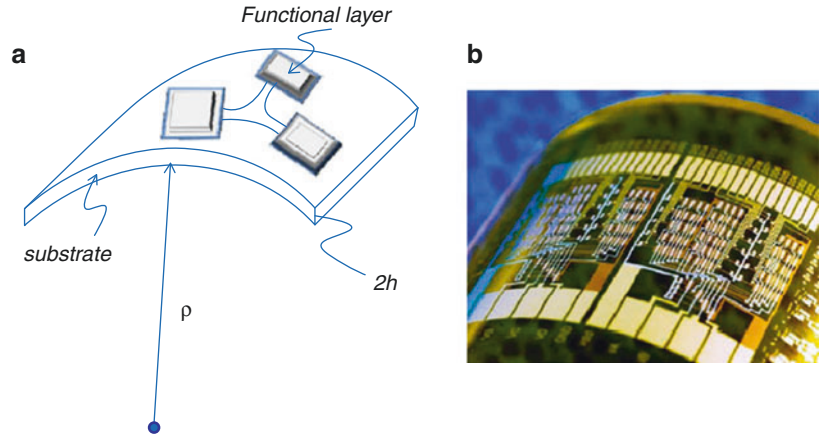
semiconductors) and development of doped α-Si in the 1970s for applications in photocopier, solar cells, and displays to pave the way for flexible electronics. The 1990s saw several demonstrations of the flexible electronics and optoelectronic components and systems based on various types of organic polymers. Bell Labs demonstrated fully functional flexible ICs with several thousand TFTs, and researchers at Kodak, Cavendish Laboratory, and UCSB prototyped flexible organic photovoltaics (OPVs) and organic light emitting diodes (OLEDs) on a plastic substrate. Since 2000, broad/innovative use of soft-lithography and transfer printing – originally developed in the eighteenth century to transfer images engraved in Cu plates via special paper to nonplanar porcelain substrates – has broaden the range of materials suitable for the flexible electronics. Today, transfer printing [3] can be used to create flexible electronic components based on almost any material with potentials applications like low-cost electronic readers, RF identification tags, sensors for structural integrity, etc. Bright, multicolored, flexible OLED displays from Sony and Samsung are approaching commercial viability and several companies are exploring the more difficult price/performance trade-off of OPVs. Cell phones provide a particularly interesting platform and offer many opportunities such as hinged displays and sliding keyboards. Additional applications to continuously monitor the vital signs of a patient as a part of implantable, biocompatible, and organ-conformal flexible electronics have been envisioned, but not fully realized yet.



Flexible Electronics, Fig. 1 *Landscape* of electronics showing classical, transparent, and stretchable regimes. Flexible electronics is subset of the stretchable electronics, and textile electronics is subset of the flexible electronics

Mechanical Properties of Flexible Electronics

Flexible electronics relies on the basic premise that a substrate of thickness $2\ h$ can be bent with the radius of curvature ρ, so long the strain, $\varepsilon = \frac{h}{\rho}$, developed does not exceed the yield strain εc of the film (Fig. 2a). Actually, the physics is no different from that of three-layered composite bows used in Roman and Tartar wars that could be repeatedly stretched thousands of times without loss of function [4]. Indeed, this is why one can fold a 100 μm thick paper or plastic substrate (with elastic yield strength of 0.1 %) around a cylinder with $\rho = 10$ cm with no detrimental



Flexible Electronics, Fig. 2 (**a**) Schematic of a flexible substrate with integrated functional layer. $2\ h$ is the substrate thickness and ρ is the radius of curvature; (**b**) CNT thin-film-transistors-based circuitry printed on a flexible substrate [2]

effects [5]. Typical substrates for flexible electronics include polymers, thin foils of metal, fabrics, and even paper. In addition to flexibility, such substrates should be relatively inexpensive [6], mechanically robust (with reasonably high melting temperature, T_g) and resistant to creep, free from structural defects like pits and pinholes, thermally conductive, but electrically passive. Polymer substrates such as Kaptonz polymide, Polytetraflouroethylene (PTFE), Polyethylene terephthalate (PET), and Polyethylene naphthalate (PEN) have been used for various applications.

Once a thin flexible substrate with adequate structural integrity has been identified, a second layer of electronics is bonded or deposited to the substrate (Fig. 2b). Flexibility is unaffected because the active layer is only several μm thick. Delamination can be avoided if the intra-layer bonding strength γ is sufficiently strong so that the strain (ε) developed due to flexing remains below the critical strain εc of the bilayer, i.e.,

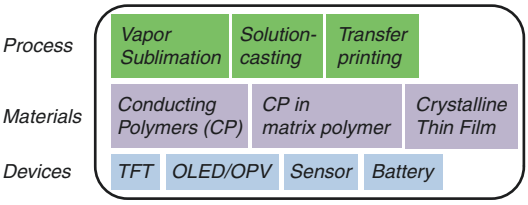
$$\varepsilon < \varepsilon = \varepsilon_p + \sqrt{2\frac{\gamma}{2L}}$$

Here, ε_P is the Poisson buckling ratio [4] and L is the linear dimension of the substrate. For a reasonable bonding strength of the substrate and electronic layers, the device may deform 20–30 times the yield strength without delamination.

Active Layer of Flexible Electronics

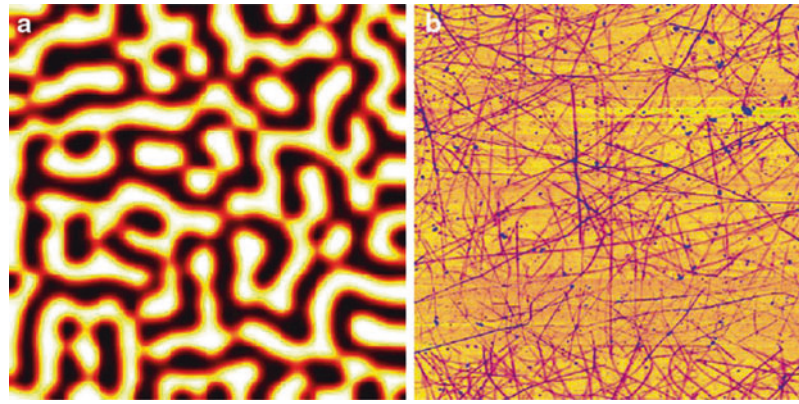
Materials and Devices in the Active Layer

The functional layer of flexible electronics may be composed of electronic and optoelectronic components like TFTs [7, 8], OPVs [9], OLEDs, batteries, sensors, etc. (Fig. 3). The simplest TFTs use active, single-carrier (hole) transport layer (i.e., p-type or acceptor) along with metal contacts and necessary interconnects. Various long-chain polymers like poly(methylmethacrylate) (PMMA), poly p-phenylene vinylene (PPV), polyaniline (PANI), poly



Flexible Electronics, Fig. 3 Processing techniques, materials, and device applications of flexible electronics

(3,4-ethylenedioxythiophene) (PEDOT), poly 3-hexylthiophene (P3HT), as well as smaller organic molecules like pentacene, rubrene, tetracene have been used for TFT applications [7, 8, 10, 11]. The organic solar cells for photovoltaic applications (OPV) are based on blended mixture of p-type and n-type polymers [9, 12], capped by transparent and reflecting electrodes (Fig. 4a). In addition to the p-type polymers discussed above, n-type (donor) polymer used in OPV applications [13] include various fullerene derivatives like (phenyl-[6, 6]-C61)-butyric acid methyl ester (PCBM). In OLEDs, donor and acceptor layers are stacked to create a p-n heterojunction diode. In many OLED applications, Aluminum tris [8-hydrooxquinoline] (Alq3) acts both as electron transport as well as light emission layer. In addition to organic polymers as discussed above, TFT, LED, and PV components of arbitrary complexity (and performance at par with single crystalline counterparts) can also be created by stacks of thin crystalline films by transfer-printing method (see below). Recently, batteries and chemical/biological sensors have been adapted to the flexible substrate. Finally, Au, Ag, and ITO are used to interconnect these active components of the functional layer. ITO films have high electrical conductivity and high optical transparency which make them a good candidate for the transparent electronic applications, but their stiffness and brittleness pose challenge for the flexible electronics. Flexible or stretchable interconnects such as conducting polymers, micro-fabricated thin films, CNT-ionic liquid gels have been explored to maintain the flexibility of the electronics [8, 14].



Flexible Electronics, Fig. 4 (**a**) Blended mixture of polymers [9, 12] (**b**) Network of carbon nanotubes as a channel region of flexible transistors using spin-casting method [13]

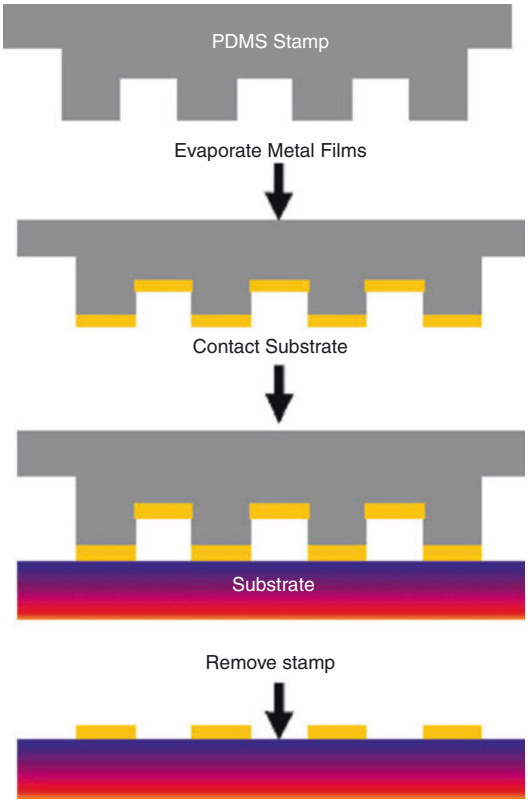
Fabrication of Active Layer

The key challenge in flexible electronics is that the deposition techniques for the functional layer must be compatible with the flexible substrate with low *Tg* and that the weakly bonded "van der Waal" polymers with interatomic forces decays as $\frac{1}{R^6}$ (in contrast, covalent bonding decays as inverse of square radius, i.e., $\frac{1}{R^{62}}$). These deposition techniques can be broadly classified as vapor-phase sublimation, solution casting, or solid state transfer printing (or a combination thereof). The vapor-phase sublimation (VPS) technique can deposit thin films of desired material onto the flexible substrate by thermal sublimation of a nearby (∼few cm apart) solid source. Films of arbitrary complexity can be created by changing the solid source, and intermixing of the layers can be avoided because solid diffusion is strongly suppressed at low temperature. Despite the challenges of uniformity over large areas, VPS has been used for creating large area displays. In the second approach, which is based on the solution casting, the conducting polymers (CPs) [15] are dissolved into a carrier solvent (with appropriate viscosity, surface tension, and anneal temperature) and deposited/cured onto the flexible substrate. Direct deposition of CP onto the flexible substrate (without the carrier solvent) leads to islanding and unacceptable morphology. Variants of solution processing techniques [16] include screen printing, ink-jet printing [17], spin casting, etc. Screen printing technique uses a stencil method where printable materials are first transferred to open areas of a woven mesh or screen and then printable materials are squeezed or rolled to cover the open areas leaving a desired imprint on the substrate [16]. Ink-jet printing creates desired patterns on the substrate by propelling ink droplets of different sizes, while spin casting uses centrifugal force to spin the printing material on the substrate (Fig. 4b).

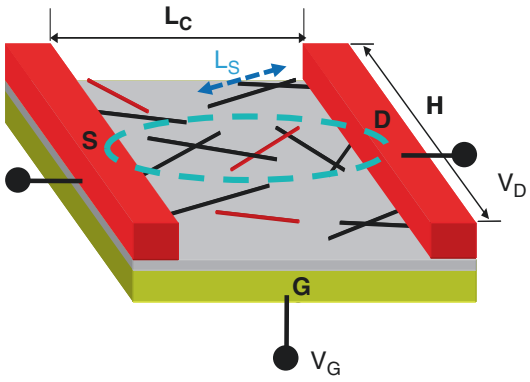
Finally, the solid *transfer-printing* method enormously broadens the range of materials available for flexible electronics [5, 9]. In this technique, high-quality materials are first grown on a temporary donor substrate (Fig. 5). A silicone rubber stamp containing a replica of the desired circuit is first inked by pressing hard and then quickly lifted from the substrate. The inked pad is then softly pressed against the flexible substrate and pulled-away slowly to let the ink stay at the final substrate [3]. All sorts of materials like crystalline Si, CNT, GaAs, etc., have been transferred. Indeed, multiple transfer printing can create substrate of unprecedented complexity.

Electrical Transport of Active Layer

The active layer can be composed of (1) electrically conducting polymer (CP), (2) insulating polymers loaded with conducting fillers, and/or (3) thin films or ribbons of crystalline materials. Advances to the theoretical understanding of these soft materials have been central to the development of flexible electronics [10, 11]. The linear backbone of CP alternates between single and double C-C bonds. The sp^2 hybridization of these C–C bonds produces (1) covalent and localized σ bonds for mechanical stability and (2) out-of-plane,



Flexible Electronics, Fig. 5 Steps of the solid transfer-printing technique [3]



Flexible Electronics, Fig. 6 Schematic of a CNT network transistor. Network of semiconducting (*black*) and metallic (*red*) tubes connecting source (*S*) and drain (*D*) electrodes of the transistor. A percolating pathway of CNTs is indicated by *dashed ellipsoid* region. Channel length (*LC*), tube length (*LS*), device width (*H*), gate voltage (*VG*), and drain voltages (*VD*) are indicated

delocalized, *pz* orbital for electrical conduction (bandwidth 1–3 eV). For long-range conduction, the electrons hop among adjacent chain via weakly coupled *pz* states (0.1 eV). Electrons in these 1D, low-dielectric constant CP interacts strongly with each other and they distort the floppy lattice as they move from site to site, therefore excitonic and polaronic effects are important. In general, the transport in the crystalline organic films could be mapped back to theory of crystalline materials with appropriate redefinition of density of states (DOS), mobility, etc., while transport in amorphous organic molecules – much like α-Si – is often described by "Gaussian disorder model" as electrons hop from one molecule to other in a landscape defined by Gaussian DOS [10].

In many applications such as spin-cast TFTs (Fig. 4b), it is easier to dope a nonconducting polymer matrix (NPM) by conducting fillers (CFs) such as carbon nanotubes (CNT) or other metallic particles. These NPM/CNT network [13, 18] conducts by percolation or hopping at amazingly small fill fraction (<0.2). CFs dictates the electrical performance, while NPM controls the solution properties for ease of printing. Specifically, for the electrical transport, the areal tube-density (*D*) must exceed the percolation threshold (D_c) so that $D > D_c = \frac{4.26^2}{\pi L_s^2}$, where *Ls* is the average length of the conducting polymer or CNTs (Fig. 6). The nonlinear multielectrode percolation theory for TFTs with CNTs or CNT-doped polymers [19] is now well developed and can be succinctly summed as

$$I_D = g\left(\frac{L_c}{L_S}, D_C L_S^2\right) \times f(V_D, V_G)$$

where *ID* is the drain current for a TFT, *VD* and *VG* are drain and gates voltages, and *LC* is the channel length and *LS* is the length of CNTs. The geometric function $g \equiv \left(\frac{L_s}{L_C}\right)^{3\left(DL_S^2\right)}$ and the voltage-dependent function $f \propto C_{ox}[(V_G - V_D)V_T - V_D^2/2]$ are given by classical long-channel TFT formula. Many of the surprises in transport through such system (e.g., (channel) length-dependent mobility!) can be attributed to non-ohmic conduction at percolation threshold.

F

For applications in organic solar cells, both n and p-type organic molecules are co-blended in a solvent and demixed by annealing (through spinodal decomposition) to produce interpenetrating donor/acceptor network [20]. The interface field at the donor/acceptor boundary dissociates the excitons before they are lost to self recombination and the percolating network shepherd electron/hole to their respective contacts. Indeed, if conductive fillers is now added so that it is miscible in n-type polymers, but not in p-type polymers, then double percolation allows further reduction in percolation threshold with a better control of overall morphology. Such double-percolating films based on poly (3octylthiophene) and ethylene-co-vinyl acetate blend has been used for TFT application and is known to percolate at <0.1 fill factors!

The third and final option for active layer involves ribbons of thin crystalline films released from the substrate by chemical etching and then grafting these films onto the flexible substrate by transfer-printing method. Theories of bulk transport apply, however, the roles of surface states and defects associated with the release process are not yet fully understood.

In addition to material used for active transport, flexible electronics demand interconnects either be vapor deposited (Au, Ag, ITO) or ink-jet printed (e.g., Ag ink) at relatively low temperature, subsequently patterned by soft-lithography or benign etching. The transport properties of such thin polycrystalline film have been investigated by many groups, although the reliability of such films under repeating flexing remains to be seen.

Conclusions

Like any other novel technology, the commercial viability of flexible electronics is hard to assess. Mechanical reliability associated with crack formation during repeated flexing, ingress of moisture, etc., and electrical reliability issues related to polymer degradation, continued phase segregation, surface passivation, contact diffusion, and low efficiency are important challenges. Many advances in integrated full color OLED displays from SONY and Samsung suggest that these technical challenges are not impossible to address, the question remains however if they can be addressed economically.

Cross-References

- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanoscale Printing

References

1. Hu, L., Pasta, M., Mantia, F., Cui, L., Jeong, S., Deshazer, H., Choi, J., Han, S., Cui, Y.: Streatchable, porous, and conductive energy textiles. Nano Lett. **10**, 708–714 (2010)
2. Cao, Q., Kim, H., Pimparkar, N., Kulkarni, J., Wang, C., Shim, M., Roy, K., Alam, M., Rogers, J.: Medium-scale carbon nanotube thin-film integrated circuits on flexible plastic substates. Nature **24**, 495–500 (2008)
3. Zaumseil, J., Meitl, M.A., Hsu, J.W.P., Acharya, B.R., Baldwin, K.W., Loo, Y.L., Rogers, J.A.: Three-dimensional and multilayernanostructures formed by nanotransfer printing. Nano Lett. **3**(9), 1223–1227 (2003)
4. Gordon, J.: Structures: Or Why Things Don't Fall Down, 2nd edn. Da Capo Press, Cambridge (2003)
5. Wang, S., Xiao, J., Song, J., Ko, H., Hwang, K.-H., Huang, Y., Rogers, J.: Mechanics of curvilinear electronics. Soft Matter **6**, 5757–5763 (2010)
6. Forrest, S.: The path to ubiquitous and low-cost organic electronic appliances on plastic. Nature **428**, 911–917 (2004)
7. Alam, M., Dodabalapur, A., Pinto, M.: A two dimensional simulation of organic transistors. Trans. Electron Device **44**, 1332–1337 (1997). Special Issue on Organic and Polymeric Devices
8. Lugli, P., Erlen, C., Pecchia, A., Brunetti, F., Latessa, L., Bolognesi, A., Csaba, G., Scarpa, G., Carlo, A.: The simulation of molecular and organic devices: a critical review and look at future development. Appl. Phys. A. **87**, 593–598 (2007)
9. Peumans, P., Uchida, S., Forrest, S.: Efficient bulk heterojunction photovoltaic cells using small-molecular-weight organic thin films. Nature **425**, 158–162 (2003)
10. Yam, P.: Plastics get wired. Sci. Am. **273**, 82–87 (1995)
11. Malliaras, G., Friend, R.: An organic electronics primer. Phys. Today **58**(5), 53–58 (2005)

12. Ray, B., Nair, P., Garcia, R., Alam, M.: Modeling and optimization of polymer-based bulk heterojunction (BH) solar cells. In: Proceedings of the International Electron Devices Meeting, Baltimore (2009)
13. Kumar, S., Murthy, J., Alam, M.: Percolating conduction in finite nanotube networks. Phys. Rev. Lett. **95**(6), 066802 (2005)
14. Rogers, J., Someya, T., Huang, Y.: Materials and mechanics for stretchable electronics. Science **327**, 1603–1606 (2010)
15. Gustafsson, G., Cao, Y., Treacy, G.M., Klavetter, F., Colaneri, N., Heeger, A.J.: Flexible light-emitting diodes made from soluble conducting polymers. Nature **357**, 477 (1992)
16. Pardo, D.A., Jabbour, G.E., Peyghambarian, N.: Application of screen printing in the fabrication of organic light-emitting devices. Adv. Mater. **12**(17), 1249–1252 (2000)
17. Sekitani, T., Noguchi, Y., Zschieschang, U., Klauk, H., Someya, T.: Organic transistors manufactured using inkjet technology with subfemtoliter accuracy. Proc. Natl. Acad. Sci. **105**, 4976–4980 (2008)
18. Kumar, S., Blanchet, G., Hybertsen, M.S., Murthy, J., Alam, M.: Performance of carbon nanotube-dispersed thin-film transistors. Appl. Phys. Lett. **89**, 143501 (2006)
19. Pimparkar, N., Cao, Q., Kumar, S., Murthy, J., Rogers, J., Alam, M.: Current-voltage characteristics of long-channel nanobundle thin-film transistors: a bottom-up perspective. Electron Device Lett. **28**(2), 157–160 (2007)
20. Yu, G., Gao, J., Hummelen, J., Wudl, F., Heeger, A.: Polymer photovoltaic cells: enhanced efficiencies via a network of donor/acceptor heterojunction. Science **270**, 1789–1791 (1995)

Flexible Sensors

▶ Organic Sensors

Flexure Mechanisms

▶ Compliant Mechanisms

Flexures

▶ Compliant Mechanisms

Floccules

▶ Biofilms in Microfluidic Devices

Fluctuating Hydrodynamics

▶ Dissipative Particle Dynamics, Overview

Foams

▶ Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)

Shinji Matsui
Laboratory of Advanced Science and Technology for Industry, University of Hyogo, Hyogo, Japan

Synonyms

Focused-ion-beam-induced deposition (FIB-ID)

Definition

Focused ion beam (FIB) can be used as a deposition tool for fabricating nanostructures. FIB chemical vapor deposition (FIB-CVD) uses ion energy to initiate and localize chemical vapor deposition (CVD) in a specific location by a direct-writing technique. When a local gas ambient is produced near a surface, the impact of energetic ions on that surface can induce chemical reaction. The ion-induced chemistry occurs in an adsorbed layer than in the gas above the surface.

FIB-CVD Process

In 1979, Seliger et al. reported Ga ion beams down to 100-nm diameter with current density in the focal spot of 1.5 A/cm^2 [1]. In 1990, Wargner et al. demonstrated pillars and walls with high aspect ratios achieved using FIB-CVD [2]. In 2000, Matsui et al. demonstrated three-dimensional (3-D) nanostructure fabrication by FIB-CVD [3].

Matsui et al. used two commercially available FIB systems (SMI9200, SMI2050, SII NanoTechnology Inc., Tokyo, Japan) with a Ga^+ ion beam operating at 30 keV. The FIB-CVD used a precursor of *phenanthrene* ($C_{14}H_{10}$) as the source material. The beam diameter of SMI9200 was about 7 nm and that of SMI2050 was about 5 nm. The systems were equipped with the gas nozzles. The nozzles were set a distance of 40 μm from each other and positioned about 300 μm above the substrate surface. The inside diameter of a nozzle was 0.3 mm. The *phenanthrene* gas pressure during pillar growth was typically 5×10^{-5} Pa in the specimen chamber, but the local gas pressure at the beam point was expected to be much higher. The crucible of the source was heated to 85 °C. The FIB is scanned in order to be able to write the desired pattern via computer control, and the ion dose is adjusted to deposit a film of the desired thickness. The experiments were carried out at room temperature on a silicon substrate.

The deposited film was characterized by observing it with a transmission electron microscope (TEM). A thin film of carbon (200 nm thick) was deposited on a silicon substrate by 30 keV Ga^+ FIB using *phenanthrene* precursor gas. The cross sections of the structures created and its electron diffraction patterns were observed by using a 300 kV TEM. There were no crystal structures in the TEM images and diffraction patterns. It was therefore concluded that the deposited film was amorphous carbon (a-C).

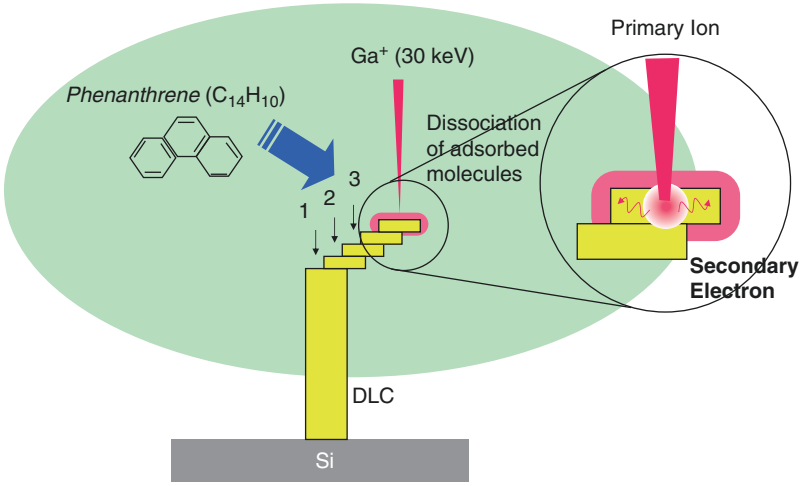
Raman spectra of the a-C films were measured at room temperature with 514.5-nm line of an argon ion laser. A relatively sharp Raman band at 1,550 cm^{-1} and a broad-shouldered band at around 1,400 cm^{-1} were observed in the spectra excited by the 514.5-nm line. Two Raman bands were plotted after Gaussian line shape analysis. These Raman bands, located at 1,550 cm^{-1} and 1,400 cm^{-1}, originate from the trigonal (sp^2) bonding structure of graphite and tetrahedral (sp^3) bonding structure of diamond. This result suggests that the a-C film deposited by FIB-CVD is a diamond-like amorphous carbon, which has attracted attention due to its hardness, chemical inertness, and optical transparency.

The coordination of carbon atoms in the carbon-based material formed by CVD deposition of *phenanthrene* assisted by Ga FIB was investigated by the measurement of near-edge X-ray absorption fine-structure spectra of the carbon *K*-edge over the excitation energy range 275–320 eV [4]. Novel peak observed at 289.0 eV was assigned to the 1 s → σ* transition of carbon neighboring to the residue gallium. The material formed by this method was found to be Ga-doped diamond-like carbon (DLC) that consists of a high sp^3-hybridized carbon.

The atomic fraction of the FIB-CVD DLC film has been determined as C: Ga: H = 87.4 at.%: 3.6 at.%: 9.0 at.% using Rutherford backscattering spectrometry (RBS) and elastic recoil detection analysis (ERDA) [5]. Hydrogen content of FIB-CVD DLC film was relatively lower than that of DLC films formed by other CVD methods.

Three-Dimensional Nanostructure Fabrication

Beam-induced CVD is widely used in the electrical device industry in repair of chips and masks. This type of deposition is mainly done on two-dimensional (2-D) pattern features, but it can also be used to fabricate a three-dimensional (3-D) object. Koops et al. demonstrated a nanoscale 3-D structure construction [6] by applying electron-beam-induced amorphous carbon deposition to a micro-vacuum tube. However, FIB-CVD seems to have many advantages for the fabrication of 3-D nanostructures [3]. The key issue to realizing such 3-D nanostructures is the short penetration depth of the ions (a few tens of nm) into the target material, where the penetration depth of the ions is much shorter than that of



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 1 Fabrication process for 3-D nanostructure by FIB-CVD

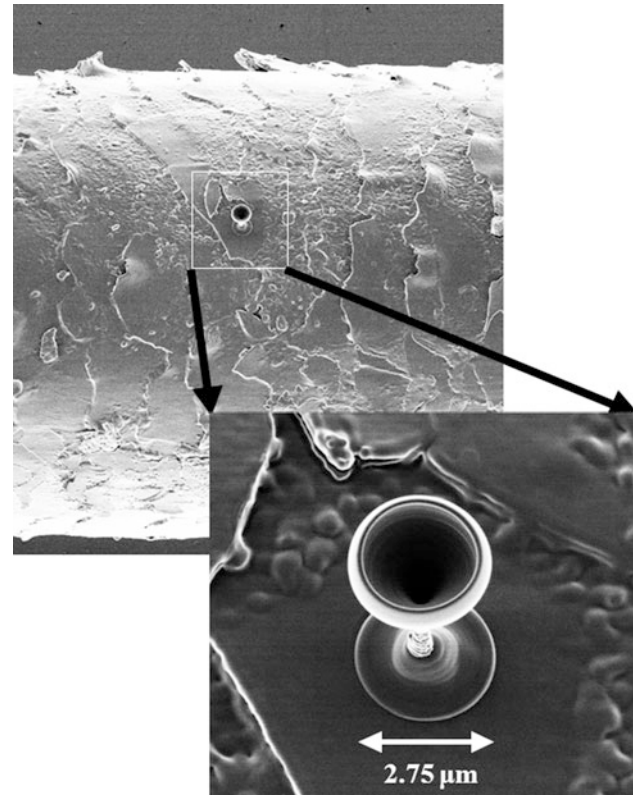
electrons (several microns). This short penetration depth reduces the dispersion area of the secondary electrons with a few eV induced by ion-beam irradiation to the molecules adsorbed on a substrate, and so the deposition area is restricted to roughly several tens nanometers.

The process of fabricating 3-D structures by FIB-CVD is illustrated in Fig. 1 [3]. In FIB-CVD processes, the beam is scanned in digital mode. First, a pillar is formed on the substrate by fixing the beam position (position 1). After that, the beam position is moved to within a diameter of the pillar (position 2) and then fixed until the deposited terrace thickness exceeds the range of the ions (a few tens nm). This process is repeated to make 3-D structures. The key point to making 3-D structures is to adjust the beam scan speed so that the ion beam remains within the deposited terrace, which means that the terrace thickness always exceeds the range of the ions. The growth in the x- and y-directions is controlled by both beam deflectors. The growth in the z-direction is determined by the deposition rate; that is, the height of structure is proportional to an irradiation time when a deposition rate is constant.

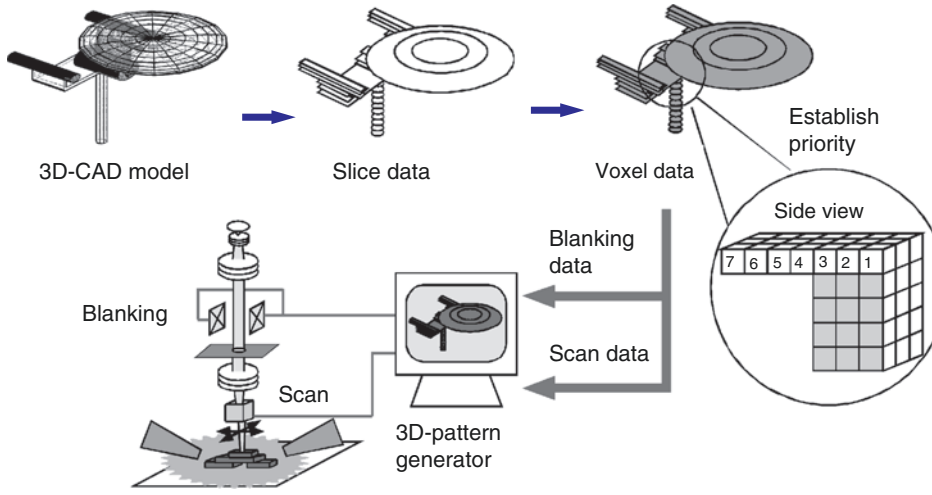
To demonstrate the possibilities of a new field called microstructure plastic arts using FIB-CVD, a "micro-wineglass" was created on a human hair as a work of microstructure plastic arts as shown in a scanting ion microscope (SIM) of Fig. 2. A micro-wineglass with an external diameter of 2.75 μm and a height of 12 μm was formed. The fabrication time was 600 s at a beam current of 16 pA. This beautiful micro-wineglass shows the potential of the field of microstructure plastic art.

The 3-D structure is built up as a multilayer structure. In the first step of this 3-D pattern-generating system, a 3-D model of the structure, designed using a 3-D CAD system (3-D DXF format), is needed. The 3-D CAD model, which is a surface model, is cut into several slices, as shown in Fig. 3. The thickness of the slices depends upon the resolution in the *z*-direction (the vertical direction). The x- and y-coordinates of the slices are then used to create the scan data (voxel data). To fabricate the overhanging structure, the ion beam must irradiate the correct positions in the correct order. If the ion beam irradiates a voxel located in midair without a support layer, the ions intended for the voxel will be deposited on the substrate. Therefore, the sequence of irradiation is determined, as shown in Fig. 3. The scan data and blanking signal therefore include the scan sequence, the dwell time, the interval time, and the irradiation pitch. These parameters are calculated from the beam diameter, x-y resolution, and z resolution of fabrication. The z resolution is proportional to the dwell time and inversely proportional to the square of the irradiation pitch. The scan data are passed to the beam deflector of the FIB-CVD as are the blanking data. The blanking signal controls the dwell time and interval time of the ion beam.

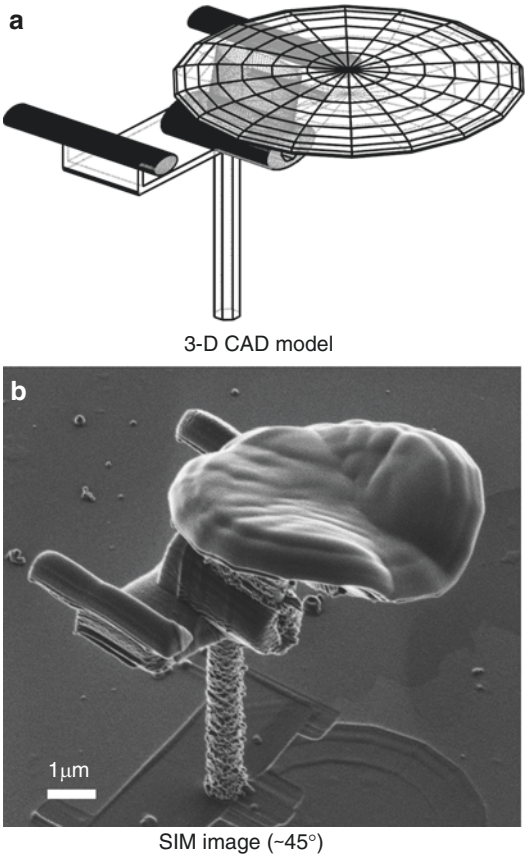
Figure 4 shows a 3-D CAD model and a SIM image of the starship Enterprise NCC-1701D (from the television series *Star Trek*) which was



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 2 Micro-wineglass with an external diameter of 2.75 µm and a height of 12 µm on a human hair



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 3 Data flow of 3-D pattern-generating system for FIB-CVD



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 4 "Micro Starship Enterprise NCC-1701D," 8.8 μm long

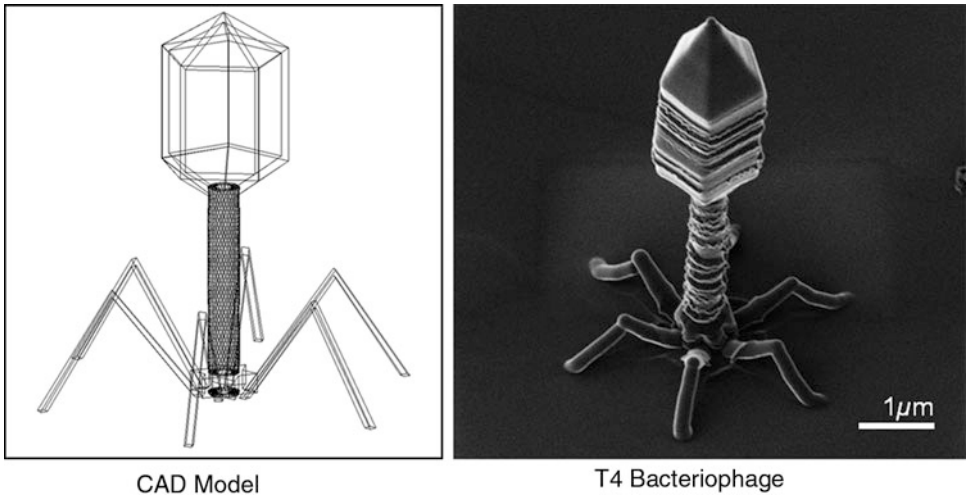
fabricated by FIB-CVD at 10 ~ 20 pA [7]. The nanospaceship is 8.8 μm long and was realized at about a 1:100 000 000 scale on silicon substrate. The dwell time (t_d), interval time (t_i), irradiation pitch (p), and total process time (t_p) were 80 μs, 150 μs, 2.4 nm, and 2.5 h, respectively. The horizontal overhang structure was fabricated successfully.

Figure 5 shows a 3-D CAD model and a SIM image of the "nano T4 bacteriophage," which is an artificial version of the virus fabricated by FIB-CVD on silicon surface. The size of the artificial nano "T4 bacteriophage" is about ten times as that of the real virus.

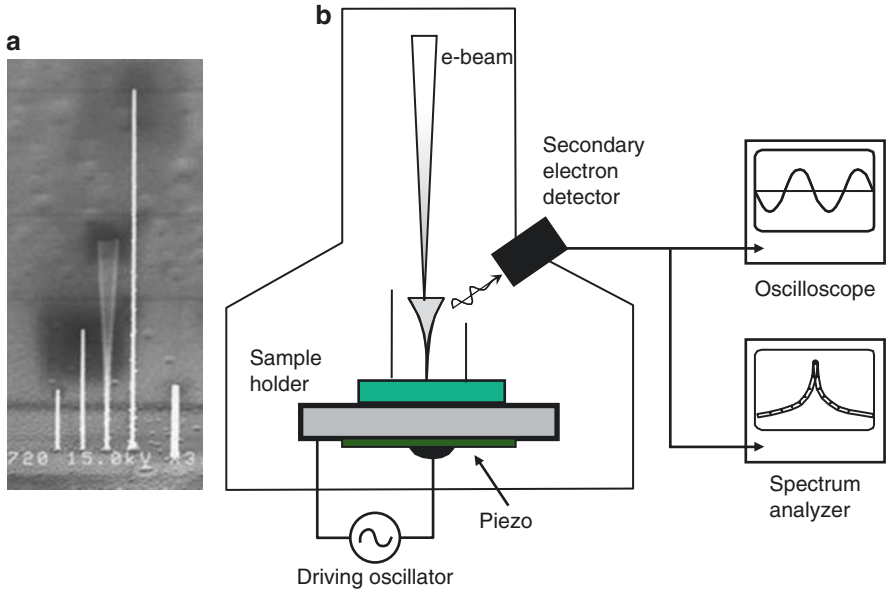
Characteristics of Materials Formed by FIB-CVD

Young's Modulus

An evaluation of the mechanical characteristics of such nanostructures is needed for material physics. Buks and Loukes reported a simple but useful technique [8] for measuring the resonant frequencies of nanoscale objects using a SEM. The secondary-electron detector in the SEM can detect frequencies up to around 4 MHz, so the sample vibration is measured as the oscillatory output signal of the detector. Buks and Loukes used this



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 5 T-4 bacteriophage fabricated by FIB-CVD



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 6 (**a**) SEM image of the vibration. The resonant frequency was 1.21 MHz. (**b**) Schematic diagram of the vibration monitoring system

technique to evaluate the Casimir attractive force between the two parallel beams fabricated on a nanoscale. We evaluated the mechanical characteristics of DLC pillars in terms of the Young's modulus, determined using resonant vibration and the SEM monitoring technique [9, 10].

The system setup for monitoring mechanical vibration is shown in Fig. 6b. The mechanical vibration was induced by a thin piezoelectric device, 300 μm thick and 3 mm square. The piezo device was bonded to the sidewall of the SEM's sample holder with silver paste. The sample holder was designed to observe cross sections in the SEM (S5000, Hitachi) system. Therefore, the pillar's vibration was observed as a side-view image, as shown in Fig. 6a. The range of vibration frequencies involved was 10 kHz up to 2 MHz, which is much faster than the SEM raster scanning speed. Thus the resonant vibrations of the pillars can be taken as the trace of the pillar's vibration in the SEM image. The resonant frequency and amplitude were controlled by adjusting the power of the driving oscillator.

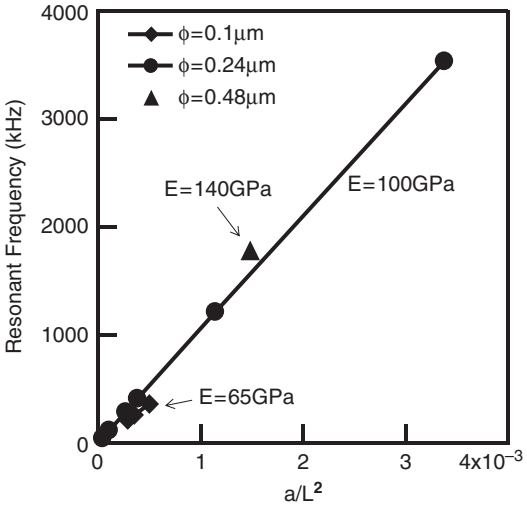
Several pillars with various diameters and lengths were fabricated. The DLC pillars with the smallest diameter of 80 nm were grown using point irradiation. Larger-diameter pillars were fabricated using an area-limited raster scan mode. The typical SEM image taken during resonance is shown in Fig. 6a. The FIB-CVD pillars seemed very durable against the mechanical vibration. This kind of measurement usually requires at least 30 min, including a spectrum analysis and photorecording, but the pillars still survived without any change in resonance characteristics. This durability of the DLC pillars should be useful in nanomechanical applications.

The resonant frequency f of the pillar is defined by Eq. 1 for a pillar with a square cross section and Eq. 2 for that with a circular cross section:

$$f_{\text{square}} = \frac{a\beta^2}{2\pi L^2}\sqrt{\frac{E}{12\rho}} \tag{1}$$

$$f_{\text{circular}} = \frac{a\beta^2}{2\pi L^2}\sqrt{\frac{E}{16\rho}} \tag{2}$$

where a is the width of the square pillar and/or the diameter of the circular-shaped pillar, L is the length of the pillar, ρ is the density, and E is the Young's modulus. The coefficient β defines



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 7 Dependence of resonant frequency on the pillar length

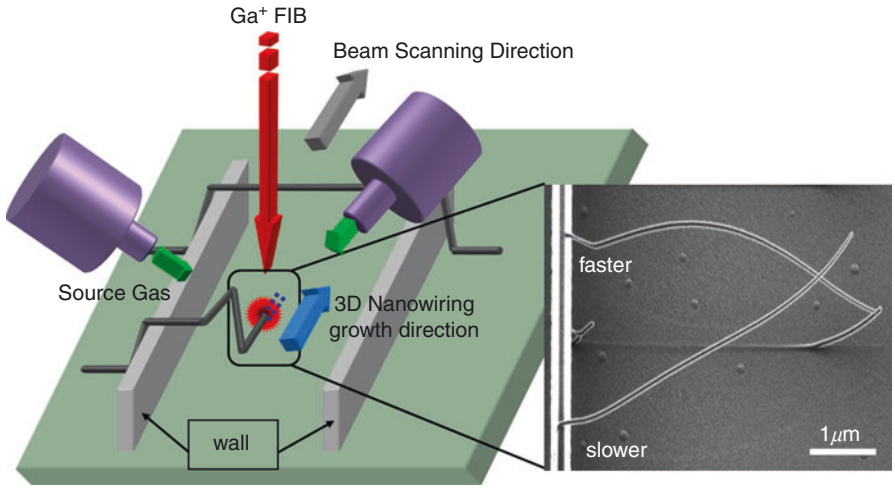
the resonant mode and the $\beta = 1.875$ for the fundamental mode. Equation 1 was used for pillars 240 nm wide and 480 nm wide, and Eq. 2 was used for pillars grown by point-beam irradiations. The relationship of the resonant frequency to the Young's modulus, which depends on the ratio of the pillar diameter to the squared length, is summarized in Fig. 7. All of the pillars evaluated in this figure were fabricated using the SMI9200 FIB system under rapid growth conditions. Typical growth rates were about 3 μm/min to 5 μm/min for the 100-nm diameter and 240-μm-wide pillars and 0.9 μm/min for the 480-nm-wide pillars. When calculating the line of Fig. 7, it was assumed that the density of the DLC pillars was about 2.3 g/cm^3, which is almost identical to that of graphite. The inclination of the line in Fig. 7 indicated the Young's modulus for each pillar. The Young's moduli of the pillars were distributed over a range from 65 GPa to 140 GPa, which is almost identical to that of normal metals. Wider pillars tended to have larger Young's moduli.

Electrical Resistivity

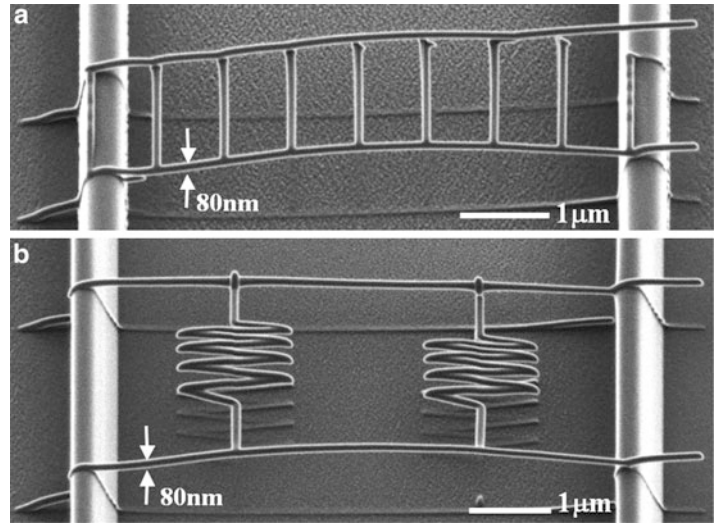
All experiments were carried out in the SMI9200 FIB system using a beam of 30 kV Ga^+ ions. The beam was focused to a spot size of 7 nm at a beam current of 0.4 pA, and it was incident perpendicular to the surface. The pattern drawing system (CPG-1000: Crestec Co., Tokyo, Japan) was added to the FIB apparatus to draw any patterns. Using the CPG, it is possible to control beam scan parameters such as scanning speed, x-y direction, and blanking of the beam, and so 3-D free-space nanowiring can be performed [11].

Figure 8 illustrates the free-space nanowiring fabrication process using both FIB-CVD and CPG. When *phenanthrene* ($C_{14}H_{10}$) gas or tungsten hexacarbonyl [$W(CO)_6$] gas, which is a reactive organic gas, is evaporated from a heated container and injected into the vacuum chamber by a nozzle located 300 μm above the sample surface at an angle of about 45° with respect to surface, the gas density of the $C_{14}H_{10}$ or $W(CO)_6$ molecules increases on a substrate near the gas nozzle. The nozzle system creates a local high-pressure region over the surface. The base pressure of the sample chamber is 2×10^{-5} Pa and the chamber pressures upon introducing $C_{14}H_{10}$ and $W(CO)_6$ as a source gas were 1×10^{-4} and 1.5×10^{-3} Pa, respectively. When a Ga^+ ion beam is irradiated onto the substrate, $C_{14}H_{10}$ or $W(CO)_6$ molecules adsorbed on the substrate surface are decomposed, and carbon (C) is mainly deposited onto the surface of the substrate. The direction of deposition growth can be controlling through the scanning direction of the beam.

After the two walls were formed in Fig. 8, free-space nanowiring was performed by adjusting the beam scanning speed. The x- and y-scanning directions and the beam scanning speed were controlled by the CPG. The height in the z-direction was proportional to the irradiation time. Deposition is made to occur horizontally by scanning the beam at a certain fixed speed in the direction of a plane. However, when the beam scanning speed is faster than the nanowiring growth speed, it grows downward or drops, and conversely when the scanning speed is too slow, the deposition grows slanting upward. Therefore, it is very important to carefully control the beam scanning speed when growing a nanowire horizontally. It turns out that the optimal beam scanning speed to realize a nanowire growing horizontally, using two $C_{14}H_{10}$ gas guns, is about 190 nm/s. The expected pattern resolution archived using FIB-CVD is



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 8 Fabrication of DLC free-space wiring using both FIB-CVD and CPG

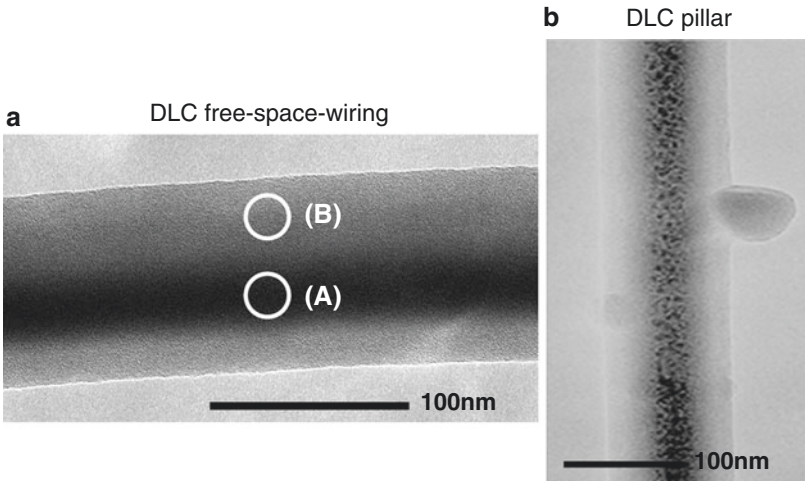


Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 9 (**a**) DLC free-space wiring with a bridge shape. (**b**) DLC free-space wiring with parallel resistances

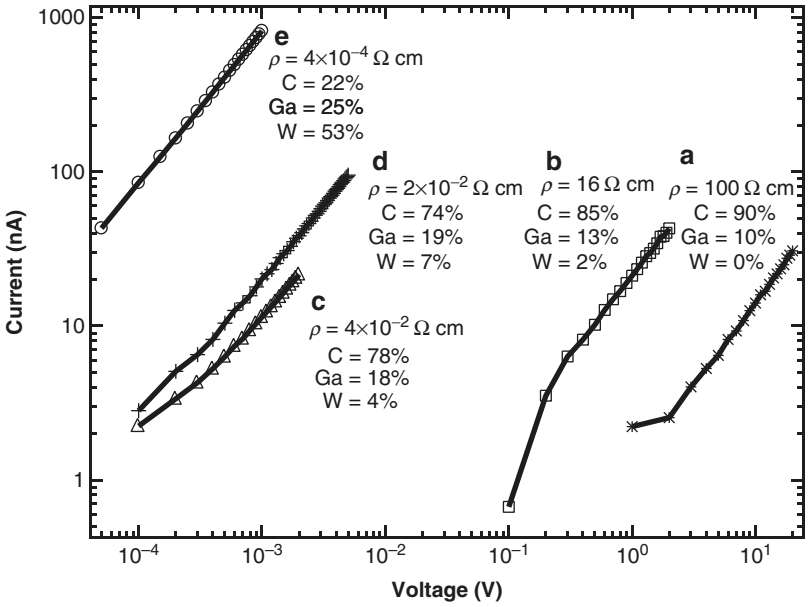
around 80 nm, because both the primary Ga^+ ion and secondary-electron scattering occur over distances around 20 nm [9, 12].

Figure 9 shows examples of free-space nanowirings fabricated by FIB-CVD and CP-G. All of the structures shown were fabricated using $C_{14}H_{10}$ gas as a precursor gas. Figure 9a shows nanobridge free-space wirings. The growth time was 1.8 min and the wiring width was 80 nm. Figure 9b shows free-space nanowires with parallel resistances. The growth time was 2.8 min, and the wiring width was also 80 nm. From these structures, it is possible to fabricate nanowiring at an arbitrary position using FIB-CVD and CPG.

The free-space wiring structures were observed using 200 keV TEM. The analyzed area was 20 nmΦ. Figure 10a and b shows TEM images of DLC free-space wiring and a pillar, respectively. It became clear from these EDX (energy-dispersive X-ray spectroscopy) measurements that the dark part (A) of Fig. 10a corresponds to the Ga core, and the outside part (B) of Fig. 10a corresponds to amorphous carbon. This free-space wiring therefore consists of



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 10 TEM images of (**a**) DLC free-space wiring and (**b**) DLC pillar



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 11 Electrical resistivity measurement for nanowiring. Electrical resistivity ρ was calculated by I–V curve. Elemental contents C, Ga, W were measured by SEM-EDX

amorphous carbon with a Ga core. The center position of the Ga core is actually located below the center of the wiring. However, in the case of the DLC pillar, the Ga core is located at the center of the pillar. To investigate the difference between these core positions, the Ga core distribution in free-space wiring was observed in detail by TEM. The center position of the Ga core was about 70 nm from the top, which was 20 nm below the center of the free-space wiring. An ion range of 30 kV Ga ions into amorphous carbon was calculated using TRIM (transport of ions in matter), which was 20 nm. The calculation indicates that the displacement of the center of the Ga core in the nanowiring corresponds to the ion range.

The electrical properties of free-space nanowiring fabricated by FIB-CVD using a mixture of $C_{14}H_{10}$ and $W(CO)_6$ were measured [11]. Nanowirings were fabricated on an Au electrode. These Au electrodes were formed on a 0.2-μm-thick SiO_2 on Si substrate by an EB lithography and lift-off process. Two-terminal electrode method was used to measure the electrical resistivity of the nanowiring. Figure 11a shows

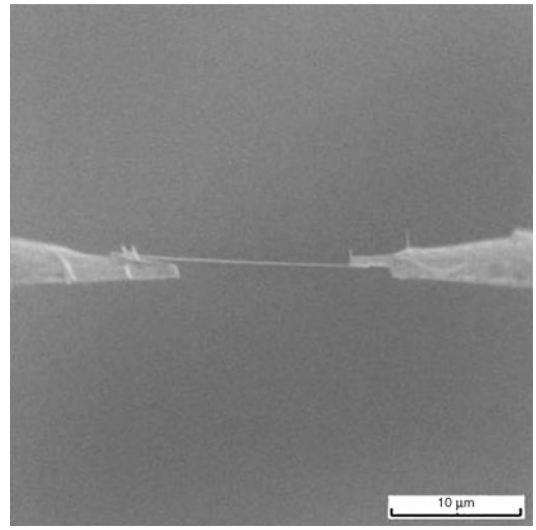
the nanowiring fabricated using only $C_{14}H_{10}$ source gas. The growth time here was 65 s and the wiring width was 100 nm. Next, $W(CO)_6$ gas was added to $C_{14}H_{10}$ gas in order to obtain a lower electrical resistivity. Figure 11b, c, and d corresponds to increasing $W(CO)_6$ contents in the gas mixture. The $W(CO)_6$ content rate was controlled by the sublimation temperature of $C_{14}H_{10}$ gas. As the $W(CO)_6$ content was increased, the nanowiring growth time and width become longer: (b) was 195 s and 120 nm, (c) was 237 s and 130 nm, and (d) was 296 s and 140 nm. Finally, the free-space nanowiring using only $W(CO)_6$ was fabricated.

The electrical resistivity of Fig. 11a fabricated using only $C_{14}H_{10}$ source gas was 1×10^2 Ωcm. The elemental contents were 90 % C and 10 % Ga, which were measured using a SEM-EDX spot beam. The *I–V* curves (b), (c), and (d) correspond to increasing $W(CO)_6$ content in the gas mixture. Increasing $W(CO)_6$ content, the electrical resistivity decreases as shown in *I–V* curves (b–d). Moreover, the Ga content also increased because the growth of nanowiring slowed down; the irradiation time of Ga^+ FIB became longer. The electrical resistivities of the *I–V* curves (b), (c), and (d) were 16×10^{-2}, 4×10^{-2}, and 2×10^{-2} Ωcm, respectively. The electrical resistivity of (e), which was fabricated by using only $W(CO)_6$ source gas, was 4×10^{-4} Ωcm. Increasing the Ga and W metallic content decreases the electrical resistivity, as shown by the SEM-EDX measurements seen in Fig. 11. These results indicate increasing the metallic content results in lower resistivity.

Examples of Applications

Electron Biprism

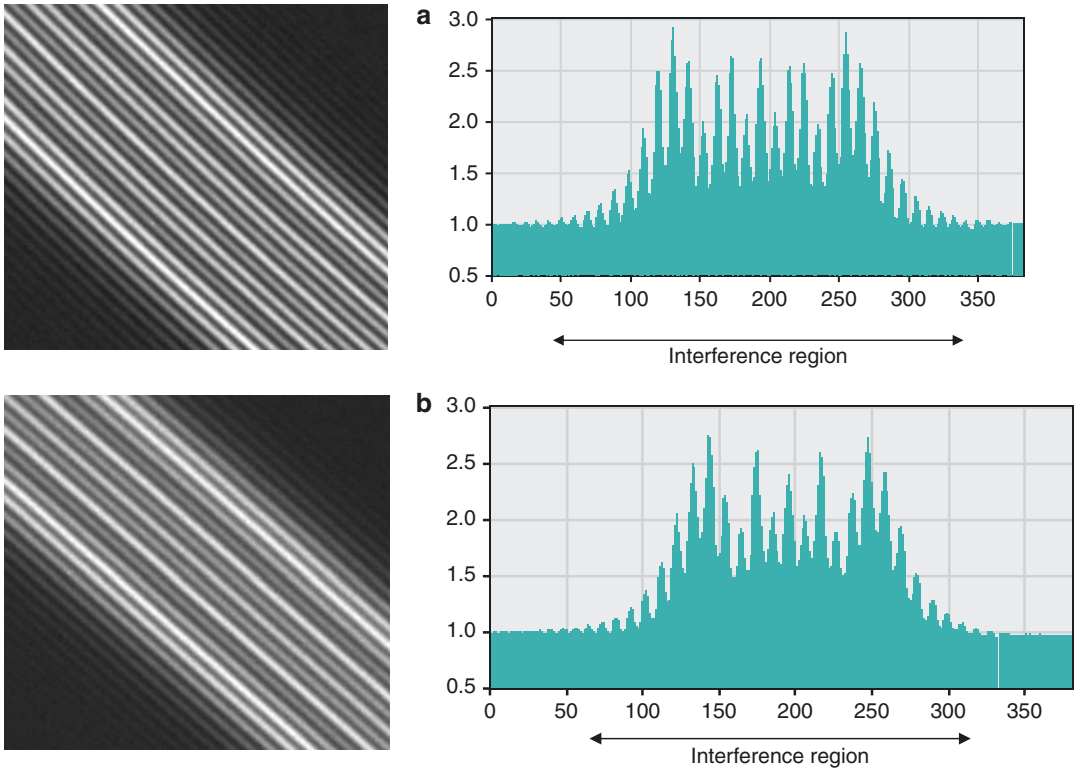
Electron holography is a useful technology for direct observation of electrical and magnetic fields at nanoscale and also has an efficient property of showing useful information by detecting the phase shift of the electron wave due to the electrical and magnetic field. The technique necessarily needs an electron biprism, which plays an important role of dividing electron wave into reference wave and objective wave. The biprism is composed of one thin filament and two ground electrodes.



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 12 Electron biprism fabricated by FIB-CVD

It is important to fabricate a filament as narrow as possible to obtain an interference fringe with a high contrast and good fringe quality. However, fabricating the filament with a diameter below 500 nm is very difficult, because a conventional electron biprism is fabricated by pulling a melted glass rod by hand. To overcome this problem, we introduce a new fabrication technique of the electron biprism using FIB-CVD and evaluate the characteristics of the new biprism [13]. Figure 12 shows an SEM micrograph of the FIB-CVD biprism. We successfully fabricated DLC wiring with smooth surface in between W rods by free-space-wiring fabrication technology of FIB-CVD. The 80-nm DLC thin wiring works as the filament of the biprism. The diameter and length of the filament are 80 nm and 15 µm, respectively.

Figure 13 shows the interference fringes obtained using the biprism with a filament of

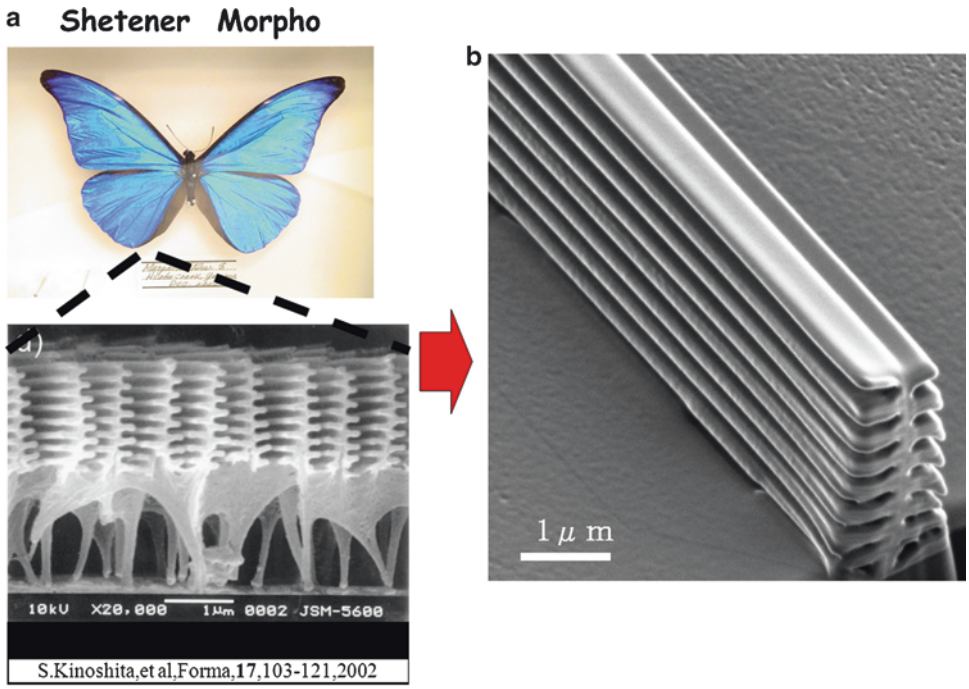


Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 13 Interference fringes and corresponding fringe profiles (**a**) obtained using the biprism with diameter of 80 nm and (**b**) obtained using the biprism with diameter of 400 nm

(a) 80-nm diameter and (b) 400-nm diameter and corresponding fringe profiles. The applied-prism voltages were 20 V. The filament with 400-nm diameter, close to the standard size used in the conventional electron biprism, was fabricated by Pt-sputter coating onto the 80-nm-diameter filament. The interference fringes were successfully obtained. Moreover, an interference region of the fringe obtained using the biprism with the 80-nm-diameter filament is larger than that of the fringe obtained using the biprism with the 400-nm-diameter filament. These results demonstrate an adequacy of the thin filament fabricated by FIB-CVD, and the new biprism will be very useful for an accurate observation with a high contrast and good fringe quality in electron holography.

Brilliant Blue Observation from a Morpho-Butterfly-Scale Quasi-Structure

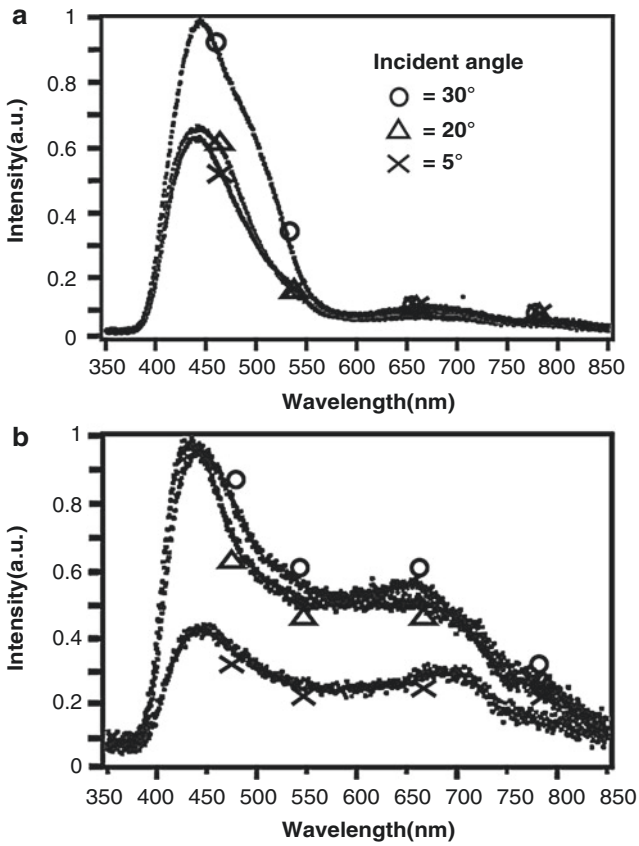
The Morpho butterfly has mysteriously brilliant blue wings, and the source of this color has been an interesting scientific problem for a long time. Through an intriguing optical phenomenon, the scales reflect interfered brilliant blue color for any incidence angle of white light. This color is called a structural color, meaning that it is not caused by pigment reflection [14]. When we observed the scales with SEM (Fig. 14a), we found three-dimensional (3-D) nanostructures with 2-μm height, 0.7-μm width, and a 0.22-μm grating pitch on the scales. These nanostructures caused the optical phenomenon in the same way as the play of color is produced in an opal and iridescence is produced by a jewel beetle.



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 14 Morpho-butterfly scales. (**a**) Top view optical microscope image of Morpho butterfly. Cross-sectional view SEM image of Morpho-butterfly scales. (**b**) Inclined-view SIM images of Morpho-butterfly-scale quasi-structure fabricated by FIB-CVD

We fabricated the Morpho-butterfly-scale quasi-structure with a commercially available FIB system (SMI9200: SII NanoTechnology Inc.) using a Ga^+ ion beam operating at 30 kV [15]. The beam diameter was about 7 nm at 0.4 pA. The FIB-CVD was done using a precursor of $C_{14}H_{10}$. Figure 14b is SIM image of the Morpho-butterfly quasi-structure fabricated by FIB-CVD using 3-D CAD data. This result demonstrates that FIB-CVD can be used to freely fabricate the quasi-structure. The Morpho-butterfly-scale quasi-structure was made of DLC. The reflectivity and transmittance of a 200-nm-thick DLC film deposited by FIB-CVD, measured by the optical measurement system at a wavelength close to 440 nm (the reflection peak wavelength of the Morpho butterfly), were 30 % and 60 %, respectively. The measured data thus indicated that the DLC film had high reflectivity near 440 nm, which is important for fabrication of an accurate Morpho-butterfly-scale quasi-structure.

We measured the reflection intensity from Morpho-butterfly scales and the Morpho-butterfly-scale quasi-structure through optical measurement and compared their characteristics. In this measurement system, white light from a halogen lamp was directed onto a sample with incident angles ranging from 5 to 45°. The reflection was concentrated by an optical microscope and analyzed by a commercially available photonic multichannel spectral analyzer system (PMA-11: Hamamatsu Photonics K.K.). The intensity of incident light from the halogen lamp had a wavelength with peak intensity close to 630 nm. Figure 15a, b, respectively, show the reflection intensity from Morpho-butterfly scales and the quasi-structure. Both had a wavelength whose peak intensity was near 440 nm and showed very similar reflection intensity spectra for the various incidence angles. We have thus successfully demonstrated that a Morpho-butterfly-scale quasi-structure fabricated by FIB-CVD can show nearly the same optical characteristics as real Morpho-butterfly scales.



Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD), Fig. 15 Intensity curves of reflection spectra. (**a**) Morpho-butterfly scales. (**b**) Morpho-butterfly-scale quasi-structure

Cross-References

► Electron-Beam-Induced Chemical Vapor Deposition

References

1. Seliger, R. L., Kubena, R. L., Olney, R. D., Ward, J. W., Wang, W.: High-resolution, ion-beam processes for microstructure fabrication. J. Vac. Sci. Technol. **16**, 1610 (1979)
2. Wargner, A., Levin, J. P., Mauer, J. L., Blauner, P. G., Kirch, S. J., Long, P.: X-ray mask repair with focused ion beams. J. Vac. Sci. Technol. **B8**, 1557 (1990)
3. Matsui, S., Kaito, T., Fujita, J., Komuro, M., Kanda, K., Haruyama, Y.: Three-dimensional nanostructure fabrication by focused-ion-beam chemical vapor deposition. J. Vac. Sci. Technol. **B18**, 3181 (2000)
4. Kand, K., Igaki, J., Kato, Y., Kometani, R., Saikubo, A., Matsui, S.: Surface modification of PTFE by synchrotron radiation under the O_2 gas atmosphere. Radiat. Phys. Chem. **75**, 1850 (2006)
5. Igaki, J., Saikubo, A., Kometani, R., et al.: Elementary analysis of diamond-like carbon film formed by focused-ion-beam chemical vapor deposition. Jpn. J. Appl. Phys. **46**, 8003 (2007)
6. Koops, H. W.: Characterization and application of materials grown by electron beam induced deposition. Jpn. J. Appl. Phys. **33**, 7099 (1994)
7. Hoshino, T., Watanabe, K., Kometani, R., et al.: Development of three-dimensional pattern-generating system for focused-ion-beam chemical-vapor-deposition. J. Vac. Sci. Technol. **B21**, 2732 (2003)
8. Buks, E., Roukes, M. L.: Stiction, adhesion energy, and the Casimir effect in micromechanical systems. Phys. Rev. **B63**, 033402 (2001)
9. Fujita, J., Ishida, M., Sakamoto, T., Ochiai, Y., Kaito, T., Matsui, S.: Observation and characteristics of mechanical vibration in three-dimensional nanostructures and pillars grown by focused ion beam chemical vapor deposition. J. Vac. Sci. Technol. **B19**, 2834 (2001)
10. Ishida, M., Fujita, J., Ochiai, Y.: Density estimation for amorphous carbon nanopillars grown by focused ion beam assisted chemical vapor deposition. J. Vac. Sci. Technol. **B20**, 2784 (2002)

F

11. Morita, T., Kometani, R., Watanabe, K., et al.: Free-space-wiring fabrication in nano-space by focused -ion-beam chemical vapor deposition. J. Vac. Sci. Technol. **B21**, 2737 (2003)
12. Fujita, J., Ishida, M., Ochiai, Y., Ichihashi, T., Kaito, T., Matsui, S.: Focused ion beam -induced fabrication of tungsten structures. J. Vac. Sci. Technol. **B20**, 2686 (2002)
13. Nakamatsu, K., Yamamoto, K., Hirayama, T., Matsui, S.: Fabrication of fine electron biprism filament by free-space-nanowiring technique of focused-ion-beam (FIB) + chemical vapor deposition (CVD) for accurate off-axis electron holography. Appl. Phys. Express **1**, 117004 (2008)
14. Vukusic, P., Roy Sambles, J.: Photonic structures in biology. Nature **424**, 852 (2003)
15. Watanabe, K., Hoshino, T., Kanda, K., Haruyama, Y., Matsui, S.: Brilliant blue observation from a morpho-butterfly-scale quasi-structure. Jpn. J. Appl. Phys. **44**, L48 (2005)

Focused-Ion-Beam-Induced Deposition (FIB-ID)

► Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)

Force Measurement

► Micromachined 3-D Force Sensors

Force Modulation in Atomic Force Microscopy

Walter Arnold
Department of Material Science and Technology, Saarland University, Saarbrücken, Germany
Physikalisches Institut, Göttingen University, Göttingen, Germany

Definition

Force modulation in atomic force microscopy is a technique to measure tip–surface interactions which in turn are determined by local elastic restoring forces, by local frictional forces, and by local adhesion between a tip and the surface under inspection. The tip or sample is oscillated at a given frequency and pushed into the repulsive regime. Data on local forces can be acquired along with topography, which allows comparison of both height and material properties.

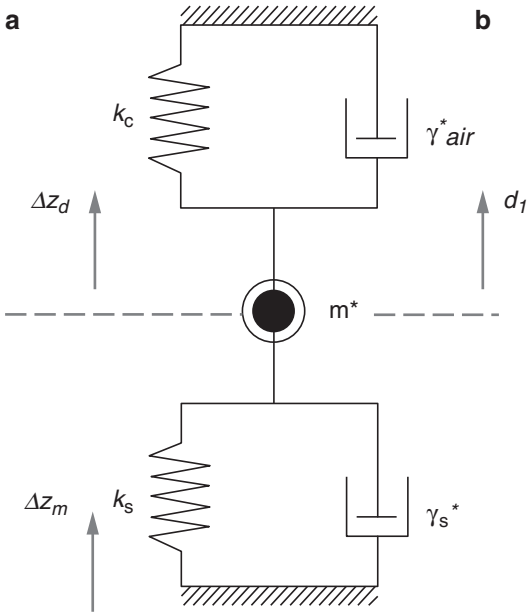
Overview

In atomic force microscopy (AFM), a microfabricated elastic beam with a sensor tip at its end is scanned over the sample surface and generates high-resolution images of surfaces. Dynamic modes, where the cantilever or the sample surface is vibrated, belong to the standard equipment of most commercial instruments. With a variety of these techniques, such as force modulation microscopy, scanning local-acceleration microscopy, scanning micro-deformation microscopy, or pulsed force microscopy, images can be obtained whose contrast depends on the elasticity, friction, and adhesion of the tip–sample contact.

The quantitative determination of the Young's modulus of a sample surface with an AFM is a challenge. Especially when stiff materials such as metals or ceramics are encountered, the image contrast due to elasticity becomes very low, because the spring constants of common AFM cantilevers, ranging from 0.01 to 70 N/m, are then much lower than the tip–sample contact stiffness. In this contribution, the various techniques are discussed to allow elasticity, anelasticity, and plastic deformation measurements or in more general terms force measurements. A number of efforts have been undertaken to combine acoustical imaging with atomic force microscopies. These are near-field imaging systems exploiting the contrast from modulation of the forces between tip and surfaces of the component under test. There is a manifold of factors determining the obtainable spatial resolution and the contrast when working in the near field of an antenna, here the tip of an AFM [1].

Force Modulation Technique

In the force modulation technique [2], the z-position of the sample is modulated with amplitudes Δz_m on the nm scale, ranging to some 10 nm (Fig. 1a). Therefore, at a constant static load applied to the cantilever, the total force exerted on the surface of the sample is modulated as well and the ensuing cantilever amplitude Δz_d depends on the local contact stiffness k_s:



Force Modulation in Atomic Force Microscopy, Fig. 1 Principle of force modulation (**a**) and scanning local-acceleration microscopy (**b**). In force modulation, the z-piezo of the AFM instrument is modulated by an amplitude Δz_m. The cantilever follows with an amplitude Δz_d which depends on the local stiffness k_s of the material and on the stiffness of the cantilever, k_c; see Eq. 1. The *dashpot* represents viscoelasticity and adhesion on the contact, as well as damping of the cantilever in air; (**b**) scanning local-acceleration microscopy (*SLAM*) [4] is identical with force modulation for frequencies below the cantilever contact-resonance frequency. For frequencies above the resonance frequency, there is a still an amplitude response of the cantilever, which is given by the accelerated mass; see Eq. 2, which corresponds to the effective mass m^* of the cantilever which is $m^* \approx m/4$ [7, 11] where m is the mass of the cantilever. In [4], the authors also discuss the cases when either the suspension of the cantilever or the cantilever itself is excited by a dynamic force inducing a corresponding cantilever amplitude

$$\frac{\Delta z_d}{\Delta z_m} = \frac{k_s}{k_s + k_c} \tag{1}$$

where k_c is the stiffness of the cantilever. In the force modulation technique, one measures the corresponding vertical photodiode output signal ΔV_m as well as the signal ΔV_d at a given static load without modulation. Using the Hertzian contact mechanics for the tip–sample contact and a calibration sample of given indentation modulus M_s, it is possible to calibrate the photodiode output signal in order to obtain the factor R [2]:

$$M_s = \frac{\sqrt{2} k_c}{a_c [R(\Delta V_m / \Delta V_d) - 1]} \tag{2}$$

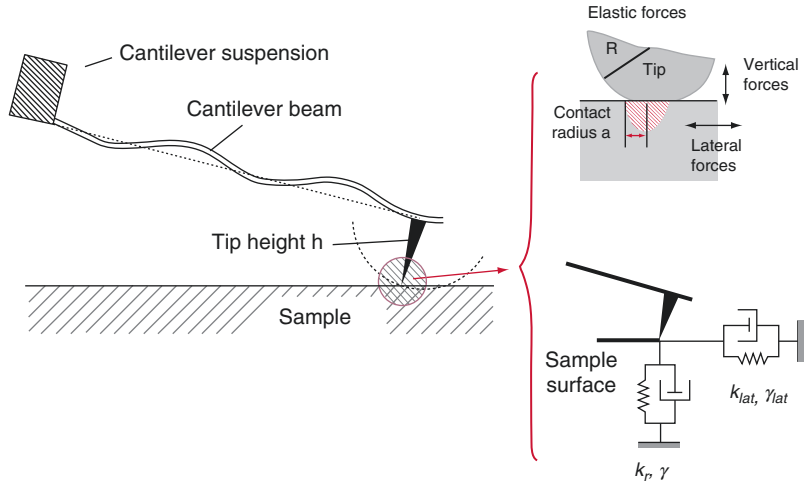
Here, a_c is the contact radius, also called sample depression radius. If the experimental parameters are set such that larger depressions occur, the modulation force develops a signal at the double frequency due to the nonlinearity of a Hertzian contact [3]. Furthermore, the adhesion in the contact, as well as the viscoelasticity of the examined material, leads to a phase shift between the force excitation signal and the output signal of the AFM photodiode.

Local-Acceleration Microscopy

From a mechanical engineering point of view, the heart of an AFM consists of the cantilever and a tip in contact with the sample. Provided that the suspension of the cantilever as well as the scanning unit is infinitely stiff, the oscillations of these AFM components may be described in a simplified way. If one assigns to the cantilever an effective mass m^*, one can describe the oscillatory behavior of the whole system with the replacement model shown in Fig. 2. Neglecting the damping effects, at frequencies smaller compared to the resonance frequency of the system in contact with a sample, the amplitude d_1 of the cantilever oscillation is given by [4]

$$\frac{d_1}{\Delta z_m} = \frac{k_s}{k_s + k_c} \tag{3}$$

F



Force Modulation in Atomic Force Microscopy, Fig. 2 Mechanical model of the AFM cantilever as a rectangular elastic beam. The contact itself is described as two parallel linear springs with spring constants or contact stiffnesses k_r and k_{lat} and two damping dashpots γ and γ_{lat} for both vertical and lateral contact forces. The cantilever length is $L = L_1 + L_2$, L_2 is the distance between the tip position and the end of the cantilever, and h is the length of the tip. Due to technical reasons, the cantilever is inclined by an angle α which is $\approx 10°$. Force modulation occurs by excitation of the sample surface with vertical displacement by an attached transducer to the bottom of the sample (AFAM or CR-AFM technique) or by exciting the cantilever suspension (UAFM). The cantilever oscillation may be excited as well by a thin-film piezoelectric transducer or built-in Schottky barrier. The frequencies of the contact resonances are independent of the type of excitation; however, the cantilever amplitudes depend on the type of excitation

where Δz_m is again the excitation amplitude of the sample. Equation 3 is equivalent to Eq. 1, and it leads after calibration, using the Hertzian contact theory again, to Eq. 2, i.e., it encompasses the force modulation. Regarding the high-frequency limit above, the contact resonance of the system leads to

$$\frac{d_1}{z_m} = \frac{k_s}{m^* \omega^2} \tag{4}$$

This means that the amplitude of the cantilever is determined by the accelerated local mass; hence, the name scanning local-acceleration microscopy (SLAM) was coined. This method has no relation with the so-called scanning laser acoustic microscope whose acronym is SLAM as well [1]. There are a number of related AFM techniques based on force modulation where instead of a cantilever with its indenting tip, a capacitive load–displacement indenter is used, allowing one to perform, besides force modulation, nanoindentation at loads higher than the loads possible with an AFM [5].

Pulsed Force Microscopy

In the pulsed force microscopy, the force–distance curve is rapidly scanned through point by point at the surface, and parameters are determined which depend on elasticity, adhesion, and other surface properties [6]. For force modulation AFM caused by piezoelectricity or magnetic forces, see the entries in this volume on piezo-force microscopy and magnetic force microscopy.

Ultrasound and Atomic Force Microscopy

A more complete description of the oscillatory behavior of the core parts of an AFM can be obtained if one considers the cantilever as a finite beam with its flexural, torsional, and

lateral eigenmodes. There are several synonyms for this approach: atomic force acoustic microscopy (AFAM), ultrasonic atomic force microscopy (UAFM), contact-resonance AFM (CR-AFM), or similarly with only minimal differences in the underlying physics of the theory of operation. There are some technical differences in the realization of the various approaches which will not be discussed here, however. This holds in particular for the commercially available instruments. In these types of force modulation techniques, one measures the resonances of atomic force cantilevers with the tip contacting the specimen surface [7–11] instead of measuring amplitudes. AFAM, UAFM, and CR-AFM emerged as tools for the measurement of elastic properties of the near-surface zone of materials with nanometer resolution. In these techniques, force modulation occurs by excitation of the sample surface with vertical displacement by an attached transducer to the bottom of the sample (AFAM or CR-AFM technique) or by exciting the cantilever suspension (UAFM). The cantilever may be excited as well by a thin-film piezoelectric transducer or by a built-in Schottky barrier. The frequencies of the contact resonances are independent of the type of excitation; however, this is not the case for the amplitudes [11]. This holds for SLAM as well [4]. Furthermore, the increased dynamic stiffness of the cantilevers at the higher modes allows for quantitative measurements and imaging of stiff materials.

From such measurements, one can derive the local indentation modulus M of the sample under test, using a suitable mechanical model relating the resonance frequencies of the oscillating cantilever to the tip–sample contact stiffness $k_s{}^*$, which in turn is related to the local indentation modulus M. The indentation modulus is an elastic constant which accounts for the compressive and the shear deformations in the tip–sample contact zone between isotropic or anisotropic materials originally used for nanoindentation [8].

Vibrations of beams are treated in many textbooks of acoustics, in textbooks of engineering mechanics, and with regard to applications to AFAM, UAFM, or CR-AFM in publications, for example, in references [7, 9, 11–13]. The interaction forces normal to the sample surface are represented as a linear spring with a characteristic spring constant k_s, which is the negative derivative of the tip–sample force in the equilibrium position:

$$k_s = -\left.\frac{\partial F(z)}{\partial z}\right|_{z=z_e} \tag{5}$$

where z is the tip–sample distance, $F(z)$ is the tip–sample interaction force, and z_e is the equilibrium position. The lateral contact stiffness $k_{s,\mathrm{lat}}$ is defined analogously. The ensuing mechanical model of a vibrating cantilever in AFAM is presented in Fig. 2. The superposition of vertical and lateral tip–sample interaction forces may be represented by a set of springs and dashpots.

For a beam of uniform cross section, the equation of motion for flexural vibrations is a fourth-order differential equation. Taking into account the boundary conditions for an oscillating cantilever, either free or in contact with the sample, one obtains the characteristic equation from the equation of motion for which the solutions $\kappa_N L$ ($N = 1, 2, 3\ldots$) yield the wave numbers κ_N of an infinite set of flexural modes. The dispersion relations are used to calculate from the resonance frequency f_N of the N^{th} mode [7, 9, 11–13] the contact stiffness. When approaching the sample surface, the tip first senses long-range attractive forces before contact and in contact repulsive tip–sample forces. If the static load $F_0 = k_c \times d_c$, applied to the tip by a cantilever deflection d_c, is large enough, attractive forces may be neglected.

If one takes viscoelasticity of the examined materials into account, the local contact stiffness k_s^* becomes complex and can be written as [14]

$$k_s^* = k_r + ik_i = k_r + i\omega\gamma, \tag{6}$$

where γ is the contact damping. In order to obtain Q_{loc}^{-1} caused by the local contact damping from the total damping Q^{-1} of the cantilever, one has to view the cantilever as a distributed mass system where a flexural wave mode of order N propagates back and forth along the cantilever. The damping of this wave is caused (i) by internal friction in the

F

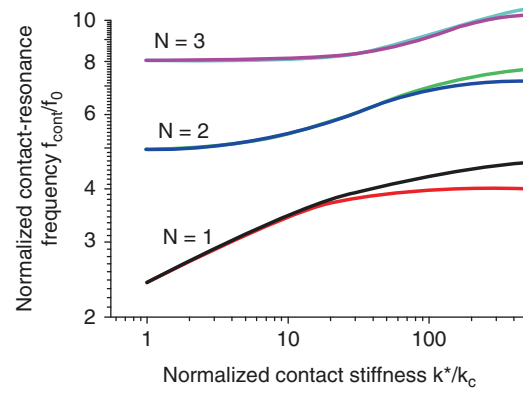
stress field of the cantilever tip extending into the sample [15], (ii) by the motion of the cantilever in the surrounding air [7], and (iii) by damping in the contact due to adhesive and frictional forces [16]. Procedures to separate the various parts are discussed in [14, 17]. For the Nth cantilever mode, the local Q_{loc}^{-1} is given by [18]

$$Q_{N,\,\text{loc}}^{-1} = \frac{M''_S}{M'_S} = \frac{\omega\gamma}{k_r} = \frac{\beta_N}{\alpha_N}\frac{L}{L_1}\frac{\omega_{c,N} \times (\kappa_N L)^2}{\omega_{\text{free},N}} \tag{7}$$

Here, M'_S is the real part of the indentation modulus of the sample, M''_S is its imaginary part, α_N and β_N are the real and imaginary parts of the wave vector κ_N describing the cantilever's motion, $\omega_{c,N}$ is the contact-resonance frequency of mode N, and $\omega_{\text{free},N}$ is the free resonance frequency of mode N. Furthermore, L is the length of the cantilever and L_1 designates the position of the cantilever tip; see Fig. 2. For the first cantilever flexural mode, Eq. 8. reduces to [18]

$$Q_{1,\,\text{local}}^{-1} \approx 3.516\frac{\beta_1}{\alpha_1}\frac{L}{L_1}\frac{\omega_{c,1}}{\omega_{\text{free},\,1}} \tag{8}$$

There is also a lateral contribution γ_{lat} to the damping in the contact [14]. The procedures described in [14, 17, 18] yield both local damping and local stiffness k_r and allow imaging of Q_{loc}^{-1}. Typical total Q-values observed are $Q \approx 100$. For the contact oscillations alone, i.e., after subtraction of the damping of the cantilever beam in air, the Q-values are higher [18]. For a spring-coupled cantilever with the ratio $k_{\text{lat}}/k_r = c_k$, and for an inclined cantilever, one obtains a characteristic equation or a dispersion curve whose solution can be found in [11–13]. A graphic display of a special dispersion curve is shown in Fig. 3. The ratio of the lateral contact stiffness to the vertical contact stiffness has been calculated for an elastic contact to be $k_{\text{lat}}/k_r = c_k = 4G^*/E^*$ [19], where E^* and G^* are the reduced Young's modulus and shear modulus of the contact zone, respectively. With a Poisson ratio of 0.4, one obtains $c_k \approx 0.85$.



Force Modulation in Atomic Force Microscopy, Fig. 3 Normalized contact-resonance frequency f_N/f_0 for the first three cantilever modes ($N = 1, 2, 3$) as a function of normalized contact stiffness k_r/k_c according to the cantilever dispersion equations (for small damping $k^* = k_r$). Due to the shape of the dispersion curve, variations in contact resonances may entail a large variation in contact stiffness and consequently in indentation modulus. Depending on the slope of the dispersion curve, higher modes ($N = 2$; $N = 3$) may become more sensitive to measure a local indentation modulus quantitatively [11]. Each dispersion curve display two values, one for $c_k = 0$ (*lower curves*) and one for $c_k = 0.85$ (*upper curves*). Parameters used for the calculations were $L_1 + L_2 = 137$ µm, $L_2 = 15.1$ µm, $h = 17$ µm, and $\alpha = 11°$ (According to M. Büchsenschütz-Göbeler, University of Göttingen, 1. Physikalisches Institut, 2010, with kind permission)

As a first step in a measurement, one obtains from the ratio $f_{\text{cont}}/f_{\text{free}}$ of the cantilever employed the normalized contact stiffness k_r/k_c, using the appropriate parameters in the dispersion curve. Knowing k_s^*, both the real and the imaginary parts, k_r and k_i, and the contact area or contact radius of the tip–sample contact, one can determine the local indentation modulus M' using the relation [8, 11, 13, 14]:

$$k_r = \frac{2}{\sqrt{\pi}}M'\sqrt{A} \tag{9}$$

$$k_i = \frac{2}{\sqrt{\pi}}M''\sqrt{A} \tag{10}$$

where A is the contact area and M' and M'' are the real and the imaginary parts of the indentation modulus. In case of an isotropic material, these

are equivalent to the reduced modulus of elasticity E^* and to the loss modulus $E^{*\prime\prime}$, respectively.

Indentation Modulus

For isotropic solids E^* is given by

$$\frac{1}{E^*} = \frac{1-\nu_s^2}{E_s} + \frac{1-\nu_{\text{tip}}^2}{E_{\text{tip}}} \tag{11}$$

Here, $E_{s,\text{tip}}$ are the Young's modulus of the sample and of the tip, and $\nu_{s,\text{tip}}$ are the corresponding Poisson ratios. In the case of anisotropic solids, an indentation modulus M is introduced, which can be calculated from the elastic single-crystal constants in a given direction [20]. If there exists a three- or fourfold rotational symmetry axis perpendicular to the boundary, the contact area is circular. The required symmetry holds for silicon sensor tips, which are oriented in (001) crystallographic direction. In this case, Eq. 11 can be replaced by

$$\frac{1}{E^*} = \frac{1}{M_S} + \frac{1}{M_{\text{tip}}} \tag{12}$$

where M_S and M_{tip} are the indentation moduli of the sample and the tip, respectively [8]. For isotropic bodies, the indentation modulus is $M = E/(1-\nu^2)$, and Eq. 12 becomes Eq. 11.

Finally, neglecting the effect of damping, the elastic properties of the sample can be extracted by comparing measurements in order to obtain information on the contact area or the contact radius:

$$E_s^* = E_r^* \left(k_s^*/k_r^*\right)^m \tag{13}$$

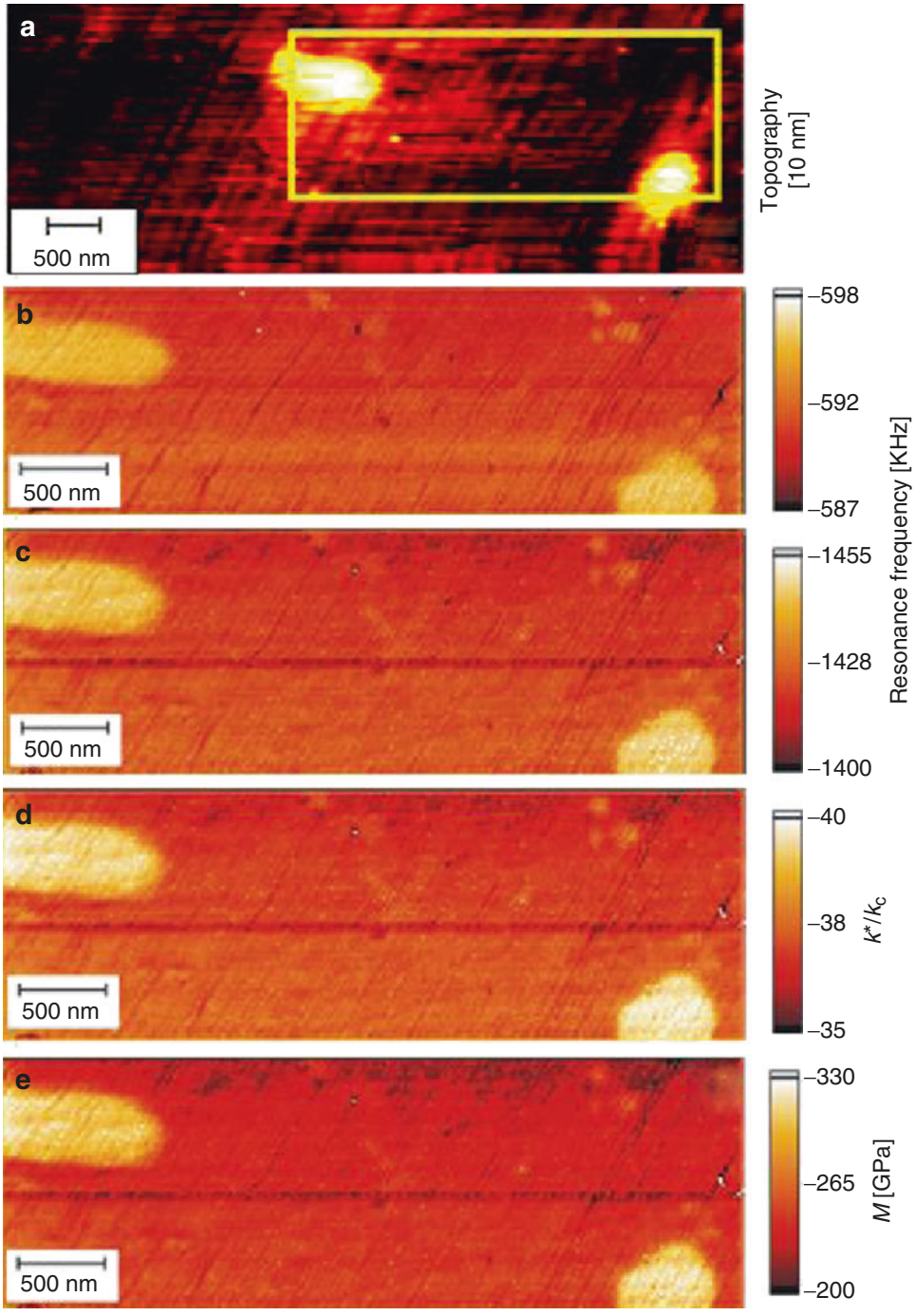
Here, r and s refer to the reference sample and the unknown sample, respectively, and m describes the tip geometry. For a flat punch $m = 1$ and for a spherical tip $m = 3/2$ [8, 9, 21]. When making comparative measurements, for example, on nanocrystalline materials, on polycrystalline materials, or on metallic glasses, one may use as well elastic data from ultrasonic velocity measurements to calculate the mean macroscopic value of M_s [8, 10, 22].

An AFAM image is shown in Fig. 4 demonstrating the sequence of data evaluation. The first and the second contact-resonance spectra were obtained at every point of the area examined in a modified 9Cr–1Mo ferritic steel with $M_{23}C_6$ carbide precipitates [10]. The distribution of k_r/k_c values was derived using the contact-resonance spectra and the mechanical model shown in Fig. 2 with the appropriate dispersion curve. As can be seen, the AFAM technique is a tool for the determination of the modulus of submicron-sized precipitates with a spatial resolution of less than 10 nm in polycrystalline materials. The spatial resolution is given by the contact radius a_c (or depression radius; see above) which depends on the tip radius R_t, the applied static force F_0 acting on the tip including the forces of adhesion F_0, and the reduced elastic modulus E^* of the contact. In case of Hertzian contact with a spherical tip, this value is given by

$$a_c = \sqrt[3]{3F_0R_t/4E^*} \tag{14}$$

This resolution has been tested and exploited in several applications of contact-resonance AFM [8–10, 13, 21].

When in an AFAM setup, the amplitude of the surface vibrations is increased, and a DC offset of the cantilever develops, i.e., the sensor tip lifts off the sample surface. This rectifying effect due to the nonlinear force distance curve has been known since the early experiments combining ultrasonic technology with AFM and was exploited in ultrasonic force microscopy (UFM) [25]. The advantage of this technique compared to AFAM, UAFM, or CR-AFM is that only a low detection bandwidth of the cantilever optical detection system of the AFM is required. However, as both elasticity and adhesion contribute to the image contrast in a nonlinear fashion [11, 12, 23–25], their separation in a quantitative manner remains difficult. Also, the UFM technique can be employed to study frictional properties [26]. Using the torsional cantilever contact resonances, one can obtain data on the shear elasticity as well as on frictional properties in the contact [16]. Finally, scanning microdeformation



Force Modulation in Atomic Force Microscopy, Fig. 4 Sequence of AFAM or contact-resonance measurements, here on a 9Cr–1Mo ferritic steel specimen containing $M_{23}C_6$-type carbides: (**a**) The topography image allows one to judge the suitability of the surface area for AFAM measurement. If the topography is too rough, the contact area A of the tip–sample contact will influence the results; see Eq. 8; (**b**, **c**) cantilever contact-resonance frequencies measured for the first ($N = 1$) (**b**) and (**c**) the second contact resonance ($N = 2$); (**d**) from the contact resonances, k_r/k_c is determined using the appropriate dispersion curve for the cantilever employed; (**e**) the indentation modulus M is determined by

microscopy [25] is analog to AFAM, UAFM, and CR-AFM in its technical realization, except that tips with larger radii are used, and hence the contact radius a_c is micrometer range instead of the nanometer range.

Subsurface Imaging

There is a lot of interest to image with atomic force microscopy subsurface features, which is difficult because AFM is a technique for surface characterization by its operational principle. Therefore, it seems straightforward to combine ultrasound with its penetrating capability with AFM for this purpose, and indeed there are efforts to achieve subsurface acoustical imaging by using various schemes of AFAM or UAFM.

Subsurface imaging with an AFM is possible if the physical quantity measured exhibits a gradient into the material within the penetration of the stress field emanating from the tip–sample contact. For example, the stiffness of the material changes at the surface when there are defects, dislocations, variations of film thicknesses, or changes of material properties within a certain depth [9, 27–29]. Cavities can be seen at depths of about 200–500 nm and having diameters of ≈300 nm. The depth range is roughly given by the extent of the stress field which the tip exerts on the surface or the stress field of the defect or their joint action and therefore on the static load. As is the case for all interior acoustical imaging techniques, the contrast determines whether an image of a given defect is obtained with an acceptable signal-to-noise ratio. Here, the ratio k_{def}/k_r of the stiffness of the defect relative to the host material determines the contrast. As a rule of thumb, one can assume that the depth range z is about three times the contact radius a_c if the Hertzian contact mechanics between tip and surface prevails, although in certain cases one can detect much deeper [29]. Recently, a subsurface imaging scheme combining ultrasound and AFM [30] called scanning near-field ultrasonic holography (SNFUH) has been reported; however, the origin of the contrast is still a matter of debate and is not further discussed here. Finally, a scheme has been reported in an AFAM setup that is meant to render possible the detection of small defects in materials much below the Hertzian stress field, using force modulation at 1 GHz [31].

Cross-References

- ▶ AFM, Noncontact Mode
- ▶ AFM Probes
- ▶ AFM, Tapping Mode
- ▶ Atomic Force Microscopy
- ▶ Dynamic Force Measurements on Modified Surfaces with AFM
- ▶ Friction Force Microscopy
- ▶ Magnetic Nanostructures and Spintronics
- ▶ Mechanical Properties of Nanocrystalline Metals
- ▶ Mechanical Properties of Nanostructures
- ▶ Nanocrystalline Materials
- ▶ Nanomechanical Resonant Sensors and Fluid Interactions
- ▶ Nano-Piezoelectricity
- ▶ Piezoelectric Effect at Nanoscale
- ▶ Piezoresponse Force Microscopy and Spectroscopy
- ▶ Plasticity Theory at Small Scales
- ▶ Ultrasonic Machining

Force Modulation in Atomic Force Microscopy, Fig. 4 (continued) calibration with a reference sample according to Eq. 13 in order to determine the contact area A or contact radius a_c; see Eq. 8. The reference sample may be the host material itself, and the global values for the elastic moduli for calibration may be determined by ultrasonic time-of-flight data. As can be seen in (**e**), the distribution of the indentation modulus of the $M_{23}C_6$-type carbide precipitate becomes visible as bright islands having a value of $M = 310 \pm 10$ GPa. The contact radius given in Eq. 14 determines the spatial resolution as well as the topography of the sample. Typical values range from nm to some 10 nm (Data in modified form from Kumar et al. [10] and from additional data, with kind permission by AIP and by A. Kumar, IGCAR, Kalpakkam, India)

References

1. Zinin, P., Arnold, W., Weise, W., Berezina, S.: Theory and applications of scanning acoustic microscopy and scanning near-field acoustical imaging. In: Kundu, T. (ed.) Ultrasonic Nondestructive Evaluation: Engineering and Biological Material Characterization. CRC Press, Boca Raton (2012), and references contained therein
2. Maivald, P., Butt, H.-J., Gould, S.A.C., Prater, C.B., Drake, B., Gurley, J.A., Elings, V.B., Hansma, P.K.: Using force modulation to image surface elasticities with the atomic force microscope. Nanotechnology **2**, 103–106 (1991)
3. Radmacher, M., Tilmann, R.W., Gaub, H.E.: Imaging viscoelasticity by force modulation with the atomic force microscope. Biophys. J. **64**, 735–742 (1993)
4. Burnham, N.A., Gremaud, G., Kulik, A.J., Gallo, P.-J., Oulevey, F.: Materials properties measurements: choosing the optimal scanning probe microscope configuration. J. Vac. Sci. Technol. **B14**, 1308–1312 (1996)
5. Syed Asif, S.A., Wahl, K.J., Colton, R.J., Warren, O. L.: Quantitative imaging of nanoscale mechanical properties using hybrid nanoindentation and force modulation. J. Appl. Phys. **90**, 1192–1200 (2001), and references contained therein
6. Krotil, H.-U., Stifter, T., Marti, O.: Concurrent measurement of adhesive and elastic surface properties with a new modulation technique for scanning force microscopy. Rev. Sci. Instrum. **71**, 2765–2771 (2000)
7. Rabe, U., Janser, K., Arnold, W.: Vibrations of free and surface-coupled atomic-force microscope cantilevers: theory and experiment. Rev. Sci. Instrum. **67**, 3281–3293 (1996)
8. Kopycinska-Müller, M., Caron, A., Hirsekorn, S., Rabe, U., Natter, N., Hempelmann, R., Birringer, R., Arnold, W.: Quantitative evaluation of elastic properties of nano-crystalline nickel using atomic force acoustic microscopy. Z. Phys. Chem. **222**, 471–498 (2008), and references contained therein
9. Yamanaka, K., Kobari, K., Tsuji, T.: Evaluation of functional materials and devices using atomic force microscopy with ultrasonic measurements. Jpn. J. Appl. Phys. **47**, 6070–6076 (2008), and references contained therein
10. Kumar, A., Rabe, U., Hirsekorn, S., Arnold, W.: Elasticity mapping of precipitates in polycrystalline materials using atomic force acoustic microscopy. Appl. Phys. Lett. **92**, 183106 (2008)
11. Rabe, U.: Atomic force acoustic microscopy. In: Bushan, B., Fuchs, H. (eds.) Applied Scanning Probe Methods, vol. 2, pp. 37–90. Springer, Berlin (2006), and references contained therein
12. Turner, J.A., Hurley, D.C.: Ultrasonic methods in contact atomic force microscopy. In: Placko, D., Kundu, T. (eds.) Ultrasonic Methods for Material Characterization. Instrumentation, Mésure, Métrologie, vol. 3, pp. 117–148. Lavoisier, Cachan (2003), and references contained therein
13. Stan, G., Cook, R.F.: Mapping the elastic properties of granular Au films by contact resonance atomic force microscopy. Nanotechnology **19**, 235701 (2008)
14. Huya, P.A., Hurley, C.D., Turner, J.A.: Contact-resonance atomic force microscopy for viscoelasticity. J. Appl. Phys. **104**, 074916 (2008)
15. Caron, A., Arnold, W.: Observation of local internal friction and plasticity onset in nanocrystalline nickel by atomic force acoustic microscopy. Acta Mater. **57**, 4353–4363 (2009)
16. Scherer, V., Reinstädtler, M., Arnold, W.: Atomic force microscopy with lateral modulation. In: Fuchs, H., Bhushan, B., Hosaka, S. (eds.) Applied Scanning Probe Methods, pp. 75–150. Springer, Berlin (2003), and references contained therein
17. Killgore, J.P., Yablon, D.G., Tsou, A.H., Gannepalli, A., Turner, J.R., Proksch, R., Hurley, D.C.: Viscoelastic property mapping with contact resonance force microscopy. Langmuir **27**, 13983 (2011)
18. Wagner, H., Büchsenschütz-Göbeler, M., Luo, Y., Kumar, A., Arnold, W., Samwer, K.: Measurement of local internal friction in metallic glasses. J. Appl. Phys. **115**, 134307 (2014); Erratum: J. Appl. Phys. **115**, 134307 (2014)
19. Mazeran, P.E., Loubet, J.L.: Normal and lateral modulation with a scanning force microscope, an analysis: implication in quantitative elastic and friction imaging. Tribol. Lett. **7**, 199–212 (1999)
20. Vlassak, J.J., Nix, W.D.: Indentation modulus of elastically anisotropic half-spaces. Philos. Mag. **A67**, 1045–1056 (1993)
21. Stan, G., Price, W.: Quantitative measurements of indentation moduli by atomic force acoustic microscopy using a dual reference method. Rev. Sci. Instrum. **77**, 103707 (2006)
22. Wagner, H., Bedorf, D., Küchemann, S., Schwabe, M., Zhang, B., Arnold, W., Samwer, K.: Local elastic properties of a metallic glass. Nat. Mater. **10**(6), 1–4 (2012). doi:10.1038/nmat3024
23. Yamanaka, K., Ogiso, H., Kolosov, O.: Ultrasonic force microscopy for nanometer resolution subsurface imaging. Appl. Phys. Lett. **64**, 178–180 (1994)
24. Muraoka, M., Arnold, W.: A method to evaluate local elasticity and adhesion energy based on nonlinear response of AFM cantilever vibration. JSME Int. J. **A44**, 396–405 (2001)
25. Vairac, P., Boucenna, R., Le Rouzic, J., Cretin, B.: Scanning microdeformation microscopy: experimental investigations on non-linear contact spectroscopy. J. Phys. D Appl. Phys. **41**, 155503 (2008), and references contained therein
26. Cuberes, M.T.: Nanoscale friction and ultrasound. In: Gnecco, E., Meyer, E. (eds.) Fundamentals of Friction and Wear, pp. 49–71. Springer, Berlin (2007), and references contained therein
27. Yaralioglu, G.G., Degertekin, F.L., Crozier, K.B., Quate, C.F.: Contact stiffness of layered materials for

ultrasonic force microscopy. J. Appl. Phys. **87**, 7491–7496 (2000)
28. Hurley, D.C., Kopycinska-Müller, M., Langlois, E.D., Kos, A.B.., Barbosa III, N.: Mapping substrate/film adhesion with contact-resonance-frequency atomic force microscopy. Appl. Phys. Lett. **89**, 021911 (2006)
29. Striegler, A., Koehler, B., Bendjus, B., Roellig, M., Kopycinska-Mueller, M., Meyendorf, N.: Detection of buried reference structures by use of atomic force acoustic microscopy. Ultramicroscopy **111**, 1405–1416 (2011)
30. Shekhawat, G.S., Dravid, V.P.: Nanoscale imaging of buried structures via scanning near-field ultrasound holography. Science **310**, 89–92 (2005)
31. Hu, S., Su, C., Arnold, W.: Imaging of subsurface structures using atomic force acoustic microscopy at GHz frequencies. J. Appl. Phys. **109**, 084324 (2011)

Force Spectroscopy

▶ 3D Micro-/Nanomanipulation with Force Spectroscopy

Force-Detected Magnetic Resonance

▶ Magnetic Resonance Force Microscopy

Forests

▶ Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

Fracture Toughness

▶ Nanomechanical Properties of Nanostructures

Friction

▶ Nanotribology

Friction Force Microscopy

André Schirmeisen[1], Udo D. Schwarz[2] and Hendrik Hölscher[3]
[1]Institute of Applied Physics, Justus-Liebig-University Giessen, Giessen, Germany
[2]Department of Mechanical Engineering, Yale University, New Haven, CT, USA
[3]Karlsruher Institut für Technologie (KIT), Institut für Mikrostrukturtechnik, Karlsruhe, Germany

F

Synonyms

Lateral force microscopy (LFM)

Definition

The friction force microscope (FFM) is a derivative of the atomic force microscope (AFM), which is optimized to measure the lateral forces that are acting between a sharp, nanoscopic tip and a surface during scanning of the tip. Ever since it was demonstrated that the frictional force can be measured at the atomic scale in 1987 by Mate et al. [1], this technique had a tremendous impact on understanding fundamental, molecular scale friction and lubrication processes and has been instrumental in defining the field of nanotribology.

Overview

In basic physics textbooks, the frictional force F_{friction} between two bodies in contact is typically described by the following phenomenological law of friction:

$$F_{\text{friction}} = \mu F_{\text{load}}, \quad (1)$$

where F_{load} represents the external loading force and μ the friction coefficient, which depends only on the actual combination of materials in contact. This equation is essentially the condensed form of observations first noted by Leonardo da Vinci

(1452–1519). His findings had been fallen into oblivion after his death, only to be independently rediscovered by Guillaume Amontons (1663–1705), who reported them to the French Académie Royale des Sciences on December 19, 1699. Hence, Eq. 1 is commonly referred to as *Amontons' law*. It holds surprisingly well especially in the case of dry friction or Coulomb *friction*, which represents the case of sliding without lubrication between the two interacting bodies. Unfortunately, it cannot be derived from first principles up to now. Even worse, the entire understanding of the fundamental mechanisms of friction on the atomic scale is still relatively limited since most macroscopic and microscopic frictional effects are usually dominated by the influence of wear, plastic deformation, lubrication, surface roughness, and surface asperities. Macroscopic friction experiments are therefore difficult to analyze in terms of a universal theory. As this complexity of phenomena prevented the observation of pure wearless friction for many years, attempts to rigorously analyze the origins of friction on the molecular and atomic scale had been largely unsuccessful.

In the last decades, however, the field of *nanotribology* was established by introducing new experimental tools, which made the nanometer and atomic scales accessible to tribologists [2–4]. One of the basic ideas of nanotribology is that for a better understanding of friction in macroscopic systems, the frictional behavior of a single-asperity contact should be investigated first. It is the hope that once the atomic-scale manifestations of friction at such a nanometer-sized single asperity have been clarified, macroscopic friction could be explained with the help of statistics, i.e., by the summation of the interactions of a large number of small individual contacts, which form the macroscopic roughness of the contact interface.

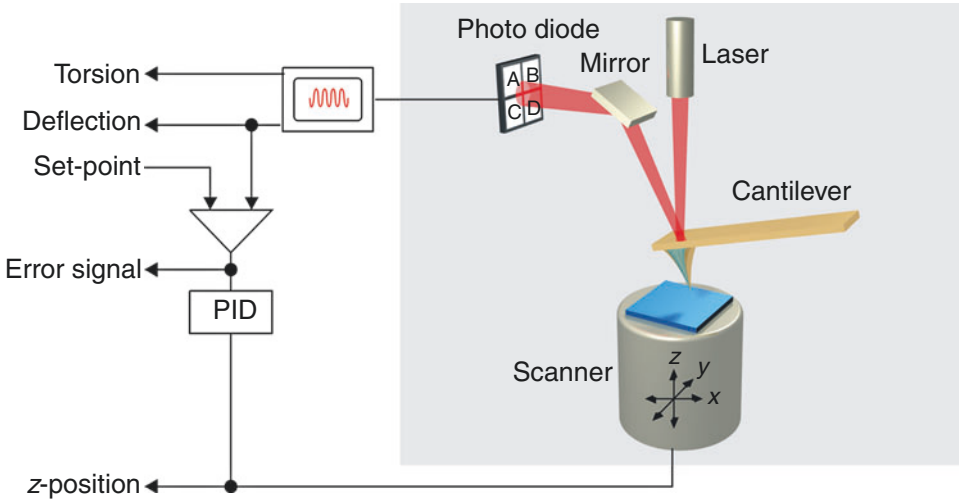
The first important step toward such experiments at the atomic level was done when the *surface force apparatus* was equipped with special sample stages enabling the measurement of lateral forces between two molecularly smooth surfaces sliding against each other [5]. The largest step forward concerning the reduction of the asperity size, however, was accomplished with the development of the *friction force microscope* (FFM) [1]. The FFM is derived from the *atomic force microscope* (AFM) and enables the local measurement of lateral forces by moving a sharp tip, representing (approximately) a *point contact*, over a sample surface.

Technical Setup of a Friction Force Microscope

The experimental setup of an atomic or friction force microscope (AFM or FFM) is based on a simple idea. It detects forces acting between a sample surface and a sharp tip that is mounted on a soft leaf spring (the so-called cantilever). Such cantilevers are produced by standard microfabrication techniques, mostly with rectangular or V-like shapes, from silicon or silicon nitride. Spring constants for the bending of cantilevers typically used for AFM measurements are about 0.01–1 N/m. The lateral spring constants for the torsion of the cantilever are in the range between 10 and 100 N/m.

A feedback system, which controls the vertical z-position of the tip on the sample surface, keeps the deflection of the cantilever (and thus the force between tip and sample) constant. Moving the tip relative to the sample in the x-y-plane of the surface by means of piezoelectric drives, the actual z-position of the tip is recorded as a function of the lateral x-y-position with (ideally) sub-Ångström precision. The obtained data represents a map of equal forces, usually interpreted as the surface topography. The conversion of the cantilever deflection into the normal force is performed by applying Hooke's law. The data can be analyzed and visualized through computer processing. A more detailed description of the method and the generally used instrumentation can be found in the books of Meyer [6] and Bhushan [4].

The by far most widespread detection scheme for atomic force microscopy experiments is the laser beam deflection technique sketched in Fig. 1. As explained in the figure caption, not only the deflection but also the torsion of the cantilever can be measured. Atomic force microscopes that can



Friction Force Microscopy, Fig. 1 Experimental setup of an atomic force microscope/friction force microscope based on the so-called laser beam deflection method. Cantilever deflection (i.e., normal force) and torsion (friction) of the cantilever are measured simultaneously by detecting the lateral and vertical deflection of a laser beam while the sample is scanned in the *x*-*y*-plane. The laser beam deflection is determined using a 4-quadrant photodiode: If *A*, *B*, *C*, and *D* are proportional to the intensity of the incident light of the corresponding quadrant, the signal $(A + B) - (C + D)$ is a measure for the deflection and $(A + C) - (B + D)$ a measure for the torsion of the cantilever. A schematic of the feedback system is shown by *solid lines* on the *left*. The actual deflection signal of the photodiode is compared with the set point chosen by the experimentalist. The resulting error signal is fed into the PID controller ("*PID*" stands for a feedback electronics that uses a proportional, an integral, and a differential control part), which moves the *z*-position of the scanner in order to minimize the deflection signal

record bending and torsion of the cantilever simultaneously are referred to either as *friction force microscopes* (FFMs) or, alternatively, as *lateral force microscopes* (LFMs).

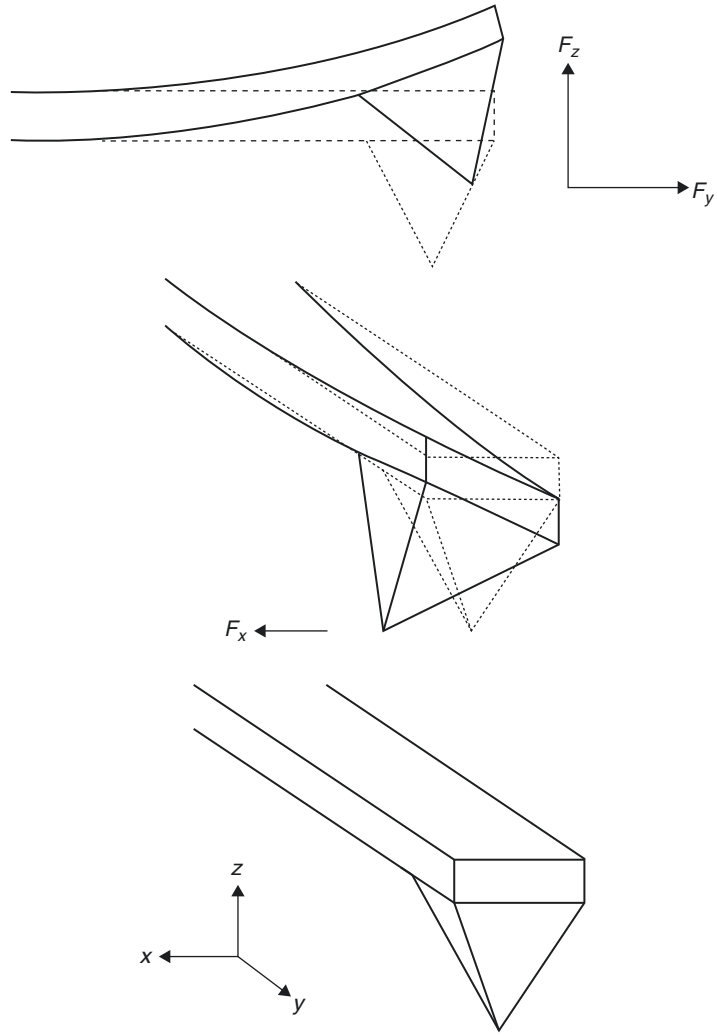
To be able to collect quantitative data, the correct calibration of the bending and the torsion of the cantilever (see Fig. 2) is an important issue. Many parameters have to be known, such as the cantilever dimensions, the elastic modulus of the cantilever material, the tip length, the exact position on the cantilever backside where the laser beam is reflected, and the sensitivity of the 4-quadrant photodiode. Different procedures for the quantitative analysis of lateral force microscopy experiments have been introduced; for a detailed review, see Munz [7].

As already mentioned in the introduction, macroscopic friction experiments are often difficult to analyze due to the complexity of the experimental system. Measurements using friction force microscopy, however, can be carried out in nearly every environment (ambient conditions, vacuum, or liquids) with well-defined tips representing approximative point contacts. Since loading forces can be kept very low, effects like wear and plastic deformation can be prevented in most cases. These features make friction force microscopy a powerful tool for nanotribology.

An example of a friction force microscopy measurement is shown in Fig. 3a, b. The images represent a measurement of a DPPC film adsorbed on a mica substrate. The lateral force was simultaneously recorded with the topography and shows a contrast between the DPPC film and the substrate. This effect can be attributed to the different frictional forces on DPPC and the mica substrate and is frequently used to obtain a chemical contrast on flat surfaces.

Atomic-Scale Friction and the Prandtl-Tomlinson Model

Despite the fact that much of the current knowledge on the origins of friction has been acquired during the last two decades, many of the newly



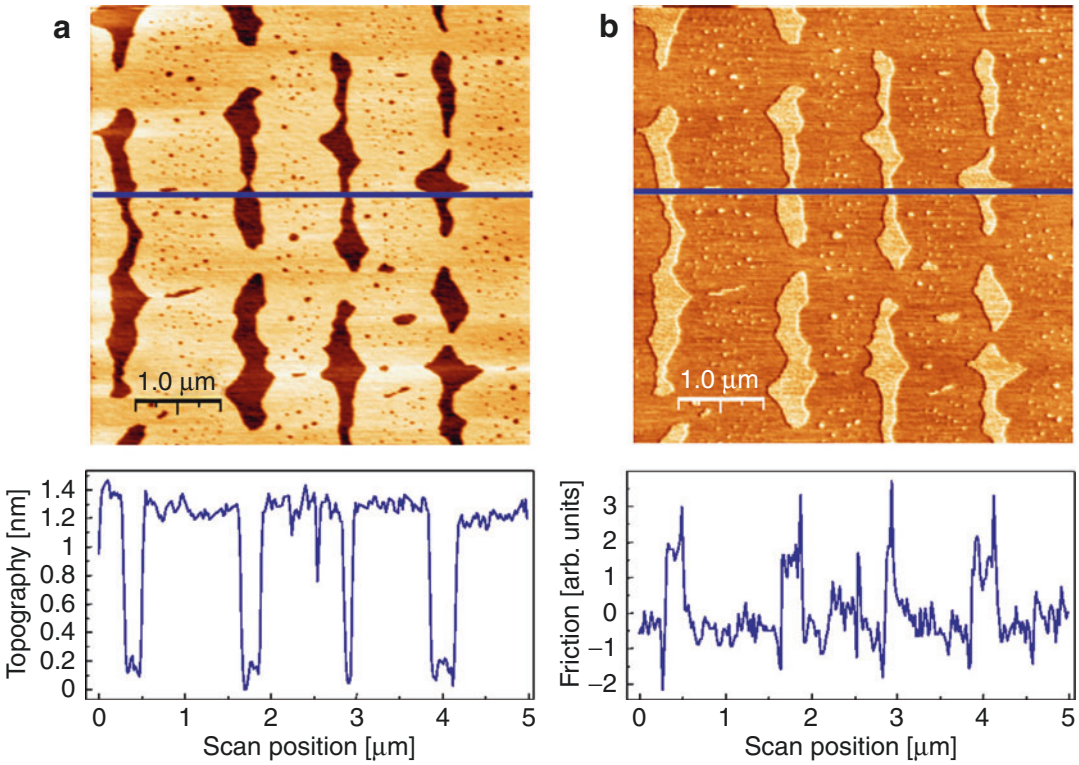
Friction Force Microscopy, Fig. 2 When the cantilever tip is sliding over the surface, there are two main modes of deflection. The *top* image shows the normal bending mode which is related to the normal forces exerted by the tip. The *middle* image shows the torsion of the cantilever beam, which is due to the lateral forces between tip and sample, i.e., the frictional forces

observed atomic-scale effects can actually be understood in the framework of a surprisingly old model, the so-called Prandtl-Tomlinson model [8, 9]. A schematic representation of this simple spring model is shown in Fig. 4. A point-like tip is elastically coupled to a main body M with a spring possessing a spring constant c_x in x-direction and interacts with the sample surface via a periodic potential $V_{\text{int}}(x_t)$, where x_t reflects the actual position of the tip. During sliding, the body M is moved in x-direction with a velocity V_{M}.

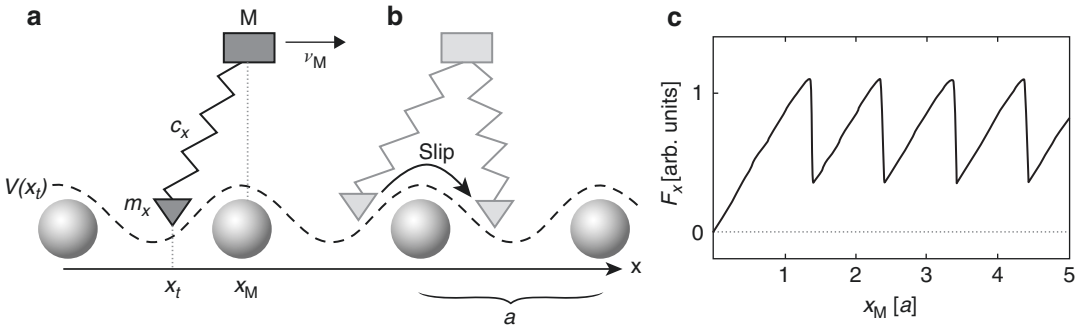
The equation of motion for the tip in the interaction potential can be written as

$$m_x \ddot{x}_t = c_x(x_{\text{M}} - x_t) - \frac{\partial V_{\text{int}}(x_t)}{\partial x_t} - \gamma_x \dot{x}_t, \qquad (2)$$

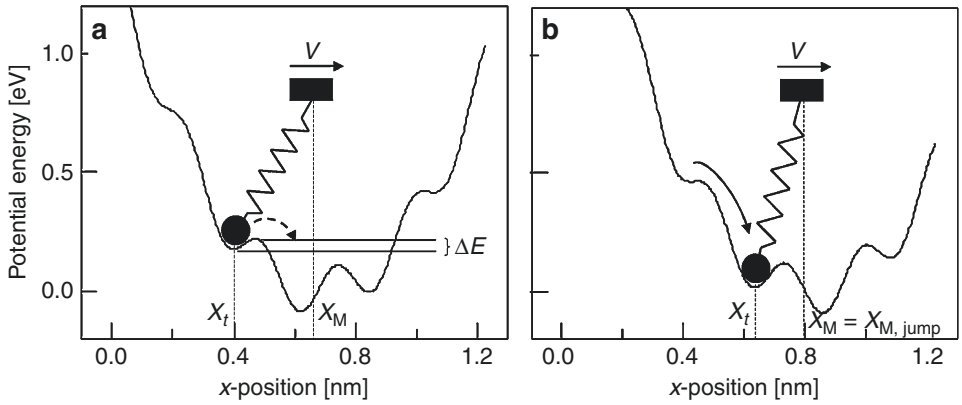
where m_x is the effective mass of the system and $x_{\text{M}} = V_{\text{M}}t$ the equilibrium position of the spring at the time t. The last term on the right side is a simple velocity-dependent damping term with a damping constant γ_x. It describes the energy dissipation of the tip independent from the exact physical nature of the actual dissipation channel (i.e., phonons or electronic excitations). The solution of this differential equation is the path of the tip $x_t(t)$. The *lateral force* F_x to move the tip in x-direction can be calculated from $F_x = c_x(x_{\text{M}} - x_t)$,



Friction Force Microscopy, Fig. 3 (**a**) Atomic force microscopy image obtained in contact mode of a monomolecular DPPC (L-*a*-dipalmitoyl-phosphatidylcholine) film adsorbed on mica. The image is color-coded, i.e., *dark areas* represent the mica substrate and *light areas* the DPPC film. (**b**) The simultaneously recorded friction image shows lower friction on the film (*dark areas*) as on the substrate (*light areas*). The graphs at the *bottom* represent single *scan lines* obtained at the positions marked by a *dark line* in the images at the *top*



Friction Force Microscopy, Fig. 4 (**a**) A simple model for a tip sliding on an atomically flat surface based on the Prandtl-Tomlinson model. A point-like tip is coupled elastically to the body M by a spring with spring constant c_x in x-direction; x_t represents its position within an external potential $V(x_t)$ with periodicity a. If $x_t = x_M$, the spring is in its equilibrium position. For sliding, the body M is moved with a velocity v_M in x-direction. (**b**) A schematic view of the tip movement in a sinusoidal interaction potential. If condition (5) holds, the tip shows a characteristic "stick-slip"-type movement, i.e., it jumps from one potential minimum to another. (**c**) If the tip moves with stick-slips over the sample surface, the lateral force F_x manifests as a sawtooth-like function



Friction Force Microscopy, Fig. 5 Potential diagram of the Prandtl-Tomlinson model at zero temperature. The indicated net potential represents the sum of a fixed sinusoidal interaction potential and a parabolic spring potential with its minimum at the position x_M, which is moving with the velocity v from *left* to *right* (cf. Eq. 3). In (**a**), where $x_M \approx 0.66$ nm, the tip is trapped in a local minimum and is separated from the next minimum to the right by an energy barrier of height ΔE. For (**b**), the tip base (and with it the minimum of the parabolic potential) has moved to the *right* ($x_M \approx 0.8$ nm). As a consequence, the energy barrier has vanished, which causes the tip to "jump" to the next local minimum. At nonzero temperature, however, there is a finite probability that even in (**a**) the tip might jump over the energy barrier due to thermal activation (*dashed arrow*)

whereas the *friction force* F_{friction} is defined as the averaged lateral force $\langle F_x \rangle$.

Within this model, the point-like tip represents the average of the actual tip-sample contact or single-asperity contact, which might include hundreds of atoms. Alternatively, the actual tip-sample contact can also be treated as a system consisting of many individual atomically sharp tips interacting through springs, which leads to more complex models. A prominent example is the model proposed by Frenkel and Kotorova [10], which is based on chains of elastically coupled atoms exposed to a periodic potential. However, here the discussion will be restricted to the simple model introduced above, which has been proven to successfully describe the tip motion for many materials.

If the sliding velocities are low, analytical solutions of Eq. 2 can be derived with the assumption that the tip is always in its stable equilibrium position, which corresponds to the actual minimum of its total energy:

$$E_{\text{tot}} = \frac{1}{2} c_x (x_t - x_M)^2 + V_{\text{int}}(x_t). \qquad (3)$$

For better clarity, the corresponding energy is plotted in Fig. 5 for two positions x_M and a sinusoidal interaction potential. In Fig. 5a, the tip is trapped in a local energy minimum. An energy barrier of height ΔE prevents the tip from reaching the next energy minimum to the right. Since the tip base M moves along the x-direction, the energy barrier diminishes with time until the local minimum vanishes at a certain position $x_M = x_{M,\text{jump}}$. As a result, the tip "jumps" to the next local minimum where it is trapped again. This situation is sketched in Fig. 5b. The "jump height," i.e., the maximum force due to spring tension just before the "slip" event occurs (cf. Fig. 4c), is a direct function of the tip-sample interaction potential V_{int}, the spring constant c_x, and the lattice constant a. Due to the periodicity of the interaction potential, this "stick" and "slip" mechanism repeats as long as the support M is scanned over the sample surface. For finite temperatures, it has to be considered that the tip might sometimes jump thermally activated over a nonvanishing potential barrier ΔE to reach the next local minimum. This case will be discussed later.

For further analysis, the two conditions for a local minimum ($\partial E_{\text{tot}}/\partial x_t = 0$ and $\partial^2 E_{\text{tot}}/\partial x_t^2 > 0$) must be considered. Then the path of the tip $x_t(x_M)$ can be determined from the solution of the equation

$$c_x(x_M - x_t) = \frac{\partial V_{\text{int}}(x_t)}{\partial x_t}. \tag{4}$$

For a stiff spring, there is only one solution for all x_M, resulting in a continuous tip movement and vanishing friction, since the averaged lateral force $\langle F_x \rangle$ is zero. However, the tip movement changes dramatically if the condition

$$c_x < -\left[\frac{\partial^2 V_{\text{int}}}{\partial x_t^2}\right]_{\min} \tag{5}$$

is fulfilled. Now, the tip moves discontinuously in a "stick-slip"-type motion over the sample surface, jumping from one potential minimum to another. This specific movement of the tip results in a sawtooth-like function for the lateral force F_x (see Fig. 3c). Since the averaged lateral force $\langle F_x \rangle$ is nonzero in this case, a finite friction force F_{friction} is necessary to move the body M in x-direction.

In fact, it is exactly this behavior that frequently occurs in atomic-scale friction force microscopy experiments. Observed first by Mate et al. [1] for graphite(0001), experimentalists found atomic-scale stick-slip motion on many different materials such as mica, NaF, TeS_2, KBr, MoS_2, $MoTe_2$, as well as metal surfaces like Cu and Au. Figure 6 shows an example of stick-slip friction on highly oriented pyrolytic graphite (HOPG).

A complete transition from stick-slip to pure sliding was reported by Socoliuc et al. [11] for the NaCl(001) surface. Depending on the actual loading force of the tip, the measured lateral forces changed from a sawtooth-like shape to a continuous curve where forward and backward forces overlap. As a consequence, there is no hysteresis, and the frictional force reduces to ultralow values as expected from the above-discussed model. Assuming a simple sinusoidal potential of the form

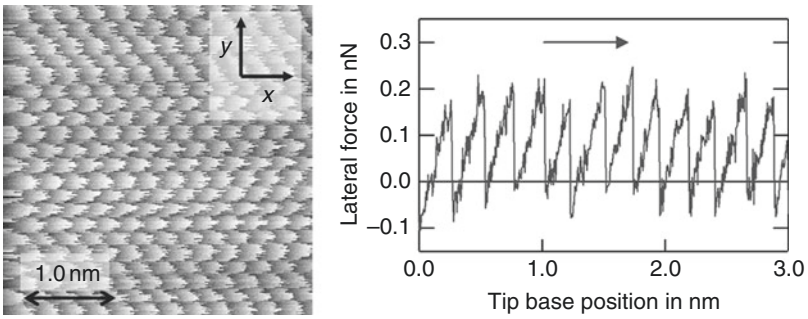
$$V_{\text{int}} = -\frac{E_0}{2}\cos\left(\frac{2\pi}{a}x_t\right) \tag{6}$$

to describe the effective tip-sample interaction, the variation of the loading force can be inherently considered by the amplitude E_0. Inserting this potential into condition Eq. 5, it is useful to define the parameter

$$\eta = \frac{2\pi^2 E_0}{c_x a^2}. \tag{7}$$

If $\eta > 1$, condition Eq. 5 is fulfilled and stick-slip occurs. Continuous sliding and no dissipation are observed for all $\eta \leq 1$.

So far, only the jump of the tip into the immediate neighboring potential minimum was considered. However, larger slips into the next but one minimum are also possible. Furthermore, there might be more than one minimum and maximum within one unit cell. In such a case, the stick-slip behavior depends on the actual loading force since



Friction Force Microscopy, Fig. 6 On the *left side*, the lateral force signal of an FFM scan on HOPG is shown (3 × 3 nm). On the *right side*, the graph shows a single lateral force trace from the scan where the expected stick-slip behavior is clearly seen by the sawtooth-type line profile. Measurements were done under ultrahigh vacuum conditions

F

this parameter influences the height of the interaction potential.

Since stick-slip movement has been identified as the underlying mechanism for atomic-scale friction, it would be of preeminent technological importance if strategies could be found that allow to prevent or suppress this effect. It has been known for several years that the actuation of the tip-sample contact by small oscillations reduces friction. A recent study by Socoliuc et al. [12] demonstrated elegantly the possible efficiency of such an approach on the atomic scale. In their experimental setup, a small oscillatory force was added to the normal force by an external electrostatic potential applied between cantilever and sample. If that oscillatory force was modulated with a frequency equivalent to one of the vertical resonance frequencies of the tip-sample contact, the stick-slip motion of the cantilever and thus friction itself could be reproducibly switched on or off.

The original Prandtl-Tomlinson model considered only the one-dimensional movement of a tip along the x-direction. As real surfaces are two-dimensional, any somehow realistic model has to include at least the x-y-periodicity of the crystalline lattice. A two-dimensional extension of the Prandtl-Tomlinson model is discussed in detail in [13].

Thermal Activation and the Velocity Dependence of Friction

As shown above, the Prandtl-Tomlinson model and its extensions are quite successful in describing the qualitative overall behavior of atoms at nanometer-scale dry, single-asperity contacts in relative motion. This theory is based on the idea that the tip "slips" from one energy minimum to the next if the energy barrier ΔE between the current minimum and the next one vanishes because of the changes induced by the cantilever movement on the total energy Eq. 3. Since these jumps are typically much faster than the scan velocities applied in friction force microscopy, the frictional forces can be assumed to be nearly constant for scan velocities well below the slip velocities.

However, this simple picture does not hold at finite temperatures. As already indicated in Fig. 5, the tip might actually jump over a nonvanishing energy barrier ΔE due to thermal activation, leading to reduced frictional forces. Even though this effect has already been discussed in the original publication by Prandtl [8], it is still a topic of current research.

The velocity dependence on atomically flat surfaces has been measured by FFM on different surfaces; most authors report a logarithmic increase of the frictional force with velocity, which might be explained by a phenomenological "thermally activated Eyring model" [14].

Nonetheless, recent theoretical analyses of these phenomena are often based on a thermally activated Prandtl-Tomlinson model. A good starting point is the following "master equation" [15]:

$$\frac{dp(t)}{dt} = -f_0 \exp\left(-\frac{\Delta E(t)}{k_B T}\right) p(t), \qquad (8)$$

where $\Delta E(t)$ reflects the energy barrier at the time t and f_0 the resonance frequency of the tip in its actual minimum. The probability that the tip does *not jump* is denoted as $p(t)$.

For the analysis of experimental data, the variable t must be changed to the actual lateral force F_x [16]. After this transformation, one finds

$$\frac{dp(F_x)}{dF_x} = -\frac{f_0}{cxv_M} \exp\left(-\frac{\Delta E(F_x)}{k_B T}\right) p(F_x). \qquad (9)$$

The maximum of the probability for jumping can be obtained from the condition d^2p/dF_x^2 and leads to

$$\frac{\partial \Delta E}{\partial F_x} = -\frac{k_B T f_0}{c_x v_M} \exp\left(-\frac{\Delta E(t)}{k_B T}\right). \qquad (10)$$

This equation can be solved for an analytical description of ΔE. To do so, [15, 16] assumed a linear dependence of the energy barrier on F_x. However, most jumps occur very close to the

critical position, where the energy barrier is relatively small and more adequately described by a function of the type $\Delta E = \frac{1}{\beta}(F_{\max} - F_x)^{3/2}$, where $F_{\max}$ is the maximum possible lateral force occurring at zero temperature and β a constant. Introducing this relationship into Eq. 10, one gets an implicit function for the frictional force [17]:

$$\ln\left(\frac{v_0}{v_M}\right) - \frac{1}{2}\ln\left(1 - \frac{F_x}{F_{\max}}\right) = \frac{1}{\beta k_B T}(F_{\max} - F_x)^{3/2}. \quad (11)$$

Here, the parameter v_0 has been defined as

$$v_0 = \frac{2f_0\beta k_B T}{3c_x\sqrt{F_{\max}}} \quad (12)$$

and is usually referred to as the "critical velocity" [18].

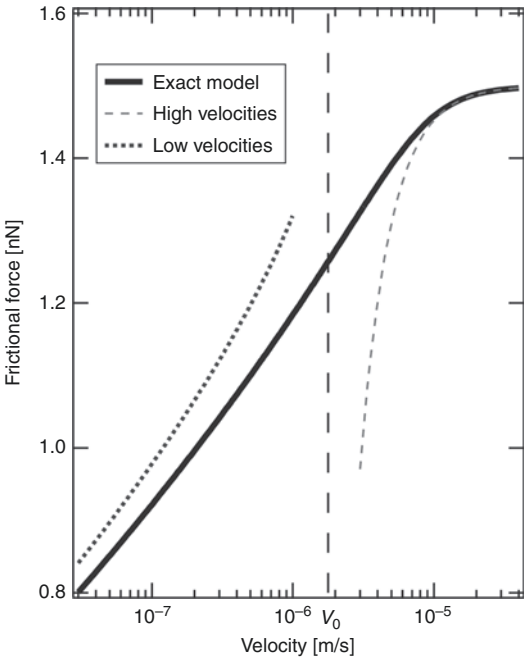
The nature of v_0 as a "critical velocity" is best explored by approximating Eq. 11 for a "low-velocity" and a "high-velocity" regime relative to v_0. As shown in Ref. [18], this approach results in

$$F_{\text{low}}(v_M) = F_{\max} - \left(\beta k_B T \ln\left(\frac{v_0}{v_M}\right)\right)^{2/3} \quad (13)$$

for the low velocity limit and in

$$F_{\text{high}}(v_M) = F_{\max}\left(1 - \left(\frac{v_0}{v_M}\right)^2\right) \quad (14)$$

for the high velocity limit. Figure 7 compares the resulting force curves for the exact and approximated solutions using the parameters $T = 300$ K, $c_x = 1.0$ N/m, $f_0 = 25$ kHz, $\beta = 10^6$ N$^{3/2}$ J^{-1}, and $F_{\max} = 1.5$ nN, leading to a "critical velocity" of $v_0 = 1.78 \times 10^{-6}$ m/s. The solid line represents the exact solution of the implicit Eq. 11. Note that the frictional force increases logarithmically until it reaches a plateau for velocities higher than the critical velocity v_0. The logarithmic increase can be described by the approximation Eq. 13, while the dependence near the plateau is well reproduced by Eq. 14. Therefore, v_0 represents the velocity where the logarithmic dependence transitions toward the plateau value. The vertical dashed line in Fig. 7 labels the "critical velocity" v_0 and marks this border. A comparison of these theoretical curves to experimental force versus velocity curves for different loads on mica can be found in Ref. [17].



Friction Force Microscopy, Fig. 7 The frictional tip-sample force in the ramped creep model. The *solid line* represents the exact equation (Eq. 11). The *dotted* and *dashed lines* represent the approximations for *low* (Eq. 13) and *high* (Eq. 14) velocities, respectively

Future Research Directions

While it seems that the stick-slip processes of atomic friction are sufficiently well understood by now, the nature of the energy dissipation occurring during the "slip" event is still unclear. Usually it is assumed that phonons are excited, which eventually thermalize and thus dissipate the energy stored in the spring during the "stick" process. However, the exact underlying mechanisms are unclear. Recently several experiments

F

with the FFM have been conducted as an explicit function of the sample temperature, in order to investigate in detail the above-discussed effect of thermal activation of the friction processes. Those results point toward an even more complex situation, with a frictional enhancement effect at cryogenic temperatures [19].

Of further interest is the question, how friction depends on the true area of contact. Since the contact geometry of the tip apex is usually poorly defined, the alternative approach of nanoparticle manipulation has been developed. Here the FFM tip is used to slide well-defined nanoparticles along the surface while simultaneously measuring the lateral force signal, which in those cases is mainly due to the interfacial friction between the particle and the surface.

This allows to determine the true area dependence of friction from nano- to microscales [20].

Thus, in summary, the method of friction force microscopy has significantly advanced the understanding of microscopic friction mechanisms, which nurtures the hope that eventually the full control of frictional properties of moving surfaces in mechanical contact can be reached.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Nanotribology

References

1. Mate, C.M., McClelland, G.M., Erlandsson, R., Chiang, S.: Atomic scale friction of a tungsten tip on a graphite surface. Phys. Rev. Lett. **59**, 1942–1945 (1987)
2. Bhushan, B., Israelachvili, J.N., Landmann, U.: Nanotribology: friction, lubrication and wear at the atomic-scale. Nature **374**, 607–616 (1995)
3. Urbakh, M., Klafter, J., Gourdon, D., Israelachvili, J.: The nonlinear nature of friction. Nature **430**, 525–528 (2004)
4. Bhushan, B., Marti, O.: Scanning probe microscopy – principle of operation, instrumentation, and probes. In: Bhushan, B. (ed.) Nanotribology and Nanomechanics – An Introduction, pp. 41–115. Springer, Berlin/Heidelberg (2005)
5. Israelachvili, J.N., Tabor, D.: The shear properties of molecular films. Wear **24**, 386–390 (1973)
6. Meyer, E., Hug, H.-J., Bennewitz, R.: Scanning Probe Microscopy – The Lab on a Tip. Springer, Berlin (2004)
7. Munz, M.: Force calibration in lateral force microscopy: a review of the experimental methods. J. Phys. D Appl. Phys. **43**, 063001 (2010)
8. Prandtl, L.: Ein Gedankenmodell zur kinetischen Energie der festen Korper. Z. Angew. Math. Mech. **8**, 85–106 (1928)
9. Tomlinson, G.A.: A molecular theory of friction. Philos. Mag. J. Sci. **7**, 905–939 (1929)
10. Frenkel, Y., Kontorova, T.: On the theory of plastic deformation and twinning. Phys. Z. Sowjetunion **13**, 1–10 (1938)
11. Socoliuc, A., Bennewitz, R., Gnecco, E., Meyer, E.: Transition from stick-slip to continuous sliding in atomic friction: entering a new regime of ultralow friction. Phys. Rev. Lett. **92**, 134301 (2004)
12. Socoliuc, A., Gnecco, E., Maier, S., Pfeiffer, O., Baratoff, A., Bennewitz, R., Meyer, E.: Atomic-scale control of friction by actuation of nanometer-sized contacts. Science **313**, 207–210 (2006)
13. Hölscher, H., Schirmeisen, A., Schwarz, U.D.: Principles of atomic friction: from sticking atoms to superlubric sliding. Philos. Trans. R. Soc. A **366**, 1383–1404 (2008)
14. Bouhacina, T., Aime, J.P., Gauthier, S., Michel, D., Heroguez, V.: Tribological behavior of a polymer grafted on silanized silica probed with a nanotip. Phys. Rev. B **56**, 7694–7703 (1997)
15. Gnecco, E., Bennewitz, R., Gyalog, T., Loppacher, C., Bammerlin, M., Meyer, E., Guntherodt, H.-J.: Velocity dependence of atomic friction. Phys. Rev. Lett. **84**, 1172–1175 (2000)
16. Gnecco, E., Bennewitz, R., Gyalog, T., Meyer, E.: Friction experiments on the nanometre scale. J. Phys. Condens. Matter **13**, R619–R642 (2001)
17. Riedo, E., Gnecco, E., Bennewitz, R., Meyer, E., Brune, H.: Interaction potential and hopping dynamics governing sliding friction. Phys. Rev. Lett. **91**, 084502 (2003)
18. Gyalog, T., Gnecco, E., Meyer, E.: Stick-slip motion on the atomic scale. In: Gnecco, E., Meyer, E. (eds.) Fundamentals of Friction and Wear on the Nanoscale, pp. 101–115. Springer, Berlin (2006)
19. Barel, I., Urbakh, M., Jansen, L., Schirmeisen, A.: Multibond dynamics of nanoscale friction: the role of temperature. Phys. Rev. Lett. **104**, 066104 (2010)
20. Dietzel, D., Feldmann, M., Schwarz, U.D., Fuchs, H., Schirmeisen, A.: Scaling laws of structural lubricity. Phys. Rev. Lett. **111**, 235502 (2013)

Friction-Reducing Sandfish Skin

Werner Baumgartner
Johannes Kepler University, Linz, Austria

Synonyms

Officinal skink; Sand skink; *Scincus officinalis*; *Scincus scincus*

Definition

The sandfish (*Scincus scincus*) is a lizard having the remarkable ability to move in desert sand in a swimming-like fashion over significant distances at remarkable velocities up to about 300 mm s^{-1} when submerged in sand. It is well adapted through evolution to live in loose aeolian sand by different morphological, physiological, and behavioral features. The skin of the sandfish exhibits a very small friction coefficient for sand and high abrasion resistance.

Biological Background

The scincid lizard genus *Scincus* [1], a group within the order Squamata, is distributed over the huge desert belt ranging from the African west coast (Morocco to Senegal) through the Sahara and the Arabian Peninsula into Jordan, Iraq, and SW Iran [1–3]. The sandfish *Scincus scincus* [1] (Fig. 1) is known for its ability to move through desert sand very fast, as if it were swimming [1–6]. This is in contrast to other desert reptiles which can bury themselves in the sand to hide from predators, but do not move in the sand over significant distances. It was found that the sandfish exhibits an in-plane meandering motion and reaches velocities of up to 300 mm s^{-1} [5, 6] when completely submerged in sand. It is currently under debate whether or not the limbs are used to propel the animal forward. Two groups did studies using different approaches and obtained different results. One group used nuclear magnetic resonance (NMR) imaging in sand [5] finding limb usage, whereas the second group used X-ray imaging in plastic granulate finding the sandfish to adpress the limbs onto the body during movement [6]. However, it might well be that both studies are right. First of all, different subspecies of *Scincus scincus* might be adapted to



Friction-Reducing Sandfish Skin, Fig. 1 An adult sandfish in the hand of the researcher. The snout-vent length (length of the animal from snout to anus) is about 120 mm

different sand types (grain size, friction coefficient, microstructure) resembling their origin from different deserts. Furthermore, the sandfish might have several different modes of locomotion, for example, fast flight from predators and normal, low energy-consuming movement. This has to be clarified in future studies.

Anyhow, the life underneath loose, aeolian sand requires morphological key adaptations, such as a shovel-shaped snout with the lower jaw wedged into the upper jaw, reduced ear openings and nostrils, a subquadrangular cross section of the body, and strongly developed limbs with fringed digits and toes. These adaptations are described by early authors [1–3]. However, other adaptations to life underneath the sharp-edged sand are subject to current research. Besides the locomotion [5, 6], the sensory system, and the breathing mechanisms [7], the integument of the sandfish, especially the scales, are of interest and have been examined in recent years. The latter cover the entire body in an imbricate manner, typical for Squamata [8].

Generally, the integument (from Latin *integere* = to cover) is the organ system that covers the body and protects it from water loss and/or damage. The integument consists of the skin and its derivatives like scales, feathers, hairs, and nails and has a variety of functions: besides mechanical protection and prevention of water loss from lower tissue layers, it serves also for temperature regulation and as a transmitter for tactile stimuli. The integument of lizards (Fig. 2) consists of several layers, the most important being the epidermis, the dermis, and the subcutis [8]. The subcutis is the connective tissue that covers the underlying muscle layers. The dermis consists of connective tissue which, in the case of some reptiles like the sandfish, contains bony plates; these plates are called *osteoderms* and resemble the shape of the covering scales. The epidermis is organized as scales and consists of an outer beta-layer mostly composed of beta-keratin, an inner alpha-layer built up by alpha-keratin, and a meso-layer separating the former two layers. The beta-layer is covered by the so-called Oberhäutchen which often exhibits particular micro- and nanostructures (microdermatoglyphics). Sandfishes, like all other squamates, renew their horny (keratinized) epidermis by periodic shedding (ecdysis). For *Scincus scincus* the ecdysis interval is of the order of 2–3 months.

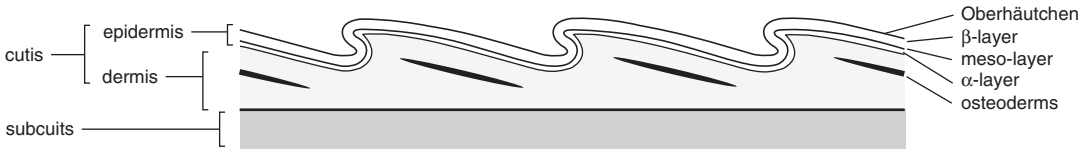
Low Friction and Abrasion of Sandfish Scales

The integument and especially the scales of the sandfish exhibit some interesting properties. It was found that the friction between sand and sandfish scales is outstandingly low.

Generally, dry friction resists relative lateral motion of two solid surfaces in contact. The two regimes of dry friction are static friction between nonmoving surfaces and dynamic friction between moving surfaces. Coulomb friction [9] is an approximate model used to calculate the force of dry friction. Dynamic friction is governed by the equation:

$$F_f = \mu \cdot F_n \tag{1}$$

where F_f is the force exerted by friction, μ is the friction coefficient which is a property of the contacting materials, and F_n is the normal force exerted between the surfaces. The direction of the frictional force is opposite to the direction of motion.



Friction-Reducing Sandfish Skin, Fig. 2 Principle layered assembly of lizard skin

Friction-Reducing Sandfish Skin, Table 1 Values for friction angle Θ and friction coefficient μ for sand on different materials. *Eumeces schneideri*, the Berber skink, is a close relative of the sandfish that does not perform sand swimming. Values from [4, 10, 11] are mean values from at least 10 independent measurements; the standard deviation is below 1°

Material	Θ	μ
Sandfish scales	21°	0.38
Eumeces scales	35°	0.70
Glass	28°	0.53
Polished steel	25°	0.46
Teflon	36°	0.73
Nylon	30°	0.57

For the sandfish, the friction was quantified by determination of the so-called friction angle Θ, i.e., the lead angle of an inclined plane covered with the material under investigation at which loose sand starts to slide off. The friction angle is related to the friction coefficient μ as

$$\mu = \tan(\Theta) \tag{2}$$

As shown in Table 1, the friction angle and thus also the friction coefficient for the matching of sand and different materials are lowest for the sandfish scales when compared with technical materials or with the scales of other reptiles like a closely related but not sand-swimming lizard, the Schneider's skink (or Berber skink) *Eumeces schneideri*. Generally it was found that sand-swimming lizards show low friction similar to the sandfish, while non-sand-swimming species have significantly higher friction [12].

Determination of the friction angle is a very simple method for determining the friction of granular media on a native surface. However, one should keep in mind that a certain degree of rolling of the granules may take place leading to an underestimate of the friction coefficient. However, if one is interested in a comparison of the friction on different materials, this effect is negligible.

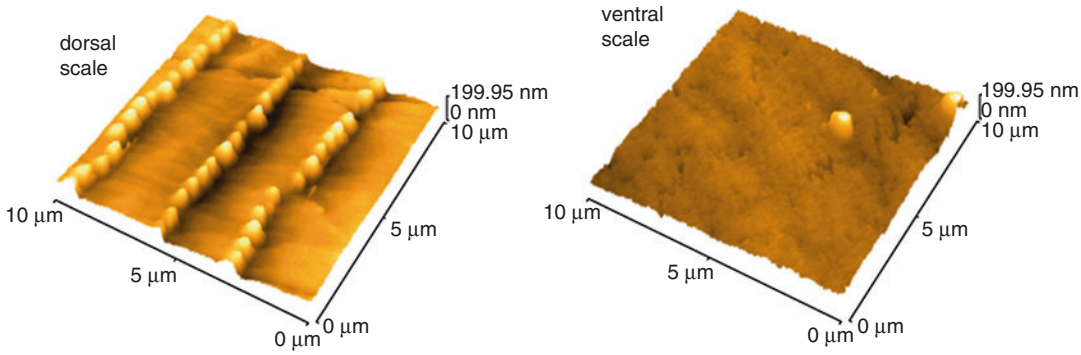
Hand in hand with the low friction, the scales show an impressive resistance to abrasion by sand. Hardly any engineering material can withstand abrasion through repeated interaction with fine quartz sand. However, the sandfish scales outperform Teflon, glass, and even steel in terms of abrasion resistance. This was measured by quantifying the increase of surface roughness after mild sand blasting [10, 11]. For this sand was poured through a funnel from above on the animal or material samples. The height was adjusted to 100–300 mm. While for glass some seconds of this sand blasting are sufficient to scratch the surface, the sandfish scales show no detectable scratches even after 60 min of this treatment.

This is consistent with the observation that the sandfish normally appears to have a gleaming surface. This is remarkable bearing in mind that the animal moves through sharp-edged sand at high velocities.

Structural Properties of Sandfish Scales

Often, the outstanding properties of biological surfaces can be understood based on their micro- or nanostructure or hierarchical surface structuring. Indeed scanning electron microscopy (SEM) as well as atomic force microscopy (AFM [13]) revealed that all dorsal scales have an interesting topography, namely, sawtooth-shaped striae with a comblike edge [4, 12] (Fig. 3). The edges are separated by ~5 μm and the step height is ~250 nm. The comblike arranged features have dimensions of approximately 50–200 nm. Although interesting, these structures were found to have nothing to do with the low friction and the abrasion resistance of the sandfish scales. These structures are less pronounced on the lateral scales and are completely absent from the ventral scales where the surface is completely flat. However, the resistance against abrasion and the friction parameters were found to be about identical on the dorsal and on the ventral scales.

Furthermore, the sawtooth-shaped striae were also found on all other Scincidae [12] investigated so far, such as the before-mentioned *Eumeces schneideri.* These animals, however, do not show the low-friction behavior like the sandfish. This indicates that the surface structure (i.e., the topography) is not responsible for the remarkable



Friction-Reducing Sandfish Skin, Fig. 3 AFM image of a dorsal and a ventral sandfish scale. Notably the dorsal scales exhibit sawtooth-shaped striae with a comblike edge while the ventral scales are mainly flat without micro- or nano-ornamentation

tribological properties but may be important for other aspects of living in the desert sand.

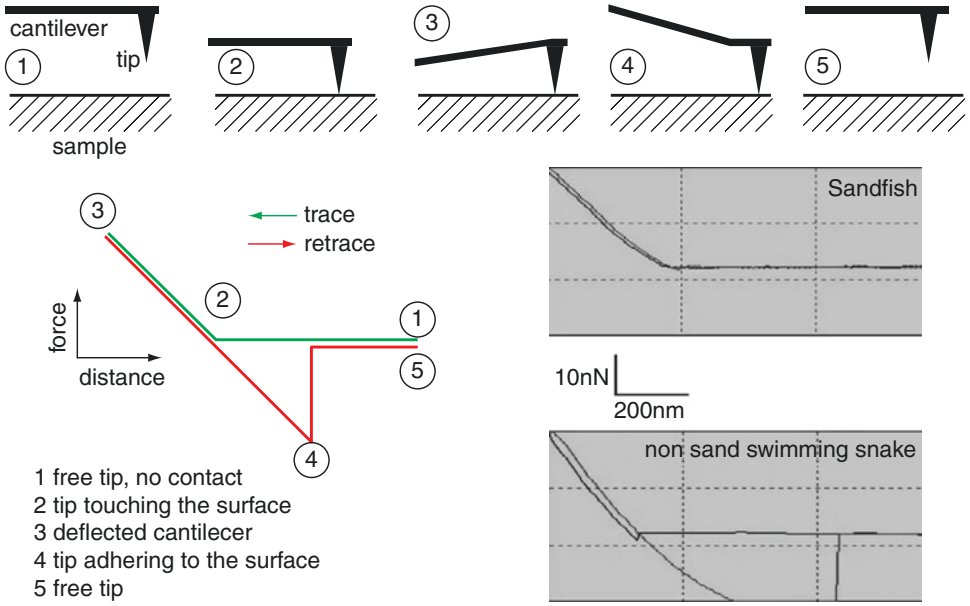
Several functional biomaterials like wood, arthropod cuticle, and bones reveal interesting material properties by structural anisotropies, for example, by fiber enforcement [9]. However, transmission electron microscopy and force modulation AFM [13] allow one to search for material inhomogeneities. Such studies show that the scales are more or less mechanically uniform and do not exhibit detectable anisotropies with respect to either stiffness or elastic modulus [4].

Mechanical Properties of Sandfish Scales

The stiffness and linked to this the reduced (or effective) elastic modulus can be determined by nanoindentation [13, 14]. An indenter of well-characterized shape (typically a pyramid) is pressed at known force into the surface and the total and sustaining (i.e., plastic) deformations are determined. Based on the knowledge of the indenter shape and area, load-displacement curves allow the determination of the resilience, the elastic modulus (Young's modulus) [4, 13, 14], and the hardness of the scale material of the sand skink. The elastic recovery of sandfish scales is comparable to values of other reptiles and at 70+ % exceeds standard technical materials like aluminum alloy (40 %) and steel (X6CrNi MoTi17-12-2, 44 %) [4]. The reduced elastic modulus values measured for the sandfish scale are on the order of about 0.5–1 GPa [14]. This is significantly higher than the elastic modulus of reptile scales of non-sand-swimming reptiles, but only a fraction of the reduced elastic modulus of stainless steel [4, 14]. The tensile strength is comparable to aluminum alloy.

In order to investigate the adhesive properties of sandfish scales, so-called force-distance measurements [4, 10–12, 14] were performed (Fig. 4). In this mode of operation, an AFM tip is first approached ("trace") and then retracted ("retrace") from a probe surface in a cyclic movement at a fixed lateral position. Attractive forces between tip and probe are evident from the jump-on of the tip to the probe at short distances during trace ("pull on") and from the sticking of the tip to the probe, followed by a concomitant jump-off at a critical force ("pull off"), during retrace. For a good overview of the nature of forces occurring between tip and probe, see [15].

As shown in the recorded force-distance curves [4] (Fig. 4), no interaction between the AFM tip and the sandfish scales was observed. Typically, on other surfaces, an interaction between the surface and the tip due to, e.g., van der Waals forces or electrostatic interaction takes place which yields a characteristic discontinuity (jump) in the force-distance data [13, 14]. However, this was never observed when measuring sandfish scales [4]. The interaction force, if present, cannot be distinguished from noise. This is remarkable and



Friction-Reducing Sandfish Skin, Fig. 4 Schematic of force-distance cycles and example of force-distance cycles obtained for sandfish scales and scales of the non-sand-swimming snake *Pantherophis guttatus*

in contrast to force-distance measurements on scales from related but not sand-swimming reptiles where clear adhesion-evoked jumps in the data can be observed [4]. The tips used consisted of silicon or silicon nitride, both showing the same results, resembling pretty much the properties of quartz sand. Thus, a specific property of sandfish skin is the missing adhesion with silicates, which appears to be, at least in part, responsible for the low friction of the lizard in the desert sand. Due to the nanometer-sized tip radius, it is reasonable to say that the surface structures observed, i.e., the striae and the comblike edges, are not responsible for the effect. The absence of any attractive force between tip and scale was observed on ventral, lateral, and dorsal scales taken from different animals.

Chemical Properties of Sandfish Scales

The chemical composition of sandfish scales is purely organic, lacking hard inorganic material like silicates or lime [4]. This was found using energy-dispersive X-ray spectroscopy (EDX). To stimulate the emission of characteristic X-rays from a specimen, a high-energy beam of electrons is focused onto the sample. The resulting X-rays, i.e., the *bremsstrahlung*, can be analyzed with respect to the wave lengths allowing the clarification of the atomic composition. The dominant detectable elements, i.e., the elements with atomic mass above beryllium, were carbon (83.5 $\pm$ 2 %), nitrogen (9.1 $\pm$ 1 %), oxygen (5.9 $\pm$ 1 %), and, interestingly, a rather high amount of sulfur (1.5 $\pm$ 0.3 %). The sulfur is most likely involved in a high degree of molecular cross-linkage via disulfide bonds between or within proteins. Disulfide bonds in proteins are formed between the thiol groups of cysteine residues. The other sulfur-containing amino acid, methionine, cannot form disulfide bonds. The disulfide bond stabilizes the folded form of a protein in several ways:

(i) It holds two parts of the protein covalently together, biasing the protein toward the folded topology.
(ii) The disulfide bond may form a hydrophobic core of the folded protein, i.e., local hydrophobic residues may aggregate around the disulfide bond through hydrophobic interactions.

(iii) The abovementioned two effects allow the disulfide bond to link segments of the protein chain, lowering the effective local concentration of water molecules. Water molecules attack hydrogen bonds and tend to break up some secondary structure. Thus, disulfide bonds stabilize secondary structure in its vicinity.

Thus, it is most likely that the higher sulfur content is responsible for the significantly higher stiffness of sandfish scales compared to the scales of non-sand-swimming reptiles.

The proteins are mainly β-keratins and small amounts of α-keratins and matrix proteins which are typically present in reptile scales in small amounts [8, 11, 12]. This was shown by SDS-PAGE and Western blotting. Furthermore, one corresponding gene has been sequenced recently [12]. Generally, keratin refers to a family of fibrous proteins which are a key structural material making up rigid and insoluble strong unmineralized tissues found throughout the vertebrates. The fibers, with diameters of 7–15 nm, can form networks and meshes by intermolecular interactions thereby yielding the mechanical properties. An interesting difference between keratins in the scales of sandfish and other non-sand-swimming lizards is the high glycosylation of the sandfish keratins. While non-sand-swimming reptiles exhibit only small traces of sugar side chains of the proteins, the β-keratins of the sandfish are highly and complex glycosylated [12, 14]. Sugar side chains are an extremely flexible, complex, and powerful chemical tool in protein biology used to modify the properties of proteins or protein complexes. It is evident that the sugars play an important role in the reduction of friction as sandfish scales after sugar removal show high friction toward sand [4]. It was shown that these glycans are sufficient to reduce the friction of sand to different materials like glass [11] or different polymers [14]. How exactly the sugars perform this task is still enigmatic. The exact structure, composition, and function of this sugar side chains in the sandfish scales have to be determined in future studies.

Future Directions for Research

The sandfish and closely related lizards appear to be an interesting object of research with respect to many different aspects. Of course the low friction and the high abrasion resistance of the scales are highly interesting. Detailed studies on the exact chemical composition and the 3D structure, especially of the sugar side chains of the β-keratins, might allow the development of low friction and abrasion-resistant technical surfaces. Besides this, the function of the sawtooth-shaped striae on the dorsal scales should be investigated. However, the life in aeolian sand might even hide further secrets which are worth being discovered.

Cross-References

- ▶ AFM
- ▶ Atomic Force Microscopy
- ▶ Bioadhesion
- ▶ Force Modulation in Atomic Force Microscopy
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Nanotribology
- ▶ Shark Skin Drag Reduction

References

1. Linnaeus, C.: Systema naturæ per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis, pp. 1–824. Laurentii Salvii, Holmiæ (1758)
2. Arnold, E.N., Leviton, A.E.: A revision on the lizard genus *Scincus*. Bull. Br. Mus. Nat. Hist. (Zool) Lond. **31**, 189–248 (1977)
3. Arnold, E.: Identifying the effects of history on adaptation – origins of different sand-diving techniques in lizards. J. Zool. **235**, 351–388 (1995)
4. Baumgartner, W., Saxe, F., Weth, A., Hajas, D., Sigumonrong, D., Emmerlich, J., Singheiser, M., Bohme, W., Schneider, J.M.: The sandfish's skin: morphology, chemistry and reconstruction. J. Bionic. Eng. **4**, 1–9 (2007)
5. Baumgartner, W., Fidler, F., Weth, A., Habbecke, M., Jakob, P., Butenweg, C., Böhme, W.: Investigating the locomotion of the sandfish in desert sand using NMR-imaging. PLoS One **3**(10), e3309 (2008)
6. Maladen, R.D., Ding, Y., Li, C., Goldman, D.I.: Undulatory swimming in sand: subsurface

locomotion of the sandfish lizard. Science **325**, 314–318 (2009)
7. Vihar, B., Wolf, C., Böhme, W., Fiedler, W., Baumgartner, W.: Respiratory physiology of the sandfish (Squamata: Scincidae: Scincus scincus) with special reference to subharenal breathing. Salamandra **51**(2), (2015), in press
8. Alibardi, L., Toni, M.: Immunological characterization and fine localization of a lizard beta-keratin. J. Exp. Zool. B Mol. Dev. Evol. **306**(B), 528–538 (2006)
9. Vogel, S.: Comparative Biomechanics: Life's Physical World. Princeton University Press, Princeton (2003)
10. Staudt, K., Saxe, F., Schmied, H., Böhme, W., Baumgartner, W.: Sandfish inspires engineering. Proc. SPIE **7975**, 79751B (2011)
11. Staudt, K., Böhme, W., Baumgartner, W.: Comparative investigations of the sandfish's β-keratin (Reptilia: Scincidae: Scincus scincus). Part 2: Glycan-based friction reduction. J. Biomimetics Biomater. Tissue Eng. **16**, 1–9 (2012)
12. Staudt, K., Saxe, F., Schmied, H., Soeur, R., Böhme, W., Baumgartner, W.: Comparative investigations of the sandfish's β-keratin (Reptilia: Scincidae: Scincus scincus). Part 1: Surface and molecular examinations. J. Biomimetics Biomater. Tissue Eng. **15**, 1–16 (2012)
13. Ikai, A.: The World of Nano-Biomechanics: Mechanical Imaging and Measurement by Atomic Force Microscopy. Elsevier, Oxford (2007)
14. Vihar, B.: Mimicking the abrasion resistant sandfish epidermis. Dissertation RWTH-Aachen University (2015)
15. Israelachvili, J.: Intermolecular and Surface Forces, 3rd ed. Academic press, Waltham, San Diego, Oxford, Amsterdam (2011)

Fuel Cell

► Nanomaterials for Electrical Energy Storage Devices

Fullerenes for Drug Delivery

Robert D. Bolskar
TDA Research, Inc., Wheat Ridge, CO, USA

Synonyms

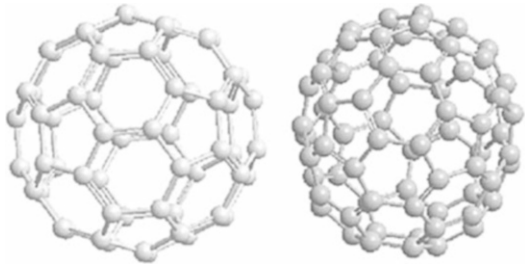
Buckminsterfullerene; C_{60}; Metallofullerene; Nanoscale drug vector

Definition

Fullerenes are nanomolecular carbon cages that can serve as platforms for the delivery of drugs and imaging agents. There are several functional modes in which fullerenes can act as drug delivery nanomaterials. Fullerenes can have a direct bioactivity such as antioxidant activity when surface functionalized. Fullerenes can also serve as drug vectors or drug delivery scaffolds with non-covalent linkages or with covalent linkages between the fullerene and a bioactive moiety. Endohedral fullerenes are a subclass of fullerenes that contain atom(s) in their interior spaces. Endohedral fullerenes offer a robust shell that can prevent contact between the body and the entrapped atom(s), serving, for example, in contrast enhancement for magnetic resonance imaging and as a carrier for radionuclides.

Brief Introduction to Fullerenes and Endohedral Fullerenes

Fullerenes, the third allotrope of carbon after diamond and graphite, are closed-cage carbon molecules initially discovered in the gas phase during investigations of the laser desorption of graphite [1]. Fullerenes are the simplest among the newer forms of nanocarbons, as fullerenes are molecular while the other nanocarbons (including single-walled carbon nanotubes, multi-walled carbon nanotubes, carbon nanohorns, carbon onions, graphene, etc.) are decidedly not. C_{60} is the archetype fullerene of the fullerene family with its icosahedral symmetry and all-carbon cage in the shape of a soccer ball (Fig. 1). Fullerenes with larger cages, including C_{70}, C_{74}, C_{76}, C_{78}, C_{84}, etc., are termed "higher fullerenes." The fullerene cages are comprised of hexagons and pentagons, with alternating carbon–carbon bonds of double bond character (at the hexagon–hexagon junctions) and of single bond character (at the hexagon–pentagon junctions). Most known fullerene structures obey the so-called isolated pentagon rule (IPR) in which fullerenes avoid having directly adjacent pentagons to minimize cage strain. Thus, fullerenes of certain symmetries



Fullerenes for Drug Delivery, Fig. 1 Ball and stick depictions of *Ih*-C_{60} (*left*) and *D5h*-C_{70} (*right*) structures

are more stable and thus more prevalent than others. The larger the fullerene cage, the greater number of possible structural isomers there are.

Several processes are now used to produce fullerenes, including the direct current arc discharge (arcing) of graphite in an inert, reduced pressure atmosphere and the combustion of hydrocarbons in sooting flames. The arc process can be used to produce fullerenes on multigram scales, while the combustion process is capable of generating far larger quantities; for example, Frontier Carbon Corporation of Japan has recently scaled up the combustion-based production of fullerenes to multiton quantities. The fullerenes are produced as mixture of the different C_{2n} combined with non-fullerene carbonaceous soot. C_{60} is the dominant fullerene product, with smaller amounts of higher fullerenes produced. Many of the fullerenes are soluble in organic solvents, particularly in methylated and halogenated arenes. The soluble C_{2n} fullerenes like C_{60} and C_{70} have large HOMO-LUMO gaps, reflective of their stability. Solvent extraction is commonly used to retrieve fullerenes from the as-produced soot matrix, and different soluble C_{2n} cages and their isomers can be separated by chromatography. There are certain fullerenes with small-HOMO-LUMO gaps that cause them to be insoluble as neutral species, including D_{3h}-C_{74} and I_h-C_{80}. Interestingly, insoluble C_{2n} fullerenes with $2n > 100$ comprise a large fraction of fullerenic combustion-produced soot. Fullerenes are generally electronegative in nature due to the effects their curved surfaces have on their electronic structures. Fullerene electronegativity has a dominating influence on their chemical reactivities. Fullerenes are redox active, with multiple reduced (negatively charged) states generally more accessible than oxidized (positively charged) states.

Besides the empty fullerenes of different cage sizes and symmetries, endohedral fullerenes comprise another major structural variant. Fullerenes, being pseudo-spherical closed-cage molecules, have a hollow interior space protected from the exterior. C_{60}, the archetypical fullerene, is about 0.7 nm in diameter, so its interior space is large enough to hold up to several atoms. Endohedral fullerenes with metals inside – the "metallofullerenes" – are the most commonly studied endohedrals [2]. Metallofullerene formulae are denoted by "M@C_{2n}," with M the metal, C_{2n} the fullerene cage, and the @ symbol indicating the encapsulation of M. Metallofullerenes are commonly produced by the carbon arc method using graphite rods doped with metals or metal oxides of Sc, Y, La, and the lanthanide series (and some actinide) elements. The metallofullerene arc product contains a mixture of empty and endohedral fullerene cages, much of which can be solvent extracted from the soot. Compared to empty fullerenes, endohedral fullerenes are produced in much smaller amounts. Whether soluble or insoluble, metallofullerenes need to be separated from the empty fullerenes having similar properties. The soluble metallofullerenes, such as M@C_{82}, can be separated from soluble C_{2n} with chromatography. The insoluble M@C_{2n} such as M@C_{60}, M@C_{70}, and M@C_{74} require different processes for access. In general, the soluble M@C_{82} species are less prevalent than the combined insoluble M@C_{2n} species. The properties, structures, and reactivities of numerous mono-metal metallofullerenes such as La@C_{82} and dimetallofullerenes such as La_2@C_{80} have been well characterized [2]. In general, the highly electropositive endohedral metal atom(s) donate electrons to the electronegative fullerene cages to form metal cations entrapped in a fullerene anion (fulleride). Lanthanides in metallofullerenes are almost always in the M^{3+} state (with several exceptions). The multi-electron donation to the fullerene cage is the reason in many cases for

why metallofullerenes have properties that are significantly different from their empty fullerene counterparts.

Some years after metallofullerenes were discovered, a new metallofullerene variant was discovered – the "trimetallic nitride templated" metallofullerenes or "TNTs" (also termed metallic nitride endohedral metallofullerenes, clusterfullerenes, trimetaspheres, etc.) [2]. These metallofullerenes contain a central nitrogen atom (nitride) bonded to three metals ions with a general formula of $M_3N@C_{2n}$. $Sc_3N@C_{80}$ was the first of its type reported, and many other $M_3N@C_{2n}$ species have since been produced and isolated with lanthanides. The electron transfer from the endohedral nitride cluster stabilizes their fullerene cages. TNTs are produced by incorporating a nitrogen source (including N_2, NH_3, cyanamide or guanidinium salts) into the process of arcing metal-doped graphite rods. The TNTs are soluble in organic solvents (similar to the $M@C_{82}$ metallofullerenes), and they can be extracted and chromatographically purified. Improvements have been found that increase the relative fraction of TNT metallofullerenes to empty fullerenes produced by the arc synthesis. Mixed-metal TNT clusters have also been produced and isolated, as well as structures with non-IPR cages.

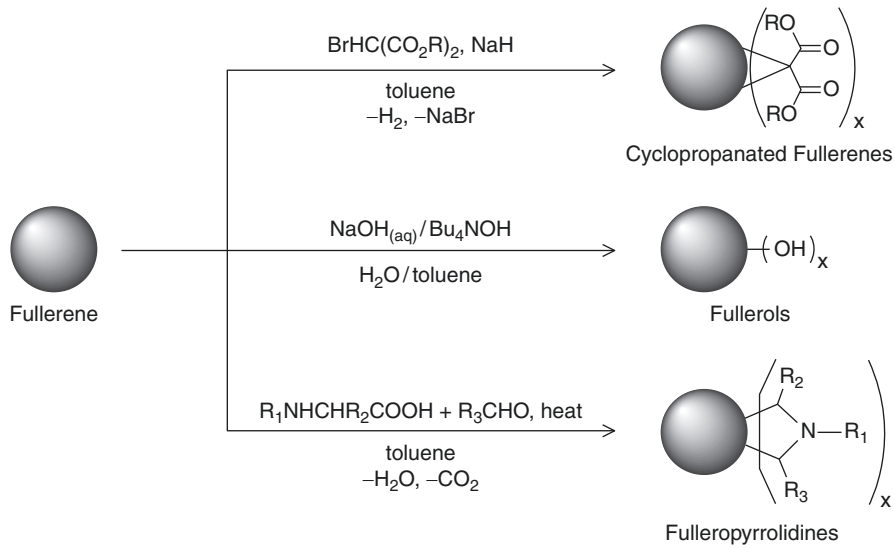
The diversity of fullerene structures is expanding through discoveries of new cages and different endohedral contents, and exceptions to fullerene structural "rules" are becoming more prevalent. Many new empty C_{2n} and endohedrals have been isolated in minute quantities from solvent extracts of the mixed products of arc production, including non-IPR isomers and fullerenes beyond those with standard C_{2n}-only cages. Examples include $C_{2n}Cl_x$ derivatives (sub-C_{60} cages such as $C_{50}Cl_{10}$ and non-IPR cages intermediate to standard fullerenes, i.e., $C_{64}Cl_4$, $C_{72}Cl_4$, etc. recovered in small quantities via in situ chlorination reactions) and endohedrals with novel clusters ($La_2@C_{72}$, $Sc_2@C_{66}$, $Sc_3N@C_{68}$, $M_2C_2@C_{2n}$, $Sc_4C_2@C_{80}$, $M_4O_2@C_{80}$, $Sc_2(\mu_2\text{-}O)@C_{82}$, $M_2S@C_{2n}$, $Sc_3NC@C_{80}$, etc.).

Alternatively, certain metal and nonmetal endohedral fullerenes can be produced by non-arc methods. $Li@C_{60}$, an endohedral produced by ion implantation of C_{60}, was recently isolated and characterized (as a monocation salt with $SbCl_6^-$). Nonmetal endohedrals include noble gas endohedrals (produced in small quantities under very high pressures), $N@C_{2n}$, and fullerenes containing small molecules such as H_2 and H_2O that are produced via open-cage derivatives into which small molecules are inserted, followed by synthetic cage closure.

Fullerene Derivatives

A minimum level of derivatization is required to use fullerenes in drug delivery applications due to their all-carbon surfaces. Native, unmodified fullerenes are soluble in organic solvents (particularly aromatic solvents such as benzene, toluene, halogenated arenes, etc.) but they are essentially insoluble in water, the solvent of life. Soon after they were first produced in macroscopic amounts, a rich chemistry for fullerene functionalization evolved [1, 3]. Given their closed cage structures, fullerene derivatization chemistry is naturally dominated by addition reactions, and their chemistry is strongly influenced by their electronegative natures. Prominent, useful derivatization methods include polyaddition of free radicals, nucleophilic attack (attachment of anionic and neutral nucleophiles), and abundant styles of cycloaddition reactions across the fullerene cage carbon–carbon double bonds. Derivatizations adding more than one group to the fullerene surface give mixtures of isomeric products, which can be separated using chromatography.

The derivatizations can give fullerenes increased solubility in organic media or in aqueous media depending on the choice of attached functionality. Among the many water-soluble fullerene derivatives that can be synthesized [4], three main synthetic approaches have shown to be particularly valuable (Fig. 2). These three are cyclopropanation (in various forms, but particularly the Bingel–Hirsch reaction of halomalonates and base with fullerenes in organic solution, to form methanofullerenes), polyhydroxylation



Fullerenes for Drug Delivery, Fig. 2 Three commonly used fullerene derivatization reactions

(synthesis of fullerols, typically using biphasic organic/caustic media), and the cycloaddition of azomethine ylides (1,3 dipolar cycloaddition to form fulleropyrrolidines), also known as the Prato reaction.

Typical solubilities for hydrophilic fullerene derivatives in water range up to ca. 100 mg C_{60} mL^{-1}. Several uncharged derivatives were synthesized having much higher aqueous solubilities (for example, malonodiserinolamide C_{60} derivatives with >240 mg C_{60} mL^{-1} and β-cyclodextrin-appended C_{60} derivatives with ca. 350 mg mL^{-1}). Water-soluble fullerene derivatives are well known to spontaneously aggregate intermolecularly. Their nanocluster aggregates have hydrodynamic diameters ranging from tens to hundreds of nm (typically with pH dependence) as determined with dynamic light scattering and other techniques.

Separately, underivatized fullerenes can be formulated in water as dispersions, suspensions, and colloidal "nanoparticle clusters," etc.; some are prepared by ultrasonication or precipitation of native C_{60} powders. Non-derivatized fullerenes have also been coaxed into water via non-covalent association with cyclodextrins and related species. With the emphasis on drug applications of functionalized derivatives, these are not generally discussed in this entry.

Drug Delivery with Nanomaterials and Fullerenes

Drug delivery with new nanotechnologies is poised to make significant improvements to human health care in the areas of selective, targeted imaging and individualized therapies [5]. The properties of nanomaterials, including their sizes, shapes, surfaces, and material compositions are well recognized to offer potentially valuable advances over traditional drug delivery. In particular, the imaging and treatment of primary and malignant cancers would benefit greatly from an improved specific drug targeting paradigm. Major goals of selective targeting of a drug agent are to increase effectiveness of the agent while decreasing nonspecific side effects, notably for the chemotherapeutics used in cancer treatment. For example, the chemotherapeutic action of the anticancer drug paclitaxel derives from its inhibition of cellular mitosis. However, the cytotoxic effects of this drug do not discriminate between cancerous cells and noncancerous

cells, which causes deleterious side effects when systemically administered. Additionally, efficient administration and circulatory delivery of otherwise hydrophobic drugs like paclitaxel is problematic. Drug delivery with carefully designed nanovectors can help to overcome poor solubility, side effects, and circulation half-life limitations of the drug molecules, along with lowering the required dosages and deliver more of the drug at the disease site (i.e., tumors, specific organs, etc.).

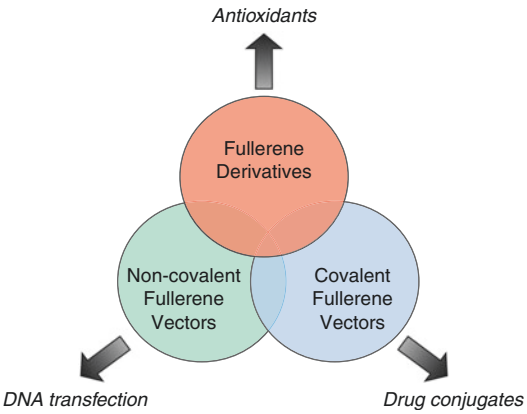
With their all-carbon surfaces, functionalization or derivatization is necessary to endow fullerenes with properties such as water solubility and bioavailability that are essential for their use in biomedicine. When suitably derivatized for water solubility, fullerenes have been shown to be relatively nontoxic both in vitro and in vivo, with toxicity generally decreasing in concert with increased surface functionalization. Biodistribution and circulation characteristics are naturally strongly affected by the composition of the derivative groups. Hydrophobic (underivatized) surface sections do remain on some fullerene derivatives, which may benefit cell membrane crossing of the fullerene with its drug payload. The attached functional groups also serve as a malleable interface for tuning the desired drug delivery and/or action.

Fullerenes as a central vector for drug delivery have valuable advantages over many of the other types of nanoparticulate and nanocrystalline systems being investigated for use in biomedicine. A distinct advantage is that in comparison to typical nanoparticulate materials, derivatized fullerenes have a reproducible nature. This is because the compositions and surface chemistries of almost all nanocrystalline and nanoparticulate materials are not precisely reproducible from particle to particle and preparation to preparation down to the level of individual atoms, as fullerenes and fullerene derivatives (with their molecular nature) are. Their relatively small size, which is regulated by derivatization, is another important advantage for in vivo applications. Fullerenes, even when derivatized, are smaller than many of the nanoparticulate and nanocrystalline systems, some of which are many tens or hundreds of nanometers in diameter. Furthermore, the sizes and shapes of fullerene derivative aggregates, when they do form in aqueous media, are reversible and fluxional in ways not possible with hard nanocrystals and nanoparticles. Fullerenes are also highly versatile in terms of cage structure, surface functionalization, and endohedral contents, making them a well-defined but flexible chemical scaffolding. These properties and advantages enable different styles of medicinal chemistry and drug delivery with fullerenes [6].

F

Exohedral Fullerene Derivatives for Biomedicine and Drug Delivery

There are at least three functional modes by which fullerene derivatives can be employed as drug delivery nanomaterials. First, the fullerene derivative alone can serve as a biomedically active agent. Secondly, the fullerene derivative can serve as a non-covalent carrier of bioactive drug species. Thirdly, the initial fullerene derivative can be secondarily derivatized via covalent attachment of drug molecule. Note that there can be overlap among these three categories with some examples of fullerene derivatives and their drug applications; the delineations are based on structure as well as function (Fig. 3). Prominent examples in these three categories of this currently young but actively developing field are discussed below.



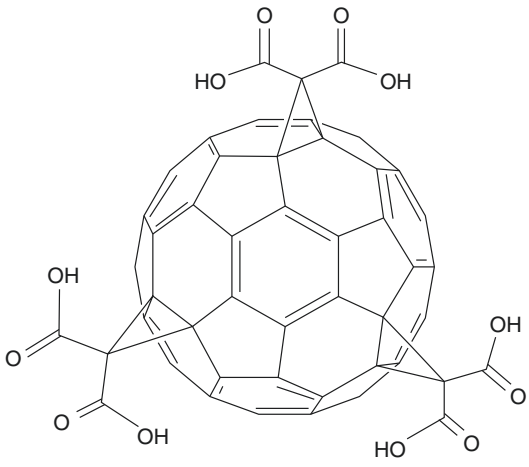
Fullerenes for Drug Delivery, Fig. 3 The three major categories of drug delivery with fullerene derivatives

Surface-Derivatized Fullerenes with Drug Applications

This category of fullerene drug nanomaterial includes those functionalized derivatives that have a biologic action without deliberate external non-covalent association or covalent attachment of drug molecules or drug agents. These actions are derived from special physical properties inherent to fullerenes such as their size, electronegativity, reactivity, photochemistry, etc.

A highly promising biomedical application of fullerene derivatives is their use as antioxidants [7]. Fullerenes and their derivatives are electronegative and readily react with free radicals. With these properties, fullerenes efficiently scavenge the damaging reactive oxygen species (ROS) that are the targets of antioxidants. These include the superoxide radical anion $O_2^{\bullet -}$, the hydroxyl radical $HO^{\bullet}$, and singlet oxygen $^1O_2$, which are produced in settings of oxidative stress and are implicated in inflammation, cell death, neurodegeneration, ischemia/reperfusion, and other disease states.

The most important antioxidant fullerene derivative studied so far is the water-soluble tris-malonate C_{60} derivative $C_{60}[C(COOH)_2]_3$ with C_3 symmetry, commonly termed "C_3" (Fig. 4). This carboxylated C_{60} derivative has been shown to have neuroprotective and life-extending effects in vivo. C_3 afforded life-extending effects in mutant rats having a type of model amyotrophic lateral sclerosis (ALS). C_3 acts as a superoxide dismutase mimetic, serving to destroy superoxide not through stoichiometric scavenging, but through catalytic action. The carboxylated fullerene C_3 also exhibits neuroprotective properties in cell cultures and has beneficial effects when administered to animals bearing neurodegenerative disorders, including Parkinsonism.



Fullerenes for Drug Delivery, Fig. 4 Depiction of C_3–$C_{60}[C(COOH)_2]_3$, a water-soluble antioxidant fullerene derivative

Also under study as antioxidants are fullerene-malonate derivatives having different numbers and arrangements of attached carboxylate groups, fullerols, dendrimer-functionalized C_{60} with terminating carboxy groups, and polyalkylsulfonated fullerenes. Dendrimer-functionalized C_{60} derivatives have been proposed as antioxidants, typically with a singly attached dendrimeric group to minimize disruption of the fullerene cage pi-electrons. The dendro-fullerene termed "DF-1" having carboxy-terminated dendritic branches was evaluated as a radioprotector in live zebrafish embryos. DF-1 afforded similar protection against overall and organ-specific ionizing radiation damage as a clinically approved drug (amifostine) does, with improved action when DF-1 was administered prior to or immediately after the radiation exposure. Tests of DF-1 in live mice showed only modest radiation protection, however, and did not show selective radiation protection in normal cells versus tumor cells. Water-soluble hexasulfobutylated C_{60} derivatives have also been studied as free radical scavengers. Factors that can influence the antioxidant activity of fullerene derivatives include the composition of the functional groups, the positioning of these groups, and their intermolecular aggregation properties. Continued studies of water-soluble fullerene antioxidants are warranted due to the high therapeutic potential of these nanomaterials; so far, the carboxy derivative C_3 has shown the greatest antioxidant utility and neuroprotective effects.

The biologically relevant properties of $C_{60}(OH)_x$ fullerols have been the subject of numerous detailed studies. In one study, the antioxidant properties of fullerols were shown to be at

least partially a result of a binding interaction that blocks glutamate receptors. The radical scavenging ability of fullerols were indicated in studies that showed reduced acute toxicity of the chemotherapeutic doxorubicin (an anthracycline DNA-intercalating anticancer drug), showing potential for mitigating doxorubicin side effects and serving as a cardioprotective agent. $C_{60}(OH)_x$ fullerols also have been shown to have tumor-inhibitory effects via immunostimulation in live mice-bearing tumor models (including activation of T-cells and macrophages) [8]. The low cytotoxicity of the fullerols indicates that the antitumor effect is not the result of direct fullerol toxicity in the cancer cells. Fullerols were also reported to inhibit tumor metastasis in mice models of mammary cancer. Further research is warranted to examine the mechanism behind the antitumor action of fullerols.

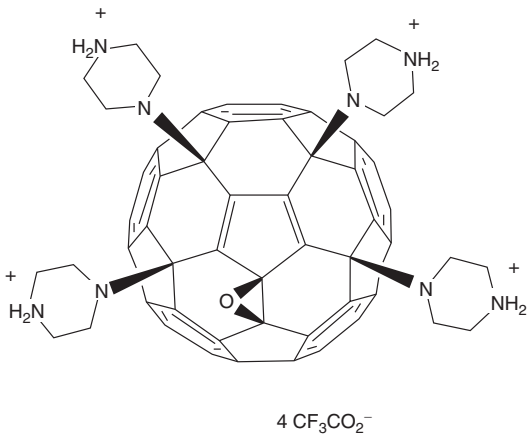
Besides antioxidant radical scavenging, fullerenes and their derivatives also can act as ROS generators upon light irradiation, due to their high, long-lived triplet state yields upon light absorption. In particular, fullerenes are promising photosensitizing agents for photodynamic therapy (PDT), with their ability to produce singlet oxygen (Type I photochemical mechanism) as well as superoxide (Type II photochemical mechanism) [9]. In vitro studies with C_{60} derivatives confirmed they can serve as PDT agents for killing cancer and other cell types, particularly with cationic pyrrolidinium derivatives. The photochemical production of ROS by fullerene derivatives is also being investigated for the photodynamic inactivation of viruses including enveloped viruses, influenza, and HIV protease, as well as in antimicrobial applications.

Fullerene Derivatives with Non-covalent Drugs and Interactions

In the second category, fullerene derivatives with drug delivery applications can be based on non-covalent association of a drug, targeting substrate, or a protein to the fullerene derivative's surface. Fullerenes and fullerene derivatives are well known to form non-covalent intermolecular complexes with species such as polyaromatics, porphyrins, and many electron-rich molecules. The types (and strengths) of these supramolecular interactions vary, and include hydrophobic interactions (particularly for underivatized fullerenes or the underivatized surface portions of lightly derivatized fullerenes), pi-pi interactions, van der Waals interactions, and electrostatic interactions, as well as hydrogen bonding between functional groups on the fullerene derivative and the substrate. Non-covalent intermolecular fullerene interactions are also evident with amphiphilic fullerene derivatives.

Non-covalent interactions can be exploited in medicinal uses of fullerene derivatives. One of the first reported examples in this category was antiviral activity, specifically HIV protease inhibition by C_{60} derivatives. The HIV protease enzyme is crucial for virus replication and ensuing infection of new cells. Its active catalytic site is a relatively hydrophobic cavity or "pocket" that is ca. 1 nm in diameter and very close in size to a C_{60} molecule. With the aid of molecular modeling, C_{60} derivatives were designed to bind in this active site (aided by van der Waals interactions between the fullerene and the cavity) and inhibit the virus [10]. In vitro tests of the first derivative confirmed binding and inhibition, prompting the design and synthesis of new C_{60} derivatives with stronger inhibitory effects [6].

Fullerene derivatives can also be used as non-covalent carriers for the intracellular delivery of nucleic acids. There is great interest in developing new, nonviral artificial vectors for gene transfection, for both research purposes and for gene therapy and vaccination with DNA. Liposomal systems are currently used but do not have high stability in serum, making them inefficient transfection vectors. C_{60} derivatives with attached protonatable amino groups were found to condense DNA into stable nanoclusters with contribution from both electrostatic and van der Waals forces. In vitro tests indicated these fullerene-DNA nanoassemblies transfected cells with DNA at a higher efficiency than a standard liposomal agent. It was recognized that both DNA binding and releasing capabilities were needed in the fullerene derivative for it to efficiently



Fullerenes for Drug Delivery, Fig. 5 A cationic tetraamino C_{60} derivative used for in vivo DNA transfection

transfect. Using the tetra(piperazino) fullerene epoxide derivative depicted in Fig. 5, in vivo DNA transfection in mice was successfully demonstrated without acute toxicity [11]. The C_{60} derivative transfected the green fluorescent protein (GFP) gene to the lung, liver, and spleen. Transfection of an insulin-2 gene reduced the blood glucose levels of live mice. Recently, a different class of water-soluble fullerene derivatives based on cyclopropanated C_{60} structures (including neutral, cationic, and anionic functional groups) were also tested as DNA transfection agents. In vitro tests showed transfection was most efficient with the cationic derivatives. Fullerene derivatives appear to have a high potential for diagnostic and therapeutic gene transfection applications.

Non-covalent interactions between fullerenes and antibodies are also potentially useful for enacting selective targeting of fullerene derivatives. Fullerene immunoconjugates can be either covalent or non-covalent in nature, as was demonstrated in a seminal study of C_{60} immunoconjugates for selective targeting of melanoma with the murine anti-gp240 melanoma antibody ZME-018 [12]. Initially, a covalent conjugate of a C_{60} derivative to the antibody via a disulfide linker was synthesized in several different fullerene-to-antibody ratios. However, upon detailed investigation, it was found that covalent coupling between the C_{60} derivative and the antibody was not necessary to form a stable and useful immunoconjugate; it happened spontaneously upon mixing the antibody and the fullerene derivative (a highly water-soluble C_{60}-malonodiserinolamide) together. ELISA binding curves using antigen-positive cells showed that at high C_{60} derivative-to-antibody molar ratios the antibody to antigen binding is not significantly reduced (relative to the covalent C_{60}-antibody conjugate and a control nonspecific IgG antibody). Non-covalent assembly of fullerene immunoconjugates avoids difficulties inherent in synthetically designing a covalent antibody conjugate, a process that can decrease the antibody's target affinity. It should also be noted that an anti-C_{60} antibody Fab fragment and its binding to fullerene have been well characterized. Non-covalent conjugate formation between fullerene derivatives and antibodies may be a general phenomenon. Other proteins and enzymes with which there is experimental evidence of binding or interactions of C_{60} and/or derivatives include serum albumin (HSA and BSA), nitric oxide synthase, potassium ion channels, proteinases, cytochrome P450, glycosidases, and acetylcholinesterase (AChE).

Fullerene derivatives can also be employed to form non-covalent supramolecular assemblies for drug containment and delivery. As noted earlier, water-soluble fullerene derivatives typically aggregate into nanoclusters. In addition to simple, reversible aggregation, amphiphilicity has been designed into derivative structures for forming fullerene-based micelles, vesicles, and liposomes. A mixed carboxy-dendrimer/polyalkyl functionalized C_{60} amphiphilic derivative (termed AF-1) self-assembled at 70 °C into 100–200 nm vesicles suitable for encapsulating hydrophobic drug molecules inside. Paclitaxel was loaded into the hydrophobic interiors of the AF-1 liposomes, and in vitro tests showed this drug delivery system to suppress growth of MCF-7 human breast cancer cells at a level comparable to Abraxane, a clinically approved injectible formulation of (otherwise hydrophobic) paclitaxel bound to albumin. In different work, an amphiphilic C_{70} bisadduct derivative was used to

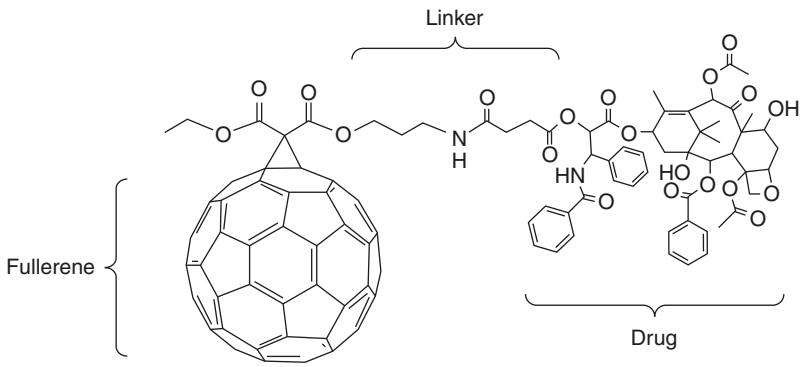
prepare stable, high-fullerene content liposomes co-formed with auxiliary membrane-forming phospholipids. These liposomes retained the fullerene molecule's antioxidant properties, including quenching of reactive oxygen species, while also inhibiting lipid peroxidation. Non-covalent delivery of drug molecules with amphiphilic fullerene structures is very promising.

Fullerene Derivatives with Covalently Attached Drugs and Bioactive Moieties

The third category includes the medicinal fullerene materials that have direct covalent connection(s) between a derivatized fullerene surface and separate, pharmaceutically active component(s). A fundamental nanoparticle drug delivery paradigm is "nanosubstrate-linker-drug." In fullerene-based covalent drug delivery assemblies, these components are (1) the fullerene cage, (2) the surface derivative groups/linker unit, and (3) the attached drug or otherwise biomedically active molecule(s). Derivatized fullerenes with covalently attached drug agents are highly amenable to introducing biofunctions such as controlled drug release while simultaneously providing for selective targeting into one overall design.

The first example of the fullerene-drug covalent conjugate class is a C_{60}-paclitaxel conjugate with slow paclitaxel release for liposomal-based aerosol drug delivery in lung cancer treatment [13]. Paclitaxel was conjugated to a C_{60} derivative via a linker having a hydrolysable ester group for slow release of the free drug (Fig. 6). The conjugate was stable in aqueous 10 % DMSO at physiological pH, and displayed a hydrolysis half-life for paclitaxel release of about 80 min upon exposure to bovine plasma. Stable C_{60}-paclitaxel conjugate liposomes with dilauroylphosphatidycholine (DLPC) were prepared and tested for their antitumor activity with human epithelial lung cancer A594 cells. The mean IC_{50} value for C_{60}-paclitaxel-DLPC (410 nM) was comparable to a paclitaxel-DLPC-only liposome (253 nM). This demonstrated viability of fullerene drug delivery, and combining this type of system with a targeting strategy (i.e., non-covalent antibody conjugation) would offer a powerful new way of delivering prodrug paclitaxel and potentially other chemotherapeutics selectively to cancer cells.

In other studies, doxorubicin was covalently conjugated to C_{60} with hydrophilic ethylene oxide linkers attached to a cyclopropanated C_{60} derivative. In vitro tests with MCF-7 human breast cancer cells showed the fullerene conjugate's cytotoxicity was comparable to native doxorubicin. One-photon confocal and two-photon fluorescence microscopies were used to image the doxorubicin-C_{60} conjugate in the MCF-7 cells. The conjugate distributed in the cytoplasm, in contrast to the localization of native doxorubicin in the cell nucleus. In different work, C_{60}-fullerols were used as the platform derivative to make covalent fullerene conjugates with releasable doxorubicin. These fullerol-doxorubicin conjugates also had co-attached poly(ethylene) glycol units and possibly residual surface hydroxyl groups. These conjugates were subjected to a



Fullerenes for Drug Delivery, Fig. 6 A "fullerene-linker-drug" conjugate assembly for drug delivery (C_{60}-paclitaxel)

variety of in vitro tests with cancer cells, including melanoma cells. Cellular uptake into the lysosomal compartment was demonstrated with fluorescent-labeled conjugates with one cell line. An in vivo test of the doxorubicin-fullerol conjugates in mice with model melanoma tumors showed increased therapeutic index for tumor inhibition and animal weight gain relative to free doxorubicin, which also inhibited tumor growth but led to decreased animal weight. When the doxorubicin-fullerol conjugate was administered to live zebrafish embryos, dose-dependent angiogenesis inhibition was noted. Angiogenesis inhibition in mice-bearing melanoma tumor models was also reported for the conjugate. The above studies with C_{60} conjugates of paclitaxel and doxorubicin indicate that fullerene-vectored delivery of chemotherapeutics will have a high potential for clinical utility with further development.

Fullerenes have been covalently attached to various other biologically relevant and drug moieties, including sugars and carbohydrates, cholesterol, adamantyl groups, the antibiotic vancomycin, viral nanoparticles, nucleosides, and photosensitizer molecules. Bone-targeted fullerene drugs for osteoporosis treatment have been studied, including C_{60} derivatives with appended bis(phosphonate) functionalities. Other potential biomedical uses of fullerene derivatives include contrast and diagnostic agents for medical imaging. Fullerene derivatization chemistry offers means to exohedrally tag fullerene cages with payloads of the active components of contrast agents and/or radioisotopic tracers. A water-soluble, nonionic hexa-iodinated C_{60} derivative was been tested as an x-ray contrast agent and it offered x-ray attenuation equivalent to that of a clinical (non-fullerene) triiodoaryl agent.

Endohedral Fullerene Derivatives for Drug Delivery

Endohedral fullerenes and metallofullerenes can carry and deliver atoms and metal ions that serve as active agents in bioimaging and therapy. Salient metallofullerene properties include a lack of exchange of their interior ion(s) with the exterior media (prohibiting transmetallation), electron transfer from the interior metal(s) to the fullerene cages to form anionic-charged cages with entrapped metal cations, and a cage derivatization chemistry similar to empty fullerenes, but with solubility and reactivity differences in some cases.

With the endohedral contents sequestered inside of the fullerene cage, metallofullerenes can participate in the same fullerene drug delivery concepts noted in Fig. 3. Metallofullerenes also can serve in applications dependent upon the metal's properties and/or derived from interactions between the fullerene cage and its contents. These include magnetic resonance imaging (MRI) contrast agents with endohedral paramagnetic ions and as radiotherapeutic agents holding radioisotopes in their interior cavities.

Metallofullerene Magnetic Resonance Imaging (MRI) Contrast Agents

Clinical magnetic resonance imaging (MRI) is a non-invasive imaging technique that uses the nuclear magnetic resonance signals of protons in tissues and organs to obtain 3-D anatomical information in patients. MRI has high spatial resolution, but has low intrinsic sensitivity. Paramagnetic contrast agents are used to increase the sensitivity of MRI via signal amplification. These agents interact with water protons via inner sphere and outer sphere mechanisms to increase the proton relaxation rates (i.e., decrease the T_1 and or T_2 relaxation times). The characteristic relaxation-enhancing potency of an agent is referred to as "relativity," and the r_1 and r_2 relaxivities vary as a function of agent concentration according to $(1/T_x)_{obsd} = (1/T_x)_d + r_x[\text{agent}]$. Clinically, r_1 is of greater interest due to its image brightening (positive contrast) effect. The paramagnetic Gd^{3+} ion, with its seven unpaired 4*f*-electrons and favorable electron spin relaxation characteristics, is the leading choice for T_1-active relaxation agents. Because the free Gd^{3+} ion is toxic, chelating molecules are required to safely hold the ion, and there are several approved Gd^{3+} chelate agents currently used in ca. 35 % of MRI scans.

A major goal in contrast agent development is increased relaxivity for both higher sensitivity (and thus smaller required dosages) and to enable advanced applications in cellular and molecular imaging. Gd-containing endohedral metallofullerenes, or "gadofullerenes," were recognized early as potential MRI contrast agents with their unique encapsulation of the Gd^{3+} ion. The MRI contrast-enhancing properties of water-solubilized gadofullerenes of $Gd@C_{82}$ and $Gd@C_{60}$ have been investigated in a number of leading studies of polyhydroxylated (fullerol) and carboxylated derivatives [14]. Note that the paramagnetism of $Gd@C_{2n}$ gadofullerenes derives not only from the encapsulated Gd^{3+} ion, but also from the unpaired electron donated by Gd to the fullerene molecular orbital (Gd donates three electrons to the fullerene). While the clinical Gd chelate contrast agents have r_1 relaxivities near 4 mM^{-1} s^{-1}, gadofullerene r_1 relaxivities are significantly higher. For example, $Gd@C_{82}(OH)_x$ and $Gd@C_{60}(OH)_x$ derivatives have reported r_1 values ranging from 20 to 98 mM^{-1} s^{-1}, and a carboxylated $Gd@C_{60}$ derivative had a r_1 maximum of 24 mM^{-1} s^{-1}. Some of the variation in r_1 values derives from the use of different measurement conditions, including temperature and magnetic field. However, solution conditions including sample pH and dissolved salt content had a strong influence on relaxivities as well.

The water-proton relaxation mechanisms of gadofullerene derivatives have recently been examined in a series of detailed experiments [15]. A strong pH dependence on relaxivity was found for the gadofullerenes, with higher relaxivities at lower pH and lower relaxivities at higher pH. Dynamic light scattering and cryo-transmission electron microscopy (TEM) measurements revealed that the gadofullerenes form intermolecular aggregates or nanoclusters with hydrodynamic diameters of ca. 700–800 nm at pH = 7.4, with larger aggregates at low pH while smaller at high pH. This trend mirrors the relaxivity trend as a function of pH, as well as the maximum relaxivity peaks seen for gadofullerenes in relaxivity measurements taken at different magnetic fields (nuclear magnetic resonance dispersion (NMRD) profiles). This latter feature, with broad relaxivity maxima at fields between 0.7 and 1.4 T, is characteristic of long rotational correlation times of larger species with slow tumbling, such as nanocluster aggregates. High dissolved salt concentrations were found to break up gadofullerene derivative nanoaggregates (and consequently decrease relaxivity), with a reported disaggregation half-life of ca. 30 min in 100 mM phosphate buffer. With inner-sphere relaxation mechanisms precluded (direct water-Gd^{3+} coordination is blocked for gadofullerenes), only outer-sphere relaxation can operate. Major factors contributing to the high gadofullerene relaxivities include large paramagnetic surface areas derivatized with hydrophilic and hydrogen-bonding functional groups, nanocluster aggregation that slows rotation, and most significantly, interstitial confinement of water molecules in the clusters. Rapid exchange of interstitial water molecules with bulk water contributes to the high relaxivities. The type of functionalization and number of attached groups employed on gadofullerene surfaces clearly has direct and strong influence on the relaxivities of gadofullerene MRI contrast agents.

Several in vitro cellular imaging studies have been conducted with $Gd@C_{2n}$ gadofullerene derivatives. In one, HeLa cells, mesenchymal stem cells (MSC), and macrophages were labeled with $Gd@C_{82}(OH)_{40}$ using protamine sulfate as a transfection agent to achieve labeling up to 4×10^7 Gd atoms/cell. With the cellular viability remaining high, MR imaging revealed excellent intracellular T_1 weighted contrast. The agents' localization in cytoplasmic endosomes was confirmed by electron microscopy. In another study, marrow stromal cells (MSC) and NIH 3 T3 stem cells were nonspecifically labeled with $Gd@C_{60}[C(COOH)_2]_{10}$, achieving an efficient 10^{11} Gd^{3+} ions per cell on average without adding a transfection agent. The cells remain viable, and an acid strip test indicated the gadofullerenes were intracellular or deep in cellular membranes. T_1-weighted imaging at 1.5 T revealed that labeled cells had MR signal intensities ca. 250 % greater than unlabeled cells. These results demonstrate that gadofullerenes can internalize into cells

and provide efficient T_1-weighted enhancement for potential in vivo cell tracking by MRI.

In vivo T_1-weighted MR imaging tests of $Gd@C_{2n}$ derivative agents have also been conducted. Tests of $Gd@C_{82}(OH)_x$ compounds in mice showed strong signal enhancements. For the compound with $x = 40$, the lung, liver, and spleen had the strongest enhancements, and for the compound reported as $x = 16$, both liver enhancement and renal excretion were seen. A different biodistribution was obtained by MRI in a rat with $Gd@C_{60}[C(COOH)_2]_{10}$. By MRI, this carboxylated derivative showed renal excretion within 1 h, less liver uptake than the gadofullerol compounds and a biodistribution similar to that of clinical Gd chelate agents. These tests indicate different biodistribution tendencies for different functionalization styles, and showed that small animals tolerated the gadofullerene agents without complications.

Gadofullerene derivatives more complex than the carboxylated and hydroxylated structures are beginning to be investigated, including $Gd@C_{82}O_6(OH)_{16}(NHCH_2CH_2COOH)_8$ and the mixed phosphonated derivative $Gd@C_{82}O_2(OH)_{16}(C(PO_3Et_2)_2)_{10}$. Toward more advanced conjugates, a mixed $Gd@C_{82}$ derivative bearing surface hydroxyl and carboxylate groups was covalently conjugated to an anti-green fluorescence protein (GFP) antibody with an amide bond. This immunoconjugate retained antibody activity and displayed a modestly higher relaxivity than the unconjugated $Gd@C_{82}$ derivative.

Water-soluble derivatives of TNT metallofullerenes with Gd ions in their endohedral metal-nitride clusters are also being investigated as MRI contrast agents, particularly with $Gd_3N@C_{80}$ [16]. Unlike lanthanide $M@C_{2n}$, $Gd_3N@C_{80}$ lacks cage-centered unpaired electron(s), so its paramagnetism comes only from the entrapped Gd^{3+} ions of the Gd_3N cluster. The mixed derivative $Gd_3N@C_{80}[DiPEG5000(OH)_x]$ (DiPEG5000 is $[COO(CH_2CH_2O)_{114}CH_3]_2$) had a r_1 maximum (at 2.4 T) of 143 mM^{-1} s^{-1} on a per C_{80} cage basis (48 mM^{-1} s^{-1} on a per Gd basis). T_1-weighted in vivo MR imaging was performed with this agent (after direct brain infusion) in a rat having a glioma tumor model. In the normal brain and tumor tissues, this $Gd_3N@C_{80}$ agent gave high contrast at relatively low concentration with better tumor margin differentiation than was obtained with a standard Gd chelate agent. Related $Gd_3N@C_{80}[DiPEG(OH)_x]$ mixed derivatives with PEG molecular weights of 350 and 750 had r_1 maxima (at 2.4 T) of 237 and 232 mM^{-1} s^{-1} on a per C_{80} cage basis (79 and 77 mM^{-1} s^{-1} on a per Gd basis), respectively.

$Gd_3N@C_{80}(N(OH)(CH_2CH_2O)_nCH_3)_x$ (n = 1,3,6; $x \approx 10$–22) derivatives were characterized, with $n = 6$ having an r_1 value of 205 mM^{-1} s^{-1} on a per C_{80} cage basis (68 mM^{-1} s^{-1} on a per Gd basis). In vivo MRI with the $n = 6$ derivative in a mouse showed this agent remained in the vasculature, as expected with its derivatization motif. $Gd_3N@C_{80}(OH)_{26}(CH_2CH_2COOM)_{16}$ (M = Na or H), with $r_1 = 207$ mM^{-1} s^{-1} on a per C_{80} cage basis (69 mM^{-1} s^{-1} on a per Gd basis), displayed a hydrodynamic diameter of ca. 160 nm and unimodal size distribution in pH = 7 water. Mixed-metal Gd/Sc TNT derivatives have also been synthesized and studied as MRI contrast agents, including $Sc_xGd_{3-x}N@C_{80}O_m(OH)_n$ (x = 1, 2; $m \approx 12$; $n \approx 26$) and a pyrrolidino $Sc_xGd_{3-x}N@C_{80}$ (x = 1, 2) derivative with functionality suited for conjugation. These are of interest for their relaxivity properties and for the greater abundance of $Sc_xGd_{3-x}@C_{80}$ (x = 1,2) as compared to $Gd_3N@C_{80}$.

With r_1 relaxivities an order of magnitude higher than the clinical chelate agents and their stable sequestration of the Gd^{3+} ion, gadofullerene derivatives show strong promise as MRI contrast-enhancing agents for intracellular labeling and in vivo imaging. Exohedral derivatization chemistry has a very strong influence on the relaxivity and intermolecular aggregation properties of the gadofullerene MRI agents. In vivo use will require a balance between relaxivity performance and biodistribution characteristics. The properties of gadofullerene MRI contrast agents are well-suited for applications requiring long circulation times (i.e., blood pool agents and targeted agents) and for multimodal imaging and theranostic agents.

Metallofullerene Radiopharmaceuticals

Nuclear medicine with endohedral radioisotopes, including radionuclear therapy and isotopic tracing and/or imaging, is also of interest with fullerenes as carriers. Metallofullerenes offer carbon-cage confinement of the radionuclide without the potential dissociation of coordinatively held isotopes in a chelate. This containment may in select cases extend to daughter radioisotopes, which is another potentially significant advantage over standard chelation.

Radioisotopic endohedral fullerenes can be generated by at least four methods, including neutron activation of metallofullerenes, direct arc synthesis with radioisotope-doped graphite rods, nuclear recoil implosion and ion implantation. In an early study of metallofullerene radiopharmaceuticals, an in vivo study was conducted with radioactive Ho metallofullerenes. Non-radioactive $^{165}Ho@C_{82}$ was neutron activated to the β-emitting $^{166}Ho@C_{82}$ species. This radio-metallofullerene was water-solubilized by polyhydroxylating its surface, and its biodistribution in mice was characterized by radiochemical methods.

Fullerene surface chemistry can be used to advantage for transport of endohedral radionuclide(s) for radionuclear therapy with both α- and β-emitting radionuclides. Toward this, carboxylate derivatives of the metallofullerene $^{212}Pb@C_{60}$ (produced by recoil following decay of a short-lived parent radioisotope) were investigated as radiocarriers suited for α-particle radioimmunotherapy [17]. TNT metallofullerenes have also been proposed for nuclear medicine agents holding endohedral ^{117}Lu, a β-emitter [18]. A mixed (poly)hydroxy/alkylcarboxy derivative of arc-generated $^{177}Lu_xLu_{(3-x)}N@C_{80}$ was conjugated to an interleukin-13 peptide, demonstrating a potentially targetable radiofullerene agent for brain cancer. In addition to the above examples, fullerene derivatives with exohedral radioisotopes have been reported. These have uses as radiotracers, but exohedral radiofullerenes do not confer the special encapsulation advantages that endohedrals do. Practical application of therapeutic radionuclear endohedrals for treating cancer, etc. will require a balance between the demands of circulation, targeting, and accumulation of a fullerene conjugate versus the inherent half-life and the amount of linear energy transfer upon decay of the chosen radioisotope.

Other Biomedical Applications of Endohedral Fullerenes

Other biomedical applications proposed for lanthanide metallofullerenes include their use as x-ray contrast agents (exploiting the x-ray absorbing properties of heavy metal ions) and as photoluminescent agents (endohedrals encapsulating fluorescing lanthanide cations, e.g., $Er_2@C_{82}$, $Er_3N@C_{80}$, $Er_2C_2@C_{82}$). Current hurdles to realizing these applications include low X-ray contrast and low photoluminescence intensities, respectively.

Metallofullerene fullerol derivatives have also been tested for antitumor activity [19]. As noted earlier, $C_{60}(OH)_x$ fullerols showed antioxidant and antitumor activities in both in vitro and in vivo tests, and related results have been reported from similar tests in small animals with $Gd@C_{82}(OH)_x$ ($x \approx 22$) materials. The derivative $Gd@C_{82}(OH)_{22}$, when compared to the chemotherapeutics CTX (cyclophosphamide) and cisplatin in vivo, exhibited a very high antitumor effect, shrinking tumors at low dose levels. These findings were unexpected because this fullerene derivative is inherently non-cytotoxic (and note that there is no toxicity from free Gd^{3+}). Spin-trapping experiments monitored with electron paramagnetic resonance spectroscopy showed $Gd@C_{82}(OH)_{22}$ scavenged reactive oxygen species and inhibited lipid peroxidation in liposomes. It was suggested that the antitumor activity may be related to the metallofullerol's ability to regulate oxidative stress (by helping to restore liver and kidney functions) and ability to stimulate Th1 cytokine-releasing immune responses. $Gd@C_{82}(OH)_{22}$ was also found to have antiangiogenic properties in live tumor-bearing mice models. Another study indicated that $Gd@C_{82}(OH)_{22}$ has potential to help reverse prostate tumor resistance to cisplatin through

restorative effects on endocytosis [20]. Comparisons of the antitumor activity of metallofullerene derivatives to empty fullerene congeners merit further study (particularly because of the greater availability of empty fullerene derivatives vs. metallofullerenes). Special properties of metallofullerenes afforded to them by the interior metal ion(s) and the electron transfer from metal to cage may be relevant to the mechanism of the observed systemic, indirect antitumor actions.

There are recent indications of greater diversity in fullerene cage and endohedral structures than was initially realized during the early years of fullerene discovery and characterization. These "new" fullerenes include those with cages smaller than C_{60}, non-IPR cages, and diverse, unsymmetrical multi-atom endohedral contents. Their physical properties such as polarity, polarizability, electronegativity, etc. may differ enough from previously known fullerenes to give substantially different attributes for medicinal chemistry and drug delivery. It remains to be seen the extent to which these special structures could contribute to biomedicine, due to the extremely small quantities in which they are produced and isolated.

Future Directions

The special physical properties of fullerenes including size, reactivity, electronegativity, endohedral encapsulation of ions, etc. present opportunities for development of novel drug action and drug delivery. Major advantages of fullerenes as nanovehicles for drug delivery include the reproducible chemistry of molecules, dimensions at the lower end of the nanoscale, diverse exterior covalent and non-covalent chemistries, and endohedral encapsulation of atoms and ions inside of the closed fullerene carbon cages. C_{60} as the "parent" fullerene has been the most investigated fullerene vector in biomedical applications, but the other fullerenes and metallofullerenes are expected to have an expanding role.

The multiverse of fullerene surface chemistry facilitates combining drug agents and therapeutic actions into a single fullerenic theranostic platform. Studies of combined imaging and therapy with multifunctional fullerene nanomaterials are currently underway in several laboratories. Still, the capabilities of fullerenes in drug delivery have only begun to be developed, as most of the examples noted in the preceding sections have not yet moved beyond the basic proof-of-concept stage. Amphiphilic fullerene structures, selective targeting (using conjugated peptides and monoclonal antibodies and their fragments), and fullerene-drug conjugates that are combined with MRI active or radionuclear endohedral fullerenes are topics that are anticipated to be productive. Binding and inhibitory interactions of fullerenes with endogenous proteins must be considered carefully when designing fullerene-based drugs intended for systemic administration in vivo. Continued emphasis on derivative design and synthesis of bioactive fullerene conjugates is needed in the near term. In vivo characterization, pharmacokinetics, long-term toxicity, and other preclinical pharmacology studies will need greater attention before the unique fullerene-based drug delivery nanomaterials can achieve clinical impact.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Functionalization of Carbon Nanotubes
- ▶ Microfabricated Probe Technology
- ▶ Nanoencapsulation
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Nanotechnology
- ▶ Ultrashort Carbon Nanotubes

References

1. Kadish, K.M., Ruoff, R.S.: Fullerenes Chemistry, Physics, and Technology. Wiley Interscience, New York (2000)
2. Chaur, M.N., Melin, F., Ortiz, A.L., Echegoyen, L.: Chemical, electrochemical, and structural properties of endohedral metallofullerenes. Angew. Chem. Int. Ed. **48**, 7514–7538 (2009)
3. Hirsch, A., Brettreich, M.: Fullerenes: Chemistry and Reactions. Wiley-VCH, Weinheim (2005)

4. Nakamura, N., Isobe, H.: Functionalized fullerenes in water. The first 10 years of their chemistry, biology, and nanoscience. Acc. Chem. Res. **36**, 807–815 (2003)
5. Sakamoto, J.H., van de Ven, A.L., Godin, B., Blanco, E., Serda, R.E., Grattoni, A., Ziemys, A., Bouamrani, A., Hu, T., Ranganathan, S.I., De Rosa, E., Martinez, J.O., Smid, C.A., Buchanan, R.M., Lee, S.Y., Srinivasan, S., Landry, M., Meyn, A., Tasciotti, E., Liu, X., Decuzzi, P., Ferrari, M.: Enabling individualized therapy through nanotechnology. Pharmacol. Res. **62**, 57–89 (2010)
6. Bosi, S., Da Ros, T., Spalluto, G., Prato, M.: Fullerene derivatives: an attractive tool for biological applications. Eur. J. Med. Chem. **38**, 913–923 (2003)
7. Beuerle, F., Lebovitz, R., Hirsch, A.: Antioxidant properties of water-soluble fullerene derivatives. In: Cataldo, F., Da Ros, T. (eds.) Medicinal Chemistry and Pharmacological Potential of Fullerenes and Carbon Nanotubes, pp. 51–78. Springer, Berlin (2008)
8. Zhu, J., Ji, Z., Wang, J., Sun, R., Zhang, X., Gao, Y., Sun, H., Liu, Y., Wang, Z., Li, A., Ma, J., Wang, T., Jia, G., Gu, Y.: Tumor-inhibitory effect and immunomodulatory activity of fullerol $C_{60}(OH)_x$. Small **4**, 1168–1175 (2008)
9. Mroz, P., Tegos, G.P., Gali, H., Wharton, T., Sarna, T., Hamblin, M.R.: Fullerenes as photosensitizers in photodynamic therapy. In: Cataldo, F., Da Ros, T. (eds.) Medicinal Chemistry and Pharmacological Potential of Fullerenes and Carbon Nanotubes, pp. 79–106. Springer, Berlin (2008)
10. Friedman, S.H., DeCamp, D.L., Sijbesma, R.P., Srdanov, G., Wudl, F., Kenyon, G.L.: Inhibition of the HIV-1 protease by fullerene derivatives: model building studies and experimental verification. J. Am. Chem. Soc. **115**, 6506–6509 (1993)
11. Nakamura, E., Isobe, H.: In vitro and in vivo gene delivery with tailor-designed aminofullerenes. Chem. Rec. **10**, 260–270 (2010)
12. Ashcroft, J.M., Tsyboulski, D.A., Hartman, K.B., Zakharian, T.Y., Marks, J.W., Weisman, R.B., Rosenblum, M.G., Wilson, L.J.: Fullerene (C_{60}) immunoconjugates: interaction of water-soluble C_{60} derivatives with the murine anti-gp240 melanoma antibody. Chem. Commun. **28**, 3004–3006 (2006)
13. Zakharian, T.Y., Seryshev, A., Sitharaman, B., Gilbert, B.E., Knight, V., Wilson, L.J.: A fullerene-paclitaxel chemotherapeutic: synthesis, characterization, and study of biological activity in tissue culture. J. Am. Chem. Soc. **127**, 12508–12509 (2005)
14. Bolskar, R.D.: Gadolinium endohedral metallofullerene-based MRI contrast agents. In: Cataldo, F., Da Ros, T. (eds.) Medicinal Chemistry and Pharmacological Potential of Fullerenes and Carbon Nanotubes, pp. 157–180. Springer, Berlin (2008)
15. Bolskar, R.D.: Gadofullerene MRI contrast agents. Nanomedicine **3**, 201–213 (2008)
16. Fatouros, P.P., Corwin, F.D., Chen, Z.J., Broaddus, W. C., Tatum, J.L., Kettenmann, B., Ge, Z., Gibson, H.W., Russ, J.L., Leonard, A.P., Duchamp, J.C., Dorn, H.C.: In vitro and in vivo imaging studies of a new endohedral metallofullerene nanoparticle. Radiology **240**, 756–764 (2006)
17. Diener, M.D., Alford, J.M., Kennel, S.J., Mirzadeh, S.: $^{212}Pb@C_{60}$ and its water-soluble derivatives: synthesis, stability, and suitability for radioimmunotherapy. J. Am. Chem. Soc. **129**, 5131–5138 (2007)
18. Shultz, M.D., Duchamp, J.C., Wilson, J.D., Shu, C.Y., Ge, J., Zhang, J., Gibson, H.W., Fillmore, H.L., Hirsch, J.I., Dorn, H.C., Fatouros, P.P.: Encapsulation of a radiolabeled cluster inside a fullerene cage, $^{177}Lu_xLu_{(3-x)}N@C_{80}$: an interleukin-13-conjugated radiolabeled metallofullerene platform. J. Am. Chem. Soc. **132**, 4980–4981 (2010)
19. Liang, X.J., Chen, C., Zhao, Y., Jia, L., Wang, P.C.: Biopharmaceutics and therapeutic potential of engineered nanomaterials. Curr. Drug Metab. **9**, 697–709 (2008)
20. Liang, X.J., Meng, H., Wang, Y., He, H., Meng, J., Lu, J., Wang, P.C., Zhao, Y., Gao, X., Sun, B., Chen, C., Xing, G., Shen, D., Gottesman, M.M., Wu, Y., Yin, J.J., Jia, L.: Metallofullerene nanoparticles circumvent tumor resistance to cisplatin by reactivating endocytosis. Proc. Natl. Acad. Sci. U. S. A. **107**, 7449–7454 (2010)

Functionalization

▶ Nanostructures for Surface Functionalization and Surface Properties

Functionalization of Carbon Nanotubes

Gerard Tobias[1], Ernest Mendoza[2] and Belén Ballesteros[3]
[1]Institut de Ciència de Materials de Barcelona (ICMAB-CSIC), Bellaterra, Barcelona, Spain
[2]Centre de Recerca en Nanoenginyeria, Universitat Politècnica de Catalunya, Barcelona, Spain
[3]ICN2 – Institut Català de Nanociència i Nanotecnologia, Bellaterra, Barcelona, Spain

Synonyms

Chemistry of carbon nanotubes; Decoration of carbon nanotubes; Filling of carbon nanotubes; Modification of carbon nanotubes

Definitions

The functionalization of carbon nanotubes consists of the attachment of organic or inorganic moieties to their tubular structure. Through the functionalization of carbon nanotubes, it is possible to modulate their physicochemical properties, increasing their ease of dispersion, manipulation, and processability, among others. The different approaches available for the modification of carbon nanotubes can be classified into four main groups: covalent functionalization, noncovalent functionalization, external decoration with inorganic materials, and endohedral filling.

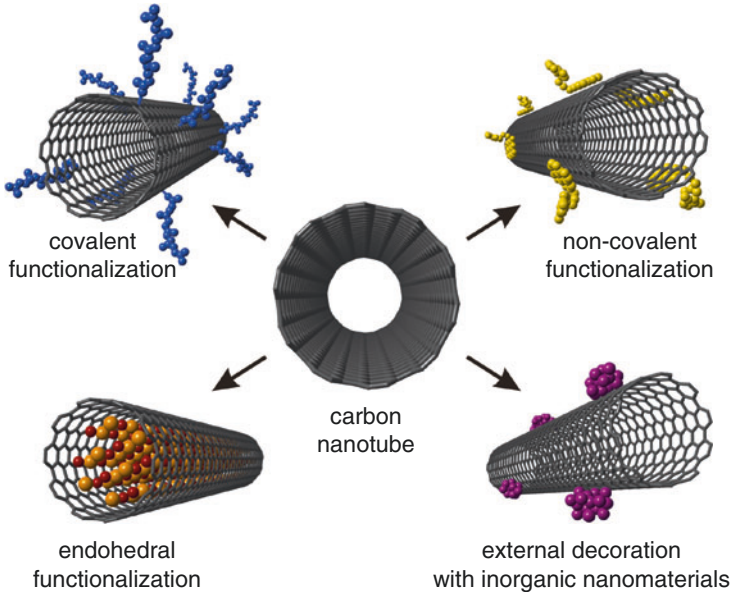
Introduction

Carbon nanotubes (CNTs) are quasi-1D structures formed by one (single-walled CNTs, SWCNTs) or more (multiwalled CNTs, MWCNTs) graphene layers rolled up forming a cylinder. The structure of a graphene layer consists of sp^2 hybridized carbon atoms in a hexagonal network with in-plane σ bonds and out of plane π-orbitals. This particular structure is responsible for CNTs' and graphene's unique properties such as their high Young's modulus, high thermal conductivity, and unique electronic properties. The length of CNTs is typically in the size of micrometers with diameters that range from about 1 to 100 nm. Defect-free CNTs are almost chemically inert. Being an advantage for some applications, this chemical inertness together with their intrinsic poor dispersibility in organic solvents and aqueous media represents a major challenge toward their practical application. Furthermore, as-produced SWCNTs and few layered MWCNTs align parallel to each other forming bundles due to van der Waals interactions, which further aggregate into entangled networks. In this context, the functionalization of CNTs is a key step toward increasing their dispersibility, reactivity, processability, and even biocompatibility [1]. The functionalization of CNTs allows the modification of the structural framework and the creation of supramolecular complexes. The main strategies employed for the modification of CNTs can be classified into four groups: covalent functionalization, noncovalent functionalization, external decoration with inorganic materials, and endohedral filling. Some of these approaches have even been applied simultaneously on a given sample of CNTs, resulting in CNTs bearing multiple functionalities. A schematic representation of these four functionalization strategies is presented in Fig. 1.

Covalent Functionalization

Covalent functionalization of CNTs has been widely used to increase their ease of dispersion, purification, conjugation with biomolecules, and anchoring to polymeric composites and to tailor their properties for specific applications [2]. A major advantage of this route is that a covalent bond is much more stable than other methods such as adsorption on the surface. Generally speaking, there are two strategies for the covalent functionalization of CNTs: direct sidewall functionalization and defect group functionalization.

On the one hand, direct sidewall functionalization is based on the rehybridization of a sp^2 carbon atom into a sp^3 configuration to form a covalent bond between the attacking species and a carbon atom of the CNT scaffold. On the other hand, defect group functionalization is based on anchoring the desired functionalities through appropriate chemical groups introduced in preexisting or intentionally created defects on the CNT scaffold. Importantly, both methods create a disturbance of the CNT tubular structure, the main difference is that defect functionalization will be primarily located (at low defect density creation) at the CNT tips where oxidation is favored compared to the sidewall, as opposed to direct sidewall functionalization where the functionalities will be installed on the CNT sp^2 network, i.e., along the outer wall. Therefore, the major drawback of both methods is the disruption of the CNT sp^2 aromatic system, leading to the modification and degradation of the intrinsic properties of the material such as the transport. Noteworthy, the creation of defects in SWCNTs arises more structural problems than in the case of



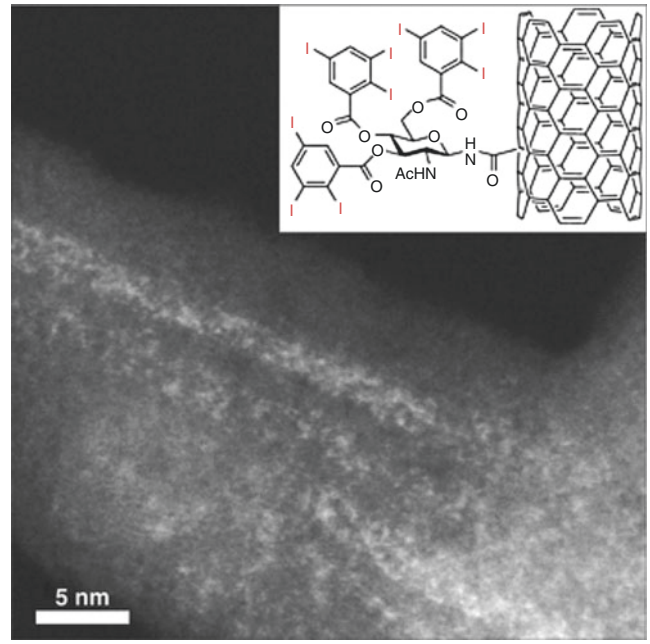
Functionalization of Carbon Nanotubes, Fig. 1 Schematic representation of different strategies for the functionalization of carbon nanotubes

MWCNTs where the inner layers sustain the mechanical integrity of the CNTs. Both strategies have been used to integrate CNTs in diverse applications such as tuning the electronic properties through the coordination with active moieties, their introduction to polymeric composites, or in the biomedical field to fabricate biosensors, drug delivery systems, or scaffolds for tissue engineering.

Defect Group Functionalization

Generally, the defect functionalization route is based on the chemical transformation of defects intrinsically present or generated through chemical oxidation. The generation of surface chemical functionalities on the sp^2 carbon network to anchor specific moieties is a challenging process since a perfect CNT is nearly a 1D cylindrical aromatic macromolecule and therefore expected to be almost chemically inert. The most common approach for the generation of defects on the CNT structure is oxidation either by liquid-phase (i.e., refluxing in HNO_3), gas-phase (i.e., O_2 plasma), or even electrochemical oxidation. It has been proved that the oxidation process starts at topological defects of the sp^2 network and the CNT tips (which ideally are two hemifullerenes) because these are more reactive toward oxidative chemical attacks than a perfect sp^2 carbon network. Hence, the oxidation process yields oxygen-bearing groups, initially in the intrinsic defects and the CNT tips, but increasing the duration or the energy of the treatment also results in the creation of new defects in the CNT scaffold. The oxidative process results in the creation of carboxylic groups and other oxygen-bearing functionalities. These functional groups created at the CNT surface serve as the starting point for further functionalization [3].

Following this approach, CNTs with carboxylic functions have been used as precursors to introduce other functional groups through esterification or amidation reactions. Generally, the carboxylic groups generated by oxidation of the CNT structure are converted to acyl chloride groups via reaction with thionyl or oxalyl chloride, followed by the addition of an alcohol or amine. An alternative approach goes via carbodiimide coupling chemistry [4]. A variety of compounds that facilitate the covalent attachment are available (cross-linking reagents), being the most commonly used the 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDC) and *N, N′*-dicyclohexylcarbodiimide (DCC). The role of



Functionalization of Carbon Nanotubes, Fig. 2 Z-contrast annular dark-field STEM image of a bundle of SWCNTs covalently functionalized with carbohydrates (Adapted with permission from Ref. [5]. Copyright 2007 American Chemical Society)

EDC/DCC is to transform the carboxylic group into a better leaving group through the formation of an unstable active-ester intermediate. In the final step, the ester is displaced by the amine group of the given moiety, resulting in the formation of an amide bond. For instance, the formation of amide bonds with amine-terminated proteins is of interest in the biomedical field. A Z-contrast annular dark-field scanning transmission electron microscopy (ADF-STEM) image of carbon nanotubes functionalized with carbohydrates is presented in Fig. 2. For ease of detection, carbohydrates have been tagged with iodine atoms through the addition of 2,3,5-triiodobenzoyl moieties. In Z-contrast ADF-STEM, the intensity of an element is ca. proportional to the square of its atomic number. Therefore, iodine atoms appear bright white, whereas carbon appears pale gray. The analysis of functionalized CNTs by means of electron microscopy techniques allows their visual detection down to the atomic scale and valuably complements spectroscopic techniques widely used for the characterization of organic compounds such as Raman, infrared (IR), ultraviolet (UV), and nuclear magnetic resonance (NMR), which provide information of the bulk material.

Direct Sidewall Functionalization

This functionalization strategy is based on the rehybridization of an sp^2 carbon atom of the CNT sidewall to an sp^3 configuration and the formation of a covalent bond with a chosen moiety. The amount of synthetic routes developed following this procedure during the last years is extensive [1, 6]. One of the first sidewall functionalization reactions was the fluorination, with C_2F having the highest degree of fluorination. The fluorinated CNTs serve as starting point in many reactions where fluorine atoms are replaced and derived, i.e., alkyl groups can replace the fluorine atoms using Grignard or organolithium reagents. Following the same idea, CNTs have also been chlorinated, brominated, and iodinated. Cycloaddition reactions are also being widely employed for the direct sidewall functionalization of CNTs, for instance, to render them dispersible and also for the attachment of biomolecules. These include, for instance, the carbene [2 + 1] cycloaddition to pristine CNTs, the 1,3-dipolar cycloaddition reaction, or the so-called Bingel [2 + 1] cyclopropanation. As an example of the potential application of the formation of nanohybrids, radical addition with diazonium salts has allowed the preparation of

donor–acceptor systems by decorating the CNTs with porphyrins. Both electrophilic and nucleophilic additions have also been achieved onto the CNT backbone. For instance, electrophilic addition of alkyl halides results in the formation of alkyl and hydroxyl groups, and as an example of the latter, nucleophilic addition of amine-based nucleophiles leads to amino-functionalized CNTs. Other nonconventional methods for the direct sidewall functionalization would include microwave irradiation, ball milling, ozonolysis, and plasma activation [7].

The attachment of biomolecules onto the CNT skeleton is getting an increased attention since it allows their integration in biological systems. Both defect and direct sidewall functionalization are being employed. It is also worthy to mention the use of covalently functionalized CNTs to act as fillers in polymers in order to enhance the mechanical or electrical properties of the composites. For this application it is fundamental to have a good dispersion of the CNTs in the polymeric matrix as well as an efficient load transfer from the matrix to the CNTs. Both of these requirements have been fulfilled, by both defect group and sidewall functionalization, by grafting the polymeric fibers covalently to CNTs or promoting the polymerization from CNTs covalently functionalized with monomers or initiators. These composites have potential applications where lightweight robust materials are needed, such as aerospace science and for sportive material, and are also being studied for field effect transistor devices, selective detection of gas traces, and photovoltaic devices, among others.

In summary, there has been an extensive effort in developing routes toward the covalent functionalization of CNTs at both the sidewalls and the tips. These findings in combination with the organic chemistry toolbox have lead to tailor-made properties for various applications in many fields in nanotechnology.

Noncovalent Functionalization

Carbon nanotubes' noncovalent functionalization consists of the adsorption or wrapping of various functional molecules on the tubular surface of the CNTs. This approach of functionalization is particularly interesting because it is based on π-π stacking, van der Waals, or charge–transfer interactions, and so it preserves the extended π-network of the tubes. Another characteristic of noncovalent functionalization which might be valuable for certain applications is its reversibility. The removal of the adsorbed functional moieties might be achieved by changing external conditions such as solvent, pH, temperature, light irradiation, and redox conditions [8].

Most of the research on noncovalent functionalization has aimed to achieve the dispersibility of CNTs by the exfoliation of bundles to obtain individual CNTs, which are otherwise very difficult to disperse homogeneously in solution. In the past few years, preparation of highly stable suspensions of individually dispersed nanotubes has been achieved in both aqueous and organic solutions.

Typically, the functionalization involves ultrasonication in solution followed by centrifugation and filtration steps. Ultrasonication produces a high-shear environment that results in the formation of gaps or spaces at the bundle ends. These gaps are unstable; so, in the absence of a dispersing agent, the nanotubes reattach through van der Waals attraction. However, when a dispersing agent is present in the solution, these spaces or gaps at the bundle ends are propagated by the dispersant adsorption, ultimately separating the individual nanotube from the bundle [9].

A wide range of compounds have been used for the noncovalent functionalization of carbon nanotubes, including surfactants, polymers, biomolecules, and polynuclear aromatic compounds, which has been detailed below. Besides, CNTs have also been modified with other compounds such as ionic liquids, dyes, macrocyclic host molecules, and phosphines, among others. Depending on their final application, these noncovalently functionalized CNTs have been used in some cases for further modification with biomolecules, inorganic nanoparticles, organic moieties, and others.

Surfactants

In general, the interaction of carbon nanotubes with surfactants takes place when hydrophilic parts of surfactants interact with the solvent and hydrophobic parts are adsorbed onto the nanotube surface, preventing the CNT aggregation into bundles and ropes and thus dispersing them. Therefore, by choosing the appropriate surfactant hydrophobic and hydrophilic groups, one can tailor the dispersibility of nanotubes to fit specific applications.

Regarding the hydrophilic groups, a range of anionic, cationic, and nonionic surfactants have been used to functionalize nanotubes. In the case of charged surfactants such as sodium dodecyl sulfate (SDS) and dodecyltrimethylammonium bromide (DTAB), negatively and positively charged, respectively, the stabilization of the dispersion is due to electrostatic repulsion between the micelles. On the other hand, for charge-neutral surfactants, such as Triton, the dispersion is achieved due to the large solvation shell created by hydrophilic tails of surfactants covering the nanotube. Therefore, for nonionic surfactants, the ability to suspend nanotubes appears to be mostly related to the size of the hydrophilic group, with heavier hydrophilic groups being capable of dispersing more nanotube material.

Hydrophobic parts of the surfactants interact directly with the nanotube surface. Surfactants with simple alkyl chains such as SDS form hydrophobic interactions with the nanotubes which are nonspecific. Therefore, the surfactant molecules are only loosely assembled around the nanotubes and the formed micelles are relatively unstable. The length and the shape of the alkyl chain are important factors, with longer and more branched alkyl groups providing better dispersions. However, too bulky hydrophobic groups show reduced debundling efficiency, since their penetration into the small spaces between the tubes is hindered. When the hydrophobic parts contain an aromatic or a naphthenic group, an especially strong interaction takes place, since the π-π stacking interactions they establish with the CNTs are more specific and directional. For instance, SDS and sodium dodecylbenzenesulfonate (SDBS) have the same length of the alkyl chain, but the latter has a phenyl ring attached between the alkyl chain and the hydrophilic group, which makes SDBS surfactant more effective for dispersing CNTs than SDS.

Although surfactants are efficient dispersants of CNTs, their use for biomedical applications is limited since they are known to permeabilize plasma membranes and have a toxicity of their own.

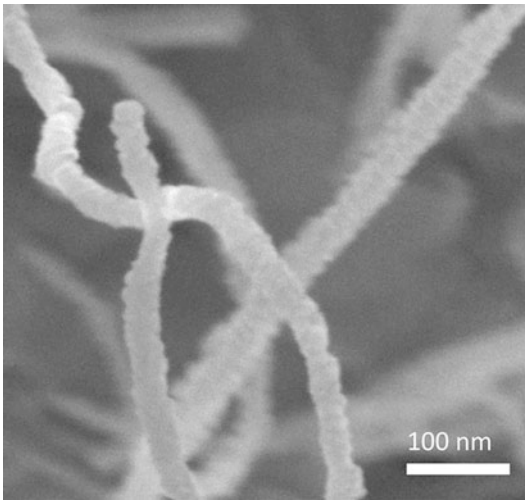
Polymers

Polymers are used for the preparation of CNT dispersions in both water and organic solvents and have also been shown to enable separation of nanotubes from carbonaceous and metal impurities. Both wrapping and loose adsorption of polymers around CNTs have been described. The former is driven by specific interactions which eliminate the hydrophobic interface between the tubes and the solvent.

Steric stabilization has been proposed to account for the stability of highly concentrated dispersions of polymer-coated CNTs. In the appropriate solvent conditions, polymer chains tend to avoid each other to maintain the conformational space available to each of them. Then, polymer-coated CNTs experience a repulsive force which results in a thermodynamically stable nanotube dispersion, together with the exfoliation of ropes into individual tubes.

A problem for polymers is that they might be forced into energetically unfavorable conformations due to the mechanical rigidness of CNTs. To overcome this, some polymers can wrap around nanotubes in a helical or multi-helical way, thus minimizing the conformational strain. This wrapping of the polymer ropes around the tube lattice can occur in a well-ordered periodicity.

Noncovalent polymer functionalization of CNTs has also been extensively studied for the preparation of CNT–polymer composites, since CNTs are considered ideal reinforcing materials due to their exceptional mechanical, thermal, and electrical properties. CNTs have been used to reinforce a wide variety of polymers, including epoxy, acrylates, polyelectrolytes, poly(ethylene glycol), polycarbonates, and conjugated polymers, among others [10]. Figure 3 presents a scanning electron



Functionalization of Carbon Nanotubes, Fig. 3 Scanning electron microscopy image of a sample of MWCNTs covered with polystyrene through noncovalent interactions (Reprinted with permission from Ref. [11]. Copyright 2006, American Institute of Physics)

microscopy image of MWCNTs covered with polystyrene by noncovalent interactions.

Biomolecules

The dispersion of CNT with biological molecules is desired for their integration into living systems and also for the potential use of carbon nanotubes as new types of biosensor materials. Different strategies have been used for the attachment of biomolecules to CNTs, namely direct immobilization and attachment through bifunctional linkers based, for instance, on a pyrene moiety. The biomolecules used for the direct noncovalent functionalization of CNTs include simple saccharides and polysaccharides, proteins, oligopeptides, phospholipids, and nucleic acids [6].

Proteins have a high affinity with the CNTs sidewalls. They are natural polyampholytes containing both hydrophobic and hydrophilic domains, and their hydrophilicity depends on the amino acid sequence and pH conditions. Sonication-assisted debundling and dispersion of CNT–protein hybrids in aqueous and/or polar media have been widely studied using high-molecular-weight proteins, such as streptavidin. Amphiphilic oligopeptides are also efficient dispersants of CNTs. The presence of amino acids with aromatic groups into the peptide sequence, like tryptophan, phenylalanine, or tyrosine, increases the degree of dispersion of the CNTs due to the π-π stacking interactions between the aromatic moieties of a peptide and the CNT graphitic surface. On the other hand, nucleic acids are ideal candidates to form supramolecular complexes based on π-stacking between the aromatic bases and the CNT surface. For instance, DNA strands form strong interactions with CNT by self-assembly processes, rendering stable dispersions using simple sonication procedures.

F

Polynuclear Aromatic Compounds

Condensed aromatic derivatives carrying a hydrophilic or hydrophobic moiety can disperse CNTs in aqueous and organic media, respectively. Specific π-π interactions between the aromatic system and the CNT graphitic structure lead to a strong adsorption of the polynuclear aromatic compounds onto the CNTs. The strong stacking interactions between polynuclear species and nanotubes have been used for the controlled placement of the CNTs onto various surfaces and for the subsequent attachment of biomolecules or inorganic nanoparticles.

Pyrene is a molecule with a large aromatic system that forms very strong π-π interactions with nanotubes. Pyrene derivatives have been extensively used for the functionalization of CNTs in both water and organic solvents, resulting in highly stable dispersions, which, in some cases, have been subsequently used for the attachment of biomolecules. Substituted anthracenes are also useful dispersants, although their adsorption is not as strong as for pyrenes. Heterocyclic polyaromatic molecules such as porphyrins and phthalocyanines also have strong interactions with CNTs through π-stacking. However, these interactions can be weakened if a metal ion is incorporated into the porphyrin or phthalocyanine system, as the electronic state of the molecule is changed. Noncovalent attachment of porphyrins to CNTs individualized in micelles and their subsequent functionalization have shown to result in highly stable assemblies while preserving the electronic properties of the CNTs [12].

External Decoration of Carbon Nanotubes with Inorganic Nanomaterials

The interest of decorating CNTs with inorganic compounds is twofold. On the one hand, it is becoming a powerful technique for the labeling and detection of defects and functionalities on the external sidewalls. A careful study of the CNTs labeled with inorganic elements allows the determination of the exact location, distribution, and density of defects or functionalities on the CNT backbone down to the atomic scale. The most commonly used techniques for the characterization of these labeled CNTs include atomic force microscopy (AFM), scanning tunneling microscopy (STM), high-resolution transmission electron microscopy (HRTEM), and scanning transmission electron microscopy (STEM). On the other hand, and this is the area that it is actually attracting most interest, the formation of hybrid materials CNT–inorganic often results in improved performance, for a given application, with respect to the properties of each of the individual components. CNTs are excellent materials for the formation of nanocomposites due to their high surface area, mechanical strength, thermal conductivity, and chemical inertness. In general, these nanocomposites consist of inorganic nanoparticles (NPs) decorating the external CNT sidewalls either as bulk powder or vertically aligned CNTs [13]. Few examples also report on the complete coverage of the CNT by the inorganic material, thus forming coaxial carbon-inorganic nanotubes [14]. These CNT–inorganic nanocomposite materials have potential application in a wide variety of areas, including catalysis, functional components in supercapacitors and batteries, sensors, optoelectronic devices, and hydrogen storage.

Different strategies have been employed for the formation of such nanohybrids which can be classified in two main groups, depending on the stage in which the NPs are formed: (a) pre-synthesized NPs can be attached to the CNTs in the so-called ex situ approach, and (b) the NPs can be formed directly on the CNT surface in the so-called in situ approach. A key point for the further development and application of these hybrid systems resides in the development of synthetic methodologies that allow a scaling up of the production process while keeping their properties and performance reproducible.

Ex Situ Approach: Decoration of CNTs with Pre-synthesized NPs

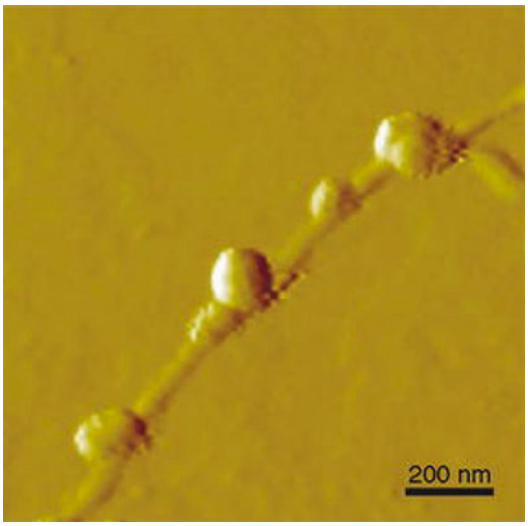
This approach goes through several steps, namely, the preparation of the NPs, chemical modification of the CNTs and/or the preformed NPs, and finally their linkage. The formation of the NPs is carried out independently, without the presence of CNTs, and thus, their size and shape can be more easily tailored. The morphology of the resulting NPs will in turn determine their physical properties. These pre-synthesized NPs can be attached to the CNT walls by either covalent bonds or through noncovalent interactions.

In the case of covalent bonding, through organic linkers or biomolecules, the amount of NPs that can be connected to the CNT walls is controlled by the number of functionalities available on the surface. These are well-defined chemical bonds that allow the formation of stable and robust hybrids. A wide range of already reported functionalization strategies and readily available linkers can be employed to connect CNTs with NPs. As detailed above, carboxylic acid groups on the sidewalls of the CNTs are being extensively used as anchor groups for further modification of the CNT structures. Therefore, a major coupling reaction used for the attachment of pre-synthesized NPs onto the CNTs goes through the esterification or amidation of previously oxidized CNTs employing an organic linker, such as 2-aminoethanethiol or ethylenediamine. Other reactions, such as the Bingel reaction, have also been employed for the attachment of NPs onto CNTs. The linkage of NPs using biomolecules is getting an increased attention due to their high specific recognition and biocompatibility, which are relevant for the application of these hybrid materials in the biomedical field. DNA, proteins, and viruses have been used to assist the formation of hybrid structures. Inorganic NPs that have been covalently attached onto the CNTs include Au, CdSe, CdTe, PbSe, TiO_2, MnO_2, SiO_2, and MgO.

As already explained, sidewall covalent functionalization of CNTs might disrupt the structural periodicity which can cause, for instance, the lost of the intrinsic optoelectronic CNT properties. An alternative method for the formation of hybrid materials, which preserves the intrinsic CNT properties, consists of taking advantage of noncovalent interactions between the CNTs and the NPs. Although the resulting hybrid materials are not as robust as with the covalent approach, its beauty resides in the experimental simplicity. A wide variety of materials, including quantum dots (e.g., CdTe, CdSe, CdS), metals (e.g., Pt, Au, Co), and metal oxides (e.g., TiO_2, SiO_2, Al_2O_3, ZrO_2, Fe_3O_4), have been attached by noncovalent interactions. For instance, these employ polymer wrapping, including polyelectrolytes, resulting in either positively or negatively charged CNTs with improved dispersibility which can then be decorated with NPs of the opposite charge by electrostatic interactions. Absorbance of organic molecules onto the CNT surface can also enhance the interaction of CNTs with NPs either through hydrophobic interactions or $\pi - \pi$ stacking. Molecules containing conjugated π-aromatic structures such as derivatives of pyrene, porphyrins, phthalocyanines, benzyl alcohol, and triphenylphosphines have been employed for the latter.

In Situ Approach: Direct Formation of NPs onto the CNT Surface

In the in situ approach, the CNTs serve as template and stabilizer for the formation of the NPs. It is a straightforward and effective method for the preparation of these nanocomposites, allowing a large degree of NP coverage. It also avoids the presence of an organic protecting shell, commonly used in the ex situ approach for the stabilization of the NPs and their linkage to the CNTs, which might impair their performance. On the other hand, the grown NPs are not so strongly bonded to the CNTs and it is not so easy to control their shape and dimensionality which depend on the influence of the CNTs in the nucleation and growth process. A major advantage is that the inorganic compounds can be deposited to different extents going from discrete NPs to a continuous layer, resulting in the complete coverage of the CNTs with inorganic nanotubes, which in turn can either have a crystalline or amorphous structure. Several techniques have been employed for the deposition of NPs onto the CNT surfaces, including electrodeposition, sol–gel processing, hydrothermal treatment, and gas-phase deposition. The types of inorganic materials that have been grown onto the CNTs include metals (e.g., Ag, Pt, Pd, Au), metal oxides (e.g., TiO_2, Fe_2O_3, ZnO, NiO, CeO_2, Al_2O_3, SnO_2), and metal chalcogenides (e.g., CdSe, CdTe, CdS, ZnS) and few examples on metal carbides (e.g., TiC, TaC, NbC, WC) and metal nitrides (e.g., TiN, Fe_2N, Si_3N_4). Although each material and technique used for the deposition requires certain experimental conditions, a widely used protocol would consist of depositing the inorganic material directly onto previously functionalized CNTs through a chemical or electrochemical redox reaction of the precursors. Metal salts and organometallic compounds are commonly used as precursors for the in situ decoration of CNTs with inorganic materials. An AFM image of a SWCNT decorated with in situ-grown Ag nanoparticles is presented in Fig. 4.

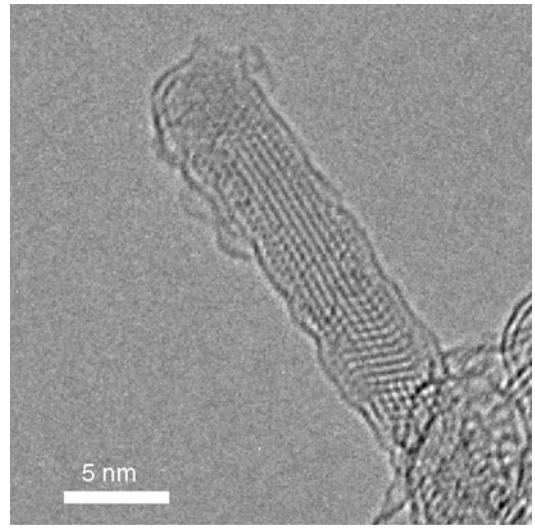


Functionalization of Carbon Nanotubes, Fig. 4 AFM image of a SWCNT decorated with silver nanoparticles [15] (Reproduced with permission from the Royal Society of Chemistry)

Endohedral Functionalization: Encapsulation of Materials Inside Carbon Nanotubes

A completely different approach for the functionalization of carbon nanotubes consists of filling their inner cavities with host materials. The potential applications of filled carbon nanotubes range from nanoelectronic devices to drug delivery systems going, for instance, through functional ceramics. A bandgap modulation of carbon nanotubes has been observed upon encapsulation of different materials, thus making filled CNTs good candidates for molecular electronics. Unprecedented structures and properties compared to those of the same material in the bulk form have been reported for the confined compounds, opening up new opportunities for the development of novel functional nanocomposites. An area in which the containment of materials into CNTs is getting an increased attention is in the biomedical field since their inner cavity can be filled with therapeutic or imaging cargos, while the outer surface can be modified to improve their dispersibility, biocompatibility, and site-selectivity [16].

Back in 1992, Pederson and Broughton predicted on the basis of computer simulations that open-ended nanotubes should act as "nanostraws" and draw in molecules from vapor or fluid phases. In the following years, several works reported on the encapsulation of materials inside CNTs [17]. The inner cavities of the CNTs can be filled either during their synthesis or in a post-synthesis filling process. Since a variety of metals catalyze the synthesis of CNTs, it is possible to tune the conditions of the synthesis to encapsulate mainly metals and metal carbides inside CNTs. A much wider variety of materials has been successfully filled into CNTs in a post-synthesis process, including water, fullerenes and their derivatives, organic molecules and biomolecules, metals and alloys, metal oxides, metal halides, and metal carbides. Once encapsulated the encaged material typically forms nanoclusters or nanowires within the inner cavities of the CNTs. Recently, the formation of single-layered inorganic nanotubes in the interior of CNTs has been reported, further expanding the possibilities of these systems [18].



Functionalization of Carbon Nanotubes, Fig. 5 HRTEM image of a CNT filled with CuBr (Reproduced with permission from Nature publishing group [20])

Several methodologies allow the encapsulation of materials into CNTs which can be generally classified in the so-called gas route and liquid route [19]. The gas route is a straightforward method for filling CNTs. It consists of heating a mixture of open-ended CNTs and the material to be encapsulated above the sublimation temperature of the chosen material. This process takes place in a sealed ampoule under vacuum. Using the same experimental protocol, if the material to be encapsulated is heated up above its melting temperature, the CNTs are then filled by capillary wetting via the liquid route. A HRTEM image of a CNT filled with a metal halide using this approach is presented in Fig. 5. Another methodology also within the liquid route consists of stirring open-ended CNTs in a solution containing the chosen material, either dissolved or suspended. The solution enters into the inner cavities drawing the desired material into the CNTs. In a subsequent step, the solvent is dried leaving behind filled CNTs.

Cross-References

- ▶ Active Carbon Nanotube-Polymer Composites
- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae
- ▶ In Vivo Toxicity of Carbon Nanotubes
- ▶ Synthesis of Carbon Nanotubes
- ▶ Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering
- ▶ Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

References

1. Battigelli, A., Ménard-Moyon, C., Da Ros, T., Prato, M., Bianco, A.: Endowing carbon nanotubes with biological and biomedical properties by chemical modifications. Adv. Drug Deliv. Rev. **65**, 1899–1920 (2013)
2. Bekyarova, E., Sarkar, S., Wang, F., Itkis, M.E., Kalinina, I., Tian, X., Haddon, R.C.: Effect of covalent chemistry on the electronic structure and properties of carbon nanotubes and graphene. Acc. Chem. Res. **46**, 65–76 (2013)
3. Gonzalez, A.B.., Mendoza, E., Pellicer, E., Alsina, F., Fernandez-Sanchez, C., Lechuga, L.M.: Discriminating the carboxylic groups from the total acidic sites in oxidized multi-wall carbon nanotube by means of acid–base titration. Chem. Phys. Lett. **462**, 256–259 (2008)
4. Wong, L.S., Khan, F., Mickfield, J.: Selective covalent protein immobilization: strategies and application. Chem. Rev. **109**, 4025–4053 (2009)
5. Hong, S.Y., Tobias, G., Ballesteros, B., El, O.F., Errey, J.C., Doores, K.J., Kirkland, A.I., Nellist, P.D., Green, M.L.H., Davis, B.G.: Atomic-scale detection of organic molecules coupled to single-walled carbon nanotubes. J. Am. Chem. Soc. **129**, 10966–10967 (2007)
6. Karousis, N., Tagmatarchis, N., Tasis, D.: Current progress on the chemical modification of carbon nanotubes. Chem. Rev. **110**, 5366–5397 (2010)
7. Vazquez, E., Giacalone, F., Prato, M.: Non-conventional methods and media for the activation and manipulation of carbon nanoforms. Chem. Soc. Rev. **43**, 58–69 (2014)
8. Di Crescenzo, A., Ettorre, V., Fontana, A.: Non-covalent and reversible functionalization of carbon nanotubes. Beilstein J. Nanotech. **5**, 1675–1690 (2014)
9. Strano, M.S., Moore, V.C., Miller, M.K., Allen, M.J., Haroz, E.H., Kittrell, C., Hauge, R.H., Smalley, R.E.: The role of surfactant adsorption during ultrasonication in the dispersion of single-walled carbon nanotubes. J. Nanosci. Nanotechnol. **3**, 81–86 (2003)
10. Bilalis, P., Katsigiannopoulos, D., Avgeropoulos, A., Sakellariou, G.: Non-covalent functionalization of carbon nanotubes with polymers. RSC Adv. **4**, 2911–2934 (2014)
11. Watts, P.C.P., Lyth, S.M., Mendoza, E., Silva, S.R.P.: Polymer supported carbon nanotube arrays for field emission and sensor devices. Appl. Phys. Lett. **89**, 103113 (2006)
12. Clavé, G., Delport, G., Roquelet, C., Lauret, J.-S., Deleporte, E., Vialla, F., Langlois, B., Parret, R., Voisin, C., Roussignol, P., Jousselme, B., Gloter, A., Stephan, O., Filoramo, A., Derycke, V., Campidelli, S.: Functionalization of carbon nanotubes through polymerization in micelles: a bridge between the covalent and noncovalent methods. Chem. Mater. **25**, 2700–2707 (2013)
13. Van Hooijdonk, E., Bittencourt, C., Snyders, R., Colomer, J.-F.: Functionalization of vertically aligned carbon nanotubes. Beilstein J. Nanotech. **4**, 129–152 (2013)
14. Eder, D.: Carbon nanotube-inorganic hybrids. Chem. Rev. **110**, 1348–1385 (2010)
15. Azamian, B.R., Coleman, K.S., Davis, J.J., Hanson, N., Green, M.L.H.: Directly observed covalent coupling of quantum dots to single-wall carbon nanotubes. Chem. Commun. 366–367 (2002)
16. Martincic, M., Tobias, G.: Filled carbon nanotubes in biomedical imaging and drug delivery. Exp. Op. Drug Deliv.12, 563–581 (2015)
17. Sloan, J., Kirkland, A. I., Hutchison, J. L.,Green, M. L. H.: Integral atomic layer architectures of 1D crystals inserted into single walled carbon nanotubes. Chem. Commun. 1319–1332 (2002)
18. Cabana, L., Ballesteros, B., Batista, E., Magén, C., Arenal, R., Oró-Solé, J., Rurali, R., Tobias, G.: Synthesis of PbI_2 single-layered inorganic nanotubes encapsulated within carbon nanotubes. Adv. Mater. **26**, 2016–2021 (2014)
19. Monthioux, M., Flahaut, E.: Meta- and hybrid-CNTs: a clue for the future development of carbon nanotubes. Mater. Sci. Eng. C **27**, 1096–1101 (2007)
20. Hong, S.Y., Tobias, G., Al-Jamal, K.T., Ballesteros, B., Ali-Boucetta, H., Lozano-Perez, S., Nellist, P.D., Sim, R.B., Finucane, C., Mather, S.J., Green, M.L.H., Kostarelos, K., Davis, B.G.: Filled and glycosylated carbon nanotubes for in vivo radioemitter localization and imaging. Nat. Mater. **9**, 485–490 (2010)

Functionalized Nanomaterials

- ▶ Nanostructured Functionalized Surfaces

Fundamental Properties of Zinc Oxide Nanowires

Jia Grace Lu
Departments of Physics and Electrophysics, University of Southern California, Los Angeles, CA, USA

Synonyms

Characterizations of zinc oxide nanowires for nanoelectronic applications

Definition

Charge transport under different temperature and optical excitations in semiconducting nanowires is strongly influenced by the size confinement, surface effect, and impurity states.

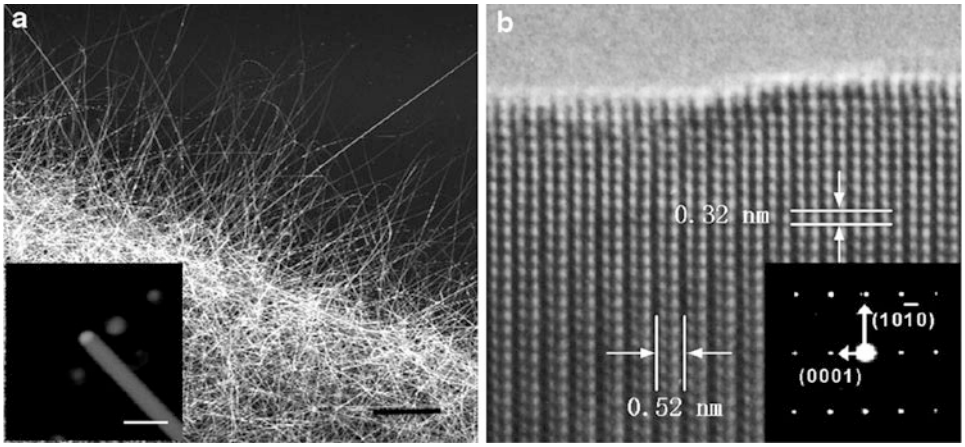
Overview

When the physical size approaches fundamental length scales, such as Bohr radius or localization length, the material properties alter from their traditional bulk properties and exhibit unique characteristics. In the case of quasi-1D (Q1D) nanowire structures with high aspect ratio, there is strong fundamental interest in elucidating the charge conduction mechanisms, e.g., the effects of size confinement, surface contribution, and impurity states. This basic understanding helps to evaluate their potential as the building blocks for future nanoscale electronics. This entry presents the basic electrical transport properties correlated with the optical measurements of metal oxide nanowires, focusing on ZnO material.

Synthesis of Single Crystalline ZnO Nanowires

ZnO has a stable wurtzite crystal structure. The zinc atoms are tetrahedrally coordinated with four oxygen atoms, which can be considered as two interpenetrating hexagonal closely packed lattices of zinc and oxygen. Single crystalline ZnO nanowires are typically synthesized via catalytic chemical vapor deposition (CVD) technique that is governed by a vapor–liquid–solid mechanism. In this process, various nanoparticles or nanoclusters are used as catalysts, such as Au, Cu, Co, Sn, etc. The formation of eutectic alloy droplet occurs at each catalyst site, followed by the nucleation and growth of solid nanowire upon the supersaturation of the liquid droplet. Incremental growth of the nanowire taking place at the droplet interface pushes the catalyst upward. The growth direction is along the crystal plane that is the most stable and has the lowest surface free energy. Figure 1a shows the scanning electron micrographs of ZnO nanowires synthesized by thermal evaporation of pure Zn powder at 700 °C in low oxygen concentration (<2 %) using Au catalyst nanoparticles [1]. Figure 1b is a high-resolution TEM image (HRTEM), and the selected area electron diffraction (SAED) pattern shown in the inset denotes single crystalline nature with [0001] growth direction.

Impurity doping is often achieved by adopting a laser ablation method in the CVD system. It can not only control the thermodynamic balance, thus influencing the crystal formation, but also tune the electron transport properties. The incorporations of Mg, Ga, In, Sn, Ni, Al, Mn, and Co in ZnO nanowires have been successfully demonstrated. One doping method places pure zinc source powder (99.99 %) in the center of the horizontal CVD furnace, while the dopant metal is loaded at the upstream of the quartz tube. The dopant source is constantly ablated by Nd:YAG (yttrium aluminum garnet) laser pulses. The doped nanowires grow via the VLS process on catalyst-coated silicon substrates. The energy dispersive X-ray (EDX) spectrum and HRTEM analyses are used to show that the impurity atoms are substitutionally doped into the wurtzite crystal lattice and the nanowires maintain the same growth direction along the *c*-axis ([0001] direction).



Fundamental Properties of Zinc Oxide Nanowires, Fig. 1 (**a**) A SEM image shows the as-grown ZnO nanowires. Inset: a ZnO nanowire with a diameter of 50 nm terminated with a Au nanoparticle (scale bar, 100 nm). (**b**) HRTEM image of a ZnO nanowire with lattice spacing 0.52 nm. The *inset* SAED indicates that the growth direction is [0001] along the *c*-axis

Material Properties of Undoped ZnO Nanowires

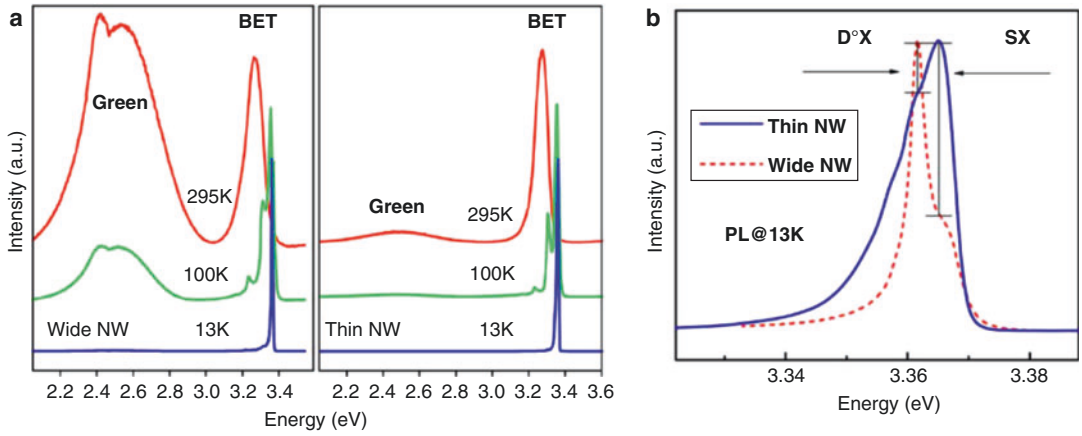
ZnO has a direct wide bandgap of 3.37 eV at room temperature. Compared with other wide bandgap materials, it has a larger exciton binding energy (60 meV) that ensures efficient exciton emissions at room temperature. In addition, due to the small size comparable to its Debye length, ZnO nanowire has demonstrated superior sensitivity to surface chemical adsorbents. As a result, ZnO nanostructures have attracted intensive research interest in developing nanoscale transparent electronics, sensors, actuators, optoelectronics, spintronics, and energy-harvesting devices.

ZnO is known to be natively *n*-type doped accompanied with slight self-compensation. The electron donors originate from its native defects, such as oxygen vacancies, zinc interstitials, or incidental hydrogen [2]. They form impurity bands with shallow donor levels around 30–60 meV [3–5] and deep levels around 0.7–2.3 eV [6, 7] below the conduction band edge E_c. The impurity states have been probed by various spectroscopy techniques such as photoluminescence, cathodoluminescence, and electroluminescence.

Photoluminescence Property

Low-temperature photoluminescence (PL) measurement is one of the most sensitive tools to investigate radiative defects in materials. From the spectra peaks and decay time measurements, one can determine the different emission mechanisms, the energy bandgap, defect levels, and their relative density.

As feature sizes shrink into the nanometer scale, optical properties of metal oxide nanowires begin to reveal size effects. Comparing the ratio of the deep level to near-band-edge emission of different sized ZnO nanowires, it is found that the ratio is directly related to the wire diameter to length ratio [8, 9]. The temperature-dependent PL spectra of ZnO nanowires with different diameters (d) are shown in Fig. 2a. The green luminescence band at ~2.5 eV, originated from bulk defects, is much stronger for wide nanowire ($d >$ 100 nm) than the thinner wires ($d <$ 30 nm). Figure 2b shows the zoomed-in view of the band-edge transition (BET) emission at 13 K for different diameters. The emission from wide nanowires shows a sharp peak at 3.362 eV arising from the donor-bound excitons and a minor peak at 3.366 eV as a result of surface-bound exciton. The emission peak at 3.366 eV prevails in the



Fundamental Properties of Zinc Oxide Nanowires, Fig. 2 (**a**) PL spectra of wide (*left*) and thin (*right*) ZnO nanowires at different temperatures show that the bulk defect states (*green*) are more pronounced in wide nanowires. Intensity is normalized with respect to the band-edge transition peak. (**b**) A close-up view of band-edge transition (BET) of thin and wide nanowires at 13 K, indicating the blue shift in thinner wires due to the predominant surface-bound exciton emission

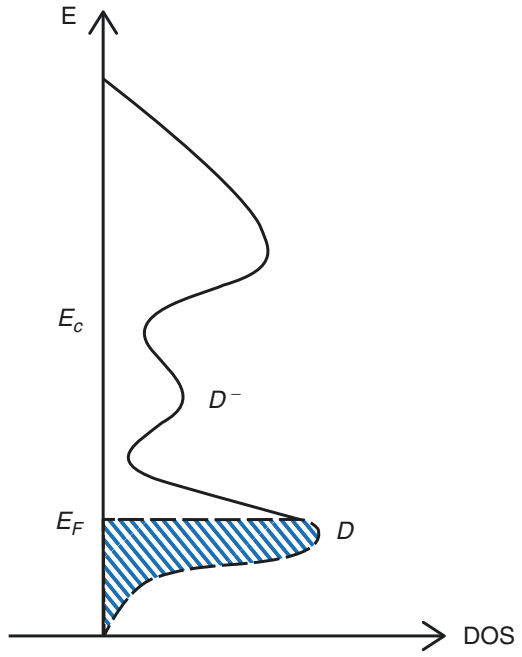
thinner wires. This blue shift is a direct result of finite size effect, due to the emerging of the surface-bound exciton emission that becomes dominant in smaller diameter nanowires.

Electrical Properties

Charge transport properties in the Q1D system are of profound interest due to the unique phenomena that occur when size reduces to nanometer scale and are also of technological importance for utilizing them as the building blocks for potential electronic devices. The contribution of surface-related properties will become increasingly important due to the enhanced surface-to-volume ratio that is inversely proportional to the nanowire radius. The optical and transport properties are closely related to the inherent native defects for the undoped nanowires and extrinsic impurities for the doped nanowires.

The underlying electronic band structure giving rise to conduction mechanisms that originate from the native *n*-type doping accompanied with self-compensation. In the case of moderate to heavy doping, the Coulomb fields arising from the ionized donors and the compensating acceptors result in the dispersion of the donor levels into impurity bands [10]. Moreover, the impurity band further splits into a lower *D* band and an upper *D* band, as depicted in the Fig. 3 [11–13]. The lower *D* band is attributed to singly charged donors, while the upper *D* band arises from neutral donors. The Fermi level, E_F, situates within the lower D band.

In order to study the electrical transport properties of the nanowires, metallic electrodes are patterned onto individual nanowires. The interface between nanowire and metal contact plays a critical role in transport behavior and device performance. Low resistance ohmic contact is necessary to elucidate the intrinsic conduction property. To reduce the contact energy barrier, titanium (Ti) with a work function of ~4.33 eV, which is close to the ZnO electron affinity of ~4.35 eV, is often selected as the adhesive layer to form ohmic contact with native *n*-type ZnO [14, 15]. The intrinsic electrical properties are determined from four-probe conductivity measurement and three-terminal field-effect transistor (FET) response [1]. Figure 4a represents the logarithmic scaled conductivity of a ZnO nanowire with diameter around 50 nm as a function of reciprocal temperature. The charge conduction consists of a combination of thermal activation and variable range hopping (VRH) conduction channels. For $T > 100$ K, the conduction shows Arrhenius behavior with thermally activated transport.



Fundamental Properties of Zinc Oxide Nanowires, Fig. 3 Band diagram depicting the density of states (DOS) in ZnO. The electronic levels below E_F (*shaded region*) are occupied, while the levels above are unoccupied

The conductivity can be expressed as $\sigma \sim \exp(-E_a/k\ T)$, where E_a is the activation energy (around 30 meV), attributed to the shallow donor levels below the conduction band edge. For $T <$ 20 K, Mott's variable range hopping model governs the transport, with conductivity modeled by $\sigma \sim A\exp(-T_0/T)^{1/(d+1)}$, where T_0 represents the characteristic temperature that depends on the localization length and the density of states at the Fermi level and d is dimensionality. The best fit yields $d = 3$, manifesting 3D conduction. For intermediate temperature range, 20 K $< T <$ 100 K, the charge transport is contributed by the activation of electrons from the Fermi level E_F to the upper D band.

Figure 4b shows an atomic force microscope (AFM) image of a ZnO nanowire FET on a SiO_2/p^{++} Si substrate. Based on simple electrostatic model, taking an infinitely long thin wire of length L and radius r ($L >> r$) located at a distance h above a conducting plane, the backgate capacitance is can be estimated to be

$$C = \frac{2\pi\varepsilon_0\varepsilon_r L}{\ln[(2h+r)/r]}$$

where ε_r is the effective dielectric constant for the wire between air and SiO_2 plane [16]. In general, the field-effect mobility (μ) can be expressed using the following expressions:

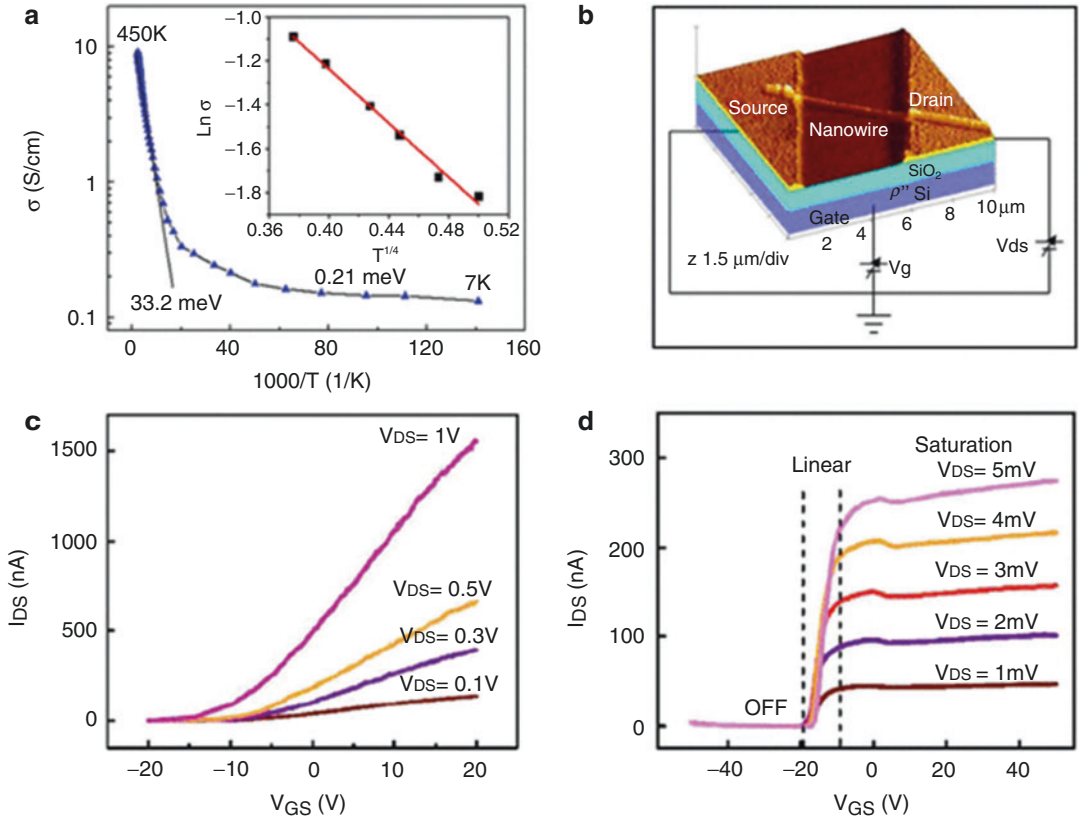
$$\mu = \frac{\partial I_{DS}}{\partial V_g}\frac{L^2}{CV_{DS}}$$

From the gate voltage-dependent source-drain conductance characteristics (Fig. 4c), one estimates the field-effect mobility to be around 30 cm^2/Vs and an electron concentration around $\sim 10^{18}$ cm^{-3} [17]. It is known that the surface defects act as scattering and trapping centers, thus affecting the charge transport. Surface passivation has been applied on metal oxide nanowires. Park et al. have reported an electron mobility of 1,000 cm^2/Vs after coating the nanowires with polyimide passivation layer to reduce the electron scattering and trapping at surface [18]. Later, a CMOS-compatible process is utilized to coat a SiO_2/Si_3N_4 bilayer on ZnO nanowire FETs. The results show dramatically enhanced electron mobility up to 4,000 cm^2/Vs (Fig. 4d). This performance improvement indicates that the ZnO nanowire device holds promising potential in memory and logic gate devices and other high-speed electronics application.

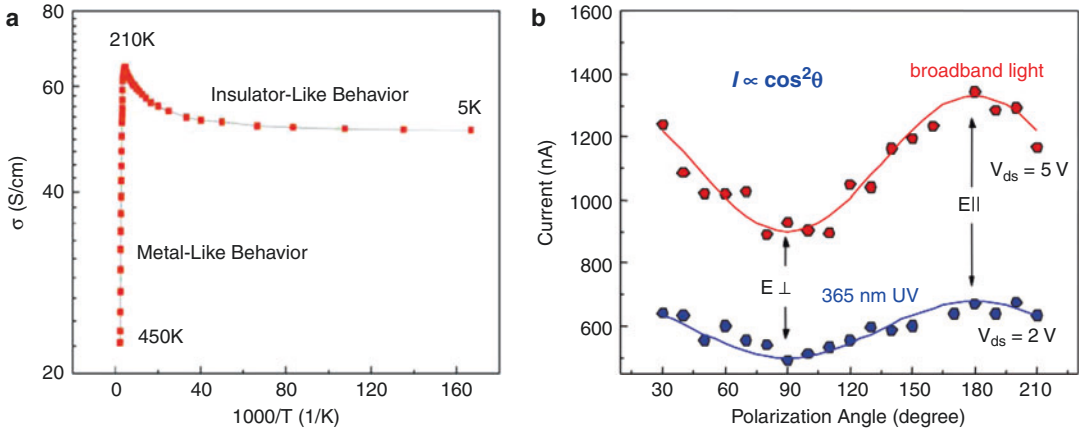
Photoconductance Properties

Under light illumination, the conductivity of semiconducting nanowires typically increases due to excess photogenerated charge carriers. Interestingly, pronounced metal-insulator transitions have been observed in undoped and doped ZnO nanowires. Figure 5a shows the temperature-dependent conductivity of ZnO nanowire upon continuous UV ($\lambda_{UV} = 365$ nm, $E_{UV} = 3.4$ eV) irradiation [19]. The conductivity is significantly increased by two orders of magnitude at low temperature. As temperature rises, thermally activated charge carriers contribute further to an

F



Fundamental Properties of Zinc Oxide Nanowires, Fig. 4 (**a**) Conductivity versus reciprocal temperature. *Inset*: $\ln\sigma$ vs $T^{-1/4}$ plot in the low temperature range showing a linear fit to the 3D VRH model. (**b**) An AFM image of a nanowire placed on a Si backgate under a SiO_2 dielectric layer with source and drain electrodes patterned onto the wire. (**c**) I_{DS}-V_{GS} curves of a ZnO nanowire FET without surface treatments showing typical *n*-type semiconducting behavior. (**d**) I_{DS}-V_{GS} of a surface-treated nanowire FET exhibits significantly increased ON/OFF ratio, diminished subthreshold swing, and enhanced mobility



Fundamental Properties of Zinc Oxide Nanowires, Fig. 5 (**a**) Photoconductivity of ZnO nanowire under UV irradiation reveals a metal-to-insulator transition at 210 K. (**b**) Photocurrent as a function of the light polarization angle for both UV (*blue curve*) and halogen light (*red*)

increase in conductance. Yet, as the carrier concentration approaches a critical value, an insulator-to-metal transition occurs. The strong surface scattering and Coulomb interactions among the dramatically increased charge carriers confined in the Q1D structure lead to a decrease in the conductance as the temperature rises.

In light polarization measurement, it is found that the photoconductance of ZnO nanowires is sensitive to the angle of the incident light as shown in Fig. 5b. The photocurrent is maximized when the electrical field of the incident light is parallel to the nanowire long axis and minimized at 90°. The large polarization response is attributed to the large dielectric contrast between the nanowire and surrounding air. The polarization-dependent behavior can be potentially utilized for optical switch, photodetector, and polarizer using aligned nanowire arrays.

Material Properties of Extrinsic Doped ZnO Nanowires

ZnO nanostructures have been doped with various elements, with the goal to tune their optical, electrical, and magnetic properties. Due to the abundant impurity states, the mechanisms of photoconduction are much more complex than the simple generation of electron–hole pairs across the direct bandgap. Therefore, it is challenging to elucidate the electrical properties of these individual nanowires and the contributions from impurity levels and size confinement. In addition, much more in-depth understanding is needed on the metal-to-semiconductor transitions in heavily doped systems. In this section, a study on Ga-doped ZnO (ZnO:Ga) nanowires is described below.

Electrical Properties

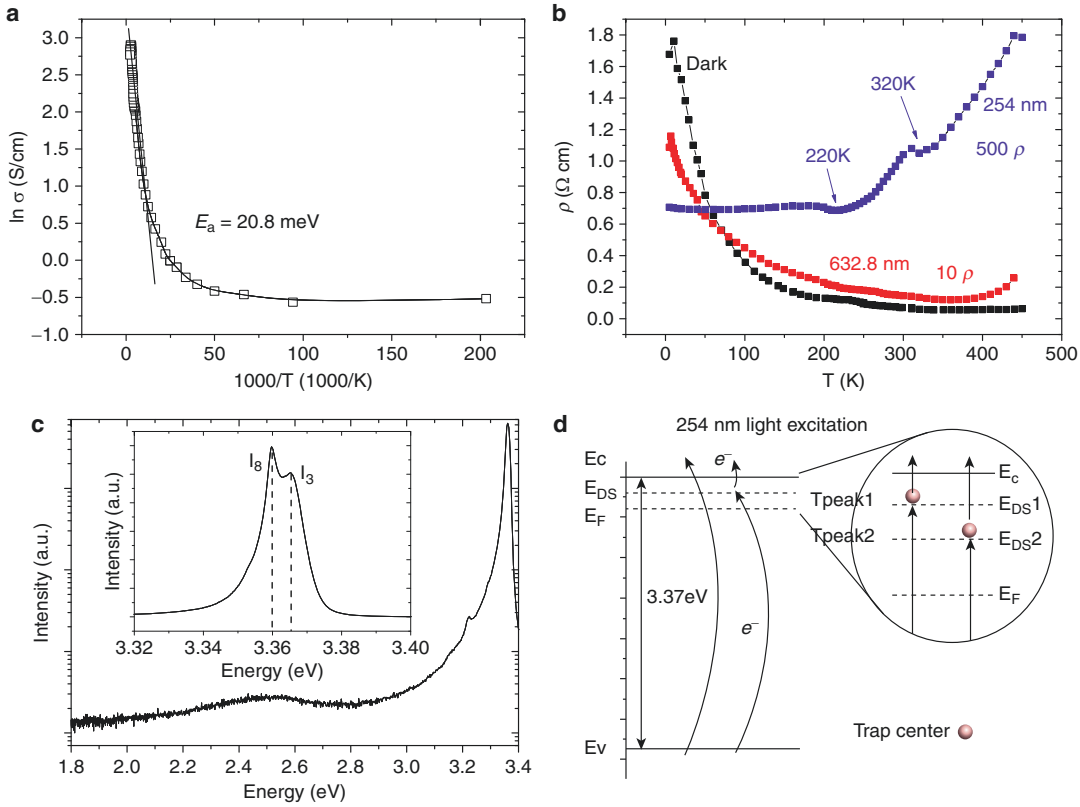
The electrical transport properties are investigated on single doped nanowires (~60 nm in diameter) in a three-terminal configuration, as in a typical field-effect transistor with a global backgate underneath a 500-nm SiO_2 dielectric layer on top of a Si substrate.

Photoconductance and Photoluminescence Properties

Figure 6a is the semilogarithmic plot of conductivity versus reciprocal temperature under dark condition. It is shown in the earlier section that the electrical transport of undoped ZnO nanowires at relatively high temperature range is mainly governed by thermal excitation [13, 19]. Here, the linear region shows such Arrhenius behavior fitted with $\sigma \propto e^{-E_a/kT}$, where E_a, k, and T are the activation energy, Boltzmann constant, and absolute temperature, respectively. As a result, E_a is extracted from the slope to be ~20.8 meV, corresponding to the shallow donor levels arising from the native and extrinsic defects [2, 5].

Figure 6b plots the resistivity of ZnO:Ga nanowire as a function of temperature under dark, He–Ne laser (λ_{ph} = 632.8 nm, E_{ph} = 1.96 eV), and UV (λ_{UV} = 254.0 nm, E_{UV} = 4.88 eV) illuminations. The phenomenon of persistent photoconductivity is typically observed in ZnO channels, originating from the surface band bending due to oxygen adsorption [20, 21]. The resistivity behaviors in dark and in continuous 632.8-nm laser (red curve in Fig. 6b) excitation follow the semiconducting-like behavior, i.e., resistivity decreases with the increase of temperature. In sharp contrast, under continuous UV illumination (violet curve in Fig. 6b), the temperature-dependent resistivity changes distinctively to a metallic-like behavior. The considerable conductance increase is due to both thermal excitation and photogeneration; and the resistivity increase with the increase of temperature manifests the significantly enhanced Coulomb interactions and surface scattering in the quasi-1D channels.

To calculate the charge concentration, the field-effect model shown earlier will not be strictly valid in this case of degenerate carriers. As shown in Ref. [16], the conductivity σ varies with resistance R as $d\sigma/\sigma = -(dR/R) = (2/3)(dN/N)$, where N is the total number of carriers in the wire and changes with gate voltage via $d(eN) = C\, dV_g$. The gate capacitance per unit length is estimated to be 0.42 pF/cm. From the measurements at $T = 4.9$ K and the drain source bias $V_{ds} = 0.2$ V, the carrier concentration n in



Fundamental Properties of Zinc Oxide Nanowires, Fig. 6 (**a**) Semilogarithmic plot of conductivity versus reciprocal temperature under dark condition. (**b**) Temperature-dependent resistivity in dark and under (632.8 and 254.0 nm) light illuminations. Note: for comparison convenience on the same plot, the resistivity data under laser and UV irradiation is multiplied by 10 and 500, respectively. (**c**) PL spectrum of ZnO:Ga nanowires measured at 13 K, showing a weak *green* luminescence band at ~2.5 eV (496 nm). *Inset*: the high-resolution plot centered at the band-edge luminescence represents donor-bound exciton emission. (**d**) Schematic band diagram of electron excitation under 254-nm UV ($E_{ph} = 4.88$ eV) illumination. Zoomed-in diagram shows the native and extrinsic impurity levels, giving rise to the corresponding resistance peaks due to trapping and detrapping processes

dark (for the wire with $L = 10$ μm and $r = 30$ nm) is calculated to be ~8.72×10^{18} cm^{-3}. From measurements on He–Ne laser ($\lambda_{ph} = 632.8$ nm, $E_{ph} = 1.96$ eV) and UV ($\lambda_{UV} = 254.0$ nm, $E_{UV} = 4.88$ eV) illuminations, the carrier concentrations are estimated to be 1.78×10^{19} cm^{-3} and 6.88×10^{19} cm^{-3}, respectively. Under irradiation, excess electrons are excited to the conduction band and form a degenerate electron gas.

In Fig. 6b, two distinct resistivity valleys are clearly displayed at 220 and 320 K, respectively. A close examination shows that the 632-nm laser illumination also induces these resistivity valleys at the same temperatures, although the effect is of much smaller magnitude than of the 254-nm UV irradiation. These two reproducible valleys signify sharp transitions at two electronic states, and they are attributed to the existence of impurity levels below the conduction band edge, leading to charge trapping and detrapping processes.

In order to verify this contribution from the impurity states, photoluminescence measurements using He–Cd laser ($\lambda = 325$ nm) incident excitation are conducted to compare with the photoconductance data. Low-temperature PL spectrum in Fig. 6c displays a weak green luminescence band at around 2.5 eV, owing to the

presence of deep lying defect states such as oxygen vacancies and structural defects [22, 23]. A weak UV emission shoulder at 3.2206 eV ($\lambda = 384$ nm) can be reproducibly observed in the PL data, which is attributed to the donor-acceptor pair emissions in ZnO [24]. An intense luminescence of donor-bound excitons around the band edge remains sharp as in undoped emission PL [23, 25], confirming that the Ga dopants have been well integrated into the ZnO lattice. With a careful examination of the bound exciton emission (inset of Fig. 6c), one sees that the peak splits into two distinct lines, manifesting that the Ga doping induces shallow donor levels [8]. The small peak at 3.3654 eV is caused by ionized donor-bound exciton, typically assigned as the I_3 line, while the prominent emission line at 3.3597 eV can be indexed to the I_8 line, attributed to the exciton emission bound to Ga donors [25, 26].

Comparing the photoconductance and photoluminescence measurements, the conduction mechanisms in ZnO nanowires can be now understood consistently. Both thermal and optical excitations contribute to the conduction in the nanowire. As a general trend under continuous UV irradiation, metallic behavior is revealed. The strong Coulomb interactions among the dramatically enhanced charge carriers confined in the Q1D nanostructure cause the resistance to increase as the temperature rises. The resistance valleys originate from the defect levels. To illustrate the process, Fig. 6d depicts the band schematic diagram and the effect under 254-nm UV illumination (4.88 eV). E_V, E_C, and E_{DS} denote, respectively, the valence band edge, the conduction band edge, and the shallow defect levels which arise from the native defects and external impurity dopants. The existence of E_{DS} levels serves as trapping and detrapping centers. At low temperatures, the photoexcited electrons from valence band get trapped in the emptied localized defects states (E_{DS} levels). As the temperature increases to ~220 K, the electrons occupied in the $E_{DS}1$ state become detrapped and thermally excited to the conduction band, inducing an increase of the electron concentration, giving rise to the first resistivity valley. The second valley occurs as temperature rises further when the trapped electrons are excited from the $E_{DS}2$ level to the conduction band. Considering the consistent valley position at 220 and 320 K, two E_{DS} levels can be estimated to be 19.0 and 27.6 meV below conduction band (Fig. 6d). In addition, the difference between the two defect levels (8.6 meV) falls in the range of the difference between the I_3 and I_8 photoemission peaks (5.7 meV). This elucidates that these resistivity valleys originate from the shallow defect states due to native and Ga donors.

Future Outlook

In the fundamental studies of ZnO nanowires, there is still much awaiting to be explored. Some important questions on the intrinsic conduction mechanisms via impurity states, surface contribution, and quantum effects remain to be answered (e.g., thermal activation, hopping conduction, weak localization, metal-insulator transition, etc.). In addition, when the diameter further reduces, scalability on the critical length scale (50 nm) and dimensional transitions from 3D, 2D, to 1D need to be thoroughly examined. Moreover, single electronic effect is expected to play an important role in low dimensions due to the strong Coulomb interactions. The quantum confinement will provide the means to manipulate individual quantized states of electrons, spins, and phonons. With a comprehensive characterization and quantitative analysis of the unique properties, one advances the knowledge necessary for device design to achieve high performance and low-power consumption based on the nanostructures.

In the trend of device miniaturization, nanowires have attracted immense interests and possess great potentials in device applications. Among various nanomaterials, ZnO nanowire has been one of the most enthralling systems due to its remarkable functionalities. With the achievements in synthesizing and characterizing high-quality nanowires, extensive research effort

has been invested on nanoscale ZnO-based device fabrications. ZnO nanowire FETs, as the elementary building blocks of next-generation transparent electronic circuitry, have been constructed with different gate configurations of back, top, side, and surrounding gate architectures. ZnO nanowire FET can be implemented into either *n*-type depletion or enhanced-mode operations that are both essential in logic circuit construction. Yet, large-scale device integration remains a crucial challenge. The realization of logic circuits based on ZnO nanowires leads toward an important step in nanoelectronics. However, cascading into multilevels of logic gates with high gain and efficient interconnection for hierarchical assembly demands additional investigations. In summary, these fascinating applications encourage further research efforts to address the technical challenges and to unravel the scientific puzzles that are vital for the meaningful development of future technology applications.

Cross-References

- ▶ Chemical Vapor Deposition (CVD)
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanostructured Materials for Sensing
- ▶ Nanotechnology
- ▶ Optical and Electronic Properties

References

1. Chang, P.C., Fan, Z., Chien, C.J., Stichtenoth, D., Ronning, C., Lu, J.G.: High-performance ZnO nanowire field effect transistors. Appl. Phys. Lett. **89**, 133113 (2006)
2. Seghier, D., Gislason, H.P.: Characterization of donor states in ZnO. Phys. B Condens. Matter **401**, 404–407 (2007)
3. Meyer, B.K., Sann, J., Hofmann, D.M., Neumann, C., Zenner, A.: Shallow donors and acceptors in ZnO. Semicond. Sci. Technol. **204**, S62–S66 (2005)
4. Look, D.C., Reynolds, D.C., Sizelove, J.R., Jones, R. L., Litton, C.W., Cantwell, G., Harsch, W.C.: Electrical properties of bulk ZnO. Solid State Commun. **1056**, 399–401 (1998)
5. Look, D.C., Hemsky, J.W., Sizelove, J.R.: Residual native shallow donor in ZnO. Phys. Rev. Lett. **8212**, 2552–2555 (1999)
6. Janotti, A., Van de Walle, C.G.: Oxygen vacancies in ZnO. Appl. Phys. Lett. **8712**, 122102 (2005)
7. Selim, F.A., Weber, M.H., Solodovnikov, D., Lynn, K. G.: Nature of native defects in ZnO. Phys. Rev. Lett. **998**, 085502 (2007)
8. Chang, P.C., Chien, C.J., Stichtenoth, D., Ronning, C., Lu, J.G.: Finite size effect in ZnO nanowires. Appl. Phys. Lett. **9011**, 113101 (2007)
9. Shalish, I., Temkin, H., Narayanamurti, V.: Size-dependent surface luminescence in ZnO nanowires. Phys. Rev. B **6924**, 245401 (2004)
10. Hung, C.S., Gliessman, J.R.: Resistivity and hall effect of germanium at low temperatures. Phys. Rev. **965**, 1226 (1954)
11. Nishimura, H.: Impurity conduction in the intermediate concentration region. Phys. Rev. **1383A**, A815 (1965)
12. Shklovskii, B.I., Efros, A.L.: Electronic properties of doped semiconductors. Springer, Berlin (1984)
13. Chiu, S.-P., Lin, Y.-H., Lin, J.-J.: Electrical conduction mechanisms in natively doped ZnO nanowires. Nanotechnology **20**, 015203 (2009)
14. Kim, S.Y., Jang, H.W., Kim, J.K., Jeon, C.M., Park, W. I., Yi, G.C., Lee, J.L.: Low-resistance Ti/Al ohmic contact on undoped ZnO. J. Electron. Mater. **318**, 868–871 (2002)
15. Fan, Z.Y., Wang, D.W., Chang, P.C., Tseng, W.Y., Lu, J.G.: ZnO nanowire field-effect transistor and oxygen sensing property. Appl. Phys. Lett. **8524**, 5923–5925 (2004)
16. Thompson, R.S., Li, D.D., Witte, C.M., Lu, J.G.: Weak Localization and electron–electron Interactions in Indium-Doped ZnO nanowires. Nano Lett. **912**, 3991–3995 (2009)
17. Martel, R., Schmidt, T., Shea, H.R., Hertel, T., Avouris, P.: Single- and multi-wall carbon nanotube field-effect transistors. Appl. Phys. Lett. **7317**, 2447–2449 (1998)
18. Park, W.I., Kim, J.S., Yi, G.C., Bae, M.H., Lee, H.J.: Fabrication and electrical characteristics of high-performance ZnO nanorod field-effect transistors. Appl. Phys. Lett. **8521**, 5052–5054 (2004)
19. Chang, P.C., Lu, J.G.: Temperature dependent conduction and UV induced metal-to-insulator transition in ZnO nanowires. Appl. Phys. Lett. **9221**, 212113 (2008)
20. Cammi, D., Ronning, C.: Persistent photoconductivity in ZnO nanowires in different atmospheres. Adv. Condens. Matter Phys. **184120** (2014)
21. Bao, J., Shalish, I., Su, Z., Gurwitz, R., Capasso, F., Wang, X., Ren, Z.: Photoinduced oxygen release and persistent photoconductivity in ZnO nanowires. Nanoscale Res. Lett. **61**, 404 (2011)
22. Leiter, F.H., Alves, H.R., Hofstaetter, A., Hofmann, D.M., Meyer, B.K.: The oxygen vacancy as the origin of a green emission in undoped ZnO. Phys. Status Solidi B **2261**, R4–R5 (2001)
23. Stichtenoth, D., Ronning, C., Niermann, T., Wischmeier, L., Voss, T., Chien, C.J., Chang, P.C.,

Lu, J.G.: Optical size effects in ultrathin ZnO nanowires. Nanotechnology **1843**, 435701 (2007)
24. Zhang, B.P., Binh, N.T., Segawa, Y., Wakatsuki, K., Usami, N.: Optical properties of ZnO rods formed by metalorganic chemical vapor deposition. Appl. Phys. Lett. **838**, 1635–1637 (2003)
25. Meyer, B.K., Alves, H., Hofmann, D.M., Kriegseis, W., Forster, D., Bertram, F., Christen, J., Hoffmann, A., Strassburg, M., Dworzak, M., Haboeck, U., Rodina, A.V.: Bound exciton and donor-acceptor pair recombinations in ZnO. Phys. Status Solidi B **2412**, 231–260 (2004)
26. Johnston, K., Henry, M.O., McCabe, D., McGlynn, E., Dietrich, M., Alves, E., Xia, M.: Identification of donor-related impurities in ZnO using photoluminescence and radiotracer techniques. Phys. Rev. B **73**, 165212 (2006)

G

Gallium Arsenide (GaAs)

► Physical Vapor Deposition

Gamma-Butyrolactone (GBL)

► SU-8 Photoresist

Gammatone Filters

► Bioinspired CMOS Cochlea

Gas Etching

► Dry Etching Processes

Gas-Phase Molecular Beam Epitaxy (Gas-Phase MBE)

► Physical Vapor Deposition

Gas-Phase Nanoparticle Formation

Yuehai Yang and Wenzhi Li
Department of Physics, Florida International University, Miami, FL, USA

Synonyms

Nanoscale particle; Nano-sized particle

Definition

In the materials science community, the term "nanoparticle" is generally used to indicate particles with atomic domains spatially confined to less than 100 nm since the physical properties of materials start to change significantly due to a variety of confinement effects in this size range. Another commonly used concept, "cluster," indicates smaller particle with a diameter of only several nanometers, corresponding to 10^4 molecules or atoms. Sharing the same features with other types of nanostructured materials, nanoparticles also have significant atom fractions associated with interfacial environments, and interactions between their constituent domains. Gas phase particle formation refers to production of particles from individual atoms or molecules in the gas phase.

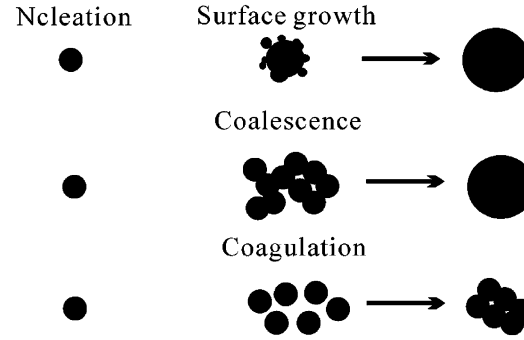
© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Overview

In the last several decades, intensive research has been performed on nanostructured materials, that is, materials with grain sizes less than 100 nm. Nano-sized particles are considered as the starting point for preparations of an extensive variety of nanostructured materials. Meanwhile, nanoparticles themselves have played a significant role in novel technology development [1]. As such, they have been prepared by a variety of methods and their synthesis is a crucial part of rapidly growing nanoscale research.

Gas phase synthesis technique is well known for producing a wide range of nanoparticles. Gas phase particle generation is captivating because it excludes the wet byproducts of liquid phase processes and the as-prepared product particles promptly separated from the gas stream are in high purity without post-growth treatment. The gas phase synthesis has been extensively developed and widely applied in industries. Meanwhile, the commercialization of new products has experienced a considerable progress and the commercial nanoparticles produced by this method comprise the largest share of the market. In the synthesis of nanoparticles from atomic/molecular precursors, one wants to be able to control different aspects of this condensed ensemble, in which the most important one is the size and size distribution of the resultant particles. The reason, as indicated above, is that the material parameters of a solid experience a dramatic change when cluster size becomes less than a certain threshold value. A property will be altered if the entity is confined within a space smaller than the critical length of the entity responsible for this property. Therefore, the product quality and application characteristics of nanoparticles strongly depends on the distribution of particle size and on the particle morphology. Well-controlled gas phase routes are able to synthesize nearly spherical and nonporous particles with high purity. The resulting particles are small in size with narrow size distribution.

In gas phase synthesis of nanoparticles, conditions include usual situations of a supersaturated vapor, or chemical supersaturation in which it is thermodynamically favorable for the gas phase molecules to undergo chemical reactions and condensation. If the reaction and condensation kinetics permit, particles will nucleate homogeneously when the degree of supersaturation is sufficient. Once nucleation occurs, the reaction/condensation of the gas phase molecules will occur on the resulting particles and will keep relieving the remaining supersaturation, since then the nucleation process stops. With this gas phase conversion process, particles are built all the way up to the desired size.



Gas-Phase Nanoparticle Formation, Fig. 1 Gas phase particle formation mechanism

Although the processes of gas phase nanoparticle synthesis vary from one to another, they all have in common the fundamental aspects of particle formation mechanisms. In gas phase reactors, three major formation mechanisms determine the particle formation within the first few milliseconds of the synthesis process, and will determine the final characteristics of particulate products. The influences of these mechanisms on particle formation, growth, and final morphology are shown in Fig. 1.

1. The precursor chemical reactions form product clusters by nucleation, and then the clusters grow into particles by surface growth [2].
2. Because of the sintering processes, coalescence occurs in the high-temperature zones of the reactor to reduce aggregation and enhance the formation of spherical particles [3].
3. Also, coagulation occurs inevitably in the gas phase synthesis with high particle concentrations, after the formation of particles in a gas

phase; their coagulation rate is weakly dependent on particle size and proportional to the square of particle number concentration.

If temperature is sufficiently high, the coalescence of particles is much faster than their coagulation to produce spherical particles. On the contrary, loose agglomerates are formed with open structures.

Most advances in this field closely relate to gaining better understanding and better control of particle aggregation and coalescence in order to produce nonporous particles with desired size and narrow size distribution. Gas phase synthesis is one of the best techniques in controlling nucleation-condensation growth and collecting/handling nanoparticles afterward. There are various parameters such as flame temperature, precursor residence time, ions, gas concentration, cooling rate, and additives which can affect the process of coagulation and sintering, and consequently, the characteristics of the resulting particles with respect to the size monodispersity. In addition to the particle size and size distribution, the composition of the constituent phases and the nature of the interfaces created between constituent phases are also of crucial importance if the particles are constituted by more than one material. The interplay among size, composition, and interfaces determines the properties of nanoparticles.

Specific Synthesis Methodologies

The physical and chemical processes for nanoparticle formation in gas phase are qualitatively understood. The process has been used in the formation of millions of tons of particles in nano-/submicroscale every year in different industries. Particle generation in the gas phase is carried out either by precursor gases reaction or by evaporation. To generate the high temperature for these processes, high-energy sources such as Joule heating, hydrocarbon flames, plasmas, ions, electrons, or laser beams are needed. Therefore, different kinds of furnace, flame, plasma, and laser reactors have been used to facilitate gas phase reactions and produce the desired materials.

Gas Phase Condensation (for Metal Nanoparticles)

To achieve supersaturation, the most straightforward method is to heat certain solids and evaporate them into background gases, and then reduce the temperature by mixing the vapors with cold gases. Formed by homogeneous nucleation in the vicinity of the gas source, the clusters grow in this process by gas phase coalescence of the source atoms. It has been found that the average particle size of the resulted nanoparticles increases with the increase of gas pressure and the applied inert gas mass. Experimentally, it has also been found that the particle size distribution is lognormal. This has been explained theoretically by the particle's growth mechanisms [4]. This method is also well suited for producing metal nanoparticles as for most metals, the evaporating temperature is attainable and evaporation rates are reasonable. Furthermore, compounds of the evaporated metal can be prepared by including a reactive gas in the cold gas stream [5].

In the gas-condensation process, it has been studied [6] that the formation of the atom clusters is essentially controlled by three fundamental rates:

1. The evaporation rate
2. The rate that the hot atoms lose their energy in the condensation region
3. The removal rate of the nucleated clusters from the supersaturation region

The evaporation-condensation process functions as a form of distillation. Therefore, with volatile impurities removed in this process, the purity of the material can be further improved. However, there are a few limitations for this synthesis method: (1) it is difficult to control composition in the case of synthesizing more complex materials by this method and (2) this process is restricted to materials which have low melting points or high vapor pressures.

Nanoparticle Formation in Flames

So far, in the nanoparticle synthesis, the most commercially successful approach is carrying out the synthesis process within a flame, which provides the energy to evaporate the precursors so that the heat needed to induce reaction and particle

nucleation is produced by the combustion reactions. Every year, this method alone produces millions of metric tons of nanostructured materials, such as carbon black and metal oxides. Since the flame provides oxidizing environment, it has been primarily used to make oxides materials.

High temperatures (up to 2,400 °C) can be realized in the flames and the precursor concentration in the flame can be quite high due to the high energy density. In the highest temperature region where the primary particles are formed, the residence time only ranges between 10 and 100 ms. After this region, the only things that can be influenced are the particle size and the aggregation morphology. The attainable particle sizes using flame synthesis is from a few nanometers up to 500 nm. The specific surface areas of these particles can go 400 m^2/g and higher.

Temperature profile, reactor residence time, and reactant concentration are the three crucial parameters for generating the desired particles. Unfortunately, the coupling of the flame chemistry to the nanoparticle production makes this process very complex and difficult to control. Every adjustment to the feed flow changes all three parameters simultaneously. Therefore, it is impossible to change these three parameters independently in flame reactors.

Furnace Flow (Hot-Wall) Reactor

Furnace flow systems employ tubular furnace-heated reactors for initiating the synthesis reaction. Temperatures for reaction are often set below 1,700 °C. The gas composition is freely selectable with variable concentrations. Usually, the system pressure is atmospheric, but it can also be set at other values and used as a process parameter.

With high efficiency, the reactor has a relatively simple design. The process of this technique is precisely controllable and flexible with respect to gas composition and system pressure. Therefore, it allows the production of particles with specific characteristics. It also allows different kinds of particle production in the range from atomic to micrometer dimensions. However, it often causes a high degree of aggregation at high aerosol concentrations. Moreover, compared to other methodologies, it has high energy requirements.

Plasma Reactor

Supersaturation and particle nucleation can also be achieved by injecting the precursors into thermal plasma to provide the energy needed for inducing reactions. It is quite common to deposit thin films and coatings by plasma spraying and thermal plasma vapor (chemical or physical) deposition. The most commonly used electrical methods to produce plasmas are high-intensity arcs and inductively coupled high-frequency discharge. Solid precursors are conveyed in the plasma where they are fully evaporated and decomposed into atoms in the plasma center by the high temperatures. The precursor vapors are then cooled rapidly by mixing with quenching gas at the end of the plasma. As a result of high supersaturation, ultrafine particles are obtained.

Ion Sputtering

The method of sputtering is to vaporize materials by bombarding a solid surface with inert gas ions shot in high velocities. The sputter source, for example, an ion gun, usually works in systems with of 10^{-3} mbar or even higher vacuum, since a higher pressure will block the sputtered material transportation. Electrons can be used to replace ions for the same purpose. The major advantage of sputtering is that the material being heated is mainly the target material, and thus the composition of the sputtered material is that of the target. However, this process needs to be carried out at relatively low pressures which can cause difficulties for further nanoparticles processing in gas phase.

Spark Discharge

High-current spark between two solid electrodes is used as another means of vaporizing metals. In the presence of an inert background gas, the electrode material is evaporated to produce nanoparticles until it reaches the breakdown voltage. This technique is very useful for materials with a high melting point (e.g., Si or C), which cannot be evaporated in a furnace. The spark (arc) formed across the electrodes can only vaporize a small amount of metal to produce very small amounts of nanoparticles, but this process is relatively reproducible.

Laser Ablation

Instead of evaporating certain material to produce a supersaturated vapor, a pulsed laser can be utilized to evaporate a piece of spatially confined material. The laser beams are focused onto the center of the reaction chamber where the precursor (e.g., coarse metal oxide powder) is fluidized, resulting in a power supply of a few kilowatts in a volume of only a few cubic millimeters. The precursor powder is evaporated by the focused laser, and nanoparticles are formed by the vapor condensation. This method can be used to vaporize materials that cannot readily be evaporated even though laser ablation can generally only produce small amounts of nanoparticles. For solids having high vaporization temperatures, typically ceramics and metal oxides, from which nanoparticles cannot be synthesized by standard gas phase processes, the high-power intensity of the laser opens a gate for them to be used as precursors. The nanoparticles synthesized by laser have potentials for new applications since their morphologies are significantly different from typical pyrogenic oxides due to the high cooling rates.

Nanoparticle Collection

Among the collection devices, the simplest one is a small-size pore filter for separating nanoparticles from a gas stream. The disadvantage is that impurities and defects can result from the incorporation of the filter material parts into the nanocrystallined particles. Therefore, by employing thermophoretic forces, a collection device was designed to apply a permanent temperature gradient, which results in the separation of nanoparticles from the gas that has been used. In addition, collection of the nanoparticles in liquid suspension can be used to enhance the stability of the collected nanoparticles against compositional changes, sintering, and agglomeration.

Limitation of Gas Phase Process

1. Hazardous gaseous reactants and its byproducts are apparent disadvantages of gas phase nanoparticles synthesis.
2. Nanoparticles agglomeration is inevitable in the gas phase synthesis. However, by capping the particles with appropriate ligands in the liquid phase synthesis, the dispersion of nanoparticles can be indefinitely stabilized.
3. In gas phase synthesis, it is often difficult to produce composite materials with uniform chemical compositions because heterogeneous compositions within an individual particle or from particle to particle can be caused by differences in the vapor pressures, chemical reaction rates of the reactants, and nucleation/growth rates of the products.

Engineering Applications

Nanoparticles have been applied or been evaluated for use in many fields by employing their electrical, magnetic, thermal, optical, and mechanical properties. Nanoparticles synthesized in gas phase have also emerged into engineering fields such as heterogeneous catalysis [7], biomaterials, electroceramics manufacturing, dental materials, and fuel cell membrane construction. The advantage of nanoparticle application is that it incorporates many effects related to their sizes. These effects include “the quantum size effects” (electronic effects caused by delocalized valence electron confinement), “the many-body effect” (altered cooperative atom phenomena), for example, lattice melting or vibrations, and suppression of the lattice-defect mechanisms, such as dislocation generation and migration within the confined grain sizes. The capability of engineering a large variety of useful technological applications will be impacted by possibly size-selected nanoparticle synthesis and assembling the nanoparticles into novel materials with unique and improved properties including controlled mechanical, electronic, optical, and chemical properties. Microelectronics and even medical applications [8] will profit from developing this nanoparticle formation process.

Future Research

Based on the newly arisen questions and challenges, research in this field underlines the need

for synthesizing nanoscale particles (especially mixed oxides particles) with precisely controlled characteristics. Future research includes topics such as the scale-up synthesis of particles with controlled properties, verification of mesoscopic chemistry relationships that can be applied in research, and nano-thin coated nanoparticles made in large quantities. In this endeavor, scientists and engineers should accumulate their knowledge and contribute to small particles generation, granulation, flocculation, crystallization, and classification.

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates
- ▶ Electric Field–Directed Assembly of Bioderivatized Nanoparticles
- ▶ Electric-Field-Assisted Deterministic Nanowire Assembly
- ▶ Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms
- ▶ Fate of Manufactured Nanoparticles in Aqueous Environment
- ▶ Genotoxicity of Nanoparticles
- ▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Mechanical Properties of Nanocrystalline Metals
- ▶ Nanoparticles
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Perfluorocarbon Nanoparticles
- ▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity
- ▶ Self-assembly of Nanostructures
- ▶ Synthesis of Gold Nanoparticles
- ▶ Synthesis of Subnanometric Metal Nanoparticles

References

1. Afzaal, M., Ellwood, K., Pickett, N.L., O'Brien, P., Raftery, J., Waters, J.: Growth of lead chalcogenide thin films using single-source precursors. J. Mater. Chem. **14**, 1310–1315 (2004)
2. Pratsinis, S.E., Spicer, P.T.: Competition between gas phase and surface oxidation of $TiCl_4$ during synthesis of TiO_2 particles. Chem. Eng. Sci. **53**, 1861–1868 (1998)
3. Johannessen, T., Pratsinis, S.E., Livbjerg, H.: Computational fluid-particle dynamics for the flame synthesis of alumina particles. Chem. Eng. Sci. **55**, 177–191 (2000)
4. Granqvist, C.G., Buhrman, R.A.: Log-normal size distributions of ultrafine metal particles. Solid State Commun. **18**, 123–126 (1976)
5. Wegner, K., Walker, B., Tsantilis, S., Pratsinis, S.E.: Design of metal nanoparticle synthesis by vapor flow condensation. Chem. Eng. Sci. **57**, 1753–1762 (2002)
6. Siegel, R.W. In: Cahn, R.W. (ed.) Materials Science and Technology, vol. 15, p. 583. VCH, Weinheim (1991)
7. Chiang, W.H., Sankaran, R.M.: Synergistic effects in bimetallic nanoparticles for low temperature carbon nanotube growth. Adv. Mater. **20**, 4857–4861 (2008)
8. Boisselier, E., Astruc, D.: Gold nanoparticles in nanomedicine: preparations, imaging, diagnostics, therapies and toxicity. Chem. Soc. Rev. **38**, 1759–1782 (2009)

Gecko Adhesion

Elmar Kroner[1] and Eduard Arzt[1,2]
[1]INM – Leibniz Institute for New Materials, Saarbrücken, Germany
[2]Saarland University, Saarbrücken, Germany

Synonyms

Dry adhesion; Fibrillar adhesion

Definition

The gecko adhesion system is a dry adhesive based on a fibrillar surface pattern and allows easy, repeatable, and residue-free detachment.

Conventional Adhesives and the Gecko Adhesion System

Adhesives are an important ingredient of modern technology. Their function is to reliably connect

two different objects with each other. Although other techniques may be used to assemble two objects (e.g., screwing or welding), the application of adhesives has many benefits. For example, they provide a more uniform distribution of stresses along joints, can be applied at ambient (or moderately elevated) temperature, or can fulfill additional requirements such as acoustical damping or energy dissipation. Adhesives have also been key to new technologies, e.g., in the development of laminated fiber-reinforced materials. The need for adhesives in special applications has led to the development of more than 250,000 adhesives in use today.

Nature has also evolved adhesives for numerous purposes. A special case is a reusable adhesive in which attachment, detachment, and reattachment occur without loss in adhesion strength; insects, spiders, tree frogs, and geckos can stick to walls and ceilings, attaching and detaching their feet rapidly and repeating this for thousands of times without losing the ability to stick. Geckos belong to the heaviest animals supporting such an adhesive system, which exhibits the following characteristics:

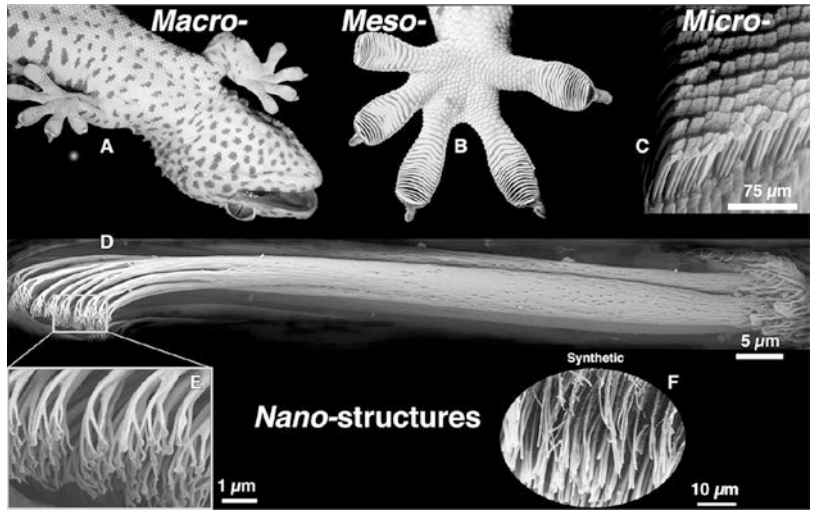
- *High adhesion to smooth and rough surfaces based on physical interactions*: Unlike conventional adhesives, which are usually based on chemical bonds or mechanical interlocking, gecko adhesives work without pretreatment of the surface, e.g., by cleaning or roughening.
- *Adhesion barely depends on the chemical properties of the surface*: Conventional adhesives often have to be tailored to different surfaces to achieve high adhesion.
- *Directional adhesion, allowing easy detachment*: Unlike conventional adhesives, gecko adhesives enable anisotropic adhesion properties. While sticking occurs in one direction, detachment happens in another.
- *High reversibility of adhesion*: The reversibility of conventional adhesives is poor and seldom lasts more than a few attachment cycles. Gecko adhesives can attach and detach hundreds or thousands of times.
- *Residue-free detachment*: Most conventional chemistry-based adhesives leave traces of residues on the surface during removal. Note that in a recent study phospholipids were found as gecko "footprints," but the source of the lipids and their role in adhesion are not yet clear [1].
- *Self-cleaning properties*: Dust and particles, which stick to conventional adhesives, can be removed easily in gecko adhesives.

Mimicking the attachment system of geckos has attracted the attention of a worldwide research community. A successful development could lead to significant improvements in countless fields of application or even lead to new technologies. After about a decade of research, first applications begin to appear on the horizon.

History of Gecko Adhesion Research

The ability of geckos to climb up walls and run along ceilings on different materials and various degrees of roughness has fascinated humans for thousands of years. Geckos were first mentioned in writing by the Greek philosopher *Aristotle* (350 B.C.), who stated in his work *Historia Animalium*: "The woodpecker [. . .] can run up and down a tree in any way, even with the head downwards, like the gecko-lizard." However, the reason behind this astonishing adhesion performance was discovered only with the development of electron microscopy in the twentieth century.

A brief study of the gecko toe structure using scanning electron microscopy was published by Hiller and Blaschke in 1967 [2]. They showed that gecko toes are covered with fine hairs (called *setae*), tens of micrometers in length and several micrometers in thickness. These *setae* branch into even finer hairs (*spatulae*) and end in plate-like tips with diameters of about 200 nanometers (nm) (Fig. 1). The results of their observation led them to the now refuted conclusion that the surface structure of the gecko toes provides high adhesion and friction due to mechanical interlocking on micro and nanorough surfaces. In 1968, Hiller found a correlation between water contact angle and attachment forces of gecko toes and corrected his statement from the



Gecko Adhesion, Fig. 1 Adhesion organ of the gecko (**a**). The toes display fine lamellae (**b**), which consist of thousands of setae (**c**, **d**). These setae branch into very fine tips, called spatulae (**e**) [4]. Similar structures can be fabricated using modern microfabrication techniques (**f**) (Reproduced with kind permission of K. Autumn and The Royal Society Publishing)

previous year by indicating that their function depends on "adhesion processes" rather than mechanical interlocking. In 1980, Stork [3] published an extensive experimental study on the beetle *Chrysolina polita*, which also suggests an adhesion system based on surface structures similar to those found in geckos. In his paper, he identified the main contribution to adhesion as molecular interactions, possibly with additional contribution from capillary forces.

Two decades later, research on the gecko adhesion system has developed to such an extent that it has become a prime topic in biomimetics. In their 2000 publication, Autumn et al. [5] measured the adhesion force of single gecko *setae*. They extrapolated that, for all *setae* in perfect contact, a single gecko foot could adhere with a shear force of approximately 100 Newton (N). Triggered by this astonishing sticking ability, many key advances have been made in recent years:

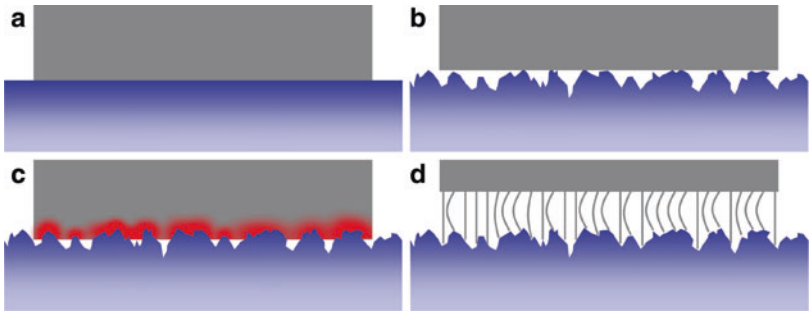
- Modern micro- and nanofabrication methods allowed bioinspired surface patterns to be produced with increasingly complex geometries.
- New methods for structural and mechanical characterization enabled extensive studies on the adhesion behavior of micropatterned surfaces.
- New micromechanical models helped explain the "gecko effect" and served as design guides for successful artificial surface patterns.
- The properties of gecko-inspired adhesives, i.e., high physical adhesion with high reversibility and repeatability, rapid and easy detachment, and self-cleaning properties, provoked interest from several application sectors, e.g., for robot locomotion, microfabrication, biomedical use, or sports articles.

While the gecko-inspired adhesion system offers the most spectacular characteristics, other adhesion systems in nature, e.g., of tree frogs, spiders, and insects, are at least as interesting as the gecko system. An overview of adhesion in nature can be found in [6].

Contact Mechanics of Patterned Surfaces

To understand the gecko attachment system in its complexity, it is necessary to apply principles of contact mechanics to the special case of patterned surfaces. The adhesive interaction between two surfaces can be investigated in a static (no movement allowed) or dynamic (movement allowed, e.g., peeling) system.

Static system for adhesion – The interplay between attractive and repulsive forces determines the strength of the adhesive force between two touching objects. Attractive forces can arise, e.g., from molecular interactions due to van der Waals forces, electrostatic or magnetic forces,



Gecko Adhesion, Fig. 2 Contact scenarios of two bodies: (**a**) two smooth surfaces with maximum contact area, (**b**) a smooth and a rough surface, both rigid, (**c**) as in (**b**) but with elastic deformation of the upper solid (high stresses are schematically indicated in *red*), and (**d**) a patterned surface which can adapt to a rough counterpart with only a small elastic energy penalty

capillary forces, or gravity. For the gecko adhesion system, the main contribution is the presence of molecular interactions as demonstrated by Stork [3]. More recent studies have shown that humidity effects also contribute to adhesion; absorbed water may reduce surface roughness [7] or influence the material properties of the gecko structures [8], both leading to increased adhesion. Repulsive forces can result, e.g., from electrostatic and magnetic repulsion or – most important in the case treated here – the energy stored in the elastically deformed bodies. If the attractive forces (positive per definition) and repulsive forces (negative) are now balanced, adhesion occurs if the attractive forces are larger than the repulsive forces. If the surfaces adhere to each other, an *external* force is necessary to detach the two surfaces. This force is defined as having a negative sign and is called pull-off force P_c.

Van der Waals forces, being the result of dipole interactions, are short range and become significant only for distances below several nanometers. Surface chemistry has only a small influence on the magnitude of the attractive forces compared to the distance and is fixed for a contact pair. Thus, the only parameter which can be modified to increase the attractive interactions is the distance between the contacting surfaces.

For illustration, consider some special cases: when perfectly smooth surfaces, e.g., the flat sides of two silicon wafers, are brought into contact, the true contact area is large and elastic deformation is unnecessary; it will then be impossible to detach the wafers from each other without damaging them (this effect is exploited in a processing step called "wafer bonding"); see Fig. 2a. However, if a stiff solid is placed onto a rough surface, the true contact area will be negligible (Fig. 2b), and it will not stick to it; a small roughness is sufficient to significantly decrease the contact area and, due to their extremely short force range, the attractive forces. In order to increase the contact area, an external force compressing the two surfaces is necessary. However, elastic deformation stores energy in the two objects and causes repulsive forces, resulting in vanishing adhesive forces (Fig. 2c).

To obtain significant adhesion, it is necessary to increase the true contact area while minimizing the elastic energy penalty. This can be realized by designing an adhesive with an *effective* Young's modulus as low as possible. The pioneering adhesives scientist *Carl A. Dahlquist* defined the limiting effective Young's modulus to achieve "tackiness" to be below 100 kilopascal (kPa). Such low values may be obtained by either choosing very soft adhesives, usually soft polymeric materials, or by introducing surface structures with high aspect ratios (Fig. 2d), where the effective Young's modulus is decreased due to additional bending of the structures. Although the contact area is decreased by patterning the surface, the reduction of the repulsive forces due to the gain in compliance can be higher, thus leading to an increased adhesion force.

The fabrication of low modulus materials seems to be tempting as a complicated microstructuring process can be avoided. However, such materials have several drawbacks for reversible and repeatable attachment. First, low modulus materials usually wear quickly, which would significantly decrease the repeatability of the adhesive performance. Second, soft materials tend to transfer material to the counter surface, e.g., free oligomers in soft polymers, disqualifying them as residue-free. Third, continuous flat surfaces will contaminate easily with dust particles and dirt, necessitating a periodic cleaning process when applied in dirty environments. Fourth, due to the isotropy within the surface plane, the adhesion will be isotropic as well. Thus, there will be no preferred detachment direction, which is a key property of the gecko adhesion system to allow locomotion. And finally, under dynamic attachment and detachment conditions, the soft non-patterned surface will be inferior to patterned surfaces, which will be explained in the following paragraph.

Dynamic system for adhesion – Patterned surfaces are superior in adhesion compared to flat ones, especially when the dynamics of detachment is considered. If detachment is initiated for a flat adhesive, the elastically stored strain energy freed in the process is available to drive the detachment front further. For a disrupted surface, however, the strain energy stored in single fibrils will be lost on detachment and will not contribute to advance the detachment front (Fig. 3a). An adhesive contact can be described as a fracture problem, where the initiation of a detachment front is regarded as crack initiation and the advancing detachment corresponds to crack propagation. In fracture processes, the initiation of a crack is usually energetically more difficult than crack propagation. Once initiated, a crack will propagate easily. This analogy can be translated to flat and patterned adhesives. For a flat adhesive, the detachment front needs to be initiated only once and will subsequently propagate. However, for a patterned surface, new detachment cracks need to be initiated for each single contact, leading to an energetically more difficult detachment process.

Size Matters: The Principle of "Contact Splitting"

In 1971, Johnson, Kendall, and Roberts [9] introduced a theory of contacts between soft elastic spheres, which soon became the most highly cited publication for contacts between soft materials, now known as the JKR theory. By balancing elastic, potential, and surface energy, they obtained a theoretical expression for the external pull-off force P_c required to separate two spherical bodies:

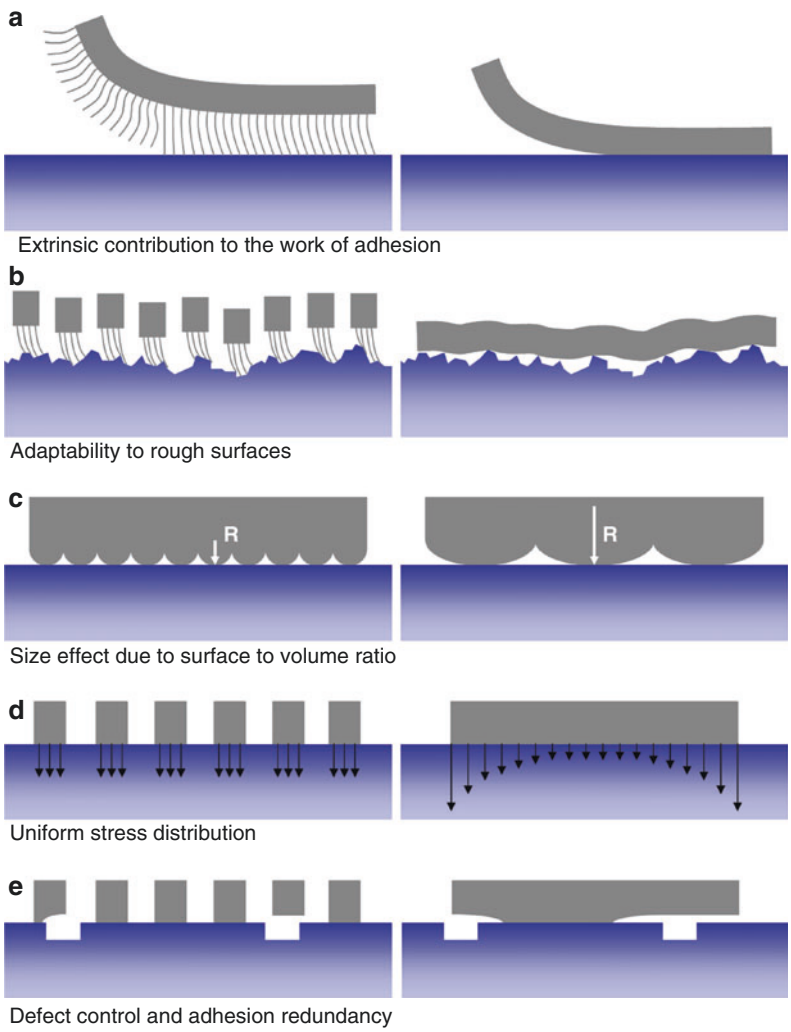
$$P_c = -\frac{3}{2}\pi\gamma R \tag{1}$$

Here, R is the relative radius of curvature defined by $1/R = 1/R_1 + 1/R_2$, R_1 and R_2 being the radii of the two contacting spheres, and γ the "work of adhesion," defined as $\gamma = \gamma_1 + \gamma_2 - \gamma_{12}$, where γ_1 and γ_2 are the surface energies of the two contacting materials, respectively, and γ_{12} is the interfacial energy. The negative sign in Eq. 1 occurs because tensile forces are defined to be negative in the contact mechanics literature. A thorough derivation of the JKR theory, additional theories for soft elastic contacts, and adhesion problems of other contact geometries such as flat contacts can be found in the books by Johnson [10] and Maugis [11].

In 2003, Arzt et al. [12] showed that Eq. 1 results in a generic size effect for adhesion: the pull-off force for a large contact is smaller than the sum of the pull-off forces for many smaller contacts together, covering the same apparent contact area. They termed this concept the "principle of contact splitting." One consequence of this principle, e.g., is the following: the number of equally sized spherical contacts n is proportional to the contact area A, which itself is proportional to the square of a length. Considering Eq. 1, it follows that the pull-off force is proportional to the square root of the number of contacts:

$$P_c \sim R \sim \sqrt{A} \sim \sqrt{n} \tag{2}$$

Thus, with increasing number of contacts, the pull-off force increases as well. Other contact

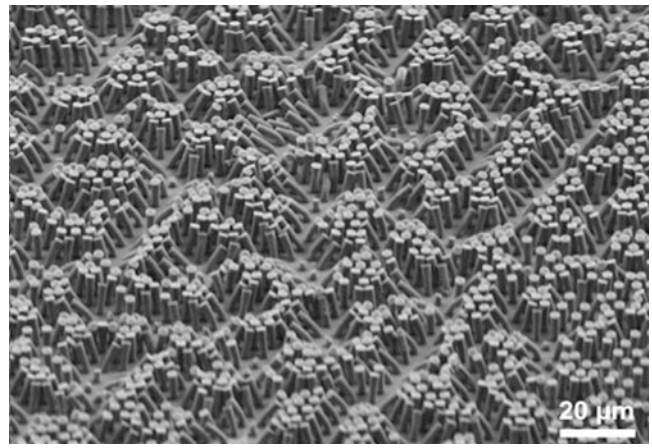


Gecko Adhesion, Fig. 3 Some theoretical explanations for the principle of "contact splitting," i.e., why arrays of small discrete contacts show increased adhesion compared to single large contacts (Schematics after [14])

geometries than spherical contacts lead to different exponents for the number of contacts but still result in an increasing pull-off force for increasing number of contacts. This simple model offered an explanation for the experimental observation that with increasing animal body weight, the size of their contact elements decreased. Their prediction fitted well to about six orders of magnitude in animal body weight, although the effect mainly reflected the differences between different animal groups (flies, beetles, spiders, and geckos), but not within these groups. It was later pointed out by Federle [13] that the data included a transition from fluid mediated to dry adhesion, which can give rise to additional effects.

The experimental finding that smaller contacts showed increased adhesion compared to larger ones has triggered numerous theoretical explanations (Fig. 3). They may be summarized as follows (classification after [14]):

- Extrinsic contribution to the work of adhesion (Fig. 3a): Patterned surfaces are more resistant to peeling because the stored strain energy in a pillar just before detachment is not available to drive detachment to the next fibril.
- Adaptability to rough surfaces (Fig. 3b): A patterned surface with long fibrils can adapt better to rough surfaces with less storage of elastic strain energy.



Gecko Adhesion, Fig. 4 "Condensation" of polydimethylsiloxane pillars. At a certain aspect ratio of the pillars, the attractive forces between the single pillars overcome their resistance to bending, leading to a collapse of the structures (Picture reproduced with permission of C. Greiner)

- Size effect due to surface-to-volume ratio (Fig. 3c): The volume, controlling the storage of elastic energy, decreases more rapidly than the gain in surface energy, favoring smaller contacts.
- Uniform stress distribution (Fig. 3d): Below a critical size of a contact, the stress distribution within the contact area becomes uniform, resulting in adhesion strengths near the theoretical values.
- Defect control and adhesion redundancy (Fig. 3e): If the adhesion is controlled by defects, smaller contacts will be more tolerant to defects than large contacts.

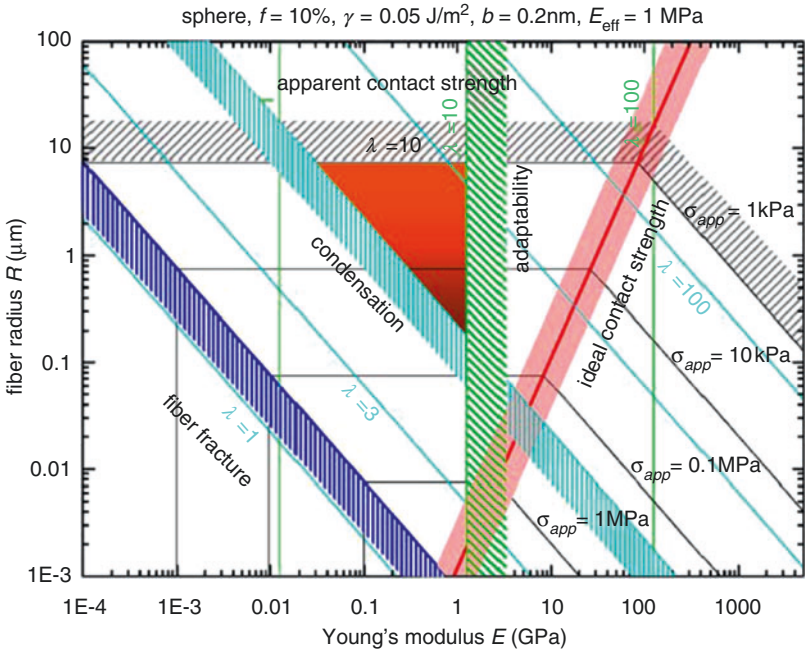
These considerations have led to the conclusion that a contact should be split into densely packed elements which should be as long and thin as possible to achieve maximum adhesion. However, if densely packed pillars exceed a critical aspect ratio (ratio between pillar length and diameter), they may become unstable with regard to "condensation" (also called "bundling" or "collapse"), as shown in Fig. 4. The reason for condensation of pillars is the low bending resistance combined with inter-pillar adhesion. This effect leads to a trade-off between pillar length, diameter, and elastic properties which can be conveniently displayed in "adhesion design maps" (Fig. 5). Illustrating the theoretical optimum in the parameter space, these maps show that, e.g., "condensation" is the overriding limitation for the refinement of a patterned adhesive. To adhere to surfaces with higher asperities, longer pillars would be needed, requiring large diameters to avoid condensation. However, there is a solution to this problem which can be found in the natural gecko adhesive: a hierarchical structure.

Role of Hierarchical Geometry for Gecko Adhesion Systems

On the macroscopic scale, a gecko has four legs with five toes each to adapt to large asperities in the range from centimeters to millimeters. On the microscopic scale, the *lamellae*, *setae*, and *spatulae* can adapt to asperities in the range of several tens of microns down to the submicrometer range. This hierarchical structure combines the advantages of small terminal contacts with stability against condensation. The influence of hierarchy on adhesion has been modeled repeatedly. Besides finite element method simulations and analytical calculations, models based on a hierarchical coupling of springs have been proposed [16].

Anisotropic Adhesion

Another key feature of geckos is their ability to rapidly detach their feet from a surface, which is a necessary feature for swift locomotion. The key to high adhesion and rapid detachment lies in the combination of their biomechanics of motion

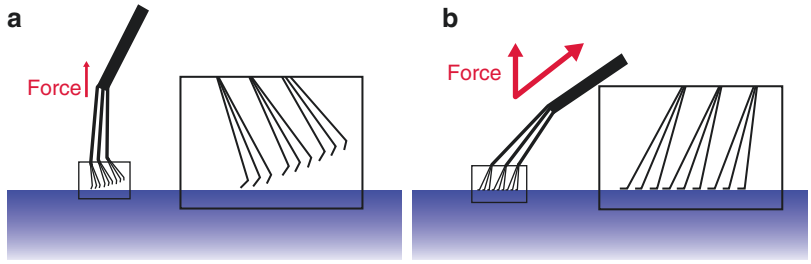


Gecko Adhesion, Fig. 5 Adhesion design map to predict successful surface patterns. The map spans a space containing material properties (Young's modulus) and geometrical parameters (radius of the fibers). Limitations are drawn in, such as fiber fracture (the adhesive strength is stronger than the material strength, causing the fiber to break rather than detach), ideal contact strength (maximum adhesion force generated by van der Waals forces), fiber condensation, and adaptability of the surface pattern. The *light blue lines* indicate the aspect ratio λ. The theoretical optimum can be found at the lower apex of the *orange triangle* [15] (Reproduced with permission of E. Arzt and Elsevier)

and the anisotropy of their adhesive structures. A closer look into their adhesive structures (Fig. 1) reveals slightly bent features with plate-like tips. When these features form contact with a surface under normal loading, the plate-like tips will be ill oriented to the surface and form insufficient contact (Fig. 6a) with low pull-off force. On application of a shear force, the structures will bend, allowing the plate-like tips to orient along the surface (Fig. 6b). The tips will then form a large contact area, enhancing adhesive forces. This so-called frictional adhesion allows the gecko to combine high sticking forces with fast detachment: orientation of the tips by pulling induces adhesion, whereas pushing or a release of the pull causes rapid pull-off. Additionally, the special anatomy of the gecko toe allows hyperextension, i.e., rolling of the toes upward, which simplifies detachment further. This shear-induced adhesion also explains the paradox that a surface provides high adhesion and, at the same time, has self-cleaning properties, i.e., no dust or particles stick to the attachment pad; the attachment organs are only adhesive if a shear force is applied. In the absence of a shear force, the adhesion is low, hindering the adhesion of dust particles to the surface structure.

Artificial Gecko Adhesives

Over the last decade, scientists around the world have tried to mimic the gecko adhesion system with artificial structures, usually fabricated from polymeric material. The fabrication methods reflect the strategies employed in microfabrication: bottom-up patterning (assembling structures from atomic or molecular materials), template-less top-down patterning (e.g., removing material from bulk), and top-down patterning



Gecko Adhesion, Fig. 6 Hierarchical gecko structures in contact with a smooth surface. If a normal force is applied, the tips of the hierarchical structures will be misaligned with the surface, leading to a low pull-off force (**a**). By applying a shear force, the structure tips become oriented, forming a large contact area and enhancing adhesion (**b**)

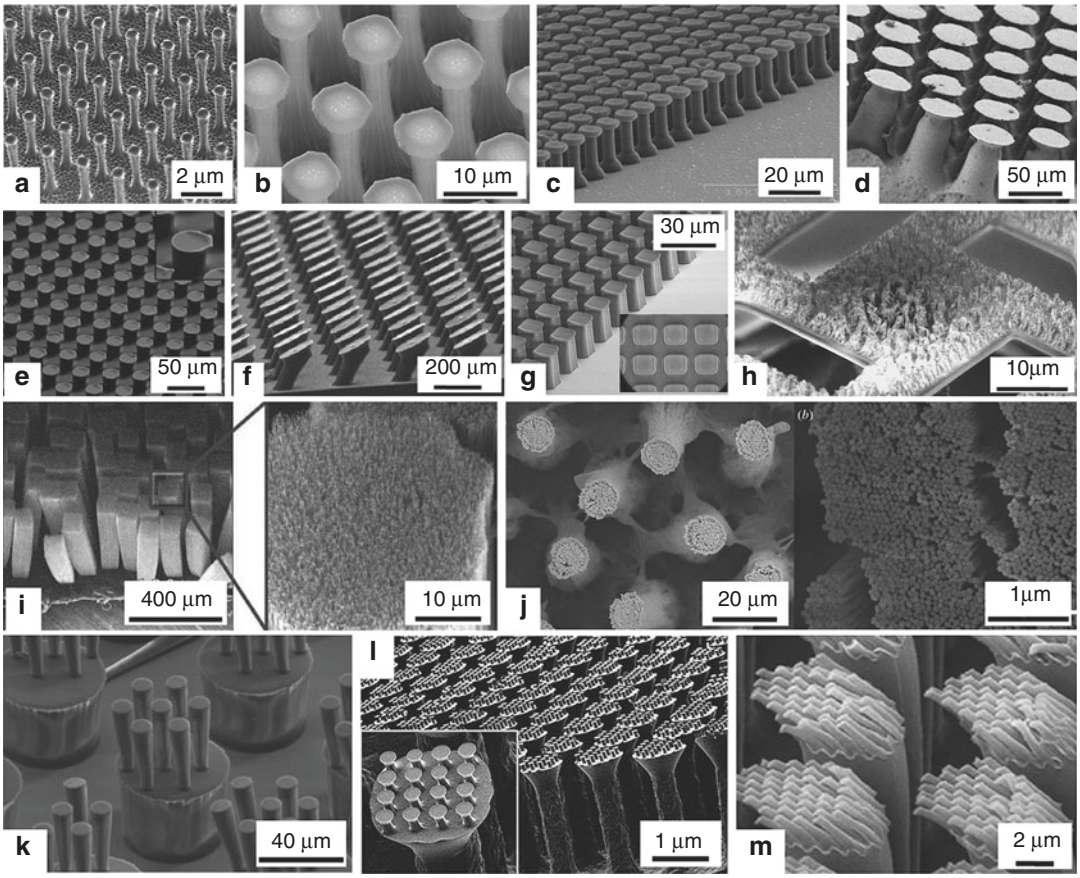
using a template (e.g., molding of pre-patterned surfaces). Figure 7 shows examples of recently fabricated gecko-inspired adhesives using different processes.

An example of bottom-up patterning is the fabrication of oriented carbon nanotubes on a substrate. Although carbon nanotube-based adhesives can provide high shear forces, the compressive preload necessary to trigger adhesion sometimes even exceeds the resulting normal and shear force.

Template-less top-down methods are based on either irradiation techniques to cross-link or decompose a polymer, such as UV light, lasers, and electron or ion beams, or on "mechanical" methods, i.e., cutting, writing, or melting of polymers. These techniques allow high precision in fabrication of surface structures, and features below 1 μm are accessible. Although some techniques allow the fabrication of larger areas in a short time (e.g., interference lithography), template-less top-down techniques are usually serial processes, where one structure is fabricated after another. This largely increases the fabrication time to achieve samples of feasible size.

The most widely used fabrication techniques are based on top-down structuring using a template. A template is usually fabricated by photolithography and is filled by a liquid polymer ("molding"). Although this process is usually more complex than the template-less techniques, it has a large benefit: the template may be used many times. If desired, the fabricated samples may be processed further, especially to modify the tip geometry of the pillars. Del Campo et al. [21] showed systematically in 2007 that the tip geometry of bioinspired adhesive structures is crucial to obtain high adhesive forces. They found that pillars with spherical, flat punch, or suction-cup tips showed relatively low adhesive forces, while spatular or mushroom-shaped tips showed high adhesion. To mimic the gecko adhesion system to its full extent, several of the fabrication processes presented above may be repeated to obtain hierarchical structures with controlled tip geometry. A current overview of fabrication methods to obtain patterned surfaces is given in the recently published book by del Campo and Arzt [30]. A review of recently fabricated gecko-like structures with high adhesion was published by Samento and Menon in 2010 [31]. The publication by Hu and Xia gives a nice overview about rational design of gecko-inspired surfaces, considering fabrication processes, materials selection, and concepts for switchability [32].

Although the fabrication of gecko-inspired adhesives is already well advanced in the laboratory, there is still plenty of room to improve the performance of such artificial dry adhesives. For example, experimental studies of the adhesion performance on rough surfaces – ideal territory for a gecko – are rare. In addition, surface specimens with high adhesion seldom exceeded several square centimeters in area, and the methods used are not feasible for large-scale fabrication. To establish a product based on gecko-inspired adhesives, new processes are now being developed which allow fabrication of surfaces with highly controlled structures in a cost-efficient way.



Gecko Adhesion, Fig. 7 Gecko-inspired adhesives, fabricated to achieve different structure geometry. (**a**–**d**) Symmetric structures [17–20], (**e**–**g**) asymmetric structures [21–23], and (**h**–**m**) hierarchical structures [24–29] (Reprints with kind permission of Nature Publishing Group (**a**), Elsevier (**b**, **h**), American Institute of Physics (**c**), The Royal Society Publishing (**d**), American Chemical Society (**e**, **l**), WILEY-VCH (**f**, **k**), IOP Publishing Ltd. (**g**, **j**), National Academy of Science, USA (**i**, **m**))

Applications

Possible applications of gecko-inspired adhesives can be identified by exploiting their key properties, i.e., high reversible adhesion, easy detachment, residue-free contact, and self-cleaning properties. Such adhesives are being implemented into a locomotion system for robots enabling them to overcome obstacles. Such a robot would be able to access terrain which would be difficult or dangerous for humans. Another related application would be the development of "gecko-grabbers," e.g., in fabrication plants. It is likely that gecko adhesives could find a large field of application in household or sports articles, e.g., gecko-fastening strips or goalkeeper gloves.

A true breakthrough for gecko-inspired adhesives, however, would be in the class of "intelligent materials." These new kind of adhesives would not only have the function of high reversible (and switchable) adhesion, but would combine these properties with other functions. For example, gecko-inspired adhesives could be designed in such a way that they adhere *selectively* to different materials, allow a reversible adhesion to glass, and serve as an optical coupling at the same time, or stick to skin and decompose with time, which could replace conventional plasters.

Research over the last decade has laid the foundations for such ideas; now technological development must follow to make them come true.

Cross-References

- ▶ Bioadhesion
- ▶ Bioadhesives
- ▶ Biomimetics
- ▶ Biomimetics of Marine Adhesives
- ▶ Disjoining Pressure and Capillary Adhesion
- ▶ Gecko Effect

References

1. Hsu, P.Y., Ge, L., Li, X., Stark, A.Y., Wesdemiotis, C., Niewiarowski, P.H., Dhinojwala, A.: Direct evidence of phospholipids in gecko footprints and spatula-substrate contact interface detected using surface-sensitive spectroscopy. J. R. Soc. Interface **9**, 657–664 (2012)
2. Hiller, U., Blaschke, R.: Zum Haftproblem der Gecko-Füsse. Naturwissenschaften **54**, 344–345 (1967)
3. Stork, N.E.: Experimental analysis of adhesion of *Chrysolina polita* (Chrysomelidae: Coleoptera) on a variety of surfaces. J. Exp. Biol. **88**, 91–107 (1980)
4. Autumn, K., Gravish, N.: Gecko adhesion: evolutionary nanotechnology. Phil. Trans. R. Soc. A. **366**, 1575–1590 (2008)
5. Autumn, K., Liang, Y.A., Hsieh, S.T., Zesch, W., Chan, W.P., Kenny, T.W., Fearing, R., Full, R.J.: Adhesive force of a single gecko foot-hair. Nature **405**, 681–685 (2000)
6. Gorb, S.N.: Functional Surfaces in Biology 2: Adhesion Related Phenomena, vol. 2. Springer, Heidelberg (2010)
7. Huber, G., Mantz, H., Spolenak, R., Mecke, K., Jacobs, K., Gorb, S.N., Arzt, E.: Evidence for capillarity contributions to gecko adhesion from single spatula nanomechanical measurements. Proc. Natl. Acad. Sci. U. S. A. **102**, 16293–16296 (2005)
8. Puthoff, J.B., Prowse, M.S., Wilkinson, M., Autumn, K.: Changes in materials properties explain the effects of humidity on gecko adhesion. J. Exp. Biol. **213**, 3699–3704 (2010)
9. Johnson, K.L., Kendall, K., Roberts, A.D.: Surface energy and the contact of elastic solids. Proc. R. Soc. Lond. A. **324**, 301–313 (1971)
10. Johnson, K.L.: Contact Mechanics. Cambridge University Press, Cambridge (1987)
11. Maugis, D.: Contact, Adhesion and Rupture of Elastic Solids. Springer, Heidelberg (2000)
12. Arzt, E., Gorb, S., Spolenak, R.: From micro to nano contacts in biological attachment devices. Proc. Natl. Acad. Sci. U. S. A. **100**, 10603–10606 (2003)
13. Federle, W.: Why are so many adhesive pads hairy? J. Exp. Biol. **219**, 2611–2621 (2006)
14. Kamperman, M., Kroner, E., del Campo, A., McMeeking, R.M., Arzt, E.: Functional adhesive surfaces with "Gecko Effect": the concept of contact splitting. Adv. Eng. Mater. **12**, 335–348 (2010)
15. Spolenak, R., Gorb, S., Arzt, E.: Adhesion design maps for bio-inspired attachment systems. Acta Biomater. **1**, 5–13 (2005)
16. Kim, T.W., Bhushan, B.: Effect of stiffness of multi-level hierarchical attachment system on adhesion enhancement. Ultramicroscopy **107**, 902–912 (2007)
17. Geim, A., Dubonos, S., Grigorieva, I., Novoselov, K., Zhukov, A., Shapoval, S.: Microfabricated adhesive mimicking gecko foot-hair. Nat. Mater. **2**(7), 461–463 (2003)
18. Davies, J., Haq, S., Hawke, T., Sargent, J.P.: A practical approach to the development of a synthetic Gecko tape. Int. J. Adhes. Adhes. **29**, 380–390 (2009)
19. Kim, S., Sitti, M.: Biologically inspired polymer microfibers with spatulate tips as repeatable fibrillar adhesives. Appl. Phys. Lett. **89**, 261911 (2006)
20. Varenberg, M., Gorb, S.: Close-up of mushroom-shaped fibrillar adhesive microstructure: contact element behaviour. J. R. Soc. Interface **5**, 785–789 (2008)
21. Del Campo, A., Greiner, C., Arzt, E.: Contact shape controls adhesion of bioinspired fibrillar surfaces. Langmuir **23**, 10235–10243 (2007)
22. Murphy, M.P., Aksak, B., Sitti, M.: Gecko-inspired directional and controllable adhesion. Small **5**(2), 170–175 (2009)
23. Sameoto, D., Menon, C.: Direct molding of dry adhesives with anisotropic peel strength using an offset lift-off photoresist mold. J. Micromech. Microeng. **19**, 115026 (2009)
24. Northen, M.T., Turner, K.L.: Meso-scale adhesion testing of integrated micro- and nano-scale structures. Sensor. Actuator. A. **130–131**, 583–587 (2006)
25. Ge, L., Sethi, S., Ci, L., Ajayan, P.M., Dhinojwala, A.: Carbon nanotube-based synthetic gecko tapes. Proc. Natl. Acad. Sci. U. S. A. **104**(26), 10792–10795 (2007)
26. Kustadi, T.S., Samper, V.D., Ng, W.S., Chong, A.S., Gao, H.: Fabrication of a gecko-like hierarchical fibril array using a bonded porous alumina template. J. Micromech. Microeng. **17**, N75–N81 (2007)
27. Del Campo, A., Arzt, E.: Design parameters and current fabrication approaches for developing bioinspired dry adhesives. Macromol. Biosci. **7**, 118–127 (2007)
28. Murphy, M.P., Kim, S., Sitti, M.: Enhanced adhesion by Gecko-inspired hierarchical fibrillar adhesives. Appl. Mater. Interface **1**(4), 849–855 (2009)
29. Jeong, H.E., Lee, J.-K., Kim, H.N., Moon, S.H., Suh, K.Y.: A nontransferring dry adhesive with hierarchical polymer nanohairs. Proc. Natl. Acad. Sci. U. S. A. **106**(14), 5639–5644 (2009)
30. del Campo, A., Arzt, E.: Generating Micro and Nanopatterns on Polymeric Materials. Wiley-VCH, Weinheim (2011)

31. Samento, D., Menon, C.: Recent advances in the fabrication and adhesion testing of biomimetic dry adhesives. Smart. Mater. Struct. **19**, 103001 (2010)
32. Hu, S., Xia, Z.: Rational design and nanofabrication of gecko-inspired fibrillar adhesives. Small **8**, 2464–2468 (2012)

Gecko Effect

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Gecko feet; Reversible adhesion

Definition

Geckos have the largest mass and have developed the most complex hairy attachment structures among climbing animals that are capable of smart adhesion – the ability to cling to different smooth and rough surfaces and detach at will. These animals make use of about three million microscale hairs (setae) (about 14,000/mm^2) that branch off into nanoscale spatulae, about three billion spatulae on two feet. The so-called division of contacts provides high dry adhesion. Multiple-level hierarchically structured surface construction provides the gecko with the compliance and adaptability to create a large real area of contact with a variety of surfaces.

Overview

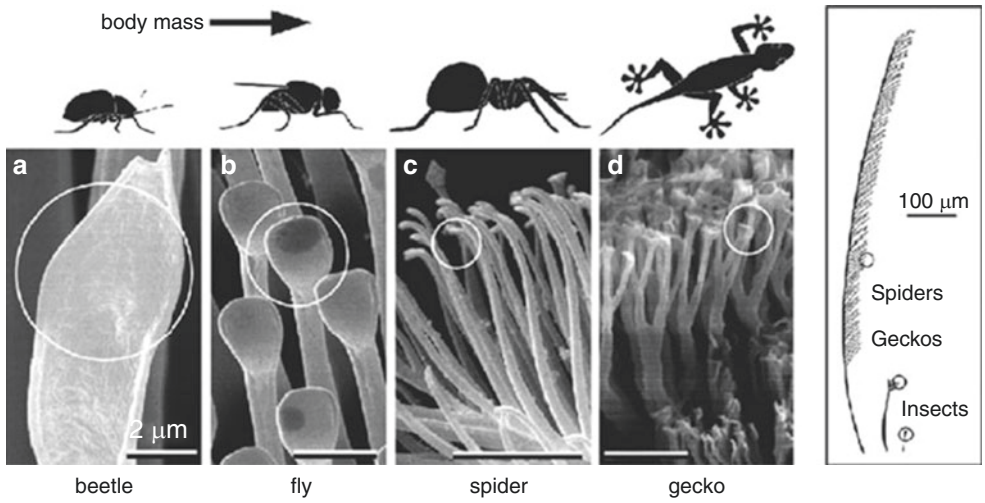
The leg attachment pads of several animals, including many insects, spiders, and lizards, are capable of attaching to and detaching from a variety of surfaces and are used for locomotion, even on vertical walls or across the ceiling [1–4]. Biological evolution over a long period of time has led to the optimization of their leg attachment systems. This dynamic attachment ability is referred to as reversible adhesion or smart adhesion [5]. Many insects (e.g., beetles and flies) and spiders have been the subject of investigation. However, the attachment pads of geckos have been the most widely studied due to the fact that they have the highest body mass and exhibit the most versatile and effective adhesive known in nature.

Although there are over 1,000 species of geckos [6, 7] that have attachment pads of varying morphology [8], the Tokay gecko (*Gekko gecko*) which is native to southeast Asia, has been the main focus of scientific research [9–11]. The Tokay gecko is the second largest gecko species, attaining lengths of approximately 0.3–0.4 m and 0.2–0.3 m for males and females, respectively. They have a distinctive blue or gray body with orange or red spots and can weigh up to 300 g [12]. These have been the most widely investigated species of gecko due to the availability and size of these creatures.

There is great interest among the scientific community to further study the characteristics of gecko feet in the hope that this information could be applied to the production of micro/nanosurfaces capable of recreating the adhesion forces generated by these lizards [2–4]. Common man-made adhesives such as tape or glue involve the use of wet adhesives that permanently attach two surfaces. However, replication of the characteristics of gecko feet would enable the development of a superadhesive tape capable of clean, dry adhesion. These structures can bind components in microelectronics without the high heat associated with various soldering processes. These structures will never dry out in a vacuum – a common problem in aerospace applications. They have the potential for use in everyday objects such as adhesive tapes, fasteners, and toys and in high technology such as microelectronic and space applications. Replication of the dynamic climbing and peeling ability of geckos could find use in the treads of wall-climbing robots.

Tokay Gecko

There are two kinds of attachment pads – relatively smooth and hairy. Relatively smooth



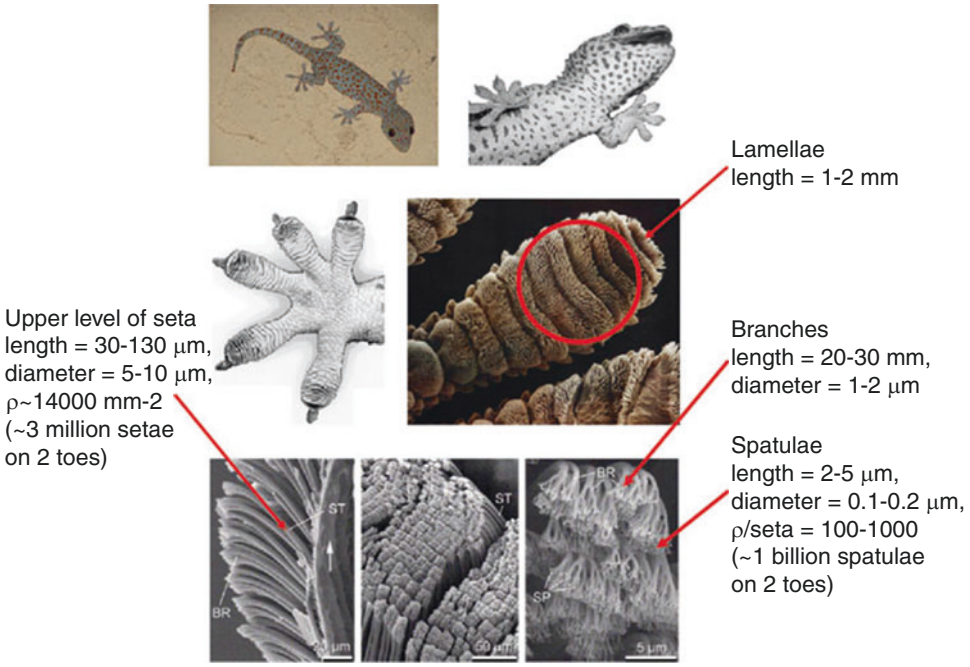
Gecko Effect, Fig. 1 Terminal elements of the hairy attachment pads of a beetle, fly, spider, and gecko shown at two different scales [13]

pads, so-called arolia and euplantulae, are soft and deformable and are found in tree frogs, cockroaches, grasshoppers, and bugs. The hairy types consist of long deformable setae and are found in many insects (e.g., beetles, flies), spiders, and lizards. The microstructures utilized by beetles, flies, spiders, and geckos have similar structures and can be seen in Fig. 1. As the size (mass) of the creature increases, the radius of the terminal attachment elements decreases. This allows a greater number of setae to be packed into an area, hence increasing the linear dimension of contact and the adhesion strength [13, 14].

The explanation for the adhesion properties of gecko feet can be found in the surface morphology of the skin on the toes of the gecko. The skin is comprised of a complex hierarchical structure of lamellae, setae, branches, and spatulae [8]. Figure 2 shows various SEM micrographs of a gecko foot, showing the hierarchical structure down to the nanoscale. Figure 3 shows a schematic of the structure, and Table 1 summarizes the surface characteristics. The gecko attachment system consists of an intricate hierarchy of structures beginning with lamellae, soft ridges 1–2 mm in length [8] that are located on the attachment pads (toes) that compress easily so that contact can be made with rough, bumpy surfaces. Tiny curved hairs known as setae extend from the lamellae with a density of approximately 14,000 per square millimeter [22]. These setae are typically 30–130 μm in length, 5–10 μm in diameter [8, 9, 19, 20], and are composed primarily of β-keratin [17, 25] with some α-keratin component [26]. At the end of each seta, 100–1,000 spatulae [8, 9] with typically 2–5 μm in length and a diameter of 0.1–0.2 μm [8] branch out and form the points of contact with the surface. The tips of the spatulae are approximately 0.2–0.3 μm in width [8], 0.5 μm in length, and 0.01 μm in thickness [21] and garner their name from their resemblance to a spatula.

The attachment pads on two feet of the Tokay gecko have an area of about 220 mm^2. About three million setae on their toes that branch off into about three billion spatulae on two feet can produce a clinging ability of about 20 N required to pull a lizard down a nearly vertical (85°) surface) [10] and allow them to climb vertical surfaces at speeds over 1 m/s with the capability to attach and detach their toes in milliseconds. In isolated setae, a 2.5 μN preload yielded adhesion of 20–40 μN, and thus the adhesion coefficient, which represents the strength of adhesion as a function of preload, ranges from 8 to 16 [27].

Typical rough, rigid surfaces are able to make intimate contact with a mating surface only over a very small portion of the perceived apparent area



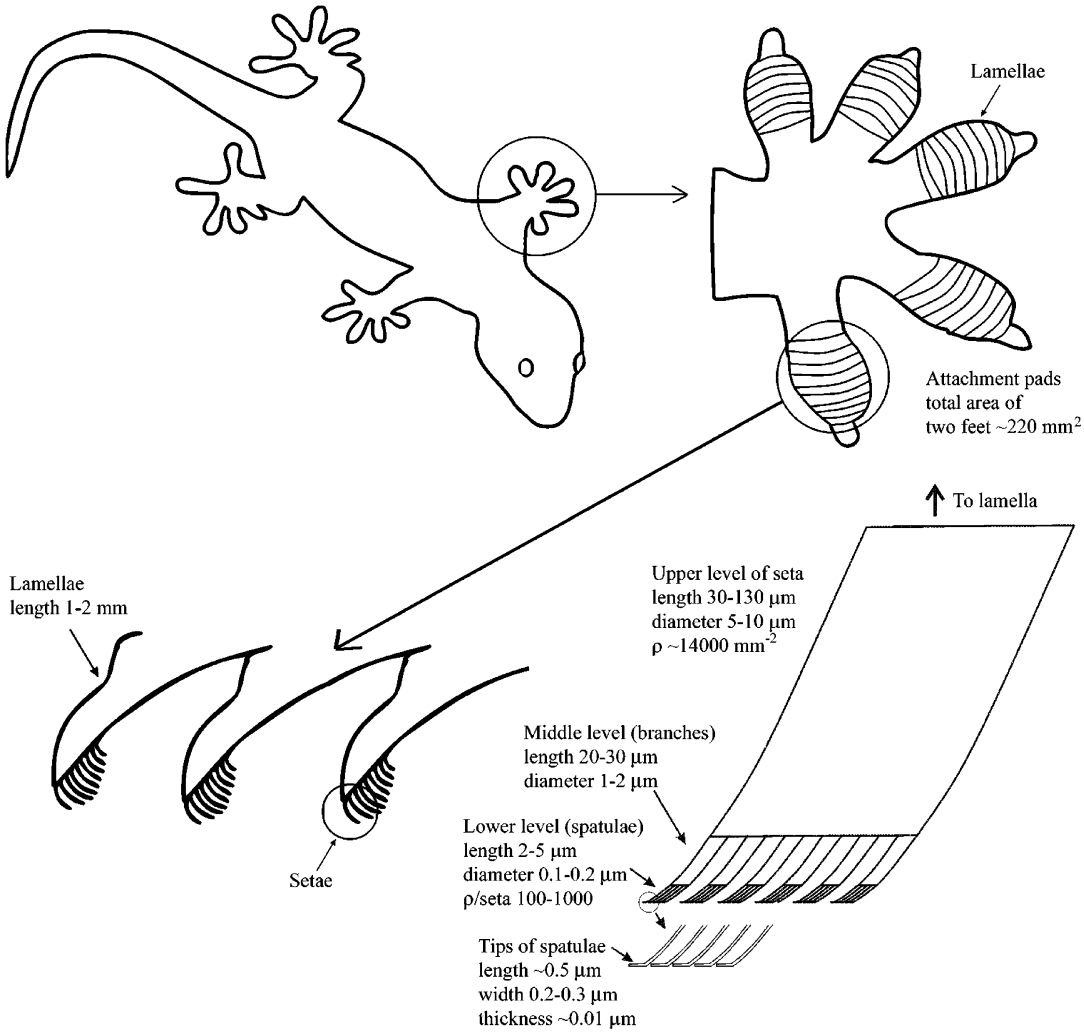
Gecko Effect, Fig. 2 The hierarchical structures of a Tokay gecko foot: a gecko foot [15] and a gecko toe [11]. Two feet contain about three million setae on their toes that branch off into about three billion spatulae. Scanning electron microscope (*SEM*) micrographs of the setae and the spatulae in the bottom row [16]. *ST* seta, *SP* spatula, *BR* branch

of contact. In fact, the real area of contact is typically two to six orders of magnitude less than the apparent area of contact [4, 28, 29]. Autumn et al. [27] proposed that divided contacts serve as a means for increasing adhesion. The surface energy approach can be used to calculate adhesion force in the dry environment in order to calculate the effect of division of contacts. Adhesion force of a single contact is proportional to a linear dimension of the contact. For a constant area divided into a large number of contacts or setae, the adhesion force increases linearly with the square root of the number of contacts (self-similar scaling) [13].

Attachment Mechanisms

When asperities of two solid surfaces are brought into contact with each other, chemical and/or physical attractions occur. The force developed that holds the two surfaces together is known as adhesion. In a broad sense, adhesion is considered to be either physical or chemical in nature [3, 4, 28–34]. Chemical interactions such as electrostatic attraction charges [35] as well as intermolecular forces [9] including van der Waals and capillary forces have all been proposed as potential adhesion mechanisms in gecko feet. Others have hypothesized that geckos adhere to surfaces through the secretion of sticky fluids [36, 37], suction [37], increased frictional force [38], and microinterlocking [39].

Through experimental testing and observations conducted over the last century and a half many potential adhesive mechanisms have been eliminated. Observation has shown that geckos lack any glands capable of producing sticky fluids [36, 37], thus ruling out the secretion of sticky fluids as a potential adhesive mechanism. Furthermore, geckos are able to create large adhesive forces normal to a surface. Since friction only acts parallel to a surface, the attachment mechanism of increased frictional force has been ruled out. Dellit [39] experimentally ruled out suction and electrostatic attraction as potential adhesive mechanisms. Experiments carried out in vacuum did not show a difference between the adhesive



Gecko Effect, Fig. 3 Schematic of a Tokay gecko including the overall body, one foot, a cross-sectional view of the lamellae, and an individual seta. ρ represents number of spatulae

forces at low pressures compared to ambient conditions. Since adhesive forces generated during suction are based on pressure differentials, which are insignificant under vacuum, suction was rejected as an adhesive mechanism [39]. Additional testing utilized X-ray bombardment to create ionized air in which electrostatic attraction charges would be eliminated. It was determined that geckos were still able to adhere to surfaces in these conditions, and therefore, electrostatic charges could not be the sole cause of attraction [39]. Autumn et al. [15] demonstrated the ability of a gecko to generate large adhesive forces when in contact with a molecularly smooth SiO_2 microelectromechanical system (MEMS) semiconductor. Since surface roughness is necessary for microinterlocking to occur, it has been ruled out as a mechanism of adhesion. Two mechanisms, van der Waals forces and capillary forces, remain as the potential sources of gecko adhesion.

Multilevel Hierarchy for Adaptability to Rough Surfaces

In order to study the effect of the number of hierarchical levels in the attachment system on attachment ability, models with one [5, 40, 41],

Gecko Effect, Table 1 Surface characteristics of Tokay gecko feet (Young's modulus of surface material, keratin = 1–20 GPa[a,b])

Component	Size	Density	Adhesive force
Seta	30–130[c–f]/ 5–10[c–f] length/ diameter (μm)	~14,000[h,i] setae/mm^2	194 μN[j] (in shear) ~20 μN[j] (normal)
Branch	20–30[c]/1–2[c] length/ diameter (μm)	–	–
Spatula	2–5[c]/ 0.1–0.2[c,g] length/ diameter (μm)	100–1000[c, d] spatulae per seta	–
Tip of spatula	~0.5[c,g]/ 0.2–0.3[c,f]/ ~0.01[g] length/ width/ thickness (μm)	–	11 nN[k] (normal)

[a]Russell [17]
[b]Bertram and Gosline [18]
[c]Ruibal and Ernst [8]
[d]Hiller [9]
[e]Russell [19]
[f]Williams and Peterson [20]
[g]Persson and Gorb [21]
[h]Schleich and Kästle [22]
[i]Autumn and Peattie [23]
[j]Autumn et al. [15]
[k]Huber et al. [24]

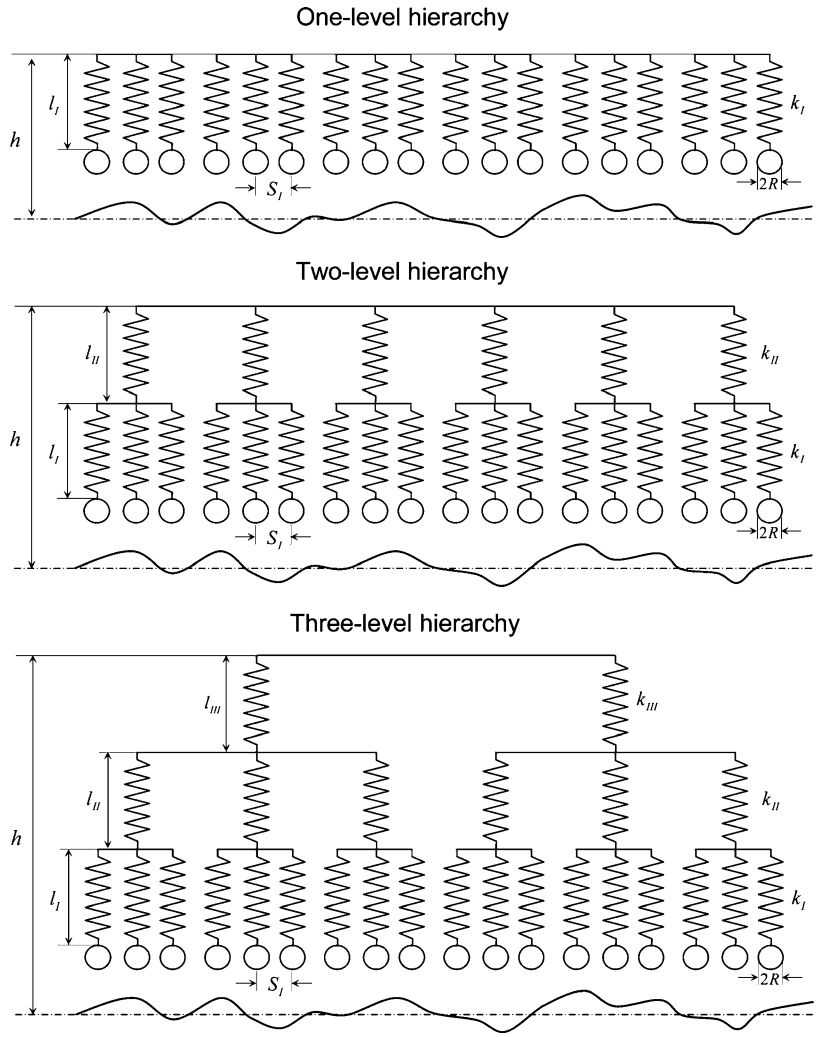
two [5, 40, 41], and three [40, 41] levels of hierarchy were simulated (Fig. 4). The random rough surfaces used in the simulations were generated by a computer program [28, 29]. Figure 5 shows the calculated spring force–distance curves for the one-, two- and three-level hierarchical models in contact with rough surfaces of different values of root mean square (RMS) amplitude σ ranging from $\sigma = 0.01$ μm to $\sigma = 10$ μm at an applied load of 1.6 μN which was derived from the gecko's weight. When the spring model is pressed against the rough surface, contact between the spring and the rough surface occurs at point A; as the spring tip presses into the contacting surface, the force increases up to point B, B′, or B″. During pull off, the spring relaxes, and the spring force passes an equilibrium state (0 N); tips break free of adhesion forces at point C, C′, or C″ as the spring moves away from the surface. The perpendicular distance from C, C′, or C″ to zero is the adhesion force. Adhesion energy stored during contact can be obtained by calculating the area of the triangle during the unloading part of the curves. The data in Fig. 5 show that adhesion forces for various levels of hierarchy are comparable when in contact with the smooth surface. However, the adhesion force for the three-level hierarchy is the largest when in contact with rough surface. Thus, multilevel hierarchy facilitates adaptability to a variety of rough surfaces.

Peeling

Although geckos are capable of producing large adhesion forces, they retain the ability to remove their feet from an attachment surface at will by peeling action. The orientation of the spatulae facilitates peeling. Autumn et al. [15] were the first to experimentally show that the adhesion force of gecko setae is dependent on the three-dimensional orientation as well as the preload applied during attachment. Due to this fact, geckos have developed a complex foot motion during walking. First, the toes are carefully uncurled during detachment. The maximum adhesion occurs at an attachment angle of 30° – the angle between a seta and mating surface. The gecko is then able to peel its foot from surfaces one row of setae at a time by changing the angle at which its setae contact a surface. At an attachment angle greater than 30°, the gecko will detach from the surface.

Self-Cleaning

Natural contaminants (dirt and dust) as well as man-made pollutants are unavoidable and have the potential to interfere with the clinging ability of geckos. Particles found in the air consist of particulates that are typically less than 10 μm in diameter while those found on the ground can often be larger [42, 43]. Intuitively, it seems that the great adhesion strength of gecko feet would cause dust and other particles to become trapped in the spatulae and that they would have no way of being removed without some sort of manual

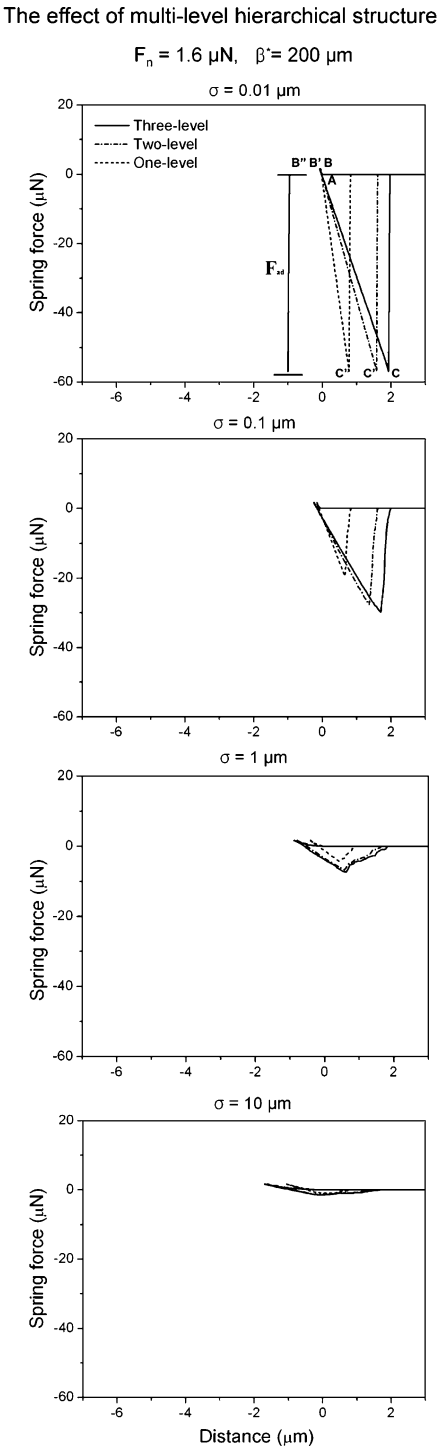


Gecko Effect, Fig. 4 One-, two-, and three-level hierarchical spring models for simulating the effect of hierarchical morphology on interaction of a seta with a rough surface. In this figure, *lI, II, III* are lengths of structures, *sI* is space between spatulae, *kI, II, III* are stiffnesses of structures, *I, II*, and *III* are level indexes, *R* is radius of tip, and *h* is distance between upper spring base of each model and mean line of the rough profile [40]

cleaning action on behalf of the gecko. However, geckos are not known to groom their feet like beetles [44], nor do they secrete sticky fluids to remove adhering particles like ants [45] and tree frogs [46], yet they retain adhesive properties. One potential source of cleaning is during the time when the lizards undergo molting, or the shedding of the superficial layer of epidermal cells. However, this process only occurs approximately once per month [47]. If molting were the sole source of cleaning, the gecko would rapidly lose its adhesive properties as it was exposed to contaminants in nature [48]. They reported that gecko setae become cleaner with repeated use – a phenomena known as self-cleaning.

Fabrication of Biomimetic Gecko Skin

Based on the studies reported in the literature, the dominant adhesion mechanism utilized by gecko and some spider attachment systems appears to be van der Waals forces. The hierarchical structure involving complex divisions of the gecko skin (lamellae-setae-branches-spatulae) enable a large number of contacts between the gecko skin and mating surface. These hierarchical fibrillar microstructured surfaces would be capable of reusable dry adhesion and would have uses in a wide range of applications from everyday objects such as adhesive tapes, fasteners, toys, microelectronic and space applications, and treads



Gecko Effect, Fig. 5 Force–distance curves of one-, two-, and three-level models in contact with rough surfaces with different σ values for an applied load of 1.6 μN

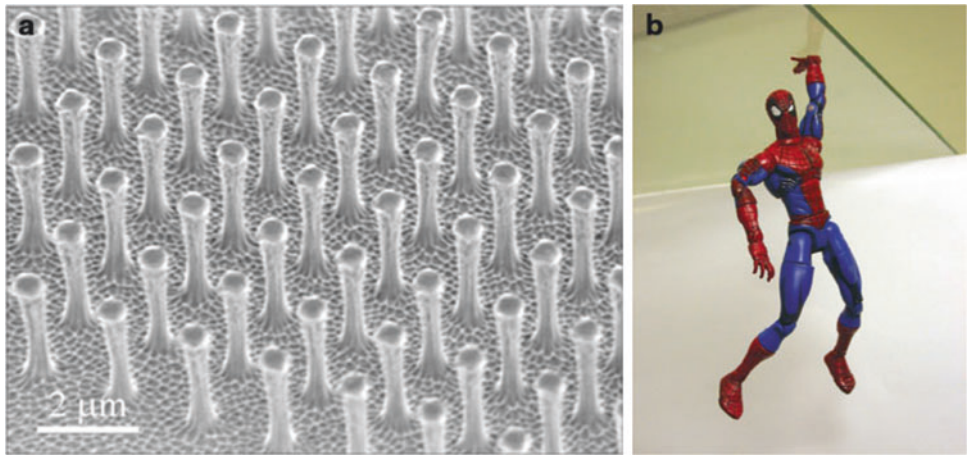
of wall-climbing robots. The development of nanofabricated surfaces capable of replicating this adhesion force developed in nature is limited by current fabrication methods. Many different techniques have been used in an attempt to create and characterize bio-inspired adhesive tapes. Attempts are being made to develop climbing robots using gecko-inspired structures [49–52].

As an example, Geim et al. [53] created arrays of polyimide nanofibers using electron-beam lithography and dry etching in oxygen plasma (Fig. 6a (left)). By using electron-beam lithography, thermal evaporation of an aluminum film, and lift-off, an array of nanoscale aluminum disks were prepared. These patterns were then transferred in the polyimide film by dry etching in oxygen plasma. A 1 cm^2 sample was able to create 3 N of adhesive force under the new arrangement. This is approximately one-third the adhesive strength of a gecko. They fabricated a Spiderman toy (about 0.4 N) with a hand covered with molded polymer nanohairs, which could support its weight Fig. 6b.

As another example, Bhushan and Sayer [54] characterized the surface roughness, friction force, and contact angle of a structured sample and compared the results to an unstructured control. The macroscale coefficient of kinetic friction of the structured sample was found to be almost four times greater than the unstructured sample. This increase was determined to be a result of the structured roughness of the sample and not the random nanoroughness. It is also noteworthy that the static and kinetic coefficients of friction were approximately equal for the structured sample. It is believed that the divided contacts allow the broken contacts of the structured sample to constantly recreate contact.

Closure

The adhesive properties of geckos and other creatures such as flies, beetles, and spiders are due to the hierarchical structures present on each creature's hairy attachment pads. Geckos have developed the most intricate adhesive structures of any of the aforementioned creatures. The attachment



Gecko Effect, Fig. 6 (**a**) An array of polyimide nanohairs and (**b**) a Spiderman toy (about 0.4 N) with a hand covered with the molded polymer nanohairs, clinging to a glass plate [53]

system consists of ridges called lamellae that are covered in microscale setae that branch off into nanoscale spatulae, about three billion spatulae on two feet. The so-called division of contacts provides high dry adhesion. Multiple level hierarchically structured surface construction plays an important role in adapting to surface roughness bringing the spatulae in close proximity with the mating surface. These structures, as well as material properties, allow the gecko to obtain a much larger real area of contact between its feet and a mating surface than is possible with a non-fibrillar material. Two feet of a Tokay gecko have about 220 mm^2 of attachment pad area on which the gecko is able to generate approximately 20 N of adhesion force. Although capable of generating high adhesion forces, a gecko is able to detach from a surface at will – an ability known as smart adhesion. Detachment is achieved by a peeling motion of the gecko's feet from a surface.

There is a great interest among the scientific community to create surfaces that replicate the adhesion strength of gecko feet. These hierarchical fibrillar microstructured surfaces would be capable of reusable dry adhesion and would have uses in a wide range of applications from everyday objects such as adhesive tapes, fasteners, toys, microelectronic and space applications, and treads of wall-climbing robots.

References

1. Gorb, S.: Attachment Devices of Insect Cuticles. Kluwer, Dordrecht (2001)
2. Bhushan, B.: Adhesion of multi-level hierarchical attachment systems in gecko feet. J. Adhes. Sci. Technol. **21**, 1213–1258 (2007)
3. Bhushan, B. (ed.): Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
4. Bhushan, B. (ed.): Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
5. Bhushan, B., Peressadko, A.G., Kim, T.W.: Adhesion analysis of two-level hierarchical morphology in natural attachment systems for "smart adhesion". J. Adhes. Sci. Technol. **20**, 1475–1491 (2006)
6. Kluge, A.G.: Gekkotan lizard taxonomy. Hamadryad **26**, 1–209 (2001)
7. Han, D., Zhou, K., Bauer, A.M.: Phylogenetic relationships among gekkotan lizards inferred from C-mos nuclear DNA sequences and a new classification of the Gekkota. Biol. J. Linn. Soc. **83**, 353–368 (2004)
8. Ruibal, R., Ernst, V.: The structure of the digital setae of lizards. J. Morphol. **117**, 271–294 (1965)
9. Hiller, U.: Untersuchungen zum Feinbau und zur Funktion der Haftborsten von Reptilien. Z. Morphol. Tiere **62**, 307–362 (1968)
10. Irschick, D.J., Austin, C.C., Petren, K., Fisher, R.N., Losos, J.B., Ellers, O.: A comparative analysis of clinging ability among pad-bearing lizards. Biol. J. Linn. Soc. **59**, 21–35 (1996)
11. Autumn, K.: How gecko toes stick. Am. J. Sci. **94**, 124–132 (2006)
12. Tinkle, D.W.: Gecko. In: Cummings, M. (ed.) Encylcopedia Americana, vol. 12. Grolier, Norwich (1992), 359 p

13. Arzt, E., Gorb, S., Spolenak, R.: From micro to nano contacts in biological attachment devices. Proc. Natl. Acad. Sci. U. S. A. **100**, 10603–10606 (2003)
14. Federle, W.: Why are so many adhesive pads hairy? J. Exp. Biol. **209**, 2611–2621 (2006)
15. Autumn, K., Liang, Y.A., Hsieh, S.T., Zesch, W., Chan, W.P., Kenny, T.W., Fearing, R., Full, R.J.: Adhesive force of a single gecko foot-hair. Nature **405**, 681–685 (2000)
16. Gao, H., Wang, X., Yao, H., Gorb, S., Arzt, E.: Mechanics of hierarchical adhesion structures of geckos. Mech. Mater. **37**, 275–285 (2005)
17. Russell, A.P.: The morphological basis of weight-bearing in the scansors of the tokay gecko. Can. J. Zool. **64**, 948–955 (1986)
18. Bertram, J.E.A., Gosline, J.M.: Functional design of horse hoof keratin: the modulation of mechanical properties through hydration effects. J. Exp. Biol. **130**, 121–136 (1987)
19. Russell, A.P.: A contribution to the functional morphology of the foot of the tokay, *Gekko gecko*. J. Zool. (Lond.) **176**, 437–476 (1975)
20. Williams, E.E., Peterson, J.A.: Convergent and alternative designs in the digital adhesive pads of scincid lizards. Science **215**, 1509–1511 (1982)
21. Persson, B.N.J., Gorb, S.: The effect of surface roughness on the adhesion of elastic plates with application to biological systems. J. Chem. Phys. **119**, 11437–11444 (2003)
22. Schleich, H.H., Kästle, W.: Ultrastrukturen an Gecko-Zehen. Amphibia-Reptilia **7**, 141–166 (1986)
23. Autumn, K., Peattie, A.M.: Mechanisms of adhesion in geckos. Integr. Comp. Biol. **42**, 1081–1090 (2002)
24. Huber, G., Gorb, S.N., Spolenak, R., Arzt, E.: Resolving the nanoscale adhesion of individual gecko spatulae by atomic force microscopy. Biol. Lett. **1**, 2–4 (2005)
25. Maderson, P.F.A.: Keratinized epidermal derivatives as an aid to climbing in gekkonid lizards. Nature **2003**, 780–781 (1964)
26. Rizzo, N., Gardner, K., Walls, D., Keiper-Hrynko, N., Hallahan, D.: Characterization of the structure and composition of gecko adhesive setae. J. R. Soc. Interface **3**, 441–451 (2006)
27. Autumn, K., Sitti, M., Liang, Y.A., Peattie, A.M., Hansen, W.R., Sponberg, S., Kenny, T.W., Fearing, R., Israelachvili, J.N., Full, R.J.: Evidence for van der Waals adhesion in gecko setae. Proc. Natl. Acad. Sci. U. S. A. **99**, 12252–12256 (2002)
28. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
29. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
30. Bikerman, J.J.: The Science of Adhesive Joints. Academic, New York (1961)
31. Zisman, W.A.: Influence of constitution on adhesion. Ind. Eng. Chem. **55**(10), 18–38 (1963)
32. Houwink, R., Salomon, G.: Effect of contamination on the adhesion of metallic couples in ultra high vacuum. J. Appl. Phys. **38**, 1896–1904 (1967)
33. Israelachvili, J.N.: Intermolecular and Surface Forces, 2nd edn. Academic, San Diego (1992)
34. Bhushan, B.: Tribology and Mechanics of Magnetic Storage Devices, 2nd edn. Springer, New York (1996)
35. Schmidt, H.R.: Zur anatomie und physiologie der geckopfote. Jena. Z. Naturwiss **39**, 551 (1904)
36. Wagler, J.: Naturliches System der Amphibien. J. G. Cotta'schen Buchhandlung, Munich (1830)
37. Simmermacher, G.: Untersuchungen uber haftapparate an tarsalgliedern von insekten. Z. Wiss. Zool. **40**, 481–556 (1884)
38. Hora, S.L.: The adhesive apparatus on the toes of certain geckos and tree frogs. J. Asiat. Soc. Beng. **9**, 137–145 (1923)
39. Dellit, W.D.: Zur anatomie und physiologie der Geckozehe. Jena. Z. Naturwiss **68**, 613–658 (1934)
40. Kim, T.W., Bhushan, B.: The adhesion analysis of multi-level hierarchical attachment system contacting with a rough surface. J. Adhes. Sci. Technol. **21**, 1–20 (2007)
41. Kim, T.W., Bhushan, B.: Effect of stiffness of multi-level hierarchical attachment system on adhesion enhancement. Ultramicroscopy **107**, 902–912 (2007)
42. Hinds, W.C.: Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles. Wiley, New York (1982)
43. Jaenicke, R.: Atmospheric aerosol size distribution. In: Harrison, R.M., van Grieke, R. (eds.) Atmospheric Particles, pp. 1–29. Wiley, New York (1998)
44. Stork, N.E.: A comparison of the adhesive setae on the feet of lizards and arthropods. J. Nat. Hist. **17**, 829–835 (1983)
45. Federle, W., Riehle, M., Curtis, A.S.G., Full, R.J.: An integrative study of insect adhesion: mechanics of wet adhesion of pretarsal pads in ants. Integr. Comp. Biol. **42**, 1100–1106 (2002)
46. Hanna, G., Barnes, W.J.P.: Adhesion and detachment of the toe pads of tree frogs. J. Exp. Biol. **155**, 103–125 (1991)
47. Van der Kloot, W.G.: Molting. In: Cummings, M. (ed.) Encylcopedia Americana, vol. 19, pp. 336–337. Grolier, Norwich (1992)
48. Hansen, W.R., Autumn, K.: Evidence for self-cleaning in gecko setae. Proc. Natl. Acad. Sci. U. S. A. **102**, 385–389 (2005)
49. Autumn, K., Dittmore, A., Santos, D., Spenko, M., Cutkosky, M.: Frictional adhesion, a new angle on gecko attachment. J. Exp. Biol. **209**, 3569–3579 (2006)
50. Daltorio, K.A., Gorb, S., Peressadko, A., Horchler, A.D., Ritzmann, R.E., Quinn, R.D.: A robot that climbs walls using micro-structured polymer adhesive. In: Proceedings of 30th Annual Meeting of the Adhesion Society, pp. 329–331. (2007)
51. Aksak, B., Murphy, M.P., Sitti, M.: Gecko inspired micro-fibrillar adhesives for wall climbing robots on micro/nanoscale rough surfaces. In: Proceedings of 2008 I.E. Conference on Robotics and Automation, pp. 3058–3063. Pasadena (2008)

52. Cutkosky, M.R., Kim, S.: Design and fabrication of multi-materials structures for bio-inspired robots. Phil. Trans. R. Soc. A **367**, 1799–1813 (2009)
53. Geim, A.K., Dubonos, S.V., Grigorieva, I.V., Novoselov, K.S., Zhukov, A.A., Shapoval, S.Y.: Microfabricated adhesive mimicking gecko foot-hair. Nat. Mater. **2**, 461–463 (2003)
54. Bhushan, B., Sayer, R.A.: Surface characterization and friction of a bio-inspired reversible adhesive tape. Microsyst. Technol. **13**, 71–78 (2007)

Gecko Feet

▶ Gecko Effect

Gel Chemical Synthesis

▶ Sol–Gel Method

Genotoxicity of Nanoparticles

Fabrice Nesslany[1] and Laila Benameur[2]
[1]Laboratory of Genetic Toxicology, Institut Pasteur de Lille, Lille Cedex, France
[2]Department of Biomedical Engineering, McGill University, Montreal, QC, Canada

Definition

Genotoxicity is toxicity against cell's genetic material affecting its integrity (DNA damage). It may lead to mutagenicity and carcinogenicity in some circumstances.

Context of the Genotoxicity Assessment of Nanomaterials

As previously mentioned, the genotoxic properties of nanomaterials (NMs) are often closely linked to oxidative damage to DNA and proteins caused by oxidative stress resulting from hyper production of reactive oxygen species (ROS) and reactive nitrogen species (RNS). NM-induced oxidative stress is perhaps the most broadly developed and accepted mechanism for the potential toxic activity of NPs (nanoparticles). NP-mediated ROS and RNS production mechanisms can be classified into three groups: intrinsic production, production by interaction with cell targets, and production mediated by the inflammatory reaction. The three groups share responsibility for most of the primary (direct or indirect) or secondary genotoxic effects so far observed with NMs [1].

A critical assessment of the current knowledge about NM genotoxicity revealed that some standard models of mutagenesis and genotoxicity are either not well adapted or even not suitable for the investigation of NMs, for instance, the bacterial cell wall possibly being a barrier for many NMs; accordingly, the Ames test [2] is not relevant for assessing the mutagenicity of NMs due to a high risk of false-negative results [3–5]. The bacterial tests are probably insufficient to ensure that bacterial DNA is exposed and in consequence cannot be considered as robust. In contrast, regulatory gene mutation tests in mammalian cells [6], using rodent cells (L5178Y, CHO, V79) presenting deficiencies (weak amount of detoxification enzymes, expression of p53, etc.), may overestimate observed effects and lead to an incorrect assessment. Furthermore, their nonhuman origin could also impede their relevance as rodent cell lines are considered more susceptible to give misleading positive results [7]. Therefore, the use of such gene mutation tests is highly debatable.

For in vivo genotoxicity, the standard tests recommended by regulations in a first intention target hematopoietic cells [8, 9]. Nevertheless, the bone marrow does not represent the most exposed tissue to NMs; thus, the relevance of these tests is highly questionable. In return the in vivo comet assay and the micronucleus test appear more relevant when carried out on target organs either primary exposed, for instance, the intestine or lung [10], or exposed after translocation as the liver.

Consequently, in vitro genotoxicity tests, which so far have most frequently and successfully been used, for genotoxicity testing of NMs

were the comet assay and the micronucleus test [11, 12]. As a matter of fact, the in vitro micronucleus assay was seen to be advantageous as it is capable of detecting chromosomal damage in the form of clastogenicity (e.g., triggered by ROS) and aneugenic effects such as physical disturbance of spindles/mitotic apparatus [13]. The in vitro comet test appears to be well suited to point out primary DNA lesions.

Many other technical issues may be at the origin of drawbacks making the in vitro tests on NMs poorly predictive:

- Possible interaction of NM suspensions with the test systems, particularly with the culture medium (complex medium with the presence of protein, salts, growth factors, and other nutrients).
- The use of very high levels of doses as recommended by the corresponding OECD guidelines.
- Interferences with the methods of determination of the cytotoxicity. For example, the interference does not seem to affect the enzymatic reaction directly but lies rather in the insoluble nature of MTT – formazan; therefore, it is strongly recommended to assess cytotoxicity data with at least two or more independent test systems [14].
- To be sure that the responses are qualitatively and/or quantitatively "nanospecific," it is interesting to compare both the nano and the microscale materials in their ability to induce genotoxicity.
- Influence of cell agglomeration on genotoxicity response.
- Absence of standard positive and negative NMs allowing proofing the sensitivity and the specificity of the implemented test systems.
- Because the properties of NMs are highly dependent on how they are synthesized, their detailed characterization is necessary for accurate comparison among different studies.
- Possible secondary genotoxic effects [15, 16] etc.

Many publications show that NMs may be genotoxic in in vitro test systems [4, 5, 17]. However, the question of nonspecific responses or "false-positive" results, i.e., positive in in vitro tests but negative in in vivo assays, is often raised and should ideally be addressed. For instance, Donaldson et al. [18] concluded that in vitro genotoxicity with non-biopersistent fibers is observed only at high doses that are non-physiologically and not achievable in vivo. Following these authors, care needs to be taken in in vitro genotoxicity studies to ensure that doses used are plausible. High doses used in vitro are sufficient to cause proinflammatory, cytotoxic, and even genotoxic effects, while much lower doses attained in vivo are sufficient only to induce antioxidant defenses. Furthermore, because genotoxicity studies mainly use short-term treatments, the exposure time is insufficient for biopersistence to play a role in shortening the long fibers or dissolving them. This means that even non-biopersistent long fibers exert effects because of long fiber-type effects. This is particularly obvious with non-biopersistent glass fibers, which are found to be as genotoxic in vitro as asbestos. Following Schulz et al. [19], genotoxic effects were observed in vitro for 2 SiO_2 while these results were not reproduced in vivo. Chen et al. [20] concluded that for the genotoxicity assessment of TiO_2, in vitro test systems generated more positive results than in vivo assays. The in vitro genotoxicity may be not relevant in vivo due to antioxidant defense mechanisms [21].

Based on these observations, it was deemed useful to come back to the results generated in the field of the Nanogenotox European Project (2010–2013) entitled *Safety evaluation of manufactured nanomaterials by characterisation of potential genotoxic hazard* [22]. The goal of Nanogenotox was to build a robust methodology in terms of sensitivity and specificity with alternative test(s) for determining the genotoxic hazard of manufactured NMs.

For this purpose, preliminary physicochemical characterizations of each of the tested 14 NMs (6 MWCNT, 4 TiO_2, and 4 SiO_2) were performed, and the protocol for dispersion was determined. Then, toxicokinetics, in vitro and in vivo genotoxicity were assessed. The objective of the in vitro genotoxicity studies was to produce

in vitro data on NMs using standard tests, i.e., MLA/TK test [6] and the micronucleus test [23] and modified assays (FPG-modified alkaline comet assay), using various cell lines (human pulmonary, intestinal, dermal cells) that are theoretically representative of the target organs (i.e., primary-exposed organs or organs exposed after translocation). A round-robin test was also carried out to assess the inter-lab qualitative and quantitative reproducibility on several selected MNs using the most promising in vitro assays. Like in vitro tests, standard in vivo genotoxicity tests, i.e., micronucleus test on rat bone marrow [8] and other assays (alkaline comet assay on lung, liver, spleen, kidney, colon, and alveolar macrophages), were carried out.

The overall results obtained in in vitro genotoxicity assays (see Tables 1 and 2) show that many MNs were slightly genotoxic, possibly due to indirect mechanisms not yet fully understood. However, the genotoxicity observed in some cases (e.g., SiO_2 on Caco-2 cells) was not reproducible in an independent assay (see Table 2). Furthermore, reports from Nanogenotox European Project reach the conclusion that full-thickness 3D skin models could not be systematically recommended for MNs hazard assessment of genotoxicity and that BEAS-2B cells appear to perform somewhat better than Caco-2 cells [22]. The results obtained during the round-robin test were heterogeneous as depicted in Table 3.

Regarding results obtained in in vivo genotoxicity tests, most data indicated no genotoxic potential whatever the NM investigated even if some genotoxic effects were observed in few organs during the comet assays, without any real dose–response relationship. Few effects linked to oxidation (from FPG-modified comet assay) were observed. All NMs were negative in the micronucleus test on bone marrow except after intravenous exposure with a SiO_2 at the high dose of 20 mg/kg b.w/day (X3) with a mean of micronucleated cells of 0.00183 versus 0.00060 in the corresponding negative control. However, this increase was very slight and no dose–response relationship was observed [22]. Anyway, within the same family (SiO_2, TiO_2, and CNT), both genotoxicity and toxic injuries varied according to the NM. From these conclusions, several comments and issues appear regarding the selected endpoints, the experimental procedures, and the results.

Test Systems

Strong discrepancies were noted depending on the cell type used. For instance, all TiO_2 were found positive on NHEK cells while they were all negative on 16HBE cells. These results underline the importance of the characterization of the cell lines used (origin, antioxidant status, and their capability of internalization and externalization of NMs, etc.). Beyond this mandatory point, the conditions of culture (e.g., presence of serum), the state of cells, etc., should also be controlled. A thorough physicochemical characterization of the tested NM in the used culture media should be performed. Importantly, the efficiency and kinetics of the uptake of NMs in the selected cell type should be done.

Endpoints

From the overall in vitro results, we can wonder what is the weight of an in vitro comet assay (DNA primary damage) that is positive while the in vitro micronucleus is negative; in other words, what is the predictivity of the in vitro comet assay? At the opposite, what kind of mechanisms of action (MoA) occurs when there is no induction of DNA primary damage but chromosomal aberrations? (Aneuploidy? Other events, e.g., indirect oxidation, mitochondrial disturbance, depletion of calcium, etc.?).

Is the in vitro comet assay really well adapted for the genotoxicity assessment of NMs as such, or does it need other adaptations? If performed, which duration of treatment should be recommended or what is the relevance of a long-term treatment depending on the cell type? If not performed, which second endpoint should be investigated? As neither the Ames test nor the standard gene point mutation assays on rodent cells are appropriate, there is a real need for implementing a gene point mutation test on human cell lines.

Genotoxicity of Nanoparticles, Table 1 Outcome of in vitro genotoxicity assays in pulmonary cell lines (from Nanogenotox European Project [22], Final report of Deliverable 5: In vitro testing strategy for nanomaterials)

		BEAS 2B					16 HBE					A549				
Nanomaterial		Micronucleus[a]	Comet, 3 h	Comet, 24 h	Comet FpG, 3 h	Comet FpG, 24 h	Micronucleus[b]	Comet, 3 h	Comet, 24 h	Comet FpG, 3 h	Comet FpG, 24 h	Micronucleus[c]	Comet, 3 h	Comet, 24 h	Comet FpG, 3 h	Comet FpG, 24 h
TiO_2	NM-102	–	+	+			–	+	–			+	+			
	NM-103	–	–	–			–	–	–			–	–	–		
	NM-104	–	–	–			–	–	–			–	–	–		
	NM-105	–	–	–			–	–	+			–	+	–		
SiO_2	NM-200	–	+		(+)		–	+	–	–	–	–/–	(+)	–	–	–
	NM-201	–	(+)		–		–	–	–	–	–	+/+	+	(+)	–	(+)
	NM-202	–	+		+		–	–	–	–	–	+/+	+	(+)	+	–
	NM-203	(+)	+		+		–	–	–	–	–	–/(+)	–	+	–	+
MWNT	NM-400	(+)	–	–	–	–	–	–	–			(+)	–	–		
	NM-401	+	–	–	–	–	–	–	–			–	–	–		
	NM-402	+	–	–	–	–	–	–	–			+	–	–		
	NM-403	+	–	–	–	–	–	–	–			–	–	–		
	NRCWE-006	+	–	–	–	–	–	–	–			+	–	–		
	NRCWE-007	+	–	–	–	–	–	–	–			+	–	–		

Outcome: + Positive: a statistically significant increase with ≥2 doses or a statistically significant increase at high dose and a dose-dependent increase, (+) Equivocal: a statistically significant increase with 1 dose, no dose-dependent increase, – Negative: used to separate outcome of two experiments
Gray box, not performed
[a]Treatment for 48 h, Cyt-B added at 6 h
[b]Treatment for 41 h, no Cyt-B used
[c]Treatment for 24 h (TiO_2, NRCWE-006, and NRCWE-007), Cyt-B added at 6 h

Results

Several NMs displayed a slight genotoxic activity in vitro. As the effect was weak, it is not easily reproducible and the hypothesis of indirect mechanisms was proposed.

Otherwise, results demonstrated the lack of intra- and inter-lab reproducibility as noted during the round-robin test. One of the hypotheses is that the method of dispersion may affect the size of agglomerates, the sedimentation, the cell exposure and consequently the level of cytotoxicity, and the choice of doses for genotoxicity assessment. In any case, it shows that there is a need to reiterate the assays.

Interestingly, the production of both in vitro and in vivo data on the same NMs allowed a first comparison of in vitro versus in vivo results and the estimation of the predictivity of the in vitro genotoxicity testing (see Tables 4 and 5).

Genotoxicity of Nanoparticles, Table 2 Outcome of in vitro genotoxicity assays in intestinal, lymphatic, and epidermal cell systems (from Nanogenotox European Project [22], Final report of Deliverable 5: In vitro testing strategy for nanomaterials)

Nanomaterial		Intestinal					Lymphatic		Epidermal				
		Caco-2					HuLy	L5178Y TK+/–	NHEK			Skin 3D	
		Micronucleus[a]	Comet, 3 h	Comet, 24 h	Comet FpG, 3 h	Comet FpG, 24 h	Micronucleus[b]	Mutation	Micronucleus[c]	Comet, 3 h	Comet, 24 h	Comet , 3 h	Comet , 24 h
TiO_2	NM-102	–	–	+			(+)	–	+	(+)	(+)	–	–
	NM-103	–	–	(+)			+	–	+	(+)	(+)	–	–
	NM-104	–	–	–			+	–	+	(+)	(+)	–	–
	NM-105	–	–	+			–	–	+	(+)	(+)	–	–
SiO_2	NM-200	+/–[1]	+	+	(+)	+	–	–					
	NM-201	+/–[1]	–	(+)	–	–	–	–					
	NM-202	+/–[1]	(+)	(+)	+	+	–	–					
	NM-203	+/–[1]	+	+	+	(+)	–	–					
MWCNT	NM-400	(+)	–	–	–	–	–	–					
	NM-401	+	–	–	–	–	–	–					
	NM-402	+	–	–	–	–	(+)	–					
	NM-403	(+)	–	–	–	–	+	–					
	NRCWE-006	–	–	–	–	–	+	–					
	NRCWE-007	+	–	–	–	–	–	–					

HuLy human lymphocytes
Outcome: + Positive: a statistically significant increase with ≥2 doses or a statistically significant increase at high dose and a dose-dependent increase, (+) Equivocal: a statistically significant increase with 1 dose, no dose-dependent increase, – Negative
Gray box, not performed
[a]Treatment for 52 h, Cyt-B added at 24 h
[b]Treatment for 30 h, Cyt-B added at 6 h
[c]Treatment for 54 h, Cyt-B added at 6 h
[d]Three independent experiments performed by the same laboratory yielded a positive outcome twice and a negative outcome once

Genotoxicity of Nanoparticles, Table 3 Outcome of the round-robin test (from Nanogenotox European Project [22], Final report of Deliverable 5: In vitro testing strategy for nanomaterials)

Partner	TiO_2 NM-102		SiO_2 NM-203		MWCNT MN-403		ZnO NM-110	
	Comet	CBMN	Comet	CBMN	Comet	CBMN	Comet	CBMN
Caco-2 cells								
ANSES	–	–	–	+	–	+	–	+
NRCWE	+		+	+	+	+		+
BfR	–	+	–	–	–	+	–	+
IPL	–	–	+	–	–	–	–	+
RIVM		–	–	+				+
INRS	+	–	+	–	(+)	+		+
BEAS 2B Cells								
IMB- BAS	+	+	–	+	+	+	+	+
FIOH	–	–	+	–	–	–	+[1]	–
NIOM	+	(+)	+	+	+	–	+	–
UAB	+	–	+	–	–	–	–	+
IPH	+	–	–	+	–	–	–	–
INSA	+	–	–	–	+	–	+[2]	–

Outcome: + Positive: a statistically significant increase with ≥2 doses or a statistically significant increase at high dose and a dose-dependent increase, (+) Equivocal: a statistically significant increase with 1 dose, no dose-dependent increase, – Negative
Gray box, no data
[a]In 3-h treatment (24-h treatment was negative)
[b]In 24-h treatment (3-h treatment was negative)

The comparison of in vitro and in vivo results displayed a weak correlation of in vitro/in vivo genotoxicity results. However, no comparison was possible for the spleen, liver, and kidney for which in vivo positive responses were observed as no in vitro tests were carried out on the corresponding specific cells.

This observation raises the question of the limits of these assays and the occurrence of "false-negative" or "false-positive" results. Many hypotheses could explain these differences:

In Vitro Positive While In Vivo Negative

- Time of treatment/exposure
  - Exposure at more long-term and at higher dose(s).
  - Intracellular dose of NM is important but depends on the mode of exposure/dispersion and on the presence or not of a dispersant. Cellular uptake could possibly serve as a measure of dose for comparison among experiments and test systems.
- Form in situ different from the one in vitro

  In cell cultures, many metal-based NMs induce weak increases in DNA damage that seem continuously formed when particles are internalized: It is unclear whether the same phenomenon could also occur in vivo and results in a low-level but prolonged genotoxic injury. Furthermore, the corona, i.e., the coating proteins, formed in vitro in culture media (and even after dispersion before instillation) is very different from that produced naturally whatever the route of entry [18]. The mechanism of genotoxic action may differ depending on the amount of proteins present in the test systems in vivo and in vitro [12]. Agglomerates of different sizes and shapes may also have different effects on cells.
- Characteristics of cell types

  The absence or a low exocytosis capability may increase the intracellular dose in vitro

Genotoxicity of Nanoparticles, Table 4 Correlation between in vitro genotoxicity assays in pulmonary cell lines and in vivo comet assay on lung and bronchoalveolar liquid cells

		BEAS 2B			16 HBE			A549			*In vivo*
		μnucleus	Comet 3 h	Comet 24 h	μnucleus	Comet 3 h	Comet 24 h	μnucleus	Comet 3 h	Comet 24 h	Comet **Lung / BAL**
TiO_2	NM-102	–	+	+	–	+	–	–	+	–	– / –
	NM-103	–	–	–	–	–	–	–	–	–	– / –
	NM-104	–	–	–	–	–	–	–	–	–	– / –
	NM-105	–	–	–	–	–	+	–	+	–	– / ++
SiO_2	NM-200	–	+		–	+	–	– / –	(+)	–	– / –
	NM-201	–	**(+)**		–	–	–	+ / +	+	(+)	– / –
	NM-202	–	+		–	–	–	+ / +	+	(+)	– / –
	NM-203	(+)	+		–	–	–	– / (+)	–	+	– / –
MWCNT	NM-400	(+)	–	–	–	–	–	(+)	–	–	– / –
	NM-401	+	–	–	–	–	–	–	–	–	– / +
	NM-402	+	–	–	–	–	–	+	–	–	– / –
	NRCWE-006	+	–	–	–	–	–	+	–	–	– / (+)

BAL bronchoalveolar liquid
Outcome: + Positive: a statistically significant increase with ≥2 doses or a statistically significant increase at high dose and a dose-dependent increase, (+) Equivocal: a statistically significant increase with 1 dose, no dose-dependent increase, – Negative

compared to the one observed in vivo. Otherwise, it is important to note that cell lines from a same origin or from different tissues could be more or less sensible to NM exposure due to variation of metabolic routes, receptors on the cell surface, antioxidant and DNA repair capabilities, the presence of different enzymes/hormones, etc. [24, 25], meaning that the final choice should be clearly argued. On the other side, the weak level of antioxidant systems could artefactually increase the level of genotoxic effect.

- Organ(s) not actually exposed in vivo

  An in vivo negative response could be expected when an in vitro positive response is observed on cells that are not representative of a target tissue because it is not exposed in vivo at least by the route of exposure.

In Vitro Negative While In Vivo Positive

In the same way, the characteristics of the cell types could be at the origin of such discrepancies if cells are not able of internalization hindering their exposure. In vivo effects were observed in tissues that were not specifically investigated in vitro (liver, kidney, spleen, etc.).

Finally, as in vitro tests are not able to display secondary genotoxicity, it may explain the difference between in vitro and in vivo results.

Genotoxicity of Nanoparticles, Table 5 Correlation between in vitro genotoxicity assays in intestinal and blood cell lines and in vivo comet assay on colon and blood cells

		Caco-2			HuLy	Comet in vivo	
			Comet				
		μnucleus	3 h	24 h	μnucleus	Colon	Blood
TiO_2	NM-102	–	–	+	(+)	+	–
	NM-103	–	–	(+)	+	++	–
	NM-104	–	–	–	+	+	–
	NM-105	–	–	+	–	–	–
SiO_2	NM-200	+/–	+	+	–	–	–
	NM-201	+/–	–	(+)	–	–	–
	NM-202	+/ –	(+)	(+)	–	–	–
	NM-203	+/–	+	+	–	–	–
M W C N T	NM-400	(+)	–	–	–	–	–
	NM-401	+	–	–	–	–	–
	NM-402	+	–	–	(+)	–	–
	NRCW E-006	–	–	–	+	–	–

HuLy human lymphocytes
Outcome: + Positive: a statistically significant increase with ≥2 doses or a statistically significant increase at high dose and a dose-dependent increase, (+) Equivocal: a statistically significant increase with 1 dose, no dose-dependent increase, – Negative: used to separate outcome of two experiments

Conclusions

The sensitivity and the specificity of the in vitro toxicity tests should be a priority in order to increase their predictivity and to avoid unnecessary in vivo assays. In that view, all the results and hypothesis previously mentioned should be used to improve the predictivity of the in vitro genotoxicity tests and also to ensure that the whole strategy currently implemented is really appropriate for NM assessment. Thus, improvements could be found at the menu level by questioning the current strategy based on regulatory tests but also at the recipe level by bringing methodological modifications in order to obtain more specific responses.

1. Modification of the Current Strategy Based on Regulatory Testing

   First, the in vitro micronucleus test seems very interesting as it can be performed on target cells even in other than the ones recommended in standard conditions including human cells and it was considered as the most predictive in the Nanogenotox European Project [22]. However, it is mandatory to demonstrate the absorption of NMs by the cell line used.

   Otherwise, the question of the level of predictivity of the in vitro comet assay as it is currently carried out should be raised. If performed, the experimental design should take into account all the methodological considerations or modifications that could be made to the test procedure to improve its applicability or reliability for testing NMs as described in point 2 hereafter.

   Regarding gene point mutation tests, the Ames test [2] is considered as not relevant for assessing the mutagenicity of NMs due to a high risk of false-negative results. Furthermore, the MLA/TK assay [6] appears to be poorly adapted mainly as it uses rodent cell that grows in suspension. Therefore, no standard gene point mutation assays seems appropriate. However, more subtle genetic changes can also arise in response to a genotoxin such as point mutations including base substitutions, frameshifts, or small amplifications and deletions. Thus, it was suggested that mammalian mutation assays should also be considered in a test battery, particularly as it appears that the genotoxicity of some NMs failed to be revealed by the in vitro micronucleus assay [26]. There is a real need for

implementing a gene point mutation test on human cell lines but also for developing new test systems (see point 3).

Finally, the in vivo bone marrow micronucleus assay could not be systematically recommended as it detects substances that are genotoxic at a systemic level while high dose levels in bone marrow are often unlikely. In any case, negative results are relevant only when there is a proof of exposure. It would be useful to perform other in vivo tests that are applicable for NMs (in vivo micronucleus test in the lung and colon that may be target organs).

Do the in vitro genotoxicity tests as actually performed predict the in vivo genotoxicity and/or the carcinogenesis of NMs? In fact, there are very few comparative studies to answer this question but following Roller [27], there is no correlation between the probability of a positive response in vitro and the carcinogenic potency. As concluded by the Nanogenotox European Project [22], further comparative in vivo data for evaluating the performance of in vitro genotoxicity assays are needed.

Overall, this confirms the need of adaptation of the OECD guidelines for genotoxicity. This task has been initiated during the two **last WPMN meetings** (Workshop on the Genotoxicity of Manufactured Nanomaterials), held in Ottawa (November 2013) and Paris (October 2014).

2. **Methodological considerations or modifications to improve the test systems**

For the standard or new tests deemed appropriate, there are some methodological considerations or modifications that could be made to the test procedures to improve the applicability or reliability for testing NMs. This includes the following areas:

- Adoption of standard protocols for dispersion knowing that the nature of the dispersant used directly influence the quality of the suspension. For instance, NMs are very poorly dispersed in pure water and commonly form large aggregates, whereas it is not the case when using some form of protein (e.g., bovine serum albumin) to disperse NM [18]. Furthermore, a thorough physicochemical characterization of each NM in all the used culture media should also be performed.
- Selection of appropriate cell lines: The choice of the target cells could be issued from the toxicokinetics study that should define which organ(s) is(are) actually exposed and to which extent. Knowing that, the most sensitive cell system, preferably of human origin and p53 efficient, should be used. Importantly, the efficiency and kinetics of the uptake of NMs in the selected cell type should be done.
- Dose level: If determined, the extent of exposure of the target organ could be used as a basis for calculating the dose level to be actually tested. Anyway, the highest possible dose to be tested will be dictated by the dispersion protocol and the highest concentration ensuring a stable stock suspension.
- The reproducibility of the results obtained in vitro should be checked.
- The selection of appropriate positive (and negative) controls (nano and non-nanoscale) may be of particular interest to ensure the sensibility and the specificity of the test system.
- The incompatibility with some experimental conditions should be checked, e.g., simultaneous co-exposure of NMs and cytochalasin B that inhibits endocytosis [28].

3. **Development of new genotoxicity tests**

Beyond a regulatory strategy, new in vitro tests that may be particularly suitable for testing NMs should be further developed. Indeed, there may be gaps in in vitro testing, for example, the lack of detection of genotoxicity induced by secondary mechanisms. These new in vitro assays that might be considered include genomics and proteomics as related to detection of immune system effects, oxidative stress, inflammation, and other pathways. These technologies also can be used to assess pathway interactions [13].

The implementation of high-throughput platforms investigating simultaneously several

endpoints (e.g., cytotoxicity, impedance, γ-H2AX, oxidation, apoptosis, telomere length measurement, DNA damage repair system, etc.) may be of particular interest for screening rapidly and cheaply the toxicity of NMs, for instance, in the field of Safer by Design. In the same way, the QNTR (Quantitative Nanostructure–Toxicity Relationship) tools should be developed.

In any case, the rules laid down previously should also be applied for these future methods.

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles
- ▶ Nanoparticle Cytotoxicity
- ▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity

References

1. Benameur, L.: Genotoxicity of nanoparticles. In: Encyclopedia of Nanotechnology, pp. 952–962. Springer, New York (2012)
2. OECD Guideline for the Testing of Chemicals No. 471: Bacterial reverse mutation test, 21 July 1997
3. Balasubramanyam, A.: In vitro mutagenicity assessment of aluminium oxide nanomaterials using the Salmonella/microsome assay. Toxicol. In Vitro **6**, 1871–1876 (2010)
4. Landsiedel, R.: Genotoxicity investigations on nanomaterials: methods, preparation and characterization of test material, potential artifacts and limitations – many questions, some answers. Mutat. Res. **681**, 241–258 (2009)
5. Singh, N.: NanoGenotoxicology: the DNA damaging potential of engineered nanomaterials. Biomaterials **30**, 3891–3914 (2009)
6. OECD Guideline for the Testing of Chemicals No. 476: In vitro mammalian cell gene mutation test, 21 July 1997
7. Fowler, P.: Reduction of misleading ("false") positive results in mammalian cell genotoxicity assays. I. Choice of cell type. Mutat. Res. **742**(1–2), 11–25 (2012)
8. OECD Guideline for the Testing of Chemicals No. 474: Mammalian erythrocyte micronucleus test, 21 July 1997
9. OECD Guideline for the Testing of Chemicals No. 475: Mammalian bone marrow chromosome aberration test, 21 July 1997
10. Muller, J.: Clastogenic and aneugenic effects of multi-wall carbon nanotubes in epithelial cells. Carcinogenesis **29**, 427–433 (2008)
11. Landsiedel, R.: Gene toxicity studies on titanium dioxide and zinc oxide nanomaterials used for UV-protection in cosmetic formulations. Nanotoxicology **4**, 364–381 (2010)
12. Magdolenova, Z.: Mechanisms of genotoxicity. A review of in vitro and in vivo studies with engineered nanoparticles. Nanotoxicology **3**, 233–278 (2014)
13. Pfuhler, S.: Genotoxicity of nanomaterials: refining strategies and tests for hazard identification. Environ. Mol. Mutagen. **4**, 229–239 (2013)
14. Worle-Knirsch, J.M.: Oops they did it again! Carbon nanotubes hoax scientists in viability assays. Nano Lett. **6**, 1261–1268 (2006)
15. Savolainen, K.: Risk assessment of engineered nanomaterials and nanotechnologies – a review. Toxicology **269**(2–3), 92–104 (2010)
16. Greim, H.: Genotoxicity testing of nanomaterials – conclusions. Nanotoxicology **4**, 421–4 (2010)
17. Gonzalez, L.: Genotoxicity of engineered nanomaterials: a critical review. Nanotoxicology **2**, 252–273 (2008)
18. Donaldson, K.: Possible genotoxic mechanisms of nanoparticles: criteria for improved test strategies. Nanotoxicology **4**, 414–420 (2010)
19. Schulz, M.: Genotoxic effects of nanomaterials in lung tissue in vitro and in vivo. The Toxicologist **120**, 315–316 (2011)
20. Chen, T.: Genotoxic evaluation of titanium dioxide nanoparticles *in vivo* and *in vitro*. Toxicol. Lett. **3**, 314–319 (2014)
21. Umweltbundesamt. Carcinogenicity and mutagenicity of nanoparticles – assessment of current knowledge as basis for regulation (2014)
22. Nanogenotox European Project (March, 2010 – February, 2013). http://www.nanogenotox.eu/
23. OECD Guideline for the Testing of Chemicals No. 487: In Vitro Mammalian Cell Micronucleus Test, 23 July 2010
24. Karlsson, H.L.: Copper oxide nanoparticles are highly toxic: a comparison between metal oxide nanoparticles and carbon nanotubes. Chem. Res. Toxic. **21**, 1726–1732 (2008)
25. Dusinska, M.: In vitro and in vivo toxicity test methods. Chapter 4. Critical evaluation of toxicity tests. In: Fadeel, B., Pietroiusti, A., Shvedova, A. (eds.) Adverse effects of engineered nanomaterials: exposure, toxicology and impact on human health, pp. 63–84. Elsevier, London, UK (2012)
26. Doak, S.H.: In vitro genotoxicity testing strategy for nanomaterials and the adaptation of current OECD guidelines. Mutat. Res. **745**, 104–111 (2012)

27. Roller, M.: In vitro genotoxicity data of nanomaterials compared to carcinogenic potency of inorganic substances after inhalational exposure. Mutat Res. **727**, 72–85 (2011)
28. Doak, S.H.: Confounding experimental considerations in nanogenotoxicology. Mutagenesis **24**(4), 285–93 (2009)

Gold Nanorods

Yang Lu[1], Shang Xu[1] and Jun Lou[2]
[1]Department of Mechanical and Biomedical Engineering, City University of Hong Kong, Kowloon, Hong Kong, P. R. China
[2]Department of Materials Science and NanoEngineering, Rice University, Houston, TX, USA

Synonyms

Cylindrical gold nanoparticles; Short (low aspect ratio) gold nanowires

Definition

A gold nanorod is defined as nanometer-sized gold particles that are rod shaped. These solid, cylindrical nanoparticles have uniform diameters ranging from a few nanometers to hundreds of nanometers, with aspect ratios (length/diameter) larger than 1 but typically smaller than 10. Owing to their unique physical properties, gold nanowires/nanorods have many attractive applications in several areas, such as in life science, electronic and electromechanical devices, sensors, and solar cells. In the field of nanomedicine, the typical sizes for gold nanorods were around tens of nanometers in diameters and hundreds of nanometers in lengths. The crystalline structures of these gold nanomaterials can be either polycrystals or single crystals with different orientations, depending on the growth methods and conditions.

Key Research Findings

Surface Plasmon Resonance (SPR)

One of the most remarkable features of gold nanorods (NRs) is its unique and tunable localized surface plasmon resonance (LSPR) properties. When metals are reduced to nanoscale dimensions (a few nanometers to hundreds of nanometers), interesting physical properties such as LSPR emerge due to the restrictions imposed on their electron mobility, an effect which also strongly depends on the overall shape of the metal. For nano-sized metallic particles, especially noble metal nanoparticles (such as gold, silver, platinum), they have the ability to confine resonant photons within their small particle size to induce the localized surface plasmon oscillations of conduction-band electrons. The surface plasmon is thus resonant at a specific frequency of the incident light and is called surface plasmon resonance (SPR). For spherical gold nanoparticles, the SPR properties can be described by Mie's theory [1], which has one specific SPR frequency (wavelength $\sim$ 520 nm) for strong absorption, while for nanorods, the SPR oscillates along two directions of the rods: longitudinal and transverse directions. The excitation of the surface plasmon oscillation along the short (transverse) axis induces an absorption band in the visible light region at wavelength (520–530 nm) similar to that of spherical gold nanoparticles, referred to as the transverse band. The transverse band is insensitive to the aspect ratio of the nanorods, while the excitation of the surface plasmon oscillation along the longitudinal direction induces a much stronger absorption band in the longer wavelength region, referred to as the longitudinal band. The longitudinal SPR band of gold nanorods is very sensitive to the aspect ratio of the nanorods. Slight increases in the aspect ratio greatly redshift the absorption maximum (λ_{max}, nm) of the longitudinal band from visible to the near-infrared (NIR) region with a dramatic increase in the band intensity. According to the Gans' theory [2], which is an extension of Mie's theory to treat elongated ellipsoids, the absorption maximum of the longitudinal

band is linearly dependent on the aspect ratio (R) of the nanorods by the following relationship $\lambda_{max} = 95R + 420$ [3]. This optical tunability of the nanorods in the NIR region via the varying rod aspect ratio provides great opportunities for in vivo medical applications, since the absorbance of the surrounding tissue in this region is comparatively low.

Photoluminescence Properties

Usually bulk metals (such as gold, silver, and copper) have extremely weak fluorescence due to their rapid nonradiative electron–hole recombination. However, it has been observed that for short nanorods (aspect ratio < 3), the quantum efficiency of the photoluminescence (PL) is in the range of 10^{-4} to 10^{-3}, about six orders of magnitude greater than that found in the bulk [3]. Although the luminescence quantum efficiency of gold nanorods is several orders of magnitude lower than that of fluorescent labels such as organic dyes or semiconductor nanorods, their larger absorption cross sections compensate for their low quantum efficiency, making them high-contrast imaging agents [4]. In addition, the PL wavelength maximum is found to increase linearly with increasing aspect ratio (less than 3). The quantum efficiency increases quadratically for aspect ratios below 3 and then begins to diminish thereafter. This can be understood by the fact that, as the SPR absorption shifts to longer wavelengths as the nanorod aspect ratio (AR) is increased, the extent of the overlap between the interband processes and the SPR absorption band will be constantly changing. When less than 3, increases of AR result in greater overlap responsible for an emission enhancement. At some point, such as AR < 3.5 shown in the experimental works by Eustis and El-Sayed, further increases in AR diminish the degree of overlap resulting in an emission decline.

However, recently scientists have found that the quantum efficiency of long nanorods (length >200 nm, AR > 10) rapidly increases when the lengths are larger than 200 nm and their corresponding fluorescence intensity has been significantly enhanced [5]. In fact, the actual PL responses for metallic nanoparticles are much more complicated and could be explained as multistep processes [3, 5]. In brief, the strong dependency of photon emission on the nanorod length, combined with the fact that similar enhancements are not observed for sphere nanoparticles, provided evidences that the strong surface fields of the longitudinal surface plasmon oscillations are the origin of the fluorescence enhancement. This is also supported by the polarization of the emission being along the long rod axis [3]. Recent studies on luminescence spectrum of single nanorods with aspect ratios ranging from 1 to ~3.5 showed about an order of magnitude increase in quantum yield for nanorods with longitudinal plasmon wavelengths longer than 650 nm, compared to nanospheres. Polarization-sensitive measurements further proved that the emission is dominated by a polarized contribution that coincides with the longitudinal plasmon [4]. The excitation of interband absorption simultaneously excites the longitudinal surface plasmon of gold nanorods, resulting in enhanced local electric fields amplifying both exciting and emission fields. The enhanced fluorescence of gold nanorods thus provides opportunities for imaging and diagnostics applications.

Photothermal Properties

Unlike aforementioned radiative properties, photothermal properties of gold nanorods are nonradiative processes, in which the absorbed light is converted into heat via a series of photophysical processes [6]. Basically, these photophysical processes begin with a fast loss of phase of the coherently excited electrons on the few femtosecond timescale and then followed by energy loss to the phonon bath in the crystal lattice of gold nanorods. The relaxation process is size and shape independent and also independent for both the transverse and longitudinal surface plasmon in the nanorods. Meanwhile, the lattice cools down by passing heat to the surrounding medium via phonon–phonon relaxation of ~100 ps [3], which leads to the competitive process of cooling and heating of the whole nanorods. If the heating rate is much faster than the cooling

rate, at appropriate timescale, heat can be accumulated within the lattice causing a local temperature rise of the nanorods in a short time, which could provide an effective means for selective photothermal therapy.

Synthesis and Surface Modifications

Although the synthesis of spherical gold nanoparticles can be dated back to 1857 by Faraday, fabrication of gold nanorods with well-controlled shape and crystalline structure emerged only during past decade [3]. Five typical gold nanorod/nanowire synthesis approaches will be described here: (1) seed-mediated growth [7, 8], (2) electrochemical synthesis [9], (3) photochemical reduction [10], (4) lithographic methods [11, 12], and (5) template-assisted growth [13].

Seed-Mediated Growth

In 2001, Murphy and coworkers [7] prepared colloidal gold nanorods by the addition of citrate-capped small gold nanospheres to a bulk $HAuCl_2$ growth solution obtained by the reduction of $HAuCl_4$ with ascorbic acid in the presence of cetyltrimethylammonium bromide (CTAB) surfactant and silver ions. This method was improved by the same group with a three-step procedure in the absence of silver nitrate for larger aspect-ratio (up to 25) nanorods. In a typical procedure, the gold nanospheres/nanorods in the previous step are sequentially used as "seeds" for the next step growth. Later on, El-Sayed and coworkers further improved the method by replacing sodium citrate with a stronger CTAB stabilizer in the seed formation process and utilizing silver ions to control the aspect ratio of gold nanorods [8]. These modifications resulted in reproducible formation of nanorods with aspect ratios ranging from 1.5 to 4.5 with desired lengths. To grow longer nanorods with aspect ratios ranging from 4.6 to 10, a binary surfactant mixture composed of benzyldimethylhexadecylammonium chloride (BDAC) and CTAB was used, with appropriate silver ion concentrations. The growth of gold nanorods via seed-mediated method has been explained in two mechanisms [3]: One is the surfactant-preferential-binding-directed growth proposed by Murphy and coworkers based on high-resolution TEM (HRTEM) imaging, which showed the atomic structure of the nanorods during various formation stages. It is believed that the intrinsic structural twinning in the face-centered cubic (fcc) metallic nanoparticles causes symmetry breaking to produce anisotropic nanoparticles. And the following anisotropic growth into nanorods is due to the preferential binding of the CTA^+ head group to faceted seed nanoparticles on the {100} faces [3]. The other theory is the electric-field-directed growth mechanism proposed by Mulvaney and coworkers, in which the $AuCl^{2-}$ ions bound to cationic micelles diffused to seed nanoparticles at high curvature points due to higher electric double-layer gradient at these points and thus leads to nanoparticle growth in rod shape. Also, it might be noted that the effect of the silver ion content is not always to increase the aspect ratio of NRs. Nikoobakht et al. showed that by increasing the concentration of the $AgNO_3$ in the growth solution to more than a specific value, the length of the NRs started to decrease. The negative effect of the silver ions in higher concentration was probably due to their interaction with the bromide counterion of the surfactant monomers [8]. Finally, the size of Au nanorods can be tailored by carefully adjusting other growth conditions, such as the pH of the growth solution, the growth temperature, the compositions of the reagents, and the nanostructures of the seeds, in the seed-mediated growth process. For instance, a very recent study has demonstrated that seed-mediated growth of Au nanorods could be achieved with high yields and uniformity through the use of aromatic additives in addition to CTAB [14]. However, even if the growths follow the same protocols and procedures, the shape and size of Au nanorods may vary among different batches, which is one of major challenges that hinder the potential industrial applications of Au nanorods [15].

Electrochemical Synthesis

Gold nanorods have also been prepared via electrochemical oxidation/reduction within a simple

two-electrode type of cell [9]. A gold metal plate is used as the anode and a platinum plate is used as the cathode in the electrochemical cell. Both electrodes are immersed in an electrolytic solution consisting of a cationic surfactant, hexadecyltrimethylammonium bromide (CTAB), and a rod-inducing cosurfactant (tetraoctylammonium bromide, TOAB). The CTAB serves not only as the supporting electrolyte but also as the stabilizer for nanoparticles to prevent their further growth. During the synthesis, the bulk gold metal is converted from the anode to form gold nanoparticles mostly at the interfacial region of the cathode surface and within the electrolytic solution. A current-controlled electrolysis is used throughout the process and an appropriate amount of acetone added into the electrolytic solution is necessary [9]. The synthesis is conducted under an ultrasonication (for dissipating the nanorods away from cathode) and a controlled temperature (typically at 38 °C). It is suggested that the role of acetone is to facilitate the incorporation of cylindrical shape-inducing cosurfactant into the CTAB micellar framework and inducing the cylindrical growth to form the Au–CTAB–TOAB system [3, 9]. The typical rod-inducing cosurfactant (e.g., TOAB) employed is an overall much more hydrophobic cationic surfactant. Similarly, by inserting a silver plate into the solution and releasing silver ions, the aspect ratio of nanorods can be varied.

Photochemical Reduction

In this method, auric acid is bound to rodlike cationic micelle surfactants to form ion pairs and then excited by UV light to reduce metal ions to form metals via electron transfer from metal ions to ligands [3]. It is proposed that the rod formation is due to a two-step aggregation process [10], that is, metal nuclei aggregation to form primary particles and then primary crystal aggregation to form rodlike structures due to specific crystal face stabilization by surfactant micelles. In this method, addition of sodium chloride electrolyte, the concentration of silver ions, and the UV light wavelength/intensity can adjust the growth speed, size, and structures of the nanorods.

More recently, several other techniques have been developed based on the similar mechanisms as electrochemical or photochemical approach. Ultrasonication of growth solution containing CTAB, gold chloride, silver nitrate, and/or ascorbic acid was shown to induce rod formation. Theoretically, ultrasonic irradiation led to the formation of hydroxyl and other radicals which could reduce gold ions to form the initial nuclei necessary for rod growth [16]. Also, the radicals can be formed by irradiation of a growth solution similar to that used in photochemical synthesis [17].

Lithographic Methods

The most commonly utilized lithography technique for fabricating gold nanorods is electron-beam lithography (EBL) [11]. Nanoscale patterns written with an electron beam allow for the formation of nanometer-scale openings in the resist through which gold can be deposited. By liftoff of the remaining resist and the unwanted gold deposited on it, free-standing gold nanorods can be obtained with high-quality and well-controlled geometry. Another approach, so-called focused ion beam (FIB) lithography, utilizes a rastered ion beam to sputter away unwanted portions of a continuous film leaving behind nanostructures of the desired shape [12]. Both methods can fabricate samples with unparalleled control over the size, shape, and alignment. These two methods break the size limit of as-fabricated nanostructure compared to traditional photolithography due to diffraction limit of light. State-of-the-art electron-beam lithography systems can even fabricate Au nanorods of diameters down to ~10 nm [18]. However, unlike previous mentioned chemical/photochemical/electrochemical synthesis methods, the crystalline structure cannot be well controlled. In addition, both methods are very time consuming and expensive.

Template-Assisted Growth

Template methods involve the use of nanoporous track-etched polycarbonate or anodized aluminum oxide (AAO) membrane [13] as the template where gold atoms are deposited by electrochemical reduction of gold ion precursor in an electrochemical cell. The subsequent chemical etching of

the template membrane gives free-standing gold nanorods. The diameter of the nanorods is determined by the pore diameter of the membrane template. For AAO templates, uniform diameters ranging from tens of nanometers to hundreds of nanometers can be prepared by controlling the anodization process. The length of the rods is controlled by varying growth time and applied voltage/current (the amount of the gold deposited), and thus, the aspect ratio of the nanorods can be tuned accordingly. Similar to lithographic methods, crystalline structure is less controllable in this method. However, template method provided similar excellent control over the size (diameter, length, aspect ratio) at a much lower cost. In addition, it is possible to grow multisegment nanorods (such as Au–Ni bimetal nanorods [19]) by switching templates between different electrolyte solutions, which brings opportunities to further introduce new functions to the original gold nanorods.

In summary, the first three approaches are bottom-up methods and the latter two are considered as top-down methods. Among these, the seed-mediated growth is the most popular method for the synthesis of colloidal gold nanorods due to the simplicity of the procedure, high quality and yield of nanorods, ease of particle size/aspect ratio controlling, and flexibility for structural modifications [3]. However, this method, as well as other bottom-up methods which involve surface surfactant, usually results in a double layer of cetyltrimethylammonium bromide (CTAB), which is problematic for bio-conjugation, nonspecific absorption of DNA, cytotoxicity, and stability [20]. These factors have severely limited the use of gold nanorods in biological and nanomedicine applications. Therefore, it is necessary to apply surface modification for gold nanorods fabricated by the bottom-up methods, to increase stability, facilitating surface chemistry, tuning properties, and broadening the practical applications.

Surface Modifications

Surface modifications of gold nanorods have attracted considerable interest in recent years, and generally two types of surface modifications have been developed.

One is to introduce or exchange surface surfactant (e.g., CTAB) by other chemical or biological functional groups on nanorod surface. For example, Takahashi et al. [21] utilized phosphatidylcholine (PC) to replace the CTAB and successfully reduced the cytotoxicity. More recently, researchers applied ligand exchange and customization by using thiol group [20] and found that the functionalized nanorods had very strong binding with DNA which demonstrated potential applications in biomedical fields.

The other is to coat nanorods surface with hard inorganic materials. Firstly, gold nanorods themselves are ideal templates for further growth of other metallic nanomaterials. By further growth of silver, or platinum, or palladium thin layer(s), the surface properties of Au nanorods could be modified or improved by forming such core–shell structures. For example, silver coating can significantly enhance Raman scattering [3]. Secondly, one could also coat gold nanorod surface with iron oxide or silicon oxide hard layers. SiO_2 is of special interests since it is highly biocompatible, stable, and hydrophilic, and Au nanorods coated with a thin layer of silicon oxide can be easily adopted for further surface functionalizations [22].

Examples of Applications

Drug/Gene Delivery and Selective Release

The mechanism is to use the shape transformation and heating effect of gold nanorods under exposure to near-infrared (NIR) laser pulses to result in the drug or DNA release due to surface atom rearrangement or breaking of the chemical bonds. For example, Wijaya et al. used gold nanorods with two different lengths to carry two different DNAs [23]. When laser with appropriate wavelength is applied, selective release of DNA could be made and the releasing rate was up to 50–80 %. In the work by Wu et al. [24], the thiolated gene of enhanced green fluorescence protein (EGFP) was covalently conjugated to gold nanorods. Under laser irradiation, gold nanorods underwent significant shape transformation, resulting in the release of EGFP DNA.

Molecular Imaging

Gold nanorods can be ideal imaging probes or contrast agents for molecular imaging applications, due to their unusual optical properties at NIR region. Compared to conventional contrast agents – such as radionucleotide in PET and SPECT, gadolinium compounds in MRI, and dye molecules in optical fluorescence imaging [3] – gold nanorods can act as NIR fluorescence probe and have two main advantages: firstly, near-IR laser introduces less damage to living cells; secondly, the penetrating depth for NIR laser in biological tissues is much larger than that for visible light, up to hundreds of mm. These make near-IR imaging with gold nanorods ideal for in vivo diagnostics. For example, Durr et al. modified the surface of gold nanorods with anti-epidermal growth factor receptor (anti-EGFR) antibody and attached them onto cancel cell selectively [25]. By performing two-photon fluorescence (TPF) on the three-dimensional tissue with cancer cells, researchers found the fluorescence intensity due to gold nanorod labeling was about three orders of magnitude higher than the intensity of cancer cell fluorescence illumination [25].

Surface Plasmon Resonance Sensing

Due to the fact that the SPR frequency highly depends on the dielectric constant of the surrounding medium, increasing the dielectric constant of the surrounding medium leads to a redshift of the SPR wavelength. Therefore, the SPR wavelength either by absorption or scattering provides great opportunity to monitor the changes of the local environment of the nanoparticles and thus can be used for sensing. Compared to gold nanospheres, the unique transverse and longitudinal SPR characteristics for gold nanorods provide higher accuracy: the longitudinal SPR absorption is very sensitive to the change of surrounding environment while the transverse SPR will have no obvious redshift and can be served as a reference. In addition, the characteristic SPR responses for gold nanorods with different aspect ratios can be utilized for higher efficient diagnostics. For example, Yu et al. attached three types of antibody molecules to three different gold nanorods (with different AR), and multiplexing detection of three targets was successfully realized via the observation of three LSPR redshifts [26].

Plasmonic Nanoantenna

Plasmonic nanostructures, especially gold nanorods, possess strong light absorption and scattering capabilities and narrow line widths, so that they can control and manipulate light–matter interactions at the nanoscale and thereby act as nanoantennas [15]. As nanoantenna, far-field radiation can be captured by the gold nanorods and transferred to highly confined electromagnetic fields surrounding them, which allows for nanoscale optical signal processing. Reciprocally, gold nanoantenna can also transfer the localized near-field radiations from a nanoscale object into far-field optical signals. These far-field signals will be labeled with the characteristics of the original plasmon, such as the frequency, polarization, and spatial radiation direction, which can be modulated by the longitudinal plasmon mode or introducing hybrid nanostructures and metamaterials. Recent studies further show that 2D [27] or 3D [28] arrays of appropriately tuned gold nanorods can also function as Yagi–Uda antenna, where different sized gold nanorods work as feeds, reflectors, and directors. This nano-optical Yagi–Uda antenna can significantly improve the directionality of emission, from broad spatial distributions into a narrow forward angular cone.

Surface-Enhanced Raman Scattering (SERS)

Raman scattering has been used for biological detection since the first interpretable laser-excited Raman spectrum of lysozyme by Lord et al. in 1970 [3]. However, the small Raman scattering cross sections of most biological molecules lead to very weak Raman signal, which hinder their biomedical applications. Discovered by Fleischmann et al. in 1974, surface-enhanced Raman scattering (SERS) can enhance the Raman signal by six orders of magnitude: To maximize Raman intensity, the distance between the molecules and the metal surface and the metallic nanoparticle shape are two major considerations. For example, Nikoobaht et al. observed SERS enhancement factors on the order of 10^4–10^5 for the molecules on the nanorods

(25 nm 10 nm), while no such enhancement on the nanospheres [29]. In addition, nanorods can be synthesized with different aspect ratios so as to overlap the longitudinal surface plasmon band to the excitation source spectrum to obtain the largest enhancement [30]. As demonstrated by Jana et al., optimum length scale is 20–100 nm irrespective of nanoparticle size and shape [31].

For in vivo biosensing and biodetection, Raman-active molecule-coded nanorods are transferred to organic tissues and thereafter detected by measuring Raman scattering signals. In the study by Bhatia et al., Au nanorods were coded with several Raman-active molecules, such as IR-792 perchlorate, DTTC-765, and DTDC-655, and then injected into the mice implanted with bilateral human MDA-MB-435 tumors. Under near-infrared illumination, the Raman-active molecule-coded Au nanorods were found to give intense in vivo Raman signatures with concentrations down to attomoles. This remote sensing ability in tandem with the photothermal conversion-based therapy has made Au nanorods an excellent multiplex biomedical platform [32].

Photothermal Therapy

The recent interest in the use of gold nanorods in photothermal therapy rooted from the advantages of their small size, excellent stability, high efficiency, and tunable absorption in the biological window. Gold nanorods become very popular also because they can absorb light in the NIR window (650–900 nm) where light penetration is optimal due to minimal absorption by water and hemoglobin in the tissue as discussed earlier. In addition, these plasmonic nanoparticles could enable dual imaging and therapy functions at the same time. For example, Huang et al. provided an in vitro demonstration of gold nanorods as contrast agents for both molecular imaging and photothermal cancer therapy [3]. In their work, nanorods were conjugated to anti-epidermal growth factor receptor (anti-EGFR) monoclonal antibodies and incubated in cell cultures with a nonmalignant epithelial cell line (HaCat) and two malignant oral epithelial cell lines. The anti-EGFR antibody-conjugated nanorods binded specifically to the surface of the malignant-type cells with a much higher affinity due to the overexpressed EGFR on the cytoplasmic membrane of the malignant cells [3]. As a result of the strongly scattered red light from gold nanorods in dark field, observed using a laboratory microscope, the malignant cells were clearly visualized and diagnosed from the nonmalignant cells. It was found that, after exposure to continuous red laser at 800 nm, malignant cells require about half of the laser energy to be photothermally destroyed than the nonmalignant cells. Thus, both efficient cancer cell diagnostics and selective photothermal therapy were realized at the same time.

Future Directions

Despite the fact that gold nanorods have already exhibited a wide range of promising applications and great potentials in disease diagnosis and treatment, there remain plenty of things to be done before they can actually advance into the realm of clinical trials. First of all, although some of the biomedical applications, such as photothermal cancer therapy, have been successfully demonstrated in animal experiments, more in vivo experimental data are definitely needed in order to establish a complete set of parameters for translational research and clinical practice. Secondly, considerable refinements need to be made, for example, by further introducing hyper-structures to the nanorods, such as porous gold nanostructures, or advanced surface coatings, so that drug/gene delivery efficiency and imaging/sensing mechanisms could be correspondingly tuned. Last but not the least, rigorous studies on the long-term effects, such as distribution and fate of gold nanorods in the human body, should be carried out carefully.

Cross-References

- ▶ Nanomedicine
- ▶ Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites

References

1. Mie, G.: Considerations on the optics of turbid media, especially colloidal metal sols. Ann. Phys. **25**, 377 (1908)
2. Gans, R.: Form of ultramicroscopic particles of silver. Ann. Phys. **47**, 270 (1915)
3. Huang, X., Neretina, S., El-Sayed, M.A.: Gold nanorods: from synthesis and properties to biological and biomedical applications. Adv. Mater. **21**, 4880–4910 (2009)
4. Yorulmaz, M., Khatua, S., Zijlstra, P., Gaiduk, A., Orrit, M.: Luminescence quantum yield of single gold nanorods. Nano Lett. **12**, 4385–4391 (2012)
5. Li, C.Z., Male, K.B., Hrapovic, S., Luong, J.H.T.: Fluorescence properties of gold nanorods and their application for DNA biosensing. Chem. Commun. **31**, 3924–3926 (2005)
6. Link, S., El-Sayed, M.A.: Spectral properties and relaxation dynamics of surface plasmon electronic oscillations in gold and silver nanodots and nanorods. J. Phys. Chem. B **103**, 8410–8426 (1999)
7. Jana, N.R., Gearheart, L., Murphy, C.J.: Seed-mediated growth approach for shape controlled synthesis of spheroidal and rod-like gold nanoparticles using a surfactant template. Adv. Mater. **13**, 1389–1393 (2001)
8. Nikoobakht, B., El-Sayed, M.A.: Preparation and growth mechanism of gold nanorods (NRs) using seed-mediated growth method. Chem. Mater. **15**, 1957–1962 (2003)
9. Yu, Y.Y., Chang, S.S., Lee, C.L., Chris Wang, C.R.: Gold nanorods: electrochemical synthesis and optical properties. J. Phys. Chem. B **101**, 6661–6664 (1997)
10. Kim, F., Song, J.H., Yang, P.: Photochemical synthesis of gold nanorods. J. Am. Chem. Soc. **124**, 14316–14317 (2002)
11. Billot, L., Chapelle, L.D.L., Grimault, A.S., Barchiesi, A.V.D., Bijeon, J.L., Adam, P.M., Royer, P.: Surface enhanced Raman scattering on gold nanowire arrays: evidence of strong multi- polar surface plasmon resonance enhancement. Chem. Phys. Lett. **422**, 303–307 (2006)
12. Babu Dayal, P., Koyama, F.: Polarization control of 0.85 mm vertical-cavity surface-emitting lasers integrated with gold nanorod arrays. Appl. Phys. Lett. **91**, 111107 (2007)
13. Sander, M.S., Tan, L.S.: Nanoparticle arrays on surfaces fabricated using anodic alumina films as templates. Adv. Funct. Mater. **13**, 393–397 (2003)
14. Ye, X.C., Jin, L.H., Caglayan, H., Chen, J., Xing, G.Z., Zheng, C., Doan-Nguyen, V., Kang, Y.J., Engheta, N., Kagan, C.R., Murray, C.B.: Improved size-tunable synthesis of monodisperse gold nanorods through the use of aromatic additives. ACS Nano **6**, 2804 (2012)
15. Chen, H., Shao, L., Li, Q., Wang, J.: Gold nanorods and their plasmonic properties. Chem. Soc. Rev. **42**, 2679 (2013)
16. Okitsu, K., Sharyo, K., Nishimura, R.: One-pot synthesis of gold nanorods by ultrasonic irradiation: the effect of pH on the shape of the gold nanorods and nanoparticles. Langmuir **25**(14), 7786–7790 (2009)
17. Abidi, W., Selvakannan, P.R., Guillet, Y., Lampre, I., Beaunier, P., Pansu, B., Palpant, B., Remita, H.: One-pot radiolytic synthesis of gold nanorods and their optical properties. J. Phys. Chem. C **114**(35), 14794–14803 (2011)
18. Koh, A.L., McComb, D.W., Maier, S.A., Low, H.Y., Yang, J.K.W.: Sub-10 nm patterning of gold nanostructures on silicon-nitride membranes for plasmon mapping with electron energy-loss spectroscopy. J. Vac. Sci. Technol. B **28**, C6O45 (2010)
19. Salem, A.K., Searson, P.C., Leong, K.W.: Multifunctional nanorods for gene delivery. Nat. Mater. **2**, 668–671 (2003)
20. Wijaya, A., Hamad-Schifferli, K.: Ligand customization and DNA functionalization of gold nanorods via round-trip phase transfer ligand exchange. Langmuir **24**, 9966–9969 (2008)
21. Takahashi, H., Niidome, Y., Niidome, T., Kaneko, K., Kawasaki, H., Yamada, S.: Modification of gold nanorods using phosphatidylcholine to reduce cytotoxicity. Langmuir **22**(1), 2–5 (2006)
22. Pastoriza-Santos, I., Perez-Juste, J., Liz-Marzan, L. M.: Silica-coating and hydrophobation of CTAB-stabilized gold nanorods. Chem. Mater. **18**, 2465–2467 (2006)
23. Wijaya, A., Schaffer, S.B., Pallares, I.G., Hamad-Schifferli, K.: Selective release of multiple DNA oligonucleotides from gold nanorods. ACS Nano **3**(1), 80–86 (2009)
24. Chen, C.C., Lin, Y.P., Wang, C.W., Tzeng, H.C., Wu, C.H., Chen, Y.C., Chen, C.P., Chen, L.C., Wu, Y.C.: DNA-gold nanorod conjugates for remote control of localized gene expression by near infrared irradiation. J. Am. Chem. Soc. **128**(11), 3709–3715 (2006)
25. Durr, N.J., Larson, T., Smith, D.K., Korgel, B.A., Sokolov, K., Ben-Yakar, A.: Two-photon luminescence imaging of cancer cells using molecularly targeted gold nanorods. Nano Lett. **7**(4), 941–945 (2007)
26. Yu, C., Irudayaraj, J.: Multiplex biosensor using gold nanorods. Anal. Chem. **79**(2), 572–579 (2007)
27. Kosako, T., Kadoya, Y., Hofmann, H.F.: Directional control of light by a nano-optical Yagi–Uda antenna. Nat. Photonics. **4**, 312 (2010)
28. Dregely, D., Taubert, R., Dorfmüller, J., Vogelgesang, R., Kern, K., Giessen, H.: 3D optical Yagi–Uda nanoantenna array. Nat. Commun. **2**, 267 (2011)
29. Nikoobakht, B., Wang, J., El-Sayed, M.A.: Surface-enhanced Raman scattering of molecules adsorbed on gold nanorods: off-surface plasmon resonance condition. Chem. Phys. Lett. **336**, 17–23 (2002)
30. Orendorff, C.J., Gearheart, L., Jana, N.R., Murphy, C. J.: Aspect ratio dependence on surface enhanced

Raman scattering using silver and gold nanorod substrates. Phys. Chem. Chem. Phys. **8**, 165–170 (2006)
31. Jana, N.R., Pal, T.: Anisotropic metal nanoparticles for use as surface-enhanced Raman substrates. Adv. Mater. **19**, 1761–1765 (2007)
32. von Maltzahn, G., Centrone, A., Park, J.H., Ramanathan, R., Sailor, M.J., Hatton, T.A., Bhatia, S.N.: SERS-coded gold nanorods as a multifunctional platform for densely multiplexed near-infrared imaging and photothermal heating. Adv. Mater. **21**, 3175–3180 (2009)

Graphene

Debdeep Jena
Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, USA

Synonyms

Bilayer graphene; Graphite; Nanoribbons

Definition

Graphene is defined as an atomically thin planar allotrope of carbon. Each carbon atom in this two-dimensional (2D) crystal is bonded to three other carbon atoms, forming a hexagonal structure. Graphene is the building block of graphite, carbon nanotubes, and fullerenes. It is the thinnest crystal known at this time.

Overview

The various solid forms of carbon have been known and used since thousands of years. In three-dimensional (3D) form, charcoal is the prototype solid fuel used to sustain life through the ages. Lavoisier first showed in 1772 that diamond was composed entirely of carbon by "burning" diamond in oxygen and producing carbon dioxide. This 3D crystalline form of carbon is the hardest, electrically insulating, and most thermally conductive solid known. These properties of diamond form the basis for its many uses. Graphite is a layered solid form of carbon, and is more stable than diamond. It is remarkably different from diamond in its physical properties. It is an electrical conductor, exhibits high anisotropy in its electrical and thermal conductivities, and is very refractory. Because it is capable of carrying copious amounts of electrical current, graphite was instrumental in the discovery of the light bulb, which converts electrical energy into photons for illumination. It is still used in arc lamps and where high currents and high temperature operation is necessary. The uses of graphite are more diverse, ranging from neutron moderation in nuclear reactors to the pencil lead, which exploits its layered structure. The rich range of physical properties of the various forms of carbon emerges from seemingly minor differences in the microscopic arrangements of the carbon atoms.

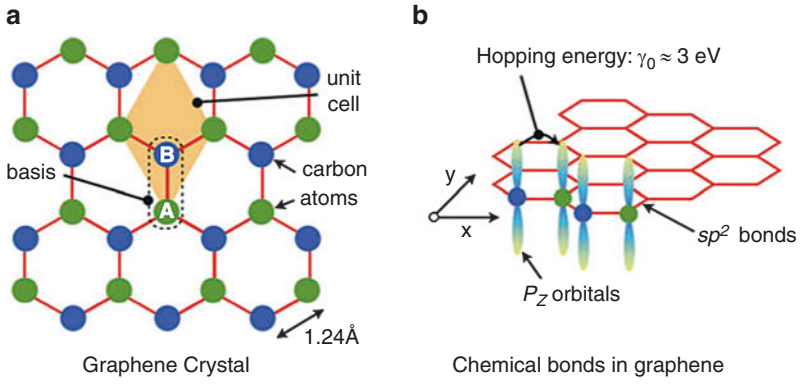
With the advent of electron microscopy, it became possible to investigate the atomic structure of solids made of carbon. Lower dimensional forms of carbon were then discovered. In semiconductor heterostructures, such as those made of silicon or GaAs, man-made quantum-size effect materials were created starting from 3D bulk semiconductors to 2D quantum wells, 1D quantum wires, and 0D quantum dots. Thus, nanostructures based on semiconductor heterostructures developed from large to small. In sharp contrast, the first carbon nanostructure discovered was zero-dimensional (0D) fullerenes (C_{60}), followed by 1D carbon nanotubes. The 2D form of carbon, graphene, has been known since the 1940s to be the building block of graphite. It was also found subsequently to be the building block of fullerenes and nanotubes. But by itself, a single layer of graphene was believed to be thermodynamically unstable. Based on the Mermin–Wagner theorem in solid-state physics, the energy cost for spontaneously broken symmetries in 2D crystals is low. This means that a perfect "flat" 2D crystal can increase its entropy by forming ripples or crumples. Thus, to a certain extent, the isolation of 2D atomic crystals in 2004 was a surprise. Using scotch tape, single layer flakes of graphene were isolated from graphite, and their electronic properties were investigated [1]. More than the isolation of single

layers of graphene, the findings of the nature of electron motion in it have far-reaching consequences on the fundamental understanding of the electronic properties of solids. These findings have opened a new door to aspects of symmetry in condensed matter physics. Rapid progress has occurred in exploiting these properties in applications. To understand the origins of the peculiar properties of graphene, its atomic structure is discussed next.

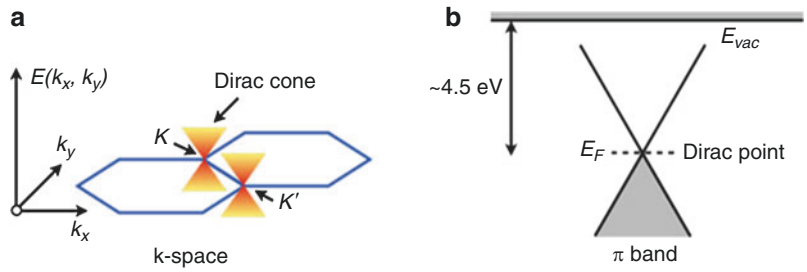
Crystal Structure of Graphene

Carbon is the sixth element in the periodic table, meaning it has six protons in its nucleus. Charge neutrality requires the presence of six electrons. The six electrons fill the lowest three orbitals in the arrangement $[1s]^2[2s]^2[2p]^2$, where the $|s\rangle$ orbitals are spherically symmetric, and $|p\rangle$ orbitals are elongated along a linear axis. When carbon atoms are brought close to each other, the two electrons in the innermost $[1s]^2$ orbitals remain close to the nucleus, and do not participate in the bonding process. The outermost orbitals start interacting, and are capable of forming chemical bonds. The four electrons in the outermost $[2s]^2[2p]^2$ orbitals can be shared between atoms in a variety of ways, each leading to a different form of carbon. For example, when four nearest carbon atoms share all four electrons of a carbon atom equally, the bond is sp^3. An sp^3 bond is tetrahedral, and is inherently 3D. Carbon atoms bonded in this fashion arrange periodically in 3D space, forming the crystal structure of diamond. Carbon is rather versatile in its bonding possibilities; when combined with hydrogen atoms, electrons in carbon atoms can form sp^2 or sp bonds as well. This means carbon atoms can form 2D planar molecules (e.g., the hexagonal molecule benzene or C_6H_6), or long one-dimensional chains (e.g., in conductive polymers). This versatility of electrons in carbon is the root of the complexity of organic molecules that lead to the emergence of life itself.

When a carbon atom shares electrons with *three* nearest neighbors, the bonding is sp^2, written as $|sp^2\rangle = a|s\rangle + b|px\rangle + c|py\rangle$, where *a, b, c* are constants. This bonding leaves one electron behind for every carbon atom. The three chemical bonds are separated by 120°, and are planar. This is the crystal structure of graphene, depicted in Fig. 1. The energies of the electrons forming the sp^2 bonds bunch together into bands called the σ band. The σ band is responsible for the perfectly planar nature of graphene, and determines its structural and vibrational properties, such as its thermal conductivity and Young's modulus, among others. Thus, graphene is a 2D crystal with hexagonal arrangements of carbon atoms. The atomic spacing is $a_{cc} \sim 0.124$ nm. As indicated in Fig. 1, the smallest repeating unit that can be translated to cover the entire 2D space consists of not one, but *two nearest-neighbor carbon atoms*. Though the atoms are physically identical (they are both carbon), they are labeled as A and B in Fig. 1 since the "mechanical" symmetry can be broken by the electron wavefunction associated with them. The A-B atom combination is the two-atom "basis" of the graphene crystal, similar



Graphene, Fig. 1 Crystal structure and chemical bonds of graphene. (**a**) Hexagonal crystal structure showing the atomic basis of two equivalent carbon atoms, the interatomic distance, and the unit cell. (**b**) Chemical bonds, showing the sp^2 hexagonal structure, and the p_z orbitals that impart graphene its conductivity



Graphene, Fig. 2 Allowed energies for electrons in graphene in the k-space, and bandstructure. (**a**) The hexagonal symmetry of electrons in the momentum space, and the Dirac cones originating from the mobile p electrons. (**b**) The bandstructure, showing electron states filled in the p band till the Dirac point energy, which is 4.5 eV below the vacuum energy level

to that in 3D diamond or silicon. The primitive repeating unit is shown as a parallelogram. By virtue of being 1-atom thick, graphene is the "thinnest" crystal. It is impermeable to small gas molecules. The structural and chemical rigidity of graphene is protected by symmetry: to form a chemical bond to graphene, one needs to convert sp^2 bonds to sp^3. This requires buckling of the crystal since the bonds have to change from 2D to 3D, requiring the movement of atoms. If and when this is achieved, the material is no longer graphene. In addition, carbon is the lightest in group-IV of the periodic table, and the strong σ bonds lead to the highest in-plane thermal conductivity. However, the σ band does not directly influence the electronic properties of graphene. Electrons in sp^3 bonds and bands resulting from them determine the electronic *and* structural properties in conventional 3D semiconductor crystals (e.g., Si, GaAs, diamond). In graphene, the electronic and structural properties arise from different bands. For understanding the electronic properties of graphene, attention is focused back to the *electron that was left out* in the formation of the sp^2 bond.

Electronic Properties of Graphene

The electrons left out in the chemical bonding process are from the $|p_z\rangle$ orbitals of the carbon atoms forming graphene. The linear axis of the $|p_z\rangle$ orbital protrudes outward from the plane, as shown in Fig. 1. These electrons can hop between the nearest neighbor carbon atoms. The energy characterizing this hopping is γ_0 ~3.0 eV. Thus, their wavefunctions become delocalized. Their allowed energies bunch together to form what are called π bands. Electrons in the π bands can be visualized as sliding over the hexagonal carbon atom structure of graphene. These electrons impart graphene its exceptional electrical conductivity.

The allowed energies of the delocalized π electrons depend on their wavelength, in accordance with the rules of quantum mechanics. According to the de-Broglie relation, the wavelength is inversely proportional to momentum. The momentum is given by $\hbar k$, where $\hbar$ is Planck's constant and $k = 2\pi/\lambda$, where λ is the wavelength. The allowed energies of electrons $E(k_x, k_y)$ can be then plotted in the k-space, as shown in Fig. 2. By taking into account the basis of two atoms, the wavefunctions of electrons in graphene are represented as "spinors," which are two-component matrices of the form $\frac{exp[i\mathbf{k}\cdot\mathbf{r}]}{\sqrt{2}}\begin{pmatrix}1\\ exp[i\theta]\end{pmatrix}$, where $\mathbf{k} = \mathbf{x}k_x + \mathbf{y}k_y$, and $tan\,\theta = k_y/k_x$. Using a tight-binding procedure, the energy spectrum $E(k_x, k_y)$ is given by the equation $\hbar v_F\left(\vec{\sigma}\cdot\mathbf{k}\right)|\psi\rangle = E\left(k_x, k_y\right)|\psi\rangle$, where $\vec{\sigma} = \mathbf{x}\sigma_x + \mathbf{y}\sigma_y + \mathbf{z}\sigma_z$ is a vector composed of the three Pauli spin matrices. The term $v_F = 3a_{cc}\gamma_0/2\hbar \sim 10^8$ cm/s is a parameter characterizing the bandstructure, with dimensions of velocity. This Hamiltonian has a closer analogy to the Dirac equation rather than the Schrodinger equation used for conventional semiconductors. The Dirac Hamiltonian consists of Pauli spin matrices in the kinetic energy term, which implies that electrons have spin-like properties based

on their momentum. This property is termed "pseudo-spin," and may be pictured as a vector whose direction depends on the momentum. In traversing a path around a constant-energy circle in the Dirac cone, the wavefunction accumulates a phase of π due to the pseudospin. This phase is called Berry's phase, and has consequences on scattering and on quantum Hall effects.

The energy bandstructure obtained from this "Dirac equation" for the π band is given by $E(k_x,k_x)=E_F\pm\gamma_0\sqrt{1+4\cos\left(\sqrt{3}k_xa/2\right)4\cos\left(k_xa/2\right)+4\cos^2\left(k_ya/2\right)}$, where $a=\sqrt{3}a_{cc}$ is the lattice constant of graphene. Here, E_F is the Fermi energy, which for pristine undoped graphene is ~4.5 eV below the vacuum energy level. This is the electron affinity or the work function of pure graphene (Fig. 2). The positive sign is for the conduction band that lies above E_F, and the negative sign is for the valence band. One can note that $E(0,4\pi/3a)=E_F\pm0$ eV, which implies that there is *no energy bandgap* in the π band of graphene. The points in the k-space such as $K=(0,4\pi/3a)$ where the conduction and valence bands meet are called the Dirac points. These points form a hexagonal network in the k-space as indicated in Fig. 2, with the nearest-neighbor points labeled as K and K'. In analogy to traditional semiconductors, graphene is a direct bandgap semiconductor. Since each K point is shared between three hexagons, there are two distinct points in the smallest repeating unit cell in k-space, called the Brillouin zone. Thus, in analogy to traditional semiconductors, graphene has a valley degeneracy of 2.

The Fermi energy is at the Dirac point for pure graphene, meaning at zero temperature and in the absence of doping, all electronic states below the Dirac point are occupied, and all above it are empty. Electrical conductivity thus occurs in graphene due to transitions of electrons between allowed states close to the Dirac point. In analogy to the effective mass approximation for semiconductors, expanding the energy dispersion around the Dirac point, say around $K=(0,4\pi/3a)$, yields the rather interesting relation $E(k_x,k_x)-E_F\approx\hbar v_F\sqrt{k_x^2+k_y^2}=\hbar v_F|\mathbf{k}|$. This is the equation of a cone, and owing to its origin in the Dirac-like Hamiltonian, referred to as the Dirac cone. The conical form of the energy dispersion remains valid for energies exceeding a few eVs from the Dirac point energy, which is the range of interest for all low-energy electronic and optical properties. A few intriguing aspects of this energy dispersion can be identified by comparison with traditional semiconductors.

In traditional semiconductors, the kinetic energy of electrons near a band edge E_C is proportional to the square of the electron quasi-momentum $E(k)-E_C=(\hbar k)^2/2m^*$, with the magnitude of the kinetic energy parameterized by an effective mass distinct from the mass of free electrons. The effective mass m * is a parameter that absorbs the details of the crystal potential. It replaces the complex motion of free electrons in a crystal by that of an electron with a different mass in free space. This is a considerable simplification, one that is heavily leveraged in the design of electronic and optical devices such as transistors, light-emitting diodes, and lasers based on traditional semiconductors. In stark contrast, the kinetic energy of electrons in graphene depends on its momentum *linearly* $E(k_x,k_x)=v_F\cdot\hbar|\mathbf{k}|$, where Fermi energy at the Dirac point is used as reference. Wallace originally derived this linear form of the bandstructure of graphene in 1947 [2].

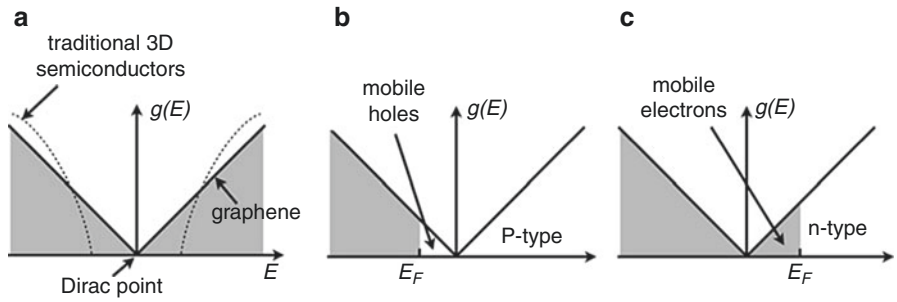
A linear dependence of kinetic energy on momentum is characteristic of light or photons. The energy dispersion of electrons in graphene is identical to that of photons, with the speed of light $c=3\times10^8$ m/s replaced by the "Fermi velocity" $v_F\approx10^6$ m/s. The maximum speed of electrons in graphene is $v_F/c=1/300$ that of light in free space. In traditional semiconductors, particularly those of small bandgaps, the velocity of carriers can also approach these speeds. A consequence of the linear energy dispersion is the difficulty in associating an effective mass to electrons in graphene. The effective mass is *defined* as the curvature of the energy dispersion. By definition, the curvature of a linear curve is zero. Thus, electrons in graphene are referred to as "massless Dirac fermions." The "massless" feature should be understood in the light of the discussion above; electrons in graphene are identical to electrons in

any other material. Unlike *truly massless photons*, they need external driving forces such as electric or magnetic fields to carry currents. *What is unique about graphene is the linear energy dispersion of electrons*. In the dispersion of electrons are hidden a number of symmetries, each leading to many of its unique features.

Carrier Statistics in Graphene

The absence of an energy bandgap in graphene is related to the light-like linear dispersion. The linear dispersion also implies that the transport properties of filled states in the conduction band (see Fig. 2) and empty states in the valence bands (holes) are similar. The conduction band is a mirror-reflection of the valence band. This symmetry is special. Two-dimensional electron or hole gases in semiconductor quantum wells have multiple subbands to occupy, and can be truly 2D at very low energies or at low temperatures as long as only one subband is occupied. In stark contrast, electrons and holes in 2D graphene are truly 2D at all temperatures, since there are no higher subbands near the Dirac point. Based on the energy dispersion, the density of states (DOS) of graphene is given by $g(E) = 2|E|/\pi(\hbar v_F)^2$, which is linear in energy. Three-dimensional semiconductors have density of states that depend on energy as $g(E) \sim \sqrt{E - E_0}$, whereas the density of states of 2D subbands in quantum wells are energy-independent. The DOS of 2D graphene is shown in comparison to traditional 3D semiconductors in Fig. 3. Based on the DOS, the electron and hole densities in pristine graphene are given by $n_0 = p_0 = (\pi/6) \cdot (kT/\hbar v_F)^2$, where T is the temperature and k is the Boltzmann constant. At room temperature, there are $\sim 8 \times 10^{10}/\text{cm}^2$ electrons in the conduction band due to interband thermal excitation, leaving behind an equal number of holes in the valence band. These carriers are mobile, and lead to a finite conductivity of graphene at room temperature even in the absence of intentional doping. They set the *lower limit* of mobile carrier concentration in 2D graphene: with any changes such as doping or field-effect, the net mobile carrier density can only exceed this number. The $\sim T^2$ dependence of intrinsic interband thermal generation of carriers is in stark contrast to traditional semiconductors, where the density depends on the bandgap as $\exp(-E_G/2kT)$ where E_G is the bandgap.

When graphene is electrostatically doped, the Fermi energy moves away from the Dirac point into the valence or the conduction band, making it p-type or n-type. Under these conditions, the carrier concentrations are given by $n = E_F^2/\pi(\hbar v_F)^2$. Thus, the Fermi energy changes as the square root of the carrier concentration, $E_F = \hbar v_F \sqrt{\pi n}$. At a 2D carrier concentration of $10^{12}/\text{cm}^2$, $E_F \sim 0.12$ eV from the Dirac point, and increases to $E_F \sim 0.37$ eV at $10^{13}/\text{cm}^2$. The Fermi energy is tunable over a wide range, which implies a tunable work function for 2D graphene. The electrons or holes thus formed in graphene are responsible for its electrical conductivity. The current per unit width that flows in a graphene



Graphene, Fig. 3 Density of states of pure graphene and p-type and n-type graphene. (**a**) Density of states of graphene is linear and symmetric around the Dirac point energy. It has no energy gap, unlike conventional semiconductors. (**b**) and (**c**) P-type and n-type graphene created by moving the Fermi level into the valence and conduction band, respectively

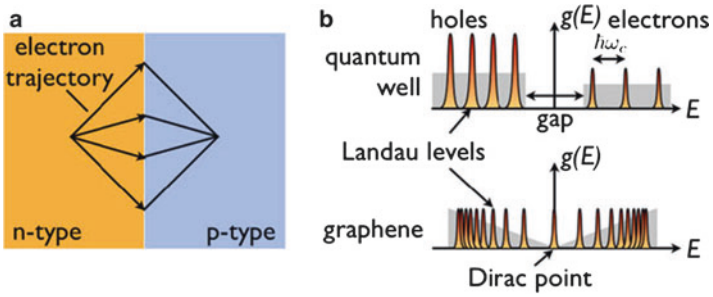
sheet in response to a voltage applied across contacts made to graphene is given by $I/W = qn\langle v\rangle$, where $\langle v\rangle$ is the ensemble-averaged velocity. In the absence of any scattering, the velocity of *all* carriers in graphene is the Fermi velocity v_F, by virtue of its linear bandstructure. However, since a net current flows only along the electric field, the effective velocity is only a fraction of the Fermi velocity obtained appropriately by angular averaging. The current drive in clean defect-free graphene is typically high, owing to the high Fermi velocity. In the presence of defects, however, the conductivity suffers due to scattering.

Carrier Transport in Graphene

The mobility of electrons in graphene is typically high, if the material is pure. Electron and hole mobilities well in excess of 50,000 cm^2/Vs have been measured at room temperature in the cleanest samples [3]. The high mobility is typically degraded due to scattering from impurities that are not inherent to graphene itself. For example, charged impurities located outside the plane of graphene result in Coulombic scattering of mobile carriers. Similarly, integrating a dielectric in close proximity to graphene allows electrons in graphene to excite phonon modes in them, resulting in "remote" phonon scattering [4]. When these scattering processes are suppressed, the mobility of carriers in graphene is found to increase. The high mobilities lead to long mean-free paths, and ballistic transport is possible over micrometer length scales. At high voltages, when electrons gain kinetic energies high enough to excite the intrinsic optical phonon modes of graphene, they collide with the carbon atoms and create optical phonons. By virtue of the lightness of carbon atoms, the optical phonon energy is high (~160 meV), implying that carriers cannot emit optical phonons until they acquire an equivalent kinetic energy, resulting in high velocities.

The high mobility and velocity of carriers leads to high currents in field-effect transistor-like device applications. Due to the lack of a bandgap, the demonstrated current modulation in 2D graphene typically does not exceed a factor of 10X, currently restricting the feasibility of 2D graphene in electronic digital switching. However, applications that do not require a high current modulation stand to benefit from the high current drive and the high speed. Unlike traditional semiconductors, the symmetry of the bandstructure and the absence of a bandgap implies that graphene is as conductive when p-type as when it is n-type. Thus, when the Fermi energy is tuned from the valence band to the conduction band, the current characteristics change from hole conduction to electron conduction symmetrically; this property is termed ambipolar. The ambipolar symmetry is useful for passive applications in electronic circuits such as mixers to generate higher harmonics of oscillatory signals. In most traditional semiconductors, the valence band effective mass is higher than the conduction band effective mass, and the ambipolar nature is not protected by symmetry, as in graphene.

Adjacent regions of graphene can be electrostatically gated with opposite polarity, forming p-n junctions as shown in Fig. 4. In such a case,



Graphene, Fig. 4 Electron "lensing" in graphene, and Landau levels in traditional semiconductors compared to those in 2D graphene. (**a**) Electron transport in graphene mimics light; a p-n junction behaves as a lens for electrons. (**b**) At high magnetic fields, the density of states of graphene collapses into Landau levels that are quite different from what is observed in 2D electron–hole systems in conventional semiconductors

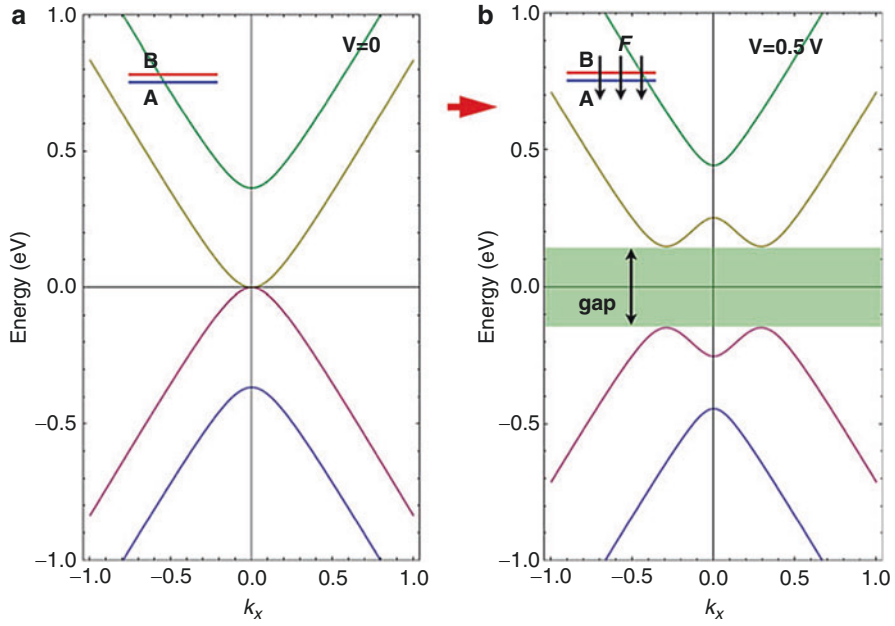
when electron waves are launched from a point in the n-region at various angles to the interface, they have to conserve energy and momentum when they enter the p-region. Since they have to tunnel from the conduction band in the n-side to the valence band in the p-side, the component of momentum parallel to the junction is flipped, resulting in the collimation of the electron waves. The tunneling process in graphene is unique owing to its Dirac dispersion, and is called Klein tunneling. This sort of electron "lensing" is a consequence of the light-like energy dispersion of graphene [5]. It requires the phase of the electron waves to be coherent throughout the transport process, and thus requires long mean free paths and low scattering. Similarly, electrons can be reflected from similar interfaces, and by combining junctions, can be "waveguided" from one region to another through all-electrical controls. The similarity of the electron energy dispersion to that of light makes such processes possible in graphene.

The quantum Hall effect in graphene offers a signature of its energy dispersion. As shown in Fig. 4, in traditional semiconductor quantum wells, 2D electron and hole gases have different effective masses, and hence different density of states. Upon the application of a high magnetic field B, the DOS collapse into uniformly spaced Landau levels, with energies peaking at $E_n = \hbar\omega c\left(n+\frac{1}{2}\right)$ for each band, where $\omega c = qB/m^*$ is the cyclotron frequency and $n = 0, 1, 2, \ldots$ is an integer. As the height of each Landau level and their separation is changed by varying the magnetic field, the Fermi level passes through peaks and minima of the Landau level, causing the longitudinal magnetoresistance to oscillate. The transverse resistance is the Hall resistance. The quantized Hall *conductance* in traditional semiconductor 2DEGs is given by $\sigma_{xy} = \frac{2q^2}{h}n$. The Landau levels of 2D graphene, on the other hand, are given by $E_n = \pm v_F\sqrt{2q\hbar B}\sqrt{n}$, bunching closer as n increases, as shown in Fig. 4. The resulting Hall conductance is given by $\sigma_{xy} = \frac{4q^2}{h}\left(n+\frac{1}{2}\right)$. For $n = 0$, there is a Landau level at the Dirac point for graphene, and the half-integer quantum Hall conductance is a signature of both the Dirac cone dispersion and the Berry phase of π for electron wavefunctions [6, 7]. The half-integer quantum Hall effect has been observed in single layer graphene grown by various methods, and the precision of the quantization is high.

Bilayer Graphene

When two layers of graphene are stacked, the mutual interaction changes the electron energy dispersion, and hence its electronic properties. Bilayer graphene should not be considered as two individual layers of graphene, but one composite electronic system. Bernal stacking is a particular type of stacking, where the A atoms in one layer (Fig. 1) are vertically above the B atoms of the other. For this form of bilayer graphene, the energy dispersion changes from the Dirac cone feature of a single layer to a parabolic bandstructure characterized by an effective mass mass $m* \approx 0.033m_0$, but still with a zero bandgap. This bandstructure is shown in Fig. 5. A remarkable phenomenon occurs upon the application of an electric field vertical to the graphene bilayer. The bandstructure deforms to that shown in Fig. 5b. An energy bandgap appears, which is tunable with the applied electric field [8]. The electric field breaks the symmetry of electron wavefunctions in the direction perpendicular to the bilayer, and this broken symmetry results in a bandgap. Such field-tunable energy bandgaps are rare in solid state systems, and can be expected to lead to novel applications. Energy bandgaps opened in this fashion up to ~0.2 eV have been experimentally measured [9].

The symmetry of the bandstructure of graphene has also led to the proposal of interlayer tunneling devices such as the BISFET [10]. This device mimics Bose-condensation of electron–hole pairs (excitons) in two graphene layers close enough to allow strong interactions, yet far enough to prevent the formation of a bilayer. It exploits collective many-body effects in an all-electronic system at room temperature. As the number of graphene layers stacked increases, the energy bandstructure evolves toward that of bulk graphite. Graphite is a



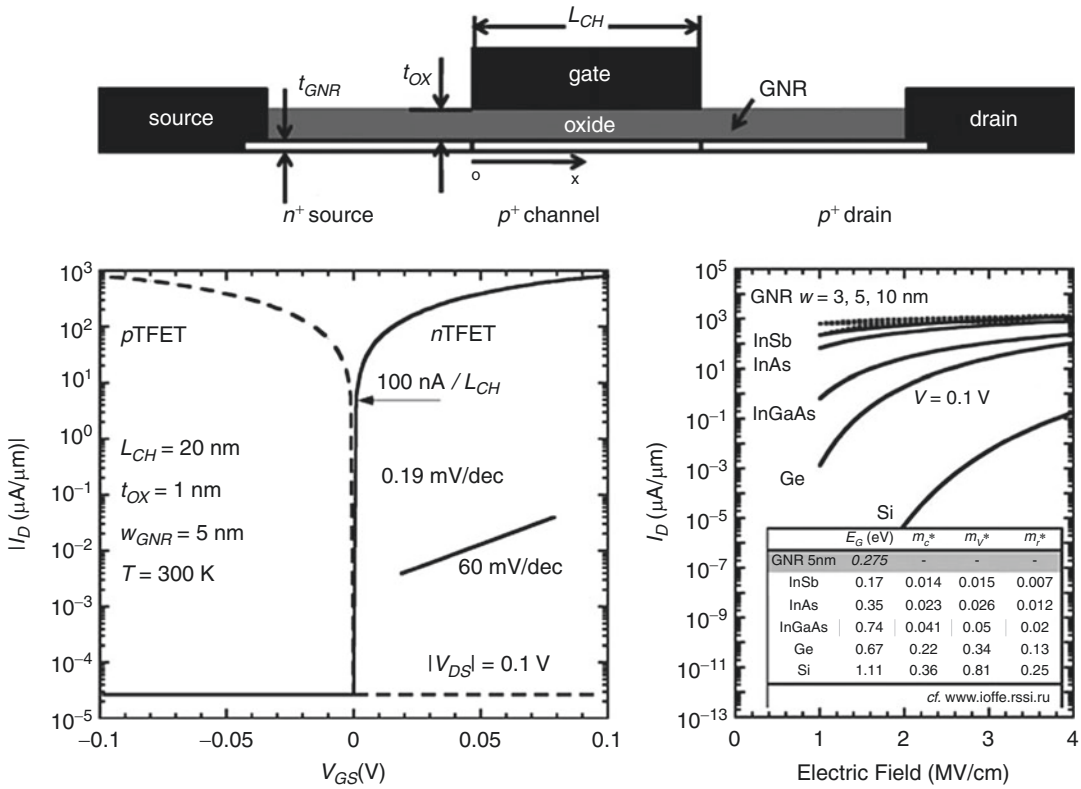
Graphene, Fig. 5 Bandstructure of bilayer graphene in the absence (**a**), and the presence (**b**) of an external field. Bilayer graphene is one of the first materials that has an electrically tunable energy bandgap

semimetal, where the conduction band minimum and the valence band maximum occur at different points in the k-space, and they overlap in energy. The semimetallic nature of graphite is responsible for its high current carrying capacity.

Graphene Nanostructures

When graphene sheets are "cut" into thin nanoscale ribbons, the size confinement results in the selection of a finite number of modes that can propagate freely along the axis of the ribbon. These structures are called graphene nanoribbons (GNRs) [11]. The edge of a GNR can be armchair or zigzag, and depending on the edge and the width, the allowed propagating modes can be metallic or semiconducting. In this aspect, GNRs are similar to carbon nanotubes (CNTs), except that the nanotube has no edges. A CNT is a rolled up GNR. Semiconducting GNRs have a bandgap $E_g \approx 1.4/W$ eV where W is the width of the GNR in nm. Since GNRs lie flat on substrates, they can dissipate heat efficiently, and are capable of high current drives. In addition, the existence of a bandgap implies they can be used for electronic switching. GNRs retain the symmetry of the bandstructure of the graphene sheet they are carved from, in the sense that the valence band of a GNR is a mirror image of the conduction band. Thus, semiconducting GNRs can be ambipolar, with a high current modulation. In the process of fabrication, if the edges of GNRs are rough, scattering from this disorder can reduce the electron mobility and carrier velocity. Nevertheless, GNRs can prove attractive for making electronic devices in the future.

If a p-n junction can be fabricated from GNRs, then the structure is very attractive for an unconventional field-effect transistor. Recently, a lot of research has focused on lowering the energy cost of digital switching by transistors. An attractive candidate for replacing conventional silicon field-effect transistor relies on the interband tunneling of electrons. The device is called a tunneling field-effect transistor or TFET; it promises digital switching of current at smaller voltages [12]. High current drives are required in transistors to maintain high performance.



Graphene, Fig. 6 Graphene nanoribbon–based tunneling field-effect transistors. Proposed device structure and predicted performance showing vast advantages over state-of-the-art silicon CMOS devices

The tunneling currents in traditional semiconductors are typically low due to the wide bandgaps. However, devices based on GNRs can deliver among the highest current drives among all homojunction TFETs [13]. Figure 6 shows the concept of the device, along with its predicted energy band diagram and switching performance. Owing to the symmetry of the bandstructure, a GNR TFET makes complementary operation feasible, a feature difficult to attain in similar devices made of traditional semiconductor heterostructures. In addition, the absence of out-of-plane bonds and the planar nature implies that GNRs and graphene are simpler to integrate with existing semiconductor technologies.

Optical Properties of Graphene

Interaction of light with graphene is understood from its energy spectrum. Single layers of graphene are found to absorb a well-defined fraction of visible light passing through them by interband excitation of electrons. This fraction is given by $\pi q^2/\hbar c = \pi \cdot \alpha \sim 2.3$ % where α is the fine structure constant [14]. The absorption coefficient is *independent* of the photon energy over a large spectral window. The photogenerated electrons and holes can lead to photocurrent, and aided by the high mobility, 2D graphene can be used for fast photodetectors that operate over a large spectral window. The ease of integration of graphene with traditional semiconductors makes it attractive for applications in optical communication devices.

Electrons in graphene can also absorb photons of very long wavelengths and small energy (in the THz spectral window). This happens by the process of free-carrier absorption, and requires momentum scattering, as in a traditional semiconductor. Since the number of photons absorbed is related to the circumference of the Fermi circle in

the k-space, by tuning the Fermi energy of graphene, the number of free carriers available for such absorption can be tuned over a large range. Thus, graphene is attractive for making THz modulators at a given frequency [15]. Further, graphene can be patterned into periodic structures similar to photonic crystals to enhance the absorption of certain modes and orientations of incident THz light. Such structures called graphene metamaterials are similar to photonic crystals made of dielectrics, and hold much promise for devices exploiting light–matter interaction.

Since graphene has zero-bandgap with a high mobility, an electromagnetic wave can create waves of electrons and holes that support its propagation. Such modes where free carriers on the surface of a conductor carry part of the electromagnetic energy of an electromagnetic wave are called surface plasmon modes. The high mobility leads to a low loss, and plasmons are thus useful for confining and controlling the flow of light in dimensions shorter than its wavelength. Such sub-wavelength waveguiding is attractive for on-chip optical communications in the future. This can enable transistors far from each other on a microprocessor chip, or between chips to exchange information by using waveguided light instead of electrons in conductors that can be lossy and have *RC* bandwidths.

Methods of Fabrication

Single-layer graphene flakes were first isolated by mechanical exfoliation from graphite. They were transferred on to substrates with ~300 nm SiO_2 on conductive silicon. The dielectric contrast of the substrate at optical wavelengths made it possible to observe the atomically thin crystal under a microscope. Since this initial breakthrough, several methods of preparing graphene have been developed [16]. Large-area single-layer graphene is formed on the surface of single crystal SiC wafers by sublimation. Under specific conditions, the few layers of Si and C on the surface dissociate, Si atoms leave the crystal, and the remaining carbon atoms rearrange to form graphene. A larger-scale method of fabricating single-layer graphene is by chemical-vapor deposition of carbon on heated copper (or other metallic) foils [17]. Owing to the low solubility of carbon in copper, it "precipitates" out, forming a single layer of graphene on the surface. The process is self-limiting, and leads to single-layer graphene of large dimensions. The graphene is subsequently transferred on to insulating substrates by removing the metallic foil. Growth and fabrication of graphene are still at their infancy. The methods used to grow graphene, and the resulting crystalline quality is expected to improve with time.

Other Applications

At present, the high conductivity of graphene and its ultrathin nature makes it attractive for transparent conductive electrodes for flexible electronics. The high in-plane thermal conductivity of graphene can be harnessed for effective heat spreading in various nanoscale structures; graphene can scale this usage of graphite down to nanoscale dimensions. The light mass of the atoms in graphene, the high bond strength, and atomically thin structure implies that microelectro-mechanical (MEMs) or nanoelectro-mechanical (NEMs) structures made from graphene can operate at very high frequencies, and extend the range of applicability of existing devices. The ease of field-effect of graphene also makes it especially conducive to the fabrication of gas and charge sensors in the field-effect geometry.

Other 2D Crystals and Future Directions

The symmetry of the bandstructure of graphene led to the proposal, and eventual discovery of a new class of materials called topological insulators [18]. These materials have a graphene-like Dirac cone energy dispersion for surface states, whereas the bulk states have energy bandgaps as in traditional semiconductors. The valley degeneracy of such materials is one, so in a sense they are "half" of graphene. Electron motion in the surface bands of topological insulators are

protected against scattering by geometric constraints, and are currently under much scrutiny.

The discovery of graphene also opened the door to the search for other 2D crystals, and a number of them have been identified. Among them, boron nitride (BN) is an excellent example. Boron and nitrogen straddle carbon in the periodic table, and is the compound semiconductor analogue of graphene. However, the broken symmetry in the crystal basis itself creates a very large bandgap, approaching ~5.4 eV. Thus, BN is an excellent insulator, and quite distinct from other insulating dielectrics since it is composed of sp^2 bonds. The absence of out-of plane or broken bonds implies that it is chemically and electronically more inert than its 3D insulating counterparts. Indeed, graphene when placed on BN in planar fashion exhibits high electron mobilities and this is attributed to the reduced electronic coupling between the two crystals. The transition-metal dichalcogenides (such as MoS_2) are an analogous family of 2D crystals that exhibit semiconducting as well as metallic properties. However, these 2D crystals are not perfectly planar, as their chemical bonds are different from sp^2; they involve d-orbitals. The presence of d-orbitals leads to a richer range of electronic phenomena, ranging from metallic, semiconducting [19], magnetic, to superconducting. Combining this rich range of electronic properties of 2D crystals in heterostructures is expected to lead to new generations of more efficient and fundamentally new devices in the future. Combining 2D crystals by stacking is expected to be simpler than 3D materials since there are no out-of-plane bonds to join, or break [20].

Cross-References

- Active Carbon Nanotube-Polymer Composites
- Active Plasmonic Devices
- Carbon Nanotube-Metal Contact
- Carbon Nanotubes for Chip Interconnections
- Carbon-Nanotubes
- Chemical Vapor Deposition (CVD)
- Flexible Electronics
- Fullerenes for Drug Delivery
- Light Localization for Nano-Optical Devices
- Modeling Thermal Properties of Carbon Nanostructure Composites
- Nanostructure Field Effect Transistor Biosensors
- Nanostructured Materials for Sensing
- Nanostructures for Photonics
- Nanotechnology
- Physical Vapor Deposition
- Terahertz Technology for Nano Applications
- Thermal Conductivity and Phonon Transport

References

1. Novoselov, K.S., Jiang, D., Schedin, F., Booth, T.J., Khotkevich, V.V., Morozov, S.V., Geim, A.K.: Two-dimensional atomic crystals. Proc. Natl. Acad. Sci. **102**, 10451–10453 (2005)
2. Wallace, P.R.: The band theory of graphite. Phys. Rev. **71**, 622–634 (1947)
3. Dean, C.R., Young, A.F., Meric, I., Lee, C., Wang, L., Sorgenfrei, S., Watanabe, K., Taniguchi, T., Kim, P., Shepard, K.L., Hone, J.: Boron nitride substrates for high-quality graphene electronics. Nat. Nanotechnol. **5**, 722–726 (2010)
4. Konar, A., Fang, T., Jena, D.: Effect of high-κ gate dielectrics on charge transport in graphene-based field effect transistors. Phys. Rev. B **82**, 115452–115458 (2010)
5. Cheianov, V.V., Fal'ko, V., Altshuler, B.L.: The focusing of electron flow and a veselago lens in graphene p-n junctions. Science **315**, 1252–1255 (2007)
6. Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Katsnelson, M.I., Grigorieva, I.V., Dubonos, S.V., Firsov, A.A.: Two-dimensional gas of massless Dirac fermions in graphene. Nature **438**, 197–200 (2005)
7. Zhang, Y., Tan, Y.-W., Stormer, H.L., Kim, P.: Experimental observation of the quantum hall effect and Berry's phase in graphene. Nature **438**, 201–204 (2005)
8. McCann, E.: Asymmetry gap in the electronic band structure of bilayer graphene. Phys. Rev. B **74**, 161403–161406 (2006)
9. Zhang, Y., Tang, T., Girit, C., Hao, Z., Martin, M.C., Zettl, A., Crommie, M.F., Shen, Y.R., Wang, F.: Direct observation of a widely tunable bandgap in bilayer graphene. Nature **459**, 820–823 (2009)
10. Banerjee, S.K., Register, L.F., Tutuc, E., Reddy, D., MacDonald, A.H.: Bilayer pseudospin field-effect transistor (BiSFET): a proposed new logic device. IEEE Electron Device Lett. **30**, 158–160 (2009)
11. Han, M.Y., Ozyilmaz, B., Zhang, Y., Kim, P.: Energy band-gap engineering of graphene nanoribbons. Phys. Rev. Lett. **98**, 206805 (2007)

12. Zhang, Q., Zhao, W., Seabaugh, A.: Low-subthreshold-swing tunnel transistors. IEEE Electron Device Lett. **27**, 297–299 (2006)
13. Zhang, Q., Fang, T., Seabaugh, A., Xing, H., Jena, D.: Graphene nanoribbon tunnel transistors. IEEE Electron Device Lett. **29**, 1344–1346 (2008)
14. Nair, R.R., Blake, P., Grigorenko, A.N., Novoselov, K. S., Booth, T.J., Stauber, T., Peres, N.M.R., Geim, A. K.: Fine structure constant defines visual transparency of graphene. Science **320**, 1308 (2008)
15. Sensale-Rodriguez, B., Fang, T., Yan, R., Kelly, M., Jena, D., Liu, L., Xing, H.: Unique prospects of graphene-based THz modulators. Appl. Phys. Lett. **99**, 113104–113106 (2011)
16. Berger, C., Song, Z., Li, X., Wu, X., Brown, N., Naud, C., Mayou, D., Li, T., Hass, J., Marchenkov, A.N., Conrad, E.H., First, P.N., de Heer, W.A.: Electronic confinement and coherence in patterned epitaxial graphene. Science **312**, 1191–1196 (2006)
17. Li, X., Cai, W., An, J., Kim, S., Nah, J., Yang, D., Piner, R., Velamakanni, A., Jung, I., Tutuc, E., Banerjee, S.K., Colombo, L., Ruoff, R.S.: Large-area synthesis of high-quality and uniform graphene films on copper foils. Science **324**, 1312–1314 (2009)
18. Hasan, M.Z., Kane, C.L.: Topological insulators. Rev. Mod. Phys. **82**, 3045–3067 (2010)
19. Radisavljevic, B., Radenovic, A., Brivio, J., Giacometti, V., Kis, A.: Single-layer MoS_2 transistors. Nat. Nanotechnol. **6**, 147–150 (2011)
20. Novoselov, K.S.: Nobel lecture: graphene: materials in the flatland. Rev. Mod. Phys. **83**, 837–849 (2011)

Graphene Laser Writing

► Nanostructures and Characteristics of Graphene and Graphene Oxides

Graphene Nanostructures for Memristive Devices

Samuele Porro
Applied Science and Technology Department, Politecnico di Torino, Torino, Italy

Synonyms

MEM devices; Memory resistors; Memristors; Resistive switching

Definition

Memristive nanoscale devices can generate intense fields by the application of relatively low voltages. This peculiar property allows fast, nonvolatile, and low-energy electrical switching as well as the possibility of retaining the internal resistance state according to the history of applied voltage and current. Memristors are predicted to revolutionize the current approaches in computer electronic architecture with their application, for instance, as resistive random access memory and for neuromorphic artificial intelligence. The use of graphene nanostructures for memristive switching systems offers an exciting alternative to other classes of materials, such as transition metal oxide and organic thin films.

Overview

The development of nanoscale structures that find applications in the fabrication of devices for computer science is going to revolutionize the semiconductor industry and in general the way in which computing systems are built. A continual improvement of device performances can be obtained by their miniaturization, and the development of nanometer-sized devices allows pushing the Moore's law beyond the actual physical limits, enabling a new era for computing. Due to their high scalability down to the nanometer scale and the possibility of application in nonvolatile memory and low-power resistive switching devices, memristors perfectly fit in this scenario. Although predicting and controlling the stochastic chemical/physical mechanisms that lie behind the memristive phenomena is currently a research challenge [1], many different working devices were already reported in the recent literature. Besides the various perovskite oxide materials, binary transition metal oxides (like TiO_2, Ta_2O_5, NiO), chalcogenides, and organic materials [2, 3], in the latest years, a growing attention was directed to carbon-based devices, in particular to those based on graphene and graphene oxide [4], due to the unique properties of this class of inherently nanoscale materials.

Classification of Graphene and Graphene Oxide

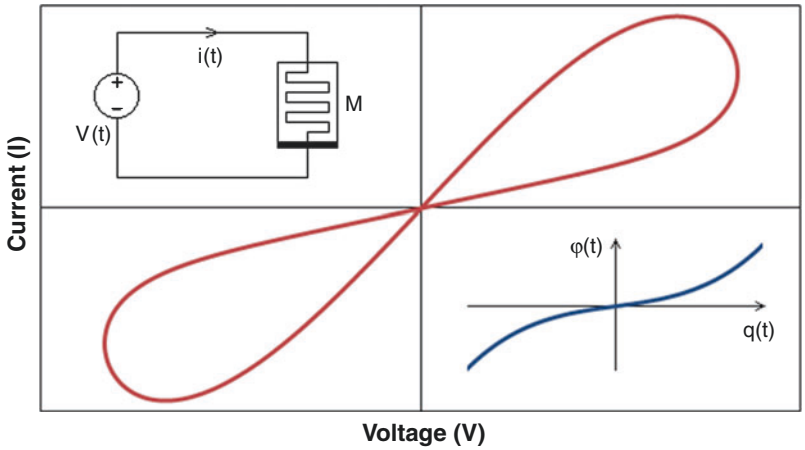
After the discovery of graphene, strikingly unconventional phenomena originated by the behavior of electrons as massless relativistic particles were predicted and measured [5]. These include interdisciplinary properties, such as excellent electron mobility at room temperature, high thermal conductivity, and superior mechanical properties [6], which make these materials particularly appealing for many applications, e.g., as active material in electronic devices such as sensors, batteries, supercapacitors, and energy devices [7, 8] or as a filler in the fabrication of multifunctional nanocomposite materials, in both polymeric or inorganic matrices [9–11].

Many approaches for the synthesis of graphene and graphene-based systems have been reported up to now. The most developed include chemical vapor deposition [12], epitaxial growth via thermal decomposition of silicon carbide [13], and mechanical or liquid phase [14] exfoliation of graphite. In addition, several methods for cost-effective and large-scale production of graphene were implemented following "chemical" processing routes, which involve the exfoliation of previously oxidized bulk graphite (the Hummer's method and many variation of the same) to obtain single or few layers of graphene oxide (GO) [15]. These routes allow low-cost and high-yield production of GO, which is highly hydrophilic and thus easily dispersible in organic solvents and even in simple water (unlike graphene). Many approaches for the reduction of GO to graphene-like sheets (reduced graphene oxide, RGO) were reported, including chemical and electrochemical processes, thermal annealing, and other unconventional methods such as microwave or photo-irradiation treatments [16–18]. These methods exploit the richness of reactive oxygen functional groups on graphene oxide surfaces, by removing the oxygen-containing groups with the recovery of a conjugated structure.

The intrinsic characteristics of GO make it a suitable candidate for the fabrication of memristive devices in substitution of other thin-layered oxide structures. For example, GO sheets are heavily oxygenated presenting functional groups on their basal planes (mainly hydroxyl and epoxide) and at the sheet edges (carbonyl and carboxyl) [19]. All those oxygen functionalities randomly placed over GO sheets are responsible for GO electrically insulating behavior, with characteristics comparable to other thin-layered oxide materials, with the advantage of being atomically thin. Moreover, the two dimensionality of GO permits scaling beyond the current limits of semiconductor technology, which is a key aspect for high-density fabrication of memristive devices. In fact, being GO truly atomically thin, the scalability of GO-based devices is better than that of organic and inorganic thin-film memristors. GO sheets can readily swell and disperse in water because the surface functional groups make it strongly hydrophilic. Exploiting this property, it is possible to synthesize GO- and RGO-based printable inks [20] which can be used for the direct 2D fabrication of memristors, for example, using inkjet printing technology which is greatly advantageous compared to other techniques, as already demonstrated by several works on polymer composites [21, 22].

Background of Memristors

The memristor can be defined as the fourth two-terminal circuit fundamental passive component, as postulated by Leon O. Chua already in 1971 [23] on the basis of a simple logic deduction. Prof. Chua noticed that the four fundamental circuit variables (current I, voltage V, magnetic flux φ, and charge q) should be invariably coupled by six different mathematical relations, while at the time only five were known. The first two were given by integrating over time voltage and current, obtaining, respectively, magnetic flux and charge. Other three relations were determined by the three classical circuit elements known at the moment: the resistor, connecting I and V; the capacitor, connecting q and V; and the inductor, connecting I and φ. A sixth missing relation connecting q and φ was postulated by Chua introducing the concept of memristor, indicating a resistor with memory, through the following equations:



Graphene Nanostructures for Memristive Devices, Fig. 1 The characteristic I-V curve for the ideal memristor, representing a hysteretic loop pinched in the origin. The circuital symbol for memristor is reported in the *top left inset*. The nonlinear relationship between q and φ characteristic of memristive devices is reported in the *bottom right inset* (Reprinted with permission from Ref. [26]; copyright © 2015 Elsevier)

$$V(t) = M(q(t))I(t)$$

with

$$M(q(t)) = \frac{d\varphi(q)}{dq}$$

where the memristance M is defined in units of Ohms. Even though at an instant time t_0 the memristor acts like a normal resistor, the memristance value depends on the time integral of its current from ∞ to t_0 and, therefore, on the past history of the system, explaining the concept of "resistor with memory" which characterize the memristor. In addition, the relation between the magnetic flux and the charge must not be linear; otherwise, the memristance would become a normal resistance, losing its peculiarity.

After several conceptual works from the same author [24], the concept of memristor remained undemonstrated, and a real working device was fabricated only recently in 2008 by HP Labs, with a work that related the theoretical idea of Prof. Chua with a physically working device made of titanium dioxide [25]. Following this milestone experimental work, many other memristive materials and device architectures were developed in the latest years. The number of materials showing this technologically interesting behavior is enormous and still growing, including binary transition metal oxides, e.g., TiO_2, Cr_2O_3, FeO_x, and NiO; perovskite-type complex transition metal oxides that are variously functional, paraelectric, ferroelectric, multiferroic, and magnetic, e.g., $(Ba,Sr)TiO_3$, $Pb(Zr_xTi_{1-x})O_3$, $BiFeO_3$, and $Pr_xCa_{1-x}MnO_3$; large band gap high-k dielectrics, e.g., Al_2O_3 and Gd_2O_3; non-oxide thin-layered materials, such as higher chalcogenides (e.g., In_2Se_3 and In_2Te_3); organic materials such as intrinsically conducting polymers, e.g., polyaniline; and, finally, also graphene oxide and graphene-based materials.

Following the fact that memristors are nanoscale devices, the application of relatively low voltages can generate very intense fields, which induce a nonlinear behavior and reversible ionic transport. The evidence of this effect is increasing by reducing the thickness of memristive materials.

Among the peculiar features of memristors, one of the most relevant is that if $M(q(t)) \geq 0$ the electronic component is passive, like the resistor, which implies low power consumption in a real device application. The characteristic I-V curve of memristors is a typical frequency-dependent Lissajous figure represented by a hysteretic loop pinched in the origin (Fig. 1).

From this characteristic it is evident that the memristor can be used as a nonvolatile memory, since for $V = 0$, and also $I = 0$, that is to say that

once a particular internal state is settled, it can be conserved indefinitely if no other input signals are provided. These proprieties make the memristor an extremely powerful element which is not reproducible with a combination of the other passive components. Memristors can find applications in many fields like resistive random access memories (RRAM) [27], programmable logic [28], neural networks, and neuromorphic systems [29]. In particular, memristors are predicted to revolutionize in the near future the current approaches in computer electronics architecture, and they are indicated as the first brick to create neuromorphic devices and artificial intelligence.

The memristor exhibits two different stable resistance states, referred to as "high-resistance state" (HRS) in which the device shows low electrical conductance and "low-resistance state" (LRS) in which the device shows high conductance. The system can make a transition from HRS to LRS (or vice versa) by applying a SET (or RESET) voltage V_{SET} (or V_{RESET}). Two types of switching can be defined: the bipolar switching, where SET and RESET processes depend on the polarity of the applied voltage, and the unipolar switching, where they depend on the amplitude of the applied voltage rather than on its polarity. Finally, the performance of a memristor can be expressed in terms of retention and endurance [30]. The retention indicates how long the internal state of a device can be held when it is not powered, with typical required values for nonvolatile memory applications exceeding 10 years. The endurance indicates how many cycles can be supported by the device before reaching breakdown, with typical values between 10^3 and 10^7 cycles depending on the application.

Memristive Devices Based on Graphene Nanostructures

Memristors can be fabricated using several device structure, among which the most common is the metal/insulator/metal (MIM) structure, which applies to an extremely wide range of insulating active layers sandwiched between two metal electrodes. As previously stated, graphene oxide can be easily dispersed in common polar solvents including simple water. The stable GO solutions can be easily deposited on bottom electrode materials using many techniques with a precise control of GO layer thickness (tens of nanometers), like spin coating, drop casting, vacuum filtration, inkjet printing, or Langmuir-Blodgett deposition, usually followed by thermal annealing to improve crystallinity and uniformity, at temperatures lower than 200 °C in inert environment to avoid GO reduction. The metallic top electrode is usually deposited by thermal evaporation or sputtering through a shadow mask.

The electrical behavior of the so-fabricated graphene-based memristors can be explained following two different mechanisms [31]. The first is related to an oxygen ion diffusion driven by the electrical field and Joule heating; the second is associated to the formation of a metallike filament due to the diffusion of species from the top electrode under a bias voltage.

Switching Mechanisms

The oxygen ion diffusion is considered as the main cause of the resistance switching, therefore of the memristive behavior, in devices based on graphene oxide. However, different interpretations of the physical phenomena can be given. In fact, the memristive behavior can be related to a bulk effect, meaning a structural modification of the entire GO layer. On the other hand, the phenomenon can be localized and associated to a modification of the contact resistance with the top electrode. In the switching mechanism based on the oxygen ions drift, the contact resistance between GO and the top electrode is of fundamental importance. The device is, as fabricated, in HRS due to the presence of a thick insulating metal oxide layer. The application of a negative bias induces a migration of oxygen ions from the oxide interface layer to the GO layer. The thickness reduction causes the device to switch to LRS.

Generally, fabrication parameters, such as the choice of deposition technique for GO and top electrode, choice of electrode materials, etc., influence the hybridization state of GO, contributing to determine differences in the behavior of

devices and influencing the choice of switching mechanism.

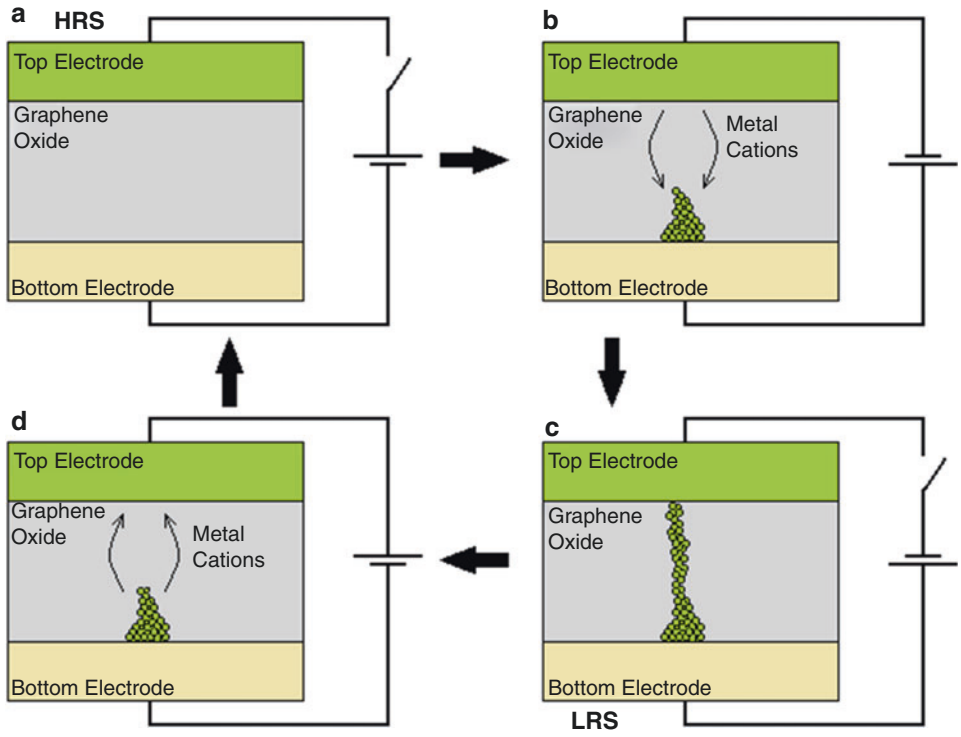
When GO is sandwiched between two metal electrodes (e.g., aluminum) [32], it can happen that the GO layer does not participate directly in the switching phenomena, which depend on the presence of a metal oxide-insulating film at the interface with the top electrode (which creates a contact resistance). In this case GO is relegated to the role of oxygen ions reservoir. The physical adsorption of oxygen in the GO structure can explain the switching events caused by the redox reaction between GO and metal oxide layer and also the migration of oxygen ions from and to the GO layer. The memristance switching can be explained as follows. The as-fabricated device shows a thick metal oxide interface layer (HRS). After the application of a negative voltage to the top electrode, the device switches to the LRS because of the formation of metallike conductive path in the metal oxide-insulating layer, originated from the oxygen ions diffusion into the GO film. By applying a positive voltage, the back diffusion of the ions from GO restores the original resistance state. The resistive switching can also be associated to a local reversible rearrangement of a highly resistive thin region of the GO film near the interface with the metallic top electrode, rather than to the presence of a metal oxide film. This area is mainly composed of a continuous heavily sp^3 hybridized GO layer, causing the as-fabricated device to be in HRS. In the second step, the application of an external voltage causes the migration of oxygen ions, providing the formation of a sp^2 hybridized cluster filament inside the highly resistive thin region, which allows switching to LRS. On the other hand, the resistive switching can be associated with the modification of the hybridization (oxidation) state of the entire GO layer, rather than to the interface between GO and top electrode. This happens when there exists a conducting path in the GO layer due to the presence of oxygen vacancies and electron traps. In this case, the memristor in its pristine form is in LRS. Through the application of an external bias, this initial state can be altered. Applying a positive voltage, the recombination between oxygen ions and vacancies induces further oxidation of the graphene layer thus the switching from LRS to HRS. By reversing the bias polarity, back diffusion of the oxygen ions is obtained, restoring the initial high conductive state.

The oxygen ion drift model could also explain the different speed of SET and RESET operations that are often measured in these type of devices, being SET speed (typical value of 100 μs) significantly slower than the RESET speed (typical value of 100 ns) [33]. The oxygen diffusivity is strictly related to the hopping energy barrier, which depends on the injection and extraction of electrons from the GO layer, which in turns is strictly related to the electron density in the insulating layer. In particular, when a negative bias is applied to the top electrode (RESET), electrons are injected into the GO, increasing the energy barrier, thus reducing the oxygen diffusivity and speed. In the same way, when the bias voltage is reversed (SET), the electrons are extracted from the GO, reducing the energy barrier, thus increasing the ions' movement speed.

In addition to the oxygen ion drift mechanism, the memristive behavior of GO-based devices can be explained by the formation of a metal filament within the insulating layer (Fig. 2) [34]. Following this mechanism, HRS and LRS states, respectively, show nonlinear and linear resistance variations respect to temperature. During the SET process, a positive bias voltage applied to the top electrode induces a high electric field inside the device, which causes the migration of the metal ions originating from the oxidation of the top electrode to the bottom electrode through the GO film. Once metal ions arrive to the bottom electrode, they amass inducing a reduction process through the insulating layer producing the growth of a metal filament. When the grown filament reach the top electrode, the resistance state of the device switches abruptly from the HRS to the LRS. An inversion of bias voltage causes the rupture of the filament with the back diffusion of the metal ions, restoring the HRS.

Characteristics of Devices

One of the most crucial aspects in the fabrication of MIM structures is the choice of materials for both top and bottom electrodes. Many different



Graphene Nanostructures for Memristive Devices, Fig. 2 (**a**) Scheme of the switching mechanism based on metal filaments formation, with device in its pristine state. (**b**) The application of a positive bias induces the oxidation of the *top* electrode metal, originating cations that migrate toward the *bottom* electrode, where after the reduction the formation of metal filaments is promoted. (**c**) When the grown filaments reach the top electrode, a short circuit is generated and the device switches from HRS to LRS. (**d**) Inverting the external bias, the opposite process is achieved (Reprinted with permission from Ref. [26]; copyright © 2015 Elsevier)

combinations of electrode materials can be used, without absolute certainty of measuring memristive properties [35]. Usually, the memristive behavior arises from an asymmetry of the electrode's working function, being one (active electrode) lower than the other (inert electrode). However, it is worth noticing that in some cases of material choice for the electrodes, local oxidation of the electrode at the interface might play a decisive role. The formation of an oxide layer under one of the electrode could also determine an asymmetry in structure. Another parameter that may influence the memristive characteristics is the surface roughness of GO layers after the deposition. Typically, a low surface roughness of GO is obtained after deposition on smooth bottom electrodes, like indium-tin oxide (ITO) and Pt. On the other hand, materials like Au and TaN exhibit higher roughness and surface cracks, which are quite detrimental to the device characteristics. In fact, the presence of cracks at GO surface can facilitate the penetration of the bottom electrode's atoms into the active layer and hinder the switching performance. This further underlines that the contact resistance can have a great importance in determining the switching proprieties of devices.

Bottom electrodes can be fabricated also using a semiconducting material like p-doped silicon in place of metal, thus realizing a MIS (metal-insulator-semiconductor) device. Semiconductors are intrinsically rectifying, having the advantage of suppressing cross talk between adjacent memory cells, without the need of introducing any additional rectifying element (e.g., diodes) like in many MIM devices.

Most devices show a bipolar switching behavior with voltages in the order of few volts, which is

desirable for low power applications, also displaying a great retention (higher than 10^4 s) and good endurance.

Although not all devices are activated through a forming process, the initial state of GO is not fixed on a specific state, and the forming process is often required to stabilize the switching operation and to determine the consequent electrical behavior (unipolar or bipolar) [36].

The memristive properties can be maintained on flexible substrate, e.g., on polyethersulfone, and can be used for electronic flexible application, without displaying any appreciable degradation of their behavior [37].

Graphene Oxide Composite Devices

Despite the good results that can be obtained with GO-based MIM structures, the device characteristics can be further improved by the development of innovative composite materials based on polymeric matrices filled with graphene or graphene oxide, which may improve not only the electrical but also the structural characteristics of future flexible devices. In this class of materials, the formation of a covalent bonding with the polymer is achieved exploiting the oxygen functional groups normally present on surfaces [38]. The composite materials are normally used as active layer sandwiched between conductive electrodes, exhibiting the classical bistable resistive switching. The memristive property of these materials is associated to an electron transfer between graphene (graphene oxide) and the polymer molecules [39]. In particular, as-fabricated devices are generally in the HRS, since the presence of functional groups (epoxide and carbonyl) produces a reduction of the electrical conductivity due to scattering events. By applying an external voltage with adequate amplitude, the transport of electrons from the polymeric molecules (highest occupied molecular orbital, HOMO) to GO (lowest unoccupied molecular orbital, LUMO) occurs by intramolecular charge transfer. This transfer results in delocalization of electrons in the π-conjugated system, reducing the initial GO to graphene (reduced graphene oxide, RGO) and increasing the device conductivity (LRS). By reversing the voltage, the electrons are extracted from the RGO layer and returned to the polymer, restoring the initial HRS. Both states are highly stable even if the external voltage is removed or a V_{read} is applied, and high retention time values can be normally obtained ($>10^4$ s). An alternative switching mechanism for this class of materials relies on charge-trapping events [40]. When carriers coming from the metallic top electrode are injected into the composite layer and localized by an energy well, they cause the formation of space charges. The generated internal electric field interferes with the external applied field, inhibiting further charge injection and leading to the HRS. When the external field is reversed, the charges are freed and the internal field disappears, and the state of the device returns to LRS. Finally, memristive properties can be measured also on functionalized GO sheets, sandwiched between two conductive layers as in the previously described examples [41]. In this case, the switching property is attributed to the movement of the counterions of functionalized groups when immersed in an electric field, while functionalized groups remain bounded to the GO framework. During LRS, those counterions are distributed along the insulating layer inducing a high conductivity. When the voltage is reversed, the resistance of the device is increased.

Similar to previously described GO-based devices, also devices based on GO/polymer composites show bipolar switching behavior. Generally, the resistance ratio values measured on GO/polymer composite devices are higher than in other types of GO-based devices, with typical values in the order of 10^5, while other characteristics do not show relevant differences. Their good retention values indicate them for employment as memristive devices in RRAM applications.

Other Graphene-Based Structures

In addition to the various structures so far presented, a great number of other memristive systems based on graphene or other carbon-based structures such as carbon nanotubes or amorphous carbon can be implemented [42].

For example, a few layered graphene deposited on a heavily doped silicon substrate can be used. After an initial forming process, which induces a soft breakdown in the structure, a physical

nano-gap within the graphene film can be obtained, setting the device in HRS. The unipolar resistance switching can be associated to the formation/rupture of a monoatomic carbon chain connecting the two edges of graphene separated by the nano-gap [43]. A different interpretation can also be given for this behavior. The applied voltage not only induces the rupture of the graphene film but also causes a morphological alteration of the underlying SiO_2 layer, with the formation of Si nanocrystals inside the underlying insulating layer. Therefore, the switching behavior can be also associated to the reversible thermal-assisted reduction/oxidation process at the breakdown region of the SiO_2 substrate, rather than to the formation of a carbon chain in the nano-gap. The presence of the Si nanocrystals brings the device to the LRS, while its absence in the SiO_x matrix entails the HRS.

Finally, GO can be also combined with other mixed oxide materials as active layers, to create multilayer structures. This combination ensured a bipolar switching behavior even using symmetric electrodes, for instance, in a metal/GO/mixed oxide/metal configuration. The insertion of mixed oxide layers also allows obtaining stable retention at relatively high temperatures (up to 85 °C) without degradation of materials. The exchange of oxygen ions between GO and the other insulating layer is considered to be the main cause for memristive switching.

Concluding Remarks

The purpose of this entry is to give an overview of the basic principles of memristor devices and the use of graphene-based systems for their implementation. The development of reliable memristive systems is a fundamental technological innovation that will find application in artificial intelligence as resistive random access memory and neuromorphic computing. Several examples of working memristive device based on graphene systems can be fabricated, exhibiting good switching proprieties (high switching ratio and low SET and RESET voltage) as well as good retention and endurance times.

Cross-References

▶ Graphene
▶ Synthesis of Graphene

References

1. Yang, J.J., Strukov, D.B., Stewart, D.R.: Memristive devices for computing. Nat. Nanotechnol. **8**, 13–24 (2013)
2. Yang, Y., Choi, S.H., Lu, W.: Oxide heterostructure resistive memory. Nano Lett. **13**, 2908–2915 (2013)
3. Chen, Y., Liu, G., Wang, C., Zhang, W., Li, R.W., Wang, L.: Polymer memristor for information storage and neuromorphic applications. Mater. Horiz. **1**, 489–506 (2014)
4. Chen, Y., Zhang, B., Liu, G., Zhuang, X., Kang, E.T.: Graphene and its derivatives: switching ON and OFF. Chem. Soc. Rev. **41**, 4688–4707 (2012)
5. Geim, A.K., Novoselov, K.S.: The rise of graphene. Nat. Mater. **6**, 183–191 (2007)
6. Zhu, Y., Murali, S., Cai, W., Li, X., Suk, J.W., Potts, J. R., Ruoff, R.S.: Graphene and graphene oxide: synthesis, properties, and applications. Adv. Mater. **22**, 3906–3924 (2010)
7. Fowler, J.D., Allen, M.J., Tung, V.C., Yang, Y., Kaner, R.B., Weiller, B.H.: Practical chemical sensors from chemically derived graphene. ACS Nano **3**, 301–306 (2009)
8. Sacco, A., Porro, S., Lamberti, A., Gerosa, M., Castellino, M., Chiodoni, A., Bianco, S.: Investigation of transport and recombination properties in graphene/titanium dioxide nanocomposite for dye-sensitized solar cell photoanodes. Electrochim. Acta **131**, 154–159 (2014)
9. Liu, G., Neoh, K.G., Kang, E.T.: Dispersible graphene oxide–polymer nanocomposites. In: Mittal, V. (ed.) Polymer-Graphene Nanocomposites, pp. 179–210. RSC Publishing, Cambridge, UK (2012)
10. Giardi, R., Porro, S., Chiolerio, A., Celasco, E., Sangermano, M.: Inkjet printed acrylic formulations based on UV-reduced graphene oxide nanocomposites. J. Mater. Sci. **48**, 1249–1255 (2013)
11. Porro, S., Giardi, R., Chiolerio, A.: Real-time monitoring of graphene oxide reduction in acrylic printable composite inks. Appl. Phys. A **117**, 1289–1293 (2014)
12. Reina, A., Jia, X., Ho, J., Nezich, D., Son, H., Bulovic, V., Dresselhaus, M.S., Kong, J.: Large area, few-layer graphene films on arbitrary substrates by chemical vapor deposition. Nano Lett. **9**, 30–35 (2008)
13. Virojanadara, C., Syväjarvi, M., Yakimova, R., Johansson, L.I., Zakharov, A.A., Balasubramanian, T.: Homogeneous large-area graphene layer growth on 6H-SiC(0001). Phys. Rev. B **78**, 245403 (2008)
14. Coleman, J.N.: Liquid exfoliation of defect-free graphene. Acc. Chem. Res. **46**, 14–22 (2012)

15. Dreyer, D.R., Park, S., Bielawski, C.W., Ruoff, R.S.: The chemistry of graphene oxide. Chem. Soc. Rev. **39**, 228–240 (2010)
16. Stankovich, S., Dikin, D.A., Piner, R.D., Kohlhaas, K. A., Kleinhammes, A., Jia, Y., Wu, Y., Nguyen, S.T., Ruoff, R.S.: Synthesis of graphene-based nanosheets via chemical reduction of exfoliated graphite oxide. Carbon **45**, 1558–1565 (2007)
17. Schniepp, H.C., Li, J.L., McAllister, M.J., Sai, H., Herrera-Alonso, M., Adamson, D.H., Prud'homme, R. K., Car, R., Saville, D.A., Aksay, I.A.: Functionalized single graphene sheets derived from splitting graphite oxide. J. Phys. Chem. B **110**, 8535–8539 (2006)
18. Roppolo, I., Chiappone, A., Bejtka, K., Celasco, E., Chiodoni, A., Giorgis, F., Sangermano, M., Porro, S.: A powerful tool for graphene functionalization: benzophenone mediated UV-grafting. Carbon **77**, 226–235 (2014)
19. Lerf, A., He, H., Forster, M., Klinowski, J.: Structure of graphite oxide revisited. J. Phys. Chem. B **102**, 4477–4482 (1998)
20. Le, L.T., Ervin, M.H., Qiu, H., Fuchs, B.E., Lee, W.Y.: Graphene supercapacitor electrodes fabricated by inkjet printing and thermal reduction of graphene oxide. Electrochem. Commun. **13**, 355–358 (2011)
21. Bocchini, S., Chiolerio, A., Porro, S., Accardo, D., Garino, N., Bejtka, K., Perrone, D., Pirri, C.F.: Synthesis of polyaniline-based inks, doping thereof and test device printing towards electronic applications. J. Mater. Chem. C **1**, 5101–5109 (2013)
22. Chiolerio, A., Bocchini, S., Porro, S.: Inkjet printed negative supercapacitors: synthesis of polyaniline-based inks, doping agent effect, and advanced electronic devices applications. Adv. Funct. Mater. **24**, 3375–3383 (2014)
23. Chua, L.O.: Memristor – the missing circuit element. IEEE Trans. Circuit Theory **18**, 507–519 (1971)
24. Chua, L.O., Kang, S.M.: Memristive devices and systems. Proc. IEEE **64**, 209–223 (1976)
25. Strukov, D.B., Snider, G.S., Stewart, D.R., Williams, R.S.: The missing memristor found. Nat. Lett. **453**, 80–83 (2008)
26. Porro, S., Accornero, E., Pirri, C.F., Ricciardi, C.: Memristive devices based on graphene oxide. Carbon **85**, 383–396 (2015)
27. Chua, L.O.: Resistance switching memories are memristors. Appl. Phys. A **102**, 765–783 (2011)
28. Borghetti, J., Snider, G.S., Kuekes, P.J., Yang, J.J., Stewart, D.J., Williams, R.S.: Memristive switches enable stateful logic operations via material implication. Nature **464**, 873–876 (2010)
29. Sah, M.P., Hyongsuk, K., Chua, L.O.: Brains are made of memristors. IEEE Circuit Syst. Mag. **14**, 12–36 (2014)
30. Waser, R., Dittmann, R., Staikov, G., Szot, K.: Redox-based resistive switching memories – nanoionic mechanisms, prospects, and challenges. Adv. Mater. **21**, 2632–2663 (2009)
31. Waser, R., Aono, M.: Nanoionics-based resistive switching memories. Nat. Mater. **6**, 833–840 (2007)
32. Nho, H.W., Kim, J.Y., Wang, J., Shin, H.J., Choi, S. Y., Yoon, T.H.: Scanning transmission X-ray microscopy probe for in situ mechanism study of graphene-oxide-based resistive random access memory. J. Synchrotron Radiat. **21**, 170–176 (2014)
33. Wang, L.H., Yang, W., Sun, Q.Q., Zhou, P., Lu, H.L., Ding, S.J., Zhang, D.W.: The mechanism of the asymmetric SET and RESET speed of graphene oxide based flexible resistive switching memories. Appl. Phys. Lett. **100**, 063509 (2012)
34. Zhuge, F., Hu, B., He, C., Zhou, X., Liu, Z., Li, R.W.: Mechanism of nonvolatile resistive switching in graphene oxide thin films. Carbon **49**, 3796–3802 (2011)
35. Hong, S.K., Kim, J.E., Kim, S.O., Choi, S.Y., Cho, B. J.: Flexible resistive switching memory device based on graphene oxide. IEEE Electron Device Lett. **31**, 1005–1007 (2010)
36. Hong, S.K., Kim, J.E., Kim, S.O., Cho, B.J.: Non-volatile memory using graphene oxide for flexible electronics. In: Proceedings of the 10th IEEE Conference on Nanotechnology (IEEE-NANO), pp. 604–606. (2010)
37. Jeong, H.Y., Kim, J.Y., Kim, J.W., Hwang, J.O., Kim, J.E., Lee, J.Y., Yoon, T.H., Cho, B.J., Kim, S.O., Ruoff, R.S., Choi, S.Y.: Graphene oxide thin films for flexible nonvolatile memory applications. Nano Lett. **10**, 4381–4386 (2010)
38. Liu, G., Zhuang, X., Chen, Y., Zhang, B., Zhu, J., Zhu, C.X., Neoh, K.G., Kang, E.T.: Bistable electrical switching and electronic memory effect in a solution-processable graphene oxide-donor polymer complex. Appl. Phys. Lett. **95**, 253301 (2009)
39. Zhuang, X., Chen, Y., Liu, G., Li, P.P., Zhu, C.X., Kang, E.T., Neoh, K.G., Zhang, B., Zhu, J.H., Li, Y.X.: Conjugated-polymer-functionalized graphene oxide: synthesis and nonvolatile rewritable memory effect. Adv. Mater. **22**, 1731–1735 (2010)
40. Wu, C., Li, F., Zhang, Y., Guo, T., Chen, T.: Highly reproducible memory effect of organic multilevel resistive-switch device utilizing graphene oxide sheets/polyimide hybrid nanocomposite. Appl. Phys. Lett. **99**, 042108 (2011)
41. Bhunia, P., Hwang, E., Min, M., Lee, J., Seo, S., Some, S., Lee, H.: A non-volatile memory device consisting of graphene oxide covalently functionalized with ionic liquid. Chem. Commun. **48**, 913–915 (2012)
42. Hwang, S.K., Lee, J.M., Kim, S., Park, J.S., Park, H.I., Ahn, C.W., Lee, K.J., Lee, T., Kim, S.O.: Flexible multilevel resistive memory with controlled charge trap band n-doped carbon nanotubes. Nano Lett. **12**, 2217–2221 (2012)
43. Standley, B., Bao, W., Zhang, H., Bruck, J., Lau, C.N., Bockrath, M.: Graphene-based atomic-scale switches. Nano Lett. **8**, 3345–3349 (2008)

Graphene NEMS

Changyao Chen
Argonne National Laboratory, Lemont, IL, USA

Synonyms

Carbon nanotube NEMS; NEMS; NEMS mass sensors; NEMS resonant mass sensors; RF MEMS/NEMS

Definition

Graphene nanoelectromechanical systems (NEMS) are devices that consist of single or multilayer graphene which is free to move along the out-of-plane direction and utilize the electrical and mechanical properties of graphene for various applications, such as mass or force sensing.

Overview

Graphene, as the first two-dimensional (2D) materials that can be explicitly prepared, offers both excellent electrical and superior mechanical attributes, both of which are desired for NEMS. Shortly after the first isolation of single-layer graphene in 2004, the field of graphene NEMS was spearheaded by researchers from Cornell University in 2007: they successfully actuated and detected the motion of doubly clamped graphene mechanical resonators. In the following years, advancements in fabrication, signal transduction, and modeling have been expanding the area of graphene NEMS. Applications of graphene NEMS include self-sustained mechanical oscillators, pressure sensors, and electromechanical switches.

A graphene NEMS device is usually composed of two components: a suspended piece of graphene sheet and electrical contacts to the graphene. The electrical connections are often used to tune the electrical or mechanical properties of graphene, as well as for signal transduction. When external stimulus is applied, such as pressure or electrostatic forces, graphene will deflect statically or vibrate resonantly. Different readout schemes are developed to track the response of graphene NEMS, including optical interferometry and electrical transduction.

Graphene obtained from conventional methods, such as mechanical exfoliation, chemical vapor deposition (CVD), and epitaxial growth, can all be used to construct graphene NEMS. There is no qualitative difference in NEMS properties among those devices.

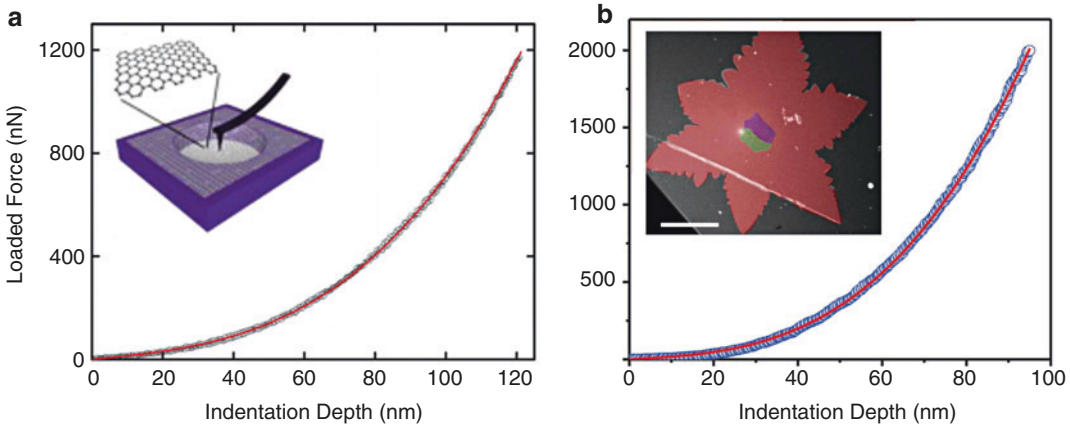
Unlike bulk materials, the thickness of the single-layer graphene cannot be measured by the distance between top and bottom surface, for which such surfaces are not well defined. However, in practice, researchers often use the interplane spacing of graphite, 0.34 nm, as the thickness of single-layer graphene. The lateral dimension of the suspended graphene, to this day, is usually from 1 to 10 μm. The separation between graphene and substrate ranges from tens of nanometers to few microns.

Various schemes can be applied to stimulate and detect the motion of graphene NEMS. Compared with other NEMS, for example, carbon nanotube (CNT)-based systems, graphene possesses the unique electrical and geometric advantages that allow simpler signal transduction. Some commonly used techniques such as capacitive transduction for MEMS resonators are demonstrated in graphene NEMS as well.

Graphene Mechanics

Since the thickness of graphene is at the atomic limit, both molecular level and continuum mechanics can be applied to model its mechanical properties. However, due to the computational expense of molecular simulations, the lateral dimension of the system is limited to tens of nanometers (few thousand carbon atoms), which is much smaller than any practical samples. Therefore, continuum mechanics with nonlinear considerations is more convenient to interpret experimental results and predict behaviors of graphene NEMS.

The most basic property of continuum mechanics is the relation between force and



Graphene NEMS, Fig. 1 Force vs displacement for circular-clamped graphene. (**a**). Experimental result of pristine graphene [1]. *Inset*: schematic of indentation setup. (**b**). Experimental result of CVD graphene [2]. *Inset*: false color dark field transmission electron microscopy (DF-TEM) images of CVD graphene; scale bar: 20 μm

displacement, commonly known as Hooke's law for bulk materials:

$$F = kx, \tag{1}$$

where F is the force exerted on the understudied subject, x is the corresponding deformation, and k is the spring constant. However, for graphene, profound nonlinearity emerges as the thickness approaches atomic limit; therefore, the relation between F and x takes the general nonlinear form of

$$F = Ax + Bx^3, \tag{2}$$

where A and B are linear and nonlinear coefficients, whose exact forms and values depend on the geometry and boundary conditions. For example, in the cases of doubly clamped strips and circular drums, Eq. 2 will take the form of

$$F = \left(30.78\frac{wt^2}{L^3}E^{2D} + 12.32\frac{t}{L}\sigma_0^{2D}\right)x + 2.67\frac{w}{L^3}E^{2D}x^3, \tag{3}$$

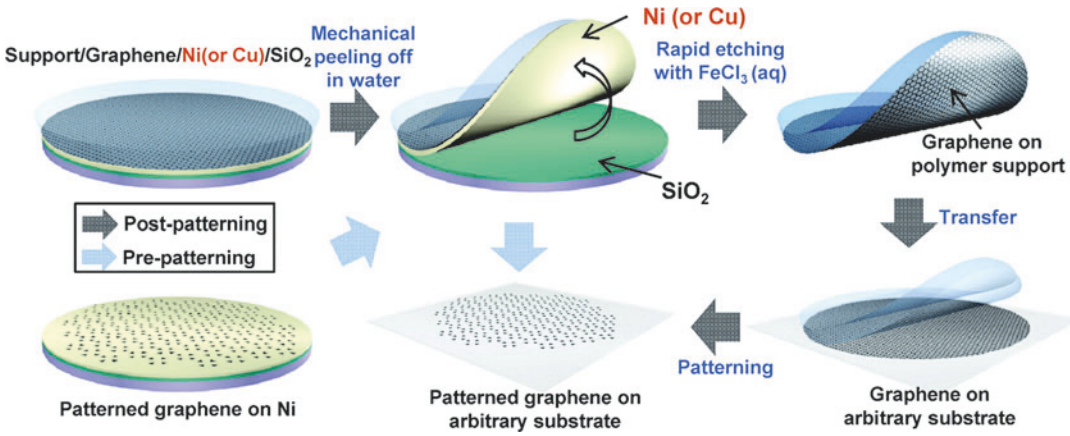
$$F = \left(1.37\frac{\pi t^2}{R^2}E^{2D} + \pi\sigma_0^{2D}\right)x + 1.06\frac{1}{R^2}E^{2D}x^3, \tag{4}$$

where w, t, and L are the width, thickness, and length of the strips, R is the radius of the drum, E^{2D} is the 2D Young's modulus, and σ_0^{2D} is the 2D pretension, both with the unit of force/length. Here, Poisson's ratio is taken as 0.165.

For exfoliated single-layer graphene, E^{2D} is experimentally determined [1] to be of 340 N/m (or equivalent 3D Young's modulus of 1 TPa), while the pretension is in the range of *O(1)* N/m. Polycrystalline graphene derived from chemical methods shows weaker mechanical strength, but if care is taken during the synthesis process, they can be as strong as pristine ones [2], as shown in Fig. 1.

From the above force versus displacement equations, it is clear that for samples with μm-size lateral dimensions, the bending rigidity (first terms with E^{2D} in the parentheses) is negligible compared to the pretension; therefore, they can be ignored. In doing so, the graphene can be modeled as membrane, whose mechanical properties such as resonant frequency are largely determined by tension/strain.

The truly unique mechanical property of graphene is its large breaking strength, i.e., the maximum strain it can sustain before breaks. In most materials, defects and flaws are abundant, and the cracks can grow and propagate from the



Graphene NEMS, Fig. 2 Schematic of transfer process of CVD graphene [6]

sites of defects easily; therefore, such imperfections dictate the maximum strength. A good example is the (intentionally introduced) notch on the plastic packages to facilitate tear opening. The inevitable existence of the defects in bulk materials weakens their breaking strength from intrinsic values of $E/9$, which is the strength of the atomic bonds [3]. Being atomically flat, graphene provides a platform of defect-free material at microscopic level. Indentation measurements, of both pristine and CVD graphene, have confirmed the breaking strength of graphene is indeed close to its intrinsic value. The breaking strength of graphene is 42 N/m, corresponding to a bulk value of 130 GPa. The measurements also reveal that the failure of graphene occurs deep into the nonlinear elastic regime, with a breaking strain of 25 %.

The large elastic modulus of graphene directly leads high resonant frequency of graphene NEMS (>10 MHz), but the atomic thinness (hence negligible bending rigidity) still allows facile frequency tuning through tension, and the large breaking strength ensures wide tuning range. Furthermore, the small mass of graphene also results in small spring constant (on the order of 1 N/m) while still maintaining high resonant frequency (high bandwidth). Such compliant system is desired for ultrasensitive applications, such as force and charge sensing.

Fabrication of Graphene NEMS

There are two major steps in graphene NEMS fabrication: graphene synthesis and release of the suspended structure. Historically, graphene can only be obtained through mechanical exfoliation from bulk graphite ("Scotch tape method"). After patterning electrodes by electron beam lithography, the silicon oxide layer below graphene is etched away by hydrofluoric (HF) acid, while the electrodes also serve as anchoring purpose. HF acid travels much faster along the interface between graphene and silicon oxide than the direction perpendicular to the interface; therefore, it is possible to release micron-size graphene structure with only tens of nanometers of graphene–substrate separation.

In 2009, few research groups developed graphene synthesis methods through CVD, using either nickel (Ni) [4] or copper (Cu) [5] as precursors, therefore enabling wafer-scale production. Cu-based CVD become more popular due to high yield and high quality. After growth, an intermediate support such as polymethyl methacrylate (PMMA) or polydimethylsiloxane (PDMS) is applied on top of the as-growth graphene/Cu foil, before the whole support/graphene/Cu foil combo is immersed into copper etchant to release the support/graphene stack. Finally, the support/graphene stack is stamped against target substrate,

followed by removal of support medium. Figure 2 shows the generic process flow of the transfer procedures. The steps to pattern and suspend CVD graphene are similar to previous described in the case of pristine graphene.

Compared to the aforementioned top-down methods, there are also bottom-up methods to prepare suspended graphene structure. In such cases, target substrates with pre-patterned trenches or holes are prepared in advance. The graphene is then placed onto the intended location with procedures similar to the CVD graphene transfer. The advantage of these top-down approaches is that the target substrates can be made through conventional photolithography process, therefore improving the device throughput. However, it is difficult to include additional wet process after the transfer without collapsing the graphene; therefore, the shapes of graphene need to be patterned beforehand.

There are hybrid techniques between the top-down and bottom-up approaches: the target substrate is pre-patterned with designed structures, but without etched trenches of holes. Only after graphene transfer and additional lithographic patterning, the graphene is finally released [7]. Such hybrid scheme combines the advantages of both top-down and bottom-up methods, but the fabrication is more laborious.

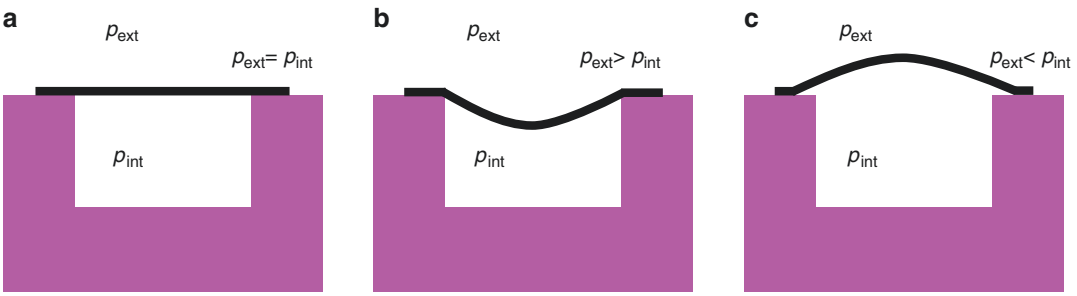
Static Graphene NEMS

Static graphene NEMS devices track the steady-state response of graphene after static deformation, where the deformation is introduced by external perturbations, such as force or pressure. These experiments are usually carried out in order to study fundamental properties of graphene; examples include indentation tests for the measurements of mechanical modulus [1, 2, 8] and "balloon/blister" tests to determine the adhesion between graphene and underlying substrate [9].

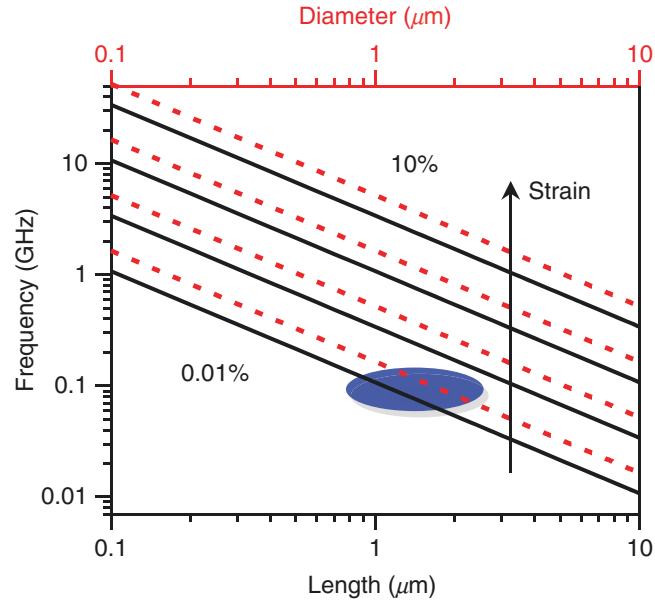
In the "balloon/blister" tests, as shown in Fig. 3, graphene and the underlying substrates form a microcavity, with gas molecules trapped inside. Since graphene is impermeable to all standard gases [10], gas can only escape the cavity through substrate, which is a slow process. Depending on the difference between the external and internal pressure of the cavity, the suspended graphene can bend either inward or outward. If the graphene is buckling outward, part of it will be peeling off away from the edge of the substrate. The peeling rate reveals that adhesion between graphene and the substrate is ultra strong, acting like liquid–solid-like interaction [9].

Applications of static graphene NEMS exploits the impermeability of graphene and the piezoresistive effect. Smith et al. demonstrated pressure sensors built from single-layer graphene [11], with gauge factor of 2.92. The device structure is very similar to the graphene balloon described above, with the cavity dimension of 64 (width) × 6 (length) × 0.64 (depth) μm, and the active area (graphene) is 64 × 6 μm.

There are preliminary studies on graphene switches [12, 13], showing switching voltage



Graphene NEMS, Fig. 3 Graphene balloon/blister, (**a**). External pressure p_{ext} equals internal pressure p_{int}, (**b**). $p_{ext} > p_{int}$, where graphene is bent inward, (**c**). $p_{ext} < p_{int}$, where graphene is buckled outward. Note that graphene is peeled from the edge of the substrate



Graphene NEMS, Fig. 4 Resonant frequencies for doubly clamped (*black solid line*) and circular dram (*red dashed line*) graphene resonators as a function of size, for strains of 0.01 %, 0.1 %, 1 %, and 10 %. The *blue* regime shows areas with experimental demonstrations [14]

smaller than 5 V. However, the reliability of such devices is not satisfying, with life cycle less than 30.

Dynamic Graphene NEMS

Given the high Young's modulus, small size, and mass, graphene is the excellent candidate for resonant NEMS. The resonant frequency of graphene resonator f_0 is given as [14]

$$f_0 \propto \frac{1}{\ell}\sqrt{\frac{E^{2D}}{\rho^{2D}}\varepsilon}, \tag{5}$$

where ℓ is the characteristic length, E^{2D} is the 2D Young's modulus, $\rho^{2D} = 7.9 \times 10^{-19}$ kg/μm^2 is the 2D mass density of single-layer graphene, and ε is the strain across the graphene sheet.

Figure 4 shows the calculated resonant frequency for doubly clamped and circular drum graphene resonators, for different sizes with various strains. For micron-size devices, it is possible to reach resonant frequency higher than 1 GHz with strain about 1 %, which is still far below the intrinsic breaking point. However, there is no deterministic method to manipulate the strain in graphene resonators yet, and the strain is largely determined from fabrication process. The blue shadow in Fig. 4 indicates the range of existing experimental achievements.

Signal transduction. The most challenging task of dynamic graphene NEMS lies in signal transduction, namely, (1) how to excite the resonant motion and (2) how to convey the mechanical vibration to easily measurable quantities (such as voltage or current) with sufficient signal-to-noise ratio (SNR).

Among these two, the actuation is relatively straightforward: graphene is being suspended over the substrate (usually conductive); therefore, they form the structure similar to parallel-plate capacitors, with geometric capacitance C_g. When both DC voltage V_g^{DC} and AC voltage $v_g^{ac}(\omega)$ at frequency ω are applied across the capacitor, the electrostatic interaction results in force exerted on graphene of

$$F \propto C_g' v_g^{ac}(\omega) V_g^{DC}, \tag{6}$$

where C_g' is the spatial derivative of the capacitance. Such capacitive drive scheme is commonly used in other MEMS and NEMS devices and works very well for graphene NEMS too. In the case when electrostatic actuation is not applicable, photothermal effect and piezo shaker (of the

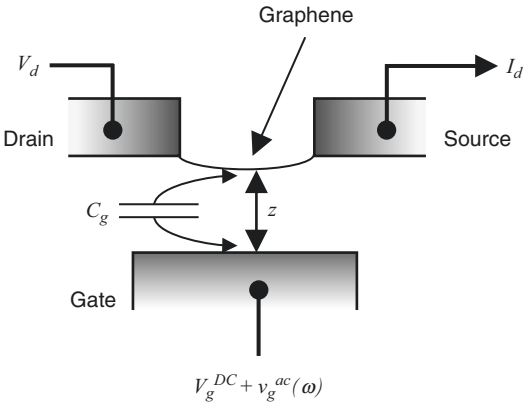
whole sample) can also be used to induce mechanical vibration.

Contrary to the simplicity of actuation, the detection of resonant graphene NEMS is more difficult. Both optical and electrical techniques are employed, with details discussed below.

The optical implementation of graphene NEMS detection exploits the Fabry–Pérot cavity formed between the suspended graphene and the underlying substrate. A laser beam is directed perpendicular to the graphene membrane, part of which is reflected by graphene, with the rest being passed through and being reflected again by the substrate. The interference pattern of these two reflected signals, recorded by a fast photodiode, is sensitive to the cavity length, hence the mechanical motion of the graphene. This interferometry technique is used to demonstrate the very first graphene mechanical resonator [15].

Albeit conceptually simple, the setup of optical detection is rather complicated and stringent, and it is almost impossible to realize such scheme at integrated circuit (IC) level. Therefore, all electrical readout techniques are more preferable. However, the conventional capacitive transduction approach cannot be directly applied to graphene NEMS, due to the overwhelming parasitic capacitance (in the order of pF) compared to the active capacitance. For example, the geometric capacitance for 1 by 1 μm^2 graphene with 300 nm vacuum gap is 30 aF, which is about five orders of magnitude smaller than the parasitic capacitance. Without active noise reduction mechanism, it is very difficult to obtain adequate SNR to resolve the resonance.

The electrical detection of graphene resonator utilizes its semiconducting properties: in the schematic setup in Fig. 5, when a gate electrode is nearby to modify the carrier density n of graphene, its conductance G will change accordingly. The magnitude of such variations is enhanced by the large electronic mobility of graphene. Given the "parallel-plate capacitor" structure, the mechanical resonance will alter the geometric capacitance C_g, hence the carrier density through $Aen = C_g V_g^{DC}$, where A is the size of the graphene and e is the electron charge.



Graphene NEMS, Fig. 5 Schematic of electrical readout of graphene resonators

Therefore, the mechanical motion is imprinted to the time-varying current across graphene as

$$\frac{dI_d}{dz} = V_d V_g^{DC} \frac{dG}{dV_g^{DC}} \frac{C_g'}{C_g}. \tag{7}$$

In the above equation, dG/dV_g^{DC} is proportional to the electron mobility, which in high-quality graphene can be orders of magnitude larger than other NEMS constituents.

The first demonstration of all-electrical transduction of graphene resonators [16], however, did not directly profit from Eq. 7. The authors used conductive silicon wafer as global back gate and therefore introduced inevitable large parasitic capacitance between the metal bond pads and the silicon wafer. To circumvent this limitation, the authors employed amplitude modulation (AM) frequency mixing method to down-convert the radio-frequency (RF) signal to low-frequency range ($\sim$ kHz). The amplitude of the low frequency still preserves the mechanical signature. Variations of such heterodyne techniques have since been applied for electrical detection, such as frequency modulation (FM) [17] and frequency up-conversion [18].

In order to capitalize on the simple relation outlined in Eq. 7, the critical step is to design the device structure that would minimize the capacitive cross talk. Xu et al. achieved this by employing local backgate on top of insulating

silicon substrate [19] and hence reduce the parasitic capacitance by two orders of magnitude. The RF current through the device is given as

$$I_d = j\omega C_g \left(\frac{C_{\text{tot}}}{C_g} v_g^{ac}(\omega) - V_g^{DC} \frac{C_g'}{C_g} \delta z \right) + V_d \frac{\mathrm{d}G}{\mathrm{d}V_g^{DC}} \left(v_g^{ac}(\omega) - V_g^{DC} \frac{C_g'}{C_g} \delta z \right), \quad (8)$$

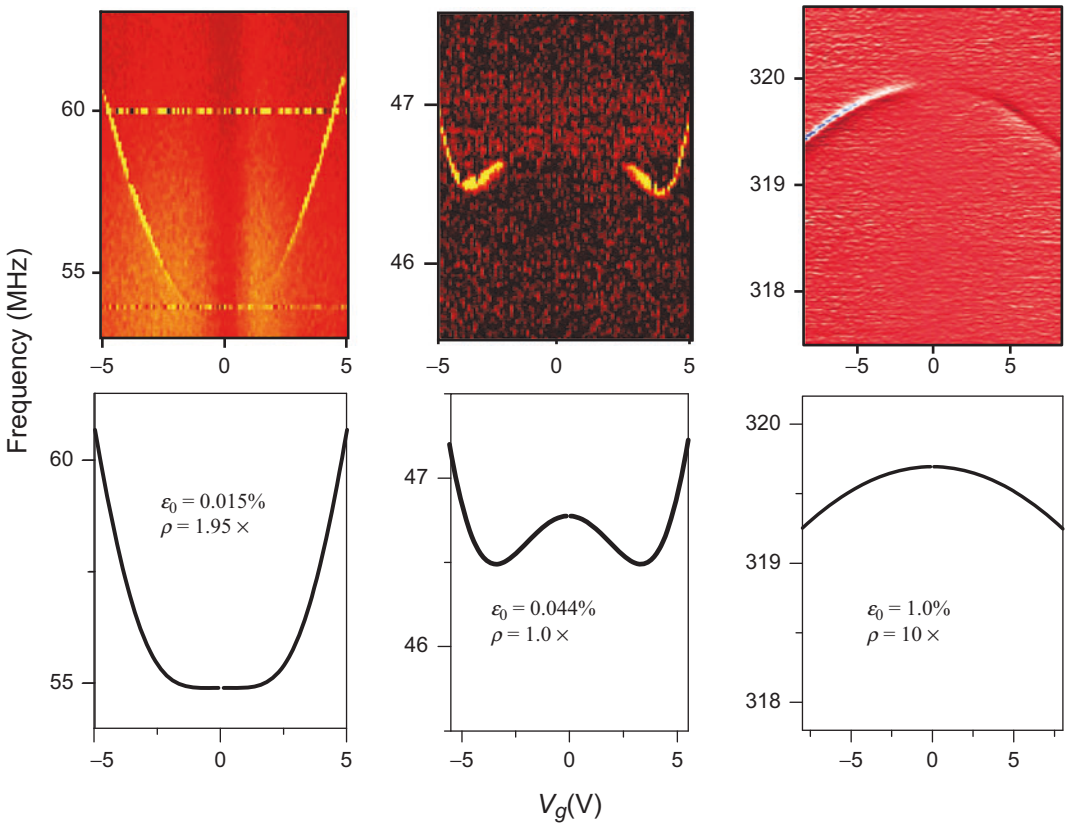
where $j = \sqrt{-1}$, C_{tot} is the total capacitance of the system, including both parasitic and motional components, and δz is the vibration amplitude of the graphene membrane. Terms in the first parentheses originate purely from capacitive transduction, which are the same as the electromechanical current used in silicon-based RF MEMS. Because of the small lateral size of the graphene sample, this current is extremely small here. The two terms in the second parentheses are the immediate consequence of Eq. 7, together with electrical feedthrough. The total current I_d at different frequency ω is monitored with a vector network analyzer (VNA) and does not require any additional heterodyne mixing. The advantage of such resonant channel transistor configuration (graphene acts as the active channel) is the improvement of measurement speed, by more than three orders of magnitude compared with the frequency mixing methods.

Frequency tuning. Due to the atomic thinness, the bending rigidity of graphene can be ignored; hence, the resonant frequency is dictated by in-plane strain, according to Eq. 5. Additionally, its ability to withstand extreme large strain before breaking opens up the potential for wide-frequency tuning through strain engineering: frequency tuning up to 400 % has been demonstrated [16]. A detailed analysis [20] shows that even at atomic limit, one can still apply continuum elastic mechanics to model the dynamic response. In typical device arrangements, where the suspended graphene is partially or fully clamped, its in-plane strain can be modified through electrostatic attraction by different V_g^{DC}. For example, the spring constant k of a doubly clamped graphene resonator can be described as

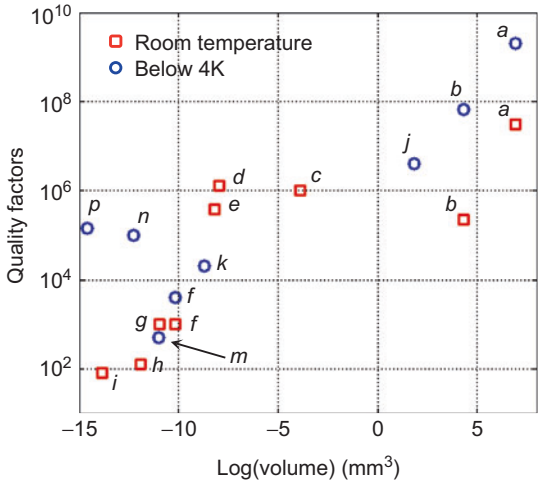
$$k = \frac{16E^{2D}w}{3L}\varepsilon_0 + \frac{256E^{2D}w}{3L^3}\left[z_e\left(\varepsilon_0, V_g^{DC}\right)\right]^2 - \frac{1}{2}C_g''\left(V_g^{DC}\right)^2, \quad (9)$$

where w and L are the width and length of the suspended portion, ε_0 is the built-in strain, z_e is the maximum static deflection of graphene at equilibrium, and C_g'' is second spatial derivative of geometric capacitance. There are three contributions in Eq. 9: the first term is governed by built-in strain and is independent of V_g^{DC}. The second term reflects the effect of external strain induced by applied V_g^{DC}: with larger $|V_g^{DC}|$, the membrane will bend more toward the substrate; hence, z_e will increase. Note that the growing rate of z_e is also determined by the built-in strain ε_0. The third term describes the nonlinearity of electrostatic interaction, which will introduce spring constant softening with larger $|V_g^{DC}|$. This softening effect is commonly observed and well understood in MEMS resonators. With small built-in strain ε_0, the second term in Eq. 9 dominates, and the resonant frequency increases monotonically with $|V_g^{DC}|$; for large ε_0, the third term in Eq. 9 prevails, and the resonance decreases monotonically with $|V_g^{DC}|$. In the case of intermediate ε_0, the resonant frequency first decreases and then increases with $|V_g^{DC}|$. All the three scenarios are observed in experiments and agree well with the model prediction (Fig. 6).

Quality factor. Another important property of resonant systems is the damping rate, and it is commonly quantified by quality factor $Q = 2\pi E/\Delta E$, where E is the time-averaged energy stored in the resonator and ΔE is energy lost per cycle. Large values of Q are preferred for applications such as mass sensing and communication. However, as the size of the resonators shrinks toward nanometer size, the quality factors usually suffer, as shown in Fig. 7. The decrease in Q is mostly due to the higher surface energy dissipation, as the surface-to-volume ratio in NEMS is much bigger than MEMS. For doubly clamped graphene resonators with length less than 5 μm, the quality factors are about 10^2 at room temperature, for both mechanically exfoliated and CVD samples [16, 17]. However, the quality

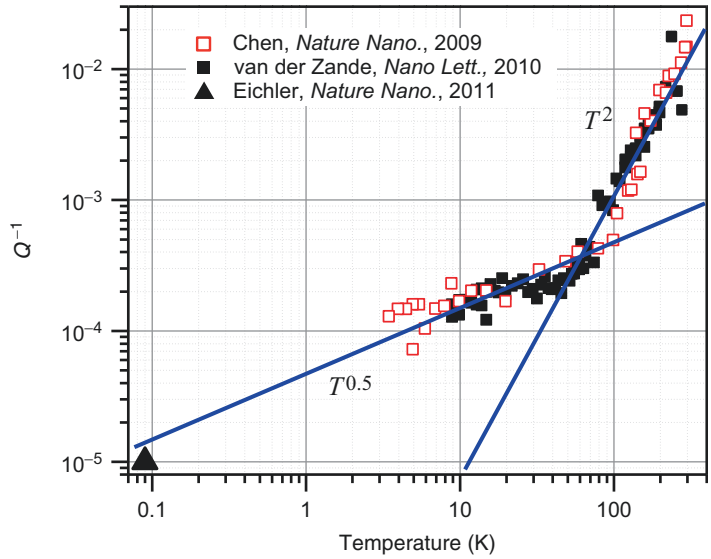


Graphene NEMS, Fig. 6 Frequency tuning of doubly clamped, micron-size graphene resonators by V_g^{DC}. *Top panel*, experimental data; *bottom panel,* continuum elastic model fitting, with only mass density and built-in strain as free parameters. Here the mass density is given as multiple of the intrinsic density of graphene [14]



a. McGuigan, *J. Low Temp. Phys.*, 1978
b. Duffy, *J. Appl. Phys.*,1990
c. Mihailovich, *Sensor Actuat. A*, 1995
d. Verbridge, *Appl. Phys. Lett.*, 2008
e. Gavartin, *Nature Comm.*, 2013
f. Hutchinson, *Appl. Phys. Lett.*, 2004
g. He, *Nano Lett.*, 2008
h. Chen, *Nature Nano.*, 2009
i. Sazonova, *Nature*, 2004
j. Klitsner, *Phys. Rev. B*, 1987
k. Cleland, *Appl. Phys. Lett.,* 1996
m. Huang, *Nature*, 2003
n. Eichler, *Nature Nano.,* 2011
p. Huttel, *Nano Lett.,* 2009

Graphene NEMS, Fig. 7 Survey of quality factors of NEMS with different sizes



Graphene NEMS, Fig. 8 Inverse quality factor Q^{-1} of graphene resonators at different operating temperatures, collected from different experiments [16, 17, 22]. *Blue* lines are merely guide to the eyes, showing different temperature dependencies

factors for circularly clamped drum-type graphene resonators show values as high as 2500 at room temperature [21]. Although the size of the drum resonators is bigger in this case (diameter of up to 25 μm), the larger values of Q still point the direction for improving quality factors by removing unclamped edges.

The quality factors of graphene resonators showed drastic increase upon cooling, reaching values higher than 10^5 when operated below 100 mK [22–24]. Its dependence on temperature, collected from experiments conducted by different research groups, is shown in Fig. 8. Between 100 K and room temperature, the energy dissipation Q^{-1} scales roughly as T^2, and below 100 K, the temperature dependence changes to approximately $T^{0.5}$. Similar trends at low temperatures [14] have been observed in carbon nanotube resonators and other NEMS resonators made from diamond and gallium arsenide; however, the explicit mechanism that can well predict the experimental results is still lacking.

The strong temperature dependence above 100 K is likely caused by thermal expansions of substrate. The samples shown in Fig. 8 are all made through HF acid under etching, as the graphene-aided diffusion of HF results in suspended metal contacts. When the temperature drops, these overhanging metals experience thermal contraction, imposing strain to the attached graphene membrane. While the in-plane strain only weakly affects the energy dissipation, it will increase the resonant frequency remarkably, hence boosting the apparent quality factor. The thermal expansion coefficients of metal become much smaller below 100 K, and the intrinsic dissipation mechanisms start to dominate.

Applications

Mass sensing. Due to the low mass and extremely large surface-to-mass ratio, graphene resonators are well suited for mass sensing applications. In general, the resonant frequency ω_0 can be written as

$$\omega_0 = \sqrt{\frac{k}{m_{\text{eff}}}}, \tag{10}$$

where m_{eff} is the effective mass. Therefore, the frequency shift $\delta\,\omega_0$ is

$$\delta\,\omega_0 = \frac{1}{2}\frac{\omega_0}{k}\delta k - \frac{1}{2}\frac{\omega_0}{m_{\text{eff}}}\delta m. \tag{11}$$

For bulk MEMS-based resonators, the added mass δm has little or no effect on spring constant (first term in Eq. 11); therefore, the result of mass loading is purely downshift of ω_0. However, this assumption is no longer valid for graphene resonators: the deposition of adsorbates interacts

strongly with graphene, and it causes an increase of strain, therefore resonant frequency. Study has shown that [16] at low strain limit, the resonant frequency indeed increases upon mass loading, while at high strain limit, the resonant frequency downshifts as expected. The experimentally demonstrated mass responsivity, $\Re_m = \frac{1}{2\pi}\delta\omega_0/\delta m$, which quantifies how sensitive the resonant frequency to mass change, is 1.3 Hz/zg at low strain limit and −2.1 Hz/zg at high strain limit.

The minimal resoluble mass $\delta m_{\min}$, on the other hand, also depends on the dynamic range (*DR*, in unit of dB), quality factor (Q) of the underlying resonator, and the measurement bandwidth *(BW)*. It can be approximated as

$$\delta m_{\min} \approx \Re_m^{-1} \times 10^{-DR/20}\sqrt{\frac{\omega_0}{Q}BW}. \qquad (12)$$

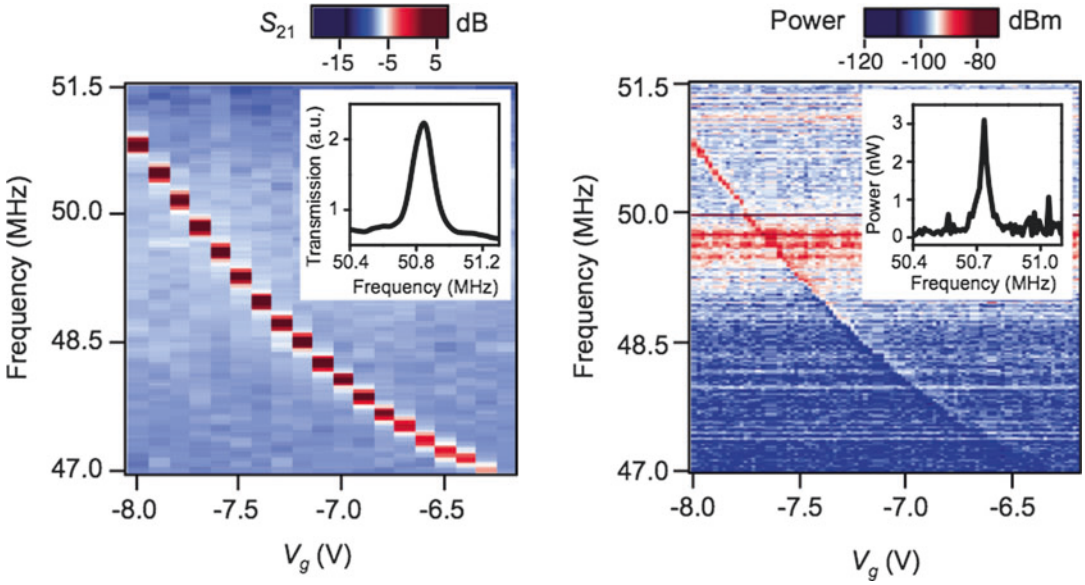
For a graphene resonator with resonance below 10 MHz and Q larger than 1000 at room temperature, the equivalent mass sensitivity is about 0.1 zg/Hz$^{0.5}$. Improvements of quality factors and strain engineering are expected to further enhance the mass sensitivity.

It is worth mentioning that sensors made from carbon nanotube have shown superior mass resolution [25]; however, the detection mechanism requires cryogenic environment and the measurement bandwidth is limited at kHz range. The simple implementation of graphene resonators allows for practical applications at room temperature that could achieve comparable sensitivity.

Self-sustained oscillators. The resonator structures that are considered so far all require an external AC stimulus for actuation. On the contrary, an oscillator structure that consists of a passive resonator and an active feedback component can be operated without such prerequisite. The mechanical motions are initiated by thermal noise, fed into the positive feedback loop, and the amplified output of the feedback loop is applied back to the resonator as actuation, completing the closed loop. This "ring-up" of the vibration will continue until certain nonlinearities, either of the resonator or of the amplifier, start to provide negative feedback, after which the system will oscillate around a given frequency. During these processes, only DC power supplies are needed to compensate the energy lost during each oscillation cycle, and the oscillator will return a stable AC signal output.

Barton et al. [26] have recently demonstrated self-oscillation of graphene mechanical resonators, using photothermal back-action as feedback scheme. The graphene and the substrate form a Fabry–Pérot cavity as described previously, while graphene serves as the "moving mirror." While the cavity length is modulated by the vibration of graphene membrane, so is the absorbed power of the incident laser by graphene. The absorption of light causes periodic heating and therefore changes in strain, giving rise to an induced optomechanical force parallel to the direction of vibration but with certain time delay. The time delay can be tuned by the equilibrium cavity length and laser wavelength, while the amplitude of the optomechanical force is controlled by the laser power. If the time delay is chosen such that the optomechanical force acts in-phase with damping, it will suppress the thermal motion of the membrane, resulting in optomechanical cooling. The authors showed that they are able to cool the fundamental mode of a room-temperature graphene resonator to an effective temperature of 100 K, with 2 mW of laser power. On the other hand, if the time delay causes the optomechanical force and the damping to be out of phase, it will decrease the effective energy dissipation, giving rise to larger apparent Q and higher vibrational amplitude. At sufficiently high optomechanical force, this negative damping is able to compensate for the internal energy loss, therefore inducing self-sustained oscillation without AC driving force. The oscillation frequency can also be tuned by V_g^{DC}, as shown by the authors.

Parallel to this all-optical implementation, as in the case of graphene resonators, Chen et al. [27] later reported the all-electrical realization of graphene mechanical oscillators, as shown in Fig. 9. The authors used regular radio-frequency amplifiers and phase shifter in their feedback loop, to provide adequate amplitude gain tuning and time delay control. Aside from the voltage-controlled oscillation, the benefits of electrical transduction enable direct signal processing. In this case, the



Graphene NEMS, Fig. 9 *Left panel*: open-loop response (resonator setup) shows frequency tuning through applied gate voltage V_g. *Inset*: transmission spectrum at $V_g = -8$ V. *Right panel*: closed-loop response (oscillator setup) of the same device, demonstrating the identical frequency tuning characteristics. *Inset*: power spectrum of the oscillation at $V_g = -8$ V (Figure adapted from Chen et al. [27])

graphene oscillators served as frequency modulator (FM), which up-converted audio waveform into commercial FM broadcast band. The transmitted signal was successfully demodulated and the original song track was recovered.

Optomechanical coupling with graphene resonators. The high-quality factors Q of graphene resonators (>100,000) at low temperatures triggered interests of exploiting graphene for applications in quantum regime. Recently, Weber et al. [24] reported coupling between high-Q graphene resonators (three to four layers) and high-Q microwave cavities and demonstrated detection of mechanical motion and manipulation of the coupling strength. Song et al. [28] designed similar optomechanical system and used back-action cooling to achieve 50 ± 6 mechanical quanta in a bilayer graphene resonator at 60 mK of bath temperature.

Singh et al. [23] then took one step further: by carefully preparing their microwave cavities that coupled with a 30-layer thick graphene resonator, the authors observed optomechanically induced reflection (OMIR) and optomechanically induced absorption (OMIA). The OMIA phenomena can be used for microwave photon storage. Given the high mechanical quality factor (≈220,000), the decoherence time is estimated up to 10 ms.

Cross-References

- C-MEMS
- Carbon Nanotube NEMS
- Chemical Vapor Deposition (CVD)
- CMOS-MEMS
- Graphene
- NEMS Mass Sensors
- Optomechanics

References

1. Lee, C., Wei, X., Kysar, J.W., Hone, J.: Measurement of the elastic properties and intrinsic strength of monolayer graphene. Science **321**, 385–388 (2008)
2. Lee, G.-H., et al.: High-strength chemical-vapor-deposited graphene and grain boundaries. Science **340**, 1013–1016 (2013)
3. Griffith, A.A.: The phenomena of rupture and flow in solids. Philos. Trans. R. Soc. Lond. A. **221**, 163–198 (1921)
4. Kim, K.S., et al.: Large-scale pattern growth of graphene films for stretchable transparent electrodes. Nature **457**, 706–710 (2009)

5. Li, X., et al.: Large-area synthesis of high-quality and uniform graphene films on copper foils. Science **324**, 1312–1314 (2009)
6. Lee, Y., et al.: Wafer-scale synthesis and transfer of graphene films. Nano Lett. **10**, 490–493 (2010)
7. Lee, S., et al.: Electrically integrated su-8 clamped graphene drum resonators for strain engineering. Appl. Phys. Lett. **102**, 153101 (2013)
8. Frank, I., Tanenbaum, D.M., Van der Zande, A., McEuen, P.L.: Mechanical properties of suspended graphene sheets. J. Vac. Sci. Technol. B **25**, 2558–2561 (2007)
9. Koenig, S.P., Boddeti, N.G., Dunn, M.L., Bunch, J.S.: Ultrastrong adhesion of graphene membranes. Nat. Nanotechnol. **6**, 543–546 (2011)
10. Bunch, J.S., et al.: Impermeable atomic membranes from graphene sheets. Nano Lett. **8**, 2458–2462 (2008)
11. Smith, A.D., et al.: Electromechanical piezoresistive sensing in suspended graphene membranes. Nano Lett. **13**, 3237–3242 (2013)
12. Milaninia, K.M., Baldo, M.A., Reina, A., Kong, J.: All graphene electromechanical switch fabricated by chemical vapor deposition. Appl. Phys. Lett. **95**, 183105 (2009)
13. Liu, X., et al.: Large arrays and properties of 3-terminal graphene nanoelectromechanical switches. Adv. Mater. **26**, 1571–1576 (2014)
14. Chen, C., Hone, J.: Graphene nanoelectromechanical systems. Proc. IEEE **101**, 1766–1779 (2013)
15. Bunch, J.S., et al.: Electromechanical resonators from graphene sheets. Science **315**, 490–493 (2007)
16. Chen, C., et al.: Performance of monolayer graphene nanomechanical resonators with electrical readout. Nat. Nanotechnol. **4**, 861–867 (2009)
17. van der Zande, A.M., et al.: Large-scale arrays of single-layer graphene resonators. Nano Lett. **10**, 4869–4873 (2010)
18. Song, X., et al.: Stamp transferred suspended graphene mechanical resonators for radio frequency electrical readout. Nano Lett. **12**, 198–202 (2011)
19. Xu, Y., et al.: Radio frequency electrical transduction of graphene mechanical resonators. Appl. Phys. Lett. **97**, 243111 (2010)
20. Atalaya, J., Isacsson, A., Kinaret, J.M.: Continuum elastic modeling of graphene resonators. Nano Lett. **8**, 4196–4200 (2008)
21. Barton, R.A., et al.: High, size-dependent quality factor in an array of graphene mechanical resonators. Nano Lett. **11**, 1232–1236 (2011)
22. Eichler, A., et al.: Nonlinear damping in mechanical resonators made from carbon nanotubes and graphene. Nat. Nanotechnol. **6**, 339–342 (2011)
23. Singh, V., et al.: Optomechanical coupling between a multilayer graphene mechanical resonator and a superconducting microwave cavity. Nat. Nanotechnol. **9**, 820–824 (2014)
24. Weber, P., Güttinger, J., Tsioutsios, I., Chang, D., Bachtold, A.: Coupling graphene mechanical resonators to superconducting microwave cavities. Nano Lett. **14**, 2854–2860 (2014)
25. Chaste, J., et al.: A nanomechanical mass sensor with yoctogram resolution. Nat. Nanotechnol. **7**, 301–304 (2012)
26. Barton, R.A., et al.: Photothermal self-oscillation and laser cooling of graphene optomechanical systems. Nano Lett. **12**, 4681–4686 (2012)
27. Chen, C., et al.: Graphene mechanical oscillators with tunable frequency. Nat. Nanotechnol. **8**, 923–927 (2013)
28. Song, X., Oksanen, M., Li, J., Hakonen, P.J., Sillanpää, M.A.: Graphene optomechanics realized at microwave frequencies. Phys. Rev. Lett. **113**, 021404 (2014)

Graphene Synthesis

▶ Synthesis of Graphene

Graphite

▶ Graphene

Greases

▶ Boundary Lubrication

Green Tribology and Nanoscience

Michael Nosonovsky[1] and Bharat Bhushan[2]
[1]Department of Mechanical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, USA
[2]Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Ecotribology

Definition

Tribology (from the Greek word τρίβω "tribo" meaning "to rub") is the branch of science and technology concerned with interacting surfaces in relative motion and with associated matters (as friction, wear, lubrication, and the design of bearings). Green tribology is the science and technology of the tribological aspects of ecological balance and of environmental and biological impacts.

Research Areas

Green tribology deals with the ecological aspects of technology involving friction, lubrication, wear, solid surfaces, coatings, as well as applications of this technology. This makes green tribology closely related to the novel areas of tribology: nanotribology, biotribology, and biomimetics. The term "green tribology" was suggested in about 2009 by several engineers and economists. In 2010, the first scientific volume was published by Philosophical Transactions of the Royal Society [1]. A number of symposia and conference sections devoted to ecological and green tribology took place starting from approximately the same time, as well as books prepared for publication [2].

Green or environment-friendly tribology emphasizes the aspects of interacting surfaces in relative motion, which are of importance for energy or environmental sustainability or which have impact upon today's environment. This includes tribological technology that mimics living nature (biomimetic surfaces) and thus is expected to be environment-friendly, the control of friction and wear that is of importance for energy conservation and conversion, environmental aspects of lubrication (including natural oil lubrication and self-lubrication), surface modification techniques, and tribological aspects of green applications, such as wind-power turbines, tidal turbines, or solar panels. A number of tribological problems could be put under the umbrella of "green tribology," and they are of mutual benefit to one another.

Green tribology can be viewed in the broader context of two other "green" areas: green engineering and green chemistry. The U.S. Environmental Protection Agency (EPA) defines green engineering as "the design, commercialization, and use of processes and products that are technically and economically feasible while minimizing (1) generation of pollution at the source (2) risk to human health and the environment." Green chemistry, also known as sustainable chemistry, is defined as the design of chemical products and processes that reduce or eliminate the use or generation of hazardous substances.

Biomimetic and Self-Lubricating Materials/Surfaces

Biomimetic materials are also usually environmentally friendly in a natural way since they are a natural part of the ecosystem. For this reason, the biomimetic approach in tribology is particularly promising. In the area of biomimetic surfaces, a number of ideas have been suggested, including the Lotus-effect, gecko-effect, micro- and nanostructured surfaces, texturing, self-lubrication, hydrophobicity, oleophobicity, etc.

Environmental engineers have only just started paying attention to biomimetic surfaces. One research indicates that while the use phase benefits are apparent, production burdens can outweigh them when compared with other cleaning methods, so a more thoughtful and deliberate use of bio-inspiration in sustainable engineering is needed. Clearly, more studies are likely to emerge in the near future.

Biodegradable Lubrication

In the area of environment-friendly and biodegradable lubrication, several ideas have been suggested: This includes the use of natural (e.g., vegetable-oil-based or animal-fat-based) biodegradable lubricants for engines, hydraulic applications, and metal cutting applications. In particular, corn, soybean, coconut oils have been used so far (the latter is of particular interest in

tropical countries such as India). These lubricants are potentially biodegradable, although in some cases chemical modification or additives for best performance are required. Vegetable oils can have excellent lubricity, far superior than that of mineral oil. In addition, they have a very high viscosity index and high flash/fire points. However, natural oils often lack sufficient oxidative stability, which means that the oil will oxidize rather quickly during use, becoming thick and polymerizing to a plastic-like consistency. Chemical modification of vegetable oils and/or the use of antioxidants can address this problem.

Powder lubricants and, in particular, boric acid lubricants, in general, tend to be much more ecologically friendly than the traditional liquid lubricants. Boric acid and MoS_2 powder can also be used as an additive to the natural oil. Friction and wear experiments show that the nanoscale (20 nm) particle boric acid additive lubricants significantly outperformed all of the other lubricants with respect to frictional and wear performance. In fact, the nanoscale boric acid powder-based lubricants exhibited a wear rate more than an order of magnitude lower than the MoS_2 and larger-sized boric acid additive based lubricants.

Self-replenishing lubrication that uses oil-free environmentally benign powders for lubrication of critical components such as bearings used in fuel cell compressors and expanders.

Tribology of Renewable Energy Sources

The tribology of renewable sources of energy is a relatively new field of tribology. Today, there are meetings and sections devoted to the tribology of wind turbines at almost every tribology conference, and they cover certain issues specific for these applications. Unlike in the case of the biomimetic approach and environment-friendly lubrication, it is not the manufacturing or operation, but the very application of the tribological system which involves "green" issues, namely, environmentally friendly energy production. The following issues can be mentioned [1].

Wind power turbines have a number of specific problems related to their tribology and constitute a well-established area of tribological research. These issues include water contamination, electric arcing on generator bearings, issues related to the wear of the mainshaft and gearbox bearings and gears, the erosion of blades (solid particles, cavitation, rain, hail stones), etc.

Tidal power turbines are another important way of producing renewable energy, which involves certain tribiological problems. Tidal power turbines are especially popular in Europe (particularly, in the U.K.), which remains the leader in this area, although several potential sites in North America have been suggested. There are several specific tribological issues related to the tidal power turbines, such as their lubrication (seawater, oils, and greases), erosion, corrosion, and biofouling, as well as the interaction between these modes of damage.

Besides tidal, the ocean water flow energy and river flow energy (without dams) can be used with the application of special turbines, such as the Gorlov helical turbine, which provides the same direction of rotation independent of the direction of the current flow. These applications also involve specific tribological issues.

Geothermal energy plants are used in the USA (in particular, at the Pacific coast and Alaska); however, their use is limited to the geographical areas at the edges of tectonic plates. In 2007, they produced 2.7 GW of energy in the USA, with Philippines (2.0 GW) and Indonesia (1.0 GW) at the second and third place. There are several issues related to the tribology of geothermal energy sources which are discussed in the literature [1].

Principles

Nosonovsky and Bhushan [1] formulated the principles of green tribology, which belong to the three areas, suggested in the preceding section. Some principles are related to the design and manufacturing of tribological applications (3–10), while others belong to their operation (1–2 and 11–12). Tradition is followed and limited the number of principles is limited to 12.

1. *Minimization of heat and energy dissipation*. Friction is the primary source of energy

dissipation. According to some estimates, about one third of the energy consumption in the USA is spent to overcome friction. Most energy dissipated by friction is converted into heat and leads to the heat pollution of atmosphere and the environment. The control of friction and friction minimization, which leads to both the energy conservation and prevention of the damage to the environment due to the heat pollution, is a primary task of ribology. It is recognized that for certain tribological applications (e.g., car brakes and clutches) high friction is required; however, ways of effective use of energy for these applications should be sought as well.

2. *Minimization of wear* is the second most important task of tribology which has relevance to green tribology. In most industrial applications, wear is undesirable. It limits the lifetime of components and therefore creates the problem of their recycling. Wear can lead also to catastrophic failure. In addition, wear creates debris and particles which contaminate the environment and can be hazardous for humans in certain situations. For example, wear debris generated after human joint replacement surgery is the primary source of long-term complications in patients.
3. *Reduction or complete elimination of lubrication and self-lubrication*. Lubrication is a focus of tribology since it leads to the reduction of friction and wear. However, lubrication can also lead to environmental hazards. It is desirable to reduce lubrication or achieve the self-lubricating regime when no external supply of lubrication is required. Tribological systems in living nature often operate in the self-lubricating regime. For example, joints form essentially a closed self-sustainable system.
4. *Natural lubrication* (e.g., vegetable-oil based) should be used in cases when possible since it is usually environmentally friendly.
5. *Biodegradable lubrication* should also be used when possible to avoid environmental contamination.
6. *Sustainable chemistry and green engineering principles* should be used for the manufacturing of new components for tribological applications, coatings, and lubricants.
7. *Biomimetic approach* should be used whenever possible. This includes biomimetic surfaces, materials, and other biomimetic and bio-inspired approaches since they tend to be more ecologically friendly.
8. *Surface texturing* should be applied to control surface properties. Conventional engineered surfaces have random roughness, and the randomness is the factor which makes it extremely difficult to overcome friction and wear. On the other hand, many biological functional surfaces have complex structures with hierarchical roughness, which defines their properties. Surface texturing provides a way to control many surface properties relevant to making tribosystems more ecologically friendly.
9. *Environmental implications of coatings* and other methods of surface modification (texturing, depositions, etc.) should be investigated and taken into consideration.
10. *Design for degradation* of surfaces, coatings, and tribological components. Similar to green chemistry applications, the ultimate degradation/utilization should be taken into consideration during design.
11. *Real-time monitoring*, analysis, and control of tribological systems during their operation should be implemented to prevent the formation of hazardous substances.
12. *Sustainable energy applications* should become the priority of the tribological design as well as engineering design in general.

Cross-References

- ▶ Biomimetics
- ▶ Nanotribology

References

1. Nosonovsky, M., Bhushan, B.: Green tribology: principles, research areas and challenges. Philos. Trans. R. Soc. A **368**, 4677–4694 (2010)
2. Nosonovsky, M., Bhushan, B. (eds.): Green Tribology Biomimetics, Energy Conservation, and Sustainability. Springer, Heidelberg (2012)

Green's Function Integral Equation Method

Thomas Søndergaard
Department of Physics and Nanotechnology, Aalborg University, Aalborg Øst, Denmark

Synonyms

Lippmann-Schwinger integral equation method

Definition

The Green's function integral equation method (GFIEM) is a method for solving linear differential equations by expressing the solution in terms of an integral equation, where the integral involves an overlap integral between the solution itself and a Green's function. In particular, within nanotechnology the method is frequently applied to calculate scattering of light.

Overview

The Green's function integral technique can be used to solve a linear inhomogeneous differential equation in real space such as

$$\theta_0 \varphi(\mathbf{r}) = \pm f(\mathbf{r}), \tag{1}$$

where θ_0 is an operator, $\mathbf{r}$ is the position, $f(\mathbf{r})$ is a given source term, and $\varphi(\mathbf{r})$ is the function that should be calculated such that Eq. 1 is satisfied. A solution can in principle be obtained straightforwardly from a Green's function $g(\mathbf{r}, \mathbf{r}')$ of the operator θ_0. The Green's function satisfies the equation

$$\theta_0 g(\mathbf{r}, \mathbf{r}') = \pm \delta(\mathbf{r} - \mathbf{r}'), \tag{2}$$

where $\delta(\mathbf{r} - \mathbf{r}')$ is the Dirac delta function. A solution to the inhomogeneous equation (1) can be given straightforwardly by an overlap integral between the Green's function and the source term, i.e.,

$$\varphi(\mathbf{r}) = \int g(\mathbf{r}, \mathbf{r}') f(\mathbf{r}') d^3 r. \tag{3}$$

It should be recognized that there can be many solutions to Eq. 2 and thereby many possible Green's functions. Usually, it is also required that the solution $\varphi(\mathbf{r})$ must satisfy certain boundary conditions, and in many cases it is possible to choose the specific Green's function that will automatically make the expression (3) satisfy these boundary conditions. An example of the above procedure, which is usually not discussed in the context of Green's functions, is the electrostatic potential $\varphi(\mathbf{r})$ due to a distribution of charges in free space, in which case the potential must satisfy the equation [1]

$$\nabla^2 \varphi(\mathbf{r}) = -\rho(\mathbf{r})/\epsilon_0, \tag{4}$$

where $\rho(\mathbf{r})$ is the charge distribution, and ϵ_0 is the vacuum permittivity. The appropriate Green's function for finding the solution being zero at infinite distance from the source region is given by

$$g(\mathbf{r}, \mathbf{r}') = 1/4\pi \, |\mathbf{r} - \mathbf{r}'|, \tag{5}$$

which is a solution to

$$\nabla^2 g(\mathbf{r}, \mathbf{r}') = -\delta(\mathbf{r} - \mathbf{r}'), \tag{6}$$

and it satisfies the boundary condition that it becomes zero at infinity. Using Eqs. 3, 4, and 5 it is now straightforward to write up the well-known solution for the potential [1]

$$\varphi(\mathbf{r}) = \frac{1}{4\pi\epsilon_0} \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d^3 r'. \tag{7}$$

This may all seem very simple. However, this is only true if the Green's function is known and is in a form that is easy to use. Analytic expressions for Green's functions are usually only known for a few simple cases. Furthermore, in many cases the Green's function has a singularity that requires

special attention in a numerical solution scheme. Consider an equation of the form

$$(\theta_0 + \Delta\theta)(\varphi_0(\mathbf{r}) + \Delta\varphi(\mathbf{r})) = 0, \qquad (8)$$

where the operator $\theta_0 + \Delta\theta$ is the sum of an unmodified operator with a known Green's function having appropriate boundary conditions and a modification $\Delta\theta$. The solution $\varphi(\mathbf{r}) = (\varphi_0(\mathbf{r}) + \Delta\varphi(\mathbf{r}))$ should be in the form of the sum of a given solution $\varphi_0(\mathbf{r})$ for the case that $\Delta\theta = 0$ and a modification $\Delta\varphi(\mathbf{r})$ that must disappear when $\Delta\theta = 0$. This can be rewritten in the form

$$\theta_0(\varphi(\mathbf{r}) - \varphi_0(\mathbf{r})) = -\Delta\theta\varphi(\mathbf{r}), \qquad (9)$$

since $\theta_0\varphi_0(\mathbf{r}) = 0$. Equation 9 now has the form of an "inhomogeneous" equation, and a solution can be obtained using Eq. 3, i.e.,

$$\varphi(\mathbf{r}) = \varphi_0(\mathbf{r}) + \int g(\mathbf{r}, \mathbf{r}')\Delta\theta(\mathbf{r}')\varphi(\mathbf{r}')d^3r', \qquad (10)$$

which is an integral equation since $\varphi(\mathbf{r})$ appears on both sides. One possible solution scheme is to divide the domain in which $\Delta\theta \neq 0$ in small elements and to assume that φ is constant within each element, in which case the discretized equation (10) becomes a linear system of equations in the discrete values of φ that can be put on matrix form and solved numerically on a computer.

In the following, integral equations of the form (10) shall be constructed for a range of simple cases of interest for scattering problems in optics and electromagnetics. The first case that will be considered is the scalar wave equation in one and two dimensions. Then, the numerically more challenging case of the vector wave equation in three dimensions shall be considered. The examples are for the case of monochromatic light with angular frequency ω. The real solution is obtained from the complex solutions obtained in the following by multiplying with $\exp(i\omega t)$ and then taking the real part of the resulting expression. Throughout this entry the factor $\exp(i\omega t)$ will be suppressed.

Scalar Green's Function Integral Equation Method for One-Dimensional Scattering Problems

While the type of problems considered in this section can readily be solved by other methods such as, e.g., a transfer-matrix method [2], the case of a scalar one-dimensional scattering problem nevertheless provides a very instructive and transparent exemplification of the Green's function method and serves as a good introduction to the more complex cases. As an example, consider light propagating from left to right being normally incident on a dielectric barrier (Fig. 1), a case that was previously considered in Ref. [3].

The incident field $E_0(x)$ is a solution to the wave equation for a structure where the dielectric barrier is not present, i.e.,

$$\left(\frac{\partial^2}{\partial x^2} + k_0^2\epsilon_{\text{ref}}\right)E_0(x) = 0. \qquad (11)$$

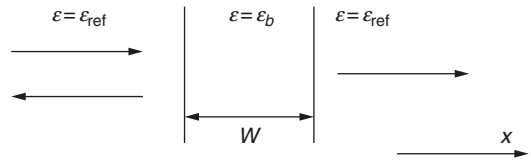
A natural choice of incident field could be a plane wave propagating along the positive x-axis, i.e.,

$$E_0(x) = E_0 e^{-ik_0 n_{\text{ref}} x}, \qquad (12)$$

where $n_{\text{ref}} = \sqrt{\epsilon_{\text{ref}}}$. The equation to be solved is the wave equation for the full structure including the barrier, namely,

$$\left(\frac{\partial^2}{\partial x^2} + k_0^2\epsilon(x)\right)E(x) = 0, \qquad (13)$$

with the boundary condition that the solution must be the sum of the given incident $E_0(x)$ and a



Green's Function Integral Equation Method, Fig. 1 Illustration of scattering of light by a dielectric barrier of dielectric constant ε_b and width w. The dielectric constant of the reference medium (without the barrier) is ε_{ref}

scattered field component that propagates away from the barrier. Equations 11 and 13 can be rewritten in the form of the following "inhomogeneous" equation

$$\left(\frac{\partial^2}{\partial x^2} + k_0^2 \epsilon_{\text{ref}}\right)(E(x) - E_0(x)) = -k_0^2(\epsilon(x) - \epsilon_{\text{ref}})E(x). \tag{14}$$

The relevant Green's function for the reference medium must satisfy the equation

$$\left(\frac{\partial^2}{\partial x^2} + k_0^2 \epsilon_{\text{ref}}\right) g(x, x') = -\delta(x - x') \tag{15}$$

along with the radiating boundary condition, namely, that the solution must propagate away from the position x'. Equation 15 and this boundary condition uniquely determines the Green's function which is given by

$$g(x, x') = g(x - x') = \frac{1}{2ik_0 n_{\text{ref}}} e^{-ik_0 n_{\text{ref}}|x - x'|}. \tag{16}$$

A solution of Eq. 14 can then be obtained by solving the integral equation

$$E(x) = E_0(x) + \int g(x, x') k_0^2 (\epsilon(x') - \epsilon_{\text{ref}}) E(x') dx'. \tag{17}$$

Due to the properties of the chosen Green's function, Eq. 17 will give the solution of interest since it satisfies the appropriate boundary condition, namely, that the total field must be the sum of the given incident field and a scattered field component propagating away from the scattering geometry.

An advantage in a numerical solution scheme of Eq. 17 compared with other methods, such as finite-difference methods, is that initially it is sufficient to consider only the x, x' where $\epsilon - \epsilon_{\text{ref}} \neq 0$. A simple approach is to, e.g., discretize the barrier in Fig. 1 in N elements sampled at positions $x_1, x_2, \ldots x_N$ with distance Δx between sampling points and to assume that within each element the field is constant and given by $E_1, E_2, \ldots E_N$, respectively. In that case, Eq. 17 can be approximated with the discrete linear system of equations

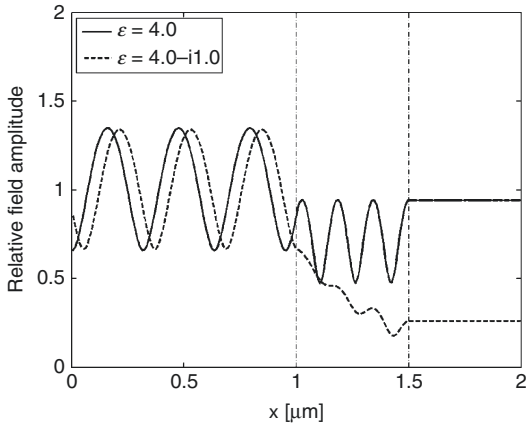
$$E_i = E_{0i} + \sum_j g_{ij} k_0^2 (\epsilon_j - \epsilon_{\text{ref}}) E_j \Delta x, \tag{18}$$

where $g_{ij} = g(x_i, x_j)$. This equation can be put in matrix form and readily solved on a computer. The calculated field inside the barrier can then be used in Eq. 17 to obtain the field at all other desired positions outside the barrier by direct integration over the field inside the barrier. Since a homogeneous reference medium is considered, the transmission through the barrier can, e.g., be calculated by considering $E(x)$ for a position x being larger than the barrier positions from the ratio $T = |E(x)/E_0(x)|^2$. Similarly, the reflection can be calculated by considering a position x being smaller than the barrier positions from the ratio $R = |(E(x) - E_0(x))/E_0(x)|^2$.

Finally, a few examples of calculations obtained with this method are presented that may also serve as an exercise for the reader, namely, the electric field in the case of scattering (reflection) from a dielectric barrier (Fig. 2) and reflection and transmission spectra for light incident on a layered structure or grating designed to strongly reflect red light (Fig. 3). These examples are very similar to instructive examples in Ref. [3].

Scalar Green's Function Integral Equation Method for Two-Dimensional Scattering Problems

In this section, the method from the previous section is extended to a two-dimensional scattering problem, namely, scattering of light by a very long (nano)wire (Fig. 4). Consider a situation with light propagating in the xy-plane and a wire oriented along the z-axis, and that the electric field is invariant along the z-axis and polarized along the z-axis, i.e., $\mathrm{E}(\mathbf{r}) = \hat{z}E(x, y)$. In that case the wave



Green's Function Integral Equation Method, Fig. 2 Electric field magnitude normalized relative to the incident field, i.e., $|E(x)/E_0(x)|$ for an incident wave with wavelength 633 nm (red light) on a dielectric barrier of width 500 nm and dielectric constant with and without absorption losses, namely, $\varepsilon = 4$ or a case with losses $\varepsilon = 4 - i$

equation for the electric field reduces to the following scalar wave equation:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_0^2 \epsilon(x,y)\right) E(x,y) = 0, \qquad (19)$$

where $\epsilon(x, y)$ is the dielectric constant of the total structure including the wire. Outside the wire the dielectric constant is given by ϵ_{ref}. Thus, also in this case is $(\epsilon(x,y) - \epsilon_{\text{ref}})$ the modification made to a reference structure, and it is only nonzero for positions inside the wire. The boundary condition when solving the differential equation (19) is also in this case that the total field $E(x, y)$ should be the sum of a given incident field $E_0(x, y)$ and a scattered field, where the scattered field is propagating away from the scattering object. The given incident field must be a solution to the wave equation in the case when there is no nanowire, i.e.,

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_0^2 \epsilon_{\text{ref}}\right) E_0(x,y) = 0. \qquad (20)$$

For simplicity, the incident field can, e.g., be a plane wave propagating in the direction given by $\hat{s} = \hat{x}\cos\theta + \hat{y}\sin\theta$, i.e.,

$$E_0(x,y) = E_0 e^{-ik_0 n_{\text{ref}}(x\cos\theta + y\sin\theta)}. \qquad (21)$$

From Eqs. 19 and 20, the following equation is obtained:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_0^2 \epsilon_{\text{ref}}\right)(E(x,y) - E_0(x,y)) = -k_0^2(\epsilon(x,y) - \epsilon_{\text{ref}})E(x,y). \qquad (22)$$

While the method can be used for wires with any geometry, examples will be given for a wire with (a) rectangular and (b) circular cross section (see Fig. 4).

The appropriate Green's function is a solution to the equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + k_0^2 \epsilon_{\text{ref}}\right) g(x,y;x',y') = -\delta(x-x')\delta(y-y'), \qquad (23)$$

with the boundary condition that the solution must correspond to a field propagating away from the position $\mathbf{r}' = (x', y')$. With these requirements, the Green's function is uniquely determined to be the cylindrical wave given by

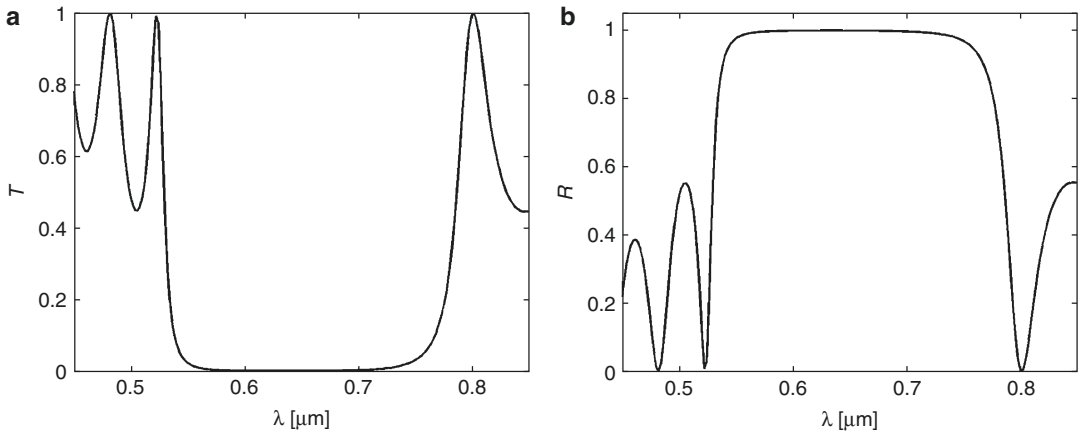
$$g(\mathbf{r},\mathbf{r}') = \frac{1}{4i} H_0^{(2)}(k_0 n_{\text{ref}}|\mathbf{r} - \mathbf{r}'|), \qquad (24)$$

where $H_0^{(2)}(z)$ is the Hankel function of order 0 and second kind. From Eqs. 22 and 23, the following integral equation is straightforwardly obtained:

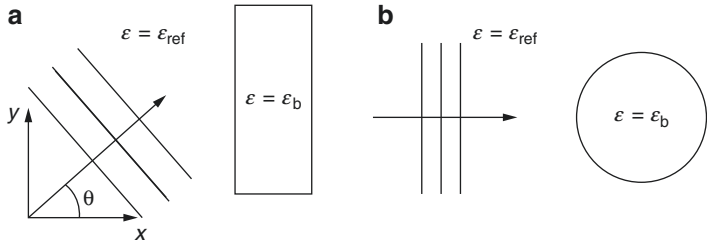
$$E(\mathbf{r}) = E_0(\mathbf{r}) + \int g(\mathbf{r},\mathbf{r}') k_0^2 (\epsilon(\mathbf{r}) - \epsilon_{\text{ref}}) E(\mathbf{r}') d^2 r', \qquad (25)$$

where the field has the desired form of the sum of the given incident field and a scattered part propagating away from the scatterer due to the properties of the chosen Green's function.

The simplest approach for solving Eq. 25 is also in this case to discretize the structure in,



Green's Function Integral Equation Method, Fig. 3 Reflection and transmission spectra for light of wavelength λ being incident on a layered structure with alternating seven layers of high refractive index $n_H = 2.32$ and thickness 633 nm/$4n_H$ and six layers of a smaller refractive index $n_L = 1.38$ and thickness 633 nm/$4n_L$. The surrounding material is air



Green's Function Integral Equation Method, Fig. 4 Schematic of a plane wave being incident on a wire with (**a**) rectangular and (**b**) circular cross section

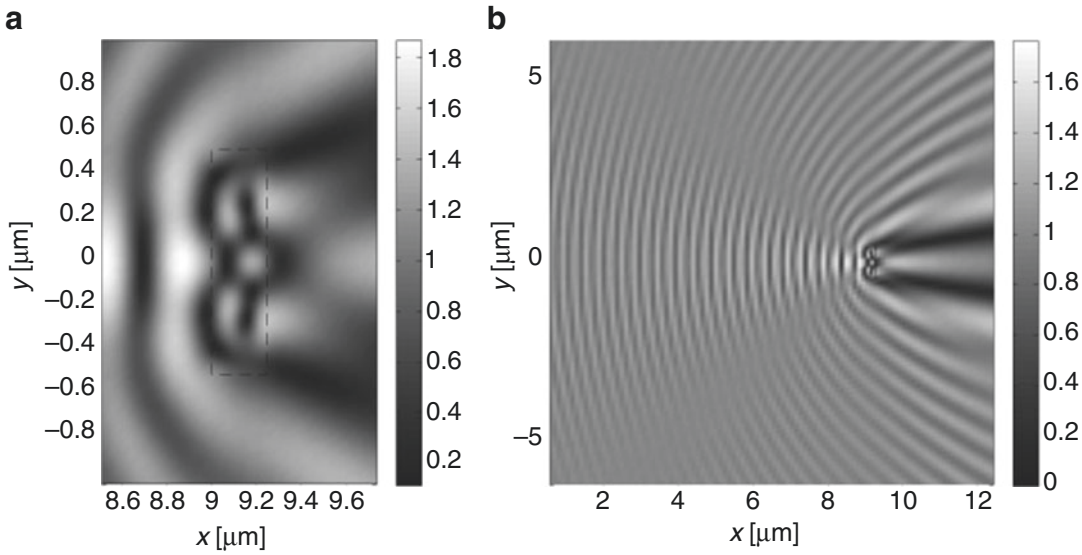
e.g., N square shaped area elements all having the same area ΔA and center at $\mathbf{r}_1$, $\mathbf{r}_2$, ... $\mathbf{r}_N$, respectively. As an approximation, it can be assumed that the field values are constant within each element and denoted E_1, E_2,..., E_N. Thereby, the following linear system of equations is obtained:

$$E_i = E_{0,i} + \sum_j g_{ij} k_0^2 (\epsilon_j - \epsilon_{\text{ref}}) E_j \Delta A, \qquad (26)$$

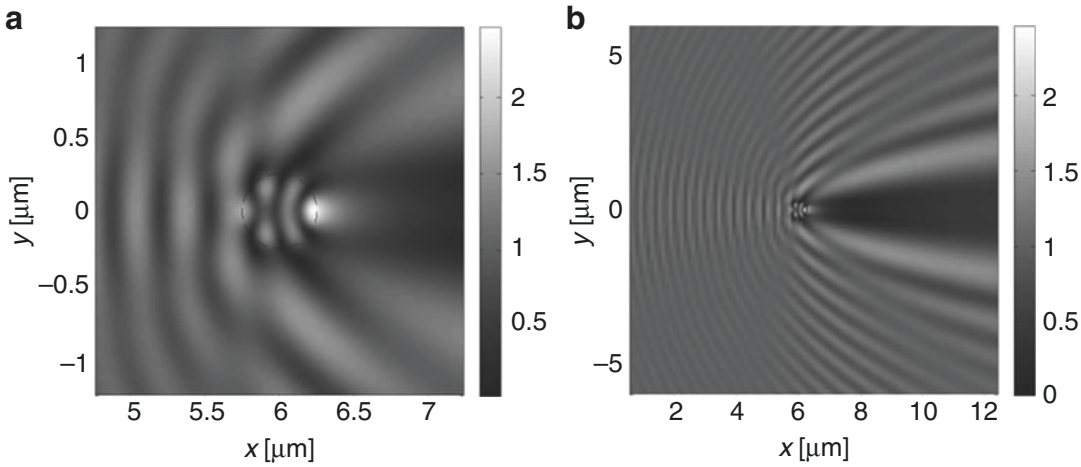
where it is, e.g., possible to use $g_{ij} = g(\mathbf{r}_i, \mathbf{r}_j)$ if $i \neq j$ and $\epsilon_j = \epsilon(\mathbf{r}_j)$. It may be slightly more accurate to use $g_{ij} = \frac{1}{\Delta A} \int_{\text{element } j} (\mathbf{r}_i, \mathbf{r}') d^2 r'$. However, this does not make a significant difference unless $i = j$. In that case, however, due to the singularity of the Green's function it is not possible to use $g_{ij} = g(\mathbf{r}_i, \mathbf{r}_j)$. Instead it can be acceptable to approximate the square element with a circular element having the same area and center, in which case the integral can be calculated analytically. The radius of the circle then is $a = \sqrt{\Delta A/\pi}$. This leads to the approximation

$$\begin{aligned} g_{ii} &\approx \frac{2\pi}{\Delta A} \int_{r=0}^{a} \frac{1}{4i} H_0^2(k_0 r) r dr \\ &= \frac{1}{2i(k_0 a)^2} \left[k_0 a H_1^{(2)}(k_0 a) - i \frac{2}{\pi} \right]. \end{aligned} \qquad (27)$$

The linear system of Eq. 26 can be formulated in matrix form and solved on a computer. As a small example that may also serve as an exercise for the reader, a few results are presented for scattering from the rectangular nanowire (Fig. 5) and a circular nanowire (Fig. 6).



Green's Function Integral Equation Method, Fig. 5 Magnitude of electric field normalized with the magnitude of the incident plane wave for a plane wave being scattered by a rectangular nanowire. The light wavelength is $\lambda = 633$ nm and the nanowire widths are 0.25 and 1.0 μm, respectively. The nanowire has dielectric constant $\varepsilon = 4$ and is located in free space with dielectric constant $\varepsilon_{\text{ref}} = 1$. The calculation was carried out with 10×40 area elements inside the scatterer



Green's Function Integral Equation Method, Fig. 6 The same as Fig. 5 except that light is incident on a circular nanowire with diameter 0.5 μm

Dyadic Green's Function Integral Equation Method for Three-Dimensional Problems

For a general three-dimensional scattering problem where light is scattered by one or more objects such as, e.g., a nanoparticle, the total electric field vector $\mathbf{E}$ must satisfy the equation

$$\left(-\nabla\nabla\cdot+\nabla^2+k_0^2\epsilon(\mathbf{r})\right)\mathbf{E}(\mathbf{r})=\mathbf{0}, \tag{28}$$

where $\epsilon(\mathbf{r})$ is the dielectric constant of the total structure consisting of a reference structure and one or more modifications representing the scattering objects. The incident field $\mathbf{E}_0$, which is the field in the case when no modification has been made to the reference structure, is a solution to the equation

$$\left(-\nabla\nabla\cdot+\nabla^2+k_0^2\epsilon_{\text{ref}}\right)\mathbf{E_0}(\mathbf{r})=\mathbf{0}. \tag{29}$$

A dyadic Green's tensor $\mathbf{G}$ for the reference medium is defined as a solution to the equation

$$\left(-\nabla\nabla\cdot+\nabla^2+k_0^2\epsilon_{\text{ref}}\right)\mathbf{G}(\mathbf{r},\mathbf{r}')=-\mathbf{I}\delta(\mathbf{r}-\mathbf{r}'), \tag{30}$$

where $\mathbf{I}$ is the unit dyadic tensor ($\mathbf{I}\cdot\mathbf{A}=\mathbf{A}$ for any vector function $\mathbf{A}$). In the case of a homogeneous reference medium (ϵ_{ref} being independent of the position) the dyadic Green's tensor can be expressed in terms of the scalar Green's function

$$g(\mathbf{r},\mathbf{r}')=\frac{e^{-ik_0n_{\text{ref}}|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{31}$$

satisfying

$$\left(\nabla^2+k_0^2\epsilon_{\text{ref}}\right)g(\mathbf{r},\mathbf{r}')=-\delta(\mathbf{r}-\mathbf{r}'). \tag{32}$$

The dyadic Green's function is for this special case given by

$$\mathbf{G}(\mathbf{r},\mathbf{r}')=\left(\frac{1}{k_0^2\epsilon_{\text{ref}}}\nabla\nabla+\mathbf{I}\right)g(\mathbf{r},\mathbf{r}') \tag{33}$$

which satisfies Eq. 30 and the radiating boundary condition that it corresponds to propagation away from the position $\mathbf{r}'$.

The resulting integral equation for the scattering problem is then given by

$$\mathbf{E}(\mathbf{r})=\mathbf{E}_0(\mathbf{r})+\int\mathbf{G}(\mathbf{r},\mathbf{r}')k_0^2(\epsilon(\mathbf{r}')-\epsilon_{\text{ref}})\cdot\mathbf{E}(\mathbf{r}')d^3r'. \tag{34}$$

An approach for solving Eq. 34 which is similar to the discrete-dipole approximation method [4] is again to discretize Eq. 34 into a number of volume elements where the electric field vector and the dielectric constant are assumed constant. This leads to the discretized version of Eq. 34 given by

$$\mathbf{E}_i=\mathbf{E}_{0,i}+\sum_j\mathbf{G}_{i,j}\cdot\left(\epsilon_j-\epsilon_{\text{ref},j}\right)\mathbf{E}_j, \tag{35}$$

where

$$\mathbf{G}_{i,j}=\int V_j^G(\mathbf{r},\mathbf{r}')k_0^2d^3r'. \tag{36}$$

The integral in Eq. 36 is rather difficult in the case when $i=j$ due to the singularity of the dyadic Green's tensor. An useful approach for calculating the integral (36) in that case is to exploit the expressions (32) and (33) and transform the volume integral into a surface integral with the surfaces placed at a distance from the singularity, i.e.,

$$\mathbf{G}_{i,j}=-\frac{\mathbf{I}}{\epsilon_{\text{ref}}}+\frac{1}{\epsilon_{\text{ref}}}\oint_{\text{surface of }V_i}(\hat{n}'\nabla'-\mathbf{I}\hat{n}'\cdot\nabla')g(\mathbf{r},\mathbf{r}')d^2r', \tag{37}$$

where $\hat{n}'$ is the surface outward unit normal vector. In the case when $i\neq j$, it might be acceptable to use

$$\mathbf{G}_{i,j}\approx\mathbf{G}\left(\mathbf{r}_i,\mathbf{r}_j\right)k_0^2V_j. \tag{38}$$

In the event that the difference in the dielectric constant in Eq. 34 is large, such as is, e.g., the case with metal particles in a homogeneous dielectric medium, the numerical solution of Eq. 34 becomes challenging. If the structure is divided into cubic volume elements, the representation of a curved surface becomes poor since it will be represented with a stair-cased approximation. In this case, convergence becomes slow at best but may not be achievable with any reasonable manageable number of discretization elements [7]. However, convergence can be achieved for this case as well if discretization elements are chosen that accurately represents the curved surfaces [7].

Further Reading

For a deeper study of the Green's function integral equation method for electromagnetic scattering problems, the reader is referred to the Refs. [5–7]. For a deeper study of the theory of Green's functions, the reader is referred to Refs. [8–10]. In the case of scattering from a two-dimensional periodic array of scatterers being, e.g., placed on a planar surface the appropriate Green's function is presented in, e.g., Ref. [11]. Another Green's function integral equation method where integrals are over the surface of scatterers instead of over the interior of the scatterers is available for two-dimensional scattering problems in, e.g., Refs. [6, 7] and more recently for three-dimensional scattering problems in Ref. [12]. The Green's function for a reference medium involving two semi-infinite homogeneous media separated by a planar interface is given in, e.g., [5, 7].

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Finite-Difference Frequency-Domain Technique
- ▶ Finite-Difference Time-Domain Technique

References

1. Reitz, J.R., Milford, F.J., Christy, R.W.: Foundations of Electromagnetic Theory, 4th edn. Addison-Wesley, New York (1992)
2. Klein, M.V., Furtak, T.E.: Optics, 2nd edn. Wiley, New York (1986)
3. Martin, O.J.F., Dereux, A., Girard, C.: Iterative scheme for computing exactly the total field propagating in dielectric structures of arbitrary shape. J. Opt. Soc. Am. A **11**, 1073–1080 (1994)
4. Draine, B.T.: The discrete-dipole approximation and its application to interstellar graphite grains. Astrophys. J. **333**, 848–872 (1988)
5. Novotny, L., Hecht, B.: Principles of Nano-optics. Cambridge University Press, Cambridge (2007)
6. Jin, J.: The Finite Element Method in Electromagnetics, 2nd edn. Wiley, New York (2002)
7. Søndergaard, T.: Modeling of plasmonic nanostructures: green's function integral equation methods. Phys. Status Solidi B **244**, 3448–3462 (2007)
8. Tai, C.-T.: Dyadic Green's Functions in Electromagnetic Theory. Intext Educational Publishers, London (1971)
9. Morse, P.M., Feshbach, H.: Methods of Theoretical Physics. McGraw-Hill, New York (1953)
10. Economou, E.N.: Green's Functions in Quantum Physics. Springer, Berlin (1979)
11. Kobidze, G., Shanker, B., Nyquist, D.P.: Efficient integral-equation-based method for accurate analysis of scattering from periodically arranged nanostructures. Phys. Rev. E **72**, 056702 (2005)
12. Kern, A.M., Martin, O.J.F.: Surface integral formulation for 3D simulations of plasmonic and high permittivity nanostructures. J. Opt. Soc. Am. **26**, 732–740 (2009)

Growth of 1-D Oxide Nanostructures

Adrian Carretero-Genevrier[1] and Narcis Mestres[2]
[1]Institute des Nanotechnologies de Lyon (INL-CNRS), Lyon, France
[2]Consejo Superior de Investigaciones Científicas (CSIC), Institut de Ciencia de Materials de Barcelona (ICMAB), Barcelona, Spain

Synonyms

Preparation of complex oxide unidimensional nanostructures; Synthesis of multicationic 1-D oxide nanostructures

Definitions

One-dimensional (1D) oxide materials are anisotropic nanocrystals with large aspect ratios (length/diameter). Typically, they display diameters of 1–200 nm and lengths up to several tens of micrometers. Given their specific shape, 1D oxides are found in the literature under different names: nanowires, nanorods, nanotubes, nanobelts, nanofibers, whiskers, etc.

To date, different approaches have been proposed to achieve 1D growth of oxides with

different degrees of control over the growth process. These approaches are based on (i) taking advantage of the inherent anisotropic crystallographic structure of certain solids to promote the 1D growth, (ii) using organic additives or surfactants to kinetically control the oriented attachment of some crystallographic facets of a seed to finally form 1D nanostructures, (iii) using hard or soft templates (with 1D porosity) to directly grow nanowires and/or nanotubes, or (iv) controlling the condition of supersaturation to modify the tendency of seed to growth.

Multinary oxides, also named functional complex oxides, are materials that contain two or more metals in a precise ratio. The preparation of these oxides is to date a key challenge in materials science development.

Introduction

1D nanomaterials have attracted significant attention due to their inherent anisotropic nature, high surface area, structure-related properties resulting from confinement effects, and the possibility to be incorporated into devices. In addition, mixed-metal oxides are key components for many technological applications. These functional materials often contain one or more transition-metal elements whose oxidation states are central to define their magnetic, electronic, or redox properties. Transition-metal oxides exhibit rich physical properties expanding from insulators to semiconductors, metals, and superconductors. Their bandgap span from the visible to the ultraviolet energy regimes and their magnetic properties can be diamagnetic, paramagnetic, ferromagnetic, and antiferromagnetic. Their dielectric properties are shown to be notorious in both the low-k and high-k regimes, and they can also be ferroelectric and piezoelectric. All these properties make 1D nanostructures of complex metal oxides excellent candidates for a wide range of applications including the fabrication of advanced and more efficient devices associated with solar cells, energy storage, fuel cells, communications, information storage, sensors, catalysis and medical imaging, diagnosis, and detection. As a result, to date, different methods of synthesis are being investigated to generate 1D nanostructures of functional oxides with a precise control over their chemical composition, size, shape, and crystallinity.

The preparation of nanostructures can be distinguished between two well-defined approaches: (i) *top-down approach* which follows conventional lithography and, more recently, new and sophisticated electron and ion beam lithography to develop such nanostructures and (ii) *bottom-up approach* where defined complex nanostructures are "built up" through self-assembly processes, i.e., they use smaller units to create the nanometer-sized features with the desired shapes and/or dimensions. Top-down techniques are usually based on patterning a given shape using lithographic methods (including photolithography, electron beam lithography, electrochemical dip-pen lithography, etc.) that are thereafter removed by wet chemical etching, focused ion beam (FIB) milling, or reactive ion etching (RIE). However, these methods tend to be rather expensive, technically challenging, tedious, time-consuming, and relatively low throughput. In contrast, bottom-up chemical synthesis methods offer an alternative strategy to produce nanostructures with large material diversity, easy setup and good control over stoichiometry and doping, and the possibility of coating large and uniform areas.

Materials with a highly anisotropic crystallographic structure grow along a well-defined crystallographic direction. As a result, the 1D morphology can be spontaneously achieved. In this context, chemical solution-based synthesis methods have been proposed as the main routes to achieve a 1D growth in oxide materials. The approaches based on chemical solutions include (i) the use of organic additives or surfactants that kinetically control the oriented attachment of some crystallographic facets of a seed to finally form 1D nanostructures, (ii) the use of oriented attachment of nanocrystals to form 1D nanostructures, (iii) the confinement of 1D growth by using a hard or soft template with 1D porosity, or (iv) the control of supersaturation to modify the tendency of the seeds growth.

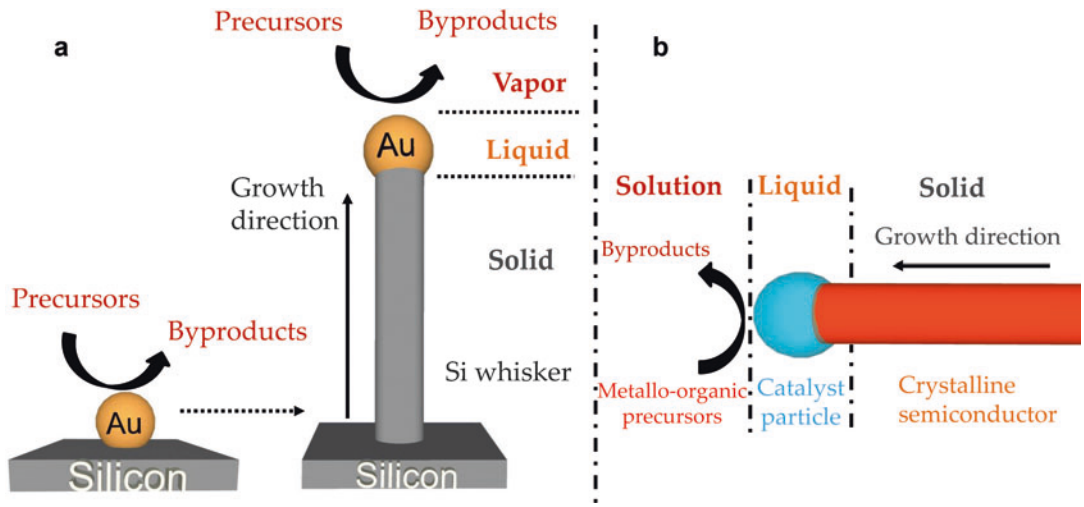
Catalytic Growth of Nanowires: Vapor–Liquid–Solid (VLS) and Solution–Liquid–Solid (SLS) Methods

The synthesis of 1D nanostructures using vapor–liquid–solid (VLS) and solution–liquid–solid (SLS) methods through confinement under a liquid drop has been the center of large efforts (see Fig. 1). These methods have been especially used to prepare nanowires of semiconductors such as Si, GaAs, and ZnO [1]. Briefly, a liquid droplet is used as a catalyst seed. As a result, each liquid droplet will work as a soft template to strictly limit the lateral growth of an individual wire. The precursor is then introduced as a vapor (VLS) or as a solution (SLS) that will be dissolved into the catalyst droplet. When the droplet is supersaturated with a given dissolved species, the nanowire starts to grow. The diameter of the nanowire is consequently determined by the size of the catalyst droplet.

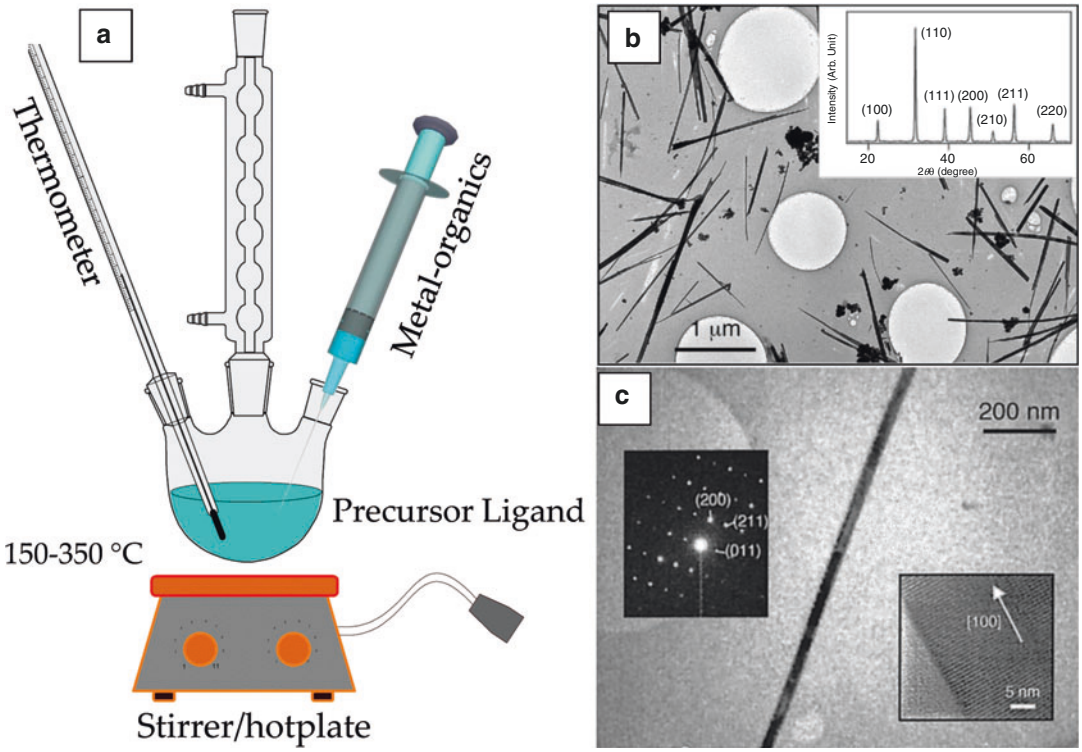
The VLS and SLS methods are better suited to produce single elements and binary semiconductors and oxides than complex materials such as ternary or quaternary oxides. The main reason for that is the need of more complex precursors and the corresponding difficulty in controlling the growth conditions and stoichiometry at the atomic level. Additionally, VLS or SLS growth of nanowires is restricted to systems that can form eutectic with catalysts at growth temperature. Unfortunately, for many of complex oxides, there is very limited information about the formation of eutectic liquids. In the case of multicomponent oxides, a precise control over the stoichiometric composition is possible only by chemical solution methods. Interestingly, when combined with template-mediated synthesis, it offers the possibility to produce aligned unidirectional and uniformly sized 1D oxide nanostructures over large areas which is highly attractive for device fabrication and the study of collective phenomena.

Chemical Solution Synthesis

1D complex functional oxide nanostructures can be produced, among other methods, by wet chemical methods. In this case, it is possible to obtain a large diversity of multielemental oxides with a number of cations that can exceed over four or five elements. Typically, soft chemistry provides high melting point compounds with a desired composition and structure which at low or mediate temperatures are otherwise difficult to form with a homogeneous composition. Therefore, these synthesis methods make it possible to



Growth of 1-D Oxide Nanostructures, Fig. 1 Schematic illustration of the VLS and SLS mechanisms for the growth nanowires. (**a**) VLS mechanism for the growth under chemical vapor deposition conditions. (**b**) SLS mechanism for the analogous growth under solution (Reprinted with permission from Ref. [1]. Copyright (2006) American Chemical Society)



Growth of 1-D Oxide Nanostructures, Fig. 2 (**a**) Schematic illustration of a typical solution-phase decomposition setup. Clip art was performed from ChemSketch freeware. (**b**) Transmission electron microscopy (TEM) image of $BaTiO_3$ nanorods. Inset: powder X-ray diffractogram of the reaction product composed of randomly oriented $BaTiO_3$ nanorod ensembles. (**c**) TEM image of a 30 nm diameter $BaTiO_3$ nanorod. *Left inset*: electron diffraction pattern obtained from the same nanorod. *Right inset*: A high-resolution TEM image of the nanorod that shows lattice fringes perpendicular to the [100] direction (Reprinted with permission from Ref. [3]. Copyright (2002) American Chemical Society)

control and to keep fixed the stoichiometry already present in the solution in a very reproducible way. Moreover, these methods avoid long-time calcinations at high temperatures, and, finally, they permit the production of large amounts of 1D nanostructure powders.

Wet chemistry includes (i) solution-phase decomposition, (ii) electrochemical techniques (including deposition and electrospinning), or (iii) hydrothermal, molten salt and template based synthesis methods. In all cases, the precursor compounds play a significant role. In this context, the fundamental purpose of wet chemistry is to design specific new precursors or to modify the existing ones in order to develop novel 1D oxide nanostructures.

Solution-Phase Decomposition Synthesis

This methodology uses the decomposition of bimetallic alkoxide precursors in the presence of coordinating ligands. As a result, it is possible to successfully obtain multicationic oxide nanowires. The main advantage of this procedure is the possibility to control the correct product stoichiometry and to avoid high-temperature treatment steps. Figure 2 shows a schematic illustration of the typical reaction apparatus involved in the synthesis of solution-phase oxide nanowires [2]. Notice that rapid injection of organometallic reagents in a hot coordination solvent produces temporally discrete homogeneous nucleation. The decomposition and nucleation occurs rapidly upon injection.

Following this approach, Urban et al. successfully synthesized [001]-$BaTiO_3$ ferroelectric nanowires [3]. The authors used an excess of H_2O_2 that was added at 100 °C to a heptadecane solution containing a 10:1 molar ratio of BaTi $[OCH(CH_3)_2]_6$ to oleic acid. The reaction mixture was then heated to 260 °C for 6 h, producing the formation of nanowires. In this case, the anisotropic growth takes place after the precursor decomposition and crystallization in a structured inverse micelle medium formed by precursors and oleic acid under the reaction conditions. These nanowires were found to be ferroelectric, exhibiting hysteresis loops with a coercive field of 7 kV/cm^{-1}.

Hydrothermal Synthesis

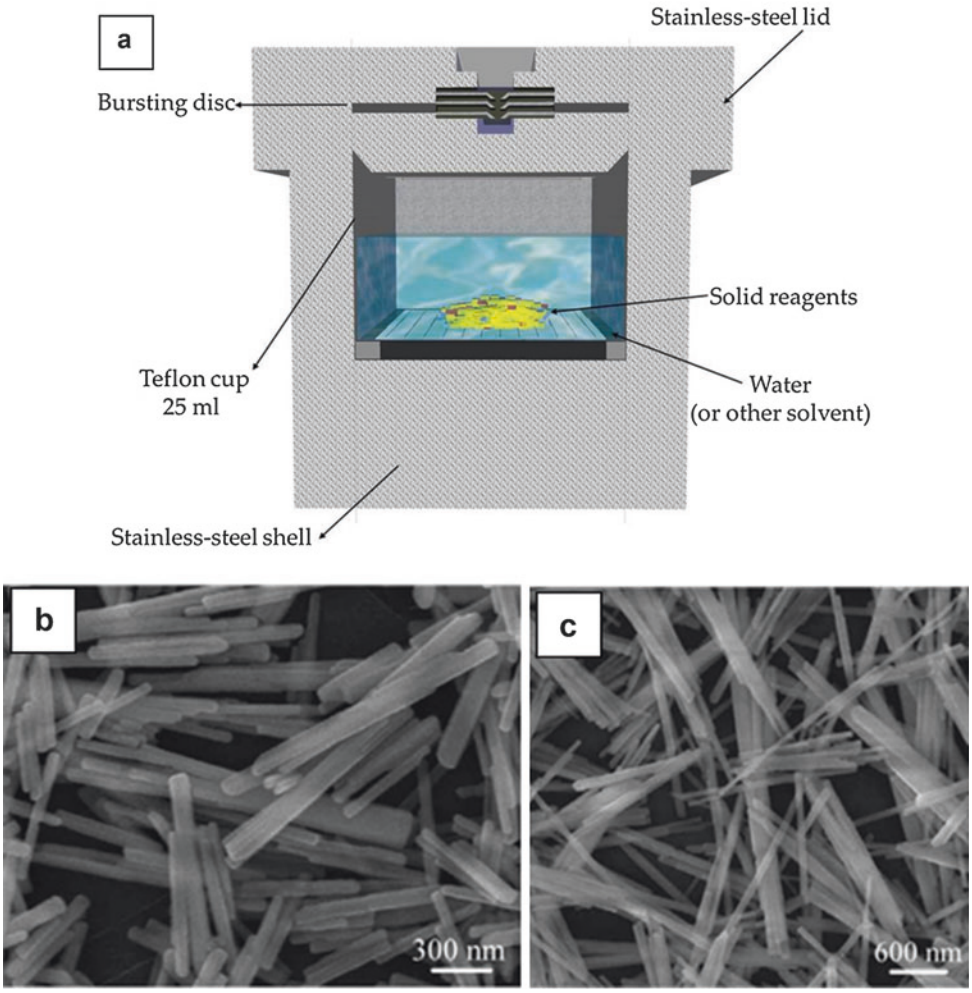
Hydrothermal synthesis is based on heating the chemical solution in a sealed vessel (autoclave). This procedure will increase the autogenous pressure inside the vessel beyond atmospheric pressure as the temperature exceeds the boiling point of the solvent [4]. As a result, the solubility and reactivity of the precursors used in the material synthesis are increased. When water is used as solvent, the process is called hydrothermal, whereas in the case of using a nonaqueous solvent, the process is termed solvothermal. Importantly, the hydrothermal synthesis exploits the enhanced solubility and reactivity of almost all inorganic substances in solvents at elevated temperatures and pressures, therefore, allowing the subsequent crystallization of the dissolved material/precursor ions from the solution under these conditions.

The hydrothermal synthesis is typically performed by placing the selected precursor, the mineralizer, and the additives all together with the solvent in the autoclave, as illustrated in Fig. 3. Normally, stainless steel autoclaves with a Teflon® lining are used which allows to avoid any eventual contamination from the steel. Precursors can be oxides, hydroxides, or salts of the elements required to build up the material. The autoclave is heated either by placing it in an oven or by using microwaves. The phase composition and morphology of the final product can be tuned by small changes in the chemical parameters such as the type of precursors, concentration, pH as well as process parameters like temperature and reaction time.

Only for materials with an anisotropic crystallographic structure, the 1D growth along one specific crystallographic direction can be directly obtained by hydrothermal synthesis. As an example, Fang et al. [5] were able to grow single-crystal lanthanide orthophosphate (e.g., $LaPO_4$) 1D nanowires by a simple hydrothermal synthesis route. In this case, the anisotropic crystal structure of these compounds facilitates the crystal growth along the *c*-direction. SEM images of two different synthesized $LnPO_4$ nanowires by hydrothermal treatment are displayed in Fig. 3. In most cases, and especially for materials with a cubic or pseudocubic crystal structure, the 1D directed growth approach has to be applied in order to achieve a 1D growth. This includes using precursors with a 1D nanostructure, taking advantage of appropriate organic additives (surfactants), or using the principle of oriented attachment of nonspherical nanocrystals. As an example, single-crystalline $KTa_{0.25}Nb_{0.75}O_3$ (KTN) nanorods with an orthorhombic perovskite structure were synthesized via a polymer-assisted hydrothermal method [6]. The preferred crystallographic orientation of the nanorods was found to be controllable by using different types of polymers in the process (see Fig. 4). As a result, the [001]-oriented KTN nanorods were obtained upon the use of polyvinyl alcohol (PVA), and the [110]-oriented ones were formed with the use of a combination of polyacrylic acid (PAA) and PVA. Since surfactants are capable to stabilize certain surfaces by selective adsorption, the difference in the growth rate between different crystallographic directions can be accentuated, therefore, giving rise to nanorods oriented along different growth directions.

Molten-Salt Synthesis

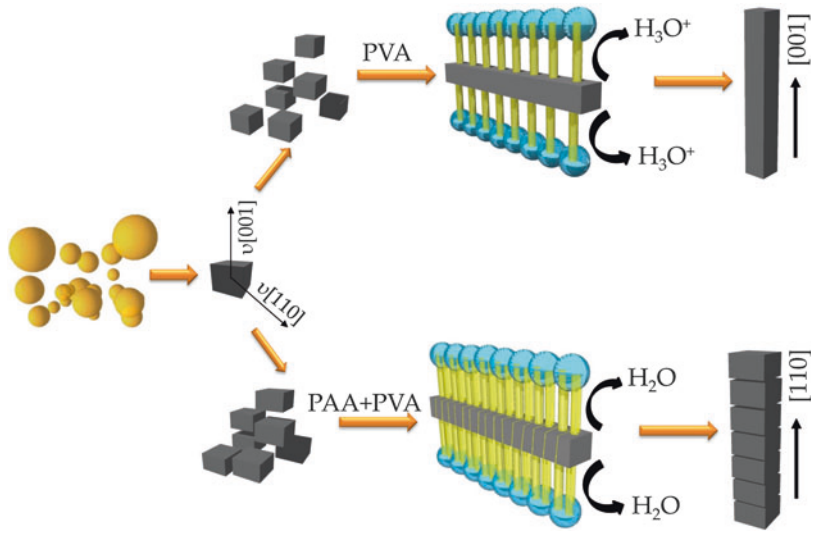
The molten-salt synthesis (MSS) method is straightforward and very versatile. In addition, it offers cost-effective approaches suitable for obtaining crystalline, chemically pure, and single-phase nanostructures at low temperatures. The main processing steps associated to the MSS



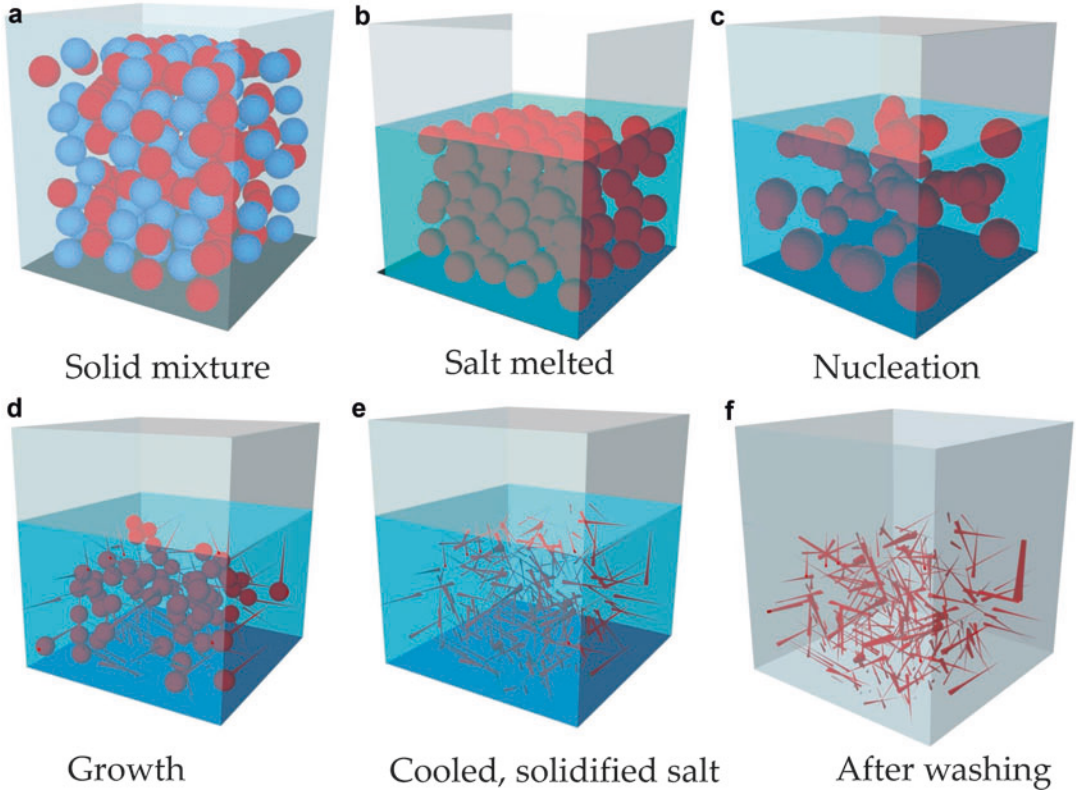
Growth of 1-D Oxide Nanostructures, Fig. 3 (**a**) Schematic representation of a typical stainless steel autoclave used for the hydrothermal synthesis (Adapted from Ref. [4]). SEM images of the as-synthesized (**b**) $GdPO_4$ and (**c**) $TbPO_4$ nanowires by hydrothermal synthesis (Reprinted with permission from Ref. [5]. Copyright (2003) American Chemical Society)

method are illustrated in Fig. 5. Essentially, oxides and/or other appropriate precursors according to the desired compound are mixed with either a given salt (e.g., NaCl) or an eutectic mixture of salts (e.g., $NaNO_3/KNO_3$, NaCl/KCl). Then, they are fired at a temperature above the melting point of the salt medium to form a molten flux. At this temperature and under these specific conditions, the precursor molecules will disperse, dissociate, diffuse, and then rapidly rearrange throughout the salt. In this way, the particles of a given metal-oxide phase are formed through an initial nucleation step followed by a growth process that strongly depends on the identity of the salt species, the quantity of salt used in the reaction medium, the temperature at which the reaction is performed, and the reaction duration. After cooling down, the salt itself can be simply eliminated by washing out with an aqueous solution. Upon drying, it is unlikely to obtain agglomerated nanowires after applying the appropriate processing.

Molten-salt synthesis has been used to prepare pristine $BaTiO_3$ nanowires with diameters ranging from 50 to 80 nm and with aspect ratios ranging from 1 to larger than 25 [7, 8].



Growth of 1-D Oxide Nanostructures, Fig. 4 Schematic illustration showing the possible formation of different growth directions of $KTa_{0.25}Nb_{0.75}O_3$ nanorods in the hydrothermal system using PVA and polyacrylic acid (PAA) + PVA as additives, respectively (Adapted from Ref. [6])



Growth of 1-D Oxide Nanostructures, Fig. 5 Schematics displaying the different steps involved in the molten-salt synthesis of metal-oxide nanowire materials (Adapted from Ref. [8])

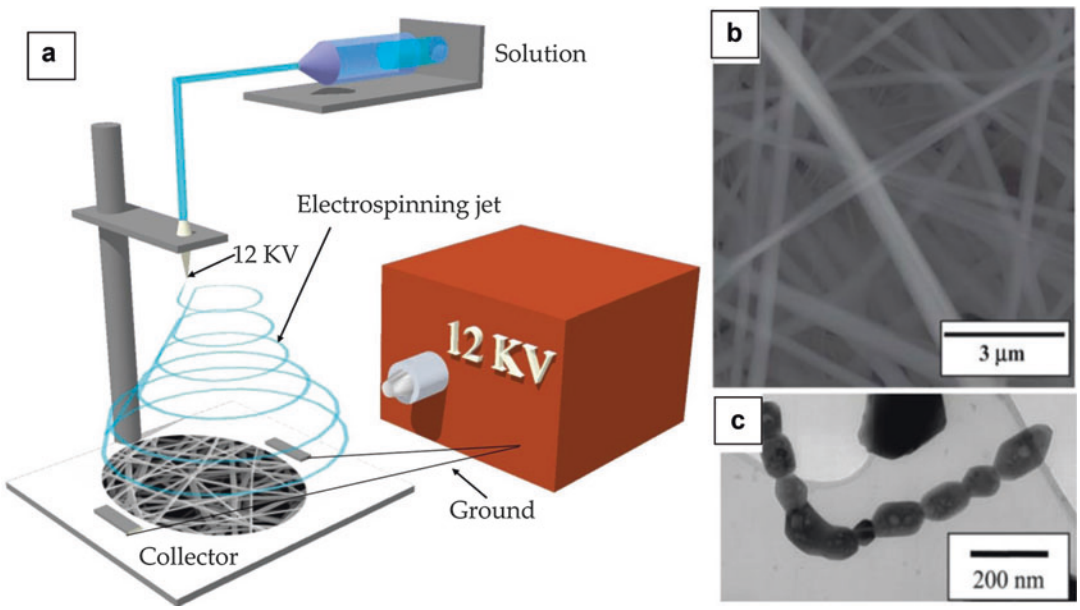
Electrospinning Synthesis

Electrospinning is another feasible solution-based process for the synthesis of 1D nanomaterials, in particular, to obtain nanofibers with very high surface area-to-volume ratios. These nanofibers can be synthesized under different morphologies. Specifically, solid, porous, hollow, core–shell, and compoundA/compoundB type have been successfully obtained using this technique. Typically, the electrospinning setup is reader simple and requires only a syringe with a metal needle, a conductive (grounded) collector, and a high-voltage power supply [9], as displayed in Fig. 6. Briefly, a solution or melt of the desired material is placed in the syringe. The application of a sufficiently strong electric field to the needle (electrode) results in a cone-like deformation of the solution droplets, from which a jet protrudes toward the collector (counter electrode). During the ejection from the needle toward the collection plate, solid nanofibers of the material can be formed on the collection plate by either solidification of the melt or by evaporation of the reaction solvent. As-prepared fibers often require an additional thermal annealing step in order to achieve sufficiently high crystalline materials.

As an example, we show, in Fig. 6, $BaZrO_3$ nanofibers that were electrospun from a 0.25 M metal cation solution containing barium acetate and zirconium (IV) 2,4-pentanedionate dissolved in glacial acetic acid [10]. Polyvinylpyrrolidone was added to increase the solution viscosity and to further allow the formation of hollow fibers during electrospinning. The microstructural evolution during thermal annealing at temperatures between 600 °C and 1,400 °C revealed a crystal growth evolution. Strong grain coarsening was also observed above 1,200 °C, leading to linear and branched nanoparticle chains that were formed from the finer initial hollow nanofibers. The faceting of the nanoparticles took place under low-energy surfaces.

Template-Assisted Synthesis

The synthesis of 1D nanostructures using templates supposes a very convenient and



Growth of 1-D Oxide Nanostructures, Fig. 6 (**a**) Schematics of a typical electrospinning setup using a grounded static collector. Reprinted with permission from reference [9]. (**b**) Typical SEM image showing as-spun $BaZrO_3$ fibers and (**c**) TEM image of faceted $BaZrO_3$ nanoparticle chains formed after crystallization at 1,400 °C (Adapted from Ref. [10]. Reproduced with permission from The Royal Society of Chemistry (RSC) on behalf of the Consejo Superior de Investigaciones Cientificas (CSIC) and the RSC)

versatile method. This technique represents a conceptually straightforward approach to the generalized synthesis of nanostructures first introduced by Martin [11]. In this method, the template is used as scaffolds in which are possible to synthesize materials with different morphologies. The obtained material within the template's pores is composed of a nanostructure that matches that of the template's morphology. The chemical and physical properties of the template membranes (i.e., pore size and geometry) enable a high degree of control over the dimensions of the resulting 1D nanostructures.

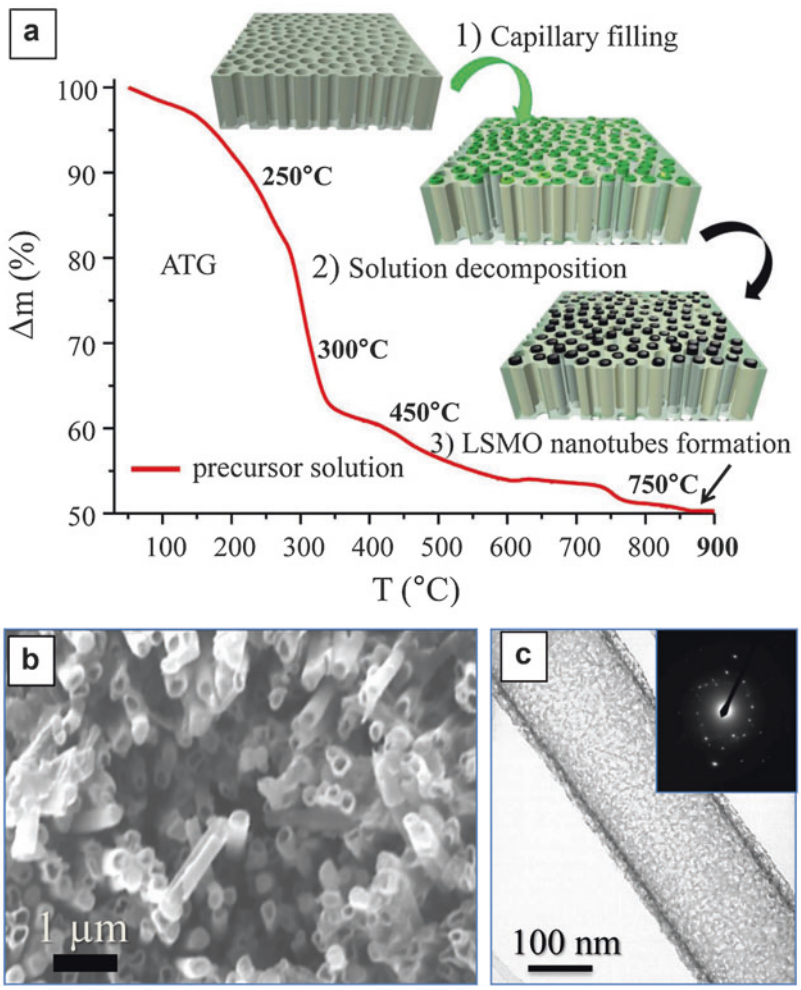
Many different types of template membranes have been successfully used, and it is generally agreed that this synthesis method is a relatively simple, reasonably high-throughput, and cost-effective procedure and, therefore, of general applicability. There are two main types of templates: "soft" templates and "hard" templates. Naturally occurring gels, micelles, and chitin scaffolds are considered as soft templates. Other soft templates include DNA strands, polymer matrices, and reverse micelles. Hard templates are typically associated with materials such as track-etched polymeric membranes and anodic alumina membranes which are both commercially available.

Direct Template Filling The direct filling of a template with a precursor solution is one of the most straightforward and versatile methods for preparing nanowires or nanorods. In the following section, we present some interesting examples on the preparation of 1D nanostructures of mixed-metal oxides by combining sol–gel processing and hard nanoporous templates.

A key issue for the synthesis of 1D nanostructures following this method is the preparation of stable starting precursor solutions which can be tuned according to the desired composition and its ability to fill the template nanopores by capillary force. Metallic nitrates which are completely soluble in water and possess long-term stability are often the precursors of choice. Good control over viscosity, wettability, stoichiometry, and stability of the precursor solution has been identified as crucial parameters for the template-mediated synthesis. In some cases, a polymerizing agent such as ethylene glycol (EG) can be used in order to reach the optimum viscosity. After template filling by capillary forces, different thermal treatments under an oxygen-rich atmosphere are further applied in a tubular oven for the ultimate phase formation and crystallization. However, the same precursor solution may have a very different wetting behavior when filling high surface tension templates such as nanoporous alumina or low surface tension track-etched porous polymer templates with low surface tension. This scenario gives rise to the formation of nanorods or nanotubes, respectively.

Anodic Aluminum Oxide Membrane Templates A recent example, showing the applicability of template-mediated synthesis, is the formation of self-standing $La_{0.7}Sr_{0.3}MnO_3$ nanotubes that were synthesized using a sol–gel precursor solution combined with porous alumina templates [12]. The nanotubes were prepared from a precursor solution containing 0.5 mol/L of lanthanum, strontium, and manganese nitrates in the proportion 07:0.3:1, respectively, in water. In this case, commercially available anodic aluminum oxide (AAO) membrane templates (Whatman®, pore diameters 100 and 200 nm) were used. The templates were immersed in the precursor solution at room temperature for different holding times ranging from 1 h to 56 h. Subsequently, the templates containing the precursor gel were heated at high temperatures by using a tubular furnace in oxygen atmosphere. The resulting nanotubes displayed outer diameters of 100 nm or 200 nm and wall thicknesses of approximately 20 nm−40 nm (see Fig. 7), depending on the template pore dimension with lengths of few micrometers. The synthesized $La_{0.7}Sr_{0.3}MnO_3$ nanotubes were polycrystalline due to the heterogeneous nucleation within the AAO pore walls and ferromagnetic with a Curie temperature of 350 K, i.e., close to the reported $La_{0.7}Sr_{0.3}MnO_3$ bulk value.

Track-Etched Polymer Templates One of the main drawbacks of using hard templates for the synthesis of 1D nanostructures relates to the ultimate template removal which is a requirement for

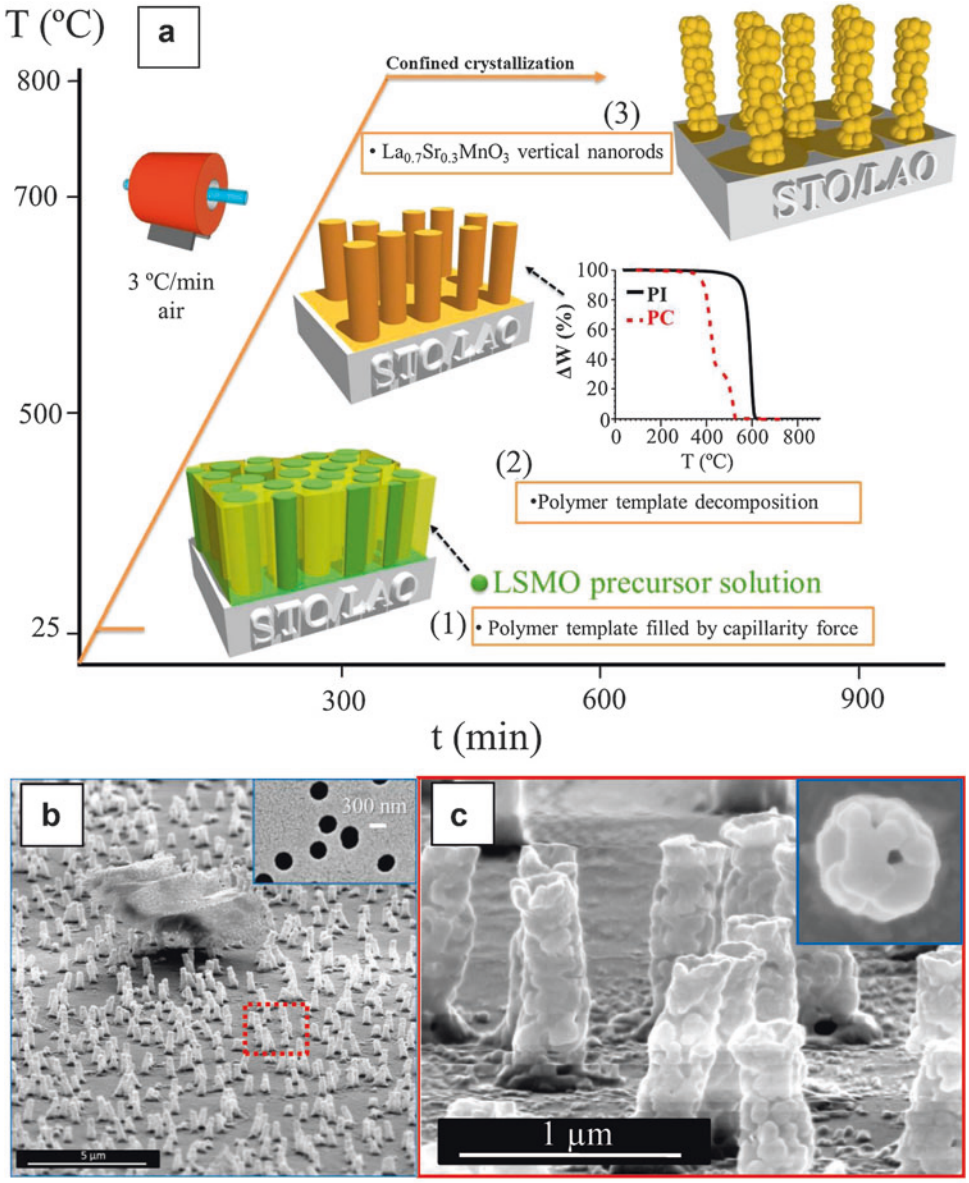


Growth of 1-D Oxide Nanostructures, Fig. 7 (**a**) Thermogravimetric analysis of the nitrate precursor solution, heating ramp 3 °C/min until a maximum temperature of 900 °C was reached. Schematic representation of AAO nanopore filling process and posterior nanotube formation at elevated temperatures is described. (**b**)Top view field emission SEM image of vertically aligned nanotubes obtained using an AAO membrane with 200 nm pore size (800 °C, 5 h) after partial removal of the AAO template. (**c**) TEM image of an isolated $La_{0.7}Sr_{0.3}MnO_3$ nanotube, the inset shows the corresponding selected area electron diffraction (SAED) pattern (Reproduced from Ref. [12] with permission from the Centre National de la Recherche Scientifique (CNRS) and Elsevier, with kind permission from Springer Science and Business Media)

many technological applications. However, polymer templates display low decomposition temperatures (300–500 °C) which enable their easy elimination during growth processes when the right thermal conditions are chosen. In the following, we will present some examples concerning the preparation of 1D nanostructures of mixed-metal oxides by combining sol–gel processing and track-etched nanoporous polymer templates.

Vertical polycrystalline $La_{0.7}Sr_{0.3}MnO_3$ nanorods can be synthesized using chemical solution deposition assisted by polycarbonate (PC) or polyimide (PI) nanoporous templates on $SrTiO_3$ (STO) and $LaAlO_3$ (LAO) substrates [13]. In this case, the key point to maintain the verticality of the nanorods synthesized at high temperatures was the similar perovskite crystallographic structures of $La_{0.7}Sr_{0.3}MnO_3$ nanorods and substrates. The difference in the lattice parameters between $La_{0.7}Sr_{0.3}MnO_3$ and the substrates was minimal, yielding a nominal tensile lattice mismatch of $\varepsilon \approx 0.9$ % for STO and a larger but compressive magnitude of $\varepsilon \approx -2$ % for LAO substrates. Consequently, filling nanoporous track-etched polymers (≈1 μm thick) supported on STO and LAO substrates with the nitrate-based precursor solution under a thermal treatment at 800 °C resulted in a final nanostructured perovskite substrate with vertically oriented nanorods of 1 μm high, as summarized in Fig. 8.

In this case, the nanorods were polycrystalline and were indexed according to the standard



Growth of 1-D Oxide Nanostructures, Fig. 8 (**a**) Schematics of the different stages during the formation process of $La_{0.7}Sr_{0.3}MnO_3$ vertical nanorods on STO or LAO substrates. (*1*) Track-etched nanoporous polymer template supported on a LAO or STO substrate filled with the $La_{0.7}Sr_{0.3}MnO_3$ precursor solution. (*2*) At mild temperatures (500–600 °C), the polymer template is decomposed as thermogravimetric analysis indicates. Thus, surface was fully covered with vertical amorphous nanorods that kept the position of initial pores. (*3*) At high temperatures (800 °C), the crystallization on $La_{0.7}Sr_{0.3}MnO_3$ nanorods is promoted. (**b**) Tilted FEG-SEM image of $La_{0.7}Sr_{0.3}MnO_3$ vertical nanorods grown on top of a (100)-STO substrate. $La_{0.7}Sr_{0.3}MnO_3$ nanorods have a 250 nm of diameter and a density of 1×10^8 cm^{-2}. (**c**) FEG-SEM enlarged image of $La_{0.7}Sr_{0.3}MnO_3$ nanorods with a diameter of 250 nm grown on top of a STO single crystal. The inset shows a FEG-SEM image showing the top view details of $La_{0.7}Sr_{0.3}MnO_3$ nanorod grain size (Reproduced with permission from The Royal Society of Chemistry (RSC) on behalf of the Centre National de la Recherche Scientifique (CNRS) and the RSC)

rhombohedral structure and space group $R-3c$ of $La_{0.7}Sr_{0.3}MnO_3$ perovskite. They were ferromagnetic with Curie temperature (Tc) values around ~360 K, i.e., close to the reported bulk $La_{0.7}Sr_{0.3}MnO_3$ value. Altogether, these results prove the capability of the chemical solution deposition method combined with track-etched polymer templates to generate homogeneous nanostructured surfaces with vertical $La_{0.7}Sr_{0.3}MnO_3$ nanorods displaying interesting magnetic properties.

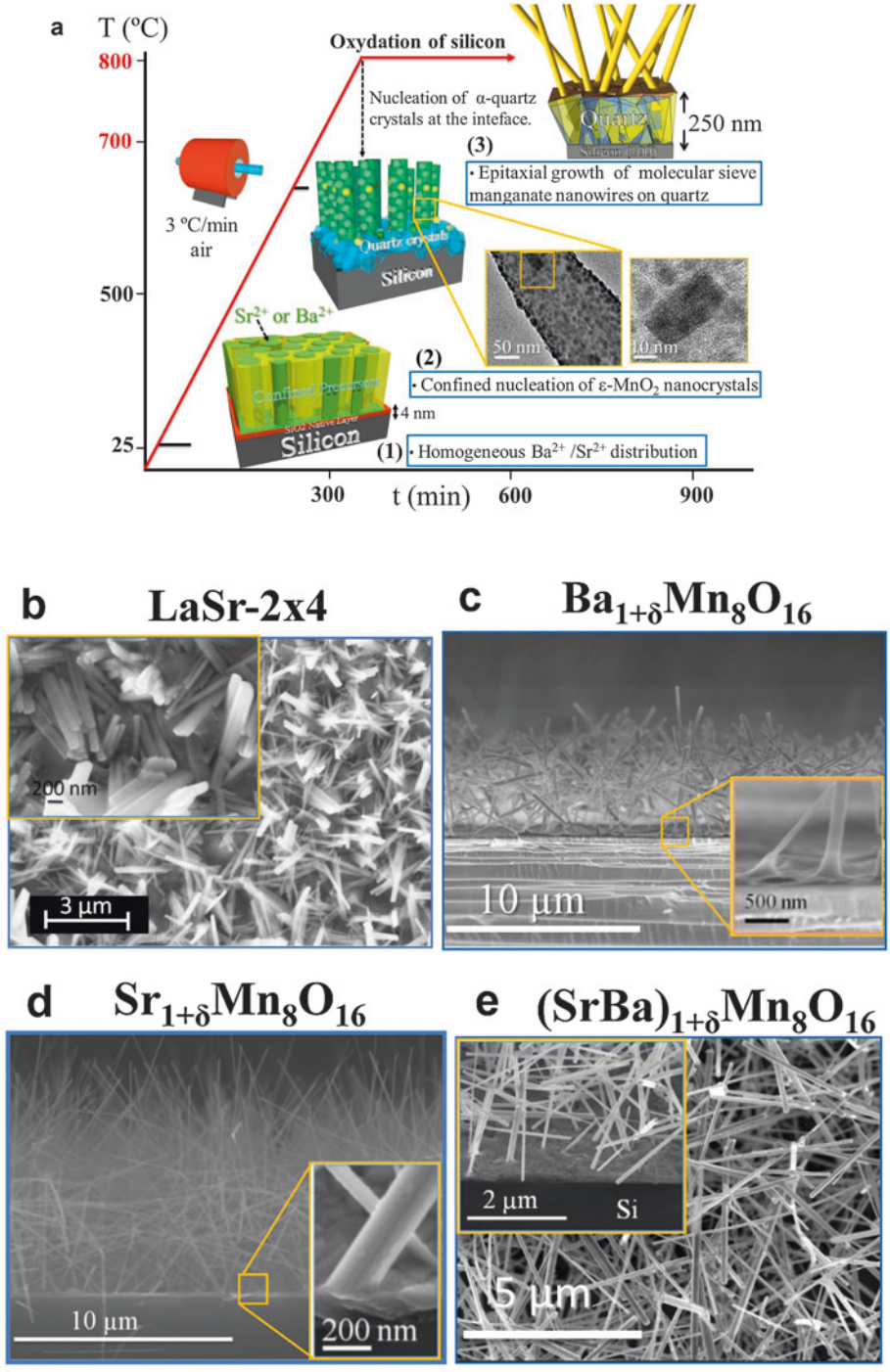
Epitaxially Stabilized Metastable Phases of 1D Oxide Nanostructures It has been shown that polymer template nanopores can be used not only as a mere physical spatial constrains but also as nanoreactors in which confined nucleation favors the stabilization of particular metastable seed nanostructures. Additionally, once the polymer template is decomposed during thermal treatment, the atomic structure of the substrate can be effectively used as a second template for phase stabilization and epitaxial growth. In the following, we present two examples of the fundamental role of the substrate as a determinant for polymorph stabilization, geometry, and orientation of the grown nanostructured oxides.

Following the same synthesis method as described above and the same precursor solution, the stabilization of a different polymorph of $La_{0.7}Sr_{0.3}MnO_3$ with highly interesting magnetic properties was achieved by using silicon substrates. In this case, the nanowire growth mechanism entails the controlled growth of an α-quartz thin film at the silicon surface which allowed octahedral molecular sieve (OMS) nanowires of different compositions to stabilize and crystallize [14, 15]. Manganese oxide OMS are 1D open framework structures with nanometric tunnel sizes. The tunnels are built up by edge-shared and corner-shared [MnO_6] octahedral units leading to different pore size materials. The shape of the tunnels is expressed as the number of constituting octahedral units ($n \times m$) and is characteristic of each porous manganese oxide.

Figure 9 shows a detailed scheme of the proposed mechanism for the formation of OMS nanowires on top of silicon substrates, based on the combination of soft chemistry routes and confined epitaxial growth. First, precursor solutions with a composition that can be tuned in order to fill the polymer template nanopores by capillary forces were used in step 1 in Fig. 9. The confinement imposed by the polymer template in high aspect ratio nanopores induced the formation of ε-MnO_2 nanoparticles at low temperatures (~500 °C) during the calcination of the precursor solution, step 2 in Fig. 9. These ε-MnO_2 phase nanoparticles served as seeds for the further growth of manganate nanowires at higher temperatures, once the polymer template has been eliminated. The synthesis of manganate structures required sufficiently high temperatures to enhance atomic mobility and, thus, to allow the diffusion of the alkaline earth and rare earth cations into the empty 1D channels of the ε-MnO_2 structure.

The second essential requirement for the growth of manganese oxide-based OMS nanowires on silicon is the role of the α-quartz layer acting as an epitaxial template for the stabilization of the nanowires. α-quartz thin films were obtained after thermally activated devitrification of the native amorphous silica surface layer assisted by the heterogeneous catalysis driven by alkaline earth Sr^{2+} cations present in the precursor solution [14]. This polycrystalline α-quartz film rendered the convenient interface lattice matching to the complex oxide nanostructures for the epitaxial growth of vertically oriented LaSr- 2 × 4 manganate nanowires at 800 °C, step 3 in Fig. 9, as revealed by scanning transmission electron microscopy (STEM) analysis.

The proposed unit cell of LaSr-2 × 4 nanowires based on the X-ray diffraction and STEM analysis was monoclinic with lattice parameters $a = 13.8$ Å, $b = 5.7$ Å, $c = 21.8$ Å, and $\beta = 101$ ° where the long axis of the nanowires is along the b crystallographic direction. High-resolution TEM images revealed that the orientation relationship can be described as (010) LaSr-2 × 4// (010) α-quartz and an in-plane epitaxial relationship given by [20–2] LaSr-2 × 4// [−101] α-quartz. Interestingly, these synthesized LaSr-2 × 4 nanowires exhibited ferromagnetic ordering with strongly enhanced Curie temperature ($T_c > 500$ K) that probably results from the new



Growth of 1-D Oxide Nanostructures, Fig. 9 (**a**) Schematics of the different stages of the crystallization process of single-crystalline OMS nanowires on silicon. (*1*) Track-etched nanoporous polymer template supported on a SiO_2/Si substrate filled with the chemical precursor solution allowing a homogeneous distribution of Ba^{2+} or Sr^{2+} melting agent. (*2*) At mild temperatures (500–600 °C) MnO_2 nanoparticles nucleated and stabilized due to the confinement in high aspect ratio nanopores (see *inset* HRTEM images showing the nucleation of MnO_2

crystallographic order between the La^{3+} and Sr^{2+} cations ordered inside the 2 × 4 1D channels defined by the MnO_6 octahedral and from the mixed valence of manganese [16].

This synthetic approach based on track-etched polymer templates technology combined with a chemical solution deposition route provides direct access to the monolithic integration of composition controlled vertical OMS nanowires, epitaxially grown on silicon substrates. As a result, this procedure has been successfully extended to other nanowire compositions including $Ba_{1+\delta}Mn_8O_{16}$, $Sr_{1+\delta}Mn_8O_{16}$, and $(BaSr)_{1+\delta}Mn_8O_{16}$ [17], as displayed in Fig. 9.

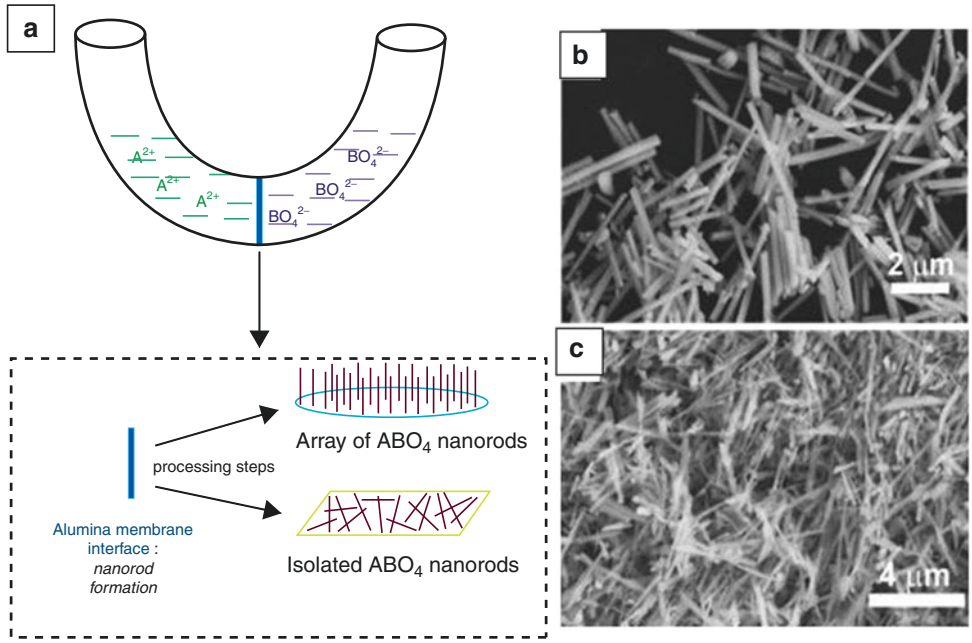
The directional growth of LaSr-2 × 4 nanowires when using gadolinium-doped CeO_2 substrates with the fluorite structure illustrates the determinant influence of the substrate on the nanostructure characteristics [18]. In this case, using the same growth method described above at temperatures above 1,000 °C, the system evolved toward its equilibrium state to form horizontal and oriented LaSr-2 × 4 nanowires with the [010] nanowire's longitudinal axis parallel to the [110] and [1–10] directions of the substrate. The large intrinsic anisotropy of the monoclinic LaSr-2 × 4 structure is what promoted the 1D growth along the nanowire [010] direction (the direction of the smallest interplanar distance). This fact combined with the low isotropic residual misfit along the <110> directions of the CeO_2 substrate (the estimated residual misfit is of 0.74 %) permitted the self-assembly of LaSr-2 × 4 nanowires along these directions [18].

The Template-Assisted U-Tube Method The template-assisted U-tube method is an excellent illustration on how the pores of a nanoporous hard template may act as nanoreactors. In this case, it is possible to fabricate 1D oxide materials in a more responsible and environmentally friendly manner. This approach is based on using a double-diffusion setup in which a desired membrane is sandwiched between the precursor half-cells of a U-shaped tube, as schematically shown in Fig. 10. Precursor solutions are suitably added to both sides of the so-called U-tube cell. The diffusion causes the precursor ions to enter into the pore channels where they react, thereby, forming 1D structures (nanorods or nanotubes) [19]. In this case, the template nanopores also act as nanoreactors. This technique is readily compatible with a variety of common chemical reactions, such as reduction of metal precursors or precipitation reactions, therefore, offering the possibility to synthesize a wide range of nanomaterials.

The U-tube setup has been used to grow high purity single-crystalline $BaWO_4$ nanorods and $BaCrO_4$ nanorods at large scale with different controllable size [19]. Alumina template membranes with pore sizes of 100 nm or 200 nm were mounted between the two halves of a U-tube cell. The half-cells were then filled with equimolar solutions of $Ba(NO_3)_2$ and either Na_2WO_4 or Na_2CrO_4 solutions to generate barium tungstate and barium chromate, respectively. After immersion times of up to 12 h at room temperature, the precursors diffused into the nanopores of the alumina template, resulting in the nanorods formation.

Template-Assisted Electrophoretic Deposition The use of template-based electrophoretic deposition has been demonstrated to be a very effective way to grow nanowires and nanotubes of complex oxides. Electrochemical deposition, also known as electrodeposition, involves the oriented diffusion of charged reactive species through a solution when an external electric field is applied along with the reduction of the charged growth species at the growth or deposition surface (which also serves as an electrode). A thin

Growth of 1-D Oxide Nanostructures, Fig. 9 (continued) nanoparticles in confined precursor nanorods). (*3*) Next devitrification of the SiO_2 layer and nucleation of α-quartz took place at the interface where epitaxial growth of OMS nanowires on the quartz is promoted at high temperatures (800 °C). FEG-SEM image of (**b**) LaSr-2 × 4 OMS nanowires, (**c**) $Ba_{1+\delta}Mn_8O_{16}$ nanowires, (**d**) $Sr_{1+\delta}Mn_8O_{16}$ nanowires, and (**e**) $SrBa_{1+\delta}Mn_8O_{16}$ nanowires grown on silicon substrate (Reproduced with permission from The Royal Society of Chemistry (RSC) on behalf of the Centre National de la Recherche Scientifique (CNRS) and the RSC)

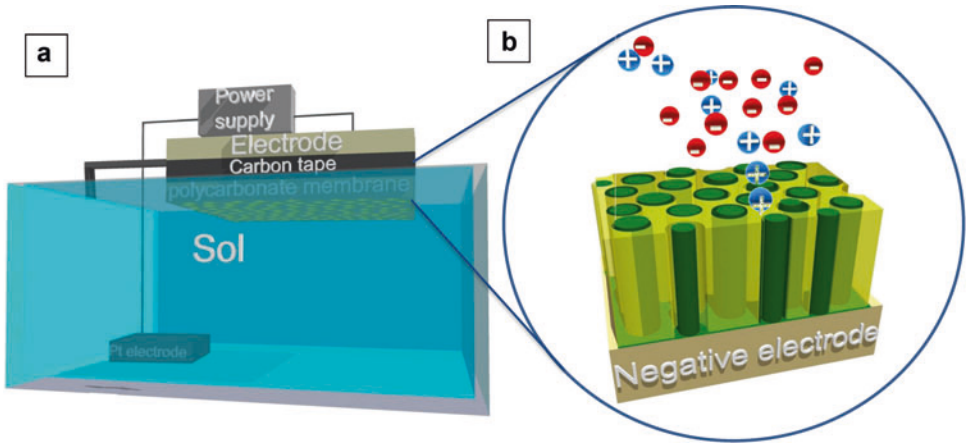


Growth of 1-D Oxide Nanostructures, Fig. 10 (**a**) Schematic representation of the U-tube setup to synthesize 1D oxide nanostructures in which a wide variety of precursor solutions can react within the nanosized pores of a membrane, often composed of anodic alumina or polycarbonate. Typical SEM micrographs of (**b**) $BaWO_4$ nanorods and (**c**) $BaCrO_4$ nanorods grown by this method (Reprinted with permission from Ref. [19]. Copyright (2004) American Chemical Society)

conducting metal film is first coated on one side of the porous membrane to serve as the cathode for the electroplating. In general, this method is only applicable to electrically conductive materials such as metals, alloys, semiconductors, and electrically conductive polymers. In the case of simple oxides, electrodeposition of the metal into the pores is followed by oxidation. Examples of this method are ZnO, In_2O_3, and SnO_2 nanowires.

However, direct electrodeposition is not suitable for multicomponent oxide nanowires as different cations would have different ionic sizes and diffusivity. Therefore, the electrophoretic deposition technique has been widely explored, particularly for the deposition of ceramic and organo-ceramic materials onto a cathode from colloidal dispersions. When dispersed in a polar solvent or an electrolyte solution, the surface of nanoparticles develops an electrical charge. Upon the application of an external electric field, charged particles are set in motion, as illustrated in Fig. 11. This type of motion is referred to as electrophoresis. Electrophoresis in the nanopores of a template offers a cost-effective and simple alternative to synthesize 1D complex oxide nanostructures.

Limmer et al. combined sol–gel preparation with electrophoretic deposition to produce nanorods of various complex oxides [20]. In this case, a conventional sol–gel processing was applied to the synthesis of various sols. By controlling the appropriate preparation of the sol, nanometer particles of a desired stoichiometric composition were formed and electrostatically stabilized by pH adjustment. Using radiation tracked-etched polycarbonate membranes with an electric field of ≈ 1.5 V/cm, they could grow nanowires with diameters ranging from 40 nm to 175 nm and lengths of 10 μm, corresponding to the thickness of the membrane. Samples were then dried at ∼100 °C for several hours and then placed in an oven and fired at 700 °C for 1–30 min. This step was set to burn off the PC membranes, to make the nanorods dense and to crystallize the material. Nanowires of many technologically important oxides like $BaTiO_3$ and $Pb(Ti,Zr)O_3$



Growth of 1-D Oxide Nanostructures, Fig. 11 (**a**) Diagram of the experimental setup used for growth of nanowires by sol–gel electrophoresis. One electrode is a Pt mesh, and the other electrode is Al. The PC membrane is attached to this electrode with a piece of double-sided conductive (carbon) tape to provide a conductive path. This electrode is placed on top of and just in contact with the sol, and the sol is drawn into the membrane pores by capillary action. (**b**) Schematic showing the filling process. Positively charged sol particles (*blue spheres*) are moving electrophoretically toward the negative electrode, depositing at the bottom of the pore, while the negatively charged counterions (*red spheres*) are moving in the opposite direction (Adapted from Ref. [20])

perovskites and of the layered perovskite $Sr_2Nb_2O_7$ have been prepared by this method.

In this context, $PbZrO_3$ nanotubes where produced by using electrophoresis in AAO templates [21]. The difference in the final nanostructures may be related to the different electrochemistry, according to the type of template used.

Acknowledgments This work was partially funded by a PEPS project of Cellule Energie INSIS-CNRS (1D-RENOX) to ACG and the Spanish MICINN (MAT2011-28874-c02-01, Consolider NANOSELECT (CSD2007-00041)) and Generalitat de Catalunya (SGR2014-00753). ACG acknowledges Ecole Centrale de Lyon for the financial support under the BQR 2015 project.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Electrospinning
- ▶ Growth of Oxide Nanowires
- ▶ Nanomaterials
- ▶ Nanoparticles
- ▶ Nanotechnology
- ▶ Nanowires
- ▶ Self-Assembly
- ▶ Self-Assembly of Nanostructures
- ▶ Sol–Gel Method

References

1. Wang, F., Dong, A., Sun, J., Tang, R., Yu, H., Buhro, W.E.: Solution−liquid–solid growth of semiconductor nanowires. Inorg. Chem. **45**(19), 7511–7521 (2006)
2. Murray, C.B., Kagan, C.R., Bawendi, M.G.: Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. Annu. Rev. Mater. Sci. **30**(1), 545–610 (2000)
3. Urban, J.J., Yun, W.S., Gu, Q., Park, H.: Synthesis of single-crystalline perovskite nanorods composed of barium titanate and strontium titanate. J. Am. Chem. Soc. **124**(7), 1186–1187 (2002)
4. Walton, R.I.: Subcritical solvothermal synthesis of condensed inorganic materials. Chem. Soc. Rev. **31**(4), 230–238 (2002)
5. Fang, Y.-P., Xu, A.-W., Song, R.-Q., Zhang, H.-X., You, L.-P., Yu, J.C., Liu, H.-Q.: Systematic synthesis and characterization of single-crystal lanthanide orthophosphate nanowires. J. Am. Chem. Soc. **125**(51), 16025–16034 (2003)
6. Hu, Y.-M., Gu, H.-S., Zhou, D., Wang, Z., Chan, H.L.-W., Wang, Y.: Orientation-control synthesis of $KTa_{0.25}Nb_{0.75}O_3$ nanorods. J. Am. Ceram. Soc. **93**(3), 609–613 (2010)
7. Mao, Y., Banerjee, S., Wong, S.S.: Large-scale synthesis of single-crystalline perovskite nanostructures. J. Am. Chem. Soc. **125**(51), 15718–15719 (2003)

G

8. Einarsrud, M.-A., Grande, T.: 1D oxide nanostructures from chemical solutions. Chem. Soc. Rev. **43**(7), 2187–2199 (2014)
9. Teo, W.E., Ramakrishna, S.: A review on electrospinning design and nanofibre assemblies. Nanotechnology **17**(14), R89–R106 (2006)
10. Calleja, A., Granados, X., Ricart, S., Oro, J., Arbiol, J., Mestres, N., Carrillo, A.E., Palmer, X., Cano, F., Tornero, J.A., Puig, T., Obradors, X.: High temperature transformation of electrospun $BaZrO_3$ nanotubes into nanoparticle chains. Cryst. Eng. Commun. **13**(24), 7224–7230 (2011)
11. Martin, C.R.: Nanomaterials: a membrane-based synthetic approach. Science **266**(5193), 1961–1966 (1994)
12. Carretero-Genevrier, A., Frontera, C., Hassini, A., Oro-Sole, J., Moreno, C., Obradors, X., Puig, T., Mestres, N.: Chemical solution growth of $La_{0.7}Sr_{0.3}MnO_3$ nanotubes in confined geometries. J. Sol–Gel Sci. Technol. **73**(3), 620–627 (2015)
13. Carretero-Genevrier, A., Gázquez, J., Puig, T., Mestres, N., Sandiumenge, F., Obradors, X., Ferain, E.: Vertical nanostructures: vertical (La,Sr)MnO_3 nanorods from track-etched polymers directly buffering substrates (Adv. Funct. Mater. 6/2010). Adv. Funct. Mater. **20**(6) (2010)
14. Carretero-Genevrier, A., Gich, M., Picas, L., Gazquez, J., Drisko, G.L., Boissiere, C., Grosso, D., Rodriguez-Carvajal, J., Sanchez, C.: Soft-chemistry-based routes to epitaxial alpha-quartz thin films with tunable textures. Science **340**(6134), 827–831 (2013)
15. Carretero-Genevrier, A., Puig, T., Obradors, X., Mestres, N.: Ferromagnetic 1D oxide nanostructures grown from chemical solutions in confined geometries. Chem. Soc. Rev. **43**(7), 2042–2054 (2014)
16. Carretero-Genevrier, A., Gazquez, J., Idrobo, J.C., Oro, J., Arbiol, J., Varela, M., Ferain, E., Rodriguez-Carvajal, J., Puig, T., Mestres, N., Obradors, X.: Single crystalline $La_{0.7}Sr_{0.3}MnO_3$ molecular sieve nanowires with high temperature ferromagnetism. J. Am. Chem. Soc. **133**(11), 4053–4061 (2011)
17. Carretero-Genevrier, A., Oro-Sole, J., Gazquez, J., Magén, C., Miranda, L., Puig, T., Obradors, X., Ferain, E., Sanchez, C., Rodriguez-Carvajal, J., Mestres, N.: Direct monolithic integration of vertical single crystalline octahedral molecular sieve nanowires on silicon. Chem. Mater. **26**(2), 1019–1028 (2013)
18. Carretero-Genevrier, A., Gazquez, J., Magen, C., Varela, M., Ferain, E., Puig, T., Mestres, N., Obradors, X.: Chemical synthesis of oriented ferromagnetic LaSr-2 × 4 manganese oxide molecular sieve nanowires. Chem. Commun. (Camb.) **48**(50), 6223–6225 (2012)
19. Mao, Y., Wong, S.S.: General, room-temperature method for the synthesis of isolated as well as arrays of single-crystalline ABO_4-type nanorods. J. Am. Chem. Soc. **126**(46), 15245–15252 (2004)
20. Limmer, S.J., Seraji, S., Wu, Y., Chou, T.P., Nguyen, C., Cao, G.Z.: Template-based growth of various oxide nanorods by Sol–Gel electrophoresis. Adv. Funct. Mater. **12**(1), 59–64 (2002)
21. Nourmohammadi, A., Hietschold, M.: Template-based electrophoretic growth of $PbZrO_3$ nanotubes. J. Sol–Gel Sci. Technol. **53**(2), 342–346 (2010)

Growth of Carbon Nanotubes

▶ Synthesis of Carbon Nanotubes

Growth of CNTs

▶ Synthesis of Carbon Nanotubes

Growth of Oxide Nanowires

▶ Growth of Silica Nanowires

Growth of Silica Nanowires

Praveen Kumar Sekhar[1] and Shekhar Bhansali[2]
[1]Electrical Engineering, School of Engineering and Computer Science, Washington State University Vancouver, Vancouver, WA, USA
[2]Department of Electrical and Computer Engineering, Florida International University, Miami, FL, USA

Synonyms

Growth of oxide nanowires; Growth of silicon dioxide nanowires; Manufacturing silica nanowires; Nucleation of silica nanowires

Definition

The evolution of nano (one-billionth of a meter)-sized material made of silicon and oxygen in a

wire-like morphology is referred to as growth of silica nanowires. The evolution may be from a substrate or freestanding.

Introduction

Nanotechnology continues to induce a paradigm shift in electronic, chemical, biomedical, sensor, and related industries [1]. The use of nanoscience enables (a) superior performance of the materials and devices, (b) increased sensitivity levels in sensors, (c) faster response in devices, and (d) high-density synthesis of materials and technology platforms leading to cost-effective manufacturing. These properties stem from the fact that nanostructures have enhanced surface area of volume ratios, allowing surface phenomenon to dominate inertial phenomenon. Consequently a small variation in their dimensions alters the surface area/volume ratio and leads to significant changes in physical, optical, and electrical properties, even resulting in new class of materials and devices. Additionally, the large surface-to-volume ratios and their low dimensions make them ideal candidates to probe events at submolecular and atomic scale. One of the prominent areas of research in nanotechnology is the synthesis of nanomaterials [2].

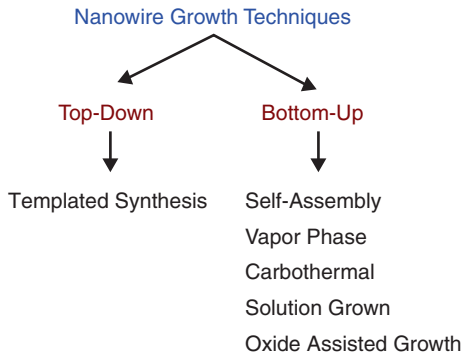
Of the numerous one-dimensional (1D) nanostructures, silica (SiO_2) nanowires are a subject of intense research [3, 4] for numerous applications such as localization of light, low-dimensional waveguides for functional microphotonics, scanning near field optical microscopy (SNFOM), optical interconnects in integrated optical microchip, sacrificial templates, biosensors, and optical transmission antennae. Silica nanowires are made up of silicon and oxygen, the two most abundant elements in the earth's crust. Their combination as bulk silica has been used in a wide variety of applications ranging from colloid chemistries to MEMS and CMOS applications [5, 6].

The surface of silica can react with various compounds such as amine, carboxyl, or thiol groups to make it functional for biosensing applications [7]. Although silica nanowires have been observed to grow at the ends of CVD reactors since the 1960s [4], they were not actively exploited till 2004. The properties of silica nanowires are unique compared to bulk or colloidal silica [8]. As nanowires would be a critical building block of any nanosystem, the ability to synthesize them at a desired location in a controlled manner is of significant importance.

This chapter reviews silica nanowire growth and the growth mechanisms. First, a general overview of the "top-down" and "bottom-up" synthesis of nanowires is presented. After brief discussion of the various growth mechanisms in "top-down" and "bottom-up" synthesis, the vapor liquid solid (VLS) technique, a preferred method to grow silica nanowires, is described in detail. Next, two new approaches in selective growth of silica nanowires using a combination of VLS and oxide-assisted growth (OAG) mechanism are presented along with a discussion on the mass manufacturing aspects of the nanoscale wires.

Nanowire Nucleation Techniques

Understanding the growth mechanism of nanowires is critical for enabling high-density and controlled synthesis of silica nanowires. The structure and morphology of the silica nanowires is dictated by the growth mechanism used. Figure 1 schematically illustrates the different techniques of growing nanowires in general. Top-down nanowires are routinely



Growth of Silica Nanowires, Fig. 1 Classification of nanowire growth methodologies

grown using templates. The preferred templates for the growth are anodized aluminum oxide (AAO), track etched polymers, and in some instances porous silicon. Bottom-up approach is an additive process, which builds on material from an atomic scale (molecule by molecule or atom by atom) to evolve one-dimensional nanostructures. Bottom-up approaches include vapor phase growth, solution-grown mechanism, oxide-assisted growth, carbothermal reduction technique, and self-assembly process. All of these growth techniques require nucleation of the wire.

Nucleation Techniques: Top Down and Bottom Up

There exist two generic methods for the growth of nanostructures, namely, the "top-down (TD)" and the "bottom-up (BU)" approach. TD approach involves machining and etching techniques to remove material from a bulk form to a nanoscale size. Templated method falls in the category of the TD approach in the formation of nanowires. On the other hand, BU approach is an additive process, which builds on material from an atomic scale (molecule by molecule or atom by atom) to evolve one-dimensional nanostructures.

Top Down: Templated Method

The "template" is a fundamental structure within which a network forms. Deposition of the material in the template and the removal of the template leaves a nanowire made from deposited material. The external morphology of the resulting nanowire can, when deposited carefully, replicate the structural details of the template. Various predefined structures such as polymer membranes, anodic alumina, and porous silicon of fixed geometry have been used as templates. Once the template is fabricated, nanowires of desired geometry (diameter dictated by the pore size and length determined by the pore length) are derived. Electrochemical methods, electroless techniques, chemical vapor deposition (CVD), sol–gel and chemical polymerization–based deposition mechanism are used to fill the pores. The porous part or the scaffold is selectively removed using appropriate etching chemistry leaving behind the nanowires. The template method is convenient and simple. Nanowires of different compounds can be easily produced harnessing the properties pertaining to specific applications. Christian [9] reported a novel iron-based surfactant to form rod-like micelle structures, which may be, used to template the formation of silica nanowires with diameters less than 5 nm. The main drawbacks of this method are the inability to produce single crystalline wires and limited geometries (i.e., fixed diameter and length).

Bottom Up: Self-assembly Process

Self-assembly process is a viable bottom-up route for manufacturing nanostructures. It is defined as the creation of material from its constituent components in a spontaneous natural manner, i.e., by an interaction between the components or by a specific rearrangement of them that proceeds naturally without any special external impetus. Kievsky and Sokolov [10] reported a novel modification of acidic synthesis of nanoporous silica fibers that assembled fibers of uniform in size and shape. Each fiber was shaped like a hexagonal cylinder of approximately 2.0 μm in diameter and approximately 4.8 μm in length. The yield of the assembled fibers was virtually 100 %.

BU: Solution-Grown Method

In this method, molecular capping agents (which control the kinetics of crystal growth by allowing selective nucleation on certain facets) along with suitable seed materials are dispersed in a solvent (a suitable chemical compound comprising material of interest) medium. The solvent is then heated and pressurized until it becomes a supercritical fluid. When the concentration of the solvent exceeds solubility limit, the seed ejects the elemental compound in the solvent as nanowires. This forms the basis of solution-grown nanowires. Kita et al. [11] successfully synthesized silicon oxide nanowires using a low-temperature hydrothermal treatment (200 °C) without the use of metallic catalysts. Scanning electron microscopy (SEM) and transmission electron microscopy (TEM) revealed nanowires having uniform distribution and straight morphology with diameters ranging from 50 to 100 nm and lengths up to ca. 60 μm. The nanowires were produced by

hydrothermal treatment based on the original concept of slow structure formation by the dissolution of bulk glass source materials (borosilicate glasses), which can be used to supply a continuous source solution.

BU: Carbothermal Reduction Technique

In this method, elevated temperatures are used which causes activated carbon to behave as a catalyst thus reducing metal oxides/nitrides/carbides to their corresponding nanowires with release of carbon monoxide. Equations 1 and 2 represent the carbothermal reduction technique.

$$\text{Metal oxide} + \text{C} = \text{Metal sub-oxide} + \text{CO} \quad (1)$$

$$\text{Metal sub-oxide} + \text{O}_2 = \text{Metal oxide nanowires} \quad (2)$$

Lin and Lin [12] achieved the growth of SiO_2 nanowires via carbothermal reduction of CuO/C powders at a temperature of 1,000–1,100 °C in flowing Ar with the flow rate of 100 sccm.

BU: Oxide-Assisted Growth (OAG) Mechanism

Metal catalysts are not required for the synthesis of nanowires by this bottom-up fabrication method. Stoichiometric or non-stoichiometric silicon oxide (Si*x*O*y*) in its vapor form induces the nanowire growth. For example, the growth of SiO_2 nanowires [13] was promoted by SiO vapor derived from Si powder targets or SiO_2 thin films. Equations explaining the oxide-assisted growth mechanisms are as follows:

$$\text{Si}_x\text{O} \rightarrow \text{Si}_{x-1} + \text{SiO}(x > 1) \quad (3)$$

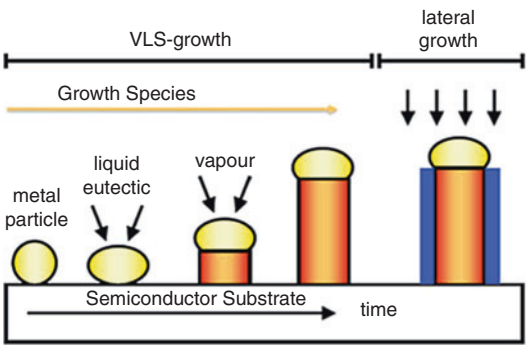
$$2\text{SiO} \rightarrow \text{Si} + \text{SiO}_2 \quad (4)$$

These decompositions result in the precipitation of Si nanoclusters, which act as the nuclei of the silicon nanowires covered by shells of silicon oxide.

BU: Vapor Phase Techniques

Typically, nanowires or one-dimensional nanostructures can be fabricated by promoting the crystallization of solid-state structures along one direction. This can be achieved either by the growth of intrinsically anisotropic crystal structure or the use of templates with one-dimensional morphologies directing the formation of one-dimensional nanostructures or the introduction of a liquid–solid interface thereby reducing the symmetry of a seed or by self-assembly of zero-dimensional nanostructures. The latter phenomenon forms the basis for vapor phase nucleation techniques. The growth of nanowires from the vapor phase techniques requires that the starting reactants be a gas phase species. Though there are many subcategories of vapor phase techniques, the preferred mechanism for nanowire growth referred to as the vapor liquid solid (VLS) technique will be discussed in the next section.

Vapor Liquid Solid (VLS) Mechanism Wagner and Ellis first discovered the VLS method in 1964 [14]. In the VLS method, the nanowire is grown in a desired orientation when a metal catalyst reacts with the "nanowire material" in its gaseous phase at high temperatures. This method is based on the eutectic solidification of a metal and semiconductor. The catalyst also known as a second-phase material is used to confine the crystal growth onto a specific orientation. The catalyst alloys itself with the growth material. This acts as a trap to the growth species, which results in the precipitation at the growth surface in a single direction. Figure 2 depicts a typical VLS growth.



Growth of Silica Nanowires, Fig. 2 Schematic of the VLS growth mechanism

The VLS growth is facilitated by (a) the mass transport of whisker growth species in the vapor phase to the vapor liquid phase, (b) the chemical reaction on the vapor-liquid interface, (c) the dissolution and diffusion through the liquid alloy phase, and (d) the precipitation on the solid–liquid interface.

Common conditions for a VLS mediated growth are as follows:

(a) The catalyst must form a liquid solution with the crystalline material to be grown at the deposition temperature.
(b) The distribution coefficient of the catalyst must be less than unity at the deposition temperature.
(c) The equilibrium vapor pressure of the catalyst over the liquid droplet must be very small.
(d) The catalyst or impurity must be inert chemically.
(e) The interfacial energy, i.e., the wetting characteristics, plays a very important role. For a given volume of liquid droplet, a small wetting angle results in a large growth area, leading to a large diameter of nanowires.
(f) For a compound nanowire growth, one of the constituents can serve as the catalyst.
(g) For controlled unidirectional growth, the solid–liquid interface must be well defined crystallographically.

In VLS growth, the growth species in its vapor form passes through the substrate in the presence of a catalyst (metal particle), which then diffuses into a liquid droplet. The liquid's surface has a large accommodation coefficient, so it is a preferred site for deposition. Saturated growth species in the droplet diffuse and precipitate at the interface between the substrate and liquid thus growing into a nanowire. The VLS growth method is not restricted by the type of substrate materials and catalysts, thus enabling the nanowires to be either single crystalline, polycrystalline, or amorphous. In general, the catalyst plays a key role in the VLS process.

It forms a liquid droplet as a preferred site to adsorb the vapor pieces of the source materials. The desired one-dimensional (1D) nanostructure precipitates, nucleates, and grows from the saturated liquid droplet. The selected catalysts in the VLS process should meet two demands. First, the catalyst should form the liquid eutectic alloy with a 1D nanostructure material under the growth temperature. The liquid alloy phase reduces the activation energy at the liquid–solid interface and enhances the growth rate appreciably. Second, the catalyst should be more stable than the components of the desired 1D nanostructure. This sustains the continuous growth as long as the vapor species is available.

VLS Kinetics Typically a catalytic agent is coated onto a growth substrate and is placed in a vacuum furnace of high temperature. Vapor atoms of the nanowire material are passed through it. The lowest temperature and pressure are chosen such that they cause the catalyst to form alloyed eutectic phase seeds with the lowest concentration of the growth species (GS). This causes the tip of the seed to be semi-molten. The semi-molten structure of the tip has a soft lattice and loose interatomic configurations to allow adsorption of the GS vapor.

If there is a lateral temperature gradient, or a sudden change of temperature, it leads to random breakup of the catalyst droplets. Also if there is an oversupply of the GS, it leads to oscillation of surface energies, which enables multiple nucleations. Lastly, the gas of the GS can cause hydrodynamic stress on the droplet thus causing droplet breakup. The temperature and pressure are very crucial for the VLS mechanism to succeed. The growth of the nanowires will proceed only when the concentration of the GS is above the equilibrium solubility. Although the evaporation of the catalyst does not change the composition of the saturated liquid composition, it reduces the total volume of the liquid droplet. Unless more catalyst is supplied, the volume of the liquid droplet reduces and growth will stop when all the catalyst is evaporated.

The orientation of the nanowires depends on the substrate texture and the diameter of wire. Moreover, it also depends on the nanowire composition, as materials prefer to orient in a direction minimizing their enthalpy of formation.

Silica Nanowires: A Review of Growth Investigations

Over years, silica nanowires have been synthesized using growth methodologies discussed in the previous sections. Among the various techniques, vapor phase approaches have been routinely used in applications that require mass manufacturing and high-density growth. A comprehensive review [15] on silica nanowire growth (for the past 12 years) is presented in Fig. 3.

Selective Growth of Silica Nanowires: New Approaches

Ion Implantation–Based Seeding Approach

Of the several methods used to produce nanowires, the VLS mechanism proposed by Wagner and Ellis for silicon whisker growth has proved to be very reliable for growing silica nanowires. A simulation study [16] indicated the feasibility of forming catalyst nanoclusters in silicon using ion implantation. Ion implantation is a low-temperature technique for the introduction of impurities (dopants) into semiconductors, and is preferred over diffusion. It is a process of introducing energetic and charged particles into a substrate. Implantation of ions into a target may alter the physical or chemical properties of the host via damage induced by the energy loss of the incident ions. In this surface modification technique, ions are extracted and accelerated toward the target. Before penetrating the host, the charged ions are separated according their mass-to-charge ratio. Then the desired species is directed at the target. The foreign atoms enter the crystal lattice, collide with the host atoms, lose energy, and finally come to rest at some depth within the substrate. The average penetration depth is determined by the species, substrate material, and the initial energy of the ions. This depth has some distribution, as the collisions with the target atoms are random. Figure 4 shows the ion implantation setup.

In this section, ion implantation as a new method to form catalyzing seed nanoclusters at or below the surface of Si and subsequent growth of silica nanowires is presented [17]. The ion implantation–based seeding approach possesses several advantages over conventional methods of catalyst placement such as (a) the size and distribution of the catalyst can be accurately controlled by varying the ion dose, energy, and dose rate, (b) the placement of catalyst being an integral part of the substrate lattice lends for an impurity-free nanowire growth, and (c) majority of the element across the periodic table can be implanted giving rise to hybrid nanostructures (metal/dielectric). Such an approach was demonstrated by implanting Pd ions into Si through a mask.

Methodology In this investigation, prime grade Si (100) wafers of 2″ diameter were used as substrates for Pd ion implantation performed with a metal vapor vacuum arc (MEVVA) ion implanter. The MEVVA ion source is a high current device, which can readily produce ion beam currents of several hundred milliamperes on target. Extraction potential up to 50 kV is used to implant metallic elements. Ion beams of most metallic elements either individually or in combination can be generated with a MEVVA source.

In this study, the lowest possible extraction voltage to produce a stable ion beam in the implanter (of 10 kV) was chosen to ensure maximum surface concentration of implanted Pd. Since the MEVVA ion source produces ions with multiple charge states, the energy of the ions emanating from the source depends on the charge state. For the Pd ion implantation, the energy was estimated to be 19 keV using a mean charge state of 1.9. Circular shadow mask (Al) was used to define identical and well-separated zones of ion-implanted regions. Dose was varied from 5×10^{12} to 3×10^{16} Pd ions/cm^2. The sample was introduced into the furnace around 1,100 °C and heated for the desired time (7–60 mins) in Ar followed by cooling to room temperature. Further, the growth temperature was varied.

Discussion After heating at 1,100 °C for 60 min, the ion-implanted region showed a dense growth of rather long and interwoven nanowires as shown in Fig. 5. The inset shows the unimplanted area at the same magnification confirming the absence of nanowires due to unavailability of Pd seeds.

Substrate	Type of Catalyst and Preparation	Growth Conditions			Geometry		Mechanism	Reference et al.
		Temperature (C)	Time (min)	Ambient	Dia. (nm)	Height and Orientation		
Glass, Si,Al_2O_3	Spu-Au, 15–90nm	CVD 350–1000	30	Ar	72	Spring Morphology	VLS	Wang L., Nanotech. 2006
Si (111)	Ga, Ni Power $Ni(NO_3)_2.H_2O$	950	30	Ar/NH_3	30	20 nm–20 μm, Jelly fish, core-shell architecture	VLS	Cai, X.M., JAP 2005
Si	Ga Molten Droplets	700–950	10–60	NH_3	50	> 30 μm, Self assembled cone	VLS	Luo, S., Small 2006
Si	NiO Powder	1000–1100	50–100	NH_3	50–100	Ultra long mm	SLS	Park, B., Surf.Rev. Lett 2003
Fused Silica Capillary	2 Chloro-1,1,2 TriFluoroEthylMethyl Ether	400	720	-	10–100	5–200 nm	VLS/VS	Woldegiorgis A., Electrophoresis 2004
Silica	Si, C and Co	1500	60	Ar	20–120	10–250 μm, Flower Like Nanofibers	Solid Phase/ Gas solid Reaction	Zhu, Y.Q., J.Mat.Chem 1998
Silica	Si/Si Powder	1400,Pyrolysis	12	N_2	80	[110], Core-Shell	VS	Z.L. Wang, Adv Mat 2000
AAO	$SiCl_4$, Wash CCl_4	Hydrolysis			P.Dia	P.Length	Surface Chem, Solgel	Kovtyukhova, Adv Mat 2003
Fused Silica/Si (111)	-	1800, HFCVD	20	$SiCl_4+H_2$	80	3.5 μm, Catalyst Free, Template Free	-	Jih-Jen Wu, Av Mat 2002
Si	Molten Ga	520–900	5–300	Ar/O_2	15	10 μm	Extended VLS	L.Dai. Phy.Lett A 2005
Si (111)	Sn	980	2	Ar	100–200	50–100	VLS	S.H.Sun, Solid. State. Commun. 2003
Si	MgO+Graphite Powder and Au	1100	45	Ar	60–90	Long	VLS	S.Kar, Solid. State. Commun. 2005
Si	Ga/In_2O_3	400	120	Ar/NH_3	60	15	VLS	J.C. Wang., Solid. State. Commun. 2003
Si	Si/Al_2O_3	1150	8–60	Ar	20–100	mm	VLS	X.S.Fang, Appl.Phys A 2005
SiO_2	Au	1100	60	Ar	25–100	mm	SVLS/OAG	Houlier Jnl of Phase Equ and Diffusion 2007
Si (100)	Co NPs	1100	120	Ar/CH_4	5–60	-	Coils	F.Wang, JPC, 2007
Crucible	SiO (P) + Fe	850	240	Ar	15	mm	VLS	D.P.Yu, APL, 1998
Si	GaN	1150	300	Ar	15–300	10–40 μm	Extended VLS	Z.L.Wang, JACS 2001
Pellets	SiO_2, $FeNO_3.9H_2O$, Active C Ball Mill	1350	180	Ar	60–110	>100 μm	Carbothermal Redn +VLS	X.C. Wu, Chem.Phy. Lett 2001

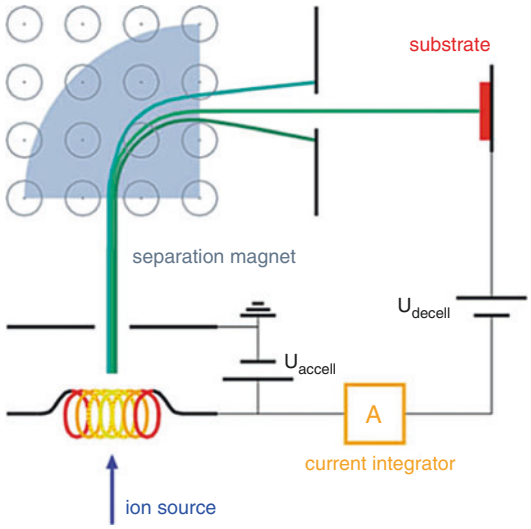
Growth of Silica Nanowires, Fig. 3 (continued)

Substrate	Type of Catalyst and Preparation	Growth Conditions			Geometry		Mechanism	Reference et al.
		Temperature (C)	Time (min)	Ambient	Dia. (nm)	Height and Orientation		
Si		800–1300	30	H_2/H_2O	50	Long NWs Aligned	Thermal Growth	J.Q.Hu, Chem.Phy. Lett 2003
AAO	TEOS + Ethanol	1 Day Aging, 200 C-1 Day	-	-	22–30	6 μm		Zhang, J.Mat Sci.Lett 1999
Si	Pd/Au, 1:1 Spu	1100	40	N_2	100		VLS	Elechiguerra, Applyphys A 2001
Si (111)	Ni–40 nm	950		Ar/H_2	20	10 mm	SLS	Yan, Chem Phys Lett 2000
TiN/Ni/SiO_2/ Si		1050	20	H_2/Ar		-	Solid state reaction	K-H Lee, Chem Phy Lett 2004
Si + PMMA Mask	Ni	1100	20	Ar/H_2	30	-	SLS	Ju Hung Hsu, Nanotech nology 2006
Si	NiO, WO_3	1100	60	Ar	20–100	-	SLS + Carbothermal	B.T.Park, Nanotech, 2006
Si	Sn + SiO	1320	420	Ar/H_2	230	mm	VLS+OAG	H.Wang Angew. Chem.Int 2005
Si	Ga	1050	60	Ar/H_2	10–20	0.5–1.5 μm	VLS	Yiqin Chen, Nanotech 2006
Si	CuO/C	1000	60	Ar	5–500	-	VLS + Carbothermal	Yu-Chiao Lin, IEEE Nano 2005
Si	Thermal Evap, Laser Ablation, SiO Powder	800–1300	30	Ar	60–120	10 μm	[111]	J-Q.Hu, Adv.Mat 2003
Si	Ni + In_2O_3	650	45	N_2/NH_3	10–20	5–20 μm	VLS	Renzlin Ma, Chem.Phy. Lett 2003
Si (111)	Ni Sputter	1000	60	Ar/SiH_4	9	10 μm	VLS	Niu, Physica E 2004
Crucible	SiO + SnO	1180	240	Ar	2500	100	VLS	J.Hu, Small 2006
Si	Zinc Carbonate, Hydroxide and Graphite	1100	60		100	Several μm	VLS	Q.Wei, Solid.State Communc 2006
Graphite	Si Powder	1200	30	Ar/O_2	300	500	VLS	K.S Wenger, J.Mat.Chem 2003
Colloidal Particles	TEOS, Ethanol, HCl, H_2O	80	30		1000		Solution Grown	S-S Choi, J.Mat.Sci Lett 2003

Growth of Silica Nanowires, Fig. 3 (continued)

Substrate	Type of Catalyst and Preparation	Growth Conditions			Geometry		Mechanism	Reference et al.
		Temperature (C)	Time (min)	Ambient	Dia. (nm)	Height and Orientation		
Ceramic Boat	SiO Powder + H_2O	1200	60	Ar	100	mm	VLS	Y.Hao, APL, 2005
Alumina	Si Powder, SiO_2, Pbo, C	1200	90	Ar/O_2	710	Various Morphology	VLS	J.Zhang, Jnl.Cryst. Growth, 2001
Si	Au	1000	180	N_2/H_2	30–150	10000	SLS	Paulose M., Jnl.Nanosci and Nanotech.
Si	Fe	1200			8–20	> 1 μm, [111]	Core-shell	Lieber, Science 1998

Growth of Silica Nanowires, Fig. 3 Literature review of silica nanowire growth

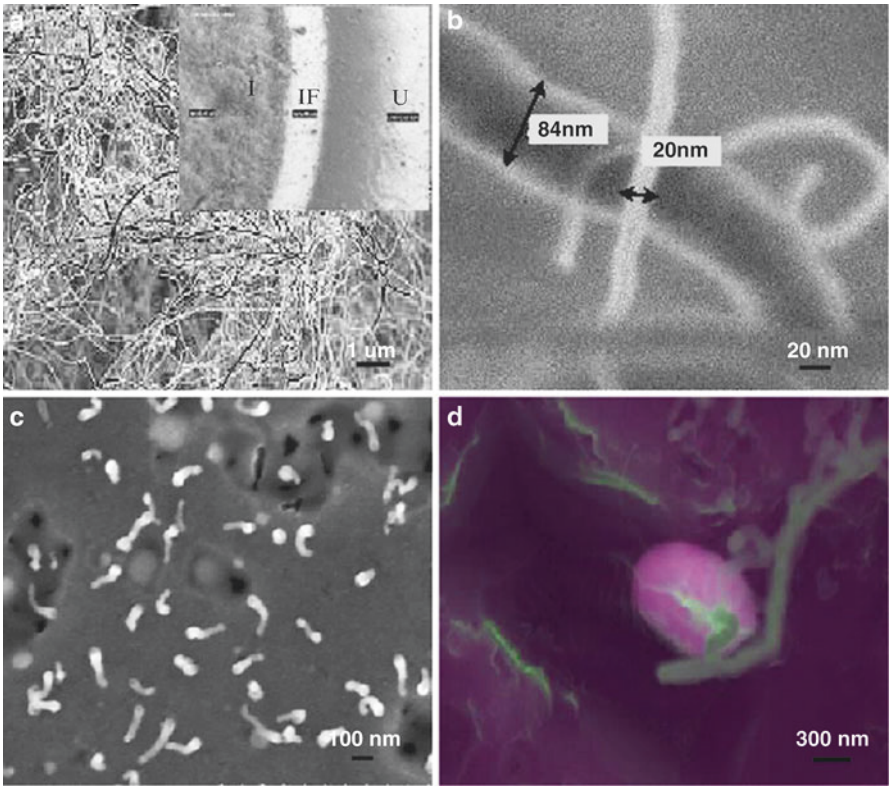


Growth of Silica Nanowires, Fig. 4 Ion implantation setup (Courtesy: Wikipedia)

Also, Fig. 5a demonstrates the selective nucleation of nanowires on the implanted region. The interface region is very narrow and sharp indicating the ability to directly locate nanowires on the implanted areas. The nanowires grow with diameters ranging from about 15 to 90 nm, as for example seen in Fig. 5b. The length of the nanowires was observed to vary up to about 50 μm corresponding to an aspect ratio of up to 3,000. The morphology of the nanowires depends on the growth time as evident from Fig. 5c (after 7 min) captured at the early stage of the growth with nanowires emanating from their parent seeds. Figure 5d is a composite constructed from an overlay of two SEM images recorded in backscattered mode (pink color) and secondary mode (green color). The pink areas therefore show atomic number contrast. The pink oval seed is confirmed to be rich in Pd by spot energy dispersive spectroscopy (EDS) measurements on and away from it.

Figure 5d therefore captures the "birth" of a silica nanowire (green color) growing from a Pd-rich seed. A much thinner nanowire can also be seen to emerge from the same seed. The composition of the nanowires was measured with EDS (at low electron energy of 5 kV to ensure that majority of X-rays come from nanowires and not from Si substrate) and compared with that of a control sample of quartz analyzed under identical EDS conditions (Fig. 6a). Thus the composition was found to be close to SiO_2 with a slight deficiency of oxygen.

The phase identification was done by X-ray diffraction (XRD) on the same sample (Fig. 6b), which did not show the peak observed in quartz sample indicating that the nanowires are amorphous. The temperature was also found to be a critical factor for the growth of nanowires. At a lower temperature of 1,010 °C for the same growth time of 60 mins, only a very early stage of nucleation and growth was observed. At still lower temperature of 910 °C (Fig. 7a), no growth of nanowires is observed. Pd ion dose was also found to be another factor critical for nucleation of Pd-rich seeds and nanowires.

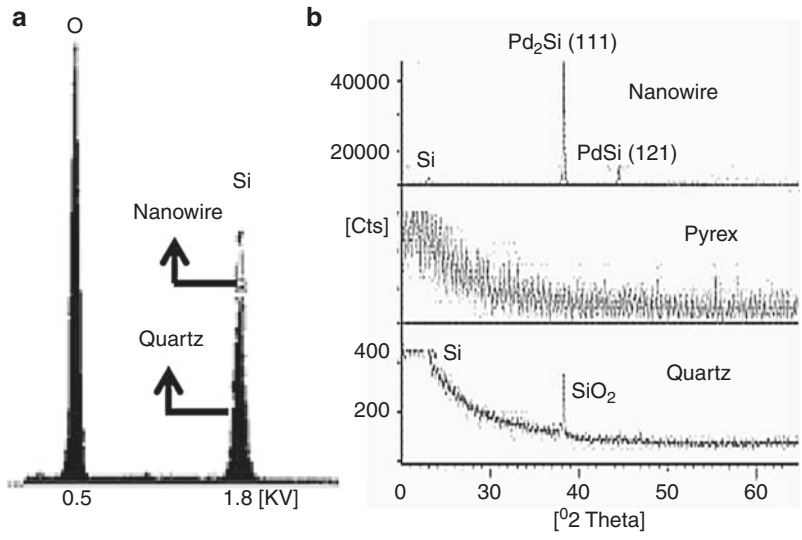


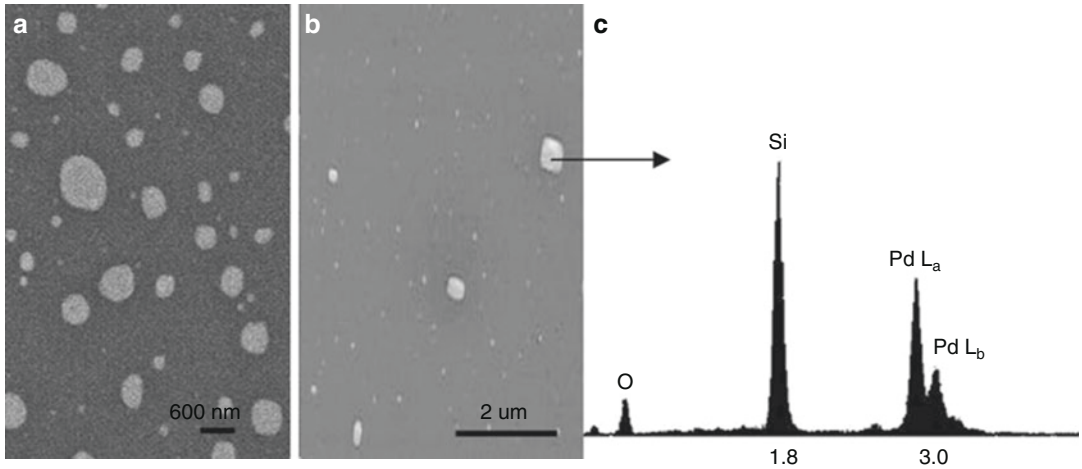
Growth of Silica Nanowires, Fig. 5 Selective growth of nanowires at a constant Pd ion dose of 3×10^{16} ions/cm^2 after heating at 1,100 °C for 60 and 7 min. (**a**) Dense wire growth after 60 min; *inset*: Selective growth on implanted regions, inset: *I* Implanted region showing nanowire growth, *IF* Interface, *U* Unimplanted region devoid of nanowire growth, (**b**) variation in morphology of wires in (**a**), (**c**) early stage of nanowire growth (after 7 min), and (**d**) composite of backscattered and scanning electron image highlights the elemental contrast between seed and wire

As the dose was reduced from 3E16/cm^2 (Fig. 8a) to 1E16/cm^2 (Fig. 8b), coiled nanowires of reduced density were seen. The density of the wires was estimated by counting the wires in five 4×4 μm (high dose), 7×7 μm regions (low dose) of the SEM and averaging them. The observation show an estimated 2E8 wires/cm^2 in sample implanted with a dose of 3E16/cm^2 (Fig. 8a) and 1E8 wires/cm^2 in sample implanted with at a dose of 1E16/cm^2 (Fig. 8b). These observations confirm that implantation dose is a critical factor in determining the density of the nanowires. However, at a dose of 5E12 ions/cm^2, no growth of nanowires (inset in Fig. 8b) or Pd-rich seeds was observed.

Moreover, the flow of Ar gas was found to be critical in producing growth of silica nanowires. In another experiment at 1,100 °C, the Ar gas was shut off after the furnace temperature reached 1,100 °C, and then the sample was introduced into the furnace and kept there for 60 min, after which the furnace was allowed to cool down to room temperature and the sample was taken out for further study by SEM and XRD. The SEM results showed islands of bright appearance.

Larger islands with rectangular morphology (Fig. 7b) were confirmed to be consisting of Pd_2Si composition by performing a spot EDS (Fig. 7c) on them. The composition of smaller islands could not be determined accurately because their size was much smaller compared to the excitation volume for the EDS. They appear to be the Si-rich PdSi phase. Away from islands, no Pd was found. There were no nanowires on this



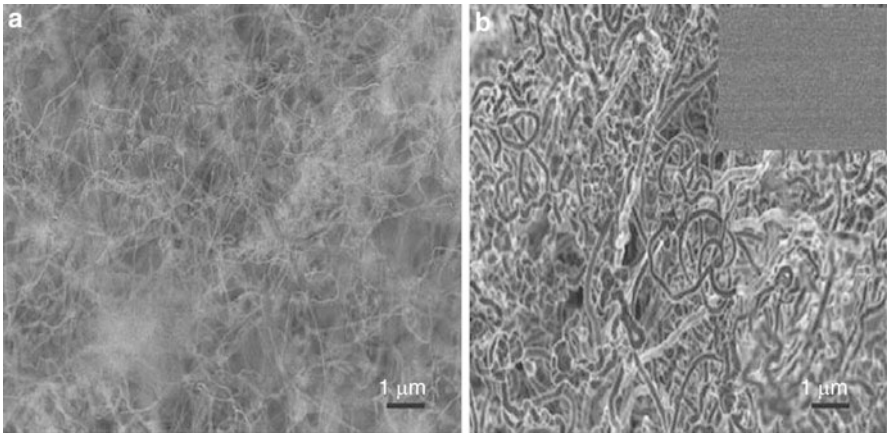


Growth of Silica Nanowires, Fig. 6 Elemental composition and texture analysis of the synthesized nanowires. (**a**) Comparison of EDS spectra on quartz with EDS on nanowire sample revealing SiOx ($x < 2$) stoichiometry and (**b**) XRD showing their amorphous nature

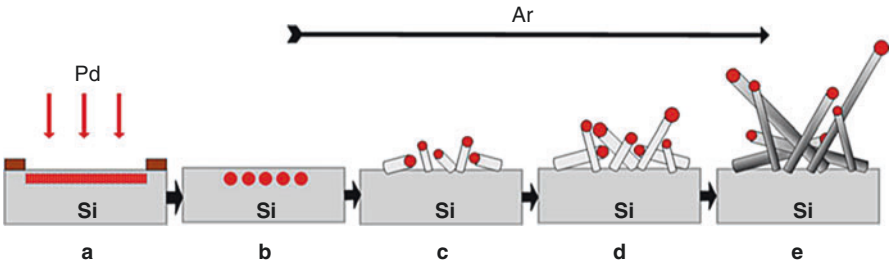
Growth of Silica Nanowires, Fig. 7 Conditions of no nanowire growth at a constant Pd ion dose of $3 \times 10^{16}/\text{cm}^2$. (**a**) 910 °C, Ar flow – absence of nucleation, (**b**) 1,100 °C, no Ar flow – formation of large whitish islands, and (**c**) spot EDS confirming the presence of palladium silicides on the large islands

sample. XRD obtained from this sample shows the presence of Pd_2Si and PdSi phases. The absence of silica nanowires proves the role of Ar as a carrier gas in transporting Si vapor to the reaction site resulting in the whisker growth by a VLS mechanism.

The physical model to explain the nanowire growth is (Fig. 9) as follows: the ion dose needs to be high enough to produce nanoclusters of Pd within the silicon matrix amorphised by ion damage. As the implanted sample is heated in Ar atmosphere, the interdiffusion at the Pd-Si interface produces two phases of PdSi and Pd_2Si, which eventually become liquid as the melting point of the silicides is approached at 972 °C or 1,331 °C, respectively. Since the maximum temperature used in the study is 1,100 °C, only liquid drops of PdSi phase are formed. These liquid drops may now coalesce to form larger sized liquid drops and migrate toward the surface of the solid Si substrate. The amorphised surface layer of silicon would also crystallize



Growth of Silica Nanowires, Fig. 8 Effect of implanted dose on the growth of nanowires heated at 1,100 °C for 60 min. (**a**) 3×10^{16} ions/cm^2 – high-density thinner nanowires and (**b**) 1×10^{16} ions/cm^2 – thicker coiled nanowires of reduced density, *inset*: 5×10^{12} ions/cm^2 – absence of nanowires



Growth of Silica Nanowires, Fig. 9 Schematic of the proposed physical model explaining ion implantation–based seeding approach. (**a**) Implanted Pd on Si (100) through a mask, (**b**) formation of PdSi cluster, and (**c**–**e**) VLS growth of silica nanowires

concurrently. When these liquid drops are exposed to the Si vapor being swept across by the carrier gas (Ar), the VLS model of whisker growth comes to play and Si nanowires begin to grow. Some nanoclusters, especially those on which nanowire growth has not yet started, agglomerate by diffusion/Oswald ripening processes, become bigger in size and nucleate wires. The growth process proceeds continuously by diffusion and precipitation of silicon atoms across the liquid-nanowire interface. Since small amount of oxygen is inevitably present in the ambient as impurity, these thin silicon nanowires may oxidize rapidly from outside and eventually get fully oxidized.

The observed oxygen deficit from the stoichiometric SiO_2 may be consistent with some unoxidized silicon around the center of the nanowires. The occurrence of selective VLS growth of nanowires only on the Pd ion–implanted regions has been validated by a careful selection of ion dose, growth temperatures, and Ar gas flow. The VLS growth can stop when either of the following conditions are met – no more vapors of Si arrive or the temperature drops below the melting point of the material of the liquid drop, or the implanted ion dose is below that required for Pd cluster formation. However, the origin of the observed morphology of silica nanowires is unclear. The nanowires are produced with a variety of diameters and have much larger lengths forming a network of entangled structures. Since the physical model is based on VLS growth of Si nanowires followed by their oxidation, one can expect the dominant growth direction to be (110) on a (100) substrate

surface so that the nanowires would point out at ±45° with respect to surface plane. However, the SEM views of the early stage of growth do not show such large angles. They appear to be coming out almost along the surface at much smaller angles. This observation appears to be in agreement with recent reports of diameter-dependent growth directions of epitaxial Si nanowires grown on Si (100) substrates using Au as liquid catalyst facilitating VLS growth [18]. It was found that Si nanowires grown epitaxially on Si (100) via the VLS growth mechanism change their growth direction from <111> to <110> at a crossover diameter of approximately 20 nm.

A model has been proposed [19] for the explanation of this phenomenon. Based on this model, it can be hypothesized that most of the nanowires with diameters greater than 20 nm grow along <211> or <111> or a mixture of these growth directions and therefore lie close to the substrate plane (within about ±18°). As the diameter of the nanowires changes, so does the growth direction leading to the observed entanglement of the nanowires.

A similar change with decreasing diameter of the nanowires from <111> to <110> direction has also been observed for non-epitaxial Si nanowires. In summary, a new method for selective growth of silica nanowires on silicon wafers using ion implantation–based seeding approach is presented. Since ion implantation is compatible with the standard top-down CMOS process, this method holds a great potential for facilitating directed bottom-up growth of nanowires on silicon wafers. It may also provide an opportunity to combine the top-down and bottom-up technologies for creation of robust nanoscale devices. It may therefore open doorway to a large number of applications in wide areas of new technologies such as on-chip optoelectronics, biosensors, micro-antennae, and metallic nanotubes. The only limitation of this technique is scarce availability of ion implanters.

Metal Thin Film–Based Catalysis

A metal thin film catalysis–based silica nanowire growth circumvents the need for ion implantation as a tool for the placement of the catalyst. Also, the ability to produce stable nanowires at desired locations in a controlled manner on planar substrates is of critical importance for fabricating nanodevices such as nanoFETs, nanotransistors, and microphotonic nanosystems.

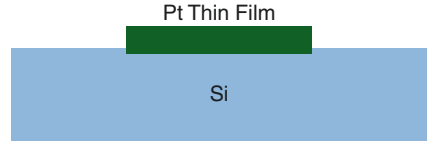
In this section, selective growth of silica nanowires in silicon using Pt thin film is presented [20]. This synthesis approach for localization offers several advantages such as the tunable size and distribution of the parent nanoclusters and a cost-effective solution for growing large-scale arrayable nanowires.

Methodology Prime grade 2″ n-type silicon wafers were used as substrates for subsequent Pt deposition.

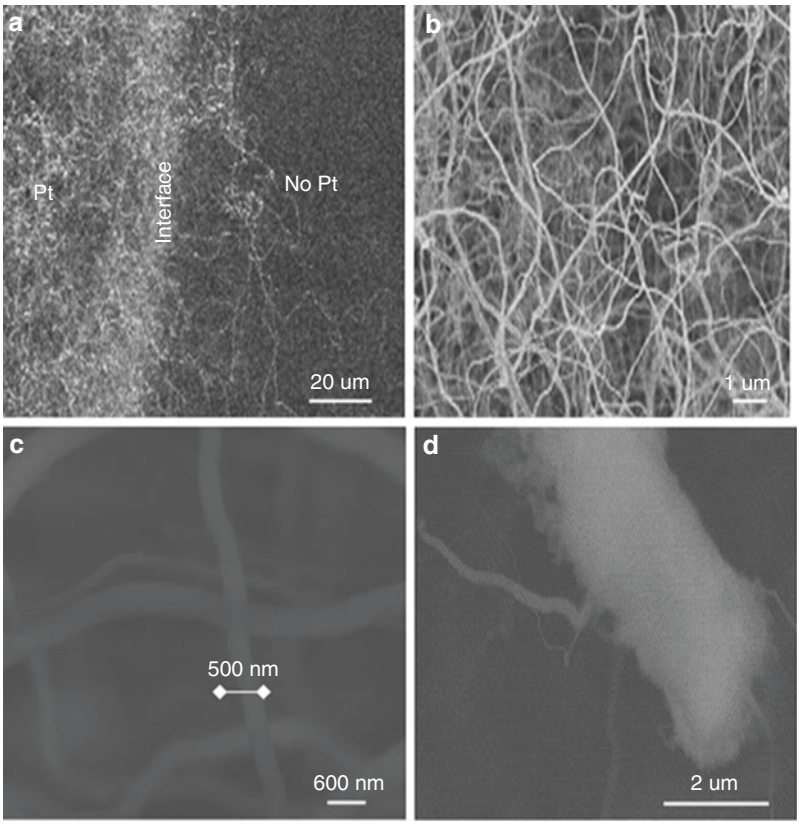
Target Pt thickness of 2, 3, 5, 10, and 100 nm was sputtered onto Si substrate (Fig. 10).

The temperature of heating is chosen based on the Pt-Si phase diagram. High-purity Ar was chosen as carrier gas and was set to 30 sccm throughout the course of the experiment. Initially the furnace was flushed with Ar to minimize interference from gaseous impurities. Then the sample is introduced into the furnace and ramped up to 1,200 °C. Once the furnace reached 1,200 °C, the heating process was timed.

Discussions After heating at 1,200 °C for an hour, regions deposited with Pt (20 nm) showed dense growth of interwoven nanowires. Figure 11a demonstrates selective growth of nanowires identifying the nucleation area and regions devoid of nanowire due to absence of Pt film. A higher magnification SEM micrograph taken in the growth region (Fig. 11b) illustrates nanowires of varying morphology. The nanowires grew at an average diameter of 200–500 nm (Fig. 11c) with lengths varying from 20 to 75 μm giving rise to high aspect ratio nanostructures for biological



Growth of Silica Nanowires, Fig. 10 Platinum thin film deposited on a silicon substrate

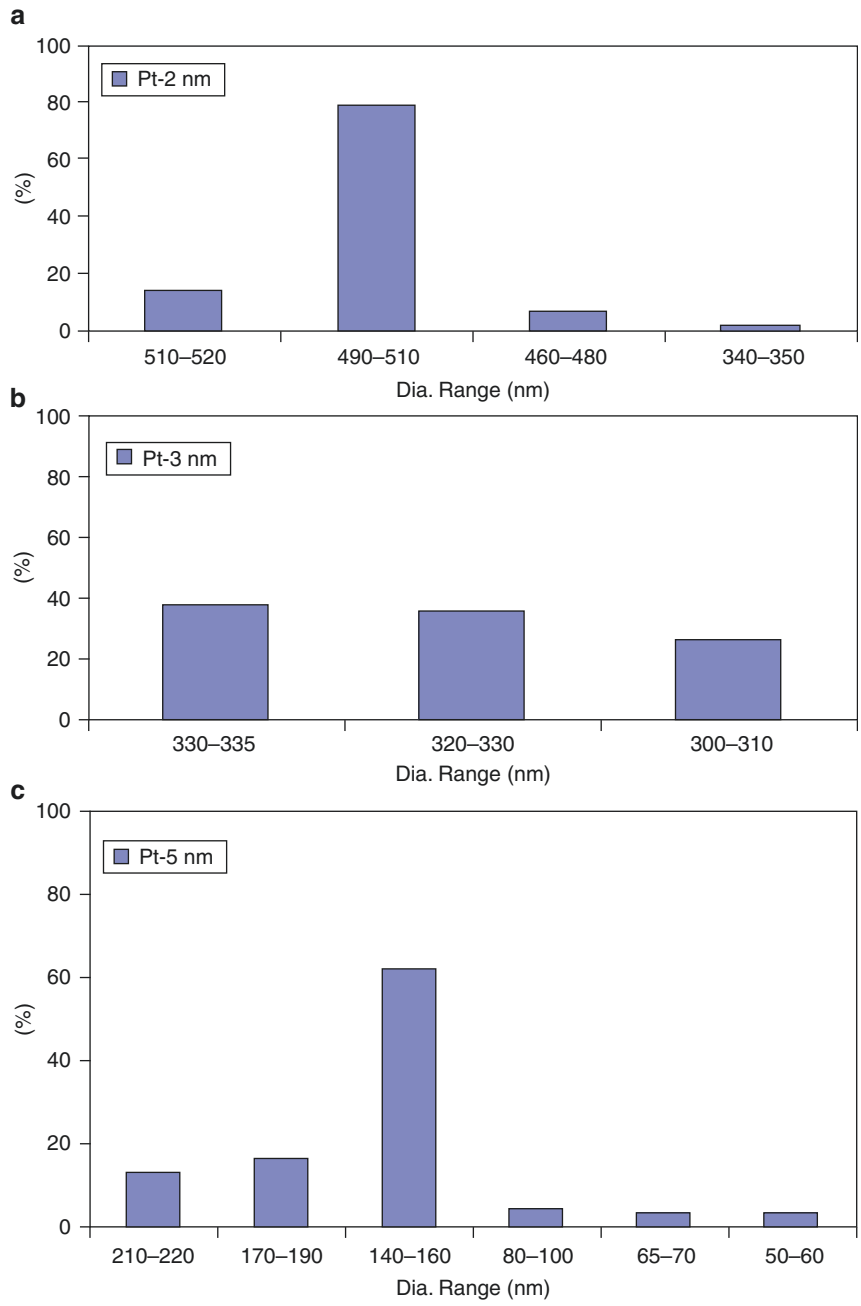


Growth of Silica Nanowires,

Fig. 11 Selective growth of nanowires after heating at 1,200 °C for 60 min for a Pt thickness of 2 nm. (**a**) Localized growth of nanowires, (**b**) higher magnification image in area of nanowire growth indicating the varied morphology of nanowires, (**c**) size of nanowires – diameter ranging from 50 to 500 nm, and (**d**) early stage of nanowire growth captured after heating at 1,200 °C for 15 min

applications. Figure 11d is a schematic capture of birth of several nanowires from their parent seed. The composition of these nucleated nanostructures from low-energy EDS was found to be SiO_2 closely matching with the EDS spectra of a control sample of quartz analyzed under similar test conditions. Phase identification through XRD revealed the amorphous (as compared to a quartz sample) nature of the wires along with the formation of PtSi, which acts as a seed catalyzing the nanowire growth. The thickness of platinum layer was found to be a critical factor affecting nanowire growth. As the thickness was increased to 3 nm (keeping other experimental conditions constant), the number density of the nanowires reduced. Further reduction in number density was observed when scaled up to 5 nm. At a thickness of 10 nm, silicide nanoclusters were formed indicating the absence of nanowire nucleation. When the deposited platinum thickness was around 100 nm, micron-sized clusters were seen.

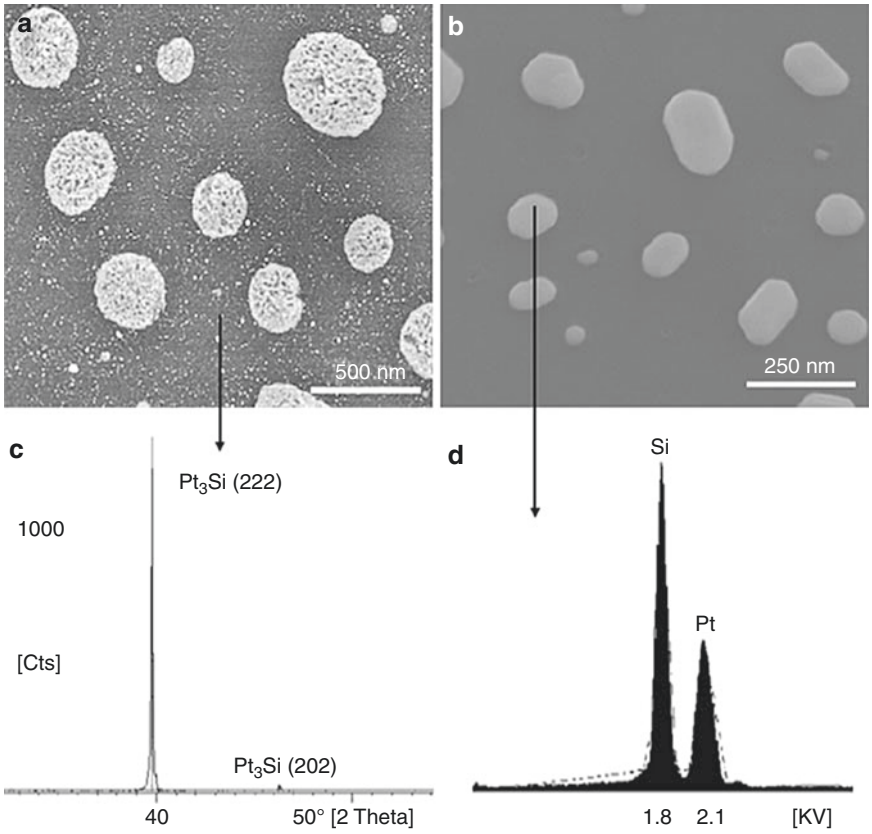
Figure 12 illustrates the statistical nature of the growth process through a frequency histogram. All the nanowire samples with different Pt thicknesses were heated to 1,200 °C for 1 h. For a Pt thickness of 2 nm, the ensemble diameter average was found to be 500 nm with a standard deviation (σ) of 6.7 over the 79 % range. A uniform distribution with an average wire diameter of 319 nm ($\sigma = 9.1$) was observed for the 3 nm Pt sample. As the platinum thickness increased to 5 nm, the nucleated nanowires exhibited a wide range of (60–220 nm) diameter variation explained by the Oswald ripening process. The average diameter was found to be 146 nm ($\sigma = 7.4$) over the 62 % range. Exploiting this selectivity of nanowire growth to platinum film thickness, a Si substrate with differential Pt coating (differential thickness achieved through repeated application of shadow mask) resulted in the nucleation of nanowires while simultaneously forming optically conducting silicide in a single furnace treatment. This would facilitate a low-loss



Growth of Silica Nanowires, Fig. 12 Percentage histogram indicating the statistical nature of the growth process with varying diameters for different Pt thicknesses. (**a**) Pt – 2 nm: ensemble average diameter = 500 nm, $\sigma = 6.7$; (**b**) Pt – 3 nm: ensemble average diameter = 319 nm, $\sigma = 9.1$; and (**c**) Pt – 5 nm: ensemble average diameter = 146 nm, $\sigma = 7.4$

hybrid nanointerconnect scheme in silicon for future integrated optical systems. The heating temperature was also found to be a critical factor influencing nanowire growth.

At a lower growth temperature of 950 °C for 1 h (Fig. 13a), spherical islands appeared surrounded by nanoclusters. These islands were identified as Pt_3Si particles through XRD.



Growth of Silica Nanowires, Fig. 13 Conditions of no nanowire growth for samples deposited with 2 nm Pt thickness. (**a**) 950 °C, 1 h – XRD indicating the formation of Pt_3Si (**c**) and (**b**) 1,200 °C, 1 h, No Ambient – EDX spectra revealing PtSi atomic stoichiometry of seeds (**d**)

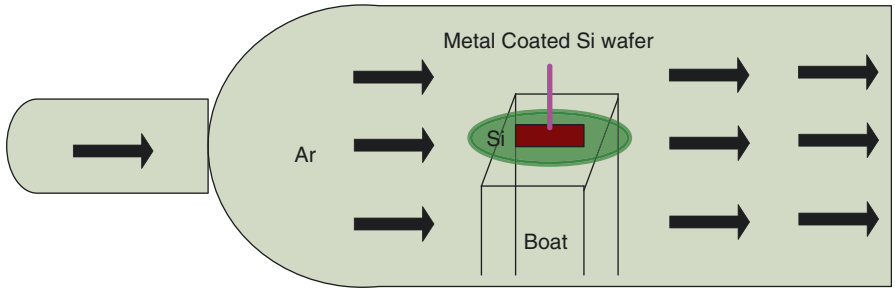
Testing the role of ambient in nanowire growth, the flow of Ar was intentionally shut off while ramping up the furnace to 1,200 °C. When the sample was analyzed in a SEM, rectangular islands of bright appearance (Fig. 13b) were observed. Larger islands were confirmed to be PdSi by composition through an EDS on them. Smaller islands could not be resolved elementally due to limitation on the excitation volume for the EDS. The absence of silica nanowires proves the role of Ar as a carrier gas in transporting Si vapor to the reaction site resulting in the whisker growth by a VLS mechanism.

In summary, selective growth of silica nanowires on silicon catalyzed by Pt thin film was demonstrated. Platinum silicide acts as the seed facilitating the bottom-up nucleation of the oxide nanowires. The nanoscale optical wires nucleate following a VLS model validated through a careful set of growth conditions. This approach represents a simple, flexible, and cost-effective solution for synthesizing large-scale silica nanowires for wide variety of future applications that include functional microphotonics, high-density optical nanointerconnects, high-resolution optical heads, and nanoantennae. In the next section, the manufacturing aspects of silica nanowires will be discussed.

Manufacturing Aspects of Silica Nanowire Growth

Overview

Exploring the manufacturing aspects of nanostructures can enable the transition from a



Growth of Silica Nanowires, Fig. 14 Schematic of the furnace and experimental setup

laboratory-based research into a commercial product. Scalable manufacturing methods are required to harness the benefits of nanotechnology for a wide variety of applications. Understanding the manufacturing aspects of nanomaterials requires growth model and growth rate estimation. This section details on (a) a growth model derived from studying silica nanowire nucleation as a function of heating time, (b) nanowire growth rate estimation via weight differential of the Si substrate before and after growth, and (c) reusability of the Si substrate for nanowire growth. The growth mechanism of silica nanowires on Si via Pd metal thin film catalysis was investigated. A growth model was derived by sequentially varying the nanowire growth time at a constant temperature. The derived growth model for mass manufacturing nanowires was verified by a combination of experiments such as (a) annealing the metal thin film–coated silicon wafer in the absence of gas flow, (b) removing the nanowires in HF to inspect the substrate after growth, and (c) vary the Ar flow rate. In addition, the growth rate was estimated by weight differential technique while identifying distinct nanowire growth regimes. Finally, the reusability of the silicon substrate for further nanowire growth was evaluated.

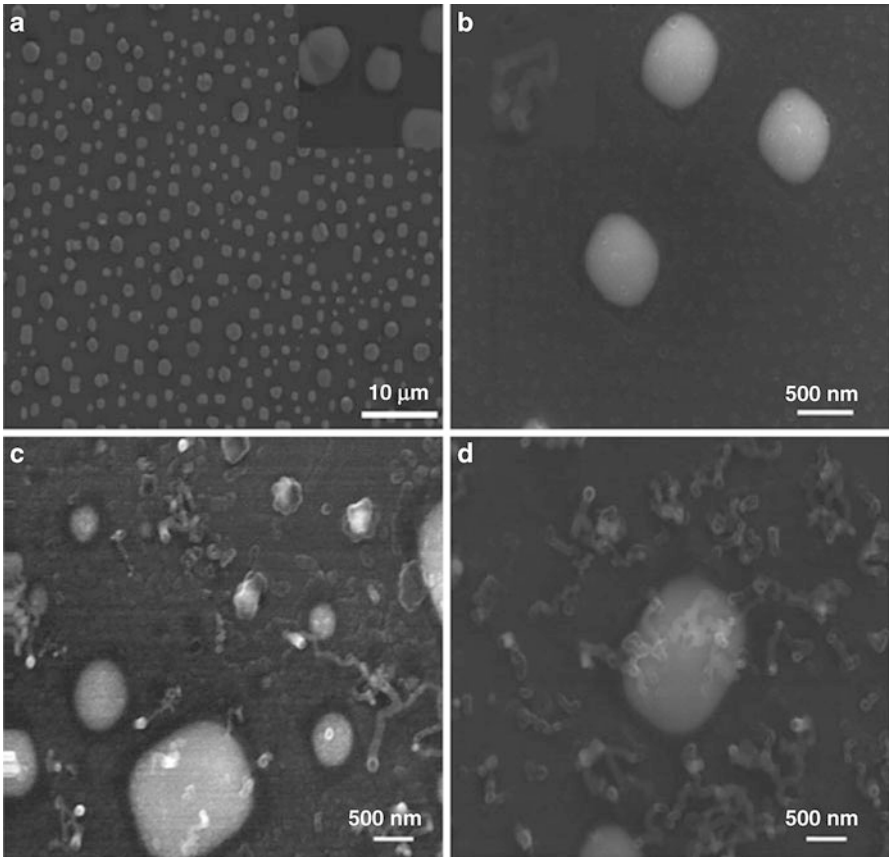
Methodology

Prime grade Si wafer pieces (2 × 2 cm) were used as substrates for metal catalyst deposition. Pd of thickness of 5 nm was sputtered onto Si substrate, which served as the catalyst. The growth temperature of nanowires was chosen to be 1,100 °C based on Pd–Si phase diagram. High-purity Ar was chosen as carrier gas in an open tube furnace and was set to 30 sccm throughout the course of the experiment. Initially the furnace was flushed with Ar to minimize interference from gaseous impurities. The furnace set up with the sample for nanowire growth has been illustrated in Fig. 14.

The base support Si wafer acts as Si vapor source. Then, the heating process was timed after the furnace reached the desired temperature. The growth time was varied between 10 and 90 min in order to derive the growth model. For the estimation of growth rate, the heating time was extended up to 210 min. Further, the growth model was validated by (a) heating the metal coated Si substrate in the absence of Ar flow, (b) removing the nanowires via HF treatment and then inspecting the substrate, and (c) varying the Ar flow rate to observe the change in nanowire growth, if any. Finally, the reusability of the Si substrate was evaluated by removing the nanowires grown on Si substrate by HF solution. Then, the Si substrate was again introduced into the furnace to check for nanowire growth. After the heating experiments, the samples were observed under Hitachi S480 SEM. EDS was also performed to identify elemental compositions in addition to XRD.

Discussions

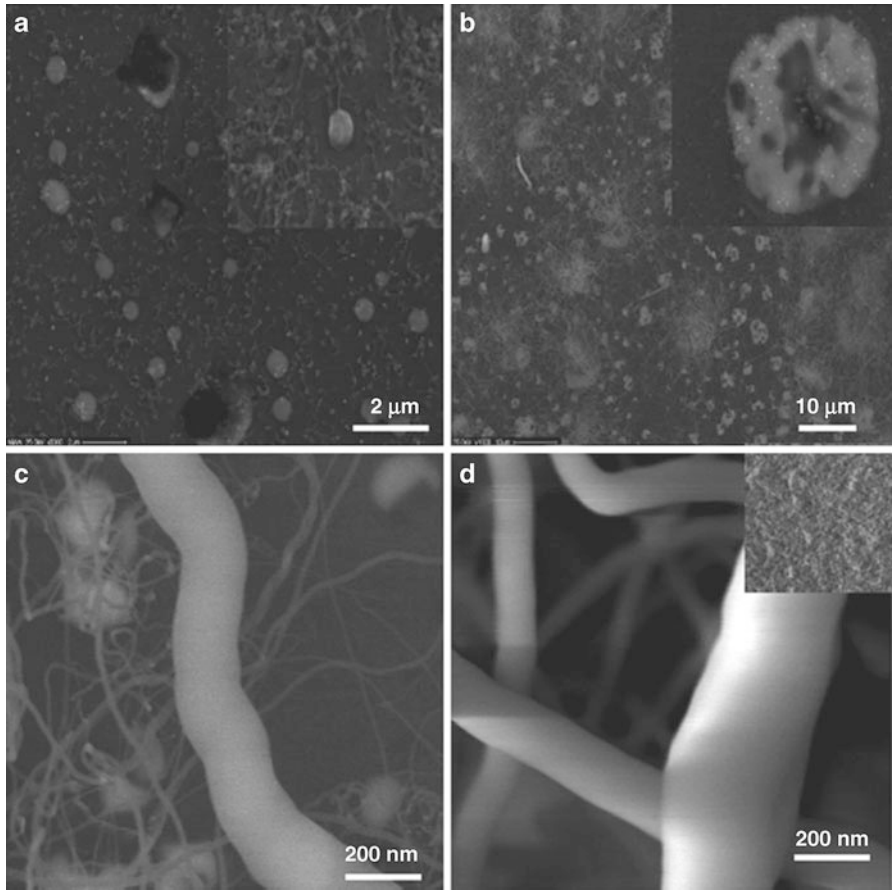
Figure 15 is a SEM composite micrograph showing the morphology of Pd coated Si substrate heated at 1,100 °C for various time intervals.



Growth of Silica Nanowires, Fig. 15 SEM micrograph illustrating the morphology of the Pd-coated Si wafer heated at 1,100 °C. (**a**) 10 min, (**b**) 15 min, (**c**) 20 min, and (**d**) 30 min

Inspecting the sample after heating 10 min, nanoclusters of various shapes and sizes were observed (shown in Fig. 15a). An EDS inspection on one of the nanoclusters indicates the presence of Pd, Si, and O. Further, the XRD data indicates the presence of crystalline Pd as well as the formation of palladium silicide. Also, a small layer (Fig. 15a inset) covering the wafer surface surrounding the island was noticed. An EDS analysis on the thin layer with quantification indicated the layer to be a SiO_2 film. The extremely low partial pressure and concentration of oxygen in high-purity Ar tank was responsible for the formation of the oxide layer at 1,100 °C. A deliberate introduction of oxygen at 0.01 sccm in addition to Ar completely oxidized the substrate without forming the islands or the nanowires. Hence, it is postulated that the thermodynamic equilibrium between the atmosphere (in an open tube furnace) and the low oxygen partial pressure in the tube is critical for nanowire growth. The sample inspected after a heating cycle of 15 min revealed surface roughening (Fig. 15b) in the form of etched pits both on the surface of Si and palladium silicide island. This has been attributed to the disassociation of the thin oxide layer to volatile SiO with continuous supply of oxygen.

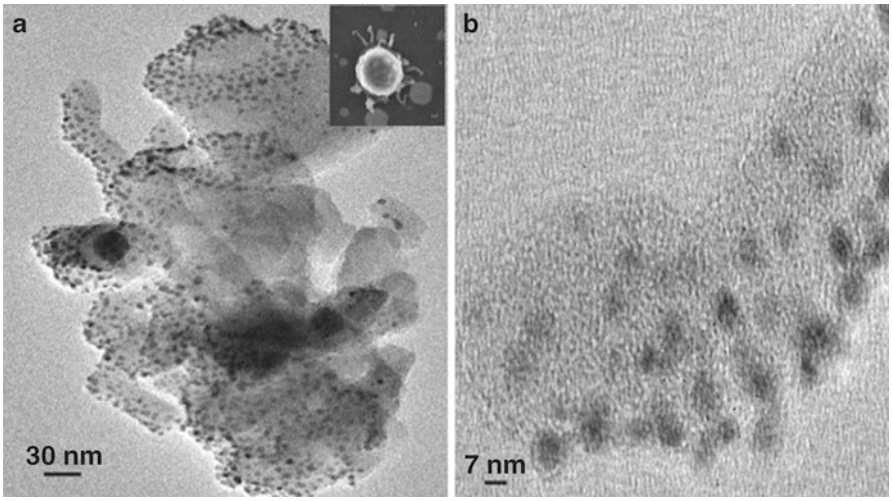
A similar observation of surface roughening of a silicon wafer during annealing in argon ambient at an extremely low oxygen partial pressure was reported by Suzuki [21]. The author concluded that the surface roughening in the form of square pits occur when the oxygen partial pressure is lower than 3×10^{-5} atm. Figure 15c shows the



Growth of Silica Nanowires, Fig. 16 SEM micrograph illustrating the morphology of the Pd-coated Si wafer heated at 1,100 °C. (**a**) 45 min, (**b**) 60 min, (**c**) 75 min, and (**d**) 90 min

SEM micrograph of a sample heated for 20 min. The nucleation of nanowires could be observed along with ripening of the seeds to induce its birth. Localized Si etching was also observed along the catalyst periphery and subsequent sinking action of the nanoclusters was seen. As the eutectic alloy formation proceeds at high temperature, a tiny amount of Si diffuses from the substrate. The silicide formation consumes silicon from the surrounding which gives the appearance of Si etching. Peng et al. [22] utilized the metal-induced Si etching behavior along with nanolithography to synthesize ordered Si nanowires. As the nanoclusters vary in size distribution (several μm to nm), some of the nanowires seem to originate from silicon. The number density of nanowires increased twofold upon inspecting a sample heated at 1,100 °C for 30 min. EDS spectra on a bunch of nanowires away from the palladium nanocluster indicate Si and O composition. Extracting the stoichiometry by background subtraction and in comparison with a standard quartz substrate reveals SiOx ($1.8 < x < 2$) stoichiometry indicating a deficiency in silicon.

The number density of nanowires increased for samples heated for 45 min (Fig. 16a). Also, the inset in Fig. 16a shows the reduction in the size of nanoclusters at the expense of nanowire growth. Upon inspected by transmission electron microscope, the catalyst seems to travel up the nanowire as it grows (Fig. 17). Table 1 indicates the progressive increase in the number density of



Growth of Silica Nanowires, Fig. 17 TEM micrograph showing migration of Pd catalyst particles captured at an early growth stage (3E16 ions/cm^2, 1,100 °C, 7 min). (**a**) "Birth" of nanowires, nucleated wires of various sizes and shapes concentrated with Pd particles, *Inset*: scanning electron microscopy (SEM) micrograph highlighting the early growth stage, Ostwald ripening of the seeds observed. (**b**) Magnified image of a single nanowire for estimating the number density of the wires: 1 × 10^{13} particles/cm^2

Growth of Silica Nanowires, Table 1 Comparison of number density of silica nanowires as a function of heating time

Heating time (mins)	Number density of nanowires (×10^{12} wires/m^2)
20	1
30	2
45	3
60	5
75	6
90	8

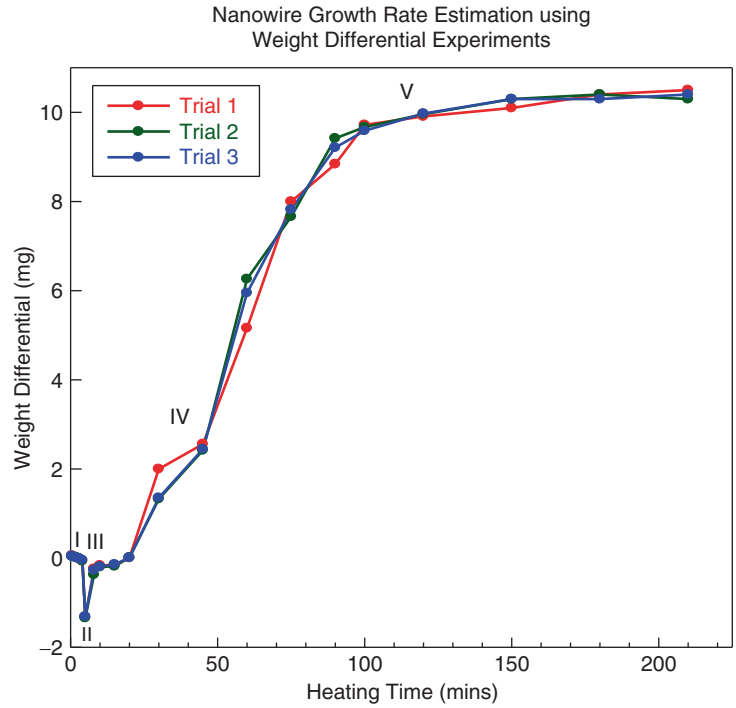
nanowires with increase in heating time. Figure 16b shows bunches of nanowires along with secondary nucleation (inset) from the nanocluster. High aspect ratio nanostructures were seen on samples heated to 75 min (Fig. 16c) along with size reduced parent nanoclusters. Samples heated for 90 min at 1,100 °C show intertwined nanowires Fig. 16d devoid of nanoclusters. The bushy growth of nanowires has been shown in the inset.

Next the growth rate (Fig. 18) was estimated by measuring the weight differential of Si substrate before and after nanowire growth as function of growth time. Five distinct growth regimes, namely, (1) thin oxide growth, (2) Si etch pit formation, (3) nanowire nucleation, (4) nanowire growth, and (5) nanowire growth saturation, have been demarcated.

Based on the weight differential data, it can be concluded that SiO*x* nanowires are conducive to mass manufacturing when heated at 1,100 °C for a heating time above 100 min. Further, the reusability of the substrate was investigated by etching the nanowires off the Si substrate in hydrofluoric acid (HF) and using the processed substrate again for nanowire growth (1,100 °C, 90 min). The nanowires grow at the expense of the catalyst in the sense the catalyst gets consumed inside the nanowires and the Si surface is depleted of the catalyst nanoparticles. Based on these experimental evidences, a growth model was derived and is illustrated in Fig. 19.

Some of the key features of silica nanowire growth via metal thin film-based catalysis and implantation-based seeding approach as inferred from the growth model are as follows:

(a) The low concentration level of oxygen in Ar is the key to the growth of silica nanowires in an open tube furnace.



Growth of Silica Nanowires, Fig. 18 Schematic illustrating the growth rate (weight differential) of silica nanowires as a function of heating time

Absence of nanowire formation was noted in the case of a minute increase in the oxygen level as mentioned earlier. Though the concentration magnitude could not be estimated, the partial pressure of oxygen must be below 3×10^{-5} atm. obtained from earlier report [19] for nanowire growth.

(b) The source of silicon originates both from the substrate (when the metal consumes Si for the alloy formation leaving exposed facets of Si) and the support wafer. Also, the initial oxide layer is transformed into silicon monoxide upon heating the sample at 1,100 °C. Presence of SiO vapor aids nanowire growth by supplying both vapor phase Si and O. Such hypothesis is concurrent with the prior investigation on thermodynamic equilibrium between Si and SiO at 1,100 °C [23].

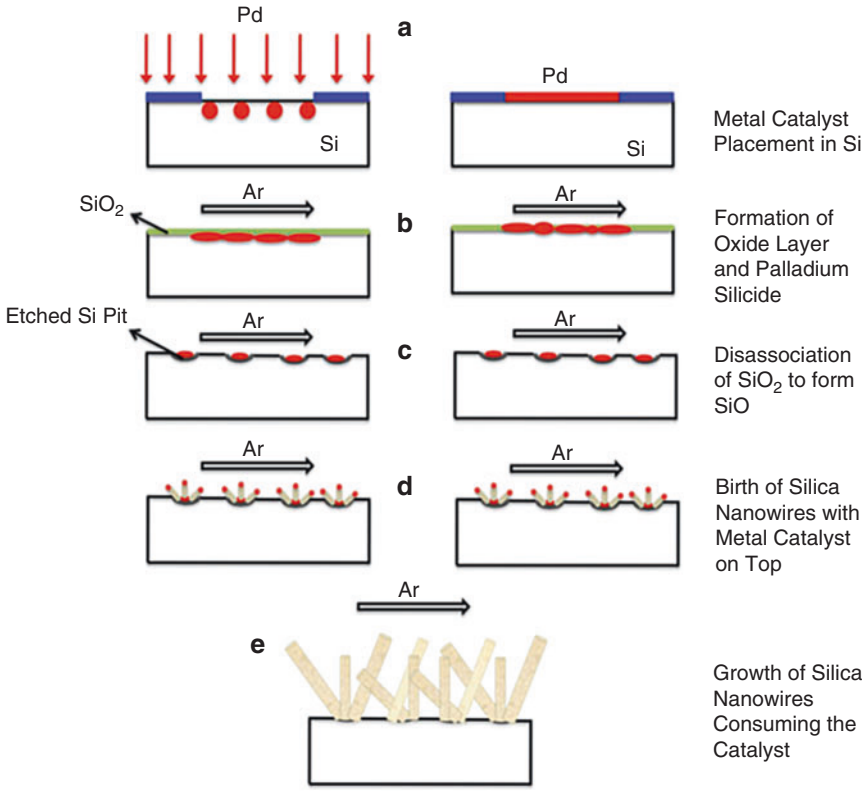
(c) A combination of growth models governs nanowire growth. The presence of metal catalyst at the tip of nanowires, formation of silicide in tandem with the need for Ar suggests Vapor Liquid Solid (VLS) model. To rule out the Solid Liquid Solid (SLS) model, the nanowires were mechanically removed and then the substrate was inspected for any silicide nanocluster. The absence of catalyst at the sample surface negates the possibility of SLS model. Also, the formation of a thin oxide layer and role of SiO inducing nanowire growth signifies oxide-assisted growth (OAG) mechanism.

Conclusions

In summary, the manufacturing aspect of silica nanowire growth was investigated. Nanowires were found to nucleate following a combination of VLS and OAG mechanism. For a given thickness of Pd thin film and heating temperature of 1,100 °C, the nanowire growth seems to saturate after a heating time of 100 min. It has been observed that the nanowires grow at the expense of the catalyst negating the reusability for the substrate.

Summary

In this entry, silica nanowire growth was discussed in view of their potential for several



Growth of Silica Nanowires, Fig. 19 Growth model for silica nanowires on Si (**a**) metal catalyst placement in Si using ion implantation or sputtered thin film, (**b**) after sample introduction into furnace – silicide formation along with a thin oxide layer, (**c**) disassociation of oxide layer to form SiO and formation of Si etch pits with increasing heating time, (**d**) birth of silica nanowires with metal catalyst on top, and (**e**) growth of silica nanowires of high aspect ratio

unique applications such as localization of light, low-dimensional waveguides for functional microphotonics, SNFOM, optical interconnects on a futuristic integrated optical microchip, sacrificial templates, biosensors, and optical transmission antennae. The "bottom-up" and "top-down" approaches in growing nanowires were discussed with emphasis on the VLS-mediated growth. Further, the research on selective growth of silica nanowires on silicon using metal ion implantation–based seeding approach and metal thin film catalysis was presented. While the former technique has the advantage of precise placement of catalyst and near to uniform growth, the latter offers an inexpensive way to mass manufacture them. A unique feature tying both the synthesis methodologies is the use of simple open tube furnace in a high-purity Ar ambient unlike the use of harmful gases like silane or complex instrumentation such as a CVD system.

Time sequence study of the silica nanowire growth reveals a combination of VLS and OAG mechanism responsible for nanowire nucleation. Silica nanowires grown using such a combinatorial mechanism exhibit high aspect ratio (~3,000:1) and are amorphous in nature. Growth rate characteristics indicate an initiation period, prolific growth regime, and a steady state phase.

Structural characterization at a high resolution indicates catalyst metal nanoclusters along its body suggesting the formation of a hybrid (metal/dielectric) nanostructure, critical for sensing applications. Finally, a brief highlight on manufacturing aspects of silica nanowires was presented.

Future Directions

Producing an ordered array of vertically aligned nanowires is critical for many applications. In particular, the biosensing applications require silica nanowires to be grown vertically. Vertically aligned nanowires are easier to manipulate and offer greater surface coverage. Silica nanowires can be grown vertically or in an aligned fashion using electric field–assisted assembly, porous alumina templates, and self-assembly using bioengineered interfaces. Also, growing nanowires at low temperature on various substrates is critical for device applications. Apart from the growth aspects, investigating silica nanowires using Local Electrode Atom Probe (LEAP) will yield interesting results on the elemental composition and texture along the length of the wire. The atom probe is a combination of a field ion microscope and a mass spectrometer. Looking at the nanowires at their early stage of growth using LEAP will provide fundamental details on nucleation and the role of catalyst inducing growth. Moreover, these results can give an insight on the growth model.

Cross-References

- ▶ Biomimetic Synthesis of Nanomaterials
- ▶ Electric-Field-Assisted Deterministic Nanowire Assembly
- ▶ Gas-Phase Nanoparticle Formation
- ▶ Nanostructured Functionalized Surfaces
- ▶ Nanostructured Materials for Sensing
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Optical Techniques for Nanostructure Characterization
- ▶ Scanning Electron Microscopy

References

1. Foster, L.E.: Nanotechnology: Science, Innovation, and Opportunity. Prentice Hall, Upper Saddle River (2009)
2. Cao, G.: Nanostructures and Nanomaterials: Synthesis, Properties, and Applications. Imperial College Press, London (2004)
3. Sui, C.-H., Wu, P.-H., Zhao, P.: Waveguiding properties and applications of silica nanowires in various environments. Optoelectron. Lett. **5**, 241–243 (2009)
4. Choi, D., Mcllroy, D., Nagler, J., Aston, E., Hrdlicka, P., Gustin, K., Hil, R., Stenkamp, D., Branen, J.: One-dimensional silica structures and their applications to the biological sciences. In: Kumar, C. (ed.) Nanostructured Oxides, pp. 83–103. Wiley-VCH, Weinheim (2009)
5. Stathis, J.: Reliability limits for the gate insulator in CMOS technology. IBM J. Res. Dev. **46**, 265–285 (2002)
6. Ollier, E.: Optical MEMS devices based on moving waveguides. IEEE J. Sel. Topics Quantum Electron. **8**, 155–162 (2002)
7. Sekhar, P.K., Ramgir, N.S., Bhansali, S.: Metal-decorated silica nanowires: an active surface-enhanced Raman substrate for cancer biomarker detection. J. Phys. Chem. C **112**, 1729–1734 (2008)
8. Tsuneyuki, S., Tsukada, M., Aoki, H.: First-principles interatomic potential of silica applied to molecular dynamics. Phys. Rev. Lett. **61**, 869–872 (1988)
9. Christian, P.: Preparation of ultrafine silica nanowires using iron based surfactant. Langmuir **26**, 1405–1407 (2010)
10. Kievsky, Y., Sokolov, I.: Self-assembly of uniform nanoporous silica fibers. IEEE Trans. Nanotechnol. **4**, 490–494 (2005)
11. Kita, T., Nishimoto, S., Kameshima, Y., Miyake, M.: Characterization of silicon oxide nanowires prepared by hydrothermal treatment using slow dissolution of bulk glass source materials. J. Am. Ceram. Soc. **93**, 2427–2429 (2010)
12. Lin, Y.-C., Lin, W.-T.: Growth of SiO_2 nanowires without a catalyst via carbothermal reduction of CuO powders. Nanotechnology **16**, 1648–1654 (2005)
13. Rao, C.N.R., Deepak, F.L., Gundaiah, G., Govindaraj, A.: Inorganic nanowires. Prog. Solid State Chem. **31**, 5–147 (2003)
14. Wagner, R.S., Ellis, W.C.: Vapor-liquid-solid mechanism of single crystal growth. Appl. Phys. Lett. **4**, 89–90 (1964)
15. Sekhar, P.K., Abd Bhansali, S.: Manufacturing aspects of oxide nanowires. Mater. Lett. **64**, 729–732 (2010)
16. Muralidhar, G.K., Bhansali, S., Pogany, A., Sood, D. K.: Electron microscopy studies of ion implanted silicon for seeding electroless copper films. J. Appl. Phys. **83**, 5709–5713 (1998)
17. Sood, D.K., Sekhar, P.K., Bhansali, S.: Ion implantation based selective synthesis of silica nanowires on silicon wafers. Appl. Phys. Lett. **88**, 143110–143113 (2006)
18. Schmidt, V., Senz, S., Gosele, U.: Diameter-dependent growth direction of epitaxial silicon nanowires. Nano Lett. **5**, 931–935 (2005)
19. Schmidt, V., Senz, S., Gosele, U.: The shape of epitaxially grown silicon nanowires and the influence of line tension. Appl. Phys. A **80**, 445–450 (2005)

20. Sekhar, P.K., Sanbandam, S.N., Sood, D.K., Bhansali, S.: Selective growth of silica nanowires in silicon catalysed by Pt thin film. Nanotechnology **17**, 4606–4613 (2006)
21. Suzuki, T.: Effect of annealing a silicon wafer in argon with a very low oxygen partial pressure. J. Appl. Phys. **88**, 6881–6886 (2000)
22. Peng, K., Zhang, H., Lu, A., Wong, N.B., Zhang, R., Lee, S.-T.: Ordered silicon nanowires via nanospehere lithography and metal-induced etching. Appl. Phys. Lett. **90**, 163123–163125 (2007)
23. Bahloul-Hourlier, D., Perrot, P.: Thermodynamics of the Au–Si–O system: application to the synthesis and growth of silicon–silicon dioxide nanowires. J. Phase Equilib. Diffus. **28**, 150–157 (2007)

Growth of Silicon Dioxide Nanowires

▶ Growth of Silica Nanowires

GW and BSE Approaches

▶ Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches

Gyroscopes

Alexander A. Trusov
Department of Mechanical and Aerospace Engineering, The University of California, Irvine, CA, USA

Synonyms

Inertial sensors; Motion sensors; Orientation sensors; Rate sensors; Rotation sensors

Definition

Gyroscopes are physical sensors that detect and measure the angular motion of an object relative to an inertial frame of reference. The term "gyroscope" is attributed to the mid-nineteenth century French physicist Leon Foucault who named his experimental apparatus for Earth's rotation observation by joining two Greek roots: *gyros*, rotation and *skopeein*, to see. Unlike rotary encoders or other sensors of relative angular motion, the unique feature of gyroscopes is the ability to measure the absolute motion of an object without any external infrastructure or reference signals. Gyroscopes allow untethered tracking of an object's angular motion and orientation and enable standalone Attitude Heading Reference Systems (AHRS). Combining three gyroscopes with three accelerometers in a complete 6-axis Inertial Measurement Unit (IMU) enables self-contained Inertial Navigation Systems (INS) for navigation, guidance, and dead reckoning.

All *gyroscopes* can be divided into two main categories, depending on whether the angular velocity or orientation is being measured [1]. *Rate gyroscopes* measure the angular velocity, or the rate of rotation of an object. *Angle gyroscopes*, also called *Whole Angle* or *Rate Integrating* gyroscopes, measure the angular position, or orientation of an object directly. While devices sensitive to the angular acceleration are used in some applications, these sensors are typically not referred to as gyroscopes, but rather as angular accelerometers. Essentially all currently existing Micro-Electro-Mechanical-Systems (MEMS) gyroscopes are of the rate measuring type and are typically employed for motion detection (e.g., in consumer electronics and automotive safety devices) and motion stabilization and control (e.g., in smart automotive steering and antenna/camera stabilization systems).

True INS and AHRS applications rely on the continuous tracking of the object's orientation. Measurement of the angular position can be accomplished either by numerical integration of a rate gyroscope's output, or by using an angle gyroscope which effectively integrates the rotation rate by virtue of its internal dynamics and outputs the angle information directly. When a rate gyroscope is used to track the orientation, its output signal is integrated over time together with the associated errors and noise, leading to fast buildup of the orientation angle drifts

(e.g., white noise in the angular rate signal results in $1/f^2$ drift, or random walk, of angle). Successful realization of standalone gyroscope-based INS and AHRS thus requires either angle gyroscopes or rate gyroscopes with extremely stable and low noise output.

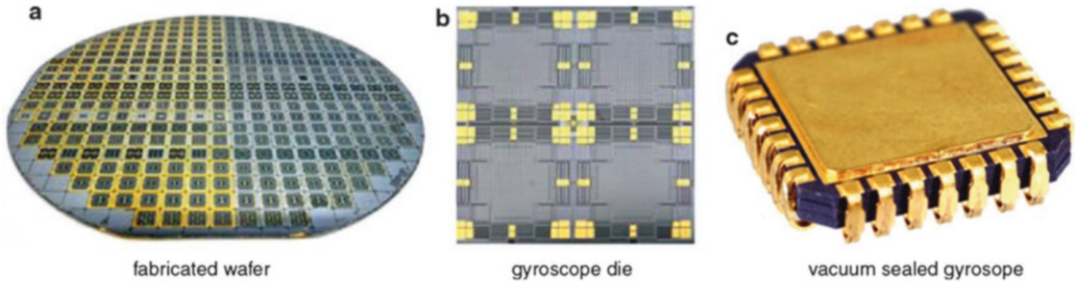
Historical Overview of Vibratory Gyroscope Technologies

Early work by Leon Foucault during the mid-nineteenth century explored two different design paradigms for angle measuring mechanical gyroscope based on either a spinning or vibrating mass. While the spinning mass approach was the dominant method of mechanical gyroscope construction from its inception well into the second half of the twentieth century, it is not well suited for MEMS implementation due to the technological limitations in the manufacturing of precision, low friction bearings. Few designs of spinning mass MEMS gyroscopes using electrostatic levitation have been reported in the literature [2, 3] without yet achieving commercial success due to the inherent instability of the mechanical system and necessity for a sophisticated control system.

The vibrating mass approach, illustrated by the popular Foucault pendulum experiment, exploits the exchange of energy between different axes of vibration due to the Coriolis effect. This architecture, at present referred to as the Coriolis Vibratory Gyroscope (CVG) [4] remained largely a scientific curiosity for almost a century until the introduction of a functional vibratory gyroscope by Sperry in the mid-twentieth century [5] followed by successful commercialization of quartz tuning fork gyroscopes by BEI Technologies in the late-twentieth century [6], and very high volume deployment of silicon MEMS CVGs in the early twenty-first century. Today, silicon vibratory rate gyroscopes with capacitive transduction comprise the majority of MEMS gyroscopes in development and production, with some research groups and manufacturers pursuing quartz devices with piezoelectric transduction or silicon devices with alternative transduction mechanisms such as inductive or electromagnetic [7]. Figure 1 shows photographs of a wafer-level batch fabricated silicon-on-insulator (SOI) gyroscope with capacitive actuation and detection designed, fabricated, and packaged at the University of California, Irvine [8].

Coriolis Vibratory Gyroscope Dynamics

In this section, the principles of operation of Coriolis vibratory gyroscopes are derived from the basic concepts of classical mechanics. While a planar, z-axis vibratory gyroscope is the focus of the explanation, the discussions are generic in nature and equally apply to devices with other architectures, including torsional, out-of-plane, and sensors with multiple modes of vibration for simultaneous detection of rotation in several axes. Let $S\xi\eta\zeta$ denote an inertial reference frame, in which a non-inertial frame $Oxyz$ is moving with a linear acceleration $A_0 = (A_0x,\ A_0y,\ A_0z)$ and an angular velocity $\Omega = (\Omega_x,\ \Omega_y,\ \Omega_z)$, where the components of the



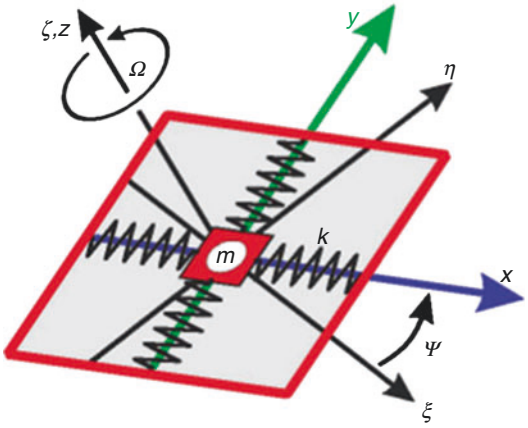
Gyroscopes, Fig. 1 Photographs of a wafer-level fabricated silicon-on-isolator gyroscope with capacitive transduction designed, fabricated, and packaged at the University of California, Irvine MicroSystems Laboratory, http://mems.eng.uci.edu/

vectors are given with respect to the moving frame. Coordinates of a point mass in the inertial and non-inertial reference frames are given by vectors $\rho = (\xi, \eta, \zeta)$ and $r = (x, y, z)$, respectively. Newton's equation of motion of the point mass m under the action of force vector $F = (F_x, F_y, F_z)$ with respect to the inertial reference frame $S\xi\eta\zeta$ are given by $\rho'' = A_{abs} = F/m$, where the prime denotes derivative with respect to time. According to the rules of differentiation in a moving frame:

$$\begin{aligned} r'' &= A_{rel} \\ &= F/m - 2[\Omega \times r'] - (A_0 + [\Omega \times [\Omega \times r]] \\ &\quad + [\Omega' \times r]). \end{aligned} \tag{1}$$

This differential vector equation provides a mathematical foundation for the Coriolis vibratory gyroscopes. The Coriolis cross-product of the angular velocity Ω and the coordinate vector $r = (x, y, z)$ governs the coupling and exchange of energy between x, y, and z non-inertial axes. This effect allows measuring the input angular velocity Ω by observing the vibration pattern of the proof mass m relative to the non-inertial device reference frame $Oxyz$.

A theoretical model of a single axis Coriolis vibratory gyroscope is shown in Fig. 2. A point mass m forms the Coriolis sensitive proof mass and is constrained to motion in the Oxy plane under the influence of elastic forces of the suspension and inertial forces, caused by the motion of the $Oxyz$ non-inertial reference frame. To simplify the conceptual discussion, it is assumed that the proof mass suspension produces a linear field of elastic forces $Fs = (k_x x, k_y y)$. Also neglected are the projection of the gravity field on the Oxy plane, out-of-plane Oz axis dynamics, and acceleration of the origin O. These assumptions are justified by high out-of-plane stiffness of typical bulk micromachined structures and large frequency separation between the external and Coriolis accelerations. After linearization with respect to the input angular velocity vector Ω components Ω_x, Ω_y, Ω_z, Eq. 1 can be simplified to the following:



Gyroscopes, Fig. 2 Theoretical model of a single z-axis vibratory gyroscope consisting of a proof mass m suspended in the x-y plane. The xyz non-inertial frame of reference associated with the sensor is moving relative to the inertial frame $\xi\eta\zeta$ with an angular velocity $\Omega = (0, 0, \Omega_z)$. Coriolis force coupling between the x and y coordinates causes energy exchange which is used to detect the input rate Ω_z

$$\begin{aligned} x'' + x{\omega_x}^2 - y{\Omega'}_z - 2\Omega_z y' &= F_x, \\ y'' + y{\omega_y}^2 + x{\Omega'}_z + 2\Omega_z x' &= F_y, \end{aligned} \tag{2}$$

where ω_x and ω_x are the mechanical natural frequencies of the x- and y-mode, respectively. The following discussion will consider two different modes of operation of Coriolis vibratory gyroscopes – rate measuring and angle measuring.

Rate Gyroscope Operation

Rate gyroscopes are operated with inherent nonsymmetry between the Ox axis, designated as the drive-mode, and the Oy axis, designated as the sense-mode. The drive-mode is operated in the forced vibrations mode, where the excitation force F_x is a sinusoidal waveform with amplitude f and angular frequency ω_d. The sense-mode of a rate gyroscope can be operated either open-loop or in a force-to-rebalance closed loop, where a feedback force is generated to suppress the sense-mode vibrations and is simultaneously used as the measure of the input rate. Dynamics of

a rate gyroscope with an open-loop sense-mode can be derived from Eq. 2 as:

$$x'' + x{\omega_x}^2 - y\Omega'_z - 2\Omega_z y' = f/m\,sin\,(\omega_d t),$$
$$y'' + y{\omega_y}^2 + x\Omega'_z + 2\Omega_z x' = 0. \tag{3}$$

In practical implementations of vibratory rate gyroscopes, stabilization of the drive-mode velocity is desired to reduce the effect of manufacturing tolerances and operational conditions on the scale factor of the sensor. This is accomplished by an electronic feedback system which regulates the amplitude f of the driving force by means of an *Automatic Gain Control (AGC)* to maintain constant amplitude of the drive-mode motion. The frequency ωd of the drive-mode feedback force is generated by a *Phase-Lock Loop (PLL)* which locks to the drive-mode mechanical resonance frequency ωx. An alternative approach for drive-mode operation is through positive feedback of the velocity signal x', which effectively destabilizes the drive-mode and maintains resonance of the drive-mode. In the velocity feedback configuration, the need for a separate frequency source is eliminated, however often at the cost of increased jitter and noise in the Coriolis signal demodulation block. Simplified equations for a vibratory rate gyroscope with a closed-loop drive-mode and an open-loop sense-mode (see Fig. 3) are derived from Eq. 3 as:

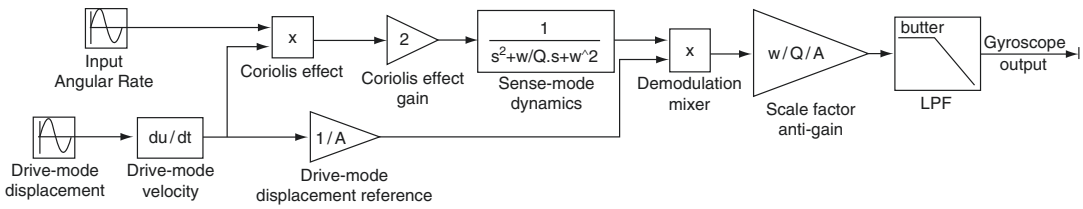
$$x = sin\,(\omega_x t),$$
$$y'' + y'\left(\omega_y/Q_y\right) + y\omega_y^2 = -2\Omega_z x', \tag{4}$$

Here, an energy dissipation term, defined by the sense-mode quality factor Q_y, is now included in the sense-mode dynamics, while the term $x\Omega_z'$ is omitted assuming a slow-varying input Ω_z.

Figure 4 illustrates the motion of the proof mass relative to the device reference frame in the presence of input rotation. The drive-mode is a resonator driven to a constant amplitude of vibrations at a fixed frequency by means of a feedback system. During rotation of the device, the Coriolis effect causes coupling of energy from the drive-mode to the sense-mode. The Coriolis force applied to the sense-mode is given by the cross-product of and is proportional to the drive-mode velocity x' and the input angular rate Ω_z. Since the drive-mode velocity is a sinusoidal signal with a fixed frequency, the Coriolis effect results in the *Amplitude Modulation (AM)* of the input angular rate by the drive-mode velocity. To measure the input angular rate, displacement (or, equivalently, velocity) of the Coriolis force induced sense-mode vibrations is measured by means of an amplitude demodulation.

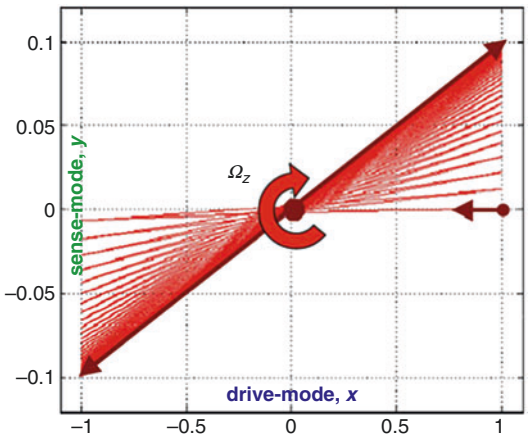
Coriolis vibratory rate gyroscopes are commonly divided into two classes depending on the relative matching or mismatching between the drive-mode and sense-mode resonant frequencies. When the drive frequency is within the bandwidth of the sense-mode resonance peak, the mechanical sensitivity of the gyroscope is maximized proportionally to the sense-mode quality factor Q_y. However, the sensitivity and *Signal-to-Noise (SNR)* advantages of this configuration come at the trade-off of the open loop rate sensor bandwidth, which in this case is inversely proportional to the sense-mode Q_y. An alternative design approach is the mismatching of the drive frequency from the sense-mode resonance by a prescribed amount.

In order to illustrate the dynamic response of a mode-mismatched rate gyroscope to a



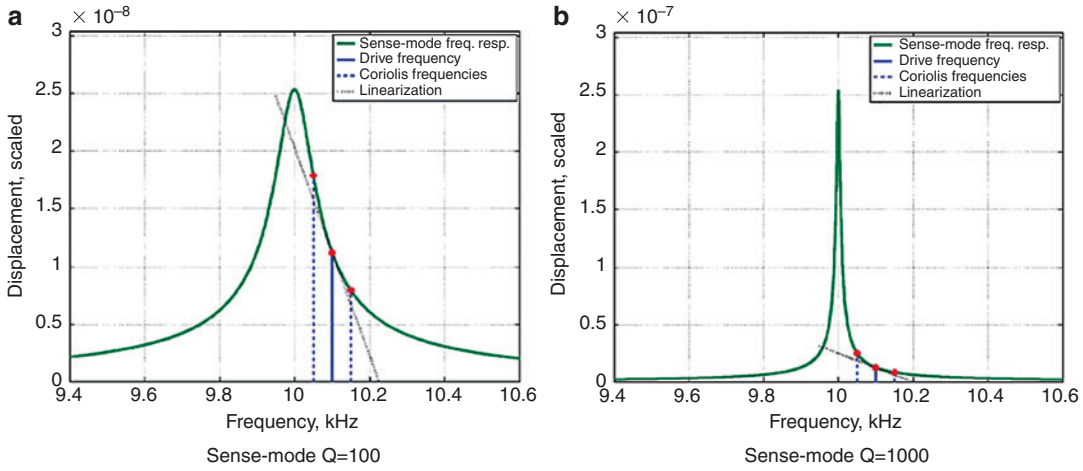
Gyroscopes, Fig. 3 Block diagram of a Coriolis vibratory rate gyroscope with closed-loop drive- and open-loop sense-modes. Input rotation rate is amplitude modulated by the drive-mode velocity due to the Coriolis effect

time-varying input angular velocity is considered in the form of $\Omega_z = cos(\omega_{in}t)$ and the output scale factor is studied as a function of the input frequency ω_{in}. According to Eq. 4, the sense-mode is excited by $-2\Omega_z x' = \omega_d[cos((\omega_d - \omega_{in})t) + cos((\omega_d - \omega_{in})t)]$. This means that for a given $\Omega_z = cos(\omega_{in}t)$ input, the sense-mode has two forcing components of equal amplitudes at frequencies $(\omega_d \pm \omega_{in})$. This concept is illustrated with numerical modeling in Fig. 5, where the sense-mode frequency is set to 10 kHz with a quality factor of 100 and 1,000, the drive-mode frequency is 10.1 kHz (i.e., 1 % mismatching), and the input angular rate frequency is 50 Hz.

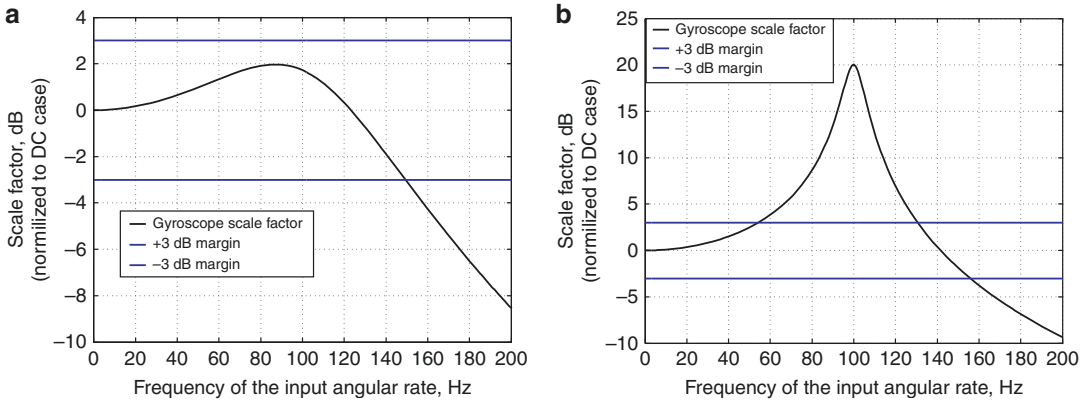


Gyroscopes, Fig. 4 Operation of a vibratory rate gyroscope is based on the excitation of the sense-mode by the Coriolis force, stemming from the cross-product of the drive-mode velocity and the input rotation rate. Sense-mode vibrations present an Amplitude Modulated (*AM*) signal of the input rotation rate with the drive-mode frequency serving as the carrier

As shown in Fig. 3, the signals at two Coriolis frequencies $(\omega_d \pm \omega_{in})$ both pass through the sense-mode dynamics and are then mapped back into a single frequency ω_{in} by the demodulation block. Figure 6 shows the input–output frequency response of a mode-mismatched vibratory gyroscope highlighting two distinct cases defined by the relationship between ω_y, Q_y, and ω_d. For a relatively low sense-mode Q_y, the gyroscope's 3-dB rate bandwidth includes the resonant peak at the $(\omega_d - \omega_y)$ frequency and the bandwidth exceeds the frequency mismatch between the drive- and sense-modes. For relatively high sense-mode Q_y, the resonant peak at the $(\omega_d - \omega_y)$ frequency escapes the gyroscope bandwidth, which in this case is given by approximately half of the drive- and sense-mode frequency mismatch. A detailed discussion on the mechanical design of vibratory rate gyroscopes is given in [9], where a multi-degree of freedom mechanical design of drive- and sense-modes is proposed as a way to increase the sensor robustness and



Gyroscopes, Fig. 5 Frequency domain representation of a mode-mismatched vibratory rate gyroscope response to a sinusoidal input angular rate with a 50 Hz frequency. The simulation assumes sense-mode resonant frequency of 10 kHz and drive-mode excitation at 10.1 kHz. (**a**) Sense-mode Q = 100. (**b**) Sense-mode Q = 1,000



Gyroscopes, Fig. 6 Input rate bandwidth of a mode-mismatched vibratory rate gyroscope. The simulation assumes sense-mode resonant frequency of 10 kHz and drive-mode excitation at 10.1 kHz. (**a**) Sense-mode at 10 kHz, Q = 100, 10.1 kHz drive (**b**) Sense-mode at 10 kHz, Q =1000, 10.1 kHz drive

bandwidth. A different approach to resolving the quality factor versus bandwidth trade-off is the instrumentation of a vibratory gyroscope for direct angle measurements, which is based on the *x*-*y* symmetric free vibrations regime.

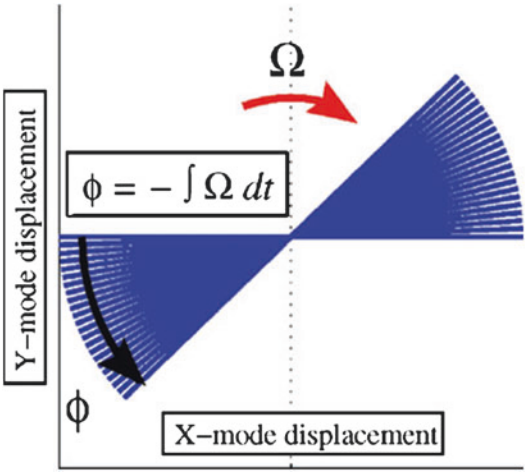
Angle Gyroscope Operation

The resolution and sensitivity of MEMS vibratory gyroscopes are often improved by maximizing *Q*-factors and reducing frequency separation between the two modes. Mode-matched silicon gyroscopes with *Q* of 100,000 have been demonstrated to provide sub-degree-per-hour bias stability [10]. However, mode matching of a high-*Q* rate gyroscope typically limits the open-loop sense-mode range to below 50 deg/s and rate bandwidth to 1 Hz. The traditional approach is forced-to-rebalance closed-loop operation of the sense-mode. While this approach extends the input–output bandwidth of a rate gyroscope, the signal-to-noise advantage of the high-quality factor is only preserved at low frequencies of the input angular rate. Two solutions have been proposed. The first approach is to design a relatively low-*Q* gyroscope with sub-micron capacitive gaps required for efficient transduction at a very high resonant frequency of 3 MHz [11]. The second approach is to operate an ultrahigh *Q*-factor gyroscope in the free vibrations mode, enabling direct angle measurements with mechanically unlimited sensing range and bandwidth.

Whole Angle operation dictates stringent requirements for the structural frequency and Q-factor symmetry, calling for new design architectures and fabrication methods. For example, the macroscale hemispherical resonator gyroscope (HRG) with sub-arcsecond angle resolution operates based on a mechanical element with isotropic *Q*-factors as high as 26 million [12]. Achieving the required level of damping and stiffness symmetry across process variations of conventional silicon MEMS technologies is very challenging. The latest advances in design and packaging of silicon MEMS devices enabled vibratory gyroscopes with *Q*-factors of one million [8], invigorating the development of angle measuring MEMS gyroscopes.

An ideal angle gyroscope, such as the classical Foucault pendulum, is a 2-D isotropic mass-spring system vibrating with the natural frequency ω, where $\omega = \omega_x = \omega_y$. In the presence of the inertial rotation with the rate Ω_z, the equations of motion can be derived from Eq. 2 by assuming very low dissipation and absence of any external forces as:

$$\begin{aligned} x'' + x\omega^2 - y\Omega_z' - 2\Omega_z y' = 0, \\ y'' + y\omega^2 + x\Omega_z' + 2\Omega_z x' = 0. \end{aligned} \tag{5}$$



Gyroscopes, Fig. 7 Illustration of the whole-angle mode: the orientation of the vibrations axis remains fixed in the inertial space, enabling direct angle detection

The whole-angle mode is based on the free precession of the axis of vibrations in response to the inertial rotation, Fig. 7. Orientation of the free vibrations axis remains fixed in the inertial space, providing an instantaneous orientation reference. Unlike the conventional rate measuring mode, where the axis of vibration is effectively locked to the intended drive direction, the whole-angle mode is not mechanically limited in rotation rate range or bandwidth. Changes in the precession angle Θ can be detected by monitoring x and y displacement components according to $\Theta = atan\ (y/x)$.

The above conceptual analysis of the whole-angle operating mode assumes isotropic stiffness and negligible damping. Detailed error model analysis shows that the key sources of the angle error are the frequency mismatch $\Delta\omega = \omega_x - \omega_y$ and energy dissipation time constant mismatch $\Delta(1/\tau) = 1/\tau_x - 1/\tau_y$. Minimization of the damping asymmetry drift requires very low dissipation structures with isotropic Q-factors. For instance, the damping asymmetry $\Delta(1/\tau)$ of 1e-5 rad/s with τ of 160 s (equivalently, $\Delta Q/Q$ of 0.1 %) is equivalent to the angle drift of 1 deg/h. These considerations motivate the advancement of fabrication of flat silicon structures with post-fabrication trimming and electrostatic tuning. Novel 3-D fabrication approaches for axisymmetric shell resonators with inherent smoothness and isotropy provide another promising possibility [13].

Gyroscopes, Table 1 Typical performance parameters of state-of-the-art commercial MEMS rate gyroscopes Ref. [16]

Parameter	Value
Rate resolution	5–10 deg/h/√Hz
Temperature sensitivity of bias	10 deg/h/deg C
Bias repeatability (temperature, turn on)	30 deg/h
Temperature sensitivity of scale factor	250 ppm/°C
Scale factor repeatability (temperature, turn on)	400 ppm

Historical Milestones of MEMS Gyroscopes

Since their initial introduction two decades ago in 1991 [14], the resolution of silicon micromachined gyroscopes has been improving by approximately a factor of 10 every 2 years [15]. Most recent literature on state-of-the-art vibratory MEMS gyroscopes with electrostatic/capacitive transduction reports precision on the order of 10 deg/h [16] with a trend of achieving better than 1 deg/h in controlled conditions [10] (Table 1). Despite the remarkable continuing improvement in the angular rate resolution achieved by micromachined vibratory gyroscopes, measurement repeatability, accuracy, and long-term drift still limit deployment in a variety of field applications where autonomous and repeatable operation is required over extended time and varying environments. Currently, there are several companies producing tactical grade inertial sensors and systems based on the silicon MEMS technology, Table 2; inertial grade systems are yet to be demonstrated. A comprehensive review of MEMS vibratory rate gyroscope commercialization trends are presented in [9]; here, some of the most significant milestones published in literature are reviewed.

In 2002, Analog Devices reported a surface micromachined polysilicon gyroscope with a

Gyroscopes, Table 2 Performance requirements for three grades of gyroscopes Ref. [15]

Parameter, unit	Rate grade	Tactical grade	Inertial grade
Angle random walk (ARW), deg/√h	>0.5	0.5–0.05	<0.001
ARW-equivalent rate resolution, deg/h/√Hz	>30	30–3	<0.06
Bias drift, deg/h	10–1,000	0.1–10	<0.01
Scale factor accuracy, ppm	100–1,000	10–100	<1
Full scale linear range, deg/s	50–1,000	>500	>400
Bandwidth, Hz	>70	100	100

50 deg/h random bias instability achieved at the averaging time below 1 min [17]. The single-chip integrated device was not vacuum packaged, which in combination with a 4 μm thin structural layer translates into a relatively low sense-mode quality factor on the order of 20. This family of gyroscopes has since gained a widespread acceptance and popularity in consumer and automotive markets, with new gyroscopes introduced every few years. According to Table 2, this device falls into the middle of the Rate grade category, as defined in [15]. Publication [17] also projected that the random bias can be improved by approximately 80 % by introducing new electronics design, suggesting a mechanical-thermal limit on the level of 10 deg/h. The newer generation design of the gyroscope has four electrically coupled, low quality factor surface micromachined polysilicon Coriolis vibratory gyroscopes integrated on the same die together with the electronics in order to suppress the effect of external vibrations; improvement of the bias instability was also reported.

A higher performance polysilicon micromachined gyroscope from Bosch was reported in 2007 in [18]. The device has a 1 deg/h random bias instability, showing its potential for the Tactical grade performance category, Table 2. However, this high sensitivity occurs only at a long, averaging times on the order 160 min. The minimal detectable signal at practical bandwidths was lower (30 deg/h for a 10 Hz bandwidth, and 100 deg/h for a 100 Hz bandwidth). This device was not integrated with electronics on the die level (hybrid integration of the MEMS and the electronics dies was done on a package level), and its low noise level is attributed to vacuum packaging and high sense-mode quality factor.

A 2006 publication [16] from Charles Stark Draper Laboratory provides an insightful review of performance characteristics of a few additional commercial MEMS gyroscopes. According to this paper, the uncompensated thermal sensitivity of higher performance quartz tuning fork gyroscopes from Custom Sensors and Technologies (formerly Systron Donner/BEI) was reported at 300 ppm/°C, achieved primarily due to extremely low sensitivity of the material's properties to temperature. Analog Devices' automotive grade angular rate sensor ADXRS150 scale factor uncompensated thermal sensitivity was quoted as 1,700 ppm/°C. Honeywell's dissolved wafer, single crystal silicon on glass, device based on the Draper laboratory design achieves a 250 ppm/°C scale factor temperature coefficient.

Surveying the literature confirms that in recent years performance of the commercial MEMS gyroscopes continued to improve, reaching the criteria of Tactical grade category; however, performance of these gyroscopes is still two to three orders of magnitude away from the Inertial grade. The reported progress often comes with costly refinements of the fabrication and vacuum packaging processes. Most reported commercial gyroscopes have intricate electronic compensation for such parameters as temperature drift and aging. From the mechanical structure point of view, the majority of commercial gyroscopes consist of several coupled, anti-phase-driven vibratory gyroscope elements (2- or 4-tine tuning fork architecture) to suppress external vibrations.

Striking breakthroughs in very large scale commercialization of very low cost (<$1 per axis) MEMS rate gyroscopes and accelerometers have been achieved during 2005–2010 by InvenSense, ST Microelectornics, and other suppliers targeting consumer electronics devices. In contrast to the automotive or defense applications with stringent requirements for reliability and high accuracy, size and price are the main drivers

for the consumer electronics applications. Current emphasis in low-cost inertial sensor commercialization is on a System-in-a-Package (SiP) and ultimately single-chip integration of 6 axis of inertial sensing augmented with a 3 axis magnetometer and a pressure sensor. Kalman filter-based sensor fusion is employed to produce useful 6-axis position and orientation information by taking advantage of the inherent sensory redundancy in the 10 degrees of freedom produced by these sensors.

Systematic Performance Parameters

In this section an explanation is provided of several most important terms used to describe and quantify the characteristics of Coriolis vibratory gyroscopes [4] with emphasis on deterministic and stochastic aspects of performance or rate sensors.

Scale factor (SF) is the ratio between the output and input signals in the linear regime of measurements. For rate gyroscopes with analog voltage output, *SF* is often expressed in V/(deg/s) or V/(deg/h). Several nonideal properties of *SF* are typically characterized, including sensitivity to temperature (in ppm/°C) and linear accelerations (in ppm/g), and *SF* turn-on repeatability (in ppm for 1-σ or 3-σ statistics). Common sources of *SF* error in MEMS gyroscopes are fabrication imperfections, drifts of drive- and sense-mode resonant frequencies and quality factors, and drifts of electronics gains, especially due to turn-on self-heating.

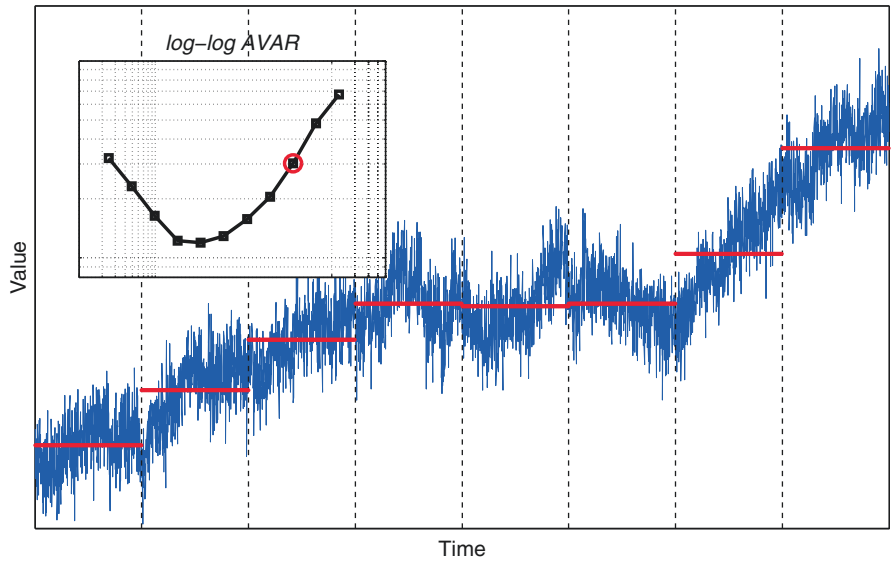
Bandwidth (BW) is the range of the input signal's frequencies which can be measured well by the gyroscope. BW is often specified assuming a ±3 dB margin around the gain (Scale Factor) of the gyroscope for a nonzero constant input rate. In general, the gyroscope's input–output BW should be distinguished from the individual mechanical bandwidths of the drive- or the sense-modes of vibration. A more complete picture of the gyroscopes dynamic response is given by a frequency response plot.

Bias is the apparent output of the rotation sensor for zero input rate assuming specific fixed operational conditions (e.g., temperature). For rate gyroscopes *Bias* is expressed in deg/s or deg/h units. Typical associated nonidealities are the sensitivity of bias to temperature (in deg/s/°C or deg/h/°C) and acceleration (deg/s/g or deg/h/g), *Bias* turn-on repeatability (in deg/s or deg/h for 1-σ or 3-σ statistics). A major sources of *Bias* error in MEMS gyroscopes is mechanical quadrature – the undesired sense-mode vibration caused by the fabrication imperfections. While phase-sensitive demodulation suppresses the quadrature signal by several orders of magnitude, minute phase drifts in the system cause so-called quadrature spillover error – the erroneous change of bias due to the partially unsuppressed quadrature. Even though random noise and drift effects are often referred to as bias, it is important to draw a clear distinction between systematic and random changes of output.

Random Noise Performance Parameters

Bias Instability (BI) refers to the additive error in the gyroscope's output with stochastic or random characteristics. Different random noise modes dominate the overall gyroscope noise level depending on the averaging time, which in turn defines the useful sensor bandwidth. The Allan deviation method is often used to identify and quantify random noise modes with different autocorrelation properties and expose their effect when the output signal is averaged or integrated over time.

Allan deviation analysis of a time domain signal $w(t)$ consists of computing its Allan deviation (also know as root Allan variance) $\sigma(\tau)$ for different integration time constants τ and then analyzing the characteristic regions and log–log scale slopes of the $\sigma(\tau)$ curve to identify different noise modes, i.e., random additive components of the signal with different autocorrelation power laws. The first step of the analysis is to acquire a time history $w(t)$ of the gyroscope's output using an experimental setup or a numerical computer simulation. The second step is to fix the integration time constant τ, divide the time history of the signal into K number of clusters of width τ, then average the



Gyroscopes, Fig. 8 Illustration of the Allan deviation analysis procedure using a simulated random process with a white noise (−1/2 slope in the Allan deviation) and a random walk component (integral of white noise, +1/2 slope in the Allan deviation)

signal over each cluster to obtain $w_{i\,=\,1\ldots K}$, and finally compute the Allan variance defined as one half of the average of the squares of the differences between the successive averaged values: $\sigma^2(\tau) = 0.5 < (w_{i+1} - w_i)^2 >$. The described sequence of steps yields an estimated value of the Allan deviation $\sigma(\tau)$ for the chosen integration time constant τ, Fig. 8. To obtain the whole $\sigma(\tau)$ curve, the computation is repeated multiple times for a sequence of τ values. Typically, the integration time constants τ are iterated through the multiples or powers of the time discretization step (inverse of the signal sampling rate). While an Allan deviation curve for a signal contains essentially the same information as the signal's power spectral density (*PSD*), it presents the information in an alternative and often more convenient way. The power laws of *PSD* and Allan deviation are related to each other in the following way:

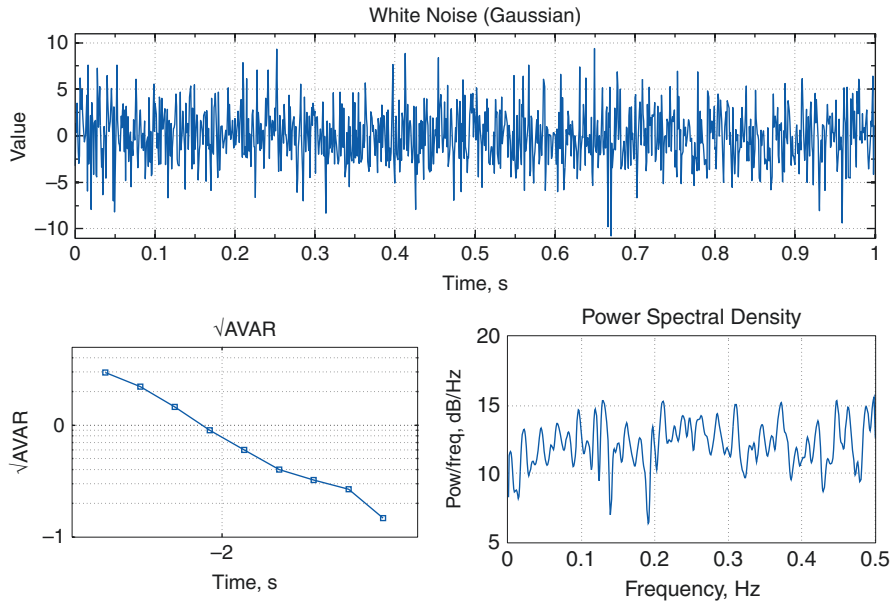
$$PSD(f) \propto f^{\alpha} \text{ is equivalent to, } \sigma(\tau) \propto \tau^{\beta}, \quad \text{where } \beta = -(\alpha + 1)/2 \text{ and } f = 1/\tau. \tag{6}$$

Table 3 presents a classification of the main noise components in MEMS gyroscopes based on their power spectral density and Allan deviation power laws.

White noise is a random signal with a constant power spectral density. While this definition presents a mathematical abstraction not physically possible due to its infinite total power, it gives a useful model of random processes with very short characteristic autocorrelation times. For example, the random motion of particles due to mechanical-thermal noise at room temperature has a characteristic cutoff frequency of approximately 6 THz. Important characteristic properties of a white noise type random process in time, frequency, and Allan deviation domains are shown in Fig. 9 using a simulated signal. Power spectral density of the white noise is independent of frequency, i.e., $PSD(f) \propto f^{\alpha}$ with $\alpha = 0$. As expected from Eq. 6, the numerically estimated Allan deviation of the white noise is governed by $\sigma(\tau) \propto \tau^{\beta}$ with $\beta = -1/2$. In the Allan deviation plot, the rate white noise is identified by fitting the $-1/2$ sloped part of the log-log scale curve with $y = ax^{-1/2}$ and estimating the value of a. In rate gyroscopes, the amount of white noise is quantified in terms of its power spectral density with deg/s/√Hz or deg/h/√Hz units. The value is often referred to as the *rate resolution* (same units) or *Angle Random Walk* (converted to deg/√h).

Gyroscopes, Table 3 Classification of random noise modes as applied to rate gyroscopes

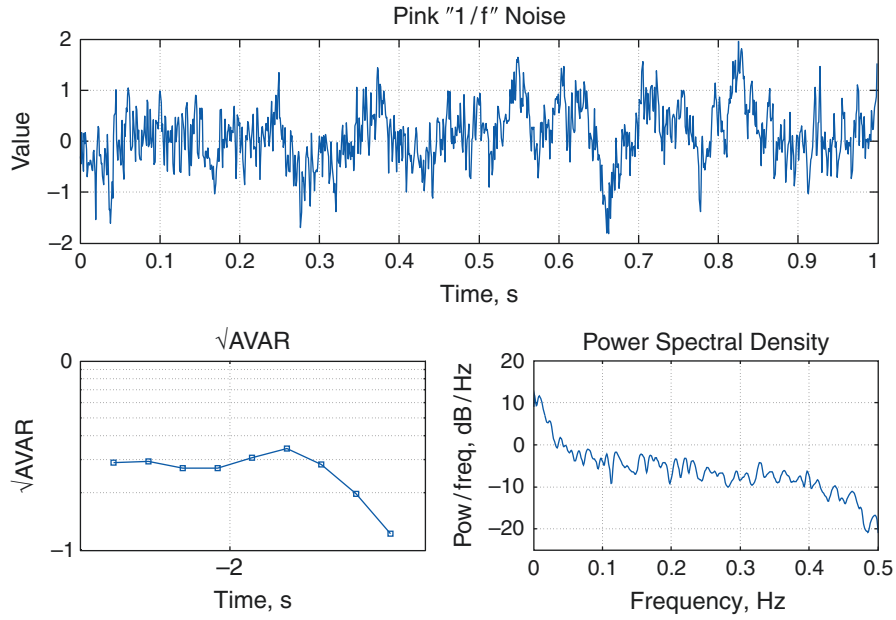
Spectral type	Example sources	*PSD(f)* power law f^{α}	$\sigma(\tau)$ power law τ^{β}	Averaging over time	Associated gyroscope parameters
White	Johnson–Nyquist thermal noise	f^{0}	$\tau^{-1/2}$	Good	Rate resolution in deg/s/√Hz or deg/h/√Hz
					Angle random walk (ARW) in deg/√h
Pink	Electronics flicker	f^{-1}	τ^{0}	Neutral	Flicker, bias instability in deg/s or deg/h
Red (Brown)	White noise accumulation	f^{-2}	$\tau^{+1/2}$	Bad	Rate Random Walk (RRW) in deg/s*√Hz or deg/h*√Hz



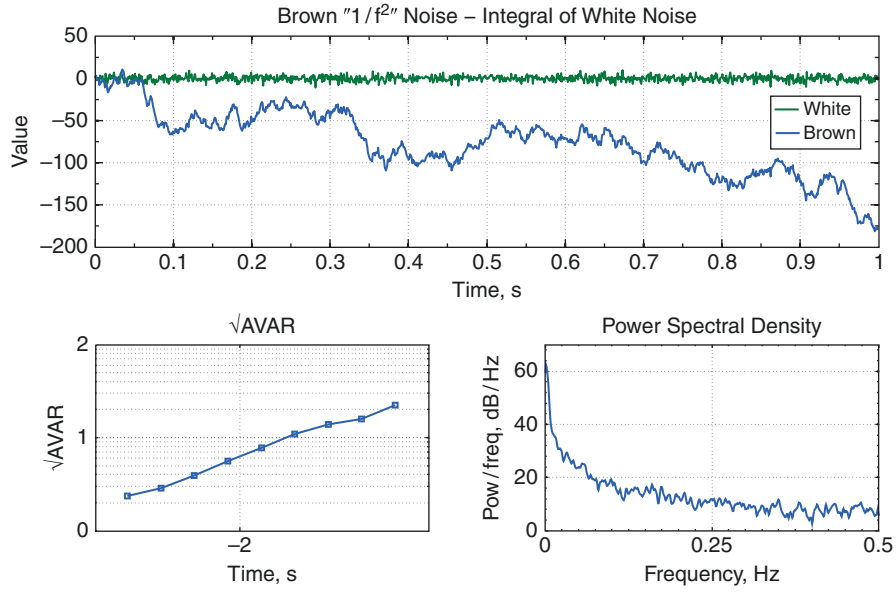
Gyroscopes, Fig. 9 Illustration of the $1/f^{0}$ white noise properties in time, frequency, and Allan deviation domains using a numerical example

Random processes with $1/f$ *PSD* are naturally the next class of noise to consider. *Flicker noise*, commonly encountered in electronic circuits, falls into this category. Figure 10 illustrates time, frequency, and Allan deviation domain properties of $1/f$, or pink noise using a computer simulation. A pink noise can be obtained by low-pass filtering a white noise with a −3 dB per octave filter. Pink noise can be distinguished from a white noise by more visible low-frequency fluctuations. As expected from Eq. 6, Allan deviation has essentially a flat profile in the case of pink $1/f$ noise. In rate gyroscopes, $1/f$ noise often defines the smallest value of the Allan deviation, which is referred to as the *Bias Instability* and has units of deg/s or deg/h.

The third important type of random noise is red noise characterized by $1/f^{2}$ *PSD*. This type of random process is called a *Random Walk* and occurs whenever a white noise is accumulated over time. While not correct from the color spectral parallel, red $1/f^{2}$ noise is sometimes referred to as Brown noise, in honor of Robert Brown who first studied Brownian motion. Figure 11 illustrates time, frequency, and Allan deviation domain properties of $1/f^{2}$ *Random Walk* obtained



Gyroscopes, Fig. 10 Illustration of the *1/f* ("*pink*") noise properties in time, frequency, and Allan deviation domains using a numerical example



Gyroscopes, Fig. 11 Illustration of the $1/f^2$ ("red" or "Brown") noise properties in time, frequency, and Allan deviation domains using numerical example

by numerical integration of a white noise signal. Another way to obtain a random walk process is to filter white noise with a −6 dB per octave filter. Random walk signal shows strong low-frequency fluctuations. The power spectral density of a random walk is given by $PSD(\mathrm{f}) \propto f^{\alpha}$ with $\alpha = -2$ and its root Allan deviation profile is governed by $\sigma(\tau) \propto \tau^{\beta}$ with $\beta = +1/2$. In the Allan deviation

plot, the rate random walk is identified by fitting the +1/2 sloped part of the curve with $y = bx^{+1/2}$ and estimating the value of b. In rate gyroscopes, $1/f^2$ noise is typically referred to as the *Rate Random Walk* (RRW) and quantified in deg/s*√Hz or deg/h*√Hz.

When a rate gyroscope is used to track or maintain the angular orientation of an object, the rate signal is integrated over time together with the associated white noise of rate, flicker noise of rate, and random walk of rate. Upon integration, the three rate noise sources with $1/f^0$, $1/f^1$, and $1/f^2$ densities produce, respectively, drifts of angle with $1/f^2$, etc., power spectral densities. Each of these errors of angle grows fast with time, leading to degradation of angle accuracy. Angle measuring (rate integrating) micromachined gyroscope can potentially provide much better accuracy and stability since the time integration of the noise is avoided.

Cross-References

- ▶ Basic MEMS Actuators
- ▶ Capacitive MEMS Switches
- ▶ Laterally Vibrating Piezoelectric Resonators
- ▶ MEMS Packaging
- ▶ Piezoelectric MEMS Switch

References

1. Shkel, A.M.: Type I and Type II micromachined vibratory gyroscopes. In: Position, Location, and Navigation Symposium, San Diego, 25–27 Apr 2006, pp. 586–593. IEEE/ION, doi:10.1109/PLANS.2006.1650648. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1650648/isnumber=34615 (2006)
2. Nakamura, S.: MEMS inertial sensor toward higher accuracy and multi-axis sensing. In: Sensors, 30 Oct–3 Nov 2005, pp. 4. IEEE. doi:10.1109/ICSENS.2005.1597855. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1597855&isnumber=33608 (2005)
3. Damrongsak, B., Kraft, M.: A micromachined electrostatically suspended gyroscope with digital force feedback. In: Sensors, 30 Oct–3 Nov 2005, pp. 4. IEEE, Irvine. doi:10.1109/ICSENS.2005.1597720. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1597720&isnumber=33608 (2005)
4. IEEE standard specification format guide and test procedure for coriolis vibratory gyros. IEEE Std 1431–2004, pp. 0_1-69 doi:10.1109/IEEESTD.2004.95744. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1394849&isnumber=30352 (2004)
5. Barnaby, R.E., Chatterton, J.B., Gerring, F.H.: General theory and operational characteristics of Gyrotron angular rate tachometer. Aeronaut. Eng. Rev. **12**(11), 31–36 (1953)
6. Madni, A.M., Costlow, L.E., Knowles, S.J.: Common design techniques for BEI GyroChip quartz rate sensors for both automotive and aerospace/defense markets. IEEE Sensors J. **3**(5), 569–578 (2003). doi:10.1109/JSEN.2003.817728. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1234893&isnumber=27682
7. Acar, C., Schofield, A.R., Trusov, A.A., Costlow, L.E., Shkel, A.M.: Environmentally robust MEMS vibratory gyroscopes for automotive applications. IEEE Sensors J. **9**(12), 1895–1906 (2009). doi:10.1109/JSEN.2009.2026466. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=%26arnumber=5297817%26isnumber=5290388
8. Trusov, A.A., Prikhodko, I.P., Zotov, S.A., Schofield A.R., Shkel, A.M.: Ultra-high Q silicon gyroscopes with interchangeable rate and whole angle modes of operation. IEEE Sensors 2010, pp. 864–867, 1–4 Nov 2010. doi:10.1109/ICSENS.2010.5690867. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5690867&isnumber=5689839
9. Acar, C., Shkel, A.M.: MEMS Vibratory Gyroscopes: Structural Approaches to Improve Robustness. Springer, New York/London (2009). ISBN ISBN 0387095357, 9780387095356
10. Zaman, M.F., Sharma, A., Zhili Hao., Ayazi, F.: A mode-matched silicon-yaw tuning-fork gyroscope with subdegree-per-hour Allan deviation bias instability. J. Microelectromech. Syst. **17**(6), 1526–1536. doi:10.1109/JMEMS.2008.2004794. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4636724&isnumber=4681904 (2008)
11. Sung, W.-K., Dalal, M., Ayazi, F.: A 3 MHz spoke gyroscope with wide bandwidth and large dynamic range. In: 2010 I.E. 23rd International Conference on Micro Electro Mechanical Systems (MEMS), Wanchai, 24–28 Jan 2010, pp. 104–107. doi:10.1109/MEMSYS.2010.5442554. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5442554&isnumber=5442276
12. Rozelle, D.: The hemispherical resonator gyro: from wineglass to the planets. In: Proceeding 19th AAS/AIAA Space Flight Mechanics Meeting, pp. 1157–1178 (2009)
13. Zotov, S.A., Prikhodko, I.P., Trusov, A.A., Shkel, A.M.: 3-D micromachined spherical shell resonators with integrated electromagnetic and electrostatic transducers. In: Solid-state sensors, actuators, and

microsystems workshop 2010, Hilton Head Island, South Carolina, 6–10 Jun 2010
14. Greiff, P., Boxenhorn, B., King, T., Niles, L.: Silicon monolithic micromechanical gyroscope. In: 1991 international conference on solid-state sensors and actuators. Digest of technical papers, TRANSDUCERS '91, San Francisco, 24–27 Jun 1991, pp. 966–968. doi:10.1109/SENSOR.1991.149051. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=149051&isnumber=3940
15. Yazdi, N., Ayazi, F., Najafi, K.: Micromachined inertial sensors. Proc. IEEE **86**(8), 1640–1659 (1998). doi:10.1109/5.704269. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=704269&isnumber=15216
16. Weinberg, M.S., Kourepenis, A.: Error sources in in-plane silicon tuning-fork MEMS gyroscopes. J. Microelectromech. Syst. **15**(3), 479–491 (2006). doi:10.1109/JMEMS.2006.876779. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1638473&isnumber=34353
17. Geen, J.A., Sherman, S.J., Chang, J.F., Lewis, S.R.: Single-chip surface micromachined integrated gyroscope with 50°/h Allan deviation. IEEE J. Solid-State Circuits **37**(12), 1860–1866 (2002). doi:10.1109/JSSC.2002.804345. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1088117&isnumber=23646
18. Neul, R., Gomez, U.-M., Kehr, K., Bauer, W., Classen, J., Doring, C., Esch, E., Gotz, S., Hauer, J., Kuhlmann, B., Lang, C., Veith, M., Willig, R.: Micromachined angular rate sensors for automotive applications. IEEE Sensor. J. **7**, 302–309 (2007). doi:10.1109/JSEN.2006.888610. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4066977&isnumber=4049782

H

Hardbake (HB)

► SU-8 Photoresist

Hardness

► Nanomechanical Properties of Nanostructures

Hard-Tip Soft-Spring Lithography

► Dip-Pen Nanolithography

Harnessing Disorder at the Nanoscale

Juan Sebastian Totero Gongora and
Andrea Fratalocchi
PRIMALIGHT, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Synonyms

Anderson localization; Coupled mode theory; Disordered plasmonics; Dissipative cavity; Product of random matrices; Random matrix theory

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Definition

In nanotechnology, any fabrication process naturally introduces disorder and randomness at the nanoscale. Disorder is typically unwanted in applications, as it is associated with unpredictable degrees of freedom that are difficult to control. However, it can be demonstrated that disorder supports fascinating and counterintuitive phenomena, which can be harnessed at the nanoscale for the realization of specific functionalities.

Introduction

It is a common perception that the presence of disorder influences negatively the properties of a physical system. Nevertheless, there are several cases in which disorder plays a positive role and can be exploited to develop novel applications. In general, the introduction of disorder significantly increases the complexity of the system under examination, but, as is often seen in nature, an increase in complexity also corresponds to a richer physical scenario. As a result, in recent years, great efforts have been devoted to the possibility of harnessing complexity and controlling disorder in optical systems at the nanoscale [1–3]. An important example is provided by the study of the disorder-induced localization phenomena. Starting from the initial works of Anderson on the insulation properties of a disordered conductor [4], researchers have been able to

generalize and study the occurrence of weak and strong localization in different physical contexts involving wave propagation in complex media. In photonics, the occurrence of Anderson localization has been observed and analyzed extensively by investigating the spectral properties of disordered photonic crystals [5, 6]. At the same time, localization phenomena play a key role in the presence of metal structures. It is well known that when an electromagnetic wave impinges on a metallic surface, it can excite a distribution of surface plasmon polaritons (SPP). By tuning the geometrical properties of the plasmonic structure, the SPP distribution can be tailored to produce a strong enhancement of the electromagnetic field, whose spatial extent can go far below the diffraction limit. The generation of highly localized hotspots plays an important role in many plasmonic applications and is usually achieved by considering symmetric and regular geometries (e.g., nano-tips). However, the investigation of the effects of disorder on a plasmonic system is of great interest, and in this respect, several studies have been conducted on the localization properties of fractal clusters, particles aggregates, and disordered surfaces [7, 8]. Due to the uncertainty produced by randomness and chaos, the study of these systems requires the development of rigorous theoretical models to treat disorder and complexity. In this entry, a disordered plasmonic is modeled by considering a set of interacting cavities. Differently from previous approaches, the modeling explicitly considers the presence of radiative and dissipative losses, hence providing a model which can be easily generalized to describe many different realistic nanoscale systems. The evolution of the system is investigated by deriving a set of coupled mode (CM) equations ruling the distribution and propagation of energy among the cavities. In the stationary regime, the CM equations can be rewritten as a recursive product of real-valued matrices by applying the transfer matrix formalism. In the presence of disorder, the transfer matrices of the system become random variables, and their general properties can be determined by means of random matrix theory (RMT). By definition, a random matrix (RM) $\mathbf{M}$ is a $D \times D$ matrix whose elements M_{ij} are random variables distributed according to a given probability distribution $\mu[\mathbf{M}]$. RMs were first introduced in the 1950s to interpret the fluctuations of the nuclear energy levels, but they have rapidly become a powerful investigation tool for random processes and disordered systems [9]. Furthermore, as in the case under examination, many physical phenomena and systems involve the introduction and manipulation of products of random matrices (PRM) [10]. The localization properties of a disordered system described by a PRM are usually characterized by introducing the Lyapunov characteristic exponent (LCE) γ_m of the PRM which, in analogy with dynamical systems, determines the exponential growth or decay of the PRM eigenfunctions and, consequently, of any vector evolving according to it. By applying the theory of PRMs, one can derive an analytic expression of γ_m for the relevant case of a binary distribution of different cavities. The calculation, which is performed by considering a suitable statistical mechanics approach, allows for the detection of an optimal amount of disorder maximizing the degree of localization of the system. The results from the theoretical model can be verified by resorting to numerical ab initio simulations, in which a chain of metallic nanoparticles under external illumination plays the role of a one-dimensional set of open cavities. In the nanoplasmonic system, disorder is introduced by varying the radius of a fixed number of spheres along the chain. The relative number of different spheres depends on a probability parameter p that controls the amount of disorder. The degree of localization in the system is determined by defining a suitable localization measure for the electromagnetic energy whose value is averaged over different realizations of disorder. In very good agreement with the theoretical prediction, the numerical results confirm the existence of an optimal value of disorder which maximizes the localization of energy in the system.

This entry is organized as follows. In section "Coupled Mode Theory of a One-Dimensional Chain of Disordered Cavities," the properties of a set of coupled open cavities are modeled by defining a set of CM equations, which in the stationary regime are rewritten as a product of

real-valued transfer matrices. In section "Random Matrix Theory: Analytic Estimation of the Localization Length," following a brief introduction to random matrix theory, the analytic expression for the LCE of the system is presented, which can be used to obtain the amount of disorder maximizing the localization properties of the system. In section "Nanoplasmonic Localization: Results from FDTD Simulations," the theoretical predictions are verified by performing a large set of massively parallel finite-differences in time-domain simulations on a chain of metallic nanoparticles, whose results are in excellent agreement with the RMT model.

Coupled Mode Theory of a One-Dimensional Chain of Disordered Cavities

A large number of nanoscale systems can be modeled by considering a random set of interacting open cavities. A cavity, or resonator, is a structure capable of storing and accumulating energy at a given resonance frequency ω [11]. The resonance frequency corresponds to a localized mode (i.e., a standing wave) of the electromagnetic field inside the cavity. The amplitude of the localized mode is described by a complex function $a_n(t)$, in such a way that the energy stored in the cavity corresponds to $|a(t)|^2$. A cavity which can continuously lose energy by means of radiative emission is called an open cavity. A dissipative cavity, conversely, depletes energy due to internal losses (e.g., metal absorption). Radiative and dissipative losses, as can be easily understood, limit the ability of storing energy inside the cavity. A quantitative measure of the latter is provided by the quality factor Q of the cavity, which is defined as

$$Q = \frac{\omega W}{P_d} \tag{1}$$

where W is the energy stored in the cavity and P_d is the power depleted by radiative and dissipative losses. Simple arguments show that the quality factor is indeed a measure of the characteristic decay time τ of the electromagnetic energy inside the cavity, which is defined as $\tau = 2Q/\omega$. It follows that the properties of the cavity are completely determined by the oscillating frequency ω and the decay time τ, which can be grouped by defining a complex characteristic frequency

$$\Omega = \omega + i\frac{1}{\tau} \tag{2}$$

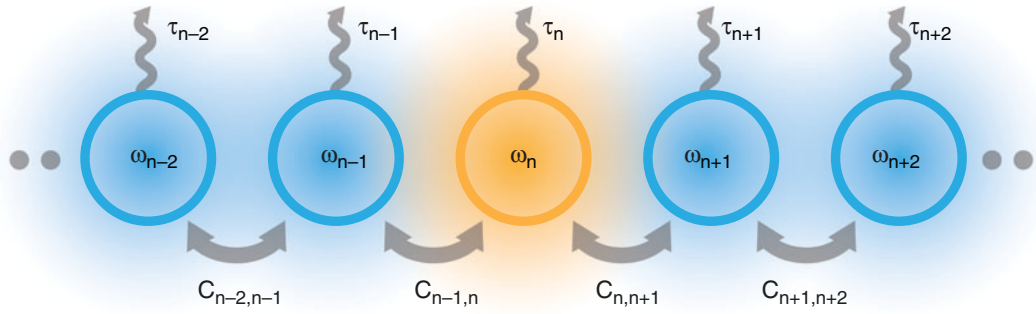
in order to simplify the notation. If one considers a set of distinct cavities with coupled resonance modes, electromagnetic energy can be further transferred among the single resonators. The exchange of energy among the cavities is described by introducing a set of coupling constants $C_{n,\,n'}$, with n, n' the indices identifying the interacting cavities. As a significant case study, one can consider the case of a one-dimensional chain of N-coupled cavities whose characteristic frequencies Ω_n are complex-valued random variables (Fig. 1). The system is chosen to be one-dimensional in order to provide a simpler notation, but all the results can be extended to higher dimensions with some algebraic effort [10, 12]. This model is particularly suitable to describe the distribution and localization of energy in a disordered plasmonic system, in which radiative and dissipative losses play an important role. Starting from an ordered chain of identical cavities, each one described by a characteristic frequency Ω_0, the presence of disorder can be introduced as a perturbation in the characteristic frequencies. As a result, the expression for the fluctuating frequencies reads

$$\Omega_n = \Omega_0 + \delta\Omega_n \tag{3}$$

where $\delta\Omega_n$ is a random variable distributed according to a given probability distribution $\mu[\delta\Omega]$. The evolution of the total system can be studied by applying coupled mode theory (CMT), which provides a set of evolution equations for the complex amplitudes of every resonator in the chain. By assuming nearest-neighbor interactions, CM equations read

$$\partial_t a_n = i\Omega_n a_n + iC_{n+1}a_{n+1} + iC_{n-1}a_{n-1}, \tag{4}$$

H



Harnessing Disorder at the Nanoscale, Fig. 1 Chain of open cavities. Every cavity is characterized by a different oscillation frequency ω_n and a different decay time τ_n. The exchange of energy along the chain is ruled by the coupling coefficients $C_{n,n'}$, which depend on the spatial coupling of the single-cavity modes

where $n \in [1, N]$ and

$$C_n = \begin{cases} 1 & n \in [1, N], \\ 0 & \text{otherwise}, \end{cases} \tag{4a}$$

C_n are coupling factors which are assumed to be constant along the chain, provided an opportune normalization of the CM equations allows for their inclusion in the complex frequencies Ω_n. Equation 4 can be rewritten in operatorial form by defining a state vector $(\mathbf{a}(t) = [a_1(t) \cdots a_N(t)]^\top$ and a tridiagonal operator $\mathbf{A}$:

$$\partial_t \mathbf{a}(t) = i\mathbf{A} \cdot \mathbf{a}(t)$$
$$\mathbf{A} = \begin{pmatrix} \Omega_1 & 1 & 0 & \cdots & 0 \\ 1 & \Omega_2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & \Omega_N \end{pmatrix} \tag{5}$$

Given an initial condition, e.g., external illumination, the electromagnetic energy distributes along the chain of nanocavities until the system reaches a stationary state, in which the external pumping is balanced by the losses of the system. In the stationary regime $\partial_t \mathbf{a}(t) = 0$, the dynamical problem expressed by Eq. 5 becomes an eigenvalue problem for the operator $\mathbf{A}$. However, due to the introduction of radiative and dissipative losses, the operator $\mathbf{A}$ is not Hermitian, and its eigenvalues $\boldsymbol{\varepsilon}$, which are solution of the eigenvalue equation $\mathbf{A} \cdot \mathrm{a} = \boldsymbol{\varepsilon}\mathbf{a}$, are complex variables. Without losing any generality, the eigenvalue problem can be written in recursive form as

$$a_{n+1} = (\boldsymbol{\varepsilon} - \Omega_n)a_n - a_{n-1} \tag{6}$$

which is analogous to a discretized Schrödinger equation in the presence of a random complex potential. Despite the simple formulation, the real version of Eq. 6 has been extensively used to investigate localization phenomena in various physical systems. Following the standard approach and by defining the vector $\phi_n = [a_n \;\; a_{n-1}]^\top$, Eq. 6 can be rewritten in the transfer matrix formalism as

$$\phi_{n+1} = \mathbf{C}_n \phi_n, \qquad \mathbf{C}_n = \begin{bmatrix} (\boldsymbol{\varepsilon} - \Omega_n) & -1 \\ 1 & 0 \end{bmatrix} \tag{7}$$

where $\mathbf{C}_n$ is a complex-valued transfer matrix. The advantage of the transfer matrix approach is that it allows for a direct control of the propagation of energy along the chain of resonators. By iterating Eq. 7, in fact, one obtains

$$\phi_{N+1} = \mathbf{Q}_N \phi_1, \qquad \mathbf{Q}_N = \prod_{n=1}^{N} \mathbf{C}_n \tag{8}$$

where the vector ϕ_1 contains the initial conditions of the system (e.g., the input signal) and where the total transfer matrix $\mathbf{Q}_N$ is simply the product of the single-transfer matrices. It follows immediately that in the presence of disorder, the propagation of energy along the chain is ruled by a product of matrices with random elements and can be investigated by means of random matrix theory. However, the general results of random

matrix theory hold for real-valued matrices. In order to overcome this difficulty and provided the matrices $\mathbf{C}_n$ are symplectic matrices, a real-valued model can be deducted from Eq. 7 by considering separately the real and imaginary parts of the cavity amplitudes

$$a_n(t) = p_n(t) + iq_n(t) \tag{9}$$

and rewriting Eq. 4 as a system of $2N$-coupled differential equations. In the stationary regime, the evolution of the system is ruled by the following set of coupled recursive equations:

$$\begin{cases} p_{n+1} = \sigma'_n p_n - \sigma''_n q_n - p_{n-1} \\ q_{n+1} = \sigma''_n p_n - \sigma'_n q_n - q_{n-1} \end{cases} \tag{10}$$

in which the complex variable $\sigma_n = \boldsymbol{\varepsilon} - \Omega_n$ was introduced to simplify the notation and σ'_n and σ''_n are the real and imaginary parts of σ_n, respectively. Using the transfer matrix formalism, Eq. 10 can be rewritten as

$$\begin{aligned} \psi_{n+1} &= \mathbf{T}_n \psi_n, \psi_n = \begin{bmatrix} p_n \\ q_n \\ p_{n-1} \\ q_{n-1} \end{bmatrix}, \\ \mathbf{T}_n &= \begin{bmatrix} \sigma'_n & -\sigma''_n & -1 & 0 \\ \sigma''_n & \sigma'_n & 0 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}. \end{aligned} \tag{11}$$

where $\mathbf{T}_n$ is a real-valued 4×4 matrix with $\det \mathbf{T}_n = 1$. In analogy with Eq. 8, the total transmitted field reads

$$\psi_N = \mathbf{P}_{N\psi 1} \tag{12}$$

where the transfer matrix of the system

$$\mathbf{P}_n = \left[\prod_{n=1}^{N-1} \mathbf{T}_n \right] \tag{13}$$

has determinant one and is again the product of all the single-transfer matrices. In order to obtain a better insight into the physics underlying the model, one can solve Eq. 10 in the absence of disorder. By considering $\delta\Omega_n \equiv 0$, Eq. 7 can be easily solved by looking for a solution of the form

$$p_n + iq_n = Ae^{-ikn} + Be^{ikn} \tag{14}$$

where $k = k' + ik''$ is a complex wave number and A and B are complex constants depending on the boundary conditions. By substituting Eq. 14 into Eq. 10 and solving, one obtains

$$\begin{cases} \sigma' = 2\cos k' \cosh k'' \\ \sigma'' = -2\sin k' \sinh k'' \end{cases} \tag{15}$$

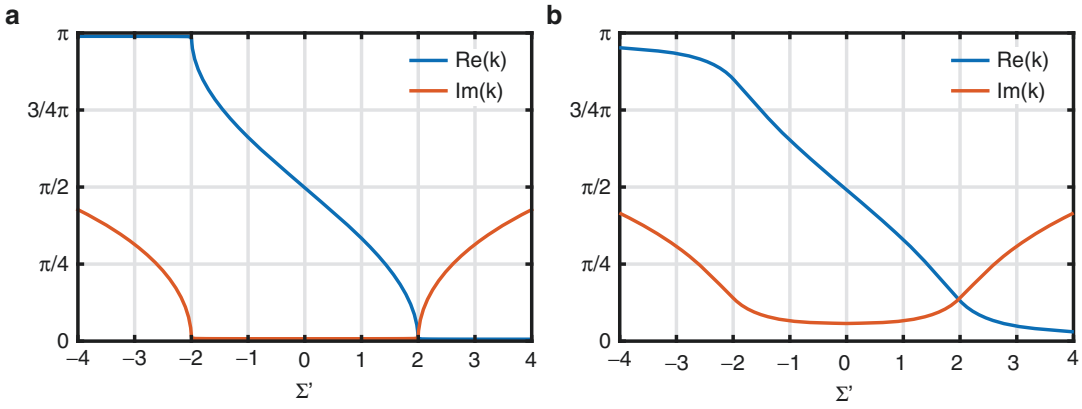
which corresponds to

$$\sigma' + i\sigma'' = 2\cos(k' + ik''). \tag{16}$$

Equation 16 relates the eigenvalues σ_n to the wavevector k of the corresponding eigenmodes. Therefore, it can be used to determine whether the solution associated with each value of σ_n is a propagating or exponentially localized solution. In the absence of any loss, σ_n is a purely real quantity, and, consequently, Eq. 16 has two kinds of solutions (see Fig. 2a). When $|\sigma'| \leq 2$, which is known as the allowed band of the system, the wave number k is a real quantity and the eigenmode of the total structure is a propagating wave. Conversely, when $|\sigma'| > 2$, the wave number k becomes a purely imaginary quantity, which corresponds to an exponentially attenuated eigenmode. As no propagation is possible, this condition defines the forbidden band of the structure. In the presence of losses, σ_n becomes a complex variable, whose imaginary part is always negative. The real part of σ_n is again divided into a forbidden and allowed region, but the presence of losses produces a solution which is exponentially attenuated also for values of σ'_n deep inside the allowed band $|\sigma'| \leq 2$ (see Fig. 2b).

Random Matrix Theory: Analytic Estimation of the Localization Length

In a disordered chain of cavities, the propagation and localization properties of the system depend



Harnessing Disorder at the Nanoscale, Fig. 2 Dispersion relation and energy bands of the system (**a**) in the absence of losses and (**b**) considering $\Sigma' + i\Sigma'' \equiv \sigma' + i\sigma''$ and $\sigma_n'' = -0.4$

on the amount of disorder introduced, which is directly related to the random frequency distribution $\mu[\delta\Omega]$. The transfer matrices $\mathbf{T}_n$ and the total product matrix $\mathbf{P}_N$ become random noncommutative variables, whose general properties can be studied by employing random matrix theory. As in the case of standard random variables, the global properties of a large set of RMs can be determined by means of limit theorems. In the case of a PRM, these limit theorems provide useful and universal estimates for the growth of the eigenfunctions of the final matrix and, consequently, for any state vector evolving according to it. In the presence of disorder and in analogy with chaotic dynamical systems, a generic vector $\psi_N = \mathbf{P}_N\psi_1$ will be related to the initial condition ψ_1 by an exponential growth of the form $|\psi_N| = |\psi_1|\exp(\widetilde{\gamma}N)$ when N is sufficiently large. The growth rate for a specific PRM is then given by

$$\widetilde{\gamma} = \frac{1}{N}\log\frac{|\psi_N|}{|\psi_1|} = \frac{1}{N}\log\mathrm{Tr}\mathbf{P}_N \tag{17}$$

and $\widetilde{\gamma}$ is called the Lyapunov characteristic exponent (LCE) of the PRM $\mathbf{P}_N$. Due to the random nature of the PRM, the LCE defined in Eq. 17 depends on the specific sequence of RMs composing $\mathbf{P}_N$. A general extension of Eq. 17 can be obtained by averaging over all the possible sequences of RMs and by taking the limit $N \to \infty$, namely,

$$\gamma_m \equiv \lim_{N\to\infty} \frac{1}{N}\langle\log\mathrm{Tr}\mathbf{P}_N\rangle \tag{18}$$

where γ_m is defined as the maximum LCE of the PRM and

$$\langle f[\mathbf{M_n}]\rangle = \int f[\mathbf{M}_n]\left(\prod_{n=0}^{N} \mathrm{d}\mu[\mathbf{M}_n]\right) \tag{19}$$

represents an average over the ensemble of all possible random sequences of matrices, i.e., over all the possible realizations of disorder. The properties of γ_m naturally descend from the properties of the RMs composing the product (e.g., domain of definition, sign, symmetric properties), and a general analytic expression of γ_m is known only for a limited number of RM classes. As mentioned earlier, however, if the random matrices are independent and identically distributed and their probability meets some regularity requirements, the existence conditions and the properties of γ_m can be determined by means of limit theorems. More specifically, the existence of a maximum LCE is ensured by the Furstenberg-Kesten theorem, which in the case of real random matrices with determinant, one assures that the maximum LCE always exists and is a positive nonrandom quantity, namely,

$$\gamma_m = \lim_{N\to\infty} \frac{1}{N} \langle \log \mathrm{Tr}(\mathbf{P}_N \psi) \rangle > 0 \qquad (20)$$

where ψ is a generic vector [13]. Moreover, it can be shown that the average over disorder in Eq. 20 can be performed in any statistical ensemble and γ_m is a self-averaging quantity. As already mentioned, an analytical calculation of the maximum LCE is feasible only in a limited number of cases, and generally one has to fall back to approximate methods or numerical estimates of γ_m. As an example of an exactly solvable system, one can consider the case of a finite sequence of identical matrices, which in the theoretical model corresponds to an ordered chain of identical cavities. In this case, the total transfer matrix of the system assumes the simple form $\mathbf{P}_N = \mathbf{T}^N$, and a direct computation of the LCE yields

$$\lim_{N\to\infty} \frac{1}{N} \log \mathrm{Tr} \mathbf{T}^N = \log |\mu_1| \equiv \gamma_m \qquad (21)$$

where μ_1 is the largest nondegenerate eigenvalue of the matrix $\mathbf{T}$. The proof of Eq. 21 is straightforward: as the matrix $\mathbf{T}$ is invertible and diagonalizable, it can be decomposed as $\mathbf{T} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ where $\mathbf{D}$ is the diagonal eigenvalue matrix $\mathbf{D} = \mathrm{diag}[\mu_1, \mu_2, \mu_3, \mu_4]$ and $\mathbf{P}$ is the eigenvector matrix. Taking the n-th power yields

$$\mathbf{T}^N = \mathbf{P}\mathbf{D}\left(\mathbf{P}^{-1}\mathbf{P}\right)\mathbf{D} \ldots \mathbf{D}\mathbf{P}^{-1} = \mathbf{P}\mathbf{D}^N\mathbf{P}^{-1} \qquad (22)$$

which in the limit of large N can be rewritten as $\mathbf{T}^N \to |\mu_1|^N \tilde{\mathbf{D}}$ where $\tilde{\mathbf{D}}$ is an $O(1)$ matrix and from which trivially follows Eq. 21. Once disorder is introduced in the system, however, the equality in Eq. 20 holds only in the limit $N \to \infty$, i.e., in the case of an infinite sequence of RMs. Hence, the LCE γ_1 is a self-averaging quantity only in the thermodynamical limit, and in the case of finite sequences, the ratio $\log \mathrm{Tr}\, \mathbf{P}_n$ fluctuates around γ_1, with fluctuations which are in general related to the statistical ensemble of sequences under consideration. It is useful in these cases to define a set of generalized Lyapunov characteristic coefficients of order k, whose expressions read

$$\mathcal{G}(k) = \lim_{N\to\infty} \frac{1}{N} \log \left\langle \mathrm{Tr}(\mathbf{P}_N)^k \right\rangle \qquad (23)$$

and which are related to the maximum LCE by $\gamma_m = \partial_q \mathcal{G}(q)_{q=0}$. Equation 23 is an example of an annealed average, in contrast with the quenched average from Eq. 20 [14]. From a strict mathematical point of view, annealed averages are easier to perform, as they do not require the averaging of the logarithmic function. The problem of dealing with logarithmic averages is a well-known issue in statistical mechanics of disordered systems, and in the recent years, a wide number of theoretical approaches to their computation have been developed. These methods, which relate the generalized LCE $\mathcal{G}(k)$ to the maximum LCE of the PRM, rely on the approximation of the logarithmic function by either expanding it into a power series or considering limit identities [10]. The most immediate approach, known as the weak disorder expansion, is based on the introduction of a multiplicative control parameter ε for the fluctuating frequencies in Eq. 3. As a result, the random transfer matrices assume the form

$$\mathbf{T}_n = \mathbf{T}_0 + \varepsilon \tilde{\mathbf{T}}_n \qquad (24)$$

where $\mathbf{T}_0$ is the transfer matrix of the ordered system and $\tilde{\mathbf{T}}_n$ is a matrix containing the fluctuating frequencies. When ε is small, the total PRM and consequently the logarithm in Eq. 20 can be expressed in terms of powers of ε, and the maximum LCE can be computed up to the desired order in ε. Another standard approach to the logarithmic average is based on the replica trick, which in the random matrix case is expressed by the identity

$$\left\langle \mathrm{Tr}(\mathbf{P}_N)^k \right\rangle = \mathrm{Tr} + \left\langle (\mathbf{P}_N)^{\otimes k} \right\rangle = \mathrm{Tr}\left[\overline{\mathbf{T}_n^{\otimes k}}\right]^N \qquad (25)$$

where $\mathbf{A}^{\otimes k} = \mathbf{A} \bigotimes \mathbf{A} \cdots \bigotimes \mathbf{A}$ represents the k-th tensor power of the matrix $\mathbf{A}$ and $\overline{\cdots}$ is the average over the probability distribution $\mu[\delta\Omega_n]$. In Eq. 25, due to the direct average of the tensor product matrix, the matrix $\overline{\mathbf{T}_n^{\otimes k}}$ is a nonrandom

quantity, and by referring to Eq. 18, the generalized LCE from Eq. 23 reads

$$\mathcal{G}(k) = \log\mu_{\otimes} \tag{26}$$

where $\mu_{\otimes}$ is the largest eigenvalue in modulus of the tensor matrix $\overline{\mathbf{T}_n^{\otimes k}}$. Apart from these methods, a very good estimate of the maximum LCE can be obtained by employing the microcanonical method, which can be easily described by considering a set of RMs characterized by a binomial probability distribution:

$$\mathbf{T}_n = \begin{cases} \mathbf{T}_A & \text{with probability } p \\ \mathbf{T}_B & \text{with probability } 1-p \end{cases} \tag{27}$$

where $p \in [0,1]$. This case is very relevant because in the majority of the experiments, the system is composed of a random aggregation of two materials, such as in the case of metallic scatterers dispersed in a host medium. The results for the binomial distribution, moreover, can be easily generalized to a generic continuous probability distribution, as shown in [12]. In a random sequence of N matrices extracted with probability given by Eq. 27, N_A and N_B denote the number of occurrences of the matrices **A** and **B**, respectively. In analogy with the statistical mechanics formalism, if N_A and N_B are fixed across the different disorder realizations, then the statistical ensemble is microcanonical like, while on the contrary, if the number of occurrences is variable, it is canonical like. The microcanonical ensemble exhibits some crucial advantages over the canonical. First of all, it can be shown that the finite N fluctuations are significantly reduced. Second, in the microcanonical ensemble, the annealed average can be computed exactly by means of the binomial formula:

$$(\mathbf{T}_A + x\mathbf{T}_B) = \sum_{k=0}^{N} \binom{N}{k} \mathbf{T}_A^k (x\mathbf{T}_B)^{N-k} \tag{28}$$

where $x \in \mathbb{R}$. The product $\mathbf{P}_N$, then, is expressed by the derivative

$$\frac{\partial^{N_B}}{\partial x^{N_B}}(\mathbf{T}_A + x\mathbf{T}_B)|_{x=0} = \binom{N}{N_B} \mathbf{T}_A^{N_A}\mathbf{T}_A^{N_B} \tag{29}$$

and this allows to write the annealed average in the microcanonical ensemble as

$$\langle \mathrm{Tr}\mathbf{P}_N \rangle_m = \binom{N}{N_B}^{-1} \frac{1}{N_B!} \mathrm{Tr}\left(\frac{\partial^{N_B}}{\partial x^{N_B}}(\mathbf{T}_A + x\mathbf{T}_B)\Big|_{x=0} \right) \tag{30}$$

which follows from the commutativity of the trace operator. Provided all matrices Tn are diagonalizable, Eq. 30 can be simplified by considering

$$(\mathbf{T}_A + x\mathbf{T}_B) = \mathbf{P}\mathbf{D}_1(x)\mathbf{P}^{-1} \tag{31}$$

where $\mathbf{D}_1(x) = \mathrm{diag}[v_1(x),\ v_2(x),\ v_3(x),\ v_4(x)]$ is the diagonal matrix containing the eigenvalues $v_i\ (x)$ of the matrix $(\mathbf{T}_A + x\mathbf{T}_B)$. From Eq. 31 follows immediately (cf. Eq. 22)

$$(\mathbf{T}_A + x\mathbf{T}_B)^N = \mathbf{P}\mathbf{D}_1(x)^N\mathbf{P}^{-1} \overset{N\to\infty}{\to} v_M(x)^N\tilde{D}(x) \tag{32}$$

where v_M is the maximum eigenvalue of $(\mathbf{T}_A + x\,\mathbf{T}_B)$ and $\tilde{\mathbf{D}}(x)$ is again a $O(1)$ matrix. By substituting Eq. 32, taking the analytic continuation of x, and applying the Cauchy integral formula

$$\left.\frac{\partial^\alpha f(z)}{\partial x^\alpha}\right|_{z=z^*} = \frac{\alpha!}{2\pi i}\oint_c \mathrm{d}z \frac{f(z)}{(z-z^*)\alpha+1} \tag{33}$$

where $\mathcal{C}$ is a closed contour surrounding $z = 0$, the annealed average reads

$$\langle \mathrm{Tr}\mathbf{P}_N \rangle_m = \binom{N}{N_B}^{-1} \mathrm{Tr}\left(\oint_c \frac{\mathrm{d}z}{z}\tilde{D}(x)\exp\left[N\log\frac{v_M(x)}{z^{(1-p)}}\right] \right) \tag{34}$$

Due to the presence of the exponential function, in the limit $N \to \infty$, the complex contour integral can be evaluated by means of the stationary phase method:

$$Z = \lim_{N\to\infty} \oint_c \mathrm{d}z\ f(z)e^{NS(z)} \propto e^{NS(w)} \tag{35}$$

where w is the solution of the saddle point equation $\partial_z S(z) = 0$. In the case of Eq. 34, the final expression for the annealed average reads

$$\langle \mathrm{Tr}\mathbf{P_N}\rangle_m \propto \binom{N}{N_B} \exp[N\log v_M(w) - (1-p)\log w] \tag{36}$$

where, again, w_p is the solution of

$$\left.\frac{\partial \log v_M(z)}{\partial \log z}\right|_{z=w_p} = (1-p). \tag{37}$$

Substituting into Eq. 20, one finally obtains

$$\widetilde{\gamma}_m = \lim_{N\to\infty}\frac{1}{N} = (1-p)\log(1-p) + p\log(p) + \log v_M(w_p) - (1-p)\log w_p \tag{38}$$

where the binomial coefficient is expressed using the Stirling approximation formula. As shown in [12], the LCE of the starting PRM is very close to the real part of $\widetilde{\gamma}_m$. Hence, Eq. 38 allows for a precise calculation of the LCE in the microcanonical ensemble. More than that, the availability of an analytic estimate for γ_m opens up interesting perspectives. As an example, Eq. 38 can be used to determine the optimal degree of disorder for the localization of the system. Once the structure of the matrices $\mathbf{T}_A$ and $\mathbf{T}_B$ has been determined, the amount of disorder is controlled by the probability p, with $p = 0$ corresponding to N matrices of type $\mathbf{T}_B$ and $p = 1$ corresponding to N matrices of type $\mathbf{T}_A$. By finding the zeros of the derivative

$$\frac{\partial \widetilde{\gamma}_m}{\partial p} = 2\tan^{-1}(1-2p) - \log\omega_p, \tag{39}$$

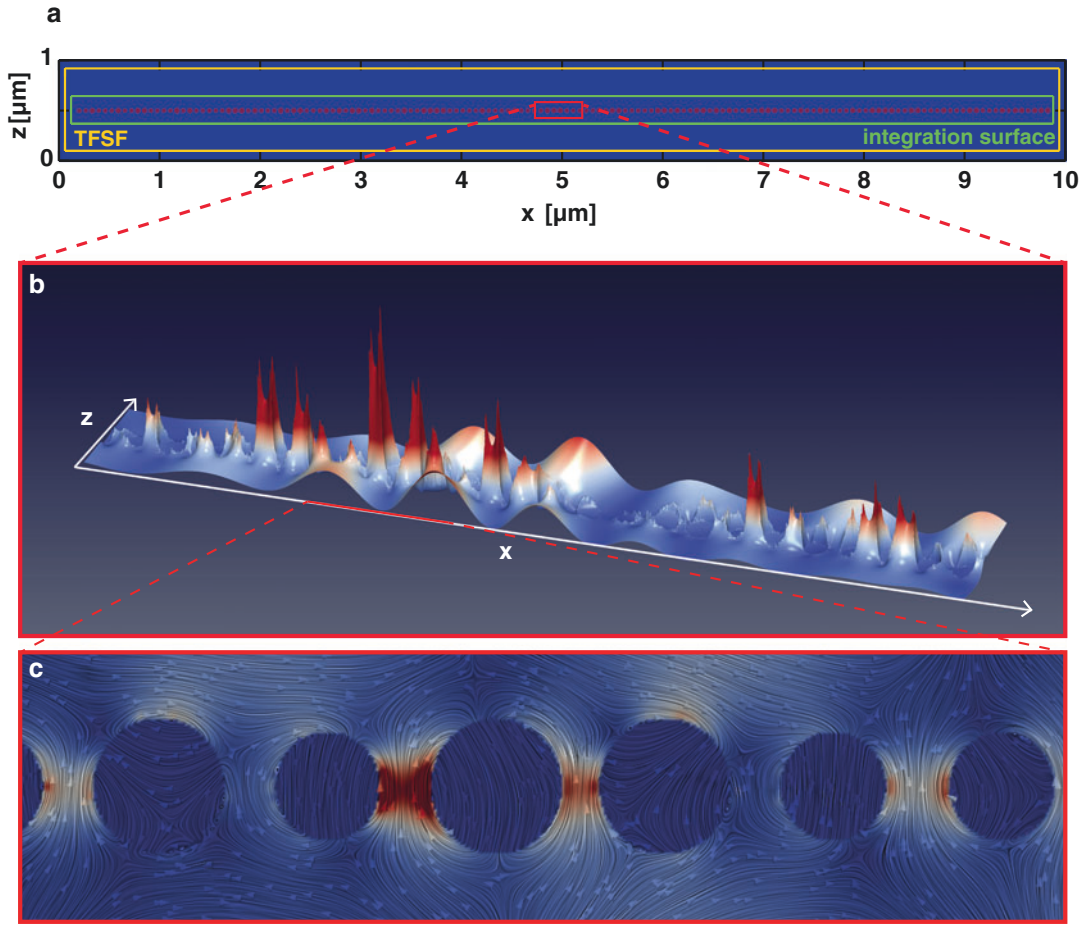
one obtains that in the case of a binomial distribution, the LCE γ_m has a single minimum for $p = p_c$. As the minimum in the LCE corresponds to a maximum in the localization length $l_c = \gamma_m^{-1}$, the system will exhibit the maximum degree of localization at the critical probability [15].

Nanoplasmonic Localization: Results from FDTD Simulations

The propagation of energy across a stationary one-dimensional set of open cavities can be simulated by considering a chain of metallic nanoparticles under external illumination. The transfer matrices for this system can be obtained by assimilating every nanoparticle with a hertzian dipole. By expressing the dipole equations in the Fourier domain, the amplitudes a_n of the elements of the chain evolve according to Eq. 7 with a transfer matrix of the form

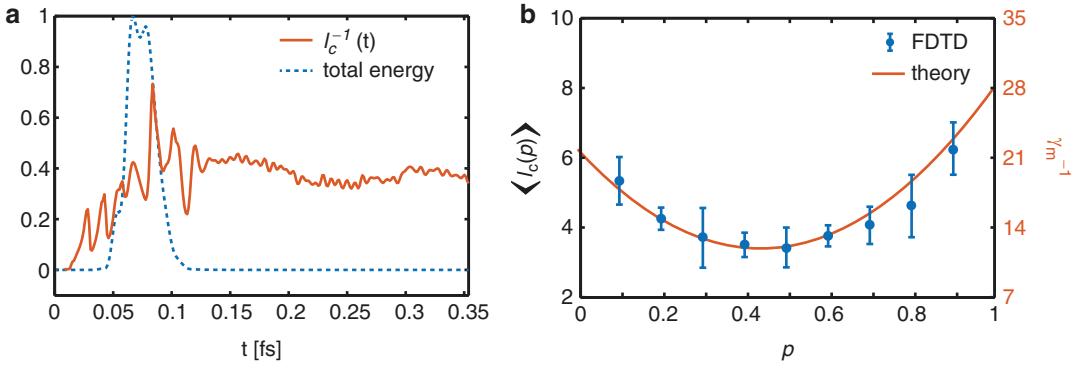
$$\mathbf{C}_n = \begin{bmatrix} \frac{1}{\gamma\alpha_n^2}\left(\omega^2 - \omega_0^2 - i\left[\frac{\omega}{\tau_e} - \frac{\omega^3}{\tau_R\omega_0}\right]\right) & -1 \\ 1 & 0 \end{bmatrix} \tag{40}$$

where ω is the oscillation frequency of the dipoles, ω_0 is the plasma frequency of the metal, $\tau_{e,R}$ are the electronic and radiative decay times, γ is a polarization-dependent term, and α_n are coupling constants depending on the relative position of the nanoparticles [16, 17]. The evolution of the electromagnetic fields in such system can be rigorously investigated by employing the finite-differences in time-domain (FDTD) algorithm [18]. The FDTD algorithm allows for the solution of Maxwell's equations simultaneously in space and time with arbitrary precision. Due to the presence of the curl operator in Maxwell's equations, the components of the electric **E** and magnetic **H** fields are staggered on a specifically designed Yee grid, and the computation proceeds in time according to a leapfrog scheme. The simulations included in this entry have been performed using the massively parallel FDTD simulator NANOCPP, which has proved to be able to scale up to hundreds of thousands of processors [2, 19]. The computational domain was organized as follows. The computational grid was set as a two-dimensional (x, z) box with dimensions 10 × 1 μm and discretized on a 10,000 × 1000 regular mesh (resolution 1 nm) with UPML boundary conditions simulating an open space (Fig. 3a). A one-dimensional chain of



Harnessing Disorder at the Nanoscale, Fig. 3 FDTD simulation of a disordered chain of metallic nanoparticles. (**a**) Computational domain. The *yellow line* represents the total-field region of the TFSF source. The *green line* represents the surface over which the inverse participation ratio $\langle l_c \rangle^{-1}$ is computed. (**b**) Electromagnetic energy distribution in proximity of a localized region. (**c**) Poynting vector distribution in a subregion of (**b**). The streamlines have been computed by applying the line integral convolution, and the energy distribution is plotted for the sake of completeness

N equidistributed Ag nanoparticles was placed along the axis $z = 0.5$ μm, and the metallic nanochain was illuminated on the long side by an impinging plane wave, represented by means of the total-field/scattered-field formalism (yellow box), and whose polarization was chosen to be parallel to the axis of the chain (TM polarization). As requested by the microcanonical hypothesis, for every value of the probability p, the nanochain was formed as a random permutation of $N_A = pN$ nanoparticles of radius 25 nm and $N_B = N - N_A$ nanoparticles of radius 20 nm. In order to excite all the plasmonic resonances of the metallic nanospheres, the waveform of the incident wave was taken as a Gaussian pulse with waist $\Delta\lambda = 200$ nm and centered in $\lambda = 400$ nm. Due to the presence of a broadband excitation, the permittivity of the Ag nanoparticles was represented using a broadband critical-point (CP) model with a set of four complex-conjugate Lorentz-Drude poles, which is implemented in the FDTD code using an unconditionally stable auxiliary equation algorithm [20, 21]. In a typical



Harnessing Disorder at the Nanoscale, Fig. 4 Inverse participation ratio from FDTD simulations. (**a**) Single realization with $p = 0.4$. The IPR (*orange line*) tends to a stationary value once the external excitation leaves the system. The total energy (*blue line*) is reported for the sake of completeness. (**b**) Localization length as a function of the amount of disorder p. For each value, 20 different realizations of disorder were considered. The FDTD results (*blue dots*) are in strong agreement with the theoretical results from Eqs. 38 to 40 (*orange line*)

H

simulation, the incident wave travels across the computational domain, exciting the plasmonic resonances of the nanoparticles composing the metallic chain, which behave as resonant cavities. As the incident wave leaves the computational domain, the excited modes of the nanoparticles start interacting. A first insight on the dynamics of the system can be obtained by computing the electromagnetic energy:

$$\mathcal{E} = \int dx \ \varepsilon \mathbf{E}^2 + \mu \mathbf{H}^2. \tag{41}$$

Figure 3b shows the typical distribution of energy in a small region of a disordered chain with $p = 0.4$. As can be easily evinced from the figure, part of the energy is localized in proximity of the nanospheres, while a large amount of energy rapidly leaves the system due to radiative emission. In the presence of disorder, however, the system tends to create highly localized hotspots. The transfer of energy can be analyzed by computing the Poynting vector $\mathbf{S} = \mathbf{E} \times \mathbf{H}$, whose distribution in proximity of a localized hotspot is shown in Fig. 3c. The degree of localization of the system can be measured by computing the normalized inverse participation ratio $\langle l_c \rangle^{-1}$ (IPR), which is inversely proportional to the localization length l_c and which is defined as

$$\langle l_c(p) \rangle = \left\langle \frac{1}{V} \frac{\left(\int \mathrm{d}x \ \mathcal{E} \right)^2}{\int \mathrm{d}x \ \mathcal{E}^2} \right\rangle_p \tag{42}$$

where $\langle \cdots \rangle_p$ represents an average over different realizations of the sequence at fixed p [9]. The integral in Eq. 41 is performed over a surface containing the nanochain (Fig. 3a, green box). Figure 4a reports the typical evolution of the localization length (orange line) and of the total energy in the integration surface (blue line) for a disordered chain of $N = 150$ nanoparticles with $p = 0.4$. As can be evinced from the total energy evolution, the incident wave enters the integration region at $t = 0.05$ fs and leaves it at $t = 0.1$ fs. In the presence of the external excitation, the IPR has strong oscillations related to the excitation of the plasmon modes on the spheres. Once the incident radiation leaves the integrating region, however, the IPR converges to a stable value. The final IPR for the single realization can then be computed by averaging over time. In order to verify the presence of a single minimum in the localization length of the disordered chain, a large set of massively parallel simulations was performed for different values of the probability p. The results of the simulation campaign are shown in Fig. 4. In order to compute the average IPR, 20 different realizations of disorder were considered for each

value of the probability p. In order to verify the theoretical model, the maximum LCE of the system (orange line) was computed using Eqs. 38 and 40 and considering the parameters from [16]. Remarkably, the FDTD results are in strong agreement with the theoretical model, and in both cases, the localization length $\langle l_c \rangle^{-1}$ exhibits a single minimum in proximity of the critical probability $p_c = 0.4$. It is worthwhile remarking that n does not contain any fit parameter and the theoretical calculation of γ_m depends only on the physical parameters of the system.

Conclusions

In this work, it was shown how a generic nanoscale system can be studied by considering a set of interacting open cavities, whose evolution is expressed in terms of a real-valued transfer matrix formalism. In the presence of disorder, the localization properties of the system can be evaluated by resorting to random matrix theory. By employing the microcanonical approach, an analytic expression for the localization length of a binary set of cavities was obtained as a function of the amount of disorder. Such analytic expression allowed to obtain an optimal degree of disorder which provides the higher degree of localization in the system. The existence of such an optimal degree of disorder has been verified by simulating a one-dimensional chain of metallic nanoparticles using the FDTD algorithm. This system, which can be described exactly with the transfer matrix approach, is analogous to a one-dimensional chain of open cavities, and the localization length can be estimated by computing the IPR of the system. The numerical results, which show a very good agreement with the theoretical predictions, represent an interesting example of how an accurate theoretical model, coupled with a rigorous treatment of the disorder in the system, can provide quantitative insights into the localization properties of a nanoplasmonic system. The existence of an optimal degree of disorder, moreover, opens up interesting perspectives for any plasmonic application requiring highly localized hotspot distributions.

Cross-References

- ▶ Finite-Difference Time-Domain Technique
- ▶ Light Localization for Nano-optical Devices
- ▶ Metal Nanoparticles from First Principles
- ▶ Nanofocusing of Light
- ▶ Nanoparticles
- ▶ Nanostructures for Energy
- ▶ Nanostructures for Photonics

References

1. Fratalocchi, A., Conti, C., Ruocco, G.: Three-dimensional ab-initio investigation of light-matter interaction in mie lasers. Phys. Rev. A. **78**, 013806 (2008). doi:10.1103/Phys-reva.78.013806
2. Liu, C., Di Falco, A., Molinari, D., Khan, Y., Ooi, B.S., Krauss, T.F., Fratalocchi, A.: Enhanced energy storage in chaotic optical resonators. Nat. Photonics **7**, 474 (2013)
3. Liu, C., Di Falco, A., Fratalocchi, A.: Dicke phase transition with multiple superradiant states in quantum chaotic resonators. Phys. Rev. X **4**, 021048 (2014)
4. Anderson, P.W.: Absence of diffusion in certain random lattices. Phys. Rev. **109**, 1492 (1958)
5. Wiersma, D.S., Bartolini, P., Lagendijk, A., Righini, R.: Localization of light in a disordered medium. Nature **390**, 671 (1997)
6. Conti, C., Fratalocchi, A.: Dynamic light diffusion, three-dimensional Anderson localization and lasing in inverted opals. Nat. Phys. **4**, 794 (2008)
7. Shalaev, V.M.: Nonlinear Optics of Random Media – Fractal Composites and Metal-Dielectric Films. Springer Tracts in Modern Physics, vol. 158. Springer, Berlin (2000)
8. Shahbazyan, T.V., Stockman, M.I.: Plasmonics: Theory and Applications. Springer, Dordrecht (2014)
9. Haake, F.: Quantum Signatures of Chaos. Springer Series in Synergetics, vol. 54, 3rd edn. Springer, Berlin (2010)
10. Crisanti, A., Paladin, G., Vulpiani, A.: Products of Random Matrices: In Statistical Physics, Softcover Reprint of the Original 1st ed. 1993 edition ed. Springer S.l, (2012)
11. Haus, H.A.: Waves and Fields in Optoelectronics. Prentice Hall, Englewood Cliffs (1984)
12. Deutsch, J.M., Paladin, G.: Product of random matrices in a microcanonical ensemble. Phys. Rev. Lett. **62**, 695 (1989)
13. Comtet, A., Texier, C., Tourigny, Y.: Lyapunov exponents, one-dimensional Anderson localization and products of random matrices. J. Phys. A **46**, 254003 (2013)
14. Mezard, M., Parisi, G., Virasoro, M.: Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications. World Scientific Lecture

Notes in Physics, vol. 9. World Scientific Press, Teaneck (1986)
15. Molinari, D., Fratalocchi, A.: Route to strong localization of light: the role of disorder. Opt. Express **20**, 18156 (2012)
16. Brongersma, M.L., Hartman, J.W., Atwater, H.A.: Electromagnetic energy transfer and switching in nanoparticle chain arrays below the diffraction limit. Phys. Rev. B **62**, R16356 (2000)
17. Dal Negro, L., Feng, N.-N.: Spectral gaps and mode localization in Fibonacci chains of metal nanoparticles. Opt. Express **15**, 14396 (2007)
18. Taflove, A., Hagness, S.C.: Computational Electrodynamics: The Finite-difference Time-domain Method. Artech House, Boston (2005)
19. Liu, C., van der Wel, R.E.C., Rotenberg, N., Kuipers, L., Krauss, T.F., Di Falco, A., Fratalocchi, A.: Model dispersive media in finite-difference time-domain method with complex-conjugate pole- residue pairs. Nat. Phys. **11**, 358 (2015)
20. Han, M.H., Dutton, R.W., Fan, S.H.: IEEE Microwave Wireless Compon. Lett. **16**, 119 (2006)
21. Totero Gongora, J.S., Fratalocchi, A.: Ab-initio techniques for light matter interaction at the nanoscale. In Computational Chemistry Methodology in Structural Biology and Material Sciences, Apple Academic Press, Oakville (to appear, expected Dec. 2015)

Heat

▶ Thermal Cancer Ablation Therapies Using Nanoparticles

Heat Capacity

▶ Nanocalorimetry

Heat Conduction

▶ Heat Transfer in Semiconductor Nanostructures
▶ Thermal Conductivity and Phonon Transport

Heat Conductivity

▶ Thermal Conductivity and Phonon Transport

Heat Removal

▶ Microchannel Flows as Heat Sinks

Heat Transfer Enhancement

▶ Microchannel Flows as Heat Sinks

Heat Transfer in Semiconductor Nanostructures

Jaeho Lee
Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

Synonyms

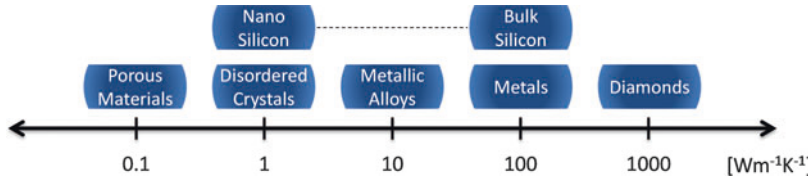
Heat conduction; Nanowires; Phonon transport; Phononic crystals; Silicon thin films; Thermal conductivity

Definition

Thermal conduction is heat transfer by microscopic motions of matter. Phonon is a quantized form of atomic vibrations, and phonon transport is responsible for thermal conduction in crystalline materials such as silicon.

Overview (Physical Principles)

Heat naturally flows from hot to cold, as governed by the Second Law of Thermodynamics. The heat flow by microscopic motions in solids is known as thermal conduction, which is described by the Fourier Law:



Heat Transfer in Semiconductor Nanostructures, Fig. 1 The thermal conductivity ranges about four orders of magnitude for common solid materials. Silicon may have about two orders of magnitude change in the thermal conductivity from bulk to nanostructures. In other words, silicon can be considered as a good thermal conductor or a good thermal insulator depending on the size

$$q = -k\nabla T \tag{1}$$

where q is the heat flux [W/m^2], k is the thermal conductivity [W/mK], and T is the temperature [K]. The heat flux is the rate of thermal energy passing through a unit area that is perpendicular to the heat flow direction. The thermal conductivity is a fundamental material property, which represents material's ability of transferring heat by conduction. The Fourier Law indicates that the thermal conductivity governs heat flows and temperature variations in solids. Thermal conduction in solid materials is normally carried by electrons and phonons, and the thermal conductivity can be expressed as

$$k = k_e + k_p \tag{2}$$

where k_e is the electron thermal conductivity and k_p is the phonon thermal conductivity. The electron thermal conductivity represents heat transport by electrons and is important in metals, semimetals, and degenerate semiconductors. Since the same electrons carry both electricity and heat, the electron thermal conductivity is proportional to the electrical conductivity according to the Wiedemann-Franz Law:

$$k_e = \sigma LT \tag{3}$$

where σ is the electrical conductivity [S/m], L is the Lorenz number [W/S-K^2], and T is the temperature [K]. The Wiedemann-Franz Law allows thermal conduction studies in metallic materials to take advantage of rich information known about electron transport. Phonon transport physics is relatively not well understood but responsible for thermal conduction in most semiconductors and insulators. With focus on thermal conduction in silicon nanostructures, the thermal conductivity in this entry strictly refers to the phonon thermal conductivity. The thermal conductivity can be simply expressed as

$$k = \frac{1}{3}Cv\Lambda \tag{4}$$

where C is the heat capacity [J/m^3-K], v is the group velocity [m/s], and Λ is the mean free path [m]. The heat capacity describes amount of thermal energy carriers per unit volume, and the group velocity describes how fast they are traveling. The one-third factor accounts for the average group velocity in three-dimensional geometric space. The mean free path is the average distance the thermal energy carriers travel before collision. The thermal conductivity is an intrinsic material property, which is independent of geometry, but this is only true when the dimensions are greater than the mean free path. Nanomaterials or bulk materials with nanoscale features that are comparable or smaller than the mean free path present strong size dependence. While the thermal conductivities of solid materials range about four orders of magnitude, silicon may have about two orders of magnitude change in the thermal conductivity from bulk to nanostructures (Fig. 1).

Thermal Conductivity Derivation Using Kinetic Theory

The thermal conductivity expression Eq. 4 is derived from the kinetic theory of gases.

Assuming ideal gas particles under a temperature gradient, the particle flux in the positive x-direction (ϕ_x) can be expressed as

$$\phi_x = \frac{1}{2} n v_x \tag{5}$$

where n is the particle density, v_x is the average velocity component, and the factor 1/2 accounts for the half of particles traveling in the negative x-direction. The particle flux can be converted to energy flux by accounting for the particle energy. As a particle travels from a position with the temperature T to another position with the temperature $T + \Delta T$, the particle absorbs the energy $c\Delta T$ where c is the heat capacity per particle. The energy flux (q) is a product of the particle flux and the energy absorption per particle. For particles traveling the opposite direction, from $T + \Delta T$ to T, the same amount of energy is released:

$$q_{x(T\to T+\Delta T)} = \frac{1}{2} n v_x c \Delta T \tag{6a}$$

$$q_{x(T+\Delta T\to T)} = -\frac{1}{2} n v_x c\Delta T \tag{6b}$$

The particles fluxes are identical in both directions, and there is no net flux of particles. There is a net flux of energy by particles under the temperature gradient, and the energy flux in the direction of heat flow can be expressed as

$$q_x = q_{x(T+\Delta T\to T)} - q_{x(T\to T+\Delta T)} = -n v_x c\Delta T \tag{7}$$

Assuming the particles thermalize and experience the temperature change (ΔT) through collisions over the average distance (Λ), the temperature change can be expressed as

$$\Delta T = \Lambda \frac{dT}{dx} \tag{8}$$

where the average distance between collisions is known as the mean free path. For an isotropic case, where the average velocity in all directions has the same speed, the group velocity can be expressed as

$$v^2 = v_x^2 + v_y^2 + v_z^2 = 3v_x^2 \tag{9}$$

where v_x, v_y, and v_z are the average velocity of particles in three-dimensional space. With Eqs. 8 and 9 and using the volumetric heat capacity term ($C = nc$), the energy flux expression takes the form:

$$q_x = -\frac{1}{3} C v \Lambda \frac{dT}{dx} \tag{10}$$

Based on the Fourier Law in Eq. 1, the simple thermal conductivity expression can be derived from the kinetic theory:

$$k = -\frac{1}{3} C v \Lambda \tag{11}$$

Thermal Conductivity Derivation Using Boltzmann Transport Equation

While the simple thermal conductivity model may provide essence of thermal conduction in bulk materials, thermal conduction in nanostructures requires rigorous thermal conductivity modeling based on the Boltzmann transport equation [1–5]. The Boltzmann transport equation (BTE) considers phonons as particles with energy and momentum. The BTE model describes the conservation of the number of particles using a distribution function $f(x,p,t)$, which is a function of geometric space (x), momentum space (p), and time (t). The general form of the BTE can be expressed as

$$\frac{\partial f}{\partial t} + v \cdot \nabla_x f + \frac{F}{m} \cdot \nabla_p f = \left(\frac{\partial f}{\partial t}\right)_{\text{collision}} \tag{12}$$

where the first term on the left side of the equation represents the total time rate of change of number of particles, the second term represents particle diffusion in geometric space due to phonon group velocity (v), the third term represents acceleration in momentum space due to external forces (F), and the right side of the equation represents

time rate of change of number of particles due to scattering. The first and the third terms can be dropped out in general heat transfer problems by assuming steady-state problems with no external force. The scattering term on the right-hand side of the equation can be replaced by the departure of the distribution function from equilibrium divided by a relaxation time τ, which incorporates all phonon scattering events under the relaxation time approximation. The relaxation time is the average time between collisions. The steady-state BTE with single relaxation time assumption and in the absence of any external force is given as

$$v \cdot \nabla_x f = -\frac{f - f_{\text{eq}}}{\tau} \tag{13}$$

The equilibrium function for phonons follows Bose-Einstein statistical distribution; therefore, the number of phonons with a particular frequency ω at an equilibrium temperature T is given by the equation

$$f_{\text{eq}} = \frac{1}{\exp\left(\frac{\hbar\omega}{k_B T}\right) - 1} \tag{14}$$

For small deviation from the equilibrium, $f = f_{\text{eq}} + \delta f$, it is assumed that

$$\nabla_x f \approx \nabla_x f_{\text{eq}} \tag{15}$$

Using the chain rule,

$$\nabla_x f_{\text{eq}} = \frac{\partial f_{\text{eq}}}{\partial T} \nabla_x T \tag{16}$$

Therefore, the BTE can be reduced to

$$f = f_{\text{eq}} - v\tau \cdot \frac{\partial f_{\text{eq}}}{\partial T} \nabla_x T \tag{17}$$

under the assumptions of a steady-state system with no external fields, and the departure of the system from equilibrium is small. The heat flux can be expressed as the product of group velocity and energy density integrated over all phonon frequencies:

$$q_x = \int v_x \, \hbar\omega \, D_p f \, d\omega \tag{18}$$

where v_x is the group velocity in the x-direction and D_p is the density of states. The density of states represents the number of quantized states available for phonon occupation. Assuming an isotropic case, $v_x = v_y = v_z$, the average group velocity over directions where all particles move in the x-, y-, or z-direction, each with the same speed, v_g

$$v_x^2 = \frac{v_g^2}{3} \tag{19}$$

Using the BTE expression, the heat flux equation becomes

$$q_x = \int -\frac{1}{3} \, v_g^2 \tau \, \hbar\omega \, \frac{\partial f_{\text{eq}}}{\partial T} \, D_p \, d\omega \, \frac{\partial T}{\partial x} \tag{20}$$

The equilibrium function from the BTE solution drops out of the integration because the equilibrium function contributes an equal amount of energy in all directions. Based on the Fourier Law for heat conduction, the thermal conductivity can be expressed as

$$k = \int \frac{1}{3} \, v_g^2 \tau \, \hbar\omega \, \frac{\partial f_{\text{eq}}}{\partial T} \, D_p \, d\omega \tag{21}$$

While the solution of the above equation provides the thermal conductivity of solids, the BTE model has limitations in its usage. Firstly, the solution is applicable for particles. The particle distribution function *f*(r,p,t) represents the probability of finding a particle with the position, the momentum, and the time. It is assumed that the momentum and the position are given simultaneously, which neglects quantum mechanical nature. The BTE is applicable for describing energy transport by particles, not waves. This allows the calculation of ensemble averages for particle density, current density, and energy density. Secondly, the solution is applicable for steady-state heat transfer problems or when the time scale of interest is much larger than the relaxation time of the energy carrier, in which case the first term can

be neglected. Thirdly, the solution is applicable for problems with length scale greater than the phonon mean free path.

In the case that the group velocity and the relaxation time are independent of frequency, the BTE solution reduces to the simple kinetic theory model:

$$k = \frac{1}{3} v_g^2 \tau \int \hbar\omega \frac{\partial f_{eq}}{\partial T} D_p \, d\omega \tag{22}$$

Phonon heat capacity by definition is obtained by differentiating the internal energy with respect to temperature:

$$C = \int \hbar\omega \frac{\partial f_{eq}}{\partial T} D_p \, d\omega \tag{23}$$

This reverts to the simple kinetic theory model for the thermal conductivity:

$$k = \frac{1}{3} C v_g \Lambda \tag{24}$$

Phonon Mean Free Path Modeling for Silicon Nanostructures

Thermal conductivity modeling requires computation of heat capacity, group velocity, and mean free path of energy carriers. While the heat capacity and the group velocity are based on dispersion relations, the mean free path accounts for size effects in nanostructures. The mean free path can be modeled in the framework of Matthiessen's Rule, in which scattering processes are parallel and independent from each other. The phonon mean free path in silicon can be expressed as

$$\Lambda^{-1} = \Lambda_U^{-1} + \Lambda_D^{-1} + \Lambda_B^{-1} \tag{25}$$

where Λ_U is the mean free path limited by phonon-phonon or Umklapp scattering, Λ_D corresponds to phonon-defect scattering, and Λ_B corresponds to phonon-boundary scattering.

There are two types of phonon-phonon scattering. Normal phonon-phonon scattering conserves energy and crystal momentum. The normal process redistributes crystal momentum, and there is no resistance to heat flow. Umklapp phonon-phonon scattering conserves energy but not crystal momentum. The Umklapp process, enabled by anharmonic lattices, makes phonon transport diffusive and creates thermal resistance, which is responsible for finite thermal conductivity in solids. The Umklapp scattering mean free path frequency can be expressed as

$$\Lambda_U = \frac{v_g}{A \, T \, \omega^2 \exp\left(-\frac{B}{T}\right)} \tag{26}$$

where A and B are fitting parameters. The $A = 1.4 \times 10^{-19}$ s/K and the $B = 152$ K are the fitted parameters that reproduce the bulk silicon thermal conductivity [6]. At high temperatures, the Umklapp scattering mean free path term begins to scale inversely with temperature, which corresponds to increasing population of phonons and increasing phonon-phonon scattering events. For this reason, at room temperature and above, the thermal conductivity of silicon decreases with increasing temperature. The mean free path due to phonon-defect scattering can be expressed as

$$\Lambda_D = \frac{v_g}{D \, \omega^4} \tag{27}$$

where D can be a fitting parameter. For nearly pure silicon, the parameter D mainly depends on the isotope concentration and has been determined as $D = 1.32 \times 10^{-45}$ s^3. For doped silicon, dopant species affect the thermal conductivity depending on their physical properties, and this can be expressed as [7]

$$D = \frac{nV^2}{4\pi v_g^3} \left(\frac{\Delta M}{M}\right)^2 \tag{28}$$

where n is the dopant concentration, V is the atomic volume of the host material, M is the atomic mass of the host material, and ΔM is the mass difference between the host and the dopant atoms. The mass differences between silicon and phosphorus and boron, and arsenic atoms are 2.89, 17.27, and 46.84 g/mol, respectively [4].

Therefore, arsenic is expected to cause the largest thermal conductivity reduction followed by boron and phosphorus for a given doping concentration. The frequency dependence resembles Rayleigh scattering, in which phonon waves do not interfere with defects that are much smaller than the wavelength. This scattering becomes important at low temperatures when Umklapp process diminishes, but when temperature is low enough that phonons have wavelength greater than defects and the defect scattering becomes less important. The phonon-boundary scattering term is important for low temperatures.

Phonon-boundary scattering can have strong dependence on the sample size. In the Casimir limit [9], the mean free path limited by phonon-boundary scattering becomes equivalent to the characteristic length scale. For polycrystalline materials, the characteristic length scale corresponds to the grain size. For single crystalline materials, the characteristic length scale is given by the sample size. For thin films or nanowires, the characteristic length scale may correspond to film thickness or diameter, respectively. For thin films, the phonon-boundary scattering term has been the control knob for matching thermal conductivity models with experimental data. For thin films, specular and diffuse scattering can be accounted by

$$\Lambda_B = d\left(\frac{1+p}{1-p}\right) \tag{29}$$

where d is the film thickness and p is the specular coefficient, which represents the probability of specular phonon reflection from the sample boundary ($0 \le p \le 1$). For $p = 0$, phonon-boundary scattering is assumed completely diffuse. In that case, the boundary is an entirely rough surface that diffusely, or randomly, scatters all incident phonons according to its surface temperature, similar to a blackbody emitting absorbed radiation. This effectively thermalizes the phonons, such that they contribute to the equilibrium phonon distribution. For $p = 1$, phonon-boundary scattering is assume completely specular. In that case, the boundary is a perfectly smooth surface that completely reflects all incoming phonons. The specular phonon-boundary scattering presents no thermal resistance and is applicable for bulk materials. The specular coefficient may be expressed in terms of the phonon wavelength and surface roughness for qualitative understanding [7, 8]:

$$p = \exp\left(-\frac{16\pi^3\eta^2}{\lambda^2}(\cos\theta)^2\right) \tag{30}$$

where λ is the phonon wavelength, η is the surface roughness, and θ is the angle of incidence with respect to the normal. This illustrates that a surface or interface appears rough for phonon wavelengths shorter than the average surface roughness height but smooth for phonon wavelengths much longer than the surface roughness. A given surface may diffusely reflect high-frequency phonons and specularly reflect low-frequency phonons. At low temperatures, the excited phonon modes tend to have low frequencies and long wavelengths, so a boundary appears smooth and specularly reflects some phonons. At higher temperatures, phonons occupy the more energetic modes with high frequencies and short wavelengths, so the effectively rough boundary diffusely scatters most phonons. This illustrates that a surface or interface appears rough for phonons with small incident angles but smooth for phonons with large incident angles. While the equation can be incorporated into the solution of the Boltzmann equation and integrated over all angles of incidence, the expression only for normally incident phonons is commonly incorporated in boundary scattering models for silicon films. Ziman suggested that this expression can underestimate p for incident waves at a large or glancing angle. While the expression provides qualitative understanding of wavelength and surface roughness relationship, the specular coefficient is also commonly set to a constant value to fit the thermal conductivity data for thin films.

For nanowires with completely diffuse boundaries, known as the Casimir limit [9], the average boundary scattering mean free path equals the diameter:

$$\Lambda_B = D \tag{31}$$

where D is the nanowire diameter. For rectangular nanobeams in the Casimir limit, boundary scattering mean free path can be expressed as

$$\Lambda_B = \frac{2}{\sqrt{\pi}}\sqrt{w\,t} \tag{32}$$

where w is the width and t is the thickness of the rectangular nanobeams.

For nanostructures with arbitrary heat flow directions and complex geometries, the angular dependence needs to be accounted by converting the Boltzmann transport equation solution in the frequency space (Eq. 21) into an integral in spherical coordinates. For cross-plane heat flows with vertical boundary scattering at top and bottom interfaces bounded in the length L, the average distance traveled ballistically by a phonon before hitting the top or bottom surface can be expressed as [10]

$$\Lambda_B = \frac{L}{2|\cos\theta|} \tag{33}$$

where L is the length scale in the direction of heat flow and θ spans from 0 to π. For cross-plane heat flows with lateral geometric confinement, i.e., nanowires, the average distance traveled ballistically by a phonon before hitting the side can be expressed as [10]

$$\Lambda_B = \frac{4D}{3\pi|\sin\theta|} \tag{34}$$

For nanowires with length shorter than the mean free path or superlattice nanowires, boundary scattering terms for both perpendicular and parallel interfaces are needed. The average distance traveled ballistically by a phonon before hitting the lateral dimension D or hitting the boundaries of dimension L in the direction of heat flow can be expressed as

$$\Lambda_B = \left(\left(\frac{L}{2|\cos\theta|}\right)^{-1} + \left(\frac{4D}{3\pi|\sin\theta|}\right)^{-1}\right)^{-1} \tag{35}$$

Phonon Mean Free Path Spectrum in Silicon

Phonon mean free path value facilitates thermal conductivity predictions in nanostructures and is the key parameter for evaluating scattering mechanisms. The phonon mean free path in silicon can be obtained using experimental values. One naïve approach is to use the kinetic theory (Eq. 4) with the known values for the thermal conductivity, heat capacity, and the group velocity:

$$\Lambda = \frac{3k}{Cv_g} \tag{36}$$

For silicon at room temperature, the specific heat is 1.66×10^6 J/m^3/K, the group velocity is 6400 m/s, and the thermal conductivity is 150 W/mK [5]. The phonon mean free path using Eq. 36 is about 40 nm. However, the thermal conductivity data for silicon thin films and nanowires show strong size dependence at dimensions much greater than 40 nm. The discrepancy is due to the severe underestimation of the phonon mean free path. Equations 4 and 36 neglect phonon frequency dependence and assume all phonons carry the same energy with the given heat capacity and the group velocity. Such a simple kinetic theory model may be a good approximation in highly disordered materials where the phonon mean free path is short and the phonon transport agrees with Einstein's minimum thermal conductivity model. However, in other materials such as silicon with broad phonon spectrum and strong frequency dependence, the heat capacity and the group velocity should be placed inside the integral, leading to

$$\Lambda_{avg} = \frac{3k}{\int Cv_g\,d\omega} \tag{37}$$

where Λ_{avg} is the average mean free path of phonons in the gray approximation. With the measured thermal conductivity value and the frequency-dependent heat capacity and the group velocity, the average mean free path is calculated to be about 250 nm. This implies

that the thermal conductivity size dependence becomes appreciable when the dimensions are 250 nm or less. The average mean free path value explains the size-dependent thermal conductivity observed in thin films and nanowires [11, 12]. Such a gray mean free path model is a good approximation in systems where the real mean free path distribution is narrow, including ideal gases and free electron gases. However, in other systems with strongly frequency-dependent scattering, the distributions can be quite broad. To better understand the potential for thermal conductivity reduction by boundary scattering, it is important to understand the range of mean free paths that contribute to heat transfer in a bulk material. The greater the fraction of heat carried by phonons with long mean free paths in bulk, the greater the thermal conductivity reduction in a given nanostructure. This effect can be quantified by rewriting the thermal conductivity expression to integrate over mean free path as the independent variable [5], leading to

$$k = \int -\frac{1}{3} C v_g \Lambda \frac{d\omega}{d\Lambda} \, d\Lambda \tag{38}$$

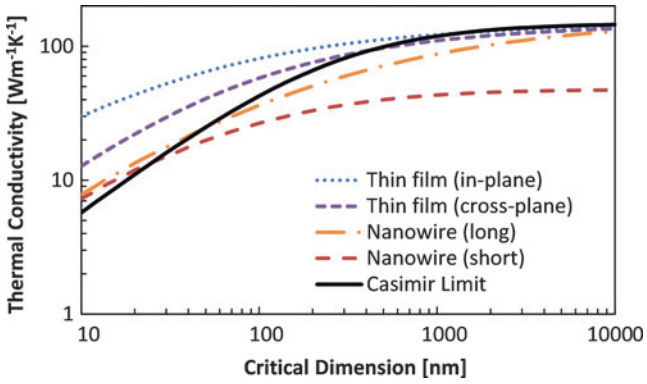
This expression quantifies the range of mean free path that contributes to thermal conduction. A complementary perspective is the thermal conductivity accumulation function:

$$\alpha = \frac{1}{k_{bulk}} \int_0^{\Lambda_\alpha} -\frac{1}{3} C v_g \Lambda \frac{d\omega}{d\Lambda} \, d\Lambda \tag{39}$$

which represents the fraction of the total thermal conductivity, ranging from 0 to 1, due to carriers with mean free path less than Λ_α. This calculation shows that 50 % of the heat in bulk silicon at 300 K is carried by phonons of mean free path up to 1 μm, which is greater than the average mean free path estimated from Eq. 37. The discrepancy is due to long-wavelength phonons that carry the heat in much greater distance than the short-wavelength phonons. In other words, the use of the average mean free path could underestimate the contribution of long-wavelength phonons. Furthermore, for up to 90 % of the heat, mean free paths up to 12.8 μm must be considered, and over two orders of magnitude of mean free paths are important for carrying the heat in silicon. This provides a useful guideline to the range of mean free paths that are important for heat conduction in a given system, and for silicon, the important mean free path ranges from 1 nm to 10 μm. The mean free path spectrum provides the relevant information for understanding phonon transport in nanostructures. The phonon mean free path models allow clear understanding of thermal conduction in nanoscale materials and systems and have great implications for ballistic phonon transport in silicon nanostructures (Fig. 2).

Quantum Effects and Future Directions for Research

While modeling thermal conduction in silicon nanostructures using bulk silicon dispersion relations is known to provide an accurate description of phonon transport, some experimental results do not agree well with the classical models, possibly due to quantum effects. The semiclassical models predict that the phonon scattering expressions based on the bulk dispersion relations are valid for nanowires with diameter greater than 10 nm and at temperatures higher than 40 K [3]. In very small dimensions or at very low temperatures, the phonon dispersion relations may become different from that in the bulk material. The change in the phonon dispersion relations can reduce the group velocity and results in further reduction in the thermal conductivity. The silicon nanostructures may become low-dimensional systems, in which frequency dependence of the scattering rates can be different due to confined dispersion relations. The quantum phonon transport effects, such as phonon confinement [14] or phononic bandgap formation [6], may influence heat conduction in very small dimensions or materials with nanoscale features. Recent experiments in silicon



Heat Transfer in Semiconductor Nanostructures, Fig. 2 The thermal conductivity of silicon nanostructures calculated from the solution of the Boltzmann transport equation. The critical dimension for thin films is thickness and for nanowires is diameter. For short nanowires, the length is defined as 100 nm, which may correspond to vertical interfaces with 100 nm pitch in superlattice nanowires. The Casimir limit for nanowires with completely diffuse boundary scattering yields the lowest thermal conductivity

nanostructures showed values that are much smaller than the lowest values predicted by the Boltzmann transport equation [6, 13]. These surprising results have led to theoretical investigations of whether the Casimir limit of diffuse surface scattering can be violated. Since the Boltzmann transport equation is based on particle transport, the models may fail to predict the thermal conductivity when the dimensions become comparable to phonon wavelength. Periodic features may result in phonon backscattering, which is equivalent to a negative p value [15]. Artificial spatial periodicity such as rough nanowires or phononic crystals can in principle result in the appearance of phononic bandgaps, which prohibit transport of phonons with certain frequencies. With periodic patterns, the zone folding effect may provide a low group velocity in a reduced Brillouin zone. These quantum effects on the phonon dispersion, particularly in the long-wavelength range, may be observed directly in experiments, and this requires further theoretical development. Thermal conduction in silicon nanostructures requires further investigation on whether phononic bandgap formation or another quantum effect becomes appreciable even at room temperature, which may lead to a breakthrough in phononics research and applications.

With the advance of nanotechnology and shrinking electronic devices, the thermal conductivity of silicon has renewed interest in understanding phonon transport in silicon, in particular to size effect in silicon nanostructures. The thermal conductivity of silicon thin films and nanowires has demonstrated significant reduction due to increased scattering of phonon on physical boundaries. In some cases such as rough nanowires and periodic structures, the thermal conductivity has been reported lower than the minimum thermal conductivity predicted by boundary scattering, and this calls for better understanding of phonon transport in silicon nanostructures. Understanding the mechanisms controlling the thermal conductivity at the nanoscale is of considerable importance in nanoelectronics and thermoelectrics applications.

References

1. Callaway, J.: Model for lattice thermal conductivity at low temperatures. Phys. Rev. **113**, 1046–1051 (1959)
2. Holland, M.G.: Analysis of lattice thermal conductivity. Phys. Rev. **132**, 2461–2471 (1963)
3. Mingo, N.: Calculation of Si nanowire thermal conductivity using complete phonon dispersion relations. Phys. Rev. B **68**, 113308 (2003)
4. McConnell, A.D., Goodson, K.E.: Thermal conduction in silicon micro- and nanostructures. Ann. Rev. Heat Tran. **14**, 129–168 (2005)

5. Dames, C., Chen, G.: In: Rowe, D.M. (ed.) Thermoelectrics Handbook, Macro to Nano. Taylor & Francis, New York (2006)
6. Hopkins, P.E., Reinke, C.M., Su, M.F., Olsson, R.H., Shaner, E.A., Leseman, Z.C., Serrano, J.R., Phinney, L.M., El-Kady, I.: Reduction in the thermal conductivity of single crystalline silicon by phononic crystal patterning. Nano Lett. **11**, 107–112 (2011)
7. Ziman, J.M.: Electrons and Phonons. Oxford University Press, Oxford, UK (1960)
8. Beckmann, P., Spizzichino, A.: The Scattering of Electromagnetic Waves from Rough Surfaces, pp. 29–81. Pergamon Press, New York (1963)
9. Casimir, H.B.G.: Note on the conduction of heat in crystals. Physica **5**, 495–500 (1938)
10. Sellan, D.P., Landry, E.S., Turney, J.E., McGaughey, A.J.H., Amon, C.H.: Phys. Rev. B **81**, 214305 (2010)
11. Ju, Y.S., Goodson, K.E.: Phonon scattering in silicon films with thickness of order 100 nm. Appl. Phys. Lett. **74**, 3005–3007 (1999)
12. Li, D.Y., Wu, Y.Y., Kim, P., Shi, L., Yang, P., Majumdar, A.: Thermal conductivity of individual silicon nanowires. Appl. Phys. Lett. **83**, 2934–2936 (2003)
13. Hochbaum, A.I., Chen, R.K., Delgado, R.D., Liang, W.J., Garnett, E.C., Najarian, M., Majumdar, A., Yang, P.D.: Enhanced thermoelectric performance of rough silicon nanowires. Nature **451**, 163–167 (2008)
14. Balandin, A., Wang, K.L.: Significant decrease of the lattice thermal conductivity due to phonon confinement in a free-standing semiconductor quantum well. Phys. Rev. B **58**, 1544–1549 (1998)
15. Moore, A.L., Saha, S.K., Prasher, R.S., Shi, L.: Phonon backscattering and thermal conductivity suppression in sawtooth nanowires. Appl. Phys. Lett. **93**, 083112 (2008)

Heatuators

▶ Thermal Actuators

Helical Nanobelt

▶ Nanorobotics for NEMS Using Helical Nanostructures

Heterogeneous Catalysts

▶ Simulation of Supported Metal Clusters

Heterogeneous Integration

▶ Self-Assembly for Heterogeneous Integration of Microsystems

HGN

▶ Hollow Gold Nanospheres

Hierarchical Mechanics

▶ Mechanical Properties of Hierarchical Protein Materials

High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles

Norbert Schell
Institute of Materials Research, Helmholtz-Zentrum Geesthacht, Geesthacht, Schleswig-Holstein, Germany

Synonyms

High-energy X-rays; Nanopowders; Structure of nanoparticles

Definition

High-energy synchrotron radiation does not have a clear and generally accepted definition. Its unique characteristic is the *high penetration into material*, and it depends on what is regarded as high penetration – millimeter or centimeter – and in what material, lightweight or comprising heavy atoms. For practical reasons which find their ways into special constructional solutions, one can regard the X-ray regime between 50 and

High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles, Table 1 Penetration depths for some elements plus water in the energy regime 50–150 keV

Element (atomic number)	Penetration depth at 50 keV	Penetration depth at 100 keV	Penetration depth at 150 keV
N (7)	41 m	53 m	61 m
H_2O (2/8)	3 cm	5 cm	10 cm
Si (14)	1.0 cm	2.3 cm	2.9 cm
Ti (22)	0.2 cm	0.8 cm	1.3 cm
Fe (26)	709 μm	0.37 cm	0.67 cm
Cu (29)	436 μm	0.25 cm	0.5 cm
Pd (46)	94 μm	615 μm	1648 μm
Au (79)	73 μm	101 μm	279 μm

150 keV as high-energy synchrotron radiation, thus a wavelength regime 0.08–0.24 Å. A look into the literature reveals a penetration of several 100 μm up to centimeters into most materials [1]. Table 1 lists the penetration depths (intensity falling off to e^{-1}) of some materials in the aforementioned energy regime.

With even higher energy, the gain in material penetration is usually offset by increased background Compton scattering and fluorescence radiation (also from shielding materials and slits). Visiting the home pages of (large) synchrotron radiation sources worldwide [http://www.lightsources.org/regions], one finds that they all offer corresponding specialized beamlines, i.e., experimental capabilities, for this regime. And usually, only the technical specifics of each instrument determine the lower (and likewise upper) boundary of this regime, where the limit is fluent down to even 30 keV with correspondingly lower penetration depths.

The *impact* of high-energy synchrotron radiation in *characterizing nanoparticles* is based on the general properties of X-rays interacting with matter, enhanced by the characteristics of synchrotron radiation (high brilliance and linear polarization), the details of which can be found in a number of textbooks [2, 3]. Obviously, those general properties and specific enhancements are not limited to characterizing only "nanoparticles" but also "non-nanomaterial." And indeed, this impact of high-energy synchrotron radiation is felt as a new force in many fields of materials research, physics, chemistry, and life sciences, dealing not exclusively with nanoparticles. However, this entry will stress the application of high-energy synchrotron radiation in the field of nanoparticles, their structural characterization, as well as the potential in understanding their formation, their domains of activity, and their workings in processes, e.g., catalysis.

Technical Peculiarities of High-Energy Synchrotron Radiation

There exist a number of *white beam* or *pink beam* techniques, i.e., techniques using a broad energy bandwidth out of the white synchrotron radiation beam; however, the overwhelming majority in scattering or imaging techniques relies on monochromatic beams, i.e., beams with a small energy bandwidth. They are usually created by reflection of perfect Si crystals (grating-based monochromatization for soft X-rays is omitted here). To keep the crystals perfect, they have to be cooled since the powerful synchrotron radiation sources (bending magnets, wigglers, or undulators) deposit enormous heat onto them with corresponding temperature and subsequently lattice gradients. Normally, they are kept under vacuum, and one uses two crystals in a fixed exit geometry where the reflection is in the vertical direction in order to keep the linear polarization of the synchrotron radiation (in the plane of the storage ring). Using Bragg's law "$2d \sin\Theta = \lambda$" with the lattice constant $d^{Si}_{111} = 3.135$ Å, one gets typical scattering angles of approximately 9° for an energy of 12 keV (wavelength $\lambda = 1$ Å). For 50–150 keV, the angles shrink to 2–0.8°. Hereby, the penetration length into Si is above 1 cm. Those two effects change completely the technical realization of monochromators for high-energy synchrotron radiation. (One can increase the scattering angle by choosing different lattice planes, e.g., Si(311) instead of Si(111), thereby, doubling it, but this reduces on the other hand the bandwidth and in consequence the transmitted flux.)

High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles,

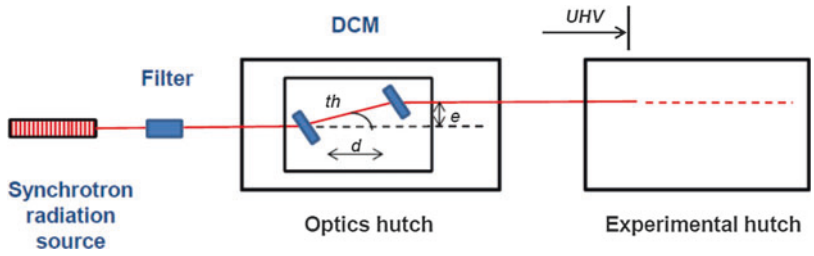


Fig. 1 Optics scheme for high-energy synchrotron radiation in top view

Figure 1 shows a typical optics scheme for high-energy synchrotron radiation delivery to an experiment. Small Bragg angles *th(eta)* lead to large footprints of the incoming beam, making Bragg reflection on perfect (Si) crystals prohibitively expensive. Therefore, one chooses a transmission mode, i.e., Laue geometry for the crystals (the radiation penetrates anyhow deep inside). Since the fixed exit angle *e* (between incoming white beam and monochromatic outgoing beam) is determined by external factors (construction or radiation protection), the distance *d* between the crystals gets large. For material stiffness reason, both crystals cannot any longer sit on the same rotation axis to guarantee the same scattering angle for monochromatization, so typically both crystals have separate rotation axis with the second crystal (downstream) sitting on a slide. But now some advantages make the beam delivery easier: small rotation ranges tend to be more precise and less expensive to manufacture. When concentrating on high-energy X-rays, one can filter out the lower energies by absorption; thus, the heat load onto the first crystal tremendously decreases allowing conventional water cooling (instead of liquid nitrogen) leading to a less expensive and more stable beam at the sample point. And as the bulk and not just the surface layers of the crystals is used for monochromatization, the vacuum requirements (surface contamination with water or organic deposits) are less severe, making interventions (repairs, changes) on the crystal(s) easier and faster. Actually, the high penetration of hard X-rays makes radiation delivery to the experiment easy: somewhere before the sample location normally the ultra-high vacuum (UHV) is changed to just vacuum or even air atmosphere which allows unhindered sample environment access as will be discussed later.

Another peculiarity of high-energy synchrotron radiation is that, normally, no longer a vertical scattering geometry is chosen, but a horizontal one. The reason is to keep the smaller vertical source size and to, only insignificantly, increase (through unavoidable aberration effects in the scattering plane) the much larger horizontal one. (As the linear polarization is proportional to the square of the cosine of the scattering angle *theta*, it practically stays above 90 % for theta $< 9°$ [3].)

As the critical angle for total external reflection is proportional to the wavelength λ, the use of perfect mirrors for the suppression of higher harmonics is prohibitively expensive which means that a monochromator is usually the only essential optical element in bringing the hard X-ray beam to the sample. Unfortunately, sometimes higher harmonics lead to unwanted artifacts in phase analysis. Practically, one depends on the decreasing flux for higher-energy synchrotron radiation or can slightly detune the crystals to suppress higher harmonics (but then loses flux). In computational view, hard X-rays allow the use of kinematical theory and the negligence of extinction and multiple Bragg scattering effects. In order to avoid chromatic aberration, the crystals are often bent into Rowland circle geometry, and depending on the experimental requirements, there exist various schemes to push the energy resolution up into the 1 eV region or down to several hundred eV to increase the bandwidth. Usually, a channel-cut design is used for high resolution (often as a second monochromator in front of the sample), while a controlled lattice constant gradient is used for increased bandwidth, i.e., flux increase (e.g., by doping Si crystals with Ge, by bending or defining heat gradients inside the monochromator crystals, or by using multilayer crystals with a thickness gradient).

Structural Characterization In Situ

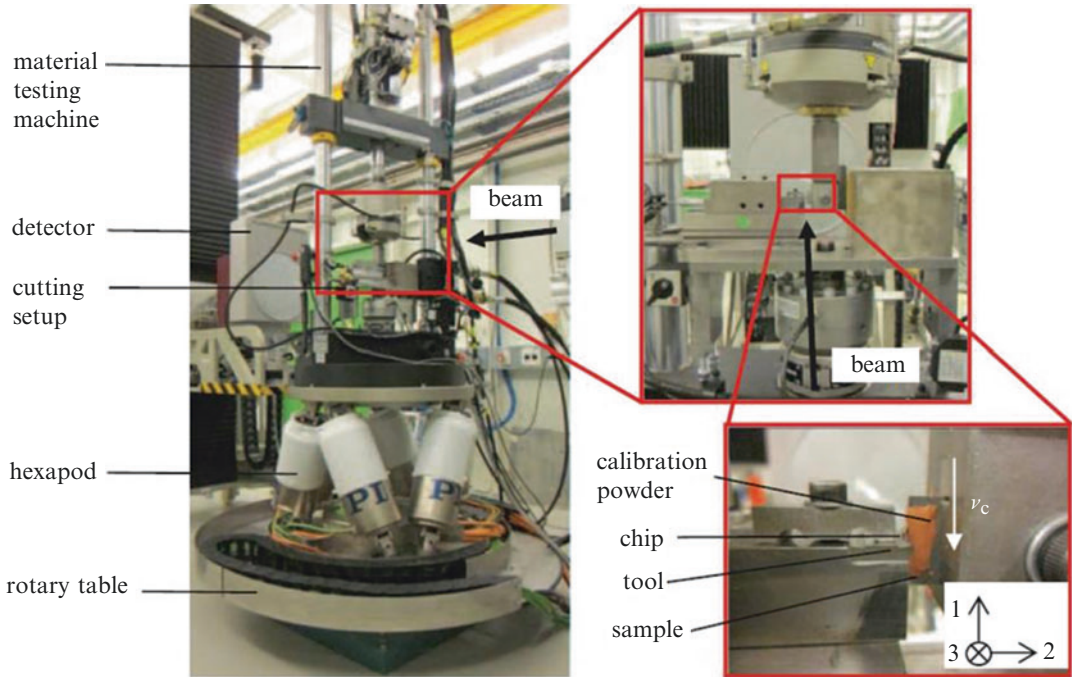
Nanoparticles are *per definitionem* characterized by their size or size distribution and their morphology being in the nm range. In contrast to characterizing methods like ▶ scanning electron microscopy (SEM), ▶ transmission electron microscopy (TEM), or atomic force microscopy (AFM), X-rays demand less effort in sample preparation, and hard X-rays are always required either if the morphology gets amorphous – short-range order requiring large reciprocal space for signal detection – or if the particles are embedded in a bulk. With small sample volumes, high-energy synchrotron radiation is then the method of choice. And as the method is destructionless, the nanoparticles are described *in their natural environment*. In a wider sense, the structural characterization comprises the phase of the material and its stress or strain state compared to a corresponding reference substance (if available).

This harmless sounding term "embedded in their natural environment" is a consequence of the high penetration of high-energy X-rays. Another consequence of high energy or small wavelength is the fact that the scattering signal is squeezed into a small solid angle which can be captured in a single shot with the currently available large and fast two-dimensional (2D) detectors [4, 5]. And this combination makes all the difference. The brilliance of synchrotron radiation allows for a high local resolution with sub-micrometer (focused) beams and corresponding small sample volumes, the fast readout of 2D detectors down to several 100 Hz (and ever increasing) allows for a high time resolution (with no need for scanning, e.g., a diffractometer), and the high penetration generally allows for widely varying sample environments, i.e., buried structures are accessible in reaction chambers. This in situ characterization of materials is currently booming, driven by the rapid technical advances in detector development, and it comprises not only X-ray techniques but also microscopy and neutron scattering [6]. Nanoscaled materials are especially prone to fast dynamic processes as their influences propagate quickly across the intrinsically small dimensions. And this starts with the synthesis and the growth of nanostructures [7]. Figures 2, 3, and 4 exemplify the easy accessibility of high-energy synchrotron radiation for various sample environments which often can be commercially available (with none or minor modifications), here for mechanical testing of samples under load, dynamical scratch testing, and laser beam welding.

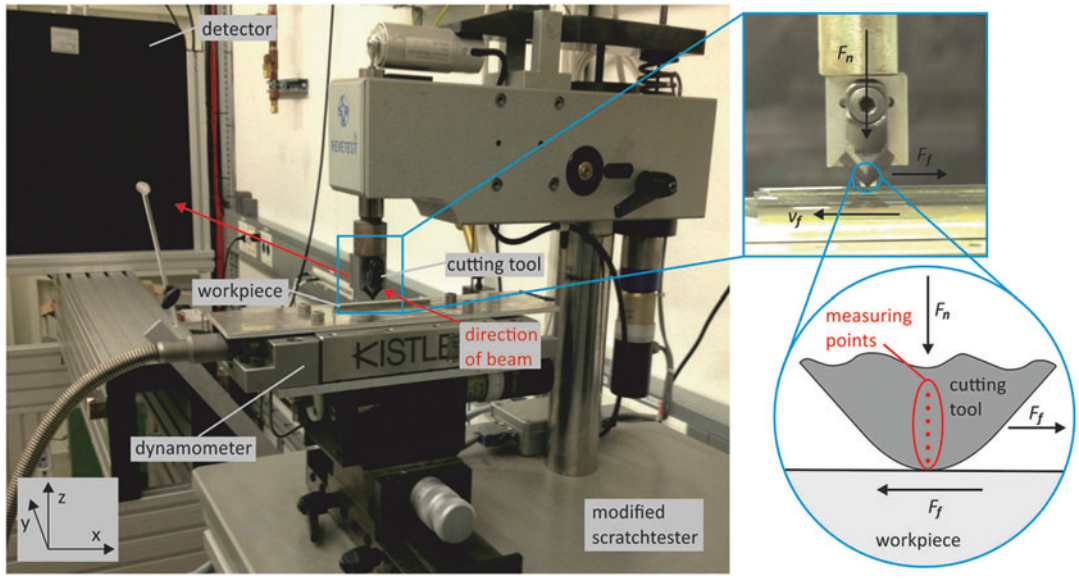
As high-energy synchrotron radiation is highly penetrating, there do not exist in principle any restrictions in characterizing *any* synthesis method with it: "breaking down of bulk material into nano-sized structures or particles" in a mill is not different in accessing with high-energy X-rays as "building up material from the bottom in atom-by-atom processes" whether in a gas phase or a liquid solution or in vacuum, independent of external temperature, pressure, or magnetic fields.

High-Energy Surface X-ray Diffraction

For nanoparticles due to their sizes, the influences of their surfaces are playing a dominant role. Understanding the interaction between the surfaces – solid, liquid, or buried – and their surroundings is thus crucial in understanding the properties of nanoparticles. And high-energy synchrotron radiation offers unique possibilities. While bulk crystalline matter leads to Bragg peaks of scattered X-rays, a surface leads to crystal truncation rods (CTRs) perpendicular to the surface. Changing in-plane periodicity leads to additional superstructures in the intensity distributions [10]. The measured intensities are several orders of magnitude less than for Bragg peaks. To gain complete information, one has to rotate the sample, i.e., rotate the Ewald sphere in reciprocal space, which for low-energy photons is time consuming. For high photon energies, 2D cuts of reciprocal space can be acquired in large 2D detectors leading to a significant speed up of data taking, approximately two orders of magnitude. Gustafson et al. [11], e.g., were thus able to measure the intensities of surface diffracted beams to follow the surface oxidation that accompanied the changes in a palladium surface during the catalytic oxidation of CO with O_2 at 85 keV in a chamber allowing the careful control of pressure, gas flow, and temperature.



High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles, Fig. 2 A huge loading device for 100 kN (Instron 8800) in order to investigate the chip formation in steel at beamline P07 at PETRA III, DESY [8]



High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles, Fig. 3 A modified scratch tester for the in situ determination of internal stresses in mixed ceramic cutting tools during friction testing at P07B of PETRA III, DESY [9]



High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles, Fig. 4 An in-house built laser welding chamber at P07 at PETRA III, DESY, allowing specific protective gas environments for studying phase transformations and strain development under realistic, i.e., industrially applied laser welding of, e.g., titanium aluminides (the laser, cooling, and heating tubes are not yet connected)

The abovementioned example does not state that surface characterization is or will be dominated by high-energy X-ray diffraction – there exists a wide range of powerful techniques that are used to characterize the surfaces of materials. The reader is referred to introductory guides in the literature [12]. On the other hand, the example shows the maturity and sensitivity reached and the complexity in process characterization with the help of high-energy synchrotron radiation diffraction.

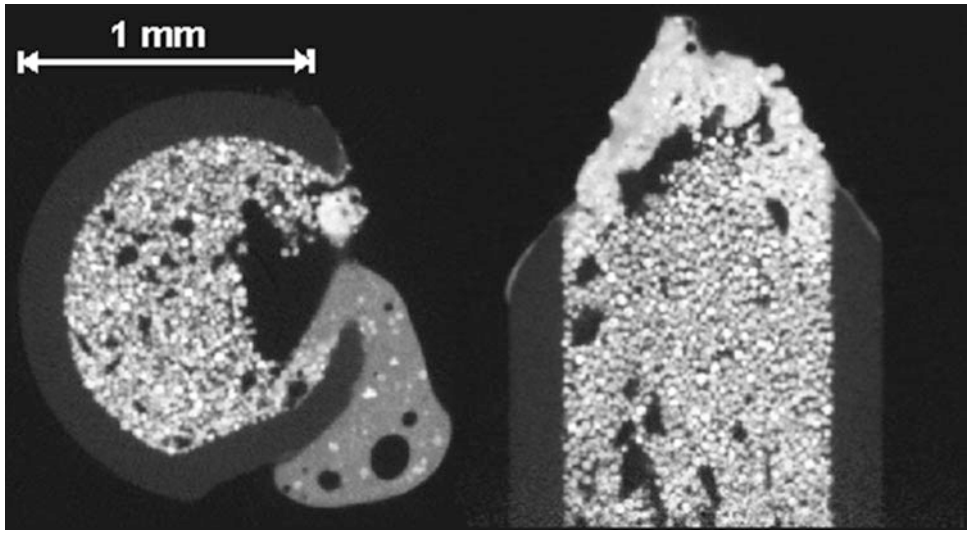
High-Energy Small-Angle X-ray Scattering

When nanoparticles are defined by their nm size or size distribution, their high-energy X-ray diffraction pattern in the here discussed regime of 50−150 keV (wavelength regime 0.08–0.24 Å) is automatically a small-angle X-ray scattering (SAXS) pattern. A more subtle definition regards SAXS as a scattering technique for structures which are much larger than the wavelength of the interacting radiation giving information about the size, shape, and orientation of those structures in a sample. The most important feature is the potential for analyzing the inner structure of disordered systems as SAXS is dependent on the average electron density difference between a particle (or a pore) and the surrounding matrix. A plethora of general [13] and specialized literature examines the method and its applications including chemical aggregation, defects, ferromagnetic correlations in magnetism, segregation, polymers, biological samples, etc.

If the particle or matrix is highly absorbing or put into a special environment (furnace, reaction chamber, stress rig, etc.), high-energy X-rays are needed to penetrate the sample (or environment) – this is the same argument as given above for wide-angle scattering.

Imaging Techniques with High-Energy X-rays

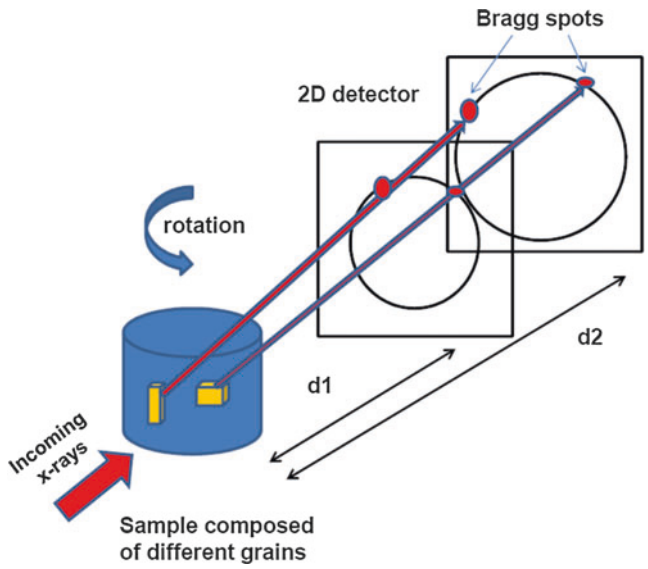
While the SAXS signal is measured as small intensity away from the trajectory of the incoming beam after interaction, caused by average electron density differences between particle and surrounding matrix, imaging techniques produce slices of a



High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles, Fig. 5 Feasibility study for SRμCT at 120 keV at beamline P07 at the storage ring PETRA III, DESY, Germany. The sample is a cored wire, with outer material Ni, inner particles WC/Co of 25–40 μm diameter, reconstructed volume (2.0 × 2.1 × 2.3) mm^3, pixel size 2.4 μm, and spatial resolution 5 μm. The tomogram allows the characterization of material properties in dependence of the manufacturing parameters (Courtesy F. Beckmann from Helmholtz-Zentrum Geesthacht, in cooperation with J. Nellesen, M. Abdulgader, and W. Tillmann from the Institute of Materials Engineering, Department of Mechanical Engineering, Technical University Dortmund)

scanned object to see what is inside. Most prominent is the X-ray computed tomography (X-ray CT) based on the absorption or the phase contrast of the different materials comprising the sample [14]. This technique needs generally *big data* handling as the resolution increases with the number of projection scans, and it is dependent on the accuracy of the rotation axis, i.e., its wobble which moves the sample – the smaller it is, the higher the accuracy. This has two severe consequences: the method is slower in comparison with single-shot scattering 2D images (where the information is immediately obvious), i.e., the number of "processes" to study is lower, and any sample environment is restricted in its weight. Since rotation axes are easier to handle in the vertical, high-energy synchrotron radiation tomography prefers a vertical scattering geometry as the monochromator aberration in the horizontal plane leads to artifacts of slightly different absorption when reconstructing the 3D image. Figure 5 shows a feasibility study for using synchrotron radiation micro-computed tomography (SRμCT) at 120 keV, necessary because of the high atomic number sample.

Interesting enough, one can image the inside of a sample likewise with X-ray diffraction spots, a technique labeled three-dimensional X-ray diffraction microscopy, short 3DXRD [15]. Its principle is shown in Fig. 6. Again the method is time consuming and calculation intensive, since one has to track typically 10–1000 individual grains with their distinct sizes, orientations, and strain states. The currently dedicated instruments for grain mapping (at, e.g., APS-1ID, ESRF-Id11, DESY-P07) are working in the micrometer size and resolution, but with technical detector and software development, the future capabilities will be driven to resolution in the several 10 nm range. And great efforts are under way to follow the behavior of the single grains, especially, under loading conditions to validate the widespread use of finite element (FE) models used in materials engineering. As nanoparticles often stiffen materials, this investigating technique will have applications for their characterization in the industrially important field of materials engineering.



High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles, Fig. 6 Scheme for 3DXRD. An incoming photon is diffracted into a specific Bragg spot which is caught in a 2D detector at different distances d1 and d2 which allows a back traction of the specific grain inside the sample. A stepwise rotation by 180° allows finally all grains to form the complete Debye-Scherrer rings in both detectors, i.e., for both distances. The individual spot (size, substructure, intensity) depends on the individual grain (size, orientation, strain, and stress) which caused the Bragg spot. Thus, one can map completely the grains inside a polycrystalline sample

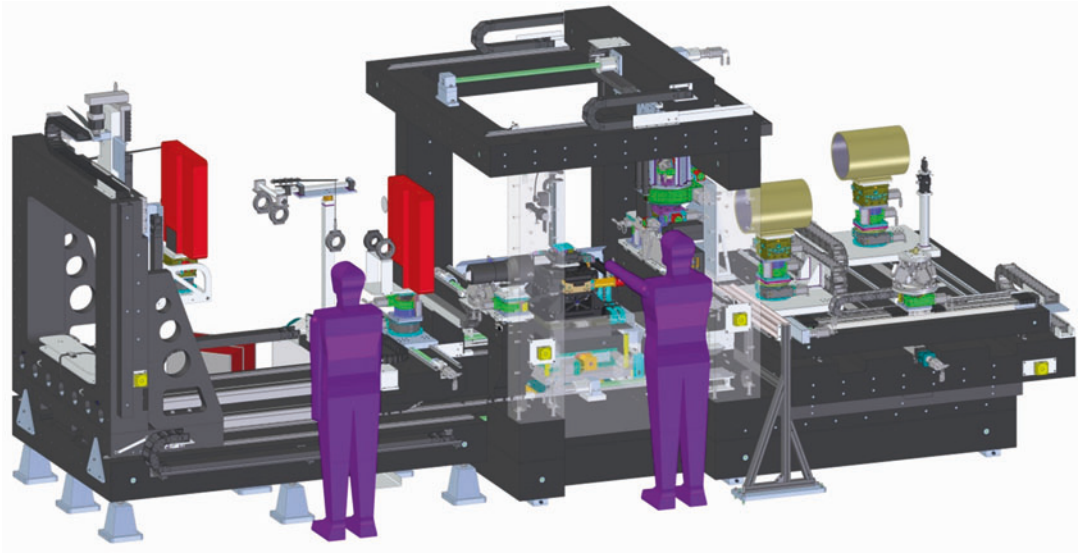
Since the technical demands of SRμCT and 3DXRD are quite similar, a single experimental setup can realize both techniques, and as they are complementary, they can be combined to improve resolution and information (omitting inelastic scattering whereby a photon is either absorbed or scattered) which is done in diffraction contrast tomography (DCT) [16]. Figure 7 shows the design of a dedicated grain mapping instrument which also allows DCT [17].

Medicine and Life Sciences

Imaging techniques play a most prominent role in medicine and life sciences. Especially, radiographs are done in the energy regime 30–150 keV. Those fields are left out of this entry as the main impact is delivered in hospital environments relying on laboratory sources and not synchrotron radiation. However, leading techniques are today often developed at synchrotron radiation sources before they get a transfer to laboratory X-ray sources. In a nice reverse trend, many 2D detectors, originally designed for medical applications, find now their way to research at synchrotron radiation sources. While nanoparticles find their way, e.g., into implants, they can then be handled under the aspect of materials research with the all abovementioned impact of high-energy synchrotron radiation characterization. Likewise the importance of surfaces or chemical reactions in biology – membranes, tissue growth, etc. – can be handled by the respective topics talked about above. Even *bioreactors* can be seen as just other in situ *reaction chambers* in order not to blow up this entry.

Only one more advantage of high-energy synchrotron radiation for life sciences is given: the merit of high penetration was often stressed to give a glimpse of the many possibilities being offered to characterize processes; in life sciences, this weak interaction can have the tremendous advantage of less radiation damage, thus avoiding or at least reducing the impact of the very characterization, itself.

A final word should be given to a certain ambiguity of definition. Biomolecules and protein



High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles, Fig. 7 Design concept of the dedicated 3DXRD instrument at the high-energy materials science beamline P07 at the storage ring PETRA III, DESY. The beam is coming from the right and can be focused either by Kirkpatrick-Baez (*KB*) mirrors or compound reflective lenses (*CRLs*). The sample sits on a precision rotation axis in the center around which an optical microscope for alignment and the near-field X-ray detectors can be positioned – the latter allow the back tracing of single grains. The far-field detectors on an optical bench to the left allow detailed strain characterization and also SRμCT and DCT

structures are not regarded here as nanoparticles although they span the size range of nanoparticles. Thus, likewise, the field of protein crystallography is omitted which anyhow is dominated by structural characterization with X-rays below 30 keV. And, in consequence, the emerging field of structural biology with its aim to understand the connection between structure and function of biological macromolecules is omitted. Its *impact* of course is enormous; it will be developed mostly with free electron lasers in the soft X-ray regime – which is no longer classical synchrotron radiation [2]. However, there will be sources in the future which reach into the aforementioned regime of 50–150 keV with spontaneous emission [18]. But that should be another entry.

Cross-References

- ▶ Nanoparticles
- ▶ Small-Angle Scattering of Nanostructures and Nanomaterials

References

1. Henke, B.L., Gullikson, E.M., Davis, J.C.: X-ray interactions: photoabsorption, scattering, transmission, and reflection at E = 50–30000 eV, Z = 1–92. At. Data Nucl. Data Tables **54**(2), 181–342 (1993)
2. Als-Nielsen, J., McMorrow, D.: Elements of Modern X-ray Physics, 2nd edn. Wiley, Chichester (2011)
3. Lifshin, E. (ed.): X-ray Characterization of Materials. Wiley-VCH, Weinheim (2008)
4. Thompson, A.C.: X-ray detectors. In: Thompson, A.C. (ed.) X-ray Data Booklet, pp. 4/32–4/39. Lawrence Berkeley National Laboratory/University of California, Berkeley (2009)
5. He, B.B., Preckwinkel, U., Smith, K.L.: Fundamentals of two-dimensional X-ray diffraction (XRD2). In: JCPDS-International Centre for Diffraction Data 2000. Advances in X-ray Analysis, International Centre for Diffraction Data, PA, USA, vol. 43, pp. 273–280 (2000)
6. Ziegler, A., Graafsma, H., Zhang, X.F., Frenken, J.W. M. (eds.): In-Situ Materials Characterization. Springer Series in Materials Science, vol. 193. Springer, Berlin/Heidelberg (2014)
7. Dubrovski, V.G.: Nucleation Theory and Growth of Nanostructures. NanoScience and Technology. Springer, Berlin/Heidelberg (2014)
8. Brömmelhoff, K., Henze, S., Gerstenberger, R., Fischer, T., Schell, N., Uhlmann, E., Reimers, W.:

Space resolved microstructural characteristics in the chip formation zone of orthogonal cut C45E steel samples characterized by diffraction experiments. J. Mater. Process. Technol. **213**, 2211–2216 (2013)
9. Eichenseer, C., Hartig, C., Schell, N., Hintze, W.: In-situ determination of internal stresses in mixed ceramic cutting tools during friction testing using synchrotron radiation. Prod. Eng. Res. Dev. **8**, 513–519 (2014)
10. Feidenhans'l, R.: Surface structure determination by X-ray diffraction. Surf. Sci. Rep. **10**, 105–188 (1998)
11. Gustafson, J., Shipilin, M., Zhang, C., Stierle, A., Hejral, U., Ruett, U., Gutowski, O., Carlsson, P.A., Skoglundh, M., Lundgren, E.: High-energy surface X-ray diffraction for fast surface structure determination. Science **343**, 758–761 (2014)
12. Brune, D., Hellborg, R., Whitlow, H.J., Hunderi, O.: Surface Characterization. Wiley-VCH, Weinheim (2008)
13. Feigin, L.A., Svergun, D.I.: Structure Analysis by Small-Angle X-ray and Neutron Scattering. Springer, Berlin (2013)
14. Buzug, T.M.: Computed Tomography. Springer, Berlin (2008)
15. Poulsen, H.F.: Three-Dimensional X-ray Diffraction Microscopy. Springer, Berlin (2004)
16. Reischig, P., King, A., Nervo, L., Vigano, N., Guilhem, Y., Palenstijn, W.J., Batenburg, K.J., Preuss, M., Ludwig, W.: Advances in X-ray diffraction contrast tomography: flexibility in the setup geometry and application to multiphase materials. J. Appl. Crystallogr. **46**, 297–311 (2013)
17. Schell, N., King, A., Beckmann, F., Fischer, T., Müller, M., Schreyer, A.: The high energy materials science beamline (HEMS) at PETRA III. Mater. Sci. Forum **772**, 57–61 (2014)
18. Altarelli, M., et al. (eds.): XFEL, The European X-ray Free-Electron Laser. Technical Design Report. DESY 2006–097 (2006)

High Index Fluid Immersion

▶ Immersion Lithography

High-Density Testing Cards

▶ MEMS High-Density Probe Cards

High-Energy X-Rays

▶ High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles

Higher-Order Plasticity Theory

▶ Plasticity Theory at Small Scales

High-Pressure Carbon Monoxide (HiPCO)

▶ Chemical Vapor Deposition (CVD)

HIV Quantification

Benjamin L. J. Webb[1] and David Holmes[2]
[1]Division of Infection and Immunity, University College London, London, UK
[2]London Centre for Nanotechnology, University College London, London, UK

Viruses

Definition

Viruses are small biological infectious particles that can infect other biological organisms, and replicate only inside living cells [1]. They can only reproduce by invading and using host-cell machinery. Viruses infect most organisms, including animals, plants, and bacteria. Human immunodeficiency virus (HIV) is a retrovirus [2], meaning that its genetic material is in the form of RNA. To reproduce, it incorporates its genetic material into specific host cell's genome. To do this, the viral RNA must be reverse transcribed by a viral-encoded enzyme into DNA. HIV is the cause of acquired immune deficiency syndrome (AIDS) in humans [3].

Overview

The term "virus" was coined by Martinus Beijerinck in 1898, following his work on the discovery of the tobacco mosaic virus. There are estimated to be 10^{31} viruses existing on Earth, most of which are bacteriophages. It is estimated that 5000 viral genotypes are present in just 200 L

of seawater. Of the many different types of virus, over 5000 viruses have been described in detail. Currently, there are 241 recognized virus genera [1].

Whether viruses are life, or organic structures that interact with living organisms, is moot. That they possess genes, evolve by natural selection, and undergo self-assembly, suggests they are life. Yet, though they have genes, they possess no cellular structure, which is often stated a criterion that defines life. Since viruses have no metabolism, and cannot naturally reproduce outside a host cell, they are parasites that utilize life forms in order to maintain their own form of existence.

Structure

Viruses are the smallest parasites, ranging from 20 to 300 nm in diameter (some filoviruses are up to 1400 nm in length), making them visible only by electron microscopy.

Viral particles (*virions*) consist of genetic material, made from either DNA or RNA, encapsulated in a protein coat called the capsid. Some viruses possess a cell-derived membrane from the infected host, sometimes impregnated with viral-encoded proteins as receptors for cell entry.

Virally encoded protein monomers self-assemble to form the capsid. These form the morphological virus type, of which there are four: helical, icosahedral, enveloped (these viruses are envelopes in a modified host membrane), and complex (a complex arrangement of other types, with additional structures) [1].

Genome

Viruses display more genomic diversity than plants, animals, or bacteria. Viral genomes can be circular or linear, depending on the group, and is not a function of the form of nucleic acid. Most genomes are either single-stranded or double-stranded, but some are both (Table 1).

Most RNA viruses, and some single-stranded DNA viruses, possess single strands of genetic material that is either positive sense or negative sense, defined according to whether complementary to the viral mRNA. Positive-sense viral RNA is in the same sense as viral mRNA and can be translated by the host cell. Negative-sense viral RNA requires conversion into positive-sense RNA.

HIV Quantification, Table 1 Viruses can be grouped according to the genome type they possess (DNA or RNA), if single or double-stranded, whether their RNA is messenger RNA (mRNA) (+) or complementary to mRNA (−), and if they possess reverse transcriptase (*RT*)

Group	Virus type
I	dsDNA viruses
II	ssDNA viruses
III	dsRNA viruses
IV	(+) ssRNA viruses
V	(−) ssRNA viruses
VI	ssRNA-RT viruses
VII	dsDNA-RT viruses

The viral genome size varies greatly between species; from <2 kilobases (kb) to over 1.2 megabases (Mb), which encodes over 1000 proteins. RNA viruses generally have smaller genomes than those of DNA viruses, due to a higher error-rate when replicating. To compensate for this error rate, RNA viral genomes are split into smaller molecules – reducing the possibility of replication errors which would render the virus defunct.

Viruses use the machinery of a host cell to produce multiple copies, which are assembled within the infected cell. There are six basic stages in the life cycle of viruses: attachment to the target cell, penetration into that cell, uncoating of capsid to release genetic contents, replication of viral genome, assembly of new particles, and release of new virions into extracellular fluid. The specifics of each stage vary considerably between different types of virus. RNA viruses usually replicate their genome in the cytoplasm whereas, for most DNA viruses, this occurs in the nucleus. After release of newly synthesized virions, the host cell usually dies.

Viruses and Human Disease

Viruses can be transmitted vertically (from mother to child), or through the more common horizontal (from person to person) method. They can also be transmitted by blood-sucking insects (vectors), coughing and sneezing (e.g., influenza), the

fecal-oral route (e.g., norovirus), from person to person contact, food, and water. HIV, among other viruses, can be transmitted through sexual contact and exposure to infected blood by contaminated needle sharing or transfusion, and from a pregnant woman to her baby during pregnancy, delivery, and breast feeding.

Viruses have variable "host ranges," meaning the species that they can infect. Additionally, within a specific host, the virus may only be able to infect certain cell types (e.g., CD4+ cells for HIV). Responses to viral infection are variable. Some infections cause acute illness following a brief incubation period, whereas others produce minor symptoms or are asymptomatic. The immune system can clear many viral infections, but some viruses can remain undetected in a latent state. Here, viral RNA or DNA remains in host cells up to years later without causing disease. During this "latent" period, viral infections may still be transmissible. An immunosuppressed state may reactivate the latent virus, producing symptoms. Herpes viruses and HIV are viruses that produce latent infections.

Viruses exist globally, but their infectious influence on humans is limited by genetic resistance, immunization, public health control measures, and antiviral drugs. "Zoonotic" viruses can cross species. These originated in animals, with humans as secondary or accidental hosts. Hundreds of different viruses infect humans. They usually spread via respiratory and enteric excretions, but also by blood transfer, sexual activity, and arthropod vectors.

Many human diseases are caused by viruses, from mild (e.g., chickenpox), to serious (e.g., AIDS). They cause disease at the organism, and cellular, level. For the latter, this is primarily by cell lysis. For large organisms, if enough cells are killed, the effects can lead to serious disease and death. Some viruses that can exist in a latent state, such as herpes simplex virus (HSV), are harmless to an organism in this state. However, on activation, they cause symptoms of infection. Other viruses, such as hepatitis B, cause chronic infections because they are able to evade the immune system.

Some viruses are oncogenic, causing predisposition to cancers (e.g., Hepatitis B virus and hepatocellular carcinoma). The clinical diagnosis of human viral infections can be through well-known viral syndromes (e.g., measles), epidemiologically (e.g., during influenza epidemics), or by specific laboratory diagnosis (e.g., HIV). The latter is important not only for the infected individual, but also when the agent is a public health threat.

The body's first line of defense against viruses is the innate immune system, which is the intrinsic system comprising specialized cells and proteins. This doesn't provide long-term protection since it doesn't adapt to the pathogen. The adaptive immune system, as its name suggests, *does* adapt, but, unlike the innate immune system, requires a period of time after exposure to the pathogen in order to become active. Here, specific antibodies are generated that bind to the virus rendering it non-infectious. The production and secretion of antibodies is part of the humoral immune response. This differs from the cell-mediated immune response, which involves T-cells. However, together they form the adaptive immune system. A virus may infect a specific cell type called an antigen-presenting cell (APC) (B-cell, macrophage, dendritic cell). These cells then "present" a viral component to a naïve T-cell. These become primed to activate a specific B-cell, which produces antibodies against the virus first encountered by the APC (hence, the link). Some viruses, such as HIV, can avoid the innate response and takes a hold in the host before the adaptive response is able to block infection. It is for this reason that effective vaccines are so eagerly strived for.

Prevention and Treatment

Since viruses utilize host cell metabolic pathways to replicate, the problem of eliminating them must be tackled with compounds non-toxic to the host. Currently, attempts to prevent and treat viral infections is through vaccination (for immunity), and antiviral drugs (to block the viral life cycle).

Following immunization with an inactive or dead component of a virus, the immune system is stimulated to produce antibodies against that virus. Should that individual become infected, their immune system is prepared to halt the viral

life cycle before it starts. Commonly used viral vaccines include those against influenza, measles, and mumps. Immunoglobulins (antibodies) can also be given to passively immunize susceptible or immunocompromised individuals.

Once developed, vaccines are a cheap and effective way to prevent infections. They have produced profound improvements in viral infection–associated morbidity and mortality (e.g., for polio). To date, vaccines are available to prevent over 13 viral infections of humans. Vaccines may be live but attenuated virus (weakened and, therefore, not harmful), killed viruses, or viral proteins.

Significant investment and progress is being made into antiviral drug development. These drugs are directed at various phases of viral replication. Drug targets include the site of virion–host cell interaction, the uncoating process for viral nucleic acids, and viral factors required for viral replication.

Interferons are used as antiviral compounds. They are compounds released from infected host cells in response to viral or other foreign antigens. They block translation and transcription of viral RNA, and stop viral replication without disturbing host cell functions. They upregulate innate antiviral proteins, such as members of the TRIM family of proteins.

DNA analogues are the basis of typical antiviral drugs. Viruses incorporate these into their genomes during replication and, as a result, newly synthesized DNA is dysfunctional and the viral life cycle is blocked. Other drugs target different stages of the viral life cycle. For HIV, the virus-encoded protease enzyme is essential for infectivity, as it cleaves other viral proteins into active forms. Drugs called protease inhibitors inhibit this enzyme, rendering the virus, in effect, dead.

HIV

Overview

HIV is of particular interest because it is the cause of AIDS [3]. An individual can become infected with HIV by a number of possible routes: sex (vaginal, anal, oral) with an infected individual, sharing infected needles (drug use, tattoos, piercings), transfusion of infected blood, and from mother to child during pregnancy, birth, and breast feeding [3].

HIV is a virus that exists as two main types: HIV-1 and, the less virulent, HIV-2. HIV-1 was discovered in the early 1980s as the agent that causes the AIDS [4–6]. HIV-1 is thought to have zoonosed to humans from simian immunodeficiency virus (SIV_{cpz})-infected chimpanzees, perhaps through the eating of bush meat in Sub-Saharan Africa (SSA). HIV-2 zoonosed from sooty mangabeys monkeys, infected with another virus, SIV_{sm}.

HIV-1 infects 33.4 million infected individuals (2009), 27 million of who are in SSA (2/3 of all infections). Since records began, 25 million people have died from HIV-1-related complications. However, this number is probably much higher if one considers the difficulty to produce accurate numbers (considering the resources of the, mostly 3rd world, nations primarily affected) [7].

HIV is the 6th largest world cause of death, and the 4th largest in low-income countries (in 2004). In 2007, there were 2.1 million deaths (550,000 children), and 2.5 million new infections from HIV. In 2008, there were 2.0 million deaths, and 2.7 million new cases, so even after 30 years after detection of the causative agent of AIDS, the virus still presents a significant global health concern. Though in 1st world countries there is, as will be shown, efficient diagnostic and treatment measures, it is in the 3rd world where progress is essential. Indeed, it is estimated that 83 % of Kenyan cases are not diagnosed (2009). Ninety percent of global pediatric HIV is in SSA. Half of HIV-infected infants less than 18 months old in SSA will die due to AIDS.

Rather than splitting the world into social and economic groups, and evaluating the global problem on these grounds, if one considers the planet as a singular entity in which the virus can proliferate, then a more effective assault would surely ensue. Due to the nature of HIV replication, it has very high genetic variability. Unchecked in SSA, drug-resistant forms could emerge which, due to the ever-shrinking national borders, might impact

strongly on globally infected numbers. Thus, the problem of HIV is not national, but international. However, the UN published "The Millennium Development Goals Report" in 2010 [8], indicating that the global spread of HIV appears to have stabilized in most regions, and more people are surviving longer (see Fig. 1).

Human HIV infection is associated with social stigma. In 2004, 38 % of HIV diagnosed individuals in the USA developed AIDS in less than 1 year. Infected individuals may be reticent to seek help, and/or may do so only when symptoms are prominent (when the virus is less likely to respond favorably to antiviral drugs).

In the USA, more than one million are infected by HIV. In the UK (2005), 55,000 were infected, with an additional 20 % suspected undiagnosed. There were also high rates of drug resistance. In 2010, the number infected in the UK raised to 83,000, with 500 deaths/year. Thus, even with all facilities in place, HIV poses a problem [5, 6, 9].

Phylogenetics

HIV is a group V1 (ssRNA-RT) virus. It is a member of the family *Retroviridae*, which are the retroviruses [2]. These are RNA viruses, which replicate in a host cell using the enzyme reverse transcriptase to produce DNA from an RNA viral genome. In the *Retroviridae* subfamily *Orthoretroviridae* are the lentiviruses, including, among others, HIV and SIV. Lentiviruses are characterized by long incubation period in the infected host. HIV consists of two species, the commoner, highly virulent and highly infectious, globally prevalent, HIV-1 form, and HIV-2, which is less virulent or infectious, and is prevalent in West Africa.

HIV-1 is divided into three main groups based on the sequence of the entire viral genome. Group M ("major") accounts for more than 90 % of HIV/AIDS worldwide. Group N ("non-M, non-O") occurs only in Cameroon. Group O ("outlier") is not usually seen outside West-central Africa.

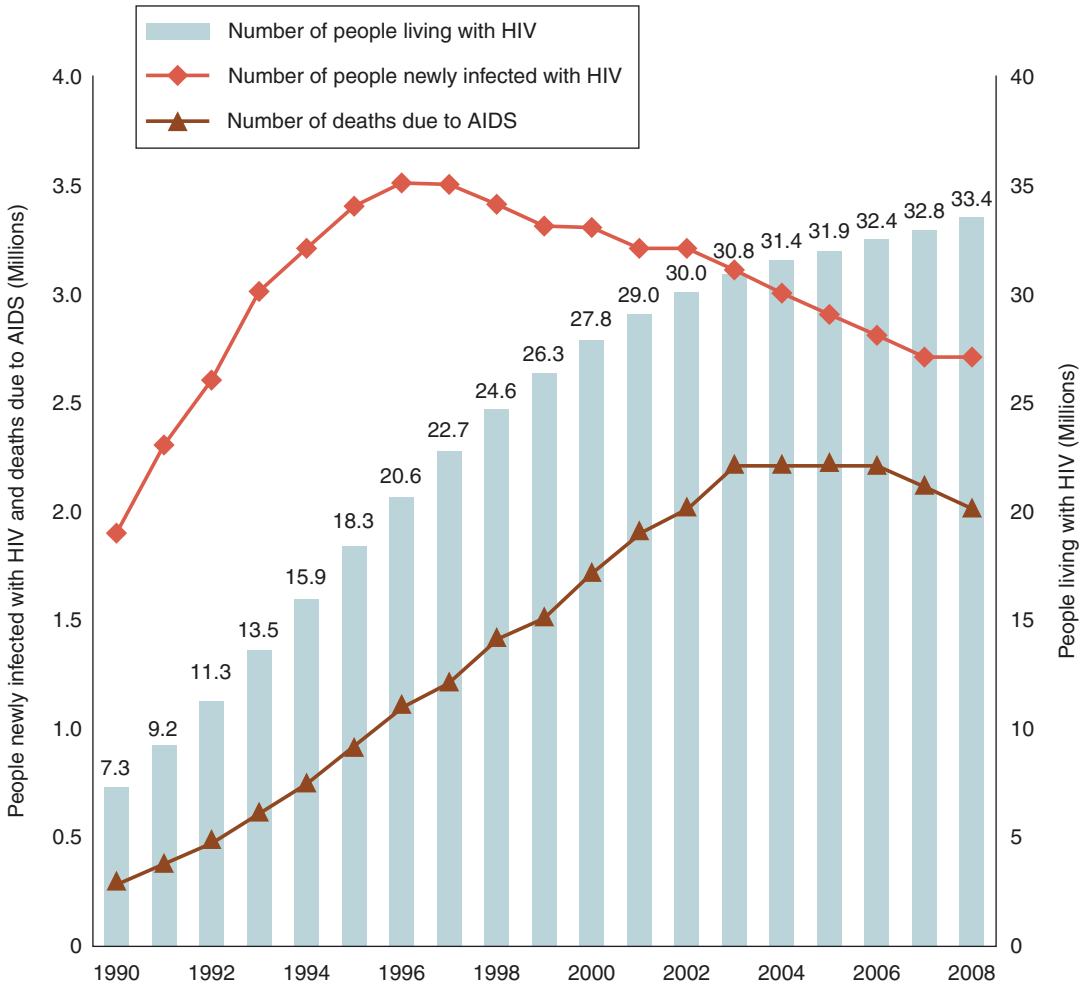
Group M is further classified into clades (subtypes), again based on the whole sequence. Clade A is common in West Africa. B is the dominant form in Europe, the Americas, Japan, Thailand, and Australia. C is the dominant form in Southern Africa, India, and Nepal, and accounts for nearly half of all worldwide HIV. There are also, Clades D, E, F, G, H, I, J, K. In addition, there are circulating recombinant forms (CRF), such as CRF02, which is clade A/G. These CRFs show genetic characteristics of more than one clade.

Since HIV has a very high genetic variability, it demonstrates huge diversity. This is a result of its fast replication cycle, generating about 10^{10} virions every day, and high mutation rate (about 3×10^{-5} per nucleotide base per cycle of replication). Thus, there are more individual variants of HIV in one infected individual than all variants of influenza in a population during an epidemic. It is this variation that makes it such a difficult agent to tackle with vaccines [10], and antiviral drugs [5, 9].

Genome and Structure

The diploid HIV-1 genome consists of two (+) RNA molecules of about 9 kilobases (kb) in length. These genomic RNAs are packaged into nascent HIV-1 particles and become reverse transcribed in infected cells to generate viral DNA for integration into the host genome. The genome encodes proteins, which are expressed from different spliced forms of a single RNA transcript, and are essential for infection. The same 9 kb HIV-1 proviral genome produces 3 mRNA transcripts encoding all the proteins necessary for the viral life cycle. As the transcripts are made from three different frames of the genome, the virus maximizes the information encoded in its genome.

The 55 kDa Gag polyprotein consists of the proteins MA (matrix), CA (capsid), spacer protein p2, NC (nucleocapsid), spacer peptide p1, and the p6 protein. Gag cleavage produces the individual proteins and occurs after virus assembly through maturation. Pol is made as a 160 kDa Gag-Pol protein, which is cleaved to produce Gag fused to protease (PR), reverse transcriptase (RT), and integrase (IN). The Env protein is cleaved into gp120, the surface protein (SU), and gp41, the transmembrane protein (TM). These proteins are essential for attachment to, and infection of, host cells. They exist together as a trimeric spike assembly. Additional auxiliary proteins



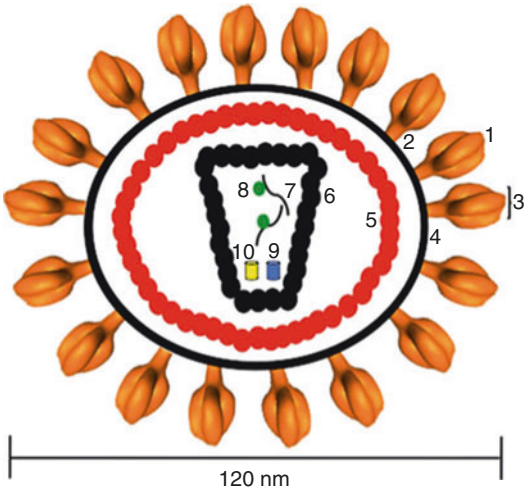
HIV Quantification, Fig. 1 Graph showing the number of people living with HIV, number of people newly infected with HIV, and number of AIDS deaths worldwide, 1990–2008 (Millions)

encodedare Vif (viral infectivity factor), Vpr (viral protein R), Vpu (viral protein unique), Rev (regulator of virion expression), Tat (transactivator of transcription), and Nef (negative factor). These serve additional functions to perpetuate the life cycle: Vif blocks innate antiviral proteins, Vpr enhances infectivity, whereas Nef and Vpu downregulate CD4.

An individual HIV-1 particle is near spherical and 100–120 nm in diameter. It consists of an outer membrane of infected host cell composition (acquired when budding from the host cell) impregnated with approximately 5–50 gp120/41 glycoproteins (for further infections) arranged as trimeric spikes. This surrounds a protein core of about 5000 p24 CA monomers arranged as a genome-housing cocoon, called the core. About 5000 MA proteins fill the space between the CA and membrane. The two RNA genomes are arranged in the core with about 250 copies each of IN, PR, RT, and variable quantities of auxiliary proteins [9] (Fig. 2).

Life Cycle

HIV-1 cell entry starts by trimeric spike binding to the susceptible cell's CD4 receptor. Further interaction with co-receptors (CCR5 or CXCR4) leads to fusion of the host cell and viral membrane,



HIV Quantification, Fig. 2 Schematic representation of a typical HIV-1 particle. *1* gp120 surface glycoprotein, *2* gp41 transmembrane glycoprotein, *3* trimeric spike (trimer of 1 + 2), *4* infected cell–derived viral membrane, *5* matrix protein, *6* p24 capsid core, *7* viral RNA genome (two copies), *8* reverse transcriptase enzyme, *9* integrase enzyme, *10* protease enzyme

involving the viral gp41. Following fusion, the CA core enters the cytoplasm and undergoes uncoating to reveal the viral genome. These two single-stranded (+) RNAs are copied into one strand of dsDNA by RT. The viral DNA is transported to the nucleus and is integrated into the host genome by the IN. Proviral transcripts encoding Rev and Tat are spliced, and their mRNAs are exported for protein synthesis in the cytoplasm. Imported back into the nucleus, these ensure efficient viral transcription and transcript export to the cytoplasm. The viral mRNAs are translated into Gag, GagPol, and Env proteins and full-length viral genome RNA is packaged into virions. Assembly and budding of new virus is mediated by host cell machinery (ESCRT-I/II/III) at the plasma membrane. Once released, HIV-1 undergoes a maturation process where protease can cleave Gag and produces cone shaped capsids within a glycoprotein/membrane cover for further infection [6].

HIV is of such interest because of the disease, AIDS, it causes. It is a disease of the immune system, such that infection leads to a progressive reduction in its effectiveness. Unchecked infections produce an ever-increasing susceptibility of the individual to opportunistic infections and tumors. Indeed, infections that would pose no threat to healthy immune systems are a potential risk to the life of those whose immune system has been compromised by HIV.

How HIV Causes Disease

Though it is widely known that HIV causes AIDS, the specifics of how it incapacitates the immune system is more specialized knowledge. It achieves this in a number of ways: primarily, since HIV infects CD4+ cells (helper T-cells, dendritic cells (DC), and macrophages), the effect on the immune system is mediated through the removal of the required amount of CD4+ cells to maintain efficient functionality. Macrophages and DCs function as, among other things, antigen-presenting cells (APCs). These internalize antigen (such as a virus), by phagocytosis or receptor-mediated endocytosis, then display a fragment of it bound to a class II major histocompatibility complex (MHC) molecule, on their membrane. Specific, naïve, CD4+ T-cells recognize and bind this complex and become activated, proliferate, and release cytokines. Later, the T-cell may bind a specific B-cell presenting the same antigen and, through a similar interaction, stimulate its conversion into an antibody-producing plasma cell. These antibodies bind to, and result in the destruction of, the virus. Thus, removal of the CD4+ cells removes the ability of antiviral (or any pathogen) antibody production.

HIV-infected cells release new virus and, as a result, this lyses and kills the cell. Thus, the specific infection allows the production of new virus, but also the destruction of the cell enabling its production. The integration of the viral genome into the host cell genome stimulates (auto) killer signals within the cell, and also downregulates anti-apoptotic signals, both promoting infected cell death. Infected cells may present viral fragments associated with MHCI. This presentation of viral component in infected cells causes cell destruction of those cells by cytotoxic CD8+ T-cells. The virus accessory proteins Nef and Vpu decrease CD4 levels in infected cells, rendering those cells ineffective APC mediators.

This effect presumably evolved to prevent other HIV particles in the infected host infecting an already infected cell.

It serves no purpose for an organism to evolve that infects another organism and kills it before the infecting organism can effectively survive. Hence, though HIV destroys cells it infects, it doesn't kill the organism for up to many months (even when non-treated), and during this time can pass to other individuals by the means described. Indeed, even if treated with drugs, or if the immune system is able to remove all serum virus (e.g., by CD8+ T-cell-mediated effects), the latent virus can be reactivated to restore serum levels and enable infection of other cells and individuals.

Prevention and Treatment

HIV differs from many viruses in that it evades the immune system by constantly changing the amino acid sequence of its glycoprotein envelope. This is essential for viral survival because the glycoprotein is on the external surface of the virus, and "visible" to the immune system, and is necessary for entry into susceptible cells.

The rapid replication of HIV, married with the high error rate, produces a large number of different particles. This results in particles whose glycoproteins differ too. Some particles will be ineffective at binding CD4 and infecting cells, rendering that virus replication–defective. Some may be replication-competent, but may present glycoproteins of a structure that are detected by neutralizing antibodies that have developed in the infected host. These antibodies block further virus–CD4 interaction by this virus. Some viruses may be produced with an envelope spike structure that the neutralizing antibody cannot recognize. These viruses can infect and result in the production of many viruses resistant to neutralization.

Thus far, the only effective neutralizing antibodies that have been produced have been isolated from HIV-infected individuals (e.g., 4E10). No comparably effective neutralizing antibodies have been produced from individuals immunized with components of an HIV particle. In effect, no effective vaccine has been produced [10].

Though clade A viruses are prevalent in Asia, C in SSA, and B in Europe and the USA, any vaccine would ideally be able to block multiclade viral infection. This is far from simple, in part due to the variability in the spike structure and the immune evasion techniques the virus has evolved.

HIV vaccine development differs from that of other classic vaccines for a number of reasons: classic vaccines mimic natural immunity against reinfection. AIDS patients tend not to recover. Classic vaccines tend to be dead whole organisms, or those that are attenuated. HIV-1 is not antigenic if killed, and a live virus vaccine is unsafe. Since HIV may remain latent for long periods before causing AIDS, it can present as a hidden target, unlike those of classic vaccines. High-risk individuals may encounter HIV every day, and in different forms, whereas classic vaccines tend to protect against rarely encountered pathogens, which are less variable in sequence. Finally, the site of infection that most classic vaccines protect against is the respiratory and gastrointestinal tract. HIV rarely uses these routes [10].

Interestingly, live attenuated simian immunodeficiency virus (SIV) vaccines can confer effective protection against wild-type SIV challenge. Therefore, mindful of the positives gained from the primate effectiveness of vaccines, it is essential that one understands the early immune response to HIV-1 to enhance the possibility of generating efficient human vaccine designs. It is estimated that after initial infection there are less than 10 days in which virus-infected cells can be eradicated. After that, the virus spreads, integrates, and generates non-eradicable reservoirs of latent virus. Thus, only if the virus is prevented from infecting *any* host cells can the individual evade a successful infection.

Immunogens based on the gp120 envelope antigen represent, logically, ideal vaccine targets. gp120 stimulates a humoral immune response, including weak neutralization of the HIV strain from which it is derived. However, vaccines based on this target might conceivably be worse than useless, as they could elicit antibodies that enhance infection by targeting macrophages and dendritic cells via components of the immune system (complement and Fc receptors)

[6]. Additionally, MHCII presentation of the vaccine gp120 would result in, from the perspective of preventing HIV infection, unwanted induced proliferation of the HIV target CD4+ T-cells. However, the spike is the ideal vaccine target since it is the one whose blockage would prevent initial infection and the production of latent viral reservoirs.

Once HIV infected, the individual's prognosis is determined, somewhat, by antiretroviral drug therapy intervention. These can be broadly classified by the phase of the viral life cycle that the drug addresses. Nucleoside and nucleotide reverse transcriptase inhibitors (NRTI) inhibit reverse transcription by incorporation into newly synthesized viral DNA. This prevents further elongation. Non-nucleoside reverse transcriptase inhibitors (NNRTI) bind and inhibit RT directly. Protease inhibitors (PIs) bind and inhibit protease, which is essential for the assembly of new virions through its function of cleaving nascent viral proteins. Another target is integrase. Integrase inhibitors bind and inhibit the enzyme, preventing integration of viral DNA into the infected cell genome. Entry (or fusion) inhibitors block viral binding to target cells. Maturation inhibitors act at the stage of the viral life cycle where the capsid polyprotein is cleaved into mature capsid protein (p24). Thus, the core assembly is blocked, resulting in primarily non-infectious particles.

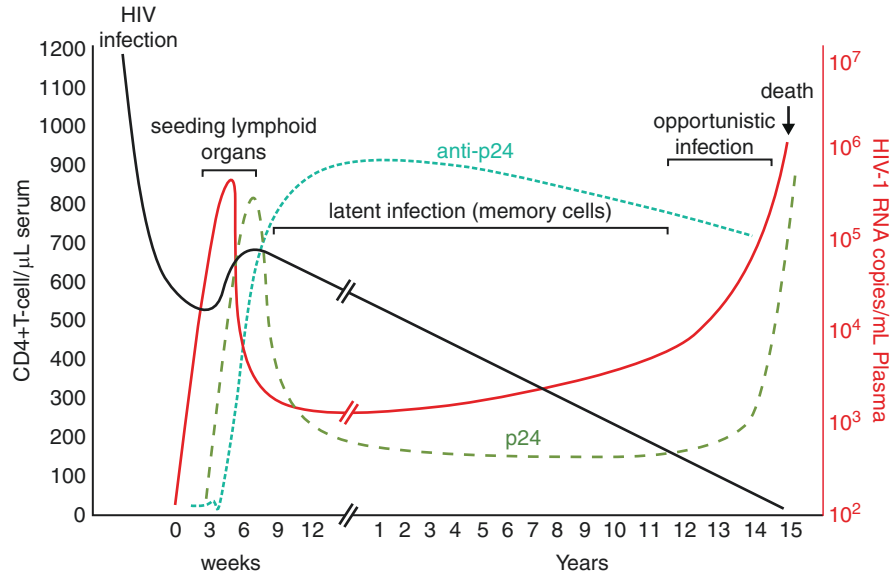
Highly active antiretroviral therapy (HAART) is the name given to the combination of antiretroviral drugs used to treat HIV infection. In the USA, there are different regimens (2009): an NRTI and NNRT1 are given in combination, or an NRTI and an integrase inhibitor or protease inhibitor. The choice and combination of the increasing number of drugs to choose from is often a combination of national guidelines together with the specific doctor-patient case. However, in resource-limited settings (such as SSA), the World Health Organization (WHO) has defined that HIV-infected adults should start drug therapy when HIV infection has been confirmed and one of the following conditions is present:

- There is clinically advanced HIV disease
- There is WHO Stage IV HIV disease (irrespective of the CD4+ T-cell count)
- There is WHO Stage III disease (consider CD4 + T-cell counts to assist decision – if less than 350/μL (mm^3) serum then start therapy)
- There is WHO Stage I or II HIV disease (CD4 T-cell counts less than 200/μL serum)

Of the 35+ million people infected with HIV, only 1.3 million of these people receive antiretroviral treatment. Drug treatment and diagnostics are closely linked. The decision on any rational treatment regimen starts when particular markers of infection (e.g., CD4+ T-cell count) reach a certain value. Diagnostic techniques determine these values. HAART drug regimens cannot completely eradicate HIV infection (latent infection is impervious to drugs). HAART, if adhered to, and with appropriate lifestyle choices (if applicable), can suppress HIV to levels below detection levels of standard clinical assays. This can be maintained for many years. Yet, if HAART is withdrawn, virus levels quickly rise and CD4+ T-cell levels fall (see Fig. 3), leading to AIDS.

Memory cells are a subset of CD4+ T-Cells. An HIV-infected individual, even if on successful HAART, is likely to possess memory cells with genome-integrated HIV. These long-lived cells possess the latent virus for the duration of their lifetime (many years, even decades). Thus, an infected individual could, theoretically, become free from HIV with effective HAART treatment until the latent memory cell population died. This, however, is unlikely.

The prognosis of an infected individual depends on many factors: HIV subtype, treatment, time treatment starts after infection, immune system fitness, age, location, resources, and so on. However, without treatment, the median survival time after infection with HIV is about 10 years, and the median survival rate after diagnosis of AIDS in resource-limited settings where treatment is not available ranges between 6 and 19 months. HAART has reduced the HIV-related death rate by 80 %, and raised life expectancy for newly diagnosed HIV-infected person to 20–50 years.



HIV Quantification, Fig. 3 The clinical determinant for antiretroviral therapy is the CD4+ T-cell count and the viral load. CD4+ T-cells are quantified by FACS, and virus by the genome levels. Other components of HIV infection are also measured. Serum viral p24 protein, and anti-24 antibody levels in serum are measured as indicators of the extent of, and response to, infection

Nanotechnological Methodologies for HIV Detection/Quantification

In addition to vaccine and antiretroviral drug development, there is a powerful need for cheap and accurate HIV diagnostic sensors. Nanotechnological methods are increasingly associated with designs of such devices. To fit all potential markets (e.g., hospitals, doctor's surgeries, and developing countries), such devices should rapidly, and accurately, diagnose and monitor HIV in resource-limited environments. In developing countries electricity may not be available. Thus, battery (such as mobile phone batteries), or solar-powered devices are advantageous.

The need for accurate diagnostics married to limited access to fully functional diagnostic clinics is driving the development of point-of-care (POC) technologies for early diagnostics. Early diagnosis reduces the risk of onward transmission from people who are not aware of their infection, and provides important information on the dynamics of epidemic, transmission networks to guide public health intervention programs. In the developing world, antiretroviral drugs are used without monitoring, leading to resistance and compromising the full potential of such drugs. The development of affordable, portable, HIV diagnostics, and smart monitoring devices, will benefit millions of HIV-infected individuals.

The main clinical virological marker is "viral load" (copies of RNA/mL plasma); p24 is another. The immunological markers are HIV-specific antibodies (against gp120 and p24). The CD4+ T-cell count is the main immunological marker against which drug therapy is judged, since it is an indicator of the immune system's response to infection (see Fig. 3).

Currently, POC enzyme-linked immunosorbent assay (ELISA) tests identify HIV infection through blood from thumb pricks, and can be used in non-hospital settings. These are insensitive and lack specificity. Well-funded laboratories perform HIV diagnoses by HIV antibodies and p24 in serum, to ensure detection of the early stages (first 2–4 weeks) of infection when p24 antigen is present. POC tests are limited to detection of antibody (at present).

HIV viral load is measured using real-time polymerase chain reaction (RT-PCR). For the CD4+ T-cell count, fluorescent-activated cell

sorting (FACS) and microscopy are used. Both require sophisticated instrumentation, sample preparation, amplification steps, and experienced personnel. This is expensive, so monitoring is restricted to tertiary centers.

Point-of-care HIV devices have not yet been implemented, both because of the apparent poor compatibility of gold standard methodologies (RT-PCR, FACS, ELISAs, each requiring specialized reagents and instrumentation), and because of the significant challenges imposed by resource-limited environments: tests should be robust, simple to operate, require a minimum sample volume, yet provide a rapid answer. There should be a low probability of false positive or false negative results. The device and reagents should tolerate variable storage conditions (certain settings may not possess refrigerators and freezers), yet be affordable. Microfluidic PCR systems have been developed. Devices which perform on-chip RT-PCR (incubation with RT, PCR components, and thermocycling) and chemiluminescent detection of the amplicon, all carried out within 1 h have been demonstrated [11]. However, such devices are relatively expensive. Certain microfluidic PCR platforms under development, using isothermal amplification, do not require thermocycling or the use of RT. So far, they have been qualitative in their amplification. The use of integrated optical detection could aid quantification and robustness of these devices as well as reduce cost. There is scope for an HIV context here.

Since the gold standards for evaluating HIV infection are the viral load (measured by qPCR of the viral genome), and flow cytometric counting of CD4+ T-cells, even if a POC device does not measure these specific virological and immunological aspects of an HIV infection, the aspects they *do* measure should quantitatively correlate to these values, such that the data obtained will be as useable in an HAART response context. For example, if one chooses to use the viral envelope spike to quantify viruses in the POC device, it must give readouts that are quantitatively correlative to the value that would be obtained if one were to perform qPCR on the same sample.

In addition to FACS, ELISA, and qPCR, current laboratory diagnostic techniques used for HIV status measurements include western blotting (a type of ELISA, but which can signify the target protein's mass). Simple POC lateral flow tests are available; these are chromatography techniques that rely on capillary action to drive the flow of sample. These are simple devices, which use strips with an absorbent layer coated with specific antibodies. A small drop of fluid sample causes antibody-antigen pairs to diffuse to a detection zone, providing an optical indication and a diagnosis. Though cheap and simple, their use for HIV is limited, especially in terms of sensitivity and quantification. Saliva and urine tests are also used, but carry the disadvantages of ELISAs and lateral flow tests, as well as the fact that HIV levels are low in these fluids compared to serum [11].

The recent advances in microelectronics have delivered chips with nanometer-sized features comparable with a single HIV virion (*circa* 120 nm). This convergence in the size of natural "bottom-up" self-assembled systems and "top-down" fabricated objects offers tremendous opportunities to probe pathogens with unprecedented sensitivity.

Cross-References

- ▶ Biosensors
- ▶ Microfluidic Whole-Cell Biosensor
- ▶ Nanoparticles
- ▶ Nanotechnology
- ▶ On-Chip Flow Cytometry

References

1. Knipe, D.M., Howley, P.M. (eds.): Fields Virology, 5th edn. Lippincott Williams & Wilkins, Philadelphia (2006)
2. Coffin, J.M., Hughes, S.H., Varmus, H.E. (eds.): Retroviruses. Cold Spring Harbor Laboratory Press, New York (1999)
3. Emini, E. (ed.): The Human Immunodeficiency Virus: Biology, Immunology and Therapy. Princeton University Press, Princeton (2002)
4. Coffin, J., Haase, A., Levy, J.A., Montagnier, L., Oroszlan, S., Teich, N., Temin, H., Toyoshima, K.,

Varmus, H., Vogt, P., Weiss, R.: What to call the AIDS virus? Nature **321**(6065), 10 (1986)
5. Ho, D.D., Bieniasz, P.D.: HIV-1 at 25. Cell **133**, 561–565 (2008)
6. Weiss, R.A.: Gulliver's travels in HIVland. Nature **410**, 963–967 (2001)
7. UN AIDS Epidemic Update (December 2007)
8. UN Millennium Development Goals Report. (2010)
9. Frankel, A.D., Young, J.A.T.: HIV-1: fifteen proteins and an RNA. Annu. Rev. Biochem. **67**, 1–25 (1998)
10. McMichael, A.J., Borrow, P., Tomaras, G.D., Goonetilleke, N., Hughes, B.F.: The immune response during acute HIV-1 infection: clues for vaccine development. Nat. Rev. Immunol. **10**, 11–23 (2010)
11. Desai, D., Wu, G., Zaman, M.H.: Tackling HIV through robust diagnostics in the developing world: current status and future opportunities. Lab Chip **11**, 194–211 (2010)

Hollow Gold Nanoshells

► Hollow Gold Nanospheres

Hollow Gold Nanospheres

Chun Li[1] and Jin Z. Zhang[2]
[1]Department of Experimental Diagnostic Imaging-Unit 59, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[2]Department of Chemistry and Biochemistry, University of California, Santa Cruz, CA, USA

Synonyms

HGN; Hollow gold nanoshells

Definition

Spherical gold nanoparticles with a hollow interior. This class of photothermal conducting nanomaterials has found applications in photothermal ablation therapy, drug delivery, photoacoustic imaging, and surface-enhanced Raman scattering spectroscopy.

Key Principles

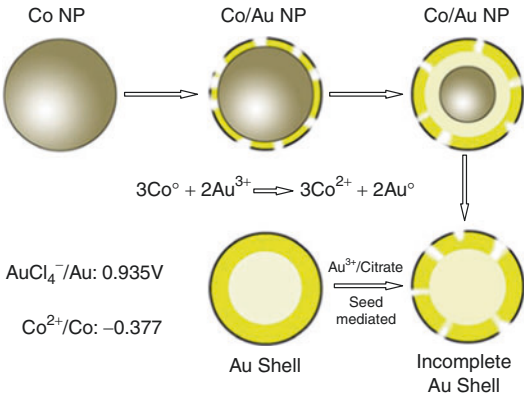
Noble metal (e.g., gold and silver) nanomaterials possess novel optical and electronic properties useful for various potential biomedical applications, including sensing, imaging, drug delivery, and therapy. In particular, their strong surface plasmon resonance (SPR) absorption and high photothermal conversion efficiency are well suited for photothermal ablation (PTA) therapy application. Ideal metal nanostructures for PTA should not only have strong and tunable SPR, especially in the near-infrared (NIR) range, for better tissue penetration but also be easy to deliver, have low toxicity, and be amenable to bioconjugation for actively targeting specific cancer cells. Several different types of gold nanostructures have been utilized for PTA therapy, including nanorods, nanoshells, nanocages, nanoprisms, and hollow nanospheres. Among these structures, hollow gold nanospheres (denoted as HAuNS or HGNs) represent second-generation gold nanostructures with unique structural features, optimal optical properties, and efficient photothermal profiles. These HAuNS are utilized in a variety of applications, including as photothermal-converting agents for PTA therapy, as contrast agents for photoacoustic imaging, as substrate for surface-enhanced Raman scattering (SERS), and as drug carriers for light-activatable delivery of anticancer agents and small interfering RNA (siRNA).

HAuNS have the unique combination of small size (outer diameter 30–50 nm), spherical shape, hollow interior, and a strong and tunable absorption band at a NIR-region wavelength as a result of their highly uniform structure. These properties make HAuNS ideally suited for targeted delivery after systemic administration. Small size is important because smaller particles extravasate more easily out of tumor blood vessels. The optimal size for intracellular uptake is about 50 nm for gold nanoparticles. Strong light absorption and a narrow absorption band are important for PTA because these properties directly affect heat-conversion efficiency. Tunable absorption is useful for matching the selected laser wavelength and minimizing the absorption of NIR light by normal tissues.

Synthesis and Characterization

HAuNS are synthesized by the cobalt nanoparticle–mediated reduction of chloroauric acid (Fig. 1) [1]. Briefly, cobalt template nanoparticles were synthesized by adding 2500 μL of sodium borohydride (1 M) into 1000 mL of deoxygenated aqueous solution of sodium citrate (2.8 mM) and cobalt chloride (0.4 mM). The slightly pink solution turned brown upon the addition of sodium borohydride, indicating the formation of cobalt nanoparticles. The solution was allowed to stand at room temperature for 45 min under constant argon flow until complete hydrolysis of the sodium borohydride. In a separate vial, 600 mL of an aqueous solution of chloroauric acid (0.04 mM) was deoxygenated by bubbling with argon gas. Subsequently, 800 mL of cobalt nanoparticle solution was transferred to the chloroauric acid solution under argon protection and magnetic stirring. The cobalt immediately reduced the chloroauric acid to gold ions that adhered to the surface of cobalt nanoparticles, while at the same time, it was oxidized to cobalt oxide. Any remaining cobalt core was further oxidized by air, resulting in the final formation of the HAuNS.

HAuNS synthesized from sodium citrate are coated with citric acid and are negatively charged. Polyethylene glycol (PEG) is normally conjugated to the surface of HAuNS to improve the colloidal stability of the nanoparticles. A transmission electron microscopy (TEM) image of PEG-HAuNS shows that the nanoparticles had a mean outer diameter of 43–46 nm and a mean shell thickness of 2–4 nm (Fig. 2a). The resonance absorbance of these nanoparticles was tuned to peak at 800 nm (Fig. 2b). According to the Mie theory and the plasmon resonance theory [2], the optical absorption coefficient of gold nanoshells increases, while the optical scattering coefficient decreases, with decreasing diameter of the core. The peak wavelength experiences a red shift (increase in wavelength) with decreasing shell thickness. Figure 2c shows the calculated optical absorption and scattering coefficients of HAuNS

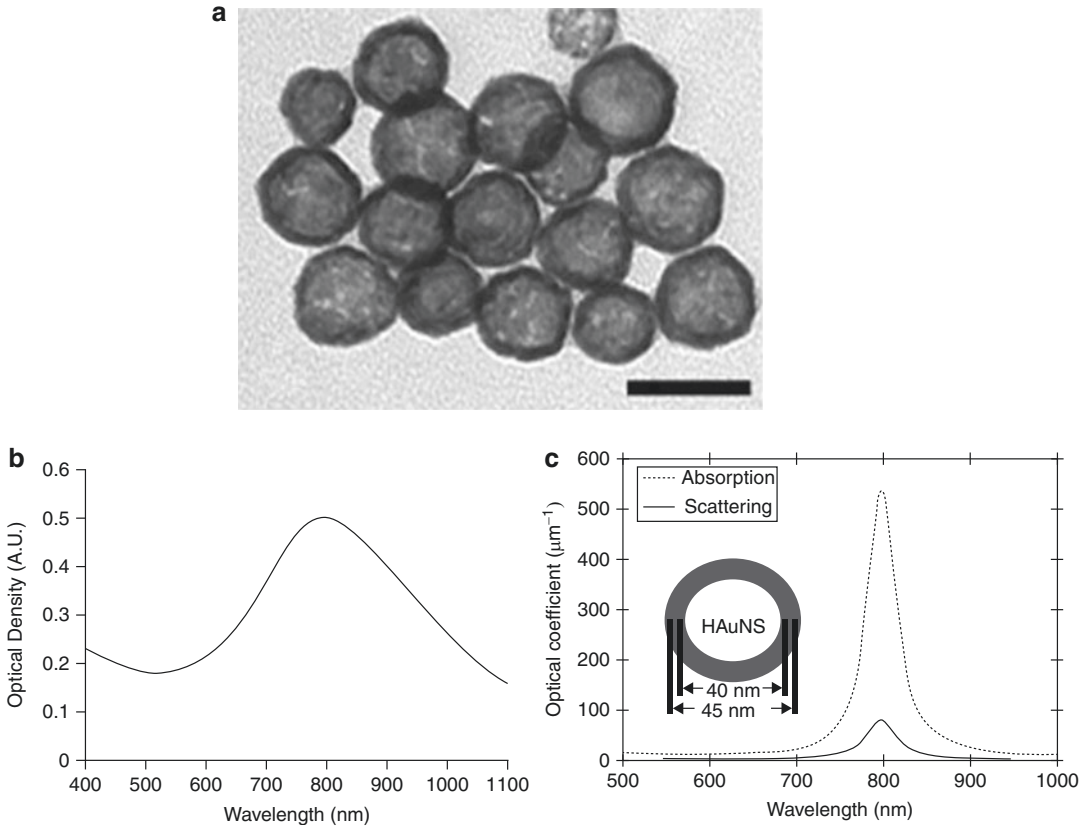


Hollow Gold Nanospheres, Fig. 1 Schematic illustration of the synthesis scheme, in which Co metal nanoparticles (*NP*) are used as a template for making HAuNS via the redox reaction shown in the *middle* of the figure

(i.e., absorption and scattering cross sections per unit volume). With an outer diameter of 45 nm and shell thickness of 2.5 nm, HAuNS have an absorption coefficient of ~540 μm^{-1} at 800 nm. The extinction of HAuNS is dominated by absorption, owing to their small size. The theoretical dimensions of HAuNS match well with the measured dimensions shown in the TEM image (Fig. 2a).

Femtosecond (fs) hole-burning studies have been carried out to understand how the HAuNS structure can be affected by pulsed laser excitation, which is important for the study of HAuNS as potential drug carriers with light as a trigger for release [3]. It was found that upon irradiation with an fs laser in the NIR (e.g., ~800 nm), HAuNS can be affected by two possible processes: (1) conversion of NIR-absorbing HAuNS to bluer light–absorbing HAuNS and (2) breakdown of NIR-absorbing HAuNS directly into solid Au nanoparticles. The branching ratio between these two processes depends on the peak power of the fs laser pulses, with a higher peak power favoring the second process.

The surface chemistry of Au is extremely versatile, which results in the formation of Au-S bonds when gold nanoparticles are mixed with molecules containing a mono-sulfhydryl (SH) group or molecules with multiple thiol



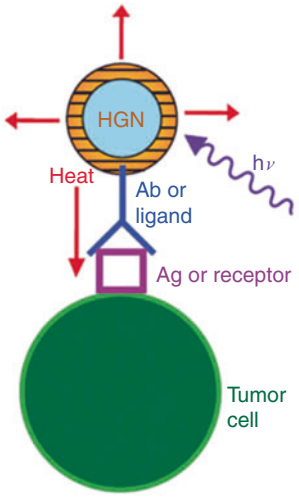
Hollow Gold Nanospheres, Fig. 2 (**a**) TEM image of PEG-HAuNS (Bar = 50 nm). The average outer diameter was 45 nm, and thickness of the shell was 2.5 nm. (**b**) Experimental absorbance spectrum of PEG-HAuNS in water, which peaked at 800 nm. (**c**) Theoretically calculated absorption and scattering spectra of PEG-HAuNS with water core having a dielectric permittivity of 1.77. The inner core radius was 20 nm, and the gold shell thickness was 2.5 nm, which are consistent with the values shown in (**a**) (Reproduced with permission from Ref. [12])

groups that act as anchoring ligands. HAuNS can also be coated with molecules attached to thioctic acid (TA), also known as alpha-lipoic acid, a naturally occurring compound [4]. TA is particularly well suited for gold surface modification because as a cyclic disulfide-containing compound, it gives added stability by providing two gold-sulfur bonds per molecule in surface-functionalized gold. Various molecules have been successfully attached to HAuNS, including the radiometal chelation agents diethylenetriamine pentaacetic acid (DTPA) and 1,4,7,10-tetraazacyclododecane-1,4,7,10-tetraacetic acid (DOTA), fluorescent dye, and siRNA, owing to the Au-S chemistry [5–7].

Applications

Targeted Photothermal Ablation Therapy

The basic principle behind photothermal ablation therapy (PTA) is illustrated in Fig. 3. HAuNS target cancer cells via specific interactions, such as antibody-antigen or ligand-receptor binding. Upon absorption of NIR radiation, HAuNS quickly release heat that destroys the cancer cells bound to them locally. Strong optical absorption and quick and highly efficient photothermal conversion at the cancer site are critical to the high efficacy of PTA. Mechanisms of heat-induced cell death that have been proposed in previous studies include heat shock of proteins, inactivation of



Hollow Gold Nanospheres, Fig. 3 Illustration of the principle of PTA. HAuNS target cancer cells via antibody (*Ab*)-antigen (*Ag*) interaction. Quick release of local heat by HAuNS upon absorption of light destroys the cancer cells bound to the HAuNS

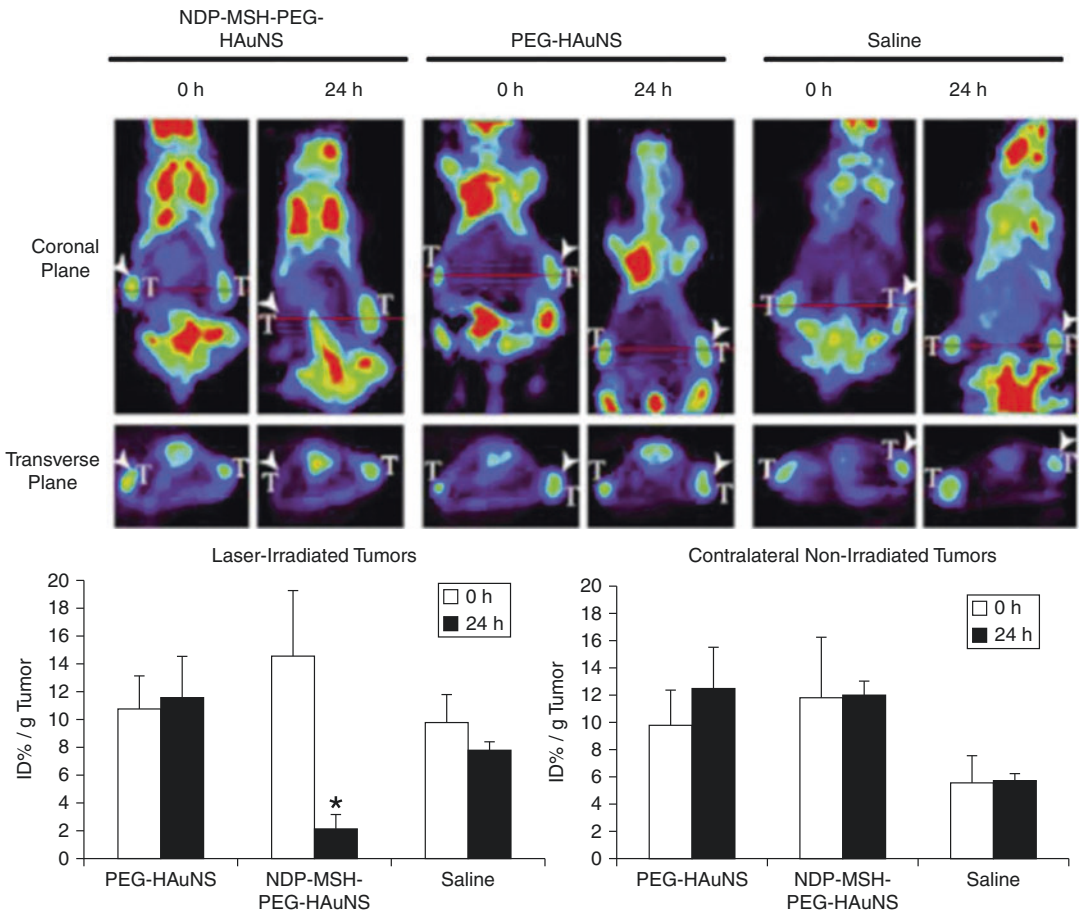
membrane proteins and mitochondria, depolymerization of cytoskeletal filaments, membrane blebbing, and plasma membrane disruption due to cavitation effect, which was also suggested to be responsible for bubble formation. The high effectiveness of PTA indicates that the local temperature at or near the metal nanostructure surface is likely much higher than the average temperature usually measured using thermocouples. Photothermal coupling agents can increase the efficiency of PTA, reducing the required energy dose of laser light and thereby lowering the potential for damage to surrounding normal tissues. Among the candidates for photothermal coupling agents are gold nanostructures.

One of the major challenges in using gold nanostructures for PTA is how to maximize the efficiency of their in vivo delivery to the target sites after systemic administration. The feasibility of tumor targeting using HAuNS has been demonstrated by studies in which HAuNS were coated with antibodies, [7] peptides, [5], and small-molecular-weight organic compound [6]. To demonstrate the feasibility of using HAuNS to target epidermal growth factor receptor (EGFR)–expressing cells, HAuNS were covalently conjugated to an antibody directed against EGFR, cetuximab (C225) [7]. The resulting C225-HAuNS conjugate was selectively bound to A431 human squamous carcinoma cells that overexpress EGFR. Selective cell killing was observed in the cells treated with C225-HAuNS and NIR laser but not in the cells treated with C225-HAuNS alone, NIR laser alone, or IgG-HAuNS plus NIR laser. C225-HAuNS had significantly higher uptake in the vicinity of tumor microvessels than did IgG-HAuNS control after intravenous injection into nude mice bearing subcutaneous A431 tumors. Because monoclonal antibodies are large-molecular-weight compounds, dispersion of C225-HAuNS into the tumor mass after the nanoparticles were extravasated outside the tumor vasculature was limited, and the difference in uptake between C225-HAuNS and IgG-HAuNS in A431 tumors was not statistically significant. In contrast, conjugating HAuNS to a small-molecular-weight peptidyl analog of α-melanocyte-stimulating hormone, [Nle4,D-Phe7]α-MSH (NDP-MSH, molecular weight 1769 Da) targeting melanocortin type-1 receptor (MC1R), led to a remarkably high tumor uptake value (~12.6%ID/g) (Fig. 4). Successful PTA of melanoma in mice was demonstrated using a laser energy dose as low as 30 J/cm^2 with HAuNS targeted to MC1R, [5] supporting the use of HAuNS as an efficient, targetable photothermal conducting agent.

A number of in vitro studies, including TEM and cell viability studies, have been conducted to better understand the fundamental interactions between HAuNS and cells. For example, in the in vitro PTA study of melanoma cells, the HAuNS were selectively taken up by B16/F10 cells via MC1R receptor-mediated endocytosis [5]. The conjugate NDP-MSH-PEG-HAuNS coated with an MC1R agonist NDP-MSH peptide was critical for selectivity (Fig. 5).

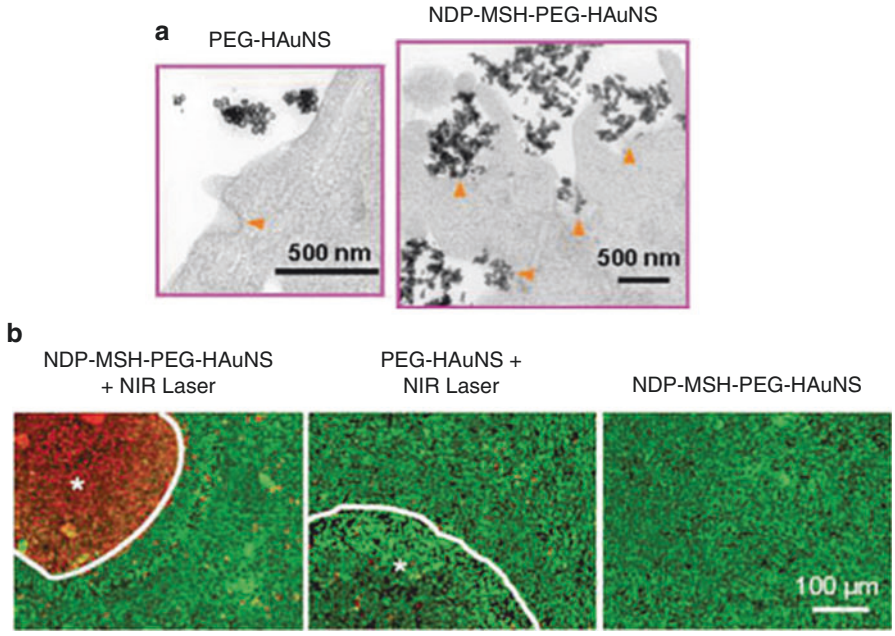
Drug Delivery

HAuNS-mediated temperature elevation may be used to modulate the release of cytotoxic and genetic agents contained in HAuNS and



Hollow Gold Nanospheres, Fig. 4 In vivo photothermal ablation with targeted NDP-MSH-PEG-HAuNS induced selective destruction of B16/F10 melanoma in nude mice. [^{18}F]fluorodeoxyglucose positron emission tomography (*PET*) imaging shows significantly reduced metabolic activity in tumors after photothermal ablation in mice pretreated with NDP-MSH-PEG-HAuNS, but not in mice pretreated with PEG-HAuNS or saline. [^{18}F]fluorodeoxyglucose PET was conducted before (0 h) and 24 h after near-IR laser irradiation (0.5 W/cm^2 at 808 nm for 1 min), which was commenced 4 h after i.v. injection of HAuNS or saline. *Arrowheads* in the figure indicate tumors (*T*) irradiated with near-IR light. [^{18}F]fluorodeoxyglucose uptake (%ID/g) before and after laser treatment is shown graphically at the *bottom*. Bars, SD ($n = 3$). *, $P < 0.01$ for %ID/g posttreatment versus %ID/g pretreatment (Reproduced with permission from Ref. [5])

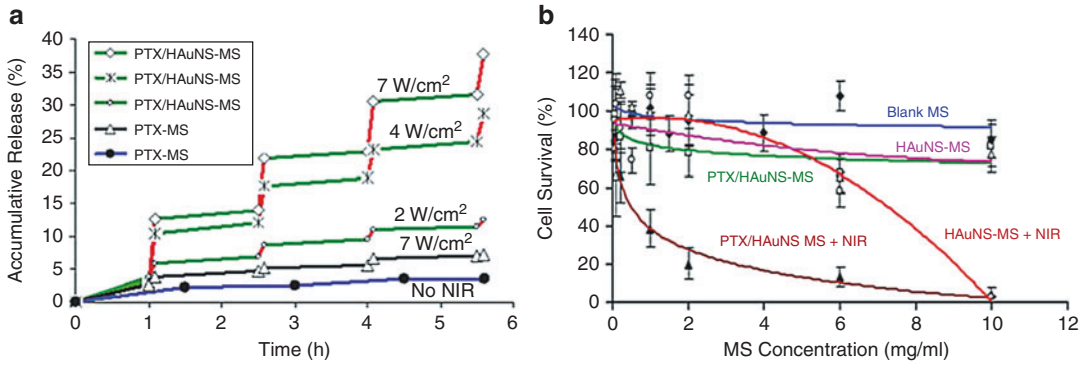
microspheres [8, 9]. To evaluate the feasibility of NIR light–modulated drug release from microspheres, biocompatible, biodegradable poly(lactic acid-co-glycolic acid) (PLGA) microspheres containing paclitaxel (PTX) and HAuNS (PTX/HAuNS microspheres) were prepared using a modified water-in-oil-in-water (W1/O/W2) double-emulsion solvent evaporation method. The resulting PTX/HAuNS microspheres had the same concentration as HauNS alone and so were able to elevate the temperature of water to the same extent, indicating that the PLGA polymer matrix did not prevent NIR light from reaching HAuNS embedded in the microspheres. Significantly, PTX released from these PLGA microspheres could be modulated by controlling the laser output power and the HAuNS content in the microspheres. Figure 6a shows the characteristics of PTX release from PTX/HAuNS microspheres and from microspheres containing PTX alone using various NIR laser output powers. The microsphere solution (10 mg/ml)



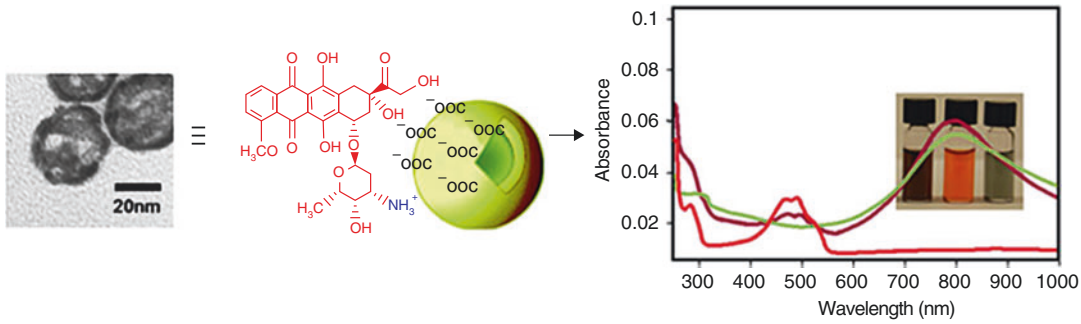
Hollow Gold Nanospheres, Fig. 5 (**a**) Transmission electron micrographs of B16/F10 melanoma cells showing selective cellular uptake of NDP-MSH-PEG-HAuNS. NDP-MSH-PEG-HAuNS were seen in coated pits (*arrowheads*) and endosomes (data not shown). Non-targeted PEG-HAuNS were found only outside the cell membrane. (**b**) Cell viability after NIR irradiation. B16/F10 cells were treated with NDP-MSH-PEG-HAuNS plus NIR light (808 nm, 32 W/cm^2, 3 min) (*left*), PEG-HAuNS plus NIR laser (*middle*), or NDP-MSH-PEG-HAuNS alone (*right*). Viable cells were *stained green* with calcein AM; dead cells were *stained red* with EthD-1. Only in the population of cells treated with NDP-MSH-PEG-HAuNS plus NIR laser were most cells dead. *Circled areas* labeled with *asterisks* represent laser-irradiated areas (Reproduced with permission from Ref. [5])

was irradiated repeatedly using an NIR laser over a period of 5 min, followed by a 1.5-h interval with the laser turned off. Rapid release of PTX from PTX/HAuNS-MS was observed when the NIR laser was on, whereas little PTX was released when the laser was turned off. The percentage of accumulated release of PTX over the course of the experiment approached 40 % when PTX/HAuNS-MS were irradiated using the NIR laser four times for 5 min each. In comparison, when PTX-MS was subjected to the same treatment protocol, the percentage of accumulated release of PTX was less than 7 % over the entire experimental period. Unirradiated PTX-MS had a 3.6 % accumulated release of PTX. These data indicate that the release of PTX from PTX/HAuNS-MS can be triggered by irradiation with an NIR laser, with excellent temporal and spatial control.

A significant difference in the cell-killing effect of PTX/HAuNS-MS with and without NIR irradiation was found (Fig. 6b). For example, 86.5 % and 67.4 % of the cells were viable after 72 h of incubation with PTX/HAuNS-MS alone at microsphere concentrations of 1 mg/ml and 10 mg/ml, respectively. However, when PTX/HAuNS-MS was combined with NIR irradiation, only 38.8 % and 3.9 % of the cells survived treatment at microsphere concentrations of 1 mg/ml and 10 mg/ml, respectively. Furthermore, about 90 % of the cells were viable after treatment with MS containing HAuNS alone and no PTX (HAuNS-MS) at the low concentration range (0.05–4.0 mg/ml) regardless of NIR exposure. At the high concentration range (6.0–10.0 mg/ml), combined treatment with HAuNS-MS and the NIR laser resulted in a significant cell-killing effect (Fig. 6b), which can be



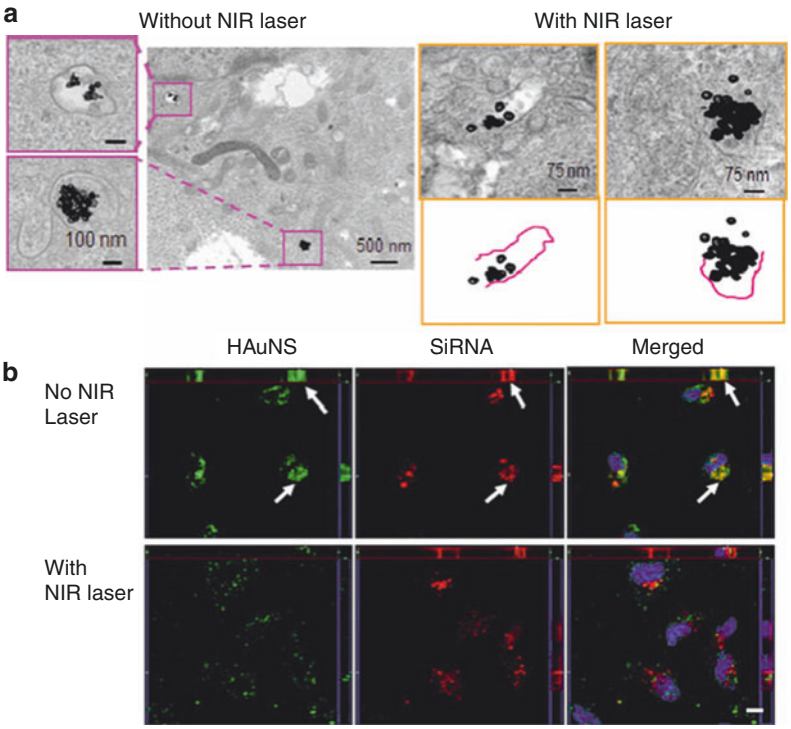
Hollow Gold Nanospheres, Fig. 6 *NIR light triggered release of PTX from PTX/HAuNS-MS. Red lines* represent the 5-min period during which the microspheres were irradiated with NIR laser. (**a**) PTX release profiles from PTX/HAuNS-MS containing 2.0×10^9 HAuNS/mg polymer at NIR output power of 7 W/cm^2 (), 4 W/cm^2 (), and 2 W/cm^2 (o), and PTX release profiles from PTX-MS (without HAuNS) at NIR output power of 7 W/cm^2 (Δ) and without NIR exposure (●). (**b**) *Cytotoxicity in the presence and absence of NIR irradiation*. MDA-MB-231 breast tumor cells were incubated with various MS formulations for 72 h. For NIR treatment, cells were irradiated with NIR laser four times at an output power of 2 W/cm^2 for 3 min each. Cancer cells were killed primarily by PTX released from PTX/NS-MS triggered by an NIR laser at low MS concentration, while efficient cell killing was achieved with combined PTX/HAuNS-MS and NIR irradiation at higher MS concentrations mediated by both increased free PTX released from the MS and increased photothermal effect (Reproduced with permission from Ref. [8])



Hollow Gold Nanospheres, Fig. 7 Hollow gold nanospheres with strong surface plasmon absorption in the near-infrared region formed stable complexes with doxorubicin, with an extremely high payload (Reproduced with permission from Ref. [9])

attributed to the photothermal effect of HAuNS-MS at high microsphere concentrations owing to increased local temperatures. These results demonstrate that the combination of photothermal effect and chemotherapy is more efficient than thermal ablation alone in killing tumor cells over a wide range of microsphere concentrations.

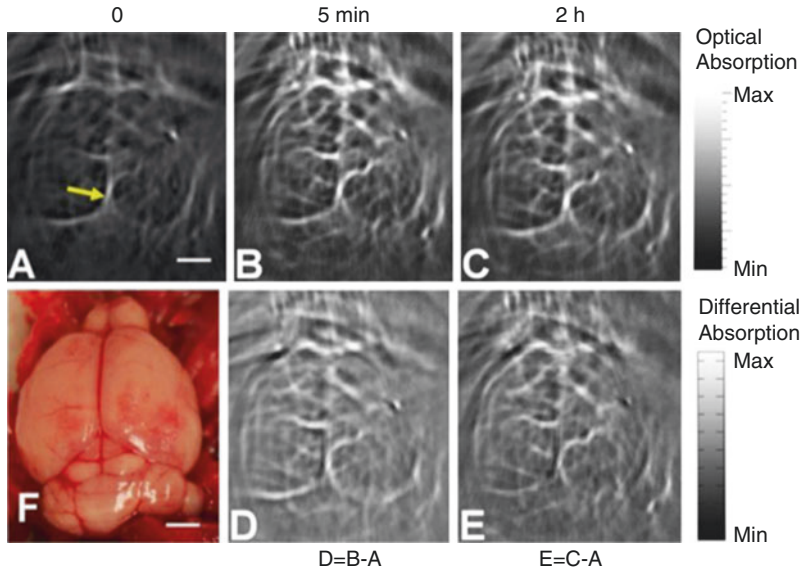
In addition to the comixing of HAuNS and drugs in the matrix of microspheres, anticancer drugs can also be directly incorporated into HAuNS and their release modulated with NIR light. For example, as high as 63 % doxorubicin (DOX) by weight (~1.7 μg DOX/μg Au) could be loaded onto PEG-coated HAuNS because DOX adhered to both the outer and the inner surfaces of HAuNS (Fig. 7). Irradiation with an NIR laser induced photothermal conversion, which triggered rapid DOX release from the DOX-loaded HAuNS [9]. The release of DOX was also pH dependent, with more DOX released in aqueous solution at a lower pH. Similar to microspheres containing HAuNS and PTX, significantly greater



Hollow Gold Nanospheres, Fig. 8 Endolysosomal escape of HAuNS-siRNA after NIR irradiation. (**a**) TEM images show F-HAuNS-siRNA nanoparticles located in the intact endocytic vesicles in HeLa cells not exposed to NIR light. After laser treatment, parts of the vesicle membranes disappeared, resulting in nanoparticle escape. (**b**) Z-stack images show dissociation of $siRNA_{NF-\kappa B}$ from HAuNS after laser irradiation. In cells without NIR irradiation, the siRNA colocalized with HAuNS (*arrows*). *Red*, Dy547-labeled siRNA; *green*, scattering signal from HAuNS visualized under a dark field; *blue*, cell nuclei counterstained with DAPI. Bar, 10 μm (Reproduced with permission from Ref. [6])

cell killing was observed when MDA-MB-231 cells incubated with DOX-loaded HAuNS were irradiated with NIR light, which was attributable to both HAuNS-mediated photothermal ablation and the cytotoxicity of released free DOX.

The excellent photothermal-conducting properties of HAuNS also provide a unique opportunity for mediating controlled release of siRNA**C** [6]. Termed "photothermal transfection," this novel siRNA delivery technology involves three steps: (1) internalization of siRNA-coated HAuNS through receptor-mediated endocytosis; (2) selective thermal disruption of the endolysosomal membrane and escape of the siRNA-coated HAuNS from endolysosomes upon NIR laser irradiation; and (3) detachment of siRNA from the nanoparticle carrier and release of free siRNA into the cytosol. Unlike traditional nanoparticulate polyelectrolyte-mediated RNAi, NIR laser-induced RNAi is confined to the illuminated area, which should reduce off-target effects. In addition, unlike UV/Vis light–induced RNAi, NIR laser–induced RNAi is translatable to the clinic because NIR can penetrate 3–5 cm into soft tissues. In a proof-of-concept study (Fig. 8), it was demonstrated that HAuNS coated with folic acid and loaded with NF-κB p65 siRNA were located in the endocytic vesicles in HeLa cells expressing folate receptors with intact membranes. However, when the cells were irradiated with an NIR laser, the endocytic membrane was disrupted, resulting in escape of the nanoparticles into the cytoplasm and detachment of the NF-κB siRNA from the HAuNS. A significant reduction in NF-κB p65 protein expression was demonstrated both in vitro and in vivo and sensitization



Hollow Gold Nanospheres, Fig. 9 Noninvasive PAT imaging of a mouse brain in vivo employing PEG-HAuNS and NIR light at a wavelength of 800 nm. Photoacoustic image acquired (**a**) before, (**b**) 5 min after, and (**c**) 2 h after the intravenous injection of PEG-HAuNS. (**d**) and (**e**): Differential images that were obtained by subtracting the preinjection image from the postinjection images (Image D = Image B–Image A; Image E = Image C–Image A). *Arrow*, middle cerebral artery. Bar = 2 mm. (**f**) Open-skull photograph of the mouse brain cortex obtained after the data acquisition for PAT. Bar = 2 mm (Reproduced with permission from Ref. [12])

to chemotherapy through enhanced apoptotic response upon laser irradiation [6]. Thus, it is possible to create multifunctional HAuNS with laser-activatable release of anticancer agents and siRNA molecules.

Photoacoustic Imaging

Noninvasive molecular and functional imaging techniques show promise for detecting and monitoring various physiological and pathological conditions in animals and, ultimately, humans. Among these imaging modalities, photoacoustic imaging offers higher spatial resolution and allows deeper tissues to be imaged than most optical imaging techniques [10, 11]. Photoacoustic tomography (PAT), also referred to as optoacoustic tomography (OAT), is a novel, hybrid, nonionizing imaging modality that combines the merits of both optical and ultrasonic imaging methods. PAT detects absorbed photons ultrasonically through the photoacoustic effect: A short-pulsed laser irradiates biological tissues and induces wideband ultrasonic waves (photoacoustic waves) as a result of transient thermoelastic expansion. Because HAuNS display strong resonance absorption tuned to the NIR range, HAuNS are a naturally efficient contrast agent for photoacoustic imaging. To demonstrate the utility of PEG-HAuNS as a photoacoustic contrast agent, PAT images of the cerebral cortex of a mouse before and after a single injection of PEG-HAuNS are presented in Fig. 9 [12]. Because of the low intrinsic absorption of oxyhemoglobin and deoxyhemoglobin at 800 nm, the photoacoustic signal was weak, with a low signal-to-noise ratio that showed only large vessels distributed along the mouse brain cortical surface, such as the middle cerebral artery (Fig. 9a, *arrow*). However, 5 min after the injection of PEG-HAuNS, the PAT image revealed the brain vasculature with much greater clarity, especially small blood vessels (Fig. 9b). This enhanced clarity was attributed to the strong photoacoustic signal generated with PEG-HAuNS. The circulating PEG-HAuNS enhanced the contrast between the

blood vessels and the brain parenchyma. This is better appreciated in a differential image (Fig. 9d) obtained by subtracting the preinjection image (Fig. 9a) from the postinjection image (Fig. 9b) pixel by pixel. At 2 h postinjection, the photoacoustic images remained essentially unchanged, indicating a significant amount of PEG-HAuNS circulating in the blood (Fig. 9c, e). Compared with PAT images based on the intrinsic optical contrast in nude mice, the PAT images acquired within 2 h after intravenous administration of PEG-HAuNS showed the brain vasculature with greater clarity and detail. The image depicted brain blood vessels as small as $\sim$ 100 μm in diameter using PEG-HAuNS as the contrast agent. These results indicate that PEG-HAuNS is a promising contrast agent for PAT, with high spatial resolution and enhanced sensitivity.

Surface-Enhanced Raman Scattering (SERS)

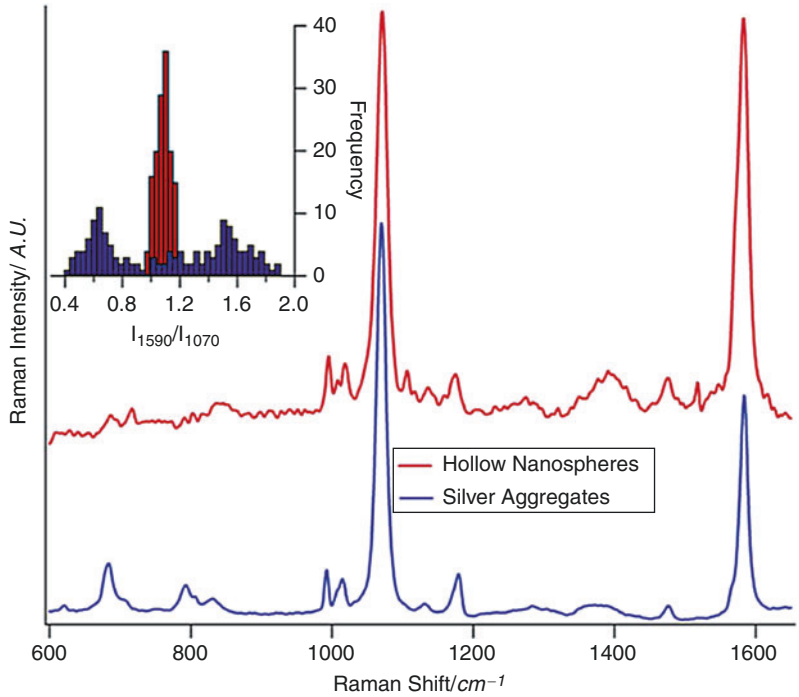
Another important application of HAuNS is surface-enhanced Raman scattering (SERS), an extremely useful method of chemical and biochemical analysis and detection. SERS is based on the enhancement of Raman scattering of an analyte molecule near or on a roughened metal substrate surface. The origin of the Raman enhancement has been largely attributed to an enhanced electromagnetic (EM) field at the metal surface due to increased absorption of the incident light. Normal Raman scattering has the key advantage of molecular specificity or selectivity that allows for the identification of unique samples but has the major limitation of very small signal or low quantum yield of scattering (on the order of 10^{-7}) [13]. SERS, discovered in the mid-1970s, has made Raman scattering more popular for practical applications by offering the combined advantages of molecular specificity *and* a large signal [14, 15].

SERS spectra are very sensitive to the metal substrate. The size and shape, as well as the morphology, of the metal substrate will affect the sensitivity and the consistency of SERS measurement. Much effort has been put into improving SERS spectra consistency in practical measurement by developing metal substrates that are uniform or monodisperse in size and shape. HAuNS have been demonstrated to be an excellent SERS substrate in this regard [1]. The unique optical and structural properties of HAuNS afford several advantages for SERS application, including small size, spherical shape, and tunable and narrow absorption from the visible to near-IR region [1]. Furthermore, compared to aggregates or nanoparticles with inhomogeneous distribution of size and shape, the HAuNS afford SERS spectra that are highly consistent, measured by the peak ratio of SERS spectra of single HAuNS. This consistency is a direct result of the narrow surface plasmon absorption band and high structural uniformity of HAuNS. Figure 10 shows a comparison of SERS spectra of mercapto benzoic acid (MBA) obtained using HAuNS vs. normal silver nanoparticle aggregates. The inset shows a histogram detailing the distribution of the peak ratio of two SERS peaks for the two different substrates. The HAuNS show a much narrower, and thus better, distribution (red), indicating a highly uniform structure and narrow SPR, and thereby, consistent SERS for a single nanostructure [16].

While the HAuNS are indeed very uniform and show highly consistent SERS, they afford a relatively weak SERS signal compared to Ag nanoparticles [16]. Ideally, one would produce hollow silver, instead of gold, nanospheres to improve SERS sensitivity. To achieve this, HAuNS have been used as seed templates and coated with a thin layer of Ag on the outer surface [17]. This approach has resulted in the successful generation of unique hollow gold-silver double nanoshells that show a significantly shifted and broadened plasmon band. These double nanoshells have shown greatly increased Raman enhancement over HAuNS, which is attributed to a combination of the increased volume of metal and the improved enhancement factor of silver.

Future Directions

Multifunctional HAuNS combined with PTA therapy, drug delivery, or RNAi capability provide a great opportunity for multimodal therapy of



Hollow Gold Nanospheres, Fig. 10 Single-particle SERS spectrum of MBA on HAuNS (*red trace*, *top*) and silver aggregates (*blue trace*, *bottom*). The *inset* is a histogram of the relative intensity of the two most intense peaks of MBA at 1070 and 1590 cm^{-1} of 150 HAuNS (*red bars*) and 150 silver aggregates (*blue bars*) (Reproduced with permission from Ref. [16])

a variety of solid tumors. Future effort should be focused on the clinical translation of this innovative therapeutic approach. A number of hurdles must be overcome before these nanoparticles are administered to patients. In particular, the acute and long-term toxicity of PEG-HAuNS needs to be carefully documented. Selective delivery of a sufficient amount of HAuNS should lead to a more selective photothermal effect localized to the tumor. The integration of imaging probes into the HAuNS to make it possible to visualize these nanoparticles using noninvasive imaging techniques such as magnetic resonance imaging and nuclear imaging may help monitor the spatiotemporal distribution of the HAuNS in solid tumors and, thus, enable accurate pretreatment thermal dose calculation. To improve the photothermal conducting efficiency of HAuNS, the SPR band should be further reduced. The SPR bandwidth directly impacts PTA efficiency; HAuNS with a narrower band have higher overall photothermal conversion and improved PTA performance. It is also desired to improve the synthesis to make it more easily reproducible and on large scales. For example, PVP has been found to improve the synthesis reproducibility of NIR-absorbing NAuNS significantly when used as a stabilizing agent to slow down the HAuNS growth [18]. Scaling up the synthesis is another important direction for further work in the future.

Cross-References

- Carbon-Nanotubes
- Gold Nanorods
- Nanomedicine
- Optical Properties of Metal Nanoparticles
- Optical Techniques for Nanostructure Characterization
- Thermal Cancer Ablation Therapies Using Nanoparticles

References

1. Schwartzberg, A.M., Olson, T.Y., Talley, C.E., Zhang, J.Z.: Synthesis, characterization, and tunable optical properties of hollow gold nanospheres. J. Phys. Chem. B **110**, 19935–19944 (2006)

2. Mie, G.: Beiträge zur optik trüber medien, speziell kolloidaler metallösungen. Ann. Phys. **25**, 377–445 (1908)
3. Wheeler, D.A., Newhouse, J., Wang, H., Zou, S., Zhang, J.Z.: Optical properties and persistent spectral hole burning of near infrared-absorbing hollow gold nanospheres. J. Phys. Chem. C **114**, 18126–18133 (2010)
4. Zhang, G., Yang, Z., Lu, W., Zhang, R., Huang, Q., Tian, M., Li, L., Liang, D., Li, C.: Influence of anchoring ligands and particle size on the colloidal stability and in vivo biodistribution of polyethylene glycol-coated gold nanoparticles in tumor-xenografted mice. Biomaterials **30**, 1928–1936 (2009)
5. Lu, W., Xiong, C., Zhang, G., Huang, Q., Zhang, R., Zhang, J.Z., Li, C.: Targeted photothermal ablation of murine melanomas with melanocyte-stimulating hormone analog-conjugated hollow gold nanospheres. Clin. Cancer Res. **15**, 876–886 (2009)
6. Lu, W., Zhang, G., Zhang, R., Flores 2nd, L.G., Huang, Q., Gelovani, J.G., Li, C.: Tumor site-specific silencing of NF-kappaB p65 by targeted hollow gold nanosphere-mediated photothermal transfection. Cancer Res. **70**, 3177–3188 (2010)
7. Melancon, M.P., Lu, W., Yang, Z., Zhang, R., Cheng, Z., Elliot, A.M., Stafford, J., Olson, T., Zhang, J.Z., Li, C.: In vitro and in vivo targeting of hollow gold nanoshells directed at epidermal growth factor receptor for photothermal ablation therapy. Mol. Cancer Ther. **7**, 1730–1739 (2008)
8. You, J., Shao, R., Wei, X., Gupta, S., Li, C.: Near-infrared light triggers release of paclitaxel from biodegradable microspheres: photothermal effect and enhanced antitumor activity. Small **6**, 1022–1031 (2010). Weinheim an der Bergstrasse, Germany
9. You, J., Zhang, G., Li, C.: Exceptionally high payload of doxorubicin in hollow gold nanospheres for near-infrared light-triggered drug release. ACS Nano **4**, 1033–1041 (2010)
10. Xu, M.H., Wang, L.H.V.: Photoacoustic imaging in biomedicine. Rev. Sci. Instrum. **77**, 041101 (2006)
11. Oraevsky, A.A., Karabutov, A.A.: Optoacoustic tomography. In: Vo-Dinh, T. (ed.) Biomedical Photonics Handbook, vol. PM125, pp. 1–34. CRC Press/Taylor and Francis Group, Boca Raton (2003)
12. Lu, W., Huang, Q., Ku, G., Wen, X., Zhou, M., Guzatov, D., Brecht, P., Su, R., Oraevsky, A., Wang, L.V., Li, C.: Photoacoustic imaging of living mouse brain vasculature using hollow gold nanospheres. Biomaterials **31**, 2617–2626 (2010)
13. Ferraro, J.R., Nakamoto, K.: Introductory Raman Spectroscopy, p. 370. Academic, Boston (1994)
14. Fleischmann, M., Hendra, P.J., McQuillan, A.J.: Raman spectra of pyridine adsorbed at a silver electrode. Chem. Phys. Lett. **26**, 163–166 (1974)
15. Albrecht, M.G., Creighton, J.A.: Anomalously intense Raman-spectra of pyridine at a silver electrode. J. Am. Chem. Soc. **99**, 5215–5217 (1977)
16. Schwartzberg, A.M., Oshiro, T.Y., Zhang, J.Z., Huser, T., Talley, C.E.: Improving nanoprobes using surface-enhanced Raman scattering from 30-nm hollow gold particles. Anal. Chem. **78**, 4732–4736 (2006)
17. Olson, T.Y., Schwartzberg, A.M., Orme, C.A., Talley, C.E., O'Connell, B., Zhang, J.Z.: Hollow gold-silver double-shell nanospheres: structures, optical absorption, and surface enhanced Raman scattering. J. Phys. Chem. C **112**, 6319–6329 (2008)
18. Preciado, S., Wang, D., Wheeler, D., Newhouse, R., Hensel, J., Schwartzberg, A., Wang, L., Zhu, J.J., Barboza-Flores, M., Zhang, J.Z.: Highly reproducible synthesis of hollow gold nanospheres with near infrared surface plasmon absorption using PVP as stabilizing agent. J. Mater. Chem. **21**, 2344–2350 (2011)

H

Hot Filament Chemical Vapor Deposition (HFCVD)

▶ Chemical Vapor Deposition (CVD)

Hot-Wall Thermal Chemical Vapor Deposition

▶ Chemical Vapor Deposition (CVD)

How to Synthesize/Make/Manufacture/Fabricate/Produce Graphene

▶ Synthesis of Graphene

Hybrid Network Method

▶ Dynamic Clamp

Hybrid Optoelectric Manipulation

▶ Optoelectrically Enabled Multiscale Manipulation

Hybrid Opto-electric Technique

▶ Rapid Electrokinetic Patterning and Its Applications

Hybrid Photovoltaics

▶ Hybrid Solar Cells

Hybrid Solar Cells

Matthew T. Lloyd
National Renewable Energy Laboratory, Golden, CO, USA

Synonyms

Hybrid photovoltaics; Organic–inorganic heterojunctions; Organic–inorganic photovoltaics; Polymer–metal oxide solar cells

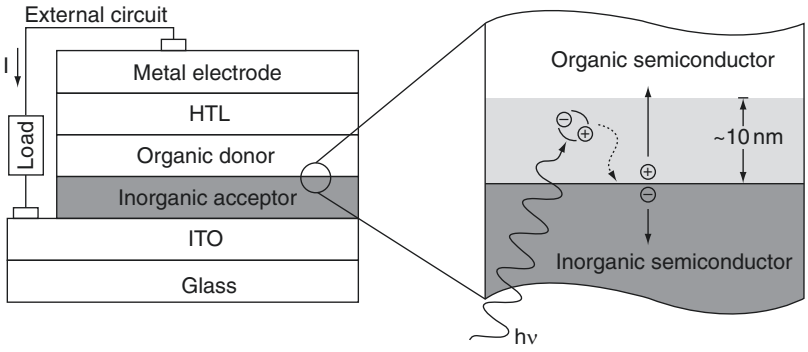
Definition

A hybrid solar cell is a photovoltaic device relying on charge transfer at the interface between two semiconductors, one being organic and the other being inorganic.

Principles of Operation

A prototypical hybrid photovoltaic device is composed of an organic semiconductor and an inorganic semiconductor. The interface between these two materials forms a type II heterojunction, where the polymer typically serves as the electron donating material and the inorganic semiconductor is the electron acceptor. In a device, these materials form a layered stack that is sandwiched between two electrodes. An opaque metallic electrode forms the back contact and a conductive semitransparent material, such as indium tin oxide (ITO), permits light to enter the front side of the device (Fig. 1).

The principle of operation is similar in nature to both a dye-sensitized solar cell and a polymer-fullerene organic photovoltaic (OPV) (see also ▶ Organic Photovoltaics: Basic Concepts and Device Physics). As in OPV devices, the conjugated polymer layer is primarily responsible for light absorption. Photons of energy greater than the optical band gap of the organic semiconductor are absorbed to create an excited state particle, known as an exciton. Consisting of a Coulombically bound electron–hole pair, excitons carry no net charge and move in a random walk uninfluenced by internal electric fields. Due to a short lifetime, excitons diffuse only a few nanometers (5–15 nm) within the organic layer before recombining [1]. By introducing a high electron affinity material, such as an inorganic semiconductor, charge separation becomes energetically favorable. An exciton reaching an organic–inorganic interface during its lifetime may dissociate by electron transfer to the inorganic semiconductor. It is also possible for excitons generated in the inorganic to dissociate at the heterojunction interface via hole transfer to the organic material. In either case, these charges tend to recombine unless they can be further separated. The driving force for separation comes from diffusion or the built-in electric field within the device. The donor/acceptor phases then serve the dual purpose of transporting charge (see also ▶ Conduction Mechanisms in Organic Semiconductors) to the respective electrodes – hole transport occurs via the highest occupied molecular orbital of the organic semiconductor, and electron transport via the conduction band of the inorganic semiconductor. Carrier transport out of the device must be sufficiently fast to compete with recombination between charges at the interface of adjacent materials. High charge carrier mobility is therefore an important parameter to ensure efficient device operation. Several recent reviews on hybrid solar cells provide further details on the physics of device operation [1–5].



Hybrid Solar Cells, Fig. 1 Schematic depiction of a planar heterojunction (or bilayer) organic–inorganic hybrid device (*left*). Light absorption by the polymer generates excitons (shown in the detail to the *right*), which diffuse approximately 5–15 nm to the heterojunction interface. After electron transfer to the conduction band of the inorganic semiconductor, electrons traverse the acceptor layer for collection at a transparent conducting electrode, typically indium tin oxide (ITO) (see also ▶ Electrode–Organic Interface Physics). Holes remain in the organic layer for transport to and collection by a metallic back contact. A hole transport layer (*HTL*) may be employed to modify the properties of the electrode, such as work function or surface roughness

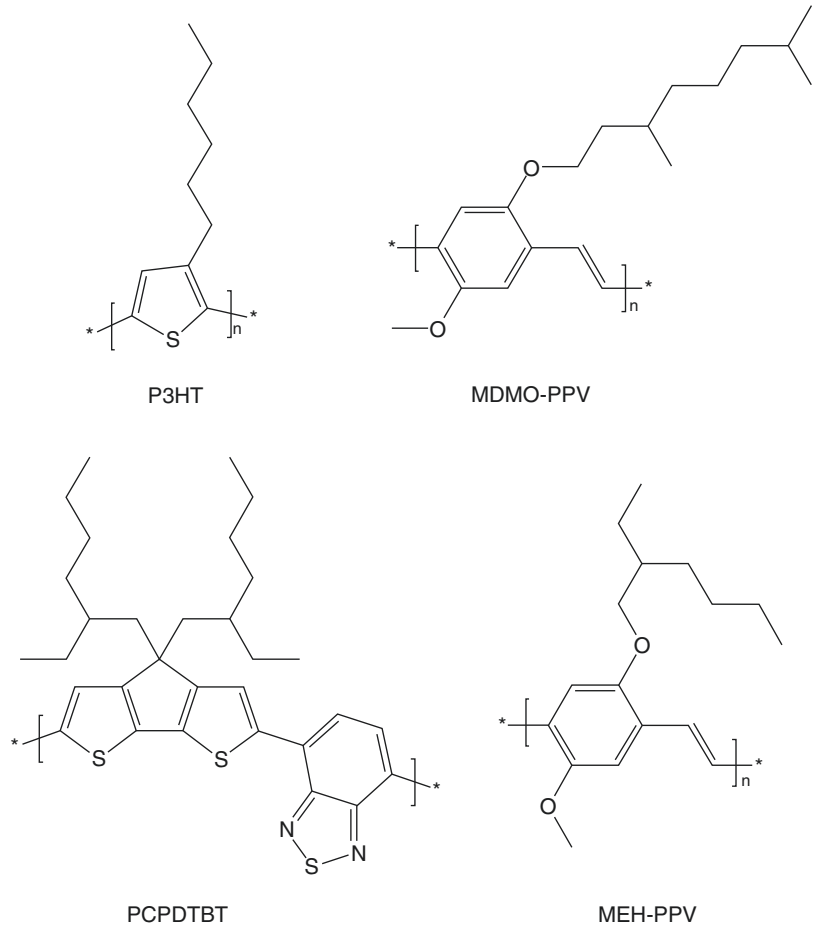
H

Materials

Organic semiconductors used in hybrid devices are typically the same as those found in OPV devices and similar processing methods and material properties are required in hybrid devices. For many polymeric semiconductors, such as regioregular P3HT, proper ordering of the conjugated backbone is needed to yield optimum charge carrier mobility (10^{-2} to 10^{-3} cm^2/Vs, measured in P3HT thin film transistors). During deposition of the organic, interactions at the polymer–substrate and polymer–metal oxide interface play an important role in controlling polymer order. Interfacial modification with SAMS (see ▶ Self-assembled Monolayers) of various organic compounds can be employed to influence polymer ordering [6]. Processing parameters such as solvent evaporation rate, post-deposition thermal and solvent annealing treatment, also strongly influence the properties of organic semiconductors. Due to a high absorption coefficient ($\sim 10^5$ cm^{-1}), light absorption by the organic component is generally good; however, absorption bands tend to be narrow and typically peak in the visible spectrum, leaving much of the red to NIR portion of spectrum unabsorbed. Much effort continues to be devoted to broadening the absorption profile and decreasing the optical bandgap by way of synthetic modification of the polymer backbone. Shown in Fig. 2, PCPDTBT is an example of one such low bandgap polymer that has been developed in recent years. In terms of energy-level requirements, most p-type organic materials have a much smaller electron affinity than the inorganic electron acceptor, ensuring a sufficient energetic offset for exciton dissociation between the two materials. Of course, a low bandgap donor material also imposes new constraints on the conduction band energy level of the acceptor to ensure sufficient driving force for charge separation.

For laboratory-scale devices, deposition of organic semiconductors is most often carried out by spin casting films from solution. Large area deposition is possible using other solution processing techniques such as slot-die coating, screen printing, and spray coating. In the case of donor-acceptor blends or bulk heterojunctions, additional thermal or solvent annealing steps are employed to optimize donor–acceptor phase separation. The molecular structure for several commonly used organic semiconductors is shown in Fig. 2.

For the inorganic component, materials selection is used to optimize light absorbance and maximize charge carrier mobility. Deposition of inorganic materials can be carried out by numerous methods. Using sol–gel chemistry, dense

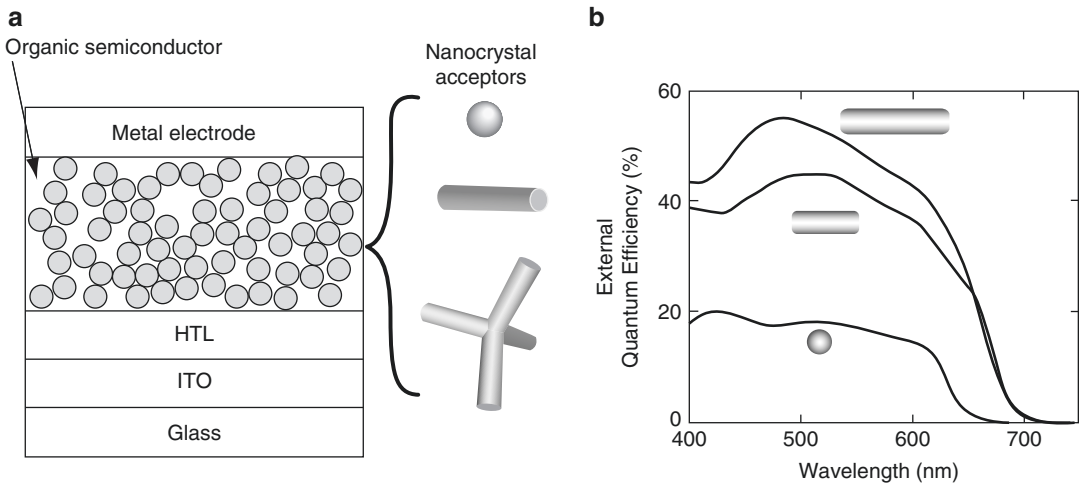


Hybrid Solar Cells, Fig. 2 Molecular structure and chemical abbreviation for several typical polymeric semiconductors. The full chemical names are as follows: poly (3-hexylthiophene) (P3HT); poly[2-methoxy-5-(3′, 7′-dimethyloctyloxy)-1,4-phenylenevinylene] (MDMO-PPV); poly [2,6-(4,4-bis-(2-ethlyhexyl)-4 *H*-cyclopenta[2,1,*b*;3,4-*b*]-dithiophene)-*alt*-4,7-(2,1,3-benzothiadiazole)] (PCPDTBT); poly [2-methoxy-5-(2-ethylhexyloxy)-1,4-phenylenevinylene] (MEH-PPV)

polycrystalline films can be processed from solution and thermally cured. In another common preparation method, soluble nanoparticles of various shapes can be dispersed in polymer solutions and spin cast as a two-component solution. Ordered inorganic nanostructures can be grown using hydrothermal techniques followed by solution phase or melt infiltration of the organic material. Similarly, ordered nanostructures can be templated using block copolymers or deposited using a variety of vapor growth methods. Below is a list of many inorganic semiconductors that have been employed in hybrid devices in varying capacities:

- CdS
- CdSe
- CdTe
- $CuInS_2$
- $CuInSe_2$
- InP
- PbS
- PbSe
- Si
- SnO_2
- TiO_2
- ZnO

During the synthesis of inorganic nanoparticles, chemical methods can be employed to tailor nanoparticle size, aspect ratio, and type of branching [2]. Colloidal nanoparticles, such as CdSe, can be nucleated from organometallic precursors in a high-temperature (250–350 °C) mixture of surfactants. Nanoparticles are typically capped with organic ligands that passivate surface



Hybrid Solar Cells, Fig. 3 (**a**) Schematic representation of an organic-nanoparticle bulk heterojunction photovoltaic device. The nanocrystalline acceptors are frequently spherical particles, but may be also synthesized in a variety of high aspect ratio shapes, such as nanorods or tetrapods, to improve electron transport properties. (**b**) External quantum efficiency (or spectral response) for a hybrid nanocrystalline bulk heterojunction device employing nanorod acceptors of increasing aspect ratio. Photocurrent collection is enhanced as the number of hopping steps between particles is reduced. (Reprinted with permission from [9]. Copyright 2002 Science/AAAS)

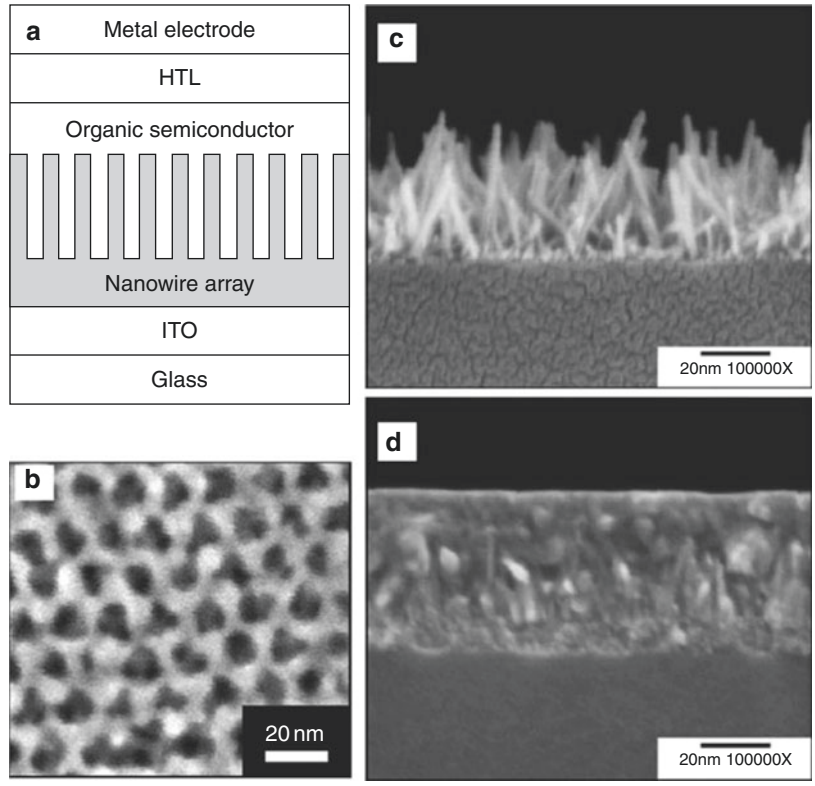
states and ensure a stable dispersion when mixed with a polymer in organic solvents. Unfortunately, many solubilizing ligands also act as insulating shells that disrupt particle–particle electronic communication. After nanoparticle synthesis, residual surfactants can be stripped away to improve contact to the nanoparticles surface. The presence of residual surfactant influences charge transfer efficiency, nanoparticles aggregation, and charge transport, ultimately limiting the performance of organic–inorganic composite photovoltaic devices.

Planar, Bulk Heterojunction, and Nanostructured Architectures

The need for nanoscale features in hybrid photovoltaics stems from the fact that exciton diffusion is limited to a few nanometers. Excitons generated further than a diffusion length from the organic–inorganic interface will not contribute to photocurrent generation. In a bilayer or planar heterojunction (see Fig. 1), the optical absorption depth is a factor of 10 greater than the exciton diffusion length in most semiconducting polymers. The fraction of absorbed light near the heterojunction interface in a bilayer is too low; hence, devices based on this architecture cannot achieve useful power conversion efficiencies.

A significant development to this technology came with the fabrication of the bulk heterojunction (Fig. 3a), which is the basis for most high efficiency all-organic photovoltaic devices [7, 8]. A bulk heterojunction is comprised of a bi-continuous interpenetrating network of the electron accepting and donating material, ideally with length scales on the order of an exciton diffusion length. In principle, an exciton generated anywhere within the polymer phase will be able to reach a heterojunction interface within its lifetime.

Although the nanoparticle–polymer bulk heterojunction dramatically enhances interfacial area, high-efficiency devices also require effective charge transport of the electron to the electrode. Spherical nanoparticles provide an electron percolation pathway; however, charges encounter multiple hopping events between discrete nanoparticles, reducing overall carrier mobility. By increasing the aspect ratio of the particle, electrons encounter fewer hops during extraction from



Hybrid Solar Cells, Fig. 4 (**a**) Ideal ordered nanostructure where the spacing between nanowires is approximately an exciton diffusion length and the thickness of the organic layer is sufficient to absorb 100 % of the incident light. (**b**) Scanning electron microscopy (*SEM*) image showing a TiO_2 mesoporous lattice, deposited using a block-copolymer templating techniques. (**c**) SEM image depicting a hydrothermally grown ZnO nanowire array. (**d**) A ZnO nanowire array after infiltration with an organic semiconductor (Figure 4b reprinted with permission from [11]. Copyright 2003 Wiley-VCH. Figure 4c, d reprinted with permission from [10], copyright 2006 Elsevier)

the active layer. As shown in Fig. 3b, high aspect ratio nanoparticles have been shown to significantly increase the photocurrent generation in hybrid devices due to their higher effective mobility [9]. Unfortunately, unbound nanorods suffer from poor morphological arrangement where most rods orient parallel to the substrate. Using synthetic techniques, it is possible to form nanoparticles in the shape of a tetrapod [2]. One arm of the tetrapod, which is always normal to the substrate, will provide a high-mobility pathway away from the center of the device reducing the number of particle–particle hops before collection by the electrode.

A variant of the bulk heterojunction is possible using nanofabrication techniques. Instead of a randomly formed percolating inorganic network, one can intentionally deposit inorganic semiconductors to form rigid, three-dimensional networks. As shown in Fig. 4, arrays of inorganic nanowires or nanotubes are ideally separated by an exciton diffusion length and long enough to achieve maximum light absorption by an infiltrated polymer layer [10]. The nanorods provide a high-mobility charge transport pathway along a single crystal of inorganic material. Similarly, block copolymers can be used as a structure-directing template to form an inorganic semiconductor into a continuous three-dimensional lattice (Fig. 4b) [11]. In both cases, the polymeric donor can be filled or infiltrated with either polymer solutions or solid polymer that is softened by thermal processing.

Advantages of Hybrid Technology

Inorganic acceptor materials display desirable characteristics that hold promise for increasing the overall performance of low-cost, solution-processed photovoltaics. The inorganic semiconductors that form the electron transport networks found in hybrid cells provide significantly higher (often orders of magnitude) charge carrier mobility compared to most n-type organic semiconductors. Many of these materials can be synthesized

and deposited from solution, and processing steps occur at temperatures compatible with polymeric substrates. Material toxicity and abundance are important factors to consider for large-scale deployment of hybrid photovoltaics. Many *n*-type materials being investigated for hybrid cells, such as TiO_2 and ZnO, are nontoxic and derived from highly abundant precursors.

Ordered inorganic nanostructures afford the opportunity for various types of interfacial modification. Inorganic nanostructures can be modified by coating with thin films, such as another inorganic layer or by an organic SAM [3]. This is unique to the hybrid system where the nanostructured inorganic forms a rigid scaffolding that is amenable to chemical modification specifically at the charge separation interface. The polymeric donor is then infiltrated or back-filled as normal. Organic interfacial treatments have been shown to enhance polymer morphology at the charge separation interface [6]. Other SAM layers establish interfacial dipoles that influence charge transfer efficiency [12]. Thin inorganic interfacial layers may be instrumental in the passivation of surface states and act as barriers to recombination. With better understanding of interfacial effects, modification of the heterojunction may become an important route to optimize the efficiency of hybrid devices.

A significant advantage of hybrid technology comes from the enhanced absorption spectrum of the acceptor material compared to that of fullerene derivatives typically used in OPV devices, which exhibit negligible absorption in the visible spectrum. Many inorganic materials under investigation for hybrids are direct bandgap semiconductors with high extinction coefficients. Devices utilizing low bandgap inorganic particles (such as CdTe) demonstrate light absorption by the acceptor phase that contributes to photocurrent generation beyond that of the donor alone. In this way, the complementary absorption spectra can maximize the absorption overlap with the solar spectrum. Additionally, quantum confinement effects in nanoscale inorganic materials can also be exploited in hybrid solar cells [4]. When the Bohr radius of an exciton (the distance between the bound electron–hole pair) exceeds the diameter of the nanocrystal, material properties are dominated by quantum effects. By further reducing the particle size, one can systematically increase the size of the bandgap of the semiconductor relative to the bulk material. This permits tuning of both the peak absorption and the energy levels for charge transport. Manipulation of the particle size may become an important tool for ensuring the necessary driving force for charge separation as well as optimizing absorption of the solar spectrum.

Challenges and Future Research Directions

Despite significant progress in understanding the fundamental mechanisms at work in hybrid photovoltaics, in practice the power conversion efficiency of hybrid devices remains low. The record efficiency for polymer-fullerene OPV devices measures 8.3 %, while hybrid devices remain below the 3 % mark. Several areas have been identified as key to improving device efficiency. The first issue concerns a discrepancy in the magnitude of the charge carrier mobility in nanoparticle thin films. Despite high mobilities generally known for inorganic semiconductors, measurements for films of pure nanoparticles or nanostructured features indicate mobilities are several orders of magnitude lower than typical bulk values. Phenomena taking place in nanostructured films, such as multiple interparticle hopping events or charge trapping, may lead to detrimental transport properties. In addition, attention to the carrier mobility of the organic semiconductor is important to achieve higher device efficiency. Carrier mobility of the polymer is vulnerable to changes induced by interaction with the surface of the inorganic as well as polymer chain confinement after infiltration of the polymer within ordered nanostructures.

Second, charge transfer events occurring at organic–inorganic heterojunctions remain poorly understood. A great deal of research seeks to understand how parameters, such as polymer order, spatial orientation, surface dipoles, and bonding, influence the efficiency of interfacial charge separation. A better understanding of these

organic–inorganic interface effects will likely lead to improved exciton dissociation efficiency and decrease interfacial recombination rates.

Finally, as is the case for OPV, in order to reach maximum efficiency one has to achieve control over the donor–acceptor phase-separated network as well as develop reliable methods to characterize these networks on the nanometer scale [13]. To date, the largest improvements in device efficiency originate with methods that increase the donor–acceptor interfacial area. Although current hybrid solar cell efficiencies lag behind those produced with other material systems, the attractive material properties and desirable processability of inorganic semiconductors help maintain interest for further development of hybrid cells.

Cross-References

- ▶ Conduction Mechanisms in Organic Semiconductors
- ▶ Electrode–Organic Interface Physics
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Organic Photovoltaics: Basic Concepts and Device Physics
- ▶ Self-assembled Monolayers

References

1. Greenham, N.C.: Chapter 6: Hybrid polymer/nanocrystal photovoltaic devices. In: Brabec, C., Dyakonov, V., Scherf, U. (eds.) Organic Photovoltaics: Materials, Device Physics, and Manufacturing Technologies, pp. 179–210. Wiley, Weinheim (2008)
2. Milliron, D., Gur, I., Alivisatos, A.: Hybrid organic – nanocrystal solar cells. MRS Bull. **30**, 41–44 (2005)
3. Bouclé, J., Ravirajan, P., Nelson, J.: Hybrid polymer–metal oxide thin films for photovoltaic applications. J. Mater. Chem. **17**, 3141 (2007)
4. Borchert, H.: Elementary processes and limiting factors in hybrid polymer/nanoparticle solar cells. Energy Environ. Sci. **3**, 1682–1694 (2010)
5. Coakley, K.M., Liu, Y., Goh, C., McGehee, M.D.: Ordered organic–inorganic bulk heterojunction photovoltaic cells. MRS Bull. **30**, 37–40 (2005)
6. Hsu, J.W.P., Lloyd, M.T.: Organic/inorganic hybrids for solar energy generation. MRS Bull. **35**, 423 (2010)
7. Halls, J., Walsh, C., Greenham, N., Marseglia, E., Friend, R., Moratti, S., Holmes, A.B..: Efficient photodiodes from interpenetrating polymer networks. Nature **376**, 498–500 (1995)
8. Yu, G., Gao, J., Hummelen, J., Wudl, F., Heeger, A.: Polymer photvoltaic cells – enhanced efficiencies via a network of internal donor-acceptor heterojunctions. Science **270**, 1789–1791 (1995)
9. Huynh, W., Dittmer, J., Alivisatos, A.: Hybrid nanorod-polymer solar cells. Science **295**, 2425–2427 (2002)
10. Olson, D., Piris, J., Collins, R., Shaheen, S., Ginley, D.: Hybrid photovoltaic devices of polymer and ZnO nanofiber composites. Thin Solid Films **496**, 26–29 (2006)
11. Coakley, K.M., Liu, Y., Mcgehee, M.D., Frindell, K. L., Stucky, G.D.: Infiltrating semiconducting polymers into self-assembled mesoporous titania films for photovoltaic applications. Adv. Funct. Mater. **13**, 301–306 (2003)
12. Goh, C., Scully, S.R., McGehee, M.D.: Effects of molecular interface modification in hybrid organic–inorganic photovoltaic cells. J. Appl. Phys. **101**, 114503–114512 (2007)
13. Oosterhout, S.D., Wienk, M.M., van Bavel, S.S., Thiedmann, R., Koster, L.J.A., Gilot, J., Loos, J., Schmidt, V., Janssen, R.A.J.: The effect of three-dimensional morphology on the efficiency of hybrid polymer solar cells. Nat. Mater. **8**, 1–7 (2009)

Hydrophilic

▶ Surface Engineering, Tailored Wettability, and Applications

Hydrophobic

▶ Surface Engineering, Tailored Wettability, and Applications

Hyperelasticity

▶ Superelasticity and the Shape Memory Effect

Hyperthermia Treatment

▶ Magnetic Nanoparticles for Biomedical Applications

Icephobicity

► Micro-/Nanostructured Icephobic Materials

Imaging the Human Body Down to the Molecular Level

Hans Deyhle, Georg Schulz and Bert Müller
Biomaterials Science Center (BMC), University of Basel, Basel, Switzerland

Synonyms

Three-dimensional imaging of human tissues

Definition

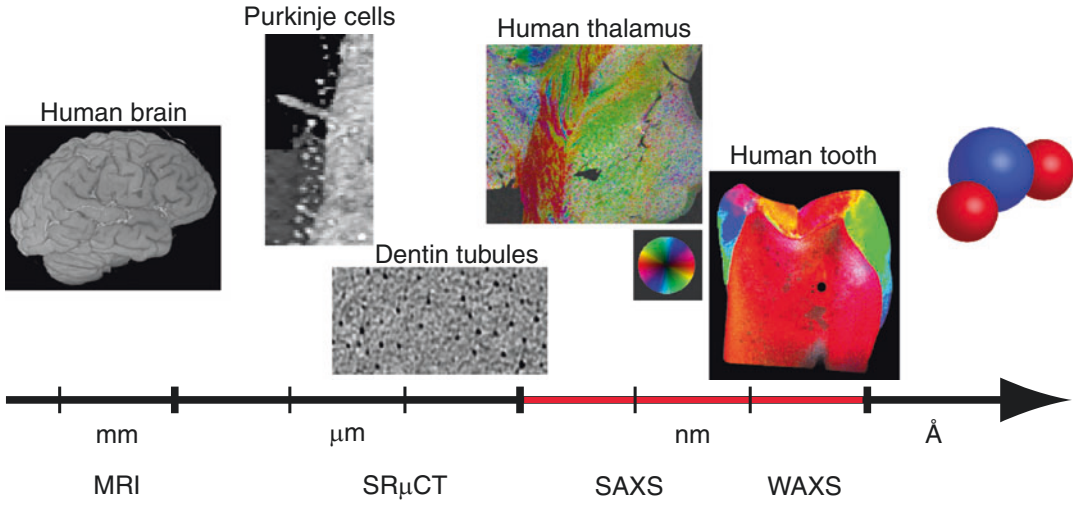
The human body consists of nanometer-sized units including proteins, apatite crystallites, collagen, and myelin fibers. Imaging, here, means the identification, localization, and quantification of these units within the human body.

Overview

The human body consists of about 10^{27} molecules. This number is so huge that it is impossible to determine their location or even only to store this amount of data. Using a logarithmic scale (see Fig. 1), one realizes that a biological cell with an extension of about 10 μm includes as many molecules like the human body biological cells. Even the number of cells within the human body is huge and beyond our imagination. The number of stars in the Milky Way, for example, is thousand times smaller than this number. Therefore, for imaging the human body on the nanometer scale, one has to restrict to predefined parts of the body or to take advantage of symmetries or periodicities as known from the atomic view of single crystals in solid-state physics.

Three-dimensional (3D) imaging in daily clinical practice of well-equipped hospitals reaches a spatial resolution of a fraction of a millimeter, as indicated by the magnetic resonance imaging (MRI) in Fig. 1. Hence, the spatial resolution of medical imaging facilities is far from the molecular scale. In engineering and research, however, methods are known that allow imaging structures down to the world of atoms. As shown in the overview of Fig. 1, micro-computed tomography (μCT) is well established in materials science and yields a spatial resolution as low as a fraction of a micrometer in real space. The necessary dose for the subcellular level, however, is so high that only postmortem studies are permitted. Although the spatial resolution is high enough to visualize individual biological cells, the contrast is, contrary to hard tissues, too low. Monochromatic X-ray beams, as applied in synchrotron radiation-based μCT (SRμCT), provide much better contrast.

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1



Imaging the Human Body Down to the Molecular Level, Fig. 1 The selected imaging techniques (MRI, SRμCT, SAXS/WAXS) cover length scales down to the molecular level

Therefore, SRμCT has allowed visualizing individual cells in human tissue after appropriate staining [1]. Grating-based phase-contrast SRμCT has been shown to give rise to even orders of magnitude better contrast for human tissues [2]. Here, non-stained, individual Purkinje cells located in the brain come to light (see Fig. 1). In order to make visible nanostructures between 1 nm and 100 nm in real space, however, X-ray optics have to be integrated into the setup and the accessible volumes become as small as for electron-based techniques.

Diffraction and scattering methods are established techniques to quantitatively characterize the arrangement of atoms or molecules in reciprocal space [3]. Therefore, they give rise to average values within the illuminated volume. The combination of small probing beams and scanning provides spatially resolved data for the nanometer range and below. X-ray-based methods are especially suitable for human tissues. They are termed scanning small-angle X-ray scattering (scanning SAXS) and scanning wide-angle X-ray scattering (scanning WAXS). While SAXS covers the entire nanometer range, WAXS depicts the interatomic distances on the sub-nanometer scale (see Fig. 1).

Basics

X-ray computed tomography (CT) is a widespread imaging technique in radiology. Multi-slice CT (MSCT), for example, is used to identify tumors and to plan radiation therapy treatments. In dental medicine, cone-beam computed tomography (CBCT) is a standard imaging technique. The 3D images are used, for instance, to decide if enough bone is present for implant placements. Figure 2 shows, as an example, the 3D rendering of a part of a human skull acquired with the Accuitomo 80 CBCT system (Morita, Kyoto, Japan) recorded with an acceleration voltage of 70 kV and a beam current of 4 mA. 512 radiographs were acquired over 360°. The voxel size of the tomography data was $0.25 \times 0.25 \times 0.16\ \text{mm}^3$.

To improve the spatial and density resolution of the X-ray-based imaging techniques, dangerous doses are required, and in vivo measurements become impossible. Hence methods such as μCT are rather found in the field of materials science than medicine. Nevertheless, postmortem studies are common in many research laboratories and provide exhaustive morphological information on the micrometer scale. Several μCT systems are built at synchrotron radiation facilities, which

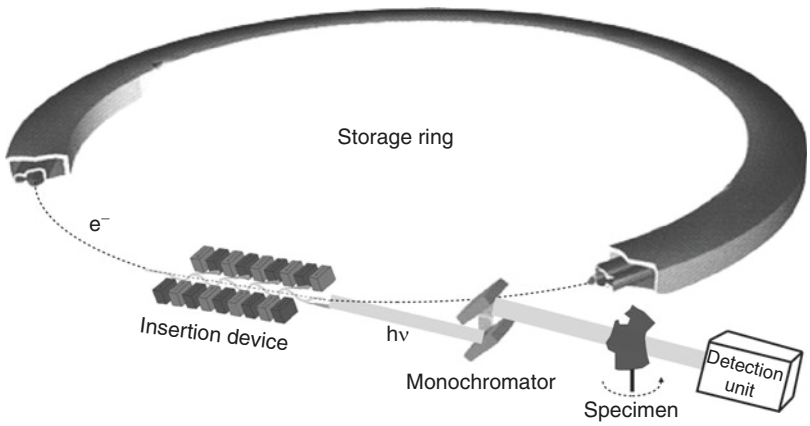
provide highly brilliant X-rays for different kinds of imaging purposes. The main components of such a facility are schematically represented in Fig. 3. The highly energetic electrons or positrons circulate in the storage ring and through insertion devices, where the intense X-ray beams are generated. The insertion devices are classified as bending magnet (first generation), wiggler (second generation), undulator (third generation), and free-electron laser (fourth generation). The difference of the insertion devices is – besides the increasing intensity – inter alia, the incoherent (first and second generations) and coherent (third and fourth generations) interference of the produced X-rays. The high intensity of synchrotron light allows realizing a tunable X-ray source. For that, a monochromator selects the predefined photon energy. The μCT end station contains the sample manipulator and the detector.



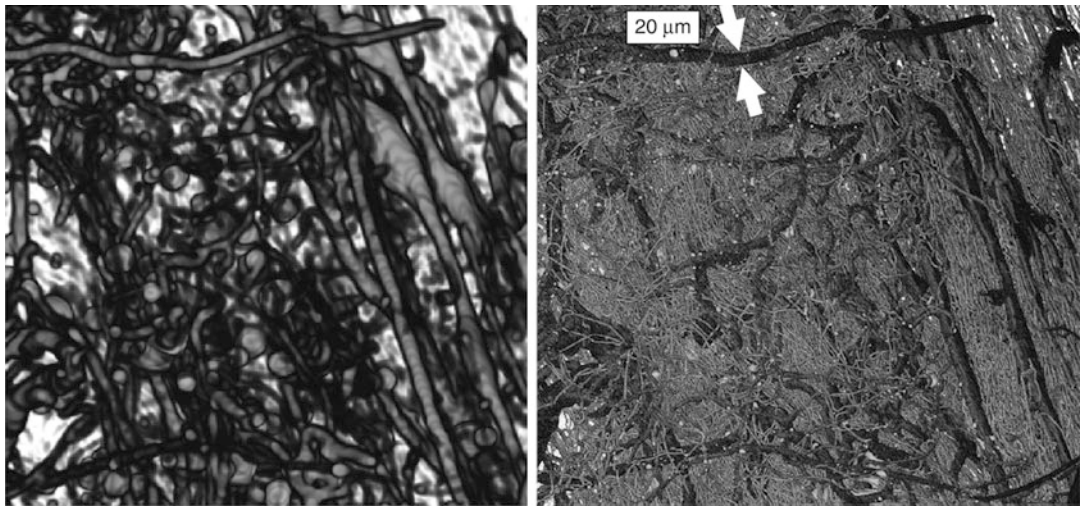
Imaging the Human Body Down to the Molecular Level, Fig. 2 The 3D rendering of a human skull generated from CBCT data impressively shows the hard tissues of the teeth and bone

The high X-ray intensity allows pushing the resolution of tomography at synchrotron radiation sources toward the sub-micrometer range, at the cost, however, of a reduced field of view. Figure 4 illustrates the SRμCT imaging capabilities for a polymer cast of the vascular system from a human C51 tumor, acquired at the TOMCAT beamline (Swiss Light Source, Paul Scherrer Institut, Switzerland) at a photon energy of 15 keV. In the top image, the isotropic voxel length corresponds to 5.9 μm. With these settings the whole tumor could be imaged. Capillaries with a diameter smaller than 10 μm, however, are not identified. Therefore, local tomography is often performed as demonstrated on the image of the same tumor below. Here, only a part of the specimen is scanned. The voxel length corresponded to 0.74 μm, and even the smallest capillaries having a diameter below 4 μm can be recognized.

Absorption tomography, as described above, is particularly suited for hard tissues like the bone or teeth, as these provide enough contrast in the X-ray images. For soft tissues, e.g., the human



Imaging the Human Body Down to the Molecular Level, Fig. 3 Layout of a synchrotron radiation facility consisting of a storage ring, an insertion device, and the end station for tomographic imaging



Imaging the Human Body Down to the Molecular Level, Fig. 4 Cast of the vascular network of a C51 tumor visualized by SRµCT – top global tomography; bottom local tomography

brain, absorption-contrast SRµCT has a limited success, because the different soft tissues generally present very similar X-ray absorption properties. For this kind of tissue, phase-contrast SRµCT promises excellent results where, contrary to the absorption contrast, the real part of the refractive index often expressed in terms of its decrement from unity $\delta(x,y,z)$ is measured. For X-ray energies far away from the absorption edges, it relates to the electron density distribution $\rho_e(x,y,z)$:

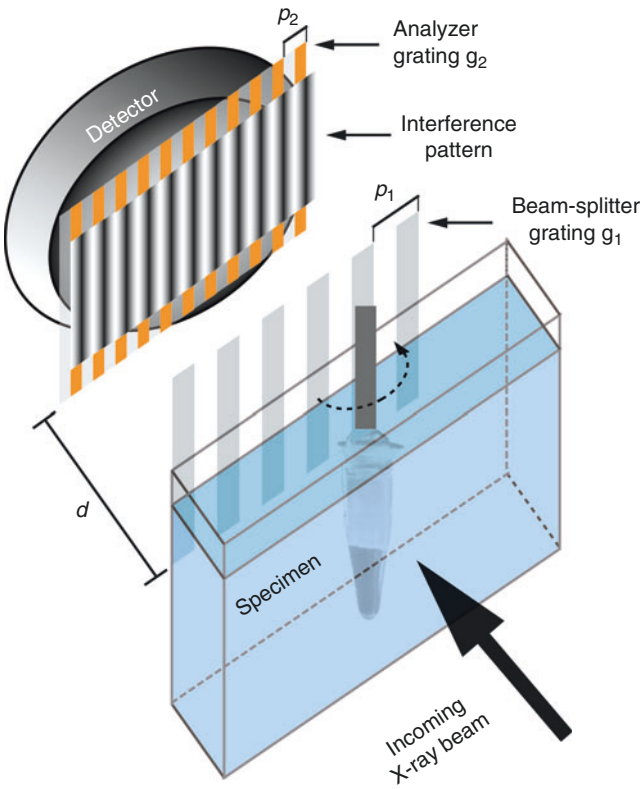
$$\delta(x,y,z) = \frac{r_e \lambda^2}{2\pi} \rho_e(x,y,z) \qquad (1)$$

where r_e denotes the classical electron radius and λ the X-ray wavelength [4]. In particular, a very high sensitivity can be reached by the recently developed grating-based, phase-contrast SRµCT [4]. A schematic view of the setup is given in Fig. 5. The grating interferometer is composed of a beam-splitter grating g_1 consisting of silicon stripes with periodicity p_1 and an analyzer grating g_2 consisting of strongly X-ray-absorbing stripes, often gold, with periodicity p_2. The periodicity of the grating g_2 has the half value of p_1. The grating g_1 induces a periodic fringe pattern in the X-ray wave-front intensity distribution with extreme values of the contrast at so-called Talbot distances from the beam-splitter grating. The Talbot distances are found at

$$d_n = \frac{n p_1^2}{8\lambda}, \qquad (2)$$

where n is a positive integer. The intensity difference between fringe maxima and minima is largest at the odd Talbot distances from the beam-splitter grating, while it is vanishing at even Talbot distances [5]. When performing a phase-contrast µCT experiment, the distance between g_1 and g_2 is adjusted to a Talbot distance with odd Talbot order. A specimen in the beam slightly deflects the incoming X-rays according to its refractive indices and leads to a disturbed fringe pattern. The analyzer grating permits detecting these changes using the phase-stepping method, where several images at different grating g_2 positions are recorded [4]. The quantitative relation between the local beam propagation $\alpha(y,z)$, the wave phase shift $\Phi(y,z)$, and the decrement of X-ray refractive index $\delta(x,y,z)$ can be obtained by the equation

$$\alpha(y,z) = \frac{\lambda}{2\pi} \frac{\partial \Phi(y,z)}{\partial y} = \int_{-\infty}^{\infty} \frac{\partial \delta(x,y,z)}{\partial y} dx. \qquad (3)$$



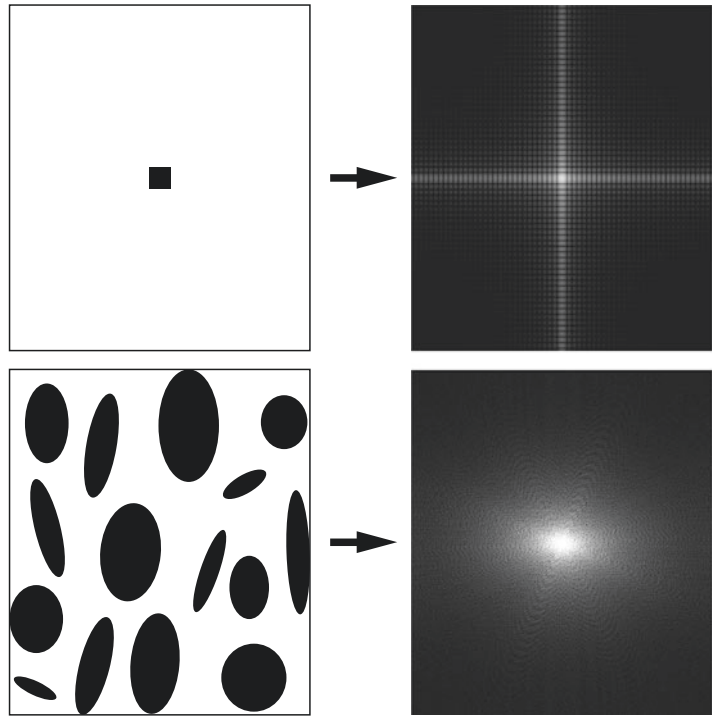
Imaging the Human Body Down to the Molecular Level, Fig. 5 Scheme of a grating-based interferometer for X-ray imaging at a third-generation synchrotron radiation source

The investigation of human tissues in the nanometer range by means of X-rays is available through scattering experiments, which belong to the reciprocal imaging techniques (see also ▶ Small-Angle Scattering of Nanostructures and Nanomaterials). In SAXS experiments, a focused X-ray beam is scattered to small angles between approximately 0.2 and 20 mrad. In order to record these angles with enough precision, the distance between the specimen and detector amounts to several meters. The scattering signal is closely related to the Fourier transform of the electron density distribution of the specimens [3]. Figure 6 shows a schematic representation of the relationship between specimens (on the left) and scattering signal (on the right). In the first example of Fig. 6, the particle that causes the scattering signal cannot easily be recognized from the scattering pattern. The information concerning particle shape, size, and orientation, however, can be extracted. When different particles scatter simultaneously, the generated signal contains contributions from all of them, yielding a mean value from the illuminated volume.

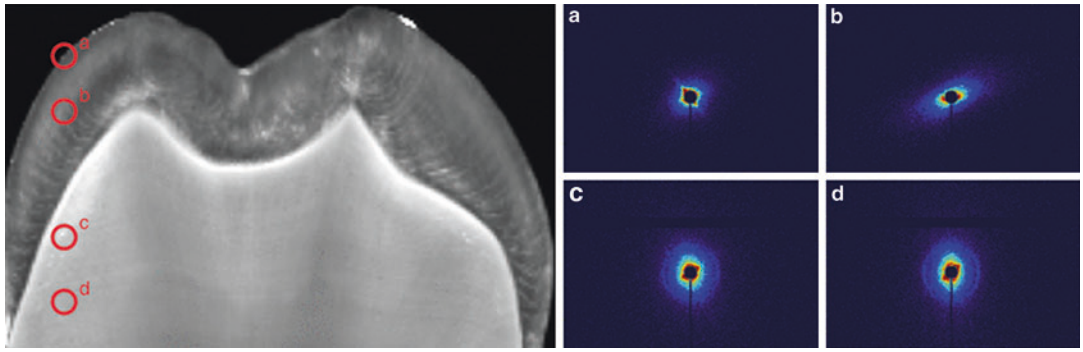
The continued improvement in detector technology [6] enables the acquisition of several tens of scattering patterns per second, allowing collecting scattering data over macroscopic areas in short time with a scanning approach [7]. Figure 7 elucidates the principle of scanning SAXS. A micro-focus X-ray beam is raster scanned along *x*- and *y*-directions, while the scattering signal is recorded. Scanning with steps of several 10 μm, the nanostructure of the specimen in reciprocal space is revealed with micrometer resolution in real space.

Key Research Findings

Grating-based phase-contrast SRμCT provides excellent measurement sensitivity for soft tissues. Figure 8 shows three virtual cuts of a human cerebellum acquired at the beamline ID19

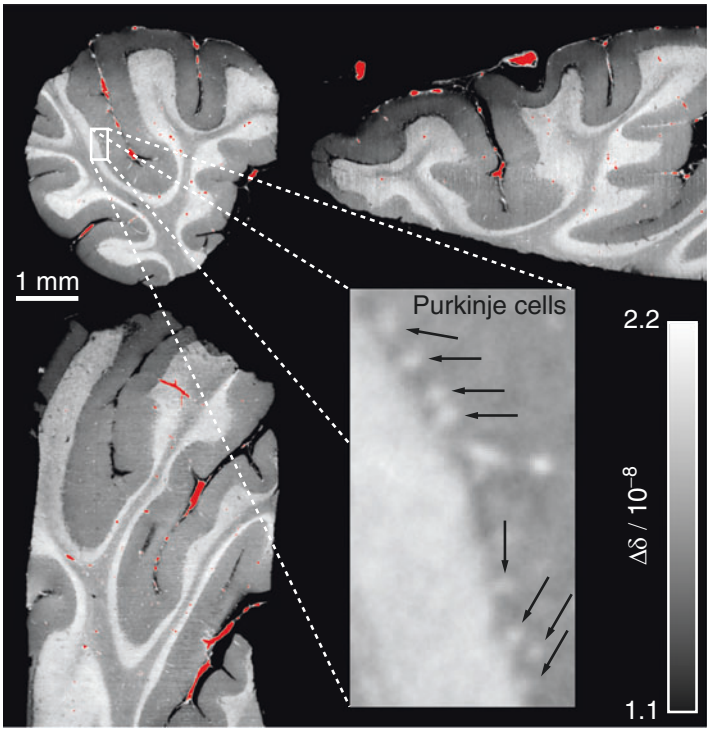


Imaging the Human Body Down to the Molecular Level, Fig. 6 Schematic representation of the information obtained from small-angle X-ray scattering measurements



Imaging the Human Body Down to the Molecular Level, Fig. 7 Principle of scanning SAXS, elucidated on a tooth slice. The specimen is raster scanned through the beam, and at each raster point a SAXS pattern is recorded. On the *right*, the scattering patterns corresponding to points a through d are shown

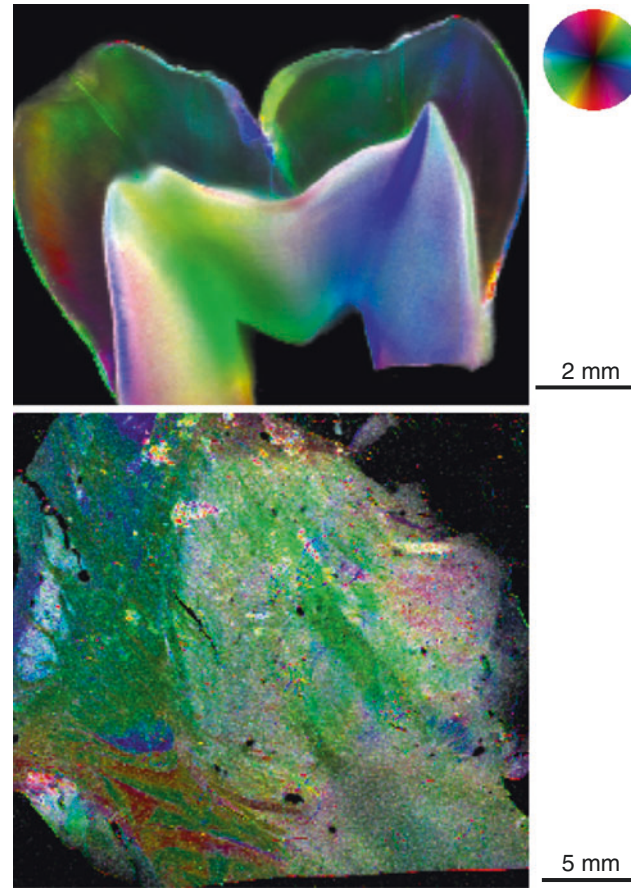
(European Synchrotron Radiation Facility, France) [8] in phase-contrast mode with a photon energy of 23 keV. For the experiment, the specimen was fixated in 4 % formalin solution. The period of the beam-splitter grating p_1 was 4.785 µm and that of the analyzer grating p_2 2.400 µm. The experiment was carried out at the ninth Talbot distance; thus a distance d of 479.4 mm between the two gratings was adjusted. The Eppendorf container with the cerebellum in formalin solution was fixed on the high-precision rotation stage and immersed in a water bath (see Fig. 5) in order to minimize artifacts owing to X-ray phase curvature induced by the container surface. For the detection, a lens-coupled scintillator and charge-coupled device (CCD) system using a FReLoN 2 K (Fast-Readout, Low-Noise, ESRF Grenoble, France) CCD with 2048 × 2048 pixels with the effective pixel size of 5.1 µm was used. Four phase-stepping images



Imaging the Human Body Down to the Molecular Level, Fig. 8 Three orthogonal virtual cuts of a human cerebellum acquired at the beamline ID19 (European Synchrotron Radiation Facility, France) using grating-based phase-contrast SRμCT. The red-colored structures correspond to blood vessels. The inset illustrates the visualization of individual, non-stained Purkinje cells having diameters of around 50 μm

were taken over one period of the interferometer fringe pattern at each projection angle. With an exposure time of 1 s per image, 1501 radiographs were recorded over a range of 360°. The spatial resolution of the experimental data was estimated by means of Fourier analysis. The calculation resulted in (16.5 ± 0.5) μm for the processed projections and (20 ± 1) μm for the reconstructed tomogram. Figure 8 demonstrates that the phase-contrast results allow for a clear differentiation between white (inner gray-colored layer) and gray matter and even a distinction within the gray matter, namely, the stratum moleculare (dark region of the gray matter) and the stratum granulosum (bright region of the gray matter). The red-colored regions correspond to blood vessels. The inset in Fig. 8 illustrates the visualization of individual Purkinje cells, which underlines the performance of the technique in postmortem brain imaging. The detection of individual Purkinje cells with spherical diameters of 40–70 μm in the surrounding soft tissues without the application of any stain or contrast agent is a novelty in CT imaging [2].

Human tissues often exhibit strong structural anisotropies of their nanometer-sized components, which are usually related to their function. SAXS allows inspecting hard and soft tissues [9–12], as illustrated in Fig. 9, where processed data from a slice of a human molar (top image, scattering from 162 to 183 nm features) and of a human thalamus (bottom image, scattering from 98 and 110 nm features) are shown. The mean orientation of the scattering signal is according to the color wheel. The brightness relates to the abundance of nanostructures in the corresponding ranges, while color saturation codes the nanostructures anisotropy (cp. Chapter "Nanodentistry"). The tooth was scanned at a photon energy of 18.6 keV with 50 μm raster scan steps in both *x*- and *y*-directions, while the thalamus data were acquired at 11.2 keV with a step width of 75 μm. For both experiments, the distance between the specimen and detector was set to 7.1 m. The scanning SAXS measurements reveal a high degree of anisotropy on the nanometer level in the tooth hard tissues. The orientation of the calcium phosphate crystallites in the



Imaging the Human Body Down to the Molecular Level, Fig. 9 Processed SAXS data of slices of a human molar in the range corresponding to 162 to 183 nm and of a human thalamus in the range between 98 and 110 nm

enamel is found to be mainly perpendicular to the tooth surface, thus pointing in the direction of maximal mechanical load. The crystallites in dentin are, however, mainly oriented parallel to the tooth surface. It is assumed that this organization is responsible for the extraordinary mechanical performance and lifetime of the human teeth. The bright region on the right-hand side of the brain slice of Fig. 9 is the thalamus, which consists mainly of gray matter, and the so-called nuclei, which belong to the central nervous system. Within the thalamus, only a few small regions containing highly oriented nanostructures can be found. In contrast to gray matter, white matter consists of long myelinated nerve cells (axons), which transfer the nerve impulses between the nuclei. This tissue can be seen on the left-hand side of the brain slice, where the colors are saturated.

Future Research Directions

The human body is fascinating because of its huge complexity. Clinical CT facilities reach a resolution of a fraction of a millimeter. This limit will not be much shifted toward true micrometer scale, as (i) the amount of data has to be (semi-) automatically analyzed and (ii) the dose becomes dangerous for the patient. Dose reduction has been one of the major research topics in clinical CT. For postmortem visualization, μCT in absorption-contrast mode has been developed to a standard technique. The contrast for soft tissues is so weak that phase-contrast methods are required. Fine-tuning of the most advanced systems is a prerequisite to reveal cellular structures within the organs of the human body. For the visualization of the nanostructures, however, scattering methods seem to be much better suited.

Besides the well-established 2D projection data, several approaches for the 3D tomographic reconstruction of tissue anisotropy and orientation were recently presented [13–16]. These techniques require sophisticated reconstruction algorithms and extended data acquisition times in the order of tens of hours. The wealth of information obtained from such data will however warrant increased interest in these techniques, leading to the development of optimized, faster, and more powerful acquisition protocols.

The structural information of the human body down to the nanometer level will provide a sound basis for developing nature analogue biomaterials and biomimetic implants [17] to be launched for the benefit of patients. These biomimetic products will exhibit micro- and nanostructures including their preferential orientation (anisotropy) as found in nature and revealed by highly advanced imaging techniques applying sophisticated software tools for data treatment and medical image analysis. Learning from nature will lead to solutions of health problems in our aging society.

Cross-References

- ▶ AFM in Liquids
- ▶ Atomic Force Microscopy
- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Biomimetics
- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ Models for Tumor Growth
- ▶ Nanodentistry
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanomedicine
- ▶ Nanostructured Functionalized Surfaces
- ▶ Nanotechnology in Cardiovascular Diseases
- ▶ Scanning Electron Microscopy
- ▶ Selected Synchrotron Radiation Techniques
- ▶ Self-Repairing Materials

References

1. Lareida, A., Beckmann, F., Schrott-Fischer, A., Glueckert, R., Freysinger, W., Müller, B.: High-resolution X-ray tomography of the human inner ear: synchrotron radiation-based study of nerve fiber bundles, membranes, and ganglion cells. J. Microsc. **234**, 95–102 (2009)
2. Schulz, G., Weitkamp, T., Zanette, I., Pfeiffer, F., Beckmann, F., David, C., Rutishauser, S., Reznikova, E., Müller, B.: High-resolution tomographic imaging of a human cerebellum: comparison of absorption and grating based phase contrast. J. R. Soc. Interface **7**, 1665–1676 (2010)
3. Guinier, A., Fournet, G.: Small Angle Scattering of X-Rays. Wiley, New York (1955)
4. Weitkamp, T., Diaz, A., David, C., Pfeiffer, F., Stampanoni, M., Cloetens, P., Ziegler, E.: X-ray phase imaging with a grating interferometer. Opt. Express **13**, 6296–6304 (2005)
5. Weitkamp, T., David, C., Kottler, C., Bunk, O., Pfeiffer, F.: Tomography with grating interferometers at low-brilliance sources. Proc. SPIE 6318, 63180S (2006)
6. Heinrich, B., Bergamaschi, A., Brönnimann, C., Dinapoli, R., Eikenberry, E.F., Jhonson, I., Kobas, M., Kraft, P., Mozzanica, A.: PILATUS: a single photon counting pixel detector for X-ray applications. Nucl. Instrum. Methods Phys. Res. A **607**, 247–249 (2009)
7. Bunk, O., Bech, M., Jensen, T.H., Feidenhans'l, R., Binderup, T., Menzel, A., Pfeiffer, F.: Multimodal x-ray scatter imaging. New J. Phys. **11**, 123016 (2009)
8. Weitkamp, T., Tafforeau, P., Boller, E., Cloetens, P., Valade, J.-P., Bernard, P., Peyrin, F., Ludwig, W., Helfen, L., Baruchel, J.: Status and evolution of the ESRF beamline ID19. AIP Conf. Proc. **1221**, 33–38 (2010)
9. Müller, B., Deyhle, H., Bradley, D., Farquharson, M., Schulz, G., Müller-Gerbl, M., Bunk, O.: Scanning x-ray scattering: evaluating the nanostructure of human tissues. Eur. J. Nanomed. **3**, 30–33 (2010)
10. Deyhle, H., Bunk, O., Müller, B.: Nanostructure of healthy and caries-affected human teeth. Nanomedicine **7**, 694–701 (2011)
11. Deyhle, H., White, S.N., Bunk, O., Beckmann, F., Müller, B.: Nanostructure of the carious tooth enamel lesion. Acta Biomater. **10**, 355–364 (2014)
12. Gaiser, S., Deyhle, H., Bunk, O., White, S.N., Müller, B.: Understanding nano-anatomy of healthy and carious human teeth: a prerequisite for nanodentistry. Biointerphases **7**, 4 (2012)
13. Georgiadis, M., Guizar-Sicairos, M., Zwahlen, A., Trüssel, A.J., Bunk, O., Müller, R., Schneider, P.: 3D scanning SAXS: a novel method for the assessment of bone ultrastructure orientation. Bone **71**, 42–52 (2015)
14. Liebi, M., Georgiadis, M., Menzel, A., Schneider, P., Kohlbrecher, J., Bunk, O., Guizar-Sicairos, M.: Nanostructure surveys of macroscopic specimens by small-angle scattering tensor tomography. Nature **527**, 349–352 (2015)
15. Malecki, A., Potdevin, G., Biernath, T., Eggl, E., Willer, K., Lasser, T., Maisenbacher, J., Gibmeier, J., Wanner, A., Pfeiffer, F.: X-ray tensor tomography. Europhys. Lett. **105**, 38002 (2014)

16. Schaff, F., Bech, M., Zaslansky, P., Jud, C., Liebi, M., Guizar-Sicairos, M., Pfeiffer, F.: Six-dimensional real and reciprocal space small-angle X-ray scattering tomography. Nature **527**, 353–356 (2015)
17. Mülle, B.: Biomimetics and medical implementations. In: Bar-Cohen, Y. (ed.) Biomimetics: Nature-Based Innovation. Taylor & Francis Group, Boca Raton (2011)

Immersion Lithography

Roger H. French
Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, OH, USA

Synonyms

High index fluid immersion; Liquid Immersion lithography; Optical lithography by liquid immersion

Definition

Immersion lithography utilizes liquids with refractive indices >1 (the index of air) to enable sub-45-nm feature size (or half-pitch) patterning for semiconductor integrated circuits. An alternative to conventional "dry" optical lithography offering finer features, this method employs innovative materials at both the photomask and imaging stack levels.

Overview

In addition to the entry on ▶ Immersion Lithography Materials, herein, the reader is directed to two recent review papers which surveyed the state of the art in immersion lithography, including photomasks and wafer level materials [1] and immersion lithography tools [2].

Basic Methodology

A modern immersion lithography tool, a scanner, is shown schematically in Fig. 1 such that the different basic elements are visible. The illuminator, which prepares the ArF excimer laser light (the light source for 193.6 nm lithography) is on the right, the photomask (which contains the desired circuit layout pattern) is on the left above the large cylindrical projection lens, and the dual wafer stages are shown below. The left wafer stage undergoes alignment procedures while the right wafer stage, under the last lens element of the projection lens, is exposing the fields of the wafer in water immersion conditions. Once the wafer completes its exposure, the two wafer stages exchange positions, and exposure of the second wafer stage begins.

Microfabrication and Immersion Lithography

Microfabrication – first developed for integrated circuit processing – has relied on photolithography as a critical pattern transfer process. From the first transistor in 1947 and the first integrated circuit (IC) or semiconductor device in 1959, the basic microfabrication process steps have been deposition, etch, implant, and patterning [3]. Patterning has included photolithography in either contact or proximity modes (where the photomask and the IC have the same 1:1 feature size ratio) as well as the now dominant projection mode (where there is a 4:1 or 5:1 optical reduction of the mask features to the IC feature sizes). Rayleigh's equations provide a simple relationship between the stepper/scanner optical parameters (such as the lithographic wavelength (λ), the lens numerical aperture (NA), and the imaging medium's refractive index (1 for air and higher for immersion fluids)) and the feature size (FS) and depth of focus (DOF) which can be achieved. These equations also include a process latitude factor (k) where k_1 corresponds to the *FS* (Eq. 1) and k_2 to the *DOF* (Eq. 2), and where the scanner's numerical aperture is defined as $NA_{\text{System}} = NA_{\text{Dry}}\, n_{\text{imaging Medium}}$.[4] Microfabrication's progression down this feature size path (as presaged by Moore's Law [5]) for complementary metal oxide semiconductor (CMOS) devices has been enabled by technological advances in optics, usable lithographic wavelengths, and, most recently, immersion fluids to allow projection lens numerical apertures above 1. During this



Immersion Lithography, Fig. 1 ASML TwinScan XT:1950i Scanner showing the illuminator, the reticle (photomask) and reticle handler, the projection lens and the wafer stages. The numerical aperture of this water immersion stepper is 1.35, and the productivity is >140 wafers per hour imaged (Image courtesy of ASML Inc., Veldhoven, Netherlands)

evolution, microfabrication and optical lithography have also expanded beyond semiconductor ICs to include flat panel displays, microelectronic machines and devices, photovoltaic solar cells, and other applications, further driving the motivation to improve optical lithography's capabilities. In addition, as IC feature sizes have decreased, physical limits on materials and processes have led to new materials and device architectures, such as the IC interconnect conversion from aluminum to dual damascene copper metallization, the transition from silica interconnect dielectrics and gate dielectrics, and new gate architectures (such as the strained gate or 3D gate geometries of FinFET devices [6]). Microfabrication's widening applications, and the incorporation of new materials, process technologies, and architectures into IC fabrication, demonstrate the broad platform that microfabrication represents in materials research, as well as the large technology base it provides to the semiconductor, display, and photovoltaic industries. See Table 1 for lithographic half-pitch generations with technology options and imaging parameters.

$$FS = k_1 \lambda / NA \tag{1}$$

$$DOF \approx k_2 \lambda / NA^2 \tag{2}$$

Key Research Findings or Key Attributes and Advantages

Phase Shift Photomasks

With the advent of subwavelength lithography, the photomask's (see Fig. 2) imaging properties (i.e., the "object" in the scanner which is imaged into the photoresist [7]) have attained critical importance, such that full 3D imaging codes are used, and the opaque chrome film used in binary intensity masks (BIM) has changed to become

Immersion Lithography, Table 1 Lithographic half-pitch generations with technology options and imaging parameters

Hp (nm)	Technology	Commercialized?	λ (nm)	Index	NA	k1
400	365 Dry	Yes	365	1	0.6	0.66
300	365 Dry	Yes	365	1	0.65	0.53
250	248 Dry	Yes	248	1	0.63	0.64
180	248 Dry	Yes	248	1	0.8	0.58
130	248 Dry	Yes	248	1	0.8	0.42
90	248 Dry	Yes	248	1	0.93	0.34
90	193 Dry	Yes	193.4	1	0.75	0.35
90	193 Dry	Yes	193.4	1	0.85	0.40
90	157 Dry	No	157	1	0.75	0.43
65	157 Dry	No	157	1	0.85	0.35
65	193 Dry	Yes	193.4	1	0.93	0.31
65	193 H2O SP	Yes	193.4	1.44	0.93	0.31
55	193 H2O SP	Yes	193.4	1.44	1.2	0.34
45	193 H2O SP	Yes	193.4	1.44	1.35	0.31
38	193 H2O SP	Yes	193.4	1.44	1.35	0.27
38	193 Gen2 SP	No	193.4	1.64	1.55	0.30
32	193 Gen2 SP	No	193.4	1.64	1.55	0.26
32	193 Dry DP	No	193.4	1	0.85	0.14
32	193 H2O DP	Yes	193.4	1.44	1.35	0.22
32	EUV	No	13.5	1	0.25	0.59
28	193 Gen3 SP	No	193.4	1.9	1.8	0.26
22	193 H2O DP	No	193.4	1.44	1.35	0.15
22	EUV	No	13.5	1	0.32	0.52

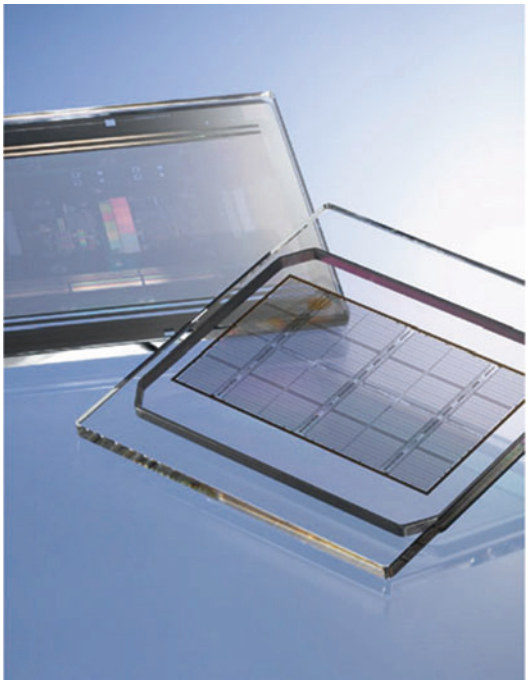
both thinner and more opaque. In addition, introduction of resolution enhancement techniques such as phase shift masks (PSM) and optical proximity corrections (OPC) have broken the simple and long-standing idea that the mask (object) is a replica of the desired circuit layout (image). Instead, the full scanner imaging must now be considered, including the properties of both the object and the image stack. This results in a complex optimization that is tied to the desired circuit pattern (such as contact layer, gate, interconnect), the particular scanner used, the particular photomask and type (BIM, Attenuating PSM, Alternating PSM, etc.), and the wafer level imaging stack. Lithography has now become computationally intensive, with iterative refinements needed to achieve accurate fidelity and yield in the fabrication facility ("fab"). These requirements lead to detailed thin film optimization of binary and phase shift masks, and of the pellicle's effects on lithographic imaging (e.g., optical proximity corrections), and also on overall mask flatness. In the 1970s or 1980s, before these advances, imaging was considered simple; now, mask and wafer level materials development must be closely coupled to lithographic image modeling to achieve the desired effects, while minimizing undesirable artifacts.

Pellicles: Dust Protectors for Photomasks

Pellicles are thin polymer membranes supported on aluminum frames and mounted on a photomask's patterned film side – typically 6 mm above the masking film (Fig. 2). They assure that any dust particles falling onto the photomask are out of the object plane of the objective lens, and therefore are not present in the image formed in the photoresist layer by the photomask. Pellicles are one of the critical elements that permit defect-free lithographic printing.

Development and Performance of Immersion Lithography

Immersion lithography has led to radical changes in lithography tools, that is, the design of the



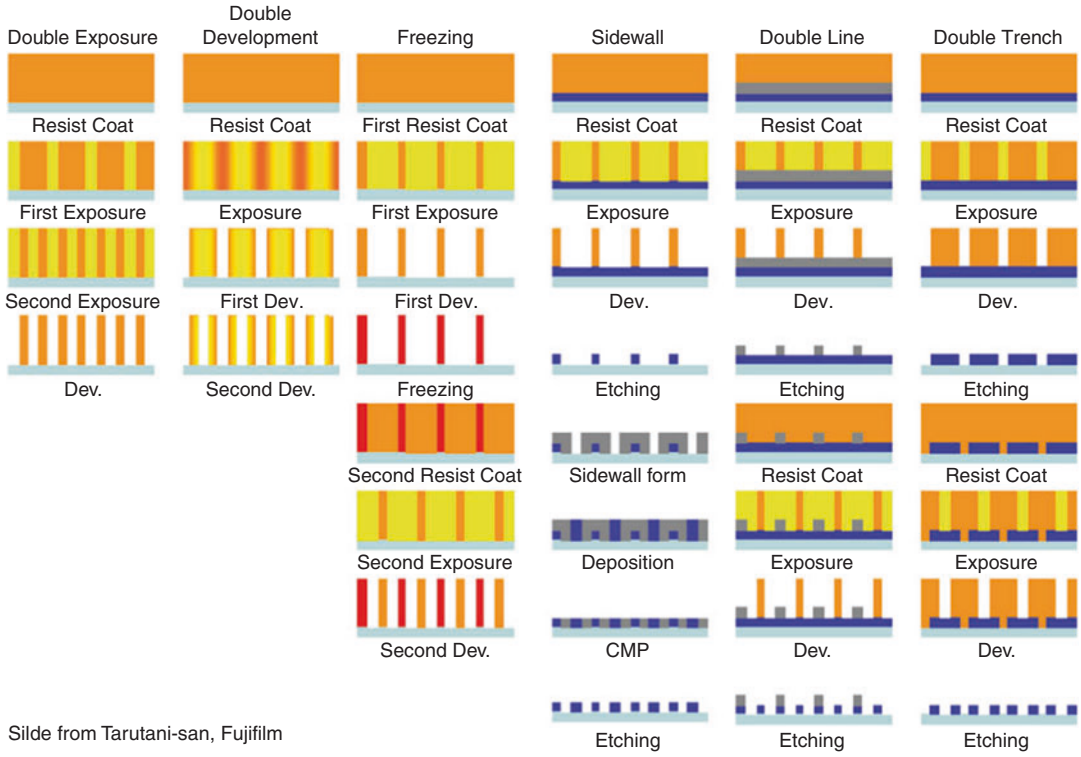
Immersion Lithography, Fig. 2 Two photomasks with pellicles. The *upper* mask is a conventional binary intensity mask with the mask pattern and the pellicle on the *lower* surface of the 6″ × 6″ ultrahigh purity fused silica photomask substrate. The *lower* mask in the figure is a negative tone phase shift mask, and the mask pattern and the pellicle are on the *upper* surface of the substrate (Photograph courtesy of Toppan Photomasks Inc. Round Rock Texas)

optical scanner, so that the immersion fluid can be contained between the projection lens' last lens element and the photoresist coated silicon wafer, thereby permitting the high wafer stage translation speeds needed to produce high process throughput (greater than 120 wafers per hour). Scanner manufacturers (ASML [8], Nikon [9, 10], and Canon) have chosen water as the first-generation immersion fluid and have evaluated many different approaches to immersion lithography's technical challenges. For example, different immersion fluid containment methods were considered, including a fluid bath in which the wafer would reside, a showerhead in which the fluid would continuously flow and would be deposited and removed from the wafer, and another in which the fluid would be deposited on the wafer but flow over the wafer edges to a drain. The final solution seems to be the showerhead where the fluid is flowing continuously, with the water temperature maintained to very high precision to minimize any imaging artifacts arising from water's refractive index change with temperature (i.e., dn/dT). Another critical aspect, necessary to control the number of immersion-specific defects per wafer, is to carefully control the showerhead's movement over the wafer's edge. Another important consideration is the track: the wafer processing and development system directly attached to the photolithographic scanner, which is used to coat the imaging stack (a bottom antireflection coating (BARC), the photoresist layer, and sometimes a topcoat) and perform any necessary post-apply thermal bake steps. The track then supplies the wafers into the scanner at a rate of more than 120 wafers an hour, and receives the exposed wafers back for postexposure bake, development in TMAH (Tetramethyl ammonium hydroxide), and spin drying. Typically, photoresist is removed from a small rim around the wafer's edge, using an edge bead remover in the original photoresist coating process, since the wafer's high speed and fast acceleration, relative to the stationary showerhead, can lead to particulate delamination from the wafers edge, generating immersion defects on the final imaged wafer.

Ongoing immersion lithography materials research has included developing next generation immersion fluids [11] (beyond immersion fluids for use in 157 nm lithography [12]), which have higher refractive indices than water, and so further increase the scanner's numerical aperture.

Double Exposure Immersion Lithography

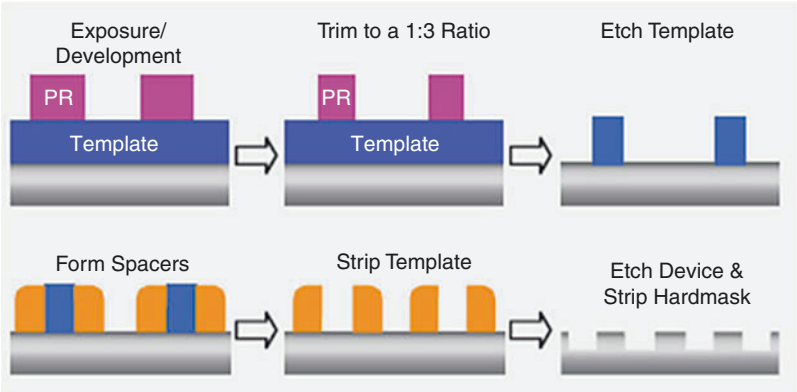
Strategies to further increase resolution by optical means alone may have been exhausted, since decreasing the wavelength beyond 193 nm leads to extreme ultraviolet (EUV; see "► EUV Lithography" herein) lithography (also known as "soft-x-ray projection lithography"; see below). Also, increasing the numerical aperture to 1.55 with high index fluids will only allow single patterning down to 32 nm half pitch, even at the most aggressive k factors. The most likely alternative would be some form of double patterning, where in one variation the desired patterns are split between two masks, such that more facile feature printing



Immersion Lithography, Fig. 3 Approaches for double patterning lithography, going from least complex in terms of processing steps (but most material demanding) on the *left*, to most complex process (but least material demanding) on the *right* (After Ref. [13])

is done at twice the pitch. The second mask would be aligned so that the second feature set is printed between the first set, providing final features with target dimensions at the desired half pitch. For example, if one wanted to print 30 nm lines and spaces, one would start by printing 30 nm lines with 90 nm spaces, then print another set of 30 nm lines with 90 nm spaces over the original image, but aligning the exposures with a 60 nm offset. One can imagine that the costs associated with double patterning lithography would be substantial given the extra processing steps and materials required, yet today some chip makers are contemplating triple and quadruple patterning.

However, since EUV lithography has not progressed as quickly as initially expected, and optical lithography with water or high index fluid immersion has shown much potential, applying some double patterning method to immersion lithography would provide a means to improve resolution to keep pace with integrated circuits' leading edge feature size requirements. Several industrial laboratories and academic institutions are developing their own approaches for accomplishing this double patterning task. A few of these approaches are shown schematically in Fig. 3 [13]. The last two columns depict the double patterning method described above, where two exposure steps, two development steps, and two etching steps are required for pitch doubling. This method's costs have prompted researchers to scout other alternatives that could eliminate some of the extra steps. The most challenging approach is a double exposure method shown in the first column. In this method, the two mask patterns are exposed consecutively into the same material, providing better overlay accuracy (since the wafer is not removed from the scanner chuck). This requires a material that can behave in an optically nonlinear manner, and to date no such material exists, although some conceptual and simulation work has been reported.



Immersion Lithography, Fig. 4 Generic process flow for positive tones (*lines*) spacer mask patterning (From Ref. [14])

Another alternative uses a self-aligned double patterning (SADP) technique, or sidewall spacer technique, to achieve pitch doubling [14]. In this process, shown schematically in Fig. 4, the first optical exposure pattern is etched into a sacrificial layer. Another material (e.g., silicon nitride) is then deposited conformally on this spacer's sides to a thickness that defines the final pattern. In positive tone spacer processes, the first sacrificial layer is removed and the pattern formed by the deposited material is etched into the final hard mask or substrate. Excellent quality lines down to 22 nm half pitch have been obtained [15]. Although this process involves only a single resist application, it does add a number of process steps, contributing to the approach's complexity. As with the double development approach, this technique may not find general use, since it is applicable only to uniformly periodic structures.

Future Directions Beyond Optical Lithography

Nanofabrication

During photolithographic patterning's ongoing development for microfabrication, there have been parallel efforts on novel new technologies to displace photolithography. For example, x-ray proximity lithography, ion beam and e-beam projection, and e-beam direct write lithography have all been the focus of large and substantial research and development efforts. However, other than the critical role that e-beam direct write has played in photomask patterning for optical lithography, none of these technologies has succeeded in displacing optical projection lithography. E-beam direct write has been pursued in both single e-beam and massively parallel multibeam modes – the latter to improve the technology's pattern transfer rate and to compete with projection optical lithography's productivity.

Soft Lithography

There has been much recent work on "soft lithography": imprint lithography whereby either stamping with an elastomeric patterned mask, or hard mask contacting into either photo or thermally polymerizable materials, allows direct pattern transfer [16]. This nanoimprint technology, discussed in an article in this volume by Willson, has shown the ability to make very fine feature sizes. There is also recent work on dip pen lithography, or scanning probe lithography, which (like e-beam direct write) utilizes an atomic force microscope tip to directly write the pattern, either with a single probe, or with a probe array. In addition, there is much research on both self-assembly of small molecules and oligomers, as well as "templated growth": using small structural features (which may be introduced by patterning), to construct complex functional materials or devices using materials growth on the template.

EUV Lithography

There has been substantial research in the past 15 years on soft-x-ray projection lithography

(now called EUV lithography) using 13.4 nm light and multilayer reflective mirror optics; this is now considered a "next generation lithography" to replace optical projection lithography.

It appears that the next generation lithography, determined when it is successfully brought to commercial-scale IC fabrication, will be either EUV lithography, imprint lithography, or massively parallel e-beam direct write [16]. But it is certain that 193 nm lithography will be extended as far and as long as it is commercially viable; this ongoing extension will make use of as many optical enhancements (phase shift, polarization) and variations such as double (and perhaps triple or quadruple) patterning, thus providing an ongoing set of R&D challenges for materials scientists.

Cross-References

- Dip-Pen Nanolithography
- DUV Photolithography and Materials
- EUV Lithography
- Nanoimprint Lithography
- Nanoimprinting
- Nanoscale Printing
- Self-Assembly
- Self-Assembly of Nanostructures
- Stereolithography
- SU-8 Photoresist

References

1. French, R.H., Tran, H.V.: Immersion lithography: photomask and wafer-level materials. Vol. on Materials advances for next-generation microelectronics. Annu. Rev. Mater. Res. **39**, 93–126 (2009)
2. Sewell, H., Mulkens, J.: Materials for optical lithography tool application. Vol. on Materials advances for next-generation microelectronics. Annu. Rev. Mater. Res. **39**, 127–153 (2009)
3. Deal, B.E., Crossley, P.A.: Microstructure fabrication in electronic devices. Annu. Rev. Mater. Sci. **11**, 321–351 (1981)
4. Lin, B.J.: The k_3 coefficient in nonparaxial l/NA scaling equations for resolution, depth of focus and immersion lithography. J. Microlith. Microfab. Microsys. **1**, 7–12 (2002)
5. Moore, G.E.:Progress in digital integrated electronics. In: International Electron Devices Meeting 1975 Technical Digest, pp. 11–13 (1975). Reprinted in Schellenberg, F.M. (ed.) Selected Papers on Resolution Enhancement Techniques in Optical Lithography. SPIE Milestone Series, vol. MS 178, p. 179. SPIE Press, Bellingham (2004)
6. Solomon, P.M.: Device innovation and material challenges at the limits of CMOS technology. Annu. Rev. Mater. Sci. **30**, 681–697 (2000)
7. Schellenberg, F.M., Adam, K., Matteo, J., Hesseling, L.: Electromagnetic phenomena in advanced photomasks. J. Vac. Sci. Technol. B **23**, 3106–3115 (2005)
8. Mulkens, J., Flagello, D., Streefkerk, B., Graeupner, P.: Benefits and limitations of immersion lithography. J. Microlith. Microfab. Microsys. **3**, 104–114 (2004)
9. Owa, S., Nagasaka, H.: Advantage and feasibility of immersion lithography. J. Microlith. Microfab. Microsys. **3**, 97–103 (2004)
10. McCallum, M., Kameyama, M., Owa, S.: Practical development and implementation of 193 nm immersion lithography. Microelectron. Eng. **83**, 640–642 (2006)
11. French, R.H., Sewell, H., Yang, M.K., Peng, S., McCafferty, D., Qiu, W., Wheland, R.C., Lemon, M. F., Markoya, L., Crawford, M.K.: Imaging of 32-nm 1:1 lines and spaces using 193-nm immersion interference lithography with second-generation immersion fluids to achieve a numerical aperture of 1.5 and a k1 of 0.25. J. Microlith. Microfab. Microsys. **4**, 031103 (2005)
12. Kunz, R.R., Switkes, M., Sinta, R., Curtin, J.E., French, R.H., Wheland, R.C., Kao, C., Mawn, M.P., Lin, L., Wetmore, P., et al.: Transparent fluids for 157-nm immersion lithography. J. Microlithogr. Microfabr. Microsys. **3**, 73 (2004)
13. Tarutani, S., Tsubaki, H., Kanna, S.: Development of materials and processes for double patterning toward 32 nm node ArF immersion lithography. J. Photopolym. Sci. Technol. **21**(5), 685–690 (2008)
14. Bencher, C.: SADP: the best option for <32 nm NAND flash. Nanochip Technol. J. **2**, 8–13 (2007)
15. Raub, A.K., Li, D., Frauenglass, A., Brueck, S.R.J.: Fabrication of 22 nm half-pitch silicon lines by single-exposure self-aligned spatial-frequency doubling. J. Vac. Sci. Technol. B Microelectron. Nanometer Struct. Process. Meas. Phenom. **25**(6), 2224–2227 (2007)
16. Costner, E.A., Lin, M.W., Jen, W.-L., Willson, C.G.: Nanoimprint lithography materials development for semiconductor device fabrication. Annu. Rev. Mater. Res. **39**, 155–180 (2009)
17. Lin, B.J.: The ending of optical lithography and the prospects of its successors. Microelectron. Eng. **83**, 604–613 (2006)

Immersion Lithography Materials

Roger H. French
Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, OH, USA

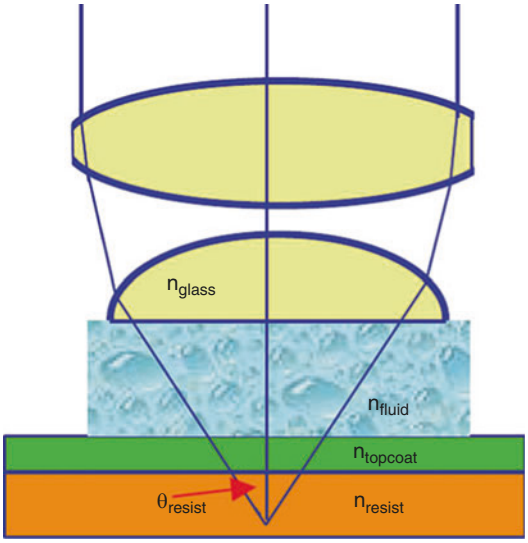
Definition

Immersion lithography utilizes liquids with refractive indices >1 (the index of air) to enable sub-45-nm feature patterning for semiconductors. An alternative to conventional dry optical lithography offering finer features, this method employs innovative materials at all imaging stack levels.

Overview

In addition to the entries on Lithography and Immersion Lithography, herein, the reader is directed to two recent review papers which surveyed the state of the art in immersion lithography, including photomasks and wafer-level materials [1] and immersion lithography tools [2].

Using optical lithography in semiconductor fabrication, the main materials development has traditionally been polymeric photoresists, which changed substantially each time the lithographic wavelength decreased: from 436 nm (G-line), to 365 nm (I-line), to 248 nm (DUV), and now to 193 nm. For example, the shift to DUV (KrF excimer laser) coincided with the introduction of chemically amplified resists. The transition to 193 nm (ArF excimer laser), along with a failed transition to 157 nm (F_2 excimer laser), expanded greatly the number and type of materials needed above and below the projection lens in lithographic scanners. These new materials include both photomask materials as well as new types of wafer-level materials, such as immersion fluids, top coats, and photoresists (Fig. 1). These new materials – photomasks, pellicles, photoresists, bottom antireflection coatings, and topcoats – coupled with the introduction of immersion lithography, have enabled much larger optical scanner numerical aperture lenses (up to 1.35 NA with water as the immersion fluid), enabling 45 nm half-pitch feature printing using 193 nm light.



Immersion Lithography Materials, Fig. 1 Immersion lithography imaging stack showing the projection lens optics including the last lens element contacting the immersion fluid, a topcoat layer on *top* of the photoresist layer (From [8])

Immersion lithography materials have become a broad, diverse, and complex family developed to help the industry's advance to 32 nm half-pitch feature sizes. At the same time, the possible uses of second-generation immersion fluids, double or triple patterning, and novel resist processes indicate that the need for new and optimized materials will not be reduced in the future.

Types of Materials Used

Photomasks are typically transparent fused silica blanks covered with a pattern defined in a chrome metallic absorbing film.

Pellicles are a thin transparent film membrane stretched over a frame that is glued over one side of the photomask to serve as a dust protector.

Photoresist is a light-sensitive polymer formulation which is spin coated onto a silicon wafer and used to define features in an integrated circuit.

Bottom antireflection coatings are an inorganic or organic thin film used between the photoresist and the silicon wafer so as to control the reflectance of light from the wafer back into the photoresist.

Topcoats are a polymer coating used above the photoresist layer to reduce interactions between the photoresist and the immersion fluid.

Key Attributes and Advantages

Attenuating Phase Shift Photomasks

Attenuating phase shift masks (AttPSM) have been widely accepted in lithography for providing phase contrast imaging without the extra process steps (e.g., trim masks) needed for strong phase shifting approaches such as alternating phase shift masks. Attenuating phase shift photomasks have been developed for I-line, DUV, and 193-nm lithography; the I-line masks met with commercial acceptance, while thin films based on molybdenum and silicon ("MoSi") have become the dominant materials used at DUV and 193-nm wavelengths. In these photomasks, the thin film attenuates the light to 6 % or 9 % of the mask's illuminating intensity, and also phase shifts this light by 180° relative to the light transmitted through the openings (the mask's lines) in the thin film. This attenuated phase shifted light then destructively interferes with the light transmitted by the lines, thereby reducing the lithographically printed line width.

AttPSMs for 248 and 193 nm lithography use essentially two component materials mixed in a microstructure so as to produce broadly tunable optical materials. Compositionally graded films of MoSi (more accurately, compounds of Mo and Si with O, N, and C) have become the accepted materials at these lithography wavelengths. An alternative AttPSM consisting of multilayer films of AlN with TiN (optical superlattices) produced by sequential multilayer ion-beam deposition, where the bilayer thickness was kept to less than λ/20 (i.e., bilayer thicknesses smaller than 10 nm), was also developed but not commercialized. These films were optically homogeneous, and behaved optically as uniform films, even though the AlN and TiN amounts in each bilayer were the variables used to produce optical tenability [3]. This material family produced APSMs with transmissions from 6 % to 20 % with 180° phase shift for both 248 and 193 nm. Materials design of new APSMs continues as a research active field [4].

Polymers for Photomask Pellicles

In G-line or I-line lithography, the pellicle membranes were initially fabricated from high molecular weight nitrocellulose polymers, with the membrane prepared by spin coating from a polymer solution onto a glass plate. The aluminum frame was then adhesively attached to the polymer membrane, followed by membrane release from the glass plate, then trimming the membrane edges at the frame and optical inspection of the pellicle for any defects. The pellicle thickness is chosen so that the membrane's interference fringes give a maximum transmission, such that the pellicle serves as an optical element with ∼100 % transmission. With the advent of DUV, lithograph amorphous fluoropolymers such as Teflon AF [5] or Cytop [6] were developed, because their optical absorption edges are at much shorter wavelengths than nitrocellulose, which was no longer sufficiently transparent. The pellicle membrane's thickness can be optimized to give its peak interference transmissions at multiple wavelengths; for example, a thickness on the order of 0.8 μm can be used for I-line and DUV lithography.

Materials in the Lithographic Imaging Stack

The immersion lithography materials required on the image side (i.e., below the projection lens) also span a wide variety of materials and classes. These materials include the immersion fluids, topcoats (used to separate the fluid from the imaging photoresist film), the photoresist layer, and an antireflection layer underneath. These are discussed in order below.

Immersion Fluids

Immersion fluids give enhanced imaging performance in the same way that water or oil

immersion techniques improve an optical microscope's resolution. When imaging is performed in a fluid which has a refractive index >1 (the value for air), the light's wavelength is effectively reduced by the fluid's index. The fluids must be optically transparent (i.e., have a low optical absorbance at the lithography wavelength) and have a high refractive index. The first immersion fluids studied were to be used to extend 157-nm lithography [6], but it is immersion lithography at 193 nm (using water) that has been successfully commercialized.

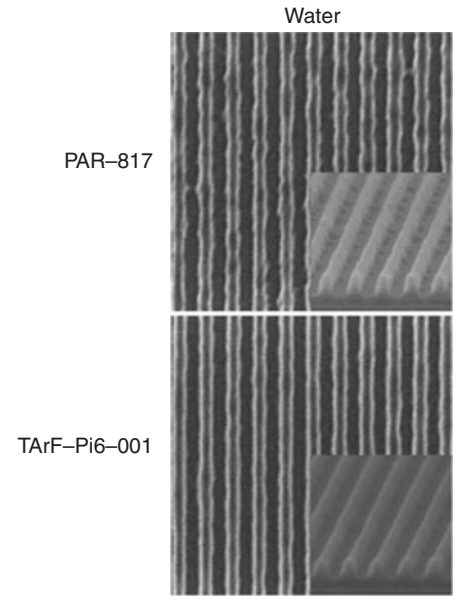
Water-based immersion lithography using 193 nm ArF illumination provides optical solutions as far as the 45-nm node (with a k_1 of 0.31), but is not able to achieve the 38- or 32-nm nodes as currently defined. Achieving these lithographic nodes will require new, higher refractive index fluids to replace the water used in first-generation immersion systems. Therefore, there has been extensive research to develop second- ($1.44 < n_{193nm} < 1.65$) and third-generation ($1.65 < n_{193\ nm} < 1.8$) 193 nm immersion fluids. A number of such second-generation high-index fluids have been developed for immersion lithography at 193 nm, with low optical absorbance and refractive indices up to 1.664. To understand different fluid classes' behavior and performance, spectral index measurements are used to characterize the index dispersion, coupled with Urbach [7] absorption edge analysis and Lorentz Oscillator modeling [8].

Immersion Fluid–Resist Interactions

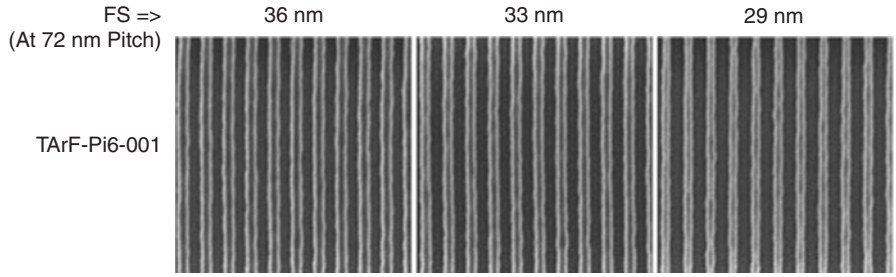
Using water or another higher index-matching liquid between the last lens element and the photoresist introduces new issues not previously encountered in conventional "dry" lithography. Much of the early work identified the key issues related to imaging under water immersion, and later, the same issues were addressed for the higher index organic liquids [9]. Of particular concern is the extent of resist components' extraction from the resist film into the immersion liquid. Components in the resist formulation such as photoacid generators (PAGs) and their photogenerated acids, dissolution modifiers, and bases may all leach from the film. The amount of leaching is of course a function of the resist film's component concentration, the rate of the components' diffusion into the liquid, the solubility of the components, and the time of fluid/resist contact. Leaching has two effects. First, the resist's imaging properties, such as sensitivity, are changed, and performance is usually degraded. Secondly, leached components, especially photoacid generators, may undergo further photochemistry and adsorb onto the system optics [10], a topic of particular concern to the stepper manufacturers – which has led to their setting maximum tolerable PAG limits on the order of $1.0–5.0 \times 10^{-12}$ mol/cm^2-s. The PAG amount detected in water for some commercially available resists has either come very close to or exceeded this limit. This result has prompted further work on designing immersion topcoats or topcoatless resists (vide infra) to slow or prevent leaching. It has also led to work to pre-rinse the resist surface with water (since most of the PAG leaching occurs in the first few seconds of contact), which has been reported to remove over 90 % of the PAG at the water/resist interface [11]. With second-generation immersion fluids, very little PAG leaching has been detected, reportedly about two orders of magnitude lower than the recommended limit.

Another key immersion imaging issue is the effect of fluid penetration into the resist film. Sub-surface immersion liquid may modify the kinetics of the acid-catalyzed deprotection reaction (especially if water carries with it species that can quench photoacid). It can also act as a polymer film plasticizer (especially true for organic liquids diffusing into organic resist polymers), changing the film's mechanical properties and diffusion-related properties. With water as the immersion fluid, quartz crystal microbalance and optical reflectance techniques have shown that 200 nm films may swell up to 5 nm in thickness. There is also some evidence that second-generation immersion fluids cause up to 2 nm in initial swelling and 10 nm of dissolution of some commercial resists. Despite these observations, there seem to be only subtle effects on sub-45-nm line and space interference imaging of some commercially available resists with both water and second-generation immersion fluids (vide infra).

Resist component extraction into the immersion fluid and fluid permeation into the resist may each cause unwanted defects in the final resist image, and much work has been done to quantify and reduce the new defect types seen under immersion imaging. These defect types may be classified into five general categories: process-induced defects, "watermark" defects, defects arising from bubbles in the fluid, drying stain defects, and particle-borne defects. Defects related to processing may originate from fluid carrying particles from the wafer edge, especially if the topcoat, resist, or BARC peels at the edge, and the fluid is scanned beyond this edge point. Steps to rinse and clean the bevel have shown marked improvement in defect counts. Fluids that permeate into the resist film can generate "watermark" defects, which include swelling of lines and T-topping. Bubbles (which could originate from gaseous by-products from the resist or entrapment of bubbles from fast scanning at the advancing edge of the fluid/resist interface) may be a source of undeveloped resist regions. Resist component leaching into the fluid, in conjunction with fluid droplet evaporation, may lead to drying stain defects. This is also possible if the fluid is not pure, with evaporation leading to surface residues. Finally, fluid or environmental particles can cause printed defects if present during exposure. What is clear is that optimization in both processing and immersion-related materials are necessary to improve defect performance. These improvements have resulted in defect densities below 0.06/cm^2 for water immersion. Water immersion lithography printing of 36 nm lines and spaces (i.e., 72-nm pitch) is shown in Fig. 2, using a 2-beam interference lithography tool, and shows good imaging performance with some line edge roughness. Current commercial photoresists on production immersion lithography perform even better than this. For high-index immersion, preliminary defect studies have concentrated on both the high-index fluids, cleanliness and fluid misted droplets' interaction with several commercially available dry and water immersion photoresists [12]. Imaging under a second-generation immersion fluid, with an index of refraction at 193 nm of ~1.64 is shown



Immersion Lithography Materials, Fig. 2 Top-down and cross-sectional SEMs of 72 nm pitch lines imaged on a two-beam interferometer under water on two photoresists (After [12])



Immersion Lithography Materials, Fig. 3 Top-down SEMs of 72 nm pitch lines imaged using a second-generation immersion fluid, IF132, with a fluid index of refraction of 1.64, on a two-beam interferometer as a function of increasing exposure dose from *left* to *right* (After [12])

Immersion Lithography Materials, Table 1 Summary of lithographic half-pitch generations with technology options and imaging parameters

Hp (nm)	Technology	Light source	Lens material	Pellicle	Photomask	Fluid	Topcoat	Photoresist
400	**365 dry**	Hg lamp	Glass	Nitrocellulose	Binary intensity mask	None	None	Novolak
300	**365 dry**	Hg lamp	Glass	Nitrocellulose	Binary intensity mask	None	None	Novolak
250	**248 dry**	KrF laser	SiO_2	Fluoropolymer	Binary intensity mask	None	None	Styrenic systems
180	**248 dry**	KrF laser	SiO_2	Fluoropolymer	Binary intensity mask	None	None	Styrenic systems
130	**248 dry**	KrF laser	SiO_2	Fluoropolymer	Attenuated PSM	None	None	Styrenic systems
90	**248 dry**	KrF laser	SiO_2	Fluoropolymer	PSM	None	None	Styrenic systems
90	**193 dry**	ArF laser	SiO_2	Fluoropolymer	PSM	None	None	Methacrylic systems
90	**193 dry**	ArF laser	SiO_2	Fluoropolymer	Attenuated PSM	None	None	Methacrylic systems
90	**157 dry**	F2 laser	CaF_2	Fluoropolymer	Binary intenstiy mask	None	None	Fluoropolymer
65	**157 dry**	F2 laser	CaF_2	Fluoropolymer	PSM	None	None	Fluoropolymer
65	**193 dry**	ArF laser	SiO_2	Fluoropolymer	PSM	None	None	Methacrylic systems
65	**193 H2O SP**	ArF laser	SiO_2	Fluoropolymer	PSM	Water	Fluorinated	Methacrylic systems
55	**193 H2O SP**	ArF laser	SiO_2	Fluoropolymer	PSM	Water	Fluorinated	Methacrylic systems
45	**193 H2O SP**	ArF laser	SiO_2	Fluoropolymer	PSM	Water	Fluorinated	Methacrylic systems
38	**193 H2O SP**	ArF laser	SiO_2	Fluoropolymer	PSM	Water	Fluorinated	Methacrylic systems
38	**193 Gen2 SP**	ArF laser	SiO_2 and LuAG	Fluoropolymer	PSM	Hydrocarbons	Unknown	Methacrylic systems
32	**193 Gen2 SP**	ArF laser	SiO_2 and LuAG	Fluoropolymer	PSM	Hydrocarbons	Unknown	Methacrylic systems
32	**193 dry DP**	ArF laser	SiO_2	Fluoropolymer	PSM	None	None/CEL	Spacer/freezing systems
32	**193 H2O DP**	ArF laser	SiO_2	Fluoropolymer	PSM	Water	Fluorinated	Methacrylic systems
32	**EUV**	EUV LPLS	None – reflective mirrors	None	Binary intensity reflective mask	None	None	Styrenic systems
28	**193 Gen3 SP**	ArF laser	SiO_2 and LuAG	Fluoropolymer	PSM	Unknown	Unknown	HI resists
22	**193 H2O DP**	ArF laser	SiO_2	Fluoropolymer	PSM	Water	Fluorinated	Methacrylic systems
22	**EUV**	EUV LPLS	None – reflective mirrors	None	Binary intensity reflective mask	None	None	Styrenic systems

in Fig. 3 with increasing exposure doses producing photoresist lines from 36 nm down to 29 nm. The dramatic ability of immersion lithography to enable the printing of such sub-resolution features can be seen by considering that 36 nm lines printed with 193 nm light correspond to imaging at λ/5, a resolution limit far beyond the diffraction limit. See Table 1 for a broad summary of lithographic half-pitch generations with technology options and imaging parameters.

Immersion Topcoats and Topcoatless Photoresists

Fears of lens contamination, increased defects, and resist performance degradation with water immersion lithography prompted much research activity on novel topcoats, and the so-called topcoatless photoresists. Topcoats are designed to reduce both water permeation into the photoresist as well as resist component leaching (especially PAG), and also to increase water's receding contact angle at fast scan speeds (to reduce water droplet formation that may lead to imaging defects). More hydrophobic topcoats provide the increased receding contact angle to allow higher scanning velocities before droplet formation occurs. Topcoats showing over 100° receding contact angles will enable speeds up to 900 mm s^{-1}. Ideally, the topcoat would dissolve in the same aqueous base developer (TMAH; tetramethyl ammonium hydroxide) used in the resist development step, but some topcoat versions were designed for nonaqueous solvent development, which adds an extra processing step. Many topcoats incorporate fluorine-containing functional groups for their increased hydrophobicity, as well as the same compound types used effectively for absorbance reduction in 157 nm technology [13]. These fluoropolymers have also been used in water immersion resist materials – either as additives to current 193 nm materials or as block copolymers [14]. These systems are designed so that a topcoat is not necessary, but instead incorporate the topcoat's hydrophobic properties directly into the photoresist, eliminating an extra topcoat spinning step. These topcoatless resists depend on the self-segregation of low surface energy fluoro-components to the air interface during spinning and baking, providing the required hydrophobicity for water immersion imaging. A novel method to make the block copolymers relies on RAFT polymerization to make the first block composed of 193 nm photoresist monomers, followed by a second block composed of a fluorinated monomer. Blends of this block copolymer (even at very small fractions) with a standard non-fluorinated methacrylate terpolymer show high receding contact angles. One variant to this system, which could further reduce the defect levels, is by modifying the fluorinated block to be developer soluble – which would also help with developer wettability.

Future Directions

As shown here, immersion lithography materials research has defined the critical application requirements and the nature and properties of the materials, which have been developed and commercialized into today's production immersion lithography. Materials which were not successfully commercialized have been included to show how successful technological developments have followed "failed" technology research efforts. An in-depth understanding of the intrinsic and extrinsic contributions to both the optical absorbance and the refractive index (or more generally, the complex index of refraction) of inorganic, polymeric, and molecular materials in the UV and VUV spectral regions has been developed. This scientific basis is essential to future successful development and commercialization of optimized materials and technology.

Cross-References

- ▶ Dip-Pen Nanolithography
- ▶ DUV Photolithography and Materials
- ▶ EUV Lithography
- ▶ Nanoimprint Lithography
- ▶ Nanoimprinting
- ▶ Nanoscale Printing
- ▶ Self-Assembly

▶ Self-Assembly of Nanostructures
▶ Stereolithography
▶ SU-8 Photoresist

References

1. French, R.H., Tran, H.V.: Immersion lithography: photomask and wafer-level materials. Annu. Rev. Mater. Res. **39**, 93–126 (2009). Vol. on "materials advances for next-generation microelectronics"
2. Sewell, H., Mulkens, J.: Materials for optical lithography tool application. Annu. Rev. Mater. Res. **39**, 127–153 (2009). Vol. on "materials advances for next-generation microelectronics"
3. Carcia, P.F., French, R.H., Reilly, M.H., Lemon, M.F., Jones, D.J.: Optical superlattices – a strategy for designing phase-shift masks for photolithography at 248 nm and 193 nm: application to AlN/CrN. Appl. Phys. Lett. **70**, 2371–2373 (1997)
4. Lai, F.: Optimized HT-AttPSM blanks using Al_2O_3/TiO_2 multilayer films for the 65 nm technology node. J. Vac. Sci. Technol. B **22**, 3097–3101 (2004)
5. Yang, M.K., French, R.H., Tokarsky, E.W.: Optical properties of Teflon® AF amorphous fluoropolymers. J. Micro/Nanolith. MEMS MOEMS **7**, 033010-1–033010-9 (2008)
6. Kunz, R.R., Switkes, M., Sinta, R., Curtin, J.E., French, R.H., Wheland, R.C., Kao, C.C., Mawn, M. P., Lin, L., Wetmore, P., Krukonis, V., Willams, K.: Transparent fluids for 157 nm immersion lithography. J. Microlith. Microfab. Microsyst. **3**(1), 73–83 (2004)
7. Urbach, F.: The long-wavelength edge of photographic sensitivity and of the electronic absorption of solids. Phys. Rev. **92**, 1324 (1953)
8. French, R.H., Sewell, H., Yang, M.K., Peng, S., McCafferty, D., Qiu, W., Wheland, R.C., Lemon, M. F., Markoya, L., Crawford, M.K.: Imaging of 32-nm 1:1 lines and spaces using 193-nm immersion interference lithography with second-generation immersion fluids to achieve a numerical aperture of 1.5 and a k_1 of 0.25. J. Microlith. Microfab. Microsyst. **4**, 031103 (2005)
9. Tran, H.V., French, R.H., Adelman, D.J., Feldman, J., Qiu, W., Wheland, R.C., Brubaker, L.W., Fischel, B. E., Fones, B.B., Lemon, M.F., Yang, M.K., Nagao, O., Kaku, M., Mocella, M., Schmieg, J.J.: Evaluation of next generation fluids for ArF immersion lithography beyond water. J. Photopolym. Sci. Technol. **20**(5), 729–738 (2007)
10. Liberman, V.L., Rothschild, M., Palmacci, S.T., Grenville, A.: Impact of photoacid generator leaching on optics photocontamination in 193-nm immersion lithography. J. Microlith. Microfab. Microsyst. **6**, 013001-1–013001-7 (2007)
11. Gaugiran, S., Feilleux, R., Sourd, C., Warrick, S., Farys, V., Cruau, D., Benndorf, M., Mortini, B.: Leaching mechanisms in immersion lithography with or without top coat. Microelectron. Eng. **84**, 1054–1057 (2007)
12. Tran, H.V., Hendrickx, E., Van Roey, F., Vandenberghe, G., French, R.H.: Fluid – photoresist interactions and imaging in high index immersion lithography. J. Micro/Nanolith. MEMS MOEMS **8**(3), 033006 (2009)
13. Feiring, A.E., Crawford, M.K., Farnham, W.B., Feldman, J., French, R.H., Junk, C.P., Leffew, K.W., Petrov, V.A., Qiu, W., Schadt III, F.L., Tran, H.V., Zumsteg, F.C.: New amorphous fluoropolymers of tetrafluoroethylene with fluorinated and non-fluorinated tricyclononenes. Semiconductor photoresists for imaging at 157 and 193 nm. Macromolecules **39**, 3252–3261 (2006)
14. Allen, R.D., Brock, P.J., Sundberg, L., Larson, C.E., Wallraff, G.M., Hinsberg, W.D., Meute, J., Shimokawa, T., Chiba, T., Slezak, M.: Design of protective topcoats for immersion lithography. J. Photopolym. Sci. Technol. **18**(5), 615–619 (2005)

Impedance Cytometry

▶ Electrical Impedance Cytometry

Impedance Measurement of Single Cells

▶ Single-Cell Impedance Spectroscopy

Impedance Sensing Techniques for Biomedical Microdevices

Vaishnavi Srinivasaraghavan and Masoud Agah
Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA, USA

Synonyms

Bioimpedance sensing; Electric cell-substrate impedance sensing (ECIS)

Definition

The use of electrical impedance as a means of detecting biological events or samples of interest

in microdevices for biomedical applications. Examples include electric cell-substrate impedance sensing (ECIS), electrochemical sensing of analytes in biofluids, and cell impedance cytometry.

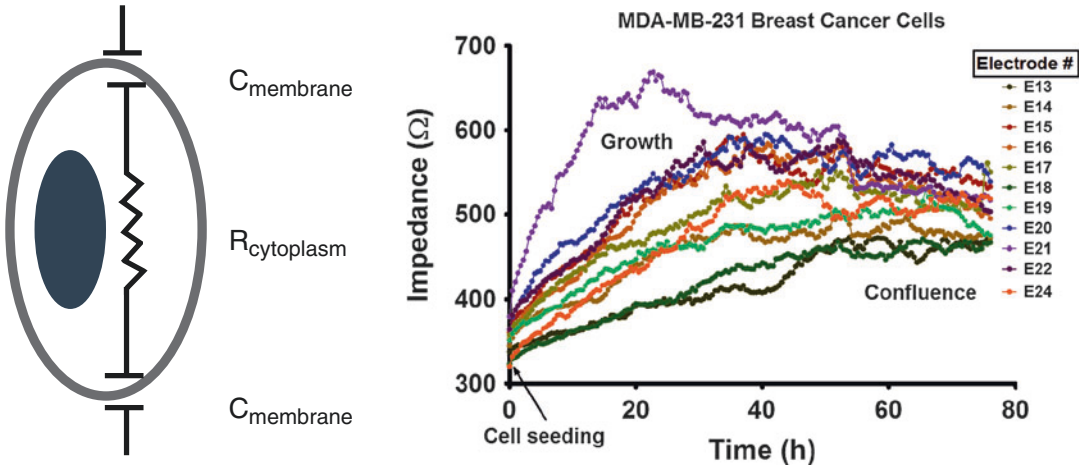
Introduction

The advent of microfabrication tools and techniques coupled with the increasing need for low-cost, rapid sensing platforms for medical diagnostics has paved the way for a new field that we have come to know as biomedical microdevices. Research in this field has made significant headway aided by collaborations between engineers, scientists from the life sciences, and physicians who will be the end users of this technology. There are many features that make this technology promising for widespread use in the future. These devices are inexpensive to manufacture and hence can be disposed after single use which is standard in hospitals and clinics to avoid cross-contamination. Low cost also ensures low-income people as well as patients in developing countries can have ready access to this technology. Also, the sample and reagent volumes are extremely small compared to conventional tests performed in the laboratory due to the small size of the device. This not only reduces costs but also makes the reaction time much lower enabling rapid analyses. In addition, high-throughput devices which can screen samples in parallel can be manufactured using this technology that will further lower analysis times or increase sample capacity. The technology has been used to create faster and more efficient miniaturized devices for multiple existing assays in the lab. But the real success of this technology is seen in the numerous research reports on new techniques, assays, and platforms for studying and detecting a wide variety of biological samples. This chapter specifically focusses on using impedance sensing within these devices as a means to capture biological events or analytes in the sample.

Impedance refers to the resistance offered to the flow of electric current. Mathematically, it is a complex number where the real part is the electrical resistance (in ohms) and the imaginary part is the reactance (also in ohms) which is a function of the frequency (in hertz) at which the measurement is made. It can also be expressed in polar format as a magnitude and a phase angle. Impedance is typically measured using an electrode pair made of a highly conductive material such as gold or platinum. These are the metals of choice because they are inert and do not adversely affect biological cells. The impedance of biological samples is a function of their size, composition, and proximity to the electrodes. Hence, it is common to treat the surface of the gold electrodes to capture the sample or analytes of interest thereby increasing its sensitivity. In order to measure the impedance, a constant voltage is supplied by an impedance analyzer or spectroscope, and the corresponding current is measured. Conversely, it is also possible to supply a constant current and measure the voltage drop across the electrode pair. These instruments also have the capability to generate these signals over a range of frequencies enabling what is called impedance spectroscopy or the analysis of impedance over a spectrum of frequencies. The following sections discuss some of the ways impedance sensing can be successfully employed as a detection strategy in microdevices for biomedical applications.

Electric Cell-Substrate Impedance Sensing

Electrical cell-substrate impedance sensing or ECIS is a technique that is used to monitor live biological cells in tissue culture. The term was coined by the inventors of the technique, Ivar Giaever and Charles R. Keese, and is currently the trademark of Applied Biophysics [1]. The electrical impedance signals are recorded from cells that are adherent on the surface of planar microelectrodes. These electrodes can be on the surface of regular treated polystyrene petri dishes or well plates used for tissue culture or on any other surface suitable for cell culture like glass or silicon with an appropriate chamber to hold media during measurements. The technique can be used to make measurements from any animal cell line



Impedance Sensing Techniques for Biomedical Microdevices, Fig. 1 The equivalent circuit model for a single cell (*left*) and the changes in electrical impedance magnitude as live breast cancer cells attach and grow on microelectrodes (*right*)

that grows as a monolayer by attachment. ECIS is based on the ability of cells to block the flow of electric current when attached to a conductive surface. The dominant current paths through the cell monolayer vary in different ranges of the frequency spectrum. The cell membrane is made of a phospholipid double layer which acts like a capacitor, and the cytoplasm of the cell is mainly made of a conductive gel-like substance which can be modeled as a resistor as shown in Fig. 1. At lower frequencies, the current passes through the cell-substrate gap and the space between adjacent cells called the cell-cell junction region. However, at higher frequencies the current can couple through the cell and thus probe the electrical properties of the cell membrane and the interior. The ECIS impedance measurements can be made continuously at a single frequency or multiple frequencies at regular time intervals.

Measurements are always made from a pair of electrodes, and they come in different configurations. For instance, the electrode pair can have one very large counter electrode which is ten times or more larger in area than one very small sensing electrode. A number of different shapes including circular, square, rectangular, branch, and other shapes have been used for the sensing and counter electrode design in ECIS devices. Another commonly used configuration is the interdigitated electrode pair in which two identical comb-shaped electrodes are interlocked with each other as shown. In this configuration, both electrodes play the role of sensing and counter electrodes. There are other commercial systems on the market that employ these electrodes in making impedance measurements from cell monolayers.

The impedance values recorded are directly related to the cell coverage on the electrodes. This makes it suitable for measuring the growth and proliferation of cells in culture. Using the impedance values recorded, it is possible to infer the rate of growth, the doubling time, and the time when the culture reaches confluence. So, one of the applications of ECIS is the remote monitoring of cell growth. To ensure good cell attachment, the electrodes are typically treated with oxygen plasma in a plasma cleaner followed by coating with extracellular matrix proteins like collagen or fibronectin just before the addition of cells. Alternately, cell culture medium can be added to the electrodes and the serum proteins in the medium self-adsorb onto them providing the same effect. However, for long-term monitoring, it is desirable to pre-coat the electrodes before adding cells to them. Figure 1 shows the characteristic change in the impedance profile as cells attach and grow on microelectrodes. The baseline impedance is the impedance when there are no cells on the

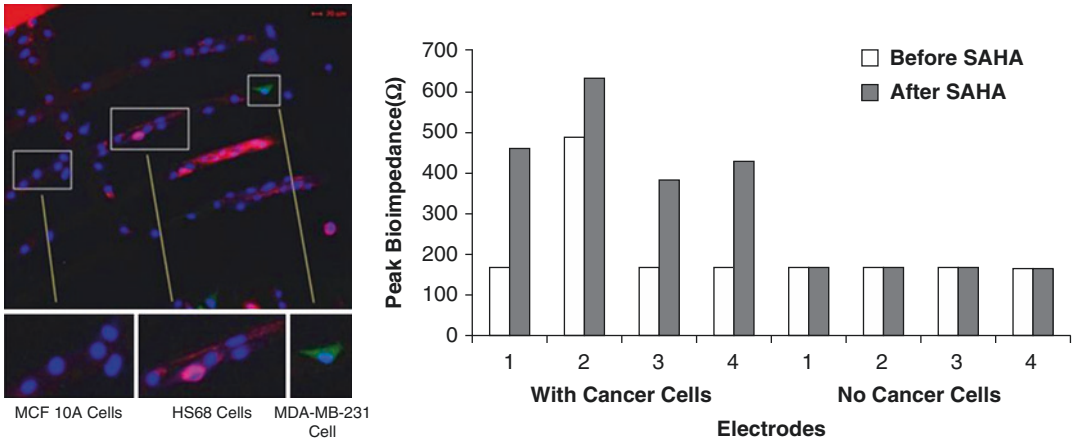
electrode. As they grow, the impedance increases as the number of cells on the electrode increases. When the area on the electrodes is completely filled with cells, the impedance saturates to a maximum value indicating they have reached confluence.

On the same note, if cells begin to die, they become rounded, and their attachment to the electrodes decreases till they completely detach from them. When this happens, it results in a drop in the impedance values which in essence is a reversal in the rise seen due to cell attachment. This is the basis for the use of ECIS to monitor events like apoptosis or programmed cell death. This mechanism of detection can also be used to measure the cytotoxicity of drugs or compounds in culture. It is particularly useful in screening for the optimal drug concentration or dose that should be used in treatments.

The technique also lends itself to a host of other applications since many cellular events are attachment based. For example, cell motility or the movement of cells in culture can be measured using ECIS which lets us compare closely related cell lines for activity in culture. Wound healing assays have also been developed in which the cells on top of the electrode are removed by electroporation and subsequent cell migration to the wounded site is measured. A short pulse of a higher voltage or current spanning a few minutes is applied which kills the cells located on the electrode causing them to detach accompanied by a fall in the impedance value. When the surrounding cells migrate onto this newly available free area on the electrode, the impedance measurements pick it up by reflecting a rise in the values recorded. Alternately, cell adhesion on the electrode area is actively prevented by applying a pulsating electric field. When the area around the electrode has a confluent cell culture, the electric field is turned off, and the impedance is monitored as cells migrate into this region. Electroporation when used on smaller timescales can result in the enlargement of pores in the cell membrane causing cells to take up molecules, particles, and compounds from the surrounding medium that normally would not make their way into the cell interior. Electroporation assays have been developed in conjunction with ECIS that measure the optimal electroporation conditions that allow for the uptake of larger molecules of drugs or nanoparticles with fluorophores for imaging. Other applications such as monitoring the binding of certain molecules to the receptors on the cell surface resulting in specific cellular events can also be done through real-time ECIS measurements.

In the last few decades, ECIS has become a useful tool to study the electrical properties of cancer cells. Although it is believed that there are biological, chemical, or physical differences between normal and cancer cells, the depth of our knowledge with respect to the similarities and differences in their electrical properties was quite limited. It has become increasingly clear, as more reports on the electrical properties of various types of cancers are published, that the impedance of cancer and normal cells are indeed different. Recently, an assay for cancer cell detection in a simulated biopsy sample was reported which employs the use of an anticancer agent called suberoylanilide hydroxamic acid (SAHA) that specifically targets the cancer cell cytoskeleton and causes an increase in the cell area on the electrode which can be captured through ECIS measurements. Figure 2 shows an electrode with normal fibroblast cells found in the extracellular matrix of tissues and normal and cancerous breast cells simultaneously cultured on it. It also displays the impedance changes recorded on electrodes with and without cancer cells demonstrating detection using this mechanism. Cancer cell invasion through a confluent layer of endothelial cells has also been monitored using real-time ECIS measurements. As mentioned before, the technique can also be used to screen the efficacy and optimal dose of anticancer agents on different types of cancer cells. These tests, in the future, could be carried out on patient samples to provide individualized reports of drug effectiveness which can then be used to make treatment plans for the patient.

Thus, ECIS is a label-free, noninvasive method that can be used to monitor live cell cultures in real-time. A number of cellular events based on cell attachment can be detected using



Impedance Sensing Techniques for Biomedical Microdevices, Fig. 2 Image of an electrode (*left*) with fluorescently stained normal fibroblast cells (*orange*) found in the extracellular matrix of tissues and normal (*red*) and cancerous (*green*) breast cells. Impedance changes (*right*) recorded on electrodes with and without cancer cells under the influence of an anticancer agent (suberoylanilide hydroxamic acid) (Srinivasaraghavan V, Strobl J, and Agah M (2012) Bioimpedance rise in response to histone deacetylase inhibitor is a marker of mammary cancer cells within a mixed culture of normal breast cells. Lab on a Chip 12.24: 5168–5179 – (Reproduced by permission of The Royal Society of Chemistry http://dx.doi.org/10.1039/C2LC40778G))

this technique. It has become an easy and useful approach to study the electrical properties of cancer cells and also lends itself for integration into existing devices and technologies which need a detection mechanism for cell attachment-based events.

Impedance Cytometry

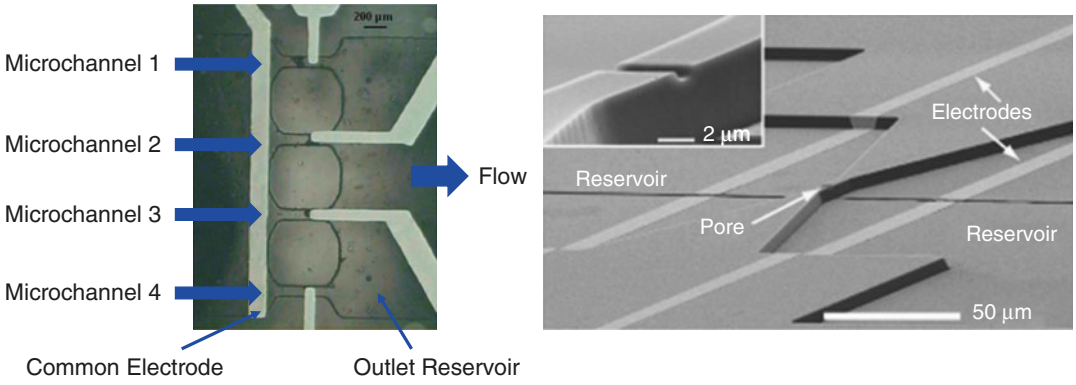
The general meaning of cytometry is making measurements from cells. Impedance measurements from flowing or suspended single cells can be used as a sensing mechanism to count and measure relative size and dielectric properties of cells [2]. Unlike ECIS, this method does not require the cells to be attached to the electrodes. Another difference is the analysis of single cells as opposed to cell populations as seen before. Measurements from single cells greatly increase the speed of measurements and also provide insight into the heterogeneity of the cell population. A number of devices that we will discuss in this part of the chapter fall under the category of microfluidics. Microfluidic channels in devices are usually fabricated using a process called soft lithography. In this process, the channel is first fabricated on a silicon wafer using standard lithography techniques with a thick photoresist like SU-8 to obtain the replica of the patterns of the microfluidic channel network. Since the channels in the final device need to be hollow to allow for liquid flow, the silicon replica, also referred to the master or the mold, has the channels that stand out from the surface of the wafer and look like they have been embossed. A moldable polymer such as PDMS can then be poured on the silicon master, cured, and peeled off to obtain the microfluidic device in PDMS. Of course, this device needs to be bonded to a glass slide or coverslip so the channels can be sealed with the openings to the channels punched out. A strong, irreversible bond between glass and PDMS is obtained by treating the surfaces with oxygen plasma.

There exist different types of electrodes that can be incorporated in the design of these devices. The first is the Ag/AgCl electrodes that have been used in electrochemical cells for many years now. They can either be purchased as wires commercially and just immersed into the reservoirs to apply a voltage or fabricated in the lab by evaporating a layer of Ag on a glass substrate and electrochemically growing a layer of AgCl.

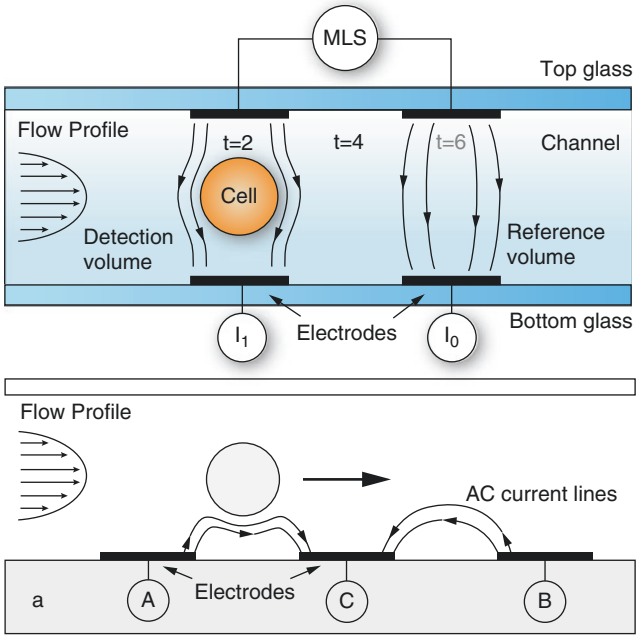
Another commonly used technique is to integrate metal electrodes usually made of gold at the bottom of the channel by evaporation. This is the preferred method because it easily integrates into the fabrication process flow. Another interesting alternative is the use of liquid metal electrodes integrated into microdevices for impedance measurements.

One of the first examples of a cytometer is the Coulter counter in which two chambers are connected by a small orifice with one large electrode in each chamber. The chambers are filled with a conductive liquid with particles or cells. As these particles move between the chambers through the orifice, the impedance measured increases. Also, a larger particle would block the current more when moving through the orifice resulting in a higher impedance change. Conversely, a smaller particle will not occupy as much space in the orifice and therefore block lesser current causing a lower change in impedance. A smaller orifice will have more sensitivity and can detect smaller particles as well. One of the limiting factors of the traditional Coulter counter is that the measurements are made only with a DC supply. The first microscale Coulter counters designed work on the same principle, but on a smaller scale, and were fabricated on silicon, glass, and PDMS. These devices had a sensing pore region instead of an orifice for detection. As these devices have evolved, although they use the same principle of detection, they no longer adhere strictly to the design of the Coulter counter. A number of different designs have been used including higher-throughput devices with multiple channels each with its own set of test electrodes to increase the number of particles analyzed and a channel with a flow-based programmable pore size that enables testing samples of different sizes on the same devices and tuning the sensitivity based on application. Figure 3 shows some examples of microfluidic devices based on the concept of the Coulter counter used for bioparticle analyses. Using the microscale Coulter counter and its design variants, a number of different samples of interest have been analyzed. Examples include mammalian cells, red and white blood cells, and nanoparticles that capture specific antibodies even single-molecule analyses of DNA, RNA, and others.

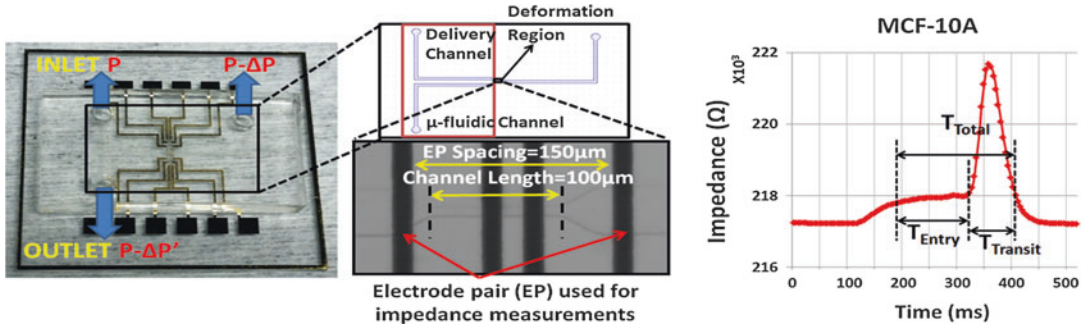
When the above concept is extended to include AC impedance measurements at multiple frequencies, the devices can be used to probe the dielectric properties of the cells. Figure 4 shows an example of such a device where the cells are flowing through a microfluidic channel that is larger than their size and gold electrodes integrated into the channel measure the changes as single cells pass them. The electrodes can be positioned laterally on the same surface or on opposite surfaces in the microchannel. In both cases, an impedance change is recorded due to the change in the current path as shown in the schematic in Fig. 4. It also needs to be reiterated that the sensitivity of the impedance measurements is related to the frequency at which they are made. At lower frequencies, the current path bends around the cell causing the measurements to be sensitive to size of the cell. However, at higher frequencies, the currents pass through the cells and are thus sensitive to cell membrane and cytoplasm properties. The technique has been proven to be sensitive enough to differentiate between cells of different sizes like red and white blood cells and closely matched cells like healthy and diseased red blood cells as wells as normal and cancer cell lines of the same tissue. Although the flow in microfluidic channels is planar and it is possible to measure the impedance of single cells without any additional design, active focusing and trapping mechanisms are often incorporated into the design to decrease any clogging issues and increase the rate and confidence of single-cell measurements. For instance, two auxiliary channels are present on either side of the input channel and flow to focus the cells coming from the input channel as a row of single cells. Another strategy is to apply a negative pressure through a tiny pore located between the electrodes to ensure that a single cell is positioned right in the middle of the electrodes. This also ensures that there are no discrepancies in the measurements due to the cell position being partial in one direction toward one of the electrodes. Dielectrophoretic trapping has also been used to position single cells between the



Impedance Sensing Techniques for Biomedical Microdevices, Fig. 3 Examples of fabricated microscale Coulter counters for bioparticle analyses (*Left* – Zhe J, Jagtiani A, Dutta P, Hu J, Carletta J (2007) A micromachined high throughput Coulter counter for bioparticle detection and counting. J Micromech Microeng 17:303–313 – Reproduced by permission of IOP Publishing. All rights reserved. http://dx.doi.org/10.1088/0960-1317/17/2/017; *Right* – Reproduced with permission from Saleh OA, Sohn LL (2001) Quantitative sensing of nanoscale colloids using a microchip Coulter counter. Rev Sci Instrum 72:4449–4451. Copyright 2001, AIP Publishing LLC. http://dx.doi.org/10.1063/1.1419224)



Impedance Sensing Techniques for Biomedical Microdevices, Fig. 4 Flow cytometry to probe to the dielectric properties of single cells flowing in a microfluidic channel using impedance spectroscopy (*Left* – Reproduced with permission from Gawad S, Sun T, Green NG, Morgan H (2007) Impedance spectroscopy using maximum length sequences: application to single cell analysis. Rev Sci Instrum 78:054301. Copyright 2007, AIP Publishing LLC. http://dx.doi.org/10.1063/1.2737751; *Right* – Reproduced from Gawad S, Schild L, Renaud Ph (2001) Micromachined impedance spectroscopy flow cytometer for cell analysis and particle sizing. Lab Chip 1:76–82 with permission of The Royal Society of Chemistry. http://dx.doi.org/10.1039/B103933B)



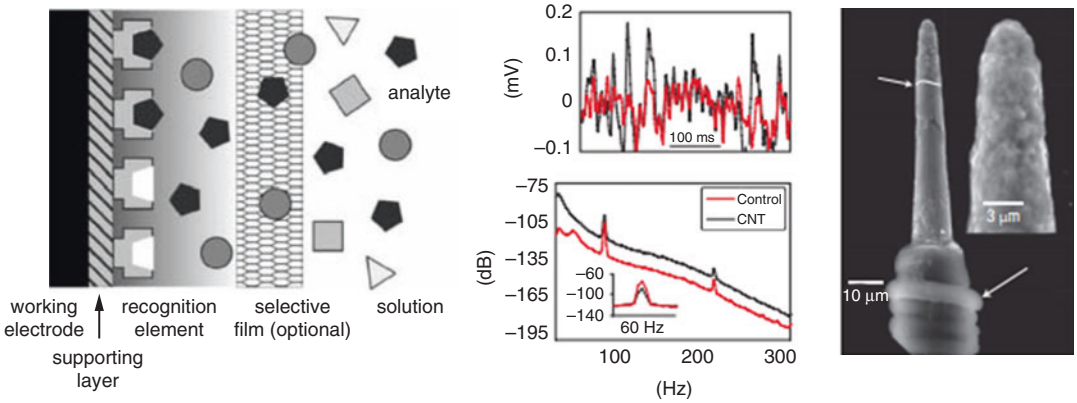
Impedance Sensing Techniques for Biomedical Microdevices, Fig. 5 Electrodes integrated into the microchannel (*left*) enable simultaneous measurement of electrical and mechanical properties from single cells (*right*) (© 2014 IEEE. Reprinted, with permission, from Srinivasaraghavan V, Aggarwal D, Babahosseini H, Nakidde D, Strobl J and Agah M, "Analyses of single-cell mechanoelectrical properties via microfluidics," IEEE Sensors 2014, November 2014, Valencia, Spain, pp. 1772–1775. http://dx.doi.org/10.1109/ICSENS.2014.6985368)

measurement electrodes by incorporating more electrodes through which the electric fields required for dielectrophoresis can be supplied. Figure 5 shows a device that has electrodes integrated into the narrow microfluidic channel that is smaller than the size of the cell, in which impedance measurements can be used to simultaneously measure the mechanical and the electrical properties of single cells. A constant negative pressure is applied at one end of the channel which forces the cells to deform and travel through it. The width of the impedance profile gives the time that single cells take to traverse through the channel which in turn is a measure of the mechanical properties of these cells. Additionally, the impedance measurements made in such a channel are more sensitive to the internal properties of the cell, especially at higher frequencies, because the leakage current around the cell is minimized.

Hence, impedance cytometry can be integrated into the microfluidic platform to generate a wide array of microdevices that have applications over a wide spectrum ranging from particle counting, size determination, and single-cell and molecule analyses.

Electrochemical Sensors

The working principle of electrochemical sensors is the same as what we have seen so far. In essence, when the analyte of interest is in the sensing region of the device, the impedance measurements change, and it is detected. However, we discuss this here separately to provide a summary of the strategies that one can use to add a sensing layer on top of the electrodes to effectively sensitize them to specific analytes of interest [3]. Figure 6 shows an illustration summarizing the idea. Self-assembled monolayers are used, for instance, to immobilize the biorecognition element on the electrode surface. Thiol end groups spontaneously chemisorb and form well-organized monolayers on gold electrodes. Thus, the recognition compound can be functionalized chemically to self-assemble on the surface. Polymer and polyelectrolyte films capable of binding to biomolecules have also been used successfully to capture target analytes. The field of biosensors for immunosensing has grown largely using this detection mechanism. The selective binding between antibody and antigen is exploited, and either the antibody or the antigen that complements the antigen or antibody in solution is immobilized on the surface of the electrode. When the complementary molecule binds to the antibody or antigen, the impedance changes, and detection is achieved. In some sensors, selected enzymes are deposited on the surface of electrodes by combining them with other materials or compounds. The catalytic action of the enzyme on the biomolecule of interest results



Impedance Sensing Techniques for Biomedical Microdevices, Fig. 6 Strategies to achieve specificity (*left*) and improving sensitivity (*right*) in impedance sensing (*Left* – with kind permission from Springer Science +Business Media: Analytical and Bioanalytical Chemistry, The use of electrochemical impedance spectroscopy for biosensing, 391.5, 2008, 1555–1567, Lisdat, F, and Schäfer D, Fig. 5a. http://dx.doi.org/10.1007/s00216-008-1970-7; *Right* – Reprinted by permission from Macmillan Publishers Ltd: Nature Nanotechnology Keefer, Edward W., Barry R. Botterman, Mario I. Romero, Andrew F. Rossi, and Guenter W. Gross. "Carbon nanotube coating improves neuronal recordings." Nature nanotechnology 3, no. 7 (2008): 434–439, copyright 2008. http://dx.doi.org/10.1038/nnano.2008.174)

in detection through a change in the measurement from the electrode. It is also possible to analyze DNA concentration or base-pair mismatches using electrochemical sensors by comparable strategies. Similarly, in the detection of cancer cells which overexpress certain receptors, the ligand that binds to these receptors is immobilized on the electrodes in an effort to sensitize these electrodes to cancer cells. Electrochemical sensors have also been developed for the detection of glucose and urea levels in samples obtained from patients.

To increase the overall sensitivity of electrodes to impedance measurements, a common strategy is to increase the surface area of the electrode by depositing or growing carbon nanotubes (CNTs), coating the surface with conductive nanoparticles or composites of them. These coatings have shown to be effective in improving the sensitivity of measurements from neurons, electrochemical sensors for analysis of human body fluids as well as cancer cells. Figure 6 shows an example of an electrode that is coated with CNTs and the improvement in performance compared to uncoated control electrodes from neuronal recordings.

Conclusion

This chapter gives an overview of the impedance sensing techniques than can be incorporated into biosensors for medical applications. This technique can be used to analyze a wide variety of bioanalytes like single biomolecules, DNA or RNA strands, compounds like glucose and urea, and single mammalian cells and cell populations. Immobilization of biorecognition molecules on the surface of the electrode can be used to further sensitize them to target analytes. Impedance sensing is a noninvasive, label-free, highly sensitive, inexpensive, and rapid technique that can be used as a detection strategy in hand held biomedical microdevices for point-of-care diagnostics. It can be used to automate many measurements made conventionally with bulky and expensive microscopes and can be used by existing medical professionals and staff with minimal training. It can also be easily integrated within the existing framework of smartphones and computers which enables faster transfer of information obtained, for faster results and prompt care for patients. This new field which is at the

intersection of engineering, chemistry, medicine, and biology holds tremendous promise and potential to solve the challenges and problems in healthcare in the twenty-first century.

Cross-References

- ▶ Biosensing
- ▶ Cell Adhesion
- ▶ Impedance Cytometry
- ▶ Impedance Spectroscopy for Single-Cell Analysis
- ▶ Microelectrodes

References

1. Giaever, I., Keese, C.R.: Monitoring fibroblast behavior in tissue-culture with an applied electric-field. Proc. Natl. Acad. Sci. U. S. A. **81**, 3761–3764 (1984)
2. Sun, T., Morgan, H.: Single-cell microfluidic impedance cytometry: a review. Microfluid. Nanofluid. **8**, 423–443 (2010)
3. Lisdat, F., Schafer, D.: The use of electrochemical impedance spectroscopy for biosensing. Anal. Bioanal. Chem. **391**, 1555–1567 (2008)

Impedance Spectroscopy for Single-Cell Analysis

- ▶ Single-Cell Impedance Spectroscopy

Implementing Alternating Nanolaminates in Trenched Energy Storage Systems

Sylvia W. Thomas, Jing Wang and Paula A. Algarin
Department of Electrical Engineering, University of South Florida, Tampa, FL, USA

Synonyms

Laminates: multilayered structures; Nanocomposite materials; Nanocomposites; Nanolayered coatings; Nanolayers

Overview

The National Academy of Engineering, in its Grand Challenge on "making solar energy economical," places emphasis on this statement, "However advanced solar cells become at generating electricity cheaply and efficiently, a major barrier to widespread use of the sun's energy remains: the need for storage" [1].

For energy storage devices, such as the battery and the supercapacitor, the most common problems are shelf/cycle life, lower efficiency, and high leakage current [2, 3]. These are tremendous challenges for an integrated circuit (IC) industry seeking greater IC functionality, smaller component size, and low power consumption using standard supercapacitors, ultracapacitors, or electrochemical capacitors for on-chip energy storage. In addition, the continuous need to reduce the component size of IC elements has not only begun to challenge the capabilities of traditional photolithography, but has prompted intense research to incorporate nanometer scale structures into current semiconductor technologies.

Nanostructures have unique and attractive characteristics, consisting of spatial arrangement, high surface area and density, and versatile shapes. Nanostructures have several forms: nanowires, nanochannels, nanotubes, nanobeams, nanopores, nanofilms, nanofibers, and nanolaminates. The atomic sizes for these structures vary from 0.1 to 100 nm or 5 to 1,000 atoms. In the past 10 years, nanostructures, particularly nanolaminates, have received significant attention for employability in supercapacitors/energy storage systems. Limitations of these energy storage systems have been plagued by nonconforming layers, pin holes, high leakage currents, low dielectric constants, and cost; all of which are directly related to the selection and quality processing of the materials and the nanostructure architecture. In an effort to address these limitations, researchers have employed alternating nanolaminate layers of high-k dielectric and wide bandgap materials (e.g., hafnium oxide (HfO_2), aluminum oxide (Al_2O_3), titanium oxide (TiO_2)) in energy storage systems to achieve higher capacitance density, higher breakdown

voltage, and lower leakage current. The high payoff for energy storage is improved equivalent planar capacitance (EPC) performance for existing supercapacitors/energy storage systems.

Equivalent planar capacitance (EPC) is the normalized planar footprint area of a capacitor (capacitance per planar area) and allows comparison between different capacitor structures. Capacitor structures being planar, embedded, or even trenched.

Theory

A nanolaminate, or nanolayered coating, consists of two or more alternating layers of dielectric films consisting of ultrathin oxide multilayers having individual layer thickness ranging from 0.1 to 100 nm. Nanolaminates have also been referred to as nanocomposites of materials with distinct properties, e.g., brittle/soft or high-k/wide bandgap materials. Particularly, nanolaminates of high-k dielectric materials have been investigated for energy storage applications [4]. Dielectric materials exhibiting high dielectric constants have recently gained considerable attention for their potential applications in microelectronics, such as capacitors and memory devices [5–8].

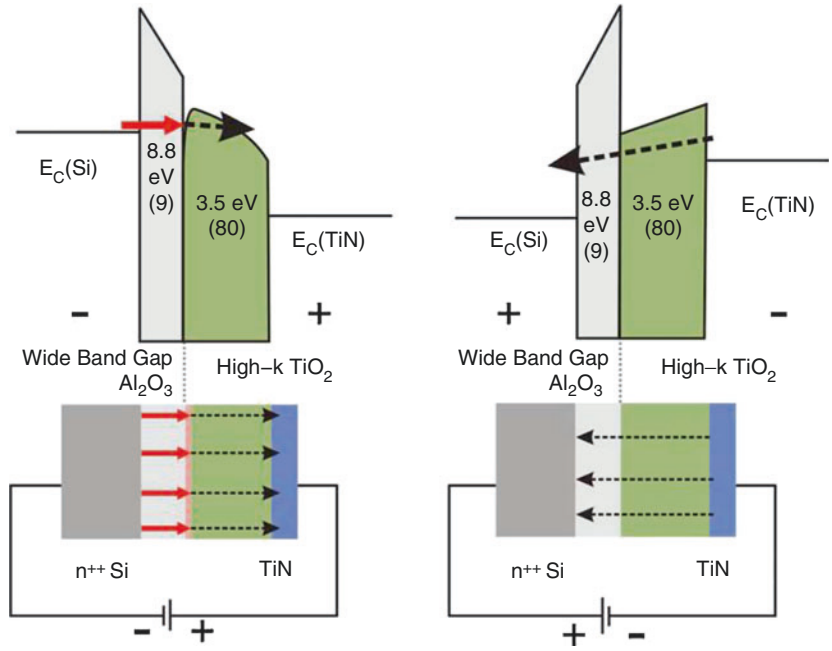
Currently studied high-k materials involve oxides of both transition or lanthanide metals and their alloys, including representative transition metal oxides such as TiO_2, Al_2O_3 [9, 10], and HfO_2 and rare-earth oxides, such as La_2O_3 and Y_2O_3. For applications of solid state supercapacitors in microchip-embedded devices, the higher the dielectric constant, the larger the energy can be stored. Excellent candidates as high dielectric materials include single oxides (e.g., HfO_2, ZrO_2, Ta_2O_5, TiO_2), binary oxides (e.g., $Ti_xAl_{1-x}O_y$, $Hf_xAl_{1-x}O_y$), and perovskite oxides (e.g., $BaTiO_3$, $SrTiO_3$).

By combining and complementing the desirable properties from several materials, ALD-enabled nanolaminates of alternating high dielectric constant (k) and wide bandgap (Eg) materials, theoretically, can achieve low leakage, higher breakdown, and a very high dielectric constant based on the Maxwell–Wagner (M-W) interfacial dielectric effect [10, 11]. When we adapt the analogy of the M-W approach, we can apply electron injection and charge accumulation, as in a depletion region, in a nanolaminate stack of wide bandgap (Al_2O_3) and high-k (TiO_2) material. A negative bias is applied to a Si substrate, and there is enhanced electron tunneling through the Al_2O_3/TiO_2 nanolaminate dielectric stack as indicated by the solid red line in Fig. 1. The electrons reach the interface at a lower E_f and cause higher leakage and low breakdown. When we apply the negative bias to the TiN, there is less direct electron tunneling, because the barrier height of the Al_2O_3 at 8.8 eV is 5.3 eV higher than the TiO_2 (3.5 eV); therefore, it is probable the leakage will be lower and the breakdown higher.

In 2007, Shi at al. proposed a configuration in which Al_2O_3 and TiO_2 have been investigated as high-k materials to replace SiO_2 as a gate oxide and a high-capacitance oxide-based capacitor for electronics [9]. High dielectric constants, in the range of 100–1,000, were achieved with planar Al_2O_3/TiO_2 nanolaminate layers [10, 11], as those featured in Fig. 2. Argonne National Labs investigated these Al_2O_3/TiO_2 nanolaminates, and results for dielectric constant measurements approached a k of 1,000 [12]. In an effort to reduce leakage and expand the range of operation, an interfacial Al_2O_3 layer was deposited prior to the top electrode. The Al_2O_3/TiO_2 nanolaminates on Si were compromised by charge injection. It should be noted that interfacial behavior dominates the ability of the multiple material lamina to reach peak performance.

Nanolaminate interfaces can be between materials of distinct properties, and in designing the laminates, the selection of materials should consider these properties, such as modulus, permittivity, conductivity, bandgap, and dielectric constant. Particularly, interfaces with structural defects, e.g., voids, cracks, and delamination, can lead to very leaky/lossy devices. Even though these defects may not propagate throughout the nanolaminate stack, but be localized, there could still be impacts on device performance.

In addition to alternating nanolaminates in a planar configuration, an alternative way to achieve high-capacitance capacitors is to increase

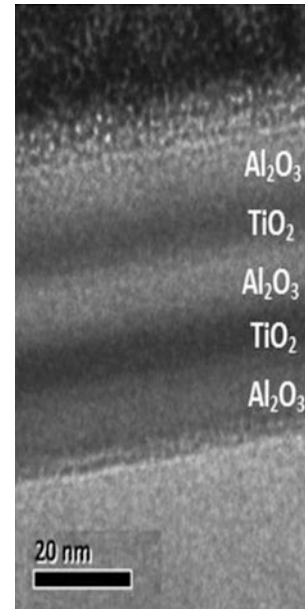


Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 1 Energy band diagram of nanolaminate stack of Al_2O_3/TiO_2 with the Maxwell–Wagner effect (Reproduced with permission from AIP Publishing LLC [10])

the effective surface area by fabricating 3D structures, such as porous structures with high aspect ratios, or trenches. The formation of high-aspect-ratio nanostructures can be attained through the use of non-lithographically self-organized nanotemplates, such as porous anodic aluminum oxides (AAO) [13]. Nanostructured electrodes for electrochemical double layer capacitors were fairly effective in enhancing capacitance, which was attributed to an increase in the surface areas [14].

Energy storage technologies span from conventional capacitors to batteries, inclusive of a region of supercapacitors. Supercapacitor technology has emerged with the potential to enable major advances in energy storage by utilizing higher-surface-area electrode materials and thinner dielectrics to achieve capacitances several orders of magnitude larger than conventional capacitors [2].

In principle, supercapacitors have higher power density than batteries and a charging time



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 2 Cross-sectional TEM of alternating Al_2O_3 and TiO_2 nanolaminates

that is ten times shorter than a battery [1, 2]. However, supercapacitors are not as efficient as batteries in energy storage, as they tend to draw high leakage current. Despite this drawback, supercapacitors can be fabricated to make their energy densities more comparable to batteries; high energy density can be achieved when the ratio of the effective surface area to the electrode volume is high.

A combination of two processing techniques can be used to enhance surface-to-volume ratios, atomic layer deposition (ALD) and anodic aluminum oxide (AAO). Arrays of nanocapacitors using ultrathin oxide layers or nanolaminates can create a high-power-density energy storage device. It is worth noting that this combination of self-assembly, ordering, uniformity, and conformality on a nanoscale level is extremely advantageous for future ICs.

Key Processes

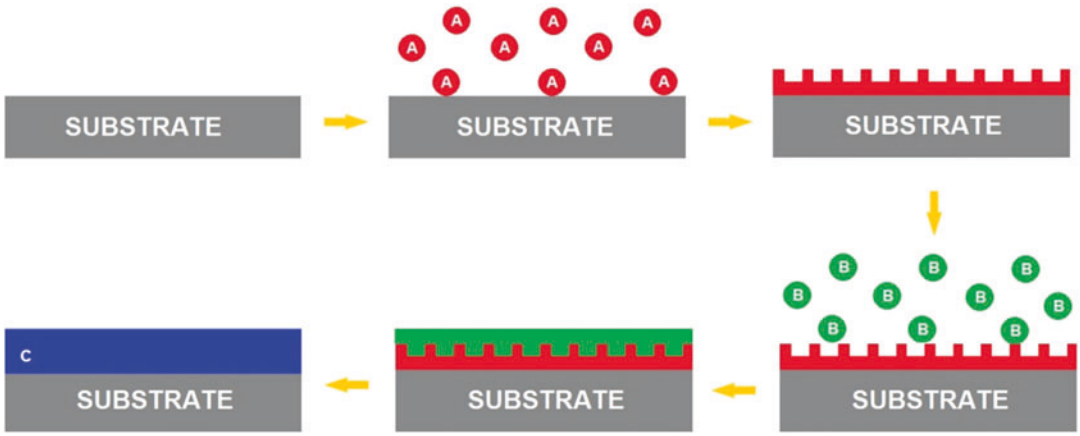
Atomic Layer Deposition (ALD). ALD is the technique of choice for depositing highly conformal thin films and enabling nanoscale device fabrication. The benefits of ALD are its unique ability for depositing ultrathin films on complex-shaped substrates with a thickness control at the atomic level and high conformality [15, 16]. Other processes used for nanolaminate formation include physical vapor deposition (PVD), dip coating, doctor blading, spray deposition, Langmuir–Blodgett, or chemical vapor deposition (CVD) methods. PVD methods, inclusive of evaporation and sputtering, have limited ability to coat structures with high aspect ratios and poor step coverage. CVD methods have problems with gas-phase reactions and impurities that interfere with nucleation or form defect voids during material growth. The ALD deposition technique is based on surface self-limiting reaction and deposition occurrences through alternating exposure of precursors and reactants, which provides precise control of the deposition thickness, even in ultrathin films and high-aspect-ratio holes and trenches [15–17]. As trench aspect ratio development continues to increase from 1:25 to greater than 1:200, ALD remains the most effective technique for deposition into the trenches.

ALD has basic characteristics of self-assembly, in which the surface reactions drive atoms and molecules to arrange themselves into functional structures. The ALD film growth process takes place in cycles, one cycle consisting of the following five steps, which are repeated until the desired thickness is obtained: (a) a self-terminating reaction of the first reactant A (precursor exposure), (b) an evacuation or purge to remove the non-reacted reactants and reaction products, (c) a self-terminating reaction of the second reactant/precursor B, (d) an evacuation or purge, and (e) after the reaction being terminated and all the by-products removed, a formation of monolayer of material C (Fig. 3). It is obvious that the self-termination feature enables extraordinary control over film thickness and also quality (pin holes, defects, and clustering are strongly reduced due to the selective interaction at the A/B interface). Another central advantage is the ability to coat 3D features with a conformal layer [17, 18].

ALD has emerged in the semiconductor industry as the technique to deposit very thin, highly insulating (high-*k*) oxides to serve as dielectrics in trench capacitor structures. ALD has been successfully applied to the depositing of oxide materials, including insulators and other ultrathin metal-oxide films for device applications. Currently the most investigated materials are HfO_2, ZnO, TiO_2, and/or Al_2O_3, because of their electrical and mechanical properties. Furthermore, for nanolaminates, the material property of interest is the dielectric constant, k, and the bandgap, as seen in Table 1.

ALD has been used to process layer deposition of alternating high-k dielectrics (needed for high capacitor density) and wide bandgap material (needed for low capacitor leakage) stacks of TiO_2 and Al_2O_3 of 3 nm layer thickness. Leakage and charge storage are not only dependent upon the material composition but is directly impacted by the nanolaminate stack thickness as well. TEM of a TiO_2 and Al_2O_3 stack indicates good uniformity, no pin holes, or interface spacing, as in Fig. 4.

Aluminum Anodic Oxide (AAO) Processing. Anodization is an electrolytic



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 3 Atomic layer deposition process illustrating the surface nucleation of thin film A and thin film B conforming to patterned layer

Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Table 1 List of ALD material parameters of high-k dielectrics

Materials	Dielectric constant	Bandgap (eV)	Growth rate (nm/cycle)
TiO_2	70	3.5	0.03
HfO_2	25	5.8	0.1
Al_2O_3	9	8.8	0.05
ZnO	2.5	3.33	0.08
AlN	9	6.2	

oxidation process in which the metal surface is converted primary to an adherent oxide coating having a desirable thickness and structure. Aluminum anodic oxide (AAO) templates, formed on aluminum by anodizing in acid electrolytes, are prepared following a chemical process. The AAO nanotemplates typically consist of densely packed high-aspect-ratio pores aligned almost perfectly to each other as in Fig. 5. The first published work in the field was Masuda et al. [13]. These papers describe the procedure to fabricate AAO pores on Al foil with ordered hexagonally parallel nanotubes using a two-step anodization technique.

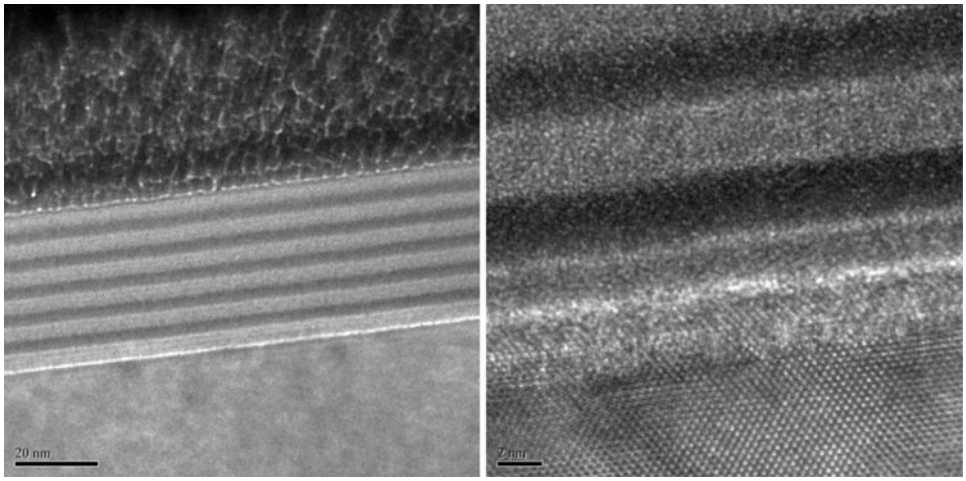
Parameters governing the successful preparation of an ideal ordered porous structure are:

1. Anodization voltage and current, which determines pore diameter, cell dimension, and pore distribution. The voltage applied in the electrodes is constant during the whole process.
2. Electrochemical bath temperature.
3. Type and concentration of electrolyte.

A two-step anodization method is used to create AAO templates:

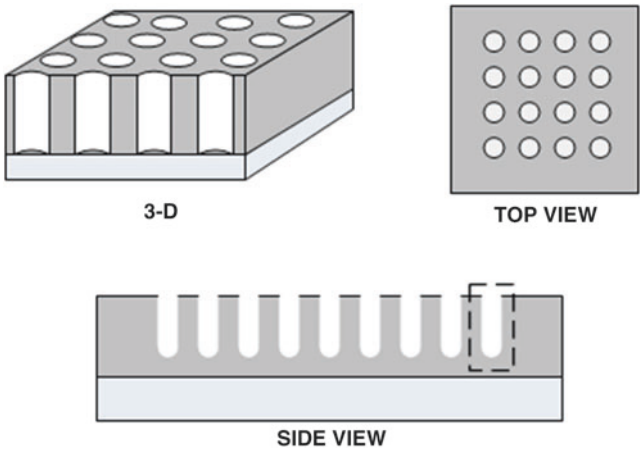
1. In the first step of the process, a thick porous alumina layer gradually grows over a very thin dense barrier oxide layer, which acts as an interface between the pure Al surface and porous alumina oxide layer. In the next step the pore array is created, which means that pore diameter is a function of the first anodization. Also, AAO pore diameter increases linearly with the anodization time (characterization must be made in order to find the limiting time in which the ordered array starts to dissolve).
2. The porous oxide layer that is formed is removed and leaves behind an ordered concave textured structure on the surface. The second step begins the Al anodized process again under the same conditions. This acts on the concave sites and forms a deep uniform pore growth, which means that pore height is a function of the second anodization.

The AAO template allows controllability of inner pore distance, pore diameter, and pore depth by changing the anodization conditions. A large-area AAO template can be generated



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 4 ALD stacks of Al_2O_3 (*white*) and TiO_2 (*dark*) of 3 nm thickness

Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 5 Schematic layout of ideal AAO porous structure



using pure aluminum (99.99 wt% pure metal) foil of 0.1 mm thickness. Aluminum foil can be ultrasonically cleaned in ethanol and methanol and dried at 100 °C. Using a solution of phosphoric acid H_3PO_4 (85 wt%), sulfuric acid H_2SO_4 (98 wt %), and water H_2O (2:2:1 by volume), the foil needs to be electrochemically polished at a constant voltage of 10 V for 14 min at room temperature to remove micro-roughness from the surface, which can affect the final quality of the films. A solution of 0.3 M oxalic acid (C_2H2O_4) can be used to perform an initial anodization at a voltage of 40 V for 90 s. After anodization the aluminum template may exhibit a surface roughness due to the internal stress of the different atoms during the anodization process, which may lead to a defective formation of the structures in the porous alumina template. The as-prepared AAO template is placed in an aqueous mixture of phosphoric acid (6 wt%) and chromic acid (0.2 M) to remove anodic oxide formed in the initial anodization. Table 2 gives processing details to be used to form an AAO template. Finally, the sample can be further anodized under the same conditions as the initial anodization for 1 h to create deeper pores of high aspect ratio as in Figs. 6 and 7.

Since aluminum foil is a very flimsy substrate material, it is very difficult to go through the photolithography process without bending/damaging the AAO template. In order to fabricate devices on the AAO templates, an anodic bonding technique should be used to attach the aluminum structure to a borosilicate glass [19] as in Fig. 8.

High-purity, mechanically polished aluminum foils (99.99 % metal basis) and borosilicate glass are recommended. Aluminum pieces can be mechanically polished using alumina particles of 1 μm diameter. The pieces are then ultrasonically cleaned with methanol and dried at 100 °C. To secure bonding, the Al/glass are set between two plates of stainless steel (electrodes) and heated to 380 °C. After reaching the desired temperature, a dc voltage of 950 V was applied for 20 min. Subsequently, the setup is cooled down to room temperature to avoid cracking of the glass.

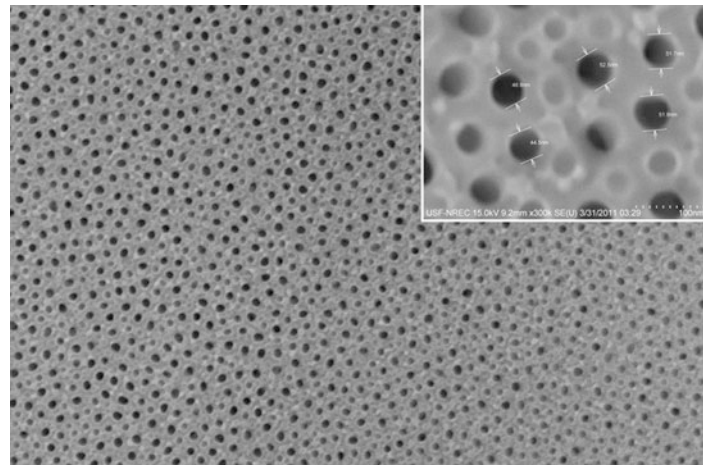
Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Table 2 Processing parameters for AAO template

Process parameters	**Range/ values**
Voltage – determines pore diameter, distribution	35 ∼ 45 V
1st anodization time – contributes to pore size and cell dimension	90–120 s
2nd anodization time – contributes to pore size and cell dimension	1 h
Pressure	Atmosphere
Temperature	RT
Electrolyte mixture	**Range/ values**
Phosphoric acid – contributes to the removal of anodic oxide	6–8 wt%
Chromic acid – contributes to the removal of anodic oxide	0.2 M
Oxalic acid – contributes in the initial formation of pores	0.3 M

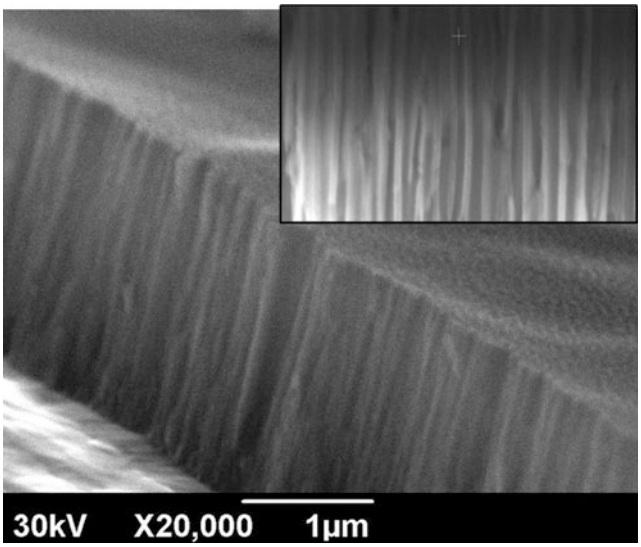
Photolithography Processing (PR). A five-level mask can also be used to fabricate nanocapacitors using alternating nanolaminates. Test structures on the mask set are necessary to characterize and analyze the nanocapacitors of planar and trench configuration per Fig. 9.

A top-down view of planar nanolaminate capacitors can be seen in Fig. 10.

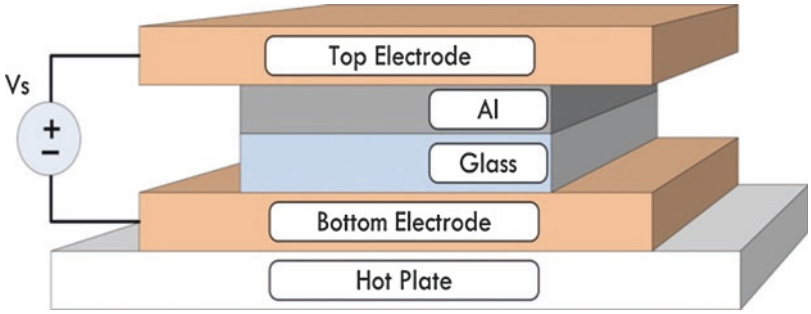
A deep reactive-ion etching (DRIE) system can provide advantages when used for the processing of high-aspect-ratio trenches. The DRIE has the ability to efficiently etch vertical structures in Si wafers with high aspect ratios. DRIE technology is based on the Bosch process, which uses high-density plasma to alternatively etch silicon. Some of the DRIE process advantages include high resolution and cleanliness, process control, use of small amounts of chemicals, and the avoidance of handling acids and solvents. However, in some structures, there can be a residual etch-resistant polymer that remains on the sidewalls, which could impact the trench width uniformity throughout the depth of the trench. SEM or TEM of a



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 6 SEM image of aluminum foil after two-step anodization process (pore size ∼50 nm)



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 7 Cross-sectional SEM image of aluminum foil after two-step anodization process (pore height ~2 µm)



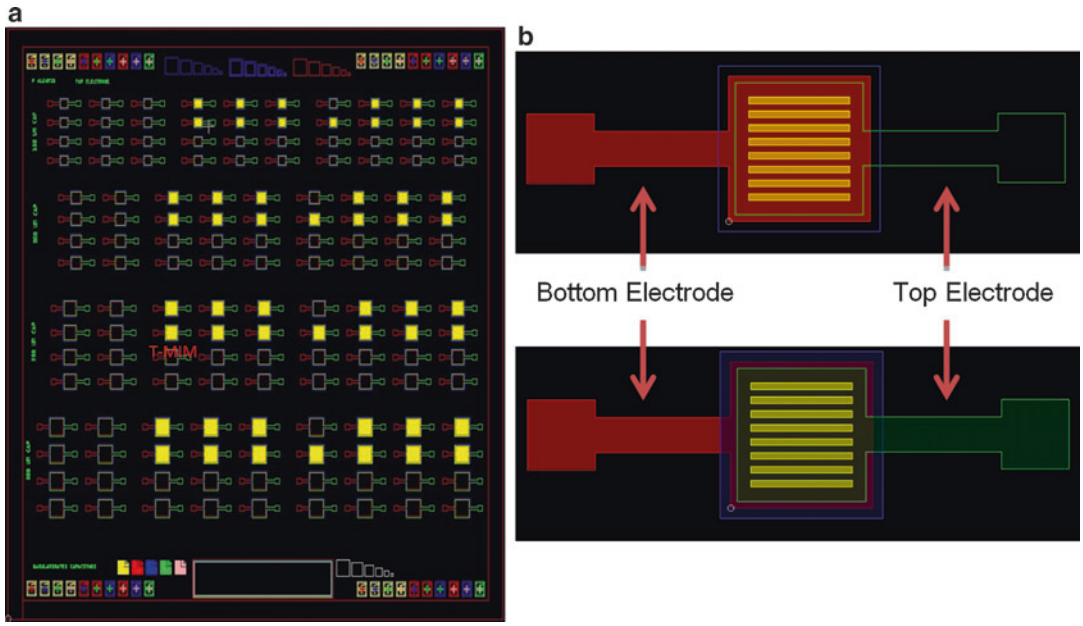
Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 8 Schematic of anodic bonding setup

cleaved etched wafer can provide measurements to determine aspect ratios. Figure 11 shows high AR trenches fabricated in Si ranging in width from 5 microns to 11 microns and in height from 65 microns to 90 microns.

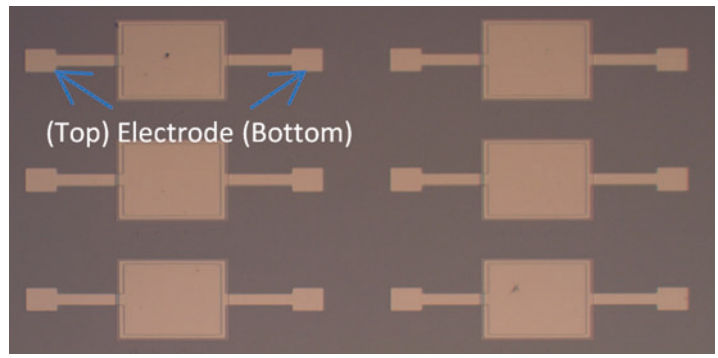
Future Directions

Nanolaminate development for nanocapacitor structures requires the integration of multiple disciplines, from materials to physics to chemistry to engineering. Since nanolaminates capitalize on superior properties of each lamina, these structures have the potential to serve as the basis for the next-generation energy storage devices that make use of densely packed interfaces (multilayer/nanolaminates) and have the capability of delivering power to renewable energy sources. These nanostructures are becoming increasingly important for ICs due to properties such as high surface area and downscaling of the technology (achieving densely packed structures from thin films).

When fabricated within a high-aspect-ratio template (AAO or PR processing) using ALD nanolaminates, nanocapacitors filled with alternating high-k and wide bandgap materials will have the ability to simultaneously achieve high capacitance density, high breakdown voltage, and lower leakage current, as compared to a planar capacitor or single layer filled trench capacitor.



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 9 (**a**) Top-down view of nanocapacitor layout with test structures. (**b**) Layout of bottom electrode (*red*), top electrode (*green*), and trench design (*yellow*)

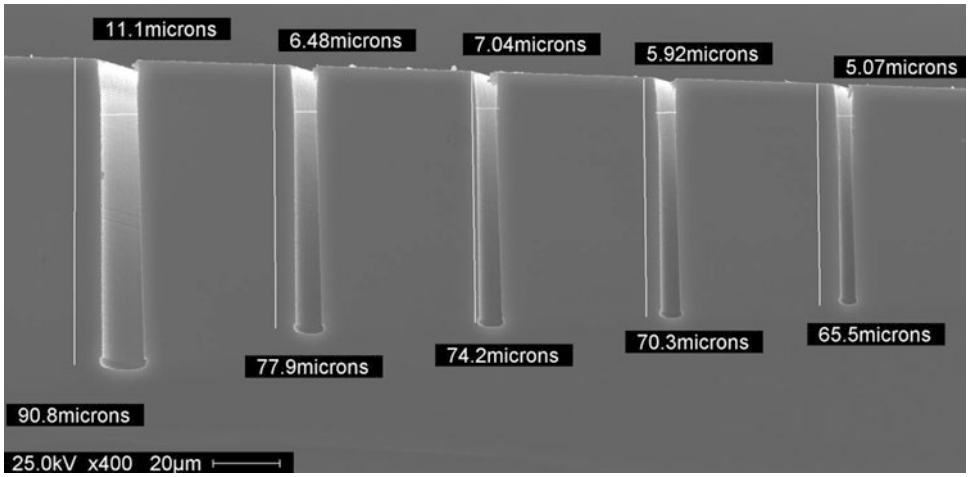


Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 10 Top-down view of patterned nanolaminate stack

ALD processing of trench-refilled capacitors have shown an incremental increase in the capacitance per unit area of 10 μF cm^{-2} and 100 μF cm^{-2} [4]. This significantly exceeds previously reported values for capacitors in porous templates [14].

Nanocapacitors with ALD nanolaminates enhance surface-to-volume ratios, improve energy storage properties, and maintain high power density. As systematic research and development continues, the fabrication of nanolaminates opens a new way to fabricate high dielectric constant materials for energy storage applications.

Creation and implementation of nanolaminate nanocapacitor structures has the potential to transform the life span and cost efficiency of energy storage devices for global solutions by 10 times. ALD and AAO technologies can address the current challenge of material reliability/durability, lower processing cost, and novel utilization of alternating nanolaminates in a trench structure. The high payoff for energy storage will improve



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 11 Trench etch using DRIE of 2,400 W RF power

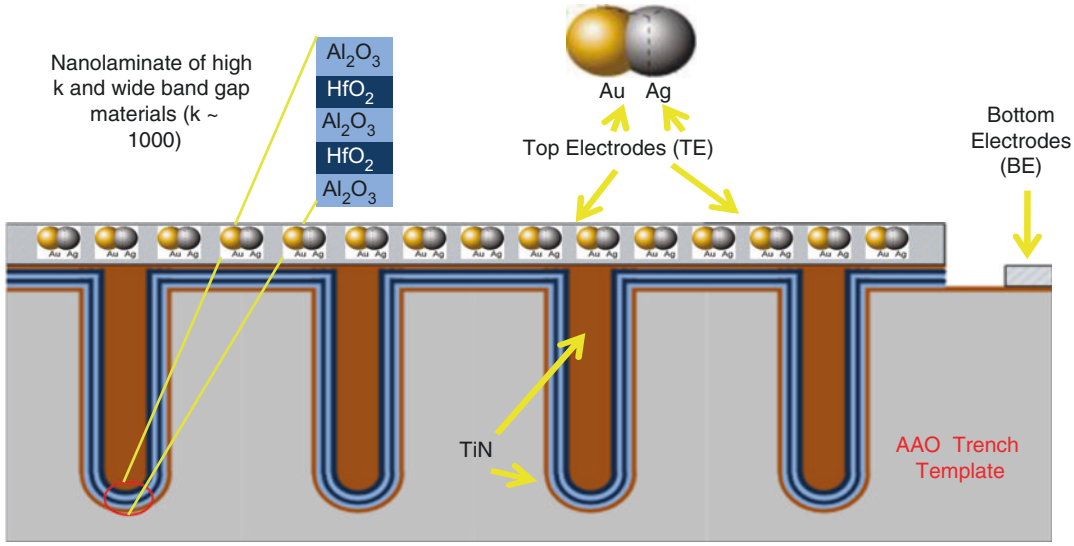
current equivalent planar capacitance (EPC) performance of existing supercapacitors by one order of magnitude from 100 μF cm^{-2} up to ~1,000 μF cm^{-2} using a high-aspect-ratio nanolaminate trench design. EPC can be estimated knowing the dimensions of a porous AAO structure and the thicknesses of the bottom electrode and nanolaminate (insulator) layers.

It has been studied that supercapacitors that use very large surface area electrodes have increased capacitances. The capacitance and energy density of a supercapacitor scales linearly with electrode surface area. Most of the work has been concentrated on the use of carbon-based materials in electrochemical supercapacitors. Carbon-based materials, like graphene, which is conductive, noncorrosive, thermally stable, inexpensive, and potentially high in surface, have been used as electrodes. Gomez et al. used a graphene nanocomposite to achieve a capacitance of 300–500 F/g. Other forms are carbon aerogels, carbon nanotubes, and carbon black particles. Therefore, a carbon-based material that has large surface area, such as graphene, may increase capacity. Two issues that must be addressed include high-temperature processing of the carbon-based materials and conformality.

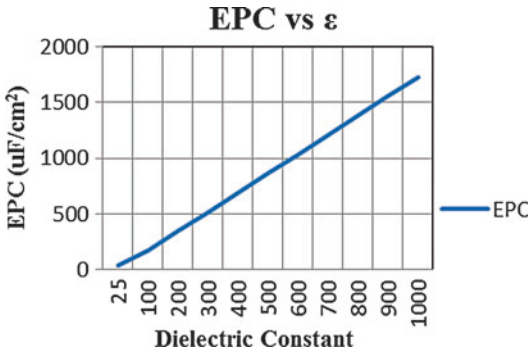
In an effort to implement enhanced electrodes that will conform to a nanocapacitor nanolaminate structure, ALD processing of metal electrodes can be performed. Several materials can be considered as potential choices to be used in fabricating the electrodes, such as platinum (Pt), nickel (Ni), aluminum nitride (AlN), ruthenium (Ru), titanium nitride (TiN), carbon nanotubes (CNTs), or nanoparticles.

Through the utilization of a trenched structure, nanolaminates, and enhanced metal electrodes, a trench metal-insulator-metal nanocapacitor structure can be achieved. Modeling the effect of area on the capacitance using geometrical parameters of the structure leads to the normalization of capacitance per planar area using the following EPC equation $EPC = D * \frac{2\pi\kappa\varepsilon_0 L}{In\left(\frac{b}{a}\right)} + \frac{\kappa\varepsilon_0}{t_i}$ where D is the pore density, k is the dielectric constant of the insulator, ε_0 is the permittivity of free space, L is the pore depth, and t_i is the thickness of the insulator. Constants, a and b, are dependent upon the thickness of the ALD layers and the radius of the pores [4].

Therefore, based on the EPC formula given for a capacitor formed using AAO pores, a calculated theoretical EPC value for a trenched configuration filled with alternating nanolaminates and having advanced electrodes (Fig. 12) can potentially yield a capacitor density up to ~1,000 μF cm^{-2} as presented in Fig. 13.



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 12 Nanolaminate filled trench capacitor (T-MIM) structure using anodic aluminum oxide template and enhanced electrodes



Implementing Alternating Nanolaminates in Trenched Energy Storage Systems, Fig. 13 Theoretical EPC values plotted as a function of dielectric constant (25–1,000) for trench capacitor of alternating nanolaminates

Challenges that persistent in the technology are the need for material modeling of inter- and intrafacial behavior at the molecular level, profiling of nanolaminate stack performance, and implementing inexpensive, scalable manufacturing of these versatile structures. When these challenges are met, technological advancements using nanolaminates will progress toward a new generation of nanocapacitors for 3D electronic device energy storage and implantable biomedical system-on-a-chip energy storage.

Cross-References

- ALD
- Atomic Layer Deposition
- Maxwell–Wagner Effect

References

1. National Academy of Engineering, Grand Challenges for Engineering, www.engineeringchallenges.org
2. Halper, M., Ellenbogen, J.: Supercapacitors: A Brief Overview. MITRE Nanosystems Group, McLean, VA (2006)
3. Conway, B.E.: Electrochemical Supercapacitors: Scientific Fundamentals and Technological Applications. Kluwer-Plenum, New York (1999)
4. Banerjee, P., Perez, I., Henn-Lecordier, L., Lee, S.B., Rubloff, G.W.: Nanotubular metal-insulator-metal capacitor arrays for energy storage. Nat. Nanotechnol. **4**, 292–296 (2009)
5. Huang, C., Cheng, C.-H., Lee, K.-T., Liou, B.-H.: High-performance metal-insulator-metal capacitor using quality properties of high- TiPrO dielectric. J. Electrochem. Soc. **156**(4), G23–G27 (2009)
6. Jeon, W., Chung, H.S., Joo, D., Kang, S.W.: TiO_2/Al_2O_3/TiO_2 Nanolaminated thin films for DRAM capacitor deposited by plasma-enhanced atomic layer deposition. Electrochem. Solid-State Lett. **11**, H19–H21 (2008)
7. Gerritsen, E., et al.: Evolution of materials technology for stacked-capacitors in 65 nm embedded-DRAM. Solid-State Electron. **49**, 1767–1775 (2005)

8. Kemell, M., et al.: Si/Al_2O_3/ZnO: Al capacitor arrays formed in electrochemically etched porous Si by atomic layer deposition. Microelectron. Eng. **84**, 313–318 (2007)
9. Shi, L., Yin, J., Yin, K.B., Gao, F., Xia, Y.D., Liu, Z. G.: An investigation into ultra-thin pseudobinary oxide $(TiO_2)_x(Al_2O_3)_{1-x}$ films as high-k gate dielectrics. Appl. Phys. A Mater. Sci. Process. **90**, 379–384 (2007)
10. Jinesh, K.B., Lamy, Y., Klootwijk, J.H., Besling, W.: Maxwell-Wagner instability in bilayer dielectric stacks. Appl. Phys. Lett., **95**, 122903 (2009), AIP Publishing LLC. doi:10.1063/1.3236532
11. Wei, L., et al.: Giant dielectric constant dominated by Maxwell–Wagner relaxation in Al_2O_3/TiO_2 nanolaminates synthesized by atomic layer deposition. Appl. Phys. Lett. **96**, 162907 (2010)
12. Auciello, O.H., Lai, B-K., Lee, G., Katiyar, R.S.: Nanolaminates of Al_2O_3/TiO_2 with Giant dielectric constant low leakage – low loss: extended frequency operation for new-generation nanoelectronics and energy storage devices. Uchicago Argonne LLC., USPTO US 20130264680 A1 (2013)
13. Masuda, H., Fukuda, K.: Ordered metal nanohole arrays made by a two-step replication of honeycomb structures of anodic alumina. Science **268**, 1466–1468 (1995)
14. Sohn, J., et al.: Fabrication of high-density arrays of individually isolated nanocapacitors using anodic aluminum oxide templates and carbon nanotubes. Appl. Phys. Lett. **87**, 123115 (2005)
15. George, S.M.: Atomic layer deposition: an overview. Chem. Rev. **110**, 111–131 (2010)
16. Ritala, M., Leskelä, M.: Atomic layer deposition. In: Handbook of Thin Film Materials edited by H.S. Nalwa, Academic Press, San Diego, **1**(2), p. 103 (2002)
17. Ladanov, M., Algarin-Amaris, P., Matthews, G., Ram, M., Thomas, S., Kumar, A., Wang, J.: Microfluidic hydrothermal growth of ZnO nanowires over high aspect ratio microstructures. Nanotechnology **24**(37) p. 1–9 (2013)
18. Ladanov, M., Algarin Amaris, P., Villalba, P., Emirov, Y., Matthews, G., Ram, M.K., Thomas, S., Wang, J., Kumar, A.: Effects of the physical properties of atomic layer deposition grown seeding layers on the preparation of ZnO nanowires. J. Phy. Chem. Solids, Ms. Ref. No.: PCS-D-13-00179R1 **74**, 1578-1588 (2013)
19. Xing, Q.F., Sasaki, G., Fukunaga, H.: Interfacial microstructure of anodic bonded Al/glass. J. Mater. Sci. Mater. Electron. **13**, 83–88 (2002)

Imprint Lithography

► Nanoscale Printing

In Situ STM

► Electrochemical Scanning Tunneling Microscopy

In Situ Transmission Electron Microscopy: Mechanical Testing

Tobin Filleter[1] and Allison M. Beese[2]
[1]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada
[2]Department of Materials Science and Engineering, Pennsylvania State University, University Park, PA, USA

Definition

In situ transmission electron microscopy (TEM) mechanical testing is a technique used to measure the mechanical properties and behavior of small structures while simultaneously imaging them using a TEM.

Introduction

The transmission electron microscopy (TEM) is one of the most powerful tools used in materials science to unveil the structural and compositional properties of materials [1]. The coupling of conventional TEM with complimentary techniques, which are used to further investigate TEM samples directly inside the TEM chamber during imaging, has led to a broad field of "in situ TEM" that spans areas that include electrical, thermal, electrochemical, optical, and mechanical testing. The latter area of in situ TEM mechanical testing takes advantage of the high spatial resolution of the TEM to characterize strain and deformation in materials down to atomic scales, while providing measurements of force and stress acting on the samples under a number of different mechanical loading configurations. Although this field has been significantly expanded recently

due to advancements in the miniaturization of mechanical testing systems, the first implementation of this class of techniques was in the 1950s through the application of simple straining holders that enabled mechanical strain to be applied to materials while being imaged under the TEM electron beam [2]. For a comprehensive in-depth and historical review of the field of in situ TEM mechanical testing, the reader is referred to the recent review article by Legros [2].

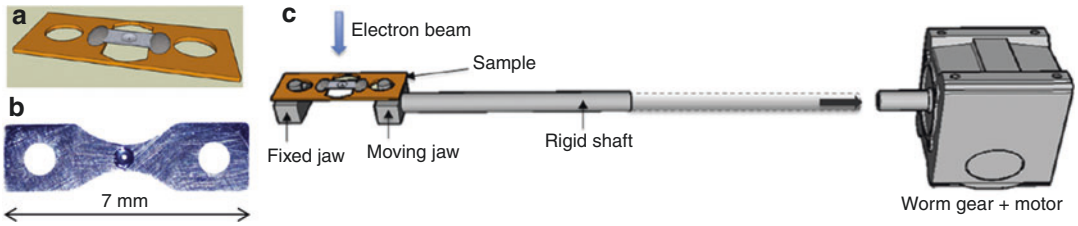
Since its inception several decades ago, in situ TEM mechanical testing has made notable advances to enable quantitative mechanical testing configurations, which include (i) tension and compression, (ii) nanoindentation, (iii) bending, and (iv) friction and wear testing [3–5]. These various mechanical testing configurations, executed within the full view of the TEM, have enabled a detailed understanding of the mechanics of materials with micro- and nano-confined structures elucidating a number of size-scale phenomena. By conducting mechanical testing in situ TEM, researchers can take advantage of the many analytical materials characterization techniques allowed by TEM to study materials properties, such as local atomic resolution imaging of dislocations and defects, selected area electron diffraction (SAED) to identify crystalline structure, and x-ray energy-dispersive spectroscopy (XEDS) and electron energy loss spectroscopy (EELS) to identify local elemental composition and electronic structure [1]. Notable applications include the in situ TEM study of 1D nanotubes and nanowires [6], indentation of nanograined materials [4], deformation of 2D ultrathin films, tensile properties of carbon-based fibers, and tribology of hard coatings [7]. Toward advancing in situ TEM mechanical testing, a number of key challenges and limitations continue to be considered and addressed when developing testing techniques. These include constraints on sample size and geometry, electron beam-induced sample modification and deposition effects, electron beam energy constraints, limits on the rate of TEM image acquisition, and environmental limitations. For example, the traditional limitation of conducting TEM experiments only in a high-vacuum environment has been addressed recently by the emergence of high-resolution environmental TEMs as well as in situ TEM liquid cells for conducting experiments under a variety of controlled environmental conditions.

The following sections detail in situ TEM mechanical testing devices and characterization techniques as well as the most prominent mechanical testing modes and important mechanical phenomena that they have unveiled to date.

In Situ Mechanical Testing Devices and Techniques

A wide range of mechanical testing techniques have been developed to allow for tension, compression, indentation, shearing, and bending of materials in situ a TEM with real-time visualization of strains, dislocations, and deformation mechanisms. When forces can also be measured, these methods allow for fundamental understanding of the constitutive behavior of materials at the nanoscale.

Relatively basic mechanical testing has been performed in situ TEM imaging using traditional tensile stages where a sample is fixed at one end and loaded in tension by moving the other gripped end of the sample. One of the first in situ TEM tensile tests was performed on 300 μm thick aluminum foils by Wilsdorf. In these experiments, a TEM holder was modified such that a sample could be gripped between a stationary pin on the holder and a pin attached to a moving shaft internal to the shaft of the TEM holder [8]. The shaft could be driven manually by a screw device or through a hydraulic system allowing for the application of tension to a dogbone-shaped specimen (Fig. 1). While this setup provided for the first time the direct visualization of dislocations moving through a foil being deformed in situ the TEM, it did not allow for the measurement of forces. Gatan has since developed commercial straining holders that allow for the deformation of films clamped by pins at two ends, using screw-driven actuator to pull one side of the film, also with no direct measurement of force. In situ TEM tensile testing can also be conducted on 1D nanostructures, such as nanotubes and nanowires,



In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 1 Schematic of first in situ TEM tensile testing setup. (**a**) Schematic of tensile sample on grips to interface with TEM holder. (**b**) Example of the tensile sample geometry. (**c**) Schematic of sample installed between fixed jaw and moving shaft (Images and caption adapted from [2])

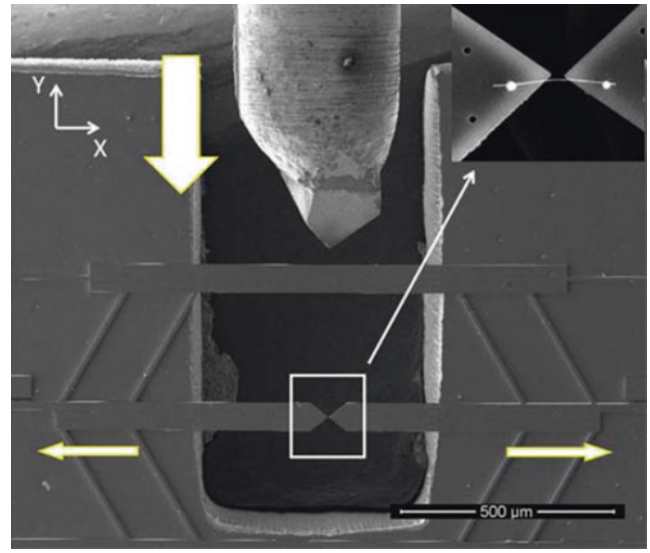
using in situ TEM 3-axis nanomanipulators (e.g., nanofactory) with integrated atomic force microscopy (AFM) like force sensors. These systems allow for the straining of the sample using a piezo-driven manipulator and the measurement of forces using either the direct measurement of deflection of the force sensor beam via TEM images or through electronic piezoresistive force sensor readouts. Similar in situ TEM manipulators with integrated scanning tunneling microscopy (STM) probes can also be used for in situ TEM friction and wear measurements.

Both passive and active microelectromechanical systems (MEMS)-based devices have been developed to apply deformation to, and measure loads through, 1D and 2D materials. Haque and Saif developed MEMS-based devices consisting of opposing shuttles across which to mount a sample [9]. Tension is applied to the sample by connecting one end of the MEMS chip, and thus one shuttle, to a piezoelectric actuator. The other shuttle is connected to a set of folded silicon beams of known stiffness, fixed to the TEM holder. Sample displacements are measured by the relative motion of the opposing shuttles, and the forces are calculated by measuring the deflection of the folded silicon beams of known stiffness.

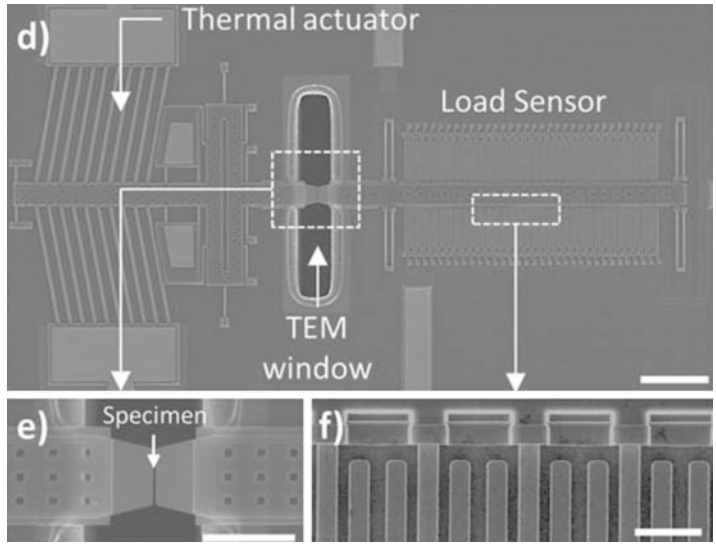
Lu and coworkers developed a push-to-pull micro-mechanical silicon-based device that converts the compression actuation of a nanoindenter to a tensile load on a sample of interest (Fig. 2) [10]. The device consists of opposing shuttles across which a sample is mounted. The shuttles are attached to oppositely pointing chevron beams, which are connected to an outer frame. Once the sample spans the shuttles, a nanoindenter compresses the outer frame, pushing the shuttles apart and applying a tensile load to the sample. The displacement can be directly measured in the TEM, and the force can be calculated by subtracting the force due to the stiffness of the device from the forces measured.

Espinosa and coworkers have developed two types of active MEMS-based tensile and compressive testing stages [6]. Both basic types consist of two opposing shuttles across which the sample is mounted. One of the shuttles is attached to a set of interdigitating combs that act as a capacitive force sensor, while the other shuttle is attached to an actuator. The actuator in one type of device types is an electotrostatic comb drive in which a voltage is applied between the stationary comb and the movable comb attached to the shuttle, in order to provide a displacement to the nanosample. The second type of actuator consists of chevron beams attached to the actuating shuttle (Fig. 3). The chevron beams are heated up via Joule heating as an electrical current is sent through the beams, causing them to elongate. For tensile loading, the chevron beams are pointed such that as they expand, they pull the shuttle away from the force sensing shuttle, and for compressive loading, they are pointed toward the force sensor so that as they heat up, they compress or indent the sample. Displacements on the order of 1 nm can be applied to the sample, with resolutions in load of approximately 12 nN depending on the stiffness of the force sensor.

Commercial TEM holders that allow for the indentation of samples or the application of

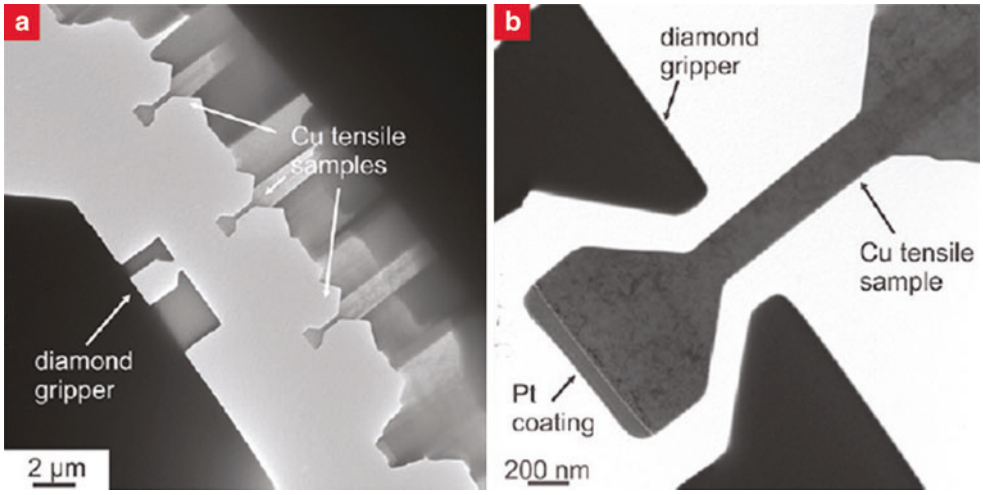


In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 2 SEM image of push-to-pull MEMS-based testing system. The sample is mounted across the two pointed shuttles highlighted in the inset. The nanoindenter is shown coming from the top of the image. When the nanoindenter moves downward and impinges on the top bar of the frame, the oppositely pointing chevron beams push the shuttles apart, applying tension to the mounted sample (Image from [10])



In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 3 SEM images of a tensile MEMS-based testing stage designed by Espinosa and coworkers. (*Top*) MEMS device actuated by thermal heating, where the chevron beams on the left are heating, causing the left shuttle to apply tension to the sample mounted across the opposing shuttles. Scale bar: 100 micrometers (*bottom left*) The right side is a capacitive load sensor (*bottom right*) (Figure and caption adapted with permission from [6])

compressive loads in situ the TEM have been developed. Nanofactory developed a modified TEM holder that allows for indentation into a sample and the measurement of force through a micromachined two-plate capacitive load sensor [3]. However, in this configuration the depth of the indenter is not calculated. Hysitron developed a commercially available picoindenter. The sample is mounted to a fixed end of the holder, and the nanoindenter tip moves during the test. The holder has 3-axis coarse positioning as well as a three-dimensional piezoelectric actuator for fine positioning. The Hysitron design incorporates load sensing through a capacitive



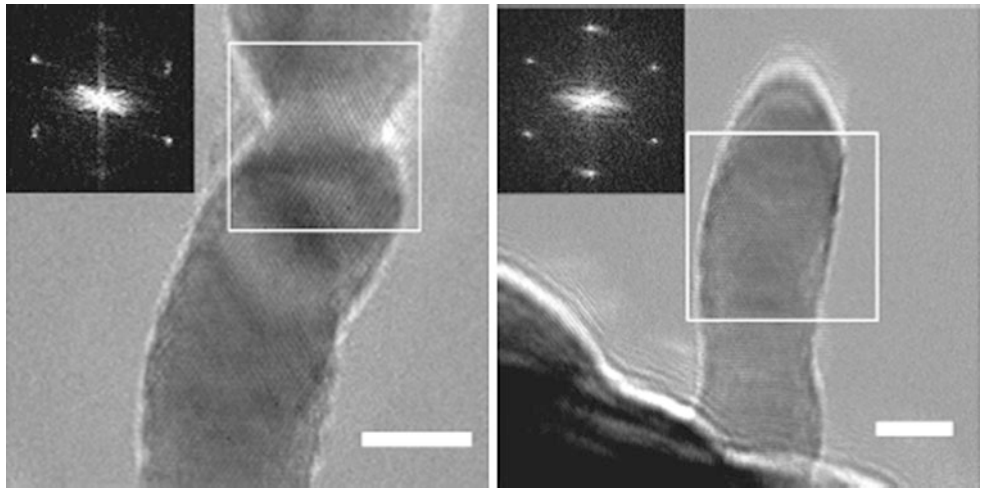
In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 4 (**a**) TEM image showing multiple Cu tensile samples on the right and the diamond tip used to grip the samples. (**b**) Higher-resolution TEM image showing the configuration of the sample and gripper before testing (Image from [11])

transducer, which consists of an electrostatic load generator to apply force and a capacitive displacement sensor that measures the depth into the sample that the tip has traveled [3]. Hysitron has also developed a push-to-pull MEMS-based device, which can be used to convert the compressive load of the nanoindenter into a tensile load across opposing shuttles. While these devices may be used to indent samples using traditional Berkovich-shaped tips, the tips can be replaced by flat punches to compress materials. In addition, Kiener and Minor modified the Hysitron nanoindenter tip using focused ion beam milling to machine a conductive diamond grip to engage and pull on the shoulders of a traditional dogbone-shaped tensile sample, allowing for the device to also be used to probe tensile properties of nanomaterials (Fig. 4) [11].

Indirect testing modes have also been used in situ TEM to measure mechanical behavior including beam-induced heating deformations and beam-induced resonating of suspended structures. The heating of the beam itself has been used to induce film bending, thereby applying strains. The first such experiment was performed by Hirsch, Horne, and Whelan in 1956, in which they watched dislocations move through aluminum as their aluminum film was deformed due to thermal strains. The electron beam has also been used to induce resonant oscillations of nanostructures such as CNTs and nanowires. In this case, the amplitude of the oscillation of the free end of the nanostructure with one end fixed to the sample holder can be measured to determine the elastic modulus [2].

Simultaneous TEM Characterization Tools

One of the most beneficial aspects of in situ TEM mechanical testing is the ability to obtain high-resolution TEM (HRTEM) images of specimen under deformation during mechanical testing. This enables a direct identification of any changes in structure, the nucleation and movement of dislocations, the presence of defects, as well as fracture behavior. Figure 5 shows an example of HRTEM images recorded during the mechanical testing of a cold weld between two gold nanowires. The crystalline structure of both the gold nanowires and the cold-welded region between the two nanowires is clearly observed in the atomically resolved TEM images. The insets in Fig. 5 show fast Fourier transforms (FFT)



In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 5 HRTEM images of a cold-welded gold nanowire before (*left*) and after (*right*) mechanical pulling experiment to break the cold-welded bond between the nanowires. Scale bars: 5 nm (Images and caption adapted from [12])

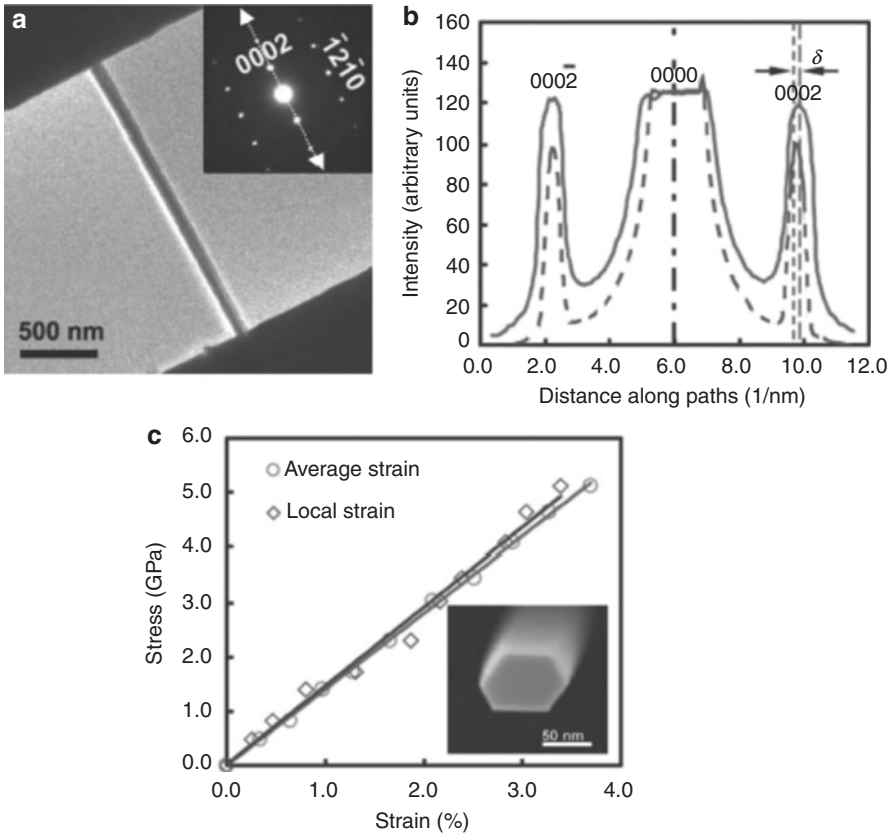
calculated for the images revealing the diffraction patterns that can be used to identify the crystalline structure and measure the lattice spacing.

Diffraction of materials subjected to in situ TEM mechanical testing can also be obtained directly using selected area electron diffraction patterns collected by TEM. In this TEM mode, a parallel beam of electrons is passed through the sample and an aperture inserted in the electron beam path is used to select the area of the sample for which the diffraction pattern is desired. Local diffraction measurements of the specimen under study can be used to determine the crystalline structure and the lattice spacing and identify defects. For in situ TEM mechanical testing, local diffraction patterns can be very useful in particular for measuring the local atomic strain at different regions of the deformed specimen. Figure 6 shows an example of a local strain measurement conducted for in situ TEM tensile testing of ZnO nanowires [6].

Elemental and chemical analysis, as well as electronic structure, of specimen subjected to mechanical deformation can also be conducted using in situ TEM mechanical testing combined with XEDS and EELS [1]. XEDS detects the x-rays generated when the electron beam interacts with the sample allowing for the determination of the relative elemental constituents of the sample based on the characteristic energies of the emitted x-rays [1]. EELS, on the other hand, detects the energy distribution of electrons that have passed through a sample allowing for a determination of the chemical bonding and electronic structure of the sample by analysis of the spectrum of the electron energies and in particular the energy lost to inelastic scattering [1]. With application to in situ TEM mechanical testing, EELS has been used, for example, to study changes in the electronic structure of carbon-based nanostructures such as carbon nanotubes that are subjected to bending deformations [5]. EELS has also been used to identify changes in the ratio of sp^2 vs. sp^3 bonding in diamond-like carbon films subjected to mechanical sliding between a tungsten tip and the films in situ TEM [7].

Tensile Testing Applications

Tensile testing in situ TEM observation has been performed on a variety of materials including 1D nanostructures and 2D films, providing real-time observation of dislocations and deformation mechanisms with applied force.



In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 6 Use of diffraction capabilities in the TEM to perform mechanical characterization. Figures (**a**–**c**) correspond to uniaxial testing of ZnO nanowires. Figure (**a**) shows the TEM image of the nanowire being pulled at two ends and its diffraction pattern. An intensity profile along the indicated path is shown in (**b**). The peaks indicate the position of the diffraction spot, which shifts by a reciprocal distance δ as a result of strain. Local strain can be computed from measurement of δ. Comparison of this local strain with the average strain (**c**) allows discarding slippage effects in the grips of the tensile testing device (Figure and caption adapted with permission from [6])

The commercial Gatan straining holder has been used to elucidate the deformation mechanisms in bulk metallic glass with dimensions on the order of 100 nm, showing that on this scale, these materials deform similar to crystalline materials. That is, they do not deform via the predominant shear banding or nanocrystallization seen in bulk metallic glasses, which have much reduced ductility with respect to their nanoscale counterparts [13]. It has also been used to investigate the deformation mechanics of freestanding nanocrystalline nickel films, showing that dislocation motion is the major mechanism for deformation, with evidence of dislocations in 20 nm grains [14]. MEMS devices have also been applied to study thin films. Haque and Saif co-fabricated MEMS devices and aluminum thin films to circumvent problems with sample transfer and mounting. With this technique, they measured the loading and unloading behavior of the films and experimentally showed reverse Hall–Petch behavior as the yield strengths decreased with reduced grain size [15]. They also noted the absence of dislocations in the samples with grain sizes of 50 nm or less, pointing to a possible grain boundary-based deformation and failure mechanism.

Tensile testing has been applied to a wide variety of 1D nanostructures in situ TEM including CNTs and metallic and semiconducting nanowires [6]. Cumings and Zettl performed

some of the first tensile experiments on single multiwalled carbon nanotubes (CNTs) by attaching a nanomanipulator to one end of a CNT and pulling the inner shells out. They did not measure forces but demonstrated that the inner shells could be pulled out and pushed back in multiple times without inducing damage in the CNT shells, concluding that the interactions between CNTs are low friction and wear-free [6]. MEMS-based tensile testing techniques have been used to measure mechanical properties and deformation mechanisms in 1D nanowires and nanotubes. For example, Lu et al. utilized their push-to-pull MEMS chip with a nanofactory nanoindenter to perform a tensile test on a single nickel nanowire, demonstrating the translation of their original tests in the SEM to in situ TEM testing to allow for acquisition of structural information. The MEMS devices developed by Espinosa et al. have also been used to conduct tensile testing on CNTs, CNT bundles, carbon nanofibers, as well as semiconducting and metallic nanowires [6]. Peng et al. used a MEMS-based testing stage to measure the strength of multiwalled carbon nanotubes (CNTs). Through the direct measurement of the CNT structures during testing, an average fracture strength of over 100 GPa for a single CNT shell was measured, which agrees with theoretical calculations for the strength of pristine CNTs [6]. In addition, a unique strain hardening mechanism was demonstrated for penta-twinned silver nanowires via in situ TEM observation of dislocation nucleation and propagation during tensile testing [6].

Ultrathin 2D films have also been recently studied via in situ electron microscopy tensile testing. Zhang and coworkers measured the fracture toughness of graphene [16] using a MEMS-based device similar to that described in [10]. They performed in situ scanning electron microscope (SEM) tensile tests on graphene that had been pre-cracked by micromachining a crack using FIB, showing that Griffith's fracture theory of brittle fracture applies to graphene, whose fracture toughness was measured to be 4.0 ± 0.6 MPa $\times$ $m^{0.5}$. Although the tensile experiments were performed in the SEM, imaging in the TEM was critical to confirm the structure of the graphene on the device and to determine when monolayer or bilayer graphene was tested [16]; furthermore, the approach for testing 2D films can, in principle, be extended for direct in situ testing in TEM.

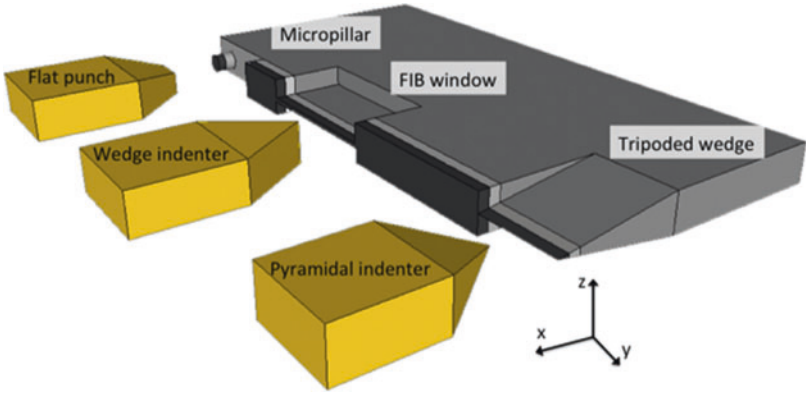
In an alternative tensile testing configuration, by using an adapted Hysitron nanoindenter in which the indenter tip was machined to interface with tensile samples, Kiener and Minor deformed 100–200 nm copper single crystals in two different orientations [11]. They showed that in the sub-micrometer size range, spiral defect sources were operational as was the gradual loss of defects, and both contributed to the hardening behavior of the single crystals. This was contrary to prior studies that proposed either one of the two mechanisms as possible hardening mechanisms, but not both simultaneously.

Indentation/Compression Testing Applications

Indentation and compression testing has been performed using nanoindentation of pillars, local indentation of films and grains, and compression of spherical particles (Fig. 7).

The pioneering micropillar compression tests were performed by Uchic et al., on 0.5–40 μm diameter pillars of nickel and nickel alloys in the SEM using nanoindentation. In this work, they clearly showed a size dependence of plasticity [17]. This work has since been extended to micro/nanopillars in situ the TEM to further elucidate size-dependent plasticity behavior and deformation mechanisms at even smaller length scales.

Shan and coworkers performed compression of nanopillars of nickel, which had diameters on the order of 100–200 nm. Using a TEM holder with a capacitive displacement sensor, they were able to correlate features in the compressive stress–strain curve with real-time experimental observations of dislocation motion. They showed that mechanical annealing was present, meaning that as mechanical force was applied, dislocations left the pillars, resulting in dislocation-free structures. Subsequent deformation was irregular as dislocation sources were activated and then



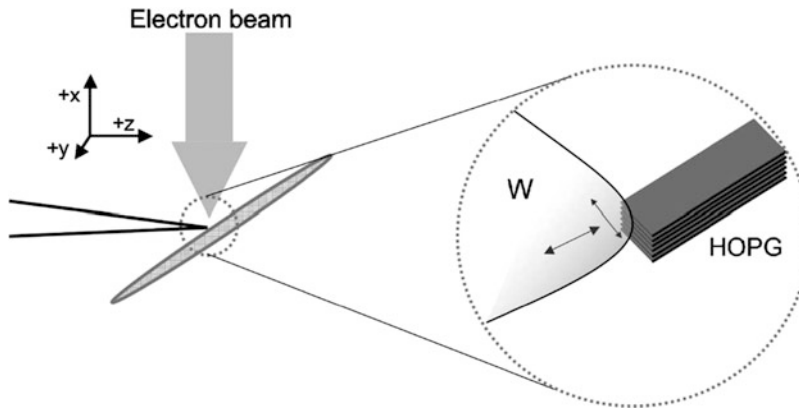
In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 7 Configurations for indentation of pillars using a flat punch, and thinned bulk materials, or films using a wedge indenter or pyramidal indenter (Figure adapted with permission from [2])

exhausted. Yu et al. performed compression tests on titanium alloy single crystal pillars and showed that deformation twinning was the dominant mechanism until the sample size decreased below 1 μm, at which point deformation twinning was no longer active and dislocation plasticity dominated the deformation [18].

Although a great deal of the in situ TEM indentation studies has focused on pillars, several other structures have also been studied. For example, Deneen and coworkers used a custom nanoindentation holder to perform in situ compression of silicon nanospheres, where they clearly visualized elastic, plastic, and fracture behavior of the silicon [19]. Minor and coworkers performed nanoindentation on individual grains in polycrystalline aluminum [20]. By measuring the forces and directly visualizing dislocations, they found that the submicron aluminum grains that were plastically deformed through nanoindentation to a nonzero dislocation density of approximately 10^{14} m^{-2} were able to support shear stresses approaching the theoretical shear strength of aluminum. These results are in stark contrast to the previously held understanding in mechanics that in order to achieve theoretical shear strengths, a material would need to be dislocation-free.

Friction/Wear Testing Applications

One of the greatest challenges in studying friction and wear of single asperity contacts at the micro- and nanoscale is obtaining a detailed knowledge of the size, geometry, and structure of the tribological contact. The application of atomic force microscopy (AFM) has enabled measuring friction forces and imaging surface wear tracks at the nanoscale; however, the shape, size, and structure of the AFM probe is not usually well known. In order to address this limitation, in situ TEM techniques have been developed that enable high-resolution imaging of a tribological contact before, during, and after sliding. The in situ TEM techniques for investigating friction and wear typically use a scanning tip in either a scanning tunneling microscopy (STM) or scanning probe microscopy (SPM) configuration within the TEM chamber. This allows for a feedback signal that controls the interaction between the tip and surface to establish contact and, in the case of SPM, measure the forces acting on the tip during sliding. Of the two configurations, STM is the most straightforward to implement inside TEM as it utilizes a ridged tip and only requires electrical measurements of the tunneling current between the tip and sample. While it does allow for TEM imaging of a sliding contact, it cannot provide quantitative lateral force measurements. SPM, on the other hand, can provide quantitative force measurement; however, the force sensors are more difficult to integrate into the TEM chamber. In particular, beam deflection AFM which is often used for ex situ nanoscale friction and wear measurements requires a flexible cantilever beam, four-quadrant photodetector, laser beam, and optical alignment system to all be integrated into the small available space within



In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 8 STM-TEM holder schematic. A tungsten STM tip is shown on the left in contact with a graphite (HOPG) sample with the contact region exposed to the TEM electron beam. Nanomanipulation allows for microstructural characterization during sliding and indentation (Figure and caption reproduced with permission from [22])

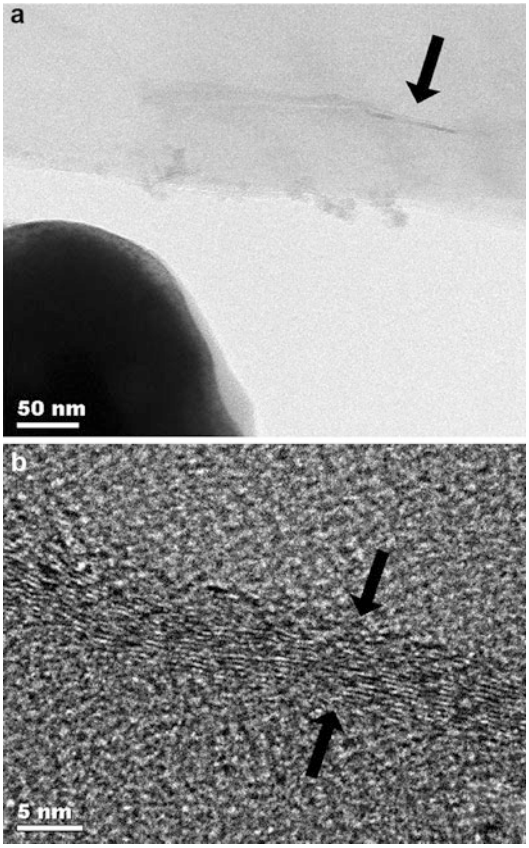
the TEM chamber. While technically difficult, beam deflection AFM in situ TEM has been achieved with dedicated instrumentation [21]. In addition, the application of in situ TEM friction and wear studies requires overcoming the key challenge of having an electron transparent contact. This is accomplished by orienting the contact at an angle with respect to the TEM electron beam and utilizing thin samples as can been seen in Fig. 8.

In situ TEM STM can be used to study the friction and wear of a number of materials including, but not limited to, graphite, diamond-like carbon, and metal films [5, 22]. Figure 9 shows an example of an in situ TEM friction and wear study conducted on graphite. The sharp tungsten STM probe is used as a sliding contact on the graphite surface. Figure 9a shows the sharp tip and a wear track on the graphite samples after 100 cycles. The great benefit of high-resolution in situ TEM imaging is demonstrated in Fig. 9b which shows a TEM image of the same wear track in which the layered atomic structure of the graphite sample is revealed. Metal thin films and diamond-like carbon films studied in situ TEM using the same STM techniques have also shown evidence of mechanical sliding-induced structural changes. In the case of polycrystalline metal films, in situ TEM sliding can result in recrystallization and grain growth, whereas sliding on diamond-like carbon films can result in graphitization of the carbon films which changes the tribological properties [7].

As mentioned above, in addition to qualitatively investigating friction and wear in situ TEM, quantitative force measurements of sliding contacts can also be achieved using integrated force sensors. This has been achieved by both AFM force measurements in situ TEM and MEMS-based force sensing in situ TEM. Beam deflection AFM has been used to measure the lateral forces acting on sliding contacts (e.g., Si-Si contacts) in situ TEM for which the contact size and sliding distance can be directly measured from the simultaneously recorded HRTEM images [21]. Despite the limited progress to date quantitative friction and wear testing, in situ TEM remains a promising area for further technological development.

Future Directions for Research

New directions of in situ TEM mechanical testing include extending the existing tools to additional mechanical testing configurations that are different to those already discussed above. For example, the



In Situ Transmission Electron Microscopy: Mechanical Testing, Fig. 9 In situ TEM friction/wear measurements of graphite. Bright field (**a**) and high-resolution (**b**) images of the wear track after 100 passes with the tungsten probe (Figure and caption adapted with permission from [22])

in situ TEM nanomanipulator-based techniques with integrated force sensors that have previously been used primarily for tensile testing have recently been utilized to study the interfacial properties layered 2D materials, such as molybdenum disulfide, via in situ TEM nanomechanical cleavage of atomic layers [23]. Further implementations of the MEMS-based testing devices described in the previous sections applied to 2D materials in situ TEM, which can characterize fracture properties, for example, are also desired. In addition to applying the techniques to other mechanical testing modes, there is a motivation to improve the functionality of the current testing configurations. Two particular directions that are of great interest are improving the temporal resolution of in situ TEM mechanical testing and controlling the testing configurations. With regard to testing configurations, in situ TEM MEMS-based tensile testers with integrated feedback control schemes are being explored, which would enable displacement controlled tensile testing of nanostructures. With regard to temporal resolution, one of the major current limitations for in situ TEM mechanical testing is that conventional TEM imaging is often restricted to ~1 ms temporal acquisition [24]. This can be a major disadvantage to in situ TEM mechanical testing as often deformations in the materials (e.g., the nucleation and propagation of dislocations) can evolve at very high rates. In addition, the slow capture of TEM images limits the capabilities of the technique to study high strain rate mechanisms, which can play an important role in the mechanics of many material types. Recent developments in dynamic TEM (DTEM) have enabled temporal resolution of TEM imaging on the order of ~1 μs to 1 ns between subsequent image captures for studying dynamic processes (i.e., nucleation/growth and phase transformation) which are envisioned to be integrated with in situ mechanical testing in the future [24].

Cross-References

- ▶ AFM Force Sensors
- ▶ Atomic Force Microscopy
- ▶ Carbon Nanotubes
- ▶ Electron-Beam-Induced Deposition
- ▶ Force Measurement
- ▶ Friction Force Microscopy
- ▶ Hierarchical Mechanics
- ▶ Indentation
- ▶ Mechanical Behavior of Nanostructured Metals
- ▶ MEMS
- ▶ Nanoindentation
- ▶ Nanomaterials
- ▶ Nanomechanics
- ▶ Scanning Probe Microscopy
- ▶ Transmission Electron Microscopy
- ▶ Wear

References

1. Williams, D.B., Carter, C.B.: Transmission Electron Microscopy: A Textbook for Materials Science. Springer, New York (2009)
2. Legros, M.: In situ mechanical TEM: seeing and measuring under stress with electrons. C. R. Phys. **15** (2–3), 224–240 (2014)
3. Warren, O.L., et al.: In situ nanoindentation in the TEM. Mater. Today **10**(4), 59–60 (2007)
4. Nili, H., et al.: In situ nanoindentation: probing nanoscale multifunctionality. Prog. Mater. Sci. **58**(1), 1–29 (2013)
5. Ross, F.M.: In situ transmission electron microscopy. In: Hawkes, P.W., Spence, J.C.H. (eds.) Science of Microscopy. Springer-Verlag, New York (2007)
6. Espinosa, H.D., Bernal, R.A., Filleter, T.: In-situ TEM electro-mechanical testing of nanowires and nanotubes. Small **8**, 3233–3252 (2012)
7. Merkle, A.P., et al.: In situ TEM studies of tribo-induced bonding modifications in near-frictionless carbon films. Carbon **48**(3), 587–591 (2010)
8. Wilsdorf, H.G.F.: Apparatus for the deformation of foils in an electron microscope. Rev. Sci. Instrum. **29**(4), 323–324 (1958)
9. Haque, M.A., Saif, M.T.A.: In-situ tensile testing of nano-scale specimens in SEM and TEM. Exp. Mech. **42**(1), 123–128 (2002)
10. Lu, Y., Ganesan, Y., Lou, J.: A multistep method for in situ mechanical characterization of 1D nanostructures using a novel micromechanical device. Exp. Mech. **50**(1), 47–54 (2010)
11. Kiener, D., Minor, A.M.: Source truncation and exhaustion: insights from quantitative in situ TEM tensile testing. Nano Lett. **11**(9), 3816–3820 (2011)
12. Lu, Y., et al.: Cold welding of ultrathin gold nanowires. Nat. Nanotechnol. **5**(3), 218–224 (2010)
13. Guo, H., et al.: Tensile ductility and necking of metallic glass. Nat. Mater. **6**(10), 735–739 (2007)
14. Mitra, R., Chiou, W.A., Weertman, J.R.: In situ study of deformation mechanisms in sputtered free-standing nanocrystalline nickel films. J. Mater. Res. **19**(4), 1029–1037 (2004)
15. Haque, M.A., Saif, M.T.A.: Mechanical behavior of 30–50 nm thick aluminum films under uniaxial tension. Scr. Mater. **47**(12), 863–867 (2002)
16. Zhang, P., et al.: Fracture toughness of graphene. Nat. Commun. **5** (2014)
17. Uchic, M.D., et al.: Sample dimensions influence strength and crystal plasticity. Science **305**(5686), 986–989 (2004)
18. Yu, Q., et al.: Strong crystal size effect on deformation twinning. Nature **463**(7279), 335–338 (2010)
19. Deneen, J., et al.: In situ deformation of silicon nanospheres. J. Mater. Sci. **41**(14), 4477–4483 (2006)
20. Minor, A.M., et al.: A new view of the onset of plasticity during the nanoindentation of aluminum. Nat. Mater. **5**, 697–702 (2006)
21. Kizuka, T., et al.: Simultaneous observation of millisecond dynamics in atomistic structure, force and conductance on the basis of transmission electron microscopy. Japn. J. Appl. Phys. Part 2-Lett. **40**(2B), L170–L173 (2001)
22. Merkle, A.P., Marks, L.D.: Friction in full view. Appl. Phys. Lett. **90**(6), 064101 (2007)
23. Tang, D.M., et al.: Nanomechanical cleavage of molybdenum disulphide atomic layers. Nat. Commun. **5**, 3631 (2014)
24. Browning, N.D., et al.: Recent developments in dynamic transmission electron microscopy. Curr. Opin. Solid State Mater. Sci. **16**(1), 23–30 (2012)

In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles

Meghan E. Samberg[1], Zhoumeng Lin[2] and Nancy A. Monteiro-Riviere[3]

[1]Arteriocyte, Inc., Cleveland, OH, USA

[2]Institute of Computational Comparative Medicine, Department of Anatomy and Physiology, Kansas State University, Manhattan, KS, USA

[3]Nanotechnology Innovation Center of Kansas State, Department of Anatomy and Physiology, Kansas State University, Manhattan, KS, USA

Synonyms

Biodistribution; Nanoparticles; PBPK modeling; Pharmacokinetics; Silver (Ag); Toxicity

Definition

Silver nanoparticles (Ag-nps): Any shaped particle that has silver as its majority constituent and one dimension ranging from 1 to 100 nm; its surface may be modified with intricate or conjugated complexes.

Nanotechnology has enabled the incorporation of silver nanoparticles (Ag-nps) into hundreds of products used as antimicrobial agents in both professional and personal healthcare fields. The quantity of Ag-nps produced annually is rapidly

increasing and fueled by the growing markets for products that incorporate them. This increased usage will lead to the accumulation of Ag-nps in the air, water, soil, and within organisms. Additionally, possible human and environmental toxicological risks associated with Ag-np exposure may arise during material fabrication, handling, usage, and waste disposal. For example, up to 100 % of the Ag-nps incorporated into socks to kill the bacteria associated with foot odor may leak out during washings [1]. These released Ag-nps have the potential to harm helpful bacteria in wastewater treatment facilities or endanger aquatic organisms in lakes and streams, and hand washings increase the risks associated with human interactions. Consequently, due to the increasing contact between these products and the human body, it is important to determine the potential risks Ag-nps pose in order to adequately impose policies and regulations on their product inclusion.

In the United States, the Environmental Protection Agency (EPA) regulates all self-identified antimicrobials under the Federal Insecticide, Fungicide, and Rodenticide Act (FIFRA). However, regulation is currently challenged by insufficient characterization and standardized test methodology due to the highly varied synthesis methods for Ag-nps. In an effort to assess current hazards, exposure and risk information pertaining to nanomaterials including Ag-nps, the European Union (EU) has established the project "Engineered Nanoparticles: Review of Health and Environmental Safety (ENRHES)." Regarding Ag-nps, the goal of both of these agencies is to consider the various sources, pathways of exposure, and impacts on health and the environment in order to generate information and data suitable for risk assessment. With this in mind, the aim of this article is to provide an overview of Ag-np uses, potential health hazards, in vitro and in vivo studies, mechanism of toxicity, and pharmacokinetics.

Evolution of Silver Use

In its bulk form, silver (Ag) has been used for thousands of years in jewelry, currency, dental alloys, and photography and, most importantly, as an antimicrobial agent. Ancient Phoenicians were known to place silver coins into their water jugs as a preservative. Until the advent of penicillin, solutions of Ag nitrate formed the mainstay of antibiotics, specifically efficacious for preventing bacterial growth in open wounds and burns. Even today, Ag sulfadiazine creams remain the standard of care in burn wards to place on a patient's denuded skin for the prevention of antibacterial infections. Solutions of microparticle colloidal solutions have been incorporated into water purification filters. At the nanoscale, the technologically important and unique magnetic, electrical, and optical properties of Ag have given rise to products for targeted drug delivery, diagnosis, detection, and imaging. However, the enhanced antibacterial activity of Ag at the nanoscale has been most valuable in the professional and personal healthcare fields through the incorporation of Ag-nps into hundreds of products, including surgical and food-handling tools, clothing, cosmetics, contact lens cases and toothbrushes, catheters, and even children's toys. In fact, according to the database of the Project on Emerging Nanotechnologies of the Woodrow Wilson International Center for Scholars, Ag-nps have consistently been identified as the most abundantly used material in all nanotechnology consumer products since 2006 and are still the most common in 2014 [2]. According to this database, the concentration of Ag-nps within consumer products from dietary supplements to cosmetics is in the range of <6–10,000 ppm. Currently, three trials on Ag-nps are registered as completed, and at least two are ongoing at the repository of clinical trials maintained at the National Institutes of Health (www.ClinicalTrials.gov). The completed studies have tested their benefits against the growth of microorganisms of a nanosilver-containing gel for hand cleaning (SilvaSorb®) and a central venous catheter embedded with nanosilver wall (AgTive®), as well as assessing the toxicity of ingested 10 and 32 ppm Ag-np solutions. Ongoing studies are focused on the incorporation of Ag-nps into fluoride dental rinses for the remineralization of teeth and reduction of dental biofilms in pediatric populations.

Silver Nanoparticles

As is the case for most nanoparticles, Ag-nps exhibit unique physicochemical properties that differ from, and provide distinct advantages over, those of their bulk metallic constituent. These augmented properties are likely due to the increased surface area to volume ratio at the nanoscale. A key reason for the ubiquitousness of Ag-nps is the ease of their generation which may be achieved by several different methods including spark discharging, electrochemical reduction, chemical reduction, solution irradiation, laser ablation, and cryochemical synthesis [3]. Depending on the method of synthesis, Ag-nps may take on a variety of different shapes such as spheres, rods, cubes, hollow tubes, multifacets, and films.

In its bulk metallic form, Ag is inert and exerts no biocidal action, but ionizes in the presence of water or tissue fluids to release antimicrobial Ag ions. Unlike other heavy metals, the speciation of Ag does not seem to play a large part of its toxicity; while Ag^{+} is highly reactive, the toxicity of Ag^{3+} or Ag^{2+} is unlikely due to the transient nature of these forms of Ag, particularly in an aqueous environment. Furthermore, there is no discernable difference between Ag ions released from nanoparticles and those generated by any other source such as through electrolysis or dissolution of ions from Ag salts. However, the rate of Ag ion production, as well as the distribution of Ag in tissue, may differ substantially between Ag-nps and other sources of Ag ions due to the ability of Ag-nps to release ions into specific tissues, cell membranes, or cells. Additionally, once a nanoparticle is located within a site, it has the ability for protracted release of Ag ions, thus increasing the potential toxic impact. Furthermore, agglomeration of multiple Ag-nps may change the rate of Ag ion offloading from their surface.

The driving force for Ag ion dissolution depends on the metal solubility within a given environment as well as the concentration gradient between the particle surface and the bulk solution phase. A nanoparticle with higher solubility such as ZnO may be more cytotoxic than one with low solubility, such as TiO_2 [4]. The rate of nanoparticle dissolution is considered to be proportional to the specific surface area, but additional factors such as surface curvature, roughness, nanoparticle aggregation, functionalization and surface modification, and the solution in which they are suspended may also be involved. Additionally, ionization of Ag or Ag compounds is enhanced by electric currents. In particular cases, it may be possible to moderate the dissolution of Ag-nps and their toxic effects via the coating on the surface [5]. Axiomatically, the functionalization and surface coating of Ag-nps will affect their endocytosis and exocytosis potentials and mechanisms, intracellular location, and agglomeration/aggregation status.

Antimicrobial Activity

Ag-nps are an effective killing agent against a broad spectrum of both Gram-positive and Gram-negative bacteria and also against antibiotic-resistant bacterial strains and a variety of fungi. Gram-negative bacterial species are responsible for nosocomial infections and include *Acinetobacter*, *Escherichia*, *Pseudomonas*, *Salmonella*, and *Vibrio.* Gram-positive bacterial strains include *Bacillus*, *Clostridium*, *Enterococcus*, *Listeria*, *Staphylococcus*, and *Streptococcus*. Antibiotic-resistant bacteria that have been killed by Ag-nps include strains such as *Enterococcus faecium* and methicillin-resistant and vancomycin-resistant *Staphylococcus aureus.*

The antimicrobial activity of Ag-nps may be explained by several mechanisms: (1) excessive binding of Ag ions and Ag-nps may prevent the uptake of essential nutrients to the cell, ultimately leading to cell death; (2) Ag ion entry into the cell by competitive binding with essential metals such as Ca^{2+}, Mg^{2+}, and Mn^{2+}; or (3) Ag transport and irreversible accumulation in the cell could occur by complexation with ligands or substrates. Once inside the cells, Ag ions can inhibit respiration or bind and condense DNA. Ultimately, the biocidal activity of Ag ions is likely caused by a synergistic effect between the binding of Ag ions to the cell wall, Ag uptake and subsequent accumulation in the cell, and Ag interference with critical biomolecules within the cell. The Ag ions released from Ag-nps, as well as the Ag-nps themselves, bind to the electron-donor receptors present on the

bacterial cell membranes such as disulfide, amino, imidazole, carbonyl, and phosphate. Inactivation of membrane-related enzymes like phosphomannose isomerase results in denaturation of the bacterial cell envelope and its functional capacity to regulate the inward diffusion of nutrients and limits the effusion of essential electrolytes and metabolites. The predominant intracellular effect of Ag probably lies in its ability to impair key intracellular enzyme systems by replacing trace metals and electrolytes, leading to defective respiratory pathways and RNA and DNA replication.

In a study by Samberg et al., Ag-nps of different sizes, surface conditions, and synthesis methods were evaluated in *Escherichia coli*, Ag-resistant *Escherichia coli*, *Staphylococcus aureus*, methicillin-resistant *Staphylococcus aureus* (MRSA), and *Salmonella* sp. [6]. The study yielded a range of minimal inhibitory concentrations and was based mainly on whether the Ag-nps solutions were contaminated by particle synthesis reducing agents. While freshly synthesized Ag-nps were toxic to all bacterial strains at concentrations between 3.0 and 8.0 ppm, the washed Ag-nps and the carbon-coated Ag-nps were toxic to all bacterial strains at concentrations between 64.0 and 1,024.0 ppm [6]. The results were supported by others that show size-dependent antimicrobial activity of Ag-nps with diameters ranging from 1 to 450 nm against both Gram-negative and Gram-positive bacterial strains [7]. In addition to size and concentration, shape-dependent antimicrobial activity of Ag-nps shows that between spherical, rod-shaped, and truncated triangular nanoplates, the truncated triangular Ag-nps displayed the strongest antibacterial activity [7]. The majority of the data suggests that most pathogenic organisms are killed in vitro at Ag ion concentrations of 10–40 ppm, with particularly sensitive organisms susceptible to 60 ppm Ag ions [8].

For maximal antimicrobial efficacy, it is therefore a critical function of Ag-nps to release a steady amount of Ag ions from their surface. Generally, the smaller the nanoparticle, the larger the surface area to volume ratio, which provides a more efficient means for antibacterial activity even at very low concentrations [9, 10].

In Vitro Toxicity

Ideally, the concentration of Ag-nps required to inhibit and eradicate microbial growth should be less than the concentration required for human cell cytotoxicity. The physicochemical and structural properties of Ag-nps play a major role in their interactions with cells, and variations in these properties among different Ag-nps may result in different toxicological effects. Therefore, it is strongly recommended that the physicochemical properties of Ag-nps, an important component of the study design, are reported in any risk assessment and health study. Manufactured Ag-nps have many physicochemical characteristics that include their size and size distribution, shape, surface area, porosity, zeta potential, solubility, crystallinity, purity, degree of agglomeration, surface functionalization (e.g., inorganic or organic coatings, etc.), surface chemical adsorption, and stability.

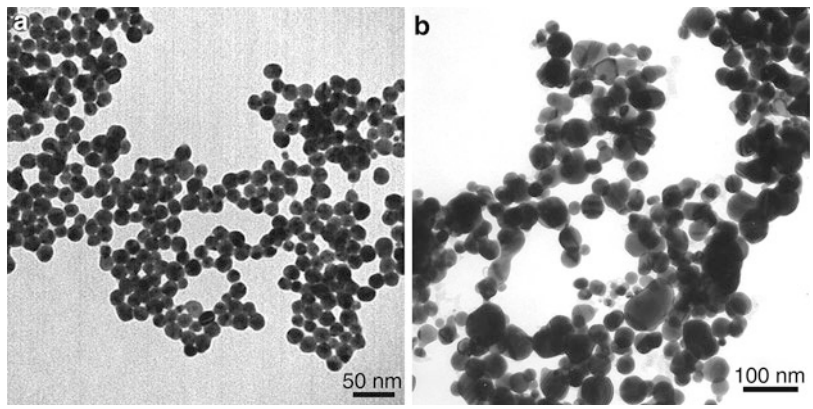
Several techniques that can be used to determine the size and size distributions of Ag-nps include atomic force microscopy (AFM), transmission electron microscopy (TEM), scanning electron microscopy (SEM), and dynamic light scattering (DLS). DLS may also be used to measure the zeta potential of Ag-nps. The size of Ag-nps is often reported from the nominal primary particle size as defined by the manufacturer. However, as was found in our study that investigated the cytotoxicity and inflammatory potential of Ag-nps of various sizes, synthesis methods, and surface conditions to human epidermal keratinocytes (HEK) in vitro and to porcine skin in vivo, as well as their antibacterial efficacy against a range of both Gram-positive and Gram-negative bacteria, the size of Ag-nps is not always correctly reported and may be observed in Table 1 [6, 11].

Due to various factors, Ag-nps tend to agglomerate in suspension and often have a larger hydrodynamic diameter when measured by DLS as compared to their physical diameter as measured by TEM. An example of a few of the Ag-nps investigated in previously mentioned studies may be observed in Fig. 1, and uniformity differences between Ag-nps of varying synthesis method may be noted [6, 11]. Therefore, it is

In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles, Table 1 Physicochemical properties of Ag-nps

Synthesis method	Dispersal conditions	MDD (nm)	DLS diameter (nm)[a]	TEM diameter (nm)[a]	Supplied conc. (mg/ml)	Particle conc. (particles/ml)	Zeta potential (mV)[a]
Reduction	Freshly synthesized, DI water[b]	20	30.8 ± 0.6	22.4 ± 2.6	0.2	2.41×10^{12}	−29.7
Reduction	Freshly synthesized, DI water[b]	50	47.7 ± 0.5	49.4 ± 6.2	0.2	4.44×10^{11}	−27.8
Reduction	Freshly synthesized, DI water[b]	80	75.5 ± 1.0	79.2 ± 8.0	0.2	7.09×10^{10}	−33.2
Reduction	Washed, 2 mM PB	20	25.5 ± 0.4	21.4 ± 3.1	2.86	1.89×10^{14}	−46.0
Reduction	Washed, 2 mM PB	50	43.7 ± 1.1	50.0 ± 5.9	3.45	5.01×10^{12}	−44.3
Reduction	Washed, 2 mM PB	80	79.9 ± 28.0	77.0 ± 6.0	2.79	1.07×10^{12}	−43.7
Laser ablation	Powder	25	149.0 ± 89	27.2 ± 10.3	N/A	N/A	−24.0
Laser ablation	Powder	35	167.0 ± 110	37.0 ± 11.6	N/A	N/A	−29.0

DI deionized water, *PB* phosphate buffer, *MDD* manufacturer-designated diameter, *DLS* dynamic light scattering, *TEM* transmission electron microscopy, *conc.* concentration, *N/A* not applicable
[a]Data are expressed as mean±SD, in DI water
[b]5.55 mg/ml formaldehyde present in solution



In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles, Fig. 1 Transmission electron micrographs of Ag-nps: (**a**) 20 nm washed Ag-nps; (**b**) 25 nm carbon-coated Ag-nps

important to use multiple techniques to fully characterize Ag-nps prior to their testing. The Brunauer, Emmett, and Teller (BET) technique is useful in characterization by measuring the surface area and porosity of Ag-nps through the adsorption of gas molecules onto the surface of the nanoparticle.

The most commonly used method to assess the cytotoxicity of Ag-nps is by cell viability. In general, the percentage of dead cells is directly proportional to the toxicity of Ag-nps. Cell viability assays are typically chemical based and depend on differential inclusion, exclusion, or conversion of a dye or dye precursor which can only be

enzymatically converted to detectable dye in living cells but not dead cells. In this case, the toxicity can be evaluated by quantitating colorimetric or fluorescent dyes that include MTT, MTS, 96 AQueous One (96AQ), alamarBlue, lactate dehydrogenase (LDH), Live/Dead, and neutral red assays. Since nanoparticles are capable of adsorbing chemicals onto their surface, they may interact and interfere with the ability of the assay dye to react appropriately and produce false results [12, 13]. Ag-nps above 1 ppm of various sizes and surface conditions were shown to interfere with MTT, 96AQ, and alamarBlue assays, with alamarBlue having the least interaction. Therefore, prior to initiating a study and selecting a particular assay as a viability biomarker, it is important to investigate any potential interactions between the nanoparticle and assay dye.

Presently, there is no consensus on the cytotoxicity of Ag-nps; however, the majority of publications do show reduced cell viability following Ag-np exposure. The variability of the cytotoxicity is linked to several characteristics, such as the synthesis method and purity of the Ag-nps, the concentration of released Ag ions from the nanoparticle, and the physicochemical properties of the nanoparticles such as size, shape, and surface functionalization. In vitro cell studies have demonstrated size-dependent Ag-np cytotoxicity and the generation of reactive oxygen species (ROS) and pro-inflammatory cytokines to various Ag-nps in a variety of human cell lines such as colon carcinoma, fibrosarcoma, skin carcinoma, lung fibroblast, glioblastoma, hepatoma, alveolar, and keratinocyte [7, 11, 14, 15]. Confocal laser scanning microscopy measurements confirmed what has been observed in many studies, that smaller Ag-nps ($\leq$20 nm) easily penetrated the cells causing significantly more damage, whereas larger Ag-nps ($\geq$40 nm) were only rarely internalized, resulting in lower toxicity [15]. Ag-nps of 20 nm were shown to be toxic to human hepatoblastoma C3A cells, following exposure to 2.5 $\mu g/cm^2$ for 24 h [16]. At sublethal concentrations, Ag-nps affected hepatocyte homeostasis by reducing albumin release and increasing mRNA expression of IL-8, IL-1RI, TNF-α, and intercellular adhesion molecule 1 (ICAM-1). Exposure of 15 and 100 nm Ag-nps to rat liver cells for 24 h at 5–10 ppm resulted in a significant decrease in mitochondrial function and membrane integrity. Additional toxic responses seen in vitro are glutathione depletion, mitochondrial changes, and destruction of cell membranes. However, Ag-np cytotoxicity does not always correlate to other stress-related effects. Contrary to other studies, exposure of human peripheral blood mononuclear cells to 1–2.5 nm diameter Ag-nps at $\geq$15 ppm for 72 h resulted in a decrease in proliferation but suppression of IFN-γ, TNF-α, and IL-5 [17].

Contaminants, stabilizers, or residual impurities created during the synthesis of Ag-nps, on their surface or in their surrounding solution, can significantly influence their cytotoxicity. Impurities may be the major contributor to the toxicity of the Ag-nps, which may cause misleading assumptions regarding the actual toxicity of the Ag-nps. Therefore, the purity of the Ag-nps should be evaluated using inductively coupled plasma-mass spectroscopy (ICP-MS) or high-performance liquid chromatography (HPLC). In the previously mentioned study by Samberg et al., freshly synthesized Ag-nps were shown to contain formaldehyde at 5.55 mg/ml in their solution and were solely responsible for toxicity in HEK [11]. In addition, an amorphous carbon matrix that effectively coated the Ag-nps was found to alleviate these cytotoxic effects.

The mechanism responsible for eukaryotic cell toxicity to Ag-nps is similar to the mechanism responsible for its antibacterial effects. It is likely to involve bound forms of oxidized Ag-nps on the surface of a target cell, the uptake and release of high concentrations of Ag ions into the cell and within cellular compartments, the disruption of integral cellular processes through catalytic or physical processes (such as protein or enzyme binding and subsequent denaturation), and the generation of ROS and induced oxidative stress. Under normal conditions, mitochondria generate ROS at low frequencies which are easily neutralized by antioxidants such as glutathione. The dissolution of Ag ions from the Ag-np surface is accompanied by the generation of ROS by disruption of the electronic and ionic flux, perturbation

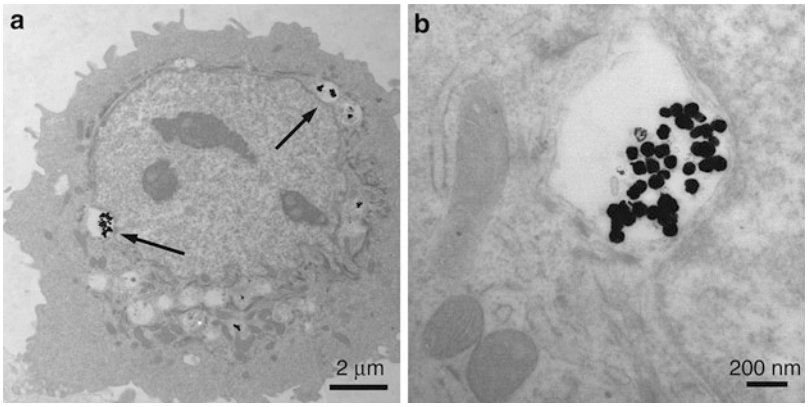
of the permeability transition pores, and depletion of glutathione. A state of oxidative stress may occur and result in adverse biological consequences such as membrane and DNA damage, protein denaturation, and apoptosis. The Ag cations released from Ag-nps bind strongly to electron-donor groups of biological molecules containing sulfur, oxygen, and nitrogen. The Ag ion absorbed into cells induces synthesis of cysteine-rich metal-binding protein metallothionein. Ag avidly binds these proteins to form stable complexes. Increased metallothionein favors the uptake of the key trace elements zinc and copper which in turn promote RNA and DNA synthesis, cell proliferation, and tissue repair. Metallothionein thus provides a double action: to protect tissues against the potential toxic effects of a xenobiotic metal and to promote healing.

To define the mechanism of Ag-nps toxicity, the cellular uptake and subcellular localization of Ag-nps must be understood. The cytotoxicity may be caused by Ag-np cellular uptake and the subsequent interaction of Ag-nps with individual cellular components, which leads to cell malfunction and ultimately death. Because Ag-nps are electron dense, their localization within cells is easily observed using TEM without secondary staining. The cellular uptake of Ag-nps has been shown in many different cell types, with endocytosis as the primary mechanism for cellular uptake. Our study investigated the cytotoxicity, inflammatory potential, and cellular uptake into HEK following 24 h exposure to 25 and 35 nm carbon-coated Ag-nps, as well as to either freshly synthesized or thoroughly washed 20, 50, and 80 nm Ag-nps at concentrations ranging from 0.00054 to 1.7 ppm [11]. Although cytotoxicity was only observed for the freshly synthesized Ag-nps, small agglomerates of each of the Ag-nps were observed within cytoplasmic vacuoles. Another study showed that sets of naked or silica-coated Ag-nps ranging in size from 20 to 120 nm in diameter acquired different protein coronas depending on the serum proteins they were exposed to [18]. These coronas affected their uptake into HEK compared to control Ag-nps not exposed to serum proteins and bring to question how in vitro assays using naked Ag-nps should be interpreted relative to what would be expected to occur in vivo. An example of the endocytosis of 80 nm washed Ag-nps is seen in Fig. 2.

In Vivo Toxicity

Although in vitro toxicity studies are an important first step toward assessing the potential risks associated with Ag-np exposure, the test conditions in vitro do not correspond to the actual conditions within an organism and may lead to results that do not reflect the interactions that will arise in vivo. In vivo studies allow for the investigation of long-term effects of Ag-np exposure, systemic and local distribution in serum and tissues, retention, excretion, and any histological changes in tissues. Therefore, these studies provide more relevant information regarding Ag-np toxicity in humans that is not available through in vitro studies.

The wide variety of uses of Ag-nps allows exposure through various routes of entry into the body. Due to their size, Ag-nps have the potential to impact the respiratory tract, gastrointestinal tract, and skin and enter the systemic circulation or even the central nervous system. Target cell and organ uptake of various nanoparticles have been investigated recently, and it is indicated that several factors influence the ability of Ag-nps to produce toxic effects in vivo, such as their solubility, their ability to bind to biological sites, and the degree to which complexes are formed and sequestered or metabolized and excreted [9]. Theoretically, Ag can be deposited in any tissue of the human body, but the skin, brain, liver, kidneys, eyes, and bone marrow have received the greatest attention. Ag absorbed into human tissues from antiseptic respiratory sprays, implanted medical devices, wound dressings, or indwelling catheters can reach systemic circulation, primarily as a protein complex. The most common health effects associated with chronic exposure to Ag are a permanent gray or blue-gray discoloration of the skin known as argyria, found primarily in sun-exposed areas. Biochemical transformations in the body can affect Ag bioavailability, toxicity, and risk. Although chronic Ag ingestion produces argyrial deposits in the skin, it has been suggested that they are secondary particles and not translocated primary particles. For centuries, the prevailing view is that the use of Ag compounds



In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles, Fig. 2 Transmission electron micrographs depicting cellular uptake of Ag-nps into HEK: (**a**) 80 nm washed Ag-nps within cytoplasmic vacuoles of an HEK; (**b**) higher magnification of the 80 nm washed Ag-nps within the cytoplasmic vacuole. *Arrows* point to Ag-nps

and Ag-nps in products is safe, because Ag in any form has not shown toxicity or carcinogenicity to the immune, cardiovascular, nervous, or reproductive systems.

Ingestion is one of the primary routes for entry for Ag compounds, colloidal Ag, and Ag-nps. Ag-nps are routinely ingested as dietary supplements for the homeopathic treatment of allergies, arthritis, chronic fatigue, colitis, dermatitis, diabetes, gastrointestinal distress reliever, Lyme disease, pneumonia, shingles, skin cancer, sore throat, syphilis, tonsillitis, tuberculosis, and warts. Normal dietary intake of Ag from drinking water and trace amounts in foods is estimated at 70–90 µg/day. Additionally, Ag-nps may be introduced into the gastrointestinal tract after being shed from coatings on food-handling machinery or directly from Ag-nps sprayed onto produce for the prevention of food-spoiling bacteria [19]. The homeopathic ingestion of Ag-nps has led to various in vivo studies; generally, exposure to Ag and Ag-nps over time leads to distribution in various tissues without causing toxicity. Oral exposure to 10 ppm (5–10 nm, 100 µg/day) or 32 ppm (25–40 nm, 480 µg/day) of commercial Ag-np solution for 3–14 days did not cause clinical changes in human metabolic profiles [20]. Chronic symptoms from prolonged intake of low doses of Ag salts are fatty degeneration of the liver and kidneys as well as changes in blood cells [7]. IV injection in rats with 4, 10, 20, and 40 mg/kg of Ag found that the highest concentration of 40 mg/kg caused significant increases in liver enzymes and blood serum ROS levels and Ag-np deposition in the liver and kidney, concluding that a dose of >20 mg/kg was toxic to rats [21].

The respiratory system is a major portal of entry for Ag-nps. When inhaled, Ag-nps may be deposited in any region of the respiratory system, be taken into cells, enter the bloodstream, and circulate to potentially sensitive target organs. Inhalation of dusts or fumes containing Ag occurs primarily in occupational settings, such as industrial plants involved in Ag chemical manufacturing, jewelry manufacturing, Ag reclamation, and production of tableware. Due to the small diameter of Ag-nps, Brownian diffusion results in deep penetration of Ag-nps into the lungs and diffusion to the high lung surface area presented in the alveolar region. Several 28-day and 90-day inhalation studies with rats have demonstrated that inhalation of Ag-nps of approximately 15 nm resulted in slight increases in neutral mucins, with some hepatic necrosis and Ag-np accumulation in the lung and liver [7]. Oropharyngeal aspiration exposure to pristine Ag-nps at 20 µg once a week for 5 weeks showed an increase in neutrophils in the bronchoalveolar lavage fluid of mice. However, no adverse effects were found in mice

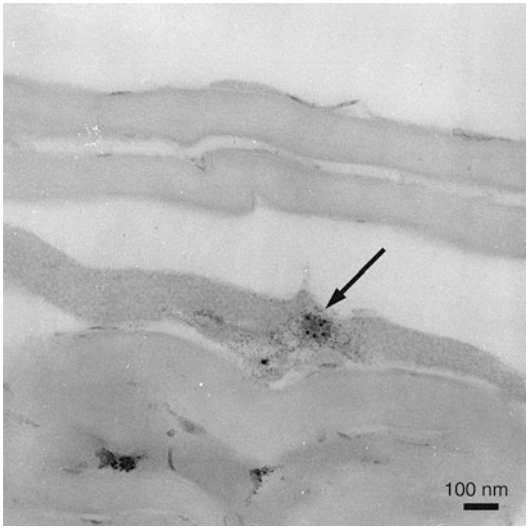
exposed to the same amount of Ag-nps embedded in a paint matrix [22]. Inhalation of 15 nm Ag-nps (179 μg/m^3) 6 h/day for 4 days in rats caused a 175-fold increase in neutrophils, an increase in cellular markers, a fivefold increase in cytokines, and a 1.5-fold increase in glutathione in the lungs 24 h postexposure, all of which disappeared by 7 days postexposure. In contrast, no adverse effects were observed after exposure to a similar mass of 410 nm Ag-nps (179 μg/m^3), suggesting a clear size-dependent effect after inhalation of Ag-nps [23].

Another major route of exposure is the skin. Skin exposure to Ag-nps is particularly high due to their increasing inclusion into textiles, burn wound creams, jewelry, acupuncture needles, catheters, coatings onto high-contact surface areas, and disinfectant sprays. Although Ag compounds have long been used in wound dressings, the incorporation of Ag-nps into wound dressings is on the rise. The release of Ag ions from the wound dressings is generally triggered by wound fluids and tissue exudates. While it has been deemed that 10–40 ppm is required for appropriate antimicrobial action, the Ag content of dressings currently available varies from $<$0.1 mg/cm^2 to more than 1 mg/cm^2 with the amount of Ag ions released ranging from $<$1 to $>$70 ppm. Several studies have evaluated the toxicity of Ag from Ag-np-containing wound dressings and have shown elevated serum and liver enzyme levels as well as argyria [7]. However, absorption through intact skin is low ($<$1 ppm) since much of the free ion is precipitated as Ag sulfide in the superficial layers of the stratum corneum. Daily repetitive dosing for 14 days with both freshly synthesized and thoroughly washed 20 and 50 nm Ag-nps topically applied onto the backs of pigs showed Ag-np penetration only in the superficial layers of the stratum corneum [11]. Ultrastructural studies with TEM and confirmed by EDS detected Ag within the stratum corneum as observed in Fig. 3 [11].

When Ag-nps pass the barriers of the lung, intestine, or skin to reach systemic circulation, further tissue distribution may occur. Once Ag-nps reach systemic circulation and are exposed to biological fluids, they adsorb onto their surface amino acids, proteins, sugars, or salts within biological environments due to the surface reactivity of Ag-nps and may also undergo biochemical transformations. The adsorption of molecules onto the surface of Ag-nps influences cellular interactions and may also eliminate toxic potency of Ag-nps. Ag absorbed from the intestines accumulates in a transitory fashion in the liver, kidneys, brain, lungs, and bone marrow with minimal to no risk. Ag absorption from the gastrointestinal tract or from antiseptic sprays can accumulate in the cornea of the eye and skin to present as argyria in these tissues upon exposure to UV radiation. Argyrosis and argyria attributable to fine deposits of Ag metal or Ag sulfide are permanent, but not life threatening. Additionally, night vision problems, bowl pain, respiratory trouble, reduced creatinine clearance, and tarnished corneas have been associated with workers exposed to Ag dust [24]. In the liver and kidneys, Ag granules accumulate in the cytoplasm of phagocytic cells, hepatocytes, and renal tubular epithelium bound mostly in lysosomes. Ag is released into bile ducts and urinary ducts for excretion in the feces and urine, respectively. Greater concern relates to the influence of chronic Ag ingestion or Ag sulfadiazine therapy on the bone marrow and circulating neutrophils. For therapeutic applications of Ag-nps targeted to brain disorders, Ag-nps must be able to cross the blood-brain barrier (BBB) without harming the integrity of the brain. Several studies have been conducted using murine models and have shown localized edema to the brain 24 h after Ag-np systemic exposure, indicating the need for further research [24].

Pharmacokinetics

The potential and degree of in vivo toxicity of Ag-nps primarily depend on their pharmacokinetic characteristics and target organ dosimetry [25, 26]. The pharmacokinetics of Ag-nps has been shown to be dose, exposure route, species, and gender dependent [26]. For example, citrate-coated 7.9 nm Ag-nps showed a plasma half-life of 99 h (1 mg/kg) after IV injection and 30 h (10 mg/kg) for oral exposure in rats [27], whereas in rabbits after IV administration, the half-lives



In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles, Fig. 3 Transmission electron micrograph depicting 20 nm washed Ag-nps only in the superficial layers of the stratum corneum following a 14-day repetitive in vivo topical dosing. *Arrow* points to Ag-nps

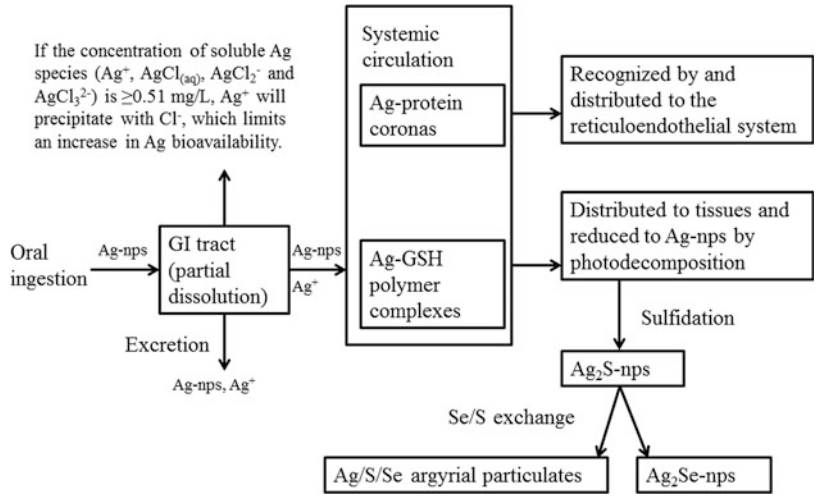
were 11.7 and 16.3 days in the 5 and 0.5 mg/kg groups, respectively [28]. The fact that the half-life of citrate-coated 7.9 nm Ag-nps is three to four times shorter in rats than in rabbits could be explained, in part, by the longer transit time in the systemic circulation of larger animals than in smaller rodents, which modulates the extent of nanoparticle biocorona formation, which is supported by pharmacokinetic models [29, 30]. Interestingly, following IV injection of 120 mg/kg 90.5 nm Ag-np agglomerates, the half-life was 15.6 h in male mice and 29.9 h in females, indicating a sex difference [31].

While exposure routes for Ag-nps may include oral, dermal, and inhalational exposures, it should be noted that the bioavailability is generally low and depends on the particle size, dose, surface coating, and soluble fraction. Following a single oral exposure in rats, the bioavailability of citrate-coated 7.9 nm Ag-nps was 1.2 % and 4.2 % in the 1 and 10 mg/kg, respectively [27]. For dermal exposure, a study showed skin penetration of 25 nm Ag-nps in both intact and damaged in vitro human skin in static diffusion cells but with a low flux compared to the absorption rate for metal powders. TEM confirmed the presence of Ag-nps in the stratum corneum and outermost surface of the epidermis but not in the dermis suggesting a small fraction of the Ag-nps had dissolved and diffused through the skin layers as elemental Ag [32]. The metabolic pathways for Ag-nps following oral exposure are depicted in Fig. 4 [26, 33]. Briefly, in the stomach, Ag-nps can be ionized to form Ag^+, but the dissolution is incomplete due to the limited gastric residence time (10–240 min). Ag^+ and other soluble complexes [AgCl(aq), $AgCl_2^-$, and $AgCl_3^{2-}$] can be absorbed by the gastrointestinal tract into the systemic circulation, where Ag^+ can bind to proteins with thiol groups such as serum albumin and small thiol molecules like glutathione (GSH). These complexes generate H^+ and GS-Ag which then produce Ag-GSH complexes which then distribute throughout the body. Ag-thiol complexes are reduced to Ag-nps upon UV photodecomposition or at a slower rate with visible light. Besides thiols, Ag-nps can be sulfidated to form Ag_2S-nps, which can react with selenium (Se) to generate Ag_2Se-nps and Ag/S/Se argyrial secondary particulates and not the primary particles [33].

As previously mentioned, the biodistribution of Ag-nps is size dependent and primarily to the liver, followed by the kidneys, spleen, and other organs, regardless of exposure routes [26]. In the liver, Ag-np deposition has been detected in several cell types, including Kupffer cells, hepatocytes, and sinusoidal endothelium cells [16]. In the kidney, Ag-nps were observed in all regions of the cortex, medulla, inner medulla, and cortical glomeruli [34]. For Ag-nps larger than 80 nm, the distribution to the spleen has been shown to be high [31, 35]. Ag-nps up to 110 nm could be distributed to and persistent in the brain for 2 months following oral exposure [36].

Available evidence suggests that the biliary pathway is the major elimination route for Ag-nps. Following a single IV injection of citrate-coated 7.9 nm Ag-nps in rabbits, the 7-day cumulative fecal excretion amounts were 8.3 % and 16.7 % of the injected dose in the 5 and 0.5 mg/kg groups, respectively, whereas only <0.05 % of the injected dose was excreted via the urine for both dose groups [28].

I



In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles, Fig. 4 Proposed metabolism pathways of Ag-nps in the body after oral exposure. *GI* gastrointestinal, *GSH* glutathione, *S* sulfur, *Se* selenium

Physiologically Based Pharmacokinetic Models

One approach that can be used to correlate the external exposure dose and observed toxicity with pharmacokinetics and target organ dosimetry, thereby establishing a dose-response relationship for Ag-nps, is physiologically based pharmacokinetic (PBPK) modeling. PBPK modeling is a computational technique that simulates the absorption, distribution, metabolism, and elimination (ADME) of chemicals, drugs, or nanoparticles in the body using mathematical equations. The advantages of PBPK models include their exposure dose-, route-, and species-extrapolation power and the ability to predict target organ concentrations. In addition, PBPK models allow the inclusion of species-dependent and nanoparticle-specific mechanistic information to more accurately simulate the pharmacokinetics. This is particularly important for nanoparticles because the pharmacokinetics of nanoparticles could be affected by multiple factors, including size, surface coating, surface charge, protein binding, exposure route, dose, and species.

Currently, there are two PBPK models for Ag-nps [35, 37]. Lankveld et al. developed the first PBPK model for 20, 80, and 110 nm Ag-nps in rats following daily IV injections for 5 days [35]. The model was composed of five compartments: blood, liver, spleen, kidneys, and remaining tissues. Distribution of Ag-nps in the liver, kidney, and spleen was described by two state variables: one is freely available for exchange between tissues and blood, and the other is irreversibly bound to tissues. Model simulations were well correlated with experimental data used for model calibration. No apparent relationship was found between key parameter values and corresponding nanoparticle diameters, suggesting that not only the size but other properties such as surface chemistry and protein absorption may be involved. Although this model was not validated with external data, it provides a valuable foundation for subsequent PBPK model development.

Bachler et al. reported a comprehensive and validated PBPK model for ionic Ag and 15–150 nm Ag-nps [37]. It consisted of 12 compartments (blood, brain, spleen, bone marrow, testes, muscles, heart, lungs, liver, skin, kidneys, and intestines) and simulated 4 exposure routes (IV, dermal, oral, and inhalation). The model described uptake of Ag-nps from the blood to the organs using transfer functions based on organ-specific mechanisms (i.e., the GSH concentration, capillary wall type, and phagocytosis efficiency) and used a same release rate to describe the transport from the tissue to the blood for all organs, except the brain and testes, for which smaller uptake and release rates were used. The model adequately predicted most of available pharmacokinetic data for Ag-nps, and model simulations offered valuable mechanistic insights into

the pharmacokinetics of Ag-nps across various exposure scenarios. Notably, this model was successfully extrapolated from rats to humans and was used to predict internal exposures based on five different human exposure scenarios. The predicted results were shown to be helpful for exposure and risk assessment of Ag-nps in humans. While being a useful model, it was built based upon a number of assumptions and there remained larger uncertainties. As proposed by Bachler et al., additional experimental studies are still needed to further refine this model [37].

The application of nanoparticles in personal and healthcare fields has the potential to have a positive effect on human wellness. Accordingly, Ag-nps are becoming widely used components in medical devices, healthcare products, and personal care items to provide antibacterial and antifungal action. While experience has shown that metallic Ag and Ag compounds are generally safe in use and effective in controlling pathogenic organisms, Ag-nps are capable of unprecedented entry into cellular compartments with protracted release of Ag ions. This ability makes Ag-nps particularly useful in eliminating harmful microbes, but also potentially harmful to humans. While some in vitro and in vivo studies have shown toxicity related to Ag-nps, others have also shown an overall lack of toxicity, and this may be due to the synthesis method, dose, exposure paradigm, various physicochemical properties of particular Ag-nps, in vivo pharmacokinetics and target organ dosimetry, species, or even type of assay used to assess the toxicity. Therefore, it is important for future studies to determine the physicochemical properties of Ag-nps that make them toxic. Additionally, because toxicity of Ag-nps is highly correlated with their pharmacokinetics, it is critical to further investigate the pharmacokinetics of Ag-nps and the influences of various physicochemical properties. PBPK models of Ag-nps have shown to be helpful in predicting target organ concentrations following a variety of exposure scenarios and in human risk assessment. New PBPK models with fewer modeling assumptions and less uncertainty in model parameters remain to be developed. Manufacturers of new products incorporating Ag-nps should envisage providing a balance between Ag ions released for antimicrobial purposes while maintaining a minimal human toxicity threshold.

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ Nanoparticle Cytotoxicity

References

1. Benn, T.M., Westerhoff, P.: Nanoparticle silver released into water from commercially available sock fabrics. Environ. Sci. Technol. **42**(11), 4133–4139 (2008)
2. Project on Emerging Nanotechnologies. Consumer Products Inventory. Retrieved 17 Nov 2014, http://www.nanotechproject.org/cpi (2014)
3. Chen, X., Schluesener, H.J.: Nanosilver: a nanoproduct in medical application. Toxicol. Lett. **176**, 1–12 (2008)
4. Brunner, T.J., Wick, P., Manser, P., Spohn, P., Grass, R.N., Limbach, L.K., Bruinink, A., Stark, W.J.: In vitro cytotoxicity of oxide nanoparticles: comparison to asbestos, silica, and the effect of particle solubility. Environ. Sci. Technol. **40**(14), 4374–4381 (2006)
5. Monteiro-Riviere, N.A., Orsière, T.: Toxicological impacts of nanomaterials. In: Wiesner, M.R., Bottero, J.-Y. (eds.) Environmental Nanotechnology – Applications and Impacts of Nano-Materials, pp. 395–444. McGraw Hill, New York (2007)
6. Samberg, M.E., Orndorff, P.E., Monteiro-Riviere, N. A.: Antibacterial efficacy of silver nanoparticles of different sizes, surface conditions and synthesis methods. Nanotoxicology **5**(2), 244–253 (2011)
7. Wijnhoven, S.W.P., Peijnenburg, W.J.G.M., Herberts, C.A., Hagens, W.I., Oomen, A.G., Heugens, E.H.W., Roszek, B., Bisschops, J., Gosens, I., Van De Meent, D., Dekkers, S., De Jong, W.H., Van Zijverden, M., Sips, A.J.A.M., Geertsma, R.E.: Nano-silver: a review of available data and knowledge gaps in human and environmental risk assessment. Nanotoxicology **3**(2), 109–138 (2009)
8. Lansdown, A.B..: Silver in health care: antimicrobial effects and safety in use. Curr. Probl. Dermatol. **33**, 17–34 (2006)

I

9. Monteiro-Riviere, N.A., Tran, C.L.: Nanotoxicology-Characterization, Dosing and Health Effects. Informa Healthcare, New York (2007)
10. Monteiro-Riviere, N.A., Tran, C.L.: Nanotoxicology-Progress Toward Nanomedicine. CRC Press, Taylor and Francis, Boca Raton (2014)
11. Samberg, M.E., Oldenburg, S.J., Monteiro-Riviere, N. A.: Evaluation of silver nanoparticle toxicity in skin in vivo and keratinocytes in vitro. Environ. Health Perspect. **118**(3), 407–413 (2010)
12. Monteiro-Riviere, N.A., Inman, A.O., Zhang, L.W.: Limitations and relative utility of screening assays to assess engineered nanoparticle toxicity in a human cell line. Toxicol. Appl. Pharmacol. **234**(2), 222–235 (2009)
13. Monteiro-Riviere, N.A., Inman, A.O.: Challenges for assessing carbon nanomaterial toxicity to the skin. Carbon **44**, 1070–1078 (2006)
14. Christensen, F.M., Johnston, H.J., Stone, V., Aitken, R.J., Hankin, S., Peters, S., Aschberger, K.: Nano-silver – feasibility and challenges for human health risk assessment based on open literature. Nanotoxicology **4**, 1–12 (2010)
15. Miethling-Graff, R., Rumpker, R., Richter, M., Verano-Braga, T., Kjeldsen, F., Brewer, J., Hoyland, J., Rubahn, H.G., Erdmann, H.: Exposure to silver nanoparticles induces size- and dose-dependent oxidative stress and cytotoxicity in human colon carcinoma cells. Toxicol. In Vitro **28**, 1280–1289 (2014)
16. Gaiser, B.K., Hirn, S., Kermanizadeh, A., Kanase, N., Fytianos, K., Wenk, A., Haberl, N., Brunelli, A., Kreyling, W.G., Stone, V.: Effects of silver nanoparticles on the liver and hepatocytes in vitro. Toxicol. Sci. **131**, 537–547 (2013)
17. Shin, S.H., Ye, M.K., Kim, H.S., Kang, H.S.: The effects of nanosilver on the proliferation and cytokine expression by peripheral blood mononuclear cells. Int. Immunopharmacol. **7**, 1813–1818 (2007)
18. Monteiro-Riviere, N.A., Samberg, M.E., Oldenburg, S.J., Riviere, J.E.: Protein binding modulates the cellular uptake of silver nanoparticles into human cells: implications for in vitro to in vivo extrapolations? Tox. Lett. **220**, 286–293 (2013)
19. Silver, S.: Bacterial silver resistance: molecular biology and uses and misuses of silver compounds. FEMS Microbiol. Rev. **27**, 341–353 (2003)
20. Munger, M.A., Radwanski, P., Hadlock, G.C., Stoddard, G., Shaaban, A., Falconer, J., Grainger, D.W., Deering-Rice, C.E.: In vivo human time-exposure study of orally dosed commercial silver nanoparticles. Nanomedicine **10**, 1–9 (2014)
21. Tiwari, D.K., Jin, T., Behari, J.: Dose-dependent in-vivo toxicity assessment of silver nanoparticle in Wistar rats. Toxicol. Mech. Methods **21**(1), 13–24 (2011)
22. Smulders, S., Luyts, K., Brabants, G., Landuyt, K.V., Kirschhock, C., Smolders, E., Golanski, L., Vanoirbeek, J., Hoet, P.H.: Toxicity of nanoparticles embedded in paints compared with pristine nanoparticles in mice. Toxicol. Sci. **141**, 132–140 (2014)
23. Braakhuis, H.M., Gosens, I., Krystek, P., Boere, J., Cassee, F.R., Fokkens, P., Post, J., van Loveren, H., Park, M.: Particle size dependent deposition and pulmonary inflammation after short-term inhalation of silver nanoparticles. Part. Fibre. Toxicol. **11**, 49 (2014)
24. Arvizo, R.R., Bhattacharyya, S., Kudgus, R., Giri, K., Bhattacharya, R., Mukherjee, P.: Intrinsic therapeutic applications of noble metal nanoparticles: past present and future. Chem. Soc. Rev. **41**(7), 2943–2970 (2012)
25. Riviere, J.E.: Pharmacokinetics of nanomaterials: an overview of carbon nanotubes, fullerenes and quantum dots. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. **1**, 26–34 (2009)
26. Lin, Z., Monteiro-Riviere,N.A., Riviere, J.E.: Pharmacokinetics of metallic nanoparticles. WIREs Nanomed. Nanobiotechnol 7, 189–217 (2015)
27. Park, K., Park, E.J., Chun, I.K., Choi, K., Lee, S.H., Yoon, J., Lee, B.C.: Bioavailability and toxicokinetics of citrate-coated silver nanoparticles in rats. Arch. Pharm. Res. **34**, 153–158 (2011)
28. Lee, Y., Kim, P., Yoon, J., Lee, B., Choi, K., Kil, K.H., Park, K.: Serum kinetics, distribution and excretion of silver in rabbits following 28 days after a single intravenous injection of silver nanoparticles. Nanotoxicology **7**, 1120–1130 (2013)
29. Riviere, J.E.: Of mice, men and nanoparticle biocoronas: are in vitro to in vivo correlations and interspecies extrapolations realistic? Nanomedicine (Lond) **8**, 1357–1359 (2013)
30. Sahneh, F.D., Scoglio, C.M., Monteiro-Riviere, N.A., Riviere, J.E.: Predicting the impact of biocorona formation kinetics on interspecies extrapolations of nanoparticle biodistribution modeling. Nanomedicine (Lond) 10, 25–33 (2015)
31. Xue, Y., Zhang, S., Huang, Y., Zhang, T., Liu, X., Hu, Y., Zhang, Z., Tang, M.: Acute toxic effects and gender-related biokinetics of silver nanoparticles following an intravenous injection in mice. J. Appl. Toxicol. **32**, 890–899 (2012)
32. Larese, F.F., D'Agostin, F., Crosera, M., Adami, G., Renzi, N., Bovenzi, M., Maina, G.: Human skin penetration of silver nanoparticles through intact and damaged skin. Toxicology **255**, 33–37 (2009)
33. Liu, J., Wang, Z., Liu, F.D., Kane, A.B.., Hurt, R.H.: Chemical transformations of nanosilver in biological environments. ACS Nano **6**, 9887–9899 (2012)
34. Kim, W.Y., Kim, J., Park, J.D., Ryu, H.Y., Yu, I.J.: Histological study of gender differences in accumulation of silver nanoparticles in kidneys of Fischer 344 rats. J. Toxicol. Environ. Health A **72**, 1279–1284 (2009)
35. Lankveld, D.P., Oomen, A.G., Krystek, P., Neigh, A., Troost-de Jong, A., Noorlander, C.W., Van Eijkeren, J. C., Geertsma, R.E., De Jong, W.H.: The kinetics of the tissue distribution of silver nanoparticles of different sizes. Biomaterials **31**, 8350–8361 (2010)

36. van der Zande, M., Vandebriel, R.J., Van Doren, E., Kramer, E., Herrera Rivera, Z., Serrano-Rojero, C.S., Gremmer, E.R., Mast, J., Peters, R.J., Hollman, P.C., Hendriksen, P.J., Marvin, H.J., Peijnenburg, A.A., Bouwmeester, H.: Distribution, elimination, and toxicity of silver nanoparticles and silver ions in rats after 28-day oral exposure. ACS Nano **6**, 7427–7442 (2012)
37. Bachler, G., von Goetz, N., Hungerbuhler, K.: A physiologically based pharmacokinetic model for ionic silver and silver nanoparticles. Int. J. Nanomedicine **8**, 3365–3382 (2013)

In Vitro Toxicity

► Nanoparticle Cytotoxicity

In Vivo Microscopy Analysis

► Intravital Microscopy Analysis

In Vivo Toxicity of Carbon Nanotubes

Sophie Lanone
Inserm U955, Équipe 4, Université Paris Est Val de Marne (UPEC), Créteil, France
Service de pneumologie et pathologie professionnelle, Hôpital Intercommunal de Créteil, Créteil, France

Definition

Carbon nanotubes (CNT) are cylinders of one or several (up to 20) concentric graphite layer(s) (single- or multiwall carbon nanotubes respectively – SWCNT and MWCNT). Their diameter is in the order of the nanometer, and they can measure up to several micrometers in length.

Introduction

Carbon nanotubes (CNT) represent emblematic nanomaterials, which have generated a worldwide and highly competitive scientific research activity. They are a family of materials featuring a large range of diameter, length, number of walls (single- and multiwalled CNT; SWCNT and MWCNT, respectively), and metallic impurities coming from the catalyst used. Since their description in 1991, the understanding of their unique physicochemical properties, including mechanical, thermal, and electrical conductivity, led to the actual and potential development of applications in many areas such as aerospace, automobile, nanoelectronics, and medicine. Because of this increasing amount of applications, the question of potential health effects consequent to CNT exposure has been rapidly raised. In this chapter, what is currently known about in vivo toxicity of CNT will be exposed, particularly at the respiratory level, as well as the general underlying mechanisms proposed in the literature, highlighting what are the potential determinants of CNT toxicity. Finally, the remaining issues regarding the understanding of CNT in vivo toxicity will be exposed.

General Overview of CNT Toxicity In Vivo

As for other manufactured nanomaterials, studies aimed to evaluate in vivo toxicity of CNT have been mainly conducted after pulmonary exposure. This is motivated by the fact that lungs are one of the three commonly considered ports of entry for nanomaterials journey into the human body (with skin and intestinal tract), because of their direct contact with the environment. The existing literature regarding health effects of CNT related to respiratory exposure is relatively new, as the first study came out in 2004 [1]. Since then, there have been an always-increasing number of studies dealing with respiratory toxicity of CNT, mainly after exposure of mice or rats to either SWCNT or MWCNT. Initially, animals have been exposed by intratracheal administration, but intrapharyngeal and, more recently, inhalation have been now adopted as pulmonary route of administration for CNT. A large range of doses has been used so far: 10–500 μg/mouse for intratracheal administration, 0.3–30 mg/m^3 for inhalation exposure, for

example, and the duration of exposure has also been wide (from 24 h to 6 months). The majority of these studies have focused on selected specific immediate and/or long-term responses of lung tissue to the presence of CNT: the occurrence of pulmonary inflammation, the formation of granulomas, and the development of lung fibrosis. Therefore, they are not necessarily representative of the entire set of modifications that occur in exposed lungs. This does not facilitate a precise and exact understanding of CNT toxicity. However, general facts can be established.

The induction of an inflammatory response has been frequently described after pulmonary exposure to CNT. This inflammation is an early event, occurring as early as 6–24 h after the initial exposure, with the recruitment of neutrophils in the bronchoalveolar lavage (BAL) fluid [2]. Such neutrophil-driven infiltration is transient, and usually resolved within 15 days after the initial exposure. This acute-phase inflammation is accompanied by the release of pro-inflammatory cytokines, including tumor necrosis factor (TNF) alpha, interleukin (IL) -1ß, -6, monocyte chemoattractant protein (MCP) -1, or macrophage inflammatory protein (MIP) -2 (or CXCL-2) [2–4]. This increased production of inflammatory mediators is observable at the level of the BAL fluid, as well as in whole lung tissue homogenates (RNA and protein levels). This robust acute-phase inflammation, consistent with a foreign body response, is often followed by the formation of multifocal granulomas. The granulomas contain macrophages, and are usually surrounding CNT clusters, but can also appear distal, where dispersed CNT are believed to be present. The presence of granulomas has been documented to be a persistent event following CNT administration; granulomas can be still observable 6 months after the initial exposure [2]. The third biological consequence of CNT exposure that is well documented in the current literature is the development of pulmonary fibrosis. This pathological event can occur both within granulomas and as diffuse interstitial and septal fibrosis distal to granulomas. Fibrosis development has been suggested to appear as soon as 15 days after the initial exposure to CNT, and the fibrotic lesions can still be persistent after 6 months [4, 5]. The histological alteration is accompanied by the expression of markers of extracellular matrix deposition such as collagen I and III or increase in hydroxyproline lung content, as well as the production of profibrotic mediators such as transforming growth factor (TGF) beta.

Oxidative stress, defined by an imbalance between oxidants production and antioxidants defense, in favor of oxidants, is proposed to be a major mechanism underlying CNT's biological effects. This has been demonstrated by several means, in various experimental settings. For example, biomarkers of oxidative stress such as lung protein thiols or malone dialdehyde (MDA) content are increased as soon as 24 h after the initial exposure to CNT [6, 7]. Moreover, the total lung content in 4 hydroxynonenal (4-HNE), a product of lipid peroxidation, is rapidly increased after pharyngeal aspiration of SWCNT in mice [4], as well as protein carbonyls [7]. The importance of oxidant generation has been further confirmed by studies using mice on a vitamin E deficient diet. This diet led to a decreased antioxidant defense. Exposure of such mice to CNT induced a higher cellular inflammatory response (total cell count in BAL, as well as total neutrophil content), accompanied by an increased secretion of two major inflammatory cytokines, TNF and IL-6. Moreover, mice on vitamin E deficient diet developed an exaggerated fibrosis response, as demonstrated by increased TGF beta levels in the lungs, and increased thickness of alveolar walls [6]. Finally, the use of $gp91^{phox-/-}$ mice confirmed the implication of oxidants in CNT biological effects, as the absence of a functional NAD(P)H oxidase in these mice was associated to a diminished induction of collagen deposition, footprint of fibrosis development, in response to SWCNT administrated by pharyngeal aspiration 28 days earlier [8].

Beside these well-documented effects of CNT after pulmonary administration, more recent studies not especially focused on these selected specific immediate and/or long-term responses of lung tissue to the presence of CNT have been published. A few studies describe systemic effects of CNT after administration via pulmonary route.

After inhalation via whole body exposure, 0.3, 1 or 5 mg/m^3, 6 h a day, for 7 or 14 days, Mitchell and collaborators [9] have demonstrated that systemic immunity was affected, despite an absence of pulmonary effects in terms of inflammation, granuloma formation or fibrosis, and tissue injury. In this study, the authors demonstrate that spleen-derived T cells show a suppressed T cell–dependent antibody response, a decreased proliferation of T cells following stimulation, and an altered natural cell killer cell killing activity. This was the first evidence of systemic effects induced by CNT secondary to pulmonary exposure. Because of the experimental setting utilizing a whole body exposure, an ingestion of CNT because of the cleaning of the animal's fur and skin and the subsequent access of CNT to the systemic circulation cannot be ruled out. Other studies, utilizing a less physiologically relevant route of exposure (pharyngeal aspiration, intratracheal administration) allow concluding with better certainty on extrapulmonary effects of CNT. Indeed, it has been described, as soon as 4 h after the initial exposure, the presence of increased systemic levels of inflammatory proteins (IL-1ß, CXCL1, CXCL2, IL-8rß, S100a8, Mac-1) as well as stress response markers (hypoxia-inducible factor 3 (HIF-3a), matrix metalloproteinase (MMP)-9, arginase II, osteopontin, colony stimulating factor-1 (CSF-1), and insulin growth factor receptor 1 (IGF-1R)) [10]. These markers were detected in lung as well as total blood and/or in serum, but, interestingly, some of them were not detected in the lungs of the exposed animals. In addition, total and active levels of plasminogen activator inhibitor 1 (PAI-1), a procoagulant acute-phase protein which is involved in inhibition of the fibrinolytic cascade, were significantly increased in the lungs as well as in the plasma of CNT-exposed animals. This study demonstrates that 4 h after CNT deposition in the lungs, local and systemic inflammatory as well as prothrombotic responses are activated [9, 10]. Of importance, this kind of systemic response, if it is chronic and persistent, may trigger or exacerbate cardiovascular dysfunction and disease, such as atherosclerosis. Unfortunately, no evaluation of the presence of CNT outside the lungs of the animals was performed. Therefore, no conclusion can be made on the necessity or not for CNT to translocate outside the lung to have systemic effects.

A limited number of studies have considered the carcinogenic potential of CNT. A study performed on mice exposed to SWCNT by inhalation showed the occurrence of mutation in the k-ras oncogene locus within the first week of exposure, and until the end of the experiment 4 weeks later. These findings were not confirmed in a study using the pharyngeal aspiration route of exposure, implying that exposure route is a critical parameter to consider in terms of influencing the biological effects of CNT [3]. Due to the physical and chemical durability and fibrous shape of CNT, concerns have been raised that they may exhibit potentially significant health hazards similar to asbestos, such as the development of mesothelioma. To date, no such effects of CNT have been described in response to pulmonary exposure of laboratory animals. However, intraperitoneal as well as intrascrotal administration of MWCNT in mice or rats have been described to lead to the formation of abdominal and/or thoracic mesothelioma [11, 12]. Moreover, it has been recently described that CNT can rapidly reach the pleura after pulmonary exposure to CNT by inhalation of pharyngeal aspiration. From these studies, it is clear that more toxicological research is necessary to determine the effective carcinogenic potential of CNT, and that some of the studies suggest proceeding with caution.

Finally, more recent studies investigated the effect of CNT exposure in the context of existing pathologies such as asthma or bacterial infections, or along with exposure to environmental contaminants such as ozone. Indeed, exposure to CNT occurring concomitantly with an installed pulmonary disease or pathogenic infections may modify (and potentially enhance) the natural pathological response. In murine models of asthma, SWCNT as well as MWCNT have been described to increase the early inflammation and the susceptibility to develop airway fibrosis compared to CNT alone or ovalbumin sensitization only. Moreover, CNT induced an increased expression of immunoglobulins (Ig) related to allergic response in blood [13, 14]. Effects of an exposure to CNT concomitantly

with bacterial infection have been explored using Gram positive (*Listeria monocytogenes*) and Gram negative (lipopolysaccharide) bacteria. In both cases, CNT exposure in combination with bacterial infection was able to induce an increased airway fibrosis as compared to bacterial infection alone [15, 16]. An increased inflammatory acute and/or chronic response is less reproducibly evidenced. Interestingly, pharyngeal administration of SWCNT following an exposure to *Listeria monocytogenes* induced a decreased bacterial clearance. Altogether, these studies demonstrate that CNT may interfere with existing allergies and with natural responses against pathogenic infections.

Determinants of CNT Biological Effects In Vivo

Several processes are commonly used for CNT synthesis: arc discharge, chemical vapor deposition (CVD), and laser ablation. These methods share in common that energy (which can be either electrical, thermal, or high intensity light) is supplied to a carbon source (carbon monoxide, alcohols or hydrocarbons for example), which enables the further production of carbon atoms that recombine together to result in the generation of CNT. CNT synthesis is therefore a complex process resulting in a heterogeneous population of CNT presenting a whole set of various physical and chemical characteristics. Moreover, post-manufacturing processes such as (partial) removal of residual catalyst content may lead to further altered physicochemical characteristics of the initially produced CNT. From the existing literature concerning biological effects of CNT, it clearly appears that several factors mainly related to the physicochemical characteristics of CNT are major determinants of their subsequent biological effects. The exact physicochemical determinant(s) conditioning CNT biological effects is/are unknown for now, but these factors include, for example, the length, catalytic residues nature and content, surface properties, etc., of CNT. As a general comment, and before going further into details, care must be taken regarding the experimental design of the studies aimed to determine the importance of (a) specific physicochemical determinant(s), in particular concerning the origin of the CNT utilized to perform the study. Indeed, as exposed earlier, fabrication processes for CNT synthesis are highly complex, and can lead to the production of a large variety of CNT. Therefore, trying to compare results obtained with CNT issued from different origins (producers, fabrication process) can be hazardous when it comes to giving straightforward conclusions.

Length has been accepted to influence fiber clearance, because it dictates the ability of phagocytic cells in general, and macrophages in particular, to completely internalize and further process the fibers. The influence of CNT length on their biological effects has been therefore assessed in a few studies, and from these data, it appears that shorter CNT have less biological effects than long CNT, in terms of reactive oxygen species production, inflammatory mediators release, and inflammatory cell recruitment [17, 18]. Longer CNT promote frustrated phagocytosis, with a failure of macrophages to entirely engulf CNT. This has been demonstrated for CNT longer than 5 μm [19]. However, several issues must be raised concerning the effect of length in CNT biological effects, apart from those mentioned above. First of all, due to technical issues, it is very difficult to measure accurately CNT length. Indeed, they have a tendency to entangle and aggregate, which complicates the measurement. Therefore, microscopy techniques (usually transmission electron microscopy, TEM) are often used to demonstrate (but not quantify) the differences of CNT length. Moreover, shortening of CNT is often obtained by grinding raw CNT materials, by the use of an agate ball mill, a process which can introduce structural defects in CNT [20]. This would lead to the modification of not only the length of CNT, but also other physicochemical characteristics of the CNT. Interestingly, Wako and collaborators recently demonstrated that grinding CNT could modify inflammatory response in terms of localization of macrophages infiltrates [21]. Indeed, they showed that

intratracheal administration of ground MWCNT in rats lead to the infiltration of MWCNT-laden macrophages in the alveoli, whereas unground MWCNT-laden macrophages accumulated in the pulmonary interstitium. This was associated with the formation of smaller aggregates in the case of ground MWCNT, as compared to unground MWCNT.

The question of the dispersion and subsequent potency for CNT to form aggregates in solution is often proposed as an important determinant of their biological effects. What is currently believed is that well-dispersed CNT preferentially induce the development of fibrosis, whereas less dispersed CNT lead to the formation of granuloma. Better dispersion of CNT can be achieved by the use of dispersing agents (such as surfactants), solvents, attachment of functional groups to CNT's surface (functionalization process), or mechanical treatment, such as grinding or sonication. Moreover, the administration mode can also have an impact on CNT dispersion, as seen in a study by Li and coworkers demonstrating the formation of small aggregates after inhalation of CNT, as compared to the formation of larger aggregates consequent to intratracheal instillation of similar amounts of CNT [22].

Finally, and directly resulting from their production process, the amount of residual catalyst particles present in CNT is believed to play an important role in determining CNT biological effects. Iron, nickel, and cobalt are the most common catalyst particles primarily used during CNT manufacture. Their amount in the final CNT product varies with the manufacturing process, as well as the posttreatments undertaken to purify as produced raw CNT. Several studies aimed at understanding the role of CNT metal content in their biological effects suggest that the metal content (and particularly that of iron) of CNT can contribute to these effects probably via the induction of an oxidative stress. However, these studies are not consistent in their conclusions, as some show that a high content in catalyst residues is associated to increased biological effects, and some others report the opposite findings. Differences between studies may relate to the bioavailability of the catalyst residues in the various CNT samples; some catalyst residues are engulfed inside the carbon structure of CNT, although in some other cases, catalyst residues are present at the surface of the carbon structure, being therefore more accessible and prone to play a role in CNT's biological effects. Finally, the purification process (high temperature, strong acidification process) used to get (at least partly) rid of catalyst residues may introduce other physicochemical modifications in the so-called purified CNT, and therefore complicate the interpretation of the results.

Remaining Issues Regarding CNT Toxicity

Several specific end points have been either not or only poorly considered until now regarding CNT toxicity. This essentially concerns the translocation of CNT as well as their further processing after pulmonary exposure. If it is quite clear given the actual literature on the subject that CNT can have extrapulmonary effects after pulmonary administration (whatever the exposure route chosen), the underlying mechanisms are quite unclear, particularly regarding the need of the translocation of CNT to systemic circulation or other organs to have an extrapulmonary effect. There are some evidences of biopersistence of CNT after pulmonary exposure; Muller and collaborators early demonstrated that 60 days after the initial intratracheal administration, roughly 80 % of MWCNT were found inside the lungs of the exposed rats [5]. Interestingly, only 36 % of the initial dose of MWCNT was found inside the lungs when the authors considered ground MWCNT. Unfortunately, no indication on the localization (organ, body fluids) of the CNT has been achieved in this study. When considering the biodistribution of CNT, it is important to have in mind that observation of CNT in a biological context is a complex matter. Optical and conventional electronic microscopy observations are of limited interest, especially in the case of small diameter SWCNT, because of limitations related to the spatial resolution of these techniques and to

the low contrast of these materials [23]. One possibility is to graft a functionalized group to the raw CNT, but this could affect the intrinsic physicochemical characteristics of the CNT, therefore modifying their biological effects. Another possibility is to take advantage of the residual catalyst present in the CNT. However, it requires highly demanding techniques, such as synchrotron-based X-ray fluorescence microscopy (μXRF) [24], not convenient to use on a routine basis.

Another important remaining issue in the comprehensive understanding of CNT's biological effects is the question of their *devenir* and potential biotransformation in a biological context, despite the earlier-mentioned indications of biopersistence, along with evidences of CNT penetration inside cells. Such data are very important to obtain since CNT biotransformation products could have their own biological effects and therefore participate in CNT's reactivity and further toxicity. A few groups have demonstrated, essentially in test tubes, the possible modifications of SWCNT [25, 26]. However, more extensive research is clearly needed regarding this matter.

Conclusions

The analysis of the existing literature regarding biological effects following respiratory exposure to CNT allows to conclude that CNT per se can induce inflammatory response, granuloma formation, and development of pulmonary fibrosis. CNT can also modulate biological response to preexisting pathologies such as asthma and bacterial infections. The exact determinants of these effects still remain to be clearly identified, although intrinsic physicochemical characteristics of CNT such as length, catalyst residues nature and content, and dispersion state (along with aggregation potential and distribution pattern) seem to be of importance.

Several critical issues remain, and the development of analytical tools for the detection and characterization of CNT should be able to bring some further insights in the understanding of CNT biological effects.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Chemical Vapor Deposition (CVD)
- ▶ Detection of Nanoparticulate Biomarkers in Blood
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Genotoxicity of Nanoparticles
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity

References

1. Lam, C., James, J.H., McCluskey, R., Hunter, R.: Pulmonary toxicity of single-wall carbon nanotubes in mice 7 and 90 days after intratracheal instillation. Toxicol. Sci. **77**, 126–134 (2004)
2. Tabet, L., Bussy, C., Setyan, A., Simon-Deckers, A., Rossi, M.J., Boczkowski, J., Lanone, S.: Coating carbon nanotubes with a polystyrene-based polymer protects against pulmonary toxicity. Part. Fibre Toxicol. **8**, 3 (2010)
3. Johnston, H.J., Hutchison, G.R., Christensen, F.M., Peters, S., Hankin, S., Aschberger, K., Stone, V.: A critical review of the biological mechanisms underlying the in vivo and in vitro toxicity of carbon nanotubes: the contribution of physico-chemical characteristics. Nanotoxicology **4**, 207–246 (2010)
4. Shvedova, A.A., Kisin, E.R., Mercer, R., Murray, A.R., Johnson, V.J., Potapovich, A.I., Tyurina, Y.Y., Gorelik, O., Arepalli, S., Schwegler-Berry, D., Hubbs, A.F., Antonini, J., Evans, D.E., Ku, B.K., Ramsey, D., Maynard, A., Kagan, V.E., Castranova, V., Baron, P.: Unusual inflammatory and fibrogenic pulmonary responses to single-walled carbon nanotubes in mice. Am. J. Physiol. Lung Cell. Mol. Physiol. **289**, L698–L708 (2005)
5. Muller, J., Huaux, F., Moreau, N., Misson, P., Heilier, J.F., Delos, M., Arras, M., Fonseca, A., Nagy, J.B., Lison, D.: Respiratory toxicity of multi-wall carbon nanotubes. Toxicol. Appl. Pharmacol. **207**, 221–231 (2005)
6. Shvedova, A.A., Kisin, E.R., Murray, A.R., Gorelik, O., Arepalli, S., Castranova, V., Young, S.H., Gao, F., Tyurina, Y.Y., Oury, T.D., Kagan, V.E.: Vitamin E deficiency enhances pulmonary inflammatory response and oxidative stress induced by single-walled carbon nanotubes in C57BL/6 mice. Toxicol. Appl. Pharmacol. **221**, 339–348 (2007)
7. Shvedova, A.A., Kisin, E.R., Murray, A.R., Johnson, V.J., Gorelik, O., Arepalli, S., Hubbs, A.F., Mercer, R.

R., Keohavong, P., Sussman, N., Jin, J., Yin, J., Stone, S., Chen, B.T., Deye, G., Maynard, A., Castranova, V., Baron, P.A., Kagan, V.E.: Inhalation versus aspiration of single walled carbon nanotubes in C57bl/6 mice: inflammation, fibrosis, oxidative stress and mutagenesis. Am. J. Physiol. Lung Cell. Mol. Physiol. **95**, L552–L565 (2008)
8. Shvedova, A.A., Kisin, E.R., Murray, A.R., Kommineni, C., Castranova, V., Fadeel, B., Kagan, V.E.: Increased accumulation of neutrophils and decreased fibrosis in the lung of NADPH oxidase-deficient C57BL/6 mice exposed to carbon nanotubes. Toxicol. Appl. Pharmacol. **231**, 235–240 (2008)
9. Mitchell, L.A., Gao, J., Wal, R.V., Gigliotti, A., Burchiel, S.W., McDonald, J.D.: Pulmonary and systemic immune response to inhaled multiwalled carbon nanotubes. Toxicol. Sci. **100**, 203–214 (2007)
10. Erdely, A., Hulderman, T., Salmen, R., Liston, A., Zeidler-Erdely, P.C., Schwegler-Berry, D., Castranova, V., Koyama, S., Kim, Y.-A., Endo, M., Simeonova, P.P.: Cross-talk between lung and systemic circulation during carbon nanotube respiratory exposure. Potential biomarkers. Nano Lett. **9**, 36–43 (2008)
11. Sakamoto, Y., Nakae, D., Fukumori, N., Tayama, K., Maekawa, A., Imai, K., Hirose, A., Nishimura, T., Ohashi, N., Ogata, A.: Induction of mesothelioma by a single intrascrotal administration of multi-wall carbon nanotube in intact male Fischer 344 rats. J. Toxicol. Sci. **34**, 65–76 (2009)
12. Takagi, A., Hirose, A., Nishimura, T., Fukumori, N., Ogata, A., Ohashi, N., Kitajima, S., Kanno, J.: Induction of mesothelioma in p53+/− mouse by intraperitoneal application of multi-wall carbon nanotube. J. Toxicol. Sci. **33**, 105–116 (2008)
13. Inoue, K., Koike, E., Yanagisawa, R., Hirano, S., Nishikawa, M., Takano, H.: Effects of multi-walled carbon nanotubes on a murine allergic airway inflammation model. Toxicol. Appl. Pharmacol. **237**, 306–316 (2009)
14. Ryman-Rasmussen, J.P., Tewksbury, E.W., Moss, O. R., Cesta, M.F., Wong, B.A., Bonner, J.C.: Inhaled multi-walled carbon nanotubes potentiate airway fibrosis in murine allergic asthma. Am. J. Respir. Cell Mol. Biol. **40**, 349–358 (2008)
15. Cesta, M.F., Ryman-Rasmussen, J.P., Wallace, D.G., Masinde, T., Hurlburt, G., Taylor, A.J., Bonner, J.C.: Bacterial lipopolysaccharide enhances PDGF signaling and pulmonary fibrosis in rats exposed to carbon nanotubes. Am. J. Respir. Cell Mol. Biol. **43**, 142–151 (2009)
16. Shvedova, A.A., Fabisiak, J.P., Kisin, E.R., Murray, A.R., Roberts, J.R., Tyurina, Y.Y., Antonini, J.M., Feng, W.H., Kommineni, C., Reynolds, J., Barchowsky, A., Castranova, V., Kagan, V.E.: Sequential exposure to carbon nanotubes and bacteria enhances pulmonary inflammation and infectivity. Am. J. Respir. Cell Mol. Biol. **38**, 579–590 (2008)
17. Kolosnjaj-Tabi, J., Hartman, K.B., Boudjemaa, S., Ananta, J.S., Morgant, G., Szwarc, H., Wilson, L.J., Moussa, F.: In vivo behavior of large doses of ultrashort and full-length single-walled carbon nanotubes after oral and intraperitoneal administration to Swiss mice. ACS Nano **4**, 1481–1492 (2010)
18. Muller, J., Huaux, F., Fonseca, A., Nagy, J.B., Moreau, N., Delos, M., Raymundo-Pinero, E., Beguin, F., Kirsch-Volders, M., Fenoglio, I., Fubini, B., Lison, D.: Structural defects play a major role in the acute lung toxicity of multiwall carbon nanotubes: toxicological aspects. Chem. Res. Toxicol. **21**, 1698–1705 (2008)
19. Poland, C.A., Duffin, R., Kinloch, I., Maynard, A., Wallace, W.A., Seaton, A., Stone, V., Brown, S., Macnee, W., Donaldson, K.: Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot study. Nat. Nanotechnol. **3**, 423–428 (2008)
20. Fenoglio, I., Greco, G., Tomatis, M., Muller, J., Raymundo-Pinero, E., Beguin, F., Fonseca, A., Nagy, J.B., Lison, D., Fubini, B.: Structural defects play a major role in the acute lung toxicity of multiwall carbon nanotubes: physicochemical aspects. Chem. Res. Toxicol. **21**, 1690–1697 (2008)
21. Wako, K., Kotani, Y., Hirose, A., Doi, T., Hamada, S.: Effects of preparation methods for multi-wall carbon nanotube (MWCNT) suspensions on MWCNT induced rat pulmonary toxicity. J. Toxicol. Sci. **35**, 437–446 (2010)
22. Li, J.G., Li, W.X., Xu, J.Y., Cai, X.Q., Liu, R.L., Li, Y. J., Zhao, Q.F., Li, Q.N.: Comparative study of pathological lesions induced by multiwalled carbon nanotubes in lungs of mice by intratracheal instillation and inhalation. Environ. Toxicol. **22**, 415–421 (2007)
23. Porter, A.E., Gass, M., Muller, K., Skepper, J.N., Midgley, P.A., Welland, M.: Direct imaging of single-walled carbon nanotubes in cells. Nat. Nanotechnol. **2**, 713–717 (2007)
24. Bussy, C., Cambedouzou, J., Lanone, S., Leccia, E., Heresanu, V., Pinault, M., Mayne-L'hermite, M., Brun, N., Mory, C., Cotte, M., Doucet, J., Boczkowski, J., Launois, P.: Carbon nanotubes in macrophages: imaging and chemical analysis by x-ray fluorescence microscopy. Nano Lett. **8**, 2659–2663 (2008)
25. Allen, B.L., Kichambare, P.D., Gou, P., Vlasova, I.I., Kapralov, A.A., Konduru, N., Kagan, V.E., Star, A.: Biodegradation of single-walled carbon nanotubes through enzymatic catalysis. Nano Lett. **8**, 3899–3903 (2008)
26. Kagan, V.E., Konduru, N.V., Feng, W., Allen, B.L., Conroy, J., Volkov, Y., Vlasova, I.I., Belikova, N.A., Yanamala, N., Kapralov, A., Tyurina, Y.Y., Shi, J., Kisin, E.R., Murray, A.R., Franks, J., Stolz, D., Gou, P., Klein-Seetharaman, J., Fadeel, B., Star, A., Shvedova, A.A.: Carbon nanotubes degraded by neutrophil myeloperoxidase induce less pulmonary inflammation. Nat. Nanotechnol. **5**, 354–359 (2010)

In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles

Jorge Boczkowski
INSERM U955 Eq04, University Paris Est Val de Marne (UPEC), Créteil, France

Synonyms

Nanoparticles toxicity examined in vivo in animals

Definition

This topic examines the effect in animals of exposure to titanium dioxide and gold nanoparticles. The effects on different organs are examined along with an analysis of the involved mechanisms.

Introduction

This chapter examines the in vivo toxicity of titanium dioxide (TiO_2) and gold nanoparticles (NPs), the former being representative of metal oxide NPs. This emanates from the finding that research has been primarily performed using these NPs, and thus they have the richest database of information relating to their toxicity.

The effects of the different NPs will be described taking into account the way of exposure.

In Vivo Effects of TiO_2 NPs

Introduction

Metal oxide NPs can be composed of a variety of diverse materials, including titanium, zinc, cerium, aluminum, and iron oxides. When considering metal oxide particles, most available studies have focused on the toxicity of TiO_2 due to its widespread exploitation involving their photocatalytic, dielectric, and optical properties. TiO_2 is a component of sunscreens, pigments and paints, solar panels, and catalyst supports [1].

Titanium dioxide exists in a variety of crystal structures. The most actively investigated are its rutile, anatase, and brookite forms. All are studied for their photocatalytic and photoelectrochemical applications; however, the difference between their performances in applications is often related to the crystal structure. In the bulk form, rutile is the most stable, while, in the NP form, anatase appears to be the most stable structure. The surface energy of anatase is lower than for rutile and brookite. Such surface energy considerations have been used to accurately predict structural phase transformations at the nanoscale between the three systems. The cytotoxic properties of TiO_2 NPs are also dependent on their phase composition. Hence, Sayes et al. [2] demonstrated that anatase TiO_2 NPs (3–10 nm) were 100 times more toxic than an equivalent sample of rutile TiO_2. Moreover, reactive oxygen species (ROS) generation under UV illumination correlated well with the observed biological responses. The reason for these differences may be that anatase TiO_2 possesses a higher efficiency in the generation of ROS [3]. While the latter studies does not rule out particle size as a parameter that can influence in vitro toxicity of NPs, it appears that size is less or equally important than phase composition in the specific case of TiO_2 NPs (see below for further discussion).

Systemic Effects

Skin Penetration

Kiss et al. [4] investigated the in vivo penetration of TiO_2 NPs via human skin transplanted to immunodeficient mice. Furthermore, they measured the in vitro effects of nanoparticles on various functional properties of numerous epidermal and dermal cells in culture. Using various nuclear microscopy methods, the authors provided evidence that TiO_2 NPs in vivo do not penetrate through the intact epidermal barrier. However, they also report that TiO_2, when exposed directly to cell cultures in vitro, exerts significant and cell type–dependent effects on such cellular functions as viability, proliferation, apoptosis, and differentiation.

Since TiO_2 is included in some sunscreen formulations to physically block ultraviolet radiation,

Sadrieh et al. [5] conducted a dermal penetration study in minipigs with three TiO_2 particles (uncoated submicron-sized, uncoated nano-sized, and dimethicone/methicone copolymer-coated nano-sized) applied 5 % by weight in a sunscreen. These and control formulations were topically applied to minipigs at 2 mg cream/cm^2 skin (4 applications/day, 5 days/week, 4 weeks). Each cream was applied uniformly beginning at the neck and covered the entire dorsal surface and ventral abdomen ending at the base of the tail. Skin (multiple sites), lymph nodes, liver, spleen, and kidneys were removed, and the TiO_2 content was determined (as titanium) using inductively coupled plasma mass spectroscopy. An increase in titanium in the left inguinal lymph nodes following treatment with uncoated nano-TiO_2 and an increase in the right inguinal lymph nodes of minipigs treated with submicron TiO_2 was observed. When the data were combined for left and right inguinal lymph nodes for the same minipig, there were statistically significant increases in titanium from 0.758 μg/g in the control group to 1.133 and 1.232 μg/g in the submicron and uncoated nano-TiO_2 treatment groups, respectively. This pattern of increased titanium was not evidenced in the other lymph nodes analyzed. There were no statistical differences in the titanium levels in the prescapular or submandibular lymph nodes taken from any of the TiO_2-treated minipigs. When all the lymph node data were combined, there were no statistical differences in the lymph node titanium concentrations between the treatment groups. Titanium levels in liver from treated animals were not increased over the values in control animals. These data and the data from lymph nodes suggest an absence of transdermal passage of the particles. The epidermis from minipigs treated with sunscreens containing TiO_2 showed elevated titanium. Increased titanium was detected in abdominal and neck dermis of minipigs treated with uncoated and coated nanoscale TiO_2. Using electron microscopy energy-dispersive X-ray analysis, all three types of TiO_2 particles were found in the stratum corneum and upper follicular lumens in all treated skin samples (more particles visible with coated nanoscale TiO_2). Isolated titanium particles were also present at various locations in the dermis of animals treated with all three types of TiO_2-containing sunscreens; however, there was no pattern of distribution or pathology, suggesting the particles could be the result of contamination. At most, the few isolated particles represent a tiny fraction of the total amount of applied TiO_2. The authors conclude that there is no significant penetration of TiO_2 nanoparticles through the intact normal epidermis.

Adachi et al. [6] morphologically investigated animal skin exposed to water/oil emulsion containing 10 wt% TiO_2 NPs. After 4 h, exposed skin was investigated by light microscopy, confocal laser scanning microscopy, and electron microscopy with energy-dispersive X-ray spectrometry (EDX). Light microscopic evaluation was also performed on the exposed skin after 24, 72, and 168 h. Light microscopy did not show any morphological and immunohistochemical changes in the skin. Electron microscopy revealed that the most TiO_2 particles were localized in the interfollicular stratum disjunctum and the keratinized layer of follicular infundibulum. No TiO_2 particles were detected in the viable skin, which was confirmed by EDX. Furthermore, the authors demonstrated a specific TiO_2 affinity to the follicular opening area by light microscopy and low-vacuum scanning electron microscopy with EDX. The authors conclude that TiO_2 particles neither penetrate into viable cell layers nor biologically cause any cellular changes.

Apart from the isolated particles found by Sadrieh et al. [5] in the dermis, these studies do not demonstrate that TiO_2 NPs can penetrate through the intact epidermal barrier. However, the penetration across an altered skin must be still investigated.

Central Nervous System Accumulation and Effects

Wang et al. [7] investigated the distributions of nano- and microparticles (diameters 80 and 155 nm respectively) in the olfactory bulb of mice after nasal instillation (500 μg of TiO_2 nanoparticles suspension every other day for 30 days). These authors demonstrated that both types of TiO_2 particles were taken up by the olfactory bulb via the primary olfactory neurons and then accumulated in the olfactory nerve layer, olfactory ventricle, and granular cell layer of the olfactory bulb. The TiO_2 content was increased in

all investigated brain regions (olfactory bulb, cerebral cortex, hippocampus, and cerebellum), with the most significant increases seen in the hippocampus. The presence of TiO_2 in hippocampus was furthermore accompanied by changes in neuron morphology and increased amount of GFAP-positive cells in the CA4 region. Signs of oxidative stress were documented in all regions of the brain. Interestingly, in general, anatase TiO_2 gave rise to stronger effects than the rutile form. Similar results were observed with other NPs, including gold NPs [8]. As stated by Simkó and Mattsson [9], it seems that NPs can translocate to the nervous system through sensory nerves. Translocation of 20 nm particles is 2–10 times higher in the human olfactory bulb than in rats. Thus, the translocated nanoparticles in humans can enter into the deeper brain structures in short exposure time. However, based on the limited data available, it is presently difficult to assess to what extent accumulation in the brain via axonal transport is a realistic possibility.

Since synaptic plasticity in hippocampus is believed to be associated with certain high functions of CNS, such as learning and memory, Gao et al. [10] investigated the effects of developmental exposure to TiO_2 NPs on synaptic plasticity in rat's hippocampal dentate gyrus (DG) area using in vivo electrophysiological recordings. The input/output (*I*/*O*) functions, paired-pulse reaction (PPR), field excitatory postsynaptic potential, and population spike amplitude were measured. The results of the study show that the *I*/*O* functions, PPR, and long-term potentiation were all attenuated in lactation TiO_2 NPs-exposed offspring rats compared with those in the control group. However, in the pregnancy TiO_2 NPs exposure group, only PPR was attenuated significantly. These findings suggest that developmental exposure to TiO_2 NPs could affect synaptic plasticity in offspring's hippocampal DG area in vivo, which indicates that developmental brains, especially in lactation, are susceptible to TiO_2 NPs exposure.

Effects on Liver

Ma et al. [11] investigated the effects of TiO_2 NPs in liver after intraperitoneal administration in mice. In this study, the authors injected TiO_2 NPs (5 nm diameter, anatase) into the abdominal cavity of ICR mice for consecutive 14 days, and liver inflammatory response was investigated. The results showed the obvious TiO_2 accumulation in liver along with DNA damage, histopathological changes, and hepatocytes apoptosis. These findings were associated with an impaired function (only in the case of the higher doses of TiO_2). TiO_2 NPs significantly altered the mRNA and protein expressions of several inflammatory cytokines, including NF-κB, macrophage migration inhibitory factor, TNF-α, Il-6, Il-1β, Il-4, and Il-10. The authors conclude that the inflammatory responses and liver injury may be involved in TiO_2 NPs-induced liver toxicity.

Pulmonary Exposure to TiO_2 NPs

Different studies investigated pulmonary effects of TiO_2 NPs (see [12] for review). For clarity purposes, they are classified into those examining inflammation and those examining carcinogenesis.

Lung Inflammatory Responses to TiO_2 NPs

Grassian et al. [13] examined the acute pulmonary toxicity of TiO_2, with a primary particle size of 2–5 nm, in vivo using a murine model. Mice exposed to an aerosol of 0.77 or 7.22 mg/m^3 NPs demonstrated minimal lung toxicity or inflammation. At a higher dose, 8.88 mg/m^3, negative health effects were observed immediately as well as at week 1 or 2 postexposure: higher counts of total cells and alveolar macrophages in the bronchoalveolar fluid compared with controls, but after 3 weeks postexposure the mice recovered.

Warheit et al. [14] studied pulmonary toxicity in rats exposed, by intratracheal instillation, to two forms of nano-sized rutile TiO_2 particles (uf-1 or uf-2, similar diameter about 100 nm), and one 80/20 anatase/rutile TiO_2 particle (uf-3, diameter about 25 nm). The authors ranked the pulmonary inflammation/cytotoxicity/cell proliferation and histopathological responses as follows: uf-3 > uf-1 = uf-2. The exposure to a positive control (quartz) and to a lesser degree, anatase/rutile TiO_2 (uf-3) produced the most significant effects. It was concluded that anatase structure could be linked to the adverse effects and that inhaled rutile TiO_2 NPs particles are

expected to have a low-risk potential for producing adverse pulmonary health effects.

Since the translocation of NPs from the pulmonary airways into other pulmonary compartments or the systemic circulation is controversially discussed in the literature, Muhlfeld et al. [15] investigated the distribution of TiO_2 NPs in rat lung after pulmonary exposure. Animals were exposed during 1 h to a Spark-generated aerosol of TiO_2 NPs (median diameter 22 nm). The authors found that NPs distribution within four compartments (air-filled spaces, epithelium/endothelium, connective tissue, capillary lumen) was not random at 1 h or at 24 h after exposure. At 1 h, the connective tissue was the preferential target of the particles. At 24 h, the NPs were preferentially located in the capillary lumen. It was established that the studied NPs can from move from one compartment to another with relative ease but that this does not happen randomly. The authors propose that a small fraction of nano-TiO_2 are rapidly transported from the airway lumen to the connective tissue and subsequently released into the systemic circulation.

Geiser et al. [16] quantified the uptake of TiO_2 NPs (20 nm) by lung macrophages at 1 h and 24 h after a 1 h aerosol inhalation. After 1 h, four particles were found in three macrophage profiles, and 47 particles in 27 macrophage profiles at 24 h. Model-based data analysis revealed an uptake of 0.06–0.12 % TiO_2 NPs by lung-surface macrophages within 24 h. These results are in line with the sporadic and rather unspecific uptake of nano-TiO_2 by macrophages and the fact that phagocytosis plays an insufficient role for clearance in peripheral lungs.

Since exposure to ambient NPs is associated with increased risk of childhood and adult asthma, Scuri et al. [17] investigated whether the exposure to TiO_2 NPs is associated with events involved in asthma pathophysiology. They focused on neurotrophin-driven airway responses because these molecules are key regulatory elements of neuronal development and responsiveness of airway sensory neurons. Indeed, changes in their expression are associated with bronchoconstriction, airway hyperresponsiveness, and airway inflammation, which are hallmarks of asthma. The results of the study show that in vivo exposure to an aerosol of TiO_2 NPs (12 mg/m^3; 5.6 h/day for 3 day) produced upregulation of lung neurotrophins in weanling (2-week-old) and newborn (2-day-old) rats but not in adult (12-week-old) animals compared to controls. This effect was associated with increased airway responsiveness and upregulation of growth-related oncogene/keratine-derived chemokine (GRO/KC; CXCL1, rat equivalent of human interleukin [IL]-8) in bronchoalveolar lavage fluid. These data show for the first time that exposure to nanoparticles upregulates the expression of lung neurotrophins in an age-dependent fashion and that this effect is associated with airway hyperresponsiveness and inflammation. The authors conclude about the presence of a critical window of vulnerability in response to TiO_2NPs exposure in earlier stages of lung development, which may lead to a higher risk of developing asthma.

Lung Carcinogenic Effects of TiO_2 NPs

TiO_2 has recently been reclassified by the IARC as group 2B carcinogen: "possibly carcinogenic to humans." The reason for this new classification stems from the fact that high concentrations of micro and nano ($<$2.5 μm and $<$100 nm respectively) TiO_2 dust can cause respiratory tract cancer in exposed rats (see Ref. [18] for review). Tumors associated with experimental exposure carcinogenic agents are generally of two types: those originating from alveolar type II cells, which are called *bronchoalveolar tumors*, and squamous or epidermoid tumors, which are associated with bronchiolarization and considered to arise from areas with squamous metaplasia. In general, the tumors induced by TiO_2 exposure include bronchioloalveolar adenomas, non-keratinizing and keratinizing squamous cell tumors, bronchioloalveolar carcinomas, squamous cell carcinomas, and adeno-squamous carcinomas. However, it should be noted that epidemiologic studies of workers exposed to pigment-grade TiO_2 conducted thus far have not been able to detect an association between occupational exposure to TiO_2 and an increased risk for lung cancer (see Ref. [18] for a review).

Mechanisms of Action of TiO_2

Two recent studies investigated mechanisms of the main adverse effects of TiO_2 NPs: inflammation and carcinogenesis.

Two studies demonstrated that TiO_2 particles induce inflammation via a cellular signaling pathway know as "inflammasome." Yazdi et al. [19] showed that TiO_2 and SiO_2 NPs, but not ZnO NPs, activate the NLR pyrin domain containing three (Nlrp3) inflammasome, leading to IL-1β release and, in addition, induce the regulated release of IL-1α. Unlike other particulate Nlrp3 agonists, TiO_2 NPs-dependent-Nlrp3 activity does not require cytoskeleton-dependent phagocytosis and induces IL-1 α/β secretion in non-phagocytic keratinocytes. Pulmonary exposure to TiO_2 NPs in mice provoked lung inflammation, which was strongly suppressed in IL-1R- and IL-1α -deficient animals. The authors concluded that the inflammation caused by TiO_2 in vivo is largely caused by the biological effect of IL-1α. Similar results were described concerning micrometric TiO_2 particles [20], suggesting that the effects of TiO_2 on the inflammasome are not strictly related to the nano size.

Troullier et al. [21] investigated TiO_2 NPs (21 nm diameter, 75 % anatase, 25 % rutile)-induced genotoxicity, oxidative DNA damage, and inflammation in a mice model. The authors treated wild-type mice with TiO_2 NPs in drinking water and determined the extent of DNA damage using the comet assay, the micronuclei assay, and the gamma-H2AX immune-staining assay and by measuring 8-hydroxy-2′-deoxyguanosine levels and, as a genetic instability endpoint, DNA deletions. They also determined mRNA levels of inflammatory cytokines in the peripheral blood. The main results of this study show that TiO_2 NPs-induced 8-hydroxy-2′-deoxyguanosine, gamma-H2AX foci, micronuclei, and DNA deletions. The formation of gamma-H2AX foci, indicative of DNA double-strand breaks, was the most sensitive parameter. Inflammation was also present as characterized by a moderate inflammatory response. Together, these results describe TiO_2 NPs-induced genotoxicity in vivo in mice possibly caused by a secondary genotoxic mechanism associated with inflammation and/or oxidative stress.

Conclusion

All together, the different described articles show that TiO_2 NPs can induce inflammation and genotoxicity in vivo in rodents. They can also translocate from one organ to other (from nose to central nervous system, for example) and elicit a systemic response from a localized way of exposure (inflammatory markers in blood after oral exposure for example). In studies, when comparison between anatase and rutile NPs was performed, the former appear as more toxic than the latter TiO_2 NPs. These data suggest that there is matter of concern about a potential risk of inflammation, cancer, or genetic disorders, especially for people occupationally exposed to high concentrations of TiO_2 NPs and that it might be prudent to limit ingestion of TiO_2 NPs through nonessential drug additives, food colors, etc.

In Vivo Effects of Gold NPs

Introduction

In contrast with TiO_2, gold has different medical applications. Gold salts such as sodium gold thiomalate are still considered important as disease modifying antirheumatic agents (see Ref. [22] for a review). The precise modes of action of gold compounds in rheumatoid arthritis have not been clarified, but an interaction with thiol groups on proteins and cell membranes is thought to be important. However, long-term accumulation of gold salts in the body may result in adverse or toxic effects. On the other hand, intra-articular delivery of gold NPs (13 nm) was recently shown to have therapeutic potential for ameliorating collagen-induced arthritis in rats, a model that mimics the clinical and histologic features of human rheumatoid arthritis [23]. Gold NPs were shown to exert anti-angiogenic activities in this animal model, which could contribute to the attenuation of arthritis. These findings are in agreement with other studies showing that that gold NPs (5 nm) are able to bind specifically to certain growth factors such as vascular endothelial growth factor (VEGF), presumably through cysteine residues of the heparin-binding domain, resulting in inhibition of angiogenesis in a mouse model [22].

In addition to their anti-inflammatory effect, gold NPs are used often as drug carriers [24]. The relative stability of the ligand-gold bonding outside the cell combined with its decreased stability within cells, partially owing to the high intracellular concentrations of glutathione, also contributes to making gold NPs good candidates for drug delivery and drug release. This kind of NPs is a hybrid material featuring an inorganic core typically surrounded by an organic monolayer.

The physical properties of the core NP such as the size can influence its organ distribution. However, the organic monolayer of the NP is fundamentally important in modifying the interactions of the NP with its environment. The monolayer acts as a barrier between the NP core and the environment, and effectively protects and stabilizes the core's integrity. Furthermore, the reactivity and solubility of the NPs are dictated by the chemical nature of the monolayer periphery.

There is currently limited information available regarding the in vivo toxicity of gold NPs. In the main, the available studies focus on the tissue distribution of gold NPs following intravenous or intratracheal administration. Moreover, the study using gold NPs in collagen-induced arthritis in rats cited above did not investigate their toxicity and their distribution and elimination from the body [23]. One can argue that gold NPs should be poorly toxic in vivo since studies of gold NPs (35 nm) in vitro have demonstrated their low toxicity, at least for murine RAW 264.7 macrophages, with no hampering of cell functionality [25]. However, studies on in vivo toxicity of gold NPs need to be performed.

Gold Distribution After Intravenous Administration

Semmler-Behnke et al. [26] observed that 24 h following the intravenous administration of gold NPs (18 nm diameter), they were completely removed from the blood and were preferentially located within the liver and spleen. However, 1 nm NPs were excreted within urine (8.6 %) and feces (5 %), with a lower level of accumulation of particles within the spleen and liver, when compared to their larger counterparts, and a proportion of particles (3.7 %) remained within the blood at 24 h. Consequently, it is apparent that particle size influences the behavior of particles within the body. Although the percentages seem small, it is relevant that they may equate to a large number of particles (depending on particle size), which may be capable of exerting toxicity at secondary target sites.

Sonavane et al. [27] evaluated the tissue distribution of gold NPs (15, 50, 100, and 200 nm) within mice, 24 h following intravenous exposure (at a dose of 1 g/kg). For all particle sizes, the highest accumulation of particles was observed within the liver, with a smaller fraction localized within the lung, kidney, and spleen. However, 15 and 50 nm NPs were distributed to a greater extent and were even observed within the brain, indicating that they have the potential to cross the blood–brain barrier or that they have the potent to gain access to the brain via neuronal transport. The size of NPs also impacted on the extent of particle distribution so that there was a greater accumulation of smaller particles within organs when compared to their larger counterparts. The translocation of 200 nm particles was observed to be minute.

Similarly, de Jong et al. [28] investigated the distribution of gold particles (10, 50, 100, and 250 nm) within rats, 24 h following intravenous injection. As observed previously, the preferential site of particle (of all sizes) accumulation was the liver and spleen. The 10 nm particles were also observed to be more widely distributed within a variety of organs, with the distribution of larger particles restricted to the liver and spleen.

Gold Distribution After Intratracheal Administration

Semmler-Behnke et al. [26] observed that following the intratracheal administration, 1 nm gold NPs are observed to cross the air/blood barrier of the lungs much more efficiently than 18 nm gold NPs. The NP accumulation pattern in the secondary target organs differs strongly from those seen after direct intravenous injection. From this it can be concluded that the route of exposure determines gold NPs accumulation in various organs.

Another study showed that inhaled 20 nm nanogold NPs (2×10^6 particles/cm^3) can accumulate in the olfactory bulb of rats [8]. The exposure for 5 days resulted in a significant increase of gold NPs in the olfactory bulb (8 ng Au/g body weight). After 15 days of exposure, significant accumulations of gold particles were detected in the septum and entorhinal cortex. Both brain structures receive direct neuronal projections from the olfactory bulb and are important in attention and new memory formation.

Conclusion

Overall, these studies show that gold NPs size is able to influence the distribution of particles following intravenous exposure. In some studies [27], the dose administered was high, and it is therefore necessary to evaluate if particles actually accumulate within the suggested target organs to a great extent at lower, more relevant doses. In addition, it is necessary to determine if particle accumulation has toxicological consequences for the different secondary target sites.

General Conclusion

The main difference between TiO_2 and gold NPs are related to the know toxicity of TiO_2 NPs in some circumstances, whereas this effect seems less important in the case of gold NPs. However, more studies are needed to clearly precise the determinants of such effects in order to protect workers and consumers to an accidental exposure to NPs used for industrial applications and to improve utilization of these NPs in nanomedicine.

Cross-References

- ▶ Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates
- ▶ Genotoxicity of Nanoparticles
- ▶ Nanomedicine
- ▶ Nanoparticle Cytotoxicity

References

1. Hagfeldt, A., Graetzel, M.: Light-induced redox reactions in nanocrystalline systems. Chem. Rev. **95**, 49–68 (1995)
2. Sayes, C.M., Wahi, R., Kurian, P.A., et al.: Correlating nanoscale titania structure with toxicity: a cytotoxicity and inflammatory response study with human dermal fibroblasts and human lung epithelial cells. Toxicol. Sci. **92**, 174–185 (2006)
3. Mo, S., Ching, W.: Electronic and optical properties of three phases of titanium dioxide: rutile, anatase, and brookite. Phys. Rev. B **51**, 13023–13032 (1995)
4. Kiss, B., Biro, T., Czifra, G., et al.: Investigation of micronized titanium dioxide penetration in human skin xenografts and its effect on cellular functions of human skin-derived cells. Exp. Dermatol. **17**, 659–667 (2008)
5. Sadrieh, N., Wokovich, A.M., Gopee, N.V., et al.: Lack of significant dermal penetration of titanium dioxide from sunscreen formulations containing nano- and submicron-size TiO_2 particles. Toxicol. Sci. **115**, 156–166 (2010)
6. Adachi, K., Yamada, N., Yamamoto, K., et al.: In vivo effect of industrial titanium dioxide nanoparticles experimentally exposed to hairless rat skin. Nanotoxicology **4**, 296–306 (2010)
7. Wang, J., Chen, C., Liu, Y., et al.: Potential neurological lesion after nasal instillation of TiO_2 nanoparticles in the anatase and rutile crystal phases. Toxicol. Lett. **183**, 72–80 (2008)
8. Yu, L., Lanry, Y., Ong, C., et al.: Translocation and effects of gold nanoparticles after inhalation exposure in rats. Nanotoxicology **1**, 235–242 (2007)
9. Simko, M., Mattsson, M.O.: Risks from accidental exposures to engineered nanoparticles and neurological health effects: a critical review. Part. Fibre Toxicol. **7**, 42 (2010)
10. Gao, X., Yin, S., Tang, M., et al.: Effects of developmental exposure to TiO_2 nanoparticles on synaptic plasticity in hippocampal dentate gyrus area: an in vivo study in anesthetized rats. Biol. Trace Elem. Res. (2011). doi:10.1007/s12011-011-8990-4
11. Ma, L., Zhao, J., Wang, J., et al.: The acute liver injury in mice caused by nano-anatase TiO_2. Nanoscale Res. Lett. **4**(2), 1275–1285 (2009)
12. Johnston, H.J., Hutchison, G.R., Christensen, F.M., et al.: Identification of the mechanisms that drive the toxicity of TiO_2 particulates: the contribution of physicochemical characteristics. Part. Fibre Toxicol. **6**, 33 (2009)
13. Grassian, V.H., O'Shaughnessy, P.T., Adamcakova-Dodd, A., et al.: Inhalation exposure study of titanium dioxide nanoparticles with a primary particle size of 2 to 5 nm. Environ. Health Perspect. **115**, 397–402 (2007)
14. Warheit, D.B., Webb, T.R., Reed, K.L., et al.: Pulmonary toxicity study in rats with three forms of ultrafine-TiO_2 particles: differential responses related to surface properties. Toxicology **230**, 90–104 (2007)

15. Muhlfeld, C., Geiser, M., Kapp, N., et al.: Re-evaluation of pulmonary titanium dioxide nanoparticle distribution using the "relative deposition index": evidence for clearance through microvasculature. Part. Fibre Toxicol. **4**, 7 (2007)
16. Geiser, M., Casaulta, M., Kupferschmid, B., et al.: The role of macrophages in the clearance of inhaled ultrafine titanium dioxide particles. Am. J. Respir. Cell Mol. Biol. **38**, 371–376 (2008)
17. Scuri, M., Chen, B.T., Castranova, V., et al.: Effects of titanium dioxide nanoparticle exposure on neuroimmune responses in rat airways. J. Toxicol. Environ. Health A **73**, 1353–1369 (2010)
18. Borm, P.J., Schins, R.P., Albrecht, C.: Inhaled particles and lung cancer, part b: paradigms and risk assessment. Int. J. Cancer **110**, 3–14 (2004)
19. Yazdi, A.S., Guarda, G., Riteau, N., et al.: Nanoparticles activate the nlr pyrin domain containing 3 (nlrp3) inflammasome and cause pulmonary inflammation through release of il-1alpha and il-1beta. Proc. Natl. Acad. Sci. U. S. A. **107**, 19449–19454 (2010)
20. St Pierre, C.A., Chan, M., Iwakura, Y., et al.: Periprosthetic osteolysis: characterizing the innate immune response to titanium wear-particles. J. Orthop. Res. **28**, 1418–1424 (2010)
21. Trouiller, B., Reliene, R., Westbrook, A., et al.: Titanium dioxide nanoparticles induce DNA damage and genetic instability in vivo in mice. Cancer Res. **69**, 8784–8789 (2009)
22. Fadeel, B., Garcia-Bennett, A.E.: Better safe than sorry: understanding the toxicological properties of inorganic nanoparticles manufactured for biomedical applications. Adv. Drug Deliv. Rev. **62**, 362–374 (2010)
23. Tsai, C.Y., Shiau, A.L., Chen, S.Y., et al.: Amelioration of collagen-induced arthritis in rats by nanogold. Arthritis Rheum. **56**, 544–554 (2007)
24. Arvizo, R., Bhattacharya, R., Mukherjee, P.: Gold nanoparticles: opportunities and challenges in nanomedicine. Expert Opin. Drug Deliv. **7**, 753–763 (2010)
25. Shukla, R., Bansal, V., Chaudhary, M., et al.: Biocompatibility of gold nanoparticles and their endocytotic fate inside the cellular compartment: a microscopic overview. Langmuir **21**, 10644–10654 (2005)
26. Semmler-Behnke, M., Kreyling, W.G., Lipka, J., et al.: Biodistribution of 1.4- and 18-nm gold particles in rats. Small **4**, 2108–2111 (2008)
27. Sonavane, G., Tomoda, K., Makino, K.: Biodistribution of colloidal gold nanoparticles after intravenous administration: effect of particle size. Colloids Surf. B Biointerfaces **66**, 274–280 (2008)
28. De Jong, W.H., Hagens, W.I., Krystek, P., et al.: Particle size-dependent organ distribution of gold nanoparticles after intravenous administration. Biomaterials **29**, 1912–1919 (2008)

Indentation

▶ Nanomechanical Properties of Nanostructures

Induced Charge Electroosmosis

▶ AC Electroosmosis: Basics and Lab-on-a-Chip Applications

Inductively Coupled-Plasma Chemical Vapor deposition (ICP-CVD)

▶ Chemical Vapor Deposition (CVD)

Inertial Sensors

▶ Gyroscopes

Influence of Defects in Photocatalysis

Valentin G. Bessergenev[1] and Eberhard Burkel[2]
[1]CCMar – Centre of marine science and FCT, Universidade do Algarve, Faro, Portugal
[2]Institute of Physics, University of Rostock, Rostock, Germany

Synonyms

Defects; Nanomaterials; Photocatalysis; Surface science

Definition

Nanoscience has recently evolved as a major research direction of our modern society resulting from an ongoing effort to miniaturize at the

nanoscale processes that currently use microsystems. Toward this end, it is well admitted that the bottom-up approach is more and more replacing the classic top-down one, a strategic move that is common to several areas of the nanosciences including optoelectronics, sensing, medicine, and catalysis. The latter discipline certainly is a key for the development of starting chemicals, fine chemicals, and drugs from raw materials.

Unique properties, different from the bulk materials, can be achieved when the sizes of particles reach the nanometer scale. As it will be shown, for nanometer-scaled metals, the electronic structure is changed. Metal clusters of a few nanometers in sizes show dielectric properties with the discrete electronic structure. These changes in electronic structure drastically change the catalytic properties of metallic nanoclusters. Catalysis by metallic nanoclusters on metal oxide surfaces plays an increasing role in the modern technology. The oxidation of carbon monoxide at around room temperature is the most famous reaction known for gold catalysts. Haruta's group discovered in 1987 that gold is a unique catalyst for this reaction if the gold metal particles are smaller than 5 nm and if they are on oxide supports. Since then, extensive and intensive fundamental works have been published, and new applications, reaching from air purification (gas masks, gas sensors, indoor air quality control) to hydrogen purification for fuel cells, have been developed. However, the CO oxidation mechanism is not yet elucidated in spite of the huge number of papers being published.

Another class of materials, the wide band gap semiconductors represented by metal oxides, made rapid and significant advances in a series of applications. The tendency in the development of the nanomaterial's technology is the shift from application of the passive nanostructures, which are designed to perform only one task, to a system of multitasking nanostructures, like actuators and sensors, and, further, to integrated nano-systems which can contain thousands of interacting components. The combination of experimental methods of surface structure studies such as SEM, AFM, STM, and spectroscopic techniques XPS, UPS, and AES (auger electron spectroscopy), as well as SIMS and TPD (temperature-programmed desorption) with theoretical modeling methods, creates a very powerful instrument for the reconstruction of real surfaces and for the understanding of their roles in catalysis, including the role of irregularities such as defects or additional species acting as promoters or inhibitors.

It should be taken into account that a photocatalytic transformation is essentially a dynamical process in which typical time intervals lie in the nanoscale level.

Defects of various types can play very important roles in enhancing the photocatalytic activity, by creating additional electron levels that take positions between the valence and the conduction bands. Defects create local changes in potential energy, which can be considerated as quantum well (or quantum dot), where an excess electron can occupy additional levels of energy. Oxygen vacancies and various dopants (metallic and nonmetallic) inclusions are examples of such defects. Taking the lattice relaxation into account, the size of these defects is in nanometer range.

Introduction

Unique electronic and chemical properties can be manifested in solids when they reach the nanometer scale. These changes, which include the appearance of discrete electronic structures, the modification of the crystal structure, and the alteration of chemical reactivities manifests themselves as new physical and chemical properties not observed in the "bulk" form of the material, so far. The development of understanding and the ability to control the key features of nanometer-scaled catalysts is one of the main scientific challenges, which could lead to the long-sought goals of optimum catalytic activities combined with highly specific selectivities. It is well established that ultrasmall metal clusters deposited on solid substrates have catalytic properties distinct from known properties of large bulk-like particles, as illustrated by the selective oxidation of propylene to propylene oxide by the usage of gold clusters, alkene and arene hydrogenation catalysis, and CO

oxidation. In these examples, the catalytic properties improve as the clusters become smaller.

Wide band gap semiconductors represented by metal oxides opened tremendous opportunities in science and technology because of the new properties exhibited by these nanomaterials. However, there are challenging problems in theoretical physics associated with these new properties.

Furthermore an important aspect of the photocatalysis is that the photocatalytic transformation is essentially a dynamic process [1, 2]. In this process, an initial excitation of the catalyst surface or the adsorbed molecule is provoked by the adsorption of photons, which is a very fast process (10^{-15} s), followed by the internal conversion of energy (10^{-14}–10^{-13} s) and the de-excitation by fluorescence (10^{-9}–10^{-5} s) and/or intersystem crossing (10^{-6} s) of the energy or the transportation of electrons and holes created during primary excitation to the surface. The dynamic equilibrium is the factor that determines the efficiency of the photocatalytic processes.

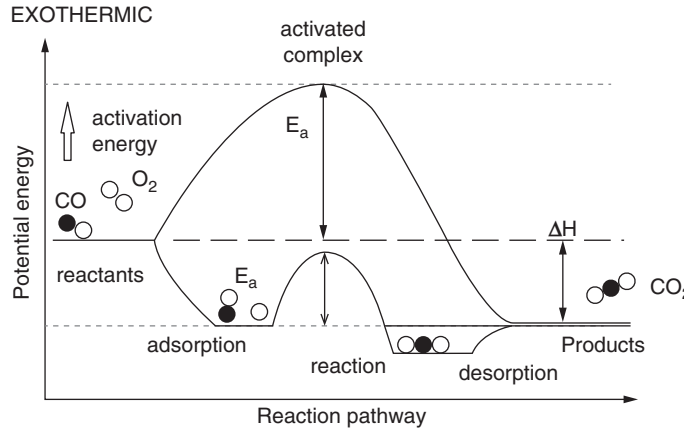
Technology of Nanomaterials and Catalysis

In order to show how surface science can be of help in the understanding of catalytic and, as a particular case, photocatalytic processes, it is useful to consider what the essence of catalysis is. Here a heterogeneous catalysis will be considered where different environment gas or liquid and a solid surface interact. A reaction in heterogeneous catalysis is a cyclic event in which molecules adsorb on the surface react on the active centers of the surface and in which, finally, products of the reaction desorb from the surface. The latter is recovered in its original form at the end of the cycle. As an example the transformation of carbon oxide into carbon dioxide can be considered ($CO + 1/2O_2 \rightarrow CO_2$). This reaction is important in automotive exhaust catalysis and usually carried out over a platinum or a palladium catalyst. Figure 1 shows the potential energy diagrams for non-catalytic and catalytic pathways explaining how catalysts accelerate the reaction between CO and O_2.

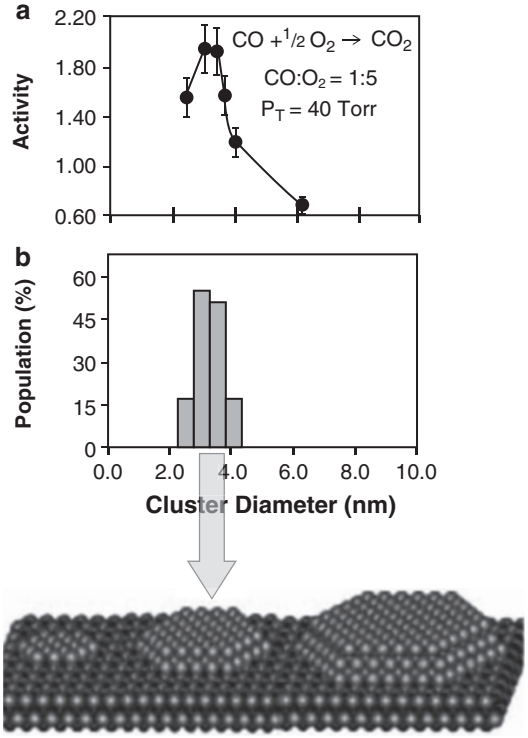
The non-catalytic reaction occurs when CO and O_2 molecules collide with sufficient energy to overcome the high potential barrier of the O–O bond of oxygen molecule splitting ($E_{splitt} \approx 498$ kJ/mol). In the initial stage of the catalytic reaction, CO and O_2 species adsorb to the catalyst loosing entropy and lowering the free energy because the adsorption is an exothermic process. Upon adsorption on the metallic surface, the O_2 molecule dissociates spontaneously for two reasons: firstly, the electrons from metal fill antibonding orbitals of the O_2 molecule and, secondly, the bonding of two O atoms to the surface is energetically more favorable than that of one O_2 molecule. In the next stage, the reaction between CO and O atoms occurs on the surface of the catalyst, however, with an energy barrier that is significantly lower than that for the non-catalytic reaction. In the final stage, the product CO_2 desorbs from the catalyst, making the surface available for another reaction cycle.

Recent results on the catalytic oxidation of carbon oxide by Au clusters [3, 4] on a TiO_2 surface can further illustrate the influence of the cluster size and the substrate on the catalytic activity. Figure 2 shows the activity of the CO oxidation at 350 K as a function of the size distribution of Au clusters in the range of 2.0–4.0 nm that are specifically two atoms thick. The sizes of the Au clusters deposited onto a TiO_2 (110) – (1 × 1) single crystal were measured by scanning tunnelling microscopy (STM) in parallel with the kinetic rate of the $CO \rightarrow CO_2$ reaction. For each point in Fig. 2, a particular Au cluster size was prepared [4]. The data of cluster sizes and surface coverage obtained from STM were then used to calculate the turnover frequency (TOF) expressed as (product molecules–CO_2) × (total Au atoms on surface sites)$^{-1}$sec^{-1} [4]. As it can be seen in Fig. 2, the activity of the Au/TiO_2 catalysts exhibits a maximum activity at an average Au cluster size of ~3.5 nm.

Electronic structure of Au clusters is of great importance for catalytic activity and has been studied by scanning tunnelling spectroscopy (STS) [4]. Figure 3 shows a plot of the band gap measured by STS over the cluster surfaces used for the $CO \rightarrow CO_2$ reaction. There is a clear correlation between



Influence of Defects in Photocatalysis, Fig. 1 Potential energy diagrams for the reaction $CO + O_2 \rightarrow CO_2$ showing that the catalytic pathway offers a significantly lower activation energy



Influence of Defects in Photocatalysis, Fig. 2 (**a**) The activity for CO oxidation at 350 K as a function of Au cluster size supported on TiO_2 (110). A 1:5 $CO:O_2$ mixture was used at a total pressure of 40 Torr. Activity is expressed as (product molecules) × (total Au atoms)$^{-1}$ × sec^{-1} (**b**) (From Goodman [4]. © 2004 Marcel Dekker, Inc)

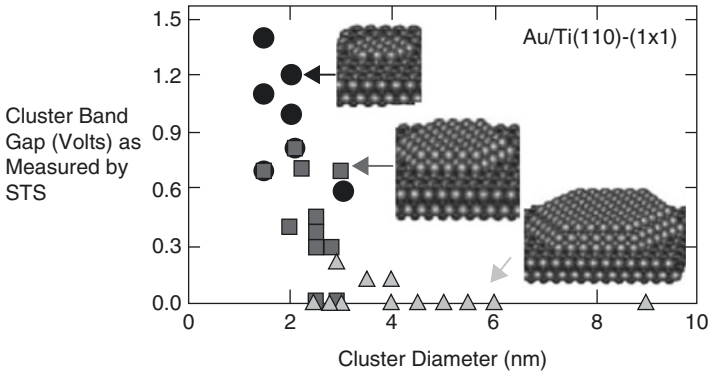
the onset of the catalytic activity and the observed metal – nonmetal transition in the supported Au clusters [4]. The STS-measured band gaps of this group of clusters range from 0.2 to 0.6 V.

From XPS measurements, it was also concluded that for adsorbed Au, the Au d bands are much closer to the Fermi level (E_f) because of a charge polarization in the interfacial region and a subsequent increase of the potential in the ad layer. Such a large energy shift of the Au 5*d* band toward the E_f should strongly alter the chemical surface properties of Au/TiO_2 from those of a clean bulk Au surface [3, 4]. At this point, an importance of the material of the substrate can be emphasized because this shift can be different for various materials.

Surface Science and Photocatalysis

Heterogeneous photocatalysis differs from catalysis by the usage of light as a special source of energy for the initial excitation of the system that is followed by the subsequent electron transfer and/or energy transfer to a reactive center. A reactive center can be a molecule or a surface reactive site. The subsequent de-excitation processes lead to chemical reactions in heterogeneous photocatalytic processes.

Various semiconductors and wide band gap semiconductors were studied, since in 1972; Fujishima and Honda discovered the photocatalytic splitting of water on TiO_2 electrodes [5]. Numerous research efforts in understanding the fundamental processes and in enhancing the efficiency of semiconductor-based



Influence of Defects in Photocatalysis, Fig. 3 Cluster band gaps were measured by STS as a function of Au cluster size supported on TiO_2 (110). The band gaps were obtained while the corresponding topographic scan was acquired on various Au coverages ranging from 0.2 ML to 4.0 ML. •: two dimensional (2D) clusters; □: 3D clusters, 2-atom layers in height; Δ: 3D clusters, 3-atom layers or grater in height (From Goodman [4]. © 2004 Marcel Dekker, Inc)

I

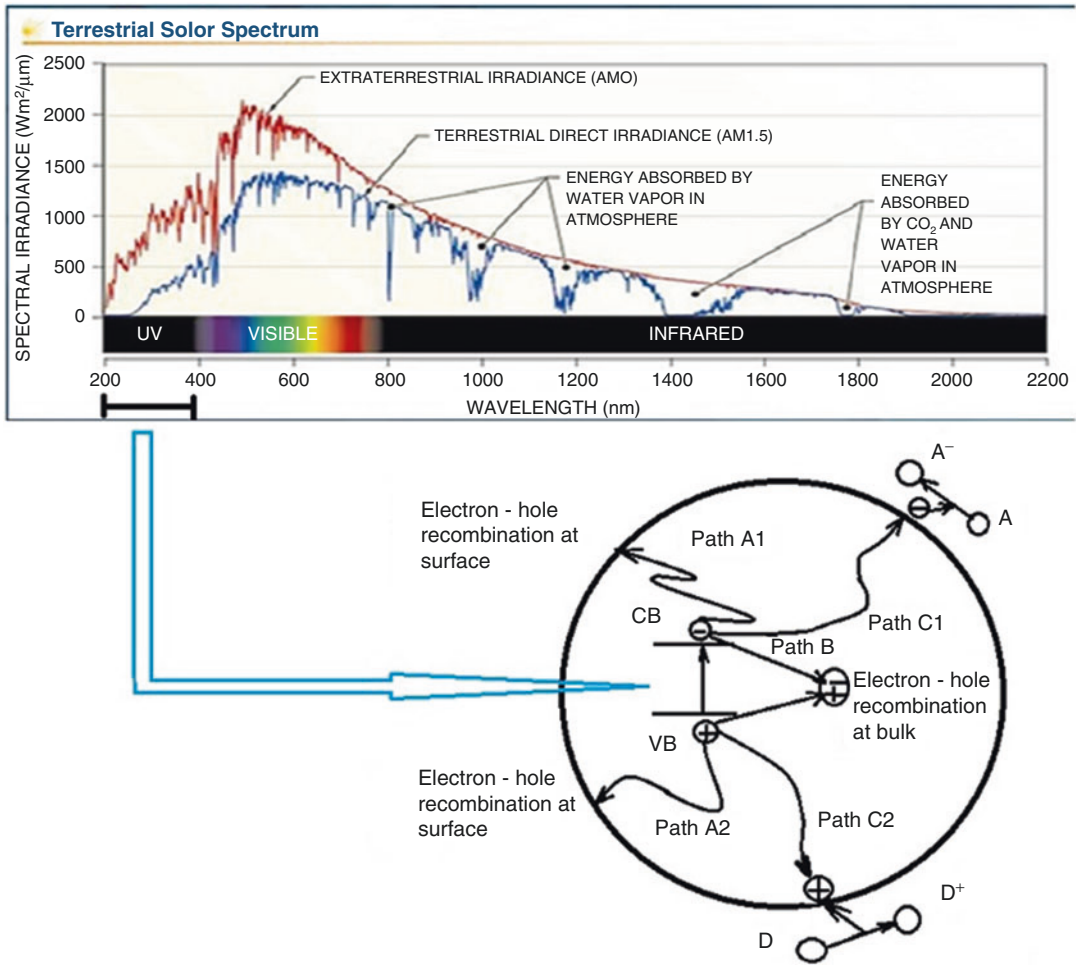
photocatalysts were undertaken by chemists, physicists, and chemical engineers.

It is generally accepted that the theoretical model of photocatalysis on TiO_2 surface consists of different consecutive steps where each one is essential for the activity and the efficiency of the photocatalyst [1, 2]. The initial step of the photocatalytic process consists in the generation of electron–hole pairs upon irradiation of the material with photons of energies that are at least equal to those of the band gap values. In this step, parameters like the relation between adsorption/reflection coefficients and the quantum efficiency of the electron–hole generation are very important. The ability to measure the actual absorbed light is very difficult in heterogeneous systems due to the scattering of light by the semiconductor surface. Due to the high absorbance coefficient for $\lambda < 400$ nm ($E_{phot} \sim \Delta_{Egap}$), it is usually assumed that all the light is absorbed.

In the second step, the formed electron–hole pairs can either recombine in the bulk or diffuse to the surface, where they can participate in chemical reactions. In this step, the lifetime and the velocity of the electron–hole recombination are crucial. Charge carrier traps are used to promote the trapping of electrons and holes at the surface leading to recombination suppressing and to more efficient charge transfer processes.

The next step is the creation of $H^{\bullet}$* and $OH^{\bullet}$* radicals as a result of the electron–hole interaction with water and, finally, a probable multiple step reaction of organic compounds with active radicals. In this step, the hydrophilic properties of the photocatalytic surfaces are crucial. The electron transfer process is more efficient if the species are pre-adsorbed on the surface. The cyclic process of photocatalysis beginning with electron–hole excitation by light is shown in Fig. 4, schematically.

An important issue in heterogeneous photocatalysis is the role of surface defects in determining the activity and the selectivity of the transition metal oxide-based catalysts. The existence of reduced surface centers and, hence, the presence of oxygen vacancies are sometimes necessary for any activity. Ultimately, the potential of transition metal oxides in an application is related to the ability of the metal atoms to assume multiple oxidation states. For example, on TiO_2 surfaces, various unsaturated Ti^{3+},Ti^{2+},Ti^{+} cations can be generated by vacuum annealing, Ar^{+} sputtering, chemical reduction (using H_2), or high-energy ball milling. Oxygen vacancies at the oxide surfaces and in the bulk alter the geometric and electronic structure and, thus, the chemical properties of these oxides. The expectation that the understanding of fundamental properties in the microscopic scale of these defects will



Influence of Defects in Photocatalysis, Fig. 4 Schematic photoexcitation and de-excitation events during photocatalysis

help to elucidate their effects on the catalyst functionality has been the driving force of experimental and theoretical studies of reduced materials.

There are two basic approaches developed over last decades – the first is based on AFM, STM, SXRD (surface X-ray diffraction), and LEED (low-energy electron diffraction) methods to understand the reconstruction of various facets and defects containing surfaces in the atomic scale [6, 7] and – the second is based on the obtained data from this reconstruction, by using DFT (density functional theory) to acquire the knowledge about the electronic structures of these surfaces [8].

Titanium dioxide crystallizes in three major different structures – rutile (tetragonal, D_{4h}^{14} – $P4_2/mnm$, a = b = 4.584 Å, c = 2.953 Å), anatase (tetragonal, D_{2h}^{19} – $I4_1/amd$, a = b = 3.782 Å, c = 9.502 Å), and brookite (rhombohedral, D_{2h}^{15} – Pbca, a = 5.436 Å, b = 9.166 Å, c = 5.135 Å) [6]. Other structures exist as well; however, only rutile and anatase play significant roles in applications of TiO_2. Both TiO_2 structures are wide band gap semiconductors. The energy gap for rutile is equal ~3.0 eV and for anatase is 3.4 eV. In both structures, the basic building block consists of a titanium atom surrounded by six oxygen atoms in more or less distorted octahedral

configurations. In each structure, the two bonds between the titanium and the oxygen atoms at the vertices of the octahedrons are slightly longer. The total energy of periodic TiO_2 slabs was calculated (see [6] and references therein) using a self-consistent ab initio method. The (110) surface has the lowest surface energy, and the (001) surface the highest. The thermodynamic stability of the (100) surface was found to be stable with the respect to forming (110) facets and the (001) surface was found almost unstable with the respect to the formation of macroscopic (1×1) (011) facets. TiO_2 (110) facet reconstruction is now well accepted as quintessential model of oxide surface and is being used routinely as a starting point for exploring new phenomena on oxides. The combination of spectroscopic, diffraction, and imaging techniques, applied by many different groups worldwide over the course of last decades, has delivered a very complete picture of the rutile (110) surface. The stoichiometric surfaces and many of the different configurations of defects are now well characterized [6, 7]. Other orientations of rutile and of anatase TiO_2 are still less well understood. For anatase, the (101) and the (100)/(010) surface planes are found in powder materials, together with some (001) planes. The anatase (101) surface was calculated to have the lowest surface energy, even lower than the rutile (110) surface [6]. Figure 5 shows the bulk truncated (110) and (101) oriented (1×1) surfaces of rutile and anatase [8].

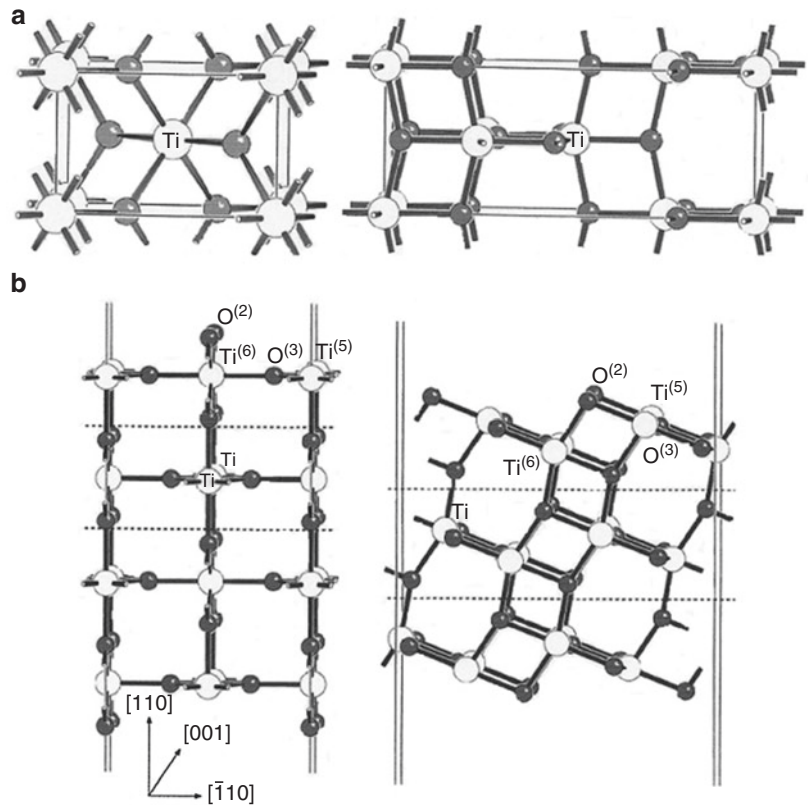
Both surfaces have sixfold- and fivefold- coordinated Ti atoms ($Ti^{(6)}$ and $Ti^{(5)}$), as well as threefold- and twofold-coordinated O atoms ($O^{(3)}$ and $O^{(2)}$). The removal of the latter, so-called bridging oxygen atoms, is the subject of much interest and debate [8].

Experimental Results on Oxygen Vacancies

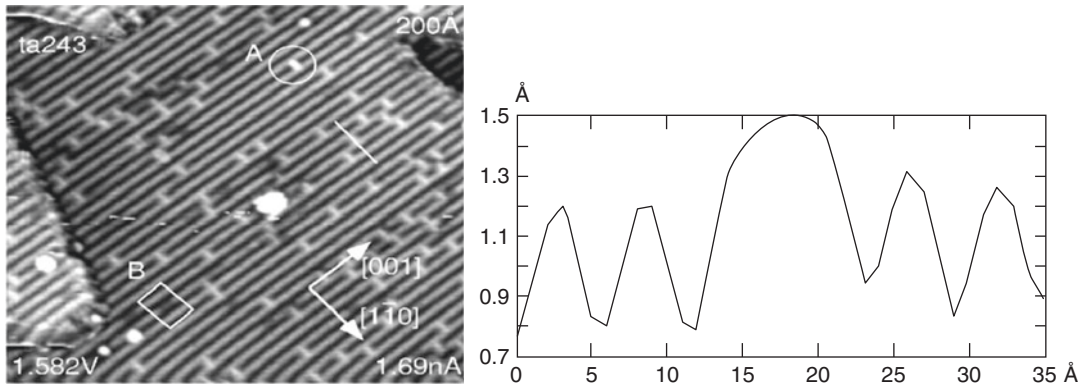
Experimentally, it was found that the titanium–oxygen phase diagram is very rich with a variety of crystal structures [6, 7]. The defect structure of TiO_{2-x} is a complex issue because a family of stable Magnéli phases is found to occur between Ti_2O_3 and TiO_2. At low concentrations of oxygen vacancies ($x < 10^{-4}$), they appear as point defects, while for higher degrees of reduction, crystallographic shear (CS) planes are formed. Reduction of TiO_2 occurs by heating in vacuum accompanied by color changes that are attributed to conduction electrons trapped at Ti sites forming EPR(electron paramagnetic resonance)-active Ti^{3+} species. The characteristic EPR signal attributed to Ti^{3+} ions was also observed after excitation of anatase TiO_2 samples with UV light. Reduced TiO_2 crystals exhibit n-type conductivity. There are overwhelming spectroscopic and chemical evidences of presence of oxygen defects in samples of rutile and anatase annealed in vacuum. They are attributed to vacancies of bridging oxygen atoms with concentrations of several percents. Anatase surfaces show, however, fewer tendencies to release of lattice oxygen atoms. As an example the (110) surface of rutile obtained by STM [6] is shown in Fig. 6.

There are two kinds of atomic scale features [5]. These appear as short bright features centered on dark rows (label A) and connecting neighboring bright rows and as dark spots on the bright rows (label B). The feature labeled A have a density of (7 ± 3) % per surface unit cell, consistent with the O vacancy density estimated from spectroscopic results [6]. The size of the bright spots is of order ~5 Å. When the sample was exposed to oxygen atmosphere, the density of the bright spots (A) decreased. This leads to the assumption that the bright spots can be attributed to oxygen vacancies. The dark features (B) are less common (surface concentration of ~(1–2) %) than oxygen vacancies, and their nature is yet unclear [6].

The ultraviolet photoemission spectroscopy (UPS) spectra of single crystal surfaces contain a band gap feature at ~1 eV below the conduction band edge which has been interpreted in terms of occupied Ti 3d states that can be supported by resonant behavior of UPS across the range of photon energies corresponding to the Ti 3p–3d excitation threshold [8] (and references therein). Moreover, UPS measurements on (110) rutile surface reduced by Ar bombardment showed a band gap state at 0.7–0.9 eV below the conducting band for low defect concentrations, moving toward the bottom of the conducting band as the defect concentration increases [8]. Furthermore, a shoulder



Influence of Defects in Photocatalysis, Fig. 5 (**a**) Bulk structures of rutile (*left*) and anatase (*right*) TiO_2 phases. (**b**) The bulk truncated (110) and (101) surfaces of rutile (*left*) and anatase (*right*), respectively. Shown are four- and three-layer (1 ×) slabs, respectively. The dotted lines indicate the repeating unit (one-layer) in the direction perpendicular to the surface (From Ganduglia-Pirovano et al. [8]. © 2007 Elsevier B.V)



Influence of Defects in Photocatalysis, Fig. 6 STM image (200 × 200 Å) of a TiO_2 (110) surface, sputtered and annealed in UHV to 1,100 K for 10 min, showing point defects. Features labeled "A" have been assigned as oxygen vacancies. The position of the line scan shown is indicated in the image (From Diebold [6].© 2002 Elsevier Science B.V)

in the Ti 2p XPS spectra can be attributed to Ti^{3+} ions, and the lack of surface conductivity can be interpreted as the localized nature of the additional electrons.

Now, when the key experimental features on the properties of the reduced systems are mentioned, one question of fundamental interest is what is the electronic structure of reduced

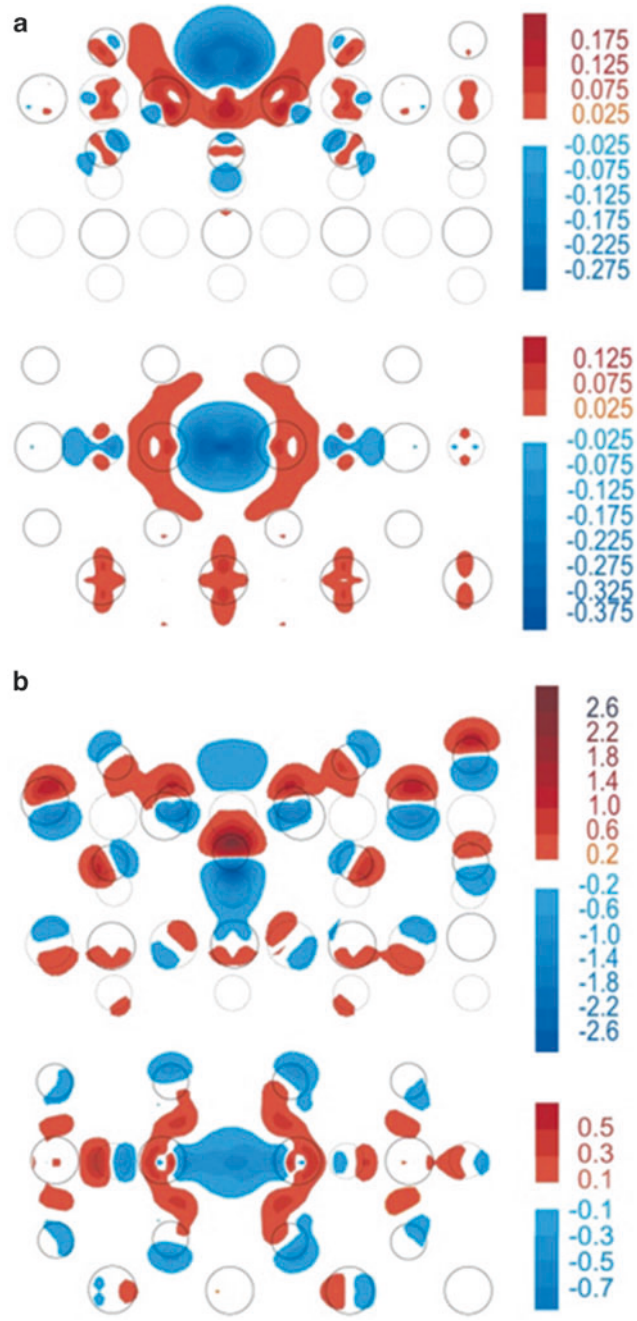
TiO_{2-x} system? In other words, when neutral oxygen atoms are removed, which states do the electrons, left in the system, and populate? What is the energy required to form the oxygen defects?

Theoretical (Numerical) Approaches on Reduced Systems

Here, some principal results obtained for reduced TiO_2 surfaces will be mentioned. Detailed information on the application of modern numerical methods of electronic structure calculations can be seen in [8] (and references therein). There are basically two different approaches to study solids, bulk, and surfaces. One is when the Schrödinger equation is numerically solved for the atoms of a unit cell containing a fixed number of atoms with periodic boundary conditions. Surfaces are modeled by repetition of a two dimensional periodic slab of the crystal together with a sufficiently thick vacuum layer. The presence of a point defect in the surface unit cell results in its periodic repetition and as a consequence leads to a high concentration of surface defects ($\Theta_{def} = 1$ that means the presence of a defect in each unit cell) that is much higher than those encountered in experiments [8]. Increasing the size of the unit cell in the surface plane reduces the defect concentration ($\Theta_{def} < 1$). Another approach is the use of cluster models where only a finite number of atoms is considered to describe a part of a solid system whereas the rest of the crystal is treated by embedding techniques. In this approach an infinite dilution of defects can be achieved. Density functional theory (DFT) is a widely used technique for the calculations of the properties of solids and molecules and can be used in both periodic and cluster approaches. Various exchange-correlations functional were used, including local density approximation (LDA), generalized gradient approximation (GGA), and hybrid functional such as Becke–Lee–Yang–Parr (B3LYP) and the Hartree–Fock (HF) along with other approaches that also have been used. There are some known failures in calculations caused by an application of LDA and GGA functional such as underestimation of band gaps, the gross overestimation of electron delocalization and metallic character, the underestimation of energy barriers for molecular reactions, etc. [8]. These problems are a consequence of employing approximations to the unknown exchange–correlation functional and the lack of knowledge about exact cancellation of the Coulomb self-interaction. These inadequacies can be partly removed by applying hybrid functional like B3LYP (or HSE03 and PBE0, see [8]). Compared to LDA and GGA functional, more accurate descriptions of the electronic structure have been achieved.

Reduced TiO_2 Rutile (110) Surface

There are two possible types of states that the electrons can occupy in TiO_2 upon reduction, namely, delocalized conduction band states formed by Ti 3d states, which leads to a metallic ground state, or localized 3d states in the band gap connected with insulating ground states that corresponds to the experimental observations (Ti^{4+} ions in TiO_2 are in d^0 state and Ti^{3+} ions formed by oxygen removal are in the d^1 state) [8]. There is an important aspect in the numerical approaches consisting in the consideration of the relaxation of the surface structure upon reduction. The surface model contained layers that all were allowed to relax, which means that all atoms can be displaced from the positions occupied in nonreduced model (these displacements are usually of about ~0.1 Å). The term *layer* is used for the repeating O – Ti_2O_2 – O unit (1 L) shown in Fig. 5. An increase in the computational resources allows to consider thicker (up to 7 L) slabs and lower defect concentrations (down to $\Theta_{def} = 1/6$). Among all periodic studies on reduced TiO_2 (110) that used gradient-corrected functional and ultrasoft pseudopotentials, here, the most systematic done by Rasmussen et al. (Ref. [18] in [8]) can be mentioned. The spin-polarized hybrid PW91 functional was used for calculations and investigated defect concentrations in the range $1/6 < \Theta_{deff} < 1/2$ using slabs of varying thickness (3L – 6L) and (1 × 6), (1 × 4), (1 × 3), and (1 × 2) and c(2 × 6), c(2 × 4), and c(2 × 2) unit cells [8]. The centered cells correspond to $\Theta_{def} = 1/6$, 1/4, and 1/2, respectively. The use of both primitive and centered cells of increasing sizes allowed for estimation of defect interactions



Influence of Defects in Photocatalysis, Fig. 7 Change in the charge density upon creation of a single defect in a $p(1 \times 4)$ cell with (**a**) unrelaxed and (**b**) relaxed geometries as calculated using PW 91. *Top* is a cut through the slab along the bridging oxygen rows. *Bottom* is a top view. The induced charge densities were integrated over the perpendicular direction to the plane shown, units are electrons/Å^2 (From Ganduglia-Pirovano et al. [8]. © 2007 Elsevier B.V)

along and across the bridging rows. The geometry of relaxations was also considered in detail [8]. Rasmussen et al. evaluated the actual changes in charge density upon defect creation and found that important changes in the electronic structure are strongly related with structural rearrangements. Figure 7a shows the induced charge density changes upon reduction for the (1×4) structure ($\Theta_{def} = 1/4$) with unrelaxed atom positions. The excess electrons distributed relatively near the defect site mostly at two closest surface located Ti atoms.

However, when atoms were allowed to relax, the excess electrons spread out (see Fig. 7b). The delocalization of the excess electrons increased upon relaxation; however, a significant amount still localized at the nearby Ti atoms. The calculated density of states of the reduced systems shows a defect-related occupied state right below the Fermi level, however, *not* separated by a gap from the conduction band as was found for other types of exchange– correlation functional [8]. The use of different approximations to exchange and correlation leads not only to considerable disagreements in the calculated electronic structure but also to considerable differences in defect formation energy [8]. Despite of these considerable variations, the most appropriate value of single defect formation energy at the rutile TiO_2 (110) can be estimated of ~3 eV [8].

Reduced TiO_2 Anatase Surfaces

Relative to the rutile (110) surface much less is known about anatase surfaces despite the much more applications in photocatalysis as a highly active substance. In correspondence to the rutile (110) surface, the reduced anatase (101) surface obtained by removal of twofold-coordinated oxygen atoms ($O^{(2)}$ in Fig. 5) was theoretically studied; however, hybrid studies have not yet been reported [8]. For such a study a partially reduced (101) surface using a 2L slab and a (1 × 3) unit cell with the spin-polarized PW91 functional was used [8]. One of the bridging oxygen atoms on the surface was removed, ($\Theta_{def} = 1/6$). The term *layer* corresponds to O_2-Ti_4O_4-O_2 unit (see Fig. 5). Atomic relaxations were allowed with the six outermost TiO_2 units fixed at their bulk positions [6]. The calculated defect formation energy was 4.7 eV (calculations that were represented by other groups lead to the values 4.38 and 4.4 eV). The relaxation produced distribution of the excess electrons similar to the results in Fig. 7 for the rutile (110) surface.

Some conclusions can be made from the above mentioned theoretical approaches to describe properties of the most stable surfaces of rutile (110) and anatase (101). Firstly, the relaxation patterns are complex for the rutile and anatase. For the rutile TiO_2 (110) surface, for example, atomic displacements are up to ~0.3 Å perpendicular to the surface as was found with gradient-corrected functionals [8]. Sizable relaxations extend up to the two outermost slab layers and lead to a *wider* defect site. This defect site can be considered as nano-sized object where excess electronic density can play an important role for catalytic and photocatalytic activity. Secondly, the energy of defect formation was estimated of about 2.7–3 eV for rutile and about 4.4–4.7 eV for anatase that are consistent with the experimentally observed lower tendency of anatase to release atomic oxygen under reducing conditions. Thirdly, the TiO_2 surface defect interactions are repulsive and are fairly strong along the [001] direction in rutile (110) surface. That corresponds to the fact that oxygen vacancies are observed as solitary nano-sized objects (see Fig. 6). Fourthly, it should be mentioned, however, that spin-polarized DFT (periodic) within the LDA or GGA approximations *fails* to correctly describe the experimentally observed localized nature of the excess electrons in reduced TiO_2 systems [8]. For the rutile TiO_2 (110) surface in particular, the localization of the excess electrons is consistent with theoretical studies at the Hartree–Fock (HF) level (periodic) and with hybrid approaches, namely, HFLYP and B3LYP, and one that combines an empirically fitted percentage of exact exchange with LDA (13 % Fock). Both B3LYP and the hybrid LDA +13 % Fock give the position of the occupied Ti 3d states upon reduction below the bottom of the conduction band (0.6–1.1 eV) that is close to the experiment [8].

In conclusion of this part, a hierarchical difference between two parts of modern knowledge about the properties of TiO_2 surfaces can be emphasized. There are sufficiently well-established reconstructions for the most stable surfaces of rutile and anatase, especially, the rutile (110) surface, including surfaces containing various surface defects [6, 7]. The results of theoretical modeling of the electronic properties of these surfaces, basically rutile (110) and anatase (101), are still dependent on the functionals used for modeling [8]. However, this situation is a point for the next step of studies of surface chemistry of various species adsorbed on TiO_2 surfaces.

Surface Chemistry of TiO_2

This field of the surface chemistry of TiO_2 is very active, with a recent emergence of theoretical calculations of adsorbate studies on various species – metals and nonmetals on TiO_2. It is evident that any modification of the TiO_2-based photocatalysts resulting in an improving of their efficiencies will represent a breakthrough in the field. For this purpose, many scientific works have appeared during recent years. A thorough analysis of the different approaches is beyond the scope of this issue and can be found, for example, in [6, 9, 10] (and reference therein); only some selected results will be summarized below.

One class of the reports is on the doping of TiO_2 with various transition metals (V, Cr, Mn, Fe, Co, Ni, Au, Ag, Pt, Nb, Cu, W, etc.) to lower the band gap energy. It was shown by ab initio band calculations based on the density functional theory (DFT) that an electron-occupied level occurs and electrons are localized around each dopant. As the atomic number of the dopant increases the localized level shifts to lower energy, thus, for Co the energy level is sufficiently low to be at the top of the valence band while other metals produce mid-gap states. In some experimental studies, it was shown that doping with metal ions such as V^{5+}, Cu^{2+}, Fe^{3+}, and W^{6+} leads to an enhanced photoactivity, whereas in other studies it was shown to lead to a reduction of the photoactivity. This fact reflects the situation that this area of modern research is not yet well established with common approaches.

Another class of reports is on the TiO_2 doping with nonmetals such as N, S, F, C, Br, or Cl. The main idea of nonmetals doping is to create electron–hole trapping centers or charge separating surfaces, as well as, to shift the optical absorption into the visible region.

A third class of reports is dealing with the increase of the adsorption of organic molecules by increasing the hydrophilicity under UV light illumination. It can be can supposed that, for example, on TiO_2:Fe surfaces, both of these mechanisms can be activated.

Metal and Metal Oxide Overlayers on TiO_2

It has been pointed out that metal reactivity toward oxygen is a very important parameter for metal overlayer or metal clusters growth on single-crystalline TiO_2 surfaces [6]. Generally, if a metal M is deposited onto TiO_2, then M should reduce the substrate and turn to be oxidized if the reaction

$$M + TiO_2 \rightarrow MO_x + TiO_{2-x}$$

is thermodynamically favorable (if the change of standard free energy ΔH_f^0 is negative). The formation of lower oxidation states from TiO_2 to Ti_2O_3 and from Ti_2O_3 to TiO is possible and ΔH_f^0 is equal to – 364 kJ/mol and – 483 kJ/mol for these reactions, correspondingly [6]. Thus, the ΔH_f^0 per mole of oxygen of oxide formation should be more negative than the above mention values, for such a reaction to take place. The heats of oxide formations were compiled by Campbell and are reproduced in Table 1 [6].

The strength of the reduction reaction scales with the reactivity of the metallic overlayer. Such a solid-state oxidation/reduction reaction was observed for Co but not for Ni (the heat oxide formation of ~250 kJ/mol, see Table 1). It should be taken into account that the values ΔH_f^0 for the TiO_2 to Ti_2O_3 and the Ti_2O_3 to TiO transformations were taken for the bulk values, so for the surface experiments, the smaller value can be acceptable. The oxidation/reduction reactions at the interface result in a substantial rearrangement of atomic positions and/or charge redistribution. Ionization of the overlayer metal introduces band gap states, or like in case of alkali metals, it leads to the formation of gap states [6]. Up to date only very few theoretical calculations of reactive metal adsorption on TiO_2 were done (see references in [6]). However, these studies took into account only one adsorption site (on top of fivefold coordinated Ti atoms) and no relaxation, which can be very important as it can be seen from the above discussions, in clear contradiction with the experimental results.

Growth morphology is an important subject of discussions about the mechanism of metal growth that can proceed in layer-by-layer fashion

Influence of Defects in Photocatalysis, Table 1 Heat of formations of the most stable oxide of the metal

ΔH_f^0 in kJ/mol of O	Metal
>0	Au
0 to −50	Ag, Pt
−50 to −100	Pd
−100 to −150	Rh
−150 to −200	Ru, Cu
−200 to −250	Re, Ni, Co, Pb
−250 to −300	Fe, Mo, Sn, Ge, W
−300 to −350	Rb, Cs, Zn
−350 to −400	K, Cr, Nb, Mn
−400 to −450	Na, V
−450 to −500	Si
−500 to −550	Ti, U, Ba, Zr
−550 to −600	Al, Sr, Hf, La, Ce
−600 to −650	Sm, Mg, Th, Ca, Sc, Y

(Frank–van de Merwe growth mode) or in cluster fashion growth (Volmer–Weber growth mode). Very rough simplification for understanding which mechanism will be preferable is comparing the difference between the surface free energy of the clean substrate, $F_{substrate}$, and surface free energy of the metal, F_{metal}, with interfacial free energy $F_{interface}$. If $F_{interface} > (F_{substrate} - F_{metal})$, cluster growth should take place. Because the surface energy of virtually all clean metals is higher than that of TiO_2 ($\sim$0.35 J/m^2), hence, cluster growth should occur [6]. The tendency to wet the substrate should also correlate with the numbers given in Table 1. The various types of morphology are shown in Fig. 8. The growth morphologies at the right side of the figure were deduced from combined XPS and LEIS (low-energy ion scattering) measurements [4].

The correlation between heat of oxide formation, interfacial reaction, and growth mode is very simple but neglect several important details. The growth modes discussed above are valid at thermodynamic equilibrium, but normally metal deposition is performed at low temperatures (mostly at room temperature) of the substrate for the vacuum evaporation technique. The substrate is considered rigid and its structure and stoichiometry stable upon metal deposition. Atomically sharp interface are assumed, while highly reactive metals clearly extract oxygen from the substrate. So that different morphology of metallic layers can be obtained by usage of different methods of deposition, chemical vapor deposition (CVD), for example.

Inorganic Molecules

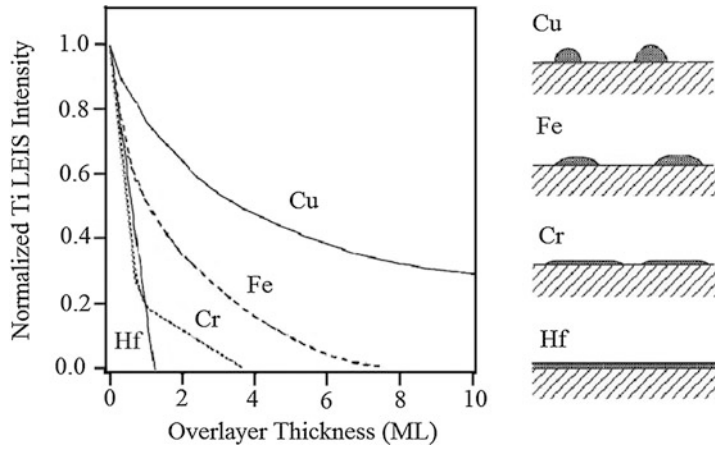
The interaction of TiO_2 with molecules of hydrogen, water, oxygen, carbon monoxide, carbon dioxide, and a lot of other species is a very important subject of studies for catalysis and photocatalysis. Here only limited important cases will be considered, and for more comprehensive review, [6, 10] (and references therein) can be recommended.

Hydrogen

The reduction of metal oxide powders in H_2 atmosphere is typically employed in catalysis. It was shown, however, that TiO_2 surfaces do not interact strongly with molecular hydrogen. High doses of H_2 (10^5 L) at room temperature caused additional emission peaks in the valence band region in ultraviolet photoemission spectra (UPS) [6, 7]. It was also noted that oxygen vacancies act as special adsorption sites for hydrogen. Despite the weak interaction between molecular hydrogen and TiO_2, reduction reactions occur when a TiO_2 surface is annealed in H_2 atmosphere. In contrast to molecular hydrogen, atomic H sticks to the TiO_2 surface at room temperature.

Water

Water vapor, in the ambient, interacts with TiO_2 surfaces, and surface hydroxyls can easily affect adsorption and reaction processes. The problem of the rutile 110 surface activity was recently studied both experimentally and theoretically. Based on the experimental results, it has been variously proposed that water (a) adsorbs molecularly on the 110 surface and only dissociates at defect sites, (b) adsorbs dissociatively at low coverages and, thereafter, molecularly, and (c) either adsorbs molecularly or dissociatively depending on temperature [6]. One point of consensus is that if dissociation does occur, it only occurs at small concentrations. It was recently calculated using the first principles of density functional method (DFT) that a monolayer of H_2O on the rutile 110 surface would contain water in both



Influence of Defects in Photocatalysis, Fig. 8 Trends for the attenuation of the Ti LEIS signals for four selected overlayers on TiO_2 (110) and schematic drawing of the initial stages of film growth for Cu (3D clusters), Fe (flat islands), Cr (2D islands followed by 3D growth), and Hf (formation of a continuous overlayer) (From Diebold [6].© 2002 Elsevier Science B.V)

molecular and dissociated forms. Based on these results, it is possible to conclude that the rutile (110) surface should be photocatalytically active. This photocatalytic activity that is lower than that for anatase was confirmed experimentally. However, these calculations do not reproduce other experimental results. Placing an H_2O molecule in a bridging oxygen vacancy did not lead to spontaneous dissociation on the TiO_2 (110) surface. Also it was concluded that single OH groups are not stable on TiO_2 (110) surfaces [6].

The adsorption of water on anatase surfaces has so far been mostly studied theoretically using DFT and first-principle molecular dynamic calculations. It was found by Selloni and coworkers (see reference [70] in [6]) that water adsorbs molecularly on an anatase (101) surface. Thermal-programmed desorption (TPD) and XPS studies confirm this prediction. Calculations of water adsorption on the anatase (001) surface predict dissociative adsorption, but they have not yet been confirmed by experiment.

Carbon Monoxide and Carbon Dioxide

Adsorption of CO on metal-promoted (covered with nano-sized Au clusters) TiO_2 surfaces is of particular interest in the context of low-temperature CO oxidation to CO_2 that has been discussed above. In contrast to metal-promoted surfaces, the adsorption of CO on the clean TiO_2 (110) surface is far less studied experimentally. It was found that oxygen vacancies can be special adsorption sites for H_2 and CO. The adsorption of CO on TiO_2 (110) was treated with different computational techniques, and the results are summarized in [6, 11]. It was found that adsorption with C-end down is energetically much more favorable than the alternative orientation. Throughout these calculations the C–O bond distances are rather similar, but the adsorption energies vary considerably from experimental results, being on the lower end of the given range [6, 11].

Carbon dioxide interacts only weakly with a clean rutile (110) surface. On the clean surface, CO_2 was found to bind to oxygen vacancies slightly more strongly than to regular lattice sites, with desorption temperatures in TPD at ~166 K, and ~137 K, respectively [6]. CO_2 adsorption is blocked by water; when H_2O is dosed on a CO_2-covered surface, the CO_2 is displaced.

Concluding this part of issue, it can be noted that hierarchically theoretical understanding of the interaction of surface with metals and even simple inorganic molecules is still, at present, less powerful for predictions, although this theory is based on detailed knowledge on surface structure and on less detailed knowledge about electronic structure of surface.

Heterogeneous Photocatalysis on TiO_2 Surface

General Considerations

In a heterogeneous photocatalysis system, photo-induced molecular transformations or reactions

take place at the surface of a catalyst. The reactions accomplished through photocatalysis can involve oxidation and oxidative cleavages, reduction, isomerization, polymerization, etc. Photocatalysis can be generally divided into two classes of processes [1, 2]. When the initial photoexcitation occurs in an adsorbed molecule, which then interact with surface in the ground state, the process is referred to as catalyzed photoreaction. When the initial photoexcitation take place in the catalyst surface and photo-excited catalyst then transfers an electron or energy into a ground state of molecule, the process is referred to as a sensitized photoreaction [1]. The initial excitation of the system is followed by electron transfer and/or energy transfers to the adsorbed molecule and subsequent de-excitation via chemical reactions in the heterogeneous photocatalysis process. Schematically this process is shown in Fig. 4.

The primary excitation process results in electron in the conducting band and a hole in the valence band. When TiO_2 is in contact with an electrolyte, the Fermi level equilibrates with the redox potential of the redox couple. The resulting Schottky barrier drives the electron and the hole in different directions. Electrons and holes when transferred across the interface are capable of reducing and oxidizing adsorbed species, forming a singly oxidized electron donor and a singly reduced electron acceptor illustrated by following reactions [6, 12]:

$$TiO_2 + h\nu \rightarrow TiO_2(e^-, h^+)$$

$$TiO_2(h^+) + RX_{ads} \rightarrow TiO_2 + RX_{ads}^{\bullet +}$$

$$TiO_2(h^+) + H_2O_{ads} \rightarrow TiO_2 + OH_{ads}^{\bullet} + H^+$$

$$TiO_2(h^+) + OH_{ads}^- \rightarrow TiO_2 + OH_{ads}^{\bullet}$$

$$TiO_2(e^-) + O_{2,ads} \rightarrow TiO_2 + O_2^{\bullet -}$$

$$TiO_2(e^-) + H_2O_{2,ads} \rightarrow TiO_2 + OH^- + OH_{ads}^{\bullet}$$

These processes result in anion or cation radicals which can undergo further reactions. Hydroxyl radicals are generally considered the most important species in photocatalytic degradation of organics. It is generally accepted that hole captures occur directly through OH_{ads}^- and not via water. It has been observed that photocatalytic activity of TiO_2 can be completely suppressed in absence of an electron scavenger such as molecular oxygen. In the absence of suitable electron or hole scavenger (adsorbed species as charge carrier trapping site), the charge carriers recombine at the surface (cf, path A1/A2 of Fig. 4) and/or in the bulk of the semiconductor (cf, path B of Fig. 4) to dissipate (heat) energy. On the other hand, when a suitable scavenger is available, the charges migrate to the surface of the TiO_2 particle and initiate interfacial redox reaction with the adsorbed species. Thus, TiO_2 photosensitizes the reduction of an electron acceptor (A) and the oxidation of an electron donor (D) forming anionic (A−) (cf, path C1 of Fig. 4) and cationic (D+) (cf, path C2 of Fig. 4) species. In nanostructured semiconductors, the surface-to-volume ratio is very high, and hence the reactive surface is also higher in nanomaterial regimes. Therefore, the surface adsorption and interfacial redox reaction can be enhanced by using nanostructured semiconductors. This semiconductor-assisted redox reaction is at the core of the heterogeneous photocatalysis. Because the conduction band of TiO_2 is almost completely isoenergetic with reduction potential of oxygen in inert solvents, adsorbed oxygen serves as an efficient trap for photogenerated electrons. The resulting species, superoxide, $O_2^{\bullet -}$, is highly active and can react with other adsorbed molecules [6]. The electron–hole recombination occurs very fast, resulting in low quantum efficiency that is one of the main impediments for use of TiO_2.

Dynamics of Photocatalytic Processes

In an aqueous solution, picoseconds and nanosecond laser photolysis was employed to study the dynamics of charge carrier trapping and recombination in ~60 Å colloidal particles of TiO_2 (see Ref. [30] in [1]). The trapping of the conduction band electrons is a very fast process and occurs within 30 ps. By contrast, the trapping of the valence band holes is a much slower process and takes about 250 ns. The recombination of trapped

electrons with free or trapped holes takes place in the 10^{-11} s to 10^{-6} s time interval. At low charge carrier concentrations, the mean lifetime of a single electron – hole pair – is ~30 ns in a TiO_2 particle, and the hole trapping can compete with recombination process [1]. In the trapped state, holes are relatively less reactive in comparison with trapped electrons. At a high concentration of electron–hole pairs, the charge carriers recombine within ~1 ns. Therefore, the interfacial carrier trapping should be very rapid in order to achieve effective photochemical conversion. This can only happened when the electron and hole trapping species are pre-adsorbed on the catalyst surface.

Photocatalysis on Single-Crystalline TiO_2

In UHV-based (UHV – ultra-high vacuum) photocatalytic studies, vacancies in the bridging oxygen positions of TiO_2 (110) play a major role [6, 7]. As discussed above, these vacancies serve as adsorption sites for O_2 molecules. Depending on surface preparation, two different defect states, termed "α_1" and "α_2" state, can be created on the TiO_2 (110) surface, which are both active for CO to CO_2 oxidation. The threshold for photoexcitation is 3.1 eV, indicating that excitations within the TiO_2 substrate are important.

Nanostructured TiO_2 Materials

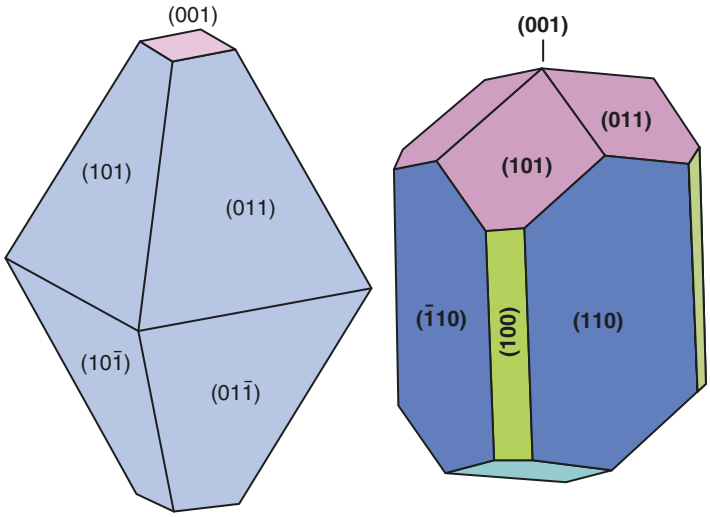
Numerous applications require TiO_2 to be in contact with a solid, liquid, or gaseous electrolyte; consequently, TiO_2 becomes a prime candidate for the development of nanoscale architectures. The properties of TiO_2 permit to use this material quite well to building tiny structures in all the sorts of sizes and shapes [13–15]. For example, TiO_2 nanopowder (size ~10 nm) of anatase crystal modification is typically produced in sol-gel process, where a titanium alkoxide or halide ($TiCl_4$, TiF_4) is hydrolyzed, often in presence of a template such as nano-spheres, nano-rods, or porous media.

Nanopowders

Based on the calculations of surface energies of various orientations and knowledge about thermodynamic stability, the equilibrium shape of macroscopic TiO_2 crystals was constructed for rutile and anatase crystals, as shown in Fig. 9.

The relation between bulk and surface energy in nanomaterials is different from that for the crystalline materials of macroscopic size, being surface energy comparable with that of bulk for nanoparticles. As a result the shape of nanoparticles can be different from that of macro-crystals shown in Fig. 9; however, some low-energy terminals should be present. TiO_2 nanoparticles have specific advantages in the enhancement of light and molecule adsorption due to the large fraction of surface atoms. The particle size at which the interface enhancement of the adsorption becomes significant is around 20 nm. Specific surface area can be calculated by using the Brunauer, Emmet, and Teller (BET) equation through nitrogen adsorption isotherms collected at 77 K. Typical specific areas for TiO_2 particles ranged from ~100 to 50 m^2/g, for the particle sizes from ~1 to 20 nm correspondingly. The surface and bulk irregularities naturally occur during the preparation process, and they present very different types. The nature of surface defect sites depends on the method of the chemical preparation. One of the most often used in photocatalysis is the industrially prepared powder Degussa P25. The technology of preparation of P25 is the gas-flame synthesis. Degussa P25 TiO_2 powder consists of particles with average size of the particles ~20 nm and represents a mixture of the anatase/rutile and amorphous forms with very different photocatalytic activities with the anatase form being much more active than the rutile form.

It is known that nanopowder particles are complex objects and contain pores of various sizes that can be classified as mesopores and as micropores. An adsorption measurement exhibits several structural features. It is possible to conclude the particle shape from the shape of the isotherm and the porous structure from the type of the hysteresis. Porous structures of nanoparticles strongly depend on conditions of heat treatment. A strong increase of adsorbed gas molecules occurs up to a relative pressure of $p/p(0) = 0.05$ where a monomolecular coverage is reached. The higher the quantity of adsorption is at this point, the bigger is the specific



Influence of Defects in Photocatalysis, Fig. 9 The equilibrium shape of a macroscopic TiO_2 crystals, anatase (*left*) and rutile (*right*) (From Diebold [6].© 2002 Elsevier Science B.V)

surface area. Typical adsorption branches are shown in Fig. 10.

At higher relative pressures the adsorption branches consist of a small increase up to p/p (0) = 0.7. The slope of the curve depends on the mesoporous structure. In mesopores it is possible to build up a multimolecular coating, while in micropores it is not. The curve of the powder treated at 600 °C in vacuum shows a bigger slope than the 200 °C in vacuum. The conclusion is that powders annealed at higher temperatures include more mesopores, which is also illustrated in Fig. 11.

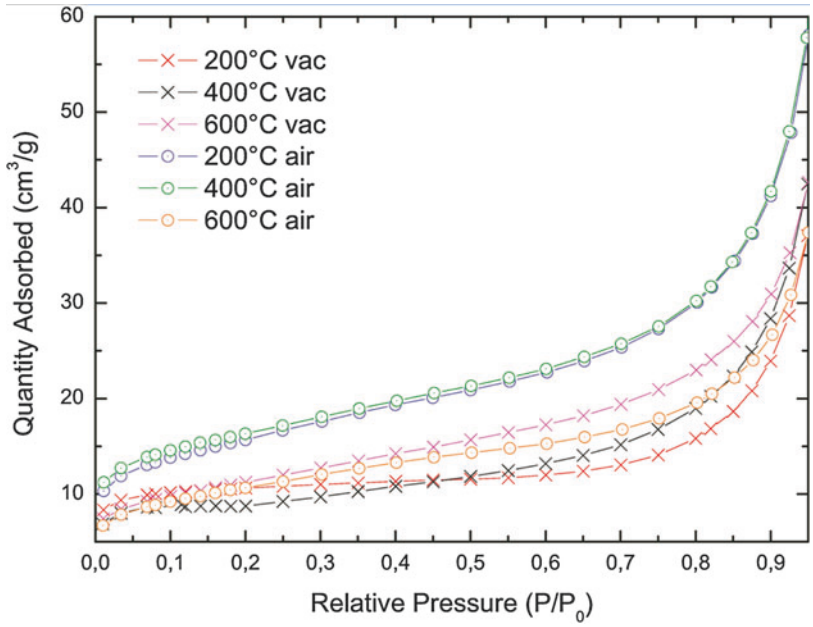
At relative pressures near 1, structural information for the external surface area and macropores is included, because only for these structures a further adsorption is possible. The pore size distribution, Fig. 11, shows a maximum of pores between 1.7 and 4 nm. The air-treated powders up to 600 °C contain a higher amount of detectable micropores, between 1.7 and 2.0 nm diameter, as well as mesopores.

Modified Nanopowders

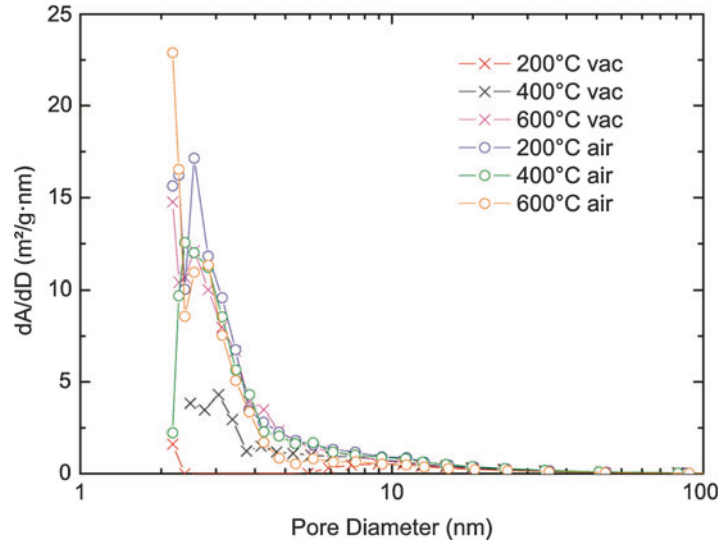
As it can be seen in Fig. 4, only an ultraviolet part of solar spectrum can be used for photocatalytic applications. In order to increase the limited optical adsorption of TiO_2 under sunlight, there have been persistent efforts to vary the chemical composition of TiO_2 by adding controlled metal or nonmetal impurities that generate intermediate donor or acceptor states in the band gap. Different from impurity incorporation, vacuum annealing that produces Ti^{3+} species in TiO_2 has also been made. Through doping, the solar adsorption characteristics of TiO_2 have been improved. However, this improvement not always was accompanied by improvement in photocatalytic activity. The optical adsorption of TiO_2 can be modified as the result of electronic transitions from the dopant 2*p* or 3*p* orbitals to the Ti 3*d* orbitals. At present, nitrogen-doped TiO_2 exhibits the greatest optical response to solar radiation.

An alternative approach to improve visible and infrared optical adsorption by engineering the disorder nanophase TiO_2 with simultaneous dopant incorporation has been developed [6]. In its simplest form, a disorder-engineered nanophase TiO_2 consists of two phases: a crystalline TiO_2 quantum dot or nanocrystal as a core and a highly disordered surface layer where dopants are introduced, Fig. 12 [6].

To introduce disorders into nanophase TiO_2, first a porous network of TiO_2 nanocrystals with a few nanometers in diameter was created. A disordered layer on the nanocrystal surface was then generated by hydrogenation [6]. A shift in the onset of adsorption in such disorder-engineered TiO_2 nanocrystals, from the ultraviolet to near-infrared, has been observed. This shift was accompanied by a dramatic color change and substantial enhancement of solar-driven photocatalytic activity.



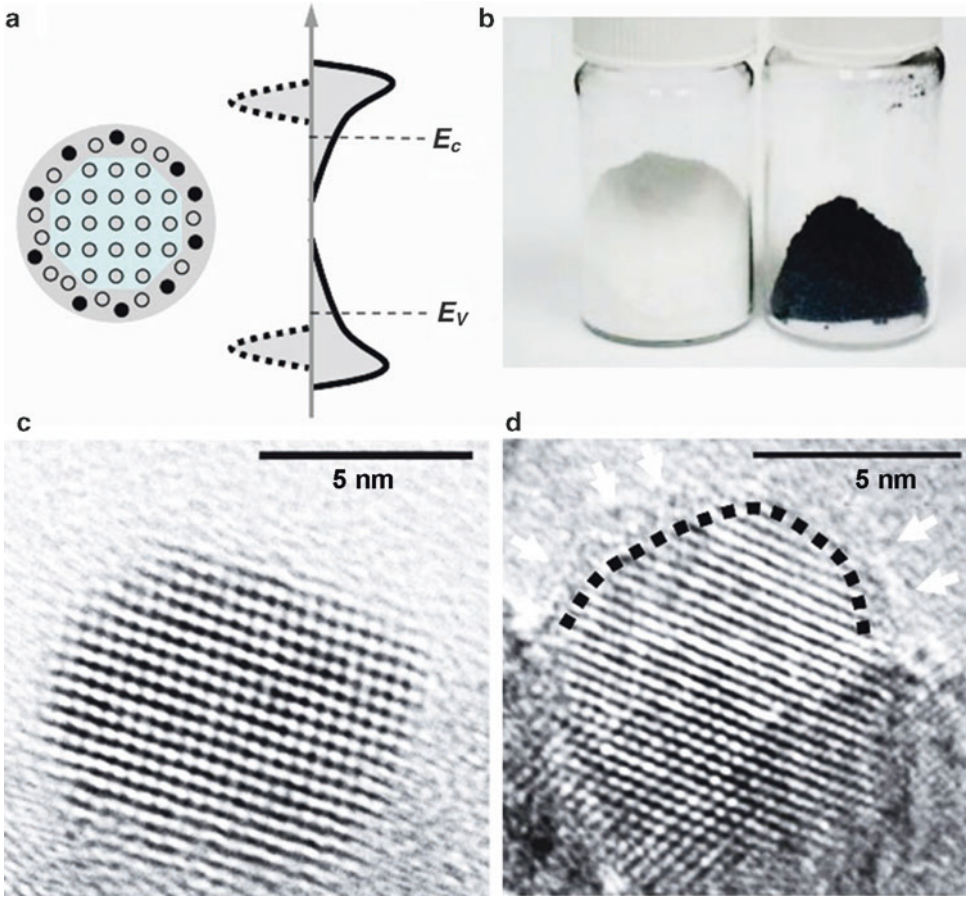
Influence of Defects in Photocatalysis, Fig. 10 Adsorption branches of N2-adsorption measurements at 77 K (From Bessergenev [16]. © 2015 Elsevier Science B.V)



Influence of Defects in Photocatalysis, Fig. 11 Pore size distribution of selected samples (From Bessergenev [16]. © 2015 Elsevier Science B.V)

Conclusion

In heterogeneous photocatalysis, the efficiency for chemical transformation depends on the photoactivity of the adsorbed molecules and the catalyst substrates. The catalyst only provides an ordered environment for a more efficient molecular interaction and does not participate in the chemical reaction process. The role of the semiconductor, in particularly of wide band gap TiO_2, is either to provide energy levels to mediate electron transfer between adsorbed molecules, by temporarily accommodating an electron, or behave as both electron



Influence of Defects in Photocatalysis, Fig. 12 (**a**) Schematic illustration of the structure and electronic DOS of a semiconductor in the form of a disorder-engineered nanocrystal with dopant incorporation. Dopants are depicted as black dots, and disorder is represented in the outer layer of the nanocrystal. The conduction and valence levels of a bulk semiconductor, EC and EV, respectively, are also shown, and the bands of the nanocrystals are shown at the *left*. The effect of disorder, which creases broadened tails of states extending into the otherwise forbidden band gap, is shown at the *right*. (**b**) A photo comparing unmodified white and disorder-engineered black TiO_2 nanocrystals. (**c**, **d**) are HRTEM images of TiO_2 nanocrystals before and after hydrogenation, respectively. In (**d**), a short *dashed curve* is applied to outline a portion of the interface between the crystalline core and the disordered outer layer (marked by *white arrows*) of *black* TiO_2. (From Chen et al. [17] © 2011 American Association for the Advancement of Science)

donor and electron acceptor (by photogenerated hole in the valence band). Defects of various types can play very important roles in enhancing the photocatalytic activity by creating additional electron levels that take positions between the valence and the conduction bands. Defects create local changes in potential energy, which can be considerated as quantum well (or quantum dot), where an excess electron can occupy additional levels of energy. Oxygen vacancies and various dopant (metallic and nonmetallic) inclusions are examples of such defects. Taking into account lattice relaxations, the sizes of these defects are in the nanometer range. Metallic nanoclusters on the catalyst surface are another example of nanoscaled objects where transitions between continuous and discrete electron spectra take place.

References

1. Linsebigler, A.L., Lu, G., Yates Jr., J.T.: Photocatalysis on TiO_2 surface: principles, mechanisms, and selected results. Chem. Rev. **95**, 735–758 (1995)
2. Schneider, J., Matsuoka, M., Takeuchi, M., Zhang, J., Horiuchi, Y., Anpo, M., Bohnemann, D.W.: Understanding TiO_2 photocatalysis: mechanisms and materials. Chem. Rev. **114**, 9919–9986 (2014)
3. Gates, B.C.: Supported metal clusters: synthesis, structure, and catalysis. Chem. Rev. **95**, 511–522 (1995)
4. Goodman, D.W.: Catalysis by supported gold nanoclusters. Dekker Encycl. Nanosci. Nanotechnol. (2004)
5. Fujishima, A., Honda, K.: Electrochemical photolysis of water at a semiconductor electrode. Nature **238**, 37–38 (1972)
6. Diebold, U.: The surface science of titanium dioxide. Surf. Sci. Rep. **48**, 53–229 (2003)
7. Henderson, M.A.: A surface science perspective on TiO_2 photocatalysis. Surf. Sci. Rep. **66**, 185–297 (2011)
8. Ganduglia-Pirovano, M.V., Hofmann, A., Sauer, J.: Oxygen vacancies in transition metal and rare earth oxides: current state of understanding and remaining challenges. Surf. Sci. Rep. **62**, 219–270 (2007)
9. Astruc, D. (ed.): Nanoparticles and Catalysis. Wiley – VCH, Weinheim (2008)
10. Carpenter, M.A., Mathus, S., Kolmakov, A. (eds.): Metal Oxide Nanomaterials For Chemical Sensors. Springer, New York (2013)
11. Li, K., An, X., Park, K., Khraisheh, H., Tang, J.: A critical review of CO_2 photoconversion: catalysts and reactors. Catal. Today **224**, 3–12 (2014)
12. Nosaka, Y., Nosaka, A.Y.: Identification and roles of the active species generated on various photocatalysts. In: Pichat, P. (ed.) Photocatalysts and Water Purification: From Fundamentals to Recent Applications. Wiley – VCH, Weinheim (2013)
13. Banerjee, A.N.: The design, fabrication, and photocatalytic utility of nanostructured semiconductors: focus on TiO_2 – based nanostructures. Nanotechnol. Sci. Appl. **4**, 35–65 (2011)
14. Wang, X., Li, Z., Shi, J., Yu, Y.: One – dimensional titanium dioxide nanomaterials: nanowires, nanorods, and nanobelts. Chem. Rev. **114**, 9346–9384 (2014)
15. Fattakhova–Rohlfing, D., Zaleska, A., Bein, T.: Three – dimensional titanium dioxide nanomaterials. Chem. Rev. **114**, 9487–9558 (2014)
16. Bessergenev, V.G., Mateus, M.C., Botelho do Rego, A.M., Hantusch, M., Burkel, E: An improvement of photocatalytic activity of TiO_2 Degussa P25 powder. Applied Catalysis A: General **500**, 40–50 (2015)
17. Chen, X., Liu, L., Yu, P.Y., Mao, S.S.: Increasing solar adsorption for photocatalysis with black hydrogenated titanium dioxide nanocrystals. Science **331**, 746–750 (2011)

Influence of Surface Engineering on Toxicity of Nanoparticles

► Effect of Surface Modification on Toxicity of Nanoparticles

Influence of Surface Functionalization on Toxicity of Nanoparticles

► Effect of Surface Modification on Toxicity of Nanoparticles

Infrared Plasmonics Beyond Noble Metals

► Terahertz and Infrared Plasmonics with Unconventional Materials

Injectable Hydrogel

Moon Suk Kim[1] and Kinam Park[2]
[1]Department of Molecular Science and Technology, Ajou University, Suwon, South Korea
[2]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, IN, USA

Synonyms

Injectable in situ forming gel

Definition

Injectable hydrogel is based on the idea that certain biomaterial could be injected as a liquid and then form an in situ solid gel.

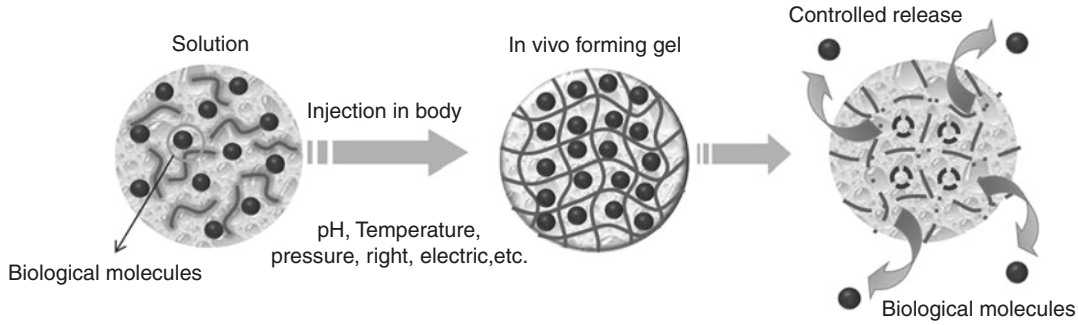
Overview

Systemic administration of a drug in a bolus, one of the most widely used methods of medication, frequently destroy healthy cells and tissues, leading to numerous undesirable side effects, such as loss of function or reduced efficiency of various organs [1]. One means to circumvent these limitations is drug delivery using sustained-release systems, which release medication at an appropriate rate for a defined duration to a desired place in the body. While a number of sustained-release strategies have been researched, hydrogels have drawn increasing attention as a candidate delivery system for the sustained release of medication. Hydrogels are composed of hydrophilic macromolecules bridged through hydrogen bonds, covalent bonds, van der Waals interactions, or physical bonds that form a three-dimensional macromolecule network structure. Because of their hydrophilic nature, such gels can swell in an aqueous environment and are highly biocompatible, exhibiting characteristics similar to live tissue when fully hydrated.

Hydrogels designed to gel in response to a given stimulus are manufactured with a large surface area so that water-diffusion distances are short; accordingly, they exhibit very rapidly induced water-swelling characteristics. Factors that have been explored as external gelation stimuli include physical factors, such as temperature, light, electricity, pressure, and sound, as well as chemical stimuli, such as changes in ionic strength or composition, molecule recognition, and covalent binding. Hydrogels with such stimulus-induced characteristics can make a phase transition from solution to gel (sol-gel) through reversible volume changes in a matter of minutes or even seconds. Some of the stimuli described above, notably temperature, can also be applied to stimulate hydrogel formation within the human body. Hydrogels based on such a system are associated with minimal side effects and hold tremendous promise as a drug-delivery system and biomaterial for regenerative medicine applications. A large number of macromolecules have been reported as potential stimulus-induced hydrogels. The temperature-induced hydrogel can be administered as an injection and is the most widely researched in several biomedical applications. Specially, hydrogel solution at room temperature can easily incorporate various biological molecules by simple mixing and offer the advantage of simple injection procedures, thus facilitating patient compliance and quality of life via comfort (Fig. 1) [2]. This entry focuses on the types of injectable hydrogels that show a characteristic physical phase-transition response to temperature and highlights their potential human applications.

Key Principles

Among various macromolecules that are the subject of ongoing research are sensitive to temperature stimuli; specifically, they yield macromolecule solutions that exhibit phase-transition characteristics in response to changes in temperature. Macromolecules with hydrophilic functional groups are soluble in aqueous solutions or swell in water and generally increase in water solubility as the temperature increases. As hydrophobic components, such as methyl, ethyl, and propyl moieties, are introduced to hydrophilic functional groups, the macromolecule acquires a lower critical solution temperature (LCST), such that water solubility decreases with increases in temperature. The hydrogel produced from such macromolecules form gels at a temperature higher than the LCST; these types are called thermosensitive hydrogels. Macromolecules composed of hydrophilic and hydrophobic segments in diverse ratios exhibit a phenomenon in which the hydrophilic segments dissolve in water due to the dominance of hydrogen-bonding strength. But as the temperature rises, the cohesive strength of the hydrophobic component dominates over the hydrogen-bonding power, forming a hydrogel through cohesion. Such basic dynamics can be used to modulate LCST by controlling the length of the unit molecule, that is, by controlling the relative ratio of hydrophilic and hydrophobic segments.



Injectable Hydrogel, Fig. 1 Schematic image of injectable hydrogels

Basic Methodology

Types of Injectable Hydrogels in Synthetic Polymers

Poly(*N*-Isopropylacrylamide)

Poly(N-Isopropylacrylamide) (PNIPAAm) is a representative and widely used LCST material, exhibiting a sol-gel phase transition of 32 °C. The LCST of PNIPAAm can easily be controlled by conjugating various hydrophobic molecules. Combining highly hydrophobic copolymers reduces the LCST. For example, poly(*N*,*N*-diethylacrylamide) (PDEAAm) is a temperature-sensitive macromolecule composed of a vinyl group, an isopropylacrylamide group, and a carboxyl group that has an LCST in the range of ~25–35 °C. Poly(*N*-(L)-(1-hydroxymethyl)propylmethacrylamide) is also a temperature-sensitive macromolecule with optical activation properties. Poly(*N*-acryloyl-*N*'-alkylpiperazine) is a macromolecule that is sensitive not only to temperature but also to pH, with an LCST of approximately 37 °C. PNIPAAm and its derivatives can be utilized in various ways on demand [3].

Poloxamers

A triblock copolymer of poly(ethylene glycol) (PEG) and poly(propylene glycol) (PPG) in the form of PEG-PPG-PEG is called a poloxamer (ICI) or pluronic (BASF) [4]. The PEG/PPG ratio of poloxamers ranges from 1:9 to 8:2, and the molecular weight varies widely (from 1,100 to 14,000). An aqueous poloxamer solution exhibits a phase transition at ~4–30 °C, reflecting transition from single molecules to a micelle as the temperature increases. The poloxamer hydrogel is a structure in which the micelle is packed according to certain structural changes. Such characteristics cause the structure to easily erode in the presence of an excess of water, resulting in short-duration (~1 day) hydrogels that are not suitable for use in long-term applications. Depending on the lengths of PEG and PPG blocks, the temperature or temperature range of the sol-gel phase transition and the minimum concentration required for gel formation can be controlled.

A PEO-b-PBO copolymer, which uses poly (butylene oxide) (PBO) instead of PPO, has also been studied [5]. PBO is more hydrophobic than PPO so that the minimum threshold temperature for the copolymer is very high. Other materials, such as poly(L-lactic acid) (PLLA) and poly(D, L-lactic acid-co-glycolic acid) (PLGA), can also substitute for the PPO block.

PEG-Polyesters

PEG-PLGA-PEG was developed using the biodegradable macromolecule PLGA in place of the PPO used in a poloxamer to create a stable, long-lasting injectable gel system [6]. This triblock copolymer is in a sol state at room temperature and forms a gel at body temperature. In an aqueous solution, single PEG-PLGA-PEG molecules pack into micelles and the triblock copolymer exhibits a transition to gel as the micelles condense. As the temperature increases, the number of molecules per micelle and the size of micelles increase, and the gravitational force

between macromolecules increases to form a gel. Unlike the poloxamer, this hydrogel can decompose within the body, and the formed hydrogel can last over a month, allowing it to be used in long-term applications. The sol-gel transition temperature is very important in determining the suitability of such triblock copolymers as bioactive molecules delivery system; importantly, the transition temperature can be modulated by adjusting PLGA and PEG lengths.

One drawback of a PEG-PLGA-PEG system is that the copolymer is difficult to prepare as an aqueous solution. This limitation is largely overcome by introducing poly(caprolactone) (PCL), a biodegradable crystalline macromolecule to yield PEG-PCL [7]. An aqueous solution of PEG-PCL shows a concentration-dependent sol-gel transition as the temperature increases within a range of 20–60 °C. While the physical form of previous temperature-sensitive macromolecules is similar to that of taffy, PEG-PCL is in the form of a powder at room temperature. This valuable property allows PEG-PCL block copolymers to dissolve as an aqueous solution in a relatively short period of time. PEG-PCL has the same advantage, but because of its high degree of crystallization, it tends to harden into a gel from the sol state at room temperature within an hour. Thus, although it is easy to incorporate biologic molecules in the macromolecule aqueous solution and inject it, the entire process must be completed rapidly.

PEG-Sebacic Acid Polyester (SA)

The viscosity of most temperature-sensitive macromolecules in a gel state is ~100–10,000 Pa, making them difficult to inject in gel form. PEG-sebacic acid polyester (PEG-SA), containing SA with a flexible methylene unit and PEG as a hydrophilic component, is a soft gel system that is injectable in sol form [8]. When prepared as an aqueous solution of an appropriate concentration, PEG-SA undergoes a phase transition from sol to gel at human body temperature. When a hydrophilic drug is delivered using hydrogel, the initial amount of drug released may increase during the transition from sol to gel. Therefore, a soft gel system that is injectable in gel form can be useful in delivering hydrophilic drugs.

Polyphosphazenes

Normal organic macromolecules have a backbone composed of C–C bonds. Such macromolecules are of limited use because the short covalent bonds between carbons allow for limited motion, causing stiffness of the chain and making it difficult to maintain elasticity and flexibility over a wide temperature range. These molecules also typically have limited resistance to oxidation or high temperature and are often inflammable. In order to overcome such limitations, researchers have explored various inorganic macromolecules. One such inorganic macromolecule is polyphosphazene, which has an N-P double-bonded skeletal structure that can be substituted with various groups, has a high stability against chemicals, is not very combustible, and maintains its function over a wide range of temperatures. It also degrades within a living body, so it can be applied in various ways as a biocompatible material.

Introducing hydrophilic and hydrophobic materials to poly(dichlorophosphazene) allows the synthesis of temperature-sensitive macromolecules with LCST behavior [9]. A temperature-sensitive macromolecule in which α-amino-ω-methyl-PEG as the hydrophilic material and DL-leucine ethyl ester(LeuOeEt), L-isoleucine ethyl ester, or L-valine ethyl ester as hydrophobic amino acids were introduced into poly (organophosphazenes) has been reported. This material is a biodegradable macromolecule with LCST behavior that shows sol-gel transition as the temperature increases. Physical bridges between macromolecules in solution are formed through hydrophobic bonding between hydrophobic amino acids, resulting in the formation of a gel. By changing features of the molecule or solution conditions, such as the composition of substitution groups, types, and lengths of amino acid esters and pH, the temperature range and characteristics of the gel can easily be controlled [10].

Types of Injectable Hydrogels in Natural Polymers

Polysaccharides

Naturally occurring biopolymers such as gelatin, agarose, and gellan benzyl ester are

temperature-sensitive materials. These form a double-spiral structure in aqueous solutions, and the points of contact between spiral structure strands create the physical bridges that allow hydrogels to form. Above the melting point, these biopolymers exist as aqueous solutions; as the temperature drops, they form hydrogels.

Gelatin is the protein generated by the breakdown of the triple-spiral structure of collagen into single-stranded molecules. Gelatin in solution, unlike other polysaccharides, changes from a random coil into a triple-spiral structure, establishing physical bridges at this time to form a gel. The hydrogel formed in this way is not stable to temperature changes, so it is not suitable for biomedical application. One strategy for creating a stable hydrogel from gelatin involves the introduction of chemical bridges that prevent the molecule from forming a double-spiral structure again, even after the temperature is lowered. However, this approach raises the issue of irreversible gel formation. An alternative method that has been suggested for forming a stable yet reversible hydrogel is the introduction of chitosan to gelatin.

Chitosan has a pKa of 6.5, so it is not soluble in water at a pH greater than 6.5. But the introduction of the hydrophilic PEG (>2,000 Da) increases the solubility of chitosan in water and allows a temperature-sensitive macromolecule that remains in solution at room temperature, but forms a hydrogel at 37 °C, to be produced.

Unlike most natural macromolecules, which form gels at a low temperature, cellulose derivatives exhibit sol-gel transitions as the temperature rises. Cellulose itself is not quite soluble in water, but can be made readily water soluble by introducing a hydrophilic component. By appropriately adjusting the hydrophobic/hydrophilic ratio of such a cellulose derivative, it is possible to produce a material that can transition from sol to gel.

Artificial Polypeptides

Dual copolymers composed of a hydrophilic block of poly(L-lysine) or poly(L-glutamate) and a hydrophobic block of poly(L-Leucine) or poly(L-valine) undergo sol-gel transition as the temperature increases and form hydrogels through micelles, similar to PEO-PPO block copolymers. The main differences between such dual polymers and existing synthetic macromolecules are formations of hydrogels at a relatively low macromolecule concentration and stability at a high temperature (90 °C). Using gene recombination methods, it is possible to prepare artificial polypeptides. Elastin-like polypeptide (ELP), composed of Val-Pro-Gly-Xaa-Gly amino acid repeat units, is a temperature-sensitive macromolecule produced using such a gene recombination approach. Any amino acid except proline can take the place of Xaa, which is a "guest residue." ELP exhibits a temperature-sensitive phase transition: It is in an aqueous state below the transition temperature, but above that temperature, hydrophobic interactions cause a conglomeration phenomenon that leads to precipitation.

Key Research Findings

Injectable Hydrogels as Drug Carriers

Oral or systemic administration of anticancer drug solutions results in the spread of drug throughout the body; only a very small amount is actually delivered to the cancer tissues where anticancer activity is needed. This leads to insufficient levels of drug at the tumor site, requiring dosing strategies that increase drug concentrations in normal tissues to toxic levels, causing serious systemic side effects. In addition, anticancer drugs generally have a low bioavailability and short half-life, so they must be injected frequently. In order to overcome these limitations of conventional chemotherapy, researchers have focused on drug-delivery methods that effectively deliver the injected drugs to the cancer tissues and consistently release the appropriate amount of anticancer drugs for the desired length of time.

Dramatic improvements in injectable hydrogel drug-delivery systems have set the stage for a groundbreaking turning point in anticancer treatment. These include the recent development of a delivery system that uses biodegradable injectable hydrogels to create a drug depot under the skin or

in cancer tissues that releases hormones or chemotherapeutic agents for cancer treatment.

Because injectable hydrogels are in aqueous solution at room temperature, it is easy to mix them with various anticancer drugs. Such a solution containing drugs and temperature-sensitive, phase-transitional macromolecules can simply be injected under the skin or into the cancer tissue, immediately forming a depot for anticancer drugs. Because no surgery is required, patient pain and discomfort are minimized. Oncogel, developed by MacroMed, uses PEG-PLGA-PEG (ReGel) to produce paclitaxel anticancer drugs in aqueous solutions [11]. Clinical trials of partial injection of Oncogel for solid tumors are currently ongoing in the USA and Europe.

Various injectable hydrogels composed of MPEG-polyester containing anticancer drugs (e.g., paclitaxel, docetaxel, doxorubicin) have been produced as aqueous solutions at room temperature for injection. In the system as currently applied, the hydrogels that form in situ upon injection create an anticancer drug depot that slowly releases anticancer drugs over a long period of time, eventually decomposing within the body. Once anticancer drugs have been completely released, the injected hydrogel is excreted from the body as a degradation product that is completely harmless to the human body.

In recent years, protein drugs developed as a result of advances in molecular biological recombination technology have been shown to possess excellent selectivity and efficacy while minimizing side effects. As such, they have received considerable attention for their potential to actually cure various diseases rather than merely treat symptoms. Although these advances have made wide commercialization of such protein drugs possible, these drugs, unlike low-molecular-weight drugs, have large, complex structures and are generally very unstable within a living body. Most protein treatments are currently administered systemically as repetitive injections, because protein drugs are easily degraded by first-pass metabolism when taken orally and thus exhibit a very short half-life and low bioavailability. These characteristics of protein drugs have motivated the development of systems to overcome the limitations of existing protein delivery systems. A number of strategies are currently being explored to achieve efficient systemic delivery with reduced injection frequency. Among the approaches that have been suggested as a solution for effective protein delivery is the use of biodegradable and injectable hydrogels.

Injectable hydrogels containing bioactive molecules like proteins has developed and has confirmed their long-term release characteristics in vivo. Importantly, the incorporated proteins are relatively stable, persisting for extended periods of time within the macromolecule.

Injectable Hydrogels as Transplants in Regenerative Medicine

Individuals in modern industrial societies face the prospects of destruction or escalation of organ tissues due to a variety of causes, including accidents and congenital organ diseases, creating a demand for organ-failure solutions. The ultimate solution is organ regeneration through tissue engineering, a topic that has received considerable recent research attention. Artificial organ development through tissue engineering consists of three important factors: (1) a regenerative cell source, such as stem cells; (2) a biomaterial for the production/maintenance of the organ form; and (3) a physiologically active substance, such a as cytokine or growth factor that promotes the growth of regenerative cells. Creating artificial organs from cells incorporated into biodegradable macromolecules requires that the cell-macromolecular complex does not provoke immunorejection, that an extracellular matrix is secreted to form the new organ, and that the macromolecule is gradually biodegraded and excreted as water and carbon dioxide. Injectable hydrogels have shown particular promise as a transporter for the production of tissue-engineered, artificial organs.

The hydrogels composed of MPEG-polyester or natural materials incorporating adult stem cells derived from bone marrow, muscle, and fat have been developed. Prepared as an aqueous solution and injected into the living body, this material was observed to undergo a sol-gel transition in situ and was found to form a gelled scaffold containing growing stem cells that persisted within the body

for at least 2 months. Inclusion of a bone differentiation-promoting substance resulted in the induction of bone formation at the implantation site. Such results highlight the virtually unlimited potential of injectable hydrogels as transplant transporters in tissue engineering applications.

Future Directions of Injectable Hydrogels

Research on injectable hydrogels and related fields is in progress at various top research institutes around the world and will certainly continue to expand. The key research advances to look for in the future are the development of materials with improved curative efficacy, convenience, and absence of side effects when applied to humans. These developments can be expected to proceed hand in hand with a focus on disease cures, life extension, and investigation of basic life phenomena. Improved understanding of the structural properties of injectable hydrogels that contribute to biocompatibility and therapeutic efficacy, as well as the design and development of original synthesis technology, will facilitate these efforts. At this time, a concerted research effort by workers in various fields is urgently needed to develop and apply original technologies that might help to realize the tremendous potential of injectable hydrogels.

Cross-References

- ▶ Biomimetics
- ▶ Self-Assembly
- ▶ Smart Hydrogels
- ▶ Sol–Gel Method

References

1. Temming, K., Fretz, M.M., Kok, R.J.: Organ- and cell-type specific delivery of kinase inhibitors: a novel approach in the development of targeted drugs. Curr. Mol. Pharmacol. **1**, 1–12 (2008)
2. Yu, L., Ding, J.: Injectable hydrogels as unique biomedical materials. Chem. Soc. Rev. **37**, 1473–1481 (2008)
3. Ruel-Gariépy, E., Leroux, J.C.: In situ-forming hydrogels – review of temperature-sensitive systems. Eur. J. Pharm. Biopharm. **58**, 409–426 (2004)
4. Kabanov, A.V., Alakhov, V.Y.: Pluronic block copolymers in drug delivery: from micellar nanocontainers to biological response modifiers. Crit. Rev. Ther. Drug Carrier Syst. **19**, 1–72 (2002)
5. Elsabahy, M., Perron, M.E., Bertrand, N., Yu, G.E., Leroux, J.C.: Solubilization of docetaxel in poly(ethylene oxide)-block-poly(butylene/styrene oxide) micelles. Biomacromolecules. **8**, 2250–2257 (2007)
6. Jeong, B., Bae, Y.H., Lee, D.S., Kim, S.W.: Biodegradable block copolymers as injectable drug-delivery systems. Nature **388**, 860–862 (1997)
7. Kim, D.Y., Kwon, D.Y., Kwon, J.S., Kim, J.H., Min, B.H., Kim, M.S.: Injectable in situ-forming hydrogels for regenerative medicines. Polym. Rev. **55**, 407–452 (2005)
8. Edlund, U., Albertsson, A.C.: Polyesters based on diacid monomers. Adv. Drug Deliv. Rev. **55**, 585–609 (2003)
9. Baillargeon, A.L., Mequanint, K.: Biodegradable polyphosphazene biomaterials for tissue engineering and delivery of therapeutics. Biomed. Res. Int. 2014, 761373 (2014)
10. Lakshmi, S., Katti, D.S., Laurencin, C.T.: Biodegradable polyphosphazenes for drug delivery applications. Adv. Drug Deliv. Rev. **55**, 467–482 (2003)
11. Elstad, N.L., Fowers, K.D.: OncoGel (ReGel/paclitaxel)-clinical applications for a novel paclitaxel delivery system. Adv. Drug Deliv. Rev. **10**, 785–794 (2009)

Injectable In Situ Forming Gel

▶ Injectable Hydrogel

Inorganic Nanoparticle Radioenhancers for Oncology

▶ Inorganic Nanoparticle Radiosensitizers for Cancer Therapies

Inorganic Nanoparticle Radiosensitizers for Cancer Therapies

Malgorzata J. Rybak-Smith and
Helen E. Townley
Department of Engineering Science, University of Oxford, Begbroke, UK

Synonyms

Inorganic nanoparticle radioenhancers for oncology; Inorganic nanoparticle radiosensitizers for oncology

Definitions

Inorganic particles – chemical compounds that do not contain carbon.

Nanoparticles – particles sized between 1 and 100 nm.

(Inorganic) Radiosensitizers – (inorganic) chemical molecules or drugs that have the capacity to increase the therapeutic effect of irradiation when administered together with irradiation.

Cancer therapy (oncology) – treating cancer with one of the modalities, such as surgery, radiotherapy, or cancer drugs (chemotherapy, biological treatment, hormone therapies, and others). Radionuclides (radioactive metal elements) can also be used in oncology.

Radiation Therapy

Based on information available on the World Cancer Research Fund International website, the number of expected cancer cases around the world is going to reach *21 million* by *2030* [1]. At present, surgery, radiation therapy, and chemotherapy are the most common treatments in oncology. Radiotherapy still remains a major nonsurgical modality for cancer therapy [2, 3]. It uses high-energy radiation, such as X-rays or γ-rays (external radiotherapy), to treat tumor. More than 50 % of cancer patients undergo radiation therapy during their treatment [3]. Skin cancers and cervix and lung carcinomas are examples of early cancers curable with radiation therapy alone [3]. However, most cancers (e.g., breast, bladder, and locally advanced lung carcinomas) need radiotherapy in combination with other modalities [3]. Radiotherapy achieves its therapeutic effect by inducing different types of cell death: apoptosis, necrosis, senescence (a state of permanent loss of cell proliferative capacity), autophagy, or mitotic catastrophe [3].

Tumors are usually hypoxic and in the absence of oxygen can become radioresistant [4]. Therefore, another major drawback of using radiation therapy is recurrence of tumor [4]. In the absence of oxygen, living cells surrounding tumor are less sensitive to irradiation than in the presence of oxygen [4]. Another obstacle of radiotherapy is that tumors do not receive sufficient dose of irradiation. Moreover, if the tumors are not targeted precisely, the healthy cells surrounding them can be destroyed by irradiation. Thus, the major drawback of using radiotherapy as well as other modalities in cancer therapy is destruction of healthy cells surrounding the tumor.

Radiosensitizers

Radiosensitizers can enhance radiation effects in tumors and allow a much higher dose to be delivered specifically to the tumor cells. An example of an organic tumor radiosensitizing drug is nicotinamide plus carbogen, which was found to enhance radiotherapy of experimental tumors [2]. Another approach is to use radioprotectors which can chemically protect healthy tissues. Amifostine appears to be beneficial when used in clinical trials of head and neck squamous cell carcinoma [2]. Molecules with electron affinity such as oxygen molecules or nitro compounds (nitroxyl-free radicals) are among radiosensitizers [4]. Molecular oxygen is the most well-known

radioenhancer [4]. The oxygen molecules can generate Reactive Oxygen Species (ROS) which can interact with DNA [4]. That is why the blood substitutes with increased oxygen concentration or high-pressure oxygen tanks can be used in the clinics to overcome tumor resistance. One of the most novel approaches is to use compounds (e.g., trans-sodium crocetinate – TSC) which can reoxygenate hypoxic brain tumor tissue (glioblastoma tumors) [5]. Another group of radiosensitizers are DNA-based analogues, e.g., halogenated pyrimidines, and also the membrane-active drugs [4]. The thiol reactive agents or silver salts are compounds which affect radiosensitivity and are called miscellaneous compounds [4]. Another approach to overcome the radioresistance problem is the use of hypoxic cell radiosensitizers, such as metronidazole or misonidazole [4].

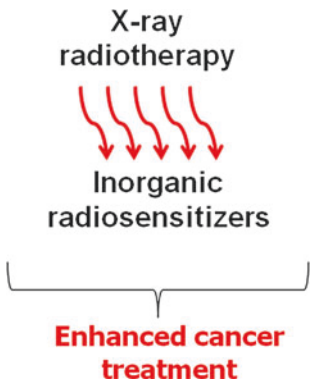
For decades, researchers have tried different approaches to improve the outcome of radiotherapy [4], but there is still room for improvement. How to design the most effective radiosensitizers? The ideal radiosensitizers should not only have high radiotherapy enhancement but also have good tumor targeting capacity, good biocompatibility (appropriate size and surface properties), and efficient renal clearance. Also, they should enhance the radiation dose without being toxic, so the ideal radiosensitizers should be inert in the absence of irradiation. The majority of molecules used in radiotherapy are organic radiosensitizers. However, various inorganic nanoparticles (NPs) attracted lots of attention for their interesting physicochemical properties, such as crystal structure, various geometries or coordination numbers, etc. and therefore have great potential to be used as effective radioenhancers. It is worth mentioning that without radiation, the inorganic nanoparticle radiosensitizers would present no therapeutic value in radiotherapy.

Nanosized Metal-Based Molecules

Recently, Pottier et al. reviewed the novel use of metals as nanosized radioenhancers [6]. The nanosized metal-based molecules have potential to be used in the clinics [6]. The metals that are most commonly used in oncology are platinum, gold, and ruthenium [6]. The novel metal-based pharmaceutical compounds represent an attractive approach in oncology. Nevertheless, the main limitation of using such compounds can be their toxicity [6]. Metal oxide NPs (such as hafnium dioxide, HfO_2) represent an interesting alternative to metal NPs [6]. Organic platinum (Pt) compounds, among them cisplatin, oxaliplatin, carboplatin, tetraplatin, are leading radioenhancers [6]. Some of the organic ruthenium (Ru) complexes (e.g., Nami-A and KP1019) are currently in clinical development. Radionuclides can also be used as attractive anticancer agents [6]. The α-particle emitters, β-particle emitters, and Auger electron emitters have energies ranging from 5 to 9 MeV, 50–2,300 keV, and eV-keV, respectively [6]. BEXXAR [I-131 (β emitter) tositumomab] or ZEVALIN [Y-90 (β emitter) ibritumomab tiuxetan] are examples of commercially approved radioimmunotherapy agents [6]. The agents can treat indolent B-cell lymphoma [6].

Inorganic Nanoparticle-Enhanced X-ray Therapy

An up-to-date review of inorganic radiosensitizers is presented here. Inorganic radiosensitizers can enhance radiation effects in tumors and allow much higher dose of irradiation to be delivered specifically to the tumor cells (as schematically presented in Fig. 1). The summary of various approaches for enhancing the radiosensitization in cancer cells was recently reviewed by Kwatra et al. [7] The term *Nanoparticle Enhanced X-ray Therapy* (NEXT), which is defined as the use of nanoparticles as radiosensitizers, was introduced [7]. The most common approaches involve using metal-based NPs, quantum dots (CaF, LaF, ZnS, or ZnO), supraparamagnetic iron oxides (magnetite Fe_3O_4 and magemite γ Fe_3O_4), and nonmetal-based NPs (i. e., silicon or fullerene (C_{60})-based) [7]. The use of quantum dots is limited to superficial cancers (photodynamic therapy) [8]. The metal-/metal oxide–based nanoparticles include gold NPs (GNPs), gadolinium-based NPs, titanium-based NPs (TiO_2), silver-based NPs, and



Inorganic Nanoparticle Radiosensitizers for Cancer Therapies, Fig. 1 Inorganic nanoparticles when activated by X-rays can increase the efficacy of conventional radiotherapy

hafnium-based NPs (HfO_2) [7]. The summary of the most commonly used inorganic radiosensitizers is shown in Fig. 2.

Radiotherapy Enhancement with GNPs

High-Z elements are materials with high atomic number. For example, the Z value for gold is 79. The dose of irradiation can be increased when a high-Z material is present in the targeted cancer tissue [2]. Gold is an exceptional absorber of X-ray irradiation and therefore the most studied radioenhancer [2].

Basic Physics of Metal-Enhanced Radiotherapy

Due to X-ray interaction with matter, a number of processes can take place, including Compton scattering, fluorescent photons, Compton electrons, pair (positron–electron) production, Rayleigh scattering, photoelectric effect, and Auger electrons (Fig. 3). Photoelectric effect changes as (Z/E) [6] (E represents the incident photon energy; Z is the atomic number) [2]. Photoelectric effect is a dominant interaction of high-Z elements with matter at energies below 0.5 MeV [2]. It was estimated that the relative effect of gold (Z = 79) to soft tissue (Z = 7.4) in the energy range above gold's K-edge (80.75 keV) is approximately 1,217 [2]. Fluorescent photons can travel longer distances (up to centimeters) than Auger electrons (about 10 nm), which are weakly bound electrons [2]. In Rayleigh scattering (also called elastic), photons collide with the entire atom, not single electrons [2]. Therefore, the amount of energy transferred is insignificant. Its cross section is very small in tissues at energies above 60 keV and in gold above 200 keV. The Rayleigh scattering is elastic with no energy deposited in the tissue, and therefore it is not suitable for cancer therapy [2]. Pair production (positron and electron) occurs at high photon energies (more than 1 MeV); its cross section in tissue below 3 MeV is not significant [2].

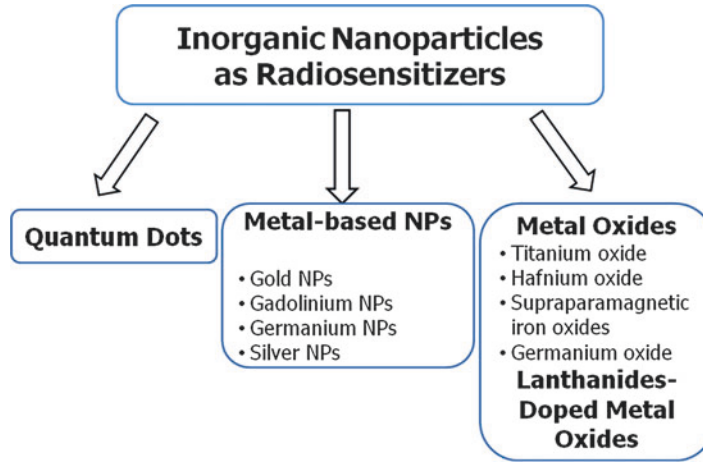
Optimal Beam Energy in Radiotherapy

The optimal beam energy for use with a dose-enhancing contrast agent has to be taken into account in the radiotherapy [2]. Currently, most radiotherapy machines use 6–25 MV electron linacs (linear particle accelerators) for the production of X-rays [2]. This does not apply for treating superficial cancers [2].

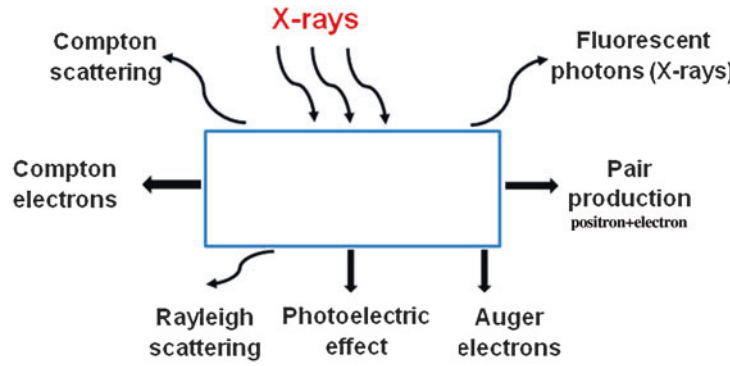
One of the factors for selecting the optimal irradiation energy is penetration of X-rays to be effective for deep tumors (liver, brain). The advantage of using gold over iodine (Z = 53) is its greater absorption at higher energies (about three times greater than iodine at 100 eV) [2]. Some of the theoretical calculations of dose enhancement in the literature have been done using Monte Carlo calculations [2, 21].

Generation of Reactive Oxygen Species

Reactive Oxygen Species (ROS) typically include the superoxide radical ($O_2^{\cdot -}$), hydrogen peroxide (H_2O_2), and the hydroxyl radical ($^{\cdot}OH$). Lifetimes and diffusion distances of various ROS are presented in Table 1 [9]. ROS as very reactive oxygen molecules can cause cellular damage, including oxidation of lipids, proteins, and DNA



Inorganic Nanoparticle Radiosensitizers for Cancer Therapies, Fig. 2 Examples of inorganic nanoparticles which have potential to be used for enhancing the radiosensitization in cancer therapy



Inorganic Nanoparticle Radiosensitizers for Cancer Therapies, Fig. 3 Consequences of interaction of X-rays with matter [2]

Inorganic Nanoparticle Radiosensitizers for Cancer Therapies, Table 1 Lifetimes and diffusion distances of Reactive Oxygen Species [9]

ROS	Symbol	Lifetime	Diffusion length
Hydroxyl Radical	**$HO^{\cdot}$**	100 nsec	4.5 nm
Singlet Oxygen	**$^1O_2$**	3.7 msec	82 nm
Superoxide Radical	**$O_2^{\cdot -}$**	50 msec	320 nm

resulting in apoptotic and necrotic cell death due to mitochondrial dysfunction [10]. Platinum complexes are the most commonly used compounds that have a great potential of targeting radiosensitizers to DNA [11]. However, other metals, such as palladium, rhodium, or ruthenium, can also be used for this purpose [11].

The major mechanism of radiation-induced cell death is ROS-mediated DNA damage. In a human body, ROS are generated as a consequence of the radiolysis of water. The most chemically active radical $^{\cdot}OH$, formed due to the radiolysis of water, readily reacts with biological molecules including cellular DNA to initiate radiation-induced apoptosis [10]. Metallic nanoparticles can induce the production of ROS leading to oxidative stress even in the absence of irradiation, for example, GNPs sized 1.9 nm [10]. When NPs are used in combination with ionizing radiation, they can cause enhanced ROS production [10]. GNPs are versatile NPs as they can serve as a platform for various biomedical and diagnostic purposes. Nonetheless, GNPs were found not to be inert in the absence of irradiation, which can have various biological consequences and limit their use [10].

Effect of Irradiation on Cell Cycle

Cell cycle is a vital mediator of sensitivity to the biological effects of irradiation [10]. In general, cells show different radiation sensitivity during different phases of the cell cycle [10]. Late S-phase cells are the most radioresistant, while mitotic cells are the most sensitive [10]. Cells respond to irradiation by activating cell cycle checkpoints in G1, S, and G2 phases [10]. Metal-based compounds block cells at the most radiosensitive phase: G2/M phase [11]. Lin et al. showed that nano-Germanium caused a higher level of DNA damage and cell cycle arrest in G2/M phase [12].

GNPs have been reported to cause alterations in cell cycle distributions [10]. Roa et al. demonstrated acceleration in the G0/G1 phase and accumulation of cells in G2/M phase in cells exposed to the irradiated glucose-capped GNPs [10]. Geng et al. observed an increase in the G2/M cell population for the 14 nm glucose-capped GNP-treated cells followed by irradiation at 6 MV X-rays [10]. Nuclear targeted GNPs have been shown to cause accumulation of cells in G1 phase and disruption of the G1/S transition inducing apoptosis in cancer cells [10]. These reports would suggest changes in cell cycle kinetics induced by GNPs, either additive toxicity as a result of increased sub-G1 population or accumulation in G2/M as a mechanism of GNP radiosensitization [10].

Summary of In Vitro Studies on GNPs

Jain et al. summarized results of in vitro studies of gold nanoparticle radiosensitization with ionizing radiation [13]. The impact of GNP size, concentration, and radiation energy on in vitro radiosensitization in HeLa cells was shown by Chitharni et al. [13]. The studies showed that the radiosensitization is size dependent. Three different GNPs sized 50 nm, 14 nm, and 74 nm were tested. The studies showed that the greatest radiosensitisation was observed with 50 nm GNPs with 220 kVp (peak kilovoltage (kVp) is the maximum voltage applied across an X-ray tube. It determines the kinetic energy of the electrons accelerated in the X-ray tube and the peak energy of the X-ray emission spectrum) X-rays [13]. Rahman et al. reported in vitro radiosensitizing effects of 1.9 nm GNPs (also called Aurovist) on the bovine aortic endothelial cells (BAEC) exposed to GNPs prior to irradiation with 80 kVp or 150 kVp X-rays [13].

Surface properties of NPs are a vital element in in vitro radiosensitization. Zhang et al. used 30-nm citrate or thioglucose-coated GNPs to test the GNP on the prostate cell line (DU145). The cells were irradiated with 200 kVp X-rays after 24 h exposure with the GNP [19]. As consequence, a threefold increase in GNP uptake was observed in the glucose-capped GNPs [13]. A study by Roa et al. compared thioglucose and cysteamine-capped GNPs (sized 10.8 nm) in MCF-7 breast cancer and MCF10a normal breast cells irradiated with 200 kVp, ^{137}Cs or ^{60}Co radiation [13]. A supra-additive effect with GNPs and radiation was observed with thioglucose GNPs and a single dose of irradiation [13]. Mechanisms of GNP sensitization in DU145 cells exposed to the thioglucose GNPs and irradiated with 2 Gy ^{137}Cs γ-rays was also investigated (original studies done by Roa et al.) [13]. The combination of the thioglucose GNPs and irradiation caused changes to the cell cycle compared to the untreated prostate cancer cells [13]. The proportion of the prostate cells in the G2/M phase was 29.9 % in the cells treated with the GNPs and irradiation compared to 18.4 % for untreated controls [13]. Chien et al. tested high concentrations of 20-nm GNP ranging from 50 to 400 mg ml^{-1} on the CT-26 mouse colorectal cell lines. Radiosensitizing effect with the cells in combination with 6-MeV electrons was observed [13].

Gold is a high-Z metal, therefore GNPs can absorb kilovoltage X-rays more effectively. GNPs have been shown to cause radiosensitization at kilovoltage and megavoltage photon energies. In the cancer therapy, most patients are treated with megavoltage (MV) photons produced by a clinical linear accelerator [13]. Studies by Chithrani et al. on HeLa cells and Liu et al. on CT-26 murine cancer cells have reported

in vitro GNP sensitization at megavoltage photon energies [19]. Chithrani's study used 50-nm citrate-coated GNPs [13]. The concentration of NPs also plays an important role in the observed radiosensitization [13]. Liu et al. used high concentrations of 6.1-nm PEGylated GNP [13]. Radiosensitization with 6-MeV and 12-MeV electron energies was also observed by Rahman et al. The researchers tested high concentrations (0.5 mM) of 1.9 nm GNPs in their studies [13]. More examples of in vitro studies of various GNPs with ionizing radiation can be found in the review by Jain et al. [13]. The same authors summarized in vivo studies of gold nanoparticle radiosensitization with ionizing radiation [13].

Summary of In Vivo Studies on GNPs

Hainfeld et al. reported the results of various in vivo studies on Balb/C mice bearing EMT-6 murine breast cancer tumors. In the studies, high concentrations of Aurovist were used [13]. Various approaches were tested: GNPs alone, radiation at 30 Gy alone, and combination of GNPs and radiation at 30 Gy [13]. The NPs were injected intravenously prior to irradiation [13]. The results showed that nonirradiated GNPs had no effect on tumor progression in mice. Irradiation at 30 Gy (alone) caused tumor growth delay [13]. The combination of irradiation at 30 Gy and GNPs caused a significant reduction in tumor growth [13]. In another experiment carried out by the same research group, mice were treated with two different concentrations of GNPs: 1.35 g of Au kg^{-1} or 2.7 g of Au kg^{-1} GNPs followed by irradiation at 26 Gy. About a half of mice treated with lower concentrations of GNPs and 86 % of mice treated with higher concentrations of GNPs followed by irradiation survived for 1 year. Interestingly, only 20 % of mice treated only with the irradiation at 26 Gy survived the treatment. None of the mice treated with either concentrations of GNPs survived the treatment. The GNPs were found to accumulate preferentially in the tumor periphery (EPR effect) [13]. The NPs do not accumulate in the main tumor mass [13]. Biodistribution studies by Hainfeld et al. showed that gold accumulated in tumors after 24 h. However, it was also found in the liver and spleen suggesting uptake by the RES, which may be reduced by PEGylation or active targeting [13].

The same research group carried out another experiment on mice bearing highly radioresistant murine squamous cell carcinoma, SCCVII. There were significant changes in tumor progression in mice treated with Aurovist or radiation only. The most significant tumor growth delay in mice was observed when Aurovist was combined with radiation using 68 keV photons [13].

Also, using of some coated GNPs gave very promising results in vivo. For example, 13-nm citrate-coated GNPs were tested in a mouse model with B16F10 murine melanoma cells. Significant in vivo tumor growth delay and increased survival were noted when 36 mg kg^{-1} citrate-coated GNPs were injected 24 h before irradiation at the dose of 25 Gy. There was no antitumor effect observed in control animals treated with the citrate-coated GNPs only (original studies performed by Chang et al.) [13].

However, some of the coated NPs did not show any significant difference in vivo. For example, gadolinium-coated GNPs, sized 5 nm, were tested in mice bearing MC7-L1 murine breast cancer cells (original work done by Hebert et al.) [13]. Mice were injected with GNPs 20 min before irradiation at 10 Gy [13]. Interestingly, the researchers did not report any difference in survival times in the group treated with radiation alone and the group of animals treated with GNPs and irradiation [13].

Designing the Most Effective GNP Radiosensitizer

Concentration. At higher concentrations, GNPs can be toxic. Moreover, the high concentration decreases cell growth rate [14].

Size. Large GNPs cannot be cleared by kidneys while small GNPs can be filtrated by kidneys quickly. Only GNPs sized 1–100 nm can be internalized by cells [14].

Surface properties. GNPs can be functionalized with PEG, glucose, DNA, peptides, etc. to improve

their biocompatibility. For example, glucose-functionalized GNPs can be much more effectively internalized by cells – as a consequence this enhances radiosensitivity [14].

Localization in the cell. GNPs should be localized inside the cell to improve their radiosensitivity [14].

Dose of irradiation. GNPs can show radiosensitization with kV, KeV, MeV, and MV X-rays [14].

Type of cell line. Radiosensitizing effect of GNPs depends on cell type [14].

Gadolinium Nanoparticles as Theranostic Agents

Mowat et al. showed in vitro that ultrasmall (5 nm) gadolinium-based NPs can be internalized in radioresistant cancer cells (human glioblastoma cell line U87) [15]. The particles consisted of a gadolinium oxide core coated by polysiloxane shell and were functionalized by diethylenetriaminepentaacetic acid (DTPA) [15]. The U87 cells were incubated with various concentrations of the gadolinium NPs, ranging from 0 mM to 10 mM followed by irradiation at 10 Gy. It was reported that the single-cell gel electrophoresis comet assay (that measures the number of DNA damages that occur during irradiation) proved the gadolinium NPs to be effective radiosensitizers [15]. Moreover, the authors showed that the same NPs can be efficient radiosensitizers at the energy 6 MV used in clinical practice [15]. Agents that allow precise defining of the tumor target to allow enhanced radiotherapy treatment without affecting neighboring healthy tissues are called "*theranostic*" agents [16]. Sancey at al. showed that the use of theranostic gadolinium-based agents (AGuIX) can improve the outcome of radiotherapy [16]. The theranostic gadolinium-based nanoparticles were comprised of polysiloxane and surrounded by gadolinium chelates (DTPA or DOTAGA – 1,4,7,10-tetraazacyclododecane-1-glutaric acid-4,7,10-triacetic acid) covalently grafted to the polysiloxane inorganic matrix [16]. The NPs consisted of high content of Gd (15 %) and had small size distribution (<5 nm) [16]. The authors reported that the presence of several gadolinium ions in the same nanoparticle could lead to further amplification of the Auger cascade [16]. In vivo efficacy of such AGuIX NPs was reported in heterotopic and orthotopic tumors [16]. It was reported that AGuIX NPs when injected intravenously into rats bearing 9 L glioma served as theranostic agents [16].

Recently, Miladi et al. reported gadolinium-based NPs (GBNs) comprised of polysiloxane and DTPA for potential theranostic use in vivo [17]. The same authors demonstrated radiosensitizing effects of the GBNs in various radioresistant human Head and Neck Squamous Cell Carcinoma (HNSCC): SQ20B, FaDu, Cal33 cell lines [17]. The GNPs had a small size distribution, about 2.9 nm. The GBNs showed a synergistic effect combined with a 250 kV energy irradiation in vitro. As a consequence of the interaction of the GBPs with photons, Auger electrons and ROS are produced that enhance the effect of irradiation [17]. The authors reported significant ROS generation 24 h after GBN treatment combined with irradiation [17]. The same authors confirmed the radiosensitizing effect of GBNs in conjunction with irradiation in vivo. Tumor growth in SQ20B-bearing nude mice was delayed by using the GBNs and a single 10Gy dose of irradiation [17].

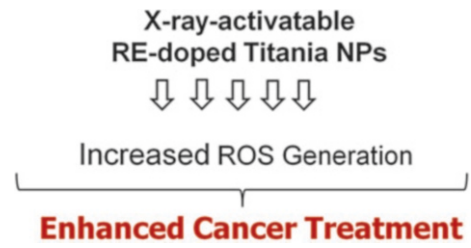
Rare Earth Elements

The effectiveness of radiotherapy can also be achieved by using rare earth (RE) elements [18]. The RE elements have a high Z number ranging from $Z = 57$ (lanthanum) to $Z = 71$ (lutetium) [18]. They comprise the 15 lanthanides in Mendeleev's periodic table with the addition of scandium and yttrium [18]. The improvement in delivering high dose of irradiation can be achieved by doping inorganic nanoparticles, such as titania NPs (titanium dioxide – TiO_2) with one or a combination of RE [18]. The exceptional properties of the various rare earth elements and their potential application in vivo for cancer imaging and cancer therapy were described by Townley [18].

RE-Doped Titania Nanoparticles as Radioenhancers

Titania (TiO_2) nanoparticles can have one of three crystal structures: brookite, anatase, and rutile. Anatase is the most photoactive among all the polymorphs of titania. TiO_2 can be doped with at least one lanthanide, such as Gadolinium (Gd), Europium (Er), or Erbium (Eu). Incorporation of the RE elements into the titania crystal structure (RE:TiO_2) is designed to optimize localized energy absorption from a conventional X-ray which has a maximum of absorption at 60 MeV [9]. The incorporation of the dopant into the TiO_2 can enhance the effect of radiation therapy [9]. The particulate characteristics of the doped titania NPs permits direct intratumoral administration. Tumor vasculature is hyperpermeable with pore sizes ranging from 50 to 600 nm. This allows the NPs to accumulate in a tumor tissue (passive targeting). The NPs can be well retained within the tumor due to the lack of drainage from solid tumors. RE:TiO_2 can be activated by X-ray radiation to generate Reactive Oxygen Species (ROS), which can be effective in killing cancer cells. As such, treatment with these RE-doped NPs can be used to enhance the efficacy of conventional radiotherapy. Incorporation of the NPs in to tumor tissue will permit the extension of radiotherapy to currently untreatable tumors deep within the body (as schematically presented in Fig. 4). Moreover, the short lifetimes and diffusion distances of ROS (shown in Table 1) would enable cancerous cells to be targeted with very limited damage to neighboring healthy cells [9]. Therefore, the RE:TiO_2 can serve as novel theranostic agents.

In vitro studies. Various compositions on NPs consisting of 1, 5 and 10 % of Gd as well as tridoped titania NPs consisting of 1 % of Er and Eu and various concentrations of Gd (1, 5 and 10 %) were tested on the breast carcinoma cell line (MCF-7) and the rhabdomyosarcoma (RMS) cell lines: RH30 and RD [9]. Rhabdomyosarcoma is the most common soft-tissue sarcoma among children and adolescents [9]. There are two major subtypes of RMS: the alveolar subtype (RH30) and the embryonal subtype (RD) [9].



Inorganic Nanoparticle Radiosensitizers for Cancer Therapies, Fig. 4 Rare earth elements (*RE*) doped titania (TiO_2) when activated by X-ray radiation, can increase the efficacy of conventional radiotherapy without causing additional damage to neighbouring healthy tissues. The nanoparticles ROS treatment can be extended to currently untreatable tumors deep within the body

The alveolar subtype has a poorer outcome than the embryonal subtype [9]. All tested NPs were internalized by the cancer cells [9]. However, they were unable to penetrate the nuclear membrane [9]. All tested cell lines showed reduction in cell proliferation in the presence of RE-doped titania NPs followed by X-ray irradiation except of RD cells which did not show statistically significant reductions [9]. MCF-7 cells showed significant reductions in cell proliferation for all NP compositions tested [9]. All tested RE-titania NPs were shown to be inert in the absence of irradiation [9]. Different amounts of ROS are produced by different cancer cell lines which can explain the different responses of the cell lines to the treatment [9]. The RH30 cells together with 3Gy irradiation showed statistically significant reductions in cell proliferation [9]. RH30 showed a 67 % and 66 % decrease in cell proliferation after treatment with the NPs consisting of 10 % Gd and tridoped sample containing the same molar percentage of Gd but additionally doped with 1 % Eu and 1 % Er followed by irradiation at 3Gy, respectively [9]. In the clinics, the total dose of radiation is not applied as a single dose but is fractionated. This allows the noncancerous cells some time to recover between irradiations [9]. RH30 cells were irradiated with two fractions of irradiation (two times 3Gy), and then the proliferation was determined [9]. Again, various compositions of NPs were tested, and a statistically significant

reduction in RH30 cell proliferation was reported [9]. For one of the tested samples, consisting of 5 % of Gd, 1 % Eu, and 1 % Er, there was a 75 % decrease in proliferation compared to untreated cells (control samples) [9].

In vivo studies. Preliminary preclinical studies in vivo by using tridoped RE:TiO_2 gave very promising results [19]. The tested NPs had a primary crystallite size ranging from 3 to 10 nm and were shown by XRD to be single-phase anatase [19]. Their hydrodynamic diameter of 65 nm allows the cellular uptake and effective generation of ROS [19]. The nanoparticles were silica coated to improve their biocompatibility and prevent aggregation in aqueous suspension [19]. Townley et al. found that intratumoral injection of the silica-coated three-doped titania nanoparticles (5 % Gd, 1 % Er, 1 % Eu:TiO_2@Si) can decrease A549 lung adenocarcinoma tumor volume in mice by around half when nanoparticle-augmented radiotherapy was used, compared to radiation alone [19].

Nanoparticle-Augmented Radiotherapy

Nanoparticle-augmented radiotherapy is an attractive, noninvasive, and novel method of using radiotherapy to activate NPs. The nanoparticle-augmented radiotherapy would allow fewer fractions of radiation in the clinics. Moreover, already existent hospital facilities and radiotherapy equipment could be used in the novel radiotherapy. The nano-sized particles are of a great advantage in this therapy (EPR effect), and this will avoid clearance by the kidneys and provide good particle retention in tumor. The surface of the NPS can be functionalized to improve their biocompatibility and prevent aggregation. If NPs have sizes less than 65 nm, they can enter a cell [20]. If the NPs have sizes less than 50 nm, they can enter an organelle of a cell [20]. The silica coating is thin (1–2 nm) and porous [9]. If the coating does not completely cover the surface of the metal oxide, it allows more efficient doses of irradiation to reach the molecule, generate enhanced ROS, and therefore allow a more effective radiotherapy outcome [9].

Other Inorganic Radiosensitizers

Lin et al. reported the use of *germanium* (*Ge*) *NPs* as potential radioenhancers [7]. Germanium oxide (GeO_2), Ge NPs (nano-Ge), and bis (2-carboxyethylgermanium) sesquioxide (^{132}Ge) were used in the study. The nano-Ge were found to be nontoxic to cells in the absence of irradiation [12]. ^{132}Ge did not enhance the radiosensitivity of Chinese hamster ovary (CHO) K1 cells, but inorganic Ge compounds enhanced cellular radiosensitivity [12]. The authors also reported a difference in solubility of the Ge compounds and their efficiency in cellular uptake. GeO_2 are uptaken by the cells much more efficiently than ^{132}Ge [12]. The study showed the advantage of using inorganic Ge nano-sized materials over the organic Ge compounds [12].

Chu et al. reported a radiosensitising effect of novel inorganic materials, *hydroxyapatite nanoparticles* (nano-HAPs) [21]. The NPs, sized about 50 nm, were synthesized by the sol-coagel method [21]. Nano-HAPs were tested on human glioblastoma U251 and breast tumor brain metastatic MDA-MB-231BR cells [21]. The nano-HAPS enhanced the radiosensitivity of the cells in vitro. Moreover, in a subcutaneous model, the NPs caused a larger than additive increase in tumor growth delay [21]. It was proposed that the same NPs used in an orthotopic model significantly reduced tumor growth. Moreover, the NPs extended the prolongation of survival induced by irradiation [21].

Silver NPs (Ag NPs) can also be potentially used as radiosensitizers [22]. The radiosensitising effect of various Ag NPs was reported for some cell lines, such as U231 breast cancer cells, A549 lung cancer cells, MGC803 gastric cells, and glioma cells and was size dependent [22]. For example, 20 nm and 50 nm Ag NPs performed better than 100 nm Ag NPs on the human U251 glioma cells (original work done by Xu et al.) [22]. There are few reports regarding the radiosensitizing effect of Ag NPs in vivo. One of the studies on rats bearing C6 glioma (original work done by Lin et al.) showed that the combination of a single dose of 10 Gy radiation and 10 μg and 20 μg Ag NPs had much higher survival times compared to

the animals treated with radiation alone [22]. The mechanism of Ag NP radiosensitization is not fully known yet [22].

Conclusion

The inorganic nanoparticle radioenhancers can be used as a platform for various applications of the NPs in oncology, especially in radiotherapy. The EPR effect allows the NPs to accumulate in the tumor. GNPs are the most commonly studied radiosensitizers. In practical terms, the main limitation of using GNPs in cancer therapy is their toxicity associated with the high concentrations of GNPs used. Interestingly, some inorganic compounds can be more effective radiosensitizers than the corresponding organic compounds. This was reported for germanium NPs. Another attractive alternative to improve the outcome of radiotherapy is to use metal oxides, such as titanium dioxide or hafnium dioxide, as potential radioenhancers. The titanium dioxide NPs, doped with one or more lanthanides, can enhance radiotherapy. The X-ray-activable RE-doped NPs can generate enhanced ROS production. The enhanced ROS produced can effectively damage tumor structure without affecting neighboring healthy cells. The advantage of using the NPs is that they are inert in the absence of irradiation so can serve as theranostic agents. The NP ROS treatment can be extended to currently untreatable solid tumors deep within the body. Also, as recently reported, various functionalized gadolinium NPs can be used as theranostic agents. In most studies, the NPs were found nontoxic at the concentration at which it exhibited some effect in combination with radiations and found to be effective even on radioresistant cancer cell lines.

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms
- ▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Nanoparticles
- ▶ Nanoparticle Cytotoxicity
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity
- ▶ Quantum Dot Toxicity
- ▶ Synthesis of Gold Nanoparticles

References

1. www.wcrf.org
2. Hainfeld, J.F., Dilmanian, F.A., Slatkin, D.N., Smilowitz, H.M.: Radiotherapy enhancement with gold nanoparticles. J. Pharm. Pharmacol. **60**, 977–985 (2008)
3. Baskar, R., Lee, K.A., Yeo, R., Yeoh, K.-W.: Cancer and radiation therapy: current advances and future directions. Int. J. Med. Sci. **9**, 193–199 (2012)
4. Shenoy, M.A., Singh, B.B.: Chemical radiosensitizers in cancer therapy. Cancer Invest. **10**(6), 533–551 (1992)
5. Sheehan, J., Cifarelli, C.P., Dassoulas, K., Olson, C., Rainey, J., Han, S.: Trans-sodium crocetinate enhancing survival and glioma response on magnetic resonance imaging to radiation and temozolomide. J. Neurosurg. **113**(2), 234–239 (2010). doi:10.3171/2009.11.JNS091314
6. Pottier, A., Borghi, E., Levy, L.: New use of metals as nanosized radioenhancers. Anticancer Res **34**, 443–454 (2014)
7. Kwatra, D., Venugopal, A., Anant, S.: Nanoparticles in radiation therapy: a summary of various approaches to enhance radiosensitization in cancer. Transl. Cancer Res. **2**(4), 330–342 (2013)
8. Shao, L., Gao, Y., Yau, F.: Semiconductor quantum dots for biomedical applications. Sensors. **11**, 11736–11751 (2011). doi:10.3390/s111211736
9. Townley, H.E., Rapa, E., Wakefield, G., Dobson, P.J.: Nanoparticle augmented radiation treatment decreases cancer cell proliferation. Nanomedicine NBM. **8**, 526–536 (2012)
10. Butterworth, K.T., McMashon, S.J., Currell, F.J., Prise, K.M.: Physical basis and biological mechanisms of gold nanoparticle radiosensitization. Nanoscale **4**, 4830–4838 (2012)
11. Skov, K.A., Adomat, H., Desmond, C.K., Farrell, N.P.: Assessment of DNA binding of Platinum-radiosensitizer complexes by inhibition of restriction enzymes. Chem. Biol. Interact. **62**, 117–129 (1987)
12. Lin, M.-H., Hsu, T.-S., Yang, P.-M., Tsai, M.-Y., Perng, T.-P., Lin, L.-Y.: Comparison of organic and inorganic germanium compounds in cellular radiosensitivity and preparation of germanium nanoparticles as

radiosensitizers. Int. J. Radiat. Biol. **85**(3), 214–226 (2009)
13. Jain, S., Hirst, D.G., O'Sullivan, J.M.: Gold nanoparticles as novel agents for cancer therapy. Br. J. Radiol. **85**, 101–113 (2012)
14. Babaei, M., Ganjalikhani, M.: A systematic review of gold nanoparticles as novel cancer therapeutics. Nanomed. J. **1**(4), 211–219 (2014)
15. Mowat, P., Mignot, A., Rima, W., Lux, F., Tillement, O., Roulin, C., Dutreix, M., Bechet, D., Huger, S., Humbert, L., Barberi-Heyob, M., Aloy, M.T., Armandy, E., Rodriguez-Lafrasse, C., Le Duc, G., Roux, S., Perriat, P.: In vitro radiosensitizing effects of ultrasmall gadolinium based particles on tumour cells. J. Nanosci. Nanotechnol. **11**, 1–7 (2011)
16. Sancey, L., Lux, F., Kotb, S., Roux, S., Dufort, S., Bianchi, A., Crémillieux, Y., Fries, P., Coll, J.-L., Rodriguez-Lafrasse, C., Janier, M., Dutreix, M., Barberi-Heyob, M., Boschetti, F., Denat, F., Louise, C., Porcel, E., Lacombe, S., Le Duc, G., Deutsch, E., Perfettini, J.L., Detappe, A., Verry, C., Berbeco, R., Butterworth, K.T., McMahon, S.J., Prise, K.M., Perriat, P., Tillement, O.: The use of theranostic gadolinium-based nanoprobes to improve radiotherapy efficacy. Br. J. Radiol. (2014). doi:10.1259/bjr.20140134
17. Miladi, I., Aloy, M.-T., Armandy, E., Mowat, P., Kryza, D., Magné, N., Tillement, O., Lux, F., Billotey, C., Janier, M., Rodriguez-Lafrasse, C.: Combining ultrasmall gadolinium-based nanoparticles with photon irradiation overcomes radioresistance of head and neck squamous cell carcinoma. Nanomed. Nanotechnol. (2014). doi:10.1016/j.nano.2014.06.013
18. Townley, H.E.: Applications of the rare earth elements in cancer imaging and therapy. Curr. Nanosci. **9**(5), 686–691 (2013)
19. Townley, H.E., Kim, J., Dobson, P.J.: In vivo demonstration of enhanced radiotherapy using rare earth doped titania nanoparticles. Nanoscale **4**, 5043–5050 (2012)
20. Dobson, P.J., Wakefield, G., Townley, H.E.: Particles for the treatment of cancer in combination with radiotherapy. United States Patent Application Publication, US2012/0282185 A1, 8 Nov 2012
21. Chu, S.-H., Karri, S., Ma, Y.-B., Feng, D.-F., Li, Z.-Q.: In vitro and in vivo radiosensitization induced by hydroxyapatite nanoparticles. Neuro Oncol. (2013). doi:10.1093/neuonc/not030
22. Su, X.-Y., Liu, P.-D., Wu, H., Gu, N.: Enhancement of radiosensitization by metal-based nanoparticles in cancer radiation therapy. Cancer. Biol. Med. **11**, 86–91 (2014)

Inorganic Nanoparticle Radiosensitizers for Oncology

► Inorganic Nanoparticle Radiosensitizers for Cancer Therapies

Insect Adhesion

► Adhesion in Insects

Insect Flight and Micro Air Vehicles (MAVs)

Hiroto Tanaka[1,2], Benjamin M. Finio[1], Michael Karpelson[1], Néstor O. Pérez-Arancibia[1], Pratheev S. Sreetharan[1], John P. Whitney[1] and Robert J. Wood[2]
[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[2]School of Engineering and Applied Sciences, Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA, USA

Synonyms

Flapping-wing flight; Micro air vehicles (MAVs); Robotic insects

Definition

Micro air vehicles are generally considered to be blown 15 cm in characteristic dimension and weigh on the order of 10 g. Small MAVs, vehicles 1 g or less, have historically relied on insect flight for inspiration and have utilized a variety of flapping-wing morphologies.

This entry discusses sub-gram MAVs inspired by flying insects.

Introduction

Many insect species are agile fliers. Yet they operate in a way fundamentally different from man-made aircraft. Although there are many, sometimes subtle, differences between the flight apparatus of individual species, in general, insects have one or two pairs of wings, driven in multiple

rotational degrees of freedom by flight musculature, and powered by metabolic processes which convert chemical energy for flight. Engineers striving to produce a robotic insect must solve similar challenges, not only for the tasks which the device could perform, but also for the open scientific questions one could use such a device to answer. Therefore, a key question is: how can one manufacture an effective replica of a flying insect? This entry describes the reproduction of some of the key functionalities of a flying insect not as an end "design," but to prove the viability of the core technologies necessary for such a robot.

The integrated circuit revolution of the 1950s and 1960s now enables the majority of the consumer electronics that are enjoyed daily. As these techniques evolved in the 1980s to include electromechanical components, an even greater space of applications emerged including sensors, optics, and even actuation [1]. Microrobots have been made using MEMS (microelectromechanical systems) surface and bulk micromachining techniques [2, 3]. However, there are many drawbacks to use integrated circuit (IC) and MEMS technologies to create robotic insects. First is the dramatic difference between the material properties of silicon and insect tissue: the former being rigid and brittle while the latter exhibits a large range of material properties, is generally quite resilient, and is approximately the density of water. Second, although the suite of techniques for high resolution machining is an appealing aspect of MEMS processes, the resulting structures are typically "2.5D," with high aspect ratio components being extremely challenging in terms of machining or requiring hinged structures [4]. Finally, although MEMS foundries exist (e.g., the Multi-User MEMS Process, www.memscapinc.com), cost and turnaround time are generally prohibitive to rapid prototyping. With the advent of "mesoscopic" manufacturing methods [5], researchers have demonstrated key components of the flight apparatus of robotic insects [6] and recently, the first demonstration of a 60 mg flapping-wing device which can produce thrust greater than its body weight [7] has proven the feasibility of creating insect-scale flying robots.

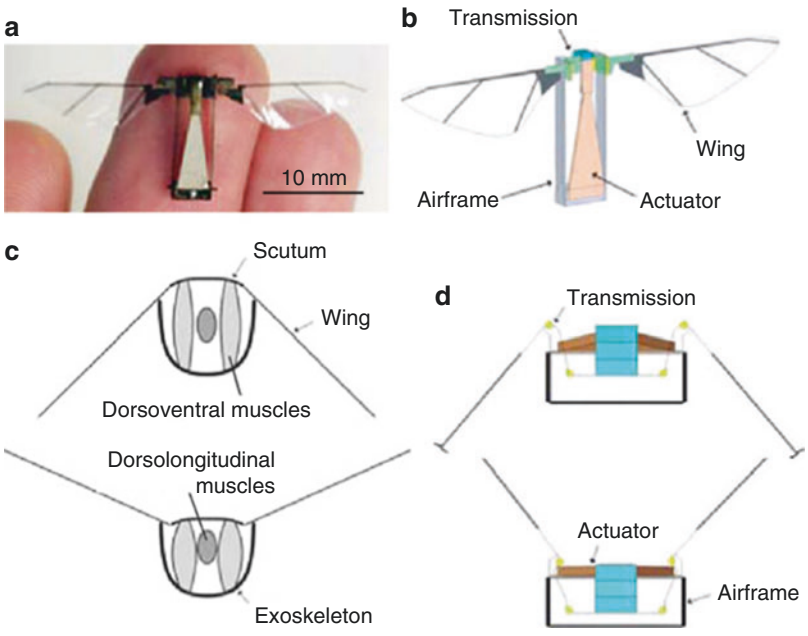
This entry discusses the challenges in developing the components necessary to achieve stable flight of a robotic insect. Topics include aeromechanics, fabrication, actuation, design, control, and power. This entry mostly considers the physical instantiation of a robotic insect. There are numerous additional aspects of a robotic insect that would be required in order for the device to perform useful tasks in natural environments. These topics will be outlined in the conclusion.

Overview of a Robotic Insect: "RoboBees"

The "RoboBee" project is aiming to build a two-wing robotic flying insect for a variety of tasks including search and rescue and assisted agriculture (e.g., pollination). This section describes the state of the art for the physical instantiation of a robotic bee which consists of composite-reinforced airfoils, a flexure-based transmission, piezoelectric actuation, and a rigid composite airframe. Key to the success of any flapping-wing design is an understanding of the aerodynamics for vehicles of this type and scale; this is discussed in section "Aerodynamics of Flapping-Wings at Low Reynolds Numbers."

Aeromechanical Components of a RoboBee

The current RoboBee prototype has four primary mechanical components: thorax, airframe, bimorph piezoelectric actuator, and a pair of airfoils [7], as shown in Fig. 1a, b. To fabricate this bee-sized robot, flexure-based structures are utilized in place of more traditional revolute bearings. A single bending actuator pushes and pulls the upper part of the mechanical thorax in a similar configuration to Dipteran dorsoventral muscles (Fig. 1c, d). This bioinspired flexure-based transmission is capable of converting small (hundreds of microns – limited by the motion of the piezoelectric actuator) actuator motions into large flapping angles (over 100° peak-to-peak) with a minimum of moving parts and with feature sizes appropriate to the scale of a bee and with extremely high transmission efficiencies. This is otherwise rather difficult, if not impossible, to

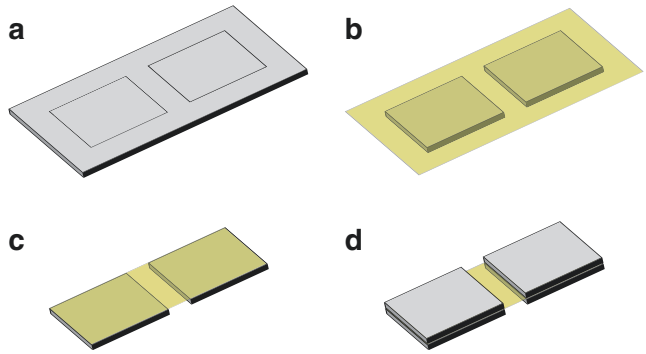


Insect Flight and Micro Air Vehicles (MAVs), Fig. 1 (**a**) A prototype RoboBee illustrating scale. (**b**) Schematic of the components of RoboBee. (**c**) Cross-sectional schematic of Dipteran thorax. (**d**) Mechanical thorax of a RoboBee

achieve with a conventional crank mechanism using a rotational motor, rotational bearings, and gears due to part size, assembly accuracy, and friction. The flexure-based structure was utilized not only for transmission hinges but also for precise fabrication and assembly of micro parts. Rigid carbon fiber frames connected with the flexure hinges were fabricated in a plane and then assembled into 3D shapes by folding. This enables precise fabrication and assembly without limiting the materials that can be utilized in the structures. The fabrication process for the flexure-based components is illustrated in Fig. 2. Rigid face sheets are laser micromachined using a diode-pumped solid-state laser from carbon fiber reinforced prepregs (preimpregnated with a catalyzed but uncured resin) with precisions and features sizes on the order of 10 μm. Thin sheets of polyimide – chosen for mechanical resilience and thermal stability – are also laser micromachined and laminated to the uncured prepreg face sheet using a vacuum tool to ensure uniform pressure distribution. The process is repeated for the opposite face sheet. Airfoils can also be created in this manner but without the need for face sheets on either side of a thin film polymer membrane. Previous versions have relied on passive flexure hinges at the base of the wing to give rise to the second degree of freedom of the wings: pronation and supination (collectively termed wing rotation). Using these techniques, prototypes as small as 60 mg (15 mm wing length) have been demonstrated [7]. Although all power electronics reside off-board the robot, this device achieved liftoff with vertical rails at a wingbeat frequency of 110 Hz.

Aerodynamics of Flapping-Wings at Low Reynolds Numbers

Flight at small scales is challenging not only in the development of the mechanism to more the airfoils, but also in the scale-dependent energetic cost of manipulating the fluid. At the scale of insects, with typical Reynolds numbers in the range of 100–10,000 [8], fluid mechanics can be dominated by viscous effects and laminar flow. For example, given a 10 mm long, 5 mm average chord wing operating in air at a 100 Hz flapping frequency with 90° stroke amplitude is operating at an average Reynolds number of approximately 500. In steady flight, these low Reynolds number conditions can dramatically reduce lift-to-drag ratios compared to larger scales and higher velocities. Therefore, there has been significant interest



Insect Flight and Micro Air Vehicles (MAVs), Fig. 2 Fabrication process for flexure-based articulated structures consisting of carbon fiber rigid plates and a polyimide hinge. First, composite face sheets are laser machined (**a**) and laminated with a polyimide film (**b**). This film is laser cut (**c**) and the structure is laminated to a second face sheet (**d**)

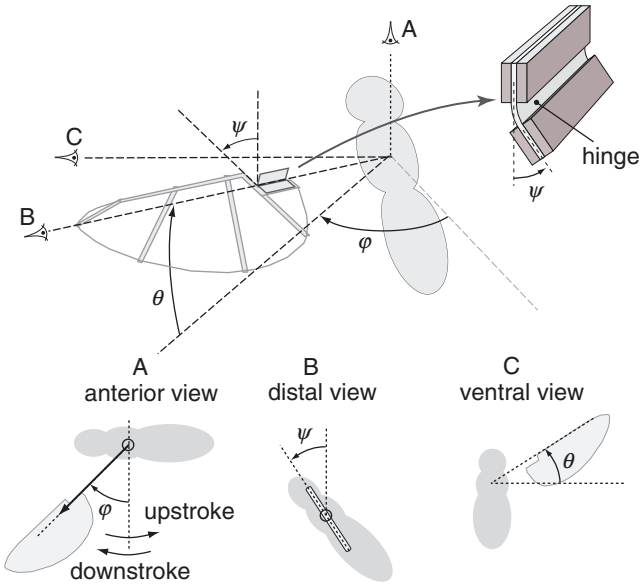
in the study of the "unsteady" mechanisms that insect use to enhance lift production [9].

Flapping kinematics are characterized by large angles of attack at midstroke and high rates of rotation during supination and pronation of the wing. Massive separation of boundary layers, possible reattachment, and strong vortex shedding at both the leading and trailing edges are possible features of this type of flow. In hover, the wing interacts with its own wake, potentially leading to significant time-dependent aerodynamic forces. Needless to say, the wing dynamics and aerodynamic effects involved are complex and the accurate prediction or measurement of forces and flow fields demands sophisticated flow simulations and precise experiments.

Many different aerodynamic effects have been identified, which are reviewed in [10]. Often, experiments attempt to explain the aerodynamics for a single wing at a single operating point, usually with the aim of identifying the important aerodynamic mechanisms for a particular insect species. Recent work by Lentink and Dickinson [11] reverses this trend; they investigate aerodynamic performance under the systematic variation of wing aspect ratio, Reynolds number (Re) and flapping amplitude using a scaled-model mechanical flapping device. As interest in flapping-wing MAVs rises, the study of flapping flight from a design, rather than biological perspective, becomes more important. Instead of studying the behavior and performance of a particular species, research must uncover the important parameters in wing configuration and flapping kinematics that most strongly influence vehicle-level performance for a range of mission profiles and vehicle sizes.

The wing design of a hovering MAV entails not only the capacity for lift generation, but, more importantly, the power required for flight. Hovering is extremely energy-intensive and aerodynamic efficiency is a primary concern. Of equal concern is minimizing the total vehicle mass. Simplification of the wing flapping mechanism and reducing the number of actuated degrees of freedom can lead to great weight savings. To this end, a successful configuration has been developed, where each wing is driven in a single flapping degree of freedom, $\Psi(t)$, allowing wing rotation to happen passively. Figure 3 illustrates a wing hinge that allows passive rotation. Clear evidence exists that some insects exhibit passive rotation [12], but there is no consensus that all insects use passive rotation, or that it is necessary for optimal flapping flight. Nevertheless, passive rotation has a demonstrated record of success in flapping-wing MAVs [7]. Prediction of wing performance, in the case of passive rotation, now becomes an aeromechanics problem, as the aerodynamic forces, wing inertia tensor and wing-hinge elastic properties jointly determine the kinematics and resulting forces.

A class of flapping-wing aerodynamic analysis techniques, the so-called blade-element methods

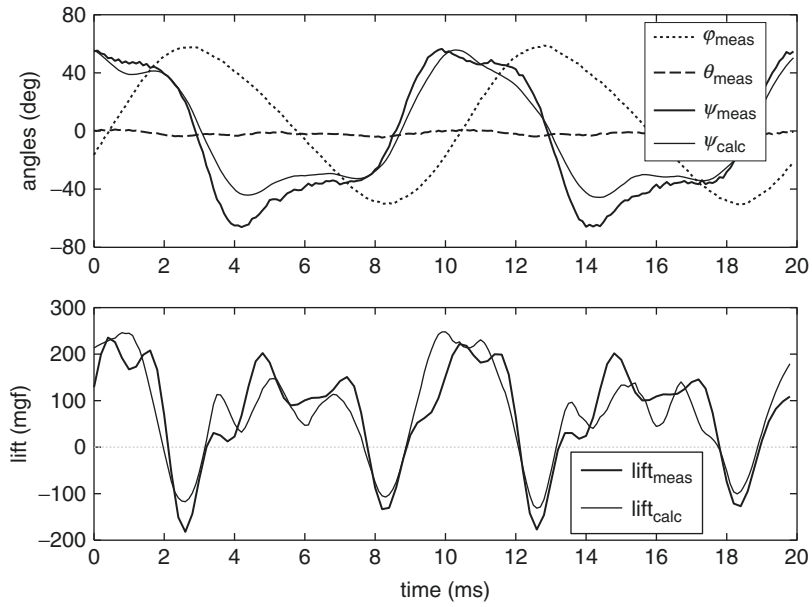


Insect Flight and Micro Air Vehicles (MAVs), Fig. 3 Angles used to specify wing kinematics, shown for a *left* wing and *vertical* body orientation. Current robotic insect prototypes utilize a flexure at the base of the wing (*upper right*) to allow passive rotation. The driving spar (flapping input) mounts to the *top* link of the hinge, the wing to the *bottom*

are fundamentally a form of dimensional analysis, applied in a careful and systematic fashion. The blade-element method is a quasi-static technique for applying nondimensional experimental and computational results to wings of different scales and operating conditions. The success of this method is critically limited by the quality of the accompanying force and moment coefficients and the method's quasi-static assumptions. The blade-element method begins by dividing the wing into chordwise strips along a flapping wing. The wing must be divided into strips because the relative local velocity varies linearly along the wing. Figure 3 illustrates a wing, flapping in hover, and breaks the kinematics down into flapping (Ψ), out-of-plane deviation (θ) and rotational (ψ) degrees of freedom. Using a simple expression based on dimensional analysis, for the local pressure on the wing, forces and moments arise after a straightforward integration over the entire wing. The key to the blade-element method is experimentally determining the force and moment coefficients, and ensuring that any time-dependence of these coefficients, as a result of unsteady effects, is negligible.

The Reynolds number (Re) for hovering insect flight is typically between 100 and 10,000. This range indicates laminar flow, and a suite of experimental and computational techniques can be employed to predict wing trajectories, deformations, and forces. Scaled-model experiments (e.g., [9]), which mechanically flap scaled-up wings in high-viscosity liquids so as to match Reynolds number, have been very successful in identifying important aerodynamic mechanisms and characterizing the performance of actual insects. However, it is not possible to match the Reynolds number (aerodynamic forces) and the inertial forces simultaneously in a scaled-model experiment. Thus, the scaled-up approaches are not appropriate for studies of passive rotation, or studies of wing deformation. Working with at-scale mechanical flapping mechanisms and micro-force sensors is an experimental challenge, but recent advances in these areas [7, 13] have enabled passive rotation and wing-deformation studies at scale.

The experiment described in [14] consists of a one-axis wing driver mounted to a capacitive force sensor. High-speed video was used to reconstruct the detailed wing trajectory. These kinematics were then used to generate blade-element force predictions, which could be compared to the experimentally measured forces. Figure 4 illustrates a comparison for one particular wing, flapping at 100 Hz. Hinge stiffness, flapping



Insect Flight and Micro Air Vehicles (MAVs), Fig. 4 Wing kinematics and lift at 100 Hz flapping (Measured and calculated). Measured kinematics are plotted unfiltered

frequency, and flapping amplitude were varied, and comparisons with blade-element predictions were made for each configuration.

Future work in support of flapping-wing MAV development will focus on continued experimental exploration of the wing design space, with a goal of minimizing the power required to generate the needed lift, under the vehicle mass and manufacturability constraints imposed by the microfabrication technology used, and the power sources available. The design and testing of flexible wings is of particular interest in the continuing quest for higher aerodynamic efficiency.

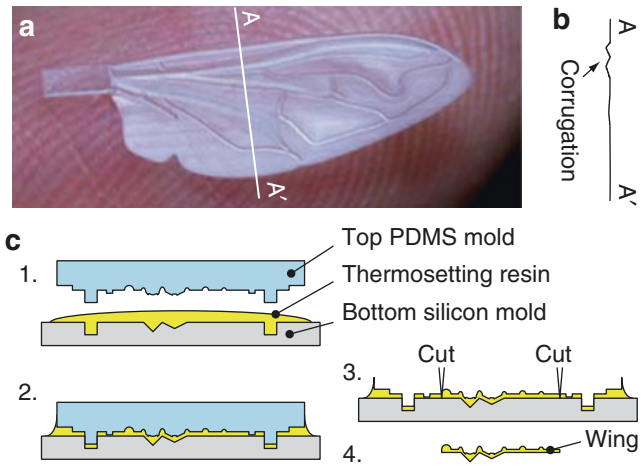
Toward Autonomy

Although a prototype RoboBee achieved takeoff along vertical guide rails, there are myriad challenges which need to be addressed to achieve full autonomy. These topics include, but are not limited to, airfoil design and fabrication, actuation and thoracic mechanics, and control. These topics will be addressed in the following sections.

Biomimetic Flexible Wings

Flexible 3D polymer wings which mimic insect wing morphologies are being developed in addition to the previous rigid carbon fiber based wings described in section "Aeromechanical Components of a RoboBee." Real insect wings are flexible and can passively deform when subject to aerodynamic load and inertia force. This flexibility is determined by the wing structure which consists of a thin cuticle membrane supported by a network of tubular cuticle veins [15]. Some insects also have 3D wing profiles such as the corrugation present in Dipteran wings. Actual effects of these wing morphologies on the aerodynamic performance of insect-sized flapping flight, however, are still unknown. By fabricating biomimetic wings, the aim is to develop an understanding of the structure-function relationship for these systems and exploit this understanding to create more effective wings for a robotic insect. Figure 5a shows a fabricated polymer wing mimicking the morphological features of a hoverfly wing [16].

As with a real hoverfly wing, the artificial polymer wings consist of veins with various thicknesses and corrugated profiles (Fig. 5b). Polymer veins and membranes with complex 3D profiles are simultaneously formed by a single molding process with a pair of molds: a bottom rigid (Silicon) mold and a top compliant (PDMS: Polydimethylsiloxane) mold (Fig. 5c). The technical challenge



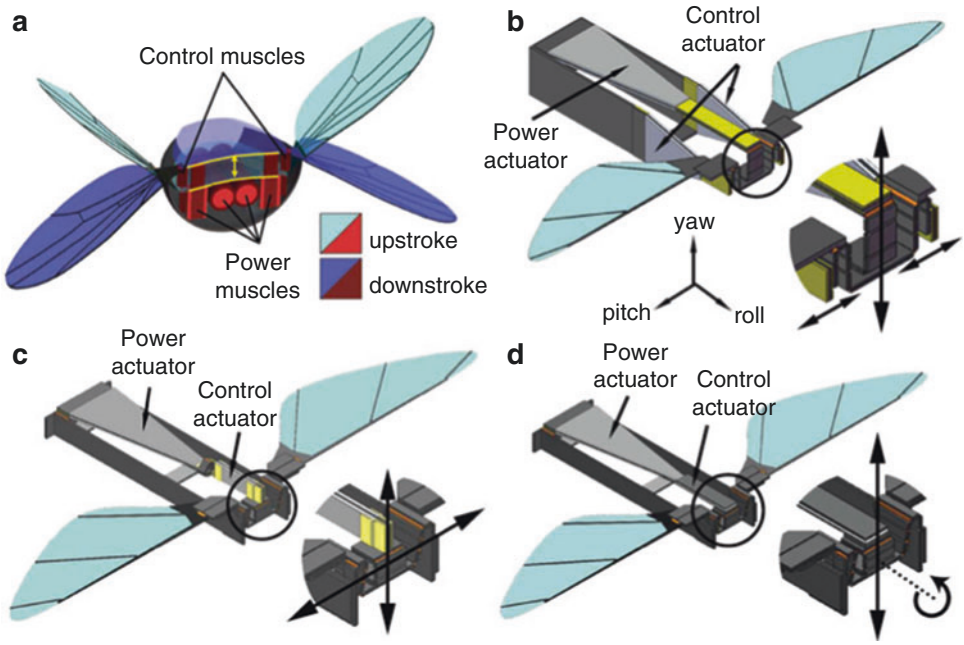
Insect Flight and Micro Air Vehicles (MAVs), Fig. 5 (**a**) Photo of an artificial corrugated polymer wing. (**b**) Measured corrugated profile. (**c**) Schematic of fabrication process flow

lies in the development of molds with various 3D microscale features: veins, membranes, and corrugation. To realize arbitrary surface profiles in a mold, a custom DPSS (diode-pumped solid-state) laser machining system is used. The surface of a silicon wafer is directly etched with the scanning pulsed laser layer-by-layer to create a 3D negative mold. The spot size of the laser pulse is 10 μm and the etching rate is approximately 5 μm per layer, which is sufficient resolution for the veins and corrugation. The bottom negative mold is directly fabricated with this scanning laser ablation method, and the top PDMS mold is created by casting PDMS on another laser-ablated positive silicon mold.

The thickness of the artificial veins ranges from 50 to 125 μm and the maximum height of the corrugation is 100 μm. Since a hoverfly has vein thickness of 10–70 μm and a maximum corrugation height of 150 μm, this fabrication method enables biomimic insect wing morphologies in the same scale. The spanwise stiffness of the artificial wing, which is measured by a point-load bending test, is 2.2×10^{-7} Nm^2. This value has the same order of magnitude of that of *E. dimidiate* (from dried specimens, measured using the same point-load bending test), 5.3×10^{-7} Nm^2, which demonstrates that the artificial wing is statically similar to actual insect wings. These biomimetic wings are useful not only as airfoils for robotic insects, but also for biological studies on insect flight. Unlike experiments with live insects, microfabricated airfoils enable parametric studies of geometric features and experimentally test the relationship between geometry and performance.

Actuators and Thoracic Mechanics

Traditional fixed and rotary wing aircraft use either control surfaces (rudders, ailerons, etc.) or tail rotors to steer. Although insects may use their legs or abdomen to shift their center of mass or aerodynamic profile, their primary steering mechanism is the use of asymmetric wing kinematics [8]. While large, central "power" muscles are responsible for flapping the wings and generating the mechanical power required to stay aloft, in Dipteran insects much smaller "control" muscles at the base of each wing can cause small kinematic changes in wing motion that result in net body torques, allowing the insect to maneuver. A similar approach is used with piezoelectric actuators in the next generation prototype RoboBee (Fig. 6). The wing kinematic parameters, stroke angle ψ, rotation angle ψ, and deviation angle θ are defined in Fig. 3. Since wing dynamics are typically about an order of magnitude faster than body dynamics (i.e., it may take ten wing strokes to complete a 90° turn, although this is not always the case), variation in average parameters over the entire stroke can be sufficient to generate mean body torques. For example (to list only a few possibilities): (a) moving the mean stroke angle for both wings forward of the



Insect Flight and Micro Air Vehicles (MAVs), Fig. 6 (**a**) Cross-sectional view (artist's rendering, not anatomically correct) of a Dipteran insect thorax, (**b**) thorax design with two separate control actuators for changing stroke amplitude, (**c**) design with a "hybrid" actuator consisting of two bending piezoelectric elements, (**d**) another hybrid actuator with one bending element and one twisting element. All three designs achieve the same result of differential stroke amplitude

body's center of mass will result in a pitch torque, (b) differentially increasing the stroke amplitude of one wing to be larger than that of the opposite wing will result in a roll torque, and (c) tilting the stroke plane of one wing forward while tilting the stroke plane of the other wing backward will result in a yaw torque.

Previous prototype RoboBee designs only had one central power actuator, a bending piezoelectric cantilever. Subsequent designs included two smaller control actuators at the base of each wing, allowing controlled stroke amplitude asymmetry [17]. The systems are designed such that very small perturbations to the mechanical transmission connecting the power actuator to the wings by the control actuators can cause large changes in wing trajectory – that is, amplitude differences or stroke-plane deviations on the order of tens of degrees. Several variations of the design for controlled amplitude asymmetry are presented in Fig. 6.

While the ability to modify wing kinematics in a manner similar to that used by Dipteran insects has been verified in [17], actual torque measurements will be required to fully verify aerodynamic models and generate useful feedback control laws for vehicle stabilization and navigation. Work in this area is ongoing.

Altitude Control

To date, detailed strategies addressing the problem of altitude control of robotic insects have not been reported. Recent results obtained at the Harvard Microrobotics Laboratory show that by transforming the original problem of altitude control into one of average lift-force control, controllers capable of enforcing specified trajectories along the vertical axis can be synthesized. The main idea is that enough information about the subsystems composing the robotic insect can be gathered a priori, using well-known identification methods, such that, during constrained vertical flight, only an altitude sensor is required.

In this context, there are two subsystems relevant from a control perspective. One is the bimorph actuator, used to transduce electrical

into mechanical power, mechanically connected to the insect's airframe. The other is the mapping from the actuator tip displacement to the average lift force generated by the passive rotation of the wings, assuming a sinusoidal excitation of the bimorph actuator. Here, the first subsystem is labeled as P, such that

$$Pu + v = y, \tag{1}$$

where, P is the open-loop dynamics, u is the voltage signal applied to the actuator, y is the output displacement of the actuator tip, and v is an output disturbance signal, representing the aggregated effects of all the disturbances affecting the system, including unmodeled aerodynamic forces produced by the wing flapping.

The second subsystem is a static map, labeled as Ψ, such that

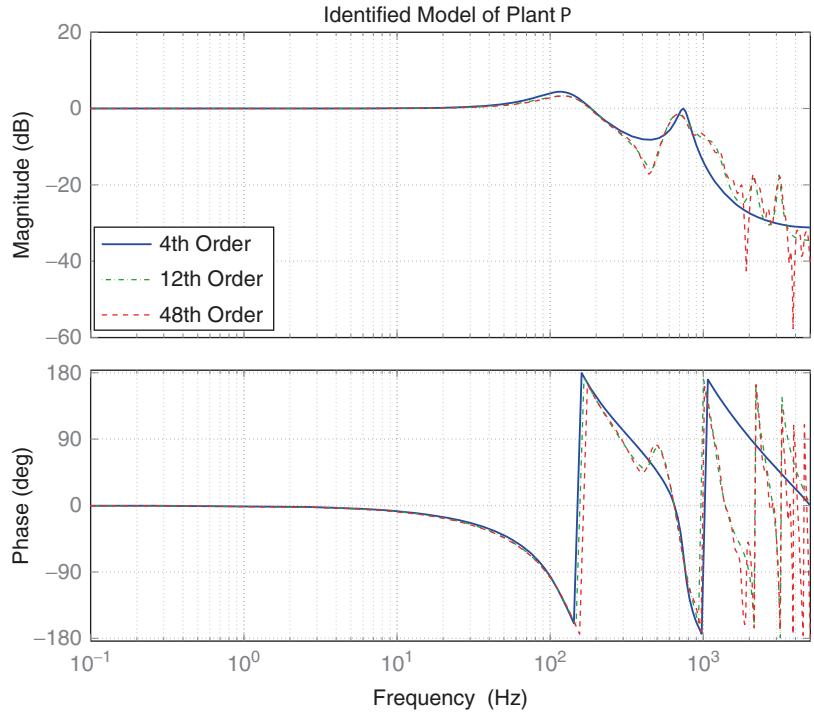
$$\Psi(A, f) = \Gamma_L, \tag{2}$$

where, A and f are the amplitude and the frequency of the displacement output $y(t) = A \sin(2\pi ft)$, respectively. The signal Γ_L is the average lift force produced by the flapping of the insect's wings. Typically, instantaneous forces produced by the wings are oscillatory signals with a positive average value. Thus, hovering occurs if Γ_L is approximately equal to the weight of the robotic insect, and vertical motion occurs if Γ_L is larger than the sum of the insect's weight and the vertical aerodynamic drag. When using digital signal processors for control, Γ_L can be estimated as $\Gamma_L^{(N_L)}(t) = \frac{1}{N_L}\sum_{i=0}^{N_L-1} \gamma_L(kT_s - iT_s)$, where $\gamma_L(t)$ is the instantaneous force measured using a force sensor like the one described in [13], Ts is the sampling time, and $0 < N_L \in \mathbb{Z}^+$, is the number of samples.

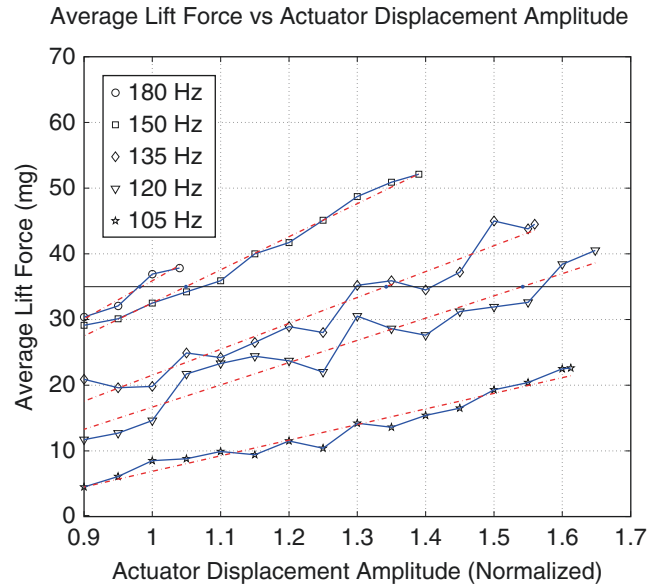
With the use of an appropriate sensor to measure the actuator tip displacement, a discrete-time *linear time-invariant* (LTI) model of P, labeled as $\hat{P}$, can be found using standard system identification methods. Similarly, with the use of a force sensor, a displacement sensor and a feedback controller capable of enforcing specified trajectories $y(t)$, a lookup table with some relevant points of Ψ can be found through experiments. The estimated mapping $\hat{P}$ and the estimated lookup table associated to Ψ are contingent on a particular microrobotic flapping system, and therefore, great variability should be expected. As examples, in Figs. 7 and 8, an identified model $\hat{P}$ of P, and a graphic representation of an estimated lookup table are shown. These plots where obtained from experiments performed on a single-wing flapping microrobot, similar to the one described in [14]. In Fig. 8, the amplitudes of the signals $y(t)$ are normalized such that a constant input $u(t) = 1$ generates an output equal to 1.

The main idea for designing and implementing controllers is depicted in Fig. 9. Here, the idea is that a desired signal altitude $x_d(t)$ is arbitrarily chosen. Simultaneously, the actual altitude $x(t)$ is measured using an altitude sensor. Both signals are filtered through an *upper level control law* in order to generate a desired average lift-force $\Gamma_d(t)$ which in turn is filtered through an estimated lookup table, like the one graphically shown in Fig. 8, to generate a desired displacement of the actuator tip with the form $y_d(t) = A_d \sin(2\pi f_d t)$. Since A_d and f_d are unknown a priori, the idea of using some type of adaptive cancelation scheme is reasonable. This follows from defining $r(t) = -y_d(t)$ and the error $e_y(t) = y(t) + r(t)$, and then treating $r(t)$ as a disturbance to be rejected. Notice that under the assumption $v(t) = 0$, if $r(t)$ is completely canceled by $y(t)$, the error $e_y(t)$ vanishes and $y(t)$ follows perfectly the desired displacement $y_d(t)$. The other element to be noticed in Fig. 9 is that in order to avoid the need for a displacement sensor during flight, the identified model $\hat{P}$ is used to compute an estimate $\hat{y}(t)$ of $y(t)$, and then, an estimate of the control error, $\hat{e}_y(t) = \hat{y}(t) + r(t)$, is filtered through a *lower level adaptive control law* to generate the control signal $u(t)$ which is the input to the physical system.

To test the proposed method, a single-wing flapping microrobot, similar to the one in [14], is employed in the implementation of hardware-in-the-loop experiments. Since the main idea is to demonstrate lift control using the adaptive scheme in Fig. 9, a simple open-loop upper level control law is used. The objective is to follow an average lift-force signal, $\Gamma_L(t)$, such that a 70-mg robotic



Insect Flight and Micro Air Vehicles (MAVs), Fig. 7 Bode diagram of an identified discrete-time model $\hat{P}$ of the plant P. A 48th-order model is shown in *red*, reduced 12th and 4th-order models are shown in *green* and *blue*, respectively

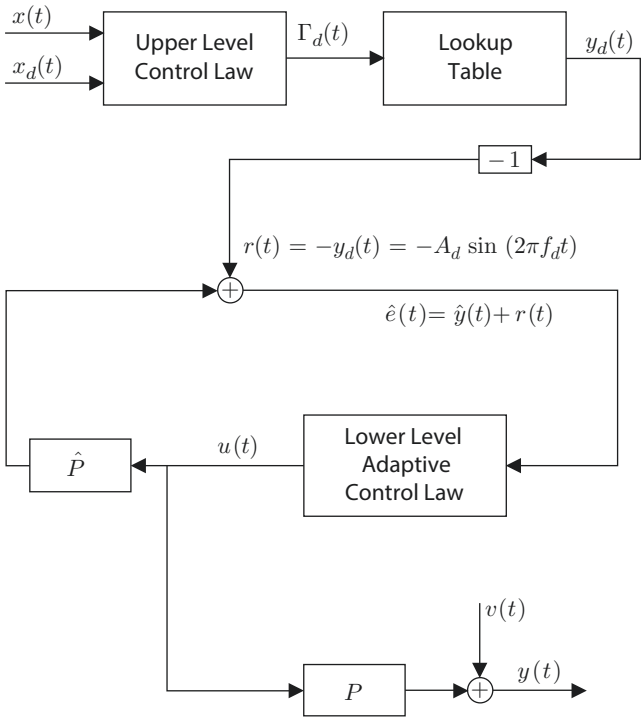


Insect Flight and Micro Air Vehicles (MAVs), Fig. 8 Empirical relationship between lift-force and actuator displacement amplitude, with f taking values of 105, 120, 135, 150, and 180 Hz

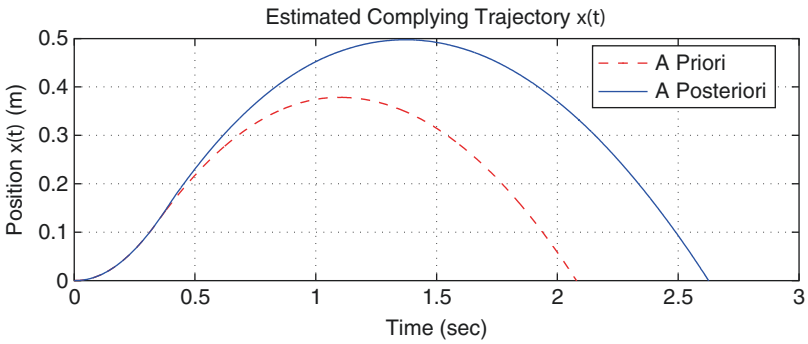
insect would move from 0 to 0.3 m and then return to 0 m in no more than 3 s. Using the simple dynamical model of the insect

$$\gamma_L - mg = m\ddot{x}, \tag{3}$$

and the experimental data obtained for plotting Fig. 8, through computer simulation the complying a priori trajectory in Fig. 10 was found. Also according to the simulation, the a priori trajectory in Fig. 10 is achievable by tracking the desired



Insect Flight and Micro Air Vehicles (MAVs), Fig. 9 Control scheme



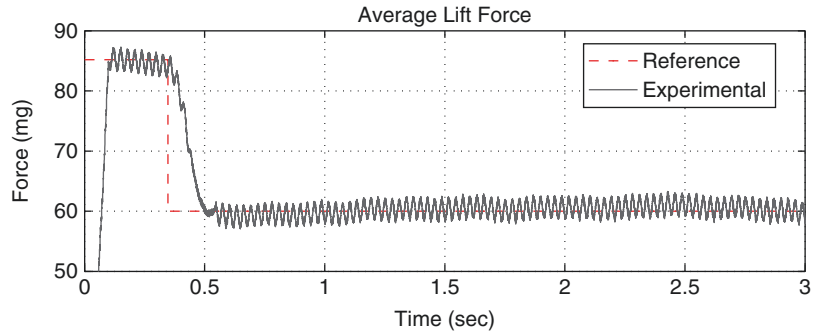
Insect Flight and Micro Air Vehicles (MAVs), Fig. 10 A priori and a posteriori estimated corresponding trajectories

lift-force signal in red in Fig. 11, where $N_L = 1{,}000$.

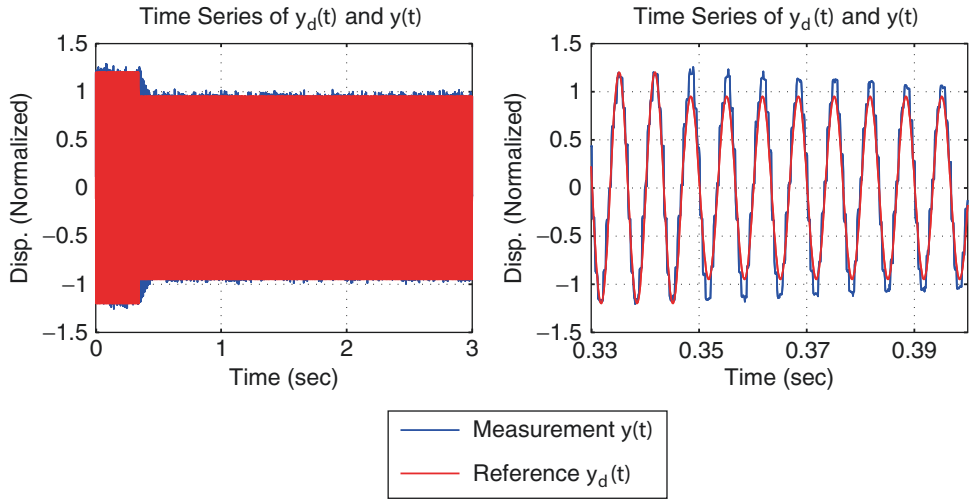
The resulting experimental average lift force is plotted in blue in Fig. 11, which using the control strategy in Fig. 9, results from choosing $y_d(t) = A_d \sin(2\pi \cdot 150\ t)$, with $A_d = 1.2$ for $t \in [0, 0.347]$ s and $A_d = 0.95$ for $t \in [0.347, 5]$ s. The time series of the reference, $y_d(t)$, and output, $y(t)$, are shown in Fig. 12. Here, on the left the complete signals are compared, and on the right the transition from $A_d = 1.2$ to $A_d = 0.95$ is shown. Notice that $y(t)$ is capable of following $y_d(t)$ and that the transition is smooth, because P is under the control of the feedforward scheme in Fig. 9. According to the hardware-in-the-loop experiments, the estimated resulting *a posteriori* trajectory is shown in blue in Fig. 10.

Passive Balancing Transmission

Fully-actuated high-bandwidth control of wing trajectory is a desirable feature for an insect-scale flapping-wing robots and is a conventional approach to attaining control and stability in flight. However, mass and power constraints in



Insect Flight and Micro Air Vehicles (MAVs), Fig. 11 Reference and experimentally obtained average lift force

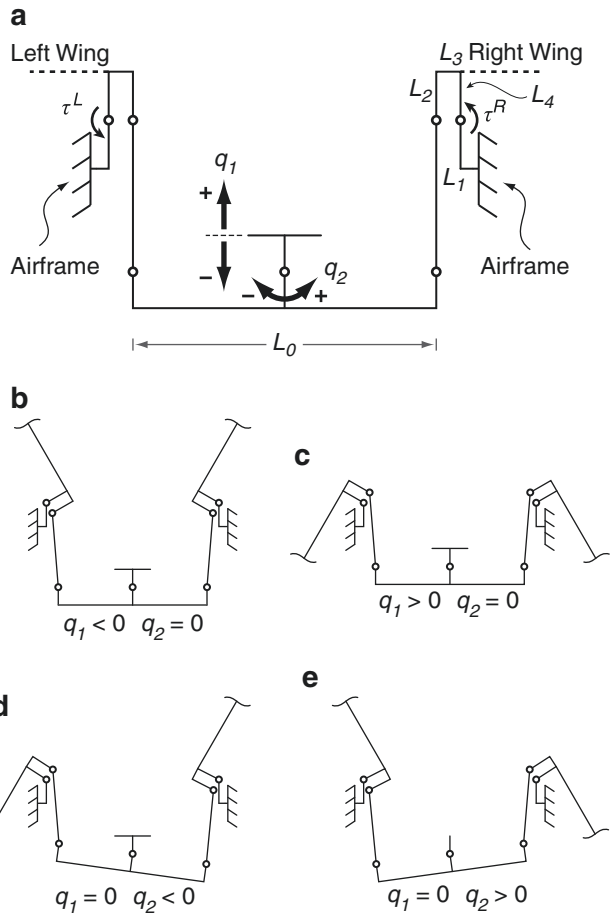


Insect Flight and Micro Air Vehicles (MAVs), Fig. 12 Comparison of time series $y_d(t)$ and $y(t)$ generating the average lift force in Fig. 11. *Left Plot:* complete series. *Right Plot:* transition from $A_d = 1.2$ to $A_d = 0.95$

such devices motivate the investigation of unconventional control methodologies. A novel methodology, called Passive Aeromechanical Regulation of Imbalanced Torques (PARITy), relies on passive elements to achieve control objectives and has shown much promise in simplifying flight control for a flapping-wing robotic insect.

Passively compliant or under-actuated mechanisms have been used in robotic insects to reduce mechanical complexity, exemplified by the passive hinges allowing the wings of the prototype RoboBee to rotate in flight. The benefits of passive elements, however, can extend beyond reducing mechanical complexity to aid in flight control. The Drag PARITy is an example of such an under-actuated transmission and has been used successfully in the RoboBee to eliminate roll torques imparted to the airframe during operation [18].

Referring to Fig. 13, the Drag PARITy introduces a second degree of freedom to the standard 1DOF RoboBee transmission. Motion along the standard degree of freedom q_1 results in a symmetric flapping motion of the wings, while the additional degree of freedom q_2 allows differential motion between the wings, coupling the upstroke of one wing with the downstroke of the other. This additional degree of freedom is not associated with any actuator; instead, motion along q_2 is determined purely by the dynamics of the system. In this fashion, wing trajectories are passively altered in response to asymmetric wing loading,



Insect Flight and Micro Air Vehicles (MAVs), Fig. 13 (**a**) Planar kinematics of the drag PARITy transmission. Holding $q_2 = 0$ and allowing q_1 to oscillate between (**b**) negative and (**c**) positive values produces a symmetric flapping motion. Holding $q_1 = 0$ and allowing q_2 to take (**d**) negative and (**e**) positive values produces a differential flapping motion, coupling the upstroke of one wing with the downstroke of the other

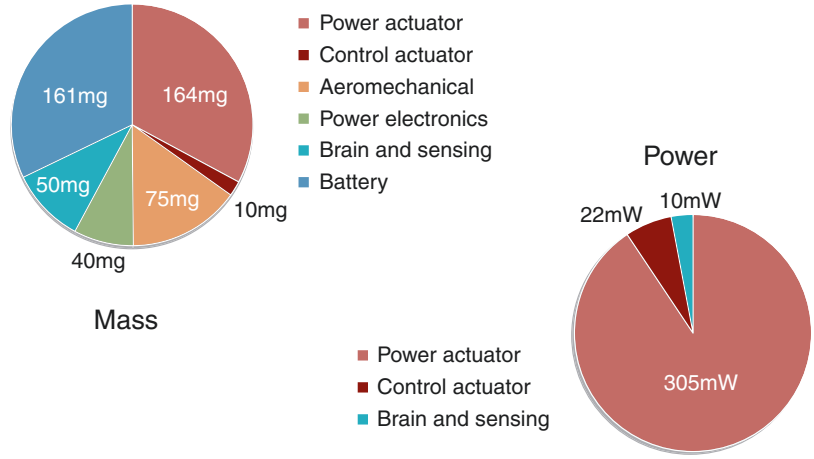
resulting in a reduction in roll torque experienced by the body of the robot.

Research into PARITy is motivated by the realization that wing trajectories are not of fundamental importance to a flapping-wing device. Ideally, the wings would execute whatever trajectories are necessary to impart the desired forces and torques on the airframe. The fundamental importance of forces and torques over wing kinematics suggests the use of force and torque control instead of kinematic wing control. Passive elements such as those demonstrated by Drag PARITy can alter wing kinematic trajectories on a short timescale in response to forces and torques. In this manner, PARITy enabled devices are expected to lower performance requirements on mass and power limited active controllers.

In a PARITy system, active control is achieved not by altering the wing trajectories directly, but by modulating the dynamics of the short timescale passive systems. For example, instead of balancing roll torques, an active control input could bias the system to attain a 10 % higher reaction torque from one wing as compared to the other. It is anticipated that enabling such direct control over realized forces and torques will greatly simplify the control problem for mass-limited flapping-wing aeromechanical platforms.

Power Considerations

Power is a key challenge in achieving autonomous operation for insect-sized flapping-wing robots. The stringent weight and power requirements of insect-scale flight, especially when hovering, make it difficult to achieve autonomous operation without a thorough understanding of



Insect Flight and Micro Air Vehicles (MAVs), Fig. 14 Mass and power budgets for a hypothetical 500 mg robot

the interactions and tradeoffs between various components. There are three primary subsystems involved in flight energetics: the aerodynamic components (wings), the power actuators, and the energy source. These subsystems are linked by two transduction mechanisms: the mechanical transmission, which serves as the interface between the actuator and the aerodynamic components, and the power electronics, which serve as the interface between the energy source and the actuator. The remaining subsystems may include: structural, control electronics, sensing, communications, and energy harvesting components.

In the study of power and flight energetics, all subsystems except for the aerodynamic components, the power actuators, and the energy source are classified as payload. Payload is divided into two categories: static payload, which has a fixed mass, and proportional payload, where the mass scales as a percentage of the total mass of the robot.

Continuing advances in the design and characterization of the various subsystems have enabled system-level models of power and energy. The benefits of system-level modeling include analysis of design tradeoffs, optimization for various performance metrics (e.g., flight time), comparison to biological insects and other aerial platforms, and the ability to estimate the mass and power consumption of system components throughout the design space. For example, Fig. 14 shows the mass and power breakdown for a hypothetical 500 mg (total body mass, with all components integrated) flapping-wing robotic insect, the target vehicle of the RoboBees project [19].

Energy Source

Promising energy sources for robotic insects include batteries, fuel cells, ultracapacitors, and solar cells. At present, lithium polymer batteries are the only developed, commercially available technology that can satisfy the energy and power requirements of flapping-wing robotic insects. According to the latest models, current battery technology can yield devices that can operate for less than a minute before recharging. A number of new energy storage technologies are currently under development, including micro-solid-oxide fuel cells, lithium batteries with silicon nanowire anodes, and lithium-air batteries. Some of these technologies have the potential to exceed the energy density of current lithium polymer batteries by over an order of magnitude, which will greatly increase the operating time of insect-sized vehicles.

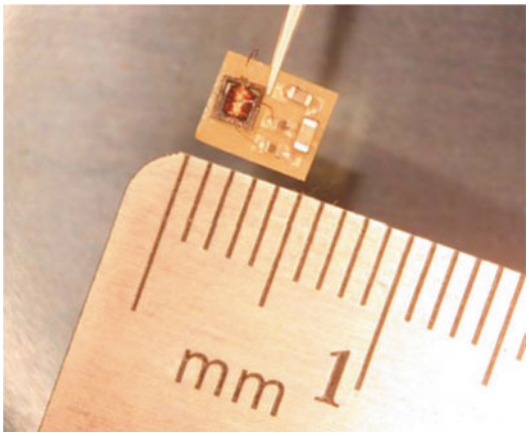
Power Electronics

The wing flapping motion in the RoboBee is powered by piezoelectric actuators, which offer robust mechanical performance, high efficiency, and high-power density at small scales, and are expected to outperform conventional actuators, such as DC motors, in insect-scale devices [6]. From the perspective of power electronic design, piezoelectric actuators present two

challenges. Firstly, only a fraction of the input electrical energy is converted to mechanical output, while the remainder is stored in the capacitive structure of the actuator and must be recovered to maximize system efficiency. Secondly, the actuators require high voltages to operate; for example, in the RoboBee, the actuators are driven in the range of 200–300 V. At the same time, prospective energy sources have output voltages below 5 V, while connecting many unit cells in series to obtain high voltage is generally not practical because the packaging overhead causes a significant reduction in energy density.

The power electronics must convert the low input voltage from the energy source into a high-voltage drive signal and recover unused energy from the actuator. Careful selection of switching circuit topologies amenable to miniaturization while retaining high efficiency has enabled milligram-scale step-up voltage converters and drive stages, such as the 20 mg voltage converter in Fig. 15. To minimize weight, the power circuits use a lightweight flexible substrate, chip-scale or bare-die semiconductors, and custom laser-micromachined magnetic components [20].

In order to drive piezoelectric actuators as efficiently as possible, a custom integrated control circuit is used to modulate the high-voltage switches in the power circuits in a way that applies the required drive signal to the actuator terminals while recovering unused electrical energy. The control circuit is manufactured in 130 nm CMOS technology and consumes less than 200 μW, resulting in a negligible effect on system efficiency. The ultimate vision for a milligram-scale power electronics package involves the integration of the power circuits with the control circuit in chip-scale or bare-die form (Fig. 16).



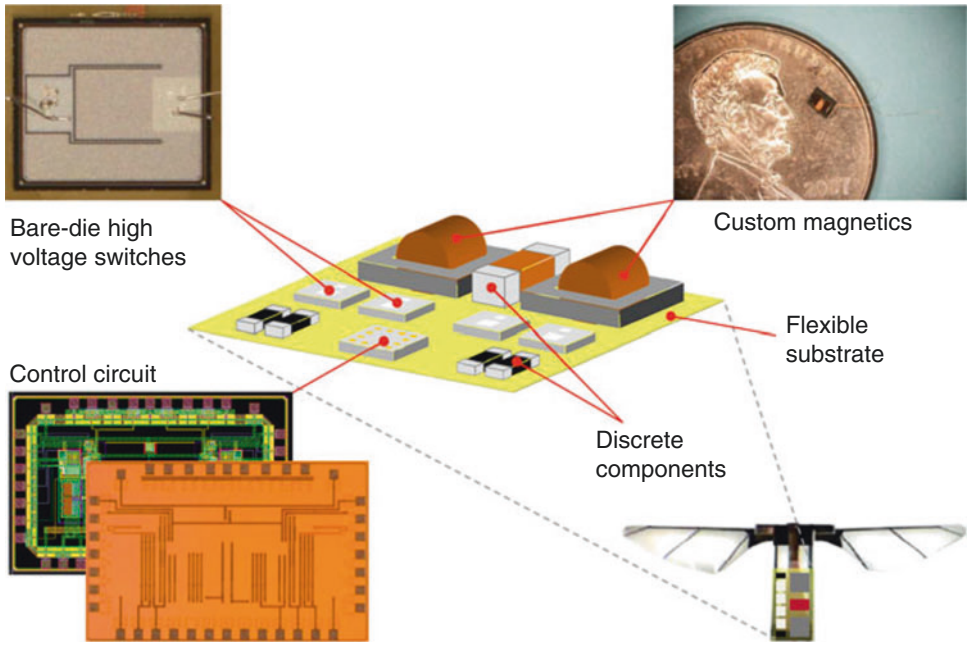
Insect Flight and Micro Air Vehicles (MAVs), Fig. 15 Twenty milligram voltage converter capable of stepping up a 3.7 V Li-poly battery to 200 V at 70 mW with external control functionality. The converter uses an 8 mg custom laser-micromachined flyback transformer

Conclusion

This entry has described enabling technologies for the aeromechanical platform of a flapping-wing robotic insect. This has only focused on the goal of achieving (a) thrust greater than weight, (b) power autonomy, and (c) stable hovering. There are myriad additional topics that need to be addressed in order for these devices to perform useful tasks in natural environments. Sensors for exteroception and proprioception need to be developed and migrated onboard. Low-power computer architectures for flight control and behaviors must reconcile the needs of processing sensory information and calculating actuator responses with the low available power, mass, and volume for anything beyond the flight apparatus. Finally, new methods to program these devices will need to overcome the relative simplicity of the individuals in order to create robust behaviors across a large group of robots.

Beyond the space of applications for insect-like robots, the constituent components are invaluable to help elucidate numerous open scientific questions. Examples include the use of wing fabrication techniques to study the structure-function relationship in natural airfoils and studies of control techniques for under-actuated and computation-limited systems. Various technologies are also motivated by the challenges of robotic insects. Some of these are under development, such as milligram-scale power, low-power computation, novel sensors, and batch meso and microfabrication techniques, while others are more longer-term such as energy harvesting methods.

I



Insect Flight and Micro Air Vehicles (MAVs), Fig. 16 Power electronics vision

Acknowledgments This work was supported by the National Science Foundation, the Office of Naval Research, the Army Research Laboratory, the Air Force Office of Scientific Research, and the Wyss Institute for Biologically Inspired Engineering. Any opinions, findings, and conclusions or recommendations expressed in this entry are those of the author and do not necessarily reflect the views of the National Science Foundation.

Cross-References

- ▶ Basic MEMS Actuators
- ▶ Biomimetics
- ▶ Compliant Mechanisms
- ▶ Nanorobotics

References

1. Petersen, K.: Silicon as a mechanical material. Proc. IEEE **70**, 420–457 (1982)
2. Yeh, R., Kruglick, E., Pister, K.: Surface-micromachined components for articulated microrobots. J. Microelectron. Mech. Syst. **5**(1), 10–17 (1996)
3. Donald, B., Levey, C., McGray, C., Paprotny, I., Rus, D.: An untethered, electrostatic, globally controllable mems micro-robot. J. Microelectron. Mech. Syst. **15**(1), 1–15 (2006)
4. Pister, K., Judy, M., Burgett, S., Fearing, R.: Microfabricated hinges. J. Sens. Actuator A Phys. **33**, 249–256 (1992)
5. Wood, R., Avadhanula, S., Sahai, R., Steltz, E., Fearing, R.: Microrobot design using fiber reinforced composites. J. Mech. Des. **130**(5), 052304 (2008)
6. Wood, R.J., Steltz, E., Fearing, R.S.: Optimal energy density piezoelectric bending actuators. J. Sens. Actuator A Phys. **119**(2), 476–488 (2005)
7. Wood, R.L.: The first flight of a biologically-inspired at-scale robotic insect. IEEE Trans. Robot. **24**(2) (2008)
8. Dudley, R.: The Biomechanics of Insect Flight: Form, Function, Evolution. Princeton University Press, Princeton (2000)
9. Dickinson, M.H., Lehmann, F.-O., Sane, S.P.: Wing rotation and the aerodynamic basis of insect flight. Science **284**, 1954–1960 (1999)
10. Sane, S.P.: The aerodynamics of insect flight. J. Exp. Biol. **206**(23), 4191–4208 (2003)
11. Lentink, D., Dickinson, M.H.: Rotational accelerations stabilize leading edge vortices on revolving fly wings. J. Exp. Biol. **212**(16), 2705–2719 (2009)
12. Ennos, A.R.: The inertial cause of wing rotation in Diptera. J. Exp. Biol. **140**, 161–169 (1988)
13. Wood, R.J., Cho, K.-J., Hoffman, K.: A novel multi-axis force sensor for microrobotics applications. Smart Mater. Struct. **18**(12), 125002 (2009)

14. Whitney, J.P., Wood, R.J.: Aeromechanics of passive rotation in flapping flight. J. Fluid Mech. **660**, 197–220 (2010)
15. Wootton, R.J., Herbert, R.C., Young, P.G., Evans, K. E.: Approaches to the structural modelling of insect wings. Philos. Trans. R. Soc. Lond. B Biol. Sci. **358**, 1577–1587 (2003)
16. Tanaka, H., Wood, R.J.: Fabrication of corrugated artificial insect wings using laser micromachined molds. J. Micromech. Microeng. **20**, 075008 (2010)
17. Finio, B.M., Wood, R.J.: Distributed power and control actuation in the thoracic mechanics of a robotic insect. Bioinspir. Biomim. **5**, 045006 (2010)
18. Sreetharan, P.S., Wood, R.J.: Passive aerodynamic drag balancing in a flapping wing robotic insect. J. Mech. Des. **132**(5), 051 006–051 016 (2010)
19. Karpelson, M., Whitney, J., Wei, G.-Y., Wood, R.: Energetics of flapping-wing robotic insects: towards autonomous hovering flight. In: Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, Taipei, pp. 1630–1637 (2010)
20. Karpelson, M., Whitney, J., Wei, G.-Y., Wood, R.: Design and fabrication of ultralight high-voltage power circuits for flapping-wing robotic insects. In: Applied Power Electronics Conference and Exposition (APEC), 2011 Twenty-Sixth Annual IEEE, Fort Worth (2011)

Insect Infrared Sensors

Helmut Schmitz[1], Helmut Soltner[2] and Herbert Bousack[3]
[1]Institute of Zoology, University of Bonn Poppelsdorfer Schloss, Bonn, Germany
[2]Zentralabteilung Technologie, Forschungszentrum Jülich GmbH, Jülich, Germany
[3]Peter Grünberg Institut, Forschungszentrum Jülich GmbH, Jülich, Germany

Synonyms

Biomimetic infrared detector; Thermal infrared detector

Definition

Technical thermal infrared (IR) detectors are defined as sensors that detect incident IR by the induced increase of the temperature of a detector target. The temperature increase can be measured by many different methods, e.g., change of electrical conductivity, capacitance, and resistance. Some animals use this detection mechanism, e.g., snakes and beetles.

Overview

Biology and Behavior of Pyrophilous Insects

Fire-loving (pyrophilous) insects depend on forest fires for their reproduction. Such insects approach ongoing fires and invade the burnt area immediately after a fire. For the long-range navigation toward a fire as well as for the short-range orientation on a freshly burnt area, these insects have special sensors for smoke and infrared (IR) radiation. Whereas the olfactory receptors for smoke are located on the antennae, the IR receptors are housed in extra-antennal sensory organs, which can be found on the thorax or on the abdomen. It is suggested that smoke-derived odors and IR information converge on brain neurons which, in turn, control and direct flight toward the forest fire.

Two genera of jewel beetles (family Buprestidae) can be classified as pyrophilous: about a dozen species of the genus *Melanophila*, which are distributed nearly all over the world except for Australia [1], and the "fire beetle" *Merimna atrata*, which is endemic to Australia [2]. Despite the fact that *Melanophila* and *Merimna* show nearly the same behavior and belong to the same family of jewel beetles, their IR receptors are totally different from each other [3].

On the freshly burnt area, the males of both genera often stay on the stems of trees close to burning or glowing wood or hot ashes. As soon as they become aware of a conspecific female, they try to copulate vigorously. After mating, the females deposit the eggs under the bark of burnt trees. The main reason for the pyrophilous behavior is that the wood-boring larvae of *Melanophila* and *Merimna* can only develop in the wood of burnt trees. As a morphological specialty, both pyrophilous buprestid genera are equipped with antennal smoke receptors and thoracic or abdominal IR organs.

Another pyrophilous beetle can be found in Australia: the "little ash beetle" *Acanthocnemus nigricans* (family Acanthocnemidae). This inconspicuous beetle is only 4 mm long and highly attracted by hot ashes. However, its biology is nearly unknown. Obviously, *Acanthocnemus* also depends on fires for its reproduction and is equipped with a pair of sophisticated prothoracic IR receptors [4]. Recently, IR receptors have also been discovered in a few pyrophilous members of the flat bug genus *Aradus* (Heteroptera, Aradidae) [5]. Interestingly, the IR receptors of *Aradus* bugs are very similar to those described in *Melanophila* beetles. Consequently, fire detection is the compelling precondition for the survival of all pyrophilous insect species mentioned above. However, the outbreak of a forest fire is highly unpredictable. Therefore, pyrophilous beetles and bugs should be able to detect fires from distances as large as possible. Furthermore, when flying over a burnt area in search for a landing ground, the small insects have to avoid "hot spots" with dangerous surface temperatures above about 60 °C.

The Photomechanic IR Receptors in Pyrophilous Beetles and Bugs

Structure of Photomechanic IR Receptors

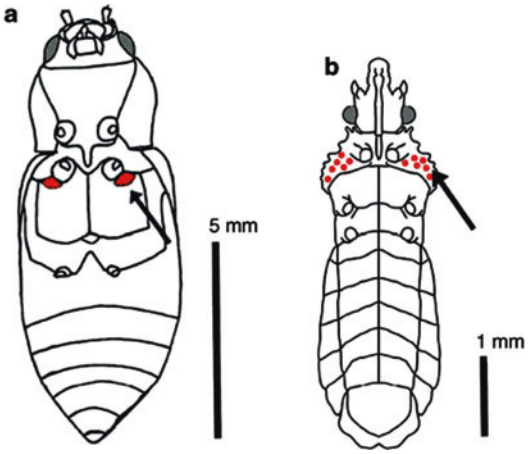
The performance of photomechanic insect IR receptors has been best studied in *Melanophila* beetles, which inhabit the Palaearctic as well as the Nearctic parts of the world. As mentioned, beetles of both sexes approach forest fires because their brood depends on burnt wood as larval food [6]. In the northern boreal forests, where most *Melanophila* species can be found, a fire statistically occurs only every 50–200 years in any one area. Therefore, it has to be postulated that the sensory organs, which are used for fire detection, have been subjected to a strong evolutionary pressure especially with regard to sensitivity. The single IR receptors (called sensilla in insects) are situated in two pit organs, which are located on the third thoracic segment (Fig. 1a). Each IR organ houses about 70 IR sensilla, which are closely packed together at the bottom of the pit [1, 7] (Fig. 2a).

In pyrophilous bugs of the genus *Aradus*, about a dozen IR sensilla are located on the lateral sides (the so-called propleurae) of the prothorax (Fig. 1b). Here, receptors are loosely interspersed between the common hair mechanoreceptors, which cover the entire outer body surface of the bug (Fig. 2b).

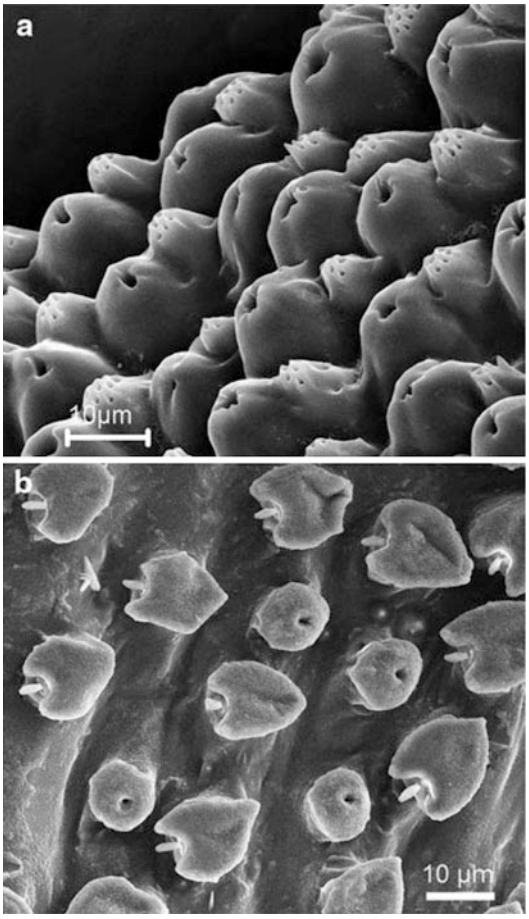
From the outside, a single IR sensillum in *Melanophila* beetles and *Aradus* bugs can be recognized by a hemispherical dome with a diameter of about 12–15 μm (Fig. 2a, b). The dome consists of a thin cuticle, which represents the outer boundary of a spherical internal cavity. The cavity is almost completely filled with a tiny cuticular sphere with a diameter of about 10 μm. Based on transmission electron microscopical observations, it has been described that the internal sphere consists of at least two different zones: (1) an outer lamellated zone consisting of exocuticle – the lamellated appearance is caused by many layers of chitin fibers with a periodically changing orientation, and (2) an internal microfluidic core (Fig. 3a, b). From below, the sphere is innervated by a single sensory cell. The outermost tip of the sensory dendrite is anchored in the fluidic core. All morphological as well as all physiological data available so far have demonstrated that this cell is a ciliary mechanoreceptor [1, 5].

Receptor Function

Currently, two models of sensillum function can be found in the literature. However, these models are inconsistent with each other. (1) More than 12 years ago, a so-called photomechanic principle was established [7]. The authors proposed that the biomolecules (i.e., proteins and chitin) of the cuticular sphere strongly absorb mid-IR radiation. In a way not described in detail, the resulting thermal expansion of the cuticular sphere is measured by the mechanoreceptor. (2) Assuming the presence of a large air-filled cavity inside the sphere, Evans [8] proposed in the second model that IR radiation enters this cavity by a small apical waveguide with a diameter of about 1.5 μm. Due to the absorption of IR photons at the inner cuticular walls of the cavity, the enclosed air is heated up and expands. In a way not further specified, the resulting increase in gas pressure should stimulate the mechanoreceptor.

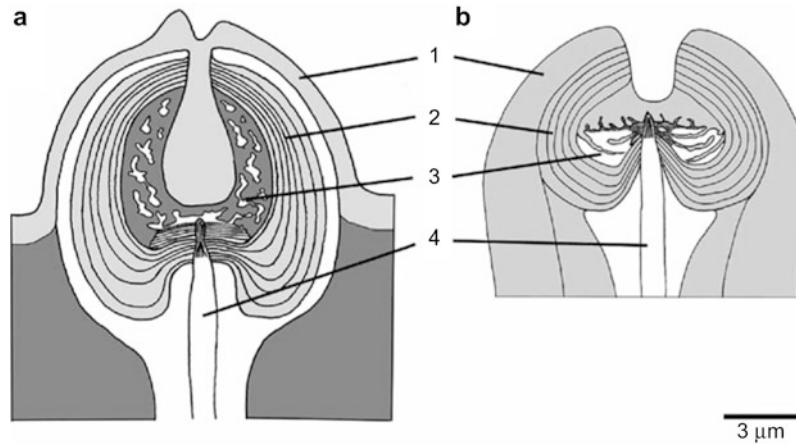


Insect Infrared Sensors, Fig. 1 Ventral view onto (**a**) *Melanophila acuminata* beetle and (**b**) *Aradus albicornis* bug. Position of IR organs and receptors are indicated by *arrows*. Legs and antenna omitted



Insect Infrared Sensors, Fig. 2 (**a**) View into an IR organ of a *Melanophila* beetle. Hemispherical IR sensilla are accompanied by porous wax glands. (**b**) IR sensilla on the propleura of an *Aradus* bug interspersed between hair mechanoreceptors with short bristles

All recent findings have clearly demonstrated that the thoracic infrared (IR) sensilla of the pyrophilous jewel beetle *Melanophila acuminata* most likely have evolved from hair mechanoreceptors (sensilla trichodea) [1]. Hair mechanoreceptors, which can be found in any insect, have a bristle of different length. In insects, the pressure-sensitive dendrite of a single mechanosensitive sensory cell innervates the base of the bristle. As a result, even a slight bending of the bristle stimulates the tip of the dendrite. To further elucidate the sensory transduction mechanism, the morphology of IR sensilla and of neighboring hair mechanoreceptors was investigated by using electron microscopical techniques (SEM, TEM) in combination with focused ion beam milling (FIB). It was assumed that any deviation from the bauplan of a hair mechanoreceptor is of special importance for the transformation of IR radiation into a mechanical stimulus. Thus, the structures supposed to be relevant for stimulus uptake and transduction were homologized. Compared to hair mechanoreceptors, minute spheres have been developed instead of bristles in photomechanic IR sensilla (see Fig. 3). Most probably, IR radiation absorbed by the proteins, the chitin fibers, and the water of the sensillum heats up the sphere, which immediately causes thermal expansion especially of the liquid inside the microfluidic core. Because the shell of the sphere consists of stiff exocuticle additionally reinforced by many layers of chitin fibers, the only compliant structure in the sphere is the membrane of the tip of the mechanosensitive dendrite. Due to the air-filled tracheal system inside the beetle's body, most of the cell body of the mechanosensitive cell always is at ambient pressure. Only the outermost tip of the sensory dendrite of the mechanosensor (cf. Fig. 3) is enclosed inside the sphere. Therefore, any increase in pressure inside the sphere will cause a cross compression of the dendritic tip which is compensated by the cell volume below the sphere. By deflection of



Insect Infrared Sensors, Fig. 3 Schematic drawings of cross sections through photomechanic IR receptors of (**a**) *Melanophila* beetles and (**b**) *Aradus* bugs (*1*): outermost layer of hard exocuticle (*2*): lamellated exocuticular shell of internal sphere (*3*): fluidic core inside the sphere (*4*): tip (dendrite) of mechanosensor

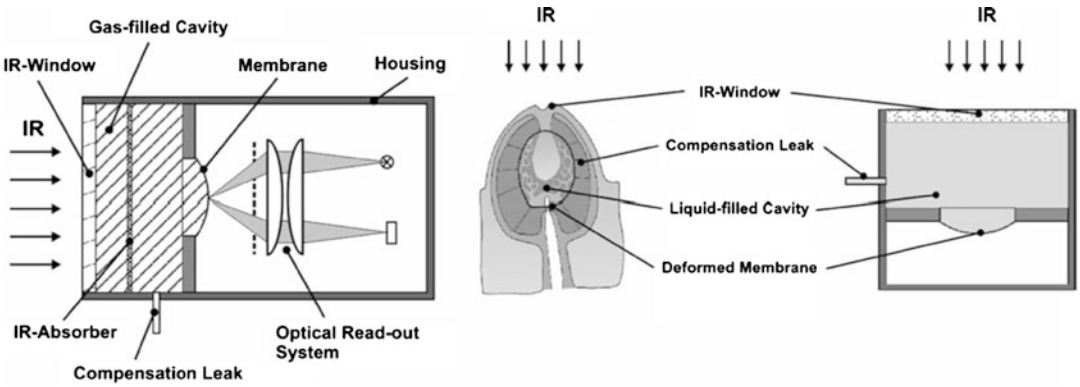
the dendritic membrane in the range of a few nanometers, stretch-activated ion channels will be opened. Accordingly, an IR sensillum acts as a microfluidic converter of infrared radiation into an increase in internal pressure inside the sphere, which is measured by the mechanosensitive neuron. In the IR sensilla of *Melanophila* beetles, fine nanocanals were found, which radially run through the outer exocuticular shell of the internal sphere (canals are too small to be shown at right dimensions in Fig. 3a). If ambient temperature slowly changes, a compensatory exchange of fluid from the inner core of the sphere prevents an unspecific stimulation of the mechanosensory cell. The principle of transforming IR radiation into a bioelectrical signal has been termed photomechanic. Although it has not been observed that the *Melanophila* beetles turn right and left in flight to search for incoming IR radiation, it has been proposed that beetles approach a fire by so-called tropotaxis (i.e., by comparing the input of two symmetrically arranged receptors on either side of their body). If the wings are used to chop incoming IR radiation has not been investigated so far.

Most probably, photomechanic IR sensilla have been developed independently in beetles and bugs. Because the products of the evolutionary process are very similar in *Melanophila* beetles and *Aradus* bugs (Figs. 2 and 3), it can be proposed that a photomechanic IR sensillum represents a highly optimized sensor for mid-IR radiation. This provides strong evidence that mechanoreceptors are promising candidates for the development of IR sensors.

Biomimetic Approach

Golay Sensor

After the clarification of the function of the biological IR sensor of the beetle, the underlying principle was analyzed mathematically to be able to design a technical sensor. For this purpose, the pneumatic Golay sensor [9] was investigated as a simple model (see Fig. 4, left). This sensor consists of an internal gas-filled cavity, which is closed on one side by a window and on the other side by a thin membrane. IR radiation enters through the window and heats up the gas by absorption. The deflection of the membrane caused by the expanding gas can be read out by an optical system [9], a capacitive detector [10], or a tunneling displacement transducer [11]. To enhance the IR absorption in the gas, the cavity is equipped with an additional absorber. Reflecting walls of the cavity are another means to enhance the absorption. The temperature changes of the gas caused by the absorbed IR radiation are in the range of mK. Because of slow variations of the ambient temperature in the range of a few K, it is absolutely necessary to integrate a leak, which compensates this influence due to an exchange of the gas with a reference volume.



Insect Infrared Sensors, Fig. 4 *Left*: principle of a gas-filled Golay sensor with optical readout. For explanations, see text. *Right*: comparison of the sensillum with the model of the sensor

Sensor Model of the IR Receptor

The model of the sensor is shown in Fig. 4, right. Similar to the sensillum, the sensor contains an internal water-filled cavity. The presence of a liquid instead of gas is a major difference between the sensillum and the Golay sensor. The cavity of the technical sensor is etched into a silicon wafer and is closed on the one side by a window (Suprasil, [12]) and on the other side by a thin silicon membrane. The IR radiation being absorbed produces a change in pressure or volume, respectively, due to the change of the state of the water. The deflection of the membrane caused by this pressure increase can be read out by, e.g., a capacitive detector or a tunneling displacement transducer.

Sensor Without Compensation Leak

For calculating the change of the liquid pressure in the cavity based on the temperature profile, the equation of state must be solved. Because the pressure in the cavity depends upon the two independent variables temperature and volume, its total change is given by

$$\Delta P = \beta \cdot \Delta T_{mean} - \frac{1}{\kappa \cdot V} \cdot \Delta V \qquad (1)$$

with ΔP as the pressure increase, $\beta = (\partial \mathrm{P}/\partial \mathrm{T})_{\mathrm{V}}$ as the isochoric tension coefficient, $\Delta \mathrm{T}_{\mathrm{mean}}$ as the mean temperature increase averaged over the cavity volume, $\kappa = -(1/\mathrm{V}) \cdot (\partial \mathrm{T}/\partial \mathrm{P})_{\mathrm{T}}$ as the isothermal compression coefficient, V as the volume of the cavity, and ΔV as the volume increase. For water (25 °C, 1 bar), $\beta = 5.68 \cdot 10^5$ Pa/K and $\kappa = 4.5 \cdot 10^{-10}$ 1/Pa [13].

The increase of the volume ΔV of the cavity results in the tiny deflection of the membrane. The local deflection y of this circular membrane caused by a pressure difference can be calculated as a function of the radial distance r with the shell theory [14, 15]

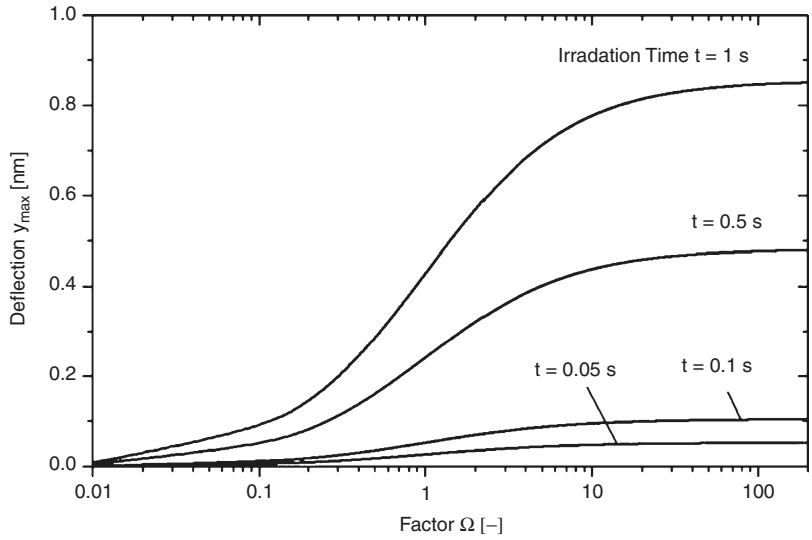
$$y(r) = \frac{\Delta P}{64 \cdot D} \cdot \left(R^2 - r^2\right)^2$$
$$D = \frac{E \cdot t_P^3}{12 \cdot (1 - v^2)} \qquad (2)$$

with R as the radius of the membrane and the cavity, D as the flexural stiffness of the membrane, E as the Young's modulus of elasticity, t_P as the thickness of the membrane, and v as the Poisson's ratio.

Equation 2 is a good approximation for small membrane deflections, i.e., $y_{max}/t_p < 1$. The corresponding volume change ΔV can be calculated in relation to the pressure increase ΔP by using Eq. 2, assuming equal radius of cavity and membrane (see Fig. 5):

$$\Delta \mathrm{V} = \frac{\pi \cdot R^6 \cdot \Delta P \cdot (1 - v^2)}{16 \cdot E \cdot tp^3} \qquad (3)$$

The combination of Eqs. 1 and 3 yields a relationship for ΔP, which considers the influence of temperature and volume. It is assumed that



Insect Infrared Sensors, Fig. 5 Maximum central deflection y_{max} of a circular membrane as function of factor Ω and irradiation time for a water-filled cavity. IR power density, 10 W/m^2; diameter and height of the cavity, 0.5 mm

there is no additional volume change in the rest of the cavity:

$$\Delta P = \frac{\beta \cdot \Delta T_{mean}}{1 + \Omega}$$
$$\Delta T_{mean} = \frac{\int_0^{H_c} \Delta T(z) \cdot dz}{Hc} \qquad (4)$$
$$\Omega = \frac{R^4 \cdot (1 - \nu^2)}{16.E.t_{p^3}.H_c.\kappa}$$

with H_C as the height of the cylindrical cavity, ΔT(z) as the temperature as a function of the axial coordinate z of the cavity and time, and ΔT_{mean} as the mean temperature increase averaged over the cavity.

The factor Ω characterizes the change of state of the liquid inside the cavity due to a temperature increase: for $\Omega \to 0$, which corresponds to an extremely hard membrane, the change of state is isochoric with a maximal pressure increase. For $\Omega \to \infty$, as for an extremely soft membrane, the change of state is isobaric with maximal volume increase. The transition between these two cases is at $\Omega \approx 1$. The maximal deflection y_{max} of the circular membrane can be calculated as a function of the factor Ω by using the Eqs. 2 and 4:

$$y_{\max} = \frac{3 \cdot \alpha \cdot \Omega \cdot T_{mean} \cdot Hc}{1 + \Omega} \qquad (5)$$

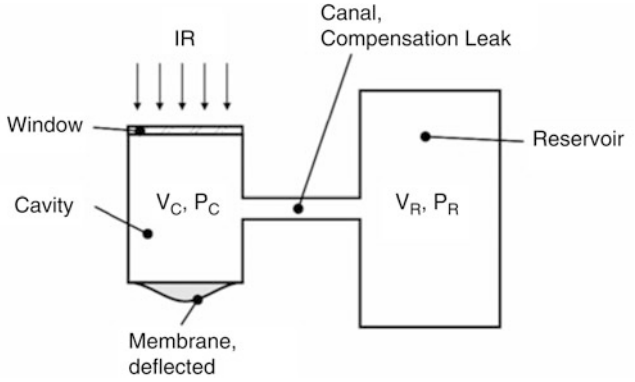
with $\alpha = \beta \cdot \kappa$ as the isobaric thermal expansion coefficient.

Figure 6 shows the maximal deflection as a function of the factor Ω according to Eq. 5 for an IR power density of 10 W/m^2 (IR window without absorption loss assumed) for a cavity with a height and a diameter of 0.5 mm, respectively, filled with water. Obviously, the maximal deflection is achieved for the isobaric case with a very soft membrane. Due to the very low deflection of less than 1 nm, a capacitor as readout must be extremely sensitive. As an alternative, an optical readout or the use of a tunneling displacement transducer may be useful.

With regard to Eqs. 4 and 5, the maximal deflection of the membrane can be increased by using a liquid with a high thermal expansion coefficient α, e.g., $120 \cdot 10^{-5}$ 1/K for methanol versus $26 \cdot 10^{-5}$ 1/K for water, and a high mean temperature increase ΔT_{mean} due to a low product of heat capacity and density in a small cavity. Additionally, the IR window and the cavity material should have a low heat conduction to avoid heat losses during measurement.

Sensor with Compensation Leak

In the *Melanophila* IR receptor, the inner sphere is enclosed by a thin layer of fluid. Therefore, the nanocanals in the shell of the sphere allow the exchange of fluid in and out of the microfluidic



Insect Infrared Sensors, Fig. 6 Model of the IR sensor with cavity (pressure P_C, volume V_C) and reservoir (pressure P_R, volume V_R). The canal of the compensation leak has a radius R_L and a length L

compartment in the sphere. Thus, any internal pressure change which may be caused by the slowly changing ambient temperature can be compensated. Golay sensors also use such compensation leaks for compensating changes of ambient temperatures [9].

However, due to the compensation leak, the pressure increase in the cavity is reduced depending on the flow resistance of the leakage through the canal. For simulating this effect, an analysis based on the model in Fig. 5 was carried out. A mass and energy balance for the subsystem's cavity, canal, and reservoir can be used to derive the formulas for the time-dependent development of temperature and pressure in the cavity.

For a laminar flow, the volume flow in the canal can be described by a Hagen-Poiseuille flow [16]:

$$\dot{V}_L = \frac{\pi \cdot R_L^4}{8 \cdot \eta \cdot L} \cdot (P_C(t) - P_R(t)) \tag{6}$$

with $\dot{V}_L$ the volumetric flow rate, R_L the radius of canal, L the length of canal, $P_C(t)$, $P_R(t)$ the time-dependent pressure in the cavity and the reservoir, and η the dynamic viscosity.

The mass balance for the cavity and the reservoir gives

$$\begin{aligned} \frac{d}{dt} m_C &= -\rho_L \cdot \dot{V}_L \\ \frac{d}{dt} m_R &= \rho_L \cdot \dot{V}_L \end{aligned} \tag{7}$$

with m_C, m_R the mass of fluid inside cavity and reservoir, respectively, and ρ_L the mean fluid density in the canal.

The density ρ_L in the Eq. 7 is the mean value between the density in the cavity ρ_c, and the density in the reservoir ρ_R. Equation 1 for an isothermal change of state and a mass balance for the canal gives

$$\begin{aligned} \frac{\rho_C}{\rho_R} &= e^{k \cdot (P_C - P_R)} \approx 1 + k \cdot (P_C - P_R) \\ \rho_L &= \frac{(\rho_C + \rho_R)}{2} \approx \rho_R \cdot \left(1 + k \cdot \frac{(P_C - P_R)}{2}\right) \end{aligned} \tag{8}$$

It is obvious that for small pressure differences between cavity and reservoir in the order of some Pa, the density in the canal is constant due to the magnitude of the compression coefficient κ. In Eq. 7, the volume of the cavity is time-dependent due to the additional volume caused by the deflection of the membrane, so that

$$\begin{aligned} \frac{d}{dt} m_C &= \frac{d(\rho_C \cdot V_C(t))}{dt} \\ &= \frac{d\rho_C}{dt} \cdot V_C(t) + \frac{dV_C(t)}{dt} \cdot \rho_C \end{aligned} \tag{9}$$

with $V_C(t)$ the time-dependent volume of the cavity and m_C the time-dependent mass of the fluid in the cavity.

Equation 3 yields

$$\frac{dV_c}{dt} = \frac{\pi \cdot R^6 \cdot (1 - \nu^2)}{16 \cdot E \cdot t_P^3} \cdot \frac{dP_c}{dt} \tag{10}$$

Equations 6, 7, 8, 9, and 10 and the simplification $\kappa \cdot (P_C - P_R) \approx 0$, which is 10^{-8}–10^{-9} for water

and a pressure difference of some 10–100 Pa, yield two differential equations for the pressures P_C and P_R:

$$\begin{aligned}(1+\Omega).\frac{dP_C}{dt}-\beta\cdot\frac{dT_C}{dt}+\frac{P_C}{\tau_C}-\frac{P_R}{\tau_C}=0\\ \frac{dP_R}{dt}+\frac{P_R}{\tau_C}+\frac{P_C}{\tau_R}=0\end{aligned} \tag{11}$$

where the time constants τ_C and τ_R are defined as

$$\begin{aligned}\tau_C&=\frac{8\cdot\eta\cdot L\cdot V_C\cdot k}{\pi\cdot R_L^4}\\ \tau_R&=\frac{8\cdot\eta\cdot L\cdot V_R\cdot k}{\pi\cdot R_L^4}\end{aligned} \tag{12}$$

Energy Balance and Pressures for the Adiabatic Cavity

The temperature gradient in Eq. 11 can be found with an energy balance for the cavity. The first law of thermodynamics assuming an adiabatic cavity results in

$$\begin{aligned}I_0\cdot A+\frac{dW_{12}}{dt}-\rho_C\cdot\dot{V}_L\cdot C_P\cdot\Delta T_C\\ =\frac{d}{dt}(\rho_C\cdot V_C(t)\cdot C_P\cdot\Delta T_C)\end{aligned} \tag{13}$$

with A the cross section of the cavity and the IR window, W_{12} the total work of the system, $\dot{V}_L$ the volume leakage flow rate in Eq. 6, $V_C(t) = V_C + dV_C(t)$ the time-dependent volume regarding Eq. 10, V_C the initial volume of the cavity at $t = 0$, and C_P the heat capacity of the fluid in the cavity.

The total work W_{12} consists of two parts: the energy stored in the membrane due to deflection, $W_{Membrane}$, and the pressure-volume work done on the environment, W_{PV}. Here, as usual, a negative work means that the work is done by the system. The energy stored in the membrane is [15]

$$\begin{aligned}W_{Membrane}=-\frac{(P_C-P_0)^2\cdot\pi\cdot R^6}{384\cdot D}\\ \cdot\left[1+0.244\cdot\left(\frac{y_{max}}{t_P}\right)^2\right]\end{aligned} \tag{14}$$

with P_0 the ambient pressure.

The first summand in the bracket of Eq. 14 is due to the energy of bending, and the second summand considers the energy of stretching the middle plane of the membrane. For small deflections where the maximum deflection of the membrane y_{max} is smaller than its thickness t_P, the energy of stretching the middle plane can be neglected.

The pressure-volume work W_{PV} can be calculated using for dV the additional volume due to the deflection of the membrane, see Eq. 3, with

$$\begin{aligned}W_{PV}&=-\int P_0\cdot dV=-P_0\cdot\frac{\pi\cdot R^6\cdot(P_C-P_0)\cdot(1-\upsilon^2)}{16\cdot E\cdot t_p^3}\\ &=-\frac{P_E\cdot(P_C-P_0)\cdot\pi\cdot R^6}{192\cdot D}\end{aligned} \tag{15}$$

Calculating the derivatives in Eq. 13 gives

$$\begin{aligned}\frac{dW_{12}}{dt}&=-\Omega\cdot V_C\cdot k\cdot P_C\cdot\frac{dP_C}{dt}\frac{d}{dt}(\rho_C\cdot V_C(t)\cdot C_P\cdot\Delta T_C)\\ &=\rho_C\cdot C_P\cdot V_C\cdot\left[(\Omega+1)\cdot k\cdot\Delta T_C\cdot\frac{dP_C}{dt}\right.\\ &\left.+(1-\alpha\cdot\Delta T_C)\cdot\frac{dT_C}{dt}\right]\end{aligned} \tag{16}$$

With Eq. 16 and some straightforward algebra, Eq. 13 yields

$$\begin{aligned}\left[(\Omega+1)\cdot k\cdot\Delta T_C+\frac{\Omega}{\rho_C\cdot C_P\cdot H_C}\right]\cdot\frac{dP_C}{dt}\\ +(1-\alpha\cdot\Delta T_C)\cdot\frac{dT_C}{dt}\\ +\frac{k}{\tau_C}\cdot(P_C-P_R)\cdot\Delta T_C=\frac{I_0}{\rho_C\cdot C_P\cdot H_C}\end{aligned} \tag{17}$$

Obviously, because $\alpha\cdot\Delta T_C \ll 1$, it can be neglected in Eq. 17. The influence of the heat loss due to the flow in the canal is represented by the last part of the left side in Eq. 17. The heat loss increases with decreasing time constant τ_C, which means a lower flow resistance and a higher mass stream between cavity and reservoir. The heat loss relative to the irradiated IR energy, e.g., for $I_0 = 10$ W/m^2, a pressure difference of about

Insect Infrared Sensors, Table 1 Dimensions and material properties [13, 17]

Radius R of the cavity and the membrane	0.25 mm
Height H_C of the cavity	0.5 mm
Time constant τ_C of the cavity	1.9 ms
Thickness t_P of the membrane	1 μm
Length L of the canal	1 mm
Young's modulus of elasticity E of the membrane	160×10^9 N/m^2
Poisson's ratio of the membrane ν	0.279
Flexural stiffness of the membrane D	$1.45 \cdot 10^{-8}$ Nm
Isochoric tension coefficient β of the fluid	5.7×10^5 Pa/K
Isothermal compression coefficient κ of the fluid	4.5×10^{-10} 1/Pa
Isobaric thermal expansion coefficient α of the fluid	2.6×10^{-4} 1/K
Density ρ of water	998 kg/m^3
Heat capacity C_P of water	4190 J/kg·K
Dynamic viscosity η of water	10^{-3} Pa

10 Pa, a temperature difference of about 10 mK, and the values in Table 1 is

$$\frac{k \cdot \rho_C \cdot C_P \cdot H_C}{\tau_C \cdot I_0} \cdot (P_C - P_R) \cdot \Delta T_C = 10^{-6} \div 10^{-7} \tag{18}$$

Regarding this estimate, the heat loss due to the flow in the canal can be neglected also, depending on the magnitude of the time constant τ_C. The relation of the derivate of the total work dW_{12}/dt and the absorbed IR energy is about 10^{-3}–10^{-4} after one second for an adiabatic cavity without compensation canal. Therefore, the influence of the total work of the system will be not neglected.

With the simplifications and the introduction of the power density factor γ, Eq. 17 results in

$$\left[(\Omega+1) \cdot \alpha \cdot \Delta T_C + \frac{\alpha \cdot \Omega}{\rho_C \cdot C_P} \cdot P_C\right] \cdot \frac{dP_C}{dt} + \beta \cdot \frac{dT_C}{dt} = \gamma$$
$$\gamma = \frac{I_0 \cdot \beta}{\rho_C \cdot C_P \cdot H_C} \left[\frac{Pa}{s}\right] \tag{19}$$

Introducing the temperature gradient in Eq. 19 in the first differential equation of Eq. 11 can be written as

$$\left(1 + \Omega + \frac{\alpha \cdot \Omega}{\rho_C \cdot C_P} \cdot P_C\right) \cdot \frac{dP_C}{dt} + \frac{P_C}{\tau_C} - \frac{P_R}{\tau_C} = \gamma$$
$$\frac{dP_R}{dt} + \frac{P_R}{\tau_R} - \frac{P_C}{\tau_R} = 0 \tag{20}$$

The Eq. 20 can be transformed for $\Omega = 0$ (isochoric, no deflection of the membrane) and $\gamma = 0$ (no IR power density) into the same equation published earlier [18].

For water as fluid, the magnitude $\frac{\alpha}{\rho_C \cdot C_P}$ of is about 10^{-10} [m^3/J], which can be neglected in Eq. 20.

For this case results

$$(1 + \Omega) \cdot \frac{dP_C}{dt} + \frac{P_C}{\tau_C} - \frac{P_R}{\tau_C} = \gamma$$
$$\frac{dP_R}{dt} + \frac{P_R}{\tau_R} - \frac{P_C}{\tau_R} = 0 \tag{21}$$

The analytical solution of Eq. 21 is

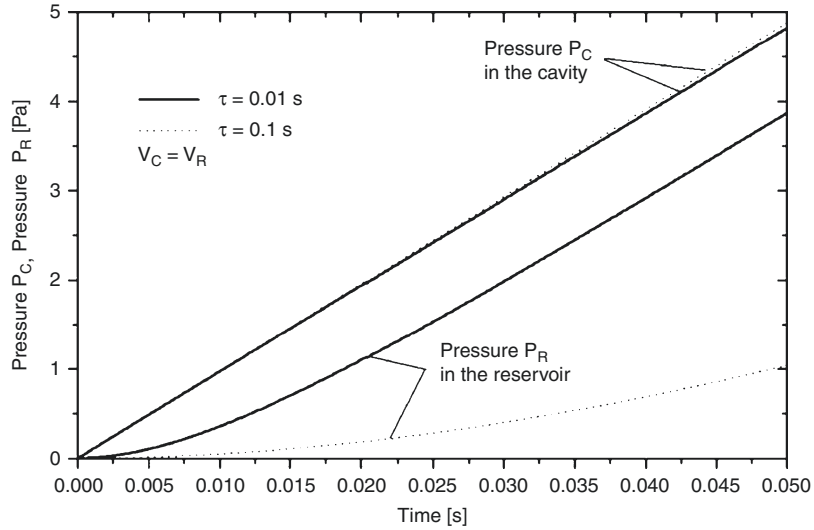
$$P_C(t) = \frac{\tau}{\tau_C \cdot (1+\Omega)} \cdot \left[\gamma \cdot \frac{\tau_C}{\tau_R} \cdot t + P_R(t=0) + \frac{\tau_C}{\tau_R} \cdot (1+\Omega) \cdot P_C(t=0) - \Phi + (\Delta P(t=0) + \Phi) \cdot e^{-\frac{t}{\tau}}\right]$$

$$P_R(t) = \frac{\tau}{\tau_C \cdot (1+\Omega)} \cdot [P_R(t=0) + \frac{\tau_C}{\tau_R} \cdot (1+\Omega) \cdot P_C(t=0) + \gamma \cdot \frac{\tau_C}{\tau_R} \cdot (t-\tau) - \frac{\tau_C}{\tau_R} \cdot (1+\Omega) \cdot (\Delta P(t=0) + \Phi) \cdot e^{-\frac{t}{\tau}}]$$

$$\Delta P(t=0) = P_C(t=0) - P_R(t=0)$$
$$\tau = \frac{\tau_C \cdot (1+\Omega)}{1 + \frac{\tau_C}{\tau_R} \cdot (1+\Omega)} \Phi = \gamma \cdot \tau_C \cdot \left(\frac{\tau}{\tau_R} - 1\right) \tag{22}$$

The pressure difference ΔP_{CR} between cavity and reservoir using Eq. 22 yields

$$\Delta P_{CR}(t) = P_C(t) - P_R(t) = \frac{\tau \cdot \gamma}{1+\Omega} \cdot \left(1 - e^{-\frac{t}{\tau}}\right) \tag{23}$$



Insect Infrared Sensors, Fig. 7 Time-dependent pressure in the adiabatic cavity and the reservoir with respect to the values in Table 1 and $\Omega = 50$, $\gamma = 5000$ ($I_o = 18$ W/m^2), $\tau = 0.01$ s, and $\tau = 0.1$ s for equal size of cavity V_C and reservoir V_R

Based on the values in Table 1 for the cavity and the canal, the time-dependent pressure in the cavity and the reservoir was calculated (see Fig. 7).

Here, the volume of the cavity V_C and the volume of the reservoir V_R are equal, and the diameter of the canal is varied, which gives two different time constants: $\tau = 0.01$ s for a larger diameter and 0.1 s for a smaller diameter, respectively. As expected, with a larger diameter of the canal, the pressure in the reservoir increases faster, and even with the larger diameter of the canal, the reduction of the pressure increase in the cavity is negligible due to the small mass stream leaving the cavity.

For an optimal readout of the sensor, the deflection of the membrane should be maximized; that means the compensation should have only a small influence on the pressure increase in the cavity. Figure 8 shows the influence of the diameter of the canal (as parameterized by the variable time constant) and the factor Ω on the maximal deflection of the membrane for equal sizes of the volumes of cavity and reservoir, respectively. For the calculation of the maximal deflection, Eq. 2 with $r = 0$ and the cavity pressure according to Eq. 22 for $t = 0.5$ s was used.

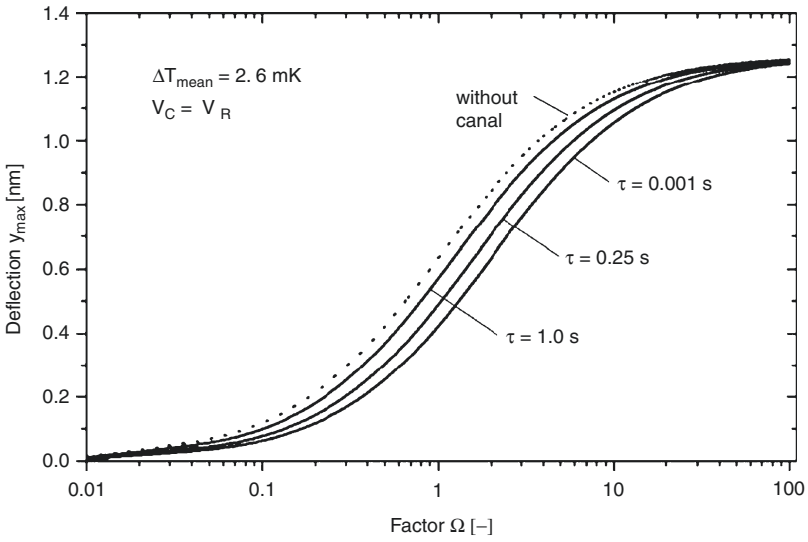
With a reduction of the flow resistance due to the increasing diameter of the canal or a shorter length of the canal (shorter time constant), the deflection is only less influenced, especially in the medium range of factor Ω from 0.1 to 10. Obviously, the change of the deflection is limited with smaller time constant, and the run of the curves does not change for time constant smaller than 0.001 s. In this case, the flow resistance is so small that no pressure difference between cavity and reservoir exists, and the two volumes act as one volume. The curve progression of the maximal membrane deflection of a cavity without canal using Eq. 5 is the limit case for a canal with a very high time constant (high flow resistance). This means the Eqs. 5 and 2 using the pressure according to Eq. 22 coincide for a time constant $\tau \rightarrow \infty$.

The maximal deflection for an increasing volume of the reservoir and a constant value of the time constant (constant flow resistance of the canal) is shown in Fig. 9. In this case, the deflection is reduced in a wide range of factor Ω, especially due to the low time constant. The reduction can be decreased by a higher time constant. In any case, the reduction is less with a high factor Ω, corresponding to a very soft membrane.

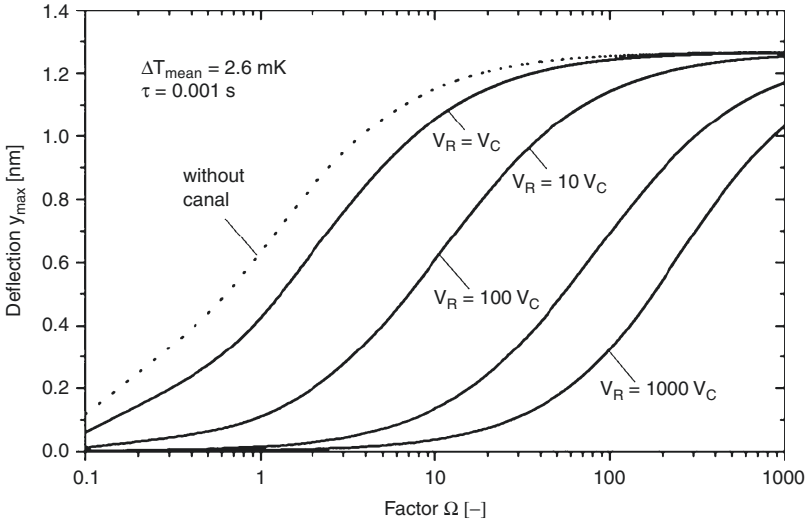
For analyzing the influencing factors on the flow rate in the canal, the Eqs. 6 and 23 are combined:

$$V_L = \frac{\pi \cdot R_L^4}{8 \cdot \eta \cdot L} \cdot \frac{\tau \cdot \gamma}{1 + \Omega} \cdot \left(1 - e^{-\frac{t}{\tau}}\right) \qquad (24)$$

Because the power density factor γ, see Eq. 19, is proportional to the IR power density I_0, and for an



Insect Infrared Sensors, Fig. 8 Maximal deflection of the membrane as function of the factor Ω and the time constant τ (variable flow resistance in the canal) for equal volumes of cavity and reservoir, V_C and V_R, at time t = 0.5 s. γ = 3000 (I_0 = 10 W/m^2), mean temperature ΔT_{mean} = 2.6 mK



Insect Infrared Sensors, Fig. 9 Maximal deflection of the membrane as function of the factor Ω and different relations of volumes of cavity and reservoir, V_C/V_R, for equal time constant τ = 0.001 s (constant flow resistance in the canal) at time t = 0.5 s. γ = 3000 (I_0 = 10 W/m^2), mean temperature ΔT_{mean} = 2.6 mK

adiabatic cavity the relation between I_0 and the temperature change in the cavity is

$$I_0 = \frac{m \cdot C_P}{A} \cdot \frac{dT_C}{dt} \tag{25}$$

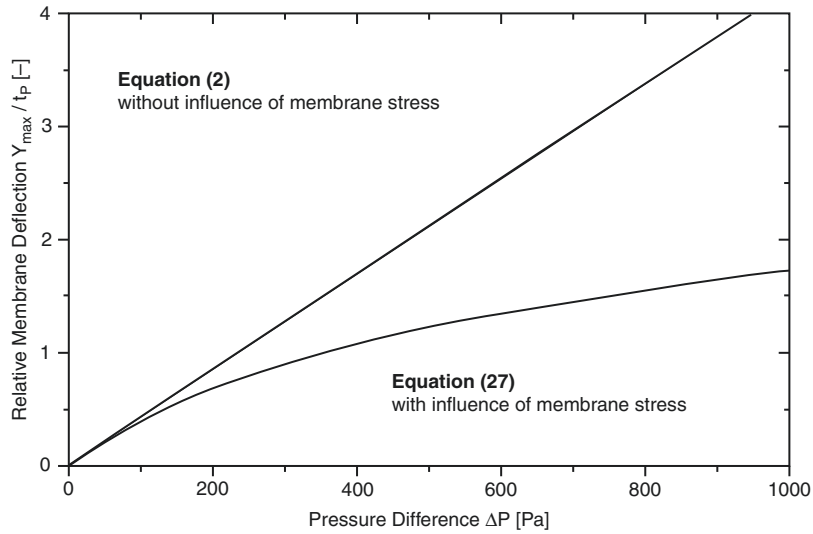
the following result is obtained:

$$\dot{V}_L \propto \frac{dT_C}{dt} \cdot \left(1 - e^{-\frac{t}{\tau}}\right) \tag{26}$$

This means, for short times $t \ll \tau$ the volumetric flow rate is about zero and the pressure increase in the cavity due to a sudden IR radiation is not influenced. For long times $t \gg \tau$ (slow ambient temperature increase), the volumetric flow rate is proportional to the temperature change in the cavity.

Compensation of Ambient Temperature Changes

The effect of the compensation will be demonstrated for a cavity based on the data in Table 1.



Insect Infrared Sensors, Fig. 10 Relative membrane deflection at r = 0 as function of applied pressure difference comparing the influence of membrane stress. The data of the membrane are taken from Table 1

For the layout of the compensation, an ambient temperature increase of 10 K in 6 h, corresponding to an IR power density I_0 of about 1 W/m^2 for an adiabatic cavity, is assumed. A cavity without temperature compensation will build up a pressure difference of at least 900 Pa during 6 h, probably higher because of the smaller deflection of the membrane due to the nonlinearity of Eq. 2. As a consequence of this high pressure bias, the membrane will be out of the linear range ($y_{max}/t_P < 1$) because the ratio of the maximal deflection y_{max} to the membrane thickness t_P is about 4 µm/1 µm, and a small pressure increase due to an incoming IR radiation will only result in a reduced additional deflection of the membrane. This effect can be explained with an improved equation compared to Eq. 2, which is also applicable for $y_{max}/t_P > 1$ [15] because it takes into account the membrane stress:

$$\frac{y_{\max}}{t_P} = \frac{\Delta P}{64 \cdot D \cdot t_P} \cdot \frac{R^4}{1 + 0.488 \cdot \left(\frac{y_{\max}}{t_P}\right)^2} \qquad (27)$$

The implicit Eq. 27 changes over to Eq. 2 with r = 0 for $y_{max}/t_P \rightarrow 0$. Figure 10 shows the comparison of the Eqs. 2 and 27. Obviously, Eq. 2 is only applicable for $y_{max}/t_P < 1$.

For the evaluation of the compensation effect a canal of 1 mm length and a diameter of 10 µm connected to a reservoir volume with constant ambient pressure, P_0 is assumed. This can be achieved, e.g., with a reservoir volume that is connected to the environment with a soft membrane so that the pressure in the reservoir is equal to the ambient pressure. A constant pressure in the reservoir is equivalent to $V_R \gg V_C$, and the time constant in Eq. 22 reduces to $\tau = \tau_C \cdot (1 + \Omega)$. With the data in Table 1 and Eq. 23, a constant pressure difference $P_C - P_0$ in the cavity of 0.1 mPa results for $t \gg \tau$. This low pressure bias results in a $y_{max}/t_P \ll 1$, and the membrane can be used in the linear range.

Assuming an increase of the pressure in the cavity of 1 Pa due to an IR source without compensation, the pressure in the cavity will increase from 900 Pa to 901 Pa, and according to Eq. 27 the maximum deflection y_{max} will only increase by about 1 nm. With compensation, the pressure will increase from 0.1 mPa to 1.0001 Pa and the deflection in the linear range is increased by 4.2 nm. This means, with an appropriate compensation the deflection of the membrane can be clearly improved, because it can operate in its linear range.

Future Directions of Research

Like in the photomechanic insect IR receptors, the technical model features very small displacements of the membrane in the sensing device in the range of 1 nm and less. The insects, however, have the advantage of possessing a very sensitive mechanoreceptor, the mechanosensitive neuron, which can detect minute deflections of its membrane [19]. It seems that this sensitivity of the beetle is difficult to achieve for a technical sensor.

For a miniaturized technical sensor, an appropriate technical readout mechanism with a high resolution of 1 nm may be necessary. So far, a sensor with a capacitor as readout is near to completion in the *Center of Advanced European Studies and Research* (caesar), Bonn. Several readout methods will be evaluated, but it has to be stressed that this search must yield a rugged and cost-effective design in order to be able to compete with existing IR sensors.

An obvious method to enhance the displacement of the membrane is to choose liquids with optimal thermal properties, e.g., high thermal expansion coefficient and low heat capacity. Here, several liquids permit an improvement. Methanol is a good candidate compared to n-pentane and toluene because of better handling during the filling process of the cavity [20].

Additionally, care has to be taken in selecting appropriate wall materials with the same low thermal conduction necessary to preserve the thermal energy inside the cell. Until now, silicon is used due to its well-known manufacturing technology for microsystems. As an alternative, microinjection molding of plastic materials with low heat conductivity are under investigation.

Cross-References

- ▶ Biomimetics
- ▶ Biosensors
- ▶ CMOS MEMS Biosensors

References

1. Schmitz, A., Sehrbrock, A., Schmitz, H.: The analysis of the mechanosensory origin of the infrared sensilla in *Melanophila acuminata* (Coleoptera; Buprestidae) adduces new insight into the transduction mechanism. Arthropod Struct. Dev. **36**, 291–303 (2007)
2. Poulton, E.B.: The habits of the Australian "fire-beetle" *Merimna atrata*, Lap. et Gory. Trans. Entomol. Soc. Lond., pt.1, proc: iii–iv (1915)
3. Schmitz, H., Trenner, T.: Electrophysiological characterization of the multipolar thermoreceptors in the "fire-beetle", *Merimna atrata* and comparison with the infrared sensilla of *Melanophila acuminata* (both Coleoptera, Buprestidae). J. Comp. Physiol. A **189**, 715–722 (2003)
4. Kreiss, E.J., Schmitz, A., Schmitz, H.: Morphology of the prothoracic discs and associated sensilla of *Acanthocnemus nigricans* (Coleoptera, Buprestidae). Arthropod Struct. Dev. **34**, 419–428 (2005)
5. Schmitz, A., Schätzel, H., Schmitz, H.: Distribution and functional morphology of photomechanic infrared sensilla in pyrophilous flat bugs of the genus *Aradus* (Heteroptera, Aradidae). Arthropod Struct. Dev. **39**, 17–25 (2010)
6. Evans, W.G.: Notes on the biology and dispersal of *Melanophila* (Coleoptera: Buprestidae). Pan-Pacific Entomol. **38**, 59–62 (1962)
7. Vondran, T., Apel, K.-H., Schmitz, H.: The infrared receptor of *Melanophila acuminata* DE GEER (Coleoptera: Buprestidae): ultrastructural study of a unique insect thermoreceptor and its possible descent from a hair mechanoreceptor. Tissue Cell **27**(6), 645–658 (1995)
8. Evans, W.G.: Infrared radiation sensors of *Melanophila acuminata* (Coleoptera: Buprestidae): a thermopneumatic model. Ann. Entomol. Soc. Am. **98**, 738–746 (2005)
9. Golay, M.J.E.: A pneumatic infra-red detector. Rev. Sci. Instrum. **5**, 357–362 (1947)
10. Chevrier, J.B., Baert, K., Slater, T.: An infrared pneumatic detector made by micromachining technology. J. Micromech. Microeng. **5**, 193–195 (1995)
11. Kenny, T.W.: Micromachined infrared sensors using tunneling displacement transducers. Rev. Sci. Instrum. **67**, 112–128 (1996)
12. Quartz Glass for Optics: Data and Properties. Data sheet Heraeus Quartzglas GmbH, Hanau (2011)
13. Weast, R.C.: Handbook of Chemistry and Physics, Chapter F-5f, 62nd edn. CRC-Press, Boca Raton (1981)
14. Szabo, I.: Höhere Technische Mechanik: Nach Vorlesungen. Springer, Berlin (2001)
15. Timoshenko, S.P., Woinowsky-Krieger, S.: Theory of Plates and Shells, 2nd edn. McGraw-Hill, New York (1959)
16. Spurk, J.H., Aksel, N.: Fluid Mechanics, 2nd edn. Springer, Berlin\Heidelberg\New York (2008)

I

17. Beadle, W.E., Tsai, J.C.C., Plummer, R.D.: Quick reference Manual for Silicon Integrated Circuit Technology. Stuttgart (1985)
18. Bousack, H., Schmitz, H., Offenhäusser, A.: Design of a fluidic infrared detector based on the infrared sensilla in the beetle *Melanophila acuminata*. In: Borgmann, H. (ed.) 11th International Conference on New Actuators, pp. 786–789. HVG Hanseatische Veranstaltungs-GmbH, Bremen, (2008)
19. Dettner, K., Peters, W. (eds.): Lehrbuch der Entomologie. Gustav Fischer-Verlag, Stuttgart (1999)
20. Schossig, M., Norkus, V.: Measurements at Institut für Festkörperelektronik. Technische Universität, Dresden (2010)

Insect-Inspired Visual Systems and Visually Guided Behavior

Paul Graham[1] and Andrew Philippides[2]
[1]Centre for Computational Neuroscience and Robotics, School of Life Sciences, University of Sussex, Brighton, UK
[2]Centre for Computational Neuroscience and Robotics, Department of Informatics, University of Sussex, Brighton, UK

Definition

Many insects are capable of impressive visually guided behaviors despite low-resolution vision and limited neural resources. These constraints have led to elegant neural and behavioral strategies which are of interest to engineers seeking to emulate insect-level performance with lightweight hardware. Such insect-inspired visual systems – in which engineers incorporate design principles derived from insect behavior – have consequently been adopted for many tasks, from the biomimetic design of visual hardware through to the development of parsimonious control algorithms for autonomous agent control.

Introduction

The insect world offers many examples of remarkable visually guided behavior. Two such behaviors which are no less remarkable for their familiarity are the precise, extremely fast flight control of flies and the long-distance visual navigation of bees and ants. These behaviors are all the more impressive because of the limitations that insects overcome in terms of visual resolution and available neural processing power. For the biomimetic engineer, the visual systems of insects therefore provide aspirational examples of parsimonious and economical strategies for visual tasks which can be evidently performed with minimal lightweight hardware.

Investigations of insect vision over many years and at various levels of study, from optical anatomy to behavioral observation, have shown how visually guided behaviors are the product of dedicated circuits which have evolved in specific ecological niches. This has led to the tight coupling of sensory systems, visual processing, and behavior. The basic organizational principles of insect visual systems are reviewed and at each stage it is considered how these principles might influence the design of artificial systems.

The Insect Visual Pathway

In insects the visual pathway begins with the compound eye, which differs fundamentally from simple eyes (i.e., single-lens eyes like human eyes) because there are multiple lenses with differing viewing directions but a fixed focal length (Fig. 1). There are two basic forms of compound eye [2], apposition and superposition. This entry focuses on the former as diurnal insects, the majority of which have apposition-type compound eyes, produce well-studied behaviors that have inspired engineers, such as precise and fast flight control, collision detection, and navigation.

In the apposition compound eye (Fig. 1), each lens has its own photoreceptors and this lens-photoreceptor unit is called an ommatidium ("little eye"). Compound eyes are severely constrained in terms of spatial resolution, because diffraction limits the resolution possible for a given aperture size. For instance, to replicate foveal human visual acuity, which is of the order

1 min of arc, a compound eye would have to have many more lenses and each lens would have to be much bigger, leading to a bulky sphere with a diameter of over 10 m. This limitation means that the highest visual acuity in insects is found in the large eyes of the dragonfly at about 0.5°. To mitigate this limitation, the density of ommatidia is not uniform, but is adapted to the environment, mode of locomotion, and behavior. Arthropods that move close to flat surfaces such as water or mud flats have greater acuity in the equatorial part of the visual field, which views the horizon where landmark information is to be found. Insects that fly in dense vegetation have a low density of ommatidia in the lateral parts of the eye to increase the detection of motion signals from objects moving from front to back across the eye's surface. Similarly, insects that need to resolve and track small targets, such as potential mates or rivals, have evolved specialized high-acuity eye regions adapted for encoding small and often fast-moving targets. Despite the limitations in acuity, compound eyes still offer advantages for insects and potentially engineers too. The hardware of the compound eye allows for a better temporal resolution (up to 300 Hz rather than 30 Hz for human vision) and because of the convex shape of a compound eye, insects can create a very wide field of view (FOV). These properties are particularly beneficial for processing motion information (which is discussed in section The Insect Visual Pathway).

The parallel output from the ommatidia is processed further in the lamina and medulla (Fig. 1). In these structures, information remains retinotopically organized. Although the functional organization of these structures is still not fully understood, it is generally taken that information is high-pass filtered in the lamina to emphasize dynamic components of the input, before subsequent extraction of local motion information in the medulla [3]. It is thought that motion information is extracted using correlational filters, where the signal from one point in space is correlated with time-delayed signals from adjacent points in space. These supposed filters are known as elementary motion detectors (EMDs).

One step further along the visual pathway are the lobula and lobula plate. Here there is massive convergence of the motion filtered signals from all parts of the visual field. These brain regions house a limited number of highly specialized neurons that are at the service of specific behaviors. For instance, single cells act as matched filters for particular patterns of optic flow generated in flight or for the motion of small objects against patterned backgrounds. The output of these cells often passes directly to flight muscles without further processing in the central brain. This highlights the economical, specialized nature of insects' visual circuits [1, 3].

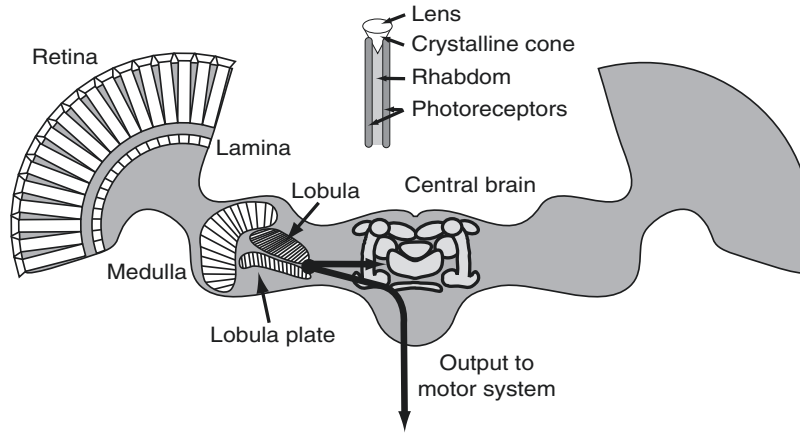
Artificial Compound Eyes

Artificial compound eyes are desirable because they promise a wide FOV without the bulk of single-lens imaging systems. It is true that a large FOV can be achieved with multi-camera systems, panoramic mirrors, or even mirror and lens hybrid arrangements that are cheap and easy to use. Indeed, these kinds of panoramic imaging systems are widely used tools for biologists using modeling and robotic approaches to study animal visual behavior [4]. They can also be mounted on medium to large wheeled [5] and flying [6] robots for autonomous navigation. However, future miniature platforms, such as micro unmanned aerial vehicles (micro UAVs) need to minimize payload and thus require much smaller visual systems and so approaches inspired by compound eyes have been explored. Moreover, artificial compound eyes of a comparable scale to insect eyes would have many applications in space-limited situations.

There have been some successes replicating the optical arrangements of compound eyes. One approach is to use arrays of microlenses to project onto CCD or CMOS cameras. The basic result is a sandwich of a microlens array, an aperture array, and a photodetector array. These flat artificial compound eyes, which can be constructed using MEMS (microelectromechanical systems) construction technology, have the advantage of being very small and thin, but do not necessarily

I





Insect-Inspired Visual Systems and Visually Guided Behavior, Fig. 1 (*Left)*. The compound eyes of a robber fly (*Holcocephala fusca*) (Photo: Magnus Manske, used with permission under a Creative Commons license). (*Right)* A schematic of the major regions in a typical insect visual pathway (Redrawn after [1]). (*Inset)* The major components of a single ommatidium

have a large FOV. 3D microlens arrays can improve the FOV to a maximum of 180° [7]. However, a practical issue remains with systems of this type, owing to the difficulty in aligning the different components of the system.

One of the more striking attempts to create an artificial compound eye, owing to its stunning similarity to a real insect eye, comes from Jeong and colleagues [8]. They avoid the problem of aligning the components of a visual system by creating a self-aligning system. To construct their eye, they make a mold of the outer surface of a compound eye – a hemisphere consisting of thousands of bumps in a hexagonal array. This compound eye mold is filled with an epoxy resin, and when struck by a beam of light, each bump in the epoxy structure acts as a lens and focuses light on the material below, hollowing out a narrow cavity. By this technique, waveguides (similar to the rhabdom in the insect ommatidium, Fig. 1) are constructed in perfect alignment with each lens. The end product is remarkably similar in size, shape, and optical performance to a honeybee ommatidium.

The aforementioned approaches to building artificial compound eyes mainly focused on taking advantage of the optical, size, and FOV advantages that come with this eye design. A second approach to building insect-inspired eyes comes from neuromorphic engineering where conventional optics are used in tandem with hardware (VLSI or FPGA) which mimics the early stages of visual processing, most commonly motion extraction. So-called "eyes on a chip" successfully achieve the low weight and fast parallel processing required in artificial systems [1, 9]. Floreano et al. [10] have taken this approach and, by using flexible printed circuit

boards, have produced a convex artificial eye (named CurvACE) that has a hemispheric FOV. The CurvACE sensor has been shown to be well suited to the wide-field extraction of motion information which is important for many insect behaviors as discussed in the next section.

Using Motion Information

Fast, precise insect behaviors, such as the ability of flies to avoid being swatted, of hoverflies to accurately maintain their position, and of locusts to avoid predator attacks, come from the capability of insect brains to rapidly extract global patterns of local motion in behaviorally relevant ways. The pattern of visual motion experienced by an animal depends on three factors: the animal's motion, the structure of the environment, and the movement of objects in the world. This means that insects can use motion signals to extract information about their own movements as well as the structure of the environment. In this section, some basic uses of motion information are considered.

Flight Control

For flying insects, controlling flight is complicated by air currents; thus, an insect's actual movements cannot be directly inferred from its motor output. To monitor the success of an intended movement or to check for unintended displacement, insects use the pattern and amplitude of the visual motion they experience. Behavioral experiments with bees and neurophysiological experiments with flies have demonstrated the ingenious tricks employed by insects when using visual motion cues to control flight behavior.

Mandyam Srinivasan and colleagues [6] have spent many years observing bees flying through patterned tunnels. These experiments show that bees maintain a central position in a tunnel by balancing the image motion, or optic flow, perceived by each eye. The optic flow, F, generated by a simple object at a distance d from an observer moving with an instantaneous velocity v can be calculated as:

$$F = -\omega + \frac{v}{d}\sin\theta$$

where ω is the optic flow generated by self-rotation and θ is the angle of the object relative to the direction of travel. In the absence of self-rotation (often assumed to be subtracted out through measurement by gyroscopic sensors) balancing the optic flow on each side of the bee (i.e., at +/− θ) would mean that d would be the same on both sides (as v is constant for both eyes). Similarly, maintaining constant optic flow requires that if the tunnel narrows, the bee must lower its velocity. This means that in more cluttered environments, the bee automatically responds appropriately by slowing down. This simple and elegant solution of balancing optic flow thus results in a bee being able to chart a safe path through a cluttered environment *without* knowing its own speed or measuring the distance to nearby objects.

A similar control scheme is used by bees to control grazing landings on flat surfaces. When landing, bees control their forward flight speed using the angular velocity perceived in the lower part of their visual field. As the bee descends, the ground appears to move at a higher angular velocity. By controlling forward thrust to maintain a fixed angular velocity, the bee "automatically" slows as its altitude falls. When the bee finally touches down, its forward speed is zero, thus ensuring a smooth landing. This algorithm also allows the bee to maintain a constant angle of descent throughout the landing and does not require explicit measurement of flight speed, flight altitude, or estimated time to contact.

Flight Control for UAVs

The fixed-point control mechanisms observed in bee flight are clearly attractive to engineers, and following Srinivasan's work, there have been multiple instantiations of corridor-centering robots. In addition, the observation that bees use

this type of control to maintain a fixed height above a surface has led to the robust control of UAVs capable of "terrain following". Notice that in the above equation, if an insect is flying parallel to the ground, the optic flow signal from the area directly beneath it will be strong and the distance, *d*, is the height of the insect above the ground. A UAV can therefore maintain a constant height over undulating terrain by altering its forward velocity so that the optic flow measured from a downward-facing camera remains constant. This type of control has been implemented on free-flying helicopters and fixed-wing aircraft [6, 11].

Crucial to the success of these systems is the extraction of reliable optic flow signals. In general, these rely on simple hardware EMDs. However, going a stage further down the visual pathway and modeling the wide-field integration of the cells in the lobula plate has led to alternative control systems for UAVs. As described above, these cells take input from a large number of EMDs in the medulla, and electrophysiological studies indicate that they respond preferentially to particular global patterns of optic flow, such as an expanding flow field caused by pure forward motion. Individual cells therefore act as matched filters for translation and rotation about different axes providing signals that insects can use to monitor movement and correct for unintended displacements or rotations [12].

Engineers have used an analog of the lobula plate system to generate fixed-point control algorithms for both flying and wheeled robots [13]. The output of a given neuron in the lobula plate can be interpreted as a comparison between the current instantaneous flow field and the flow field that would be generated by the particular movement that the neuron is configured to detect. This can be modeled as an inner product between a template flow field pattern and the instantaneous flow field. With this technique it is possible to extract information such as forward and vertical velocity and pitch rate. Further, spatial Fourier components of appropriate filter outputs can be fed back into the system and used in fixed-point control algorithms for UAV stabilization.

Collision Avoidance

The depth cues available to insects are an impoverished subset of those that vertebrates can use. Two main cues are available: motion parallax and looming. Insects do not use these cues to generate a universal depth map. Rather, individual cues are used in specific visual circuits to control specific behaviors. For avoiding obstacles in free flight, many insects use parallax cues and have evolved a specific movement strategy to simplify the perception of parallax information. During translatory movements, objects closer to the animal move faster across the retina, but this signal can be swamped by the rotatory flow fields (where objects at all distances move across the retina with the same speed) produced by rotation. Many insects therefore move in a saccadic manner where periods of straight flight are alternated with very fast turns. In addition, small head movements are used to compensate for body rotation, such that the eyes are rotating for the least amount of time. The utility of this strategy was highlighted in a seminal piece of insect vision-inspired biorobotics [14]. Franceschini and colleagues implemented an array of electronic EMDs on the perimeter of a large wheeled robot. The EMDs were tuned to respond to the parallax generated by objects that were on a collision course. This was possible as the robot used a fixed movement direction relative to its sensors and used a saccadic movement strategy, i.e., pure translation alternated with fast pure rotation. Given these two constraints the robot was able to move rapidly around a flat environment while avoiding obstacles.

An alternative distance cue comes from looming, also known as image expansion. The mechanisms of looming detection have been well studied in the locust collision-avoidance response which is triggered by image expansion [15]. In locusts, a specialist neuron in the lobula called the lobula giant movement detector (LGMD) responds to any movement within a large visual area, but is most sensitive to objects directly approaching the eye. The firing rate of this neuron increases throughout an object's approach, but responses decrease if the approaching target

deviates from a collision course by a small amount. The response properties of the LGMD indicate that the cell is sensitive to the movement and total length of luminance boundaries in the image. All these parameters can be derived by conflating the output of EMDs in the medulla and therefore the LGMD neuron acts as a matched filter for impending collision using only a few processing stages.

Because the LGMD is able to reliably signal impending collisions without the need to explicitly represent the target object's physical size, distance, or approach speed, it represents an interesting model for collision avoidance in artificial systems. For instance, artificial neural networks have been constructed to account for the responses of the locust LGMD neurons and detect impending collisions on wheeled robots. Essentially, the edges of an expanding object trigger feed-forward excitation, followed a short time later by laterally spreading inhibition, with both excitatory and inhibitory signals converging on the LGMD. So long as the edges continue expanding rapidly enough to "outrun" the pursuing inhibitory signals, the LGMD's response continues to grow in strength. This model has been further developed and tested for collision warning systems in automobiles [16]. However, in order to produce satisfactory performance in an automobile context, the model had to be significantly augmented. This highlights a particular problem with biomimetics. In this example, the sensory environment and behavioral niche for locusts and cars are very different. In general, it should not be assumed that bioinspired systems will work perfectly in novel contexts.

Economical Visual Navigation

Efficient and robust navigation using visual cues is an important capability for many animals as well as autonomous robots. By using simple strategies, insects demonstrate excellent navigational ability albeit with small brains and low-resolution vision. The most famous example of this is view-based homing or snapshot guidance [17] in which insects navigate back to a location by remembering a retinotopic view of the world, or snapshot, stored at the goal position. Subsequent navigation is achieved by moving so as to make the current view of the world more similar to the remembered goal view. In general, algorithmic instantiations of this process work by iteratively comparing a parameterized view of the world from the agent's current position with a similarly parameterized view stored when it was at the goal location. Differences between the current and goal views are then used to derive an approximate direction to the goal. For engineers, one attraction of insect-style view-based homing is that these strategies can work without identifying objects in the world. That is, they can work using only the appearance of the world, rather than needing to identify and subsequently match features between views, vastly simplifying the visual task [18].

One of the simplest and most elegant implementations of snapshot homing for robotic navigation is the average landmark vector (ALV) model [5]. The ALV model works by first identifying visual features from a 360° panoramic view of the environment. The view of the world from that position is then represented as the average of unit vectors pointing from the agent to each identified feature, i.e., the ALV. The algorithm works since the vector difference between the current ALV and an ALV stored at the goal position points approximately back to the goal. While this approach does extract features, it does not attempt to match them and thus the feature extraction can be very coarse and has been shown to work using one feature per pixel (essentially, without feature extraction). This idea of encoding a visual scene as the visual center of mass of image features, or even the center of mass of pixel intensities, is surprisingly robust and is a general purpose method for retinotopic encoding of a visual scene.

While the ALV is a particularly parsimonious visual homing algorithm, at the heart of all algorithms of this type is the fact that, within a region local to the goal, the difference between the goal and surrounding views increases smoothly with distance from the goal until it starts to plateau [19]. The size of the region before the differences plateaus, called the catchment area of an image, is determined by the distribution of object distances.

I

To home successfully, the agent thus needs only to calculate the local gradient of the difference between current and goal images (whether processed or unprocessed) and move in the direction which decreases this difference. While the local gradient can be calculated by sampling the world, it has been shown that the gradient can also be estimated using matched filters which approximate the effects of forward and sideway movements on the current image by effectively performing the inverse of elementary optic flow calculations. This method has been shown to work robustly in natural environments and can be implemented with simple parallel processing steps and so should be amenable to hardware implementation.

As well as using visual memory to search for a discrete goal, insects can also use vision to guide long routes [20]. Unlike search for a discrete goal where an agent needs to be attracted to the goal from any location around it, when following a route an agent instead needs to set a direction that is roughly parallel to the overall direction of the route section at their current location. Because compound eyes cannot be moved relative to an insect's head, an insect's direction of movement specifies its viewing direction, and so if an insect makes similar movements when at similar locations, it will experience similar views. Thus, in practical terms, for an agent to recapitulate a route, it must simply move in the direction that appears visually most familiar to the views it experienced when it previously navigated that route [21]. This simple heuristic can be underpinned by parsimonious computational architectures [22] allied to stereotypical behaviors [23]. For instance, Baddeley et al. [24] show how during route learning the views experienced along a route can be used to update the weights in a neural network using the Infomax algorithm. After training the network can be used to gauge the familiarity of any scene presented to it and thus can be used to underpin familiarity-based route navigation.

While there are many insect-inspired navigation algorithms, the key similarity is that the agent never calculates where it is at any point but rather what it should do. This class of algorithm thus highlights how navigation can be performed without explicitly calculating position, setting it apart from map-based approaches such as simultaneous localization and mapping or SLAM [25] and allowing navigation to proceed with low computation.

Conclusion

Insects are an existence proof that low-resolution vision and a small brain are not a barrier to the production of exquisite visually guided behavior. What's more, insects provide specific design blueprints for emulating this performance in artificial systems [1]. In summary, the early stages of insect vision are massively parallel and fast, and much visual behavior is controlled by neural matched filters that integrate across many of these parallel inputs. However, a significant lesson for engineers is that insects' visual systems have been fine-tuned by evolution such that the interactions between optics, neurons, behavioral routines, and the environment are optimized for a specific ecological niche. The niche of an artificial system has to be considered when applying biomimetic design principles.

Cross-References

▶ Biomimetics
▶ Insect Flight and Micro Air Vehicles (MAVs)

References

1. Webb, B., Harrison, R.R., Willis, M.A.: Sensorimotor control of navigation in arthropod and artificial systems. Arthropod. Struct. Dev. **33**, 301–329 (2004)
2. Land, M.F., Nilsson, D.-E.: Animal Eyes. Oxford University Press, Oxford (2002)
3. Sanes, J.R., Zipursky, S.L.: Design principles of insect and vertebrate visual systems. Neuron **66**, 15–36 (2010)
4. Webb, B.: Using robots to understand animal behavior. Adv. Study Behav. **38**, 1–58 (2008)
5. Lambrinos, D., Ralf Möller, R., Labhart, T., Pfeifer, R., Wehner, R.: A mobile robot employing insect strategies for navigation. Robot. Auton. Syst. **30**, 39–64 (2000)
6. Srinivasan, M.V., Thurrowgood, S., Soccol, D.: From visual guidance in flying insects to autonomous aerial

vehicles. In: Floreano, D., Zufferey, J.C., Srinivasan, M.V., Ellington, C. (eds.) Flying Insects and Robots. Springer, Berlin (2010)
7. Song, Y.M., Xie, Y., Malyarchuk, V., Xiao, J., Jung, I., Choi, K.-J., Liu, Z., Park, H., Lu, C., Kim, R.-H., Li, R., Crozier, K.B., Huang, Y., Rogers, J.A.: Digital cameras with designs inspired by the arthropod eye. Nature **497**(7447), 95–99 (2013)
8. Jeong, K.-H., Kim, J., Lee, L.P.: Biologically inspired artificial compound eyes. Science **312**, 557–561 (2006)
9. Floreano, D., Zufferey, J.C., Srinivasan, M.V., Ellington, C. (eds.): Flying Insects and Robots. Springer, Berlin (2010)
10. Floreano, D., Pericet-Camara, R., Viollet, S., Ruffier, F., Brückner, A., Leitel, R., Buss, W., Menouni, M., Expert, F., Juston, R., Dobrzynski, M.K., L'Eplattenier, G., Recktenwald, F., Mallot, H.A., Franceschini, N.: Miniature curved artificial compound eyes. Proc. Natl. Acad. Sci. U. S. A. **110**(23), 9267–9272 (2013)
11. Franceschini, N., Ruffier, F., Serres, J.: A bio-inspired flying robot sheds light on insect piloting abilities. Curr. Biol. **17**, 329–335 (2007)
12. Krapp, H.G.: Neuronal matched filters for optic flow processing in flying insects. Int. Rev. Neurobiol. **44**, 93–120 (2000)
13. Humbert, J.S., Conroy, J.K., Neely, C., Barrows, G.: Wide-field integration methods for visuomotor control. In: Floreano, D., Zufferey, J.C., Srinivasan, M. V., Ellington, C. (eds.) Flying Insects and Robots. Springer, Berlin (2010)
14. Franceschini, N., Pichon, J.M., Blanes, C.: From insect vision to robot vision. Philos. Trans. R. Soc. B. **337**, 283–294 (1992)
15. Rind, F.C., Simmons, P.J.: Seeing what is coming: building collision-sensitive neurones. Trends. Neurosci. **22**, 215–220 (1999)
16. Stafford, R., Santer, R., Rind, F.C.: The role of behavioural ecology in the design of bio-inspired technology. Anim. Behav. **74**, 1813–1819 (2007)
17. Cartwright, B.A., Collett, T.S.: Landmark learning in bees. J. Comp. Physiol. A **151**, 521–543 (1983)
18. Horridge, G.A.: What does an insect see? J. Exp. Biol. **212**, 2721–2729 (2009)
19. Zeil, J., Hofmann, M. I., Chahl, J.S.: Catchment areas of panoramic snapshots in outdoor scenes. J. Opt. Soc. Am. A **20**, 450–469 (2003)
20. Graham, P.: Insect navigation. In: Breed, M.D., Moore, J. (eds.) Encyclopedia of Animal Behavior, pp. 167–175. Academic Press, Oxford (2010)
21. Philippides, A., Baddeley, B., Cheng, K., Graham, P.: How might ants use panoramic views for route navigation? J. Exp. Biol. **214**(3), 445–451 (2011)
22. Philippides, A., Graham, P., Baddeley, B., Husbands, P.: Using neural networks to understand the information that guides behavior: a case study in visual navigation. In: Artificial Neural Networks, pp. 227–244. Springer, New York (2015)
23. Wystrach, A., Philippides, A., Aurejac, A., Cheng, K., Graham, P.: Visual scanning behaviours and their role in the navigation of the Australian desert ant *Melophorus bagoti*. J. Comp. Physiol. A **200**(7), 1–12 (2014)
24. Baddeley, B., Graham, P., Husbands, P., Philippides, A.: A model of ant route navigation driven by scene familiarity. PLoS Comput. Biol. **8**(1), e1002336 (2012)
25. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part I. Robot. Autom. Mag. IEEE **13**(2), 99–110 (2006)

Insects

► Arthropod Strain Sensors

Integrated Approach for the Rational Design of Nanoparticles

Paolo Decuzzi
Department of Translational Imaging, and Nanomedicine, The Methodist Hospital Research Institute, Houston, TX, USA

Synonyms

Optimization of nanoparticles

Definition

The integrated approach for the rational design (IARD) of nanoparticles is a framework combining in silico mathematical modeling, in vitro and in vivo experiments with the objective of identifying an optimal nanoparticle that can efficiently localize within the diseased tissue (tumor, atherosclerotic plaque, injured vessel, and so on) while avoiding nonspecific sequestration by the organs of the reticuloendothelial system (RES). The size, shape, surface properties, and mechanical stiffness (*4S design parameters*) are the independent design parameters to be optimized that affect the vascular, extravascular, and subcellular behavior of systemically injected nanoparticles.

Introduction

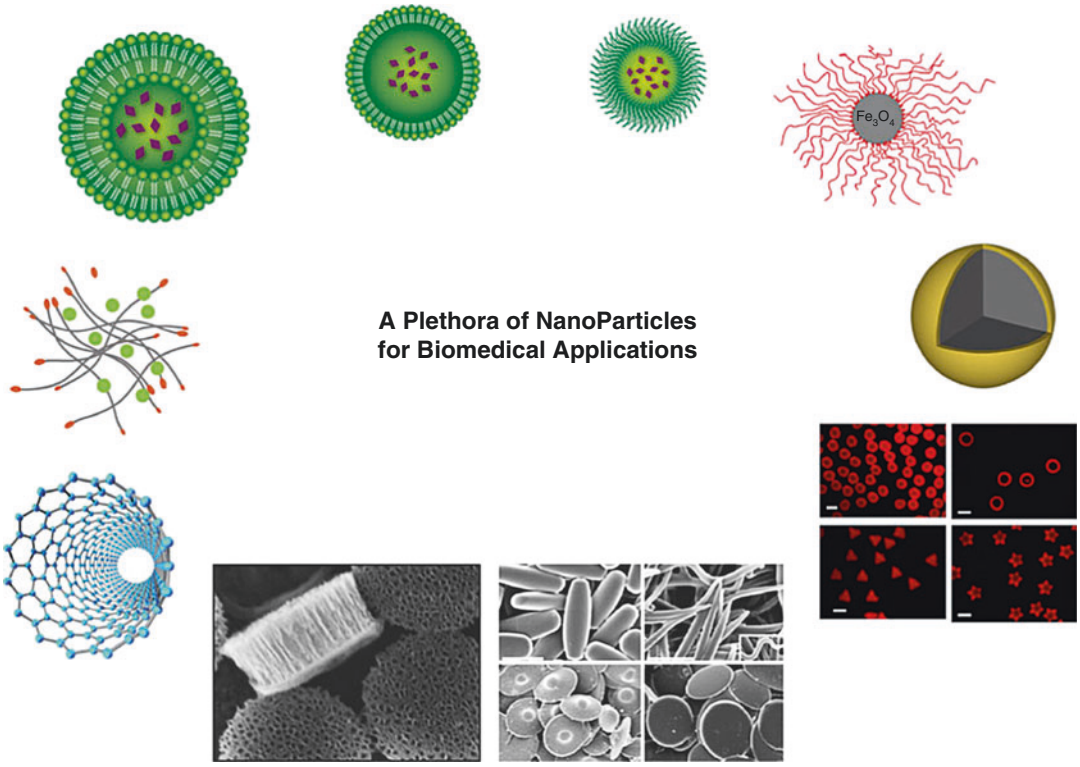
Nanoparticles (NPs) are man-made objects small enough to be injected intravascularly and navigate safely within the circulatory system and are designed to perform several crucial functions [1, 2]. They can carry multiple drug molecules and contrast agents from the site of injection to the biological target (diseased tissue) thence providing their therapeutic action and contrast enhancement. Many classes of NPs have been developed for diverse applications, including the treatment and imaging of cancer and other diseases. The advantage of using NPs over freely administered molecules is threefold: (1) *higher specificity*, increased active molecules reaching the target; (2) *multifunctionality*, loading and codelivery of several and different drug molecules and contrast agents; and (3) *engineerability*, multiple design parameters that can be properly tailored during the NP synthesis process.

Geometry, surface properties, and composition often change among the different classes of NPs. This is shown in Fig. 1 where a few representative NPs, developed over the last 20 years, are exhibited. In the top left corner, the archetypical NP – a liposome – is depicted with a spherical shape, a diameter of about 100 nm and is composed of a lipid double-layer. Proceeding in a clockwise direction, a micelle is shown which is still spherical, with a diameter of few tens of nanometers, and is composed of a lipid single-layer. Figure 1c represents a spherical polymeric NP. These NPs usually have a size varying in a wide range, from tens to a few hundreds of nanometers, and can be synthesized using several different types of polymers. Continuing, two examples of metallic NPs are presented: a superparamagnetic iron oxide (SPIO) particle (Fig. 1d), whose magnetic core is generally spherical with a characteristic size of 5–20 nm, and a gold nanoshell (Fig. 1e), comprising an inner spherical silica core (110–130 nm) coated with a thin (5–10 nm) gold film. So far, only spherical NPs have been featured which are generally obtained by the precipitation and self-assembly of salts, lipids, and polymeric chains. More recently, methods have been proposed for the synthesis of nonspherical NPs made out of different materials. These include the use of hydrogel-based templates (Fig. 1f), PRINT technology, thermoforming, and mechanical molding (Fig. 1g) for the synthesis of polymeric nano- and microparticles, offering a variety of shapes, sizes, and compositions. Nonspherical particles have been also fabricated out of silicon (Fig. 1h), combining lithographic techniques and electrochemical etching; and using gold and carbon-based materials, leading to nanorods and nanotubes, respectively (Fig. 1i). Finally Fig. 1j presents an interesting example of a hybrid NP obtained by combining bacteriophages (organic component) with gold nanospheres (nonorganic component). Clearly, the variety of size, shape, and composition is enormous and, in addition, the surface of most NPs can be decorated with multiple and diverse functional groups and polymeric chains and molecules, making the overall picture even more complex.

Size, shape, surface properties, and mechanical stiffness (*the 4S design parameters*) have been shown to affect the vascular, intravascular, and subcellular transport of NPs and their functions [3, 4]. It is imperative for the development of NPs to accurately control the 4S design parameters during manufacturing and to identify the optimal combination (or set of combinations) that would allow the systemically injected NPs to efficiently and preferentially reach the biological target. This indeed is the objective of the integrated approach for the rational design (IARD) of nanoparticles, a framework where in silico mathematical models, in vitro and in vivo experiments are used together to solve an optimization problem (Fig. 2): identify the 4S design parameters that maximize NP lodging within the diseased tissue (*efficient accumulation*) and minimize nonspecific sequestration by the organs of the reticuloendothelial system (RES) (*preferential accumulation*).

In Silico Mathematical Modeling

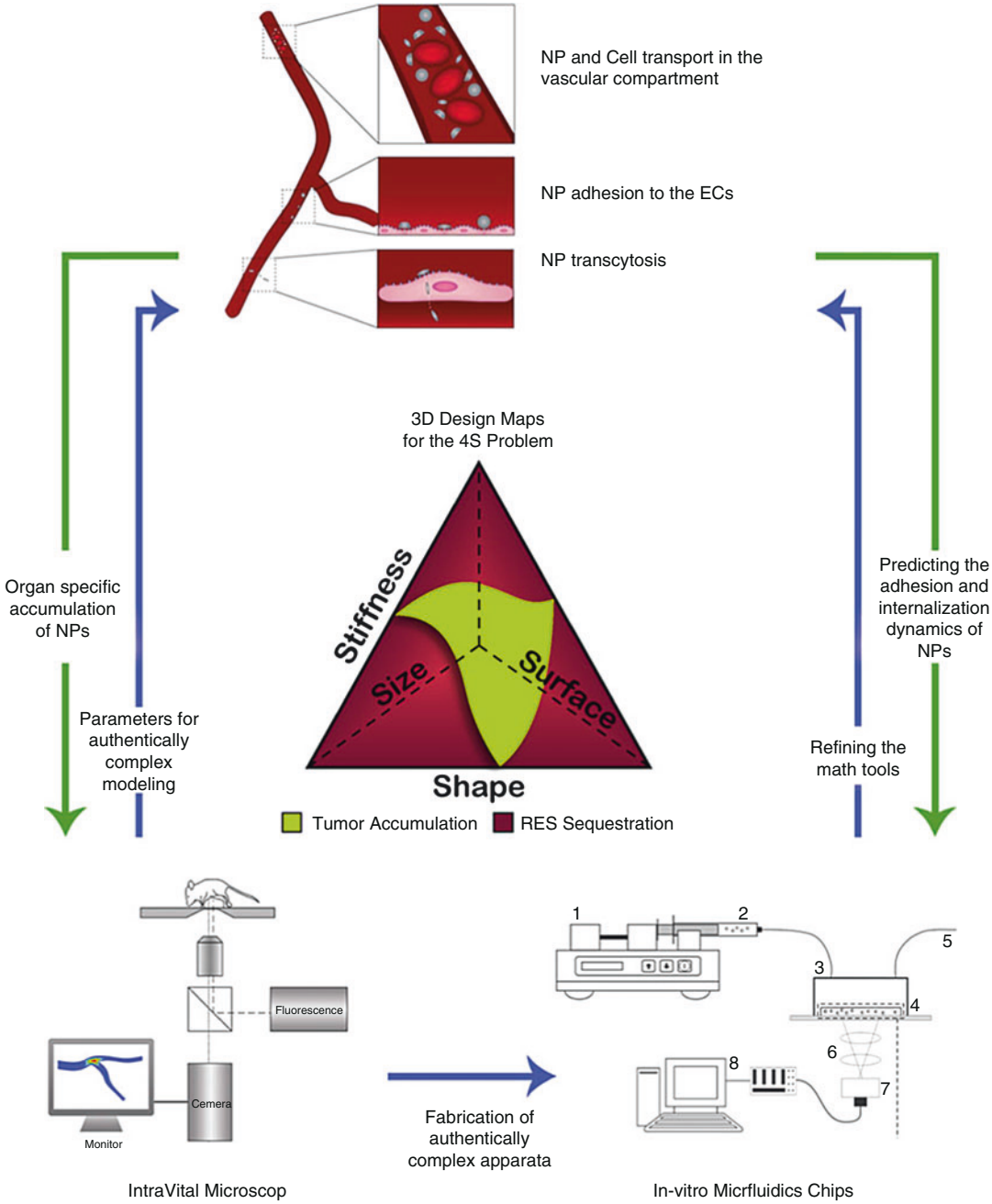
An intravascularly injected NP navigates the circulatory system transported by the blood flow. During the journey, NPs interact with red blood



Integrated Approach for the Rational Design of Nanoparticles, Fig. 1 A plethora of nanoparticles for biomedical applications. Over the last 20 years, nanoparticles with different size, shape, surface properties, mechanical stiffness, and composition have been developed

cells (RBCs), the most abundant cells in the circulation; with cells of the immune system, both circulating monocytes and resident macrophages, such as the Kupffer cells in the liver and the splenic macrophages; and with endothelial cells (ECs), lining the vessel walls. Different computational tools can be employed to study the vascular transport and cell/NP interaction, depending on the complexity of the problem and desired accuracy of the solution.

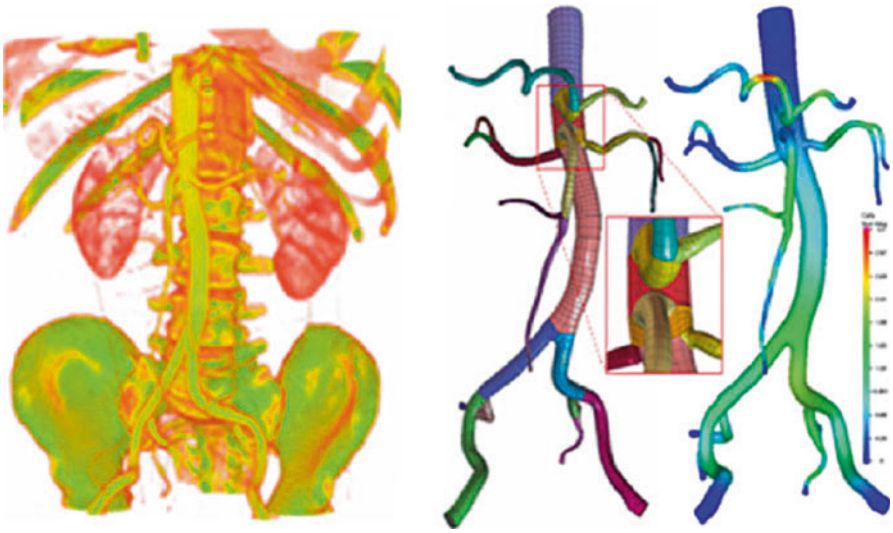
In the macrocirculation (vessel diameter >300 μm), NPs can be considered as passive scalars advecting and diffusing in a quasi-Newtonian fluid. Under these conditions, a classical computational scheme, based on finite element methods (FEMs) and finite volume methods (FVMs), can be used quite efficiently to solve the transport problem. Recently, the isogeometric analysis (IA) has been introduced as a generalization of FEM which allows one to improve the functional representation of the geometry in the computational problem, simplifying the mesh refinement and increasing the accuracy of the solution [5]. In particular, patient-specific vascular models can be accurately constructed by identifying a set of procedures to create solid NURBS-based vascular models directly from patient-specific MRI/CT data (Fig. 3) [6]. This requires four main steps: preprocessing, where the quality of the original patient-specific data is improved using imaging processing techniques; path extraction, where the skeleton of the arterial system is identified using Voronoi and Delaunay diagrams; control mesh construction, where hexahedral NURBS control meshes are generated by sweeping a templated quad mesh of a circle along the arterial path; solid NURBS construction, which allows one to generate a 3D exact representation of the patient-specific



Integrated Approach for the Rational Design of Nanoparticles, Fig. 2 A schematic representation of the integrated approach for the rational design of nanoparticles. Mathematical tools, in vitro experiments, and intravital microscopy are used synergistically to solve the 4S problem, presented in the quaternary diagram in the *middle*

data, and, eventually, the reconstructed 3D geometry is used for the final numerical analysis.

The IA framework has been applied successfully to solve several nonlinear and complex physical problems. It can be effectively used to model transport problems in the macrocirculation. A 3D Navier–Stokes solver for blood flow is coupled with a linear scalar advection/diffusion

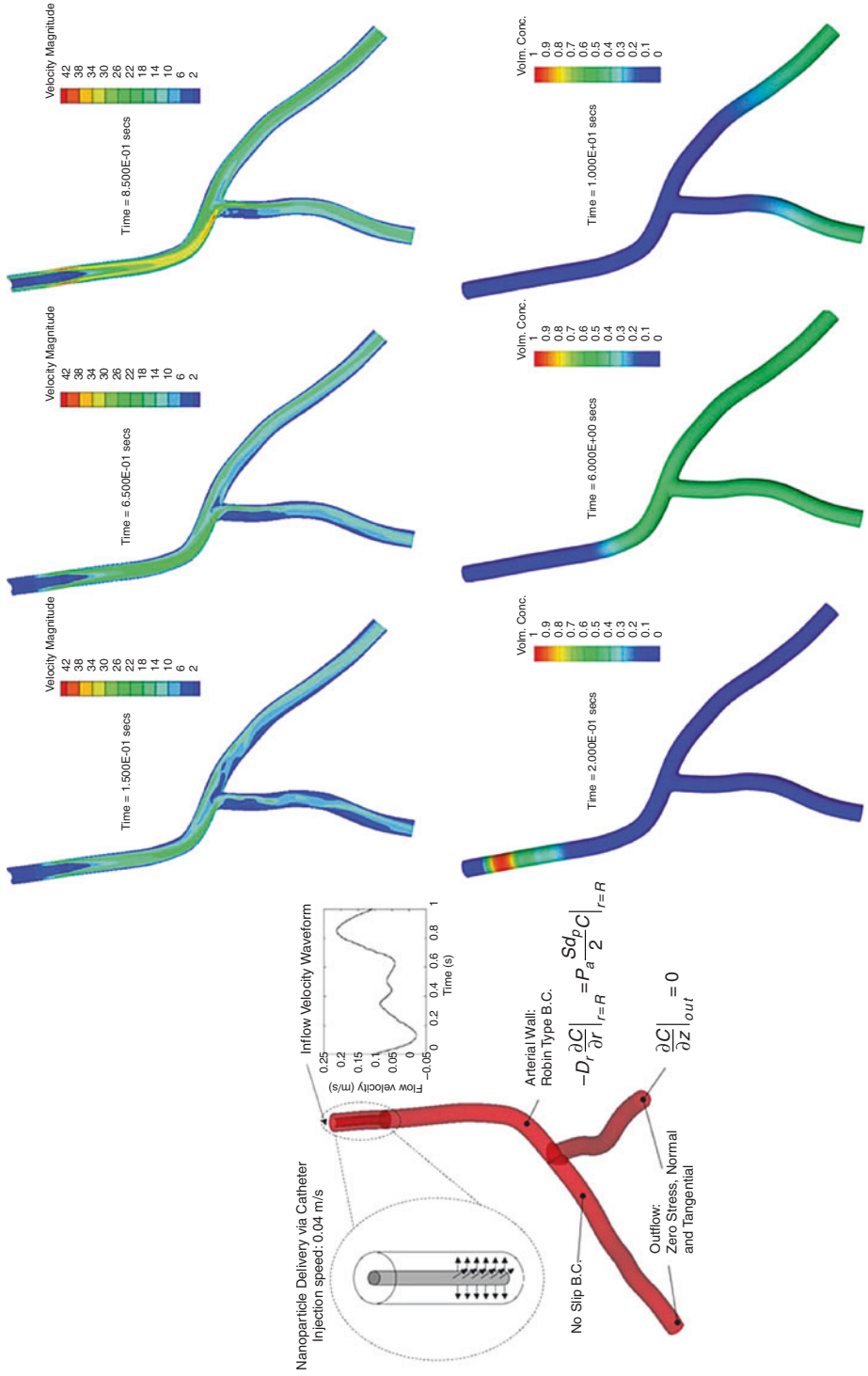


Integrated Approach for the Rational Design of Nanoparticles, Fig. 3 Patient-specific modeling of the abdominal aorta. The original image is divided into 26 patches, and each color represents one different patch. The image to the *right* presents data on the flow field within the vessels (Adapted from [6])

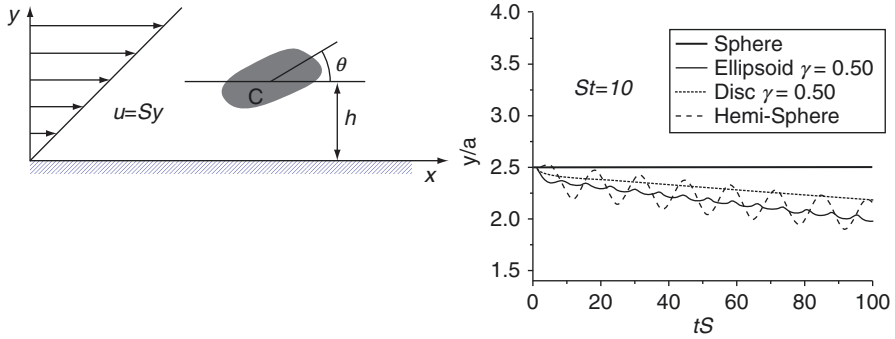
equation for studying the blood flow and transport of passive scalars within the authentically complex arterial vasculature. The spatial and temporal variations of the wall shear rate (WSR), wall shear stress (WSS), velocity profile, pressure field, and volumetric concentration of drug molecules and NPs can be estimated in patient-specific geometries. A classical problem solved within the IA framework is exemplified in Fig. 4, where a catheter is placed at the inlet of a coronary artery and releases drug molecules in the flow. The 3D Navier–Stokes solver allows one to quantify the magnitude of the velocity field within the vasculature in the presence of the catheter (top, right) at different time points. Using the derived flow field, the transport equation is solved for time and spatial evolution of the drug molecules released from the catheter (bottom, right). The IA can be implemented with an efficient parallel-solution strategy.

In the microcirculation (vessel diameter $<$300 μm), NPs interaction with RBCs and with the ECs has to be also taken into account. Different techniques have been proposed including semianalytical models [7, 8], the immersed boundary method (IBM) [9], the immersed finite element method (IFEM) [10, 11], dissipative particle dynamics (DPD) [12], and Lattice Boltzman (LB) [13] approaches. In the following, the focus will be on semianalytical methods and IFEM.

RBCs tend to preferentially accumulate in the core of the vessels leaving a "cell-free layer" in proximity of the wall where the flow is mostly laminar and linear with a shear rate S. Such a layer has a thickness varying with the vessel lumen diameter and mean blood velocity, ranging from few microns in small capillaries (10 μm and more in diameter) to few tens of microns in arterioles (100 μm and more in diameter). The dynamics of NPs moving in the cell-free layer can be effectively described using semianalytical methods. The geometry of the problem is presented in Fig. 5 (left), where an arbitrary shaped particle is immersed within a linear laminar flow in a semi-infinite 3D region bounded by an infinite flat plane at $y = 0$, reproducing the blood vessel wall [8]. In this specific example, a sphere, an ellipsoid, a discoid, and a hemisphere are considered. For the fluid motion, mass and momentum conservation are imposed, whereas the NP is required to obey Newton's law for linear and rotational motion. For the case presented in Fig. 5, it is clearly seen that nonspherical particles would drift laterally across



Integrated Approach for the Rational Design of Nanoparticles, Fig. 4 Patient-specific modeling for the transport of drug molecules in a coronary artery. The fluid velocity field (*right*, *top*) and the volumetric concentration of drug molecules (*bottom*, *right*) are shown at different time points. A typical inflow velocity profile of a beating heart is used (*left*, *inset*)

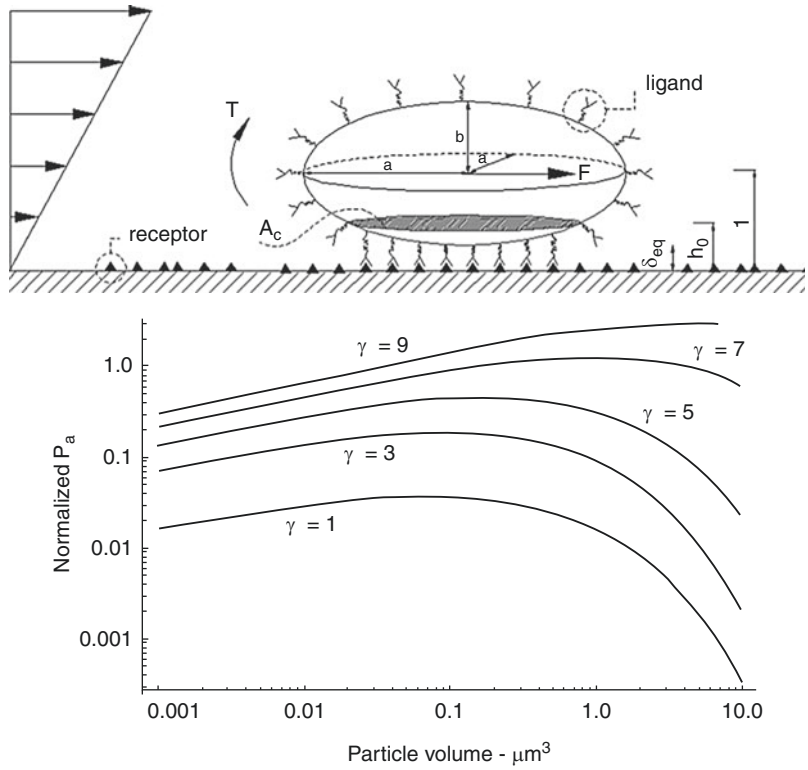


Integrated Approach for the Rational Design of Nanoparticles, Fig. 5 Shape-specific NP dynamics in proximity of a vessel wall. Schematic of the problem (*left*) and trajectory of NPs with different shapes (*right*). Nonspherical NPs are predicted to drift laterally across the streamlines toward the vessel wall (y = 0) (Adapted from [7])

the streamlines. Such a lateral drifting has been named *margination dynamics*, and it is affected by the NP size, shape, and composition and by the local hydrodynamic conditions. Also, it has been demonstrated that the smaller the aspect ratio, the larger is the drifting velocity having fixed other parameters.

Clearly, a marginating NP has a larger likelihood of interacting with the vessel walls to recognize the diseased tissue. However, this does not directly imply adhesion. The propensity for firm adhesion to the vessel walls under flow can be related to a specific mathematical quantity: the probability of adhesion P_a [14]. This is affected by biophysical parameters, such as the local hydrodynamic conditions (wall shear rate S) and the density of receptor molecules (m_r) expressed on the cell membrane, and by the size, shape, surface properties, and mechanical stiffness of the particle. The NP surface properties include the density of ligand molecules (m_l) distributed over the particle surface, the electrostatic surface charge and decoration with polymeric chains which would affect nonspecific interactions at the cell/particle interface (van der Waals, steric, double-layer electrostatic), and the ligand-receptor affinity k_a. Mechanical stiffness is relevant too because it affects the distribution of the hydrodynamic forces over the multiple interfacial molecular bonds. Using in silico models, it has been demonstrated that oblate spheroidal particles exhibit a much higher strength of adhesion (P_a) compared to spherical particles under the same biophysical conditions and for the same volume V (i.e., payload). The variation of P_a with V, for different values of the particle aspect ratio γ, is shown in Fig. 6: the P_a of a spheroid with $\gamma > 2$ is orders of magnitude larger than for a sphere ($\gamma = 1$), for a fixed V [14]. As γ increases, the oblate spheroid tends in the limit to a thin disk. The large adhesion observed is explained by the larger surface area exposed (i.e., larger number of molecular bonds) to cells and lower hydrodynamic dislodging forces of a "squashed" spheroid compare with a sphere of equivalent volume.

In order to analyze the interaction of NPs with RBCs, a more sophisticated computational approach is required, such as the immersed finite element method (IFEM). This technique was developed by integrating the concept of immersed boundary, finite element, and mesh-free methods. Following the IFEM framework, the immersed bodies (i.e., NPs and circulating cells) are modeled as flexible and highly deformable structures using a Lagrangian mesh, whereas the fluid (plasma) is solved in a fixed Eulerian mesh. The cell/cell and cell/NP interactions can be modeled using a Morse potential, and Kramer's reaction rate is employed for describing the ligand-receptor binding at the interface. Within the IFEM framework, the dynamics of immersed bodies is coupled with the surrounding incompressible viscous fluid



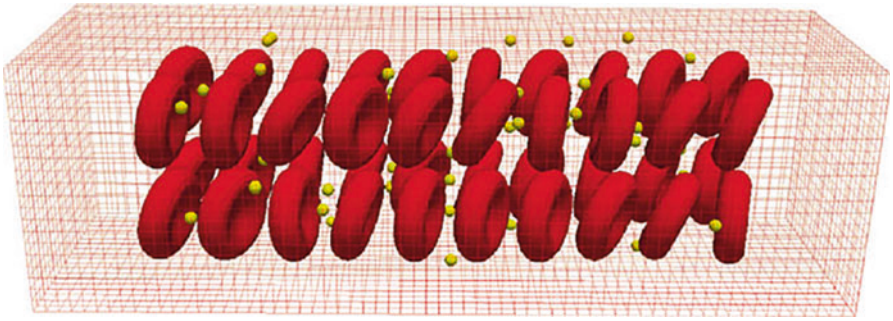
Integrated Approach for the Rational Design of Nanoparticles, Fig. 6 Shape-specific NP adhesion under flow. Schematic of the problem (*top*) and normalized probability of adhesion Pa as a function of the particle volume V and for different values of the particle aspect ratio γ (*bottom*). For a fixed volume, non spherical particles are predicted to adhere more avidly to the vessel wall under flow (Adapted from [14])

but avoiding expensive remeshing or mesh adaptivity of the fluid domain. With this approach, the dynamics of NPs can be studied in whole blood, meaning in the presence of a large group of deformable RBCs under different flow conditions. An example is presented in Fig. 7, where NPs are dispersed within a train of RBCs moving in a capillary. The relative position of the RBCs and NPs, and their orientation can be estimated at different time points as blood flows from left to right.

Finally, the cellular uptake of NPs is again affected by their size, shape, surface physicochemical properties, and mechanical stiffness. The size effect has been clearly shown for spherical particles, where the radius not only affects the uptake rate, in that the larger is the radius and the longer is the time for cell uptake, but it also affects the internalization mechanisms [15]. The characteristic time for NPs to be wrapped by a cell membrane can be related to their size, shape, and surface physicochemical properties. A mathematical model for receptor-mediated internalization based on an energetic analysis shows that a threshold radius exists below which particle uptake does not occur because it is energetically unfavorable. The same analysis shows that the surface physicochemical properties of the NP related to that of the cell membrane can dramatically increase or decrease the uptake rate [16]. A mathematical model has been developed to predict the rate of uptake for ellipsoidal particles as a function of their aspect ratio. Spherical or ovoidal particles can be more rapidly internalized by cells compared to elongated particles [17]. These results agree qualitatively well with experimental data [18].

Microfluidic Chips and Intravital Microscopy

Microfluidic chips (MFCs) and intravital microscopy (IVM) analysis are used to validate and refine the mathematical models. MFCs can be effectively used in conjunction with semianalytical models and IFEM for studying the NP behavior

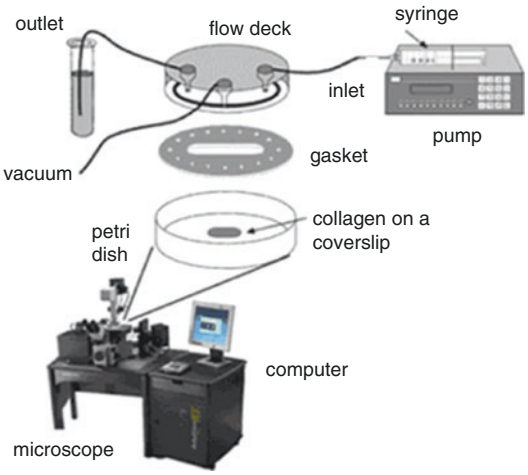


Integrated Approach for the Rational Design of Nanoparticles, Fig. 7 Immersed finite element modeling of NP/RBC interaction in a capillary flow

under flow. IVM is used for the final validation in vivo, for elucidating the mechanisms regulating NP dynamics, and for acquiring input parameters for the computational models. These include patient-specific data on the vascular geometry, blood flow condition, blood vessel thickness and diameter, expression of specific receptor molecules.

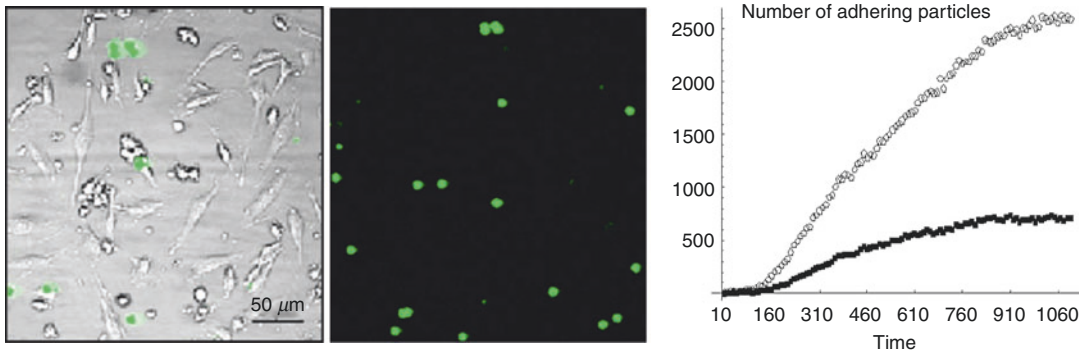
The microfluidic chip apparatus presented in Fig. 8 can be used to estimate the rate of rolling and adhesion of nanoparticles and circulating monocytes [19]. This consists of (1) the MFC (inlet, flow deck, gasket, coverslip, and outlet), (2) a fluorescent confocal microscope, (3) a digital camera, (4) a pumping system, (5) connecting tubings, and (6) a PC. The chip is installed over the microscope stage. An external pumping system (syringe pump) is connected to the chip through tubings and connection kits. The working fluid (physiological solution with and without RBCs) is continuously infused into the chip through a main syringe pump at a constant velocity defined by the desired wall shear rate S, whereas a bolus of the sample particles is injected within a short period of time through an additional syringe. The chip is installed on the stage of the microscope, with the transparent slide upright facing the objectives. Lenses with different magnification ratios are used depending on the size of the region of interest (ROI), varying from about $100 \times 100\ \mu m^2$ for a dry $40\times$ with digital focusing up to $2 \times 2\ \mu m^2$ with a $10\times$ objective.

Parallel plate flow chambers have been used to characterize the behavior of circulating cells, as



Integrated Approach for the Rational Design of Nanoparticles, Fig. 8 Experimental setup for a microfluidic chip. The transport dynamics of NPs and cells can be studied under controlled biophysical conditions

leukocytes and platelets, and micro-/nanoparticles in vitro. By decorating the NP with fluorescent molecules, the transport along the chamber and the adhesive dynamics to the bottom cell layer can be analyzed over time. Based on the imaging analysis, the surface density of adhering NPs and the rate of adhesion and decohesion can be measured over time and compared with those derived theoretically. Images from a typical flow chamber experiment are given in Fig. 9, showing, from left to right, fluorescent beads adhering to the cellular layer (imaged in bright and fluorescent field) and



Integrated Approach for the Rational Design of Nanoparticles, Fig. 9 Outcome of a typical MFC experiment. From *left* to *right*: bright field and fluorescence image of NPs adhering to a layer of HUVECs, fluorescent image identifying the NPs, time-dependent adhesion of NPs under flow

the final result with the variation of the number of beads adhering to a cell layer over time.

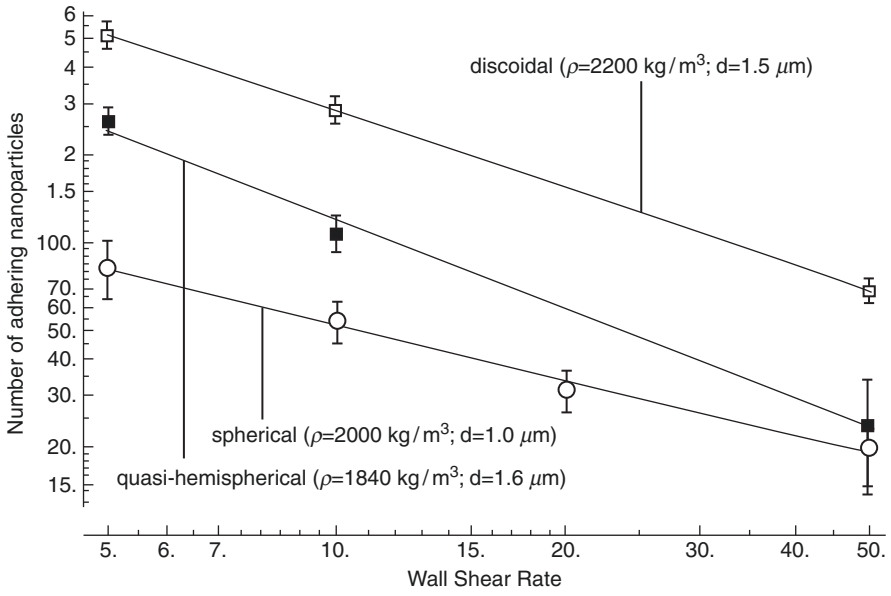
The higher propensity of nonspherical particles to marginate and adhere more avidly to a substrate under flow has been also verified experimentally in vitro, as in Fig. 10, where the number of particles depositing over a layer of fibronectin under a laminar flow in a parallel plate flow chamber has been measured for three different particle geometries (sphere, quasi-hemisphere, and discoidal particles) and different shear rates S [19]. These experimental results support the notion that nonspherical particles can more effectively marginate toward the vessels and more avidly adhere under flow.

Flow chambers can be heated up to 37 °C or cooled down to 4 °C to perform internalization experiments over time under flow or in a quiescent medium. Therefore, the same apparatus can be also used to measure the uptake rates of NPs as a function of their geometrical and physicochemical properties.

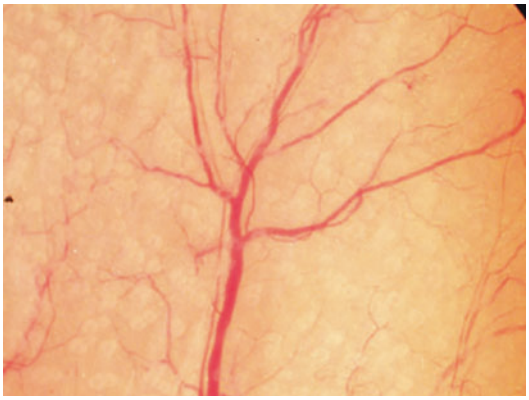
Intravital microscopy (IVM) is used to collect dynamic, quantitative data in vivo with cellular and subcellular resolution. Techniques have been developed that allow the dynamics of individual cells and nanoparticles to be monitored and quantified in real time in live small animals [20]. A variety of organs can be studied acutely or longitudinally via surgical manipulations, including the aorta and arteries, the liver, spleen, and solid tumors. Nanoparticle and cell localization relative to blood vessels and specific cells of interest (as the ECs) can be visualized with the selection of appropriate vital dyes and/or transgenic animal strains. The behavior of cells and nanoparticles within the microvasculature can be correlated to local flow conditions by simultaneously tracking fluorescently tagged red blood cells in the region of interest. Cell/particle trajectory and velocity can be quantified across multiple frames.

The architecture of vascular networks can be determined using intravascularly injected fluorescent tracers, as tetramethylrhodamine isothiocyanate–labeled dextran (TRITC-dextran) with a molecular weight of 155 kDa. The tracers perfuse the vessels after tail vein injection, and images are taken at sufficiently short rates. A sample image of a vascular network from an intravital microscopy analysis is shown in Fig. 11. From the reconstructed network, the following morphological parameters can be estimated from off-line imaging analysis: volume density (volume of vessels per unit volume of tissue), vessel diameter, lineal density (length of vessels per unit volume of tissue), branch point density (number of vessel bifurcations per unit tissue volume), segment length (vessel length between branch points), and tortuosity index (segment path length/shortest distance between branch points).

The permeability of the vascular walls can be assessed by injecting calibrated fluorescent nanoparticles with increasing size, ranging from 100 nm up to 1 μm. After clearance from the



Integrated Approach for the Rational Design of Nanoparticles, Fig. 10 MFC analysis of NPs with different shapes. Discoidal NPs exhibit the higher propensity to marginate and firmly adhere to the chamber substrate under flow (Adapted from [19])



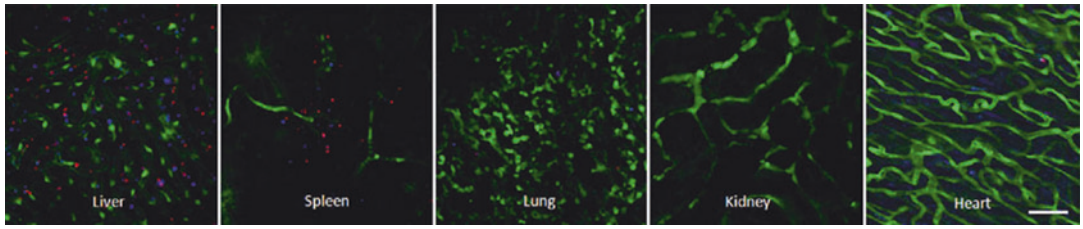
Integrated Approach for the Rational Design of Nanoparticles, Fig. 11 Complex vessel architecture imaged in a mouse using intravital video microscopy

circulation, the extent of particle extravasation can be observed and the pore cutoff size determined as the range between the largest particle size that could extravasate and the smallest size that could not extravasate. In addition to this, the vascular permeability within a ROI can be quantified by analyzing the extravascular diffusion of fluorescent blood tracers, as fluorescein-labeled dextran. The permeability parameter determined through the procedures described above can be used in the mathematical models.

The dynamics of intravascularly injected NPs can be monitored in real time. In particular, the surface density of adhering particles can be measured within the ROI and compared with the mathematical predictions (Fig. 12) [20].

Future Directions

The integrated approach for the rational design (IARD) will help in developing patient-specific therapeutic and imaging agents with superior efficiency and efficacy through the use of (1) multiscale computational tools, for the analysis of the vascular and extravascular transport of nanoparticles, cells, and molecules; (2) the fabrication of sophisticated fluidic chips, resembling the biological complexity of authentic vascular networks; and (3) in vivo imaging, for tracking the dynamics of circulating agents and their performance.



Integrated Approach for the Rational Design of Nanoparticles, Fig. 12 Organ-specific and time-dependent accumulation of systemically injected NPs in tumor-bearing mice. Intravital microscopy allows us to quantify the rate and level of accumulation of NPs in multiple organs (Adapted from [20])

Acknowledgments The author wants to acknowledge the collaboration with Dr. Thomas J.R. Hughes from the University of Texas at Austin, Dr. Yongjie Zhang from Carnegie Mellon University, and Dr. Shaolie Hossain at The Methodist Hospital Research Institute on the development of the IA for the analysis of molecule and nanoparticle transport. The author wants also to acknowledge the collaboration with Drs. Wing Kam Liu, Tae-Rin Lee, and Adrian Kopacz from Northwestern University on the development of the immersed finite element method for the analysis of RBC/nanoparticle interaction.

Cross-References

▶ Adhesion
▶ Nanoparticles

References

1. Ferrari, M.: Cancer nanotechnology: opportunities and challenges. Nat. Rev. Cancer **5**, 161–171 (2005)
2. Peer, D., Karp, J.M., Hong, S., Farokhzad, O.C., Margalit, R., Langer, R.: Nanocarriers as an emerging platform for cancer therapy. Nat. Nanotechnol. **2**, 751–760 (2007)
3. Decuzzi, P., Pasqualini, R., Arap, W., Ferrari, M.: Intravascular delivery of particulate systems: does geometry really matter? Pharm. Res. **26**(1), 235–243 (2009)
4. Godin, B., Driessen, W.H., Proneth, B., Lee, S.Y., Srinivasan, S., Rumbaut, R., Arap, W., Pasqualini, R., Ferrari, M., Decuzzi, P.: An integrated approach for the rational design of nanovectors for biomedical imaging and therapy. Adv. Genet. **69**, 31–64 (2010)
5. Bazilevs, Y., Calo, V.M., Zhang, Y., Hughes, T.J.R.: Isogeometric fluid–structure interaction analysis with applications to arterial blood flow. Comput. Mech. **38**(4), 310–322 (2006)
6. Zhang, Y., Bazilevs, Y., Goswami, S., Bajaj, C.L., Hughes, T.J.R.: Patient-specific vascular NURBS modeling for isogeometric analysis of blood flow. Comp. Methods Appl. Mech. Eng. **196**(29–30), 2943–2959 (2007)
7. Lee, S.Y., Ferrari, M., Decuzzi, P.: Design of bio-mimetic particles with enhanced vascular interaction. J. Biomech. **42**(12), 1885–1890 (2009)
8. Lee, S.Y., Ferrari, M., Decuzzi, P.: Shaping nano-/micro-particles for enhanced vascular interaction in laminar flows. Nanotechnology **20**(49), 495101 (2009)
9. Peskin, C.S., McQueen, D.M.: Computational biofluid dynamics. Contemp. Math. **141**, 161–186 (1993)
10. Zhang, L., Gerstenberger, A., Wang, X., Liu, W.K.: Immersed finite element method. Comput. Method. Appl. Mech. Eng. **193**(21–22), 2051–2067 (2004)
11. Zhang, L.T., Wagner, G.J., Liu, W.K.: A parallelized meshfree method with boundary enrichment for large-scale cfd. J. Comput. Phys. **176**, 483–506 (2002)
12. Pan, W., Fedosov, D.A., Caswell, B., Karniadakis, G. E.: Predicting dynamics and rheology of blood flow: a comparative study of multiscale and low-dimensional models of red blood cells. Microvasc. Res. **82**(2), 163–170 (2012)
13. Sun, C., Munn, L.L.: Lattice Boltzmann simulation of blood flow in digitized vessel networks. Comput. Math. Appl. **55**(7), 1594–1600 (2008)
14. Decuzzi, P., Ferrari, M.: The adhesive strength of non-spherical particles mediated by specific interactions. Biomaterials **27**(30), 5307–5314 (2006)
15. Chithrani, B.D., Ghazani, A.A., Chan, W.C.: Determining the size and shape dependence of gold nanoparticle uptake into mammalian cells. Nano Lett. **6**, 662–668 (2006)
16. Decuzzi, P., Ferrari, M.: Design maps for nanoparticles targeting the diseased microvasculature. Biomaterials **29**(3), 377–384 (2008)
17. Decuzzi, P., Ferrari, M.: The receptor-mediated endocytosis of nonspherical particles. Biophys. J. **94**(10), 3790–3797 (2008)
18. Champion, J.A., Mitragotri, S.: Role of target geometry in phagocytosis. Proc. Natl. Acad. Sci. U. S. A. **103**(13), 4930–4934 (2006)

19. Gentile, F., Chiappini, C., Bhavane, R.C., Peluccio, M., Cheng, M.C., Liu, X., Ferrari, M., Decuzzi, P.: Margination dynamics of non-spherical particles within micro-capillaries. J. Biomech. **41**, 2312–2318 (2008)
20. van de Ven, A.L., Kim, P., Haley, O., Fakhoury, J.R., Adriani, G., Schmulen, J., Moloney, P., Hussain, F., Ferrari, M., Liu, X., Yun, S.H., Decuzzi, P.: Rapid tumoritropic accumulation of systemically injected plateloid particles and their biodistribution. J. Control. Release **158**(1), 148–155 (2011)

Integrated Biosensors

► CMOS MEMS Biosensors

Integrated Micro-acoustic Devices

Tian-Ling Ren[1], Yu-Feng Wang[2] and Yi Yang[1]
[1]Institute of Microelectronics, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, Haidian, China
[2]Institute of Microelectronics, Tsinghua University, Beijing, China

Synonyms

Micromachined acoustic devices

Definition

Integrated micro-acoustic devices are silicon-based acoustic transducers which are integrated with IC modules such as preamplifiers or analog-to-digital converters (ADCs).

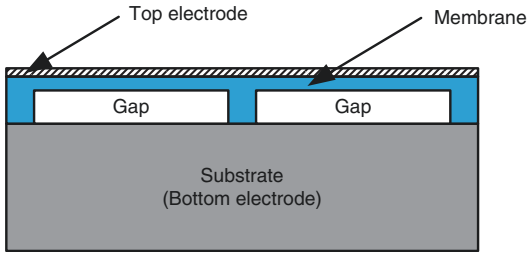
Overview

An integrated micro-acoustic device consists of an acoustic transducer and an integrated circuit (IC) module. The functional part is the acoustic transducer, and it can either be fabricated with the IC module on the same chip or be packaged with the IC module on the same substrate. The transducer part is fabricated on a silicon wafer by MEMS techniques. The fabrication process is usually compatible with IC fabrication. Due to the adoption of semiconductor technologies, integrated micro-acoustic devices have many advantages such as small size, low cost, and high yield. In terms of applications, integrated micro-acoustic devices can be divided into micro-audio devices and micro-ultrasonic devices. The working frequency range of micro-audio devices is from 20 Hz to 20 kHz. Micro-audio devices include MEMS microphones, micro-receivers, and micro-speakers. Commercial MEMS microphones are very mature MEMS products and they have a wide variety of applications in mobile phones, MP3 and other consumer electronics. The working frequency range of micro-ultrasonic devices is typically greater than 20 kHz. They are also known as micromachined ultrasonic transducers (MUTs). Due to their miniature size and high performance, micro-ultrasonic devices have great potential to be used in medical imaging systems or distance measuring devices.

According to the transduction principle, there are three types of integrated micro-acoustic devices: capacitive, piezoelectric, and piezoresistive. Major manufacturers producing integrated micro-acoustic devices include Analog Devices, Akustica (AKU200x), Knowles Electronics, and Memstech (MSMx).

Capacitive Micro-Acoustic Devices

Capacitive micro-acoustic devices are based on electrostatic principle [1–3]. The basic structure is a parallel-plate capacitor (Fig. 1). They convert sound waves into electrical signals by moving an active membrane separated from the silicon substrate by a very thin air gap. The top electrode and the highly doped substrate are conductive, and they form a parallel-plate capacitor. When sound waves hit the membrane, it vibrates, inducing an alternating current from the capacitor. When the capacitor is applied with an alternating voltage, the active membrane will move periodically by electrostatic force and transmit sound waves.



Integrated Micro-acoustic Devices, Fig. 1 Basic structure of capacitive micro-acoustic device

According to the operation frequency, the capacitive micro-acoustic devices can be categorized as capacitive MEMS microphones and capacitive micromachined ultrasonic transducers.

Capacitive MEMS Microphones

Capacitive MEMS microphones are the most successful, commercial integrated micro-acoustic devices which work at the audio frequency (20 Hz–20 kHz). The MEMS microphones are also called microphone chips or silicon microphones. The pressure-sensitive diaphragm is formed directly on a silicon chip by MEMS or CMOS techniques and is usually accompanied with IC modules such as preamplifiers or analog-to-digital converters (ADCs). Due to the adoption of semiconductor techniques, MEMS microphones have some advantages, compared with traditional ECMs (electret condenser microphones), including greater reliability, higher tolerance to mechanical vibration, smaller footprint and digital signal output. Akustica and Knowles are the first two MEMS microphone makers reporting volume shipments. In 2006, Akustica announced a single-chip digital MEMS microphone that has a downsized die that is smaller than 1 mm^2 and Knowles announced a two-chip digital MEMS microphone that includes an ADC on its ASIC in the same year. Later in 2008, Analog Devices Inc. (ADI) and other companies started to enter the MEMS microphone market. A unique aspect of ADI's design is a particle filter constructed by perforating a silicon cover over the microphone's diaphragm. This structure can effectively prevent the dust from reaching the delicate silicon diaphragm and thus improves the performance. Due to the hot sales of smart phones and other consumer electronics, the demands of MEMS microphone increase rapidly. The Yole Développement predicts that the shipment of MEMS microphones will exceed one billion units in 2013 [4]. Now the major manufacturers producing MEMS microphones are Wolfson Microelectronics (WM7xxx), Analog Devices, Akustica (AKU200x), Infineon (SMM310 product), Knowles Electronics, Memstech (MSMx), NXP Semiconductors, Sonion MEMS, AAC Acoustic Technologies, and Omron.

Capacitive Micromachined Ultrasonic Transducers

Capacitive micromachined ultrasonic transducers (cMUTs) are the kind of transducers that can convert ultrasound into electrical signal and transmit ultrasound when they are applied with an alternating voltage. cMUTs have been subject to extensive research for the last two decades. Although they were initially developed for air-coupled applications, today their main application space is medical imaging and therapy. The typical operation frequency of cMUTs used for medical imaging is from several to tens of megahertz.

cMUTs, invented in the mid-1990s by the research group of Stanford University [5], have come a long way in the last two decades and recently reached the market for medical ultrasound imaging [6]. The research team at Stanford University is still one of the most active groups in the design, fabrication, and application of cMUTs [7]. In the early years of research in this field, the main focus was on the basic device fabrication and understanding device operation. Several fabrication processes based on standard surface micromachining techniques have been developed [8–12]. An alternative cMUT fabrication method based on wafer bonding was developed later in 2003 [13]. Equivalent circuit models for CMUTs have been developed to help the design of arrays for practical applications [14–16]. Finite element analysis has been used to understand transducer characteristics (especially cross talk issues) and to optimize transducer response [17–21].

Integration of ultrasonic transducer arrays and integrated circuits is highly desired in many applications, especially for transducer arrays with small elements such as in 2D arrays and arrays for use at the end of a catheter [7]. Several methods have been developed to integrate cMUT arrays with electronic circuits including monolithic integration [22–25] and multichip integration [26, 27]. Recently, flexible cMUT arrays are successfully designed and fabricated using wafer-bonding and PDMS filling techniques [28]. Ultrasound imaging with a forward-looking 2D array and a ring array at the end of a catheter are achieved [7].

Piezoelectric Micro-acoustic Devices

Piezoelectric micro-acoustic devices are based on piezoelectricity and converse piezoelectricity [29–31]. The most widely used piezoelectric materials are lead zirconate titanate (PZT), AlN, and ZnO. Most researches of the piezoelectric micro-acoustic devices focus on the supersonic range. These devices are called piezoelectric micromachined ultrasonic transducers (pMUTs). pMUTs can convert ultrasound into electric signal and will transmit ultrasound when applied with an alternating voltage. The pMUTs working at 40–80 kHz are used in positioning or distance measuring systems, and high frequency pMUTs (typically several or tens mega Hertz) are used for medical imaging systems. In terms of vibration mode, piezoelectric micro-acoustic devices can be divided into d_{33} mode and d_{31} mode.

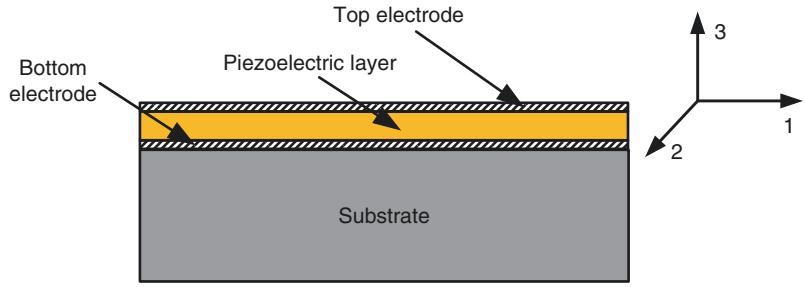
d_{33} Mode

A piezoelectric micromachined ultrasonic transducer working at the d_{33} mode is an acoustic resonance device. The basic structure is shown in Fig. 2. When the sound waves hit the device, the strain of the piezoelectric layer will generate charges on the electrodes due to the piezoelectricity. According to the converse piezoelectricity, the electric field in direction "3" will cause strain in direction "3," so the thickness of the piezoelectric layer will be changed periodically under a sine voltage signal, in other words, the layer vibrates. So the device will transmit sound waves under electrical stimulus. When the phase difference between the sound wave in the surface and the sound wave reflected from the bottom electrode is the multiple of 2π, the device will be in the resonance state. So the resonance frequency is decided by the thickness of the piezoelectric layer.

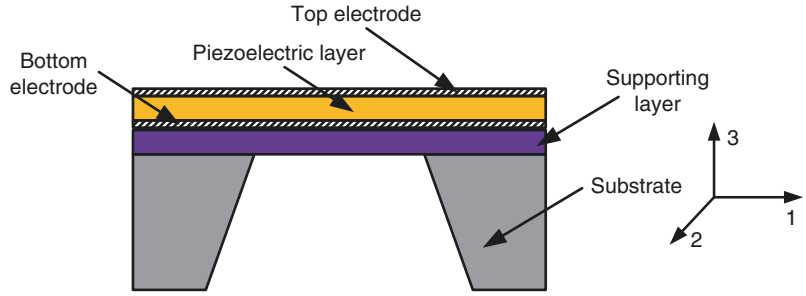
The piezoelectric layer is the key part of pMUTs. The methods to deposit piezoelectric materials include sputter, pulse laser deposition (PLD), MOCVD, Sol–gel, and bulk ceramic bonding. The piezoelectric layers formed by these methods are either very thin (less than 10 μm) or very thick (more than 100 μm). The resonance frequency of pMUTs in d_{33} mode is decided by the thickness of the piezoelectric layer which means pMUTs of certain frequency bands are very difficult to fabricate because of the thickness gap between thin-film deposition and bulk ceramics. Research group of Technical University of Denmark proposed a novel method to deposit thick piezoelectric films [32]. In their work, the piezoelectric elements are formed by deposition of PZT paste into etched features of the silicon substrate such that the depth of these features determines the element thickness and hence the resonance frequency. The process leaves a near-planar surface which is ideal for further wafer-level processing such as top electrode and interconnect formation. Using this method, they fabricated pMUTs with piezoelectric layer thicknesses ranging from 15 to 50 μm. The pulse-echo experiments on these transducers demonstrated their abilities to emit and receive ultrasound waves.

d_{31} Mode

A piezoelectric micromachined ultrasonic transducer working at the d_{31} mode is a mechanical resonance device. The basic structure of the piezoelectric micro-acoustic device at d_{31} mode is a multilayer membrane which consists of a top electrode, a piezoelectric layer, a bottom electrode, and a supporting layer (Fig. 3). The supporting layer can be either silicon or silicon dioxide. When sound waves hit the membrane, the membrane will deform. Thus there will be charges induced on the electrodes due to piezoelectricity.



Integrated Micro-acoustic Devices, Fig. 2 Piezoelectric micro-acoustic device of d_{33} mode



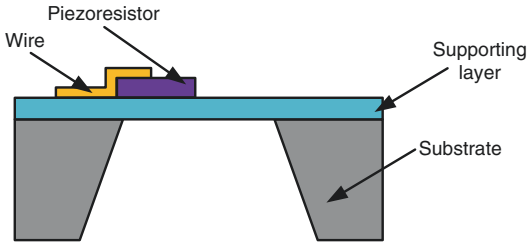
Integrated Micro-acoustic Devices, Fig. 3 Piezoelectric micro-acoustic device of d_{31} mode

According to the converse piezoelectricity, the piezoelectric layer will stretch or compress in direction "1" and direction "2" when it is exercised by an electric field in direction "3" with an opposite phase. So the multilayer membrane will bend periodically by an alternating voltage and transmit sound waves. The resonance frequency is determined by the thickness and size of the multilayer membrane.

d_{31} mode is widely used because its design parameter is more flexible. In the early years of research in pMUTs, SiO_2 or Si_3N_4 serves as the supporting layer. The membrane is released by bulk micromachining processing using wet etching. But it is difficult to control the thickness and residual stress of SiO_2 or Si_3N_4, and these parameters will heavily influence the resonance frequency and the performance of the pMUTs. Researchers in the Research Triangle Park and Duke University use the device layer of SOI as the supporting layer, and the membrane is released using deep reactive ion etch (DRIE) [33]. They have successfully fabricated 2D pMUTs arrays using these techniques, and obtained B-mode images of three metal wires spaced 5 mm apart. Research group of Nanyang Technological University also uses the device layer of SOI as the supporting layer, but the piezoelectric layer is formed by a PZT/silicon wafer-bonding technique, and the bulk PZT is polished down to 10 μm using chemical mechanical polishing [34]. A perforated damping backplate is employed to broaden the acoustic frequency spectrum. Since the resonance of the pMUT array for medical imaging is very high (several to tens mega Hertz), it is necessary to reduce the spacing between elements in the array. Research group of Tsinghua University proposed a new fabrication process based on an SOI/Si wafer-bonding technique [35]. Shallow pits are etched on a Si wafer, then the device layer of an SOI wafer is bonded with the Si wafer, and after removal of the handle layer of SOI, the membrane is released. The element's size and location are defined by the etched pitch. High element density and low cross talk are achieved by these techniques.

Piezoresistive Micro-Acoustic Sensors

Piezoresistive micro-acoustic sensors are based on piezoresistive effect [36–38]. This kind of



Integrated Micro-acoustic Devices, Fig. 4 Basic structure of piezoresistive micro-acoustic device

sensors cannot generate sound, they can only convert sound wave to electric signal. The basic structure of piezoresistive micro-acoustic sensors is shown in Fig. 4. The supporting layer can be either a diaphragm or a cantilever, and it will vibrate when it is hit by sound waves. The strain of the piezoresistive layer will cause the change of resistance due to piezoresistive effect. Polycrystalline Si (poly-Si) is the most common piezoresistor. Typically, a Wheatstone bridge configuration is adopted to convert resistance changes into voltage signals.

Poly-Si piezoresistors on silicon nitride (SiN) were used in early demonstrations of piezoresistive micro-acoustic sensors. However, the gauge-factor of poly-Si is relatively low compared with single-crystal silicon (sc-Si). Research group of Hong Kong University of Science and Technology proposed a novel fabrication process to substitute poly-Si with sc-Si [39]. They bond SOI with a Si wafer which is deposited with silicon nitride. Exfoliation of the handle layer of SOI is accomplished by heat-treatment at 550 C for 30 min. After that the device layer of the SOI which is sc-Si is transferred to the silicon nitride. The diaphragm is released from the front side through cavity etching holes using aqueous potassium hydroxide (KOH). This technique can effectively reduce the footprint. The mode shape and the resonance frequency of the microphone are measured and recorded using an optical vibrometer. For the fundamental mode, the resonance frequency is 520 kHz.

The piezoresistive micro-acoustic sensors can also be used to detect sound source. Researchers in Korea Advanced Institute of Science and Technology developed a piezoresistive multicantilever microphone to locate the sound source by the ITD (interaural time difference) method [40]. In their method, the device layer of SOI wafer is served as the cantilever and boron-doped silicon is used as piezoresistor. Cantilevers with different lengths have different resonance frequencies, so the piezoresistive multicantilever microphone is a broad frequency device. Experiment results show that they can successfully detect the azimuthal angle of sound sources.

Summary

Capacitive micro-acoustic devices are the most mature micro-acoustic devices, mainly because of their excellent performances and ease of integration with IC modules. The capacitive MEMS microphones have already been commercialized. And the demands of MEMS microphones grow rapidly as the conventional electret microphones are substituted. Researches of piezoelectric micro-acoustic devices are mainly focused on piezoelectric micromachined ultrasonic transducers. Although a lot of prototypes with good performance and ability for medical imaging have been fabricated by researchers, it is still not commercialized yet. This is partially because the deposition of the piezoelectric materials is not so compatible with the standard IC process. Different from the capacitive and piezoelectric micro-acoustic devices which can convert sound wave and electric signal into each other, the micro-acoustic devices based on piezoresistive effect cannot generate sound wave but can only be acoustic sensors.

Cross-References

▶ Piezoresistivity

References

1. Jin, X.C., Ladabaum, I., Khuri-Yakub, B.T.: The microfabrication of capacitive ultrasonic transducers. J. Microelectromech. Syst. **7**(3), 295–302 (1998)
2. Jin, X.C., Ladabaum, I., Degertekin, F.L., Calmes, S., Khuri-Yakub, B.T.: Fabrication and characterization of surface micromachined capacitive ultrasonic

I

immersion transducers. J. Microelectromech. Syst. **8**(1), 100–114 (1999)
3. Knight, J., McLean, J., Degertekin, F.L.: Low temperature fabrication of immersion capacitive micromachined ultrasonic transducers on silicon and dielectric substrates. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**(10), 1324–1333 (2004)
4. Ding, W.: MEMS Microphone. Market & Technology Report, Yole Développement (2011)
5. Haller, M.I., Khuri-Yakub, B.T.: A surface micromachined electrostatic ultrasonic air transducer. In: Proceedings of the IEEE Ultrasonics Symposium, Cannes, vol. 2, pp. 1241–1244 (1994)
6. Sako, A., Ishida, K., Fukada, M., Asafusa, K., Sano, S., Izumi, M.: Development of ultrasonic transducer "Mappie" with cMUT technology. MEDIX **51**, 31–34 (2009)
7. Khuri-Yakub, B.T., Oralkan, O.: Capacitive micromachined ultrasonic transducers for medical imaging and therapy. J. Micromech. Microeng. **22**(5), 1–11 (2011)
8. Schindel, D.W., Hutchins, D.A.: The design and characterization of micromachined air-coupled capacitance transducers. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **42**, 42–50 (1995)
9. Haller, M.I., Khuri-Yakub, B.T.: A surface micromachined electrostatic ultrasonic air transducer. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **43**, 1–6 (1996)
10. Eccardt, P.C., Niederer, K., Scheiter, T., Hierhold, C.: Surface micromachined ultrasound transducers in CMOS technology. In: Proceedings of the IEEE Ultrasonics Symposium, San Antonio, vol. 2, pp. 959–962 (1996)
11. Eccardt, P.C., Niederer, K., Fischer, B.: Micromachined transducers for ultrasound applications. In: IEEE Ultrasonics Symposium. Proceedings, Toronto, vol. 2, pp. 1609–1618 (1997)
12. Cianci, E., Foglietti, V., Caliano, G., Pappalardo, M.: Micromachined capacitive ultrasonic transducers fabricated using silicon on insulator wafers. In: Proceedings of the 27th International Conference on Micro- and Nano-Engineering, Grenoble, vol. 61–62, pp. 1025–1029 (2002)
13. Huang, Y., Ergun, A.S., Haeggstrom, E., Badi, M.H., Khuri-Yakub, B.T.: Fabricating capacitive micromachined ultrasonic transducers with wafer-bonding technology. J. Microelectromech. Syst. **12**, 128–137 (2003)
14. Ladabaum, I., Jin, X., Soh, H.T., Atalar, A., Khuri-Yakub, B.T.: Surface micromachined capacitive ultrasonic transducers. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **45**, 678–690 (1998)
15. Caronti, A., Caliano, G., Iula, A., Pappalardo, M.: An accurate model for capacitive micromachined ultrasonic transducers. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **49**, 159–168 (2002)
16. Lohfink, A., Eccardt, P.C.: Linear and nonlinear equivalent circuit modeling of CMUTs. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **52**, 2163–2172 (2005)
17. Bozkurt, A., Ladabaum, I., Atalar, A., Khuri-Yakub, B.T.: Theory and analysis of electrode size optimization for capacitive microfabricated ultrasonic transducers. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **46**, 1364–1374 (1999)
18. Wojcik, G., Mould, J., Reynolds, P., Fitzgerald, A., Wagner, P., Ladabaum, I.: Time-domain models of MUT array cross-talk in silicon substrates. In: Proceedings of the IEEE Ultrasonics Symposium, Rome, pp. 909–914 (2000)
19. Roh, Y., Khuri-Yakub, B.T.: Finite element analysis of underwater micromachined ultrasonic transducers. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **49**, 293–298 (2002)
20. Yaralioglu, G.G., Ergun, A.S., Khuri-Yakub, B.T.: Finite-element analysis of capacitive micromachined ultrasonic transducers. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **52**, 2185–2198 (2005)
21. Bayram, B., Kupnik, M., Yaralioglu, G.G., Oralkan, O., Ergun, A.S., Lin, D., Wong, S.H., Khuri-Yakub, B. T.: Finite element modeling and experimental characterization of crosstalk in 1D CMUT arrays. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **54**, 418–430 (2007)
22. Cheng, X., Lemmerhirt, D.F., Kripfgans, O.D., Zhang, M., Yang, C., Rich, C.A., Fowlkes, J.B.: CMUT-IN-CMOS ultrasonic transducer arrays with on-chip electronics. In: Proceedings of the Solid-State Sensors, Actuators and Microsystems Conference, Denver, pp. 1222–1225 (2009)
23. Noble, R.A., Davies, R.R., King, D.O., Day, M.M., Jones, A.R.D., McIntosh, J.S., Hutchins, D.A., Saul, P.: Low temperature micromachined cMUTs with fully-integrated analogue front-end electronics. In: Proceedings of the IEEE Ultrasonics Symposium, Munich, vol. 2, pp. 1045–1050 (2002)
24. Daft, C., Calmes, S., da Graca, D., Patel, K., Wagner, P., Ladabaum, I.: Microfabricated ultrasonic transducers monolithically integrated with high voltage electronics. In: Proceedings of the IEEE Ultrasonics Symposium, Montreal, vol. 1, pp. 493–496 (2004)
25. Gurun, G., Qureshi, M.S., Balantekin, M., Guldiken, R., Zahorian, J., Sheng-Yu, P., Basu, A., Karaman, M., Hasler, P., Degertekin, L.: Front-end CMOS electronics for monolithic integration with CMUT arrays: circuit design and initial experimental results. In: Proceedings of the IEEE Ultrasonics Symposium, Beijing, pp. 390–393 (2008)
26. Wygant, I., Zhuang, X., Yeh, D., Oralkan, O., Ergun, A.S., Karaman, M., Khuri-Yakub, B.T.: Integration of 2D CMUT arrays with front-end electronics for volumetric ultrasound imaging. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **55**, 327–342 (2008)
27. Wodnicki, R., Woychik, C.G., Byun, A.T., Fisher, R., Thomenius, K., Lin, D.-S., Zhuang, X., Oralkan, O., Vaithilingam, S., Khuri-Yakub, B.T.: Multi-row linear cMUT array using cMUTs and multiplexing

electronics. In: Proceedings of the IEEE Ultrasonics Symposium, Rome, pp. 2696–2699 (2009)
28. Zhuang, X.F., Lin, D., Oralkan, O., Khuri-Yakub, B. T.: Fabrication of flexible transducer arrays with through-wafer electrical interconnects based on trench refilling with PDMS. J. Microelectromech. Syst. **17**(2), 446–452 (2008)
29. Lee, S.S., Ried, R.P., White, R.M.: Piezoelectric cantilever microphone and microspeaker. J. Microelectromech. Syst. **5**(4), 238–242 (1996)
30. Ko, S.C., Kim, Y.C., Lee, S.S., Choi, S.H., Kim, S.R.: Piezoelectric membrane acoustic devices. In: Proceedings of the 15th IEEE International Conference on Micro Electro Mechanical Systems, Las Vegas, pp. 296–299 (2002)
31. Zhu, Y.P., Ren, T.L., Yang, Y., Wu, X.M., Zhang, N. X., Liu, L.T., Tan, Z.M., Wang, H.N., Cai, J., Wang, S. D., Li, Z.J.: Novel ferroelectrics-based micro-acoustic devices and their ultrasonic applications. In: Proceedings of the IEEE International Electron Devices Meeting, San Francisco, pp. 51–54 (2004)
32. Pedersen, T., Thomsen, E.V., Zawada, T., Hansen, K., Lou-Moeller, R.: New technique for fabrication of high frequency piezoelectric micromachined ultrasound transducers. In: Proceedings of the IEEE International Ultrasonics Symposium, Roma, pp. 2115–2118 (2009)
33. Dausch, D.E., Castellucci, J.B., Chou, D.R., von Ramm, O.T.: Theory and operation of 2-D array piezoelectric micromachined ultrasound transducers. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **55**(11), 2484–2492 (2008)
34. Wang, Z.H., Miao, J.M., Tan, C.W.: Acoustic transducers with a perforated damping backplate based on PZT/silicon wafer bonding technique. Sens. Actuator. A-Phys. **149**(2), 277–283 (2009)
35. Wang, Y.F., Ren, T.L., Yang, Y., Chen, H., Zhou, C.J., Wang, L.G., Liu, L.T.: High-density pMUT array for 3-D ultrasonic imaging based on reverse-bonding structure. In: Proceedings of the IEEE 24th International Conference on Micro Electro Mechanical Systems, Cancun, pp. 1035–1038 (2011)
36. Burns, F.P.: Piezoresistive semiconductor microphone. J. Acoust. Soc. Am. **29**(2), 248–253 (1957)
37. Schellin, R., Hess, G.: A silicon subminiature microphone based on piezoresistive polysilicon strain-gauges. Sens. Actuator. A-Phys. **32**(1–3), 555–559 (1992)
38. Schellin, R., Mohr, R.: A monolithically-integrated transistor microphone – modeling and theoretical behavior. Sens. Actuator. A-Phys. **37–38**, 666–673 (1993)
39. Zhou, Z.J., Rufer, L., Wong, M.: Aero-acoustic microphone with layer-transferred single-crystal silicon piezoresistors. In: Proceedings of the 15th International Conference on Solid-State Sensors, Actuators and Microsystems, Denvern, pp. 1916–1919 (2009)
40. Kim, Y.C., Lee, S.S., Kim, J.Y., Park, Y.J.: Sound source localization using piezoresistive Multicantilever microphones. In: Proceedings of the IEEE Sensors, Atlanta, pp. 780–783 (2007)

Integration

► CMOS MEMS Fabrication Technologies

Integration of Nanostructures Within Microfluidic Devices

Meng Lian[1], Bernadeta Srijanto[2], Prachya Mruetusatorn[3,4], Mitchel J. Doktycz[1,2] and Scott T. Retterer[1,2,3]
[1]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[3]Department of Electrical Engineering and Computer Science, University of Tennessee Knoxville, Knoxville, TN, USA
[4]Oakridge National Laboratory, Oak Ridge, TN, USA

Synonyms

Multiscale fluidics; Nanofabrication; Nanofluidics; Nanomaterials; Plasmonics

Definition

Nanostructures encompass a broad class of materials with characteristic or functional lengths below 100 nm. For the purposes of this entry, nanostructures will be defined more broadly to include structures with functional dimensions at the submicron scale. The distinguishing characteristic of the materials discussed here is that their integration within a microfluidic platform is attained through a combination of microfabrication and material synthesis processes that allow the unique functionality of the structure or array of structures to be incorporated directly within the microfluidic system of interest. Generally, nanostructures or nanomaterials are considered to be zero dimensional (nanodots or nanoparticles) or one dimensional (nanowires

and nanorods). Conventionally two-dimensional materials, such as thin-films, are not considered nanostructures, though their synthesis is fundamental to microfabrication and nanostructure modification. This entry includes examples of nanostructured thin films or patterned thin films into discussion as well as monolithic structures that are created using a combination of nano- and microfabrication techniques.

Concepts

Nanostructured materials possess properties that are significantly different than their bulk counterparts. As a result, these materials have garnered a great deal of attention from the scientific community as well as the popular media. Increasingly, these materials are also finding their way into commercial products. Often, the material properties of these nanostructures will depend on the atomic structure of the material of interest, but are enhanced or perturbed by reductions in scale that amplify phenomena that occur at interfaces or result from material defects. Other trends in nanomaterial properties are material independent, relating to the scale of the material rather than its atomic composition. For example, the high surface area-to-volume ratio of nanostructured materials enhances the role that interfacial phenomena play in dictating the interplay between these materials and their environment. Thus, these materials are often the focus of interfacial studies that examine surface adsorption, wettability, and catalytic activity. The scale of these materials also makes them mechanically sensitive, allowing small changes in material mass to quickly impact both the static and dynamic properties of the material. One of the most captivating features of nanomaterials is that they exist and function at the same scale as the most basic elements of biological systems, providing a new approach to interfacing "hard" inorganic and "soft" biological materials. For example, nanoporous materials have pore sizes that approach the scale of biomolecules such as nucleic acids and proteins. This can facilitate both size and charge selective transport as a means of controlling or modulating chemical communication in biological systems, creating a chemical or fluid interface to biology. Because nanomaterials are smaller than most biological cells, they can also provide a model physical interface for cellular systems. The surface chemistry and physical structure of such nanostructured materials can be tuned to influence cell attachment, growth, and even differentiation.

As individual entities or disordered collections of objects, nanostructures clearly exhibit unique properties. Additional functionality can often be achieved when these materials are arranged into ordered ensembles of interacting components. For many applications, ensembles of nanostructures have been created to simply expand the overall coverage and effective size of the nanomaterial array. In such cases the result of the ensemble is to simply scale the phenomena exhibited by any one element over a larger area. Alternatively, the ensemble can exhibit new properties that one might not predict based on the behavior of a single element. Such collective phenomena are often dependent upon the arrangement of materials within the system. Interactions, exchange or coupling of materials, can be tuned or modulated through discrete patterning and assembly techniques that allow control of relative distance and orientation of nanostructures. Because the local environment can dramatically impact such exchanges, changes in the environment can affect the collective behavior of the ensemble. Thus, changes in the collective behavior of ensembles of nanostructures may be indicative of changes in the local environment, making them ideal sensors.

Because of their scale and enhanced functionality, nanostructure ensembles or arrayed nanomaterials are ideal for integration into microfluidic and Lab-on-a-chip (LOC) platforms. Their function can be particularly useful for platforms that are intended to interface with biological materials. As described above, these ensembles can be used as sensing elements to detect changes within the local fluid environment or within a biological system confined within the microfluidic platform. They can be used as passive elements to impart changes in the local environment, shaping fluid flow, molecular transport, reaction kinetics, and electromagnetic fields

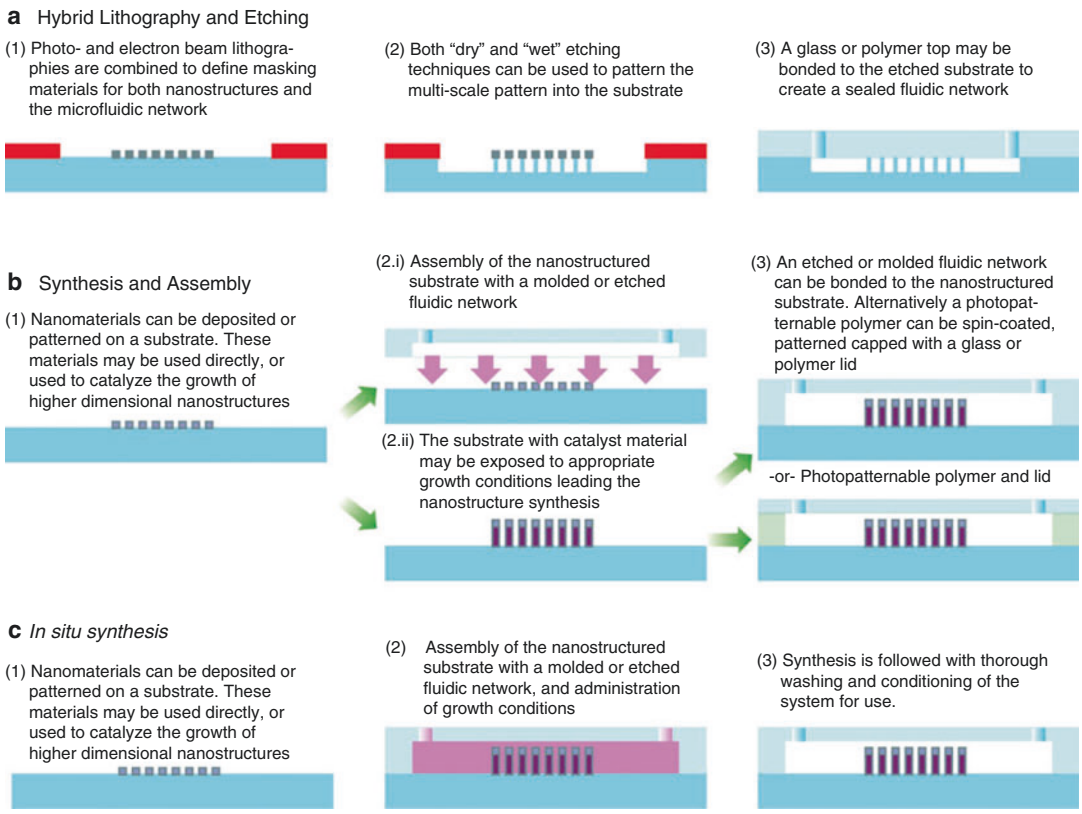
within the system. Furthermore, nanostructure ensembles can be designed to undergo directed changes, respond to changes in their environment, or be actively manipulated to change their functionality. Such alterations can cause local perturbations or lasting changes within the system. Altogether, ensembles of nanostructures provide a unique path toward integrating higher level sensing and actuation into LOC and microfluidic platforms.

Methodology

One of the ongoing challenges of nanoscience is addressing the need to interface with the broad range of nanomaterials and nanoscale structures that one is capable of synthesizing. In order to take full advantage of the tremendous properties that these materials exhibit, one needs to develop technologies that allow the manipulation and interrogation of these materials at time scales that are relevant to the processes in which they are involved. Multiscale fabrication techniques combine conventional microfabrication technologies, honed through decades of integrated circuit and semiconductor manufacturing, with emerging methods for nanomaterial synthesis and assembly, to create integrated platforms in which functional nanomaterial ensembles can be created and taken advantage of.

The strategies applied to the creation of nanostructures within microfluidic devices are lumped into two categories. Top-down strategies involve the step-by-step addition, patterning, and removal of bulk materials and substrate until a desired shape is achieved. Top-down techniques include microfabrication and patterning techniques conventionally used in the semiconductor industry, encompassing both surface and bulk micromachining processes. The resolution of the patterning techniques used to facilitate these processes can be used to create nanostructures that are as small as 5–10 nm, though the true limitations of such processes are dependent on structure composition. In contrast, bottom-up strategies involve the precise placement of nanoscale building blocks, template molecules, and catalysts at molecular levels to build two- and three-dimensional structures. High temperature or solution based synthesis techniques may be involved in the growth of nanostructures or nanomaterials from a patterned catalyst or initiator. The process usually involves the self-assembly of materials into energetically favorable arrangements with a very fine resolution (sub-nanometer range). Both covalent and non-covalent bonds have been utilized to order molecules into nanostructures. The primary challenge in bottom-up approaches is that by nature of the assembly and growth processes involved, long-range order can be difficult to achieve. Typically a combination of bottom-up and top-down approaches can be combined to integrate ensembles of nanoscale structures into functional microfluidic platforms (Fig. 1).

There are three fundamental fabrication processes that are involved in the top-down approach. They are: deposition, lithography, and etching. Combination of these three processes in a layer-by-layer sequence allows for the creation of quasi-three-dimensional structures. As more layers are added, more complex architectures can be created. Deposition processes typically involve the creation of metal, dielectric, or semiconductor thin films using physical vapor deposition processes such as evaporation and sputtering, as well as chemical vapor deposition and oxidation. Common patterning techniques include photolithography, electron beam lithography, and X-ray lithography. These processes rely on the selective exposure and subsequent development of a typically polymeric "resist" material to create a protective layer for material etching processes. Emerging lithographic techniques include soft lithography, nanoimprinting and scanning probe lithography. Soft lithography and nanoimprinting are based on the use of a stamp made of elastomer or hard materials, such as etched silicon, to pattern a protective resist. These processes often require the use of a "master" or "mold" that is created using other lithographic techniques. Each of these techniques balances practical advantages and disadvantages, often sacrificing speed or up-front cost for higher resolution. Etching processes include wet chemical etching techniques and reactive dry processes, carried out in highly energetic plasmas.



Integration of Nanostructures Within Microfluidic Devices, Fig. 1 Methods for integrating nanostructures into microfluidic networks

Detailed overviews of top-down manufacturing processes are provided elsewhere [1, 2]. With respect to the creation of nanoscale structures and ensembles within microfluidic devices, top-down processes can be carried out to create monolithic structures with both micro- and nanoscale features. This is accomplished using sequential electron beam and photolithographic processes to define nanoscale features and microfluidic channels in compatible resist materials respectively. Once patterned, a single directional etch process can be used to transfer this pattern into a silicon, quartz, or glass substrate.

In contrast, bottom-up processes follow the approach of placing small components, at the molecular scale, into more complex assemblies. Predictable and well-controlled growth of nanostructures comprised of carbon (nanotubes, nanohorns, nanofibers) provides an example of such processes. In the case of nanofibers, for example, a catalyst material such as Ni is deposited on a surface. When heated and exposed to a carbon-rich environment, the carbon assembles into distinct cone-shaped sheets of graphene that stack to form nanofibers of carbon with a Nickel catalyst particle at the tip. When formed on a thin-film of Ni, these nanofibers grow as randomly arranged forests, derived from the dewetting of the Ni film and random formation of Ni catalyst particles during the initial heating process. Alternatively, top-down lithographic patterning processes can be used to pattern Ni into distinct shapes, that is, disks and lines, to confine the formation of nanofibers to specific positions. Ultimately it is this combination of top-down and bottom-up processes that provide the most promise for the integration of nanostructures into functional microfluidic devices.

Two enabling approaches to the creation of microfluidic devices with nanostructured ensembles or elements have been the alignment and patterning of SU-8 polymer and assembly of polydimethylsiloxane (PDMS) microfuidic networks, created using soft lithographic techniques [3], to create microfluidic networks that are aligned to nanomaterials synthesized or patterned on glass and silicon substrates. This approach requires careful alignment of microfluidic structures and precise control of nanostructure placement to achieve integration, making it a laborious and low-yield process. Nonetheless, it can provide the quickest route to creating multiscale fluidic systems. PDMS structures can be assembled directly on top of nanostructure ensembles, using an appropriate microscope, manipulator or steady pair of hands, and adequate alignment marks or guides. Alternatively, if the nanostructures are robust and can withstand a polymer spin-coating process, SU-8, or any other photopatternable polymer, can be spin-coated on the sample upon which the nanoscale structures are grown or deposited, and selectively exposed to yield a microfluidic network aligned to the existing nanostructure array.

Applications

A range of nanomaterials have been integrated into and used within microfluidic platforms to create complex sensing and actuation platforms for monitoring and modulating the environment within, and even around, the microfluidic or LOC device. One way to classify or examine these applications is to group them based on the use of nanomaterials as sensors, passive elements, or active modulators of their environment. The examples provided here are not intended to be comprehensive, but rather illustrative of the methods used to create these multiscale platforms, and the unique phenomena that can be taken advantage of in order to realize new levels of functionality within microfluidic platforms.

Nanostructure Arrays as Sensing Elements

The field of plasmonics, the study of charge density oscillations confined to the surface of metallic materials, has garnered a great deal of attention in recent years. Using nanostructured metal materials, including nanoparticle arrays or nanoporous thin metal films, scientists have utilized both plasmon resonance and Surface Enhanced Raman Scattering (SERS) techniques to develop highly sensitive biosensors [4]. Surface plasmon resonance results when plasmons are excited on the surface of metallic nanoparticles by an external electromagnetic field. The frequency at which adsorption, or excitation, occurs is highly dependent on particle size and distribution, particle spacing, and local refractive index. Accordingly arrays of metallic nanoparticles integrated within microfluidic devices have been used to detect the presence of specific molecules within a fluid sample. As these molecules adsorb, or are captured, onto the surfaces of arrayed nanoparticles, local changes in refractive index at the particle surface resulting in changes to the adsorption spectra of the array are indicative of the molecule's presence. Hiep et al. [5] describe the utilization of gold-capped nanoparticles and nanopore arrays to achieve a highly sensitive platform for protein and gene sensors. Chemical modification of the arrayed nanoparticles or pores makes them sensitive to the presence of specific antigens. In this case an antibody to insulin was attached to the particles, and shifts in the absorption peak were observed as insulin, in varied concentrations, was pushed through the microfluidic device. Changes in the Local Surface Plasmon Resonance (LSPR) signal were recorded over time for concentrations as low as 100 ng/mL. Nanoparticle arrays were created via a random dispersion of amine functionalized silica particles onto an oppositely charged (COO− terminated) gold-coated glass slide. A capping layer of gold was added to create the metallic nanoparticle arrays. Assembly of the capped nanoparticle arrays with a molded PDMS microfluidic channel allowed for facile production of the sensor platform. Many similar examples exist within the literature. Most of these utilize electron beam or colloidal patterning of thin metal films via lift-off or etching to create plasmonic materials comprised of metal nanostructure ensembles or porous metal films with discretely spaced and sized pores. Following production of

I

the plasmonic material, SU-8 processing or assembly with PDMS channels can be used to integrate the material into a functional sensing platform.

In an example that illustrates the power of combining top-down and bottom-up synthesis techniques, Kim et al. [6] describe the in situ synthesis of ZnO nanowire arrays by carrying out a controlled hydrothermal reaction within an assembled microfluidic device. In this process, zinc acetate dihydrate and ZnO thin films were used as texture seeds to catalyze ZnO nanowire growth within a microfluidic channel. Photolithography and lift-off processing were used to pattern ZnO seed films and local microheaters that helped drive localized formation of ZnO nanowires. Once patterning was completed, a PDMS channel was bonded to the substrate. Nanowire formation was demonstrated within the channels as either global or local heating was administered while flowing the nanowire precursor solution through the microfluidic device. Following in situ synthesis, the nanowire arrays were used to detect local changes in pH and physically trap particles against fluid flow.

Nanostructures as Passive Elements

Ensembles of nanoscale structures can be used to shape and perturb microflows, influence molecular transport and even modulate chemical reactions within microfluidic systems by simply acting as obstructions that can comprise "cages" or "sieves" within a microfluidic device. In addition to influencing physical flow within a microfluidic device, these obstructions can also be used to perturb or focus electromagnetic fields, creating field concentrations that can influence particle or cellular motion within the fluid flow. For example, Cao et al. [7] describe the use of nanoscale obstructions within a microfluidic device to improve the injection of biomolecules into sets of nanochannels, providing a transition from micro- to nanofluidics, preventing the accumulation and clumping of biomolecules at the nanofluidic inlets. Park et al. [8] developed three-dimensional interconnected pore networks inside microfluidic channels by combining single prism holographic and photolithography techniques. Arranged into a face-centered cubic lattice with a lattice constant of 650 nm, the pores continuously split and recombined the fluid flow to improve mixing efficiency by 84 %.

Hybrid lithography techniques that combine electron beam and photolithography along with anisotropic (directional) silicon etching have been used to simultaneously create arrays of nanoscale structures and microfluidic devices in silicon. These nanoscale structures include cellular-scale microreactors with nanoscale slits as well as arrays of silicon posts with controlled spacing. In both cases, modification of nanoscale structures using plasma enhanced chemical vapor deposition and vapor phase treatment with alkylsilane reagents was carried out to modulate the surface chemistry of the nanostructures and channel. Choi et al. [9] used the chemically functionalized pillar arrays to examine the effects of synthetic crowding and volume exclusion on the transport of green fluorescent protein (GFP) and "supercharged" GFP variants. Fluorescence recovery after photobleaching and computational modeling of molecular transport made it possible to quantify the effects of surface charge and molecular binding on diffusion within the nanostructured silicon post arrays. Retterer et al. [10] demonstrated the fabrication of nanoporous silicon bioreactors using hybrid lithography and etching processes, demonstrating the size selective retention of DNA and reactive enzymes within the nanostructured reactor while delivering small molecules, continuously, to the reaction system. Subsequently, Siuti et al. [11] demonstrated the use of the same reactor within a microfluidic network to carry out cell-free synthesis of GFP using cell-free extract and a GFP encoding plasmid. Protein synthesis was carried out for over 24 h under continuous flow of amino acids and ATP. Lavrik et al. [12] have used similar patterning and etching techniques to create arrays of nanopillars within microfluidic devices for use in chemical separations carried out on a microfluidic chip.

In addition to shaping physical flow and diffusion processes, arrays of insulating nanostructures can be used to shape electromagnetic fields. Arrays of nanoscale elements can be used to

"squeeze" the electric field generated through a microfluidic device into areas of high field over a short distance, creating a very high field gradient, and correspondingly high dielectrophoretic force [13]. The forces resulting from the constricted electric field can be used to concentrate bacteria for analysis or align and assemble nanostructures for further use.

Nanostructures as Actuators

Electrical and optical addressability of nanostructure arrays or ensembles makes it possible to monitor local changes within microfluidic systems by performing electrochemical measurements with nanofiber electrodes [14], monitoring the adsorption spectra of plasmonic materials [4, 15], or analyzing nanowire impedance [8]. While applications that utilize nanostructure ensembles as sensors are practical and appealing, just as intriguing is the possibility of using ensembles of nanostructures to actively and dynamically induce changes within the microfluidic system. Fletcher et al. [16] integrated patterned arrays of carbon nanofibers into microfluidic devices by patterning SU-8 microchannels onto arrays of nanofibers. Prior to patterning the SU-8 channels onto the device, an additional metal patterning step was used to create a set of electrodes that incorporated the carbon nanofibers. Subsequent electropolymerization of polypyrole onto the fibers allowed shrinking and swelling of the nanofiber profile via actuation of the polypyrole coating. Given the nanoscale gaps between the fibers, on-demand capture and release of fluorescent proteins within the microfluidic chip was demonstrated. Future work exploring the integration of functional polymer architectures that shrink or swell in response to changes in environmental factors such as pH or temperature may facilitate the development of smart systems that require no input from an outside user.

A Developing Field

The convergence of nanoscale materials and microfluidics has created a rich field in which sensors and actuators that are created at the cellular and subcellular scale can be integrated into fluid environments that are ideal for the study and manipulation of biological systems [17–19]. This has allowed the development of biological interfaces that can be used to probe cellular systems over prolonged periods under dynamic conditions that were previously impossible to create, quantify, and visualize. This has led to new insights into biological function and the emergence of a host of techniques that promise to facilitate new levels of molecular diagnostics and therapeutic strategies that can be administered within the field and at the point-of-care.

Acknowledgments The preparation of this entry was performed at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US DOE under Contract No. DE-AC05-00OR22725.

I

Cross-References

- ▶ Biomimetic Synthesis of Nanomaterials
- ▶ Micro- and Nanofluidic Devices for Medical Diagnostics
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanoscale Properties of Solid–Liquid Interfaces
- ▶ Nanostructured Materials for Sensing
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Self-Assembly of Nanostructures

References

1. Kovacs, G.T.: Micromachined Transducers Sourcebook. McGraw-Hill, New York (1998)
2. Madou, M.: Fundamentals of Microfabrication. CRC Press, Boca Raton (2002)
3. Qin, D., Xia, Y., Whitesides, G.: Soft lithography for micro- and nanoscale patterning. Nat. Protoc. **5**(3), 491–502 (2010)
4. Anker, J.N., Hall, W.P., Lyandres, O., Shah, N.C., Zhao, J., Van Duyne, R.P.: Biosensing with plasmonic nanosensors. Nat. Mater. **7**, 442–453 (2008)
5. Hiep, H.M., Yoshikawa, H., Saito, M., Tamiya, E.: An interference localized surface plasmon resonance biosensor based on the photonic structure of Au nanoparticles and SiO_2/Si multilayers. ACS Nano **3**(2), 446–452 (2009)

6. Kim, J., Li, Z., Park, I.: Direct synthesis and integration of functional nanostructures in microfluidic devices. Lab Chip **11**, 1946–1951 (2011)
7. Cao, H., Tegenfeldt, J.H., Austin, R.H., Chou, S.Y.: Gradient nanostructures for interfacing microfluidics and nanofluidics. Appl. Phys. Lett. **81**, 3058–3060 (2002)
8. Park, S.G., Lee, S.K., Moon, J.H., Yang, S.M.: Holographic fabrication of three-dimensional nanostructures for microfluidic passive mixing. Lab Chip **9**, 3144–3150 (2009)
9. Choi, C.K., Fowlkes, J.D., Retterer, S.T., Siuti, P., Iyer, S., Doktycz, M.J.: Surface charge- and space-dependent transport of proteins in crowded environments of nanotailored posts. ACS Nano **4**(6), 3345–3355 (2010)
10. Retterer, S.T., Siuti, P., Choi, C.K., Thomas, D.K., Doktycz, M.J.: Development and fabrication of nanoporous silicon-based bioreactors within a microfluidic chip. Lab Chip **10**(9), 1174–1181 (2010)
11. Siuti, P., Retterer, S.T., Doktycz, M.J.: Continuous protein production in nanoporous, picoliter volume containers. Lab Chip **11**, 3523–3529 (2011)
12. Lavrik, N.V., Taylor, L.C., Sepaniak, M.J.: Enclosed pillar arrays integrated on a fluidic platform for on-chip separations and analysis. Lab Chip **10**(8), 1086–1094 (2010)
13. Pohl, H.A.: Dielectrophoresis: The Behavior of Neutral Matter in Nonuniform Electric Fields. Cambridge University Press, Cambridge (1978)
14. Guillorn, M.A., McKnight, T.E., Melechko, A., Merkulov, V.I., Britt, P.F., Austin, D.W., Lowndes, D.H., Simpson, M.L.: Individually addressable vertically aligned carbon nanofiber-based electrochemical probes. J. Appl. Phys. **91**, 3824–3828 (2002)
15. Malic, L., Veres, T., Tabrizian, M.: Nanostructured digital microfluidics for enhanced surface plasmon resonance imaging. Biosens. Bioelectron. **26**(5), 2053–2059 (2011)
16. Fletcher, B.L., Retterer, S.T., Mcknight, T.E., Melechko, A.V., Fowlkes, J.D., Simpson, M.L., Doktycz, M.J.: Actuatable membranes based on polypyrrole-coated vertically aligned carbon nanofibers. ACS Nano **2**(2), 247–254 (2008)
17. Kim, S.M., Lee, S.H., Suh, K.Y.: Cell research with physically modified microfluidic channels: a review. Lab Chip **8**, 1015–1023 (2008)
18. Kumar, C. (ed.): Microfluidic Devices in Nanotechnology: Fundamental Concepts. Wiley, Hoboken (2010)
19. Kumar, C. (ed.): Microfluidic Devices in Nanotechnology: Applications. Wiley, Hoboken (2010)

Intelligent Drug Delivery Microchips

► Stimuli-Responsive Drug Delivery Microchips

Interaction of the Immune System with Nanoparticles

Kirsten M. Pondman[1,2], Robert B. Sim[3,4] and Uday Kishore[1]
[1]Centre for Infection, Immunity and Disease Mechanisms, College of Health and Life Sciences, Brunel University London, London, UK
[2]Neuro Imaging, MIRA Institute, University of Twente, Enschede, NB, The Netherlands
[3]Department of Pharmacology, University of Oxford, Oxford, UK
[4]Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, UK

Synonyms

Nanoimmunology; Nanovaccinology

Definition

In most applications of nanomedicine, nanoparticles will come into contact with the skin, the gastrointestinal tract, the bloodstream, or other tissues in the body. Therefore, the nanoparticles will interact with the immune system. Interactions range from opsonization by immune or nonimmune proteins, uptake by phagocytic cells, and likely generation of antibodies. These immune processes can cause adverse inflammatory conditions and affect the intended target and delivery of the drugs by nanoparticles. Nanotherapeutics can also be strategically designed to interact with specific components of the immune system, either to stimulate or suppress the immune responses. These properties are useful for prophylactic vaccination and treatment of inflammatory and autoimmune diseases.

Introduction

Due to their advantageous physicochemical properties and increased targeting specificity, nanoparticle-based drugs can be effective at

lower dose, with less deleterious side effects. It is, thus, understandable that nanoparticles are frequently proposed for clinical applications, especially targeted drug delivery [1, 2]. In order to be applied in clinics, the nanoparticles have to be safe for patients. Therefore, nanoparticle non-toxicity to cells and tissues is a highly discussed topic in literature. In addition, studies examining interactions of nanoparticles with the immune system and consequent immunotoxicity are on the rise currently. The immune system can significantly influence drug and tissue targeting, that can be achieved with nanotherapeutics. If the immune system is activated and phagocytic cells in tissues and blood are actively ingesting the nanoparticles, they will not be available to deliver a therapeutic to a tumor or other target. Furthermore, if the nanoparticles residing in an area cause the recruitment of immune cells, this may lead to extensive inflammation and immunopathology.

Interaction between nanoparticles and the immune system is like a double-edged sword. This can have desirable as well as detrimental consequences. In vaccination strategy, the stimulation of the immune system by nanoparticles can be exploited to increase the efficiency of the treatment. However, when the immune system is overly stimulated by the nanoparticles causing inflammation, or when the nanoparticles suppress the immune response, it would impair normal challenge to infectious agents.

The Immune System

The immune system's main task is to protect against attack from infectious non-self (viruses, bacteria, fungi, and parasites) and clear from the body altered-self such as apoptotic, necrotic, and transformed cells [3]. It uses a complex network of intercommunicating proteins, cells, and organs to recognize and eliminate a potential pathogen. Recognition of the pathogens is mediated by binding of specific surface receptors to the targets, which trigger downstream responses, ultimately leading to the elimination of the target such as pathogens, apoptotic, and tumor cells. The response of the immune system has to be carefully balanced. A weakened immune response can make the individual succumb to infection, whereas an exaggerated immune response (as in hypersensitivity reactions and autoimmune disorders) can lead to a spectrum of diseases [4].

The immune system consists of two wings: the innate and the adaptive immune system. The innate immune system operates rapidly and is broadly specific. Recognition of the pathogens is achieved by pattern recognition receptors (PRRs), which recognize molecular motifs on the targets, also referred to as PAMPs (pathogen-associated molecular patterns). PAMPs can be charge clusters, neutral sugars, vicinal hydroxyl groups, or acetyl groups. The recognition proteins of the innate immune system work through multiple low-affinity binding. PRRs can be found on cellular membranes (Toll-like receptors; TLRs), in the cytoplasm (NOD-like receptors), and free in the blood plasma (complement proteins; acute phase proteins such as C-reactive protein, serum amyloid protein, and pentraxins; collectins and ficolins).

In the adaptive immune system, the immune cells will interact with the targets through highly specific interactions leading to either antibody production via B cells, or cell-mediated immunity via macrophage activation. In both the innate and adaptive immune responses, interaction with cellular receptors will trigger the synthesis and release of cytokines by the cell and can lead to phagocytosis, and migration and activation of more immune cells. When a pathogen interacts with a dendritic cell (DC), the cell can mature toward an antigen-presenting cell (APC), which migrates to the lymph nodes where they activate T lymphocytes, with the specific receptors for the presented epitope. These T cells can either be cytotoxic T cells, specialized in killing their virus-infected cells, or stimulate B lymphocytes to produce antibodies toward the pathogen [5].

Humoral factors are essential in the immune response toward nanoparticles. Among these, the complement system is a principal player. The complement system consists of over 30 proteins, interacting in three different pathways: the classical, alternative, and lectin pathway, each of which is initiated by recognition of a different array of targets. In the classical pathway, C1q is the recognition subcomponent. In the lectin pathway, this

function is performed by mannose-binding lectin (MBL). Binding of these recognition proteins activates C4 and C2 to form C3 convertase, which is able to cleave C3 into C3a and C3b. In the alternative pathway, coating of the target with C3b is initiated by a conformational change in C3 via spontaneous hydrolysis of C3. C3b bound onto a target is an inducer of phagocytosis, also known as opsonization, and can also trigger further activation of the complement system culminating in a graded association of components C5, C6, C7, C8, and C9 forming the C5–9 or membrane attack complex (MAC), which disrupts the lipid bilayer of cells [6, 7].

Immunotoxicity

The interactions between nanoparticles and the immune system and its consequences vary greatly. The effects and the pathways involved are influenced by factors, such as size, shape, surface charge, hydrophobicity, and coating of the nanoparticles. Upon entry in the body, the nanoparticle surface changes quickly due to adsorbed proteins. This corona contains a range of proteins, which alters over time. Initially, the most abundant proteins from the serum bind to the surface, which are slowly replaced by those with the highest affinity toward the surface of the nanoparticle. In addition, several proteins will attach only onto other proteins in the corona. The physical characteristics of the resulting nanoparticle–protein complexes can be vastly different from those of the native nanoparticle, and examining these differences is essential in order to understand eventual distribution and clearance of the nanoparticles and the biological responses that follow. The binding of proteins can increase the nanoparticles' effective size and change its effective surface charge, which, in turn, will affect uptake by phagocytic cells and interactions with PRR on the cell surface.

Complement proteins are commonly part of the protein corona found on nanoparticles. As described above, these proteins are part of the innate immune response and are able to trigger phagocytosis and enhance immune responses toward the nanoparticles. Surface charge and distribution are the most important factors that influence the extent and the pathway of complement activation. Complement protein fragments, C3b and C4b, can react with nucleophilic groups (OH, NH_2, SH) on the surface of a nanoparticle. Complement fragment C3b is the main opsonizing agent in the complement cascade. Binding of large numbers of C3b molecules onto a nanoparticle will result in rapid elimination of the particles by uptake via phagocytic cells, which will limit the availability of the nanoparticle for drug delivery to the target tissue/organ. Besides opsonization and membrane lysis, complement activation supports humoral- and cell-mediated adaptive immunity through augmenting functions and activation of antigen-presenting cells such as dendritic cells, B and T cells [8]. Furthermore, by-products of the activated complement system (C3a, C4a, and C5a) are potent anaphylatoxins, which can attract cellular infiltration at the site of infection or tissue injury, exaggerating inflammation. A wide range of nanoparticles, including carbon nanotubes (CNTs), iron oxide nanoparticles, fullerenes, polymer particles, and dendrimers, have been shown to activate complement, predominantly via the classical pathway [9–11]. The complement classical pathway is initiated by binding of C1q to the target self- and nonself ligands. The most potent binders of C1q are antigen–antibody complexes. Therefore, the antibodies used for targeting via a nanoparticle platform specifically can induce complement activation via the classical pathway. Nanoparticles containing an array of antibodies or antigens in a highly organized manner can be rapidly recognized by the innate immune system, specially complement recognition subcomponents and, therefore, removed from circulation.

Various coatings have been proposed to limit the formation of a protein corona on nanoparticles, the most well-known example being polyethylene glycol (PEG). This polymer has been shown to prolong the circulation time of the nanoparticles in the blood by reducing their uptake by resident phagocytes in the spleen and liver. Steric hindrance to proteins as well as complement adsorption by PEG leading to reduced

opsonization is the most likely explanation for the reduced phagocytosis of the PEGylated nanoparticles. However, specific antibodies against PEG have been detected following exposure to PEGylated nanoparticles, which would suggest stimulation of the adaptive immune system. Whereas some PEG-derivatives reduce complement activation, other commonly used coatings such as dextran-enhanced complement activation mainly through the lectin pathway [10].

In general, neutral nanoparticles are less likely to be phagocytized than charged nanoparticles (with cationic or anionic surface charges). Larger nanoparticles (over 50 nm) are convenient targets for phagocytosis due to a greater possibility of multiple interactions between cell surface receptors and nanoparticles. The optimal size for uptake is cell-dependent. Dendritic cells (DC) preferentially take up virus-sized particles (20–200 nm), while macrophages prefer larger particles 0.5–5 μm) [12]. Following the uptake of the nanoparticles by phagocytic cells such as macrophages, the clearance route/mechanism depends mainly on the biodegradability of the nanoparticle. While biodegradable nanoparticles are digested and cleared, nonbiodegradable nanoparticles may accumulate inside the cells and tissues for extended periods. Such accumulated nanoparticles may cause granulomas and excessive inflammation that can potentially cause carcinomas. These effects have mainly been studied for carbon nanotubes (CNTs), where after inhalation, granulomas were found in the lungs [13].

Nanoparticles can also interact with other components of the immune system in addition to phagocytes and complement. Few nanoparticles have been shown to alter the cytokine levels expressed by immune cells. An example is Resovist, a dextran-coated iron oxide nanoparticle, which on its own was shown to shift the cytokine balance from helper T cell 1 (Th1) to Th2 immune response, thereby lowering IFN-γ and increasing IL-4 synthesis in cultured splenocytes. Th1 cytokine response promotes cell-mediated immunity via activation of macrophages, whereas Th2 cytokine (including IL-4, IL-5, and IL-13) promotes humoral immune response leading to the production of antibodies. Fullerenes (C_{60} derivatives) had a similar effect but did not interfere with IFN-γ production; it enhanced INF-α and suppressed IL-6 and IL-17 production, suggesting stimulation of regulatory T cells (Treg). Polyvinyl alcohol (PVA)-coated iron oxide nanoparticles interfered with antigen processing by dendritic cells, thereby limiting pro-inflammatory cytokine production and T-cell proliferation.

The immunomodulatory effects are not always clear-cut, and often, a mixed response can be observed. Both pro- and anti-inflammatory effects are induced by gold nanoparticles (AuNP). For example, it has been shown that AuNPs are able to bind IL-1β and TGF-β1 directly, thereby downregulating several important pro-inflammatory pathways. However, at the same time, it was shown that following uptake of AuNPs, dendritic cells secrete pro-inflammatory cytokines, TNF-α, IL-12, and IL-6 [14].

The process of immunomodulation by nanoparticles can be even more complex. CNTs after being administered by inhalation enhanced TGF-β production by alveolar macrophages in mice. This TGF-β was then distributed throughout the body and triggered COX and IL-10 production in the spleen, altering the antibody production by B lymphocytes [15]. In a series of studies, the interplay between complement activation by nanoparticles and the synthesized cytokines has recently been analyzed [11]. These studies showed that complement deposition considerably enhanced uptake of CNTs by immune cells known to overexpress complement receptors such as macrophages and B cells. qPCR and multiplex array analyses showed serum-dependent downregulation of TNF-α and IL-1β and upregulation of IL-12 by CNTs, revealing IL-10 as a crucial regulator during nanoparticle–immune cell interaction. It appears that CNTs offer diverse molecular patterns to the complement system and subsequent handling and immune response by antigen-capturing and antigen-presenting cells, which can shape the innate and adaptive immune axis. Factor H, a negative regulator of the alternative pathway and also known to inhibit classical pathway for certain

C1q ligands, was partly able to inhibit C1q-mediated classical pathway activation by CNTs. C1q and the recombinant globular head modules of C1q, but not factor H, enhanced uptake of CNTs by U937 macrophage-like cell line and human macrophages. Binding of C1q and its globular head modules to CNTs was able to downregulate the production of pro-inflammatory cytokines and transcriptional factor, NF-κB, and NLRP3 inflammasome.

Not only the coating of the nanoparticle can cause immunotoxicity, but the core material can also have an influence, especially when the nanoparticle is degraded inside the body or cells. An example is cadmium, which is known to cause T-cell toxicity by inhibiting proliferation. It is not clear if this effect is also caused by cadmium-containing quantum dots.

Adverse effects of immunosuppression by nanoparticles include myelosuppression, thymic suppression, and lowered immune responses to infections and cancer [16]. The increased retention time and barrier-crossing properties of nanotherapeutics could increase the myelosuppression of the nanotherapeutics. This leads to decreased production of erythrocytes, lymphocytes, and platelets in the blood.

Immunomodulation as Treatment

Specific stimulation of the immune system can be beneficial in defense against an invading pathogen or clearance of damaged self. Vaccination involves administration of an antigenic material, or an inactivated/attenuated pathogen to evoke a limited adaptive immune response and generate long-term immunological memory toward a specific pathogen. By attaching pathogen-derived antigens onto nanoparticles, their stability and immunogenicity are increased. It also enables targeted delivery of the antigens and slow release, allowing for more efficient vaccination. A limited number of nanoparticle-based vaccines have been approved for human use, and their numbers are increasing steadily [12]. The role of the nanoparticle in the vaccination can be to carry the antigen to the target cell and aid in the phagocytosis of the antigen into the cell. In addition, nanoparticles can also have an important role in activating an immune pathway that will enhance antigen processing and presentation.

The main principle of nanoparticle-based vaccination strategy is to deliver multiple copies of an antigen array mimicking the way it would be presented on a virus or pollen to the antigen-presenting cell such as DCs, inducing their maturation and subsequent effector mechanisms. Local complement activation by nanoparticles may be desirable in order to enhance the efficacy of a nanoparticle-based vaccine as an adjuvant.

Therapeutic vaccination is used for cancer treatment and has also been proposed for Alzheimer's disease and hypertension. In cancer treatment, the objective is to stimulate the host immune system to fight the cancer cells. Furthermore, several tumors have developed mechanisms/strategies to downregulate the immune response. These include secretion of anti-inflammatory cytokines such as IL-10 and TGF-β, suppression of DC maturation, or engagement/recruitment of Treg cells [14]. By reversing these escape mechanisms, the immune system can be tailored toward mounting a robust anti-cancer immune response. Prophylactic nanovaccinology is commonly applied for the prevention of infectious diseases [12]. Nanoparticles can be used to present arrays of antigens and thereby improve the immunogenic efficacy, which is also dependent on size and surface of the nanoparticles.

Gold nanoparticles (AuNPs) have been used as antigen carriers in the case of antigens for influenza A virus, respiratory syncytial virus, foot-and-mouth disease, and human immunodeficiency virus (HIV-1) [12]. Due to the nature of HIV-1, it is undesirable to inject deactivated viruses. Thus, gp120 structural motif was grafted onto AuNPs to mimic the presentation of oligomannosides by HIV-1 to DCs, ultimately inhibiting HIV-1 infection of $CD4^+$ T cells.

Virus-like particles (VLP) are frequently proposed as ideal candidates for nanovaccines. VLPs are self-assembling, biocompatible capsid proteins that form a nanocage with a virus-like structure. Therefore, VLPs naturally interact with the immune system in a similar way viruses would which gives

rise to a potent immune response, while avoiding the risks involved with natural (deactivated) viruses, especially in immunosuppressed or immunocompromised individuals.

Immunosuppression can be beneficial when nanoparticles are used to treat immune-mediated pathologies (e.g., inflammatory and autoimmune disorders) and also to prevent overactive immune responses in allergy and transplant rejections. Nanoparticle-based therapeutics have been proposed for rheumatoid arthritis, diabetes, uveoretinitis, and thyroiditis. Several nanoparticles do not need the addition of a therapeutic ligand to cause immunomodulation. Several polyamidoamine (PAMAM) dendrimers were found to have anti-inflammatory effects; the extent depended on their size and surface modifications. The effects were most likely caused by inhibition of cyclooxygenases (COX). Hydrophilic fullerenes (C_{60} derivative) were shown to inhibit type I hypersensitivity responses in mast cells and basophils and in a mouse model of anaphylaxis.

A major pathway of immunosuppression is through inhibition of T lymphocytes. This can be achieved either by inhibiting their interaction with APCs or by inhibiting T-cell proliferation. Available agents to target either the antigen presentation or IL-12-mediated signal transduction appear to have poor stability and hydrophobicity; both issues can be circumvented by the use of nanoparticle platform. Unfortunately, the specificity of T-cell suppression is not very high; therefore, undesired off-target immunosuppression can result. Improvements can be made by targeting specific subpopulations of T cells, e.g., regulatory T cells, by tuning the cytokine profile toward stimulation of the proliferation of this type. These T cells are responsible for self-tolerance and downregulation of the immune response; therefore, they can be very useful in the treatment of autoimmune disorders such as multiple sclerosis [16].

Corticosteroids are very helpful for the treatment of chronic inflammatory disorders. However, due to rapid elimination from the body, they have to be administered frequently and in high dose. Therefore, the main goal of nanotherapeutics has been to improve the circulation time of the drugs. Prolonged circulation and increased drug accumulation in inflamed tissues have been achieved with PEGylated polymeric nanoparticles loaded with betamethasone phosphate. Interestingly, liposomes filled with glucocorticoids altered the drug target from T lymphocytes to macrophages and an induced anti-inflammatory cytokine response [16].

Immune cell recruitment is one of the key factors in inflammation. By inhibiting the signaling pathways, cellular infiltration at the inflamed region can be reduced, thereby limiting the inflammation. In leukocyte migration, an essential step is the adherence of leukocytes to the endothelial cells lining blood vessels. By inhibiting this interaction, migration of leukocytes can be avoided [16].

Another approach to treat inflammatory conditions is to alter the cytokine expression profiles of immune cells. This can be achieved by delivering small interfering RNA (si-RNA) to the target cells. By altering specific gene transcription, the cells will be either stimulated to produce anti-inflammatory cytokines (e.g., IL-10) or decrease pro-inflammatory cytokine production (e.g., TNF-α). This approach does require the si-RNA to enter the cell nucleus, which adds another great challenge to the barriers to be crossed by the nanoparticles [17].

Tests

Immunomodulation is not easily monitored in vitro, as it involves many interacting pathways and long-term variations and changes. Therefore, long-term, systematic, multi-parameter in vivo studies would be required to evaluate various immunological aspects of nanotechnology. These tests are time consuming and costly. Therefore, most immunological studies presented in literature focus on single aspects of the immune response, e.g., complement activation or cytokine response by a single cell type.

Clearly, nanoparticles often do not evoke a singular immune response; often, both pro- and anti-inflammatory cytokine production are influenced by the presence of nanoparticles. This effect can be caused by the fact that nanoparticles can enter

different cells through various receptors and pathways. Therefore, in one cell type, they might evoke a pro-inflammatory response, while in another cell, an anti-inflammatory or no response can be seen [16].

No consensus has been reached about the optimal way to test the immune response induced by nanoparticles. The most applicable and internationally recognized standard available is ISO-10993, which recommends using hemolysis, thrombogenicity, and complement activation (in vitro) tests to examine hematocompatibility of medical devices. However, potentially different applications of highly variable nanoparticles available may require specific regulations. Other essential tests involve confirmation of the sterility of the nanoparticle formulation, and the absence of endotoxins, which can be achieved by LAL assays (limulus amebocyte lysate). Contaminations of the nanoparticles might evoke immune responses, which are not related to the nanoparticles. In vitro phagocytosis assays are useful for an initial prediction of the reticuloendothelial system response toward the particles and can give an indication of the distribution of the particles toward liver and spleen. Thus, in vivo studies will be essential to confirm these predictions since many other factors will influence the biodistribution and clearance of the nanoparticles.

It is widely known that nanoparticles may interfere with the outcome (both false negatives and false positives) of the test, mainly the high optical absorbance of the particles will interfere with all dye and fluorescence-based assays. Therefore, tests should be calibrated for use with the specific nanoparticles.

Conclusions

The immune systems can have a major influence on the biodistribution, efficiency of treatment, and the adverse effects caused by nanoparticles. The main parameters influencing the interactions of nanoparticles with the immune system are the size, surface charge, and chemistry of the nanoparticles. Several nanoparticles have been shown to cause significant inflammatory responses, while others seem to suppress the immune response through pathways involving the complement system.

Nanovaccinology for immunosuppression as well as immunostimulation is still in a very early stage of development. Thus, there is a clear argument to include immune responses to nanoparticles as specialized tests, leading to accurate and reproducible results, with the possibility to compare a range of nanoparticles.

References

1. Thanou, M.: Nanoparticles for drug and gene delivery. In: Roberts, G.C.K. (ed.) Encyclopedia of Biophysics, pp. 1686–1691. Springer, Berlin (2013)
2. Dobrovolskaia, M.A., Aggarwal, P., Hall, J.B., McNeil, S.E.: Preclinical studies to understand nanoparticle interaction with the immune system and its potential effects on nanoparticle biodistribution. Mol. Pharm. **5**, 487–495 (2008)
3. Janeway, C.A., Travers, P., Walport, M., Shlomchik, M.J.: Immunobiology, 5th edn. Garland Science, New York (2001)
4. Lydyard, P.M., Whelan, A., Fanger, M.W.: Instant Notes Immunology. BIOS Scientific, Oxford (2000)
5. Abbas, A.K., Lichtman, A.H.: Cellular and Molecular Immunology. Saunders, Philadelphia (2009)
6. Kolev, M., Le Friec, G., Kemper, C.: Complement – tapping into new sites and effector systems. Nat. Rev. Immunol. **14**, 811–820 (2014)
7. Law, S.K.A., Reid, K.B.M.: Complement: In Focus. Oxford University Press (1995)
8. Carroll, M.C.: The role of complement and complement receptors in induction and regulation of immunity. Annu. Rev. Immunol. **16**, 545–568 (1998)
9. Rybak-Smith, M., et al.: Recognition of Carbon Nanotubes by the Human Innate Immune System. In: Klingeler, R., Sim, R.B. (eds.) Carbon Nanotubes for Biomedical Applications, pp. 183–210. Springer, Heidelberg (2011)
10. Moghimi, S.M., et al.: Material properties in complement activation. Adv. Drug Deliv. Rev. **63**, 1000–1007 (2011)
11. Pondman, K.M.: Pristine and Modified Nanoparticles: Complement Activation and Immunomodulation. Enschede (2014)
12. Zhao, L., et al.: Nanoparticle vaccines. Vaccine **32**, 327–337 (2014)
13. Donaldson, K., et al.: Pulmonary toxicity of carbon nanotubes and asbestos – similarities and differences. Adv. Drug Deliv. Rev. (2013). doi:10.1016/j.addr.2013.07.014. [pii] S0169-409X(13)00168-3
14. Almeida, J.P.M., Figueroa, E.R., Drezek, R.A.: Gold nanoparticle mediated cancer immunotherapy. Nanomed. Nanotechnol. Biol. Med. **10**, 503–514 (2014)

15. Mitchell, L.A., Lauer, F.T., Burchiel, S.W., McDonald, J.D.: Mechanisms for how inhaled multiwalled carbon nanotubes suppress systemic immune function in mice. Nat. Nanotechnol. **4**, 451–456 (2009)
16. Ilinskaya, A.N., Dobrovolskaia, M.A.: Immunosuppressive and anti-inflammatory properties of engineered nanomaterials. Br. J. Pharmacol. **171**, 3988–4000 (2014)
17. Zhou, J., Shum, K.T., Burnett, J.C., Rossi, J.J.: Nanoparticle-based delivery of RNAi therapeutics: progress and challenges. Pharmaceuticals **6**, 85–107 (2013)

Interface Excess Free Energy

▶ Surface Tension Effects of Nanostructures

Interfacial Charging Effect

▶ Maxwell–Wagner Effect

Interfacial Energy and Chemical Potential at Nanoscale

▶ Surface Energy and Chemical Potential at Nanoscale

Interfacial Investigation of Protein Films Using Acoustic Waves

Hwall Min[1], Ping Kao[1] and Srinivas Tadigadapa[2]
[1]The Pennsylvania State University, University Park, PA, USA
[2]Department of Electrical Engineering, The Pennsylvania State University, University Park, PA, USA

Synonyms

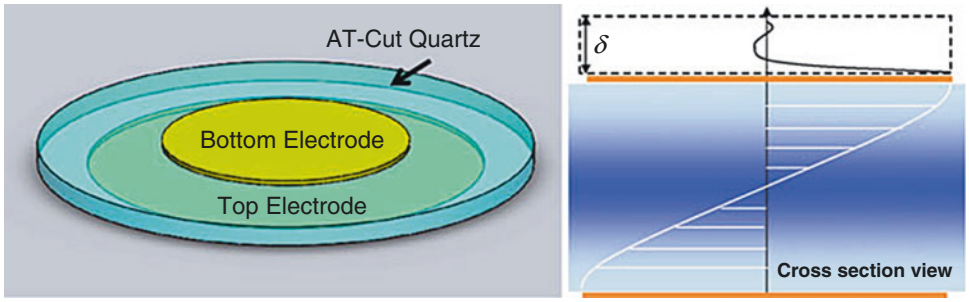
Density; Modulus; Self-assembled protein layers; Thickness; Viscosity

Definition

Bulk acoustic wave quartz crystal resonator can be used for the investigation of the material properties such as the thickness, density, viscosity, and modulus of protein films adsorbed on their surface via self-assembly processes.

Overview

Understanding the phenomenon of the aggregation and assembly of biological molecules such as proteins, lipids, and other biomaterials at the liquid–solid interface is of enormous practical consequence and is not yet well understood. For example, protein adsorption to materials is known to mediate the foreign body response to medical implants in the body. Bacterial adhesion to biomaterials has been found to be mediated by the adsorption of proteins such as fibronectin onto the material surface, creating an interface suitable for adhesion, proliferation, and biofilm formation in medical implants [1–5]. Furthermore, what appears to be driving the host response is not only the chemistry of the material, but the microstructure of individual features within the material onto which host cells can attach. Miniaturized bulk acoustic shear wave resonators exhibit high sensitivity to viscoelastic loading within a few tens of nanometers from the electrode surface and are ideally suited for in situ monitoring of the adsorption of proteins and characterization of their viscoelastic properties. While several techniques are currently available for studying biomolecular assembly processes at interfaces such as atomic force microscopy (AFM), optical reflectometry, surface plasmon resonance (SPR), and techniques based upon protein labeling (e.g., radioactive isotopes) most of these techniques are optimized to either identifying the molecular species or characterization of simple properties such as aerial mass or number density. Acoustic wave sensing techniques based on quartz crystal microbalances (QCM), in principle, are able to characterize not only the physical properties of the layer but also viscoelastic properties of the layer. As will be



Interfacial Investigation of Protein Films Using Acoustic Waves, Fig. 1 Schematic drawing of a quartz crystal resonator showing the *top* and *bottom* electrodes. Application of a sinusoidal electrical stimulus across the electrodes sets-up a shear wave across the thickness of the quartz and resonance occurs at the frequency for which the thickness of the quartz is equal to half the acoustic wavelength

demonstrated in this article, the technique can be used to study adsorption of proteins on plane and textured surfaces with resolution approaching few angstroms to nanometers of the assembling layer.

Principle of Operation

A typical thickness shear mode quartz crystal resonator is schematically illustrated in Fig. 1. It consists of a thin slice of single-crystal, piezoelectric quartz, with very large lateral dimensions in comparison to its thickness, which is sandwiched between two metal electrodes. Application of a sinusoidal electric field using the electrodes sets up a shear wave through the thickness of the quartz. Under the assumptions of ideal free-free boundary conditions of the electrode surfaces, the device exhibits a resonance behavior when the thickness of the quartz slab is half the wavelength of the shear acoustic wave. The fundamental resonance frequency of a quartz resonator can simply be given by

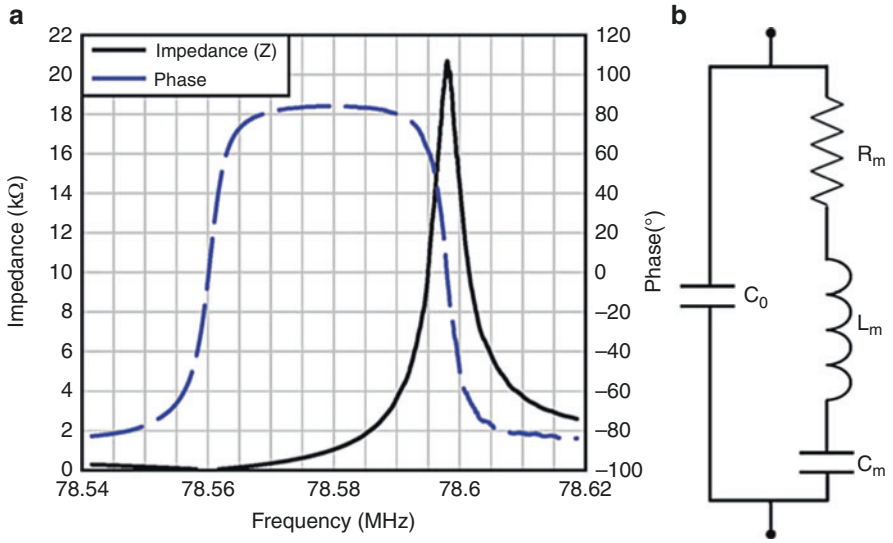
$$f_0 = \frac{1}{2t_q}\sqrt{\frac{\mu_q}{\rho_q}}, \tag{1}$$

where t_q is the thickness of the quartz slice and $\mu_q = 2.947 \times 10^{10}$ kg/m-s^2 and $\rho_q = 2{,}650$ kg/m^3 are the shear modulus and density of quartz, respectively.

Figure 2a shows a typical resonance characteristic of a quartz resonator. Under the assumptions of (1) the quartz crystal is an infinitely large, thin plate, (2) shear wave traveling along the thickness of the quartz plate, (3) negligible material anisotropy, and (4) massless electrodes of negligible thickness, an equivalent circuit representation of the resonator can be realized and is shown in Fig. 2b. At resonance, the equivalent electrical circuit of the QCM is modeled as consisting of a capacitor C_0 in parallel with the motional arm of the circuit consisting of a series circuit of L_m, C_m and R_m and is commonly referred to as the Butterworth-Van Dyke (BVD) model. The values of these parameters can be derived and are given as [6].

$$\left.\begin{aligned} C_0 &= \frac{\varepsilon_{22}}{h}A \\ C_m &= \frac{8Ae_{26}^2}{\pi^2 h\mu_q} \\ L_m &= \frac{\rho_q h^3}{8Ae_{26}^2} \\ R_m &= \frac{h\eta_q\pi^2}{8Ae_{26}^2} \end{aligned}\right\} \begin{aligned} &\text{Static Capacitance} \\ &\text{Motional Arm} \end{aligned} \tag{2}$$

where $\varepsilon_{22}(=3.982 \times 10^{-11}$ F/m) is the permittivity of quartz, A is the area of the energized region of the quartz plate (electrode region), h is the thickness of the quartz plate, $e_{26}(=-0.095$ C/m$^2)$ is the piezoelectric coefficient of quartz, $\mu_q(=2.947 \times 10^{10}$ N/m) is the piezoelectrically stiffened quartz elastic constant, and $\eta_q(=3.5 \times 10^{-4}$ kg/m-s) is the viscosity of quartz.



Interfacial Investigation of Protein Films Using Acoustic Waves, Fig. 2 (**a**) Typical resonance curve obtained for a micromachined quartz resonator. The thickness of the resonator is ~21 μm. (**b**) The equivalent circuit representation of the quartz resonator at resonance shows the motional arm consisting of a series LCR circuit connected in parallel with a static capacitance C_0

The exceptionally low phase-noise characteristically exhibited by quartz resonators allows for a precision of few parts in 10^{14} and is the primary reason for their widespread use in frequency control applications. The phase noise is quantitatively specified in terms of the Q-factor defined as the half-width at the full maximum of the resonance curve or in terms of an energy dissipation factor D, which is inversely proportional to the decay time constant and is also equal to the $1/Q$-factor. The resonance frequency and the Q-factor of the resonance are significantly affected by any surface load placed on the quartz crystal. Generally speaking, loading on the resonator surface can be of three types: (1) pure elastic loads (rigid solids), (2) pure viscous loads (liquids), and (3) viscoelastic loads (polymers) and/or a sequential combination of any of these. The problem of interaction of the transverse shear wave with these surface loads has been treated using three fundamental approaches: (1) evaluation of electrical admittance characteristics upon mechanical perturbations on the resonators [7, 8], (2) using continuum mechanics with appropriate material properties and boundary conditions for each added material [9, 10], and (3) the energy transfer model in which the quartz resonator and the deposited film are treated as a compound resonator, and energy balance methods are used to evaluate the perturbed characteristics of the resonator [11]. These approaches are based on a one-dimensional treatment of the problem since the lateral size of the resonators is typically much larger than the relevant dimensions such as the penetration depth of the acoustic waves in the direction perpendicular to the surface of the resonators.

Elastic Loading

The concept of measurement of aerial mass density using a resonating quartz crystal was presented in 1959 by Sauerbrey. For an unloaded resonator with a resonance frequency of $f_0(0)$, the decrease in frequency, Δf, upon loading of a rigid mass of small thickness in comparison to the resonator thickness and under no-slip condition can be given by

$$\Delta f = -\left(\frac{2f_0^2(0)}{\sqrt{\rho_q \mu_q}}\right)\frac{\Delta m}{A} = -\left(\frac{1}{t_q^2(2\rho_q)}\sqrt{\frac{\mu_q}{\rho_q}}\right)\frac{\Delta m}{A}, \qquad (3)$$

where A is the area of the electrode on the quartz crystal and Eq. 1 has been used to simplify the formula. The most common commercially available quartz microbalances (QCMs) have a resonance frequency of 5–10 MHz and typically consist of 25.4–12 mm diameter and 333–167 μm thick AT-cut quartz discs with gold electrodes on each face and exhibit a mass sensitivity of 17.7–4.42 ng · cm^{-2} Hz, respectively. From Eq. 3 it can be seen that the sensitivity to aerial mass density of the QCM can be increased by increasing its resonance frequency. Based on empirical data, the smallest frequency noise in an AT-cut quartz resonator has been approximated to be $\sim 12 \times 10^{-2} f_0^2$. Thus, the highest aerial mass-density resolution that can be achieved by a given QCM can be calculated by dividing this expression by the mass–sensitivity factor in Eq. 3, that is,

$$\left.\frac{\Delta m}{A}\right|_{ultimate} = 0.6 \times 10^{-20} * \sqrt{\rho_q \mu_q} = 5.6 \times 10^{-15}, \quad (4)$$

where is $\left.\frac{\Delta m}{A}\right|_{ultimate}$ is given in g/cm^2. In finite electrode sized resonators, energy trapping is necessary to maintain a high Q-factor and is determined by the effectiveness of the total internal reflection of vibrations at the edge of the electrodes to confine the acoustic energy to within this region only. For the effective suppression of the spurious modes, the rule-of-thumb thickness of quartz to electrode radius ratio is given by

$$\Delta f_{pb} < 0.744 f_{q0} \frac{1}{N^2} \left(\frac{t_q}{r_q}\right)^2, \quad (5)$$

where $f_q 0$ is the expected resonance frequency of the bare quartz crystal of thickness t_q before the mass loading due the electrodes, and Δf_{pb} is the plate back frequency, that is, the frequency after the deposition of the top and bottom electrodes of radius r_q, and N is the overtone number. This implies that as the quartz resonator thickness is reduced, the electrode diameter must also be reduced in order to effectively suppress spurious modes. Excessive decrease in the electrode diameter, however, will result in acoustic energy leakage and poor confinement of energy resulting in small Q-factors. These opposing requirements entail a design compromise. Results show that for 50–90 MHz resonators with 100 nm of gold electrodes on both faces, the optimal electrode diameter ($2r_q$) to resonator thickness (t_q) ratio is ~10–20. Large deviations from this value either way result in resonators with poor characteristics.

Viscous Loading

When used for applications such as biosensing and electrochemical studies, where one of the resonator surfaces is in contact with a liquid medium, the viscous drag effects need to be considered and now determine the ultimate device performance. Behavior of quartz resonators under purely viscous loading was first reported by Kanazawa and Gordon in 1985. Under such conditions, the fundamental mode QCM frequency shift is given by

$$\Delta f = -\frac{f_0^{3/2}}{\sqrt{\pi \rho_q \mu_q}} \sqrt{\rho_{liq} \eta_{liq}}, \quad (6)$$

where ρ_{liq} and η_{liq} are the viscosity and density of the liquid, respectively. The change in the quality factor ΔQ of the resonator under viscous loading conditions can be given by

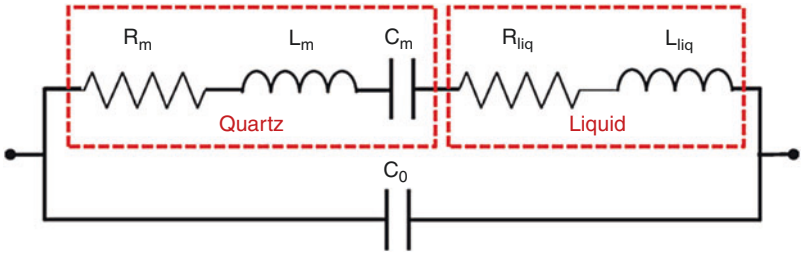
$$\Delta Q = \frac{\pi \rho_q \mu_q}{2\sqrt{\pi f_0 \rho_{liq} \eta_{liq}}} - \frac{\rho_q \mu_q^2}{2\pi f_0 \eta_q}, \quad (7)$$

For operation in liquids, the modified BVD model is shown in Fig. 3. Two additional terms arising from the liquid mass-loading and viscous damping appear in the motional arm of the circuit.

The additional equivalent circuit elements due to the liquid loading can be written as

$$\left.\begin{aligned} L_{liq} &= \frac{\omega_s L_m}{N\pi} \sqrt{\frac{2\rho_{liq}\eta_{liq}}{\omega \mu_q \rho_q}} \\ R_{liq} &= \frac{\omega_s L_m}{N\pi} \sqrt{\frac{2\omega\rho_{liq}\eta_{liq}}{\mu_q \rho_q}} \end{aligned}\right\} \quad (8)$$

The expression for the ultimate aerial mass resolution of the QCM under liquid mass loading conditions is now modified and can be given as



Interfacial Investigation of Protein Films Using Acoustic Waves, Fig. 3 The equivalent model of a quartz resonator when loaded by a liquid of density ρ_{liq} and viscosity η_{liq}. An additional resistance and inductance term appear in the motional arm representing the effect of this load on the resonator characteristics. For a typical resonator, $R_{liq} \gg R_m$ and $L_m \gg L_{liq}$

$$\left.\frac{\Delta m}{A}\right|_{ultimate} = 5.64 \times 10^{-8} \frac{\sqrt{\rho_{liq}\eta_{liq}}}{\sqrt{f_0}}, \qquad (9)$$

where all quantities are specified in SI units. For example, in pure water loading condition, the resolution now varies between 2.529 and 0.596 pg/cm^2 for a 5–90 MHz resonator. Clearly, the ultimate mass resolution in liquids is 10^2–10^3 times lower as compared to the case of mass resolution in air. In the case of liquid operation, it is quite clear that a high-frequency resonator is likely to provide a higher mass resolution. However, often in liquid applications, one is interested in measuring viscosity and density changes in liquid rather than adsorbed mass changes. In this case, the functional dependence of Δf on the density and viscosity of the liquid, can be written as [12]

$$\frac{\Delta\eta_{liq}}{\eta_{liq}} + \frac{\Delta\rho_{liq}}{\rho_{liq}} = 4 \times 10^{-7} \qquad (10)$$

for the ultimate resolution of viscosity and density using a QCM. Since most liquid experiments are accompanied by simultaneous changes in the density and the viscosity of the liquid, Eq. 10 gives the combined minimum resolution in both parameters. Furthermore, in liquid medium applications, the shear wave rapidly damps out as it travels through the thickness of the liquid, and consequently the QCM typically samples a layer of thickness equivalent to the decay length $\delta_{liq} = \sqrt{\eta_{liq}\pi\rho_{liq}f_0}$. For a commercial 5 MHz QCM, the typical decay length in water is ~250 nm, which is large in comparison to the thickness of molecular or even polymer and biomolecular (e.g., proteins) films which can be in a thickness range of 10–50 nm. As a result, a commercial QCM not only samples the adsorbed film but is also significantly affected by the viscous liquid layer above it. Small changes in the viscoelastic properties of adsorbed layers are often masked by the viscous loading from the liquid overlayer. However, if the QCM thickness is decreased, its fundamental resonance frequency increases inversely as a function of the thickness of quartz, and thus the decay length in liquid also decreases consequently. For example, if the thickness of a quartz resonator is decreased from 330 μm (commercial 5 MHz resonator) to 24 μm (miniaturized 66–69 MHz resonator), the decay length in water is reduced to ~68 nm. These factors are critical to the improved abilities of miniaturized devices for probing the viscoelastic properties of adsorbed layers compared to standard commercial QCMs.

Viscoelastic Loading in Liquid Ambient

The QCM frequency shift resulting from the deposition of viscoelastic layer in a viscous liquid ambient can be analyzed using a continuum mechanics approach, for example, as developed by Kasemo and coworkers [10]. In order to model this situation, the QCM surface is considered to be in intimate contact with a layer of continuous viscoelastic slab with an infinitely thick Newtonian liquid overlayer on one of its surfaces. Under the assumption that the thickness of the bulk liquid layer is much larger than the decay length of the acoustic wave in the liquid, the frequency,

Q-factor, and dissipation factor changes with respect to liquid loading conditions can then be written as

$$\Delta f \approx -\frac{1}{2\pi\rho_q t_q} \times \left(t_{visc}\rho_{visc}\omega - 2t_{vise}\left(\frac{\eta_{liq}}{\delta_{liq}}\right)^2 \frac{\eta_{visc}\omega^2}{\mu_{vise}^2 + \omega^2\eta_{visc}^2} \right) \tag{11}$$

$$\Delta Q \approx -2\pi f_0 \rho_q t_q \frac{\left(2t_{vise}\frac{\mu_{visc}\omega}{\mu_{visc}^2 + \omega^2\eta_{visc}^2}\right)}{\left[1 + \left(2t_{visc}\left(\frac{\eta_{liq}}{\delta_{liq}}\right)\frac{\mu_{visc}\omega}{\mu_{visc}^2 + \omega^2\eta_{visc}^2}\right)\right]}, \tag{12}$$

$$\Delta Q \approx \frac{1}{2\pi f_0 \rho_q t_q}\left(2t_{visc}\left(\frac{\eta_{liq}}{\delta_{liq}}\right)^2 \frac{\mu_{visc}\omega}{\mu_{visc}^2 + \omega^2\eta_{visc}^2}\right), \tag{13}$$

where μ_{visc}, η_{visc}, and ρ_{visc} are the elastic modulus, viscosity, and density of the viscoelastic layer, respectively, and t_{visc} is its thickness and δ_{liq} is the penetration depth of the acoustic wave in the liquid.

Basic Methodology

Micromachined Quartz Resonator Arrays

Figure 4a shows a schematic illustration of the fabricated resonator array with eight resonators in the array. The creation of an abrupt step in the quartz substrate for each pixel acoustically and energetically isolates each of these for the shear mode operation [13, 14]. Each resonator is individually addressed through backside (etched-side) electrodes which are extended to the rim of the sensor array chip. The top electrode is common to all the pixels of the array. In order to effectively confine the acoustic energy in each pixel, the pixel thickness to diameter of the electrodes ratio has been carefully optimized as per the conditions described in Eq. 4 [15]. The mounting and packaging of the array and the interconnection were also carefully optimized to maintain a high-quality factor of the resonators. Figure 4b shows the packaged device placed on the liquid test cell. For the fabricated 25–18 μm thick pixels, the fundamental resonance frequency is expected to be in the range of 66.72–90 MHz. Fabricated devices were mounted in a specially designed Teflon test cell capable of liquid testing experiments.

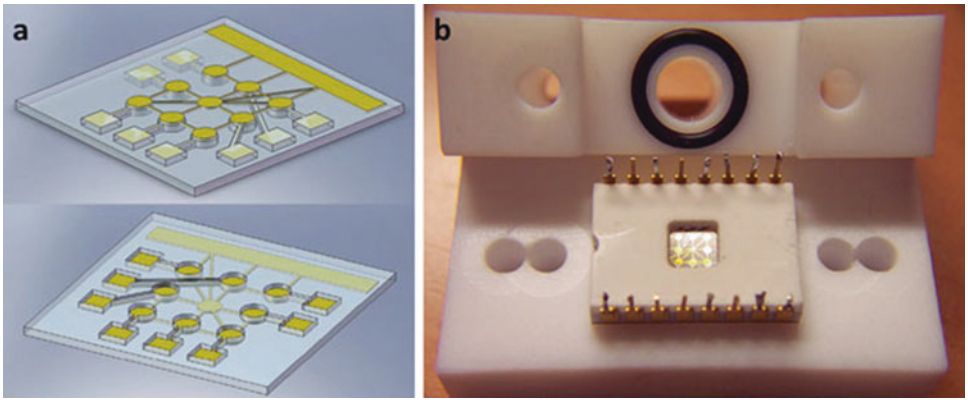
Key Research Findings

Impedance Characteristics in Air and Liquid Loading Conditions

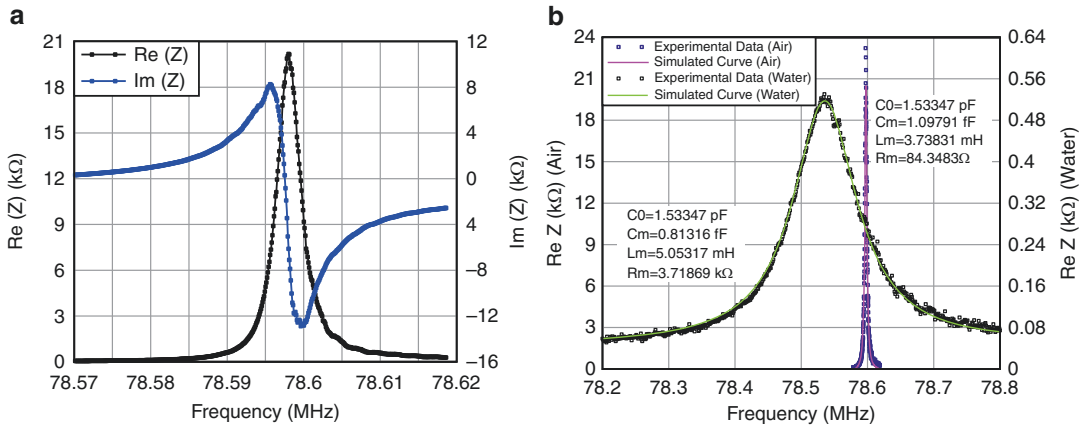
Figure 5a shows the at-resonance real and imaginary impedance curves for one of the eight pixels of the array in air. All the fabricated pixels show similar impedance characteristics. The equivalent BVD circuit model for the quartz resonator yields the following expression for the real part of the impedance (Re Z):

$$\left.\begin{aligned} \mathrm{ReZ} &= \frac{R_m}{(1 - B_0 X_m)^2 + (B_0 R_m)^2} \\ &\text{Where} \\ \mathrm{B}_0 = \omega\, C_0, &\text{ and } X_m = \left(\omega L_m - \frac{1}{\omega C_m}\right) \end{aligned}\right\}, \tag{14}$$

The 801 data points of experimentally measured Re Z near resonance was fitted to the equivalent BVD circuit model of the resonator with C_0, C_m, L_m, and R_m as parameters. Using nonlinear fitting function in Mathematica® software, the best fit was used to obtain the values of the four equivalent circuit parameters. Figure 5b shows the experimental data and the fit obtained in air and water and the values of the resonator parameters obtained using this method. As can be seen, an excellent fit was obtained using this technique and the method was followed for all concentration of water–glycerol mixtures as well. Using this method and Eq. 8 the values of the $(\rho_{liq}\eta_{liq})^{0.5}$ for the various water-glycerol concentrations has been deduced and are listed Table 1 along with the equivalent circuit parameters deduced by nonlinear fitting.



Interfacial Investigation of Protein Films Using Acoustic Waves, Fig. 4 (**a**) Schematic illustration showing the topside and the bottom-side of the eight pixel resonator array used in this study. The top electrode is common to all the eight pixels whereas the resonators are individually accessed using the independent bottom electrodes. (**b**) An optical photograph of a packaged sensor placed in the Teflon test cell



Interfacial Investigation of Protein Films Using Acoustic Waves, Fig. 5 (**a**) Real and imaginary components of the impedance of one of the pixels of the array at resonance. (**b**) The experimentally obtained real part of impedance and the fit obtained to the data using non-linear regression to the expression in Eq. 14 is used to extract the motional circuit element parameters. The resulting simulated curve shows a very close fit to the obtained data in both air and water ambient

Investigation of Interfacial Layer: Protein Adsorption Studies

Micromachined quartz resonators were used to measure the changes in the frequency and the *Q*-factor of the various pixels upon adsorption for five globular proteins as a function of concentration in the solution phase. The five proteins investigated are human serum albumin (HSA), immunoglobulin G (IgG), human fibrinogen (Fib), alpha-2-macroglobulin (AMG), and immunoglobulin M (IgM) at the fundamental and the third resonance modes.

Human serum albumin (HSA), immunoglobulin G (IgG) from rabbit serum, and fibrinogen (Fib) were used as received from Sigma Aldrich and were of highest purity available (>96 %). Alpha-2-macroglobulin (AMG) and immunoglobulin M (IgM) were from MP Biomedicals

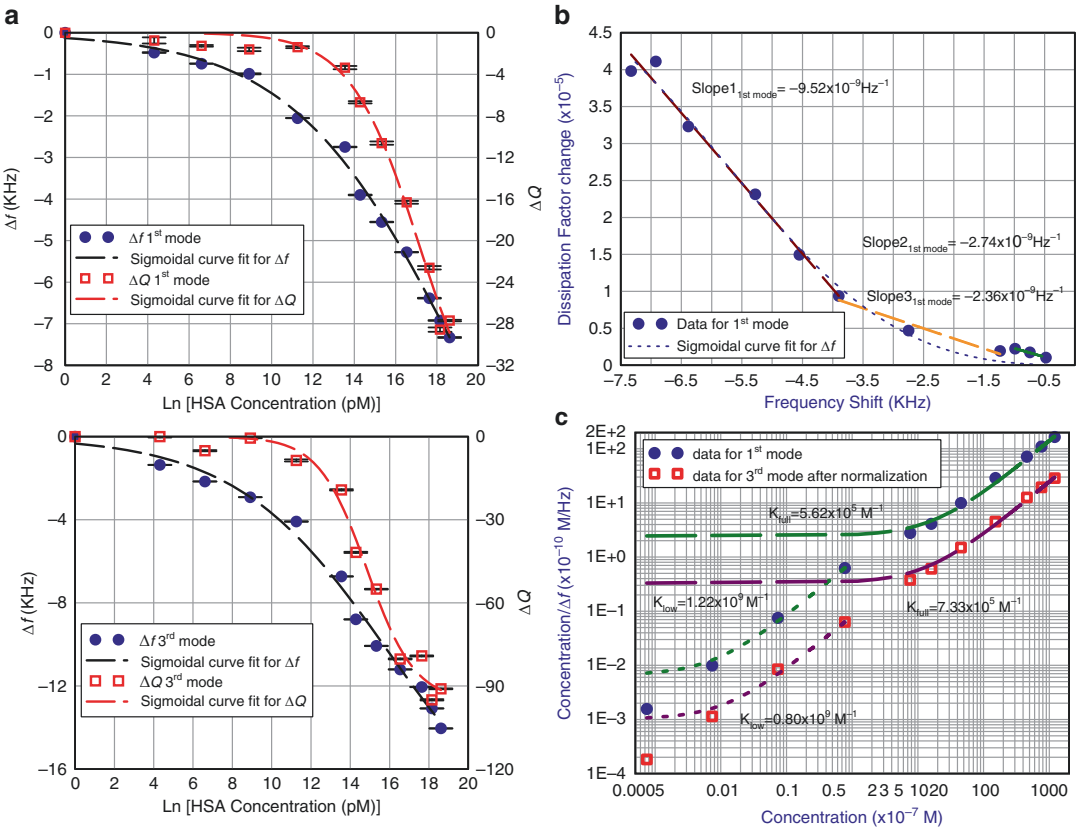
Interfacial Investigation of Protein Films Using Acoustic Waves, Table 1 The extracted equivalent circuit parameters of the quartz resonator in air and various liquid loading conditions. These values are used to calculate the density-viscosity product of various water-glycerol mixture concentrations and compared to the listed values in literature. The method used here has a maximum error of ~28 %

	C_0 (pF)	C_m (fF)	L_m(mH)	R_m (Ω)	R_{liq} (Ω)	$\sqrt{\rho_{liq}\eta_{liq}}\Big\vert_{Calc}$	$\sqrt{\rho_{liq}\eta_{liq}}\Big\vert_{List}$	%error
Air	1.5332	1.09644	3.74238	79.5620				
Water	1.5332	0.80954	5.07558	3,725.65	3,646.08	1.2872	1.0015	28.52
10 % glycerol	1.5332	0.77299	5.31636	3,697.47	3,617.91	1.2195	1.1575	5.36
20 % glycerol	1.5332	0.83761	4.90795	4,392.42	4,312.86	1.5752	1.3608	15.75
30 % glycerol	1.5332	0.81377	5.05300	5,076.26	4,996.70	1.7729	1.6379	8.24
40 % glycerol	1.5332	0.78477	5.24207	6,282.81	6,203.25	2.1224	2.0303	4.53
50 % glycerol	1.5332	0.78756	5.22758	8,232.40	8,152.84	2.7988	2.6103	7.22

and used without further purification. Different HSA, IgG, and fibrinogen concentration solutions were made by diluting a 10 mg/ml stock solution. AMG and IgM were made by diluting a 5 mg/ml stock solution. The 10 mM phosphate buffer solution (PBS) was made by dissolving phosphate buffered saline powder (Sigma Aldrich) into 18 MΩ-cm DI water (Millipore Milli-Q system; Barnstead international). The micromachined quartz resonator pixels were cleaned by three cycles of exposure to ultraviolet (UV) light and ozone; each 15 min long and followed by thorough rinsing with ethanol and immersion in ethanol for 1 h.

The electrodes were exposed to 1 mM hexadecanethiol solution for 24 h. This procedure has been demonstrated to result in the formation of a hexadecanethiolate self-assembled monolayer (HD-SAM). Given the thin (~2 nm), chemically attached, dense nature of the HD-SAM layer, it can be considered as a rigid extension of the electrode. The adsorbed globular protein layer attaches to the HD-SAM surface via van der Waals forces with each molecule surrounded by water and in contact with the phosphate buffer saline solution. The protein layer is considered to be viscoelastic in nature whereas the surrounding PBS solution, containing only a very low-volume fraction of protein molecules, can be represented as a Newtonian fluid. Furthermore, the change in frequency as well as the change in the Q-factor at the fundamental and third mode of the resonator upon adsorption for each protein at various concentrations was measured. Experimental results show a decrease in the frequency and the Q-factor at both the fundamental mode and third overtone of operation with increasing concentration for each protein. Using expressions given by Eqs. 11, 12, and 13, the expected frequency and Q-factor changes can be calculated. In order to model the system, the thickness, density, elastic modulus, and viscosity of the protein layer are considered as parameters and varied to fit to the experimental data. Detailed results and analysis of these measurements are presented elsewhere [16]. Here, the results are briefly summarized.

Figure 6a shows a plot of the resonance frequency shift and quality factor change upon adsorption of HSA at the fundamental and the third overtone modes versus HSA solution concentration. All sets of data show a very close fit with high (>0.99) R^2 value to a sigmoidal function. The observed change in the dissipation factor ΔD versus the change in frequency Δf for various concentrations is plotted in Fig. 6b. The slopes reveal the value of $\partial D/\partial f$ in different concentration regions indicating a varying loss and rigidity of the adsorbed film assembled on the resonator electrode at different concentrations. A low value of $\partial D/\partial f$ indicates a highly rigid, low dissipation film. In contrast, a high value of $\partial D/\partial f$ indicates a soft, dissipative film. In the low-concentration region, a small value of $\partial D/\partial f$ is obtained indicative of a low dissipation. It is reasonable to suggest that the acoustic wave is very slightly damped by protein molecules in low surface coverage region. With increasing surface coverage, the damping of acoustic wave by protein molecules



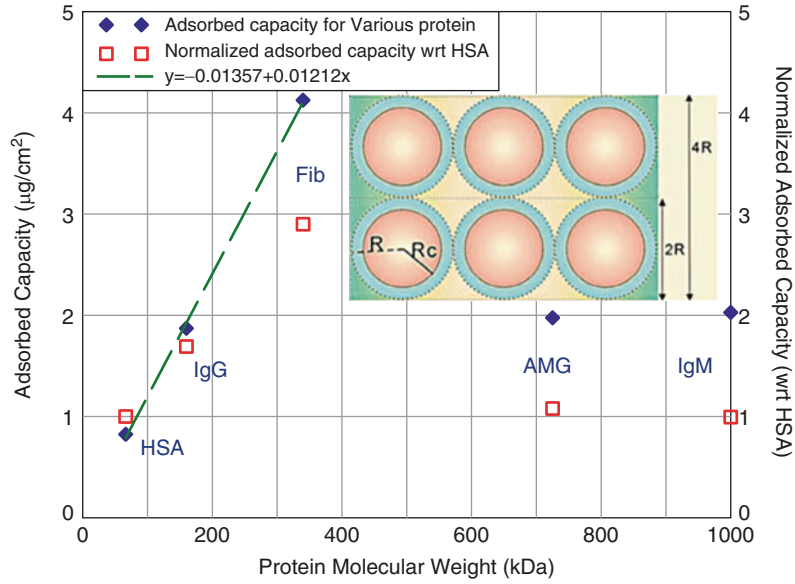
Interfacial Investigation of Protein Films Using Acoustic Waves, Fig. 6 (**a**) Shows the experimentally measured frequency and quality factor shift upon adsorption of human serum albumin (*HSA*) in phosphate buffer solution on a hydrophobic CH_3-terminated surface. The experimentally measured shifts at the fundamental mode (*top*) and third overtone (*bottom*) as a function of the concentration of HSA are shown. (**b**) Shows the dissipation factor ($1/\Delta Q$-factor) plotted as a function of the frequency shift at the various concentrations of HSA. Distinct regions with different slopes can be identified indicating a change in the viscoelastic behavior of the absorbed protein film. (**c**) Langmuir model plot of adsorbed HSA mass, represented by $C/\Delta f$ versus HSA concentration (C), for the entire concentration range tested. The *long-dash line* is a linear fit to all concentration range of data. The *dotted line* is a linear fit in the low concentration region of the data. The spread in the data with respect to the linear fit indicates a deviation from the fit only be seen in the low-concentration region

continuously increases. In the high-concentration (surface coverage) region, the slope of $\partial D/\partial f$ is markedly different from its value at low concentrations. This may be interpreted as gradual conformational changes during the self-assembly process upon reaching critical intermolecular interactions amongst the molecules in the adsorbing film. The reduction of $\partial D/\partial f$ near the saturation region implies a slightly rigid HSA film upon the completion of the adsorption process. The same experiments were repeated for other globular proteins listed and the analyzed using Eqs. 11, 12, and 13.

Table 2 summarizes the physical and viscoelastic properties of the various proteins that were extracted from the QCM measurements [16]. As a simple, preliminary way to model the protein adsorption trends, the proteins can be treated as spheres with a radius $r_v = 6.72 \times 10^{-8}$ MW$^{1/3}$ where the volume of the core protein is considered to be proportional to their molecular weight [17]. For a core protein, including a shell of water molecules, the hydrated protein radius is 1.3 times the core protein radius. These values are in close agreement with the thickness per layer deduced from the QCM results. A linear

Interfacial Investigation of Protein Films Using Acoustic Waves, Table 2 Summary of the protein thickness, density and viscoelastic parameters obtained from the frequency and Q factor shift at the first and third mode of operation [16]

Proteins	Thickness (nm)	Density (kg/m^3)	Viscosity (mPa-s)	Modulus (MPa)
HSA	7.5	1,100	3.5	1.8
IgG	18	1,040	5.5	6.7
Fib	37.5	1,100	2.7	0.42
AMG	19	1,040	4.8	1
IgM	19.5	1,040	2.3	2



Interfacial Investigation of Protein Films Using Acoustic Waves, Fig. 7 Absorbed capacity (density multiplied by thickness) versus protein molecular weight. *Inset* shows the core-shell packing construction

relationship between the MW of the proteins and adsorbed capacity (density x thickness (in µg/cm^2) of the protein layer) on the QCM electrode is observed for three of the five proteins, as shown in Fig. 7. Furthermore, normalization of the adsorbed capacity with respect to HSA (i.e., adsorbed capacity of given protein/adsorbed capacity of HSA) gives a simple integral number for the remaining two of the four proteins measured (see Fig. 7). HSA is known to adsorb on a surface as a single layer, which implies that the integral values obtained for two of the four proteins can be interpreted as adsorption of IgG as two layers and Fibrinogen as three layers. From the estimated value of the hydrated layer, AMG and IgM proteins are thought to adsorb as a single layer.

The chemical and physical properties of the interface and ambient strongly influence the degree of protein adsorption at a solid interface. The relative strength of the thermodynamic adhesion of the protein to the CH_3-SAM surface has been reported in the following order (Fib > Alb > IgG) [18]. Furthermore, the surface concentration of IgG adsorbed on SAMs has been shown to decrease in the order: $CH_3 > C_6H_5OH > COO- > NH_2 > OH$ [19]. Langmuir theory provides a general approach to discussing the adsorption kinetics of proteins by including the interactions between surface active sites and adsorbed species. However, extensive debates on the applicability of Langmuir model for protein adsorption can be found in literature. From the observed irreversibility of the macroscopic adsorption, as demonstrated by observations

that dilution of the protein solution does not lead to a corresponding desorption of the protein films, it can be concluded that a large barrier to desorption exists. For example, addition of pure PBS solution on top of an adsorbed (saturated) protein layer for as long as 24 h showed no observable frequency and Q-factor change.

Clark et al. reported similar observations where, the degree of reversibility of protein adsorption decreased with increasing contact time [20]. This desorption irreversibility has been attributed to time-dependent interfacial conformational changes of the adsorbing protein molecules which result in a strong interaction with the surface. As a consequence of this irreversibility, for any given concentration of the protein solution, the adsorption process leads to saturation coverage after long enough time. However, at low concentrations of proteins, the adsorption process is extremely slow and requires very long times to approach saturation coverage – thus giving the appearance of a concentration-dependent surface coverage akin to Langmuir isotherm. Attempted modifications to Langmuir model of protein adsorption have not been very successful at explaining the kinetics of protein adsorption processes. For now, the Langmuir type of model remains the best way to empirically analyze adsorption data and has been used to interpret the results of HSA adsorption shown in Fig. 6c. It is important to realize that the high-frequency micromachined bulk acoustic resonators described here, with reduced acoustic penetration depths, are able to resolve the interfacial effects in the adsorbed films with high sensitivity.

In the log-log plots of $C/\Delta f$ versus concentration C for HSA, it is clear that only a few data points lie on a straight line. It appears that the data can be fitted by two linear regions of different slope and intercept. The calculated K value has a large variation in the two regions of concentration. To interpret these variations in the K value, the theory needs to include not only adsorption process, but also reorientation, spreading, spatial conformations, and their kinetics at the various concentrations. The ΔD versus Δf graph, Fig. 6b, shows the dissipative effects in the adsorbing film as a function of concentration and relates more directly to reorientation, spreading, and spatial conformation effects. The presence of more than one linear region with different slopes as seen in a plot of $C/\Delta f$ versus concentration C and ΔD versus Δf suggests a competition between adsorption and dissipative processes which are expected to be determined by the native form of the protein.

To compare the effect of surface functionalization on protein adsorption process, QCM gold electrodes were functionalized to produce a high-density carboxyl-terminated hydrophilic surface by chemisorption of 16-mercaptohexadecanoic acid. Identical frequency and Q-factor change measurements and analysis were undertaken for the adsorption of three globular proteins on the COO^- terminated electrode surface. The observed results show similar overall behavior for the adsorption of the globular protein layer on to COO^- surfaces as on hydrophobic surface. The measured change in frequency and Q-factor at the fundamental and third resonance modes were once again fitted to Eqs. 11, 12, and 13 with the thickness, density, elastic modulus, and viscosity of protein layer as parameters. Table 3 summarizes the physical and viscoelastic properties of the various proteins that were extracted from these measurements. However, the parameters of each protein on hydrophobic and charged hydrophilic surfaces differ due to the differences in the interfacial packing of spherical protein molecules on the two surfaces.

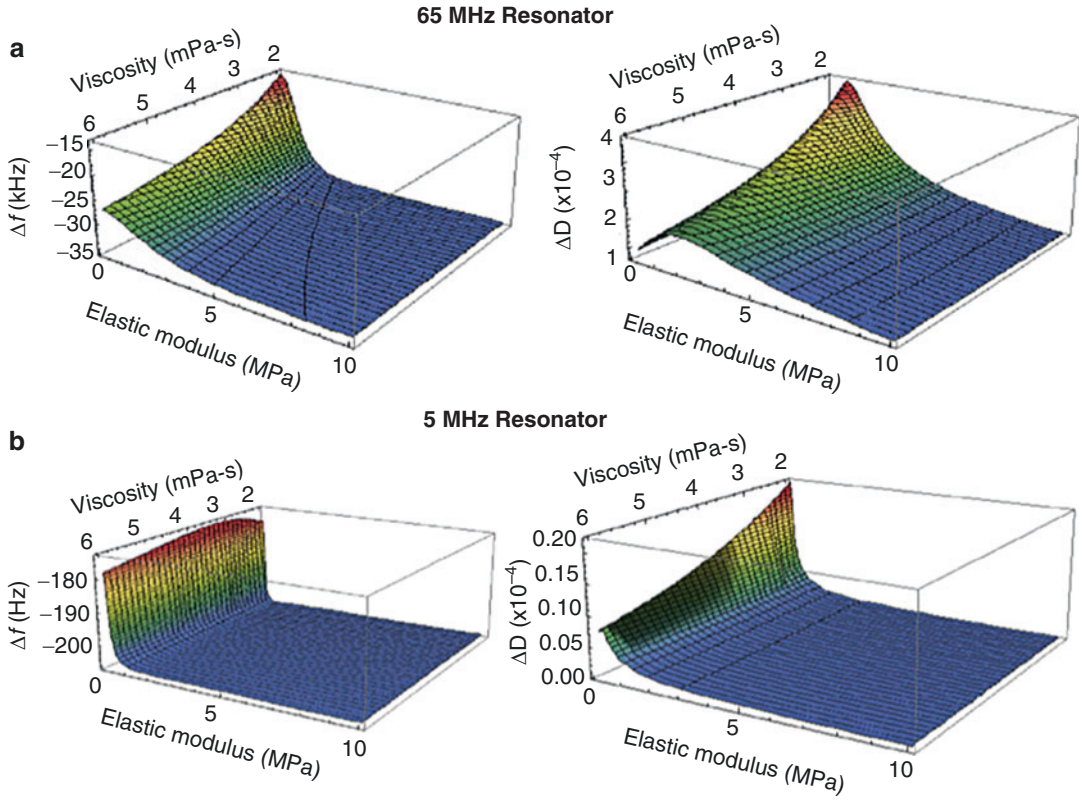
Furthermore, normalization of the surface adsorbed capacity with respect to HSA (i.e., adsorbed capacity of given protein/adsorbed capacity of HSA) gives a simple integral number for the other two proteins on COO^- SAM. The two proteins on hydrophilic surface indicate three layers for IgG, and four layers for fibrinogen on hydrophilic surface.

Future Directions for Research

In conclusion, it has been demonstrated that high-frequency micromachined quartz resonators are especially suited for monitoring small changes in the viscoelastic loading in films adsorbed on their surface. Figure 8 shows the dependence of the frequency and Q-factor shift as a function of the

Interfacial Investigation of Protein Films Using Acoustic Waves, Table 3 Summary of the protein thickness, density and viscoelastic parameters obtained from the frequency and Q factor shift at the first and third overtone of operation

Surface	Proteins	MW (kDa)	Protein diameter (nm)	Thickness (nm)	Layer (s)	Thickness/layer (nm)
Hydrophobic	HSA	66.3	7.1	7.5	1	7.5
	IgG	160	9.5	18	2	9
	Fib	340	12.2	37.5	3	12.5
	AMG	725	15.7	19	1	19
	IgM	1,000	17.5	19.5	1	19.5
Hydrophilic	HSA	66.3	7.1	7.5	1	7.5
	IgG	160	9.5	30	3	10
	Fib	340	12.2	50	4	12.5



Interfacial Investigation of Protein Films Using Acoustic Waves, Fig. 8 Comparison of the frequency and dissipation factor sensitivity of (**a**) 65 MHz and (**b**) 5 MHz resonators to variations in the viscosity and modulus of the adsorbed films with 35 nm of thickness and 1,040 kg/m^3 density. Clearly, the 65 MHz resonator has a better resolution over wider range of values of these properties

viscosity and elastic modulus of a 35 nm thick protein film adsorbed on a commercial 5 MHz and a micromachined 65 MHz resonator. It is clear that the miniaturized resonators provide a greater resolution of these properties. A large variation in the *Q*-factor and the frequency for small changes in the viscoelastic properties of the adsorbed film is highly desirable for accurate

Interfacial Investigation of Protein Films Using Acoustic Waves, Table 4 The sensitivities(resolution) of thickness, density, viscosity, and elastic modulus based on the measurement of five selected proteins

Thickness (nm)	Density (kg/m^3)	Viscosity (mPa-s)	Elastic modulus (MPa)
<0.1 nm	1.98	0.05	0.05

determination of the material properties of the films and for the observation of conformational changes in these nanometer thin films. Based on the measurement of five selected globular proteins, the sensitivities (resolution) of thickness, density, viscosity, and elastic modulus are summarized in Table 4. Careful interpretation of the frequency and *Q*-factor changes can provide a sensitive tool for the analysis and investigation of the biomolecular films in terms of film assembly, packing density, and other conformational changes.

Cross-References

- ▶ Concentration Polarization at Micro-/Nanofluidic Interfaces
- ▶ Integrated Micro-acoustic Devices
- ▶ Laterally Vibrating Piezoelectric Resonators
- ▶ Nanomechanical Resonant Sensors and Fluid Interactions
- ▶ Piezoresponse Force Microscopy and Spectroscopy

References

1. Simpson, K.H., et al.: Measurement of adhesive forces between *S. epidermidis* and fibronectin-coated surfaces using optical tweezers. Lasers Surg. Med. **31**(1), 45–52 (2002)
2. Willcox, M.D., et al.: Bacterial interactions with contact lenses; effects of lens material, lens wear and microbial physiology. Biomaterials **22**(24), 3235–3247 (2001)
3. Bruinsma, G.M., van der Mei, H.C., Busscher, H.J.: Bacterial adhesion to surface hydrophilic and hydrophobic contact lenses. Biomaterials **22**(24), 3217–3224 (2001)
4. Bundy, K.J., et al.: Measurement of fibroblast and bacterial detachment from biomaterials using jet impingement. Cell Biol. Int. **25**(4), 289–307 (2001)
5. Montanaro, L., et al.: Adhesion of a *Staphylococcus aureus* strain to biomaterials does not select methicillin-resistant mutants. New Microbiol. **24**(1), 57–61 (2001)
6. Janshoff, A., Wegener, J., Sieber, M., Galla, H.-J.: Double-mode impedance analysis of epithelial cell monolayers cultured on shear wave resonators. Eur. Biophys. J. **25**, 93–103 (1996)
7. Bandey, H.L., et al.: Modeling the responses of thickness-shear mode resonators under various loading conditions. Anal. Chem. **71**(11), 2205–2214 (1999)
8. Lucklum, R., Hauptmann, P.: The quartz crystal microbalance: mass sensitivity, viscoelasticity and acoustic amplification. Sens. Actuators B **B70**(1–3), 30–36 (2000)
9. Rodahl, M., Kasemo, B.: On the measurement of thin liquid overlayers with the quartz-crystal microbalance. Sens. Actuators A **54**(1–3), 448–456 (1996)
10. Voinova, M.V., Rodahl, M., Jonson, M., Kasemo, B.: Viscoelastic acoustic response of layered polymer films at fluid–solid interfaces: continuum mechanics approach. Phys. Scr. **59**(5), 391 (1999)
11. Mecea, V., Bucur, R.V.: The mechanism of the interaction of thin films with resonating quartz crystal substrates: the energy transfer model. Thin Solid Films **60**(1), 73–84 (1979)
12. Kao, P., Allara, D.L., Tadigadapa, S.: Fabrication and performance characteristics of high-frequency micromachined bulk acoustic wave quartz resonator arrays. Meas. Sci. Technol. **20**(12), 124007 (2009)
13. Rabe, J., et al.: Monolithic miniaturized quartz microbalance array and its application to chemical sensor systems for liquids. IEEE Sensor. J. **3**(4), 361 (2003)
14. Kao, P., et al.: A micromachined quartz resonator array for biosensing applications. J. Microelectromech. Syst. **18**(3), 522–530 (2009)
15. Kao, P., et al.: Human serum albumin adsorption study on 62-mhz miniaturized quartz gravimetric sensors. Anal. Chem. **80**(15), 5930–5936 (2008)
16. Kao, P., Allara, D., Tadigadapa, S.: Study of adsorption of globular proteins on hydrophobic surfaces. Sensor. J. IEEE **11**(11), 2723–2731 (2011)
17. Ostuni, E., et al.: Adsorption of proteins to hydrophobic sites on mixed self-assembled monolayers. Langmuir **19**(5), 1861–1872 (2003)
18. Kidoaki, S., Matsuda, T.: Adhesion forces of the blood plasma proteins on self-assembled monolayer surfaces of alkanethiolates with different functional groups measured by an atomic force microscope. Langmuir **15**(22), 7639–7646 (1999)
19. Silin, V., Weetall, H., Vanderah, D.J.: SPR studies of the nonspecific adsorption kinetics of human IgG and BSA on gold surfaces modified by self-assembled monolayers (SAMs). J. Colloid Interface Sci. **185**(1), 94–103 (1997)
20. Cao, Y.C., Jin, R., Mirkin, C.A.: Nanoparticles with Raman spectroscopic fingerprints for DNA and RNA detection. Science **297**(5586), 1536–1540 (2002)

Interference Modulator (IMOD) Technology

▶ MEMS Display Based on Interferometric Modulation Technology

Intravital Microscopy Analysis

Pilhan Kim[1] and Seok H. Yun[1,2,3]
[1]Graduate School of Nanoscience and Technology, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea
[2]Wellman Center for Photomedicine, Department of Dermatolog, Harvard Medical School and Massachusetts General Hospital, Boston, MA, USA
[3]The Harvard–Massachusetts Institute of Technology Division of Health Science and Technology, Cambridge, MA, USA

Synonyms

In vivo microscopy analysis

Definition

Intravital microscopy is a type of microscopy performed upon the living organism to directly observe dynamic phenomena in high resolution enough to distinguish cellular and subcellular features.

Overview

Intravital microscopy is an emerging type of microscopic imaging technique which can be performed in the living organisms to directly visualize dynamic cellular phenomena in its natural environment. During the last decade, intravital microscopy has evolved to become an indispensable technique to address key questions in a wide area of biomedical fields such as molecular biology, cell biology, immunology, neuroscience, and tumor biology. Real time direct visualization of dynamic cellular processes in action under its natural environment in vivo by intravital microscopy has provided unprecedented insights that were impossible to obtain by static observation of thinly sliced chemically fixed tissues ex vivo by conventional light microscopy (i.e., histology, immunohistochemistry). In addition, although in vitro model studies have enabled many significant findings with its flexibility and manipulability, it fails to reconstitute the cellular environment in specific tissue under physiological condition that can only be found in in vivo models, which makes the intravital microscopy analysis unique and invaluable for many biomedical investigations. For the field of nanotechnology toward biomedical applications such as nanomedicine and nanoparticles for drug delivery, the utilization of intravital microscopy that can monitor the dynamic behaviors of individual nanoparticles in vivo is highly attractive.

Principles

Intravital microscopy is defined as a type of microscopy that can directly image the living organism. In practice, categorized by the combination of key techniques to implement intravital microscopy, there are two most widely used forms of microscopies, *fluorescence confocal laser-scanning microscopy* and *fluorescence two-photon laser-scanning microscopy*. All major microscopy companies such as Carl Zeiss, Olympus, Nikon, and Leica have their own versions of these two types of microscopies and all of them share same working principles to construct image – fluorescence detection and sectioning.

Fluorescence

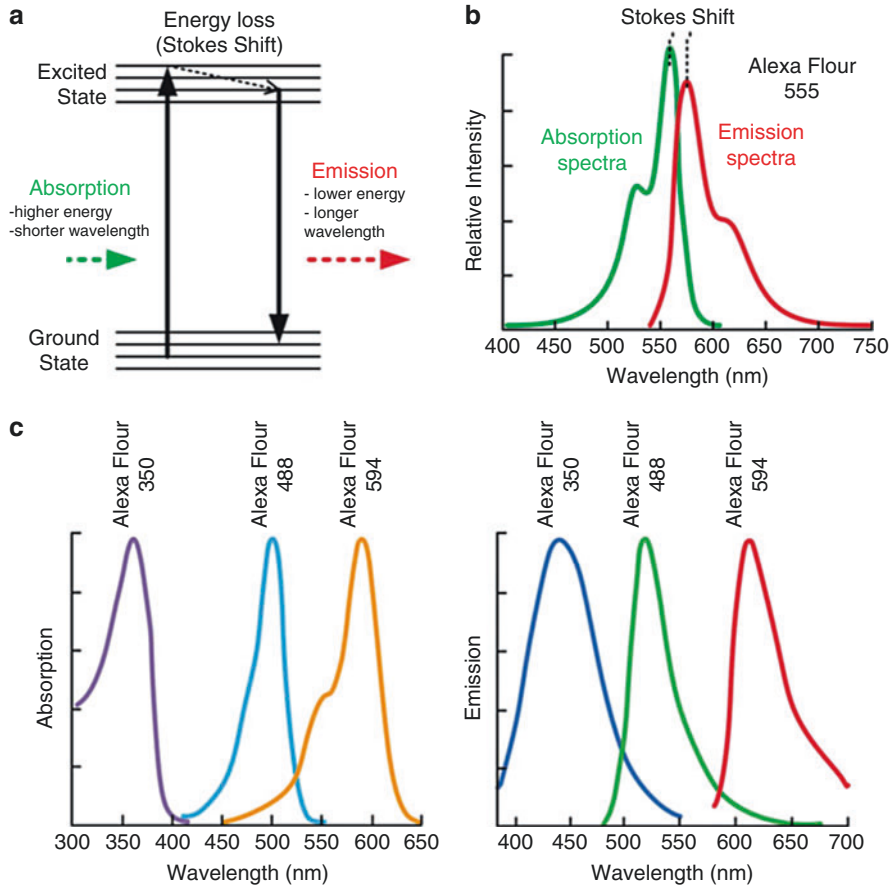
Currently, intravital microscopy is mostly performed by observing *fluorescence* emitted from a *fluorophore* as a contrast to create an image, which is categorized as *fluorescence microscopy* [1, 2]. It can reveal only objects of interests tagged by fluorophore from a black background, providing high selectivity, sensitivity, and

contrast. Over the past decades, tens of thousands of fluorescent probes were synthesized by organic chemists to provide a toolbox to label virtually any components in biological systems. In addition, the development of fluorescent proteins, most notably green fluorescent protein (GFP), has allowed the intrinsic endogenous tagging of numerous components in living systems by genetic engineering without the hassle of devising a labeling scheme using exogenous probes. Lastly, rapid technical advances in light sources and detectors optimized for fluorescence imaging make fluorescence microscopy to become an ideal technique for the observation of all biological specimens either alive (in vivo) or fixed (ex vivo) by allowing the detection of molecules at very small concentration with superior signal-to-background noise ratio.

Fluorescence is the emission of light from a fluorophore through after its absorption of light. Figure 1a shows typical electronic states of a fluorophore in Jablonski diagram. Each electron state has multiple vibrational states. The energy of absorbed photon makes an electron to be excited, elevating to a higher energy level. Subsequently, the electron returns to the ground state and emits a lower energy photon. Typically, the wavelength of emitted light is longer than that of absorbed light due to the energy loss involved in the vibrational state transition and the difference is called Stokes shift. This is the critical property that enables the fluorescence imaging to have high contrast and high sensitivity by completely filtering out the unwanted background light spectrally without affecting the signal emission light. Since the energy transition by absorption and emission can occur between multiple vibration states in each state, a fluorophore typically has a broadened spectrum of absorption and emission as shown in Fig. 1b. These spectral properties of a fluorophore are determined by its electronic energy structure that can be tuned to have many different emission wavelengths, that is, colors. During past decades, a variety of colorful fluorophores have developed in ultraviolet (UV), visible, and near-infrared (NIR) spectrum from 300 to 800 nm in wavelength. Fortunately, the emission spectrum of most fluorophores is spectrally narrow enough that it is possible to use multiple fluorophores with different colors at the same time as illustrated in Fig. 1c. This enables simultaneous visualization of multiple components tagged by different color fluorophores, which is extremely useful to analyze complex processes involving interaction between components, for example, interaction between a nanomaterial and a cell.

Fluorophores

Common fluorophores with different colors for multicolor imaging are summarized in Table 1. Fluorescein Isothiocyanate (FITC) and Tetramethylrhodamine (TAMRA) are traditional fluorescent dyes that have been successfully utilized for many years. The quantum yield of fluorescein is very high and its physical and chemical characteristics have been well known. But, fluorescein is very sensitive to environmental factors such as pH, changing its emission intensity. 4′,6-diamidino-2-phenylindole (DAPI) and Hoechst dyes are very popular to stain DNA and chromatin by its binding property to the double helix. And their blue fluorescence makes them to be easily combined with traditional dyes with green and red fluorescence. Alexa Fluor dyes, marketed by invitrogen™, have several improved features such as higher quantum yield than spectrally similar traditional dyes, enhanced photostability under continuous excitation and pH insensitivity. In addition, the absorption spectra of derivatives are well matched to common laser lines and water solubility is high. These advantages make the Alexa Fluor dyes to be one of the most popular dyes for many fluorescence imaging applications, but its relatively high price is the drawback. Cyanine dyes, Cy2, Cy3, Cy5, Cy5.5, and Cy7 are more environmentally stable and thus less sensitive to pH and organic media mounting samples. The cyanine dyes generally have broader absorption spectra than Alexa Fluor family, making them somewhat more versatile in the choice of laser excitation. As described before, due to the unique advantage of endogenous labeling, fluorescent proteins in various colors in blue (BFP), cyan (CFP), green (GFP), and red (RFP, DsRed) are widely used. Qdot®, marketed by invitrogen™, is



Intravital Microscopy Analysis, Fig. 1 Fundamentals of fluorescence and fluorophore. (**a**) Jablonski diagram depicting the electronic states of a fluorophore and energy transition involved in fluorescence. (**b**) Typical spectral properties of fluorophore, absorption, and emission spectra of Alexa Flour 555, a fluorophore marketed by invitrogen™, are illustrated. (**c**) Example of fluorophores for multicolor fluorescence imaging, Alexa Fluor 350, 488, and 594 from invitrogen™ is illustrated. Emission spectra of these fluorophores are largely nonoverlapping, which minimize the cross talk between them

Intravital Microscopy Analysis, Table 1 Common fluorophores for intravital microscopy

Color	Blue	Green	Orange	Red	NIR
Emission wavelength	450–480 nm	500–520 nm	570–600 nm	660–700 nm	760–850 nm
Fluorophore		GFP	DsRed	APC	
	CFP	YFP	RFP	Cy5	
	DAPI	FITC	TAMRA	Cy5.5	Cy 7
	Hoechst	Cy2	Cy3	Qdot 705	Qdot 800
		Alexa Flour 488	Alexa Flour 555	Alexa Flour 647	

a newly emerging semiconductor quantum dot with diameter of tens of nanometers. Unlike typical organic fluorophores that have highly specific spectral absorption profiles, quantum dots have a unique absorption property that absorption efficiency continuously increases with shorter wavelength, independent of emission wavelength. Thus, multiple kinds of quantum dots of different

colors can be excited by single UV light source. Another unique feature is that the emission wavelength of quantum dot is determined by the dot size, which enables the manufacturing of quantum dot with emission at NIR range. And the photostability of quantum dot is several orders of magnitude higher than organic dyes. These exceptional properties and the reduced scattering/absorption by most biological tissue in NIR range make the NIR quantum dot to have great potential for in vivo observation of living animal models, even for clinical applications such as fluorescence-guided surgery of the human patient [3]. To perform a fluorescence imaging with a certain fluorophore, it is crucial to select a light source whose output wavelength is closely matched to the peak wavelength of absorption spectra of fluorophore. Common light sources type and output wavelengths are summarized in Table 2.

Sectioning

Traditional fluorescence microscopy with simple configuration of Kohler illumination and wide-field detection can provide excellent images of thin transparent specimen, mostly sliced tissue sample after fixation or living cell cultured in vitro [1, 2]. However, for much thicker sample such as live tissue in mouse in vivo, it fails to reveal microscopic-scale details at the focus plane due to the reduced contrast by the fluorescence arise from fluorophores in out-of-focus plane, because the entire sample is excited indiscriminately. Thus, intravital microscopy must equip with a capability that can exclusively acquire fluorescence signal only from the focus by eliminating the out-of-focus fluorescence signal, which is termed as *sectioning* and this type of microscopy is categorized as *sectioning microscopy*. The removal of unnecessary fluorescence provides a tomographic snapshot of thick living sample with greatly enhanced contrast, which enables three-dimensional (3D) reconstruction by integrating the image data at multiple focal planes computationally.

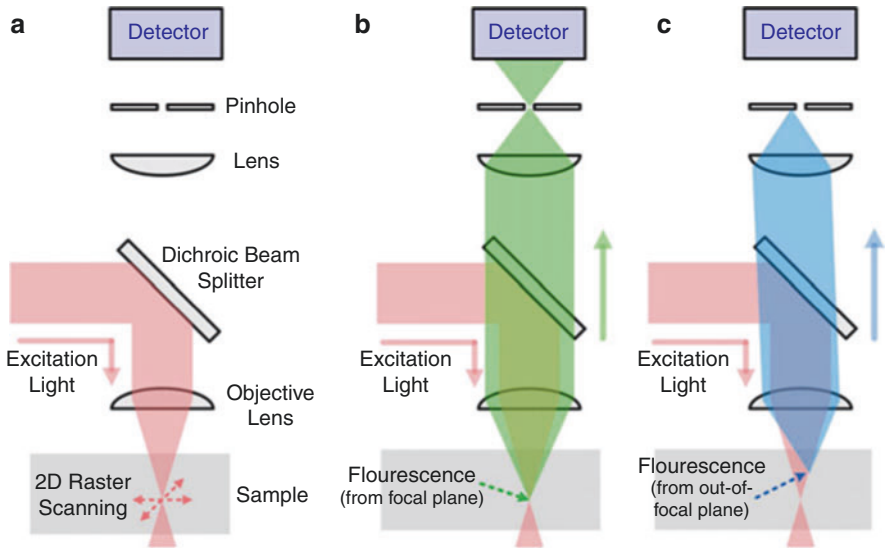
As a sectioning microscopy, there exist two major types of microscopies, *confocal microscopy* [3, 4] and *two-photon microscopy* [5, 6]. In both cases, the illumination of sample with the light source to excite a fluorophore is performed by two-dimensional (2D) raster-pattern scanning of tightly focused beam, which is termed as *scanning*. To achieve high resolution with *scanning*, it is crucial to implement light illumination that is focused to the smallest spot possible, the diffraction-limited spot at focus. This tight focusing is normally created by sending collimated beam to pass the high power, high numerical aperture (NA) objective lens that transforms the beam to converge. The higher the NA of objective lens and the shorter the wavelength of light for excitation, the smaller the spot becomes to the diffraction-limit that is calculated to 0.61 λ/NA in radius, where λ is the wavelength of the light. Although this diffraction-limited spot at the focal plane ensures the highly localized fluorescence signal generation for high-resolution information, the excitation of fluorophore happens nonselectively all along the converging and diverging beam path out of focal plane. Thus, it is necessary to implement rejection mechanism to block the fluorescence from out-of-focus plane. For example, in confocal microscopy, this is achieved by placing a pinhole in front of the detector at a plane optically conjugated to the focal plane such that pinhole aperture is aligned with the focus. This arrangement makes the only fluorescence arise from the focus exclusively pass the pinhole to reach the detector.

Confocal Microscopy

The operation principle of confocal microscopy for fluorescence imaging with sectioning is illustrated in Fig. 2. The arrangement of major components including pinhole, objective lens, sample, dichroic beam splitter, and detector as well as the path of excitation light are illustrated in Fig. 2a. The dichroic beam splitter is a specially engineered optical component that reflects shorter wavelength excitation light while transmitting longer wavelength fluorescence. Excitation light under 2D scanning is directed to the sample by the dichroic beam splitter and focused by the objective lens to create a tightly focused spot inside the sample. Figure 2b illustrates the light path of fluorescence that is emitted from the focal plane.

Intravital Microscopy Analysis, Table 2 Common light source and output wavelength for fluorescence detection

	UV	Violet	Blue	Green	Yellow	Red
Light source type	Output wavelength (nm)					
Argon-ion laser	351, 364	–	457, 488	514	–	–
Krypton-ion laser	–	413	–	531	569	647
Helium-neon laser	–	–	–	543	594	633
DPSS laser	365	405	488	532	561	641
Mercury-arc lamp	365	405, 436	–	546	579	–
Xenon-arc lamp	–	–	467	–	–	–
Metal-halide lamp	365	435	495	520, 545	575	625

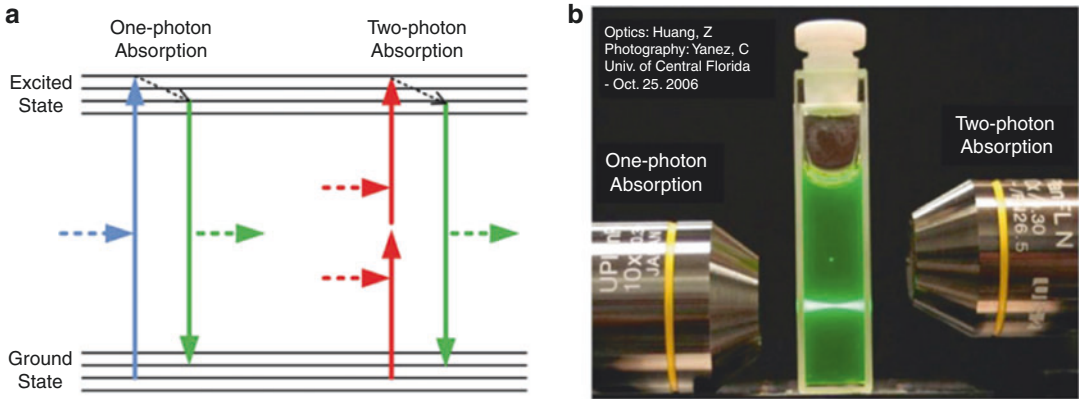


Intravital Microscopy Analysis, Fig. 2 Principle of confocal microscopy. (**a**) Schematic layout of confocal microscopy for fluorescence imaging with sectioning. (**b**) Light path of fluorescence emitted from the focus. (**c**) Light path of fluorescence emitted from the out-of-focus

This in-focus fluorescence signal is back-collected by the objective lens and then separated by the dichroic filter to transmit toward the detector. Then the fluorescence signal passes the pinhole of which the aperture is precisely aligned to the optical conjugation point with the focus of excitation light inside the sample and reaches the detector as illustrated in Fig. 2b. In contrast, the fluorescence emitted from the out-focus plane is focused onto the nontransparent part of the pinhole and thus fails to reach the detector as illustrated in Fig. 2c. The effectiveness of pinhole in terms of sectioning is greatly related to the size of the opening in the pinhole. Normally, for the observation of a thin specimen, it is recommended to match aperture size to the calculated spot size at focus, 0.61 λ/NA in radius (often referred to Airy Disk size). But, for in vivo observation of a thick living specimen, increasing the aperture size up to 2–4 Airy Disk size often provides a significantly brighter image with small penalty in resolution due to excessive light scattering.

Two-Photon Microscopy

On the other hand, *two-photon microscopy* achieves *sectioning* based on a nonlinear optical process between excitation light and fluorophore, two-photon absorption. It is a phenomenon that



Intravital Microscopy Analysis, Fig. 3 Principle of two-photon microscopy. (**a**) Jablonski diagram illustrating one- and two-photon absorption. (**b**) Photograph showing the spatial confinement of fluorescence generation with two-photon absorption, compared to one-photon absorption

occurs when two individual photons arrive at the same fluorophore simultaneously (within ~0.5 fs interval) and are absorbed together to combine their energy to excite the fluorophore. Since the energy of photon is inversely proportional to the wavelength, two absorbed photons substituting one photon of twice the energy should have twice the wavelength. For example, a fluorophore normally absorbing UV light at 350 nm can be excited by two-photon absorption with NIR light at 700 nm. Figure 3a illustrates the two-photon absorption along with a one-photon absorption. As depicted, once fluorophore is excited in either of absorption processes, it returns to the ground state in the exact same manner emitting a fluorescence photon. Due to the requirement that two photons hit the same fluorophore within an extremely short interval, compared to one-photon absorption, a light with about a million times higher intensity is required to generate the same level of fluorescence with two-photon absorption. In practice, to achieve this high intensity, the excitation light facilitating two-photon absorption has to be extensively concentrated not only in space by a high NA objective lens but also in time. Mode-locked Titanium-sapphire laser generating ultrashort pulse light of 10–100 fs pulse width at about 80 MHz repetition rate is the most common light source for two-photon microscopy. Even with this level of concentration in time domain, the high photon density sufficient for two-photon absorption is achieved only at the very small volume at the focus. In other words, the fluorescence excited by two-photon absorption is generated only from the focal plane as shown in Fig. 3b, which facilitates intrinsic *sectioning* without additional pinhole in front of detector like confocal microscopy. In Fig. 3b, a quartz cuvette with the solution of fluorophore, fluorescein having green fluorescence, is placed between two identical objective lenses. From the right-side objective lens ultrashort pulse light with 200 fs pulse width and 780 nm wavelength is delivered into the cuvette, demonstrating the highly confined generation of fluorescence signal by two-photon absorption. In contrast, when illumination light of 380 nm wavelength tuned for one-photon absorption is delivered through the left-side objective lens, the fluorescence is generated indiscriminately in whole illuminated volume. Table 3 summarizes the features of confocal and two-photon microscopy.

Laser Scanning

The most common method to implement a *scanning* for commercial confocal microscope is using two galvanometer-based oscillating mirrors to deflect the angle of reflected light. Typically, one mirror scans the excitation light along the "fast axis" (mostly horizontal direction) and the other

Intravital Microscopy Analysis, Table 3 Comparison between confocal and two-photon microscopy

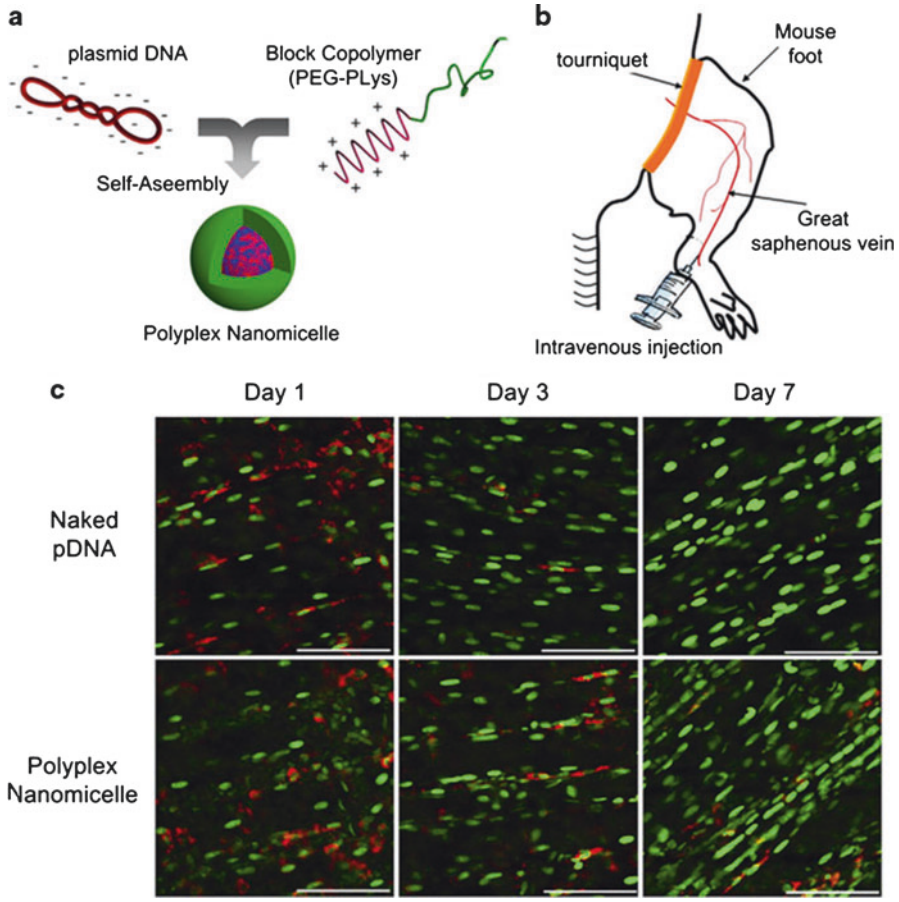
	Confocal microscopy	Two-photon microscopy
Sectioning principle	Pinhole rejecting out-of-focus fluorescence	Nonlinear fluorescence generation exclusively in focus
Light source type	Continuous-wave laser	Femtosecond pulsed laser
Light source wavelength	360–640 nm	700–1050 nm
Excitation process	Single-photon absorption	Two-photon absorption
Pinhole before detector	Yes	No
De-scanning	Yes	No

mirror scans the "slow axis" (mostly vertical direction) in much slower speed to sequentially cover the 2D plane. Naturally, the speed of image acquisition is limited by the maximum scanning speed of the fast-axis scanner. The typical scanning speed of the galvanometer-based oscillating mirror is from 500 Hz to 1 kHz, which corresponds to 1–2 frames per second with 512 × 512 pixels. However, for the in vivo observation such as intravital microscopy, it is often very useful to have a capability of video-rate image acquisition at 24–30 frames per second. Fast dynamics such as blood circulation, muscle reflex, and neural signal are several examples. One common method to achieve video-rate scanning is to use the resonant oscillating mirror that can scan at very high speed from 7 to 14 kHz by sacrificing scanning linearity. One commercial confocal microscope employing this resonant scanning is available in market, Nikon's A1R-A1 confocal microscope. Another method is using a polygonal mirror rotating at very high speed from 400 to 900 rev/s, which corresponds to 14.4–32.4 kHz with a 36-facet polygonal mirror. In practice, for in vivo observation, the image acquisition speed often needs to be compromised to obtain a high-quality image depending on the brightness of the fluorescence signal, since the achievable signal-to-noise ratio is proportional to the square root mean of exposure time per pixel with given signal intensity and frame rate.

Applications

In the last decade, with its unique capability of visualizing cellular dynamics in vivo, intravital microscopy has become an indispensable tool for many fields of biology, such as immunology [14], cancer biology [15], neuroscience [16], and neuropathology [17]. In the field of nanotechnology, there have been tremendous amount of efforts to develop new functionalized nanoparticles for biomedical applications such as gene or drug delivery, diagnosis, and therapeutics. In the process of translation from the chemical bench to the bed in clinic, extensive analysis using experimental animal, especially mouse, to assess the efficacy and safety in vivo is commonly required. The traditional microscopy-based analysis of ex vivo tissue sample provides only static information at a single given time of animal death, like a snapshot. The intravital microscopy-based analysis of live mouse can provide dynamic information over a period of hours to weeks in a single animal, which enables direct monitoring of various biological processes involving nanoparticles introduced into the animal in chronological view, like a movie. Here, two examples of intravital microscopy-based analysis of nanoparticle behavior in a live mouse will be introduced.

The first example is the in vivo tracking of polyplex nanomicelle administered for gene delivery to skeletal muscle over a week [7]. Skeletal muscle has been studied as an ideal target of gene delivery for therapeutic protein production, such as growth factor [8]. Since direct injection of naked plasmid DNA (pDNA) into muscle suffers very low efficacy, a virus-based gene delivery system, viral vector, has been developed. However, several fundamental limitations of utilizing the virus as a gene carrier such as the size limitation of integrated gene, the difficulty in virus manufacture, and most critically, the immunogenicity driven by virus have limited its clinical application. To overcome these limitations, a



Intravital Microscopy Analysis, Fig. 4 In vivo tracking of polyplex nanomicelle hydrodynamically administered for gene delivery to thigh muscle over a week in live mice. (**a**) Polyplex nanomicelle composed of plasmid DNA and block copolymer. (**b**) Hydrodynamic gene delivery to the thigh muscle by intravenous injection and tourniquet. (**c**) Longitudinal in vivo tracking of Cy5-labeled pDNA (*red*) distributed in the muscle fibers of Histone-GFP mice, in which every nucleus of cells expresses GFP (*green*). Scale bars are 100 μm (Modified from Ref. [7])

nonviral gene carrier, such as nanomicelle and porous nanoparticle, has been actively developed [9].

In this example, the polyplex nanomicelle is synthesized by a self-assembly process in the mixture of pDNA and block copolymer as illustrated in Fig. 4a. To image the pDNA in vivo, it is fluorescently labeled by red fluorophore, Cy5. As an experimental animal model to monitor the delivery of Cy5-labeled pDNA in vivo, a transgenic Histone-GFP fusion mouse (B6.Cg-Tg (HIST1H2BB/EGFP) 1 Pa/J, Jackson Lab) expressing GFP in the nucleus of every cell is used. To administer the nanomicelle safely, a Histone-GFP mouse is anesthetized. The blood circulation of the right foot is isolated by placing tourniquet as illustrated in Fig. 4b. Then the polyplex nanomicelle containing Cy5-labeled pDNA solution (300 μl solution with 50 μg pDNA) is injected intravenously through the Great saphenous vein in 5 s. Due to the tourniquet limiting the blood circulation, the hydrostatic pressure in the right foot is greatly increased and enhances the gene delivery. After 5 min, the tourniquet is released. For comparison, the same amount of naked pDNA is injected into another mouse as a control.

The gene delivery efficacy of naked pDNA and polyplex nanomicelle in terms of distribution in the

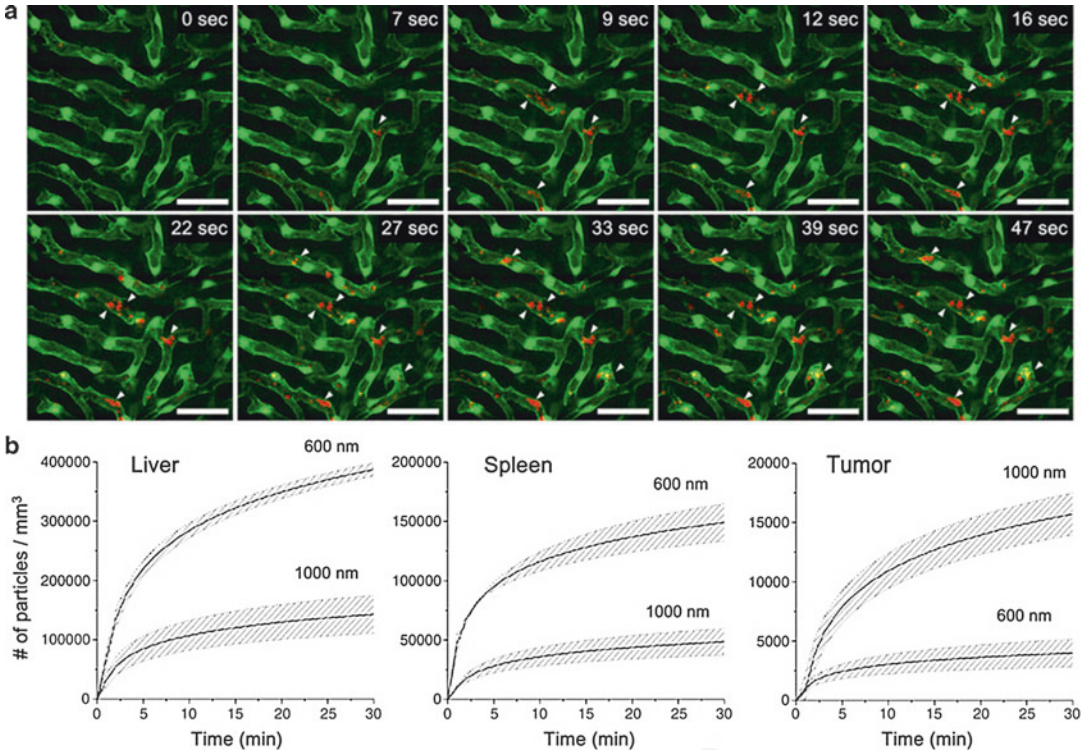
muscle and retention over time is monitored by intravital microscopy analysis using a custom-built video-rate confocal microscopy system [10, 11]. To evaluate the time-dependent change, each mouse is imaged at multiple time points of days 1, 3, and 7. In each in vivo imaging session, the mouse is anesthetized and placed on the heated plate of motorized XYZ stage. The thigh muscle is carefully exposed by skin incision without damaging fascia. GFP and Cy5 are excited by 491 and 635 nm continuous-wave DPSS laser, respectively, and detected by a photomultiplier tube through a 520 ± 17 nm and a 670 ± 17 nm band-pass filter, respectively. After the imaging, the skin incision is closed by 6–0 polyethylene suture for the follow-up imaging at later time point and triple antibiotic ointment is applied to avoid infection.

Figure 4c shows the representative images at each time point. On day 1, for both of naked pDNA and nanomicelle, the Cy5-labeled pDNA (red) is widely distributed, revealing no difference. With image analysis to identify the co-localization of GFP and Cy5 fluorescence, the ratio of nuclei with delivered pDNA can be assessed. For naked pDNA, the Cy5-labeled pDNA is significantly decreased on day 3 and completely disappeared on day 7. In contrast, for the nanomicelle, a much larger amount of pDNA is detectable until day 7. This intravital microscopy analysis clearly demonstrates the greatly improved gene delivery efficacy of nanomicelle over naked pDNA.

The second example is the in vivo tracking of systemically injected porous silicon nanoparticles at reticuloendothelial system (RES) organs – liver and spleen [12]. Drug encapsulating nanoparticles have been extensively developed to improve the pharmacokinetic features, safety, and selective accumulation to target site, compared to freely administered molecule. For anticancer treatment, the common strategy to maximize the accumulation of nanoparticles loaded with chemotherapeutic agents to the tumor relies on the enhanced permeation and retention (EPR) effect [13]. However, unfortunately, the RES in liver and spleen that works like a net catching blood circulating foreign materials such as bacteria take up the most of injected nanoparticle. As part of the immune system fighting against the environment, the RES functions so efficiently that it has been a major hurdle ultimately limiting the accumulation of nanoparticles to the tumor site, which rarely exceeds 1 % ID/g.

In this example, the accumulation of porous silicon nanoparticles in two major RES organs of liver and spleen is monitored in real time in vivo by intravital microscopy analysis [12]. As described in the previous example, a custom-built video-rate confocal microscopy system is used [10, 11]. Figure 5a is the image sequence taken at the liver from the very moment of intravenous injection of spherical-shaped porous silicon nanoparticles with diameter of 500 nm. To track the individual nanoparticles, it is fluorescently labeled by Alexa Flour 555. To visualize the liver sinusoid simultaneously, a transgenic Tie2-GFP mouse (STOCK Tg(TIE2GFP) 287Sato/J, Jackson Lab) expressing GFP under the direction of endothelial-specific receptor promoter Tie2 was used. To access the liver with high NA objective lens, a small incision is made on the abdominal skin and peritoneum of anesthetized Tie2-GFP mouse. Then the nanoparticle is injected into the blood stream by retro-orbital injection and image acquisition is started at the same time. After 7 s, the injected nanoparticle starts to appear in the liver, then rapidly accumulates as shown in Fig. 5a. Interestingly, a few particles initially attach to an endothelial cell of liver sinusoid, then they quickly attract many particles to aggregate as marked by arrowhead. This high spatial and temporal analysis based on intravital microscopy suggests that inhibiting the initial attachment of the nanoparticle to the endothelial cell of RES organs like the liver or minimizing the aggregation of the nanoparticle can be a good strategy to avoid these unwanted accumulations.

One approach to modulate the characteristics of nanoparticles is to modify their shape from spherical to anisotropic such as the plateloid (disk) or cylindroid (rod) shape as well as their size. As an example, the accumulation of two



Intravital Microscopy Analysis, Fig. 5 In vivo real time monitoring of time-dependent accumulation of systemically administered porous silicon nanoparticle in live mice. (**a**) Image sequence taken at the liver from the very moment of intravenous injection of the nanoparticle (*red*). Time-dependent accumulation of nanoparticles is clearly visualized in high resolution. Liver sinusoid (*green*) is simultaneously visualized by GFP expression under the direction of endothelial-specific receptor Tie2. Scale bars are 50 μm. (**b**) Time-dependent accumulation of plateloid shape porous silicon nanoparticles with two different sizes of 600 × 200 and 1000 × 400 nm in the liver, spleen, and melanoma tumor. Data is plotted as a mean curve ± 1 SD ($n = 4$ animal) (Modified from Ref. [12])

plateloid-shaped nanoparticles with different sizes of 600 × 200 and 1000 × 400 nm in the liver, spleen, and tumor is monitored by similar intravital microscopy analysis. Tumor is generated by subcutaneous injection of 10^6 mouse melanoma cells (B16F10) at the flank of live Tie2-GFP mouse, then left for 1 week to allow the tumor to grow in 3–5 mm size. Figure 5b shows the time-dependent accumulation of plateloid-shaped nanoparticles over 30 min from the systemic administration by intravenous injection. For the first 5 min, the number of accumulated nanoparticles rapidly increases and then slows down (600 nm) or begins to level off (1000 nm) in all three sites. In the RES organs, liver, and spleen, the 600 nm particles accumulate much faster than the 1000 nm particles in about three times. Interestingly, this trend becomes opposite in the tumor that the 1000 nm particles accumulate in about four times faster, later confirmed by traditional ex vivo analysis like ICP-AES or histological observation.

Perspectives

Intravital microscopy to visualize a dynamic cellular process in a living animal with submicron resolution can provide valuable new insights that are unobtainable by other analyses. During last

decade, this unique capability makes intravital microscopy to be one of most popular tool for many fields of biology, such as immunology [14], cancer biology [15], neuroscience [16], and neuropathology [17]. Naturally, in the field of nanotechnology aiming biomedical application involving living organism, it is clear that intravital microscopy analysis will emerge as a new powerful tool. As illustrated in the examples, for the fields of nanomedicine and nanoparticles for drug delivery, it is demonstrated that intravital microscopy analysis can provide new insights to improve the current understanding about interactions of nanomaterial with cellular compartments in live animal. On the other hand, for the field of intravital microscopy, nonorganic optical probes based on nanotechnology such as fluorescent quantum dot and nonlinear optical signal generating nanoparticles emerges as a novel contrast agent with highly desirable features like enhanced photostability and near-infrared emission. Combined with rapid technical advances in laser source, optical detector, and micro-optics, these new nanoprobes can greatly expand the applications of intravital microscopy.

Cross-References

- ▶ Confocal Laser Scanning Microscopy
- ▶ Nanomedicine
- ▶ Nanoparticles

References

1. Diaspro, A.: Optical Fluorescence Microscopy. Springer, New York (2010)
2. Lichtman, J.W., Conchello, J.-A.: Fluorescence microscopy. Nat. Methods **2**, 910–919 (2005)
3. Pawley, J.: Handbook of Biological Confocal Microscopy. Springer, New York (2006)
4. Conchello, J.-A., Lichtman, J.W.: Optical sectioning microscopy. Nat. Methods **2**, 920–931 (2005)
5. Masters, B.R., So, P.T.C.: Handbook of Biomedical Nonlinear Optical Microscopy. Oxford University Press, New Work (2008)
6. Helmchen, F., Denk, W.: Deep tissue two-photon microscopy. Nat. Methods **2**, 932–940 (2005)
7. Itaka, K., Osada, K., Morii, K., Kim, P., Yun, S.-H., Kataoka, K.: Polyplex nanomicelle promotes hydrodynamic gene introduction to skeletal muscle. J. Control. Release **143**, 112–129 (2010)
8. Blau, H.M., Springer, M.L.: Muscle-mediated gene therapy. N. Engl. J. Med. **333**, 1554–1556 (1995)
9. Lu, Q.L., Bou-Gharios, G., Partridge, T.A.: Non-viral gene delivery in skeletal muscle: a protein factory. Gene Ther. **10**, 131–143 (2003)
10. Kim, P., Puoris'haag, M., Cote, D., Lin, C.P., Yun, S.-H.: In vivo confocal and multiphoton microendoscopy. J. Biomed. Opt. **13**, 010501 (2008)
11. Kim, P., Chung, E., Yamashita, H., Hung, K.E., Mizoguchi, A., Kucherlapati, R., Fukumura, D., Jain, R.K., Yun, S.-H.: In vivo wide-area cellular imaging by side-view endomicroscopy. Nat. Methods **7**, 303–305 (2010)
12. van de Ven, A.L., Kim, P., Haley, O., Fakhoury, J.R., Adriani, G., Schmulen, J., Moloney, P., Hussain, F., Ferrari, M., Liu, X., Yun, S.-H., Decuzzi, P.: Rapid tumoritropic accumulation of systemically injected plateloid particles and their biodistribution. J. Control. Release (2012, in press)
13. Maeda, H., Wu, J., Sawa, T., Matsumura, Y., Hori, K.: Tumor vascular permeability and the EPR effect in macromolecular therapeutics: a review. J. Control. Release **65**, 271–284 (2000)
14. Germain, R.N., Miller, M.J., Dustin, M.L., Nussenzweig, M.C.: Dynamic imaging of the immune system: progress, pitfalls and promise. Nat. Rev. Immunol. **6**, 497–507 (2006)
15. Jain, R.K., Munn, L.L., Fukumura, D.: Dissecting tumor pathophysiology using intravital microscopy. Nat. Rev. Cancer **2**, 266–276 (2002)
16. Kerr, J.N.D., Denk, W.: Imaging in vivo: watching the brain in action. Nat. Rev. Neurosci. **9**, 195–205 (2008)
17. Misgeld, T., Kerschensteiner, M.: In vivo imaging of the diseased nervous system. Nat. Rev. Neurosci. **7**, 449–463 (2006)

Ion Beam

▶ Physical Vapor Deposition

Ion Beam Etching (IBE)

▶ Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Ion Beam Figuring (IBF)

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Ion Beam Smoothing (IBS)

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Iridescent Colors (Organisms Other Than Animals)

► Nanostructures for Coloration (Organisms Other Than Animals)

Isothermal Calorimetry

► Nanocalorimetry

K

Kelvin Probe Force Microscopy

Bharat Bhushan and Manuel L. B. Palacio
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Scanning Kelvin probe force microscopy; Scanning surface potential microscopy

Definition

Kelvin probe force microscopy (KPFM) is an atomic force microscopy (AFM)-based technique used to measure the surface potential and work function properties of a material surface interacting with a conductive AFM probe.

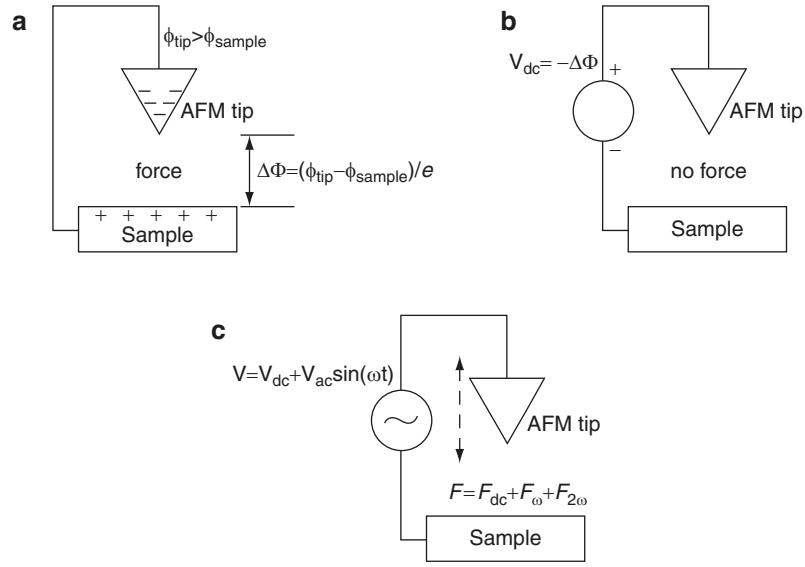
Overview

The applications of atomic force microscopy (AFM) have continuously evolved to include the measurement of properties aside from surface topography. To date, the AFM is widely used for the characterization of surface/near-surface properties such as friction, adhesion, scratch, wear, fabrication/machining, electrical properties, local deformation, nanoindentation, local surface elasticity, viscoelasticity, and boundary lubrication.

The mapping of surface or near-surface electrical properties using the AFM is possible through the use of an electrically conducting tip, which acts as a movable top electrode that can locally set a potential difference between the tip and the sample of interest. As the cantilever tip is scanned above the sample, the tip can sense the response to the electrical potential. An advantage of conducting electrical AFM is that the topography and electrical properties can be obtained simultaneously during the experiment [2].

The contact, intermittent (tapping) and noncontact operating modes, along with the use of a conductive tip, extended the application of AFM from topographic imaging to electrical surface characterization [2–4, 6]. The ability to detect the surface potential difference between the tip and sample, with the tip lifted a certain distance above the sample, has led to Kelvin Probe Force Microscopy (KPFM). The ability for capacitance measurements, such as in *n*-type silicon, led to Scanning Capacitance Microscopy (SCM). Deflection detection for electromechanical response of ferroelectric films led to Piezoresponse Force Microscopy (PFM). The detection of current through oxide films led to Conductive AFM (CAFM) and on doped samples as a function of the load led to Scanning Spreading Resistance Microscopy (SSRM). The studies of current flowing into constrictions in

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1



Kelvin Probe Force Microscopy,
Fig. 1 Schematic of the KPFM principle: (**a**) electrostatic potential and interaction force between a conducting tip and a sample (for illustration, assume that $\varphi_{tip} > \varphi_{sample}$), (**b**) external DC voltage applied to nullify the force, and (**c**) external AC voltage with adjustable DC offset is applied to the tip which leads to its vibration [5]

two-dimensional semiconductor structures led to Scanning Gate Microscopy (SGM). The detection of frequency dependence force applying a voltage across a sample in $BiFeO_3$ and *p*-doped silicon led to Scanning Impedance Microscopy (SIM). A recently reported new technique to map capacitance, impedance, and dielectric constant variations is called Scanning Microwave Microscopy (SMM). This technique studies the electromagnetic tip–sample interactions using a specially designed conductive probe to send and receive microwave signals up to GHz frequencies.

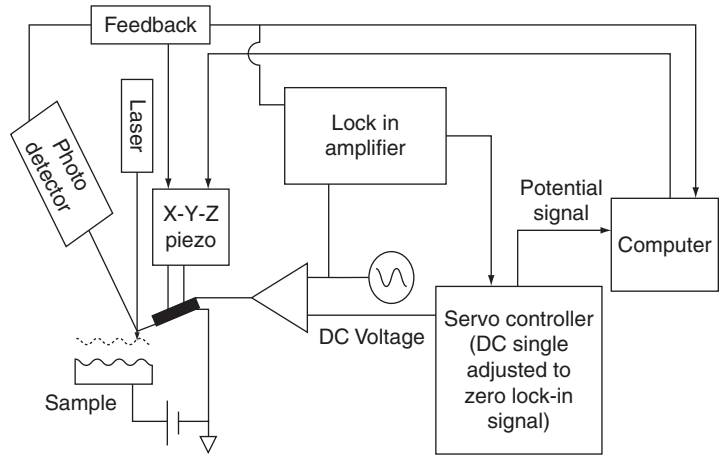
The focus of this entry is on Kelvin probe force microscopy. Since the surface potential measurements are highly sensitive to surface charge variations, KPFM is widely used in detecting the dopant concentration and the dopant type (either *p*-type or *n*-type) on semiconductor materials. Non-semiconductor applications of KPFM include the characterization of adsorbates on Si substrates, intermetallic grain formation on metal films, nanotribology studies, and the degradation of inorganic materials for batteries, among others.

Kelvin Probe Force Microscopy Principles

Kelvin probe force microscopy is used to measure the surface potential with the AFM in noncontact mode, and is based on the Kelvin probe principle [3, 4, 10]. To describe the operating principle behind KPFM, consider a tip and sample interaction as shown in Fig. 1. When the tip and sample are electrically connected (Fig. 1a), electrons flow from the material with the lower work function to the material with the higher work function. Due to the difference in the work function of the electrically connected tip and the sample, an electrostatic contact potential difference (or surface potential) is created between the tip and the sample. The value of this contact potential difference ($\Delta\Phi$) is given by the following equation:

$$\Delta\Phi = \frac{\Phi_{tip} - \Phi_{sample}}{e} \tag{1}$$

where Φ_{tip} and Φ_{sample} are work functions of the tip and the sample, respectively, and e is the magnitude of the charge of one electron. The contact potential difference provides material contrast information, because it is sensitive to adsorbed molecules on the surface or the presence of various phases within a given material. Electrostatic force is created between the tip and sample under the influence of this surface potential difference and the separation-dependent local capacitance C of the tip and sample. This force is given by



Kelvin Probe Force Microscopy,
Fig. 2 Schematic of the interleave scan method used in KPFM, where the first pass records the height profile performed in tapping mode, and the second pass in lift mode monitors the variations of electrostatic force at a constant height [12]

$$F = \frac{1}{2}(\Delta\Phi)^2 \frac{\partial C}{\partial z} \tag{2}$$

where z is the distance between the tip and sample.

In KPFM, an AC voltage signal ($V_{AC}\sin(\omega t)$) with an adjustable offset DC voltage (V_{DC}) is applied to the conducting tip (Fig. 1b and c). As shown in (Eq. 3), the electrostatic force is the sum of three components, namely, the DC force, a harmonic force with the same frequency of the AC tip bias (the first harmonic), and a harmonic force at double the frequency of the AC signal (the second harmonic):

$$\begin{aligned} F &= \underbrace{\frac{1}{2}\frac{\partial C}{\partial z}\left\{(\nabla\Phi + V_{DC})^2 + \frac{{V_{AC}}^2}{2}\right\}}_{\text{DC term}} \\ &+ \underbrace{\frac{\partial C}{\partial z}(\Delta\Phi + V_{DC})V_{AC}\sin(\omega t)}_{\omega \text{ term}} F \\ &= \underbrace{\frac{1}{4}\frac{\partial C}{\partial z}{V_{AC}}^2\cos(2\omega t)}_{2\omega \text{ term}} \end{aligned} \tag{3}$$

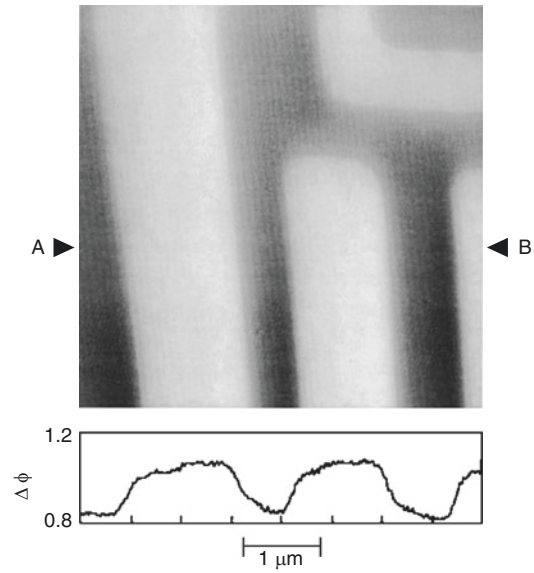
The electrostatic force is used to adjust the externally applied voltage in order to compensate the contact potential until the force between the tip and the sample is nulled. This is attained when the difference between the work functions of the tip and sample is equal to the external voltage. The presence of different materials on the sample surface leads to the formation of different material interfaces (with the tip), each having their own value of the contact potential difference. This leads to surface potential contrast [9].

In the initial implementation of KPFM, the topography and potential are measured simultaneously using two feedback circuits [13]. Subsequent modifications to the setup involve the sequential measurement of topography and potential in the so-called lift mode, which is regarded to improve on the lateral resolution of the topographic image [9]. A schematic of the two-pass lift or interleave scan method used in KPFM experiments is shown in Fig. 2. On the first pass, surface height information is recorded in the standard tapping mode where the oscillating cantilever lightly taps the surface. On the second pass, the feedback loop controlling the vertical piezo is turned off, and the tip is lifted from the surface and traced following the surface profile acquired in the first pass and at constant tip–sample separation distance [7, 8]. An oscillating potential is applied to the DC-biased tip (the tip oscillation is electrically driven). The oscillating potential frequency is equal to the cantilever's resonance frequency. An extra feedback loop is used to adjust the DC bias on the tip to nullify the first harmonic component of the electrostatic force [5, 9].

A more recent implementation of the KPFM experiment involves the detection of the surface potential using a frequency modulation technique instead of the amplitude modulation method. The main difference between these two methods is that amplitude modulation is sensitive to the

electrostatic force, while frequency modulation is sensitive to its gradient. Instead of having one feedback control/lock-in amplifier system that sequentially records the topography and surface potential, a separate lock-in amplifier is used in the frequency modulation KPFM experiment to detect surface potential. The frequency modulation method has been shown to measure a potential difference in close agreement with macroscopic measurements, and is thus regarded as a more quantitative potential measurement technique compared to amplitude modulation [16].

As mentioned earlier, KPFM is widely used in characterizing the doping of semiconductors, and it can also be utilized for non-semiconductor materials. Examples of KPFM applications are described next.



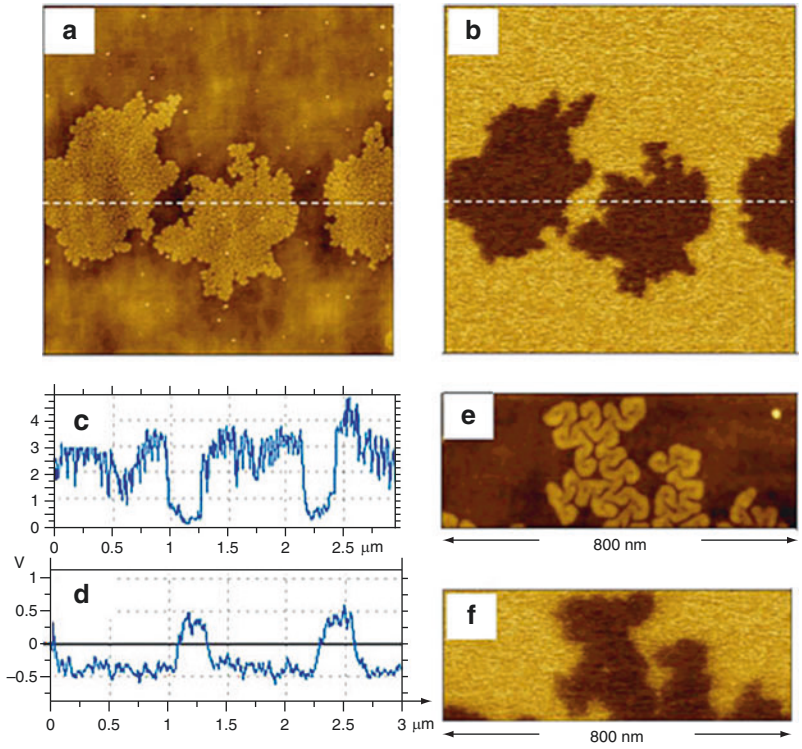
Kelvin Probe Force Microscopy, Fig. 3 Surface potential image of a silicon *pn* junction (*above*), with the *bright* and *dark areas* corresponding to the *p*-type and *n*-type regions, respectively. The cross section (*below*) refers to the scan defined by the line A-B on the image, where Δφ is the surface potential relative to the probe surface potential [11]

Applications of KPFM

An example of the applicability of KPFM to semiconductor materials is shown in Fig. 3, which is a surface potential image of a *pn* junction on a silicon wafer created through ion implantation. Below the image is a corresponding cross section of the scan defined by the line A-B on the image [11]. The contrast in the surface potential image is induced by the charge carriers introduced by the ion implantation process. The brighter features correspond to areas with a higher surface potential. In this case, the surface potential in the *p*-type region is higher than in the *n*-type region. This is attributed to the bending of energy bands at the subsurface as a result of the trapping of charge carriers in surface states [11].

KPFM has been used in the characterization of perfluoroalkyl alkanes self-assembled on substrates such as graphite, mica, and silicon. As an example, Fig. 4 shows images of $F_{14}H_{20}$ adsorbates on a Si substrate [1]. Topography and surface potential images of the adsorbate layers show clusters in Fig. 4a and b. In the surface potential image, the difference between the adsorbates and silicon shows a high contrast. In Fig. 4c and d, the profiles of the white lines marked in Fig. 4a and b are displayed. The profiles indicate that there is a large surface potential drop of ± 0.5 V within the adsorbate clusters as a result of possible repulsion of the adsorbate composition. A correlation between the surface potential and the topography image suggests that the surface potential corresponds to the adsorbate clusters. The magnified images in Fig. 4e and f also revealed a particular shape in the cluster, spiral structures. In the example KPFM images, the capacity of the technique to detect different surface potentials on the sample is shown.

Figure 5 shows a KPFM application to recognize constituent materials in a metal alloy film [14]. In this example, an alloy consisting of Al and Cu were deposited as 5 μm lines on a silicon substrate. The lines in the contact potential difference image in Fig. 5a contain features that are not present in the topographic image shown in Fig. 5b. In the potential image, the dark features contained within the lines correspond to the Al/Cu intermetallic that forms on the grain boundaries, which has a lower contact potential difference relative to the surrounding Al regions. The clear



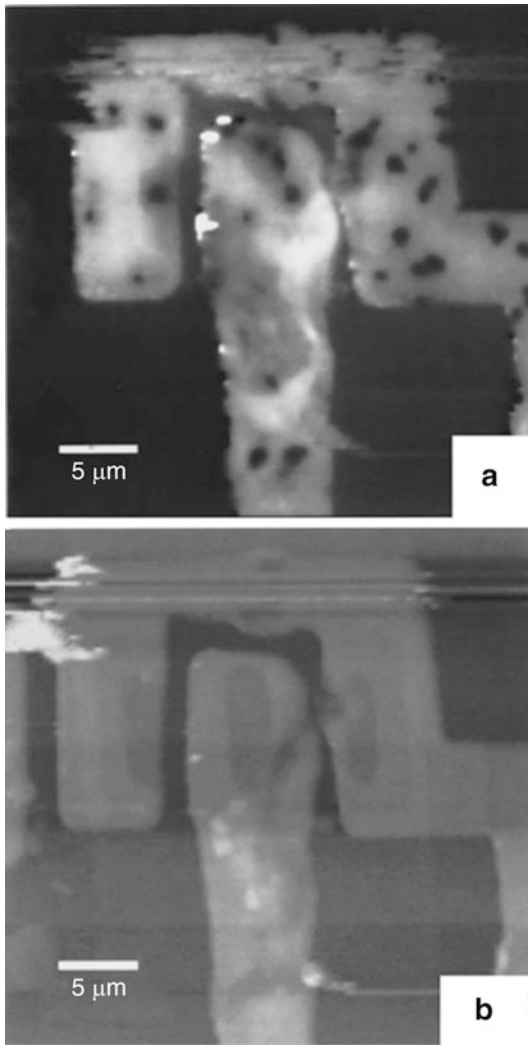
Kelvin Probe Force Microscopy,
Fig. 4 KPFM applied to studies of $F_{14}H_{20}$ adsorbates on Si: (**a**) topography, (**b**) surface potential image of $F_{14}H_{20}$ adsorbates on Si, (**c**) and (**d**) the cross-section profiles taken along the directions marked with the *dashed white lines* in the images in (**a**) and (**b**), respectively, and (**e**) and (**f**) zoomed-in topography and surface potential images on one of the locations shown in (**a**) and (**b**), respectively [1]

contrast between the Al/Cu and Al shows the capacity of KPFM as a technique to identify the presence of different materials on alloy surfaces at the nanoscale.

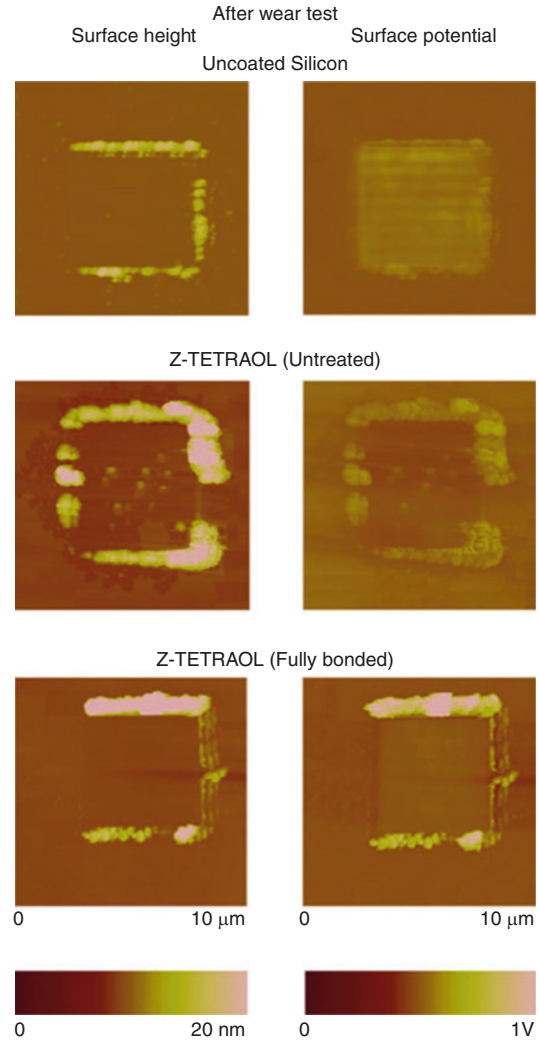
KPFM can also be applied to nanotribological studies, specifically for studies involving wear initiation. In Fig. 6, surface height and surface potential data (taken after nanoscale wear tests were performed) are shown for silicon surfaces coated with the perfluoropolyether lubricant Z-TETRAOL. The lubricant is applied using two different techniques, such that the samples are either untreated or fully bonded. Data for the uncoated silicon surface subjected to the same wear test procedure is shown for reference [15].

The surface height and surface potential images in Fig. 6 show contrast on both sets of images. Compared to the uncoated silicon surface, reduction in the debris is observed in the partially bonded samples. This indicates that the lubricant is able to provide wear protection. The pileup of debris observed in the untreated sample could be a combination of removed fragments of Si as well as the weakly bound lubricant. The untreated sample with its mobile lubricant fraction exhibits a lower surface potential change compared to the fully bonded sample, which only has immobile lubricant molecules. This is attributed to lubricant replenishment by the mobile lubricant molecules present on the untreated sample. In general, a considerable surface potential change is observed on the wear region when: (1) the lubricant has been fully removed from the substrate; (2) the native SiO_2 layer has been abraded from the surface; (3) wear has caused subsurface structural changes; and (4) charges build up as the charges are unable to dissipate into the surrounding material [15].

Finally, the application of KPFM in studying material degradation is illustrated for the aging of lithium-ion batteries. The surface potential maps of the unaged and aged $LiFePO_4$ cathode samples are shown in Fig. 7 [12]. The maps were collected by applying +1.0 V from an external DC voltage source. On the surface potential maps for both the unaged and aged sample, a small difference in the contrast is observed, suggesting an almost uniform dissipation of charge over the surface of the samples under the externally applied voltage. This



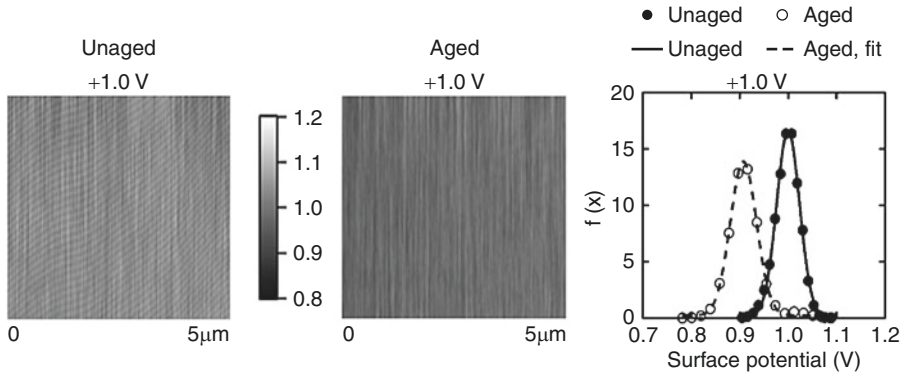
Kelvin Probe Force Microscopy, Fig. 5 KPFM study of various intermetallics in a metal alloy: (**a**) surface potential image of Al/Cu film showing contrast between Al and the Al/Cu intermetallic (*the dark points*); the scan field is 40 × 40 μm, (**b**) AFM topography image [14]



Kelvin Probe Force Microscopy, Fig. 6 Surface height and surface potential maps after wear tests for an uncoated silicon wafer and two types of Z-TETRAOL-lubricated silicon wafers (untreated and fully bonded). The *brighter areas* correspond to higher values of both the height and surface potential

indicates that under the external source the sample tends to charge uniformly even in an aged condition. The uniform charging is good for the cell as it avoids large local currents and subsequent damage to the cathode. The last column of Fig. 7 shows the distribution of the collected data points in unaged and aged cathodes. Then a normal probability density function is used to fit the histogram data. The unaged sample had mean values of the surface potential almost equivalent to the external applied voltage. The mean values of surface potential on the aged sample are lower than that of the unaged sample. Thus, even though the externally applied voltage was the same for the unaged and aged samples, the charge sustained on the surface of the aged cathode is less than that sustained on the unaged cathode [12].

The work function of the tip is constant in each surface potential map, since the same kind of tip is used in both samples. Therefore, Fig. 7 shows the



Kelvin Probe Force Microscopy, Fig. 7 Surface potential map of the unaged (*left column*) and the aged (*middle column*) $LiFePO_4$ cathode samples with external voltage of + 1 V. The *right column* shows the normal probability density distribution of the surface potential for the unaged and aged samples, indicating that the mean value of the surface potential decreases after aging

change in the work function of the aged sample relative to the unaged sample. The decreased surface potential in the aged sample is an indication of surface modification as a result of the phase change of the cathode material upon degradation [12]).

Cross-References

- ▶ AFM Probes
- ▶ AFM, Tapping Mode
- ▶ Atomic Force Microscopy

References

1. Alexander, J., Magonova, S., Moeller, M.: Topography and surface potential in Kelvin force microscopy of perfluoroalkyl alkanes self-assemblies. J. Vac. Sci. Technol. B **27**, 903–911 (2009)
2. Avila, A., Bhushan, B.: Electrical measurement techniques in atomic force microscopy. Crit. Rev. Solid State Mater. Sci. **35**, 38–51 (2010)
3. Bhushan, B.: Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
4. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
5. Bhushan, B., Goldade, A.V.: Measurements and analysis of surface potential change during wear of single-crystal silicon (100) at ultralow loads using Kelvin probe microscopy. Appl. Surf. Sci. **157**(2000), 373–381 (2000)
6. Bonnell, D.: Scanning Probe Microscopy and Spectroscopy Theory, Techniques, and Applications, 2nd edn. Wiley, New York (2001)
7. DeVecchio, D., Bhushan, B.: Use of a nanoscale Kelvin probe for detecting wear precursors. Rev. Sci. Instrum. **69**, 3618–3624 (1998)
8. Elings, V., Gurley, J.: US Patent No. 5,266,801, (1993)
9. Jacobs, H.O., Knapp, H.F., Muller, S., Stemmer, A.: Surface potential mapping: a qualitative material contrast in SPM. Ultramicroscopy **69**, 39–49 (1997)
10. Kelvin, L.: Contact electricity of metals. Philos. Mag. **46**, 82–120 (1898)
11. Kikukawa, A., Hosaka, S., Imura, R.: Silicon *pn* junction imaging and characterizations using sensitivity enhanced Kelvin prone force microscopy. Appl. Phys. Lett. **66**, 3510–3512 (1995)
12. Nagpure, S.C., Bhushan, B., Babu, S.S.: Surface potential measurement of aged Li-ion batteries using Kelvin probe microscopy. J. Power Sources **196**, 1508–1512 (2011)
13. Nonnenmacher, M., O'Boyle, M.P., Wickramasinghe, H.K.: Kelvin probe force microscopy. Appl. Phys. Lett. **58**, 2921–2923 (1991)
14. O'Boyle, M.P., Hwang, T.T., Wickramasinghe, H.K.: Atomic force microscopy for work functions on the nanometer scale. Appl. Phys. Lett. **74**, 2641–2642 (1999)
15. Palacio, M., Bhushan, B.: Ultrathin wear-resistant ionic liquid films for novel MEMS/NEMS applications. Adv. Mater. **20**, 1194–1198 (2008)
16. Zerweck, U., Loppacher, C., Otto, T., Grafstroem, S., Eng, L.M.: Accuracy and resolution limits of Kelvin probe force microscopy. Phys. Rev. B **71**, 125424 (2005)

Key Point Detection

- ▶ Waypoint Detection

L

Lab-on-a-Chip

► Electrical Impedance Tomography for Single-Cell Imaging
► Micro- and Nanofluidic Devices for Medical Diagnostics

Lab-on-a-Chip (LOC)

► Lab-on-a-Chip for Studies in *C. elegans*

Lab-on-a-Chip for Studies in *C. elegans*

Nuria Vergara-Irigaray and Michèle Riesen
Department of Genetics Evolution and Environment, Institute of Healthy Ageing, University College London, London, UK

Synonyms

Lab-on-a-chip (LOC); Microfluidic device; Microsystem; Microtrap; Worm chip

Definitions

Lab-on-a-chip (LOC) is a general term used to describe microdevices that integrate and scale down single or multiple laboratory functions and processes to a miniaturized chip format. The main characteristic of LOC devices is that they are able to handle extremely small fluid volumes down to less than pico-liters. This feature enables the analysis of very small samples in short time. LOC devices have recently been developed as micro-engineering approaches to small-size animal model research offering several advantages over conventional approaches, such as control of the microenvironment, automation, and high-throughput analysis.

Caenorhabditis elegans (*C. elegans*) is a small, transparent, nonparasitic nematode that is used as a model organism for the study of multiple biological processes such as development, genetics, and neurobiology. This is mainly due to the fact that this animal is small in size, has a rapid life cycle, and it is easy to manipulate and cultivate in controlled laboratory conditions. In addition, there are many human disease models available in this model organism providing an in vivo environment for the study of these diseases over other approaches such as cell culture.

Overview

The possibility to perform laboratory operations on a small scale using miniaturized lab-on-a-chip devices has become in recent years a real tool for which biologists show an increasing interest. The use of small volume samples leads to a reduction

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

in the time taken to perform an experiment and analyze the result; it also reduces the cost of the analysis. Together with the ability to automate sample manipulation and analysis, the use of LOC devices brings with it the potential to analyze very large populations of animals and to perform high-throughput analysis, which could not be done prior to the development of these new techniques.

C. elegans has been used as a model organism for the last 40 years. The multiple characteristics that make this nematode an excellent model organism for the study of a wide range of biological processes will be discussed further in this text.

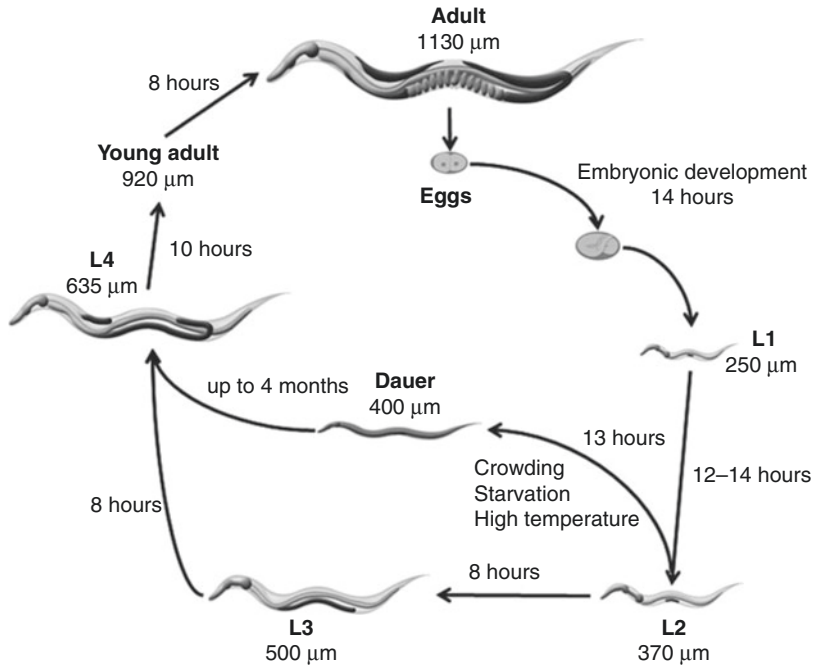
In this entry, an introduction to the use of *C. elegans* as a model organism in a range of different biological scenarios will be provided, outlining the basic requirements of LOC devices to be used for the study of multicellular organisms and highlighting the technical challenges they have had to overcome. Finally, some approaches that so far have been of particular value for the study of *C. elegans* will be shown.

Introduction to *C. elegans*

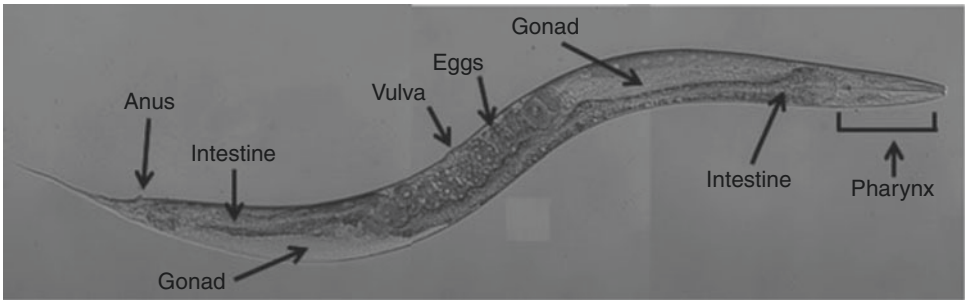
C. elegans Physiology

Caenorhabditis elegans (*C. elegans*) is a small, transparent, free-living nematode which lives in temperate soil environments feeding on bacteria and fungi. It can often be found in compost heaps, fallen fruits, and some invertebrates, but hardly anything is known about its life in the wild.

C. elegans is small. The length of the adult hermaphrodite worm is around 1.2–1.5 mm and the diameter 30–40 μm. It also has a very short life cycle (compared to other model organisms) of about 3 days under optimal conditions of food and temperature (Fig. 1). The life cycle of these



Lab-on-a-Chip for Studies in *C. elegans*, Fig. 1 Representation of the life cycle of *C. elegans*. The different stages of growth: egg, L1, L2, L3, L4, and adult are shown. Adult worms are fertile and capable of laying eggs. These are laid outside during the gastrula stage. If worms are deprived of food in larval stages or environmental conditions are not optimal they will enter a larval stage known as dauer, in which development is arrested. When food becomes available development will continue. Numbers along the *arrows* indicate the length of time the animal spends at a certain stage. The length of the animal at each stage is marked underneath the stage name in micrometers (μm) (Reproduced from [28] with permission)



Lab-on-a-Chip for Studies in *C. elegans*, Fig. 2 Anatomical features of an adult hermaphrodite *C. elegans*. *Arrows* point to the main features of the worm: the pharynx, intestine, vulva, and gonad

nematodes starts with the fertilization of the eggs within the adult hermaphrodite. The fertilized eggs are laid when they are at the gastrula stage and contain about 40 cells. Embryogenesis occurs in 12–14 h after which the animal goes through four juvenile stages (L1–L4) each of which ends in a molt. Following the final molt, the animals reach adulthood at which point they become fertile producing about 300 progeny each over the course of 3 days. When environmental conditions are not optimal (e.g., high temperature, overcrowded cultures, or starvation), the L3 stage is replaced by a dauer larval stage that corresponds to the infectious stage of parasitic nematodes. The dauer of free-living soil nematodes is sometimes associated with some invertebrates. Dauer larvae are thin and are unable to feed because their mouths are closed with a cuticle but survive high temperatures and other harsh conditions much better than adult worms. Most interestingly, they can remain viable for 3 months and resume the L4 reproductive stage when they encounter more favorable environmental conditions. Adult worms have an average life span of 2–3 weeks.

C. elegans displays two sexes: hermaphrodites and males. In culture, males are produced spontaneously at low frequency (about 0.1 %) but their frequency and crossing rate in the wild are unknown. The hermaphrodite produces both sperm and oocytes and can therefore either self-fertilize or be fertilized by a male.

The worms are cylindrical in shape. They essentially consist of an inner and outer tube separated by the pseudocoelomic space. The inner tube contains the gonad, pharynx, and intestine and the outer tube contains neurons, muscle, excretory system, hypodermis, and cuticle. Figure 2 shows a photograph of the anatomical features of a gravid adult *C. elegans* with the pharynx, intestine, and the reproductive system highlighted.

The hermaphrodite and male worms consist of a constant number of somatic cells of 959 and 1,031, respectively. The haploid genome size is relatively small (9.7×10^7 base pairs) about 30 times smaller than the human genome and contains near 20,000 genes. In 1998, *C. elegans* became the first multicellular organism to have its genome fully sequenced. About 35 % of *C. elegans* genes have human homologs. The genome consists of five pairs of autosomal chromosomes and a pair of sex chromosomes: XX in the hermaphrodite and XO in the male.

C. elegans development has been fully characterized, meaning the complete cell lineage of the worm has been recorded. All the cell divisions required to generate a specific group of differentiated cells has been documented for the whole animal [1].

C. elegans as a Model Organism

In 1974, Sydney Brenner established *C. elegans* as a model organism when working on the genetics of the animal development (including neural development), behavior, and physiology [2]. Brenner, with his exhaustive labor, set up the basis for today's worm research and as recognition for this work he was awarded the Nobel Prize in Physiology or Medicine in 2002.

There are several important features that make *C. elegans* an attractive organism not only for the study of gene regulation and gene function, but also for the study of many other different relevant biological aspects including development, cell biology, behavior, and neurobiology. *C. elegans* is easy and cheap to maintain in the laboratory. It grows on agar plates feeding on *Escherichia coli* lawns. It can also be grown in liquid cultures in large quantities. *C. elegans* is a multicellular organism which goes through a complex developmental process, including embryogenesis, morphogenesis, and growth to adulthood. As mentioned before, it has a short life cycle which is an advantage for developmental studies. It is transparent; therefore, it is possible to use fluorescent makers in order to track different cell types and tissues and to perform high-resolution microscopy.

C. elegans is one of the simplest organisms with a nervous system comprised of only 302 neurons. Nevertheless, it exhibits complex behavioral modalities, such as habituation, sensitization and conditioning in response to stimulus. *C. elegans* nervous system has been completely mapped and all neuronal connections are known. Using microfluidic platforms it has been possible to perform precise neuronal ablations and study the phenotypic and behavioral consequences of this in a high-throughput manner.

In comparison to other model organisms *C. elegans* has a relative short lifespan representing a useful model for the study of aging. Due to its small size it is possible to perform high-throughput drug screening assays. For example, *C. elegans* provides a useful model for the investigation of the mode of action of anthelmintics (group of compounds to treat infections caused by parasitic nematodes). As mentioned earlier, *C. elegans* displays a particular mode of reproduction with self-fertilizing hermaphrodites and facultative males which makes it an interesting model for population genetic studies. Males arise infrequently in *C. elegans* populations, about 0.1 %. This percentage can be increased through heat-shock of L4 hermaphrodites. Once males have been obtained, a population containing up to 50 % males can be maintained through mating.

C. elegans also has the advantage of being amenable to genetic manipulation, through mutagenesis, transgene expression, and RNA interference (RNAi) partly because of its rapid life cycle, small size, and ease of laboratory cultivation. When investigating molecular mechanisms in *C. elegans* either forward or reverse approaches can be taken. Reverse genetic techniques are hypothesis driven. Analysis of the phenotypes of mutants with specific mutations can be used to inform about the role of genes in a drug's mode of action. Alternatively, RNA interference can be used to knock down the expression of specific genes allowing, for example, the study of morphological or behavioral changes associated with the gene.

A number of microfluidic approaches that exploit several of these *C. elegans* features have been described and will be further discussed in the text.

Introduction to Worm Chip Technology

Lab-on-a-chip (LOC) devices provide the ability to perform multiple laboratory operations in a miniaturized chip format. This is mainly due to the fact that they handle extremely small fluid volumes reducing the amount of sample and time needed for a particular analysis. Worm chips, as they are commonly described, are microfabricated LOC devices capable of manipulating single worms or a population of worms and their environment in a precise and automated fashion.

The advantages provided by LOC devices depend very specifically on their applications. In general: (1) they handle very low fluid volumes reducing the cost and sample volume needed for the analysis; (2) they are very efficient systems because they provide high control of the experimental conditions (e.g., thermal control, short diffusion distances); (3) they allow high-throughput analysis because being very compact, LOC systems are capable of massive parallelization; and (4) they can be mass produced with low fabrication costs, allowing cost-effective disposable chips.

LOC devices belong to a broader group of small mechanical devices known as MEMS (Micro-Electro-Mechanical-Systems). Worm chips usually integrate MEMS components to provide flow control (microchannels, pumps, mixers, and valves) and sensing capabilities to the devices (optical or electrochemical sensors) with microfluidics platforms (to deal with and manipulate tiny volumes of fluid samples) [3].

Worm Chip Materials and Fabrication Technologies

Worm chips can be fabricated from many types of material including various polymers, glass, or silicon, or combinations of these materials. A broad variety of fabrication technologies are used for LOC device fabrication; the most frequent being single layer or multilayer soft lithography.

Soft lithography-based microfluidic devices are typically made in polydimethylsiloxane (PDMS). This polymer offers several benefits for biological research including biocompatibility, high permeability to gas, a chemical inert surface, and optical transparency down to 300 nm. The chips normally have a closed architecture as they are sealed with a glass slide, a PDMS layer, or an agar plate. Worm chips that include MEMS are normally fabricated using standard silicon and polymer micromachining processes, which requires expertise in microfabrication and access to a clean room.

Photolithography is the most common method of fabrication of microfluidic devices. Typically it consists of the transfer of a pattern into a photosensitive material by selective exposure to a radiation source such as light. Following the creation of a template (replica mold) PDMS can be cast against it to form the microfluidic devices. This process is cost effective and can be performed in any traditional laboratory setting with the mold being used repeatedly [4].

Worm Chip Technical Challenges

The study of multicellular organisms using microfluidic devices presents some technical challenges that when controlled allow high-throughput phenotyping, screening, imaging, and sorting of small animals in an automated manner. In particular, the organisms need to be handled efficiently and the microenvironment created for the experiment needs to be precisely controlled. Environment parameters, to be controlled, include temperature, nutrients, dissolved gas, and stimuli delivered to the animals.

Worm chips have been designed for the detailed analysis of single worm phenotype. The manipulation in this type of chips involves complete or partial immobilization of the worm. The movement of the worm can be restricted physically or by applying glue, cold, or CO_2. Some examples of immobilization of the worms on chip will be discussed further in section "Key Research Findings: Applications of LOC Devices to the Study of C. elegans." Other worm chips have been designed for the analysis of large populations of animals and do not restrict worm movement. These types of chips are mainly used to study worm behavior and locomotion [5].

Key Research Findings: Applications of LOC Devices to the Study of *C. elegans*

The development of LOC devices for the study of small animals is a new area of research still in progress. Worm chips capable of performing high-throughput studies show great potential to provide a better understanding of key aspects of the biology of nematodes. In this section, a few examples of microfluidic devices designed to study different aspects of the worm biology will be shown and some technical challenges will be highlighted.

Immobilization of Single Animals: Suction Pipette, Glue, Physical Restraint (Worm Trap), Suction Microchannel (Aspiration)

As discussed above, one of the challenges that LOC devices have had to overcome in order to be able to study relevant aspects of the biology of *C. elegans* is the total or partial immobilization of the whole animal. *C. elegans* is a very active animal. It constantly explores the environment while searching for food (foraging behavior). In the laboratory *C. elegans* is normally cultivated on agar plates where it exhibits a sinusoidal

forward and backward movement of unpredictable direction. Locomotion behavior is often measured and quantified by researchers to describe the healthy status of the worm. Immobilization of the animal is a key element in order to get high-resolution images of various anatomical features or to image cellular development and gene expression. In addition, immobilization of the worm is required to precisely deliver stimuli taking control of the microenvironment conditions for the experiment, to perform laser ablations of individual neurons for studying neural networks, axon regeneration, or monitoring neuronal responses at the presence of various stimuli.

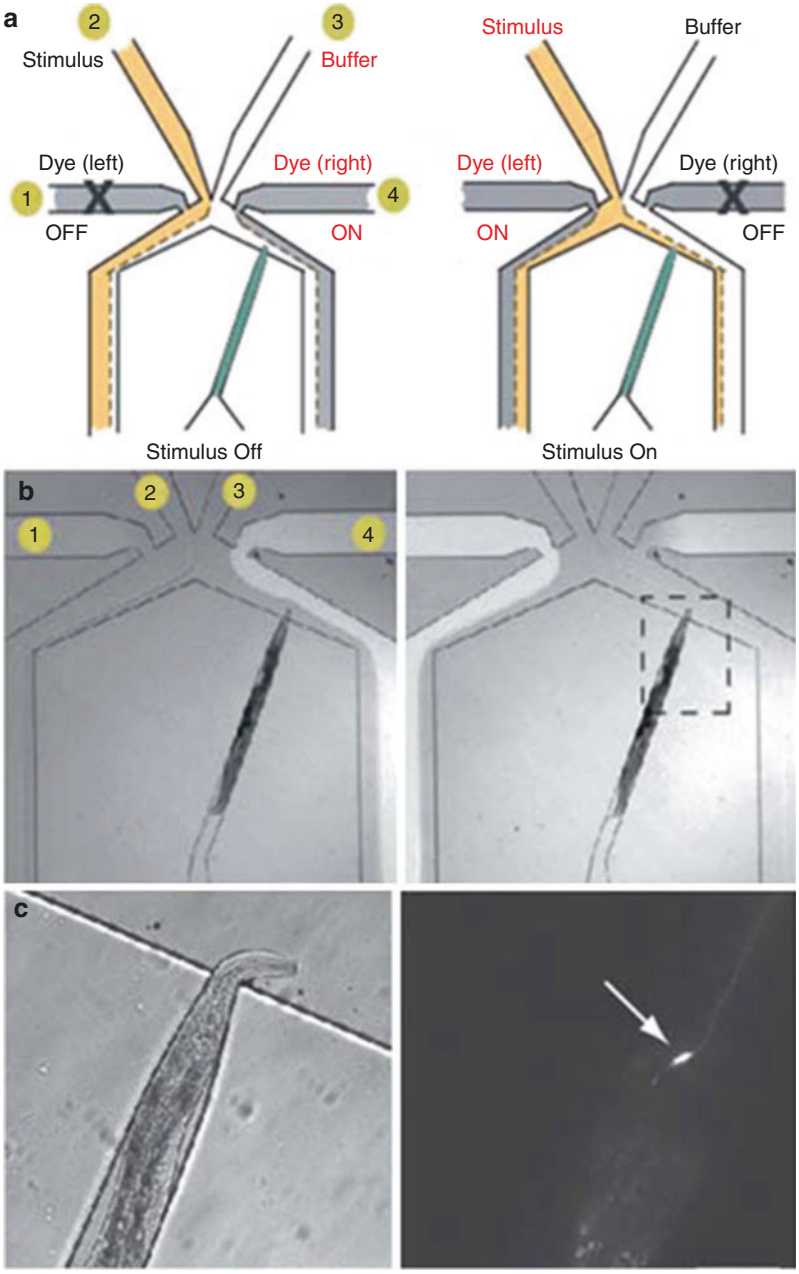
One of the first microfluidic devices developed to study biological questions using *C. elegans* addressed the issue of immobilization of the worm in a flowing solution; this was achieved by holding the animal by the tail using a suction pipette. Forward and backward movement was recorded in response to different chemical stimuli using an array of photocells connected to a multichannel chart recorder [6]. This allowed, for the first time, to control the timing of the stimulus presented to the worm and behavioral events were recorded with relatively high time resolution thus allowing the determination of changes in behavior. With this device the author demonstrated the mechanism of chemotaxis and described for the first time that the frequency of change of direction varied with stimulus intensity in the appropriate direction and the animal does adapt rapidly.

C. elegans evaluates the quality of its environment through a complex and versatile chemosensory system. Peripheral chemosensation is highly divergent between mammals and *C. elegans*, but regulatory and behavioral mechanisms appear to be more conserved across species. *C. elegans*' chemosensory behavior was tested again a few years later on a device where the worm was semi-restrained. The anterior part of the worm was glued to the substrate and the worm's behavior was recorded and analyzed quantitatively [7]. The use of mutant worms in this study provided insight into the understanding of the neural basis of behavior and represented a new technique to use in combination with genetic or neuronal manipulations.

More sophistication in the design of the LOC devices was shown by Chronis et al. In their device to measure olfactory neuron activity, the worm was physically immobilized in a narrow space (i.e., no glue or anesthetic was used) (Fig. 3). This device combined the concept of worm trap (used as well to analyze multiple worms at the same time [8]) with a microfluidic chemical delivery system and allowed to measure the activity of chemosensory neurons in response to chemosensory stimuli [9]. The same authors also designed a chip to measure neural activity in relation to behavior. In this device the worms, one at a time, were able to move forward and backward and neural activity could be measured simultaneously for the first time. The results that emerged from this study had a high biological relevance. Prior to this work, nobody had shown neural activity in real time; this revealed important finds relating to the behavior of the olfactory neurons. Olfactory neurons are activated in the absence of the stimuli (food odor) and remain deactivated when the food odor appears. The amount of activity of the neuron depends on how strong the odor is and how long it lasts.

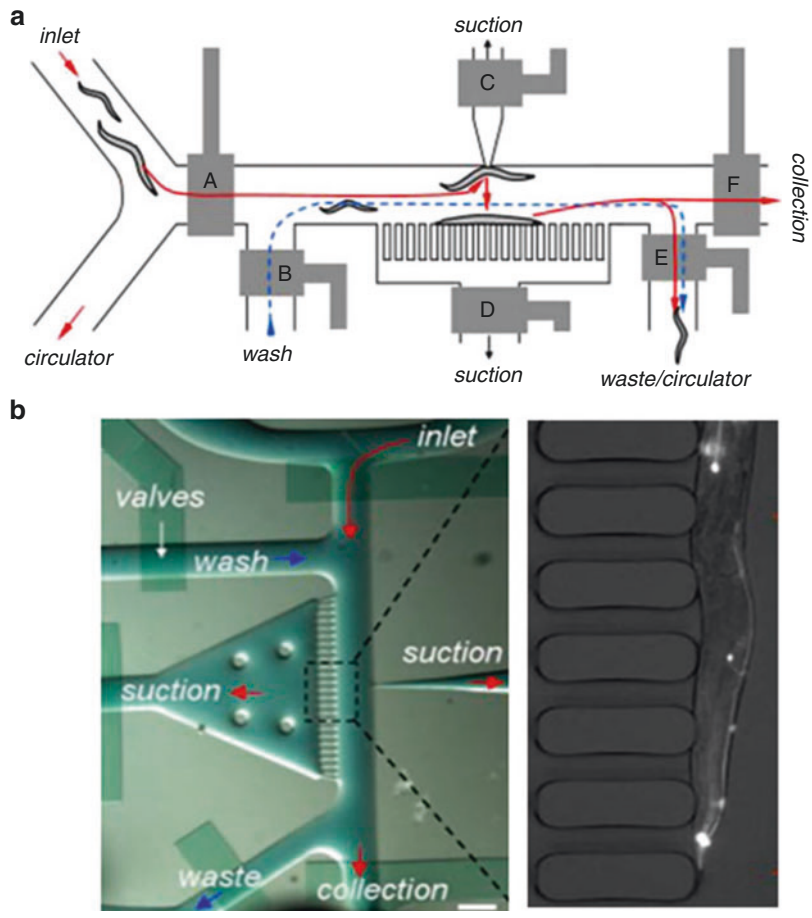
A microsystem for high-throughput whole animal sorting and screening at subcellular resolution was developed by Rohde et al. [10]. The device was capable of isolating and immobilizing single worms for the screening of phenotypic features at subcellular resolution in live animals (Fig. 4). The animals were exposed to biochemical compounds and high-resolution time-lapse imaging of many individual worms on a single chip, without the need of anesthesia, was performed. Individual worms were capture in an immobilization chamber via suction by a microchannel held at a low pressure and imaged using high-resolution optics. After image acquisition and processing, the worm can be released and directed to the appropriate collection channels according to its phenotype.

Zeng et al. presented in 2008 an improved immobilization system for single animals that allowed performing techniques such as femtosecond-laser manipulation (laser microsurgery) and three-dimensional two-photon



Lab-on-a-Chip for Studies in *C. elegans*, Fig. 3 Lab-on-a-chip device to measure olfactory neuron activity in response to chemosensory stimuli. (**a**) The chip consists of two channels (*1* and *4*) loaded with fluorescent dyes that are used to redirect the chemical stimulus (applied in channel *2*) to the nose of the worm. (**b**, **c**) Higher magnification photographs of a worm trapped in the chip. The *arrow* points to a sensory neuron which is expressing a fluorescent marker in response to the stimulus delivered (Reproduced from [5, 9] with permission)

microscopy (which cannot be done on anesthetized animals) [11]. They described their immobilization method as fast and noninvasive because it did not require the use of anesthesia or cooling and allowed high-throughput on-chip worm manipulation at subcellular level. The device combined aspiration channels to capture and align the animals in a linear position (achieving partial immobilization) and a "sealing membrane" that flexed on top of the captured animal forming a tight seal which completely restrained the motion in a linear orientation. This method allowed highly demanding subcellular resolution imaging and manipulation techniques (such as laser microsurgery for precision ablation with minimal collateral damage and 3D two-photon microscopy



Lab-on-a-Chip for Studies in *C. elegans*, Fig. 4 Microfluidic worm sorter. (**a**) The sorter consists of control channels and valves (labeled with the letters *A–F*) that direct the flow of worms in different directions in the flow channels. Animals enter the chip through the inlet channel and can be continuously recirculated. The immobilization chamber captures one worm each time via suction by multiple microchannels held at low pressure and subsequently, the image acquisition and processing are performed. The worm is either collected or directed to the waste, depending on its phenotype. (**b**) Image of the worm sorter and a worm trapped in the system. Bright spots along the animal show mechanosensory neurons expressing a fluorescent reporter (Reproduced from [10] with permission)

including two-photon excitation fluorescence or TPEF) that required long periods of stable immobilization.

Immobilization of Multiple Animals: Arrays of Worm Traps, Chip-Gel Hybrid

The first micro-engineered array of clamps for worms (worm trap) was designed by Hulme et al. with the purpose of investigating the physiology and behavior of large populations of worms. The immobilization of the worms in this device was achieved by an array of wedge-shape channels (clamps) where the worms were trapped and oriented by the direction of the microfluidic flow. The purpose of this work was to identify mutants affecting anatomical structure or body size distribution of a population of worms. One major advantage of this device was that the worms could be released after the analysis by reversing the flow direction. Therefore repeated sampling of the same population of worms at different times could be performed. The authors measured life span and brood size from the released worms [8]. This type of worm chip provides a high-throughput system for morphological analysis with potential to be used for extremely delicate operations such as microsurgery; and with the combination of high-resolution optical imaging could be used for fluorescence imaging. An improved modified version of this system was used to culture *C. elegans* in liquid, in separate chambers from L4 throughout the entire life span of the worm until death [12]. The device included microfluidic worm clamps to allow the temporary and reversible immobilization of each worm, making it possible to monitor changes in body size and locomotion throughout the life span of the animal.

In 2008, Chung et al. presented the first microsystem with complete automation of sample handling for high-resolution microscopy phenotyping and sorting of *C. elegans* [13]. This device could analyze and sort several hundred worms per hour using a brief and reversible immobilization by transiently cooling to 4 °C. This system in comparison with the COPAS (Complex Object Parametric Analyzer and Sorter), which is a fluorescence-activated cell sorter modified to handle small organisms, represents a much more advanced system not limited by the tissue level. The automated microsystem is capable of high-throughput high-resolution microscopy and allows sorting based on phenotype. In this case, the authors based the sorting in the expression of a reporter in neurons.

Almost simultaneously another chip to perform laser ablations was also described [14]. The nanoaxotomy chip is a microfluidic device used to sever axons in *C. elegans* using ultrashort laser pulses with high precision and monitor axonal regeneration activity. This chip presented several advantages over the chips used previously. As mentioned before, immobilization of the animal was done without the need of glue or anesthetic therefore not interfering with physiological processes. This system used a two-layer microfluidic approach with a thin membrane between the two channels. When the top layer was pressurized the membrane deflected quickly immobilizing the worm. The chip had great flexibility in terms of being able to analyze worms from L4 to adult size; furthermore, the worms did not need a recovery period after surgery, permitting immediate behavioral studies. The chip also allowed control of the sample population.

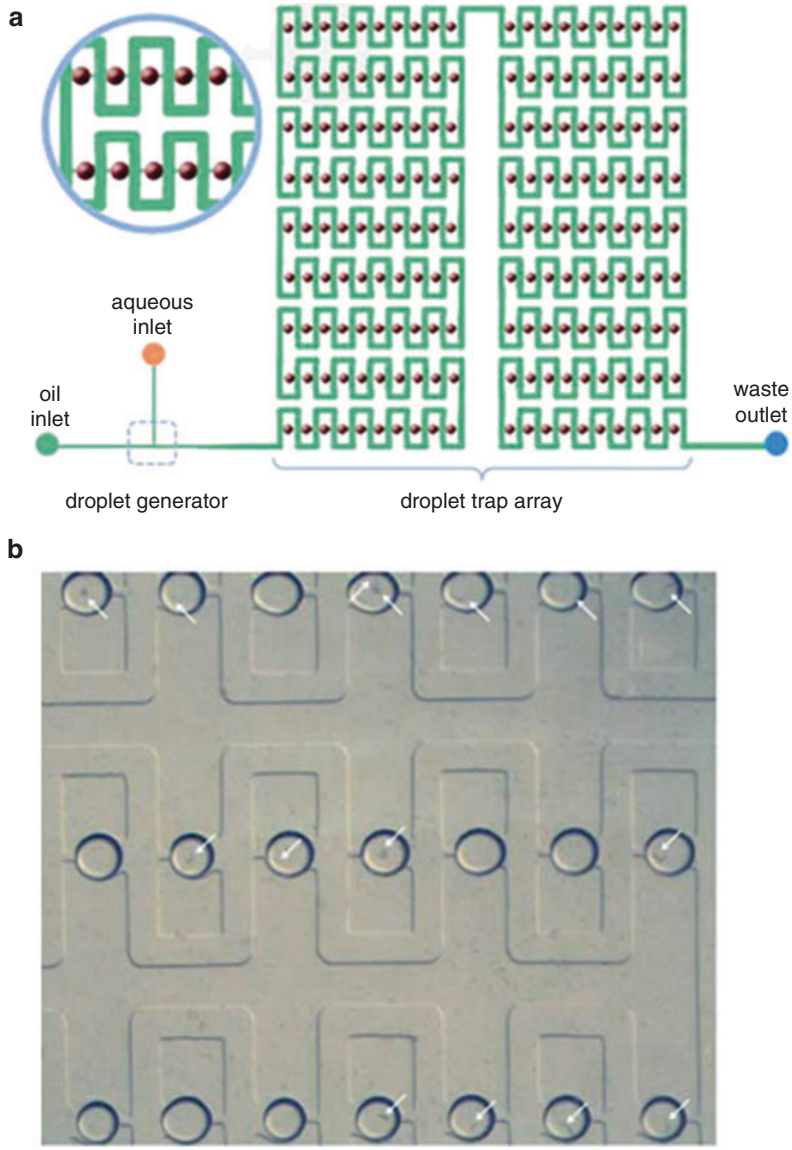
In 2009, Chokshi et al. explored the use of CO_2 to immobilize *C. elegans* on a microfluidic chip and characterized the effect of this immobilization method on worm behavior. The authors also tested mechanical immobilization through a deformable membrane of PDMS that compressed the worm to mechanically restrict its movement (compressive immobilization). They found that CO_2 immobilization was effective and efficient for up to 2 h with low photobleaching, whereas the mechanical approach was best for shorter periods of time. Locomotion was measured as well as gene reporter expression [15, 16].

In 2010, Krajniak et al. presented a chip-gel hybrid microfluidic device for developmental studies. The novelty of this device was the use of a biocompatible polymer called Pluronic F127 as the immobilization agent in combination with a microfluidic platform. This polymer was reversibly transformed into a gel in a process that depends on temperature. The gel phase was sufficient to immobilize the animals. While the animals were not imaged they were cultured in individual chambers containing media with nutrients required for development. This device allowed culturing, individual tracking and repeated imaging of *C. elegans* at high-resolution throughout development [17].

No Immobilization: Microchamber for Small Volume Liquid Culture, CD Platform, Droplet Generator, Artificial Dirt

The appearance of the first platform to maintain and monitor *C. elegans* in a microfluidic environment arrived almost 25 years after the first example of a microfluidic device for the study of the worm. This device was designed for the experimental study of *C. elegans* in space. It consisted of a chamber able to support a tiny volume of liquid culture of *C. elegans* for long periods of time with a membrane to allow gas exchange [18]. Shadow images were acquired by a camera attached at the bottom of the microchamber. The authors measured the activity of *C. elegans* as a function of temperature finding that thrashing (stroke frequency) increased with temperature validating previous results obtained with manual counting.

As an alternative to the immobilization of the animal with glue, other platforms for culture and behavioral observation have also been used. In 2007, Kim et al. [19] described a compact disc microfluidic system for the automatic cultivation of *C. elegans*. This system showed to be proficient in automated feeding, waste removal, and live-animal microscopy using centrifugal force-driven fluidics. This system could analyze and control a population of worms in one chamber (up to three generations could be monitored over time) and measure the rate of development, brood size, and



Lab-on-a-Chip for Studies in *C. elegans*, Fig. 5 Droplet-based microfluidic system for single worm analysis. (**a**) The device consists of a parallel series of microdroplets (droplet trap array) continuously generated by shearing a dispersed aqueous phase with a constant oil phase at the T-junction droplet generator; the inset shows a magnified view of ten traps with droplets. (**b**) Image of an array of droplets encapsulated with worms. The position of individual worms is indicated by *white arrows* (Reproduced from [21] with permission)

the rate of population growth. As shown previously, with other microsystems, locomotion could also be measured in single worms as an indicator of the physiological state of the worm. The CD cultivation system showed the capacity to incorporate diverse molecular, biochemical, and pharmacological assays for real-time physiological and behavioral test of live animals. This system has also been used to study the effect of hypergravity and the ability of *C. elegans* to withstand mechanical stress. The device affords an opportunity to investigate molecular pathways that transduce mechanical forces into biological regulators and to identify genetic targets of mechanical stress in the context of the whole animal [20].

Droplet-based microfluidic systems represent another type of chips for individual worm assay [21]. These systems consist of a droplet generator and a droplet trap array, which allows the encapsulation of individual worms into parallel series of droplets of tiny volumes (Fig. 5). The authors studied worm swimming behavior in response to a neurotoxin. This work represents for the first

time droplet encapsulation of small organisms. The main disadvantage is that the worms only survived for a maximum of 2 h.

Artificial dirt illustrates novel experimental microfluidic devices to study *C. elegans* neurobiology and behavior. Artificial dirt devices mimic the natural environment where *C. elegans* can be found in contrast to the planar surface of an agar plate. The authors argue that this fails to reflect the complexity of *C. elegans* natural environment with complicated stimulus delivery. Artificial dirt devices are thin and transparent and consist of microchambers and agarose-free channels that mimic the moist soil environment with rapid delivery of fluid-borne stimuli. The animals are able to crawl as they would on agar plates and are compatible with optical recording. Artificial dirt devices were made to accelerate studies of the neuronal basis of behavior in *C. elegans* [22].

Others: Live-Animal FACS, Maze, Optofluidic On-Chip Platform

A curious example of the broad applicability of worm chips is the device designed by Qin et al. to study behavioral plasticity of an animal using a microfluidic platform [23]. These authors studied the tendency of the worm to explore different environments in the presence or absence of food in one of the outlet chambers. They measured locomotion rates and compared the results obtained from mutant and wild-type worms.

Live-animal FACS (la-FACS) is a Fluorescence-Activated Cell Sorting (FACS) based method to sort live *C. elegans* larvae. Up until recently FACS only allowed sorting of *C. elegans* embryos [24]. The la-FACS is able to obtain large quantities of pure mutant populations. This method has a great potential for expanding the arsenal of high-throughput tools available in *C. elegans* [25].

There are many further examples were *C. elegans* has been used. For example in 2006, Heng et al. used *C. elegans* and a technique described as optofluidic microscopy (OFM) on chip. L1 worms were euthanized to demonstrate the feasibility of the method to perform high-resolution imaging.

Future Directions of Research

The development of LOC systems over the last few years has been spectacular. The future of microfluidic devices offers a range of interesting and new opportunities in worm research. Worm chips offer several advantages over conventional approaches including the maintenance of well-controllable microenvironments, creation of reproducible experimental conditions, and automation of tedious experimental protocols. However, some improvements are still needed in order to overcome the challenges arising from working with small volumes and multicellular living organisms such as *C. elegans*.

In addition to imaging different features of the worm, most of the microfluidic devices described so far have focused on measuring locomotion behavior. However, *C. elegans* displays many other types of behavior which have not been exploited thus far using a worm-chip platform. One example is pharyngeal pumping behavior.

The *C. elegans* pharynx is required for the ingestion and mechanical digestion of food. There is a basement membrane isolating the pharynx from the rest of the animal, making it an almost completely self-contained system. The pharyngeal nervous system coordinates the pharyngeal muscle cells allowing food to be sucked in, broken down, and passed into the intestine. This process is referred to as pharyngeal pumping. The pharyngeal pumping of *C. elegans* can either be counted optically or electrically. The method of recording pumping electrophysiologically gives what is referred to as the electropharyngeogram or EPG. This is an extracellular recording that can be performed on intact worms or on isolated pharynxes. As well as producing a readout of pump rate, this technique can also be used to provide information on neuronal activity of the pharynx. The EPG trace has excitatory and inhibitory components that can be attributed to the function of specific pharyngeal neurons. Measuring EPGs is a laborious and tedious process and it takes a long time to analyze a single worm. The possibility to perform EPG-recordings in a high-throughput and automated fashion remains to be exploited using microfluidic platforms.

RNAi represents another technique that has not been exploited using microfluidic devices yet.

RNAi is a powerful tool to investigate gene function in *C. elegans* broadly used to do worm research in many different areas such as development or aging. So far nobody has used this tool on a microfluidic platform in order to perform a genetic screen. In this scenario the combination of RNAi and a microfabricated worm sorter device would result in a very powerful platform.

Further Reading

Almost everything you want to know about *C. elegans* itself can be found online. Wormbase is a complete database about the worm with connections to data of human and other model organisms. Wormbase contains complete cell lineages as well as the genetic and physical maps and the annotated sequence of the worm. Also all web references can be found there: http://www.wormbase.org/.

The latest reviews describing microfluidic devices applied for the study of *C. elegans* are listed next: [5, 26, 27].

Cross-References

- Applications of Nanofluidics
- MEMS Adaptive Optics
- MEMS Display Based on Interferometric Modulation Technology
- MEMS on Flexible Substrates
- Microfabricated Probe Technology
- Micropumps
- Nanoscale Properties of Solid–Liquid Interfaces
- On-Chip Flow Cytometry
- RNAi in Biomedicine and Drug Delivery
- Thermal Cancer Ablation Therapies Using Nanoparticles

References

1. Riddle, D.L., Blumenthal, T., Meyer, B.J.: *C. elegans* II, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor. (1997)
2. Brenner, S.: The genetics of *Caenorhabditis elegans*. Genetics **77**(1), 71–94 (1974)
3. Bhushan, B. (ed.): Springer Handbook of Nanotechnology, 3rd edn. Springer, Berlin (2010). 1964
4. Herold, K.E., Rasooly, A.: Lab-on-a-Chip Technology: Fabrication and Microfluidics, vol. 1. Caister Academic Press, Norfolk (2009)
5. Chronis, N.: Worm chips: microtools for *C. elegans* biology. Lab Chip **10**(4), 432–437 (2010)
6. Dusenbery, D.B.: Responses of the nematode *Caenorhabditis-elegans* to controlled chemical-stimulation. J. Comp. Physiol. **136**(4), 327–331 (1980)
7. Faumont, S., Miller, A.C., Lockery, S.R.: Chemosensory behavior of semi-restrained *Caenorhabditis elegans*. J. Neurobiol. **65**(2), 171–178 (2005)
8. Hulme, S.E., et al.: A microfabricated array of clamps for immobilizing and imaging *C.elegans*. Lab Chip **7**(11), 1515–1523 (2007)
9. Chronis, N., Zimmer, M., Bargmann, C.I.: Microfluidics for in vivo imaging of neuronal and behavioral activity in *Caenorhabditis elegans*. Nat. Methods **4**(9), 727–731 (2007)
10. Rohde, C.B., et al.: Microfluidic system for on-chip high-throughput whole-animal sorting and screening at subcellular resolution. Proc. Natl. Acad. Sci. U. S. A. **104**(35), 13891–13895 (2007)
11. Zeng, F., Rohde, C.B., Yanik, M.F.: Sub-cellular precision on-chip small-animal immobilization, multi-photon imaging and femtosecond-laser manipulation. Lab Chip **8**(5), 653–656 (2008)
12. Hulme, S.E., et al.: Lifespan-on-a-chip: microfluidic chambers for performing lifelong observation of *C. elegans*. Lab Chip **10**(5), 589–597 (2010)
13. Chung, K.H., Crane, M.M., Lu, H.: Automated on-chip rapid microscopy, phenotyping and sorting of *C. elegans*. Nat. Methods **5**(7), 637–643 (2008)
14. Guo, S.X., et al.: Femtosecond laser nanoaxotomy lab-on-a-chip for in vivo nerve regeneration studies. Nat. Methods **5**(6), 531–533 (2008)
15. Chokshi, T.V., Ben-Yakar, A., Chronis, N.: CO_2 and compressive immobilization of *C. elegans* on-chip. Lab Chip **9**(1), 151–157 (2009)
16. Chokshi, T.V., Bazopoulou, D., Chronis, N.: An automated microfluidic platform for calcium imaging of chemosensory neurons in *Caenorhabditis elegans*. Lab Chip **10**(20), 2758–2763 (2010)
17. Krajniak, J., Lu, H.: Long-term high-resolution imaging and culture of *C. elegans* in chip-gel hybrid microfluidic device for developmental studies. Lab Chip **10**(14), 1862–1868 (2010)
18. Lange, D., et al.: A microfluidic shadow imaging system for the study of the nematode *Caenorhabditis elegans* in space. Sens. Actuators B-Chem. **107**(2), 904–914 (2005)
19. Kim, N., et al.: Automated microfluidic compact disc (CD) cultivation system of *Caenorhabditis elegans*. Sens. Actuators B-Chem. **122**(2), 511–518 (2007)

20. Kim, N., et al.: Gravity force transduced by the MEC-4/MEC-10 DEG/ENaC channel modulates DAF-16/FoxO activity in *Caenorhabditis elegans*. Genetics **177**(2), 835–845 (2007)
21. Shi, W.W., et al.: Droplet-based microfluidic system for individual *Caenorhabditis elegans* assay. Lab Chip **8**(9), 1432–1435 (2008)
22. Lockery, S.R., et al.: Artificial dirt: microfluidic substrates for nematode neurobiology and behavior. J. Neurophysiol. **99**(6), 3136–3143 (2008)
23. Qin, J.H., Wheeler, A.R.: Maze exploration and learning in *C.elegans*. Lab Chip **7**(2), 186–192 (2007)
24. Stoeckius, M., et al.: Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression. Nat. Methods **6**(10), 745–751 (2009)
25. Fernandez, A.G., et al.: Automated sorting of live *C. elegans* using la-FACS. Nat. Methods **7**(6), 417–418 (2010)
26. Crane, M.M., et al.: Microfluidics-enabled phenotyping, imaging and screening of multicellular organisms. Lab Chip **10**, 1509–1517 (2010)
27. Taylor, A.M., Jeon, N.L.: Micro-scale and microfluidic devices for neurobiology. Curr. Opin. Neurobiol. **20**(5), 640–647 (2010)
28. Altun, Z.F., Hall, D.H.: Introduction to *C. elegans* anatomy. In: Altun, Z.F., Hall, D.H. (eds.) Worm Atlas. Cold Spring Harbor Laboratory Press (2009). ISBN 978-087969794-5

Lab-on-a-Chip Whole-Cell Biosensor

▶ Microfluidic Whole-Cell Biosensor

Lamb-Wave Resonators

▶ Laterally Vibrating Piezoelectric Resonators

Lamé Mode MEMS Resonators

Joshua E-Y Lee
Department of Electronic Engineering, City University of Hong Kong, Kowloon Tong, Hong Kong

Synonyms

Lamé mode resonators; Lateral shear mode resonators; Wine-glass mode resonators

Definition

A Lamé mode MEMS resonator comprises a microscale square plate designed to be excited to vibrate in a particular lateral antisymmetric mode within the plane of fabrication. The Lamé mode can be seen as two opposite sides of the square plate expanding, while the other two opposite sides on the orthogonal axis contract. The corners of the square plate are nodes in the Lamé mode such that the square plate can be conveniently clamped for minimal anchor loss. The Lamé mode is a planar pure shear vibration mode.

Modal Analysis

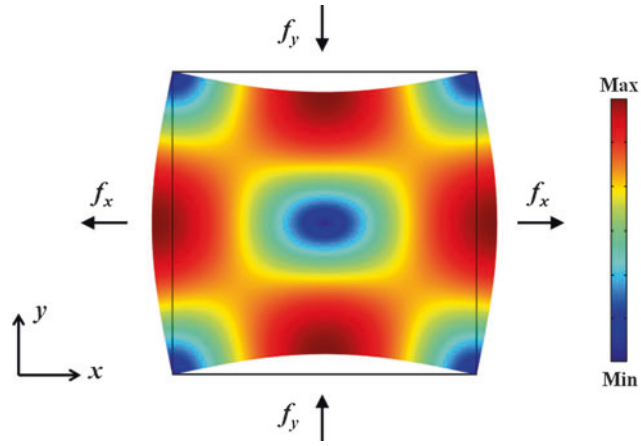
As illustrated in Fig. 1, the Lamé mode is typically actuated by forces that are out of phase between adjacent sides of the square plate. Given that the plate is thin, whereby the length (L) of the plate is much greater than its thickness (H), we ignore displacements along the thickness (z-direction) of the plate. The resulting Lamé mode shape can then be mathematically defined for just the in-plane axes:

$$x-\text{direction}: \quad u(x,y) = \sin\left(\frac{\pi x}{L}\right)\cos\left(\frac{\pi y}{L}\right) \tag{1a}$$

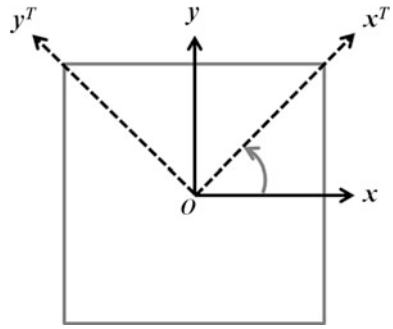
$$y-\text{direction}: \quad v(x,y) = -\sin\left(\frac{\pi y}{L}\right)\cos\left(\frac{\pi x}{L}\right) \tag{1b}$$

Note that the amplitudes of displacement in either axis have been normalized to one. Neglecting strain components associated with the thickness z-direction assuming thin plate theory, only the lateral strain components remain. This includes the normal strains along the x and y axes denoted by ε_x and ε_y, respectively, and the shear strain in the x-y plane denoted by γ_{xy}. From Eq. 1, we obtain the lateral strain components for the Lamé mode:

$$\varepsilon_x(x,y) = -\varepsilon_y(x,y) \tag{2a}$$



Lamé Mode MEMS Resonators, Fig. 1 Illustration of the Lamé mode shape and requirement for antiphase actuation forces applied through adjacent capacitive gap electrodes



Lamé Mode MEMS Resonators, Fig. 2 Depiction of rotated planar coordinates for deriving the shear characteristics of the Lamé mode

$$\gamma_{xy}(x,y) = 0 \tag{2b}$$

By applying planar strain coordinate transformation using Mohr's circle of strain, we find that maximum shear strain occurs at 45° to the x and y axes while the normal strains become zero. In this respect, the Lamé mode can be described as a lateral shear mode characterized by a shear wave along the diagonals of the square plate. Hence, it is convenient to perform a 45° rotation to the planar coordinates when deriving the shear characteristics of the Lamé mode as illustrated in Fig. 2. The transformed mode shape according to the rotated planar coordinates is then given by:

$$\text{Rotated } x^T - \text{direction} : u^T\left(x^T, y^T\right) = -\frac{1}{\sqrt{2}} \sin\left(\frac{\sqrt{2}\pi}{L} y^T\right) \tag{3a}$$

$$\text{Rotated } y^T - \text{direction} : v^T\left(x^T, y^T\right) = -\frac{1}{\sqrt{2}} \sin\left(\frac{\sqrt{2}\pi}{L} x^T\right) \tag{3b}$$

The superscript T denotes the transformation of a variable according to the coordinate rotation. Solving the relevant equation of motion, we obtain the lumped mechanical parameters of mass (m) and linear spring constant (k) for the Lamé mode:

$$m = \frac{1}{2}\mu H L^2 \tag{4a}$$

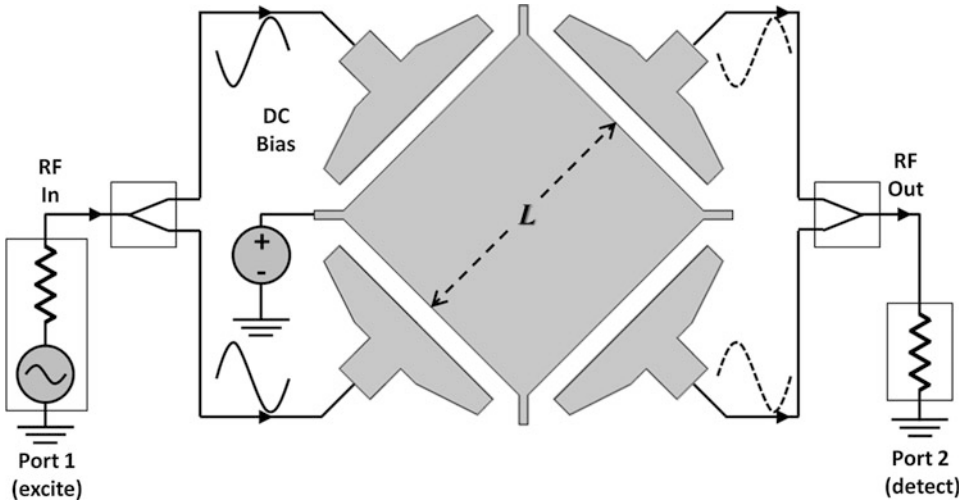
$$k = \pi^2 G_{x^T y^T} H \tag{4b}$$

$G_{x^T y^T}$ refers to the shear modulus with respect to the rotated coordinates, in other words along the lengths of the square plate. In addition, μ denotes the density of the structural material. From Eq. 4, we obtain the resonant frequency of the Lamé mode [1, 2]:

$$f_0 = \frac{1}{\sqrt{2}L}\sqrt{\frac{G_{x^T y^T}}{\mu}} \tag{5}$$

Electromechanical Transduction

The Lamé mode can be electrostatically actuated through capacitor gap electrodes fabricated on



Lamé Mode MEMS Resonators, Fig. 3 Fully differential biasing configuration to electrically excite and detect the Lamé mode using capacitive transduction

each side of the square plate. The displacement of the resonator can similarly be picked up electrically through the capacitor gaps. A commonly used configuration of electrically interfacing a Lamé mode resonator is depicted in Fig. 3. In the configuration shown in Fig. 3, a DC bias voltage is applied to the structure, while a differential pair of AC voltages is applied to adjacent electrodes. The combination of a DC bias voltage (V_{dc}) on the structure and a differential pair of AC voltages (v_{in}) generates a pair of forces that are out of phase to match the antisymmetry of the Lamé mode in order to excite the Lamé mode. The total electrostatic force imposed on the resonator from two sides, taking into account the nonuniform edge displacement of the Lamé mode, is given by

$$F = \frac{4}{\pi}\frac{\varepsilon_o HL}{g^2} V_{dc} v_{in} \tag{6}$$

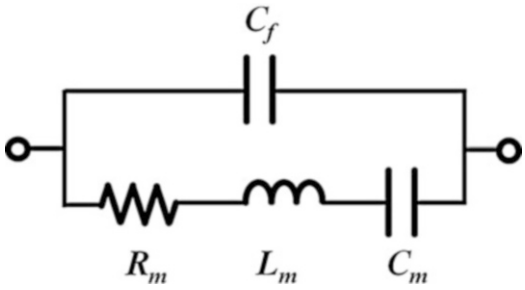
The above equation approximates the length of the electrode to be the same as the length of the square plate. Note that g denotes the nominal gap of the capacitive transducer under zero displacement conditions and ε_o is the permittivity in vacuum.

The lateral amplitude displacements at the sides of the square plate, denoted by a lumped variable x, are detected through the remaining capacitive gap electrodes. As shown in Fig. 3, the same DC bias voltage is dropped across both capacitive gaps; thus the same charge is stored on both gaps when the square plate is at rest. Modulation of the gap resulting from exciting the Lamé mode results in a change in capacitance. Given that the gap change in one gap is opposite to the adjacent gap, the resulting motional currents between the two electrodes will also be out of phase. By detecting the motional currents from the two electrodes differentially, the total motional current for an amplitude displacement x at the driving radian frequency of ω is given by

$$i_{out} = \frac{4}{\pi}\frac{\varepsilon_o HL}{g^2} V_{dc} \omega x \tag{7}$$

As in the case when considering the actuation port, the above expression once again takes into account the nonuniform edge displacement of the Lamé mode.

Like most other capacitive MEMS resonators, the electromechanical transfer function of the Lamé mode MEMS resonator can be represented by the Butterworth van Dyke (BVD) circuit model as depicted in Fig. 4. Note that η is the transduction factor common to each of the electromechanical lumped parameters in the BVD model. For the Lamé mode when biased with the setup shown in Fig. 3, η is given by:



Lamé Mode MEMS Resonators, Fig. 4 Butterworth van Dyke (BVD) equivalent circuit model to describe the electromechanical transfer function of a Lamé mode MEMS resonator

$$\eta = \frac{4}{\pi}\frac{\varepsilon_o HL}{g^2}V_{dc} \tag{8}$$

The equivalent circuit model depicted in Fig. 4 includes a capacitor (C_f) between the input and output ports. C_f captures the effect of feedthrough between the input and output ports through a parasitic capacitor [3]. The physical sources of C_f could include, though by no means exhaustively, the direct overlapping capacitance of the capacitive gap transducer, capacitive coupling through the substrate, interconnects, and the electrical package holding the die. Although the value of C_f may be relatively small (pF for packaged devices) [4], due to the comparatively weak electromechanical coupling for capacitive transducers, the electrical path through C_f becomes significant and could potentially obscure the output motional current. The fully differential configuration in Fig. 3 as such has the benefit of suppressing feedthrough [5] by providing feedthrough cancelation at the input and output ports, thus enhancing the electrical characterization of the MEMS resonator [6].

Quality (*Q*) Factor and Dissipation

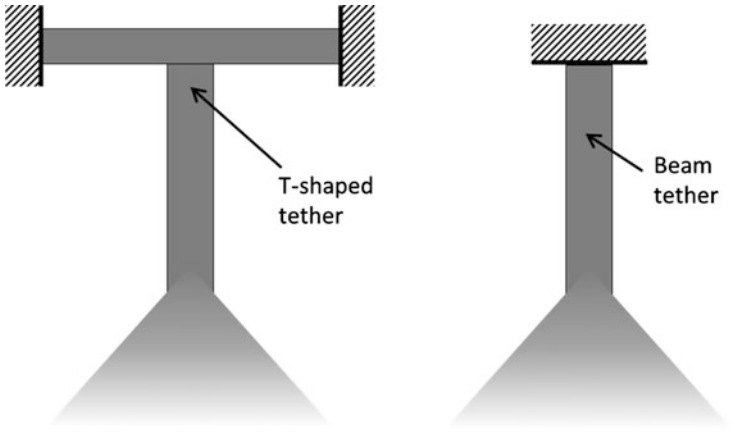
A MEMS oscillator is realized by embedding a MEMS resonator in an amplifier loop to achieve self-sustained oscillation. MEMS oscillators have made inroads during the last decade in starting to replace quartz crystals as timing references [7]. Compared to quartz crystals, MEMS oscillators offer the advantage of small form factor and greater compatibility of their fabrication with complementary metal-oxide-semiconductor process technology. High-quality (Q) factor benefits the close-to-carrier phase noise performance of the oscillator. From the view of high Q, silicon Lamé mode resonators are highly attractive in that they exhibit Qs that are close to the fundamental material limit set by phonon-phonon interaction losses. Examples of Q in the millions for MHz silicon Lamé mode resonators have been reported in the literature [8–10]. To see what governs Q in Lamé mode resonators, we can consider the various sources of damping since the net Q of a resonator is determined by the sum of all dissipative sources:

$$\frac{1}{Q_{\text{net}}} = \frac{1}{Q_{\text{anchor}}} + \frac{1}{Q_{\text{TED}}} + \frac{1}{Q_{\text{viscous}}} + \frac{1}{Q_{\text{AKE}}} \tag{9}$$

Q_{anchor}: Clamping structures known as tethers are required to physically support the free-standing square plate to the underlying substrate. As the square plate vibrates, some of the energy propagates out through the supporting tethers and into the substrate where they are lost. This is referred to as anchor loss and is accounted for by Q_{anchor}.

Q_{TED}: As a solid deforms, regions in tension become cooler, while regions in compression become hotter. This temperature gradient results in irreversible heat flow. The energy loss due to this mechanism is known as thermoelastic damping (TED) and is accounted for by Q_{TED}. In other words, the driving mechanism for TED is a lack of thermal equilibrium associated with thermal gradients resulting from a strain field.

Q_{viscous}: As the square plate deforms, fluid between the capacitive gaps is squeezed and results in resistance to the vibration. The loss associated with squeezing fluid through the gaps and friction between the solid and air is accounted for by Q_{viscous}. Q_{viscous} is related to the pressure in which the resonator is operating in. For silicon Lamé mode resonators up to tens of MHz, it has been shown that Q increases as the pressure is reduced until about a Torr,



Lamé Mode MEMS Resonators, Fig. 5 Stick diagram illustrating two geometries of annexed tethers connecting the resonator to the anchors: (**a**) T-shaped tether and (**b**) beam tether

beyond which the pressure appears to have little effect on Q [9]. Hence in operating a Lamé mode resonator at pressures much lower than 1 Torr, Q_{viscous} is no longer a dominant factor in determining the actual Q.

Q_{AKE}: This represents the losses due to phonon-phonon interactions and constitutes the fundamental limit on Q for a given material. For silicon, Q_{AKE} is bound by the fQ product in the order of 10^{13} below 1 GHz resonance [11, 12].

At room temperature, anchor loss does not appear to be the limiting factor on Q. The Lamé mode has five nodes in total: four nodes from each of the corners of the square plate and the last node at the center of the square plate. As such, the square plate can be conveniently clamped at the corners with little interference to the Lamé mode. Various studies have been carried out to assess the effect of the tether design on Q. The most common tether designs include beam tethers and T-shaped tethers as illustrated in Fig. 5. Keeping the size of the square plate the same but supporting the resonator by a T-shaped tether or beam tether, experimental results have suggested that no significant difference in Q between the two tether designs can be observed [13]. The results imply that the shape of the tether has no effect on Q. In a more recent study which focused on variations made to the T-shaped tether, it was found that changing the geometry of the T-shaped tether had no impact on Q either [14]. Taken together, the experimental results suggest that anchor loss does not play a dominant role in determining Q at room temperature.

With respect to TED, it can be shown that the condition for TED is absent in the Lamé mode. This can be seen by considering the volumetric changes caused by the strain field related to the Lamé mode. The fractional change in volume at a given location on the square plate can be found by summing the strains in all three directions. Assuming thin plate theory, the strain along the thickness of the plate can be neglected, leaving only the lateral strains. As seen from Eq. 2a, the lateral strain components are equal and opposite everywhere in the square plate for the Lamé mode. The result is zero volumetric change everywhere in the square plate. As such, the thermal gradients that give rise to TED are absent in the Lamé mode. This isochoric characteristic of the Lamé mode leads to no TED in theory [15]. However, as we will see in the following section, introducing holes into the square-plate annuls this isochoric property of the Lamé mode and consequently introduces TED.

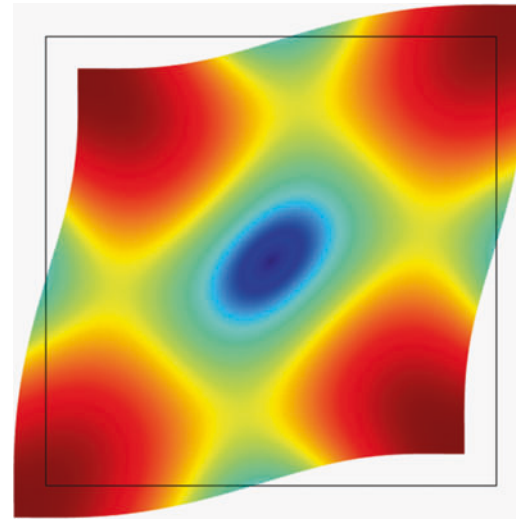
Effect of Etch Holes on *Q*

In the fabrication process of bulk mode resonators in SOl technology, etch holes are often employed. The etch holes are introduced to facilitate the timed HF etch of the buried oxide (BOX) layer underneath the bonded silicon device layer that

defines the resonator structure. Undercutting the structure by etching the BOX releases the resonator from the handling substrate. It has been found that adding etch holes to an otherwise plain square-plate resonator can massively reduce Q by over 95 % [16]. Through the use of finite element (FE) simulations, it has been found that the cause of this drop is TED. The Lamé mode shape is uniquely isochoric, which means that the fractional change in volume is zero at every point of the square plate. This can be shown from Eqs. 2a and 2b, whereby adding up the strain in the x-direction and y-direction at any given point on the square plate yields zero as the answer. Hence in theory, there is no TED in the Lamé mode. By FE analysis, it can be shown that adding holes all over the square plate results in a nonzero sum of x-strain and y-strain manifested as adjacent variations of expansion and contraction. By solving coupled thermoelastic equations using FE analysis, Q_{TED} can be computed for various hole patterns and compared against the measured values of Q reported by Shao [17]. Interestingly, the deviation between the measured Q and FE computed Q was found to be larger where the holes were concentrated in parts of square plate as compared to all over the square plate, suggesting that TED is accompanied by other sources of dissipation.

Effects of Crystal Orientation

Single-crystal silicon is anisotropic with respect to its mechanical properties. On an n-type (100) wafer, a square plate is aligned to the <110> axis when the sides of the plate are parallel to the primary flat. Rotating the same square plate 45° relative to the primary flat aligns the resonator to the <100> axis. Changing the orientation changes the resonant frequency of the Lamé mode MEMS resonator due to the change in the associated shear modulus. With reference to the stiffness matrix for single-crystal silicon, the shear modulus for a square plate aligned to the <110> axis is defined by C_{44}. In the <100> axis, the shear modulus is defined by $(C_{11}–C_{12})/2$ [18]. Apart from differences in resonant



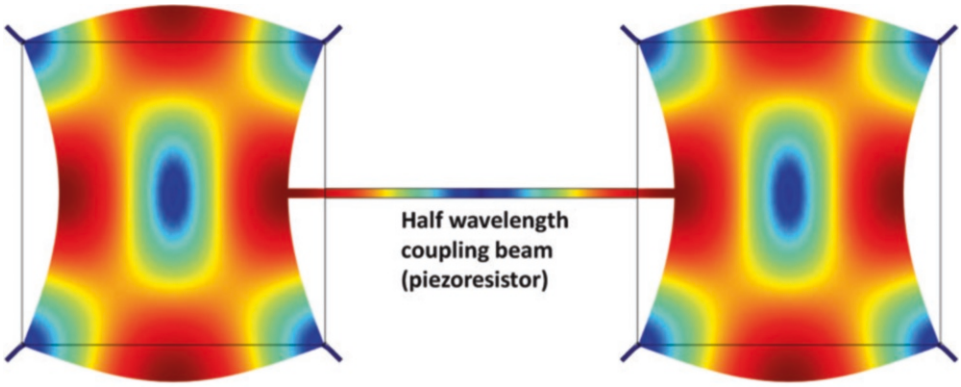
Lamé Mode MEMS Resonators, Fig. 6 Illustration of the face shear mode shape

frequency, the choice of alignment for a silicon Lamé mode resonator to the <110> or <100> axis makes a difference to both its power handling limit and the nature of nonlinear behavior. The effect of orientation has been observed in lateral shear modes, namely, the Lamé mode and the face shear (FS) mode (illustrated in Fig. 6). Both rely on the shear modulus to determine the spring constant of the resonant mode, but their relation to the crystal orientation properties is exactly opposite. As such the nature of the observed nonlinear behaviors in relation to crystal orientation has been found to be opposite. This is summarized in Table 1 based on the measurement results of 800 μm wide square plates driven sufficient hard to observe hysteresis in their electric frequency response.

The observed opposite nature of nonlinearity in relation to crystal orientation is best explained by the intrinsic nonlinear moduli of silicon itself. Geometric nonlinearities alone are unable to explain the difference if the material properties were assumed to be unaffected by crystal orientation because then the resulting nonlinear behavior would be of the same nature (either hardening or softening in both cases). If the nonlinearities were attributed to the capacitive gap transducer, then it

Lamé Mode MEMS Resonators, Table 1 Summary of nonlinear behavior of lateral shear modes in relation to crystal orientation

	Lamé		Face shear (FS)	
Orientation of edge alignment	Shear modulus	Nonlinearity	Shear modulus	Nonlinearity
<110>	C_{44}	Spring softening	$(C_{11}–C_{12})/2$	Spring hardening
<100>	$(C_{11}–C_{12})/2$	Spring hardening	C_{44}	Spring softening



Lamé Mode MEMS Resonators, Fig. 7 Mode shape of a pair of coupled Lamé mode resonators where the coupling beam in the middle acts as piezoresistor through which a bias current flows through. The Lamé mode is detected through measuring changes in the resistivity seen as a modulated current

should be noted that capacitive gap transducers only produce spring softening effects [19].

Mechanically Coupled Lamé Mode MEMS Resonators

Despite very high *Q*s even in the millions possible in Lamé mode MEMS resonators, unless nanoscale gaps can be reliably fabricated, the poor electromechanical transduction based on capacitive drive and sense leads to high motional resistance. As an alternative readout method, the piezoresistive property found in single-crystal silicon can be used as means to convert the lateral deformations into an output motional current via strain fields. On this note, piezoresistive sensing as a readout method has been shown to be highly effective for boosting the transduction of silicon resonators relative to capacitive sensing. However, piezoresistive sensing is not an immediately favorable option for the Lamé mode when one considers its associated stain field. The Lamé vibration mode shape comprises regions of tension and compression, thus greatly reducing the net change in stress required for piezoresistive sensing to be effective. One way of exploiting piezoresistivity in the case of the Lamé mode is to couple a pair of identical Lamé mode resonators mechanically and use the coupling spring as the piezoresistive strain gauge [20] as depicted in Fig. 7. The coupling spring is a longitudinal free-free beam with a length that is half the acoustic wavelength of the Lamé mode resonators. This mechanical coupling design, together with a biasing configuration used to apply a desired combination of electrostatic forces, helps synchronize the coupled pair of square-plate resonators to resonate in phase as illustrated in Fig. 7 at the same frequency.

The plates are excited to vibrate in the Lamé mode just like in a single square plate through capacitive gap electrodes. As the pair of square-plate resonators resonates in phase, the

mechanical coupling beam compresses or elongates accordingly like a longitudinal free-free beam resonator. This has been confirmed by FE simulations, which shows that the deformation profile follows a sine function that corresponds to half of the acoustic wavelength. It has been shown that transmission peak magnitude increases linearly with the bias current level, in agreement with theory. The Q of the coupled resonator pair was a million, which is 20 % lower than the measured Q for a single Lamé mode resonator. But in terms of transduction, compared to a single Lamé mode resonator, the piezoresistive coupled resonator pair produced a peak resonance that was 28 dB higher with 14 mA bias current despite the slight drop in Q [20]. It has also been confirmed experimentally that the increase in transduction is due to the piezoresistive effect through the coupling beam and not the square plates [21]. This device concept also presents an interesting frequency scaling characteristic in that the electromechanical coupling efficiency is independent of the size of the square plates.

Cross-References

- ▶ Anchor Loss in MEMS/NEMS
- ▶ Bulk Acoustic Wave MEMS Resonators
- ▶ CMOS-MEMS Resonators

References

1. Basrour, S., Majjad, H., Coudevylle, J.R., de Labachelerie, M.: Simulation and characterization of high Q microresonators fabricated by UV-LIGA. In: Proceedings of the 4th International Conference on Modeling and Simulation of Microsystems, Hilton Head Island, 19–21 Mar 2001
2. Ekstein, H.: Free vibrations of anisotropic bodies. Phys. Rev. **66**(5–6), 108–118 (1944)
3. Lee, J.E.-Y., Seshia, A.A.: Parasitic feedthrough cancellation techniques for enhanced electrical characterization of electrostatic microresonators. Sensors Actuator A Phys. **156**(1), 36–42 (2009)
4. Xu, Y., Lee, J.E-Y.: Differential-input piezoresistively-sensed square-extensional mode resonator for parasitic feedthrough cancellation. In: Proceedings of the 16th International Conference on Solid-state Sensors, Actuators and Microsystems, Beijing, 5–9 June 2011
5. Bhave, S.A., Gao, D., Maboudian, R., Howe, R.T.: Fully-differential poly-SiC Lamé-mode resonator and checkerboard filter. In: Proceedings of the 18th IEEE International Conference on Micro Electro Mechanical Systems, Miami, 30 Jan–3 Feb 2005
6. Xu, Y., Lee, J.E-Y.: Feedthrough cancellation in micromechanical square resonators via differential transduction. In: Proceedings of the 2010 I.-E. International Conference on Electron Devices and Solid-state Circuits, Hong Kong, 15–17 Dec 2010
7. van Beek, J.T.M., Puers, R.A.: A review of MEMS oscillators for frequency reference and timing applications. J. Micromech. Microeng. **22**(1), 013001 (2012)
8. Lee, J.E.-Y., Yan, J., Seshia, A.A.: Low loss HF band SOI wine glass bulk mode capacitive square-plate resonator. J. Micromech. Microeng. **19**(7), 074003 (2009)
9. Khine, L., Palaniapan, M.: High-Q bulk-mode SOI square resonators with straight-beam anchors. J. Micromech. Microeng. **19**(1), 015017 (2009)
10. Lee, J.E-Y., Seshia, A.A.: Square wine glass mode resonator with Q of 4 million. In: Proceedings of the 7th IEEE Conference on Sensors, Lecce, 26–29 Oct 2008
11. Ghaffari, S., et al.: Quantum limit of quality factor in silicon micro and nano mechanical resonators. Sci. Rep. (2013). doi:10.1038/srep03244
12. Chandorkar, S.A., et al.: Limits of quality factor in bulk-mode micromechanical resonators. In: Proceedings of the 21st IEEE International Conference on Micro Electro Mechanical Systems, Tucson, 13–17 Jan 2008
13. Lee, J.E.-Y., Yan, J., Seshia, A.A.: Study of lateral mode SOI-MEMS resonators for reduced anchor loss. J. Micromech. Microeng. **21**(4), 045010 (2011)
14. Xu, Y., Lee, J.E-Y.: Evidence on the impact of T-shaped tether variations on Q factor of bulk-mode square-plate resonator. In: Proceedings of the 7th IEEE International Conference on Nano/micro Engineered and Molecular Systems, Kyoto, 5–8 Mar 2012
15. Chandorkar, S.A., et al.: Multimode thermoelastic dissipation. J. Appl. Phys. **105**(4), 043505 (2009)
16. Tu, C., Lee, J.E.-Y.: Thermoelastic dissipation in etch-hole filled Lamé bulk-mode silicon microresonators. IEEE Electron Device Lett. **33**(3), 450–452 (2012)
17. Shao, L., Palaniapan, M.: Effect of etch holes on quality factor of bulk-mode micromechanical resonators. Electron. Lett. **44**(15), 938–939 (2008)
18. Zhu, H., Lee, J.E-Y.: Orientation dependence of nonlinearity and TCf in high-Q shear-modes of silicon MEMS resonators. In: Proceedings of the 2014 I.-E. International Frequency Control Symposium, Taipei, 19–22 May 2014
19. Zhu, H., Lee, J.E.-Y.: Reversed nonlinear oscillations in Lamé-mode single crystal silicon microresonators. IEEE Electron Device Lett. **33**(10), 1492–1494 (2012)
20. Xu, Y., Zhu, H., Lee, J.E-Y.: Piezoresistive sensing in a strongly-coupled high Q Lamé mode silicon MEMS resonator-pair. In: Proceedings of the

2014 I.E. International Frequency Control Symposium, Taipei, 19–22 May 2014
21. Zhu, H., Xu, Y., Lee, J.E.-Y.: Piezoresistive readout mechanically-coupled Lamé mode SOI resonator with Q of a million. J. Microelectromech. Syst. **24**(4), 771–780 (2015)

Lamé Mode Resonators

► Lamé Mode MEMS Resonators

Laminates: Multilayered Structures

► Implementing Alternating Nanolaminates in Trenched Energy Storage Systems

Landmark Detection

► Waypoint Detection

Large Unilamellar Vesicle (LUV)

► Liposomes

Laser Beam Machining

Richa Agrawal and Chunlei (Peggy) Wang
Department of Mechanical and Materials Engineering, Florida International University, Miami, FL, USA

Synonyms

Direct laser writing; Laser cutting; Laser machining; Laser micromachining; Laser scribing; Laser writing; Multiphoton fabrication; Two-photon lithography

Definition

Laser beam machining (LBM) is a nonconventional machining process, which broadly refers to the process of material removal, accomplished through the interactions between the laser and target materials. The processes can include laser drilling, cutting, grooving, writing, scribing, ablation, welding, cladding, milling, and so on. LBM is a thermal process, and unlike conventional mechanical processes, LBM removes material without mechanical engagement. In general, the workpiece is heated to melting or boiling point and removed by melt ejection, vaporization, or ablation [1]. The high degree of monochromaticity enables the laser beam to be focused to very precise locations and makes it extremely useful for machining with high precision. Depending upon the feature size of the laser tool, laser machining can be categorized into micromachining and macromachining. Typically, the laser systems used for micromachining are normally pulsed beams with an average power lower than 1 kW, whereas the ones for macromachining are continuous wave laser beams ranging up to several kW [2]. The applications of laser beam machining include manufacturing processes such as laser cutting, drilling, welding, ablation, and surface texturing. Since laser machining processes are generally thermal in nature, varied laser durations result in different thermal effects; for instance, shorter laser pulses in the order to femtoseconds help in minimizing thermal effects such as melting and burr formation and therefore minimizing the after-process measures [2]. For polymers, however, laser machining processes are more light driven than heat driven, and therefore, the process is more polymerization based. Polymers are of wide interest in the microfabrication and biomedical industries and contribute a large portion in the relatively new but very popular laser micromachining technique called *direct laser writing* also known as *two-photon lithography* or *multiphoton lithography*. The latter is a maskless technique and employs solid-state lasers in conjunction with galvanostatic scanners and focusing optics in order to minimize the beam diameter and

directly use the laser beam to "write" the pattern onto the material. Direct laser writing is based on nonlinear absorption of light direct toward a tight volume of photosensitive materials; the absorption of light leads to local photopolymerization. The beam can be moved in order to generate the desired pattern. The redundancy of the use of photomasks and complex optical systems makes direct laser writing very popular for microfabrication.

Laser Beam Machining Fundamentals

Introduction to Lasers

The history of lasers dates back to early twentieth century, when Planck propounded the "quantum theory of light." The theoretical foundation of laser was led by Einstein in 1917, using Planck's laws of radiation and probability coefficients for absorption and spontaneous and stimulated emission of electromagnetic radiation. In 1953, Charles Town along with graduate students produced what was known as the "maser (microwave amplification by stimulated emission of radiation)," which was essentially a microwave amplifier and similar to laser but used microwaves instead of visible or infrared light. The first practical laser was demonstrated by Theodore Maiman in 1960. Maiman's laser was based on optical pumping of a synthetic ruby crystal and generated a pulsed red radiation at 694 nm [3]. Later in 1960, Javan, Bennett, and Heriott developed the first gas laser that used a mixture of helium and neon. The first diode laser was developed by Robert Hill; the laser was made of gallium arsenide and emitted a near-IR radiation of 850 nm. Much advancement has been made in laser technology since then.

Features of a laser beam: A laser beam unlike conventional light is *monochromatic*, *directional*, *coherent*, and *bright*. The features of a laser beam are briefly described below:

Monochromaticity – The term monochromatic refers to a "single color" and for light refers to a single wavelength or frequency. In practice, however, no light is ideally monochromatic including a laser beam; nonetheless, a laser is considered to be monochromatic because the laser consists of very close, discrete, and narrow spectral lines [1].

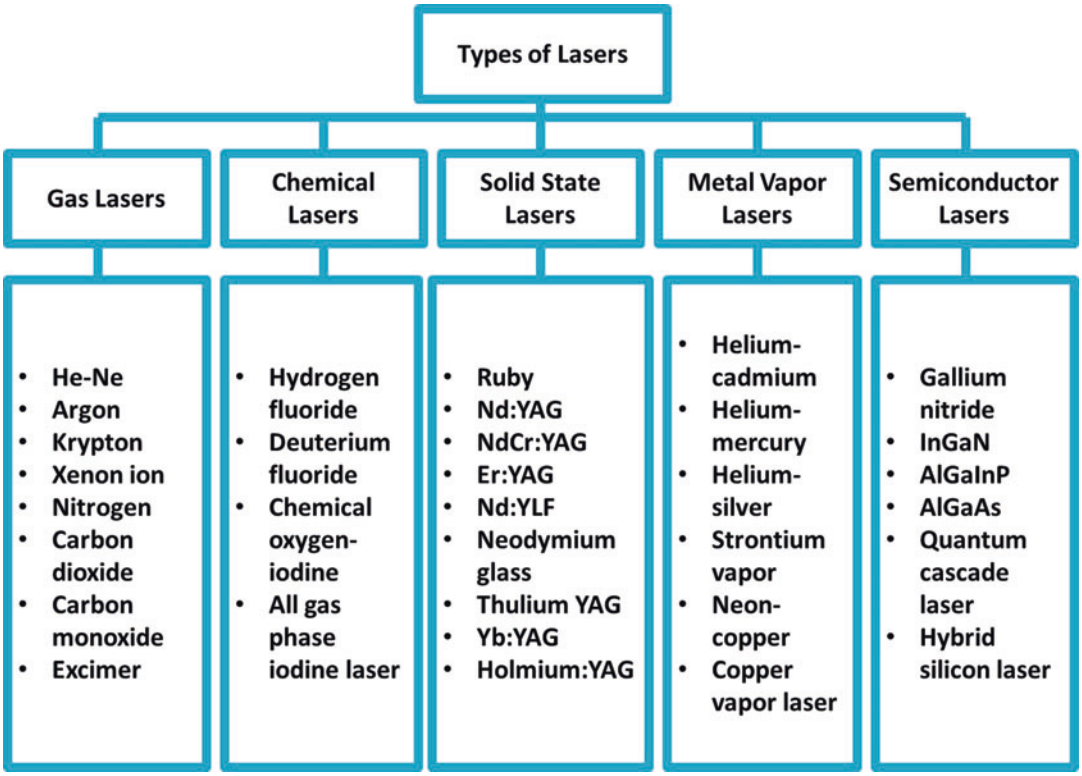
Directionality – The laser beam has a very small divergence angle, which enables the beam to be focused to a very small spot over a long distance. Owing to the striking directionality of laser beams, the output of a laser beam is almost parallel and can therefore be used to transmit data for diagnostic and communication purposes.

Coherence – The laser beam is coherent because there exists a fixed phase relationship between electric fields at different locations across the beam profile (spatial coherence) and between electric fields at one location at different times (temporal coherence).

Brightness – Brightness measures the ability of a laser oscillator to emit high optical power per unit area per unit solid angle [1]. Since the brightness of the laser beam is inversely proportional to the angle, the smaller the divergence angle, the brighter the laser beam.

Types of lasers used for laser beam machining: Gas lasers such as CO_2 and excimer lasers, along with solid-state lasers such as Nd:YAG and YAG lasers and femtosecond lasers, are some of the most popular lasers. The active media of *solid-state lasers* is obtained by embedding transition metals (Ti, Nd, Yb, Er, etc.) into insulating host lattices. The energy levels of the active ions are responsible for lasing actions, while the thermal and physical properties of the host material are crucial for determining the efficiency of the laser operation [3]. In Nd:YAG (neodymium-doped yttrium aluminum garnet) lasers, Nd^{3+} replaces a tiny fraction of yttrium ions in the host crystal of YAG.

Gas lasers are divided into three main categories: atomic (He:Ne), ionic (Ar ion gas laser), and molecular (CO_2 gas laser, nitrogen gas lasers, excimer lasers), depending upon the nature of active medium. These lasers are pumped by *electrical discharge*. After the passage of discharge current, the electrons are accelerated between the electrodes and thereafter collide with the entities



Laser Beam Machining, Fig. 1 Different types of widely used lasers

(atoms, ions, molecules) in the active medium; this causes the transitioning of electrons to higher-energy levels in order to achieve population inversion and stimulated emission [3]. The first gas laser was the He:Ne (helium-neon) gas laser invented by Ali Javan and William R Bennett in 1960.

– The *carbon dioxide laser* is one of the most widely used gas lasers that operates at the infrared wavelength (9.4–10.6 μm) and can provide power in the order of hundreds of kW. It is typically used in the continuous wave mode and is mostly used in industrial applications such as cutting, welding, drilling, engraving, etc.
– *Excimer lasers* – Excimer lasers utilize excimer molecules such as ArF, KrF, XeCl, etc.; essentially an excimer is a combination of rare excimer lasers that operate in the ultraviolet region by a fast electrical discharge in a high-pressure mixture of a rare gas and a halogen gas. The combination of the two determines the final wavelength of the laser. The typical pulse energies for excimer lasers range from a few hundred millijoules to 1 J. The average power ranges from 10 W to 1 kW. Applications of excimer lasers include lithography, ablation, and microfabrication.

Femtosecond lasers – A femtosecond laser is an ultrafast laser emitting optical pulses with a duration well below one picosecond (with 1 ps $= 10^{-12}$s), i.e., in the domain of femtoseconds (with 1 fs $= 10^{-15}$s). The generation of such short pulses is achieved with the technique of mode locking. Methods for producing mode locking in a laser may be classified as either an active method using an external signal to induce optical modulation or a passive method using self-modulation of the light via saturable absorbers. Femtosecond lasers are widely used in direct laser writing.

Figure 1 categorizes different types of widely used lasers. Other types of lasers include X-ray

lasers, dye lasers, nuclear pumped lasers, electron lasers, etc.

Modes of laser operation: Depending upon the desired application, lasers can be operated in different modes. Typically for laser macromachining, *continuous* high-power laser sources are required, while for micromachining, *shorter pulses* are required, and therefore broadly laser modes can be classified into (i) continuous and (ii) pulse generation modes. *Continuous wave* mode operation of laser essentially means that the laser is being pumped continuously and therefore emits light continuously. Laser pulses can on the other hand be generated by using gain switched, Q-switched, and mode-locked modes; *gain switching* is regarded as the most direct method to generate laser pulses. Upon the initiation of pumping, the population inversion builds up, and the laser starts to oscillate once critical inversion is reached. The process continues until pumping is switched off. Pulse durations typically vary from 10 to 10 ms [4]. *Q-switching*, on the other hand also known as giant pulse formation, is used to produce a pulsed output beam usually with extremely high peak power (of the order of gigawatts). This technique provides a single strong and a short pulse of laser radiation [3]. *Mode locking* is used to generate ultrashort laser pulses in the picosecond (10^{-12} s) to femtosecond (10^{-15} s) range. This technique relies on production of a fixed phase relationship between the longitudinal modes of the laser resonator. Mode locking of lasers can be achieved by active (uses an optical modulator) and passive (uses an absorber) mode-locking techniques.

Types of Laser Beam Machining Processes

Laser Welding

Welding processes that utilize lasers as heat sources are known as *laser welding*. Essentially welding process is a "joining" or "fusion" process of similar or dissimilar materials. Laser beam welding is a versatile technique that finds applications in a host of industries including the aerospace, automotive, biomedical, and electronics. In the welding process, the laser beam is directed toward the work material, whereupon the laser energy is converted into heat energy and surface melting occurs and therefore laser welding is a "pyrolytic" process, where the light energy from the laser beam is converted into thermal energy as a result of photon-electron interactions upon the incidence of the laser on the workpiece. Depending upon the temperature of the surface and the thermal characteristics of the workpiece material, (i) thermal stresses in the heat-affected zone (HAZ) can be generated, (ii) the surface of the material can melt, or (iii) the surface of the workpiece can vaporize. Laser welding is categorized into (i) *keyhole* and (ii) *conduction welding*. In keyhole welding, the laser is focused to achieve very high-power densities at the workpiece. The metal vaporizes at the center where the intensity of the laser is the highest, leaving behind a pit or a "keyhole" in the molten metal pool. Also, the metal vapors transfer the laser energy into the sides of the pit, resulting in a deep aspect ratio structure. Conduction welding on the other hand are shallower; in conduction weld processes, most of the laser energy is absorbed by the surface of the material, while the heat transfer occurs in the rest of the material by heat conduction and hence the name "conduction welding." The HAZ is typically larger for conduction welds as compared to the keyhole welds.

Depending upon the type of application, laser beam welding can be carried out in both continuous and pulsed modes. Solid-state lasers (Nd:YAG) and gas lasers such as CO_2 lasers are widely popular as laser sources for welding processes. Despite their longer IR wavelength, which is typically not absorbable by metals, CO_2 lasers are widely used in laser welding given their ability to generate directional beams with very high powers. The high-intensity beams with smaller spot size particularly are suited for keyhole welding. CO_2 lasers find application in industrial welding of glass, aluminum alloys, and also welding of disparate materials. Pulsed Nd:YAG lasers, on the other hand, produce energy in the order of a few millijoules per pulse and are popular for welding of metals including stainless steel and aluminum alloys and also for spot welding applications. Laser welding is also popular in the microdomain. Laser microwelding offers a

number of advantages over conventional joining methods: (i) the outstanding control of laser allows for the weld energy to be delivered only where it is required; (ii) laser provides stronger welds with smaller number of welds; (iii) the welds are aesthetically pleasing; and (iv) lasers provide better hermetic welds which are desired for microapplications [5]. In addition to Nd:YAG lasers, fiber lasers are also highly popular for microwelding purposes.

Laser Ablation

Laser ablation, in the broadest sense, refers to a process of removal of the material with the aid of a laser beam as the energy source; the material is removed by vaporization or sublimation upon the incidence of the laser beam. Ablation mechanisms are distinguished into *photolytic* and *pyrolytic* processes. In *photolytic processes* the chemical bonding energy is overcome by the use of light energy. Photolytic ablation is observed in the case of polymers where the large polymeric molecules are broken into smaller monomers. When all the laser energy is used to overcome the binding energy, the ideal case is known as *cold ablation*. However, in the case of metals, the laser energy is first absorbed by the metal electrons and transferred into thermal energy which can thereafter induce thermal stresses, melt the surface, or evaporate the surface depending upon the surface temperature and material properties.

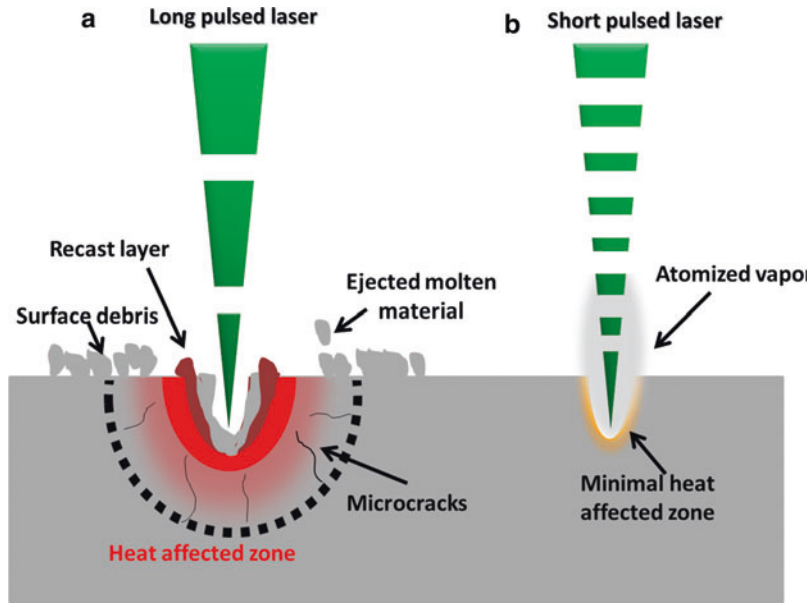
Ablation of organic polymers: As mentioned earlier, ablation of organic polymers is a photolytic process, which means that the energy from the light is enough to overcome the energy "gluing" the atoms together. The relationship between energy and frequency (Planck-Einstein's relation) is given as $\boldsymbol{E = h\nu = hc/\lambda}$, where E is the energy, h is the Planck's constant, v is the frequency, c is the velocity of light, and λ is the wavelength. For organic polymers, the bond strengths of O–H, C–H, and C–C are 4.7 eV, 4.25 eV, and 3.7 eV, respectively. These bond energies can easily be overcome by excimer lasers such as KrCl (225 nm ~ 5.5 eV) and ArF (193 nm ~ 6.42 eV). The removal of these materials is a three-step process: (i) the laser photons are absorbed at a characteristic depth of ~0.2 μm [4]; (ii) longer chains of polymers are broken into shorter chains; and (iii) the material is removed in the form of vapors or small particles. In addition fluency of the laser beam must be controlled in order to obtain finer ablation features.

Ablation of metals: Laser ablation of metals is a pyrolytic or a photothermal process. Given the high thermal conductivities and low melting temperatures of metals, microstructuring of metals is rather challenging with conventional laser pulses longer than *1 ns*, metal ablation is accompanied with large heat-affected zones (HAZ) and molten throw-outs, and therefore machining metals with long pulsed lasers becomes a hassle as the achievable quality and precision is quite inferior. Figure 2 shows a qualitative difference between using long laser pulses and short laser pulses for machining purposes. A significant improvement in metal ablation can be realized by using ultrashort laser pulses. Ti/sapphire femtosecond lasers are quite popular for metal ablation [6].

Ablation of diamond: Diamond is a hard to process material, given its transparency in a wide range of wavelengths. However at high-power densities, diamond transitions into graphite, which absorbs the laser power and is thereafter subsequently removed by ablation. Diamond machining is achieved by using microsecond Nd:YAG lasers and nanosecond excimer lasers. The layers of graphite or amorphous carbon that are found on the diamond surface post machining can be removed via polishing [4].

Laser Cutting

As suggested by the name, laser cutting is a technology that employs a laser as the cutting tool. The process involves a high-power laser to cut materials based on a computer-based design. Upon the incidence of the laser, the material is removed by burning, vaporizing, or melting. The advantages of laser cutting over conventional mechanical cutting include the noninvolvement of a mechanical saw to cut the material, which significantly reduces contamination. Moreover since the laser beam does not wear, laser cutting gives high precision. Depending upon the application, CO_2 lasers



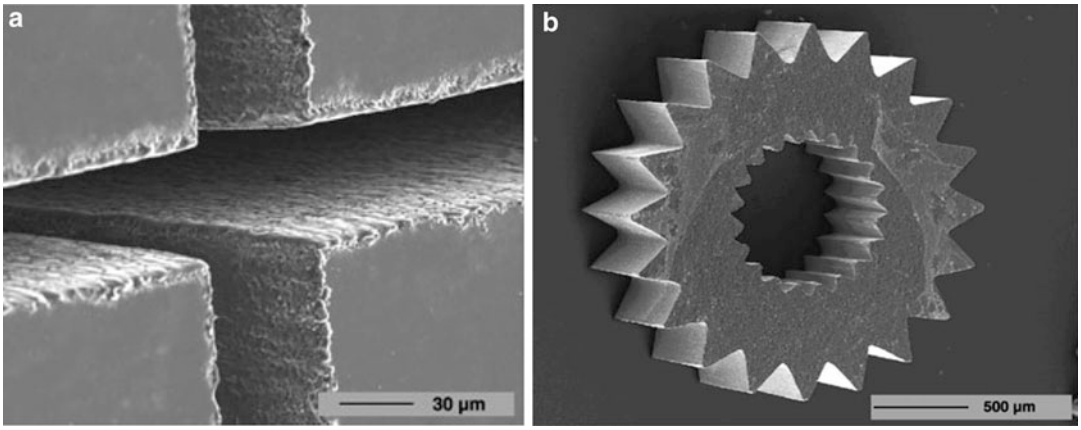
Laser Beam Machining, Fig. 2 An illustration of heat-affected zones (*HAZ*) created by using (**a**) a long pulsed laser and (**b**) a short pulsed laser. Notice that the HAZ is much more pronounced for the longer pulse since the relatively longer period of laser irradiation allows for the conduction of heat resulting in large thermal stresses eventually leading into cracking, generation of surface debris from the solidification of the melt ejection, and also voids. The short pulsed laser, on the other hand, has a much smaller HAZ because the laser pulse time is short for thermal transport, leaving behind a much precise and aesthetic structure

and Nd:YAG lasers are very popular for cutting purposes. Nd:YAG lasers are more popular when higher power is required.

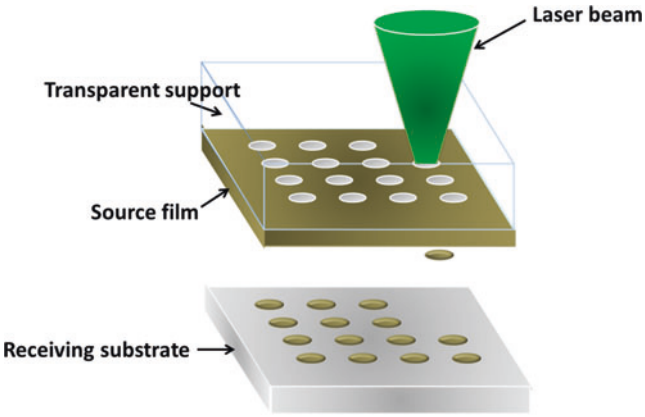
Laser cutting of silicon: Silicon micromachining is extremely popular owing to the dominance of silicon-based devices in microelectronics and medical industries. Silicon in the form monocrystalline wafers dominates the microelectronics and semiconductor industries. Both mono- and polycrystalline silicon along with amorphous silicon are used in high-performance solar cells. High-precision structuring of silicon can be achieved by the use of diode pumped Q-switched Nd:YAG lasers. The short pulses and high-energy short wavelengths allow for alleviated thermal effects including molten material deposition and HAZ at the edges [2]. Shorter laser pulses ensure reduced heat-affected depth and melting and also prevent prolonged and intense plasma heating. The high-energetic short wavelengths Nd:YAG lasers allow for precision and quality structuring of silicon. The accuracy of structuring depends on wavelength, pulse duration, intensity of beam spot, and beam guiding technique. Nd:YAG lasers have been used for cutting silicon in various geometries and cuts. As noted by Klotzbach et al. [2], using a combination of appropriate solid-state laser along with a flexible and well-controllable scanner, suitable structures for micromechanics and microfluidics can also be obtained. Figure 3 shows some of the silicon structures created with UV laser (355 nm); the clean edges in Fig. 3b were obtained with hydrofluoric acid treatment [2].

Laser-Induced Forward Writing

Laser-induced forward writing abbreviated as LIFT uses a laser beam and employs an optically transparent support in order to transfer a thin film to a substrate placed 25–100 μm below the film. Figure 4 demonstrates a schematic of the LIFT technique. The technology of LIFT was introduced by Bohandy et al. [7] in 1986 to deposit copper and silver over silicon and fused silica [8]. The use of the laser pulse leads to vaporization of the material that recondenses on the



Laser Beam Machining, Fig. 3 (**a**) Cross cuts in 550 µm silicon wafer (**b**) toothed gear created with a UV laser [2]

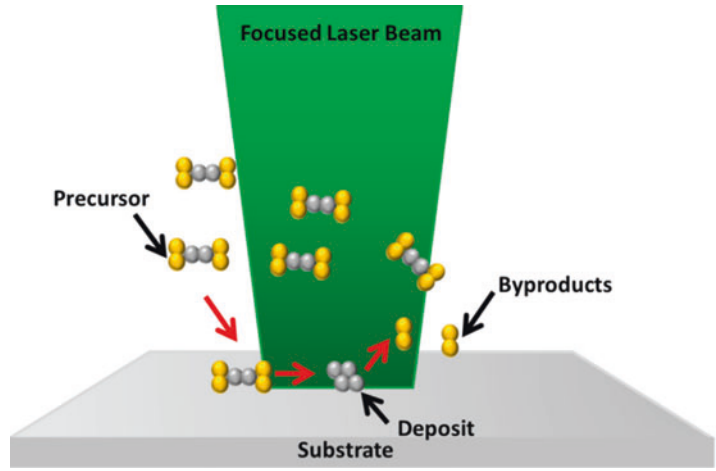


Laser Beam Machining, Fig. 4 Schematic of laser-induced forward transfer (*LIFT*) process. The image on the right shows the final pattern

substrates. The technique of LIFT is popular owing to the simplicity and versatility which enables processing of a wide variety of materials and also allows for patterning of structures with a wide range of geometries.

One of the applications of LIFT is the transfer of transparent materials. Majority of materials are transparent in the visible-IR range, which makes their laser machining difficult; however, the transfer of transparent materials becomes feasible using high-intensity femtosecond lasers which takes place as the two-photon absorption phenomenon. LIFT has shown to be appropriate for patterning a variety of materials including metals such as gold, aluminum, and titanium, oxides such as alumina and ZnO, and high-temperature superconductors [9].

Typically conventional LIFT technique makes use of high-intensity ultrashort laser pulses which may not be appropriate for biological applications as the high-intensity lasers can be harmful for cells, proteins, or other biomolecules. This shortcoming is overcome by pre-coating of the donor substrate with a layer, which is known as the dynamic release layer (DRL), also known as the sacrificial layer. Metals such as titanium, gold, platinum, or chromium are mostly used as DRLs. UV-sensitive photopolymers such as triazene polymers (TP) have also been studied as sacrificial layers [8]. Therefore, the application of LIFT is highly suitable in biomedical fields; the applications include printing of soft materials for biosensors and sensors. Other applications include transfer and printing of liposomes, DNA, etc.



Laser Beam Machining, Fig. 5 Schematic representation of laser chemical vapor deposition (*LCVD*) process

Matrix-assisted pulsed laser evaporation direct write: Matrix-assisted pulsed laser evaporation direct write (MAPLE DW) utilizes laser ablation in order to transfer material from one substrate to another. MAPLE DW is widely used to fabricate electronic devices which utilize matrices containing metals and dielectric or resistive components and also to pattern 3D structures such as chemical-sensitive materials, proteins, antibodies, cells, etc. The MAPLE DW is widely used to fabricate electronic devices which utilize matrices containing metals and dielectric or resistive components and also to pattern 3D structures such as chemical-sensitive materials, proteins, antibodies, cells, etc. MAPLE DW was originally developed by the US Naval Research Laboratory for fabricating electronic devices and later was extended to biological applications [10]. For biological applications, MAPLE DW has been used to transfer hamster ovaries, human osteoblast cells, etc. Patz et al. [11] used MAPLE DW to pattern B35 neuroblast-like cells onto and within polymerized extracellular matrix substrates.

Laser Chemical Vapor Deposition

Laser chemical vapor deposition abbreviated as LCVD is a popular laser direct writing process. Essentially in an LCVD process, a laser beam is used to heat the substrate locally and induce chemical vapor deposition reaction. Chemical vapor deposition (CVD) is a technique that is used to deposit coatings or for growing layers on a substrate and is widely used in the semiconductors, optics, and refractory materials [12]. In the conventional CVD process, gaseous substances are passed over a heated substrate where the reactants either decompose or combine with other constituents to form a layer of the substrate [12]. The advantages of using a laser source are manifold. For instance, the use of laser offers high spatial control and also the availability of non-equilibrium heating and cooling rates. LCVD processes have been used since early 1980s.

LCVD processes are used for fabrication for both two- and three-dimensional structures. The applications of LCVD includes mask and circuit repair, fabrication interconnects, construction of ohmic contacts, localized coating, and also corrosion-resistant coatings [12]. Figure 5 shows the schematic of LCVD process. The directed laser beam locally increases the temperature of the substrate placed in a chamber that is filled with precursor gas. The precursor gas thereafter is adsorbed on the heated substrate and decomposes into the deposit and by-products.

The deposition rate of LCVD depends on factors such as gas pressure (linearly), power density of the laser (linearly), and the scanning speeds (inversely). For a Gaussian beam profile, the thickness of the laser deposited film is given as [8]

$$h(v_s, t) = \frac{\sqrt{\pi} R_0 r t}{\sqrt{\pi r} + 2(v_s, t)} \qquad (1)$$

where h(v_s,t) is the thickness of the deposited film, R_0 is the diffusion-limited growth rate, r is the radius of the laser beam, t is the time, and vs is the scan speed.

Direct Laser Writing

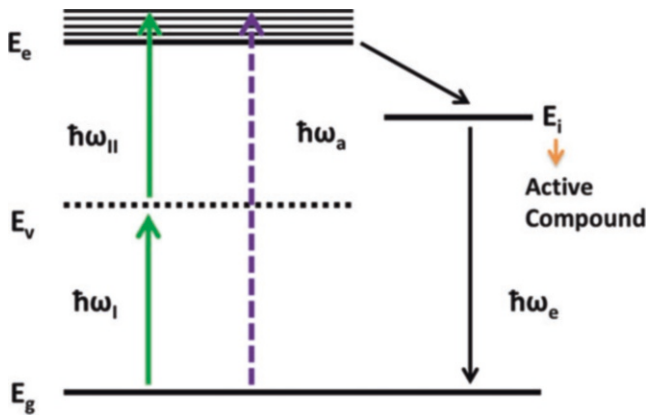
Multiphoton lithography, also known as direct laser writing, is a powerful three-dimensional printing technology with its ability to realize resolution sub 100 nm. The words multiphoton lithography and direct laser writing are used interchangeably in this chapter. The history of direct laser writing dates back to the early 1980s with its first use in Lawrence Livermore National Laboratory and AT&T Bell Laboratories to create 1-2D features on microelectronic circuits. Later, further advancements were made in direct laser writing technique with the creation of 3D features for microapplications at Max Planck Institute and US Naval Research Laboratory [8]. Since then, much advancement has been made in direct laser writing. The process relies on nonlinear two- or multiphoton absorption; upon the incidence of ultrafast lasers, the photoactive material absorbs the photons in a nonlinear fashion. The absorption of light can initiate photochemical processes in the material including photopolymerization and photoactivation.

Two-Photon Absorption Figure 6 shows one- and two-photon absorption. Two-photon absorption was first predicted by Maria Goppert-Mayer in 1931 and was later verified with the detection of two-photon excited fluorescence that was observed in a europium-doped crystal. The two-photon absorption process can be extrapolated to three or more photons. E_g represents the ground state, and E_e represents the excited energy level. In order to reach a higher-energy level, the molecule can absorb one photon with energy $\hbar\omega_a$. According to the Frank-Cordon principle, the transition in the electronic states is faster relative to the correlated change in nuclear motion [13]. Therefore, the molecule undergoes internal conversion and reaches an intermediate energy level E_i, whereupon the molecule finally relaxes to the ground state releasing energy measuring $\hbar\omega_e$. The emitted energy $\hbar\omega_e$ is lower than the absorbed energy, and therefore the radiation emitted by the molecule upon relaxation is longer than the absorbed wavelength (red shift). Two-photon absorption on the other hand is the excitation of the molecule by absorption of two photons, and therefore in the two-photon absorption process, the energy difference between the ground state and excited state is given as

$$E_e - E_g = \Delta E = \hbar\omega_I + \hbar\omega_{II} \tag{2}$$

In the two-photon absorption process, a photon with energy $\hbar\omega I < \Delta E$ interacts with the molecule in the ground state and excites it two a virtual level with energy E_v. While the molecule is the in E_v state, a second photon with energy $\hbar\omega II$ is absorbed, and the molecule is transitioned into the final excited level E_e:

$$\hbar\omega_a = \hbar\omega_I + \hbar\omega_{II} \tag{3}$$



Laser Beam Machining, Fig. 6 Energy level diagram for single- and two-photon absorption processes. The dotted violet line shows molecule excitation by one photon, while the two green lines ($\hbar\omega_I$ and $\hbar\omega_{II}$) show two-photon absorption

The degenerate case, where $\hbar\omega_I = \hbar\omega_{II} = \frac{1}{2}\hbar\omega_a$, i.e., the two photons have similar energy, is of practical significance. In terms of technical application, this means only one laser source can initiate two-photon absorption.

Photopolymerization Polymerization processes refer to conversion of small monomers into large polymers through a series of chain reactions, and "photopolymerization," as the name suggests, is light-aided polymerization. The curing source is not limited to lasers operating in the visible, UV, or near-IR radiation range but can extend to x-rays, microwaves, and γ-rays. The starting materials are the monomers or oligomers in the liquid state. Upon light incidence, the monomers or oligomers can be solidified by either polymerization or cross-linking. The polymerization process occurs through a series of chain reactions, while in cross-linking, cross-links are formed with chemical bonds [14]. The distinction between the two processes is the quantum yield, which is defined as:

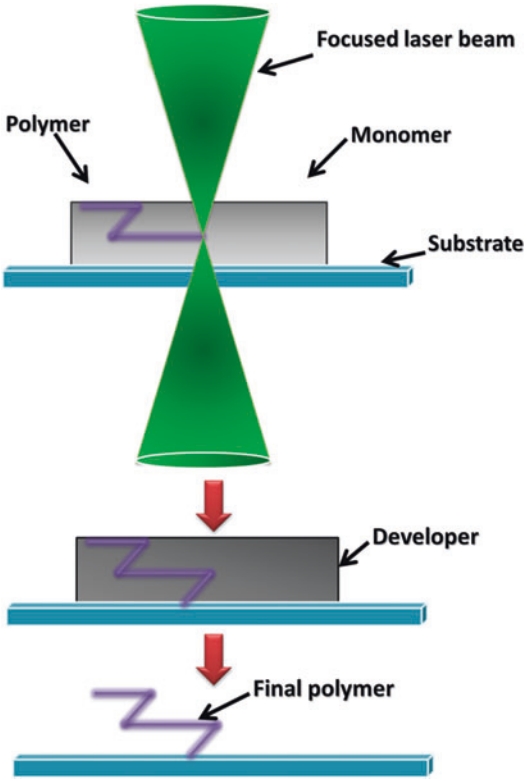
$$\text{Quantum Yield} = \frac{\text{Number of polymerized units}}{\text{Number of photons utilized}} \tag{4}$$

For cross-linking, the quantum yield is less than 1, since the addition of each monomer unit requires one photon, while in the polymerization process, the quantum yield can range in thousands since the mechanism relies on chain reactions [14]. Typically in order to increase the efficacy of initiation and enhance the quantum yield of photopolymerization, additives such as photoinitiators and photosensitizers are used.

A photoinitiator is a molecule that absorbs a photon and the forms initiating species. A photosensitizer on the other hand is a molecule that absorbs the photon energy and then transfers the energy to a photoinitiator. Depending upon the type of photoinitiator used, photoinitiation processes can be divided into radical initiated and cation initiated. The polymerization process involves various steps that include (i) photoinitiation, (ii) chain propagation, and (iii) termination. Majority of the photopolymerization processes utilize radical photoinitiators. An example of radical type polymerization process will be the double-bond addition of acrylates; ring opening of epoxides on the other hand will be an example of cationic-type polymerization [14].

Quite akin to the conventional photopolymerization processes, two-photon polymerization utilizes photoinitiator molecule excitations to induce chemical reactions between the photoinitiator molecules and the monomeric units [14]. The difference between the two processes is that the two-photon polymerization utilizes the two-photon absorption for the molecule excitation. As explained earlier, two photons are absorbed almost concurrently with the creation of a virtual second for several femtoseconds. Two-photon absorption is a nonlinear optical process meaning that the transition rate is a nonlinear function of the light intensity. Figure 7 shows two-photon laser polymerization process. Two-photon polymerization offers several advantages over conventional polymerization processes for microscale device fabrication. For instance, most materials are transparent to near-IR radiation which is desirable for two-photon polymerization processes. The fact that near-IR radiation is quite low in energy and does not harm biological cells when applied at low intensities makes two-photon polymerization particularly suitable to life sciences and biological applications. Other benefits include the ability of two-photon polymerization processes being able to set up in regular environment and not requiring cleanroom facilities and also their ability to produce feature structures similar to those fabricated using conventional photolithography, deep-reactive etching which require devoted cleanroom work. Moreover, two-photon polymerization does not require masks or molds; it has the ability to directly pattern computer-generated designs, unlike conventional lithography that is a mask projected technique.

Experimental Setup Figure 8 shows a typical experimental setup for direct laser writing. The basic components are (i) an ultrafast laser, (ii) beam motion system, (iii) beam focusing optics, (iv) beam intensity control, and control software [15].



Laser Beam Machining, Fig. 7 Schematic illustration of two-photon polymerization process – first the liquid starting material is patterned using the laser tip as a pen. After the material has been patterned, the unsolidified material is developed, and the remnant photoresist is cleaned leaving behind the desired polymeric pattern

High photon densities are required to initiate nonlinear optical effects, and therefore ultrashort pulsed lasers typically femtosecond lasers in the visible or near-IR radiation are used. A femtosecond laser is an ultrafast laser emitting optical pulse duration in the domain of femtoseconds. Such short pulses are generated by using the lasers in mode-locked mode. The benefits of ultrashort lasers are multifold: (i) the extremely short laser pulses allow for an unprecedented degree of sculpting control of the desired structures, and (ii) the thermal transport and electron-phonon interactions which typically take about nanoseconds to picoseconds to occur are almost impossible to occur given the extremely short-time frame of the pulses and therefore the absence of thermal processes minimizes the surrounding damages. A common laser used for DLW is the Ti/Sapphire laser which operates at around 780 nm. Other examples include YAG lasers operating at green wavelength. The pulse length for the latter is sub-picosecond, and the laser repetition is in the kHz range [15].

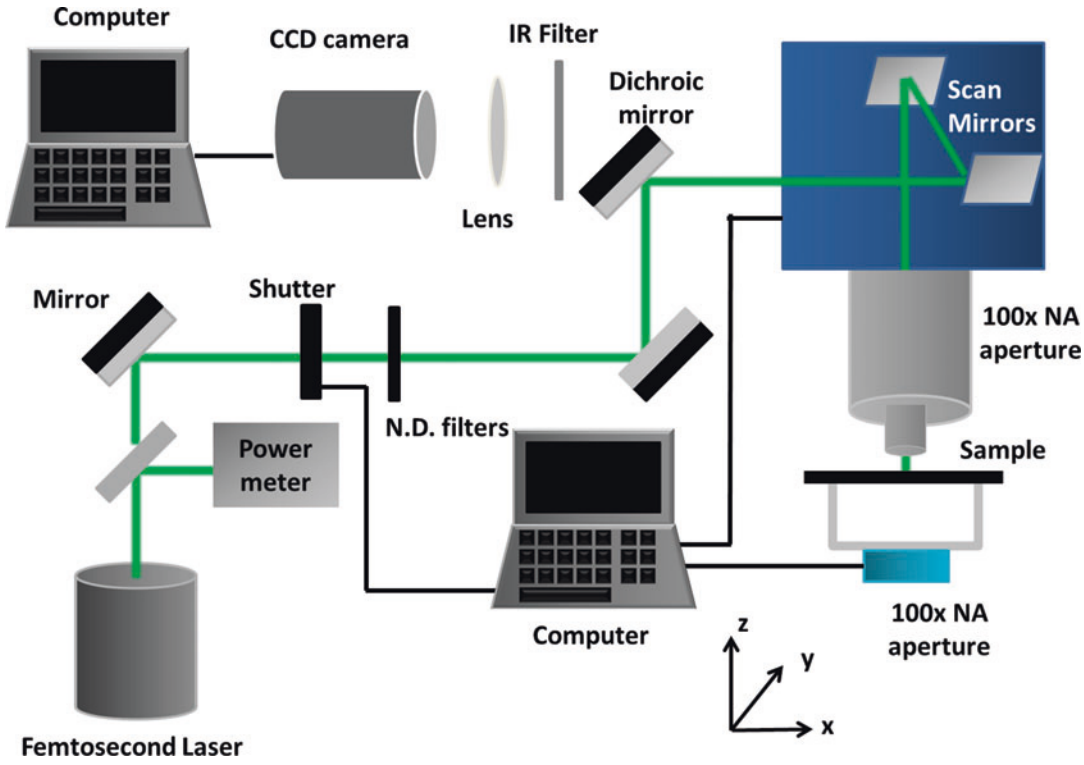
To fabricate the desired pattern, either the beam focus or the sample should be moved. The beam focus motion is achieved by using galvanometric stages, while the sample motion is controlled by using XYZ stages. Most experimental setups are equipped with both systems; this gives the flexibility to move either or both for precise patterning.

The energy distribution is controlled by the focusing optics. The size of the "voxel" is determined by the energy distribution [13–15]. A voxel in general can be understood as a pixel in a three-dimensional space. Pertaining to laser machining, a voxel refers to the volume of the material that is influenced by the laser irradiation. In terms of photopolymerization, a voxel refers to the insoluble volume that is created by the incidence of the laser source at a single spot for a certain amount of irradiation time [13].

The beam intensity is controlled by using neutral density filters, attenuators, waveplates, and polarizers. Control software, on the other hand, essentially controls the components in the experimental setup; these can include 3DPoli or laser scanning software such as Scanning Applications Modules-light (SAM-light) [15].

Applications Laser writing is a maskless fabrication technique and is therefore a one-step process. The ease of direct laser writing makes it very popular in the fields of microelectronics industry; photonics; engineering disciplines including mechanical, biomedical, and microengineering; and life sciences.

Microelectronics: Direct laser writing finds major applications in the *microelectronics industry* which includes fabrication of devices such as capacitors, transistors, resistors, batteries, and integrated chips and also interconnects between the circuit board and microbatteries [4, 8]. Additionally, direct laser writing is used to fabricate microscale inductors, which can be further applied to filters, transformers, and oscillator

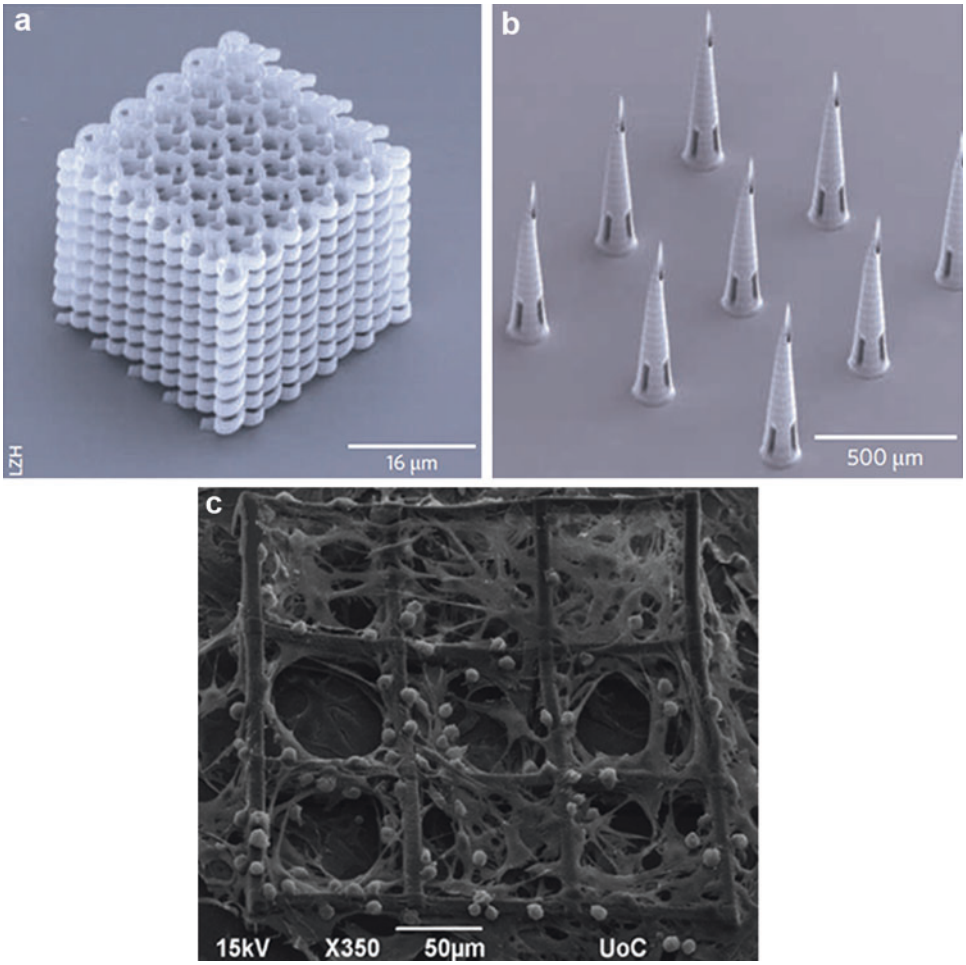


Laser Beam Machining, Fig. 8 Typical experimental setup of direct laser writing

circuits. They can be used for generating magnetic fields for mechanical actuation and sensing [16].

Biomedical and life sciences: Direct laser writing is widely used in *biomedical and life sciences*. With direct laser writing, it is possible to create chemically and physically defined cell environments. Direct laser writing provides the possibility to fabricate synthetic structures that mimic the complex human tissue structures, which in turn can be transplanted in case of trauma or injury. Laser-guided writing has made it possible to trap and deliver cells within microscale resolution onto arbitrary surfaces such as human umbilical vein cells which have been deposited in two and three dimensions using laser writing [8]. Ossicular replacement prostheses have also been fabricated using laser-guided writing utilizing two-photon polymerization. Direct laser writing is used to make 3D scaffolds for biomedical applications which can be further modified with desired cell type for the desired application. Yet another application of direct laser writing in biomedical field is the construction of microneedles, which can be used for drug delivery purposes.

Photonics: Photonics is another field, where multiphoton lithography is very popular. Direct laser writing is ideal for creating both planar and nonplanar components of waveguide structures, fiber optics, and synthesis of photonic crystals [16]. Photonic crystals are microstructures with periodical distribution of refractive indices, which are optical analogues of semiconductors with a photonic bandgap. For photonic-crystal application, the period should match that of the wavelength of interest; for instance, for visible and near-IR radiation, the lattice constant should be several hundreds of nanometers [14]. Two-photon polymerization offers great promise to process 3D structures including photonic-crystal optoelectronic devices. SU-8, one of the most popular negative tone photoresists, was used by Deubel et al. [17] to fabricate photonic crystals at telecomm wavelengths. Multiphoton fabrication has been used to fabricate crystals with complex geometries



Laser Beam Machining, Fig. 9 (**a**) A spiral photonic-crystal fabricated using Zr/Si hybrid sol-gel SZ2080 [18], (**b**) test micro-needle array for transdermal drug delivery [18], (**c**) a mineralized scaffold with pre-osteoblastic cells [19]

like diamond unit cell structures, slanted pore structures, etc. Figure 9 shows some of the structures created through direct laser writing.

Summary and Outlook

This entry explained different types of lasers used for laser beam machining processes and different types of laser beam machining processes including laser welding, ablation, cutting, laser chemical vapor deposition, laser-induced forward transfer, MAPLE direct writing, as well as multiphoton lithography. It was seen that laser machining offers several advantages such as (i) the structures obtained by ultrashort laser pulses are highly precise and aesthetic with very little damage by thermal transport; (ii) structuring with lasers does not involve mechanical engagement of lasers, and the final pattern has lower contamination levels if the same processes were carried out with a mechanical tool; (iii) with the right source of laser, it is possible to machine almost any kind of material including metals, ceramics, and polymers; (iv) given the high efficiency of laser, mass production becomes highly feasible with identical replicas; (v) typically laser cutting processes consume less energy as compared to plasma cutting technology; (vi) lasers are extremely popular in biomedical and life sciences applications, given the high precision and nanoscale

resolution of the scaffolds and devices created by laser technology that allows for better control and selectivity of delivery of biomolecules such as proteins and cells; (vii) laser writing equipment for two-photon polymerization does not require devoted cleanroom setup and can be installed in regular environments; and (viii) the advantage of direct laser writing being a maskless technique saves extra time and manual efforts as opposed to traditional mask projection-based photolithographic procedure. However, despite the plethora of advantages of laser beam machining, there are some disadvantages associated with laser beam machining. For instance, power consumption is quite high for laser machining processes and the processes should be extremely timely, otherwise the high intensity lasers may lead to extreme thermal effects and damage to the material; also there exist problems such as handling of samples with arbitrary shapes, maintenance of proper imaging and focusing on the sample surface and proper programming of the system to ensure high degree of accuracy.

Future developments for laser beam machining include more cost-effective laser manufacturing. The use of shorter wavelength lasers in addition to ultrashort pulses to machine a large range of materials with very high precision is also a part of future works. Of high interest is the application of nanostructured materials in the biological sciences and microelectronics, and therefore one of the highly desired prospects for laser micromachining is achieving smaller resolution and feature sizes in the near future.

Cross-References

- ▶ Mechanisms of Laser Ablation Synthesis of Nanoparticles in Liquid Solution
- ▶ Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing

References

1. Sun, S., Brandt, M.: Laser beam machining. In: Davim, J.P. (ed.) Nontraditional Machining Processes, pp. 35–96. Springer, London (2013)
2. Klotzbach, U., Lasagni, A.F., Panzner, M., Franke, V.: Laser micromachining. In: Lasagni, F.A., Lasagni, A.F. (eds.) Fabrication and Characterization in the Micro-nano Range. Advanced Structured Materials, vol. 10, pp. 29–46. Springer, Berlin/Heidelberg (2011)
3. Singh, S.C., Zeng, H., Guo, C., Cai, W.: Lasers: fundamentals, types and operations. In: Singh, S.C., Zeng, H., Guo, C., Cai, W. (eds.) Nanomaterials: Processing and Characterization with Lasers, pp. 1–34. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim (2012)
4. Meijer, J.: Laser beam machining (LBM), state of the art and new opportunities. J. Mater. Process. Technol. **149**, 2–17 (2004)
5. Naeem, M.: Microwelding performance comparison between a low power (125W) pulsed Nd: YAG laser and a low power (100–200 W) single mode fiber laser. Proceedings of the 3rd Pacific International Conference Optics Application of Lasers, Beijing (2008)
6. Nolte, S., Momma, C., Jacobs, H., Tunnermann, A., Chichkov, B.N., Wellegehausen, B., Welling, H.: Ablation of metals by ultrashort laser pulses. J. Opt. Soc. **14**(10), 2716–2722 (1997)
7. Bohandy, J., Kim, B.F., Adrian, F.J.: Metal-deposition from a supported metal-film using an excimer laser. J. Appl. Phys. **60**, 1538–1539 (1986)
8. Hon, K.K.B., Li, L., Hutchings, I.M.: Direct laser technology – advances and developments. CIRP Annals Manuf. Technol.. **57**, 601–620 (2008)
9. Palla-Papavlu, A., Valentina, D., Libbert, T., Dinescu, M.: Laser induced forward transfer for materials patterning. Romanian Rep. Phy. **63**, 1285–1301 (2011)
10. Gittard, S.D., Narayan, R.J.: Laser direct writing of micro and nano-scale medical devices. Expert Rev. Med. Devices **7**(3), 343–356 (2010)
11. Patz, T.M., Doraiswamy, A., Narayan, R.J., He, W., Zhong, Y., Bellamkonda, R., Modi, R., Chrisey, D.B.: Three-dimensional direct writing of B35 neuronal cells. J. Biomed. Mater. Res. B **78**(1), 124–130 (2006)
12. Allen, S.D.: Laser chemical vapor deposition. J. Appl. Phys. **52**(11), 6501–6505 (1981)
13. Engelhart, S.: Direct Laser Writing. In: Schmidt, V., Belegratis, M.R. (eds.) Laser Technology in Biomimetics, pp. 13–65. Springer, Berlin/Heidelberg (2013)
14. Sun, H.B., Kawata, S.: Two-photon photopolymerization and 3D lithographic microfabrication. In: NMR.3D Analysis. Photopolymerization, pp. 169–273. Springer, Berlin/Heidelberg (2004)
15. Selimis, A., Mironov, V., Farsari, M.: Direct laser writing: principles and materials for scaffold 3D printing. Microelectron. Eng. **132**, 83–89 (2015)
16. LaFratta, C.N., Fourkas, J.T., Baldacchini, T., Farrer, R.A.: Multiphoton fabrication. Angew..Chem. **46**, 6238–6258 (2007)
17. Deubel, M., Freymann, G., Wegener, M., Pereira, S., Busch, K., Soukoulis, C.M.: Direct laser writing of three-dimensional photonic-crystal templates for telecommunications. Nat. Mater. **3**, 444–447 (2004)

18. Farsari, M., Chichkov, B.N.: Materials processing: two-photon fabrication. Nat. Photonics **3**, 450–452 (2009)
19. Terzaki, K., Kalloudi, E., Mossou, E., Mitchell, E.P., Forsyth, V.T., Rosseeva, E., Simon, P., Vamvakaki, M., Chatzinikolaidou, M., Mitraki, A., Farsari, M.: Mineralized self-assembled peptides on 3D laser-made scaffolds: a new route toward 'scaffold on scaffold' hard tissue engineering. Biofabrication **5**, 045002 (14pp) (2013)

Laser Cutting

► Laser Beam Machining

Laser Interactions with Particles in Solution

► Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids

Laser Machining

► Laser Beam Machining

Laser Micromachining

► Laser Beam Machining

Laser Scanning Confocal Microscopy

► Confocal Laser Scanning Microscopy

Laser Scribing

► Laser Beam Machining

Laser Tweezers

► Optical Tweezers

Laser Tweezers Using Nanoapertures in Metal Films

Abhay Kotnala and Reuven Gordon
Electrical and Computer Engineering, University of Victoria, Victoria, BC, Canada

Synonyms

Aperture trap

Definition

The technique uses nanoapertures in metal films for optical trapping and manipulation of Rayleigh particles, that is, particles that are significantly smaller than the wavelength of light, typically in the 1–100 nm range. The nanoaperture trap works on the principle of self-induced back action, where the presence of the particle in the trap significantly alters the electromagnetic environment (i.e., light transmission) to favor trapping.

Conventional Laser Tweezers

The laser tweezer is an optical tool to immobilize and manipulate small objects using light. Single-beam laser tweezers employ a highly focused laser beam to trap and manipulate micron-sized particle in three dimensions. The diffraction-limited spot produces optical forces around tens of piconewtons, which enables it to hold and manipulate even biological objects like cells, bacteria, and viruses without damage. This was first shown by Ashkin where he used a single-beam laser tweezer to trap dielectric particles down to ~100 nm [1]. Modified beam profiles, such as Bessel beams or vortex beams, have been used for further manipulation of micrometer-sized particles. Nonsymmetric beams are capable of transferring angular momentum from light to particles for their rotation. The laser tweezers have applications in the field of nanoscience, materials, and nanotechnology. For a more detailed insight into

the principle of general laser tweezer, the reader can refer to the section on ▶ **Optical Tweezers**.

Stable Trapping of Rayleigh Particles

The high-intensity laser spot in an optical tweezer creates a potential well to trap the particles. The stable trapping of Rayleigh particles in an optical trap requires the potential depth of the trap to be sufficiently large to overcome the fluctuating energy of the particles due to Brownian thermal motion [2]. In this case, the particle is unlikely to escape due to Brownian motion if

$$U \gg k_B T$$

where k_B is the Boltzmann constant, T is the temperature, and U is the potential depth of the trap.

The trap potential can be formulated based on the perturbative approximation, where the particle does not significantly change the surrounding electromagnetic field and can be treated as a point dipole. The electromagnetic field constituting the laser beam induces a dielectric polarization (p) in the trapped particle, which depends on the field and the permittivity of the medium. This dipole moment p induced at the sphere by a uniform electric field E is given analytically as [3]

$$p = 4\pi {n_m}^2 \varepsilon_o r^3 \left(\frac{m^2-1}{m^2+2}\right) E$$

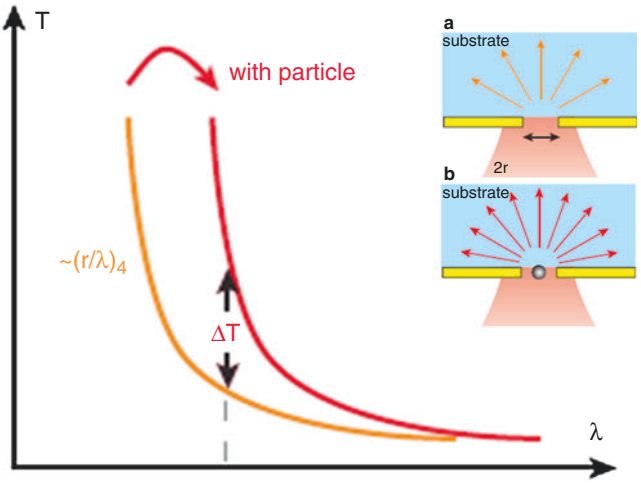
where n_m is the refractive index of the surrounding medium, m is the relative refractive index of the particle, with $m = \frac{n_p}{n_m}$, n_p is the refractive index of the particle, and r is the radius of the sphere. The overall trapping potential can be written as [2]

$$U = -p.E = -\frac{2\pi n_m r^3}{c}\left(\frac{m^2-1}{m^2+2}\right) I$$

The trapping potential is proportional to the third power of the radius of the particle (r^3) and the intensity $I = \frac{1}{2}\varepsilon_0 c|E|^2$ at the trapping point. Hence, the laser tweezer suffers from an inherent limitation when trapping particles in the Rayleigh regime of sizes less than 100 nm. Also, the Stokes drag force in a homogeneous liquid environment scales linearly with particle diameter, giving an additional dependence on the size. (This does not change the trapping potential, but does rescale the time unit, so stable trapping is achieved for less time.) Therefore, stable trapping of particles for this range of sizes using the conventional approach requires higher power, which can have damaging effects on the particle. This is a major issue especially when dealing with biological specimens. Tethering of these small particles to larger micron-sized beads is used as an alternative, which is not always favorable as it sometimes restricts the free motion of the particle while adding complexity to the process [4]. For trapping small dielectric particles, one has to look beyond the perturbative regime and look for a situation where even a small particle can induce a significant change to the electromagnetic field and play a positive role in the trapping process. This can be achieved by nanoapertures in metal films which can overcome these challenges for tweezing particles in the low Rayleigh domain. The presence of the particle in the aperture increases the local electromagnetic field intensity which increases the potential well depth required for stable trapping, without the need of increased laser power.

Bethe Theory

Nanoaperture refers to an aperture in a metal film of size much smaller than the wavelength of light. The small dimension of the aperture compared to the wavelength of light cuts off the propagation of light. This is due to the inability of the propagating wave to satisfy the boundary conditions. Thus, the light is diffracted at the edges of the subwavelength aperture. Hans Bethe first studied the transmission of light through a subwavelength hole in a metal screen, where the light transmitted through the circular aperture is approximated by the emission of magnetic dipole [5]. The transmission through such an aperture is approximated as



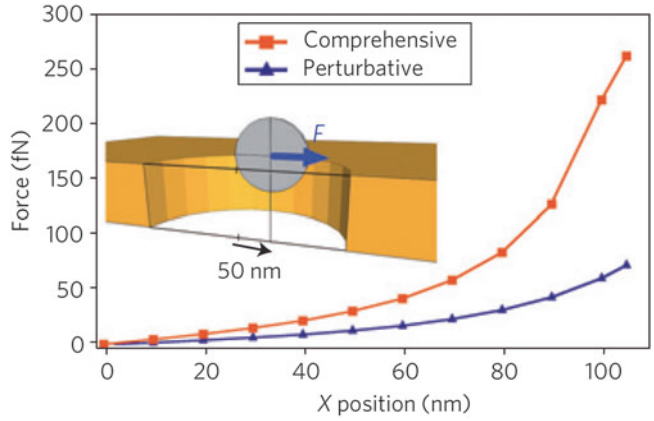
Laser Tweezers Using Nanoapertures in Metal Films, Fig. 1 Optical transmission through a single subwavelength hole: (**a**) without particle and (**b**) transmission enhanced with a dielectric particle in the hole (dielectric loading). The presence of the dielectric particle makes the hole optically larger through dielectric loading, redshifting the transmission curve, and giving the change ΔT in transmission

$$T = \frac{128\pi^3 z_0 H_0^2 r^6}{27{\lambda_o}^4}$$

where Z_o is the free-space impedance, λ_o is the wavelength in free space, r is the hole radius, and H_o is the magnetic field of the incident wave. Normalizing with respect to the area of the circular aperture, according to Bethe's theory the transmission through a subwavelength circular aperture is inversely proportional to the fourth power of the wavelength $T \propto \frac{r^4}{\lambda^4}$ as shown in Fig. 1a. If the aperture is surrounded by a dielectric medium with a refractive index n, the wavelength in the medium is scaled as $\lambda = \frac{\lambda_o}{n}$. Therefore, a larger optical transmission is expected for the same aperture size as shown in Fig. 1b. Since the optical transmission through a subwavelength aperture is very sensitive to the aperture size (scales with the fourth power of the size), any change, including a dielectric loading by a particle, effectively makes the hole "appear" larger to light (the wavelength of light shrinks in inversely with an increase in the refractive index) inside the aperture. This gives a significant change to the optical transmission. Thus, the nanoaperture shows the characteristics of working beyond the perturbative regime by having a strong scaling with the local dielectric environment and therefore is used as a tweezer for small particles in the range of few nanometers.

Nanoaperture Tweezer: Principle, Force Calculation, and Detection

Nanostructures concentrate light into highly localized and intense fields known as hot spots. This is due to the coupling of the propagating field to the evanescent field which unlike the propagating field can be concentrated well below the diffraction limit. This results in strong gradient force which is a necessary but not sufficient condition for stable trapping of the particle. For particles sizes of less than 100 nm, the nanostructures fail to fulfill the potential energy requirement to compensate for high energy events and result in unstable trapping. Since the potential energy is a function of the electric field intensity at the trapping point, it can only be increased by increasing the incident laser power. Nanoapertures not only provide the strong gradient force necessary for optical trapping but also provide a solution to work beyond the perturbative regime where the particle itself plays a major role in trapping. The trapped particle provides a dynamic feedback by increasing the magnitude of the trapping potential in order to compensate the high energy events associated with the trapped particle, which results in stable trapping. This phenomenon is termed as self-induced back action (SIBA) [6]. It is based on the high sensitivity of the nanoaperture to the local environment as shown by Bethe's aperture theory



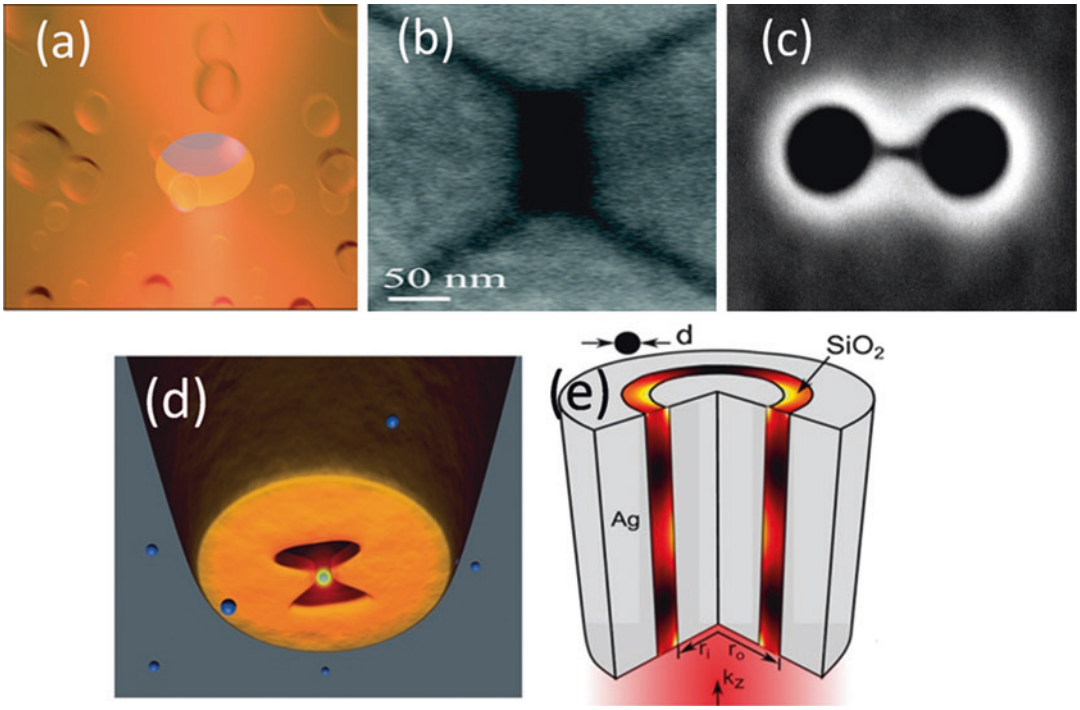
Laser Tweezers Using Nanoapertures in Metal Films, Fig. 2 The numerically computed optical force in the nanohole trap based on finite-difference time-domain (FDTD) simulations in trapping 100 nm polystyrene spheres. Two physical formulations are compared: the comprehensive MST analysis and the perturbative gradient force approximation. It is found that the perturbative gradient force approximation is no longer a good approximation in the nanohole optical trap, and the MST analysis predicts a much larger optical force than the gradient approximation does [6]

above. The presence of a small particle in the aperture significantly increases the electric field intensity which gives a decreased potential energy (increased magnitude) and thus provides a deeper potential well. The deep potential results in strong and efficient trapping of the nanoparticle using much less optical power. Thus, the nanoaperture tweezer provides an inherent feedback mechanism with simple geometry without requiring any outside feedback mechanism [6]. An early work noted increased optical forces inside an aperture in a metal film, but it was limited to particles much greater than 100 nm [7]. The SIBA is based on a more general phenomenon and has been applied with different nanoaperture designs [6, 8–11] and photonic crystals [12]. These geometries have made it possible to trap objects in the range of few nanometers using very small laser powers.

The force calculation for nanoaperture traps using perturbative optical force formulation does not provide a good approximation. This is because the dielectric particle produces a strong change to the ambient field. The rigorous Maxwell stress tensor (MST) analysis is used in this case. This has been done for a subwavelength circular aperture which shows much larger optical forces than calculated using the perturbative approach as shown in Fig. 2 [6].

Several nanoaperture designs have been used with the idea of trapping smaller and smaller particles. Some of the geometries such as circular aperture [6], bowtie aperture [11], rectangular aperture [10], coaxial aperture [9], and double nanohole aperture [8] are shown in Fig. 3. These nanoapertures have been able to show trapping of very small particles down to the size of 1 nm including the trapping of single proteins, DNA strands of a few base pairs (bp), quantum dots, and so on. Also, the planar and small size of nanoaperture make them suitable candidate for dense optical integration and use in a planar microfluidic environment. It also makes it possible to integrate it with other optical techniques like fluorescence microscopy and Raman spectroscopy.

The detection of the trapped particle in the nanoaperture tweezers is usually based on the change in the optical transmission signal through the nanoaperture. The normal optical transmission through a nanoaperture is mainly through diffraction and low, but when a particle of size smaller than the aperture and larger refractive index than the surrounding medium is trapped, the aperture becomes optically larger than the physical size due to dielectric loading. This leads to an increase in transmission indicative of the particle being trapped in the nanoaperture. An example of the

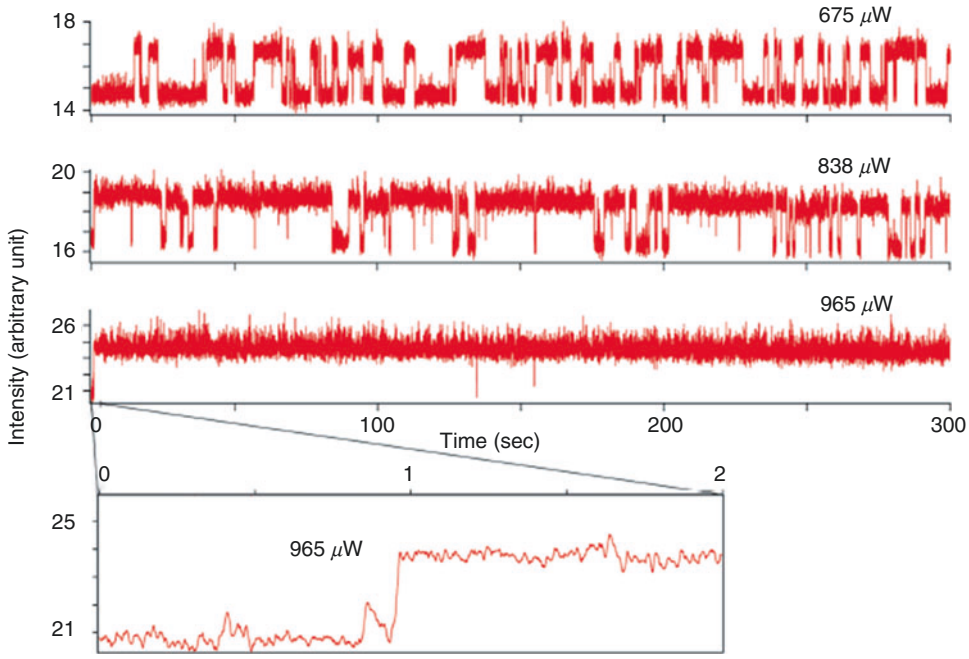


Laser Tweezers Using Nanoapertures in Metal Films, Fig. 3 Nanoaperture designs. (**a**) Circular aperture, (**b**) rectangular plasmonic nanocavity, (**c**) double nanohole aperture, (**d**) bowtie nanoaperture at the fiber tip, (**e**) coaxial nanoaperture [6, 8–11]

trapping based on the detection of change in transmission through a circular aperture is shown in Fig. 4. It is clearly shown that a 100 nm particle can be trapped stably using less than a milliwatt of laser power.

Advantages of Nanoaperture Tweezers

The nanoaperture tweezers provide some additional features compared to the regular laser tweezers, which makes it an important tool for detection, sensing, and studying of very small nanoparticles and biological molecules. The nanoaperture tweezers are able to efficiently trap very small nanoparticles and biological molecules without using tethers as compared to regular tweezer which mostly require tethers to trap and study small biological molecules. This allows the particle to have uninhibited natural motion in a given biological environment, that is, in free solution. The nanoaperture tweezer provides a high signal-to-noise ratio due to the background-free operation from the aperture, as well as a trivial detection approach. Also, the high sensitivity of the transmission signal to the physical attributes and motion of the trapped specimen makes it effective to detect trapping and interaction dynamics, without the need of fluorescent labeling. Thus, nanoaperture tweezers provide a label-free study of the dynamics of the trapped particle. It can be easily integrated into existing laser tweezer microscopy systems. The nanoaperture in metal film efficiently removes heat as the metal film acts as a heat sink which is crucial to prevent damage of biological particles due to heat. This has been shown by experiments which show very small rise in temperature for an aperture, several orders of magnitude smaller than metal nanoparticles [13]. Thus, nanoaperture tweezers provide a real-time, single-molecule, label-free, and free-solution application platform. In addition, it is simple, scalable, and low cost.



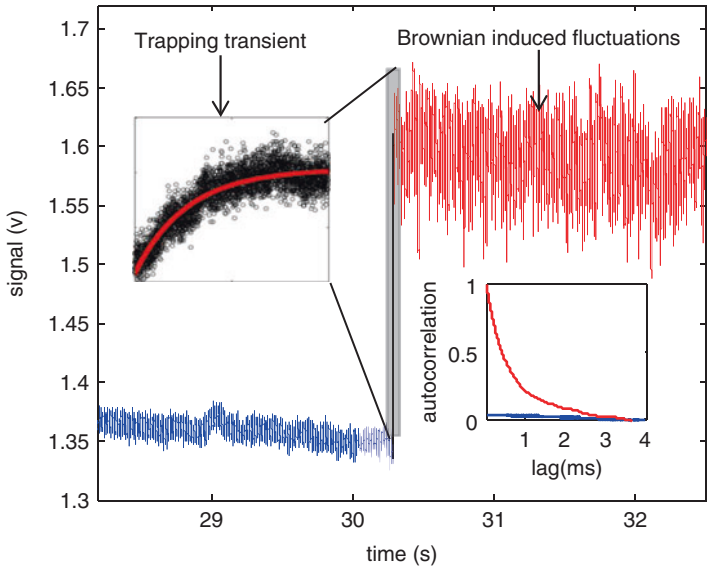
Laser Tweezers Using Nanoapertures in Metal Films, Fig. 4 Optical transmission signal through the circular nanohole optical trap showing trapping of 100 nm polystyrene spheres for different incident optical power of 675, 838, and 965 μW. The sudden jumps in the signal indicate individual trapping events

Nanoaperture Performance Quantification

The performance of an optical tweezer is evaluated by its trap stiffness. The trap stiffness is a measure of the strength of the trap and shows the efficiency of the tweezer in trapping specific nanoparticles. Techniques like the power spectrum method, the equipartition method, optical potential analysis, the drag force method, and direct measurement of the optical force are used to determine the stiffness of the optical trap. These techniques require precise measurement of the position fluctuations of the particle in the optical trap for calculation of trap stiffness. Nanoaperture trap stiffness can be calculated experimentally using basically two approaches: autocorrelation of Brownian-induced intensity fluctuations and trapping transient analysis [14]. These techniques are analogous to the power spectrum method and the step response method, respectively, used for calculating the stiffness of a conventional laser tweezer. The major difference is that the nanoaperture traps measure directly the intensity fluctuations using a photodiode as compared to conventional tweezer, which measures the position of the particle in the trap using a quadrant photodiode (QPD) or a video microscope. The latter approach is not suitable for nanoapertures as it is challenging to determine the precise position of very small particles (1–100 nm) in the nanoaperture using the QPD or video microscope.

Autocorrelation of Brownian-Induced Intensity Fluctuation and Trapping Transient Analysis

The transmission through the nanoaperture is sensitive to the position of the particle in the aperture, and so Brownian motion-induced fluctuations of the transmission signal can be used to calculate the trap stiffness. Direct observation of the intensity fluctuations reveals little information, as changes at the millisecond time scale are buried in noise. The autocorrelation of the trapped



Laser Tweezers Using Nanoapertures in Metal Films, Fig. 5 Illustration of the trapping transient and the autocorrelation of Brownian-induced intensity fluctuations (*inset*) obtained from a typical trapping event of 20 nm polystyrene nanosphere using double nanohole nanoaperture tweezer (*blue*, untrapped state; *red*, trapped state). *Insets* show expanded region of trapping transient with an exponential fit to calculate the time constant ($\boldsymbol{\tau_t}$) and autocorrelation of the trapped and untrapped transmission signal [14]

transmission signal can be used to measure characteristic time constant (τ_a) which is directly related to trapping stiffness (κ).

Trapping transient technique is analogous to the step response method related to the drag force of the fluid as compared to the autocorrelation analysis based on thermal vibrations. The step response method measures the time constant of the particle to return to the trap center due to trapping forces as opposed to the viscous drag, while the present technique calculates the time constant (τ_t) of the transition from the untrapped to trapped state for calculation of stiffness.

Figure 5 shows an example of a trapping event of 20 nm polystyrene bead using nanoaperture. The insets show the trapping transient and autocorrelation of Brownian-induced intensity fluctuations used to find the respective time constant for calculation of trap stiffness.

Trap Stiffness Calculation

The time constant obtained from the trapping transient and the autocorrelation analysis, $\boldsymbol{\tau_t}$ & $\boldsymbol{\tau_a}$, are used to evaluate the stiffness of the nanoaperture trap. The motion of a particle in an optical trap under the influence of thermal fluctuation and drag force can be modeled by an overdamped Langevin equation as

$$\frac{dx(t)}{dt} = \frac{\kappa}{\gamma}x(t) + \left(\frac{2k_BT}{\gamma}\right)^{\frac{1}{2}}\zeta(t)$$

where $x(t)$ is the particle displacement from the equilibrium position, κ is the stiffness of the optical trap, γ is the Stokes drag coefficient, and $\zeta(t)$ is the white noise. Using this equation, τ_a and τ_t are related to the trap stiffness, as [15]

$$\tau_a = \tau_t = \frac{\gamma}{\kappa}$$

For the particles trapped at the nanoaperture site, additional drag force can arise due to the wall effects from the wall of the aperture. The boundary effect is included in the calculation by using a modified Stokes drag coefficient, estimated using Faxén's law as

$$\gamma = \frac{6\pi\eta r}{1 - \frac{9}{16}\left(\frac{r}{h}\right) + \frac{1}{8}\left(\frac{r}{h}\right)^2 + \frac{45}{256}\left(\frac{r}{h}\right)^4 + \frac{1}{16}\left(\frac{r}{h}\right)^5}$$

where η is the viscosity of the medium and h is the distance from the center of the nanosphere to the wall.

It should be noted that for an accurate measurement (10 % precision) of the time constant from autocorrelation, the measurement time should be more than $\sim$1000 τ. This is important for precise

Laser Tweezers Using Nanoapertures in Metal Films, Table 1 Comparison of trap stiffness for different optical traps scaled for 10 nm dielectric sphere

Structure	Method[a]	Stiffness	Scaled stiffness ($\times \left(\frac{10\ nm}{r}\right)^3$)
Nanopillar	Exp.	0.013	0.013×10^{-3}
Slot waveguide	Sim.	0.2	0.001
Rectangular plasmonic aperture [10]	Sim.	0.018	0.018
Plasmonic nanoblock pair	Exp.	4	0.03
Double nanohole aperture [8]	Exp.	0.1	0.1
Slot waveguide	Sim.	0.131	0.131
Cylindrical aperture [6]	Sim.	6.6	0.4
Coaxial plasmonic aperture [9]	Sim.	0.19	1.56 (0.36[b])
Conventional optical tweezer	Sim.	0.03	0.3×10^{-4}

[a]Exp., experimental; Sim., simulation
[b]Scaled by Clausius–Mossotti. Unit: fN/nm.mW

calculation of the stiffness value obtained from autocorrelation due to Brownian fluctuations. Table 1 shows a comparison of the trap stiffness of different laser tweezers. The values have been scaled for 10 nm polystyrene particles. It is clear from this table that nanoapertures are 3–4 orders of magnitude more efficient than conventional tweezers. There are also other non-aperture-type nanostructures for enhanced trapping; however, those require high local field intensities.

Applications of Nanoaperture Tweezers

Nanoparticle Sensor

One of the proposed applications of nanoaperture tweezers is the sensing of nanoparticle size, concentration, and refractive index. The idea is derived from simple diffusion theory based on simple microfluidic conditions. The diffusion time (τ) of the nanoparticle in the trapping solution inside a microfluidic chip is directly proportional to the time to trap (τ_{trap}). Time to trap is the time taken by the nanoaperture tweezer to trap the nanoparticle after the laser beam is turned on, and a statistical analysis of time to trap can be used to quantify the concentration and size of the nanoparticles in the solution [16]. Furthermore, it is possible to determine the refractive index of the particle by examining the percentage change in the transmission intensity at the trapping instant which scales with the Clausius–Mossotti factor.

The diffusion coefficient D is defined as

$$D = \frac{l^2}{\tau}$$

where l is a characteristic diffusion length and τ is a characteristic diffusion time. Considering the simple case of spherical particles, the diffusion constant can also be expressed using the Stokes–Einstein equation as

$$D = \frac{k_B T}{6\pi\eta r}$$

Since the time to trap scales linearly with the diffusion time (i.e., faster diffusion makes trapping occur faster since the nanoparticle visits the trapping region more quickly), this gives

$$\tau_{trap} \propto r$$

Therefore, there is a linear dependence of time to trap with size of the nanoparticles. The above assumption is made for spherical nanoparticles, but the idea can be extended for particles of other defined shapes by making necessary change in the diffusion constant term in the above expressions wherever necessary.

The average diffusion length to the trap is related to the concentration, c, as

$$l \propto c^{\frac{-1}{3}}$$

Using the definition of diffusion constant from above, we get

$$\tau_{trap} \propto c^{\frac{-2}{3}}$$

Thus, the time to trap of a particle in a nanoaperture tweezer under simple microfluidic consideration would show a -0.67 power-law dependence with the concentration of the nanoparticles.

The refractive index of the particle trapped using nanoaperture tweezer can be determined by observing the percentage change in the transmission signal when the particle is trapped. One of the factors which determine the amount of change in transmission level as the particle goes from the untrapped to trap state is the refractive index contrast of the particle $(n_p - n_m)$. This is attributed to trapping efficiency for particles which is expected to scale, at least to first order, as the Clausius–-Mossotti factor K given by

$$K = \frac{\varepsilon_p - \varepsilon_m}{\varepsilon_p + 2\varepsilon_m}$$

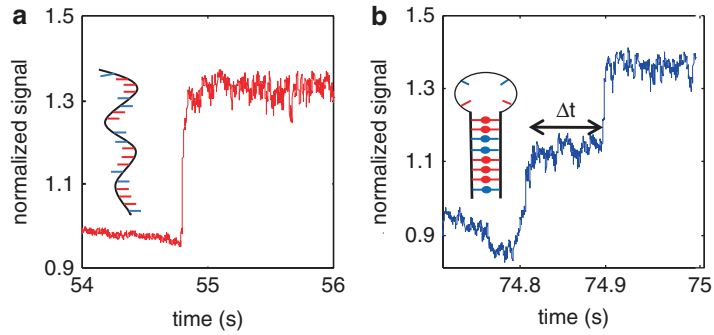
where ε_p is the dielectric permittivity of the particle and ε_m is the dielectric permittivity of the surrounding medium.

Based on these results, with suitable calibrations it is possible to determine the refractive index of the particle by observing the percentage change in the transmission signal at the onset of trapping to a previously known sample. It is important to make sure that both the particles are of similar shape and size, and the trapping is done on the same nanoaperture for accurate measurements. This is done to minimize the possible errors in measurements, as these parameters also affect the change in transmission signal at the instant of trapping.

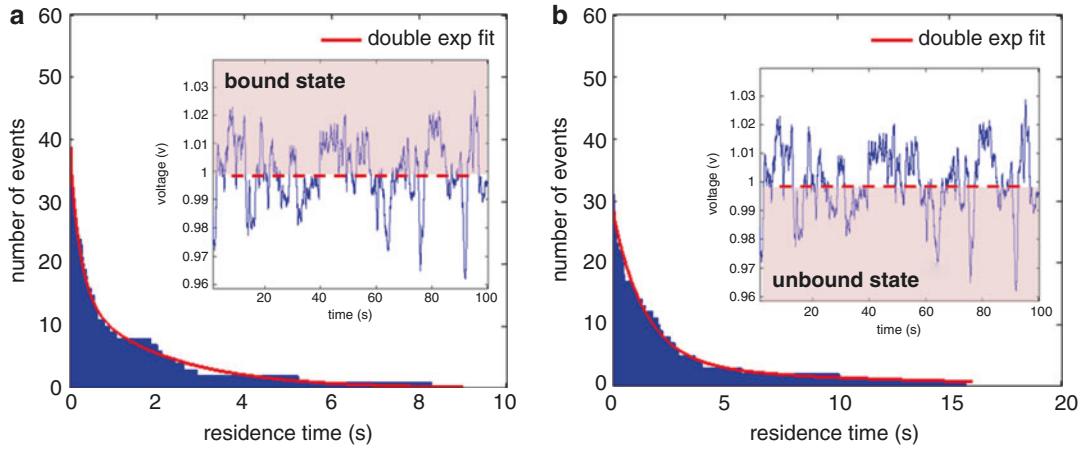
The nanoaperture tweezers as sensor are mainly limited by the acquisition time. This is because it requires a large sample space of the trapping events to accurately predict the time to trap value with maximum precision. This has been a major challenge for low-concentration sensors in general. This limitation can be overcome by imaging multiple traps in a narrower channel, thereby parallelizing the process. Some deviation from simple microfluidic theory may arise at higher concentrations or due to hydrodynamic effects. The potential applications are in size selectivity in heterogeneous samples and isolation/analysis (e.g., spectroscopy) of particles once trapped. This could be useful in the study of virus populations in biological samples, where the population levels vary over many orders of magnitude and there is great heterogeneity in the population.

Biomolecular Dynamics and Interactions

The nanoaperture tweezer is used as a label-free and free-solution single-molecule probe for biomolecular studies. The conventional optical tweezers mostly use tethering and fluorescent labeling of the biomolecules for their studies. Tethering is used to hold the biological specimen in the trap without using damaging laser powers. Fluorescent labels are used to monitor dynamic processes, such as protein binding and unzipping. These techniques suffer from some inherent limitations, such as modification of the natural state of protein–DNA, restriction of free motion, heating of the molecule, photobleaching, poor signal-to-noise ratio, large particles (e.g., DNA strands), as well as adding cost and complexity. The nanoaperture tweezer can trap very small biological particles like proteins, small DNA strands of few base pair or other ligands, without needing tethers and while using small power levels (in the mW range). Valuable information regarding the dynamics and interaction of the biological molecules can be extracted from the transmission signal obtained during the trapping of the given biological specimen. This is because of a high sensitivity of the detected signal to the change in size, shape, conformational structure, and motion of the particle in the trap. The parameters of interest are the changes in the magnitude and time scales of transmission intensity due to the molecular dynamics. A statistical analysis of these parameters provides information that is useful in understanding the dynamics at single-molecule



Laser Tweezers Using Nanoapertures in Metal Films, Fig. 6 (**a**) Single-strand DNA trapping event with no intermediate step. (**b**) A hairpin DNA trap event showing the unzipping with an intermediate step of ∼ 0.1 s [18]

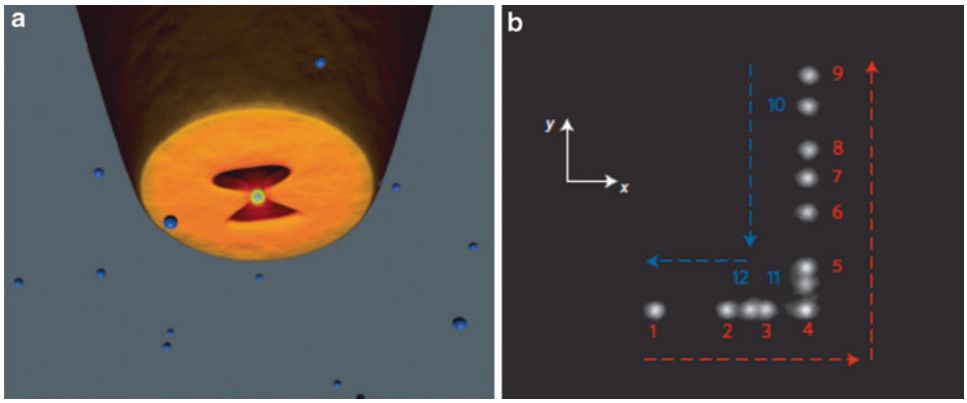


Laser Tweezers Using Nanoapertures in Metal Films, Fig. 7 (**a**, **b**) Histograms of residence times of human serum albumin (*HSA*) molecule in the bound and unbound states, respectively, as obtained from the transmission signal (*inset*). The *inset* shows the interaction of HSA with tolbutamide during trapping in the double nanohole aperture [20]

level and in real time. In this respect, the nanoaperture tweezers have shown the unfolding of proteins such as bovine serum albumin (BSA) and also unzipping of hairpin DNA [17, 18]. Figure 6 shows the unzipping of the hairpin DNA using a double nanohole aperture tweezer. There are two-step changes in the transmission signal, where the first step corresponds to the trapping and the second to the increase in the polarizability of the molecule due to elongation of DNA.

Another approach of simply comparing the time constants of the autocorrelation functions of the transmission signal can be used to distinguish between the bounded and unbound form of proteins [19]. Small-molecule binding to a protein alters its molecular dynamics. Aperture trap can study the real-time binding kinetics of protein – small molecule interactions and determine the disassociation constant. This can be done by using a statistical approach of measuring the resident time in the bound and unbound states as shown in Fig. 7 [20]. The nanoaperture tweezer system can detect the bound and unbound forms from the transmission signal as the protein in the bound state has a higher polarizability and hence higher transmission through the aperture, compared to the protein in the unbound state.

This makes the nanoaperture tweezer an exciting tool for studying dynamics of biomolecules in a label-free way and also shows interactions among them like protein–DNA interactions and protein–small molecule interactions.



Laser Tweezers Using Nanoapertures in Metal Films, Fig. 8 Three-dimensional manipulation of a single 50 nm polystyrene sphere using bowtie aperture at the fiber tip. (**a**) Schematic of experimental configuration with the laser directly coupled into the fiber. (**b**) Experimental time trace of the reflected signal showing trapping of a single 50 nm polystyrene sphere. [11]

The information obtained from these studies could aid researchers in the study of diseases and drug discovery applications.

Three-Dimensional Manipulation

The nanoaperture trap can be fabricated at the end of a scanning near-field optical microscopy (SNOM) tip or an optical fiber to trap and manipulate nanoparticles. For this, a metallic film is deposited at one end of the fiber, and the nanoaperture is fabricated on the film to form the optical trap. The trap laser is transmitted through the other end of the fiber, and the assembly is used to trap the particle of interest. Due to the unique configuration of the setup, the particle can be moved in three dimensions by moving the fiber with the aperture, after the particle is trapped. This is because the alignment of the laser is not disturbed during the movement of the fiber, which prevents the release of the particle from the trap. A bowtie aperture at the end of a tapered fiber has been used to trap and translate polystyrene spheres of size 50 nm in three dimensions as shown in Fig. 8. The ability to manipulate nanoparticles with high precision will open applications in the field of nanoscience, noninvasive manipulation of biological molecules, and material technology [11].

Key Research Findings

Nanoaperture tweezer can trap and manipulate Rayleigh particles of sizes 1–100 nm. The SIBA mechanism results in efficient trapping of the nanoparticles with small incident laser powers in the mW range. The detected trapping signal due to Brownian fluctuation of the particle in the trap and during the transition from untrapped to trapped state are used to quantify the efficiency of the aperture tweezers. The aperture traps provide high signal-to-noise ratio (SNR), greater sensitivity of the detected signal to trapped particle characteristics as compared to conventional tweezers, high temporal resolution, and natural heat sink, which make it suitable for different biological applications. Nanoaperture trap are used for sensing and studying molecular dynamics of proteins, DNA, and protein–small molecule interactions. The integration of nanoaperture tweezer with the existing fiber technologies, microfluidics, and other optical techniques can make nanoaperture tweezers a nano-optical tool for multiple applications in a single platform.

Cross-References

▶ Optical Tweezers

References

1. Ashkin, A., Dziedzic, J.M., Bjorkholm, J.E., Chu, S.: Observation of a single-beam gradient force optical trap for dielectric particles. Opt. Lett. **11**, 288–290 (1986)
2. Nieminen, T. A., Knöner, G., Heckenberg, N. R. & Rubinsztein Dunlop, H.: Physics of optical tweezers. In: Laser Manipulation of Cells and Tissues, Academic Press, United States. vol. 82, pp. 207–236. (2007)
3. Molloy, J.E., Padgett, M.J.: Lights, action: optical tweezers. Contemp. Phys. **43**, 241–258 (2002)
4. Bryant, Z., et al.: Structural transitions and elasticity from torque measurements on DNA. Nature **424**, 338–341 (2003)
5. Jackson, J.D.: Classical Electrodynamics. Wiley, New York (1998)
6. Juan, M.L., Gordon, R., Pang, Y., Eftekhari, F., Quidant, R.: Self-induced back-action optical trapping of dielectric nanoparticles. Nat. Phys. **5**, 915–919 (2009)
7. Okamoto, K., Kawata, S.: Radiation force exerted on subwavelength particles near a nanoaperture. Phys. Rev. Lett. **83**, 4534–4537 (1999)
8. Pang, Y., Gordon, R.: Optical trapping of 12 nm dielectric spheres using double-nanoholes in a gold film. Nano Lett. **11**, 3763–3767 (2011)
9. Saleh, A.A.E., Dionne, J.A.: Toward efficient optical trapping of sub-10-nm particles with coaxial plasmonic apertures. Nano Lett. **12**, 5581–5586 (2012)
10. Chen, C., et al.: Enhanced optical trapping and arrangement of nano-objects in a plasmonic nanocavity. Nano Lett. **12**, 125–132 (2012)
11. Berthelot, J., et al.: Three-dimensional manipulation with scanning near-field optical nanotweezers. Nat. Nanotechnol. **9**, 295–299 (2014)
12. Mirsadeghi, S.H., Young, J.F.: Ultrasensitive diagnostic analysis of Au nanoparticles optically trapped in silicon photonic circuits at sub-milliwatt powers. Nano Lett. **14**, 5004–5009 (2014)
13. Melentiev, P.N., Afanasiev, A.E., Kuzin, A.A., Baturin, A.S., Balykin, V.I.: Giant optical nonlinearity of a single plasmonic nanostructure. Opt. Express **21**, 13896–13905 (2013)
14. Kotnala, A., Gordon, R.: Quantification of high-efficiency trapping of nanoparticles in a double nanohole optical tweezer. Nano Lett. **14**, 853–856 (2014)
15. Reif, F.: Fundamentals of Statistical and Thermal Physics. McGraw-Hill, New York (1965)
16. Kotnala, A., DePaoli, D., Gordon, R.: Sensing nanoparticles using a double nanohole optical trap. Lab Chip **13**, 4142–4146 (2013)
17. Pang, Y., Gordon, R.: Optical trapping of a single protein. Nano Lett. **12**, 402–406 (2012)
18. Kotnala, A., Gordon, R.: Double nanohole optical tweezers visualize protein p53 suppressing unzipping of single DNA-hairpins. Biomed. Opt. Express **5**, 1886–1894 (2014)
19. Al Balushi, A.A., Zehtabi-Oskuie, A., Gordon, R.: Observing single protein binding by optical transmission through a double nanohole aperture in a metal film. Biomed. Opt. Express **4**, 1504–1511 (2013)
20. Al Balushi, A.A., Gordon, R.: A label-free untethered approach to single-molecule protein binding kinetics. Nano Lett. **14**, 5787–5791 (2014)

Laser Writing

▶ Laser Beam Machining

Laser-Assisted Tip-Based Heating and Fabrication

▶ Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing

Laser-Assisted Tip–Substrate Interaction

▶ Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing

Laser-Reduced Graphene Oxide

▶ Nanostructures and Characteristics of Graphene and Graphene Oxides

Laser-Scribed Graphene

▶ Nanostructures and Characteristics of Graphene and Graphene Oxides

Lateral Force Microscopy (LFM)

▶ Friction Force Microscopy

Lateral Shear Mode Resonators

▶ Lamé Mode MEMS Resonators

Lateral Thinking

▶ Biornametics: Architecture Defined by Natural Patterns

Laterally Vibrating Piezoelectric Resonators

Gianluca Piazza
Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA

Synonyms

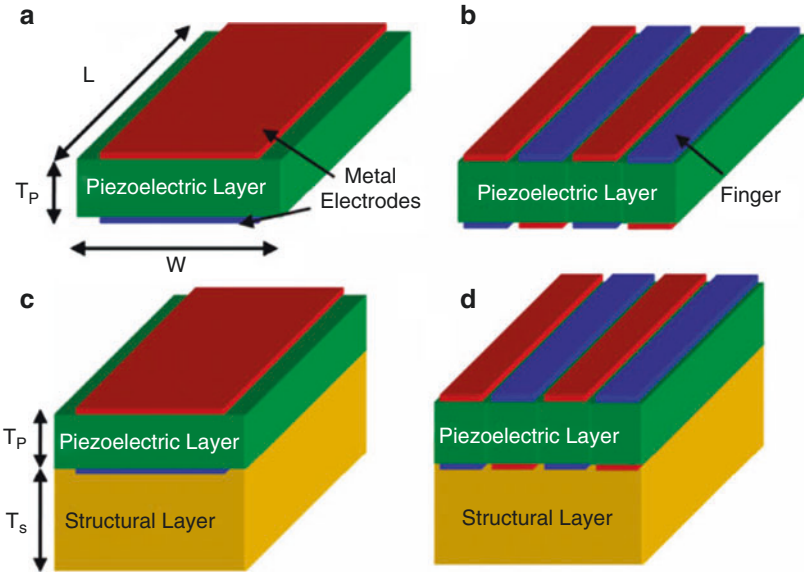
Bulk acoustic wave MEMS resonators; Contour-mode resonators; Lamb-wave resonators

Definition

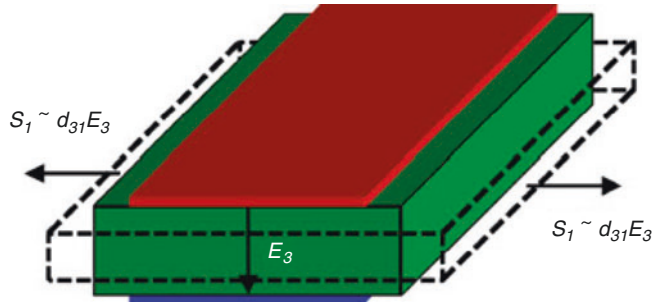
A laterally vibrating piezoelectric resonator is a particular class of microelectromechanical (MEM) devices, whose resonant motion is transduced via thin-film piezoelectrics and occurs primarily in the in-plane lateral directions of the structure (e.g., width extensional or length extensional motion). The resonant frequency is therefore set by the lateral dimensions of the device, which can operate in the few MHz to several GHz range on the same chip and is generally employed for the making of filters and oscillators.

Principle of Operation

Laterally vibrating piezoelectric resonators are formed either entirely by a thin-film piezoelectric film sandwiched between metallic electrodes [1] or by a composite constituted by the same piezoelectric film stacked on top of another structural layer such as a semiconducting or insulating material (Fig. 1). As shown in Fig. 2, the application of a potential across the piezoelectric film induces a lateral strain in the plane of the device. When the frequency of the excitation signal coincides with the mechanical resonance of the piezoelectric/



Laterally Vibrating Piezoelectric Resonators, Fig. 1 Mock-up view of laterally vibrating piezoelectric resonators formed either by (**a**, **b**) pure piezoelectrics ((**a**) representing a single finger and (**b**) the so-called multi-finger embodiment) or (**c**, **d**) composite stack ((**c**) representing a single finger and (**d**) the so-called multi-finger embodiment) of a piezoelectric and semiconducting or insulating material (structural layer)



Laterally Vibrating Piezoelectric Resonators, Fig. 2 Principle of transduction of a laterally vibrating piezoelectric resonator. The application of an electric field, E_3, across the film thickness induces lateral expansion and contraction of the piezoelectric layer. The induced strain (s_1) is approximately proportional to the electric field via an equivalent piezoelectric coefficient (d_{31})

composite body, the device exhibits enhanced amplitude of vibration, which is translated to an electrical signal by the direct piezoelectric effect and generates a charge proportional to the amplitude of the displacement.

The main geometry of the laterally vibrating structure is designed such that one of the in-plane lateral dimensions, W, is used to set the resonant frequency of the structure:

$$f_o = v_{ac}/2W \tag{1}$$

where v_{ac} is the acoustic velocity of the piezoelectric/composite that forms the resonator, and depends on the square root of the ratio of the Young's modulus and mass density of the material. This is the main important feature of laterally vibrating resonators. In fact, the ability to set frequency via the in-plane lateral geometrical variables of the structure makes it possible to lithographically define multiple frequency devices on the same substrate and effectively produce components that are capable of operating over a broad spectrum and dynamically reconfigure radio receivers. Furthermore, as with transistors, lithography scaling is accompanied by an increase in the frequency of operation of the resonator.

The second in-plane lateral dimension, L, and the piezoelectric film thickness, T_p, are used to properly define the capacitance, C_o, of the piezoelectric transducer and therefore its electrical impedance.

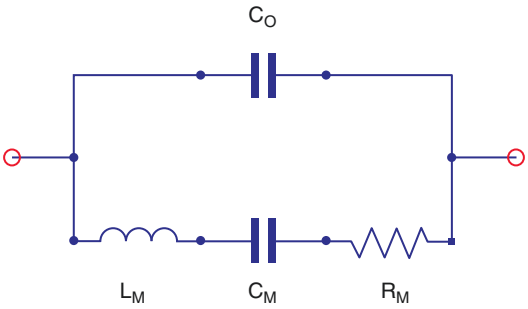
$$C_o = \varepsilon \frac{WL}{T} \tag{2}$$

In the case of a composite structure, an additional design parameter is set by the thickness of the main structural layer, T_s. This layer has generally been introduced to enhance the quality factor, Q, (an inverse measurement of the mechanical damping in a resonator) of the pure piezoelectric device. In various demonstrations, materials with intrinsically high Q have been used, such as silicon [2–6], diamond [7], and silicon carbide [8]. Although beneficial, its insertion introduces a trade-off between Q and the electromechanical coupling coefficient, k_t^2, a parameter that describes how effectively the input electrical energy in the resonator is converted into mechanical (i.e., motion). In the case of pure piezoelectrics, the k_t^2 is set exclusively by the mode of vibration and the material properties. In the case of a composite, the intrinsic electromechanical coupling is approximately scaled by the ratio of the thickness of the piezoelectric layer to the total thickness of the composite ($T_s + T_p$). Therefore, the selection of the thickness of the main structural layer, T_s depends on the resonator application and should be tailored to either enhance the device Q or maintain a certain k_t^2. The most important figure of merit (FOM) of an electromechanical resonator is in fact defined as the product of k_t^2 and Q (FOM $= k_t^2 Q$). The FOM is the main factor in determining the performance of the system

in which it will be used, and affects parameters such as insertion loss and bandwidth in a filter, and power consumption and phase noise in an oscillator. Different from the pure piezoelectric device, the composite structure permits, within fabrication limits, the tuning of k_t^2 values and Q by varying the thickness of T_s.

Various thin-film piezoelectrics have been used for making these resonators, with zinc oxide (ZnO), lead zirconate titanate (PZT), and aluminum nitride (AlN), being the most commonly available. The selection of one over the others is dictated by material properties, process compatibility, reproducibility, and specific application of interest. For example, PZT offers a very large k_t^2, but exhibits very low Q and therefore is exclusively used in a composite configuration with other materials such as silicon. AlN is easily sputter-deposited and has very reproducible properties with high Q and intermediate values of k_t^2, therefore often making it a preferred material for the fabrication of laterally vibrating resonators formed by pure piezoelectrics.

The piezoelectric layer thickness can range between 100 and 3 μm. The structural layer forming the composite can be as thick as 50 μm for low frequency (10–200 MHz) devices made out of silicon, but it is generally thinner for higher frequency (0.5–3 GHz) devices and in the few μm range if made out of materials such as SiC or diamond. The in-plane dimensions of the device are generally in the 50–200 μm range for each side if arranged in a rectangular geometry. This area is kept rather constant for devices operating at different frequencies so that the device capacitance does not change significantly. At higher frequency (>200 MHz), this is attained by effectively arraying many of the basic aforementioned elements (which are called fingers in this configuration) and exciting adjacent elements 180° out of phase with respect to each other so as to induce an overtone vibration in the overall plate. In this case, the standing wave generated into the body of the resonator closely resembles a symmetric Lamb wave. From a practical standpoint, this is achieved by simply patterning small feature electrodes on top of the piezoelectric plate and applying opposite polarity voltage to the adjacent fingers (Fig. 1b or d).



Laterally Vibrating Piezoelectric Resonators, Fig. 3 Equivalent electrical circuit for an electromechanical resonator

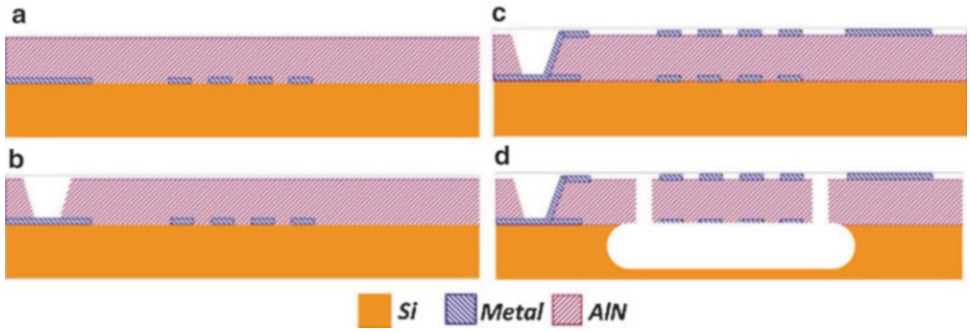
The previous sections intuitively describe the principle of operation of these resonators. More rigorously, an equivalent electrical model can be used to describe the device operation as a passive electronic component. The simplest equivalent circuit (Fig. 3) is formed by a capacitor representing the physical transducer in parallel with a series LCR, which models the mechanical portion of the resonator.

The circuit components are completely defined by the geometry of the resonator and its electromechanical coupling and are given by the following expressions:

$$R_M = \frac{1}{\omega_o C_o \frac{\pi^2}{8} k_t^2 Q} \quad C_M = \frac{8}{\pi^2} k_t^2 C_o \quad L_M = \frac{1}{\omega_o^2 C_M} \tag{3}$$

Methods of Fabrication

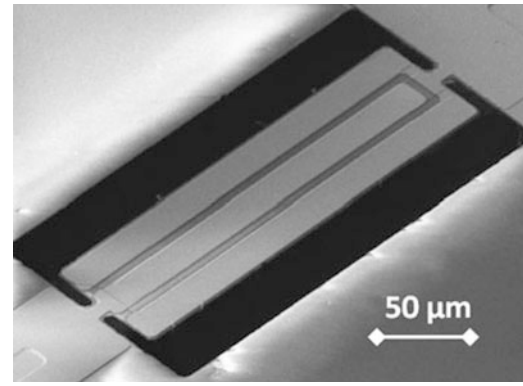
Laterally vibrating piezoelectric resonators are generally fabricated by micromachining techniques. Although each process has its own specific implementation, there are common steps. In particular, standard integrated circuit (IC) processes are used for the deposition and patterning of the metals that sandwich the piezoelectric film. The deposition of the piezoelectric film is one of the most challenging steps and requires the most attention since the most characteristics of the



Laterally Vibrating Piezoelectric Resonators, Fig. 4 Schematic representation of the fabrication process used for the making of pure AlN resonators. (**a**) Bottom electrode is deposited and patterned; AlN is sputter-deposited and (**b**) access to the bottom electrode is defined; (**c**) top electrode is deposited and patterned; (**d**) the main resonator body is defined by a dry etching step and the structure is released from the substrate

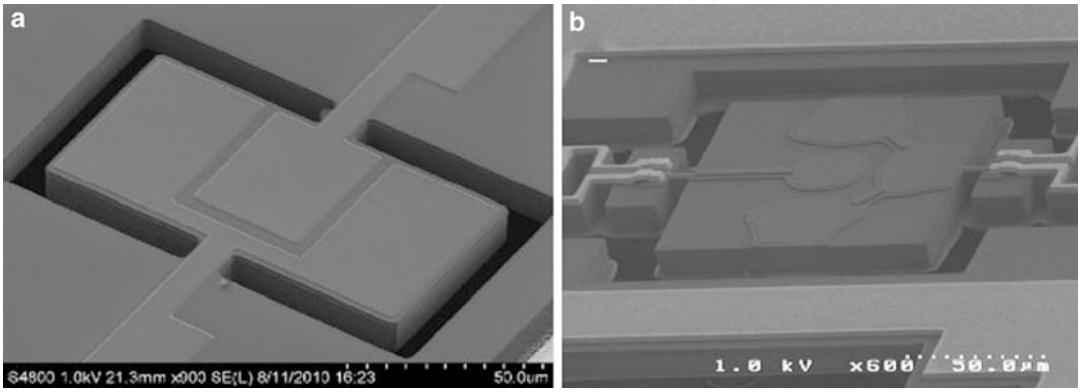
resonators are affected by the property of this film. ZnO and AlN are preferentially sputter-deposited, whereas a sol–gel technique is used for the deposition of PZT films. The as-deposited film quality is generally assessed by monitoring its orientation via x-ray diffraction measurements (note that these films are poly-crystalline). A rocking curve measurement is generally performed with the same equipment to quantitatively assess the degree of orientation. A film is considered to be of good quality when the full width half maximum (FWHM) of the rocking curve measurement is less than 2°. Although higher values of the FWHM will likely not affect the piezoelectric properties of the film, they will further increase the Q of the device (intrinsic material damping).

Specialized recipes have been developed for the etching of the piezoelectric layers, but ultimately use gases and equipment that are readily available in any IC fabrication facility. It is worth noting that compatibility with standard IC processes is of paramount importance for enabling large-scale manufacturing of these devices, which will be produced in the billions if commercialized for filtering or frequency setting applications. In addition, a longer term vision is to integrate these mechanical devices with electronics (mainly complementary metal oxide semiconductors [CMOS]) so that they are readily available to circuit designers and can be broadly employed for design of low-loss components. In this case, AlN emerges as the most suitable candidate since it is very stable and is not considered to be a contaminant; it is already employed in some CMOS facilities. A simple schematic of the fabrication process for an AlN laterally vibrating resonator is shown in Fig. 4. Figure 5 is a scanning electron micrograph of the same type of AlN devices. This is an individual resonator formed by an AlN films sandwiched by three patterned metal electrodes, so as to excite an acoustic wave in the plane of the film and operate at a frequency around 200 MHz. In this case, the device release is straightforward and is generally performed by a dry fluorine etch chemistry that removes the underlying silicon.



Laterally Vibrating Piezoelectric Resonators, Fig. 5 Scanning electron microscope image of an AlN laterally vibrating micromechanical resonator

In the case of a composite structure, an additional etch step is required to define the structural layer. Bulk micromachining of silicon via a



Laterally Vibrating Piezoelectric Resonators, Fig. 6 Scanning electron microscope images of two types of composite resonators formed by (**a**) AlN on Si (Courtesy of Prof. R. Abdolvand, Oklahoma State University) and (**b**) PZT on Si (Courtesy of Dr. R. Polcawich, Army Research Laboratory)

Bosch-type process, which etches through the entire substrate, has generally been used when Si is present. Examples of microfabricated structures are shown in Fig. 6. SiC and diamond are generally thinner and standard reactive ion etching has been employed for their patterning. In the case of a composite structure incorporating silicon, generally a silicon dioxide sacrificial layer is employed and a wet etch step is used to remove it to accomplish device release.

Main Research Accomplishment

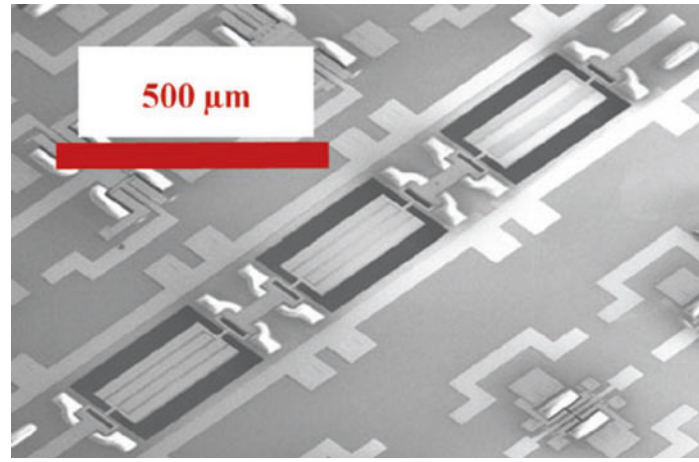
The main accomplishments in terms of device performances such as k_t^2, Q and frequency of operation are reported herein for the pure piezoelectrics and the composite laterally vibrating resonators.

In terms of pure piezoelectrics, the reproducibility of AlN deposition and its compatibility with CMOS fabrication make it the piezoelectric of choice for laterally vibrating resonators. These devices have exhibited k_t^2 around 2 % and quality factors as high as 4,000 in ambient conditions [1, 9]. These k_t^2 and Q values are sufficient to satisfy stringent power and phase noise requirements for oscillators and have proven adequate to meet the specifications for narrowband filtering. More recently the frequency of operation of these resonators has been extended around 10 GHz [10, 11] and there is no fundamental limit to operating them at higher frequencies. This implies that this technology is effectively broadband and could be employed to synthesize filters and oscillators for a wide range of communication systems that operate over different frequency bands.

In terms of composite resonators, some of the best results have been attained by the integration of AlN on Si[2]. These devices exhibit k_t^2 that are generally <1 % and have attained Qs in excess of 15,000 at low frequency (<200 MHz) and above 5,000 at about 1 GHz. It is clear that a trade-off between k_t^2 and Q is intrinsic to the composite structure over the pure piezoelectric solution and its choice depends on the particular application.

Examples of Application

The main advantage of laterally vibrating resonators over other type of resonant devices consists in the ability to define multiple frequencies on the same substrate (for a fixed film thickness) and therefore address some of the emerging needs of reconfigurable and multi-band radio solutions, which need to operate efficiently over a broad range of the electromagnetic spectrum. Currently available solutions are either too bulky or lossy to meet the needs of a broadband and widely reconfigurable radio and will ultimately be replaced by small footprint and low power devices



Laterally Vibrating Piezoelectric Resonators, Fig. 7 Scanning electron microscope image of a filter formed by electrical coupling of AlN laterally vibrating resonators

such as the laterally vibrating MEMS resonators described here.

Resonators are primarily employed as individual components that form the frequency reference in electronic oscillators or arrayed and coupled for filtering applications.

To this extent, reconfigurable and multi-frequency oscillators based on AlN laterally vibrating resonators have been demonstrated [12–14], experimentally proving that these device have high performance and could effectively enable new applications. Currently, no such solution exists commercially because the large size of the available resonators and the use of exotic substrates that are non-integrable with electronics would render them prohibitively bulky.

Simultaneously, similar pure AlN resonators have been electrically coupled [15] (Fig. 7) to form low-loss band pass filters. These devices are very good candidates to replace surface acoustic wave (SAW) filters given their smaller size and integrability with CMOS. These micromachined filters occupy 20–30 X times smaller area and offer a reduction in insertion of losses of about 2–3 dB with respect to SAW devices.

References

1. Piazza, G., Stephanou, P.J., Pisano, A.P.: Piezoelectric aluminum nitride vibrating contour-mode MEMS resonators. J. Microelectromech. Syst. **15**(6), 1406–1418 (2006)
2. Harrington, B.P., Shahmohammadi, M., Abdolvand, R.: Toward ultimate performance in GHz MEMS resonators: low impedance and high Q. In: 23rd IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2010), pp. 707–710. Hong Kong (2010)
3. Ho, G.K., Abdolvand, R., Sivapurapu, A., Humad, S., Ayazi, F.: Piezoelectric-on-silicon lateral bulk acoustic wave micromechanical resonators. J. Microelectromech. Syst. **17**(2), 512–520 (2008)
4. Humad, S., Abdolvand, R., Ho, G.K., Piazza, G., Ayazi, F.: High frequency micromechanical piezo-on-silicon block resonators. In: IEEE International Electron Devices Meeting 2003, pp. 39–3. Washington, DC (2003)
5. Chandrahalim, H., Bhave, S.A., Polcawich, R.G., Pulskamp, J.S., Kaul, R.: PZT transduction of high-overtone contour-mode resonators. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **57**(9), 2035–2041 (2010)
6. Chandrahalim, H., Bhave, S.A., Polcawich, R., Pulskamp, J., Judy, D., Kaul, R., Dubey, M.: Performance comparison of $Pb(Zr_{0.52}Ti_{0.48})O_3$-only and $Pb(Zr_{0.52}Ti_{0.48})O_3$-on-silicon resonators. Appl. Phys. Lett. **93**(23), 233504 (2008)
7. Abdolvand, R., Ho, G.K., Butler, J., Ayazi, F.: ZNO-on-nanocrystalline diamond lateral bulk acoustic resonators. In: 2007 20th IEEE International Conference on Micro Electro Mechanical Systems – MEMS'07, pp. 795–798. Kobe (2003)
8. Gong, S., Kuo, N., Piazza, G.: A 1.75 GHz SiC lateral overtone bulk acoustic-wave resonator. In: 16th International Conference on Solid-State Sensors, Actuators and Microsystems, Transducers' 2012, pp. 922–925. Beijing (2012)
9. Olsson, R.H., III, Fleming, J.G., Wojciechowski, K.E., Baker, M.S., Tuck, M.R.: Post-CMOS compatible aluminum nitride MEMS filters and resonant sensors. In: 2007 I.E. International Frequency Control Symposium Jointly with the 21st European Frequency and Time Forum, pp. 412–419. Geneva (2007)

10. Rinaldi, M., Zuniga, C., Piazza, G.: 5–10 GHz AlN contour-mode nanoelectromechanical resonators. In: 2009 I.E. 22nd International Conference on Micro Electro Mechanical Systems, MEMS 2009, pp. 916–919. Sorrento (2009)
11. Wojciechowski, K.E., Olsson, R.H., III, Nordquist, C. D., Tuck, M.R.: Super high frequency width extensional Aluminum Nitride (AlN) MEMS resonators. In: 2009 I.E. International Ultrasonics Symposium, pp. 1179–1182. Rome (2009)
12. Rinaldi, M., Chengjie, Z., Van der Spiegel, J., Piazza, G.: Reconfigurable CMOS oscillator based on multifrequency AlN contour-mode MEMS resonators. IEEE Trans. Electron Devices **58**(5), 1281–1286 (2011)
13. Zuo, C., Van Der Spiegel, J., Piazza, G.: 1.05-GHz CMOS oscillator based on lateral- field-excited piezoelectric AlN contour- mode MEMS resonators. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **57**(1), 82–87 (2010)
14. Zuo, C., Sinha, N., Van der Spiegel, J., Piazza, G.: Multifrequency pierce oscillators based on piezoelectric AlN contour-mode MEMS technology. J. Microelectromech. Syst. **19**(3), 570–580 (2010)
15. Piazza, G., Stephanou, P.J., Pisano, A.P.: Single-chip multiple-frequency ALN MEMS filters based on contour-mode piezoelectric resonators. J. Microelectromech. Syst. **16**(2), 319–328 (2007)

Lattice Boltzmann Methods for Nanofluidics

M. Sbragaglia[1], L. Biferale[1] and S. Succi[2]
[1]Department of Physics and INFN, University of "Tor Vergata", Rome, Italy
[2]Istituto per le Applicazioni del Calcolo CNR, Rome, Italy

Synonyms

Droplet microfluidics; Microfluidic devices

Definition

Lattice Boltzmann method (LBM) applications – with emphasis on hydrodynamics at the small scales (nanofluidics) – are discussed.

Introduction

The quantitative study of the flowing state of matter relies upon two major levels of description: continuum and atomistic. In the continuum picture fluids are described in terms of fields, such as density, velocity, pressure, which vary smoothly in space and time. Such a description is fairly adequate for most applications, to the point that the physics of fluids is often implicitly identified with continuum fluid mechanics. Nevertheless, for over a century, it is known that fluids (gas and liquids) are ultimately composed by a collection of individual entities, atoms and molecules, whose discrete nature becomes apparent at scales around the nanometer and below. Much less known, at least to the general public, is the existence of a third, intermediate level of description, in which fluids are represented in terms of the probability density to find a molecule with a given velocity field as a function of space and time. Lattice Boltzmann methods (LBMs) are a minimal form of Boltzmann kinetic equation which is meant to simulate the dynamic behavior of fluid flows without directly solving the equations of continuum fluid mechanics. Instead, macroscopic fluid dynamics emerges from the underlying dynamics of a fictitious ensemble of particles, whose motion and interactions are confined to a regular spacetime lattice [1–3]. Historically, the main successful applications in the context of computational fluid dynamics pertain to the weakly compressible Navier–Stokes equations [1, 2]. The LBM does not solve directly the hydrodynamic conservation equations but rather models the streaming and collision (i.e., relaxation toward local equilibria) of particles, thus offering a series of advantages [1–3], among them the capability of handling boundary conditions associated with highly irregular geometries, its nearly ideal amenability to parallel computing, and the possibility to describe nonideal fluids with phase transitions/phase separation [3, 4]. This contribution is meant to be a short summary of LBM applications with emphasis on those issues which are relevant for the modeling of the hydrodynamics at the small scales, which includes a

huge variety of applications in micro- and nanofluidics. In such situations, complexity derives from complex fluid properties (e.g., nonideal fluids, thermal fluctuations) and boundary conditions (e.g., confinement, slip).

LBM Model

The goal of LBM is to have a kinetic approach to reproduce the continuity and the Navier–Stokes equations for both density and momentum transport (repeated indexes are meant summed upon) [5]

$$\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial r_\beta}\left(\rho u_\beta\right) = 0 \tag{1}$$

$$\rho\left[\frac{\partial u_\alpha}{\partial t} + u_\beta \frac{\partial u_\alpha}{\partial r_\beta}\right] = -\frac{\partial p}{\partial r_\alpha} + \frac{\partial \sigma_{\alpha\beta}}{\partial r_\beta} + g_\alpha. \tag{2}$$

In the above equations, ρ is the fluid density and $p = c_s^2\rho$ the internal pressure of the fluid, while $c_s = \sqrt{K_B T/m}$ is the isothermal speed of sound (a constant in the model), with m and T being the mass and temperature respectively. The α-th component of the velocity $\boldsymbol{u}$ is denoted with u_α. The term g_α refers to the α-th component of the internal and external forces with density $\boldsymbol{g}$. The viscous stress tensor $\boldsymbol{\sigma}$ in Eqs. 1 and 2 is given by

$$\sigma_{\alpha\beta} = \eta_s\left(\frac{\partial u_\beta}{\partial r_\alpha} + \frac{\partial u_\alpha}{\partial r_\beta} - \frac{2}{3}\frac{\partial u_\gamma}{\partial r_\gamma}\delta_{\alpha\beta}\right) + \eta_b \frac{\partial u_\gamma}{\partial r_\gamma}\delta_{\alpha\beta}. \tag{3}$$

The continuum hydrodynamical equations emerge after a suitable multiscale Chapman-Enskog expansion of the Boltzmann kinetic equations is made [2]. The memory of the kinetic description remains in the definition of the transport coefficients η_s and η_b in terms of the relaxation to local equilibrium of the kinetic evolution (see below). The LBM equation evolves in time the discretized probability density function $f_i(\boldsymbol{r}, t)$ to find at position $\boldsymbol{r}$ and time t a fluid particle with velocity $\boldsymbol{c}_i$ according to the LBM updating scheme

$$f_i(\boldsymbol{r} + \boldsymbol{c}_i, t+1) = f_i(\boldsymbol{r}, t) + \Delta_i + \Delta_i^g \tag{4}$$

with the time step set to a unitary value. The collision operator Δ_i expresses the relaxation of the probability distribution function toward the local equilibrium $f_i^{(eq)}$

$$\Delta_i = \sum_j \mathcal{L}_{ij}\left(f_j - f_j^{(eq)}\right) \tag{5}$$

where the equilibrium distribution is a result of the projection of a local Maxwellian onto the lower-order (second) Hermite polynomials [6, 7]

$$f_i^{(eq)} = w_i \rho\left[1 + \frac{u_\alpha c_{i\alpha}}{c_s^2} + \frac{u_\alpha u_\beta\left(c_{i\alpha}c_{i\beta} - \delta_{\alpha\beta}\right)}{2c_s^4}\right] \tag{6}$$

and the weights w_i are a priori known through the choice of a suitable quadrature [8]. Moreover, in Eq. 14 the term Δ_i^g stems from the discretization of the external forcing contributions. Practical applications based on the LBM are largely dominated by the single-time relaxation (BGK form, for Bhatnagar-Gross-Krook [9]) identifying the collisional operator with a diagonal operator. This is mostly a matter of simplicity, for it is known that BGK suffers some obvious physical limitations (with just a single relaxation parameter, the transport of mass, momentum, and energy is forced to proceed at the same pace). To cope with these problems, a number of authors (see [6, 7] and references therein) have developed optimized versions of the original collision matrix in scattering form [10], now known as MRT (multiple relaxation time). The operator $\mathcal{L}_{ij}$ in Eq. 5 is constructed in such a way as to be diagonal in the so-called mode space: the basis vectors $\boldsymbol{e}_k$ of mode space are constructed by orthogonalizing polynomials of the dimensionless velocity vectors [6, 7]. The basis vectors are used to calculate a complete set of moments, the so-called modes $m_k = \sum_i e_{ki} f_i$. The lowest-order modes are associated with the hydrodynamic variables. In particular, the zeroth order momentum gives the density

$$\rho = m_0 = \sum_i f_i. \tag{7}$$

The next three momenta $\tilde{\boldsymbol{m}} = (m_1, m_2, m_3)$ are related to the velocity of the fluid

$$\boldsymbol{u} \equiv \frac{1}{\rho}\tilde{\boldsymbol{m}} + \frac{1}{2\rho}\boldsymbol{g} = \frac{1}{\rho}\sum_i f_i \boldsymbol{c}_i + \frac{1}{2\rho}\boldsymbol{g}. \quad (8)$$

The other modes are the bulk and the shear modes (associated with the viscous stress tensor) and groups of kinetic modes which do not emerge at the hydrodynamical level [6]. Since the operator $\mathcal{L}_{ij}$ is diagonal in mode space, the collisional term describes a linear relaxation of the nonequilibrium modes:

$$m_k \to (1 + \lambda_k) m_k + m_k^g \quad (9)$$

where the relaxation frequencies $-\lambda_k$ (i.e., the eigenvalues of $-\mathcal{L}_{ij}$) are related to the transport coefficients of the modes. After the Chapman-Enskog expansion [2], one gets the following expression for the hydrodynamical transport coefficients entering in the Navier–Stokes equations:

$$\eta_s = -\rho c_s^2 \left(\frac{1}{\lambda_s} + \frac{1}{2}\right) \quad \eta_b = -\frac{2}{3}\rho c_s^2 \left(\frac{1}{\lambda_b} + \frac{1}{2}\right). \quad (10)$$

The term m_k^g is related to the k-th moment of the forcing source Δ_i^g associated with the forcing term with density $\boldsymbol{g}$. While the forces have no effect on the mass density, they transfer an amount $\boldsymbol{g}$ of total momentum to the fluid in one time step. The forcing term is determined in such a way that the hydrodynamical Eqs. 1 and 2 are recovered and can be written as [11]

$$\Delta_i^g = \frac{w_i}{c_s^2}\left(\frac{2 + \lambda_M}{2}\right) g_\alpha c_{i\alpha} + \frac{w_i}{c_s^2}\left[\frac{1}{2c_s^2} G_{\alpha\beta}\left(c_{i\alpha}c_{i\beta} - c_s^2 \delta_{\alpha\beta}\right)\right], \quad (11)$$

where the components of the tensor $\boldsymbol{G}$ are defined as

$$G_{\alpha\beta} = \frac{2 + \lambda_s}{2}\left(u_\alpha g_\beta + g_\alpha u_\beta - \frac{2}{3} u_\gamma g_\gamma \delta_{\alpha\beta}\right) + \frac{2 + \lambda_b}{3} u_\gamma g_\gamma \delta_{\alpha\beta}. \quad (12)$$

Summarizing, the basic LBM is based on a very transparent stream-collide microscopic mechanism, from which the familiar advection–diffusion structure of hydrodynamic equations spontaneously emerges in the limit of small Knudsen number [2], i.e., the ratio of the molecular mean free path to the typical macroscopic length. The stream-collide paradigm offers several advantages over the advection–diffusion representation, primarily the fact that information always travels along straight constant lines, rather than spacetime varying material lines, regardless of the spacetime complexity of the fluid configuration. This is key to mathematical transparency of LBM and their ensuing computational efficiency, especially on parallel computers.

Nonideal LBM Models

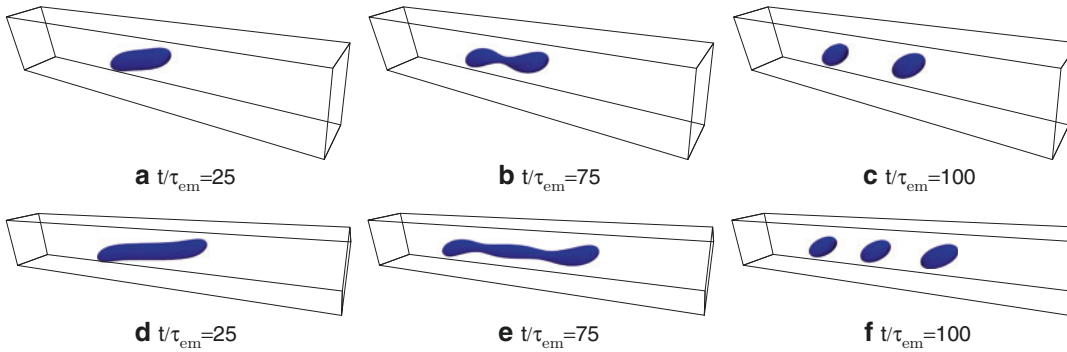
Multiphase fluids, such as droplets and bubbles, are commonly found in recent microfluidic applications, such as biotech lab-on-a-chips. One of the most brilliant outgrowths of the basic LBM theory is the extension to the case of interacting (nonideal) fluids. Recent work is also hinting at the possibility that multiphase LBM may provide a methodological framework for the description of nonideal fluid-fluid and fluid–solid interactions, which play a crucial role for microfluidic and nanofluidic applications [12]. For example, the possibility to model slip boundary conditions and wetting properties [13, 14] has been recently achieved within the framework of the LBM. More detailed comparisons between the mesoscopic technique and atomistic molecular dynamics simulations [15, 16] have pointed out that lattice Boltzmann may become a method of choice for physical problems where supramolecular details play a major role. By supramolecular, we refer to situations which escape a purely continuum treatment and yet still exhibit sufficient universality to do away with a fully atomistic description. Arguably, a wide range of multiphase flows out of equilibrium falls within this class. LBM turned out to be a very effective method to describe mesoscopic physical interactions and nonideal interfaces coupled to hydrodynamics.

Many multiphase and multicomponent LB models have been developed, on the basis of different points of view, including the Gunstensen model [17], the free-energy model [18], and the "Shan-Chen" (SC) model [4], the latter being widely used thanks to its simplicity and efficiency in representing interactions between different species and different phases [19–22]. Here we shall give a few details of the latter method which relies on the modeling of nonideal interactions directly on the lattice. The SC approach augments the LBM equation with the following force term:

$$\boldsymbol{g} - \mathcal{G}\psi(\boldsymbol{x}) \sum_{i \in NS} w_i \psi(\boldsymbol{x} + \boldsymbol{c}_i)\boldsymbol{c}_i \qquad (13)$$

where $\mathcal{G}$ is a constant of proportionality dictating the overall strength of these nonideal interactions and ψ is a density-dependent pseudopotential, incorporating the effects of potential energy interactions through a simple nearest-shell (NS) coupling. One important feature of the multiphase LBM is that phase separation is achieved automatically from the interactions, if only the interaction strength $\mathcal{G}$ exceeds a critical value. This identifies the strength $\mathcal{G}$ as an inverse temperature or more precisely as the ratio between potential and thermal energy, $k_B T$.

An interesting application to micro- and nanofluidics is the study of droplet deformation and breakup [23, 24]. These problems are relevant in the characterization of emulsions, which play an important role in a huge variety of applications, including foods, cosmetics, chemicals, and material processing [25]. Deformation, breakup, and coalescence of droplets occur during flow, and the control over these processes is imperative to synthesize the desired macroscopic behavior of the emulsion. The problem is challenging also from the theoretical point of view: it is intrinsically multiscale, as it bridges between the "microscopic" dynamics of single constituents (i.e., droplets) and the macroscopic behavior of the emulsion. Most times, the synthesis of the emulsion takes place under confinement: this is the case of microfluidic technologies, which are gaining increasing importance as a promising technological route for the emulsion fabrication [12]. The "single" droplet problem has been considered to be the simplest model: in the case of dilute emulsions with negligible droplet-droplet interactions, the dynamics of a single droplet provides indeed complete information about the emulsion behavior. An example of droplet deformation in a confined microfluidics environment is provided in Fig. 1. Other applications of LBM in the context of multiphase fluids in



Lattice Boltzmann Methods for Nanofluidics, Fig. 1 Droplet deformation in a confined microfluidic channel obtained with the LBM. The droplet is placed in a confined microfluidic environment and is sheared by the application of opposite velocities at the confining walls. Breakup takes place after the startup of the shear flow. We report the time history of droplet deformation and breakup including three representative time frames: initial deformation (*left column*); deformation prior to breakup (*middle column*); post breakup frame (*right column*). The simulation time is made dimensionless with respect to the droplet emulsion time τ_{em}, representing the characteristic time for viscosity to damp the capillary waves at the interface. The *top panel* reports an unbounded case, where the droplet breaks into two equally sized daughter droplets. The effect of confinement is visible in a different breakup mechanism (*bottom panel*), leading to a triple breakup

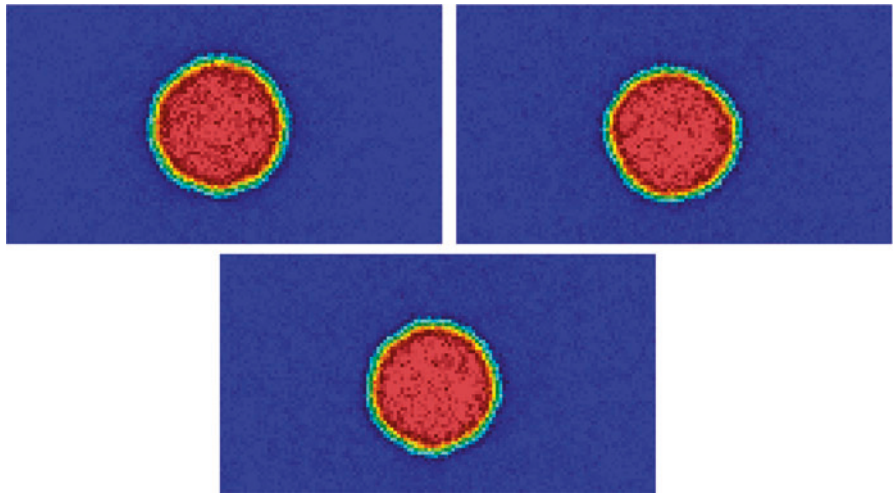
microfluidics pertain to the modeling of droplets in more complex geometries (T-shaped channels, flow-focusing devices) [26] and fluids with complex rheological properties [19].

Thermal Fluctuations

Thermal fluctuations are crucial for a proper mesoscale description of a broad variety of flows in soft matter and biological physics [27]. Theoretically, thermally fluctuating mesoscopic flows are most conveniently dealt within the framework of fluctuating hydrodynamics [28]. This approach, pioneered by Landau and Lifshitz for simple fluids, promotes the nonequilibrium fluxes to stochastic variables, thereby enabling their statistical mechanical description. The idea of including noise in LBM is also an active research field, as witnessed by the various publications of the recent years [6, 29–31]. The basic idea has been pioneered by Ladd [30], who suggested the introduction of noise on the nonconserved hydrodynamic modes, thus reproducing fluctuating viscous stresses in the corresponding hydrodynamic limit (small wavevectors). However, besides the hydrodynamic modes, there exist also higher-order degrees of freedom, the so-called ghosts [31]. The ghost sector, which is coupled to the transport sector at small length scales, acts as a sink for the thermal stress fluctuations and thereby compromises the balance between fluctuation and dissipation. Consequently, the thermalization of the fluid remains incomplete. Adhikari et al. [31] were the first to recognize the necessity to include noise on all the nonphysical ghost modes, and Dünweg et al. [6] reformulated this approach to follow a detailed-balance condition. All these implementations, however, consider ideal-gas descriptions. Only recently there was significant progress in extending fluctuating LBM to nonideal equations of state for single component fluids [29]. In the works by Gross et al. [29], a theoretical framework for the fluctuating LBM is provided based on the theory of linear regression of fluctuations due to Onsager and Machlup [32]. The fluctuating LBM augments the LBM equation with a stochastic term

$$f_i(\boldsymbol{r}+\boldsymbol{c},t+1)=f_i(\boldsymbol{r},t)+\Delta_i+\Delta_i^g+\hat{\Delta}_i. \quad (14)$$

The fluctuating terms $\hat{\Delta}_i$ are assumed to be zero-mean Gaussian random variables, uncorrelated in time and with constant variances (which may however be space dependent). The derivation of the precise expression of the noise covariance has been a central aspect of many recent works [6, 29]. In Fig. 2 we report results of fluctuating LBM simulations for a droplet in a nanofluidic channel moving under the effect of thermal fluctuations only. Thermal fluctuations in LBM represent an open research topic, with many important applications in confined microfluidic and nanofluidic channels, where the usual continuum description of hydrodynamics is coupled with small-scale noise mimicking molecular agitation. Many open ends still remain as to the implementation of fluctuating LBM, especially in the regime of strong fluctuations, where LBM could serve as an effective time-averaged molecular dynamics, with no coarse-graining in space. Such an extension is fascinating and highly nontrivial, as it involves a tangle of fundamental nanoscale physics, in fact the thermodynamics of small systems, and highly technical implementation details. Among the fundamental issues, an outstanding one is whether the fluctuation-dissipation theorem (FDT) still holds under strong confinement and, if not, how should the fluctuating LBM be modified in order to accomodate departures from FDT. Another important issue regards the extension of LBM to the regime of strong thermal fluctuations, $\theta \equiv k_B T/mc_s^2 = O(1)$. Current fluctuating LBM operate around $\theta \sim 10^{-4}$, which leaves outside many emergent nanoscale applications (Casimir effects). On a similar vein, it is still unclear how far one can take LBM down the spatial ladder before touching the barrier of molecular individualism. Such question is paramount to extend LBM to applications where subnanometric resolution is a must, such as the study of hydrogen-bond networking in water-like fluids [33]. These are only some of the basic questions that must be faced in order to further LBM research into the fascinating territory of nanoscale fluid dynamics.



Lattice Boltzmann Methods for Nanofluidics, Fig. 2 Droplet migration due to thermal fluctuations in a confined nanofluidic channel. The results are obtained with the fluctuating LBM proposed by Gross et al., which augments the traditional LBM description with stochastic terms mimicking molecular agitation [29]

Conclusions

The LBM was born as a computational alternative to the discretization of the partial differential equations of continuum mechanics (Navier–Stokes equations), hence in connection with the numerical simulation of macroscopic fluids. In the recent years, however, LBM has made proof of a remarkable versatility in describing a wide host of micro- and nanofluidic situations for which the continuum picture may become inaccurate or even physically untenable and yet sufficiently universal to dispense with atomistic details. The LBM portfolio of applications keeps expanding across scales of motion, particularly toward micro- and nanofluidics [3, 5, 19] with active thermal fluctuations [6, 29]. Realizing the full potential of the method in the latter case, however, still raises many challenges ahead.

MS and LB kindly acknowledge funding from the European Research Council under the EU Seventh Framework Programme (FP7/2007-2013) / ERC Grant Agreement no [279004]. SS kindly acknowledges funding from the European Research Council, project Nanojets.

Cross-References

▶ Droplet Microfluidics
▶ Microfluidic Devices

References

1. Succi, S.: The Lattice Boltzmann Equation for Fluid Dynamics and Beyond. Oxford University Press, New York (2001)
2. Wolf-Gladrow, D.A.: Lattice-Gas Cellular Automata and Lattice Boltzmann Models: An Introduction. Springer, New York (2001)
3. Zhang, J.: Lattice Boltzmann Method for Microfluidics: models and applications. Microfluid. Nanofluid. **10**, 1–28 (2011)
4. Shan, X., Chen, H.: Lattice Boltzmann Model for simulating multiple phases and components. Phys. Rev. E **47**, 1815–1819 (1993)
5. Sega, M., Sbragaglia, M., Kantorovich, S.S., Ivanovd, A.O.: Mesoscale Structures at Complex Fluid-Fluid Interfaces: a Novel LB/ MD Coupling. Soft Matter **9**, 10092–10107 (2013)
6. Dünweg, B., Schiller, U.D., Ladd, A.J.C.: Statistical mechanics of the fluctuating lattice Boltzmann equation. Phys. Rev. E **76**, 036704 (2007)
7. D'Humieres, D., Ginzburg, I., Krafczyk, M., Lallemand, P., Luo, L.-S.: Multiple − relaxation − time lattice Boltzmann models in three dimensions. Proc. Roy. Soc. Lond. A **360**, 367 (2000)

8. Shan, X., Yuan, X.F., Chen, H.: Kinetic theory representation of hydrodynamics: a way beyond the Navier–Stokes equation. J. Fluid Mech. **550**, 413–441 (2006)
9. Bhatnagar, P., Gross, E.P., Krook, M.: A model for collision processes in gases. I. Small amplitude processes in charged and neutral one-component systems. Phys. Rev. A **94**, 511 (1954)
10. Higuera, F., Jimenez, J.: Boltzmann Approach to Lattice Gas Simulations. Europhys. Lett. **9**, 663 (1989)
11. Guo, Z., Zheng, C., Shi, B.: Discrete lattice effects on the forcing term in the lattice Boltzmann method. Phys. Rev. E **65**, 046308 (2002)
12. Squires, T.M., Quake, S.R.: Microfluidics: Fluid physics at the nanoliter scale. Rev. Mod. Phys. **77**, 977 (2005); Tabeling, P.: Introduction a la Microfluidique. Belin, Paris (2003)
13. Benzi, R., Biferale, L., Sbragaglia, M., Succi, S., Toschi, F.: Mesoscopic Modelling of a Two-Phase Flow in Presence of the Boundaries: the Contact Angle. Phys. Rev. E **74**, 021509 (2006); Mesoscopic Two-Phase Model for Describing Apparent Slip in Microchannel Flows. Europhys. Lett. **74**, 651 (2006)
14. Harting, J., Kunert, C., Herrmann, H.: Lattice Boltzmann simulations of apparent slip in hydrophobic microchannels. Europhys. Lett. **75**, 328 (2006)
15. Sbragaglia, M., Benzi, R., Biferale, L., Succi, S., Toschi, F.: Surface roughness-hydrophobicity coupling in microchannel and nanochannel flows. Phys. Rev. Lett. **97**, 204503 (2006)
16. Horbach, J., Succi, S.: Lattice Boltzmann versus molecular dynamics simulation of nanoscale hydrodynamic flows. Phys. Rev. Lett. **96**, 224503 (2006)
17. Gunstensen, A.E., Rothman, D.H., Zaleski, S.: Lattice Boltzmann model of immiscible fluids. Phys. D **47**, 47–52 (1991); Latva-Kokko, M., Rothman, D.H.: Diffusion properties of gradient-based lattice Boltzmann models of immiscible fluids. Phys. Rev. E **71**, 056702 (2005)
18. Swift, M.R., Osborn, W.R., Yeomans, J.M.: Lattice Boltzmann Simulation of Nonideal Fluids. Phys. Rev. Lett. **75**, 830–833 (1995); Briant, A.J., Wagner, A.J., Yeomans, J.M.: Lattice Boltzmann simulations of contact line motion. I. Liquid-gas systems. Phys. Rev. E **69**, 031602 (2004); Briant, A.J., Yeomans, J.M.: Lattice Boltzmann simulations of contact line motion. II. Binary fluids. Rev. E **69**, 031603 (2004)
19. Sbragaglia, M., Benzi, R., Bernaschi, M., Succi, S.: The emergence of supramolecular forces from lattice kinetic models of non-ideal fluids: applications to the rheology of soft glassy materials. Soft Matter **8**, 10773–10782 (2012)
20. Sbragaglia, M., Benzi, R., Biferale, L., Succi, S., Sugiyama, K., Toschi, F.: Generalized lattice Boltzmann method with multirange pseudopotential. Phys. Rev. E **75**, 026702 (2007); Falcucci, G., Bella, G., Chiatti, G., Chibbaro, S., Sbragaglia, M., Succi, S.: Lattice Boltzmann models with mid-range interactions. Comm. Comp. Phys. **2**, 1071–1084 (2007)
21. Gross, M., Moradi, N., Zikos, G., Varnik, F.: Shear stress in nonideal fluid lattice Boltzmann simulations. Phys. Rev. E **83**, 017701 (2011)
22. Jansen, F., Harting, J.: From Bijels to Pickering emulsions: A lattice Boltzmann study. Phys. Rev. E **83**, 046707 (2011)
23. Xi, H., Duncan, C.: Lattice Boltzmann simulations of three-dimensional single droplet deformation and breakup under simple shear flow. Phys. Rev. E **59**, 3022–3026 (1999)
24. Komrakovaa, A.E., Shardt, O., Eskinb, D., Derksen, J. J.: Lattice Boltzmann simulations of drop deformation and breakup in shear flow. Int. J. Multiphase Flow **59**, 23–43 (2014)
25. Larson, R.G.: The Structure and Rheology of Complex Fluids. Oxford University Press, New York (1999)
26. Van der Graaf, S., Nisisako, T., Schroen, C., van der Sman, R., Boom, R.: Lattice Boltzmann simulations of droplet formation in a T-shaped microchannel. Langmuir **22**, 4144–4152 (2008)
27. Chaikin, P.M., Lubensky, T.C.: Principles of Condensed Matter Physics. Cambridge University Press, Cambridge (1997)
28. Landau, L.D., Lifshitz, E.M.: Fluid Mechanics. Pergamon, New York (1959)
29. Gross, M., Adhikari, R., Cates, M.E., Varnik, F.: Thermal fluctuations in the lattice Boltzmann method for nonideal fluids. Phys. Rev. E **82**, 056714 (2010); Gross, M., Cates, M.E., Varnik, F., Adhikari, R.: J. Stat. Mech.: Langevin theory of fluctuations in the discrete Boltzmann equation. Theory. Exp. P03030 (2011)
30. Ladd, A.J.C.: J. Fluid Mech. **271**, 285 (1994)
31. Adhikari, R., Stratford, K., Cates, M.E., Wagner, A.J.: Fluctuating lattice boltzmann. Europhys. Lett. **71**, 473 (2005)
32. Onsager, L., Machlup, S.: Fluctuations and Irreversible Processes. Phys. Rev. **36**(1505) (1953)
33. Moradi, N., Greiner, A., Melchionna, S., Rao, F., Succi, S., A hydro-kinetic scheme for the dynamics of hydrogen bonds in water-like fluids, Phys. Chem. Chem. Phys. 16, 15510–15518 (2014).

LC Filters

▶ Tunable RF MEMS Filters: A Review

Left-Handed Materials

▶ Theory of Optical Metamaterials

Lesson from Nature

► Biomimetics

Life Cycle

► Fate of Manufactured Nanoparticles in Aqueous Environment

Ligand-Directed Gold-Phage Nanosystems

Lawrence F. Bronk[1], Wouter H. P. Driessen[1], Julianna K. Edwards[1], Wadih Arap[1,2] and Renata Pasqualini[1]
[1]David H. Koch Center, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA
[2]Departments of Genitourinary Medical Oncology and Cancer Biology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA

Synonyms

Nanoassembly; Nanoscaffold

Definition

The nanoshuttle is a relatively novel nanosystem formed by the self-assembly of colloidal gold nanoparticles with M13 bacteriophage, which occurs due to electrostatic interactions. The nanoshuttle is a fractal network of phage and gold nanoparticles that maintains the phage's ability to selectively target cells and become internalized via specific cellular receptors. The nanoshuttle also has the unique ability to efficiently incorporate other molecules and nanoparticles of interest and "shuttle" them to a specific tissue in a ligand-directed manner, which has diverse applications for both medical diagnostics and therapeutic strategies.

Procedures

Nanoparticle Synthesis via Citrate Reduction

The standard nanoshuttle is created by mixing phage and gold nanoparticles and allowing time for the assembly to take place. Colloidal gold nanoparticles are synthesized via a citrate reduction method [1]. First, 1 % (w/v) solutions of auric chloride ($Au(III)Cl_3$) and sodium citrate in ultrapure deionized water are prepared. The solutions are then sterile filtered and stored at 4 °C. Clean glassware is filled with ultrapure deionized water, and a magnetic stir bar is added. Typical starting volumes of water are 100–500 mL. Preferably, the citrate reduction takes place while boiling under reflux, but alternatively, the flask can be capped with aluminum foil to reduce evaporative losses. Once the water is brought to a boil under a fume hood while stirring, the aluminum foil is carefully peeled back and 1 % of the water volume of the 1 % $Au(III)Cl_3$ solution is added and stirred for 10 s before quickly adding 0.8 % of the water volume of the 1 % sodium citrate solution. For example, if making a batch of gold nanoparticles with a starting volume of 100 mL of water, 1 mL of the 1 % $Au(III)Cl_3$ solution and 800 μL of the 1 % sodium citrate solution would be added. The aluminum cap is then placed back on the flask, and nanoparticle formation can be observed. Within 2 min, the solution changes from clear to cloudy, to dark gray, to black/dark purple and finally settles at red. The solution is allowed to boil for an additional 3 min after the color has settled before the aluminum foil cap is removed, and the solution is allowed to boil off approximately 80 % of its volume in order to concentrate the gold nanoparticle solution. Alternatively, the gold nanoparticle solution can be concentrated via centrifugation (7000 rcf for 10 min); however, this process is more time, work, and materially intensive, especially for larger batches.

The gold nanoparticle solution can then be evaluated by measuring the visible and ultraviolet

(UV–vis) absorbance spectrum from 350 to 900 nm. Two dilutions of the nanoparticle solution, 1:5 and 1:10, are tested. A narrow absorbance peak centered at 530 nm, which corresponds to a nanoparticle size of 44 ± 8 nm, should be observed. If the peak absorbance is not at 530 nm, the size of the nanoparticles can be determined following [2]. To resynthesize nanoparticles using the same materials, the size of the resulting nanoparticles can be adjusted by using different ratios of the 1 % $Au(III)Cl_3$ and the sodium citrate solutions to obtain the desired size. A solution with a lower ratio of $Au(III)Cl_3$ to sodium citrate will yield smaller nanoparticles whereas a higher ratio will yield larger nanoparticles. The optical density of the 530-nm peak for the undiluted nanoparticle solution should be between 4 and 5. The nanoparticle solution can then be sterile filtered, wrapped in aluminum foil, and stored at 4 °C for later use.

Phage Production and Quantification

Bacteriophage particles are produced and quantified following protocols described in [3].

Bottom-Up Assembly of Nanoshuttles

Once the gold nanoparticles have been prepared, the baseline nanoshuttles are synthesized by selectively adding phage, gold nanoparticles (AuNP), and either an imidazole hydrochloride solution or deionized water. There are two currently used formulations of the nanoshuttle: AuNP/phage and AuNP/phage/imidazole. The difference between the two is the presence of imidazole, a metal chelating agent that induces gold nanoparticle clustering. To produce nanoshuttles that incorporate imidazole, 100 μL of a 1 M solution of imidazole hydrochloride is added to rows B through H and 180 μL are added to row A of a standard 96-well plate. Twenty microliters of phage with a titer of approximately 1010 TU/μL is then added to each well in row A. Using a pipette, 100 μL of the phage-imidazole solution in row A is added to row B and pipetted up and down to ensure the solution is well mixed, and then 100 μL of the phage-imidazole solution from row B is added to row C. This 2× dilution process is repeated until row G. Every well in rows A–G should now have 100 μL of a phage-imidazole solution, and each row should correspond to a 2× dilution compared to its preceding row. Row H serves as a control and should contain 100 μL of 1 M imidazole only. To synthesize nanoshuttles without the presence of imidazole, deionized water is substituted for the 1 M imidazole solution. Lastly, a volume of 100 μL of gold nanoparticles is added to every well. The plate is then covered, and nanoshuttles are allowed to form overnight at room temperature. Visible nanoshuttle formation should occur within minutes of adding the gold nanoparticles. A color change from red to purple is observed in all columns after two dilutions (row C). The next day, the UV–vis absorbance spectrum of each well containing AuNP/phage is measured from 350 to 900 nm. The final nanoshuttle solution is made by mixing the rows where nanoshuttle formation has taken place. When there is an insufficient concentration of phage to keep the clustered gold nanoparticles in solution, the gold will precipitate and coat the bottom of the well. This can be observed by eye as well as on the UV–vis spectrum by comparing the near-infrared (NIR) spectrum of the control imidazole-gold row (H) to the phage-gold-imidazole rows (A–G). The gold-imidazole spectrum will have a linearly increasing absorbance throughout the NIR region whereas the phage-gold-imidazole will not. The NIR absorbance of the nanoshuttles will increase as the relative gold-imidazole to phage ratio increases. If necessary, one can determine whether the phage-gold-imidazole solution has precipitated manually by pipetting up the solution to see if it aspirates easily. If it does not, then the gold-imidazole has likely fallen out of solution. Following this procedure, rows A–F are typically combined in sterile glassware or polypropylene FACS tubes to form the nanoshuttle solution. The resulting nanoshuttle solution is then concentrated via centrifugation at 44 rcf for 2 min. The supernatant is then pipetted out and can be saved for further analysis such as phage or gold concentration, or discarded. The final solution of phage-gold-imidazole is the nanoshuttle.

Incorporation of Additional Materials into the Nanoshuttle

The baseline nanoshuttle serves as a scaffold or backbone enabling the incorporation of other materials, such as diagnostic or therapeutic compounds. So far, silicon and various magnetite nanoparticles have been successfully incorporated into the scaffold [4, 5]. To incorporate other components into the nanoshuttle, the synthesis protocol is slightly changed. The compound that needs to be incorporated is first mixed in with the gold nanoparticle solution before adding it to the phage-imidazole solutions.

Results

Biological Characterization of the Nanoshuttle

The initial characterization studies of the nanoshuttle focused on determining its physical and chemical properties. The base nanoshuttle was first synthesized by determining the phage input concentration needed to transform colloidal gold solutions into hydrogels. Filamentous phage were chosen as the backbone of the nanoshuttle because of their many unique and desirable properties including the ease of production in host bacteria compared to the complex chemistry and by-products involved in polymer synthesis. Other advantages of the phage backbone include the fact that the phage pIII minor capsid protein can be engineered to display targeting peptides, as well as the phage's innate resilience to severe conditions such as high salinity, acidic pH, high temperatures, chaotropes, and prolonged storage [1]. Altering the content of the phage-gold nanoparticle solution by adjusting the phage or gold concentration or by adding imidazole allows for precise tuning of the system for different purposes. Once the prototype nanoshuttle was formed, it was found that the phage backbone of the system still maintained its ability to infect bacteria and target mammalian cells. The nanoshuttle was first tested for biocompatibility and cytotoxicity in C17.2 murine neural stem cells. It was found that the cells incorporated into the nanoshuttle solution and continued to proliferate, indicating that the system was suitable for further biomedical study [1]. Since the initial studies, the nanoshuttle has been tested on multiple cell lines with the same biocompatibility and lack of cytotoxicity previously observed. The nanoshuttle has also been used in murine models of cancer with comprehensive toxicity studies under way.

Fractal Analysis of the Nanoshuttle

The nanoshuttle system was physically characterized using fractal dimension analysis because many self-assembly processes contain fractal geometry. Images of the nanoshuttle were acquired using transmission electron microscopy (TEM) [1]. Using the TEM images, a two-dimensional fractal analysis showed that the nanoshuttle with imidazole had a higher fractal dimension than without imidazole: 1.78 ± 0.14 versus 1.32 ± 0.12 [1, 6]. This observation corresponded to an increase in the density of gold nanoparticle packing by imidazole-induced aggregation. These results were consistent with measurements made using angle-dependent elastic light-scattering fractal dimension analysis of the nanoshuttle in solution. The nanoshuttle's fractal geometry likely improves its binding ability by making the phage pIII minor capsid protein more available to cellular receptors due to the high surface area of fractals.

Mechanism of Assembly

The binding mechanism between the phage and the gold nanoparticles was also investigated. It was hypothesized that the phage pVIII major capsid coat proteins served as the binding site for the gold nanoparticles. Given the fact that there are no typical metal-binding amino acids such as histidine on the pVIII major capsid, it was suggested that the binding must be the result of electrostatic interactions between the two components [1]. It is well known that varying the ionic conditions of the nanoparticle solution can induce gold nanoparticle aggregation. Using a competitive binding assay between an ionic solution and phage, it was found that the presence of phage reduced the amount of gold nanoparticle aggregation as the phage's input titer was increased [1]. This result implied that phage and gold nanoparticle binding

was similar to gold nanoparticle and gold nanoparticle binding. Testing the formation of the nanoshuttle in solutions of varying pH further validated this hypothesis. When the pH of the solution was greater than 7, nanoshuttle formation did not occur indicating that the mechanism for nanoshuttle formation is the electrostatic interaction between the gold nanoparticle and phage. Nanoshuttle assembly with a mutant phage was tested to further validate an electrostatic mechanism of assembly. This phage had its pVIII major capsid mutated to contain more cationic amino acid residues compared to the wild-type phage. By altering the pH of the nanoshuttle assembly solution, it was found to still be possible to induce the mutant phage to form nanoshuttles under conditions where the wild-type phage could not [7]. These results illustrate the fine control that one has over nanoshuttle formation, even with variations in its base components.

Optical Characterization of the Nanoshuttle

The optical properties of the nanoshuttle were studied by looking at the UV–vis absorbance spectrum of the system. It had been noted that the assembled nanoshuttle network contained a shift in the surface plasmon absorbance to the near-infrared (NIR) regime, indicating that the average distance between the nanoparticles is less than their diameter [8]. As expected, the more densely packed imidazole-containing nanoshuttles exhibited a larger red shift in their absorbance spectrum than the nanoshuttles without imidazole. To ensure that the NIR photons were actually being absorbed and not just scattered, heating measurements were made on the nanoshuttle solutions under NIR laser illumination. The temperature of the nanoshuttle solution dramatically increased, indicating that the NIR photons are indeed being locally absorbed via the surface plasmon absorption shift into the NIR [1]. It had previously been reported that imidazole-gold chelates have up to a 10^8 increase in NIR surface-enhanced Raman scattering (SERS) signal than imidazole alone [8]. NIR-SERS spectrum measurements of the nanoshuttle were taken to determine whether the gold-imidazole clusters still maintained this reporter property, as well as to characterize the relationship between the phage, imidazole, and gold nanoparticles. By comparing the spectra of the nanoshuttle's constituent components alone to the fully assembled nanoshuttle, it was possible to identify specific NIR-SERS bands intrinsic to the nanoshuttle.

Applications

The list of potential applications of the nanoshuttle continually grows, and the focus has become directed to possible uses within the biomedical sciences. The initial characterization studies found that the nanoshuttle can be used to enhance fluorescence and dark-field microscopy, serve as a NIR-SERS reporter, and convert near-infrared laser light into heat. The gold in the nanoshuttle increases light scattering due to its large elastic scattering cross section, leading to its use as a dark-field microscopy image contrast agent. Treating cells with the targeted nanoshuttle greatly increases dark-field image signal intensity compared with nontargeted nanoshuttle or plain nanoparticles. The discovery that the nanoshuttle had a unique NIR-SERS signal as well as heating properties led to research on noninvasively detecting and ablating tissue, using NIR-SERS and NIR laser light, respectively. These studies are ongoing.

After the initial characterization studies, several new lines of research began simultaneously, and many are still under active investigation. It was found that the gold nanoshuttles could be utilized as a scaffold for three-dimensional tissue culture [5]. Three-dimensional tissue culture better re-creates in vivo conditions in vitro than traditional two-dimensional tissue culture leading to an improved model with a higher degree of translational viability. Three-dimensional tissue culture was achieved by incorporating magnetic iron oxide (MIO) nanoparticles into the nanoshuttle and plating it on top of human glioblastoma cells. After plating, it was possible to levitate the cells using a variety of different magnets and grow them three-dimensionally. The human glioblastoma cells grown using this

method showed higher morphological and molecular similarities to orthotopic human tumor xenografts than did the cells cultured using the traditional two-dimensional method. By varying the geometry of the magnetic field, the shape of the cellular mass could be controlled and was successfully manipulated into two geometries: a traditional spheroid and a torus. Using the nanoshuttle-enabled three-dimensional tissue culture approach also enabled different types of cells to be cocultured in order to simulate the heterogeneity of in vivo tumor masses. These findings led to the start-up company, n3d Biosciences, Inc., which sells this system, known as the Bio-Assembler™, for commercial and academic use.

Future Directions for Research

The nanoshuttle is still a subject of ongoing research on both its basic properties and its applications. Current research into the nanoshuttle's basic properties involves characterizing its loading capabilities of gold and other metallic nanoparticles as well as other nanosystems. The nanoshuttle's in vivo binding properties such as specificity, toxicity, and biodistribution are also under active investigation. Research into the applications of the nanoshuttle is focused on utilizing the system as a multimodal theranostic device for the diagnosis, imaging, and treatment of a wide array of cancers. Depending on the composition of the nanoshuttle, it can act as a contrast agent for various imaging modalities such as NIR-SERS, computed tomography, and magnetic resonance imaging. Loading the nanoshuttle with different substances can also enable it to act as a therapeutic agent for cancer treatments such as targeted chemotherapy, laser ablation, and radiation therapy. Determining how the nanoshuttle can aid or enhance these imaging and treatment modalities is still under investigation. Another line of research currently being pursued is using the nanoshuttle to target other nanomedicine modalities to specific sites in the body. For example, coating other therapeutic agents such as porous silicon nanoparticles with the nanoshuttle provides a vascular targeting capacity, thereby increasing their specificity [4]. Despite the fact that the nanoshuttle is a relatively young nanosystem in the growing field of nanotechnology, it has already proven that it has many significant translational applications warranting further active research.

References

1. Souza, G.R., et al.: Networks of gold nanoparticles and bacteriophage as biological sensors and cell-targeting agents. Proc. Natl. Acad. Sci. U. S. A. **103**(5), 1215–1220 (2006)
2. Khlebtsov, N.G.: Determination of size and concentration of gold nanoparticles from extinction spectra. Anal. Chem. **80**(17), 6620–6625 (2008)
3. Christianson, D.R., et al.: Techniques to decipher molecular diversity by phage display. Methods Mol. Biol. **357**, 385–406 (2007)
4. Decuzzi, P., et al.: Intravascular delivery of particulate systems: does geometry really matter? Pharm. Res. **26**(1), 235–243 (2009)
5. Souza, G.R., et al.: Three-dimensional tissue culture based on magnetic cell levitation. Nat. Nanotechnol. **5**(4), 291–296 (2010)
6. Souza, G.R., et al.: Combinatorial targeting and nanotechnology applications. Biomed. Microdevices **12**(4), 597–606 (2010)
7. Souza, G.R., et al.: Bottom-up assembly of hydrogels from bacteriophage and Au nanoparticles: the effect of cis- and trans-acting factors. PLoS One **3**(5), e2242 (2008)
8. Souza, G.R., et al.: In vivo detection of gold-imidazole self-assembly complexes: NIR-SERS signal reporters. Anal. Chem. **78**(17), 6232–6237 (2006)

Light Emitting Diodes

Talal Al Tahtamouni
Department of Physics, Yarmouk University, Irbid, Jordan

Light-Emitting Diode (LED)

LED is an optoelectronic device; it consists mainly of a p-n junction. LEDs are particularly designed so that the radiative recombination dominates to produce a photon. This radiative

recombination process of electrically injected electrons and holes from n-type and p-type regions occurs efficiently in the active region of LEDs. The wavelength of the emitted photon depends on the bandgap energy of the semiconductor material in the active region.

The Working Principle of LED

The typical structure of LED is a p-n junction diode, mainly consisting of one layer uniformly p-doped in one side and one n-doped on the other side. Once the p-n junction has been created, a space charge layer and a built-in voltage (energy barrier) will automatically be formed due to the carrier balance process. By applying an external forward voltage to the diode, the electrons and holes in the conduction band and valance band respectively see a lower energy barrier and start the minority injection process. The electrons will be injected to the n-type layer and the holes to the p-type layer. When they reach the active region, the electrons in the conduction band recombine with holes from the valence band creating a photon. The wavelength of the emitted photons is associated to the bandgap energy of the active region.

Active Region

The active region is the core of the LED structure, and this region in the LED determines the way that the photons created by electron–hole radiative recombination are emitted. Quantum well structure is typically employed as active region in LED structure.

Quantum Well

A quantum well (QW) is a thin layer sandwiched between two larger–energy gap materials. The width of the QW corresponds to the thickness of the sandwiched layer, so the well width is controlled by changing the thickness of this layer. The QW structures are used as active regions in the LED structure because they have a small volume, therefore when carriers are injected into a QW structure, the free carrier concentration in QW is high so nonradiative recombination probability is less. QW structures give a freedom to design the transition energies through changing the well width.

Recombination

There are two fundamental recombination mechanisms in LED, radiative recombination and nonradiative recombination. During a radiative recombination, electrons and holes recombine followed by the emission of photons, whereas in a nonradiative recombination process, the energy is converted to phonons, generating extra heat to the material.

Turn-On Voltage

The LED turn-on voltage is the voltage at which the LED first emits light or at which the diffusion current starts to dominate. This turn-on voltage depends on the bandgap of the active region.

LED Efficiencies

- **Injection efficiency**
  Injection efficiency of LED is the fraction of injected current that can be captured in the active region for the recombination process.
- **Recombination efficiency**
  Recombination efficiency is the fraction of injected carriers in the active region that recombine radiatively generating photons.
- **The internal quantum efficiency**
  The internal quantum efficiency of LED is the ratio of photons generated inside the active region to the total number of electrons injected into the active region (the efficiency of converting carrier current to photons).
- **The optical extraction efficiency (light extraction efficiency)**
  The optical extraction efficiency of LED is the ratio of photons able to escape from the LED

into free space to photons generated in the active region.

- **Power efficiency (wall-plug efficiency)**
  The power efficiency of LED is the ratio of the light output power to the electrical input power.
- **The total (external) quantum efficiency**
  The external quantum efficiency of LED is the ratio of the extracted photons emitted into free space to the injected electron–hole pairs.

Cross-References

- ▶ Band Alignment
- ▶ Light Emitting Diodes
- ▶ Metalorganic Chemical Vapor Deposition (MOCVD)
- ▶ Molecular Beam Epitaxy (MBE)
- ▶ Near-Field Scanning Optical Microscopy (NSOM)
- ▶ Optical and Electronic Properties
- ▶ Polarization-Induced Effects in Heterostructures
- ▶ SEM
- ▶ Solar Cells
- ▶ Spintronic Devices

Light Localization for Nano-optical Devices

Dmitri K. Gramotnev
Nanophotonics Pty Ltd, Brisbane, QLD, Australia

Synonyms

Metallic waveguides; Nanofocusing of light; Near-field optics; Plasmonic waveguides; Sub-wavelength waveguiding

Definition

Conventional dielectric structures like dielectric waveguides, fibers, lenses, etc., do not allow localization of electromagnetic waves within the regions that are significantly smaller than the wavelength of light in the medium. For example, if a dielectric lens is used to focus a laser beam, then the minimal beam waist at the focal point will be $\sim\lambda_{vac}/(2n)$, where λ_{vac} is the wavelength in vacuum and n is the refractive index of the medium where the laser beam is propagating (and is being focused by the lens). This is the so-called diffraction limit of light. Sub-wavelength localization of light is the focusing or concentration of light energy, or guiding of a light signal within spatial regions that are much smaller than the wavelength and could be as small as a few nanometers in size [1, 2].

Major Physical Principles

The current way of achieving sub-wavelength localization of light and light signals is associated with using material with negative dielectric permittivities. Such materials involve

- Metals below plasma frequency ωp [3]
- Ionic crystals in the infrared range of spectrum between the frequencies of transverse and longitudinal optical phonons [3]
- Any other material near sufficiently sharp line of resonant absorption

However, metals below plasma frequency ωp are the most readily available and commonly used materials with negative dielectric permittivity.

Metals support a special type of electromagnetic waves that are coupled to collective oscillations of the electron plasma in the metal. These waves are called plasmons. A very important and useful physical effect is that metallic surfaces can guide a special type of plasmons that are called surface plasmons. Surface plasmons are somewhat similar to waves on a surface of water, but they occur on a surface of electron plasma – the surface that coincides with the metal interface. The field of surface plasmons exponentially decays into the metal and into the adjacent dielectric medium (i.e., vacuum). Their dispersion on a flat metal surface is given by a simple equation [4]:

$$q = \frac{\omega}{c}\sqrt{\frac{\varepsilon_d \varepsilon_m}{\varepsilon_d + \varepsilon_m}}, \quad (1)$$

where q is the wave number of the surface plasmon, ε_m and ε_d are the dielectric permittivities of the metal and the adjacent dielectric medium, respectively, ω is the plasmon frequency, and c is the speed of light in vacuum. Equation 1 shows that if ε_m is real, i.e., dissipation in the metal is neglected, the surface plasmon exists only if

$$\varepsilon_m < 0 \quad (2)$$

and

$$|\varepsilon_m| > \varepsilon_d. \quad (3)$$

Indeed, if condition (Eq. 2) is not satisfied, i.e., ε_m is positive (which occurs in metals at frequencies $\omega > \omega_p$ [2]), then $q < k_d \equiv \omega\sqrt{\varepsilon_d} = c$, where k_d is the wave number of the bulk wave in the adjacent dielectric medium. This means that the surface plasmon must leak into the bulk wave in the dielectric medium at an angle given by the ratio: $\cos(\theta) = q/k_d$. This angle is determined by the Snell's law that represents the momentum conservation in electrodynamics [5]. In this case, the plasmon is not a surface wave localized near the metal surface and guided by this surface, but is rather a leaky wave that propagates, rather than exponentially decays, in the adjacent dielectric medium. If condition (Eq. 3) is not satisfied but $\varepsilon_m < 0$, then q is imaginary (see Eq. 1) and the surface plasmon does not exist either.

Because the wave number q for a surface plasmon is larger than the wave number k_d for the bulk wave in the adjacent dielectric (see Eq. 1), the tangential component of the wave vector of the bulk wave $(\mathbf{k}_d)t$ can never be equal to the wave vector $\mathbf{q}$ of the surface plasmon. Therefore, the Snell's law (that requires the equality of the tangential components of the wave vectors of the waves involved [5]) and the momentum conservation can never be satisfied by a bulk wave and a surface plasmon (having the same frequencies) on a flat smooth metal surface. Therefore, surface plasmons on a smooth flat surface are not coupled to bulk radiation. However, surface plasmons can be generated by bulk radiation on surface nonuniformities (such as, for example, an edge of the surface – end-fire excitation of surface plasmons [1]), or in periodic gratings, or in the presence of a prism coupler in the Otto or Kretschmann geometries [4].

The recent years have seen an explosion of theoretical and experimental research in the area of surface plasmons and their applications for integrated nano-optics: development of a new generation of nano-optical sensors, near-field imaging techniques with the spatial resolution as good as a few nanometers, manipulation and trapping of small nanoparticles and separate molecules, detection and imaging of separate molecules, design of nanomechanical systems, etc.

These areas of application and research involving surface plasmons have started to rapidly expand when it was realized that metallic nanosructures could guide and concentrate plasmon energy far beyond the diffraction limit of light and within nanoscale regions as small as a few nanometers. The first demonstrations of such unique capabilities were the original papers on sub-wavelength guiding using plasmons in metal nanorods and metallic nanoholes [6], and plasmon nanofocusing in gradually tapered metal structures [7]. In fact, the smallest dimensions of a localization region for surface plasmons are limited only by the atomic structure of matter, dissipation in the metal, and spatial dispersion, i.e., nonlocal dielectric response of materials [5, 8].

Key Research Findings

Sub-wavelength Waveguiding

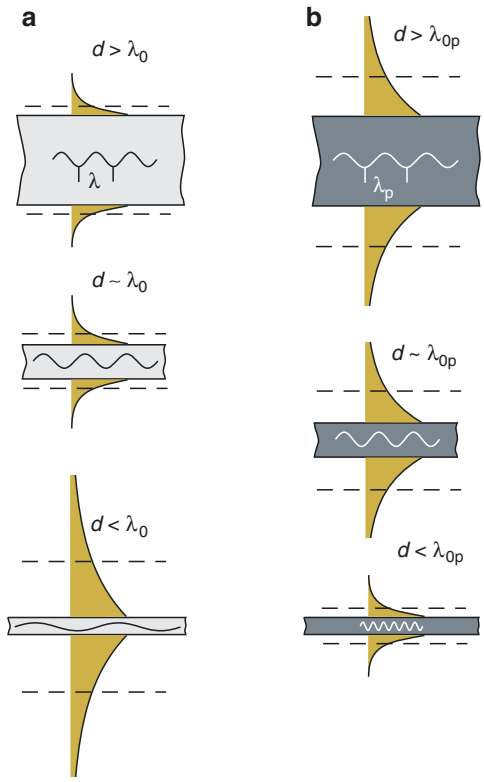
Modern electronic is fast approaching its fundamental speed and bandwidth limits. Therefore, further increase of efficiency and speed of information-processing systems relies upon the development of new approaches and physical principles. One of the most feasible current options is based on using light signals and optical circuits instead of their less efficient electronic counterparts. However, one the major current problems of integrated optics is the insufficient

level of miniaturization of optical components and interconnects, which is not even close to that typically achieved in electronics. Therefore, design of efficient sub-wavelength waveguides, interconnects, and optical devices with the localization of light signals within spatial regions with dimensions much smaller than the wavelength is the major goal of the modern integrated nanophotonics.

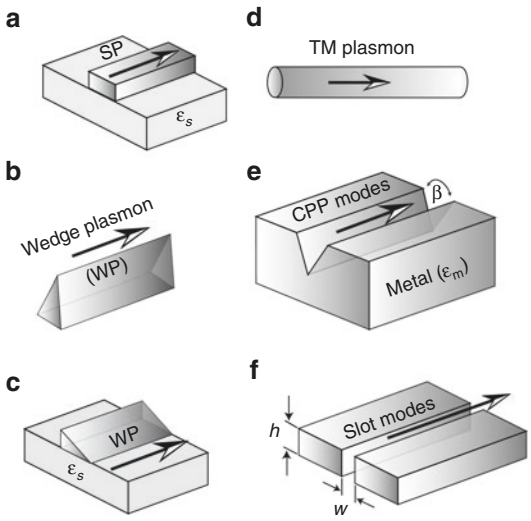
As indicated above, unfortunately, dielectric waveguides do not normally allow such localization. For example, though an optical fiber can be made arbitrarily thin, this does not lead to sub-wavelength localization of the fundamental mode guided by this fiber. As the diameter of the fiber is reduced to zero, the field of the guided mode extends further into the surrounding medium (Fig. 1a), and the size of the guided mode tends to infinity. No two such fibers can be placed close to each other, as the overlapping guided modes in these fibers will produce a significant crosstalk and signal interference.

A completely different situation occurs in sub-wavelength plasmonic waveguides, such as, for example, cylindrical metal nanowires (Fig. 1b). Decreasing diameter of a metal nanowire results in a monotonic decrease of the wavelength of the fundamental guided plasmon mode having the TM polarization. This leads to a monotonic increase of the mode localization or, equivalently, to a monotonic decrease of the dimensions of the localization region for the fundamental mode (these regions are shown by the *dashed lines* in Fig. 1).

Several different guiding plasmonic structures have been proposed and analyzed in detail. These included metal nanorods (cylindrical nanowires), nanostrips on a dielectric substrate, sharp metal wedges guiding wedge plasmons, metallic grooves guiding a special type of strongly localized plasmons called channel plasmon-polaritons (CPP modes), or nanoslots in a thin metal membrane or film [1] – see Fig. 2a–f. It is important to note that not all plasmonic modes guided by different types of metallic guiding nanostructures can experience strong sub-wavelength localization. For example, a thin metal film surrounded by a uniform dielectric can guide two different types of plasmonic modes – with symmetric and antisymmetric charge distribution across the film. If the thickness of the film is reduced to zero, it is only the plasmon with the symmetric charge distribution that experiences increased confinement to the metal film (i.e., increased localization). The other mode with the antisymmetric charge



Light Localization for Nano-optical Devices, Fig. 1 The differences between the dielectric fibers and metallic nanowires in achieving sub-wavelength waveguiding. (**a**) As the fiber diameter d decreases from infinity to zero, the wavelength of the guided fundamental mode λ monotonically increases from the wavelength in the material of the fiber λ_0 to the wavelength in the dielectric medium surrounding the fiber. Simultaneously, the typical mode size reduces as the fiber diameter is reduced to about λ_0 and then starts to increase to infinity (as the fiber diameter tends to zero) – no sub-wavelength localization. (**b**) As the diameter of a cylindrical metal nanowire d decreases from infinity to zero, the wavelength of the guided fundamental TM plasmonic mode λp monotonically decreases from the plasmon wavelength on the flat metal surface $\lambda_0 p$ to zero. Simultanously, the typical mode size also monotonically decreases to zero, leading to sub-wavelength waveguiding



Light Localization for Nano-optical Devices, Fig. 2 Major examples of different types of plasmonic sub-wavelength waveguides: (**a**) metal nanostrip on a dielectric substrate; (**b**) free-standing sharp metal wedge surrounded by a uniform dielectric, guiding strongly localized wedge plasmons; (**c**) tapered section of a metal film in the shape of a metal wedge on a dielectric substrate with the permittivities of the substrate and cladding being εs and εc, respectively; (**d**) cylindrical nanorod or cylindrical nanowire guiding strongly localized TM plasmons; (**e**) V-groove of a metal surface, guiding strongly localized channel plasmon-polaritons (CPP modes); and (**f**) slot plasmonic waveguide in a metal membrane of thin film guiding strongly localized slot modes

distribution across the film – the so-called long-range surface plasmon – becomes less localized and in the limiting case of zero thickness of the metal film transforms into a bulk wave propagating in the surrounding dielectric.

These are the only strongly localized ("short-range") plasmonic modes that present a significant interest for the design of sub-wavelength waveguides, interconnects, and nano-optical devices. The indicated plasmonic structures including metal strips, nanorods, metal grooves, wedges, and nanoslots are all capable of guiding strongly localized plasmonic modes. The localization of the guided "short-range" plasmons increases with decreasing geometrical dimensions of the guiding structure, such as slot width and film thickness for a nanoslot waveguide, taper angle for a V-shaped metallic groove or a metal wedge, diameter of a nanorod, and width and thickness of a metal strip.

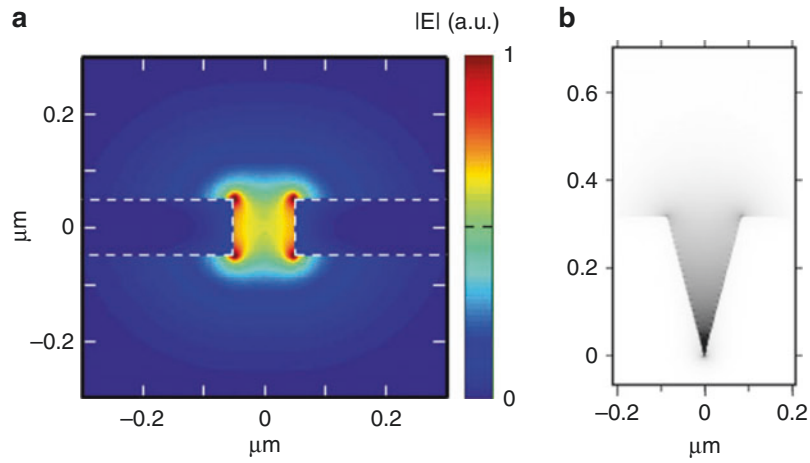
For example, in the approximation of continuous electrodynamics, the dimensions of the guiding region for a strongly localized plasmon mode near the tip of a V-shaped metal groove or metal wedge tend to zero with decreasing taper angle. In fact, there is a critical value of the taper angle $\beta = \beta c$:

$$\beta_c \approx \begin{cases} -2\varepsilon_d = \mathrm{Re}(\varepsilon_m), & \text{for V} - \text{grooces} : \\ -(\varepsilon_s + \varepsilon_c) = \mathrm{Re}(\varepsilon_m) & \text{for wedges,} \end{cases} \tag{4}$$

below which the guided mode does not exist because of its infinite localization near the infinitely sharp tip. Here, ε_d is the permittivity of the dielectric medium in the metal groove (Fig. 2e), and ε_s and ε_c are the permittivities of the dielectric substrate and cladding for a tapered section of a metal film in the form of a metal wedge (Fig. 2c).

The typical distribution of the magnitude of the electric field in the fundamental plasmonic mode of a nanoslot of 100 nm width in a silver membrane of thickness 100 nm suspended in vacuum is shown in Fig. 3a for the vacuum wavelength $\lambda_{vac} = 632.8$ nm (He-Ne laser) and metal permittivity $\varepsilon m = -16.2 + 0.5i$. The localization of the field in this case is ~150 nm, which is significantly smaller than the vacuum wavelength. Reducing the gap width and film thickness to 25 nm results in a significant reduction in the typical dimensions of the field localization region below ~50 nm. The typical field distribution in the fundamental CPP mode in a V-groove waveguide of finite depth is shown in Fig. 3b, also illustrating guiding with sub-wavelength localization.

Increasing localization of a guided plasmonic mode comes at a price of increased dissipative losses and thus decreased propagation distance. For example, for the 100 × 100 nm slot the propagation distance of the fundamental mode (Fig. 3a) appears to be ~15 μm, whereas for the 25 × 25 nm slot it is ~2.3 μm. Similar situation occurs for the other types of the plasmonic waveguides, such as V-groove and wedge waveguides [1].



Light Localization for Nano-optical Devices, Fig. 3 Typical distributions of the magnitude of the electric field in sub-wavelength plasmonic waveguides: (**a**) fundamental mode in a 100 × 100 nm slot waveguide in a silver membrane in vacuum at $\lambda_{vac} = 632.8$ nm (He-Ne laser); (**b**) fundamental mode in a 30° V-shaped 314 nm deep groove filled with vacuum on a silver surface at $\lambda_{vac} = 632.8$ nm

Significant dissipation of the propagating signals in plasmonic sub-wavelength waveguides is a major problem of integrated plasmonics. On the other hand, the typically achievable propagation distances of around at least a few microns are sufficient for the design of interconnects between highly integrated nano-optical devices, or interconnects between the existing optical telecommunication systems and nano-electronic devices and components. However, the development of all-optical information-processing integrated circuits with high level of integration will require a resolution of the dissipation problem.

One of the current approaches to the dissipation problem is to use materials with gain to compensate for the dissipation. However, the high level of dissipation of strongly localized plasmons $\sim$1,000 cm^{-1} and higher is above what could be expected to be compensated by most gain media. Therefore, dissipation compensation or partial dissipation compensation was claimed for rather weakly localized plasmonic modes [1, 9] that cannot normally be used to design nano-optical devices and interconnects. Therefore, strong dissipation of localized plasmonic modes in sub-wavelength waveguides is a standing problem that currently imposes limitations upon the design of highly integrated nanophotonic circuits for efficient all-optical signal processing.

Strong dissipation of plasmonic modes in metallic nanostructures is also an obstacle for the development of nanoscale optical sources. An efficient nanoscale optical source that would result in stimulated and coherent emission of surface plasmons would be a major technological step in nanophotonics. Such a source of surface plasmons is called "spaser" using the analogy with a "laser" as a light source based on stimulated emission of photons. A significant recent advancement in this direction was the first experimental observation of stimulated plasmon emission in gold nanoparticles covered in high-brightness luminescent dye doped silica shells [10]. Each of such gold-silica particles worked as a spaser nanocavity with the dimensions of $\sim$44 nm. Using the optical pumping, plasmon oscillations are stimulated in each of the gold-silica nanoparticles, and outcoupling of these oscillations into radiative photonic modes constituted a nanolaser operating in the visible range and having the cavity dimensions of $\sim$44 nm [10]. Coupling such a spaser cavity to a sub-wavelength plasmonic waveguide could lead to a plasmon source in a plasmonic information-processing circuit.

Plasmon Nanofocusing

A special area of interest and significant applications of strongly localized plasmons lies with plasmon nanofocusing. Nanofocusing is the process of concentration and localization of electromagnetic waves by means of tapered metallic nanostructures into regions as small as a few nanometers, i.e., far beyond the diffraction limit

of light [1, 2]. Plasmon nanofocusing occurs when a strongly localized plasmonic mode propagates along a gradual taper of a plasmonic guiding structure. For example, if a plasmon with TM polarization propagates along a tapered conical rod, it experiences a gradual increase in its localization combined with a significant local field enhancement, despite the dissipative losses in the metal [1, 2]. Disregarding the atomic structure of matter and spatial dispersion, the local field enhancement tends to infinity, and the localization region tends to zero as the infinitely sharp tip of the conical rod is approached. Realistically, the field localization region and local field enhancement are limited by the smallest practically achievable radius of the tip of the metal cone, which is typically limited by a few nanometers.

Similar nanofocusing effects can be achieved in other tapered metal structures including tapered metal gaps, metal wedges, tapered metal strips on a dielectric substrate with gradually reducing width and/or thickness of the strip, etc. All of these structures are typically characterized by a significant increase of plasmon localization and enhancement of its local field near the tip of the taper. An interesting case of nanofocusing of channel plasmon-polariton modes was considered and observed experimentally in V-shaped metal grooves with gradually reducing taper angle and depth. In this case, the channel plasmon-polariton mode guided by the groove experienced gradual increase in its localization, combined with the local field enhancement, caused by the gradually changing groove parameters [1].

Nanofocusing is typically efficient if the taper angle of the metal structure is sufficiently small, such that:

$$\left|\frac{\partial Q_1^{-1}}{\partial z}\right| << 1, \tag{5a}$$

where Q_1 is the real part of the plasmon wave number in the structure, and the z-axis is parallel to the direction of the taper. If this condition is satisfied, the variations of the plasmon wave number within one plasmon wavelength are negligible, the plasmon does not "feel" the taper, and thus does not experience noticeable reflections from the taper. This is the so-called geometrical optics approximation or, equivalently, adiabatic regime of nanofocusing. Usually, the applicability condition (Eq. 5a) for the adiabatic approximation can be replaced by a significantly less restrictive inequality

$$\left|\frac{\partial Q_1^{-1}}{\partial z}\right| < 0.8, \tag{5b}$$

which was obtained from the comparison of the rigorous numerical results with the outcomes based on the adiabatic approximation.

Condition (Eq. 5b) determines a critical taper angle for the structure βc above which the adiabatic approximation cannot be used for the analysis of plasmon nanofocusing. For a tapered nanogap in a metal medium, or a metal wedge, the values of the critical taper angle are given by Eq. 4. This highlights an interesting physical aspect that adiabatic nanofocusing in V-grooves and wedges is antagonistic to the existence of the strongly localized guided modes in these structures. This is because, in the approximation of continuous electrodynamics, the conditions for adiabatic nanofocusing ($\beta < \beta c$) and for existence of localized modes ($\beta > \beta c$) in infinitely sharp metal V-grooves and wedges are exactly opposite.

Optimal parameters for different nanofocusing structures were determined from the adiabatic approximation and rigorous numerical analysis. For example, if the taper angle is much smaller than βc, then adiabatic condition (Eq. 5a) is satisfied, but the plasmon must propagate large distances along the taper until it reaches the tip. This means that it will experience large dissipative losses, and the local field enhancement at the tip will not be the strongest. On the other hand, if the taper angle is large, so that condition (Eq. 5b) is not satisfied, then the plasmon will experience significant reflections from the taper, and only part of its energy will be able to reach the tip. This will also lead to a reduction of the achievable local field enhancement at the tip. Therefore, there is an optimal taper angle $\beta = \beta_{opt}$ at which the local field enhancement at the tip is the strongest. For example, for a gold cone in vacuum at $\lambda_{vac} = 632.8$ nm (He-Ne laser) and $\beta_{opt} \approx 35°$

[1]. Optimal taper angles for gaps (~20°) tend to be smaller than for cones, but larger than for wedges (~7°). The optimal taper angle for a tapered gap or a metal wedge is close to the critical taper angle below which adiabatic nanofocusing takes place, i.e., $\beta_{opt} \approx \beta c$. This is expected, as both the angles are physically related to the same phenomenon – the appearance of significant reflections of the plasmon from the taper.

Another important parameter is the optimal length of the taper L_{opt}. If the plasmon is generated too far from the tip, it has to propagate a large distance until it reaches the tip and starts experiencing significant focusing effects. This leads to excessive dissipative losses. If the plasmon is generated too close to the tip, it does not have "an opportunity" to take all the advantage of the nanofocusing effect to enhance its field. Therefore, there is an optimal distance from the tip of a nanofocusing structure (or the optimal length of the nanofocusing structure L_{opt}), at which a localized plasmon mode should be generated to ensure the maximal possible local field enhancement at the tip. For example, the optimal length of a nanofocusing conical metal rod can be determined by a simple equation [11]:

$$L_{opt} = \frac{e_1(\varepsilon_d + e_1)}{e_2 \varepsilon_d Q_0} \cos(\beta = 2), \qquad (6)$$

where e_1 and e_2 are the real and imaginary parts of the metal permittivity $\varepsilon_m = e_1 + ie_2$, ε_d is the permittivity of the dielectric medium surrounding the metal cone, and Q_0 is the real part of the plasmon wave number on a flat and smooth interface between the considered metal and dielectric medium. Equation 6 is applicable if the taper angle β is not too small:

$$\mathrm{Sin}\ (\beta = 2)\ >> \frac{e_2 \varepsilon_d}{e_1(\varepsilon_d + e_1)}. \qquad (7)$$

For example, Eq. 6 gives $L_{opt} \approx 10$ μm for a gold cone in vacuum at $\lambda_{vac} = 632.8$ nm, and this estimate is correct within a broad range of taper angles $20° < \beta < 160°$.

Typical achievable local field enhancements at the tip may vary for different nanofocusing structures. For example, conical rods represent one of the best options for plasmon nanofocusing, resulting in the largest local field enhancements up to ~1,600 times for a gold rod in vacuum [1] with the localization of the enhanced field within a region of ~5 nm. Tapered gaps are less efficient from this point of view and typically yield the local field enhancements of ~50 to 100 times. Metal wedges are the least efficient nanofocusing structures compared to the gaps and conical rods with the typical enhancement factors of ~10. Further enhancement of the local field at the tip can be achieved through optimizing the shape of a nanofocusing rod. For example, it has been demonstrated that if an annually converging plasmon is initially generated on a flat surface and then coupled into a conical rod (made in the form of a wizard hat), the local field enhancement at the tip could be as strong as ~10,000 [12].

Importantly, such local field enhancements should lead to a Raman signal enhancements from Raman-active molecules near the tip of the structure up to $\sim 10^{16}$, which is two orders of magnitude larger than what is required for the single molecule detection. Therefore, the analyzed nanofocusing structures may present a unique opportunity for the development of reliable structures for sensing, detection, and imaging of single molecules and traces of substances in the environment surrounding the nanofocusing structure [12].

Strong concentration of the local field near the tip of a nanofocusing structure provides not only the substantial field enhancements, but also enormous field gradients within the distances of about a few nanometers around the tip. This opens a unique opportunity for near-field trapping and manipulation of small nanoparticles and separate molecules with the efficiency up to ~4 orders of magnitude higher than those achievable by other means [12].

Examples of Application

This naturally leads to the discussion of numerous and important applications of plasmon nanfocusing. As indicated above, one of the

major areas of these applications is the development of a new generation of nano-optical sensors and near-field optical imaging techniques with the spatial resolution as good as a few nanometers – far beyond the diffraction limit of light. These will likely to include a new generation of nano-optical sensors for the detection and identification of traces of explosives, drugs, and chemicals in the environment for security and environmental monitoring. Another major area of future applications of nanofocusing will involve cellular and intercellular manipulations, imaging and testing in nanobiophotonics and medicine. These might include new techniques for DNA manipulations and diagnostics, mapping of the local refractive index in a living cell, testing of macromolecules, intercellular fluids and membranes, highly targeted delivery of electromagnetic energy to nanoscale regions of a living cell with equally localized and controlled heat discharge, etc.

A new generation of nonlinear devices is expected to emerge from plasmon nanofocusing structures, such as nonlinear all-optical switches, nanoscale bolometers, active and passive optical devices, new techniques for the investigation of the local material response to highly localized heat discharge on the femtosecond scale, etc. Another important practical application of nanofocusing structures will be related to their natural capability of efficiently coupling relatively large-scale light signals from the conventional optical communication systems to nanoscale electronic chips for further information processing. This means overcoming the major problem with the efficient delivery of light energy in the visible and infrared ranges to nanoscale devices, structures, and separate molecules. A separate area of applications and further research is related to the analysis of single plasmon propagation and focusing, quantum and nanomechanical systems.

A range of other nano-optical devices and techniques based on strongly localized plasmonic modes in metallic nanostructures has been considered and analyzed, including nanoscale plasmonic resonators, filters, couplers, optical adaptors, elements for integrated optical circuitry, nanoantennas, nanolasers, photodetectors and photovoltaic devices, heat-assisted magnetic recording, etc. [1, 2].

Cross-References

- ▶ Active Plasmonic Devices
- ▶ Nanostructures for Photonics
- ▶ Scanning Near-Field Optical Microscopy
- ▶ Surface Plasmon-Polariton-Based Detectors
- ▶ Theory of Optical Metamaterials

References

1. Gramotnev, D.K., Bozhevolnyi, S.I.: Plasmonics beyond the diffraction limit. Nat. Photonics **4**, 83–91 (2010)
2. Schuller, J.A., Barnard, E.S., Cai, W., Jun, Y.C., White, J.S., Brongersma, M.L.: Plasmonics for extreme light concentration and manipulation. Nat. Mater. **9**, 193–204 (2010)
3. Kittel, C.: Introduction to Solid State Physics, 4th edn. Wiley, New York (1971)
4. Raether, H.: Surface Plasmons on Smooth and Rough Surfaces and on Gratings. Springer, New York (1988)
5. Landau, L.D., Lifshitz, E.M.: Electrodynamics of Continuous Media. Pergamon, Oxford (1984)
6. Takahara, J., Yamagishi, S., Taki, H., Morimoto, A., Kobayashi, T.: Guiding of a one-dimensional optical beam with nanometer diameter. Opt. Lett. **22**, 475–477 (1997)
7. Nerkararyan, K.V.: Superfocusing of a surface polariton in a wedge-like structure. Phys. Lett. A **237**, 103–105 (1997)
8. Liebsch, A.: Screening properties of a metal surface at low frequencies and finite wave vectors. Phys. Rev. Lett. **54**, 67–70 (1985)
9. Gather, M.C., Meerholz, K., Danz, N., Leosson, K.: Net optical gain in a plasmonic waveguide embedded in a fluorescent polymer. Nat. Photonics **4**, 457–461 (2010)
10. Noginov, M.A., Zhu, G., Belgrave, A.M., Bakker, R., Shalaev, V.M., Narimanov, E.E., Stout, S., Herz, E., Suteewong, T., Wiesner, U.: Demonstration of a spaser-based nanolaser. Nature **460**, 1110–1112 (2009)
11. Gramotnev, D.K., Vogel, M.W., Stockman, M.I.: Optimized nonadiabatic nanofocusing of plasmons by tapered metal rods. J. Appl. Phys. **104**, 034311 (2008)
12. Gramotnev, D.K., Vogel, M.W.: Ultimate capabilities of sharp metal tips for plasmon nanofocusing, single molecule detection, near-field trapping and sensing. Phys. Rev. Lett. (to be published)

Light-Element Nanotubes and Related Structures

Wenlong Wang[1], Xuedong Bai[1] and Enge Wang[2]
[1]Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Zhongguancun, Beijing, China
[2]International Center for Quantum Materials, School of Physics, Peking University, Beijing, China

Synonyms

Boron- and/or nitrogen-doped carbon nanotubes

Definition

Light-element nanotubes, as specifically defined here, refer to boron (B) and/or nitrogen (N) substitutionally doped carbon (C) nanotubes (C-NTs), including both the selectively doped CB_x and CN_x nanotubes and the co-doped ternary system $B_xC_yN_z$ nanotubes. For either the CB_x or CN_x nanotubes, it is expected that high dopant concentrations will lead to new phases with well-defined stoichiometry; the two most eminent examples are BC_3 and C_3N_4 nanotubes. These stoichiometric nanotubes cannot be simply treated as highly B- or N-doped carbon nanotubes as their structures and electronic properties are inherently different than the pristine C-NTs.

Introduction

The discovery of carbon nanotubes (C-NTs) in the early 1990s has evoked much interest in nanotubular materials as an important frontier area of research that is still flourishing [1–4]. Shortly after the discovery of C-NTs, it was experimentally confirmed that substitutional doping of C-NTs with B and/or N heteroatoms was feasible [5, 6]. As the nearest neighbors of C in the Periodic Table and given their size proximity to C, B, and N, probably the only elements that can be substitutionally doped in a hexagonal C network without severely distorting the lattice. Furthermore, considering the isoelectronic and isostructural relationship between a B–N pair and a C–C pair, it is expected that the simultaneous B and N doping (i.e., co-doping) of C-NTs through partially substituting for C–C units in the tube-shell lattice will lead to the formation of ternary system $B_xC_yN_z$ nanotubes. To some extent, the $B_xC_yN_z$ nanotubes can be treated as random or ordered hybrids or solid solutions of C-NTs and boron nitride (BN) nanotubes [1–11]. On the other hand, selective doping of C-NTs with B or N alone will give rise to binary CB_x and CN_x nanotubes, respectively [12–15]. $B_xC_yN_z$, CB_x, CN_x nanotubes, along with pristine C and BN nanotubes, constitute the family of light-element nanotube materials.

As there have been many reviews available for C and BN nanotubes, the focus of this entry will be constrained to the doped system $B_xC_yN_z$, CB_x, and CN_x nanotubes. Theoretical and experimental aspects concerning their crystal structures, electronic structures, electrical transport properties, state-of-the-art synthesis, characterization of bonding environments, and the atomic distribution of dopants will be surveyed.

Structure and Electronic Structure

The Ternary System $B_xC_yN_z$ Nanotubes

The existence of stable layered bulk structures is usually suggestive of the existence of corresponding stable nanotubular structures. Topologically, a single-walled carbon nanotube (C-SWNT) can be constructed by rolling a single layer of graphite (i.e., graphene) into a seamless tube. Likewise, one can view a BN-SWNT as a rolled up sheet of sp^2-bonded boron nitride (h-BN). Known as the III–VI homology of C nanotubes, BN nanotubes are structurally identical to their C counterparts except that the C atoms are fully substituted by alternating B and N atoms [1, 3]. However, despite the structural similarity between C and BN nanotubes, their electronic band structures are vastly

different. It is well-known that pristine C-NTs can be either semiconducting or metallic, with electronic band structures being sensitively determined by specific tube geometries, that is, diameters and chirality, whose control remains a formidable (if not impossible) challenge for all known synthetic methodologies. By contrast, as a consequence of the ionic character of B–N bond, BN nanotubes are electrical insulators with a large electronic band gap of around 5.5 eV that is almost independent of the diameter and the chirality of the BN tube [1, 3]. It is therefore inferred that the intermediate ternary system $B_xC_yN_z$ nanotubes will exhibit tunable semiconducting properties in between metallic or small band gap C-NTs and insulating BN-NTs [7–11]. Importantly, the electronic band structures of $B_xC_yN_z$ nanotubes are governed largely by the B/C/N chemical stoichiometry rather than by the specific geometry [7–11].

As the most prominent example of $B_xC_yN_z$, BC_2N nanotubes have been theoretically studied by Miyamoto et al. through first principles and tight-binding (TB) band structure calculations [7, 8]. They showed that BC_2N nanotubes would be semiconductors and could be either p- or n-doped depending on the detailed B and N concentrations [8]. In a later study, the evolution of electronic band structures of $B_xC_yN_z$ nanotubes as a function of the B and N concentrations were systematically evaluated within a TB framework by Yoshioka et al. [9]. This relatively simple model gave essential qualitative features of the electronic structures of $B_xC_yN_z$ nanotubes with varied stoichiometry. The calculated results clearly showed the opening of a band gap for metallic (9, 0) nanotubes and the band gap increasing for a semiconducting (10, 0) nanotubes with increasing (x + y) concentration [9].

In a recent study by the present authors, Xu et al. have performed a series of ab initio band structure calculations of $B_xC_yN_z$ nanotubes in a framework of generalized gradient approximation (GGA) [11]. As a first step, various co-doped structures were searched, and as expected, it was found that B and N atoms have a strong preference to neighbor each other and form B–N pairs. The total energy of a unit cell containing an adjacent B–N pair is about 2.0–2.6 eV lower than those with separately doped B and N atoms at isolated sites. By starting from a (5,5) metallic C-SWNT, the incorporation of one B–N pair in a unit cell (with B or N component ratio of 5 %) resulted in the opening of a small band gap of about 0.1 eV at the Fermi level. With further incorporation of one more B–N pair to the unit cell (the two adjacent B–N pairs are separated by six carbon atoms), the energy gap is widened to about 0.5 eV. These results clearly revealed a transition of electronic structure of the (5, 5) nanotube from metallic to semiconducting after B–N co-doping. Other metallic SWNTs at different B–N co-doping concentrations were also calculated, and the results showed a similar band structure crossover from metallic to semiconducting, independent of the chirality of the nanotubes despite the different gap values. The gap opening is found to be fairly sensitive to B/N components and widens monotonically with increasing B/N co-doping concentration. The detailed geometry of the B/N atomic pairs may change the width of the gap slightly, but the overall trend remains the same [11].

Selectively Doped CB_x and CN_x Nanotubes

When considering selective doping of C-NTs solely with electron-deficient B or electron-rich N, one is reminded of the substitutional doping of Silicon (Si) upon which microelectronics relies. The selective doping of Si with substitutional acceptor or donor type atoms results in a sensitive shift of the Fermi level, a concept known as the rigid-band model (i.e., shifting the Fermi level without altering the band structure). However, it is important to note that the doping levels in conventional Si technology are very low, typically on the order of parts per million (ppm). For B- or N-doped nanotubes, the rigid-band model is valid only for a very low dopant concentration [12–15]. In general, as the typical dopant concentrations are on the order of a percent and tenths of a percent, the substitutional B or N doping of nanotubes cannot be solely explained in the framework of a rigid-band model, and the modifications of band structures and density of states (DOEs) must be taken into account [12–15].

Structurally, the B dopant is expected to be incorporated into the honeycomb tube-shell lattice

primarily by direct substitution of C atoms with a triple-coordinated bond [12, 13]. In the case of low B concentration, the formation of an acceptor state in the band gap just above the valence edge is expected. For instance, an early theoretical work by Yi and Bernholc in 1993 has revealed an acceptor state located at 0.16 eV above the Fermi energy for a (10, 0) nanotube with 1.25 wt% B concentration [12]. In a more recent study by Wirtz and Rubio, the authors theoretically studied the evolution of an acceptor-like level that is caused by a periodic distribution of the B dopant as a function of the B-doping ratio. They found that doping profoundly affected the physical properties [13]. Specifically, for pristine C-SWNTs, the DOS is symmetric around the band gap for the first and second van Hove Singularities (vHSs), whereas the incorporation of more B dopants led to an increasing asymmetry near the band gap due to the mixing of the *p* and *s* orbitals [13].

As far as N-doped CN_x nanotubes are concerned, the situation is a little more complicated than an intuitive expectation that the incorporation of N heteroatoms will lead to n-type doping of nanotubes with localized states above the Fermi level [6]. This is the case only if a foreign N atom directly substitutes for a C atom on the honeycomb tube-shell lattice, in a bonding environment just like that of the aforementioned B dopants [6, 14, 15]. For example, a substitutional N dopant may also generate structural defects in the tube-shell lattice and lead to a rearrangement of the neighboring C atoms. In this case, it is not immediately obvious that n-type behavior will result, because the electrical behavior depends on the details of the new geometry. As a result, the as-formed new tube-shell structure can be p type [6, 14, 15].

For either B- or N-doped nanotubes, it is expected that higher dopant concentrations may lead to new phases with defined stoichiometry, for example, BC_3 and C_3N_4 [16–18]. Just like graphene sheets, the stability of the planar structures of these stoichiometric compounds could make it possible for nanotubes with the same stoichiometry to exist. Band structure calculations have revealed that individual BC_3 nanotubes will act as semiconductors. Bundles of tubes are expected to exhibit metallic behavior due to inter-tube cohesive interactions [16]. As far as both crystallographic and electronic structures are concerned, the stoichiometric BC_3 compound nanotubes cannot simply be regarded as highly B-doped nanotubes. Similarly C_3N_4 nanotubes, which are also theoretically predicted to be well-defined semiconductors, are inherently different from substitutionally doped CN_x nanotubes [17]. To date, studies on BC_3 and C_3N_4 nanotubes are limited to theoretical modeling. The experimental synthesis of these stoichiometric compound nanotubes is still elusive.

Synthesis of $B_xC_yN_z$, BC_x, and CN_x Nanotubes

Generally speaking, there are two different strategies for the production of B- and/or N-doped nanotubes, namely, direct synthesis and post-synthetic nanotube-substitution reactions. The direct synthesis methods were in fact inherited from those that have been well established for pure C-NTs synthesis, mainly including arc-discharge, laser-ablation, and chemical vapor deposition (CVD) [19–33]. The post-synthetic nanotube-substitution reaction approach, on the other hand, refers to the use of pristine C-NTs as starting precursors that undergo carbothermal reactions with B_2O_3 vapor at elevated temperature (1,200–1,700 °C) in N_2 or NH_3 atmosphere; this reaction process will induce partial substitution of the carbon lattice by foreign B and/or N atoms, thereby resulting in the formation of doped nanotubes [34].

Synthesis of B- and N-doped C-NTs in the form of multi-walled nanotubes (MWNTs) was reported by Stéphan et al. early in 1994 using the arc-discharge method [5]. Since then, a great deal of research has been directed toward B- and/or N-doped nanotubes. However, despite the considerable progress in synthesis of doped MWNTs [1–4], substitutional doping of SWNTs turn out to be much more difficult, and it was not until very recently that reports of successful synthesis of high-quality B- and/or N-doped

SWNTs have emerged [4, 10, 11, 19–33]. Typically, B- and/N-doped MWNTs suffer from large percentages of severe topological modifications and therefore usually exhibit highly defective, bamboo-like structural features. Doped SWNTs, however, normally exhibit straight, non-buckled tubular structures, close to pristine C-SWNTs [4, 10, 11, 19–33]. For SWNTs, it is likely that the B- and/or N doping introduces topological defects in the tube-shell, but not necessarily defective, bamboo-like structures.

Direct Synthesis: Chemical Vapor Deposition

Over the past decade, catalytic decomposition of carbon-containing precursor gases over transition metal catalyst nanoparticles in a CVD growth has proved to be an elegant strategy for the controlled synthesis of MWNTs and SWNTs. In comparison with the physical evaporation approaches such as arc-discharge and laser-ablation, CVD growth holds advantages in terms of experimental facility, potential for large-scale production at low cost, and control over tube location and orientation. CVD synthesis of B- and/or N-doped nanotubes has also been extensively exploited, but the majority of reported studies have been limited to doped MWNTs. In 2006, a breakthrough in CVD growth of ternary $B_xC_yN_z$-SWNTs was made by the present authors through a bias-assisted hot filament CVD (HFCVD) growth process [10]. $B_xC_yN_z$-SWNTs growth by HFCVD was achieved over a high-activity powdery-type catalyst, namely, porous MgO supported Fe–Mo (denoted as Fe–Mo/MgO) with CH_4, B_2H_6, and ethylenediamine vapor as the reactant gases and inert He as the diluting gas. Transmission electron microscopy (TEM) imaging showed that the as-grown SWNTs had a clean and smooth morphology with diameters in the range of 0.8–2.5 nm. In some cases, double-walled nanotubes (DWNTs) with slightly larger diameter coexisted with the SWNTs, but the amount was not significant. The electron energy loss spectroscopy (EELS) spectra taken from the SWNT bundles showed the core-loss K-edges of B, C, and N located at 188, 284, and 401 eV, respectively. The sharply defined 1 s $\rightarrow \pi^*$ and 1 s $\rightarrow \sigma^*$ transition features of the C–K edge is characteristic of well-graphitized sp^2-bonding carbon networks. The B and N K-edges, though being much weaker, also showed discernible π^* peak as well as σ^* band, indicating that the B and N atoms were in the same sp^2-hybridized state as their C counterparts. Importantly, all as-grown $B_xC_yN_z$-SWNTs had straight cylindrical tube walls, a morphological feature similar to that of pristine C-SWNTs. This is in sharp contrast to the previously reported defective, "bamboo-like" tube structures of doped MWNTs. Other than SWNTs and DWNTs, the HFCVD growth also produced a small amount of carbon onions and "bamboo-like" MWNTs. These by-products were hard to remove completely but their amount could be effectively reduced by optimizing the growth parameters and the use of the Fe–Mo/MgO catalyst [10].

In recent years, there have also been a few studies concerning the synthesis of N-doped SWNTs by conventional thermal CVD. Keskar et al. reported on the growth of isolated CN_x-SWNTs on quartz and SiO_2/Si substrates via a liquid injection CVD process by using the mixture of xylene and acetonitrile (CH_3CN, N precursor) as the feeding source [19]. The catalysts they used were nanoparticles deposited onto the SiO_2/Si substrates prior to the growth. The resulting surface-grown SWNTs were characterized by resonant Raman spectroscopy. It was found that the intensity of the radial breathing mode (RBM) of the as-grown nanotubes decreased with increasing N precursor concentration in the feeding source, while the relative intensity of D/G-band increased with the appearance of the D′-band. The authors attributed these observations to the N doping in SWNTs [19]. In another CVD experiment, Villalpando-Paez et al. synthesized long strands of N-doped SWNTs through an aerosol-assisted CVD process using a mixture of ferrocene/ethanol/benzylamine, with benzylamine ($C_6H_5CH_2NH_2$) acting as the N precursor and ferrocene as the Fe catalyst precursor [20]. As the relative ratio of benzylamine to ethanol increased, the SWNT yield decreased gradually with the bundled SWNT strands getting shorter and thinner. At higher benzylamine concentrations, SWNT growth was completely inhibited. Neither EELS nor Energy

Dispersive X-ray (EDX) spectra detected the presence of N in the as-grown SWNTs,, indicating that the N-doping ratio, if any, was rather low and likely below the detection limit of these instruments (e.g., <1 wt%) [20]. In the following study, the same CN_x-SWNT samples were reexamined and the N-dopant concentration was determined to be about 0.3 at.% [21].

As for the CVD growth of B-doped SWNTs, to date there is only one reported by Ayala et al. [22]. The authors utilized triisopropyl borate as precursor and Fe-based supported catalyst in a CVD system working under controlled high-vacuum conditions. It was found that although high-quality C-SWNTs could be formed in a wide temperature range, B doping was not always successful, with less than 1 at.% B concentration confirmed [22].

Direct Synthesis: Arc-Discharge and Laser-Ablation

The carbon arc-discharge method, initially used for producing C_{60} fullerenes, led to the experimental discovery of C-NTs by Iijima in 1991 [23], and later the first synthesis of C-SWNTs [24, 25]. In 1994, Stéphan et al. initiated the direct synthesis of B- and N-doped MWNTs by means of arc-discharge [5]. One-year later, the arc-discharge synthesis of MWNTs with nominal stoichiometry of BC_2N and BC_3 was reported by the Zettl group [6]. Nevertheless, the direct synthesis of doped SWNTs in an arc-discharge experiment is not easily achieved. A study in 2002 by Droppa et al. first examined the possibility of doping SWNTs with N by the arc-discharge evaporation of graphite in a He–N_2 atmosphere in the presence of Fe–Co–Ni catalysts [26]. It was found that a pure He atmosphere could produce pure C-SWNTs bundles whereas a mixed He–N_2 atmosphere instead gave rise to nested nanofibers rather than SWNTs. As such, their attempt to produce doped SWNTs was unsuccessful [26]. The successful arc-discharge synthesis N-doped SWNTs was achieved later on by Glerup et al. [27, 28]. What is unique about their synthesis was the use of N-rich precursors (melamine, $C_3N_3H_6$) as additives in the anode rod rather than relying on a nitrogen gas-containing atmosphere [27]. With an appropriate amount of melamine, the doped SWNTs with a N content around 1 at.% could be successfully produced. The as-grown N-doped SWNTs were well graphitized and exhibited straight non-buckled nanotube walls, similar to pristine C-SWNTs. When BN powder was used as precursor, the incorporation of $\sim$ 1 at.% N into the SWNTs was also achieved but no B doping was detected by EELS. The authors concluded that B doping is more difficult than N doping in SWNTs [27].

Both arc-discharge and laser-ablation belong to the high temperature vaporization methods that have demonstrated remarkable success in the synthesis of standard C-SWNTs. The use of the laser-ablation technique for B-doped SWNT synthesis was first reported by McGuire et al. [29, 30]. In their growth experiments, ablation targets were prepared by mixing different amounts of elemental B with C paste and a Co–Ni catalyst. It was found that the amount of B in the target strongly affected the final products of laser-ablation growth. SWNT bundles could be observed in the final products only when there was no more than 3 at.% B in the target, whereas the presence of excess B precursor would severely inhibit the SWNT growth. The authors examined the as-grown SWNT bundles by EELS, but no B was detected. Nonetheless, some indirect evidence for B doping came instead from Raman spectra, which showed appreciable changes in the intensity and/or the position of the D-band, G'-Band, and RBM [29, 30]. Later on, a breakthrough in the synthesis of B-doped SWNTs was achieved by Blackburn and coworkers using the laser-ablation method [31]. In their work, a high-yield of doped SWNTs with B contents up to 1.8 at.%, as determined by EELS measurements, was successfully produced. The unique feature of their synthesis was the use of nickel boride (NiB) alloy as both the B precursor and the catalyst. NiB was loaded in the graphite targets with appropriate ratios, and laser vaporization of the composite target in a N_2 atmosphere produced B-doped SWNTs bundles that were comparable in quality to the pure C-SWNTs from traditional laser ablations synthesis. Although the B-doped SWNTs were produced in a N_2 atmosphere, EELS as

well as [34] C-NMR measurements excluded the possibility of B and N co-doping and thus the formation of ternary $B_xC_yN_z$-SWNTs [31]. Besides NiB, the authors also tested the use of elemental B as a precursor for B doping, but laser vaporization of the targets produced only low-yield SWNTs which showed little B incorporation. This comparison further confirmed the unique role of NiB alloy in high-yield laser-ablation synthesis of B-doped SWNTs [31].

More recently, Lin et al. reported the successful synthesis of N-doped SWNTs by laser vaporization [32, 33]. The N-doping concentration was determined by means of EELS in a TEM. They also performed the first combined STS-TEM study of N-doped SWNTs; however, identifying the N bonding environment was difficult from a spectral deconvolution of the N core level edge [32, 33].

In another study, the phase-separated ternary $B_xC_yN_z$-SWNTs, that is, SWNTs made of BN nanodomains in a tube-shell hexagon lattice, were synthesized by laser-ablation using a target consisting of carbon and boron mixed with a Co/Ni catalyst under N_2 atmosphere [34]. HRTEM and nanoelectron EELS (nano-EELS) provided direct evidence that B and N co-segregated with respect to C and formed nanodomains within the hexagonal lattice of the graphene layer in a sequential manner [34].

Post-synthetic Nanotube-Substitution Reaction Route

The post-synthetic nanotube-substitution reaction approach was initially reported by Han et al. and latterly by Golberg et al. [35–37] using multi-walled C-NTs as a precursor for substitution. They demonstrated successful synthesis of the multi-walled BN-NTs and $B_xC_yN_z$-NTs in bulk quantities. Unfortunately, this approach was found to be far less effective when applied to the B/N co-doping of SWNTs. The use of a substitution reaction approach for SWNTs was first reported by Golberg et al. in 1999 [38, 39]. In the synthesis, laser-ablation produced C-SWNTs bundles were employed as the starting materials, which were heated together with B_2O_3 in a flowing N_2 atmosphere. By examining the as-treated SWNTs with HRTEM and EELS, the authors investigated the effects of synthesis temperature and reaction time on the B- and N-doping levels. The maximum yield of doped SWNTs was obtained at 1,553 K for 30 min, with an accumulating B-dopant concentration up to ~ 10 at.% and N concentration up to ~ 2 at.%: effective B/N co-doping was thus achieved [38, 39]. It was also found that the corrosive B_2O_3 vapor at such high temperature severely deteriorates the tubular morphology of SWNTs, while the problem is alleviated at lower temperatures. In a later study, Borowiak-Palen et al. revisited the same B/N substitution reaction process for SWNTs. But they were able to produce only binary B doped CB_x-SWNTs, rather than the ternary B/N co-doped $B_xC_yN_z$-SWNTs. More corrosive ammonia (NH_3) was used as the N source in place of N_2 in their study. A maximum B-doping ratio up to ~ 15 at.% could be reached in the as-generated CB_x-SWNTs [40].

The main stumbling point for effective B/N co-doping of SWNTs via a post-synthetic reaction process lies in the bundling phenomena of the C-SWNTs. As a result of the extraordinary cohesive interactions between adjacent nanotubes, SWNTs have a strong tendency to spontaneously form densely bundled ropes in their native state. When such bundled C-SWNTs are subjected to the carbothermal substitution reaction process, only a small fraction of SWNTs localized at the outermost fragments of bundles will be chemically accessible to the gaseous B and N reactants, whereas the inner of SWNTs are largely intact [38–40]. Aiming at solving this problem, the present authors have recently developed an innovative wet-chemistry approach for high-yield synthesis of the ternary $B_xC_yN_z$-SWNTs in bulk quantities [41]. An initial and key step of this approach involved the unbundling and dispersion of the starting C-SWNTs (HiPCO, super-purified) in an ionic liquid (IL) at room temperature, followed by the wrapping of the unbundled SWNTs with boric acid and then carbothermal substitution reaction at an optimized temperature of 1,000 °C with a NH_3 gas flow. The as-synthesized $B_xC_yN_z$-SWNTs are of high purity and quality rivaling that of the super-purified

HiPCO C-SWNTs. TEM and EELS investigations revealed that the ternary $B_xC_yN_z$-SWNTs typically have straight and smooth tube-shell structures, and a fairly homogenous B and N concentration. As a result of B/N co-doping, the thin film electrical devices built from $B_xC_yN_z$-SWNTs exhibit good electrical performance characterized by semiconducting behaviors [41].

Depth of Characterizations: Bonding Environments and Dopant Distribution

sp^2 bonding plays a fundamental role in all light-element nanotubes made of B, C, and N. Besides the information on chemical compositions and doping concentrations, a clear understanding of the bonding configuration and atomic distribution of B and/or N dopants is an important issue. Studies in this regard have been most advanced in the N-doped C-NTs [1, 4, 14, 15]. There have been a number of theoretical works modeling the bonding configuration of N within the CN_x nanotubes. Three main structural models have been proposed: pyridine-like, pyrrole-like, and substitutional (also named as "graphitic"), among which the pyridinic and substitutional N are both sp^2 coordinated whereas the pyrrolic N is sp^3 coordinated [1, 4, 14, 15]. Experimentally, the evidence of pyridinic and substitutional N is often from the results of XPS and/or EELS measurement on CN_x-MWNTs. In a recent study on the N-doped SWNTs, Glerup et al. observed in the EELS spectra that the N K-edge splits into two weak π^* peaks located at 398.6 and 402 eV, which could be attributed to π^* states associated with the pyridinic and substitutional N, respectively [27]. It is interesting to notice that CN_x-MWNTs often exhibit "bamboo-like" morphology rather than the regular tubular morphology, suggesting the N dopants may induce local structural distortion and a high concentration of pyridinic N will lead to the formation of the compartmentalized bamboo-like structure [1, 4, 14, 15]. On the other hand, the doped SWNTs normally exhibit a non-buckled straight tubular morphology. In this vein, the N-doped SWNTs provide us with a model system to study and establish the correlation between the bond configuration of dopants and their atomic distribution. Unfortunately, relevant studies are still lacking up to date. Nevertheless, it is plausible to argue that the doped SWNTs should contain more substitutional N atoms than pyridinic N atoms, and the overall N distribution should be more homogenous in doped SWNTs than in the bamboo-like MWNTs.

As for the B-doped C-NTs, there are very limited studies available concerning the bond configuration and the atomic distribution of B dopants. In an early study, Carroll et al. proposed the presence of BC_3 island nanodomains within the B-doped MWNTs, which could explain their experimental observation of the metallic behavior of the doped nanotubes with B concentration in the range of 1–5 at.% [42]. Recently, in the B-doped SWNTs produced from substitution reaction, Fuentes et al. have also assumed the formation of BC_3 nanodomains instead of a homogeneous random B substituted C-NT [43].

In the case of the ternary $B_xC_yN_z$ nanotubes, the situation is more complicated when considering the B/N co-doping in the nanotube lattice. Two types of atomic distributions have been theoretically predicted: one proposes that the B, C, and N species are randomly and homogeneously distributed in the tube lattice with a ternary B–C–N bonding nature, whereas the other one proposes a phase separation of the C-rich and BN-rich domains [3, 4, 7–11, 34, 44–47]. In the latter case, the C/BN phase separation would result in the formation of heterojunctions, superlattices, or quantum dots, or even the formation of the "sandwich-like" structures consisting of the separated pure C and BN shells [44–47]. Given the relatively higher stability of C–C and B–N bonds over that of C–B and C–N bonds, thermodynamically there should be a tendency for such a phase separation [44–47]. However, one must note that kinetics rather than thermodynamics can govern the configuration of nanotubes, so the formation of homogenous ternary $B_xC_yN_z$ nanotubes is also possible. In earlier studies via the post-synthetic substitution route, Golberg and Terrones et al. have successfully

obtained homogeneous $B_xC_yN_z$-MWNTs [37]. In the $B_xC_yN_z$-SWNTs grown by HFCVD method, the present authors studied the atomic distribution of B, C, and N in the nanotubes by means of spatially resolved EELS and energy-filtered electron microscopy [10]. It turned out that the constituting B, C, and N elements were homogenously distributed in the SWNTs, without appreciable phase separation. In other words, the HFCVD grown $B_xC_yN_z$-SWNTs are homogenously structured with a ternary-bonding nature. As we know, the bias-assisted HFCVD growth process occurs in a highly non-equilibrium condition, as such, a kinetic-controlled nanotube growth mechanism may come into play, which leads to the formation of homogenous ternary B–C–N tube shells rather than segregated C and BN domains. Complementary evidence for the ternary-bonding nature of $B_xC_yN_z$-SWNTs also comes from the XPS characterization. In the XPS spectra of B 1 s and N 1 s core level electrons, the existence of a significant amount of B–C and C–N bonds that are prevailing over the B–N bonds are confirmed [10]. It is worth mentioning that the spatial resolution of the energy-filtered electron microscope images (by Omega filter TEM instrument) was estimated to be ~ 0.5 nm, and detecting sub-nm domains is beyond the capabilities of most state-of-the-art instruments, hence the existence of sub-nanometer fragments made of pure BN and pure C phases could not be completely ruled out.

Electrical Transport Properties

Investigating doping effects on the electrical transport properties of doped C-NTs is of fundamental importance to the use of nanotubes in electronic devices. To date, most studies in this area are theoretical, as experimental design and transport measurements are restricted by the lack of high-quality doped nanotube samples, especially SWNT samples. In a recent study, electrical transport measurements on individual CN_x-SWNTs exhibited *n*-type conduction when there was a high concentration of graphite-like nitrogen bonding [28]. On the other hand, reports on the electrical transport properties of individual CB_x-SWNTs are scarce. We note that thermopower measurements on the CB_x-SWNTs bundle exhibited sustained *p*-type characteristics [29]. Also worth mentioning is that more recently, superconductivity was observed in CB_x-SWNTs thin films [48].

As for ternary $B_xC_yN_z$ nanotubes, theoretical studies predicted that that they may display semiconducting behavior intermediated between C-NTs and dielectric BN nanotubes [7–11]. Owing to the advances in synthesizing high-quality $B_xC_yN_z$-SWNTs by HFCVD, the present authors for the first time fabricated field effect transistor (FET) devices made of $B_xC_yN_z$-SWNTs [11]. Gate-dependent transport measurements revealed that the $B_xC_yN_z$-SWNTs were *p*-type semiconductors, and importantly, by surveying a large number of devices, the abundance of semiconducting nanotubes was estimated to be more than 97 %, in contrast to the nearly two thirds abundance of semiconducting nanotubes in pristine C-SWNTs.

The influence of substitutional B and/or N dopants on the electrical transport properties of nanotubes, the localized transport phenomena, and the problem of resonant scattering are general concerns [49–51]. For isolated substitutional B or N dopants, it is commonly accepted that they will not only transfer holes or electrons to the graphitic C lattice but also give rise to quasibound impurity states [49–51]. In this vein, the electronic quantum conductance will be reduced and strong localization may set in: the mean-free path of electrons would be drastically damped around the corresponding quasibound state energies whereas pristine nanotubes display ballistic behavior [49–51]. However, in the B–N pairs co-doping tubes, Khalfoun et al. recently reported an interesting finding that that a pair of adjacent B and N atoms was practically transparent for conducting electrons close to the Femi level [52]. In other words, the substitution of a B–N pair in the C lattice does not significantly modify the quantum transport of the nanotubes at low gate voltage [52].

Concluding Remarks

Light-element nantoubes and related structures constitute a fertile research area that awaits further experimental and theoretical efforts. From a fundamental viewpoint, doped nanotubes provide model systems for understanding dopant-induced perturbations on the physical properties of one-dimensional electron systems. Meanwhile, doping C nanotubes with B and N can lead to electronic properties that are more dependent on chemistry than the specific geometry of tubes. Keeping in mind the possibility of creating an entire range of ternary nanotubes intermediated between C and BN underlies a fascinating prospect for systematical band gap engineering in the ternary $B_xC_yN_z$ nanotube systems. In particular, the semiconducting $B_xC_yN_z$-SWNTs are promising candidates for envisioned applications in nanotube-based electronics and optoelectronics, sensors, and other systems. On the other hand, there is still a long way to go toward this fascinating goal due to the challenges in producing ternary SWNTs with controlled and variable B–C–N stoichiometry.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Chemical Vapor Deposition (CVD)
- ▶ Synthesis of Carbon Nanotubes

References

1. Terrones, M., Grobert, N., Terrones, N.: Synthetic routes to nanoscale $B_xC_yN_z$ architectures. Carbon **40**, 1665–1684 (2002)
2. Ewels, C.P., Glerup, M.: Nitrogen doping in carbon nanotubes. J. Nanosci. Nanotechnol. **5**, 1345–1363 (2005)
3. Golberg, D., Bando, Y., Tang, C., Zhi, C.: Boron nitride nanotubes. Adv. Mater. **19**, 2413–2432 (2007)
4. Ayala, P., Arenal, R., Loiseau, A., Rubio, A., Pichler, T.: The physical and chemical properties of heteronanotubes. Rev. Mod. Phys. **82**, 1843–1845 (2010)
5. Stéphan, O., Ajayan, P.M., Colliex, C., Redlich, P., Lambert, J.M., Bernier, P., Lefin, P.: Doping graphitic and carbon nanotube structures with boron and nitrogen. Science **266**, 1683–1685 (1994)
6. Wengsieh, Z., Cherrey, K., Chopra, N.G., Blase, X., Miyamoto, Y., Rubio, A., Cohen, M.L., Louie, S.G., Zettl, A., Gronsky, A.R.: Synthesis of $B_xC_yN_z$ nanotubules. Phys. Rev. B **51**, 11229–11232 (1995)
7. Liu, A.Y., Wetzcovitch, R.M., Cohen, M.L.: Atomic arrangement and electronic structure of BC_2N. Phys. Rev. B **39**, 1760–1765 (1989)
8. Miyamoto, Y., Rubio, A., Cohen, M.L., Louie, S.G.: Chiral tubules of hexagonal BC_2N. Phys. Rev. B **50**, 4976–4979 (1994)
9. Yoshioka, T., Suzuura, H., Ando, T.: Electronic states of BCN alloy nanotubes in a simple tight-binding model. J. Phys. Soc. Jpn. **72**, 2656–2664 (2003)
10. Wang, W.L., Bai, X.D., Liu, K.H., Xu, Z., Golberg, D., Bando, Y., Wang, E.G.: Direct synthesis of B-C-N single-walled nanotubes by bias-assisted hot filament chemical vapor deposition. J. Am. Chem. Soc. **128**, 6530–6531 (2006)
11. Xu, Z., Lu, W.G., Wang, W.L., Gu, C.Z., Liu, K.H., Bai, X.D., Wang, E.G., Dai, H.J.: Converting Metallic Single-Walled Carbon Nanotubes into Semiconductors by Boron/Nitrogen Co-Doping. Adv. Mater. **20**, 3615–3619 (2008)
12. Yi, J.Y., Bernholc, J.: Atomic structure and doping of microtubules. Phys. Rev. B **47**, 1708–1711 (1993)
13. Wirtz, L., Rubio, A.: Band structure of boron doped carbon nanotubes. AIP Conf. Proc. **685**, 402–405 (2003)
14. Czerw, R., Terrones, M., Charlier, J.C., Blasé, X., Foley, B., Kamalakaran, R., Grobert, N., Terrones, H., Tekleab, D., Ajayan, P.M., Blau, W., Rühle, M., Carroll, D.L.: Identification of electron donor states in N-doped carbon nanotubes. Nano Lett. **1**, 457–460 (2001)
15. Kang, H.S., Jeong, S.: Nitrogen doping and chirality of carbon nanotubes. Phys. Rev. B **70**, 233411 (2004)
16. Miyamoto, Y., Rubio, A., Cohen, M.L., Louie, S.G.: Electronic properties of tubule forms of hexagonal BC_3. Phys. Rev. B **50**, 18360–18366 (1994)
17. Miyamoto, Y., Cohen, M.L., Louie, S.G.: Theoretical investigation of graphitic carbon nitride and possible tubule forms. Solid State Commun. **8**, 605–608 (1997)
18. Gracia, J., Kroll, P.: First principles study of C_3N_4 carbon nitride nanotubes. J. Mater. Chem. **19**, 3020–3026 (2009)
19. Keskar, G., Rao, R., Luo, J., Hudson, J., Chen, J., Rao, A.M.: Growth, nitrogen doping and characterization of isolated single-wall carbon nanotubes using liquid precursors. Chem. Phys. Lett. **412**, 269–273 (2005)
20. Villalpando-Paez, F., Zamudio, A., Elias, A.L., Son, H., Barros, E.B., Chou, S.G., Kim, Y.A., Muramatsu, H., Hayashi, T., Kong, J., Terrones, H., Dresselhous, G., Endo, M., Terrones, M., Dresselhaus, M.S.: Synthesis and characterization of long strands of nitrogen-doped single-walled carbon nanotubes. Chem. Phys. Lett. **424**, 345–352 (2006)

21. Elias, A.L., Ayala, P., Zamudio, A., Grobosch, M., Cruz-Silva, E., Romo-Herrera, J.M., Campos, J., Terrones, H., Pichler, T., Terrones, M.: Spectroscopic characterization of N-doped single-walled carbon nanotube strands: an x-ray photoelectron spectroscopy and Raman study. J. Nanosci. Nanotechnol. **10**, 3959–3964 (2010)
22. Ayala, P., Plank, W., Grüneis, A., Kauppinen, E., Rümmeli, M., Kuzmany, H., Pichler, T.: A one step approach to B-doped single-walled carbon nanotubes. J. Mater. Chem. **18**, 5676 (2008)
23. Iijima, S.: Helical microtubules of graphitic carbon. Nature (London) **354**, 56–58 (1991)
24. Iijima, S., Ichihashi, T.: Single-shell carbon nanotubes of 1-nm diameter. Nature **363**, 603–605 (1993)
25. Bethune, D.S., Kiang, C.H., Devries, M.S.: Cobalt-catalysed growth of carbon nanotubes with single-atomic-layer walls. Nature **363**, 605–607 (1993)
26. Droppa Jr., R., Hammer, P., Carvalho, A.C.M., dos Santos, M.C., Alvarez, F.: Incorporation of nitrogen in carbon nanotubes. J. Non-Cryst. Solids **874**, 299–302 (2002)
27. Glerup, M., Steinmetz, J., Samaille, D., Stéphan, O., Enouz, S., Loiseau, A., Roth, S., Bernier, P.: Synthesis of N-doped SWNT using the arc-discharge procedure. Chem. Phys. Lett. **387**, 193–197 (2004)
28. Krstic, V., Rikken, G.L.J.A., Bernier, P., Roth, S., Glerup, M.: Nitrogen doping of metallic single-walled carbon nanotubes: n-type conduction and dipole scattering. Europhys. Lett. **77**, 37001 (2007)
29. McGuire, K., Gothard, N., Gai, P.L., Dresselhaus, M. S., Sumanasekera, G., Rao, A.M.: Synthesis and Raman characterization of boron-doped single-walled carbon nanotubes. Carbon **43**, 219–227 (2005)
30. Gai, P.L., Stephan, O., McGuire, K., Rao, A.M., Dresselhaus, M.S., Dresselhaus, G., Colliex, C.: Structural systematics in boron-doped single wall carbon nanotubes. J. Mater. Chem. **14**, 669–675 (2004)
31. Blackburn, J.L., Yan, Y., Engtrakul, C., Parilla, P.A., Jones, K., Gennett, T., Dillon, A.C., Heben, M.J.: Synthesis and characterization of boron-doped single-wall carbon nanotubes produced by the laser vaporization technique. Chem. Mater. **18**, 2558–2566 (2006)
32. Lin, H., Arenal, R., Enouz-Vedrenne, S., Stephan, O., Loiseau, A.: Nitrogen configuration in individual CN_x-SWNTs synthesized by laser vaporization technique. J. Phys. Chem. C **113**, 9509–9511 (2009)
33. Lin, H., Lagoute, J., Chacon, C., Arenal, R., Stephan, O., Repain, V., Girard, Y., Enouz, S., Bresson, L., Rousset, S., Loiseau, A.: Combined STM/ST-S. TEM/EELS investigation of CNx-SWNTs. Phys. Status Solidi B **245**, 1986–1989 (2008)
34. Enouz, S., Stéphan, O., Cochon, J.L., Colliex, C., Loiseau, A.: C – BN patterned single-walled nanotubes synthesized by laser vaporization. Nano Lett. **7**, 1856–1862 (2007)
35. Han, W.Q., Bando, Y., Kurashima, K., Sato, T.: Boron-doped carbon nanotubes prepared through a substitution reaction. Chem. Phys. Lett. **299**, 368–373 (1999)
36. Golberg, D., Bando, Y., Kurashima, K., Sato, T.: MoO_3-promoted synthesis of multi-walled BN nanotubes from C nanotube templates. Chem. Phys. Lett. **323**, 185–191 (2000)
37. Terrones, M., Golberg, D., Grobert, N., Seeger, T., Reye-Reyes, M., Mayne, M., Kamalakaram, R., Dorozhkin, P., Dong, Z.C., Terrones, H., Rühle, M., Bando, Y.: Production and state-of-the-art characterization of aligned nanotubes with homogeneous BC_xN ($1 <= x <= 5$) compositions. Adv. Mater. **15**, 1899 (2003)
38. Golberg, D., Bando, Y., Han, W., Kurashima, K., Sato, T.: Single-walled B-doped carbon. B/N-doped carbon and BN nanotubes synthesized from single-walled carbon nanotubes through a substitution reaction. Chem. Phys. Lett. **308**, 337–342 (1999)
39. Golberg, D., Bando, Y., Bourgeois, L., Sato, T.: Large-scale synthesis and HRTEM analysis of single-walled B- and N-doped carbon nanotube bundles. Carbon **38**, 2017–2027 (2000)
40. Borowiak-Palen, E., Pichler, T., Graff, A., Kalenczuk, R.J., Knupfer, M., Fink, J.: Synthesis and electronic properties of B-doped single wall carbon nanotubes. Carbon **42**, 1123–1126 (2004)
41. Xiaoxia Yang, X., Liu, L., Wu, M., Wang, W.L., Bai, X.D., Wang, E.G.: Wet-chemistry assisted nanotube-substitution reaction for high-efficiency and bulk-quantity synthesis of boron and nitrogen co-doped single-walled carbon nanotubes (to be published).
42. Carroll, D.L., Redlich, P., Blase, X., Charlier, J.-C., Curran, S., Ajayan, P.M., Roth, S., Ruhle, M.: Effects of nanodomain formation on the electronic structure of doped carbon nanotubes. Phys. Rev. Lett. **81**, 2332–2335 (1998)
43. Fuentes, G.G., Borowiak-Palen, E., Knupfer, M., Pichler, T., Fink, J., Wirtz, L., Rubio, A.: Formation and electronic properties of BC_3 single-wall nanotubes upon boron substitution of carbon nanotubes. Phys. Rev. B **69**, 245403 (2004)
44. Blase, X., Charlier, J.C., De Vita, A., Car, R.: Theory of composite $B_xC_yN_z$ nanotube heterojunctions. Appl. Phys. Lett. **70**, 197–199 (1997)
45. Blase, X.: Properties of composite $B_xC_yN_z$ nanotubes and related heterojunctions. Comput. Mater. Sci. **17**, 107–114 (2000)
46. Kohler-Redlich, P., Terrones, M., Manteca-Diego, C., Hsu, W.K., Terrones, H., Rühle, M., Kroto, H.W., Walton, D.R.: Stable BC_2N nanostructures: low-temperature production of segregated C/BN layered materials. Chem. Phys. Lett. **310**, 459–465 (1999)
47. Suenaga, K., Collies, C., Demoncy, N., Loiseau, A., Pascard, H., Willaime, F.: Synthesis of nanoparticles and nanotubes with well-separated layers of boron nitride and carbon. Science **278**, 653–655 (1997)
48. Murata, N., Haruyama, J., Reppert, J., Rao, A.M., Koretsune, T., Saito, S., Matsudaira, M., Yagi, Y.: Superconductivity in thin films of boron-doped carbon nanotubes. Phys. Rev. Lett. **101**, 027002 (2008)

49. Choi, H.J., Ihm, J., Louie, S.G., Cohen, M.L.: Defects, quasibound states, and quantum conductance in metallic carbon nanotubes. Phys. Rev. Lett. **84**, 2917–2920 (2000)
50. Latil, S., Roche, S., Mayou, D., Charlier, J.C.: Mesoscopic transport in chemically doped carbon nanotubes. Phys. Rev. Lett. **92**, 2568051 (2004)
51. Eom, J.H., Lee, H., Im, J., Park, C., Jeong, B.W., Kim, S., Ihm, J.: Electronic structure of defects and quantum transport in carbon nanotubes. Physica B **376–377**, 7–10 (2006)
52. Khalfoun, H., Hermet, P., Henrard, L.: B and N codoping effect on electronic transport in carbon nanotubes. Phys. Rev. Lett. **81**, 193411 (2010)

Lipid Vesicles

► Liposomes

Liposomes

Seyed Moein Moghimi
Department of Pharmaceutics and Analytical Chemistry, Centre for Pharmaceutical Nanotechnology and Nanotoxicology, University of Copenhagen, Copenhagen, Denmark

Synonyms

Large unilamellar vesicle (LUV); Lipid vesicles; Multilamellar vesicle (MLV); Small unilamellar vesicle (SUV)

Definition

Liposomes are vesicular structures of colloidal dimension, consisting of amphiphilic lipid forming bilayers enclosing part of the aqueous phase in which they are dispersed. The lipids used are predominantly glycerophospholipids and/or sphingophospholipids; liposomes therefore exhibit many of the properties of natural membrane-bounded structures. Liposomes are characterized by their membrane composition, morphology, and particle size distribution. Liposomes have diverse applications in science, engineering, and medicine.

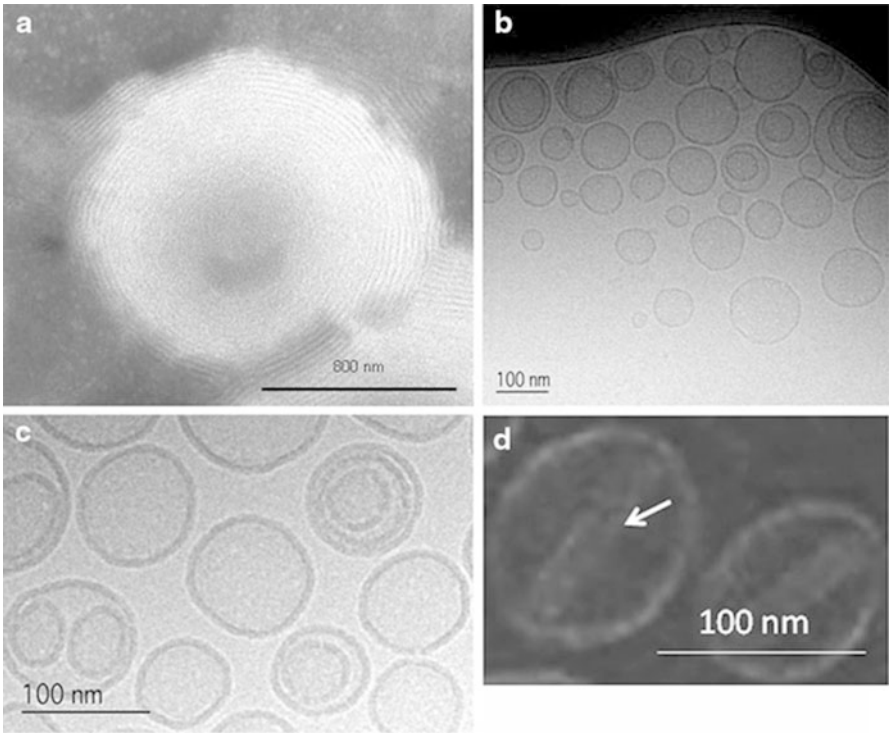
Morphology

Morphologically, one may distinguish between small (S), large (L), giant (G), unilamellar (U), oligolamellar (O), and multilamellar (ML) vesicles (V) [1, 2]. Small unilamellar vesicles (SUVs) are composed of a single lipid bilayer and are typically below 100 nm in size, whereas large and giant terminology are indicative of liposomes larger than 100 nm (but below 1 μm), and 1 μm or larger, respectively. Oligolamellar morphology refers to vesicles composed of between two and five concentric lamellae, whereas liposomes containing more than five lamellae are typically called multilamellar vesicles (MLVs). There are other combinations such as multivesicular liposomes in which smaller liposomes are entrapped in larger vesicles. Figure 1 represents different liposomal morphologies.

Many physicochemical characteristics of liposomes are controlled by their morphology and size [1, 3, 4]. These include encapsulated volume, thermotropic behavior, leakage, aggregation, deformability, and stability. For example, the fraction of encapsulated volume in LUVs is much higher than in SUVs at the same lipid concentration; the entrapped volume increases with $(r-\delta)^3$ where r is the vesicular radius and δ is the thickness of the membrane. SUVs are also less stable due to stored tension compared with LUVs. The membrane of SUVs can freeze only partially where regions of unfrozen lipids at the edges make vesicles more deformable and unstable; at high temperature phase SUVs are rigid spheres. Both LUVs and GUVs, however, become deformable in the liquid crystalline phase.

Historical Background

Liposomes were first observed by Alec Bangham and colleagues in the early 1960s [5] as lipid



Liposomes, Fig. 1 Micrographs representing different liposomal morphologies. A transmission electron micrograph (*TEM*) of MLV preparation (**a**), cryo-TEM images showing a plethora of different morphologies (SUVs, LUVs, OVs, and multivesicular liposomes) (**b** and **c**) (Courtesy of Dr. Thomas L. Andresen, Technical University of Denmark), and a cryo-TEM image of regulatory approved Doxil® (**d**) showing stacks of precipitated doxorubicin in the centre of oval-shaped vesicles (*arrow*)

spherules (the form when viewed by negative staining technique in the electron microscope) capable of trapping whatever ionic species that was dissolved in the aqueous phase at the time of their formation. They further realized that these structures were osmotically active and could exhibit different permeability properties to different molecules and ions. The term "liposome" was first used in 1967.

Historically, however, Virchow in 1854 under polarizing microscope recorded the formation of normal and helical myelin figures following swelling of dry lipids. Lehmann, at the turn of the twentieth century produced a micrograph, which today is referred to as MLVs. In 1934, a British patent also described drug depots composed of lecithin, cholesterol, and water.

Geometrical Packing of Amphiphiles

A useful parameter (but with limitations) in predicting the aggregate shape of amphiphilic molecules is a packing parameter P (2), defined as:

$$P = v/a_l$$

where v is the volume of the molecule, a is the area occupied by the polar head group, and l the length of the hydrocarbon chain(s). For single-chain amphiphiles (e.g., soaps, lysophospholipids) the polar head is relatively large with respect to the nonpolar segment and pack either into spherical $\left(P < \frac{1}{3}\right)$ or rod-like $\left(\frac{1}{3} < P < \frac{1}{2}\right)$ micelles (aggregates of colloidal dimension that exists in equilibrium with the molecules that they are formed). With double-chained lipids

(phosphatidylcholine, phosphatidylserine, dicetylphosphate, sphingomyelin) $P \sim 1$ (cylindrical shape) and pack into lamellar structures such as liposomes. Lipids with relatively small polar heads ($P > 1$, cone shaped) tend to pack into inverse structures; examples include phosphatidylethanolamine, phosphatidic acid, cardiolipin, and cholesterol. The packing model is often useful in cases of different lipid mixtures where an average P can be defined as:

$$P = \sum\nolimits_i n_i P_i$$

where *ni* are mol fractions. For instance equimolar mixtures of phosphatidylethanolamine with cholesterol or lyso lipids can form liposomes.

Methods of Liposome Preparation

There are numerous methods for liposome preparation and these have been reviewed extensively [2, 3]. The method of choice usually depends on the lipid composition, the desired morphology or the sensitivity of the material to be encapsulated. Also, the same type of liposomes prepared by different methodologies may exhibit different properties to include encapsulation efficacy, bilayer permeability, and vesicular stability in vitro and in vivo.

The most common applied method for the laboratory-scale liposome preparation is the thin-film hydration, which yields predominantly very heterogeneous populations of MLVs and some SUVs, GUVs, and GOVs. The procedure begins by dissolving lipids in an organic solvent usually chloroform (chloroform normally contains 1 % ethanol; however, the use of pure chloroform should be avoided as this may cause lipid peroxidation). If charged lipids are present, addition of methanol (20–30 vol%) aids solubilization and most lipid mixtures are soluble in this solvent combination. Methanol, however, forms hydrogen bonds with the phospholipid polar heads and is difficult to remove. Therefore, inclusion of methanol is only recommended if absolutely necessary. For instance, for preparation of giant liposomes, it is necessary that solvent evaporates slowly resulting in the formation of a very smooth lipid film; for such cases larger amounts of methanol are used. Once the lipids are thoroughly mixed in the organic solvent, the solvent is removed by rotary evaporation yielding a thin lipid film on the sides of a round bottom flask. After removal of residual solvent, by placing the flask on a vacuum pump overnight, the film is hydrated by addition of an appropriate volume of an aqueous medium (that may contain a water-soluble drug for encapsulation) and agitation from a few minutes to several hours. The temperature of the aqueous phase should be above the gel-liquid crystal transition temperature (Tc) of the lipid with the highest Tc. Lipid composition, the type of aqueous media, and energy and power of agitation all influence vesicular morphology and size distribution. For instance, charged bilayers yield smaller and less lamellar liposomes.

A relatively homogeneous population of liposomes composed of a single bilayer can be produced by extruding the MLV preparation through a filter with well-defined pore sizes under high pressure (up to 100 atm) in a high-pressure cell or alternatively by sonication (e.g., probe sonication). Repeated extrusions tend to decrease the number of lamellae. In the case of not fully hydrated liposomes large myelin figures of greater than 10 μm may grow from the cylindrical worm-like structures into solution. The nonencapsulated aqueous cargo can then be removed by dialysis or gel-permeation chromatography. Hydrophobic and amphiphilic molecules or drugs can be associated with the bilayer during any of these steps.

Other methodologies for liposome preparation include detergent depletion techniques, solvent (ethanol, ether) injection, and demulsification (reverse-phase evaporation, double emulsions) [2, 3]. In addition to these, there are methods that use already preformed vesicles to prepare vesicles with higher encapsulation efficiencies of water-soluble molecules [2, 3]. These methods are based on freeze-thawing or dehydration-rehydration of SUVs or on fusion of SUVs induced by fusogenic agents. For instance SUVs made from phosphatidylserine (PS) fuse into cylindrical rolls in the presence of Ca^{2+} at

concentrations slightly above the half of PS concentration. Such entities are referred to as cochleate cylinders; chelation of calcium by EDTA or its removal by ion exchange results in the formation of LUVs.

Some advanced nanotech-based methodologies for liposome preparation includes inkjet printing and controlled microfluidic mixing and nanoparticle determination (COMMAND) [6, 7]. The latter utilizes microfluidic hydrodynamic focusing to precisely control the convective-diffusive mixing of miscible liquids under laminar flow, which provides a method for both on-chip liposome preparation and the controlled experimental investigation of self-assembly process of lipid molecules into vesicles. A key advantage of COMMAND use is the elimination of size homogenization post-processing procedures. COMMAND allows controlled formation of vesicles in the size range of 50–150 nm with relative standard deviations ranging from <10 % for smaller vesicle distributions to <20 % for larger vesicle distributions.

A Proposed Mechanism for MLV Formation

In contact with water, the outer monolayer of dry phospholipid film hydrates more rapidly than the inner layers resulting in the formation of convex contusions, since the surface area of polar heads increases with the increasing hydration. Due to water penetration bilayers grow from contusions into tubular fibrils thus exposing polar heads to water maximally. Fibrils may further grow into helices if there are packing defects at one side of the primary contusion as this induces torque. The bilayers are subsequently stabilized into their equilibrium distance. Following agitation, tubular fibrils detach and immediately seal off their exposed edges to form MLVs.

Related Structures

Two commonly related structures to liposomes are niosomes and polymersomes. Niosomes (also known as nonionic liposomes) are composed from diacyl or monoacyl polyglycerol or polyoxyethylene-based amphiphiles in mixtures with cholesterol and prepared by similar methods to liposomes. Polymersomes are self-assembled polymer shells that are composed of block copolymer amphiphiles. By tailoring the block lengths and chemistry, the physicochemical characteristics of polymersomes such as permeability, contents release rates, and stability can be manipulated.

Biological and Medical Applications of Liposomes

Due to their versatile nature, physicochemical properties and morphological characteristics liposomes have long been served as useful model systems in many fundamental biochemistry, biophysics, cell biology, and immunological studies [2]. These include biochemical investigations of conformational and function of membrane proteins (e.g., ion pumps like ATPases), membrane fusion processes and biological recognition events, vesicular budding and pinching-off mechanisms that operate in Golgi apparatus, exocytosis and formation of synaptic vesicles, and artificial photosynthetic models.

The benefits and applications of liposomes in medicine are numerous [8–11]. The use of liposomes as drug delivery systems substantially influences pharmacokinetics and biodistribution of the drug. In several cases this can be used to improve the therapeutic efficacy and minimize drug-related adverse effects. For such applications, liposomes can be formulated as suspension, lyphophilized powder for reconstitution, aerosol, cream or lotion, and therefore practically amenable for all routes of administration into the body. However, the vast majority of liposomal medical applications uses parenteral injection [8, 10, 11], where the main applications include:

1. Improved solubility of lipophilic and amphiphilic drugs (or even emulsions).
2. Encapsulation of certain hydrophilic drugs at concentrations several fold above their

aqueous solubility as a result of drug precipitation or gel formation inside the liposome with appropriate substances encapsulated (Fig. 1d).

3. Passive targeting to macrophages in contact with the blood, notably the Kupffer cells (also known as hepatic sinusoidal macrophages) and macrophages located at the marginal zone and the red-pulp regions of the spleen [10]. This offers an unprecedented approach for precision targeting, delivery and release of metabolic enzymes, antimicrobials, toxins, immunomodulators, and immunosuppressive agents for treatment and disorders of the macrophage system as well as for modulating their accessory functions (e.g., tumoricidal function). Indeed, the idea and the application of liposomes for enzyme replacement therapy was put forward and tested soon after their discovery [12, 13]. First patients injected with liposomes in 1974.
4. Passive accumulation at pathological sites where the integrity of endothelial barrier is perturbed by inflammatory process (e.g., infection, infarction) or by tumor growth. For such applications, the surface of liposomes are often engineered with hydrophilic polymers such as poly(ethylene glycol) as this will dramatically reduce macrophage interception, which results in prolonged liposomal circulation times within the vasculature and better opportunities for liposome extravasation from the blood into pathological regions [10, 11].
5. Site-specific targeting to accessible cells and structures within the vasculature (or beyond) following grafting of biological recognition molecules such as monoclonal antibodies, saccharides, peptides, etc., to specialized lipids incorporated into the liposomal bilayer [10, 11, 14]. Here, target sites have included vascular endothelial cells, blood cells such as erythrocytes, lymphocytes, and platelets, as well as circulating tumor cells.
6. Controlled intracellular release of therapeutic agents such as nucleic acids as well as antigens with triggered-release vesicles (e.g., pH-sensitive, fusogenic liposomes, light-sensitive liposomes). For example, pH-sensitive liposome maintains stable structure at neutral pH but destabilizes at low endosomal (intracellular vesicles formed from the cell membrane, which are involved in intracellular transport) pH. This subsequently disrupts endosomal integrity causing cargo release into cytoplasm [15]. Indeed, antigen delivery to the cytosol of dendritic cells, and hence channeling into the major histocompatibility complex class I route is dramatically enhanced by triggered release in late endosomes with the aid of pH-sensitive and fusogenic liposomes.
7. Controlled extracellular drug release at pathological sites such as infections and solid tumors with specialized formulations (temperature-sensitive liposomes, phospholipase A2-sensitive liposomes).
8. Slow-release system within the blood (intravenous administration) or at selected sites following local administration.
9. As adjuvant (an immune potentiator that triggers early immune responses that subsequently assist the generation of potent and persistent adaptive immune responses) for vaccines following intramuscular and subcutaneous routes of administration. Liposomes further offer antigen protection from degradation and elimination following local injection [8].
10. Site-avoidance systems where drug encapsulation offers protection to the drug vulnerable sites (e.g., doxorubicin-mediated cardiotoxicity, amphotericin B-mediated nephrotoxicity); following intravenous injection liposomes do not accumulate appreciably in the heart, kidneys, and the brain [8].
11. Medical imaging and diagnostics following encapsulation of grafting of contrast agents (as in magnetic resonance imaging).
12. Lubrication systems for joints.
13. Artificial oxygen delivery systems.

Table 1 lists some examples of regulatory approved liposomal formulations for infusion into human subjects for treatment of selected cancers and infections. In addition to the above-

Liposomes, Table 1 Examples of regulatory approved (Food and Drug Administration, USA) liposomal products in the market

Product	Liposome type	Lipids[a]	Active ingredient	Indication(s)
Doxil® (approved 1995)	Stealth SUV	mPEG-DSPE, HSPC, cholesterol	Doxorubicin hydrochloride	Treatment of patients with ovarian cancer whose disease has progressed or recurred after platinum-based chemotherapy
				AIDS-related Kaposi's sarcoma after failure of prior systemic chemotherapy or intolerance to such therapy
				Multiple myeloma in combination with bortezomib in patients who have not previously received bortezomib and have received at least one prior therapy
DaunoXome® (approved 1996)	SUV	DSPC and cholesterol	Daunorubicin citrate	Advanced AIDS-related Kaposi's sarcoma
Ambisome® (approved 1997)	SUV	HSPC, cholesterol, DSPG, α-tocopherol	Amphotericin B	Empirical therapy for presumed fungal infection in febrile, neutropenic patients
				Treatment of *Cryptococcal* Meningitis in HIV-infected patients
				Treatment of patients with *Aspergillus* species, *Candida* species and/or *Cryptococcal* species refractory to amphotericin B deoxycholate, or in patients where renal impairment or unacceptable toxicity precludes the use of amphotericin B deoxycholate
				Treatment of visceral leishmaniasis
Depocyte® (approved 1999)	Multivesicular	Cholesterol, triolein, DOPC, DPPG	Cytarabine	Intrathecal treatment of lymphomatous meningitis

[a]*mPEG-DSPE N*-(carbonyl-methoxypolyethylene glycol 2000)-1,2-distearoyl-*sn*-glycero3-phosphatidylethanolamine sodium salt, *HSPC* hydrogenated soy phosphatidylcholine, *DSPC* distearoylphosphatidylcholine, *DSPG* distearoylphosphatidylglycerol, *DOPC* dioleoylphosphatidylcholine, *DPPG* dipalmitpoylphosphatidylglycerol

mentioned applications, liposomes are also used as in vitro diagnostics (e.g., rapid screening test for coagulation system with liposomes containing recombinant human tissue factor).

Other Applications

Liposomes have received considerable attention in cosmetics (for skin moisturizing and delivery of humectants, nutrients, enzymes, etc., in to the skin) as well as in agro-food industry (e.g., as sustained release systems for enzymes in various fermentation processes and in the preservation of cheeses, sustained release systems for biocides), environmental cleaning and protection (e.g., delivery of agents to oil spills to speed up the degradation), and paint industry.

References

1. Lasic, D.D.: The mechanism of vesicle formation. Biochem. J. **256**, 1–11 (1998)
2. Lasic, D.D.: Liposomes: From Physics to Applications. Elsevier, Amsterdam (1993)
3. Gregoriadis, G.: Liposome technology. In: Liposome Preparation and Related Techniques, vol. 1, 3rd edn. Informa Healthcare, New York (2006)
4. Gregoriadis, G.: Liposome technology. In: Entrapment of Drugs and other Materials into Liposomes, vol. 3, 2nd edn. Informa Healthcare, New York (2006)
5. Bangham, A.D., Standish, M.M., Watkins, J.C.: Diffusion of univalent ions across the lamellae of swollen phospholipids. J. Mol. Biol. **13**, 238–252 (1965)

6. Hauschild, S., Lipprandt, U., Rumplecker, A., Borchert, U., Rank, A., Schubert, R., Forster, S.: Direct preparation and loading of lipid and polymer vesicles using inkjets. Small **1**, 1177–1180 (2005)
7. Jahn, A., Stavis, S.M., Hong, J.S., Vreeland, W.N., DeVoe, D.L., Gaitan, M.: Microfluidic mixing and the formation of nanoscale lipid vesicles. ACS Nano **4**, 2077–2087 (2010)
8. Lasic, D.D., Papahadjopoulos, D.: Medical Applications of Liposomes. Elsevier, Amsterdam (1998)
9. Gregoriadis, G.: Liposome technology. In: Interaction of Liposome with the Biological Milieu, vol. 3, 3rd edn. Informa Healthcare, New York (2006)
10. Moghimi, S.M., Hunter, A.C., Murray, J.C.: Nanomedicine: current status and future prospects. FASEB J. **19**, 311–330 (1995)
11. Torchilin, V.P.: Recent advances with liposomes as pharmaceutical carriers. Nat. Rev. Drug Discov. **4**, 145–160 (2005)
12. Sessa, G., Weismann, G.: Formation of artificial lysosome in vitro. J. Clin. Invest. **48**, 76a (1969)
13. Gregoriadis, G., Ryman, B.E.: Fate of protein-containing liposomes injected to rats. An approach to the treatment of storage diseases. Eur. J. Biochem. **24**, 485–491 (1972)
14. Jølck, R.I., Feldborg, L.N., Andersen, S., Moghimi, S. M., Andresen, T.L.: Engineering liposomes and nanoparticles for biological targeting. Adv. Biochem. Eng. Biotechnol. (2010). doi:10.1007/10_2010_92
15. Drummond, D.C., Zignani, M., Leroux, J.C.: Current status of pH-sensitive liposomes in drug delivery. Prog. Lipid Res. **39**, 409–460 (2000)

Lipospheres

► Solid Lipid Nanoparticles (SLN)

Lippmann-Schwinger Integral Equation Method

► Green's Function Integral Equation Method

Liquid Atomization Through Vapor Explosion

► Spray Technologies Inspired by Bombardier Beetle

Liquid Condensation

► Disjoining Pressure and Capillary Adhesion

Liquid Film

► Disjoining Pressure and Capillary Adhesion

Liquid Immersion Lithography

► Immersion Lithography

Liquid Meniscus

► Disjoining Pressure and Capillary Adhesion

Lithographie, Galvanoformung, Abformung (LIGA)

► SU-8 Photoresist

LM10

► Nanoparticle Tracking Analysis

Local Density of States

► Surface Electronic Structure

Local Surface Plasmon Resonance (LSPR)

► Optical Properties of Metal Nanoparticles

Localized Plasmon Polaritons in the Infrared

► Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy

Log-Domain

► Bioinspired CMOS Cochlea

Logic-Embedded Vectors

► Multistage Vectors

Lotus Effect

Bharat Bhushan[1] and Michael Nosonovsky[2]
[1]Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA
[2]Department of Mechanical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Synonyms

Self-cleaning; Superhydrophobicity

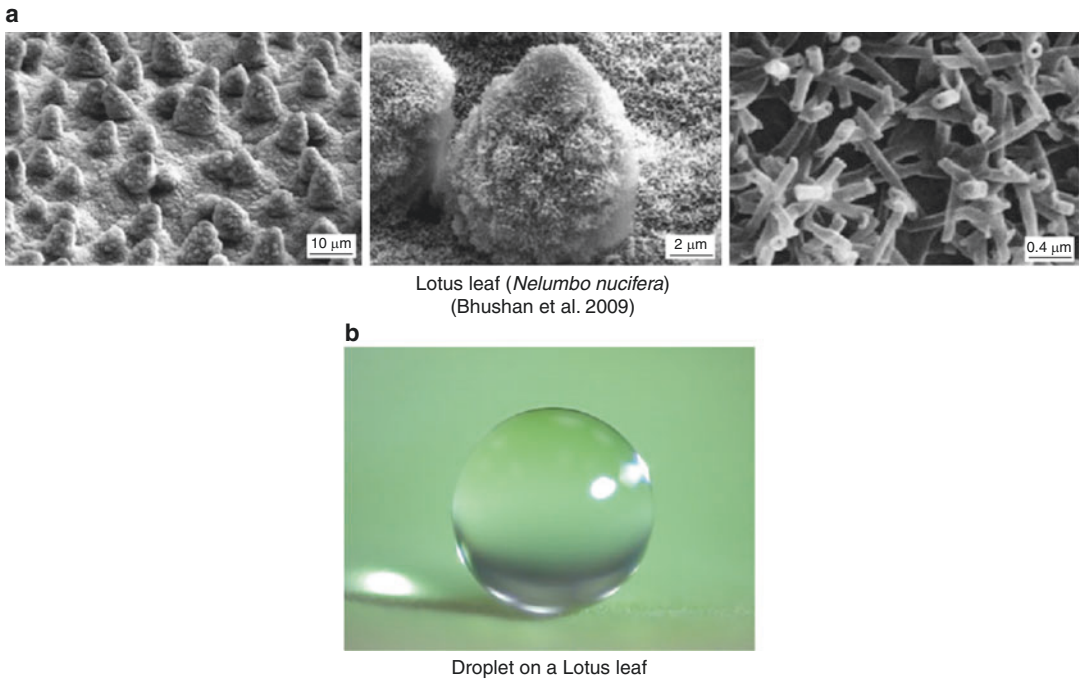
Definition

Lotus effect is surface roughness–induced superhydrophobicity and self-cleaning. These surfaces have a water contact angle (CA) larger than 150° and a low (usually <5°) contact angle hysteresis. These surfaces also exhibit low drag during fluid flow.

Introduction

Superhydrophobic surfaces with a high static contact angle above 150° and contact angle hysteresis of less than 10° exhibit extreme water repellence and self-cleaning properties [1–4]. At a low value of contact angle hysteresis, the droplets may roll in addition to slide which facilitates removal of contaminant particles. Surfaces with low contact angle hysteresis have a low water roll-off (tilt) angle, which denotes the angle to which a surface must be tilted for roll off of water drops. These surfaces exhibit low drag during fluid flow. Superhydrophobic and self-cleaning surfaces are of interest for various applications including self-cleaning windows, windshields, exterior paints for buildings and navigation of ships, utensils, roof tiles, textiles, solar panels, and applications requiring antifouling and a reduction of drag in fluid flow, e.g., in micro/nanochannels. These surfaces can also be used for energy conversion and energy conservation [1]. Condensation of water vapor from the environment and/or process liquid film can form menisci, leading to high adhesion in devices requiring relative motion. Superhydrophobic surfaces are needed to minimize adhesion between a surface and liquid.

A model surface for superhydrophobicity and self-cleaning is provided by the leaves of the Lotus plant (*Nelumbo nucifera*) (Fig. 1a) [3]. The leaf surface is very rough due to so-called papillose epidermal cells, which form asperities or papillae. In addition to the microscale roughness the surface of the papillae is also rough with submicron sized asperities composed of 3-D epicuticular waxes. The waxes of Lotus are tubules, but on other leaves waxes exist also in the form of platelets or other morphologies. Lotus leaves have hierarchical structures. The water droplets on these surfaces readily sit on the apex of nanostructures because air bubbles fill in the valleys of the structure under the droplet. Therefore, these leaves exhibit considerable superhydrophobicity (Fig. 1b). The water droplets on the leaves remove any contaminant particles from their surfaces when they roll off, leading to self-cleaning. It has been reported that nearly all superhydrophobic and self-cleaning leaves



Lotus Effect, Fig. 1 (**a**) SEM micrographs (shown at three magnifications) of Lotus (*Nelumbo nucifera*) leaf surface which consists of microstructure formed by papillose epidermal cells covered with 3-D epicuticular wax tubules on surface, which create nanostructure, and (**b**) image of water droplet sitting on the Lotus leaf [3]

consist of an intrinsic hierarchical structure. Water on such a surface forms a spherical droplet, and both the contact area and the adhesion to the surface are dramatically reduced.

Based on the Lotus effect, one of the ways to increase the hydrophobic property of the surface is to increase surface roughness, so roughness-induced hydrophobicity has become a subject of extensive investigations.

Modeling of Contact Angle for a Liquid in Contact with a Rough Surface

As stated earlier, superhydrophobic and self-cleaning surfaces should have both high contact angle and low contact angle hysteresis. Liquid may form either a homogeneous interface with a solid, or a composite interface with air pockets trapped between the solid and liquid. In this section, mathematical models which provide relationships between roughness and contact angle are discussed.

Homogeneous (Wenzel) Interface

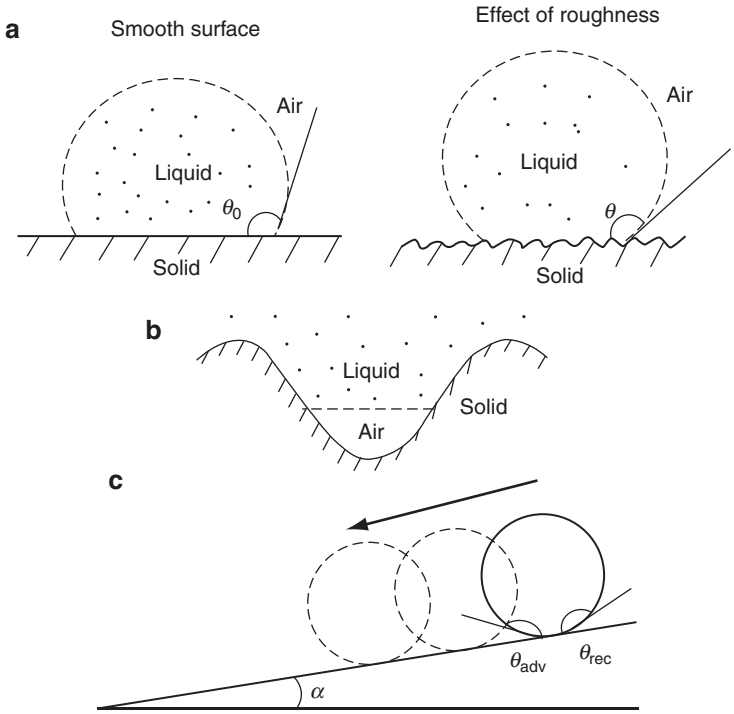
Consider a rough solid surface with a typical size of roughness details smaller than the size of the droplet as shown in Fig. 2a. For a droplet in contact with a rough surface without air pockets, referred to as a homogeneous interface, the contact angle is given as [5]

$$\cos\theta = R_f \cos\theta_0 \tag{1}$$

where θ is the contact angle for a rough surface, θ_0 is the contact angle for a smooth surface (Fig. 2a), and R_f is a roughness factor defined as the ratio of the solid–liquid area A_{SL} to its projection on a flat plane, A_F. The model predicts that a hydrophobic surface ($\theta_0 > 90°$) becomes more hydrophobic with an increase in R_f, and a hydrophilic surface ($\theta_0 < 90°$) becomes more hydrophilic with an increase in R_f.

Composite (Cassie-Baxter) Interface

For a rough surface, a wetting liquid will be completely absorbed by the rough surface cavities



Lotus Effect, Fig. 2 (**a**) Schematic of a liquid droplet in contact with (*top left*) a smooth solid surface (contact angle, θ_0) and (*top right*) a rough solid surface (contact angle, θ), (**b**) schematic of formation of a composite solid–liquid-air interface for a rough surface, and (**c**) tilted surface profile (tilt angle, α) with a liquid droplet; advancing and receding contact angles are θ_{adv} and θ_{rec}, respectively

while a non-wetting liquid may not penetrate into surface cavities, resulting in the formation of air pockets, leading to a composite solid–liquid-air interface as shown in Fig. 2b. Cassie and Baxter [6] extended the Wenzel equation for the composite interface, which was originally developed for the homogeneous solid–liquid interface. For this case, there are two sets of interfaces: a solid–liquid interface with the ambient environment surrounding the droplet and a composite interface involving liquid-air and solid-air interfaces. In order to calculate the contact angle for the composite interface, Wenzel's equation can be modified by combining the contribution of the fractional area of wet surfaces and the fractional area with air pockets ($\theta = 180^\circ$)

$$\cos\theta = R_f \cos\theta_0 - f_{LA}\left(R_f \cos\theta_0 + 1\right) \qquad (2)$$

where f_{LA} is fractional flat geometrical areas of the liquid-air interfaces under the droplet. According to Eq. 2, even for a hydrophilic surface, the contact angle increases with an increase of f_{LA}. At a high value of f_{LA}, a surface can become hydrophobic; however, the value required may be unachievable or formation of air pockets may become unstable. Using the Cassie-Baxter equation, the value of f_{LA} at which a hydrophilic surface could turn into a hydrophobic one is given as [1, 2]

$$f_{LA} \geq \frac{R_f \cos\theta_0}{R_f \cos\theta_0 + 1} \quad \text{for } \theta_0 < 90^\circ \qquad (3)$$

Contact Angle Hysteresis

The contact angle hysteresis is another important characteristic of a solid–liquid interface. If a drop sits over a tilted surface (Fig. 2c), the contact angle at the front and back of the droplet corresponds to advancing contact angle and receding contact angle, respectively. The advancing angle is greater than the receding angle, which results in contact angle hysteresis. The contact angle hysteresis occurs due to surface roughness and heterogeneity. Although for surfaces with the roughness carefully controlled on the molecular scale it is possible to achieve a contact angle hysteresis as low as $<1^\circ$, hysteresis cannot be eliminated completely, since even atomically smooth surfaces have a certain roughness and

heterogeneity. The contact angle hysteresis is a measure of energy dissipation during the flow of a droplet along a solid surface. Surfaces with low contact angle hysteresis have a very low water roll-off angle, which denotes the angle to which a surface must be tilted for roll off of water drops. Low water roll-off angle is important in liquid flow applications such as in micro/nanochannels and surfaces with self-cleaning ability.

Bhushan et al. [2] derived a relationship for contact angle hysteresis as a function of roughness, given as

$$\theta_{\text{adv}} - \theta_{\text{rec}} = (1 - f_{LA}) R_f \frac{\cos\theta_{a0} - \cos\theta_{r0}}{-\sin\theta} = (\sqrt{1 - f_{LA}}) R_f \frac{\cos\theta_{r0} - \cos\theta_{a0}}{\sqrt{2(R_f \cos\theta_0 + 1}} \quad (4)$$

For the homogeneous interface, $f_{LA} = 0$, whereas for the composite interface f_{LA} is a nonzero number. It is observed from Eq. 4 that for a homogeneous interface, increasing roughness (high R_f) leads to increasing (high values of $\theta_{\text{adv}} - \theta_{\text{rec}}$), while for a composite interface, an approach to unity of f_{LA} provides both high a contact angle and a small contact angle hysteresis. Therefore, the composite interface is desirable for superhydrophobicity and self-cleaning.

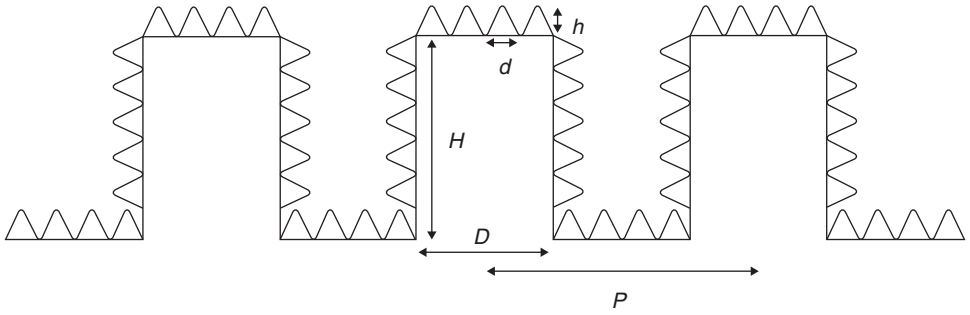
Stability of a Composite Interface and Role of Hierarchical Structure

Formation of a composite interface is a multiscale phenomenon which depends upon the relative sizes of the liquid droplet and roughness details. The composite interface is metastable, and its stability is an important issue. Even though it may be geometrically possible for the system to become composite, it may be energetically profitable for the liquid to penetrate into the valleys between asperities and form the homogeneous interface. The composite interface is fragile and can be irreversibly transformed into the homogeneous interface, thus damaging superhydrophobicity. In order to form a stable composite interface with air pockets between solid and liquid, the destabilizing factors such as capillary waves, nanodroplet condensation, surface inhomogeneity, and liquid pressure should be avoided. For high f_{LA}, a nanopattern is desirable because whether the liquid–air interface is generated depends upon the ratio of distance between two adjacent asperities and droplet radius. Furthermore, nanoscale asperities can pin liquid droplets and thus prevent liquid from filling the valleys between asperities. High R_f can be achieved by both micropatterns and nanopatterns.

Nosonovsky and Bhushan [1] have studied destabilizing factors for the composite interface and found that a convex surface leads to a stable interface and high contact angle. Also, they have suggested the effects of a droplet's weight and curvature among the factors which affect the transition. They have demonstrated that a combination of microroughness and nanoroughness (multiscale roughness) with convex surfaces can help to resist destabilization by pinning the interface. It also helps in preventing the gaps between the asperities from filling with liquid, even in the case of a hydrophilic material. The effect of roughness on wetting is scale dependent, and mechanisms that lead to destabilization of a composite interface are also scale dependent. To effectively resist these scale-dependent mechanisms, it is expected that a multiscale roughness is optimum for superhydrophobicity.

Ideal Surfaces with Hierarchical Structure

For stability of a composite interface, Bhushan et al. [2] suggested the structure of an ideal hierarchical surface as shown in Fig. 3. The asperities should be high enough that the droplet does not touch the valleys. As an example, for a structure with circular pillars, the following relationship should hold for a composite interface. For a droplet with a radius on the order of 1 mm or larger, a value of H on the order of 30 μm, and D on the order of 15 μm, a P on the order of 130 μm is optimum. Nanoasperities can pin the liquid–air interface and thus prevent liquid from filling the valleys between asperities. They are also required to support nanodroplets, which may condense in the valleys between large asperities. Therefore, nanoasperities should have a small pitch to handle nanodroplets, less than 1 mm down to few nm radius. Values of h on the order of 10 nm and d on the order of 100 nm can be easily fabricated.



Lotus Effect, Fig. 3 Schematic of structure of an ideal hierarchical surface. Microasperities consist of circular pillars with diameter *D*, height *H* and pitch *P*. Nanoasperities consist of pyramidal nanoasperities of height *h* and diameter *d* with rounded tops [3]

Artificial Lotus-Effect Surfaces

Two main requirements for a superhydrophobic surface are that the surface should be rough and that it should have a hydrophobic (low surface energy) coating. These two requirements lead to two methods of producing a superhydrophobic surface: first, it is possible to make a rough surface from an initially hydrophobic material and, second, to modify a rough hydrophilic surface by modifying surface chemistry or applying a hydrophobic material upon it. Note that roughness is usually a more important property than a low surface energy, since both moderately hydrophobic and very hydrophobic materials can exhibit similar wetting behavior when roughened [1].

In general, the same techniques that are used for micro- and nanostructure fabrication, such as lithography, etching, and deposition, have been utilized for producing superhydrophobic surfaces (Table 1). Pros and cons of these techniques are summarized in Table 2. An especially interesting development is the creation of switchable surfaces that can be turned from hydrophobic to hydrophilic by applying electric potential, heat, or ultraviolet (UV) irradiation. Another important requirement for potential applications to optics and self-cleaning glasses is the creation of transparent superhydrophobic surfaces. In order for the surface to be transparent, roughness details should be smaller than the wavelength of visible light (about 400–700 nm) [1–3].

Fabrication using Nature's Route – To verify the understanding of the mechanism responsible for the Lotus effect, first structures using nature's routes were fabricated to verify that contact angle data of the fabricated structures were comparable to that of the Lotus leaves found in nature. Bhushan et al. [3] fabricated surfaces with a hierarchical structure with micropatterned epoxy replicas, and Lotus leaf microstructure, and created a second level of structuring with wax platelets and wax tubules. Platelets and tubules are the most common wax morphologies found in plant surfaces, and exist on superhydrophobic leaves, such as Lotus and *Colocasia esculenta* leaves. The structures developed mimic the hierarchical structures of superhydrophobic leaves.

A two-step molding process was used to fabricate several structurally identical copies of the micropatterned Si surface and Lotus leaves. The technique used is a fast, precise, and low-cost molding process for biological and artificial surfaces. Nanostructure was created by self-assembly of synthetic and plant waxes deposited by thermal evaporation. Hierarchical structures fabricated with microstructured Lotus leaf replicas and micropatterned Si replicas covered with a nanostructure of Lotus wax tubules are shown as Fig. 4. The grown tubules provide the desired nanostructure on microstructured surfaces. The recrystallized Lotus wax shows tubular hollow structures, with random orientation on the surfaces. Their shapes and sizes show only a few variations. The tubular diameter varied between 100 and 150 nm, and their length varied between 1,500 and 2,000 nm.

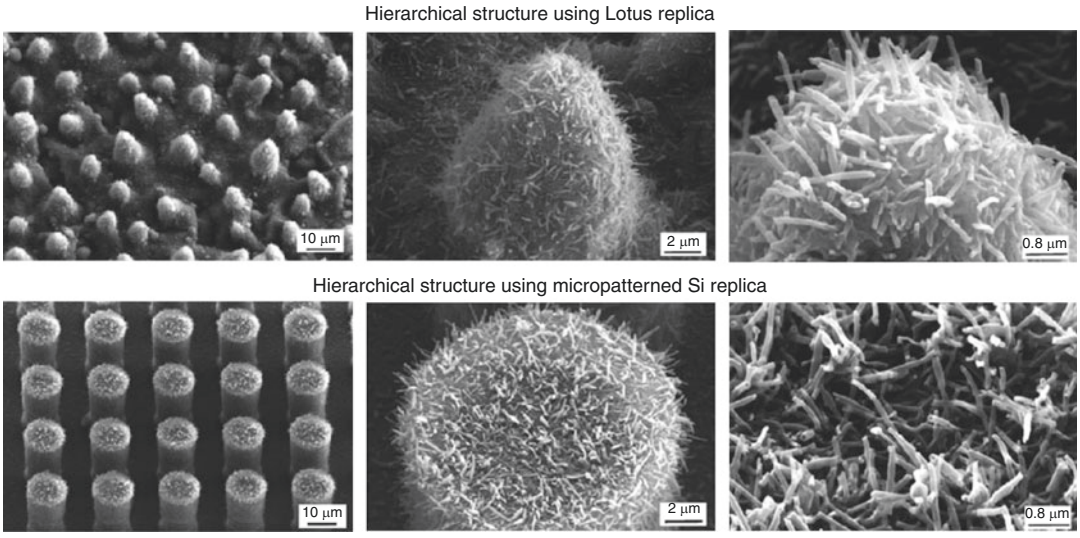
Lotus Effect, Table 1 Typical materials and corresponding techniques to produce micro/nanoroughness [2]

Material	Technique	Contact angle, °	Notes
Teflon	Plasma	168	
Fluorinated block polymer solution	Casting under humid environment	160	Transparent
PFOS	Electro- and chemical polymerization	152	Reversible (electric potential)
PDMS	Laser treatment	166	
PS-PDMS Block copolymer	Electrospining	>150	
PS, PC, PMMA	evaporation	>150	
PS nanofiber	Nanoimprint	156	
PET	Oxygen plasma etching	>150	
Organo-triethoxysilanes	Sol–gel	155	Reversible (temperature)
Al	Chemical etching	>150	
Copper	Electrodeposition	160	Hierarchical
Si	Photolithography	170	
Si	E-beam lithography	164	
Si	X-ray lithography	>166	
Si	Casting	158	Plant leaf replica
Si (Black Si)	Plasma etching	>150	For liquid flow
Silica	Sol–gel	150	
Epoxy resin with alkanes ($C_{36}H_{74}$)	Replication and self-assembly	169	Hierarchical
Polyelectrolyte multilayer surface overcoated with silica nanoparticles	Self-assembly	168	
Nano-silica spheres	Dip coating	105	
Silica colloidal particles in PDMS	Spin coated	165	Hierarchical
Au clusters	Electrochemical deposition	>150	
Carbon nanotubes	CVD	159	
ZnO, TiO_2 Nanorods	Sol–gel	>150	Reversible (UV irradiation)

Lotus Effect, Table 2 Pros and cons of various fabrication techniques [2]

Techniques	Pros	Cons
Lithography	Accuracy, large area	Slow process, high cost
Etching	Fast	Chemical contamination, less control
Deposition	Flexibility, cheap	Can be high temperature, less control

To study the effect of Lotus wax tubule nanostructures on superhydrophobicity, static contact angle, contact angle hysteresis, and tilt angle were measured on flat, microstructured Lotus replica; micropatterned Si replica; and hierarchical surfaces. Hierarchical surfaces were made of the Lotus leaf replica and micropatterned Si replica with a nanostructure of wax tubules on top. Additionally, fresh Lotus leaves were investigated to compare the properties of the fabricated structures with the original biological model. Figure 5 shows that the highest static contact angles of 173°, the lowest contact angle hysteresis of 1°, and tilt angle varying between 1 and 2° were found for the hierarchical structured Si replica. The hierarchical structured Lotus leaf replica showed a static contact angle of 171°, and the same contact angle hysteresis (2°) and tilt angles (1–2°) as the hierarchical Si replica. Both artificial

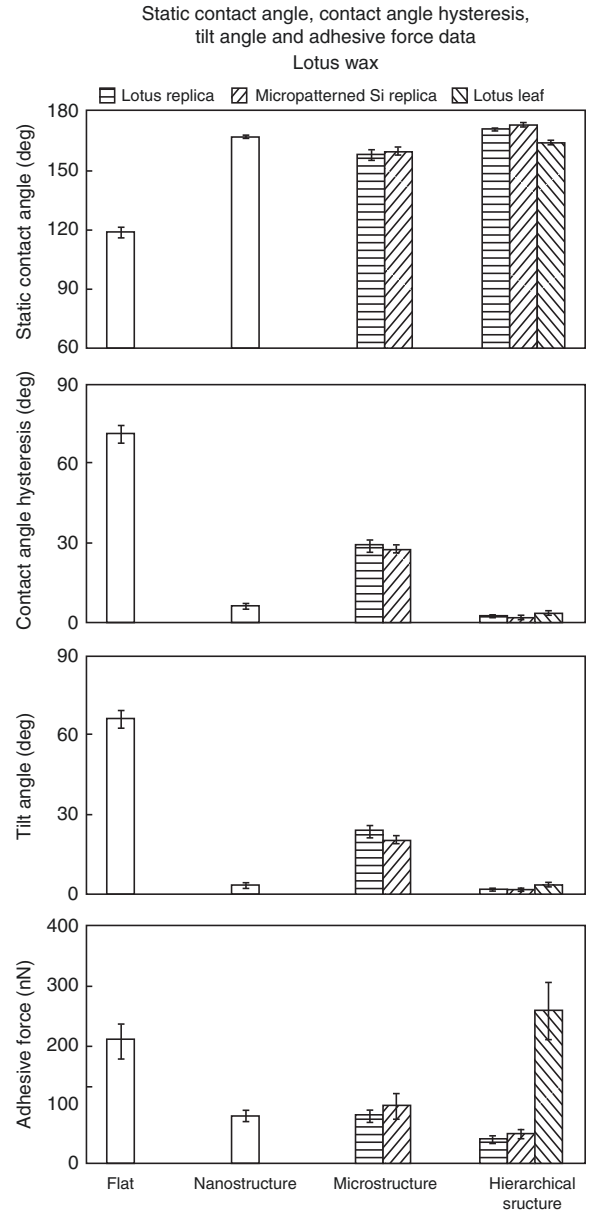


Lotus Effect, Fig. 4 SEM micrographs taken at 45° tilt angle (shown using three magnifications) of hierarchical structure using Lotus and micropatterned Si replicas. Nano- and hierarchical structures were fabricated with mass 0.8 μg mm^{-2} of Lotus wax after storage for 7 days at 50 °C with ethanol vapor [3]

hierarchical structured surfaces show similar values as the fresh Lotus leaf surface investigated here, with static contact angle of 164°, contact angle hysteresis of 3°, and a tilt angle of 3°. However, the artificial hierarchical surfaces showed higher static contact angle and lower contact angle hysteresis. Structural differences between the original Lotus leaf and the artificial Lotus leaf produced here are limited to a difference in length of wax tubules, which are 0.5–1 μm longer in the artificial Lotus leaf.

Adhesive force measured using a 15 mm radius borosilicate tip in an AFM also shows a similar trend as the wetting properties for the artificial surfaces. The adhesion force of the hierarchical surface structure was lower than that of the micro-, nanostructured, and flat surfaces because the contact between the tip and the surface was lower as a result of the contact area being reduced. However, for the fresh Lotus leaf, there is moisture within the plant material that causes softening of the leaf, and so when the tip comes into contact with the leaf sample, the sample deforms, and a larger area of contact between the tip and the sample causes an increase in the adhesive force.

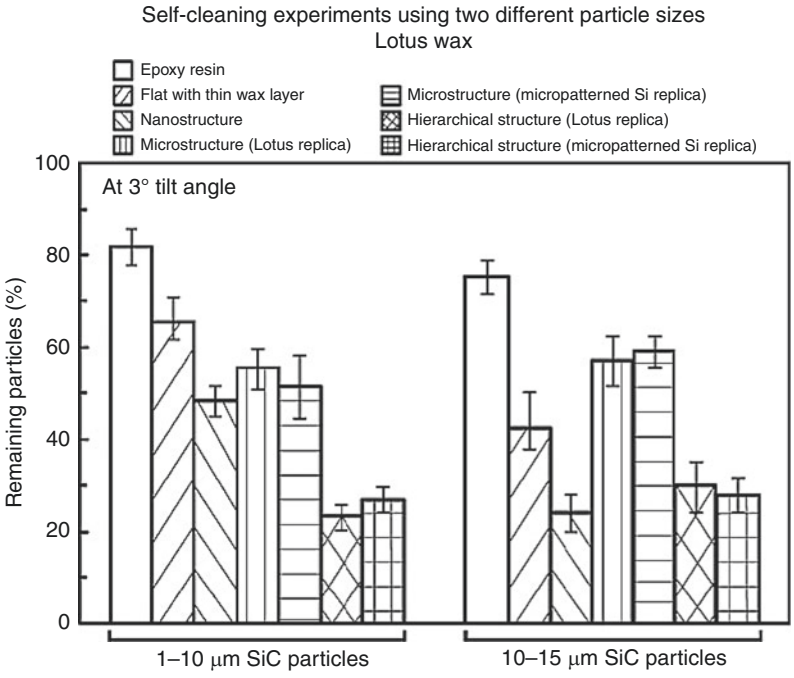
To test self-cleaning efficiency of hierarchical structured surfaces, silicon carbide particles in two different size ranges of 1–10 and 10–15 μm used as the contaminant were deposited. SiC particles were chosen because of their similarity in shape, size, and hydrophilicity to natural dirt contaminations. For the cleaning test, specimens with the contaminants were subjected to water droplets of approximately 2 mm diameter, using two microsyringes. Figure 6 shows that none of the investigated surfaces was fully cleaned by water rinsing. The data represent the average of five different investigated areas for each experiment. Most particles (70–80 %) remained on smooth surfaces, and 50–70 % of particles were found on microstructured surfaces. Most particles were removed from the hierarchical structured surfaces, but approximately 30 % of particles remained. A clear difference in particle removal, independent of particle sizes, was only found in flat and nanostructured surfaces, where larger particles were removed with higher efficiency. Observations of droplet behavior during movement on the surfaces showed that droplets rolled only on the hierarchical structured surfaces. On flat, micro-, and nanostructured surfaces, the droplets first applied did not move, but the continuous application of water droplets increased the droplet volumes and led to sliding of these large droplets.



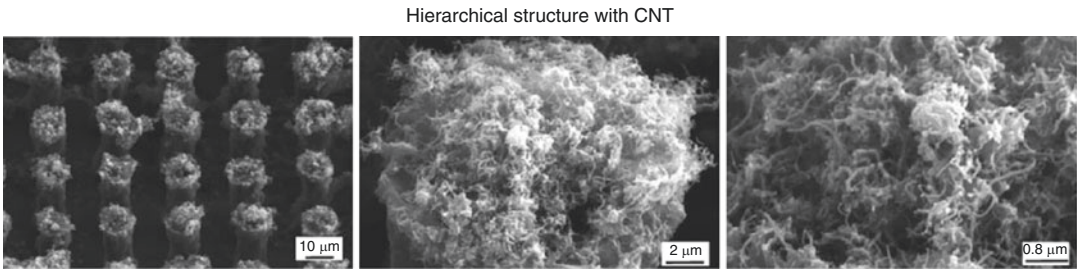
Lotus Effect, Fig. 5 (**a**–**c**) Bar charts showing the measured static contact angle, contact angle hysteresis and tilt angle on various structures fabricated with 0.8 μg mm^{-2} mass of Lotus wax after storage at 50 °C for 7 days with ethanol vapor. (**d**) Bar chart showing the adhesive forces for various structures, measured using a 15 μm radius borosilicate tip. Hierarchical structures were fabricated using Lotus and micropatterned Si replicas superimposed with nanostructures of Lotus wax tubules. The error bar represents ±1 standard deviation [3]

During this, some of the particles were removed from the surfaces. However, rolling droplets on hierarchical structures did not collect the dirt particles trapped in the cavities of the microstructures. The data clearly show that hierarchical structures have superior cleaning efficiency.

Fabrication of Durable Surfaces – For many commercial applications, the surfaces should have mechanical strength and chemical stability. Jung and Bhushan [7] prepared hierarchical superhydrophobic structures using carbon nanotube (CNT) – epoxy composite applied by spray method. Figure 7 shows SEM micrographs taken of the CNT-epoxy composite at three magnifications. The surface shows a homogeneous distribution of CNTs. As a benchmark, structured surfaces



Lotus Effect, Fig. 6 Bar chart showing the remaining particles after self-cleaning experiments applying droplets with nearly zero kinetic energy on various structures fabricated from Lotus wax using 1–10 and 10–15 μm SiC particles. The experiments on the surfaces with Lotus wax were carried out on stages tilted at 3°. The error bars represent ±1 standard deviation [3]

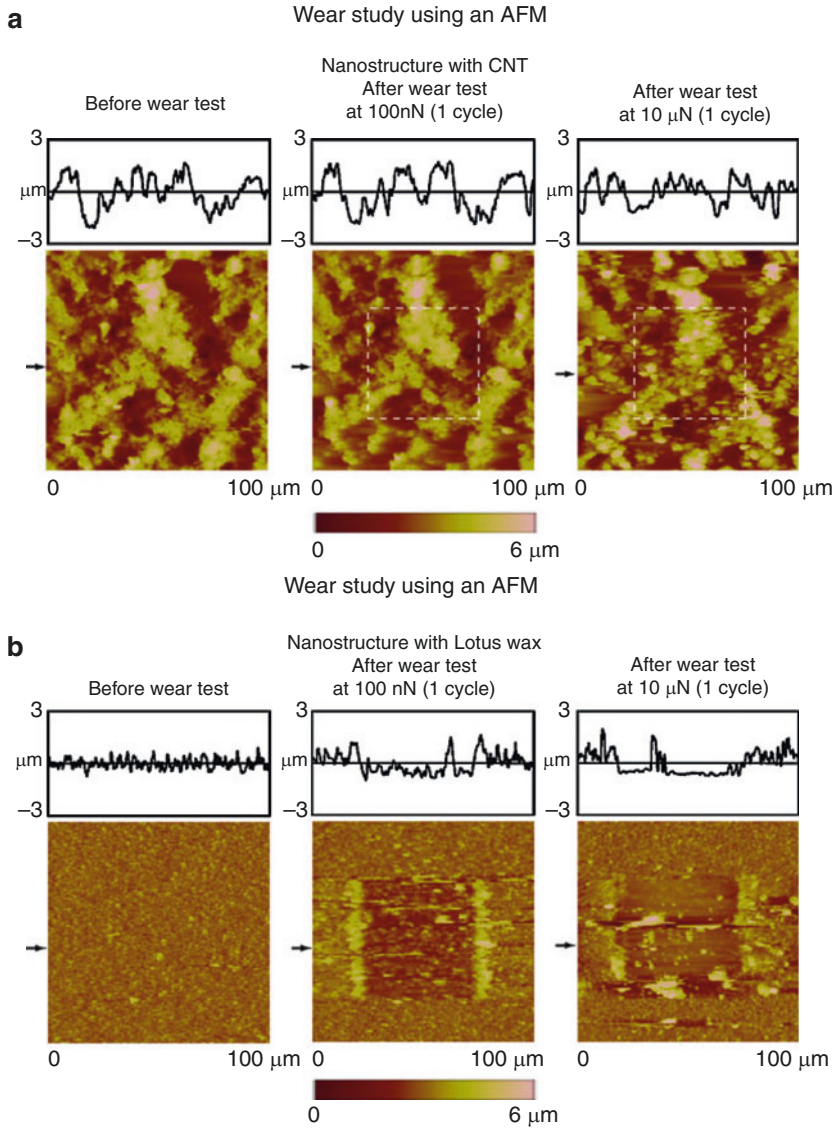


Lotus Effect, Fig. 7 SEM micrographs taken at 45° tilt angle showing three magnifications of hierarchical structures fabrication with CNT-epoxy composite [7]

with Lotus wax were also prepared to compare with the durability of CNT composite structures. To investigate the durability, wear tests were conducted by using an atomic force microscopy over 50 × 50 μm^2 area at two nominal loads of 100 nN and 10 μN. Figure 8 shows surface height maps before and after wear tests. These superhydrophobic CNT composite structures showed good mechanical durability, superior to the structured surfaces with Lotus wax. These structures may be suitable for real-world applications.

Closure

The Lotus effect provides one of the most successful examples of biomimetic surface effects. While theoretical foundations of wetting of rough surfaces have been developed decades ago, emerging applications and the need to design non-sticky surfaces has led to the intensive modeling and experimental research in the area. These findings provide new insights on the fundamental mechanisms of wetting and can lead to creation of successful nonadhesive surfaces.



Lotus Effect, Fig. 8 Surface height maps before and after wear tests with a 15 μm radius borosilicate ball at 100 nN and 10 μN for nanostructures with (**a**) CNTs and (**b**) Lotus wax using an AFM [7]

Cross-References

- ▶ Biomimetics
- ▶ Rose Petal Effect
- ▶ Wetting Transitions

References

1. Nosonovsky, M., Bhushan, B.: Multiscale Dissipative Mechanisms and Hierarchical Surfaces: Friction, Superhydrophobicity, and Biomimetics. Springer, Heidelberg (2008)

2. Bhushan, B., Jung, Y.-C., Nosonovsky, M.: Lotus effect: surfaces with roughness-induced superhydrophobicity, self-cleaning, and low adhesion. In: Bhushan, B. (ed.) Springer Handbook of Nanotechnology, 3rd edn, pp. 1437–1524. Springer, Heidelberg (2010)
3. Bhushan, B., Jung, Y.-C., Koch, K.: Micro-, nano- and hierarchical structures for superhydrophobicity, self-cleaning and low adhesion. Phil. Trans. R. Soc. A **367**, 1631–1672 (2009)
4. Bhushan, B.: Biomimetics inspired surfaces for drag reduction and oleophobicity/philicity. Beilstein J. Nanotechnol. **2**, 66–84 (2011)
5. Wenzel, R.N.: Resistance of solid surfaces to wetting by water. Ind. Eng. Chem. **28**, 988–994 (1936)
6. Cassie, A., Baxter, S.: Wettability of porous surfaces. Trans. Faraday Soc. **40**, 546–551 (1944)
7. Jung, Y.C., Bhushan, B.: Mechanically durable carbon nanotube-composite hierarchical structures with superhydrophobicity, self-cleaning, and low-drag. ACS Nano **3**, 4155–4163 (2009)

Low Carrier Density Materials for Infrared Plasmonics

► Terahertz and Infrared Plasmonics with Unconventional Materials

Low Fluid Drag Surface

► Shark Skin Effect

Low-Power

► Bioinspired CMOS Cochlea

Low-Pressure Chemical Vapor Deposition (LPCVD)

► Chemical Vapor Deposition (CVD)

LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design

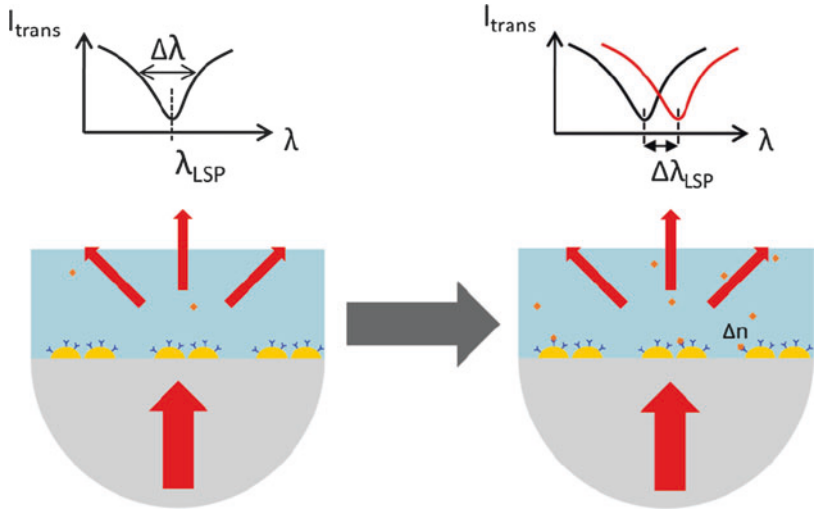
Ophélie Saison, Gaëtan Lévêque and Abdellatif Akjouj
Institut d'Electronique, de Microélectronique et de Nanotechnologie, UMR CNRS 8520, Université Lille 1, Sciences et Technologies, Villeneuve d'Ascq, France

Definition

A localized surface plasmon resonance sensing device allows probing a modification in the properties of a surface plasmon mode supported by a metal nanoparticle, in order to extract information on its local environment. For example, metal nanoparticles functionalized with a specific receptor will present a shift of their plasmon wavelength when the target molecule attaches to the metal interface.

Context

Plasmonics emerged mostly in the 1990s as an extremely rich new domain in optics, exploiting the particular optical modes supported by metal nanoparticles. When such a system is excited by an external light source, a resonant coupling between the surface charge of the metal-free electrons and the electromagnetic field scattered by the particle takes place at a specific wavelength, resulting in the excitation of a so-called localized surface plasmon (LSP) [1]. These modes are characterized by the squeezing of the optical field in very small volumes with typical dimension well below the light wavelength, where the intensity of light is enhanced. Such "hotspots" are usually localized at sharp tips of metal particles [2] and inside gaps of a particle dimer [3] or confined in between a particle and a metal film [4]. Its field of



LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design, Fig. 1 Principle of a LSPR-based biosensor. An interface is decorated with metal nanoparticles, themselves functionalized with a receptor specifically associated to a target molecule. In a reference situation where all receptors are free, the plasmon wavelength is λ_{LSP}, but it shifts by $\Delta\lambda_{LSP}$ in the presence of the target molecule which will change the environment of the nanoparticle (and its refractive index) when attached to the receptor

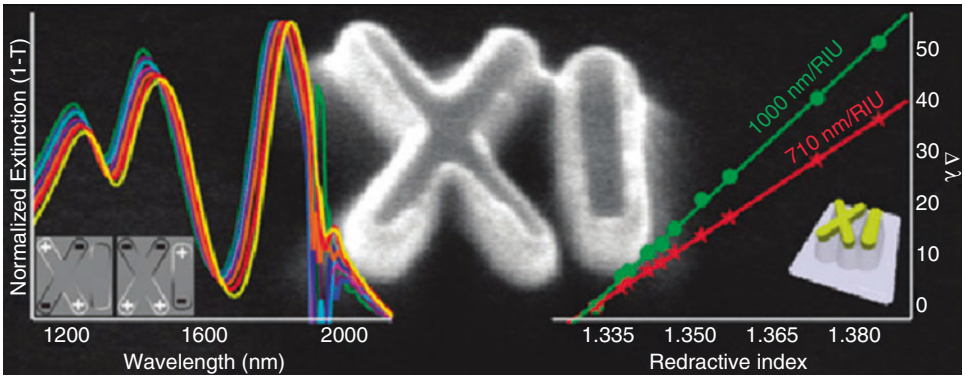
applications is very wide, ranging from simply downscaling the conduction of light on subwavelength scale [5], improving Raman spectroscopy for single molecule detection [6], increasing the efficiency of solar cells [7], creating new treatment in medicine using thermal properties of metal nanoparticles [8], or designing an entirely new type of metamaterials with optical properties not found in nature. Moreover, surface plasmon modes are very sensitive to small variations of their local environment, which make them suitable candidates for biosensing devices with very small detection volume [9–11].

The first plasmon-based sensing devices have early been proposed by Nylander et al. [12, 13] on the principle of *propagative* surface plasmon (PSP) modes, which are surface waves traveling along the extended interface between a metal and a dielectric. PSP-based biosensing devices have already been commercialized for more than 20 years now [14] and are widely used in medicine, biology, or food safety. However, they suffer from several limitations, mostly due to their cost and difficulty of use, as this technique is mostly accessible to specialists in surface chemistry and are not easily portable, which prevent them from being more widely used as in certain immunoassays which could be routinely performed and accessible to everyone. LSPs offer a parallel route to cheaper and more flexible sensing devices, as they have large enhancement of the electric field intensity near the particle, the penetration depth of light is one order of magnitude shorter than for the PSP, and the resonance wavelength can be tuned in the visible range by modifying the shape of noble metal nanoparticles. For instance, small spherical gold spheres in water support LSPs around 520 nm with electromagnetic fields localized within a few 10 nm from their surface. That wavelength will slightly shift with as a small modification of the refractive index around the particle [15], induced, for example, by adsorption of a target molecule on its surface, as pictured in Fig. 1.

The performances of sensors based on surface plasmons are usually quantified with the refractive-index sensitivity S and the figure of merit (FoM):

$$S = \frac{\Delta\lambda_{LSP}}{\Delta n}, \quad FoM = \frac{S}{\Delta\lambda}$$

The refractive-index sensitivity is equal to the shift of the excitation wavelength per refractive



LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design, Fig. 2 Nanocross/nanobar plasmonic system and their sensitivity properties, from Verellen et al. [22]

index units (RIU), whereas the figure of merit is equal to the ratio of the refractive-index sensitivity to the resonance full half width maximum (FWHM), $\Delta\lambda$. Finally, the resolution characterizes the minimum measurable variation of refractive index, in RIU.

Currently, the largest optical sensitivities are obtained with PSP-based sensors [16]. The typical device is based on the standard prism Kretschmann configuration and allows typical resolution of about $10^{-7} - 10^{-8}$ RIU, with refractive-index sensitivities of about $10^4 - 10^5$ nm/RIU. In the first place, LSP modes seem less interesting as their bulk sensitivity is much lower than for PSP modes: typically, a gold nanosphere of 50 nm in diameter exhibits an optical sensitivity of about 60 nm/RIU. However, as the detection volume is much smaller, few molecules adsorbed on the particle lead to a large wavelength shift: finally, interfacial sensitivities are much larger. A great deal of work has been done in trying to improve the sensitivity and decrease the FWHM of the plasmon resonances. The first method is to replace gold with silver, as this metal has much lower losses in the visible. This is not easy in practical because silver is easily oxidized in wet environment. In 2001, Malinsky et al. [17] demonstrated an optical sensitivity of about 150 nm/RIU for a grating of silver nanoparticles fabricated by nanosphere lithography, when particles are functionalized with self-assembled monolayers. By playing with more complex shapes, bulk refractive-index sensitivities above 800 nm/RIU were obtained in gold nanorice, nanorings, and nanocrescents [18–20]. Gold nanorice and nanocrescents showed in addition plasmonic modes located in the near-infrared region. A lot of new approaches have also been proposed more recently and demonstrated to improve the sensing performance by shifting the plasmon excitation into the near infrared. They often rely on designing complex particles supporting subradiant plasmon modes with asymmetric profiles. For example, a double split nanoring cavity designed to suppress radiative damping showed bulk sensitivities of 1200 nm/RIU [21], whereas Verellen et al. demonstrated the coherent coupling of bright and dark plasmon modes in a gold nanocross and nanobar, with bulk RI sensitivities exceeding 1,000 nm/RIU and near-infrared plasmons [22]. With a completely different approach, Lodewijks et al. [23] have recently shown that the FoM of a LSP-based sensing device could be largely increased using phase measurement. With an average refractive-index sensitivity of about 200–300 nm/RIU, they claimed linewidths one order of magnitude smaller and a FoM four to six times larger than intensity-based measurements using random ensemble of gold disks and rings (Fig. 2).

For sake of simplicity, this entry is only based on the first approaches, based on the wavelength shift of the LSP resonance.

LSPR in the Electrostatic Approximation

The basic principles of localized surface plasmon resonances and its application to sensing can be easily captured when the particle is much smaller than the wavelength of the incident light (quasistatic approximation). In that limit, a metal spherical particle of volume V with complex dielectric constant $\epsilon(\lambda)$ at wavelength λ, placed in a homogeneous transparent substrate of dielectric constant ϵ_B, is fully characterized by its polarizability $\alpha(\lambda)$:

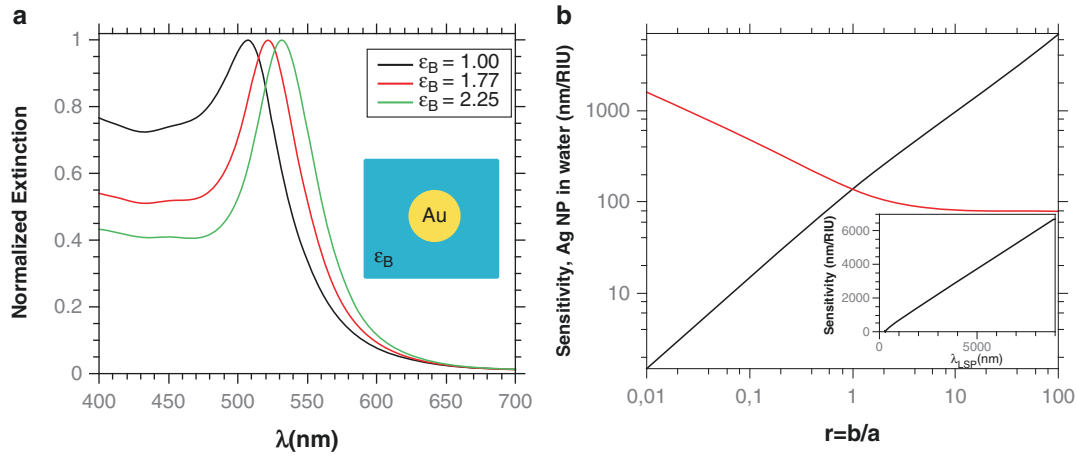
$$\alpha(\lambda) = 3V\epsilon_B \frac{\epsilon(\lambda) - \epsilon_B}{2\epsilon_B + \epsilon(\lambda)}$$

As in the visible range of wavelengths the dielectric constant of a noble metal particle is negative, the modulus of the polarizability can reach very large values when $\Re\{\epsilon(\lambda)\} \approx -2\epsilon_B$ (in case of negligible losses), where $\Re$ is the real part. The wavelength for which the polarizability modulus is maximum corresponds to the excitation wavelength of the localized surface plasmon mode, which, in this quasistatic limit, is a dipole parallel to the direction of the incident electric field.

As an illustration, Fig. 3a shows the normalized extinction spectra, proportional to the imaginary part of $\alpha(\lambda)$, for a spherical gold nanoparticle in the quasistatic approximation in air ($\epsilon_B = 1.00$), water ($\epsilon_B = 1.77$), and silica ($\epsilon_B = 2.25$). The dielectric constant for gold was taken from [24]. Hence, the LSPR is sensitive to a change in the refractive index of the environment. The refractive-index sensitivity can be estimated to about $S = 45$ nm/RIU in that situation, which gives a FoM $= 0.75$ RIU^{-1}, quite low compared to what can be achieved with more complex designs.

The refractive-index sensitivity can be improved by working with elongated nanoparticles like spheroids or nanorods. For these geometries, the electrical field is more enhanced at the tips of the particle and an increase in the particle aspect ratio results in an increase in S. For a spheroidal particle in the quasistatic approximation, the electric field created inside the particle is generally not parallel to the induced dipole. In each i direction of the three particle main axes, the particle responds to the excitation with a polarizability equal to

$$\alpha_i = V\epsilon_B \frac{\epsilon - \epsilon_B}{\epsilon_B + L_i(\epsilon - \epsilon_B)} \quad i = 1, 2, 3$$



LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design, Fig. 3 **(a)** Normalized extinction spectra of a spherical gold nanoparticle inside substrates of different dielectric constant; **(b)** Evolution of the refractive-index sensitivities of a spheroidal silver nanoparticle for the transverse (*red*) and longitudinal (*black*) LSP modes as a function of the aspect ratio r of the ellipsoid; *inset*: refractive-index sensitivity as a function of the resonance wavelength

where the coefficient L_i is a geometrical factor describing the shape of the spheroid (for a spherical particle, $L_i = 1/3$). The expressions of the depolarization factors can be found in [25]. Then, to each i direction corresponds one specific localized plasmon mode excited at one specific wavelength. Taking a Drude model for the metal, it is easy to evaluate the resonance wavelength and the refractive-index sensitivity of each plasmon mode:

$$\epsilon(\lambda) = \epsilon_\infty\left(1 - \lambda^2/\lambda_p^2\right)$$

$$\rightarrow \frac{\lambda_{LSP}^i}{\lambda_p} = \sqrt{1 + \frac{\epsilon_B}{\epsilon_\infty}\left(\frac{1}{L_i} - 1\right)}$$

$$S_i = \frac{d\lambda_{LSP}^i}{dn} = \frac{n\lambda_p\left(\frac{1}{L_i} - 1\right)}{\epsilon_\infty\sqrt{1 + \frac{\epsilon_B}{\epsilon_\infty}\left(\frac{1}{L_i} - 1\right)}}$$

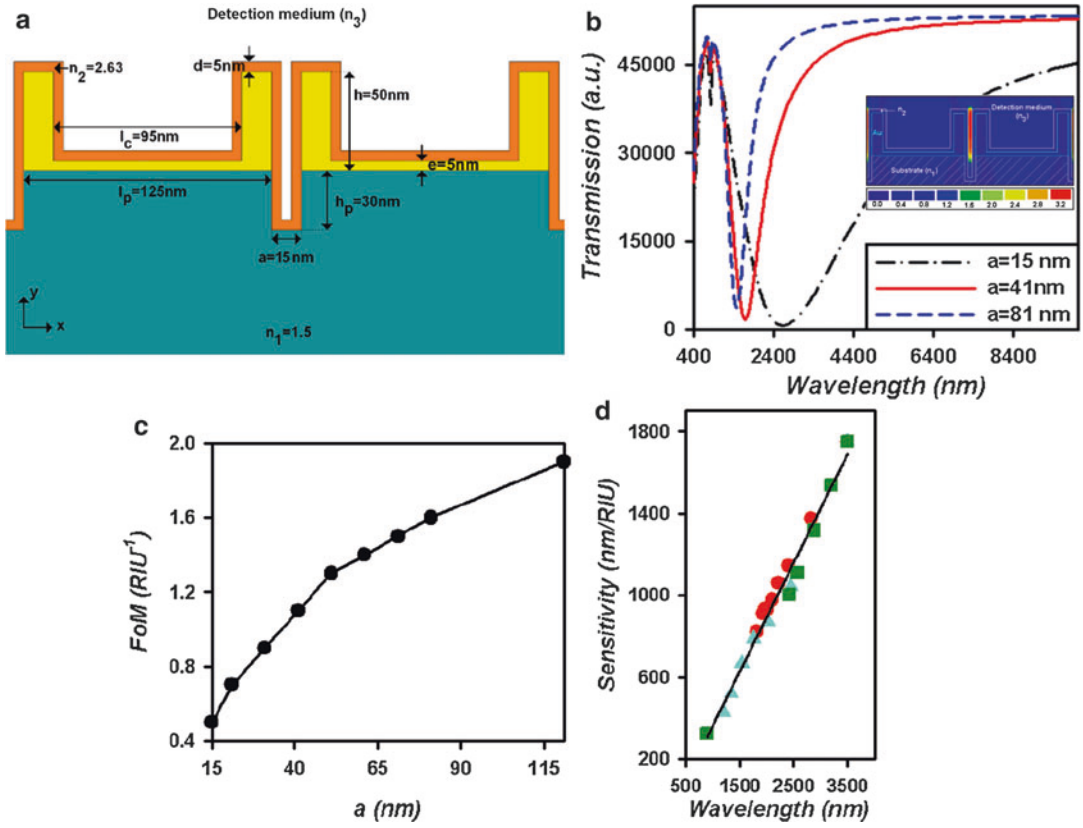
$$= \frac{1}{n}\left(\lambda_{LSP}^i - \frac{\lambda_p^2}{\lambda_{LSP}^i}\right)$$

This equation shows that, in the quasistatic approximation, the refractive-index sensitivity of the metal nanoparticle is only a function of the plasmon resonance wavelength. The dependency on the shape, included here in the coefficients L_i, occurs only through the position of the plasmon wavelength [26]. Figure 3b shows the evolution of the refractive-index sensitivity of the spheroid in water as a function of the aspect ratio $r = b/a$, b being the long axis and a the two equal short axes of the ellipsoid. The particle is supposed to be made of silver, described by the Drude model with $\epsilon_\infty = 4.0$ and $\lambda_p = 284.0$ nm. It has two degenerated dipolar plasmon modes along the section of the spheroid (transverse mode) and one along its symmetry axis (longitudinal mode). The inset shows the mode-independent evolution of the sensitivity S as a function of the particle surface plasmon wavelength, where it appears that it can reach very high values, either for very elongated or very flat particles, but at the expense of a large redshift of the surface plasmon resonance wavelength. This fact has to be taken into account in biosensor design, as they must work at a wavelength lower than 1900 nm in order to avoid absorption of water.

A More Advanced Example

As an illustration of sensor design beyond the electrostatic approximation, Fig. 4 shows the optical properties of a periodic arrangement of more complex plasmonic particles [26]. The structure consists in a planar interface superposed with an array of dielectric pillars supporting a U-shaped gold nanostructure. The whole system is eventually covered with a thin dielectric layer (Fig. 4a). The pillars decrease the spatial overlap between the substrate and the plasmon field generated at the resonance, resulting in higher refractive-index sensitivities. Typical transmission spectra obtained by finite difference time domain (FDTD) simulations are shown in Fig. 4b, for several interparticle spacing a, the detection medium being air. Such a structure shows a transmission spectrum dominated by one peak in the infrared domain, at $\lambda = 2621$ nm for $a = 15$ nm. The position of that peak is very sensitive to every parameter of the structure, in particular the spacing a, as the shift is about 1,000 nm when the interparticle distance varies from 15 to 81 nm. The origin of the large dependency of the infrared plasmon band is due to the high electrical field confinement in the U-shaped gold nanostructures. The inset of Fig. 4b shows that maximum field enhancement is observed in between the two arms of two consecutive U-shaped gold nanostructures, with a very low but not zero electrical field within the U cavity.

For the shortest interparticle spacing, the wavelength shift obtained by increasing the refractive index of the detection medium from 1.00 to 1.50 is equal to 875 nm, which corresponds to a sensitivity of $S = 1{,}750$ nm/RIU, much larger than the sensitivities reported by Verellen et al. and Liu et al. using nanocross/nanobar structures [22] or double split nanoring cavities [21]. However, the rather broad plasmon band in the near-infrared spectral region results in low FoM, as the



LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design, Fig. 4 **(a)** U-shaped plasmonic structure investigated in [26] and its geometrical parameters. In the calculations, the nanostructure is illuminated by a p-polarized plane wave from below, the transmission is computed above the structure.; **(b)** Transmission spectra computed for several values of the interparticle distance a; **(c)** Evolution of the figure of merit with a.; **(d)** Sensitivity as a function of the resonance wavelength, varying a (*red circles*), e (*blue triangles*), and l_c (*green squares*)

FWHM $\Delta\lambda \approx 3{,}400$ nm gives a figure of merit of about $FoM = 0.5$ RIU^{-1}. It compares to that obtained by Chen et al. on ensembles of gold spheres ($FoM = 0.6$) [27] but is much lower than for gold nanocrescents ($FoM = 2.4$ RIU^{-1}) [20], gold stars ($FoM = 5.4$ RIU^{-1}) [28], nanocross/nanobar structures ($FoM = 12.5$ RIU^{-1}) [22], or double split nanoring cavities by Liu et al. ($FoM = 8.5$ RIU^{-1}) [21].

The influence of all the morphological dimensions (a, e, l_c) on the interface plasmonic response is summarized in the following table. It appears that whatever the parameter is changed, the sensitivity seems to increase linearly with the resonance wavelength, following the properties of spheroidal particles in the electrostatic approximation (Fig. 4d).

a	e	l_c	λ	S	FoM
nm	nm	nm	nm	nm/RIU	1/RIU
15	5	95	2621	1750	0.5
51	5	95	1608	980	1.3
121	5	95	1409	823	1.9
15	15	95	1599	868	0.5
15	25	95	109	667	0.7
15	5	75	2427	1540	0.5
15	5	55	2221	1318	0.4

For clinical applications it will be relevant to work with plasmonic interfaces with resonance wavelength just below the water absorption band

at 1900 nm. The best compromise between sensitivity and resonance wavelength has been found for the following parameters:

- U-shaped gold nanostructure: $l = 120$ nm, $h = 50$ nm, $a = 45$ nm, $e = 10$ nm, $l_c = 90$ nm
- Dielectric pillar $l_p = 120$ nm, $h_p = 30$ nm
- Substrate and dielectric overlayer: $n_1 = 1.45$, $n_2 = 2.63$, $d = 5$ nm

The resonance wavelength is then 1,200 nm for $n_3 = 1.00$, and the corresponding refractive-index sensitivity and FoM are respectively $S = 783$ nm/RIU and $FoM = 1.4$ RIU^{-1}.

References

1. Akjouj, A., Leveque, G., Szunerits, S., Pennec, Y., Djafari-Rouhani, B., Boukherroub, R., Dobrzynski, L.: Nanometal plasmon polaritons. Surf. Sci. Rep. **68**, 1–67 (2013)
2. Martin, O.J.F.: Plasmon resonances in nanowires with a non-regular cross-section. Top. Appl. Phys. **88**, 183–210 (2003). Springer Verlag Series
3. Novotny, L., van Hulst, N.: Antennas for light. Nat. Photon. **5**, 83–90 (2011)
4. Yamamoto, N., Ohtani, S., Garca De Abajo, F.J.: Gap and Mie plasmons in individual silver nanospheres near a silver surface. Nano Lett. **11**(1), 91–95 (2011)
5. Barnes, W.L., Dereux, A., Ebbesen, T.W.: Surface plasmon subwavelength optics. Nature **424**, 824–830 (2003)
6. Fleischmann, M., Hendra, P.-J., Mequilla, A.-J.: Raman spectra of pyridine adsorbed at a silver electrode. Chem. Phys. Lett. **26**, 163–166 (1974)
7. Liu, L., Ouyang, S., Ye, J.: Gold-nanorod-photosensitized titanium dioxide with wide-range visible-light harvesting based on localized surface plasmon resonance. Angew. Chem. **125**, 6821–6825 (2013)
8. Hirsch, L.R., et al.: Nanoshell-mediated near-infrared thermal therapy of tumors under magnetic resonance guidance. Proc. Natl. Acad. Sci. **100**, 13549–13554 (2003)
9. Yao, Y., Yi, B., Xiao, J., Li, Z.: Surface plasmon resonance biosensors and its application. First International Conference on Bioinformatics and Biomedical Engineering, 1043–1046 (2007). doi:10.1109/ICBBE.2007.270
10. Szunerits, S., Boukherroub, R.: Sensing using localised surface plasmon resonance sensors. Chem. Commun. Camb. Engl. **48**, 8999–9010 (2012)
11. Saison, O., Leveque, G., Akjouj, A., Pennec, Y., Djafari-Rouhani, B., Boukherroub, R., Szunerits, S.: Plasmonic nanoparticles array for high-sensitivity sensing: A theoretical investigation. J. Phys. Chem. C **116**, 17819–17827 (2012)
12. Nylander, C., Liedberg, B., Lind, T.: Gas detection by means of surface plasmon resonance. Sens. Actuators **3**, 79–88 (1983)
13. Liedberg, B., Nylander, C., Lunstrom, I.: Surface plasmon resonance for gas detection and biosensing. Sens. Actuators **4**, 299–304 (1983)
14. Baird, C.L., Myszka, D.G.: Current and emerging commercial optical biosensors. J. Mol. Recognit. **14**, 261–268 (2001)
15. Underwood, S., Mulvaney, P.: Effect of the solution refractive index on the color of gold colloids. Langmuir **10**, 3427–3430 (1994)
16. Wijaya, E., Lenaerts, C., Maricot, S., Hastanin, J., Habraken, S., Vilcot, J.-P., Boukherroub, R., Szunerits, S.: Surface plasmon resonance-based biosensors: From the development of different SPR structures to novel surface functionalization strategies. Curr. Opin. Solid State Mater. Sci. **15**, 208–224 (2011)
17. Malinsky, M.D., Kelly, K.L., Schatz, G.C., Van Duyne, R.P.: Chain length dependence and sensing capabilities of the localized surface plasmon resonance of silver nanoparticles chemically modified with alkanethiol self-assembled monolayers. J. Am. Chem. Soc. **123**, 1471–1482 (2001)
18. Larsson, E.M., Alegret, J., Kall, M., Sutherland, D.S.: Sensing characteristics of NIR localized surface plasmon resonances in gold nanorings for application as ultrasensitive biosensors. Nano Lett. **7**, 1256–1263 (2007)
19. Wang, H., Brandl, D.W., Le, F., Nordlander, P., Halas, N.: Nanorice: A hybrid plasmonic nanostructure. Nano Lett. **6**, 827–832 (2006)
20. Bukasov, R., Shumaker-Parry, J.S.: Highly tunable infrared extinction properties of gold nanocrescents. Nano Lett. **7**, 1113–1118 (2007)
21. Liu, S.-D., Yang, Z., Liu, R.-P., Li, X.-Y.: High sensitivity localized surface plasmon resonance sensing using a double split NanoRing cavity. J. Phys. Chem. B **115**, 24469–24477 (2011)
22. Verellen, N., Van Dorpe, P., Huang, C.-C., Lodewijks, K., Vandenbosch, G.E.A., Lagae, L., Moshchalkov, V. V.: Plasmon line shaping using nanocrosses for high sensitivity localized surface plasmon resonance sensing. Nano Lett. **11**, 391–397 (2011)
23. Lodewijks, K., Van Roy, W., Borghs, G., Lagae, L., Van Dorpe, P.: Boosting the figure-of-merit of LSPR-based refractive index sensing by phase-sensitive measurements. Nano Lett. **12**, 1655–1659 (2012)
24. Johnson, P.B., Christy, R.W.: Optical constants of noble metals. Phys. Rev. B **6**, 4370–4379 (1972)
25. Bohren, C.F., Huffmann, D.R.: Absorption and scattering of light by small particles. Wiley-VCH, New York (1998)
26. Saison-Francioso, O., Lévêque, G., Akjouj, A., Pennec, Y., Djafari-Rouhani, B., Boukherroub, R.,

Szunerits, S.: Search of extremely sensitive near−infrared plasmonic interfaces: a theoretical study. Plasmonics **8**, 1691–1698 (2013)
27. Chen, H.J., Kou, X.S., Yang, Z., Ni, W., Wang, J.F.: Shape- and size-dependent refractive index sensitivity of gold nanoparticles. Langmuir **24**, 5233–5237 (2008)
28. Nehl, C.L., Liao, H.W., Hafner, J.H.: Optical properties of star-shaped gold nanoparticles. Nano Lett. **6**, 683–688 (2006)

LSPR-Based Sensing

► Refractometric Sensing Using Plasmonic Nanoparticles

Lubrication

► Nanotribology

Lumbricidae/Oligochaeta/Earthworms

► Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms

M

Macromolecular Crystallization Using Nano-volumes

Gabriela Juarez-Martinez
Centeo Biosciences Limited, Dumbarton, UK

Synonyms

Biological nano-crystallization; Microarrays/microfluidics for protein crystallization; Protein nano-crystallization

Definitions

Biological macromolecules are large molecules or biopolymers which are made of smaller organic components that are linked together; they usually include proteins, nucleic acids, carbohydrates, and lipids.

Crystal comes from the Greek word "krustallos" meaning "clear ice." Crystals are geometrical solids with regular faces and with well-defined edges. They are formed by atoms, molecules, ions, or molecular assemblies that are arranged in an ordered and repeating pattern which extends over all three spatial dimensions. Crystals formed from biological macromolecules such as proteins are smaller than their inorganic and small molecule counterparts. Bio-macromolecular crystals are also very fragile as they may have between 20 % and 80 % solvent content; hence they are prone to dehydration which leads to cracking and destruction of the crystal. For this reason, bio-macromolecular crystals need to be kept in a constant solvent saturated environment [1].

Crystallization refers to the process of forming or creating crystals, this process can be naturally or artificially induced. The term "protein crystallization" is often used instead of bio-macromolecular crystallization, and although strictly speaking it should only apply when referring to proteins. However, the crystallization techniques and methods are the same. There are many ways to crystallize a compound and the crystallization processes depend on the molecule or substrate to be crystallized. However, all crystallization process have two things in common: Firstly, all types of crystallizations are multiparametric processes, which means that a variety of factors such as temperature, pH, purity, time, and pressure, among others, can affect the outcome of the process; secondly, they all involve three steps: nucleation, growth, and cessation of growth. These steps will be reviewed later in the chapter.

Structural Biology refers to the study of tridimensional structures of biological macromolecules in order to better understand their function and interactions with other molecules at a molecular level.

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Introduction

To understand the importance of using nano-volumes to improve the crystallization of biological macromolecules, it is necessary to look at the historical motivation behind it. More than a century ago, scientists were fascinated with the creation of crystals of biological nature. In 1840, Hünefeld observed hemoglobin crystals in samples of almost dry blood. In 1851, Funke described a method to crystallize pure hemoglobin [2]. This was followed by the crystallization of urease by Sumner and pepsin by Northrop. They showed that urease and pepsin were pure proteins but also that they had an enzymatic activity. This was an important discovery, as at the time it was believed that proteins were incapable of catalyzing biological reactions, with proteins regarded as just carriers for the true catalytic molecules, enzymes. In 1946, the Nobel Prize in Chemistry was awarded to Sumner for the discovery that enzymes can be crystallized and to Stanley and Northrop for their preparation of enzymes and viruses in pure form [1].

It was not until 1934 that Bernal and Hodgkin produced the first X-ray diffraction pattern of a protein, pepsin. They showed that in order to obtain a good diffraction pattern, protein crystals should not be kept dry but studied surrounded by the liquid solution where they were grown from. This prompted the interest of many scientists and soon after the tridimensional structure of several proteins were reported including myoglobin by Kendrew et al.; hemoglobin by Perutz and Kendrew; and penicillin, vitamin B12, and insulin by Hodgkin [3].

The latest big push in protein crystallization came as a result of the various Genome projects carried out in the 1990s, when a large number of genes from different organisms were sequenced. Many of these genes coded for proteins of unknown function; consequently in the late 1900s and early 2000, a number of initiatives known as Structural Genomics Projects were set up across the world to determine in a high-throughput format the tridimensional structure of those proteins. The results of these projects have applications in many fields such as medicine, pharmacology, veterinary, and agriculture.

Determining the tridimensional structure of unknown proteins allows scientist to better understand their function and interactions with other biological molecules at a molecular level. This is done by comparing different parts of the tridimensional structure of an unknown protein to a structural database of proteins of known function, revealing similarities from which biochemical and biological function may be deduced.

Knowing the tridimensional structure of proteins of known function is also important for rational drug design, where the protein is usually involved in a disease process and can be used as a target where drug candidates can bind to inhibit, block, or modify its function. In rational drug design, scientists analyze the tridimensional structure of the target (usually a protein) and computer-generate a number of drug candidates (or fragments that can later be joined to form a drug candidate) that can bind to the target with high affinity. These drug candidates are then produced in the lab and tested against the target. The goal is to generate drugs that have higher affinity and selectivity for the target, therefore reducing the number of side effects [4, 5].

There are two main techniques for determining the tridimensional structure of a protein; these are nuclear magnetic resonance (NMR) and X-ray diffraction. The use of NMR is currently limited by the size of the macromolecule, (currently around 35 kDa) although advances in procedures may increase this limit. X-ray diffraction is the most common technique but requires the generation of biological crystals that are capable of diffracting at high resolution. Unfortunately, there is no way to know a priori if a biological molecule will crystallize or under what conditions. Hence, the common practice is to screen the target molecule against a large number of crystallization parameters including, but not limited to: pH, ionic strength, temperature, and chemical additives. Consequently, the crystallization of biological macromolecules has remained as the bottleneck of structural genomics.

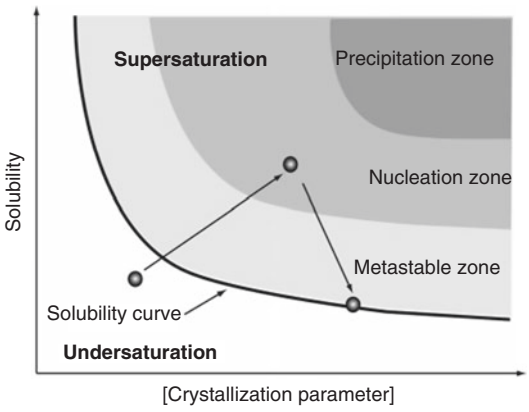
Protein Crystallization Theory

In order to create a protein crystal, it is necessary to have a highly pure and stable protein in solution and then force the protein molecules to aggregate in an ordered manner by modifying one or more of the crystallization parameters highlighted above. The formation of a protein crystal can be divided in three stages: *nucleation*, *growth*, and *cessation of growth*:

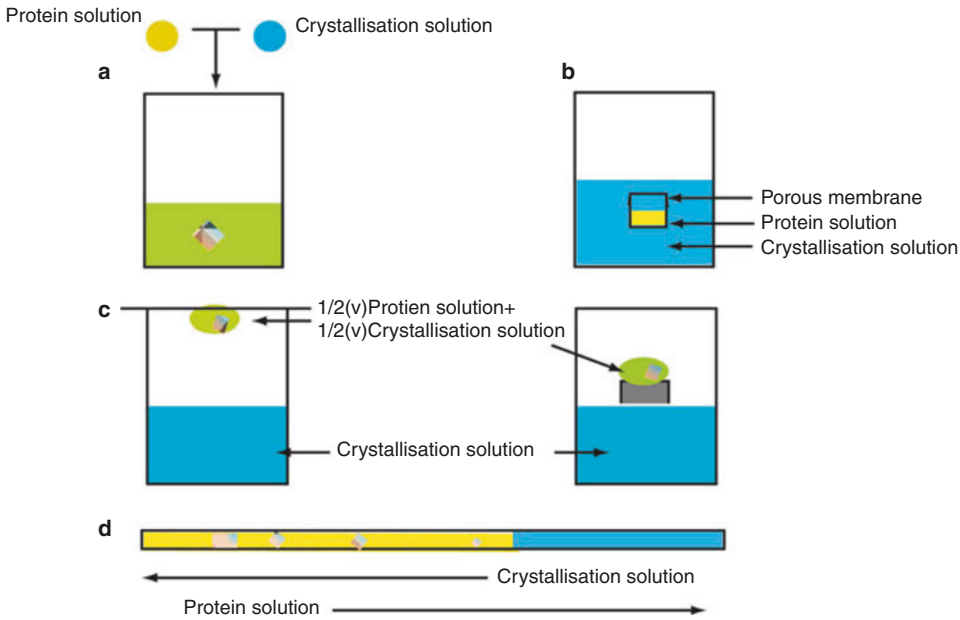
- *Nucleation*: This is the stage where small aggregates of protein are formed. If the aggregates have an internal order they will yield a macroscopic crystal, if not, an amorphous precipitate will be formed. The nucleation can be subclassified into a number of processes, namely:
  - *Primary nucleation*: It is defined as the initial aggregation of the same type of molecules as a consequence of a crystallization process at a high degree of supersaturation. Furthermore, primary nucleation can be divided into a homogeneous or heterogeneous process. The former is produced when only the target macromolecules form a crystal in the bulk solution. The latter is when other molecules or the walls of the crystallization chamber act as a nucleation center.
  - *Secondary nucleation*: It occurs after primary nucleation in the presence of the already formed crystals. It can be divided into attrition and catalytic mechanisms. In the former, the waves of diffusion or convection that transport the molecules in solution to the surfaces of the crystal collide directly with the crystal and can detach part of it. Thus, a new stable crystal nucleus is formed and can grow in this zone. The second type of secondary nucleation occurs when a defect on the surface of the crystal appears due to the addition of a new molecule in a disorientated way or due to the addition of an impurity. This creates a new nucleus, which may form a twin crystal [6].
- *Growth*: Once a stable nucleus is formed, its growth is thermodynamically favored. Molecules of smaller size than the stable nucleus are transported to the surface of the stable nucleus to crystallize [6].
- *Cessation of growth*: A crystal stops growing when the concentration of the macromolecules in solution decreases to the solubility level or when the surface of the crystal changes. These changes can be due to the addition of impurities that "poison" the surface of the crystal, such that the addition of new molecules becomes unfavorable [6].

Importance of Solubility in Protein Crystallization

In order to control the crystallization of proteins, it is important to understand phase diagrams, which describe the solubility of proteins (Fig. 1). Initially the protein is in a concentrated, stable but soluble state of maximum solubility (undersaturation zone). By altering one or more crystallization parameters, it is possible to make the system shift out of equilibrium toward the supersaturation zone. Once the system arrives at a critical point in the supersaturation where it can overcome the free activation barrier, a spontaneous nucleus will form (nucleation zone). In this zone, the solid



Macromolecular Crystallization Using Nano-volumes, Fig. 1 Schematic description of a phase diagram for the crystallization of Bio-macromolecules



Macromolecular Crystallization Using Nano-volumes, Fig. 2 Shows a schematic representation of the most common protein crystallization methods: (**a**) batch; (**b**) dialysis; (**c**) vapor diffusion and (**d**) liquid diffusion crystallization methods. The schematic does not describe the dimensions of the containers as it is intended only to describe the process; however, any of these methods can be implemented in macro- or microscale

phase cannot accumulate and the system slowly returns to equilibrium, on its way it crosses a metastable zone where no more spontaneous nucleation will occur but it allows the already formed nucleus to grow. Once the system returns to equilibrium by reaching the solubility curve, the crystal will not grow any further nor will it dissolve as the solid phase is in equilibrium with the saturated protein in solution [7].

Protein Crystallization Methods

There are four basic methods to crystallize a protein: *batch*, *dialysis*, *vapor diffusion*, and liquid *diffusion* crystallization methods. It is important to mention that no method is considered superior to any other. The choice depends on the protein involved, its quality, and the preference of the researcher [8].

Batch crystallization method (Fig. 2a): This is the simplest and oldest method. Here the solution containing the protein of interest is mixed with the crystallizing solution that is at a concentration designed to reach the supersaturation zone instantaneously, forming a number of nuclei that can grow by returning to the metastable zone [8, 9].

Dialysis crystallization method (Fig. 2b): The solution that contains the protein is separated from the crystallizing solution by a semipermeable membrane of a specific pore size, which is smaller than the size of the protein and only allows small molecules (ions, additives) into and out of the protein compartment. This exchange of molecules allows the system to reach supersaturation and then equilibrium [8, 9]. Dialysis is a method that is suitable for independently optimizing the nucleation phase from the growth phase, as the nucleation phase requires a higher concentration of protein while the growth phase requires lower protein concentration.

Vapor diffusion crystallization method (Fig. 2c): This technique involves having a drop containing the protein and the crystallizing agents in a concentration lower than that required for the formation of a crystal. Then the drop is equilibrated against a reservoir containing the crystallizing agent at a higher concentration. Equilibrium

is achieved by diffusion of the volatile species (water or organic solvents) out of the protein drop toward the more concentrated crystallizing agent reservoir. Consequently, the protein drop becomes more concentrated forcing the system into the supersaturation zone. Vapor diffusion can be achieved by hanging the protein drop from a sealing structure or by placing the drop on a plastic bridge [8].

Liquid diffusion crystallization method (Fig. 2d): This technique exploits the use of restricted geometries such as thin glass capillaries and microfluidic channels to force the protein solution and the crystallizing solution to mix only by diffusion. In this method, the protein solution and the crystallizing solution are placed next to each other in direct contact or separated by a physical buffer such as a permeable membrane, a gel, or by a physical barrier such as a valve. The two solutions mix into each other when the physical barrier is removed or by diffusing through the physical buffer. The liquid diffusion method has been implemented in two ways namely: *free interface diffusion* (*also known as liquid-liquid diffusion*) *and counterdiffusion*.

Counterdiffusion (Fig. 2d): In this method, the experiment starts at high values of supersaturation. Then when the protein solution and the crystallizing solution diffuse into each other, amorphous precipitate or microcrystals will form at the interphase. Immediately after, the concentration of protein in the vicinity of the precipitate decreases, but the molecules of the crystallizing agent will continue diffusing further into the protein solution. As the process continues, simultaneous events of diffusion and crystallization are created. As a consequence, a supersaturation gradient will be formed along the length of the glass capillary or microchannel. Typical concentration values for the protein solution vary between 4 and 20 mg mL^{-1}, the concentration for the crystallizing agent solution can be close to saturation. The length of the channel or capillary should be sufficiently long to allow a supersaturation gradient to be formed [10].

Free interface diffusion (Fig. 2d): The aim of this method is to achieve critical supersaturation for nucleation at a very slow rate in order to obtain a single nucleation event. Consequently, the concentrations of protein and crystallizing agent are lower than in the counterdiffusion method. The length of the capillaries of channels used in this method is usually shorter compared to the previous method. The free interface diffusion method can be viewed as a batch method conducted at very slow rate of mixing. This characteristic compared to the traditional batch method offers the advantage of a wider screening of crystallization conditions [10].

Unfortunately, protein crystallization is not a straightforward process, requiring the setting of several experiments to find clues on where the nucleation zone may lay. This is known as the screening process and consumes considerable amounts of protein. This is often problematic as many of the most interesting proteins are only available in limited quantities. Therefore, there is a demand for techniques that can obtain as much information on the crystallization conditions while consuming minimal amounts of protein. Once the nucleation zone is found, it is possible to move to the optimization process. Here the crystallization conditions found during the screening process (which do not necessarily result in a good quality crystal) are modified to obtain a large single crystal that can diffract at high resolution.

In practice, the most common way of searching for adequate crystallization conditions is by mixing the protein solution with already prepared commercial crystallization kits. Crystallization kits consist of a set of solutions which screen different crystallization parameters. Crystallization assays are usually performed in commercial plastic plates known as microplates. These plastic plates come in different formats, e.g., 24, 48, 96, 384, and 1,526 well plates. The first two formats are suitable for setting crystallization conditions manually while the later ones are used for automated liquid dispensing systems. Protein solution volumes used in each experiment can vary from nanoliters to a few microliters, depending on the protein and the purpose of the experiment, e.g., screening vs. optimization.

Nano-crystallization

In the 1990s, the development of automated liquid handling technology made it possible to dispense nanoliter volumes. This had a significant impact in a number of life science applications as it allowed multiple experiments to be performed in parallel while using reduced amounts of sample, which in many cases is expensive and difficult to produce or available in limited quantities. Protein crystallization samples presented an additional challenge to the automated liquid handling manufacturers as the solutions used vary greatly in terms of viscosity, vapor pressure, and surface tension. In addition, the solutions usually contain high concentrations of salts that can block the dispensing tips [11]. Despite these complications, some liquid handling robots can dispense sample volumes as small as 25 nL, although currently the majority of the systems dispense minimum volumes of 50 nL.

The use of automated liquid dispensing systems delivering nanoliter volumes in protein crystallization allowed screening of a vast number of crystallization conditions in a fast and reliable manner. Now in a matter of minutes, hundreds of experiments could be set up in a reproducible format [11]. Also nano-crystallization may produce better ordered crystals since convection is significantly reduced. Some authors argue that this may also result in new crystal forms that are not produced in larger volumes [12]. An additional advantage is that nanoliter drops are easier to inspect using automated visualization systems as they have less depth to focus on compared to microliter drops [11]. The implementation of nano-crystallization methods has taken two different avenues namely: microarrays and microfluidic devices.

The Use of Microarrays for Nano-crystallization

The easiest way to work with nano-volumes in protein crystallization is by using an automated liquid dispensing system to set up crystallization trials in standard 96 well crystallization plates. However, evaporation of the sample while the system was setting up the experiment became an issue. The traditional method of sealing a crystallization experiment involved sealing the complete top of the crystallization plate with an optical clear sealing tape after the robot had finished setting up the experiment. This problem could worsen if the sealing tape was not applied correctly leaving air gaps between the microplate contents and the environment. To help solve this problem in 1992 Chayen et al. described a method now known as "batch crystallization under oil" where the mix of protein and crystallizing agent solution is dispensed inside a few microliters of inert oil to prevent evaporation [13]. In 1996 D'Arcy et al. proposed the use of mixtures of different oils to allow controlled evaporation, thus effectively implementing a vapor diffusion crystallization method under oil [14].

Another way of implementing nano-crystallization using a microarray format was reported by Bodenstaff et al. in 2002 [15]. They described the fabrication of a microarray with 70 wells made of an epoxy resin; each well had a maximum volume of 250 nL. Inside these wells Bodenstaff et al. set up batch crystallization experiments using sample volumes from 1 nL to 250 nL. The dispensing of the sample was done by adapting a piezoelectric dispensing system and sealing was achieved by dispensing a layer of oil as reported previously by Chayen [13]. Bodenstaff showed that in practice, it was possible to crystallize in nanoliter volumes; however, setting up experiments in microsized arrays required a level of finesse, in practice, it is easier to set up nano-crystallization experiments in standard microplates.

The Use of Microfluidic Devices for Nano-crystallization

Microfluidic devices are ideal for screening a given protein against a vast number of crystallization conditions. These devices offer an additional advantage, that the mixing of the crystallization solution and the protein solution takes place by diffusion rather than convection,

resulting in higher quality crystals. In one way or another, all four basic crystallization methods have be implemented in a microfluidic format.

The list of reported microfluidic crystallization chips is large, as a number of the academic groups working in protein crystallization are now collaborating with groups involved in microfabrication technologies to come up with new microfluidic devices that better suit their research needs. The following is a brief review of current microfluidic devices used for protein crystallization, which is based on the crystallization method that they implement (batch, vapor diffusion, dialysis, and liquid diffusion). In some cases, a single device can be used to implement more than one crystallization method. The commercial microfluidic crystallization devices currently available to scientists are also presented.

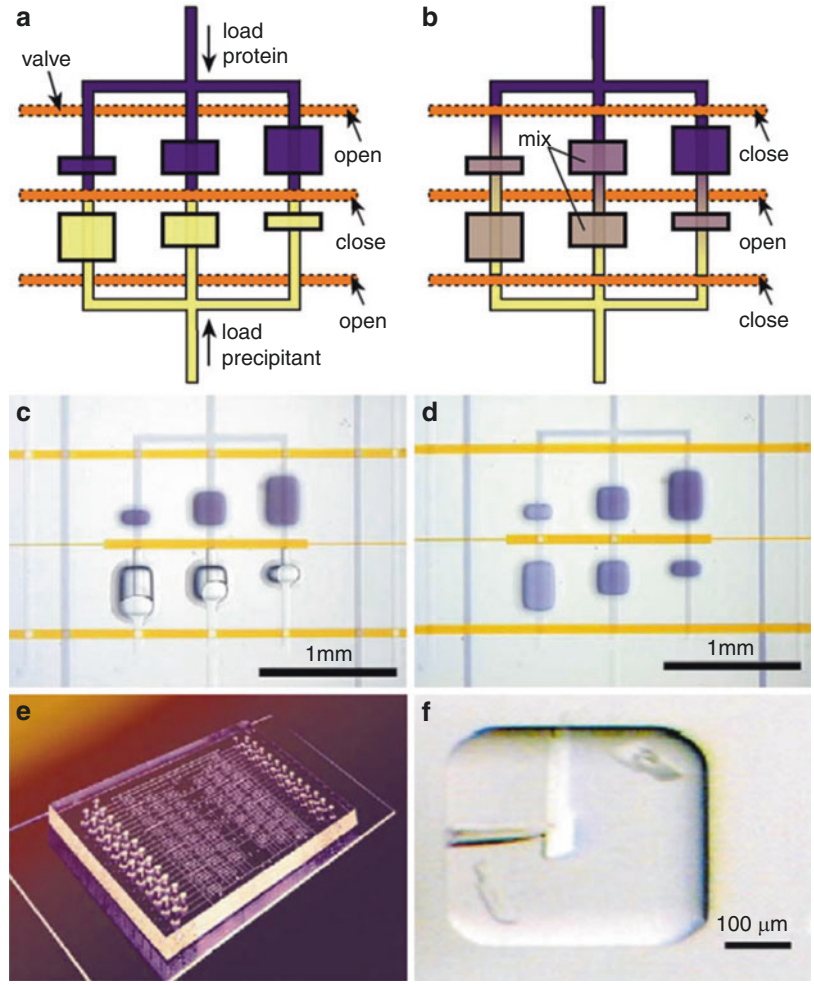
Implementation of the Liquid Diffusion Crystallization Method Using Microfluidics

The first crystallization method to be implemented in a microfluidic format was the liquid diffusion method, specifically the free interface diffusion method [16]. It could be said that the traditional counterdiffusion method in glass capillaries [10] was the beginning of crystallizing using microfluidics, since the narrow capillaries offered a reduced or free convection environment depending on the diameter used. Setting up traditional liquid diffusion experiments in glass capillaries required a reasonable level of expertise and this method was not suitable for employing automated systems in the way batch and vapor diffusion methods are. The intrinsic characteristics of the liquid diffusion methods allowed a wider screening of crystallization conditions in a single experiment. These methods were logically much more suitable for implementing in microfluidic format, which offered a good alternative to the glass capillaries [17]. Commercial versions also allowed nonexperts to set up hundreds of experiment in a fast and easy manner.

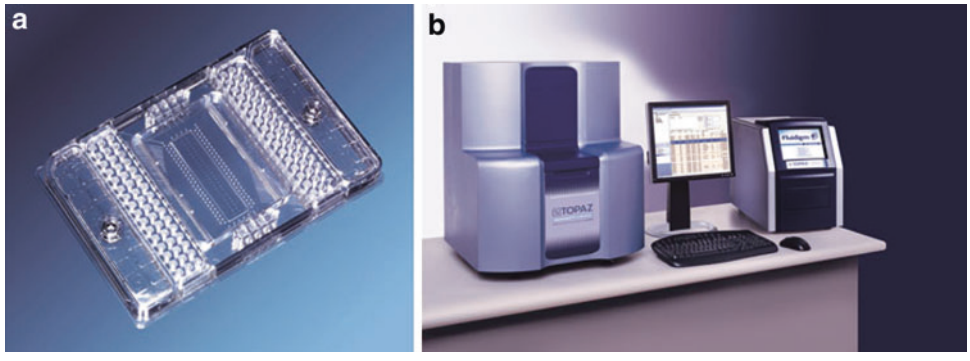
The implementation of the free interface diffusion method was presented by Hansen and Quake in 2003 [16]. Their device consisted of a glass substrate where microfluidic channels and microchambers were etched and multilayers of silicon elastomer provided flexible valves that opened and closed to allow the protein and crystallization solutions to mix at the required point. This device allowed the screening of 48 different crystallization solutions at 4 different concentrations (Fig. 3). In 2003 Fluidigm Inc., a company founded by Stephen Quake and Gajus Worthington, launched a commercial version of this device known as the TOPAZ chip (Fig. 4). Later an advanced version of this system was launched; here the glass substrate was replaced by a polymer. It combined free interface diffusion with controlled vapor diffusion, which was possible due to the gas-permeable material used in the construction of the TOPAZ chip. Fluidigm Inc., also developed software and peripheral instrumentation to facilitate liquid handling inside the TOPAZ chip as well as the inspection and analysis of the crystallization results.

Another commercial implementation of the free interface diffusion method was introduced by Microlytic Inc., where their crystallization chip, known as the Crystal Former (Fig. 5), consists of 16 channels, each 10-mm long, allowing a volume of 150 nL per channel. Each channel has two inlets at opposite ends to allow the protein solution and the crystallization solution to be loaded. The channels are covered by a thin removable foil which acts as a seal and allows extraction of the crystals that have formed.

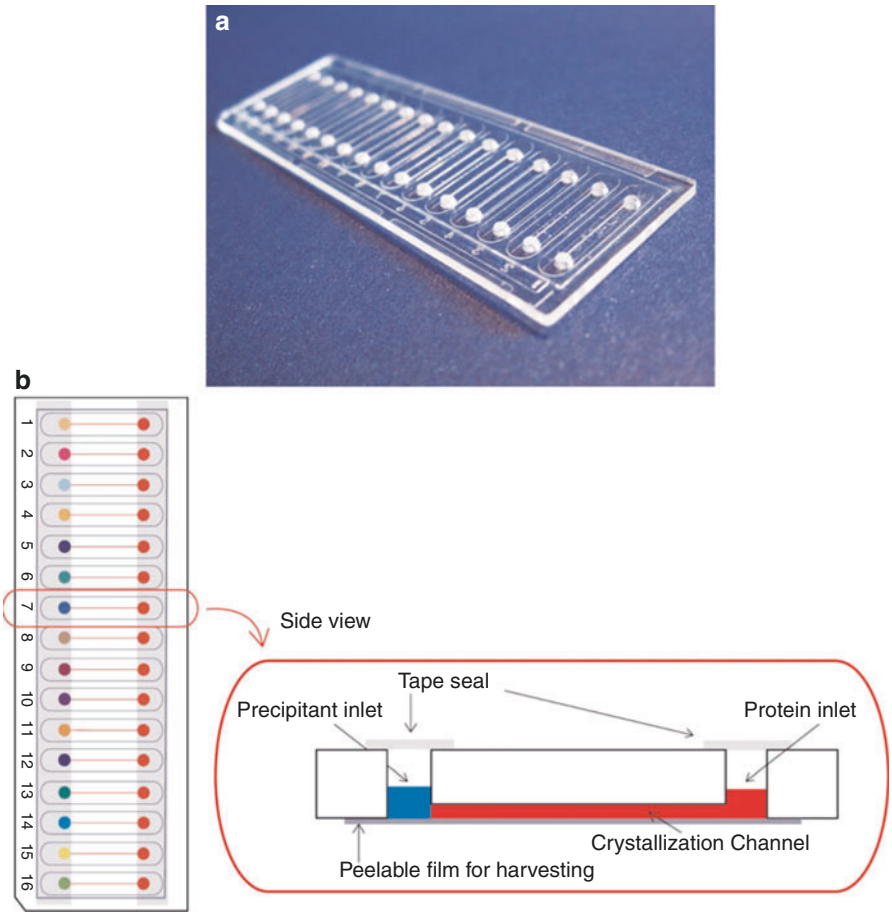
Initially commercial microfluidic chips for protein crystallization were expensive ($600 per chip); making them only affordable to large academic institutions and pharmaceutical companies. Current prices are more accessible ($50 USD per chip) to smaller laboratories. However, price and the need for microfluidic devices that fit a specific research need prompted many academic groups to develop their own chips. A recent example of this is from Dhouib and coworkers who in 2008 [18] reported the design of a microfluidic chip named the CD chip to implement the counter-diffusion crystallization method. The CD chip was later enhanced in 2009 [19]. The CD chip consists of eight microchannels with a total volume of 169 nL



Macromolecular Crystallization Using Nano-volumes, Fig. 3 Shows the free Interface diffusion (*FID*) microfluidic device designed by Hansen et al. in 2003. (**a**) Shows the loading of the protein and crystallizing solutions; (**b**) shows when the valve between the protein solution and crystallizing solution opens allowing the free interface diffusion to proceed; (**c**) and (**d**) are the micrographs of the (**a**) and (**b**) stages respectively; (**e**) shows the complete microfluidic device; and (**f**) shows a crystal of aquaporin obtained using this device (Figure and captions reprinted from Ref. [16] Copyright © 2003, with permission from Elsevier Ltd.)



Macromolecular Crystallization Using Nano-volumes, Fig. 4 Shows the complete TOPAZ system which includes (**a**) the TOPAZ chip and (**b**) the peripheral instrumentation (Images courtesy of Fluidigm Inc.)

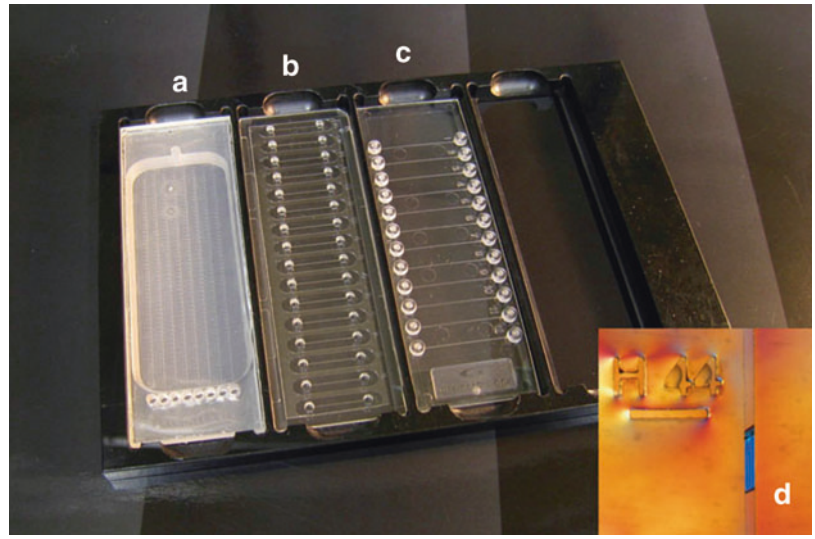


Macromolecular Crystallization Using Nano-volumes, Fig. 5 Shows (**a**) a photograph of the Crystal Former chip and (**b**) a schematic diagram of a side view of one of the microchannels (Images courtesy of Microlytic Inc.)

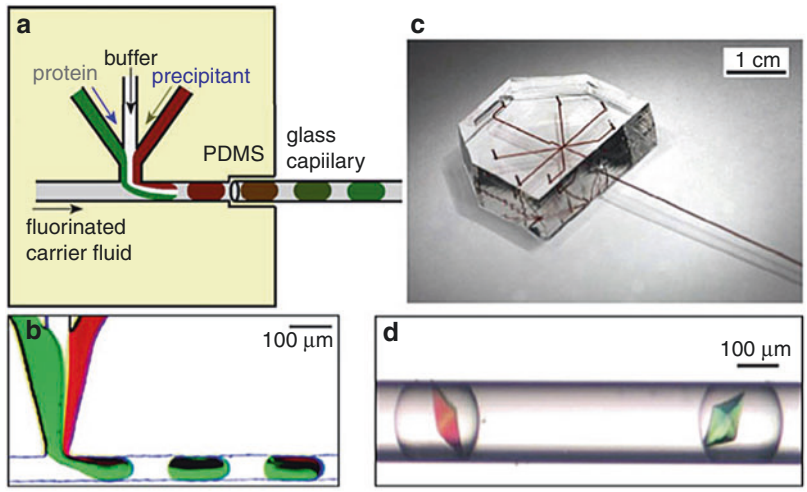
per channel. All eight channels converge at the protein solution inlet. On the opposite end of each channel another inlet is created to allow the loading of the crystallization agent. Each channel is 4.5-cm long which represented an advantage over other chips with shorter channels as the longer the channel, the larger the gradient of crystallization conditions that can be formed, hence increasing the chances of obtaining a better quality crystal. Additionally this chip was designed to be suitable for in situ X-ray diffraction analysis without the need to remove the crystals from the chip, avoiding potential physical damage. The other advantage is that a number of labels were placed along the length of the channel to facilitate locating the exact position of a given crystal when the sample is placed in the X-ray diffractometer. This device is made of cyclo-olefin copolymer (COC) which also provides good optical clarity and X-ray compatibility (Fig. 6) [19].

Implementation of Batch Crystallization Method Using Microfluidics

In 2004, Zheng et al. [20] reported a microfluidic platform for screening hundreds of crystallization conditions using less than 4 nL of protein solution and a total sample volume of 7.5 nL per crystallization experiment (Fig. 7). This device consisted of a PDMS block from which inlets and interconnecting microchannels were fabricated.

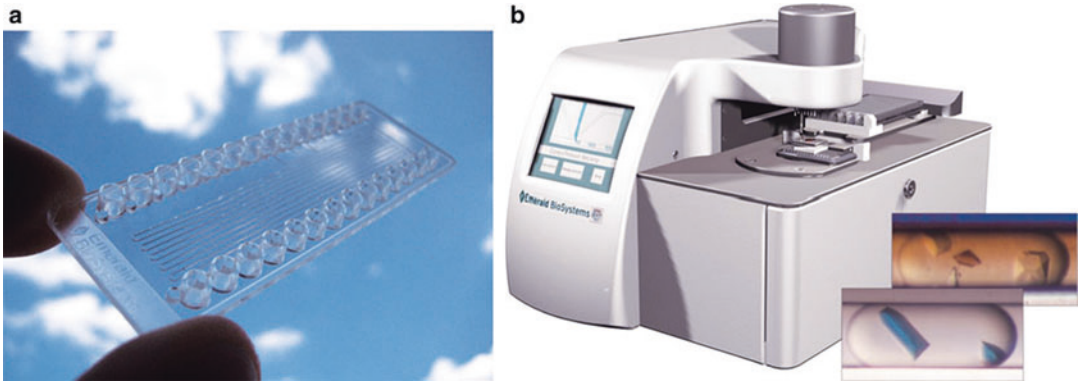


Macromolecular Crystallization Using Nano-volumes, Fig. 6 Shows (**a**) the CD chip designed by Dhouib and coworkers; (**b**) the Crystal Former chip from Microlytic Inc; (**c**) a microfluidic chip from Greiner Bio One GmbH, and (**d**) a micrograph of a thaumatin crystal grown inside one of the channels of the CD chip. It also shows a label indicating in which channel and at which position in the channel the crystal is positioned (Image courtesy of Claude Sauter)



Macromolecular Crystallization Using Nano-volumes, Fig. 7 Shows (**a**) A schematic illustration of the method of forming droplets for protein crystallization trials in a PDMS/glass capillary composite microfluidic device. (**b**) A micrograph showing the droplet formation. (**c**) A photograph of the composite device. All inlets and channels were filled with $[Fe9SCN)x]^{(3-x)+}$ solution for clarity. (**d**) A micrograph of thaumatin crystals grown inside droplets in a capillary (Figure and caption reprinted from Ref. [20] Copyright © 2004 with permission from Wiley-VCH Verlag GmbH & Co)

These channels converged to a single point where a thin glass capillary was inserted. The protein solution, buffer solution, and crystallization solutions were briefly allowed to flow through the channels into a stream of fluorinated oil (the carrier), thus creating a small drop containing a mixture of the aqueous solutions. The composition of the drops was changed by varying the flow rates of each of the aqueous streams. The flow was stopped once the glass capillary was full. The capillary was then disconnected from the PDMS block and the extremes sealed with wax. The crystals formed inside the droplets and no evaporation of the sample was observed even after 6 months of incubation without any humidity control. An advantage of this microfluidic chip



Macromolecular Crystallization Using Nano-volumes, Fig. 8 Shows (**a**) a photograph of a Crystal Card and (**b**) a photograph of the MPCS system and micrographs of protein crystals grown inside the Crystal Cards (Images courtesy of Emerald Biosystems Inc.)

was that if a good crystal was formed inside the glass capillary, the capillary could be mounted directly on the goniometer head of an X-ray diffractometer and a diffraction pattern of the protein crystal could be obtained.

In 2008, the Accelerated Technologies Center for Gene to 3D Structures created the Microcapillary Protein Crystallization System (MPCS). This is an instrument capable of generating hundreds of crystallization drops in a microfluidic format in a fast and convenient way (Fig. 8). Each crystallization drop is separated from the other by a volume or plug of oil. The main advantage of this system is its precise and finely controlled concentration gradients that can be performed over a series of drops, thus effectively creating a concentration screening for optimizing protein crystals [21].

The MPCS uses microfluidic chips known as Crystal Cards made of COC and containing two separated microfluidic channels, each channel with approximately 10 μL of useful volume; an independent crystallization experiment can be performed in each channel. The Crystal Cards together with the MPCS represented the commercial implementation of the method proposed by Zheng et al. in 2004. This also simplified the experimental process, as it was no longer required to disconnect the capillary for incubation and X-ray analysis. The Crystal Cards allow growing crystals inside the chip cards and are compatible with in situ X-ray diffraction analysis [21].

Implementation of the Vapor Diffusion Crystallization Method Using Microfluidics

The microfluidic system proposed by Zheng et al. in 2004 (Fig. 7) also allowed the implementation of the vapor diffusion method where the goal is to change the concentration of the droplet containing the protein solution by extracting water molecules from it. Zheng et al. achieved this by injecting droplets containing the mixture of the protein, buffer, and crystallization solution in a stream of water permeable oil into the glass capillary. These drops were alternated with drops containing a high concentration of salt. This procedure enabled water from the protein drop to diffuse through the water permeable oil to the drop containing a high concentration of salt until the system reached osmotic equilibrium. The ability to concentrate the protein drops allowed the crystallization process to proceed and also allowed crystals to form where the original protein and crystallization agent concentrations were too low to allow the formation of crystal nuclei.

Implementation of the Dialysis Crystallization Method Using Microfluidics

Dialysis is probably the most difficult method to implement in a microfluidic format as it requires

the careful manipulation of the liquids inside the microfluidic chip. In dialysis, the protein solution can be concentrated or diluted by allowing aqueous solution containing higher or lower concentrations of salt to be in contact with the protein solution through a water-permeable membrane. This membrane allows the passage of water molecules but not protein molecules. In 2007, Shim et al. designed a microfluidic chip known as the Phase Chip (Fig. 9) [22], which used an advanced crystal optimization method called "seeding" combined with the dialysis method.

Seeding is a crystal optimization technique that involves crushing a larger protein crystal in order to produce many microcrystals that are then used as seeds that are added to new supersaturated crystallization solutions. These seeds act as stable nucleation sites which protein molecules can use to form larger crystals under less supersaturated conditions [23].

The Phase Chip is made of PDMS and uses the method described previously by Zheng et al. to create drops of protein inside a continuous stream of oil. The Phase Chip exploits surface tension forces to guide each protein drop into wells that are located at the side of the channel, spaced 500 μm away from each other. The wells are also deeper than the flow channel. To implement the dialysis method, the bottom of each well is made of a thin layer of PDMS (15-μm thick) which allows water molecules to pass but not protein molecules. On the other side of the membrane, there is a 100-nL reservoir which can be filled with dry air, a solution with a high concentration of salt or with water. If the reservoir is filled with either dry air or salt solution, the protein solution will shrink, thus concentrating the solution, as water molecules will flow into the reservoir. Alternatively, if the reservoir is filled with pure water, the protein solution will become diluted as water molecules will now flow in the other direction.

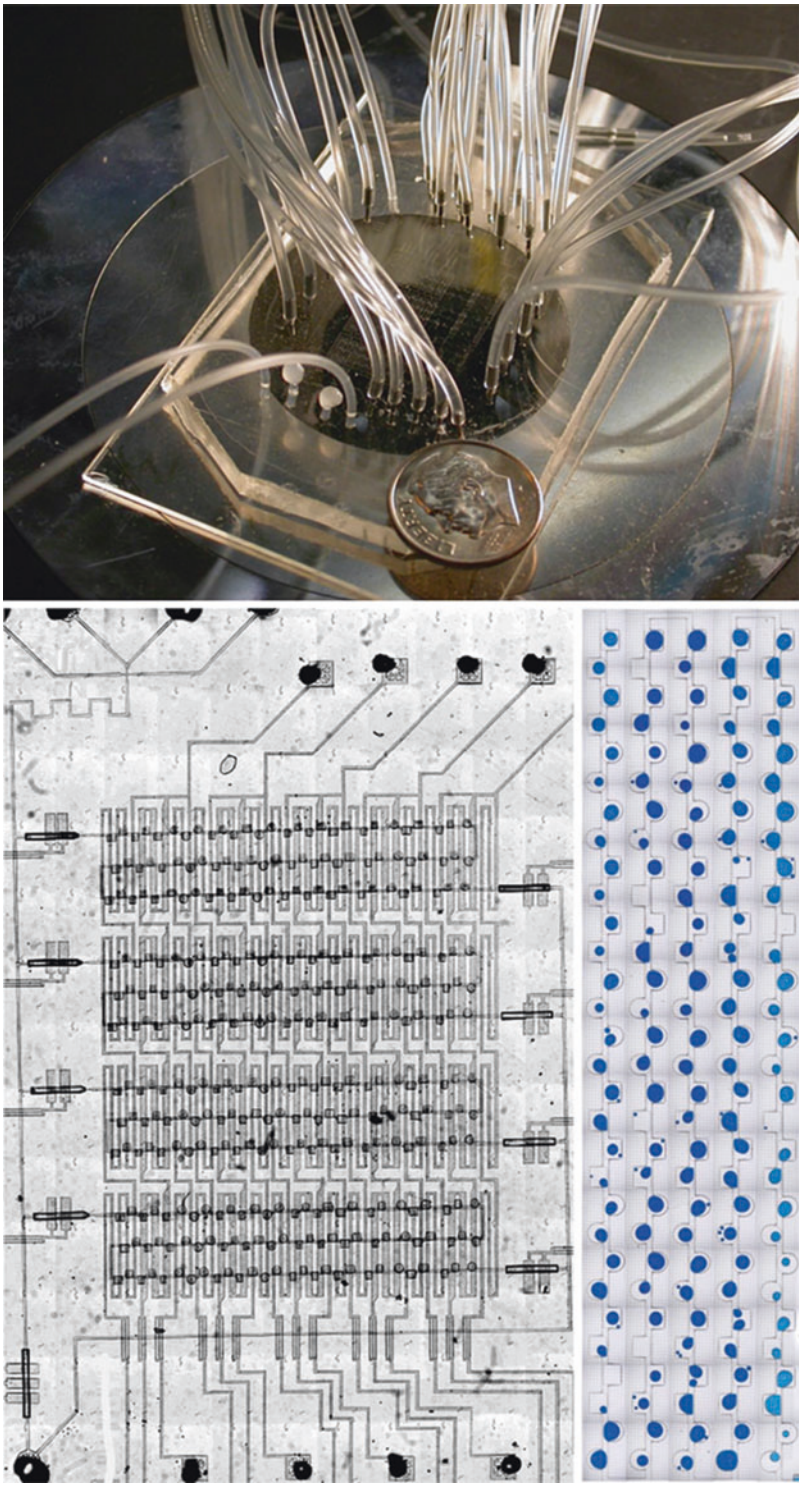
The initial goal was to fill each well with protein solution drops at different concentrations. This was achieved by changing the flow rate of the different solutions injected in the chip. Then all protein drops were concentrated by introducing dry air into the reservoir that in many cases lead to the formation of many microcrystals inside the wells. These microcrystals were considered seeds that were later transformed into large crystals by introducing pure water into the reservoir. The water molecules diffused from the reservoir to the protein well and diluted the protein solution, by doing this the smallest and least stable nuclei dissolved and only the larger microcrystals grew into even larger and higher quality crystals.

Another way to classify microfluidic devices for protein crystallization is by the type of mechanism used to create and mix the final crystallization solution. This classification was proposed by Liang and Ismagilov in 2010 [24] and classifies the devices as: *valve-based systems*, *droplet-based systems*, and *Slip Chip systems* (any combination of the above is called *hybrid systems*).

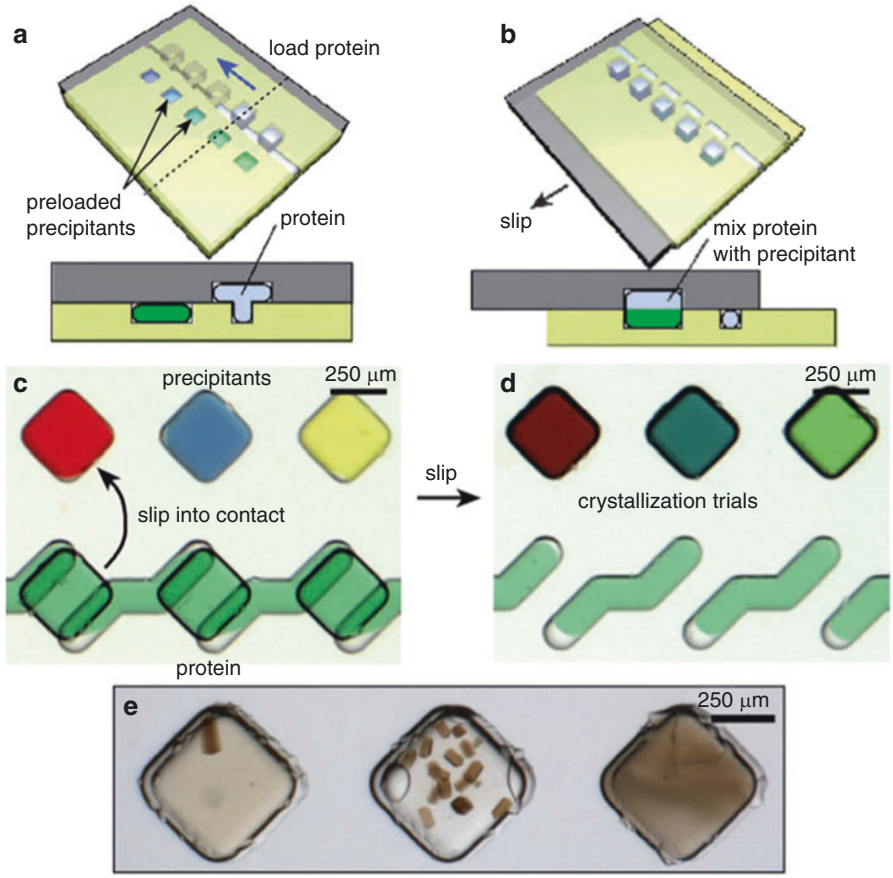
In *valve-based systems*, the protein solution and the crystallization solution are placed in different chambers within the microfluidic chip and the mixing takes place by opening a valve that connects the two chambers [24]. A good example of this is the device described by Hansen [16] and commercialized by Fluidigm Inc. as shown in Fig. 4.

In *droplet-based systems*, drops containing the protein and the crystallization agents are created by injecting these solutions into a flow of oil or other immiscible carrier fluid. Systems using this method are shown in Figs. 7 and 9. Droplet-based microfluidic systems have been useful to crystallize membrane proteins, which are traditionally difficult to work with as they are highly hydrophobic [24]. Crystallizing a membrane protein usually involves reconstituting it inside a lipidic mesophase, which is highly viscous and then incubating it against a set crystallizing solutions hoping to induce nucleation and crystal growth.

The *slip chip systems* [24, 25] consist of two different plates, each containing a number of wells. One plate is loaded with various crystallizing solutions and the other is loaded with



Macromolecular Crystallization Using Nano-volumes, Fig. 9 Shows a composed image of the Phase Chip. The *top* image shows the complete phase chip showing the inlet connections, the *bottom left* figure shows a schematic of the microfluidic circuit with the channels arranged in an "s" shape and the protein microwells arranged on the side of the channels. The *bottom right* photo shows the microwells filled with a *blue dye* (Figure reprinted from Ref. [22] Copyright © 2007 with permission from the American Chemical Society)



Macromolecular Crystallization Using Nano-volumes, Fig. 10 Shows the Slip Chip. Schematics showing (**a**) loading of protein into a Slip Chip that has already been preloaded with precipitants and (**b**) slipping to combine protein and precipitants to form crystallization trials. (**c**) Micrographs of loading a green food dye (mimicking the protein) into a Slip Chip that has already been preloaded with colored dyes (mimicking precipitants). (**d**) Micrographs of the Slip Chip after slipping to combine the solutions. (**e**) Crystals of the photosynthetic reaction center from *Blastochloris viridis* obtained using this device (Figure and caption reprinted from Ref. [25] Copyright © 2009 with permission of the Royal Society of Chemistry)

protein solutions. Then the plates slide on each other bringing both solutions into contact to form individual crystallization assays, as shown in Fig. 10. The Slip Chip represents another way of implementing the batch crystallization method and could be considered a hybrid between microarrays and microfluidic devices. The volumes of the protein solution and crystallizing solution are defined by the volumes of the corresponding wells. The wells are filled by a series of microchannels and microwells and storage under paraffin oil to prevent evaporation.

Future Trends

The crystallization process has been one of the main bottlenecks in structural biology. Now with the use of methods and technologies coming from biology, physics, biochemistry, and engineering, the screening part of the process has been improved and hundreds of crystallization conditions can be screened with small amounts of sample. This will allow us to increase our chances of finding the right crystallization conditions for a given protein or at least to find initial clues on where to start the optimization phase.

The challenge now is to focus on understanding the effect of alternative crystallization parameters that have not been sufficiently mastered and which may allow to control the crystallization process with more precision. These may include temperature, pressure, and electric and magnetic fields. Although a few academic groups have designed and built their own instrumentation to use these parameters to improve the quality of protein crystals, these instruments have not been available to the wider scientific community, with the exception of the TG40 System from Centeo Biosciences Ltd, which allows users to use temperature as an additional optimization parameter. The use of these parameters may also be useful for the crystallization of novel proteins coming from organisms that live in extreme environments such as deep-sea creatures.

Also with the incorporation of peripheral instrumentation it may be possible to analyze pre-crystallization conditions via dynamic light scattering or fluorescence techniques, allowing to know in advance if a given protein will crystallize under a given condition.

Acknowledgments The author would like to thank Phillip Dobson and Philipp Steinmann for reviewing this manuscript, to Abel Moreno, Claude Sauter, and Jose Antonio Gavira for their useful discussions and comments, and to the academic groups and companies that provided many of the images for this entry.

Cross-References

- ▶ Prenucleation Clusters
- ▶ Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures
- ▶ Structure and Stability of Protein Materials

References

1. Giege, R., Ducruix, A.: An introduction to the crystallogenesis of biological macromolecules. In: Giege, R., Ducruix, A. (eds.) Crystallization of Nucleic Acids and Proteins, pp. 1–5. Oxford University Press, New York (1999)
2. Fruton, J.S.: Molecules of Life: Historical Essay on the Interplay of Chemistry and Biology. Wiley Interscience, New York (1972)
3. Glusker, J.P., Adams, M.J.: Dorothy Crowfoot Hodgkin. Phys. Today **48**(5), 80 (1995). Obituraries
4. Goodwill, K.E., Tennant, M.G., Stevens, R.C.: High-throughput x-ray crystallography for structure-based drug design. Drug Discov. Today **6**(15), S113–S118 (2001)
5. Blundell, T.L., Jhoti, H., Abell, C.: High-throughput crystallography for lead discovery in drug design. Nat. Rev. Drug Discov. **1**, 45–54 (2002)
6. Mersmann, A., Bartosch, K.: How to predict the metastable zone width. J. Cryst. Growth **183**(1), 240–250 (1998)
7. Ries-Kautt, M., Ducruix, A.: From solution to crystals with a physic-chemical aspect. In: Giege, R., Ducruix, A. (eds.) Crystallization of Nucleic Acids and Proteins, pp. 267–312. Oxford University Press, New York (1999)
8. Chayen, N.E.: Recent advances in methodology for the crystallization of biological macromolecules. J. Cryst. Growth **198**(199), 649–655 (1999)
9. Ducruix, A., Giege, R.: Methods in crystallization. In: Giege, R., Ducruix, A. (eds.) Crystallization of Nucleic Acids and Proteins, pp. 121–148. Oxford University Press, New York (1999)
10. Otalora, F., Gavira, J.A., Ng, J.D., et al.: Counterdiffusion methods applied to protein crystallization. Prog. Biophys. Mol. Biol. **101**, 26–37 (2009)
11. Bergfors, T.: Automated liquid handling systems for submicroliter crystallization. In: Chayen, N.E. (ed.) Protein Crystallization Strategies for Structural Genomics, pp. 60–61. International University Line, La Jolla (2007)
12. Santarsiero, B.D., Yegian, D.T., Lee, C.C., et al.: An approach to rapid protein crystallization using nanodroplets. J. Appl. Crystallogr. **35**, 278–281 (2002)
13. Chayen, N.E., Stewart, P.D.S., Blow, D.M.: Microbatch crystallization under oil – a new technique allowing many small-volume crystallization trials. J. Cryst. Growth **122**(1–4), 176–180 (1992)
14. D'Arcy, A., Elmore, C., Stihle, M., et al.: A novel approach to crystallizing proteins under oil. J. Cryst. Growth **168**(1–4), 175–180 (1996)
15. Bodenstaff, R.E., Hoedemaeker, F.J., Kuil, M.E., de Vrind, H.P.M., et al.: The prospects of protein nanocrystallography. Acta Crystallogr. **D58**, 1901–1906 (2002)
16. Hansen, C.L., Quake, S.R.: Microfluidics in structural biology: smaller, faster. . .better. Curr. Opin. Struct. Biol. **13**, 538–544 (2003)
17. Sauter, C., Dhouib, K., Lorber, B.: From microfluidics to microfluidics for the crysallization of biological macromolecules. Cryst. Growth Des. **7**(11), 2247–2250 (2007)
18. Dhouib, K., Malek, C.K., Pfleging, W., et al.: Microfluidic chips for crystallization of biomacromolecules by counter-diffusion and on-chip crystal X-ray analysis. Lab Chip **9**, 1412–1421 (2009)

19. Brun, M., Dhouib, K., Morin, P., et al.: Microfluidic chips for the screening of crystallization conditions and in-situ crystallographic analyses. In: Proceedings of the 2nd European conference in microfluidics, Toulouse (2010)
20. Zheng, B., Tice, J.D., Roach, S.L., et al.: A droplet-based composite PDMS/Glass capillary microfluidic system for evaluating protein crystallization conditions by microbatch and vapor-diffusion methods with on-chip X-ray diffraction. Angew. Chem. Int. **43**, 2508–2511 (2004)
21. Gerdts, C.J., Stahl, G.L., Napuli, A., et al.: Nanovolume optimization of protein crystal growth using the microcapillary protein crystallization system. J. Appl. Crystallogr. **43**, 1078–1083 (2010)
22. Shim, J., Cristobal, G., Link, D.L., et al.: Using microfluidics to decouple nucleation and growth of protein crystals. Cryst. Growth Des. **7**(11), 2192–2194 (2007)
23. Stura, E.A.: Seeding techniques. In: Giege, R., Ducruix, A. (eds.) Crystallization of Nucleic Acids and Proteins, pp. 177–191. Oxford University Press, New York (1999)
24. Liang, L., Ismagilov, R.F.: Protein crystallization using microfluidic technologies based on valves, droplets and slip chip. Annu. Rev. Biophys. **39**, 139–158 (2010)
25. Du, W.B., Li, L., Nichols, K.P., et al.: Slip chip. Lab Chip **9**, 2286–2292 (2009)

Magnetic Assembly

▶ Magnetic-Field-Based Self-Assembly

Magnetic Dipole Emitters

▶ Optical Frequency Magnetic Dipole Transitions

Magnetic Dipole Transitions

▶ Optical Frequency Magnetic Dipole Transitions

Magnetic Electromagnetic Radiation

▶ Thermal Cancer Ablation Therapies Using Nanoparticles

Magnetic Nanoparticles

▶ Magnetic Nanoparticles for Biomedical Applications

Magnetic Nanoparticles for Biomedical Applications

Kerstin Witte[1,2], Cordula Grüttner[2], Wiktor Bodnar[1] and Eberhard Burkel[1]
[1]Institute of Physics, University of Rostock, Rostock, Germany
[2]Micromod Partikeltechnologie GmbH, Rostock, Germany

Synonyms

Biomedical applications; Contrast enhancement; Drug targeting; Hyperthermia treatment; Magnetic nanoparticles; Magnetic particle imaging; Magnetic resonance imaging; Magnetic separation; Superparamagnetism

Definition

The physical and also chemical properties of ferri- or ferromagnetic particles in the size regime of 10–1000 nm [1] vary drastically from typical bulk materials which makes them interesting for biomedical applications. In the following, a basic overview of main synthesis routes, physical properties, as well as main uses and applications in biomedicine of magnetic nanoparticles is given.

Introduction

Magnetic nanoparticles offer attractive possibilities for application in biomedicine. Living cells are in a size regime of 10 μm. Magnetic nanoparticles are smaller by three orders of magnitude, and their size is comparable to the size of

viruses, proteins, or genes, which means that they are "close to" the entity of interest [1]. When the particles are covered with specific biomolecules, they can be directed or linked to special areas in a body or in a cell. Additionally, such nanoparticles can be manipulated by an external magnetic field, which opens the possibility to direct the particles to particular regions of interest in a body. Furthermore, magnetic nanoparticles can resonantly respond to an alternating external magnetic field. This creates the possibility to transfer energy from the field to the nanoparticles.

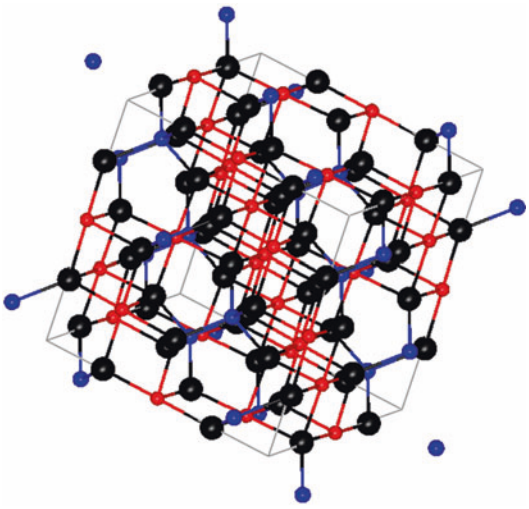
Therefore, in the recent years a lot of efforts have been done in the development of magnetic nanoparticles and the understanding of their behavior due to their unique response to an external magnetic field. Magnetic nanoparticles normally consist of a magnetic core and a shell which stabilizes the particle. The magnetic core can be arranged in two ways: as an accumulation of ferromagnetic crystallites in a nanoparticle or as randomly dispersed crystallites in a matrix shell. The shell sometimes with an additional surface coating is used in order to ensure tolerance and biocompatibility [2].

Synthesis and Stabilization

For magnetic nanoparticles, the precise control over their synthesis conditions as well as the surface functionalization is crucial, because it influences their physical and physicochemical properties and their stability and biological behavior, i.e., pharmacokinetics, toxicity, and biodistribution, drastically.

The main representatives of magnetic nanoparticles are the iron oxides, namely, magnetite (Fe_3O_4) and maghemite ($\gamma - Fe_2O_3$), due to the fact that they are biocompatible and biodegradable [3]. In magnetite, the oxygen atoms form a face-centered cubic structure. Between the oxygen atoms there are two different types of interstitial positions, tetrahedral and octahedral positions, which are occupied by iron atoms, as presented in Fig. 1.

The iron atoms on both different positions have opposite orientation of the spin. Since there are twice more iron atoms in the octahedral positions than in the tetrahedral positions, a resulting net magnetic moment is observed, leading to the ferrimagnetic behavior of this iron oxide [4]. When magnetite completely oxidizes, it is known as maghemite.



Magnetic Nanoparticles for Biomedical Applications, Fig. 1 Crystal structure of magnetite (*black*: oxygen atoms, *blue*: tetrahedral positions of iron, *red*: octahedral positions of iron)

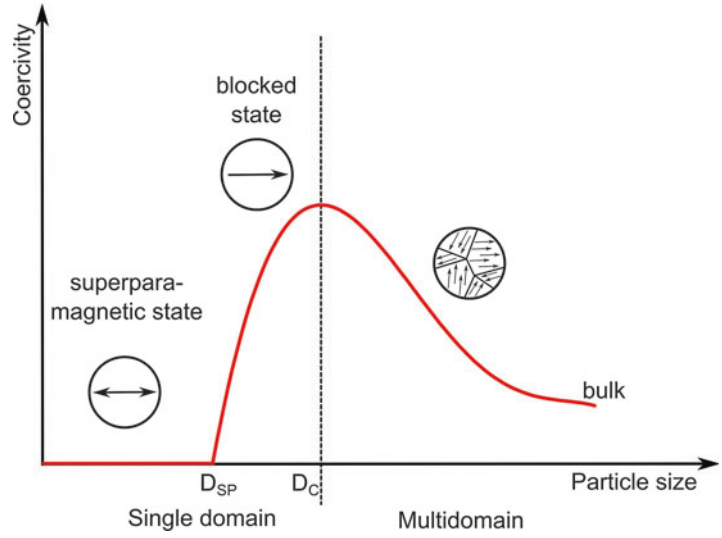
There are many routes of the synthesis of iron oxide-based magnetic nanoparticles. In general they can be classified in three groups [2]:

1. Physical methods like gas phase deposition and electron beam lithography
2. Wet chemical methods like sol-gel synthesis, oxidation methods, chemical coprecipitation, or sonochemical decomposition
3. Microbial methods

The simplest and most efficient way to obtain magnetic nanoparticles is the coprecipitation method. To fabricate magnetic iron oxide nanoparticles, a stoichiometric solution of ferrous and ferric salts is mixed in an aqueous alkaline medium [5]:

$$Fe^{2+} + 2Fe^{3+} + 8OH^- \rightarrow Fe_3O_4 + 4H_2O.$$

Magnetite tends to oxidize and an additional phase of maghemite can be formed [5]. This method



Magnetic Nanoparticles for Biomedical Applications, Fig. 2 Dependence of the coercivity on the size of a magnetic nanoparticle

generates nanoparticles with a broad particle size distribution. Synthesis of iron oxides with more uniform dimensions can be also performed in synthetic and biological nanoreactors [3].

The obtained iron oxide crystallites are extremely reactive toward water, air, as well as the living organism. This is related to the fact that they have huge specific surface which enhances their reactivity and additionally stimulates the tendency of the nanoparticles to agglomerate. To ensure a biocompatible colloid, it is usually necessary to stabilize the prepared nanoparticles by appropriate modification of their surface or by incorporation into an appropriate biocompatible matrix [5]. Several compounds with carboxylic, phosphate, and sulfate functional groups are known to bind to the surface of magnetic nanoparticles and stabilize them. Furthermore, if magnetic nanoparticles were synthesized in the presence of a stabilizing agent such as dextran, further functionalization of the surface can be necessary in order to increase the stability and to target the surface [6]. There are several coating strategies which depend on the chemical state of initial particle surface as well as the final application of the magnetic nanoparticles. In general, all strategies should have in common the achievement of hydrophilic nanoparticles, which are stable in a broad pH range and have a high ionic strength [6].

Characteristics of Magnetic Nanoparticles

The behavior of ferromagnetic material strongly depends on the domain structure of this material. Normally a ferromagnet consists of several domains separated by domain walls. When the particle size is reduced, these domains start to rearrange. A well-known example for influence of the domain structure on the macroscopic properties is the behavior of the coercivity in dependence of the particle size, as presented in Fig. 2. With the decrease of the particle size, the coercivity increases and reaches a maximum for the multidomain – single domain – transition. Below this critical size the creation of two separate domains is energetically unfavorable. With a further decrease of the particle size, the coercivity decreases drastically and reaches zero in the superparamagnetic state of the nanoparticle. Caused by thermal effects, the transition to the superparamagnetic state is not a direct phase transition.

Nanoparticles in a single domain state can be described by assuming that they exhibit an effectively uniaxial anisotropy with only one anisotropy constant, since the particles are so small, that all spins are aligned parallel [7]. The anisotropy energy can be described by

$$E = K_1 V \sin^2\theta,$$

where K_1 is the first anisotropy constant, V the particle volume, and θ the angle between the easy axis of magnetization and the magnetization vector. The anisotropy of a nanoparticle depends on several features, not only size but also shape and defects. The anisotropy energy directly depends on the volume of the particle. When the particle size decreases, the anisotropy energy can reach the same order of magnitude as the thermal energy k_BT. Due to this fact jumps between the two energetic minima are possible and the nanoparticle behaves like a paramagnetic material. The time scale of such jump can be described by the Neel relaxation time, which is an exponential function

$$\tau_N = \tau_0 \exp\left(\frac{K_1 V}{k_B T}\right),$$

where τ_0 is the Neel relaxation constant in the order of 10^{-9} s. It describes the time constant necessary for a particle to relax into an energetic minimum after a perturbation. When the Neel relaxation time of a nanoparticle is in the same order of magnitude as the measurement time, the superparamagnetic state is observed. This measurement time can be different for different measurement methods, and therefore also the transition to the superparamagnetic state can be observed for different particle sizes. When nanoparticles have the anisotropy energy significantly bigger than the thermal energy, a stable ferromagnetic state can be observed in a measurement. This state is referred to as the blocked state.

When the nanoparticles are suspended in a liquid, they are free to move and can perform physical rotations of the whole particle, which is characterized by the Brownian relaxation time [8]

$$\tau_B = \frac{3\eta V'}{k_B T},$$

where η is the viscosity of the fluid and V' the hydrodynamic volume of the particles. Since both relaxation mechanisms are possible in a solution, an effective relaxation time τ_{eff} as a linear combination of both relaxation processes can be observed [8]

$$\tau_{\text{eff}} = \frac{\tau_N \tau_B}{\tau_N + \tau_B}.$$

The relaxation behavior of nanoparticles with different diameters is dominated by different relaxation mechanisms. This is due to the fact that the nanoparticle tries to reach the equilibrium state in the shortest possible time. For small particle sizes the Neel relaxation mechanism is dominant, while for larger particle sizes the Brownian rotation mechanism is more pronounced.

There are two ways to illustrate the relaxation mechanisms – description in the time domain and in the frequency domain. Using the dynamic magnetic susceptibility, Neel and Brownian relaxation mechanisms can be identified. In an alternating magnetic field the magnetic susceptibility can be described by

$$M = \chi H = (\chi' - i\chi'')H,$$

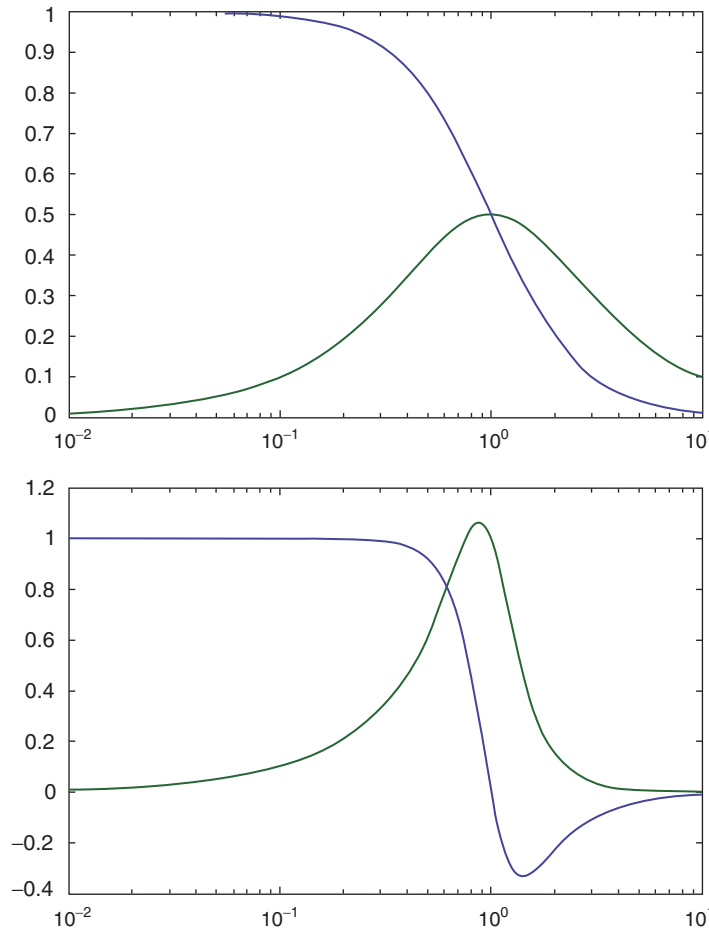
where χ' is the real and χ'' the imaginary part of the susceptibility. The behavior of the susceptibility for the Brownian relaxation and the Neel relaxation is slightly different in dependence of the frequency, as presented in Fig. 3. In both cases the resonance can be observed. In the case of Brownian motion, the resonance can be described by Debye distribution, while in the case of Neel relaxation, it can be described by Lorentz distribution.

Biomedical Applications

The unique response of magnetic nanoparticles led to the development of various applications in areas of bioscience, biotechnology, as well as environmental technology. The applications in biomedicine are mainly divided into in vitro and in vivo applications. The in vitro applications are mainly characteristic of magnetic separation procedures, while the in vivo applications are drug targeting, magnetic resonance contrast enhancement, magnetic particle imaging, and hyperthermia treatment of cancer cells.

Magnetic Separation

Magnetic separation is in biomedicine probably the most used application of magnetic



Magnetic Nanoparticles for Biomedical Applications, Fig. 3 Real part (*blue*) and imaginary part (*green*) of the susceptibility in dependence of the normalized frequency. *top* – resonance behavior in the Brownian relaxation, *bottom* – resonance behavior in Neel relaxation

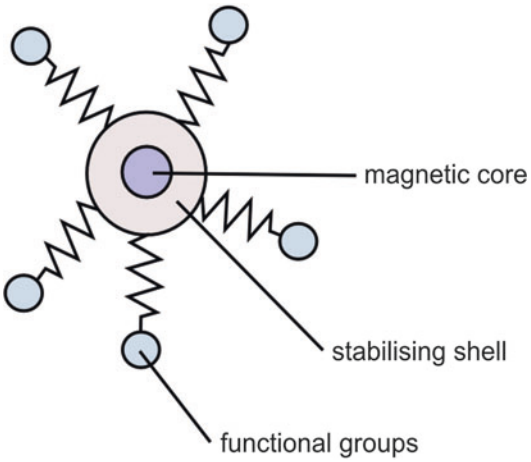
nanoparticles, since it is a simple and efficient tool for the fast isolation of target molecules, cell organelles, cells from complex biological tissues, as well as samples like blood, bone marrow, and other biological materials. The big advantage of using magnetic nanoparticles compared to other separation methods is that the biological tissue is usually pure and unaltered.

In biomedicine it is often of advantage to separate a special biological target structure from its native environment in order to achieve a concentrated sample. The magnetic nanoparticles are used as a carrier. Particles are covered with molecules which have an affinity to structure which is supposed to be isolated, as shown schematically in Fig. 4. In the incubation period the target structure is supposed to bind with the nanoparticle. Afterwards, an external magnetic field gradient is applied with an appropriate magnetic separator, and the whole magnetic complex can be rapidly isolated. After washing out the contaminants, the target structure is finally isolated [1].

Nowadays various equipment utilizing magnetic nanoparticles for the purification of DNA and RNA is commercially available. Furthermore, magnetic nanoparticles can be used in immunoassay, for example, for the detection of viruses or the assay of hormones. The use of magnetic nanoparticles for separation can reduce the duration of adsorption as well as separation steps [1, 9].

Drug Targeting

A major disadvantage of conventional chemotherapy is that it is nonspecific. Therapeutic drugs are injected intravenously, leading to a general distribution of the drug into the whole system. This also



Magnetic Nanoparticles for Biomedical Applications, Fig. 4 Schematics of a functionalized magnetic nanoparticle

affects healthy cells and tissue and not only the actual targeted tumor. The idea of using magnetic nanoparticles was developed in the 1970s and is twofold. First the systemic distribution of the drug should be reduced and with this side effects as well. Second also the required dosage should be reduced since the target is localized [10].

The physical principles of magnetic drug targeting are comparable to those of magnetic separation and are based on the magnetic force applied onto the magnetic nanoparticles via an external magnetic field. In this kind of therapy, a cytotoxic drug is attached to a magnetic nanoparticle carrier in a stable ferrofluid. When the ferrofluid is injected into the patient, the magnetic particles can be directed into the region of interest with the help of an external magnetic field gradient. Due to this fact, the drugs attached to the magnetic carrier can be concentrated at the site of the target tissue. Different ways of releasing the drug in the tumor from the carrier are possible, e.g., enzymatic activation or changes in the physiological conditions as pH value, osmolality, or temperature.

How effective such a therapy can be depends on several parameters such as the field strength, the applied gradient, or the volumetric as well as the magnetic properties of the magnetic nanoparticles used. Also parameters as the blood flow rate, the concentration of the ferrofluid, the circulation time of the nanoparticles in the blood, the position of the tumor, and the binding strength of the drug to the tumor will influence the effectiveness of such a therapy [1, 10].

Magnetic Resonance Contrast Enhancement

The magnetic resonance imaging is based on the principle of nuclear magnetic resonance of hydrogen nuclei. The target is placed in a strong homogeneous static magnetic field. Due to the magnetic field, a Zeeman splitting of the nuclear magnetic moment appears. The energy difference between the energy levels is characterized by the Larmor frequency. A transition between these two states can be achieved by applying an orthogonal high-frequency pulse with the Larmor frequency leading to a torque of the system. The precession of the magnetic moment can be described by Bloch equations, introducing two relaxation times, T_1 and T_2. The relaxation time T_1 describes the direct relaxation of the whole magnetization parallel to the external magnetic field into the energetic minimum, while the relaxation time T_2 describes the characteristic decay time of the magnetization induced in the plane orthogonal to the external magnetic field.

The images are generated by spatially encoding a nuclear magnetic resonance signal from nuclei present in an object, applying a time-varying linear magnetic field gradient. A high resolution and the ability to distinguish soft tissues are the main advantages of the magnetic resonance imaging. In some cases the contrast between different tissues is not sufficient, and additional contrast agents are required to achieve sufficient contrast in the magnetic resonance image. Compared to conventional contrast agents, nanoparticle-based contrast agents offer enhanced cellular internalization and slower clearance from tumors [11]. Furthermore conventional contrast agent can lead to severe adverse effects [6].

The function of the magnetic nanoparticles as agents for contrast enhancement can be explained considering the relaxation mechanism in hydrogen nuclei in the presence of the magnetic nanoparticles. The magnetic nanoparticles show a strong T_1 relaxation. They also introduce strong

inhomogeneities in the local magnetic field due to their high susceptibility, which also reduces the relaxation time T_2. Depending on the size of the nanoparticles, a strong influence of the T_1 or T_2 relaxation time can be observed. Smaller magnetic nanoparticles rather tend to shorten the relaxation time T_1, while bigger ones shorten relaxation time T_2, which results in dark areas in magnetic resonance images [12]. Therefore, the nanoparticles used for contrast enhancement are divided into two groups: superparamagnetic iron oxides (SPIOs) with diameters greater than 50 nm and ultrasmall SPIOs (USPIOs) with diameters smaller than 50 nm. It is worth to mention that several clinical applications in imaging of bowel, liver, or spleen have been already presented [11].

Magnetic Particle Imaging

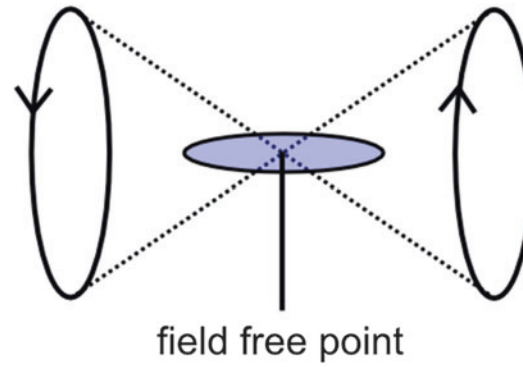
Magnetic particle imaging is a novel imaging technique first reported in 2005 [13]. This technique takes advantage of the nonlinear magnetization curves which generate harmonic response to a time-varying field. This can be detected using lock-in methods with a high degree of precision and a high signal to noise ratio [14].

The target is placed in a constant magnetic field, in which the magnetization of the magnetic nanoparticles is mainly saturated. Due to this fact there is no harmonic signal from these particles. Only a small area is kept field free, and a harmonic signal from this specific area can be detected [15]. The field free point can be generated by a Maxwell coil pair which consists of two opposing coils driven by current flowing in opposite directions, as illustrated in Fig. 5. To image a volume, the field free point has to be moved over the tested volume.

The magnetic particle imaging has a great potential for medical applications such as vascular or small intestine imaging where fast dynamic information is required and the targets are located relatively deep below the skin. This is related to the fact that the signal from magnetic particles is unaffected by intervening tissue.

Hyperthermia Treatment

Hyperthermia is based on the idea that malignant cells are more sensitive to heat than healthy tissues. Therefore an artificially induced local hyperthermia could be used to treat cancer [16].



Magnetic Nanoparticles for Biomedical Applications, Fig. 5 Schematic of the principle of magnetic particle imaging

During hyperthermia treatment magnetic nanoparticles are dispersed into the target tissue, and an external alternating magnetic field with a certain field strength and frequency is applied. This leads to a heating of the particles via Neel loss, Brown loss, or hysteresis loss [6]. Additionally, a dissipation of the heat to the surrounding tissue itself takes place. When the temperature is kept above a therapeutic threshold of 42 °C for 30 min or more, cancer tissues can be destroyed. The big advantage of using magnetic nanoparticles for such an application is that only the targeted tissue is heated while the other part of the body stays unaffected [10]. Furthermore, the particles can be functionalized by coupling to antibodies in order to enhance the uptake [17].

The main parameter determining the heating of the tissue is the specific absorption rate (SAR). It is defined as the rate at which electromagnetic energy is absorbed by a unit mass of a biological material and is proportional to the temperature increase [17]

$$\mathrm{SAR} = C_e \frac{dT}{dt},$$

where C_e is the specific heat capacity of the sample, T the temperature, and t the time. The deposition of heat into the particles by an external field is proportional to the SAR and can be described

by the frequency f, the imaginary part of the susceptibility of the magnetic nanoparticles χ'', and the square of the magnetic field strength H^2

$$P = \mu_0 \pi f \chi'' H^2.$$

This means that basically the deposition of heat into the target should be resonant for a certain magnetic field strength, and it strongly depends on both relaxation times of the nanoparticles. Still the product of frequency and applied field strength is limited for a certain coil diameter. This limitation shall ensure that a person can withstand a treatment for more than 1 h without major discomfort [17]. Due to this fact and that the resonance frequency of magnetic nanoparticles strongly depends on their size, shape, and morphology, the resonant excitation of the nanoparticles is a challenging task.

Summary and Future Prospects

The vast majority of magnetic nanoparticles used in biomedicine consists of magnetite and/or maghemite crystallites covered with a matrix shell which stabilizes them. The nanoparticles can be produced via several routes and can be additionally functionalized by adding functional groups to the surface of the matrix shell. The properties of the nanoparticles strongly depend on their morphology, size, and shape. Especially, the magnetic anisotropy of the nanoparticles is affected to a great extent by the particle size. Two kinds of relaxation phenomena can be observed for the magnetic nanoparticles, namely, the Brownian and the Neel relaxation. In an alternating external magnetic field, the relaxation of the particles can be induced, and the particles can be excited resonantly at the resonance frequencies which depend on their relaxation time.

The unique properties of magnetic nanoparticles make them an ideal tool in biomedicine. The nanoparticles functionalized with different biomolecules can be used for magnetic separation of proteins, DNA/RNA, or cells as well as for drug targeting in chemotherapies applying magnetic field gradients to localize the magnetic nanoparticles.

The particles can be used as well as the contrast enhancement medium in magnetic resonance imaging where the particles are mainly used as negative contrast agents. A novel imaging technique is the magnetic particle imaging. In comparison to the magnetic resonance imaging, the particles do not lead to a negative contrast, but to a positive contrast, where the measured signal is unaffected by intervening tissue. Furthermore, due to an alternating magnetic field, relaxation processes are induced in the nanoparticles, which leads to a heating of the particles placed inside a body. This way a hyperthermia treatment of localized cancer cells can be performed.

Presented here are the main applications of magnetic nanoparticles in biomedicine. Many different applications of magnetic nanoparticles in tissue engineering or biosensing are developing [10].

Beside the magnetic separation and the magnetic resonance contrast enhancement, the application of magnetic nanoparticles in biomedicine is not market-based yet. Drug delivery already undergoes human trials, while the hyperthermia treatment of cancer was already shown to be effective in tests on animals. Since the application of the nanoparticles is affected by numerous factors, the transfer of the hypothetical process from laboratory conditions to an actual process performed in real conditions is still a challenging task.

Cross-References

- ▶ Drug Delivery and Encapsulation
- ▶ Dye-Doped Nanoparticles in Biomedical Diagnostics
- ▶ Genotoxicity of Nanoparticles
- ▶ Magnetic Assembly
- ▶ Magnetic Nanostructures and Spintronics
- ▶ Magnetic Self-Assembly
- ▶ Nanobiosensors
- ▶ Nanomagnetism
- ▶ Nanoscale Particle
- ▶ Nanostructures for Surface Functionalization and Surface Properties

References

1. Gubin, S.P. (ed.): Magnetic Nanoparticles. Wiley-VCH Weinheim, Weinheim (2009)
2. Reddy, L.H., Arias, J.L., Nicolas, J., Couvreur, P.: Magnetic nanoparticles: design and characterization, toxicity and biocompatibility, pharmaceutical and biomedical applications. Chem. Rev. **112**, 5818–5878 (2012)
3. Laurent, S., Forge, D., Port, M., Roch, A., Robic, C., Elst, L.V., Muller, R.N.: Magnetic iron oxide nanoparticles: synthesis, stabilization, vetorization, physicochemical characterizations, and biological applications. Chem. Rev. **108**, 20064–2110 (2008)
4. Cox P.A.: Transition metal oxides: An introduction to their electronic structure and properties. In: Rowlinson J.S., Halpern J., Green M.L.H., Mukaiyama T. (eds.) Oxford University Press, New York, USA, preprint 2010. (1992)
5. Safarik, I., Horska, K., Safarikova, M.: Magnetic nanoparticles for biomedicine. In: Intracellular Delivery, pp. 363–372. Springer, Netherlands (2011)
6. Colombo, M., Carregal-Romero, S., Casula, M.F., Gutierrez, L., Morales, M.P., Böhm, I.B., Heverhagen, J.T., Prosperi, D., Parak, W.J.: Biological applications of magnetic nanoparticles. Chem. Soc. Rev. **41**, 4306–4334 (2012)
7. Skomski, R.: Simple Models of Magnetism. Oxford University Press, New York, USA (2008)
8. Astalan, A.P., Ahrentorp, F., Johansson, C., Larsson, K., Krozer, A.: Biomolecular reactions studied using changes in Brownian rotation dynamics of magnetic particles. Biosens. Bioelectron. **19**, 945–951 (2004)
9. Ito, A., Shinkai, M., Honda, H., Kobayashi, T.: Medical application of functionalized magnetic nanoparticles. J. Biosci. Bioeng. **100**, 1–11 (2005)
10. Pankhurst, Q., Conolly, J., Jones, S.K., Dobson, J.: Application of magnetic nanoparticles in biomedicine. J. Phys. D Appl. Phys. **36**, R167–R181 (2003)
11. Berry, C.C.: Progress in functionalization of magnetic nanoparticles for applications in biomedicine. J. Phys. D Appl. Phys. **42**, 224003 (2009) (9pp)
12. Geraldes, C.F., Laurent, S.: Classification and basic properties of contrast agents for magnetic resonance imaging. Contrast Media Mol. Imaging **4**, 1–23 (2009)
13. Gleich, B., Weizenecker, R.: Tomographic imaging using nonlinear response of magnetic nanoparticles. Nature **435**, 1214–1217 (2005)
14. Pankhurst, Q.A., Thanh, N.K.T., Jones, S.K., Dobson, J.: Progress in applications of magnetic nanoparticles in biomedicine. J. Phys. D Appl. Phys. **42**, 224001 (2009) (15pp)
15. Buzug, T.M., Borgert, J. (eds.): Magnetic Particle Imaging: A Novel SPIO Nanoparticle Imaging Technique, 1st edn. Springer, Heidelberg/Berlin (2012)
16. Gilchrist, R.K., Medal, R., Shorey, W.D., Hanselman, R.C., Parrott, J.C., Taylor, C.B.: Selective inductive heating of lymph nodes. Ann. Surg. **146**, 596–606 (1957)
17. Laurent, S., Dutz, S., Häfeli, U.O., Mahmoudi, M.: Magnetic fluid hyperthermia: focus on superparamagnetic iron oxide nanoparticles. Adv. Colloid Interface Sci. **166**, 8–23 (2011)

Magnetic Nanostructures and Spintronics

Alessandro Chiolerio[1] and Paolo Allia[2]
[1]Physics Department, Politecnico di Torino, Torino, Italy
[2]Materials Science and Chemical Engineering Department, Politecnico di Torino, Torino, Italy

Synonyms

Nanomagnetism; Spintronic devices

Definition

Nanostructures featuring magnetic properties and devices for the elaboration and/or storage of information based on electron spin transport.

The need of storing and elaborating binary information permeates everyday life. Today's computers are mainly charge-transfer devices; this means that inside a processing unit the information chain corresponds to a physical electron (hole) drift current, characterized by diffusion lengths and flow times that strongly limit the evolution toward faster computations and smaller sizes. Codifying binary information into a spin-polarized current could allow much faster logic operations, due to the possibility of reversing spins without stopping the carrier flow. The main difficulty of implementing spin-electronic ("spintronic") devices is that typical spin diffusion lengths are well below the hundredth of nanometer for diamagnetic metals and between some tenths of nm and some nms for ferromagnetic metals and alloys, meaning that the typical device must be nanostructured. Furthermore, since spin-flip scattering must be prevented to process

uncorrupted data, both interface quality and material purity must be extremely controlled. High and ultra-high vacuum deposition equipments are needed, and perfect control of geometries down to the nm scale requires extremely sophisticated lithographic steps. Such technological issues have slowed down the challenge of bringing spintronic devices into the core of our logic devices. On the contrary, commercial data storage still relies almost entirely on magnetic nanostructures: both granular media for hard disk drives and tunneling magnetic junction-based read/write heads are worldwide diffused.

Due to their technological relevance, the most important effects on which commercial spintronic devices are based will be considered: Giant Magnetoresistance, Tunneling Magnetoresistance and Spin Torque. Two more subsections are addressed to particularly hot topics: Semiconductor Spintronics and Data Storage. In all these cases, physical systems and/or devices are authentic nanostructures, that is, artificial, nanosized structures characterized by carefully controlled geometries in spite of their reduced dimensions. On the other hand, magnetic nanoparticles, although increasingly gaining attention because of their potential applications in different fields (ranging from ICTs to biomedicine) are excluded from this entry because the synthesis processes, even when characterized by extremely narrow size distributions, do not allow complete control on the assembly geometries, not (yet) resulting in the fabrication of bonafide ordered nanostructures. A subsection related to magnetic anisotropies at the nanoscale represents an essential premise to the rest of the work.

Magnetic Anisotropies at the Nanoscale

The magnetic properties of various classes of nanomaterials and nanostructures are strongly influenced by many types of magnetic anisotropy which determine their behavior. As a general rule, it can safely be stated that all anisotropies existing in bulk, macroscopic material play a role at the nanoscale [1]; in addition, novel types of anisotropy, typical of nanostructured materials, appear. A brief overview of the different magnetic anisotropies and their effects follows.

Shape Anisotropy

Shape anisotropy, emerging from the requirement that the magnetostatic energy of the magnetic system must be a minimum, effectively dictates the direction of the local magnetization vector in ultrathin films of a ferromagnetic material and in non-spherical magnetic nanoparticles. In thin films of a ferromagnetic material the direction of the magnetization vector is preferentially forced to stay parallel to the film plane by shape anisotropy. This property is widely exploited in spintronic devices. For similar reasons, in elongated (either ellipsoidic or acicular) nanoparticles, the zero-field magnetization stays aligned to the major axis. This anisotropy determines the coercivity of an assembly of randomly oriented magnetic nanoparticles, and plays, therefore, a central role in the magnetic response of particulate media for high-density magnetic recording and hybrid materials containing nanoparticles embedded in a diamagnetic, organic matrix such as a polymer. Finally, shape anisotropy is fundamental for patterned media also determining the local direction of magnetization and the magnetic domain structure within a magnetic nanodot or nanopillar, consequently influencing the magnetization reversal mechanism and its duration.

Magnetocrystalline Anisotropy

Magnetocrystalline anisotropy may concur in determining the local magnetization direction of magnetic nanoparticles; however, its effects often emerge only when the shape anisotropy is negligible, that is, for almost perfectly circular/spherical nanoparticles. One of the most important examples of the role played by magnetocrystalline anisotropy is provided by FePt thin films, which have been largely investigated owing to their potential as novel media for high-density perpendicular magnetic recording. In these films, shape anisotropy which would impose a parallel-to-the plane magnetization state is largely overcome by a very strong magnetocrystalline anisotropy term, such that the easy axis is pointing perpendicular to the plane. This term results from the development

of a tetragonal distorted phase (L_{10}) which is obtained by submitting cubic FePt to controlled heating.

Stress (Magnetoelastic) Anisotropy

In nanostructured materials obtained by means of deposition or growth techniques followed by lithographic process, the concentration of local stress fields can reach high levels. The magnetization direction of a thin ferromagnetic film or patterned medium can be affected by stress through magnetoelastic (magnetostrictive) coupling. The resulting stress anisotropy often plays a detrimental role on the magnetic response of a thin ferromagnetic film, because it acts to break the symmetry of the system, generating local disturbances in the magnetization pattern of the material and acting to reduce its performance.

Surface/Interface Anisotropy

In thin ferromagnetic films, specific additional forms of magnetic anisotropy exist [1]. Surface anisotropy may play a relevant role because it can be one order of magnitude larger than typical bulk anisotropy values. Surface anisotropy of transition-metal elements or alloys can be understood in terms of the different band filling of 3D states for magnetic atoms at a surface/interface, and can be either perpendicular or parallel to the film plane in dependence of the band filling level, that is, in dependence of the transition metal considered. For symmetry breaking reasons, some of the five 3D states of transition-metal atoms at a surface/interface (i.e., those whose wavefunctions extend in a direction perpendicular to the surface plane) take quite different energies with respect to bulk atoms, while other 3D states (those mainly extending along the surface plane) have almost identical energies in the bulk and at the surface. Interface anisotropy is important in magnetic multilayers also, because of the hybridization of electron states of the adjacent metals (one typically being Fe, Co, Ni; the other one being either a noble metal or an early (or late) transition element). A variety of effects emerges as a consequence of such hybridization, including strong perpendicular anisotropy.

Interlayer Exchange Anisotropy

The transmission of a magnetic anisotropy through an interface between two magnetic materials – that is, the occurrence of a common axis of easy magnetization shared by the two materials across the interface – is a peculiar effect of layered magnetic films and plays a central role in the optimization of the properties of a number of spintronic devices, including the spin valve [2]. As known from the physics of bulk magnetic systems, long range order (ferromagnetic, antiferromagnetic, ferrimagnetic) is assured by the exchange interaction between pairs of adjacent spins. Exchange interaction, a purely quantum effect, is particularly strong between spins belonging to the nearest neighbor atoms, and goes rapidly to zero with increasing inter-spin distance. When the interface between two different magnetically ordered metals is sharp, spin information can be transmitted from one material to the other and vice versa. In particular, if one material is strongly anisotropic, it can influence the anisotropy of the adjacent layer through exchange coupling. The most interesting case from the application side arises by placing together a ferromagnet and an antiferromagnet. In this case, the antiferromagnetic layer (it can be a transition metal oxide of a transition metal, such as NiO or CoO), although acting as a definite source of magnetic anisotropy for the ferromagnetic layer, is magnetically "dead," that is, not contributing at all to the magnetic response of the device in the range of magnetic fields of interest. Such a feature is exploited, for example, in spin valves, where the magnetization of one of the two ferromagnetic layers must remain fixed in space ("pinned") while that of the other layer is free to reverse under the action of the same weak magnetic field. The pinning force is typically provided by the exchange anisotropy between an antiferromagnetic layer in close contact with the pinned ferromagnetic layer [3].

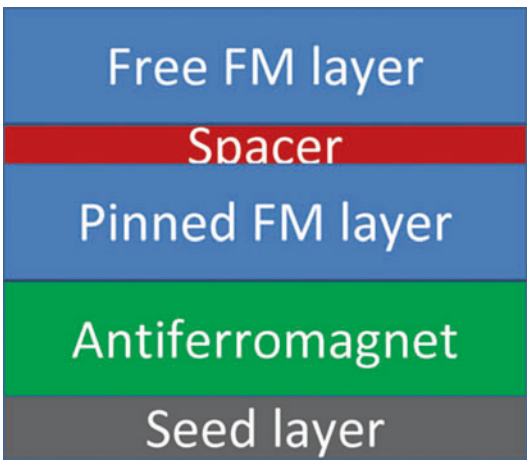
Exchange Bias

Exchange bias [4] is strictly related to exchange anisotropy, even though resulting in a different variety of applications. The effect, discovered in 1956 – that is, well before the rise of interest

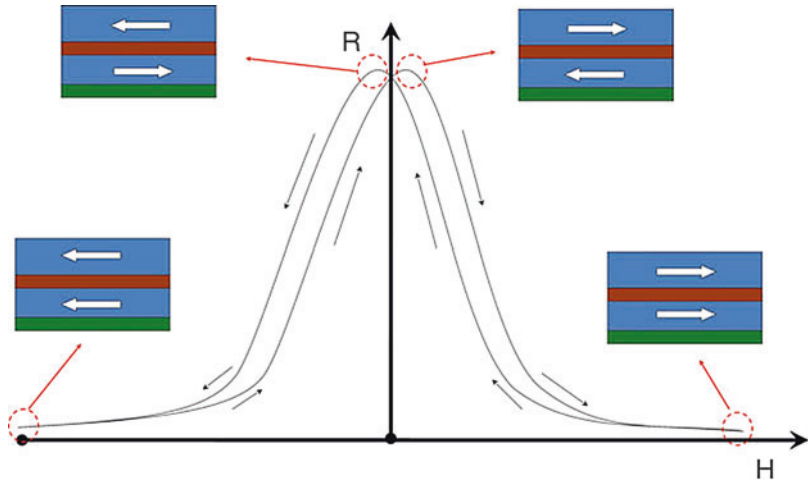
toward nanomagnetism and spintronics – is again associated to the exchange anisotropy between an antiferromagnetic and a ferromagnetic layer. After proper thermal pretreatment (e.g., field cooling through the ordering [Néel] temperature of the antiferromagnet), the hysteresis loop of the FM material is permanently shifted along the field axis. This brings about a different absolute value of the material's coercivity under decreasing and increasing fields. The mean coercivity can also increase. Possible applications of novel exchange-biased bi- or multilayers are permanent magnets and magnetic recording media. Moreover, the loop shift induced by exchange bias can act to reduce the field needed to observe giant magnetoresistance effects in spin valves or magnetic multilayers with respect to the field needed in unbiased systems.

Giant Magnetoresistance

In 1975, experiments on a Fe/GeO/Co junction led to the discovery of tunneling magnetoresistance (TMR) by M. Jullière, which was again verified in 1995 by T. Miyazaki, N. Tezuka, and J. S. Moodera. This great delay was caused by the fact that the experiment was done at cryogenic temperatures; the interest on the phenomenon was mainly academic because commercial electronics was and is still now unable to operate at such low temperatures, where the carriers in a semiconductor are typically freezed. So in 1988, experiments on layered thin films of ferromagnets alternated to a nonmagnetic metal led to the simultaneous and independent discovery of the giant magnetoresistance (GMR) by A. Fert and P. A. Grünberg, who were awarded for this discovery with the Nobel prize in 2007. They found that submitting those layered structures to a current perpendicular to the interfaces, the conduction electrons were submitted to a spin-dependent scattering whose probability depends on the relative alignment between the macroscopic magnetization of the two ferromagnetic thin films. In case they are parallel the scattering probability is low, hence the electrical resistance is low too. Typical materials used to realize a GMR stack are thin films of Co/Cu, Fe/Cr, Co/Ag, Co/Ru, $Co_{90}Fe_{10}$/Cu; a typical structure is shown in Fig. 1. When applying an external magnetic field, if the two ferromagnetic films feature different coercivities, it is possible to act on the orientation of the magnetically softer one and produce a high resistance state. Different coercivities are obtained either realizing films of different thicknesses or coupling one of them, the so-called pinned layer, to an antiferromagnet by exchange bias. Between the materials used to implement exchange bias, selected ones are NiO, FeMn, Fe_2O_3, $Mn_{76}Ir_{24}$, and MnPt. Generally the highest values of GMR, defined as total resistance variation normalized by the resistance in case of parallel alignment times 100, are obtained for epitaxial thin films working at cryogenic temperatures: GMR between 100 % and 200 % were measured at temperatures lower than 5 K, while at room temperature typical values approach 50 % at maximum. GMR systems are commercially used as magnetic field sensors, isolators and hard-disk read heads [5]. The typical electrical response is presented in Fig. 2.



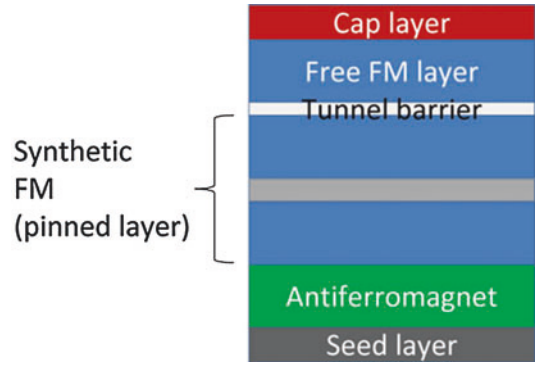
Magnetic Nanostructures and Spintronics, Fig. 1 Section of the nanostructured stack of *bottom* spin-valve featuring an antiferromagnetic layer providing exchange bias to the pinned layer and including a cap layer to prevent oxidation



Magnetic Nanostructures and Spintronics, Fig. 2 Typical giant magnetoresistance curve with sketches indicating the parallel/antiparallel configuration of magnetization in the free (*topmost blue layer*) and pinned layer (*bottom blue layer* close to the *green* antiferromagnet)

Tunneling Magnetoresistance

As already stated, TMR was observed before GMR and is still under intensive study, due to extreme industrial interests; devices showing TMR are called magnetic tunneling junctions (MTJs) and, if compared to more classical all-metal devices (such as GMR), are characterized by a high impedance, representing the optimal choice for electronic drivers operating at rather high frequencies. This is of particular interest for at least two applications: the elements of digital nonvolatile memories and the stray field sensors used in TMR read heads for high-density storage media. A TMR device is composed by a stack of two ferromagnets, one soft and one hard (see Fig. 3) separated by an insulator; in analogy to spin-valves they have a hysteretic three-state electrical response and could act as magnetically operated switches. Generally the materials used for soft electrodes are FeNi, CoFeB, or a thick layer of Co, Fe, or Ni; for the hard electrodes CoFe, CoCr, or an extremely thin layer of Co, Fe, or Ni are used; the tunnel barrier may be realized in Al_2O_3, MgO (best performances), and many other hybrid materials such as Ta_2O_5, $SrTiO_3$, and NiO [6]. Since the best electrical response is given by a single domain structure (Fig. 4), each MTJ should be nanostructured, for example, as a nanopillar with top and bottom contacts. Furthermore monocrystalline epitaxial materials, although displaying many technological problems, provide the best performances giving TMR values as high as

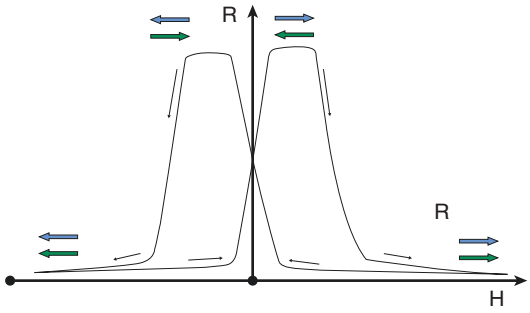


Magnetic Nanostructures and Spintronics, Fig. 3 Structure of a nanoengineered stack of a magnetic tunneling junction, comprehensive of a synthetic ferromagnetic pinned layer and a cap layer to prevent oxidation

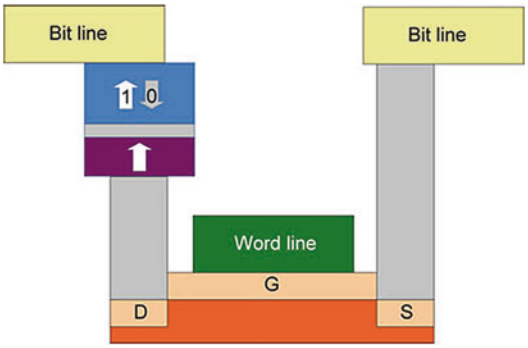
15,000 %, versus typical TMR of sputtered devices below 1,000 %. Since monocrystalline materials require extremely expensive equipments such as molecular beam epitaxy, engineers found that also sputtered materials could guarantee reasonable performances when *texturized*, that is, thermally treated in order to undergo solid state epitaxy, consisting in a pseudo-conformal growth at the interfaces between insulating barrier and ferromagnetic electrodes.

Spin-Torque Effects and Devices

Most spintronic effects have their own opposites; this circumstance engenders a variety of novel



Magnetic Nanostructures and Spintronics, Fig. 4 Typical tunneling magnetoresistance curve with a representation of the magnetization configuration in the free layer (*blue arrow*) and in the pinned one (*green arrow*)



Magnetic Nanostructures and Spintronics, Fig. 5 Structure of a spin transfer magnetic RAM, seen in section. The bit and word line connections address a single memory cell, directly switching on or off the control transistor; when it is switched on, the current emitted by the drain (*D*) transfers the perpendicular magnetization of the hard layer (*purple*) to the free one (*blue*), writing data. Reading is done by lowering the voltage on the gate (*G*) and measuring the resistance across the two bit lines, which depends on the parallel/antiparallel configuration of the MTJ electrodes

potential applications. As known, the giant magnetoresistance effect is the result of a modulation of a spin-polarized current by spin-dependent scattering. In a spin valve, the direction taken by the magnetization vector in one ferromagnetic layer can modulate the current across the device, depending on the direction taken by the magnetization of other ferromagnetic layer, that is, the magnetization affects the current.

In magnetic nanostructures, the opposite effect is observed: a strongly spin-polarized electrical current applies a torque on the magnetic moment of a ferromagnetic layer initially pointing in a different direction [7, 8]. This effect is called "spin torque" or spin-transfer switching and is exploited in various nanodevices. A typical structure resembles the spin valve: two thin magnetic layers (pinned and free) are separated by a diamagnetic spacer. When an electrical current passes through the first layer, the spin of the electrons becomes oriented in the direction of that layer. This polarized current then passes through the second layer and again the electron spins align to the direction of the magnetic moment. However, because angular momentum is conserved, the difference in the orientation of the electrons caused by the magnetic layers is seen as a torque. On the contrary, if the current is first passed through the free layer, the fixed layer works as a filter, allowing only the component of parallel magnetic moments carried by the current, and reflecting back toward the free layer the reminder; therefore, a torque again acts on the free layer. It can be easily shown that a torque can be applied to the free layer in different directions. Therefore, the torque can either force the magnetization of the free layer toward or away from the direction of the fixed layer magnetization. If the torque is large enough, it can lead to switching of the magnetic moment of the free layer. Magnetization dynamics is described by the Landau–Lifshitz–Gilbert equation, whose terms represent spin precession, magnetization damping, and spin transfer torque [9]. Depending on the direction and strength of the polarized current, the magnetization of the free layer will end either parallel to the applied field, or antiparallel to the magnetization of the fixed layer, or even will indefinitely precess around the applied field direction in steady state.

Spin torque effects are being investigated in view of their potential applications as alternative methods to efficiently write data in nonvolatile MRAMs exploiting smaller currents with respect to standard writing techniques involving transient magnetic fields [10]. ST-MRAMs are characterized by notable write endurance, good compatibility with standard CMOS processing and fast operation (Fig. 5). In fact, spin-transfer-driven switching of magnetization can take 1 ns or less.

The same effect can also be exploited in the design of fast electronic devices such as microwave (RF) oscillators microwave oscillators, converters, microwave sources and antennas. These devices are interesting because of their small size, low power consumption and good response to frequency modulation and phase locking.

However, the current density needed to achieve fast magnetization reversal is often too high to ensure a reasonable lifetime of a spin-torque device (of the order of one decade for a MRAM's operation time). Another major issue is the control over all parameters entering the Landau–Lifshitz–Gilbert equation in order to optimize the efficiency of the device.

Recently, spin torque effects and related magnetization switching have attracted much attention in connection with the development of spintronic memristors at the nanoscale. In memristors, the resistance is a function of the history of the current through and voltage across the device. Spintronic sensing elements based upon spin torque effects can be designed to explore and memorize continuously varying current and voltage applied to the device [11].

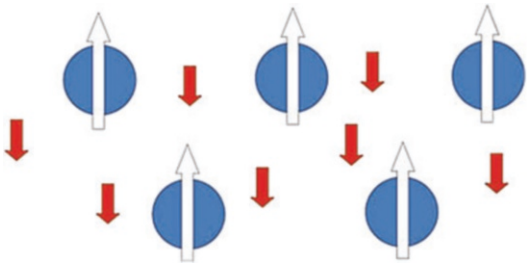
Semiconductor Spintronics

Spintronics in semiconductors – as opposed to or complementary to spintronics in metals – has been thoroughly investigated since the very beginning of research on spin-polarized currents. Semiconductor materials provide a variety of quantum effects which make them the ultimate choice for applications to micro- and nanoelectronics. Semiconductor-based solid state physics can be easily applied to nanostructures which in turn provide appropriate quantum confinement effects for carriers. Moreover the fabrication technology, based on reliable, accessible techniques of materials growth and process, has reached full maturity in the past decades.

From this viewpoint, applying spintronics concepts to semiconductor materials was a straightforward step; however, research in the field has faced various difficulties and drawbacks which have so far prevented widespread applications. The main reason being that standard semiconductors are diamagnetic materials without a self-sustained degree of spin polarization of the carriers, contrary to the case of ferromagnetic metals.

Two main strategies have been developed in order to cope with this shortcoming. The first one is based on injection of a spin-polarized current from a ferromagnetic metal such as Fe or Co into a semiconductor [12]. Of course both materials must be in the form of thin films in (more or less) close contact. In this way, a polarized current naturally produced in the ferromagnet can be injected in the semiconductor and propagate within it, although gradually losing the polarization degree. However, it can be shown that the time needed by the spin-polarized current to fully de-polarize through spin relaxation processes, although intrinsically short (being of the order of 100 ns in GaAs and up to 100 μs in GaAs quantum dots) would be enough for electrons to travel across the semiconductor over mesoscopic distances without losing spin coherence (tens to hundreds of μm), therefore allowing one to design new devices. The residual polarization degree can be experimentally checked by analyzing the light emitted in the recombination process of polarized electrons with unpolarized holes in the quantum well of a layered heterostructure (this is called the "spin-LED"). The emerging electromagnetic radiation appears to be circularly polarized. A measurement of the polarization of emitted light is therefore a measure of polarization degree of electron spins. The drawback of this approach is represented by its low efficiency, in turn determined not only by the rate of loss of spin memory within the semiconductor, but mostly by the loss of spin polarization during the injection process from the ferromagnetic metal. In fact, the Schottky barrier existing at the metal–semiconductor interface hinders the injection and reduces the efficient injection of a spin-polarized current. This difficulty can be partially – although not entirely – circumvented by adopting solutions which include, for instance, the presence of a tunnel barrier contact between the metal and the semiconductor.

The second strategy consists in building novel semiconductor materials which display



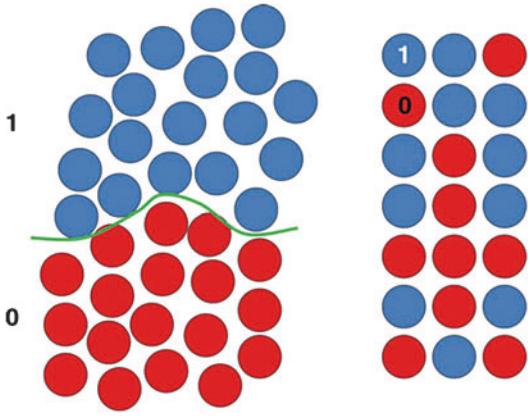
Magnetic Nanostructures and Spintronics, Fig. 6 Sketch of a dilute magnetic semiconductor spin configuration: the blue dots represent localized moments (Mn ions), while the *red arrows* represent itinerant spins (holes)

ferromagnetic properties while keeping their semiconducting features [13]. Adding moderate concentrations of specific dopants carrying a magnetic moment may result in obtaining the so-called dilute ferromagnetic semiconductors. Magnetic dopants exist; trivalent Mn atoms act as acceptors in GaAs-like semiconductors (i.e., those belonging to the (III,TM)V series, TM being a metal provided with intrinsic magnetic moment) and have a high permanent magnetic moment (entirely related to the spin quantum number $S = 5/2$). The key concept is that when magnetic Mn atoms occupy substitutional sites in the GaAs structure, they induce a polarization of the free carriers of the GaAs host. These are holes, rather than electrons, owing to the acceptor nature of dopants. Therefore, a magnetization of the itinerant carriers is sustained by the magnetization of localized moments on Mn ions; the latter is in turn sustained by that of itinerant carriers, in a sort of mutually sustained exchange effect (Fig. 6). It has been shown that, as a result, the spins of both Mn ions and mobile holes display a ferromagnetic ordering below the ordering temperature, although their magnetizations are opposite (strictly speaking, the semiconductor is therefore a ferrimagnetic rather than ferromagnetic material). However, spin-polarized, mobile holes can be transferred over long distances across the semiconductor always keeping their spin memory. A strict condition for the success of this strategy is to completely avoid clustering or agglomeration of magnetic dopants: their atoms must be evenly distributed in space. As a consequence, an upper dopant concentration limit (around 5 %) exists. Unfortunately, the Curie temperatures of these ferromagnetic semiconductors are still quite low (100 K). This has prevented widespread diffusion of dilute magnetic semiconductor materials so far.

Finally, it should be noted that many different spintronic devices based on semiconductors have been designed and/or achieved as demonstrators. Many borrow their working principles from classical optics of polarized, coherent light, exploiting the analogies existing between coherent optics and quantum coherence of ballistic electrons moving in one-dimensional semiconductor "channels" and between classical polarization of light and spin polarization. A popular spintronic device has been, for example, the spin-FET proposed by Datta and Das as early as 1990, which keeps the overall geometry and general working principle of a conventional FET, while exploiting ferromagnetic source and drain electrodes to produce a spin-polarized electron flow in the channel. The transmittance of the device is modulated by the gate voltage which may, or may not, induce a flip of the electron spin during the ballistic motion in the channel, therefore modifying the current intensity by quantum interference effects. In this interesting device, the spin flip is induced by an electrical rather than a magnetic field by thoroughly exploiting the spin–orbit interaction of ballistic electrons in the one-dimensional nanoconduit.

Data Storage and Patterned Media

Hard disk drives (HDDs) represent one of the most diffused systems for data storage, featuring extremely low costs per bit and reasonable data transfer rates. Nanostructured magnetic materials are present both in the read head and in the disc surface. The development of magnetic field sensors from magnetoresistive systems, to devices earlier based on GMR and nowadays based on TMR allowed the continuous increment in data areal density. On the other hand, recording materials, starting from structures magnetized in the disc plane, evolved to earlier structures with perpendicular magnetization, up to today's antiferromagnetically coupled layers that increase data



Magnetic Nanostructures and Spintronics, Fig. 7 Hard disk media magnetic domain structure. On the *left* conventional media for perpendicular recording, with randomly distributed magnetic domains and irregular bit contour. On the *right* patterned media, where each nanostructured element may carry one single bit of information

stability thanks to interlayer exchange coupling [14]. For what concerns recording media (Fig. 7), generally based on sputtered CoPtCrB alloys, the main problem is data retention rate: the smaller the size of a magnetic domain, the higher the effects of the demagnetizing field, the higher the switching probability. Reducing the magnetic domain size below the superparamagnetic limit would result in a useless media containing randomly distributed information. Nowadays the commercial data density is approaching an areal density of 100 Gbit/inch^2, while laboratory demonstrators are available approaching 1 Tbit/inch^2, based either on patterned media (PM) or thermally assisted magnetic recording. In particular, PM provides the key to break the superparamagnetic limit, because each nanometric magnetic domain corresponds to a confined physical grain that can be kept far from others avoiding self-demagnetization. The challenge of patterning a wide surface (some square inches) with homogeneously distributed 25-nm-big elements has been accomplished by means of *density multiplication*, a particular process where only a small fraction of the elements is patterned thanks to an electron beam system and the full feature density is achieved thanks to a block copolymer assembly guided by the underlying chemically modified former pattern [15]. Large area–patterned nanostructures are studied also for their interesting collective behavior [16], resulting in peculiar electrical properties [17] and potentially suitable to implement full magnetic logics, particularly promising for their low electrical consumption [18].

Cross-References

- ▶ Chemical Vapor Deposition (CVD)
- ▶ Electron Beam Lithography (EBL)
- ▶ Nanoparticles
- ▶ Physical Vapor Deposition
- ▶ Self-Assembly of Nanostructures

References

1. Sander, D.: The magnetic anisotropy and spin reorientation of nanostructures and nanoscale films. J. Phys. Condens. Matter **16**, R603–R636 (2004)
2. Nogues, J., Schuller, I.K.: Exchange bias. J. Magn. Magn. Mater. **192**, 203–232 (1999)
3. Chiolerio, A., Allia, P., Chiodoni, A., Pirri, F., Celegato, F., Coisson, M.: Thermally evaporated Cu-Co top spin-valve with random exchange bias. J. Appl. Phys. **101**, 123915-1–123915-6 (2007)
4. Berkowitz, A.E., Takano, K.: Exchange anisotropy – a review. J. Magn. Magn. Mater. **200**, 552–570 (1999)
5. Parkin, S.S.P.: Giant magnetoresistance in magnetic nanostructures. Annu. Rev. Mater. Sci. **25**, 357–388 (1995)
6. Yuasa, S., Djayapravira, D.D.: Giant tunnel magnetoresistance in magnetic tunnel junctions with a crystalline MgO(001) barrier. J. Phys. D Appl. Phys. **40**, R337–R354 (2007)
7. Slonczewski, J.C.: Current-driven excitation of magnetic multilayers. J. Magn. Magn. Mater. **159**, L1–L7 (1996)
8. Ralph, D.C., Stiles, M.D.: Spin transfer torques. J. Magn. Magn. Mater. **320**, 1190–1216 (2008)
9. Cros, V., Boulle, O., Grollier, J., Hamzic, A., Munoz, M., Pereira, L.G., Petroff, F.: Spin transfer torque: a new method to excite or reverse a magnetization. C. R. Phys. **6**, 956–965 (2005)
10. Sun, Z.: Spin angular momentum transfer in current-perpendicular nanomagnetic junctions. IBM J. Res. Dev. **50**, 81–100 (2006)
11. Wang, X., Chen, Y.: Spintronic memristor devices and application. In: Proceedings of Design, Automation and Test in Europe Conference and Exhibition (DATE), IEEE Conferences, pp. 667–672. Dresden (2010)

12. Schmidt, J.: Concepts for spin injection into semiconductors – a review. J. Phys. D Appl. Phys. **38**, R107–R122 (2005)
13. Awschalom, D.D., Flatté, M.E.: Challenges for semiconductor spintronics. Nat. Phys. **3**, 153–159 (2007)
14. Dietzel, A.: Hard disk drives. In: Waser, R. (ed.) Nanoelectronics and Information Technology, 2nd edn, pp. 617–629. Wiley, Weinheim (2005)
15. Density multiplication and improved lithography by directed block copolymer assembly for patterned media at 1 Tbit/in^2 and beyond, white paper WPDM08EN-01, Hitachi GST, San Jose (2008)
16. Adeyeye, A.O., Singh, N.: Large area patterned magnetic nanostructures. J. Phys. D Appl. Phys. **41**, 153001-1–153001-29 (2008)
17. Chiolerio, A., Allia, P., Celasco, E., Martino, P., Spizzo, F., Celegato, F.: Magnetoresistance anisotropy in a hexagonal lattice of Co antidots obtained by thermal evaporation. J. Magn. Magn. Mater. **322**, 1409–1412 (2010)
18. Graziano, M., Vacca, M., Chiolerio, A., Zamboni, M.: An NCL-HDL snake-clock-based magnetic QCA architecture. IEEE Trans. Nanotech. **10**, 1141–1149 (2011)

Magnetic Particle Imaging

▶ Magnetic Nanoparticles for Biomedical Applications

Magnetic Resonance Force Microscopy

Martino Poggio[1] and Christian L. Degen[2]
[1]Department of Physics, University of Basel, Basel, Switzerland
[2]Department of Physics, ETH Zurich, Zurich, Switzerland

Synonyms

Force-detected magnetic resonance

Definition

Magnetic resonance force microscopy (MRFM) is a type of magnetic resonance imaging (MRI) now capable of acquiring three-dimensional (3D) images of tiny samples with a spatial resolution better than 10 nm. By applying the techniques of scanning probe microscopy (SPM), MRFM surpasses the sensitivity of conventional, inductive MRI by eight orders of magnitude. MRFM images the interior of nanoscale objects noninvasively and with intrinsic chemical selectivity and, in 2009, it was used to capture 3D images of individual virus particles.

Overview

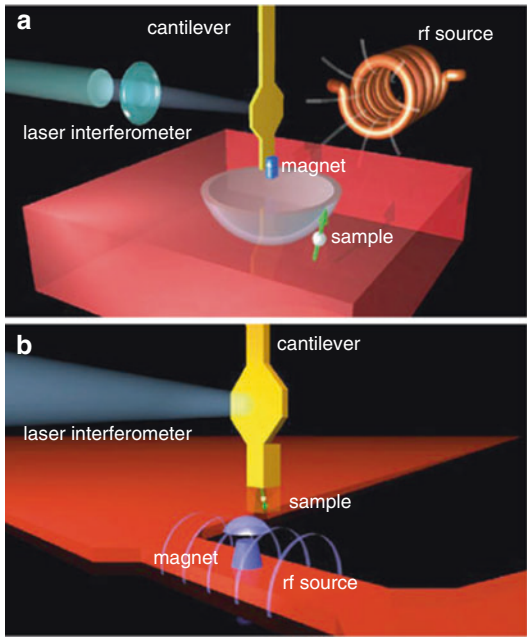
MRFM relies on the mechanical measurement of the weak magnetic force between a microscopic magnet and the magnetic moments in a sample. These moments are due to either the atomic nuclei with nonzero nuclear spin or electron spins present in a sample. For a single magnetic moment $\boldsymbol{\mu}$ in a magnetic field $\boldsymbol{B}$, this force can be expressed as:

$$\mathbf{F} = \nabla(\boldsymbol{\mu} \cdot \boldsymbol{B}) \tag{1}$$

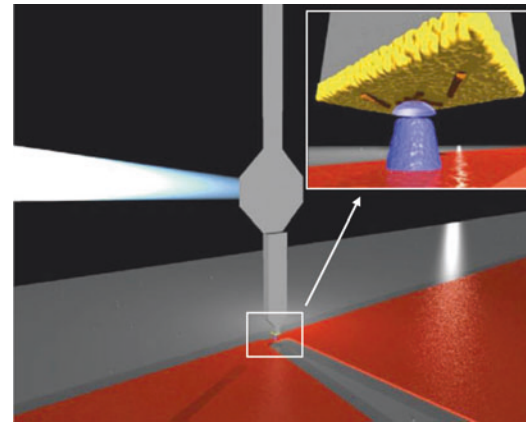
Using a compliant cantilever, one can measure the component of $\mathbf{F}$ along the cantilever's deflection direction $\hat{x}$:

$$F_x = \frac{\partial}{\partial x}(\boldsymbol{\mu} \cdot \boldsymbol{B}) = \mu \frac{\partial B_z}{\partial x} = \mu G, \tag{2}$$

where $\boldsymbol{\mu}$ points along $\hat{z}$ and $G \frac{\partial B_z}{\partial x}$ is a magnetic field gradient. First, either the sample containing nuclear or electronic moments or the nano-magnet must be fixed to the cantilever. The sample and magnet must be in close proximity, sometimes up to a few tens of nanometers from each other. A nearby radiofrequency (rf) source produces magnetic field pulses similar to those used in conventional MRI, causing the moments to periodically flip. This periodic inversion generates an oscillating magnetic force acting on the cantilever. In order to resonantly excite the cantilever, the magnetic moments can be inverted at the cantilever's mechanical resonance frequency. The cantilever's mechanical oscillations are then measured by an optical interferometer or beam deflection



Magnetic Resonance Force Microscopy, Fig. 1 Schematics of an MRFM apparatus. (**a**) Corresponds to the "tip-on-cantilever" arrangement, such as used in the single-electron MRFM experiment of 2004 [1]. (**b**) Corresponds to the "sample-on-cantilever" arrangement, like the one used for the nanoscale virus imaging experiment in 2008 [2]



Magnetic Resonance Force Microscopy, Fig. 2 Artistic view of the MRFM apparatus used for MRI of individual tobacco mosaic virus particles. Pictured is the cantilever, the laser beam used for position sensing, and the Cu microwire rf source. The *inset* shows a close-up representation of the gold-coated end of the cantilever with attached virus particles. On *top* of the microwire, is the magnetic FeCo tip with the "mushroom"-shaped resonant slice hovering *above*

detector. The electronic signal produced by the optical detector is proportional to the cantilever oscillation amplitude, which depends on the number of moments in the imaging volume. Spatial resolution results from the fact that the nano-magnet produces a magnetic field which is a strong function of position. The magnetic resonance condition and therefore the region in which the spins periodically flip is confined to a thin, approximately hemispherical "resonant slice" that extends outward from the nano-magnet (see Figs. 1 and 2). By scanning the sample in 3D through this resonant region, a spatial map of the magnetic moment density can be made. Different types of magnetic moments (e.g., ^{1}H,^{13}C, ^{19}F, or even electrons) can be distinguished due to their different magnetic resonance frequencies, giving an additional chemical contrast.

In conventional nuclear magnetic resonance detection, the sample is placed in a strong static magnetic field in order to produce a Zeeman splitting between spin states. The sample is then exposed to an rf magnetic field of a precisely defined frequency. If this frequency matches the Zeeman splitting, then the system absorbs energy from the rf radiation resulting in transitions between the spin states. The resulting oscillations of this ensemble of magnetic moments produce a time-varying magnetic signal that can be detected with a pickup coil. The electric current induced in the coil is then amplified and converted into a signal that is proportional to the number of moments (or spins) in the sample. In MRI this signal can be reconstructed into a 3D image of the sample using spatially varying magnetic fields and Fourier transform techniques. The magnetic fields produced by nuclear moments are, however, extremely small: more than one trillion (10^{12}) nuclear spins are typically needed to generate a detectable signal.

The advantage of force-detected over inductive techniques is that much smaller samples can be measured. In the latter case, the measurement can only be sensitive if the spins significantly alter the magnetic field within the pickup coil, i.e., if the spins fill a significant fraction of the coil volume.

For spin ensembles with volumes significantly smaller than $(1\ \mu m)^3$, it is extremely challenging to realize pickup coils small enough to ensure an adequate filling factor. As a result, even the best resolutions achieved by inductively detected MRI require sample volumes of at least $(3\ \mu m)^3$ [3]. Mechanical resonators, in contrast, can now be fabricated with dimensions far below a micron, such that the sample's mass (which is the equivalent to the filling volume in a pickup coil) is always significant compared to the bare resonator mass. In addition, mechanical devices usually show resonant quality factors that surpass those of inductive circuits by orders of magnitude, resulting in a much lower baseline noise. For example, state-of-the art cantilever force transducers achieve quality factors between 10^4 and 10^7, enabling the detection of forces of $aN/Hz^{1/2}$ – less than a billionth of the force needed to break a single chemical bond. In addition, scanning probe microscopy offers the stability to position and image samples with nanometer precision. The combination of these features allows mechanically detected MRI to image at resolutions that are far below 1 μm and – in principle – to aspire to atomic resolution.

Basic Methodology

Despite the large magnetic field gradients produced by specially made magnetic tips, the forces produced by small number of magnetic moments remain astoundingly small: in an achievable gradient of 10^6 T/m, a single ^1H moment, for example, produces a force of about 10^{-20} N. In order to detect forces due to small ensembles of moments, the most sensitive MRFM takes advantage of the thermally limited force resolution of specially made nanomechanical cantilevers. In addition, it utilizes the fact that the force of interest can be made to occur in a narrow bandwidth around the cantilever's mechanical resonance.

The sensitivity and resolution of MRFM hinge on a simple signal-to-noise ratio, which is given by the ratio of the magnetic force power exerted on the cantilever over the force noise power of the cantilever device. For small enough volumes of spins, one must measure statistical spin polarizations; therefore one is interested in force powers (or variances) rather than force amplitudes:

$$\mathrm{SNR} = N\frac{(\mu G)^2}{S_F \Delta f}. \tag{3}$$

Here, N is the number of spins in the detection volume, μ is the magnetic moment of the relevant spin, G is the relevant magnetic field gradient at the position of the sample, SF is the force noise spectral density set by the fluctuations of the cantilever sensor, and Δf is the bandwidth of the measurement, determined by the spin relaxation rate. This expression gives the single-shot signal-to-noise ratio of a thermally limited MRFM apparatus. The larger this signal-to-noise ratio is, the better the spin sensitivity will be.

From the four parameters appearing in Eq. 3, only two can be controlled and possibly improved. On the one hand, the magnetic field gradient G can be enhanced by using higher quality magnetic tips and by bringing the sample closer to these tips. On the other hand, the force noise spectral density SF can be reduced by going to lower temperatures and by making intrinsically more sensitive mechanical transducers.

Key Findings

The use of force-detection techniques in nuclear magnetic resonance (NMR) experiments dates back to Evans in 1955 [4], and was also used in torque magnetometry measurements by Alzetta and coworkers in the 1960s [5]. In 1991, Sidles, independent of this very early work, realized that highly sensitive magnetic resonance detection could be achieved using microfabricated cantilevers and nanoscale ferromagnets. In these early days of scanning probe miscroscopy, he proposed the MRFM technique as a method to improve the resolution of MRI to molecular length scales [6, 7]. Subsequently, the first micrometer-scale experimental demonstration using cantilevers was realized by Rugar [8], demonstrating mechanically detected electron spin resonance in

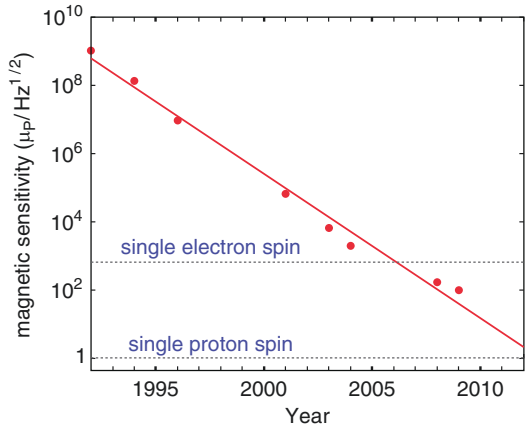
a 30-ng sample of diphenylpicrylhydrazyl (DPPH).

The visionary goal of the MRFM proposal was to eventually image molecules atom by atom, so as to directly map the 3D atomic structure of macromolecules [9]. Such a "molecular structure microscope" would have a dramatic impact on modern structural biology, and would be an important tool for many future nanoscale technologies. While this ultimate goal has not been achieved to date, the technique has undergone a remarkable development into one of the most sensitive spin detection methods available to researchers today. Among the important experimental achievements are the detection of a single electronic spin [1] and the extension of the spatial resolution of MRI from several micrometers to below 10 nm [2].

First Demonstration

The first MRFM apparatus operated in vacuum and at room temperature with the DPPH sample attached to the cantilever. A millimeter-sized coil produced an rf magnetic field tuned to the electron spin resonance of the DPPH (220 MHz) with a magnitude of 1 mT. By changing the strength of a polarizing magnetic field (8 mT) in time, the electron spin magnetization in the DPPH was modulated. In a magnetic field gradient of 60 T/m, produced by a nearby NdFeB permanent magnet, the sample's oscillating magnetization resulted in a time-varying force between the sample and the magnet. This force modulation was converted into mechanical vibration by the compliant cantilever. Displacement oscillations were detected by a fiber-optic interferometer achieving a thermally limited force sensitivity of 3 fN/$\sqrt{\text{Hz}}$.

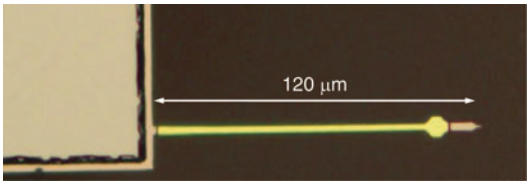
During the years following this initial demonstration of cantilever-based magnetic resonance detection, the technique has undergone a series of developments toward higher sensitivities that, as of today, is about 10^7 times that of the 1992 experiment (see Fig. 3). The following touches on two important demonstrations during this development. Several review articles and book chapters have appeared in the literature that discuss the historical progress of the technique more broadly and in richer detail [15–22].



Magnetic Resonance Force Microscopy, Fig. 3 Advances to the sensitivity in force-detected magnetic resonance over time. Remarkably, improvements have closely followed a "Moore's law" for over a decade, with the magnetic moment sensitivity doubling roughly every 8 months. *Dots* are experimental values [1, 2, 8, 10–14], and *dashed lines* indicate sensitivities of one electron and one proton magnetic moment (μ_P), respectively

Single-Electron MRFM

The measurement of a single electron spin by Rugar et al. in 2004 concluded a decade of development on the MRFM technique and stands out as one of the first single-spin measurements in solid-state physics. A variety of developments led to the exceptional measurement sensitivity required for single-spin detection. These include the operation of the apparatus at cryogenic temperatures and high vacuum, the ion-beam milling of magnetic tips in order to produce large gradients, and the fabrication of mass-loaded attonewton-sensitive cantilevers [23] (shown in Fig. 4). The thermal noise in higher order vibrational modes of mass-loaded cantilevers is suppressed compared with the noise in the higher order modes of conventional, "flat" cantilevers. Since high-frequency vibrational noise in combination with a magnetic field gradient can disturb the electron spin, the mass-loaded levers proved to be a crucial advance for single-electron MRFM. In addition, Rugar et al. developed a sensitive interferometer employing only a few nanowatts of optical power for the detection of cantilever displacement [24]. This low incident laser power is crucial for achieving low cantilever temperatures and thus



Magnetic Resonance Force Microscopy, Fig. 4 Image of an ultrasensitive mass-loaded Si cantilever taken from an optical microscope. This type of cantilever, which has a spring constant under 100 μN/m, has been used as a force transducer in the many of the latest MRFM experiments [23]

minimizing the effects of thermal force noise. A low-background measurement protocol called OSCAR based on the NMR technique of adiabatic rapid passage was also employed [10]. Finally, the experiment required the construction of an extremely stable measurement system capable of continuously measuring for several days in an experiment whose single-shot signal-to-noise ratio was just 0.06 [1].

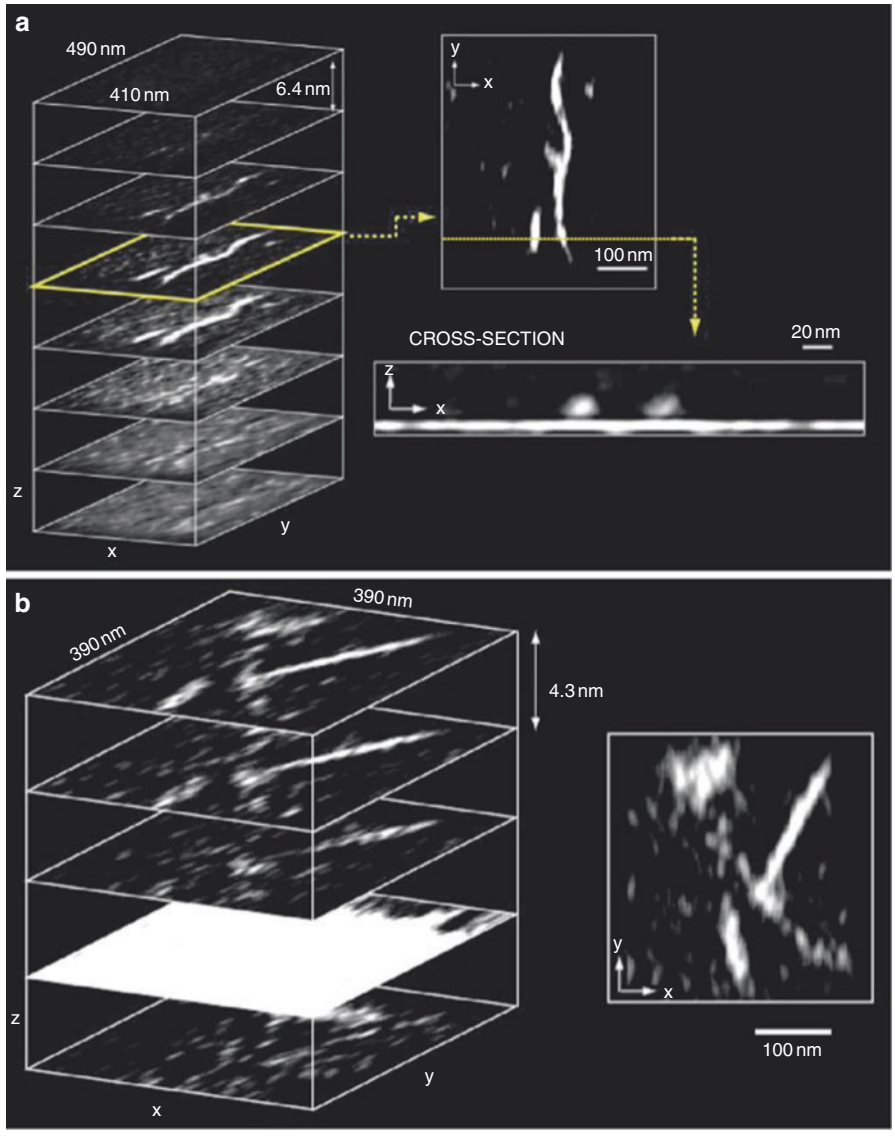
Nano-MRI

The detection of a single nuclear spin is far more challenging than that of single electron spin. The intrinsic magnetic moment of a nucleus is much smaller than that of an electron; as a result the force produced by a nuclear moment in Eq. 2 is proportionally smaller. A ^{1}H nucleus (proton), for example, possess a magnetic moment that is only $\sim 1/650$ of an electron spin moment. Other important nuclei, like ^{13}C or a variety of isotopes present in semiconductors, have even weaker magnetic moments. In order to observe single nuclear spins, it is necessary to improve the state-of-the-art sensitivity by another two to three orders of magnitude. While not out of the question, this is a daunting task that requires significant advances to all aspects of the MRFM technique. Despite these challenges, significant progress has been made. In particular, one experiment that used single tobacco mosaic virus (TMV) particles as the sample shows both the feasibility of MRI imaging with nanometer resolution and the applicability of MRFM to biologically relevant samples.

Figure 2 is a representation of the MRFM apparatus used in the experiment. The virus particles were transferred to the cantilever end by dipping the tip of the cantilever into a droplet of aqueous solution containing suspended TMV. As a result, some TMVs were attached to the gold layer previously deposited on the cantilever end. The density of TMV on the gold layer was low enough that individual particles could be isolated. Then the cantilever was mounted into the low-temperature, ultra-high-vacuum measurement system and aligned over a magnetic tip and rf field source.

After applying a static magnetic field of about 3 T, resonant rf pulses were applied to the rf source in order to flip the ^{1}H nuclear spins at the cantilever's mechanical resonance. Finally, the end of the cantilever was mechanically scanned in 3D over the magnetic tip. Given the extended geometry of the region in which the resonant condition is met, i.e., the "resonant slice," a spatial scan does not directly produce a map of the ^{1}H distribution in the sample. Instead, each data point in the scan contains force signal from ^{1}H spins at a variety of different positions. In order to reconstruct the 3D spin density (the MRI image), the force map must be deconvolved by the point spread function defined by the resonant slice. Fortunately, this point spread function can be accurately determined using a magnetostatic model based on the physical geometry of the magnetic tip and the tip magnetization. Deconvolution of the force map into the three-dimensional ^{1}H spin density can be done in several different ways; for the results presented in [2] the authors applied the iterative Landweber deconvolution procedure suggested in an earlier MRFM experiment [25, 26]. This iterative algorithm starts with an initial estimate for the spin density of the object and then improves the estimate successively by minimizing the difference between the measured and predicted spin signal maps. The iterations proceed until the residual error becomes comparable with the measurement noise.

The result of a representative experiment is shown in Fig. 5. Here, clear features of individual TMV particles, which are cylindrical, roughly 300-nm-long and 18 nm in diameter, are visible and can be confirmed against a scanning electron micrograph (SEM) of the same region. Both whole virus particles and particle fragments are



Magnetic Resonance Force Microscopy, Fig. 5 Nanoscale MRI images of tobacco mosaic virus (*TMV*) particles acquired by MRFM [2]. (**a**) The series of images *to the left* depicts the 3D ^{1}H spin density of virus particles deposited on the end of the cantilever. *Black* represents very low or zero density of hydrogen, while *white* is high hydrogen density. The *right side* shows a representative xy-plane, with several viral fragments visible, and a cross section (xz-plane) of two virus particles that reveals an underlying molecular layer of hydrocarbons covering the cantilever surface. (**b**) 3D ^{1}H spin density recorded on a different region of the same cantilever as in (**a**), showing an intact and several fragmented virus particles. The *right side* shows a representative xy-plane

observed. The imaging resolution, while not fine enough to discern any internal structure of the virus particles, constitutes a 1,000-fold improvement over conventional MRI and a corresponding improvement of volume sensitivity by about 100 million.

Comparison to Other Techniques

The unique position of MRFM among high-resolution microscopies becomes apparent when comparing it to other, more established nanoscale imaging techniques. As a genuine scanning probe

method, MRFM has the potential to image matter at atomic resolution. While atomic-scale imaging is routinely achieved in scanning tunneling microscopy and atomic force microscopy, these techniques are confined to the top layer of atoms and cannot penetrate below surfaces [27, 28]. Moreover, in standard scanning probe microscopy (SPM), it is difficult and in many situations impossible to identify the chemical species being imaged. Since MRFM combines SPM and MRI, these restrictions are lifted. The 3D nature of MRI permits acquisition of subsurface images with high spatial resolution even if the probe is relatively far away. As with other magnetic resonance techniques, MRFM comes with intrinsic elemental contrast and can draw from established NMR spectroscopy procedures to perform detailed chemical analysis. In addition, MRI does not cause any radiation damage to samples, as do electron and X-ray microscopies.

MRFM also distinguishes itself from super-resolution optical microscopies that rely on fluorescence imaging [29]. On the one side, optical methods have the advantage of working in vivo and they have the ability to selectively target the desired parts of a cell. Fluorescent labeling is now a mature technique which is routinely used for cellular imaging. On the other side, pushing the resolution into the nanometer range is hampered by fundamental limitations, in particular, the high optical powers required and the stability of the fluorophores. Moreover, fluorescent labeling is inextricably linked with a modification of the target biomolecules, which alters the biofunctionality and limits imaging resolution to the physical size of the fluorophores.

MRFM occupies a unique position among other nanoscale spin detection approaches. While single electron spin detection in solids has been shown using several techniques, these mostly rely on the indirect readout via electronic charge [30, 31] or optical transitions [32, 33]. In another approach, the magnetic orientation of single atoms has been measured via the spin-polarized current of a magnetic STM tip or using magnetic exchange force microscopy [34–36]. These tools are very valuable to study single surface atoms; however, they are ill-suited to map out subsurface spins such as paramagnetic defects. In contrast, MRFM *directly* measures the magnetic moment of a spin, without resorting to other degrees of freedom, making it a very general method. This direct measurement of magnetic moment (or magnetic stray field) could also be envisioned using other techniques, namely, SQuID microscopy [37], Hall microscopy [38], or recently introduced diamond magnetometry based on single nitrogen-vacancy centers [39–41]. So far, however, none of these methods have reached the level of sensitivity needed to detect single electron spins, or volumes of nuclear spins much less than 1 μm [42, 43]. It is certainly possible that future improvements to these methods – especially to diamond magnetometry – may result in alternative techniques for nanoscale MRI that surpass the capabilities of MRFM.

Outlook

Despite the tremendous improvements made to MRFM over the last decade, several important obstacles must be overcome in order to turn the technique into a useful tool for biologists and materials scientists. Most existing MRFM instruments are technically involved prototypes; major hardware simplifications will be required for routine screening of nanoscale samples. Suitable specimen preparation methods must be developed that are compatible with the low-temperature, high-vacuum environment required for the microscope to operate at its highest sensitivity and resolution. While this is particularly challenging for biological samples, protocols exist which could be adapted to MRFM. In cryo-electron microscopy, for example, dispersed samples are vitrified to preserve their native structure by plunge-freezing in liquid nitrogen [44]. As objects become smaller, isolation of samples and suppression of unwanted background signals from surrounding material will become increasingly important.

The conditions under which the latest MRFM imaging experiments were carried out are remarkably similar to those prevailing in cryo-electron

microscopy, the highest resolution 3D imaging technique commonly used by structural biologists today. Cryo-electron microscopy, like MRFM, operates at low temperatures and in high vacuum, requires long averaging times (on the order of days) to achieve sufficient contrast, and routinely achieves resolutions of a few nanometers [45, 46]. Unlike MRFM, however, electron microscopy suffers from fundamental limitations that severely restrict its applicability. Specimen damage by the high-energy electron radiation limits resolution to 5–10 nm if only a single copy of an object is available. Averaging over hundreds to thousands of copies is needed to achieve resolutions approaching 10 Å [47]. In addition, unstained images have intrinsically low contrast, whereas staining comes at the expense of modifying the native structure.

MRFM has the unique capability to image nanoscale objects in a noninvasive manner and to do so with intrinsic chemical selectivity. For this reason the technique has the potential to extend microscopy to the large class of structures that show disorder and therefore cannot be averaged over many copies. These structures include such prominent examples as HIV, influenza virus, and amyloid fibrils. Virtually all of these complexes are associated with important biological functions ranging from a variety of diseases to the most basic tasks within the cellular machinery. For such complexes, MRFM has the potential not only to image the 3D macromolecular arrangement, but also to selectively image specific domains in the interior through isotopic labeling.

While the most exciting prospect for MRFM remains in its application to structural imaging in molecular biology, its applications are not limited to biological matter. For example, most semiconductors contain nonzero nuclear magnetic moments. Therefore, MRFM may prove useful for subsurface imaging of nanoscale electronic devices. MRFM also appears to be the only technique capable of directly measuring the dynamics of the small ensembles of nuclear spin that limit electron spin coherence in single semiconductor quantum dots. Polymer films and self-assembled monolayers – important for future molecular electronics – are another exciting target for MRFM and its capability to image chemical composition on the nanoscale. Finally, isotopically engineered materials are becoming increasingly important for tuning a variety of physical properties such as transport and spin. Researchers currently lack a general method for noninvasively imaging the isotopic composition of these materials [48–50]; MRFM techniques could fill this void.

Acknowledgements Thanks to H. J. Mamin and D. Rugar of the IBM Almaden Research Center for their many detailed comments and very helpful discussions pertaining to this text.

Cross-References

▶ Atomic Force Microscopy
▶ Basic MEMS Actuators
▶ Imaging the Human Body Down to the Molecular Level

References

1. Rugar, D., et al.: Single spin detection by magnetic resonance force microscopy. Nature **430**, 329 (2004)
2. Degen, C.L., Poggio, M., Mamin, H.J., Rettner, C.T., Rugar, D.: Nanoscale magnetic resonance imaging. Proc. Natl. Acad. Sci. U. S. A. **106**, 1313 (2009)
3. Ciobanu, L., Seeber, D.A., Pennington, C.H.: 3D MR microscopy with resolution by 3.7 7 μm mu;m by 3.3 3 μm mu;m by 3.3 μm. J. Magn. Reson. **158**, 178 (2002); Lee, S.-C., et al.: One micrometer resolution NMR microscopy. J. Magn. Reson. **150**, 207 (2001); Glover, P., Mansfield, P.: Limits to magnetic resonance microscopy. Rep. Prog. Phys. **65**, 1489 (2002)
4. Evans, D.F.: Direct observation of static nuclear susceptibilities at room temperature. Philos. Mag. **1**, 370 (1956)
5. Alzetta, G., Arimondo, E., Ascoli, C., Gozzini, A.: Paramagnetic resonance experiments at low fields with angular-momentum detection. Il Nuovo Cimento B **62**, 392 (1967)
6. Sidles, J.A.: Noninductive detection of single-proton magnetic resonance. Appl. Phys. Lett. **58**, 2854 (1991)
7. Sidles, J.A.: Folded Stern-Gerlach experiment as a means for detecting nuclear magnetic resonance in individual nuclei. Phys. Rev. Lett. **68**, 1124 (1992)
8. Rugar, D., Yannoni, C.S., Sidles, J.A.: Mechanical detection of magnetic resonance. Nature **360**, 563 (1992)

9. Sidles, J.A., Garbini, J.L., Drobny, G.P.: The theory of oscillator-coupled magnetic resonance with potential applications to molecular imaging. Rev. Sci. Instrum. **63**, 3881 (1992)
10. Stipe, B.C., et al.: Magnetic dissipation and fluctuations in individual nanomagnets measured by ultrasensitive cantilever magnetometry. Phys. Rev. Lett. **86**, 2874 (2001)
11. Rugar, D., et al.: Force detection of nuclear magnetic resonance. Science **264**, 1560 (1994)
12. Wago, K., Züger, O., Kendrick, R., Yannoni, C.S., Rugar, D.: Low-temperature magnetic resonance force detection. J. Vac. Sci. Technol. B **14**, 1197 (1996)
13. Mamin, H.J., Budakian, R., Chui, B.W., Rugar, D.: Detection and manipulation of statistical polarization in small spin ensembles. Phys. Rev. Lett. **91**, 207604 (2003)
14. Mamin, H.J., et al.: Isotope-selective detection and imaging of organic nanolayers. Nano Lett. **9**, 3020 (2009)
15. Sidles, J.A., et al.: Magnetic resonance force microscopy. Rev. Mod. Phys. **67**, 249 (1995)
16. Nestle, N., Schaff, A., Veeman, W.S.: Mechanically detected NMR, an evaluation of the applicability for chemical investigations. Prog. Nucl. Magn. Reson. Spectrosc. **38**, 1 (2001)
17. Suter, A.: The magnetic resonance force microscope. Prog. Nucl. Magn. Reson. Spectrosc. **45**, 239 (2004)
18. Berman, G.P., Borgonovi, F., Gorshkov, V.N., Tsifrinovich, V.I.: Magnetic Resonance Force Microscopy and a Single-spin Measurement. World Scientific, Singapore/London/Hong Kong (2006)
19. Hammel, P.C., Pelekhov, D.V.: The magnetic resonance microscope. In: Kronmüller, H., Parkin, S. (eds.) Handbook of Magnetism and Advanced Magnetic Materials. Spintronics and Magnetoelectronics, vol. 5. Wiley, Chichester (2007). ISBN 978-0-470-02217-7
20. Kuehn, S., Hickman, S.A., Marohn, J.A.: Advances in mechanical detection of magnetic resonance. J. Chem. Phys. **128**, 052208 (2008)
21. Barbic, M.: Magnetic resonance force microscopy, in magnetic resonance microscopy: spatially resolved NMR techniques and applications. In: Codd, S.L., Seymour, J.D. (eds.) Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim (2009)
22. Poggio, M., Degen, C.L.: Force-detected nuclear magnetic resonance: recent advances and future challenges. Nanotechnology **21**, 342001 (2010)
23. Chui, B.W., et al.: Mass-loaded cantilevers with suppressed higher-order modes for magnetic resonance force microscopy. TRANSDUCERS, 12th International Conference on Solid-State Sensors, Actuators and Microsystems, Boston, vol. 2, p. 1120 (2003)
24. Mamin, H.J., Rugar, D.: Sub-attonewton force detection at millikelvin temperatures. Appl. Phys. Lett. **79**, 3358 (2001)
25. Chao, S., Dougherty, W.M., Garbini, J.L., Sidles, J.A.: Nanometer-scale magnetic resonance imaging. Rev. Sci. Instrum. **75**, 1175 (2004)
26. Dobigeon, N., Hero, A.O., Tourneret, J.-Y.: Hierarchical Bayesian sparse image reconstruction with application to MRFM. IEEE Trans. Image Process. **18**, 2059 (2009)
27. Hansma, P., Tersoff, J.: Scanning tunneling microscopy. J. Appl. Phys. **61**, R1–R23 (1987)
28. Giessibl, F.: Advances in atomic force microscopy. Rev. Mod. Phys. **75**, 949–983 (2003)
29. Huang, B., Bates, M., Zhuang, X.: Super-resolution fluorescence microscopy. Annu. Rev. Biochem. **78**, 993–1016 (2009)
30. Elzerman, J.M., et al.: Single-shot read-out of an individual electron spin in a quantum dot. Nature **430**, 431 (2004)
31. Xiao, M., Martin, I., Yablonovitch, E., Jiang, H.W.: Electrical detection of the spin resonance of a single electron in a silicon field-effect transistor. Nature **430**, 435 (2004)
32. Wrachtrup, J., von Borczyskowski, C., Bernard, J., Orritt, M., Brown, R.: Optical detection of magnetic resonance in a single molecule. Nature **363**, 244 (1993)
33. Jelezko, F., Popa, I., Gruber, A., Tietz, C., Wrachtrup, J.: Single spin states in a defect center resolved by optical spectroscopy. Appl. Phys. Lett. **81**, 2160 (2002)
34. Heinze, S., et al.: Real-space imaging of two-dimensional antiferromagnetism on the atomic scale. Science **288**, 1805–1808 (2000)
35. Durkan, C., Welland, M.E.: Electronic spin detection in molecules using scanning-tunneling- microscopy-assisted electron-spin resonance. Appl. Phys. Lett. **80**, 458 (2002)
36. Kaiser, U., Schwarz, A., Wiesendanger, R.: Nature **446**, 522–525 (2007)
37. Kirtley, J.R., et al.: High-resolution scanning SQUID microscope. Appl. Phys. Lett. **66**, 1138 (1995)
38. Chang, A.M., et al.: Scanning hall probe microscopy. Appl. Phys. Lett. **61**, 1974 (1992)
39. Degen, C.L.: Scanning magnetic field microscope with a diamond single-spin sensor. Appl. Phys. Lett. **92**, 243111 (2008)
40. Maze, J.R., et al.: Nanoscale magnetic sensing with an individual electronic spin in diamond. Nature **455**, 644 (2008)
41. Balasubramanian, G., et al.: Nanoscale imaging magnetometry with diamond spins under ambient conditions. Nature **455**, 648 (2008)
42. Degen, C.L.: Nanoscale magnetometry: microscopy with single spins. Nat. Nanotechnol. **3**, 643 (2008)
43. Blank, A., Suhovoy, E., Halevy, R., Shtirberg, L., Harneit, W.: ESR imaging in solid phase down to sub-micron resolution: methodology and applications. Phys. Chem. Chem. Phys. **11**, 6689–6699 (2009)
44. Taylor, K.A., Glaeser, R.M.: Electron diffraction of frozen, hydrated protein crystals. Science **186**, 1036 (1974)

45. Lucic, V., Forster, F., Baumeister, W.: Structural studies by electron tomography: from cells to molecules. Annu. Rev. Biochem. **74**, 833 (2005)
46. Subramaniam, S.: Bridging the imaging gap: visualizing subcellular architecture with electron tomography. Curr. Opin. Microbiol. **8**, 316 (2005)
47. Glaeser, R.M.: Cryo-electron microscopy of biological nanostructures. Phys. Today **61**, 48 (2008)
48. Shimizu, Y., Itoh, K.: Growth and characterization of short-period silicon isotope superlattices. Thin Solid Films **508**, 160–162 (2006)
49. Shlimak, I., Safarov, V., Vagner, I.: Isotopically engineered silicon/silicon-germanium nanostructures as basic elements for a nuclear spin quantum computer. J. Phys. Condens. Matter **13**, 6059 (2001)
50. Kelly, T., Miller, M.: Atom probe tomography. Rev. Sci. Instrum. **78**, 031101 (2007)

Magnetic Resonance Imaging

▶ Magnetic Nanoparticles for Biomedical Applications

Magnetic Self-Assembly

▶ Magnetic-Field-Based Self-Assembly

Magnetic Separation

▶ Magnetic Nanoparticles for Biomedical Applications

Magnetic-Field-Based Self-Assembly

Zhang Li[1], Bradley J. Nelson[1] and Lixin Dong[2]
[1]Institute of Robotics and Intelligent Systems, ETH Zurich, Zurich, Switzerland
[2]Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA

Synonyms

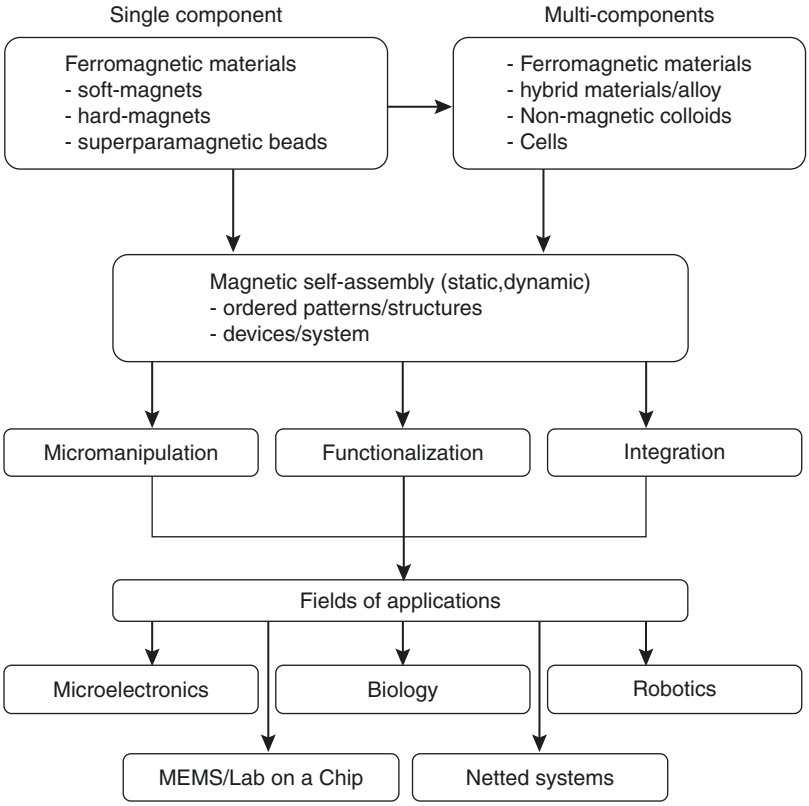
Magnetic assembly; Magnetic self-assembly

Definition

Magnetic-field-based self-assembly refers to the autonomous organization of discrete components to form ordered patterns, structures, devices, and systems driven by an external magnetic field or by magnetic fields from permanent dipoles incorporated into parts. The magnetic-field-based self-assembly approach can be applied for the assembly of individual magnetic components such as magnetic nanoparticles and nanowires. It can also be applied for the self-assembly of multiple components, such as components with complementary micromagnet arrays or a mixture of magnetic and nonmagnetic components.

Overview and Concepts

For magnetic self-assembly, a magnetic interaction between magnetic components is required [1]. The strength of the magnetic interaction force is dependent on the distance and orientation between the magnetic components, the magnetization of the materials, and the dimensions and shapes of the magnetic components. Most research studies have used either ferromagnetic materials or superparamagnetic particles as magnetic components for magnetic assembly. Generally, ferromagnetic materials are classified as either hard magnetic or soft magnetic depending on their coercivity [2]. For hard magnetic materials that are permanently magnetized, an external magnetic field may not be required for self-assembly. The magnetization of soft-magnetic materials is highly dependent on the external magnetic field, and they tend to return to their nonmagnetized state as the field is removed because of their low remanence magnetization. When the dimensions of ferromagnetic material are reduced to a single magnetic domain, superparamagnetism occurs in which the saturation magnetization can be as high as that in a ferromagnetic state but its remanence magnetization decreases to zero. Therefore, an external field is required for the magnetic self-assembly of soft magnetic or superparamagnetic components. Furthermore, shape effect, i.e., the geometrically



Magnetic-Field-Based Self-Assembly, Fig. 1 Overview of magnetic self-assembly: concept and applications

induced directional dependency of the magnetic properties within the magnetic component [2], significantly influences how the components magnetize and, hence, how soft-magnetic or superparamagnetic components assemble. In addition, the environment for self-assembly is crucial, since the components move with respect to one another [1]. One issue that can occur for micro-/nanoscale components is that van der Waals interactions and other surface forces dominate gravitational forces and other volume forces at this scale. One way to reduce these interactions is to perform self-assembly in liquid or at a liquid-air interface.

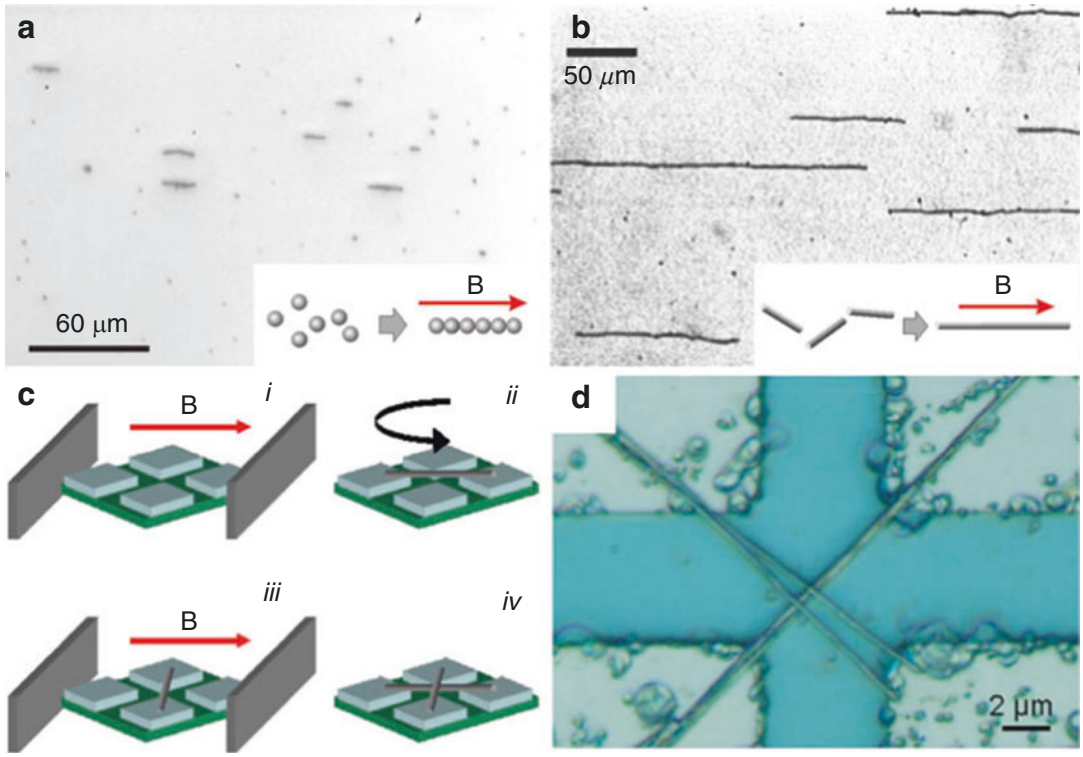
Magnetic-field-based self-assembly has many advantages [3], for instance, it is compatible for both wet and dry environments; it does not require physical contact for assembly; it is scalable in a large range, i.e., from nano-/meso- to macroscale; it has robust control of angular and lateral alignment; and, in addition, a weak-strength magnetic field is considered safe for in vivo applications.

Magnetic assembly has potential applications in different fields, such as microfluidics and lab on a chip [4–7], wireless biosensing [8], biomedical engineering [9], microelectronics [10–12], and microrobotics [13–15]. An overview of magnetic self-assembly including concepts and applications is shown in Fig. 1.

Key Research Findings

Static Self-Assembly of Single Magnetic Components

Magnetic particles and nanowires are two fundamental components for magnetic self-assembly of one-dimensional (1-D) structures. Figure 2a shows a suspension of superparamagnetic beads, after a uniform magnetic field has been applied. The beads are driven to form chains and align with the external field. Self-assembly is caused by the induced magnetic moments in the beads, and the magnetic alignment of the chain resembles an elongated



Magnetic-Field-Based Self-Assembly, Fig. 2 Magnetic self-assembly of superparamagnetic beads and Ni nanowires. (**a**) Self-assembly of 1 μm diameter superparamagnetic microbeads suspension to form straight chains. The applied field strength (*B*) is 2.5 mT. (**b**) Ni nanowire chains are formed when a low strength magnetic field is applied. The applied field strength (*B*) is 0.8 mT. The individual nanowires have a diameter of 350 nm and a length of 12 μm (Reprinted with permission from Ref. [16]. Copyright 2002, American Institute of Physics). *Insets*: schematic drawings of magnetic self-assembly of beads (**a**) and wires (**b**) when a field is applied. (**c**) The two-step alignment procedure with (*i*) alignment in a single direction, (*ii*) followed by a substrate rotation, (*iii*) applying an external magnetic field in a perpendicular direction to the assembled nanowire, and (*iv*) the second alignment step to generate a junction. (**d**) A junction of crossing nickel nanowires on a ferromagnetic electrode. The nanowires have a diameter of 200 nm. (**c**–**d** Reproduced with permission from Ref. [17]. Copyright 2007 IOP)

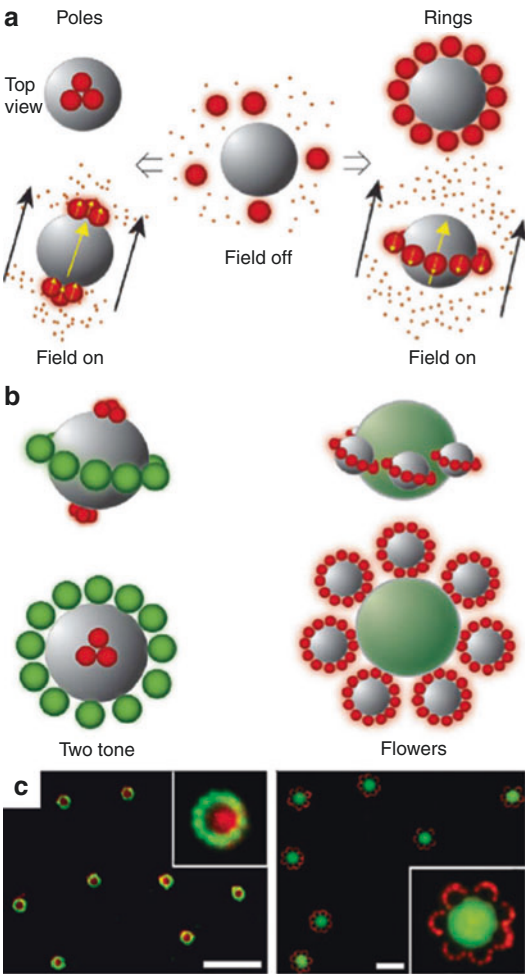
magnetic body aligned with the applied field [6]. In comparison to individual magnetic beads, which have no magnetic shape anisotropy, a magnetic nanowire has an "easy" magnetized axis along its long axis. When a magnetic field is applied to nanowire suspensions, they are driven to form chains and self-align their easy axis along the magnetic field, (see Fig. 2b) [16]. The results also indicate that to self-assemble magnetic micro-/nanoparticles and nanowires in fluid, a weak-strength magnetic field, i.e., in a range of milli-Tesla (mT), is sufficient, which is approximately three orders of magnitude lower than most clinical magnetic resonance imaging (MRI) systems. Therefore, the magnetic self-assembly approach is considered noninvasive for biomedical applications such as cell manipulation and biosensing.

The construction of 2-D or 3-D patterns using single magnetic components is not straightforward. One method to overcome this issue is to perform the assembly in steps. Magnetic assembly of ferromagnetic nanowires to create networks comprised of cross-shaped and T-shaped junctions was shown by Hangarter et al. [17]. The hierarchical 2-D patterns were obtained using two steps of magnetic alignment in which

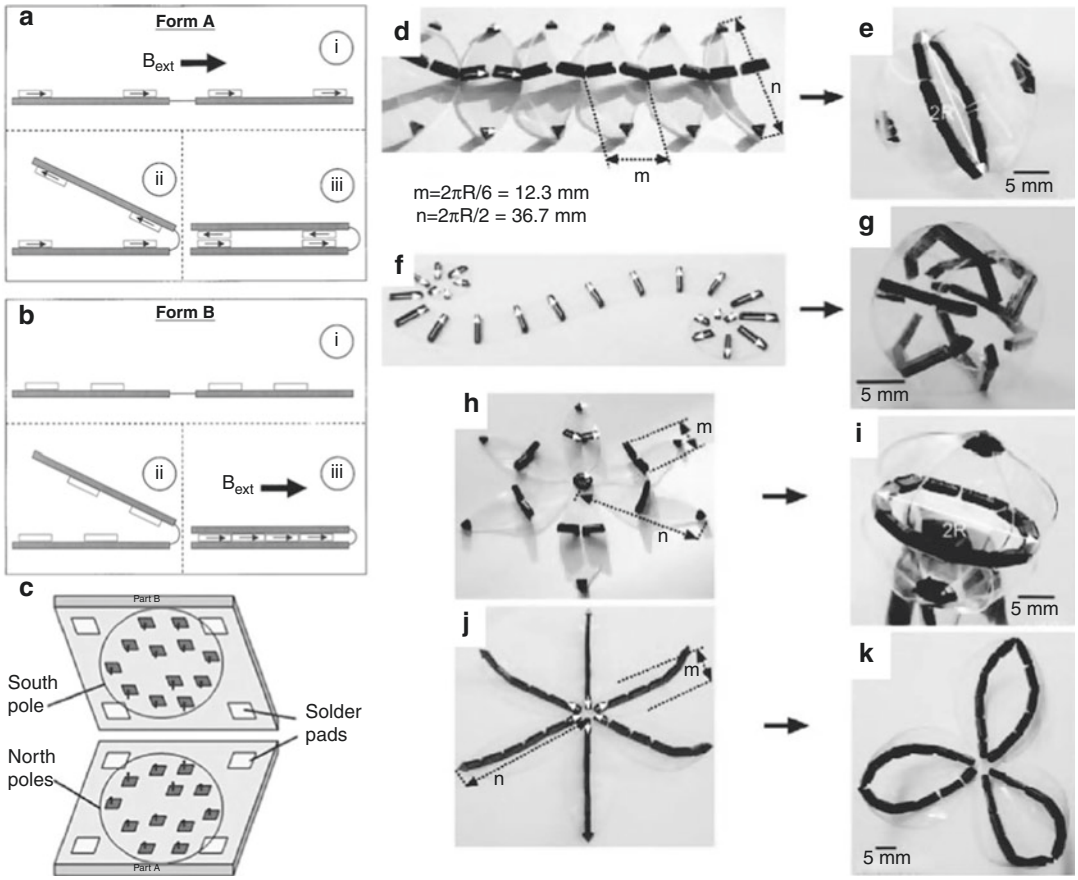
crossing angles from 30° to 90° were obtained. The magnetic self-assembly and alignment process were conducted by placing two permanent magnets around the substrate covered with droplets of the nanowire suspensions. The pattern configuration was controlled by changing the direction and the strength of the applied field in two steps. Magnetic field strengths of 28, 76, and 124 mT were applied in the experiments by adjusting the distance between the two permanent magnets [17]. It was also demonstrated that on a patterned substrate with ferromagnetic electrodes, the electrodes act as micromagnet arrays which can be interconnected by magnetic nanowires due to the dipole interaction. Magnetic nanowire cross junctions were realized by a two-step alignment as shown in Fig. 2c–d.

Static Self-Assembly of Multiple Components

For self-assembly of more complex structures than simple periodic structures or crystals, multiple components in a mixture must come together in an organized way [1]. Unlike the magnetic self-assembly, using single components such as beads and nanowires, Yellen and coworkers reported that 3-D flower-like colloidal structures with a rotational symmetry are formed using multiple components, i.e., nonmagnetic and paramagnetic colloids mixed in a ferrofluid consisting of magnetite nanoparticles suspended in water, when a static and uniform magnetic field is applied [18]. The results showed that the paramagnetic and nonmagnetic colloids interact anti-ferromagnetically in the multiple-component suspensions, leading to the formation of ring structures as shown in Fig. 3 [18]. This is because the colloids less magnetizable than the ferrofluid exhibit diamagnetic property, which implies that the particle moments prefer to align antiparallel with respect to the direction of the applied uniform field. Figure 3a, b show schematics of magnetic self-assembly from two and three colloidal components, respectively, which were mixed into ferrofluid. Figure 3c shows the experimental results of the fluorescence-labeled colloids, corresponding to the schematic in Fig. 3b, hierarchically self-assembling to ring-like structures with multipole symmetry configurations. Additionally,



Magnetic-Field-Based Self-Assembly, Fig. 3 (**a**) Illustration of magnetic assembly of two-component colloidal particle (*red* and *gray spheres*) mixed with ferrofluid (*brown dots*). No structures form without an external magnetic field. When a field is applied (*black arrows*), equatorial and polar arrangements form when the particle magnetizations, respectively, have the opposite and the same sense. Particle dipole moments are indicated by the *yellow arrows*. (**b**) Self-assembly in three-component suspensions is illustrated. "Two-tone" structures consist of simultaneous polar and equatorial arrangements assembling on a single core particle, whereas "flower" structures are formed by the fractal assembly of multiple equatorial arrangements. (**c**) Fluorescent images of multiple-component particle assembly in ferrofluid corresponding to the illustrations in (**b**). Scale bars are 20 μm (**a**–**c** Adapted with permission from Macmillan Publishers Ltd. Ref. [18]. Copyright 2008)



Magnetic-Field-Based Self-Assembly, Fig. 4 (**a**–**b**) Two forms of the alignment method using magnetic assembly. In *form A*, hard magnets are magnetized in the same direction before folding. In *form B*, soft magnets are magnetized to saturation during the folding (Reprinted with permission from Ref. [10]. Copyright 2006, American Vacuum Society). (**c**) Micromagnets array for self-assembly of two complementary parts (*part A* and *B*) (Reproduced with permission from Ref. [11]. Copyright 2007 IEEE). (**d**–**k**) Three-dimensional magnetic assembly from planar components. Three-dimensional structures (**e**, **g**, **i**, **k**) self-assembled from magnetically patterned sheets (**d**, **f**, **h**, **j**). The direction of the magnetic dipoles is indicated with *white arrows* (**d**–**k** Reproduced with permission from Ref. [19]. Copyright 2005 National Academy of Sciences, USA)

the concentration of the magnetic nanoparticles and the strength of the external magnetic field are two crucial factors for such a self-assembly process. In those tests, the applied uniform magnetic field was not higher than 100 Oe (in free air 100 Oe is equal to 10 mT).

An alternative approach to construct 3-D patterns is using the origami technique to fold planar components. The magnetic assembly approach is used for the alignment of the components with a high positioning precision and efficiency. Barbasathis et al. showed the self-alignment of thin films with nanomagnets on the surfaces [10]. The film are folded into 3-D configurations using a magnetic field. The nanomagnet arrays were prepared by lithographic patterning technique. Two types of alignment methods (form A and B) were investigated as presented in Fig. 4a, b, in which the magnets are magnetized before and after folding, respectively. There are two main advantages to perform the alignment after the folding process (i.e., form B): (1) the magnetic interaction forces are enhanced by one order of magnitude due to the applied field;

(2) since an external field is "ON" during the assembly process, it can be applied for a broader range of ferromagnetic materials and the shapes of the nanomagnets. Magnetic alignment tests of folding a freestanding Si_3N_4 membrane on its own substrate, both are patterned with 75 nm thick cobalt nanomagnets, are conducted, and the layer-to-layer alignment accuracy of approximately 250 nm was achieved [10]. The applied field strengths were in the range of several hundred milli-Teslas. The layer-to-layer magnetic self-assembly using permanent magnet arrays was reported by Arnold et al. to assemble different microscale components (see Fig. 4c), in which solder pads were designed for permanent bonding of the assembled parts [11]. Moreover, the results indicate that the magnetic bonding force is enhanced by using an array of micromagnets compared to a single piece of magnet with the same volume.

Spontaneous folding of flexible 2-D components to form freestanding hollow 3-D objects using magnetic self-assembly was demonstrated by Whitesides' group. To construct a spherical shell, they investigated elastomeric sheets with three different 2-D shape designs [19]: an equatorial cut (see Fig. 4d), an orange-peel cut (see Fig. 4f), and a flower-petal cut (see Fig. 4h, j). To self-fold the sheets, different patterns of magnetic dipoles generated by permanent magnets were examined. The self-assembly processes were performed in water or at the air-water interface with some minutes of agitation. The magnetic and mechanical interactions influence the self-assembly significantly [19].

Dynamic Self-Assembly at a Liquid-Air Interface

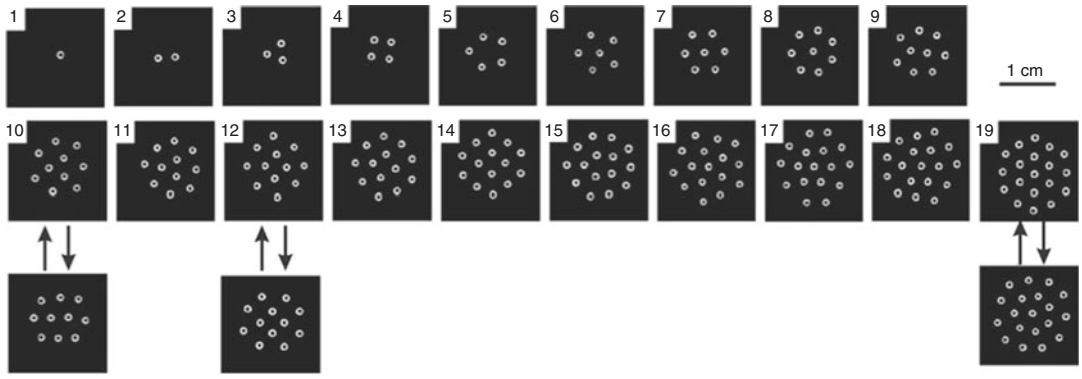
Self-assembly processes can be classified into two categories: static self-assembly and dynamic self-assembly [1]. The self-assembly approaches presented in the previous section all belong to the static self-assembly category. An assembly is defined as static when the ordered structure is in a stable equilibrium not dissipating energy [20]. Unlike the static self-assembly, in a dynamic self-assembly process, an ordered pattern/structure only occurs when it is dissipating energy [1]. Recently, dynamic self-assemblies were demonstrated using millimeter-size magnetic disks at a liquid-air interface. When a rotating magnetic field was applied by placing a rotating bar magnet underneath the container, the magnetic disks spun around their axis and created different dynamic patterns depending on the number of the disks. The dynamic self-assembly patterns are induced by the interplay between the attractive forces (magnetic interactions) and the repulsive forces (hydrodynamic interactions) [20]. In the case of a single magnetic disk, it stays at a stable position with its center coinciding with the rotational axis of the bar magnet, when the number of magnetic disks is varied from 2 to 19, various patterns are formed in a dynamic mode rotating at the liquid-air interface as shown in Fig. 5 [20]. More recently, Kwok et al. showed that, using a vertical (perpendicular to the liquid-air interface) alternating magnetic field, nickel microbeads with a diameter of 80–100 μm suspended at the liquid-air interface form complex 2-D patterns resulting from dynamic self-assembly [5]. The formation of the snake-like pattern is driven by the coupling of the hydrodynamic and magnetic interactions at the liquid-air interface. More detailed explanation of this dynamic assembly is shown in Ref. [5].

Examples of Application

Self-assembly is expected to be applied for crystallization at all scales, and in the fields of robotics and manufacturing, nanoscience and technology, microelectronics, and netted systems [1]. Numerous groups explored the applications using magnetic self-assembly, and some examples are given in the following section.

Artificial Microswimmers for Cargo Transport

Self-assembled magnetic micro-/nanobeads and nanowires have been investigated extensively for various applications. One example is the actuation of 1-D magnetic chains for individual cargo transport and manipulation of microfluid. Figure 2a shows that when a uniform magnetic field is applied, magnetic bead suspensions form 1-D



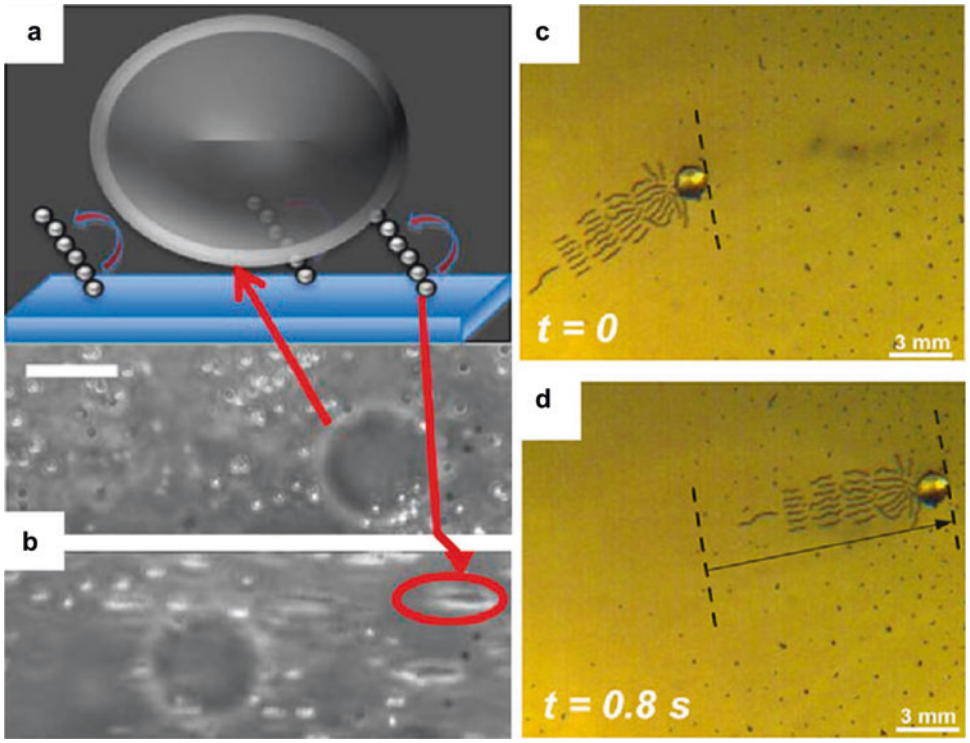
Magnetic-Field-Based Self-Assembly, Fig. 5 Dynamic patterns formed by various numbers (n) of magnetic disks rotating at the ethylene glycol/water-air interface. This interface is 27 mm above the plane of the external magnet. The disks are composed of a section of polyethylene tube (*white*) of outer diameter 1.27 mm, filled with poly (dimethylsiloxane), PDMS, doped with 25 wt% of magnetite (*black center*). All disks spin around their centers at $\omega = 700$ rpm, and the entire aggregate slowly ($\Omega < 2$ rpm) precesses around its center. For $n < 5$, the aggregates do not have a "nucleus" – all disks are precessing on the rim of a circle. For $n > 5$, nucleated structures appear. For $n = 10$ and $n = 12$, the patterns are bistable in the sense that the two observed patterns interconvert irregularly with time. For $n = 19$, the hexagonal pattern (*upper*) appears only above $\omega \approx 800$ rpm, but can be "annealed" down to 700 rpm by slowly decreasing the spinning rate. Without annealing, a less symmetric pattern exists at $\omega = 700$ rpm (Adapted with permission from Macmillan Publishers Ltd.: Ref. [20]. Copyright 2000)

chains. Upon rotation of the uniform field, the self-assembled chains rotate, resembling a microscale rotor, due to the induced magnetic torque [6]. Moreover, the rotating chains are capable of propulsion as they are near a solid surface. The change of motion from simple rotation to rotation plus translation results from wall effects. The drag on an object increases when it is closer to a solid boundary. This induces a drag force imbalance on the object resulting in propulsion along the surface in the direction of the rotating field. This type of motion is commonly referred to as tumbling motion. Rotating devices, such as self-assembled magnetic-bead-chains and magnetic nanowires, generates a localized flow around them, which implies that they can act as tools for microfluid and microobject manipulation [6, 7]. Controlled transport of dispersed microobjects such as vesicles near a solid boundary by self-assembled tumbling chains is shown in Fig. 6a, b [6]. Self-assembled magnetic structure can also perform micromanipulation at a liquid-air interface. In the previous section, it was shown that snake-like pattern can be created at the liquid-air interface by dynamic self-assembly approach. Kwok et al. reported the transport of individual 1–2 mm diameter glass or polystyrene beads using such a snake-like swimmer as shown in Fig. 6c, d. The self-propulsion of the surface microswimmer with the cargo is attributed to the breaking of symmetry of the surface vortex flow when a bead is placed near one end of the swimmer [5].

One advantage of utilizing self-assembled microswimmers for microfluid and microobject manipulation is that no complex microfabrication techniques are required, and moreover, the self-assembly and disassembly processes are reversible and controlled on demand [6].

Magnetophoretic Cell Manipulation

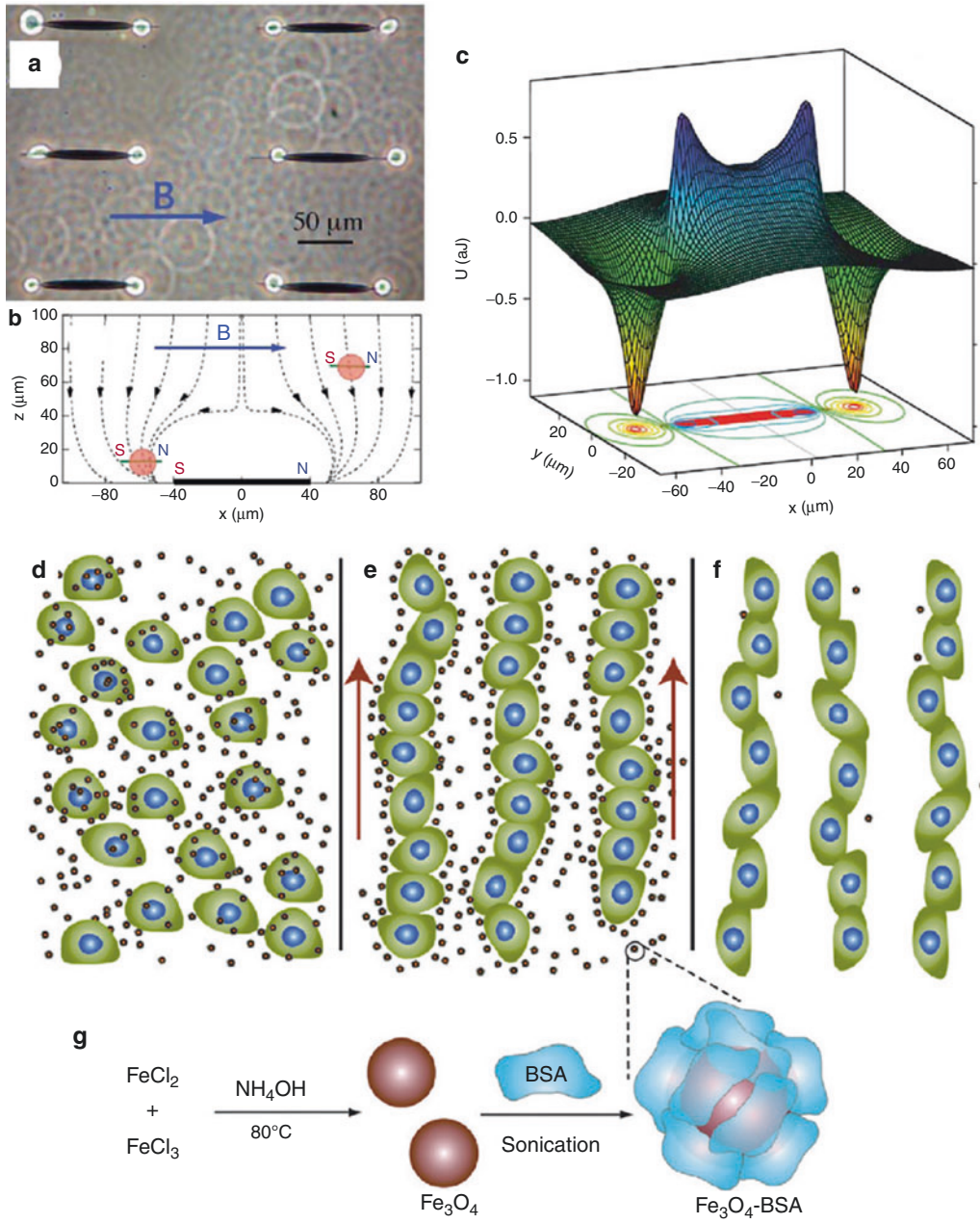
Low-strength magnetic fields have a minimal impact on living cells and biological tissues; therefore, they are promising for biomedical applications [15]. Among the different magnetic materials, magnetic nanoparticles and nanowires are interesting components for cell manipulation. In principle, two methods can be applied for magnetophoretic cell manipulations, i.e., positive and negative magnetophoresis. Positive magnetophoresis means that magnetic



Magnetic-Field-Based Self-Assembly, Fig. 6 (**a**–**b**) The presence of a flow field created by using the rotors enables the transport of large, dispersed objects like the vesicles (*large dark circle*). The flow is generated by using a carpet of surface walkers randomly positioned (*bright*, *short*, *chain-like regions* in **b**) actuated at 32 Hz. The snapshots in (**b**) are *top* view. The scale bar represents 20 μm, and the time between the two frames is 1.5 s (**a**–**b** Adapted with permission from Ref. [6]. Copyright 2010 National Academy of Sciences, USA). (**c**–**d**) Magnetic surface swimmer is composed of self-assembled 90 μm diameter Ni microbeads. The swimmer is capable of cargo transport, such as glass bead with a diameter of 1.5 mm, at a liquid-air interface using an alternating magnetic field. The input frequency is 80 Hz and the field strength is 10 mT (**c**–**d** Reprinted with permission from Snezhko, A., Belkin, M., Aranson, I. S., Kwok, W. K.: Phys. Rev. Lett., 102, 118103 (2009) (Ref. [5]). Copyright 2009 by the American Physical Society)

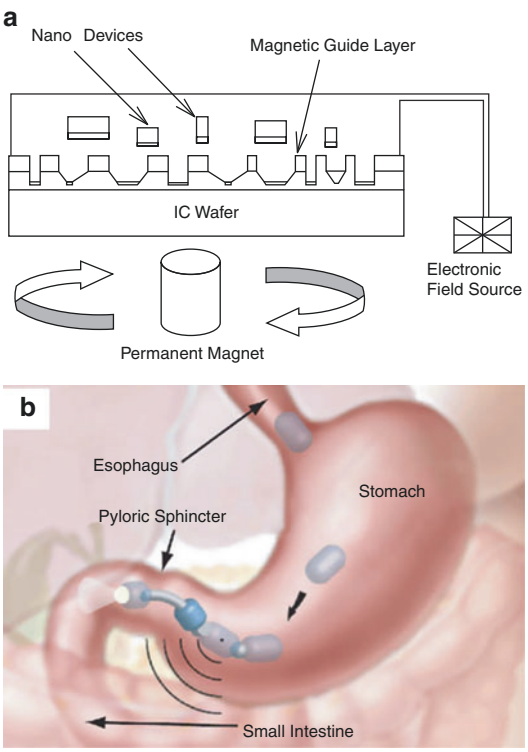
components are attracted toward the area with highest magnetic field intensity and are repelled from the area with lowest magnetic field intensity. For cell manipulation, the cells are tagged with magnetic components directly, resulting in higher permeability than the fluidic environment. The alternative approach for cell manipulation is to employ negative magnetophoresis, in which fluid, e.g., ferrofluid, with a high permeability is employed to manipulate cells. One example of 3-D colloidal structure assembly using negative magnetophoresis was shown in the last section (see Fig. 3). For cell manipulation, one advantage of the negative magnetophoresis is that it is capable of performing cell manipulation without binding the cell or uptake of the magnetic components by the cells [9]. The magnetophoretic manipulation can be achieved if the magnetic potential energy is significantly larger than the thermal fluctuation energy [9].

Figure 7a–c shows targeted patterning of 3 T3 mouse fibroblasts cells on a substrate using positive magnetophoresis. In the experiments, the cells were attached to ferromagnetic nanowires, and then self-assembled on the substrate pre-patterned with ellipsoidal permalloy micromagnet arrays. The results show that the cells are trapped separately at the two ends of the micromagnets forming an ordered 2-D pattern (see Fig. 7a). The schematics in Fig. 7b, c indicate



Magnetic-Field-Based Self-Assembly, Fig. 7 (**a**) Trapping of single cells by ellipsoidal micromagnet arrays. The field strength (*B*) is 2 mT. (**b**) Calculated settling trajectories for a spherical cell over the centerline of one of the ellipses in (**a**). (**c**) Calculated wire-ellipse interaction energy *U* at a wire height z = 3 μm. Ellipse footprint is shown in *red* in the *x*-*y* plane (**a**–**c** Ref. [4]. Reproduced with permission of the Royal Society of Chemistry). (**d**–**g**) Schematic of cell chain formation in ferrofluid using a uniform magnetic field. (**d**) Suspension of cells in ferrofluid assumes a random orientation in the absence of a magnetic field. (**e**) Suspended cells form linear arrangements in ferrofluid in the presence of magnetic field (*arrows*), where the ferrofluid particles shepherd the cells into chains due to their induced magnetic dipoles. (**f**) Linear arrangement of cells adherent to cell-adhesive surface survives and grows upon removal of ferrofluid and magnetic field. (**g**) Schematic of BSA-passivated nanoparticle synthesis (**d**–**g** Reprinted with permission from Ref. [9]. Copyright 2009 American Chemical Society)

that the two ends and the center of the micromagnet are the magnetic field intensity maxima and minima, respectively, therefore all cells are trapped at the two ends but never land at the center [4]. Three different species of cells, 3 T3, HeLa, and MCF10A breast cells, were employed in the cell manipulation tests. One example of cell manipulation using negative magnetophoresis is illustrated in Fig. 7d–f, in which the cells are self-assembled to form linear structures using ferrofluid. Krebs et al. demonstrated that using bovine serum albumin (BSA) functionalized magnetic nanoparticles in ferrofluid (see Fig. 7g) can efficiently avoid the potential cytotoxicity due to the attachment of particles onto the cells or the cellular uptake of the nanoparticles during the manipulation [9]. HUVECs (human umbilical vein endothelial cells) were employed in the tests, and it was reported that the assembled cellular chains are intact and stable after the removal of external magnetic field and ferrofluid. Organized cell structures can be utilized in tissue engineering research [9].



Magnetic-Field-Based Self-Assembly, Fig. 8 (**a**) A schematic of the magnetic-field-assisted assembly method of integrating multiple microcomponents on an IC substrate (Reproduced with permission from Springer Science + Business Media: Ref. [12]. Copyright 2008). (**b**) Artistic concept of a snake-type swallowable modular robot assembled in the stomach and moving into the intestine. Permanent magnets can be applied for self-assembly of each functional module (Reproduced with permission from Ref. [13]. Copyright 2008 IEEE)

Magnetic Assembly for Microelectronics and Microrobotics

Another promising application using self-assembly is to integrate various functional components for microelectronics. To date, robotic assembly, e.g., performing pick-and-place and alignment tasks using robotic arms, is still the main approach in industry for integration of microelectronic devices and systems. However, because the dimensions of individual components for an integrated circuit (IC) are miniaturized to a micrometer or even smaller scale, the throughput of robotic assembly is reduced significantly due to the high precision requirement in the assembly process. Moreover, van der Waals forces and other surface interactions can lead to a severe releasing problem in robotic assembly as the size of the component is reduced to a micro-/nanoscale. Therefore, magnetic self-assembly processes were explored to improve the throughput and accuracy of the assembly process for electronic devices.

One straightforward method is to pre-pattern the device surface with micromagnet arrays for the self-assembly and then interconnect the components permanently by fusing the solder pads as shown in Fig. 4c. A 3-D electrical circuit was constructed from planar components consisting of micromagnets, solder pads, copper wires, and six LEDs (light-emitting diodes), and the results show that all LEDs illuminate when a battery is connected [19]. It is notable that the reliability of the magnetic assembly can be enhanced by combination of other self-assembly method. For example, multiple-magnetic-components assembly on an IC wafer is shown schematically in Fig. 8a. The components have unique shapes that fit into the patterned surface. During the process, all components fall on the surface due to

gravity and with the help of a permanent magnet under the IC wafer, they are assembled and retained [12].

In the field of microrobotics, magnetic self-assembly is a promising approach to build modular robotic systems, which is particularly interesting for robotic endoluminal surgery [13, 14]. For minimal invasive medical diagnosis and treatment, swallowable capsule endoscopes were developed; however, the size and the functionality of the capsule are the two main challenges. One idea to overcome these issues is to swallow the individual modules one at a time and final assembly is performed inside gastrointestinal tract. This allows the size and the complexity of the final robot to be significantly greater than that of its individual modules [13]. If the modules are designed to be heterogeneous, each having a specific dedicated function, the difficulty in fabrication and packaging will be reduced, and relatively large devices could be included in the robotic system. As an example, Fig. 8b shows a snake-type robot with heterogeneous modules that assembles in the stomach and then moves through the intestine [13]. Magnetic alignment and assembly of such kind of robotic modules with permanent magnets have been investigated, and though the research is still in an infant stage, the recent development shows that the self-assembled modular capsules are a promising approach for the development of next-generation microsurgical devices [13, 14].

Cross-References

- ▶ Biosensors
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Nanorobotic Manipulation of Biological Cells
- ▶ Nanorobotics
- ▶ Self-Assembly
- ▶ Self-Assembly of Nanostructures

References

1. Whitesides, G.M., Grzybowski, B.: Self-assembly at all scales. Science **295**, 2418–2421 (2002)
2. Furlani, E.P.: Permanent Magnet and Electromechanical Devices: Materials, Analysis, and Applications. Academic, San Diego (2001)
3. Shetye, S. B.: Magnetic self-assembly of small parts. Ph.D. dissertation, Department of Electrical and Computer Engineering, University of Florida, Gainesville (2009)
4. Tanase, M., Felton, E.J., Gray, D.S., Hultgren, A., Chen, C.S., Reich, D.H.: Assembly of multicellular constructs and microarrays of cells using magnetic nanowires. Lab Chip **5**, 598–605 (2005)
5. Snezhko, A., Belkin, M., Aranson, I.S., Kwok, W.K.: Self-assembled magnetic surface swimmers. Phys. Rev. Lett. **102**, 118103 (2009)
6. Sing, C.E., Schmid, L., Schneider, M.F., Franke, T., Alexander-Katz, A.: Controlled surface-induced flows from the motion of self-assembled colloidal walkers. Proc. Natl. Acad. Sci. U. S. A. **107**, 535–540 (2010)
7. Zhang, L., Petit, T., Lu, Y., Kratochvil, B.E., Peyer, K. E., Pei, R., Lou, J., Nelson, B.J.: Controlled propulsion and cargo transport of rotating nickel nanowires near a patterned solid surface. ACS Nano **4**, 6228–6234 (2010)
8. Park, S.Y., Handa, H., Sandhu, A.: Magneto-optical biosensing platform based on light scattering from self-assembled chains of functionalized rotating magnetic beads. Nano Lett. **10**, 446–451 (2010)
9. Krebs, M.D., Erb, R.M., Yellen, B.B., Samanta, B., Bajaj, A., Rotello, V.M., Alsberg, E.: Formation of ordered cellular structures in suspension via label-free negative magnetophoresis. Nano Lett. **9**, 1812–1817 (2009)
10. Nichol, A.J., Arora, W.J., Barbastathis, G.: Thin membrane self-alignment using nanomagnets for three-dimensional nanomanufacturing. J. Vac. Sci. Technol. B **24**, 3128–3132 (2006)
11. Shetye, S.B., Agashe, J.S., Arnold, D.P.: Investigation of microscale magnetic forces for magnet array self-assembly. IEEE Trans. Magn. **43**, 2713–2715 (2007)
12. Rivero, R.D., Shet, S., Booty, M.R., Fiory, A.T., Ravindra, N.M.: Modeling of magnetic-field-assisted assembly of semiconductor devices. J. Electron. Mater. **37**, 374–378 (2008)
13. Nagy, Z., Oung, R., Abbott, J. J., Nelson, B. J.: Experimental investigation of magnetic self-assembly for swallowable modular robots (1915–1920). In: Proceedings of the 2008 IEEE/RSJ International Conferences on Intelligent Robots and Systems (IROS 2008), Nice (2008)
14. Harada, K., Susilo, E., Menciassi, A., Dario, P.: Wireless reconfigurable modules for robotic endoluminal surgery. In: Proceedings of the 2009 I.E. International Conference on Robotics and Automation, Kobe, pp. 1–7, 2699–2704 (2009)
15. Nelson, B.J., Kaliakatsos, I.K., Abbott, J.J.: Microrobots for minimally invasive medicine. Annu. Rev. Biomed. Eng. **12**, 55–85 (2010)
16. Tanase, M., Silevitch, D.M., Hultgren, A., Bauer, L. A., Searson, P.C., Meyer, G.J., Reich, D.H.: Magnetic trapping and self-assembly of multicomponent nanowires. J. Appl. Phys. **91**, 8549–8551 (2002)

17. Hangarter, C.M., Rheem, Y., Yoo, B., Yang, E.H., Myung, N.V.: Hierarchical magnetic assembly of nanowires. Nanotechnology **18**, 205305 (2007)
18. Erb, R.M., Son, H.S., Samanta, B., Rotello, V.M., Yellen, B.B.: Magnetic assembly of colloidal superstructures with multipole symmetry. Nature **457**, 999–1002 (2009)
19. Boncheva, M., Andreev, S.A., Mahadevan, L., Winkleman, A., Reichman, D.R., Prentiss, M.G., Whitesides, S., Whitesides, G.M.: Magnetic self-assembly of three-dimensional surfaces from planar sheets. Proc. Natl. Acad. Sci. U. S. A. **102**, 3924–3929 (2005)
20. Grzybowski, B.A., Stone, H.A., Whitesides, G.M.: Dynamic self-assembly of magnetized, millimetre-sized objects rotating at a liquid-air interface. Nature **405**, 1033–1036 (2000)

Magneto-optics and Plasmonics

► Magnetoplasmonics

Magnetoplasmonics

Giovanni Pellegrini[1], Valentina Bonanni[2], Giulio Campo[2], Francesco Pineider[2], Claudio Sangregorio[3], César de Julián Fernández[4], Francesca Casoli[4], Maria Grazia Manera[5], Roberto Rella[5] and Giovanni Mattei[1]
[1]Department of Physics and Astronomy, University of Padova, Padova, Italy
[2]Department of Chemistry, University of Florence, Florence, Italy
[3]CNR-ICCOM (Institute of Chemistry of OrganoMetallic Compounds – Italian National Research Council), Florence, Italy
[4]CNR-IMEM (Institute of Materials for Electronics and Magnetism – Italian National Research Council), Parma, Italy
[5]CNR-IMM (Institute for Microelectronics and Microsystems – Italian National Research Council), Lecce, Italy

Synonyms

Magneto-optics and Plasmonics

Definition

Magnetoplasmonics (MP) studies the effects arising from the interplay between plasmonic and magneto-optical (MO) phenomena typically occurring in metallic nanostructures. Plasmonic field enhancement allows for large MO activity while magnetic functionalities give access to independent control of plasmonic properties by tailoring intensity and direction of an external magnetic field.

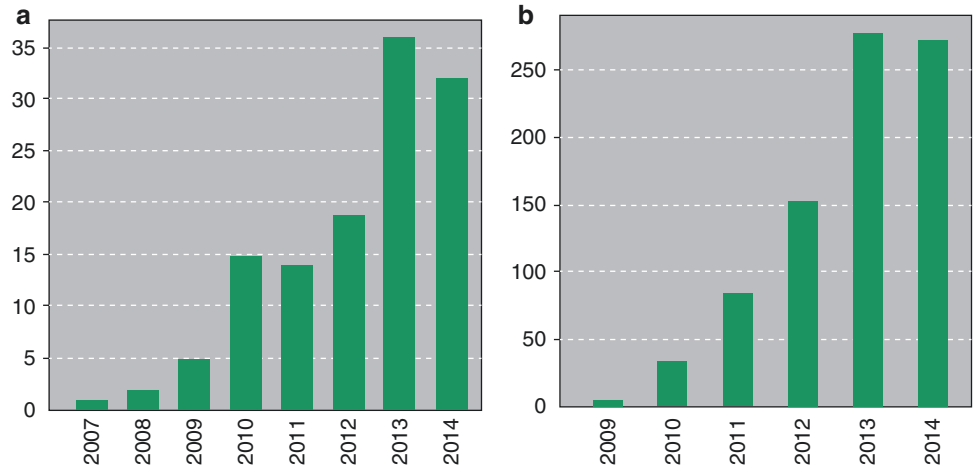
Introduction

Overview

Metallic nanostructures supporting Surface Plasmon Polaritons (SPP) and Localized Surface Plasmons (LSP) are characterized by their unique ability to control and manipulate light at the nanoscale [1]. Yet the necessity to realize nanometric integrated photonic components, such as switches or modulators, requires finding a way to control the optical properties of a nanostructure by means of an external factor. If high switching and modulation speeds are required, control mechanisms resting on temperature and voltage change must be discarded, and the all-optical option remains the main viable solution. Nevertheless, a compelling alternative to pure photonic control for active nanophotonics and plasmonics may be found in the adoption of a magnetic field modulation, which allows tailoring of optical properties as a function of the magnetic field intensity and direction. Performances comparable to all-optical configurations are expected, since magnetism can be inherently considered an ultrafast property [2, 3].

In the following we point our attention, both from the theoretical and experimental viewpoints, to magnetoplasmonics, i.e., to the entwined effects emerging from the mutual interaction of plasmonic and MO phenomena. This is a very fascinating and fast-growing research field as indicated by the bibliometric figures related to the word "magnetoplasmonics" reported in Fig. 1.

The present section is dedicated to a brief introduction to plasmonics, magneto-optics (Kerr and Faraday effects, in particular) and optimized



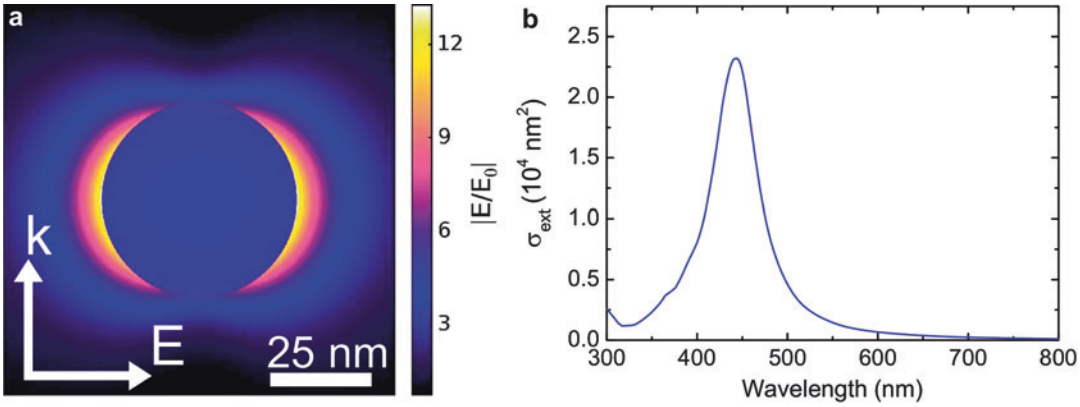
Magnetoplasmonics, Fig. 1 Number of (**a**) published papers and of (**b**) citations for the topic "magnetoplasmonics" (Source Web of Science)

solutions for MP materials. The following two sections will be devoted to the description and understanding of MP phenomena emerging in extended (SPP) and localized (LSP) structures. Finally, the last part of this contribution will focus on actual and potential applications, with special consideration to biosensing and theranostics (i.e., the combination of therapeutic plus diagnostic nanomedicine).

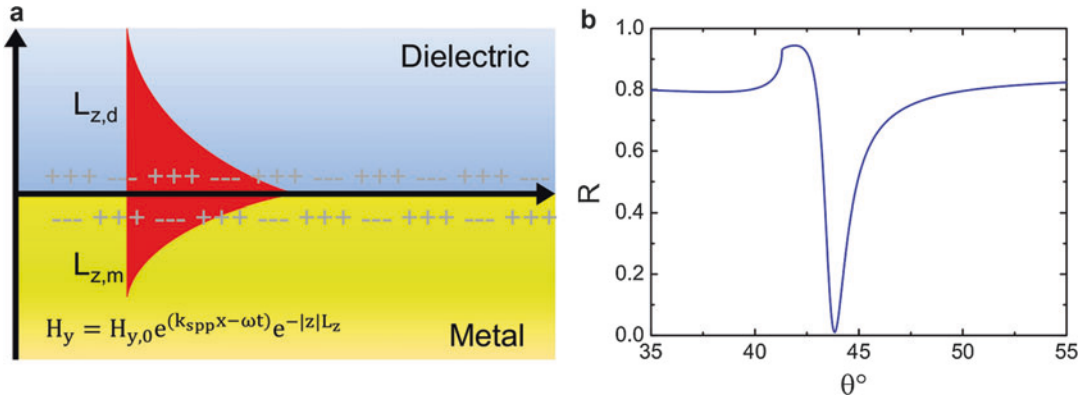
Surface Plasmon Polaritons and Localized Surface Plasmons

Surface Plasmons (SPs) are defined as surface bound electromagnetic waves coupled to collective oscillations of free charges (typically free electrons in metals). They are supported at the interface between different media, with the mandatory condition that the involved media must exhibit dielectric constants with real parts of opposite sign (e.g., a metal and a dielectric). Paramount characteristic of SPs is their strong dimensional localization at the interface; this is in practice observed for a large variety of metallic (especially noble metals) structures, since metals are often characterized by optical constants with a strongly negative real part at optical frequencies. Surface plasmons are best classified in two main categories, depending on the dimensionality of the supporting structure: LSPs are plasmon oscillations typically supported by 0-dimensional (0D) or 1-dimensional (1D) structures, i.e., by structures confined to a sub-wavelength scale in two or three spatial dimensions, such as noble metal nanoparticles and nanowires (Fig. 2a). On the other hand, SPPs are supported at planar or nanostructured interfaces such as thin films or gratings (Fig. 3a). Regardless of the structure dimensionality, the charge oscillations typical of SPs are able to focus large electromagnetic fields within nanoscale volumes well beyond the diffraction limit. Direct consequence of the field localization and enhancement is a strongly increased light-matter interaction which reveals itself in a twofold way: on one hand, large field localization allows strong interactions with molecules, light emitters and analytes placed in the field hot-spots, while conversely small perturbations of the local refractive index manifest as a sizable modification of the far-field optical properties, making nanostructures supporting SPs an ideal platform for the development of sensing applications.

The optical fingerprint of a SPP is, for instance, a sharp dip in the reflection spectrum, corresponding to a condition where energy and momentum matching between incident radiation and the SPP wave are simultaneously satisfied, as displayed in Fig. 3b. The extra momentum needed for the photon-SPP coupling is obtained in attenuated total reflection (ATR) configuration, where



Magnetoplasmonics, Fig. 2 Near-field and far-field properties of a typical plasmonic system: a 50 nm Ag nanoparticle in silica. (**a**) Dipolar local field at LSPR, (**b**) extinction cross-section



Magnetoplasmonics, Fig. 3 Near-field and far-field properties of plasmonic systems: (**a**) sketch of the field distribution for a Surface Plasmon Polariton at a metal-dielectric interface, (**b**) reflectance for a typical Au thin-film SPP

the momentum matching is provided by a glass prism (e.g., Kretschmann coupling), or with a grating coupler, where the phase matching is assisted by the periodic structure. Given the strong SPP wavevector dependence on the surrounding dielectric medium, small changes of the local environment cause large momentum changes that are visible as resonance shifts in the reflection spectrum. In this context a resonance redshift corresponds to an increase of the surrounding refractive index, whereas the opposite is true in the case of a blue shift.

The far field optical response of LSPs is associated with an intense and narrow absorption and/or scattering of the light (Fig. 2b): indeed, noble metals nanoparticles, like Au and Ag, embedded in a solution or in a glass exhibit very bright colors. Historical examples can be found in ancient glass or artifacts, where the light produces different colors depending on the direction of the observer. Moreover, induced LSP resonances (LSPRs) give rise to strong amplification and confinement of the electric field at subwavelength distances from the plasmonic surface. These properties thus make LSP based devices very promising for advanced technological applications, such as the realization of sensors and nanoantennas. The LSP features are affected by several parameters such as material, size, morphology, coating, and environment. Indeed, the interaction between

light and nanoparticles involves both absorption and scattering mechanisms, and the ratio between them depends critically on the size of the nanostructures (e.g., larger sizes produce larger scattering). Moreover, a change of the refractive index of the surrounding medium has a relevant effect in shifting the energy of the LSP resonance: normally, increasing the environment refractive index produces a redshift of the LSP wavelength. Nevertheless, the direction of this shift can be influenced by the chemical environment of the particle surface and both blue- and red-shifts have been experimentally observed. In addition, the amplitude and the width of LSP resonances are strictly related to the electronic damping described by the imaginary part of the dielectric function. All these effects provide different degrees of freedom for the design of custom LSP plasmonic materials.

The LSP phenomenon is not limited to spherical nanoparticles: it can be observed also in other confined geometries, such as ellipsoids, spheroids, disks, rods, wires, where, due to the lower symmetry, different spectrally separated plasmon resonances can be excited, corresponding to oscillations of its conduction electrons along different axes of the system.

Magneto-Optical Effects

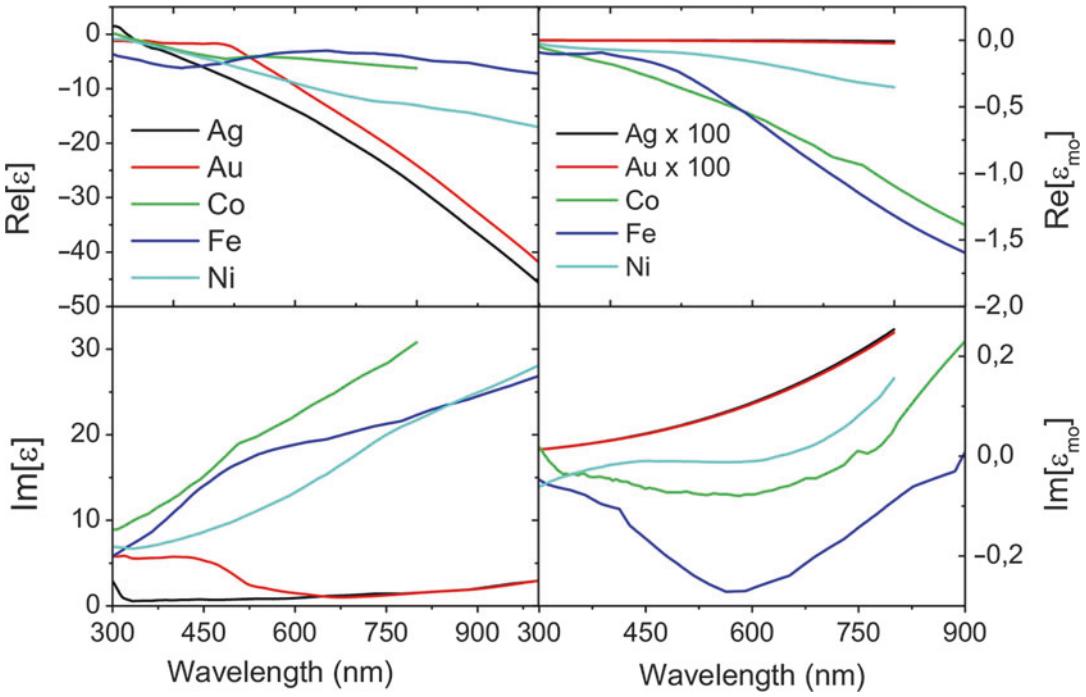
A magnetic field can modify the optical properties of materials. The modification of the light–matter interaction is detected as a variation of the polarization state or of the intensity in the transmitted (Faraday effect) or reflected (Kerr effect) light. The optical response of the medium to the external excitation of the incident light can be described by the dielectric tensor or permittivity, whose general form in a magnetic medium is [3]:

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon & \varepsilon_{xy} & \varepsilon_{xz} \\ -\varepsilon_{xy} & \varepsilon & -\varepsilon_{yz} \\ -\varepsilon_{xz} & \varepsilon_{yz} & \varepsilon \end{pmatrix} = \begin{pmatrix} \varepsilon & iQ_z & iQ_y \\ -iQ_z & \varepsilon & -iQ_x \\ -iQ_y & iQ_x & \varepsilon \end{pmatrix}. \quad (1)$$

The diagonal terms correspond to the (complex) isotropic dielectric function of the material in absence of an external magnetic field, while the off-diagonal terms are expressed as $\varepsilon_{mo} = iQ_j$, with Q_j the magneto-optical constants of the material. If dealing with paramagnetic (at room temperature) and diamagnetic materials (for which a linear relationship $\boldsymbol{M} = \chi_m \boldsymbol{H}$ exists, where χ_m is the magnetic susceptibility) the Q_j term is proportional to the j-component of the applied magnetic field, $\boldsymbol{H}$, while in the case of ferromagnets (or ferri- or antiferro magnets), for which the relationship between $\boldsymbol{M}$ and $\boldsymbol{H}$ is nonlinear and saturable, the spin–orbit coupling makes Q_j proportional to the j-component of the magnetization, $\boldsymbol{M}$.

If we first focus on the magnitude of the off-diagonal terms, which as stated above depends both on the external magnetic field and the intrinsic material properties, the importance of finding and exploiting materials with optimal MO properties in order to obtain sizable MO effects is immediately clear. In most cases, spin–orbit coupling is the mechanism responsible of the MO properties of a material: as a rule, the stronger the spin–orbit interaction the larger is the MO activity. In practice, under the effect of an external incident electromagnetic field, the spin–orbit interaction couples the spin of the electron to its motion, i.e., to its linear momentum. Furthermore, the spin also interacts with the internal magnetic field of the material thus producing a correlation between magnetic and optical properties of a magnetic material.

Quantum mechanics describes more precisely the MO effects in terms of transition probability induced by the external electromagnetic field between two states, as described by the Fermi Golden rule [3] and controlled by the dipolar selection rules for the optical transitions ($\Delta l = \pm 1$). Most magnetic metals, characterized by a large spin–orbit coupling, display very complex and rich MO spectra as different interband and intraband transitions are allowed in the Vis-NIR range. However, metals strongly absorb light in this range and the effective MO effect is limited to some tens of nanometers of light penetration from the surface. On the other hand, for magnetic oxides the MO transitions occur mainly between localized



Magnetoplasmonics, Fig. 4 Real and imaginary part of optical and MO constants for ferromagnetic and noble metal materials (All the optical constants are plotted from experimental data [2] and refs. therein), with the exception of Au and Ag MO constants (computed at 1T external magnetic field)

bands (nearby the Fermi level) and weak MO effects are generally observed, mainly triggered by crystal field or charge transfer interactions in which selection rules are partially broken.

The MO properties of noble metals are promptly described in the classical framework of the Drude model, since their physical origin rests mostly on the properties of the conduction band electrons. Noble metals MO optical constant may indeed be formulated as [2]:

$$\varepsilon_{mo} = i\frac{\omega_c(\omega_p\tau)^2}{\omega\left[(1-i\omega\tau)^2+(\omega_c\tau)^2\right]}, \qquad (2)$$

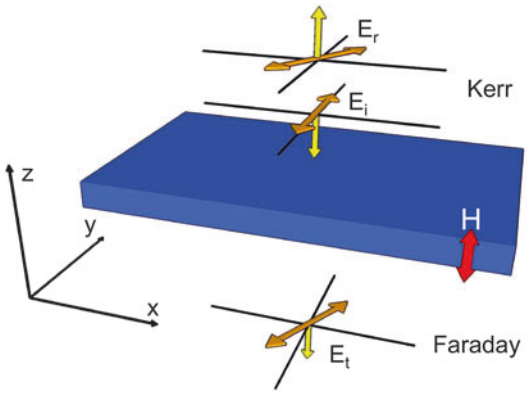
with $\omega_c = eH/m^*$ being the cyclotron frequency, $\omega_p = \sqrt{Ne^2/\varepsilon_0 m^*}$ the plasma frequency, N the electronic density, m^* the effective electron mass, e the electron charge, ε_0 the vacuum permittivity, H the external magnetic field and τ the electron relaxation time. Plugging into Eq. 2 realistic values for all the physical parameters in the case of Au and Ag, one might see that the obtained MO constants are roughly three orders of magnitude smaller than those of ferromagnetic materials such as Co, Ni, and Fe (Fig. 4 and Table 1).

If we now consider the tensorial nature of the MO dielectric constant, it is clear that not only the intensity, but also the direction of the applied magnetic field plays a fundamental role in the system optical response. In this context, MO effects are better understood if schematized as in Fig. 5. The material surface (xy plane) is illuminated by an electromagnetic radiation (linearly or circularly polarized) and the MO effects can be observed either in transmitted or reflected light. The mutual orientation among the plane of incidence (xz), the light polarization and the magnetic field directions is the distinguishing feature to classify the MO effects.

The first considered configuration, represented in Fig. 6, is characterized by z-oriented (normal to the xy plane) magnetic field. In this case the field is

Magnetoplasmonics, Table 1 Optical constants at 633 nm for typical plasmonic and magnetic metals. Both diagonal (dielectric) and off-diagonal (magneto-optic) components are reported. MO values are calculated at 1T of external field (Au, Ag) or measured at saturation (Ni, Fe, Co)

Element	$Re(\boldsymbol{\varepsilon})$	$Im(\boldsymbol{\varepsilon})$	$Re(\boldsymbol{\varepsilon}_{mo})$	$Im(\boldsymbol{\varepsilon}_{mo})$
Au	−11.74	1.26	-8.68×10^{-5}	1.23×10^{-3}
Ag	−18.28	0.48	-2.74×10^{-5}	1.26×10^{-3}
Ni	−12.93	16.36	−0.19	-1.26×10^{-3}
Fe	−1.03	17.74	−0.71	−0.24
Co	−12.49	18.45	−0.64	−0.067



Magnetoplasmonics, Fig. 5 Sketch of a Kerr and Faraday MO effects in case of a linearly polarized plane wave illuminating a MO active slab

also normal to the sample surface and thus this configuration corresponds, in reflection, to the so called Polar Kerr configuration (PMOKE). In this setup all the off-diagonal components of the dielectric tensor vanish, with the exception of the ε_{xy} and ε_{yx} terms which in this case provide cross-coupling between s ($\boldsymbol{E}$ field perpendicular to xz plane) and p ($\boldsymbol{H}$ field perpendicular to xz plane) polarization states. The direct consequence is that a linearly polarized incident plane wave is reflected or transmitted as an elliptically polarized beam, and therefore the measurable quantities are the beam rotation (ϑ) and ellipticity (ϵ), represented as the complex Kerr rotation $\Phi = \vartheta + i\epsilon = r_{ps}/r_{pp}$, where $r_{ps}(Q)$ and r_{pp} are the Fresnel reflectivity coefficients.

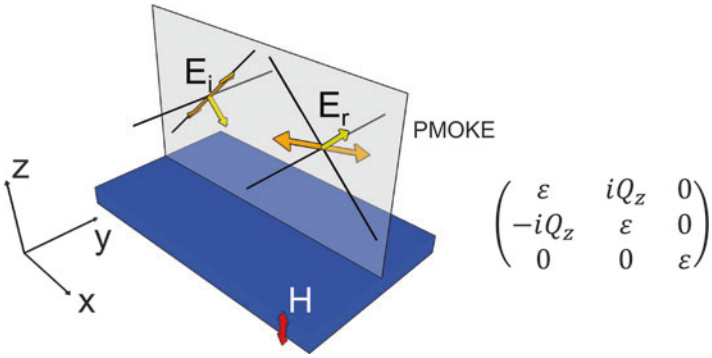
An analogous effect is seen for an x-oriented (normal to the yz plane) external magnetic field (Fig. 7), where a reflection configuration of the present arrangement is known as a Longitudinal Kerr setup (LMOKE). The coupling between s and p polarization states, provided in this case by the only nonvanishing off-diagonal coefficients ε_{yz} and ε_{zy}, finally leads to MO effects similar to what observed in the PMOKE case, where the complex Kerr rotation represents the principal observable.

Slightly different is the case of an y-oriented (normal to the xz plane) external magnetic field, also known as transverse Kerr configuration in a reflection setup (TMOKE) and Voigt or Cotton-Mouton configuration in a transmission one (Fig. 8). In this arrangement no polarization coupling is provided by the nonzero terms ε_{xz} and ε_{zx}, the only visible effect being a variation of the reflectance or transmittance of p-polarized incident light.

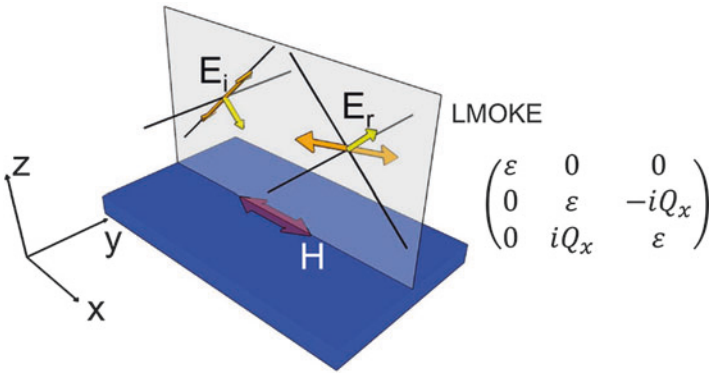
A qualitatively different MO phenomenon is finally represented by the Magnetic Circular Dichroism (MCD), which is defined as the difference in absorption (or reflection) between two spectra acquired using light with opposite helicity in the presence of an external magnetic field (usually but not necessarily parallel to the incident light direction, as depicted in Fig. 9).

Materials Viewpoint

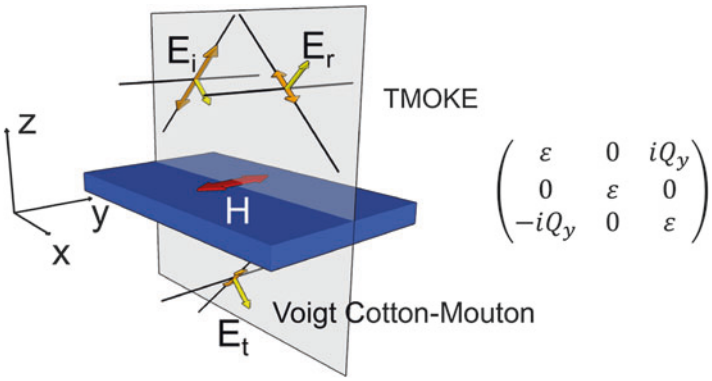
Actually there are no materials that simultaneously exhibit optimal plasmonics features and important magnetic and magneto-optical properties. If we first consider the magnetic properties of the most common plasmonic materials, Au and Ag, it emerges that they are diamagnetic in nature and thus huge unphysical magnetic fields (in the order of thousands of Teslas) are required to induce magnetizations and magneto-optical effects comparable to those of standard magnetic metals or oxides [12]. If instead we focus on the plasmonic properties of efficient magnetic materials, only the metallic ones exhibit plasmonic resonance in the Vis-NIR range. Nevertheless



Magnetoplasmonics, Fig. 6 Sketch of a typical PMOKE configuration and corresponding dielectric tensor



Magnetoplasmonics, Fig. 7 Sketch of a typical LMOKE configuration and corresponding dielectric tensor

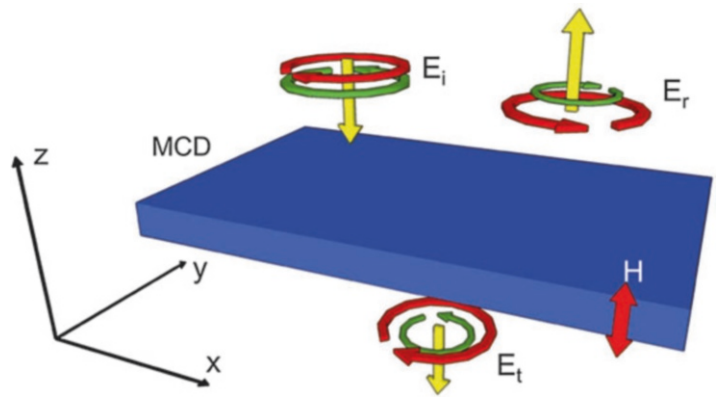


Magnetoplasmonics, Fig. 8 Sketch of typical TMOKE (reflection) and Voigt Cotton-Mouton (transmission) configurations and corresponding dielectric tensor

both interband and intraband transitions are allowed in the same range, opening relaxation channels that damp the plasmon resonance through nonradiative processes, essentially heat.

Table 1 summarizes the optical constants at 633 nm of typical plasmonic (Au, Ag) and magnetic (Fe, Co, Ni) metals [2, 4]. While all metals exhibit $Re(\varepsilon) < 0$ required for plasmonic resonances to occur, the $Im(\varepsilon)$ values are quite large for the magnetic materials, naturally explaining the strong SPR damping.

Hybrid nanostructures built upon noble and ferromagnetic metals building blocks represent at the moment the state of the art in terms of magneto-plasmonic structures. Nevertheless, a very promising solution for achieving both plasmonic and magnetic functionalities required by magnetoplasmonics can be the growth of a



Magnetoplasmonics, Fig. 9 Sketch of a typical MCD configuration. *Left* and *right* circular polarizations are transmitted and reflected with different intensity

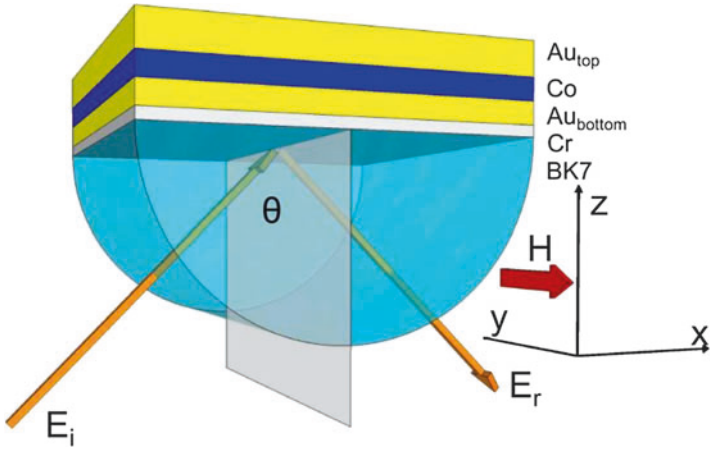
solid solution between magnetic transition metals and noble metals. However, except for the Ni-Au and Ni-Ag alloys, the two classes of metals are in general immiscible as bulk phases. Metastable solid solutions can be obtained by techniques like ion implantation, sputtering or laser ablation, yet achieving simultaneous plasmonic and MO properties in structures such as Au-Fe nanoparticles still represents a significant challenge: for instance, Au_{60}-Fe_{40} alloy nanoparticles obtained by sequential ion implantation in silica [5] exhibit a quenching of the LSPR (which is regained upon thermally induced segregation between Au and Fe) and a ferromagnetic order at room temperature. The coupling between optical and magnetic properties stems as a strong MCD signal and a photoswitchable magnetization upon irradiation with photons with wavelength of 514 nm (i.e., close to the region of Au SPR). Hence it is clear that, in spite of its importance, the research of single magnetoplasmonic metals remains a mostly unsolved challenge for the future while, as it will be shown, almost all the current studies focus their attention on hybrid nanostructures composed of magnetic and plasmonic moieties.

Magnetoplasmonics in Nanosystems Supporting Surface Plasmon Polariton Resonances

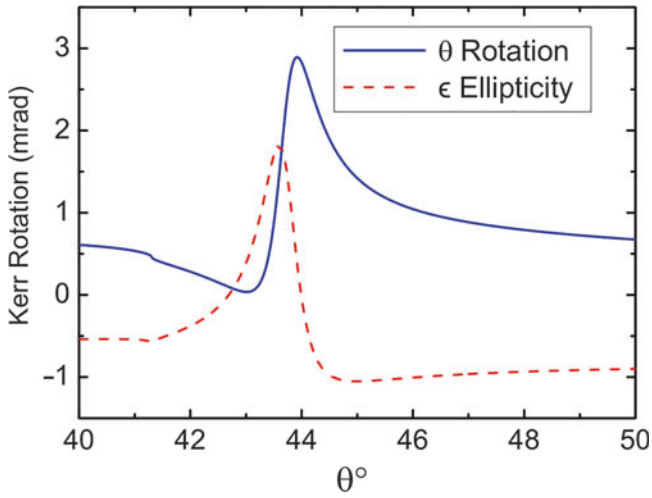
Multilayered Thin Films

In the present section we are going to discuss, from a theoretical viewpoint, the MP properties of structures supporting extended SPPs. Given the vast variety of structures capable of supporting this kind of excitation, we shall focus our attention only on a specific one, namely a noble metal-ferromagnetic multilayer, where a thin layer of Co is sandwiched between two Au layers and illuminated at 633 nm, with p or circularly polarized light, in an ATR geometry (Glass/Au/Co/Au/Air, Fig. 10). Such a structure has the advantage of being simple to analyze and study, and yet presents a richer phenomenology than that of a simple ferromagnetic or noble metal thin film. In this discussion the attention will be focused on the effect of MP phenomena on observable quantities like Kerr rotation and variations in reflectivity in PMOKE, LMOKE, and TMOKE configurations, and finally an MCD setup will be also analyzed. The effect of SPs on the MO properties will be studied first, then the modulation of SPP propagation upon change of the MO properties of the system will be scrutinized.

The PMOKE configuration is an ideal setup to analyze the effect of a SPP on the MO properties of the structure since the preferred observable in this case, the complex Kerr rotation Φ, is expressed as $r_{ps}(Q)/r_{pp}$ where r_{pp} and $r_{ps}(Q)$ are again the first two Fresnel coefficients [3]. From a quick glance to the previous expression it is clear that the excitation of a SPP is able to modulate the magnitude of Φ by strongly modifying the r_{pp} term value. In ideal light-plasmon coupling conditions we indeed have $r_{pp} \to 0$, therefore obtaining a relevant enhancement of Kerr rotation and ellipticity, as it may be seen in Fig. 11.



Magnetoplasmonics, Fig. 10 Sketch of a typical Au/Co/Au multilayer on glass substrate (e.g., BK7) in Attenuated Total Reflection Configuration (*ATR*). The Cr layer is an adhesion layer for Au on glass



Magnetoplasmonics, Fig. 11 Kerr rotation (θ) and ellipticity (ϵ) for an Au/Co/Au multilayer illuminated in p polarization at 633 nm in PMOKE configuration

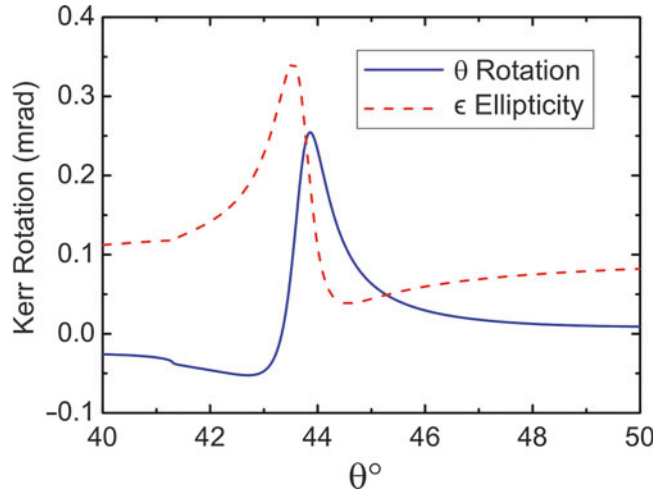
It is appropriate to underline that in this context the SPP contribution to the MO activity is purely optical, meaning that no leverage of the MO properties of the ferromagnetic material is needed to obtain the enhancement. On the other hand, for our multilayer structure the numerator term $r_{ps}(Q)$ may be expressed as [2]:

$$\left|r_{ps}(Q)\right| \propto \langle E_p E_s \rangle d \left|\varepsilon_{mo}(Q)\right| \qquad (3)$$

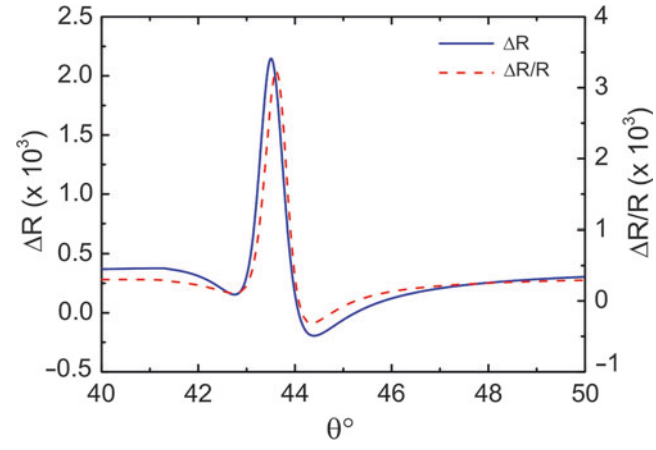
where $\langle E_p E_s \rangle$ stands for the mean value of the product of both components of the field, d is the Co layer thickness and $|\varepsilon_{mo}(Q)|$ is the MO optical constant of the ferromagnetic material. Thus, in the presence of a SPP the local-field enhancement at the metal-dielectric interface is going to boost the $\langle E_p E_s \rangle$ factor and as a consequence the MO activity and the complex Kerr ellipticity.

The overall MP performance of the system is finally determined by a delicate balance of the SPP optical r_{pp} and MO r_{ps} contributions, keeping in mind that strong plasmon resonances demand for low optical losses and therefore low Co amounts, for the multilayer considered, while on the contrary large MO contributions require a large Co thickness and field intensity inside the ferromagnetic layers, and thus high optical losses.

In close analogy to what described in the previous MO section, a similar situation is found for magnetoplasmonic LMOKE configurations. The original discussion of the PMOKE arrangement



Magnetoplasmonics, Fig. 12 Kerr rotation (θ) and ellipticity (ϵ) for an Au/Co/Au multilayer illuminated in p polarization at 633 nm in LMOKE configuration



Magnetoplasmonics, Fig. 13 Differential reflectance ΔR and ΔR/R for an Au/Co/Au multilayer illuminated in p polarization at 633 nm in TMOKE configuration

also applies to the present case, where an enhancement of the MO signal is seen in correspondence of the SPP resonance, as displayed in Fig. 12.

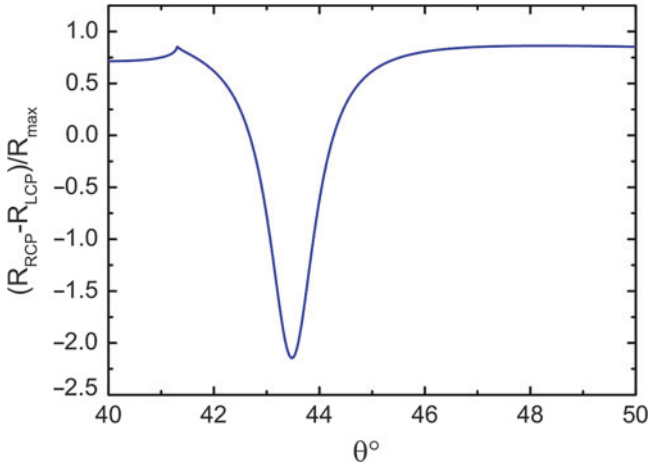
In the case of TMOKE configuration, the observable is defined as:

$$\Delta R = R(+H) - R(-H), \tag{4}$$

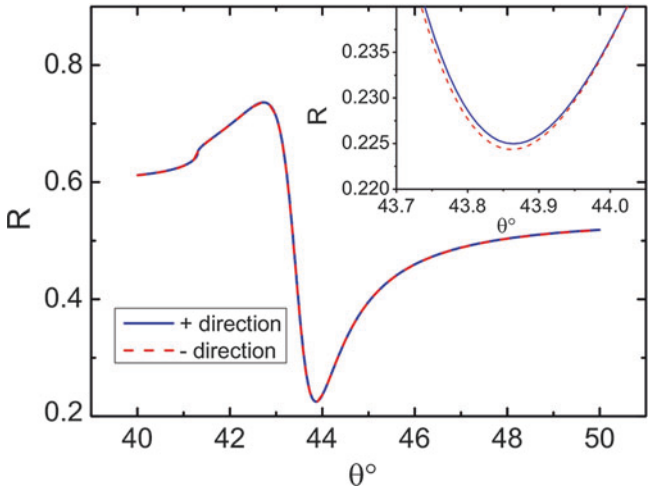
with $\boldsymbol{H}$ the applied magnetic field. In this case ΔR (Fig. 13) corresponds to a purely MO term, i.e., its magnitude is directly proportional to the ferromagnetic layer thickness and the intensity of the local-field inside that layer. An analogy to the previous cases may be recovered by introducing a denominator defined as $R = R(+H) + R(-H)$. Such a normalization factor takes small values in the presence of an optimal light plasmon coupling and therefore boosts the magnitude of the normalized TMOKE signal defined as $\Delta R/R$.

This same Kerr configuration is also suitable for monitoring the presence of plasmon enhanced MCD signals when the Au/Co/Au multilayer is illuminated with circularly polarized light. The MCD signal reported in Fig. 14 indeed shows a MP fingerprint as a spectral dip around 44°, the SPP coupling angle, where the MCD signal reaches its maximum in modulus. It is finally clear from the previous examples how SPP may influence and enhance the MO properties of a system, and how a balance of MO and SPP properties must be achieved in order to maximize the overall MP properties.

The adopted model multilayer structure is also ideal to study the magnetic field effects on SPPs.



Magnetoplasmonics, Fig. 14 MCD signal (defined as the normalized difference of reflectance for *right* and *left* circular polarization) for an Au/Co/Au multilayer illuminated at 633 nm

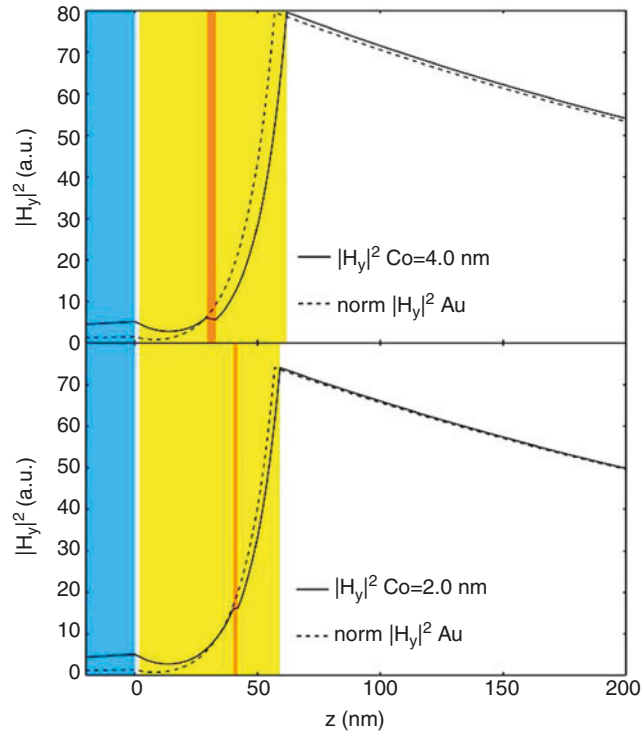


Magnetoplasmonics, Fig. 15 Reflectance spectra for an Au/Co/Au multilayer computed for the same incident angle but opposite incident directions in TMOKE configuration. The MO effects, and the consequent absence of local reciprocity, cause the k_{SPP} wavevector modulation and finally a modification of coupling angles for opposite incident directions, as seen in the inset

This phenomenology is understood by reminding that an external magnetic field modifies the optical constants of a material according to Eq. 1. Thus, in the framework of Maxwell equations, different optical constants imply a modification of the boundary conditions and finally of the electromagnetic field distribution and SPP propagation constant k_{SPP}.

More explicitly, dealing first with the TMOKE configurations for simplicity, these effects produce different SPP coupling conditions, and therefore different coupling angles, in presence of an external magnetic field. Moreover, in the absence of local reciprocity, equal incident angles but opposite incident directions lead to different coupling conditions and thus to slightly different reflectance spectra as seen in Fig. 15. Formally, the modulation of the SPP wave vector may be expressed as $k_{SPP} \sim k_{SPP,0} + \Delta k_{SPP}(Q)$, and the nonreciprocity and consequent different coupling conditions may be written as $k_{SPP,0} + \Delta k_{SPP}(Q) \neq -k_{SPP,0} + \Delta k_{SPP}(Q)$. It is finally interesting to note that for our multilayer structure the wave vector modulation can directly be written as a function of the local magnetic component of the electromagnetic field as [2]:

$$\Delta k_{SPP} \sim \frac{k_0 \varepsilon_d \varepsilon_{Au} \sqrt{-(\varepsilon_d + \varepsilon_{Au})}}{{\varepsilon_d}^2 - {\varepsilon_{Au}}^2} \times \frac{i\, \varepsilon_{mo,Co}}{\varepsilon_{Co}} \left| H_y^{Co} \right|^2, \quad (5)$$



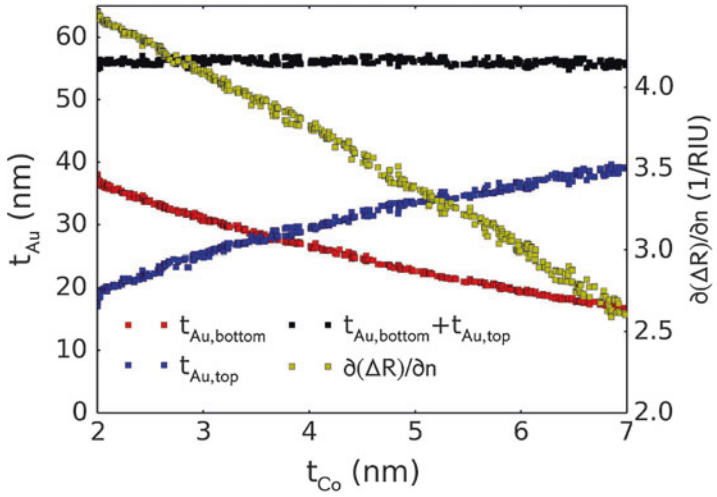
Magnetoplasmonics, Fig. 16 Magnetic field component of the local-field distribution for two different Au/Co/Au multilayers (solid lines) and Au layers (dashed lines) in ATR configuration. The SPP excitation contributes to a strong local-field enhancement in the ferromagnetic layer (From Ref. [9])

where k_0 is the light wavevector in vacuum, ε_d, ε_{Au} and ε_{Co} are the glass, Au and Co optical constants, respectively, $\varepsilon_{mo,Co}$ is the MO optical constant of Co and H_y^{Co} is the magnetic field in the Co layer, which varies strongly depending, among other things, on the Co layer position and thickness (Fig. 16).

Now focusing on PMOKE and LMOKE configurations, the first effect of a magnetic field on the SPP propagation is to introduce a Transverse Electric (TE) component in an otherwise purely Transverse Magnetic (TM) wave, in close analogy with the *s* and *p* polarization coupling outlined in the preceding paragraph. It may also be shown that the SPP wavevector k_{SPP} undergoes a more complex modulation that is quadratic in the term Q and may be finally be expressed as $k_{SPP} \sim k_{SPP,0}, + \beta Q^2$, where β is a proportionality constant and $k_{SPP,0}$ is again the unperturbed wavevector [2].

An elegant demonstration of the modulation of the SPP wavevector by an external applied magnetic field has been performed using a slit-groove plasmonic interferometer [6]: this technique potentially allows nanometre-resolved measurements of the electromagnetic field distribution inside a metal at optical frequencies. Momentum modulation of 2 % can be reached with nonoptimized Au/Co/Au multilayers and a further improvement of interface quality and the use of a high dielectric covering layer [7] increases the modulation up to a value of 12 %, comparable to the 15–20 % ones of other nonmagnetically active plasmonic modulators. Finally, Belotelov et al. [8] successfully demonstrated the magnetic field modulation of reflected light in magnetoplasmonic crystals consisting of an Au grating deposited on a magnetic garnet film. All the abovementioned experiments constitute cornerstone examples of magnetic-field-driven active plasmonics in the high frequency (GHz-THz) regime, as required by telecom applications.

It is finally worth mentioning that, while trial and error design and optimization of such MP structures is possible, the task is certainly cumbersome. In the case of our simple model multilayer structure, for instance, it is necessary to optimize at once all the structural parameters (all the layer



Magnetoplasmonics, Fig. 17 Ideal multilayer Au/Co/Au SPP sensors as a function of the Co layer thickness obtained by genetic algorithm optimization technique. On the x-axis, t_{Co} represent the Au/Co/Au Cobalt thickness. On the y-axis, $t_{Au,bottom}$ and $t_{Au,top}$ represent the Au/Co/Au Gold thickness, thus defining a sensor of sensitivity $\partial(\Delta R)/\partial n$. Thicker Co layers must be positioned farther from the metal dielectric interface in order to minimize the optical losses, while thinner Co layers may be positioned closer to the surface to maximize the MO activity. In order to obtain an ideal sensor, low optical losses take precedence on high MO activity, since thinner Co layers provide final better sensitivities (from [9])

thicknesses, i.e., four parameters) plus the light wavelength and angle of incidence, thus obtaining a six-dimensional parameter space to explore. It is then clear that a different approach is desirable to tackle the design of this kind of structures. Recently, powerful genetic optimization tools have been successfully introduced for the design of such devices [9]. This kind of approach allows an easy exploration of the whole parameter space, and finally the determination of illumination and multilayer structural conditions that are optimal for the task at hand, e.g., the maximization of the MO activity, a minimization of the optical losses, or an ideal balance between SPP and MO features. Figure 17 shows an example of this process for a MO SPP Au/Co/Au multilayer designed to be an intensity interrogated Magneto Optical Surface Plasmon Resonance (MOSPR) sensor. In this case, in order to maximize the sensor sensitivity $\partial(\Delta R)/\partial n$, where n is the monitored refractive index, the genetic algorithm outputs multiple ideal sensors, providing a different tradeoff between the MO activity (thick Co layers) and low optical losses (thin Co layers), burying the ferromagnetic layer at different depths to modulate the balance between the two [9].

As a final remark it is extremely important to underline that in a real life scenario the quality of the adopted materials and the adopted synthesis processes play a major role in the final device MP properties, with epitaxially grown materials providing vastly improved performances if compared with more traditional approaches such as evaporation and sputtering [2].

Hybrid Nanohole Arrays

Thin metallic film with 2D ordered arrays of holes with sub-wavelength size (nanohole arrays, NHA) exhibit the so-called extraordinary optical transmission (EOT), whose explanation has been inherently linked to the excitation of SPPs propagating at both interfaces. Since the periodicity of the holes provides the extra momentum needed for direct photon-SPP coupling, NHAs received growing attention also as effective platforms to investigate the coupling between plasmonic and MO effects.

Extraordinary optical transmission and enhanced Faraday rotation have been found by Belotelov et al. [10] in bilayer metal/dielectric heterostructures, where the Au metallic plate

(68 nm thick) was perforated with a periodic hole arrays (pitch $a = 750$ nm, hole size $d = 395$ nm) and the Bi-substituted yttrium, iron garnet magnetic dielectric plate (118 nm thick) was magnetized in polar geometry perpendicularly to its plane. The measured simultaneous enhancement of transmittance and MO Faraday and Kerr effect in the NIR range is a clear demonstration of the SPP-MO interplay: the incident light excites the SPP propagating as a Bloch wave at the upper surface of the metallic film. The SPP tunnels through the NHA film to excite the SPP Bloch wave at the metal/dielectric interface. Finally, the MO activity of the dielectric layer produces two effects: (i) a TM-TE conversion (i.e., a Faraday rotation), whose efficiency depends on the dielectric layer thickness and (ii) a localization of the TE mode as a guided wave into the dielectric layer. The far-field optical signal is generated by photons radiated by the scattering of the SPP at the holes.

Nonreciprocal effects in EOT from NHA embedded in MO-active medium have been theoretically predicted by Khanikaev and coworkers [11]. The nonreciprocity was shown to give rise to one-way EOT, so that the structure behaves as a one-way mirror or an optical isolator. This effect can be explained as a result of the modification of the phase-matching condition for the evanescent coupling between leaky SSP modes existing at the opposite interfaces of the NHA film, which exhibit nonreciprocal dispersion due to the coupling with the MO film.

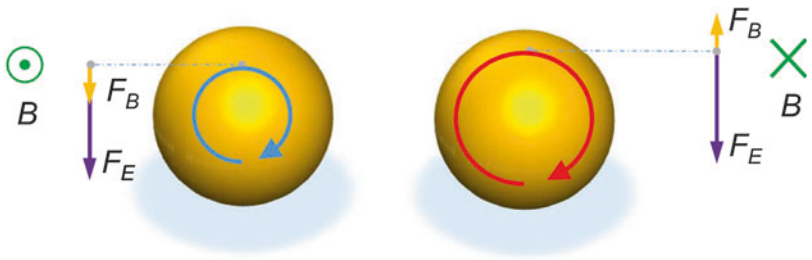
Magnetoplasmonics in Systems Supporting Localized Plasmon Resonances

Noble Metal Nanostructures

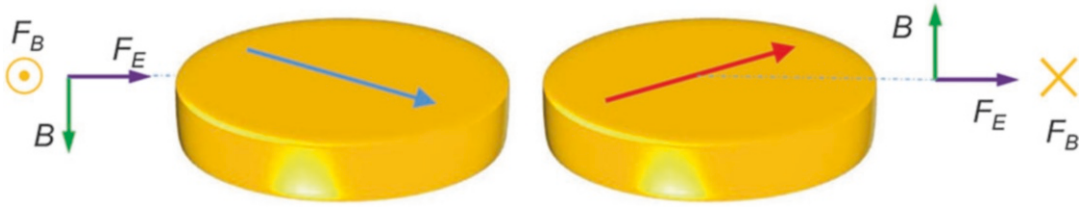
Examples of magnetic field-modulated LSPR have been recently observed and rationalized in the MO response of pure noble metal nanostructures, such as Au nanoparticles and nanodisks. Although metals such as Au, Ag, or Cu are diamagnetic, experimental results clearly showed that their magneto-optical effect is directly correlated with the LSPR excitation. Similar results were also observed in other geometries, e.g., Au dots and rods.

The influence of the magnetic field on the LPSR has been demonstrated by Pineider et al. [12] in colloidal dispersions of Au nanospheres, characterized with MO spectroscopy. The interpretation of their MCD response (i.e., the difference between two absorption spectra acquired using light with opposite helicity in the presence of an external magnetic field parallel to the incident light direction) introduces the essential role of the LSPR: conduction electrons interacting with the electric field of a light wave move as an oscillating dipole so that, if a magnetic field is also present, the oscillating charges undergo an additional Lorentz force. In the case of MCD, circularly polarized light provides to the LSPR electrons a peculiar collective rotating motion, yielding to circular plasmon resonance modes. The left- and right-polarized modes are degenerate in zero applied magnetic field, but the activation of an external magnetic field induces an energy shift, removing the degeneration. Observation of such a small energy shift by direct absorption measurements can be very challenging, whereas MCD spectroscopy easily allows its detection and measurement due to the derivative-like shape of the MCD spectrum crossing zero at the optical resonance (see Fig. 18).

Similar effects have been observed by Sepúlveda et al. [13] in arrays of Au nanodisks deposited on a glass substrate, separated by sub-wavelength distances, fabricated by large scale colloidal lithography. In this case, the system has been studied in reflection geometry by means of polar MOKE spectroscopy. It has been shown that in such structures it is possible to excite a LSPR and a sizeable MO signal related to LSPR excitation has been observed. In this case, the phenomenon has been equivalently described in terms of the linear polarization of incident light. The Lorentz force, generated by the external magnetic field, causes the LSPR direction to shift by a complex angle with respect to the incident light polarization, thus giving rise to MO effects of rotation and ellipticity (see Fig. 19).



Magnetoplasmonics, Fig. 18 Schematic representation of the Lorentz force affecting the *left*- and *right*-circular plasmon resonance in Au nanospheres



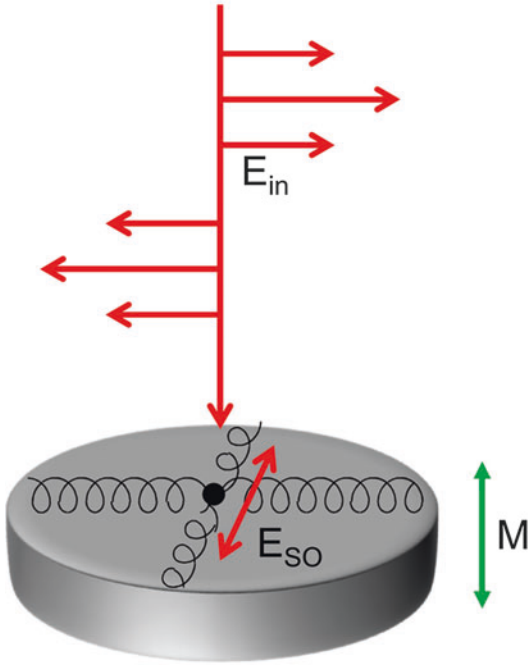
Magnetoplasmonics, Fig. 19 Schematic of the effect of the Lorentz force on linear plasmon resonance modes in Au nanodisks

Ferromagnetic Metal Nanostructures

A complementary approach to magneto-plasmonics is represented by the investigation of the phenomenon in pure ferromagnetic nanostructures. Any nearly free-electron metal can support surface plasmon resonances in the optical regime and several experimental and theoretical works have evidenced the modification in the MO response near the surface plasmon frequency. Guided by the common intent of magneto-optics in increasing the interaction between the light and the magnetic medium, also the magnetoplasmonic community focused initially on the enhancement of MO effects.

An example is given by the studies conducted by Gonzalez-Diaz et al. [2] on the MO enhancement in Ni nanowires embedded in a dielectric matrix. The amplitude of the MO signal is known to be basically proportional to the amount of the magnetic material. However, the experimental results of that study show a comparable Kerr rotation in both continuous Ni film and Ni nanowires in correspondence of the spectral region where LSPRs take place. Two important aspects have been introduced and discussed: the necessity of a theoretical MO calculation support, in this case represented by a scattering matrix method adapted to MO, and subsequently, the longstanding question about the role played by the reduction of the purely optical component in the MO enhancement. As already introduced in the SPP section, the MO response is indeed a ratio between the off-diagonal and the diagonal elements of the dielectric tensor of the material and for this reason, any reduction in the pure optical contribution results in an increase of the MO activity.

Apart from the MO enhancement, an interesting and basically new effect arising from the combination of LSPR and MO is represented by the possibility to manipulate the light phase which turns into a modified polarization of the light reflected by the nanostructure. Maccaferri et al. [14] have experimentally finalized such a tuning and have introduced and developed a complete computational model able to reproduce the experimental Kerr rotation and ellipticity of noninteracting magnetic nanostructures deposited on a nonabsorbing substrate. A key point of those studies is represented by the theoretical model employed to describe the coupling between LSPR and MO activity: Contrary to the intuitive argument that a MO enhancement is associated with the LSPR that is directly excited by the electric field of the incident light, only the LSPR excitation transverse to the electric field of the



Magnetoplasmonics, Fig. 20 A schematic of the two orthogonal damped harmonic oscillators coupled by the spin–orbit (SO) interaction in a ferromagnetic disk (Figure based on Ref. [14])

incident light, due to the spin–orbit coupling, determines the MO response observed in the ferromagnetic nanostructures (Fig. 20).

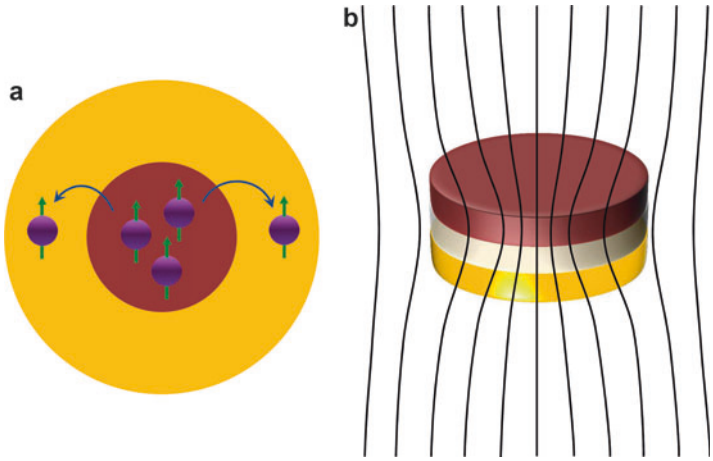
Finally, treating magnetoplasmonics in pure ferromagnetic nanostructures cannot overlook a mention of the effects of the LSPR-enhanced electric field on the nonlinear MO response. A recent example of the interplay between optical and magneto-optical second harmonic generation (MO-SHG) is the work by Valev et al. [15] where the role of LSPR is evidenced in the magnetic response of G-shaped Ni nanostructures. Upon rotating the G-shaped structures, an asymmetric MO-SHG signal is reported inverting the magnetic field direction. More generally, the nonlinear MO effects measured under SHG excitation are orders of magnitude larger than the linear MO effects. This extreme sensitivity of the nonlinear MOKE effect to weak plasmonic modes can actually be the breakthrough for effective applications in MO.

Hybrid Magnetic-plasmonic Nanostructures

Hybrid magnetic-plasmonic systems are subject of intense research due to the fact that they bear a double functionality. This is particularly exciting in view of applications in several fields, such as theranostics or heat-mediated magnetic recording. In a perspective more specific to magnetoplasmonics, they are in principle ideal platforms to observe both the effect of the magnetic component on the field-dependent behavior of LSPR, and the effect of plasmon confinement of the EM field on the MO response of the magnetic material. So far, research on hybrid magnetic-plasmonic nanosystems has been primarily focused on the latter effect, and no unambiguous report has been published to date on the direct effect of the magnetic moiety on the plasmonic one.

LSPR-enhanced MO activity has been reported for a number of systems, prepared both by chemical synthesis and through physical methods. Jain et al. [2] were among the first to describe a MO enhancement in gold@maghemite core-shell colloidal nanoparticles, which resulted in the observation of an electronic transition from maghemite located at the same energy of LSPR, and that is too weak to be observed in plain maghemite. Other similar enhancement effects have been reported for colloidal hybrid magnetic-plasmonic nanoparticles such as Co@Ag core-shells and Ag-$CoFe_3O_4$ heterodimers [2].

The finer control on morphology achievable by physical fabrication methods allows a wider range of geometries for magnetic-plasmonic hybrid systems. In addition, the more regular structures than those prepared by chemical routes are easier to model with analytical or numerical methods. Sizeable effects have been found in Au nanostructures embedded in dielectric garnet magnetic layers by Tomita et al., while Du et al. studied the size and shape effects on the MO enhancement in Au/CoPt/Au nanodisks [2]. The most accurate correlation between theory and experiment is found in the work by Armelles group [2]: by using a trilayer Au/Co/Au nanodisk geometry, they correlated various positions of the Co layer with respect to the top and bottom Au layers to the



Magnetoplasmonics, Fig. 21 (**a**) Sketch of the hybridization (gray region) of the plasmonic (shell) and magnetic (core) moieties. (**b**) Sketch of the effect on local magnetic field arising from the vicinity of a magnetic material (core) to a plasmonic structure (shell)

intensity of the MO signal, and through accurate numerical modeling they found it to be related to the local electromagnetic field intensity concentrated by the Au structures. They also studied the effect of a dielectric (silica) spacer on similar systems and found that the lower optical losses afforded by the spacer further increase the MO enhancement by LSPR.

The vicinity of a magnetic component is expected to influence the MO response of the plasmonic moiety in several ways. Here two of the main effects are considered: (i) electronic hybridization between the two materials, due to which part of the uncoupled spins of the magnetic material are transferred to the plasmonic part (Fig. 21a) and (ii) proximity effects, in which the local static magnetic field is increased with respect to the applied field due to the permeability of the magnetic material (Fig. 21b). The former phenomenon requires at least a partial metallic nature of the magnetic moiety, for electronic delocalization to take place, while the latter can in principle be observed independently of the nature of the magnetic material.

Possibly the only report of a magnetoplasmonic system in which delocalization gives rise to hybrid magnetoplasmonic modes can be found in Armelles group's work on Au/Co/Au disks prepared by colloidal lithography [2], in which a single hybrid plasmonic mode arising from both the Au and Co layers is influenced by the magnetic field, resulting in strong magneto-optical activity at the plasmon resonance peak. The authors however focused more on plasmon-mediated MO-enhancement aspects, so a detailed analysis of the phenomenon is not possible for the time being. In general, the underlying idea to observe magnetic-plasmonic hybridization is to obtain nanostructures in which the magnetic material is a conductor (metallic or semimetallic) and does not completely damp the LSPR of the noble metal structure; promising candidate structures include core-shell and heterodimeric nanoparticles prepared via colloidal chemistry methods [16] or layered magnetic-plasmonic nanodisks prepared by physical fabrication methods [2], in which the plasmonic metal can be either Au or Ag, and the magnetic metal one among the ferromagnetic ones (Fe, Co, Ni). Another possible configuration is the combination of the two materials into alloys, so that the nanostructure and its electronic properties are homogenous.

The enhancement of the effect of a magnetic field on LSPR relies on the fact that a strongly magnetic material will significantly affect the magnetic field distribution in its vicinity: if a noble metal nanostructure is found in the area of field hot-spot, an increased MO response is expected. The only stringent requirement for this phenomenon is that the MO response of the magnetic material does not significantly overlap with LSPR: in such case the magnetoplasmonic response would be drowned by the MO signal of

the magnetic material. Since no direct interaction between the two moieties is required, dielectric spacers, such as silica, can help reducing the damping introduced by strongly absorbing magnetic materials. A wide range of hybrid materials are potentially suitable for the observation of such phenomenon, chiefly gold-silica-iron oxide core-shell systems prepared following colloidal chemistry routes [16] and layered noble metal-silica-magnetic metal nanodisks prepared via physical methods [2]. Again, no direct observation has been reported so far on this effect, mainly due to fact that the opposite type of magnetoplasmonic enhancement has been sought after by the scientific community.

Applications of Magnetoplasmonics

The field of application of magnetoplasmonic devices spans a range similar to that of purely plasmonic nanosystems, aiming at improving the functionalities of traditional optical devices with novel active features triggered by the magnetic field. In particular, among the most relevant and up-to-date applications of magnetoplasmonic devices, we will address theranostics, sensing (mostly of biomolecules) and optical telecommunications.

Plasmonic and magnetic materials show, individually, extremely promising optical properties that are in some cases already exploited in applications such as diagnostics, food quality control, and environmental monitoring. Prompted by the need of further improving the functionalities of these devices, material science researchers are nowadays exploring the feasibility of combining the two classes of materials so as to realize new engineered magnetoplasmonic systems with innovative and intriguing enhanced properties.

On one hand plasmonic nanomaterials such as Au, Ag, and Pt NPs have a widespread range of applications. Gold NPs are often used in scattering-based optical imaging as contrast agents in the visible range. Moreover, as they can be engineered into various shapes such as nanoshells, nanorods, or nanocages, they can exhibit LSPR also in the NIR range. Indeed, the region around 900–1000 nm is particularly favorable for in vivo imaging of biological tissues, since shorter wavelengths (>750 nm) are extinguished by endogenous pigments or b-hemoglobin, and longer wavelengths (<1300 nm) are strongly attenuated by water. The only drawback is that the tissue penetration depth of optical imaging is limited to few millimeters. Moreover, metal NPs, thanks to their strong field localization, can be also employed to produce localized photothermal effects for selective destruction of malignant tissues, the breaking of chemical bonds, or disintegration of carrier matrices, joining optical imaging with potential therapeutic action. This combined approach, which is often called *theranostic*, is probably one of the most fascinating promises offered by nanomedicine. While Ag NPs can be employed as antibacterial agents, Au NPs can be further used for biosensing by monitoring changes of the LSPR frequency, or as proper substrates for surface enhanced Raman spectroscopy (SERS) since SH thiol groups readily bind to Au surfaces, making them a good choice for bioconjugation.

On the other hand magnetic NPs (MNPs) are already clinically employed for improving the contrasting magnetic resonance imaging (MRI), offering high resolution and excellent tissue penetration depth. However, despite the good spatial resolution, MRI is not as sensitive as optical imaging and of difficult adoption for the visualization of microscopic tissue samples. Therefore, it is highly desirable to have novel nanomaterials enabling new imaging modes for enhanced contrast specificity and complementary use. Similarly to plasmonic nanomaterials magnetic NPs may also provide local heat dissipation which can be exploited for the hyperthermic treatment of tumors. Moreover, a magnetic field gradient can be used to externally drive magnetic NPs to target tissue or to enhance uptake by cells. Finally, magnetic NPs also play a key role in many lab-on-chip systems in transport, separation and recognition of biological compounds. In these cases, the indirect recognition of a biological compound often occurs via magnetoresistive sensors.

Multifunctional Magnetoplasmonic Theranostics

By properly combining magnetic and plasmonic materials at the nanoscale, a number of desirable properties, including magnetic attraction, NIR absorption, multimodal imaging, photon scattering and easy biomolecular conjugation through stable thiolate–metal interaction can be achieved (Fig. 22). Moreover, magnetic hyperthermia and photodynamic therapy with plasmon states can be used for cancer treatment (Fig. 23).

For instance, the use of magnetoplasmonic nanostructures facilitates the simultaneous manipulation and imaging of single NPs, by taking advantage of both optical and magnetic properties. The ability of simultaneously manipulating and imaging single NPs inside living cells demonstrated their potentiality as novel bio-nanoprobes for applications in single molecule measurements, drug delivery, and in vitro or in vivo diagnostics and therapy.

Different combinations of materials at the nanoscale can be found in the literature for a wide range of applications, e.g., therapy of various types of cancer, Alzheimer, virus-induced infections, etc.. . .

Magnetoplasmonic hybrid nanosystems have been used for improving optical imaging in living cells. Zhang et al. proposed a new magnetoplasmonic nanomaterial based on the integration of one dimensional FePt and Au sections [17]. The unique, strongly polarization dependent, and highly anisotropic plasmonic response of these nanomaterials enabled the visualization of their position and orientation down to the individual segments well beyond the diffraction limit, thus demonstrating the possibility to simultaneously manipulate and image single NPs inside live cells.

Biophotonic imaging based on a dynamic mode of optical contrast has been discussed also by Wei et al., who used Au nanostars with super paramagnetic cores as novel optical contrast agents [18]. The nanostars exhibit polarization-sensitive NIR scattering and produce a frequency-modulated signal in response to a rotating magnetic field gradient. This periodic modulated signal can be converted into Fourier-domain images with a dramatic reduction in background of the recorded images.

Chen et al. prepared magnetoplasmonic core-shell nanoparticles based on a magnetic Fe_3O_4 core encapsulated in a silica shell, in turn surrounded in a crown-like fashion by Au nanoparticles. The particles are also coated by poly(ethyleneglycol) chains to extend the blood circulation time [19]. The work demonstrated the remarkable imaging and therapeutic capabilities of the NP assemblies against leukemia HL-60 cells.

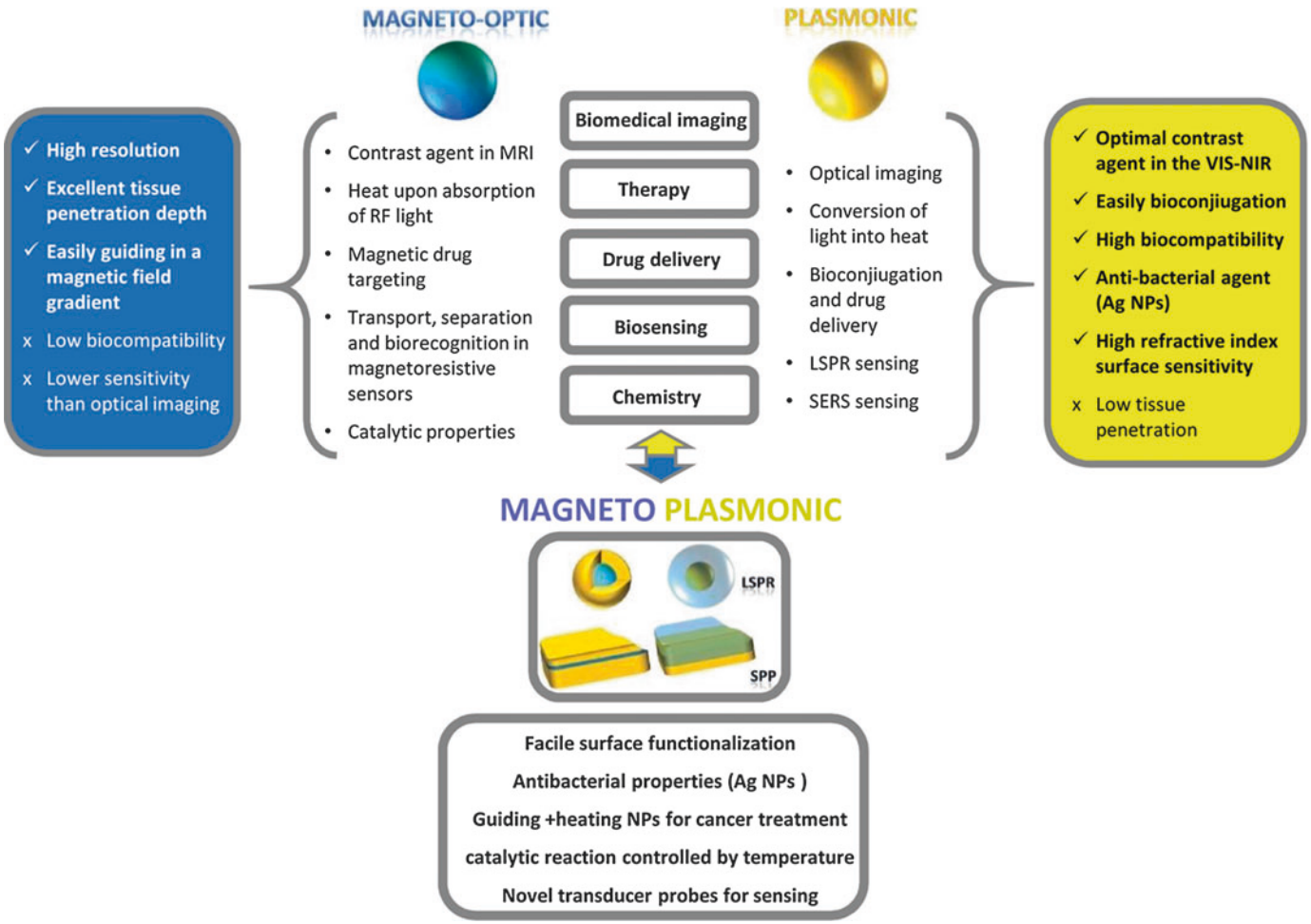
Finally, a new imaging mode, called magnetomotive photoacoustic imaging, has been proposed by Jin et al. based on iron oxide and gold-coupled core-shell NPs with well-defined structural characteristics and physical properties, simultaneously compact in size and responsive in the NIR spectrum. Beyond conventional NP-based imaging modalities such as TEM, optical imaging, MRI and photo acoustic imaging (PA), coupling of magnetic motion with photothermal conversion enables the new modality with remarkable contrast enhancement compared with conventional PA imaging using conventional NP contrast agents [20].

Magnetoplasmonic Biosensing

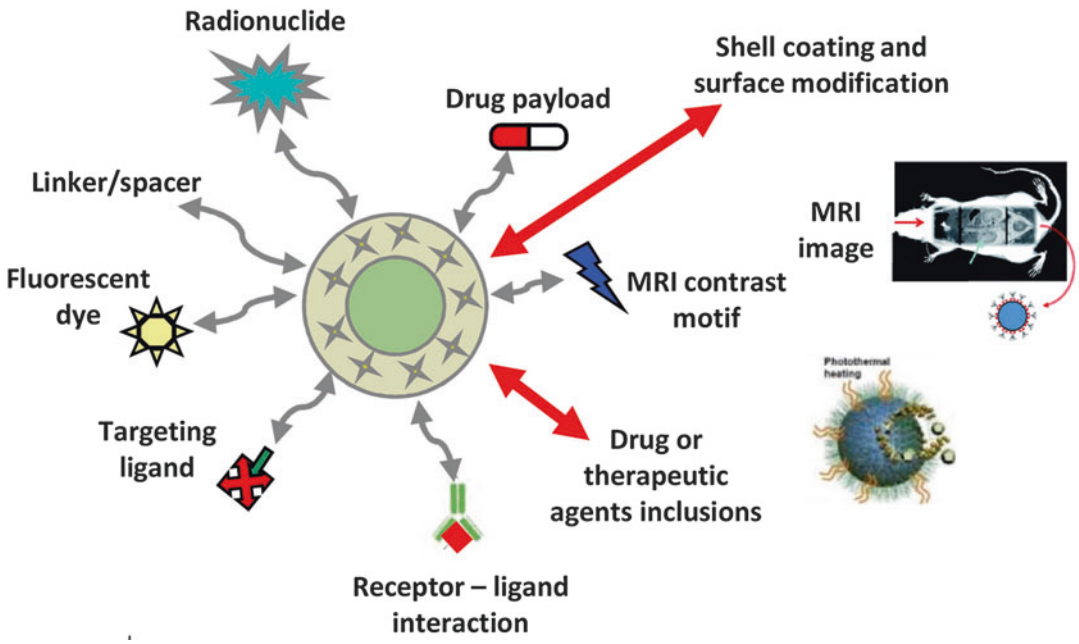
Beyond the use in theranostics, magnetoplasmonic nanomaterials attracted a huge interest as novel transducers in gas or biosensing devices.

Magnetic modulation of plasmonic properties is a novel modulation scheme for SPR–based sensor, alternative to the established mechanical and phase modulations, allowing for a better signal-to-noise ratio and, therefore, an enhanced limit of detection in any kind of sensor. This is an intriguing issue since the detection performance of the current SPR bio-sensors is often insufficient to monitor low concentrations of small biomolecular analytes (e.g., drugs, vitamins or antigens).

Sepúlveda et al. [21] first proposed the use of a proper combination of magnetic materials and plasmonic layers as novel transducers for sensing. Such a combination produced a sharp enhancement of the MO effects strongly dependent on the



Magnetoplasmonics, Fig. 22 Schematic representation of functional properties of magnetic, plasmonic and magnetoplasmonic nanomaterials and relative potential applications in biological analysis and nanomedicine



Magnetoplasmonics, Fig. 23 Scheme of multifunctional magnetoplasmonic nanoparticles for molecular imaging, drug delivery and therapy. Functionalized nanoparticles can be tailored for individualized diagnosis and treatments

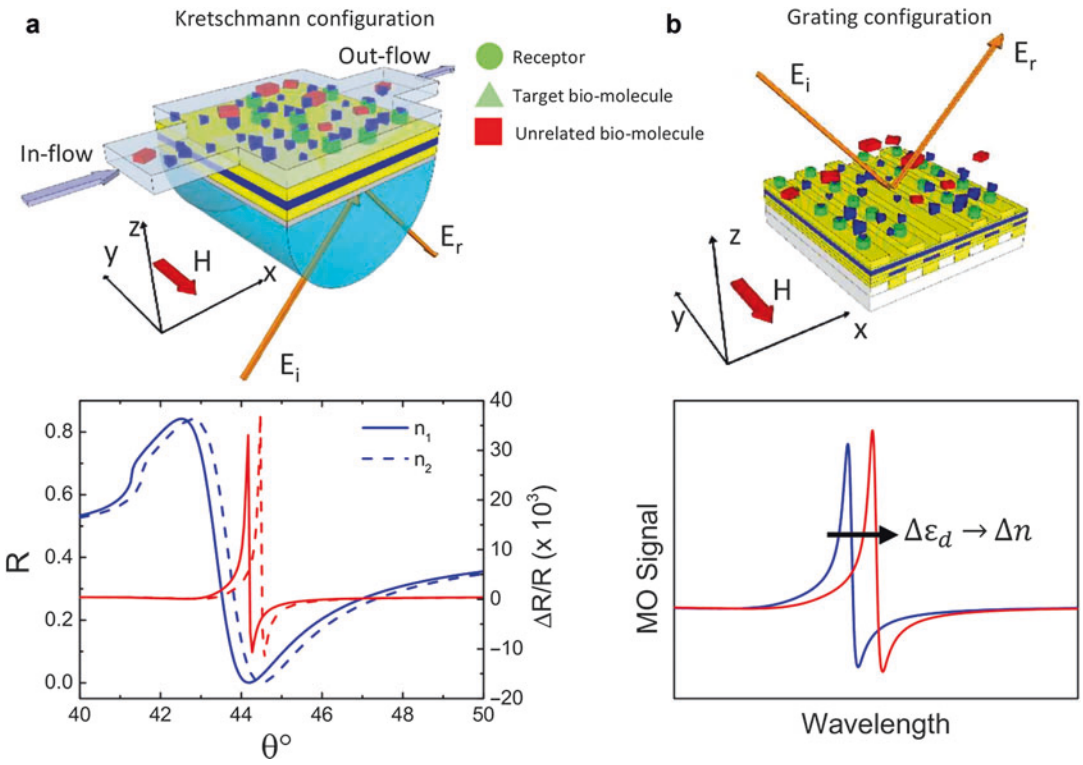
optical properties of the surrounding medium, allowing its use for biosensing applications. They demonstrated the possibility to record the adsorption of biomolecules onto the transducing multilayers structure with an increased limit of detection if compared with standard sensors (Fig. 24a).

Manera et al. [22] demonstrated the possibility to use the enhanced MO signal of metal/ferromagnetic materials trilayers for investigating biomolecular interactions in liquid phase. To this purpose, the transducer surface was functionalized with thiols for chemical anchoring of Bovine Serum Albumin molecules and the binding between immobilized molecules and their complementary target molecules was recorded. Finally higher sensing performance in term of sensitivity and lower limit of detection of the MOSPR biosensor with respect to traditional SPR sensors was demonstrated.

The same research group also demonstrated that a similar approach can be used to improve gas sensing performances of traditional SPR sensors. In this case using a thin nanostructured TiO_2 sensing layer it was shown that the new magnetoplasmonic transductor used in resonance condition had a better sensitivity than a plasmonic transductor used in a classical SPR sensor technique [23].

Martín-Becerra et al. [24] discussed theoretically the possibility to use the planar multilayer configuration as sensing transducer in a magnetically modulated plasmonic interferometer. The modulation of the SPP wavevector induced by an external applied magnetic field in Au/Co/Au trilayers represents a new physical parameter offering a higher sensitivity to refractive index changes.

A different approach was used by Chou et al. [25] to exploit strong transverse MO Kerr effect on a ferromagnetic surface Plasmon grating for developing a highly sensitive sensor. The surface plasmon grating, made of a hybridized Au/Fe/Au layer, exhibits a strongly dispersive Kerr parameter variation near the surface plasmon polariton wavelength via coherent scattering of the SPP on the grating structure. Kerr's parameter monitoring can be used to detect chemical or biological objects in a fluid medium (Fig. 24b).

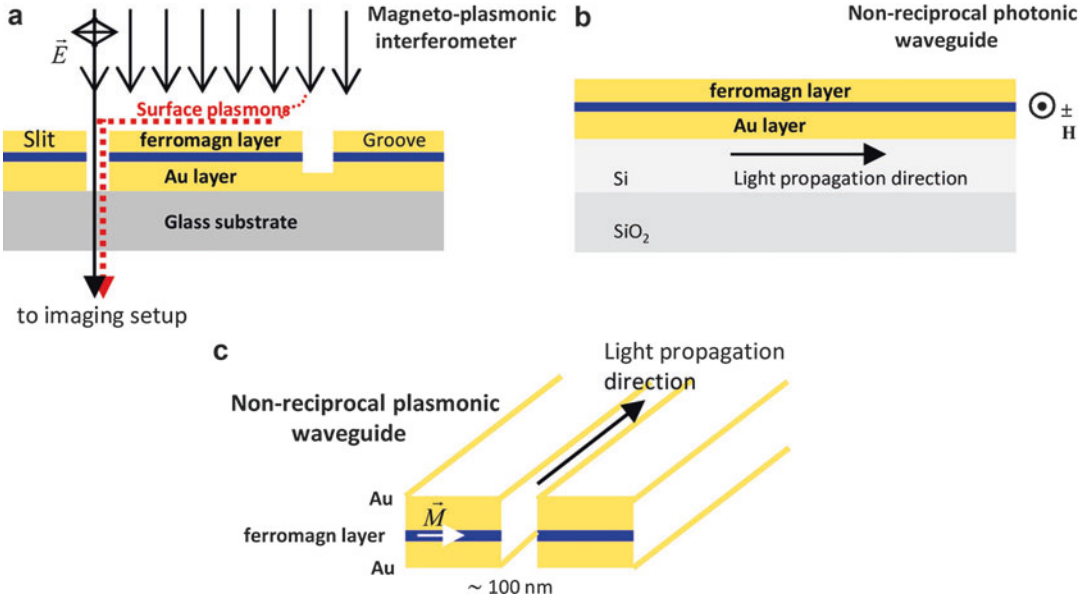


Magnetoplasmonics, Fig. 24 Examples of magnetoplasmonic nanomaterials for biosensing applications. (**a**) Magneto-optical Surface Plasmon Resonance transduction (*MOSPR*) technique achieved in a Kretschmann ATR configuration. (**b**) Magneto–optic Kerr effect on surface plasmon grating biosensors. In both cases, a proper combination of metallic and magnetic nanomaterials is chosen as transducer platform. Outer metal surface allows for easy functionalization with biorecognition molecules able to detect the presence of target molecules specifically. The detection is recognized as a change in the refractive index at the metal/dielectric interface, which can be revealed sensitively by the magnetoplasmonic signal

Optoelectronics and Telecommunications

Apart from biological and nanomedicine applications, active plasmonic systems have been designed also to control plasmon propagation in on-chip miniaturized all-optical circuits. Modulation of plasmonic properties by an external magnetic field as well as enhanced magneto-optical (MO) Faraday and Kerr effects arising in gyrotropic media attracted much attention due to their possible applications in light control at a submicron scale, in the field of telecommunications, in data storage and lastly in the new field of spinplasmonics (Fig. 25).

To this respect, magnetoplasmonics can be one of the enabling technologies for the development of innovative photonics and telecommunication devices. An excellent example is given by garnets, which are currently being employed into commercial optical insulators for telecommunications. Promising MO improvements, induced by the introduction of SPR components, were indeed observed in hybrid films composed of Au NPs dispersed in garnet films and in Au-Fe oxide hybrid NPs. On the other hand, since the garnets and the Faraday effect technologies cannot be easily integrated into photonic microtechnologies, the nonreciprocity of SPPs offers an interesting way to design alternative integrated optical isolators. Garnet films deposited on Au or Ag films or metallic magnetic films deposited on semiconductors optical waveguides are now being proposed [2], integrated in most of cases into an interferometer in Mach-Zehnder configuration. Similarly, Khurgin et al. considered propagation of a



Magnetoplasmonics, Fig. 25 Examples of application of magnetoplasmonic nanomaterials for optoelectronic devices. (**a**) Working principle of a magnetoplasmonic microinterferometer which can find application in telecommunications Ref. [6]; (**b**) working principle of a nonreciprocal photonic waveguide; (**c**) working principle of a nonreciprocal plasmonic waveguide Ref. [21]

plasmon polariton on the interface between a metal and a MO material, demonstrating that, in the presence of transverse magnetic field, a strong nonreciprocal phase shift occurs [26]. This phenomenon can be used to achieve optical isolation action in short nanoplasmonic guides with low insertion loss.

Chau et al. [27] proposed a novel high-speed MO modulator based on the optical excitation of surface magneto-plasmon waves in bismuth-substituted yttrium iron garnet (Bi-YIG) under an applied magnetic field perpendicular to the plane of incidence. The device was able to provide about 50 % efficiency modulation at GHz speed in the IR range (at 1550 nm), demonstrating a much faster switching ability over other driving agents, such as thermo-optical or electro-optical effects.

An innovative application of magneto-plasmonic materials as compact isolators for circularly polarized waves has been found by Davoyan et al. [28]. They suggested a principle for isolation of circularly polarized waves in magnetically active extreme parameter meta-materials. Using theoretical analysis and numerical simulations, they showed that metamaterials with extreme parameters, such as epsilon-near-zero materials, when merged with MO materials, become transparent for forward circularly polarized waves of a given handedness and opaque for backward propagating waves of the same handedness. The concept may find useful applications also in magnetically controlled light modulation and polarization modulation and nonreciprocal optical antennas.

Conclusions

This entry focused on understanding magnetoplasmonic effects arising from the interaction of plasmonic and MO phenomena typical of metallic and magnetic materials. Theoretical and experimental considerations show that these phenomena are strongly correlated, whereas plasmonic effects are able to enhance MO properties, and likewise magnetic materials are able to modulate plasmonic properties at high frequency. The mutual interaction that leads to

magnetoplasmonic phenomena is observed in a large variety of structures, ranging from extended or nanostructured thin-films to a broad class of nanoparticles characterized by marked structural and compositional differences. The ability to simultaneously exploit enhanced plasmonic and MO properties, and the wide diversity of emerging applications spanning disciplines such as nanomedicine, telecommunications and sensing, illustrate the importance of the fascinating and fast-growing research field of magnetoplasmonics.

Cross-References

- ▶ Active Plasmonic Devices
- ▶ Biosensors
- ▶ Local Surface Plasmon Resonance (LSPR)
- ▶ LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design
- ▶ Nanobiosensors
- ▶ Nanomagnetism
- ▶ Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods
- ▶ Nanophotonic Structures for Biosensing
- ▶ Nanostructures for Photonics
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Plasmonics
- ▶ Refractometric Sensing Using Plasmonic Nanoparticles
- ▶ Surface Plasmon-Polariton-Based Detectors

References

1. Maier, S.A.: Plasmonics: Fundamentals and Applications. Springer, New York (2007)
2. Armelles, G., Cebollada, A., García-Martín, A., González, M.U.: Magnetoplasmonics: combining magnetic and plasmonic functionalities. Adv. Opt. Mater. **1**, 2–2 (2013)
3. Zvezdin, A.K., Kotov, V.A.: Modern Magnetooptics and Magnetooptical Materials. Institute of Physics Pub., Bristol/Philadelphia (1997)
4. Palik, E.D.: Handbook of Optical Constants of Solids. Academic Press, New York (1985)
5. Bogani, L., Cavigli, L., de Julián Fernández, C. et al.: Photocoercivity of nano-stabilized Au:Fe superparamagnetic nanoparticles. Adv. Mater. **22**, 4054–4058 (2010)
6. Temnov, V.V., Armelles, G., Woggon, U., et al.: Active magneto-plasmonics in hybrid metal–ferromagnet structures. Nat. Photonics **4**, 107–111 (2010)
7. Martín-Becerra, D., Temnov, V.V., Thomay, T., et al.: Spectral dependence of the magnetic modulation of surface plasmon polaritons in noble/ferromagnetic/noble metal films. Phys. Rev. B **86**, 035118 (2012)
8. Belotelov, V.I., Kreilkamp, L.E., Akimov, I.A., et al.: Plasmon-mediated magneto-optical transparency. Nat. Commun. **4**, 1–7 (2013)
9. Pellegrini, G., Mattei, G.: High-performance magneto-optic surface plasmon resonance sensor design: an optimization approach. Plasmonics **9**, 1457 (2014)
10. Belotelov, V.I., Doskolovich, L.L., Zvezdin, A.K.: Extraordinary magneto-optical effects and transmission through metal-dielectric plasmonic systems. Phys. Rev. Lett. **98**, 077401 (2007)
11. Khanikaev, A.B.., Mousavi, S.H., Shvets, G., Kivshar, Y.S.: One-way extraordinary optical transmission and nonreciprocal spoof plasmons. Phys. Rev. Lett. **105**, 126804 (2010)
12. Pineider, F., Campo, G., Bonanni, V., et al.: Circular magnetoplasmonic modes in gold nanoparticles. Nano Lett. **13**, 4785–4789 (2013)
13. Sepúlveda, B., González-Díaz, J.B., García-Martín, A., et al.: Plasmon-induced magneto-optical activity in nanosized gold disks. Phys. Rev. Lett. **104**, 147401 (2010)
14. Maccaferri, N., Berger, A., Bonetti, S., et al.: Tuning the magneto-optical response of nanosize ferromagnetic Ni disks using the phase of localized plasmons. Phys. Rev. Lett. **111**, 167401 (2013)
15. Valev, V.K., Silhanek, A.V., Gillijns, W., et al.: Plasmons reveal the direction of magnetization in nickel nanostructures. ACS Nano **5**, 91–96 (2011)
16. Cortie, M.B., McDonagh, A.M.: Synthesis and optical properties of hybrid and alloy plasmonic nanoparticles. Chem. Rev. **111**, 3713–3735 (2011)
17. Zhang, Y., Wang, Q.: Magnetic-plasmonic dual modulated FePt-Au ternary heterostructured nanorods as a promising nano-bioprobe. Adv. Mater. **24**, 2485–2490 (2012)
18. Wei, Q., Song, H.-M., Leonov, A.P., et al.: Gyromagnetic imaging: dynamic optical contrast using gold nanostars with magnetic cores. J. Am. Chem. Soc. **131**, 9728–9734 (2009)
19. Chen, W., Xu, N., Xu, L., et al.: Multifunctional magnetoplasmonic nanoparticle assemblies for cancer therapy and diagnostics (theranostics). Macromol. Rapid Commun. **31**, 228–236 (2010)
20. Jin, Y., Jia, C., Huang, S.-W., et al.: Multifunctional nanoparticles as coupled contrast agents. Nat. Commun. **1**, 1–8 (2010)
21. Sepúlveda, B., Calle, A., Lechuga, L.M., Armelles, G.: Highly sensitive detection of biomolecules with the magneto-optic surface-plasmon-resonance sensor. Opt. Lett. **31**, 1085–1087 (2006)

22. Manera, M.G., Ferreiro-Vila, E., Garcia-Martin, J.M., et al.: Enhanced antibody recognition with a magneto-optic surface plasmon resonance (MO-SPR) sensor. Biosens. Bioelectron. **58**, 114–120 (2014)
23. Manera, M.G., Montagna, G., Ferreiro-Vila, E., et al.: Enhanced gas sensing performance of TiO2 functionalized magneto-optical SPR sensors. J. Mater. Chem. **21**, 16049–16056 (2011)
24. Martín-Becerra, D., Armelles, G., González, M.U., García-Martín, A.: Plasmonic and magnetoplasmonic interferometry for sensing. New J. Phys. **15**, 085021 (2013)
25. Chou, K.-H., Lin, E.-P., Chen, T.-C., et al.: Application of strong transverse magneto-optical Kerr effect on high sensitive surface plasmon grating sensors. Opt. Express **22**, 19794–19802 (2014)
26. Khurgin, J.B.: Optical isolating action in surface plasmon polaritons. Appl. Phys. Lett. **89**, 251115 (2006)
27. Chau, K.J., Irvine, S.E., Elezzabi, A.Y.: A gigahertz surface magneto-plasmon optical modulator. IEEE J. Quantum Elect. **40**, 571–579 (2004)
28. Davoyan, A.R., Mahmoud, A.M., Engheta, N.: Optical isolation with epsilon-near-zero metamaterials. Opt. Express **21**, 3279–3286 (2013)

Magnetron Sputtering

► Physical Vapor Deposition

Making Graphene

► Synthesis of Graphene

Manipulating

► Robot-Based Automation on the Nanoscale

Manufacturing Graphene or Graphene Manufacturing

► Synthesis of Graphene

Manufacturing Silica Nanowires

► Growth of Silica Nanowires

Marine Bioinspired Adhesives

► Biomimetics of Marine Adhesives

Marine Organisms

► Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment

Massive

► Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

Materiomics

► Mechanical Properties of Hierarchical Protein Materials

Mats

► Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

Maxwell–Bloch Simulations or Equations or Studies

► Nanopolaritonics

Maxwell–Schrödinger Simulations (or Equations)

▶ Nanopolaritonics

Maxwell–Wagner Effect

Mitsumasa Iwamoto
Department of Physical Electronics, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan

Synonyms

Interfacial charging effect

Definition

The Maxwell–Wagner effect accounts for charge accumulation at the two-material interface on the basis of the difference of charge carrier relaxation times in these two materials. Macroscopically, basic electrical properties of materials are specified using two physical parameters, dielectric constant ϵ and conductivity σ. The ratio of these two parameters, $\tau = \epsilon/\sigma$, gives relaxation time and stands for spreading time of excess free carriers in the materials. Charge accumulation is allowed at the interface between two materials with different relaxation times. When a current flows across the interface, charges are accumulated at the interface. This is nothing but the effect called the Maxwell–Wagner (MW) effect, and this effect appears in all kinds of two-material interface, such as metal–insulator interface, metal–semiconductor interface, insulator–semiconductor interface, and others. The current which flows through materials is given by $\vec{j} = \sigma \vec{E}$, where $\vec{E}$ represents the electric field and σ is conductivity of the materials. In steady state, $\nabla \cdot \vec{j} = 0$ is satisfied. This relation is also satisfied when a steady-state current flows across the two-material interface. As the electric field $\vec{E}$ is given by $\vec{E} = \vec{D}/\epsilon$, using electric flux density $\vec{D}$ and dielectric constant ϵ, the following relation concerning electric flux density $\vec{D}$ at the interface is obtained as [1]

$$\nabla \cdot \vec{D} = \nabla \cdot \epsilon \vec{E} = \nabla \cdot \left(\frac{\epsilon}{\sigma}\right) \vec{j} = \nabla\tau \cdot \vec{j} = q_s(\neq 0)$$

where $q_s(\neq 0)$ is accumulated charge density at the interface. This simple relation shows that charge $q_s(\neq 0)$ is accumulated at the interface between two materials with different relaxation times, i.e., at the meeting point satisfying the relation $\nabla(\epsilon/\sigma) = \nabla\tau \neq 0$. In addition, accumulated charge $q_s(\neq 0)$ is given as the inner product of the spatial gradient of relaxation time $\nabla\tau$ and steady-state current density $\vec{j}$. This is the Maxwell–Wagner effect and accounts for charge accumulation at all kinds of two-material interface.

Introduction

"Interface" plays an important role in electronic devices. For general capacitors with a structure of metal–insulator–metal, the metal–insulator interface is working to prevent charge injection from metals into insulator and accumulates charges on the metals, in proportion to applied voltage. For p–n junctions, the main function of the p–n interface is rectifying carrier transport across the interface and results in the formation of charge distribution that is asymmetric with respect to the interface. For field-effect transistors, mobile charges are accumulated at the gate insulator–semiconductor interface and they are conveyed along the interface. These typical examples show that "interface" is a meeting place of two different materials, where charges are accumulated. Accordingly, much attention should be paid to "interface" in electronic device fabrication. The Maxwell–Wagner (MW) effect accounts for charge accumulation at these two-material interfaces, such as metal–insulator interface, semiconductor–semiconductor interface, and insulator–semiconductor interface. The basic concept of the MW effect lies in the electromagnetic

field theory described by Maxwell's equation. In this entry entitled "Maxwell–Wagner Effect," the Maxwell–Wagner effect is explained, and then some examples of electronic devices are shown, in which the Maxwell–Wagner effect is actively working. In the following, special attention is paid to organic devices, because organic devices, including organic field-effect transistors, organic solar cells and organic electroluminescence (EL) devices, etc., are the most possible candidates of electronic devices in the twenty-first century, and the "interface" plays a dominant role in these device performances.

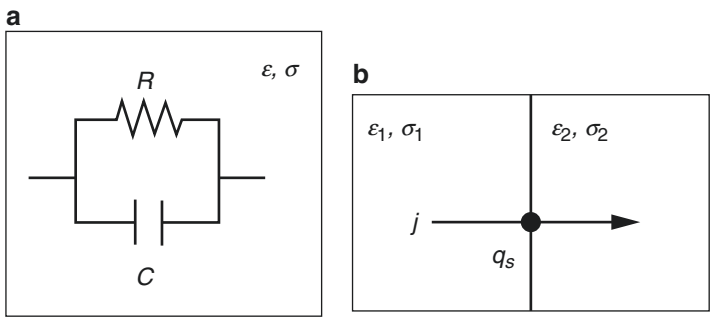
Maxwell–Wagner Effect

Macroscopic electrical properties of materials are characterized by using two physical parameters. These are dielectric constant ϵ and conductivity σ. The ratio $\tau = \epsilon/\sigma$ gives a relaxation time and stands for spreading time of excess free carriers in the materials. That is, a steady-state charge distribution is established after an elapsed time of around $\tau = \epsilon/\sigma$. In other words, this value $\tau = \epsilon/\sigma$ specifies the electrical properties of materials in terms of carrier transport. Charge accumulation is allowed at the interface between two materials with different relaxation times when a current flows across the interface. This is nothing but the effect called the Maxwell–Wagner (MW) effect, and this effect appears at all kinds of two-material interface, such as metal–insulator interface, metal–semiconductor interface, insulator–semiconductor interface, and others. Further, it is instructive to note that according to Maxwell's electromagnetic theory, a total current flowing across organic materials is the sum of the conduction current and Maxwell's displacement current. In other words, there are two paths for the current flow. This situation is described by using a parallel equivalent circuit model as illustrated in Fig. 1a, where the resistance R represents a path due to conduction current and capacitance C represents a path due to Maxwell's displacement current. R and C are given by $R = l/A\sigma$, $C = \epsilon A/l$, respectively (A: cross section, l: length of path). Consequently, the relaxation time $\tau = \epsilon/\sigma = RC$ represents response time of the equivalent circuit.

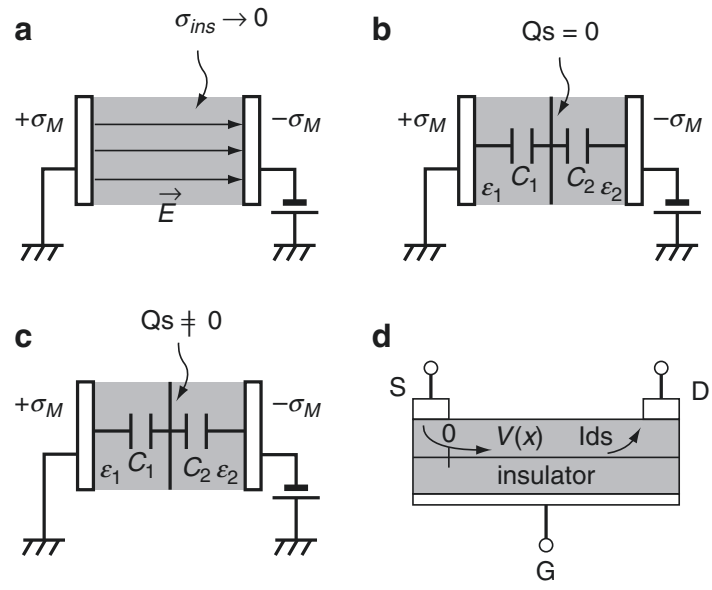
For simplicity, let us discuss a case of steady-state current flowing across the interface of the two-material system, as illustrated in Fig. 1b. The two materials are described using a set of two parameters, dielectric constant ϵ and conductivity σ, and constant steady-state current with a current density j is flowing across the interface. The current flow is given by $\vec{j} = \sigma \vec{E}$, where $\vec{E}$ represents the electric field and σ is conductivity of materials. In this situation, $\nabla \cdot \vec{j} = 0$ is satisfied over the entire region of the two-material system. The electric field $\vec{E}$ satisfies the relation $\vec{E} = \vec{D}/\epsilon$, where $\vec{D}$ is electric flux density. Hence the following relation concerning electric flux density $\vec{D}$ at the interface is obtained as [2–4]

$$\nabla \cdot \vec{D} = \nabla \cdot \epsilon \vec{E} = \nabla \cdot (\epsilon/\sigma)\ \vec{j} = \nabla\tau \cdot \vec{j} = q_s (\neq 0). \quad (1)$$

In more detail, from Eq. 1, the following relation is obtained as



Maxwell–Wagner Effect, Fig. 1 (**a**) Equivalent circuit of materials and (**b**) Maxwell–Wagner effect



Maxwell–Wagner Effect, Fig. 2 (**a**) MIM, (**b**) double-layer structure device ($Q_s = 0$), (**c**) double-layer device with Q_s, and (**d**) a three-electrode system

$$(\tau_2 - \tau_1)J = \sigma_s(\neq 0). \tag{2}$$

Here $\tau_1 = \epsilon_1/\sigma_1$ and $\tau_2 = \epsilon_2/\sigma_2$ are relaxation times of materials 1 and 2, respectively. J is steady-state current density per unit area, which flows normal to the interface, and σ_s is charge accumulated at the interface per area. Equations 1 and 2 show that charge is accumulated at the interface between two materials with different relaxation times, i.e., $\tau_1 \neq \tau_2$. In addition, accumulated charge σ_s is given as the product of time difference $\Delta\tau = \tau_2 - \tau_1$ and steady-state current density J. This is the Maxwell–Wagner effect and accounts for charge accumulation phenomena observed in all kinds of two-material interfaces.

Examples of Maxwell–Wagner Effect in Electronic Devices

Metal–Insulator–Metal System

Figure 2a shows a structure of metal–insulator–metal (MIM) device. Typical examples are seen in power cable systems, capacitors, etc., and these are modeled as shown in Fig. 2a. In MIM system, there is a metal–insulator interface. The conductivity of metal, σ_m, is generally very high, and it is infinity in the limit. On the other hand, the conductivity of insulator σ_{ins} is extremely low and it is nearly zero in an ideal case. In that situation, relaxation times of the metal and the insulator are $\tau_m \to 0$ and $\tau_{ins} \to \infty$. As a result, current flowing across the insulator becomes nearly zero in steady state, $J \to 0$, in the presence of external electric field $E = V/L$, where V is applied voltage across the insulator and L is the distance between the two electrodes. Substituting these relations into Eq. 2, the induced charge on metals $\sigma_M(=\sigma_s)$ is derived as

$$\begin{aligned}\sigma_M &= (\tau_{ins} - \tau_m)J \\ &= (\epsilon_{ins}/\sigma_{ins})\sigma_{ins}E_{ins} - (\epsilon_m/\sigma_m)\sigma_m E_m \\ &= \epsilon_{ins}E_{ins} - 0 = +\epsilon_{ins}V/L.\end{aligned} \tag{3}$$

In the derivation of Eq. 3, $E_m = 0$ is assumed with consideration of the fundamental property of metals, i.e., there is no electric field in metals. Equation 3 is the general equation of capacitors that describes induced charge on metals of MIM systems. That is, induced charge on electrodes is proportional to the applied voltage, V.

Metal–Double-Layer–Metal System (Two-Electrode System)

Figure 2b and c shows structures of metal–double-layer–metal device. Typical examples are seen in double-layer dielectric and insulator systems,

double-layer EL devices, double-layer solar cells, etc. In these systems, there is a double-layer interface. As mentioned in section "Maxwell–Wagner Effect," no charge is accumulated at the interface, when the relaxation times of the two materials are the same as $\tau_1 = \tau_2$ (see Fig. 2b), i.e., $q_s = 0$ and $\sigma_s = 0$ (see Eqs. 1 and 2). In that situation, the electric field formed in the two-material system satisfies the following relation:

$$\epsilon_1 E_1 = \epsilon_2 E_2. \tag{4}$$

Equation 4 suggests that the ratio of the electric field formed in the two materials depends on the dielectric constants of the two materials. Obviously for electrical insulation, the conductivity of insulator should be extremely small, but in a case of multilayer insulator system that satisfies the relation $\tau_1 = \tau_2 = \tau_3 \ldots$, charge accumulation at the interface is discarded even when a leakage current flows steadily. Hence the space charge field formation caused by accumulated charge is negligible, and the electric field distribution in the multilayer system is regulated only by the dielectric constants of materials used.

Generally, the electric field distribution in multilayer systems depends on distribution of accumulated charges at the interface, caused by the difference in relaxation times, i.e., $\tau_1 \neq \tau_2 \neq \tau_3 \ldots$. For a double-layer system with different dielectric relaxation times, $\tau_1 \neq \tau_2$, as illustrated in Fig. 2c, the electric field distribution depends on external electric field formed by applied voltage V as well as accumulated charges at the interface. Assuming charge $Q_s = S\sigma_s$ (S: electrode area) is accumulated at the interface, the electric fields in materials 1 and 2 are derived as follows:

$$E_1 L_1 = \frac{\sigma_M L_1}{\epsilon_1} + \frac{Q_s}{(C_1 + C_2)}. \tag{5.1}$$

$$E_2 L_2 = \frac{\sigma_M L_2}{\epsilon_2} + \frac{Q_s}{(C_1 + C_2)}. \tag{5.2}$$

Here $\sigma_M = Q_M/S$ is the charge induced on metal electrodes with $Q_M = C^* V$, where $C^* = C_1 C_2/(C_1 + C_2)$ is a series capacitance of the double-layer system with $C_1 = \epsilon_1 S/L_1$ and $C_2 = \epsilon_2 S/L_2$. Here σ_M is given by Eq. 3 with $L = L_1$, and accumulated charge is given by $Q_s = S\sigma_s$, using relation (2). That is, the electric field formed in the double-layer system changes depending on the accumulated charge at the interface, $Q_s = S\sigma_s$. From a viewpoint of device application in electronics, information on the charge Q_s and the electric field distribution is crucially important. These important information can be obtained by using novel experiments based on electric field-induced optical second-harmonic generation (EFISHG) measurement, charge modulation spectroscopy, and others.

For $\sigma_1 >> \sigma_2$, i.e., in a case of materials 1 and 2 showing conducting and insulating properties, respectively, $E_1 = 0$, $E_2 = V/L$ with $V = Q_s/C_2$ are obtained in steady state. That is, no electric field is formed in conducting material 1 and an electric field is present in material 2.

FET System (Three-Electrode System)

For double-layer devices with a sandwich structure as shown in Fig. 2, the potential along the double-layer interface is constant, while it is not the case when devices with a three-electrode configuration are discussed. As illustrated in Fig. 2d, a typical example is found in field-effect transistors, where the three electrodes are source, drain, and gate electrodes. In this system, the potential varies along the interface, depending on gate-source voltage, V_{gs}, and drain–source voltage, V_{ds}. What kind of potential distribution should be postulated? To answer this question is difficult, but a constant-slope potential distribution along the interface would be the most simple (gradual channel approximation) and most probable one. In other words, the electric field along the interface is constant and given by $-(V_{ds} - V_{gs})/L$. In that situation, the potential at $x = x$ from position of the source electrode edge is described as

$$V(x) = V_{gs} + \frac{x}{L}\left(V_{ds} - V_{gs}\right). \tag{6}$$

As the charge induced on the interface is given by $C_g(V(x) - V_{th})$ ($C_g = \varepsilon_g/d$, gate capacitance per unit area; V_{th}, threshold voltage; ε_g is the dielectric

constant of gate insulator, and d is the thickness of gate insulator), the charge induced at the interface between source and drain electrodes is given by

$$Q_{\text{int}} = C_g\left(V_{\text{gs}} - V_{ds}/2 - V_{\text{th}}\right) \tag{7}$$

per unit area. The accumulated charge given by Eq. 7 is of course the result of the Maxwell–Wagner effect at the gate insulator–active layer interface. In general FETs, the active layer is very conducting in comparison with the gate insulator; the relation $\tau_g >> \tau_{\text{act}}$ thus holds. Here τ_g and τ_{act} are dielectric relaxation times of gate insulator and the active layer. The situation of charge accumulation is very similar to the case expressed by Eq. 3 with $\tau_g = \tau_{\text{ins}}$ and $\tau_{\text{act}} = \tau_m$. In other words, the electric field in gate insulator is much higher than the field in the active layer. The drain–source current I_{ds} is the flow of charge Q_{int} at a constant velocity of $u\left(= \mu\overline{E}_{\text{ds}}\right)$ under the average electric field along the channel between drain and source electrodes, $\overline{E}_{\text{ds}} = V_{\text{ds}}/L$. Consequently, the I_{ds} is written as [2, 3]

$$\begin{aligned} I_{\text{ds}} &= Q_{\text{int}}\frac{hW}{h}\mu\frac{V_{\text{ds}}}{L} \\ &= \mu\frac{W}{L}C_g\left[\left(V_{\text{gs}} - V_{th}\right)V_{\text{ds}} - V_{\text{ds}}^2/2\right]. \end{aligned} \tag{8}$$

Here L/W and h are the channel length/width and thickness of the channel, respectively. Equation 8 is nothing but the equation of drain–source current of field-effect transistors, which is derived under gradual channel approximation on the basis of general semiconductor device physics. Nevertheless, it should be noted that Eq. 8 is derived on the basis of classical electrostatic analysis with consideration of the Maxwell–Wagner effect. In other words, charge accumulation at semiconductor–insulator interface is a result of the Maxwell–Wagner effect.

Note that in the case of constant current flowing across the FET channel, the potential distribution such as [5]

$$V(x) = V_0\sqrt{1 - \frac{x}{L}} + V_{\text{th}} \tag{9}$$

is obtained. Here V_0 is the potential below the source electrode with respect to the gate potential. Hence, the electric field along the channel is no longer constant and dependent on the position as

$$E(x) = \frac{V_0}{2\sqrt{L(L-x)}}. \tag{10}$$

In this case, the drain current is described as

$$I_{\text{ds}} = \mu\frac{W}{L}C_{\text{g}}\left[\left(V_{\text{gs}} - V_{\text{th}}\right)V_{\text{ds}} - \frac{2}{3}V_{\text{ds}}^2\right]. \tag{11}$$

The deviation of the current Eq. 11 from Eq. 8 is small, but the positional dependence of the electric fields is rather different between them. Hence, it is important to measure the electric field distribution along the FET channel.

As has been described above, the Maxwell–Wagner effect is actively working in field-effect transistors.

Maxwell–Wagner Effect as Interfacial Polarization Phenomena

As mentioned in section "Metal–Double-Layer–Metal System (Two-Electrode System)," charge σ_{s} is accumulated at the interface between two materials with different relaxation times, $\tau_1 \neq \tau_2$ (see Eq. 2). The interfacial charge accumulation process is a macroscopic dielectric polarization process, and it is called as interfacial polarization process. This process is a single polarization process, and the charge accumulation at the interface of the double-layer system illustrated in Fig. 2c is written as

$$\begin{aligned} &\sigma_{\text{s}}(t) = \sigma_s(1 - \exp(-t/\tau_{\text{MW}})) \quad \text{with} \\ &\tau_{\text{MW}} = \frac{C_1 + C_2}{G_1 + G_2} \end{aligned} \quad . \tag{12}$$

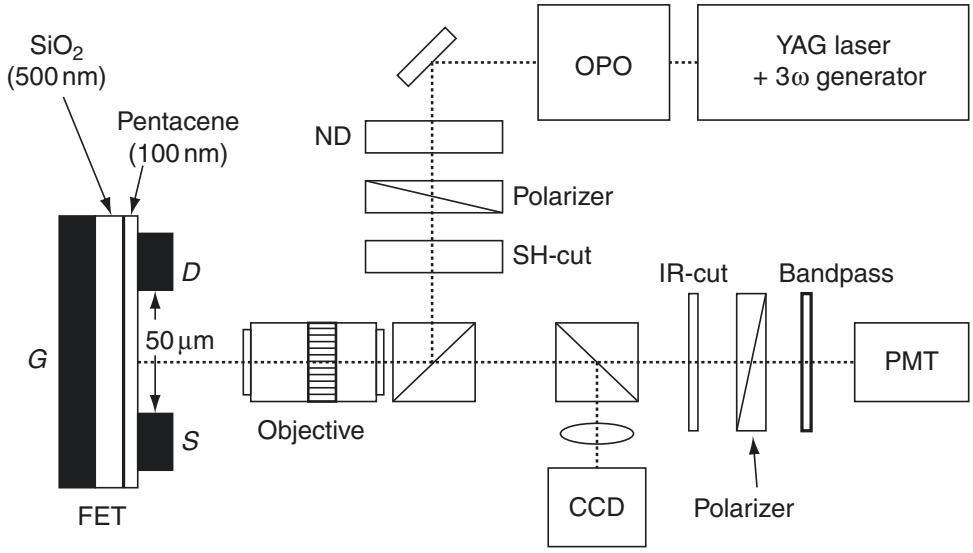
Here $C_1 = \epsilon_1 S/L_1$, $C_2 = \epsilon_2 S/L_2$, $G_1 = \sigma_1 S/L_1$, and $G_2 = \sigma_2 S/L_2$. τ_{MW} is Maxwell's dielectric relaxation time and it differs from relaxation times of the two materials, τ_1 and τ_2. It depends on not only material properties, dielectric

constant ϵ and conductivity σ, but also on geometrical factors, film thickness L and working area S. This is the characteristic of the interfacial charging caused by the Maxwell–Wagner effect.

Probing of Interfacial Charge Caused by Maxwell–Wagner Effect

As described in section "Metal–Double-Layer–Metal System (Two-Electrode System)," accumulated charge at the interface is a source of space charge field. This results in the modification of the electric field distribution in a double-layer device in accordance with amount of accumulated charge at the interface (see Eqs. 1 and 2). This scenario is also applicable to the case of any other device systems with material–material interfaces. In that situation, probing the electric field in devices is a very important research subject. Many techniques for probing the electric field distribution have been developed. Among them are electrical measurements, i.e., charge decay method, thermally stimulated current method, and so on. However, these techniques are indirect ones and cannot directly probe carrier process. On the other hand, optical method based on optical second-harmonic generation is available for directly probing carrier process related to the interfacial charging caused by the Maxwell–Wagner effect [5–10]. Figure 3 portrays the experimental setup of the optical second-harmonic generation method using a light source of YAG laser [8, 10]. The attached CCD camera visualizes carrier motion in active layer of the FET device [10], and the PMT probes the intensity of generated SHG signals.

Using this technique, the electric field distributions in electronic devices are obtained, e.g., organic field-effect transistors, electroluminescence devices, solar cells, and others [10, 11]. As charges injected from electrodes are excess charges, they act as a source of electric field. As a result, materials accepting these excess charges would be polarized. In other words, electron clouds of molecules are distorted in the presence of a space charge field arisen from injected carriers. In that situation, by probing this polarization, carrier motion in devices will be probed. The polarization to be probed is second-order nonlinear dielectric polarization that is generated in materials, due to coupling of electron clouds and electromagnetic waves. This is the electric field-induced nonlinear polarization, and this polarization is probed in the electric field-induced



Maxwell–Wagner Effect, Fig. 3 Optical setup for the SHG measurement. The sample has the FET structure

optical second-harmonic generation (EFISHG). In more detail, EFISHG is generated due to the coupling of electrons in molecules and electromagnetic waves $\mathbf{E}(\omega)$ in the presence of electrostatic local electric field $\mathbf{E}(0)$. Hence the EFISHG generation is allowed from even symmetric molecular systems, such as pentacene, tetracene, phthalocyanine, etc.

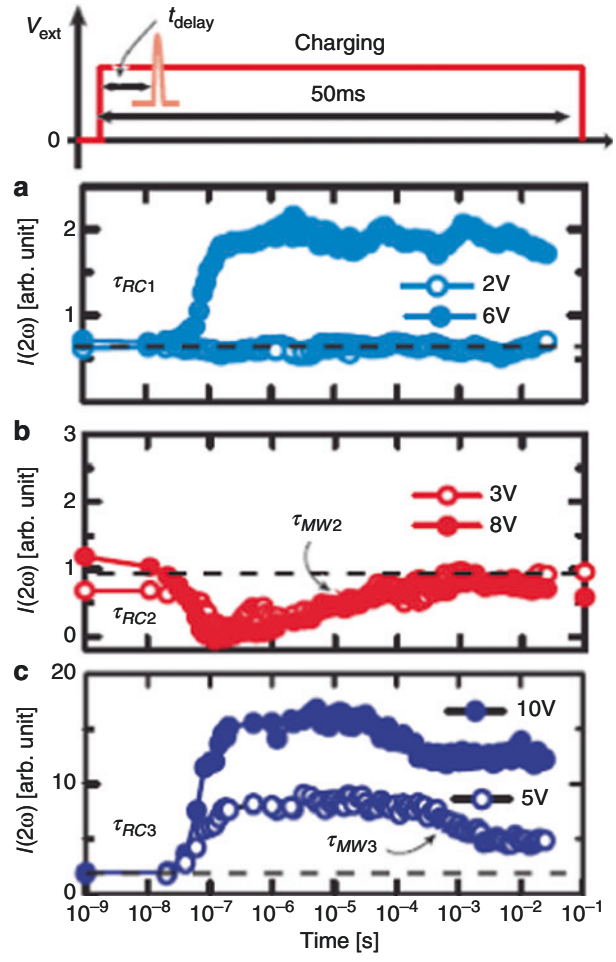
The EFISHG intensity is given as [10–13]

$$I(2\omega) \propto |\mathbf{P}(2\omega)|^2 = \left|\epsilon_0 \chi^{(3)}(0,\omega,\omega) \vdots \mathbf{E}(0)\mathbf{E}(\omega)\mathbf{E}(\omega)\right|^2, \quad (13)$$

where $\mathbf{P}(2\omega)$ is the nonlinear polarization induced in the materials, ϵ_0 is the dielectric permittivity of vacuum, $\chi^{(3)}$ is the third order nonlinear susceptibility tensor, $\mathbf{E}(0)$ is the electrostatic local field in materials, and $\mathbf{E}(\omega)$ is the electric field of incident electromagnetic wave. Generally, laser beam is used as a source of $\mathbf{E}(\omega)$. The local electric field is sum of the external electric field $\mathbf{E}_0$ and other fields such as space charge field, inner electric field due to workfunction difference of electrodes, and so on. Hence, carrier motion is probed as propagation of nonlinear polarization in the EFISHG measurements. In the following, some examples are shown.

Double-Layer Devices

Figure 4 shows transient EFISHGs generated from double-layer devices by applying a step voltage, i.e., ITO/polyimide/α-NPD/Au (Fig. 4a), ITO/polyimide/pentacene/Au (Fig. 4b),



Maxwell–Wagner Effect, Fig. 4 Transient SHG response from double-layer devices applying a step voltage: (**a**) α-NPD/polyimide device, (**b**) pentacene/polyimide device, and (**c**) α-NPD/Alq3 device

and ITO/α-NPD/Alq3/LiF/Al (Fig. 4c) [14–17]. Here pentacene and α-NPD are hole-transport layers, and Alq3 is an electron-transport layer, and polyimide is a blocking layer. The delay time (t_{delay}) in Fig. 4 represents the time interval between fundamental laser beam used as incident electromagnetic wave $E(\omega)$ and the wave front of the applied square voltage. In the SHG experiments, the SHG wavelength was chosen so that the electric field in α-NPD and pentacene layers was probed [14–17]. The black dash lines represent baselines for each device ("zero level"), with no applied voltage. At the beginning $t < 10^{-7}$ s, the three devices show similar transient characteristics. The SH intensity begins to increase at around $t \sim 10^{-7}$ s and then reaches some saturated value. This response time is independent of applied voltage and reflects charging to the electrodes. This response time of $\tau_{RC} \sim 10^{-7}$ s corresponds well with the response of the device circuit R_s C* (see section "Metal–Double-Layer– Metal System (Two-Electrode System)") for C*).

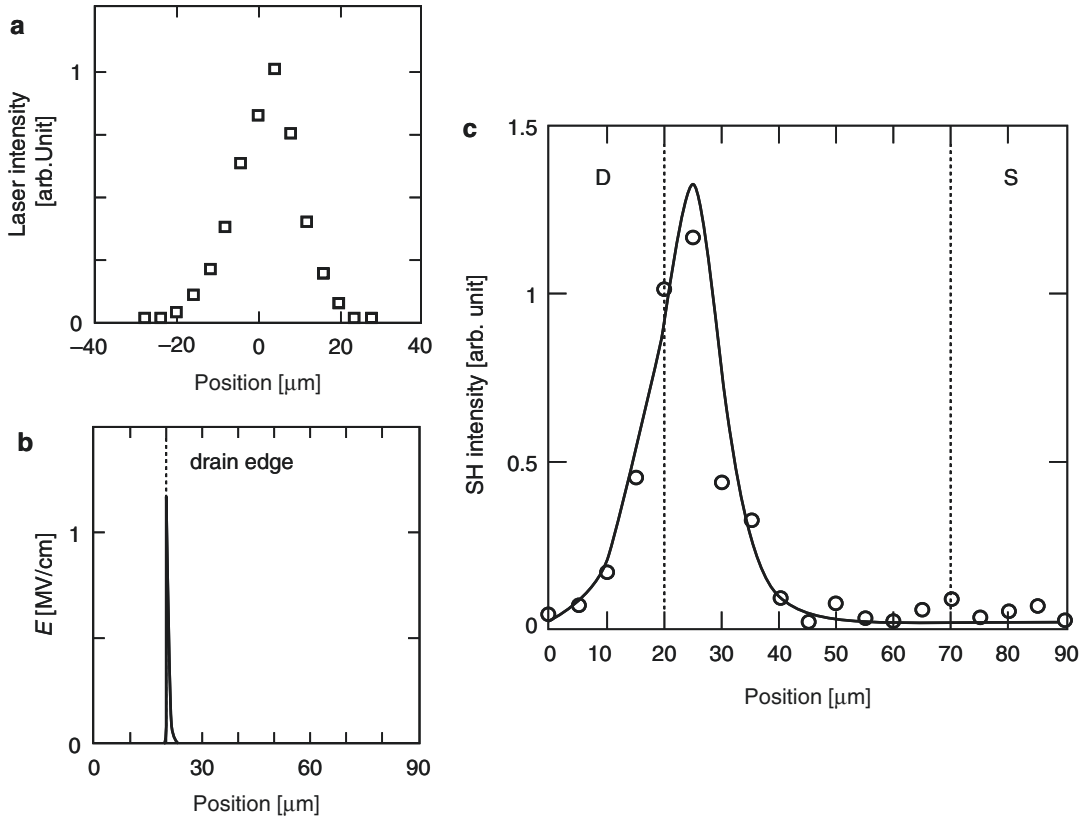
By contrast, the transient SHG in the region $t > 10^{-7}$ s depends on the device and applied voltage. Figure 4a shows the transient SHG of ITO/polyimide/α-NPD/Au. Here polyimide layer is a blocking layer for hole and electron injection, and only hole injection from the Au electrode to the α-NPD layer is allowed. However, in actual experiments, after the charging of the electrodes, the SH intensity saturates and never changes. The results suggest that there is no charge accumulation at the interface, i.e., $Q_s = 0$ (see Eq. 1). In other words, the α-NPD layer works like an insulating blocking layer. Figure 4b shows the transient SHG of ITO/polyimide/pentacene/Au devices. Here polyimide is a blocking layer. In this case, hole injection from the Au electrode is expected. Actually, after charging of the electrodes, rapid relaxation of SH intensity to zero level is significant. The decrease of the SHG intensity indicates charge accumulation at the pentacene/PI interface, i.e., $Q_s \neq 0$, as a result of the MW effect. Moreover, accumulated holes completely compensate the external electric field in the limit $t \to \infty$ (see Eqs. 1 and 2). Note that the relaxation time given by Eq. 12 satisfies the relation $\tau_{MW} \geq \tau_{RC}$ (10^{-7} s) in this experiment. Figure 4c shows the transient SHG of ITO/α-NPD/Alq3/LiF/Al device. This device is a typical two-carrier transportation OLED device with hole and electron injected from opposite electrodes. After the charging of the electrodes, these carriers are conveyed into α-NPD and Alq3 layers from opposite sides and reach at the α-NPD/Alq3 interface, where they recombine for electroluminescence. Nevertheless, from viewpoints of Maxwell–Wagner effect, electric field in this device is modified as excess charges remained at the interface. Observed SHG decaying after its saturation was due to this MW effect. In this case, the experiment shows that the relation $\tau_{MW} >> \tau_{RC}$ is satisfied. Interestingly, the electroluminescence was generated at the very time when the SHG decay was started.

OFET Devices

As described in section "FET System (Three-Electrode System)," charge is accumulated at the interface [18–20]. For double-layer devices illustrated in Fig. 2, it is easy to expect charge is homogeneously distributed at the interface, and as a result, constant potential is built at the interface. On the other hand, at the interface of three-electrode devices such as FET, the distribution is dependent on configuration structure and current flowing through the system, though the Maxwell–Wagner effect suggests that charge must be accumulated at the interface. Hence, it is a fundamental task to determine experimentally charge distribution along the channel. Probing electric field distribution along the channel is of great help and EFISHG measurement is available (see Eq. 13). Interest topics here are the charge distribution in steady state and charge accumulation process along the channel. In the following, these two points are focused.

In Steady State

As is well known, Laplace fields are theoretically calculated if the electrode configuration and its potential distributions are known a priori. Figure 5 shows the SHG experiment employed for OFETs to determine the electric field distribution along the channel. Figure 5a and b shows the intensity profile of the fundamental laser at a focal point and the field distribution calculated using the

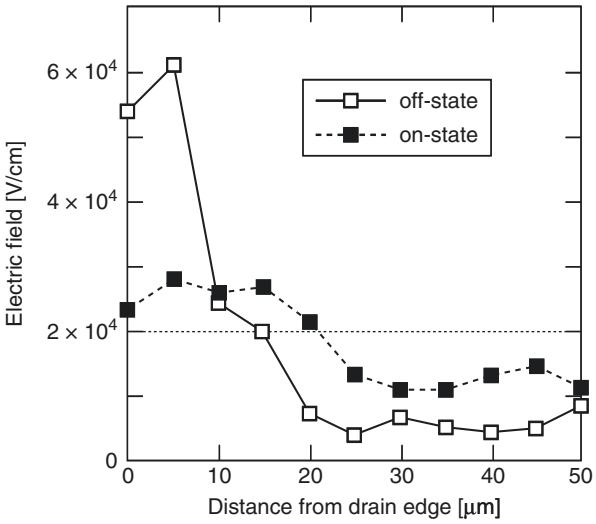


Maxwell–Wagner Effect, Fig. 5 SHG experiment. (**a**) Intensity profile of the fundamental laser at a focal point, (**b**) theoretical field distribution calculated using the finite element method, and (**c**) SHG profile estimated using (**a**) and (**b**). *Open circles* represent the observed SHG intensity at off-state of OFET device applying $V_{dg} = V_{sg} = -100$ V

finite element method on account of the actual electrode configuration. The separation between source and drain electrodes is $L = 50$ μm. At off-state of the device, i.e., without injected carriers along the channel, the potentials of the source and gate electrodes are the same, whereas potential difference between drain and gate (source) electrodes is -100 V. In this case, a Laplace field is formed in the device system, and the electric field is concentrated at the edge of drain electrodes due to edge effects. Figure 5c shows the SHG profile estimated from the convolution of the Laplace field and the intensity profile of the fundamental laser. Open circles represent the observed SHG intensity at each spot position. The theoretical curve reproduces well the experimental results, indicating actual electric field distribution in FETs is evaluated by carrying the deconvolution of the SHG profile. For simplicity, assuming the laser beam profile as delta function, the electric field distribution is roughly evaluated. Figure 6 shows the evaluated electric field distribution in the pentacene FET channel. At first glance, electric field is concentrated at the drain edge in the off-state (open squares). On the other hand, electric field distribution profile is getting flat in the on-state (filled squares), and it is approaching a constant electric field distribution (Eq. 6, gradual channel approximation).

Transient Charge Accumulation Process

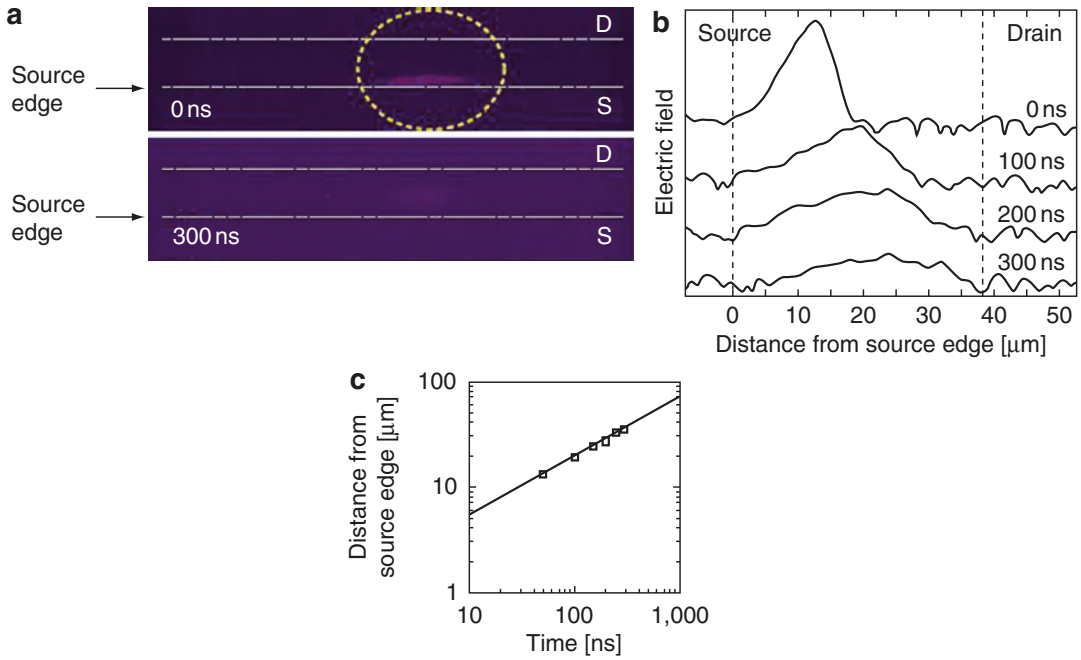
As mentioned in section "Probing of Interfacial Charge Caused by Maxwell–Wagner Effect," using a transient SHG technique can directly probe carrier motion in organic materials by monitoring the propagation of induced polarization



Maxwell–Wagner Effect, Fig. 6 Evaluated electric field distribution in the pentacene FET channel from SHG intensity: *open squares*, off-state; *closed squares*, on-state. *Dotted line* represents the average field calculated as V_{ds}/L, where L is the channel length

along with carrier motion [18–20]. According to the Maxwell–Wagner effect, at the interface between two materials with different relaxation times, the charge Q_s is accumulated. In OFETs, carriers injected from the source electrode are accumulated at the interface between active organic layer and gate insulator, and they are conveyed along the channel. Note that the injected carriers are excess charges and they are acting as a source of space charge field E_s. Hence actual local electric field E is a sum of the external field E_0 and space charge field, i.e., $E = E_0 + E_s$. Obviously, E_0 is the Laplace field and it never changes after applying voltages to the source, drain, and gate electrodes, and it is the same with that at time $t = 0$. On the other hand, E_s is the Poisson field and changing with time in accordance with carrier propagation, i.e., carrier injection, accumulation, and transport. Of course, it stabilizes in the limit $t \to \infty$, i.e., in steady state. Carrier migration along the channel is considered as an interfacial charging process and modeled using a ladder *RC* circuit model.

Visualization of carrier migration is of great help to verify the above carrier migration along the OFET channel. In such situation, transient EFISHG technique is available. Substituting $E = E_0 + E_s$ into $\mathbf{E}(0)$ of Eq. 13, it is found that SH intensity distribution along the channel changes with the distribution of space charge field E_s originating from moving carriers. For the transient SHG measurement, time-resolved microscopic SHG (TRM–SHG) measurement is useful. This approach allows to directly probe carrier motion. Figure 7 shows the TRM–SHG experiments using top-contact pentacene FET with Au source and drain electrodes. The samples were prepared on gate insulator surface with a 500 nm of SiO_2 gate insulator. The thickness of the pentacene layer was 100 nm. Figure 7a shows the TRM–SHG image obtained at various times after a positive voltage pulse of 70 V was applied to the source electrode with the gate and drain electrodes shorted and grounded. At $\tau = 0$ ns, the laser pulse coincided with the rising edge of the voltage pulse, and SHG signals were found near the edge of source electrode, indicating that carrier injection was just started, and the Laplace field E_0 is formed only around source electrode. Interestingly, as is clearly shown in the image, the emission band of the SHG signal gradually moved in the channel from source to drain electrode with time. Motion of the emission band from source electrode, not from drain, is the direct evidence of hole injection from the Au source electrode. That is, pentacene FET shows a p-type behavior. The SHG intensity profile plotted in Fig. 7b clearly shows that holes travel in direction from the Au source electrode to the drain electrode, on satisfying the square root time dependence (see Fig. 7c). Results show that carrier transient is an interface charge propagation process, and the square root time dependence verifies the RC ladder model.



Maxwell–Wagner Effect, Fig. 7 (Color online) TRM–SHG measurement. (**a**) SHG image from the FET channel for different delay times 0, 300 ns, applying positive voltage pulse to the source electrode. The area of the laser spot is indicated by the *dashed circle*. (**b**) Plots of SHG intensity profile. (**c**) Log–log scaled time evolution of the peak position of the SHG intensity

Future Directions

The Maxwell–Wagner effect is based on the classical electromagnetic theory and accounts for charge accumulation phenomena at the interface between two materials. It can widely apply to all kinds of material interfaces, from macro- to nanomaterials. Also it can apply at molecular level. Many new materials will be coming out, but physics related to charge accumulation is the same.

Cross-References

- ▶ Electrode–Organic Interface Physics
- ▶ Field Electron Emission from Nanomaterials
- ▶ Flexible Electronics
- ▶ Optical and Electronic Properties

References

1. Maxwell, J.C.: A Treatise on Electricity and Magnetism, vol. 1, 3rd edn, p. 450. Dover, New York (1954) (Chap. X)
2. Tamura, R., Lim, E., Manaka, T., Iwamoto, M.: Analysis of pentacene field effect transistor as a Maxwell-Wagner effect element. J. Appl. Phys. **100**, 114515 (2006)
3. Lim, E., Manaka, T., Tamura, R., Iwamoto, M.: Maxwell-Wagner model analysis for the capacitance-voltage characteristics of pentacene field effect transistor. Jpn. J. Appl. Phys. Part 1 **45**, 3712 (2006)
4. Lim, E., Manaka, T., Iwamoto, M.: Analysis of carrier injection into a pentacene field effect transistor by optical second harmonic generation measurements. J. Appl. Phys. **101**, 024515 (2007)
5. Weis, M., Lin, J., Taguchi, D., Manaka, T., Iwamoto, M.: The charge transport in organic field-effect transistor as an interface charge propagation: the Maxwell-Wagner effect model and transmission line approximation. Jpn. J. Appl. Phys. **49**, 071603 (2010)
6. Yamada, D., Manaka, T., Lim, E., Tamura, R., Weis, M., Iwamoto, M.: Probing of electric field in pentacene using microscopic optical second harmonic generation. J. Appl. Phys. **103**, 084118 (2008)
7. Weis, M., Manaka, T., Iwamoto, M.: Origin of electric field distribution in organic field-effect transistor: experiment and analysis. J. Appl. Phys. **105**, 024505 (2009)
8. Manaka, T., Lim, E., Tamura, R., Iwamoto, M.: Direct imaging of carrier motion in organic transistors by optical second harmonic generation. Nat. Photonics **1**, 581 (2007)
9. Lim, E., Yamada, D., Weis, M., Manaka, T., Iwamoto, M.: Probing of channel region in pentacene field effect

transistors by optical second harmonic generation. Chem. Phys. Lett. **477**, 221 (2009)
10. Iwamoto, M., Manaka, T., Weis, M., Taguchi, D.: Probing and modeling of interfacial carrier motion in organic devices by optical second harmonic generation. J. Vac. Sci. Technol. B **28**(4), C5F12 (2010)
11. Iwamoto, M., Manaka, T., Yamamoto, T., Lim, E.: Probing motion of electric dipoles and carriers in organic monolayers by Maxwell displacement current and optical second harmonic generation. Thin Solid Films **517**, 1312 (2008)
12. Shen, Y.R.: The Principles of Nonlinear Optics. Wiley, New York (1984) (Chap. 1)
13. Manaka, T., Lim, E., Tamura, R., Yamada, D., Iwamoto, M.: Probing of the electric field distribution in organic field effect transistor channel by microscopic second-harmonic generation. Appl. Phys. Lett. **89**, 072113 (2006)
14. Shibata, Y., Nakao, M., Manaka, T., Lim, E., Iwamoto, M.: Probing electric field distribution in underlayer of an organic double-layer system by optical second-harmonic generation measurement. Jpn. J. Appl. Phys. **48**, 021504 (2009)
15. Zhang, L., Taguchi, D., Li, J., Manaka, T., Iwamoto, M.: Probing of interfacial charging and discharging in double-layer devices with a polyimide blocking layer by time-resolved optical second harmonic generation. J. Appl. Phys. **108**, 093707 (2010)
16. Taguchi, D., Inoue, S., Zhang, L., Li, J., Weis, M., Manaka, T., Iwamoto, M.: Analysis of organic light-emitting diode as a Maxwell-Wagner effect element by time-resolved optical second harmonic generation measurement. J. Phys. Chem. Lett. **1**, 803 (2010)
17. Taguchi, D., Zhang, L., Li, J., Weis, M., Manaka, T., Iwamoto, M.: Analysis of carrier transients in double-layer organic light emitting diodes by electric-field-induced second-harmonic generation measurement. J. Phys. Chem. C **114**, 15136 (2010)
18. Manaka, T., Liu, F., Weis, M., Iwamoto, M.: Diffusion like electric-field migration in the channel of organic field-effect transistors. Phys. Rev. B **78**, 121302R (2008)
19. Manaka, T., Liu, F., Weis, M., Iwamoto, M.: Influence of traps on transient electric field and mobility evaluation in organic field-effect transistors. J. Appl. Phys. **107**, 043712 (2010)
20. Nakao, M., Manaka, T., Weis, M., Lim, E., Iwamoto, M.: Probing carrier injection into pentacene field effect transistor by time-resolved microscopic optical second harmonic generation measurement. J. Appl. Phys. **106**, 014511 (2009)

Mean Field Approaches to Many-Body Systems

► Ab Initio DFT Simulations of Nanostructures

Mechanical Behavior of Nanostructured Metals

► Mechanical Properties of Nanocrystalline Metals

Mechanical Characterization of Nanostructures

► Reliability of Nanostructures

Mechanical Energy Dissipation

► Surface Dissipations in NEMS/MEMS

Mechanical Properties of Biological Materials

► Mechanical Properties of Hierarchical Protein Materials

Mechanical Properties of Hierarchical Protein Materials

Markus J. Buehler[1] and Graham Bratzel[1,2]
[1]Laboratory for Atomistic and Molecular Mechanics, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Synonyms

Biomolecular mechanics; Hierarchical mechanics; Materiomics; Mechanical properties of biological materials

Definition

Mechanical properties of proteins define the behavior of the deformation and failure of proteins and protein-based biological and synthetic materials.

Overview

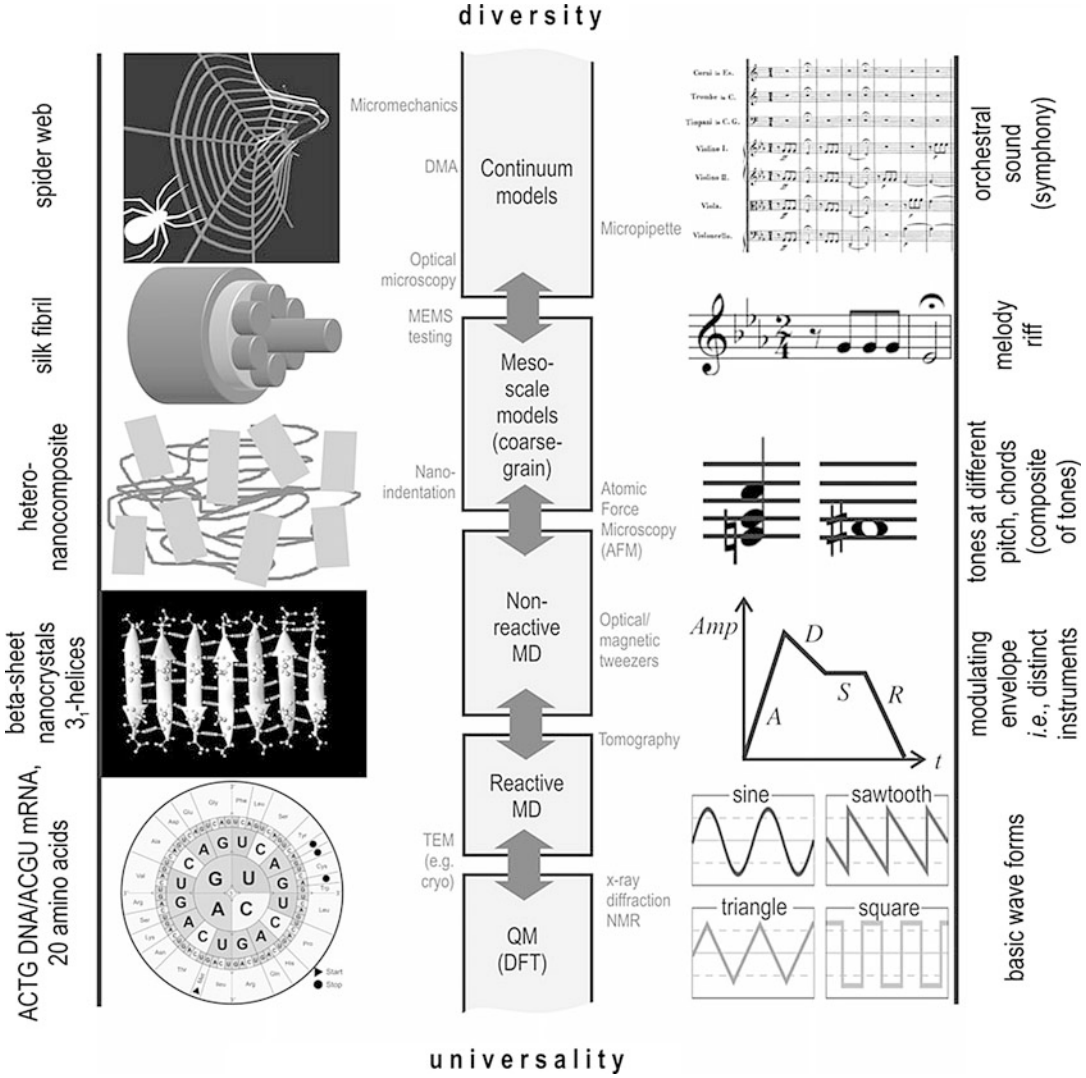
With only 20 standard amino acids as universal building blocks, evolutionary adaptation has resulted in a multitude of polypeptides and protein structures with a wide range of properties and applications [1]. A fundamental application of protein materials is mechanical, providing static and dynamic support and defenses for the organisms they comprise. To achieve specialized mechanical properties, animal tissue shows a wide range of complex, structured composites composed of proteins (e.g., collagen, keratin, chitin) and inorganic minerals (e.g., calcite, hydroxyapatite, and aragonite) [2, 3]. Instead of summarizing primarily single-molecule protein mechanics studies [4–8], the main goal of this entry is to give an introduction to the current knowledge of the nanomechanical properties of important structural protein materials and to explain the generic role of the hierarchical structures that give rise to exceptional mechanical properties that are superior to many synthetic engineering materials. For example, shell, bone, and antler are hierarchically structured composites of collagen and minerals and have toughnesses that are about one order of magnitude greater than engineering ceramics [9]. Natural polymers and polymer composites like collagen, keratin, and silk show elastic moduli and tensile strengths larger than those of engineering polymers made from similar building blocks. For example, silks have elastic moduli between 2 and 20 GPa and strengths in the range from 0.3 to 2 GPa [9]. Of synthetic polymers, only Kevlar has a higher stiffness at 200 GPa and a strength up to 4 GPa, which it achieves through a highly oriented molecular structure which is, in fact, characteristic of biological materials. Natural elastomers, including skin, muscle, cartilage, abductin, resilin, and elastin, exhibit elastic moduli and densities similar to those of engineering elastomers. Muscle is unique in combining a low modulus on the order of an elastomer with a high strength on the order of steel [9]. Natural cellular materials like cancellous bone have low densities because of the high volume fractions of voids they contain [10].

Whether natural composite, polymer, or elastomer, many biological materials achieve their notable mechanical properties through a balance of strong covalent bonds and weaker noncovalent bonds (e.g., hydrogen bonds) in a multilevel hierarchical structure that spans from nanoscopic to macroscopic length scales [11, 12]. Known as the universality-diversity paradigm [13] and summarized in Fig. 1, a limited number of universal protein structures or mineral components can provide a wide diversity of geometrical arrangements within the natural composites that give rise to exceptional bulk mechanical properties as compared to the material properties of the individual constituents. Certain specially constructed geometrical arrangements can enhance bulk properties much more than the rule of mixtures can predict. This puts the concept of geometry, assembly, and, more generally, the architecture of hierarchical structures at the forefront of the science of protein materials.

This entry will first discuss how protein materials are limited by an intrinsic strength limit at the nanoscale that exemplifies the constraints under which many biological systems operate, how these are overcome, and how such materials exemplify a powerful paradigm of how weakness is turned into strength. Next, the multilevel structures of silk, collagen fibers, and bone are presented as specific examples of how specialized geometrical arrangement of respective protein and mineral components into a hierarchical structure can overcome the strength limit at the base level to yield highly functional macroscale materials that rival engineering composite materials.

Hierarchical Structures Overcome Strength Limits

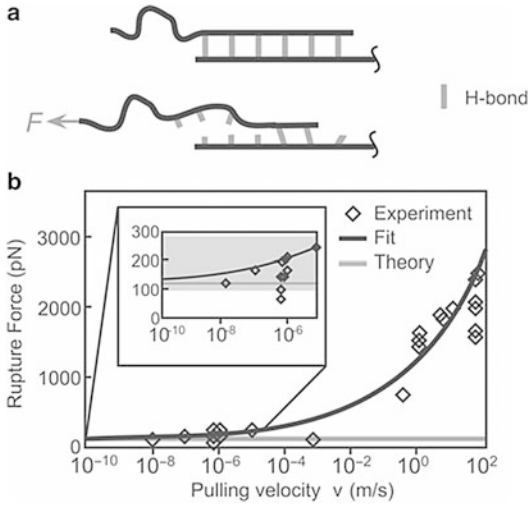
While covalent peptide bonds make up the backbone of a protein, noncovalent bonds, in particular



Mechanical Properties of Hierarchical Protein Materials, Fig. 1 Overview over the universality-diversity paradigm (*UDP*) [11, 13]. The hierarchical structure of protein materials is analogous to the construction of symphonic music. In much the same way that all music is made from a superposition of "simple" sound waves, a limited list of universal building blocks – here, amino acids of proteins – can be combined in various ways at multiple hierarchical levels to yield a diverse groups of protein-based biomaterials. Only a small subset of this diverse group exhibit exceptional mechanical properties, e.g., an orb-weaver spider web or bone, akin to Bach or Mozart

weak bonds such as hydrogen bonds (H-bonds), are indispensable to biological function in defining the native folding and stability of protein structures. The amino acid sequence defines a protein's secondary and higher-level structures via folding, but arrays of parallel adjacent H-bonds govern major universal secondary and tertiary structural motifs, chiefly α-helices and β-sheets, that organize the three-dimensional structure of polypeptides and thus determine their mechanical properties at a fundamental level. For example, the rupture of parallel interstrand H-bonds during β-sheet failure defines the true ultimate strength of such proteins (Fig. 2a for a schematic). Atomistic simulation combined with single-molecule force microscopy studies [7]

M



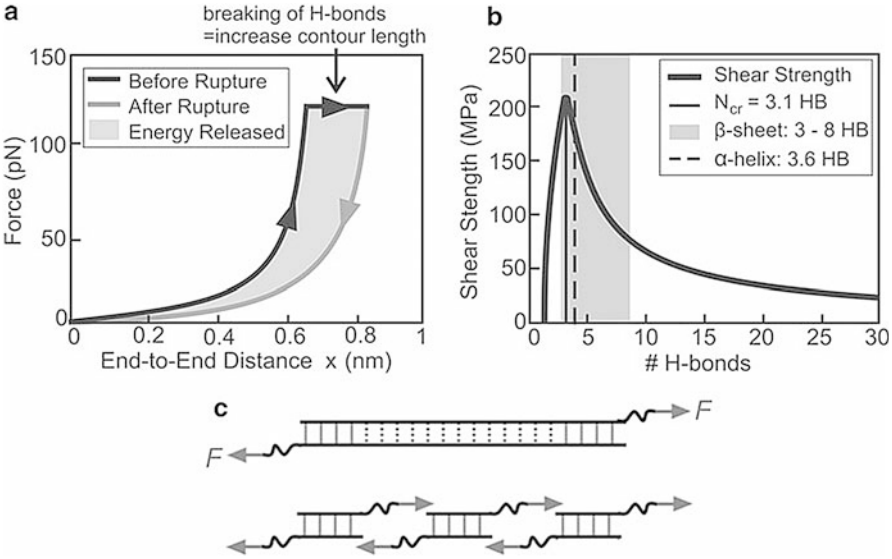
Mechanical Properties of Hierarchical Protein Materials, Fig. 2 Strength limit of H-bond clusters under shear loading (see *panel* (**a**) for schematic of geometry before and after rupture). As shown in *panel* (**b**), at vanishing pulling rates (see *inset* for a detailed view of the data), clusters of H-bonds resist a force limit of 100–300 pN. At high pulling rates, considerably more force is needed to unfold and rupture H-bonds, according to experimental and computational results from fibronectin, immunoglobin, and titin deformation (for references to source data, see Ref. [18]). This data reflects the rate dependence of failure seen in many biomaterials [18]

have shown that proteins rich in β-sheets exhibit higher rupture forces since they employ parallel strands with numerous H-bonds that act as strong mechanical clamps under shear loading. At experimental pulling rates, most mechanical proteins exhibit a rupture force of a few hundred piconewtons (pN) [14], as seen in Fig. 2. An asymptotic limit at very low pulling rates can be inferred from a limit analysis for deformation through equilibrium, which suggests an upper strength limit for the unfolding β-sheet-rich proteins [15–17] (see results in Fig. 2b). Based on this analysis for typical single-molecule length scales at a temperature of 300 K and a vanishing pulling rate, the asymptotic force of rupture is found to be on the order of 100–300 pN [18] (Fig. 2b). This asymptotic force limit depends strongly on the H-bond dissociation energy, which can vary with solvent conditions and which explains the range of strengths obtained.

The asymptotic strength limit for the unfolding β-sheet-rich proteins can be explained based on fracture analysis inspired by Griffith's analysis that was originally applied to solids, and is here used to describe the strength of protein domains [18]. We briefly review the key concepts behind this analysis that is also shown in Fig. 3. At low force levels, the elasticity of the polypeptide chain is primarily due to entropic effects (rather than energetic effects). The Marko-Siggia Worm-like Chain (WLC) model is one of the most widely used expressions to predict the entropic elasticity of the polypeptide backbone. Indeed, comparative studies provide substantial evidence that WLC theory is a good model for the behavior of individual, unconstrained polypeptide chains [19]. In the WLC theory, the force is a function of end-to-end chain length, the persistence length, the contour length, and the temperature. The asymptotic strength limit for the unfolding β-sheet-rich proteins can be obtained by considering the energy dissipation during the process of first stretching of a β-strand protein (described by the WLC theory) followed by subsequent breaking of H-bonds, where the energy released must equal the energy stored in the H-bonds (Fig. 3a); where the energy released is properly normalized to describe how many H-bonds exist per unit contour length of the protein. This analysis is similar to a fracture analysis of solids where the elastic energy stored in the material (here often described using linear elastic continuum theory and not by the WLC theory as in the case of proteins) is being dissipated into the creation of new surfaces.

For a given H-bond dissociation energy the force limit is independent of the number of H-bonds being sheared simultaneously, given that the number of H-bonds in a deformed cluster is *larger* than a critical H-bond cluster size N_{cr}, which can be calculated for each case [18, 20]. This is because it is found that no more than $N_{cr} = 3$–4 H-bonds can rupture concurrently, thus defining a critical H-bond cluster size beyond which the strength can no longer increase and at which scale the asymptotic strength limit is reached.

Therefore, while the shear strength of β-sheets with *less* than $N_{cr} = 3$–4 H-bonds benefits from



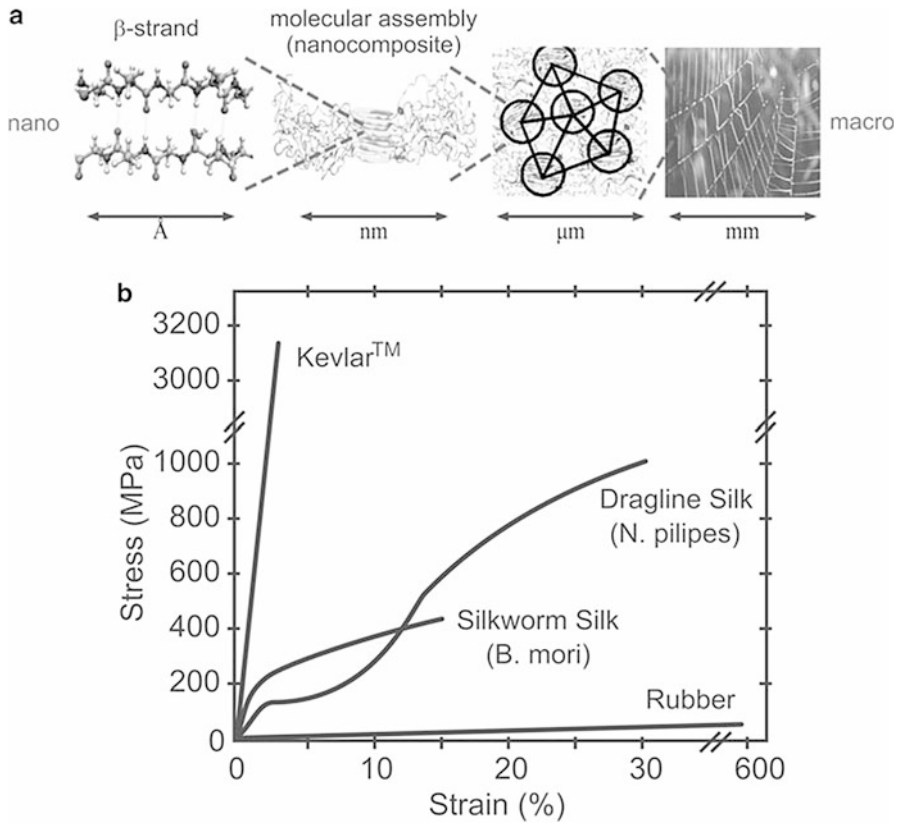
Mechanical Properties of Hierarchical Protein Materials, Fig. 3 Mechanistic view into the H-bond rupture process (**a**), and scaling of shear strength over size of H-bond cluster (**b**). (**a**) The area enclosed between loading and unloading curves is equivalent to the change in free energy, i.e., the adhesion energy per unit length. The increase in contour length is due to fracture, which involves breaking of H-bonds and as a result, the release of previously attached polymer length that now contributes to entropic elasticity. (**b**) A β-sheet of 3–4 residues in length displays the theoretical maximum shear strength, meaning that up to 3–4 H-bonds can rupture simultaneously. The structure of H-bond clusters in α-helices are in this range, with 3.6 residues per turn; and similarly, most β-sheet protein domains feature 3–8 H-bonds in each cluster. (**c**) Illustration of the physical concept behind the scaling law shown in *panel* (**b**), where the breaking of H-bond clusters is confined to a critical cluster size of $N_{cr} = 3–4$ (Figures adapted from Ref. [20])

addition of H-bonds, the shear strength (the maximum force divided by the sheared area) decays for β-sheets of 4 H-bonds or longer, as seen in b, as the H-bonds along the full length of the β-sheets do not rupture concurrently since they can only rupture in groups of $N_{cr} = 3–4$ H-bonds (see Fig. 3c for an illustration of this concept). Notably, following this scaling law, most protein materials are indeed found to intersperse many small β-sheets close in size to N_{cr}, in order to turn the weakness of the building block (H-bond) into strength by using geometry as a paradigm to achieve this goal. For example, β-sheets show a typical H-bond cluster size of 3–8 H-bonds across thousands of proteins in the Protein Data Bank, and α-helices have 3–4 H-bonds per turn [18, 20]. It is interesting to note that the scaling law of the shear strength, increasing for small values of H-bonds to a maximum value followed by a decay represents a biological material analogue of the Hall–Petch relationship known from polycrystalline engineering materials (e.g., metals) [21]. The intrinsic strength limit can only be overcome and scaled up by using structural hierarchies that circumvent β-sheet saturation and other limiting factors at critical length scales. Silk, as will be discussed in the following section, is a prime example of how H-bond cooperativity within β-sheets of controlled lengths and a hierarchical structure of networks and fibrils can overcome molecular strength limits and create unusually strong and tough protein fibers at the macroscale.

Structure and Mechanics of Exemplary Hierarchical Protein Materials

Silk

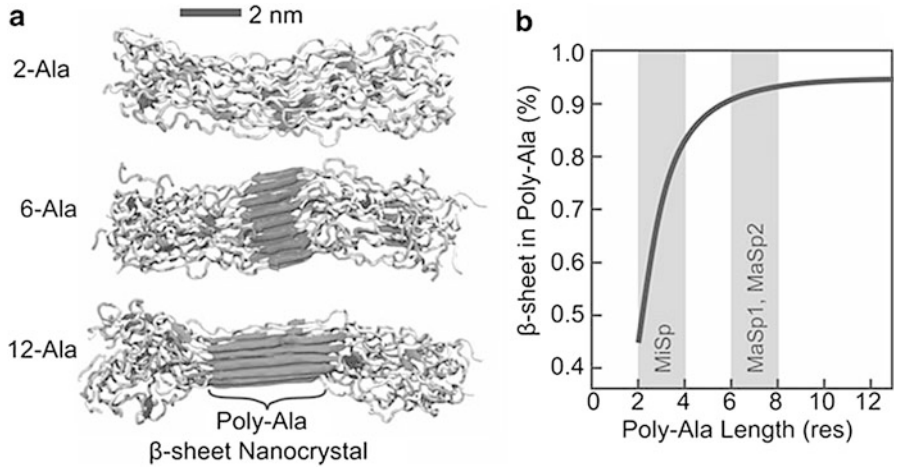
Silk is a hierarchically structured protein fiber (Fig. 4a) with a high tensile strength and great extensibility, making it one of the toughest and



Mechanical Properties of Hierarchical Protein Materials, Fig. 4 Hierarchical structure of silk from nano to macro (**a**) and a summary of stress–strain plots superior strain-hardening behavior of dragline and silkworm silks (**b**). Rubber is extensible but not strong. Kevlar™ is stiff and strong, but not tough. Silk, especially dragline silk, behave with a combination of strength and toughness not typically found in synthetic materials (Figure adapted from Ref. [31])

most versatile materials known [22–24]. In contrast to synthetic polymers based on petrochemicals, silk is spun into strong and totally recyclable fibers at ambient temperatures, low pressures, and with water as the solvent. However, biomimetic reproductions of silk remain a challenge because of silk's unique microstructural features that can only be achieved by controlled self-assembly of protein oligomers with molecular precision [25, 26]. Unlike silkworms, some spiders can use different glands to create up to seven types of silk, from the strong structural dragline to the viscoeleastic capture silk and tough egg sack casing [23]. Dragline silk, containing a high fraction of densely H-bonded domains, is used to provide the structural frame for the web and has an elastic modulus of around 10 GPa [24]. Capture silk, on the other hand, is a viscid biofilament containing cross-linked polymer networks and has an elastic modulus that is comparable to that of other elastomers [27].

The stress–strain behavior of silks is compared to Kevlar™ and rubber in Fig. 4b. While rubber is extensible, and Kevlar™ is stiff and strong, silks feature a combination of strength and toughness not typically found in synthetic materials. It is known that β-sheet crystals at the nanoscale play a key role in defining the mechanical properties of silk by providing stiff and orderly cross-linking domains embedded in a semi-amorphous matrix that consists predominantly of less orderly structures [28, 29]. These β-sheet nanocrystals, bonded by means of assemblies of H-bonds, have dimensions of a few nanometers and



Mechanical Properties of Hierarchical Protein Materials, Fig. 5 Tuning the structure of spider silk by controlling the genetic makeup of silk proteins, used to define the structure of the material at the nanoscale (obtained using Replica Exchange Molecular Dynamics). (**a**) Computational structure predictions of MaSp1 agglomeration in dragline silk show that the size of the β-sheet nanocrystal may be tuned by changing the poly-alanine length [32]. (**b**) Percentage of β-sheet nanocrystal formation in the poly-alanine region of the silk protein, measuring the success of the genetic signal to achieve the desired structure. The structure prediction indicates a critical poly-alanine length of 3–4 alanine residues for the establishment of a well-defined nanocrystal. The MaSp1 and MaSp2 proteins of dragline silk typically have poly-alanine repeats 6–8 residues in length, while the MiSp protein of capture silk has poly-alanine repeats of only 2–4 residues [33]

constitute roughly 10–15 % of the silk volume. When silk fibers are stretched, the β-sheet nanocrystals reinforce the partially extended and oriented macromolecular chains by forming interlocking regions that transfer the load between chains under lateral loading, similar to their function in other mechanical proteins [30]. The hierarchical network of spider draglines, contrasting synthetic elastomers like rubber, enables quick energy absorption and efficiently suppresses vibration during an impact [31]. Rubbers, composed of random polymer chains, display an elasticity primarily due to the change in conformational entropy of these chains. In contrast, the amorphous chains in silk filaments are extended and held in partial alignment with respect to the fiber axis in its natural state, resulting in remarkably different mechanical behaviors from rubbers [29, 30]. Mesoscale simulations reveal that the characteristic, cross-linking β-sheet nanocrystals govern the coupled effects of silk network realignment and strain-hardening behavior beyond the regime of unfolding amorphous chains [28].

Advances in nanotechnology may allow for an optimization of silk and silk-like synthetic fibers for specific applications by modifying the governing network at critical length scales. At the base of the structural hierarchy, β-sheet nanocrystal size may be tuned though genetic modification and control of the self-assembly process. Computational structure predictions of MaSp1 agglomeration in dragline silk, seen in Fig. 5a, shows that the size of the β-sheet nanocrystal may indeed be tuned by controlling the poly-alanine length [32], which in turn affects the stability and "quality" of the resulting nanostructures. A critical poly-alanine length of 3–4 residues is predicted for the amyloidization of a defined and stable β-sheet nanocrystal. A sampling of amino acid sequences of both orb-weaving and non-araneoid spiders shows that the MaSp1 and MaSp2 proteins of strong and tough dragline silk typically have poly-alanine repeats 6–8 residues in length, while the MiSp protein of elastic capture silk has poly-alanine repeats of only 2–4 residues [33]. Combined with the structure predictions, this shows

that the size of the β-sheet nanocrystals has a fundamental effect on the mechanical properties of the macroscopic silk threads (Fig. 5b). Further computational studies of spider silk proteins reveal that the nanoscale confinement of β-sheet nanocrystals in silks indeed has a fundamental role in achieving great stiffness, resilience, and fracture toughness at the molecular level of the structural hierarchy [28, 34], where H-bond cooperativity depends quite strongly on the size of the crystals and breaks down once β-sheet nanocrystals exceed a critical size. Larger β-sheet nanocrystals are softer and fail catastrophically at much lower forces owing to crack-like flaw formation. This catastrophic breakdown leads to rapid disintegration of silk fibers, whereas reducing the nanocrystal dimensions below 3 nm increases the ultimate strength and the modulus manyfold [34]. Furthermore, noncovalent H-bonds are able to reform during stick–slip deformation, allowing smaller crystals a self-healing ability until complete rupture occurs.

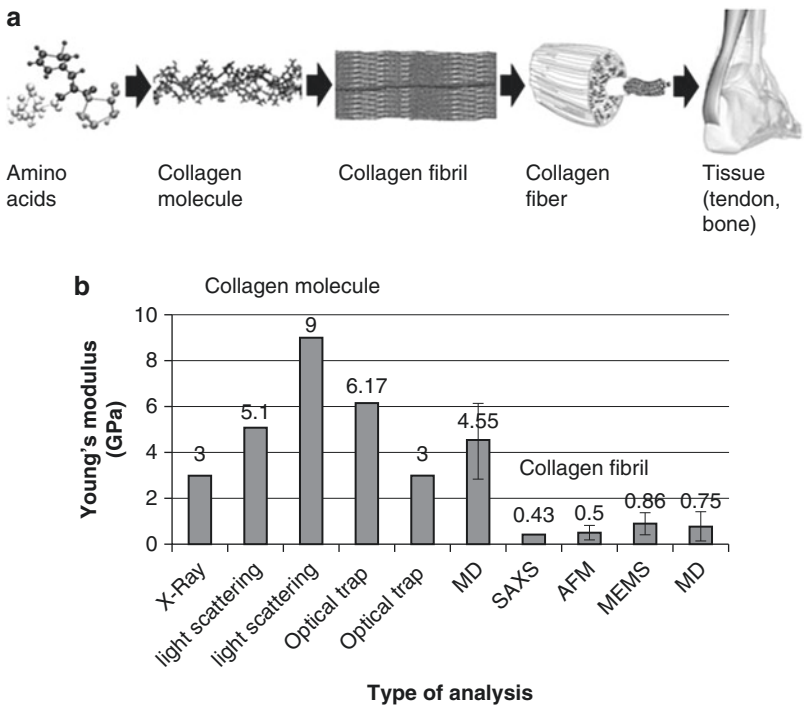
It is also noted that at the macroscale, experiments have demonstrated that when the size of β-sheet nanocrystals is reduced by moderating the reeling speed or by metal infiltration, silk shows enhanced toughness and greater ultimate strength, exceeding that of steel and other engineered materials [30]. By understanding and controlling how H-bonded silk threads can reach great strength and toughness and overcome strength limitations at the molecular scale, the behavior of a broader class of β-sheet-rich protein materials and similar synthetic composite fibers can also be tuned and adapted.

Collagen-Based Protein Materials

The concepts discussed above are important for many other protein materials. Collagen, for example, is present in various forms in almost all structural animal tissue and represents how the limitations of protein subunits can be overcome by the complex way they are assembled (Fig. 6a). Skin, tendon, and cartilage are all largely collagenous composites [2]. In skin, collagen is sandwiched between a basement membrane and an overlying keratinized epidermis. In tendon, collagen fibers form rope-like structures which make up 70–80 % of the volume, with the remainder being a combination of non-collagenous protein, polysaccharides, and inorganic salts. In cartilage, the collagen fibers are in a proteoglycan matrix with a small fraction of elastin fibers. Bone, shell, and antler are composites of calcite, hydroxyapatite, or aragonite platelets dispersed in a helical matrix of collagen. A diverse class of materials result from the universal collagen protein.

Variations in the polypeptide sequence of collagen chains gives rise to different types, termed Type I through Type XIX. Type I and Type III collagen are the most abundant, with Type I in skin and bone and Type III in blood vessel walls [2]. Three collagen chains, each of about 1,000 amino acids, arrange in the basic subunit of functional collagen, a left-handed triple-helix 300 nm long and 1.5 nm in diameter called tropocollagen. Typically, tropocollagen aggregates into a microfibril of five triple helices. About 1,000 microfibrils aggregate laterally and end to end to form a collagen fibril that displays periodic banding every 65 nm. Comparing the Young's moduli of tropocollagen and a collagen fibril, seen in Fig. 6b, shows that the structures at distinct hierarchical levels have very different mechanical properties, where single collagen molecules are much stiffer than collagen fibrils. About 500 fibrils aggregate into a collagen fiber, which then aggregate into fiber bundles. The diameters and specific hierarchical structures of fibrils and fibers vary by function among skin, tendon, and muscle. A bundled-fiber architecture connected by solvent-mediated intermolecular and intramolecular cross-linking of aldehyde groups provides stability, strength, and viscoelasticity [35, 36].

Beyond collagen fibers, a broad range of materials are available with the addition of mineral platelets within collagen. Bone and nacre belong to a family of natural ceramic composites that also includes coral, antler, and dental enamel [37, 38] that are made of mineral particles (e.g., hydroxyapatite, calcite, or aragonite) in a matrix of collagen. Their elastic moduli are lower than those of engineering ceramics, but their tensile strengths are similar, and their toughnesses are greater by a factor of 10 due to an ordered,

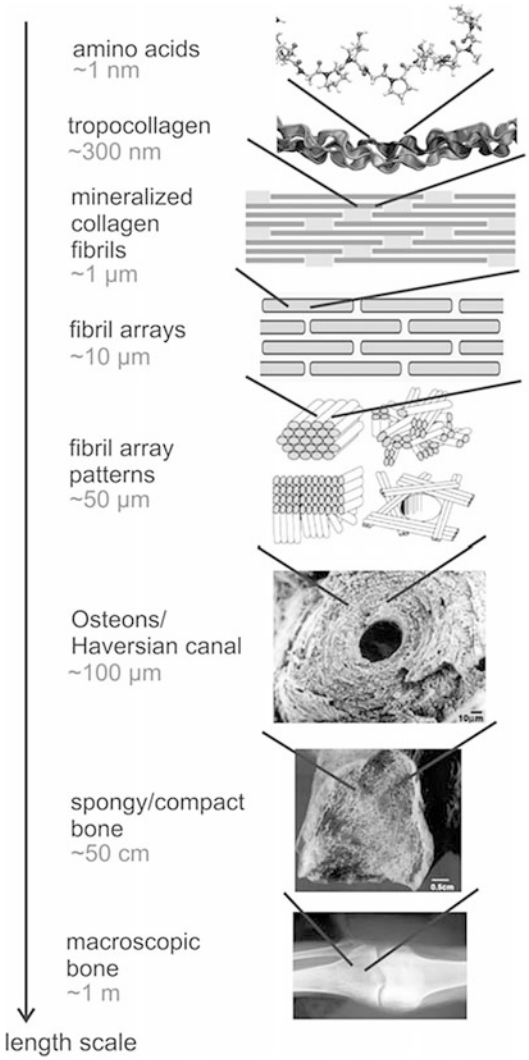


Mechanical Properties of Hierarchical Protein Materials, Fig. 6 Structure (**a**) and scale-dependent mechanical properties of collagen at the scale of single molecules and fibrils. (**a**) Hierarchical structure of collagenous tissues. The specific scales highlighted here are the collagen molecule and the collagen fibril. (**b**) A summary of small-strain Young's moduli for a single collagen molecule and a collagen fibril shows that the structure of fibrils sacrifice stiffness for increased extensibility and toughness (Data taken from various experimental and computational sources, for original references please see Refs. [40–44]). The data clearly demonstrates the importance of scale-dependent properties in biological protein materials

M

hierarchical structure [9]. The hierarchical structure of bone is outlined in Fig. 7 as an example. Bone must likewise survive cyclic compressive loading, bending, and impulses for the lifetime of large and active animals. Bone is suitable for a wide range of loading rates: the failure stress of bulk bone increases by as much as 44 % when the stress rate is increased by a factor of 10^3 [2]. In addition, since bone is continuously rebuilt, bones can adapt to creep and changes in local loading after injuries. Large bones like the femur have a structure similar to bamboo: strong and dense cortical bone tissue forms an outer sheath around a porous network of cancellous (i.e., trabecular or spongy) bone tissue with a core of marrow called the medullary cavity, and many reinforcing, longitudinal cortical tubes called osteons. Both cortical and cancellous bone tissue are differently organized structures of the same biomineral composite created by osteoblast cells, which secrete an amorphous protein matrix of predominantly Type I collagen. Over time, mineralization and ion deposition introduce reinforcing globular and plate-like crystals of carbonated hydroxyapatite [2].

There exist two cases of this matrix: woven bone, with a random arrangement of collagen fibers, is formed quickly in fetal bones and after fractures. Over time, woven bone is slowly converted to lamellar bone, made of ordered, concentric, parallel sheets of collagen fibers. Cancellous bone exhibits a wide range of local architecture and density, with volume fractions ranging from 0.05 to 0.6. Several factors affect mechanical property studies of cancellous bone, including donor age, sample hydration, anatomic



Mechanical Properties of Hierarchical Protein Materials, Fig. 7 The many hierarchical levels of bone allow this material to reach high levels of strength, toughness, as well as the ability to self-repair. Bone has at least seven hierarchical levels, each with a structure that contributes toward enhancing the mechanical property advantages available at that length scale [38]

location, and diseases such as osteoporosis, osteoarthritis, bone cancer, and *osteogenesis imperfecta*, but typical elastic moduli of cancellous bone are between 0.01 and 2 GPa and ultimate strengths between 0.1 and 30 MPa [39]. The rods and plates that form the individual trabeculae of cancellous bone have a uniform composition similar to cortical bone, but are slightly less mineralized, allowing plasticity during network deformation. Although the anisotropic mechanical properties of cancellous bone depend on the density of the network, the porous structure allows yielding at 1–2 % strain and can sustain large deformations up to nearly 50 % strain while maintaining its load-carrying capacity [2], allowing considerable impact energy dissipation before fracture.

An analysis of the molecular mechanisms of protein and mineral crystal phases under large deformation of the mineralized collagen fibrils reveal a fibrillar toughening mechanism that leads to a dramatic increase in energy dissipation compared to fibrils without a mineral phase. This fibrillar toughening mechanism increases the resistance to fracture by forming large local yield regions around crack-like defects, a mechanism that protects the integrity of the entire structure by allowing for localized failure that can later be healed. As a consequence, mineralized collagen fibrils are able to tolerate micro-cracks on the order of several hundred micrometers without causing any macroscopic failure of the tissue. At the microscale, mineralized collagen fibrils have a Young's modulus of 6.23 GPa, yield at a tensile strain of 6.7 %, and feature a fracture stress of 0.6 GPa. This shows a dramatic increase over unmineralized collagen fibrils, which only show a Young's modulus of 4.6 GPa and fracture stress of 0.6 GPa. Other deformation and fracture mechanisms exist in bone at larger hierarchical levels, including the breaking of sacrificial bonds between discrete mineralized fibrils, the formation of microcracks ahead of the crack tip, crack bridging by collagen fibrils arranged orthogonal to the crack, and the formation of large plastic zones around crack tips that contribute to an increase in bone's overall toughness [38].

Conclusion

Nature has much to offer for the next generation of both macroscale and nanoscale materials, in particular relating to the linking of vast and seemingly separated scales (from nano to macro). The expansion of the structural design space through

the concept of merging structure and material (Fig. 1) is a powerful concept that enables us to create highly functional materials through the use of relatively simple building block, a mechanism established in the universality-diversity paradigm of materials science [11, 13]. The transfer of concepts derived from biological materials, for example, mimicking of the strain-hardening behavior of spider silk, can give rise to the design of advanced functional materials with exceptional kinetic energy buffering and absorption. Developing ordered nanocomposites that are based on the structures found in bone, antler, or nacre can dramatically increase the toughness of typically brittle structures and may enable us to use abundant materials such as silica in the design of tough, elastic, and strong materials. Even the architecture of collagen's reinforced helices, fibrils, and fibers offer much inspiration in creating advanced tensile struts from seemingly inferior materials. The key issue is to better understand the mechanisms of the origin of the mechanical properties of proteins from the bottom-up, and the identification of systematic approaches to transfer the insight from biology to engineering design through an approach referred to as materiomics [11].

Cross-References

- ▶ Self-assembly
- ▶ Spider Silk
- ▶ Structure and Stability of Protein Materials

References

1. Meyers, M.A., Chen, P.-Y., Lin, A.Y.-M., Seki, Y.: Biological materials: structure and mechanical properties. Prog. Mater. Sci. **53**(1), 1–206 (2008)
2. Black, J., Hastings, G.: Handbook of Biomaterial Properties. Springer, New York (2006)
3. Wegst, U.G.K., Ashby, M.F.: The mechanical efficiency of natural materials. Philos. Mag. **84**(21), 2167–2186 (2004)
4. Ackbarow, T., Sen, D., Thaulow, C., Buehler, M.J.: Alpha-helical protein networks are self-protective and flaw-tolerant. PLoS One **4**(6), e6015 (2009)
5. Kai-Nan, A., Yu-Long, S., Zong-Ping, L.: Flexibility of type I collagen and mechanical property of connective tissue. Biorheology **41**(3), 239–246 (2004)
6. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., Walter, P.: Molecular Biology of the Cell. Garland, New York (2002)
7. Gao, M., Sotomayor, M., Villa, E., Lee, E.H., Schulten, K.: Molecular mechanisms of cellular mechanics. Phys. Chem. Chem. Phys. **8**(32), 3692–3706 (2006)
8. Gautieri, A., Uzel, S., Vesentini, S., Redaelli, A., Buehler, M.J.: Molecular and mesoscale mechanisms of osteogenesis imperfecta disease in collagen fibrils. Biophys. J. **97**(3), 857–865 (2009)
9. Ashby, M.F., Gibson, L.J., Wegst, U., Olive, R.: The mechanical properties of natural materials. I. Material property charts. Proc. R. Soc. Lond. A Math. Phys. Sci. **450**(1938), 123–140 (1995). doi:10.1098/rspa.1995.0075
10. Gibson, L.J., Ashby, M.F.: The mechanics of three-dimensional cellular materials. Proc. R. Soc. Lond. A Math. Phys. Sci. **382**(1782), 43–59 (1982). doi:10.1098/rspa.1982.0088
11. Buehler, M.J.: Tu(r)ning weakness to strength. Nano Today **5**(5), 379–383 (2010). doi:10.1016/j.nantod.2010.08.001
12. Fratzl, P., Weinkamer, R.: Nature's hierarchical materials. Prog. Mater. Sci. **52**(8), 1263–1334 (2007)
13. Ackbarow, T., Buehler, M.J.: Hierarchical coexistence of universality and diversity controls robustness and multi-functionality in protein materials. J. Comput. Theor. Nanosci. **5**(7), 1193–1204 (2008)
14. Isralewitz, B., Gao, M., Schulten, K.: Steered molecular dynamics and mechanical functions of proteins. Curr. Opin. Struct. Biol. **11**(2), 224–230 (2001)
15. Sulkowska, J., Cieplak, M.: Mechanical stretching of proteins – a theoretical survey of the protein data bank. J. Phys. Condens. Matter **19**(28), 283201 (2007)
16. Sotomayor, M., Schulten, K.: Single-molecule experiments in vitro and in silico. Science **316**(5828), 1144–1148 (2007). doi:10.1126/science.1137591
17. Ackbarow, T., Chen, X., Keten, S., Buehler, M.J.: Hierarchies, multiple energy barriers, and robustness govern the fracture mechanics of α-helical and β-sheet protein domains. Proc. Natl. Acad. Sci. U. S. A. **104**(42), 16410–16415 (2007). doi:10.1073/pnas.0705759104
18. Keten, S., Buehler, M.J.: Asymptotic strength limit of hydrogen-bond assemblies in proteins at vanishing pulling rates. Phys. Rev. Lett. **100**(19), 198301 (2008)
19. Fisher, T.E., Oberhauser, A.F., Carrion-Vazquez, M., Marszalek, P.E., Fernandez, J.M.: The study of protein mechanics with the atomic force microscope. Trends Biochem. Sci. **24**(10), 379–384 (1999)
20. Keten, S., Buehler, M.J.: Geometric confinement governs the rupture strength of H-bond assemblies at a critical length scale. Nano Lett. **8**(2), 743–748 (2008). doi:10.1021/nl0731670
21. Penel, S., Morrison, R.G., Dobson, P.D., Mortishire-Smith, R.J., Doig, A.J.: Length preferences and periodicity in β-strands. Antiparallel edge β-sheets are more likely to finish in non-hydrogen bonded rings.

Protein Eng. **16**(12), 957–961 (2003). doi:10.1093/protein/gzg147
22. Hayashi, C.Y., Lewis, R.V.: Evidence from flagelliform silk cDNA for the structural basis of elasticity and modular nature of spider silks. J. Mol. Biol. **275**(5), 773–784 (1998)
23. Hu, X., Vasanthavada, K., Kohler, K., McNary, S., Moore, A., Vierra, C.: Molecular mechanisms of spider silk. Cell. Mol. Life Sci. **63**(17), 1986–1999 (2006). doi:10.1007/s00018-006-6090-y
24. Denny, M.: The physical properties of spider's silk and their role in the design of orb-webs. J. Exp. Biol. **65**(2), 483–506 (1976)
25. Huemmerich, D., Scheibel, T., Vollrath, F., Cohen, S., Gat, U., Ittah, S.: Novel assembly properties of recombinant spider dragline silk proteins. Curr. Biol. **14**(22), 2070–2074 (2004)
26. Foo, C., Po, W., Bini, E., Huang, J., Lee, S.Y., Kaplan, D.L.: Solution behavior of synthetic silk peptides and modified recombinant silk proteins. Appl. Phys. A **82**, 193–203 (2006)
27. Gosline, J., Guerette, P., Ortlepp, C., Savage, K.: The mechanical design of spider silks: from fibroin sequence to mechanical function. J. Exp. Biol. **202**(23), 3295–3303 (1999)
28. Nova, A., Keten, S., Pugno, N.M., Redaelli, A., Buehler, M.J.: Molecular and nanostructural mechanisms of deformation. Strength and toughness of spider silk fibrils. Nano Lett. **10**(7), 2626–2634 (2010). doi:10.1021/nl101341w
29. Grubb, D.T., Jelinski, L.W.: Fiber morphology of spider silk: the effects of tensile deformation. Macromolecules **30**(10), 2860–2867 (1997). doi:10.1021/ma961293c
30. Gosline, J., Lillie, M., Carrington, E., Guerette, P., Ortlepp, C., Savage, K.: Elastic proteins: biological roles and mechanical properties. Philos. Trans. R. Soc. Lond. B Biol. Sci. **357**(1418), 121–132 (2002). doi:10.1098/rstb.2001.1022
31. Du, N., Yang, Z., Liu, X.Y., Li, Y., Xu, H.Y.: Structural origin of the strain-hardening of spider silk. Adv. Funct. Mater. **21**(4), 772–778 (2010). doi:10.1002/adfm.201001397
32. Bratzel, G.H., Buehler, M.: Sequence-structure correlations and size effects in silk nanostructure: Poly-Ala repeat of *N. clavipes* MaSp1 is naturally optimized. JMBBM (2011). doi:10.1016/j.jmbbm.2011.07.012
33. Gatesy, J., Hayashi, C., Motriuk, D., Woods, J., Lewis, R.: Extreme diversity, conservation, and convergence of spider silk fibroin sequences. Science **291**(5513), 2603–2605 (2001). doi:10.1126/science.1057561
34. Keten, S., Buehler, M.J.: Nanostructure and molecular mechanics of spider dragline silk protein assemblies. J. R. Soc. Interface **7**(53), 1709–1721 (2010). doi:10.1098/rsif.2010.0149
35. Fratzl, P.: Collagen: Structure and Mechanics. Springer, New York (2008)
36. Gautieri, A., Vesentini, S., Redaelli, A., Buehler, M.J.: Hierarchical structure and nanomechanics of collagen microfibrils from the atomistic scale up. Nano Lett. **11**(2), 757–766 (2011). doi:10.1021/nl103943u
37. Jackson, A.P., Vincent, J.F.V., Turner, R.M.: The mechanical design of nacre. Proc. R. Soc. Lond. Series B. Biol. Sci. **234**(1277), 415–440 (1988). doi:10.1098/rspb.1988.0056
38. Espinosa, H., Rim, J., Barthelat, F., Buehler, M.: Merger of structure and material in nacre and bone – perspectives on de novo biomimetic materials. Prog. Mater. Sci. **54**, 1059–1100 (2009)
39. Currey, J.: Mechanical properties of vertebrate hard tissues. Proc. Inst. Mech. Eng. Part H: J. Eng. Med. **212**(6), 399–411 (1998)
40. Sasaki, N., Odajima, S.: Elongation mechanism of collagen fibrils and force-strain relations of tendon at each level of structural hierarchy. J. Biomech. **29**(9), 1131–1136 (1996)
41. Eppell, S.J., Smith, B.N., Kahn, H., Ballarini, R.: Nano measurements with micro-devices: mechanical properties of hydrated collagen fibrils. J. R. Soc. Interface **3**(6), 117–121 (2006). doi:10.1098/rsif.2005.0100
42. Espinosa, H.D., Zhu, Y., Moldovan, N.: Design and operation of a MEMS-based nanoscale material testing system. J. Microelectromech. Syst. **16**, 1219–1231 (2007)
43. van der Rijt, J.A.J., van der Werf, K.O., Bennink, M. L., Dijkstra, P.J., Feijen, J.: Micromechanical testing of individual collagen fibrils. Macromol. Biosci. **6**(9), 697–702 (2006). doi:10.1002/mabi.200600063
44. Aladin, D.M.K., Cheung, K.M.C., Ngan, A.H.W., Chan, D., Leung, V.Y.L., Lim, C.T., Luk, K.D.K., Lu, W.W.: Nanostructure of collagen fibrils in human nucleus pulposus and its correlation with macroscale tissue mechanics. J. Orthop. Res. **28**(4), 497–502 (2009). doi:10.1002/jor.21010

Mechanical Properties of Nanocrystalline Metals

Farghalli A. Mohamed
Department of Chemical Engineering and Materials Science, University of California, Irvine the Henry Samueli School of Engineering, Irvine, CA, USA

Synonyms

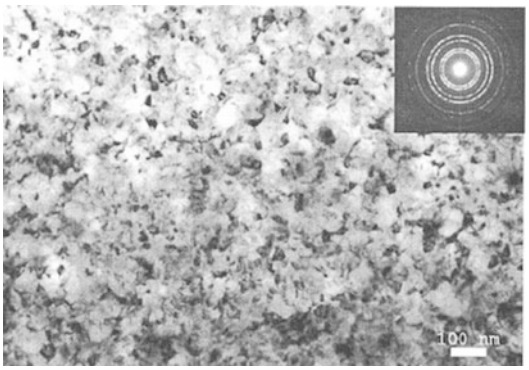
Mechanical behavior of nanostructured metals

Definition

Materials are classified into four groups in terms of the value of grain size: large-grained materials, micrograined materials, ultrafine-grained (UFG) materials, and nanocrystalline (nc) materials. Large-grained materials are characterized by grain sizes larger than 20 μm, micrograined materials by grain sizes in the range 1–10 μm, ultrafine-grained materials by grain sizes in the range 300–900 nm, and nc-materials by grain sizes less than 200 nm. The grain size ranges associated with the above classifications are approximate and are based on consideration of observations reported in literature.

Figure 1 provides a transmission electron microscopy (TEM) micrograph for the grain size in nc-nickel. The grain size ranges from 3 to 60 nm with an average value of about 24 nm.

Because of the small grain sizes characterizing nc-materials, features such as grain boundaries, junction lines, and nodes have significant volume fractions and influence properties far more strongly than in more conventional microstructures [1, 2]. Nc-materials have novel physical, chemical, biological, and mechanical properties that may have important engineering uses. As a consequence, they are attracting wide attention in materials research.



Mechanical Properties of Nanocrystalline Metals, Fig. 1 Bright field transmission electron microscopy micrograph showing the nanograin distribution in nc-nickel. The ring configuration at the *upper right corner* represents the diffraction pattern for the nc-grain structure

While the chemical and electromagnetic properties of nc-materials have received most of the attention, their mechanical properties are also important. It is both necessary and useful that these properties be thoroughly understood. Understanding mechanical properties is necessary since materials in almost all devices experience mechanical loads that must not only be supported but also be predicted to produce reliable designs. If the mechanical properties are unknown, even promising chemical or electromagnetic applications may be very difficult to achieve. Understanding the mechanical properties of nc-materials would be useful because of the following considerations:

- It is possible to make nc-materials with exceptionally high strengths for use as hard coatings or high-strength structural elements.
- The potential of using nanomaterials as machine components; if the nc-materials are used as cutting tools, localized shear should be avoided to increase the tool life.

Over the past two decades, considerable progress has been made in terms of providing a better understanding of mechanical behavior in nc-materials. In particular, considerable progress has been made in areas related to the following four aspects of the mechanical behavior of nc-materials: (a) creep behavior, (b) yielding and hardness, (c) ductility, and (d) superplasticity. Before presenting key research findings in these areas, it is useful to provide information on some of the parameters that are important in terms of characterizing mechanical behavior in materials.

Mechanical Properties Basic Parameters

Activation Volume

The activation volume, v, is the rate of decrease of the activation enthalpy with respect to flow stress at constant temperature. It provides useful insight into the nature of deformation mechanisms because it is characterized by a definite value of stress dependence for each atomistic process. Measurements of the activation volume during

deformation are made by (a) making either a small change in applied stress, and measuring the corresponding strain rate after the change, or a small change in strain rate and measuring the corresponding stress, and (b) applying the following equation:

$$v = kT\frac{\partial \ln \dot{\gamma}}{\partial \tau} \tag{1}$$

where k is Boltzmann's constant and T is the absolute temperature, τ is the shear stress, and $\dot{\gamma}$ is the shear strain rate.

Apparent Activation Energy

When the plastic deformation of a material is controlled by a thermally activated process, the strain rate for the i-the process may be expressed in terms of rate reaction theory as:

$$\dot{\gamma}_i = A_i(\tau, T)e\left[-\frac{\Delta \mathrm{H}_i(\tau, T)}{RT}\right] \tag{2}$$

where A_i is a function of stress, τ, temperature, T, and substructure; R is the gas constant; and $\Delta \mathrm{H}_i$ is the activation enthalpy. $\Delta \mathrm{H}_i$ is an important parameter in the deformation equation, and it must be determined in any attempt to identify the rate-controlling process. According to Eq. 2, $\Delta \mathrm{H}_i$ is uniquely defined by the following expression:

$$\Delta \mathrm{H}_i = R\left[\frac{\partial \ln \gamma i}{\partial\left(\frac{-1}{T}\right)}\right]_{\tau, d} \tag{3}$$

where $\boldsymbol{d}$ is the grain size. Equation 3 suggests a direct method for estimating $\Delta \mathrm{H}_i$. In deformation experiments, $\Delta \mathrm{H}_i$ is approximated by the apparent activation, $\boldsymbol{Q_a}$:

$$Q_a \approx \Delta \mathrm{H}_i = -R\left[\frac{\partial \ln \gamma}{\partial\left(\frac{1}{T}\right)}\right]_{\tau, d} \tag{4}$$

Apparent Dependence of Strain Rate on Applied Stress

The apparent dependence of strain rate on applied stress is an important parameter that receives attention in studies on the mechanical behavior of materials. In investigations focusing on high temperature deformation behavior, such dependence is represented by the stress exponent, n, which is defined as:

$$\mathrm{n} = \left.\frac{\partial \ln \dot{\gamma}}{\partial \ln \tau}\right| T, d \tag{5}$$

There is also another parameter that has been extensively used to express the relation between strain rate and stress, and is termed strain rate sensitivity, m. This parameter is given by:

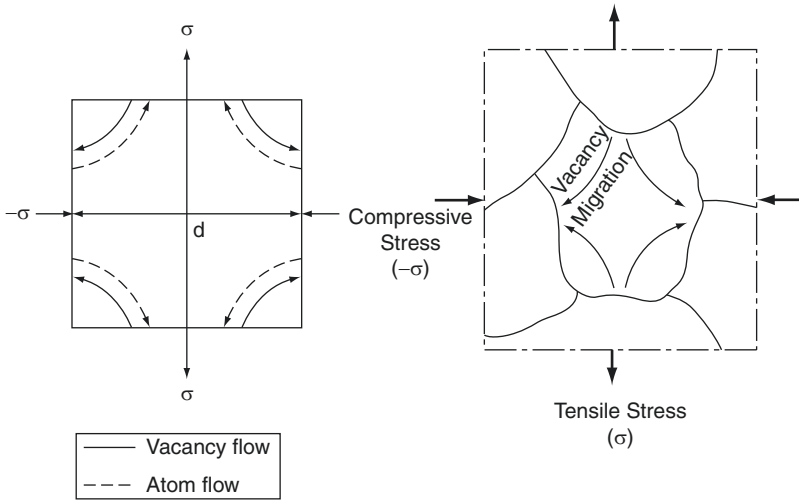
$$\mathrm{m} = \left.\frac{\partial \ln \tau}{\partial \ln \dot{\gamma}}\right| T, d \tag{6}$$

Under steady-state high deformation conditions, $n = 1/m$. In addition to mentioning briefly the aforementioned four areas, it is useful to touch upon crystal imperfections that play key roles during plastic deformation.

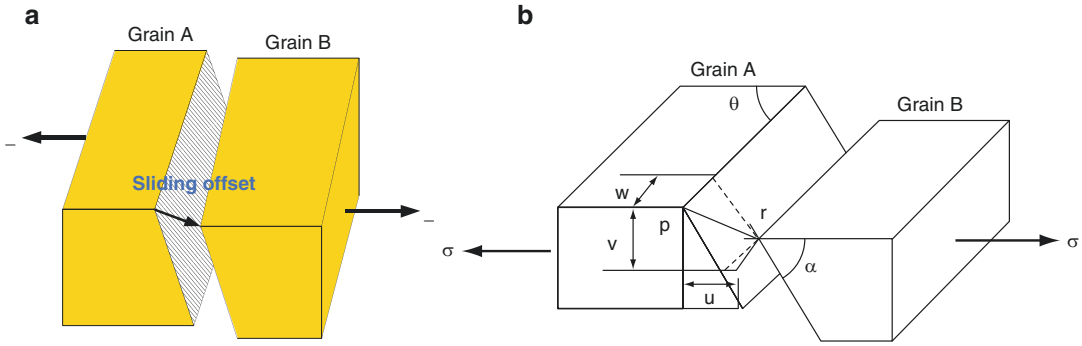
Crystal Imperfections

Metallic or ceramic crystals are not perfect. They contain imperfections [3] that include: point imperfections, line imperfections, and surface imperfections. Point imperfections include vacancies, substitutional atoms (e.g., magnesium in aluminum), and interstitial atoms (e.g., carbon in iron). Line imperfections refer to dislocations. Surface imperfections include grain boundaries.

While substitutional and interstitial atoms are responsible in many ways for strengthening in materials, the creation and diffusion of vacancies provide a mode of deformation at very low stress and small grain sizes [4]. Figure 2 can illustrate this mode of deformation, known as diffusional creep [4]. According to this figure, three steps are associated with diffusional creep: (a) the creation of a vacancy in the grain volume near the grain boundary under tensile stress, (b) the migration of the vacancy to the boundary that is under compressive stress, and (c) the annihilation of the vacancy at that boundary. The flow of vacancies is matched by an equal and opposite flow of atoms resulting in a longitudinal extension and lateral



Mechanical Properties of Nanocrystalline Metals, Fig. 2 Stress-directed diffusion of vacancies in a polycrystalline material



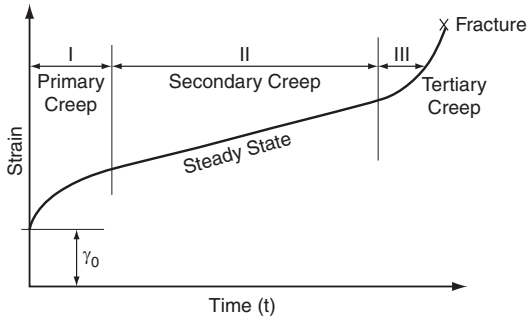
Mechanical Properties of Nanocrystalline Metals, Fig. 3 (**a**) Sliding offset as a result of boundary sliding. (**b**) Sliding displacement associated with boundary sliding

contraction of the grain with time under stress. If the migration of vacancies occurs in the interior of the grain, the deformation process is referred to as Nabarro-Herring creep. If, on the other hand, the migration of vacancies takes place in the grain boundary, the process is known as Coble creep.

Dislocations play a key role in plastic deformation. Their presence and movement in crystals account for the fact that the experimental yield strengths of materials are much lower than their theoretical shear strengths [3]. Their multiplication and interactions produce strain hardening while their annihilation results in crystal recovery. By using the details of dislocation theory, several investigators were able to develop a number of deformation mechanisms that have provided plausible interpretations for mechanical behavior under various conditions.

Grain boundaries are regions of misorientation that separate grains. At low homologous temperatures ($T/T_m < 0.2$), grain boundaries act as a source for strengthening by serving as obstacles to the propagation of deformation from one grain to the next [3]. At high homologous temperatures ($T/T_m > 0.4$), grain boundaries become mobile especially at low stresses, producing strain via grain boundary sliding. Figure 3 illustrates the process of boundary sliding. When enough stress is applied to a polycrystalline material, the grains slide past each other and the metal changes shape or deforms, a phenomenon known as "boundary sliding." As a result of boundary sliding, steps

M



Mechanical Properties of Nanocrystalline Metals, Fig. 4 An example for a normal creep curve

(offsets) are produced on the surface of a flat specimen. Detection of these steps provides evidence for the occurrence of boundary sliding, a process that may lead to the formation of voids and cavities in the material. The formation of voids in turn results in premature failure of the material. However, in micrograined materials, the strain contributed by boundary sliding may become considerable, leading to the occurrence of micrograin superplasticity [5].

Key Research Findings

Creep Behavior

A polycrystalline material subjected to an applied stress or load deforms continuously. This progressive accumulation of strain with time is known as creep [4]. Creep takes place as a result of thermal activation of deformation processes. The curve in Fig. 4 illustrates the idealized shape of a creep curve. The slope of this curve, $d\gamma/dt$, is referred to as the creep rate (or the strain rate). Following an initial rapid elongation of the specimen, the creep rate decreases with time (primary creep, or primary stage), then reaches a steady state (secondary creep, steady-state stage), and finally the creep rate increases with time until fracture occurs (tertiary creep or tertiary stage). Major differences among creep curves of various materials, however, are noted over the primary creep stage.

Nc-materials offer interesting possibilities related to many structural applications. In order to explore some of these possibilities, an understanding of the deformation of these materials under creep conditions is essential. In order to develop this understanding, *guiding information* is needed. This guiding information can be obtained from comprehensive experimental creep measurements.

During the period of 1990–2005, experimental measurements dealing with the creep behavior of nc-materials were reported. However, these measurements not only were limited in scope but also were obtained under questionable conditions. For example, the total strain measured in creep experiments conducted on nc-materials in several investigations did not exceed in most cases 0.01. With such small strains, it is difficult to ascertain whether the steady-state creep is reached. The problem of the lack of experimental measurements that would characterize creep and deformation-induced grain size growth in nc-materials was recently addressed in a detailed investigation [6], in which the creep behavior of nc-Ni was studied *for the first time* under the conditions of large strains and constant stresses. In the investigation, creep testing was performed over more than five orders of magnitude of strain rate ($10^{-9}\,s^{-1}$ to $2 \times 10^{-4}\,s^{-1}$) at 393 K (0.23 T_m, where T_m is the melting point). In addition, information regarding the stress and the temperature dependence of strain rate, the activation volume, and deformation-induced grain growth was obtained. The results of the above investigation have shown that the behavior of nc-Ni exhibited a variable stress exponent, an activation energy close to that expected for boundary diffusion, and an activation volume of about 19.5 b^3; and that upon loading nc-Ni specimens at 393 K, there is a grain growth, and this growth is driven by a combination of stress and time. However, once steady-state creep is reached, significant grain growth ceases.

Yield Strength and Hardness

One of the aspects of the mechanical behavior of nc-materials that has been extensively studied is the influence of grain size on flow shear stress, $\tau\ (= \sigma/2)$. In general, as the grain size, d, is refined, the strength increases as $1/\sqrt{d}$ according

to the Hall–Petch relation [3] that can be represented by the following expression:

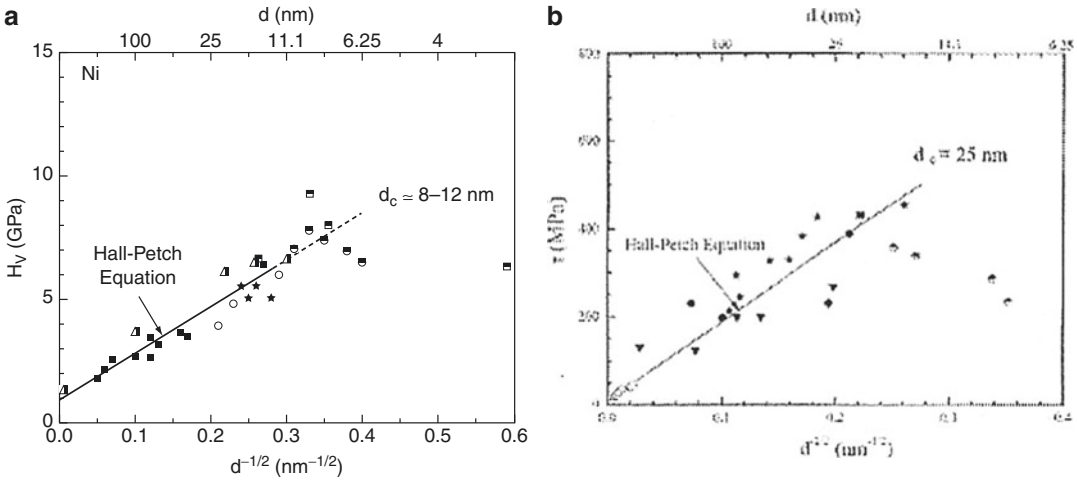
$$\tau = \tau_0 + c/\sqrt{d} \tag{7}$$

where τ_0 is a friction stress and c is a constant. Equation 7 is also valid when the yield strength is replaced by the hardness, H ($H = 6\,\tau$). The mechanistic source of the Hall–Petch relation is not entirely clear, and more than one mechanism may pertain [3]. The effect of Eq. 7 in the nanoscale regime is, however, striking; if Eq. 7 holds, decreasing d from 20 to 20 nm increases the strength by a factor of 30. In general, Hall–Petch behavior for both hardness and yield strength holds to grain sizes well within the nanoscale range. However, experimental observations have suggested that when the grain size of some nc-materials falls below a critical value in the nanoscale range, the strength may decrease, that is, nanoscale softening (an inverse Hall–Petch behavior) occurs. The observations related to nanoscale softening were initially controversial since some of them could be attributed to the presence of flaws in the material [7]. For example, the loss in strength in the nanoscale range can be the result of pores in the material whose presence leads to a decrease in both the shear stress required for deformation and the shear modulus.

Despite the aforementioned controversy and others related to contaminations and grain size measurements, three pieces of information lend support to the occurrence of softening in the nanoscale range. First, it does seem reasonable that Hall–Petch behavior would be lost at the smallest grain sizes in the nanoscale region partly because it is no longer possible to maintain dislocation structures within grains, and partly because grain boundary activity assumes significance at very small grain sizes. Second, recent experimental results reported for the variation of hardness against grain size in nickel, nickel-tungsten alloys, and copper (see Fig. 5) have provided clear evidence for the occurrence of nanoscale softening [8, 9]. Finally, the results of recent molecular dynamics simulations have shown [10] the presence of a maximum in the flow stress of copper when its grain size becomes smaller than 15 nm.

Deformation Processes

An understanding of the origin and nature of deformation processes in nc-materials is essential.



Mechanical Properties of Nanocrystalline Metals, Fig. 5 (**a**) Plot of hardness versus the inverse square root of grain size for nc-nickel. (**b**) Data on copper plotted as yield strength as a function of the inverse square root of the grain size for Cu

Such understanding is important in two ways. First, when the origin of the deformation process is uncovered in sufficient detail, it should be possible to predict the mechanical behavior under a variety of conditions (stress, temperature, and grain size). Second, when the basic deformation process is known (successfully identified or developed), it is possible to introduce microstructural features that can improve the mechanical behavior of the materials in terms of ductility, toughness, etc.

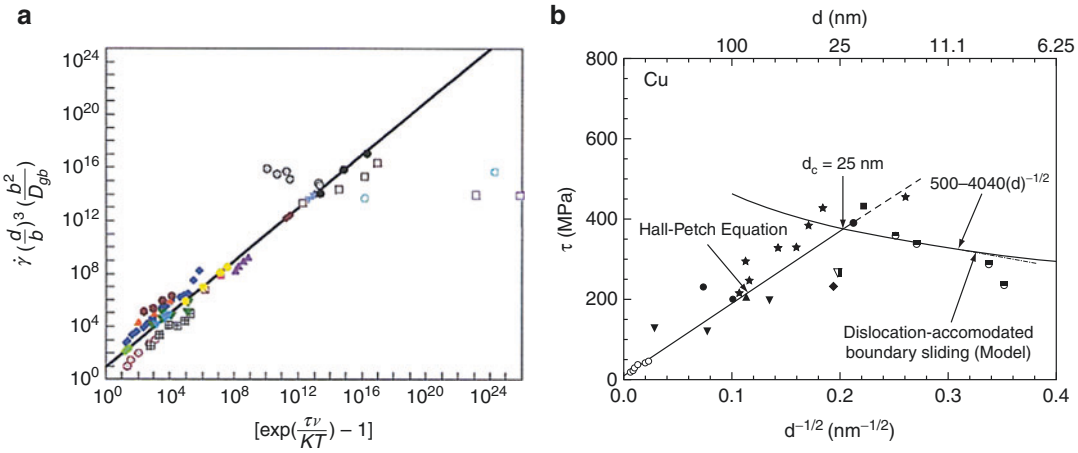
As a result of recent investigations on nc-materials, several deformation mechanisms have been proposed. These mechanisms are different in terms of concept and details. The proposed mechanisms include: *(a) Coble creep, (b) Triple Junction Model, (c) Atomic Scale Boundary Sliding Model, (d) Modified Coble creep (inclusion of a threshold stress), (e) Grain Boundary Sliding with Back Stress, (f) Boundary Sliding under the Condition of Nonlinear-Viscous Behavior*, (*g*) *Thermally Activated Grain Boundary Shearing Process*, (*h*) *Thermally Activated Grain Boundary Shearing Process*, (*i*) *Composite model involving amorphous boundary layer, (j) Model of strongest size*, *and (k) Dislocation Accommodated Boundary Sliding*. These mechanisms were reviewed in Ref. [11].

As a result of investigations on nc-materials, extensive experimental data on the mechanical behavior of nc-materials such as nc-Ni and nc-Cu have become available [11]. The availability of the experimental data on these two metals provides an opportunity to closely examine the validity of deformation mechanisms that were proposed to account for the characteristics of deformation in nc-materials.

An analysis of the data reported led to the identification of the requirements [10] that a viable deformation mechanism should meet in terms of accounting for the mechanical characteristics, which were revealed by the data. These requirements are: (a) an activation volume whose value is in the range 10–40 b^3, (b) an activation energy that is close to the activation energy for boundary diffusion but that decreases with increasing applied stress, (c) the magnitudes of deformation rates that cover wide ranges of temperatures, stresses, and grain sizes, (d) inverse Hall–Petch behavior (softening), and (e) limited ductility.

The validity of aforementioned deformation mechanisms for nc-materials has been closely examined in the light of these requirements. The results of the examination have led to the following findings [11]:

(a) Coble creep and thermally activated Grain Boundary Shearing process are ruled out as viable mechanisms not only because they predict an activation volume smaller than those experimentally reported but also because they yield unrealistic critical grain sizes for a transition from conventional Hall–Petch behavior (hardening with decreasing grain size) to inverse Hall–Petch behavior (softening with decreasing grain size).
(b) Mechanisms such as modified Coble creep and thermally activated grain boundary shearing process are not applicable since they predict a very small value of the activation volume; they predict an activation energy of *b3* energy compared to experimentally reported values of 20–40 *b3*.
(c) While the composite model involving a grain boundary layer yields critical grain sizes for a transition from conventional Hall–Petch behavior (hardening with decreasing grain size) to inverse Hall–Petch behavior (softening with decreasing grain size) that agree well with those inferred from experimental results, it fails to account for the experimental data for nc-Cu and nc-Ni.
(d) The predictions of the model of dislocation accommodated boundary sliding [11] are consistent with several deformation characteristics reported for nc-materials including the range of the activation volume, (b) the value of activation energy and its dependence on stress, (c) the trends and magnitudes of experimental data on nc-Cu and nc-Ni, and (d) the critical grain sizes for transitions from conventional Hall–Petch behavior to inverse Hall–Petch behavior in nc-Cu and nc-Ni. Figure 6a, b illustrate characteristics (c) and (d), respectively.



Mechanical Properties of Nanocrystalline Metals, Fig. 6 (**a**) Plot of $\dot{\gamma}(d/b)^3\left(b^2/D_{gb}\right)$ against $(\tau v/KT) - 1$ on a logarithmic scale showing the correlation between deformation data reported for nc-Ni and nc-Cu and the prediction of the Model of Dislocation Accommodated Boundary Sliding (*solid line* in the plot) [6]. (**b**) Yield strength as a function of the inverse square root of the grain size for copper

Ductility

Ductility is an important mechanical property that denotes the total elongation experienced by a tensile sample prior to fracture, *ef*. It is defined by the following equation [12]:

$$e_f = L - L_o/L_o \tag{8}$$

where L is the final length and L_o is the initial length.

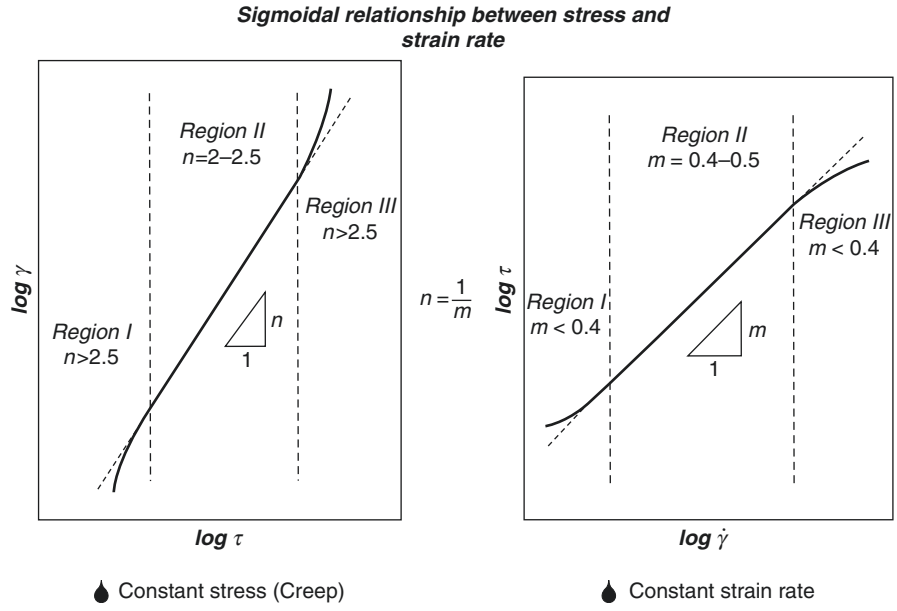
Although nc-materials always possess higher strength than their coarse-grain counterparts, they usually exhibit low ductility. For engineering applications, low ductility is undesirable for two primary reasons. First, low ductility does not permit metalworking processes to effectively produce required configurations. Second, low ductility impairs the material's ability to absorb overload, a condition that reduces the margin of safety in design. The source of low ductility nc-materials is not entirely clear, but it probably arises from either extrinsic flaws such as porosity or loss of work hardening due to the lack of dislocation interactions and accumulations [7].

Recently, the concept of improving ductility via enhancing dislocation interactions and accumulations has received considerable attention. This attention has led to several approaches that include the introduction of a bimodal grain size distribution, deformation twins, or dispersion of nanoparticles.

The role of both bimodal grain distribution and twins in achieving high ductility can be attributed to two primary factors. First, the presence of large grains enhances strain hardening, which prevents excessively large local strains (plastic deformation becomes localized), and suppresses the propagation of cracks nucleated in the surrounding regions of the very small grains. Second, twins in the interiors of the large grains in the bimodal grain distribution can block moving dislocations, a process that results in dislocation interaction and accumulation. Dislocation interactions and accumulation in turn promote strain hardening.

Superplasticity

Before discussing superplasticity in nc-materials, it is appropriate to provide a brief review for superplastic behavior in micrograin materials. Superplasticity refers to the ability of micrograined materials ($d < 10$ μm, where d is the grain size) particular materials to exhibit extensive neck-free elongations (>300 %) during deformation at elevated temperatures ($T > 0.5\ T_m$, where T_m is the melting point).



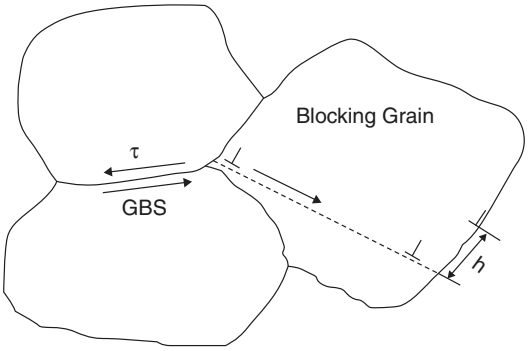
Mechanical Properties of Nanocrystalline Metals, Fig. 7 Relationship between shear stress and shear train rate for superplastic materials

The behavior of micrograin superplastic alloys has been extensively studied. As a result of these investigations, three findings are well documented [5]. First, micrograin superplasticity is a diffusion-controlled process. Second, the relationship between stress, τ, and strain rate, $\dot{\gamma}$, is often sigmoidal (see Fig. 7). Under creep testing conditions, this sigmoidal relationship is manifested by the presence of three regions: region I (the low-stress region), region II (the intermediate-stress region), and region III (the high-stress region). In region II, the stress exponent, n, is about 2 and the activation energy for creep is equal to that for grain boundary diffusion, Q_{gb}. In regions I and III, both the stress exponent, n, and the activation energy for creep, Q_c, increase ($Q_c > Q_{gb}$). Finally, in region II, maximum ductility (>300 %) occurs. Because of this characteristic, region II is often referred to as the superplastic region.

From a practical point of view, it is very desirable to extend the superplastic region to high strain rates. As reported elsewhere [5], the superplastic region is in general observed at strain rates in the range of 10^{-5}–10^{-2} s^{-1}. These rates are too slow to be economically suitable for a variety of industrial applications. Nc-materials appear to offer an interesting possibility to address this issue. According to information on micrograin superplasticity, as the grain size decreases from micrometer to nanometer, it is expected that the superplastic region would be transposed to high strain rates or observed at low temperatures such as room temperature, that is, high strain rate and/or low temperature superplasticity is possible. Despite this attractive possibility and its implications in terms of the potential for industrial adaptation of superplastic forming, attempts made so far to explore the occurrence of superplasticity have been unsuccessful.

One possible reason for the difficulty of observing superplastic flow in nc-materials is the occurrence of grain growth: it is not possible to obtain low-temperature superplasticity in pure metals in which grain sizes are less than 100 nm, since the advantage of observing superplastic flow at low temperatures is neutralized by the onset of significant grain growth in this temperature range.

A second reason can be inferred from the rate controlling mechanism for deformation in nc-materials. Recently, a new model for



Mechanical Properties of Nanocrystalline Metals, Fig. 8 Schematic diagram for the model of dislocation accommodated boundary sliding, showing that as a result of boundary sliding, dislocations are generated at a triple junction and then traverse the grain to the opposite grain boundary where they climb and are annihilated

deformation in nanocrystalline (nc) materials has been developed [6]. The development of the model was based on the concept that plasticity in nc-materials is the result of grain boundary sliding accommodated by the generation and motion of dislocations under local stresses, which are higher than applied stresses due to the development of stress concentrations. Specifically, it has been assumed that as a result of sliding of a group of grains, the shear stress becomes concentrated at any grain, triple point, or protrusion that obstructs motion of this group; that this high local stress can then generate dislocations in the blocking grain (or initiate voids); and that the generated dislocations move one by one to the opposite boundary where they climb to their annihilation sites (no dislocation-ups). For illustration, see Fig. 8. By postulating that the creep rate, $\dot{\gamma}$, is governed by the time for the climb of a dislocation along the boundary until annihilation occurs, the following rate-controlling equation was derived [6]:

$$\dot{\gamma} = 9\left(\frac{b}{d}\right)^3\left(\frac{D_{gbo}}{b^2}\right)\exp\left(\frac{-Q_{gb}}{RT}\right) \times \left[\exp\left(\frac{\tau v}{kT}\right) - 1\right] \quad (9)$$

where b is the Burgers vector, D_{gbo} is the frequency factor for grain boundary diffusion, and Q_{gb} is the activation energy for grain boundary diffusion. The stress exponent for creep, n, as estimated from Eq. 9, exhibits high and variable values; for the applicable range of stresses, $n > 5$. Accordingly, it is expected that ductility in nc-materials would be much lower than those characterizing micrograined superplastic alloys for which $n = 2$, since ductility depends on $1/n$ ($n = 1/m$) as shown by the following (5):

$$e_f\% = [\exp(C/(n-1)) - 1] \times 100 \quad (10)$$

where $C = (n - (1/n))\ ln\ (400/n)$; $n = 1/m$. As shown in reference [6], combining Eqs. 9 and 10 leads to predicting that ductility not only is low and in the range 2–10 %. From a microstructural point of view, the low ductility can be explained in terms of the details of the concept of dislocation accommodated boundary sliding [5]. According to the concept, dislocations are generated under the local stresses and move in the blocking grains to boundaries where they are annihilated. This scheme results in the absence of dislocation accumulation and interactions, that is, nanograins are not able to sustain arrays of dislocations. This leads to loss of work hardening, which in turn results in low ductility.

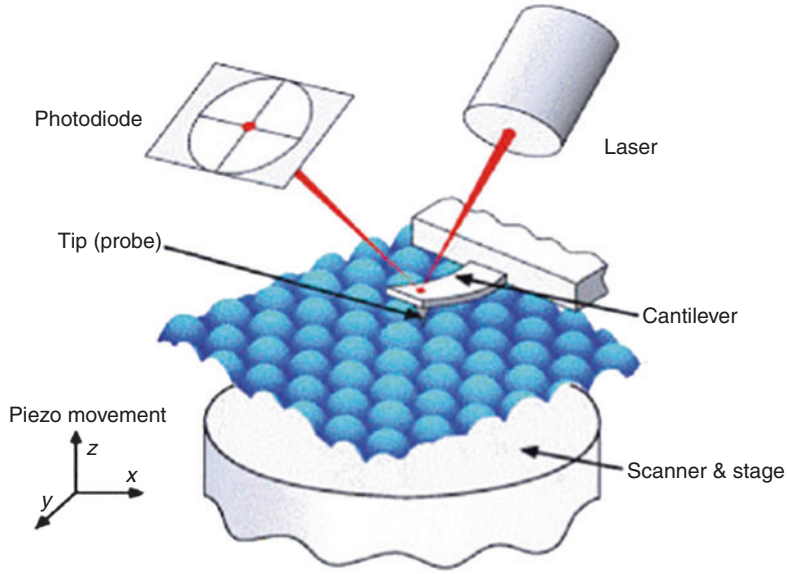
Future Research

Creep

Nc-nickel has served as a model material to study the mechanical behavior of nc-materials. However, in order to establish the generality of the results, it is necessary to investigate the deformation behavior of other nc-materials using approaches similar to those adopted in studying the creep characteristics of nc-Ni. In addition, there is a need to study the creep behavior of nc-aluminum alloys to examine their thermal stability under temperature and stress.

Grain Boundary Sliding

Atomistic computer simulations on nc-materials have revealed the dominance of boundary sliding with decreasing grain size at room temperature. Despite this evidence and despite the need to establish the role of boundary sliding during



Mechanical Properties of Nanocrystalline Metals, Fig. 9 The illustration shows how the technique of atomic force microscopy works

deformation in nc-materials, the occurrence of this process has not been fully addressed in terms of experimental measurements. It is important to experimentally examine the hypothesis that strain measured during creep in nc-metals mostly arises from sliding.

Traditional techniques (surface marker and scanning electron microscopy) that were utilized in measuring sliding micrograined superplastic alloys do not have sufficient resolution. Recently, a solution for the above problem has become available: investigators have demonstrated that atomic force microscopy (AFM) provides a powerful technique for obtaining high-resolution images of the surface topography of bulk materials. AFM works by scanning a fine ceramic or semiconductor tip over a surface. The tip is positioned at the cantilever beam shaped much like a diving board. As the tip is repelled by or attracted to the surface, the cantilever beam deflects (see Fig. 9). The magnitude of the deflection is detected by a laser that reflects at an oblique angle from the very end of the cantilever. A plot of the laser deflection versus tip position on the sample surface provides the resolution of the hills and valleys that constitute the topography of the surface. On this basis, processes such as boundary sliding that are characterized by the occurrence of offsets at the surface of a flat specimen can be detected in ultrafine-grained and in nc-materials by AFM.

Evidence of Dislocations Activity During Deformation

Transmission electron microscopy (TEM) observations on deformed samples of nc-nickel have indicated the presence of few dislocations in grain interiors. However, *the absence of evidence regarding dislocations in deformed nc-Ni is not evidence of the absence of dislocation activity during deformation*. According to the model of dislocation accommodated boundary sliding [4], dislocations are generated under the local stresses and move in the blocking grains to boundaries where they are annihilated. Also, during unloading and/or subsequent specimen preparation for TEM examination, dislocations could disappear; dislocations only need to travel a very short distance of less than 100 b to be lost at nearby grain boundaries. The hypothesis to be tested is that dislocations are generated during deformation in nc-materials and that their absence in TEM micrographs are due to their annihilation at boundaries during deformation

or during unloading and subsequent sample preparation. In order to test the validity of this hypothesis, it is suggested that strong obstacles such as nanoscale particles be introduced in the grain interior to provide effective barriers to the movement of dislocations, leading to attachment of dislocations to particles and to dislocation tangles.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Nanoparticles
- ▶ Nanotechnology
- ▶ Transmission Electron Microscopy

References

1. Gleiter, H.: Nanostructured materials: basic concepts and microstructure. Acta Mater. **48**, 1–29 (2000)
2. Suryanarayana, C.: Mechanical Alloying and Milling, p. 333. Marcel Dekker, New York (2004), Chapter #13
3. Meyers, M., Chawla, K.: Mechanical Behavior of Materials, 2nd edn, pp. 345–352. Cambridge University Press, Cambridge (2009)
4. Kassner, M.E., Perez-Prade, M.-T.: Fundamental of Creep in Metals and Alloys. Elsevier, Amsterdam (2004)
5. Mohamed, F.A.: The role of boundaries during superplastic deformation. Surf. Interface Anal. **31**, 532–546 (2001)
6. Mohamed, F.A., Chauhan, M.: Interpretation of the creep behavior of nanocrystalline Ni in terms of dislocation accommodation boundary sliding. Metall. Mater. Trans. A **37A**, 3555–3567 (2006)
7. Meyers, M.A., Mishra, A., Benson, D.J.: Mechanical properties of nanocrystalline materials. Prog. Mater. Sci. **51**, 427–556 (2006)
8. Schuh, C.A., Nieh, T.G., Iwasaki, H.: Acta Mater. **51**, 431–443 (2003)
9. Chokshi, A.H., Rosen, A., Karch, J., Gleiter, H.: Scr. Metall. **23**, 1679–1683 (1989)
10. Schiotz, J., Jacobsen, K.W.: Science **301**, 1357–1359 (2003)
11. Mohamed, F.A., Yang, H.: Deformation mechanisms in nanocrystalline materials. Metall. Mater. Trans. A **41A**, 823–837 (2010)
12. Dieter, G.: Mechanical Metallurgy. McGraw Hill Science Engineering, New York (1986)

Mechanical Properties of Nanostructures

Laurent Pizzagalli, Sandrine Brochard, Julien Godet and Céline Gérard
Department of Physics and Mechanics of Materials, Institut P′, CNRS – Université de Poitiers ENSMA, Futuroscope Chasseneuil, France

Synonyms

Elastic and plastic behavior of materials with nanometric dimensions; Properties related to applied stress or strain

Definition

In the context of materials science, the usual definition of a nanostructure implies that at least one of its dimension ranges from few nanometers to at most few hundreds. Then, according to the numbers of these dimensions, it could be (non-exhaustive list) a nanofilm, a membrane, a nanoplate, a multilayer (one dimension), a nanowire, a nanopillar, a nanobridge, a nanotube (two dimensions), or a nanoparticle, a nanograin (three dimensions). The response of these systems to an applied stress (or equivalently to a deformation) depends on their mechanical properties. Those include the elastic regime, for which changes induced in the material are reversible, and plastic regime, for which they are not.

Concepts

Nanostructures Versus Bulk Materials

The main difference between macroscale materials and nanostructures lies in the significant role of surfaces (or interfaces) for the latter. Atoms located at surfaces have typically fewer neighboring atoms than in bulk, leading to marked changes of the local

electronic structure. Structural modifications usually occur, thanks to surface relaxation or reconstruction, both mechanisms leading to surface energy minimization, but also adding an extra contribution to surface stress. Besides, surfaces are often reactive, and most of nanostructure surfaces are covered with passivants or oxide layers, except in a well-controlled environment.

Surfaces and interfaces can have a strong impact on mechanical properties of nanostructures. In the elastic regime, the moduli that link stress to strain can change compared to their bulk counterparts, due to the surface stress. The latter will either facilitate or hinder the deformation of the nanostructure, thus decreasing or increasing the elasticity coefficients. In the plastic regime, deformation mechanisms can be different in bulk and in nanostructures. For instance, the dislocation starvation process has been proposed to explain the first stages of plastic deformation in nanostructures [1]. Also, surfaces act as nucleation centers for dislocations [2]. This is also true for boundaries between nanograins, which can emit or absorb dislocations [3] or favor other deformation mechanisms like twinning for instance [4]. Furthermore surface diffusion may play an important role in the plastic deformation of thin nanowires.

In addition to surfaces/interfaces, the mechanical behavior of nanostructures is influenced by reduced dimensions. In fact, the number of preexisting defects which can act as sources for plastic deformation mechanisms scales with the volume of the nanostructure. This source exhaustion is often proposed as an explanation for the hardening in small nanostructures, as reported in many cases [5]. Reduced dimensions can also hinder common bulk mechanisms. At last, quantum confinement effects can become significant in nanostructures with a dimension of few nanometers. For instance they have been shown to influence elastic moduli [6].

Another phenomenon associated to nanostructures is phase transformation. In fact, in certain systems, phases metastable in bulk can be stabilized in nanostructures due to the large surface (or interface) to bulk ratio. This has obvious consequences on the mechanical properties. Finally, a special set of nanostructures includes systems which are characterized by a specific structure, as nanotubes or fullerenes, for example, inducing specific mechanical properties.

As a consequence of all the aforementioned points, the general response of nanomaterials to a mechanical solicitation can be much different than their bulk counterparts. This is spectacularly exemplified in the case of covalent materials like semiconductors. Although they are usually brittle at room temperature in bulk state, small systems can exhibit ductility instead [7, 8]. Overall, the mechanical behavior is greatly changed by the reduction of dimensions [9].

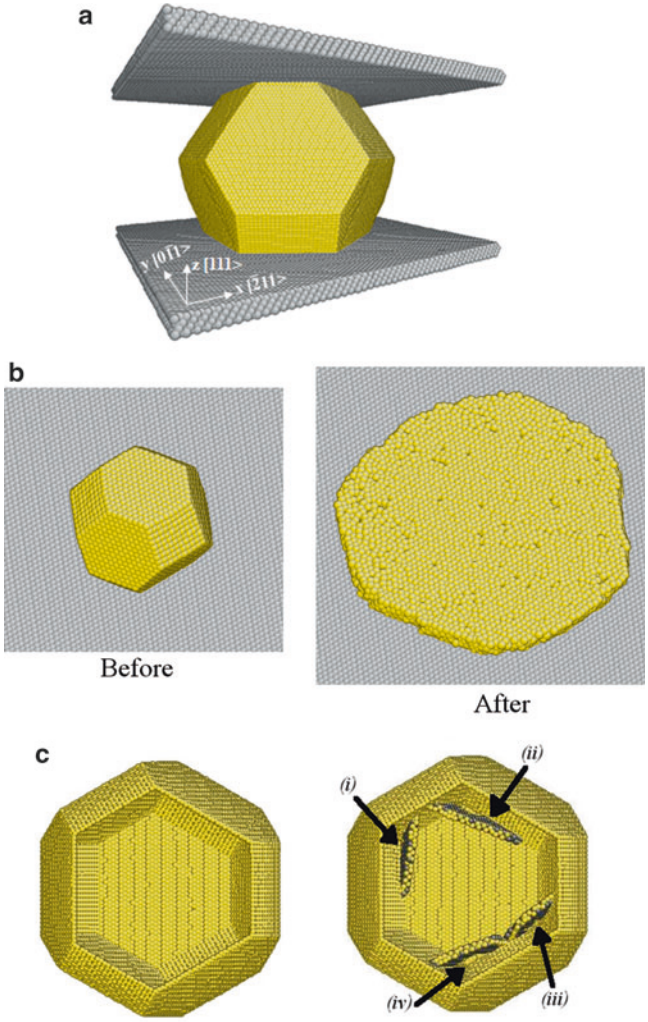
Molecular Modeling Versus Experiments

In molecular modeling, investigated system is described at an atomistic level, i.e., as a collection of atoms. Molecular modeling is complementary to experiments for several reasons. First, it allows to probe the mechanical properties of very small nanostructures, in a dimension range hardly attainable experimentally. Then, in the simulations, it is possible to follow the trajectories of all atoms in the system, which provides a wealth of information. Activated mechanisms can be thoroughly analyzed. Also, the time resolution is of the order of a femtosecond, much lower than in common mechanical testing experiments. This point is particularly important in the plastic regime, where plasticity mechanisms such as dislocation nucleation or fracture propagation can occur at very short timescale. Finally, molecular modeling allows a full control of the initial system and of the mechanical test. The effect of specific features or conditions, even beyond experimental realizations, can then be finely examined.

Methodologies

Different Theoretical Frameworks

There are two main classes of molecular modeling frameworks. In the first one, commonly called "classical," the electronic structure is not calculated and only the forces between atoms are considered. These are typically computed from empirical interatomic potentials, which are usually fitted on known bulk



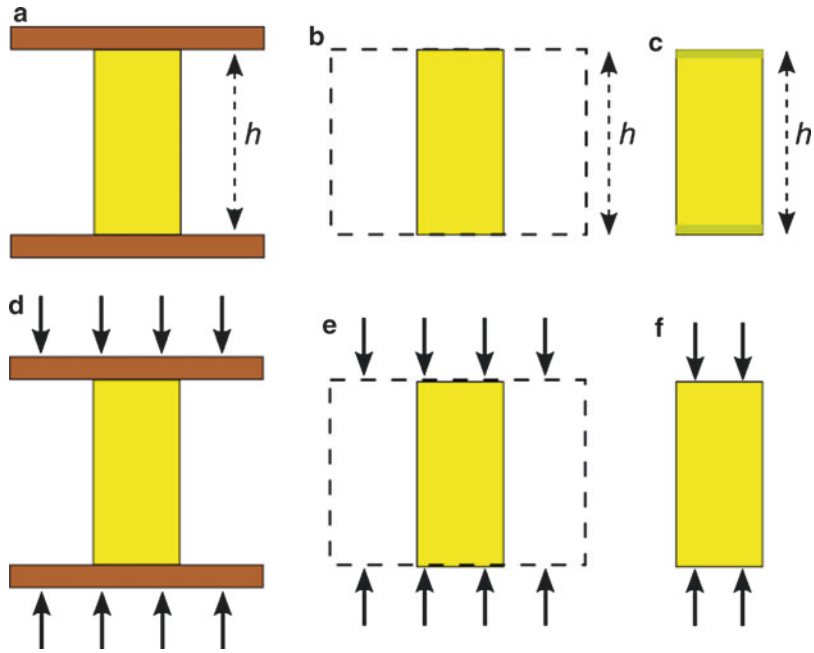
Mechanical Properties of Nanostructures, Fig. 1 Atomistic simulations of the flat punch compression of gold nanoparticles (Courtesy of D. Mordehai [10]). (**a**) *Ball-and-stick* representation of the system setup (Au nanoparticle in *gold* and flat punch indenter atoms in *gray*) for a 4.9 nm high nanoparticle. (**b**) *Top view* of the nanoparticle, before and after complete load. (**c**) *Top view* of the nanoparticle, before and after the first dislocation nucleation events (*i*)–(*iv*) during compression (only atoms in free surfaces and in the defect cores are shown; atoms in stacking faults are shaded in *gray*)

materials properties. This is often the selected approach for modeling nanostructures, since it allows to describe systems whose dimensions are within the reach of experiments. With this approach, multicomponent systems can be more difficult to deal with, since appropriate potentials must be available to model all different atomic interactions.

In the second class of molecular modeling frameworks, the electronic structure is calculated and taken into account into the modeling. There are several ways to do it with different levels of approximation. In the case of nanostructure modeling, tight-binding and density functional theory are possible choices. However, although providing accurate and reliable results, only small nanostructures with dimensions of the order of several nanometers can be dealt with these approaches. As a consequence, to date most of the molecular modeling investigations have been performed using interatomic potentials.

Technical Setup

Prior to the numerical simulation step, one has to build the initial system by setting the positions in space of all needed atoms. This allows for a total control of the geometry and composition of the nanostructure. For instance, the Fig. 1a shows a typical setup for the compression of a nanoparticle. An important issue for the system setup concerns the boundary conditions used in the simulation. The most common choice is periodic boundary conditions (PBC). It can be required



Mechanical Properties of Nanostructures, Fig. 2 Possible ways to load a nanostructure (*yellow*) in a molecular modeling simulation. (**a**) and (**d**): the load is transmitted by an "indenter" (*brown*) made of atoms. (**b**) and (**e**): the load is transmitted by the simulation cell. (**c**) and (**f**): the load is directly applied on the nanostructure. In (**a**–**c**), the mechanical testing is strain controlled by changing the dimension h, eventually rescaling atoms positions, whereas it is stress controlled in (**d**–**f**). Simulations are usually performed in strain-controlled mode, easier to use than the stress-controlled mode

depending on the selected theoretical framework. For instance, PBC are enforced when plane-waves density functional theory simulations are made. This has obvious consequences on the modeling setup. PBC can be used to model a nanostructure with one or several infinite dimensions (as one of the examples will show in the following). When PBC are used, it is important to check possible artificial size effects due to the interaction of the system with its replicas. Note that these potential size effects can often be dealt with by including a sufficient amount of void around the system, which then exhibits surfaces, simply by increasing the size of the simulation cell.

The mechanical properties of a given system can be obtained as its response to a mechanical load. Three different classes of methods can be used to apply this load. In the first one (Fig. 2a, d), not only the nanostructure is modeled with atoms but also the devices that transmit the load in a real experiment (such as plates or an indenter). This method is probably the closest to experiments and allows for taking into account the additional mechanical response of the devices or possible adhesion effects. Besides, devices can also be modeled as non-attractive or infinitely rigid. The mechanical test can be strain or stress controlled, depending whether the position of devices is fixed, or an applied stress is used. The drawback of this approach is the extra computational cost associated with the atoms used for modeling the devices. A second option is to take advantage of PBC (Fig. 2b, e). In that case, strain- or stress-controlled testing is done by simply setting the computational cell dimensions or applying the appropriate stress on the cell faces. With this method, load-transmitting devices are not considered. This is best suited to determine intrinsic nanostructure properties such as elastic constants, except for nanoparticles for which void is necessary present around the whole system. Finally, in the third method, the load is transmitted to the nanostructure by either applying a force field on selected atoms (stress controlled) or by fixing the

positions of these atoms (strain controlled). The latter is also called fixed boundary conditions and can be useful in specific cases. For instance, it does not rely on the use of PBC (as the first method), and additional atoms are not required.

Molecular Dynamics or 0 K Force Relaxation

Simulations are often performed at finite temperature using molecular dynamics. In this framework, the dynamical evolution of all atoms in the system can be calculated by solving the Newton equations of motion, interatomic forces being obtained with one of the aforementioned formalisms. In order to compare with experiments, simulations are typically done in the canonical ensemble, using a thermostat to keep the temperature to the desired value. Generally, a simulation starts by a thermalization step, followed by an increasing applied strain (or stress) on the nanostructure. Important quantities such as the energy variation, the positions of atoms, or the local stress tensor can be monitored during the test for further analyses.

Besides, zero temperature simulations can also be performed (also known as molecular mechanics). For instance, this is done to compute elastic constants or to investigate the effect of stress on the plastic deformation without the stochastic influence of thermal activation. In that case, forces are simply relaxed at a given strain (or stress), leading to a configuration of minimum energy. To be complete, one can also mention that advanced simulation techniques such as chain of states methods have also been applied for determining the mechanical properties of nanostructures [11].

Examples

Molecular modeling has been used to investigate the mechanical properties of many nanostructures. Most of these were metallic systems or nanotubes, for which there are a lot of experimental data. But the literature also includes several studies of nanostructures made of covalent or ionic materials, such as semiconductors or oxides. Furthermore, although most of the studied systems are quite ideal, there is an effort to consider more realistic or complex nanostructures in molecular modeling simulations. Those could be core/shell systems, for instance, where the nanostructures are composed of two different phases or materials.

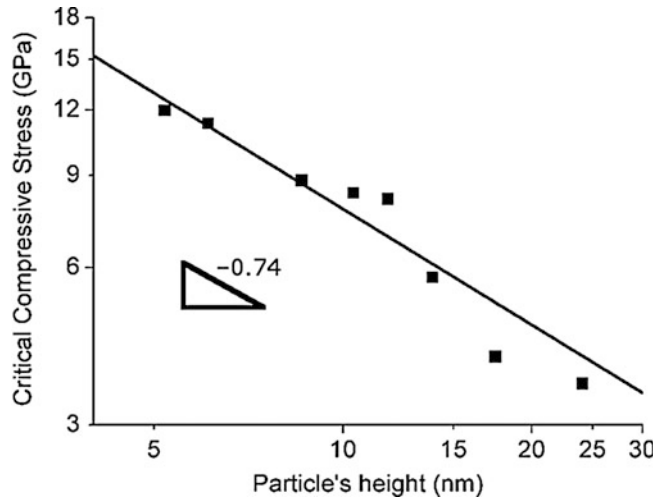
In this entry, three examples of recent works focusing on the molecular modeling of the plasticity properties of nanostructures are presented, in order to better illustrate the concepts developed in the previous sections.

Metallic Nanoparticles

The first example concerns the plastic deformation of facetted gold nanoparticles [10]. The latter can be experimentally obtained as stand-alone nanostructures on a sapphire substrate by using a solid-state dewetting process and are next compressed by a flat punch indenter. Equivalent computational systems have been built (Fig. 1a), the nanoparticle being located between two infinitely rigid gold plates, which are then brought closer during a molecular dynamics simulation, in a strain-controlled mode, with a constant rate of 1 m s^{-1}. An embedded atom model potential is used to model interatomic interactions. The procedure used here corresponds to the case (a) in Fig. 2.

After complete compression of the nanoparticle, a flake-like structure is obtained (Fig. 1b), in excellent agreement with the combined experiments. Nevertheless, the true power of the simulations is revealed when focusing on the plasticity onset. Experimentally, one can only see the termination of the elastic regime due to a sudden and large strain burst. Instead, atomistic simulations allowed to fully characterize the very first stages of plasticity. Hence it is revealed that plastic deformation first occurred by the nucleation of leading Shockley partial dislocations from the upper vertices of the nanoparticles (Fig. 1c).

Experiments as well as simulations both showed that the yield stress associated with the onset of plasticity is size dependent and well reproduced by a power law (Fig. 3). The exponents are found to be -0.77 and -0.74 in experiments and simulations, then in excellent agreement. This proves the ability of molecular



Mechanical Properties of Nanostructures, Fig. 3 Compressive stress at the point of nucleation of the first dislocation versus particle height in the simulations (Adapted from Fig. 5 in Mordehai et al. [10], courtesy of D. Mordehai). A scaling effect can be clearly seen and is best fitted by a power law with an exponent of -0.74 ± 0.08

modeling to accurately describe a complex process such as the plastic deformation of a metallic nanostructure.

Metallic Nanowires

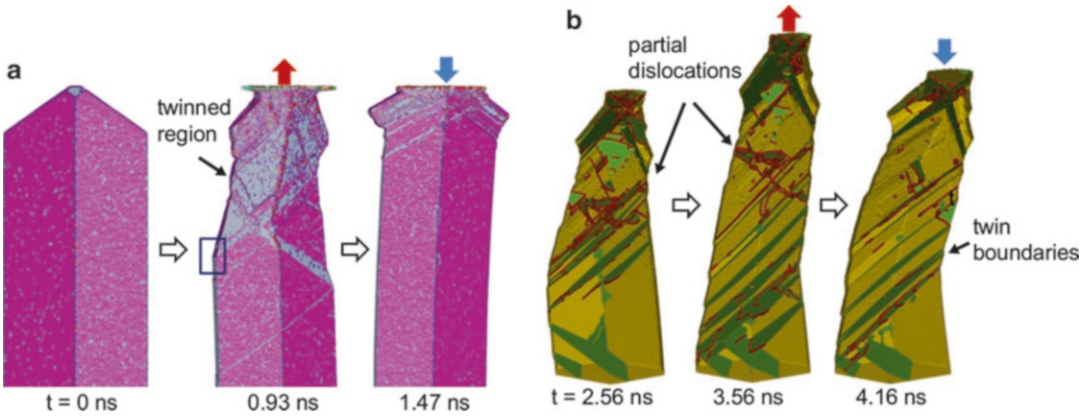
A second example concerns the plastic deformation of gold nanowires with a $\langle 110 \rangle$ orientation. Molecular dynamics simulations have been performed for nanowires of lengths 200 nm and diameters of at most 50 nm, including several millions of Au atoms, and compared to deformation experiments on slightly larger systems [12]. The flat punch indenter and the supporting surface ($SrTiO_3$ in experiments) are modeled by few infinitely rigid gold layers. The setup is then quite similar to the previous example, albeit the used EAM potentials are different and the deformation rate of 20 m s^{-1} is higher.

In both experiments and simulations, the nanowire is initially compressed until a significant amount of plastic deformation is obtained, first due to the nucleation of perfect dislocations in form of prismatic loops in both cases. This stage also allows to form a strong adhesive contact between the testing devices and the nanowire. In a second step, the nanowires are deformed in tension and yield by the formation of twins (Fig. 4). The simulation reveals that these twins result from the successive nucleation of leading partial dislocations in adjacent planes. Reversing again the deformation direction leads to a detwinning process in the nanowire in compression (Fig. 4), due to the successive nucleation of trailing partial dislocations in the twinned layers. This quite impressive result can be explained by the differences in resolved shear stresses for leading and trailing partial dislocations according to the deformation direction. As a result, it is demonstrated that these gold nanowires combine an ultrahigh strength together with a very high ductility (strains of at most 30 % were obtained).

Overall, the virtual mechanical tests performed with these large-scale molecular dynamics simulations lead to structures in excellent agreement with in situ transmission electron microscopy observations. The activated plasticity mechanisms also appear to be similar in simulations and experiments. This confirms the reliability of these simulations for the investigation of mechanical properties of nanostructures. This study also emphasizes how molecular modeling can be helpful for a fine analysis of the plasticity mechanisms.

Silicon Nanowires

As a last example, the molecular modeling of the mechanical properties of a semiconductor nanostructure is presented. J. Guénolé and coworkers have recently investigated the behavior in



Mechanical Properties of Nanostructures, Fig. 4 Snapshots from 300 K molecular dynamics simulations of a gold nanowire, showing the size increase (decrease) of the twinned regions during tension (compression). The *left* representation shows atoms belonging to twins as *gray balls* and the other as *pink balls*. The *right* representation shows later stages, with the presence of partial dislocations as *red lines* and twin boundaries in *green* (Adapted from Fig. 6 in Lee et al. [12], courtesy of E. Bitzek, S. Lee, and S.H. Oh)

compression of silicon nanowires with a ⟨123⟩ orientation, using either conjugated gradient minimization or finite temperature molecular dynamics [13]. In this work, the systems are smaller than in previous examples, with characteristic dimensions of about several tens of nanometers, but several different surface states and nanowire shapes are considered. In particular, the effect of an amorphous coating of the surface is analyzed, in order to obtain insights about the influence of an oxidized nanowire surface, or amorphized following preparation by ion beam implantation for instance.

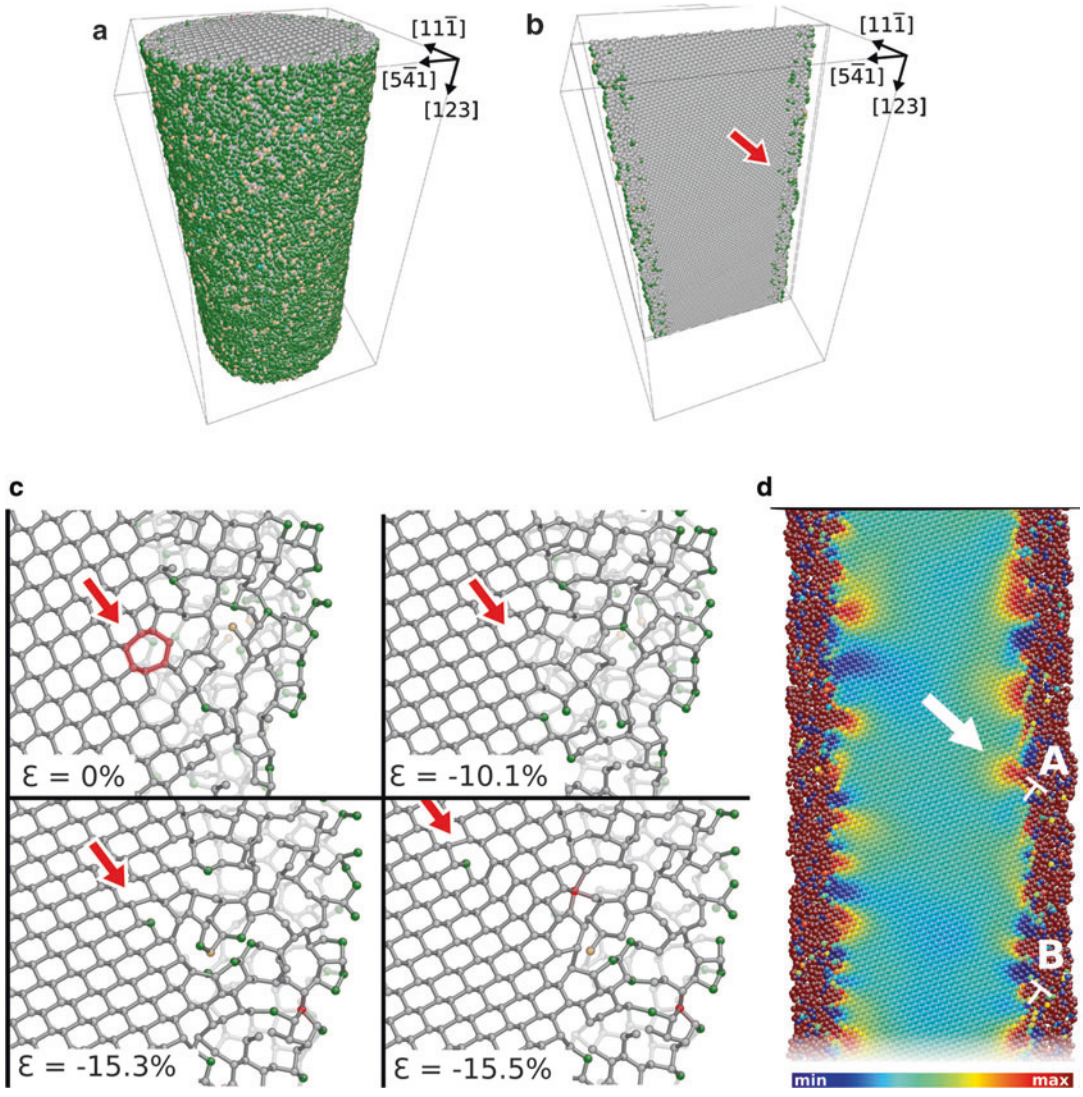
Conversely to the examples described above, no indenters are used in this study, and compression is obtained by scaling the simulation cell along the nanowire orientation (case (b) in Fig. 2). The chosen interatomic potential is specifically designed for modeling silicon in high pressure high shear conditions.

The simulations revealed that plasticity at 0 K occurred at large compressive strains, about 15–18 % depending on the surface state and nanowire shape. Dislocations are nucleated from the surface, with an increased probability in case of slightly disordered surfaces, and preferentially at edges in rhombic nanowires. However, these features all vanish when an amorphous Si coating is present. In this case, the activated plasticity mechanism remains dislocation nucleation, but it now always occurs at the interface between the a-Si layer and the c-Si core for a strain of about 15 %. Structural analysis reveals that the dislocation formation is initiated at interface sites characterized by a specific topology (Fig. 5c). Also, only those sites concentrating enough stress can be dislocation sources (Fig. 5d).

This work shows how simulations can be useful for obtaining information on nucleation sites, even when those are located deep below the surface, and to investigate many different configurations (shape, surface state). Besides, while simulations were used as complement to experiments in the previous cases, their predictive power is illustrated in the present case.

Limits and Issues

Atomistic simulations rely on several approximations and suffer from certain technical limitations. For instance, although an accurate description can be obtained using first-principles approaches, they are computationally too demanding to deal with nanostructure sizes usually found in experiments. This is why the vast majority of the available simulation works make use of classical interatomic potentials. The reliability of these potentials is an important issue. In fact, they are



Mechanical Properties of Nanostructures, Fig. 5 Atomistic simulation of the mechanical test of a Si nanowire at 0 K (Figure adapted from Guénolé et al. [13]). (**a**) 3D view of the nanowire for a compressive strain of −15.5 % strain, just before yielding (atoms colored according to the coordination number). (**b**) Same nanowire state, but only the slice including the site of plasticity onset is shown (marked by the *red arrow*). (**c**) Zoom on this plasticity mechanism for different deformation states. (**d**) Shear strain map just before yielding (the *white arrow* shows the site *A*, characterized by a stress concentration high enough to nucleate a dislocation, unlike the site *B*)

fitted on several properties of the materials and usually tested for low energy configurations, not far from stable states. Conversely, the atomistic modeling of nanostructures first requires a good description of surfaces, which include undercoordinated atoms. Moreover, mechanical testing of nanostructures in the plastic regime involves highly strained systems and distorted atomic configurations. For instance, the description of crack tip formation and propagation is one of the most difficult phenomena to reproduce with interatomic potentials, for it involves bonds that are highly stretched until the rupture. Without any references to compare with, it is not always possible to assess

the reliability of a given potential in a specific situation. Nevertheless, it seems that for fcc metals, an accurate and reliable description can be obtained using embedded atom method potentials, as shown in the aforementioned examples. For other materials such as bcc metals or covalent systems, no such consensus exists. Then, if an original mechanism is revealed by using a given interatomic potential, one may often question whether the mechanism is real or an artificial product of the use of this potential. One way to tackle this issue is to try to reproduce the same mechanism in a much smaller systems, using more accurate and reliable calculations [14]. Another approach is to employ different interatomic potentials for the same investigation. Obtaining the same outcome in all cases surely strengthen the conclusions.

Another issue is related to the size difference between systems investigated in experiments and simulations. It is possible to perform classical molecular dynamics simulations including tens and even hundreds of millions of atoms on nowadays computers (but at the expense of the timescale, see below). This corresponds to system dimensions in the range 50–100 nm at most, thus large enough for many of the experimentally investigated systems. Other systems in the submicronic scale remain hardly accessible with current atomistic simulations. However, the size issue is not very severe. In fact, on the one hand, atomistic modeling using interatomic potentials scales linearly as the number of considered atoms, and the continuous increase of available computational power should allow larger systems in the future. On the other hand, the comparison with experiments is increasingly possible because of impressive recent experimental developments. It is now more and more feasible to fabricate systems truly in the nanoscale, which can be mechanically tested (and sometimes monitored in situ). For instance it has been possible to measure the mechanical strength of nanowires with diameters of few nanometers [15].

At last, the timescale issue is probably the most critical one in dynamical molecular modeling simulations. A standard timestep in these simulations is about 1 femtosecond, because of physical arguments and numerical stability constraints. Then to reach a simulation duration of only 1 ns, 10^6 iterations are required. While this is not problematic for small systems, it severely conflicts with the above size issue when one wants to consider multimillion atom systems. Mechanical properties of materials are experimentally probed by deformation apparatus working at rates in the range $10^{-4} - 10^{-1}$ s^{-1}, typically. However, because of the timescale issue, molecular dynamics simulations of large systems are performed at best at deformation rates of 10^7 s^{-1}, thus many orders of magnitude higher. Such a huge difference is probably not an issue in the elastic regime but can be dramatic in the plastic regime. For instance the nucleation of the first dislocation, setting the onset of plasticity in the nanostructure and defining the yield point, is a rare event in the dynamical evolution of the system. Very high deformation rates mean less time for this unfrequent event to occur. As a consequence, yield strains (or stresses) are much higher in numerical simulations than in experiments. Fortunately it is possible to apply post-corrections to bring numbers in closer agreement [16]. Nevertheless, it cannot be stated with certainty that plasticity mechanisms activated in the simulations and in experiments at different stresses/strains are always the same, since counterexamples exist [17]. There have been several attempts to improve this timescale issue. For instance, few orders of magnitude can be gained by using specific techniques such as hyperdynamics, parallel replica dynamics, or temperature accelerated dynamics [18]. Another approach relies on the use of static saddle point searching methods such as ART, dimer, or nudged elastic band to investigate mechanisms [11, 19], but the comparison with experiments is much less straightforward than with molecular dynamics.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Mechanical Properties of Nanostructures
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nano-Pillars

▶ Plasticity Theory at Small Scales
▶ Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures

References

1. Greer, J.R., Nix, W.D.: Nanoscale gold pillars strengthened through dislocation starvation. Phys. Rev. B **73**, 245410 (2006)
2. Nix, W.D., Lee, S.W.: Micro-pillar plasticity controlled by dislocation nucleation at surfaces. Philos. Mag. **91**, 1084 (2011)
3. Van Swygenhoven, H., Derlet, P.M., Frøseth, A.G.: Stacking fault energies and slip in nanocrystalline metals. Nat. Mater. **3**, 399 (2004)
4. Chen, M., Ma, E., Hemker, K.J., Sheng, H., Wang, Y., Cheng, X.: Deformation twinning in nanocrystalline aluminum. Science **300**, 1275 (2003)
5. Uchic, M.D., Dinmiduk, D.M., Florando, J.N., Nix, W.D.: Sample dimensions influence strength and crystal plasticity. Science **305**, 986 (2004)
6. Cherian, R., Gerard, C., Mahadevan, P., Cuong, N.T., Maezono, R.: Size dependence of the bulk modulus of semiconductor nanocrystals from first-principles calculations. Phys. Rev. B **82**, 235321 (2010)
7. Michler, J., Wasmer, K., Meier, S., Östlund, F., Leifer, K.: Plastic deformation of gallium arsenide micropillars under uniaxial compression at room temperature. Appl. Phys. Lett. **90**, 043123 (2007)
8. Kang, W., Saif, M.T.A.: In situ study of size and temperature dependent brittle-to-ductile transition in single crystal silicon. Adv. Funct. Mater. **23**, 713 (2013)
9. Gerberich, W.W., Michler, J., Mook, W.M., Ghisleni, R., Östlund, F., Stauffer, D.D., Ballarini, R.: Scale effects for strength, ductility, and toughness in "brittle" materials. J. Mater. Res. **24**, 898 (2009)
10. Mordehai, D., Lee, S.W., Backes, B., Srolovitz, D.J., Nix, W.D., Rabkin, E.: Size effect in compression of single-crystal gold microparticles. Acta Mater. **59**, 5202 (2011)
11. Zhu, T., Li, J., Samanta, A., Leach, A., Gall, K.: Temperature and strain-rate dependence of surface dislocation nucleation. Phys. Rev. Lett. **100**, 025502 (2008)
12. Lee, S., Im, J., Yoo, Y., Bitzek, E., Kiener, D., Richter, G., Kim, B., Oh, S.H.: Reversible cyclic deformation mechanism of gold nanowires by twinning-detwinning transition evidenced from in situ TEM. Nat. Commun. **5**, 3033 (2014)
13. Guénolé, J., Godet, J., Brochard, S.: Plasticity in crystalline-amorphous core-shell Si nanowires controlled by native interface defects. Phys. Rev. B **87**, 045201 (2013)
14. Godet, J., Brochard, S., Pizzagalli, L., Beauchamp, P., Soler, J.M.: Dislocation formation from a surface step in semiconductors: an ab initio study. Phys. Rev. B **73**, 092105 (2006)
15. Kizuka, T., Takatani, Y., Asaka, K., Yoshizaki, R.: Measurements of the atomistic mechanics of single crystalline silicon wires of nanometer width. Phys. Rev. B **72**, 035333 (2005)
16. Ryu, S., Kang, K., Cai, W.: Predicting the dislocation nucleation rate as a function of temperature and stress. J. Mater. Res. **26**, 2335 (2011)
17. Ph, S.H., Legros, M., Kiener, D., Dehm, G.: In situ observation of dislocation nucleation and escape in a submicrometre aluminium single crystal. Nat. Mater. **8**, 95 (2009)
18. Voter, A.F., Montalenti, F., Germann, T.C.: Extending the time scale in atomistic simulation of materials. Annu. Rev. Mater. Res. **32**, 321 (2002)
19. Pizzagalli, L., Pedersen, A., Arnaldsson, A., Jónsson, H., Beauchamp, P.: Theoretical study of kinks on screw dislocation in silicon. Phys. Rev. B **77**, 064106 (2008)

Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices

Kwanoh Kim[1] and D. L. Fan[1,2]
[1]Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX, USA
[2]Materials Science and Engineering Program, The University of Texas at Austin, Austin, TX, USA

Synonyms

Rotary nanoelectromechanical systems (NEMS); Rotary nanomachines; Rotary nanomotors

Definition

Rotary nanoelectromechanical devices are miniaturized actuators that can generate rotational motion. They typically have a few microns to submicron feature sizes and are referred to as rotary nanomotors because of their size and function. Assembling arrays of rotary nanoelectromechanical system (NEMS) devices is a precise and efficient method to fabricate single or multiple rotary nanomotors in a desired location with

nanoentities, e.g., nanowires, nanotubes, and even biomolecules, as components or building blocks.

Miniaturization of Rotary Motors

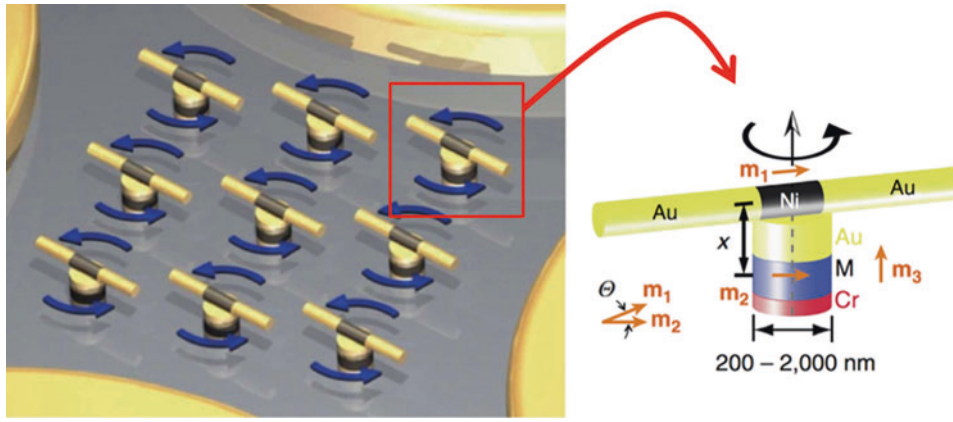
Machines that convert various types of energy sources into rotational motion are widely used in our lives, from small electric motors for household appliances to combustion engines and even large-scale turbines. Since Richard Feynman's legendary talk titled "There's Plenty of Room at the Bottom" in the late 1950s, miniature motors that could work inside the human body has been imagined and inspired researchers. The advance in microelectronics and semiconductor device fabrication technology opened up the field of microelectromechanical systems (MEMS) including a variety of sensors and actuators from hundreds of microns to a millimeter in size. The first rotary micromotors have been developed in the late 1980s [1], and even smaller devices have been reported owing to the astonishing progress in micro-/nanofabrication for the last couple of decades. However, while many techniques have been developed for the fabrication of integrated circuit (IC) microchips even with tens of nanometer feature size, it is still challenging to further reduce the size of rotary motors down to a few microns or submicron range with micromachining and conventional semiconductor device fabrication technology. Only several rotary nanomotors have been successfully fabricated using those sophisticated fabrication processes [2, 3].

Recently, assembling miniaturized electromechanical devices with nanoentities as components, or building blocks, has attracted attention from researchers. Nanoentities have obvious advantages for use in building NEMS devices. Their small footprint and unique chemical, mechanical, electromagnetic properties can significantly downsize devices and effectively improve performance. Moreover, owing to the recent progress in nanotechnology and nanomaterial synthesis process, nanoentities with desired chemical, mechanical, and electromagnetic properties as well as dimensions and geometry can be readily synthesized at low cost [4].

Assembling Rotary NEMS Devices

In order to build a NEMS device from nanoentities, it is essential to precisely bring them to targeted positions, align them in a certain orientation, and securely connect each component forming functional joints. Previously, nanomotors can be assembled without positioning and aligning nanoentities, solely relying on biochemical interactions between nanoentities [5]. However, they have low device yield and could not be suitable for systematic device fabrication for practical use, which tells why precision nanomanipulation is important. While the fabrication processes of nanoentities were well established, e.g., vapor-liquid-solid (VLS) [6], hydrothermal synthesis [7], and template-assisted electrodeposition [8], precision nanomanipulation and assembling strategies were relatively less developed due to significant technical difficulties. Recently, the field has been vigorously explored, and a variety of mechanisms generating and controlling the motion of micro-/nanoparticles have been developed, including optical, plasmonic, magnetic, electric, and acoustic tweezers [9].

Once positioned and aligned, each component should be properly connected and assembled for a working device. Depending on the purpose, working principles, and materials, different types of joints and connections are required. One effective way to accomplish this is to use magnetic interaction between components that contain magnetic materials or are comprised of magnetic nanoentities. Approaches for nanoparticle synthesis with controlled magnetic properties have been well developed owing to the vigorous research in various fields from data storage to biomedical applications [10, 11]. Due to controlled magnetic interactions, magnetic nanostructures can be efficiently integrated with other components and form joints and movable structures. In an alternative approach, chemical bonding and bioconjugation have been widely used to link surface-modified nanoparticles. These only occur with selective pairs of chemicals or biomolecules, useful for site-specific assembling process.



Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices, Fig. 1 Schematic diagrams of nanomotors in an array and an individual assembly of a nanowire on top of a patterned magnet. Reprinted by permission from Macmillan Publishers Ltd: Nature Communications ref [12], copyright 2014

Examples of Rotary NEMS Devices Assembled from Nanoentities

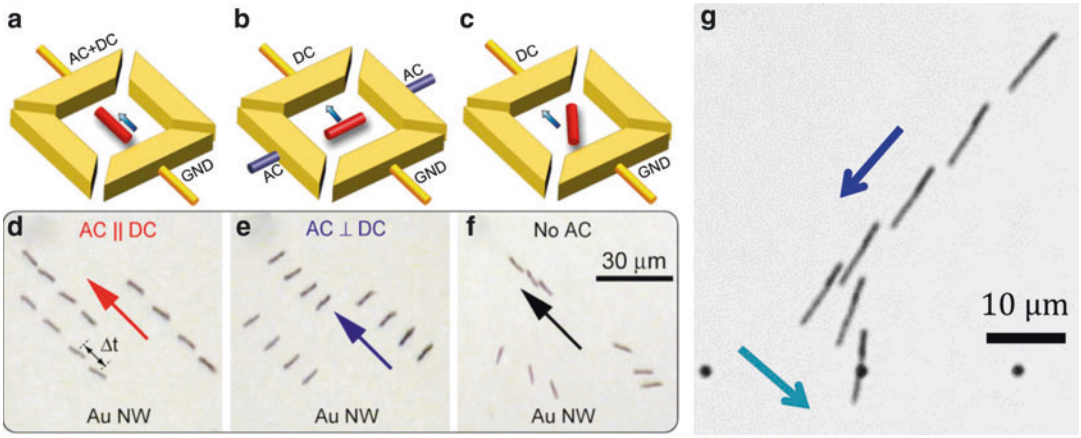
Rotary Nanomotors

Recently, a rotary nanomotor has been assembled from a nanowire and a patterned micro-/nanomagnet working as a rotor and a bearing, respectively (Fig. 1) [12]. 800-nm- to 10-μm-long Au/Ni/Au nanowires are synthesized through template-assisted electrodeposition, and patterned magnets 200 nm to 2 μm in diameter are fabricated by depositing stacks of Cr/Ni/Au thin films on a substrate. Then, the nanowires suspended in deionized (DI) water are transported to patterned magnets with the electric tweezers [13, 14].

The electric tweezers are a nanomanipulation technique to precisely control the motion of nanoentities in a liquid medium using uniform DC and AC electric fields. Nanoparticles with a charge (q) are compelled to move by electrophoretic force ($F_{EP} = qE_{DC}$) in a DC electric field (E_{DC}), while dielectrophoretic torque ($\tau_{DEP} = p \times E_{AC}$) drives nanoparticles, which have the induced dipole moment p, to rotate under a uniform AC electric field (E_{AC}) where there is no electric field gradient [15]. Based on these electrophoresis (EP) and dielectrophoresis (DEP), the electric tweezers can not only convey and transport nanoentities but also rotate them and adjust their orientation by just applying DC and AC electric fields.

The electric tweezers have two orthogonal sets of parallel electrodes that steer the motion of nanoparticles in each direction (Fig. 2a–f). This enables two-dimensional motion control, and nanoentities can be delivered to anywhere in the region surrounded by the electrodes along an arbitrary route at a desired orientation with a precision of 300 nm and a maximum speed of ~80 μm/s. Besides high precision and speed, the electric tweezers are applicable to any nanoentities regardless of their dimensions, geometry, chemical, and electromagnetic properties and can manipulate wide selection of nanoentities.

Once the nanowires are delivered to the magnetic nanopatterns, each component should be properly assembled for a working device. Having a magnetic segment, i.e., nickel, embedded in its center, the nanowire located in the vicinity of the patterned magnet is quickly attracted, attached to the patterned magnet, and even aligned along the magnetization of the patterned magnet (Fig. 2g). This saves time and effort required for the position and posture control of the nanoentities, which not only expedite the whole assembling process but also enhance the quality of the assembly. The patterned magnet is covered with a nonmagnetic layer, i.e., gold, whose thickness determines the magnetic force holding the nanowire, as magnetic force is inversely proportional to the fourth power of the distance (x) between two magnets ($F \sim 1/x^4$).



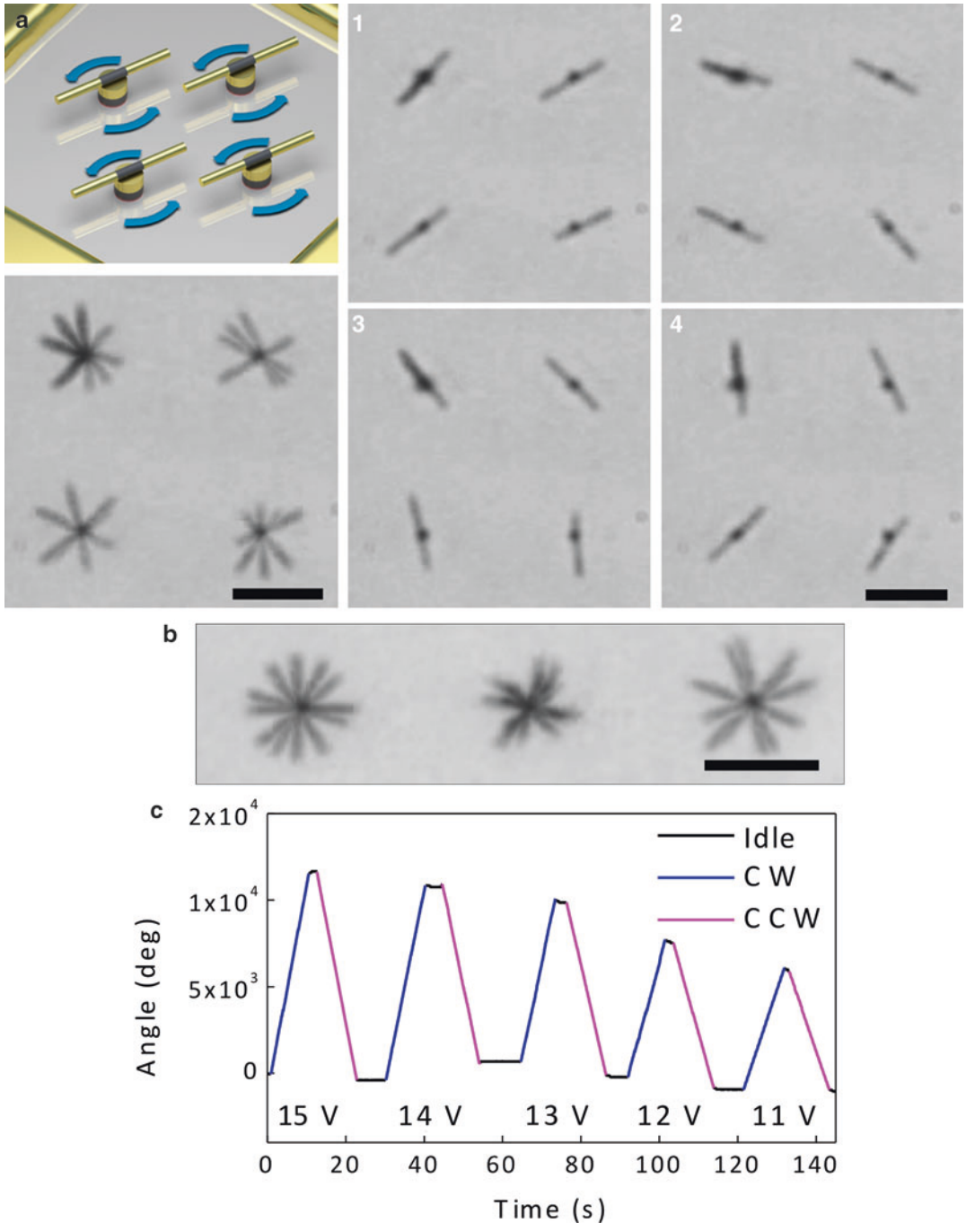
Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices, Fig. 2 (**a**–**c**) Schematic diagrams of the electric tweezers and (**d**–**f**) overlaid sequential micrographs of Au nanowires manipulated with DC and AC electric fields with different configurations. (**a** and **d**) AC||DC, (**b** and **c**) AC⊥DC, and (**e** and **f**) DC without AC. Reproduced with permission from ref [13]. Copyright 2008, AIP Publishing LLC. (**g**) Overlaid sequential micrographs of a Au/Ni/Au nanowire transported and attached to a patterned magnet using the electric tweezers. Reprinted by permission from Macmillan Publishers Ltd: Nature Communications ref [12], copyright 2014

With properly designed dimensions, the nanowire rotor can be securely anchored but still can rotate so that the device readily works as a rotary NEMS actuator.

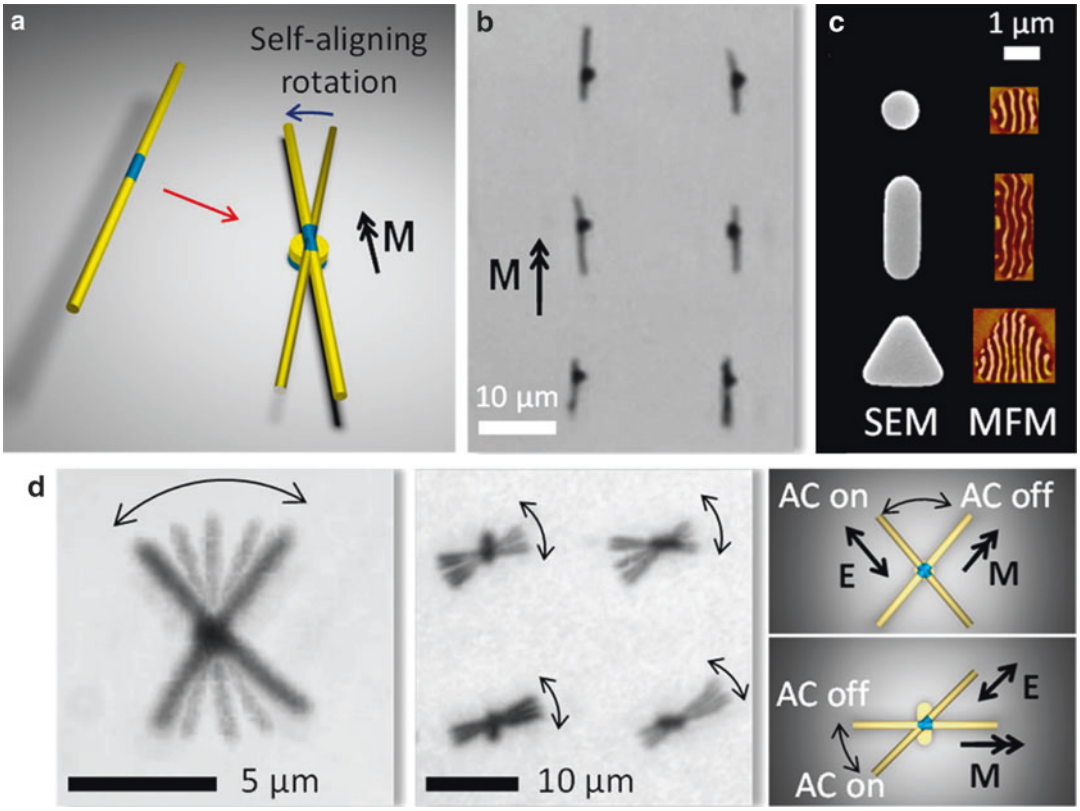
The assembled motor is driven by a rotating electric field generated by four AC voltages (V_{AC}) applied to the surrounding electrodes with 90° phase shift either clockwise or counterclockwise, and their rotation speed (ω) is easily controlled by the voltage ($\omega \sim {V_{AC}}^2$) [12, 16]. Multiple motors are assembled in ordered arrays, and their rotation can be initiated, stopped, and reversed simultaneously (Fig. 3). A motor can rotate over 18,000 rpm and continuously work for 15 h over 2.4×10^5 cycles. Based on the original device, the design including materials and magnetic configurations has been rationally modified for improved durability and controllability. As a result, the lifetime has been enhanced to 80 hours over 1.1 million cycles [17], and it has been even possible to fine control angular displacement and position of the nanomotors like step motors [18].

Rotational Nano-Oscillators

In a similar manner, nano-oscillators rotating between two different positions are assembled from a Au/Ni/Au nanowire, a patterned magnet in different shapes and geometry, e.g., disks, triangular and rectangular plates with feature size from 1 μm to 3 μm, and a set of quadruple microelectrodes [19]. Assembled nanowires in arrays are well aligned parallel to the striped magnetic domain structures uniquely remaining in the Ni magnets after the demagnetization and magnetization treatment (Fig. 4a). Their orientation can be readily altered at an arbitrary angle with the demagnetization and re-magnetization treatment in a different direction. When the nanowires are perturbed or rotated to a certain position, they return to their original position. This is caused by the magnetic interactions between the nanowires and the patterned magnets, more specifically magnetic torque exerted on the nanowires ($\boldsymbol{\tau_m} = \boldsymbol{m} \times \boldsymbol{B}$), where $\boldsymbol{m}$ and $\boldsymbol{B}$ are the magnetic moment of the nanowires and the magnetic field created by the patterned magnets, respectively. The assembled nanowires synchronously oscillate between distinct angular positions, e.g., 0° and 45°, when an AC electric field is switched on and off in a desired direction other than the initial orientation of the nanowires set by their magnetic interaction with the patterned magnets (Fig. 4b). Oscillation amplitude and



Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices, Fig. 3 Nanomotors in ordered arrays. Reprinted by permission from Macmillan Publishers Ltd: Nature Communications ref [12], copyright 2014. (**a**) Schematic diagram, overlaid, and sequential micrographs showing a 2×2 array of nanomotors rotating clockwise. (**b**) Overlaid micrographs of a 1×3 array of nanomotors rotating simultaneously and (**c**) the cumulative angle of the left nanomotor in figure (**b**) rotating at different AC voltages clockwise (CW) and counterclockwise (CCW)

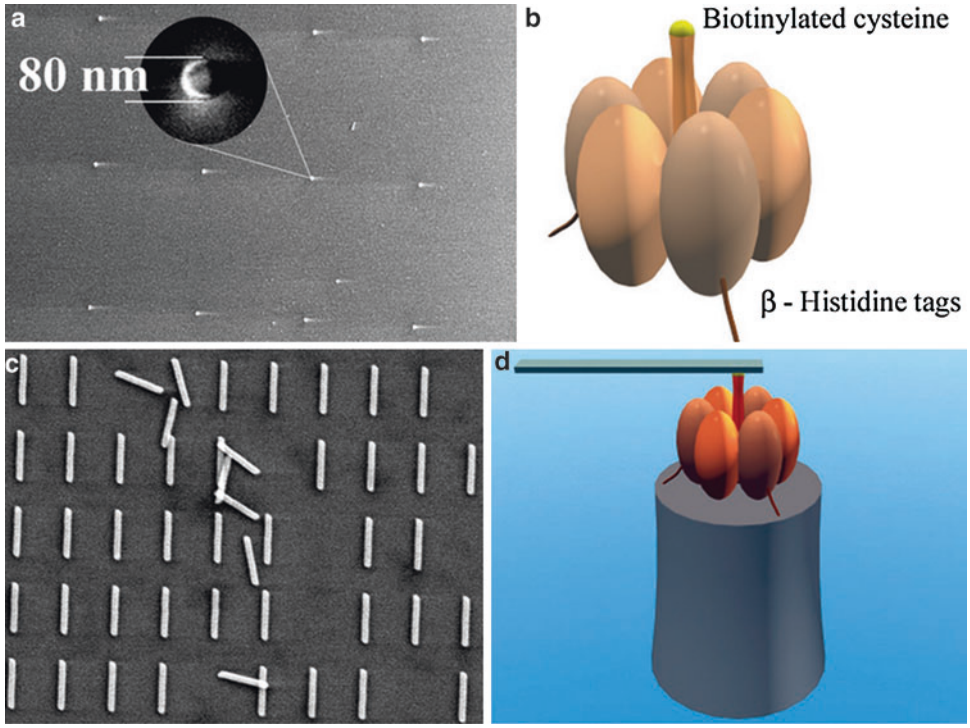


Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices, Fig. 4 Nano-oscillators. Reprinted with permission from ref [19]. Copyright 2013 American Chemical Society. (**a**) Illustration of a self-aligning nanowire with the magnetic interaction between the components. (**b**) Nanowires assembled in an ordered array with their orientation well aligned parallel to the magnetization of the patterned magnets (M). (**c**) Scanning electron microscope (SEM) images of patterned magnets in different shapes and their magnetic force microscope (MFM) images showing striped magnetic domain structures. (**d**) Overlaid sequential micrographs and schematic diagrams of oscillating nanowires in different configurations

frequency are controlled with the direction, intensity, and switching frequency of the AC electric field. A nano-oscillator can toggle over 4,000 consecutive cycles and finally stopped by the increased friction at the interface of the nanowires and the pattern magnets. One thing interesting about this type of NEMS oscillators is that their oscillation is enabled by magnetic interaction between the components and traditional mechanical springs or flexures are not required.

Organic-Inorganic Hybrid Nanomotors

In nature and biological systems, miniaturized motors and actuators commonly exist and function for various purposes with great performance and efficiency [20]. With those biological molecular motors powered by the dissociation of hydrolyzing adenosine triphosphate (ATP) in a surrounding medium, nanoscale organic-inorganic hybrid actuators have been assembled and demonstrated their rotation at fixed positions [5]. A batch of 750–1400-nm-long Ni nanorods with 150 nm in diameter and Ni nanopillars of 50–120 nm in diameter are separately patterned on substrates. The Ni nanorods are released from the substrate and suspended in a liquid medium. After the preparation of the nanowires, nanopillars, and biomolecular motors, each component is introduced and assembled sequentially, i.e., biomolecular motors on nanopillars and then

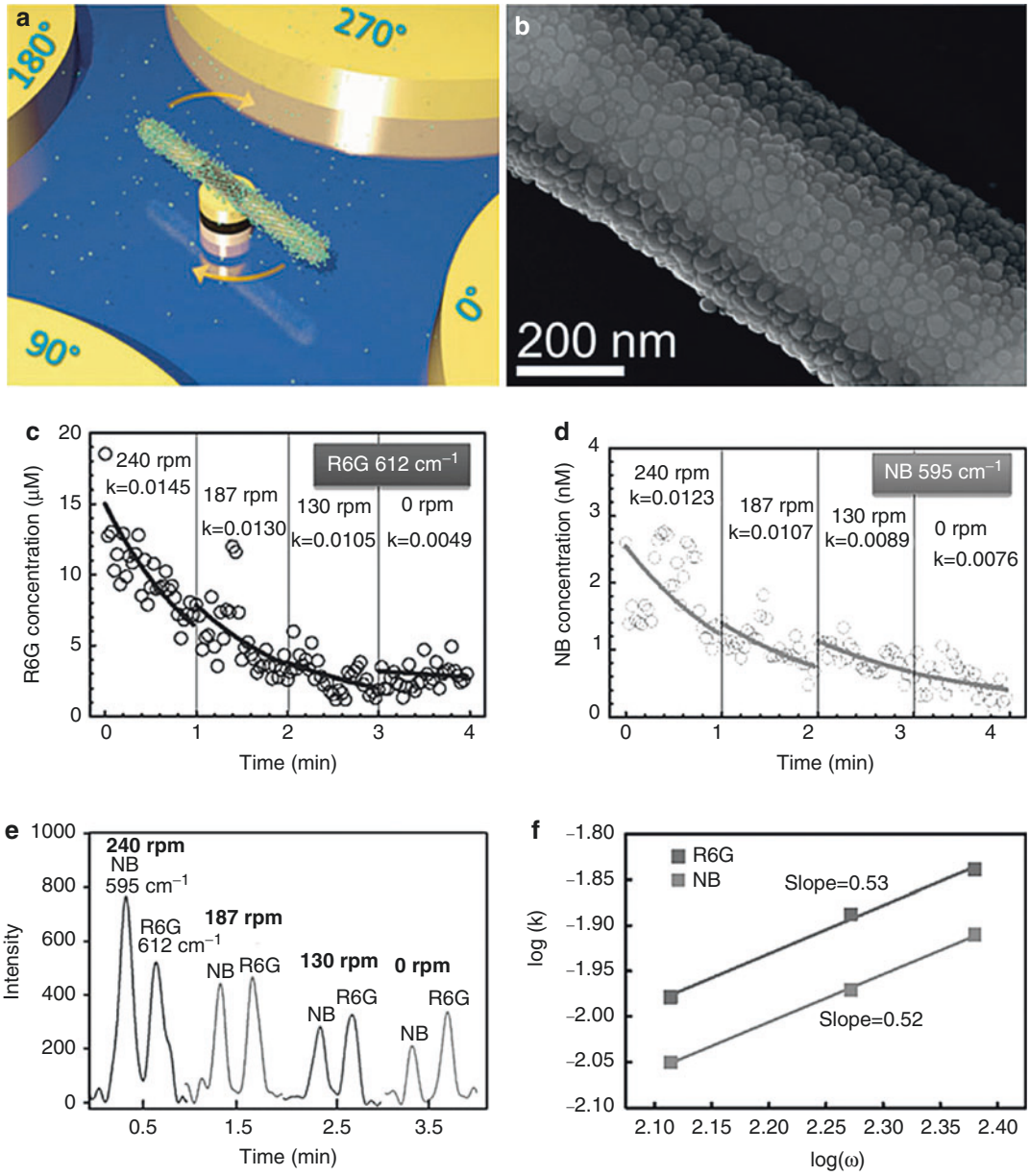


Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices, Fig. 5 Organic-inorganic hybrid nanomotors. From ref [5]. Reprinted with permission from AAAS. (**a** and **c**) SEM images of (**a**) Ni nanopillars and (**c**) nanorods. (**b** and **d**) Illustration of (**b**) a biomolecular motor and (**d**) an assembled hybrid device

nanowire rotors to the tips of biomolecular motors. They are randomly attached to each other via bioconjugation and chemical bonding and form biomolecular motors with Ni nanorod rotors on Ni pillars (Fig. 5). With the introduction of ATP to the assembled devices, the biomolecular motors drive the Ni nanorod rotors at 0.74–8.3 revolutions per second. Relying on the complex bioconjugation and random positioning of nanoentities, it has a low device yield and difficult site-specific assembling. Besides, the rotation direction and speed of the nanomotors are arduous to control because of the nature of biomolecular motors. However, the hybrid nanomotors introduced innovative mechanisms for assembling and powering rotary nanomachines, and refined device fabrication processes and improved controllability of device behavior would make them more inspiring.

Applications

Rotary NEMS devices, assembled by the electric tweezers and magnetic interaction between nanoentities, have been used for precisely tunable biochemical release with their highly controllable assembling and actuation characteristics [12, 21]. The nanowire rotors are densely coated with surface-enhanced-Raman-scattering (SERS)-sensitive Ag nanoparticles, and then biochemicals, such as pure Nile blue (NB) and multiplex of NB and Rhodamine 6G (R6G) molecules, are loaded on the surface of the nanowires (Fig. 6a, b). While a nanomotor is releasing the loaded biochemicals to the surrounding medium, time-dependent release rate is determined via Raman spectroscopy on the nanomotor, showing the remaining amount of the molecules on the surface of the nanowire rotor. The release rate (k)



Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices, Fig. 6 Tunable biochemical release. Reprinted with permission from ref [21]. Copyright 2015 Wiley-VCH. (**a**) Illustration of biochemical release from a nanomotor with a SERS-sensitive nanowire rotor. (**b**) SEM image of a nanowire rotor densely covered with Ag nanoparticles. (**c**–**f**) Characterization of biochemical release of R6G and NB controlled by the rotation. The release rate k linearly increases with the square root of the rotation speed ω

is linearly dependent on the square root of the rotation speed ($\sqrt{\omega}$) for both the single and multiplex molecule release (Fig. 6c–f). It could be explained by the fact that the thickness of the diffusion layer in the vicinity of the nanowire rotors decreases as a result of higher flow speed due to the rotation of nanomotors and this could expedite the whole biochemical release process

including their diffusion from the surface. This molecule release mechanism is applicable to any biochemical of interest in addition to the Raman-sensitive molecules, including drugs, cytokine, DNA, antigens, and antibodies.

Cross-References

- ▶ AC Electrokinetics of Nanoparticles
- ▶ Basic MEMS Actuators
- ▶ Dielectrophoresis
- ▶ Magnetic-Field-Based Self-Assembly

References

1. Tai, Y.-C., Muller, R.S.: IC-processed electrostatic synchronous micromotors. Sens. Actuators **20**, 49–55 (1989)
2. Fennimore, A.M., Yuzvinsky, T.D., Han, W.-Q., Fuhrer, M.S., Cumings, J., Zettl, A.: Rotational actuators based on carbon nanotubes. Nature **424**, 408–410 (2003)
3. Liu, M., Zentgraf, T., Liu, Y., Bartal, G., Zhang, X.: Light-driven nanoscale plasmonic motors. Nat. Nanotechnol. **5**, 570–573 (2010)
4. Xia, Y., Yang, P., Sun, Y., Wu, Y., Mayers, B., Gates, B., Yin, Y., Kim, F., Yan, H.: One-dimensional nanostructures: synthesis, characterization, and applications. Adv. Mater. **15**, 353–389 (2003)
5. Soong, R.K., Bachand, G.D., Neves, H.P., Olkhovets, A.G., Craighead, H.G., Montemagno, C.D.: Powering an inorganic nanodevice with a biomolecular motor. Science **290**, 1555–1558 (2000)
6. Wagner, R.S., Ellis, W.C.: Vapor-liquid-solid mechanism of single crystal growth. Appl. Phys. Lett. **4**, 89–90 (1964)
7. Ning, J., Xu, X., Liu, C., Fan, D.L.: Three-dimensional multilevel porous thin graphite nanosuperstructures for $Ni(OH)_2$-based energy storage devices. J. Mater. Chem. A **2**, 15768–15773 (2014)
8. Whitney, T.M., Searson, P.C., Jiang, J.S., Chien, C.L.: Fabrication and magnetic properties of arrays of metallic nanowires. Science **261**, 1316–1319 (1993)
9. Guix, M., Mayorga-Martinez, C.C., Merkoçi, A.: Nano/micromotors in (bio)chemical science applications. Chem. Rev. **114**, 6285–6322 (2014)
10. Gao, J., Gu, H., Xu, B.: Multifunctional magnetic nanoparticles: design, synthesis, and biomedical applications. Acc. Chem. Res. **42**, 1097–1107 (2009)
11. Sun, L., Hao, Y., Chien, C.L., Searson, P.C.: Tuning the properties of magnetic nanowires. IBM J. Res. Dev. **49**, 79–102 (2005)
12. Kim, K., Xu, X., Guo, J., Fan, D.L.: Ultrahigh-speed rotating nanoelectromechanical system devices assembled from nanoscale building blocks. Nat. Commun. **5**, 3632 (2014)
13. Fan, D.L., Cammarata, R.C., Chien, C.L.: Precision transport and assembling of nanowires in suspension by electric fields. Appl. Phys. Lett. **92**, 093115 (2008)
14. Fan, D.L., Zhu, F.Q., Cammarata, R.C., Chien, C.L.: Electric tweezers. Nano Today **6**, 339–354 (2011)
15. Jones, T.B.: Electromechanics of particles. Cambridge University Press, Cambridge/New York (1995)
16. Fan, D., Zhu, F., Cammarata, R., Chien, C.: Controllable high-speed rotation of nanowires. Phys. Rev. Lett. **94**, 247208 (2005)
17. Guo, J., Kim, K., Lei, K.W., Fan, D.L.: Ultra-durable rotary micromotors assembled from nanoentities by electric fields, Nanoscale, doi:10.1039/C5NR02347E (2015)
18. Kim, K., Guo, J., Xu, X., Fan, D.E.: Micromotors with step-motor characteristics by controlled magnetic interactions among assembled components. ACS Nano **9**, 548–554 (2015)
19. Kim, K., Zhu, F.Q., Fan, D.: Innovative mechanisms for precision assembly and actuation of arrays of nanowire oscillators. ACS Nano **7**, 3476–3483 (2013)
20. Schliwa, M., Woehlke, G.: Molecular motors. Nature **422**, 759–765 (2003)
21. Xu, X., Kim, K., Fan, D.: Tunable release of multiplex biochemicals by plasmonically active rotary nanomotors. Angew. Chem. Int. Ed. **54**, 2525–2529 (2015)

Mechanisms of Laser Ablation Synthesis of Nanoparticles in Liquid Solution

▶ Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids

Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids

J. J. Naddeo and D. M. Bubb
Department of Physics, Rutgers University – Camden, Camden, NJ, USA

Synonyms

Laser Interactions with Particles in Solution; Mechanisms of Laser Ablation Synthesis of

Nanoparticles in Liquid Solution; Role of Cavitation Bubble Dynamics on Nanoparticle Generation; Role of Plasma Confinement on Nanoparticle Generation

Definition

There are various complex mechanisms present in laser ablation of metallic targets in a liquid medium (PLAL). Many of these mechanisms play a key role in nanoparticle synthesis namely, plasma confinement by the liquid medium and cavitation bubble dynamics. Laser interactions with nanoparticles in colloidal solutions, may also affect particle size and composition.

Nanoparticle Synthesis by Laser Ablation in Liquid

Laser ablation describes a process by which material is removed from a surface by the application of a high-intensity laser beam. Pulsed laser ablation in a liquid medium (PLAL) can be a method used to produce a clean, stable, and relatively homogeneous dispersion of nanoparticles in a liquid medium. This process has been carried out for metals [1], semiconductors [2], and even organic materials [3] to produce colloidal solutions. One remarkable feature of this process is the very different conditions under which these solutions are formed.

Several parameters can be tuned for the laser (pulse duration, wavelength, energy, beam waist, etc.) and for the liquid (depth, composition, temperature). One of the most important quantities that can be specified is the intensity of the incident laser beam. The intensity parameter is given by

$$I = \frac{E_p}{A \cdot \tau_p}, \tag{1}$$

where E_p is the energy contained in the laser pulse, A is the area of the laser beam on the target surface, and τ_p is the duration of the laser pulse. Although this field encompasses studies that utilize continuous wave (*CW*) to *attosecond* (10^{-18} s) lasers, we will focus our attention on the pulse regime between 10 *fs* and 10 *ns*, as it it most relevant to laser synthesis of nanoparticles. For many laser systems, pulse duration is the main determinant of the intensity output. For purposes of this article, we will define ultrashort laser pulses as $\tau_p < 1\ ps$ duration and long pulses as $\tau_p > 1\ ns$. Short-pulse lasers will be those for which $1\ ns > \tau_p > 1\ ps$.

A typical experimental setup for laser ablation involves a laser beam being focused onto a metal, oxide, polymer, or semiconductor surface with concomitant ejection of material as the result of strong explosive vaporization, spallation, or in some cases photoablation. Laser ablation is very challenging and computationally expensive to model; nonetheless, much insight has been gained in recent years from the application of the two-temperature model and molecular dynamics [4].

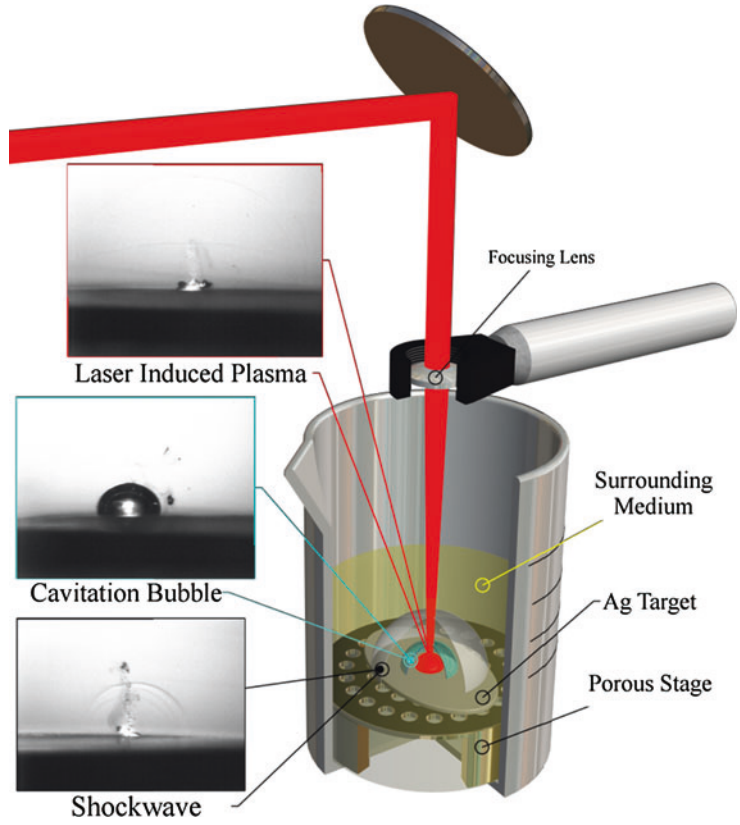
In the case of PLAL, all of the complexity of laser ablation is present, with the addition of hydrodynamic and thermodynamic processes in the liquid medium. Here the atoms and small clusters that are ejected from the surface undergo many more collisions than they would for laser ablation in air or vacuum as the result of the much smaller mean free path for the liquid. The liquid medium also induces both inertial and plasma confinement and strong cavitation near the surface. These processes strongly influence the observed particle size and distribution that is obtained.

Finally, PLAL process depicted in Fig. 1 can be modified to proceed without the solid target. In this case, the target is understood to be a colloidal liquid. This process of postproduction laser irradiation (PIR) is used to homogenize and sometimes reduce the size distribution of the nanoparticles. Typically, this is referred to as laser-induced size reduction, although this usage is not universal.

Confinement, Cavitation, and Nucleation

Plasma Confinement

Once the laser energy has been deposited in the target medium, the system relaxes from a highly nonequilibrium state. The particular relaxation mechanisms depend on the laser wavelength and



Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids, Fig. 1 Schematic of laser ablation in liquids experiment

its relationship with the target's optical constants, as well as the presence of nonlinear effects at sufficient peak intensities of the incident beam. Shorter-wavelength lasers (i.e., UV and near-UV) possess enough energy to directly break bonds, and the mechanism of nanoparticle formation is dominated by photoablation. Near-IR lasers are more efficient at igniting plasma close to the target surface through inverse Bremsstrahlung (IB) absorption. The cross sections for ion-electron and atom-electron IB absorption are given as a function of laser frequency (v):

$$\text{Ion} - \text{electron: } \sigma_{IB}(v) = \frac{3.7 \times 10^{18}}{\frac{1}{T_e^2} v^3} Z^2 \left[e^{\frac{hv}{k_B T_e}} - 1 \right] n_e n_i \tag{2}$$

$$\text{Atom} - \text{electron: } \sigma_{IB}(v) = \frac{e^2}{\pi m_e c v^2} n_e n_0 \sigma_{\text{col}} \left(\frac{8 k_B T_e}{\pi m_e} \right) \tag{3}$$

where T_e is the electronic temperature, h is Planck's constant, Z is the ionic charge, c is the speed of light in vacuum, σ_{col} is the cross section for neutral-atom collisions, m_e is the electronic mass, and n_e, n_i, and n_0 are the number density of electrons, ions, and atoms, respectively. Lasers with relatively longer pulse durations allow for the ignited plasma and late photons to temporally overlap, thus enhancing IB absorption. The plasma will rapidly expand into the surrounding liquid medium and the quenching process produce spherical nanoparticles. A fraction of these nanoparticles will be at a sufficient temperature such that they sinter together into larger aggregates, giving rise to a polymodal size distribution. Previously produced nanoparticles may interact with subsequent pulses and undergo laser-induced melting and Coulomb explosion, which tends to reduce the size. This effect of postirradiation is more intense for larger particles. For example, the absorption coefficient per unit volume at the

plasmon peak of a 4 nm *Au* NP is one-fourth the absorption coefficient per unit volume at the plasmon peak of a 30 nm *Au* NP [5]. This selective absorption will tend to fragment larger particles, while smaller particles do not absorb enough energy to explode or melt, thus homogenizing the colloidal nanoparticle solution.

Cavitation Dynamics

Approximately 100 ns -1 μ*s* after the arrival of the laser pulse, rapid heating of the target surface induces the growth of a cavitation bubble which undergoes rapid oscillation, typically for several hundred microseconds. Cavitation is typically modeled by the Rayleigh-Plesset equation, which is given as

$$\frac{P_B(t)-P(t)}{\rho_L}=R(t)\frac{d^2R(t)}{dt^2}+\frac{3}{2}\left[\frac{dR(t)}{dt}\right]^2+\frac{4v}{R(t)}+\frac{2S}{\rho_L R(t)}, \quad (4)$$

where $R(t)$ is the bubble radius as a function of time, t, after the arrival of the laser pulse, $P_B(t)$ and $P(t)$ are the pressures in the interior and exterior of the bubble, respectively, the surface tension of the bubble is S, and the kinematical viscosity is given by v. Finally, the density of the liquid medium is ρ_L.

The dynamics of this cavitation bubble on the surface of the target appear to be well described by the Rayleigh-Plesset equation for *ps* and faster pulses. For long-pulse ablation (i.e., *ns*), the thermodynamic parameters need to be modified to account for the presence of nanoparticles and ejected material [6]. The role that the bubble plays in the growth and nucleation of nanoparticles is less clear, although some reports indicate that the primary particles form during bubble expansion, while the aggregates form during collapse [7], although more recent work [8] has shown that the role of the bubble is still not completely understood. Nonetheless, cavitation plays a fundamental role in nanoparticle formation by laser ablation in liquids, whether it is through influencing the nucleation kinetics or inducing tertiary material ejection during the collapse phase.

Nucleation

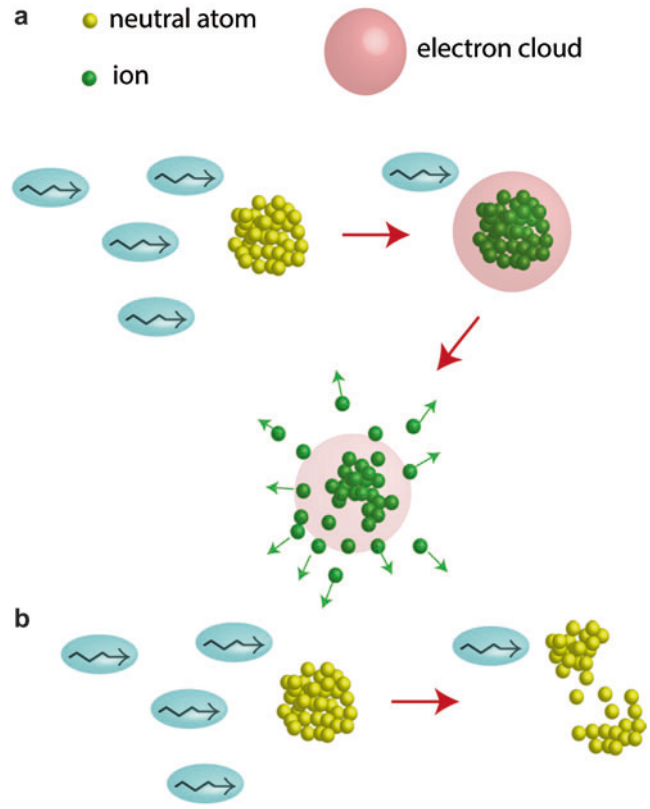
Classical nucleation theory is often invoked to explain the growth of nanoparticles in solution. So-called Lifshitz-Slyozov-Wagner (LSW) theory models this process as a diffusion-limited reaction with a rate constant given by $k=\alpha\frac{T}{\eta}$ where α is a proportionality constant, T is the solution temperature, and η is the dynamic viscosity. Then the particle radius is given as a function of time (t) as

$$R^3(t)=R_0^3+kt. \quad (5)$$

Of course, the application of any equilibrium theory to a dynamic process such as laser ablation in liquids is expected to be largely qualitative, and studies have shown that the dependence of particle volume on viscosity does not scale directly as $\frac{1}{\eta}$ [9]. Actually, for gold nanoparticles, it is found that the compressibility of the the water medium is more correlated with the hydrodynamic radius than the viscosity is. Nonetheless, LSW theory provides useful guidance for considering what occurs in solution for times long after the arrival of the laser pulse. As discussed in section "Plasma Confinement," the presence of a large number of particles in the solution can interact with subsequent laser pulses, thus leading to an increased reaction rate and larger particles.

Laser Interactions with Particles in Solution

As briefly discussed earlier, laser irradiation of nanoparticles may be performed purposefully as a means of size reduction and homogenization, or it can occur during the synthesis of nanoparticles in solution. In the latter case, this process is often viewed as a nuisance, or at least an interferent; however, in the former case such interactions can be beneficial under the right circumstances. Incident laser radiation may be absorbed by the nanoparticles through either the excitation of interband transitions or, in the case of metal and semiconductor nanoparticles, excitation of a surface plasmon resonance (SPR). With short and ultrashort laser pulses, multiphoton absorption

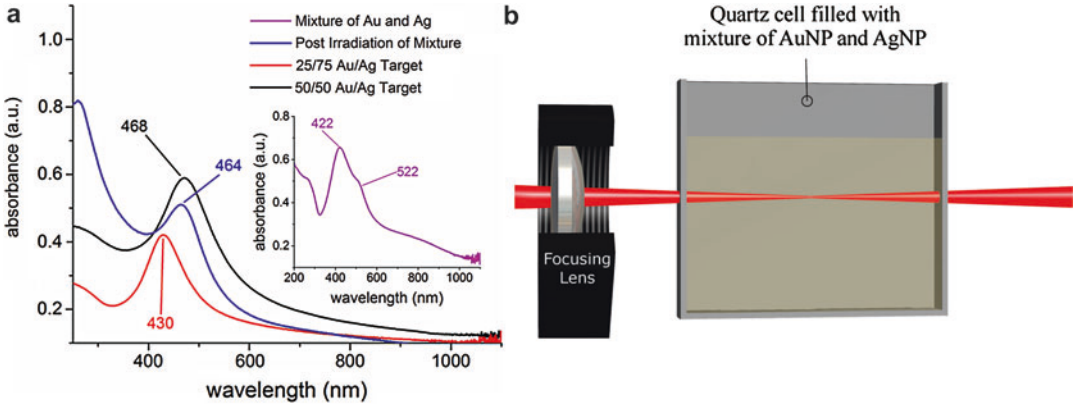


Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids, Fig. 2 (**a**) Representation of Coulomb explosion: First, the neutral nanoparticle is irradiated with a high flux of photons that are capable of ionizing individual atoms. This produces ions and a delocalized electron cloud. The ions are forcefully ejected from the core due to Coulomb repulsion. Note that in the last step the electron cloud is more faint due to interactions with the surroundings and other nanoparticles (**b**) Laser melting: absorption of photons leads to the particle exceeding its melting temperature, and it subsequently fragments

can also occur at sufficient pulse energies. It is currently believed that fragmentation of nanoparticles can take place via either thermal evaporation or Coulomb explosion – see Fig. 2. Upon absorption of the laser light, the temperature of the nanoparticle can increase to above the boiling point, thus resulting in evaporation of the outer layers of the particle. In fact, the melting of *Au* nanoparticles during pulsed irradiation has been confirmed by X-ray scattering where the loss of long-range order occurred ~ 1 *ps* after the arrival of the laser pulse [10]. Additionally, gold nanorods were converted into nanospheres by pulsed irradiation that induced melting and a subsequent Rayleigh instability [11]. This melting and rise in temperature can be accompanied by strong thermionic emission which leads to a loss of stability in the highly charged particles. The strong repulsive forces lead to fragmentation of the particle by Coulomb explosion into smaller and more stable units. It is also possible to have direct photoionization of particles leading to Coulomb explosion when very high-intensity laser pulses are used.

Nanoalloys and Nanocomposites

Composite and alloy nanoparticles can be produced by laser ablation in liquids in a number of ways. First, two colloids may be mixed together (pink line in below spectra) and irradiated to alloy the colloids (blue line in below spectra). The diagram in Fig. 3b shows a typical experimental setup used to postirradiate a nanoparticle mixture. Second, an alloy target can be used in the PLAL setup depicted in Fig. 1, producing a colloidal alloy nanoparticle solution. When followed by postirradiation, the particles are simultaneously homogenized and further alloyed. Third, a metal target can be ablated in the presence of a colloidal NP solution of the desired alloy partner, again followed by postirradiation. These procedures can easily be performed with miscible systems such as $Ag - Au$, where it is fairly easy to tune the surface plasmon resonance (SPR) over the



Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids, Fig. 3 (**a**) Extinction spectra for various alloying methods. When the target is composed of equal amounts of *Ag* and *Au* the SPR is located at $\lambda = 468$ nm; this SPR shifts to $\lambda = 430$ nm when the target is composed of mostly *Ag*. (**b**) The diagram to the right illustrates an experimental setup where a mixture of *Ag* and *Au* nanoparticles is postirradiated to induce alloying

range of $\lambda = 400$–520 nm. When the second procedure is used, it can be seen in Fig. 3a that a target with a higher *Ag* mol fraction will result in SPR closer to that of *Ag* ($\lambda = 400$ nm) (red line vs. black line).

Experimentally the former two methods discussed above produce the best alloying and are the most reproducible. Alloy nanoparticles are very desirable for nanoparticle plasmonics and surface-enhanced Raman spectroscopy (SERS) applications where one desires to tune the resulting SPR. Many other alloys and nanocomposites can be created with this technique, but it is not always trivial to characterize the particles, and high-resolution transmission electron microscopy might be necessary to adequately perform the analysis.

Conclusion

This entry focuses on the mechanism by which nanoparticles are generated during laser ablation of solid targets submersed in liquid. The role of plasma confinement and cavitation was discussed with a special emphasis on their influence on the nanoparticle size and polydispersity of the resulting colloids. The effect of laser interactions with nanoparticles was also reviewed in two contexts. First, the influence of temporal overlap between the incident laser pulse and nascent particles is argued to be a mechanism by which laser-generated nanoparticles become bi- or polymodal. Secondly, the irradiation of colloidal solutions is often performed to reduce the polydispersity within nanoparticle solutions, as well as reduce the primary particle size. The two most commonly cited mechanisms by which this size reduction and homogenization occurs are Coulomb explosion and laser-induced melting. Finally, the role of these irradiation mechanisms is discussed with respect to nanoalloying of the noble metals *Au* and *Ag*.

Cross-References

- Engineered Nanoparticles
- Metal Nanoparticles from First Principles
- Nano-sized Particle
- Nanoalloy Simulation
- Nanomaterials
- Nanoparticle
- Nanoparticles

References

1. Fojtik, A., Henglein, A.: Laser ablation of films and suspended particles in a solvent – formation of cluster and colloid solutions. Ber. Bunsen Ges. Phys. Chem. Chem. Phys. **97**(2), 252–254 (1993)

2. Simakin, A.V., Voronov, V.V., Shafeev, G.V.: Nanoparticle formation during laser ablation of solids in liquids. Phys. Wave Phenom. **15**(4), 218–240 (2007)
3. O'Malley, S.M., Mitesh, A., Borchert, J., Jimenez, R., Bubb, D.M.: Formation of rubrene nanocrystals by laser ablation in liquids utilizing MAPLE-deposited thin films. Chem. Phys. Lett. **595–596**, 171–174 (2014)
4. Leveugle, E., Ivanov, D.S., Zhigilei, L.V.: Photomechanical spallation of molecular and metal targets: molecular dynamics study. Appl. Phys. A **79**, 1643–1655 (2004)
5. Amendola, V., Meneghetti, M.: Size evaluation of gold nanoparticles by UV-vis spectroscopy. J. Phys. Chem. C **113**, 4277–4285 (2009)
6. Soliman, W., Nakano, T., Takada, N., Sasaki, K.: Modification of Rayleigh-Plesset theory for reproducing dynamics of cavitation bubbles in liquid-phase laser ablation. Jpn. J. Appl. Phys. **49**(11), 116202 (2010)
7. Wagener, P., Ibrahimkutty, S., Menzel, A., Plech, A., Barcikowski, S.: Dynamics of silver nanoparticle formation and agglomeration inside the cavitation bubble after pulsed laser ablation in liquid. Phys. Chem. Chem. Phys. **15**(9), 3068–3074 (2013)
8. Tomko, J., Naddeo, J.J., Jimenez, R., Tan, Y., Steiner, M., Fitz-Gerald, J.M., O'Malley, S.M., Bubb, D.M.: Size and polydispersity trends found in gold nanoparticles synthesized by laser ablation in liquids. Phys. Chem. Chem. Phys. **17**, 16327–16333 (2015)
9. Menendez-Manjon, A., Chichkov, B., Barcikowski, S.: Influence of water temperature on the hydrodynamic diameter of gold nanoparticles from laser ablation. J. Phys. Chem. C **114**, 2499–2504 (2010)
10. Plech, A., Kotaidis, V., Grésillon, S., Dahmen, C., von Plessen, G.: Laser-induced heating and melting of gold nanoparticles studied by time-resolved X-ray scattering. Phys. Rev. B **70**(19), 195423 (2004)
11. Zijlstra, P., James, W.M., Chon, M.G.: White light scattering spectroscopy and electron microscopy of laser induced melting in single gold nanorods. Phys. Chem. Chem. Phys. **11**(28), 5915–5921 (2009)

Mechanoreceptors

► Arthropod Strain Sensors

Medical adhesives

► Bioadhesives

Medical Nanorobotics

► Nanorobotics for Bioengineering and Medical Interventions

MEM Devices

► Graphene Nanostructures for Memristive Devices

Memory Resistors

► Graphene Nanostructures for Memristive Devices

Memristors

► Graphene Nanostructures for Memristive Devices

MEMS

► DNA Manipulation Based on Nanotweezers
► Nanotechnology

MEMS (Micro-Electro-Mechanical Systems)

► CMOS MEMS Fabrication Technologies

MEMS = Microelectromechanical Systems

► MEMS on Flexible Substrates

MEMS Adaptive Optics

Nathan Doble
The New England College of Optometry, Boston, MA, USA

Definition

MEMS: Micro-electro-mechanical Systems

MEMS are small silicon-based mechanical devices that employ semiconductor fabrication techniques in their construction. MEMS devices may take various forms: from pressure, chemical, and vibration sensors, to optical switches and mirrors through to inkjet printer heads. MOEMS – Micro-Opto-Electromechanical Systems – are a subset of MEMS devices that in addition employ optical components; an example could be a photon detector mounted on a silicon chip.

Note: MEMS/MOEMS devices have structures that can be measured at the micrometer level, whereas nanotechnology deals with scales at the level of an individual molecule (generally taken to be in the 1–100 nm range).

Adaptive Optics

Adaptive Optics (AO) are optical systems that can both measure and correct optical wavefront aberrations in real time. Real time depends on the particular AO application: in astronomy closed loop bandwidths of several hundred Hertz are required to correct the rapidly changing atmospheric turbulence. For vision science, changes in the lens/cornea and head/eye movements require a slower correction speeds (a few Hertz). Typically, an AO system has three main system components: (1) a wavefront sensor (WFS) that measures the aberration (note: there are also wavefront sensor-less AO systems), (2) a wavefront corrector – typically a deformable mirror (DM), and (3) a control computer that controls the WFS and wavefront corrector.

MEMS Adaptive Optics

An AO system that employs MEMS technology in the wavefront corrector

Introduction

Adaptive optics (AO) is a technique that measures and corrects wavefront distortions. These distortions may be due to effects of atmospheric turbulence (astronomy or free space communications); however, they may also be biological in nature (vision science or microscopy).

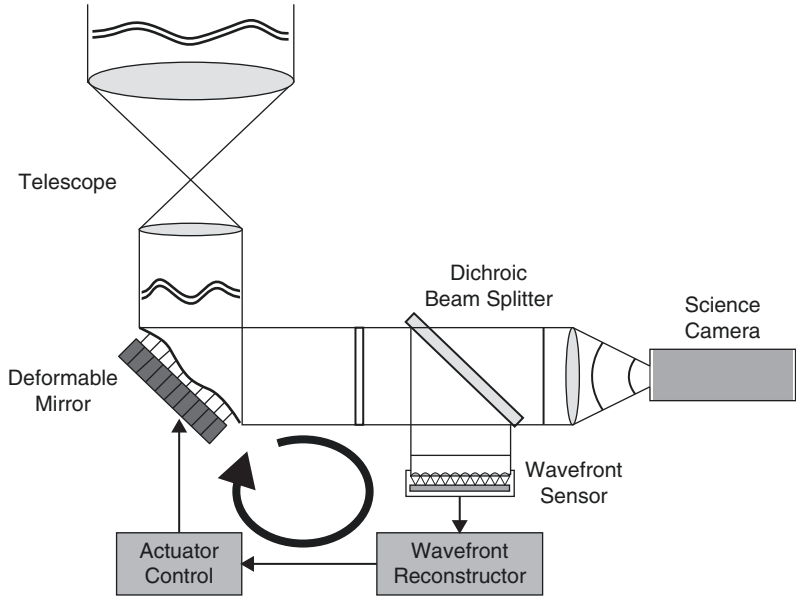
AO was first proposed by Babcock in 1953 for use in large, ground-based telescopes but implementation at that time was limited by the available technology. The initial development beginning in the 1970s was for military applications. Declassification in the early 1980s led to significant use in astronomy with all large terrestrial telescopes now using AO to enhance their image quality. Other scientific fields have benefited greatly from this work particularly in the fields of vision science and microscopy, both of which have leveraged the technology and know-how to provide AO corrected images with remarkable clarity.

This entry focuses on recent developments in AO that use micro-electro-mechanical systems (MEMS) technology. The technology itself and the results in a range of research fields are described. A detailed discussion of AO principles is, however, beyond the scope of this entry and the interested reader is referred to the reference texts [1–3].

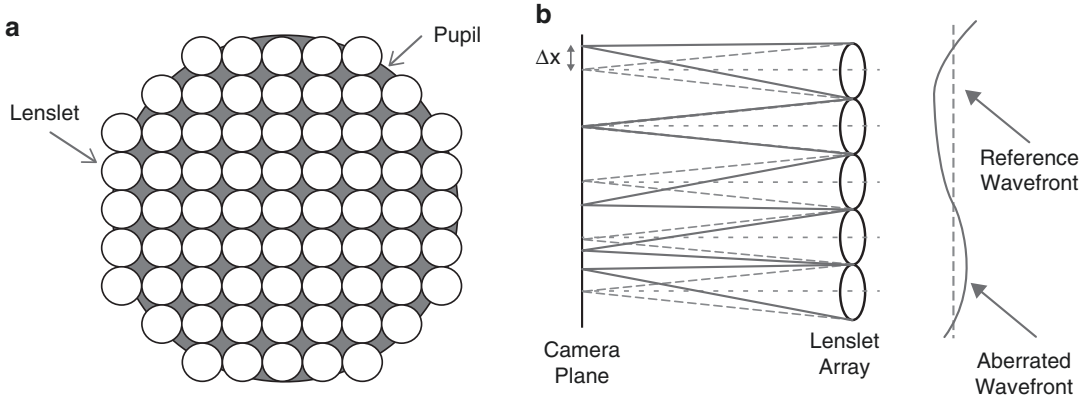
Principle and Application of Adaptive Optics (AO)

A schematic AO layout is shown in Fig. 1 and generally requires the implementation of three main subsystems:

The Wavefront Sensor (*WFS*) – This measures the optical aberrations in the incoming distorted wavefront. There are two broad WFS classes [1, 2]: (1) *direct* – in which the phase or optical path difference is measured, e.g., Hartmann-Shack (Fig. 2), curvature, and pyramid sensing, (2) *indirect* – the wavefront properties are not explicitly measured, but rather the correction signals are deduced from whole-aperture intensity measurements made at or near the focal plane. These are generally iterative in nature with



MEMS Adaptive Optics, Fig. 1 The basic AO operating principle. In this case, the distorted wavefront from a telescope is reflected off the deformable mirror (*DM*). A fraction of the reflected light is redirected to the wavefront sensor (*WFS*) where the aberrations are measured and the appropriate control signals are determined for the DM. This continual measurement and correction of the wavefront distortion is termed closed loop, the speed of which depends on the particular application. The majority of the light is used in forming the image at the science camera, the aim being to obtain planar wavefronts that focus down close to the limits imposed by diffraction. Note: there are wavefront sensor-less AO approaches that use a measure of the light intensity at the science camera to drive the AO loop and which do not require a separate WFS arm (Reproduced from Expert Review of Medical Devices, 2, 205–216 (2005) with permission of Expert Reviews Ltd)



MEMS Adaptive Optics, Fig. 2 Principle of the Hartmann–Shack WFS. (**a**) The incident pupil (*gray circle*) is sampled by a two-dimensional array of lenslets, the diameter and focal length of which are chosen to give the required measurement resolution and dynamic range. (**b**) For the reference wavefront (*gray hatched line*), a focal spot is formed by each lenslet on the camera plane and is exactly centered behind each lenslet. For an aberrated wavefront (*solid gray line*), variations in the slope between each lenslet will cause the focal spots to be shifted from the position of the reference. By determining both the Δx and Δy shifts for each lenslet, the aberrated wavefront may be reconstructed

examples being image sharpening and phase diversity.

The Control Computer – This takes the output of the WFS and converts it to voltage commands that are sent to the wavefront corrector.

The Wavefront Corrector – This device provides the conjugate correction to the measured optical distortions. The aim of any wavefront corrector is to equalize the optical path length (OPL) seen by all points on the incident wavefront, resulting in a flat wavefront after correction. The OPL is defined as the product of the refractive index, n, and the physical path length, d. Wavefront correctors can be divided into two main classes depending on how they impart the OPL change: (1) *deformable mirrors (DMs)* – these are by far the most common device used. As the name suggests these operate by locally deforming a mirrored surface, i.e., they change the physical path length, d, and (2) *liquid crystal spatial light modulators (LC-SLMs)* – these produce localized changes in their refractive index, n (d is fixed). The refractive index change is achieved through the application of a voltage that rotates the liquid crystal molecules.

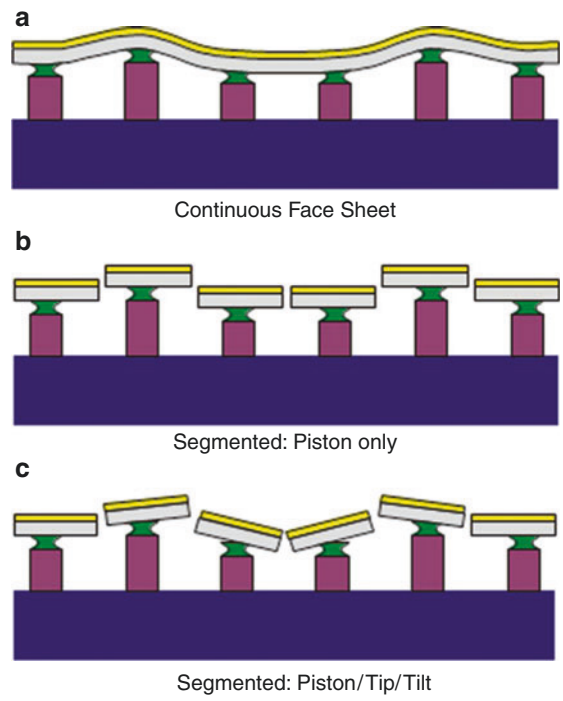
Consider Fig. 1 and the example of aberrations entering a telescope; the incident distorted wavefront is reflected by the DM with a fraction of the reflected light being redirected into the WFS arm where the aberrations are measured and the appropriate control signals are determined for the DM. The DM surface profile is then updated to exactly cancel the aberrations in the wavefront. Both the DM and WFS are placed in optical planes that are conjugate to the source of the aberration, for example, in the human eye this plane is taken to be somewhere between the cornea and lens. The majority of the light is used in forming the image at the science camera, the aim being to obtain planar wavefronts that focus down close to the limits imposed by diffraction. This continual measurement and correction of the wavefront distortion is termed closed loop control.

Within AO, MEMS technology has been used extensively in the construction of DMs and the next section details their design. Particular emphasis on the use of MEMS AO is given over to vision science as it was in this field that such technology was first used. Recently, MEMS AO has increased its breadth and the newer applications and results are described.

MEMS Deformable Mirrors

MEMS has seen considerable research interest in the development of DMs and is a field of active research and development. The basic principle of a DM is simple; however, fabricating them with the required operating parameters has proven to be extremely challenging. Figure 3 shows the three basic DM types in cross section – light would be incident from the top of the figure. Continuous face sheet (Fig. 3a) – a continuous reflective mirror surface is deformed by an underlying two-dimensional array of actuators. Piston-only DMs (Fig. 3b) are composed of a discrete array of segments, each of which may individually move in a pure piston (vertical) motion only. Figure 3c shows a piston/tip/tilt DM, an individual segment may now also move in tip and tilt in addition to piston. All the DM types in Fig. 3 show piston-like actuators under the mirror surface(s) and an example actuation mechanism could be piezoelectric. However, there are other forms of actuation, for example, electrostatic, magnetic, thermal, or voice coil. MEMS-based DMs generally use electrostatic attraction; however, other forms have been employed.

The first AO systems had physically large (>10 cm diameter), expensive (>$100 K), non-MEMS DMs that had been primarily designed for the military/astronomy communities. The DM was required to provide a modest amount of stroke (~5 μm) with quick response times (several kHz). In biomedical imaging the requirements are much different, (1) the DM diameter needs to be small (a few mm in diameter), hence minimizing the pupil magnification and allowing for the design of compact systems, (2) the DM must also be inexpensive particularly if AO is to transition into commercial products, and (3) these applications require slower response times (a few Hz), fewer actuators (or segments), but much more stroke (>50 μm) [4].



MEMS Adaptive Optics, Fig. 3 Deformable mirror (DM) types. (**a**) A continuous mirror surface is deformed by an underlying array of actuators. Many forms of actuation are possible, for example, piezoelectric, electrostatic, or magnetic. (**b**) The mirror surface is composed of a discrete array of segments each of which may move in a pure piston motion only. (**c**) Similar to (**b**) but each segment may now also move in tip and tilt in addition to piston

Structure of MEMS DMs

The first MEMS mirror was the Texas Instruments digital micro-mirror device (DMD) introduced in the late 1990s which comprises up to one million mirror segments. It is used extensively in high-quality projection systems with each mirror segment representing one pixel in the image. However, it is bistable in nature with a low fill factor and a surface profile which precludes its use in AO. Around the same time, MEMS-based DMs designed specifically for AO applications were being introduced and they can be divided into two broad classes: (1) bulk micromachined and (2) surface micromachined devices, the difference being in the way the mirrors are fabricated and assembled. The first MEMS DMs were bulk micromachined [5]; a few years later saw the introduction of the first surface micromachined devices [6].

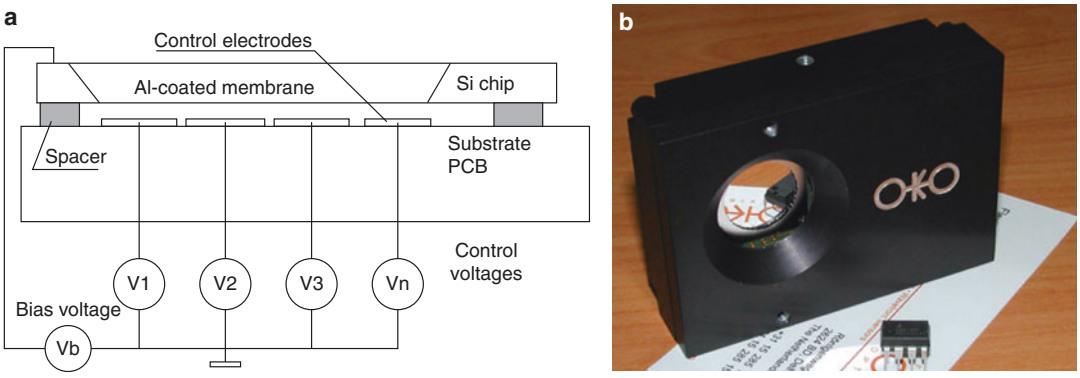
Bulk Micromachined Membrane DMs (MMDM)

In these electrostatic devices, a flexible, reflective membrane lies between a transparent top electrode and a patterned array of bottom electrodes [5]. Application of a voltage to the bottom electrodes causes the mirror surface to deflect downward. The DM has two principal components:

- *A reflective mirrored membrane*: This is made from silicon nitride and an evaporated layer of aluminum or sometimes gold to provide the reflective layer. The membrane stress can be controlled during the nitride deposition.
- *The electrode array*: The bottom set of electrodes are formed by patterning a layer of aluminum deposited onto a silicon wafer.

Wafer bonding is then used to join these two sections; this has a particular advantage as it allows fabrication of devices with a large actuator gap and hence large dynamic range. However, membrane DMs are generally used for applications which require the correction of large, low spatial frequency aberrations as it is difficult to make devices with high actuator packing densities.

A cross section through the OKO Technologies DM [5] is shown in Fig. 4. Up to 9 μm of mirror



MEMS Adaptive Optics, Fig. 4 Structure of the OKO Technologies' bulk micromachined membrane DM. (**a**) A reflective mirrored membrane is positioned above a patterned set of bottom electrodes (V_1,. . .,V_n). A voltage difference between these electrodes and the top transparent grounded electrode causes the DM surface to deflect. (**b**) The packaged DM – up to 9 μm of deflection is possible at the center of the 15 mm DM aperture. The DM has 37 electrodes with operating voltages between 0 and 260 V (DM schematic and photograph provided courtesy of OKO Technologies)

deflection is possible at the center of the 15 mm DM aperture. The DM has 37 electrodes with operating voltages between 0 and 260 V and has been used in a variety of AO applications.

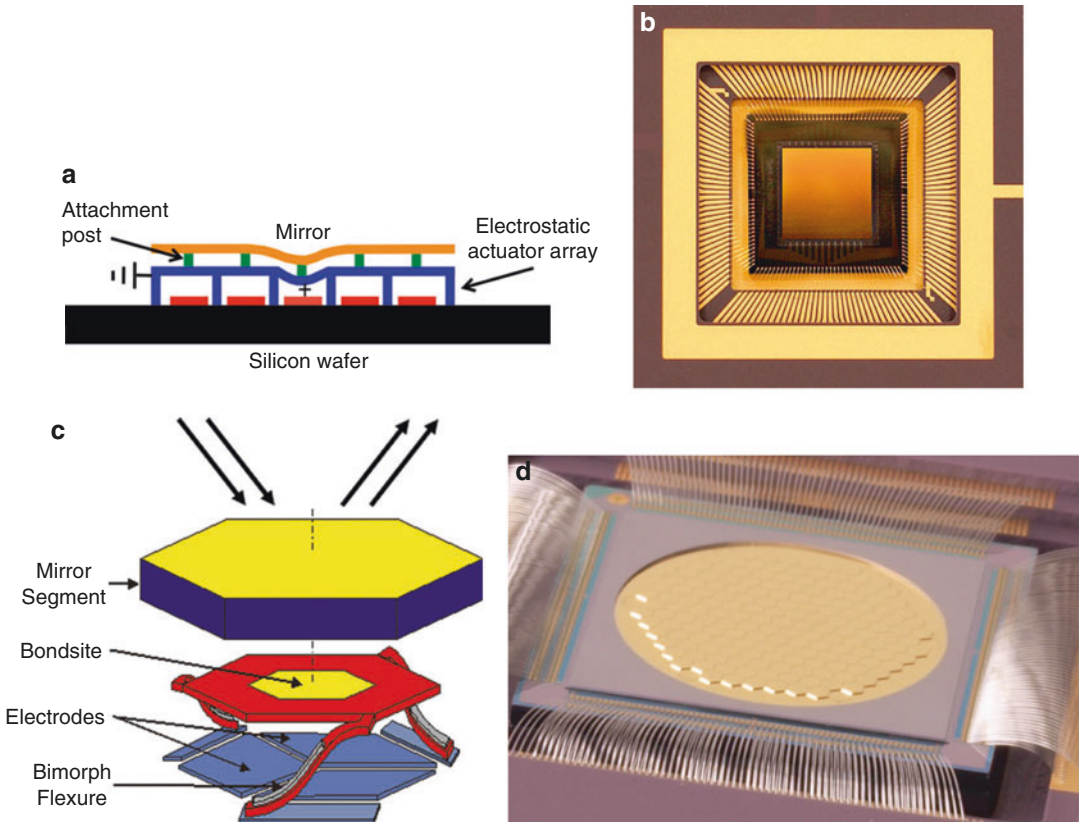
Surface Micromachined Continuous Surface MEMS Mirrors

Another class of MEMS DMs is formed by surface micromachining, in which devices are built up from a series of deposited layers. A layer is deposited, etched before the next is added, much like a layer cake. This construction technique allows for devices with broad functionality. The various layers can be (1) polysilicon – structural and conductive properties, (2) oxides – sacrificial and dielectric, (3) nitride – insulator and dielectric, and (4) various metals – good conductors and optically reflective. With surface micromachining it is more challenging to achieve the necessary dynamic range for some applications; however, devices with extremely high actuator counts are possible.

Instead of a continuous flexible membrane as in the MMDM case, these DMs allow for more localized deformations of the top mirrored surface. This permits the correction of higher order aberrations due to the ability to impart high spatial frequency patterns onto the DM. Figure 5a shows a cross section of a continuous surface MEMS DMs manufactured by Boston Micromachines Corporation (BMC). A top metal-coated mirror surface is attached by silicon posts to a locally anchored, grounded top membrane. Applying individual voltages to the bottom actuator array causes the top mirror surface to deflect. The amount of deflection is proportional to the applied drive voltage. Figure 5b shows a photograph of a gold-coated 140 actuator, 5.5 μm stroke BMC MEMS DM. Another surface micromachined MEMS DM manufactured by Iris AO Inc. [7] consists of a tiled array of hexagonal segments (Fig. 5c). Each segment is raised above a set of three diamond-shaped electrodes via the bimorph flexures. A high-quality mirror surface is bonded to the platform surface which also serves as the ground electrode. Application of voltages to the bottom three electrodes gives tip, tilt, and piston motion on a per segment basis. Figure 5d shows a fully assembled MEMS DM from Iris AO Inc. consisting of a gold-coated tiled array of 163 segments. The thin wires at the perimeter of the chip provide the voltage connections.

AO for Vision Science

A research area where the acceptance and implementation of MEMS devices has been extremely rapid is in vision science. The need for low cost



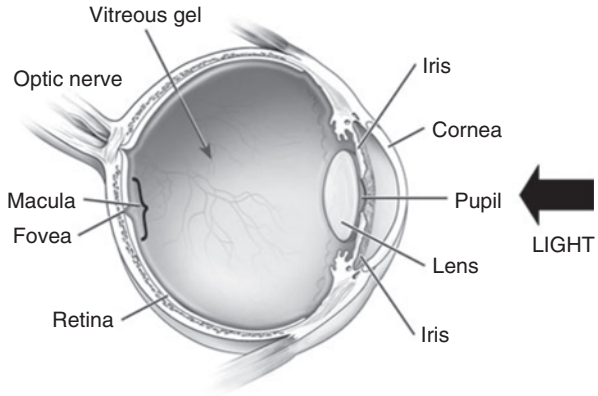
MEMS Adaptive Optics, Fig. 5 Examples of surface micromachined MEMS DMs manufactured by Boston Micromachines Corp. (BMC) and Iris AO Inc. (**a**) Cross section of a continuous surface MEMS DM from BMC. A top metal-coated mirror surface is attached by silicon posts to a locally anchored, grounded top membrane. Applying individual voltages to the bottom actuator array causes the top mirror surface to deflect. The amount of deflection is proportional to the applied drive voltage. (**b**) Photograph of a gold-coated, 140 actuator, 5.5 μm stroke BMC MEMS DM. (**c**) The Iris AO MEMS DM consist of a tiled array of hexagonal segments. Each segment is raised above a set of three diamond-shaped electrodes via the bimorph flexures. A high-quality mirror surface is bonded to the platform surface which also serves as the ground electrode. Application of voltages to the bottom three electrodes gives tip, tilt, and piston motion on a per segment basis. (**d**) A fully assembled MEMS DM from Iris AO Inc. consisting of a gold-coated tiled array of 163 segments (DM schematics and photographs provided courtesy of BMC and Iris AO Inc.)

and physically small devices has made MEMS particularly attractive.

Structure of the Eye

Vision is the primary human sense with over half of the cortex being used to process the signals coming from the visual pathways. Furthermore, the human eye is the only organ allowing a direct in vivo view of the human nervous system. The eye is therefore the subject of intense interest as a means to the early detection of a host of retinal and possibly systemic diseases.

The human eye is a complex optical structure (Fig. 6); incident light is focused by the lens and cornea onto the light-sensitive retinal surface. The primary sources of aberration are the cornea and lens, with the aberration magnitude increasing with larger pupil diameters. Upon reaching the retinal surface, the light traverses many layers before being incident on the light-sensitive photoreceptor cells (rods and cones) where the photons are absorbed. The macula or foveal region is where the cone photoreceptors have the smallest diameter (1.9–3.4 μm) and where one has the



MEMS Adaptive Optics, Fig. 6 The structure of the human eye. The primary sources of aberration are the cornea and lens, the aberration magnitude increases with larger pupil diameters. The incident light propagates to the retinal surface where it is sampled by the rod and cone photoreceptors. The photoreceptors are a few microns in diameter and are unresolvable without the application of AO (Credit: National Eye Institute, National Institutes of Health)

highest visual acuity. The signals from these photoreceptors are then processed by many intervening cell types before exiting the retina via the ganglion cells and the optic nerve. For detailed information of the physiology of the eye, the reader is referred to the references [8, 9].

The ability to visualize the individual retina layers and cells is particularly important as various retinal diseases manifest themselves in different retinal layers. Glaucoma, for example, attacks the retinal ganglion cells leading to a thinning of the nerve fiber layer. Other diseases such as retinitis pigmentosa (RP) and various cone/rod dystrophies manifest themselves in the photoreceptors. Age-related macular degeneration (AMD) first shows in the deeper retinal layers, with thickening of Bruch's membrane and/or drusen between the retinal pigment epithelium (RPE) and the underlying membrane.

Aberrations

Both the lens and cornea introduce low- and high-order spatial aberrations [3], which along with temporal instabilities (accommodation, eye and head movements) limit the ability to obtain high-resolution, in vivo images of retinal cells. These aberrations also limit an individual's visual perception of everyday scenes. Spectacles, contact lenses, or refractive surgery, for example, only provide a low-order, static correction of these aberrations, primarily focus and astigmatism. Theoretically, the lateral (transverse) resolution of the eye is given by Eq. 1:

$$r = 1.22 f\lambda / nD \tag{1}$$

where r is the distance from the center of the Airy disk to the first minima, f is the focal length of the reduced eye, λ is the imaging wavelength, n is the refractive index, and D is the pupil diameter. Hence, the highest resolution would be achieved using the shortest wavelength, λ and the largest possible aperture, D (f being fixed). For the human eye, with $D = 8$ mm, $f = 22.2$ mm, n ~1.33, and imaging at $\lambda = 550$ nm, one could theoretically achieve a resolution on the order of 1.4 μm, comparable to the size of the smallest photoreceptors. Foveal cones have center-to-center spacing ranging from 1.9 to 3.4 μm, whereas for rods, the range is 2.2–3.0 μm. However, uncorrected aberrations limit the lateral (transverse) resolution to approximately 10 μm and hence these small structures would be unresolvable.

Moreover, the aberration magnitudes change with time and hence a static correction such as spectacles or contact lenses would not be optimal. Even with static correction of the higher order aberrations, the temporal changes can reduce the retinal image contrast by 33 % and the Strehl ratio

(SR) by a factor of 3 [3]. The SR is defined as the ratio of the peak intensity in the aberrated point spread function (PSF) to that of the unaberrated case, a SR greater than 0.8 being considered diffraction limited. AO corrects both the low- and high-order aberrations in real time and has been applied to various ophthalmic imaging modalities: flash (flood) fundus cameras, confocal laser scanning ophthalmoscopes (cSLOs), and optical coherence tomography (OCT). To date, AO imaging has provided imagery of several important cell types, namely, cone, rod, RPE, and leukocyte cells in the intact, living human eye – in all cases without the use of any contrast-enhancing agents.

Brief History of AO in the Human Eye

In 1989, Dreher et al. [10] used a 13 electrode membrane DM to provide a static correction of astigmatism for use in conjunction with a scanning laser ophthalmoscope. Several years later, Liang et al. [11] implemented the Hartmann-Shack wavefront sensor (HS-WFS) for measurement of the human wavefront aberration. The same author [12] was then the first to combine a HS-WFS and a DM into one system for the active correction of the higher order ocular aberrations. Today, these systems use a variety of DM technologies with MEMS being just one subset. The next section details the application of AO to these cameras, highlighting the use of MEMS technology where appropriate.

Retinal Imaging Modalities

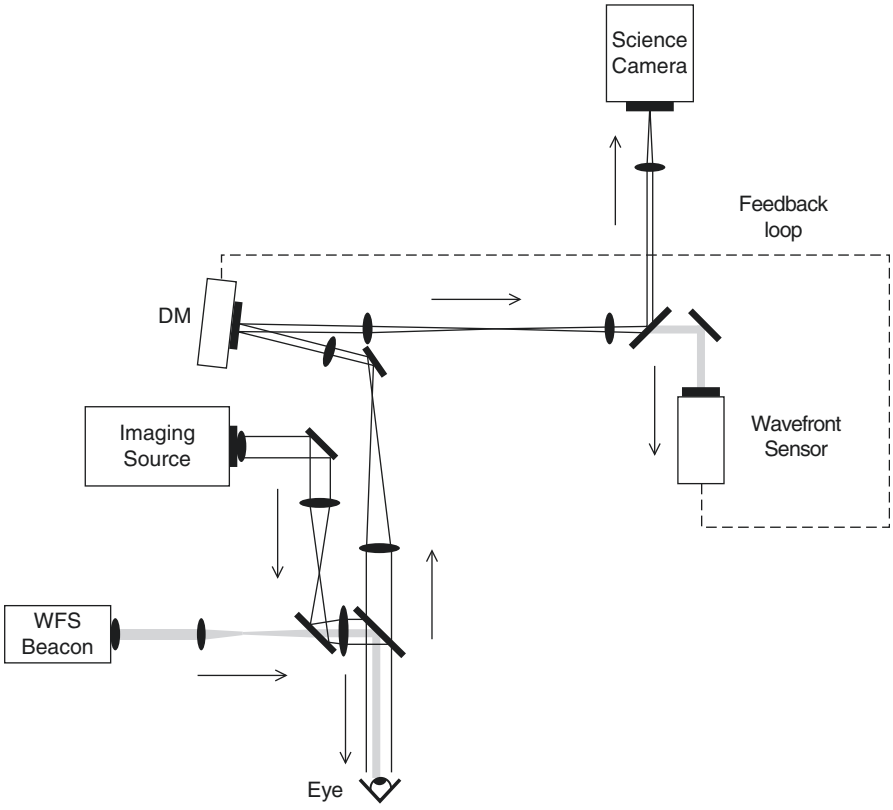
Flood-Illuminated (Flash) AO Fundus Cameras

Figure 7 shows the basic operation of a flood-illuminated (flash) AO fundus camera. Such systems take short exposure (4–6 ms) images of the retinal surface; longer exposures can introduce blurring due to eye movements. Two light sources are used: a superluminescent diode (SLD) in the near IR is used to measure the ocular aberration. A small (1 mm diameter) SLD beam (shown in gray) is incident on the cornea and focuses to a ~10 μm diameter spot on the retina. Small diameter beams are used to avoid blurring from the lens and cornea. The SLD light is scattered by the retina and the reflected light exits through the dilated pupil (typically 6–7 mm in diameter). This light is redirected by the DM and the dichroic beamsplitter into the WFS. The aberrations are measured and the DM commanded to take on the conjugate phase profile. Once the aberrations have been corrected to a suitable level (typically <0.1 μm RMS over a dilated 6–7 mm pupil), the imaging source (typically an arclamp) is triggered. The imaging wavelength (between 500 and 800 nm) is chosen through the use of an interference filter. The field of view (FOV) of the retinal images is 1–3° in diameter (~0.3–0.9 mm on the retina). Multiple images are acquired at one retinal location before post-processing to improve the signal to noise.

These systems were the first to use MEMS DM technology. In 2001, Fernandez et al. [13] used a 37-element MMDM from OKO Technologies to dynamically correct the human ocular aberration. Doble et al. [14] used a surface micromachined MEMS DM to obtain high-resolution images of the human cone photoreceptors in vivo. Figure 8 shows before and after AO images, visualization of individual photoreceptors corresponding to a transverse (lateral) resolution of approximately 3 μm at a horizontal eccentricity of 1° from the fovea are clearly visible [14]. Several other authors report high-resolution imaging of the human retina using an AO flood-illuminated fundus camera using other DM technologies [3].

Adaptive Optics Confocal Scanning Laser Ophthalmoscopes (AOCSLO)

Instead of taking a two-dimensional snapshot retinal image like those obtained with flood-illuminated fundus cameras, a SLO rasters a point of light across the retina with the reflected and de-scanned light being measured by a point detector such as an avalanche photodiode or photomultiplier tube. The confocal nature of the cSLO is implemented using a pinhole in front of the detector that is placed optically conjugate to a particular layer of the retina. This confocal pinhole then acts to block any scattered light from layers in front or behind the layer of interest, greatly improving image contrast. The raster is achieved through scanning mirrors (or rotating polygons) with typical frames rates of 20–30 Hz,



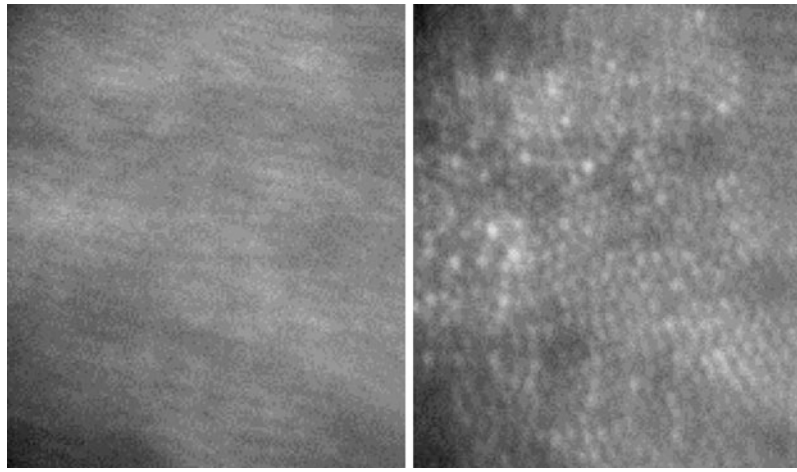
MEMS Adaptive Optics, Fig. 7 Schematic of a flood-illuminated (flash) AO fundus camera. A superluminescent diode (SLD) in the near IR is to measure the ocular aberration. A small (1 mm diameter) SLD beam (shown in *gray*) is incident on the cornea and focuses to a ~10 μm diameter spot on the retina. This light is scattered by the retina and the reflected light exits through the dilated pupil (typically 6–7 mm in diameter). This light is redirected by the DM and the dichroic beamsplitter into the WFS. Once the aberrations are corrected, the imaging source (typically an arclamp) is triggered. This source delivers a 4–6 ms retinal exposure. The wavelength (between 500 and 800 nm) is chosen through the use of an interference filter. The reflected visible light is then directed onto the science camera. Retinal images with diameter of 1–3° (0.3–0.9 mm) are acquired and post-processed to improve the signal to noise

each 512 × 512 pixels in size. cSLOs typically have lateral and axial resolutions of 5 and 300 μm, respectively. By incorporating AO, these values are improved to 2.5 μm laterally and <100 μm axially.

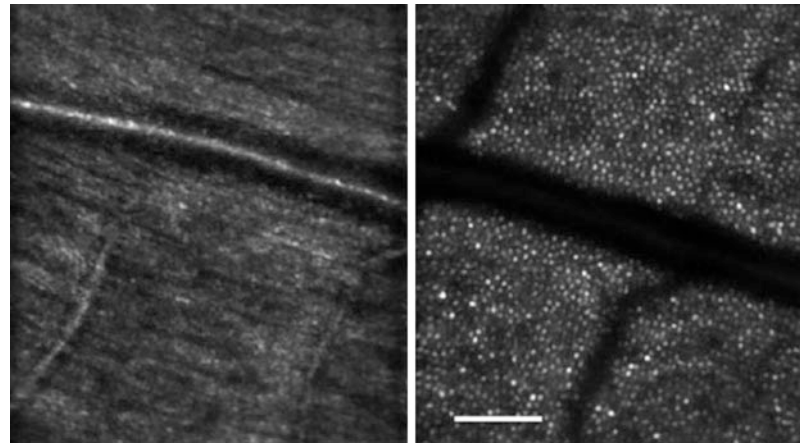
Roorda et al. [15] was the first to detail a fully closed loop AOcSLO system using a 37 channel Xinetics DM and a Shack Hartmann WFS. They demonstrated the axial sectioning capability of the system and highlighted its ability for dynamic retinal imagery by visualizing single bold cells flowing through the retinal capillaries. Newer AOcSLOs are employing MEMS DMs in conjunction with a low-order wavefront corrector such as trial lenses/Badal optometers or even a second DM. Figure 9 shows an example of the axial sectioning capability of an AOcSLO [16]. Adding in a defocus changes the retinal plane that is conjugate to the confocal pinhole, allowing for clear imagery of the retinal nerve fiber layer followed by the cone photoreceptors [17].

Adaptive Optics Optical Coherence Tomography (AO-OCT)

The final retinal imaging modality that has successfully employed AO is optical coherence tomography (OCT) [3]. OCT does not image the

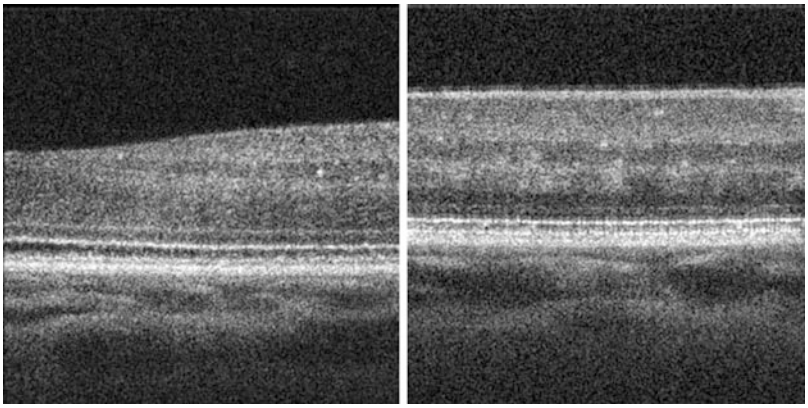


MEMS Adaptive Optics, Fig. 8 Before (*left*) and corrected (*right*) AO images of the human retina at an eccentricity of 1° using a surface micromachined MEMS DM in a flood-illuminated AO fundus camera. Cone photoreceptors are clearly visible in the corrected image; the *dark shadows* are anterior blood vessels. The horizontal field of view is 0.3° (75 μm on the retina), from Ref. [14] (Reproduced with permission of the Optical Society of America)



MEMS Adaptive Optics, Fig. 9 Images of the human retinal nerve fiber layer (*left*) and cone photoreceptors (*right*) [17] taken with the AOcSLO described by Burns et al. [16]. The scale bar is 100 μm. The ability for axial sectioning and video rate imaging of the retina is a major advantage of such systems

retina in the conventional sense but rather uses the coherence properties of light to form the image. It is an interferometric, noninvasive method that offers very high axial resolution. In OCT, the axial and lateral resolutions are decoupled. The axial resolution is a function of the source bandwidth, the wider the bandwidth the higher the resolution; however, the chromatic aberration of the eye can then become problematic. The lateral resolution of an OCT system is a function of the pupil size and hence is limited by optical aberrations. OCT systems are based on the Michelson interferometer using a broad bandwidth, low coherence light source. The light is split into and then recombined from a reference and a sample arm (which includes the eye). The first systems were operated in the time domain; today most systems work in the spectral domain whereby the recombined signal is passed through a spectrometer and the resulting spectra imaged onto a



MEMS Adaptive Optics, Fig. 10 AO-OCT images of the human retina [18] from the system described by Zawadzki et al. [19]. The high axial resolution (3.5 μm) coupled with the high lateral resolution (3 μm) afforded by AO enables exceptional 3D rendering of the living retina. *Left* −1.5° temporal retina, *right* −4.5° nasal retina. For both images the field of view is 3° with the system focused in the plane of the photoreceptors. Light would be incident from the *top* of the image

line scan camera – allowing for very rapid image acquisition.

To improve the lateral resolution, several research groups have incorporated AO into OCT systems [3]. The first systems employed large piezoelectric or bimorph DM technologies; today newer systems are employing MEMS DMs usually in conjunction with a second high stroke DM. Figure 10 shows AO-OCT images of the human retina [18] from the system described by Zawadzki et al. [19]. The high axial resolution (3.5 μm) coupled with the high lateral resolution (3 μm) afforded by AO enables exceptional 3D rendering of the living retina. Figure 10 (left) −1.5° temporal retina, right −4.5° nasal retina. For both images the field of view is 3° with the system focused in the plane of the photoreceptors. OCT images show clear stratification of the various retinal layers which are easily related to histological sections.

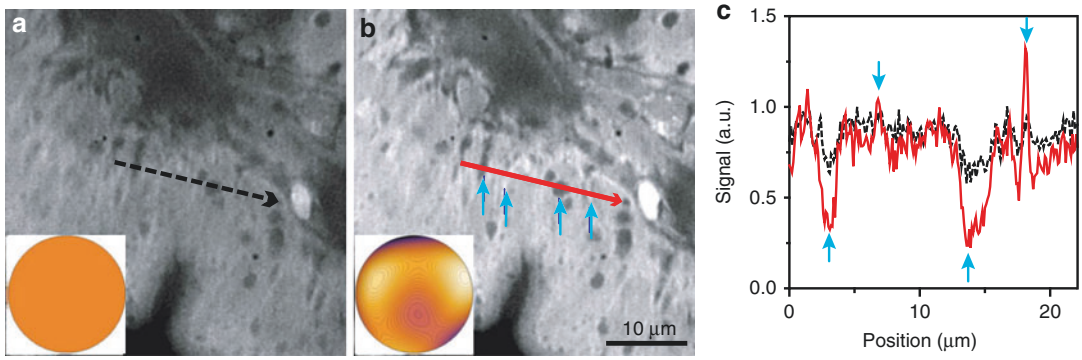
AO for Microscopy

Another application that is particularly suited to MEMS technology is the realm of biological microscopy. Like vision science, the small size of MEMS devices and the fact that they are relatively inexpensive make them an attractive choice.

Both confocal and two-photon imaging modalities are widely used in the study of cellular-level process [20, 21]. The confocal microscope works in a very similar way to that of the SLO described previously. The lens and cornea are replaced by a specially designed objective and the retina by the sample under study. A point source is rapidly scanned across the sample of interest and the reflected light is imaged onto a point detector. The two-photon microscope relies on excitation fluorescence caused by the simultaneous absorption of two photons by a molecule in the sample. A major advantage of the two-photon microscope is the very high depth penetration (around 1 mm). The excited fluorescence is very strongly localized due to the efficiency of the excitation process being proportional to the square of the intensity resulting in excellent three-dimensional imaging.

Sources of Aberrations

There are two primary sources of aberration in microscopy: (1) The sample itself is inhomogeneous with refractive index changes within it's three-dimensional structure this can lead to significant aberrations particularly when focusing through thick specimens. (2) There is a further index mismatch between the sample and the immersion media (oil or water). These aberrations cause loss of resolution, contrast, and signal.



MEMS Adaptive Optics, Fig. 11 Before (**a**) and after (**b**) AO-corrected image of mouse intestine under structured light illumination microscopy [25]. The *insets* show the phase in the pupil plane of the objective. The AO correction allows for the visualization of detail previously not observed – denoted by the *arrows*. (**c**) Shows the improvement in the recorded signal for the small features. The imaging depth was approximately 10 µm (Reproduced with permission of the Optical Society of America)

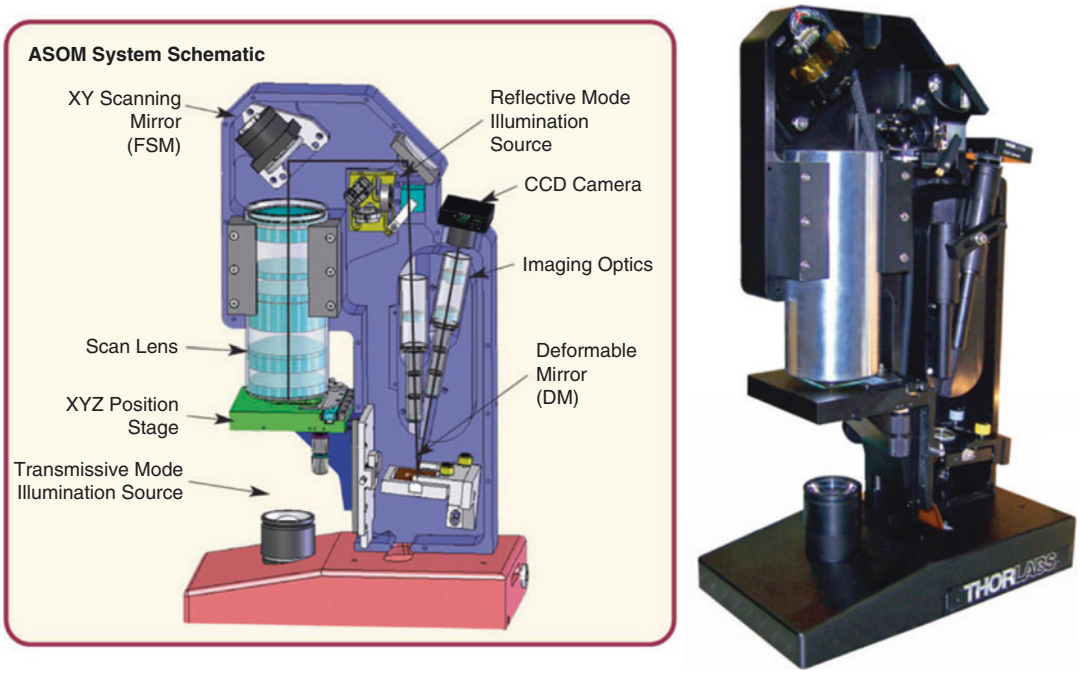
Wavefront Sensing

However, a particular challenge with microscopy is optical scatter, particularly when trying to image within thick samples. The scattering profile makes conventional WFS techniques such as Shack–Hartmann challenging to implement, as multiple layers within the sample can reflect light. In many AO microscopy systems, indirect (wavefront sensor-less) control has been applied to overcome the scatter issue. Indirect methods do not explicitly measure the wavefront aberration but instead use an iterative search algorithm to converge to the best AO correction state. Generally, a stochastic or hill climbing algorithm is used with an image quality metric (IQM). The IQM is a single figure that the search algorithm tries to optimize, example IQMs could be the Strehl ratio or simply the energy through the pinhole in a confocal system. Several authors have built on the first indirect astronomical AO systems [1, 2] and have adapted them for microscopy [22, 23]. One issue with these approaches is the convergence speed as many iterations are required to find the best solution; however, modal-based approaches reduce the number of iterations improving the convergence speed [24, 25]. Figure 11 shows before (a) and after (b) AO-corrected images of mouse intestine under structured light illumination microscopy [25]. Small details (denoted by the arrows) previously unseen were observed after the AO correction.

Indirect AO approaches have several advantages over direct schemes. Firstly, all of the light is used to form the image as none is required for the separate WFS arm; this reduces the required light levels minimizing possible phototoxic effects to the sample. Secondly, it removes any non-common path errors: in many AO systems there may be optics that are not in the optical path of the WFS; any misalignment will not be measured and hence will be uncorrected.

Using AO to Improve Microscope Capabilities

The standard role of AO is to correct for aberrations due to some intervening optical media between the object and the detector. However, it is also possible to use AO to reduce the cost of complex optical system by reducing the design constraints of certain elements. An example is the adaptive scanning optical microscope (ASOM) designed by researchers at Rensselaer Polytechnic Institute [26] and subsequently commercialized by Thorlabs. In high-power microscopes, methods to obtain a large field of view either require moving the sample itself and stitching together the small FOV images or designing wide field objective which can become prohibitively expensive for some applications. The ASOM overcomes these limitations by using a fast steering mirror and a MEMS DM to firstly scan the sample and secondly correct for the



MEMS Adaptive Optics, Fig. 12 The adaptive scanning optical microscope (*ASOM*) designed by researchers at Rensselaer Polytechnic Institute [26] and subsequently commercialized by Thorlabs. In high-power microscopes, methods to obtain a large field of view either require moving the sample itself and stitching together the small FOV images or designing wide field objectives which can become prohibitively expensive for some applications. The ASOM overcomes these limitations by using a fast steering mirror and a MEMS DM to firstly scan the sample and secondly correct for the aberrations at the various scan angles. This provides diffraction-limited imaging over a wide FOV (ASOM schematic and photograph provided courtesy of Thorlabs)

aberrations at the varying field angles. This provides diffraction-limited imaging over a wide FOV; a schematic and photograph of the system is shown in Fig. 12.

MEMS AO for Astronomy

All of the world's large ground-based telescopes require AO to overcome the degrading effects of atmospheric turbulence [1, 2]. The improvement afforded by AO has aided in every field of optical astronomical imaging. Large aperture DMs have generally been used in part to limit the demagnification required in going from the large (3–10 m) diameter primary mirrors to the rest of the optical train. In recent years, however, the increasing size and actuator counts of MEMS devices have made them attractive for astronomy.

Extreme AO

Extreme AO is the use of very high order AO systems in conjunction with coronagraphs or spectroscopy to enable imaging and measurements of extrasolar planetary systems. The aim being to detect very faint objects (planets) close to very bright sources (stars) that would otherwise be unseen. An exoplanet can be over a million times fainter than the companion star. Extreme AO systems increase the peak intensity of point-source images and remove the scattered light due to the atmosphere and the telescope optics. Overcoming these two sources of image degradation dramatically improves the contrast of these very faint objects.

An example instrument is the Gemini Planet Imager (GPI) project [27] being led by Lawrence Livermore National Laboratory, designed to image planets that are one to ten million times fainter than the companion star. GPI will use

MEMS AO to deliver diffraction-limited images and spectroscopy between 0.9 and 2.4 μm and is due to be delivered to Gemini South telescope (Cerro Pachon, Chile) in 2011. A key element of this system is a 4096 (64 × 64) actuator continuous surface MEMS DM manufactured by BMC, which will be used for the wavefront correction.

A European consortium is building a similar instrument for the European Southern Observatory. The Spectro-Polarimetric High-contrast Exoplanet Research (SPHERE) instrument [28] is due for delivery in 2011 to the Very Large Telescope (VLT) at Paranal, Chile.

Other MEMS AO Applications

There are several other applications that benefit from the small form factor and the potential low cost of MEMS devices.

Free space laser communications – Changes in the refractive index of turbulent air cause optical aberrations; these are manifest as beam spreading and scintillation causing loss of signal. These are particularly strong at ground level and become extremely challenging for horizontal, line-of-sight applications, such as laser communication and range finding. Atmospheric beam spreading requires the use of large detectors which in turn limits the data rates; by using AO to improve the beam quality, smaller, higher frequency detectors can be used. This can increase data rates by a factor of 10–100 [29].

Laser beam shaping – Many applications require precise control of the beam profile, for example, in laser machining. Beam control may be required to overcome aberrations due to thermal effects for example, or manipulation of the beam profile may be required to increase the efficiency of the material processing. AO may be implemented in either an intracavity or extracavity mode. Intracavity AO directly controls the beam profile and may be used to modify the temporal shape of the laser pulse or the transverse profile of the beam. For example, the ability to produce supergaussian profiles (so-called top hat functions) over the typical TEM00 Gaussian profile has particular attraction. Extracavity control is used for focus control or the control of optical distortions outside of the laser cavity. For laser machining applications the challenge for the use of MEMS AO is the harsh environment and the extremely high intensities incident on the active DM surface.

Optical manipulation – The effectiveness of the optical trap is limited by the aberrations (mainly spherical aberration) due to the refractive index mismatch at the glass/water interface. Similar to high NA microscopy, MEMS AO can be used to overcome this source of aberration [30].

Conclusion

The range and applicability of MEMS AO is constantly growing and evolving. It has proven highly successful in improving the performance of optical instrumentation in a variety of fields. Perhaps the biggest challenge to widespread adoption will be the cost of adding MEMS AO to a particular system. MEMS promises the ability to mass produce many millions of identical devices at low cost but to date the high-volume application is unclear. There exist low-volume, high-unit-cost applications such as microscopy and ophthalmology that make current AO technologies a viable economic proposition.

Acknowledgments The author would like to thank Drs. John S. Werner, Robert Zawadski (UC Davis), and Stephen Burns (Indiana University) for their support. Figure permission from Iris AO Inc., Boston Micromachines Corp, OKO Technologies, and Thorlabs is gratefully acknowledged.

Cross-References

▶ Basic MEMS Actuators
▶ CMOS MEMS Biosensors
▶ Confocal Laser Scanning Microscopy
▶ Optical Tweezers

References

1. Hardy, J.W.: Adaptive Optics for Astronomical Telescopes. Oxford University Press, Oxford (1998)

2. Tyson, R.K.: Principles of Adaptive Optics, 2nd edn. Academic, Boston (1998)
3. Porter, J., Queener, H., Lin, J., Thorn, K., Awwal, A.: Adaptive Optics for Vision Science. Wiley, Hoboken (2006)
4. Doble, N., Miller, D.T., Yoon, G.Y., Williams, D.R.: Requirements for discrete actuator and segmented wavefront correctors for aberration compensation in two large populations of human eyes. Appl. Opt. **46**, 4501–4514 (2007)
5. Vdovin, G.V., Sarro, P.M.: Flexible mirror micromachined in silicon. Appl. Opt. **34**, 2968–2972 (1995)
6. Bifano, T.G., Mali, R.K., Dorton, J.K., Perreault, J.A., Vandelli, M.N., Horenstein, M., Castanon, D.A.: Continuous membrane, surface micromachined, silicon deformable mirror. Opt. Eng. **36**, 1354–1360 (1997)
7. Manzanera, S., Helmbrecht, M.A., Kempf, C.J., Roorda, A.: MEMS segmented-based adaptive optics scanning laser ophthalmoscope. Biomed. Opt. Express **2**, 1204–1217 (2011)
8. Wandell, B.A.: Foundations of Vision, 1st edn. Sinauer Associates, Sunderland (1995)
9. Rodieck, R.W.: The First Steps in Seeing, 1st edn. Sinauer Associates, Sunderland (1998)
10. Dreher, A.W., Bille, J.F., Weinreb, R.N.: Active optical depth resolution improvement of the laser tomographic scanner. Appl. Opt. **28**, 804 (1989)
11. Liang, J., Grimm, B., Goelz, S., Bille, J.F.: Objective measurement of the wavefront aberration of the human eye with the use of a Hartmann-Shack wavefront sensor. J. Opt. Soc. Am. A **11**, 1949 (1994)
12. Liang, J., Williams, D.R., Miller, D.T.: Supernormal vision and high-resolution retinal imaging through adaptive optics. J. Opt. Soc. Am. A **14**, 2884 (1997)
13. Fernandez, E.J., Iglesias, I., Artal, P.: Closed-loop adaptive optics in the human eye. Opt. Lett. **26**, 746 (2001)
14. Doble, N., Yoon, G., Chen, L., Bierden, P., Singer, B., Olivier, S., Williams, D.R.: The use of a microelectromechanical mirror for adaptive optics in the human eye. Opt. Lett. **27**, 17 (2002)
15. Roorda, A., Romero-Borja, F., Donnelly, W.J., Queener, H., Herbert, T.J., Campbell, M.C.W.: Adaptive optics laser scanning ophthalmoscopy. Opt. Express **10**, 405 (2002)
16. Burns, S.A., Tumbar, R., Elsner, A.E., Ferguson, D., Hammer, D.X.: Large-field-of-view, modular, stabilized, adaptive-optics-based scanning laser ophthalmoscope. J. Opt. Soc. Am. A **24**, 1313–1326 (2007)
17. Burns, S.A.: Personal communication, April (2011)
18. Werner, J.S.: Personal communication, April (2011)
19. Zawadzki, R.J., Pilli, S., Kim, D.Y., Werner, J.S., Jones, S.M., Olivier, S.S.: Retinal imaging with a combined adaptive optics/optical coherence tomography and adaptive optics/scanning laser ophthalmoscopy system. Proc. Soc. Photo. Opt. Instrum. Eng. **7550**, 75500Z (2010)
20. Wilson, T.: Confocal Microscopy. Academic, London (1990)
21. Denk, W., Strickler, J., Webb, W.: Two-photon laser scanning fluorescence microscopy. Science **248**, 73–76 (1990)
22. Albert, O., Sherman, L., Mourou, G., Norris, T.B., Vdovin, G.: Smart microscope: an adaptive optics learning system for aberration correction in multiphoton confocal microscopy. Opt. Lett. **25**, 52–54 (2000)
23. Booth, M.J., Neil, M.A.A., Juskaitis, R., Wilson, T.: Adaptive aberration correction in a confocal microscope. Proc. Natl. Acad. Sci. U. S. A. **99**, 5788–5792 (2002)
24. Booth, M.J.: Wave front sensor-less adaptive optics: a model-based approach using sphere packings. Opt. Express **14**, 1339–1352 (2006)
25. Debarre, D., Botcherby, E.J., Booth, M.J., Wilson, T.: Adaptive optics for structured illumination microscopy. Opt. Express **16**, 9290–9305 (2008)
26. Potsaid, B., Bellouard, Y., Wen, J.T.: Adaptive scanning optical microscope (ASOM): a multidisciplinary optical microscope design for large field of view and high resolution imaging. Opt. Express **13**, 6504–6518 (2005)
27. Macintosh, B.A., Graham, J.R., Palmer, D.W., Doyon, R., Dunn, J., Gavel, D.T., Larkin, J.: The gemini planet imager: from science to design to construction. Proc. Soc. Photo. Opt. Instrum. Eng. **7015**, 701518 (2008)
28. Fusco, T., Petit, C., Rousset, G., Sauvage, J.-F., Dohlen, K., Mouillet, D., Charton, J.: Design of the extreme AO system for SPHERE, the planet finder instrument of the VLT. Proc. Soc. Photo. Opt. Instrum. Eng. **6272**, 62720K (2006)
29. Weyrauch, T., Vorontsov, M.A.: Free-space laser communications with adaptive optics: atmospheric compensation experiments. Opt. Fiber Commun. Rep. **1**, 247–271 (2008). Springer, New York
30. Ota, T., Sugiura, T., Kawata, S., Booth, M.J., Neil, M.A.A., Juskaitis, R., Wilson, T.: Enhancement of laser trapping force by spherical aberration correction using a deformable mirror. Jpn. J. Appl. Phys. **42**, L701–L703 (2003)

MEMS Capacitive Microphone

► Electrostatic MEMS Microphones

MEMS Condenser Microphone

► Electrostatic MEMS Microphones

MEMS Display Based on Interferometric Modulation Technology

Alok Govil[1], Ion Bita[1] and Evgeni Gusev[1,2]
[1]Qualcomm MEMS Technologies, Inc., San Jose, CA, USA
[2]MEMS Research and Innovation Center, Qualcomm MEMS Technologies, Inc., San Jose, CA, USA

Synonyms

Interference Modulator (IMOD) technology; mirasol® displays

Definition

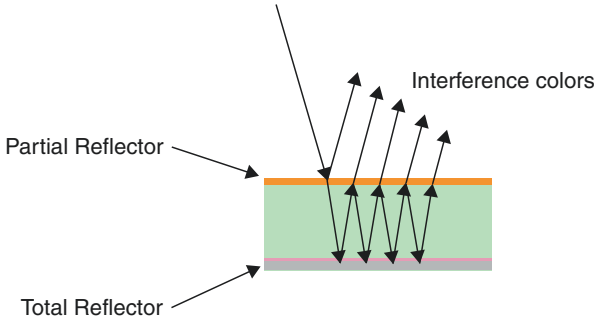
Happy "marriage" of micro/nanomachining (MEMS/NEMS) and thin-film optics brings about a new disruptive display technology, mirasol®, offering low-power operations and consistent viewing quality in various ambient lighting conditions including direct sunlight. Based on the principles of interferometric light and color modulation and microelectromechanical device operation and fabrication, the technology demonstrates unique attributes of low power consumption and reflective nature of display operations. Fast switching of the MEMS pixel elements enables video operation of mirasol displays at high refresh rates. mirasol displays can be manufactured using conventional tools and materials in standard glass TFT fabs balanced and modified for MEMS processing steps.

Introduction

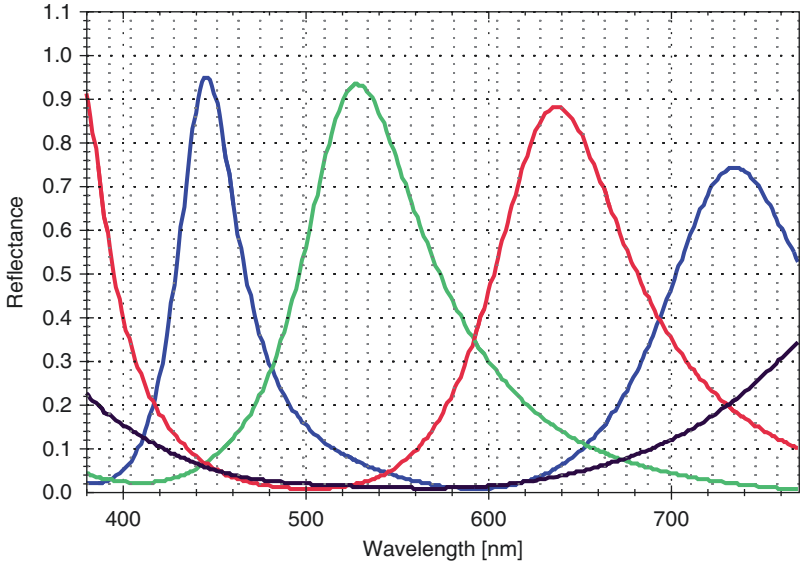
Coloration in nature is generally realized in two ways, through pigmentation or by means of controlling the flow of photons in thin optical films and/or 3D structures [1, 2]. Iridescence (originated from the Greek word *irides* – "rainbows") is an optical phenomenon caused by multiple reflections from multiple layers of optical films in which phase shift and interference of the reflections modulate the incident light by amplifying or attenuating certain wavelengths more than others. Some examples of this effect in living creatures include butterfly wings, feathers of some birds, and seashells [3].

Recent developments in micro- and nanofabrication and photonics allow one to fabricate structures that can mimic iridescence in nature by manipulating light in a controlled manner in MEMS/NEMS architectures, thus opening up a window for a variety of novel devices and applications [4–6]. The Qualcomm mirasol reflective display is a high-speed, electrostatically actuated, bi-stable MEMS device built on a transparent (e.g., glass) substrate and represents one of these emerging MEMS applications [7–10]. The basic device consists of a suspended conductive membrane serving as a mirror, over a partially reflective optical stack. There is a few hundred nanometers wide gap between the two that is filled with air. Interference between light reflected from the mirror and from the partially reflective optical stack generates vibrant color. Applying bias to this capacitive MEMS device color can cause change, for example, from color to black, thus enabling information display operation. The core of the technology, the mirasol display pixel, is designed on the principles of (1) thin-film interferometric modulation optics to produce color in the reflected spectrum of a (white) ambient light and (2) MEMS switching mechanism for changing the pixel state and, as a result, refreshing an image by applying low voltage. The same MEMS device achieves functions of color generation, fast light modulation, and memory, which are the three key elements required for making a video-capable color display. In contrast, the most commonly used display technology, liquid crystal display (LCD), uses color filters to generate colors, the liquid crystal layer with polarized light and an analyzer for light modulation, and generally thin-film transistor (TFT) circuits for memory.

Fundamentally, these principles offer significant power advantages because (1) reflected light (energy) is modulated from an ambient lighting



MEMS Display Based on Interferometric Modulation Technology, Fig. 1 "Folded" etalon



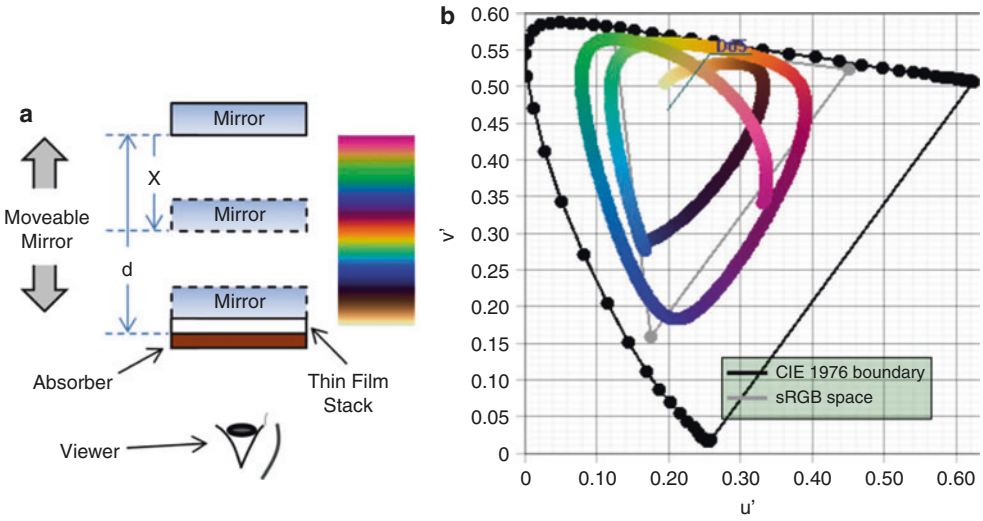
MEMS Display Based on Interferometric Modulation Technology, Fig. 2 Reflection spectra for four different air gaps corresponding to *red*, *green*, *blue*, and *black* colors respectively

source and (2) low-voltage MEMS pixel switching automatically means low power consumption during display array addressing. Therefore, this revolutionary interferometric modulator (IMOD) technology offers significant benefits in viewability, ruggedness, and power reduction, which make it attractive for a wide variety of consumer electronics applications. While LCDs in particular have also been used to make direct-view reflective displays, they have some inherent limitation especially due to the requirement of polarizer for light modulation because of which less than half of the light gets used to display images.

In this entry, basic electromechanical physics of the IMOD device and principles of color generation and modulation is described. Device fabrication and manufacturing utilizing mostly existing materials and tools available from TFT display industry will also be discussed as well as necessary modifications due to the MEMS nature of the technology. Other MEMS-based display technologies will be also reviewed.

Color Generation in IMODs

IMOD device is analogous to a Fabry Pérot interferometer [11, 12] operating in reflection mode (Figs. 1 and 3a). The air gap between the two reflectors shapes the spectrum of the reflected light as some wavelengths go though constructive interference while others through destructive interference. Figure 2 shows the example reflection spectra for IMODs for air gaps (on the order of 40–400 nm) corresponding to red, green, blue,



MEMS Display Based on Interferometric Modulation Technology, Fig. 3 IMOD colors for varied air gap. (**a**) IMOD device structure and color vs. gap. (**b**) IMOD color spiral on CIE u′-v′ chart

and black states respectively. The range of colors reached by the device as a function of the air gap can be computed using transfer-matrix method [11] and are shown on a standard CIE 1976 chromaticity diagram [13] in Fig. 3b for a given thin-film optical stack. A typical optical oxide stack has a thickness of several tens of nanometers.

Light Modulation and Electro-Mechanics

To change the color seen by the human observer, the air gap is changed by application of voltage on the two parallel plates of the IMOD device that results in electrostatic force of attraction between them given by:

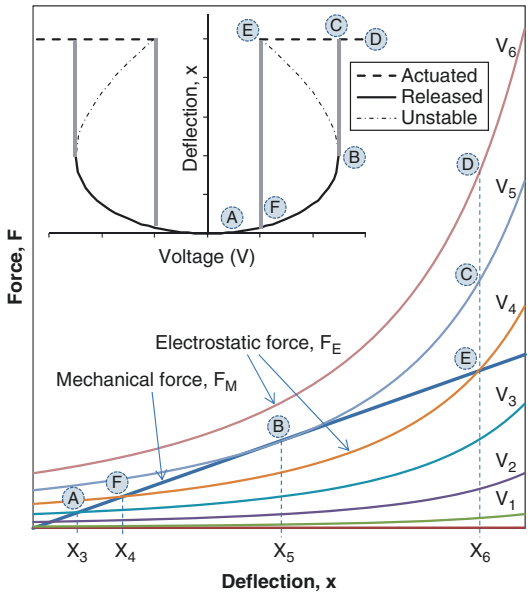
$$F_E = \frac{1}{2}QE = \frac{1}{2}A\varepsilon_0\left(\frac{V}{d-x}\right)^2$$

where F_E is the electrostatic force of attraction between the two plates, Q is the magnitude of charge on the plates, E is the electric field between them, A is the area of the plates, V is the applied voltage, d is the electrical gap in undeflected state, and x is the deflection.

Under equilibrium, this force is balanced by the mechanical restoring force F_M that follows Hooke's law:

$$F_M = K \cdot x$$

where K is the effective spring constant. This results in a cubic equation in x solving which results in the three solutions shown in Fig. 4 inset labeled actuated, released, and unstable respectively. Only two of the three solutions correspond to a stable equilibrium. With reference to Fig. 4, as the voltage is increased in magnitude starting from zero volts, IMOD starts in "released" state ('A') while the moveable plate gradually deflects toward the fixed plate. Once a threshold called pull-in or actuation voltage ('B') is crossed, the moveable plate collapses toward the fixed plate, reducing the air gap thickness to practically zero ('C') ("actuated" state). The transition to the actuated state is accompanied by a nonlinear increase of the electrostatic attraction, which increases rapidly as the air gap decreases. If the voltage now is decreased in magnitude, the electrostatic force needs to be reduced significantly before the moveable plate is able to snap back to the released state again. The threshold voltage level when the transition to the released state ('E' → 'F') happens is called release voltage. When a voltage between the actuation and release voltages called "bias voltage" is applied,



MEMS Display Based on Interferometric Modulation Technology, Fig. 4 Balance between electrical and mechanical forces leads to hysteresis

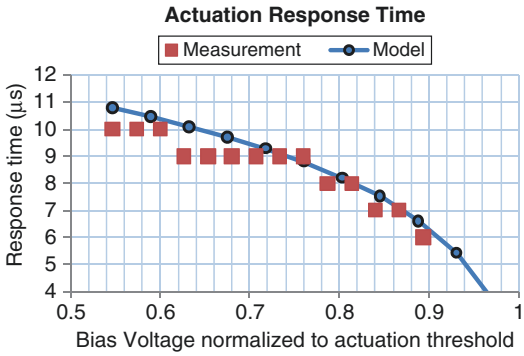
both actuated and released states are stable, akin to a memory effect where each pixel preserves the last state to which it was set.

The dynamical response time of the IMOD device can be modeled by including the terms for acceleration and squeezed-film damping into the force equation [14, 15]:

$$m\frac{d^2x}{dt^2} + \frac{\alpha\eta_0 A^2}{\left(1 + (9.638)K_n^{1.159}\right)(g-x)^3}\frac{dx}{dt} + Kx = \frac{A\varepsilon_0}{2}\left(\frac{V}{d-x}\right)^2,$$

where m is effective mass of the mirror, Kn is Knudsen number, η_0 is coefficient of viscosity of air, and α is a constant dependent on the geometry of the device [15]. Mechanical response times of a few microseconds are commonly achieved (compared to 2–40 ms for LCD), thus allowing for video capability. Figure 5 demonstrates a good match between model predictions for the response time and device measurements across a range of bias voltages.

Hysteresis behavior of the device is desirable for low-power operation as the content image

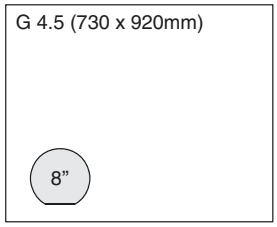


MEMS Display Based on Interferometric Modulation Technology, Fig. 5 Response time as a function of bias voltage; model vs. measurements

written on the panel does not need to be refreshed at a high frame rate as done by conventional LCD displays. It further allows passive matrix operation (thus saving panel manufacturing costs) without being subject to contrast ratio degradation as the vertical resolution of the display is increased [16].

To change the state of the IMOD device, a voltage pulse is applied to the device that takes it from the "hold" or bias voltage to either actuated or released state. Voltages much smaller in magnitude than the bias voltage are needed for this, allowing low-power image update operation.

Each mirasol pixel typically consists of three sub-pixels for red, green, and blue color channels, respectively, as traditionally used in other displays. The sub-pixels can be independently transitioned between the chosen primary color state and a black state. A common process sequence producing three separate air gaps is used to fabricate the individual sub-pixels. The color points for the sub-pixels are chosen from the IMOD color "spiral" shown in Fig. 3b considering white-balancing constraint and the trade-off between brightness and gamut pertinent to any reflective display technology. Given the bi-stable or 1-bit nature of the device, intermediate colors are generated for each sub-pixel by dividing it into multiple area-weighted IMODs and/or using spatial dithering techniques such as error diffusion commonly used by the print industry [17]. IMOD sizes used are optimized to avoid



Gen	Glass size (mm)
G5	1100 x 1300
G6	1500 x 1850
G7	1870 x 2200
G8	2160 x 2460

MEMS Display Based on Interferometric Modulation Technology, Fig. 6 Substrate size comparison – flat panel display glass generations, and 8″ (200 mm) wafer

MEMS Display Based on Interferometric Modulation Technology, Table 1 Examples of materials typically available for mirasol display manufacturing

Material	Usage
Al alloys (AlNd, AlSi, AlCu, etc.)	Gate metal, gate bus line
Ta, MoW, MoCr, Mo	Gate metal
ITO	Transparent pixel electrode
SiO_2, SiN_X	Passivation, gate dielectric
a-Si, p-Si	TFT channel

dither artifacts while still maintaining a reasonably high fill factor. Device sizes ranging from 30 to 60 μm are typically used for high-resolution displays.

Future generations of the mirasol display technology may use multistate or analog control of the air gap. This allows each IMOD to reach a larger range of colors from the color spiral shown in Fig. 3b having varied color saturation and including a natively bright white state. The ability of generating all colors with a single IMOD device not only allows better color gamut, but also fundamental improvements to the overall brightness of the display [18, 19].

IMOD Device Materials Science and Processing Fundamentals

The development of direct-view display applications of the IMOD technology has been shaped by a defining decision to leverage the set of materials and processes available in flat panel display factories in order to enable a scalable and cost-effective manufacturing platform [20]. Currently in production on G4.5 (730 × 920 mm) glass substrates, mirasol displays represent a significant development in the MEMS industry that has been traditionally based on smaller size silicon wafers. As a result of this strategy, the production IMOD MEMS device architectures had to be adapted to fit the capabilities and constraints of TFT LCD type fabrication infrastructure (Fig. 6).

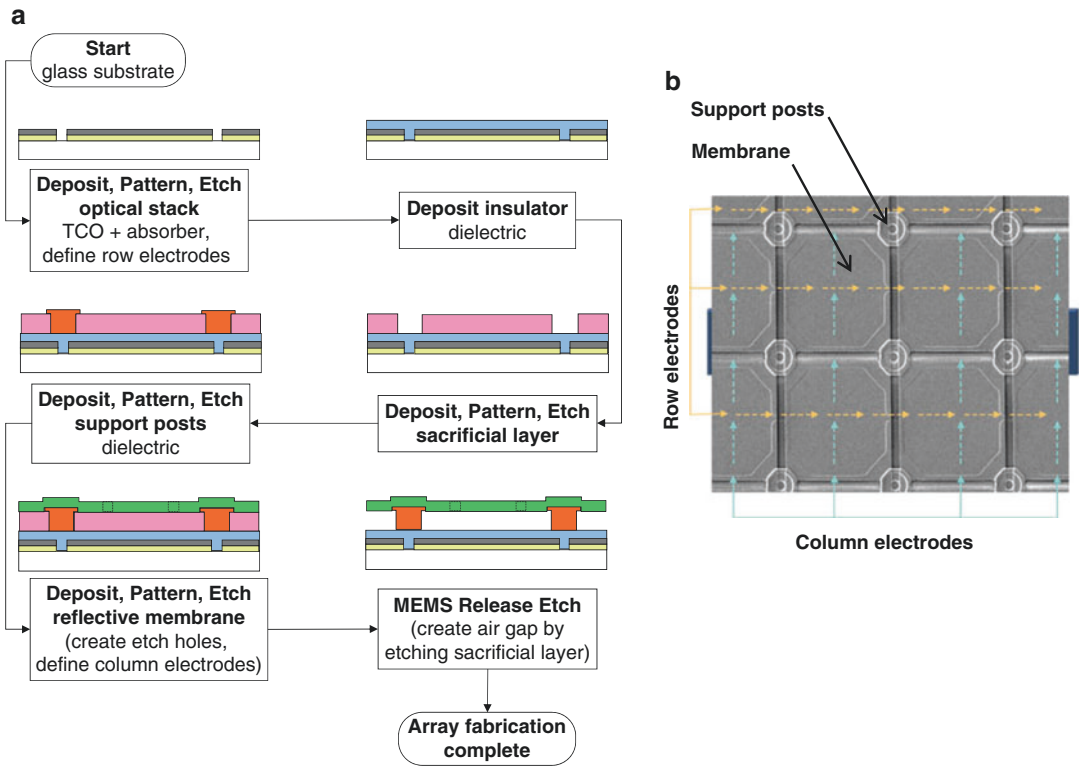
In this section, materials and processing aspects involved in fabricating the IMOD MEMS array glass panel are examined. The remaining steps toward producing finished mirasol displays are then described.

Basic Pixel Materials and Process Flow

The fabrication of the IMOD MEMS pixel is based on surface-micromachining of materials and process steps largely common to TFT array fabrication: large area deposition (Plasma-Enhanced Chemical Vapor Deposition (PECVD), sputter, etc.), photolithographic patterning, and etching (dry, wet) of dielectric and metallic layers [21]. In addition, MEMS-specific processes include forming the air gap below the moveable membrane, which is accomplished by depositing a sacrificial layer that is later etched out through appropriately designed openings in the membrane. Typical thickness of the films in the IMOD layers ranges from a several nanometers to submicron dimensions (Table 1).

A simplified process flow for IMOD pixel fabrication is shown in Fig. 7.

While there are significant similarities to standard processes used in TFT array manufacturing, a number of additional controls and optimizations are needed in order to ensure that the resulting MEMS device meets the desired operational and reliability specifications. Given the electro-opto-mechanical nature of the IMOD device, it is not surprising that additional attention is dedicated to engineering the combined optical, electrical, mechanical, and surface properties of the various layers forming the device. Some examples include the residual stress of the moveable membrane, the surface energy and topography of the surfaces on each side of the air gap [22], and the refractive index of each of the layers in the optical path [23].



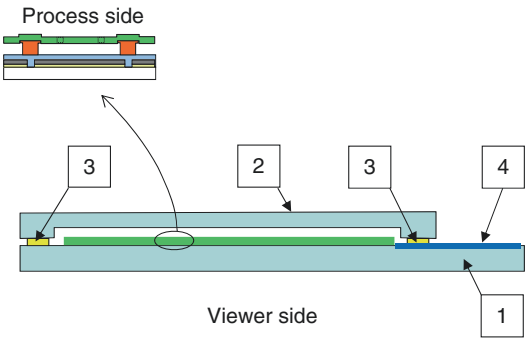
MEMS Display Based on Interferometric Modulation Technology, Fig. 7 (**a**) Pixel array glass fabrication process. (**b**) Scanning Electron Microscopy (SEM) plan view from the process side of a group of IMOD pixels

M

Packaging

The next step in mirasol display module manufacturing is the display panel fabrication. mirasol display panels are made by encapsulating the array glass with a back glass in order to protect the MEMS structures against physical damage or environment contamination (Fig. 8).

As with the pixel array fabrication, the mirasol strategy for packaging is also based on leveraging the scalable, high volume solutions that have been developed for the FPD industry. IMOD pixel array glass encapsulation is done at G4.5 plate level by using a back glass substrate (item 2 in Fig. 8), followed by singulation of the glass bi-layer to yield the individual display panels. The characteristics of the environment inside the panel package need to be controlled in order to meet the device lifetime requirements for mobile display applications. Typically, the seal is formed by using thin curable polymeric layers, with the package being at near atmospheric pressure and



MEMS Display Based on Interferometric Modulation Technology, Fig. 8 Typical mirasol display panel structure: *1* IMOD pixel array substrate, *2* glass backplate, *3* seal layer, *4* routing, and pads for display driver attachment

including a desiccant designed to have a moisture capture capacity sufficient for the display lifetime. It should also be noted that vacuum encapsulation is not required as in other traditional MEMS devices, since the pixel response time is already

MEMS Display Based on Interferometric Modulation Technology, Table 2 Typical mirasol display module components

Subsystem	Components
Display	mirasol display panel and driver chips
Optical stack	Diffuser, adhesive layers, cover glass
Touch sensor	Touch sensor panel
Front light	Front light panel and LED edge sources

very fast for display applications as discussed in the previous section.

mirasol Display Fabrication

The fully finished mirasol display product is assembled from a number of components that complement the mirasol display panel. The intrinsic attributes of IMOD technology that eliminate the need for using traditional display components such as polarizers, color filters, etc., also reduce the complexity of basic mirasol display module architectures.

As described in Table 2, at minimum, the display panel needs to be integrated with the driver chip and connected to the display interface. In practice, depending on the particular requirements for the display product, additional components are integrated within the module – optical diffusion layers to manage reflective view angle, touch panels for enabling advanced user interactions, internal light sources (front light) to enable use of the display in dark ambient conditions, and a cover lens.

Comparison to Other MEMS Display Technologies

There are several other MEMS-based display technologies, some of them can already be found in products, others are still in the development phase. They can be grouped into two broad categories, projection type displays and direct-view displays. Most existing MEMS displays fall under the projection type. They typically utilize a single micromirror or an array of MEMS mirrors to modulate light from a powerful RGB light source and hence create an image.

Texas Instruments has pioneered digital micromirror device (or DMD) which forms a core element for their digital light processing (DLP) projection display technology [24, 25]. DMD is composed of an array of MEMS mirrors on the order of 10 μm each fabricated on top of CMOS SRAM cells. Each alloy mirror made of an aluminum alloy is mounted on a yoke which is connected to two support posts by hinges. SRAM controls mirror position from flat to tilted by about 10–15°. DMD is a binary device. In one state, light from the source is reflected into the lens making the pixel appear bright on the screen. In the other state, the light is directed onto a mask which also acts as a heat sink. As a result, the pixel and its screen image appear dark. RGB color wheel is used to generate color images. Current DMD devices demonstrate billions on cycles of MEMS operation and can reliably operate in a variety of consumer products (DLP projectors, large screen projection TVs, pico-projectors for portable devices) [25].

Projection type moving/scanning mirror devices are also developed by Microvision [26] and Fraunhofer Institute [27, 28]. Both technologies exploit laser sources. In the heart of Microvision's silicon MEMS device, there is a scanning mirror connected to flexures. The MEMS scanner oscillates in two directions in order to display an image on a projection screen, pixel-by-pixel. Fraunhofer Institute approach is based on a two degrees-of-freedom MEMS resonator which can be used for horizontal deflection of the laser beam in a projection display system. The device consists of a circularly shaped silicon plate suspended by beams in the center of an elastically suspended driving plate. The moving part of the device forms two mechanically coupled resonators. The inner resonator acts as a moving mirror and the outer resonator drives this mechanical system. In a 1D implementation, both devices can be used as scanners.

The grating light valve (GLV) technology originally proposed at Stanford University [29, 30] and then further developed by Silicon Light Machines start-up [31] is based on light diffraction optics. It is a projection-based technology which operates using dynamically adjustable diffraction gratings. The GLV technology is built using silicon surface micromachining. The core element of the device consists of parallel rows of



MEMS Display Based on Interferometric Modulation Technology, Fig. 9 Photograph of a 5.7″ diagonal color mirasol display (1024 × 768 pixel resolution)

highly reflective micro-ribbons suspended above an air gap. Alternate ribbons can be actuated by applying voltage. When all ribbons are undeflected, the device acts as a mirror. When the ribbons are deflected toward the substrate and the deflection is one quarter wavelength, diffraction effects occur and incident light is reflected at a different angle. It is worthwhile to note that GLV devices are very fast with a switching speed in the nanoseconds range. Another attractive feature is that it is a contactless MEMS device which, in turn, enables robust design for reliability.

Pixtronix's digital microshutter technology targets direct-view transmissive display applications [32]. It is based on several key components: (1) MEMS module, (2) TFT backplane (similar to TFT arrays used in LCD display manufacturing), (3) backlight (optical train) plane with RGB light sources, and (4) field sequential color generation. The MEMS module includes high-speed (less than 100 μs) shutters that operate in a binary (open or closed) fashion and hence module light transmitting from the backlight plane.

Compared to above optical MEMS, mirasol distinguishes itself as a direct-view reflective display technology based on low-voltage capacitive MEMS arrays and interferometric color generation. These basic device and design principles enable low-power operations and consistent viewing quality in a variety ambient light conditions including direct sunlight.

Summary

mirasol, a new technology that mimics the way nature generates bright color by interferometric modulation, enables a new type of low-power reflective displays for portable applications with a consistent viewing experience under a wide range of ambient lighting conditions including bright sunlight [7, 8]. Interferometric modulation, which can produce color by "recycling" light from the ambient, is used in a novel display pixel architecture based on microelectromechanical devices. This unique combination of color reflection, low-voltage switching MEMS, and bi-stable pixel architectures enables a low-power consumption superior to existing conventional display technologies. Qualcomm MEMS Technologies [7] develops and commercializes mirasol displays covering the whole technology and product development cycle from early phases of research all the way to module manufacturing, with the first product introduced to the marketplace in 2008. The technology is particularly attractive for mobile devices including e-reader applications (Fig. 9). Current research activities center around new pixel architectures and materials with enhanced

display performance, innovations in the array addressing, and low-cost component research, such as the front light technology.

Cross-References

- Basic MEMS Actuators
- Capacitive MEMS Switches
- Chemical Vapor Deposition (CVD)
- CMOS MEMS Fabrication Technologies
- Dry Etching Processes
- MEMS Packaging
- Nanotechnology
- Optomechanical Resonators
- Scanning Electron Microscopy
- Structural Color in Animals

References

1. Vukusic, P.: Manipulating the flow of light with photonic crystals. Phys. Today **59**, 82 (2006)
2. Sampsell, J.B.: Causes of color: especially interference color. In: Proceedings of the Fourteenth IST Color Imaging Conference, Scottsdale, p. 90 (2006)
3. Pete, V., Sambles, J.R.: Photonic structures in biology. Nature **424**, 852–855 (2003)
4. Kovacs, G.T.A., Maluf, N.I., Petersen, K.E.: Bulk micromachining of silicon. Proc. IEEE **86**(8), 1536 (1998). doi:10.1109/5.704259
5. Bustillo, J.M., Howe, R.T., Muller, R.S.: Surface micromachining for microelectromechanical systems. Proc. IEEE **86**, 1552–1574 (1998)
6. Madou, M.J.: Fundamentals of Micro-Fabrication and Nanotechnology, 3d edn. Three-Volume Set: From MEMS to Bio-MEMS and Bio-NEMS: Manufacturing Techniques and Applications. CRC Press (2011)
7. Sampsell, J.B.: MEMS-based display technology drives next-generation FPDs for mobile applications. Inf. Disp. **22**, 24 (2006)
8. Miles, M.W.: Toward an iMoD ecosystem. In: Hilton Head 2006 – A Solid-State Sensors, Actuators and Microsystems Workshop (2006)
9. Miles, M.W., Larson, E., Chui, C., Kothari, M., Gally, B., Batey, J.: Digital Paper for Refractive Displays. SID, Boston (2002)
10. Felnhofer, D., Khazeni, K., Mignard, M., Tung, Y.J., Webster, J.R., Chui, C., Gusev, E.P.: Device physics of capacitive MEMS. Microelectron. Eng. **84**, 2158 (2007)
11. Macleod, H.A.: Thin-Film Optical Filters, 4th edn. CRC Press (2010)
12. Pulker, H.K.: Coatings on Glass, 2nd edn. Elsevier Science (1999)
13. Schanda, J.: Colorimetry: Understanding the CIE System. Wiley-Interscience (2007)
14. Veijola, T.: Equivalent circuit models of the squeezed film in a silicon accelerometer. Sens. Actuators **A48**, 239–248 (1995)
15. Bao, M.: Analysis and Design Principles of MEMS Devices. Elsevier Science (2005)
16. Alt, P.M., Pleshko, P.: Scanning limitations in liquid crystal displays. IEEE Trans. Electron Devices **ED-21**, 146 (1974)
17. Lau, D.L., Arce, G.R.: Modern Digital Halftoning, 2nd edn. CRC Press (2008)
18. Alex Henzen (iRex Technologies): Development of E-paper color display technologies. In: Society of Information Displays 2009 Digest, pp. 28–30 (2009)
19. Bita, I., Govil, A., Gusev, E.P.: IMOD/mirasol display technology. In: Chen, Janglin (John), Cranton, W., Fihn, M. (eds.) Handbook of Visual Display Technology. Springer, Berlin (2011)
20. Floyd, P.D., Heald, D., Arbuckle, B., et al.: IMOD display manufacturing. In: SID 2006 Proceedings (37), San Francisco, pp. 1980–1983 (2006)
21. Londergan, A., Gusev, E.P., Chui, C.: Advanced processes for MEMS-based displays. Proc. Asia Disp. SID **1**, 107–112 (2007)
22. Du, Y., de Groot, W.A., Kogut, L., Tung, Y.J., Gusev, E.P.: Effects of contacting surfaces on MEMS device reliability. J. Adhes. Sci. Technol. **24**, 2397–2413 (2010)
23. Gally, B.J.: White-Gamut color reflective displays using iMoD^{TM} interference technology. In: Proceedings of the SID, Seattle, p. 103 (2004)
24. Hornbeck, L.J.: Proc. Soc. Photo. Opt. Instrum. Eng. **1150**, 86 (1989)
25. Douglass, M.: DMD reliability: a MEMS success story. Proc. SPIE **4980**, 1–11 (2003)
26. Wine, D., Helsel, M.P., Jenkins, L., Urey, H., Osborn, T. D.: Performance of a biaxial MEMS-based scanner for microdisplay applications. Proc. SPIE **4178**, 186 (2000)
27. Urey, H., Wine, D.W., Osborn, T.D.: Optical performance requirements for MEMS-scanner based microdisplays. Proc. SPIE **4178**, 176–185 (2000)
28. Schenk, H., Dürr, P., Kunze, D., Lakner, H., Kück, H.: Design and modelling of large deflection micromechanical 1D and 2D scanning mirrors. Proc. SPIE **4178**, 116–125 (2000)
29. Bloom, D.M.: The grating light valve: revolutionizing display technology. Proc. SPIE **3013**, 165–171 (1997)
30. Solgaard, O.: Integrated semiconductor light modulators for fiber-optic and display applications. Ph.-D. thesis, Stanford University (1992)
31. Trisnadi, J.I., Carlisle, C.B., Monteverde, R.: Overview and applications of grating light valve-based optical write engines for high-speed digital imaging. In: SPIE Photonics West 2004, San Jose (2004)
32. Hagood, N., Barton, R., Brosnihan, T., Fijol, J., Gandhi, J., Halfman, M., Payne, R., Steyn, J.L.: A direct view MEMS display for mobile applications. In: SID 2007 Digest, paper 35.5L, pp. 1278–1281

MEMS Encapsulation

► MEMS Packaging

MEMS Fabricated Microwave Devices

► MEMS Vacuum Electronics

MEMS High-Density Probe Cards

Jung-Tang Huang[1], Kuo-Yu Lee[1], Hou-Jun Hsu[2], Pen-Shan Chao[3] and Chung-Yi Lin[4]
[1]National Taipei University of Technology, Taipei, Taiwan, ROC
[2]C.C.P. Contact Probes CO., LTD, New Taipei, Taiwan, ROC
[3]Chung-Shan Institute of Science and Technology, Taoyuan, Taiwan, ROC
[4]FormFactor, Inc, Hsinchu, Taiwan, ROC

Synonyms

Electronic contact testing cards; High-density testing cards

Definition

In the semiconductor production process, probe card is an important tool that examines whether the IC components are functioning properly before package. The IC components that do not meet the specifications will be eliminated after wafer cuts into individual chips. Probing ICs while they are still on the wafer normally requires that contact be made between the automatic test equipment (ATE) and IC. This contact is usually made with some form of mechanical probe. A set of mechanical probes will often be arranged together on a probe card, which is attached to the wafer prober. The wafer is lifted by the wafer prober until metal pads on one or more ICs on the wafer make physical contact with the probes. A certain amount of overdrive is required after the first probe makes contact with the wafer, for two reasons:

1. To guarantee that all probes have made contact (to account for non-planarity of the wafer and the probe tips).
2. To break through the thin oxidized layer (if the metal pad is aluminum) on the pad.

There are numerous types of mechanical probes available commercially: Their shape can be in the form of a cantilever, spring, or membrane, and they can be bent into shape, stamped, or made by microelectromechanical systems (MEMS) processing. Historically, probe card has evolved from cantilever type (1960), vertical type (1970), membrane type (1980), and micro-spring type (1990) to present MEMS probe card (2000∼). The evolution indicates the traditional mechanical processing methods cannot cope with the amount of I/O pad for current semiconductor IC. Only the development of near-mature MEMS technology can produce high-density probe card and even ultra high-density probe card.

Introduction

Under the trend of VLSI (ULSI), the volume of semiconductor transistor become smaller, that enhances the density of the components of integrated circuit (IC). But the technology of producing semiconductor encounters many new challenges as the volume of integrated circuit shrinks. In addition, the rapid development of electronic products requires all types of electronic products to be small, light, multifunctional, and fast. Thus, the trend of the integrated circuit will continue to set goal in developing higher density integrated products.

In recent years, the technology of developing of high-density semiconductor has been from submicron (submicro) technology to nanometer

(nano) system regulation. The pin count for I/O for semiconductor of IC has become more and more, the density has become higher, and the speed has become faster; similarly, the pin count of package's components has increased accordingly. Under the area limit, high-density integrated circuit must cope with pin count of high-density I/O; thus, traditional chips are packed with plastic or ceramic material and wire bonding to lead frame package do not meet the requirements. The structure of IC's pad has changed from peripheral arrangement to matrix arrangement or even 3-D IC stacking arrangement. The matrix type of arrangement of package can be completed through the technology of flip chip of ball matrix type arrangement, which is characterized by I/O pad using preimplantation solder ball through reflow to directly bond with substrates. This can shorten the connection path but also enhance the electrical transmission speed, and the packaging technology can meet current and future demand for electronic products.

In early period, the lead frame or flip-chip packaging methods have been continuously used for a long time. Previously, the wafer must be cut to be individual, which is integrated circuit (or bare die). The function of the probe card is to examine the features of the chips, and to filter out defective products; therefore, it can avoid completing the package steps of defective products, which might lead to additional costs and waste. In addition, the I/O pad of integrated circuits has been switched from peripheral arrangement to matrix arrangement in order to achieve the objective of higher density. The higher density and quantity of the pads has substantially increased the difficulties in examining ICs.

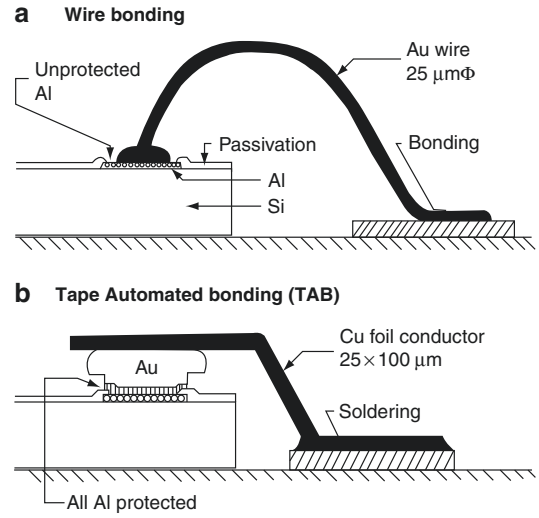
Traditional cantilever probe card has not been able to meet the testing conditions; thus, the development has shifted to probe card with matrix arrangement or the so-called vertical probe card. However, the capacity of bumping is expanding more than 25 % of compound annual growth rate as the technology of flip chips have been used widely. It is expected that there will be rapid growth on demand of integrated high-frequency matrix probe card.

Probe Cards for Peripheral Pads

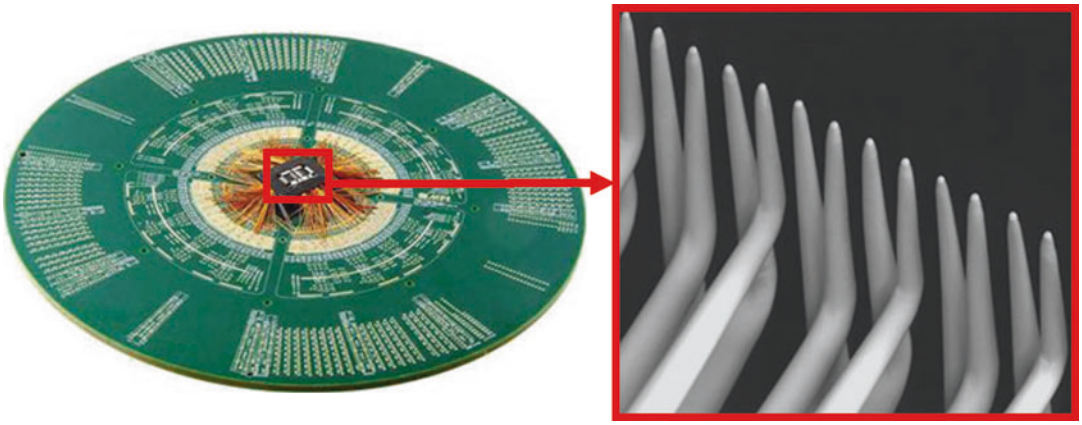
Conventional Packaging Technology

Wire bonding is the primary method of making interconnections between an integrated circuit (IC) die and a lead frame or a printed circuit board (PCB) during semiconductor device fabrication (Fig. 1a) [1]. Although less common, wire bonding can be used to connect an IC to other electronics or to connect from one PCB to another. Wire bonding is generally considered the most cost-effective and flexible interconnect technology, and is used to assemble the vast majority of semiconductor packages. Bond wires usually consist of one of the following materials: aluminum, gold, and copper. There are two main classes of wire bonding: ball bonding and wedge bonding.

Tape-automated bonding (TAB) is a process that places bare integrated circuits onto a printed circuit board (PCB) by attaching them to a polyamide or polyimide film (Fig. 1b) [1]. The mounting is done such that the bonding sites of the die, usually in the form of bumps or balls made of gold or solder, are connected to fine conductors on the tape, which provide the means of connecting the die to the package or directly to external circuits. Historically, TAB was created as an alternative to



MEMS High-Density Probe Cards, Fig. 1 Schematic of wire bonding and tape-automated bonding. [1]



MEMS High-Density Probe Cards, Fig. 2 Schematic of epoxy ring probe card [2]

wire bonding and finds common use by manufacturers in LCD display driver circuits.

Testing Technology for Peripherial Probe Pads

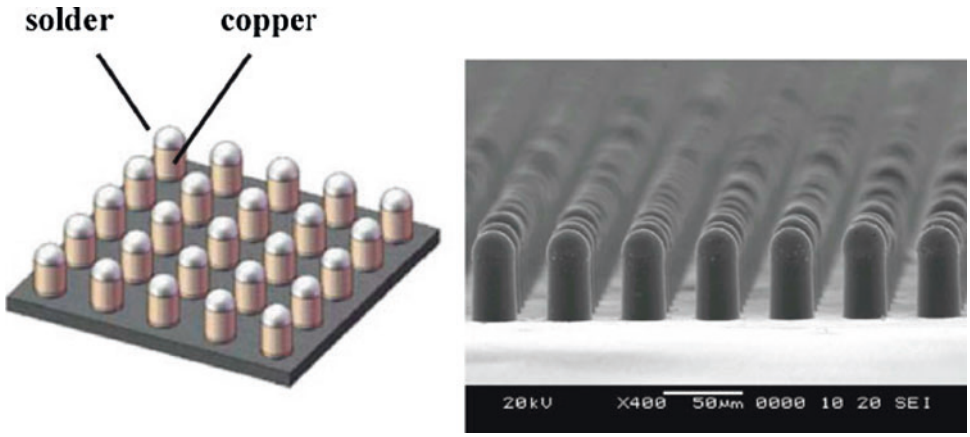
The epoxy ring probe card, chronically the industry standard for probing silicon wafers which include memory, RF and logic devices with peripheral pads, continues to dominate as the principal interface between the wafer and the ATE (Automatic Testing Equipment). The epoxy ring probe card with cantilever needles has the merits of lower cost and easy maintenance. In the past year, probably the greatest concentration of effort has been focused upon the needle design and quality associated with achieving a tighter pitch. At present, the pad pitch achieves the level to 30–40 μm. Figure 2 shows the picture of an epoxy ring probe card [2].

As for the needle material, the addition of beryllium to copper increases its strength, resulting in tougher wire but maintaining conductivity close to copper. The corresponding force for the beryllium copper needle is 1.2–1.5 g per pin at 50 μm overdrive, achieving the object of an ultra-low-force probe card. A low resistance must be achieved by the contact area at the end of probe sliding.

Fabrication procedure of the epoxy ring probe card includes Mylar drilling, needle bending, components assembling, epoxy fastening, conducting wire welding, needle pin etching, and position adjustment. In the probe card functions test, in addition to the detection of electric characteristics, the contact force and the forward displacement of the needle tip are inspected, which are both important indices to maintain the probing quality. The length of scrub mark on the weld pad is decided by the forward displacement of the needle tip. Due to the structure of cantilever needle, the arrangement of multiple needle arrays is difficult to meet the requirement of IC with matrix pads without mutual interference or non-consistent geometrical and mechanical features for each needle.

Probe Cards for Matrix Pads

Wire bonding process only provides peripheral interconnect, making a finer pad size and higher I/O count impossible. Other limitations are lower throughput and higher parasitic components, which would result in high propagation delay [3]. In tape-automated bonding (TAB), copper (Cu) leads pattern on polymer tapes are used to connect bare chips and substrates. The advantage of TAB technique is that it could test and burn-in devices prior to assembly; that is, it permits to test for known good die (KGD), provides better electrical performance, and higher circuit density compared to wire bonding. The disadvantage of TAB is that it requires additional processes to bump devices or tapes and specially



MEMS High-Density Probe Cards, Fig. 3 Schematic of solder bump package

designed tapes for each die. Similar to wire bonding, TAB provides only peripheral interconnect [4]. As the demand for cheaper, smaller, faster, portable, and yet multi-featured electronic consumer devices or products increase, wide use of flip-chip bonding in high-density electronic packaging will become indispensable [4]. Accordingly probe cards for probing matrix pads of flip chips are necessary tools for guaranteeing known good die (KGD), especially further stack IC or 3-D IC package in the future.

Solder Bump Technology

Flip-chip bonding, in essence, utilizes an array of solder bumps varying in shapes, sizes, or materials, to serve as I/O count to connect with chips or substrates rather than employ traditional lead frames or gold wires as connecting devices. This innovative packaging process invented by IBM in the early 1960s is characterized by yielding smaller substrate size, larger number of I/O count, and less pitch size between each solder bump [7]. Figure 3 shows the high-density copper pillar bumps with pitch of 50 µm and diameter of 30 µm.

Other advantages provided by flip-chip bonding include better electrical performance, shorter interconnection length, lower inductance effect, better electromagnetic shielding effect, self-alignment ability, and lower cost [8–11].

MEMS High-Density Probe Cards, Table 1 Comparison of three COB bonding processes

	Wire bonding	Tape-automated bonding	Flip-chip bonding
Area ratio	1	1.33	0.33
Weight ratio	1	0.25	0.2
Thickness ratio	1	0.67	0.52
Input/Output (Max)	300–500	500–700	>1,000
Rework	Superior	Inferior	Superior
Cost	×	>2×	0.8×

A detail comparison among wire bonding, tape-automated bonding, and flip-chip bonding is listed in Table 1.

Vertical probe cards are indispensable for testing flip chip with matrix pads due to limited space of fine pitch among pads to accommodate too many horizontal cantilever probes. At present, FormFactor Company catches the largest market share in the probe card industry. In addition, a famous vertical probe, named as Cobra, invented by IBM was initiated in late 1970s for testing C4 (Controlled Collapse Chip Connection – which is also associated with flip-chip package) process base semiconductor devices.

MEMS Technology for Matrix Pads

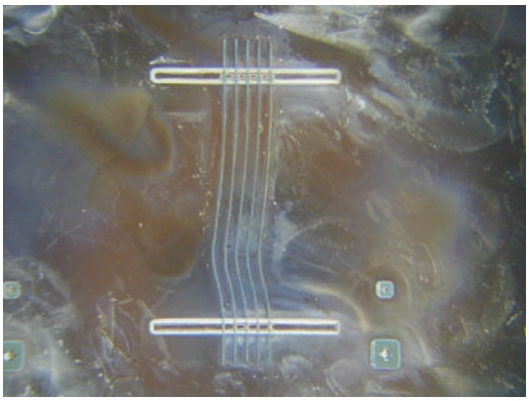
Most of the traditional methods of vertical probe production are completed through punching or drawing; thus, it is not easy to control the radius of the curves of the probe. It is then followed up with assemble work which is particularly difficult to control its accuracy. The traditional probe card has the following disadvantages:

1. There is poor coplanarity of probes due to manual assembly, which causes defective contacts that lead to excessive contact resistance and even fail conduction.
2. It is not easy to adjust the accurate position of each probe; thus, it must go back to the factory for maintenance after every period of using time.

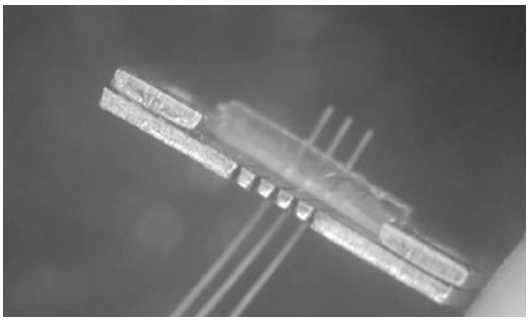
MEMS vertical probe card employs MEMS technology with electroforming alloy technology to produce spring-like probe, which enables the probe to have greater degree of compliance, and provides adequate vertical deformation (Fig. 4). For the purpose of convenience and alignment in the probe assembly process, a guide plate is indispensable (Fig. 5). The guide plate can be fabricated by using the lithography, electroforming, and DLC coating processes to complete the definition of perfect spacing between each guide hole. The insulated guide plate is then combined with the technology of packaging to produce high precision probe card. In this way, it can reduce complexity in the assembly process with accuracy at the same time.

Besides importing the technology of MEMS to produce probe card, it can also use the technology of CMOS-MEMS to design the probe's arrangement and force sensor. It can improve the issue of accuracy control for many traditional probe cards. The standard CMOS process, such as TSMC 0.35 um or 0.18 um process can provide flexibility of designing microstructure with diversification and low development cost. The main advantages are summarized in the following points:

1. The accuracy and the coplanarity can be improved by using the technology of multiple probe packaging [12, 13].



MEMS High-Density Probe Cards, Fig. 4 Schematic of packaged with the probe photoresist



MEMS High-Density Probe Cards, Fig. 5 Schematic of single row of the probe assembly to the guide

2. The FEA software used to analyze design can optimize overdrive and vertical press force to reduce contact resistance and horizontal force that scrub off oxide layer on the surface of the pad.
3. The planar or 2-D manufacturing process can produce long and thin probe with different width [12, 13].
4. Using lithography and the electroforming technology can enhance the capacity of batch production [14].
5. The multiple probe packaging with real-time detection of force sensor can decrease the recovering or repair time of probe, and reduce cost and labor [15–18].

As the production of probe becomes smaller, it is very important to know the characteristic of the

probe to find out the maximum current and resistivity that a probe can withstand. Based on the study [12], the result after number of measurements was that the probe maximum current was between 700 and 850 mA and the resistance was 35 μΩ-cm or less. This probe composes of nickel cobalt alloy, and the result of EDX quantitative ratio analysis of components was Co 28 % and Ni 70 %. The nano-indentation system had evenly gathered ten testing points in the testing chip, which yielded Young's modulus of 180 GPa for nickel and cobalt alloys, and hardness of HV 450.

MEMS Probe Evolution

In 1995, Mark Beiley et al. have used thin film structure with importing gas as a way for the probe to apply force, in order to achieve the concept of reduction of contact resistance [19, 20]. In 1997 and 1999, Yanwei Zhang et al. have used technology of MEMS to produce circular form of bimetallic microprobe which is working as a thermal actuator [21, 22]. In 2000, Ito Takahiro et al. have used SOI wafer technology to produce probe with arch-shaped beam [23].

In 2000, Dong-Seok Lee et al. have also used SOI chip and bulk micromachining technology to produce microprobe structure that the probe tip could move up to 80 μm. In 2001, Robert B. Marcus et al. brought up a new technology, which was to utilize the residual stress of bimetallic (Cu, Cr) that worked with sacrificial layer etching, and then, after annealing treatment, it would produce a warping cantilever [24]. In 2002, Bong-Hwan Kim et al. used monocrystalline silicon (100) to produce structure of three-dimensional microprobes [25, 26]. In 2002, Kenichi Kataoka et al. used electroforming nickel process that worked with sacrificial layer etching and utilized residual stress to form a warping cantilever.

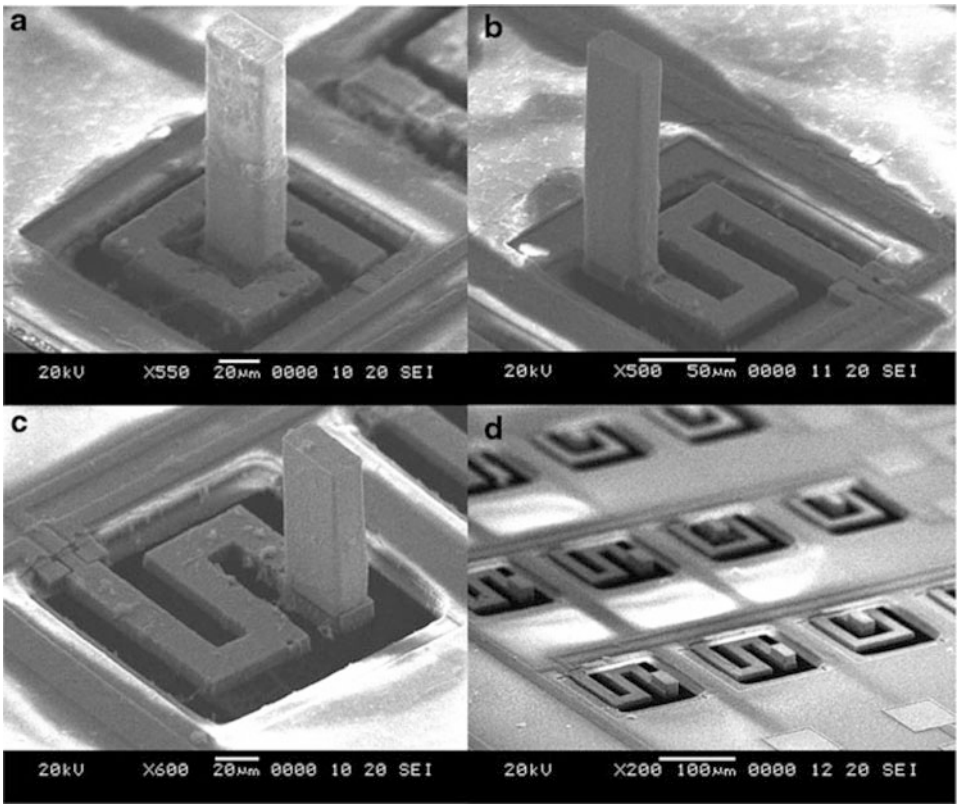
In 2003, Younghak Cho et al. used properties of anisotropic etching for KOH to etch an inverted triangle concave hole in the monocrystalline silicon, and then used deposition method to produce metal (Cr, W, Au) probe tip [27]. In 2004, K. Kataoka brought up electroformed multilayer film with three-dimensional S-shaped spring of microprobe arrangement [28]. In 2005, Si-Hyung Lee and Bruce C. Kim took advantages of two different residual stress relationships between BCB and Mo layers to produce a warp probe beam for wafer-level testing. [29]. In 2005, Young-Min Kim used KOH to etch silicon wafer and annealing process to produce microprobe [30].

In 2007, T. Namazu et al. used nickel titanium memory alloy to produce micro-actuator MEMS probe card [31]. In 2007, Seong-Hun Choe et al. used doping of P++ to silicon to make cantilever probe [32–34]. In 2008, Fei Wan et al. used wet etching and microelectroforming of nickel on a 4 in. wafer to provide 110,000 out-plane hoe-shaped probes for wafer-level IC testing [35–37]. Then all the probes were transferred onto the corresponding solder balls of the adaptor board. The contact force of the probe is 7.5 mN with deformation of 20 μm, the pitch of the probe is X-160 μm, Y-240 μm, and the contact resistance is 0.8 Ω.

In 2008, J. Liu et al. used the technology of CMOS process, MEMS etching, and microelectroforming of nickel to produce thermoelectric actuating cantilever probe [38]. The probe is to test the memory components, and has contact resistance of 10 Ω and the power consumption of 3.7 mW. In 2009, Fei Wang et al. used bulk micromachining technology as cantilever arm of the probe, while using the concept of TSV to produce tip of the probe [39–41]. It was used in wafer-level IC testing, of which the displacement was about 20–33 μm, the contacting force of 50–80 mN, contact resistance of 1 Ω. In 2009, Bong-Hwan Kim used bulk micromachining technology and electroformed nickel and cobalt alloy material to produce MEMS probe used in testing DRAM [42–45]. The probe had contacting force of 20 mN at deformation of 50 μm and resulted in contact resistance of 3.5 Ω.

Summary

The probe card produced by using CMOS process with MEMS process [46–49] has become indispensable testing components in the next generation of semiconductor (Fig. 6). It can be used in



MEMS High-Density Probe Cards, Fig. 6 SEM images of the CMOS-MEMS probe chip after ICP dry etching: (**a**) a 106-mm-tall probe tip with spiral cantilever, (**b**) a 106-mm-tall probe tip with meander cantilever, (**c**) a 42.4-mm-tall probe tip with spiral cantilever, and (**d**) the chip with a 14.8-mm-tall probe tip

the practice of high frequency, wafer-level testing, and batch of production, which can reduce the cost to meet the trend.

Regardless of how the semiconductor's technology evolve, the probe card produced by CMOS process with MEMS technology can work effectively with the IC pad size and space, especially the circuits are next to the back end of the probe, so it can reduce the transmission losses and noises. This is a technology to realize system integration on a chip (System On Chip), which cannot be reached with the traditional assembled probe.

Using CMOS technology not only significantly reduces the size of the probe, but it can also connect the external components with the circuit, which are integrated in the design of CMOS chip. Thus, it can increase the convenience of wiring that is the space transformer needed by the probe card, which can be achieved using the metallic layer in the CMOS process. The process is not only simple and without traditional assemble method, but also enhances yield rate; therefore, the cost can be significantly decreased. This will be a great support for the present and future generation of testing IC.

While the IC designer implements the IC's layout, he may also provide with the circuit layout that is required for testing IC based on the technology of the built-in testing circuits on the CMOS-MEMS probe card. These can be processes together to save cost and accelerate time for product launch at market in order to achieve ultimate technology in "Using IC to test IC." Especially when the pilot run is completed for mass production, the use of CMOS-MEMS probe card can accelerate time on product to be in the market.

References

1. Halbo, L., Ohlckers, P.: Electronic components, packaging and production. University of Oslo, Oslo (1995). ISBN 82-992193-2-9
2. http://www.mpi.com.tw/. 6 Oct 2011
3. Jittinorasett, S.: UBM formation on single die/dice for flip chip applications. Thesis of Virginia Polytechnic Institute and State University, United States, pp. 5–6. (1999)
4. Lee, S., Guo, Y.X., Ong, C.K.: Electromigration effect on Cu-pillar (Sn) bumps. In: IEEE Electronics Package, Technical Conference, Singapore, p. 135 (2005)
5. Riley, G.A.: Introduction to Flip Chip: What, Why, How. http://www.flipchips.com/tutorial01.html. 6 Oct 2011
6. Riley, G.A.: Picking your flipchip: A comparison of flipchip methods and benefits. Electron. Manuf. Eng. **15**(3 Third Quarter), 1–4 (2000)
7. http://www.esrf.fr/events/conferences/past-conferences-and-workshops/IWORID7/FinalProgramme/volpert_1.pdf
8. http://www.flipchips.com/SMErev01.htm
9. Chang, C.S., Oscilowski, A., Bracken, R.C.: Future challenges in electronics package. IEEE Circuits Devices Magazine **14**(2), 45–54 (1998)
10. Jang, S.Y., Paik, K.W.: Comparison of electroplated eutectic Bi/Sn and 9. Pb/Sn solder bumps on various UBM systems. IEEE Trans. Electron. Packag. Manuf. **24**, 269 (2001)
11. Wolf, M.J., Engelmann, G., Dietrich, L., Reichl, H.: Flip chip bumping technology-status and update. J. Nucl. Instrum. Method. Phys. Res. **565**, 291 (2006)
12. Huang, J.T., Lee, K.Y., Wu, C.S., Lin, C.Y., Shih, S. H.: Using micro-electroforming and micro-assembly technology to fabricate vertical probe card. In: The 8th International Conference on Electronic Materials and Packaging, Hong Kong, pp. 435–439 (2006)
13. Huang, J.T., Wu, C.S., Lee, K.Y., Hsu, H.J., Chao, P.S., Shin, S.H.: Batch mode fabrication on high coplanarity with pressure sensors by using MEMS. In: The 4th International Conference on Materials for Advanced Technologies, Singapore, pp. 84–87 (2007)
14. Huang, J.T., Hsu, H.J., Chao, P.S., Lee, K.Y., Wu, C. S., Shih, S.H., Lin M.Z., Lee, F.Y., Lan, Z.C.: Fabrication of a MEMS-Based cobra probe. In: IEEE International Conference on Electron Devices and Solid-State Circuits, pp. 809–812 (2007)
15. Lee, K.Y., Huang, J.T., Hsu, H.J., Chiu, M.C., Tsai, T. C., Chen, C.K.: CMOS-MEMS piezoresistive force sensor with scanning signal process circuit for vertical probe card. Sens. Actuator A **160**, 22–28 (2010)
16. Huang, J.T., Chiu, M.C., Lee, K.Y., Wu, C.S., Hsu, H. J., Chao, P.S.: Development of CMOS process compatible force sensor and its application to probe card. In: IEEE International Conference on Electron Devices and Solid-State Circuits, Taiwan, pp. 817–820 (2007)
17. Huang, J.T., Lee, K.Y., Chiu, M.C., Hsu, H.J., Shih, S. H.: The integration of CMOS process compatible force sensor and scanning signal process circuit for vertical probe card. In: International Conference on Electronic Packag, Tokyo, pp. 346–350 (2008)
18. Lee, K.Y., Huang, J.T., Hsu, H.J., Chiu, M.C., Tsai, T. C.: CMOS-MEMS piezoresistive force sensor for vertical probe card. In: The 22nd International Microprocesses and Nanotechnology Conferences, pp. 420–421 (2008)
19. Beiley, M., Leung, J., Wong, S.S.: A micromachined array probe card-fabrication process. IEEE Trans. Compon. Packag. Manuf. Technol. Part B **18**, 179–183 (1995)
20. Beiley, M., Leung, J.: A micromachined array probe card-characterization. IEEE Trans. Compon. Packag. Manuf. Technol. Part B **15**, 84–191 (1995)
21. Zhang, Y., Worsham, D., Morrow, D., Marcus, R.B.: A new MEMS wafer probe card. In: MEMS Conference, pp. 395–399. (1997)
22. Zhang, Y.: Thermally actuated microprobes for a new wafer probe card. J. MEMS **8**, 43–49 (1999)
23. Takahiro, I., Sawada, R., Higurashi, E.: Fabrication of micro IC probe for LSI testing. Sens. Actuator. A **80**, 126–131 (2000)
24. Marcus, R.B.: A new coiled microspring contact technology. In: Electronic Components and Technical Conference, Orlando (2001)
25. Kim, B.H., Park, S.J.: Fabrication of nickel electroplated Cantilever-type MEMS probe card with Through-hole interconnection. In: Microprocesses and Nanotechnology Conference, Tokyo, pp. 170–171 (2003)
26. Kim, B.H., Park, S.: A novel MEMS silicon probe card. In: MEMS Conference, Las Vegas, Nevada, pp. 368–371 (2002)
27. Cho, Y., Kuki, T.: Si-based micro probe card with sharp knife-edged tips combined metal deposition. In: Actuators and Microsystems International Conference, Boston, vol. 1, pp. 774–777 (2003)
28. Kataoka, K., Itoh, T.: Multi-layer electroplated microspring array for MEMS probe card. In: MEMS Conference, pp. 733–736 (2004)
29. Lee, S.H., Kim, B.C.: Curled micro-cantilevers using benzocyclobutene polymer and Mo for wafer level probing. Sens. Actuator. A **121**, 472–479 (2005)
30. Kim, Y.M., Yoon, H.C.: Silicon Micro-probe card using porous silicon micromachining technology. J. ETRI **27**, 433–438 (2007)
31. Namazu, T., Tashiro, Y., Inoue, S.: Ti–Ni shape memory alloy film-actuated microstructures for a MEMS probe card. J. Micromech. Microeng. **17**, 154–162 (2007)
32. Choe, S.H., Fujimoto, S., Tanaka, S., Esashi, M.: A matched expansion probe card for high temperature LSI testing. In: The International Conference on Solid-state Sensors, Actuators and Microsystems, Seoul, pp. 1259–1262 (2005)

33. Choe, S.H., Tanaka, S., Esashi, M.: A matched expansion MEMS probe card with low CTE LTCC substrate. In: Proceedings. International Test Conference, Santa Clara, 4437618
34. Tanaka, S., Fujimoto, S., Ito, O., Choe, S.H., Esashi, M.: Fabrication of microprobes on a ultrathick glass substrate with narrow-pitch electrical feedthroughs for next-generation LSI burn-in tests. In: MEMS Conference, pp. 347–350 (2008)
35. Wang, F., Li, X., Feng, S.: A MEMS probe card with 2D dense-arrayed "hoe"-shaped metal tips. J. Micromech. Microeng. **18**, 1–8 (2008)
36. Wang, F., Li, X., Cheng, R., Yang, H., Wang, Y., Feng, S., Ge, X., Chen, T., Chen, L., Sun, L.: Two-dimensional dense-arrayed probe-cards with a hoe-shaped probing-tip micromachining technique. In: MEMS Conference, pp. 343–346 (2008)
37. Wang, F., Cheng, R., Li, X.: MEMS vertical probe cards with ultra densely arrayed metal probes for wafer-level IC testing. J. Microelectromechan. Syst. **18**, 933–941 (2009)
38. Liu, J., Noman, M., Bain, J.A., Schlesinger, T.E., Fedder, G.K.: CMOS-MEMS probes for reconfigurable IC's. In: MEMS Conference, Tucson, pp. 517–518 (2008)
39. Wang, F., Li, X., Wang, Y., Feng, S.: Simultaneous formation of through wafer electrical interconnects and highly dense & uniform nickel tips for silicon-cantilever probe-cards. In: Transducers and Eurosensors – 4th International Conference on Solid-State Sensors, Actuators and Microsystems, 4300567, pp. 2051–2054 (2007)
40. Wang, F., Li, X., Feng, S.: Microcantilever probe cards with silicon and nickel composite micromachining technique for wafer-level burn-In testing. IEEE Trans. Adv. Packag **32**, 468–477 (2009)
41. Wang, F., Li, X., Cheng, R., Jiang, K., Feng, S.: Silicon cantilever arrays with by-pass metal through-silicon-via (TSV) tips for micromachined IC testing probe cards. Microelectron. Eng. **86**, 2211–2216 (2009)
42. Kim, B.H., Kim, J.B.: Design and fabrication of a highly manufacturable MEMS probe card for high speed testing. J. Micromech. Microeng. **18**, 075031 (2008)
43. Kim, B.H., Park, B.J., Kim, J.B.: Process effects of double step DRIE and Ni–Co electroplating for a trench-type cantilever probe for a fine-pitched MEMS probe card. Sens. Actuator. A: Phys. **152**, 252–260 (2009)
44. Kim, B.H., Kim, J.B.: Fabrication of a high aspect ratio thick silicon wafer mold and electroplating using flipchip bonding for MEMS applications. J. Micromech. Microeng. **19**, 065024 (2009)
45. Kim, B.H., Kim, J.B., Kim, J.H.: A highly manufacturable large area array MEMS probe card using electroplating and flipchip bonding. IEEE Trans. Ind. Electron. **56**, 1079–1085 (2009)
46. Huang, J.T., Lee, K.Y., Hsu, H.J., Wu, R.G., Tsai, T.C., Chen, C.K.: Fabrication technology of microelectro-mechanical systems probe chip compatible with standard complementary metal-oxide-semiconductor process. J. Appl. Phys. **49**, 06GN021 (2010)
47. Lee, K.Y., Huang, J.T., Hsu, H.J., Wu, R.G., Tsai, T.C.: Fabricating MEMS probe card compatible with CMOS-MEMS process. In: The 22nd International Microprocesses and Nanotechnology Conference, Sapporo, pp. 88–89 (2009)
48. Lee, K.Y., Huang, J.T., Hsu, H.J., Chen, C.K., Tsai, T.C.: Fabrication technology of CMOS-MEMS probe chip compatible with electroless nickel plating process. In: The 5th International Microsystems, Packaging, Assembly and Circuits Technical Conference, Taipei (2010)
49. Lee, K.Y., Huang, J.T., Hsu, H.J., Chen, C.K., Tsai, T. C.: Integration via structure with electroless nickel plating process to fabricate CMOS-MEMS probe chip. In: The 23nd International Microprocesses and Nanotechnology Conference (2010)

MEMS Microphone

► Electrostatic MEMS Microphones

MEMS Neural Probes

Jit Muthuswamy[1], Arati Sridharan[1] and Murat Okandan[2]
[1]School of Biological and Health Systems Engineering, Arizona State University, Tempe, AZ, USA
[2]Advanced MEMS and Novel Silicon Technologies, Sandia National Laboratories, Albuquerque, NM, USA

Synonyms

Brain implants; Microelectrodes; Neural interfaces; Prostheses

Definitions

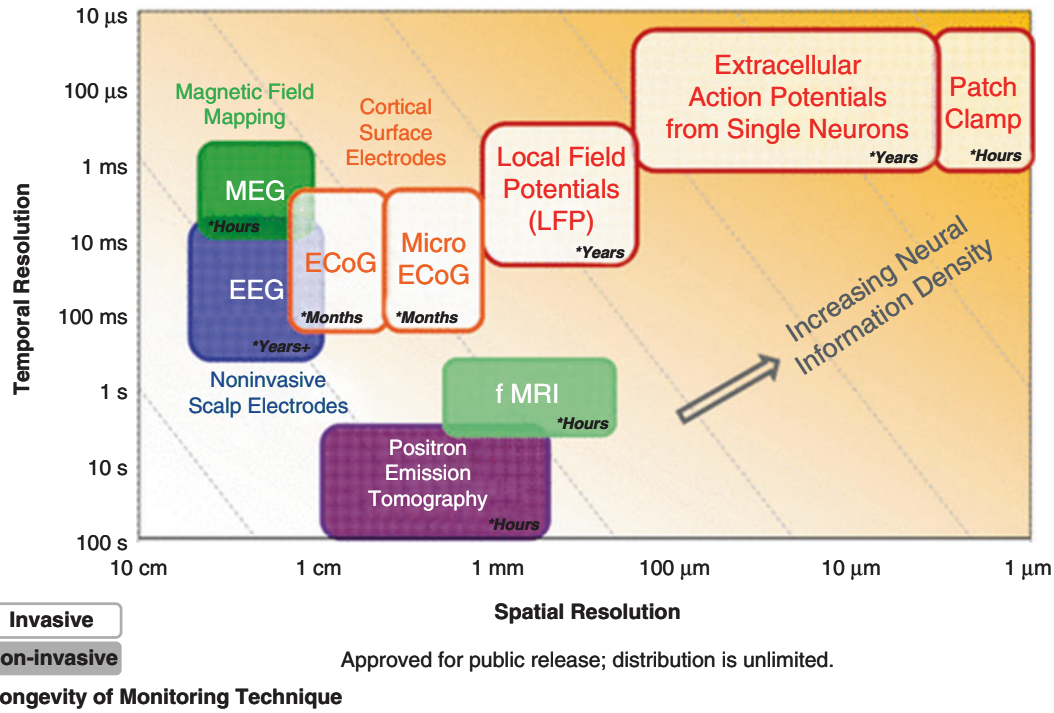
MEMS neural probes are microscale implants for the brain or peripheral nervous system that use

micro-electromechanical systems to assess function or stimulate activity in specific regions of the brain, ensembles of neurons, or neuronal fibers.

Brain Monitoring Technologies

Hypotheses regarding brain function or dysfunction are generally verified and confirmed using accurate measurements of brain activity at different scales – from whole brain regions to single cells or neurons and subcellular structures such as synapses, ion channels, etc. Such measurements are made using a variety of different techniques each with their own spatial and temporal resolution. The technologies are generally invasive or noninvasive; the noninvasive technologies include a whole spectrum of imaging modalities as illustrated in Fig. 1, electroencephalograms (EEG) from the scalp, etc.

The invasive technologies typically involve electrodes placed either on the brain (such as in electrocorticogram or ECoG) or microelectrodes implanted in specific brain regions of interest for obtaining electrical activity from the brain. This may be in the form of field potentials from ensembles of neurons or from extracellular action potentials of single or a few neurons in the vicinity of the microelectrode. Generally, electrical activity from single neurons has been a preferred currency for testing hypotheses about brain function. Over the last 10 years, exciting applications such as cortical prostheses have also been developed using electrical activity from single neurons in motor and premotor cortices of the brain. Success in this effort will lead to wider application of this theme in other brain prosthetic applications such as memory and cognitive prostheses. Neural probes of different kinds have been developed for a variety of applications such as electrical stimulation of specific neurons in deep brain stimulation (DBS) therapy, monitoring neurotransmitters and other neurochemicals, and delivering drugs. However, currently



MEMS Neural Probes, Fig. 1 Illustration of the spatio-temporal resolution of current brain monitoring technologies (Courtesy: Jack Judy, program manager, DARPA (MTO), approved for public release; distribution is unlimited)

micro-electromechanical systems (MEMS)-based technologies are primarily used for monitoring electrical activity from single neurons in mainstream research and emerging prosthetic technologies. Therefore, this chapter here will focus on neural probes used for monitoring the electrical activity of single neurons.

Implantable technologies to monitor electrical activity from single neurons have existed for over six decades. The earliest single neuronal recordings consisted of manually fabricated glass micropipettes that tapered down to few microns at one end. These pipettes were backfilled with electrolytes and a conducting wire was inserted. The tapered end was stereotactically positioned in the brain after a craniotomy in the vicinity of neurons of interest. They were initially used to record intracellular membrane potential changes [1, 2]. Insulated tungsten or platinum microelectrodes were used for recording extracellular activity [3, 4]. One of the first attempts to use such cumbersome manual approaches to obtain long-term recordings from animals was Strumwasser's bold effort in 1958 [5] to record activity of specific neurons in hibernating California ground squirrels. This was followed by several decades during which bundles of metal microwires in different configurations were manually assembled along with interconnects to record from ensembles of single neurons in the brain. The microwires were insulated along the length and the insulation at the tip of the microelectrode was etched a few tens of microns to allow an electrical interface with single neurons. The length (and consequently the electrical impedance) of the uninsulated tip or the size of the recording site determined the number of neurons that would be recorded. Microelectrodes developed using MEMS-based approaches have been a relatively recent development and have significantly evolved in complexity over the last 25 years. The development of the first neural probe using microfabrication technologies in 1970 [6] demonstrated the promise of MEMS-based systems to make neural probes more usable for uninitiated researchers. Using a manufacturing process that readily allowed batch fabrication, they dramatically improved scalability in terms of number of neurons that could be recorded and allowed possibilities for integrating multiple functionalities. Previously interconnects were manually soldered or glued. MEMS technology improved interconnect reliability. It also significantly reduced the overall form factor, which allowed animals to behave naturally.

General Features of Neural Probes

An individual neural probe must be slender (tens to hundreds of micrometers in diameter or in cross-sectional dimensions) and several millimeters long. This will allow interrogation of neurons in the cortical and deep brain structures while minimizing trauma and displacement of brain tissue during penetration and eventual implantation. Multiple probes are often integrated to form arrays of neural probes. Some studies [7, 8] have suggested that slender probes with smaller cross-sectional areas significantly decrease the reactive response of the brain tissue to the foreign body implant. However, the lower limit for slenderness appears to be that which will allow for successful penetration without bending or buckling of the probe. In addition, a smooth tapered probe is generally preferred to minimize dimpling of the meningeal layers such as the dura mater during penetration. However, the likelihood of such tissue dimpling is also dependent on the velocity of penetration and on the number of probes constituting an array. Therefore, design of the mechanical structure of the neural probe is constrained by requirements such as minimal tissue response, adequate stiffness for penetration, minimal brain tissue dimpling during penetration, etc.

One of the features of a probe used for recording from single neurons is the presence of one or more recording sites along the length of a probe. To ensure recording from one neuron or a few selected neurons, the recording sites are of similar geometrical scale as the neuronal cell bodies, which are typically 10–20 μm in diameter in the mammalian brain. Ion exchange across the cell membrane during a neuronal action potential results in current flow. The current flow in a volume conductor results in lines of electrical field intensities and spatial potential gradients being

established in the extracellular space. Microelectrodes in the extracellular space therefore register the spatial potentials (typically less than 1 mV) at any given point. The interface between the recording site (typically made of metals like platinum, platinum/iridium, gold, and also sometimes doped polysilicon) and the surrounding ion-rich extracellular media usually determines the overall electrical impedance of the probe. This impedance is with respect to a reference electrode remotely placed on or in the brain, or on the neck muscles depending on the application and the type of recording desired. The interface impedance is largely capacitive due to the spontaneous formation of an ionic double layer on the probe recording site surface upon insertion. However, faradic reactions take place at the recording site and give rise to a parallel pathway for charge transfer. This can be modeled as a resistive pathway in parallel to the double-layer capacitance.

Given the above general mechanical and electrical features of neural probes, two broad classes of MEMS micromachining techniques have been used successfully to fabricate slender neural probes and arrays of neural probes out of silicon wafers. These are bulk micromachining and surface micromachining. Silicon has been a preferred substrate material for a variety of reasons: (a) well-established micromachining methods, tools, and recipes are readily available, (b) silicon lends itself readily to subsequent integration of circuitry and other functionalities, (c) silicon has been shown to be nonreactive or inert as an implant material [9], eliciting minimal foreign body response. Directional, material, and dopant-dependent selectivity of anisotropic etching is used in both of the above classes to machine microscale mechanical structures in the form of slender neural probes. Besides these classes, MEMS neural probes have also been fabricated out of nonconductive polymers spun-coat on micromachined silicon or from metal molds. A wide variety of MEMS neural probes have been developed and tested. A few representative technologies that are broadly distinct and have been tested in long-term animal experiments are described here.

Bulk Micromachining of MEMS Neural Probes

The first MEMS-based neural probes were microscale shanks of silicon with metal recording sites along the length [10]. They were fabricated by bulk micromachining of silicon using anisotropic etching techniques. A boron-doped silicon substrate acted as an etch-stop and shanks of silicon 8–15 μm in thickness, 75–100 μm in width at the base that narrowed down to 25 μm and 1.5–3 mm length were machined in the initial versions. Each shank was capable of hosting multiple recording sites of gold with polysilicon or tantalum conductors. Currently these probes are fabricated and distributed by Neuronexus Inc., Ann Arbor, MI.

Another widely used silicon MEMS technology using bulk micromachining is a 3-D array of 10×10 microelectrodes that was first reported in 1991 by the University of Utah [11]. Aluminum (5 μm thick) was deposited on the polished surface of silicon wafers (1.7 mm thick) and then patterned into squares (325 μm sides and 400 μm separation between the centers of adjacent squares). A temperature gradient between the two sides of the wafers created a thermal migration of the eutectic droplets of aluminum and silicon with silicon recrystallizing behind the eutectic droplet trail. The net result was the creation of highly conductive trails of doped (p+) silicon within the bulk silicon wafer. Chemical and mechanical polishing of the wafer resulted in pillars of p+ doped silicon (approximately 1.5 mm long) isolated from its immediate neighbors by back-to-back p-n junctions that provided the isolation. The pillar tips were finally coated with platinum and insulated with polyimide after establishing interconnects at the back end. Bulk micromachining approaches have also been used successfully with other materials such as ceramic and diamond.

Surface Micromachining of MEMS Neural Probes

Surface micromachining allows the fabrication of neural probes that are generally thinner (2–5 μm

thick) than typical bulk micromachined probes (>8–15 μm thick). There is no evidence yet that a reduction in one of the dimensions of the neural probes leads to an improvement in performance. However, a significant reduction in the cross-sectional area of the neural probe has been shown to significantly improve the inflammatory and foreign body reaction of the tissue immediately surrounding the neural probe [7, 8]. There are other significant and unique advantages of surface micromachining over bulk micromachining such as the ability to integrate (a) complex mechanical structures such as microactuators, gears, etc. on the same die and (b) signal conditioning and control circuitry to significantly enhance the functionality of neural probes.

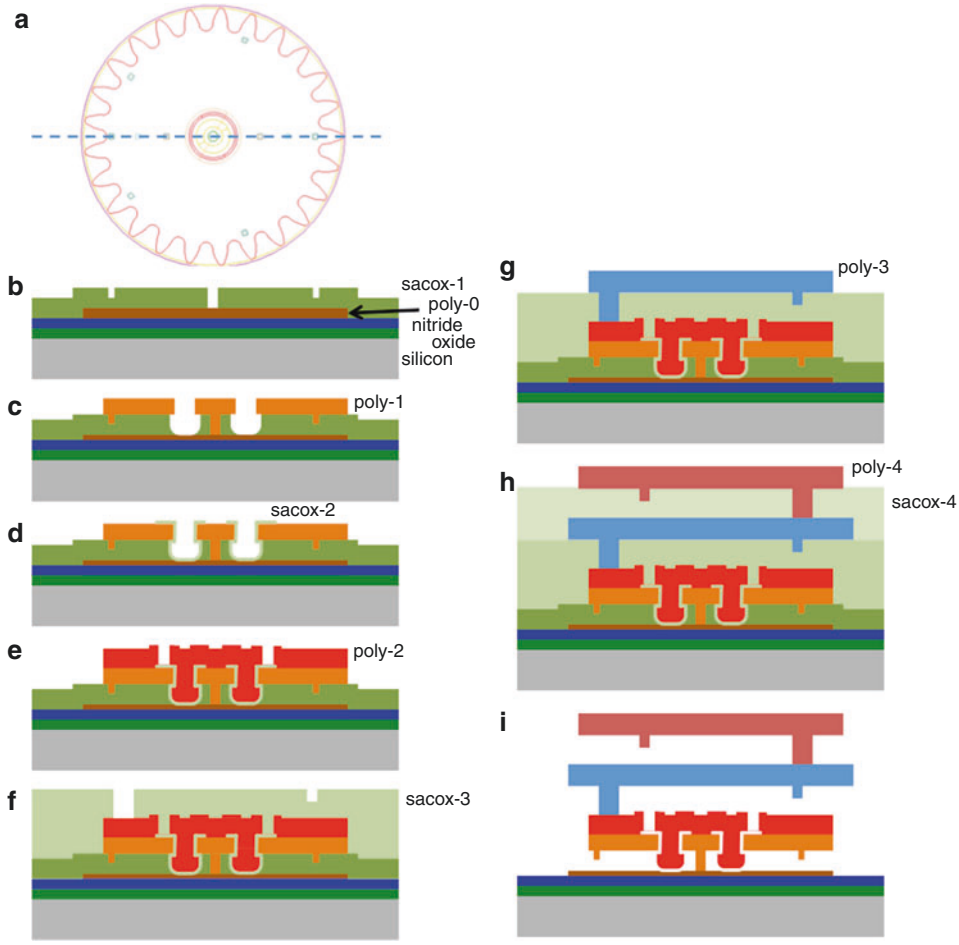
Complex mechanical structures such as microactuators and gears can potentially enhance the reliability and functionality of the neural probe in the development of autonomous neural probes. These are capable of adapting their position in the brain to compensate for any change in the quality of neuronal signal being recorded [12]. Such movable neural probes enhance the ability (a) to obtain well-isolated and stable single neuronal activity with high signal-to-noise ratio over longer durations, (b) to carefully and unambiguously monitor changes in the underlying neurobiology, (c) to monitor neurons and neuronal tracts of specific functional interest, and (d) potentially to enhance the reliability of brain prosthetic devices that require reliable operation over a life of the patient.

Innovative approaches have been used to develop neural probes using surface micromachining process. One involved the fabrication of a 3D array of polyimide pillars (1.5 mm long) with metal (Titanium on Aluminum) traces for three recording sites on each pillar, as opposed to the single recording site that was found in the Utah microelectrode array [13]. Polyimide film (10 μm) was spun coat over a 5 μm-thick film of Nickel electrodeposited on silicon. Nickel acted as a backing metal for the polyimide film. After patterning the metal interconnects and recording sites on the polyimide film, a second layer of polyimide (also 10 μm) was spun coat on top. Finally, and most interestingly, an externally applied magnetic field was used to lift the nickel backing layer and along with it the polyimide films. Similar surface micromachining approaches have been used to fabricate slender neural probes from other polymers such as benzocyclobutene (BCB).

One of the key challenges in using surface micromachining approaches to fabricate neural probes is to ensure that there be no residual stress in the thin films that constitute the substrate of the neural probe. Simultaneously, the thin films will also need sufficient stiffness to penetrate the outer meningeal layers and the brain itself and travel through the brain tissue without significant bending, buckling, or breaking. The five-layer polysilicon SUMMiTV^{TM} (*S*andia's *u*ltraplanar *m*ultilayer *m*icromachining *t*echnology) process of Sandia National laboratories, Albuquerque, NM, has been extremely useful in fabricating complex, surface micromachined, movable neural probes, and have been tested successfully in long-term animal experiments. Details of this process are explained in the remaining sections of this essay.

SUMMiT^{TM} Surface Micromachining Process

A brief summary of the SUMMiT process flow is presented here; more detailed process description can be found elsewhere [14]. An example of a multilayer mechanical structure that can be formed in this process is shown in Fig. 2a, a gear with close-tolerance hub and four mechanical layers. SUMMiT surface micromachining process is carried out on 150 mm, (100), 2–20 Ω.cm, *n*-type silicon wafers. First steps are a thermal oxidation (0.6 μm) and low-stress silicon nitride deposition (low-pressure chemical vapor deposition – LPCVD, 0.8 μm), which form the insulating stack for the following process steps. An etch through the insulating stack allows electrical connection between polysilicon layers and the substrate. First polysilicon layer (poly-0) deposited by LPCVD is 0.3 μm thick; this layer usually serves as a ground plane, mechanical anchor, electrostatic shield, and part of the



MEMS Neural Probes, Fig. 2 Plan view and cross section of multi-layer gear illustrating the process technology sequence for fabrication

electrical interconnect. All of the polysilicon layers in the process flow are heavily *n*-type doped, providing the required electrical conductivity.

Next layer is a 2 μm LPCVD sacrificial oxide layer (sacox-1). Two separate etch steps form the features for mechanical anchors, which are also electrical connections and dimples (Fig. 2b). The partial etch of oxide allows the dimples to be formed in the following polysilicon layer. This helps in preventing stiction by allowing relatively small areas to come into contact rather than the full surfaces that might be available in the structures. The following polysilicon layer (poly1) deposited by LPCVD is 1 μm thick. Mechanical features in this layer can be defined by lithography and etching at this step or also by a following etch after the next polysilicon layer, which generates features that are a laminate of the two layers. Close tolerances needed for forming hubs, captive mechanical features, sliding seals, and pressure/flow sections are formed by deposition and patterning of a conformal, thin (0.5 μm) silicon dioxide layer (sacox-2) after an isotropic oxide etch through lithographically defined and etched regions in poly-1 (Fig. 2c–d). Poly-2 layer (1.5 μm) is then deposited and patterned. In areas where sacox-2 was remaining, a close-fitting and mechanically free linkage is formed between poly-1 and poly-2 layers (Fig. 2e). Sacox-3 is then deposited and chemimechanically polished (CMP) which removes the topography that was a result of the

preceding etch and deposition steps. The final thickness of sacox-3 above poly-2 is 2 μm. The dimple and anchor etch steps prepare the dimples and mechanical connections between poly-3 and poly-2 (Fig. 2f). Poly-3 is then deposited (2.25 μm) and patterned (Fig. 2g). Following oxide layer (sacox-4) is deposited by LPCVD and planarized with a CMP step to a thickness of 2 μm above poly-3. After the dimple and anchor etch steps, final mechanical polysilicon layer (poly-4, 2.25 μm) is deposited and patterned (Fig. 2h). An aluminum metallization layer is the last layer that is deposited and patterned on poly-4 features to enable wire-bonding and low-resistance electrical connections. The release step is performed in an HF containing chemistry to remove all sacrificial oxide layers while preserving the polysilicon and metal features (Fig. 2i). In order to prevent stiction, a supercritical CO_2 drying step is used following the release etch to displace all liquids between layers without exposure to a gas-liquid interface. The release can be performed at die level or as full wafers, depending on the design and application requirements.

Further modifications of the SUMMiT process are also available, such as SwIFT [15, 16] and SFET [17, 18]. SwIFT process adds two silicon nitride layers (0.3 μm thick nitride-1 on poly-0 and 0.8 μm thick nitiride-2 below poly-3), a deep oxide cut (from poly-3 to poly-0, filled with nitride-2), and Bosch etch through the substrate to enable the creation of insulator-lined microfluidic channels, fluidic connections, and other novel features. SFET integrates NMOS transistors in the silicon substrate, which allows pre-amps and addressing circuitry to be fabricated for signal conditioning and dense-matrix addressing.

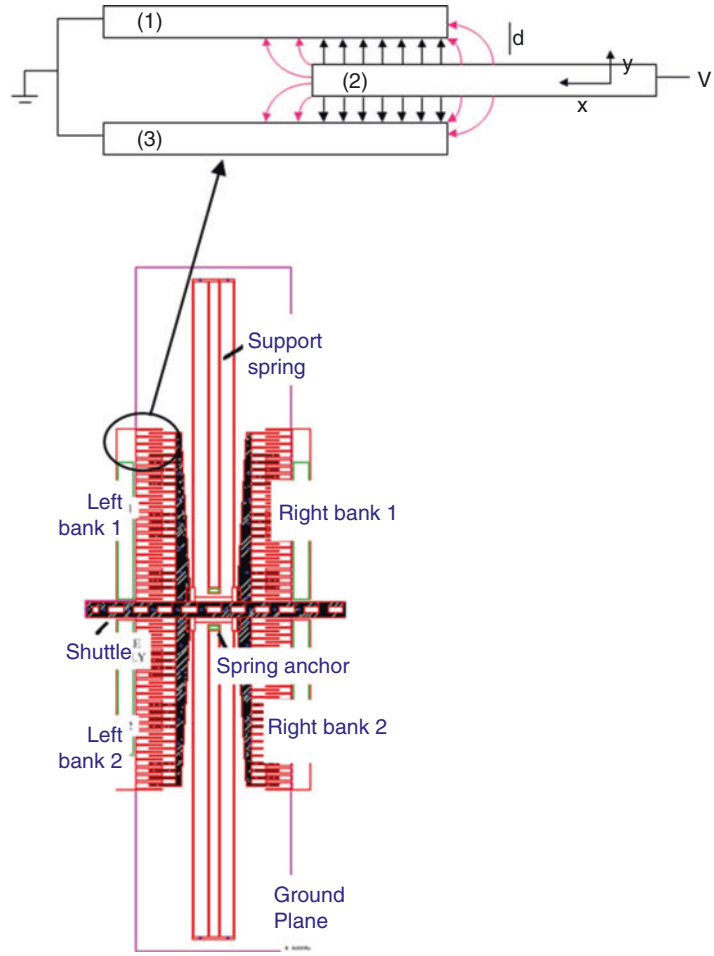
The neural probe in the form of a long, slender beam of doped ($10^{21}/cm^3$) polysilicon 7 mm long, 50 μm wide and approximately 5 μm thick is defined in the poly-1, poly-2, poly-3, and poly-4 layers of the SuMMITV™ process. Interestingly, the nonlinear I-V characteristics of doped polysilicon has an inflection point in the same range of voltages that typically separate noise signal amplitudes (10–30 μV peak-to-peak) from action-potential signal amplitudes (typically >100 μV) [14]. The above property, remarkably, results in selective suppression of noise when recording extracellular potentials from the brain [14]. The polysilicon neural probe has been demonstrated to produce significantly higher signal-to-noise ratios compared to commonly used tungsten neural probes.

MEMS Neural Probes with Electrostatic Microactuators

Integrated electrostatic microactuators have been developed for moving neural probes in the brain using comb drives. Electrostatic forces due to a fringing electric field between a movable plate and stationary plate upon application of a voltage (90–110 V) between the two plates creates a movement in one of the plates. A bank of such parallel plates forming a comb drive is used to create translation (∼30–40 μm) in the x-direction as shown in Fig. 3.

A similar arrangement is used to create translation in the y-direction. The above two systems are driven by voltages that are 90° phase shifted and the translations are coupled to a pin joint that rotates a flywheel as shown in Fig. 4. The direction of rotation of the flywheel is determined by which of the two drives (x- or y-drive) is leading in phase. The flywheel is coupled to the microelectrode either directly or through gears to move the microelectrodes in steps of 1 μm over a distance of 5 mm. The above system of electrostatically actuated microelectrodes has been successfully tested in acute rodent experiments [15] but was found to be prone to mechanical failure. There were stiction issues between the parallel plates of the comb drive, alignment issues between the gears and the microelectrode that resulted in the microelectrode getting stuck particularly during intense behavioral routines of the rodent. Besides such serious mechanical robustness issue, the electrostatic microactuators also required high voltages (90–110 V) and optimally designed voltage waveform shapes.

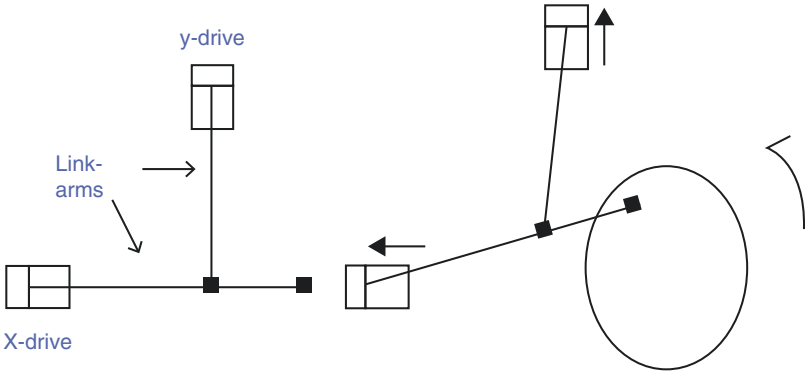
Electrothermal microactuators have, on the other hand, been demonstrated to be more robust and occupy smaller footprint on the die [16, 17]. In addition, the waveforms required to drive the electrothermal actuators are 6–10 V and standard rectangular pulses that can be



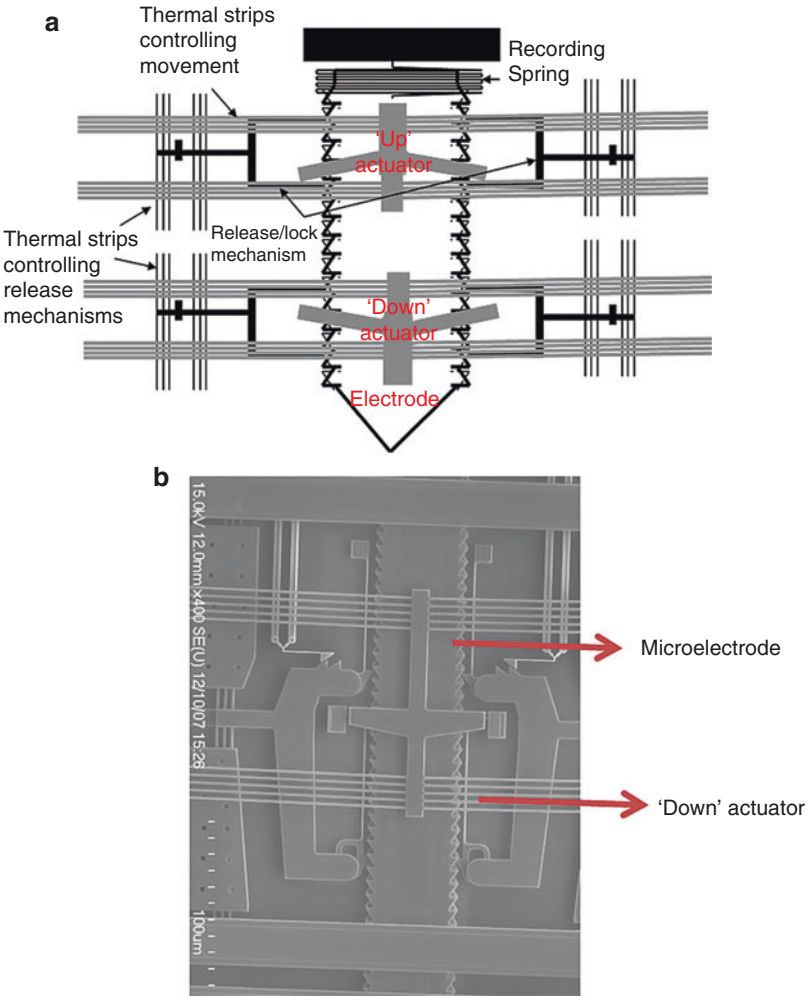
MEMS Neural Probes, Fig. 3 Schematic of the comb drive actuator and a close-up of one of the parallel plates and the associated fringe field intensity lines that make up the comb drive

readily generated. At the core of electrothermal microactuators are the electrothermal strips that consist of parallel beams of doped polysilicon that expand upon the application of voltage, displacing a shuttle in the middle of the electrothermal strips as illustrated in Fig. 5. Each microelectrode was designed with one set of actuators to move the microelectrode up, one set to move the microelectrode down, two sets of microactuators to release the lock that prevents the microelectrode from moving up, and finally two sets of microactuators to release the lock that prevents the microelectrode from moving down. The above neural probes had a displacement resolution of approximately 9 μm over a total displacement capability of over 5 mm. The electrothermal microactuators and the integrated neural probes have been successfully tested in long-term rodent experiments lasting over 13 weeks [16].

The electrothermal microactuators are operated on a bent-beam actuation design known as the chevron-type mechanism. The first generation of chevron-type MEMS actuators utilized a complex latch-based release mechanism for controlling the bidirectional movement of the electrode. Long-term in vivo rodent testing of these first-generation actuators revealed the need to improve the (a) reliability of bidirectional control of electrode movement due to out-of-plane errors in the latch mechanism and (b) force output for moving the microelectrode in a viscoelastic medium like brain tissue in the presence of inflammation and subsequent immune response. The second-generation chevron-type MEMS device address the above key issues (Fig. 6)

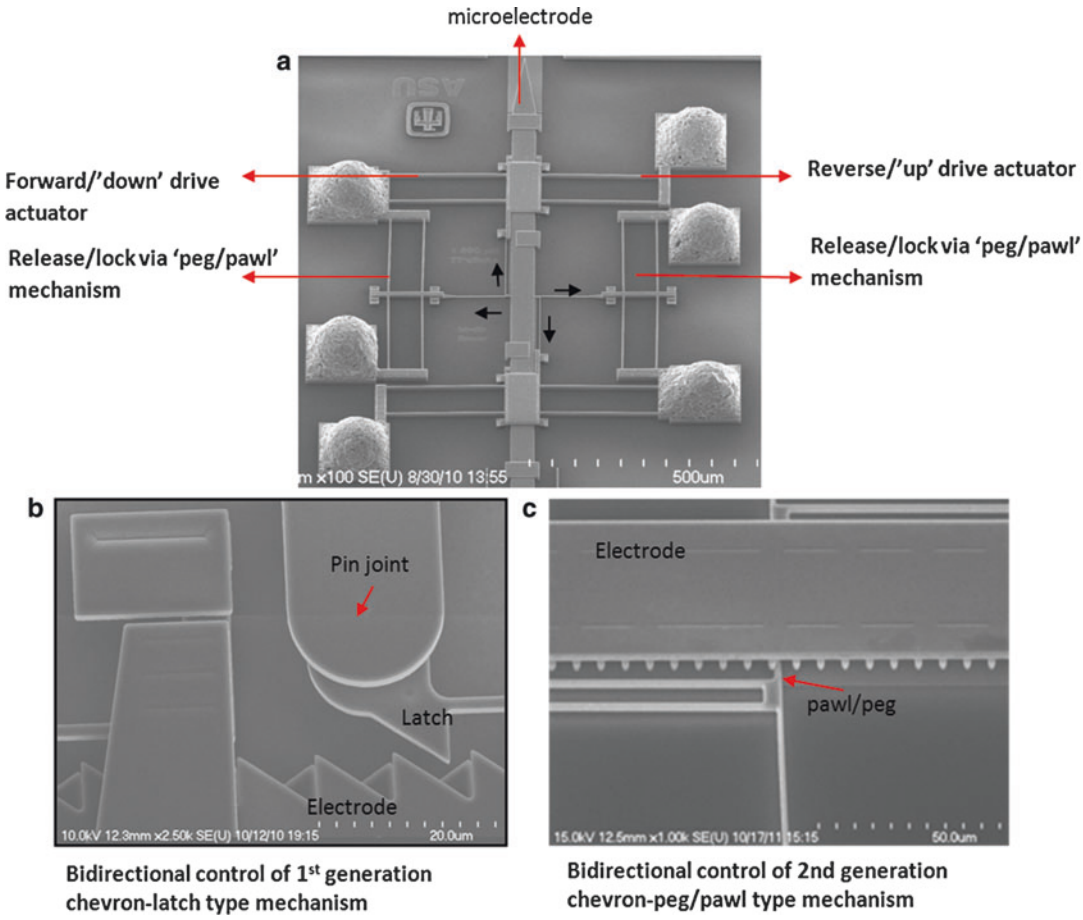


MEMS Neural Probes, Fig. 4 Schematic of the x- and y-drive connected to a pin joint and a fly-wheel. The x- and y-drives are activated at 90° phase difference to produce a net rotational motion of the fly-wheel that will eventually drive a gear or the microelectrode



MEMS Neural Probes, Fig. 5 (**a**) Schematic of a microelectrode and six electrothermal actuators associated with the microelectrode. The "Up" and "Down" actuators move the microelectrode in the center up or down respectively. (**b**) SEM of the one of the actuators showing the electrothermal heat strips and ratchets that move the microelectrode

[18]. The improved design simplified the actuation mechanism for the bidirectional movement by placing "upward" and "downward" movement actuators on either side of the electrode. To prevent movement in the opposite direction of desired direction, a "peg/pawl" was added on either of the electrodes. The new design improved the cross-sectional area upon which the MEMS



MEMS Neural Probes, Fig. 6 (**a**) SEM of the electrode actuated by the Chevron-peg/pawl type of mechanism with a pair of release/lock actuators and "up"/"down" actuators placed opposed to each other. The microelectrode has maximum displacement of 5 mm in steps of about 6.5 μm. (**b**) SEM of the 1st generation chevron-latch type mechanism controlling bidirectional movement. High precision is required as the distance between "up" and "down" teeth is 2–4 μm. (**c**) SEM of the 2nd generation chevron-peg/pawl type mechanism with contact between the pawl/peg and a single row of teeth on the microelectrode

electrothermal actuator could impose its thrust, generating more force per stroke. The comparative advantages of the new design include (a) finer displacement resolution (6.5 μm), (b) increased force output (upto 1 mN per stroke), (c) upto 76 % less power consumption, and (d) increased movement reliability up to to four million cycles.

However, the key challenge in the application of MEMS neural probes in brain prosthetic technologies as well as technologies to monitor single neurons for neurobiological applications has been the unreliable interface between the implanted microelectrode and the neurons in the brain tissue surrounding the microelectrode. Current technologies do not have the stability, consistency, or reliability to record from targeted single neurons for long periods of time. The MEMS-based movable microelectrodes with their flexibility to be repositioned offer a very promising approach to overcome the above challenges.

MEMS with Soft Interfaces

As the next generation of MEMS for brain-machine interfaces evolve, several

MEMS Neural Probes, Table 1 Typical elastic moduli of nervous tissues

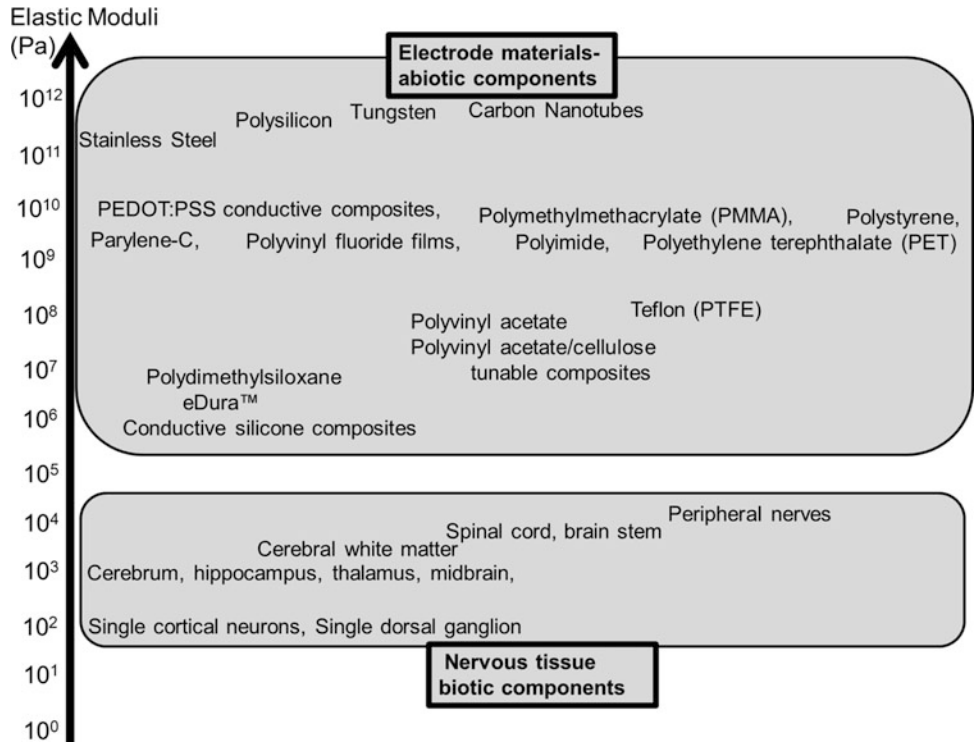
Type of nervous tissue	Elastic modulus	References
Single cortical neurons	0.2 kPa	[19]
Single dorsal root ganglion	0.9 kPa	[19]
Fresh peripheral nerves (Sciatic)	20–40 kPa	[20]
Acellular sciatic nerve	576 kPa	[21]
Spinal cord with meninges	8–36 kPa	[22]
Spinal dura mater	1.2 MPa	[23]
Cranial dura mater	420 kPa	[23]
Cerebral white matter	12–15 kPa	[24]
Cerebral gray matter	4–6 kPa	[24]
Cerebrum	1.6–11.4 kPa	[25]
Hippocampus	0.8–4.2 kPa	[25]
Caudate/ Putamen	0.51–3 kPa	[25]
Thalamus	2.8–3.0 kPa	[26]
Midbrain	2.9–3.1 kPa	[26]
Brain stem	2.9–20.52 kPa	[27] (assumed 80 % stiffer than cerebrum)

unconventional materials are currently being explored to enhance long-term reliability and performance. One such effort is directed toward the development of soft neural interfaces that match the stiffness of the surrounding brain tissue. Typical elastic moduli of nervous tissue from the periphery to the spinal cord and cortex are summarized in Table 1. The peripheral nerves tend to be slightly stiffer (20–40 kPa) compared to central nervous system tissues (1–20 kPa). The central nervous system also displays a wide range in elastic moduli in terms of regional variations (i.e., white matter – 12–15 kPa vs. gray matter – 4–6 kPa), in addition to differential anisotropy. Nevertheless, it is clear that the biotic component of the neural interface such as the brain tissue is approximately 2–3 orders of magnitude softer or less stiff compared to the abiotic component of the interface comprising of conventional materials used in neural devices (Fig. 7).

Abiotic components, such as conductive wires used in recording (i.e., tungsten, stainless steel, polysilicon) range 100–400 GPa in elastic modulus. In addition, brain tissue is typically viscoelastic in nature compared to stiffer electrode and insulation components. Therefore, there is a mechanical mismatch between the material properties of the brain and commonly used materials at the neural interface. It is widely hypothesized that this mechanical mismatch between the abiotic and biotic components of the neural interface aggravates inflammatory response and glial scarring that typically surround neural probes under chronic implantation conditions.

To make neural probes more mechanically biocompatible, several groups have designed softer materials for interfacing with the brain. A key design parameter is to increase the compliance by using elastomeric materials as a base for conductive thin films. This strategy has been successfully utilized for peripheral nerve, EEG, and ECoG applications [28–31]. Ultrathin, flexible materials with thin films of gold were used to conform to the unique surface topology of the ear [29]. Silicon nanomembrane "ribbons" were deposited on top of a highly flexible, elastomeric substrate to conform to the undulating gyri of the cortex [28].

Another common strategy has been to develop soft conductive composites using elastomeric materials. Using a composite of MED6607 silicone material and poly (3,4-ethylenedioxythiophene) (PEDOT)-based conductive polymer, a "soft" neural interface with an elastic modulus of 974 kPa was developed that showed higher neuronal attachment and reduced inflammation [32]. Elastomeric composites of poly-dimethyl siloxane (PDMS) (15–20 %) and platinum particles (with a composite elastic modulus of 10 MPa) have been tested for peripheral nerve interfaces [33, 34]. A PDMS-gold composite (2 MPa elastic modulus) has been developed that can be used as stretchable, elastomeric films for in vitro applications [35].



MEMS Neural Probes, Fig. 7 A diagram showing the elastic moduli of common materials used in neural interface applications (abiotic components) and nervous tissue (biotic components). Abiotic components such as electrodes, insulation, etc. are 2–4 orders higher in magnitude compared to biotic components, reflecting the mechanical mismatch found at the neural interface

A third, innovative strategy has been to utilize materials having tunable mechanical properties. An advantage of using materials with tunable properties is that stiffer materials have better penetration properties during insertion into the brain. Upon insertion into the brain, the material adjusts to a softer elastic modulus, leading to better biocompatibility with the nervous tissue. An example of a mechanically dynamic material is a polyvinyl acetate/cellulose-based nanocomposite inspired by the tunable stiffness properties of sea cucumbers [36]. Cellulose-based microwhiskers tune the softness of the composite material within 300 s in a hydrated environment from ~80GPa to ~12 MPa. When implanted in rodent brain models, these materials show reduced inflammation and increased biocompatibility [37, 38]. A conductive composite with Ti/Au micromachined in the cellulose-based material has been demonstrated as a potential neural probe [39]. Another example of a material with tunable mechanical properties is the thioacrylate-based shape memory alloy. These shape memory polymers modulate their material properties from 1 GPa to 18 MPa after insertion [40]. The thioacrylate-based material can be incorporated with conductive elements such as gold with platinum black or iridium for novel probes for the nervous system [40, 41]. In addition, the materials are conducive to photolithography techniques, leading to potentially innovative MEMS-based solutions for conformal and biocompatible neural devices.

Emerging High-Density Neural Probes

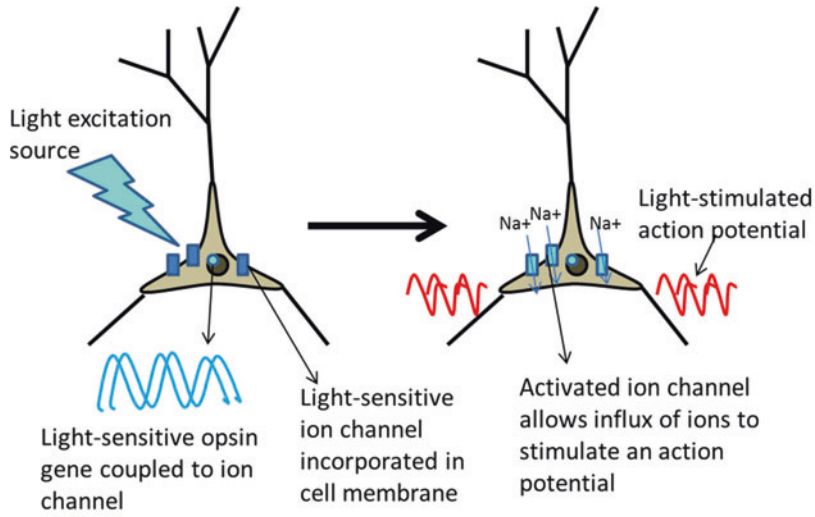
It is estimated that there are >100,000 neurons/mm^3 in the brain [42]. There is a significant need to map function in neural circuits in the human

brain across scales – from events at the molecular scale to microscale to macroscale. The idea of mapping the human brain "connectome" requires cutting-edge tools that have the capacity to record from the maximum neurons possible with high spatial and temporal resolution [43]. This ambitious goal remains one of the holy grails of neural interfaces alongside the goal of developing reliable, long-term interfaces. High-density neural probes are needed to monitor local populations across multiple anatomical regions for better understanding neuronal dysfunction and disease pathology. High-density EEG (up to 256 electrodes) and ECoG arrays (250–360 channels) are increasingly used for neural disease applications such as epilepsy and sleep disorders [28, 44, 45]. Current state-of-the-art intracortical technologies can monitor tens to hundreds of single neurons in parallel [46]. High-density Utah electrodes with the capacity of 25 electrodes/mm^2 have been reported [47]. Recent advances in CMOS and other silicon-based technologies have enabled massively parallel devices with upward of 10,000 electrodes [42]. In vitro CMOS-based systems can go up to 65,536 electrode sites (1536 electrode/mm^2), enabling highly precise subcellular to network-level resolution of neural signals [48–51]. Significant advantages using CMOS-based technologies includes monolithic integration of the (a) electrode recording readout, (b) signal processing, (c) in situ amplification, and (d) data transfer processes [52]. Key challenges in the fabrication of high-density neural probes include trade-offs with the number of electrode sites, device width and geometry, and scalability and reliability of interconnects. Innovative solutions such as double-sided fabrication of electrodes on either side of the planar substrate and using the e-beam technique to reduce lead width and separation to submicron dimensions (~290 nm) pave the way for high-density, high-resolution neural probes [53, 54]. Using hybrid e-beam lithography and conventional photolithography techniques, a highly dense electrode array (1000 electrodes) in an area of 10 × 10 μm has been fabricated and tested in animals [55]. Additional strategies for obtaining high resolution is via incorporation of silicon and other nanowires [56, 57]. These nanowire-based electrodes can serve a dual purpose of stimulation and obtaining "gold standard" intracellular recording via harmless penetration of cells.

Neural Interfaces with Integrated Optoelectronics

Neural interfaces with integrated optoelectronics are critical to serve the needs of the exciting, emerging discipline of optogenetics. Optogenetics utilizes genetic engineering techniques to genetically modify a neuron and render it light sensitive (Fig. 8) opening up the exciting possibility of stimulating the neuron using light. The light-sensitive gene typically encodes for either a cation (i.e., Na^+, H^+) or anion (Cl^-) ion channel coupled to a microbially derived, light-sensitive gene (i.e., rhodopsins). Light-sensitive genes that incorporate a biochemical modulation method (i.e., G-protein coupled receptors, cAMP modulation) for manipulating the excitatory or inhibitory response have also been demonstrated [58]. Upon excitation with an appropriate wavelength and light intensity, the ion channel pumps in the ion either hyperpolarizing or depolarizing the cell. This action leads to increased or decreased generation of action potentials and hence affects neural activity at the cellular level. Sodium (Na^+) and calcium (Ca^{2+}) ions typically have excitatory effects, while chloride (Cl^-) and proton (H^+) ions have inhibitory effects on the neural activity. There are numerous genetic constructs that have been developed and offer a wide spectrum of selectivity (350–1050 nm) [59, 60]. Typical light intensities required to stimulate a neuron are 1–5 mW/mm^2, though recent advances have increased the relativity sensitivity of the gene constructs to ~0.1 mW/mm^2 [46, 59]. Nevertheless, optogenetics offers a unique method to perturb a specific neural circuit in brain-mapping studies.

Key technological challenges in developing MEMS-based solutions for the above needs are in (a) developing scalable, low-power approaches for light source and (b) delivering light to localized neuronal populations. Typical light sources include high-intensity lasers and light-emitting



Light-sensitive gene family	Excitatory or Inhibitory	Ion channel sensitivity
Channelrhodopsin	Excitatory	Na+/ cations
Halorhodopsin	Inhibitory	Cl-
Bacteriorhodopsin	Inhibitory	H+

MEMS Neural Probes, Fig. 8 Illustration of the principle of optogenetic stimulation. A neuron is genetically modified by the addition of a light-sensitive gene. Upon excitation from a light source, the excited ion channel pumps ions into the cell. Depending on the type of ion pumped in, the cell then either hyperpolarizes or depolarizes. In the illustrated example, an excitatory response is depicted, leading to increased action potential generation. The ion channel typically comes from one of three classes of light-sensitive microbial genes: channelrhodopsin, halorhodopsin, bacteriorhodopsin

diodes (LEDs). Most systems utilize a laser coupled with optical fiber-based delivery method. However, LEDs have an advantage in that they utilize less power and can be integrated in neural interfaces readily. Recent advances in LEDs for optogenetics include utilizing multicolor diodes that were coupled to fiberoptics mounted on micromachined grooves on silicon shanks [61, 62]. This innovation allowed for multisite targeting and perturbation of neural circuits at different points. Another interesting innovation is the development of wirelessly powered, injectable GaN-based μLEDs (6.45 μm thick) that could potentially be placed inside deep brain structures of a rodent with relatively better biocompatibility compared to fiberoptics [63]. GaN μLEDs were fabricated using conventional epitaxial methods and microtransfer printed onto microneedles [64]. These μLEDs are highly versatile with independently addressable, inorganic LEDS with variable-wavelength emissions that implanted in the brain for precise, localized stimulation or used on stretchable, conformal substrates for ECoG applications [65]. Their modular nature allows for ready incorporation with multiple electronic sensors and MEMS devices. An example of a complete, injectable, multifunctional optoelectronic system is ~20 μm thick [63]. The other rapidly growing area of innovation is in developing novel optical fiber-based waveguide systems for precise and spatially selective delivery of light in a large-scale manner. Strategies for spatial targeting include glass-sharpened optrodes [61]. Another strategy is to etch the cladding at various points using a focus-ion beam for targeting neuronal populations along the length of the optrode [66, 67]. An alternative strategy is to use individual waveguides that were addressed by a matrix of micromirrors separately coupled to different light sources, allowing optical stimulation at each point with tunable wavelength and intensity [68]. Optrode arrays, such as the 3D optrode array, have also been fabricated using micromachined quartz and fused silica for optogenetic and IR application [69]. A completely microfabricated μLED coupled optrode array

(32 channel) with microneedle waveguides, a switched-capacitor stimulator, and wireless telemetry on a single platform [70] was demonstrated recently. Based on the rapid advances in optogenetic-related technologies, the development of scalable, high-throughput, flexible, and efficient optical stimulation-coupled recording electrodes is emerging as an innovative front in neural probe designs.

Future Directions

While MEMS-based neural probes have already made a significant impact on the field of neural monitoring, the immediate interface between the neural probe and the brain tissue is a key challenge that is poorly understood. So it is important to direct future efforts toward a better understanding of the events at the interface over the lifetime of the implant. On the device side, future efforts are likely to be focused on developing technologies that will allow us to scale up the number of probes (both movable and fixed). Such high-density probes are essential for monitoring large ensembles of neurons simultaneously. Integrating wireless telemetry to the above probes will also be critical to allow monitoring from the brains of awake, behaving subjects besides enhancing the long-term reliability of the probes. Finally, integrating multiple functionalities such as the ability to perform neurochemical monitoring, deliver drugs, stimulate the brain tissue electrically and/or optically will dramatically enhance our ability to assess brain function in a multimodal framework.

Cross-References

- ▶ Basic MEMS Actuators
- ▶ Biosensors
- ▶ CMOS MEMS Biosensors
- ▶ Epiretinal Prosthesis
- ▶ MEMS Packaging
- ▶ Nanorobotics
- ▶ Nanorobotics for Bioengineering
- ▶ Organic Bioelectronics

References

1. Cole, K.S., Curtis, H.J.: Electric impedance of the squid giant axon during activity. J. Gen. Physiol. **22**, 649–670 (1939)
2. Ling, G., Gerard, R.W.: The normal membrane potential of frog sartorius fibers. J. Cell. Physiol. **34**, 383–396 (1949)
3. Dowben, R.M., Rose, J.E.: A metal-filled electrode. Science **118**, 22–24 (1953)
4. Hubel, D.H.: Microelectrode for recording from single units. Science **125**, 549–550 (1957)
5. Strumwasser, F.: Long-term recording from single neurons in brain of unrestrained mammals. Science **127**, 469–470 (1958)
6. Wise, K.D., Angell, J.B., Starr, A.: An integrated-circuit approach to extracellular microelectrodes. IEEE Trans. Biomed. Eng. **17**, 238–247 (1970)
7. Seymour, J.P., Kipke, D.R.: Neural probe design for reduced tissue encapsulation in CNS. Biomaterials **28**, 3594–3607 (2007)
8. Stice, P., Gilletti, A., Panitch, A., Muthuswamy, J.: Thin microelectrodes reduce GFAP expression in the implant site in rodent somatosensory cortex. J. Neural Eng. **4**, 42–53 (2007)
9. Stensaas, S.S., Stensaas, L.J.: Histopathological evaluation of materials implanted in the cerebral cortex. Acta Neuropathol. **41**, 145–155 (1978)
10. Najafi, K., Wise, K.D., Mochizuki, T.: A high-yield IC-compatible multichannel recording array. IEEE Trans. Electron Devices **ED-32**, 1206–1211 (1985)
11. Campbell, P.K., Jones, K.E., Huber, R.J., Horch, K.W., Normann, R.A.: A silicon-based, three-dimensional neural interface: manufacturing processes for an intracortical electrode array. IEEE Trans. Biomed. Eng. **38**, 758–768 (1991)
12. Muthuswamy, J., Anand, S., Sridharan, A.: Adaptive movable neural interfaces for monitoring single neurons in the brain. Front. Neurosci. **5**, 94 (2011)
13. Takeuchi, S., Suzuki, T., Mabuchi, K., Fujita, H.: 3D flexible multichannel neural probe array. J. Micromech. Microeng. **14**, 104–107 (2004)
14. Saha, R., Jackson, N., Patel, C., Muthuswamy, J.: Highly doped polycrystalline silicon microelectrodes reduce noise in neuronal recordings in vivo. IEEE Trans. Neural Syst. Rehabil. Eng. **18**, 489–497 (2010)
15. Muthuswamy, J., Okandan, M., Jain, T., Gilletti, A.: Electrostatic microactuators for precise positioning of neural microelectrodes. IEEE Trans. Biomed. Eng. **52**, 1748–1755 (2005)
16. Jackson, N., Sridharan, A., Anand, S., Baker, M., Okandan, M., Muthuswamy, J.: Long-term neural recordings using MEMS based movable microelectrodes in the brain. Front. Neuroeng. **3**, 10 (2010)
17. Muthuswamy, J., Okandan, M., Gilletti, A., Baker, M. S., Jain, T.: An array of microactuated microelectrodes for monitoring single-neuronal activity in rodents. IEEE Trans. Biomed. Eng. **52**, 1470–1477 (2005)

18. Anand, S., Sutanto, J., Baker, M., Okandan, M., Muthuswamy, J.: Electrothermal microactuators with peg drive improve performance for brain implant applications. J. Microelectromech. Syst. **21**, 1172–1186 (2012)
19. Spedden, E.W., James, D., Naumova, E.N., Kaplan, D. L., Staii, C.: Elasticity maps of living neurons measured by combined fluorescence and atomic force microscopy. Biophys. J. **103**, 868–877 (2012)
20. Rosso, G., Liashkovich, I., Gess, B., Young, P., Kun, A., Shahin, V.: Unravelling crucial biomechanical resilience of myelinated peripheral nerve fibres provided by the Schwann cell basal lamina and PMP22. Sci. Rep. **4**, 7286 (2014)
21. Borschel, G.H., Kia, K.F., William, J., Kuzon, M., Dennis, R.G.: Mechanical properties of acellular peripheral nerve. J. Surg. Res. **114**, 133–139 (2003)
22. Zhang, H., Falkner. P., Cai, C.: In-vivo indentation testing of sheep spinal cord with meninges. In: Annual Conference on Experimental and Applied Mechanics, Costa Mesa, CA, pp. 99–104 (2015)
23. Maikos, J.T., Elias, R.A.I., Shreiber, D.I.: Mechanical properties of dura mater from the rat brain and spinal cord. J. Neurotrauma **25**, 38–51 (2008)
24. Kruse, S.A., Rose, G.H., Glaser, K.J., Manduca, A., Felmlee, J.P., Jr, C.R.J., et al.: Magnetic resonance elastography of the brain. Neuroimage **39**, 231–237 (2008)
25. Lee, S., King, M., Sun, J., Xie, H., Subhash, G., Sarntinoranont, M.: Measurement of viscoelastic properties in multiple anatomical regions of acute rat brain tissue slices. J. Mech. Behav. Biomed. Mater. **29**, 213–224 (2014)
26. van Dommelen, J.A.W., van der Sande, T.P.J., Hrapko, M., Peters, G.W.M.: Mechanical properties of brain tissue by indentation: interregional variation. J. Mech. Behav. Biomed. Mater. **3**, 158–166 (2010)
27. Arbogast, K.B., Margulies, S.S.: Material characterization of the brainstem from oscillatory shear tests. J. Biomech. **31**, 801–807 (1998)
28. Viventi, J., Kim, D.-H., Vigeland, L., Frechette, E.S., Blanco, J.A., Kim, Y.-S., et al.: Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. Nat. Neurosci. **14**, 1599–1607 (2011)
29. Norton, J.J.S., Lee, D.S., Lee, J.W., Lee, W., Kwon, O., Won, P., et al.: Soft, curved electrode systems capable of integration on the auricle as a persistent brain–computer interface. Proc. Natl. Acad. Sci. U. S. A. **112**, 3920–3925 (2015)
30. Guo, L., Guvanasen, G.S., Liu, X., Tuthill, C., Nichols, T.R., DeWeerth, S.P.: A PDMS-based integrated stretchable microelectrode array (isMEA) for neural and muscular surface interfacing. IEEE Trans. Biomed. Circuits Syst. **7**, 1–10 (2013)
31. McClain, M.A., Clements, I.P., Shafer, R.H., Bellamkonda, R.V., LaPlaca, M.C., Allen, M.G.: Highly-compliant, microcable neuroelectrodes fabricated from thin-film gold and PDMS. Biomed. Microdevices **13**, 361–373 (2011)
32. Kolarcik, C.L., Luebben, S.D., Sapp, S.A., Hanner, J., Noah, S., Kozai, T.D.Y., et al.: Elastomeric and soft conducting microwires for implantable neural interfaces. Soft Matter **11**, 4847 (2015)
33. Minev, I.R., Musienko, P., Hirsch, A., Barraud, Q., Wenger, N., Moraud, E.M., et al.: Electronic dura mater for long-term multimodal neural interfaces. Science **347**, 159–163 (2015)
34. Minev, I.R., Wenger, N., Courtine, G., Lacour, S.P.: Platinum-elastomer mesocomposite as neural electrode coating. APL Mater. **3**, 014701 (2015)
35. Graudejus, O., Morrison III, B., Goletiani, C., Yu, Z., Wagner, S.: Encapsulating elastically stretchable neural interfaces: yield, resolution, and recording/stimulation of neural activity. Adv Funct. Mater. **22**, 640–651 (2012)
36. Capadona, J.R., Shanmuganathan, K., Tyler, D.J., Rowan, S.J., Weder, C.: Stimuli-responsive polymer nanocomposites inspired by the sea cucumber dermis. Science **319**, 1370–1374 (2008)
37. Harris, J.P., Capadona, J.R., Miller, R.H., Healy, B.C., Shanmuganathan, K., Rowan, S.J., et al.: Mechanically adaptive intracortical implants improve the proximity of neuronal cell bodies. J. Neural Eng. **8**, 066011 (2011)
38. Nguyen, J.K., Park, D.J., Skousen, J.L., Hess-Dunning, A.E., Tyler, D.J., Rowan, S.J., et al.: Mechanically-compliant intracortical implants reduce the neuroinflammatory response. J. Neural Eng. **11**, 056014 (2014)
39. Hess, A.E., Capadona, J.R., Shanmuganathan, K., Hsu, L., Rowan, S.J., Weder, C., et al.: Development of a stimuli-responsive polymer nanocomposite toward biologically optimized, MEMS-based neural probes. J. Micromech. Microeng. **21**, 054009 (2011)
40. Simon, D., Ware, T., Liu, C., Musa, T., Vasudevan, S., Sloan, A., Keefer, E.W., Rennaker, R.L., Voit, W.: Thiol-ene/acrylate substrates for softening intracortical electrodes. J. Biomed. Mater. Res. Part B **102B**, 1–11 (2014)
41. Arreaga-Salas, D.E., Avendaño-Bolivar, A., Simon, D., Reit, R., Garcia-Sandoval, A., Rennaker, R.L., et al.: Integration of high-charge-injection-capacity electrodes onto polymer softening neural interfaces. ACS Appl. Mater. Interfaces **7**, 26614–26623 (2015)
42. Yazicioglu, F., Lopez, C.M., Mitra, S., Raducanu, B., Musa, S., Kloosterman, F.: Ultra-high-density in-vivo neural probes. IEEE EMBC, Chicago, IL, **2014**, 2032–3035 (2014)
43. Alivisatos, A.P., Chun, M., Church, G.M., Greenspan, R.J., Roukes, M.L., Yuste, R.: The brain activity map project and the challenge of functional connectomics. Neuron **74**, 970–974 (2012)
44. Lustenberger, C., Huber, R.: High density electroencephalography in sleep research: potential, problems, future perspective. Front. Neurol. **3**, 77 (2012)

45. Rubehn, B., Bosman, C., Oostenveld, R., Fries, P., Stieglitz, T.: A MEMS-based flexible multichannel ECoG-electrode array. J. Neural Eng. **6**, 036003 (2009)
46. Buzsaki, G., Stark, E., Berenyi, A., Khodagholy, D., Kipke, D.R., Yoon, E., et al.: Tools for probing local circuits: high-density silicon probes combined with optogenetics. Neuron **86**, 92–105 (2015)
47. Wark, H.A.C., Sharma, R., Mathews, K.S., Fernandez, E., Yoo, J., Christensen, B., et al.: A new high-density (25 electrodes/mm^2) penetrating microelectrode array for recording and stimulating sub-millimeter neuroanatomical structures. J. Neural Eng. **10**, 045003 (2013)
48. Müller, J., Ballini, M., Livi, P., Chen, Y., Radivojevic, M., Shadmani, A., et al.: High-resolution CMOS MEA platform to study neurons at subcellular, cellular, and network levels. Lab Chip **15**, 2713–2908 (2015)
49. Müller, J., Ballini, M., Livi, P., Chen, Y., Shadmani, A., Frey, U., et al.: Conferring flexibility and reconfigurability to a 26,400 microelectrode CMOS array for high-throughput neural recordings. Presented at the Transducers, Barcelona (2013)
50. Obien, M.E.J., Deligkaris, K., Bullmann, T., Bakkum, D.J., Frey, U.: Revealing neuronal function through microelectrode array recordings. Front. Neuroeng. **8**, 423 (2015)
51. Tsai, D., John, E., Chari, T., Yuste, R., Shepard, K.: High-channel-count, high-density microelectrode array for closed loop investigation of neuronal networks. Presented at the IEEE-EMBC, Milan, Italy, 7510–7513, (2015)
52. Lopez, C.M., Andrei, A., Mitra, S., Welkenhuysen, M., Eberle, W., Bartic, C., et al.: An implantable 455-active-electrode 52-channel CMOS neural probe. IEEE J. Solid-State Circuits **49**, 248–261 (2014)
53. Du, J., Riedel-Kruse, I., Nawroth, J., Roukes, M., Laurent, G., Masmanidis, S.: High-resolution three-dimensional extracellular recording of neuronal activity with microfabricated electrode arrays. J. Neurophysiol. **101**, 1671–1678 (2009)
54. Du, J., Blanche, T.J., Harrison, R.R., Lester, H.A., Masmanidis, S.C.: Multiplexed, high density electrophysiology with nanofabricated neural probes. PLoS One **6**, e26204 (2011)
55. Scholvin, J., Kinney, J.P., Bernstein, J.G., Moore-Kochlacs, C., Kopell, N., Fonstad, C.G., Boyden, E. S.: Close-packed silicon microelectrodes for scalable spatially oversampled neural recording. IEEE Trans. Biomed. Eng. (2015). doi:10.1109/TBME.2015.2406113
56. Duan, X., Fu, T.-M., Liu, J., Lieber, C.M.: Nanoelectronics-biology frontier: from nanoscopic probes for action potential recording in live cells to three-dimensional cyborg tissues. Nano Today **8**, 351–373 (2013)
57. Robinson, J.T., Jorgolli, M., Park, H.: Nanowire electrodes for high-density stimulation and measurement of neural circuits. Front. Neural Circuits **7**, 38 (2013)
58. Yizhar, O., Fenno, L.E., Davidson, T.J., Mogri, M., Deisseroth, K.: Optogenetics in neural systems. Neuron **71**, 9–34 (2011)
59. Dufour, S., Koninck, Y.D.: Optrodes for combined optogenetics and electrophysiology in live animals. Neurophotonics **2**, 031205 (2015)
60. Grosenick, L., Marshel, J.H., Deisseroth, K.: Closed-loop and activity-guided optogenetic control. Neuron **86**, 106–139 (2015)
61. Royer, S., Zemelman, B.V., Barbic, M., Losonczy, A., Buzsaki, G., Magee, J.C.: Multi-array silicon probes with integrated optical fibers: light-assisted perturbation and recording of local neural circuits in the behaving animal. Eur. J. Neurosci. **31**, 2279–2291 (2010)
62. Stark, E., Koos, T., Buzsáki, G.: Diode probes for spatiotemporal optical control of multiple neurons in freely moving animals. J. Neurophysiol. **108**, 349–363 (2012)
63. Kim, T.-i., McCall, J.G., Jung, Y.H., Huang, X., Siuda, E.R., Li, Y., et al.: Injectable, cellular-scale optoelectronics with applications for wireless optogenetics. Science **240**, 211 (2013)
64. McCall, J.G., Kim, T.-i., Shin, G., Huang, X., Jung, Y. H., Al-Hasani, R., et al.: Fabrication and application of flexible, multimodal light-emitting devices for wireless optogenetics. Nat. Protoc. **8**, 2413–2428 (2013)
65. Kim, R.-H., Kim, D.-H., Xiao, J., Kim, B.H., Park, S.-I., Panilaitis, B., et al.: Waterproof AlInGaP optoelectronics on stretchable substrates with applications in biomedicine and robotics. Nat. Mater. **9**, 929–937 (2010)
66. Pisanello, F., Sileo, L., Oldenburg, I., Pisanello, M., Martiradonna, L., Assad, J., et al.: Multipoint-emitting optical fibers for spatially addressable in vivo optogenetics. Neuron **82**, 1245–1254 (2014)
67. Pisanello, M., Patria, A.D., Sileo, L., Sabatini, B.L., Vittorio, M.D., Pisanello, F.: Modal demultiplexing properties of tapered and nanostructured optical fibers for in vivo optogenetic control of neural activity. Biomed. Opt. Express **6**, 4014–4026 (2015)
68. Zorzos, A., Scholvin, J., Boyden, E., Fonstad, C.: Three-dimensional multiwaveguide probe array for light delivery to distributed brain circuits. Opt. Lett. **37**, 4841–4843 (2012)
69. !Abaya, T., Blair, S., Tathireddy, P., Rieth, L., Solzbacher, F.: A 3D glass optrode array for optical neural stimulation. Biomed. Opt. Express **3**, 3087–3104 (2012)
70. Kwon, K.Y., Lee, H.-M., Ghovanloo, M., Weber, A., Li, W.: Design, fabrication, and packaging of an integrated, wirelessly-powered optrode array for optogenetics application. Front. Syst. Neurosci. **9**, 69 (2015)

MEMS on Flexible Substrates

Donald P. Butler and Zeynep Celik-Butler
Nanotechnology Research and Teaching Facility, The Department of Electrical Engineering, The University of Texas at Arlington, Arlington, TX, USA

Synonyms

MEMS = microelectromechanical systems

Definition

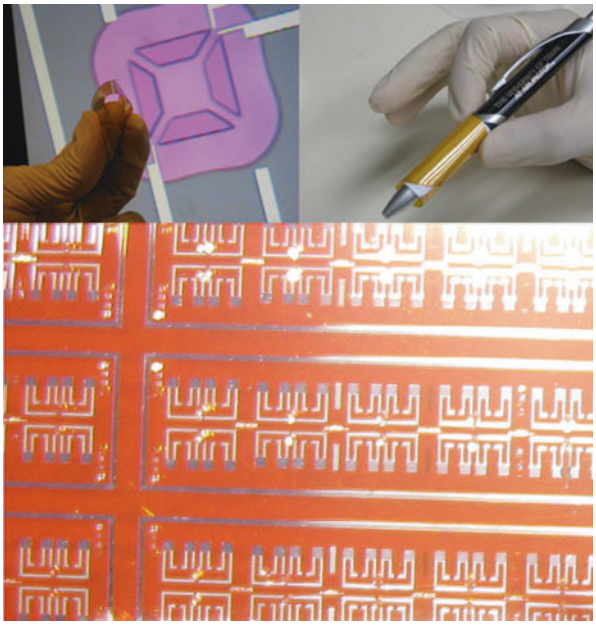
MEMS in flexible substrates can be defined as microelectromechanical devices and systems embedded in a flexible polymer or plastic film such that the device is conformal to nonplanar surfaces and objects.

Introduction

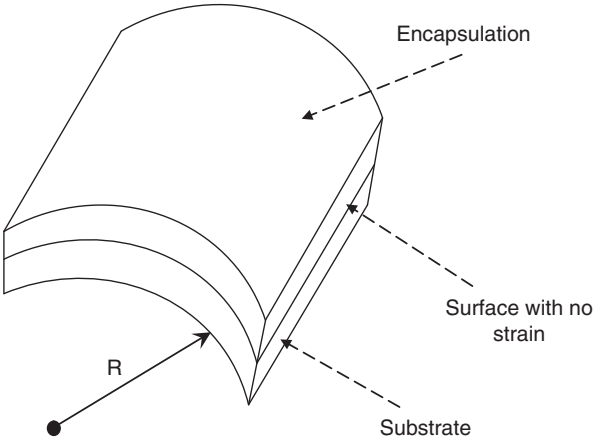
Flexible electronics [1] is a rapidly emerging field that has resulted in flexible displays using organic light-emitting diodes being inserted into the marketplace in commercial products such as cell phones. Flexible circuit boards are also seeing widespread use in commercial electronics. These flexible electronic components have the advantage of being able to conform to nonplanar objects and surfaces. They can be bent or folded to some degree. They also provide decreased size and weight when compared to their rigid conventional electronic counterparts. The use of organic semiconductors to form transistors and diodes has so far limited the performance of the components due to the low intrinsic mobility of the materials developed so far. Current organic transistors are sufficient for the low speed switching requirement of a video display but do not possess the performance of Si or SiGe transistors that are utilized for computing, communication, and signal-processing applications. This has resulted in hybrid systems constructed of both rigid Si components for high-performance applications and flexible organic transistors and diodes for the display. The recent development of graphene-based transistors [2] with high mobility and the continued research on other polymer materials for electronic applications may change this scenario in the near future by allowing for low-cost, high-speed electronics to be placed on or in flexible substrates. The inherent paradigm of flexible electronics would have potential beyond displays including sensing and communication electronics.

Microelectromechanical (MEMS) devices exist as both sensors [3] and actuators [4]. Microelectromechanical devices were first developed by bulk micromachining where the substrate was etched to form the device structure. This process has rapidly given way to surface micromachining for many devices, where the device is formed by the subsequent deposition of layers on a substrate. Some of these layers are sacrificial layers which are removed to release parts of the device giving them the freedom to move and form the electromechanical device. Surface micromachining is independent of the choice of substrate material, hence can readily be adapted to flexible substrates made of polymers and plastics so long as low-temperature processing techniques are employed that do not exceed the glass transition temperature of the substrate. This has been a challenge in developing MEMS sensors on flexible substrates since many traditional sensor designs use polycrystalline piezoresistive materials that require high temperature processing. MEMS devices in flexible substrates have applications where the ability to conform to a nonplanar surface or the savings in size and weight are advantageous. Toward this end, MEMS sensors in flexible substrates have been developed for medical and structural health-monitoring applications (Fig. 1), and also in defense. This work has been focused on MEMS sensors that measure infrared radiation, pressure/force, acceleration, strain, and other physical parameters.

Traditional electronics utilize stiff and thick substrates where the films on top must conform to the substrate. Any stress that develops is mainly

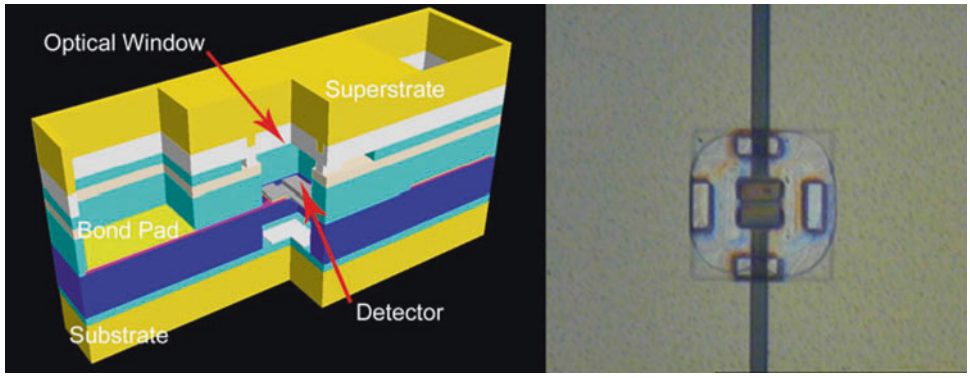


MEMS on Flexible Substrates, Fig. 1 Photographs of pressure/tactile sensors on flexible substrates illustrating the conformity of the devices. (*Top left*) A section of polyimide containing pressure sensors bent with an optical micrograph of the sensor in the background. (*Top right*) A section of polyimide containing pressure sensors applied to a pen. (*Bottom*) A closer view of the pressure sensors on the polyimide film



MEMS on Flexible Substrates, Fig. 2 The structure containing the substrate, electronic layer, and the encapsulation, bent to a radius R. A similar concept will be utilized in this investigation

concentrated in the films forming the device structure on the top surface of the substrate. In the case of flexible electronics, when the substrate is bent with the circle center closer to the *bottom* surface, the top surface goes into tension while the bottom surface goes into compression, as shown in Fig. 2. Under compression with the circle center closer to the *top* surface, a film or device may buckle and delaminate from the substrate. Under tension, a film or device may tear. At the neutral plane of the stack lies a low stress plane which experiences minimal strain. The devices are best suited to be located at this low stress plane where they are supported by a substrate layer below and a superstrate or encapsulation layer above. If the stiffness of the device layer in the middle is negligible, then the electronic layer can be made to lie on a plane of no strain as long as the following expression $Y_s d_s^2 = Y_e d_e^2$ is valid [5]. Here, the subscript e refers to the encapsulation while the subscript s refers to the substrate. Y is the Young's or tensile modulus and d is the thickness of the layer. By using an encapsulation, the electronic MEMS devices will not only lie at a low stress plane but they will be protected from the environment thus removing the need for



MEMS on Flexible Substrates, Fig. 3 (*Left*) An expanded view of a self-packaged microbolometer (not to scale: the *vertical* dimensions have been exaggerated and the bump bonds are placed next to the bolometer; the polyimide substrate and superstrate have been compressed for clarity). The IR transparent aluminum oxide window and bolometer with the aluminum mirror form an optically resonant structure to maximize absorption. (*Right*) An optical micrograph of the completed device

further packaging. Thus, the devices can be referred to as being self-packaged or device-level packaged [6–8] since the flexible substrate layer, encapsulation layer, and device layer are all fabricated using microfabrication techniques. This results in a MEMS device that is embedded in a flexible film or substrate. To produce a MEMS device layer in which parts of the device must be free to move, a local cavity is formed around the device by micromachining (Fig. 3). An alternative to device-level packaging the MEMS devices is to package the flexible encapsulation layer or superstrate to the substrate at the wafer level with a wafer bonder.

In this entry, the reader is introduced to the development of MEMS devices on flexible substrates and their packaging through the presentation of several examples including microbolometers, pressure sensors, and accelerometers. In addition, several other examples are discussed with their relationship to developing flexible electronics that include MEMS devices.

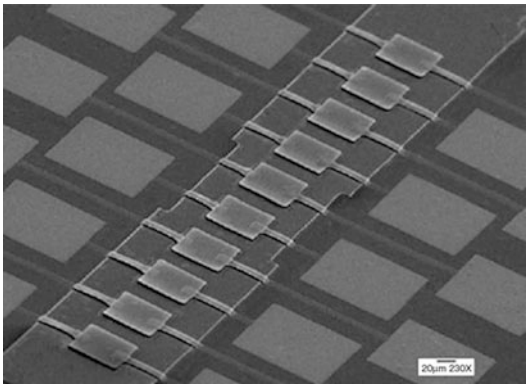
Microbolometers on Flexible Substrates

Microbolometers are thermal optical detectors whose electrical resistance changes as a result of the heating from the absorbed optical radiation. To minimize the heat lost by the microbolometer to the substrate, the resistive thermometer is suspended above the substrate through micromachining. The detectors are also placed in vacuum to eliminate the heat loss through the surrounding atmosphere. Suspending the thermometer above the substrate also allows for the formation of an optically resonant cavity to increase absorption of the detector. Microbolometers are typically used in night vision as infrared imagers operating in the 8–14 μm atmospheric transmission window.

The development of micromachined sensors on flexible substrates began with microbolometers [6, 9, 10]. There are advantages to producing curved imaging elements because the number of optical elements in the lens can be decreased, resulting in cameras that have reduced size and weight. Early work first investigated fabricating microbolometers on Kapton sheets bonded to Si carrier wafers so that conventional photolithography and semiconductor processing equipment could be used. This permitted the patterning of structures with micron precision. However thermal cycling that occurred during the microfabrication process caused distortion of the Kapton films that resulted in low device yield across the wafer [10]. The fabrication process was adapted to use polyimide (liquid Kapton, Dupont, now HD Microsystems, PI5878G) substrates that were formed by spin casting the

substrate onto a Si carrier wafer and curing the polyimide at the beginning of the microfabrication process [9]. In this case, the polyimide film did not undergo distortion during the microfabrication process, allowing for precise alignment of the device features across the wafer and a corresponding high yield. Polyimide and Kapton, for that matter, are attractive materials to use due to their flexibility and relatively high glass transition temperature, up to 400 °C. Polyimide is unattractive due to its moisture sensitivity, limiting its deployment in some applications. However, its material properties allowed for a good test bed to demonstrate the feasibility of the technology. New polymers are continually being developed for future application. Moreover, the evolution of high precision roll-to-roll processing equipment allows for high volume, large area fabrication of devices for commercial manufacturing [1]. The technique of utilizing a Si carrier wafer is more appropriate for the university research environment and small production lines.

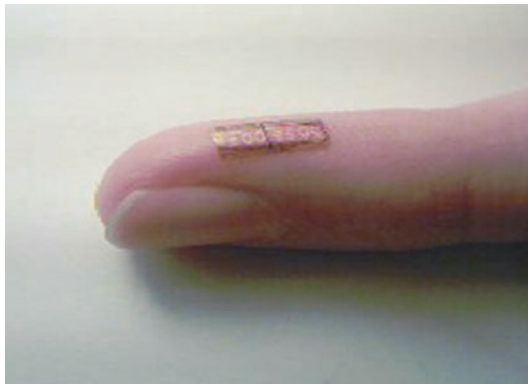
After spin casting and curing the polyimide substrate, a Si_3N_4 film was deposited to protect the substrate from future microfabrication processes and define the device plane [9]. The microbolometers were then fabricated on top of the Si_3N_4 film by using conventional photolithography and rf magnetron sputtering to deposit the metal electrodes and yttrium barium copper oxide (YBCO) thermometer. All sputter depositions were performed in a conventional turbo-pumped sputter system at room temperature. The highest temperature step was hard baking the photoresist at 105 °C. This ensured that the glass transition temperature of the polyimide substrate was not exceeded. The fabrication process involved the patterning and deposition of an Al mirror that would form an optically resonant cavity with the YBCO thermometer. A photodefinable polyimide sacrificial layer was then spin coated and cured on the wafer. The polyimide sacrificial layer was designed to be approximately 2-μm thick to maximize the infrared absorption at a wavelength of 10 μm and satisfy the approximately $\lambda/4$ relationship for maximum absorption in the suspended thermometer, where λ is the wavelength of the incident radiation. Subsequently, Ti electrode



MEMS on Flexible Substrates, Fig. 4 Micromachined YBCO microbolometer array fabricated on a spin-on polyimide substrate (liquid Kapton)

arms were patterned and deposited, followed by Au contacts on the bond pads and to the YBCO thermometer which was deposited next. The microbolometers were patterned into 1×10 arrays of 40×40 and 60×60 μm detectors. The polyimide sacrificial layer was then removed by ashing in an oxygen plasma to suspend the microbolometers above the substrate and form the optically resonant cavity with the Al mirror. The Si_3N_4 film deposited initially protected the polyimide substrate from being attacked by the oxygen plasma.

Some of the microbolometer die were gently peeled off the Si carrier wafer while others were left on the Si carrier wafer for characterization. An SEM micrograph of a section of a 1×10 array is shown in Fig. 4. Two microbolometer dies are applied to a fingertip in Fig. 5, demonstrating the conformal nature of the devices. The microbolometers were characterized by gluing the die, both attached to and removed from the Si carrier wafer, into a conventional integrated circuit package. The packaged microbolometers were placed into an evacuated cryostat with a ZnSe window. No significant difference in the responsivity and detectivity was observed between the microbolometers that were removed from the Si carrier wafer, those that remained on the Si carrier wafer, or to ones that were fabricated directly on a Si substrate. This indicated that MEMS microbolometers can be fabricated on both rigid or flexible substrates without



MEMS on Flexible Substrates, Fig. 5 Microbolometer arrays fabricated on flexible polyimide substrates applied to a finger

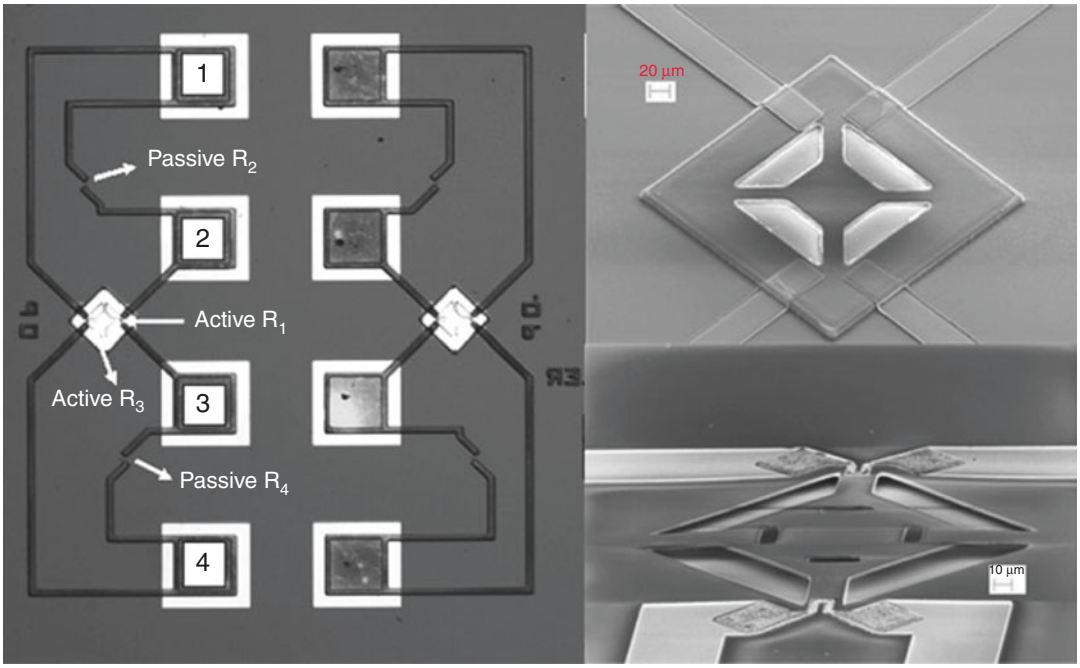
degradation of their performance. The infrared detectors showed broadband optical responsivity as high as 6.1×10^4 V/W and detectivity of 1.2×10^8 cm $Hz^{1/2}$/W at 1 μA of current bias [9].

This work also demonstrated that, in the general case, there is no need to fabricate sensors on flexible substrates if they are going to be placed in rigid, conventional integrated circuit packages. Such packaging defeats the ability of the devices to conform to nonplanar surfaces. Therefore, as a next step, device-level packaging of the microbolometers was developed [6]. The package for the microbolometer was now created during the microfabrication process, with a local vacuum cavity surrounding the microbolometer to minimize its thermal conductance with its surroundings. A schematic and optical micrograph (top view) of the device-level vacuum-packaged microbolometer is shown in Fig. 3. This device-level package flexible device would be compatible with direct integration or bonding to a flexible circuit board. In this case, the flexible die can be glued to the flexible circuit board with an appropriate adhesive with the electrical connections being made by bump bonding.

Here, the fabrication proceeded along the same lines as in the previous case up through the deposition of the YBCO thermometer. The precise details can be found in Ref. [6]. After the YBCO thermometer was deposited, a second polyimide sacrificial layer was spin coated and patterned on top of the thermometer. This was followed by an aluminum oxide, Al_2O_3, packaging layer. Al_2O_3 was chosen for its high mechanical strength and optical transparency, which extends into the infrared. The Al_2O_3 packaging layer had trench cuts or openings to facilitate the removal of the sacrificial layers above and below the thermometer by ashing in an oxygen plasma. After the sacrificial layers were removed, the trench cuts were sealed by sputtering more Al_2O_3 at 10 mTorr Ar pressure, thereby creating a vacuum cavity around the microbolometer. Once sealed, a polyimide encapsulation or superstrate could be applied to place the microbolometer at a low stress plane inside the polyimide film.

Pressure/Tactile Sensors on Flexible Substrates

Pressure/tactile sensors on flexible substrates have been developed for medical and structural health-monitoring applications. Two types of MEMS pressure sensors have been investigated: absolute pressure sensors and delta pressure sensors. Absolute pressure sensors measure the external pressure relative to a sealed reference cavity while delta pressure sensors respond to an external pressure of force relative to the ambient pressure. This entry will focus upon delta pressure sensors. MEMS delta pressure sensors operate by deflecting a suspended membrane with an external pressure or force. The deflection of the membrane can be sensed capacitively or by measuring the resulting strain induced in the membrane with piezoresistors. In the latter case, the piezoresistors can be configured into a Wheatstone bridge geometry with two active and two passive piezoresistors. This configuration provides zero output under a no load condition when the resistors are balanced and also provides temperature drift compensation by fabricating the piezoresistors from the same material with the same temperature coefficient of resistance. A confocal microscope picture of a piezoresistive delta pressure sensor is shown in Fig. 6. In the case of pressure sensors, since the sensor is designed to measure an external pressure, the



MEMS on Flexible Substrates, Fig. 6 (*Left*) A confocal microscope picture showing a piezoresistive delta pressure sensor in a Wheatstone bridge geometry. (*Top right*) A SEM micrograph of the piezoresistive delta pressure sensor in a mesa geometry. (*Bottom left*) A SEM micrograph of a delta pressure sensor showing the cavity underneath

sensor must have access to the environment through a hole in the superstrate or have a thin superstrate, less than 10-µm thick, so the external pressure is transferred to the sensor.

Two types of piezoresistors have been investigated, polycrystalline Si piezoresistors [11, 12] and nichrome piezoresistors [13]. In the case of polycrystalline Si piezoresistors, normally to achieve the polycrystalline lattice structure requires deposition or annealing temperatures that would be prohibitively high (>600 °C) for their use with flexible polymer substrates. By utilizing aluminum-induced crystallization, Si can be crystallized after its deposition by annealing at temperatures as low as 152 °C [14]. Piezoresistive Si with gauge factor ([ΔR/R]/average strain) in the range of 7–12 was obtained through this technique [15]. Alternatively, nichrome (nickel 80 %, chromium 20 %) thin films were employed as the piezoresistors [13]. The advantage of using metallic piezoresistors is their ability to be deposited at room temperature while one sacrifices the magnitude of the response due to their lower gauge factor, in the range of 1–3.

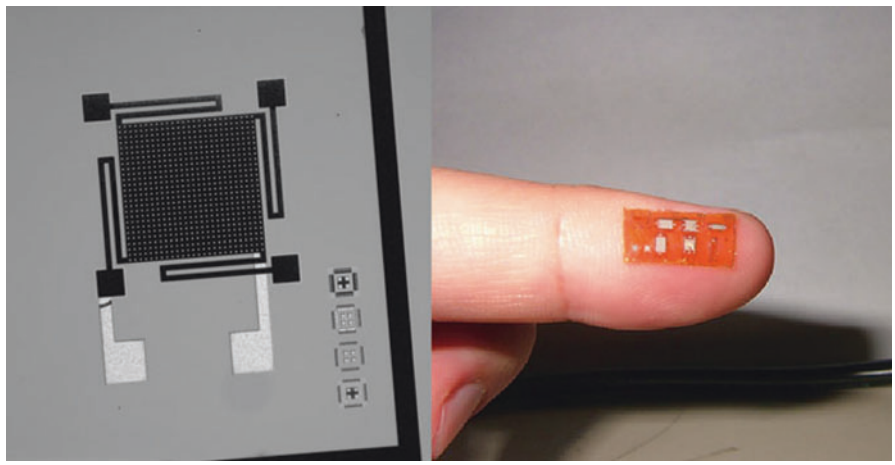
The fabrication of the delta pressure sensors on flexible substrates proceeded along lines similar to the fabrication of the microbolometers on flexible substrates described earlier. The basic fabrication began with spin coating the polyimide substrate onto a Si carrier wafer. After curing, a protective film of Si_3N_4 or Al_2O_3 was sputtered to protect the polyimide substrate during subsequent fabrication steps. Next, the polyimide sacrificial layer was spin coated and cured on the wafer. Then, the Si_3N_4 or Al_2O_3 membrane layer was deposited. The piezoresistors were then fabricated. Here, the process differed depending upon whether nichrome [13] or aluminum-induced crystallization polysilicon [12] was used as the piezoresistor. In the case of nichrome, the few tens of nanometer thick piezoresistor was deposited by rf magnetron sputtering and patterned by photolithographic liftoff to complete the piezoresistor fabrication.

For aluminum-induced crystallization of polysilicon, first a 0.5-μm-thick aluminum layer was deposited, followed by a 0.5-μm-thick amorphous Si layer. The two layers were patterned together by liftoff to define he piezoresistor geometry. Then, the wafer was annealed. During this process, the aluminum and silicon layers change positions through interdiffusion and a polycrystalline Si layer results at the bottom of the stack. The Al top layer is then removed by wet etching leaving behind an Al-doped poly-Si piezoresistor. The Al metallization layer was then deposited and patterned by liftoff to wire the active and passive piezoresistors into a Wheatstone bridge geometry. To prevent the piezoresistors from oxidizing during the removal of the polyimide sacrificial layer, a protective layer was deposited over the piezoresistor. In the case of the nichrome piezoresistors, a thin layer of Al_2O_3 was used. The polyimide sacrificial layer was ashed in an oxygen plasma to suspend the membrane above the substrate. In some devices, a thin polyimide layer was then bonded on top of the wafer to protect the pressure sensors. In other devices, a photo-patternable superstrate was spin coated on top of the pressure sensors and developed to open a small access hole to the sensor. At this point, the fabrication was complete and the sensor could be removed from the substrate.

The sensors can be designed to have various sensitivities by varying the membrane thickness and size of the suspended region. In this case, typical dimensions were 70–150 μm across. After completion of the sensor fabrication, the sensors were tested by deflecting the membrane with a probe tip in a probe station or applying a known load with a load cell [13] or atomic force microscope [16].

Accelerometers on Flexible Substrates

MEMS accelerometers on flexible substrates have applications in structural health monitoring of aircraft and aiding in the control of micro air vehicles [17]. The basic operating principle of a MEMS accelerometer is an under-damped to critically damped spring-mass system. The inertia of the mass causes a deflection when the mass is accelerated. By measuring the deflection, the acceleration can be determined [3]. The deflection can be measured by piezoresistive, piezoelectric, or capacitive transducers. Capacitive accelerometers, fabricated using a modified UV-LIGA process [18] with an electroplated Ni proof mass (Fig. 7), have been investigated. Both z-axis (perpendicular to substrate) and lateral axis (parallel to substrate) accelerometers have been



MEMS on Flexible Substrates, Fig. 7 (*Left*) An optical micrograph of a z-axis accelerometer fabricated on a polyimide substrate. (*Right*) A flexible polyimide die containing z-axis and lateral axis accelerometers conformally resting on a finger

designed to measure accelerations up to 20 *g* where *g* is the acceleration due to gravity. The details of the fabrication process are given elsewhere [17]. The fabrication will be described in brief here.

The basic fabrication process started with spin coating the polyimide substrate onto a Si carrier wafer followed by the sputter deposition of a Si_3N_4 protection layer. The accelerometers were constructed using a polyimide sacrificial layer that was patterned by photolithography and cured. An Au seed layer was utilized for electroplating the Ni proof mass. The mold for the proof mass was defined with a 6-μm-thick negative photoresist. The low stress Ni was electroplated in a nickel sulfamate solution. After electroplating, the mold resist was stripped and the Au seed layer was wet etched. Then, the polyimide sacrificial layer was ashed with an oxygen plasma to suspend the proof mass above the substrate with Ni springs defined by the photoresist mold. The dies were then packaged for testing. Some of the dies were taken off the substrate as shown in Fig. 7. Figure 7 also depicts an optical micrograph of a z-axis accelerometer.

The accelerometers were tested with a shaker that provided accelerations up to 8 g. The acceleration was calibrated through a commercially available accelerometer which was mounted beside the accelerometer under test. The z-axis accelerometers displayed a sensitivity exceeding 20 fF/g while the sensitivity of the lateral axis accelerometers was more than 10 fF/g [17]. The sensitivity was compared to identical accelerometers fabricated directly on Si substrates and found to be the same within the measurement and fabrication tolerances. The noise floors due to Brownian motion of the suspended sensor was calculated to be 2.25 μV/√Hz compared to the measured noise floor of the readout circuit of 0.13 mV/√Hz indicating that the measurement system dominated the measured noise [17].

Other MEMS on Flexible Substrates

At the University of Illinois at Urbana-Champaign (UIUC) work on incorporating electronic devices on flexible substrates [19] has also turned to micromachining in some cases [20, 21]. Using micromachining, ribbons of electronic devices [20] and interconnects [21] have been suspended above the substrate to enable both flexibility and stretchability of the electronic film. This work is separated from the work described earlier in that electronic devices such as diodes and transistors are incorporated on the flexible substrate. Micromachining is used to suspend the electronic ribbons or interconnects above the substrate with periodically spaced anchor points so that when the substrate is bent or stretched, there is room for the electronic ribbons and interconnects to deform without encountering a large strain that could damage the devices. Therefore, in this case, the devices are more traditional electronic devices rather than traditional MEMS devices described earlier. The research at UIUC may form a path for the integration of MEMS devices with electronic components such as transistors. One factor not addressed by this work is the packaging of the suspended components. Along similar lines, other investigators have also reported micromachined interconnects that are suspended above the substrate [22, 23].

Using MEMS cantilever rf switches, reconfigurable frequency-selective surfaces [24] and antennas [25, 26] have been demonstrated on flexible substrates. A surface micromachined hot-wire anemometer has also been shown [27]. In addition, a variety of MEMS biomedical devices have been developed on flexible substrates including implantable moveable electrodes for the brain [28] and a polymerase chain reaction (PCR) system for DNA amplification [29]. These devices validate the attractiveness of MEMS devices on flexible substrates for a variety of applications. Yet a common set of problems always exists and is sometimes neglected such as packaging of the MEMS devices so that they have the ability to function without losing their flexibility and can be freely handled without damage. This property is more easily achieved for small-area MEMS devices but becomes more difficult as the size increases. In addition, the design process to achieve flexibility is easier for small-area devices compared to large area devices since the

deflection and stress encountered by a small-area device is less than those experienced by a large area device for the same radius of curvature.

Future Directions and Conclusions

This work has demonstrated MEMS sensors can be fabricated on a variety of substrates from rigid Si to flexible polyimide without losing performance. The work on flexible substrates has thus far been focused on constructing MEMS using conventional electronic materials for the insulators and metals for the wiring. The key restriction has been the necessity to use low-temperature processing to maintain the polyimide below the glass transition temperature during the fabrication process. In future, the traditional electronic materials of Al_2O_3, Si_3N_4, Al, and nichrome will give way to polymer and plastic materials that can be used for the structural MEMS materials. Work has already been started on using polymer-carbon nanocomposites for the piezoresistive transducers employed in many MEMS sensors. The compatibility with organic semiconductors and graphene transistors will pave the way for the integration of transistor readout circuitry in the near future. MEMS sensors on flexible substrates provide the ability to conform to nonplanar surfaces that are attractive in medical, robotic, and aviation applications. The resulting devices have the potential to be fabricated at reduced cost, size, and weight. Flexible electronics appears to be growing in its attractiveness for a variety of applications.

Acknowledgment This material is based in part on work supported by the Air Force Office of Scientific Research, the Texas Ignition Fund, and Lockheed Martin Aeronautics Company.

Cross-References

- ▶ AFM, Tapping Mode
- ▶ Flexible Electronics
- ▶ Graphene
- ▶ Insect Flight and Micro Air Vehicles (MAVs)
- ▶ MEMS Packaging
- ▶ Piezoelectric Effect at Nanoscale
- ▶ Piezoresistivity

References

1. Wong, W.S., Salleo, A. (eds.): Flexible Electronics: Materials and Applications. Springer, New York (2010)
2. Schwierz, F.: Graphene transistors. Nat. Nanotechnol. **5**, 487–496 (2010)
3. Elwenspoek, M., Wiegerink, R.: Mechanical Microsensors. Springer, New York (2001)
4. Lobontiu, N.: Dynamics of Microelectromechanical Systems (Microsystems). Springer, New York (2007)
5. Suo, Z., Ma, E.Y., Gleskova, H., Wagner, S.: Mechanics of rollable and foldable film-on-foil electronics. Appl. Phys. Lett. **74**, 1177–1179 (1999)
6. Mahmood, A., Butler, D.P., Celik-Butler, Z.: Device-level vacuum-packaging scheme for microbolometers on rigid and flexible substrates. IEEE Sensors J. **7**, 1012–1019 (2007)
7. Rahman, M.S., Chitteboyina, M., Butler, D.P., Celik-Butler, Z., Pacheco, S., McBean, R.: Device-level vacuum packaged MEMS resonator. IEEE/ASME J. Microelectromech. Syst. **19**, 911–917 (2010)
8. Mahmood, A., Butler, D.P., Celik-Butler, Z.: Device-level vacuum packaged micromachined infrared detectors on flexible substrates. Presented at the IEEE Sensors Conference, 31 Oct–3 Nov 2005, Irvine (Proceedings)
9. Dayeh, S.A., Butler, D.P., Celik-Butler, Z.: Micromachined infrared bolometers on flexible polyimide substrates. Sens. Actuators A **118**, 49–56 (2005)
10. Yaradanakul, A., Celik-Butler, Z., Butler, D.P.: Infrared sensing microbolometers on flexible substrates. IEEE Trans. Electron Devices **49**, 930–933 (2002)
11. Patil, S.K., Celik-Butler, Z., Butler, D.P.: Nanocrystalline piezoresistive polysilicon film by aluminum induced crystallization for pressure sensing applications. IEEE Trans. Nanotechnol. **9**, 640–646 (2010)
12. Patil, S.K., Celik-Butler, Z., Butler, D.P.: Piezoresistive polysilicon film obtained by low-temperature aluminum induced crystallization. Thin Solid Films **519**, 479–486 (2010)
13. Nadvi, G.S., Butler, D.P., Celik-Butler, Z., Gönenli, İ. E.: Micromachined pressure/force sensors using nickel chromium piezoresistors (submitted for publication)
14. Radnoczi, G., Robertsson, A., Hentzell, H.T.G., Gong, S.F., Hasan, M.A.: Al induced crystallization of a-Si. J. Appl. Phys. **69**, 6394–6399 (1991)
15. Patil, S.K.: Nanocrystalline piezoresistive polysilicon film obtained by aluminum induced crystallization for pressure sensing applications. PhD dissertation, University of Texas at Arlington (2010)

16. Patil, S.K., Celik-Butler, Z., Butler, D.P.: Characterization of MEMS piezoresistive pressure sensors using AFM. Ultramicroscopy **110**, 1154–1160 (2010)
17. Gönenli, İ.E., Çelik-Butler, Z., Butler, D.P.: MEMS accelerometers on polyimide substrates. IEEE Sensors J. **11**, 2318–2326 (2011)
18. Saile, V., Wallrabe, U., Tabata, O., Korvink, J.G.: LIGA and Its Applications. Wiley-VCH, Weinheim (2009)
19. Ahn, J.-H., Kim, H.-S., Lee, K.J., Zhu, Z., Menard, E., Nuzzo, R.G., Rogers, J.A.: High-speed mechanically flexible single-crystal silicon thin-film transistors on plastic substrates. IEEE Electron Device Lett. **27**, 460–462 (2006)
20. Sun, Y., Rogers, J.A.: Structural forms of single crystal semiconductor nanoribbons for high performance stretchable electronics. J. Mater. Chem. **17**, 832–840 (2007)
21. Kim, D.-H., Choi, W.M., Ahn, J.-H., Kim, H.-S., Song, J., Huang, Y., Liu, Z., Lu, C., Koh, C.G., Rogers, J.A.: Complementary metal oxide silicon integrated circuits incorporating monolithically integrated stretchable wavy interconnects. Appl. Phys. Lett. **93**, 044102 (2008)
22. Lin, K.L., Jain, K.: Design and fabrication of stretchable multilayer self-aligned interconnects for flexible electronics and large-area sensor arrays using excimer laser photoablation. IEEE Electron Device Lett. **30**, 14–17 (2009)
23. Lin, K.L., Chae, J., Jain, K.: Design and fabrication of large-area, redundant, stretchable interconnect meshes using excimer laser photoablation and in situ masking. IEEE Trans. Adv. Packag. **33**, 592–601 (2010)
24. Coutts, G.M., Mansour, R.R., Chaudhuri, S.K.: Microelectromechanical systems tunable frequency-selective surfaces and electromagnetic-bandgap structures on rigid-flex substrates. IEEE Trans. Microw. Theory Tech. **56**, 1737–1746 (2008)
25. Kingsley, N., Ponchak, G.E., Papapolymerou, J.: Reconfigurable RF MEMS phased array antenna integrated within a liquid crystal polymer (LCP) system-on-package. IEEE Trans. Antennas Propag. **56**, 108–118 (2008)
26. Kingsley, N., Anagnostou, D.E., Tentzeris, M., Papapolymerou, J.: RF MEMS sequentially reconfigurable sierpinski antenna on a flexible organic substrate with novel DC-biasing technique. J. Microelectromech. Syst. **16**, 1185–1192 (2007)
27. Buder, U., Berns, A., Petz, R., Nitsche, W., Obermeier, E.: AeroMEMS wall hot-wire anemometer on polyimide substrate featuring top side or bottom side bondpads. IEEE Sensors J. **7**, 1095–1101 (2007)
28. Jackson, N., Muthuswamy, J.: Flexible chip-scale package and interconnect for implantable MEMS Movable microelectrodes for the brain. J. Microelectromech. Syst. **18**, 396–404 (2009)
29. Lee, D.-S., Park, S.H., Chung, K.H., Pyo, H.-B.: A disposable plastic-silicon micro PCR chip using flexible printed circuit board protocols and its application to genomic DNA amplification. IEEE Sensors J. **8**, 558–564 (2008)

MEMS Packaging

Bongsang Kim[1] and Woo-Tae Park[2]
[1]Advanced MEMS, Sandia National Laboratories, Albuquerque, NM, USA
[2]Seoul National University of Science and Technology, Seoul, South Korea

Synonyms

MEMS encapsulation

Definition

A method of protecting microelectromechanical system (MEMS) devices from the environment by means of physical enclosures.

Introduction

Packaging is a challenging and critical step in the development of MEMS products. MEMS structures are generally very fragile, have moving structures, and must interface with the environment. The requirements for packaging are application specific and differ from each other, thus standard MEMS packaging protocols are difficult to establish, leading to an increase in product cost. For many MEMS products, the cost for packaging often dominates over the device fabrication cost. In this entry, general packaging requirements for typical MEMS devices will be discussed, followed by requirements for specific applications. Then, various state-of-the-art MEMS packaging approaches will be introduced with some examples.

General Requirements for MEMS Packaging

To ensure proper operation in environments that may be encountered, packaging is required for any semiconductor electronic device. Generally,

packaging is the final step of semiconductor device fabrication before final product testing. The packaging activity encompasses a wide spectrum of materials and technologies from the back-end-of-line (BEOL) of the semiconductor manufacturing process until the final product itself. The most basic function is to distribute the electrical connections from the semiconductor chip to a circuit board. This function is usually realized by wirebond or flip chip technology. Besides establishing electrical connections, other functions include: (1) mechanical protection from the environment of use or subsequent packaging process, (2) chemical protection from the environment, and (3) thermal management of the heat generated from the chip. Depending on the device and product requirement, the importance of packaging functionality listed above varies. For example, for modern central processing unit (CPU) devices, thermal management is the most critical, while mechanical protection and chemical protection are relatively less important.

In the case of MEMS devices, thermal management is usually less problematic, since most applications draw relatively low current, and thus generate little heat. However, MEMS devices contain mechanical structures. Therefore, they tend to be more sensitive to mechanical perturbation from the environment. Also, as the size scales down, the surface-to-volume ratio increases, which makes MEMS devices more susceptible to surface forces that may develop.

Mechanical Protection

Usually, MEMS devices contain suspended microscale mechanical structures with micrometer-spaced gaps. Mechanical protection is the most obvious required functionality of MEMS packaging. Airborne particles must be kept away from structures to ensure proper function. Impact can potentially fracture fragile structures, or cause compliant structures to come into contact, leading to adhesion failure. Mechanical stress on the device structure can also jeopardize its reliability. For example, stress can lead to a change in the electrical signal in the devices using piezoresistive or piezoelectric transduction. Capacitively transduced devices can also generate unreliable output signal, if the relative positions between electrodes are shifted by stress. Therefore, one important requirement of MEMS packaging is preventing MEMS devices from exposure to direct mechanical impact or shock. For example, shock requirement becomes more important in some applications such as automotive airbag sensors (sensing range ~50 g), and consumer devices that are prone to be dropped by the user.

Chemical Protection

Chemical protection must also be provided by MEMS packaging technology. Water condensation on the micromechanical structure can create permanent failure to the MEMS device. Due to their large surface-to-volume ratio, chemical reactions on the surface can completely change the device performance. For example, surface oxidation can completely change the mechanical fatigue behavior of silicon-based MEMS devices. For such reasons, from the beginning of the MEMS device development, the Hermeticity of its packaging has been a significant issue. Early studies were focused on hermeticity against humidity or moisture [1]. Later, hermeticity against atmospheric gas species has become important for many vacuum applications such as IR sensors and RF resonators, or inertial sensors where certain level of damping has to be maintained during operation [2]. Recent research works have investigated the diffusion characteristics of different gas species and diffusion pathways in MEMS packagings [3].

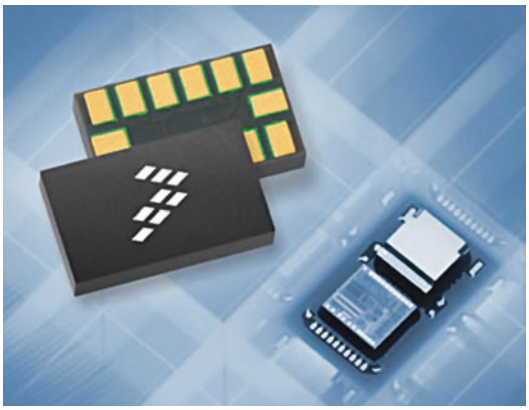
Cost-Effectiveness

Although MEMS packages have many stringent requirements (electrical connections, mechanical and chemical protection), cost minimization is required to gain commercial success. With respect to the mass market, MEMS sensors first penetrated the automotive industry. Even though the cost of the packaging could be up to 90 % of the device cost, MEMS sensors were still much more economical choice compared to the few existing solutions. In order to provide isolation from the harsh environment of automobiles (e.g., high temperature and humidity), ceramic packages that closely matched the coefficient of

thermal expansion (CTE) of silicon were first used. They were later replaced by molded packages to reduce cost. Since molded packages have a different CTE from silicon, a soft and thick die-attach such as DOW RTV and gel dome coating have been used for stress isolation (i.e., to minimize the transfer of stress from the package to the device).

In the consumer electronics market, the requirement for cost reduction became even higher. To bring the cost of the overall sensor to less than $1, cost reduction is needed in all areas of product manufacturing, and packaging is no exception. Today the most prevalent standard packages used are quad flat no leads (QFN), dual flat no leads (DFN), or land grid array (LGA) packages (testing is done in arrays before package singulation). Another important requirement for consumer electronics applications is the size of the overall package. The packages need to have the smallest possible footprint, and thin profile. The accelerometers (STMicroelectronics LIS331series) used in Apple's Iphone 4 have dimensions of 3 mm × 3 mm × 1 mm. Because of this small size, the thick and soft die-attach or the gel dome coating cannot be used anymore. Although themomechanical stress or long-term reliability is not as stringent as in automotive products, stress isolation remains an issue with the thin die-attach and gel coating. Alternatively, more effort was directed toward the MEMS sensor design to compensate the stress. One example is a single anchor design, so that no thermomechancial stress is applied due to anchor movement.

Advances in consumer electronics applications will be applied back to automotive applications, leading to further cost reduction. This is important as passenger safety requirements are becoming even more strict for automobiles. For example, in the USA, the National Highway Traffic Safety Administration (NHTSA) has already implemented front-seat airbag regulation, tire pressure monitoring systems, and will implement vehicle stability program regulation in 2012 for all vehicles. All of these safety regulations require MEMS sensor-based systems, and must be affordable for cars in all classes.



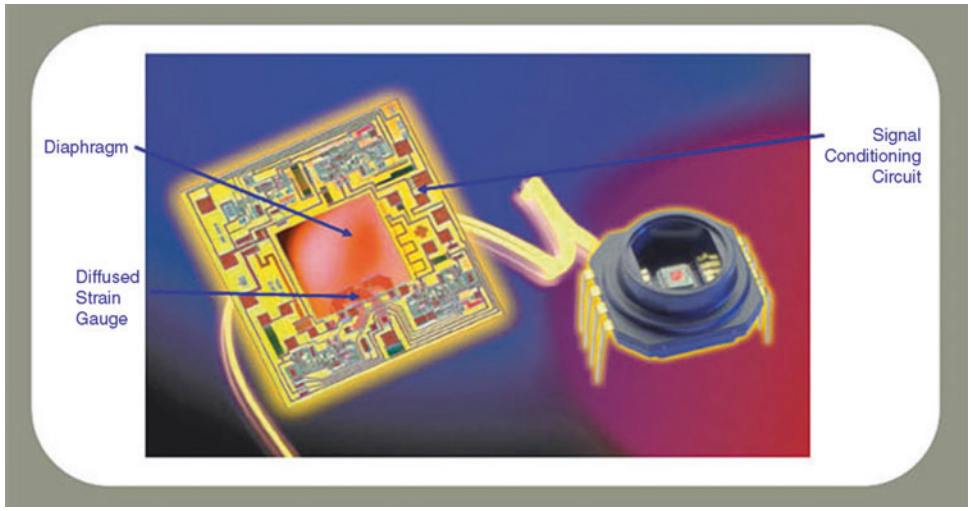
MEMS Packaging, Fig. 1 An accelerometer that is wafer level packaged with a cap wafer bonded by glass frit bonding (*left side*). Afterward, the accelerometer and ASIC are packaged in a land grid array (*LGA*) package (*right side*) [4]

Various MEMS Applications and Their Packaging Requirements

Inertial Sensors

Accelerometers, such as shown in Fig. 1, are one of the most successful MEMS devices. This success is due to MEMS technology advances and to the NHTSA airbag safety regulation (section 208) in the 1990s. Further success will be aided by recently announced side airbag regulations beginning in 2013. Accelerometers have contributed in saving millions of lives. For consumer electronics, accelerometers were used to detect "portrait-landscape" orientation for wide-screen phones. They are currently also widely used as motion control sensors in consumer electronics.

Accelerometer sensitivity can often be improved by increasing the size of released proof mass. Due to the need for a relatively large cavity, wafer bonding is preferred over other wafer-level encapsulation methods. In this approach, after device release, encapsulation is accomplished by sealing the device wafer with a separate "cap" wafer. Pre-etched wafers with cavities are bonded to the device wafer immediately after the release step (see the descriptions in the next section for more details). The gas environment during the bonding step must be controlled to establish the sensor dynamics. Nitrogen is used



MEMS Packaging, Fig. 2 Freescale pressure sensor packaged with die-level packaging [4]

as the main filling gas, because it is inert and the molecules are relatively large (compared to hydrogen or helium), so diffusion is slow. Pressure is carefully set to critically damp the mass.

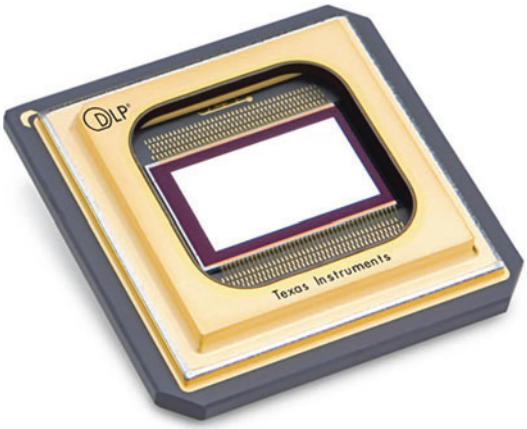
Gyroscopes have a more complex device structure than accelerometers. MEMS gyroscopes are usually designed as vibration-based structures, so high vacuum is required (<1 Torr). Because of this requirement, there are many considerations to be made. First, the bonding wafers and the bonding chamber need to be sufficiently outgassed before bonding. Outgassing from the surfaces of the wafer and chamber has high impact on the resulting vacuum level of the device. Second, the encapsulation volume must be tailored. Subsequent outgassing cannot be avoided completely and extreme outgassing steps are expensive and not manufacturable. Therefore, the encapsulation volume needs to be large enough to overcome the inevitable outgassing from the device surface. Finally, the encapsulation material needs to be hermetic. Polymer-based materials are not suitable since gas can easily penetrate the barrier over time. A getter material such as titanium or zirconium can also be used to ensure vacuum and will also help to maintain the long-term reliability of the device.

Pressure Sensors

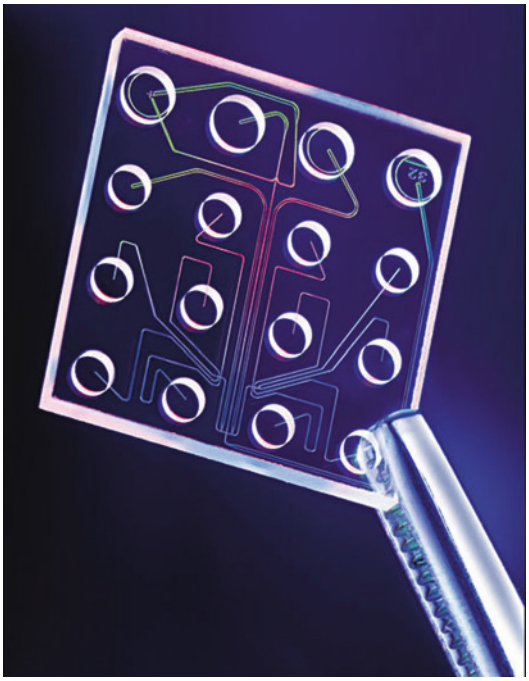
Pressure sensors are different from inertial sensors since the movable structure, the membrane, needs to be exposed to the environment during use. Another difference is that the membrane can be cleaned during the packaging process if any debris settles on it. Therefore, encapsulation need not be done immediately after the membrane is formed. Manufacturers use a soft gel type polymer to encapsulate the membrane for packaging. The gel allows the transmission of pressure to the membrane while protecting it from debris. The gel is used for packages having the pressure port close to the membrane. An example of a pressure sensor package is shown in Fig. 2. For differential pressure sensing, another hole in the backside is needed. Although most pressure sensor packages are not molded (unlike accelerometers), model LPS001D from ST microelectronics use a "holed" cap wafer on top of the membrane to realize a molded package. This new method should reduce the packaging cost and provide a standard packaging method as the accelerometers.

Optical MEMS

Packaging for many optical MEMS device must be optically accessible (transparent), which significantly limits the choice of materials. Typically,



MEMS Packaging, Fig. 3 MEMS mirror-based digital light processing (DLP™) product packaged with die-level packaging (Picture courtesy Texas Instruments)



MEMS Packaging, Fig. 4 Agilent Labchip™ packaged with a polymer bonding technology (Picture courtesy Agilent Technologies)

coating materials for AR (antireflection) surface or HR (high reflection surface) are not standard semiconductor materials, which makes the choice for packaging options even narrower. Epoxy is often used to hold optical components together; however, its use in the light pathways must be avoided because of long-term reliability and drift issues. Figure 3 shows an example of a packaged optical MEMS device, MEMS mirror-based digital light processing products from Texas Instruments.

Hermetic sealing is highly desirable since moisture can cause electrode erosion or degrade AR or HR surface coatings. Optical alignment of the packaging window with MEMS components can be critical. In addition, thermal management is important since optical MEMS systems can incorporate high laser power.

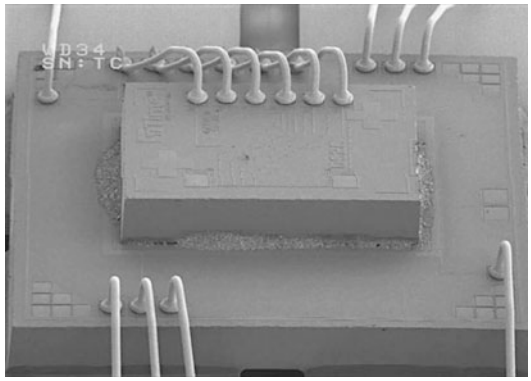
Fluidics/Bio Applications

Packaging for microfluidics requires good level of sealing to prevent leakage of the liquid in the device. In many applications, packaging materials have to be carefully selected considering their chemical compatibilities with the interfacing fluid. Also for the applications in this category, micro-to-macro interconnectors are often needed such as reservoirs and micropumps, which pose another challenge for packaging. Figure 4 shows an example of a packaged bio-MEMS product, Agilent Labchip™.

Especially for bio-related applications such as lab-on-a-chip or implantable devices, material biocompatibility is one of the most important criteria. Polymers have been widely used for the materials in such bio-related applications but usually polymers require joining processes that are different from those developed for silicon or glass-based devices. Polymer bonding particularly needs low-temperature processes that do not necessitate the use of high voltage, high pressure, or vacuum. Examples of sealing methods for polymer fluidic devices are adhesive joining, thermal welding, solvent welding, and plasma welding.

RF MEMS

RF MEMS is one of the most rapidly growing MEMS fields in market (Fig. 5). After Stanford University demonstrated long-term stability of silicon MEMS resonators [5], several silicon-based oscillator start-ups have been successfully



MEMS Packaging, Fig. 5 MEMS resonators packaged with an epitaxial silicon thin-film deposition as the cap layer. Packaged MEMS dies are bonded with CMOS circuit chips connected by wirebonds (Picture courtesy SiTime)

launched. Aluminum nitride-based MEMS filters appear to be next on the horizon. RF MEMS switches have also recently demonstrated good reliability [6], and are becoming commercialized.

The reliability of RF MEMS devices heavily rely on their packaging quality. For resonating devices for filters or oscillators, surface contamination plays a key role for their long-term frequency stabilities; therefore, hermetic packaging that can protect contaminant intrusion is essential for such applications. In addition, many resonator-based RF MEMS devices require high level of vacuum to minimize damping for high-Q, therefore hermetic vacuum packaging is needed for many RF MEMS applications. In the case of mechanical switches, the humidity has been observed to be the main cause of switch performance degradation [7].

Discera utilizes glass frit packaging, which provides required vacuum hermeticity. Discera also uses a transparent cap to allow the use of laser trimming for the absolute frequency tuning of the resonator. SiTime utilizes epitaxially deposited silicon film encapsulation. The technique provides advantages not only on the cavity cleanliness due to extremely high sealing temperature, ~1,000 °C, but also on its capability to be integrated with transistor circuits. These factors may enable huge system-level benefits such as reduction in size, cost, and noise.

Examples of MEMS Packaging Approaches

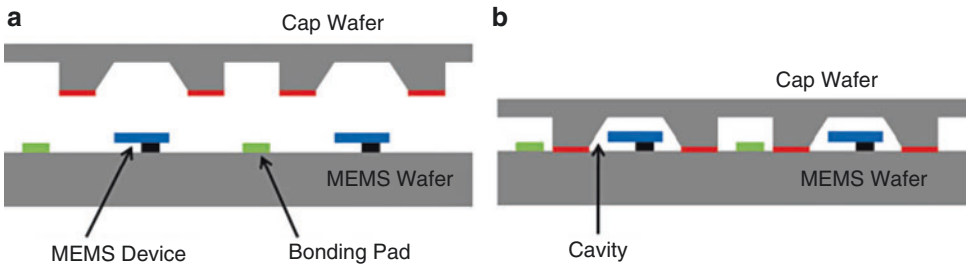
Wafer-Level Packaging

In most cases, fabricated MEMS devices are singulated from the host wafer in order to be packaged. This singulation process generally involves mechanical dicing and wet processes, which can cause yield loss of MEMS devices. Wafer-level packaging (WLP) refers to the technology of packaging MEMS devices at the wafer level, avoiding difficulties in singulation. WLP can dramatically reduce the handling and increase the yield of MEMS devices, thereby significantly decreasing the manufacturing cost. Traditionally, two approaches have been widely researched and developed for wafer-level packaging methods for MEMS devices. One is wafer-bonding packaging while the other is thin-film packaging.

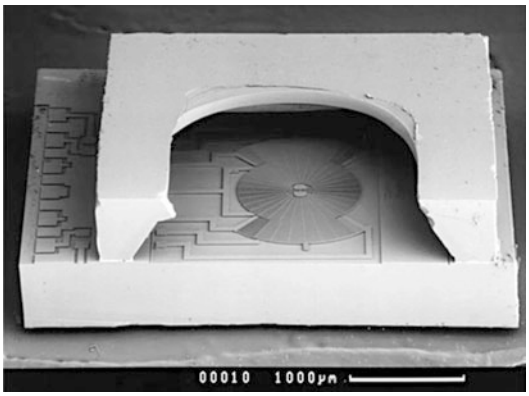
Wafer-Bonding Packaging

Wafer-bonding packaging involves fusing two separate wafers. A wafer containing MEMS devices is mechanically bonded with a separately prepared wafer having a cavity for MEMS devices, as shown in Fig. 6. A variety of bonding techniques have been introduced including silicon-silicon fusion bonding, silicon-glass bonding anodic bonding, glass frit bonding, and eutectic bonding. Silicon-silicon fusion bonding and silicon-glass anodic bonding both require high temperature, limiting the range of materials that can be present during bonding. Also, the prepared bonding surface needs to be ultra flat (<10 nm). This type of bonding is used to enclose pressure sensors and other MEMS devices before metallization is done. Recent developments, using plasma activation of the fusion bonding surfaces, are enabling low temperature or even room temperature bonding of ultra flat layers.

Glass frit bonding utilizes a glass frit that is screen printed to the wafer, fired to remove the solvents, and bonded at ~500 °C [8]. Due to the multiplicity of MEMS devices that can be packaged inside the cavity using this technique as well as the relatively cheap processing cost, this type of packaging (as shown in the example of (Fig. 7)) is the most widely used MEMS packaging method



MEMS Packaging, Fig. 6 (**a**) A MEMS wafer and a cap wafer with cavities are prepared separately. (**b**) The two wafers are bonded mechanically, realizing MEMS devices in the packaged cavities



MEMS Packaging, Fig. 7 An SEM image of a MEMS accelerometer packaged by the wafer-bonding approach (Picture courtesy Bosch RTC)

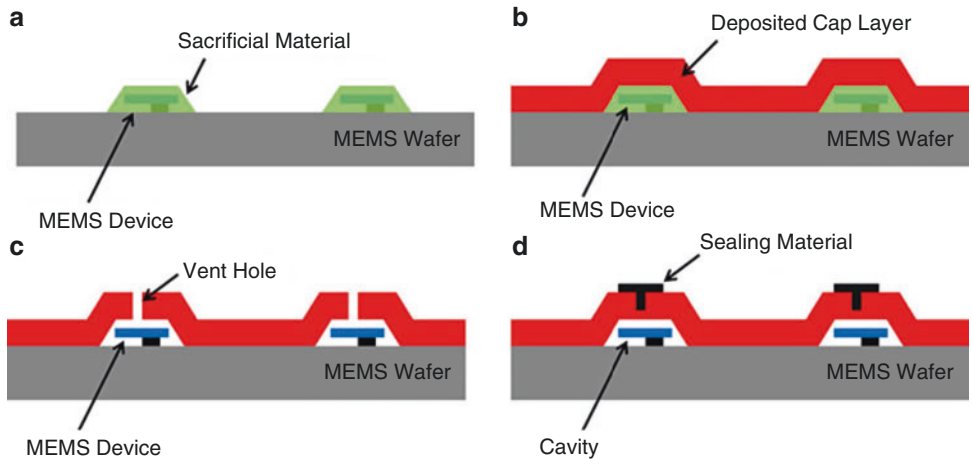
in MEMS industry. One drawback of this approach is the efficiency of the area utilization. For secure bonding, significant area on the wafer is occupied by the surfaces to be bonded, and therefore fewer devices can be fabricated in a single wafer. This limitation can be overcome by improving the control of the frit flow and reducing the "safety zone" between the device and the frit. However, to ensure hermeticity, glass requires a thicker seal width than metal or silicon.

Eutectic bonding is a more recent development that addresses the limitation of glass frit bonding without sacrificing the bonding topography flexibility of glass frit bonding. Using metal also enables thinner sealing widths since metal has 100 times less permeability compared to glass [9]. Au–Sn bonding works well with most processing conditions, but since gold is not compatible with most fabrication environments, metal layers such as Al–Ge [10] or Cu–Sn [11] are the current focus in research and manufacturing focus. Since metal has much less diffusibility than glass, a finer sealing ring can be realized, enabling chip sizes to further reduce for higher profitability.

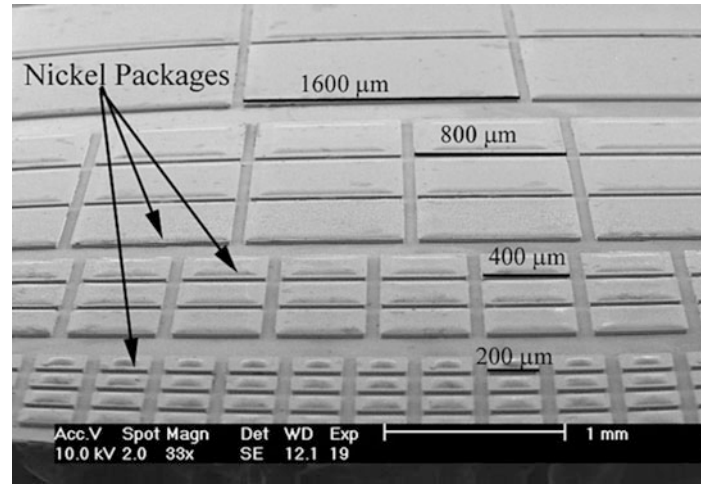
Thin-Film Packaging

Unlike the wafer-bonding approach, thin-film packaging techniques use deposited thin films as a cap layer. As shown in Fig. 8, MEMS structures are covered first by a sacrificial material, and then by a deposited cap material. Vent holes are patterned in the cap layer to remove the sacrificial material, followed by hole sealing. This approach can maximize the area utilization since the area needed for packaging is minimal; however, the choice of materials and processes becomes more problematic. For example, both MEMS device and sacrificial materials that can sustain the high temperature and the corrosive processing environment of the cap layer deposition must be selected. The current status of thin-film packaging methods is less mature relative to wafer-bonding approaches.

The University of Michigan has developed a thin-film packaging technology that utilizes nickel as the cap layer (see Fig. 9) [12]. Specifically, photoresist was used as a sacrificial layer followed by a nickel cap layer electroplating step. Solder bumping has been identified as the most effective sealing method among other approaches (localized welding and evaporated/sputtered/seals). A cavity pressure of ~1.5 Torr was measured at the time of the fabrication and a slight



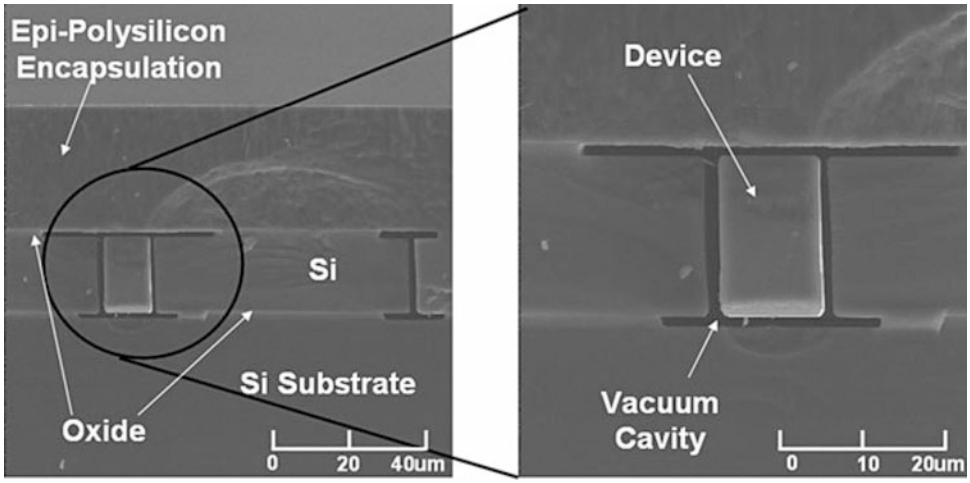
MEMS Packaging, Fig. 8 (**a**) A sacrificial material covers released MEMS devices. (**b**) A cap layer covers the sacrificial material. (**c**) Vent holes are patterned in the cap layer and the sacrificial material is removed through the vent holes. (**d**) The vent holes are sealed to realize MEMS devices in the packaged cavities



MEMS Packaging, Fig. 9 An SEM image of multiple nickel packaged MEMS devices (Reprinted by permission [12])

increase was detected during the 30 days of monitoring. At a maximum of 250 °C, this packaging technology has an advantage with respect to process temperature. This low temperature allows the realization of MEMS packaging of wafers containing CMOS circuitry without compromising their functionality; i.e., MEMS and CMOS integration is possible with CMOS first.

Epi-seal encapsulation technology (see Fig. 10) was developed in collaboration between Stanford University and Bosch [5]. The packaging utilizes epitaxially deposited silicon as both a cap and a sealing layer. Silicon dioxide is used as the sacrificial layer. By the nature of the high temperature epitaxial deposition (∼1,000 °C), most of the contaminants and native oxide are eliminated, leaving an extremely clean surface and an encapsulated cavity. Long-term hermeticity of the package was investigated and the cavity pressure remained at less than 10 mTorr over a year while exposed to atmospheric pressure. This approach has also the advantage of potentially integrating with CMOS circuitry as a MEMS-first approach. However, the materials



MEMS Packaging, Fig. 10 An SEM image of a cross section view of a device fabricated by the epi-seal encapsulation process. The polysilicon cap layer creates a hermitically sealed enclosure (Reprinted by permission [5])

and processes that can be used for the MEMS structures are limited due to the high temperature of the sealing deposition (1,000 °C).

Die-Level Packaging

Die-level packaging was the first form of MEMS packaging, and is still used for many types of devices. Although wafer-level packaging options are often preferred in many MEMS applications, such methods cannot be applied to devices such as most pressure sensors, digital light processing (DLP) mirrors, thermo-accelerometers (which need sufficient gas volume), and some quartz-based piezoelectric gyroscopes.

After the devices are released, a special singulation process (dicing) is required. Temporary protective tape with cavities can be used to encapsulate the released die prior to the regular dicing, or stealth dicing using IR laser, can be used to dice without any debris or liquid involved. After singulation, careful assembly process in a clean room environment (at least class 1,000) is needed to avoid debris on the die.

Cavity packages are used for die-level packaging. These are packages with premade cavities. After the die is attached and electrically connected, the lid is sealed by soldering or by welding. Die-level packaging is more expensive than wafer-level packaging since it requires special treatment of the die, a clean room for assembly, and more expensive packages.

Cross-References

- ▶ Biosensors
- ▶ Capacitive MEMS Switches
- ▶ Gyroscopes
- ▶ MEMS Display based on Interferometric Modulation Technology
- ▶ Micropumps
- ▶ Nanoencapsulation
- ▶ Piezoelectric MEMS Switch
- ▶ Self-assembly

References

1. Dokmeci, M.R.: Accelerated testing of anodically bonded glass-silicon packages in salt water. In: Proceeding of the 9th International Conference of on Solid-State Sensors, Actuators, and Microsystems Transducers 97, Chicago, vol. 1, pp. 283–286 (1997)
2. Gooch, R.: Wafer-level vacuum packaging for MEMS. J. Vac. Sci. Vac. Technol. A **17**(4), 2295–2299 (1999)
3. Kim, B.: Hermeticity and diffusion investigation in polysilicon film encapsulation for microelectromechanical systems. J. Appl. Phys. **105**(1), 013514–10 (2009)
4. Freescale Semiconductor 2011, MEMS-Based Sensor Technology, digital image. http://www.freescale.com/webapp/sps/site/overview.jsp?code=SNSMEMSOVERVIEW (2011). Accessed 29 Mar 2011

5. Kim, B.: Frequency stability of wafer-scale film encapsulated silicon based MEMS resonators. Sensor. Actuat. A:Phys. **136**(1), 125–131 (2007)
6. DeNatale, J.: RF MEMS reliability. In: Proceeding of the 12th International Conference of on Solid-State Sensors, Actuators, and Microsystems Transducers 2003, vol. 2, Boston, pp. 943–946 (2003)
7. Newman, H.S.: Lifetime measurements on a high reliability RF-MEMS contact switch. IEEE Micro. Wirel. Compon. Lett. **18**(2), 100–102 (2008)
8. Knechtel, R.: Glass frit bonding: an universal technology for wafer level encapsulation and packaging. Microsyst. Technol. **12**, 63–68 (2005)
9. Stroehle, D.: On the penetration of gases and water vapor into packages with cavities and a maximum allowable leak rate. In: Proceedings 15th Annual Reliability Physics Symposium, Las Vegas, pp. 101–106 (1977)
10. Park, W.-T.: Microstructure and mechanical properties of aluminum-germanium eutectic bonding with polysilicon metallization for microelectromechanical systems (MEMS) packaging. Scripta Mater. **64**, 733–736 (2011)
11. Li, L.: Cu/Sn isothermal solidification technology for hermetic packaging of MEMS. In: Proceedings of 1st International Conference on NEMS, Zhuhai, pp. 1133–1137 (2006)
12. Stark, B.: A low-temperature thin-film electroplated metal vacuum package. J. Microelectromech. Syst. **13**(2), 147 (2004)

MEMs Pumps

► Micropumps

MEMS Resonant Infrared Detectors

► MEMS Resonant Infrared Sensors

MEMS Resonant Infrared Sensors

Yu Hui and Matteo Rinaldi
Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA

Synonyms

MEMS resonant infrared detectors; MEMS resonant thermal detectors; MEMS resonant thermal sensors

Definition

A microelectromechanical system (MEMS) resonant infrared (IR) sensor is a particular class of thermal detectors that relies on a transduction scheme based on the change in vibration frequency of a microelectromechanical resonator with a temperature dependent mechanical resonance frequency. The incident IR radiation heats the MEMS resonator changing its resonance frequency. By monitoring such resonance frequency shift, the incident IR power is detected. Unlike other IR sensors, MEMS resonant IR sensors do not require cooling and are suitable for the implementation of high performance, miniaturized, and power efficient thermal imagers and spectrometers.

Principle of Operation

The core of a MEMS resonant IR detector is a microelectromechanical resonator whose resonance frequency is highly sensitive to temperature. An IR absorbing material is typically integrated on the surface of the resonator. The freestanding micromechanical resonant structure can be simply modeled as a thermal mass (with thermal capacitance C_{th}) coupled to the heat sink at a constant temperature, T_0, via a thermal conductance, G_{th} (thermal resistance, $R_{th} = 1/G_{th}$). Upon exposure to IR radiation, the absorbed power causes a large and fast increase of the device temperature, ΔT, due to the excellent thermal isolation and extremely low thermal mass of the freestanding micromechanical structure (released from the substrate):

$$\Delta T = \frac{\eta Q_p}{\sqrt{{G_{th}}^2 + \omega^2 {C_{th}}^2}} \tag{1}$$

where Q_p is the incident IR power, η is the absorption coefficient of the material stack forming the resonator (typically set by the IR absorbing material), ω is the modulation frequency of the incident IR radiation, respectively. Such IR-induced temperature rise results in a shift in the mechanical

resonance frequency of the micromechanical resonator due to the large temperature dependence of the structure equivalent Young's modulus. Absolute values of temperature coefficient of frequency, $|TCF|$, in the 10–100 ppm/K range are typically achieved in MEMS resonators [1–4]. Therefore, for IR radiation slowly varying over time ($\omega \approx 0$), the overall responsivity of the MEMS resonant detector to IR radiation can be expressed according to Eq. 2:

$$R_s = \frac{\Delta f}{Q_p} = \eta \cdot R_{th} \cdot TCF \cdot f_0 \tag{2}$$

Taking advantage of the high quality factor, Q, resonant system (unique feature compared to conventional microbolometers), extremely small IR radiation-induced frequency variations can be detected by a compact and low power electronic readout (i.e., a single transistor oscillator circuit). In fact, higher resolution and stability, compared to static measurements, is typically achieved in the MEMS domain by exploiting the high Q resonant behavior of a micromechanical structure. Such MEMS resonant sensors use frequency as the output variable, which is one of the physical quantities that can be monitored with the highest accuracy and converted to digital form by measuring zero-crossings without the need of an analog-to-digital-converter to digitize the voltage or current readout.

A crucial parameter that ought to be considered for the design and optimization of a MEMS resonant IR sensor is the noise equivalent power, *NEP*, which represents the IR power that gives a signal-to-noise ratio equal to 1 in a 1 Hz measurement bandwidth (minimum detectable IR power) and it is defined as the noise-induced frequency fluctuation, f_n, divided by the responsivity of the detector, given by:

$$NEP = \frac{f_n}{R_s} = \frac{f_n}{\eta \cdot R_{th} \cdot TCF \cdot f_0} \tag{3}$$

Among different noise sources, the fundamental limit to the frequency stability of a MEMS resonant IR sensor is set by [5]: (1) the thermal fluctuation noise associated with the spontaneous temperature fluctuations of the detector element due to the finite heat conductance G_{th} to the surroundings, (2) the background fluctuation noise due to radiative heat exchange with the environment, and (3) the thermomechanical noise originated from thermally driven random motion of the mechanical structure. Therefore, the minimization of the *NEP* values associated with each of these three fundamental noise contributions (respectively NEP_{th}, NEP_{rad}, and NEP_{mec}) is typically used to drive the design of MEMS resonant IR sensors:

$$NEP_{th} = \frac{2T_0}{\eta}\sqrt{\frac{K_B}{R_{th}}} \tag{4}$$

$$NEP_{rad} = \frac{1}{\eta}\sqrt{16A\varepsilon\sigma k_B T_0^5} \tag{5}$$

$$NEP_{mec} = \sqrt{\frac{K_B T_0}{4P_c}}\frac{1}{\eta \cdot R_{th} \cdot TCF \cdot Q} \tag{6}$$

where K_B is the Boltzmann constant, T_0 is the temperature of the resonator, A is the area of the device, ε is the emissivity, σ is the Stefan-Boltzmann constant, P_c is the power used to drive the mechanical resonance in the structure, Q is the resonator's quality factor. The total *NEP* can be expressed as:

$$NEP_{tot} = \sqrt{NEP_{th}^2 + NEP_{rad}^2 + NEP_{mec}^2} \tag{7}$$

Equations 4, 5, and 6 indicate that minimum *NEP*, thus a high resolution resonant IR detector, can be achieved by maximizing the device thermal resistance R_{th}, absorption coefficient η, temperature coefficient of frequency, *TCF*, quality factor Q, and power handling P_c. The minimization of detector *NEP* also minimizes its noise equivalent temperature difference (*NETD*) which is a widely used figure of merit for IR detector arrays, especially IR imaging systems based on focal plane arrays (*FPA*), representing the minimum temperature change of the scene that induces a detector output signal equal to the noise level [6]:

$$NETD = NEP \frac{(4F^2+1)B^{1/2}}{\eta A \gamma} \left[\int_{\lambda_a}^{\lambda_b} \frac{\partial M(\lambda)}{\partial T} d\lambda \right]^{-1} \quad (8)$$

where F is the focal ratio of the optics, B is the measurement bandwidth and γ is the transmission of the optical path between the IR source to the IR detector, and the last term on the right-end side of the equation is the radiant exitance integrated over the wavelength range of interest [6]. It is worth noting that, beyond the minimization of the *NEP*, a small focal ratio of the optics and a large absorption coefficient are generally desirable for the implementation of high resolution infrared imagers.

Another crucial parameter to be considered for the design of a high performance MEMS IR resonant sensor is its response time which is characterized by the thermal time constant, τ, of the device, defined as the product between its thermal resistance, R_{th}, and capacitance, C_{th}:

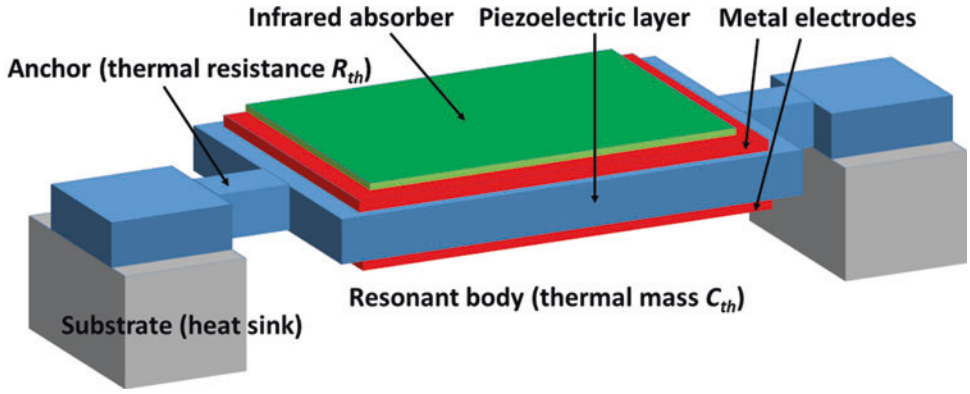
$$\tau = R_{th} \cdot C_{th} \quad (9)$$

Equation 9 indicates that by improving the thermal isolation of the sensing element from the heat sink (hence increasing R_{th}), the *NEP* of the device is reduced but its response time is increased. Therefore, a trade-off between these two important performance metrics needs to be generally considered for the design of the detector. Nevertheless, for a given thermal resistance (guaranteeing a satisfactory *NEP*) the response time of the sensor can be reduced by minimizing its thermal capacitance which directly translates into reducing the volume of the micromechanical resonant structure.

Main Research Accomplishments

Different micro- and nanoelectromechanical system (MEMS/NEMS) resonator technologies have been explored for the implementation of resonant IR sensors [1–3, 5, 7–12]. Among them, thin-film piezoelectric MEMS resonators have shown a great potential, thanks to their intrinsic efficient on-chip piezoelectric transduction mechanics. In particular, the efficient on-chip piezoelectric actuation and sensing of high frequency bulk acoustic modes of vibration in micro-/nano-plate structures (Fig. 1), instead of conventional beams, have enabled the fabrication of high frequency and high power handling resonators that can be directly connected to compact and power efficient electronic readouts [9, 13] (without the need of off-chip bulky piezoceramic disks for actuation and complex and power inefficient optical readouts for displacement detection), which is a crucial requirement for the implementation of miniaturized thermal imagers and IR sensing systems.

The idea of using piezoelectric MEMS resonators for the implementation of high performance



MEMS Resonant Infrared Sensors, Fig. 1 Mock-up view of a typical thin-film piezoelectric MEMS resonant infrared sensor

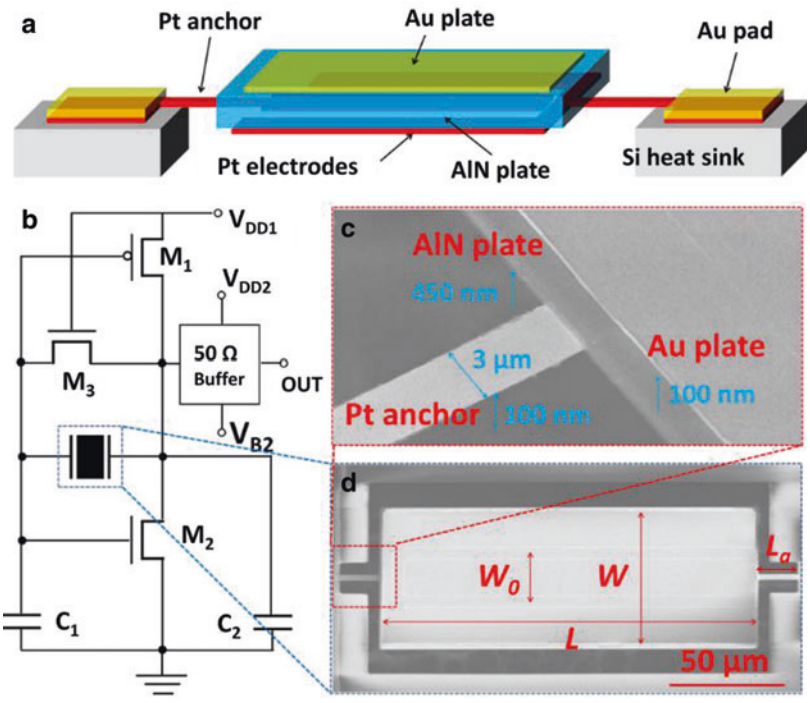
MEMS resonant infrared sensors was first introduced in 1996 [11]. Theoretical analysis predicted that an array of quartz microresonators working in the 200 MHz to 1 GHz range could be employed for the implementation of uncooled infrared imaging systems with *NETD* values lower than 10 mK. The first compelling experimental demonstration of a high performance MEMS resonant infrared sensor based on a micromachined Y-cut-quartz film (6.9 μm thick) resonator (working at 241 MHz with a quality factor, Q of 4200) was reported in 2011 [3]. Given the good IR absorbance of quartz in the 8–12 μm spectral range [3], no additional IR absorbing material was required. Thanks to the large temperature coefficient of frequency, $|TCF|$ of the Y-cut quartz microresonator (92 ppm/K) and its relatively large thermal isolation, a responsivity of 11.4 Hz/μW was demonstrated. The *NEP* of the device was extracted to be 3.9 nW/Hz$^{1/2}$, thanks to the low noise performance of the resonator (0.0445 Hz/Hz$^{1/2}$). The thermal time constant of the infrared sensor was also predicted by three-dimensional finite element method (FEM) time domain simulation and found to be 5.92 ms.

A fundamental limitation of quartz-based MEMS resonant IR detectors is the complexity associated with high density micromachining of quartz substrates (i.e., substantial thinning of the quartz substrate is required to produce a few microns thick quartz film), which poses some significant challenges in terms of yield and reproducibility of the device fabrication process, miniaturization, and dense integration of resonator arrays in a small footprint. In this context, the use of a thin-film piezoelectric material (such as aluminum nitride or gallium nitride) instead of a bulk piezoelectric material (such as quartz) can potentially reduce significantly the complexity of the microfabrication process (using conventional micromachining of thin-film materials) and enable the implementation of high performance, miniaturized, and highly integrated resonant sensors on a silicon substrate.

A MEMS resonant IR sensor based on a gallium nitride (GaN) thin-film piezoelectric resonator was recently demonstrated [2]. The core of the device consisted of a GaN (1.4 μm thick) on silicon (10 μm thick) micromechanical resonator, working at 101 MHz with a quality factor Q of 280 (the 10 μm thick silicon layer was necessary to reduce the relatively high stress of the epitaxially grown GaN thin-film). The $|TCF|$ of the GaN resonator was measured to be ~3 kHz/K (29.7 ppm/k) and the responsivity of the infrared sensor was extracted to be ~1.7 Hz/μW (1.68 %/W). The frequency stability of the device was characterized showing a ~8 Hz (~80 ppb) frequency fluctuation in a 1 Hz measurement bandwidth (~8 Hz/Hz$^{1/2}$) which translated in an estimated *NEP* of ~4.7 μW/Hz$^{1/2}$. The thermal time constant of the sensor was also experimentally measured and found to be ~556 μs.

Despite the demonstrated potential of quartz and GaN MEMS IR sensors, these technologies are fundamentally limited by the lack of cost-effective and low temperature deposition techniques capable of producing high quality piezoelectric quartz or GaN thin films on an arbitrary substrate. In this context, the aluminum nitride (AlN) MEMS technology offers some unique advantages compared with other piezoelectric resonant technologies: ultrathin (10s ~ 100 s nm) and high quality AlN film can be directly deposited on silicon substrates by low-temperature sputtering process [14], enabling the fabrication of ultralow volume resonators with excellent electromechanical performance. Furthermore, the post-CMOS compatibility of the microfabrication process used for the fabrication of AlN MEMS resonant devices [15] enables the monolithic integration of resonant sensors and the CMOS electronic readout on the same substrate which is an attractive feature for the implementation of ultraminiaturized, high performance, and low power sensing platforms.

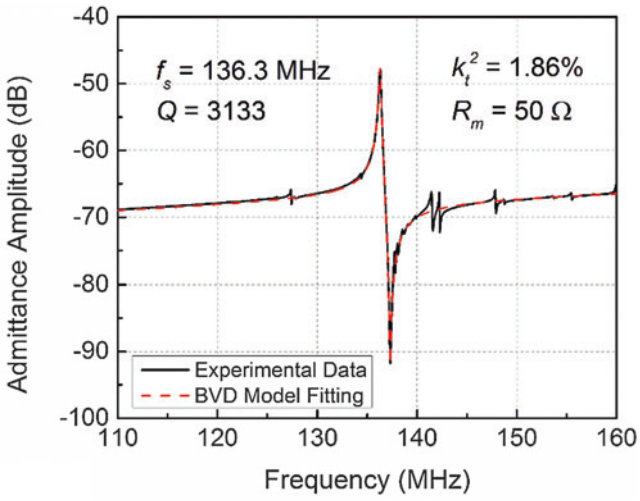
A step stone toward the development of MEMS resonant IR thermal detectors with high sensitivity, high resolution, fast response, and compact readout has been recently set with the introduction of a novel micromechanical resonant structure, based on the use of ultrathin (100 nm) fully metallic tethers to support the resonant body of an AlN piezoelectric nano-plate resonator (NPR) [9]. Such innovative device concept enables the implementation of uncooled IR



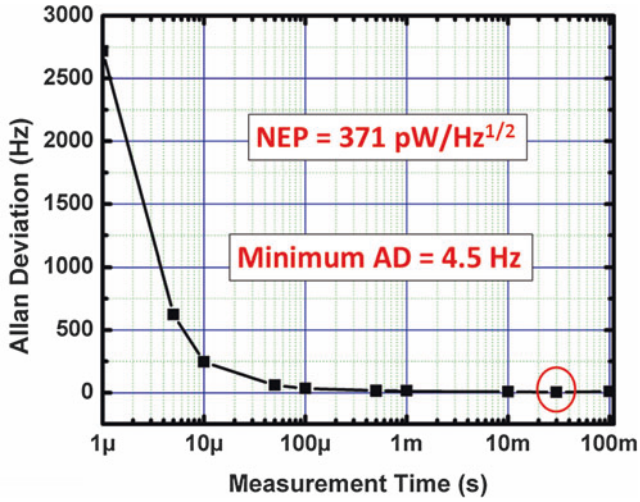
MEMS Resonant Infrared Sensors, Fig. 2 (**a**) Mock-up view of the AlN piezoelectric nano-plate resonator with nanoscale Pt anchors; (**b**) complementary metal-oxide-semiconductor readout circuit for the AlN MEMS resonant infrared sensor; (**c**–**d**) scanning electron microscopy image of the nanoscale metallic anchor and the AlN nano-plate resonator: $L =$ 200 μm, $W = 75$ μm, $W_0 = 25$ μm

detectors with a unique set of application enabling features. High $|TCF|$ (~ 30 *ppm/K*) is guaranteed by the large temperature coefficient of Young's modulus of the material stack forming the resonant nano-plate, and maximum thermal isolation of the device resonant body from the heat sink (maximum R_{th}) is readily achieved by completely removing the relatively thick piezoelectric material from the anchors, hence minimizing their thicknesses (ultimately limited by the need of a thin metal layer for low electrical resistance routing), which results in a resonant thermal detector with enhanced responsivity, R_s ($\sim Hz/nW$). At the same time, a low thermal time constant ($\sim ms$) is achieved by exploiting the unique capability of depositing high quality ultrathin AlN films ($< 1\ \mu m$) on a Si substrate by low temperature sputtering process which enables the implementation of low volume nano-plate resonant structures with a significantly reduced thermal mass (low C_{th}). Despite the resonator volume scaling, low noise performance is achieved thanks to the unique piezoelectric transduction properties of AlN ultrathin films which enable efficient on-chip piezoelectric actuation and sensing of a high quality factor ($Q > 2000$) and high frequency bulk acoustic mode of vibration (with associated high power handling, P_c) in a freestanding nano-plate, enabling the implementation of low *NEP* ($\sim pW/Hz^{1/2}$) resonant thermal detectors.

The electromechanical performance of such AlN nano-plate resonator with nanoscale Platinum (Pt) anchors was characterized by measuring its admittance amplitude versus frequency and performing Butterworth-Van Dyke (BVD) modeling fitting (Fig. 2) [16]. Despite the device volume reduction, high quality factor Q (3133) and electromechanical coupling coefficient k_t^2 (1.86 %) were attained. Such high electromechanical performance (Figure of Merit FOM $= k_t^2 \cdot Q_M \approx 58$) enabled direct wire bonding of the AlN nano-plate resonator to an oscillator circuit chip fabricated in the ON Semiconductor 0.5 μm CMOS process (Fig. 3) resulting in the first complete and compelling prototype of a high performance MEMS resonant thermal detector with a compact and power efficient Complementary Metal-Oxide-Semiconductor oscillator as direct frequency readout (Fig. 3).

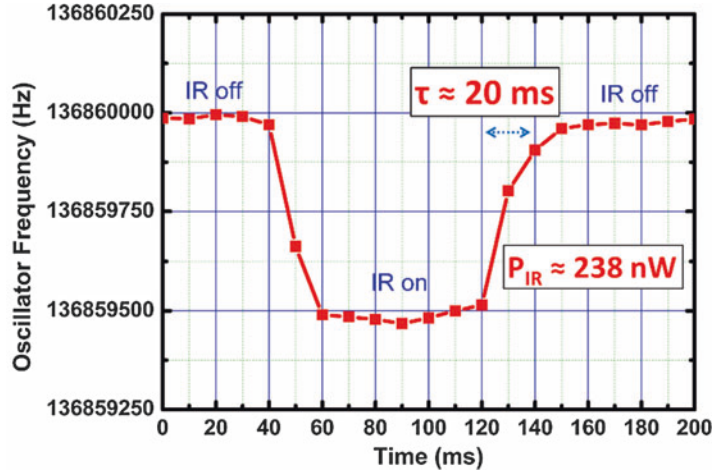


MEMS Resonant Infrared Sensors, Fig. 3 Experimental and fitted (BVD model) admittance curves of the fabricated AlN nano-plate resonator with nanoscale Pt anchors

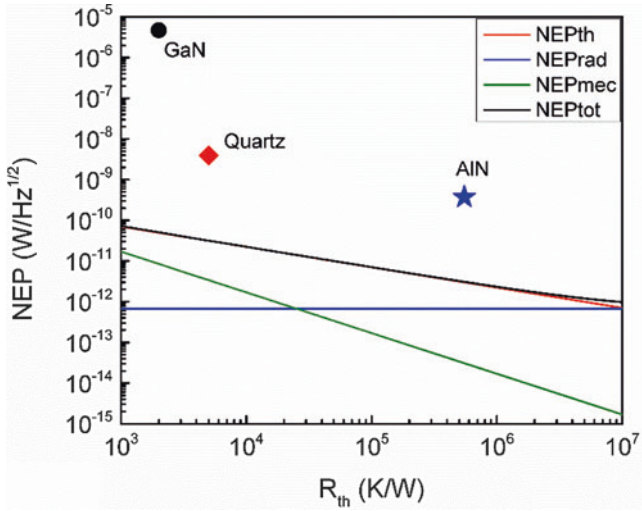


MEMS Resonant Infrared Sensors, Fig. 4 Measured Allan deviation (*AD*) of the AlN piezoelectric MEMS resonant infrared sensor

The $|TCF|$ of the AlN nano-plate resonator with nanoscale Pt anchors was measured using a temperature-controlled radio frequency probe station and found to be 3.8 kHz/K (28 ppm/K). The thermal resistance of the device was estimated by performing three-dimensional FEM simulation using COMSOL and found to be 5.5×10^5 K/W. The responsivity of the infrared sensor was calculated by multiplying the $|TCF|$ by the thermal resistance, and found to be 2.1 Hz/nW, which is one order of magnitude higher than the one achieved using conventional AlN nano-plate resonators with anchors made out of a thicker AlN/Pt material stack. The short-term frequency stability of the oscillator output signal was characterized by measuring its Allan deviation [17], which is a measurement of the noise-induced frequency fluctuation. A minimum Allan deviation was found to be 4.5 Hz for a measurement time of 30 ms (Fig. 4), translating into a noise-induced frequency fluctuation of ~33 ppb and a noise spectral density of ~0.78 Hz/Hz$^{1/2}$. The *NEP* of the infrared detector was extracted by dividing the measured noise spectral density by the responsivity and found to be ~371 pW/Hz$^{1/2}$. Such *NEP* is one order of magnitude lower than



MEMS Resonant Infrared Sensors, Fig. 5 Measured response of the AlN nanoplate MEMS-CMOS prototype to NIR radiation



MEMS Resonant Infrared Sensors, Fig. 6 Noise equivalent power (*NEP*) for different values of thermal resistance (R_{th}). The *solid lines* indicate the calculated NEP values associated with each of the three fundamental noise contributions (thermal fluctuation noise, NEP_{th}; background fluctuation noise due to radiative heat exchange with the environment, NEP_{rad}; thermomechanical noise, NEP_{mec}; and NEP_{tot} the total noise as expressed in Eq. 7), assuming: resonator area = 200 × 75 μm, $\varepsilon = 1$, $T_0 = 300$ K, $P_c = 0$ dBm, |TCF| = 30 ppm/K, $Q = 2000$. The individual data points indicate the demonstrated performance of the state-of-the-art MEMS resonant IR sensors based on quartz [3], GaN [2], and AlN [9] piezoelectric resonator technologies

the ones achieved with other piezoelectric MEMS resonant IR detector technologies [3].

The frequency response of the device prototype to near infrared (NIR) radiation was characterized using a Bruker Vertex 70 FTIR and Hyperion microscope. A broadband IR radiation from a tungsten halogen lamp (0.25–100 μm) was optically filtered (0.33–2.8 μm transmission) and chopped (6.25 Hz) before it was coupled into a Hyperion microscope used to focus the light onto the resonant body of the device. The frequency variation of the MEMS-CMOS oscillator was monitored using an Agilent 53230A frequency counter (Fig. 5). The measurement was performed in air. Despite the fact that no specific IR absorbing element was integrated with the AlN resonator, an absolute frequency shift of ~ 500 Hz was recorded when the device was exposed to NIR

radiation. Such frequency shift translates into an absorbed NIR power of ~238 nW (minimal IR absorption due to the intrinsic material loss), considering the device absolute responsivity of 2.1 Hz/nW. A thermal time constant of ~20 ms was also experimentally recorded (Fig. 5).

The demonstrated MEMS-CMOS resonant IR detector prototype, based on an AlN NPR with ultrathin Pt anchors, reduced the performance gap with the theoretical predicted limits (Fig. 6), setting a milestone toward the development of a new technology platform that, by taking advantage of the unique transduction properties and scaling capabilities of AlN nano-plate resonators, can potentially deliver fast (~ms) and high resolution ($NEP \sim 1$ pW/Hz$^{1/2}$, $NETD \sim 1$ mK) uncooled IR detectors suitable for the implementation of high performance, miniaturized, and power efficient IR imaging systems. Possible approaches to further improve the performance of these AlN MEMS resonant IR detectors include: optimization of the device thermal isolation (investigating optimal materials and geometries for the device anchors); further scaling of the device volume maintaining high electromechanical performance; proper engineering of the absorption properties of the ultrathin piezoelectric resonant structure; and investigation of strategies to reduce the effect of all the noise sources contributing to the generation of frequency fluctuations (such as the resonator flicker noise, random walk, and drifts).

Cross-References

- ► Laterally Vibrating Piezoelectric Resonators
- ► NEMS Resonant Chemical Sensors

References

1. Hui, Y., Rinaldi, M.: Fast and high resolution thermal detector based on an aluminum nitride piezoelectric microelectromechanical resonator with an integrated suspended heat absorbing element. Appl. Phys. Lett. **102**, 093501 (2013)
2. Gokhale, V.J., Rais-Zadeh, M.: Uncooled infrared detectors using Gallium Nitride on silicon micromechanical resonators. J. Microelectromech. Syst. **23**, 803–810 (2014)
3. Pisani, M.B., Ren, K., Kao, P., Tadigadapa, S.: Application of micromachined-cut-quartz bulk acoustic wave resonator for infrared sensing. J. Microelectromech. Syst. **20**, 288–296 (2011)
4. Kuypers, J.H., Lin, C.-M., Vigevani, G., Pisano, A.P.: Intrinsic temperature compensation of aluminum nitride Lamb wave resonators for multiple-frequency references. In: Frequency Control Symposium, 2008 I.E. International Honolulu, Hawaii, USA, pp. 240–249 (2008)
5. Yi, F., Zhu, H., Reed, J.C., Cubukcu, E.: Plasmonically enhanced thermomechanical detection of infrared radiation. Nano Lett. **13**, 1638–1643 (2013)
6. Rogalski, A.: Infrared Detectors, 2nd edn. CRC Press, Boca Raton (2010)
7. Zhang, X., Myers, E., Sader, J., Roukes, M.: Nanomechanical torsional resonators for frequency-shift infrared thermal sensing. Nano Lett. **13**, 1528–1534 (2013)
8. Hui, Y., Rinaldi, M.: High performance NEMS resonant infrared detector based on an aluminum nitride nano-plate resonator. In: Solid-State Sensors, Actuators and Microsystems (TRANSDUCERS & EUROSENSORS XXVII), 2013 Transducers & Eurosensors XXVII: The 17th International Conference on Barcelona, Spain, pp. 968–971 (2013)
9. Hui, Z.Q.Y., Hummel, G., Rinaldi, M.: Pico-Watts range uncooled infrared detector based on a freestanding piezoelectric resonant microplate with nanoscale metal anchors. In: Proceedings of the 2014 Solid-State Sensors, Actuators and Microsystems Workshop (Hilton Head 2014), Hilton Head Island, pp. 387–390 (2014)
10. Wang, Z., Qiu, X., Chen, S.J., Pang, W., Zhang, H., Shi, J., et al.: ZnO based film bulk acoustic resonator as infrared sensor. Thin Solid Films **519**, 6144–6147 (2011)
11. Vig, J.R., Filler, R., Kim, Y.: Uncooled IR imaging array based on quartz microresonators. J. Microelectromech. Syst. **5**, 131–137 (1996)
12. Cabuz, C., Shoji, S., Fukatsu, K., Cabuz, E., Minami, K., Esashi, M.: Fabrication and packaging of a resonant infrared sensor integrated in silicon. Sensors Actuators A Phys. **43**, 92–99 (1994)
13. Hui, Y., Rinaldi, M.: Aluminum nitride nano-plate infrared sensor with self-sustained CMOS oscillator for nano-watts range power detection. In: European Frequency and Time Forum & International Frequency Control Symposium (EFTF/IFC), 2013 Joint, Prague, Czech Republic, pp. 62–65 (2013)
14. Zaghloul, U., Piazza, G.: Synthesis and characterization of 10 nm thick piezoelectric AlN films with high c-axis orientation for miniaturized nanoelectromechanical devices. Appl. Phys. Lett. **104**, 253101 (2014)
15. Olsson, R.H., Wojciechowski, K.E., Baker, M.S., Tuck, M.R., Fleming, J.G.: Post-CMOS-compatible aluminum nitride resonant MEMS accelerometers. J. Microelectromech. Syst. **18**, 671–678 (2009)

16. Hui, Y., Nan, T., Sun, N.X., Rinaldi, M.: High resolution magnetometer based on a high frequency magnetoelectric MEMS-CMOS oscillator. J. Microelectromech. Syst. **24**, 134–143 (2014)
17. Allan, D.W.: Statistics of atomic frequency standards. Proceedings of the IEEE **54**, 221–230 (1966)

MEMS Resonant Thermal Detectors

▶ MEMS Resonant Infrared Sensors

MEMS Resonant Thermal Sensors

▶ MEMS Resonant Infrared Sensors

MEMS Resonators Integrated with Circuits

▶ CMOS-MEMS Resonators

MEMS Vacuum Electronics

Anisullah Baig, Diana Gamzina, Jinfeng Zhao, Youngmin Shin, Robert Barchfeld, Larry R. Barnett, Calvin Domier and Neville C. Luhmann Jr.
Department of Electrical and Computer Engineering, University of California, Davis, CA, USA

Synonyms

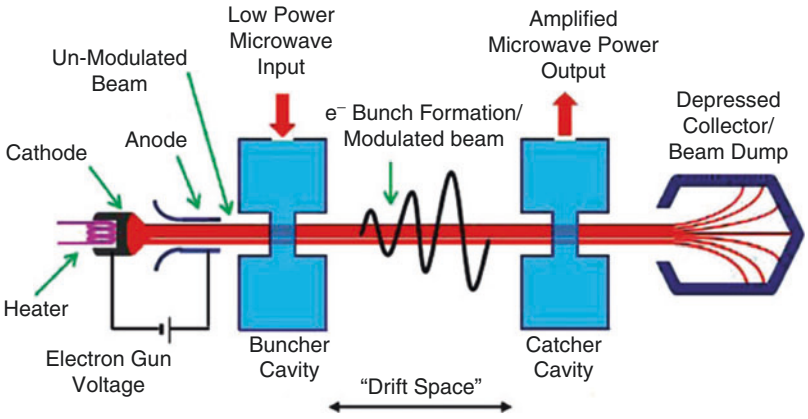
MEMS fabricated microwave devices; Microfabricated MVED (Microwave Vacuum Electron Devices); μVED (Micro Vacuum Electronic Devices)

Definition

Vacuum electronics is the area of research focused on the generation of coherent electromagnetic radiation, in which kinetic energy of the collisionless ballistic motion of electrons in vacuum is converted to electromagnetic (RF/microwave) energy [1–3]. At high frequencies (100 GHz and above), unconventional fabrication techniques for features with sizes 0.01–2 mm, tolerances <1 μm, and surface roughness 50–200 nm are employed with the various techniques generally called MEMS or micro-electro-mechanical systems.

The operating principle and constituent parts of a typical Microwave Vacuum Electron Device are shown in Fig. 1.

The heart of a Microwave Vacuum Electron Device (MVED) is a high current emission cathode. A thermionic cathode emits a high current density beam of electrons when heated. The electron beam is accelerated by the applied high voltage between the cathode and anode. This whole setup is commonly called an electron gun assembly. The first cavity of a typical Klystron is excited by a low power RF input signal using a coaxial line loop or a waveguide aperture. The electrons that lie on the crest of the RF get accelerated and the ones which lie on a trough get decelerated; this results in velocity modulation of the electron beam and subsequent bunching after appropriate drift. Due to this bunch formation, the first cavity is also called a "buncher" cavity. The electron beam passes through a "drift space" under magnetic focusing to maximize the ac current in the second cavity, called a "catcher" cavity. The ac current in the second cavity excites high intensity RF fields and amplification of the RF signal takes place. The high RF/microwave power is extracted through the output coupler attached to the output cavity. The spent electron beam goes to a copper block called collector. Klystrons typically have more than two cavities for higher gain. There are various kinds of MVEDs: Klystrons, Traveling-Wave Tube Amplifiers (TWTA), Backward Wave Oscillators, Crossed-Field Amplifiers and Oscillators, Electron Masers, and Free Electron Lasers. The scope of this entry is to give an overview of



MEMS Vacuum Electronics, Fig. 1 A typical microwave vacuum electron device (Klystron)

the recent efforts to scale MVEDs to near THz frequencies for substantial applications, where advanced fabrication methods like MEMS (micro-electro-mechanical system) are considered.

Overview

Vacuum Electron Devices

Vacuum electron devices (VEDs) have been widely employed as high-power radiation sources for communications, plasma heating, thermonuclear fusion diagnostics, and radar since the Second World War. However, vacuum electronics tended to fall behind solid-state electronics in the competition for the development of efficient low to moderate power microwave/millimeter wave sources due to the rapid advance of semiconductor technologies after 1970s. The merits of the solid-state devices, e.g., low voltage, small size, and light weight, surpassed the bulky VEDs for applications requiring low to moderate power (mW to W) in the lower frequency region ($\lesssim$30–100 GHz). Nonetheless, with the advent of a demand for powerful radiation sources in the millimeter and submillimeter wave region, the classical technologies have been again paid attention to together with advances in the so-called micro-electro-mechanical systems (MEMS) or microfabrication. Unlike the collisional conduction transport of electrons in a typical solid-state device, a large portion of electron energies can be transformed to the EM (electromagnetic) energies through collisionless convection transport in vacuum with a significant fraction of the spent beam energy available for electrostatic recovery; consequently, VEDs can provide very high radiation efficiency. Furthermore, the lower ohmic loss and the higher breakdown threshold compared to the solid-state devices make it possible to handle extremely high energy density in the high-frequency regime at room temperature.

Taken together, these advantages suggest the broad applicability of THz technology as long as some problems of miniaturizing VEDs down to millimeter and submillimeter wave scales are solved.

Terahertz (THz) Region of Electromagnetic Spectrum

Despite several potentially high impact commercial and security applications in the THz region (0.1–1 THz), the source technology is not available with reasonable power in a compact package. This so-called THz gap [4] occupies a large portion, 0.1–10 THz, of the electromagnetic (EM) spectrum that lies between the microwave and infrared region.

THz-rays (T-rays) have a number of advantages like: visible light creating a photograph, radio waves transmitting sound, and x-rays allowing us to observe inside of the human body, THz waves can create images and transmit information. The photon energies of T-rays are very low, so that they are not harmful to living organisms as opposed to x-rays. THz waves with

rotational, vibrational, and translational response modes to molecules, radicals, and ions can provide a range of unique information, which is difficult to obtain from optical, x-ray, and NMR (Nuclear Magnetic Resonance) images. The peculiarity of THz reactivity is very effective to fingerprint specific materials in the THz range. It provides conformational information closely related to the biological functions of molecules in tissues and cells and inaccessible by other technical means. While dielectrics are opaque to visible light and provide only low contrast to x-rays, T-rays can easily penetrate and image inside most dielectric materials. Furthermore, T-ray imaging can provide strongly enhanced contrast because of low scattering (Rayleigh scattering) due to their longer wavelengths, whereas large amounts of scattering in conventional optical transillumination, using near-infrared pulses, tends to blur images of objects.

Challenges in THz Devices

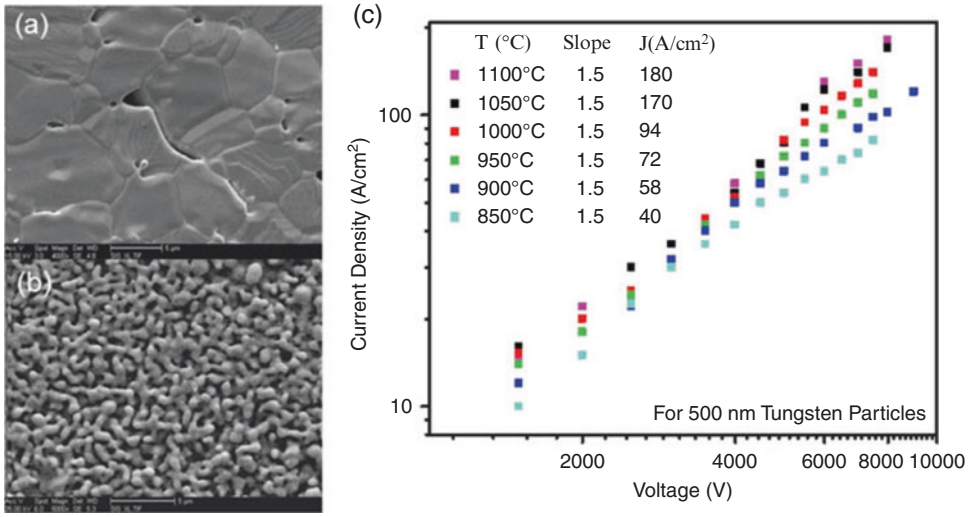
The expression μVEDs refers to high-power and high-frequency vacuum tubes based on advanced MEMS technologies. Recently, owing to high efficiency, robustness, and stability, this promising area is beginning to attract considerable attention from the THz community seeking a practical solution to advance imaging and sensing technologies. However, despite many advantages, some remaining technical issues [5] still hamper its practical applications. The most critical issues to the advancement of μVEDs are oscillation threshold and miniaturization, which severely affect device performance at increasing frequencies.

Enabling Technologies/High Current Density Electron Sources

μVEDs need high start-oscillation current density electron beams proportional to f_0^2, where f_0 is the frequency, in the millimeter and submillimeter wave region. Reduction of the electron beam cross section due to an increase of operational frequency dramatically increases the beam current density. For instance, a straightforward estimation of current density through scale-down indicates that 3 A/cm^2 at 10 GHz run corresponds to 3 × 10^4 A/cm^2 at 1 THz [6]. This is hardly achievable by conventional electron beam sources, particularly given the need to transport the electron beams with low interception. Recently, some electron emitters capable of producing a high emission current density have been challenged for THz μVED application and the electron gun issues have also been discussed [7–9]. The quality of these cathodes is usually evaluated by using scanning electron microscope images of the emitting surface (Fig. 2a, b) and by plotting testing data on logarithmically scaled I-V (Current–voltage) curves (Fig. 2c).

Scandia-Doped Tungsten Nano-composite Cathodes' Development at Beijing Vacuum Electronics Research Institute and Beijing University of Technology

Li and his colleagues at the Beijing Vacuum Electronics Research Institute (BVERI) have been investigating scandate cathodes since 1981 with a considerable portion of this work aimed at the development of impregnated scandate cathodes (I-Sc) for use in linear beam devices [10]. The work has focused on both an understanding of emission mechanisms and achieving high current density up to 100 A/cm^2 [11, 12]. Beginning in 2002, BVERI began collaborations with the materials science group at Beijing University of Technology (BJUT) with an emphasis on studies of scandia-doped dispenser (SDD) cathodes [11, 13]. In closely spaced diodes, pulsed emission in excess of 30 A/cm^2 at 850°C$_b$ and 2 A/cm^2 at 950°C$_b$ with DC loading for over 10,000 h has been obtained from SDD cathodes. In addition, the SDD cathode was tested at the Stanford Linear Accelerator Center in a cathode life test vehicle with a Pierce gun configuration, and 100 A/cm^2 for 500 h with a cathode operation temperature of 1,170°C$_b$ was achieved. Recently Liu et al. [14] at BJUT reported that SDD cathodes have reproducible high emission capability tested in open-structure Pierce electron guns. The SDD cathodes with planar emitting surfaces achieved 100 A/cm^2 SCL current densities at 950°C$_b$. For the SDD cathodes with concave emitting surfaces, 80 A/cm^2 pulsed emission has been reached at



MEMS Vacuum Electronics, Fig. 2 (**a**) Commercially available tungsten scandate cathode 311X made by SpectraMat, (**b**) scandia-doped tungsten nano-composite cathode made by UC Davis, (**c**) I-V emission curve for scandia-doped tungsten nano-composite cathode

1,025°C_b [14]. Such Sc-type cathodes offer the promise of the development of a new class of long lived and reliable microwave tubes as well as opening the way to the development of millimeter wave and THz sources. Current work is aimed at developing reliable and repeatable low-work-function scandate cathodes [8, 15]. The precise emission mechanism for scandate cathodes is still unclear. This limits the application of these cathodes. Continuing research is required to elucidate this mechanism and to establish the details of the structure and composition of the emission sites.

Scandia-Doped Tungsten Nano-composite Cathodes' Development at the University of California, Davis

Jinfeng Zhao et al. [8, 15] at UC Davis have been working on improving emission uniformity and increasing current density of porous nano-structured tungsten cathodes doped with nano-size scandium oxide and impregnated with $BaO–CaO–Al_2O_3$. Jinfeng Zhao uses sol–gel method to produce nano-size tungsten matrix doped with scandium oxide. Tungsten particle size can vary between 200 nm and 2 μm depending on desired cathode performance. The cathode matrix fabricated from the powders made by sol–gel method has high porosity, uniform grain size and scandia distribution, and open pore distribution. The uniform cathode matrix structure makes the emission surface acquire more uniform barium, oxygen, and scandium distribution, thereby improving the emission uniformity and lowering the work function of the emission surface. The UC Davis team was able to reproduce cathodes with 40 A/cm^2 current densities at 850 °C and 170 A/cm^2 current densities at 1,050 °C. Their life tests have shown that the cathode can produce 50 A/cm^2 at 1,050 °C for over 10,000 h.

Circuit Fabrication Methods/MEMS Fabrication

For the realization of a compact THz source, the lateral dimensions of the RF structure (slow wave structure or cavity) scale as ~0.1 λ_0 (~100 μm at 300 GHz) [5]. There are two fabrication constraints: (a) the fabrication tolerances for the metalized sub-mm structures should be kept typically <10 % (b) the surface roughness should be much less than the skin depth at the operating frequency. Hence, conventional machining technology is unable to meet the design specification.

MEMS Vacuum Electronics, Table 1 MEMS fabrication technologies and their typical machining capability range

Fabrication technique	Typical capability range
Conventional machining	>300 μm
EDM	80–1,000 μm
X-Ray LIGA	<1 mm
UV LIGA	10–1,000 μm
DRIE	10–500 μm
Nano wire EDM	>2 μm
Nano-CNC milling	>10 μm

MEMS fabrication has been extensively used by the semiconductor-based integrated circuit (IC) fabrication industry. More recently, vacuum electronics research has exploited the same techniques for precision fabrication of sub-mm scale, high aspect ratio micro-metallic structures. For an overview, a few techniques and their applicability range are described in Table 1. Nanoscale sinker EDM (electron discharge machining), ion beam milling, laser micro-machining, and ultrasonic machining are some of the available manufacturing techniques that could be applicable to miniaturizing of VEDs, but have not been explored extensively.

X-ray LIGA

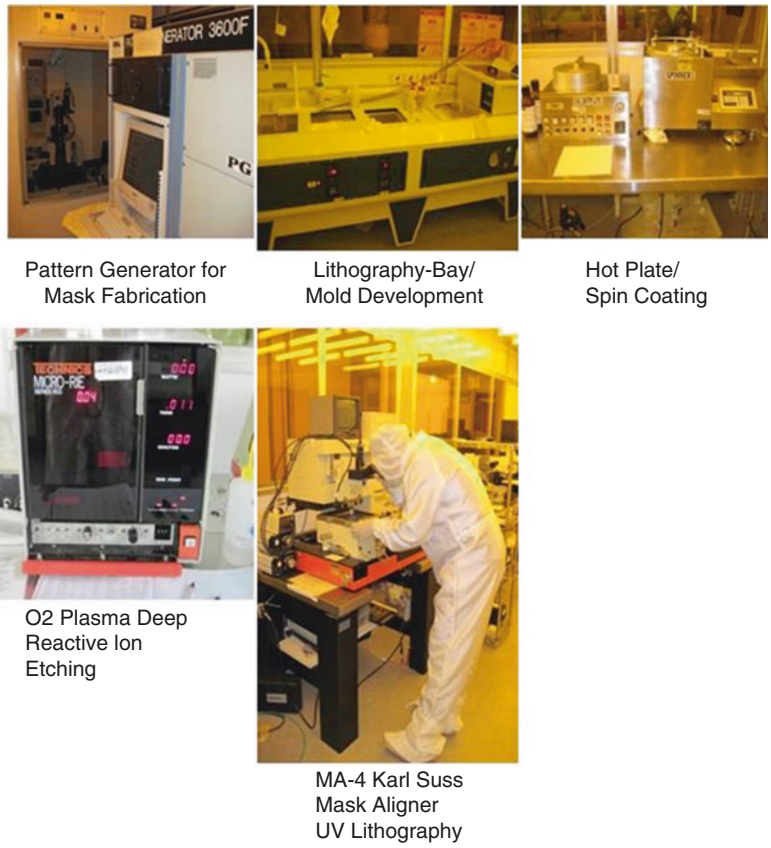
LIGA is an abbreviation of *Lithographe, Galvonoformung, und Abfornung* and translates as lithography, electroplating, and micromolding. The process uses X-rays from a synchrotron to expose a resist material through a metallic mask containing an image of the structure to be produced. Poly-methylmethacrylate (PMMA) is deposited on a metal substrate up to a maximum thickness of approximately 1 mm. In some cases, a thin titanium coating is applied to the metal substrate before deposition to improve adhesion. Following deposition (casting) it is annealed at temperatures between 110 °C and 170 °C. An X-ray mask containing the image of the desired part is placed over the PMMA, and the structure is exposed to X-rays produced by a synchrotron. Only synchrotrons can provide the dose rates required, approximately 3,000–10,000 J/cm^3, and several hours of exposure are required. Following exposure, regions of the PMMA impacted by the X-rays are removed, leaving a positive image. The removed regions are replaced by electrodeposited metal, and the structure is lapped to the desired thickness. The remaining PMMA is removed to obtain the finished part. Some issues with X-ray LIGA include: Considerable care is necessary to create high-contrast and high accuracy exposure masks, adhesion of PMMA to wafers can be troublesome, and finally, the X-ray source used for exposure is an expensive investment causing X-ray LIGA to be the least economically viable choice for inexpensive circuit production.

UV-LIGA

With the availability of improved photoresists namely SU-8 and KMPR, it became possible to use a UV (ultraviolet) source integrated into a mask aligner in a typical class-100 clean room facility that usually exists in many universities. Figure 3 shows the typical class-100 clean room equipment required for the UV-LIGA process. Figure 4 shows the LIGA process flowchart.

SU-8 LIGA

SU-8 LIGA is a lithographic process using a negative photoresist originally developed by IBM. It consists of a polymeric epoxy resin dissolved in an organic solvent together with a photo-acid generator. The SU-8 is placed on a silicon wafer to the desired thickness. It can be spun up to a thickness approaching 500 mm in a single step. Multistep applications can achieve thicknesses greater than 1 mm. Thickness nonuniformity will result in exposure nonuniformity in later processes as will any tilting of the substrate. The wafer is then baked at a low temperature to remove the solvent. SU-8 has low optical absorption in the ultraviolet (UV) range at 365-nm wavelength; consequently, the photoresist film can be used with thicknesses greater than 1 mm. When exposed to UV light, absorption of a photon reacts with the photo-acid generator to form a photo-acid acting as a catalyst. This photo-acid causes a cross-linking reaction, primarily during a postexposure bake (PEB). Cross-linking only occurs where the photo-acid is present, i.e., in regions exposed to UV radiation. The PEB



MEMS Vacuum Electronics, Fig. 3 Typical class-100 clean room equipment for UV-LIGA lithography

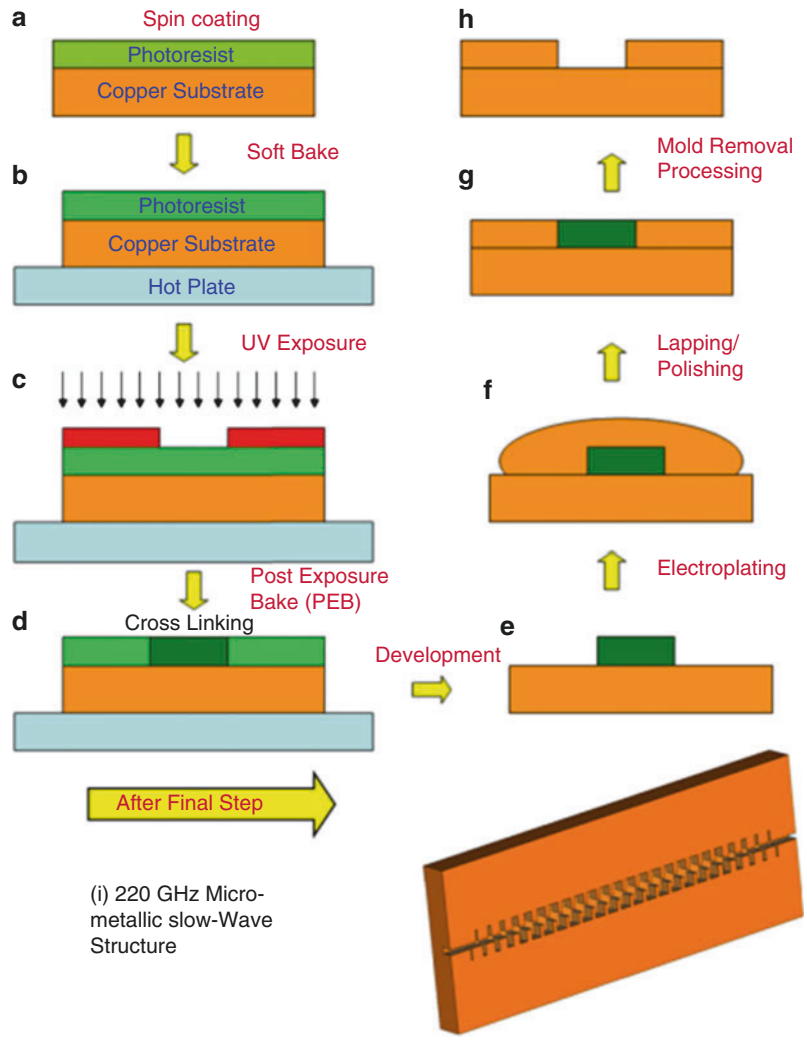
completes the cross-linking reaction. The degree of cross-linking depends on both the exposure dose and bake conditions. After PEB, the resist is developed with a two-stage process using a solvent, followed by an isopropanol wash and air dry. Impressive structures can be found in the literature; however, there are a number of significant problems with SU-8 removal. Performance is very sensitive to the process parameters. All parameters, including resist deposition, pre-exposure bake, exposure dose, post exposure bake, and final development, impact the performance. As an example, a UCD fabricated TWT circuit using an SU-8 mold is shown in Fig. 5c.

KMPR™ LIGA

KMPR is a negative tone photoresist. This chemically enhanced carboxylated epoxy resin consists of a photo-generated acid which requires a postexposure bake to fully activate the solidification process called cross-linking. For electroforming applications, this can withstand acidic solutions at relatively high temperatures compared to room temperature. UC Davis has worked on specially fabricated 4 in. Oxygen Free High Conductivity OFHC Cu wafers with highly flat (±0.5 μm) and polished surfaces. The process consists of spin coating, soft bake, UV exposure, postexposure bake, electroplating, lapping/polishing, and mold removal processing. The UC Davis team has successfully characterized the entire process to precisely fabricate 220 GHz Traveling-Wave Tube Amplifier TWTA copper circuits ~400-μm thick in a single layer process. Figure 5d shows the fully metalized half of the TWT circuit using a KMPR mold.

Si DRIE

Deep reactive ion etching (DRIE) is a process whereby high aspect ratio structures are etched into silicon [16]. These structures can be used to generate molds or serve as a mold themselves for

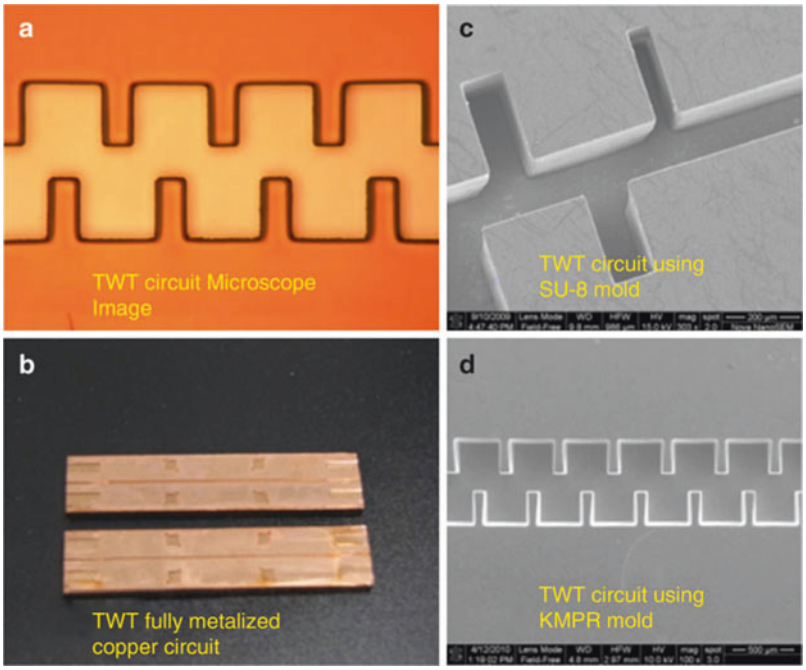


MEMS Vacuum Electronics, Fig. 4 LIGA process flow chart

generating metallic structures. Silicon structures ranging from two to hundreds of micrometers can be created. Fluorine-based chemistry is typically used because of its high etch rate (10 mm/min) and selectivity (100:1). The silicon is etched with SF_6. Anisotropy is improved by ion impact assistance coupled with sidewall passivation using flourocarbon polymer deposition and/or O_2 plasma oxidation and redeposition at the silicon sidewall surface. The flourocarbon polymer consists of a chain of CF_2 molecules similar to Teflon with a film thickness of approximately 10–50 nm. It is derived from a 4–8 s application of C_4F_8 applied as a plasma source gas. The introduction of oxygen improves anisotropy by acting as a gettering agent for carbon. The following etch cycle consists of 4–12 s of SF_6, applied as a plasma source gas. The ion assisted SF_6 removes the polymer passivation on horizontal surfaces prior to etching the underlying silicon, leaving the polymer on the sidewalls relatively intact.

Nano-Computer Numerical Controlled (CNC) Milling

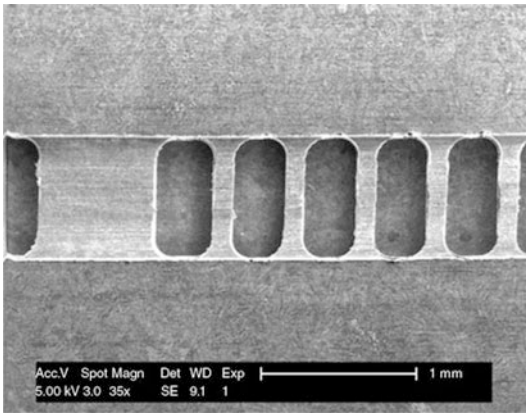
The common wisdom has been that direct machining cannot meet the sensitive tolerances of high-frequency vacuum electronic devices (dimensional tolerance <3 μm, surface finish



MEMS Vacuum Electronics, Fig. 5 UCD LIGA fabricated 220 GHz TWTA circuit in a single layer ~400 μm process (**a**) Microscope image (**b**) Fully metalized copper circuit (**c**) TWT circuit using SU-8 mold (**d**) TWT metalized circuit using KMPR mold

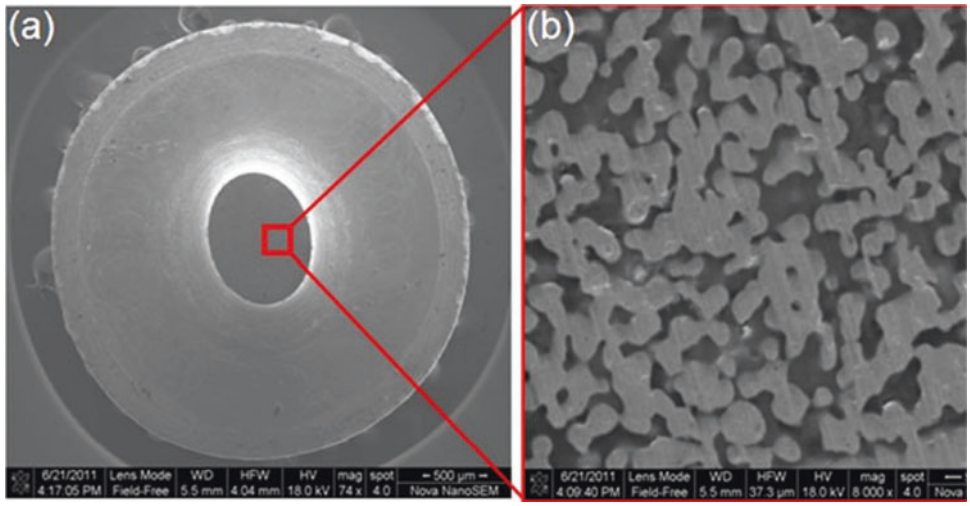
<100 nm R_a). However, since the invention of the *Mori Seiki* NN1000 "Nano-Mill" developed by Digital technology Laboratory in Davis, the UC Davis team has successfully fabricated broadband (~60 GHz), 0.22 THz traveling wave tube components using Nano CNC (computer numerical control) milling as their primary approach (Fig. 6) [17, 18]. The NN1000 Nano-Mill features air-bearing guideways, vibration isolation technology, and laser-scale position sensors that have a resolution of 34 pm, and work in concert with linear drive magnetic motors to allows position control resolution of 1 nm. The machine can use either an air spindle capable of 55,000 RPM or a fixed mandrel for linear scribing. The NN1000 is a smooth and steady platform designed for ultraprecision work with exceptionally fine surface finishes. The capabilities of this machine have yet to be fully incorporated into miniaturizing the vacuum electronics industry.

Also, the UC Davis team has used Nano CNC Milling for machining of tungsten high current density cathodes [15]. Since the particle size of the tungsten nano-composite cathodes is usually less than one micron, it is very difficult to achieve good quality emission surface using conventional cathode machining techniques (the surface becomes smudged and the pores are closed). UC Davis has shown that the NN1000 machine is also an excellent choice for creating an ideal cathode emission surface (Fig. 7).



MEMS Vacuum Electronics, Fig. 6 0.22 THz circuit machined using nano CNC milling

M



MEMS Vacuum Electronics, Fig. 7 (**a**) Scanning electron microscope image of a fully machined cathode for a 0.22 THz electron gun, (**b**) scanning electron microscope image of the emission surface of the nano-composite tungsten cathode

Key Research Findings

Two-Step X-ray LIGA at Seoul National University

Shin et al. [19] at Seoul National University were able to build the first working LIGA fabricated vacuum electron device operating at 35 GHz. To further their achievement, the group was able to fabricate a circuit for a device that would operate at 100 GHz using two-step X-ray LIGA creating three-dimensional structures with two-dimensional techniques.

Northrop Grumman 650 GHz Folded Waveguide TWT

Northrop Grumman [20] fabricated and demonstrated a Folded Waveguide TWT amplifier with regenerative feedback at 650 GHz. The device maximum output power was 16 mW at a duty factor of 1 % in the frequency range 600–675 GHz. This is significantly higher than the power capability of existing sources operating at these frequencies.

Nano-Klystron at Jet Propulsion Laboratory, CA, USA

Siegel et al. [21] at Jet Propulsion Laboratory have reported on the designing and fabrication of Nano-Klystrons at 640 and 1,200 GHz. They were able to manufacture a cavity for a 640 GHz Nano-Klystron using precision metal machining on copper and they are working on developing a manufacturing method for 1,200 GHz cavity using silicon DRIE. To achieve the current densities required for Nano-Klystrons, they are planning to use cold cathodes that are capable of generating 1,000 A/cm^2. These cathodes are made of carbon nanotubes and they are well suited for low power high current density applications.

Diamond-Based Submillimeter Backward Wave Oscillator

Dayton et al. [22] at the GENVAC Aerospace Corporation, OH, were able to produce a 300 GHz circuit by forming chemical vapor deposited diamond around lithographically constructed silicon mold. They believe that similar techniques can be easily adapted to producing 600 GHz circuits. They are applying solid-state fabrication techniques to vacuum electronics: The gun is manufactured as an integral part of the slow wave structure. In addition, they were able to apply precision fabrication and assembly techniques used in fabrication of liquid crystals to bring vacuum electronics even closer to being truly nanoscale.

The Klystrino at the Stanford Linear Accelerator Center (SLAC)

Stanford Linear Accelerator Center (SLAC) reported on the development of a "klystrino" [23] – high-power RF source at W-band intended for tabletop accelerator, radar, and communications applications. An X-ray LIGA process was used to manufacture the circuit with deep features (<3 mm) and surface finish of about 50 nm. The klystrino was designed to operate at 110 kV, 2.4 A and produce 10 kW of peak power with 1 kW of average power at 1 % duty.

220 GHz TWT by UC Davis, CPI, and Teledyne

At UC Davis, a novel 0.22 THz Traveling Wave Tube Amplifier design was conceived [24] with potential application for high data rate communications and imaging. It consists of a staggered double-vane grating array operating in the TE mode. The device performance using the MAGIC PIC code showed a power bandwidth product of >1,000 W GHz with 3–5.5 % intrinsic electronic efficiency, and >25 % bandwidth at a center frequency of 220 GHz. The shortest dimension of this 220 GHz TWTA double-vane micrometallic structure is 75-μm wide and 385-μm deep, corresponding to an aspect ratio of ~4:1. The MEMS fabricated circuits are shown in Fig. 5. The RF measurements conducted at UCD demonstrated excellent RF transmission for KMPR LIGA fabricated circuits [25] and the Nano-CNC fabricated circuits [17, 26] with S_{21} ~ 5 dB in passband (~200–265 GHz) and S_{11} ~ −10 dB.

Cross-References

- Chemical Vapor Deposition (CVD)
- CMOS MEMS Fabrication Technologies
- Dry Etching Processes
- DUV Photolithography and Materials
- Electron-Beam-Induced Deposition
- Photolithography
- Scanning Electron Microscopy
- SU-8 Photoresist
- Ultraprecision Machining (UPM)

References

1. Gittins, J.F.: Power Traveling-Wave Tubes. American Elsevier Publication, New York (1965)
2. Gilmour, A.S.: Principles of Traveling Wave Tubes. Artech House, Boston (1994)
3. Luhmann, N.C., Barker, R.J., Booske, J.H., Nusinovich, G.S.: Modern Microwave and Millimeter-Wave Power Electronics. IEEE Press/ Wiley, Hoboken (2005)
4. Siegel, P.H.: Terahertz technology. IEEE Trans. Microw. Theory Tech. **50**, 910 (2002)
5. Booske, J.H.: Plasma physics and related challenges of millimeter-wave-to-terahertz and high power microwave generation. Phys. Plasmas **15**, 055502–055516 (2008)
6. Shin, Y.-M., Zhao, J., Barnett, L.R., Luhmann Jr., N. C.: Investigation of terahertz sheet beam traveling wave tube amplifier with nanocomposite cathode. Phys. Plasmas **17**, 123105 (2010)
7. Wang, S.: Scandate cathode for TWT. Appl. Surf. Sci. **251**, 114–119 (2005)
8. Jinfeng, Z., Na, L., Ji, L., Barnett, L.R., Banducci, M., Gamzina, D., Munir, Z.A., Luhmann, N.C.: High current density and long-life nanocomposite scandate dispenser cathode fabrication. IEEE Trans. Electron Dev. **58**, 1221–1228 (2011)
9. Ives, R.L., Collins, G., Read, M., Miram, G., Marsden, D.: Electron guns for Terahertz vacuum electron sources (Invited Paper). Int. J. THz Sci. Technol. **4**(4), 230–239 (2011). Special Issue on Microfabricated THz Vacuum Electron Devices
10. Li, J., Yan, S., Shao, W., Chen, Q., Zhu, M.: Investigation and application of impregnated scandata cathodes. Appl. Surf. Sci. **215**, 49–53 (2003)
11. Li, J., Wang, H., Gao, Y., Zhang, K., Chen, Q., Yan, T., Liao, F., Wang, J., Wang, Y., Liu, W. Recent developments. In: Scandia-Doped Dispenser Cathodes. International Vacuum Electronics Conference (IVEC), Monterey, pp. 51–52 (2006)
12. Li, J., Wang, H., Yu, Z., Gao, Y., Chen, Q., Yan, T., Liao, F., Wang, J., Wang, Y., Liu, W.: Emission mechanism of high current density thermionic cathodes. In: International Vacuum Electronics Conference (IVEC), Monterey, pp. 1–2 (2007)
13. Yiman, W., Jinshu, W., Wei, L., Ke, Z., Ji, L.: Development of high current-density cathodes with Scandia-doped tungsten powders. IEEE Trans. Electron Devices **54**, 1061–1070 (2007)
14. Liu, W., Wang, Y., Wang, J., Wang, Y.C., Vancil, B.: Emission characteristics of nanosized scandia-doped dispenser cathodes. IEEE Trans. Electron Dev. **54**, 1241–1246 (2011)
15. Zhao J., Li N., Li J., Barnett, L.R., Banducci, M., Gamzina, D., Luhmann, N.C.: Scandate dispenser cathode for 220 GHz 50 W sheet beam travelling wave tube amplifier. In: Vacuum Electronics Conference (IVEC), 2011 I.E. International, Bangalore, pp. 41–42 (2011)

16. Shin, Y.-M., Barnett, L.R., Gamzina, D., Luhmann Jr., N.C., Field, M., Borwick, R.: Terahertz vacuum electronic circuits fabricated by UV lithographic molding and deep reactive ion etching. Appl. Phys. Lett. **95**, 181505-3 (2009)
17. Barchfeld, R.A., Gamzina, D., Shin, Y.-M., Barnett, L. R., Luhmann, N.C. Microfabrication of Terahertz Vacuum Electronics Devices. In: Kim, J., Alldritt, C. (Mori Seiki) (eds.) American Society for Precision Engineers Conference, University of California, Davis (2011)
18. Gamzina, D., Barchfeld, R., Barnett, L.R., Luhmann, N.C., Shin, Y.-M.: Nano CNC milling technology for terahertz vacuum electronic devices. In: Vacuum Electronics Conference (IVEC), 2011 I.E. International, Bangalore, pp. 345–346 (2011)
19. Shin, Y.-M., So, J.-K., Han, S.-T., Jang, K.-H., Park, G.-S., Kim, J.-H., Chang, S.-S.: Microfabrication of millimeter wave vacuum electron devices by two-step deep-etch x-ray lithography. Appl. Phys. Lett. **88**, 091916-3 (2006)
20. Tucek, J., Kreischer, K., Gallagher, D., Vogel, R., Mihailovich, R.: Development and operation of a 650 GHz folded waveguide source. In: Vacuum Electronics Conference, 2007. IVEC '07. IEEE International, Kitakyushu, pp. 1–2 (2007)
21. Siegel, P.H., Fung, A., Manohara, H., Xu, J., Chang, B.: Nano Klystron: A monolithic tube approach to THz power generation. In: 12th International Symposium on Space THz Technology, San Diego, pp. 81–90 (2001)
22. Dayton, J.A., Mearini, G.T., Kory, C.L.: Diamond-based sub millimeter backward wave oscillator. In: Vacuum Electronics Conference, 2004. IVEC 2004. Fifth IEEE International, Monterey, California, pp. 71–72 (2004)
23. Scheitrum, G., Arfin, B., James, B.G., Borchard, P., Song, L., Cheng, Y., Caryotakis, G., Haase, A., Stockwell, B., Luhmann, N., Shew, B.Y.: The klystrino: a high power W-band amplifier. In: International Vacuum Electronics Conference, 2000: Abstracts, Monterey, California, p. 2 (2000)
24. Shin, Y.-M., Barnett, L.R., Luhmann, N.C.: Phase-shifted traveling-wave-tube circuit for ultrawideband high-power submillimeter-wave generation. IEEE Trans. Electron Dev. **56**, 706–712 (2009)
25. Baig, A., Gamzina, D., Johnson, M., Domier, C.W., Spear, A., Barnett, L.R., Luhmann, N.C., Shin Y.-M.: Experimental characterization of LIGA fabricated 0.22 THz TWT circuits. In: Vacuum Electronics Conference (IVEC), 2011 I.E. International, Bangalore, pp. 275–276 (2011)
26. Baig, A., Shin, Y.-M., Barnett, L.R., Gamzina, D., Barchfeld, R., Domier, C.W., Wang, J., Luhmann Jr., N.C.: Design, fabrication and RF testing of near-THz sheet beam TWTA (Invited). Terahertz Sci Technol. **4**(4), 181–207 (2011)

MEMS/IC Integrated Resonator Circuits

▶ CMOS-MEMS Resonators

MEMS-Based Drug Delivery Devices

▶ Stimuli-Responsive Drug Delivery Microchips

Meniscus Adhesion

▶ Disjoining Pressure and Capillary Adhesion

Mesoscopic Modeling

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Mesoscopic Simulation

▶ Dissipative Particle Dynamics, Overview

Mesoscopic Simulations

▶ Coarse-Grained and Hybrid Simulations of Nanostructures

Metal Alloy Nanoparticles

▶ Nanoalloy Simulation

Metal Nanoparticles from First Principles

Nicephore Bonnet
National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

Synonyms

Ab initio modeling and simulation of metal nanoparticles; Atomistic and multiscale modeling of metal nanoparticles

Definition

Metal nanoparticles are particles composed of metal elements (most frequently transition metals such as Ag, Au, Co, Cu, Fe, Ni, Pd, Pt, Re, Ru) with a size below ~200 nm and an aspect ratio close to 1 (as opposed to nanowires and nanosheets). Particles smaller than 3 nm (~500–1000 atoms) are also called *clusters*.

A first-principles approach aims at computing the properties of a system by directly solving the quantum-mechanical Schrodinger equation for atomic cores and electrons without the input of empirical parameters.

Introduction

Metal nanoparticles have been used since ancient times as famously exemplified by the fine Lycurgus Cup in the British Museum. This fourth-century Roman glass cup is green in reflected light and ruby red in transmitted light owing to the presence of gold colloids embedded in the glass, where the size of the colloids does not exceed 70 nm [1]. Interestingly, the manufacture of gold-ruby glass was subsequently lost and rediscovered in Europe only in the seventeenth century.

Nowadays, metal nanoparticles have become critically important and promising components in a large variety of fields such as biological imaging and sensing, drug delivery, catalysis, thin film technology, optoelectronics, optics, and magnetic recording. The use of nanoparticles may be primarily motivated by geometric constraints, as is the case for drug delivery to the central nervous system through the brain-blood barrier. Similarly, metal nanoparticles have been used in catalyst arrays as a way to maximize the active surface area for a given amount of material, and thus increase the efficiency or reduce the cost of the device. In most cases, however, downsizing the material to the nanoscopic scale will also modify or generate new properties that are either detrimental to the intended application or on the contrary specifically targeted. For instance, the surface activity of Pt nanoparticles used in proton exchange membrane fuel cells to catalyze the oxygen reduction reaction (ORR) has been shown to deteriorate with reducing size. Therefore the optimal diameter of the Pt nanoparticles is practically around 3–5 nm before this detrimental effect starts outweighing the benefit of size reduction on the specific activity (activity per unit volume). A counterexample is offered by gold. While bulk gold is noble and does not easily adsorb molecules, gold nanoparticles have been found to be good catalysts for some applications such as CO oxidation even at room temperature.

Thus, while a vast amount of experimental work is devoted to finding easy, robust, and cost-effective synthetic routes for particles of targeted composition, shape, and size, it is equally important to develop theoretical tools that can predict their properties at the nanoscale, and first-principles techniques have a long record in this field. This entry first explains why the simulation of metal nanoparticles requires the use of first-principles techniques and particularly density functional theory (DFT). Then, it describes a series of well-established DFT-based methodologies to obtain some of the main properties of metal

Notes Readers may view, browse, and/or download material for temporary copying purposes only, provided these uses are for noncommercial personal purposes. Except as provided by law, this material may not be further reproduced, distributed, transmitted, modified, adapted, performed, displayed, published, or sold in whole or part, without prior written permission from the American Physical Society.

nanoparticles, namely, chemical and geometric stability, growth, segregation effects in the case of alloy particles, electrochemical stability, and catalytic properties.

First-Principles Calculations

The metallic bond is remarkable, first because it tends to favor high coordination numbers, up to 12 in close-packed atomic arrangements, second because it is moderately directional, which means that the bond energy has a small dependence on the bond angular variables. This has important implications when dealing, for instance, with small metal clusters (<50 atoms). First, the metallic bond in a small cluster is not saturated and thus not fully developed, and the empirical embedded-atom method (EAM), which is often a method of choice for predicting the energy of metallic systems, is then not applicable. Second, a small cluster with a given composition and number of atoms may adopt a large number of metastable configurations that are very close in energy. It is therefore necessary to be able to accurately compute the energy of each configuration at a small computational cost. The accurate treatment of atomistic systems is specifically the object of *first-principles methods*, whereby the energy of the system is determined by directly solving the quantum-mechanical Schrödinger equation without inputting empirical parameters. Given that the nuclei are much heavier than the electrons, it is in most cases reasonable to treat the former as classical particles moving in the field generated by the electron cloud, the latter being represented by the full electronic wave function $\Psi_{\mathrm{el}}(r_1, \ldots, r_{N_e})$ as a function of the coordinates r_i of the N_e electrons in the system. In this so-called Born-Oppenheimer approximation, one should then solve for Ψ_{el} in the Schrödinger equation written for any fixed position of the nuclei as [2]

$$\left(-\sum_{i=1}^{N_e}\frac{\hbar^2}{2m_e}\nabla_i^2 + V_{ext} + \sum_{i\neq j}^{N_e}\frac{q_e^2}{4\pi\epsilon_0 r_{ij}}\right)\Psi_{\mathrm{el}} = E_0\Psi_{\mathrm{el}}. \quad (1)$$

Here, the first term stands for the kinetic energy, where $\hbar$ is the reduced Planck constant and m_e is the electron mass. The second term stands for the external energy, where V_{ext} is the potential created by the nuclei and possibly by any other external source and acting on the electrons. The third term stands for the electron–electron interaction energy, where q_e is the electron absolute charge, ϵ_0 is the permittivity of free space, and $r_{ij} = |r_j - r_i|$. The total ground-state electronic energy of the system E_0 is obtained as the lowest eigen value of Eq. 1. However, in practice, the multibody character of Ψ_{el}, whereby the probability distributions of any two electrons i and j are correlated in this wave function, makes this general approach intractable even numerically, and thus calls for an approximate solution.

Hartree-Fock Approximation

In the Hartree-Fock approximation [2, 3], the electronic wave function is written as the Slater determinant of $N_e/2$ one-body wave functions, $\Psi_{\mathrm{el}} = |\Psi_1(r_1)\ldots\Psi_{N_e/2}(r_{N_e/2})|$, which satisfies the Pauli exclusion principle for a system of fermions (here, for simplicity, we assume that N_e is an even number and that the system is not spin polarized, in which case each wave function is occupied by two electrons of opposite spins). This allows separating the coordinates in Eq. 1 and to obtain the $\Psi_i(r_i)$ as the solutions of a system of $N_e/2$ coupled eigen equations, which can be solved through a self-consistent numerical scheme. This approach is also interesting because the Ψ_i can be chemically interpreted as orbitals occupied by the electrons, and an approximation of the total energy E_0 can be obtained from the orbital energies ϵ_i. One drawback of this approach, however, is that the Slater determinant form of Ψ_{el} is unable to capture all of the electronic correlation, notably leading to a poor description of the Coulomb hole at short range. As a consequence, it is known that for a metallic system, the bandwidth and the density of electronic states at the Fermi level are largely exaggerated at the Hartree-Fock level [2–4].

Density Functional Theory (DFT)

Metallic systems are better described in the framework of density functional theory [2, 3], where the ground-state electronic density $n(r)$ is taken as the primary unknown and determines the total energy of the system E_0 and the full electronic correlation. In the Kohn-Sham approach to DFT (KS-DFT), the electronic density can be in principle obtained exactly by solving the Kohn-Sham eigen equation

$$\mathcal{H}_{\mathrm{KS}}[n(r)]\Psi_i = \left(-\frac{\hbar^2}{2m_e}\nabla^2 + V_{ext} + V_{xc}[n(r)]\right)\Psi_i = \epsilon_i\Psi_i. \tag{2}$$

In the Kohn-Sham Hamiltonian $\mathcal{H}_{\mathrm{KS}}$, $n(r)$ determines the exchange-correlation potential V_{xc} and is itself determined by the Kohn-Sham orbitals Ψ_i, implying that Eq. 2 has to be solved self-consistently. Because the core electrons are usually not involved into bonding with surrounding atoms, the core electronic density may be taken within a good approximation as the sum of the core densities of isolated atoms stored in atomic pseudopotentials. The unknown is then only the valence electronic density, which drastically reduces the computational cost of the simulation. For heavy elements such as gold, the relativistic character of the core electrons modifies their interactions with the valence electrons and thus the chemical behavior of the latter, in which case a relativistic pseudopotential has to be used for the accuracy of the calculation. Also, in practice, the exact dependence of V_{xc} on $n(r)$ is unknown, and approximate functions have to be used. Most commonly, V_{xc} is expressed as a local function of $n(r)$ in the local density approximation (LDA), or as a local function of $n(r)$ and its gradient in the generalized gradient approximation (GGA). Moreover, some DFT functionals have also been developed to include the long-range van der Waals interactions, which are especially important for metal clusters.

Although the computational cost of a naively implemented KS-DFT scheme scales $\mathcal{O}(N_e^3)$, a number of approaches reduce this scaling to $\mathcal{O}(N_e^2)$ or even $\mathcal{O}(N_e)$, allowing for the treatment of systems that contain hundreds or thousands of atoms. Thus, in general DFT represents the best compromise between accuracy and computational effort for metal clusters. It is however important to note that contrary to the Hartree-Fock approach, the KS-DFT formalism is in principle affected by electronic self-interaction, whereby an electron interacts with the electronic density of its own orbital. While this has a minor impact on metallic systems where the wave functions are largely delocalized, this may be the source of substantial discrepancies for insulators and systems with a sizable HOMO-LUMO gap such as oxides and organic components. The energetics of these systems may be improved by introducing some of the Hartree-Fock correlation in the KS-DFT potential. Alternatively, in the DFT + U approach, an additional potential characterized by the U term may be selectively applied to some elements, e.g., oxygen atoms, to penalize the electronic overdelocalization induced by the self-interaction. Such hybrid approaches have been shown, for instance, to better describe the magnetic properties of transition metal complexes [2, 4].

Chemical Stability of Metal Clusters

Small metal clusters have interesting properties related to their persistent molecular character. Particularly, they exhibit distinct electronic shells, and thus tend to favor atomic arrangements and numbers of atoms (kwown as *magic numbers*) ensuring electronic shell closure [4]. An example is provided by the synthesis of monodisperse, ultrastable silver clusters protected by a thiolate layer [5]. X-ray spectroscopy has revealed a core structure consisting of a Ag_{12} hollow icosahedron within a Ag_{20} dodecahedron enclosed in a sulfur rhombicuboctahedron, six faces of which are capped with Ag_2S units, resulting in an overall octahedral particle with 44 Ag atoms. By computing the electronic structure of this system with DFT, one shows that the delocalized electrons not involved in bonding with thiolates occupy

spherical orbitals around the Ag cluster playing the role of a superatom. The number of these electrons is 18, which corresponds to the Aufbau closed-shell filling $1S^2 \mid 1P^6 \mid 1D^{10}$ and thus explains the outstanding stability of this structure. This is further supported by the relatively large HOMO-LUMO gap (0.78 eV) obtained for this structure.

Structure Optimization of Metal Clusters (<500 atoms)

The purpose of structure optimization is to minimize the energy of a particle with a given composition and number of atoms. This can be achieved, for instance, by the basin-hopping (BH) algorithm, consisting in a series of random perturbations of the atom coordinates followed by local relaxations. The new coordinates are accepted or rejected based on the value of the relaxed energy, for instance, by using the Metropolis criterion of a standard Monte-Carlo simulation. The efficiency of this algorithm is sensitive to the coordinate perturbation step, which may consist in a simple kick on the atoms or in the exchange of two atoms of distinct species. The selection of new coordinates may also be enhanced or oriented through some pattern recognition functionalities, e.g., using a nearest-neighbor analysis. The stable patterns are either entered in advance or learned during the simulation as in genetic algorithms. The approach to compute and locally minimize the energy of intermediate configurations depends on the size of the system.

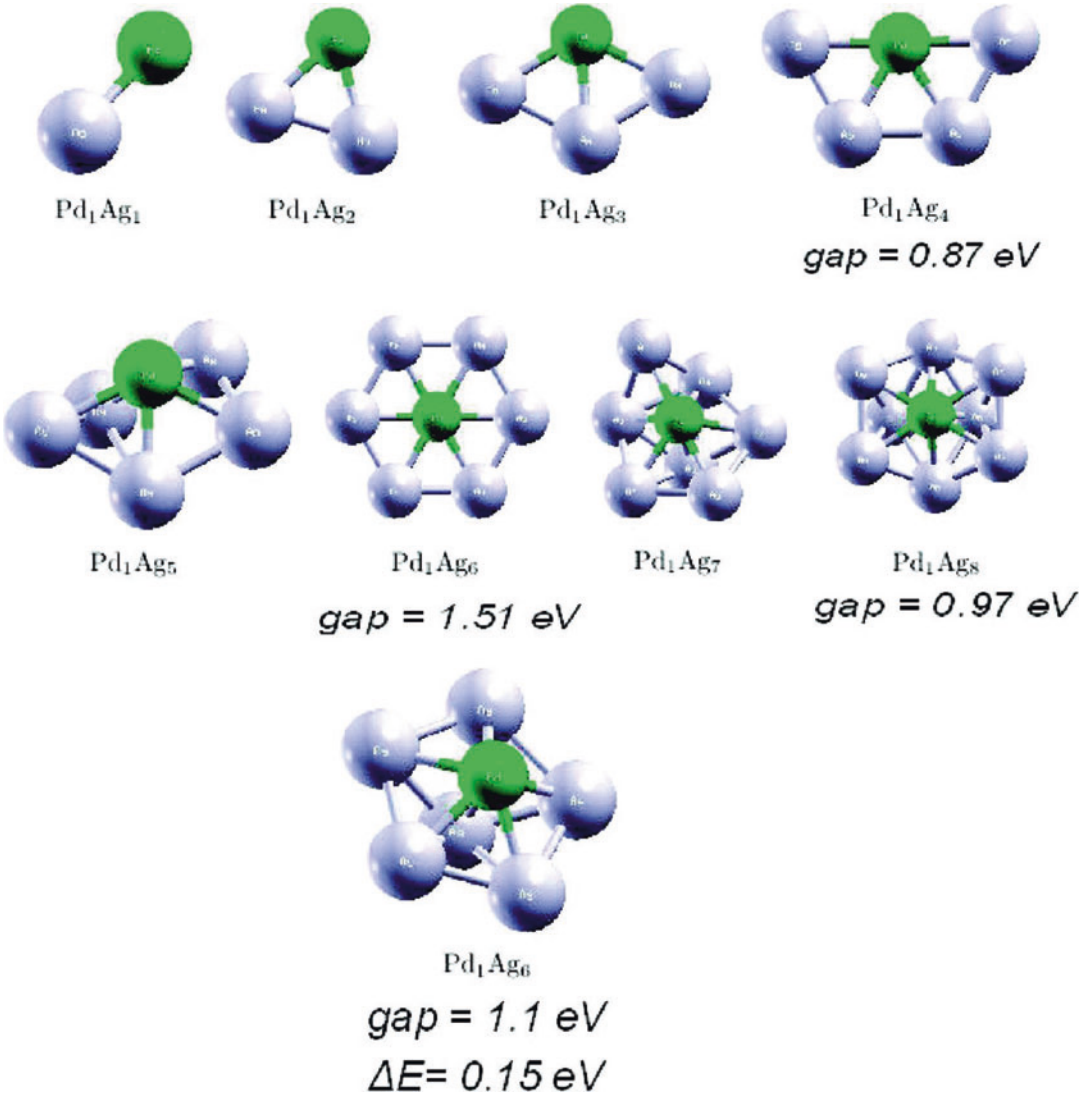
For small clusters (<50 atoms), it is necessary, and computationally affordable, to apply first-principles techniques (such as DFT) to accurately describe the specific chemistry of the system. The resulting BH-DFT algorithm has been used, for instance, to predict the stable structure of $PdAg_M$ nanoclusters as shown in Fig. 1. While in general for $M > 4$ the cluster favors a 3D structure to maximize the number of bonds, the PdAg6 exceptionally retains a planar geometry to achieve a closed-shell electronic structure in the 2D jellium model [4].

For large clusters (50–500 atoms), the computational cost implied by the large number of configurations starts precluding an all-DFT treatment. Fortunately, as the system size increases the electronic shells start to merge to build up the metallic character of the system. The vanishing of shell closure effects justifies developing and using empirical potentials (EP) to calculate the energies in the EP-BH algorithm. However, in this size range it is recommended to check a posteriori the energy ordering of the most stable configurations by DFT and possibly to refine the EP in case of a discrepancy.

DFT-Based Modeling of the Shape Stability of Metal Nanoparticles (>500 Atoms)

Beyond a few hundreds of atoms, it becomes too costly to perform first-principles calculations of the entire particle, implying the need for some modeling and multiscale approaches. Regarding the particle shape, one can use the fact that in most cases highly symmetrical shapes exposing highly packed facets such as (111) and (100) are energetically favorable. Typically, this can lead to multiply twinned particles such as icosahedra and decahedra or monocrystalline particles such as cubes, octahedra, and cuboctahedra. Thus, in the case of monoelemental particles, a general expression for the particle free energy can be obtained within this limited set of geometries as a function of the temperature T and external pressure P_{ext} through a simple thermodynamic model including the contributions of the bulk, surfaces, and twin boundaries [6]:

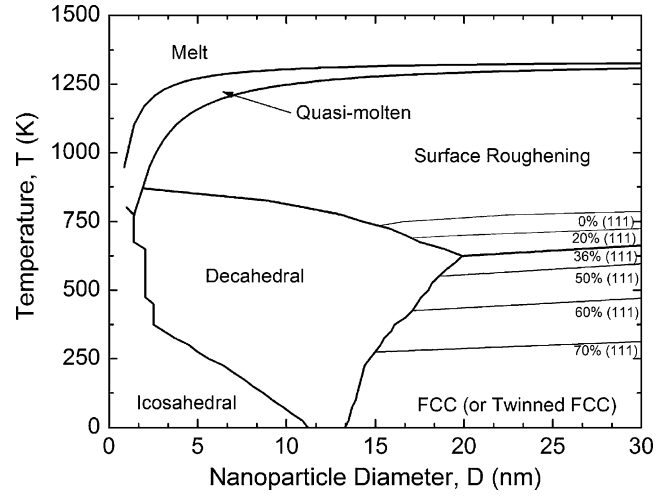
$$G(T,P_{\text{ext}}) = G_{\text{bulk}}(T) + v\left(1 - \frac{2\sum_i f_i\sigma_i(T)}{B_0 R} + \frac{P_{\text{ext}}}{B_0}\right) \times \left[q\sum_i f_i\gamma_i(T) + n\sum_\theta \left(a_\theta v_\theta(T) + \sum_\phi l_{\theta\phi}\eta_\phi(T)\right)\right], \quad (3)$$



Metal Nanoparticles from First Principles, Fig. 1 Optimal structures of $PdAg_M$ clusters as obtained from BH-DFT (Reprinted with permission from Ref. [17]. Copyright 2007 American Chemical Society)

where $G_{\text{bulk}}(T)$ is the bulk energy at normal pressure; v, the molar volume; q, the surface-to-volume ratio; f_i, the surface fraction of facet i; $\gamma_i(T)$ the surface energy of facet i; n, the density of planar defects; $v_\theta(T)$, the energy of the twin plane of area a_θ in orientation θ; and $\eta_\phi(T)$, the re-entrant line tension where the twin plane intersects with the free surface in the orientation ϕ with length $l_{\theta\phi}$. We note that by definition, $\sum f_i = 1$; This model also includes volume change effects induced by the surface stresses σ_i and external pressure P_{ext} through the Young-Laplace law, requiring the bulk modulus B_0 and the radius of the particle averaged to a sphere R. The quantities B_0, $\gamma_i(T)$, $\sigma_i(T)$, $v_\theta(T)$, and $\eta_\phi(T)$ are input parameters that need to be calculated in advance, and the optimal particle shape is then obtained by minimizing $G(T, P_{\text{ext}})$ over the f_i, a_θ, and $l_{\theta\phi}$. The surface energy of a facet γ_i may be determined at zero temperature as the quantity satisfying the relation

$$E^N_{\text{slab}} = 2\gamma_i + NE_{\text{bulk}}, \tag{4}$$



Metal Nanoparticles from First Principles, Fig. 2 Phase diagram of nanogold as obtained from relativistic DFT calculations (Reprinted with permission from Ref. [18]. Copyright 2009 American Chemical Society)

where E^N_{slab} and E_{bulk} are the total energy of an N-atom slab in the direction of the facet and the total energy of the bulk per atom, respectively, and are computed with DFT. The factor 2 accounts for the fact that the slab contains two facets. The temperature dependence of $\gamma_i(T)$ can then be extrapolated from $\gamma_i(0)$ through a semiempirical law [6]. The same kind of DFT approach may be used to determine B_0, $\nu_\theta(T)$, and $\eta_\phi(T)$, and the surface stress may be obtained through the common approximation $\sigma_i(T) \sim \gamma_i(T)$, With this approach, the phase diagram of nanogold has been obtained as a function of size and temperature based on relativistic DFT calculations, as shown in Fig. 2. At low temperature, a striking feature of this diagram is the change of the stable structure from the icosahedron, to the decahedron (>9 nm), and finally to the truncated octahedron for sizes larger than 12 nm. Below 9 nm (~20,000 atoms) the energetics are dominated by surface effects, and the icosahedron, with its 20 (111) facets, is the most stable structure. By contrast, above 12 nm the energetics are dominated by the bulk, which penalizes structures with internal twinning. Thus, the monocrystalline truncated octahedron becomes the most stable structure, in spite of the energy penalty associated with its six (100) facets [7].

The treatment of the surface tension in this approach can be refined by including the dependence of the facet energies on the lattice parameter a. The free energy of a monocrystalline structure is then written, with the temperature and pressure dependence being kept implicit,

$$G(a) = G_{\text{bulk}}(a) + vq\sum_i f_i\gamma_i(a), \qquad (5)$$

where $G_{\text{bulk}}(a)$ and $\gamma_i(a)$ have to be computed for different values of a. The particle shape is then obtained by minimizing $G(a)$ over the f_i and a. By comparing the predictions of this approach with explicit DFT calculations for small Pt particles, it has been shown that Eq. 5 is adequate to describe the surface tension effects on the relative stabilities of monocrystalline structures below ~3 nm [8], For larger particles, this differential effect becomes negligible, and it is then valid to use surface energies at the bulk lattice parameter $\gamma_i(a_{\text{bulk}})$ as in Eq. 3. As the particle size increases further, dilation effects too become progressively negligible, and thus for monocrystalline particles Eq. 3 simplifies into

$$G = G_{\text{bulk}} + vq\sum_i f_i\gamma_i. \qquad (6)$$

Wulff Construction (in the Presence of Adsorbates)

In the regime of validity of Eq. 6, the Wulff theorem states that the energy-minimizing polyhedron is such that the distance of each facet to the center of the nanoparticle is proportional to its surface energy. This approach can, for instance,

be used to predict the effect of the chemical environment on the particle shape stability. First, the facet energies in the presence of adsorbates $\gamma_i(\theta)$ are computed using the following formula:

$$\gamma_i(\theta) = \gamma_i(0) + \Delta g_{\text{ads}}(\theta). \tag{7}$$

Here θ is the adsorbate coverage, which is a vector $\theta = (\theta_1, \ldots, \theta_j, \ldots)$ if molecules of different species j are adsorbed; $\gamma_i(0)$ is the surface energy of the bare surface as given by Eq. 4; and $\Delta g_{\text{ads}}(\theta)$ is the free energy of adsorption per unit surface area. This can be computed again with a slab approach,

$$\Delta g_{\text{ads}}(\theta) = \frac{1}{A}\left(G_{\text{slab}}(\theta) - G_{\text{slab}}(0) - \sum_j N_j g_j\right), \tag{8}$$

where $G_{\text{slab}}(\theta)$ is the free energy of the slab + adsorbates; $G_{\text{slab}}(0)$ is the free energy of the bare slab; g_j is the free energy of a molecule of species j in solution or in the gas phase; and N_j is the number of molecules of species j adsorbed on the slab of surface A, which implies $\theta_j = \frac{N_j}{A}$;. The electronic energies and zero-point energies (ZPEs) may be determined by DFT, while entropy contributions are often taken from experiment or estimated through simple statistical models of adsorption [9], The coverage at equilibrium is obtained by minimizing $\Delta g_{\text{ads}}(\theta)$. This approach has been applied, for instance, to study the effect of oleic acid (OA) adsorption on the shape stability of PbSe particles [10]. OA tends to bind more strongly to the (111) surface, such that at low OA concentrations Δg_{ads} is more negative on this surface and the octahedral particle shape is stabilized. However, as the OA concentration increases, the (100) facets become also covered and stabilized by OA, which can ultimately lead to the cubic shape as shown in Fig. 3.

Growth Morphology

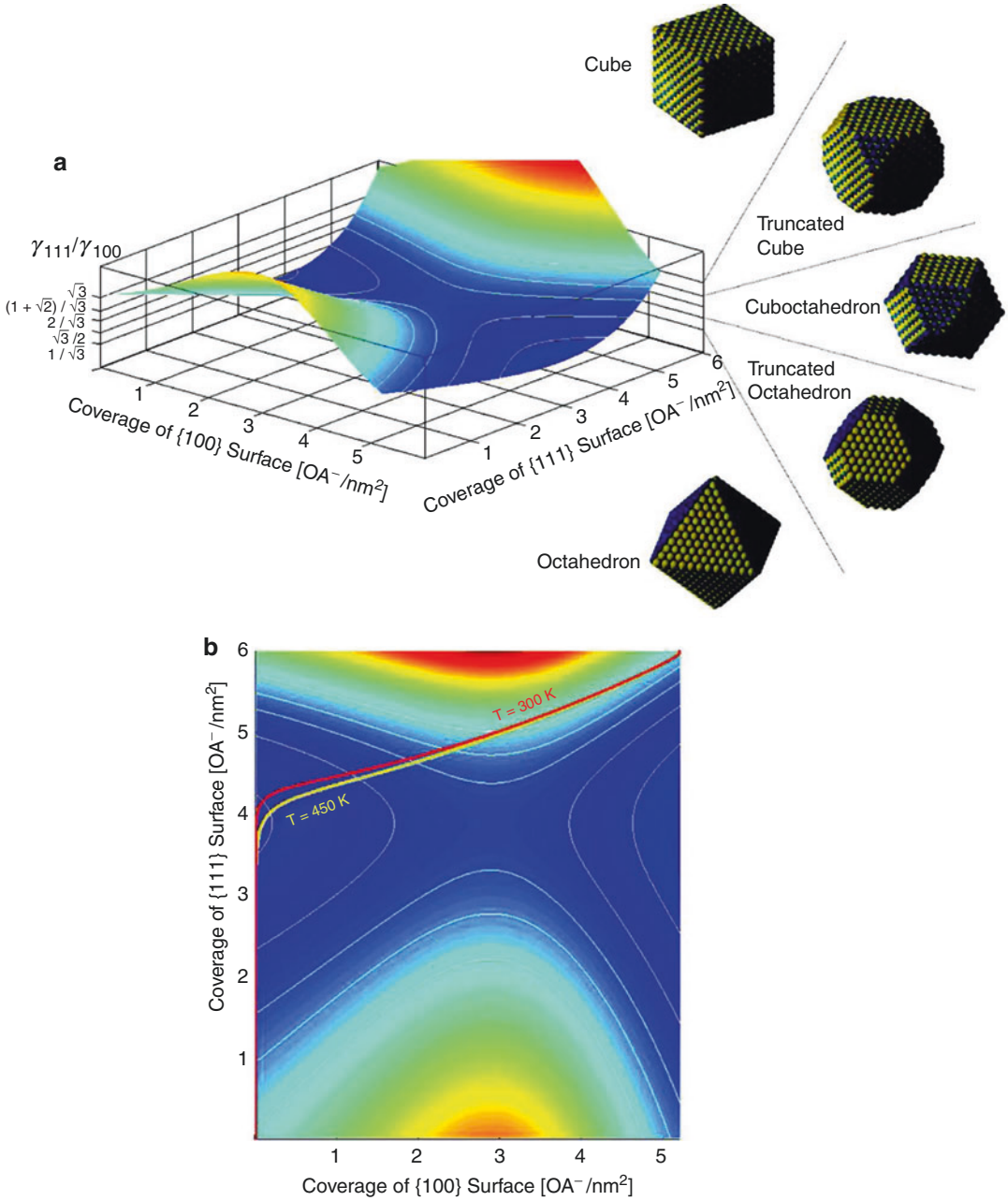
In some cases, nevertheless, such as particle bottom-up synthesis, the particle shape is mainly controlled by kinetic effects and can thus substantially deviate from the ideal thermodynamic shape. The small clusters formed upon initial nucleation and their interactions with other species, e.g., the surfactant molecules used in colloidal synthetic methods, can be explicitly treated with DFT [11]. However, as the particle size increases, the growth morphology can be predicted through the kinetic version of the Wulff theorem, whereby the distance of a facet i to the particle center is proportional to its growth rate v_i. In the framework of transition state theory, the growth rate of a facet (in units of nm/s in a specific direction) can be expressed as

$$v_i = v_0 \exp\left(-\frac{\Delta G_i^*}{kT}\right), \tag{9}$$

where v_0 is the rate of an ideal barrierless growth process and ΔG_i^* is the effective activation barrier for the growth of a monolayer. As such, ΔG_i^* depends on the specific mechanisms of atom diffusion and adsorption on the surface and can be estimated through molecular-dynamics simulations. This general approach has been demonstrated in the case of urea particle growth [12], where the facet growth process and rate are found to be orientation dependent and may furthermore be modified by the addition of simple surfactant molecules such as biuret and acetone as a way to control the particle growth morphology. The simulation times required for unveiling these mechanistic features are of the order of tenths of μs, implying the need for classical force fields, either empirical or fitted on DFT calculations. There is much interest in applying this approach to the case of metal nanoparticle growth possibly in combination with more complex surfactant molecules, as exemplified by the recent development of DFT-based effective force fields for the interaction of polyvinylpyrrolidone chains on Ag surfaces [13].

DFT-Based Modeling of Elemental Distributions in Alloy Nanoparticles

The physical and chemical properties, particularly at the surface, of alloy nanoparticles depend critically on their elemental distribution. For instance,



Metal Nanoparticles from First Principles, Fig. 3 (**a**) Surface energies and thermodynamic shapes of the PbSe particle as a function of OA coverages on (100) and (111) facets as obtained from the Wulff construction. (**b**) Isotherms for OA adsorption on (100) and (111) facets at 300 K (*red line*) and 450 K (*yellow line*), illustrating how the equilibrium shape can be varied through ligand concentration (Reprinted with permission from Ref. [10]. Copyright 2012 American Chemical Society)

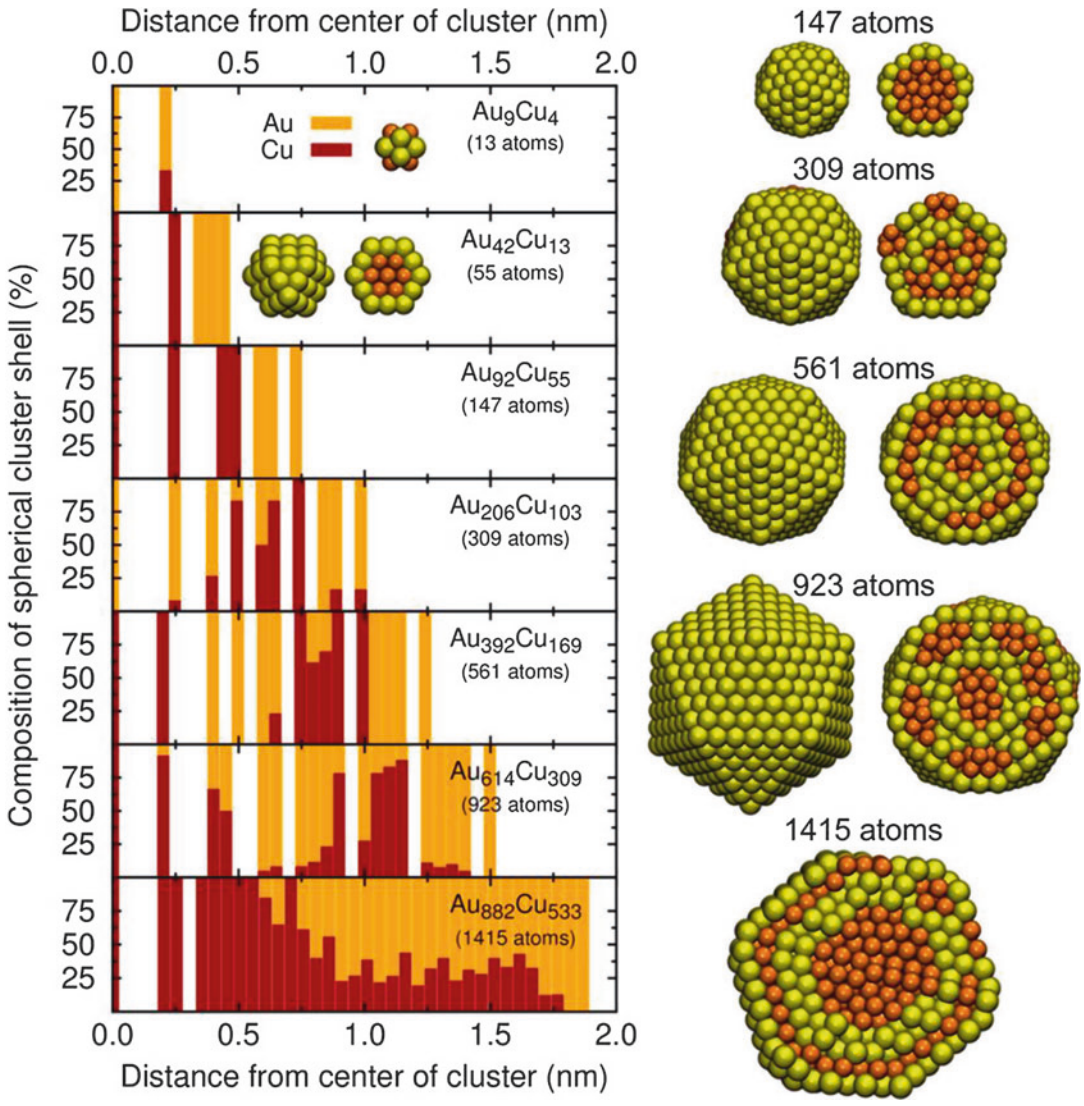
core-shell bimetallic nanoparticles, in which the constitutive elements tend to segregate from each other toward the center and surface of the particle, have been used in an increasing number of applications. The search for new stable structures can be aided by theoretical tools that, for a given chemical composition, minimize the particle energy over the large number of possible

elemental distributions. This can be achieved through a multiscale treatment consisting in the following two steps: first develop a model Hamiltonian that can provide the energy of any configuration at a reasonable trade-off between accuracy and computational cost; second minimize the value of that Hamiltonian over the configurational phase space through an optimization algorithm, such as simulated annealing. Typically, to develop the model Hamiltonian, one first builds an appropriate set of reference configurations and computes their energies from first principles; then the parameters of the Hamiltonian are fitted on this set of accurate energies. One of the most delicate aspects of this approach is the choice of the functional or structural form of the Hamiltonian. Three common forms are here briefly listed:

- **Analytical formula**: The energies are fitted on an explicit analytical expression as is the case, for instance, in the embedded-atom model (EAM), which has been extended, for instance, to the case of bialloys.
- **Cluster expansion (CE)**: The value of the energy is written as the Taylor-like expansion of two (three, four, etc.)-body interactions between first (second, third, etc.)-nearest-neighbor atoms around a set of fixed sites, where the fitting parameters are the weights in the expansion.
- **Artificial neural network (NN)**: A set of variables characterizing the system structure is fed into a neural network (NN) consisting of several layers of artificial neurons. Each neuron in a layer processes a linear combination of the outputs of the previous layer through a sigmoid function, and the predicted energy of the system is obtained as the output of the last layer consisting of a single neuron. The fitting parameters are the weights in the linear combinations between two layers.

Each form has its advantages and drawbacks. The CE approach allows a quick and easy analysis of the energy in terms of local atomic contributions, but it is constrained by its reliance on a set of fixed sites, which may not apply for systems with a large number of translational and rotational degrees of freedom. As a flexible and universal fitting scheme, NNs are useful, for instance, when the analytical approach cannot cover the entire phase space; however, the fitting error may converge slowly with the size of the NN. Thus, in some cases, it is found necessary to combine these approaches. The model Hamiltonian that is obtained must finally be cross-validated against a set of configurations not included in the original reference set, and the development workflow "reference set → model Hamiltonian → cross-validation" must be repeated self-consistently until the desired convergence is reached.

These principles have been illustrated by the prediction of the equilibrium composition of Au-Cu bimetallic nanoparticles in vacuum and water, and particularly their surface structure as a key to understand their catalytic activity for CO_2 reduction [14]. A NN is used to predict the energy of an individual atom in the particle. For transferability across all possible structures, the NN input variables representing the local environment around the atom are obtained by integrating the Cartesian coordinates of the neighbor atoms into a fixed number of radial and angular symmetry functions. The total energy of the particle is then simply obtained as the sum of the atom energies. Thus, an analytical approach (the symmetry functions) and a CE approach (summing over the atom energies) are embedded within the NN and are essential for the accuracy. The fitting parameters of the resulting model Hamiltonian are obtained by training the NN on a reference set of ~20,000 local environments including bulk and slab structures, atomic clusters, and molecules, where the reference energies are computed with DFT and include the van der Waals interactions. A NN consisting of two hidden layers of five neurons each is found to be sufficient to approximate the DFT energies within a mean absolute error of 6 meV per atom and allows computing the total energy of a 6-nm nanoparticle in less than a minute on a single core. Simulated annealing of the nanoparticle with this Hamiltonian shows that the equilibrium distribution differs whether the particle is placed in vacuum or in water. In vacuum, Au atoms tend to segregate toward the low-coordinated surface sites because gold is more noble than copper, as shown in Fig. 4. However, the presence of water lifts this



Metal Nanoparticles from First Principles, Fig. 4 Optimized compositions of Au-Cu nanoparticles in vacuum as obtained from simulated annealing using the NN potential. Au atoms tend to segregate to the surface (Reprinted with permission from Ref. [14]. Copyright 2014 American Chemical Society)

segregation effect by promoting mixed Au-Cu surface structures as a result of the stronger van der Waals interaction of water with Cu atoms.

Electrochemical Stability of Metal Nanoparticles

It is important to be able to predict the electrochemical stability of nanoparticles in aqueous environment. For instance, the dissolution of small Pt particles and coarsening into larger particles through Ostwald ripening is a phenomenon observed in practical catalyst arrays that is responsible for the performance degradation of fuel cell systems over time. While the Gibbs-Thomson effect in basic thermodynamics predicts a decreased stability with reduced particle size, there is a need to better characterize this effect at the atomic scale. The electrochemical stability of

metals as a function of solution pH and electrode potential U, also known as Pourbaix diagrams, can be determined through an established method combining first-principles techniques and a careful set of reference energies [9]. It is equally applicable to nanoparticles, and it is outlined in the following for Pt particles in aqueous solution. In the stability region, the particle may adsorb hydroxides and oxides generated by electrochemical water splitting:

$$\begin{aligned} \mathrm{Pt} + \mathrm{H_2O_{(l)}} &\rightarrow \mathrm{PtOH} + \mathrm{H^+_{aq}} + \mathrm{e^-} \\ \mathrm{PtOH} &\rightarrow \mathrm{PtO} + \mathrm{H^+_{aq}} + \mathrm{e^-}, \end{aligned} \tag{10}$$

where the adsorption of OH and O on the particle surface is thermodynamically driven by decreasing the electron chemical potential -eU and increasing the pH. Alternatively, in the instability region, the particle is subject to dissolution:

$$\mathrm{Pt} \rightarrow \mathrm{Pt^{2+}_{aq}} + 2\mathrm{e^-}, \tag{11}$$

which is also driven by increasing the electrode potential. In general, the equilibrium state of the particle at given pH and U may be determined by minimizing the total free energy over the different adsorption patterns and coverages of OH and O as well as the dissolution state. However, the free energy of liquid water and ionic species such as Pt^{2+} and H^+ is challenging to estimate from first principles. A solution then consists in assigning free energies to these species in a way that is consistent with well-known reaction energies. For water, one knows the experimental free energy of condensation $\mathrm{H_2O_{(g)}} \rightarrow \mathrm{H_2O_{(l)}}$. Thus, the chemical potential of water in the liquid phase $\mu_{\mathrm{H_2O_{(l)}}}$ can be related to that in the gas phase, which in turn is obtained as

$$\mu_{\mathrm{H_2O_{(g)}}} = E_{\mathrm{el}} + \mathrm{ZPE} - Ts_{\mathrm{H_2O_{(g)}}}, \tag{12}$$

where E_{el} and ZPE are the electronic and zero-point energies as obtained from DFT, while $s_{\mathrm{H_2O_{(g)}}}$ is the entropy of the H_2O molecule in the gas phase as taken from experimental tables. For H^+, one can use the fact that at any temperature the free energy of the reaction

$$\frac{1}{2}\mathrm{H_2} \rightarrow \mathrm{H^+_{aq}} + \mathrm{e^-} \tag{13}$$

is by definition 0 in standard conditions of pressure and concentration and at the potential of the standard hydrogen electrode (SHE). Thus, at any pH and electrode potential U, the chemical potential of the proton + electron can be written

$$\mu_{\mathrm{H^+_{aq}+e^-}} = \frac{1}{2}\mu^{\mathrm{1atm.}}_{\mathrm{H_2}} - kT\log(10)\mathrm{pH} - eU, \tag{14}$$

where U is referenced to the SHE and $\mu^{\mathrm{1atm.}}_{\mathrm{H_2}}$ is the chemical potential of a H_2 molecule in the gas phase at one atmosphere and is obtained as in Eq. 12. Likewise, the free energy of the dissolution reaction

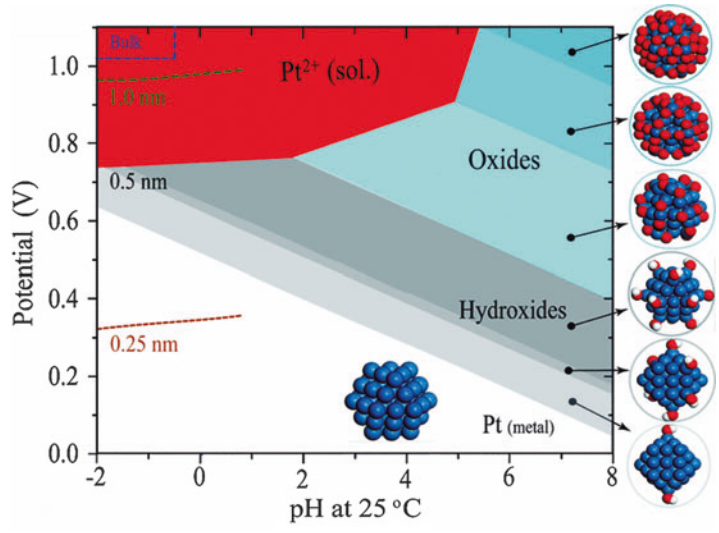
$$\mathrm{PtO} + 2\mathrm{H^+} \rightarrow \mathrm{Pt^{2+}_{aq}} + \mathrm{H_2O} \tag{15}$$

is well known experimentally at given concentrations of the ions. Consequently, the chemical potential of $\mathrm{Pt^{2+}_{aq}} + 2\mathrm{e^-}$ can be related to those of PtO, $H^+_{aq} + e^-$, and H_2O, where μ_{PtO} is dominated by the electronic energy and is obtained from DFT.

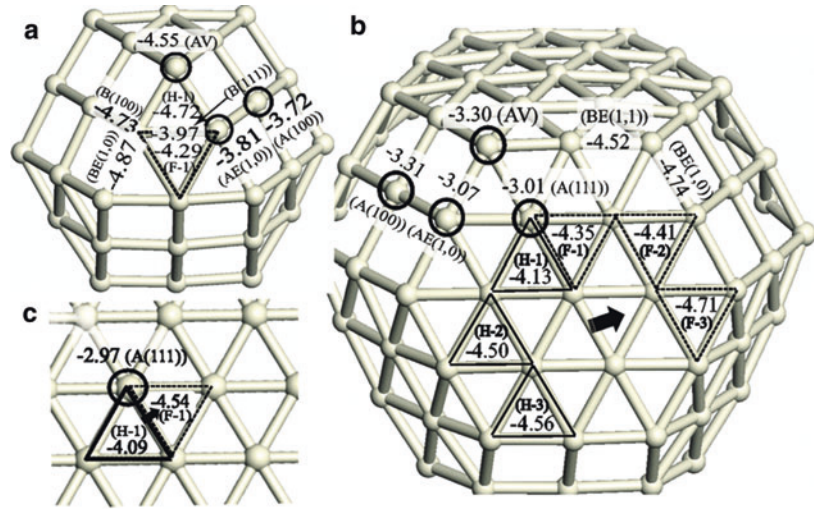
The aqueous chemical potentials obtained through this reference scheme can then be combined with the DFT energies of a set of Pt cuboctahedra with various adsorption patterns to determine the Pourbaix diagram for particle sizes between 0.25 and 1 nm, as shown in Fig. 5 [15]. Particularly, this approach allows quantifying the extension of the dissolution region at low pH and high electrode potential with respect to bulk platinum as a result of the dramatic decrease in the average cohesive energies for sizes smaller than 3 nm.

Catalytic Properties

Metal nanoparticles are commonly used in catalytic arrays as a way to maximize the active surface area per unit mass of the catalyst material.



Metal Nanoparticles from First Principles, Fig. 5 Ab initio Pourbaix diagram of a 0.5-nm Pt particle (Reprinted figure with permission from Ref. [15]. Copyright (2012) by the American Physical Society (a))



Metal Nanoparticles from First Principles, Fig. 6 Chemisorption energies of O at various surface sites of Pt nanoparticles as obtained from DFT. Both the particle size and surface site effects can be evaluated (Reprinted figure with permission from Ref. [19]. Copyright (2008) by the American Physical Sciety (a))

However, reducing the particle size may also modify the intrinsic catalytic activity owing to the larger number of low-coordinated sites sitting on the vertices and edges of the particle, which tend to exhibit stronger chemical reactivities. The binding energy of a given adsorbate on the different sites of a particle may be computed by DFT, as illustrated in Fig. 6, and the resulting effect on the overall catalytic turnover depends in general on the specific application. For instance, a stronger binding of OH on the low-coordinated sites of the particle may accelerate the CO oxidation reaction by increasing the amount of oxidant on the catalyst surface. Inversely, this effect can be deleterious for the oxygen reduction reaction by reducing the number of active sites for oxygen adsorption. The DFT approach can be equally applied to determine activation barriers on different sites of the nanoparticle [16].

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Synthesis of Gold Nanoparticles
- ▶ Synthesis of Subnanometric Metal Nanoparticles

References

1. Wagner, F.E., Haslbeck, S., Stievano, L., Calogero, S., Pankhurst, Q.A., Martinek, K.-P.: Before striking gold in gold-ruby glass. Nature **407**, 691–692 (2000)
2. Martin, R.M.: Electronic structure – basic theory and practical Methods. Cambridge University Press, Cambridge (2004)
3. Sholl, D., Steckel, J.A.: Density Functional Theory: A Practical Introduction. Wiley-Interscience, Hoboken (2009)
4. Mariscal, M.M., Oviedo, O.A., Leiva, E.P.M.: Metal Clusters and Nanoalloys – From Modeling to Applications. Springer, New York (2013)
5. Desireddy, A., et al.: Ultrastable silver nanoparticles. Nature **501**, 399–402 (2013)
6. Barnard, A.S.: Direct comparison of kinetic and thermodynamic influences on gold nanomorphology. Acc. Chem. Res. **45**, 1688–1697 (2012)
7. Gruner, M.E., Rollmann, G., Hucht, A., Entel, P.: Structural and magnetic properties of transition metal nanoparticles from first principles. In: Haug, R. (ed.) Advances in Solid State Physics, pp. 117–128. Springer, Berlin (2008)
8. Chepulskii, R.V., Curtarolo, S.: Ab initio insights on the shapes of platinum nanocatalysts. ACS Nano **5**, 247–254 (2011)
9. Karlberg, G.S., Jaramillo, T.F., Skúlason, E., Rossmeisl, J., Bligaard, T., Nørskov, J.K.: Cyclic voltammograms for H on Pt (111) and Pt(100) from first principles. Phys. Rev. Lett. **99**, 126101 (2007)
10. Bealing, C.R., Baumgardner, W.J., Choi, J.J., Hanrath, T., Hennig, R.G.: Predicting nanocrystal shape through consideration of surface-ligand interactions. ACS Nano **6**, 2118–2127 (2012)
11. Mpourmpakis, G., Vlachos, D.G.: Growth mechanisms of metal nanoparticles via first principles. Phys. Rev. Lett. **102**, 155505 (2009)
12. Salvalaglio, M., Vetter, T., Giberti, F., Mazzotti, M., Parrinello, M.: Uncovering molecular details of urea crystal growth in the presence of additives. J. Am. Chem. Soc. **134**, 17221–17233 (2012)
13. Zhou, Y., Saidi, W.A., Fichthorn, K.A.: A force field for describing the polyvinylpyrrolidone-mediated solution-phase synthesis of shape-selective Ag nanoparticles. J. Phys. Chem. C **118**, 3366–3374 (2014)
14. Artrith, N., Kolpak, A.M.: Understanding the composition and activity of electrocatalytic nanoalloys in aqueous solvents: a combination of DFT and accurate neural network potentials. Nano Lett. **14**, 2670–2676 (2014)
15. Persson, K.A., Waldwick, B., Lazic, P., Ceder, G.: Prediction of solid-aqueous equilibria: scheme to combine first-principles calculations of solids with experimental aqueous states. Phys. Rev. B **85**, 235438 (2012)
16. Jennings, P.C., Aleksandrov, H.A., Neyman, K.M., Johnston, R.L.: A DFT study of oxygen dissociation on platinum based nanoparticles. Nanoscale **6**, 1153–1165 (2014)
17. Barcaro, G., Fortunelli, A.: A Magic Pd-Ag Binary Cluster on the Fs-Defected MgO(100) Surface. J. Phys. Chem. C **111**, 11384–11389 (2007)
18. Barnard, A.S., Young, N.P., Kirkland, A.I., van Huis, M.A., Xu, H.: Nanogold: A Quantitative Phase Map. ACS Nano **3**, 1431–1436 (2009)
19. Han, B.C., Miranda, C.R., Ceder, G.: Effect of Particle Size and Surface Structure on Adsorption of O and OH on Platinum Nanoparticles: A First-Principles Study. Phys. Rev. B **77**, 075410 (2008). doi:10.1103/PhysRevB.77.075410

Metal Plating

▶ Plating

Metallic Waveguides

▶ Light Localization for Nano-optical Devices

Metallofullerene

▶ Fullerenes for Drug Delivery

Metalorganic Chemical Vapor Deposition (MOCVD)

▶ Chemical Vapor Deposition (CVD)

Metal–Organic Interfaces

▶ Electrode–Organic Interface Physics

Metal-Organic Molecular Beam Epitaxy (MOMBE)

▶ Physical Vapor Deposition

Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy

Andrea Lamberti
Department of Applied Science and Technology, Politecnico di Torino, Torino, Italy

Synonyms

Nanostructured SERS-active metal oxides

Definition

A metal oxide is a chemical compound that contains one metal element and at least one oxygen atom. If one of the characteristic dimensions of the material was at the nanoscale, thus, it was called metal oxide nanostructure. Thanks to their smart properties, metal oxide nanostructures, alone or decorated with noble metal nanoparticles, can enhance the Raman signal and are called SERS-active substrates.

Metal Oxide Nanostructure Applications

Nanomaterials are able to show completely different properties from their conventional bulk counterpart, and several applications are suitable to exploit the unique properties of these materials. Figure 1 collects the enormous variety of application of nanoparticles, but the same sectors take advantage also from different nanomaterial shapes. Among the inorganic semiconductor materials, nanostructured semiconductors have recently invaded the scientific activity related to devices for sensing, energy conversion, and energy storage. Indeed these materials can overcome the limitation imposed by the relatively poor properties of the standard electrodes, showing high surface area, superior carrier mobility, and significantly low charge recombination rates.

Metal oxide (MO) nanostructures are the focus of current research efforts in nanotechnology since they are the most common minerals on the Earth and thanks to their special shapes, compositions, and chemical and physical properties [2]. They have now been widely used in the design of energy saving and harvesting devices such as lithium-ion batteries, fuel cells, solar cells, and even transistors/FETs, light-emitting devices (LEDs), hydrogen production by water photolysis and its storage, water and air purification by degradation of organic/inorganic pollutants, bio-sensing devices, environmental monitoring by their applications in the fabrication of gas, humidity and temperature sensors, and photodetectors [3].

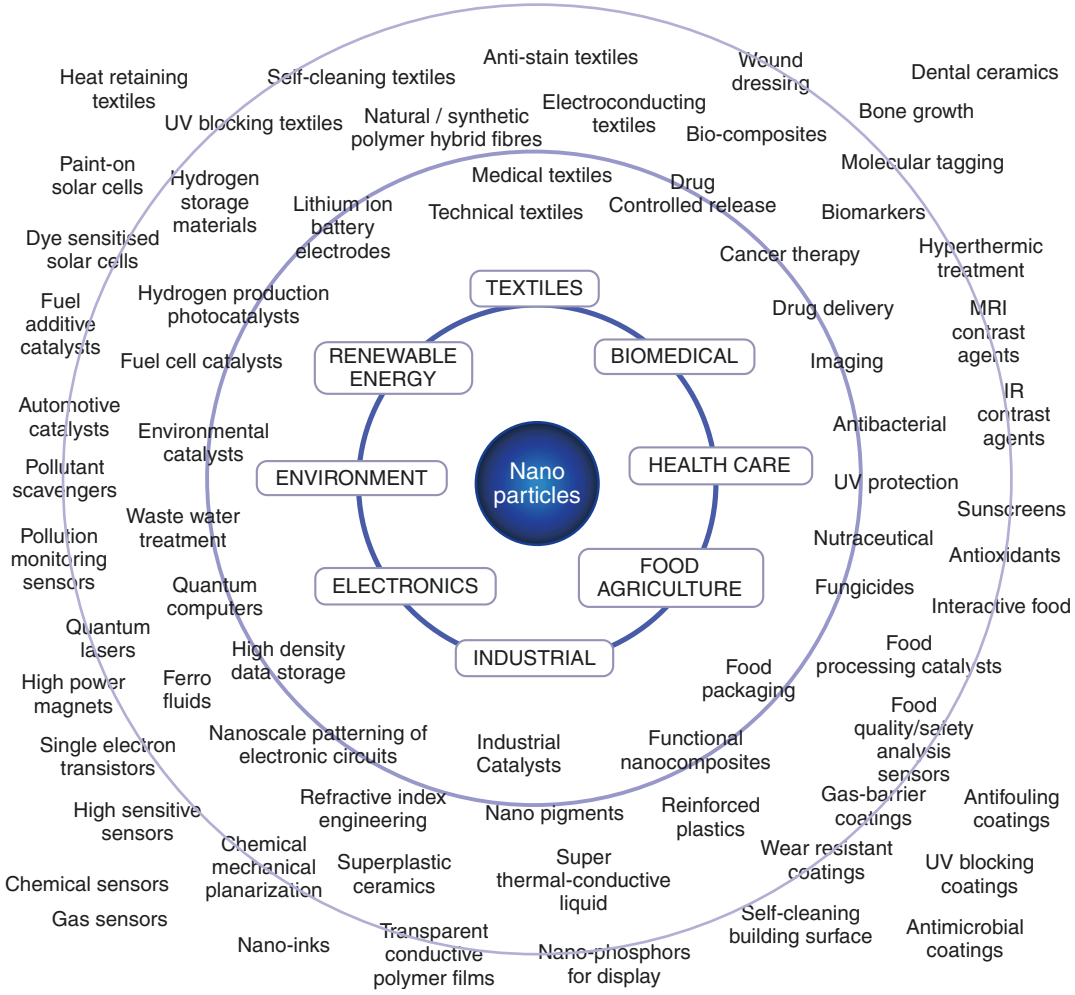
Besides the great application potential, oxide-based nanomaterials such as ZnO and TiO_2 have recently revolutionized the nanomaterial research because of the availability of soft-chemical synthesis.

Metal Oxide Nanostructure Synthesis and Deposition

Nanostructured metal oxide semiconductors can be deposited on any type of substrates such as metals, semiconductors, polymers, and flexible plastics unlike other III-V semiconductor and silicon, which requires specific and costly substrates.

All these metal oxide materials can be produced in different shapes (such as nanowires, nanobelts, nanorods, nanotubes, nanocombs, nanorings, nanohelixes/nanosprings, nanocages and nanosheets, and nanostars) depending on the synthesis routes, which can range over a wide variety of techniques [4].

The methods usually employed for metal oxide film deposition can be divided into two groups based on the nature of the deposition process, as depicted in Fig. 2. The physical methods include pulsed laser deposition (PLD), molecular beam epitaxy (MBE), and sputtering. The chemical methods comprise gas-phase deposition methods and solution techniques. The gas-phase methods are chemical vapor deposition (CVD), metal organic chemical vapor deposition (MOCVD), and atomic layer deposition (ALD), while spray



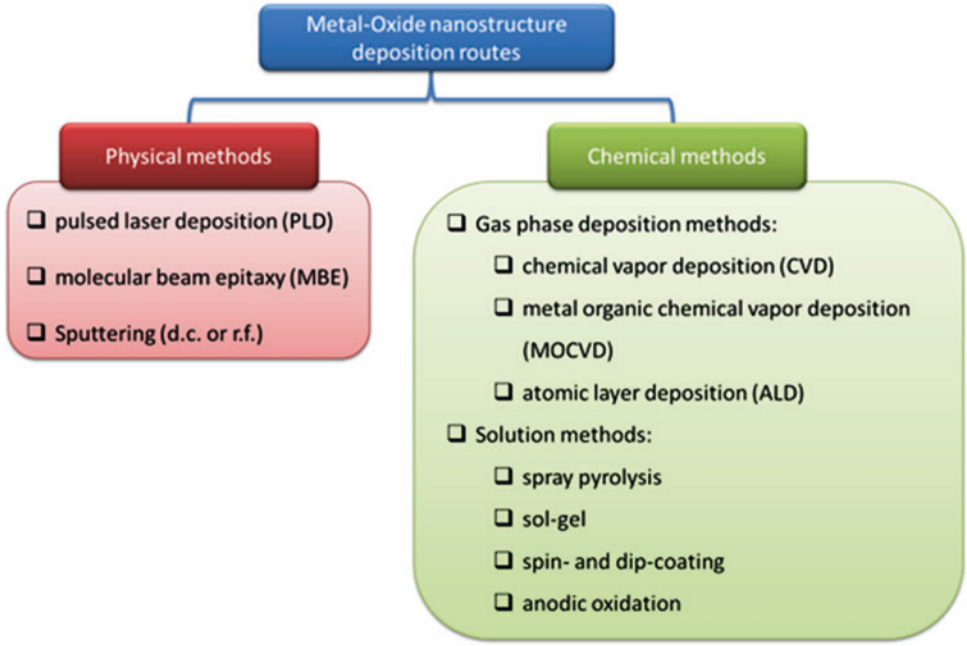
Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy, Fig. 1 Scheme illustrating all the fields of application of nanomaterials (Adapted from Ref. [1])

pyrolysis, sol–gel, and spin- and dip-coating methods employ precursor solutions.

Each method has certain advantages and disadvantages on issues ranging from ease of use to operating expense to time required for growth to quantity and quality of material produced.

Among them, the most employed to deposit metal oxide nanostructures for SERS application are:

– Sputter deposition, which involves bombardment of a target (metal or metal oxide) with a plasma (usually a noble gas, such as argon). The relatively low target temperature makes sputtering one of the most flexible deposition techniques. The sputtered atoms ejected into the gas phase are not in their thermodynamic equilibrium state, so they tend to condense back into the solid phase on the substrate surface. Film-grown rate is well controllable, allowing the reproducibility of the samples' thickness.
– Chemical vapor deposition (CVD), in which one or more precursors, usually activated by temperature, react at a substrate to form a film. The molecules must have a vapor pressure high enough to be transported to the reaction chamber, but yet they must have a high thermal



Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy, Fig. 2 Scheme collecting the metal oxide nanostructure deposition methods

stability to prevent particle formation in the gas phase. In this method, controllable film growth rate is attainable.

– Hydrothermal processing that can be defined as any heterogeneous reaction in an aqueous solvent (or nonaqueous solvent for solvothermal processing) under high pressure and temperature conditions, which induces the dissolution and recrystallization of materials that are relatively insoluble under ordinary conditions.
– Anodic oxidation, i.e., the process of forming an oxide on a metal surface by applying an electrical current or potential through a cell containing a suitable electrolyte. The metals that can be anodized to form oriented nanostructures belong to the so-called valve-metal group. The characteristics of a layer treated with anodic oxidation can vary according to the type and concentration of the electrolytes as well as the processing variables used during anodic oxidation. Compared with other synthesis approach, electrochemical anodization is a simple and convenient technique to fabricate uniform layer of vertically self-oriented nanostructures.

SERS-Active Substrates

Surface-enhanced Raman scattering (SERS) has been a growing area of interest in recent years due to its potential as label-free detection in materials science, biophysics, medical diagnostics, and molecular biology.

SERS technique is based on impressively enhanced Raman signal from Raman-active molecules that have been adsorbed onto certain metal surfaces (Au, Ag, Cu, etc.). These increases in the intensity of Raman signal have been commonly observed in the order of 10^4–10^6 and can be as high as 10^8 and 10^{14} for some particular systems [5].

There are two main mechanisms that contribute to the scattering enhancement observed in SERS: the electromagnetic and the chemical enhancement. The former is a moderate long-range effect and comes from the interaction of

free electrons on the metal surface with the excitation laser beam to yield surface plasmon resonance (SPR). The SPR is basically a collective oscillating electron wave on the metal surface. The SPR intensity on the metal surface is strictly related to the excitation light wavelength and the surface roughness of substrates and the dielectric constant of an eventual external medium at the interface of the metal structure. By acting on the shape, the size, and the spacing between metal surfaces (such as in the case of nanoparticles), it is possible to tune the SPR to obtain the optimized SERS signal from the metal nanostructures at a desired wavelength [5].

The chemical enhancement mechanism instead primarily originates from the charge-transfer (CT) mechanism involving changes to the molecule electronic states due to its chemisorption.

In noble metal nanoparticle (NP) interaction, the SPR mechanism is the main contribution. The most common types of SERS-active substrates exhibiting the largest enhancement effects are clusters of silver and gold NPs, generally used in colloidal solution.

Besides colloidal solutions, several lithographic techniques have been employed for the production of uniform and controllable SERS substrates with fully controlled shape and size. Despite the high efficiency of such substrates, the involved technology presents some drawbacks concerning cost and complexity.

Metal Oxide SERS Substrates

Recently, it has been observed that several metal oxides such as TiO_2, ZnO, α-Fe_2O_3, and CuO also exhibit SERS capability, although their enhancement effects are weaker than those of noble metal substrates. In all of these cases, CT at the semiconductor−analyte interface plays a major role in their SERS activity.

As described above, MO materials can be produced in different shapes (such as nanowires, nanorods, nanotubes, nanosheets, etc.) depending on the synthesis routes, which can range over a wide variety of techniques. The high specific surface area intrinsic in their nanostructuration allows a higher amount of metal particle loading with the consequent formation of 3D plasmon "hot spots" which could provide giant enhancement. Additionally, the 3D network can provide larger adsorbing area for analyte molecules than the standard planar substrates further increasing the Raman signal.

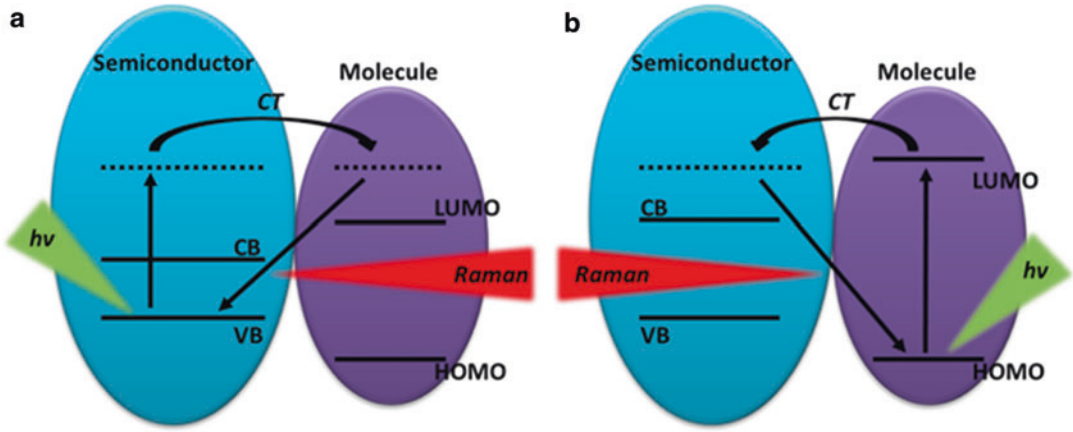
The first observation of Raman signal enhancement on metal oxide surface dates back to 1983, when Yamada and coworkers found SERS effect of pyridine molecules anchored on single crystal TiO_2 and NiO surfaces [6] attributing such improvement to the chemical enhancement mechanism. That work opened the way to the study of metal oxide surfaces as innovative SERS-active substrates, while the growth of nanoscience and nanotechnology has contributed to a rapid development of SERS-active nanostructured substrates.

Nanostructured TiO_2 indeed shows higher enhancement with respect to the single crystal counterpart: reducing the characteristic dimension and increasing the surface area, the charge transfer from TiO_2 nanocrystal to adsorbed molecule is strongly amplified.

As introduced before, the chemical enhancement mechanism implies that, when an analyte molecule was adsorbed on the surface of the SERS-active material, a chemical interaction between the molecules and the atoms of the substrate surface takes place. This interaction directly acts on the electron density distribution of the molecule in correlation with the vibrational energy states. Consequently the variation of the electron density distribution could enhance the Raman scattering cross section in respect to specific molecular vibrational modes.

According to recent literature [7], it is possible to distinguish different kinds of chemical enhancement mechanisms: the chemical-bonding enhancement, the surface complexes resonance enhancement, and the photon-induced charge-transfer (PICT) enhancement.

Among them, the PICT mechanism is the most suitable to explain the SERS from metal oxide semiconductor nanostructures. This photoactivated charge transfer can occur from the semiconductor to the molecule that is in the



Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy, Fig. 3 Scheme of the PICT process

opposite direction. To ensure its occurrence, it is necessary to actuate a thermodynamically favored condition under light excitation, i.e., the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) of the adsorbed molecule should energetically match the valence band (VB) and the conduction band (CB) of the semiconductor. This process mainly involves three electronic transitions as depicted in Fig. 3.

The PICT from the semiconductor surface to the adsorbed molecule can be summarized in the following steps (Fig. 3a):

1. An electron in the VB of the semiconductor is excited by the incident light to the CB.
2. The photogenerated electron can be rapidly transferred by resonant tunneling from the metal oxide CB to a matching energetic level above the LUMO of the molecule involving a vibrational level.
3. The as-transferred electron could turn back to the semiconductor VB (through an electron–hole recombination), thus, a vibrational quantum of energy can be involved, giving a Stokes Raman scattering process.

Likewise, the PICT from the molecule to the semiconductor involves the other following steps (Fig. 3b):

1. An electron in the HOMO of the adsorbed molecule is photoexcited to an energy level above the LUMO.
2. The photogenerated electron can be quickly transferred by resonant tunneling to a matching energetic level within the metal oxide CB.
3. The as-transferred electron could turn back to the HOMO of molecule, and a vibrational quantum of energy can be involved giving a Stokes Raman scattering process.

It should be underlined that, contrary to what happens in the other two chemical enhancement mechanisms, SERS activated by PICT does not need a molecule chemisorption on the metal oxide surface.

Several kinds of metal oxide nanostructures have been investigated in the last decades. Yang et al. [8] observed SERS signal from the 4-mercaptobenzoic acid molecules adsorbed on TiO_2 nanoparticles, thanks to the contribution of the TiO_2-to-molecule CT mechanism related to the surface state energy level of TiO_2 nanoparticles. Obviously such CT process is strongly related to the intrinsic nature of the adsorbed molecules and to the substrate surface properties.

Musumeci et al. [9] reported strong enhancement of selected Raman-active modes of biologically active molecules adsorbed on the surface of

TiO_2 NPs. In contrast with the previous example, they presented the evidence of the electronic mechanism involving molecule-to-TiO_2 charge transfer, which results in significantly larger enhancement factors.

Ivano Alessandri [10] demonstrates that a remarkable enhancement of Raman scattering can be achieved by TiO_2 shell-based spherical resonators ascribing the effect to the synergistic combination of high refractive index of the shell layer, multiple light scattering through the spheres, and related geometrical factors.

Differently from titania, SERS effect on zinc oxide was firstly observed in 1996 [11] increasing the interest of the SERS scientific community on this semiconductor.

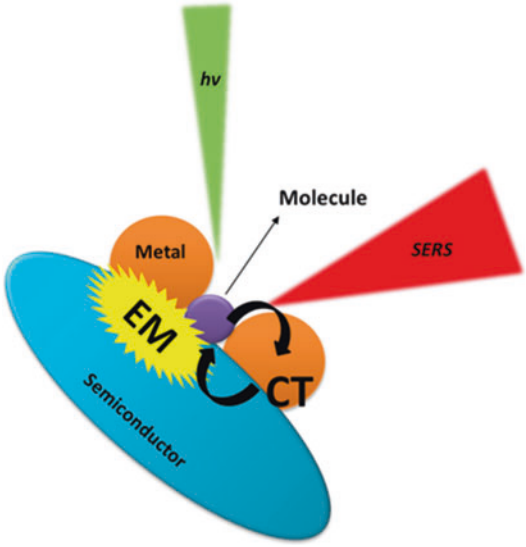
A strong CT process between ZnO nanocrystals and 4-mercaptopyridine electronic states was clearly demonstrated by Wang et al. [12] observing an excitation wavelength-dependent behavior.

SERS signal from 4-aminobenzenethiol on porous Cu_2O and Cu_2O/CuO nanowires was also demonstrated [13] showing a strong signal dependence to the Cu_2O/CuO ratio and particle size of the porous nanowires. The extraordinary enhancement of such SERS signals is attributed to numerous nanoparticles acting as hot spots on the surface of the porous nanowires.

More recently [14], surface-enhanced Raman signal of 4-mercaptobenzoic acid (4-MBA) has been detected on the surface of Cu_2O nanospheres. Simulations were employed to demonstrate the nature of the enhancement mechanisms operative between the molecules and the semiconductor. Density functional theory (DFT) calculations suggested a charge-transfer (CT) transition process between the molecules and the Cu_2O nanospheres.

Noble Metal Nanoparticle-Decorated SERS Substrates

More significant improvements can be obtained by using composites of nanostructured semiconductors and noble metals, which integrate the advantages of the electromagnetic coupling among metal nanoparticles with those of chemical enhancement arising from both noble metal and semiconductor charge transfer with adsorbed molecules (as schematically represented in Fig. 4).

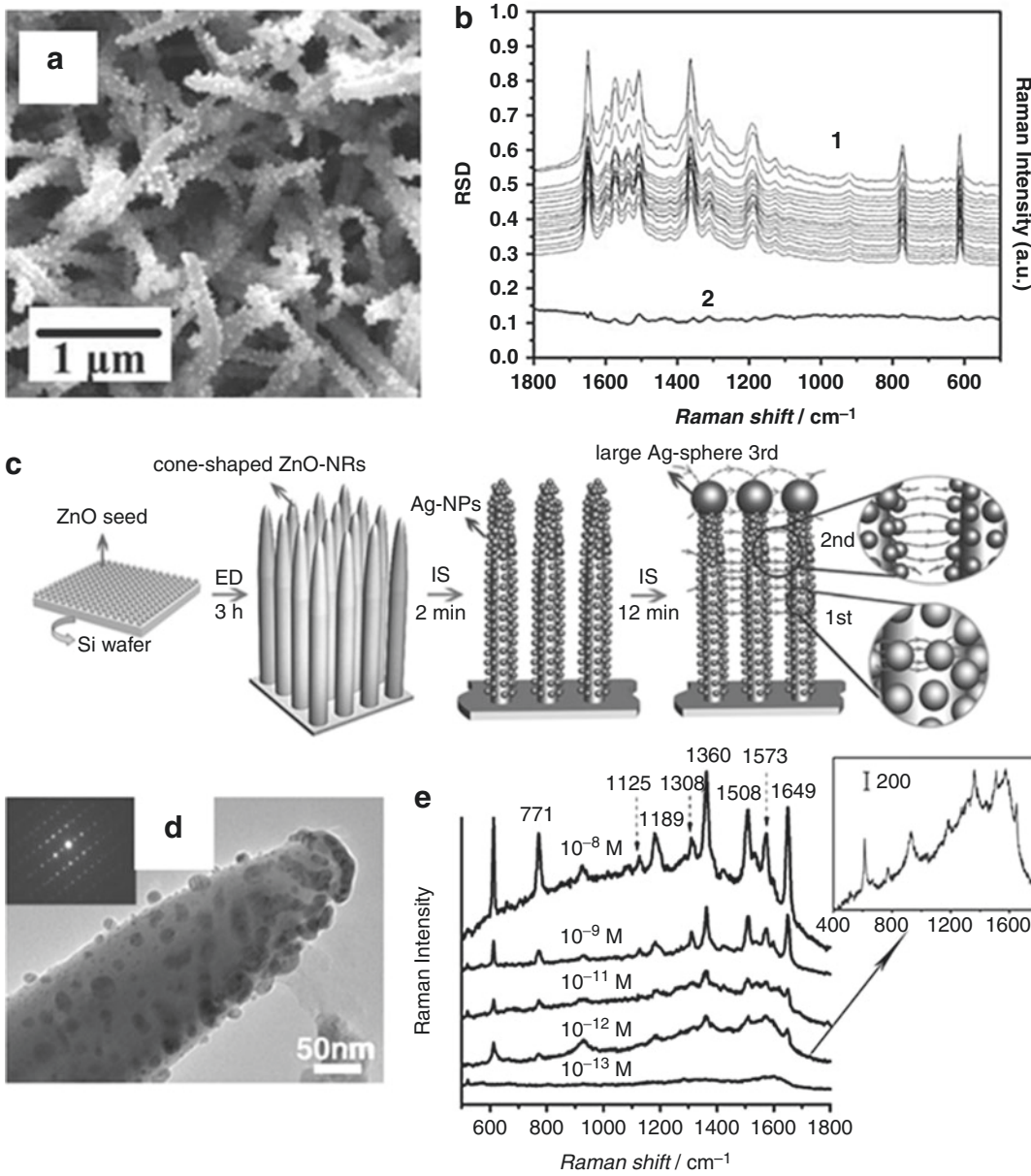


Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy, Fig. 4 Scheme of the hybrid MO-noble metal nanostructure under excitation

Thanks to their nanostructuration, the specific surface area is usually high, thus allowing an increased loading of metal particles as well as the high possibility of forming 3D plasmon "hot spots" which provide giant enhancement. Moreover, the 3D substrates can also supply larger surface area for adsorbing probe molecules than the 2D substrates.

Recently, following this concept, several composites were investigated as SERS-active substrate. Li and coworkers [15] have investigated multifunctional Au-coated TiO_2 nanotube array made by synthesis of a TiO_2 nanotube array through a template, followed by deposition of Au particles onto the TiO_2 surface. In Fig. 5a, a FESEM micrograph of the obtained composite nanostructure is reported. Such substrate exhibits superior SERS detection ability with good signal reproducibility as shown in Fig. 5b.

Despite the greater chemical stability, biocompatibility, and easier surface chemistry of gold, silver is often used since it can provide the highest efficiency in enhancing the Raman scattering due to its optical absorption properties.



Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy, Fig. 5 SEM images of Au-coated TiO_2 nanotube arrays (**a**). RSD-SERS graph: (1) a series of SERS spectra of R6G molecules collected on randomly selected 20 places of the sample and (2) the corresponding RSD value curve (**b**) (Reprinted with permission from Ref. [15]. Copyright John Wiley & Sons). Schematic image of the fabrication of an array of ZnO NRs decorated with Ag NPs (on the NR side surface) and Ag spheres (on the NR tops) on Si wafers (**c**). TEM image of the ZnO NRs after an Ag sputtering (**d**). The SERS spectra of R6G collected on the substrates exposed to different R6G concentrations (**e**) (Reprinted with permission from Ref. [16]. Copyright John Wiley & Sons)

For this reason Tan et al. [17] have chosen to fabricate three-dimensional hybrid nanostructures made of TiO_2 nanorods decorated by silver nanoparticles. The hybrid substrate exhibits excellent SERS enhancement effect due to the formation of high-density hot spots among the Ag NPs able to detect malachite green (MG) down to of 1×10^{-12} M. The experimental

results show that this substrate has high activity, uniformity, and stability. Moreover the density and sizes of Ag NPs can be easily controlled by the deposition times, which induced the decrease of the distance between Ag NPs. Ag NPs aggregations with small gap distance produce giant Raman enhancement due to strong electromagnetic coupling. There are two kinds of hot spots that contribute to the huge enhancement effect of the substrate: few hot spots are from Ag NPs adsorbed on same rod, more hot spots are from two near-crossover rods because of the 3D configuration.

A possible alternative in terms of NP-supporting material leads in zinc oxide nanostructures. Au-coated ZnO nanorods have been fabricated by Sakano and coworkers [18]. ZnO nanorod arrays on silicon substrate were coated with Au by ion beam sputtering. High SERS enhancement was observed: the Raman spectra of rhodamine 6G (R6G) as low as 1 nM were measured and explained by the field enhancement effect induced by surface plasmon polaritons of Au nanoshells coating the nanorods. Moreover, the results indicate that the SERS properties depend strongly on the geometry and the size of the nanorods that can be easily tuned acting on the growth PLD parameters.

Obviously, also silver can be coupled to ZnO as shown by Tang et al. [16] via simultaneously assembling small Ag nanoparticles (Ag NPs) and large Ag spheres onto the side surface and the top ends of large-scale vertically aligned cone-shaped ZnO nanorods (ZnO NRs), respectively, by simple physical ion sputtering (Fig. 5c). Three kinds of inter-Ag-NP gaps in 3D geometry create a huge number of SERS "hot spots" as schematically shown in Fig. 5c: "1st" stands for the gaps between the Ag NPs located on the side surface of the same NR; "2nd" stands for the gaps between the two Ag NPs located on the side surface of two neighboring NRs; and "3rd" stands for the gaps between the two large Ag spheres located on the tops of two neighboring NRs. Figure 5d shows the TEM image of the ZnO NRs after an Ag sputtering. This 3D "hot spot" network mainly contributes to the high SERS sensitivity. Moreover, the supporting chemical enhancement effect of ZnO NRs and the large surface area of the substrate also help to achieve a low detection limit for rhodamine molecule down to 10^{-12} M, as observed in Fig. 5e.

Lastly, alternative metal oxide nanostructure has been proposed by Ji et al. [19]. In their work, highly ordered porous anodic aluminum oxide (AAO) templates were fabricated using a two-step anodization approach, and then an alternating-current (AC) electrochemical deposition was used to fill AAO templates with Ag nanoparticles. High-quality SERS spectra were observed using 4-mercaptopyridine (4-MPy) as the probing molecule. Good homogeneity and stability make these substrates good candidates for SERS spectroscopy.

Intrinsic Functionalities of Metal Oxide Nanostructures for SERS

Photocatalytic Cleaning for Recyclable SERS Substrates

The Raman substrates most commonly used (Au or Ag nanoparticles on silicon) cannot be easily reused. This could be an important drawback from the economical point of view, but also from the substrates reproducibility standpoint. Indeed one of the most serious problems in SERS technique is related to the poor reproducibility of the Raman signal among identical substrates as well as in different place within the sample itself.

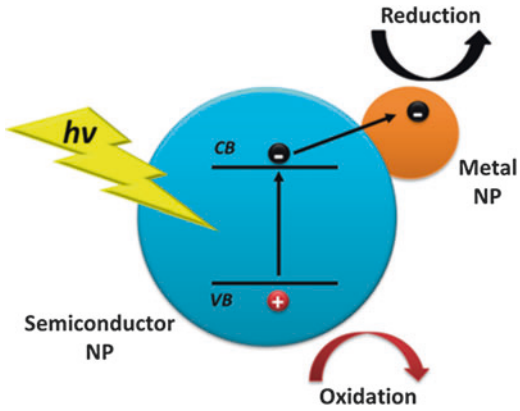
For these reasons a lot of efforts have been addressed in the last years in studying recyclable SERS substrates. As discussed above, the intrinsic photocatalytic activity of metal oxide nanostructures could be exploited to degrade the analyte molecules after the SERS analysis allowing recyclability for a new detection run.

The general mechanism of photocatalysis on MO nanostructures is based on the electron–hole pairs photogeneration under UV light, inducing the reaction with both oxygen and water molecules to promote the oxidation of organic substance.

The photocatalytic process for the degradation of organic molecules (e.g., methylene blue – MB)

can be described following a well-known mechanism – efficient photons are absorbed by MO (e.g., TiO_2), and the following reactions associated with electron and hole pair occur:

$$TiO_2 + hv\ (UV) \rightarrow\ TiO_2^*\left(e_{CB}^- + h_{VB}^+\right) \quad (1)$$



Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy, Fig. 6 Scheme of the hybrid MO-noble metal nanostructure during photocatalytic cleaning of the substrate

$$H_2O + h_{VB}^+ \rightarrow\ OH^{\bullet} + H^+ \quad (2)$$

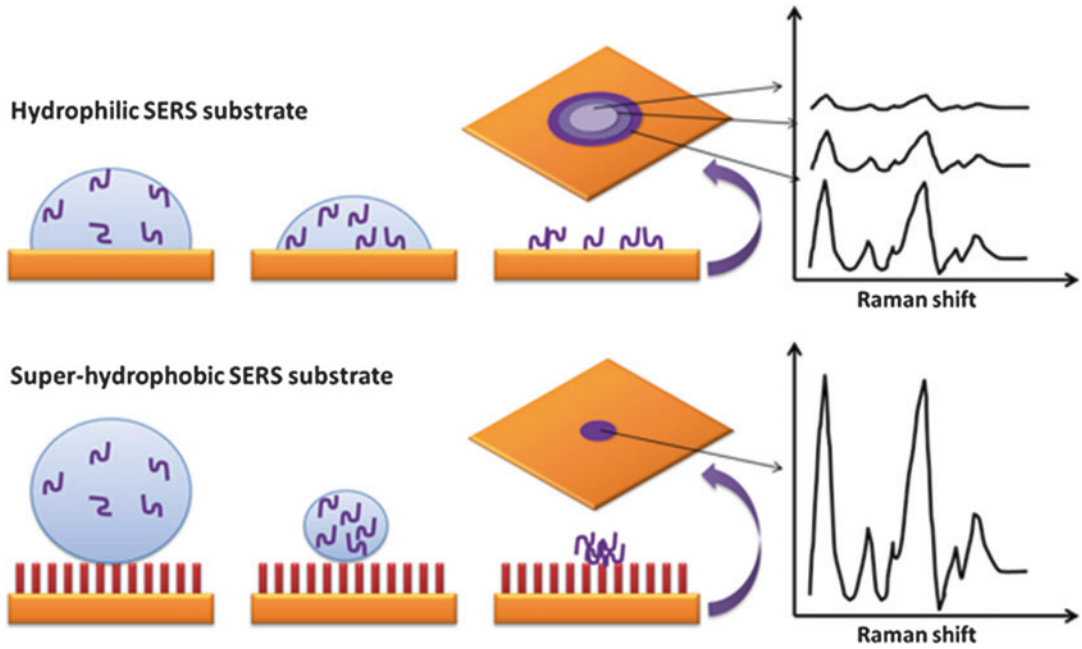
$$OH_{ad}^- + h_{VB}^+ \rightarrow\ OH_{ad}^{\bullet} \quad (3)$$

$$e_{CB}^- + O_2 \rightarrow\ O_2^{\bullet -} \quad (4)$$

$$MB + OH_{ad}^{\bullet} + OH^{\bullet} + O_2^{\bullet -} + h_{VB}^+ \rightarrow \text{Photodegraded products} \quad (5)$$

where TiO_2^* is the excited state of titania, OH_{ad}^- are adsorbed hydroxide ions, e_{CB}^- is a photoexcited electron in the conduction band and h_{VB}^+ the photogenerated hole in the valence band $OH^{\bullet}$, and $O_2^{\bullet -}$ obtained in reactions (2) and (4) are active species highly oxidative, and they can contribute significantly to photocatalytic decomposition.

Nevertheless the fast recombination of charge carriers in the semiconductor alone could reduce the resulting photocatalytic efficiency. To obtain effective molecule photodegradation, the electrons and the holes in semiconductor must be separated to suppress their recombination. The noble metal-decorated metal oxides are more efficient than their uncoated counterpart since for



Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy, Fig. 7 Scheme of the behavior of hydrophilic and superhydrophobic surfaces during analyte droplet evaporation

semiconductor/metal hybrid nanostructures with suitable energy level alignment, electrons can be quickly transferred to metal nanoparticles to achieve charge separation (see Fig. 6).

Superhydrophobic Condensation Effect

Another important property of some metal oxide nanostructures is the superhydrophobicity.

Indeed, thanks to their high surface area, the transition to Cassie–Baxter state could happen dramatically reducing the contact area between the droplet and the MO surface. This phenomenon can be exploited in SERS to achieve more sensitive detection [20].

A survey of SERS substrate literature shows that one of the major reasons for poor Raman signal reproducibility is the nonuniformity of molecule spot profile on the surface. As depicted in Fig. 7, when drop delivery was performed onto a hydrophilic surface, spots often exhibit ringlike structures (the most common are coffee stains).

Exploiting a superhydrophobic surface, the diluted analyte in the droplet could be highly concentrated after the droplet evaporation (Fig. 7). This effect is known as the superhydrophobic condensation and can lead to a further amplification of the SERS signal together with a higher signal reproducibility.

Obviously the signal amplification effect is dependent by the contact angle and the volume of water droplet delivered on the surface. Consequently, a higher contact angle and the suitable drop volume could be tuned to obtain stronger SERS signal.

Cross-References

- ▶ Fundamental Properties of Zinc Oxide Nanowires
- ▶ Nanostructured Materials
- ▶ Nanostructured Materials for Sensing
- ▶ Nanostructures for Photonics
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Optimization of Plasmonic Substrates for SERS Application
- ▶ Plasmonics
- ▶ Production of Gold Nanoparticles
- ▶ Silver (Ag)
- ▶ Surface Plasmon Nanophotonics
- ▶ Surface-Enhanced Raman Scattering for Imaging Biological Cells
- ▶ Synthesis of Gold Nanoparticles
- ▶ Titanium Dioxide

References

1. Tsuzuki, T.: Commercial scale production of inorganic nanoparticles. Int. J. Nanotechnol. **6**, 567–578 (2009)
2. Cao, J., Han, W.Q.: Functional Metal Oxide Nanostructures. In: Janotti, A., Kim, H.C. (eds.) Springer Series in Materials Science, pp. 303–335. Springer, New York (2012)
3. Umar, A., Hahn, Y.B. (eds.): Metal Oxide Nanostructures and Their Applications, vol. 5. American Scientific Publishers, Valencia (2010)
4. Srivastava, A. K. (ed.): Oxide Nanostructures: Growth, Microstructures, and Properties. CRC Press, Boca Raton (2014)
5. Kneipp, K., Moskovits, M., Kneipp, H. (eds.): Surface-Enhanced Raman Scattering: Physics and Applications, vol. 103. Springer, New York (2006)
6. Yamada, H., Yamamoto, Y.: Surface enhanced Raman scattering (SERS) of chemisorbed species on various kinds of metals and semiconductors. Surf. Sci. **134**, 71–90 (1983)
7. Wang, X., Shi, W., She, G., Mu, L.: Surface-enhanced Raman scattering (SERS) on transition metal and semiconductor nanostructures. Phys. Chem. Chem. Phys. **14**, 5891–5901 (2012)
8. Yang, L., Jiang, X., Ruan, W., Zhao, B., Xu, W., Lombardi, J.R.: Observation of enhanced Raman scattering for molecules adsorbed on TiO_2 nanoparticles: charge-transfer contribution. J. Phys. Chem. C **112**, 20095–20098 (2008)
9. Musumeci, A., Gosztola, D., Schiller, T., Dimitrijevic, N.M., Mujica, V., Martin, D., Rajh, T.: SERS of semiconducting nanoparticles (TiO_2 hybrid composites). J. Am. Chem. Soc. **131**, 6040–6041 (2009)
10. Alessandri, I.: Enhancing Raman scattering without plasmons: unprecedented sensitivity achieved by TiO_2 shell-based resonators. J. Am. Chem. Soc. **135**, 5541–5544 (2013)
11. Wen, H.: Surface enhancement of Raman and absorption spectra from cyanine dye D266 adsorbed on ZnO colloids. Mol. Phys. **88**, 281–290 (1996)
12. Wang, Y., Ruan, W., Zhang, J., Yang, B., Xu, W., Zhao, B., Lombardi, J.R.: Direct observation of surface-enhanced Raman scattering in ZnO nanocrystals. J. Raman Spectrosc. **40**, 1072–1077 (2009)
13. Wang, R.C., Lin, H.Y.: Efficient surface enhanced Raman scattering from Cu_2O porous nanowires transformed from CuO nanowires by plasma treatments. Mater. Chem. Phys. **136**, 661–665 (2012)

14. Jiang, L., You, T., Yin, P., Shang, Y., Zhang, D., Guo, L., Yang, S.: Surface-enhanced Raman scattering spectra of adsorbates on Cu_2O nanospheres: charge-transfer and electromagnetic enhancement. Nanoscale **5**, 2784–2789 (2013)
15. Li, X., Chen, G., Yang, L., Jin, Z., Liu, J.: Multifunctional Au-coated TiO_2 nanotube arrays as recyclable SERS substrates for multifold organic pollutants detection. Adv. Funct. Mater. **20**, 2815–2824 (2010)
16. Tang, H., Meng, G., Huang, Q., Zhang, Z., Huang, Z., Zhu, C.: Arrays of cone-shaped ZnO nanorods decorated with Ag nanoparticles as 3D surface-enhanced Raman scattering substrates for rapid detection of trace polychlorinated biphenyls. Adv. Funct. Mater. **22**, 218–224 (2012)
17. Tan, E.Z., Yin, P.G., You, T.T., Wang, H., Guo, L.: Three dimensional design of large-scale TiO_2 nanorods scaffold decorated by silver nanoparticles as SERS sensor for ultrasensitive malachite green detection. ACS Appl. Mater. Inter. **4**, 3432–3437 (2012)
18. Sakano, T., Tanaka, Y., Nishimura, R., Nedyalkov, N.N., Atanasov, P.A., Saiki, T., Obara, M.: Surface enhanced Raman scattering properties using Au-coated ZnO nanorods grown by two-step, off-axis pulsed laser deposition. J. Phys. D: Appl. Phys. **41**, 235304 (2008)
19. Ji, N., Ruan, W., Wang, C., Lu, Z., Zhao, B.: Fabrication of silver decorated anodic aluminum oxide substrate and its optical properties on surface-enhanced Raman scattering and thin film interference. Langmuir **25**, 11869–11873 (2009)
20. Xu, F., Zhang, Y., Sun, Y., Shi, Y., Wen, Z., Li, Z.: Silver nanoparticles coated zinc oxide nanorods array as superhydrophobic substrate for the amplified SERS effect. J. Phys. Chem. C **115**, 9977–9983 (2011)

Micro Air Vehicles (MAVs)

► Insect Flight and Micro Air Vehicles (MAVs)

Micro- and Nanofluidic Devices for Medical Diagnostics

Yitzhak Rosen[1] and Pablo Gurman[2]
[1]Superior NanoBioSystems LLC, Washington, DC, USA
[2]Materials Science Division, Argonne National Laboratory, Argonne, IL, USA

Acronyms

CDRH	Center for devices and radiological Health
DNA	Deoxyribonucleic acid
ELISA	Enzyme-linked immunosorbent assay
FDA	Food and drug administration
IBM	International business machines
MEMS	Micro-electromechanical systems
PCR	Polymerase chain reaction
PDMS	Poly(dimethylsiloxane)
PMMA	Poly(methyl methacrylate)
POC	Point of care

Synonyms

Lab-on-a-chip

Definition

The use of fluid flow in micro- and nanochannels to perform diagnostics of a variety of diseases.

Overview of Micro- and Nanofluidics

Although the science of fluid mechanics at the microscale has been around for a long time, the technology of microfluidics finds its roots during the 1970s, when pioneering work in microfluidic systems was developed at IBM labs for an ink-jet printing device and at Stanford University for the development of a gas chromatograph [1]. Microfluidics is a technology that deals with the manipulation of fluids inside microstructures (usually microchannels) (1 μm = 10^{-6} m), whereas nanofluidics is a technology that deals with the manipulation of fluids in and around nanosized objects (1 nm = 10^{-9} m) [1, 2]. Manipulation of fluids encompasses a number of issues including mixing fluids, controlling their flow profile, and directing the flow to a desired location. Additional requirements might be necessary for specific applications such as medical diagnostics, where manipulation of cells or proteins could demand a number of tasks such as cell lyses, control over biochemical reactions, analyte or cell transport and separation, analyte or cell capturing, and the integration within a single system of miniaturized molecular biology tools such as PCR or electrophoresis. Although a promising technology, the realm of micro- and nanofluidics

faces critical challenges. For example, many physical properties of fluids at the macroscale change radically at the micro- and nanoscale. Moreover, many of these physical phenomena still remain under study at the basic science level. Therefore, understanding fluid behavior at the micro- and nanoscale becomes mandatory in order to develop micro- and nanofluidic devices [1, 2].

Microfluid Mechanics: Harnessing New Physical Properties at the Microscale

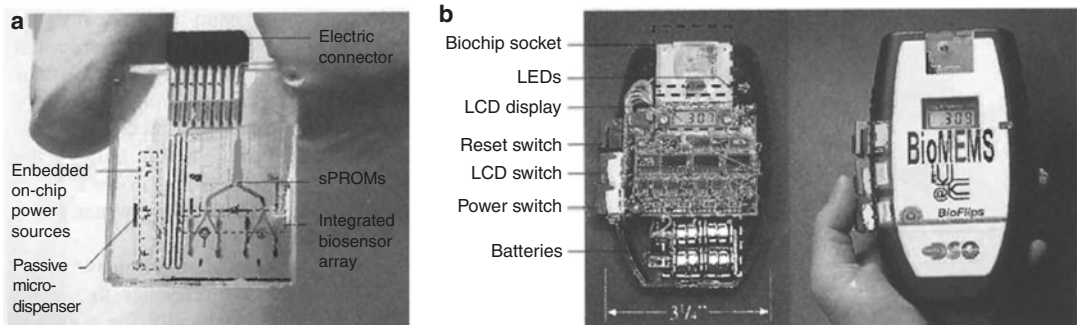
The physics of fluids on microchannels is based on a number of physical effects that appear at the micro- and nanoscale. That is, these are not seen at the macroscopic scale. The Reynolds number has a critical role in this venue. The Reynolds number (Re) relates the viscosity forces with the inertial forces, which determines whether the flow profile behaves as laminar (parallel flow lines that do not mix) or turbulent (a chaotic flow without a specific profile that mix). The Reynolds number involves several fluid properties such as mean velocity of the object relative to the fluid, density of the fluid, traveled length of the fluid, and viscosity. In addition to that, and in order to better understand microfluidic mechanics, other properties such as diffusion, fluid resistance, surface to volume ratio, and surface tension should be taken into account. Diffusion indicates the motion of a particle in a fluid through a concentration gradient. Fluidic resistance is a parameter dependent on the channel geometry. Surface to volume ratio of microstructures such as microchannels increase orders of magnitude against its macro counterparts. Surface tension, the property of water molecules to remain together when exposed to an interface like a gas, increases at the microscale as the surface to volume ratio increases [2].

Nanofluidics: Mimicking Nature

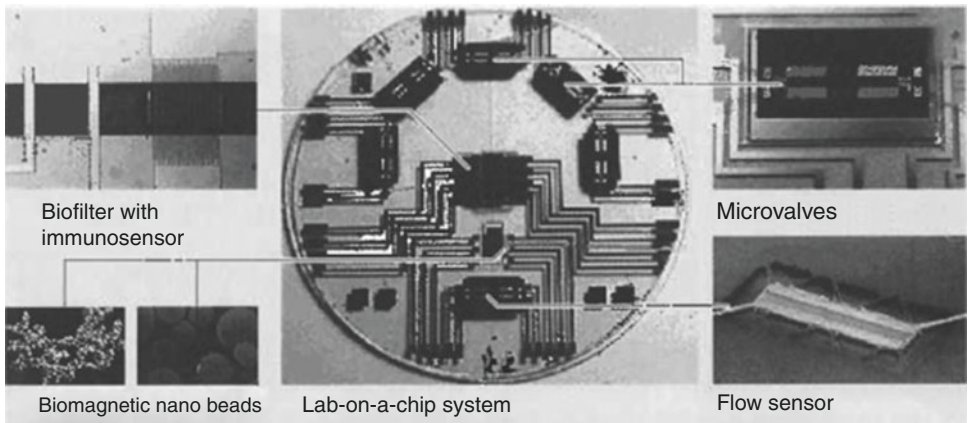
"Nanofluidics" is a term that encompasses a variety of disciplines including biochemistry, biophysics, microengineering, bioengineering, and membrane science. Biology occupies a special place among nanofluidic disciplines since nature has been developing an astonishingly diverse array of nanofluidic systems during the course of evolution. Examples include biomolecular transport inside and across the cells, lateral flow of proteins along the cell membrane, traffic through nanoscale pores from the cytoplasm to the cellular nucleus, axon transport in neurons where nutrients and neurotransmitters flow from the neuron body to the axon terminal, transport of proteins through the complex endoplasmic reticulum, and many others [3].

Nanofluidics deals with several architecture and mechanisms that rely on forces that operate on fluids at the nanoscale. Among them we can mention van der Waals forces (attractive forces that appear when two objects are closer than 2 nm), electrostatic forces that operate at distances between 1 and 100 nm and are either attractive or repulsive, and to a much lesser extent gravitational and inertial forces. As a consequence of the aforementioned, a number of phenomena arising from these nanoscale forces have been described, such as a surface-energy-related phenomenon, a shear-related phenomena, a double electrical layer phenomenon, a size-dependent phenomena, and an entropy-related phenomena. A theoretical approach on the study of nanofluidics has also been developed through the so-called DVLO theory. DVLO theory was developed taking into account the interaction of two objects at the nanoscale by considering van der Waals and electrostatic forces. The DVLO theory accounts to explain the interaction of colloidal systems and has been very successful in this regard. In addition, molecular dynamics (MD) simulations have been developed to study biological nanofluidic systems such as the study of protein transport through nanopores or nanochannels embedded in cell membranes and other biomolecular-nanostructure interactions [2].

Finally, biomimetic approaches where biological nanostructures assist in the development of nanopores and nanochannels are becoming an alternative or complimentary to the above-described techniques. As it was mentioned previously, nature provides an amazing repertoire of nanostructures that provide to be useful for nanofluidic purposes. These structures could be



Micro- and Nanofluidic Devices for Medical Diagnostics, Fig. 1 (**a**) Microfluidic cartridge to be used in an equipment at the clinical laboratory setting and (**b**) handheld microfluidic device suitable for rapid sample analysis of the sample at the bedside of the patient avoiding the need of sample transport (Developed at University of Cincinnati with permission from Bhushan, from Ref. [5])



Micro- and Nanofluidic Devices for Medical Diagnostics, Fig. 2 Microfluidic device designed for an immunoassay test (With permission from Bhushan from Ref. [5])

used directly or as templates for constructing artificial nanochannels or nanopores. Some examples include the use of zeolites found in natural sources for the synthesis of nanopore membranes, or the assembling of protein monomers to create a nanochannel; such is the case of alpha-hemolysin monomers. Other examples include the use of biological motors such as ATPase or kinesins to impulse nanofluids inside nanochannels [3].

The Impact of Micro- and Nanofluidic Devices in Medical Diagnostics

In order to harness the properties of micro- and nanofluidics for medical diagnostics, it becomes necessary to build up a system where several structures such as microchannels, microchambers, and microvalves are integrated with functional parts such as microsensors and microactuators in a single platform or chip to create a device (see Figs. 1a, b and 2), where the fluid containing the patient sample will be injected, transported, and mixed with reagents such as fluorescent labels, or where the sample will be subjected to a variety of process commonly used in the molecular biology laboratory, such as PCR and electrophoresis or biochemical reactions, in order to obtain valuable information about the patient sample.

With that being said, it is now important to understand the implications of such microdevices in the medical diagnostic arena. In medicine,

micro- and nanofluidic devices have found an important niche in the diagnostic area. This is particularly true for mass screening test to detect the presence of absence of a disease at an early stage to allow a prompt therapeutic intervention. Mass screening has a significant role in public health as is shown by the early detection of several types of cancer at an early stage (i.e., breast cancer in women, prostate cancer in men, colorectal cancer), neurological disorders such as Alzheimer disease, and infectious diseases such as AIDS, among the general population, saving the life and improving the quality of life of millions of people worldwide.

From a technological standpoint, it is important to note how many of the current test available at the clinical laboratory setting using the molecular biology toolbox are now starting to harness all the possibilities that micro- and nanofluidics bring about. An example of the later is seen in immunoassays, a kind of test used for diagnosis of several disorders such as AIDS that are now being developed in microfluidic chips (see Fig. 2). It is also important to understand the impact of micro- and nanofluidics in diagnostics from a clinical standpoint. Clinical decision making requires appropriate data in a timely manner. The diagnosis of a disease may require a comprehensive interrogation and clinical examination of the patient. In addition, complimentary information could be obtained from a series of laboratory test and imaging techniques. Laboratory test could be performed in the clinical laboratory setting (in vitro), at the bedside taking a sample from the patient (ex vivo), or inside the body of the patient (in vivo). Laboratory test are standardized according to performance parameters including sensitivity, specificity, predictive value, and accuracy of the test. Sensitivity involves the capacity of a test to detect the presence of the disease (test positive, people with the disease). Specificity is the capacity of a test to detect the absence of the disease (test negative, healthy people). Predictive value is how believable the test result is, namely, if a test result is positive, how believable is that the patient has the disease (positive predictive value), and if the test result is negative, how believable is that the patient does not have the disease (negative predictive value). Finally, accuracy accounts for how the test performs overall, that is to say among all the population tested, which percentage was correctly diagnosed as positive (carrying the disease) or negative (healthy) [4].

Microfluidics and nanofluidics may have a positive impact on sensitivity on medical diagnostics in several ways. First, using small volumes of samples that react with the reporting element in an extremely confined space, such as microchannels, nanochannels, microchambers, and nanopores, the chances of detection of small amounts of the target biomolecule (that indicates the presence of disease, also called biomarker) increase. Microfluidic and nanofluidic devices may also achieve high specificity. A high level of specificity can be achieved when incorporating nanosensors or molecular recognition moieties like antibodies or DNA probes into the micro- or nanochannels that could interact with the biomolecule target (biomarker) with high spatial resolution and affinity, by controlling separation, mixing, and other flow conditions, thus avoiding false unspecific recognition events (if the target biomolecule is not present, the test will be negative since the system will not bind to any other biomolecule than the target) at the sensor–target interface. On the other hand, predictive values are more dependent on epidemiological parameters such as prevalence of the disease and do not depend solely on the technology being used for the test. However, in terms of device performance, as microfluidics and nanofluidics will improve sensitivity and specificity of the test which in turn determines accuracy, new diagnostic devices based on these technologies will outperform current technologies as more accuracy will be obtained.

Potential Applications of Microfluidics and Nanofluidics for Medical Diagnostics

Cancer

A microfluidic device that can capture circulating cancer cells, particularly those with high metastatic ability, has been developed (Cell Search™ Circulating Tumor Cell Kit). The detection of

metastasizing cancer cells could have a prognostic value and perhaps a therapeutic value if early detection can be done [6–8].

Infectious Diseases

Identifying the infectious agent and its response to treatment is of critical importance, particularly with the increasing number of resistant infectious strains. Furthermore, a broad-spectrum treatment is usually given when there is no immediate answer of what the infectious bacterial agent is and its response to treatment. Therefore, when administering broad-spectrum treatment that is not targeted specifically against a specific agent, a possibility of developing resistance to treatment against this potent and important treatment can occur. A portable microfluidic device with high accuracy and rapidity could provide the caregiver an immediate and reliable data output about the infection and its susceptibility to treatment, assisting in early and appropriate therapeutic decisions; furthermore, early and appropriate decisions can safeguard the important broad-spectrum treatment from encountering resistance strains in the future [9].

Neurological Diseases

Earlier detection of neurological disorders such as Alzheimer disease using biomarkers would allow an earlier intervention on the patient, improving his quality of life. In addition, and until an effective therapy becomes available, it could prepare the patient's family to afford the challenge of taking care of his relative more effectively. Future systems may be able to monitor new biomarkers and their clinical therapeutic justification for their early detection [9, 10].

Critical Care

Point of care (POC) devices are of great utility in a critical care unit setting if they can accurately provide important data relevant for the resuscitation, metabolic, and respiratory status of the patient. If they can provide it very rapidly and accurately, then immediate intervention may be warranted for a prompt intervention over critical conditions, improving survival rates. The same concept applies for POC devices in remote settings where healthcare facilities are scarce or in human and natural catastrophes when the healthcare infrastructure is overwhelmed [9, 11].

Neonatology

A point of care device for measuring bilirubin in neonates has been commercialized. The system called Bilichek™ could be used at the bedside and avoids the need for transporting the sample to the central laboratory, allowing a fast decision making. The device works by measuring the levels of bilirubin in a noninvasive manner using optical technology instead of a needle. The device has been developed for newborns wherein hyperbilirubinemia is suspected [12, 13].

Advantages of Microfluidic and Nanofluidic Technologies over Standard Technologies

Some of the advantages of micro- and nanofluidic devices over existing technologies include less reactant and sample consumption (the volume to perform the reaction is limited to a volume of nanoliters, an important requirement in pediatrics wherein the amount of sample could be significantly small), automatization (the technology is integrated with microelectronic circuits that are programmed to perform a sequence of steps), parallelization (several reactions are being run along parallel microchannels and chambers), and sequencing integration (miniaturization allows to add several techniques one after the other in the same device). Lower power consumption, portability, and low cost per unit, as developed through batch microfabrication technologies, are additional assets. As described earlier, different physical phenomena take place at the micro- and nanoscale in fluids limited to micro- and nanochannels. As the size of the channel shrinks, there is an increasing chance of two molecules to meet and react. In diagnostics, this means that a sample containing the analyte would interact with the recognition system such as antibody in less time and thus the result of the diagnostic test will be obtained faster. Moreover, for diagnostic kits employing reactants, a faster mass and heat

transfer along the microstructures means faster reaction kinetics, allowing faster results. Parallelization allows the use of a single patient sample to perform a battery of test such as a metabolic profile decreasing the time for diagnostic results significantly and avoiding the need to puncture the patient several times to obtain a blood sample. Automatization avoids the need of personnel during the running of the diagnostic test protocol, which decreases the human factor error and improves results reliability. Sequencing integration avoids sample transport from one machine to another, making the process less time consuming and decreasing the chance of sample contamination and human error [14].

To better understand how micro- and nanofluidics could outperform current standard diagnostic methods, a few examples will be given. The first example relates to infectious diseases. When suspecting an infectious disease, the physician has to test antimicrobial susceptibility of the pathogen in order to take clinical decisions to initiate an effective antimicrobial therapy. Current methods based on cell culture systems are time consuming, requiring sample transportation and culture of the pathogen, taking a few days to obtain the result. By taking advantage of microfluidic systems, this time will be reduced to only 2 h [15].

The second example refers to cancer diagnostics. Current procedures for performing cancer diagnostics include obtaining a piece of the patient tissue (a biopsy) and its evaluation through microscopic analysis (histopathology). Biopsies are pieces of tissue that are usually obtained invasively from the patient. In addition, histopathology could be a time-consuming process. Microfluidic devices capable of detecting circulating tumor cells in the blood as the one described above will use a blood sample analyzer, avoiding the need to perform biopsies. Moreover, these tests will operate much faster than current histopathology analysis since they will not require time-consuming sample preparation and they will be operator independent, which will make them devoid of human error as opposed to current histological examinations that require an interpretation of the microscopy images, since the microfluidic device will operate automatically detecting very specific markers at the cell surface.

The third example in how microfluidics will make a difference with conventional technologies is the detection of microsatellite regions of chromosomes in malignant tumors using electrophoresis on chip. Microsatellites are sites of DNA with higher rates of mutations, having a prognostic value in several types of cancer such as colorectal cancer. Currently, fluorochrome-based PCR with electrophoresis is one of the techniques used for assessing the presence of mutations in microsatellite regions of chromosomes. This technique requires the use of labels that make the test costly and time consuming. Using miniaturized electrophoresis on a microfluidic chip, it is possible to achieve high sensitivity, rapidity (sample preparation and running time between 30 and 45 min versus 4–6 h in conventional systems), and cost-effectiveness because it does not require using labels and the equipment associated with it [16].

Finally, in neurological disorders such as Alzheimer disease, current methods based on microarray technology require are time consuming and expensive. A microfluidic chip for detection of single DNA mutations (single nucleotide polymorphism or SNP) has been described as a promising tool for early detection of Alzheimer disease [17].

Regulation of Diagnostic Devices

Diagnostics represents a very significant portion of the biomedical market. In order to commercialize such products, it becomes mandatory to obtain regulatory clearance. Within the United States, the Center for Devices and Radiological Health (CDRH) at the Food and Drug Administration (FDA) oversees diagnostic medical devices, controlling their commercialization by ensuring they perform in a safe and reliable manner [18].

Diagnostic devices are classified as in vitro systems or ex vivo or in vivo devices. In vitro diagnostic devices are those located in the clinical laboratory where the sample is transported

and processed. Ex vivo devices or point of care (POC) devices are those used at the bedside of the patient. Finally, in vivo devices are those devices that perform the diagnosis inside the human body such as pressure sensors inside catheters. However, FDA classifies diagnostic devices according to the general category of "in vitro diagnostics." Within this category, it is possible to find home test devices like pregnancy test, laboratory devices used in the clinical laboratory setting by the physician to measure a clinical parameter, and point of care devices such as glucose monitoring systems (ex vivo point of care devices) that are used by the physician at the bedside. Regulation of in vitro systems is based on the intended use of the devices, indication of use of the devices, and, in a major extent, the risk that the device possesses to the patient. In vitro systems, as other medical devices, are classified as class I, class II, and class III. Class I in vitro systems are considered of lower risk. Class II in vitro systems are considered medium risk and are required to comply with standards along with additional data including analytical sensitivity and specificity. Finally, class III in vitro systems are high-risk devices and, in addition to analytical sensitivity and specificity, they must provide clinical validity including clinical sensitivity and specificity of the device when using a clinical sample [17].

To date, microfluidic technology has been developed mostly for point of care diagnostics (commercial examples include the iSTAT™ system and Bilicheck™) and as disposable cartridges for clinical laboratory apparatus systems. Commercial examples of the later include the Agilent 2100 Bioanalyzer™ and the Cepheid Xpert system™ [18].

Cross-References

- ▶ AC Electroosmosis: Basics and Lab-on-a-Chip Applications
- ▶ Integration of Nanostructures Within Microfluidic Devices
- ▶ On-Chip Flow Cytometry
- ▶ Surface-Modified Microfluidics and Nanofluidics

References

1. Gravesen, P., Branejnerg, J., Jenses, O.S.: Microfluidics – a review. Micromech. Microeng. **3**, 168–182 (1993)
2. Eijkel, J.C.T., Van Der Berg, A.: Nanofluidics: what is and what can expect from it? Microfluid. Nanofluid. **1**, 249–267 (2005)
3. Mijatovic, D., Eijkel, J.C.T., Van der Berg, A.: Technologies for nanofluidic systems top down vs bottom up. Lab Chip **5**, 492–500 (2005)
4. Drobatz, K.J.: Measures of accuracy and performance of diagnostic tests. J. Vet. Cardiol. **1**(Suppl. 1), S33–S40 (2009)
5. Ahn, C.H., Choi, J.W.: Microfluidic devices and their applications to lab-on-a-chip. In: Bhusan, B. (ed.) Handbook of Nanotechnology, 3rd edn, pp. 503–530. Springer, New York (2010)
6. Sequist, L.V., Nagrath, S., Toner, M., Haber, D.A., Lynch, T.J.: The CTC chip. An exciting new tool to detect circulating tumor cells in lung cancer patients. J. Thorac. Oncol. **4**, 281–283 (2009)
7. Veridex website. http://www.veridex.com/CellSearch/Products.aspx
8. DeVita, V.T., DeVita Jr., V.T., Lawrence, T.S., Rosenberg, S.A., DePinho, R.A.: DeVita, Hellman, and Rosenberg's: Cancer: Principles and Practice of Oncology, 9th edn. Lippincott Williams & Wilkins, Philadelphia (2011)
9. Fauci, A.S., Braunwald, E., Kasper, D.L., Hauser, S. L., Longo, D.L., Jameson, J.L., Loscalzo, J.: Harrison's Principles of Internal Medicine, 17th edn. McGraw Hill, New York (2008). Various paging
10. Ropper, A., Samuels, M.: Adams and Victor's Principles of Neurology, 9th edn. McGraw Hill, New York (2009). Various paging
11. Chin, C.D., Linder, V., Cia, S.K.: Lab-on-a-chip devices for global health: past studies and future opportunities. Lab Chip **7**, 41–57 (2007)
12. Robert, M., Kliegman, M.D., Bonita, M.D., Stanton, M.D., St Geme, J., Schor, N.M.D., Behrman, R.E.: Nelson Textbook of Pediatrics, 19th edn. Elsevier Saunders, Philadelphia (2011). Various paging
13. Philips website. http://www.healthcare.philips.com/main/products/mother_and_child_care/post_natal_care/well_baby/jaundice_management/bilichek
14. Sia, S.K., Whitesides, G.M.: Microfluidic devices fabricated in poly (dimethylsiloxane). Electrophoresis **24**, 3563–3576 (2003)
15. Chen, C.H., et al.: Rapid antimicrobial susceptibility using high volume to surface ratio in microchannels. Anal. Chem. **82**(3), 1–17 (2010)
16. Odenthal, M., et al.: Analysis of microsatellite instability in colorectal carcinoma by microfluidic-chip based electrophoresis. J. Clin. Pathol. **62**, 850–852 (2009)
17. Devadhasan, J.P., Kim, S., An, J.: Fish-on-a-chip: a sensitive detection microfluidic system for Alzheimer's disease. J. Biomed. Sci. **18**(33), 1–11 (2011)
18. FDA website. http://www.fda.gov

Micro- and Nanomanipulation for Nanomanufacturing

Stefan Johansson
Department of Engineering Sciences, Uppsala University, Uppsala, Sweden

Synonyms

Dip-pen nanolithography; Microrobotics; Nanoprinting; Nanorobotics

Definition

Micro- and nanomanipulation for nanomanufacturing concern the manipulation of micrometer or nanometer sized objects to build nanostructures. In spite of the intuitive impression that the concept in this entry is limited to individual and physical handling of nanoscale objects, also parallel processing with a large degree of self-assembly is normally included in the definition. The manipulation can be made both by physical contact, e.g., electrostatic gripping, and by non-contact techniques, e.g., optical tweezers. The manipulation equipment could be any size while the tool(s) should have a manipulation resolution in the sub-micrometer range.

Essay Disposition

This essay is divided in seven main sections. After a very brief introduction to the "History" of this entry, the "Concepts" used in the text are explained to avoid misunderstandings. There are slight variations of concept definitions depending on the different authors' scientific background and the youth of the research field as well as political and societal factors. The possibilities with the entry "micro/nanomanipulation for nanomanufacturing" are given in the section "Examples of Micro/Nanomanipulation for Nanomanufacturing" while the basic equipment is presented in two different sections: "Micro- and Nanomanipulators" and "Manipulation Tools." The equipment defines and limits the present possibilities for nanomanufacturing. To give a brief understanding of differences between manipulation in the macroscopic and nanoworld, some important "Phenomena" are presented in a separate chapter. As concluding remarks some "Future Trends" are presented.

History

The interest in the atomic world dates back a long time while a first practical approach to enter the nanoworld was given by R. P. Feynman in his talk "There is plenty of room at the bottom" December 1959. He addressed a technological route to manipulate atoms and identified several of the phenomena that occur when we arrive at that size scale. He suggested scaling down, a quarter at a time, our engineering factories and in that way arrive at nanofactories. The possibility to scale down tolerances in parallel with the physical size of the objects/tools/machines is not that straightforward and most efforts towards the nanoworld have followed "shortcut" paths; the nanomanipulating instruments are themselves macroscopic, often decimeter-meter sized while the precision is in the nanometers range or even better.

Nanomanufacturing by manipulation in the broadest sense has occurred naturally in all biological systems since the dawn of biological life. Inorganic nanomanipulation dates back even further, as long as matter has existed, if we include physical and chemical processes with rearrangements at an atomic level in materials. Human beings have been able to manipulate at a nanolevel, even though not in a controlled fashion, since prehistoric times and one example is when one macroscopic object is put in friction contact with another. However, most of us consider nanomanipulation as an artificial and controlled process with spatial resolution in the nanometer range, which excludes these natural and random processes. The possibility to observe nanomanipulation has so far been the key point to get control and the "eyes" into the nanoworld are

the high resolution instruments, for instance Scanning Tunneling Microscopes (STMs) and electron microscopes. The evolution of manipulating techniques in the nanoworld has hence followed the development of the imaging techniques.

It is fair to say that the atomic manipulations began with the experiments at IBM where xenon atoms were manipulated on a Ni surface to form the word "IBM" [1]. These experiments were based on the development of the Scanning Tunneling Microscope (STM) by G. Binnig and H. Rohrer which awarded them the Nobel Prize in Physics 1986 [2]. Still, controlled manipulation at nanoscale has been possible as long as there has been means to obtain spatial resolution at that scale. With the development of scanning electron microscopes (SEMs) in the 1930s it was also possible to locally deposit materials at sub-micrometer scale even though this was rather an unwanted side effect of undesired molecules condensing/reacting on the specimen surface. Electron Beam Lithography in the nano-range by SEMs or similar instruments – making it possible to manipulate atomic bonding in the nano-range – was developed more or less in parallel with the atomic manipulations in the STM. There were also early mechanical "blind" approaches towards the nanoworld – nanomechanical testing with subsequent high resolution imaging – and one example is controlled scribing at the nanoscale during the 1980s.

During the 1990s manipulation experiments with instruments based on the STM gained momentum and also dedicated equipment for micro- and nanomanipulation were developed and presented. The aim with the dedicated nanomanipulators was more or less to provide "hands" in the nanoworld e.g., [3]. Typically these tools were more dexterous than those of the STM and the "eyes" were imaging by SEMs. During the following decade the main research within the controlled inorganic nanomanipulation was based on STM-based instruments and it was not until the launch of dedicated nanotechnology research programs in the beginning of the 2000s that the present bio/micro/nano- technology cross-disciplinary efforts started pacing up.

Concepts

Nanomanufacturing

Manufacturing of artificial structures where the final structure size in at least two dimensions is in the sub-micrometer range is commonly referred to as nanomanufacturing. Artificial is here a key word – biological processes at the nanoscale are not considered as nanomanufacturing. Biochemical synthesis with molecules in the sub-micrometer size range is the borderline of this concept, sometimes included, sometimes not. The final size of the manufactured component is not limiting. Also the fabrication of structures larger than micrometer sizes, by techniques that control sculpturing or assembly at nano-level with spatial definition in at least two dimensions, is considered nanomanufacturing.

MEMS, NEMS and MST

Microelectromechanical systems (MEMS), Nanoelectromechanical systems (NEMS) and Micro Systems Technology (MST) are based on the same concept – the use of processing technology that originates from the microelectronics industry. NEMS indicate that there are spatial dimensions in the sub-micrometer range. Often MEMS/NEMS is used for devices essentially produced with microelectronics processing equipment while processing with dedicated equipment is sometimes considered MST. Still, there is no real difference in technology behind these concepts and hence MEMS will be used synonymously for all three concepts in this essay.

Top-Down and Bottom-Up

The conventional fabrication technology used for Micro Electro Mechanical Systems (MEMS) is considered "top-down" processing since a larger workpiece, typically a silicon wafer, is sculptured by various etching techniques. Nanomanufacturing, on the other hand, is typically a "bottom-up" process since nanometer sized objects, ranging from atoms to sub-micrometer particles, are assembled to create the final product. Still, the concepts of "top-down" and "bottom-up" are not synonymous with MEMS and nanomanufacturing, respectively. Local

micro-deposition to build MEMS would be considered a "bottom-up" technique while parallel manipulation of nanoparticles by imprinting techniques is rather a "top-down" technique.

Micro and Nanorobots

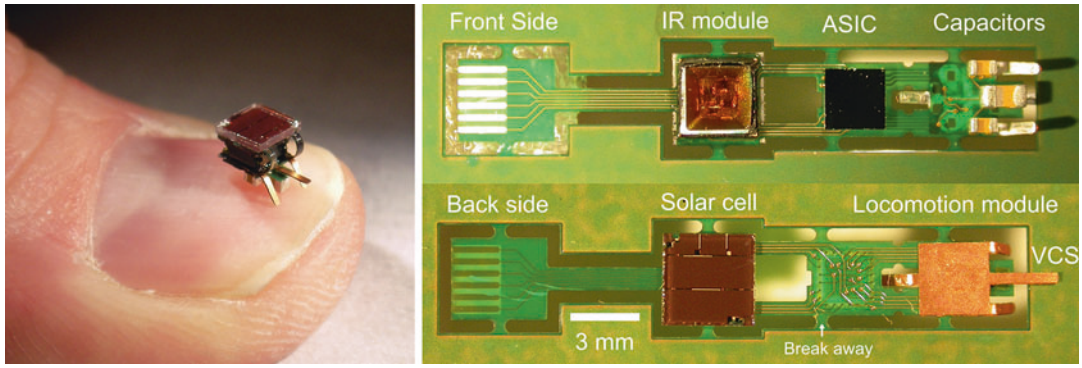
Microrobots are usually defined as robots with a physical size in the micrometer size range, i.e., less than a millimeter. Such robots can be made by slightly modified MEMS technology and microelectronics processing techniques of today (2015). The level of autonomy is not strict; both fully autonomous robots and fully teleoperated robots are included in the concept. A nanorobot is a futuristic idea of a similar artificial unit with similar properties but in a sub-micrometer size range. A modified virus could in many respects be considered an autonomous nanorobot. To distinguish the biologically based molecular components from pure artificial machines the term bionanorobotics has been coined. If we consider present research efforts, the most probable nanorobot that will occur appears to be biologically based and hence these will be considered nanorobots here. Robotic equipment of centimeter or decimeter size with manipulating capacity in the micro- or nanometer size range is here consider as micro- or nanorobotic systems, i.e., robots for micro- or nanomanipulation.

Micro- and Nanomanipulators

Microrobots

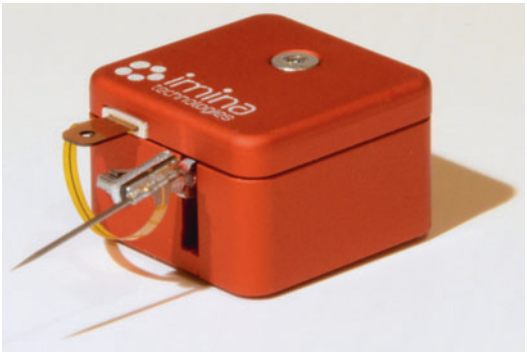
In R. P. Feynmans talk "There is plenty of room at the bottom" 1959 he suggests scaling down our present manufacturing equipment step-by-step, one quarter at a time, to eventually create a factory for nanocomponents. While the interest in micro- and nanotechnology developed in the direction he envisioned the scaling down of equipment followed another route. Even though watchmaking industry has been able to produce mechanical details well below millimeter sizes for decades, the microelectronics processing technology made it possible to go some steps further. In comparison with discrete electronics, the introduction of integrated circuits (IC) made it possible to shrink inter-component distances with at least two orders of magnitude. The size reduction of the IC structures has continued with some 10 % each year since the 1970s. The miniaturization level attained by MEMS technology is rather impressive but the tolerances of the details manufactured have not scaled down in the same manner as the sizes have been reduced. There are about the same tolerances in MEMS devices as high quality fine-mechanics details. The possibility to make scaled down manufacturing plants by MEMS technology is therefore not easily accomplished in the manner presented in R.P. Feynmans talk. To describe it in simple terms – microrobots fabricated by present MST are still too clumsy to manage any sub-micrometer manipulation/processing and are, of course, not able to build any nanofactories.

What type of manipulation can be made with microrobots of today? Figure 1 is an example of a miniaturized robot built with MEMS technology [4]. The size is $4 \times 4 \times 3$ mm and it is made with techniques that would allow for mass-production. The robot is designed to be autonomous but with a possibility to communicate with a master system. There are essentially five different modules that are assembled on a flexible printed circuit board (FPC): a solar cell, an IR communication unit, an Application Specific Integrated Circuit (ASIC), capacitor(s) and a locomotion unit. The solar cell is multifunctional – it can provide the robot with power as well as communication and sensing capabilities. Two small areas are dedicated for position identification and communication. The capacitors are used as energy storage and to stabilize the electrical power. The IR communication consists of four photoemitters and photodiodes directed in four lateral directions. The ASIC is basically a microcontroller with some specific peripherals to manage the other modules. Finally the locomotion unit consists of three legs that create motion by vibrations, tapping the leg tips against the floor. The legs are made of polyimide with a multilayered piezoelectric Poly(Vinylidene DiFluoride-TriFluoroEthylene) P(VDF-TrFE) on top. The carrier is also made of polyimide and the idea was to demonstrate that microactuators can be integrated in the carrier directly (not made in the referred article).



Micro- and Nanomanipulation for Nanomanufacturing, Fig. 1 *Left*: an autonomous microrobot of mm size (4 × 3 × 3 mm) on the nail of a fingertip (With permission from Messe Bremen/HVG Hanseatische Veranstaltungs-GmbH). *Right*: the robot is prepared for mass-fabrication by surface assembly of modules on a flexible printed circuit board and subsequent folding to form the robot [4] (Reprinted with kind permission from IOP Publishing)

Returning to the question – what can be made with microrobots – we should maybe first address the possibility to make a sub-millimeter sized robot with present technology. The answer is yes, we can make them, if we accept the robot intelligence that can be achieved with an IC of sub-millimeter size. Most of the modules in Fig. 1 have substrates which are inactive and can be removed to a large extent. If we add the active thickness of all modules we will be well below 1 mm and there are possibilities to go even further with an ample budget for processing. The next part of the question is what manipulation such a microrobot can do. As part of the locomotion module there is actually a vibrating "tool" which was designed as a touch sensor, to detect when the tool gets in contact with some physical object. We could imagine using something similar to manipulate. Would we also be able to make one tiny hand on this type of robot? The microactuator technology used [5] would not allow for anything complex at a sub-millimeter size range but it is possible to make something more advanced with refined technology. Still, the dexterity of such hands would be extremely limited and we should rather be looking at "grasping" by, for instance, electrostatic forces and releasing by tilting or vibrations. Maybe more interesting is what a large number of microrobots can do together. The tools of each robot might be more or less a needle but the combined action of several robots makes it possible to perform various manipulating tasks. If we imagine some three-four needles that can be controlled arbitrarily in space then most basic manipulating tasks can be accomplished: grasping, moving, rotating, positioning, releasing etc. The crucial points are the precision in movement by the robots, the observation/sensing capacity, the ability of moving the tool in all three dimensions and also their capability to cooperate. Without too much speculation about these engineering challenges, all of them will by themselves limit the handling to the micron range, presently. At this robot size range – sub-millimeter size – there is however nothing in the physics that introduces insurmountable obstacles. It is merely a question of solving technical challenges and the demand for access to an extensive research budget. Considering the 55 years it has taken to reach this immature stage after R. P. Feynmans talk it appears to take another century before we will manipulate the nanoworld with tiny artificial microrobots, though. The cm-sized robot by Imina Technologies, Fig. 2, is the smallest commercial robot with nanometer resolution manipulating capacity that exists today. With its piezodriven motion system it is able to manipulated objects at the nano-level with 4 degrees of freedom (DOF). It is designed for operation in vacuum making it possible to observe the manipulation in situ in a scanning electron microscope (SEM).



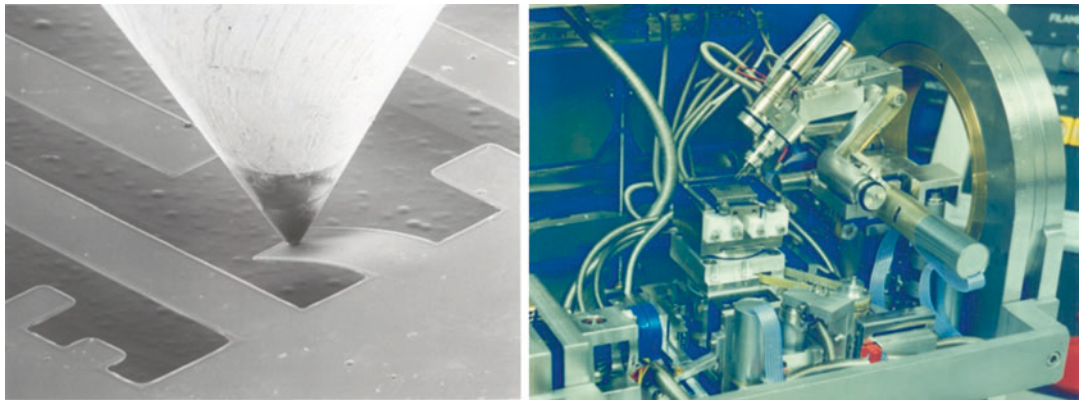
Micro- and Nanomanipulation for Nanomanufacturing, Fig. 2 A cm-sized nanorobotic manipulator from Imina Technologies. The robot has a tool holder allowing up-down movements of the tool and the robot itself can rotate and move in two orthogonal directions (Reprinted with kind permission from Imina Technologies SA)

Using biological molecules, e.g., DNA and proteins, as base elements to build nanorobots appears to be an efficient strategy solving crucial issues like propulsion. These molecules already have highly developed functionality and it would be difficult competing with artificial structures. Some ideas to build nanorobots from these base building blocks are based on nanoassembly. The present state of the art is summarized in [6] and it can be concluded that we are far from building any nanorobots and hence even further from using this type of nanorobots for nanomanufacturing. The question that arises is which path is most efficient for building nanorobots. There is already an extensive ongoing bio-chemical research regarding existing "bionanorobots," viruses, cells etc., and it might prove to be more efficient to (genetically) modify these than building artificial counterparts. A recent example of manipulating with bacteria has been shown in [7]. In the experiments the bacteria are randomly attached to microbeads with 10 μm diameter. The bacteria propulsion is controlled chemically in a digital fashion (on/off). The observed random motion of the microbeads appears to be due to both the arbitrary attachment and random propulsion. The possibility to manipulate microbjects in a controlled and efficient way is obviously far away and there are several critical issues to solve before this will occur.

The magnetically guided microsized object investigated for medical therapy [6] at nano level should maybe be classified as nanorobotic systems. These magnetic particles, without own propulsion, need to be rather large, $>$ 50 μm, to overcome the Brownian motion in a liquid (typically blood), see "Phenomena", with available magnetic field forces. An interesting approach is to synthesize magnetic crystals inside bacteria to guide these self-propelling 1–2 μm "micro/nanorobots" inside a biological body. These guided bacteria are developed to carry therapeutic drugs but we could as well imagine them carrying nanoscale base building blocks. The ability to control nanomanipulation is so far limited and the most obvious control will be by methods that resemble those of conventional (bio-) chemical processes.

Micro- and Nanomanipulators

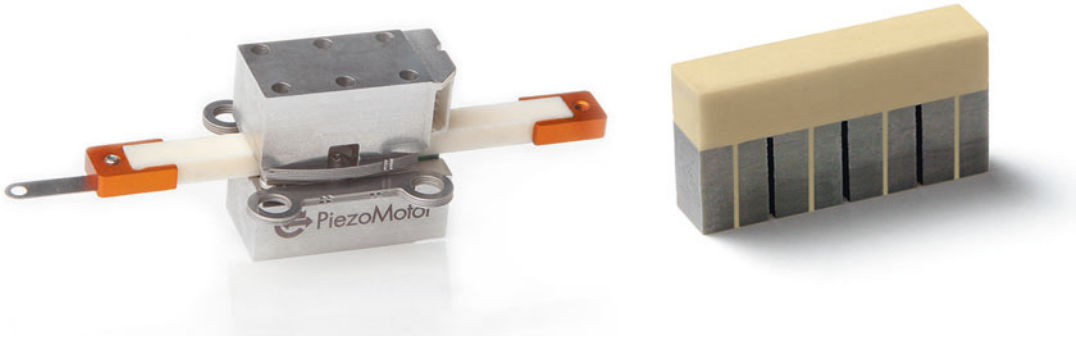
During the 1980s the field of MEMS became an important research field globally. At that time it was called micromechanics and the possibilities to build complex mechanical systems at micron level appeared endless. The two dimensional character of the structures as well as poor relative tolerances proved to be major obstacles. As part of the general characterization of the materials used in MEMS a simple micromanipulator operating in situ in a SEM was used for mechanical characterization of silicon cantilevers, Fig. 3 [8]. These tests were fairly straightforward – a simple spherical diamond tip was used to deflect the silicon elements until fracture, Fig. 3. The fascinating high fracture stress/strain of these close to perfect materials/components could be observed for the first time. The successful use of that equipment with barely μm precision can be attributed to the ability to observe the tool and device at high magnification with a large depth of focus. The SEM gave us eyes in the micro- and nanoworld and most of the mechanical errors as hysteresis, non-linearity and even back-lash can be compensated for with no or little prior knowledge. Most of us that have been working with objects in the sub-millimeter size range have noticed that it is possible to manipulate, possibly with some simple support/leverage, micro-object by the bare hand



Micro- and Nanomanipulation for Nanomanufacturing, Fig. 3 Micromanipulation in a Scanning Electron Microscope (SEM). *Left*: Silicon cantilever microstructures are fracture tested in-situ. *Right*: A nanorobotic system for operation in a SEM with a 4 DOF specimen heating table and a 5 DOF manipulation station using micromachined tweezers as tool

as long as the tool and object can be observed at relevant magnifications, often sufficiently well with a stereo microscope. There were several different micro- and nanomanipulation projects running during the 1980s [9] with focus on different applications, typically medical, micro/nanorobotics and micro/nanomanufacturing. The parallel development of Scanning probe Microscopes (SPMs) for nanomanipulation did not appear to be the desired manipulation equipment since the available 2–3 degrees of freedom (DOF) – x, y and possibly z- movements – in that instrument does not give the flexibility desired for general manipulation tasks. Full flexibility demands at least 6 DOF at the point of observation and rather complex instruments were built to accomplish this. One example of a micro/nanomanipulating equipment with 6 DOF at the observation area in the SEM is given in Fig. 3. The specimen table and the two tool stations were designed for 14 DOF to become, in some sense, the "tiny hands" in R. P. Feynmans talk. The SEM has several advantages in comparison with e.g., SPMs and real-time imaging with large depth-of-focus at high resolution is very attractive. To utilize the highest resolution some contrasting means might be necessary but the main drawback is the need for operation in a vacuum environment. The use of environmental SEMs (E-SEMs) [10] solves this to a certain extent – the need for short distance between sample and lens at higher pressures limits the workspace. The resolution is also somewhat reduced.

The mechanical design in Fig. 3 is based on DC-motors, leadscrews, linear bearings and piezotranslators. The DC motors and mechanical devices are used for the rough positioning while the piezotranslators are used to compensate for back-lash etc. Since the piezomaterials deform smoothly with applied voltage it is possible to achieve nm resolution using a sufficient voltage resolution. The piezotranslators are used to "scan" within the steps of the rough positioning. Piezoactuators are key components in both the SPM manipulators and dedicated manipulating equipment. Present piezomotors, Fig. 4, are able to both make the rough positioning by a "walking" mechanism and nm-positioning by "scanning." The motor core is a unit with four legs, Fig. 4 Right, consisting of piezoelectric bimorphs that individually can make steps, very much like the movement of a four legged animal [11]. The legs are pressed with a fairly high normal force against the drive rod (white rod in Fig. 4) and the frictional contact makes it possible to generate large driving forces. The use of piezomotors has made it possible to build compact manipulating units with high stiffness and nm-resolution. Commercial nanomanipulators are optimized for



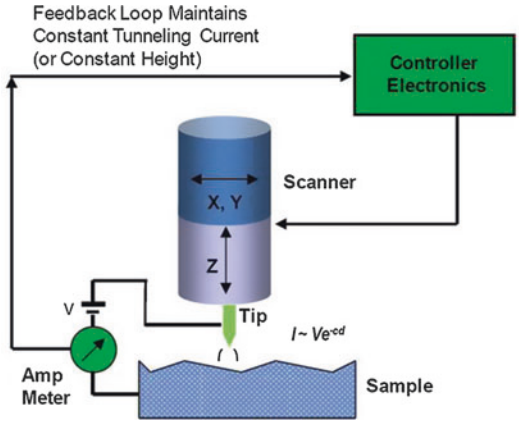
Micro- and Nanomanipulation for Nanomanufacturing, Fig. 4 *Left*: a piezoelectric "walking" motor from Piezomotor® able to drive the central rod (*white*) with sub-nanometer resolution continuously over the whole rod length. *Right*: the motor core is a four legged piezoelectric unit with bimorph legs that can both extend and bend. The legs are operating in pairs and can "walk" continuously while the in-step resolution is mainly limited by the voltage resolution of the drive electronics (Reprinted with kind permission from Piezomotor Uppsala AB)

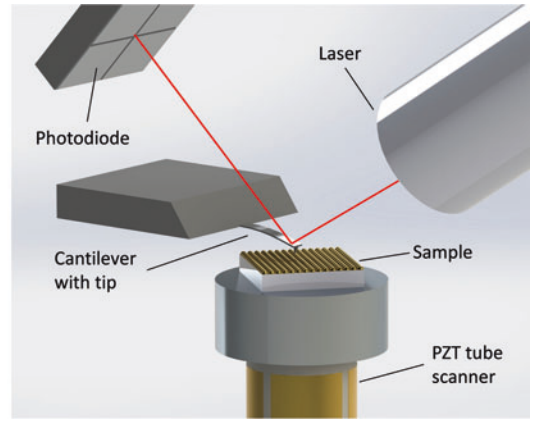
various applications and imaging instruments used. Typically each axis is guided by linear roller bearings and piezomotors or piezotranslators are used to achieve nanometer – or better – movement precision. Optical scales are most often used to get positioning resolution in the nanometer range and feed-back for controlled motion and automatic operation ("► Robot-Based Automation on the Nanoscale"). One drawback is that the speed of a piezomotor is lower than that of an electromagnetic motor, typically by one-two orders of magnitude. While a Hard Disc Drive is able to "manipulate" magnetic particles within an area of 10 × 100 nm every ns the Piezomotor needs to manipulate atoms to keep the same pace. Still, the resolution and the outstanding high speed dynamics of the piezomotors makes it possible to reach the same manipulation frequency at the very bottom-up atomic spatial dimensions. The processing tools with this resolution and speed are yet to be developed.

Scanning Probe Microscopes

The scanning probe microscopes (SPMs) originate from the scanning tunneling microscope (STM) invented by G. Binnig and H. Rohrer in the late 1970s. Initially they wanted to characterize the tunneling behavior locally, at an area less than 100 Å, and later it turned out to be one of the most versatile and accessible high-resolution imaging techniques. The instrument developed was based on piezotranslators arranged to get sub-nanometer resolution in three orthogonal directions. Typically tripods or tubes are used to achieve controlled movements with minimized vibration sensitivity, Fig. 5. The piezoelectric material deforms linearly with supplied charge (more or less linearly with voltage) and the resolution of the voltage source will determine the positioning resolution. A sharp tip – it has to end with a single atom – is positioned by the piezoactuator at a few Ångströms distance from the substrate. In the STM there is no need for a position detector since the instrument uses the tunneling current to measure the surface topography, e.g., operate in constant tunneling current mode and calculate the z-movement (orthogonal to the investigated surface) from the voltage changes in the piezotranslator. This gives a map of the surface with atomic resolution at the correct experimental conditions.

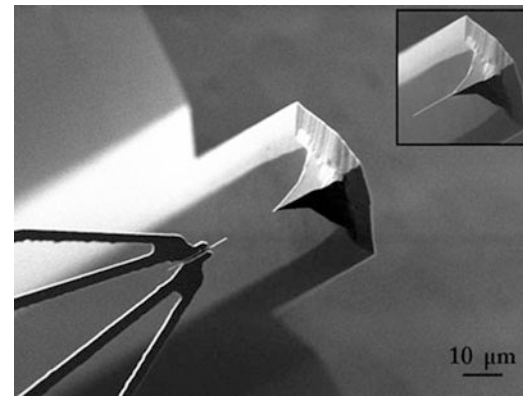
If a position sensor with nanometer resolution of the tip is added to a STM then it can operate as an Atomic Force Microscope (AFM), Fig. 5. Typically a laser beam and split photodiode are used to detect deflection of a cantilever with the probe tip. Apart from being able to image non-conducting substrates several new modes and various detection methods can be used. The tip can be in constant contact or in non-contact with the surface. There is also a so-called tapping mode where the tip is intermittently in contact





Micro- and Nanomanipulation for Nanomanufacturing, Fig. 5 *Left*: scanning tunneling microscope with a piezoelectric tube scanner able to move in three orthogonal directions with sub-nanometer resolution (Reprinted with kind permission from Springer Science and business media [12]). The tunneling current is used for feed-back to keep the atomically sharp tip at a given distance from the surface. *Right*: in an atomic force microscope the tip is mounted on a cantilever beam. The atomic forces between sample and tip will deflect the cantilever and the deflection is detected with a laser beam and split diode

with the surface. Numerous interactions can be imaged, e.g., forces, chemical and optical, and such an instrument is often called Scanning Probe Microscope (SPM). These instruments are well suited for nanometer manipulations on a surface and many nanomanipulation tasks have been explored, see ("▶ Robot-Based Automation on the Nanoscale"). Assuming a layer-by-layer technique to build nanostructures we could also consider this an instrument also for three-dimensional 3D manufacturing. One common drawback is the speed of operation; the time to scan a full image takes some tens of seconds which does not allow for real time observation and manipulation. There has been a development towards video-rate scanning by particular scanner and cantilever solutions.



Micro- and Nanomanipulation for Nanomanufacturing, Fig. 6 A CNT is manipulated with MEMS tweezers and assembled on an AFM tip in situ in an SEM [13] (Reprinted with kind permission from NanoWerk (image Özlem Sardan, DTU))

Manipulation Tools

Physical Probes

The SPMs typically have a probe tip on a cantilever. The conventional probes are made of Si and Si_3N_4 with a tip radius down to a few nm. The cone angle is typically down to some 10°. There are equally sharp tips made of Diamond Like Carbon (DLC) and more blunt polycrystalline diamond coated tips. The tips with highest aspect ratio have a Carbon Nano Tube (CNT) as tip with a radius of a few tens of nm and length of some hundreds of nm, Fig. 6.

There are several examples of miniaturized manipulation tools and the simplest is a sharp needle as in the case of SPMs. These types of tools become more versatile if several can be operated independently and close to each other.

The closeness is often limited by the fairly large manipulators, in comparison with the manipulation space, that are used so far. The nanoparticles/objects are held with electrostatic or interatomic forces and the main concern is often how to detach the particle/object from the tool. Micromachining techniques have been used to fabricate various microtweezers and one example is given in [13]. Tweezers are typically used for releasing/breaking nanostructures from carrier substrates without losing them after removal.

Nanodroplets can be ejected by both physical and non-contact means. Conventional printing techniques can be miniaturized and using nozzles of sub-micrometer size also droplets of sub-micrometer size have been printed on a substrate. Experiments with sub-micrometer capillaries have shown that it is possible to transfer some few hundred nanometer diameter droplets to a substrate by surface tension at physical contact. Using an AFM it is possible to transfer an ink by capillary forces to a substrate by e.g., heating the tip, the so called *dip-pen technology*. The printing line widths are typically 100 nm while enriched particles in the center could have a line width of 10 nm.

Non-contact Probes

Most high resolution probes can be used for non-contact manipulation of atoms or nanoparticles. Focused atom/ion- and electron beams are used both for removal and deposition of material locally with a spatial resolution in the nanometer range. Electromagnetic forces are used in many different ways. Needles are used to create local electric fields that result in manipulation of material and magnetic probes are e.g., used to manipulate the magnetization of nano-areas in present hard discs. Electromagnetic field manipulation is often based on fields close to or within the visible spectrum. Optical tweezers and two-photon processes are two particular techniques.

The technique to manipulate small particles by optical light is called optical-trapping or -tweezers ("► Optical Tweezers") and is based on the radiation pressure that forces particles towards the focused light center. This non-contact tool is particularly well suited for manipulation of biological molecules in aqueous solutions and the typical technique is to attach (biological) molecules to micrometer sized beads. The beads can be displaced with nanometer spatial resolution using e.g., piezotranslators to move the focused laser beam. The applied forces, typically in the range 0,1–100 pN, can be controlled by variations of the optical intensity or trap displacement using calibration constants for the measured displacement at given force. Using several lasers it is possible to make parallel manipulation and with super-resolution fluorescence imaging, single molecule microscopes, molecules can be observed with close to nanometer resolution.

Two-photon excitation microscopy uses a process where two photons of lower energy than needed for fluorescence, e.g., for observation of biological cells, together will cause the excitation of the atoms. One important improvement from a single photon excitation is that the spatial resolution is much better. The same technique is employed to manipulate bonds in photopolymers and both 3D-lithography and printing in the sub-micrometer range can be made. Nanoscribe GmbH has commercialized equipment with spatial resolution down to 200 nm.

Dielectric particles can be manipulated by inhomogeneous electric fields and this method is called dielectrophoresis. The forces are due to polarization effects and gradients in the electric field (in analogy with optical trapping). Using a physical electrical probe in a nanorobotic system the particles can be manipulated with nanometer precision. Typical particles sizes that can be manipulated by this technique are in the range 1–1,000 μm.

Acoustic radiation forces can also be used to manipulate particles in the micrometer or sub-micrometer range. Typically the particles are handled in MEMS based microfluidic channels and the particles are suspended in the liquid. At frequencies above about 10 MHz it is possible to trap or force particles of sub-micrometer size. The need for standing waves, typically by integrated transducers, makes this technique most advantageous for dedicated manipulation tasks in liquids.

Phenomena

In the nanoworld, the nanoparticle mass is low and inertial forces as well as gravitational forces can be neglected. Manipulation of nanoparticles will therefore be rather different from handling of macroscopic objects. As the particles are essentially without inertia a constant manipulation force is needed to move them in a desired direction. The Brownian motion of surrounding molecules will result in random impacts and the necessary manipulating forces have to be relatively large to give a controlled movement. For instance the manipulation force from optical tweezers is too weak to "hold" sub-micrometer sized particles. Most manipulation difficulties might though result from the "stickiness" of the surrounding tools, surfaces and molecules.

Our intuitive macroscopic feeling is that the adhesive force between two objects getting in contact is in most case negligible. This is partly due to the extremely small real contact area with nanometer spacing between two macroscopic objects and partly due to protective surface adsorbates and inactivated surfaces (recombined bonds). If the objects become extremely flat and have nanometer roughness the objects will however stick to each other (assuming that no microscopic particles will be trapped in between), e.g., Johansson gauges, and it might be close to impossible to part them if the surfaces are ultraflat. The situation with microscopic objects is even more particular.

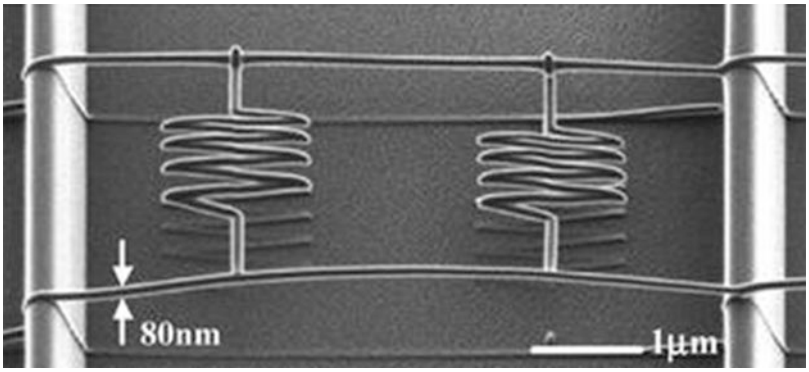
There are three main phenomena that will make small particles stick and an additional number of phenomena that will make them bond to another object. Bonding can be due to chemical reactions and formation of primary bonds between the particle(s) and object. This might be a desired process in a nanofabrication sequence but can normally be avoided for manipulation tasks in conventional environments. Typically an elevated temperature is needed to form primary bonds and this is one of the most straightforward methods to get controlled fixation. Energy supply by tools such as focused electron or ion beams can be used at the local nanoscale.

The sticking in dry environments by electrostatic and Van der Waals forces are essentially impossible to avoid with particles of micron and sub-micron sizes. The Brownian motion makes the particles constantly move and the probability that they will encounter each other – creating agglomerates – or some larger object, is high. Long-range electrostatic forces are not as strong as Van der Waals forces but will still result in strong bonds and become more important when the surface contact area is less. The third sticking phenomenon is wetting and can occur in e.g., humid environments. The adhesive forces by capillary action can be as strong as the two other sticking forces and it might be difficult to remove the particles held by capillary forces even at elevated temperatures.

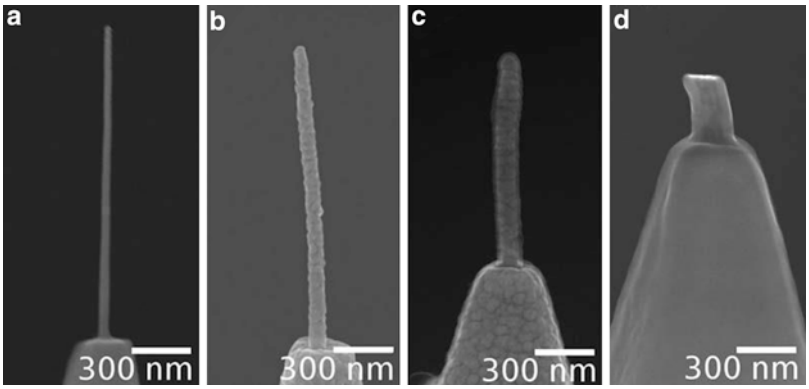
The sticking effects make it difficult to manipulate nanoparticles and even rather large microparticles. The particles stick to the tools and it is normally not necessary to use tweezers to hold the particles but the concern is rather to remove them from the tool or carrier substrate. The techniques to manipulate and in particular to release particles/objects in the nanoworld reminds more about handling objects in the macro world with all fingers and tools covered with sticky glue. In general the contact surface has to be reduced in order to release something and the strategy presented in ▶ Robot-Based Automation on the Nanoscale is illustrative.

Examples of Micro-/Nanomanipulation for Nanomanufacturing

This entry covers processes and methods that still are in an immature state. In an engineering context the word "manipulation" leads our thoughts towards gripping, moving, orienting and assembling nanoparts. The nanoworld is however different from our macroscopic world (see "Phenomena"), and we therefore act accordingly. Nanoparticles should be *guided*, preferably in parallel, to the preferred sites. If we want to assemble, then we want to *arrange* for sufficiently strong interatomic forces, or even bonds, to keep them in place. The tools are typically not based on



Micro- and Nanomanipulation for Nanomanufacturing, Fig. 7 Free-standing diamond like carbon nanostructures deposited by FIB (Reprinted with kind permission from Springer Science and business media [14])



Micro- and Nanomanipulation for Nanomanufacturing, Fig. 8 (**a**) A nanoneedle is grown by EBID on top of an AFM tip. (**b** and **c**) The tip is deposited with Au and Al2O3 and (**d**) the tip is cut with a FIB (Reprinted with kind permission from Comstock et al. [15], Copyright 2012 AIP Publishing LLC)

M

mechanical contact; the physical shape is rather used for non-contact spatial control. Several of the manipulation experiments made here follow this strategy. The examples below present some of the possibilities with present technology.

With non-contact tools as ion- and electron beams, atoms and molecules can be manipulated during simultaneous observation. The focused ion beam (FIB) instrument is one of the more common nanolabratories. This instrument resembles a SEM but with an ion beam source. The ions, usually Ga, can be focused to a spot size of some few nanometers and this beam can be used for both imaging and manipulation of atoms. Impinging ions will remove (sputter) surface atoms and the sample can be sculptured with down to nanometer precision. Atoms can also be deposited locally by the focused ion beam and the process is local activation (Chemical Vapor Deposition) of adsorbed molecules on the surface, Fig. 7. As mentioned in the "History", electron beams have been used to manipulate atoms and molecules since 1930s but mainly in top-down approaches. One example of a bottom-up process is electron beam induced deposition (EBID) which is similar to the FIB deposition process. Adsorbed molecules are decomposed by electrons at the focused electron beam on the surface and subsequently reacting with the underlying surface. The electron beam can be focused to sub-nanometer size and nanometer sized features can be deposited. In the example shown in Fig. 8 the electron beam is used to deposit a needle some tens of nanometer in diameter and the FIB is used to cut and form the tip. The conventional optical lithographic techniques – producing integrated electronics and MEMS – should be mentioned for reference. The optical light manipulates the bonds (make chemical changes) in the photoresist which allows for selective removal of the exposed pattern in a developer (chemical solution). Diffraction limits the smallest feature size and with present deep



Micro- and Nanomanipulation for Nanomanufacturing, Fig. 9 A demonstration of the possibility to manipulate CNTs on a surface with an AFM. The letter θ has been formed of CNTs with a total length of 2.5 μm (Reprinted/adapted with kind permission from Hertel et al. [16]. Copyright (1998) American Chemical Society)

ultraviolet lasers the smallest details are of the order 50 nm, i.e., well into the nanoregion.

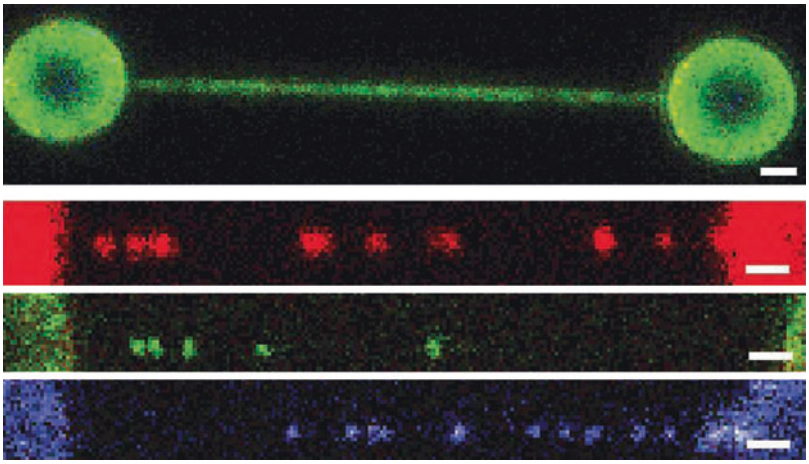
Carbon Nanotubes (CNTs), one atom thick tubes of Carbon atoms, are of great interest for scientific, as well as commercial and technical reasons. Single walled tubes can have diameters less than 10 nm and AFMs have been used to explore the possibility to manipulate these nanoobjects. The adhesive forces between the tubes and surface typically makes the tube bend when forced with the AFM probe. Apart from deforming the CNTs it is possible to translate and rotate the tubes as well as cut them – if they get pinned at two different positions. Figure 9 shows the formation of the letter θ by moving and bending the tubes.

Nanorobotic systems have been used for the assembly of CNTs on AFM tips as illustrated in Fig. 6 [13]. The nanorobotic system consists of commercial units, a 4 DOF manipulator and a 3 DOF nanostage, mounted on the SEM specimen stage. A thermally actuated microgripper (made by MST) mounted on the manipulator is used to grip and release a CNT by sideways movement causing fracture close to the attachment site at the substrate surface (this technique is explained in detail in ▶ Robot-Based Automation on the Nanoscale) The CNT is moved to get in contact with the AFM tip and thereafter attached with electron beam induced deposition (EBID).

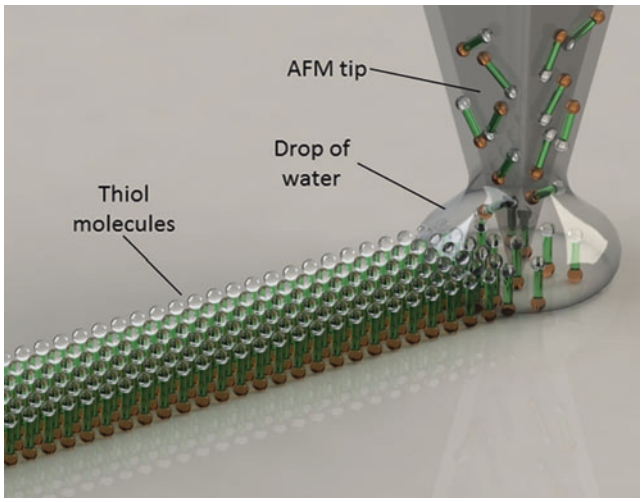
CNTs have also been manipulated by dielectrophoresis and large concentrations of CNTs can be handled in parallel in the fluid channel [17]. The technique is to create parallel electrodes in a fluidic channel and trapped tubes will align with the electric field. This allows for nanofabrication of single-layer arrays but does not allow for free manipulation of individual nanocomponents. Another parallel nanomanipulation technique that has good potential for multi-layered structures is soft- and imprinting lithography [18]. These techniques are normally not considered as manipulation but in practice multiple atoms can be manipulated, e.g., transferred or displaced, with nanometer control. These top-down processes need unique stamps/tools for each pattern, which rather categorizes them as high-volume production than scientific techniques.

Biological molecules are often studied by fluorescence imaging where fluorescent molecules are chemically linked to the molecules of interest, e.g., proteins, making it possible to image single molecules. These molecules can be manipulated by optical tweezers in their natural aqueous environment with real-time imaging. The technique is to attach the molecule to a polymer microbead and use this bead as a handle for holding and moving the molecule. Figure 10 is an example of a manipulation experiment. A DNA is attached to two microbeads and using two optical tweezers the DNA strand can be stretched. Using different fluorescence labels on three proteins the individual positions can be imaged. Improved resolution, down to the order of tens of nanometers, is achieved with superresolution techniques.

One of the techniques that might appear to be the most straightforward way to manufacture at nanoscale is dip-pen nanolithography. Nanoparticles are loaded on an AFM tip and by positioning the tip in proximity to the surface and using a wetting aqueous meniscus (humid environment) the nanoparticles can be deposited on a substrate with nanometer precision.



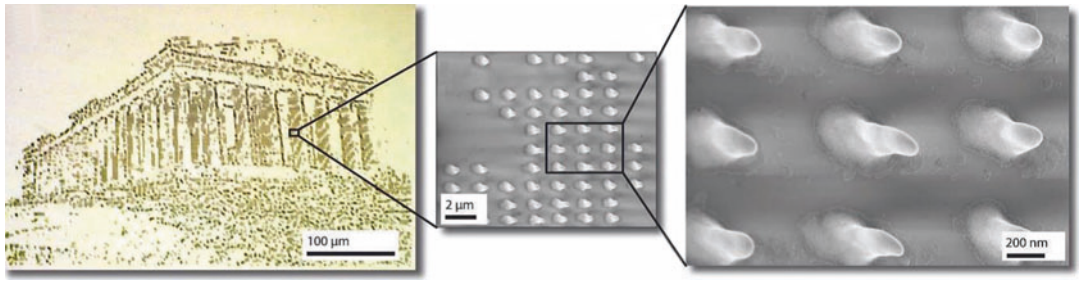
Micro- and Nanomanipulation for Nanomanufacturing, Fig. 10 Stretching of DNA attached to two microbeads (3.2 μm) by optical tweezer. Imaging of fluorescence tagged DNA and proteins by confocal microscopy. Scale bar 1 μm (Reprinted/adapted with kind permission from Macmillan Publishers Ltd: Nature methods [19])



Micro- and Nanomanipulation for Nanomanufacturing, Fig. 11 Dip-pen Nanolithography using an AFM to make a pattern with nanometer resolution. The aqueous meniscus defines the line width of the thiol molecules which form a self-assembling monolayer

The nanoparticles can also be dissolved in a polymer matrix and by heating the tip the liquid ink can be printed on the substrate. It is possible to deposit inks with line widths less than 100 nm and for particular conditions the nanoparticles can be enriched in the center of the ink. By removal of the polymer matrix line widths down to 10 nm has been presented [20]. One drawback with this method is the slow deposition speed. Typically polymer lines can be written at speeds of hundreds of μm/s while the speed for nanoparticles rather is in the order of μm/s. The speed could be increased by using tip arrays, a large number of pens operating in parallel. Arrays with 55,000 pens have been demonstrated and hence five orders of magnitude in speed improvement should be possible. It is possible to tailor the interfacial properties of the substrate within large ranges using inks of alkanethiols which form self-assembling monolayers, Fig. 11. These last two techniques are examples of nanoscale manufacturing in parallel with a large degree of self-assembly.

The last selected examples are nanoparticle printing which are techniques well suited for local and parallel manipulation of nanoparticles. There are several printing methods with sub-micron, or potential sub-micron resolution [21]. Figure 12 shows dielectrophoretic printing of Au nanoparticles using a sub-micrometer sized capillary. The Au particles suspended in a non-polar liquid is dispensed by the electric field gradient at the nozzle opening and are pinned at



Micro- and Nanomanipulation for Nanomanufacturing, Fig. 12 Dielectrophoretic printing of Au nanoparticles using sub-micrometer capillaries. The pillars are deposited with 125 ms printing at each site (Reprinted with kind permission from Schirmer et al. [22]. Copyright Wiley-VCH Verlag GmbH & Co. KGaA)

the surface by Van der Waals forces. Both dots and pillars with diameters of 200 nm have been made. There are also commercial 3D printers with sub-micrometer resolution. Nanoscribe GmbH has developed an equipment for additive fabrication based on a two-photon process which gives a printing resolution of about 0,2 µm. The instrument can handle writing areas of 100 × 100 mm by stitching scanned areas of typically 300 × 300 × 300 µm.

Future Trends

The vision of entering the nanoworld and using nanofactories to create completely new materials, machines and therapeutic devices are receiving more and more attentions as the decades pass and particularly now during the 2010s since dedicated research funding for nanotechnology is available globally. The amount of published material is, as often, not in proportion to the real progress and we are still just touching the surface of this vast and fascinating field. Piezotechnology has been the key for the progress in nanomanipulation and even though there has existed commercial actuators and equipment for more than three decades there are still several improvements to come during the coming decade(s). For instance, sub-nanometer resolution actuators with continuous movements and stable speed over a large working range are still lacking.

Nanomanufacturing was made by biochemical techniques prior to the introduction of the nanotechnology concept and the artificial micro/nanomanipulating tools have so far made little impact. The fast growth of Artificial DNA nanofabrication appears to open up new possibilities. A more extensive cross-fertilization between biotechnology, MEMS/NEMS and other nano-engineering is expected and this will give us more effective "manipulation" means. Among the artificial nanomanipulating techniques both 3D microprinting and microcontact printing might become something that we could consider as generic nanofactories. While microcontact printing is a top-down technique, 3D nanoprinting has a potential to be the general, parallel, bottom-up "manipulation" method for nanomanufacturing. It can be concluded that nanotechnology will be very important since most of the coming challenges in the society, health, environment and energy supply could benefit from nanotechnology, one way or another. The development might possibly not follow the visionary directions, though. As always, the identification of a large value application will accelerate and focus the research efforts. If nanomanipulation will play a central role is yet to be discovered.

Cross-References

- ▶ AFM Probes
- ▶ Dielectrophoresis
- ▶ Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes
- ▶ Dip-Pen Nanolithography
- ▶ DNA Manipulation Based on Nanotweezers

- ▶ Electric-Field-Assisted Deterministic Nanowire Assembly
- ▶ Electron-Beam-Induced Deposition
- ▶ Nanogrippers
- ▶ Optical Tweezers
- ▶ Physics in Laser Near-field Nanomanufacturing: Fundamental Understanding and Novel Probing
- ▶ Robot-Based Automation on the Nanoscale
- ▶ Scanning Electron Microscopy
- ▶ Scanning Tunneling Microscopy
- ▶ Self-Assembly of Nanostructures

References

1. Eigler, D.M., Schweizer, E.K.: Positioning single atoms with a scanneling tunneling microscope. Nature **344**, 524–526 (1990)
2. Binning, R., Rohrer, H., Gerber, C., Weibel, E.: Surface studies by scanning tunneling microscopy. Phys. Rev. Lett. **49**, 57–61 (1982)
3. Hatamura, Y., Nakao, M., Sato, T.: Construction of nano manufacturing world. Microsyst. Technol. **1**, 155–162 (1995)
4. Edqvist, E., Snis, N., Mohr, R.C., Scholz, O., Corradi, P., Gao, J., Dieguez, A., Wyrsch, N., Johansson, S.: Evaluation of building technology for mass producible millimeter-sized robots using flexible printed circuit boards. J. Micromech. Microeng. **19**, 075011 (2009)
5. Edqvist, E., Snis, N., Johansson, S.: Gentle dry etching of P(VDF-TrFE) multilayer micro actuator structures by use of an inductive coupled plasma. J. Micromech. Microeng. **18**, 015007 (2008)
6. Mavroidis, C., Ferreira, A. (eds.): Nanorobotics Current Approaches and Techniques. Springer Science+Business Media, New York (2013). ISBN 978-1-4614-2119-1
7. Behkam, B., Sitti, M.: Bacterial flagella-based propulsion and on/off motion control of microscale objects. Appl. Phys. Lett. **90**, 023902 (2007)
8. Johansson, S., Schweitz, J.-Å., Tenerz, L., Tirén, J.: Fracture testing of silicon microelements in situ in a scanning electron microscope. J. Appl. Phys. **63**, 4799–4803 (1988)
9. Johansson, S.: Micromanipulation for micro- and nanomanufacturing. In: Proceedings of Emerging Technologies and Factory Automation, Paris, France. vol. 3, pp. 3–8. (1995)
10. Nakajima, M., Hirano, T., Kojima, M., Hisamoto, N., Homma, M., Fukuda, T.: Direct nano-injection method by nanoprobe insertion based on E-SEM nanorobotic manipulation under hybrid microscope. In: IEEE International Conference Robotics and Automation, pp 4139–4144. Shanghai, (2011)
11. Bexell, M., Tiensuu, A.-L., Schweitz, J.-Å., Söderkvist, J., Johansson, S.: Characterization of an inchworm prototype motor. Sens. Actuat. A **43**, 322–329 (1994)
12. He, Y., Chen, S., Yu, Q.: Scanning tunneling microscope (STM). In: Wang, Q., Chung, Y. (eds.) Encyclopedia of Tribology. Springer, Berlin/Heidelberg. SpringerReference. www.springerreference.com. 2013-08-15 16:18:43 UTC. (2013)
13. Sardan, O., Eichhorn, V., Petersen, D.H., Fatikow, S., Sigmund, O., Bøggild, P.: Rapid prototyping of nanotube-based devices using topology-optimized microgrippers. Nanotechnology **19**, 495503 (2008)
14. Matsui, S.: Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD). In: Bhushan, B. (ed.) Encyclopedia of Nanotechnology. Springer, Berlin/Heidelberg. SpringerReference. www.springerreference.com. 2012-08-21 14:20:38 UTC. (2012)
15. Comstock, D.J., Elam, J.W., Pellin, M.J., Hersam, M. C.: High aspect ratio nanoneedle probes with an integrated electrode at the tip apex. Rev. Sci. Instrum. **83**, 113704 (2012)
16. Hertel, T., Martel, R., Avouris, P.: Manipulation of individual carbon nanotubes and their interaction with surfaces. J. Phys. Chem. **B102**, 910–915 (1998)
17. Shekhar, S., Stokes, P., Khondaker, S.I.: Ultrahigh density alignment of carbon nanotube arrays by dielectrophoresis. ACS Nano **5**, 1739–1746 (2011)
18. Biswas, A., Bayer, I.S., Biris, A.S., Wang, T., Dervishi, E., Faupel, F.: Advances in top–down and bottom–up surface nanofabrication: techniques, applications & future prospects. Adv. Colloid Interface Sci. **170**, 2–27 (2012)
19. Heller, I., Sitters, G., Broekmans, O.D., Farge, G., Menges, C., Wende, W., Hell, S.W., Peterman, E.J. G., Wuite, G.J.L.: STED nanoscopy combined with optical tweezers reveals protein dynamics on densely covered DNA. Nat. Methods **10**, 910–916 (2013)
20. Lee, W.K., Dai, Z., King, W.P., Sheehan, P.E.: Maskless nanoscale writing of nanoparticle-polymer composites and nanoparticle assemblies using thermal nanoprobes. Nano Lett. **10**, 129–133 (2010)
21. Ru, C., Luo, J., Xie, S., Sun, Y.: A review of non-contact micro- and nano-printing technologies. J. Micromech. Microeng. **24**, 053001 (2014)
22. Schirmer, N.C., Carmen, K., Schmid, M.S., Burg, B. R., Schwamb, T., Poulikakos, D.: On ejecting colloids against capillarity from submicrometer openings: on-demand dielectrophoretic nanoprinting. Adv. Mater. **22**, 4701–4705 (2010)

Micro Coulter

- ▶ On-Chip Flow Cytometry

Micro FACS

- ▶ On-Chip Flow Cytometry

Micro Flow Cytometry

► On-Chip Flow Cytometry

Micro Free-Flow Electrophoresis (μFFE)

Jae-Sung Kwon and Michael T. Bowser
Department of Chemistry, University of Minnesota, Minneapolis, MN, USA

Synonyms

Carrier-free electrophoresis

Definition

Micro free-flow electrophoresis (μFFE) is an electrokinetic technique that can continuously separate and focus charged analytes in the microflow of a carrier buffer where an electric field is applied. A thin sample curtain is introduced into a planar separation chamber with a flowing buffer stream. An external electric field is provided perpendicularly to the buffer flow in the separation chamber. Then the ionic components contained in the sample are laterally deflected along the applied electric field lines depending on their respective electrophoretic mobility and subsequently, collected through the outlets installed at downstream of the chamber.

Introduction

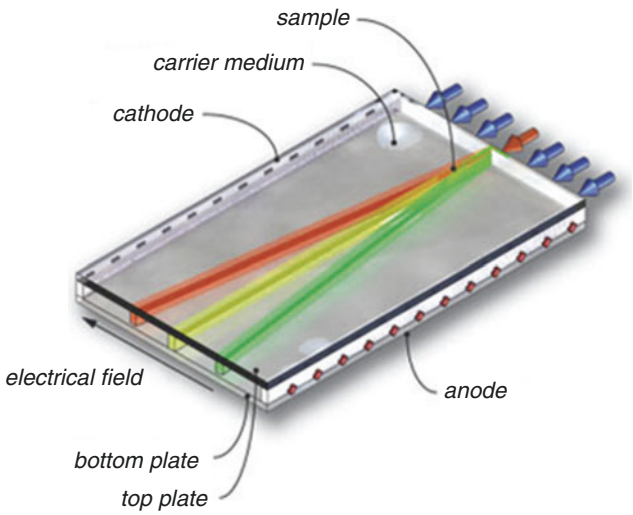
The rapid growth of biotechnology during the past 20 years has constantly emphasized the importance of efficient purification methods for screening and isolating specific constituents from a complex biological mixture. Until now, cell purification has been performed mostly for amino acids, peptides, proteins, nucleic acids, viruses, and bacteria. However the recent emergence of genetically engineered cells enhanced its significance further [1].

Various methods and techniques have been developed for cell separation and isolation [2]. Fluorescence-activated cell sorting (FACS) is the most common method of cell purification. The technique takes advantage of antibodies linked with fluorophores for the purification. The fluorescent antibodies combine with specific surface markers of target cells to be separated. The cells are separated according to the existence and nonexistence of the fluorescently labeled cell surface proteins in the application of an excitation laser and an external electric field. FACS received much attention from scientific communities because of its outstanding efficiency, accuracy, and reliability for cell purification. It led to the extensive applications of the method in the various research fields including tissue engineering and molecular biology. However FACS sometimes suffered from the availability of fluorescent antibodies to tag target cells, the ease of surface marker protein identification, and the viability of cells during FACS process. Most of proteins considered for cell separation are intracellular and hence, they are not easy to be labeled with fluorescent antibodies. If cells to be separated are in a developmental stage, finding their surface marker proteins also is very difficult. Further, since some cells are too sensitive to external stimulations, they may be damaged or killed during the FACS process. Magnetic cell sorting utilizes antibodies tagged with magnetic beads, unlike the FACS using fluorescent antibodies. Thus, the method can easily obtain targets of interest at cheap price. However, there are not many antibodies available for the magnetic cell sorting. Besides, only one kind of protein at a time can be processed by the method. And the accuracy and reliability of the cell purification is not as high as FACS. Cells can also be sorted based on their size (density purification). The introduction of centrifugation media predefined for the sorting into a cell sample rearranges target cells to be separated by forming successive liquid layers with density. It enables to separate a mixed cell population with their sizes through centrifugation at optimized rpm.

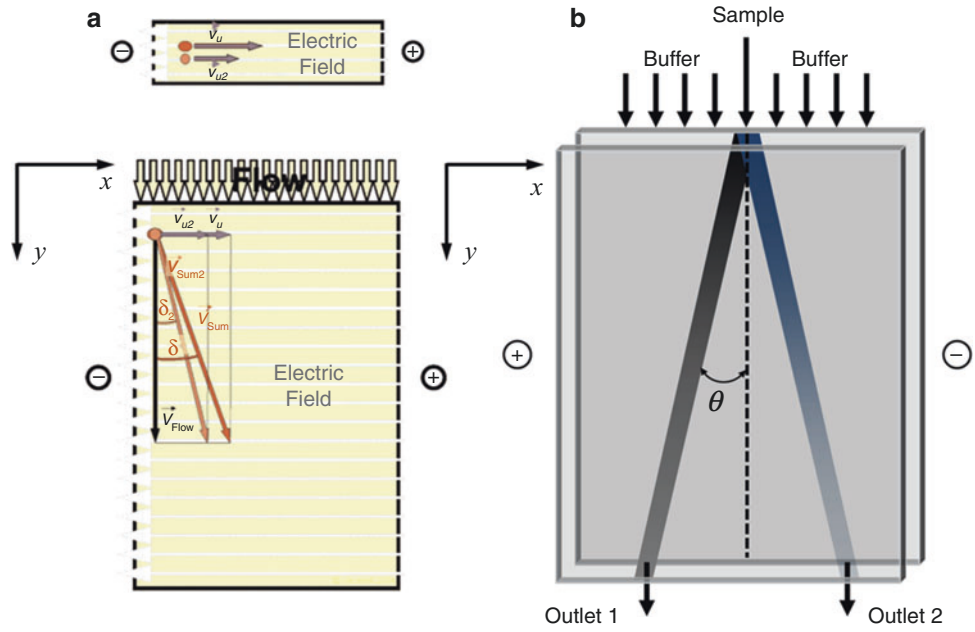
The density-based separation has the advantages of being fast, simple, and cheap. However, the method cannot be used to separate cell populations of similar size. Also it is relatively inefficient compared to other purification techniques. As introduced previously, each of the existing cell purification techniques has several limitations in screening and isolating interesting cells. Those disadvantages motivated the need of a novel separation technique or method that can perform cell purification in the simultaneous mode of ease, accuracy, high speed, and high throughput.

Since its first introduction about 50 years ago, free-flow electrophoresis (FFE) has steadily evolved to one of the most promising techniques for separation of cells, organelles, proteins, and peptides [3–5]. In FFE, analytes are introduced into the pressure-driven flow of a carrier buffer between two parallel plates and then laterally separated along their electrophoretic mobility by an electric field applied perpendicularly to the flow direction (Fig. 1). The separated ionic components are transported with the flowing buffer toward downstream of the FFE device for selective collection at outlets. The greatest advantage of FFE is that a continuous separation and collection of a sample is possible, unlike liquid chromatography and gel electrophoresis requiring repetitive sample loading. It significantly reduces the time and labor consumed for sample separation. The range of electric field intensity used for FFE operation dispels the concern about the electric breakdown of bioanalytes such as electroporation, membrane fusion, bleb formation, and cell lysis [2]. Generally electroporation and the related phenomena are induced by a high-intensity electric field of several kilovolts per centimeter, which is much stronger than the one for a FFE system. This ensures the viability of biomolecules processed by FFE. With the trend of miniaturization and the advancement of microfabrication technology, the FFE has been integrated on lab-on-a-chip (LOC) systems and which, in turn, allowed to find more benefits in some aspects [3, 4]. Compared to macroscale FFE, chip-based FFE can separate and focus analytes within a shorter time. Furthermore, the amount of a sample consumed for analysis also is small (several hundred microliter volumes). High surface area-to-volume ratio minimizes the Joule heat produced by an applied electric field. These advantages have caused wide applications of the micro-FFE (μFFE) for cell separation and isolation.

This manuscript provides a detailed theoretical and technical overview of μFFE so that readers can find more abundant μFFE-based approaches for cell purification through fundamental understanding of the technique. First, a basic principle and separation resolution of μFFE are discussed. It is followed by introducing the various operation modes of the technique and the application cases of each mode.



Micro Free-Flow Electrophoresis (μFFE), Fig. 1 Schematic of μFFE-based cell purification. Reprinted with permission from [3]. Copyright (2008) John Wiley and Sons



Micro Free-Flow Electrophoresis (μFFE), Fig. 2 Principle of μFFE-based separation. (**a**) Similarity of working mechanism between capillary electrophoresis (CE) and μFFE. Reprinted with permission from [6]. Copyright (2006) American Chemical Society. (**b**) Deflection of a sample stream by μFFE in a separation chamber. Reprinted with permission from [5]. Copyright (1994) American Chemical Society

Principle of μFFE

Cell purification by μFFE takes advantage of the difference of electrophoretic mobilities of ionic components contained in a cell sample [3, 4]. The principle is similar to capillary electrophoresis (CE) (Fig. 2a) [2, 6]. Each of cell components has their unique charge-to-size ratio (i.e., electrophoretic mobility) and accordingly, migrates along a capillary at different velocity ($\vec{v}$) under the application of an external electric field ($\vec{E}$). The velocity is given in Eq. 1:

$$\vec{v}_i = \mu_i \vec{E} \tag{1}$$

where μ_i is the electrophoretic mobility of ionic species to be separated and is defined by the size (r_i) and the charge (q_i) of the species as shown in Eq. 2:

$$\mu_i = q_i / 6\pi\eta r_i \tag{2}$$

where η indicates the dynamic viscosity of the suspending medium. The velocity difference of the cell components produces the difference of deformation angles of each component stream in the continuous flow of a carrier buffer in a μFFE system (Fig. 2b). When a cell sample is introduced into a flowing buffer solution at one point through an inlet, it is transported along y-direction with the buffer. While passing across an applied electric field, the sample recedes linearly from the buffer flow direction in x-direction. The deformation angle (θ) is defined as a ratio of the electrophoretic velocity of the sample to the flow velocity of the buffer, like in Eq. 3:

$$\tan\theta = \mu_i I / \kappa A_c v_b \tag{3}$$

where I represents the applied electric current and κ the electrical conductivity of the buffer solution, A_c the cross-section area of the separation chamber in a μFFE device, and v_b indicates the linear velocity of the buffer solution. The electric current and the

electrical conductivity can be assumed to be constant as long as certain ionic materials are not added to the system during μFFE process. Therefore the deformation angle depends only on the electrophoretic mobility of each of cell components to be separated. It makes them move along different flow paths in the separation chamber for separation and selective collection of the components. This is the principle of μFFE-based cell purification.

Band Broadening and Resolution of μFFE System

The most important and ideal thing in cell purification is to obtain desirable target materials from a cell complex in the separation mode of high throughput and high resolution. μFFE system can separate a large amount of cell samples in real time. However the separation resolution is often limited by several physical phenomena which cause the band broadening of a sample introduced [3, 4, 7]. The broadening mechanisms can be understood by discussing those that occur in a macroscale FFE system. It includes the following things: (1) Poiseuille flow profile of a carrier buffer in two fixed parallel plates, (2) Joule heating produced by an applied electric field, (3) electroosmotic flow (EOF) near the walls of a separation chamber, and (4) the difference of electrical conductivity of a cell sample and a buffer solution.

Poiseuille Flow Profile of a Carrier Buffer

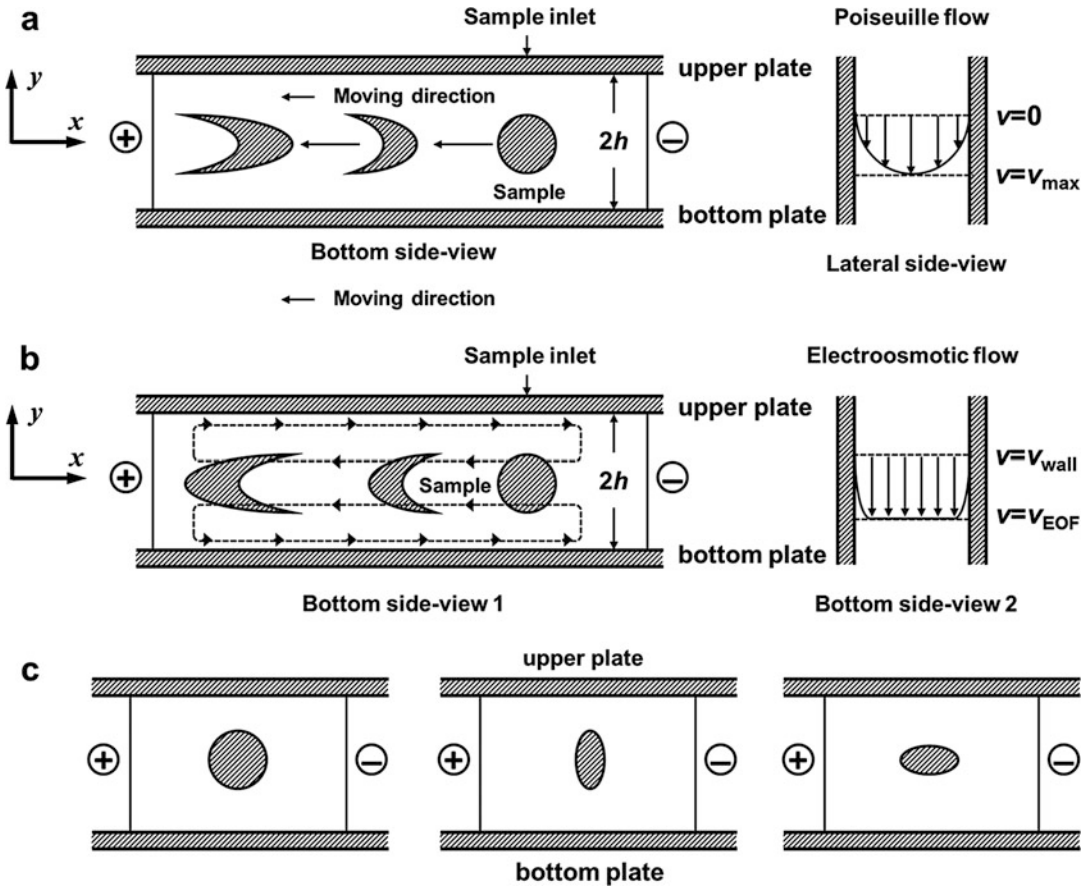
When a carrier buffer flows along a pressure gradient between two parallel plates in a FFE system, the buffer has a Poiseuille velocity profile given by Eq. 4 [5]:

$$v = -(\mathrm{d}p/\mathrm{d}x)\left(h^2/2\eta\right)\left(1 - y^2/h^2\right) \qquad (4)$$

where v represents the velocity of a carrier buffer at a distance y from the middle of the two parallel plates and η the dynamic viscosity of the buffer, $\mathrm{d}p/\mathrm{d}x$ the pressure gradient along the flow direction, and h denotes the half height between the two parallel plates. In the flow profile, the maximum velocity occurs at $y = 0$ which is the middle height of the buffer bed. At the plate surfaces ($y = \pm h$), the flow velocity is essentially zero by no-slip condition. This velocity gradient causes the ionic components in a cell sample to migrate at different velocities according to their location in the buffer. The cell components near $y = 0$ move fast toward downstream of the separation chamber, spending a short residence time in the applied electric field. The cell species around the channel walls move very slowly and accordingly, takes a long time until getting out of the electric field. The difference of the residence times results in the deformation of the sample stream in a FFE system (Fig. 3a). The sample band takes the form of a crescent while flowing downward with the buffer. The sample broadening caused by the Poiseuille velocity profile is referred as hydrodynamic distortion [3–5]. It can be minimized by reducing the diameter of the sample stream or increasing the height of the separation chamber. However those things should be addressed in the viewpoint of throughput and thermal stability of a FFE system. Reducing the sample diameter decreases the amount of a target sample collected by the FFE-based separation. Increasing the height of the separation chamber hinders the dissipation of the Joule heat produced by an applied electric field by increasing the volume of a carrier buffer occupying the chamber.

Electroosmosis (EOF)

Electroosmosis (EOF) occurring near the walls of the separation chamber also distorts the sample stream. EOF is defined as the movement of a fluid driven by an external electric field, which is relative to a stationary and charged solid surface [5]. Most substances acquire surface electric charges through electrostatic interaction when they are in contact with an electrolyte. The charged surface affects the distribution of ions in the electrolyte. While the co-ions of the surface charges are pushed from the surface, the counterions are attracted to the surface. This attraction and repulsion results in the formation of electric double layer (EDL) on the charged surface.



Micro Free-Flow Electrophoresis (μFFE), Fig. 3 Sample band broadening in a μFFE system. (**a**) Hydrodynamic distortion of a sample due to a Poiseuille flow profile of a carrier buffer. (**b**) Electrodynamic distortion of a sample due to electroosmosis (EOF) of a buffer near separation chamber walls. (**c**) Electrohydrodynamic distortion of a sample due to the electrical conductivity difference with a carrier buffer. Reprinted with permission from [5]. Copyright (1994) American Chemical Society

The mobile counter-ions in the EDL migrate toward the counter electrode in the application of an external electric field. Then they drag fluid elements at the same time with their movement due to low Reynolds number of a microflow. It leads to the generation of EOF with a plug-flow velocity profile. In a FFE system, the EOF induces the circulation flow of a carrier buffer. The flow distorts the streams of cell components to be separated (Fig. 3b). This phenomenon is termed electrodynamic distortion. The electrodynamic distortion can be suppressed by coating the surfaces of a separation chamber with non-ionic surfactants like polyethylene oxide [7].

Joule Heating

When an electric current passes through a buffer solution with electrical conductivity, the temperature of the buffer rises by Joule heating effect [8]. Generally the Joule heat can dissipate effectively in FFE systems where a shallow separation chamber is built [5, 7]. However, as the thickness of a buffer bed flowing in the chamber increases, the heat dissipation occurs slowly at the bed center

while it progresses quickly near the chamber walls due to the heat transfer to surrounding atmosphere. This temperature gradient results in natural convection that can distort or completely disturb the sample streams to be separated. The thermal instability of FFE-based cell purification caused by the Joule heat can be minimized by examining several points found in Eq. 5 [2]:

$$J = DAE^2\kappa t \tag{5}$$

where J represents the Joule heat and D the distance between two integrated electrodes, A the surface area of the electrodes, E the strength of an applied electric field, and t means the residence time of a carrier buffer in the separation chamber. The equation implies that the range of the applied electric field intensity can be set to reduce the Joule heat under given working conditions. The electric field strengths below the upper limit ensure thermally stabilized FFE-based cell purification. The stability is also supported by the use of a carrier buffer with low electrical conductivity because the Joule heat is produced proportional to the electrical conductivity of a buffer solution [8]. Short residence time of a buffer in the separation chamber decreases the amount of the Joule heat absorbed into the buffer. Therefore a high-buffer flow rate is advantageous to prevent the overheating of a FFE device.

Difference of Electrical Conductivity Between a Cell Sample and a Buffer Solution

The difference of electrical conductivity of a cell sample and a buffer solution also causes the deformation of sample streams (Fig. 3c) [3–5, 8]. The deformation is called electrohydrodynamic distortion. The electrical conductivity difference results in the gradient of local electric field intensity between the sample and the surrounding buffer. It produces a non-zero net electrical force on the interface of the two. When the sample is less conductive than the buffer, the local electric field strength in the sample is stronger than it in the buffer and accordingly, the net electrical force compresses the stream horizontally. On the contrary, when the sample is more conductive than the buffer, it lengthens the diameter of the sample stream along the width of a separation chamber. The electrohydrodynamic distortion can be eliminated by choosing a carrier buffer with similar electrical conductivity with a sample.

In addition to the above four distortions, the injection width and the diffusion of a cell sample also causes sample broadening. The total deformation of a sample band in a FFE system produced by these factors is mathematically described by the sum of variations of all the broadening sources (Eq. 6) [3, 4, 7]:

$$\sigma_T^2 = \sigma_{INJ}^2 + \sigma_D^2 + \sigma_{HD}^2 + \sigma_{ED}^2 + \sigma_{JH}^2 + \sigma_{EHD}^2 \tag{6}$$

σ_{INJ}^2, σ_D^2, σ_{HD}^2, σ_{ED}^2, σ_{JH}^2, and σ_{EHD}^2 represents the variations related to the injection width, diffusion, hydrodynamic broadening, electrodynamic broadening, Joule heating, and electrohydrodynamic broadening of a cell sample, respectively. However all of these factors need not to be considered for the prediction of sample broadening in a μFFE system. As mentioned previously, some can be eliminated or minimized by coating the separation chamber surfaces with non-ionic surfactants, increasing the buffer flow rate, or choosing the buffer solution with similar electrical conductivity with a sample. Consequently, Eq. 6 is simplified to Eq. 7:

$$\sigma_T^2 = \sigma_{INJ}^2 + \sigma_D^2 + \sigma_{HD}^2 \tag{7}$$

Equation 7 is re-expressed to Eq. 8 by taking the mathematical definitions of each variation into the equation [3, 4, 7]:

$$\sigma_T^2 = \frac{w^2}{12} + \frac{2DL}{v} + \frac{h^2 d^2 v}{105DL} \tag{8}$$

where w is the injection width of the cell sample at an inlet and D the diffusion coefficient, L the length of the separation chamber, v the linear

velocity of the carrier buffer, and h represents the height of the separation chamber. Fonslow et al. suggested the following equation for the separation resolution (R_s) in μFFE through several theoretical findings observed from Eq. 8 and their experimental demonstration [7]:

$$R_s = \frac{(d_1 - d_2)}{(w_1 + w_2)/2} = \frac{(\mu_{total,1} - \mu_{total,2})\mathrm{Et}}{2\left(\sqrt{w_{inj}^2/12 + 2D_1L/v + h^2d_1^2v/105D_1L} + \sqrt{w_{inj}^2/12 + 2D_2L/v + h^2d_2^2v/105D_2L}\right)} \quad (9)$$

For the case that the flow velocity of a carrier buffer is low, the separation resolution is determined by diffusion as in Eq. 10. When the diffusion dominates, the resolution increases with the applied electric field strength and decreases with the square root of the buffer velocity:

$$R_s = \Delta\mu E\sqrt{\frac{L}{32Dv}} \quad (10)$$

For high-buffer velocities, hydrodynamic broadening governs the separation resolution (Eq. 11). The resolution is inversely proportional to the height of the separation chamber and the square root of the buffer velocity:

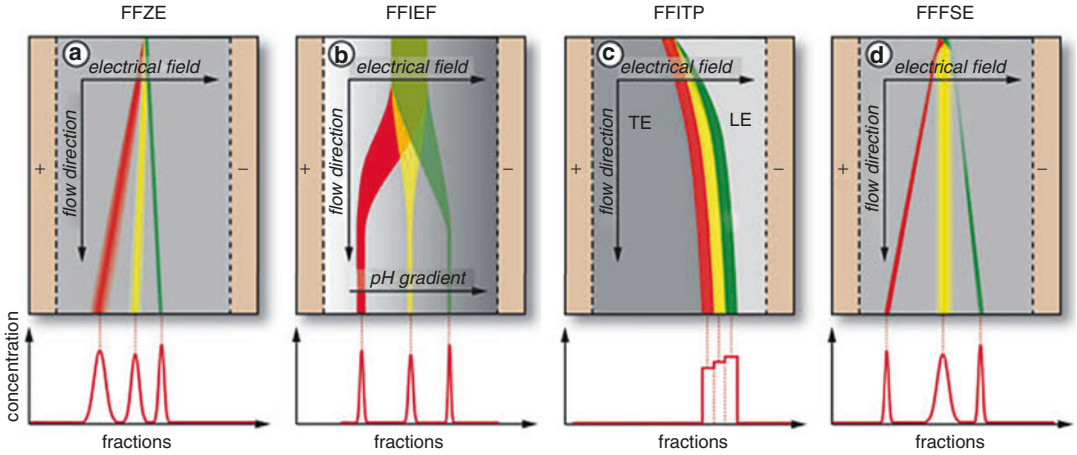
$$R_s = \frac{\Delta\mu}{4h\mu}\sqrt{\frac{105L}{v}} \quad (11)$$

μFFE Modes and their Applications

μFFE has been extensively applied in various fields because of the enormous advantages like high-throughput collection of wanted cell components and online real-time monitoring of separation processes. The applications were made mostly in four operation modes [6, 7, 9–20]: micro free-flow zone electrophoresis (μFFZE), micro free-flow isoelectric focusing (μFFIEF), micro free-flow isotachophoresis (μFFITP), and micro free-flow field-step electrophoresis (μFFFSE).

μFFZE

μFFZE is a method that separates and purifies cell samples in the uniform pH and electrical conductivity distribution of a carrier buffer in a separation chamber (Fig. 4a) [11, 12, 14, 15, 18, 20]. The separation is performed based on the difference of ionic mobilities of the sample components. μFFZE has been often utilized for selective isolation of various biomolecules. Raymond et al. separated lysine, glutamine, and glutamic acid using their microfabricated silicon device for μFFZE (Fig. 5a) [18]. The separation was achieved in 73 s by the applied electric potential of 50 V (I = 7.5 mA). Also they showed experimentally that the μFFZE separation resolution can be improved by varying electrical conductivity of the buffer solution flowing in a side bed in the device. Sodium sulfate Na_2SO_4 was used for the control of the buffer electrical conductivity. The capability of μFFZE was also applied to separate large biomolecules like bradykinin, ribonuclease A, and FITC-labeled human serum albumin in their another study. They sorted the molecules successively within 62 s by increasing applied electric field strength from 25 to 100 V/cm. The peak capacity obtained from the separation was about 8 bands/cm, and it is similar to the peak capacity of commercial FFE systems. Fonslow et al. attempted the separation of fluorescein, rhodamine 110, and rhodamine 123 using an anodic-bonded glass microchip where a μFFE module is integrated (Fig. 5b) [12]. The fluorescent dyes were selectively isolated in 9.6 s by the electric field of 259 V/cm. The separation resolution increased along the strength of the applied electric field. However, electrolysis resulting from a high electric current hindered the continuous performance of the separation. To resolve this

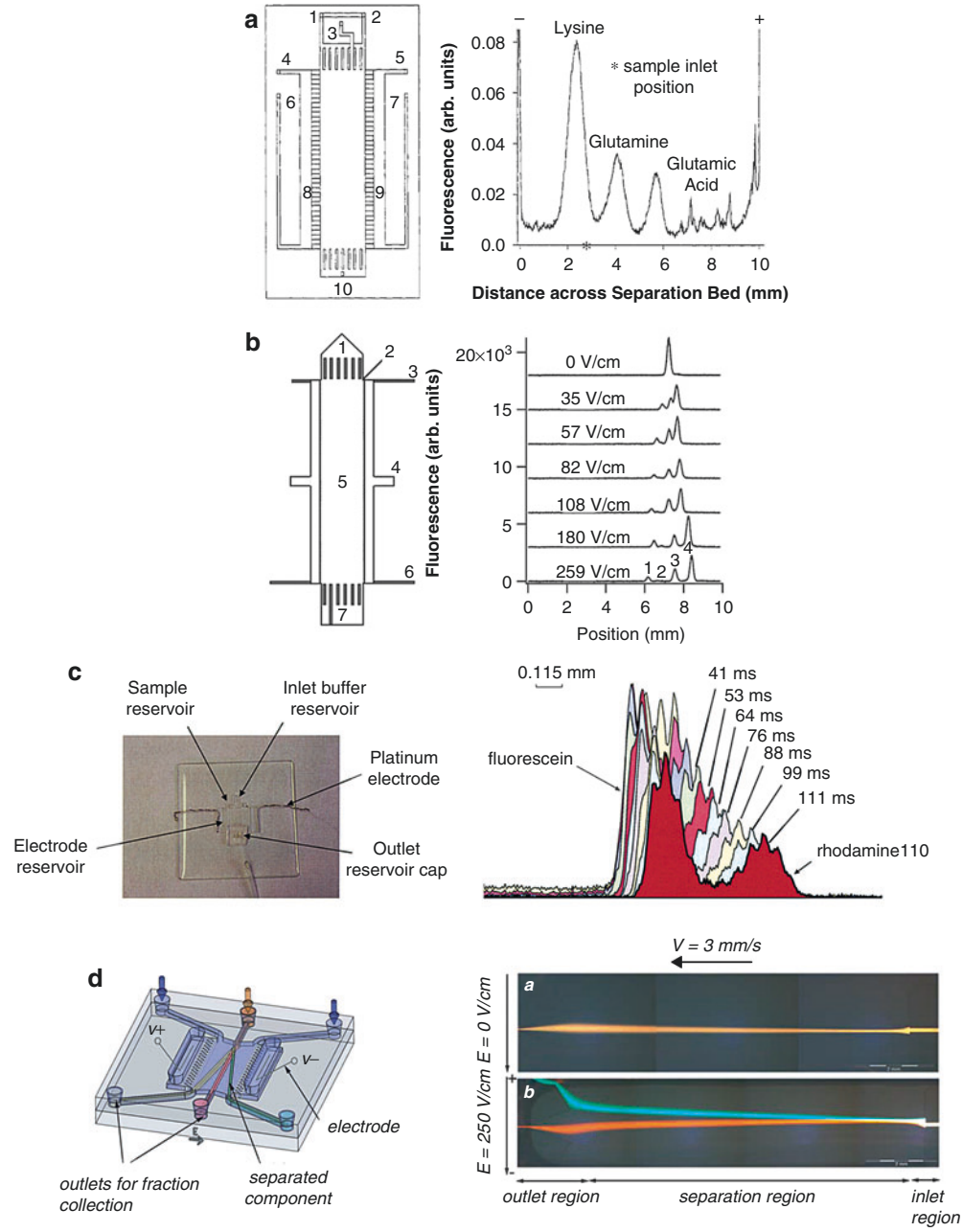


Micro Free-Flow Electrophoresis (μFFE), Fig. 4 Operation modes of μFFE. Reprinted with permission from [3]. Copyright (2008) John Wiley and Sons. (**a**) Micro free-flow zone electrophoresis (μFFZE). (**b**) Micro free-flow isoelectric focusing (μFFIEF). (**c**) Micro free-flow isotachophoresis (μFFITP). (**d**) Micro free-flow field-step electrophoresis (μFFFSE)

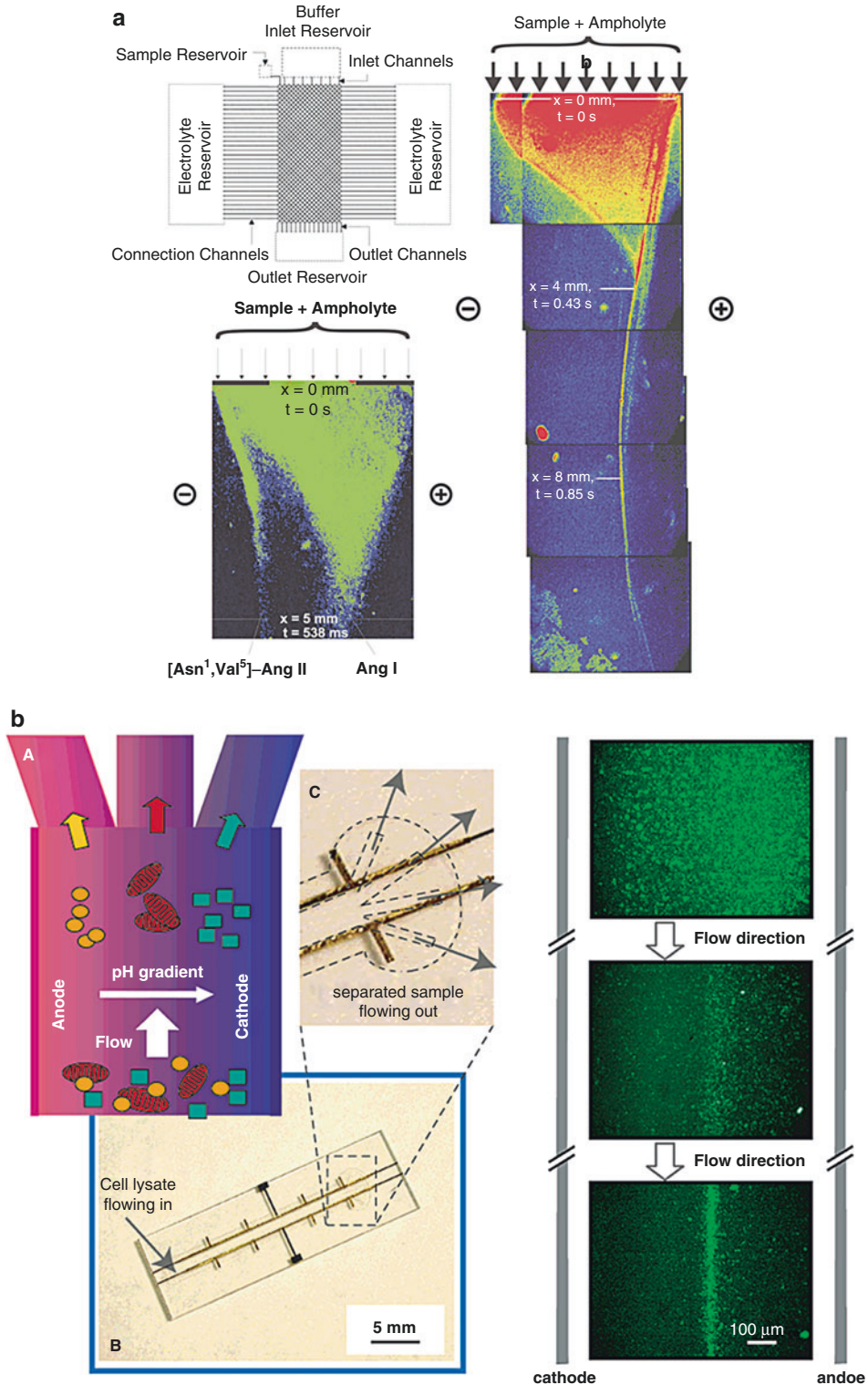
problem, they varied the depth of channels in the chip [11]. While the separation channel has 20 μm in depth, the electrode channel (i.e., side channel) has 78 μm in depth. Using these multiple-depth channels, they could eliminate the electrolysis effectively and consequently perform more stable separation compared to their previous device. Kobayasi et al. separated a mixture of horse heart cytochrome c and sperm myoglobin using the μFFZE mode [14]. The purity of the two proteins separated by 2 kV electric potential within 10 s was estimated in a HPLC system connected with the μFFZE device through a micromodule fraction separator (MFS). Zhang et al. developed a hybrid PDMS/glass-based high-speed μFFZE system where 31104 posts are built in the separation chamber (Fig. 5c) [20]. They isolated fluorescein and rhodamine 110, respectively, using the device. The isolation was completed in about 0.1 s under the electric field strength of 1.75 kV (I = 140 μA). The μFFZE system also was used to separate FITC-labeled amino acids, i.e., Arg-FITC and Gly-FITC. The electric potential of 3.5 kV was applied for the separation. Kohlheyer et al. separated a mixture of fluorescein and rhodamine B in a μFFZE glass chip with an ion-permeable membrane (Fig. 5d) [15]. Each of the separated fluorescent dye streams was continuously collected at the chip outlets by applying 180 V electric potential (I = 50 μA).

μFFIEF

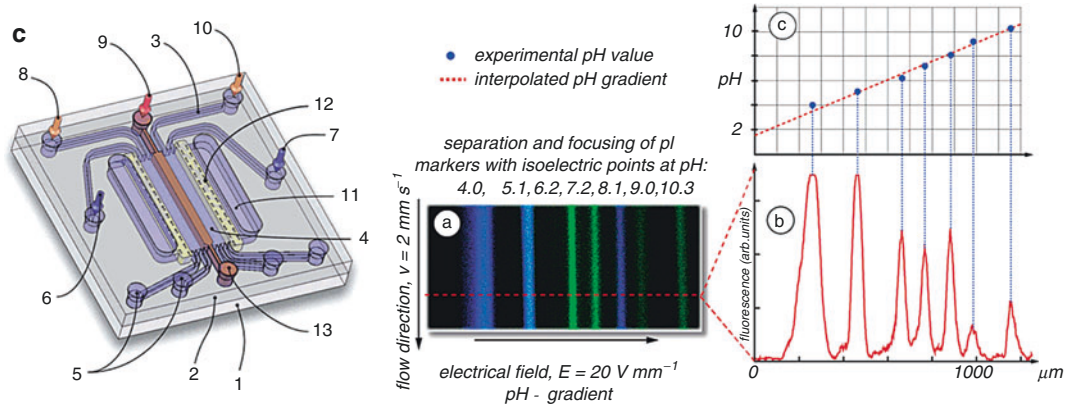
μFFIEF takes advantage of a pH gradient in a carrier buffer for separation and focusing of a cell sample (Fig. 4b) [9, 10, 15–17, 19]. This pH gradient is obtained by adding a mixture of ampholytes to the buffer. The addictive produces a linear pH gradient between an anode and a cathode in an applied electric field. The gradient direction is perpendicular to the buffer flow. In the pH gradient, sample components to be separated migrate by an applied electric field till reaching the buffer region with the same pH with their own pI (isoelectric point). At the arrival points, the components are focused stopping the lateral movement. The broadening of the focused sample streams is constant because they are in an electrochemically stabilized equilibrium state after the separation and focusing. Even though some of the sample components get out of their streams, they immediately return to their initial locations acquiring electric charges again. Therefore, the band broadenings associated with an injection width and diffusion of a cell sample need not to be considered in a μFFIEF system. Xu et al. first introduced μFFIEF-based cell purification by applying the mode to the focusing of



Micro Free-Flow Electrophoresis (μFFE), Fig. 5 Applications of μFFZE. (**a**) Separation of lysine, glutamine, and glutamic acid using a silicon μFFZE device. Reprinted with permission from [18]. Copyright (1994) American Chemical Society. (**b**) Separation of three fluorescent dyes using an anodic-bonded μFFZE device. In the figure, the peaks (1) to (4) represent fluorescein, rhodamine 110 impurity, rhodamine 110, and rhodamine 123, respectively. Reprinted with permission from [12]. Copyright (2005) American Chemical Society. (**c**) High-speed μFFZE-based separation of rhodamine 110 and fluorescein. Reprinted with permission from [20]. Copyright (2003) American Chemical Society. (**d**) Separation and focusing of fluorescein and rhodamine B in 10 mM Hepes solution using a μFFZE glass device with ion-permeable membranes. Reproduced from [15] with permission of The Royal Society of Chemistry



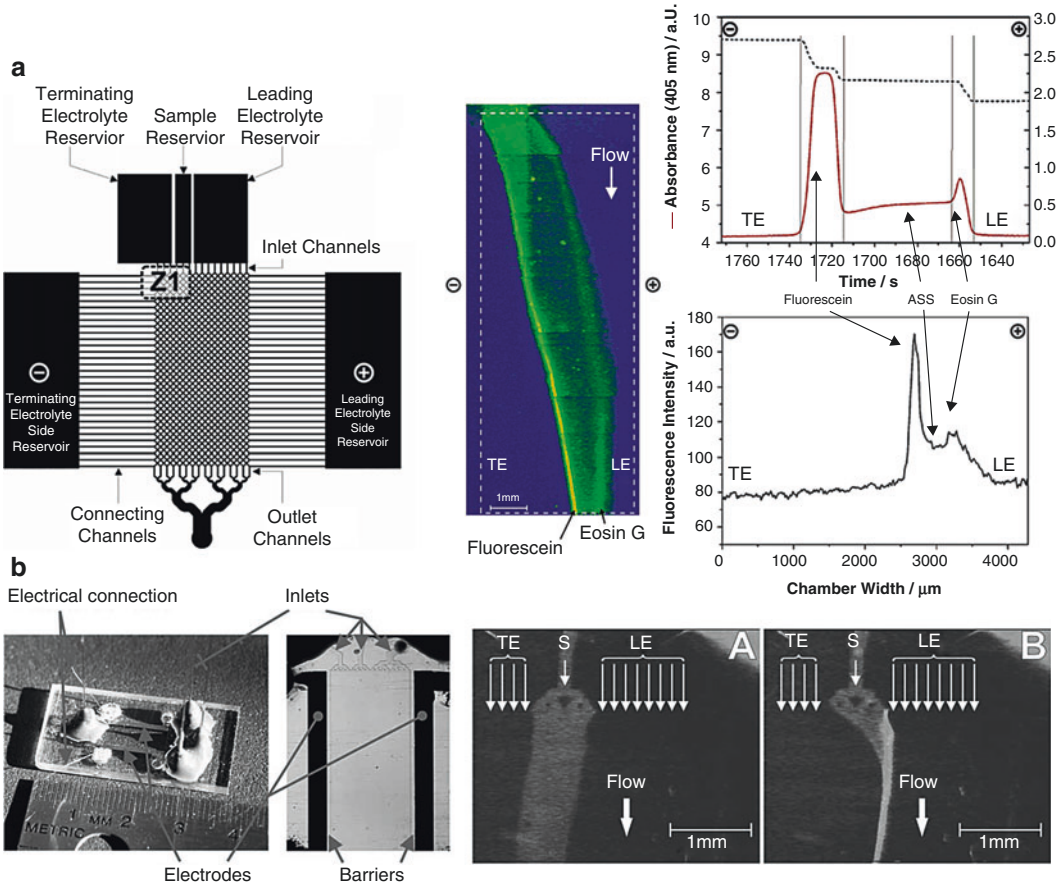
Micro Free-Flow Electrophoresis (μFFE), Fig. 6 (continued)



Micro Free-Flow Electrophoresis (μFFE), Fig. 6 Applications of μFFIEF. (**a**) Isoelectric focusing of [Asn1, Val5]-angiotensin II (*right*) and a mixture of angiotensin I and [Asn1, Val5]-angiotensin II (*left* lower end) in a μFFIEF device filled with an ampholyte solution. Reproduced from [19] with permission of The Royal Society of Chemistry. (**b**) Isoelectric focusing of a mitochondrial fraction purified from lysate of HT-29 cell. Reprinted with permission from [17]. Copyright (2004) American Chemical Society. (**c**) Separation and focusing of seven IEF markers. The IEF markers have their pIs at 4.0, 5.1, 6.2, 7.2, 8.1, 9.0, and 10.3, respectively. The device used for the experiment consists of (1) bottom chip plate, (2) top chip plate, (3) microfluidic channel, (4) separation chamber, (5) outlets, (6) low-pH sheath flow inlet, (7) high-pH sheath flow inlet, (8) ampholytes 1, (9) ampholytes 2 and sample inlet, (10) ampholytes 3 inlet, (11) electrode compartment, (12) conductive membrane, and (13) not separated sample. Reprinted with permission from [16]. Copyright (2007) American Chemical Society

[Asn1, Val5]-angiotensin II peptide labeled with 5-CR 6G (Fig. 6a) [19]. They concentrated the peptide sample 400 times through the application of 135 V/cm electric field strength and 10 μL/min flow rate. The concentration was achieved within 430 ms much faster than conventional FFIEF systems. Also they successfully separated and focused a mixture of angiotensin I and [Asn1, Val5]-angiotensin II under the same experimental conditions. Lu et al. attempted the continuous separation of subcellular organelles using μFFIEF (Fig. 6b) [17]. They focused and concentrated mitochondrial fraction purified from lysate of HT-29 cell in their IEF device consisting of electroplated electrodes and microchannels. It was completed in about 6 min by the low electric potential of 2 V. Analogous experiments were performed for not only the selective isolation and focusing of mitochondria from whole cells and nuclei but also the separation of two different mitochondrial subpopulations. Their studies were very successful in the aspect of minimizing Joule heating effect and electrolysis because low electric potential was used. However it took a long time to obtain a stable focusing of the samples. Kohlheyer et al. focused four IEF markers with pIs between pH 4 and 9 by means of the same FFE device which they had used in their previous study (μFFE glass chip with an ion-permeable membrane) [15]. The markers with their pIs at pH 4.5, 5.5, 7.6, and 8.7, respectively, were separated successively by increasing the applied electric potential from 0 to 20 V. The separation was achieved in about 3 s in the carrier ampholytes with a 500 μm width pH gradient (pH 3 ~ 10). In their another study, Kohlheyer et al. attempted to improve the separation efficiency of the μFFIEF system through the use of multiple sheath flows and pretreated ampholytes (Fig. 6c) [16]. As a result, the rapid linear pH gradient of the ampholytes buffer from pH 2.5 to 11.5 could be produced between 1.2 mm and 2 mm of the separation chamber width. In the gradient, seven IEF markers were successfully separated and focused within 2.5 s through the application of 20 V/mm electric field strength. The experiment demonstrated that their improved μFFIEF device can separate analytes with a minimum pI difference

of 0.4. However further theoretical investigation showed that the resolution can be lowered to 0.23. The value corresponds to the peak capacity of 29 peaks within 1.8 mm width. This is eight times larger than the peak capacities other researchers obtained. Albrecht and Jensen developed a μFFIEF device with a miniaturized thermoelectric cooler and porous conductive gel bridges in a focusing chamber [10]. The gel structures built by functionalized polyacrylamide not only allow ion exchange between the focusing chamber and the electrode buffer but also prevent gaseous bubbles produced by electrolysis from interrupting μFFIEF-based separation and focusing process. The thermal cooling system mitigates the Joule heating effect in the focusing chamber. The μFFIEF device was used to separate and focus a mixture of low molecular weighted pI markers and protein G-mouse IgG. Based on the results obtained from the experiment, Albrecht et al. developed a cascaded μFFIEF system with improved focusing speed and separation resolution [9]. The device performs the separation of analytes in two stages. The analytes are first processed by the steep pH gradient with low separation resolution in the first focusing chamber. The primarily separated components are processed once again by the shallow pH gradient enabling high-resolution separation in the second focusing chamber.



Micro Free-Flow Electrophoresis (μFFE), Fig. 7 Applications of μFFITP. (**a**) Separation and concentration of fluorescein, acetylsalicylic acid (ASS), and eosin G using a PDMS-based μFFITP chip. Reprinted with permission from [6]. Copyright (2006) American Chemical Society. (**b**) Isotachophoretic focusing of a fluorescein in a μFFITP device where electrostatic induction barriers are integrated. Reproduced from [13] with permission of The Royal Society of Chemistry

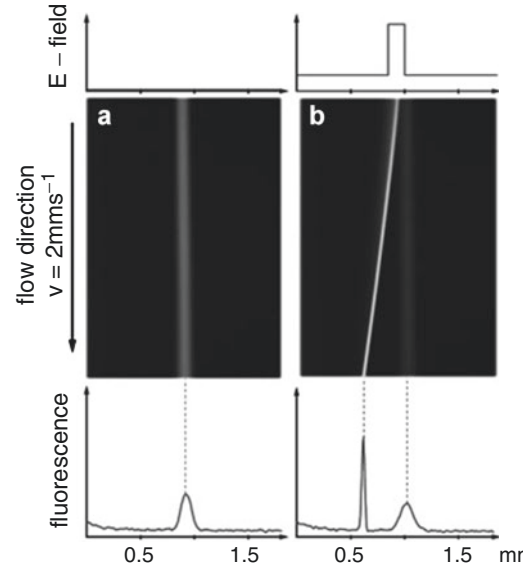
This device was used to separate and focus native model proteins, denatured proteins, and protein complexes.

μFFITP

μFFITP utilizes three buffer regions with distinct ion mobility difference for separation of analytes: leading buffer zone, terminating buffer zone, and sample zone (Fig. 4c) [3–6, 13]. While the leading zone is filled with electrolyte with higher ion mobility than analytes, the terminating buffer zone is composed of lower ion mobility electrolyte. When a sample is introduced between the two regions and an external electric field is applied to the μFFITP system, the streams of ionic components in the sample are aligned adjacent to each other along their ion mobilities and collected individually at the downstream outlets by a continuous flow of the buffers. Janasek et al. first applied μFFITP for separation of a sample mixture consisting of fluorescein, eosin G, and acetylsalicylic acid (Fig. 7a) [6]. The components were separated and concentrated within 1 min under the application of 210 V/cm electric field strength. Additionally they demonstrated the potential of μFFITP for online sample preparation through isotachophoretic separation of myoglobin and fluorescein isothiocyanate. However their device had several problems like electrolysis and low transfer efficiency of an applied electric potential into a separation chamber. To complement the defects, Janasek et al. integrated insulating barriers between the separation chamber and the electrode reservoirs (Fig. 7b) [13]. The improved μFFITP device focused a fluorescein sample successfully with the application of low electric field intensity of 18 V/mm.

μFFFSE

μFFFSE-based separation takes advantage of the difference of the local electric field intensities in a focusing chamber, which are determined by electrical conductivity of carrier buffers in the application of an external electric field (Fig. 4d) [3–5]. The focusing chamber is filled with a buffer with low electrical conductivity, being surrounded by a high electrical conductivity buffer. It produces a step gradient of local electric field intensity in the focusing chamber when applying an external electric field. Accordingly analytes introduced into the chamber exhibit sluggish movement, when they enter a weak local electric field region while migrating by the external electric field. This mechanism enables the concentration and focusing of a sample by μFFFSE. The application of μFFFSE is found in the review paper published by Kohlheyer et al. (Fig. 8) [3].



Micro Free-Flow Electrophoresis (μFFE), Fig. 8 μFFFSE-based separation and focusing of rhodamine B and fluorescein. Reprinted with permission from [3]. Copyright (2008) John Wiley and Sons

Conclusion

This manuscript was designed to introduce a theoretical and technical overview of μFFE and its applications. μFFE is a promising technique that can continuously separate and focus a target analyte and detect the components in real time. This capability has been steadily used for selective isolation of biomolecules such as amino acids, peptides, and proteins. From it, μFFE established its firm position as a powerful tool for high-speed and high-throughput cell purification. The development of the related theory and technology made until recently enhanced the ability and potential of μFFE further, resolving several primary problems

inherent in the technique. It will help to find abundant μFFE applications in disparate research and industrial fields as well as biotechnology fields.

Cross-References

- ▶ Lab-on-a-Chip
- ▶ Micro Free-Flow Electrophoresis (μFFE)

References

1. Prakash, S., Chang, T.M.S.: Microencapsulated genetically engineered live *E. coli* DH5 cells administered orally to maintain normal plasma urea level in uremic rats. Nat. Med. **2**, 883–887 (1996)
2. Amos, P.J., Bozkulak, E.C., Qyang, Y.: Methods of cell purification: a critical juncture for laboratory research and translational science. Cells Tissues Organs **195**, 26–40 (2011)
3. Kohlheyer, D., Eijkel, J.C.T., van den Berg, A., Schasfoort, R.B.M.: Miniaturizing free-flow electrophoresis – a critical review. Electrophoresis **29**, 977–993 (2008)
4. Turgeon, R.T., Bowser, M.T.: Micro free-flow electrophoresis: theory and applications. Anal. Bioanal. Chem. **394**, 187–198 (2009)
5. Roman, M.C., Brown, P.R.: Free-flow electrophoresis as a preparative separation technique. Anal. Chem. **66**, 86A–94A (1994)
6. Janasek, D., Schilling, M., Franzke, J., Manz, A.: Isotachophoresis in free-flow using a miniaturized device. Anal. Chem. **78**, 3815–3819 (2006)
7. Fonslow, B.R., Bowser, M.T.: Optimizing band width and resolution in micro-free flow electrophoresis. Anal. Chem. **78**, 8236–8244 (2006)
8. Castellanos, A., Ramos, A., González, A., Green, N. G., Morgan, H.: Electrohydrodynamics and dielectrophoresis in microsystems: scaling laws. J. Phys. D. Appl. Phys. **36**, 2584–2597 (2003)
9. Albrecht, J.W., El-Ali, J., Jensen, K.F.: Cascaded free-flow isoelectric focusing for improved focusing speed and resolution. Anal. Chem. **79**, 9364–9371 (2007)
10. Albrecht, J.W., Jensen, K.F.: Micro free-flow IEF enhanced by active cooling and functionalized gels. Electrophoresis **27**, 4960–4969 (2006)
11. Fonslow, B.R., Barocas, V.H., Bowser, M.T.: Using channel depth to isolate and control flow in a micro free-flow electrophoresis device. Anal. Chem. **78**, 5369–5374 (2006)
12. Fonslow, B.R., Bowser, M.T.: Free-flow electrophoresis on an anodic bonded glass microchip. Anal. Chem. **77**, 5706–5710 (2005)
13. Janasek, D., Schilling, M., Manz, A., Franzke, J.: Electrostatic induction of the electric field into free-flow electrophoresis devices. Lab Chip **6**, 710–713 (2006)
14. Kobayashi, H., Shimamura, K., Akaida, T., Sakano, K., Tajima, N., Funazaki, J., Suzuki, H., Shinohara, E.: Free-flow electrophoresis in a microfabricated chamber with a micromodule fraction separator: continuous separation of proteins. J. Chromatogr. A **990**, 169–178 (2003)
15. Kohlheyer, D., Besselink, G.A.J., Schlautmann, S., Schasfoort, R.B.M.: Free-flow zone electrophoresis and isoelectric focusing using a microfabricated glass device with ion permeable membranes. Lab Chip **6** (2006)
16. Kohlheyer, D., Eijkel, J.C.T., Schlautmann, S., van den Berg, A., Schasfoort, R.B.M.: Microfluidic high-resolution free-flow isoelectric focusing. Anal. Chem. **79**, 8190–8198 (2007)
17. Lu, H., Gaudet, S., Schmidt, M.A., Jensen, K.F.: A microfabricated device for subcellular organelle sorting. Anal. Chem. **76**, 5705–5712 (2004)
18. Raymond, D.E., Manz, A., Widmer, H.M.: Continuous sample pretreatment using a free-flow electrophoresis device integrated onto a silicon chip. Anal. Chem. **66**, 2858–2865 (1994)
19. Xu, Y., Zhang, C.-X., Janasek, D., Manz, A.: Sub-second isoelectric focusing in free flow using a microfluidic device. Lab Chip **3**, 224–227 (2003)
20. Zhang, C.-X., Manz, A.: High-speed free-flow electrophoresis on chip. Anal. Chem. **75**, 5759–5766 (2003)

Micro/Nano Flow Characterization Techniques

Yanwei Wang[1], Craig Snoeyink[2], Xudong Pan[3] and Steven T. Wereley[4]

[1]Mechanical Engineering, Harbin Institute of Petroleum, Purdue University, West Lafayette, IN, USA

[2]Birck Nanotechnology Center, Mechanical Engineering, Purdue University, West Lafayette, IN, USA

[3]Harbin Institute of Technology, Harbin, Heilongjian, People's Republic of China

[4]School of Mechanical Engineering, Purdue University, Birck Nanotechnology Center, West Lafayette, IN, USA

Synonyms

Astigmatic microparticle imaging; Biomimetic flow sensors; Capillary flow; Micropumps; Nano (evanescent-wave)-particle image velocimetry

Definition

Micro/nano flow characterization techniques are methods used to measure fluid motion in a spatially resolved manner over short length scales, usually ranging from 1 mm to 10 μm. Typically the fluid motion is observed through the introduction of tracer particles, and multiple experimental images are acquired and analyzed with spatial correlation methods or particle tracking algorithms. These measurement techniques are distinguished from their macroscale counterparts by the various difficulties associated with imaging at these length scales and the methods used to overcome these difficulties.

Overview

Micro/nano flow characterization techniques are used to develop spatially resolved velocity measurements of micro- and nanofluidic devices. The need for these measurement techniques has continued to grow since the development of the first micro flow characterization technique, microparticle image velocimetry (μPIV), in 1998 [1]. Since then the number and variety of methods of manipulating micro- and nanofluidic flows has exploded. It is now common to find microfluidic devices with several different forces acting on the flow that are not typically significant at macroscales. Surface tension gradients, electrophoretic effects, and electrothermal forces are all common drivers of microfluidic flows.

As a result of the increasing complexity of micro- and nanofluidic devices and the flows within them, what started with μPIV has grown into a range of techniques capable of measuring flows of varying complexity and length scales. This entry will discuss the most prominent of these flow measurement methods covering the basics of each technique as well as their applications.

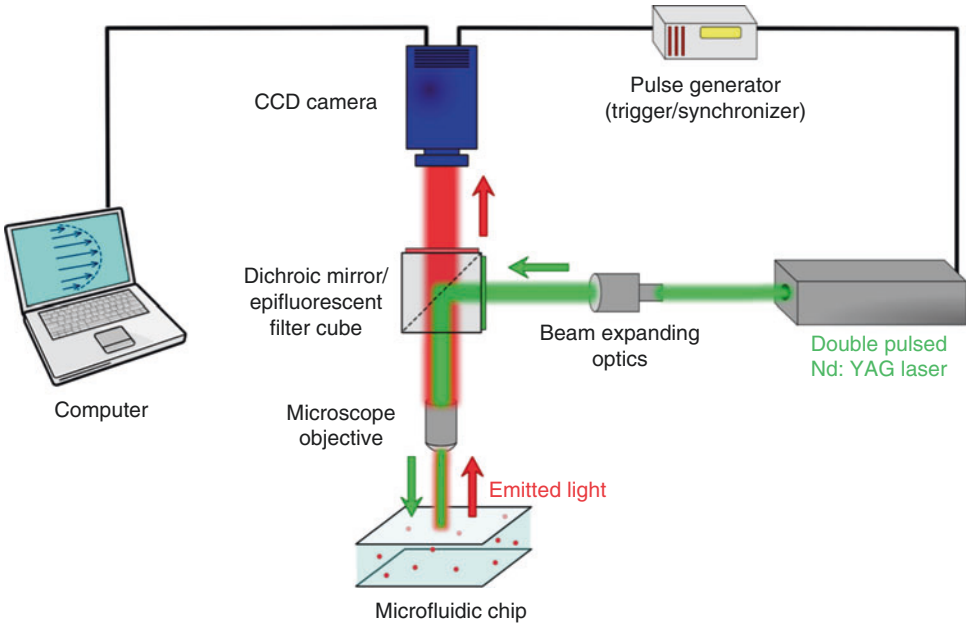
μPIV

Microparticle image velocimetry (μPIV) is the oldest of the micro/nanoscale flow characterization techniques and shares many features with its macroscale counterpart, PIV. Many thorough reviews of PIV techniques exist [2, 3], and because of the similarity between the two techniques the reader is encouraged to consider them. Here the focus will be on the unique constraints imposed on μPIV by viewing the test section through a microscope.

The first difference relates to the fact that both PIV and μPIV are two-dimensional two-coordinate measurement techniques. What this means is that the resulting velocity data is confined to single plane and each velocity vector has only in-plane components. In PIV the measurement is confined to a single plane by illuminating the sample with a laser sheet. In this way the only particles that are imaged are those in the plane of the laser. This is not practical in many micro/nanoscale applications due to the small dimensions of the fluidic channels. As a result the sample is usually volume illuminated using the microscope's built-in illumination light path. In order to restrict the field of view to as thin a measurement plane as possible, a high numerical aperture microscope objective is used [4]. The high numerical objective results in a relatively thin plane in which particles are in focus. Particles outside the focal plane of the microscope objective are defocused and contribute less to the resulting velocity field [5]. This can result in relatively large errors in the measurement if there are large velocity gradients perpendicular to the measurement plane. Choosing a microscope objective with a high numerical aperture and using small-diameter tracer particles (<1 μm) can minimize this effect.

Because the microfluidic test section is volume illuminated there is often a lot of background noise. Reflection of the illumination light off the walls of the test section can overwhelm the signal from the particles. Minimizing the background noise is typically accomplished using fluorescent tracer particles and an epifluorescence filter cube in the light path. A typical μPIV setup is shown in Fig. 1.

As in traditional PIV, a double-pulsed Nd:YAG laser is used to illuminate the sample. However, in μPIV this light is sent through an epifluorescence

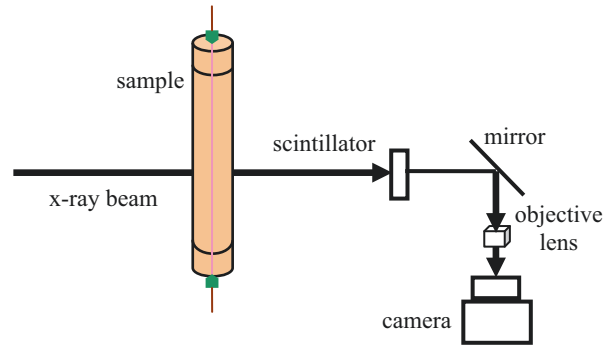


Micro/Nano Flow Characterization Techniques, Fig. 1 Schematic of conventional micro PIV system illustrating the essential components (From Williams et al. [6] with permission)

filter cube which has a dichroic mirror in it. The mirror reflects short wavelengths and passes long wavelengths. In this manner the excitation light from the laser of 532 nm is filtered from the light passing to the camera. The emitted light from the particles, usually around 630 nm, passes freely to the camera, greatly increasing the signal-to-noise ratio. Companies like Thermo Scientific, Polysciences Inc., and Sigma Aldrich offer a variety of microparticles available in a range of excitation and emission wavelengths.

As in PIV, the images in μPIV are evaluated with a conventional correlation-based algorithm. Like in PIV, a sufficient number of particle images are required in the interrogation window for reliable and accurate measurements. For practical reasons this high particle seeding can be difficult to achieve in microfluidic devices. For example, since defocused particles still contribute light to the image, increasing the number of particles in the measurement volume can have the effect of lowering the signal-to-noise ratio of the images. Additionally, both the seeding particles and the velocities being measured are small enough that the random walks of particles due to Brownian motion can contribute significantly to the uncertainty of the measurements. For both of these reasons the velocity field calculated from a single pair of images is likely to have an unacceptably high uncertainty.

This high uncertainty is easily addressed due to one of the advantages of studying microfluidic flows: The fluid mechanics can often be simplified as steady, quasisteady, or periodic due to the very low Reynolds numbers involved (usually less than one). This results in laminar Stokes flow. Due to the laminar and steady nature of many microfluidic flows, a large number of image pairs can be analyzed with a correlation technique termed ensemble averaging or correlation averaging [7]. This technique can compensate for suboptimal particle seeding and low-quality images by averaging the correlation over many image pairs, typically hundreds, thereby significantly increasing the signal-to-noise ratio. If the flow is not steady state or periodic, then the velocity analysis is restricted to only one pair of images at a time. In this instance, the measurements may contain erroneous vectors which deviate significantly in magnitude or direction from their neighbors.



Micro/Nano Flow Characterization Techniques, Fig. 2 Example application of X-ray μPIV measuring the velocities of inside flow (From Vlachos et al. [9] with pemission)

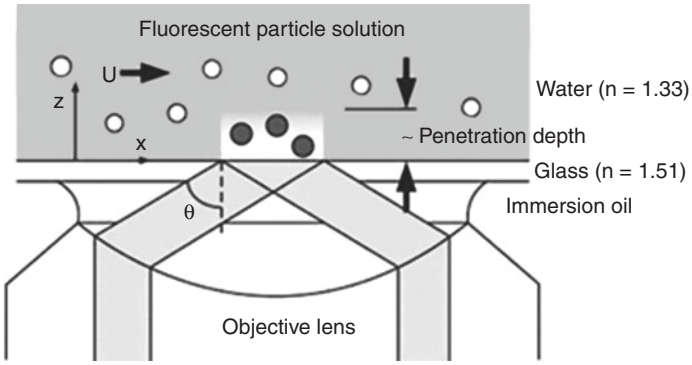
Traditional μPIV techniques have been applied successfully to a wide range of microfluidic flow field investigations. As a result of its popularity, several companies now offer turnkey μPIV systems: TSI (www.tsi.com), Dantec Dynamics (www.dantecdynamics.com), and LaVision (www.lavision.de) are some of the main suppliers of systems and related equipment. Despite its wide use, μPIV has several limitations. The first is that the calculated velocity fields are two dimensional two component, meaning that out-of-plane velocities are not measured. Second, the flow needs to be steady state or periodic such that a sufficient number of image pairs can be taken and averaged to minimize sources of measurement error. Finally, at least one surface of the system being investigated needs to be transparent in order for the tracer particles to be excited and viewed. Overcoming these limitations has been a recent area of development and will be discussed in the following sections.

X-Ray and IR μPIV

For several reasons it may be necessary to use a wavelength of light other than those in the visible spectrum. Usually this occurs when attempting to study microfluidic flows in vivo or in devices that cannot be fabricated in a clear material. For the first case it is possible to use a coherent synchrotron X-ray beam as a light source, much like medical X-rays are used to see inside the body. X-ray μPIV has been used to investigate blood flow in an opaque tube by analyzing the fringe patterns of red blood cells [8]. Even the thick hydrated collagen tissue has no effect on the quality of the phase-contact X-ray image, which can pave the way toward future experimental investigations of the vitro tissue vascular flow. A typical X-ray μPIVsketch is shown in Fig. 2 [9].

Visualizing the X-rays is achieved by passing them through a scintillating crystal and converting them to visible light capable of being imaged by a camera. One of the disadvantages of this technique is that the coherent X-rays are simply shown through the sample without focusing or conditioning. This means there is no way to restrict the measurements to a single plane; all particles in the flow are imaged with no way to determine their depth.

Moving to the other side of the visual spectrum is infrared (IR) radiation which has found use as a light source in silicon devices as a technique called infrared-μPIV (IR-μPIV). Silicon, which is transparent to infrared wavelengths, is often used in heat transfer experiments where the temperatures are too high for typical transparent microfabrication materials. Care must be taken when performing μPIV with infrared illumination. The particle size and material must be matched well to maximize scattering of light at these longer wavelengths since conventional tracers do not absorb or emit in the infrared. Additionally, silicon has a very high index of refraction ($n = 3.3$) which results in aberrations affecting the image quality. Finally, microscope objectives are designed to minimize aberrations in the visible spectrum, and so some optical aberrations are likely to be present when imaging in the infrared. For these reasons the resolution of IR-μPIV is not as good as that of conventional μPIV systems.



Micro/Nano Flow Characterization Techniques, Fig. 3 Example application of evenescent wave micro PIV demonstrating illuminating laser path and measurement region (From Jin et al. [12] with permission)

IR μPIV has been used to investigate flow around a heat sink, and the images were processed with an algorithm that was developed to improve the quality of IR recordings [10].

Evanescent Wave μPIV

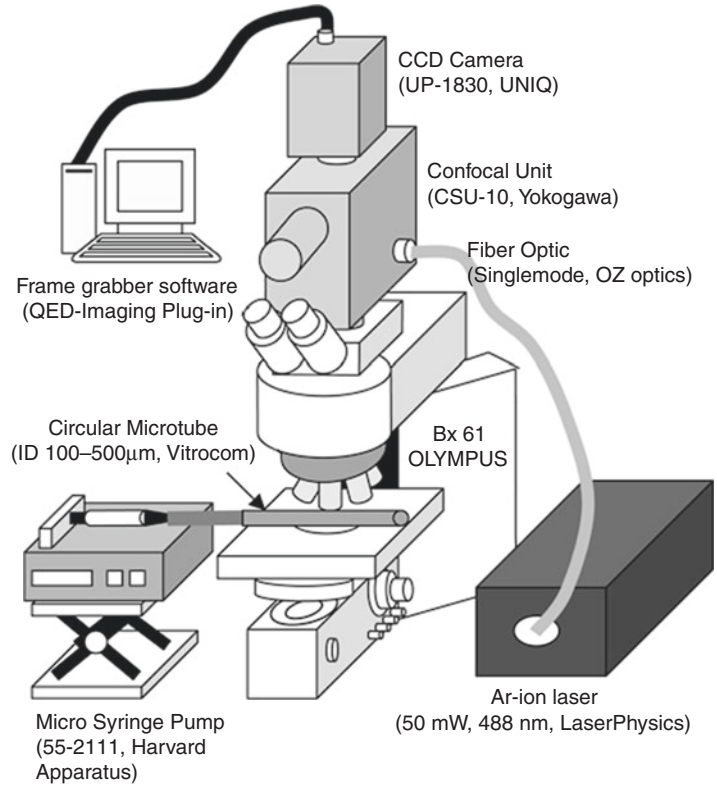
For many applications it is the fluid flowing very near a wall (<1 μm) that is of interest. Since this is less than the practical spatial resolution of μPIV, it is necessary to confine the field of view to only those tracer particles immediately adjacent to the wall. This is accomplished by using evanescent waves generated by total internal reflection (TIR) [11]. Since the intensity of the evanescent wave decays exponentially away from the wall the measurement is confined to a region only 50–250 nm thick. This small penetration depth, which can be adjusted according to the incident angle, also reduces background noise from out-of-focus tracer particles [12, 13].

This technique has been used often to study near-wall velocity profiles [12]. Since the intensity of particles varies in a consistent way according to their distance from the wall, the height of the particles can be inferred from the brightness of the particle images. Using this technique with submicron particles, researchers were able to extract information about the apparent slip length for flow in a microchannel [14] (Fig. 3). Furthermore, this technique can also be applied to study not only fluid flows but also nanoparticle dynamics, a fundamental issue for designing novel devices in fields like nanochemistry [15].

Confocal and Scanning Confocal μPIV

Confocal μPIV takes advantage of the much thinner focal plane that is characteristic of a confocal microscope. The thinner focal plane leads to several advantages: the first is a significant reduction in light from defocused particles; the second is to minimize depth of correlation (DoC) errors inherent in flows with large velocity gradients. In practice, the DoC, which is adjusted by filter parameters, is able to recover the in-plane loss of correlation [16]. It also can be modeled by a weighting function. The DoC correction has been explicitly demonstrated for a stationary flow field in a rectangular channel [17]. Adapted a high-speed confocal scanner to μPIV applications, an illustration of which is shown in Fig. 4 [18]. Usually, a high-speed video camera is used to record data in order to adjust the DoC [19].

Despite recent advances in scanning speed, confocal μPIV is limited to steady flows as it takes a relatively long time to adequately scan the depth of the measurement volume. Although the thinner focal plane allows closer spacing of the measurement planes, and thus finer depth resolution, the technique still is only able to resolve the velocity components in the measurement plane. To get an estimate of the out-of-plane velocities, it is necessary to integrate the continuity equation over the volume of interest. If the image planes are closely spaced this technique has been shown to agree with numerical simulations to within 3 % [20]. To achieve reliable, time-resolved measurements, the robust phase correlation (RPC) method



Micro/Nano Flow Characterization Techniques, Fig. 4 Example application of high-speed scanning confocal microscopy to micro PIV (From Park et al. [18] with permission)

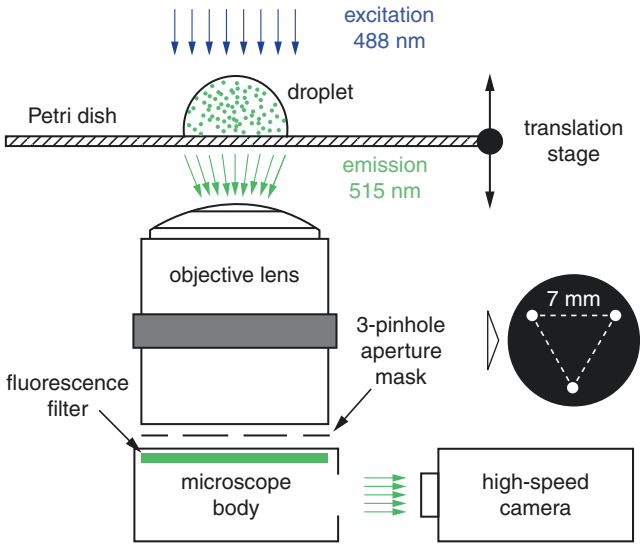
is used to increase the signal–noise ratio (SNR). RPC applies a PIV-optimized spectral filter to the cross-correlation planes resulting in correlation planes with sharper primary peaks and low background levels [21].

Stereo μPIV

Stereo particle image velocimetry allows one to measure all three velocity components but only in a single plane. Stereo PIV uses two viewpoints separated by an angle to infer the three-dimensional position of particles. This technique has been well developed and used successfully in macroscale PIV [22]. Stereo PIV hardware and software are also available commercially.

Applying this technique to microscale flows has proved considerably more challenging due to conflicts between the requirements for good spatial resolution and the accuracy of out-of-plane velocity components. Stereo μPIV gives reliable results in 2D2C (two dimension two components), but due to the unavoidable mismatch of focal planes, systematic errors will appear that cannot be compensated digitally [23]. In order to achieve high spatial resolution, it is necessary to have a large imaging aperture. However, a large imaging aperture imposes practical limits on the angle separating the two viewpoints. This is typically not a problem for macroscale applications as the cameras and optics can be placed sufficiently far away from the measurement plane such that the two viewpoints do not interfere. For microscale applications though there is simply not enough room to place two large aperture imaging optics, typically microscope objectives, separated by a small angle and close enough to view the same measurement plane.

Despite these drawbacks, the technique has been successfully developed, albeit with lower spatial resolution than typical μPIV. Bown et al. [24] reported a spatial resolution of $10 \times 10 \times 10\ \mu m^3$ when measuring the flow over a backward-facing step [24]. In practical



Micro/Nano Flow Characterization Techniques, Fig. 5 Schematic demonstrating a typical defocusing micro PIV setup (From Pereira et al. [29] with permission)

applications, stereo μPIV is used to investigate oscillatory flows, whose Reynolds number is 150 and 2.65 at inspiration and expiration, in the realistic human lung model [25]. Recently, this technique has also been used to investigate the tip leakage vortex and the flame propagation in a practically premixed flow [26, 27].

Particle Image Defocusing

Some of the complexity of stereo imaging optics can be avoided by employing a particle image defocusing technique. These techniques rely on a single light collection lens and camera and determine the three-dimensional location of the particle using the degree of defocusing of the particle image. Once the location of the particles is determined for two images, a particle tracking velocimetry (PTV) algorithm can be used to determine the velocity components. Furthermore, interference-PTV, which determines the depth of particle, is suited to situations with nonuniform illumination. It can be applied to any optical imaging system where a light source with high contrast must be located in three dimensions [28]. Since the positions of the particles are known in three dimensions, the resulting velocity vectors are also known for all three dimensions throughout the measurement volume. This is in contrast to the previously mentioned techniques for which many two-dimensional measurement slices have to be acquired in order to build up a measurement volume. One drawback though is that these techniques typically require a much lower particle seeding concentration so that the particle images do not interfere with each other. As a result, many image pairs must be acquired and combined in order to increase the spatial resolution to an acceptable level.

The first of these techniques to be applied to microscale flows was defocusing digital particle image velocimetry (DDPIV). The arrangement is shown in Fig. 5 [29]. DDPIV is employed to visualize the resist filling process, which helps to understand the resist filling behavior and mechanism [30].

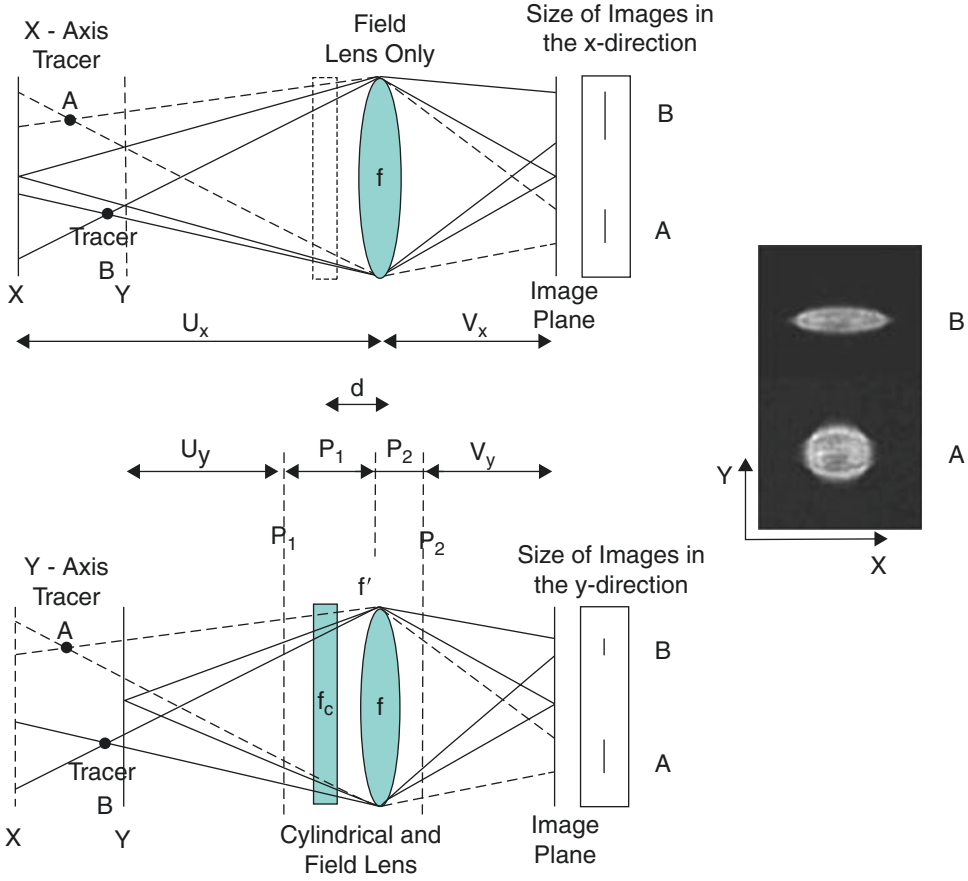
This transforms the particle image into a set of dots, the relative spacing of which can be used to determine the depth of the particle. Yoon and Kim [31] applied this technique to a backward-facing step and by averaging over 2,000 image pairs were able to achieve an average spatial resolution of 5 μm in plane and 1 μm out of plane [31].

A similar technique does away with the mask and instead uses the ring structure from the diffracted defocused image of a particle. An example of this effect is shown in Fig. 6.

The diameter of the outermost ring of the diffraction pattern can be related to the depth of



Micro/Nano Flow Characterization Techniques, Fig. 6 Images of a fluorescent bead scanned at 10 μm intervals along the optical axis with a 20× objective (From Wu et al. [32] with permission)



Micro/Nano Flow Characterization Techniques, Fig. 7 Schematic of astigmatism micro PIV demonstrating the arrangement of optical elements and the resulting ellipsoidal particle image (From Chen et al. [34] with permission)

the particle through a calibration curve obtained prior to the experiment. Once this is done the three-dimensional position of the particle can be extracted and a particle tracking algorithm used to determine the velocity components [32].

This technique has the obvious advantage of requiring only common μPIV equipment. However, in order to get good particle images it is necessary to use larger particles (>2 μm) and high magnifications. This technique also requires images with a good signal-to-noise ratio in order to determine the radius of the outermost ring of the particle image which is often quite faint. Recent research has shown that the defocusing-based 3D particle tracking method by automatic calibration can be used to track particles less than 2 μm in diameter [33].

Rounding out this class of measurement techniques is astigmatism microparticle tracking velocimetry (A-μPTV). This technique also relates the diameter of a defocused particle to its depth but adds an asymmetry to the optical path. This asymmetry causes the perimeter of the defocused particle image to form an ellipse for which the aspect ratio is a function of particle depth. In A-μPTV this asymmetry takes the form of a cylindrical lens placed between the camera and the microscope as shown in Fig. 7 [34, 35].

This technique has recently been used to measure a three-dimensional micro vortex with an estimated resolution of 3 μm. This technique has proved more robust with respect to noise and image quality as the light from the particles is typically spread out over a smaller area as compared to the other defocusing techniques [35].

To date, mPIV has been applied to an ever growing list of practical microfluidics problems. Micro-PIV has been used to investigate a wide range of fluid properties beyond simple velocity, such as micro-volume viscosity [36], fluid mixing [37], and viscoelastic microfluidic hydrogels [38]. In the future μPIV will be used in many more fields.

Cross-References

- ▶ Applications of Nanofluidics
- ▶ Capillary Flow
- ▶ Conductivity of Metal Nanowires studied by Infrared Plasmon-polariton Spectroscopy
- ▶ Detection of Nanoparticulate Biomarkers in Blood
- ▶ Dissipative Particle Dynamics, Overview
- ▶ Micro- and Nanofluidic Devices for Medical Diagnostics
- ▶ Nanoparticle Tracking Analysis
- ▶ Nanoparticles
- ▶ Nanotechnology
- ▶ Nanotechnology Applications in Polymerase Chain Reaction (PCR)
- ▶ Nanotechnology in Cardiovascular Diseases
- ▶ Nanotechnology in Pulmonary Disease
- ▶ Nanotribology
- ▶ Rapid Electrokinetic Patterning and its Applications
- ▶ Surface-Modified Microfluidics and Nanofluidics

References

1. Santiago, J.G., Wereley, S.T., Meinhart, C.D., Beebe, D.J., Adrian, R.J.: A particle image velocimetry system for microfluidics. Exp. Fluids **25**(4), 316–319 (1998)
2. Adrian, R.J.: Twenty years of particle image velocimetry. Exp. Fluids **39**(2), 159–169 (2005)
3. Lindken, R., Rossi, M., Große, S., Westerweel, J.: Micro-particle image velocimetry (μPIV): recent developments, applications, and guidelines. Lab Chip **9**(17), 2551 (2009)
4. Meinhart, C.D., Wereley, S.T., Santiago, J.G.: PIV measurements of a microchannel flow. Exp. Fluids **27**(5), 414–419 (1999)
5. Bourdon, C.J., Olsen, M.G., Gorby, A.D.: Validation of an analytical solution for depth of correlation in microscopic particle image velocimetry. Meas. Sci. Technol. **15**(2), 318–327 (2004)
6. Williams, S.J., Park, C., Wereley, S.T.: Advances and applications on microfluidic velocimetry techniques. Microfluid. Nanofluid. **8**(6), 709–726 (2010)
7. Meinhart, C.D., Wereley, S.T., Santiago, J.G.: A PIV algorithm for estimating time-averaged velocity fields. J Fluids. Eng. **122**(2), 285–289 (2000)
8. Kim, G.B., Lee, S.J.: X-ray PIV measurements of blood flows without tracer particles. Exp. Fluids **41**(2), 195–200 (2006)
9. Antoine, E., Buchanan, C., Fezzaa, K., Lee, W.K., Rylander, M.N., Vlachos, P.: Flow measurements in a blood-perfused collagen vessel using X-ray microparticle image velocimetry. PLoS One **8**(11), 1–8 (2013)
10. Han, G., Bird, J.C., Johan, K., Westin, A., Cao, Z., Breuer, K.S.: Infrared diagnostics for measuring fluid and solid motion inside silicon microdevices. Microscale Thermophys. Eng. **8**(2), 169–182 (2004)
11. Frumin, L.L., Nemykin, A.V., Perminov, S.V., Shapiro, D.A.: Plasmons excited by an evanescent wave in a periodic array of nanowires. J. Optics **15**(8), 2048–2052 (2013)
12. Jin, S., Huang, P., Park, J., Yoo, J.Y., Breuer, K.S.: Near-surface velocimetry using evanescent wave illumination. Exp. Fluids **37**(6), 825–833 (2004)
13. Lisicki, M., Cichocki, B., Dhont, J.K.G., Lang, P.R.: One-particle correlation function in evanescent wave dynamic light scattering. J. Chem. Phys. **13**(20), 204704-1–204704-7 (2012)
14. Li, H., Sadr, R., Yoda, M.: Multilayer nano-particle image velocimetry. Exp. Fluids **41**(2), 185–194 (2006)
15. Yutaka, K., Keizo, I., Kazuma, M., Takehiko, K.: Evanescent wave-based particle tracking velocimetry

for nanochannel flows Kazoe. Anal. Chem. **85**(22), 10780–10786 (2013)
16. Mustin, B., Stoeber, B.: Effect of linear image processing on the depth of correlation in micro PI-V. Exp. Fluids **55**(10), 1–18 (2014)
17. Hein, M., Wieneke, B., Seemann, R.: Calculation of the weighting function and determination of the depth of correlation in micro-piv from experimental particle images. Meas. Sci. Technol. **25**(8), 084008-1–084008-11 (2014)
18. Park, J., Choi, C., Kihm, K.: Optically sliced micro-PIV using confocal laser scanning microscopy (CLSM). Exp. Fluids **37**(1), 105–119 (2004)
19. Kikuchi, K., Mochizuki, O.: Velocity profile of thin film flows measured using a confocal microscopy particle image velocimetry system with simultaneous multi depth position. Meas. Sci. Technol. **26**(2), 1–8 (2015)
20. Pommer, M.S., Kiehl, A.R., Soni, G., Dakessian, N.S., Meinhart, C.D.: A 3D-3C Micro-PIV method. In: 2007 2nd IEEE International Conference on Nano/Micro Engineered and Molecular Systems, Bangkok, Thailand, vol. 1–3, pp. 976–981 (2007)
21. Raben, J., Klein, S., Posner, J., Vlachos, P.: Improved accuracy of time-resolved micro-particle image velocimetry using phase-correlation and confocal microscopy. Microfluid. Nanofluid. **14**(3), 431–444 (2013)
22. Arroyo, M.P., Greated, C.A.: Stereoscopic particle image velocimetry. Meas. Sci Technol. **2**(12), 1181–1186 (1991)
23. Cierpka, C., Rossi, M., Segura, R., Mastrangelo, F., Kähler, C.: A comparative analysis of the uncertainty of astigmatism-μPTV, stereo-μPIV, and μPIV. Exp. Fluids **52**(3), 605–615 (2012)
24. Bown, M.R., MacInnes, J.M., Allen, R.W.K., Zimmerman, W.B.J.: Three-dimensional, three-component velocity measurements using stereoscopic micro-PIV and PTV. Meas. Sci. Technol. **17**(8), 2175–2185 (2006)
25. Soodt, T., Pott, D., Klaas, M., Schröder, W.: Analysis of basic flow regimes in a human airway model by stereo-scanning PIV. Exp. Fluids **54**(6), 1–10 (2013)
26. Matthieu, D., Jean, D., Cécile, M.A., Mohamed, F.: Mind the gap: a new insight into the tip leakage vortex using stereo-PIV. Exp. Fluids **55**(11), 1–13 (2014)
27. Weinkauff, J., Trunk, P., Frank, J.H., Dunn, M.J., Dreizler, A., Böhm, B.: Investigation of flame propagation in a partially premixed jet by high-speed-Stereo-PIV and acetone-PLIF. Proc. Combust. Inst. **35**(3), 3773–3781 (2015)
28. Snoeyink, C., Wereley, S.: A novel 3D3C particle tracking method suitable for microfluidic flow measurements. Exp. Fluids **54**(1), 1–10 (2013)
29. Pereira, F., Lu, J., Castaño-Graff, E., Gharib, M.: Microscale 3D flow mapping with μDDPIV. Exp. Fluids **42**(4), 589–599 (2007)
30. Du, J., Wei, Z.Y., Li, S.Z., Tang, Y.P.: Resist deformation and flow fields in microimprint lithography. J. Micromech. Microeng. **23**(5), 1–12 (2013)
31. Yoon, S., Kim, K.: 3D particle position and 3D velocity field measurement in a microvolume via the defocusing concept. Meas. Sci. Technol. **17**(11), 2897–2905 (2006)
32. Wu, M., Roberts, J.W., Buckley, M.: Three-dimensional fluorescent particle tracking at micron-scale using a single camera. Exp. Fluids **38**, 461–465 (2005)
33. Winer, M.H., Ahmadi, A., Cheung, K.C.: Application of a three-dimensional (3D) particle tracking method to microfluidic particle focusing. Lab Chip **14**(8), 1443–1451 (2014)
34. Chen, S., et al.: Wavefront sensing for three-component three-dimensional flow velocimetry in microfluidics. Exp. Fluids **47**(4–5), 849–863 (2009)
35. Cierpka, C., Segura, R., Hain, R., Kähler, C.J.: A simple single camera 3C3D velocity measurement technique without errors due to depth of correlation and spatial averaging for microfluidics. Meas. Sci. Technol. **21**(4), 045401 (2010)
36. Sie, Y.S., Chuang, H.S.: A micro-volume viscosity measurement technique based on μPIV diffusometry. Microfluid. Nanofluid. **16**(1), 65–72 (2014)
37. Blois, G., Barros, J., Christensen, K.T.: A microscopic particle image velocimetry method for studying the dynamics of immiscible liquid–liquid interactions in a porous micromodel. Microfluid. Nanofluid. **18**(5), 1391–1406 (2015)
38. Kwon, J.S., Wereley, S.T.: Light-actuated electrothermal microfluidic motion: investigation and physical interpretation. Microfluid. Nanofluid. (5), (2015). http://link.springer.com/article/10.1007/s10404-015-1587-z (published on line)

Micro/Nano Transport in Microbial Energy Harvesting

Ranajay Ghosh[1], Aloke Kumar[2] and Partha P. Mukherjee[3]
[1]Department of Mechanical, Aerospace and Nuclear Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
[2]Department of Mechanical Engineering, University of Alberta, AB, Canada
[3]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Synonyms

Bacterial electrical conduction; Cellular electronic energy transfer; Microbial electrolysis cell; Microbial fuel cell

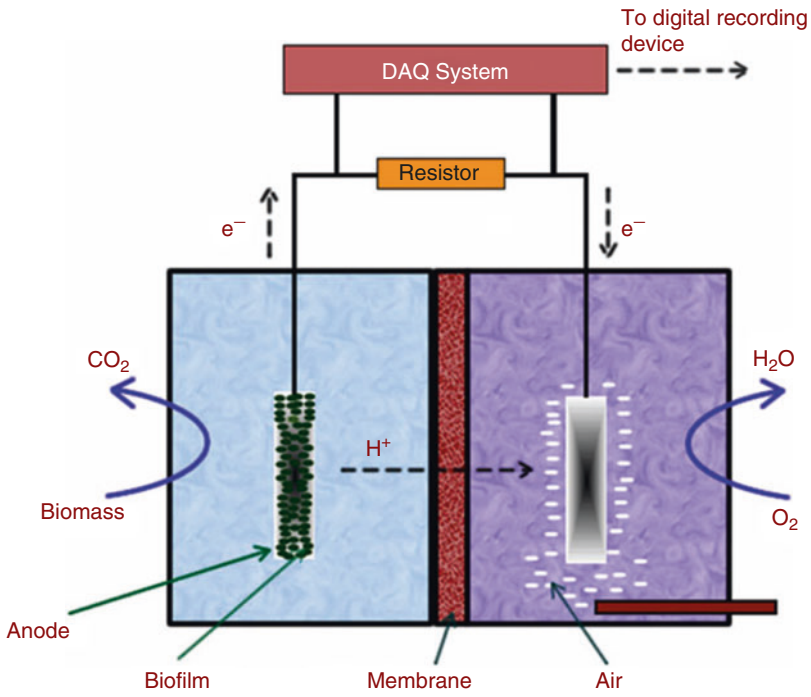
Definition

Prokaryotic cells under some conditions are capable of extracellular transport of electrons generated during their metabolic processes. When these cells donate electrons to a conductive surface (e.g., an electrode), they cause electron flow, which can be used for harvesting energy from organic substrates. Disparate multi-scale transport processes relevant to the energy generation scenario are elucidated in this entry.

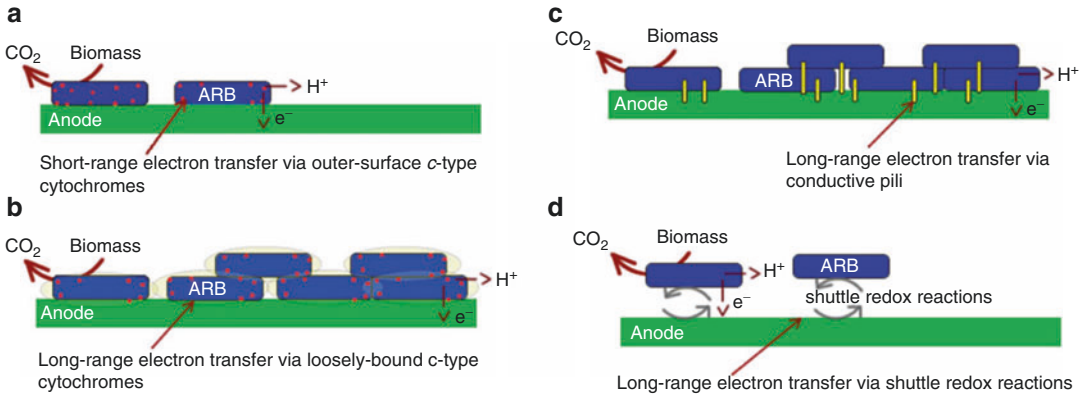
Overview

Microbes are micro organisms which survive by utilizing energy from environmental nutrients (organic substrates) through the process of respiration. These organic substrates are essentially electron donors, which carry electrons at high energy levels that are ultimately transferred to a biomolecule (called the electron acceptor) through a set of complex biochemical reactions involving the electron transport chain (ETC) [1]. Microbial energy-harvesting techniques successfully capture the electrons at various stages of the ETC and deliver these electrons to an electrode, where they can flow through an external circuit thus performing electrochemical work. Needless to say, not all microbes are capable of successful extracellular electron transport. This energy transport from the substrate to the electrode involves complex multi-scale phenomena, which include, but are not limited to, mass transport of charged species through diffusion and convection, chemical transfer through mediator species, direct conduction through bacterial nanowires, microscale aggregation of bacteria into complex continua, and nucleation of heterogeneities through mutations.

A typical design of a microbial energy harvesting device is depicted in Fig. 1. It consists of an anaerobic anode chamber and a cathode chamber which are open to air and separated from each other by an ion exchange membrane. In the anode chamber, living microbial cells harvest high-energy electrons from an electron-rich substrate generating H^+ (proton) and CO_2 in the process. At the end of energy extraction, the now low-energy electrons combine with an ambient



Micro/Nano Transport in Microbial Energy Harvesting, Fig. 1 A schematic of a typical microbial energy-harvesting device



Micro/Nano Transport in Microbial Energy Harvesting, Fig. 2 Electron transfer from a microbe in the biofilm to the anode surface can occur via various mechanisms. (**a**–**c**) Schematics of conductive transport of electron; (**d**) advection-based transport through a biochemical mediator

electron acceptor and are removed from the cell. In an aerobic respiration process, dissolved oxygen acts as the ultimate electron acceptor whereas in a usual microbial energy-harvesting device, the conducting electrode placed in the anaerobic anodic chamber acts as the ultimate electron acceptor. This entire process proceeds via a set of extremely complex chemical reactions as part of the ETC and occurs inside the mitochondria of a cell isolated from the external environment [2]. These transferred electrons can flow through an external circuit to the cathode chamber, where the H^+ is reduced to H_2O in the presence of oxygen. A device, which is capable of executing this energy-harvesting process, is schematically shown in Fig. 1. It can be noted that the design is very similar to a fuel cell and hence such a device is also known as a microbial energy-harvesting (MEH) device. The most important transport process involved in a device like the one shown in Fig. 1 is the transport of electrons to the anode surface.

Typically, microbes associate with an external electrode through a phase of aggregative living called a biofilm inside which the electron transfer takes place. The exact mechanism of this extracellular electron transport is still an active area of research [3] although two broad mechanisms exist – mediator-based and conduction-based electron transport. In the mediator-based mechanism, an intrinsically produced (or externally added) chemical called the mediator capable of freely moving into and out of the mitochondrial structure acts as an electron acceptor. It gets reduced by the electrons inside the mitochondria, diffuses out of the cell, and is finally oxidized by the electron-accepting electrode. Hence, at the end of the electron donation to the electrode, the mediator is once again free for further electron transport and the process continues. On the other hand, in the conduction-based mechanism, electrons are directly deposited onto the electrodes through extracellular electron transport like conducting pilli or other mechanisms [4, 5] obviating the need for mediators. However, at any given time it is highly likely that a combination of mechanisms described above is present. A schematic picture explaining the various mechanisms is shown in Fig. 2.

It is clear from the above discussion that efficiency of the microbial energy transport will depend critically on a number of complex biological, chemical, and physical factors, namely, species of microbes used, mechanism of electron transfer onto the electrode surface, ambient physical and chemical environment of the electrode chamber, as well as the materials used for electrodes. Each of the above-mentioned factors intimately affects the others creating coupled multiphysics operating conditions. For instance,

although the biological species being used in the electrode chamber can affect the electron transfer mechanism and thereby the chemical environment of the microbe colony, the chemical environment itself can affect the dynamics of metabolism and growth of the microbe colony. This also means that research carried out in a number of areas is relevant to understanding the science of microbial energy transport. It is appropriate therefore to classify the rapidly growing and variegated literature of microbial energy transport into the following broad areas of research:

- Microbiology and biochemistry research
- Device design
- Transport studies

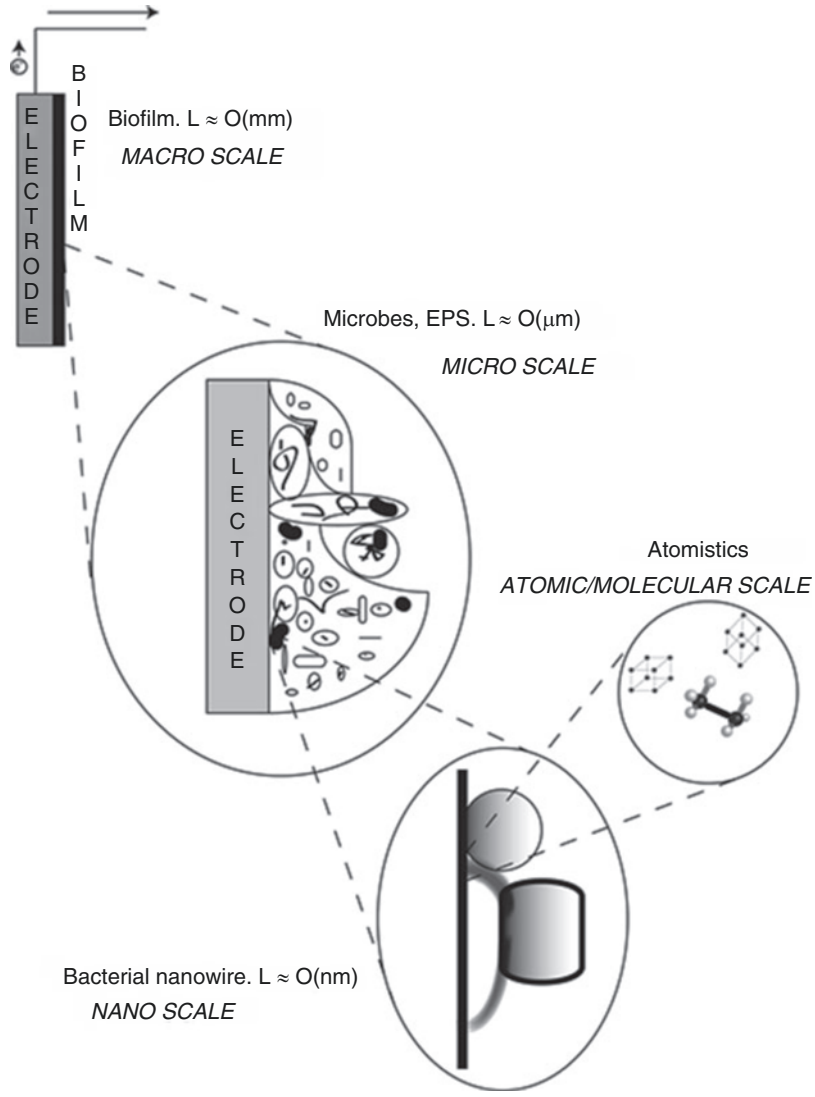
In this entry, the transport modeling aspect of microbial fuel cell energy conversion would be dealt with. The next sections describe the methodology of modeling, key findings, and future directions for research.

Method

The overall principle of operation of an MEH device was described in the previous section. It is evident that modeling microbial-level energy transfer critically depends on our ability to understand and quantify the complexities described above.

As noted in the previous section, only some microbes can participate in extracellular energy transport. Two such microbes, which have been extensively investigated, are *Geobacter sulfurreducens* and *Shewanella oneidensis*. As depicted in Fig. 1, the appropriate microbe grows in the anode chamber over the anode surface. The microbes organize themselves in an elaborate thin structure bonded to the anode where they are embedded within a self-produced matrix of extracellular polymeric substance (EPS) [6]. EPS is generally composed of extracellular DNA, proteins, and polysaccharides. The colony thus formed grows because of continuous diffusion of the nutritious substrate. This diffusion of nutritious substrate is afforded by a constant flux of biomass into the anode chamber. This substrate is utilized by the microbes for respiration under a favorable biochemical environment, producing electrons which are then transferred to the electrode (Fig. 2) [7], with the simultaneous production of H^+ ions. The H^+ ions then migrate via a proton-exchange membrane to the cathode chamber (Fig. 1). Finally, in the cathode chamber, the H^+ ions are reduced by the electrons thus completing the cycle.

A complete modeling effort for an MEH device would naturally include multi-scale transport aspects. Figures 1, 2, and 3 capture various aspects of the multi-scale charge and mass transport processes. The overall MEH device and, consequently, the fluxes and transport of the bulk fluid are macroscale. Thus, the transport of H^+ ions from the anode to cathode is seemingly macroscale in nature. However, the H^+ ions themselves need to diffuse out of the biofilm which has a typical thickness of the order of hundreds of microns. Thus modeling of transport in the biofilm has to take into account this much smaller length scale that starts at the biofilm boundary. Looking more closely at the biofilm itself, one finds that the microbes, which reside in the biofilm, are of micron to submicron resolution. On the other hand, charge transport from individual microbes to the electrode surface may encompass an even smaller length scale defined by pilli and certain protein complexes (Fig. 2). Their characteristic lengths often lie in the nanometer scale. The complete illustration of the hierarchical length scales present in the system is shown in Fig. 3. It is usual for a model to deal with variables defined on a higher spatial scale which itself approximates (or homogenizes) information from a lower scale assuming fluctuations to be averaged out. For instance, nanoscale energy transfer through pilli (bacterial nanowires) can be taken into account through a micro-level conductivity variable rather than through molecular-level simulations. The choice of scale of model depends on the available computing power as well as resolution necessary to observe critical control variables. In addition to the spatial multi-scale described above, disparity on temporal scale also



Micro/Nano Transport in Microbial Energy Harvesting, Fig. 3 Multi-scale energy transport in a microbial fuel cell

exists during a typical MEH device operation. For instance, the diffusion and convection take place at much smaller timescale than the growth of biofilm which itself grows at a much smaller timescale than the entire operating life cycle of an MEH. In addition, the chemical kinetics of the rate-determining steps of various reactions also provide another timescale. Therefore it is important to note that steady state assumption of one model may be a transient for another. This interplay between the multiple length and timescales makes the operating dynamics of an MEH device both complex as well as rich.

Electrochemistry

In contrast to fuel cells, the voltage generation of an MEH device is intimately tied to the biological properties of the microbes in addition to the chemistry of the fuel (substrates). The microbes take time to colonize the electrode by synthesizing enzymes and biological structures necessary to transfer electrons outside their cellular structure. This possibility of electron transfer causes a voltage or an electro motive force (EMF) to appear at the electrode capable of charge transport through current flow if connected to an external circuit.

In this way one can estimate the net voltage of a cell made of two such electrodes (half cells) as follows:

$$E_{emf} = E_{Cathode} - E_{Anode} \tag{1}$$

where $E_{Cathode}$ is the cathode EMF and EA_{node} is the anode EMF.

In mixed cultures, different bacteria can grow, setting different potentials. Unfortunately, predicting these potentials even for pure cultures is impossible. However, limits to the maximum voltage generation can be calculated using thermodynamic properties of the electron donors (substrates) and acceptors (oxidizers). The maximum equilibrium electromotive force that can be generated for a given chemical reaction is governed by the laws of thermodynamics and depends on the concentration of the species involved, nature of the reaction and the ambient temperature. This equation known as the Nernst equation in electrochemistry is mathematically expressed for both electrodes (cathode and anode) as:

$$\begin{aligned} E_{Cathode} &= E^0_{Cathode} - \frac{RT}{nF}\ln(\Pi_{Cathode}), \\ E_{Anode} &= E^0_{Anode} - \frac{RT}{nF}\ln(\Pi_{Anode}), \end{aligned} \tag{2}$$

where $E_i(i = Cathode,\ Anode)$ is the maximum electromotive force (EMF) of the half cell, $E^0_i(i = Cathode, Anode)$ is the standard half cell potential standardized against a standard electrode such as the normal hydrogen electrode (NHE) [8], T is the ambient temperature, n is the number of electrons transferred in the reaction, $R = 8.314$ J mol^{-1} K^{-1} is the universal gas constant, and $F = 96{,}485$ C mol^{-1} is the Faraday's constant. The reaction quotient $\Pi i(i = Cathode,\ Anode)$ is defined as:

$$\Pi_i = \frac{[products]^P}{[reactants]^r}, \qquad (i = Cathode, Anode) \tag{3}$$

Substrates oxidized through a set of complex biochemical reactions by the microbes in the anode chamber produce a number of different electron carriers. It is these electron carriers that become the immediate electron donors during the ETC from which electrons are tapped out of the cell. For instance, if the substrate is oxidized using a citric acid cycle, an electron carrier compound called NADH becomes the immediate electron donor. In the absence of oxygen in the anode chamber, NADH donates its electrons to the anode and is oxidized into NAD^+. The electron-deficient NAD^+ is then free to accept electrons from the substrate and continue the process. This electron donation process can be represented by the following chemical reaction:

$$NAD^+ + 2H^+ + 2e^- \rightarrow NADH \tag{4}$$

Using the Nernst equation described above, a theoretical limit of the potential difference produced by this reaction can be obtained as follows:

$$E_{emf} = E^0 - \frac{RT}{2F}\ln\left(\frac{[NADH]}{[NAD^+][H^+]^2}\right) \tag{5}$$

Several conclusions can be drawn almost immediately from the above equation if it is rewritten as:

$$E_{emf} = E^0 - \frac{RT}{2F}\ln(\gamma) - \frac{2.303RT}{F}pH \tag{6}$$

where the ratio of [NADH] and $[NAD]^+$ is denoted by γ and $pH = -\log[H^+]$. Equation 6 clearly indicates that the voltage generated depends on the relative intracellular chemical pool of NADH/NAD^+ as well as the acidity of the ambient quantified by the pH. Since production of the NADH/NAD^+ couple is tied to the citric acid cycle, the microbial energy production is now physically linked to the chemistry of microbial respiration. It must be noted that NADH is not the only critical electron donor species generated. The respiration process is composed of many different steps, which can produce many such electron carrier chemical species.

The electrode potentials (EMFs) can also be used to predict the direction of a chemical reaction

through their relationship with the Gibb's free energy. Thermodynamically, the electrode potential can be related to the Gibb's free energy (ΔG) as follows:

$$E^0 = -\frac{\Delta G}{nF} \tag{7}$$

Since the sign of free energy change indicates the direction of change of a system, it is evident that comparison of the EMF of half reactions enables the prediction of the direction of various chemical pathways depending on the sign of the net free energy change. Thus electrochemistry yields a powerful tool in investigating the intermediate complexities like interspecies electron transfer and the role of communities versus enzymes in setting anode potentials for certain strains of microbes [8]. This is still an active area of research with important implications in the production of bioenergy.

Charge Transport

In the previous section, the electrochemical analysis of the microbial energy transport was presented which yielded the theoretical open circuit potential (EMF) of a cell. As soon as the cell is connected to an external circuit, current flows due to the motion of electrons and power is obtained according to the following equation:

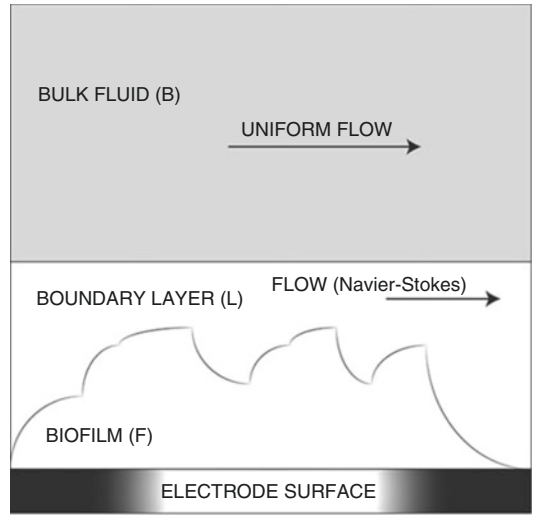
$$P = IE_{emf} \tag{8}$$

where P is the power, I is the current (often analyzed in terms of the current density j), and E_{emf} is the EMF of the cell. Flow of electrons (current flow) is related to the mechanism of charge transport as well as rate of production of electron carrier species like NADH. The production of such electron carrier species depends on the respiration process which itself depends on the diffusion of the substrates. The substrates act as the ultimate electron donors (EDs) for the entire process.

In the conduction-based mechanism, microbial nanowires play a significant role in the electron transport and hence their electrical properties are critical to charge transport processes [9]. The charge transport will be, in general, dependent on biofilm geometry, its electrical properties, as well as the electric field. Although conductivity can be treated as a biofilm material property, the geometry and potential difference depend both on the nature of electron-producing chemical reactions as well as the growth dynamics of the biofilm itself.

The amount of charge produced at any point in the biofilm depends on the concentration of the chemical species taking part in the electron-producing (and consuming) reactions. The concentration of chemical species in the biofilm is in turn governed by their mass transport dynamics. The concentration gradient of these species causes charge flow and a corresponding current determined by the electrical properties of the biofilm. The computational domain of the biofilm with boundaries is schematically shown in Fig. 4.

Mathematically, the concentration of a chemical species under an applied electric field can be expressed by the following Nernst-Planck equation [10]:



Micro/Nano Transport in Microbial Energy Harvesting, Fig. 4 Key aspects of anodic biofilm computational domain

$$\frac{\partial S_i}{\partial t} = \nabla \cdot \left(D_i \nabla S_i - \mathbf{u} S_i + \frac{D_i z_i e}{k_B T} S_i \nabla \Phi \right) + r_i \tag{9}$$

where the subscript i denotes the cardinal number of the species, S is the concentration of the dissolved substrate, D is the diffusivity, z is valence of the ionic species, $e = 1.602 \times 10^{-19}$ C is the elementary charge, $k_B = 1.38 \times 10^{-23}$ J K^{-1} is the Boltzmann constant, T is the temperature, $\mathbf{u}$ is the fluid velocity vector inside the biofilm, Φ is the potential difference corresponding to the electric field, and r_i is the species utilization rate which models the transformation processes. If the chemical species are individually electro-neutral and fluidic velocity is negligible, Eq. 9 reduces to the standard diffusion–reaction equation.

The physical interpretation of the quantity r_i is similar to the rate of a rate-limiting reaction and can have various empirical forms. If, for instance, first-order kinetics is followed, one can write:

$$r_i = k_i S_i \tag{10}$$

where k_i is the first-order rate constant and S_i is the instantaneous concentration of the ith rate-limiting species. Simple though the relation is, it suffers from runaway transformation rate with increasing concentration tending toward infinite transformation rate.

In actual systems, such infinite quantities are never encountered; instead, the reaction rates tend to a constant saturation value. The anomaly of infinite rates of first-order kinetics can be addressed by the following empirical Monod kinetics capable of showing the experimentally observed saturation-like behavior [11]:

$$r_i = r_{\max} \frac{S_i}{K_s + S_i} \tag{11}$$

where $r_{\max}$ is the maximum specific growth rate, S_i is the instantaneous concentration of the ith species, and K_s is the concentration with one half the maximum rate. It is to be noted that when species concentration is much farther away from saturation ($S_i \ll K_s$) Monod kinetics reduces to first-order kinetics as follows:

$$r_i = r_{\max} \frac{S_i}{K_s + S_i} \approx \bar{r}_{\max} S_i, \quad \bar{r}_{\max} = \frac{r_{\max}}{K_s} \text{ if } S_i \ll K_s \tag{12a, b}$$

On the other hand, the concentration gradient of the charged chemical species inside the biofilm results in a potential difference which can be written in terms of the following charge balance equation [12]:

$$\nabla \cdot \boldsymbol{j} = -\rho_{-E} \tag{13}$$

where $\boldsymbol{j}$ is the current density and $\dot{\rho}_E$ is a function denoting net time rate of charge density at a point in the biofilm implicitly dependent on the rate of reactions which produce charged species in the film. The fluidic velocity component can be solved using the incompressible Navier–Stokes equation:

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = \frac{1}{\rho_F} \nabla p + \nu \nabla^2 \mathbf{u} + \mathbf{g} \tag{14}$$

where $\mathbf{u}$ is the fluid velocity vector, $\boldsymbol{g}$ is the body force vector, ρ_F is the fluid mass density, ν is the kinematic viscosity, and p is the pressure. In addition, the incompressibility condition implies:

$$\nabla \cdot \mathbf{u} = 0 \tag{15}$$

The current density itself can be related to the potential difference in the biofilm through the Ohm's law:

$$\mathbf{j} = \kappa \nabla \Phi \tag{16}$$

where $\mathbf{j}$ is the current density vector and κ is the conductivity of the biofilm.

One can note that the mass and charge balance described above (Eqs. 9, 13, 14, 15 and 16) form a system of coupled partial differential equations (PDEs) which describe the physical processes occurring in the biofilm. In addition, the potential difference Φ inside of the biofilm is also linked to the chemistry of the respiration process through the Nernst-Planck equation described in Eq. 2.

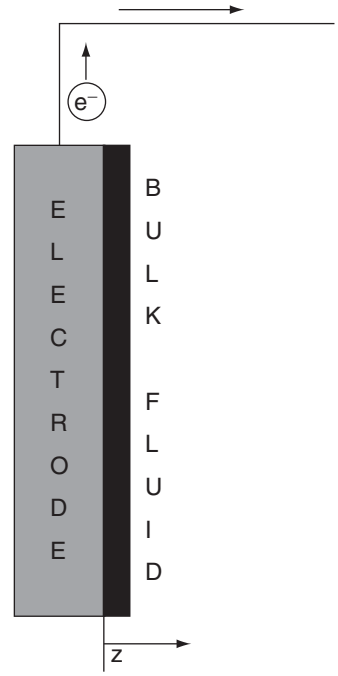
Hence the complete set of physical PDEs coupled with the chemical processes yields the complete spatial distribution of the potential difference, current density, as well as chemical species concentrations. The general solution of such a system is analytically intractable and numerical algorithms are regularly employed [13]. The total current thus generated can be computed from the current density as:

$$I = \int\int \mathbf{j} \cdot d\mathbf{A} \tag{17}$$

where I is the total current expressed as the area integral of the current density over the electrode surface.

A simplified biochemistry has been postulated by Kato and Marcus [14] in their one-dimensional, conduction-based model. The authors assume that the substrate electron donor donates its electrons to an intermediate chemical species called the electron acceptor which then deposits the electrons directly onto a solid biofilm which transfers them directly to the anode through electrical conduction. A schematic description of their simplified model is given in Fig. 5.

Another model popularized by Picioreanu and his coworkers [15] describes the charge transport assuming a mediator-based mechanism. In their model, the authors utilize the electrodic Butler-Volmer (B-V) equation to obtain currents in the biofilm. However, a complication in the said model is brought about by the stoichiometric information of the mediator species needed to solve the B-V equations. Hence, in their model a detailed description of chemical reactions is important. In addition, rather than using a one-dimensional model for diffusion, Picioreanu et al. used a more general three-dimensional model inside the biofilm.



Micro/Nano Transport in Microbial Energy Harvesting, Fig. 5 Simple one-dimensional, conduction-based energy transport

Biofilm Growth and Species Transport

The previous section described the governing PDEs which can then be integrated over the domain of the biofilm. The biofilm itself is not a static structure but continuously grows due to the consumption of nutritious substrate from the bulk liquid. In a typical MEH device, the biofilm attached to the anode is the actual energy-producing region and its growth dynamics is crucial to performance characteristics of the device.

The modeling of biofilm growth commenced in the 1970s [13] when simple one-dimensional, homogeneous models were used. Since then, biofilm modeling has encompassed increasingly sophisticated models involving both geometrical as well a functional heterogeneities to take care of the spatiotemporal multi-scale phenomena described earlier. As a biofilm grows by taking nutrients from the environment, it becomes prone to detachment due to external mechanical loads. This implies that biomass growth is enhanced by nutrient diffusion into the biofilm and impeded by the detachment phenomena. In the absence of any net external biomass flow across the biofilm, a general mass balance equation can be written as follows:

$$\dot{m}_{biomass} = \dot{m}_{production} - \dot{m}_{detachement} \tag{18}$$

where $\dot{m}_{biomass}$ denotes the net rate of biomass accumulation in the biofilm, $\dot{m}_{production}$ denotes the rate of biomass production in the biofilm, and $\dot{m}_{detachement}$ denotes the rate of detachment of biofilm due to mechanical loads.

The rate of detachment is often assumed to be proportional to the current mass of biofilm:

$$\dot{m}_{detachement} = k_{det} m_{biomass} \tag{19}$$

where k_{det} is the constant of proportionality dependent on the physics of the detachment process. The biofilm production rate depends on the respiration process which is in turn dependent on the chemistry of substrate utilization. Various models exist incorporating the chemistry of respiration into the growth model [14, 15]. In general, the production term can be assumed to have a functional dependence on substrates and biomass concentration:

$$\dot{m}_{production} = f(S_i, X_{biomass}) \tag{20}$$

where f is an appropriate production function and $X_{biomass}$ is the biomass concentration. If the biomass concentration is assumed to be constant (e.g., constant mass density of microbes) the entire rate of biomass accumulation can be directly related to the volume change by:

$$\begin{aligned} m_{biomass} &= X_{biomass} V_f, \\ \dot{m}_{biomass} &= X_{biomass} \dot{V}_f \end{aligned} \tag{21}$$

where the overdot denotes time rate of change and V is the current volume of the biofilm. The mass balance equation can now be written as:

$$\begin{aligned} \dot{V}_f &= \overline{f}(S_i) - k_{det} V_f, \\ \overline{f}(S_i) &= f(S_i, X_{biomass}) / X_{biomass} \end{aligned} \tag{22}$$

It must be noted that $\overline{f}(S_i)$ implicitly depends on the spatial coordinates due to the dynamics of respiration processes which involves the spatial gradients of the chemical species. The simplified functional form of Eq. 22 indicates a general saturation-like behavior of the biofilm growth. In other words, biofilm thickness over the anode would approximately tend to a steady state asymptotic value given sufficient time.

The above analysis implies that the differential equations derived in the earlier sections must be solved on a domain which in general changes with time. The time evolution of the domain itself depends on the variables, which are being solved in the domain creating a highly coupled set of partial differential equations. Various numerical integration schemes have been provided over the years to accurately solve these equations.

Key Research Findings

One of the key issues at hand is the competitive transport of proton and electrons taking place inside the anode chamber. In chemical terms, this means current density is affected by the pH of the solution. The chemical basis of this phenomena stems from the fundamental nature of cellular respiration. As explained before, during the process of respiration, microbes gain energy by transferring electrons from a donor to an acceptor. The electron donor generates electrons in an oxidation half reaction, while the electron acceptor consumes the electrons in a reduction half reaction. Normally, the bacteria combine the two half reactions stoichiometrically to form an overall balanced reaction. For anaerobic process taking place in anode, however, only the donor half reaction occurs, and this makes the anode biofilm's performance dependent on the pH. Excess acidity (low pH) often results from poor transport of H^+ from the biofilm into the bulk liquid from where it must pass on to the cathode chamber. This effect is shown in the work of Torres et al. [16] where a phosphate buffer was employed for the bulk fluid. The authors, from their computational models showed that the effect of pH is extremely pronounced at higher current density. The authors showed that at high current densities a slight drop in pH (increased acidity) sharply reduced the current density thus adversely affecting the performance. The total movement of H^+ ions inside the biofilm is a sum of diffusion as well as migration. Diffusion is a concentration-determined process whereas migration is

governed by the electric field. Torres et al. [16] showed that the main effect of migration comes from the buffer concentration. They found that increase in ionic concentration by adding a salt like NaCl has very little effect on the current density. Therefore, the authors concluded that anodes for MEH device applications should be designed to increase the contact between the biofilm and the bulk liquid in addition to improving the hydrodynamic conditions at the biofilm-liquid interface for better mass transport of proton out of the biofilm. The chemical effect of buffer species was further investigated by Marcus et al. [17]. The authors showed a sharp improvement in current density depending on the particular species of buffer. The effect of buffer was also shown to affect the pH profile inside the biofilm and thus indirectly control the chemical ambience of the respiration reactions. Picioreanu et al. [18] incorporated the effects of electrode morphology and bulk fluid hydrodynamics in their mediator-based, two-dimensional microbial energy transport model. The authors also confirmed the adverse effect of acidity (low pH) on the current densities. In their model which was capable of capturing the spatial distribution of the pH, the authors pointed out that mass transport through diffusion is dominant near the biofilm surface whereas electromigration is important near the electrode surface. Interestingly, they also found that increased porosity of the biofilm complex which enhanced the specific surface area did not necessarily result in higher current densities.

Future Research Direction

Mathematical modeling frameworks have two essential roles to play in the area of MEH devices. On an application-driven outlook, mathematical modeling facilitates design and maintenance of MEH devices and is thus vital to commercial adoptions of this technology. On a fundamental-level, experimentally validated models are necessary ingredients in the understanding of the complex charge-transport processes that participate in an MEH device. Today, despite strides, a complete elucidation of the interdependencies of the various mass and charge transport mechanisms is yet to be fully described. One reason is due to lack of proper understanding of all the multi-scale and multiphysics aspects of the charge-transport processes which demand a concerted effort from researchers from a number of disciplines, whereas another reason is difficulty in obtaining model parameters which have various scale-dependent transient effects.

Cross-References

- ▶ Bioadhesion
- ▶ Biofilms in Microfluidic Devices
- ▶ Micro- and Nanofluidic Devices for Medical Diagnostics
- ▶ Microfluidic Whole-Cell Biosensor

References

1. Campbell, M.K., Farrell, S.O.: Biochemistry, 6th edn. Brooks/Cole, Pacific Grove/London (2007)
2. Weeks, B.: Microbes and Society, 2nd edn. Jones and Bartlett, Sudbury (2008)
3. Nealson, K.H., Finkel, S.E.: Electron flow and biofilms. MRS Bull. **36**(5), 380–384 (2011)
4. Gorby, Y.A.: Electrically conductive bacterial nanowires produced by *Shewanella oneidensis* strain MR-1 and other microorganisms. Proc. Natl. Acad. Sci. **103**(30), 11358–11363 (2006)
5. El-Naggar, M.Y., et al.: Electrical transport along bacterial nanowires from *Shewanella oneidensis* MR-1. Proc. Natl. Acad. Sci. **107**(42), 18127–18131 (2010)
6. Wingender, J.: Microbial Extracellular Polymeric Substances: Characterization, Structure, and Function. Springer, Berlin/New York (1999)
7. Lovley, D.R.: The microbe electric: conversion of organic matter to electricity. Curr. Opin. Biotechnol. **19**(6), 564–571 (2008)
8. Logan, B.: Microbial Fuel Cells. Wiley-Interscience, Hoboken (2008)
9. Leung, K.M., et al.: Bacterial nanowires: conductive as silicon, soft as polymer. Soft Matter **7**(14), 6617 (2011)
10. Probstein, R.: Physicochemical Hydrodynamics: An Introduction, 2nd edn. Wiley, New York (1994)

11. Bae, W., Rittmann, B.E.: A structured model of dual-limitation kinetics. Biotechnol. Bioeng. **49**(6), 683–689 (1996)
12. Jackson, J.: Classical Electrodynamics, 3rd edn. Wiley, New York (1999)
13. Eberl, H.: IWA Task Group on Biofilm Modelling: Mathematical Modeling of Biofilms. IWA, London (2006)
14. Kato Marcus, A., Torres, C.I., Rittmann, B.E.: Conduction-based modeling of the biofilm anode of a microbial fuel cell. Biotechnol. Bioeng. **98**(6), 1171–1182 (2007)
15. Picioreanu, C., Head, I.M., Katuri, K.P., van Loosdrecht, M.C.M., Scott, K.: A computational model for biofilm-based microbial fuel cells. Water Res. **41**(13), 2921–2940 (2007)
16. Torres, C.I., Kato Marcus, A., Rittmann, B.E.: Proton transport inside the biofilm limits electrical current generation by anode-respiring bacteria. Biotechnol. Bioeng. **100**(5), 872–881 (2008)
17. Marcus, A.K., Torres, C.I., Rittmann, B.E.: Analysis of a microbial electrochemical cell using the proton condition in biofilm (PCBIOFILM) model. Bioresour. Technol. **102**(1), 253–262 (2011)
18. Picioreanu, C., van Loosdrecht, M.C.M., Curtis, T.P., Scott, K.: Model based evaluation of the effect of pH and electrode geometry on microbial fuel cell performance. Bioelectrochemistry **78**(1), 8–24 (2010)

Micro-/Nanodevices

► Nanotechnology

Micro-/Nanofluidic Devices

► Concentration Polarization at Micro-/Nanofluidic Interfaces

Micro-/Nanomachines

► Nanotechnology

Micro-/Nanostructured Icephobic Materials

Michael Nosonovsky
Department of Mechanical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Synonyms

Anti-icing; Deicing; Icephobicity

Definition

Icephobicity is the ability of a solid surface to repel ice or prevent ice formation due to a certain micro/nanotopographical structure of the surface and hydrophobic coating. Icephobic surfaces (i) prevent freezing of water condensing on the surface (ii) prevent freezing of incoming water, and, (iii) if ice formed, it should have weak adhesion strength with the solid so that it can be easily removed.

Occurrence and Key Findings

Although the term "icephobic" was coined in the early 1950s, only the progress in micro/nanopatterned superhydrophobic surfaces since the 2000s resulted in the emergence of micro/nanostructured icephobic surfaces which delay or suppress frost formation, bounce off incoming water droplets and, if ice still forms at the surface, reduce ice adhesion to solids [1].

Suppression of Frost Formation

Humid conditions can lead to water condensation and freezing on surfaces if the temperature is below the dew point. An important requirement for an icephobic surface is the ability to expel condensing water droplets and delay the process

of ice nucleation or frost formation on a surface. Condensation frosting occurs on a surface as a result of a formation of interdroplet frost bridges. However, so called "jumping-drop superhydrophobic surfaces" can eject condensed water droplets spontaneously using the energy released during droplet coalescence [2]. This can be achieved by designing a hierarchical superhydrophobic surface that can delay ice nucleation and suppress the interdroplet frost bridge formation.

Experiments with water droplets on superhydrophobic surfaces have shown that ice nucleation can be delayed up to 25 h for a droplet at −21 °C by carefully controlled nanoscale roughness of the surface. The coated Slippery Liquid-Infused Porous Surfaces (SLIPS) exhibited delayed ice nucleation compared to textured and hydrophobic surfaces. Water droplets condensing and coalescing on tilted SLIPS at −10 °C were seen to roll off before they could freeze [3].

Minimizing Contact Time for Incoming Droplets

Icephobic surfaces delay or prevent freezing of incoming water droplets (e.g., freezing rain) by bouncing off or breaking up the droplets. Minimizing the contact time of a droplet with a cold surface reduces the time for heat transfer and also reduces the chance of heterogeneous ice nucleation thus leading to bouncing off. Hydrophobic and superhydrophobic self-cleaning surfaces often repel incoming water droplets at room temperature. However, operation at temperatures below 0 °C poses a different challenge. As temperature decreases, the increase of the viscous force is much more significant than the increase of the surface tension forces. Therefore, a high contact angle and low contact angle hysteresis help minimizing contact time and maximizing droplet shedding.

Although the dynamic pressure of impacting droplets is small, the "water hammer-type" pressure on impact can cause the transition from the heterogeneous solid-water-air (Cassie-Baxter) to homogeneous solid-water (Wenzel) wetting regime on the surface. Droplet pinning, as a result of this wetting transition, increases the contact time as well as heat transfer through the contact area at the surface. Oblique and asymmetric impacts of droplets on hydrophobic and superhydrophobic surfaces can reduce chances of this wetting transition as well as droplet pinning [4]. Functionalized carbon nanotube superhydrophobic surfaces repel obliquely impacting water droplets at −8 °C. Nanostructured silicon superhydrophobic surface could repel low-velocity supercooled water droplets at −25 °C and thus remain ice free. Experimental studies showed that droplets freeze below −25 °C, however their removal was easy because the droplets did not undergo Cassie-Baxter to Wenzel wetting transition prior to freezing [5]. A study of hydrophobic fiber-reinforced concrete showed that it can repel obliquely impacting water droplets (Weber number ~55) at 20 °C and −5 °C [4].

Ice Adhesion to Solids

Adhesion of ice to surfaces is a result of the synergistic effect of chemical bonding, van der Wall's forces, and direct electrostatic interactions. Charges on the surface of the ice causes mirror charges in the metallic surface in contact with ice, thereby causing adhesion. Dielectrics can reduce these mirror charges, and in principle reduce ice adhesion strength. The effect of dielectric constant on ice adhesion is significant only for very low values of the dielectric constant. Polytetrafluoroethylene-coated superhydrophobic surfaces showed negligible ice adhesion. High contact angle alone does not always translate into low ice adhesion. The adhesion of ice to a superhydrophobic surface is related to the contact angle hysteresis and thus to the ice-solid interfacial area [6]. Theoretical considerations have suggested that ice adhesion is closely correlated to contact angle hysteresis, but experiments have shown this is not always the case [1]. High contact angle coupled with low contact angle hysteresis implies the presence of air pockets on the surface. These air pockets can act as stress concentrators at ice-solid interface. The size of microvoids/cracks at the ice-solid interface is critical for ice adhesion to a surface. If the cracks are not sufficiently large, the superhydrophobic surface may exhibit strong ice adhesion [7].

The latent heat released during freezing of a droplet on a surface can temporarily increase the vapor pressure in the vicinity of the solid-ice-vapor contact line. This may result in desublimation of supersaturated water vapor as a "frost halo" surrounding the three-phase contact line. The increase in ice-solid interfacial area manifests as an increase in ice adhesion to the surface [8].

Ice adhesion to a surface is quantified by the adhesion strength. A method of measuring adhesion strength of ice is by shear (or ice adhesion) test. A shear force (push or pull) is applied to the substrate-ice interface, and the ultimate shear strength at which the ice detaches is experimentally determined. Ice adhesion strength is often reduced on superhydrophobic surfaces. Ice adhesion strength to SLIPS-coated aluminum is one to two orders of magnitude less than to regular aluminum [3]. Low ice adhesion strength to lubricant-infused surfaces is due to the thick excess layer of lubricant atop the textured surface [9]. For a surface with a thermodynamically stable lubricant film (with no excess liquid completely covering the surface textures), the ice adhesion strength decreases as the surface texture density increases. A similar mechanism may be found in thermogenic icephobic plants [10].

Comparison of the Superhydrophobicity and Icephobicity

The hydrophobicity or the ability of a molecule to be repelled from a mass of water is a property that has a rigorous definition in thermodynamic terms. Thus, hydrophobicity is often compared with the lipophilicity (affinity to fats, oils, and organic liquids), and it is viewed as an "entropic" effect. In turn, the superhydrophobicity is an extreme degree of the hydrophobicity with (i) CA $>150^\circ$ and (ii) low CA hysteresis, whereas the ability to repel incoming water droplets is sometimes added as an additional independent feature of a superhydrophobic surface. Unlike the hydrophobicity, it is difficult to provide a thermodynamic definition of the icephobicity. However, most scientists agree that the property should involve (i) low ice-solid adhesion energy, (ii) a delay in freezing of supercooled droplets, and (iii) repellence of incoming water droplets by a cold surface [1].

The usual assumption is that mechanisms of ice and water adhesion are similar, and therefore, by designing a surface with low surface energy (large water contact angle), one can also achieve weak adhesion between the surface and the ice. However, ice detachment from a solid surface occurs through fracture, and therefore, it is different from the dewetting mechanism. Fracture can occur within the ice itself when the interface with the substrate is strong or at the substrate-ice interface if defects (e.g., cracks) are present. Fracture occurs in accordance with mode I (opening) or mode II (edge sliding) cracking scenarios, which correspond to normal and shear loading. The critical strength τ_{yy} above which the fracture occurs in mode I with the critical stress is related to the receding CA, θ_{rec}, as

$$\tau_{yy} = \sqrt{\frac{EG_{rec}}{\pi a}} = \sqrt{\frac{E\gamma(1 + \cos\theta_{rec})}{\pi a}} \quad (1)$$

where E is the Young modulus, G_{rec} is the surface energy of the crack during its opening, γ is the surface tension, and a is the crack length [7]. The analysis for mode II crack fracture is similar. From equation (1), we concluded that (i) shear strength correlates with the receding contact angle; (ii) the superhydrophobic surface has low shear strength because high CA, $\theta_{rec} \rightarrow 180^\circ$, corresponds to low values of $(1 + \cos\theta_{rec})$; and (iii) the effect of initial crack size is very significant. Therefore, it is important that the so-called Cassie state exists with air pockets trapped between the solid and the water droplet. When water freezes, the pockets become air voids and can serve as stress concentrators.

When a snow crystal grows, the direction of growth depends on the "roughening transition" which occurs at a certain critical temperature, so that entropy ΔS prevails over the enthalpy ΔH in the Gibbs energy ΔG formula $\Delta G = \Delta H - T\Delta S$. The combination of various directions of ice crystal growth provides the complexity of snowflake shapes and apparent "synchronization" of growth at different branches of a snowflake [1].

Micro-/Nanostructured Icephobic Materials, Table 1 The comparison of the hydrophobicity and icephobicity [1]

Property	Water	Ice
Definition of "phobicity"	Low surface energy/low adhesion	Low adhesion
	High contact angle	Low normal strength (maximum stress)
	Low CA hysteresis	Low shear strength
	Bouncing-off	Bouncing-off supercooled droplets
Interaction	Hydrophobic interaction	Crystallization of ice and snowflakes ("icephobic interactions")
Thermodynamic relationship	Minimization of $\Delta G=\Delta H-T\Delta S$	Surface-roughening transition $\Delta G=\Delta H-T\Delta S=0$ and similar
Typical manifestation of the interactions	Hydrocarbon molecules in water environment	Ice molecules in saturated vapor environment at negative temperatures
Effects	Protein folding (including fractal shape), self-organized criticality, long-range hydrophobic force, wetting transitions	Snowflakes (diverse but symmetric, fractal)

The parallelism of the superhydrophobicity and icephobicity is shown in Table 1. The three requirements for an icephobic surface (low ice normal and shear adhesion, bouncing-off droplets) are similar to the three properties for a superhydrophobic surface (high contact angle, low contact angle hysteresis, bouncing-off droplets). Both hydrophobic interactions and ice crystallization are governed by minimizing the Gibbs free energy, ΔG. Both hydrophobic and icephobic interactions are responsible for a wide range of entropic effects, from protein folding to snowflake formation.

Cross-References

► Lotus Effect

References

1. Hejazi, V., Sobolev, K., Nosonovsky, M.: From superhydrophobicity to icephobicity: forces and interaction analysis. Sci. Rep. **3**, 2194 (2013)
2. Boreyko, J.B., Collier, C.P.: Delayed frost growth on jumping-drop superhydrophobic surfaces. ACS Nano **7**, 1618–1627 (2013)
3. Kim, P., Wong, T., Alvarenga, J., Kreder, M.J., Adorno-Martinez, W., Aizenberg, J.: Liquid-infused nanostructured surfaces with extreme anti-ice and anti-frost performance. ACS Nano **6**, 6569–6577 (2012)
4. Ramachandran, R., Sobolev, K., Nosonovsky, M.: Dynamics of droplet impact on hydrophobic/icephobic concrete with the potential for superhydrophobicity. Langmuir **31**, 1437–1444 (2015)
5. Mishchenko, L., Hatton, B., Bahadur, V., Taylor, J.A., Krupenkin, T., Aizenberg, J.: Design of ice-free nanostructured surfaces based on repulsion of impacting water droplets. ACS Nano **4**, 7699–7707 (2010)
6. Kulinich, S.A., Farzaneh, M.: How wetting hysteresis influences ice adhesion strength on superhydrophobic surfaces. Langmuir **25**, 8854–8856 (2009)
7. Nosonovsky, M., Hejazi, V.: Why superhydrophobic surfaces are not always icephobic. ACS Nano **6**, 8488–8491 (2012)
8. Boinovich, L.B., Emelyanenko, A.M., Ivanov, V.K., Pashinin, A.S.: Durable icephobic coating for stainless steel. ACS Appl. Mater. Interfaces **5**, 2549–2554 (2013)
9. Subramanyam, S.B., Rykaczewski, K., Varanasi, K. K.: Ice adhesion on lubricant-impregnated textured surfaces. Langmuir **29**, 13414–13418 (2013)
10. Ramachandran, R., Nosonovsky, M.: Surface micro/nanotopography, wetting properties and the potential for biomimetic icephobicity of skunk cabbage symplocarpus foetidus. Soft Matter **10**, 7797–7803 (2014)

Micro-/Nanosystems

► Nanotechnology

Microactuators

► Basic MEMS Actuators

Microarray

Kuo-Sheng Ma[1], Yanchen Wang[2], Lucas Prater[2] and Chunlei (Peggy) Wang[3]
[1]Kennesaw State University, Marietta, GA, USA
[2]Electrical Mechatronics Engineering, Southern Polytechnic State University, Marietta, GA, USA
[3]Department of Mechanical and Materials Engineering, Florida International University, Miami, FL, USA

Synonyms

Nucleic acid amplification; Oligonucleotide synthesis

Definition

Oligonucleotide synthesis: A process to construct a piece of desired oligonucleotide sequence using a chemical additive method. It is also called de novo oligonucleotide synthesis.

Microarray: It is also called a DNA microarray or DNA chip. It usually presents in the microchip format with a variety of DNA sequences on the chip. Each spot contains multiple strands with the same sequence and forms arrays on the substrate. Biologists use DNA microarrays to genotyping biological samples or test the level of expression of the specific gene.

Introduction

Synthetic biology is an emerging field that is being applied in a wide range of disciplines. The burgeoning ideas of its applications involve gene circuits and DNA computing, biomolecule nanosensors, genome or chromosome construction, and DNA nano-robots for medical and therapeutic applications [1–10]. To achieve these goals, techniques for oligonucleotides assembly are critical. Several bioengineering methods have been developed. Among them, ligation-based assembly and PCR-based assembly are the two most commonly used methods to construct the gene-sized DNA sequences using synthesized oligonucleotides fragments as building blocks. Although these techniques are mature and routine, the bottleneck still exists on synthesis of the oligonucleotide fragments with diverse sequences. The method to chemically de novo synthesize oligonucleotides was developed in the 1980s, and the method has been commercialized and automatic. However, the capacity to provide low-cost and high-throughput products is still far behind the demands.

Microelectromechanical system (MEMS) technology has been applied in building miniaturized devices for a long time. The devices made by this technology have several benefits including low cost, less sample or reagent consumption in chemical or biological applications, easy to control and flexible expansion as well as integration in hardware. These benefits provide a potential solution to reduce cost, increase throughput, device automation and integration to the conventional oligonucleotide synthesis as well as microarray fabrications. This entry will outline several MEMS technologies that have been used on oligonucleotide synthesis and explicate their major functions in the conventional oligonucleotide synthesis.

Chemical Oligodeoxynucleotides Synthesis

The phosphoramidite reaction is currently the most widely used technique in the chemical oligonucleotide synthesis. The method was developed in the early 1980s by Caruthers et al. and has been adopted by most commercial oligonucleotide synthesizers [11, 12]. The standard phosphoramidite method is a four-step cyclic process which elongates the synthesized nucleotide chain from the 3′ end to the 5′ end. The process is initiated by tethering the first protected nucleoside on a solid support via its 3′ hydroxyl group. The 5′ end of the nucleoside is blocked via a non-reacted dimethoxytrityl (DMT) ether molecule. The commonly used solid supports are polystyrene

(PS) beads and controlled pore glass (CPG). In a single cycle of the stepwise addition of nucleotide monomers, different reagents are sequentially applied to the elongating chain.

The reactions are carried out in four steps:

1. Deprotection: The capping molecule on the 5′ end of the previous nucleotide needs to be removed to form a reactive function group on the growing chain. A weak acid is normally used in this step to remove the dimethoxytrityl (DMT) ether capping group and leaves the 5′ end hydroxyl (5′-OH) group on the chain;
2. Coupling: The desired phosphoramidite monomer couples to the chain via the 5′-OH group generated in step (1). The desired phosphoramidite monomers are activated simultaneously in solution with an appropriate activator (i.e., tetrazole) to facilitate the coupling reaction;
3. Capping: The uncoupled 5′-OH groups need to be blocked to prevent further reactions that produce error sequences. The acylating molecules are normally used in this step;
4. Oxidation: The phosphite triester linkage between two coupled oligonucleotides formed in step (2) is not strong with chemicals in subsequent reactions. An oxidizer agent is used to produce a more stable phosphotriester bond in growing chains.

This four-step cycle process is repeated to add nucleotides until the desired sequence is completed. The illustration diagram of this four-step process is depicted in Fig. 1.

In the post-synthesis process, all protection groups on the oligonucleotide chains are removed by a concentrated ammonium hydroxide solution. For DNA or gene synthesis purposes, the synthesized oligonucleotide fragments are cleaved from the solid support in the same solution with a higher reaction temperature.

The solid-phase oligonucleotide synthesis procedure has been implemented in automatic machines, and different types of synthesizers with various synthesis capacities from 1 to 1536 sequences [13–16] have been developed. In most traditional synthesizers, the required reagents are pumped into the column where the solid supports reside, and all procedures are controlled by a programmable microcontroller.

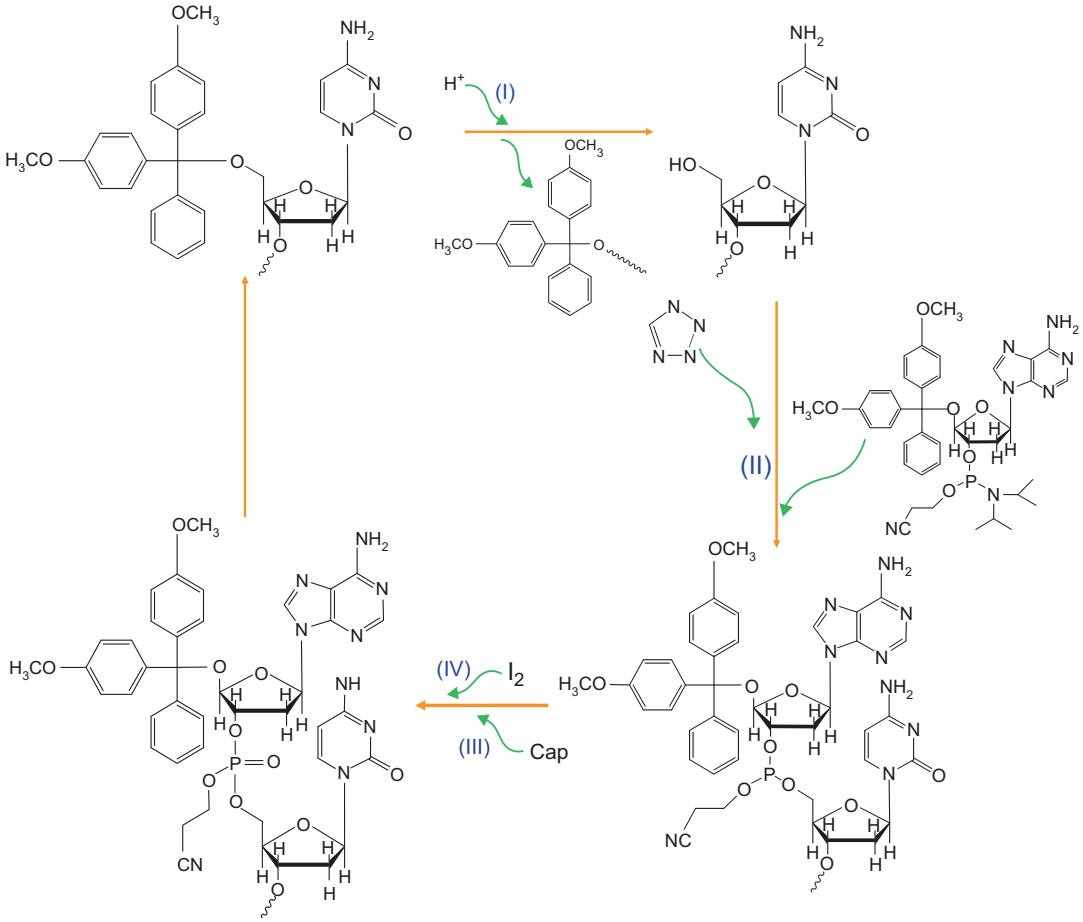
Rather than the four-step synthesis process, a two-step alternative method in oligonucleotide synthesis has been reported [17]. This two-step method simplifies the procedure by combining steps with one chemical reagent. For example, a peroxy anion agent has been used to remove 5′-end cap molecule and simultaneously transforms a phosphite triester to a more stable phosphotriester with the agent oxidation capability. This chemical reaction design has the potential for quicker and simpler automation procedures and will be a cost-efficient alternative for large-scale oligonucleotides synthesis in the future.

MEMS Technology for DNA Microarrays Fabrication

Unlike synthesizing oligonucleotides in columns, the fabrication of DNA microarrays grows oligonucleotide chains directly from the surface of substrates. Each spot contains multiple strands with the same sequence, and various sequences at different spots form the arrays. Silicon (Si) wafers, glass slides, or functionalized plastic slabs are widely used for the substrate. A conventional four-step synthesis process is also adopted yet through different mechanisms. Most of these mechanisms combine MEMS technologies to decrease the cost and further increase the throughput. Regarding the synthesis results, for microarray fabrication, oligonucleotides will remain on the substrate, but all strands can also be digested and collected from the substrate for further bioengineering applications.

Photolithography

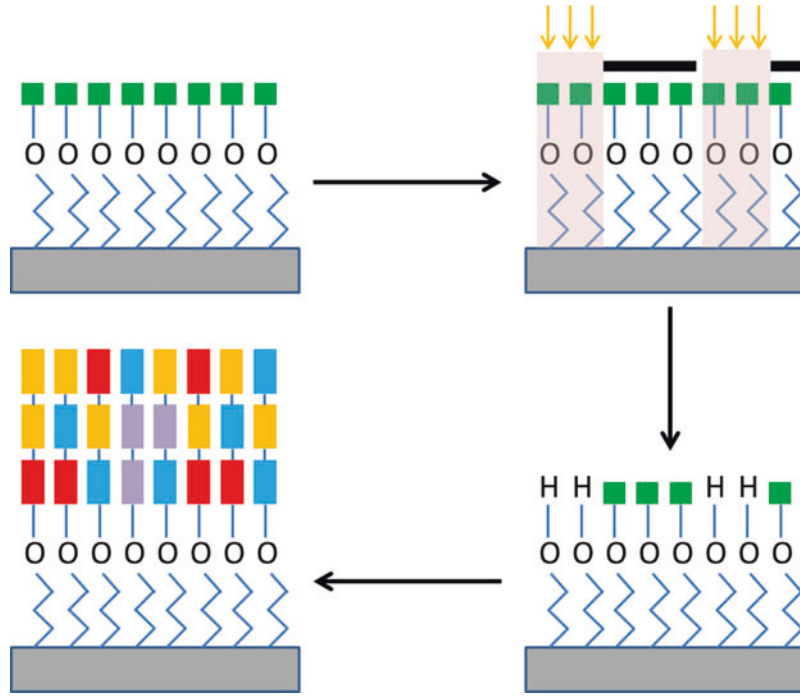
Photolithography is one the most common microfabrication procedures in constructing micron or submicron devices. In this technique, the process is exclusively used on pattern transfers through a mask to the thin-film polymer on the



Microarray, Fig. 1 Cycle of standard "phosphoramidites" oligonucleotides synthesis

substrate via a suitable light source. The masks are used to generate light patterns that dictate exposure and non-exposure areas on the polymer and form desired patterns. The concept of this technique has been applied to oligonucleotide synthesis on the microarrays [18–20]. The technique employed masks and photolabile nucleoside monomers to grow oligonucleotides in arrays on a solid substrate. In an early system, physical masks were used to generate light patterns. The light exposure on the specific array area removes photolabile capping groups from oligonucleotide chains and results an active functional terminal (i.e., deblocking). Afterward, the desired phosphoramidite monomer is applied to all arrays on the entire substrate surface. The coupling of the oligonucleotide monomer only occurs in arrays which were light exposed in the previous step and add oligonucleotides to the elongation chains. The steps repeat until all desired sequences are synthesized (Fig. 2). To perform the synthesis on diverse oligonucleotide sequence, a stack of masks with various light exposure patterns needs to be pre-manufactured based on the desired oligonucleotide sequences. Affymetrix Inc. pioneered this technique and employed the technology to fabricate the high-density probe arrays product GeneChip™ for gene sequence analysis.

As mentioned, in the previous technique, a stack of physical masks needs to be fabricated. Therefore, the cost to use this technique to synthesize oligonucleotides dramatically increases when the



Microarray, Fig. 2 Schematic illustration of DNA synthesis using photolithography technique

sequence diversity increases. In addition, in order to increase the sequence diversity in one chip, the feature size of each array needs to be decreased to increase the array density. The cost for masks is also increased when the pattern feature size decreases. Therefore, the technique is only justifiable for large-scale, fixed-sequence and high-volume microarray manufacturing. Recently, a low-cost alternative method using virtual masks to perform light patterning in the deprotection step has been developed. The technique employed "Digital Light Processing" (DLP™) technology with "Digital Micromirror Device" (DMD™) developed by Texas Instruments. The device consists of a high-density micromirror array manufactured by MEMS technology that is used as a light-reflective display by electromechanically switching each micromirror between on and off positions. Thus, the device generates the specific light pattern digitally with a computer control. Compared with using the physical masks in photolithography, the use of this device avoids the tremendous cost in physical mask fabrications. In addition, the light patterns are easily programmed through software to match the desired sequences.

In regard to chemical developments, progress has been made in DNA microarray fabrication using the photolithography technique. Chemicals such as (R,S)-1-(3,4-(methylenedioxy)-6-nitrophenyl)-ethyl chloroformate (MeNPOC) or 2-(2-nitrophenyl)propoxycarbonyl (NPPOC) have been reported [21, 22]. In another report, an alternative technique used the photo acid generator (PAG) agent to produce acids in the reaction solution when exposed to light. The generated protons (i.e., H^+) are used to eliminate the protection group on the phosphoramidite monomer and generate an active 5′-OH terminal in the oligonucleotide chain as in the conventional phosphoramidite reactions. Gao et al. [23] have successfully demonstrated using such methods in oligonucleotide synthesis.

Electrochemical Array

Instead of using photo-generated acids for the deprotection step, Egeland et al. reported using a hydrolysis reaction in an electrolyte solution to generate acids [24]. In this technique, the authors

used thin-film iridium metal to fabricate arrays of microelectrodes and used hydroquinone and benzoquinone with tetrabutylammonium hexafluorophosphate in anhydrous acetonitrile as the reaction electrolyte. When a voltage was applied to the microelectodes, a corresponding current was applied to the electrolyte. An oxidation reaction occurred at the anodes and produced acids in the solution. The acids were confined to the anode region by the adjacent cathode electrodes and directly diffused to the area of the array where the deprotection reaction would be performed. In this technique, the electrode needs to be separated from the substrate at which the oligonucleotide synthesized to prevent an oxidation reaction on oligonucleotide chains. An alternative technique performed synthesis directly on the electrode surfaces and used a porous polymer layer to protect oligonucleotide chains from oxidation damages. In this technique, due to the slow diffusion of the acid molecules in the porous polymer, the overall time for the whole synthesis procedure is extended.

Microfluidics

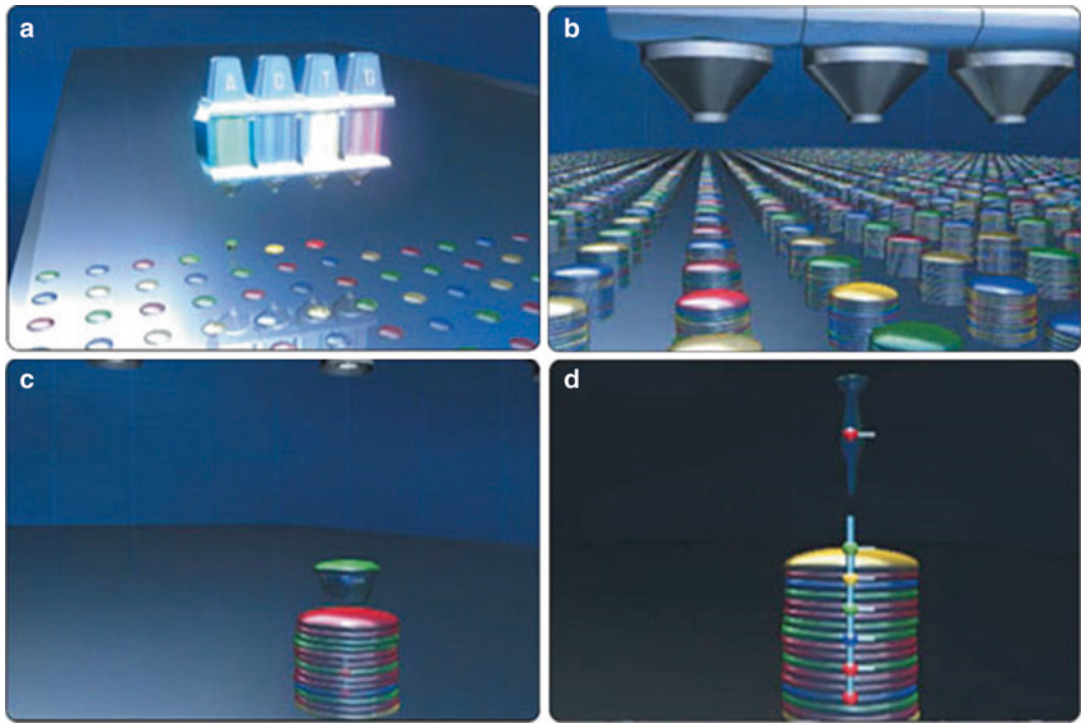
Microfluidic devices have been widely used in various applications such as the chemistry, biology, and medical fields. Although microfluidic devices can be fabricated with various materials such as silicon, glass, or quartz, the technique to fabricate these devices using elastomeric materials is most popular. This is due to the characteristics of the material such as ease and flexible fabrication processes, low cost, and versatile material properties. From application aspects, due to the miniaturized size of device components (i.e., channels and wells), the amount of sample and reagents required for each application is dramatically reduced, and the throughput is increased. Recently, several groups reported adopting microfluidic devices in microscale oligonucleotide synthesis to reduce the total cost. In those studies, the microfluidic devices were used to transport reagents used for each synthesis step, and oligonucleotides were synthesized in the micro-wells, which were fabricated by etching on the substrate. Several technical challenges have been discussed including individual addressability to each synthesis array, processes robustness and reliability, and the chemical resistance of the elastic materials used in microfluidic fabrication. Most plastic devices lost sustainability upon contact with organic solvents or harsh chemicals which are all used in synthesis procedures. For example, devices made of poly(dimethylsiloxane) (PDMS), a popular polymer material used in most microfluidic device fabrication, became swollen and had their channels clogged during synthesis while the material was in contact with the organic solvents. Moorcroft et al. [25] first attempted to solve this dilemma by modifying the synthesis protocol. They changed the solvent used in the deprotection step and the reagent used in the oxidation step. In addition, in order to immobilize linkers on the PDMS substrate, a surface treatment technique was developed. First, the microchannel surface was oxidized via a UV-ozone system to introduce a functional silanol (Si-OH) group. A subsequent silanization step was performed on the oxidized surface using vapor phase 3-glycidoxypropyltrimethoxysilane (GPTMS). Finally, PEG was introduced to link to the silane as a spacer to connect the first oligonucleotide for whole synthesis. The author reported 21-mer oligonucleotides that were successfully synthesized with this microfluidic device.

Several other strategies have also been developed to avoid PDMS swelling when it is used as the material in microfluidic oligonucleotide synthesis. Hua et al. used parylene to coat on PDMS and form a highly chemical-resistant hybrid film in microfluidic devices for the oligonucleotide synthesis [26]. Quake et al. reported using perfluoropolyether (PFPE), a high chemical-resistant polymer material, to replace PDMS as the material for fabricating the microfluidic devices for oligonucleotide synthesis. In their method, porous silica beads were used as the solid support, and all the same synthesis reagents were used in the conventional phosphoramidite method [27]. This work also demonstrated that microfluidic device consumed 60-fold less reagents as compared with the conventional column apparatus.

Inkjet Printing

Another example using the MEMS technique in oligonucleotide synthesis is inkjet printing. In this technique, an inkjet printer is employed to deliver liquid droplets of reaction reagents to specific spots on a substrate to complete the four-step cycle in the conventional oligonucleotide synthesis procedure. A piezoelectric printer head, fabricated using MEMS techniques, consists of ink reservoirs and is controlled by a piezoelectric crystal. When a voltage is applied to the crystal, the lateral contraction force causes the attached diaphragm to deflect and ejects a drop of liquid from the reservoir. The printing head is controlled by motion controllers to precisely "print" the liquid at specific areas on the substrate. Blanchard et al. first demonstrated using this mechanism to synthesize oligonucleotides [28]. In their system design, the inkjet head consisted of six liquid reservoirs with four different phosphoramidite monomers, one for activators and one for modified base solution. Afterward, other groups also demonstrated using commercial piezoelectric inkjet printers to fabricate oligonucleotide microarrays [29, 30].

One of the major technical issues discovered in this technique is that the open environment caused fast evaporation of the pico-liter liquid on the substrate. Acentonitrile is the major solvent chosen in conventional oligonucleotide synthesis. However, when using it in printing synthesis mechanism, the printing head nozzle became clogged, and crystallized dry reactants on the substrate were discovered, which deteriorated the whole synthesis process. Therefore, the solvent selection is critical. Lausted et al. [29] developed a technique using propylene carbonate (PC) with a mixture of 2-methyl glutaronitrile (MGN) and 3-methoxypropionitrile (3-MP) as the solvents in the whole synthesis procedure. These solvents showed low volatility, high vapor pressure, and



Microarray, Fig. 3 Cartoon of DNA microarray synthesis with inkjet printing technology by Agilent Technologies Inc. [37]

full compatibility with all phosphoramidite chemicals which are suitable in all oligonucleotide synthesis. Agilent Technologies Inc. has commercialized the technique to fabricate microarray for gene-expression profiling applications (Fig. 3) [31–32].

Conclusion

Oligonucleotide synthesis is the foundation of many DNA technologies including biomolecular engineering, gene circuits, nano-devices and computing, synthetic biology, and genome synthesis [1–4, 33]. However, one of the biggest barriers to commercial applications of oligonucleotides synthesis still remains on making primers for PCR (i.e., polymerase chain reaction) and mutagenesis. The reasons are high cost in the units priced for each synthesized base and low-throughput techniques used in the market with conventional synthesis methods or machines.

With the benefits from microfabrication technology, many novel synthesis technologies have been developed. As we mentioned previously, those techniques have brought benefits in low-cost device manufacturing, fewer amounts of reagents used, fast and flexible sequence synthesis, and batch or high-throughput production to the oligonucleotide synthesis and further to all DNA technology. For example, in 2002, the de novo synthesis of a functional poliovirus genome was achieved. The size of it is around 7500 bp [34]. In 2003, a functional bacteriophage genome (5386 bp) was synthesized in 2 weeks [35]. The synthesis of a genome of equal size needed months with earlier technologies. In 2008, a breakthrough genome synthesis and cloning of the *Mycoplasma genitalium* (582970 bp) was completed [36] which demonstrated a milestone in the synthetic biology discipline. It is believed that with the continued evolution of microfabrication and material technology, full automation and platform integration, further decreases in unit price, and increases in throughput, de novo DNA synthesis will become a standard technique for all DNA engineering applications.

Cross-References

- ▶ BioMEMS/NEMS
- ▶ Digital Microfluidics
- ▶ Functionalization
- ▶ Lab-on-a-Chip
- ▶ MEMS
- ▶ Micro-/Nanofluidic Devices
- ▶ Microarrays/Microfluidics for Protein Crystallization
- ▶ Microelectrodes
- ▶ Microfluidic Device
- ▶ Patterning
- ▶ Photolithography
- ▶ Synthetic Biology
- ▶ Synthetic Genomics

References

1. Martin, V.J., Pitera, D.J., Withers, S.T., Newman, J.D., Keasling, J.D.: Engineering a mevalonate pathway in *Escherichia coli* for production of terpenoids. Nat. Biotechnol. **21**, 796–802 (2003)
2. Dueber, J.E., Yeh, B.J., Chak, K., Lim, W.A.: Reprogramming control of an allosteric signaling switch through modular recombination. Science **301**, 1904–1908 (2003)
3. Park, S.H., Zarrinpar, A., Lim, W.A.: Rewiring MAP kinase pathways using alternative scaffold assembly mechanisms. Science **299**, 1061–1064 (2003)
4. McDaniel, R., Weiss, R.: Advances in synthetic biology: on the path from prototypes to applications. Curr. Opin. Biotechnol. **16**, 476–483 (2005)
5. Kobayashi, H., Kærn, M., Araki, M., Chung, K., Gardner, T.S., Cantor, C.R., Collins, J.J.: Programmable cells: interfacing natural and engineered gene networks. Proc. Natl. Acad. Sci. U. S. A. **101**, 8414–8419 (2004)
6. Bayer, T.S., Smolke, C.D.: Programmable ligand-controlled riboregulators of eukaryotic gene expression. Nat. Biotechnol. **23**, 337–343 (2005)
7. Benner, S.A., Sismour, A.M.: Synthetic biology. Nat. Rev. Genet. **6**, 533–543 (2005)
8. Elowitz, M.B., Leibler, S.: A synthetic oscillatory network of transcriptional regulators. Nature **403**, 335–338 (2000)
9. Kodumal, S.J., Patel, K.G., Reid, R., Menzella, H.G., Welch, M., Santi, D.V.: Total synthesis of long DNA sequences: synthesis of a contiguous 32-kb polyketide synthase gene cluster. Proc. Natl. Acad. Sci. U. S. A. **101**, 15573–15578 (2004)
10. Tian, J., Gong, H., Sheng, N., Zhou, X., Gulari, E., Gao, X., Church, G.: Accurate multiplex gene

synthesis from programmable DNA microchips. Nature **432**, 1050–1054 (2004)
11. Caruthers, M.H., Beaucage, S.L., Becker, C., Efcavitch, J.W., Fisher, E.F., Galluppi, G., Goldman, R., de Haseth, P., Matteucci, M., McBride, L., Stabinsky, Y.: Deoxyoligonucleotide synthesis via the phosphoramidite method. Gene. Amplif. Anal. **3**, 1–26 (1983)
12. Caruthers, M.H., Barone, A.D., Beaucage, S.L., Dodds, D.R., Fisher, E.F., McBride, L.J., Matteucci, M., Stabinsky, Z., Tang, J.-Y.: Chemical synthesis of deoxyoligonucleotides by the phosphoramidite method. Methods Enzymol. **154**, 287–313 (1987)
13. Horvath, S.J., Firca, J.R., Hunkapiller, T., Hunkapiller, M.W., Hood, L.: An automated DNA synthesizer employing deoxynucleoside 30-phosphoramidites. Methods Enzymol. **154**, 314–326 (1987)
14. Sindelar, L.E., Jaklevic, J.M.: High-throughput DNA synthesis in a multichannel format. Nucleic Acids Res. **23**, 982–987 (1995)
15. Lashkari, D.A., Hunickesmith, S.P., Norgren, R.M., Davis, R.W., Brennan, T.: An automated multiplex oligonucleotide synthesizer-development of high-throughput, low-cost DNA synthesis. Proc. Natl. Acad. Sci. U. S. A. **92**, 7912–7915 (1995)
16. Cheng, J.Y., Chen, H.H., Kao, Y.S., Kao, W.C., Peck, K.: High throughput parallel synthesis of oligonucleotides with 1536 channel synthesizer. Nucleic Acids Res. **30**, e93 (2002)
17. Caruthers, M.H.: Gene synthesis machines: DNA chemistry and its uses. Science **230**, 281–285 (1985)
18. Fodor, S.P.A., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D.: Light-directed, spatially addressable parallel chemical synthesis. Science **251**, 767–773 (1991)
19. Southern, E.M., Maskos, U., Elder, J.K.: Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. Genomics **13**, 1008–1017 (1992)
20. Barone, A.D., Beecher, J.E., Bury, P.A., Chen, C., Doede, T., Fidanza, J.A., McGall, G.A.: Photolithographic synthesis of high-density oligonucleotide probe arrays. Nucleosides Nucleotides Nucleic Acids **20**, 525–531 (2001)
21. Singh-Gasson, S., Green, R.D., Yue, Y., Nelson, C., Blattner, F., Sussman, M.R., Cerrina, F.: Maskless fabrication of light directed oligonucleotide microarrays using a digital micromirror array. Nat. Biotechnol. **17**, 974–978 (1999)
22. Richmond, K.E., Li, M.-H., Rodesch, M.J., Patel, M., Lowe, A.M., Kim, C., Chu, L.L., Venkataramaian, N., Flickinger, S.F., Kaysen, J., Belshaw, P.J., Sussman, M.R., Cerrina, F.: Amplification and assembly of chip-eluted DNA (AACED): a method for high-throughput gene synthesis. Nucleic Acids Res. **32**, 5011–5018 (2004)
23. Gao, X.L., LeProust, E., Zhang, H., Srivannavit, O., Gulari, E., Yu, P., Nishiguchi, C., Xiang, Q., Zhou, X.: A flexible light directed DNA chip synthesis gated by deprotection using solution photogenerated acids. Nucleic Acids Res. **29**, 4744–4750 (2001)
24. Egeland, R.D., Southern, E.M.: Electrochemically directed synthesis of oligonucleotides for DNA microarray fabrication. Nucleic Acids Res. **33**, e125 (2005)
25. Moorcroft, M.J., Meuleman, W.R., Latham, S.G., Nicholls, T.J., Egeland, R.D., Southern, E.M.: In situ oligonucleotide synthesis on poly(dimethylsiloxane): a flexible substrate for microarray fabrication. Nucleic Acids Res. **33**, e75 (2005)
26. Hua, Z.S., Xia, Y., Srivannavit, O., Rouillard, J.-M., Zhou, X., Gao, X., Gulari, E.: A versatile microreactor platform featuring a chemical-resistant microvalve array for addressable multiplex syntheses and assays. J. Micromechan. Microeng. **16**, 1433–1443 (2006)
27. Huang, Y.Y., Castrataro, P., Lee, C.C., Quake, S.R.: Solvent resistant microfluidic DNA synthesizer. Lab Chip **7**, 24–26 (2007)
28. Blanchard, A.P., Kaiser, R.J., Hood, L.E.: High-density oligonucleotide arrays. Biosens. Bioelectronics **11**, 687–690 (1996)
29. Lausted, C., Dahl, T., Warren, C., King, K., Smith, K., Johnson, M., Saleem, R., Aitchison, J., Hood, L., Lasky, S.R.: POSaM: a fast, flexible, open-source, inkjet oligonucleotide synthesizer and microarrayer. Genome Biol. **5**, R58 (2004)
30. Lausted, C.G., Warren, C.B., Hood, L.E., Lasky, S.R.: Printing your own inkjet microarrays. Methods Enzymol. **410**, 168–189 (2006)
31. Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H., Linsley, P.S.: Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nat. Biotechnol. **19**, 342–347 (2001)
32. Fisher, W., Mingjun, Z.: A biochip microarray fabrication system using inkjet technology. IEEE Trans. Autom. Sci. Eng. **4**, 488–500 (2007)
33. Fishman, A., Tao, Y., Rui, L., Wood, T.K.: Controlling the region specific oxidation of aromatics via active site engineering of toluene para-monooxygenase of *Ralstonia* pickettii PKO1. J. Biol. Chem. **280**, 506–514 (2005)
34. Cello, J., Paul, A.V., Wimmer, E.: Chemical synthesis of poliovirus cDNA: generation of infectious virus in the absence of natural template. Science **297**, 1016–1018 (2002)
35. Smith, H.O., Hutchison III, C.A., Pfannkoch, C., Venter, J.C.: Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. Proc. Natl. Acad. Sci. U. S. A. **100**, 15440–15445 (2003)

36. Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M.A., Merryman, C., Young, L., Noskov, V.N., Glass, J.I., Venter, J.C., Hutchison III, C.A., Smith, H.O.: Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. Science **319**, 1215–1220 (2008)
37. http://www.genomics.agilent.com/article.jsp?pageId=2011&_requestid=778696

Microarrays/Microfluidics for Protein Crystallization

▶ Macromolecular Crystallization Using Nanovolumes

Microbial Aggregations

▶ Biofilms in Microfluidic Devices

Microbial Electrolysis Cell

▶ Micro/Nano Transport in Microbial Energy Harvesting

Microbial Fuel Cell

▶ Micro/Nano Transport in Microbial Energy Harvesting

Microcalorimetry

▶ Nanocalorimetry

Microcantilever Chemical and Biological Sensors

Hashem Etayash and Thomas Thundat
University of Alberta, Edmonton, AB, Canada

Definition

Advanced sensing techniques are currently being developed so that quick and accurate detection of chemical and biological analytes can be obtained. Among those techniques, microcantilever Array sensors, have attracted considerable attention as a result of their potential as highly sensitive platforms for real-time, high-throughput, and multiplexed detection of a myriad of chemical and biological analytes. Microcantilevers are microscale beams which are anchored at one end. Molecular adsorption on a cantilever surface can generate bending as a result of adsorption-induced changes in surface stress. In addition, mass loading due to molecular adsorption results in changes in the resonance frequency of the cantilever. Chemical/biological selectivity can be achieved via immobilized receptors on the cantilever surface. The microcantilever array platform has the potential for application in many areas such as monitoring of water quality, chemical analysis, detection of food pathogens, and medical point-of-care diagnosis.

Introduction

A need currently exists for miniature devices to use in chemical and biological analysis, which can provide fast and accurate diagnoses in real time. The necessity for rapid and sensitive sensing with high selectivity in a multiplexed fashion is becoming urgent as multiple markers are required for accurate detection [1, 2]. The contamination of food, water, and plants by pathogenic bacteria, poisons, and toxic chemicals is causing concern across the world [3, 4]. Survival rates depend on early detection and treatment, both of which are

difficult to achieve when a diagnosis is based on slow, non-movable equipments. Therefore, the development of inexpensive handheld chemical/biological sensors with high selectivity and sensitivity and low power consumption is crucial for early detection in the field.

Conventional medical diagnostic tests [5], such as colony-based platforms (blood, biopsy, saliva culture, or food/product contaminated cultures), are efficient, sensitive, and accurate. However, they are time consuming, inconvenient, and cannot be miniaturized for use in field applications. High-throughput microarrays such as DNA and protein microarrays are prevalent and versatile tools in detection, but they require time- and expensive steps of labeling using fluorescent tags [6, 7]. In existing clinical diagnostic methods, results are not received quickly resulting in treatment delays, and this can have a negative impact on a patient's health and overall long-term treatment outcomes [8, 9]. Delays in the accurate detection of food contamination can affect public health while discovering and treating outbreaks. There is an urgent need for the development of cost-effective, and rapid diagnostic tools that can be mass deployed.

Miniature sensors for rapid detection are increasingly finding applications in fields such as chemical analysis, drug discovery, blood analysis, glucose monitoring [10], microbial detection, food/water analysis, and medical diagnostics [11, 12]. Chemical/biosensors are basically devices with engineered interfaces that respond to specific analytes in complex environments by measuring the variations in the physical properties of the sensing elements [13]. Generally, chemical and bio interfaces for molecular recognition are derived from receptor molecules which selectively interact with target molecules under specific conditions. Miniature sensor technologies that allow simultaneous sensing of multiple chemical and biological entities in a single experiment are, therefore, becoming essential tools in drug discovery, clinical detection, safety, and life science research. A typical sensor contains three main components: a bio-/chemical recognition element, a transduction component, and a response detector. Based on the transduction mechanisms, sensors can be classified into optical, electrochemical, electrical, and electromechanical [14]. The bio-recognition interfaces used in the biosensors are usually receptors, antibodies, proteins/enzymes, nucleotides, and other chemical recognition molecules. These elements are attached physically or chemically onto the surface of the physical sensor, and their interactions with the target analytes lead to specific sensor responses, which can be displayed using simple read-out methods.

The microcantilever, which resembles a miniature diving board, can be readily fabricated on silicon wafers and other materials using conventional microfabrication techniques. Originally, microcantilevers were introduced as displacement detection probes in atomic force microscopy (AFM) for acquiring surface topographies [15]. Soon afterwards, it was recognized that they could be used to sense changes in temperature and relative humidity [16]. The ability to fabricate arrays of extremely small structures using standard wafer-scale semiconductor processing techniques has allowed the development of practical micromechanical chemical sensors and biosensors. The microcantilevers are displacement sensors which respond to surface energy changes as a result of chemical or biomolecular interactions. Restricting adsorption of the molecules on one side of the cantilever surface results in a surface stress that can be observed as a deflection. In addition to bending, the mechanical resonance frequency of the cantilever can undergo variations due to molecular adsorption-induced mass loading. These two signals, cantilever bending and resonance frequency variation, can be monitored simultaneously using a variety of techniques.

In the mid-1990s, the ability to functionalize cantilevers with molecular probes for the generation of selective responses to targeted biochemical compounds opened a new field of biosensor technology based on electromechanical properties and surface structures [17, 18]. Despite it being relatively new, the microcantilever detection methodology has attracted significant attention as an approach for label-free sensing device. Microcantilever platform has the advantage as a

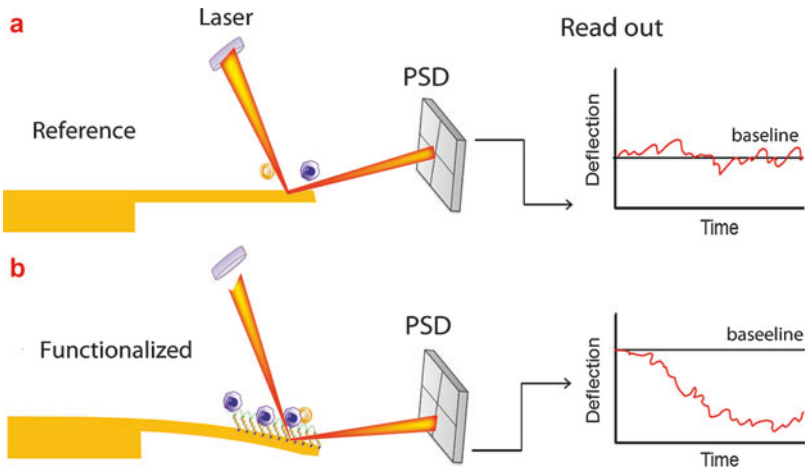
label-free detective method that is extremely sensitive, reliable, and scalable. Recent studies have shown that cantilevers are capable of sensing biomolecules in liquid sat concentrations of nano- to femtomolars [19]. Microcantilevers have been used for detection of numerous chemical and biological entities including nucleotides [20, 21], proteins [22], microbes [23, 24], as well as a wide range of toxic chemicals [25] and heavy metals [26]. Nanomechanical cantilevers have demonstrated mass resolution of a zeptogram mass when operated in a vacuum. However, the mass resolution is only in the nanogram range when operated in liquid media due to the liquid damping. The mass resolution depends on the frequency of the cantilever. The higher the resonance frequency of the cantilevers, the higher the mass resolution. The ultra-small size of the cantilevers allows the sensor to exhibit quick responses to the evolving biological and chemical deviations in timescales of micro- and milliseconds.

Cantilever Principle and Modes of Operation

Microcantilevers are beams fabricated from materials such as silicon using conventional microfabrication techniques [27]. Typical dimensions of silicon microcantilevers are 20–50 μm (width), 100–500 μm (length), and 0.3–2 μm (thickness). Nanoscale cantilevers, on the other hand, have dimensions of 600–950 nm (width), 300–650 μm (length), and 50–100 nm (thickness). For most biochemical applications, cantilevers are first vacuum deposited with a thin layer of gold (20–30 nm) on one side. The gold-coated side of the cantilever is functionalized with ligand molecules which are specific to the target analytes. Many different attachment techniques are used in functionalizing the cantilever, such as, the alkane thiol binding chemistry. During the detection process, the immobilized ligands selectively capture the target molecules from the surrounding medium, causing the cantilevers to bend, and/or change the resonance frequency, depending on the mode of operation. Since liquids heavily damp the resonance frequency, this approach is not routinely used in sensing. Adsorption-induced cantilever bending is largely independent of the medium; therefore, it is routinely used as a sensing method. The specificity of detection relies deeply on the molecular interaction between the predesigned detective agents (ligands) and the unknown sample molecules (targets). Figure 1 illustrates the basic principle of the cantilever sensor technique based on beam deflection.

Modes of Operation

- Static Mode of Operation (Surface Stress Mode)
Microcantilevers are routinely operated in a static deformation mode, where the cantilever beam



Microcantilever Chemical and Biological Sensors, Fig. 1 Basic principle of cantilever technique; (**a**) no deflection from a reference cantilever (no binding to the target analyte); (**b**) cantilever deflection as a result of the ligand-analyte binding

bends as a result of increasing or decreasing the surface stress generated by analyte-ligand interactions on the sensor. While an increase in the surface stress (compressive stress) leads to downward bending of the cantilever, a decrease in the surface stress (tensile stress) leads to upward bending of the cantilever. Since the cantilever bending reaches a steady state after attaining equilibrium during adsorption, this operational mode is called the static deformation mode. Specific adsorption of the molecules on one side of the cantilever results in a decrease in the surface free energy. The surface free energy per unit area is the surface stress and preferential molecular adsorption on one side of the cantilever which results in differential surface stress. The difference in surface stress between the functionalized surface and the passive (nonfunctionalized) surface leads to cantilever deflection (bending), the value of which can be estimated according to Stoney's formula:

$$\Delta\sigma = \left[\frac{Et^2}{3\,(1-\nu)}\,L2\right]\Delta \qquad (1)$$

where $\Delta\sigma$ is the change in surface stress (or surface energy) due to the molecular adsorption, E is the Young's modulus of the cantilever, ν is the Poisson ratio, and L and t are the length and thickness of the cantilever, respectively. From Eq. 1, it is clear that a longer cantilever generates larger bending.

- Dynamic Mode of Operation (Resonance Mode)
The dynamic mode of operation of the cantilever is based on resonance frequency of the cantilever. Adsorption of specific molecules on the cantilever surface leads to a change in the cantilever mass. The resonance parameters include resonance frequency, amplitude of vibration, quality factor (Q), and phase of vibration. The fundamental resonance frequency f_0 of a cantilever is expressed as

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{nm}}$$

where k is the spring constant of the lever, n is a geometrical parameter for the fundamental vibration of the cantilever beam, and m is the effect of mass of the cantilever.

- Bi-material Mode of Operation (Thermal Mode)
When a microcantilever is deposited with a thin layer of metal on one of its sides, it exhibits bi-material (bimetallic) properties. In the bi-material mode, the cantilever undergoes a static deflection in response to changes in temperature. The bending of a bi-material cantilever is a result of the different coefficients of thermal expansion between the metallic thin film and the underlying cantilever. This mode of operation is also referred to as the heat mode of operation due to its dependence on changes in temperature. A bi-material cantilever is an extremely sensitive thermometer. This mode is widely used for temperature-related applications such as calorimetry and thermal actuators. The temperature changes on the surface of the cantilever can be driven by catalytic reactions. The thermal sensitivity of the cantilever heating mode is orders of magnitude higher than that of conventional calorimetric methods.

Cantilever Surface Functionalization

The selectivity of a microcantilever sensor originates from the selectivity of receptor-analyte binding. The receptors are immobilized on the cantilever surface using surface functionalization techniques. The attachment of chemical or biological receptor molecules (proteins, DNA, cells, and other semiselective biochemical elements) has drastically enhanced the selectivity of several biosensors, including microcantilevers in molecular recognition applications [28, 29]. One of the advantages of surface functionalization is its ability to detect specific targets in an unknown mixture.

There are many approaches that can be used to functionalize the cantilever with specific recognition molecules depending on what the final applications of the sensor will be. On silicon cantilevers for instance, functionalization can be performed using so-called silicon chemistry [30], based on coating the cantilevers with polylysine, aminosilane, epoxysilane, or nitrocellulose.

On gold-coated cantilevers however, the functionalization can be easily implemented using thiol chemistry [31],which takes advantages of the strong covalent interaction between Au and the thiol group of the ligand molecules. The challenge of these approaches is developing a healthy reaction chemistry which can ensure the required chemical selectivity, stability, and complete bioactivity of the attached molecules. The immobilization mechanism is determined both by the chemical nature of the molecular probes and their reactive functional groups. In the case of the surface stress-based operating mode, the functionalization must be carried out only on one of the sides of the cantilever, not both. Failure to do so can result in minimal differential surface stress and insignificant cantilever deflection. To prevent that from happening, the inactive surface of the cantilever should be made passive by coating it with a material that does not interact with the analyte or other solution components. Typically in microcantilevers, one surface is made passive by salinization or grafting with inert polymer substances such as polyethylene glycol.

Functionalizing individual cantilevers is an array with different receptors for multiplexed detection is a challenge. For a single cantilever, the process of immobilization is performed simply by dipping the cantilever in a solution contains the receptor molecules. Although the procedure is simple and easy, it is not applicable to individually functionalize a cantilever in an array. Cantilever arrays are useful in multiplexed detection where a specifically functionalized individual cantilever detects a specific analyte molecule in a mixture. In general, cantilevers are operated in a differential mode where a chemically or biologically inert cantilever serves as a reference. Reference cantilevers are important since a typical cantilever is sensitive to many environmental factors, such as flow rate, mechanical vibrations, temperature, pH, etc. Many methods are currently under development to functionalize individual cantilevers in an array without affecting the neighboring ones. One of these techniques includes cantilever insertion into dimensionally matched capillary arrays containing specific receptors. Techniques based on inkjet printing are also under development. Other approaches that have been used recently include photolithographic approach, liquid drop dispensing tools, and microcontact printing.

Microcantilever as Chemical and Biological Sensors

Microcantilever sensors continue to find applications in the fields of chemistry and biomedical sciences. In recent years, cantilever biosensors have broadened their applications to include diagnostics. They have been used to detect chemical agents, volatile compounds, heavy metals, and toxic industrial gases [32]. However, specificity in detection has been a challenge. To obtain high selectivity, the cantilever has to be coated with a thin layer of materials that can interact specifically to the target analytes. Coating the cantilever with a thin layer of certain metals allows a semiselective sensing of certain toxic gases. A gold-coating layer, for instance, is used to detect mercury and thiol impurities [32]. Likewise, a palladium thin layer coating allows a semiselective detection of hydrogen [32]. Water vapor can also be sensed using a layer of hydrophilic polymers such as polyvinyl alcohol, carboxymethyl cellulose, or polyvinylpyrrolidone [32].

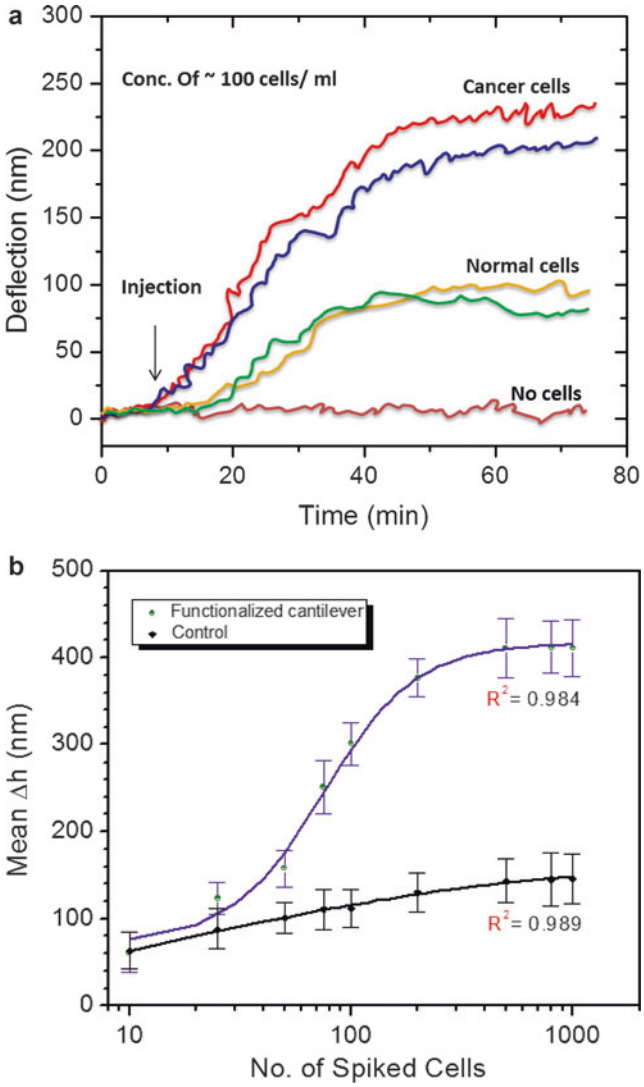
Functionalized microcantilevers have also been used to detect several toxic organic compounds such as organophosphorus compounds, dichlorodiphenyltrichloroethane (DDT), endocrine-disrupting chemicals (EDCs), hydrofluoric acid (HF), and others [32]. Microcantilevers have also been demonstrated to detect toxic heavy metals generated in water from polluted soil and ground water. Ion-specific, self-assembled monolayers (SAMs) immobilized on the cantilever surface provide selectivity. Microcantilevers are also used as an electrode in an electrochemical cell for use in the detection of metal ions. Many exciting applications of microcantilever chemical sensors with high sensitivity and selectivity are currently being demonstrated. However, simultaneous detection of multiple analytes is still in its infancy. Cantilever arrays functionalized with unique selective layers provide a real possibility of achieving this goal.

The microcantilever deflection-based sensing technique has drawn substantial attention as a biosensor platform due to its simplicity, sensitivity, and label-free detection, as well as its ability for real-time, in situ monitoring. Unlike the resonance frequency-based approach, cantilever bending is not significantly affected by the liquid medium. As biological sensing devices, microcantilever applications have included studying antigen–antibody interactions [33]; DNA hybridizations [20]; detection of superbugs such as pathogenic bacteria, fungus, and viruses; and, most recently, the monitoring and detection of cancer biomarkers and gene transmutations [2, 12]. Observations that multiple DNA hybridizations and antigen–antibody interactions can collectively produce a nanomechanical motion in a cantilever make the platform suitable for developing sensors based on DNA [34]. Several studies have shown the applicability of using cantilever sensors to detect DNA hybridizations and mismatching [21, 35, 36]. The detection is simply based on the interaction of immobilized DNA probes (oligonucleotides) with complementary DNA strands in solutions. The process also requires reference cantilevers functionalized with negative oligonucleotides that do not match the target DNA strands. It has been demonstrated that the cantilever biosensors offer interesting possibilities for label-free detection of biomolecules such as proteins and enzymes and carbohydrates based on nanomechanical motion [37]. Recent advances in cantilever technology developments have shown the ability of microcantilevers to detect a number of pathogenic microorganisms and cancer biomarkers. A large number of published works show the capability of cantilever technologies to detect bacterial cells at very low concentrations in different samples. For example, a report by *Mutharasan et al*. showed the capability of a piezoelectric-excited millimeter cantilever (PEMC) functionalized with a virulent gene stx2 to detect *E. coli O157:H7 (EC)* in both buffer and proteinous sample at very low concentrations of the bacteria ($\sim$700 cell mL^{-1}) and without any sample pretreatment steps [38]. Recently *Huber F et al.* showed a gold standard assay for detecting the BRAF gene mutation using an array of microcantilever sensors [2]. The BRAF gene mutation is a characteristic indication of malignant melanoma, the deadliest form of skin cancer. This study indicated the applicability of cantilever assays to distinguish between mutated cell lines and wild-type cell lines and gives hope for the system to be used for real-time cancer detection. Furthermore, in our unpublished results, a microcantilever sensor array functionalized with cancer-specific peptides was found to bend specifically to cancer cells over a wide range of concentrations as a result of a peptide–receptor interaction. As shown in Fig. 2, the bending response due to cancer cell adsorption was higher than that due to normal cells.

An ideal sensing device should meet a number of specific criteria such as exceptional sensitivity, specificity, and time-independent functionality for real-time monitoring. Microcantilever sensors offer the possibility of highly sensitive, real-time detection, and cost-effective analysis assays that can be used for multiplexed detection with no need for extrinsic labels. Thus far, the microcantilever array platform has shown significant progress toward becoming an advanced sensing tool for chemical and biological detection. The cantilever platform has proven to be very effective, ultrasensitive, and a time-saving tool which can potentially be used as a diagnostic device in the near future.

Challenges

Despite enormous advances over the past decade in the development of cantilever sensors, the transition from laboratory sensing to the outside world has been slow. There are many potential reasons behind this slow transition. Chemical sensing of small molecules has been limited by the challenge of selectivity. Even in array-based cantilever sensing, achieving selectivity in detection has been a challenge. Efforts are underway to develop selective interfaces in order to enhance selectivity. For microcantilever-based biosensors, the main challenge has been reproducibility. Reproducibility is affected by several factors



Microcantilever Chemical and Biological Sensors, Fig. 2 (**a**) Real-time detection of breast cancer cell lines using microcantilever functionalized with a cancer-specific peptide. Significant deflection of the cantilever as a result of interaction of the immobilized peptide with a specific receptor over expressed on the outer membrane of cancer cells. (**b**) Microcantilever bends as a function of time for different concentrations of breast cancer cell line. Microcantilever bending increases as the concentration of the cells increases, indicating a linear scaled relationship. The control experiment, where normal cells are spiked, indicates specificity of peptide binding to the cancer cell membrane receptor

which need to be resolved. Finding the ideal recognition molecule or binding receptor which can selectively and specifically capture the target analyte is still missing. Despite the fact that coating cantilevers with semiselective receptors such as antibodies, nucleic acid probes, or targeting peptides have significantly enhanced cantilever detection sensitivity and selectivity, the challenges remain in the sense that the ideal selective ligand is not available yet. Though selectivity can be achieved to some extent by coating the cantilevers with semiselective receptors, the issue of immobilizing the receptors without hampering its specificity and sensitivity is still under development. Reliable receptor immobilization techniques on an individual cantilever in an array in a reproducible manner still need to be developed. Furthermore, cantilevers can be affected by the variation of the environmental parameters such as temperature, humidity, and flow rate. In general, although a common-mode rejection approach using reference cantilevers may provide sufficient protection from changes in the environmental factors, more work is required. Systematic investigations are also needed to understand the side effects of the fluid delivery systems to enhance reproducibility.

Conclusion

Although a number of obstacles remain to be resolved, the microcantilever sensor platform offers a viable approach with high-sensitivity, high-throughput, and multiplexed detection capabilities. Development of sensors capable of determining several analytes simultaneously with proper sensitivity, accuracy, and real-time detection would lead to an ideal sensing tool in product analysis, environmental monitoring, clinical diagnostics, and biomedical point of care. So far, the instantaneous determination of multiple analytes in real time has been a great challenge in clinical diagnosis. However, recent developments in microcantilever technologies are paving the way toward real-time monitoring in complex samples with high sensitivity and selectivity using miniature handheld devices.

References

1. Farahi, R.H., Passian, A., Tetard, L., Thundat, T.: Critical issues in sensor science to aid food and water safety. ACS Nano **6**, 4548–4556 (2012)
2. Huber, F., Lang, H.P., Backmann, N., Rimoldi, D., Gerber, C.: Direct detection of a BRAF mutation in total RNA from melanoma cells using cantilever arrays. Nat. Nanotechnol. **8**, 125–129 (2013)
3. Scallan, E., et al.: Foodborne illness acquired in the United States – major pathogens. Emerg. Infect. Dis. **17**, 7–15 (2011)
4. Fakruddin, M., Mannan, K.S., Andrews, S.: Viable but nonculturable bacteria: food safety and public health perspective. ISRN Microbiol. **2013**, 703813 (2013)
5. Utzinger, J., et al.: Microscopic diagnosis of sodium acetate-acetic acid-formalin-fixed stool samples for helminths and intestinal protozoa: a comparison among European reference laboratories. Clin. Microbiol. Infect. **16**, 267–273 (2010)
6. Sorgenfrei, S., et al.: Label-free single-molecule detection of DNA-hybridization kinetics with a carbon nanotube field-effect transistor. Nat. Nanotechnol. **6**, 126–132 (2011)
7. Puard, V., et al.: Semi-quantitative measurement of specific proteins in human cumulus cells using reverse phase protein array. Reprod. Biol. Endocrinol. **11**, 100 (2013)
8. Ampil, F.L.,Caldito, G. Patient-provider delays in superior vena caval obstruction of lung cancer and outcomes. Am. J. Hosp. Palliat. Care (2013).
9. van der Waal, I.: Are we able to reduce the mortality and morbidity of oral cancer; some considerations. Med. Oral Patol. Oral Cir. Bucal **18**, e33–e37 (2013)
10. Justino, C.I., Rocha-Santos, T.A., Duarte, A.C.: Review of analytical figures of merit of sensors and biosensors in clinical applications. TrAC Trends Anal. Chem. **29**, 1172–1183 (2010)
11. Wu, G., et al.: Bioassay of prostate-specific antigen (PSA) using microcantilevers. Nat. Biotechnol. **19**, 856–860 (2001)
12. Zemp, R.J.: Nanomedicine: detecting rare cancer cells. Nat. Nanotechnol. **4**, 798–799 (2009)
13. Rajmohan Joshi, R.M.J.: Biosensors. Isha Books, New Delhi (2006)
14. Sadana, A.: Biosensors: Kinetics of Binding and Dissociation Using Fractals. Elsevier, Amsterdam/Boston (2003)
15. Binnig, G., Quate, C.F., Gerber, C.: Atomic force microscope. Phys. Rev. Lett. **56**, 930–933 (1986)
16. Thundat, T., Zheng, X.-Y., Chen, G., Warmack, R.: Role of relative humidity in atomic force microscopy imaging. Surf. Sci. **294**, L939–L943 (1993)
17. Thundat, T., et al.: Atomic force microscopy of DNA on mica and chemically modified mica. Scanning Microsc. **6**, 911 (1992)
18. Xu, X., Thundat, T.G., Brown, G.M., Ji, H.-F.: Detection of Hg^{2+} using microcantilever sensors. Anal. Chem. **74**, 3611–3615 (2002)
19. Mertens, J., et al.: Label-free detection of DNA hybridization based on hydration-induced tension in nucleic acid films. Nat. Nanotechnol. **3**, 301–307 (2008)
20. Wu, G., et al.: Origin of nanomechanical cantilever motion generated from biomolecular interactions. Proc. Natl. Acad. Sci. U. S. A. **A98**, 1560–1564 (2001)
21. McKendry, R., et al.: Multiple label-free biodetection and quantitative DNA-binding assays on a nanomechanical cantilever array. Proc. Natl. Acad. Sci. U. S. A. **A99**, 9783–9788 (2002)
22. Braun, T., et al.: Quantitative time-resolved measurement of membrane protein-ligand interactions using microcantilever array sensors. Nat. Nanotechnol. **4**, 179–185 (2009)
23. Rijal, K., Mutharasan, R.: A method for DNA-based detection of *E. coli* O157:H7 in a proteinous background using piezoelectric-excited cantilever sensors. Analyst **138**, 2943–2950 (2013)
24. Maraldo, D., Mutharasan, R.: 10-minute assay for detecting *Escherichia coli* O157:H7 in ground beef samples using piezoelectric-excited millimeter-size cantilever sensors. J. Food Prot. **70**, 1670–1677 (2007)
25. Waggoner, P.S., Craighead, H.G.: Micro- and nanomechanical sensors for environmental, chemical, and biological detection. Lab Chip **7**, 1238–1255 (2007)

26. Cherian, S., Gupta, R.K., Mullin, B.C., Thundat, T.: Detection of heavy metal ions using protein-functionalized microcantilever sensors. Biosens. Bioelectron. **19**, 411–416 (2003)
27. Nordström, M., et al.: SU-8 Cantilevers for bio/chemical sensing; fabrication, characterisation and development of novel read-out methods. Sensors **8**, 1595–1612 (2008)
28. Ding, C., Zhu, A.,Tian, Y. Functional surface engineering of C-dots for fluorescent biosensing and in vivo bioimaging. Acc. Chem. Res. (2013)
29. Martínez, M.T., et al.: Label-Free DNA Biosensors Based on Functionalized Carbon Nanotube Field Effect Transistors. Nano Lett. **9**, 530–536 (2009)
30. Letant, S.E., Hart, B.R., Van Buuren, A.W., Terminello, L.J.: Functionalized silicon membranes for selective bio-organism capture. Nat. Mater. **2**, 391–395 (2003)
31. Zimmermann, J.L., Nicolaus, T., Neuert, G., Blank, K.: Thiol-based, site-specific and covalent immobilization of biomolecules for single-molecule experiments. Nat. Protoc. **5**, 975–985 (2010)
32. Buchapudi, K.R., Huang, X., Yang, X., Ji, H.F., Thundat, T.: Microcantilever biosensors for chemicals and bioorganisms. Analyst **136**, 1539–1556 (2011)
33. Zhang, J., et al.: Development of robust and standardized cantilever sensors based on biotin/NeutrAvidin coupling for antibody detection. Sensors (Basel) **13**, 5273–5285 (2013)
34. Hansen, K.M., et al.: Cantilever-based optical deflection assay for discrimination of DNA single-nucleotide mismatches. Anal. Chem. **73**, 1567–1571 (2001)
35. Oliver, P.M., Park, J.S., Vezenov, D.: Quantitative high-resolution sensing of DNA hybridization using magnetic tweezers with evanescent illumination. Nanoscale **3**, 581–591 (2011)
36. Fritz, J., et al.: Translating biomolecular recognition into nanomechanics. Science **288**, 316–318 (2000)
37. Savran, C.A., Knudsen, S.M., Ellington, A.D., Manalis, S.R.: Micromechanical detection of proteins using aptamer-based receptor molecules. Anal. Chem. **76**, 3194–3198 (2004)
38. Sharma, H., Mutharasan, R.: hlyA gene-based sensitive detection of *Listeria monocytogenes* using a novel cantilever sensor. Anal. Chem. **85**, 3222–3228 (2013)

Microcavity Biosensor

► Whispering Gallery Mode Resonator Biosensors

Microchannel Flows as Heat Sinks

Yogesh Jaluria
Department of Mechanical and Aerospace Engineering, Rutgers University, Piscataway, NJ, USA

Synonyms

Cooling of electronic components; Heat removal; Heat transfer enhancement; Microfluidics

Definitions

Microchannel, heat sink, thermal management, slip flow, pressure head, viscous dissipation, heat transfer rate, heat transfer coefficient, Pareto frontier.

Introduction

The cooling of electronic components is one of the main barriers to developing faster, smaller, and more reliable systems. Heat removal, or *thermal management*, is critical to achieving higher data processing speeds, particularly for large-scale computers such as data centers. Because of the growing trend of assembling increasing numbers of electronic chips on a single module and multiple modules on a single board, higher power densities are encountered at both the chip and module levels. Similarly, the number of devices at chip level has increased from thousands to several millions, while the average chip size has reduced from a few millimeters to a few microns, making the thermal management problem even more challenging [1–3]. Emphasis has been placed on improving the reliability and performance of electronic devices by means of appropriate cooling techniques. As a consequence, several cooling arrangements are currently being used in the industry, of particular interest being air and

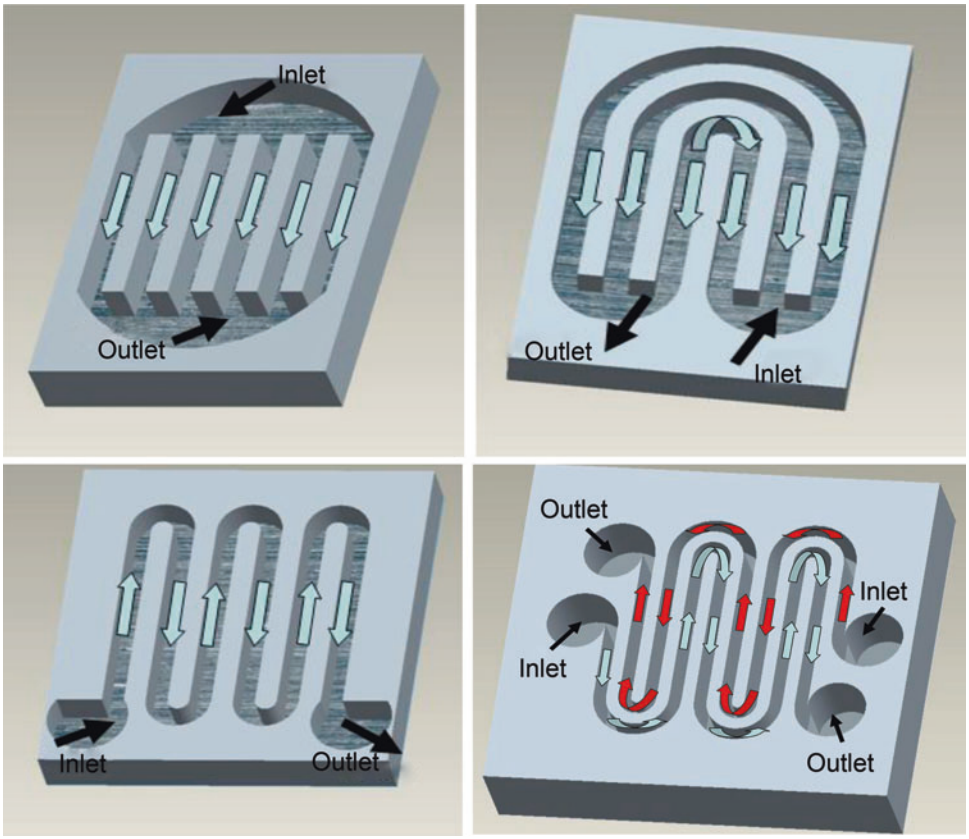
liquid cooling. The heat transfer modes associated with air cooling are largely natural, forced, and mixed convection [4]. Liquid cooling can be direct or indirect, the former being the result of physical contact between the coolant and the heat sources [5]. Convection and boiling are the most commonly used heat transfer modes in both types of liquid cooling. Though air cooling is still widely used due to availability, low cost, easy maintenance, and simple design, high heat flux inputs require liquid cooling, which has thus been increasing in popularity, particularly for high-performance and rapid response electronic systems. The system design is largely dominated by the temperature constraints, which are generally quite severe. For this purpose, it is desirable to remove the energy dissipated by the electronic components efficiently and at minimum cost.

Over the last few decades, substantial work has been done on the thermal management of electronics systems, considering new systems and devices, with increasing diversity, complexity, performance, and reliability. The challenges posed by continued reduction in physical size, while the number of processors over a given area has continued to increase, have been met by a wide variety of solutions based on air and liquid cooling. This has resulted in a phenomenal increase in our capacity to achieve a wide range of operations in computation, monitoring, and control, with fairly compact systems [6, 7]. However, as the power densities have increased, reaching into hundreds of W/cm^2 and placing stringent demands on the *heat removal* systems, the use of *microchannel* flows for heat removal has become important to remove excessive heat and control the temperature for every component to ensure reliable performance and prevent any failure or malfunction. Microchannels are generally in the range of 10–200 μm, with respect to the diameter or width, with larger channels, in the range 1–5 mm, often being termed as minichannels.

A *heat sink* is typically a finned structure in contact with the heat source and is applied as an effective arrangement to remove the heat dissipated by electronic chips and other devices. It is cooled largely by forced convection with forced air cooling the most commonly used approach for heat removal. The power dissipated by the electronic chips is conducted through the substrate of the heat sink to the fins and to the coolant. However, these are typically at lower heat flux inputs, up to a maximum of around 0.2 W/cm^2 for air and around 1 W/cm^2 for liquids.

For higher heat fluxes, ranging from 1 to 10^2 W/cm^2 and beyond, air or liquid flow without phase change in microchannel heat sinks provides a promising thermal management arrangement. Heat fluxes as high as 10^3 W/cm^2 can also be addressed by single-phase liquid flow, though the pressures generally become unacceptably high due to the small microchannel diameters needed [8, 9]. Two-phase flow with boiling can then be employed for still higher heat fluxes, without the large pressures encountered in single-phase flow, though the complexity of the system further increases due to the different flow regimes, instability, and complex nature of the basic phenomena [10]. Compared with two-phase flow, single-phase cooling systems have less complicated physical phenomena, and the system is easier to control.

Electronic cooling problems are generally considered for steady-state conditions. Assuming the power to the electronic chip is constant, and after the electronic system is turned on and kept running for a long period of time, the temperatures of the electronic chips and cooling devices are expected to reach steady state. When the thermal equilibrium condition is met, the rates of heat being transferred by conduction, convection, and radiation all remain constant. For a large electronic system, it may take a very long operation time before the system becomes steady. It may only need several minutes for a cooling system of a small electronic chip to become thermally developed. However, the maximum heat load usually appears at the start or the shutdown. In order to control the temperature and prevent the electronics from overheating, it is important to study the transient heat transfer behavior for electronic cooling as well [11]. Thus the thermal management problems involve both steady and time-dependent situations.



Microchannel Flows as Heat Sinks, Fig. 1 Different configurations of microchannels in heat sinks

The thermal performance of the heat sink can be changed by varying the configuration of the microchannels. Several configurations are sketched in Fig. 1, indicating straight, U-shaped, and serpentine microchannels with different inlets and outlets [12]. Similarly, the flow may be split into different channels, such as a tree-shaped structure or fork-shaped flow structure, in order to improve the heat removal effectiveness. This breaks up the generally developed flow in the microchannel and increases the heat transfer coefficient. Multiple microchannel heat sinks with a bifurcation along the flow direction may also be considered, so that the bifurcation can augment the heat transfer coefficient due to flow separation, secondary flow, and restarting of the boundary layer, if the system is designed properly. The aim of the design in all these cases is to enhance the heat transfer without unacceptably high pressures. Modeling and experiments have been used to provide the basic information and characteristics that may be used to design and optimize the heat removal system so that the energy dissipated is removed without exceeding the temperature constraints at any point within the electronic components. Such points, where excessive temperature may arise, are known as hot spots, and the cooling system must ensure that unacceptable localized temperature levels do not arise.

Basic Considerations

As mentioned in the preceding, many important basic and practical considerations arise and must be addressed for a satisfactory deployment of microchannel heat sinks. Some of the most important ones are

1. Heat transfer rates achieved by the microchannel flow
2. Pressure head needed for the flow to achieve desired heat transfer characteristics
3. Predictability, reliability, control, and stability of the flow
4. Required characteristics of the coolant
5. Fabrication and cost of the microchannel heat sink
6. Space and weight constraints

The focus is often placed on the heat transfer rates, or heat transfer coefficients, attained by the use of microchannels. The heat transfer enhancement, over traditional heat sinks, can be substantial, making this an attractive heat removal approach. However, as the flow rate increases to obtain higher heat transfer rates, the pressure head needed to maintain the flow also increases and can, in many cases, make the use of microchannel heat sinks difficult, if not impossible. Also, if the channel dimensions are reduced for heat transfer enhancement, the pressures increase as D^{-4}, where D is the diameter, limiting this approach for increased heat transfer rates. Similarly, fabrication costs, flow instabilities, particularly in phase change systems, and coolant leakage at high pressures can limit the advantages offered by microchannel heat sinks in terms of higher heat transfer rates.

There are several methods for fabricating microchannels. Among these are wet etching and plasma etching. In wet etching, a potassium hydroxide solution can be used as etchant for the silicon-based microchannels. Since KOH is an orientation-dependent etchant, which means it attacks {111} much more slowly than other planes, P type {110} oriented single-crystal silicon can be masked with nitride and oxide together. A magnetic stirrer is used to agitate the solution to eliminate the difference in rate of etching from the top to the bottom of the solution. Isopropanol is added to the KOH solution to improve the uniformity, though it also reduces the etch rate.

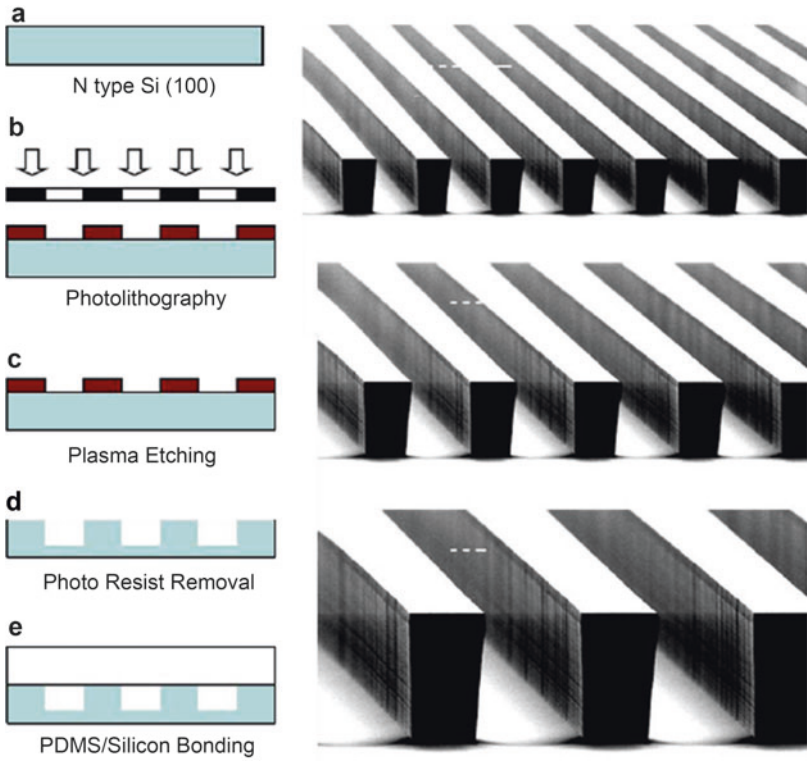
Plasma, or dry, etching yields better microchannel uniformity compared to that obtained with wet etching. The wafer preparation for this technique is easier since it involves fewer steps compared with wet etching. The basic approach and the resulting microchannels are shown in Fig. 2. Typically, the heat sinks are fabricated on a silicon wafer and are integral with the heat-dissipating electronic chip. Bare silicon is cleaned to remove organic contamination, followed by a HF dip to remove native oxide on the silicon in order to reduce undercutting of the photoresist mask. After conventional ultraviolet (UV) photolithography defines the microchannel pattern on the photoresist, the silicon wafer is etched to create microchannels by a state-of-the-art inductively coupled plasma etcher. It is optimized for deep Si etching using Bosch process. This is a high–aspect ratio etching technology, which employs a protection film deposition after the isotropic etching. A Polydimethylsiloxane (PDMS) layer may be bonded on top of the silicon microchannels to create fluid passages, as shown in Fig. 1. Openings can then be made for fluid connections [12].

Heat transfer and fluid flow in a single microchannel tube or a channel for use in a heat sink have been studied by several researchers experimentally and numerically, particularly for liquids without phase change as the coolant for chip cooling. As mentioned earlier, the heat transfer rate, or heat transfer coefficient, with the microchannel flow and the pressures needed for the flow are of particular interest. In addition, the thermal response is of interest in time-dependent problems. The overall size and weight are also of practical interest.

Modeling and Simulation

The modeling of microchannel sinks involves several important aspects because of the relative small size of the channels, generally being between a few microns to about 200 μm in diameter or width, and typically macro, or engineering, scale of the electronic components and systems. Thus, the modeling must consider

1. Flow and heat transfer at the microscale level
2. Conjugate heat transfer in the solid material surrounding the microchannels

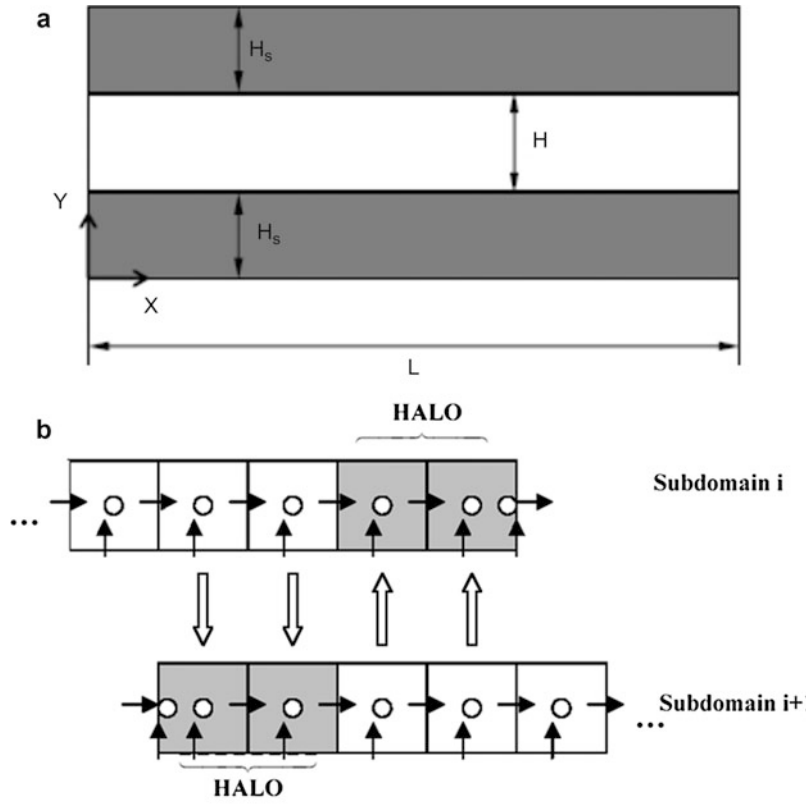


Microchannel Flows as Heat Sinks,
Fig. 2 Fabrication and packaging process for microchannel heat sinks, using plasma etching

3. Multiscale modeling, considering widely different length scales of interest
4. Time-dependent and steady-state situations

Many mathematical and numerical models have been developed to study microchannel flow and heat transfer systems. The majority of these use the Navier–Stokes equations to analyze the microscale fluid flow, as also in other microchannel flows such as those in *microfluidics*. However, unlike the macroscale channels, the ratio of fin width to channel width, as well as the ratio of the silicon substrate thickness to the channel depth, are not negligible as in microscale systems. Because of changes in temperature and pressure, the influence of the variable thermal properties on heat transfer and flow in microchannel heat sinks can not generally be neglected. Combined conduction and convection, or conjugate heat transfer, needs to be considered for numerical analysis in this case.

Let us first consider the modeling at microscale length dimensions. As an example, consider the numerical simulation of pressure-driven nitrogen slip flow in long microchannels, with conjugate heat transfer at the walls of the heat removal system under a uniform heat flux wall boundary condition, as sketched in Fig. 3. For the gas phase, the two-dimensional momentum and energy equations are solved, considering variable properties, rarefaction, which could give rise to velocity slip at the walls, thermal creep and temperature jump, compressibility, and viscous dissipation. Slip conditions generally do not arise for liquid flow. However, the basic concerns with pressure, viscous dissipation, variable properties, and fine grids to capture the flow and heat transfer characteristics have to be considered at these length scales. For conduction in the solid region, which forms the bounding materials around the channel and which are much larger in dimensions, the energy equation is solved with variable properties [13]. Thus, the two regions need to be treated with different approaches, resulting in multiscale modeling and simulation. In many practical systems, transport processes in



Microchannel Flows as Heat Sinks, Fig. 3 Nitrogen flow in a microchannel, with conjugate transport at the walls. (**a**) Sketch of the microchannel; (**b**) domain decomposition for parallel computation of the microchannel flow, as well as of heat transfer

various components or subsystems are at different length or time scales. Then the governing equations for different components are different. Figure 3b shows a domain decomposition approach that may be used for parallelization of the computation process, as well as for different components or subsystems. This allows an efficient approach to the modeling of microchannel heat sinks.

Due to rarefaction, discontinuous boundary conditions are applied at the wall, particularly for gases, depending on the Knudsen number, $Kn = \lambda/H$, where λ is the mean free path of the gas molecules and H is the channel height for two-dimensional rectangular channels or diameter for circular channels, as shown in Fig. 3. Continuum flow is assumed if the Knudsen number $Kn \leq 10^{-3}$; slip flow occurs when $10^{-3} \leq Kn \leq 10^{-1}$; transition flow arises $10^{-1} \leq Kn \leq 10$; and free molecular flow arises for $Kn > 10$. These aspects have been considered in detail in the literature [9]. For isothermal walls, the wall boundary conditions are at $y = 0$ and $y = H$, $u = \frac{2-\sigma_v}{\sigma_v}\lambda\left(\frac{\partial u}{\partial n}\right)_w + \frac{3}{4}\frac{\mu}{\rho T}\left(\frac{\partial T}{\partial x}\right)_w$, $v = 0$,

$$T_g - T_w = \frac{2-\sigma_T}{\sigma_T}\left[\frac{2\gamma}{\gamma+1}\right]\frac{\lambda}{\mathrm{Pr}}\left(\frac{\partial T}{\partial n}\right)_w$$

The coefficients σ_v and σ_T depend on properties of the gas and on the surface characteristics. They represent the fraction of the molecule's tangential momentum and energy loss through the interactions with the solid wall, respectively. The substrates, or solid regions, are governed by the energy equation,

$$c_p\left[\frac{\partial \rho T}{\partial t} + \frac{\partial \rho u T}{\partial x} + \frac{\partial \rho v T}{\partial y}\right] = \left[\frac{\partial}{\partial x}\left(k\frac{\partial T}{\partial x}\right) + \frac{\partial}{\partial y}\left(k\frac{\partial T}{\partial y}\right)\right]$$

The boundary conditions in the y-direction, i.e., at $y = 0$ and $y = 2H_s + H$, are given by $q = q_w$,

q being the heat flux and subscript w referring to the wall. The temperature at the fluid–solid interface is obtained by solving the two equations

$$T_g - T_w = \frac{2-\sigma_T}{\sigma_T}\left[\frac{2\gamma}{\gamma+1}\right]\frac{\lambda}{\mathrm{Pr}}\left(\frac{\partial T}{\partial n}\right)_w$$

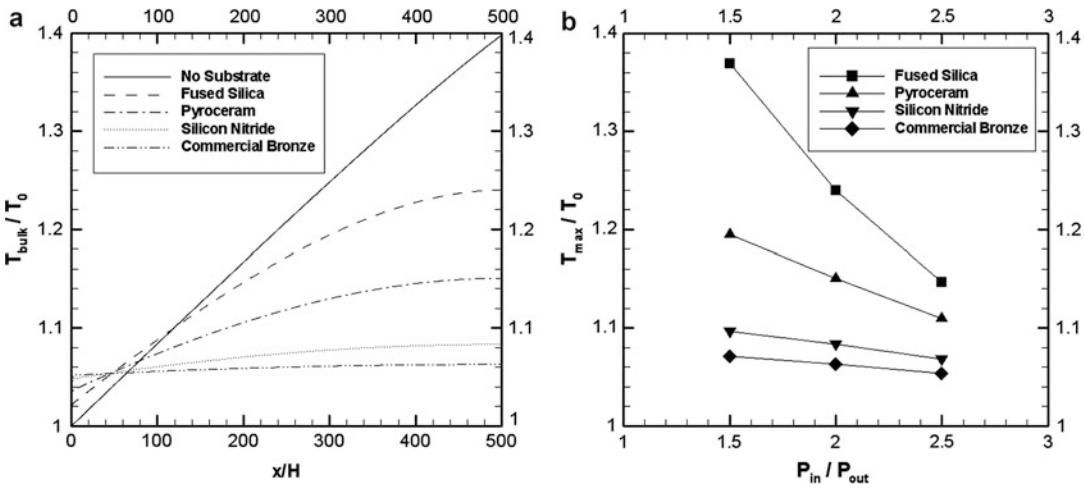
$$q_g = q_s$$

where the first equation is due to the temperature jump at the gas-solid interface and the second one arises from the fact that energy cannot be stored in the interface. Here, u, v are the velocity components along x and y coordinates, T is the temperature, subscripts g and w refer to the gas and wall, respectively, Pr is the Prandtl number, c_p is the specific heat at constant pressure, k is thermal conductivity, γ is the ratio of the two specific heats, and t is time.

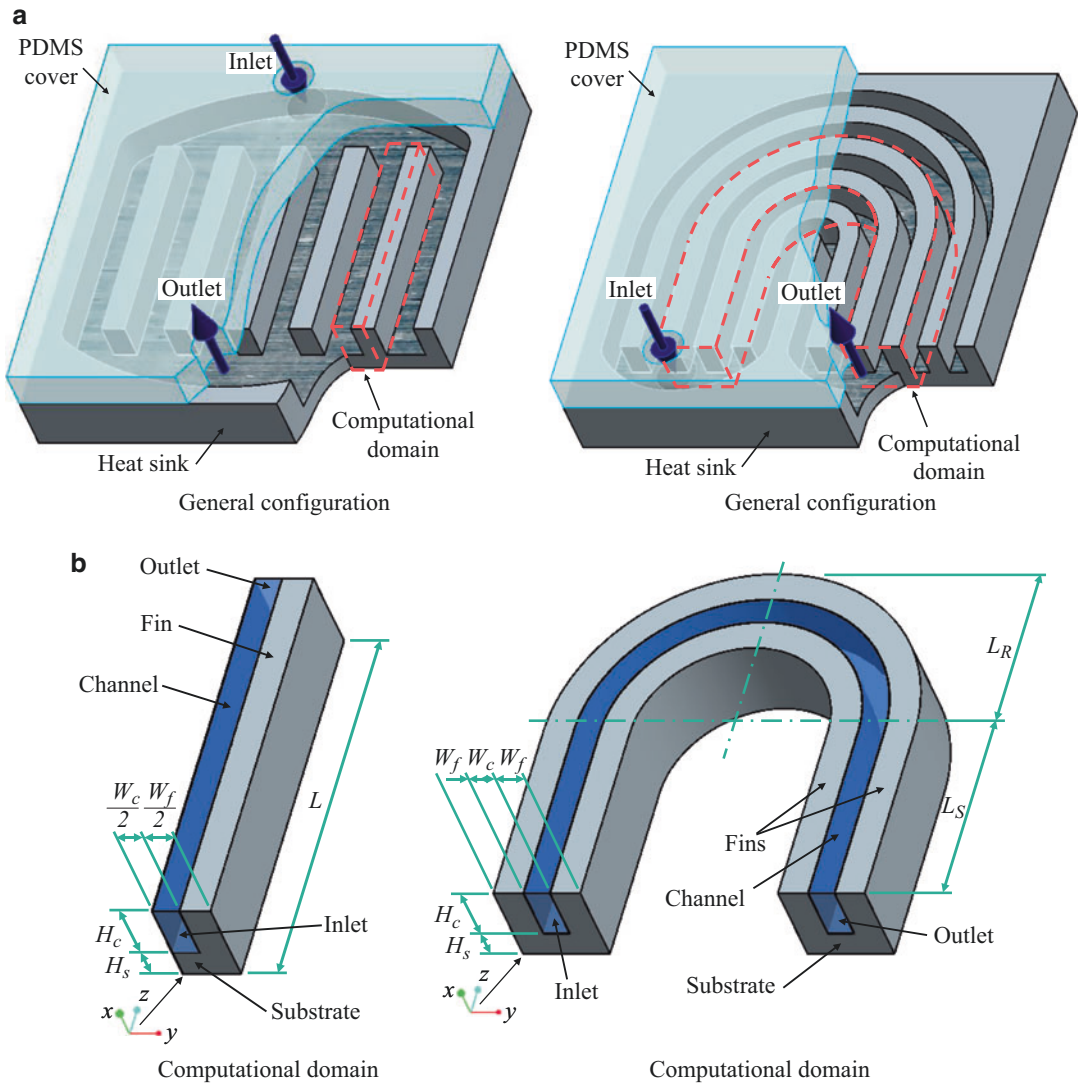
Typical results for different substrate materials including commercial bronze, silicon nitride, pyroceram, and fused silica are shown in Fig. 4. The effects of axial conduction in the substrate, material thermal conductivity, and substrate thickness are clearly seen. The substrate axial conduction leads to a flatter bulk temperature profile along the channel, lower maximum temperature, and lower Nusselt number. The effect of substrate thickness on the conjugate heat transfer is very similar to that of the substrate thermal conductivity. Thus, in terms of axial thermal resistance, the increase in substrate thickness has the same effect as that caused by an increase in its thermal conductivity [14].

The same concepts may be used for typical heat sinks using microchannel flows. Figure 5 shows close-up sketches of straight and U-shaped microchannels. The substrate is shown along with the microchannels. Thus, two different configurations are considered. Microchannel heat sinks with bent, or curved, regions can be compared with straight channels with the same surface area. Numerical models need to be validated by experimental data. For each configuration, the influence of geometry and flow rate on the heat transfer and fluid flow characteristics of the system are determined. The validated numerical models can then be used for the system design and optimization. The objective is to minimize the thermal resistance and maintain the chip temperature below a given limit of about 90 °C with an acceptable pumping power.

It is computationally intensive to model the whole microchannel heat sink with multiple microchannels. Therefore, simplified domains are used in the model, as shown. For the straight



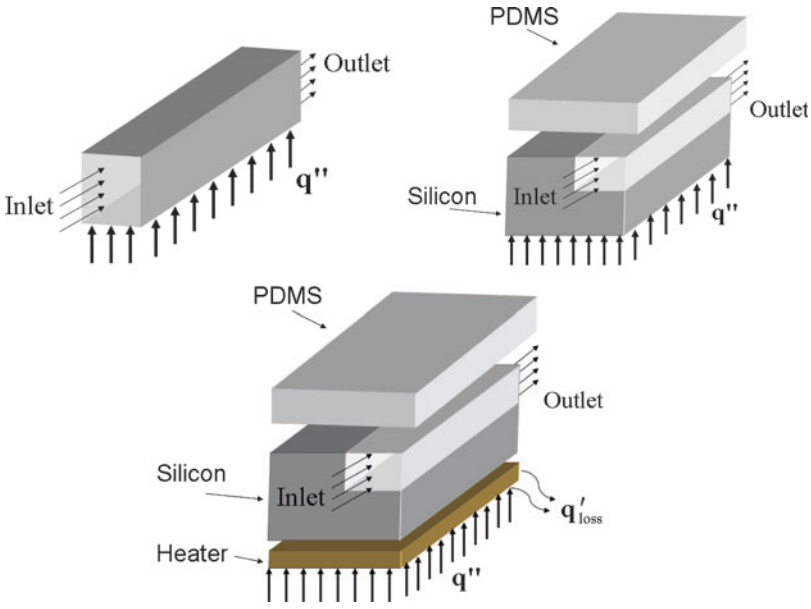
Microchannel Flows as Heat Sinks, Fig. 4 Conjugate transport in microchannels. (**a**) Bulk temperature distribution for different wall materials; (**b**) maximum temperature as a function of the imposed pressure difference for different wall materials



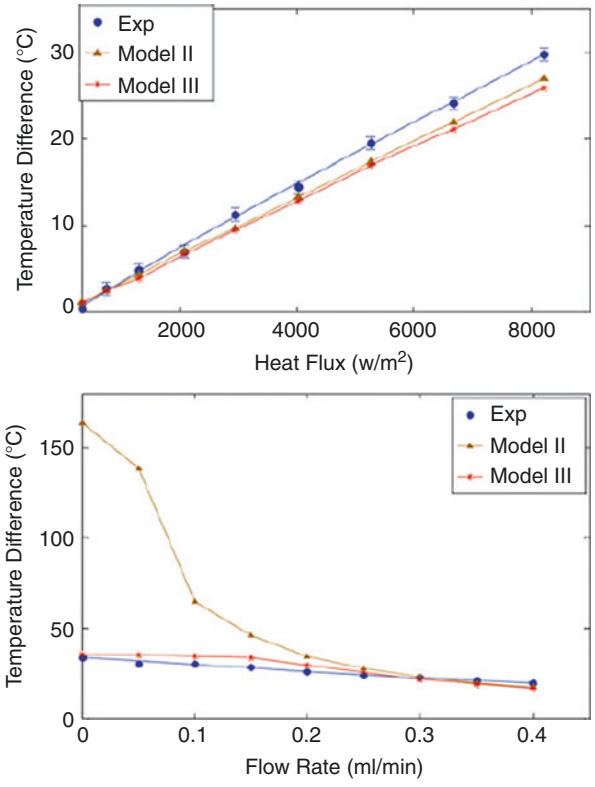
Microchannel Flows as Heat Sinks, Fig. 5 Sketch of straight and U-shaped microchannel heat sinks, along with their computational domains

microchannel heat sink, because the entire heat sink is geometrical symmetric, only half of the fin and the microchannel need to be considered. It is noted that for the U-shaped channels, the length of each channel is different. The channels located at the center of the heat sink are chosen for study as the computational domain because there is relatively little heat transfer to the sides at the center. It was shown that the channel at the center and one on either side are adequate to model the entire sink consisting of 40 microchannels or more.

Figure 6 shows straight channels embedded in a solid substrate and in contact with a heat source. Three different models are shown, from the simplest focusing on the microchannel to a complete model that considers the heater and the substrate along with the microchannel. The validity of any given model is based on comparisons with experimental data. Figure 7 shows such comparisons, indicating that an accurate modeling demands that the entire region consisting of the heater, microchannel, and substrate be considered, rather



Microchannel Flows as Heat Sinks, Fig. 6 Different mathematical and computational models for a straight microchannel heat sink



Microchannel Flows as Heat Sinks, Fig. 7 Comparison between numerical results and experimental data for the systems shown in Fig. 6

than using approximations for the boundary conditions. This simple example illustrates the simulation of different components with different length scales, followed by coupling of the results to obtain the model for the entire system. These comparisons are shown for steady operation, and similar conclusions are obtained from time-dependent investigations. This is particularly

valuable in electronic thermal management systems.

Many studies have focused on heat transfer enhancement in microchannel flows. The effects of increased flow rate and reduced diameter, or width, on enhancing the heat transfer coefficient have been mentioned earlier. The effect of the flow configuration on both transient and steady operation was discussed. Other techniques considered include enhancement features, such as pin fins, surface roughness, flow disruption, and cross mixing, in the microchannels [15]. Many such techniques disrupt the developed flow in the channel and thus lead to higher heat transfer coefficients. Appropriate fabrication techniques have also been developed for such enhancement features.

Thermal Management and Optimization

A wide range of microchannel heat sinks have been considered in the literature [16, 17] from both fundamental and applied viewpoints. Even though considerable information exists on the flow and heat transfer in microchannels for single- as well as multiphase flow, arising from boiling [18, 19], the implementation of these results into practical systems for heat removal, or thermal management, involves several other important considerations, as mentioned earlier. Certainly, among the most important is the pressure needed for the flow and the consequent concerns with pressure supply, leakage, dimensions, and cost. These must be addressed for a feasible or optimal design. In addition, cost for fabrication and maintenance of the sink, as well as the weight and volume, are important considerations. The coolant, particularly its cost, availability, dielectric and thermal properties, and interaction with materials being used is also an important consideration. Air is generally the most convenient and inexpensive coolant, though obviously the heat transfer rates are much lower than those for liquids. Liquids such as water, refrigerants R123 and R134a, fluorocarbons, and Ethylene or Propylene Glycol/water mixtures are commonly used in cooling systems for electronics. Phase change results in very high heat transfer coefficients, but the stability, control, and maintenance of the system poses its own challenges.

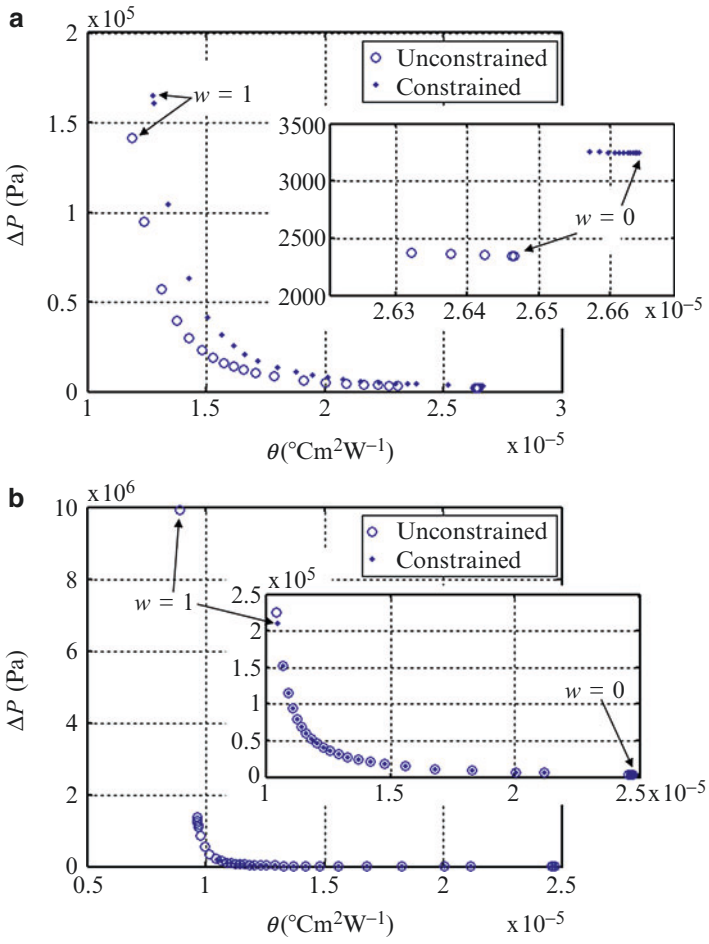
Increase in heat transfer rates for single-phase coolants can thus be achieved in many ways, as mentioned earlier. Some of these are

Reducing diameter of the microchannels
Higher flow rates of the coolant
Varying the flow configuration of the channels
Building heat transfer enhancement structures in the microchannels
Adding nanoparticles in the fluid, i.e., use of nanofluids
Increasing the roughness of the surface in the channel

There are several other approaches for heat transfer enhancement that have been considered in the literature. However, in most cases the price paid for the increase in heat removal is in terms of the increased pressure head needed for the flow.

Therefore, it is critical to optimize the microchannel heat sink with respect to these two main objectives – heat transfer rate, or heat transfer coefficient, and the pressure head. Using both numerical simulation and experiments, as outlined earlier, a large database can be built up to characterize the behavior of microchannel heat sinks. From these results, an optimization of the system, with respect to the heat transfer rate and the pressure, can be undertaken. Even though this is clearly a critical consideration in the design and use of heat sinks, it has not received as much attention as is necessary. Much of the focus has been on raising the heat transfer rates and thus removing the heat generated by the electronic components at very high heat flux levels [20].

The data obtained from experiments and numerical simulation can be used to generate response curves and simplified models that can form the basis for optimization. Since there are two parameters to be minimized or maximized, namely, the heat transfer coefficient h and the pressure Δp, it is a multiobjective optimization problem. For isothermal surfaces, we can consider the heat transfer coefficient h, and for uniform heat flux surfaces, we may consider the temperature θ. Thus, the response surfaces focus on h, or θ, and Δp, with



Microchannel Flows as Heat Sinks, Fig. 8 Pareto frontiers for multiobjective optimization of straight microchannel heat sinks for (**a**) a given number of microchannels, $N = 56$, and (**b**) varying number of microchannels

the above variables, such as channel diameter and flow rate, as the independent variables. It is possible to combine θ and Δp into a composite objective function, with different weights, and thus optimize a single objective function. But this is arbitrary and does not represent the actual problem. However, by following well-established multiobjective techniques, one could generate *Pareto frontiers* that represent the dominant solutions for the two objective functions.

Figure 8 presents some of such multiobjective optimization results in terms of Pareto frontiers. Different weights may be assigned to the two objective functions, depending on the particular system under consideration. Each point on the frontier is the result of substantial calculation to obtain the dominant solutions. Then the designer can choose a point on the curve as the desired solution. As one lowers the temperature θ in a uniform heat flux situation, or increases h in an isothermal case, the pressure needed increases. Thus, the price paid in terms of pressure for increasing the heat transfer rate is clearly seen, and the choice is made in terms of acceptable pressure and the associated components needed to supply this pressure. There may be constraints on the design, or it may be unconstrained in terms of size, weight, etc. Cost can also be built into the deliberations for an optimum.

Summary

Considerable challenges arise for heat removal from high-performance and large electronic systems due to increased energy dissipated

per unit area. Microchannel heat sinks provide a valuable means to remove these high inputs without exceeding the stringent temperature limitations generally placed on electronic components. Consequently, recent years have seen an explosion of research and development in this area, starting with air cooling and then moving on to single-phase liquid cooling and to liquids with phase change for very high heat flux levels. By using a variety of heat transfer enhancement techniques, it has been clearly demonstrated that heat flux levels as high as 1000 W/cm^2 and beyond can be handled by using liquid flow in microchannels with and without phase change. The focus has been largely on high heat transfer rates under steady conditions. However, a major problem is the increase in pressure needed for the flow and the associated concerns with delivering this pressure, leakage, and space and weight limitations. The stability, predictability, control, and maintenance of the heat removal system, particularly for phase-change methods, are also of concern. Finally, time-dependent problems are also of considerable interest for changing load and operating conditions, as well as for start-up and shutdown. Many of these aspects have been considered in the literature. Appropriate systems have been designed, optimized, and implemented, particularly for air and single-phase liquid flow. This is an active an important area of research and development with respect to electronic systems, which continue to grow in performance and versatility while reducing in the physical size. Thus the challenges posed to the heat removal systems are expected to increase and microchannel heat sinks suitably modified, designed, and optimized to meet challenges in years to come.

References

1. Steinberg, D.S.: Cooling Techniques for Electronic Equipment, 2nd edn. Wiley-Interscience, New York (1992)
2. Tummala, R.R.: Fundamentals of Microsystems Packaging. McGraw-Hill, New York (2001)
3. Joshi, Y., Kumar, P.: Energy Efficient Thermal Management of Data Centers. Springer, New York (2012)
4. Sathe, S., Sammakia, B.: A review of recent developments in some practical aspects of air-cooled electronic packages. ASME J. Heat Transfer **120**, 830–839 (1998)
5. Incropera, F.P.: Liquid Cooling of Electronic Devices by Single-Phase Convection. Wiley, New York (1999)
6. Incropera, F.P.: Convection heat transfer in electronic equipment cooling. ASME J. Heat Transfer **110**, 1097–1111 (1988)
7. Incropera, F.P., Ramadhyani, S.: Application of Channel Flows to Single-Phase Liquid Cooling of Chips and Multi-Chip Modules. Kluwer, Dordrecht (1994)
8. Tuckerman, D.B., Pease, R.F.W.: High performance heat sinking for VLSI. IEEE Electron Device Lett. **2**(5), 126–129 (1981)
9. Kandlikar, S.G., Colin, S., Peles, Y., Grimella, S., Pease, R.F., Brandner, J.J., Tuckerman, D.B.: Heat transfer in microchannels – 2012: status and research needs. ASME J. Heat Transfer. **135**, 091001-1–091001-18 (2013)
10. Kandlikar, S.G.: History, advances, and challenges in liquid flow and flow boiling heat transfer in microchannels: a critical review. ASME J. Heat Transfer. **134**, 034001-1–034001-15 (2012)
11. Conti, A., Lorenzini, G., Jaluria, Y.: Transient conjugate heat transfer in straight microchannels. Int. J. Heat Mass Transfer **55**, 7532–7543 (2012)
12. Zhang, J., Zhang, T., Prakash, S., Jaluria, Y.: Experimental and numerical study of transient electronic chip cooling by liquid flow in microchannel heat sinks. Numer. Heat Transfer **65**, 627–643 (2014)
13. Sun, Z., Jaluria, Y.: Conjugate thermal transport in gas flow in long rectangular microchannel. ASME J. Electron. Packag. **133**, 021008-1–021008-11 (2011)
14. Sun, Z., Jaluria, Y.: Numerical modeling of pressure-driven nitrogen slip flow in long rectangular microchannels. Numer. Heat Transfer **56**, 541–562 (2009)
15. Steinke, M.E., Kandlikar, S.G.: Review of single-phase heat transfer enhancement techniques for application in microchannels, minichannels and microdevices. Int. J. Heat Technol. **22**, 1–10 (2004)
16. Kandlikar, S., Garimella, S., Li, D., Colin, S., King, M. R.: Heat Transfer and Fluid Flow in Minichannels and Microchannels, 2nd edn. Butterworth-Heinemann, Waltham (2014)
17. Riaz, S., Agonafer, D.: Microchannel Heat Sink: Fabrication and Liquid Cooling System for High Heat Flux Integrated Circuits. VDM Verlag Dr. Muller, Saarbrücken (2010)
18. Kandlikar, S.G.: High flux heat removal with microchannels – a roadmap of challenges and opportunities. Heat Transfer Eng. **26**, 5–14 (2005)
19. Qu, W., Mudawar, I.: Experimental and numerical study of pressure drop and heat transfer in a single-phase micro-channel heat sink. Int. J. Heat Mass Transfer **45**, 2549–2565 (2002)
20. Zhang, J., Lin, P.T., Jaluria, Y.: Designs and optimization of multiple microchannel heat transfer systems. ASME J. Ther. Sci. Eng. Appl. **6**, 011044-1–011044-10 (2014)

Microcontact Printing

Gauri Bhave[1], Ashwini Gopal[1], Kazunori Hoshino[2] and John Xiaojing Zhang[3]
[1]Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, USA
[2]Department of Biomedical Engineering, University of Connecticut, Storrs, CT, USA
[3]Thayer School of Engineering, Dartmouth College, Hanover, NH, USA

Synonyms

Microcontact transfer; Molding; Replica molding; Soft lithography

Definition

Microcontact printing is a method of transferring patterns of various materials such as polymers, proteins, nanoparticles, etc., onto another surface. Typically a polydimethylsiloxane (PDMS) stamp is dipped in a solution of a material that has to be patterned and is brought into contact with the surface. Transfer of micrometer (μm)/nanometer (nm)-size patterns is possible by this technique.

Overview of Microcontact Printing

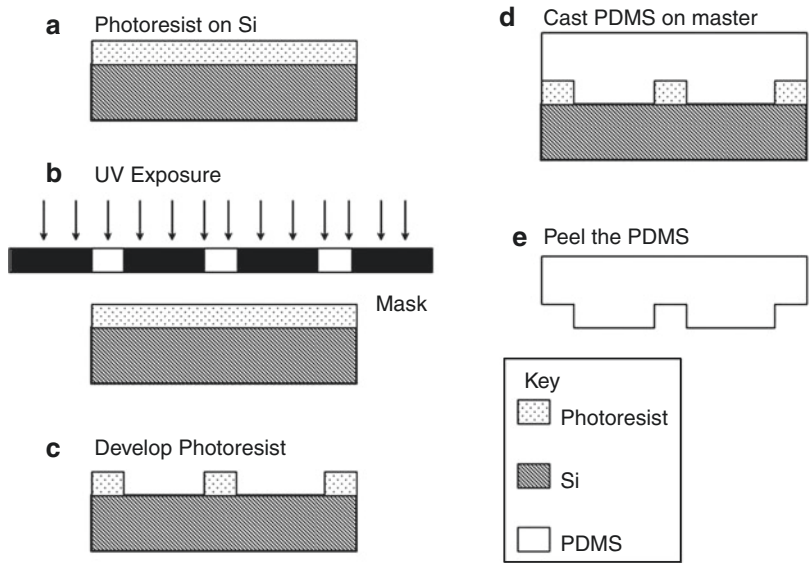
Optical lithography is a fundamental step in microfabrication technology. Almost all integrated circuits are fabricated using a standard lithography process to achieve scalable manufacturing at low cost. However, this method has some shortcomings of compatibility issues, and it cannot be used for substrates that are not flat. Furthermore, the top-down lithography process cannot be easily integrated with labile biological systems because removal of photoresist from the polymers or proteins or quantum dot layer without damaging them would be challenging. Microcontact printing overcomes most of these shortcomings and can be used as a cost-effective alternative. The term soft lithography is coined from the fact that it is a method, which uses patterned elastomer or polymer as a stamp or a mold [1]. There are a number of elastomers that can be used for patterning surfaces. PDMS is a common elastomer used for soft lithography [1]. A number of stamps can be generated from one master mold, and each stamp can be used more than one time for a particular application.

Mold Preparation

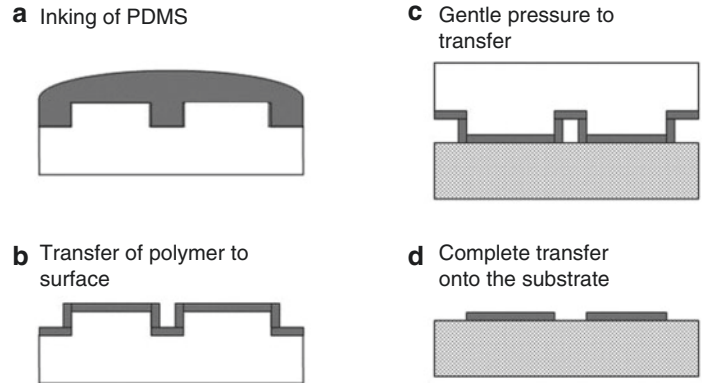
Elastomer molds are usually made with templates from silicon (Si) wafer called "master." First, a silicon master is patterned with high-aspect-ratio photoresist (SU8 or AZ 4562 or AZ 9260) or using deep reactive-ion etching (DRIE) of trenches on silicon substrate using standard semiconductor processing techniques [2]. A thin layer of photoresist is spin coated on a Si wafer. The resist is patterned by ultraviolet (UV) lithography via a mask. The resist is then developed to obtain the pattern onto the master. Polydimethylsiloxane (PDMS, Sylgard 184) is mixed with curing agent (10:1 ratio), degassed, poured onto the master, and degassed again. It is then allowed to cure at 70 °C for 30 min. The PDMS "stamp" is then removed from the master and cut into desired shape and size (Fig. 1). Patterned stamps should have the following properties: they should (a) be peeled off their mold without being damaged, (b) have the accurate replica pattern of their mold, (c) be soft enough to provide conformal contact with substrates, and (d) have patterns that are mechanically stable during inking and printing. To prevent the PDMS from sticking onto substrates, its surface is typically coated with Sigmacote (Sigma Aldrich) or exposed to CHF_3 gas. This would avoid the bonding of PDMS onto the substrate.

Printing Process

The polymer film to be patterned is picked using the PDMS stamp. During the stamping process, only the raised part of the stamp transfers the ink onto the substrate. The inked stamp is then inverted onto a substrate. A gentle pressure is applied for about 30 s such that the polymer transfers completely onto the substrate (Fig. 2). Pompe et al. [3] developed a typical inking method of PDMS stamps. A stamp-pad method was



Microcontact Printing, Fig. 1 Fabricating a PDMS mold (Adapted from Madou [2])



Microcontact Printing, Fig. 2 Microcontact printing of polymer onto a substrate (Adapted from Madou [2])

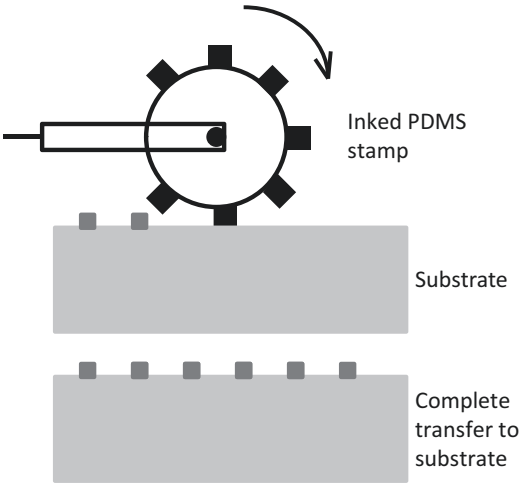
developed where the PDMS stamp was placed in contact with the ink and then the ink is transferred from the stamp to the substrate on the area of contact. This method reduces the swelling of the stamp and minimal absorption of the molecules on the surface after patterning [4]. Inking of the substrates with alkylthiols was also demonstrated by Pompe et al. [3]. The inked areas are self-passivating and have low interfacial tension that repels additional monolayers such that uniform self-assembled monolayers (SAM) are formed on the substrate. The SAM patterns act as localized barrier to wet etching. Sometimes certain conditions require microcontact printing of hydrophobic ink onto the substrate. Typically PDMS is hydrophobic in nature. To overcome this drawback, the PDMS was oxygen plasma treated or alternate polymer based on poly(ether-ether) was used. These inked stamps are patterned on flat substrates or functionalized substrates.

Microcontact printing is a stamping-based method for the transfer of molecules or polymers onto a surface [2]. The stamp is inked with the material or polymer to be transferred by inking techniques. These patterns when pressed against the substrate would selectively transfer the pattern onto the substrate. These elastomers have the capability to pattern on different topographies flat or curved [2]. Deformation of the polymer elastomer allows it to accommodate rough surface. This method is also used to make arrays of multiple units on a large area

simultaneously [5]. The capability of transferring patterns on to curved surfaces such as lens structures has been successfully demonstrated [6]. For surfaces with large radii of curvature, the deformation of the stamp allows for contact with the non-flat surface. For very small radii of curvature, the substrate can be rolled on top a flat elastomeric stamp. Another method which is especially suited to microcontact printing on large areas is "roller imprinting." A cylindrical rolling stamp [7] was inked, dried, and mounted on a roller. This setup was used for patterning surfaces up to 3 in. in diameter.

Recent developments in microcontact printing have focused on patterning with precise alignment. The basic printing process has been modified in various ways to achieve specific patterning goals. One such demonstrated method is vacuum-assisted microcontact printing. This method was demonstrated for patterning of functional materials with precise alignment. The elastomeric stamp was fabricated using double-exposure process to obtain different-height protrusions on the stamp. Printing pressure was applied using a pressure difference between the inside and outside of the stamp. Various materials like proteins and nanomaterials could be successively patterned on a large scale using this printing process [8].

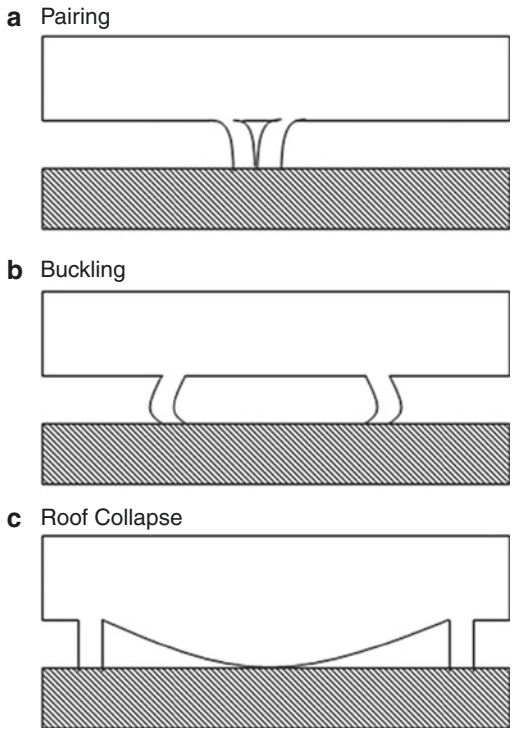
Numerous elastomers have been developed for soft-lithography process. Some of them include polydimethylsiloxane, polyurethanes, polyamides, and phenol formaldehyde polymer. The most commonly used elastomer for soft lithography is PDMS. PDMS is a silicon-based organic polymer. It is flexible, inert, transparent, durable, nontoxic, and not flammable, has modifiable interfacial properties, and allows gas to pass through. PDMS is a liquid at room temperature but can be cured into solid by cross-linking. It is commercially available from Dow Corning, known as Sylgard 184. The master for microcontact printing is typically prepared by photolithography on silicon. The PDMS stamp is peeled from the master. Adhesion of PDMS onto the master substrate is observed in certain cases after peeling. These cause unnecessary contamination to the substrates. A number of methods have been developed to avoid the adhesion of PDMS. Passivation of silicon masters with hydrogen fluoride or treatment with sodium dodecyl sulfate [9] could overcome this issue.



Microcontact Printing, Fig. 3 Microcontact printing of polymer onto a substrate with a rolling stamp (Adapted from Qin et al. [7])

If an elastomeric stamp is too soft, it loses definition to replicate microstructures. Deformation of the stamp causes distorted patterns on the substrate. Stamp deformation can occur during the removal from the template and during contact of the stamp with the substrate. This limits the resolution of patterning. Deformations of the PDMS stamp occur when the aspect ratio is high, and buckling and lateral collapse of the PDMS features can occur. When the aspect ratio is low, roof collapse is possible for the PDMS. Typical deformations that have been observed are due to pairing, buckling, or roof collapse of the structure during contact with the surface. The phenomenon is attributed to the elastomeric nature of the stamp [2]. Figure 3 shows the deformation of such stamps. Thermal expansion of PDMS makes it difficult to get high accuracy across a large area. This limits the application of soft lithography in nanofabrication. Several methods have been proposed to overcome these problems such as developing new stamp materials, adjusting the aspect ratio of the features, and the use of composite stamps [10] (Fig. 4).

There are developments of harder polymers to achieve smaller sizes by increasing the stiffness of a stamp. Thermoplastic elastomers were



Microcontact Printing, Fig. 4 Deformation of the stamps (Adapted from Madou [2])

developed for printing thiols. They are typically polymers such as poly(styrene-b-butadiene-b-styrene) and poly(styrene-b-(ethylene-co-butylene)-b-styrene) that have been used as elastomers for printing. These new materials typically have better mechanical properties, and these materials have low chance to transfer of uncured polymer impurities. These elastomer stamps are produced in minutes, using hot embossing method. They have overcome the problem of residual monomers in the stamp material like PDMS.

Poly(ether-ester) elastomers were used as stamp materials to pattern proteins on surface using lower ink concentrations. The time for inking in these elastomers also is lower compared to PDMS elastomers [11]. PDMS has a tendency to expand on exposure to organic solvents. They tend to lose their conformal shape. UV curable stamps are based on functionalized prepolymer with acrylate groups for cross-linking and different monomeric modulators [12]. These have been used for printing 100-nm hexadecanethiol patterns on gold. These elastomers do not deform on exposure to organic solvents. By varying the modulator, the mechanical properties of the polyurethane acrylate stamp can be varied.

Composite stamps consist of two layers, a stiff layer supported by a flexible layer. Stamps used a thick, flexible 184 PDMS slab to support a thin, stiff *h*-PDMS underlayer. This was used to perform soft lithography to sub-100-nm features. The preparation of these stamps was developed by Schmid et al. for microcontact printing [13].

Key Research Findings

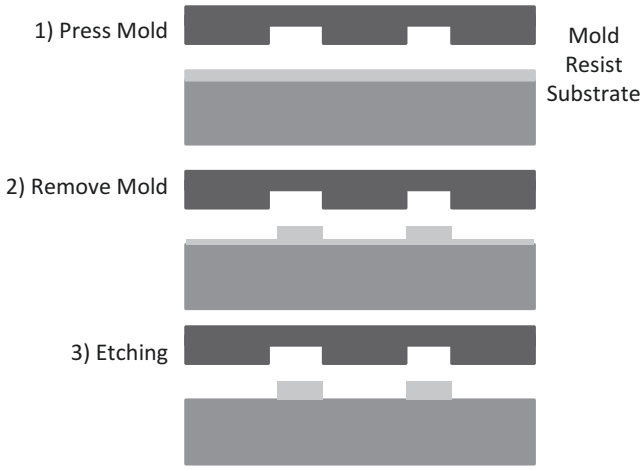
Microcontact printing method was first described in 1993 by Kumar and Whitesides [14]. The authors reported using a patterned stamp with hexadecanethiol as ink brought into contact with gold (Au)-coated surface. The hexadecanethiol molecules form a monolayer in the area of contact between the elastomeric stamp and an Au surface area.

This method was extended to numerous biomedical applications such as patterned cell culture and fabrication of microarrays for biosensor purposes. The range of inking molecules can be extended from alkylsilanes and alkylthiols to various higher molecular weight organic molecules. Printing has been performed for proteins DNA and using Langmuir-Blodgett films [15].

Patterning in Semiconductor Processing

Lithography Masks

Microcontact printing has been used to imprint self-assembled monolayers (SAM) as masks for gold etching in Si semiconductor processes. PDMS stamp was inked with an SAM solution consisting of ethanol or diethyl ether and dried with nitrogen flow [14]. The inked stamp was brought in soft contact with the substrate typically consisting of a thin Au layer. The SAM layers transferred on to the Au surface upon removal of the stamps were used as a mask layer for the gold. These masks could then be used in two ways: (i) patterning of gold using wet etching and [16]



Microcontact Printing, Fig. 5 Nanoimprint lithography demonstrated in Ref. [18]

(ii) as a mask for selective deposition of metals by CVD and electroplating [17].

Nanoimprint Lithography

In this technique, a mold with nanostructures on its surface was pressed into a thin resist cast on a substrate. Upon removal of the mold, the nanostructures on the mold were duplicated on the resist film. This was followed by an anisotropic etching process to remove the residual film. This simple two-step process is shown in Fig. 5 [18].

Gratings and Integrated Optics

Wang et al. showed the use of nanoimprint lithography in the fabrication of a double-layer metal grating as a broadband waveguide polarizer. A transverse magnetic (TM)-pass waveguide polarizer was fabricated with SiN as top layer, SiO_2, and silicon. Nanoimprint lithography, lift-off, and dry etching were used to fabricate the waveguide stripes [19]. Seekamp et al. also investigated the printing of polymer films in polymer waveguides and demonstrated the feasibility of printing a polymer waveguide with micrometer features combined with photonic crystals [20].

Lithographic Molding

Lithographic molding technique has been demonstrated to achieve patterns for use in basic microcontact printing technique. In this technique a normal microcontact printing step was followed by an anisotropic etch into the Si substrate with patterned Au as a mask. The resulting grooves and pyramids were molded again into three-dimensional PDMS stamps used for microcontact printing of smaller patterns. Applied pressure was also finely tuned to control the pattern sizes [21].

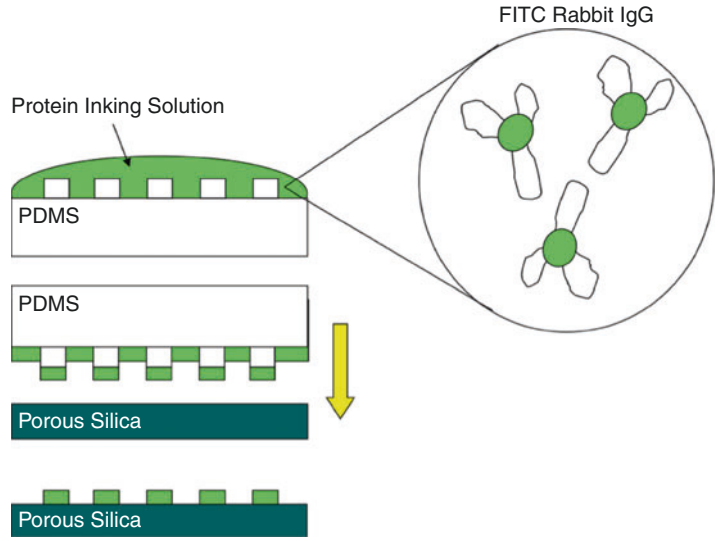
Patterning of Biomaterials

Patterning of Self-Assembled Monolayers

Microcontact printing provides the patterning of self-assembled monolayers (SAM). Many studies involving the patterning of proteins and cells using microcontact printing have used self-assembled monolayers of alkanethiols on gold. This technique is useful to pattern surfaces with multiple SAMs on a single substrate. After a SAM of an alkanethiol is patterned on a gold-coated surface by microcontact printing, the substrate is again dipped into a solution of a second alkanethiol to form a different SAM pattern. This technique has been used to make multiple islands of SAMs that adsorb proteins and cells and which are surrounded by SAMs that resist the adsorption of biomaterial. Patterns of SAMs have been used to study the adhesion of mammalian cells on surfaces [22].

Patterning of Proteins

For applications in biomedical engineering, microcontact printing has been a developed method for printing patterns of biomolecules



Microcontact Printing, Fig. 6 Microcontact printing of proteins onto porous silica substrate demonstrated in Ref. [25]

onto flat surfaces with high resolution (100 nm) [15]. Microcontact printing method was extended to the transfer of protein molecules by Bernard et al. in 2000 [23]. Microcontact printing can be used to pattern multiple proteins on the same substrate, thus allowing for biological assays with greater diagnostic power.

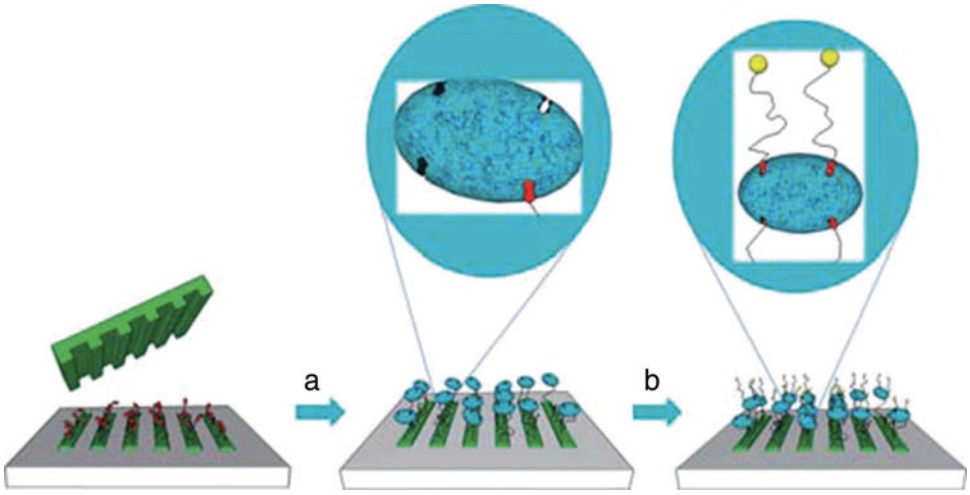
A number of possible applications would be enabled if the same surface could be patterned with proteins of different types [15]. The large-scale (>50 μm) patterning of therapeutic proteins is key to the efficient design, characterization, and production of biologics for cost-effective, high-throughput, and point-of-care detection and analysis system. The hydrophobic nature of PDMS causes the denaturation and irreversible adsorption of the proteins. This can be overcome by surface treatment of the PDMS stamps (chemical modification or oxygen plasma) or by selecting a polar polymer to prepare the stamp.

Bernard et al. used microcontact printing to pattern proteins with a resolution of 1 μm [23]. They found that a monolayer of immunoglobulin antibody (IgG) adsorbed to the PDMS stamp during the inking process and that it was transferred fully after only a few seconds of contact with glass or polystyrene. The stamped immunoglobulins were recognized by secondary antibodies that were attached to them. Kwak et al. demonstrated the patterning of protein cytochrome c using microcontact printing onto gold substrates [24].

Blinka et al. demonstrated an efficient method for protein deposition and adsorption on porous silica substrates in specific patterns using microcontact printing [25]. Relative to the chemically modified and untreated substrates, the porous silica-coated substrates provided for thicker and more uniform protein layers. Protein layers observed on the chemically treated substrates that were thinner than the smallest dimension of the stamped protein were observed, indicating denaturation (Fig. 6).

Indirect Patterning of Proteins

A simple method to place multiple proteins on a surface is to print one that binds another in solution. This has been done to pattern proteins by using biotin-avidin interaction. The avidin is first printed onto the polymeric substrate containing biotin. Avidin binds to biotin moiety with great affinity. Addition of a solution of biotin-conjugated protein leads to the capture of that protein by the available biotin-binding sites on avidin and consequently the patterning of that protein. This can be used to pattern cells [26]. A variation of this method was performed by patterning the biotinylated substrate with avidin molecules. A carboxylic acid group was coated on the polymer surface and activated using



Microcontact Printing, Fig. 7 Printing biotin ligands onto a substrate and subsequent binding of streptavidin or avidin which can be used to direct the immobilization of further biotinylated molecules (Reprinted with copyright permission from Polymer Chemistry [26])

pentafluorophenol. A stamp with the biotin amine group was brought into contact with the surface (Fig. 7).

Patterning of DNA Molecules

Patterning of biomolecules onto solid substrates is needed to provide controlled biocompatible surfaces [15]. Microcontact printing techniques have been used to deposit DNA molecules with chemically modified PDMS surface. Chemical modification enabled the DNA molecules to stick to the PDMS surface. Bernard and his coworkers demonstrated the microcontact printing of DNA on APTES-modified substrates. They demonstrated that microcontact printing has advantages compared to microspotting of DNA solutions. This method has an advantage of reduction in time and DNA material. Spotted stamps can be used to print multiple arrays as well as homogeneous sensitive dots. An alternate method for developing arrays of DNA was developed with different fragments of loading flat stamps with a multichannel microfluidic network. Binding of complementary nucleotides was observed [27].

Indirect Patterning of Cells

Patterning methods have been translated to better understanding on how cells interact with substrates. Islands of printed SAMs promote the adsorption of extracellular matrix protein. Singhvi et al. observed the cells preferentially bound to printed islands and spread to conform their size and shape [28]. Microcontact printing has helped to study the effect of the cell shape on cell division on cell axis orientation. The cell division axis orientation was determined by the adhesion pattern rather than the shape of the cell. Stamping was used to study growth in endothelial and epithelial monolayers. In this study, it was observed that the proliferation was highest in cells at the monolayer edges. These studies demonstrate that the patterning of cells can lead to new findings not possible with traditional cell culture techniques.

Microcontact Printing for Patterning of Nanoparticles, Nanorods, and Nanotubes

Nanoparticles

Patterning colloidal nanoparticle films enables patterning of desired conformations and good quality deposition of quantum dots on substrates. Quantum dot films cannot be patterned using standard process. They are inhomogeneous in nature that contains a semiconductor material and organic capping. Lithography processes are unsuitable because removal of photoresist from

the quantum dot layer would be an issue. Conventional shadow mask technique to deposit particles is undesirable. The quality of quantum dot in solution would degrade using this method. Typically deposition of uniform quantum dots on substrates is critical for creating quantum dot-based light-emitting diode systems.

Microcontact printing has also been developed as an alternate method for depositing nanoparticles by a number of research groups. This method has been used to deposit close packed monolayer of alkanethiol Au nanoparticles [29]. Microcontact printing has also been used for deposition of $(Cd_xZn_{1-x}Se)$ $Cd_yZn_{1-y}S$ core: shell on 4,40-N,N0-diphenylcarbazole (CBP) for fabrication of green diodes [30]. Typically in microcontact, a polydimethylsiloxane (PDMS) stamp is inked with the polymer and is then transferred onto a solid substrate. PDMS substrates can make conformal contact with surfaces over large areas, enabling patterning in nanometer- or micrometer-size features.

As mentioned previously, PDMS stamps have a drawback that it swells on absorption of organic solvents. This changes the shape and dimensions of the protrusions that have to be patterned. Ethanol has minimal swelling effect on the polymer compared to other organic solvents like pentanes or xylenes. They have adverse effect of swelling of the elastomer. PDMS can easily melt in contact with nonpolar solvents such as toluene and chloroform solutions. Microcontact printing is limited to polar inks that are soluble in ethanol.

In this case, conformal deposition becomes difficult. A number of methods have been developed to overcome this problem. Coe-Sullivan et al. have a modified surface of PDMS stamp by coating chemical vapor-deposited parylene-C layer [31]. The particles are then spin coated onto the PDMS surface that is microcontact printed onto the hard substrate. This enables uniform deposition of multilayer of quantum dots onto the substrates. The process has a drawback such that spin coating consumes a large amount of particles; thereby, significant amount of the material is wasted.

Gopal et al. have developed a modified microcontact printing method, where the nanoparticles resuspended in hexane are used in a modified Langmuir-Schaffer technique to create thin films of particles which are picked up by PDMS stamps and microcontact printed on silicon substrates [32]. The quantum dot thin film is self-assembled onto water surface by dispensing a hydrophobic suspension of CdSe/ZnS quantum dots in a solvent mixture of 1,2-dichloroethane and hexane. The uniform film is picked up by the PDMS stamps with nano-/micro-features and deposited onto silicon substrate (Fig. 8).

Hoshino et al. have demonstrated the deposition of quantum dots onto three-dimensional substrates with great accuracy. This method enables accurate control on the deposition of the number of particles of quantum dots onto a substrate [33].

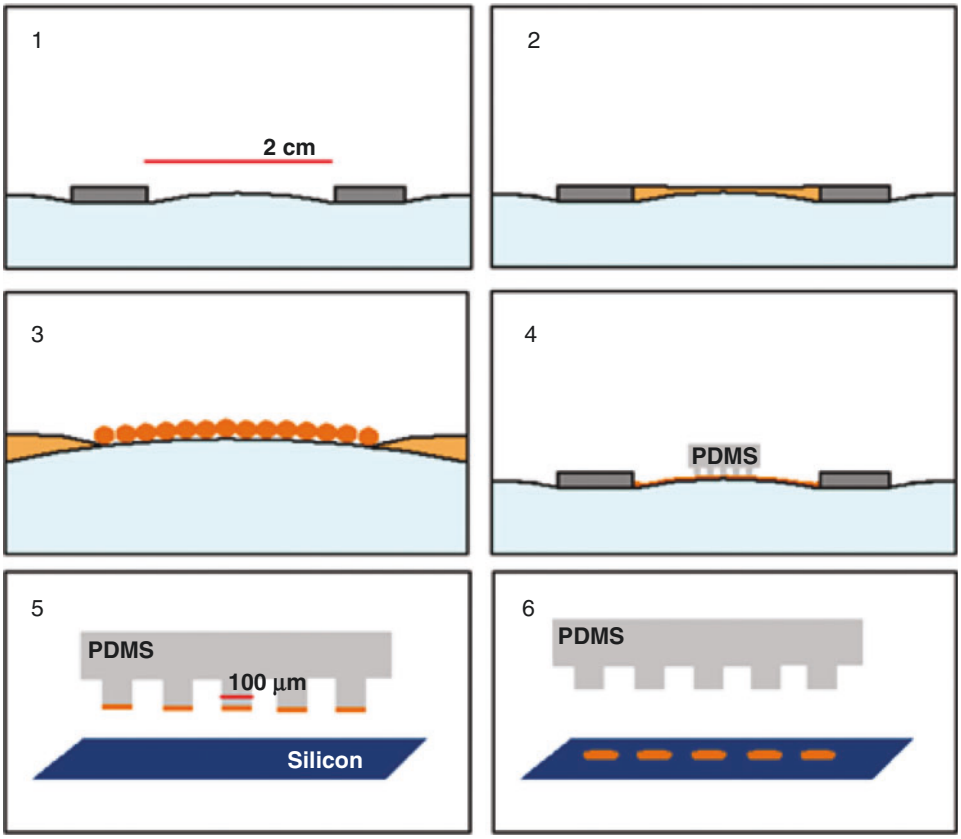
Nanorods and Nanotubes

Murphy et al. developed a method for patterning the growth of Au nanorods. This was achieved by microcontact printing gold nanoparticle seeds on a hard surface. The nanorods then have grown from the seed via seed-mediated growth process [34]. Jen et al. developed a method to deposit 1-D n-type hexaazatrinaphthylene (HATNA) via microcontact printing. The nanostructures were spin coated onto the PDMS stamp; this was allowed to form conformal contact with SiO_2/Si substrate for about 30 s. The nanostructures were transferred onto the surface without additional surface modifications.

A similar method was described for patterning ZnO nanostructures [35]: selective growth of ordered single-crystal ZnO nanostructures based on the microcontact printing of nanocrystal seeding films. The pattern transfer quality is dependent on the concentration of the inking solution. Subsequently, the growth of the ZnO nanowires was carried out with different chemical components.

Future Research Directions

Microcontact printing has the advantage that it can be carried out outside a clean room



Microcontact Printing, Fig. 8 Principle of quantum dot film self-assembly and microcontact printing. These steps illustrate total film self-assembly and microcontact printing process: 1. Convex water surface is created. 2. Colloidal suspension is poured – first, hydrophobic solvent holding the nanoparticles begins to evaporate. Because of the convex water surface, the layer in the center of the surface is the thinnest, exposing the nanoparticles and forming solvent menisci between individual exposed nanoparticles. 3. Film is self-assembled – as the solvent evaporates in the thinnest region, an attractive capillary immersion force collects the nanoparticles into densely packed hexagonal arrays called the nucleus. 4. Hydrophobic PDMS stamp is lowered to pick up the nanoparticle film. 5. PDMS stamp is lowered onto flat substrate. 6. Film is stamped onto the flat substrate [32]

environment conveniently, rapidly, and inexpensively for large features (>50 µm) for cell biology. Through various techniques, it can be scaled down to the nano-regime where it can pattern delicate inorganic nanostructures and biological molecules on a number of materials [36].

This technique is extremely flexible in terms of shapes of the patterns obtained. It enables to control the chemistry at a molecular level. Studies have been developed at reducing some of the drawbacks of microcontact printing such as stamp deformation and ink-transfer issues. Advances allow the technique to be applicable in many fields and enhance the capability of patterning nanometer-scale structures with high precision. The characteristics of microcontact printing can enable new techniques for micro-nanoscale patterning and for new generation of microdevices for biology.

Cross-References

▶ Nanoscale Printing
▶ Photolithography

References

1. Zhao, X.M., Xia, Y., Whitesides, G.M.: Fabrication of three dimensional micro-structures: microtransfer molding. Adv. Mater. **8**, 837–840 (1996)
2. Madou, M.J.: Fundamentals of Microfabrication: The Science of Miniaturization. CRC Press, Boca Raton (2002)
3. Pompe, T., Fery, A., Herminghaus, S., Kriele, A., Lorenz, H., Kotthaus, J.P.: Submicron contact printing on silicon using stamp pads. Langmuir **15**, 2398–2401 (1999)
4. Xiao, P.F., He, N.Y., Liu, Z.C., He, Q.G., Sun, X., Lu, Z.H.: In situ synthesis of oligonucleotide arrays by using soft lithography. Nanotechnology **13**, 756 (2002)
5. Decré, M.M.J., Schneider, R., Burdinski, D., Schellekens, J., Saalmink, M., Dona, R.: Wave printing (I): towards large-area, multilayer microcontact printing. In: Materials Research Society Symposium Proceedings, Boston, pp. 59–61 (2004)
6. Jackman, R.J., Wilbur, J.L., Whitesides, G.M.: Fabrication of submicrometer features on curved substrates by microcontact printing. Science **269**, 664 (1995)
7. Qin, D., Xia, Y., Whitesides, G.M.: Rapid prototyping of complex structures with feature sizes larger than 20 μm. Adv. Mater. **8**, 917 (1996)
8. Kang, H.W., Leem, J., Ko, S.H., Yoon, S.Y., Sung, H. Y.: Vacuum-assisted microcontact printing (mCP) for aligned patterning of nano and biochemical materials. J. Mater. Chem. C **1**, 268 (2013)
9. Lauer, L., Klein, C., Offenhäusser, A.: Spot compliant neuronal networks by structure optimized microcontact printing. Biomaterials **22**, 1925–1932 (2001)
10. Odom, T.W., Love, J.C., Wolfe, D.B., Paul, K.E., Whitesides, G.M.: Improved pattern transfer in soft lithography using composite stamps. Langmuir **18**, 5314–5320 (2002)
11. Trimbach, D.C., Stapert, H., van Orselen, J., Jandt, K. D., Bastiaansen, C.W.M., Broer, D.J.: Improved microcontact printing of proteins using hydrophilic thermoplastic elastomers as stamp materials. Adv. Eng. Mater. **9**, 1123–1128 (2007)
12. Yoo, P.J., Choi, S.J., Kim, J.H., Suh, D., Baek, S.J., Kim, T.W., Lee, H.H.: Unconventional patterning with a modulus-tunable mold: from imprinting to microcontact printing. Chem. Mater. **16**, 5000–5005 (2004)
13. Schmid, H., Michel, B.: Siloxane polymers for high-resolution, high-accuracy soft lithography. Macromolecules **33**, 3042–3049 (2000)
14. Kumar, A., Whitesides, G.M.: Features of gold having micrometer to centimeter dimensions can be formed through a combination of stamping with an elastomeric stamp and an alkanethiol "ink" followed by chemical etching. Appl. Phys. Lett. **63**, 2002–2004 (1993)
15. Ruiz, S.A., Chen, C.S.: Microcontact printing: a tool to pattern. Soft Matter **3**, 168–177 (2007)
16. Xia, Y., Zhao, X.-M., Whitesides, G.M.: Pattern transfer: self-assembled monolayers as ultrathin resists. Microelectron. Eng. **32**, 255 (1996)
17. Xia, Y., Whitesides, G.M.: Soft lithography. Annu. Rev. Mater. Sci. **28**, 153 (1998)
18. Guo, L.J.: Nanoimprint lithography: methods and material requirements. Adv. Mater. **19**, 495 (2007)
19. Wang, J., Schablitsky, S., Zhaoning, Y., Wei, W., Chou, S.: Fabrication of a new broadband waveguide polarizer with a double-layer 190 nm period metal-gratings using nanoimprint lithography. J. Vac. Sci. Technol. B **17**, 2957 (1999)
20. Seekamp, J., Zankovych, S., et al.: Nanoimprinted passive optical devices. Nanotechnology **13**, 581 (2002)
21. Wilbur, J.L., Kim, E., Xia, Y., Whitesides, G.M.: Lithographic molding: a convenient route to structures with sub-micrometer dimensions. Adv. Mater. **7**, 649 (1995)
22. Whitesides, G.M., Ostuni, E., Takayama, S., Jiang, X.: Soft lithography in biology and biochemistry. Annu. Rev. Biomed. Eng. **3**, 335–373 (2001)
23. Bernard, A., Renault, J.P., Michel, B., Bosshard, H.R., Delamarche, E.: Microcontact printing of proteins. Adv. Mater. **12**, 1067–1070 (2000)
24. Kwak, S.K., Lee, G.S., Ahn, D.J., Choi, J.W.: Pattern formation of cytochrome c by microcontact printing and dip-pen nanolithography. Mater. Sci. Eng. C **24**, 151–155 (2004)
25. Blinka, E., Loeffler, K., Hu, Y., Gopal, A., Hoshino, K., Lin, K., Liu, X., Ferrari, M., Zhang, X.J.: Enhanced microcontact printing of proteins on nanoporous silica surface. IOP Nanotechnol. **21**, 415302 (2010)
26. Yang, Z., Chilkoti, A.: Microstamping of a biological ligand onto an activated polymer surface. Adv. Mater. **12**, 413–417 (2000)
27. Lange, S.A., Benes, V., Kern, D.P., Hörber, J.K.H., Bernard, A.: Microcontact printing of DNA molecules. Anal. Chem. **76**, 1641–1647 (2004)
28. Singhvi, R., Kumar, A., Lopez, G.P., Stephanopoulos, G.N., Wang, D.I., Whitesides, G.M., Ingber, D.E.: Engineering cell shape and function. Science **264**, 696 (1994)
29. Santhanam, V., Andres, R.P.: Microcontact printing of uniform nanoparticle arrays. Nano Lett. **4**, 41–44 (2004)
30. Steckel, J.S., Snee, P., Coe-Sullivan, S., Zimmer, J.P., Halpert, J.E., Anikeeva, P., Kim, L.A., Bulovic, V., Bawendi, M.G.: Color-saturated green-emitting QD-LEDs. Angew. Chem. **118**, 5928–5931 (2006)
31. Kim, L.A., Anikeeva, P.O., Coe-Sullivan, S.A., Steckel, J.S., Bawendi, M.G., Bulovic, V.: Contact printing of quantum dot light-emitting devices. Nano Lett. **8**, 4513–4517 (2008)
32. Gopal, A., Hoshino, K., Kim, S., Zhang, X.: Multicolor colloidal quantum dot based light emitting diodes micropatterned on silicon hole transporting layers. Nanotechnology **20**, 235201 (2009)

33. Hoshino, K., Turner, T.C., Kim, S., Gopal, A., Zhang, X.: Single molecular stamping of a sub-10-nm colloidal quantum dot array. Langmuir **24**, 13804–13808 (2008)
34. Jana, N.R., Gearheart, L., Murphy, C.J.: Seed-mediated growth approach for shape-controlled synthesis of spheroidal and rod-like gold nanoparticles using a surfactant template. Adv. Mater. **13**, 1389–1393 (2001)
35. Wang, C.H., Wong, A.S.W., Ho, G.W.: Facile solution route to vertically aligned, selective growth of ZnO nanostructure arrays. Langmuir **23**, 11960–11963 (2007)
36. Kaufmann, T., Ravoo, B.J.: Stamps, inks and substrates: polymers in microcontact printing. Polym. Chem. **1**, 317–381 (2010)

Microcontact Transfer

▶ Microcontact Printing

Microelectrodes

▶ MEMS Neural Probes

Microelectromechanical Systems

▶ Basic MEMS Actuators

Microelectromechanical Systems (MEMS)

▶ Nanomechanical Resonators Based on III–V Semiconductors

Microfabricated MVED (Microwave Vacuum Electron Devices)

▶ MEMS Vacuum Electronics

Microfabricated Preconcentrator-Focuser in Micro Gas Analyzers

▶ Micropreconcentrator Technology for Portable Gas Chromatography System

Microfabricated Probe Technology

Takahito Ono[1] and Masayoshi Esashi[2]
[1]Department of Mechanical Systems and Design, Graduate School of Engineering, Tohoku University, Aoba-ku Sendai, Japan
[2]The World Premier International Research Center Initiative for Atom Molecule Materials, Tohoku University, Aoba-ku Sendai, Japan

Synonyms

Probe technology of scanning probe microscopy (SPM) and its fabrication technology

Definition

Microfabricated microprobes play important roles in SPM. In order to develop functional and high sensitive microprobes, microfabrication technology is employed.

Overview

Scanning Probe Microscopy

Scanning probe microscopy (SPM) families have been widely used as an essential tool in nanoscience, with applications in an increasing range of fields such as nanomaterials, bioscience, and nanoengineering [1]. Also the environment, in which dynamic SPM is required, is broadening, allowing applications such as nanolithography [2, 3], nano-oxidation [4, 5], controlled deposition of molecules [6, 7], fabrication of nanotransistors [8, 9], and molecular modification [10, 11].

Among them, operating SPM in a liquid now became a powerful measurement instrument to investigate the morphology of biological samples with a high resolution in their native solutions [12]. Also near-field scanning optical microscopy (NSOM) has been extensively studied owing to the capabilities for optical imaging with a spatial resolution beyond the diffraction limit of the light [1]. NSOM is also used as an essential tool for observing biological samples, including cells and viruses. In addition, NSOM can be applied to fluorescence imaging and spectroscopy. In these technologies, microfabricated probes play an important role. In this chapter, microprobe technology based on microfabrication will be presented.

Overview

Microfabrication of Near-Field Probe

Together with scanning tunneling microscopy (STM) and atomic force microscopy (AFM), the near-field scanning optical microscopy (NSOM) has been extensively studied owing to the capabilities for optical imaging with a spatial resolution beyond the diffraction limit of the light [13], for subwavelength photolithography [14], and for next generation optical data storage [15] using the near-field light. The most crucial part of the NSOM probe is a subwavelength size aperture at the apex of the tip. Tapered optical fiber probes are most widely used for the NSOM. However, the shape of the tip and the size of the metallic aperture of the fiber tip are not reproducible. The opening angle of the fiber tip is small. Therefore, most of the light is absorbed at the metal wall, which leads to a low optical transmission efficiency. Moreover, it is very difficult to fabricate the fiber NSOM probe in a mass production.

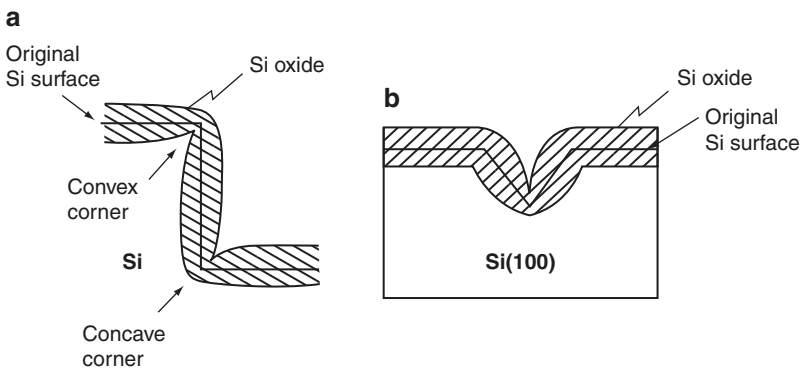
Contrarily, Si based NSOM probe can be easily fabricated in a batch process. It is very easy to combine the microfabricated NSOM probe with a well-developed AFM as an AFM/NSOM probelight [16–20]. In order to create the aperture at the apex of the tip, several technological approaches have been developed such as coating and selective etching metal at the apex of a Si_3N_4 tip [17], metal molding with an aperture at the pyramidal etched pit on Si cantilever [18], or creating the metallic aperture at the apex of a metal-coated SiO_2 tip on Si cantilever by field evaporation [19], etc. However, until recently, a simple method to fabricate the aperture for the NSOM probe with a high reproducibility for mass production has not yet been achieved. Here, a simple method using a low temperature oxidation and selective etching technique to fabricate the aperture for the NSOM probe is presented [21].

Basic Methodology

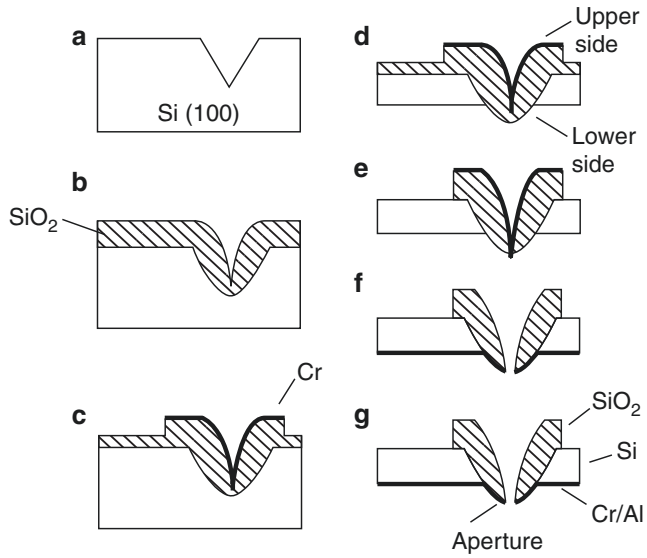
Fabrication and Evaluation of Apertured Probe for NSOM

It is known that the thickness of the oxide grown at a low temperature at convex and concave corners is thinner than that at a flat surface of Si due to the compressive stress at the corner structures (see Fig. 1) [22]. The nonuniform Si oxidation effect at the convex corner has been used for fabrication of Si conical sharp tips for the AFM or field emitter [23]. The anomaly of the thermal oxidation at the concave corner has been used to produce a mold for fabrication of a Si_3N_4 pyramidal sharp tip for the AFM [24].

On the other approach, the nonuniform Si oxidation effect at the pyramidal etched pit is useful to fabricate the aperture for the NSOM probe using the LOSE technique. The process flow of the LOSE technique is schematically shown in Fig. 2. First, a Si(100) wafer is thermally oxidized. Next, photolithography is done to define the mask patterns for etching SiO_2. In order to make a perfect pyramidal etched pit (i.e., four (111) surfaces should intersect at a point), electron beam lithography is used to define square patterns on a resist as a mask for SiO_2 opening. Consecutively, pyramidal etched pits and structure of cantilevers are formed by anisotropic etching of Si in a tetramethyl ammonium hydroxide (TMAH) (step a). Next, a SiO_2 film is thermally grown by wet oxidation at 950 °C with a thickness of about 1 μm (step b). Subsequently, a chromium (Cr) film of about 100 nm in thickness is sputtered and lifted



Microfabricated Probe Technology,
Fig. 1 Schematic of Si thermal oxidation at the convex and concave corner (**a**) and at the pyramidal etched pit formed on Si (100) (**b**)



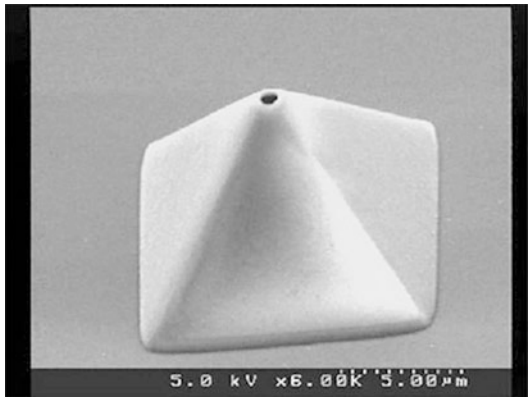
Microfabricated Probe Technology,
Fig. 2 Process flow for making aperture at the apex of the SiO_2 tip on Si cantilever for NSOM probe using the LOSE technique

off to form a protective pattern on the upper side of the oxidized pit (step c). Using this Cr pattern as a mask, the SiO_2 of about 0.6 μm in thickness is partly etched as shown in step c. The remaining oxide of about 0.4 μm in thickness is needed for following steps. Next, the Si wafer is anisotropically etched from the back side in the TMAH solution until forming Si cantilever with the SiO_2 tip at the end of the cantilever (step d). The wafer is then dipped into the buffered-HF (50 % HF:40 % NH_4F, 9 cm^3:100 cm^3) (BHF) solution for selective etching SiO_2 until the Cr protrusion comes in sight (step e). At this step, only exterior wall of the SiO_2 tip is etched in the BHF while the interior wall is protected by the Cr pattern. Finally, a metal film (Cr, or Cr-Al) is entirely deposited onto the back side of the cantilever to form an opaque layer (step g).

SEM image of a typical Cr-coated SiO_2 tip with an aperture of about 500 nm diameter fabricated using this technique is shown in Fig. 3. It can be seen that the aperture is exactly situated at the apex of the tip. Pinholes were not found on the sidewall of the tip. The aperture sizes ranging from 10 to 500 nm are experimentally achieved, corresponding to etching time ranging from 3 to 6 min in BHF at 36 °C. In the case of the patterning window for the etched pit in step a, Fig. 2 is not very square or round, double apertures can be formed. The distance between the centers of two apertures is as same as the discrepancy of the opening window of the etch pit. Normally, the size of the two apertures in the double one is identical. In order to fabricate only one small aperture, the patterning window for the etch pit

should be done by electron beam lithography. The fabricated silicon probe is bonded with a glass substrate on which an electrode is formed. A three-dimensional view of the probe is shown in Fig. 4. The probe consists of the Si cantilever with the tiny aperture at the apex of the tip and the opposed parallel electrode with a narrow gap for capacitance detection or electrostatic actuation. SEM images of the fabricated probe and a close-up view of the aperture with about 25 nm diameter are shown in Fig. 5a, b.

The throughput of several apertures is evaluated and plotted, as shown in Fig. 6. It can be seen that the throughput of the probe is several orders of magnitude higher than that of the conventional optical fibers [25, 26] and is also higher than other published throughput on the micromachined tips [27, 28]. The reason for that is the opaque layer is formed on the outer curvature of the SiO_2 tip. The outer curvature angle of the thermal SiO_2 tip grown at low temperature at the apex area is very large (almost flat in the area of about $0.5 \times 0.5\ \mu m^2$ at the apex or nearly equal to the curvature of a sphere with about 1 μm diameter). Therefore, the cut-off effect (if the incident light passes through a narrow wave-guide of which diameter is less than the wavelength, absorption of light by the coated metal results in decrease the throughput) is minimized. The laser light beam is come to the aperture without any strong losses inside the tip. Furthermore, the aperture is empty, i.e., there is no glass core as in the optical fiber tip; therefore, many modes of the light can reach to the aperture and transmit through the aperture as same as transmitting through a small aperture in a perfectly conducting plane [29]. However, for the aperture smaller than 100 nm, the throughput is still drastically decreased.



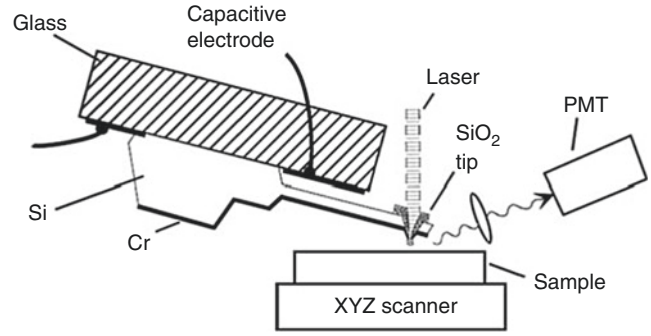
Microfabricated Probe Technology, Fig. 3 SEM image of a typical NSOM tip with an aperture of about 500 nm

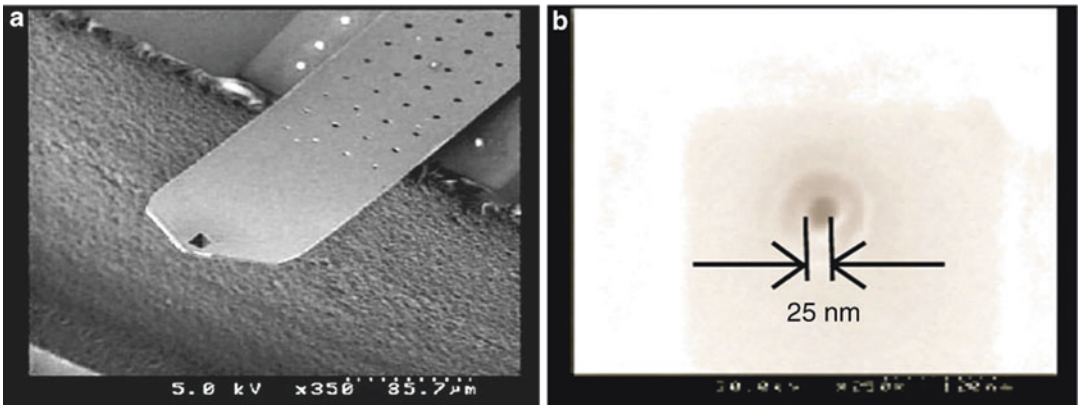
Key Research Findings

Imaging Using the Apertured Probe

Imaging with fabricated aperture probe is demonstrated. A simple characterization setup for the fabricated NSOM/AFM probe in contact and tapping modes is used. The fabricated NSOM/AFM probe is mounted on the AFM stage. The tip-sample distance is kept constant by utilizing the well-developed AFM technique. The light beam from the laser diode of 780 nm wavelength, 2 mW power is directly focused on the tip area using an objective lens. A part of the incident light is reflected by the tip's walls and is monitored by a deflection detector. The near-field light passing

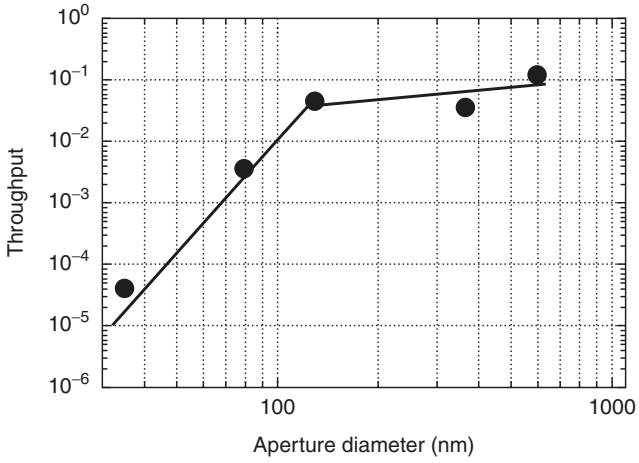
Microfabricated Probe Technology, Fig. 4 Schematic view of the NSOM/AFM probe



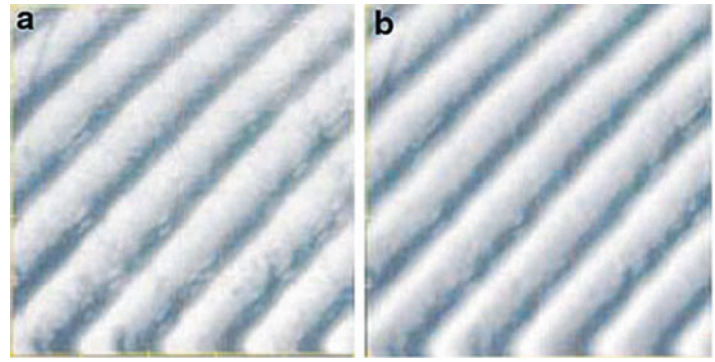


Microfabricated Probe Technology, Fig. 5 (**a**) Typical SEM image of the fabricated NSOM/AFM probe using the local oxidation and etching technique. (**b**) SEM image of an approximately 25-nm-diameter aperture at the apex of the tip on fabricated NSOM probe

Microfabricated Probe Technology, Fig. 6 Throughput of the fabricated aperture NSOM probes as a function of the aperture diameter



through the aperture is reflected by the surface of the sample, is collected by a long distance working lens, and is detected by a photomultiplier tube (PMT). The signal from the output of the PMT is amplified and acquired by the AFM controller as an optical signal. Using the described measurement setup and the fabricated probe with the aperture of about 120 nm, NSOM and corresponding AFM images in contact mode of 200 nm Cr grating on glass substrate are simultaneously observed as shown in Fig. 7a, b. It should be noted, however, that even with 120 nm aperture size, very small Cr particles (about 20 nm diameter) can be observed. The system presented here is specialized for illumination mode; however, it is possible to use the fabricated aperture tip as a receiver for collection mode NSOM. It is also expected to develop this setup for optical near-field and Raman near-field spectroscopy as well. The setup can also be used for NSOM/AFM in non-contact or tapping modes. The probe is mounted on an AFM non-contact stage on which the cantilever can be vibrated by a piezo element. The near-field signal modulated by the vibration signal is lock-in amplified using the vibration signal applied to the piezo element as a reference and acquired by the AFM controller.



Microfabricated Probe Technology, Fig. 7 (**a**) The AFM and (**b**) corresponding illumination mode NSOM images in contact mode of 200 nm Cr grating on glass substrate using the fabricated NSOM probe with the aperture of about 120 nm diameter

Overview

Quartz-Crystal SPM Probe

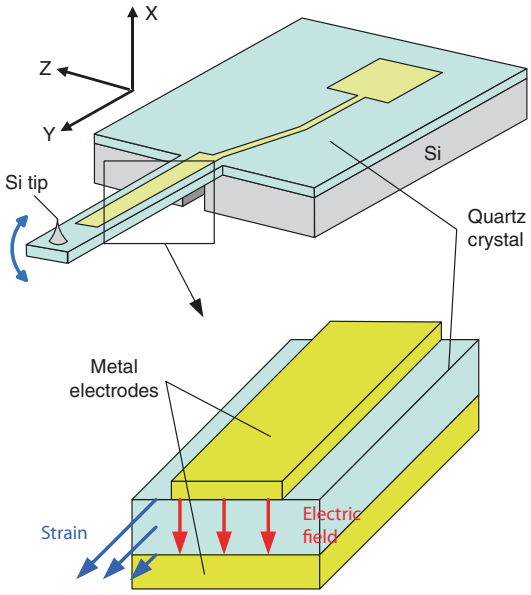
SPM can be used for imaging of a biological sample in liquid. There are several difficulties in liquid operation of dynamic-mode SPM (DM-SPM). The quality factor (Q factor) severely drops due to the viscous damping in a liquid, and that leads to low spatial resolution [30]. In addition, when the cantilever is driven indirectly with acoustic excitation, several spurious resonance peaks are observed around the resonance peak of the cantilever [31]. Thus, it is difficult to identify the "true" resonance peak of the cantilever, and in a worst case, one cannot operate DM-SPM in the liquid solution. One of the solutions for these problems is to use quartz-crystal tuning fork (QC-TF) as a force sensor of DM-AFM. QC-TF has higher Q factor than that of conventional silicon cantilevers [32–39], and one of quartz sensor [36] with both actuation and sensing performances has been available on the market from Nanosensor™ [40]. QC-TF SPM sensors with attached tip can be applied to atomic resolution imaging in high vacuum [35] also in ambient atmosphere [41], high-speed imaging [42], and can be operated even at low temperatures [43]. QC-TF are also used for other applications such as resonant biosensors operated in liquid [44], and mass sensors [45].

QC-TF can vibrate by itself by applying an AC voltage near the resonant frequency, so it expected to causes no spurious resonances. In addition, the vibration signal can be detected electrically; thus, no optics is needed. This self-sensing property is advantageous to the simplicity of the SPM system and the ability of utilizing in an opaque liquid. However, the spring constant of conventional QC-TF is on the order of a few thousands of N/m [34], which is much larger than that of a conventional silicon probe and results in reducing the force sensitivity and the injury of the soft biological sample. For SPM in liquid with high resolution and usability, self-sensing and self-actuation are desirable. Although QC-TF has both properties in these points, lower spring constant and more convenient method of the probe tip formation are needed. Additionally, the preparation of the tip requires great care, i.e., the tip has to be mounted on one of the prongs of the QC-TF piece by piece. Recently various quartz-crystal sensors (QC-S) have been developed for advanced performances, and some of them have been commercialized. Needle-type QC resonator with a high frequency ~1 MHz can improve the scan speed [33]. Cantilevered QC sensors have been developed for decreasing the spring constant [38], also cantilevered sensors with ability of self-sensing and self-actuation using three electrodes have been reported [39]. Using advanced microfabrication technology, a silicon microresonator is combined with QC-TF sensors [36]. Here, the design, the fabrication, and characterization of new quartz-crystal force sensor with a low spring constant are presented.

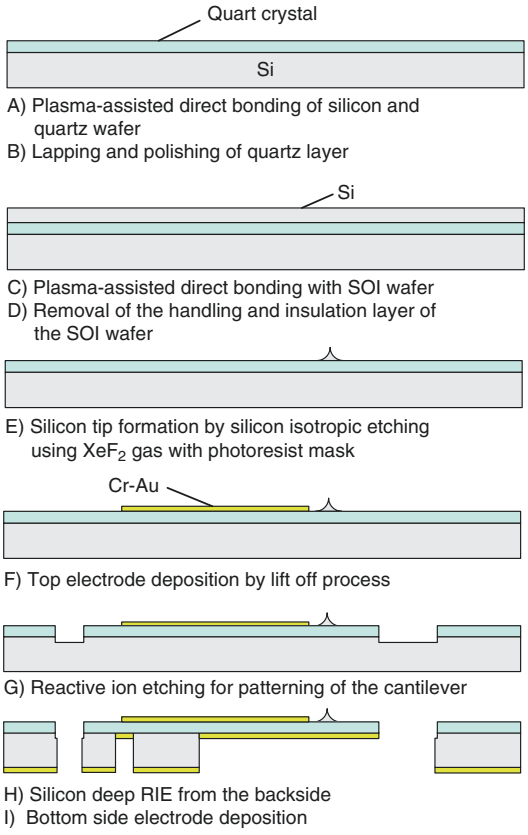
Basic Methodology

Design and Fabrication of Quartz Cantilever Sensors

Schematic view of the quartz cantilever is shown in Fig. 8 [46]. The quartz-crystal cantilever is



Microfabricated Probe Technology, Fig. 8 Schematic of the quartz-crystal cantilever (*up*) and the excitation principle of the flexural mode vibration (*bottom*)

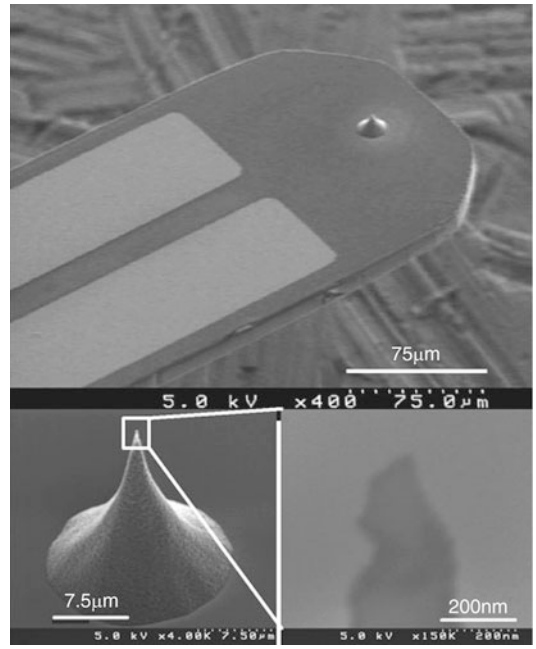


Microfabricated Probe Technology, Fig. 9 Process flow of the quartz-crystal cantilever with an integrated silicon tip

made from X-cut quartz plate that is bonded with silicon substrate. Y axis of the quartz-crystal is designed to be parallel to the longitudinal direction of the cantilever. On the both sides of the cantilever, metal electrodes are formed. The top electrode pattern size is designed to be narrower than that of bottom electrode. By applying a voltage between two electrodes, strain is generated in y direction due to inverse piezoelectric effect. The generated strain in y direction is distributed in thickness direction because of the asymmetric pattern sizes of the top and bottom electrodes; therefore, the applied AC voltage near the resonant frequency generates the flexural vibration of the cantilever (self-actuation). By measuring the generated charge using a current amplifier, vibration signal can be detected with no optics (self-sensing). A sharp silicon tip is integrated on the free end of the cantilever.

Fabrication process flow is shown in Fig. 9. Starting materials are a 100-µm-thick, both side mirror-polished X-cut quartz-crystal wafer and both side mirror-polished 200 µm-thick (100)-oriented silicon wafer. Before the bonding process, each wafer was cleaned by a megasonic cleaner using deionized water in order to remove contaminants such as particles, organic matters, and metals contaminants. Subsequently, plasma treatment for surface activation was performed on the quartz wafer using N_2 plasma with gas pressure of 0.8 mbar, RF power of 150 W, and activation time of 30 s. In similar way, the surface of the silicon wafer is activated by N_2 plasma (0.4 mbar, 100 W, 30 s). After this surface activation treatment, only silicon wafer was rinsed with deionized water. Both wafers were contacted together in ambient atmosphere at room temperature. Then, this pre-bonded wafer was annealed at 180 °C, which increases the bonding strength. Surface roughness and morphology of the bonding surfaces are critical parameters for bonding. If wafer's surface is too rough, bonding will fail despite using surface activation. A surface

roughness of less than 1 nm in Ra was needed to obtain enough strength in our experiments. The bonded quartz layer was thinned down to the designed thickness (2 ~ 10 μm) by conventional lapping and polishing techniques. Then a SOI wafer (20/1/400 μm) is bonded with the polished quartz layer, and handling layer and insulating layer are removed. Next, silicon sharp tips are formed by XeF_2 anisotropic etching using photoresist as a mask. After depositing Cr-Au with thicknesses of 20 and 100 nm, respectively, an electrode pattern was formed by lift-off process. The cantilever shape is defined by reactive ion etching (RIE) using a mixture gas of SF_6 and Xe gases with a ratio of 1/1 at process pressure of 0.27 Pa. As a masking material for RIE process, we used a thick photoresist (approximately100-μm thick). Then, cantilevers are released by silicon deep RIE from the bottom side. Then, Cr-Au with thicknesses of 20 and 100 nm, respectively, as an electrode was deposited on the entire area of the bottom side. Finally, using catalytic chemical vapor deposition, a thin SiCN film was deposited on the electrodes to prevent current leakage to a sample. Typical SEM image of the fabricated quartz-crystal cantilever is shown in Fig. 10. In this cantilever, two metal electrodes are formed on the top side for detection and excitation. The tip is electrically separated from the excitation electrode; therefore, electric field effect to a biological sample can be minimized. It can be seen that the silicon tip is very sharp and the curvature radius of the tip is below 50 nm. Sensitivity for external force depends on the spring constant of the cantilever. By using bonding and thinning technique, the thickness of the cantilever can be tuned. The smallest spring constant is 0.1 N/m (length, width and thickness of the cantilever are $l = 800$ μm, $w = 200$ μm, $t = 2$ μm, respectively). This value is approximately 10^4 times smaller than that of conventional quartz tuning fork.



Microfabricated Probe Technology, Fig. 10 SEM image of the fabricated quartz-crystal cantilever integrated with a sharp silicon tip. *Bottom left* is the magnified view of the silicon tip. The tip radius of the silicon tip is below 50 nm

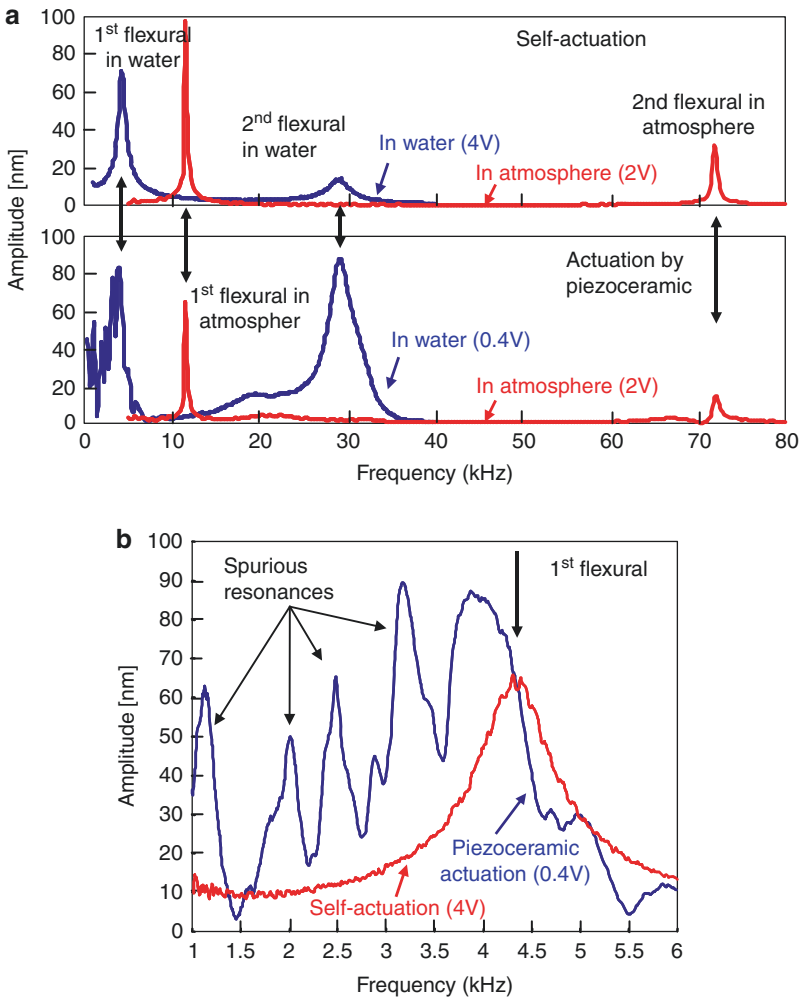
Key Research Findings

Evaluation of Quartz-Crystal Sensor

The vibration characteristics are measured in atmosphere and in water. Also mechanical response spectra actuated by the self-actuation and the acoustic actuation using piezoceramics are compared. In order to evaluate the difference between conventional excitation method and self-actuation, the quartz cantilever is excited using piezoceramics placed near the cantilever or excited by self-actuation.

The amplitude of the quartz cantilever is measured by a laser Doppler vibrometer and lock-in amplifier. In liquid, the resonant frequency and quality factor decline due to the viscosity of the liquid, as shown in Fig. 11a. The measured resonant frequencies in atmosphere and water are 11.67 and 4.7 kHz, respectively, and the quality factors are 270 and 7, respectively.

As shown in Fig. 11b, the self-actuation of the quartz cantilever using the inverse piezoelectric effects is very effective to remove spurious resonances in liquids. No spurious resonances are observed in the self-actuation, while many spurious resonances can be seen in the case of the acoustic excitation in water using the



Microfabricated Probe Technology, Fig. 11 Vibration characteristics of the X-cut quartz-crystal cantilever: (**a**) Mechanical vibration spectra in water and atmosphere; (**b**) comparison of mechanical response spectra between self-actuation and acoustic excitation using piezoceramics in water near the first resonance frequency

piezoceramics. This is advantageous for SPM operation in liquids.

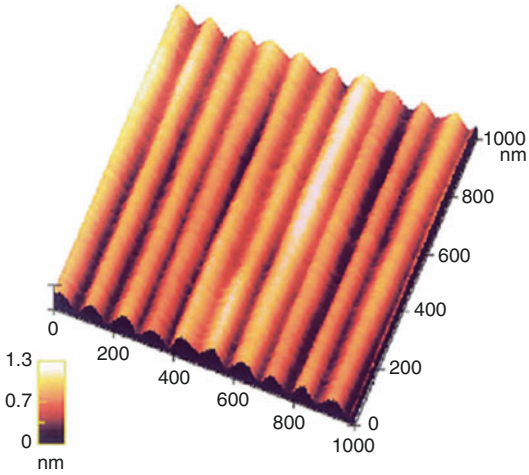
SPM imaging using a fabricated quartz cantilever was demonstrated at the amplitude-modulation SPM mode. A commercial SPM (Olympus, NV2000) was used in this experiment. A quartz cantilever (*length*: $l = 800$ μm, *width*: $w = 200$ μm, thickness: $t = 8$ μm) was actuated by applying 2 Vp-p of AC voltage, and the vibration was detected using a laser displacement sensor of the SPM. Flexural first mode with a fundamental resonant frequency of 10.7 kHz was employed for the measurement. The nominal spring constant is 4 N/m, amplitude of the cantilever is 30 nm, and the measured Q factor is 300. Figure 12 shows the demonstrated SPM image of the silicon grid pattern.

In a liquid, the Q factor of the sensor decreases due to viscous damping, which declines the sensor sensitivity. In order to increase the Q factor of the quartz-crystal cantilever, electrical Q-enhancement was tested. This method is so-called active Q control. Figure 13 shows the results of active Q control. Displacement signal of the cantilever from the laser Doppler vibrometer is added to the oscillation AC signal through an amplifier and phase shifter. The Q factors could be increased up to 6,300 from 270 in atmosphere, and up to 100 from 7 in water. Thus, it is found that the quartz-crystal cantilever is very effective

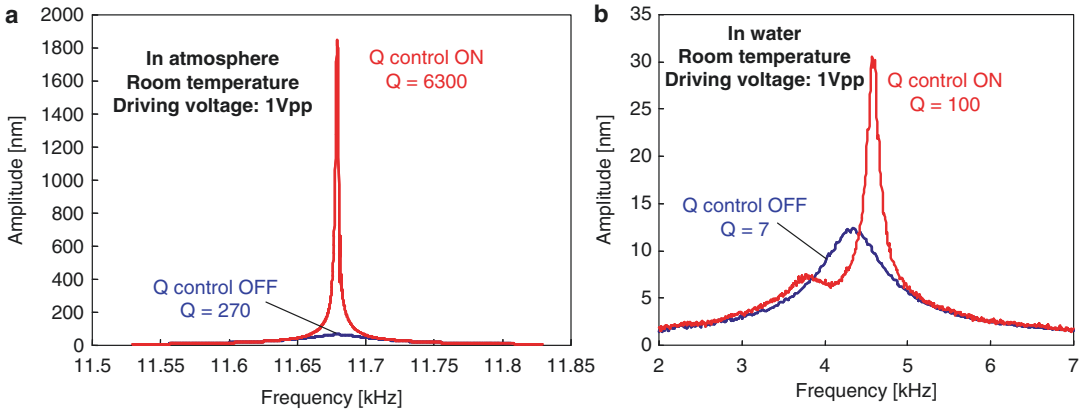
to enhance the Q factor using the self-oscillation function.

Self-sensing property was examined with coupled quartz-crystal cantilevers. Two cantilever structures are mechanically coupled via thin quartz plate formed at the cantilever support. In this coupled quartz resonator, the detection electrode and excitation electrode are electrically isolated by each other. Therefore, electrical cross talk between excitation signal and detection signal can be minimized, and from observation using an oscilloscope, no cross talk was recognized at off-resonance frequency. In addition, in this coupled resonator, the self-oscillation based on self-sensing becomes easy in comparison with the sensor with two electrodes. The dimensions of the rectangle beam are $l = 200$ μm, $w = 150$ μm, $t = 5$ μm. The calculated spring constant of the beam is approximately 50 N/m. Frequency response of the quartz resonator measured by laser Doppler vibrometer is shown in Fig. 14. This is actuated with driving AC voltage of 1Vp-p. This quartz resonator has two flexural vibration modes, i.e., 97.3 kHz of in-phase mode and 120.7 kHz of anti-phase mode. The anti-phase mode vibration exhibits higher Q factor ($Q = 732$) than that of in-phase mode ($Q = 428$), due to the suppression of the support energy loss.

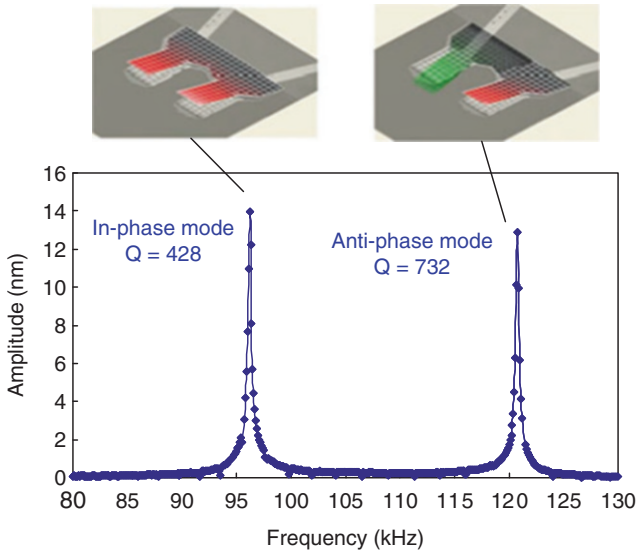
For self-sensing, AC voltage was applied to one cantilever, and the generated current from the other cantilever was detected using a current-to-voltage converter with a feedback resistance of 10 MHz. The bottom side electrode was grounded. The parallel capacitance of the quartz resonator was canceled by active bridge circuit. The output current of the current-to-voltage converter was successfully measured using a lock-in amplifier. It is expected that this quartz resonator can be applied to the SPM imaging with abilities of self-actuation and self-sensing.



Microfabricated Probe Technology, Fig. 12 Typical SPM image of the grating obtained using the fabricated quartz cantilever. Image size is 1,000 × 1,000 nm^2



Microfabricated Probe Technology, Fig. 13 Mechanical spectra of the quartz-crystal cantilever without and with Q-control method (**a**) in atmosphere and (**b**) in liquid



Microfabricated Probe Technology,
Fig. 14 Frequency response of the tuning fork–type quartz-crystal resonator measured using a laser Doppler vibrometer. The driving voltage is 1Vp-p

Cross-References

- ▶ AFM in Liquids
- ▶ AFM Probes
- ▶ Atomic Force Microscopy
- ▶ Scanning Near-Field Optical Microscopy

References

1. Jalili, N., Laxminarayana, K.: Mechatronics **14**, 907 (2004)
2. Salaita, K., Wang, Y.H., Mirkin, C.A.: Nat. Nanotechnol. **2**, 145 (2007)
3. Hoeppener, S., Maoz, R., Cohen, S.R., Chi, L.F., Fuchs, H., Sagiv, J.: Adv. Mater. **14**, 1036 (2002)
4. Albonetti, C., Martinez, J., Losilla, N.S., Greco, P., Cavallini, M., Borgatti, F., Montecchi, M., Pasquali, L., Garcia, R., Biscarini, F.: Nanotechnology **19**, 435303 (2008)
5. Losilla, N.S., Martinez, J., Garcia, R.: Nanotechnology **20**, 475304 (2009)
6. Losilla, N.S., Oxtoby, N.S., Martinez, J., Garcia, F., Garcia, R., Mas-Torrent, M., Veciana, J., Rovira, C.: Nanotechnology **19**, 455308 (2008)
7. Martinez, R.V., Martinez, J., Chiesa, M., Garcia, R., Coronado, E., Pinilla-Cienfuegos, E., Tatay, S.: Adv. Mater. **22**, 588 (2010)
8. Martinez, J., Martinez, R.V., Garcia, R.: Nano Lett. **8**, 3636 (2008)
9. Martinez, R.V., Martinez, J., Garcia, R.: Nanotechnology **21**, 245301 (2010)
10. Garcia, R., Losilla, N.S., Martinez, J., Martinez, R.V., Palomares, F.J., Huttel, Y., Calvaresi, M., Zerbetto, F.: Appl. Phys. Lett. **96**, 143110 (2010)
11. Yoshida, S., Ono, T., Esashi, M.: Nanotechnology **19**, 475302 (2008)
12. Cohen, S.R., Bitler, A.: Curr. Opin. Coll. Interface Sci. **13**, 316 (2008)
13. Betzig, E., Trautman, J.K.: Science **257**, 189 (1992)
14. Ghislain, L.P., Elings, V.B., Crozier, K.B., Manalis, S. R., Minne, S.C., Wilder, K., Kino, G.S., Quate, C.F.: Appl. Phys. Lett. **74**, 501 (1999)
15. Betzig, E., Trautman, J.K., Wolfe, R., Gyorgy, E.M., Finn, P.L., Kryder, M.H., Chang, C.-H.: Appl. Phys. Lett. **61**, 142 (1992)
16. Ruter, A.G.T., Moers, M.H.P., van Hulst, N.F., de Boer, M.: J. Vac. Sci. Technol. B **14**, 597 (1996)
17. Drews, D., Ehrfeld, W., Lacher, M., Mayr, K., Noell, W., Schmitt, S., Abraham, M.: Nanotechnology **10**, 61 (1999)
18. Mihalcea, C., Scholz, W., Werner, S., Munster, S., Oesterschulze, E., Kassing, R.: Appl. Phys. Lett. **68**, 3531 (1996)
19. NgocMinh, P., Ono, T., Esashi, M.: Rev. Sci. Instrum. **71**, 3111 (2000)
20. Minh, P.N., Ono, T., Esashi, M.: Presented at the proceeding of Eleventh IEEE International Conference on Micro Electro Mechanical Systems, 1999, Sens. Actuators A, 80 p. 163. Orlando (2000)
21. NgocMinh, P., Ono, T., Esashi, M.: Appl. Phys. Lett. **75**, 4076 (1999)
22. Maruss, R.B., Sheng, T.T.: J. Electrochem. Soc. **129**, 1278 (1982)
23. Marcus, R.B., Ravi, T.S., Gmitter, T., Chin, K., Liu, D., Orvis, W.J., Ciarlo, D.R., Hunt, C.E., Trujillo, J.: Appl. Phys. Lett. **56**, 236 (1990)
24. Akamine, S., Quate, C.F.: J. Vac. Sci. Technol. B **10**, 2307 (1992)
25. Ohtsu, M. (ed.): Near-Field Nano/Atom Optics and Technology. Springer, Tokyo (1998)

26. Valascovic, G.A., Holton, M., Morrison, G.H.: Appl. Opt. **34**, 1215 (1995)
27. Zhou, H., Midha, A., Bruchhaus, L., Mills, G., Donaldson, L., Weaver, J.M.R.: J. Vas. Sci. Technol. B. **17**, 1954 (1999)
28. Lee, M.B., Atoda, N., Tsutsui, K., Ohtsu, M.: J. Vac. Sci. Technol. B. **17**, 2462 (1999)
29. Fischer, U.C., Zapletal, M.: Ultramicroscopy **42–44**, 393 (1992)
30. Song, Y., Bhushan, B.: Ultramicroscopy **107**, 1095 (2007)
31. Humphris, A.D.L., Round, A.N., Miles, M.J.: Surf. Sci. **491**, 468 (2001)
32. King, G.M., Lamb, J.S., Nunes Jr., G.: Appl. Phys. Lett. **79**, 1712 (2001)
33. Clauss, W., Zhang, J., Bergeron, D.J., Johnson, A.T.: J. Vac. Sci. Technol. B **17**, 1309 (1999)
34. Seo, Y., Hong, S.: Mod. Phys. Lett. **19**, 1303 (2005)
35. Giessibl, F.J.: Appl. Phys. Lett. **76**, 1470 (2000)
36. Akiyama, T., Staufer, U., Rooij, N.F., Frederix, P., Engel, A.: Rev. Sci. Instrum. **74**, 112 (2003)
37. Dida, H., Shikida, M., Fukazawa, K., Murakami, S., Sato, K.E., Asaumi, K., Iriye, Y., Sato, K.A.: Sens. Actuator A **148**, 311 (2008)
38. Ono, T., Lin, Y.-C., Esashi, M.: Appl. Phys. Lett. **87**, 074102 (2005)
39. Chen, C.J., Schwarz, A., Wiesendanger, R., Horn, O., Müller, J.: Rev. Sci. Instrum. **81**, 053702 (2010)
40. http://www.nanosensors.com
41. Seo, Y., Choe, H., Jhe, W.: Appl. Phys. Lett. **83**, 1860 (2003)
42. Giessibl, F.J.: Appl. Phys. Lett. **73**, 3956 (2008)
43. Hembacher, S., Giessible, F.J., Mannhart, J.: Appl. Surf. Sci. **188**, 445 (2002)
44. Su, X., Dai, C., Zhang, J., O'Shea, S.J.: Biosens. Bioelectron. **17**, 111 (2002)
45. Zhang, J., O'Shea, S.: Sens. Actuator B **94**, 65 (2003)
46. Takahashi, A., Esashi, M., Ono, T.: Nanotechnology **21**, 405502 (2010)

Microfluidic Device

► Lab-on-a-Chip for Studies in *C. elegans*

Microfluidic Devices

► Lattice Boltzmann Methods for Nanofluidics

Microfluidic Optomechanics

Gaurav Bahl
Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Synonyms

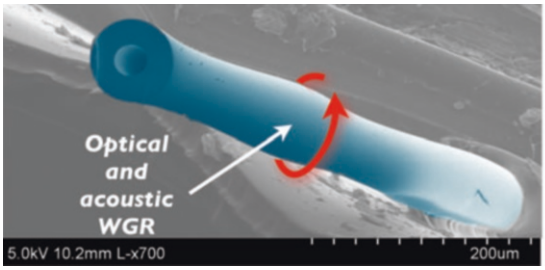
OMFR; Opto-mechano-fluidics; uFOM

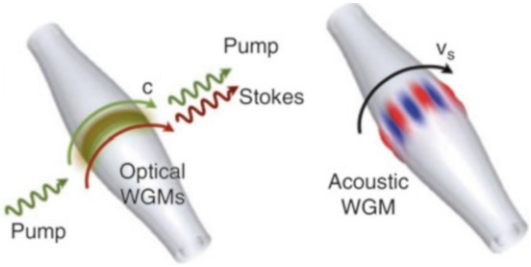
Definition

Microfluidic optomechanical resonators (opto-mechano-fluidic resonators or OMFRs) enable coupling between light, solids, and fluids by means of optical forces and opto-acoustic scattering. This is achieved through solid–fluid hybrid modes that span vibrational frequencies in the MHz–GHz range. Using these interactions, microfluidic optomechanical devices can be used to perform acoustic rheological measurements on fluids using only light.

Principle of Operation

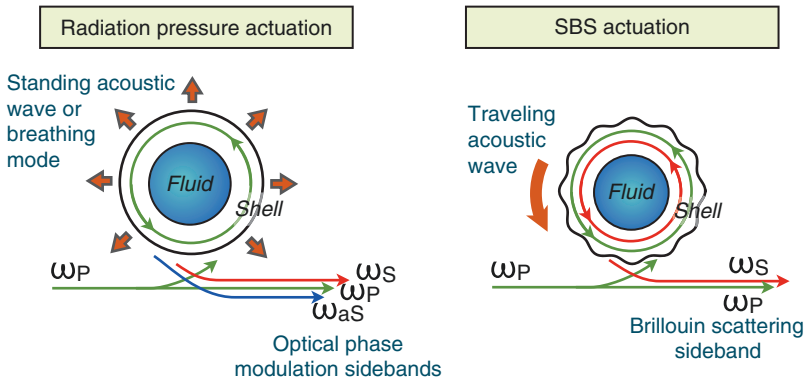
OMFRs are hollow silica microcapillary devices of diameter typically around 50–300 um, with diameter modulation along their length. These regions of larger diameter simultaneously host ultrahigh-Q optical resonant modes and high-Q acoustic resonant modes in the solid shell (see Fig. 1). Fluids infused into the hollow core of OMFRs do not directly interact with light in the shell (see *optofluidics*), provided the shell is made sufficiently thick. As a result, high optical Q can be maintained. Vibration of the solid shell, however, entrains the fluid and results in hybrid solid–fluid acoustic resonances that are influenced by the properties (speed of sound, bulk modulus, damping) of the fluid. These hybrid modes can be optically sensed by modulation of the input laser

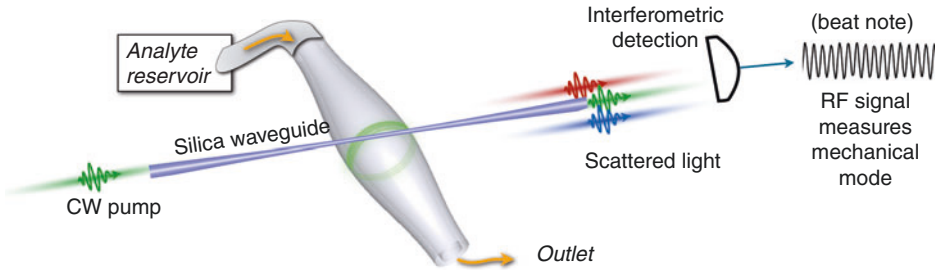




Microfluidic Optomechanics, Fig. 1 (*Left*) Electron microscope image of a microfluidic optomechanical resonator. These devices confine fluid inside while being exposed to air on the outer surface. (*Right*) The device can host both mechanical and optical resonances (Graphics adapted from [1])

Microfluidic Optomechanics, Fig. 2 (*Left*) Actuation of stationary resonant acoustic modes of the OMFR by means of optical radiation pressure. (*Right*) Actuation of traveling acoustic waves in the OMFR by means of stimulated Brillouin scattering (*SBS*)

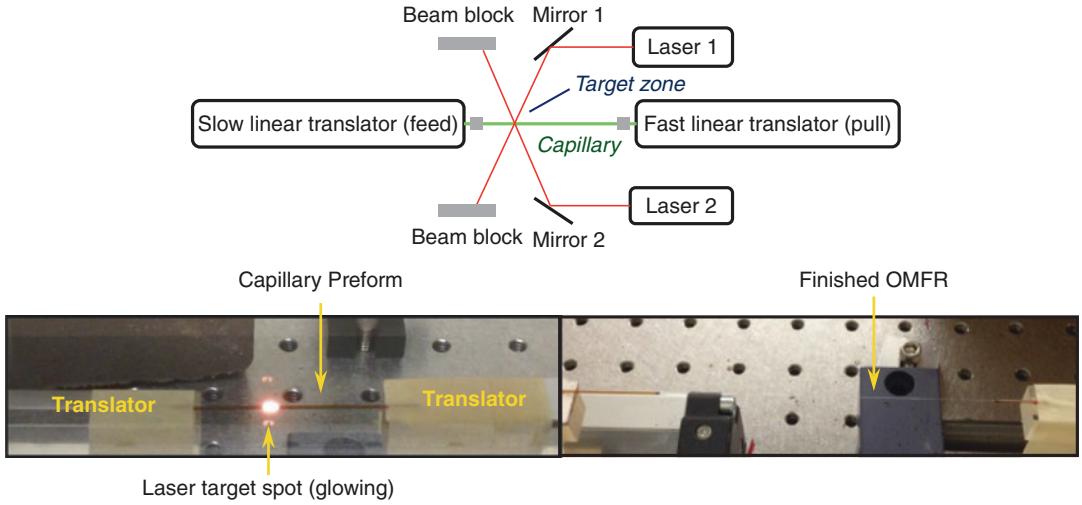




Microfluidic Optomechanics, Fig. 3 Arbitrary analytes can be flowed through an OMFR. A continuous-wave (*CW*) pump laser passing through an adjacent optical waveguide is modulated by the stochastic thermal–mechanical fluctuations and by the optically induced vibrations of the OMFR. The scattered light can be retrieved by the same waveguide, and the radio frequency (*RF*) spectrum can be monitored experimentally to measure the vibration

light by the vibration (see Figs. 2 and 3). This coupling permits all-optical measurement of fluid acoustic properties and even extends to flowing particles within the fluid. The optical sensing mechanism is sufficiently sensitive for the detection of stochastic thermal fluctuations of the vibrational modes.

In addition to sensing-only operation, light is also able to directly actuate mechanical vibrations within the device through optical radiation pressure [2] and through stimulated Brillouin scattering [3]. These modes of actuation are illustrated in Fig. 2. The optical radiation pressure actuation mechanism operates because photons carry linear



Microfluidic Optomechanics, Fig. 4 OMFRs are fabricated from silica capillary preforms by linear drawing under high-power laser heating (Graphics adapted from [5])

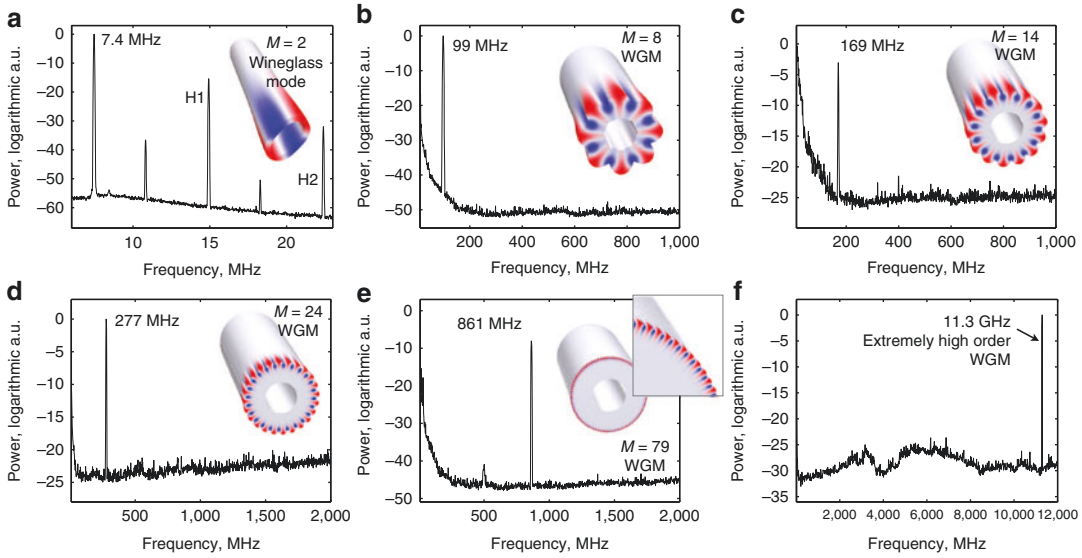
momentum that must be conserved, and when they are guided around a circular path, a centrifugal force must be generated. The simulated Brillouin scattering mechanism is a self-feedback lasing process [4] in which traveling sound waves are spontaneously generated and amplified by light.

Methods of Fabrication

OMFRs are fabricated from commercially purchased fused silica capillaries as described in [5] (see Fig. 4). The capillary is heated using CO_2 laser illumination (10.6 μm wavelength is absorbed by silica). When the glass heats up to the softening point, two linear pulling stages can draw the capillary into a thinned-down OMFR. Modulation of the CO_2 lasers permits rapid spatial variation of the local OMFR diameter, where the optical and acoustic modes can be simultaneously confined. The fabrication method is derived from existing techniques employed in optofluidics [6]. Alternative heating methods include hydrogen flame, arc discharge [7], laser glassblowing [8, 9], and even coil heaters, each having limitations on temperature ramp-up/ramp-down time constants with consequent limitation on the bubbled region geometry.

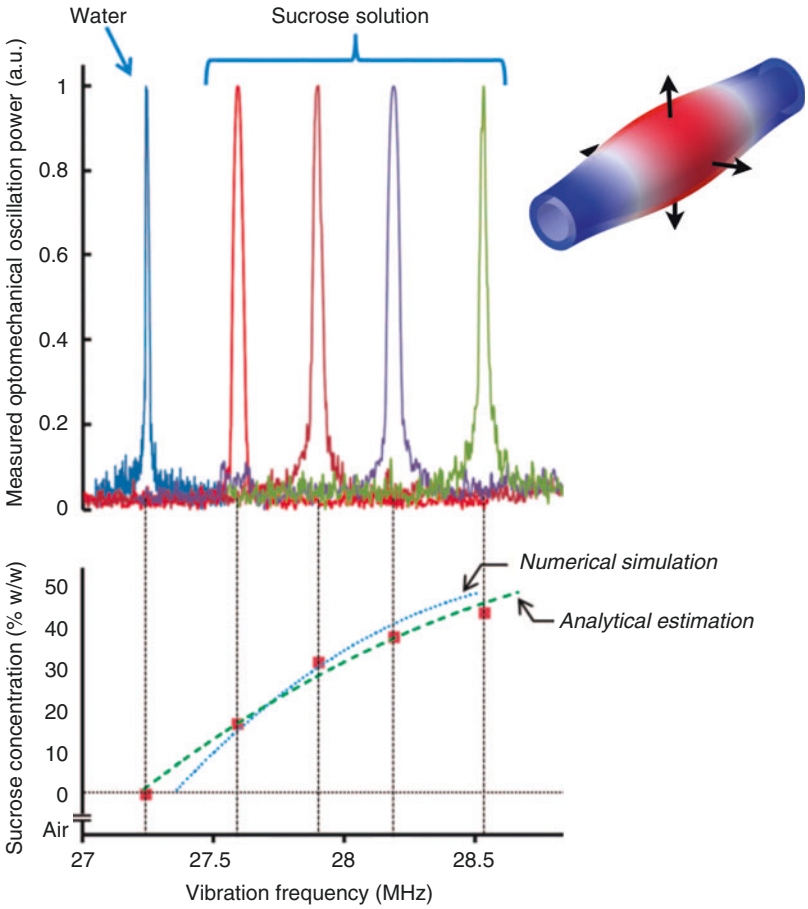
Main Research Accomplishment

The presence of fluids in the medium around an optomechanical device leads to both mechanical damping [10] and water-induced optical absorption, making optically induced oscillations challenging to actuate and observe. The development of *microfluidic* optomechanical devices has permitted, for the first time, testing of fluid-phase or fluid-suspended analytes and gases through high-frequency and high-fidelity optomechanics techniques. By placing the fluid within the device, OMFRs circumvent high mechanical damping in a similar manner to suspended microchannel resonators [11, 12]. Additionally, since the fluid is confined inside, the optical mode that remains at the exterior surface of the device remains largely unperturbed. As a result of the above characteristics, mechanical Q of several thousand and optical Q of 100 million can be readily achieved for high-sensitivity measurement of thermal fluctuations. The low optical and mechanical damping have permitted the actuation and sensing of vibrational modes in the 10's of MHz–10 GHz range (see Fig. 5).



Microfluidic Optomechanics, Fig. 5 Experimental results of the excitation of MHz–GHz vibrational modes of an OMFR with water infused within the device (Figure from [3])

Microfluidic Optomechanics, Fig. 6 Example of sucrose solution sensing through an OMFR. While the infused sucrose concentration is changed, the vibrational frequency of the OMFR breathing mode (visualized *top right*) changes monotonically (Graphics adapted from [2])



Examples of Application

OMFRs were first reported in 2013 in the work of Bahl and Kim [2, 3, 13], in which a sensitivity to fluid density properties was noted using sucrose solutions of varied sucrose weight fraction (Fig. 6). Han et al. later employed OMFRs to demonstrate pressure sensing [14] and fluid viscometry [15]. Kim and Fan experimentally have evaluated the surface sensitivity of OMFRs for potential biomolecule sensing applications [16].

Cross-References

- ▶ Cavity Optomechanics
- ▶ Microfluidic Device
- ▶ Nano-optical Biosensors
- ▶ Nano-optomechanical Systems (NOMS)
- ▶ Optical Cavity Biosensor
- ▶ Optical Resonance Biosensor
- ▶ Optical Ring Resonator Biosensor
- ▶ Optofluidic Resonators and Sensors
- ▶ Optofluidics
- ▶ Optomechanics

References

1. Bahl, G., Zehnpfennig, J., Tomes, M., Carmon, T.: Stimulated optomechanical excitation of surface acoustic waves in a microdevice. Nat. Commun. **2**, 403 (2011)
2. Kim, K.H., Bahl, G., Lee, W., Liu, J., Tomes, M., Fan, X., Carmon, T.: Cavity optomechanics on a microfluidic resonator with water and viscous liquids. Light Sci. Appl. **2**, e110 (2013)
3. Bahl, G., Kim, K.H., Lee, W., Liu, J., Fan, X., Carmon, T.: Brillouin cavity optomechanics with microfluidic devices. Nat. Commun. **4**, 2994 (2013)
4. Chiao, R.Y., Townes, C.H., Stoicheff, B.P.: Stimulated brillouin scattering and coherent generation of intense hypersonic waves. Physical Review Letters, **12**(21), 592–595 (1964)
5. Han, K., Kim, K.H., Kim, J., Lee, W., Liu, J., Fan, X., Carmon, T., Bahl, G.: Fabrication and testing of microfluidic optomechanical oscillators. J. Vis. Exp. **87**, e51497 (2014)
6. Lacey, S., White, I.M., Sun, Y., Shopova, S.I., Cupps, J.M., Zhang, P., Fan, X.: Versatile opto-fluidic ring resonator lasers with ultra-low threshold. Opt. Express **15**(23), 15523–15530 (2007)
7. Berneschi, S., Farnesi, D., Cosi, F., Conti, G.N., Pelli, S., Righini, G.C., Soria, S.: High Q silica microbubble resonators fabricated by arc discharge. Opt. Lett. **36**(17), 3521–3523 (2011)
8. Watkins, A., Ward, J., Wu, Y., Chormaic, S.N.: Single-input spherical microbubble resonator. Opt. Lett. **36**(11), 2113–2115 (2011)
9. Lee, W., Sun, Y., Li, H., Reddy, K., Sumetsky, M., Fan, X.: A quasi-droplet optofluidic ring resonator laser using a micro-bubble. Applied Physics Letters, vol. 99, p. 091102 (2011)
10. Arcizet, O., Riviere, R., Schliesser, A., Anetsberger, G., Kippenberg, T.J.: Cryogenic properties of optomechanical silica microcavities. Phys. Rev. A **80**(2), 021803 (2009)
11. Burg, T.P., Manalis, S.R.: Suspended microchannel resonators for biomolecular detection. Appl. Phys. Lett. **83**(13), 2698 (2003)
12. Burg, T.P., Godin, M., Knudsen, S.M., Shen, W., Carlson, G., Foster, J.S., Babcock, K., Manalis, S.R.: Weighing of biomolecules, single cells and single nanoparticles in fluid. Nature **446**(7139), 1066–1069 (2007)
13. Bahl, G., Kim, K.H., Lee, W., Liu, J., Tomes, M., Fan, X., Carmon, T.: Bridging two worlds: microfluidic optomechanics. Opt. Photonics News **24**(12), 39 (2013)
14. Han, K., Kim, J., Bahl, G.: Aerostatically tunable optomechanical oscillators. Opt. Express **22**(2), 1267–1276 (2014)
15. Han, K., Zhu, K., Bahl, G.: Opto-mechano-fluidic viscometer. Appl. Phys. Lett. **105**(1), 014103 (2014)
16. Kim, K.H., Fan, X.: Surface sensitive microfluidic optomechanical ring resonator sensors. Appl. Phys. Lett. **105**(19), 191101 (2014)

Microfluidic Pumps

- ▶ Micropumps

Microfluidic Whole-Cell Biosensor

Soichiro Tsuda
School of Chemistry, University of Glasgow, Glasgow, UK

Synonyms

Lab-on-a-chip whole-cell biosensor

Definition

Microfluidic whole-cell biosensors incorporate living cells, such as bacteria, inside lab-on-a-chip sensing devices for a conversion of initial inputs (e.g., environmental samples) into cellular response as primary transducer.

Overview

Whole-cell biosensors are a type of biosensor that incorporates living cells as a biorecognition element [1].

Conventionally, chemical or physical sensors, such as high-performance liquid chromatography (HPLC) or gas chromatography coupled with mass spectrometry (GC-MS), have been used for analyzing samples of interest (e.g., the composition and concentration of chemicals). Although they are highly accurate and results are reliable, such instrumentation is expensive and often requires special skills and facilities for analysis. An alternative approach in sensor technologies is sensors using biological components, such as enzymes and living cells. They are simpler and often single-purpose devices, but can be easily produced (hence inexpensive).

There are two different objectives in biosensor research: One objective is to develop biological counterparts of conventional chemical or physical sensors. Biological molecules (enzymes, antibodies, and nucleic acid fragments) are commonly used to develop such biosensors due to their substrate specificities [2]. Another objective is the toxicity assessment. While biomolecule-based biosensors are more accurate and give more detailed information about analytes, it is difficult to know the effect of the analyte on living organisms. Thus, a main application domain of biosensor using living cells or tissues is to obtain information about functional consequences of substrates being analyzed or the measurement of the total "bioavailability." The bioavailable (or bioaccessible) compound is defined as "(a molecule that) is freely available to cross an organism's cellular membrane from the medium the organism inhabits at a given time" [3], which would be useful for the understanding of, for example, the environmental effect of the compound. Potential applications of whole-cell biosensors could be the assessment of physiological effect of a pharmaceutical compound, checking if a substance is an agonist or an antagonist for a receptor, the cellular toxicity test, and the environmental pollution test [4]. As these types of information involve vital reactions inside biological cells, whole-cell biosensors are considered to be the most suitable sensor in order to investigate these questions.

Living cells employed for whole-cell biosensor development are quite diverse: bacteria, oocytes, cultured fish cells, fibroblasts, neurons, cardiac myocytes, etc. Among others, bacterial cells are commonly used due to the ease of use. For extensive lists of bacteria used for biosensor and target compounds, see [5].

In microbiologist's term, it is often assumed that "biosensor" refers to biological cells themselves. However, biosensor cells need to be actually integrated into hardware devices that allow readout of the output response of cells, when practical applications are in focus. Microfluidic or lab-on-a-chip technologies would contribute in this regard and will help further integration and miniaturization of biohybrid sensor devices.

This entry first gives an overview of whole-cell biosensors and then describes possible interfaces between cells and readout devices. Lastly, examples of actual whole-cell biosensor implementations using microfluidic techniques are given.

Types of Whole-Cell Biosensor

Whole-cell biosensors can be categorized into three groups based on the specificity to external stimuli to be analyzed [4]: specific, semi-specific, and nonspecific biosensors.

The specific whole-cell biosensor reacts only to a certain compound, and thus it is a biosensor counterpart of standard analytical chemical techniques. So far various types of specific whole-cell biosensors have been developed. They mostly rely on genetic engineering of reporter genes

with inducing promoters. Hansen and Sørensen categorized specific whole-cell biosensors into two subgroups [5]: activator-type and repressor-type. The important elements of those biosensors are DNA sequences of a cell: DNA sequences of activator-type biosensor cells contain an activator gene, promoter, and reporter genes. The activator gene constantly produces activator proteins, but they can bind to the promoter region only when a target compound to be detected is present. Binding of the activator protein to the promoter allows the reporter genes to be read and produce reporter signals (e.g., fluorescent proteins), which are picked up by detection devices (fluorescent microscopes). On the other hand, cells of the repressor-type biosensor have a repressor gene instead of activator gene. The repressor gene constantly produces repressor proteins that bind to the promoter region which prevent reporter genes to be transcribed unless a target compound is present. The target compound binds to the repressor proteins and releases them from the promoter region. This triggers the transcription of reporter genes and production of reporter signals. However, due to the genetic engineering of biosensor cells, these types of specific whole-cell biosensor can be constructed only after the gene regulation mechanisms related to target compounds are discovered.

Another type of whole-cell biosensor is the semi-specific or group-specific biosensor. This biosensor also employs gene regulation mechanisms for the detection of substance classes. A typical genetic circuit of the biosensor consists of an activator gene, promoter, and reporter gene. While only one target compound could trigger the reporter protein production in specific whole-cell biosensor, different stresses can induce the activator gene in semi-specific biosensor cells. The activator gene subsequently induces transcription of the reporter gene and production of reporter signals. Examples of stresses demonstrated to detect are antibiotics, gamma irradiation, genotoxins, oxidative stress, and UV radiation (cf. [3, 5]).

Nonspecific or general whole-cell biosensors are developed to obtain general information about the presence of toxic compounds. Cells of nonspecific biosensors constantly produce detection signals, such as bioluminescent proteins. Any condition that affects the activity of nonspecific biosensor cells, such as pH, salinity, metals, and xenobiotics, will cause degradations of the released signal, which are used as a measure to estimate toxicity. Although it is not possible to detect a specific compound with this biosensor, it gives information about, for example, the effect of a compound or environmental sample to biological cells by employing complex cascades of intracellular reactions. Such biosensor can be used to prescreen environmental samples, which will be further analyzed using specific chemical or physical sensors to track down the chemical identity of toxic substances. Nonspecific biosensors have been more commonly used for commercial environmental toxicity test, for example, the Microtox test.

Detection Methods

A primary role of biological cells employed for detection is a conversion of initial inputs (e.g., environmental samples) into cellular response as primary transducer. The next required step is to develop secondary transducer that converts the cellular response into detectable signals, as secondary transducer.

A most commonly used interface is optical detection using bacterial luciferase (Lux), firefly luciferase (Luc), aequorin, and green fluorescent protein (GFP) [6]. Among others, bacterial luciferase and GFP are commonly adopted for whole-cell biosensor use. A bacterial bioluminescence reaction involves the bacterial luciferase enzyme, oxygen, reduced riboflavin phosphate ($FMNH_2$), and a long chain fatty aldehyde. When the luciferase enzyme catalyzes the oxidation of the aliphatic aldehyde, a high-energy intermediate of $FMNH_2$, oxygen, and the aldehyde emits light primarily at a wavelength of 490 nm. A chief advantage of this bioluminescent system as bioreporter is an autonomous nature of *lux* operon. When genes required for aldehyde biosynthesis are integrated in biosensor cells, they can function to produce light completely autonomously without any additional inputs. This nature

makes the Lux bioluminescent system suitable for nonspecific whole-cell biosensor in which reporter signals (in this case, bioluminescence) are constantly produced. When the cells are exposed to toxic compounds, the bioluminescent signal will decrease depending on the viability of cells. Another widely used bioreporter, GFP, is an autofluorescent protein, first discovered in 1962 together with aequorin [7]. It has two absorption peaks at 397 nm (major) and 475 nm (minor) and emits light at 504 nm. While bacterial luciferase (*lux*) genes can be expressed only in prokaryotes without any gene sequence alterations, GFP can function in both prokaryotic and eukaryotic cells [8]. It also has been shown that GFP can be used for all three types of biosensor described above (cf. [9–11]).

So far several types of devices are developed for detection of these optical signals. Fiber-optic cable is one of the most common interfaces. Bioluminescent light can be detected by attaching bioreporter cells and photomultiplier tube (PMT) on each end of the cable. In the case of GFP detection, however, an additional cable is required for interfacing an excitation light source (e.g., tungsten lamp) and cells because GFP is a fluorescent protein. Fiber-optic cables can be used for various types of samples (liquid, solid, gaseous), but is not a suitable interface for miniaturization of the device because of size and power constraints. Toward the on-chip integration of whole-cell biosensor device, complementary metal-oxide-semiconductor (CMOS) [12], MOEMS (Micro-Opto-Electro-Mechanical System) [13], and avalanche photodiode (APD) [14] devices have been developed for the detection of bioluminescent signal.

Apart from the optical interface, electrochemical and electrical interfaces are developed. Electrochemical whole-cell biosensors monitor cellular redox events due to metabolic activity of cells by amperometric, potentiometric, and conductometric techniques. Analytes detected at working electrodes are normally natural chemical species, such as oxygen, or an introduced chemical which can exist in both the oxidized and reduced forms. In the case of amperometry, measured current is linearly dependent on the concentration of chemical species. Electrical interfaces are mainly for neuronal biosensors. Single neurons are grown on a microelectrode array, and their electrical activities are monitored by the microelectrodes [15]. A main advantage of the electrical interface is real-time detection. Neurons react to external stimuli within a second, whereas other (typically microbial) whole-cell biosensors take over minutes.

Lab-on-a-Chip Device Integration

Although biological cells are employed as sensors for detection of chemicals, cell-based biosensor research often experiences problems with reproducibility and stability of the biosensors. This is mainly because of the lack of control over the (micro)environment of cells used for detection. Microfluidic and lab-on-a-chip technologies are expected to fill the gap, and thus microfluidic-based biosensors would have a large potential to improve these problems. This section introduces examples of microfluidic biosensors using biological cells to illustrate various types of ongoing research in this field.

One of the first developments of microfluidic whole-cell biosensors is by Derek DeBusschere and Kovacs [16]. They have developed a cell cartridge consisting of polydimethylsiloxane (PDMS) fluidic chambers and integrated CMOS circuits. The cell cartridge provides temperature control system, microelectrode sensors, analog signal buffering, and digital interface for a portable handheld measurement device. Cardiomyocytes are cultured in the cartridge, and the electrophysiological activity (cell contraction) is monitored by the integrated electrodes attached to the cells. As their implementation represents, a typical usage of microfluidic devices is for compartmentalization and immobilization of sensor bacteria in microchambers. Rabner and coworkers developed a PDMS-based integrated whole-cell biosensor for water toxicity analysis [17]. This is basically a kind of semi-specific biosensor using genetically modified *E. coli* cells harboring bacterial luciferase (*lux*) genes. When cells are exposed to genotoxic chemicals in

cylindrical reaction chambers of the microfluidic device, they start to produce bioluminescent proteins as described above. Light emissions from the cells are picked up by PMT integrated in the sensor. A motorized XY stage controls the position of PMT to go through the reaction chambers one by one. Tani and coworkers developed a bacterial biosensor chip for genotoxic bioassay [18]. The chip comprised of two layers of PDMS microfluidic channels sandwiching a silicon substrate layer with 25 well arrays. A mixture of *E. coli* as sensor strains and agarose is introduced from one side of the microfluidic layer and immobilized in the 25 wells. Samples of interest with different concentrations are applied from separate lines of fluidic channels in another PDMS layer to stimulate the cells. Reactions of cells are detected as bioluminescent light emission. This whole-cell biosensor has been demonstrated to assay the genotoxicity of mitomycin C and can be used for high-throughput multiple assays of drugs or physiological substances.

Ben-Yoav and coworkers developed an on-chip electrochemical biosensor using *E. coli* and *Salmonella typhimurium* [19]. Three gold electrodes (working, counter, and reference electrodes) are patterned on a single crystal silicon substrate, and a PDMS microfluidic chamber for cell culture was mounted on the chip. This device detects a genotoxin, nalidixic acid (NA), by means of catalytic reactions in the culture solution. For example, in the presence of NA, the genetically engineered bacteria produce the reporter enzyme β-galactosidase which catalyzes the conversion of para-aminophenylphosphate (pAPP) into the electroactive species pAP. This reaction is picked up as increasing electrochemical current by the electrodes.

Integration of nerve cell culture and electrical measurement device is also one of the active areas in this field. Prasad and coworkers patterned single neurons on individual microelectrodes integrated in a microfluidic device [15], and the electrophysiological activities are monitored by the electrodes. Although a broad spectrum of chemical agents does stimulate the patterned neurons, reaction patterns of neurons to each chemical (called "signature patterns") can be identified using fast Fourier transformation and wavelet transformation analyses. Takeunchi and coworkers developed a microfluidic odorant sensor using oocytes of African clawed frog *Xenopus laevis* [20]. Oocytes are modified to express insect membrane proteins (pheromone and odorant receptors). Cells are trapped in a microfluidic channel, and the receptor activities are monitored by microelectrodes integrated in the device. If a ligand (e.g., bombykal) attaches to a receptor on the oocyte membrane, the cation channel opens up, and influx of cations changes the current across electrodes inserted in the cell. It has been shown that this biosensor qualitatively detects several chemicals with high reproducibility and high specificity with the working range of 10 nM to 1 μM.

Daunert and coworkers developed a spore-based lab-on-a-compact disk (CD) biosensor. Microfluidic channels are fabricated by micromachining on a 12 cm poly(methyl methacrylate) (PMMA). The fluidic part consists of reagent reservoirs and detection chamber connected by mixing channels. Samples and cells (in this case, spores of *B. subtilis* strain) are loaded in the reagent reservoirs and microcentrifugally move toward the detection chamber through mixing channels. When exposed to a target chemical (e.g., arsenite), genetically modified spores start to express GFP proteins which are detected by a fluorescence spectrophotometer equipped with a fiber-optic system at the detection chamber. Although detection of chemicals takes time (over 150 min) due to the germination of spores in the CD, this spore-based biosensor showed sensitive detection (1×10^{-7} M for arsenite) and excellent storage life (more than 2 years at room temperature).

Cross-References

- ▶ Nanogap Biosensors
- ▶ Nanophotonic Structures for Biosensing
- ▶ Nanostructure Field Effect Transistor Biosensors
- ▶ Nanostructured Materials for Sensing
- ▶ NEMS Resonant Chemical Sensors

- ▶ Organic Sensors
- ▶ Whispering Gallery Mode Resonator Biosensors

References

1. Wang, P., Liu, Q. (eds.): Cell-Based Biosensors: Principles and Applications. Artech House, Norwood (2009)
2. Cooper, J., Cass, T. (eds.): Biosensors. The Practical Approach Series, 2nd edn. Oxford University Press, Oxford (2004)
3. Harms, H., Wells, M.C., van der Meer, J.R.: Whole-cell living biosensors-are they ready for environmental application? Appl. Microbiol. Biotechnol. **70**(3), 273–280 (2006)
4. Bousse, L.: Whole cell biosensors. Sens. Actuators. B. Chem. **34**(1–3), 270–275 (1996)
5. Hansen, L., Sørensen, S.: The use of whole-cell biosensors to detect and quantify compounds or conditions affecting biological systems. Microb. Ecol. **42**(4), 483–494 (2001)
6. Close, D.M., Ripp, S., Sayler, G.S.: Reporter proteins in whole-cell optical bioreporter detection systems, biosensor integrations, and biosensing applications. Sensors **9**(11), 9147–9174 (2009)
7. Shimomura, O., Johnson, F.H., Saiga, Y.: Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan Aequorea. J. Cell. Comp. Physiol. **59**, 223–239 (1962)
8. Chalfie, M., Tu, Y., Euskirchen, G., Ward, W.W., Prasher, D.C.: Green fluorescent protein as a marker for gene expression. Science **263**(5148), 802–805 (1994)
9. Choi, O., Deng, K.K., Kim, N.-J., Ross, L., Surampalli, R.Y., Hua, Z.: The inhibitory effects of silver nanoparticles, silver ions, and silver chloride colloids on microbial growth. Water Res. **42**, 3066–3074 (2008)
10. Justus, T., Thomas, S.M.: Evaluation of transcriptional fusions with green fluorescent protein versus luciferase as reporters in bacterial mutagenicity tests. Mutagenesis **14**(4), 351–356 (1999)
11. Gvakharia, B., Bottomley, P., Arp, D., Sayavedra-Soto, L.: Construction of recombinant *Nitrosomonas europaea* expressing green fluorescent protein in response to co-oxidation of chloroform. Appl. Microbiol. Biotechnol. **82**, 1179–1185 (2009). doi:10.1007/s00253-009-1914-y
12. Ripp, S., Daumer, K., McKnight, T., Levine, L., Garland, J., Simpson, M., Sayler, G.: Bioluminescent bioreporter integrated-circuit sensing of microbial volatile organic compounds. J. Ind. Microbiol. Biotechnol. **30**, 636–642 (2003). doi:10.1007/s10295-003-0093-6
13. Elman, N.M., Ben-Yoav, H., Sternheim, M., Rosen, R., Krylov, S., Shacham-Diamand, Y.: Towards toxicity detection using a lab-on-chip based on the integration of moems and whole-cell sensors. Biosens. Bioelectron. **23**(11), 1631–1636 (2008)
14. Yotter, R.A., Wilson, D.M.: A review of photodetectors for sensing light-emitting reporters in biological systems. IEEE. Sens. J. **3**, 288–303 (2003)
15. Prasad, S., Zhang, X., Yang, M., Ozkan, C.S., Ozkan, M.: Neurons as sensors: individual and cascaded chemical sensing. Biosens. Bioelectron. **19**(12), 1599–1610 (2004)
16. DeBusschere, B.D., Kovacs, G.T.: Portable cell-based biosensor system using integrated CMOS cell-cartridges. Biosens. Bioelectron. **16**(7–8), 543–556 (2001)
17. Rabner, A., Belkin, S., Rozen, R., Shacham, Y.: Whole-cell luminescence biosensor-based lab-on-chip integrated system for water toxicity analysis. Microfluidics BioMEMS Med. Microsyst. IV **6112**(1), 33–42 (2006)
18. Tani, H., Maehana, K., Kamidate, T.: Chip-based bioassay using bacterial sensor strains immobilized in three-dimensional microfluidic network. Anal. Chem. **76**(22), 6693–6697 (2004)
19. Ben-Yoav, H., Biran, A., Pedahzur, R., Belkin, S., Buchinger, S., Reifferscheid, G., Shacham-Diamand, Y.: A whole cell electrochemical biosensor for water genotoxicity bio-detection. Electrochim. Acta **54**(25), 6113–6118 (2009)
20. Misawa, N., Mitsuno, H., Kanzaki, R., Takeuchi, S.: Highly sensitive and selective odorant sensor using living cells expressing insect olfactory receptors. Proc. Natl. Acad. Sci. U. S. A. **107**(35), 15340–15344 (2010)

M

Microfluidics

- ▶ Microchannel Flows as Heat Sinks

Microinjection Molding

Fred Chiou
Department of Engineering Technology, Weber State University, Ogden, UT, USA

Definition

Microinjection molding is defined as a particular injection molding process that requires a specialized molding machine that is capable of delivering

high injection speed, high injection pressure, precise shot control, uniform melt temperature, and ultrafine resolution using servo-electric drives and sophisticated controls geared specially to mold micro-miniature components [1]. In other words, as cited by Attia et al. [2], microinjection molding is the process of transferring a thermoplastic material in the form of granules from a hopper into a heated barrel so that it becomes molten and soft. The material is then forced under pressure inside a mold cavity where it is subjected to holding pressure for a specific time to compensate for material shrinkage. The material solidifies as the mold temperature is below the glass transition temperature of the polymer. After sufficient time, the material freezes into the mold shape and gets ejected, and the cycle is repeated. A typical cycle lasts between few seconds and few minutes.

Introduction

Microinjection modeling is a very specialized precision injection molding technique. The technique can be used in microfluidic devices for various applications such as chemistry, medicine, and biotechnology. Some articles, capillary electrophoresis (CE) platforms [3–7], miniaturized heat exchangers [8], and nanofilters [3] reported that microinjection molding was also used for microfluidic applications. It is also used to mold micro-miniature components for the applications of micro-optical (gratings, waveguides, and lenses – [5, 6, 8–11]), micro connectors (ceramic ferrule holder, capacitor housing, etc.), automotive (axle for cockpit instruments, etc.), Microdrive system and controls (potentiometer gear, acetal stepper motor axle for watches, micro gears, etc.), medical and surgical (blade holder, dental prosthetic, etc.), micromechanics (micro-springs, gears and miniaturized switches, catch wheels, etc. – [6, 8–10]), hearing aid (sensor holder, etc.), catheter components, clot removal catheter, communications (for devices that require connectors, gaskets, rollers, lenses, switches, etc.), watch industry, computer, electronics, sensors, and actuators [5, 6]. Surace et al. [12] summarized some examples of microinjection molding in various application fields. These examples can be categorized as shown below.

Micromechanical Parts

- Locking lever for micromechanical industry or micro switch
- Latch for the watch industry
- Catch wheel for micro switch
- Operating pin
- Gear plate for motive power engineering

Micro Gear Wheel

- Dented wheel for watch industry
- Rotor with gear wheel for watch industry
- Gear wheel for micro gear
- Spur wheel in the field of electrical technology
- Spiral gear in the field of electrical technology/metrology
- Spline in the field of electrical technology/metrology

Medical Industry

- Microfilter for acoustics, hearing aid
- Implantable clip
- Bearing shell/bearing cap
- Sensor housing implantable
- Aseptic expendable precision blade

Optical and Electronic Industries

- Coax plug/switch MID for mobile phone
- SIM card connector for mobile phone
- Pin connector for mobile phone
- Single mode and multimode ferrules

The definition of a micro-molded part can be classified by some factors such as weight and dimension. Generally speaking, if a part with the weight from 0.1 to 0.001 g and the part size is in the range of 0.1 in. in diameter, it can be categorized as a micro-molded part [2]. Kukla et al. [13], Angelov and Coulter [10], Whiteside et al. [14], Martyn et al. [15], Bibber [16], and Tom et al. [17] presented a set of criteria to classify the micro-molding parts [1]. They mentioned that a

micro-molding part has one or more of the following specifications: (1) It weighs less than 1 mg, or it is a fraction of a polymer pellet, where a pellet can be approximated to be spherical in shape with an average diameter of approximately 3 mm. (2) It is a part with microstructured regions, or more specifically, with wall thicknesses less than 100 μm. Microfluidic devices usually fall in this category. (3) It is a micro-precision part, which is a part that can have any dimensions, but has tolerances in the micrometer range, or more specifically, between 2 and 5 μm.

Michaeli et al. [18] investigated the use of microinjection molding for the assembly of hybrid structures. This process is called micro-assembly injection molding (μAIM). It is derived from the rather common processes two-component injection molding and over-molding. It is the combination of the positioned joining of inlay parts with the generation of functional microstructures. Some applications of using micro-assembly injection molding (μAIM) are movable structures, hard/soft combinations, hollow structures, optical systems, and electrically functional structures [18].

The advantages of using microinjection molding include the following features:

1. Cost-effectiveness for mass production process, especially for disposable products [3, 5, 19–21].
2. Available wide range of thermoplastics and the potential for full automation with short cycle times [3, 19, 22].
3. Good dimension control and accurate shape replication [19, 20, 22].
4. Relative low-maintenance costs of capital equipment compared to lithographic methods, for example [22].
5. Large amount of know-how knowledge and industrial information are available from conventional injection molding.

Aspect Ratio (AR)

The aspect ratio is defined as the ratio of the longer and the shorter dimensions of a micro-miniature component. For the three-dimensional components, the aspect ratio will be the longest and the shortest axis [12]. When replicating microfeatures using microinjection molding, the aspect ratio becomes an important issue, and the constraint – what is the maximum achievable aspect ratio or how small the micro-miniature component – can be replicated. Heckele et al. [23] discovered that achievable aspect ratio is a function of geometry of microfeatures, their position on the sample, the polymer type, and the process parameters. They suggested that the aspect ratio is the main factor of determining the critical minimum dimensions that can be successfully replicated using injection molding.

Hesitation Effect

The hesitation effect [12] is a phenomenon that can occur during the filling of polymers, and it is common when an injection-molded part contains different thicknesses [19]. It may also take place when microstructures with high aspect ratio (HAR), usually larger than 2, are placed on a relatively thick substrate such as the microfluidic devices [2]. The polymer melt tends to flow easier into cavities with relatively low resistance in the areas of greater cross section, while the flow stagnates at the entrance of microstructures. The causes that the melt frozen in this area is because the filling time of the substrate is usually greater than the freezing time of the microfeature. Rötting et al. [24] recommended in their literature that injection-molded parts with HAR microstructures should have a thickness in which a quick filling of the substrate can allow for filling of the micro-cavities before solidification starts.

Tools for Microinjection Molding

As discussed above, microinjection molding has been widely used for manufacturing micro-miniature components for consumer products. For example, low-cost mass production of polymer microfluidic components can be achieved using either injection molding. The technique

requires a tool to transfer the microstructures to the polymer material. The manufacturing of the tool becomes essential. There are many ways to manufacture the tools or tool inserts. The combination of photolithography, etching, and electrodeposition is commonly used for this process. The basic fabrication schemes can be summarized and listed as shown in schemes 1–4B below from Bissacco et al. [25]. Silicon, polymers, or metals can be used as the substrates. The micromachining can be subtractive (reactive-ion etching, mechanical micromachining, or laser micromachining) or additive (such as UV lithography combined with electroforming).

1. Fabrication scheme 1 – tool insert based on a metal substrate, lithography, and electroforming

Step	Process
1	Machining and cleaning of metal wafer
2	Spin coating, exposure, and development of photoresist
3	Electroforming of Ni
4	Removal of photoresist

2. Fabrication scheme 2A – silicon substrates, lithography, and electroforming

Step	Process
1	Sputtering of 50/500 Å Ti/Au on SiO_2
2	Spin coating, exposure, and development of photoresist
3	Electroforming of Ni
4	Sputtering of 50/500 Å Cr/Cu
5	Electroforming of Ni
6	Selective etching of Si, Au, and photoresist

3. Fabrication scheme 2B – silicon substrates, lithography, and electroforming

Step	Process
1	Sputtering of 50/500 Å Ti/Au on SiO_2
2	Spin coating, exposure, and development of photoresist
3	Electroforming of nickel
4	Machining of tool back
5	Selective etching of Si, Au, and photoresist

4. Fabrication 3A – utilize reactive-ion etching of silicon directly

Step	Process
1	Deposition of mask on Si
2	Spin coating, exposure, and development of photoresist
3	Selective etching of mask and removal of photoresist
4	Reactive-ion etching of Si
5	Mask removal and sputtering of 50/500 Å Ti/Au
6	Electroforming of Ni
7	Machining of tool back
8	Selective etching of Si

5. Fabrication scheme 3B – utilize reactive-ion etching of silicon indirectly

Step	Process
1	Deposition of mask on Si
2	Spin coating, exposure, and development of photoresist
3	Selective etching of mask and removal of photoresist
4	Reactive-ion etching of Si
5	Mask removal and gluing of glass support

6. Fabrication scheme 4A – based on subtractive micromachining of non-silicon substrates (positive)

Step	Process
1	Machining and preparation of substrate
2	Micromachining of substrate (positive)
3	Electroforming of Ni
4	Machining of tool back
5	Selective etching of substrate

7. Fabrication scheme 4B – based on subtractive micromachining of non-silicon substrates (negative)

Step	Process
1	Machining and preparation of substrate
2	Micromachining of substrate (negative)
3	Cleaning
4	Micromachining of substrate (negative)
5	Optional plating of Ni or Cr

Microinjection Molding, Table 1 Commonly used polymers for microinjection molding, maximum aspect ratios (AR), minimum structural thickness, and applications [1]

Polymer	Abbrev.	Maximum aspect ratios (AR)	Minimum structural thickness (μ lm)	Example of application	References
Polymethyl methacrylate	PMMA	20	20	Optical fiber connector	Piotter et al. [26]
Polycarbonate	PC	7	350	Cell container	Piotter et al. [26]
		4–8	0.2–0.5	Optical element	Wimberger-Friedl [20]
Polyamide	PA	10	50	Micro gear wheels	Piotter et al. [26]
Polyoxymethylene	POM	5	50	Filter with defined pore diameters	Piotter et al. [26]
		10	80	Microrods	Ong et al. [27]
Polysulfone	PSU	5	270	Housings for microfluidic devices	Piotter et al. [26]
Polyetheretherketone	PEEK	5	270	Housing for micropumps	Piotter et al. [26]
Liquid crystal polymers	LCP	5	270	Microelectronic devices	Piotter et al. [26]
Polyethylene (high density)	HDPE	8	125	High aspect ratio squares	Despa et al. [28]
		10	225	Microwells	Yao and Kim [19]
Cyclic olefin copolymer	COC	0.02–2	0.1–0.9	Microfluidic patterns	Mönkkönen et al. [29]
Conductively filled polyamide	PA 12-C	10	50	Housing for electrostatic microvalves	Piotter et al. [26]

Bissacco et al. [25] also presented a classification of the available process sequence alternatives and proposed a solution based on generation of the pattern by mechanical or thermal material removal operation. Their viable alternative is constituted by the introduction of mechanical and thermal processes based on material removal in the process chain for insert fabrication. They reported that the solution allows manufacturing of complex three-dimensional shapes with high aspect ratios.

Polymeric Microstructure Replication by Microinjection Molding

Among the materials that have been used for microstructured parts (such as glass, polymer, silicon, acetal, polyester, and polycarbonate), polymers show superior properties that become promising materials for microinjection molding [1]. Microstructures can be easily replicated in medium and large production by using polymer in microinjection molding. Commonly used polymers for microinjection molding, maximum aspect ratios (AR), minimum structural thickness, and applications are shown in Table 1. Several authors [3, 22, 23, 30–36] discovered that polymers have a wide variety of characteristic material properties, such as mechanical strength, optical transparency, chemical stability, and biocompatibility.

Because the costs are not greatly affected by the complexity of the design and the design complexity mostly impacts on mold design rather than

on the molding process itself [3, 23, 31, 37, 38], the cost of polymers is relatively low, especially for high-production disposable devices [30, 39].

Another advantage of polymers is their excellent replication fidelity. A typical polymer can completely fill and accurately replicate small features down to tens of nanometers if the optimum processing conditions are applied [1]. The polymer, due to its viscoelasticity, would fill the smallest mold details with a low viscosity if the shear rate is increased and polymer temperature above its T_g is maintained during the filling stage.

De Mello [22] reported that the quality of the molded surface features is almost completely dependent on the quality and precision of the mold. Some reports from the experiments have demonstrated the successful replication using injection molding, for example, submicron test features [29, 40], submicron grating optical elements [20, 41, 42], and surfaces with nanostructured patterns [43, 44] or nanoscale topography [45].

Polymers show good electrical insulation compared to silicon [7, 22]. However, polymers have some limitations such as temperature range, higher autofluorescence, and limited well-established surface modification techniques compared to glass due to their properties [1].

Material Selection for Polymeric Microfluidic Devices

Material requirements for polymeric microfluidic devices have been investigated by several authors in the literature [3, 6, 7, 22, 35]. It is suggested that the material should possess appropriate mechanical properties, thermal properties (especially when the microfluidic device is subject to elevated temperatures while in service), appropriate glass transition temperature Tg, melt temperature Tm and thermal expansion coefficient (in order to avoid softening of the device and subsequent deformation in features), good heat dissipation ability (in order to avoid local high temperatures in the device when it is subject to high temperatures while in service), chemical resistance (to certain solvents or strong acids), properties of dimensional stability and accuracy (including reproducibility, tendency to shrinkage, and warpage), good electrical insulating properties, biocompatibility and permeability (for biological applications), high mechanical strength, low viscosity, and good flow properties. Generally speaking, all standard thermoplastic materials suitable for conventional injection molding can be used for microinjection molding [9], provided that some manufacturability conditions are met [11, 32, 46, 47].

Part Shrinkage and Shape Distortions

Part shrinkage and shape distortions are common problems during the molding process. Part shrinkage causes volume changes while shape distortions result in warpage. Shrinkage happens with the decrease in part volume due to the temperature change between the demolding temperature and the ambient temperature. Geometrical changes and controlling the processing parameters can improve demolding and part shrinkage [23].

Heckele and Schomburg [23] investigated the orientation of the polymer being injected that affects the direction at which shrinkage is mostly observed. They suggested that the path of polymer injection inside the mold should be considered. In addition to demolding and feature replication, shrinkage affects shape stability in the form of induced warpage. Warpage is produced due to the nonuniformity of the shrinkage because of the residual stresses induced by the complex thermal variation inside the mold [48].

In injection molding, the position of the cooling channel is dependent on part geometry, cavity configuration, and location of ejection mechanism. So it is not always possible to design an ideal cooling system. Greener and Wimberger-Friedl [48] discovered that the variation in temperature across the molded part should be expected depending on the geometry. Cooling the mold is not always required, especially when it is desired to keep the mold temperature above the Tg of the polymer to ensure complete filling of the mold cavity.

Giboz et al. [49] suggested the method of increasing the amount of holding pressure to decrease the effect of shrinkage. This method, on the other hand, will also increase stresses inside the part. Another technique is to allow the part to thermally equilibrate inside the mold cavity, i.e., have a long cooling time. The temperature is to be approximately uniform in this condition. The part is likely to distort because of lack of homogeneity in shrinkage in precision molding, which is likely to occur if the cycle is terminated before the thermal equilibrium is reached, i.e., the part is ejected before its temperature is reduced to the ambient temperature [48]. The cost of this technique is a longer cycle time.

Besides the consideration of decreasing shrinkage, Heckele and Schomburg [23] discovered that the molded parts that need to be assembled in subsequent stages should be placed on the same mold because the degree of shrinkage will vary over time.

Micrometal Injection Molding

Microinjection molding can also be used to produce micrometal parts [50]. The process is called micrometal injection molding or μ-MIM. Authors have studied miniaturization of metal parts by using metal injection molding (MIM) and developed the advanced process of μ-MIM to fabricate ultrafine structures and realized mass production of μ-MIM products [51–53]. In metal injection molding (MIM) process, metal powder is mixed with organic binder and kneaded and pelletized [54]. The injection machine works in the same manner as with plastics: the compound pellet is plasticized and injection-molded into high-precision near-net-shaped products [55]. The process of MIM shows the advantage of high productivity of complicated three-dimensional parts.

The molding conditions such as injection speed, mold temperature, and holding pressure have the impacts on dimensional accuracy of micro-sized parts. The holding pressure can be accurately measured. The measuring device is normally equipped to the injection machine for operational control. Therefore, with the stable material, mixing, and sintering conditions, the dimensional accuracy of μ-MIM products may be improved by monitoring the cavity pressure and controlling the injected volume during injection molding.

Micro-Miniature Component Quality

Studies with different experimental conditions and test parts have been conducted to investigate the process parameters that affect the quality of

Microinjection Molding, Table 2 Some microinjection molding machines

Manufacturer	Machine model
Single-step system	
Standard technology – downscaled technology	
Standard injection molding machines for small pieces	
Arburg GmbH & Co. KG	Allrounder 170 U
Battenfeld Kunststoffmaschinen GmbH	HM 250 Micromelt
Dr. Boy GmbH & Co. KG	Boy 12 A
Demag Ergotech GmbH	IntElect 50
Engel Austria GmbH	E-Motion (X-Melt)
Krauss Maffei-Kunststofftechnik GmbH	KM 35-55 CX
Mitsui Machine Tool – Fanuc	Roboshot S-2000i 5A
Sumitomo Plastics-Machinery	SE18D
Toshiba Machine Co.	EC20P-0.4
Two-step system	
Semi microtechnology – split plasticizing and injecting screw-piston-machines for micro parts	
Cronoplast, sl bzw. ICH-Christmann	Babyplast 6/10
Ettlinger Kunststoffmaschinen GmbH	D-I-M
Ferromatik Milacron GmbH	K-TEC
Kunststoff-Zentrum Leipzig GmbH	formicaPlast M2
Sodick Plustech Co., Ltd.	TR05EH
Three-step system	
Micro technology – split plasticizing, dosing and injecting screw-piston-machines for precise, small micro parts	
Battenfeld Kunststoffmaschinen GmbH	Microsystem 50

the micro-miniature components and their optimization in addition to part geometry. Surace et al. [12] reported these main process parameters:

Mold temperature	Melt temperature
Injection speed	Injection pressure
Holding time	Holding pressure
Cooling time	

Microinjection Molding Machines

Some microinjection molding machine manufacturers and their products are listed in Table 2 [56].

Conclusion

Microinjection molding is the process of transferring a thermoplastic material in the form of granules from a hopper into a heated barrel so that it becomes molten and soft. The material is then forced under pressure inside a mold cavity where it is subjected to holding pressure for a specific time to compensate for material shrinkage. The material solidifies as the mold temperature is decreased below the glass transition temperature of the polymer. After a sufficient time, the material freezes into the mold shape and gets ejected, and the cycle is repeated. Microinjection modeling is a very specialized precision injection molding technique. The technique can be used in microfluidic devices for various applications such as micromechanical parts, micro gear wheel, and medical, optical, and electronic industries.

The advantages of using microinjection molding include cost-effectiveness for mass production process, its availability for a wide range of thermoplastics, its potential for full automation with short cycle times and good dimension control and accurate shape replication, etc. Among several materials, polymers show superior properties that become promising materials for microinjection molding. The material should possess appropriate mechanical properties, thermal properties, appropriate glass transition temperature Tg, melt temperature Tm and thermal expansion coefficient, good heat dissipation ability, chemical resistance, properties of dimensional stability and accuracy, good electrical insulating properties, biocompatibility and permeability, high mechanical strength, low viscosity, and good flow properties.

Many studies discovered that the process parameters such as mold temperature, melt temperature, injection speed, injection pressure, holding time, and holding pressure will influence the quality of the micro-miniature parts. Several microinjection molding machine manufacturers and their products are listed in this article.

Cross-References

- ▶ Micro- and Nanofluidic Devices for Medical Diagnostics
- ▶ Micro-/Nanodevices
- ▶ Micro-/Nanofluidic Devices
- ▶ Micro-/Nanomachines
- ▶ Micro-/Nanosystems
- ▶ MicroElectroMechanical Systems
- ▶ Microfluidic Device
- ▶ Microfluidic Optomechanics
- ▶ Micromachining
- ▶ Micromechanical Switches
- ▶ Micro-optomechanical Systems (MOMS)

References

1. Shah, V.: Consultek Consulting Group: Micro Injection Molding, http://www.consultekusa.com/pdf/Consulting/Process/MICRO%20INJECTION%20MOLDING.pdf
2. Attia, U.M., Marson, S., Alcock, J.R.: Micro-injection moulding of polymer microfluidic devices. Microfluid. Nanofluid. **7**, 1–28 (2009). Springer-Verlag 28 February 2009
3. Niggemann, M., Ehrfeld, W., Weber, L.: Fabrication of miniaturized biotechnical devices. In: Proceedings of SPIE – the International Society for Optical Engineering, Santa Clara, 21–22 September 1998. (1998)
4. Piotter, V., Hanemann, T., Ruprecht, R., Haußelt, J.: Microinjection Moulding of Medical Device Components. Business Briefing: Medical Device Manufacturing and Technology 2002, Business Briefings Ltd. London, United Kingdom, pp. 63–66. (2002a)
5. Piotter, V., Guber, A.E., Heckele, M., Gerlach, A.: Micro moulding of medical device components.

Business briefing: medical device manufacturing and technology, 2004, http://www.touchbriefings.com/pdf/954/piotter.pdf (2004). Accessed Feb 2007
6. Weber, L., Ehrfeld, W.: Micro-moulding – processes, moulds, applications. Kunsts. Plast. Eur. **88**, 60–63. 10 Munich (1998)
7. Becker, H., Locascio, L.: Polymer microfluidic devices. Talanta **56**(2), 267–287 (2002)
8. Weber, L., Ehrfeld, W., Freimuth, H., Lacher, M., Lehr, H., Pech, B.: Micromolding: a powerful tool for large-scale production of precise microstructures. In: Proceedings of SPIE – the International Society for Optical Engineering, Austin, 14–15 October 1996. (1996)
9. Weber. L., Ehrfeld, W.: Molding of microstructures for high-tech applications. In: Proceedings of the 56th Annual Technical Conference (ANTEC 1998). Part 3 (of 3), Atlanta, 26–30 April 1998. (1998)
10. Angelov, A., Coulter, J.: Micromolding product manufacture – a progress report. In: Proceedings of the Annual Technical Conference (ANTEC 2004), Chicago, 16–20 May 2004. (2004)
11. Ruprecht, R., Hanemann, T., Piotter, V., Haußelt, J.: Polymer materials for microsystem technologies. Microsyst. Technol. **5**(1), 44–48 (1998)
12. Surace, R., Trotta, G., Bellantone, V., Fassi, I. The Micro Injection Molding Process for Polymeric Components Manufacturing, New Technologies – Trends, Innovations and Research. ISBN: 978-953-51-0480-3 (2012)
13. Kukla, C., Loibl, H., Detter, H., Hannenheim, W.: Micro-injection moulding – the aims of a project partnership. Kunsts Plast. Eur. **88**(9), 6–7 (1998)
14. Whiteside, B., Martyn, M., Coates, P., Allan, P., Hornsby, P., Greenway, G.: Micromoulding: process characteristics and product properties. Plast. Rubber Compos. **32**(6), 231–239 (2003)
15. Martyn, M., Whiteside, B., Coates, P., Allan, P., Greenway, G., Hornsby, P.: Aspects of micromoulding polymers for medical applications. In: Proceedings of the Annual Technical Conference (ANTEC 2004), Chicago, 16–20 May 2004. (2004)
16. Bibber, D.: Advanced micromolding applications. In: Proceedings of the Annual Technical Conference (ANTEC 2005), Boston, 1–5 May 2005. (2005)
17. Tom, A., Layser, G., Coulter, J.: Mechanical property determination of micro injection molded tensile test specimens. In: Proceedings of the Annual Technical Conference (ANTEC 2006), Charlotte, 7–11 May 2006. (2006)
18. Michaeli, W., Rogalla, A., Ziegmann, C.: Processing Technologies for the Injection Moulding of Hybrid Microstructures. Macromol. Mater. Eng., vol. 279. WILEY-VCH Verlag GmbH, D-69451 Weinheim 2000, 1438-7492/2000/0106–0042. (2000)
19. Yao, D., Kim, B.: Injection molding high aspect ratio microfeatures. J. Inject. Molding Technol. **6**(1), 11–17 (2002)
20. Wimberger-Friedl, R.: Injection molding of sub-lm grating optical elements. Inject. Molding Technol. **4**(2), 78–83 (2000)
21. Pirskanen, J., Immonen, J., Kalima, V., Pietarinen, J., Siitonen, S., Kuittinen, M., et al.: Replication of sub-micrometre features using microsystems technology. Plast. Rubber Compos. **34**(5–6), 222–226 (2005)
22. De Mello, A.: Plastic fantastic? Lab. Chip. Miniatuarisation Chem. Biol. **2**(2), 31N–36N (2002)
23. Heckele, M., Schomburg, W.: Review on micro molding of thermoplastic polymers. J. Micromech. Microeng. **14**(3), R1e14 (2004)
24. Rötting, O., Röpke, W., Becker, H., Gärtner, C.: Polymer microfabrication technologies. Microsyst. Technol. **8**(1), 32–36 (2002)
25. Bissacco, G., Hansen, H.N., Tang, P.T., Fugl, J.: Precision manufacturing methods of inserts for injection molding of microfluidic systems, Micro Nano Workshop, Department of Manufacturing Engineering and Management, Technical University of Denmark, Produktionstorvet, Kgs. 2800 Lyngby, Denmark, Spring 2005 (2005)
26. Piotter, V., Mueller, K., Plewa, K., Ruprecht, R., Haußelt, J.: Performance and simulation of thermoplastic micro injection molding. Microsyst. Technol. **8**(6), 387–390 (2002)
27. Ong, N., Zhang, H., Woo, W.: Plastic injection molding of high aspect ratio micro-rods. Mater. Manuf. Process. **21**(8), 824–831 (2006)
28. Despa, M., Kelly, K., Collier, J.: Injection molding using high aspect ratio microstructures mold inserts produced by LIGA technique. In: Proceedings of SPIE—the International Society for Optical Engineering, Santa Clara, 21–22 September 1998. (1998)
29. Mönkkönen, K., Hietala, J., Pääkkönen, P., Pääkkönen, E., Kaikuranta, T., Pakkanen, T., et al.: Replication of sub-micron features using amorphous thermoplastics. Polym. Eng. Sci. **42**(7), 1600–1608 (2002)
30. Gottschlich, N.: Production of plastic components for microfluidic applications. Business briefing: future drug discovery, 2004. http://www.touchbriefings.com/pdf/855/fdd041_greiner_tech.pdf (2004). Accessed 2007
31. Xu, G., Kim, D., Koelling, K., Lee, L.: Flow dynamics in injection molding with microfeatures. In: Proceedings of the Annual Technical Conference (ANTEC 2005), Boston, 1–5 May 2005. (2005)
32. Schneider, C., Maier, G.: Small, but potent: special plastics for injection moulding microparts. Kunsts Plast. Eur. **91**(3), 27–28 (2001)
33. Gerlach, A., Knebel, G., Guber, A.E., Heckele, M., Herrmann, D., Muslija, A., et al.: Microfabrication of single-use plastic microfluidic devices for high-throughput screening and DNA analysis. Microsyst. Technol. **7**(5–6), 265–268 (2002)
34. Zhao, J., Mayes, R., Chen, G., Chan, P.S., Xiong, Z.J.: Polymer micromould design and micromoulding process. Plast. Rubber Compos. **32**(6), 240–247 (2003)

35. Song, S., Lee, K.Y.: Polymers for microfluidic chips. Macromol. Res. **14**(2), 121–128 (2006)
36. Stange, T.: Development and production of microfluidic chips made of polymers. Am. Biotechnol. Lab. **20**(8), 8–10 (2002)
37. Yu, L., Koh, C., Lee, L., Koelling, K., Madou, M.: Experimental investigation and numerical simulation of injection molding with micro-features. Polym. Eng. Sci. **42**(5), 871–888 (2002)
38. Bourdon, R.: Short cycles for injection moulded microfluidics parts. Kunsts Plast. Eur. **93**(9), 9 (2003). 11?33
39. Boone, T., HughFan, Z., Hooper, H., Ricco, A., Tan, H., Williams, S.: Plastic advances microfluidic devices. Anal. Chem. **74**(3), 78A–86A (2002)
40. Liou, A., Chen, R.: Injection molding of polymer micro- and submicron structures with high-aspect ratios. Int. J. Adv. Manuf. Technol. **28**(11–12), 1097–1103 (2006)
41. Angelov, A., Coulter, J.: The development and characterization of polymer microinjection molded gratings. Polym. Eng. Sci. **48**(11), 2169–2177 (2008)
42. Macintyre, D., Thoms, S.: The fabrication of high resolution features by mould injection. Microelectron. Eng. **41–42**, 211–214 (1998)
43. Pranov, H., Rasmussen, H., Larsen, N., Gadegaard, N.: On the injection molding of nanostructured polymer surfaces. Polym. Eng. Sci. **46**(2), 160–171 (2006)
44. Angelov, A., Coulter, J.: A feasibility study for sub-100 nm polymer injection molding. In: Proceedings of the Annual Technical Conference (ANTEC 2007), 6–11 May 2007. (2007)
45. Gadegaard, N., Mosler, S., Larsen, N.B.: Biomimetic polymer nanostructures by injection molding. Macromol. Mater. Eng. **288**(1), 76–83 (2003)
46. Weber, L., Ehrfeld, W.: Micromoulding: market position and development potential. Kunststoffe **89**(10), 192–202 (1999)
47. Clay, J., Heggs, R.: Material challenges in medical micromolding applications. In: Proceedings of the Annual Technical Conference (ANTEC 2002), 5–9 May 2002. (2002)
48. Greener, J., Wimberger-Friedl, R.: Precision Injection Molding: Process, Materials, and Applications. Hanser Gardner Publications, Cincinnati (2006)
49. Giboz, J., Copponnex, T., Me'le', P.: Microinjection molding of thermoplastic polymers: a review. J. Micromech. Microeng. **17**(6), R96–R109 (2007)
50. German, R.M.: Powder Injection Molding. MPIF, Princeton (1990)
51. Okubo, K., Ishida, M., Tanaka, S., Nishiyabu, K.: Molding Techniques for Improvement in Dimensional Accuracy of Micro Metal Injection Molded Parts, p. 32. Japan Society of Powder and Powder Metallurgy, Japan (2004)
52. Okubo, K., Matsuzaki, S., Nishiyabu, K., Piotter, V.: Molding techniques for improvement in quality of micro-MIM products. In: Proceedings of Japan Society of Polymer Processing Annual Meeting II-212, pp. 99–102. (2004b)
53. Okubo, K., Tanaka, S., Nishiyabu, K.: Molding technique for improvement in dimensional accuracy of micro-MIM products. An international conference on the powder injection molding of metals, ceramics and carbides (PIM2005), pp. 252–264. (2005)
54. Okubo, K., Tanaka, S., Ito, H.: April 2009, Springer-Verlag 2009 Microsyst. Technol. **15**, 887–892 (2009)
55. German, R.M.: Powder Metallurgy Science, p. 215. Metal Powder Industry (1996)
56. Battenfeld Kunststoffmaschinen GmbH: Microsystem Presentation, http://www.battenfeld.ru/fileadmin/templates/docs/imm/microsystem_presentation.pdf

Micromachined 3-D Force Sensors

Ryan M. Pocratsky and Maarten P. De Boer
Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

Synonyms

Force measurement; Micromachining; Transduction

Definition

Micromachined three-dimensional (3-D) force sensors measure forces from 10 Newtons (N) to less than 1 nN. This large range depends on the details of the processing technique and the application. Loads are applied to a probe or stylus that extends from the moving body. The input signal can be transduced by various methods including optical, piezoelectric, piezoresistive, and capacitive. The latter two transduction methods are the focus of this entry. The force sensor design typically has a moving body connected by flexible members to a stationary structure. Displacement or strain is measured in the flexible members and correlated to the applied force.

Introduction

This entry provides an overview of a 3-D force sensor design and a summary of recently

published work. First, the various applications for 3-D force sensors are described. Then, design considerations such as sensing methods, cross-sensitivity, flexure design, probe or stylus design, and fabrication methods are discussed. A table that summarizes published results on 3-D force sensors is included. Finally, a section on current and future work is presented.

Force Sensor Applications

Micromachined 3-D force sensors are used in a consumer application known as the Smartpen™ and in disciplines such as biomedicine, robotics, and metrology. Micromachining is a process used to produce microstructures by depositing and selectively etching thin layers. The small-scale features of this process allow precise sensor dimensions to be achieved, which is needed to obtain high force resolution. Additionally, this process allows high volume production of identical devices with integrated mechanical and electrical components. Applications of 3-D force sensors will be described in this section.

In the Smartpen™ [1], the force sensor was developed to measure microscale forces and torques. Although not micromachined, it works according to principles that are used in micromachined 3-D force sensors. The pen utilizes the force sensor to record the motion of the pen on paper and in air [1]. The pen positions can be calculated and an electronic representation generated to store handwritten notes or sketches in a computer [1]. The force range for this sensor is ±5 N with an accuracy of ±0.015 N.

The information obtained from a 3-D micro-force sensor has been used for biomedical research including force feedback in surgical instruments [2] and cell injection studies [3] and for the evaluation of a prosthesis fitting [4]. The force feedback improves the performance of the operator by improving alignment. The surgical instrument utilized a sensor with a force range of ±1 N with a resolution of 5 mN [2]. This sensor was fabricated by wire-cut electrical discharge machining (EDM). The cell injection sensor has a ±300 mN force range in the out-of-plane direction and ±50 mN range in-plane [3]. The larger force range is required for the out-of-plane direction because it is the primary needle insertion direction. The prosthesis of interest was for an above-knee amputee. The user comfort is highly conditioned by the normal and shear forces between the socket surface and the stump [4]. This application required a sensor with a force range of 7 N [4]. A short probe length (200 μm) allows a sensor array to measure the shear and normal force distribution during locomotion, thus reducing skin damage without significantly affecting the dimensions of the prosthetic. To summarize, force feedback information from 3-D force sensors has improved the performance of surgeries [2], biological studies [3], and prosthetic fittings [4].

Another major area of use for 3-D force sensors is tactile sensing for robotic grasping. Arrays of force sensors detect the normal and shear forces developed during a tactile sensing or surface texture qualification operation in robotics [5]. Previous tactile sensors utilized polymer-based techniques, in which it is difficult to control material properties in all directions [6]. Therefore, a micromachined silicon force sensor was proposed by Chu et al. [6] because of its excellent mechanical properties. From Table 1 in the comparison section, the force range for tactile sensors varies from 10 mN (1 g) to about 2 N, depending on the grasping application.

A mesoscale 3-D force sensor has been designed for a surface forces apparatus (SFA) and can be scaled for use in an atomic force microscopy (AFM) application [14]. The fabricated sensor has a force range of 2.7 N out-of-plane and 1 N in-plane with resolutions of 0.23 and 0.16 mN, respectively [14]. Common sensing techniques for currently available AFM tips include optical and piezoresistive methods. However, such sensors work only in two directions, have different sensitivities in each direction, and there is cross-sensitivity between directions [14]. The sensor developed by Kristiansen et al. [14] has little to no cross-sensitivity and similar sensitivities in each direction. The additional dimensional information from the sensor in AFM or SFA applications allows frictional forces,

Micromachined 3-D Force Sensors, Table 1 Published 3-D force sensor showing their applications, sensing method, design features, and key sensor specifications

Author	Application	Sensing type	Body type	Force/ displacement range	Sensitivity/ resolution
Diddens et al. [1]	Smartpen™	Piezoresistance	Cylindrical structure with axial and radial beams	±5 N, 0.01 mm axially	
				0.1 mm tangentially	
Berkelman et al. [2]	Microsurgical instrument	Piezoresistance	Double cross design	±1 N	5 mN
Beccai et al. [4]	Biomechanical measurements	Piezoresistance	Cross-bar with fabricated stylus	7 N	$0.032 \pm 0.001\ N^{-1}$ out-of-plane
					$0.0323\ N^{-1}$ in-plane
Sieber et al. [3]	Cell injection	Piezoresistance	Cross-bar with stylus attachment	±50 mN in-plane	0.4 mN in-plane
				0–0.3 N out of plane	2 mN out of plane
Beyeler et al. [7]	Biological micro-characterization	Differential capacitive	Flexible beams with in-plane probe	1,000 µN	1.4 µN
				2.5 µm out of plane	
Chu et al. [6]	Tactile sensing for robot fingers	Differential capacitive	Membrane with fabricated stylus	0–1 g	$0.13\ pF\ g^{-1}$ out-of-plane
					$0.32\ pF\ g^{-1}$ in-plane
Vázsonyi et al. [5]	Tactile sensing	Piezoresistance	Membrane with fabricated stylus	Tested up to 2 N	3.68 mV/V/N
Arai et al. [8]	Micromanipulation	Piezoresistance	Cross-bar with a fabricated or attached stylus	100 µm	40 nm
Bütefisch et al. [9]	Micromaterial characterization	Piezoresistance	Membrane with attached stylus	0–60 µm in-plane	3 nm in-plane resolution
				0–10 µm out of plane	5 nm out of plane resolution
Levey et al. [10]	Micro-coordinate measurement machine	Piezoresistance	Cross-bar with an attached stylus	3.6 N	5 mN
				100 µm	33 nm
Tibrewala et al.[a] [11]	Nano-positioning and measuring machine	Piezoresistance	Multiple devices: membrane, cross-bar, beam	640 ± 10 µm in-plane	11.29 ± 0.29 mV/V/mN in-plane
					12.28 ± 0.09 mV/V/mN out of plane
				140 ± 5 µm out of plane	Resolution: 9 µN and 1.4 nm
Dai et al. [12]	Micro-dimensional metrology	Piezoresistance	Membrane with stylus attachment	0.73 mN	nm scale resolution
				200 nm	
Wang et al. [13]	Dimensional metrology	Piezoresistance	Cross-bar with an attached stylus	0–1 N	0.1682 mV/µN in-plane
					0.0106 mV/µN out-of-plane
				5 µm	Resolution: 3 µN

(*continued*)

Micromachined 3-D Force Sensors, Table 1 (continued)

Author	Application	Sensing type	Body type	Force/ displacement range	Sensitivity/ resolution
Kristiansen et al. [14]	Surface forces apparatus (SFA)	Piezoresistance	Cross-bar with an attached stylus	1 N in-plane	0.16 mN and 0.84 μm in-plane
				2.7 N out of plane	0.23 mN and 0.17 μm out of plane
Viet Dao et al. [15]	Turbulent flow measurement	Piezoresistance	Cross-bar without a stylus	1,000 mN in-plane	1.32 mV/mN out-of-plane
				100 mN out of plane	0.107–0.12 mV/mN in-plane
Jin and Mote [16]	Calibration of a 3-D force sensor	Piezoresistance	Cross-bar without a stylus		10 μN in-plane
					3 μN out of plane
Muntwyler et al. [17]	Tunable force sensor, calibration	Differential capacitive	Flexible beams with in-plane probe	adjustable from ±20 to ±200 μN	30 nN

[a]Best results presented

normal forces, and surface roughness to be distinguished. Additionally, greater topographical image resolution can be achieved because the actual position of the tip can be measured.

The measurement of tip position has driven the use of 3-D force sensors in coordinate measurement machines for micro- and nanocomponents [9–12]. Typically, optical methods are used for metrology for devices because of the excellent out-of-plane resolution. However, the moderate lateral resolution of optical sensors prevents them from being useful for structures with steep sidewalls or undercuts [10]. Sensor configurations that utilize multiple 1-D sensors have been used to obtain highly accurate measurements in multiple directions, but they require a complicated setup [11]. The designed 3-D force sensors for this application [9–12] have given measurement resolutions in the range of nanometers, which is similar to the resolution of a scanning electron microscope.

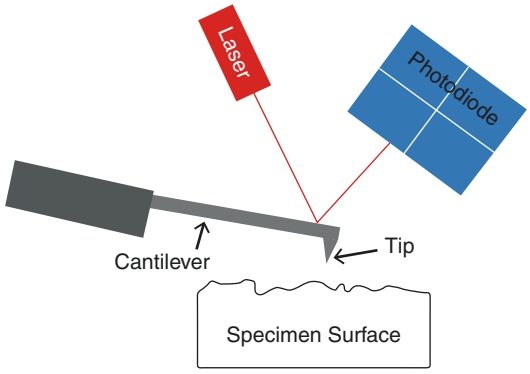
Design Considerations

There are three main components to the design of a 3-D force sensor: sensors, flexural members, and a probe or stylus. The fabrication process has significant impact on the design, performance, and production cost of the sensor. The number and thickness of fabrication layers as well as the etching process limits the possible dimensions of the flexible members. Additionally, the feature resolution through the masking process limits the performance and repeatability of the sensor.

The typical transduction methods used for micro-force sensors are optical, piezoelectric, piezoresistive, and capacitive. These methods are described next, but the optical and piezoelectric methods are used mainly in 1-D and 2-D force sensing. The other design considerations discussed in this section are cross-sensitivity, circuitry, flexure design, probe orientation, fabrication processes, and calibration.

Optical and Piezoelectric Transduction Methods

Optical and piezoelectric methods are mostly used for one- or two-dimensional force sensors. The optical method typically uses a laser reflected from the free end of the cantilever beam to a four-quadrant photodiode, which is the common method used for AFM. Figure 1 shows the typical setup for the optical method. The deflection of the cantilever changes the position at which the light is received on the photodiode. Micromachined



Micromachined 3-D Force Sensors, Fig. 1 Optical sensing method shown in a typical AFM application

cantilevers with a wide range of stiffness and tips are currently available for AFM applications. Based on the deflection and the known stiffness of the cantilever, the applied force can be calculated. The complicated and relatively large spatial setup of the optical systems prevents its extension to 3-D force sensors [18]. The piezoelectric method has also not been extended to 3-D because it can only measure dynamic loads, has a large sensitivity to temperature, and is difficult to set up for multiaxis measurements [18].

Piezoresistive and Capacitive Methods

The piezoresistive and capacitive methods are the primary transduction methods for 3-D micro-force sensors and will be described in this section.

Piezoresistive Method

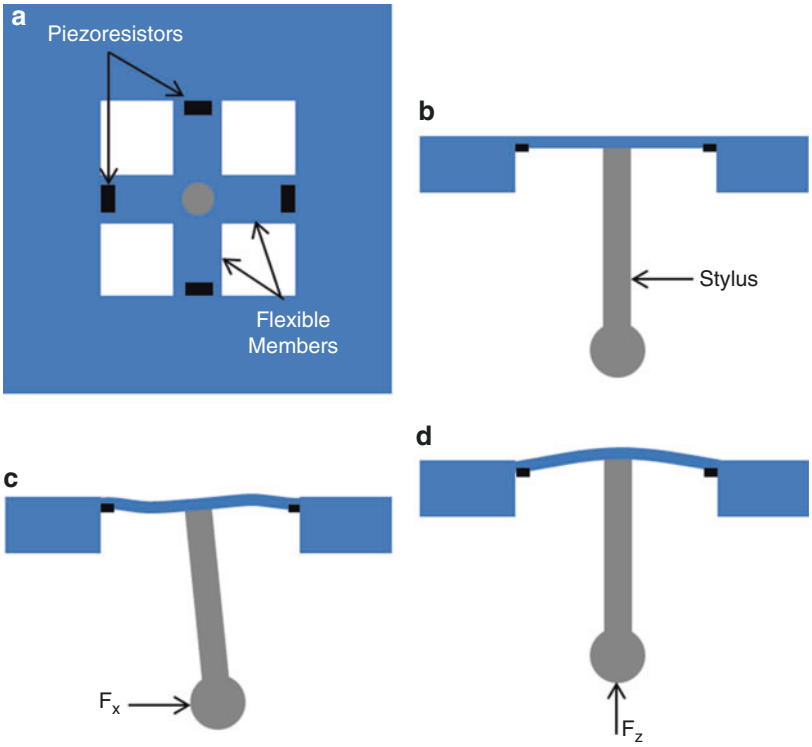
The piezoresistive method relies on strain in the flexible members to create a change in resistance. In piezoresistors, strain affects electron mobility, causing the resistance change. Piezoresistors are placed or fabricated in locations on the flexible members that experience the maximum strain to achieve the greatest sensitivity. Additionally, the flexible member design and piezoresistor locations should be optimized to reduce cross-sensitivity. Piezoresistor placement can be performed during the fabrication process or as a post-processing step. Generally, applying piezoresistors in a postprocessing step increases the variability, cross-sensitivity, and cost of the sensors. The fabrication process will be discussed briefly in the fabrication processes section below.

A common piezoresistive micro-force sensor is shown in Fig. 2. The top view, shown in Fig. 2a, shows a cross-bar flexure design with the typical location of the piezoresistors. Extending from the center of the device is the stylus, shown in Fig. 2b. Loads applied to the stylus tip cause deflections in the flexure members, which create strain in the piezoresistors. The flexure response due to in-plane loads and out-of-plane loads is shown in Fig. 2c, d, respectively. In this design, an out-of-plane load causes an equal amount of strain in all four sensors. An in-plane load applied along a flexible member direction causes equal and opposite strain in two of the four sensors. A simple algorithm can be applied to determine the amount of each.

Capacitive Method

The capacitive method relies on displacement changes, which affect parallel-plate capacitor gaps. A generic capacitive force sensor is shown in Fig. 3. Figure 3a shows the various components used in capacitive force sensor design. Flexure members and capacitors are in place for sensing in each direction. A cross section is shown in Fig. 3b to indicate the setup of the out-of-plane capacitor, C_z. If an out-of-plane force is applied to the probe tip, as shown in Fig. 3c, the inner body of the sensor rotates causing a change in capacitance C_z. Similarly, in-plane force responses are shown in Fig. 3d, e. The in-plane capacitors are shown in a differential configuration, which is explained in the circuitry section.

Capacitive force sensing is relatively insensitive to changes in light, temperature, and humidity and exhibits low drift [17]. In-plane capacitors are typically set up in a comb-finger arrangement to increase the overlap area when the device thickness is small. Out-of-plane capacitance is typically set up by a parallel-plate configuration. However, Muntwyler et al. [17] have developed a sensor that uses differential overlap area of two isolated layers for out-of-plane capacitance. The capacitors are typically placed in a location that has the greatest displacement to maximize sensitivity while isolating motion along each axis.

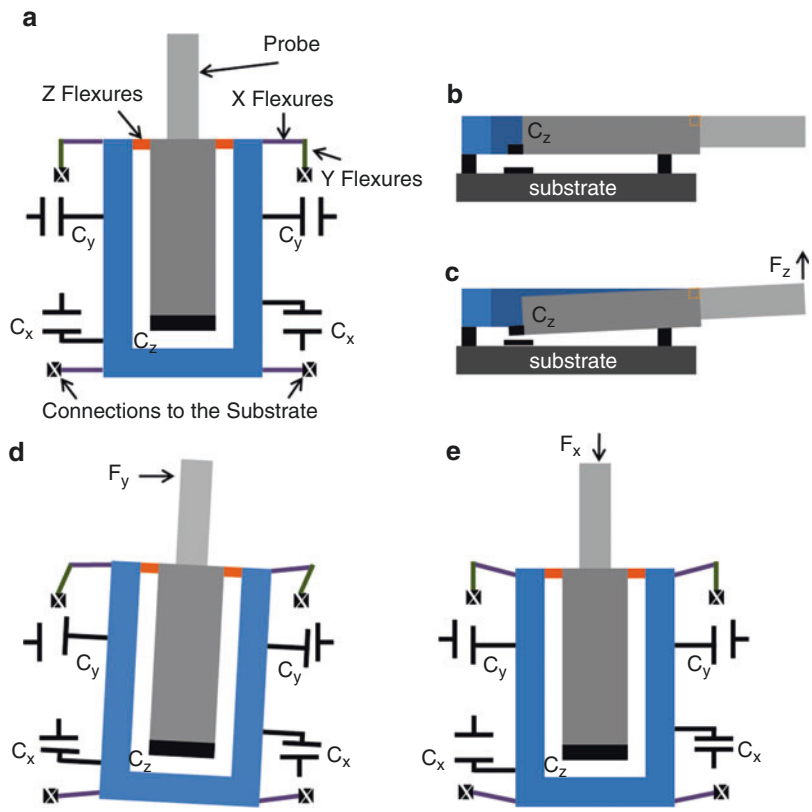


Micromachined 3-D Force Sensors, Fig. 2 Common piezoresistive micro-force sensor with cross-bar flexure design: (**a**) *top view*, (**b**) *side view*, (**c**) in-plane loading response, (**d**) out-of-plane loading response

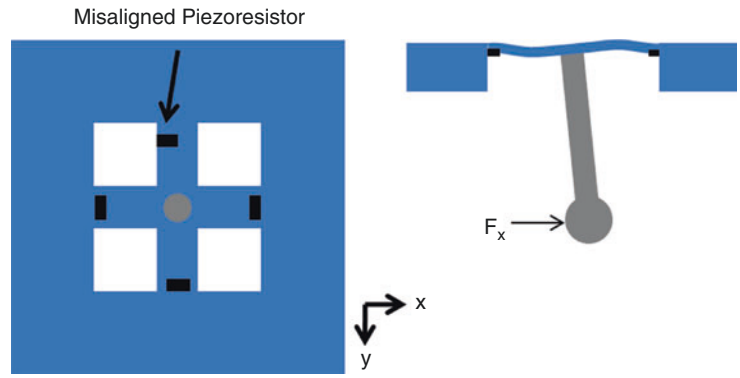
A drawback to this method is that the application of a sensing voltage to the capacitors creates an electrostatic force on the moving body. This electrostatic force generated from the input voltage signal could cause pull-in issues. Pull-in is an instability that occurs when the electrostatic force overwhelms mechanical spring stiffness. It causes a sudden displacement of the body that closes the capacitor gap, and short-circuiting will occur unless precautions are taken in the design.

Cross-Sensitivity

Cross-sensitivity occurs when a force applied in one direction causes a response in one or two perpendicular directions. It has also been referred to as cross talk or cross-coupling. It is minimized by designing a balanced force sensor with similar stiffness and sensitivity in each direction. Additionally, the location of each sensing element should be optimized to minimize cross-sensitivity. For example, Fig. 2c shows the response of flexures along the x-direction to an x-force. The response of the flexures along the y-direction is a twisting motion. The location of the y-piezoresistors is close to the device base, and therefore they experience minimal strain due to an x-force. Common contributions to cross-sensitivity include variations in material properties along the device, in flexure dimensions due to processing, in the location and fabrication of piezoresistors, and in the dimensions or motions of the parallel plates in capacitive sensors. Returning to the example described above from Fig. 2, the importance of precise positioning of piezoresistors can be seen. If the alignment of a y-piezoresistor were shifted from the center of the y-flexures, the cross-sensitivity between the x and y directions will increase. This misalignment is shown in Fig. 4. If a load is applied along the x-direction, the twisting motion in the y-flexures causes more strain in one of the y-piezoresistors than the other, which generates an output indicating a force along the y-direction. Cross-sensitivity contributes to the output noise from the sensor and reduces the reliable resolution of the sensor. Cross-sensitivity in capacitive sensors is described jointly in the Circuitry section below.



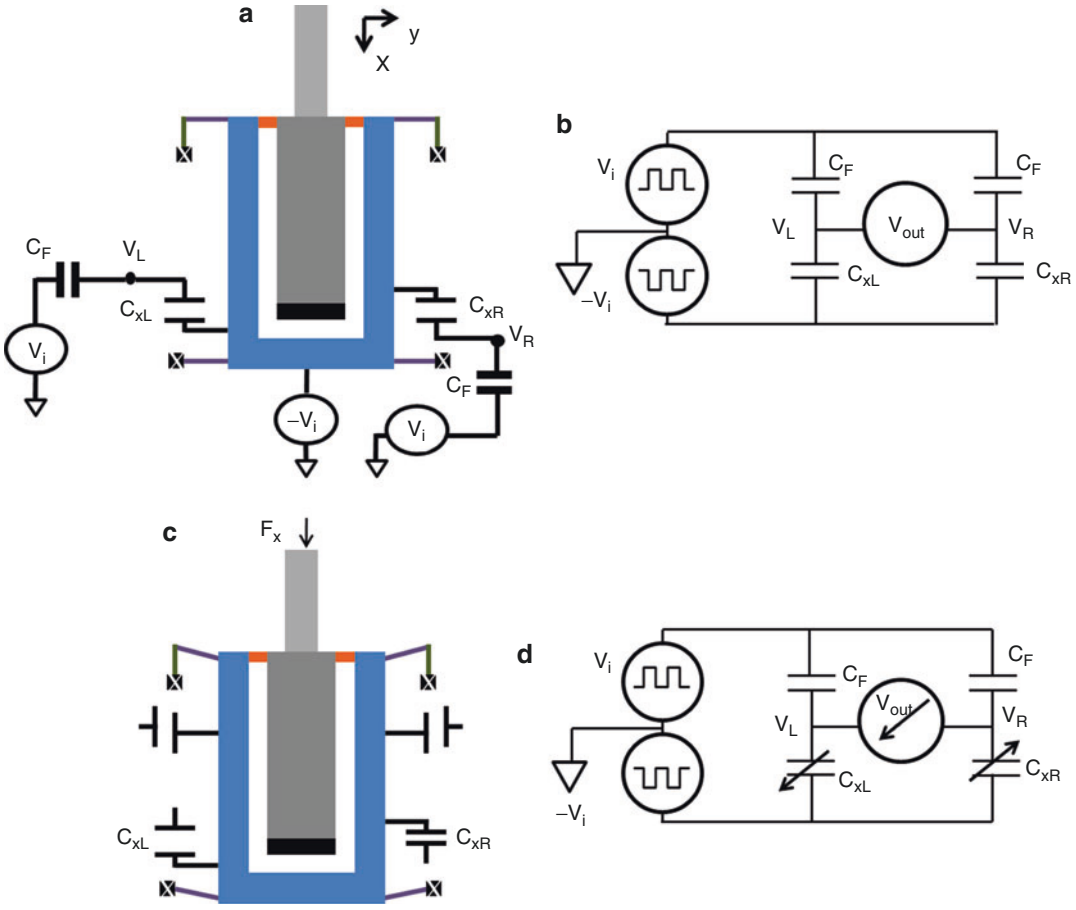
Micromachined 3-D Force Sensors, Fig. 3 Capacitive force sensor with beam flexures and in-plane probe: (**a**) *top view*, (**b**) *side view*, (**c**) out-of-plane loading response, (**d**) and (**e**) in-plane loading response



Micromachined 3-D Force Sensors, Fig. 4 Variations in the location of piezoresistors contribute to cross-sensitivity

Circuitry

Both piezoresistive and capacitive 3-D force sensors typically utilize a differential sensing configuration with a Wheatstone bridge. An example of the Wheatstone bridge circuitry for the capacitive method with differential capacitor setup is shown in Fig. 5. Figure 5a shows the capacitive device with the circuitry for the x-direction included. The Wheatstone bridge schematic is shown in Fig. 5b. The differential setup of the capacitors allows the gap of one capacitor to increase while the other decreases under the loading in the x-direction, as shown in Fig. 5c. This generates an output voltage that is correlated to the applied force. The circuit schematic shown in Fig. 5d shows the effect of the applied x-force. The left capacitance decreases, the right capacitance increases, and the output signal goes to a negative value. If displacement



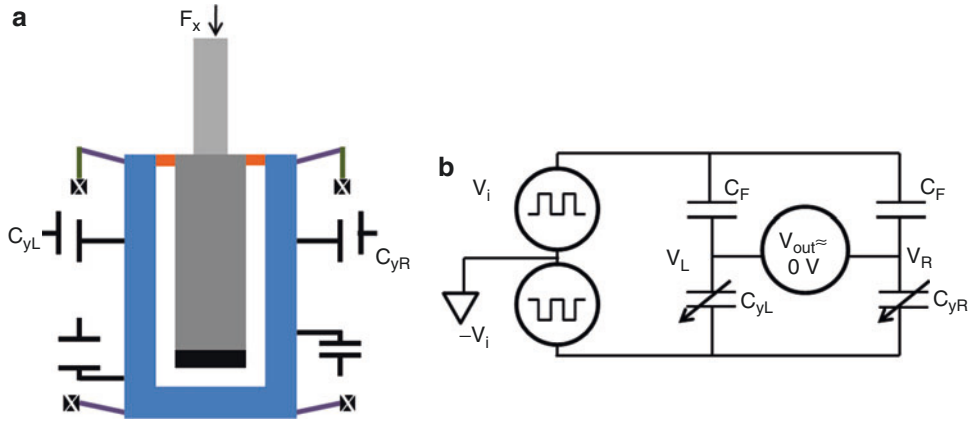
Micromachined 3-D Force Sensors, Fig. 5 Wheatstone bridge circuitry for a capacitive force sensor: (**a**) sensor circuit schematic, (**b**) Wheatstone bridge circuit, (**c**) capacitive response to in-plane loading, (**d**) Wheatstone bridge circuit response to in-plane loading

occurs in the opposite direction, the same voltage is output with the opposite sign. The differential resistive Wheatstone bridge behaves in the same way where the compression of one piezoresistor and extension of another creates the differential resistance legs of the circuit.

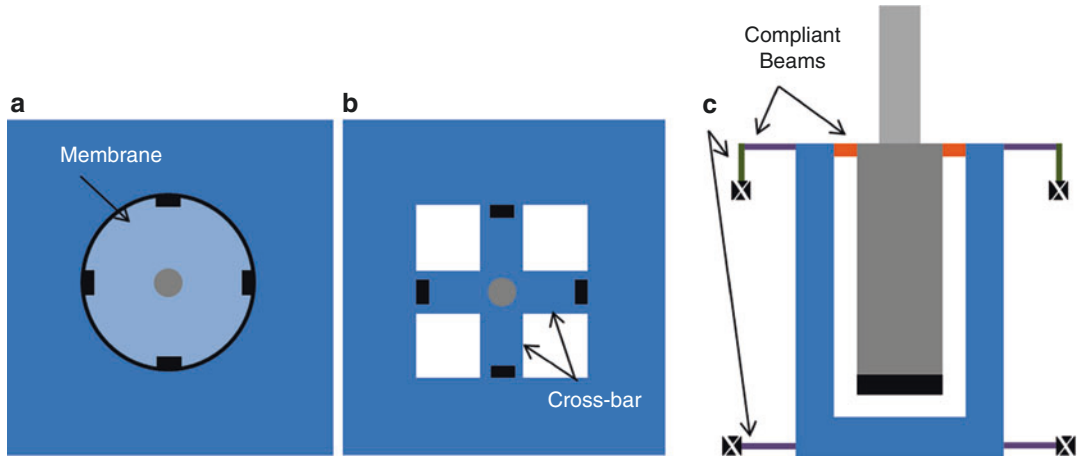
The differential sensing configuration and Wheatstone bridge circuitry can reduce cross-sensitivity, compared to measuring the output from individual sensing elements. Consider the capacitive force sensor shown in Fig. 3 with an applied load in the x-direction. Figure 6a shows the resulting displacement and effect on the y-capacitors. The overlap area of the y-capacitor plates decreases, which could result in an erroneous signal in the y-direction. However, both capacitors in the differential pair decrease by the same amount. The overall change in capacitance does not affect the output voltage for the y-direction because the same capacitance change is experienced on each side of the Wheatstone bridge, as shown in Fig. 6b.

Flexure Member Design

The design of the flexure members, or flexible members, controls the force range, displacement range, and sensitivity of the device. Three common designs are thin membrane, cross-bar, and compliant beams as shown in Fig. 7. Thin membrane, cross-bar, and combinations of these



Micromachined 3-D Force Sensors, Fig. 6 Differential capacitor configuration with applied x-load and net effect on the y-direction Wheatstone bridge circuit: (**a**) sensor response to in-plane loading, (**b**) circuit response for y-capacitors



Micromachined 3-D Force Sensors, Fig. 7 Common flexure designs: (**a**) thin membrane, (**b**) cross-bar flexures, (**c**) compliant beams

designs are typically used with the piezoresistance method, which can be seen from Table 1 in the comparison section. Compliant beams are typically used with the capacitive method. Analytical considerations determine the approximate stiffness and position of flexures. Then, finite element modeling (FEM) is used to optimize the design to achieve the desired stiffness for the target force range. The results from FEM are also used to determine the optimum placement of the piezoresistors. The flexure design is also affected by the length of the probe or stylus. A long probe tends to amplify the effect of applied loads because of the greater leverage. Therefore, the flexure members must have greater stiffness for longer probes for the same force range.

Probe Orientation

There are two probe, or stylus, orientations: in-plane and out-of-plane. Out-of-plane probes are common for piezoresistive sensors. The probe can be attached to the moving body as a postprocess operation or fabricated with the body. Attaching the probe allows for greater probe lengths than available from fabrication but introduces greater variability in position and increases cost.

Additionally, an off-center probe contributes to the cross-sensitivity of the sensor. In-plane probes are common for capacitive sensors. Since it is cofabricated, its length and position are easily controlled. However, additional processing could be required for applications that require the probe to extend beyond the edge of the substrate.

Fabrication Processes

The fabrication processes used to create the sensors in Table 1 include photolithography, multiple masks, reactive ion etching, and a wet release etch. Typically, bulk micromachining processes are used, but the piezoresistive sensors require additional processing to become functional. Piezoresistive sensors require the piezoresistors to be attached in postprocessing by epoxy or can be integrated in the fabrication process by ion implanting or diffusion. The doping of silicon with ions like boron can create a p-type piezoresistor directly on the flexure members of the device.

Calibration

Calibration of a force sensor requires a known load to be applied and correlated to the output signals. The increased importance of measuring small-scale forces has driven the need to verify these measurements. The small force metrology section of the mass and force group of the National Institute of Standards and Technology (NIST) has developed an electrostatic force balance that can be used to measure nanoscale forces and calibrate other force sensors. Their work has focused on achieving International System of Units (SI) traceability for forces below 1 mN to verify forces measured by indentation equipment and atomic force microscopes (AFMs).

For 3-D sensors, the orientation of the force vector also needs to be known to decompose it into the sensing directions. A matrix can be created to convert a vector of electrical signals to a vector of forces. Ideally, the relationship between force input and the output from the sensor would be linear. Nonlinearities in the response of the device cause additional error in the conversion matrix. For micro-force sensors, applying a known load in a known direction is a difficult task because the resolution of some sensors is smaller than that of the calibration method. However, standards for small force metrology are currently being developed. The sensors presented in Table 1 in the comparison section used the following methods for calibration.

Jin and Mote [16] used Lorentz forces generated from an aluminum cross-beam structure over the force sensor. The aluminum cross-beam is suspended over the force sensor and connected to its center where the probe would be located. This calibration structure is micromachined along with the sensor. Deposition of an aluminum layer on a patterned sacrificial area is used to create the cross-beams and connections to the sensor. Based on the applied current direction to the aluminum beams and magnetic field, forces in each direction can be generated. The analytical model of the generated force is used to calibrate the output signals from the force sensor.

Wang et al. [13] used a precision positioning micro-stage and an analytical balance to calibrate their sensor. The micro-stage has a resolution in the nanometer scale, and the balance has a resolution of 0.01 mg (0.098 μN). The stage is used to move the sensor tip into the balance. The output signals are calibrated to the measurement from the analytical balance. Butefish et al. [9] used the same method with a commercial compensation balance. Valdastri et al. [19] and Tibrewala et al. [11] used a similar method with a commercial load cell rather than a balance.

Dai et al. [12] used a method that requires a micromachined cantilever. Standards have been developed to calibrate the stiffness of micro-cantilevers. The calibrated micro-cantilever is then used to determine the stiffness of the micro-force sensor. The cantilever is moved into the force sensor probe tip along one direction. The deflection of the cantilever and the force sensor probe are measured. The known stiffness of the cantilever and these displacements are used to determine the spring constant of the force sensor. Currently, only one direction was calibrated with this method. Future work on a 3-D micro-force calibration standard will allow the other directions to be calibrated.

Comparisons

A table comparing various designs and sensing methods is shown in Table 1. The published sensors measure forces that range from less than 20 μN to 7 N with resolutions in the nanoNewton scale. The majority of 3-D force sensors utilize the piezoresistive method with a cross-bar and out-of-plane probe design.

Current Status and Future Work

Experiments that require monitoring multiple force components can be set up to use multiple single axis sensors or can be repeated multiple times to gather the necessary information. A 3-D micro-force sensor will reduce this complicated and time-consuming setup. Therefore, there is a growing need for the additional information from higher-dimensional force sensors that has created an active area of research. The sensors presented above were produced for specific research applications. Recently, Muntwyler et al. [17] have developed a sensor with an adjustable force range to allow it to be used for different applications. Future work includes SI-traceable calibration of 3-D force sensors and further improvements in resolution and force range to meet desired research study requirements.

Cross-References

- ▶ AFM Probes
- ▶ Atomic Force Microscopy
- ▶ Surface Forces Apparatus

References

1. Diddens, D., Reynaerts, D., Van Brussel, H.: Design of a ring-shaped three-axis micro force/torque sensor. Sensors Actuators A Phys. **46–47**, 225–232 (1995)
2. Berkelman, P.J., Whitcomb, L.L., Taylor, R.H., Jensen, P.: A miniature microsurgical instrument tip force sensor for enhanced force feedback during robot-assisted manipulation. IEEE Trans. Robot. Autom. **19**(5), 917–922 (2003)
3. Sieber, A., Valdastri, P., Houston, K., Eder, C., Tonet, O., Menciassi, A., Dario, P.: A novel haptic platform for real time bilateral biomanipulation with a MEMS sensor for triaxial force feedback. Sensors Actuators A Phys. **142**, 19–27 (2008)
4. Beccai, L., Roccella, S., Arena, A., Valvo, F., Valdastri, P., Menciassi, A., Carrozza, M.C., Dario, P.: Design and fabrication of a hybrid silicon three-axial force sensor for biomechanical applications. Sensors Actuators A Phys. **120**, 370–382 (2005)
5. Vázsonyi, É., Ádám, M., Dücsö, C.S., Vizváry, Z., Tóth, A.L., Bársony, I.: Three-dimensional force sensor by novel alkaline etching technique. Sensors Actuators A Phys. **123–124**, 620–626 (2005)
6. Chu, Z., Sarro, P.M., Middelhoek, S.: Silicon three-axial tactile sensor. Sensors Actuators A Phys. **54**, 505–510 (1996)
7. Beyeler, F., Muntwyler, S., Nelson, B.J.: A six-axis MEMS force-torque sensor with micro-Newton and nano-Newtonmeter resolution. J. Microelectromech. Syst. **18**, 433–440 (2009)
8. Arai, F., Sugiyama, T., Fukuda, T., Iwata, H., Itoigawa, K.: Micro tri-axial force sensor for 3D bio-micromanipulation. In: IEEE International Conference on Robotics and Automation, pp. 2744–2749. IEEE, Piscataway (1999)
9. Bütefisch, S., Büttgenbach, S., Kleine-Besten, T., Brand, U.: Micromechanical three-axial tactile force sensor for micromaterial characterization. Microsyst. Technol. **7**, 171–174 (2001)
10. Levey, B., Gieschke, P., Doelle, M., Spinner, S., Trautmann, A., Ruther, P., Paul, O.: CMOS-integrated silicon 3D force sensor system for micro component coordinate measurement machines. In: Proceedings of the IEEE Micro Electro Mechanical Systems, pp. 611–614. IEEE, Piscataway (2007)
11. Tibrewala, A., Hofmann, N., Phataralaoha, A., Jäger, G., Büttgenbach, S.: Development of 3D force sensors for nanopositioning and nanomeasuring machine. Sensors **9**, 3228–3239 (2009)
12. Dai, G., Bütefisch, S., Pohlenz, F., Danzebrink, H.U.: A high precision micro/nano CMM using piezoresistive tactile probes. Meas. Sci. Technol. **20**, 084001 (2009)
13. Wang, W., Zhao, Y., Lin, Q., Yuan, G.: A three-axial micro-force sensor based on MEMS technology. Int. J. Appl. Electromagn. Mech. **33**, 991–999 (2010)
14. Kristiansen, K., McGuiggan, P., Carver, G., Meinhart, C., Israelachvili, J.: 3D force and displacement sensor for SFA and AFM measurements. Langmuir **24**, 1541–1549 (2008)
15. Dao, D.V., Toriyama, T., Wells, J., Sugiyama, S.: Silicon piezoresistive six-degree of freedom micro force-moment sensor. Sens. Mater. **15**, 113–135 (2003)
16. Jin, W.L., Mote Jr., C.D.: Development and calibration of a sub-millimeter three-component force sensor. Sensors Actuators A Phys. **65**, 89–94 (1998)
17. Muntwyler, S., Beyeler, F., Nelson, B.J.: Three-axis micro-force sensor with sub-micro-Newton

measurement uncertainty and tunable force range. J. Micromech. Microeng. **20**, 025011 (2010)
18. Sun, Y., Nelson, B.J.: MEMS for cellular force measurements and molecular detection. Int. J. Inf. Acq. **1**, 23–32 (2004)
19. Valdastri, P., Roccella, S., Beccai, L., Cattin, E., Menciassi, A., Carrozza, M.C., Dario, P.: Characterization of a novel hybrid silicon three-axial force sensor. Sensors Actuators A Phys. **123–124**, 249–257 (2005)

Micromachined Acoustic Devices

▶ Integrated Micro-acoustic Devices

Micromachines

▶ Basic MEMS Actuators

Micromachining

▶ Micromachined 3-D Force Sensors

Micromechanical Resonators Fabricated Using CMOS-MEMS Technologies

▶ CMOS-MEMS Resonators

Micromechanical Switches

▶ Capacitive MEMS Switches

μMist®

▶ Spray Technologies Inspired by Bombardier Beetle

Micro-optomechanical Systems (MOMS)

▶ Nano-optomechanical Systems (NOMS)

Micropreconcentrator Technology for Portable Gas Chromatography System

Junghoon Yeom
Department of Mechanical Engineering, Michigan State University, East Lansing, MI, USA

List of Abbreviations

GC	Gas chromatography
GPC	Gas preconcentrator
IMS	Ion-mobility spectrometry
MS	Mass spectrometry
μGPC	Micro gas preconcentrator
DRIE	Deep reactive ion etching

Synonyms

Micropreconcentrators for Micro Gas Chromatography System; Microfabricated Preconcentrator-Focuser in Micro Gas Analyzers

Definition

A micro-preconcentrator is one of the key components in micro gas chromatography system and increasing the concentration of an analyte before detection by using high surface-area adsorbents and rapid thermal desorption of the trapped analytes.

Preconcentration refers to a process of concentrating an analyte (a substance to be identified) before analysis. It is often used by analytical chemists when a concentration of the target

analyte is too low and below the detection limit of a detector. Preconcentration is one of the sample treatment procedures, and many preconcentration methods utilize the isolation and separation techniques such as extraction, sorption, precipitation, distillation, filtration, and centrifugation. For instance, the concentration of solid or low-boiling-point analytes can be increased simply by evaporating or filtrating solvents. In this contribution, a gas-phase preconcentrator, a device that increases the net concentration of gaseous analytes, will be discussed with a particular emphasis on design and manufacturing aspects for small-scale portable applications.

A **gas preconcentrator (GPC)** is installed at the front end of an analytical system [1] to significantly improve the overall performance of the system. Examples of the analytical instruments that can benefit from GPC include gas chromatography (GC), mass spectrometry (MS), ion-mobility spectrometry (IMS), and electrochemical sensors. Concentration enhancement and subsequent improvement of the detection limit are not the only role of GPC [2]. In some cases, GPCs can purify a sample stream and reduce interferents by selective preconcentration. The use of the standard gas source in conjunction with GPC substantially facilitates calibration. Another important role of GPCs is sampling, and sample injection as a focused injected sample eluted into a column can significantly improve a chromatographic resolution.

Preconcentration Method for Gas Phase

Three main methods for gas preconcentration are cryogenic trapping, solvent extraction, and sorption trapping [3]:

(i) Cryogenic trapping typically utilizes a cooled tube filled with high surface-area glass beads through which an air with analytes flows. Cryogenic fluids such as liquid nitrogen or liquid CO_2 are used to achieve trapping temperature between $-150°C$ and $-170°C$. Trapped analytes are released from the cryotrap by direct heating, creating a concentrated band of analytes. The instrumentation involved in cryogenic cooling is however cumbersome and not suitable for portable and miniaturized system. Recently, pseudo-cryogenic trapping using micromachined Peltier coolers was implemented to generate sub-ambient temperature (+10 °C to −30 °C) but had to be combined with sorbents for efficient cryogenic preconcentration.

(ii) Solvent extraction is another popular gas sampling technique – dissolving airborne analytes in an appropriate liquid solvent for preconcentration. Dissolution of the analytes into a solvent is conducted either by bubbling an analyte-containing gas through a dispersion device (called an impinger) or by flowing a film of absorbing liquid along the inner wall of a cylindrical tube, while the gas sample passes through in a counter-flow manner (called a denuder). This solvent loaded with analytes is typically analyzed using liquid-phase detectors such as liquid chromatography and fluorescent spectroscopy, and therefore miniaturization of the solvent extraction method for portable applications appears to be impractical. Moreover, dissolution of analyte into a solvent is thermodynamically limited, and sometimes chromatographic peaks from analytes overlapping with the extract solvent peaks can be a concern.

(iii) Sorption trapping collects analytes onto solid sorbents based on sorption phenomena, including adsorption and absorption. Either active or passive sampling is used – for active sampling, pumping an air with analytes through a bed of sorbents in the order of minutes while, for passive sampling, a gas transport relying on pure diffusion taking in the order of days. Sorbents enriched with the analytes are typically desorbed by thermal desorption. The desorbed analytes form an injection plug in a carrier stream, subsequently eluted into a separation column. In comparison to cryogenic trapping or solid phase microextraction, the main disadvantages of the sorbent-based

preconcentration using conventional thermal desorption are significant thermal resistance between heating elements and sorbents and long diffusion paths, resulting in a slow analysis and high power consumption. Refocusing of analytes is often required for high-resolution chromatographic separation. Unlike the other sample trapping techniques, GPCs using solid sorbents can greatly benefit from miniaturization for the reasons to be discussed later and are relatively easier to be realized for portable applications.

Preconcentration Factor

The performance of the preconcentrator can be characterized by the preconcentration factor (or preconcentrator gain), G_{PC}, the ratio of the outlet concentration to the inlet concentration of an analyte. Various definitions of preconcentration factor have been introduced in literature, but the most straightforward concept is to take the ratio of the peak height obtained in a detector with preconcentration to the peak height obtained without preconcentration [4]. The other common definition, often used in conjunction with the injection system, is a ratio of the sampling volume, V_{sample}, to the desorption volume, $V_{desorbed}$, (which is the same as the preconcentration volume plus the dead volume).

$$G_{PC} = \frac{V_{sample}}{V_{desorbed}} = \frac{V_{sample}}{W_a \cdot \dot{Q}}$$

The desorption volume can be estimated as a product of the width of the injection peak, W_a, and the desorption flow rate, $\dot{Q}$. The column efficiency is measured by a number of theoretical plate, N_p, and the larger the plate number, the more efficient the chromatographic separation.

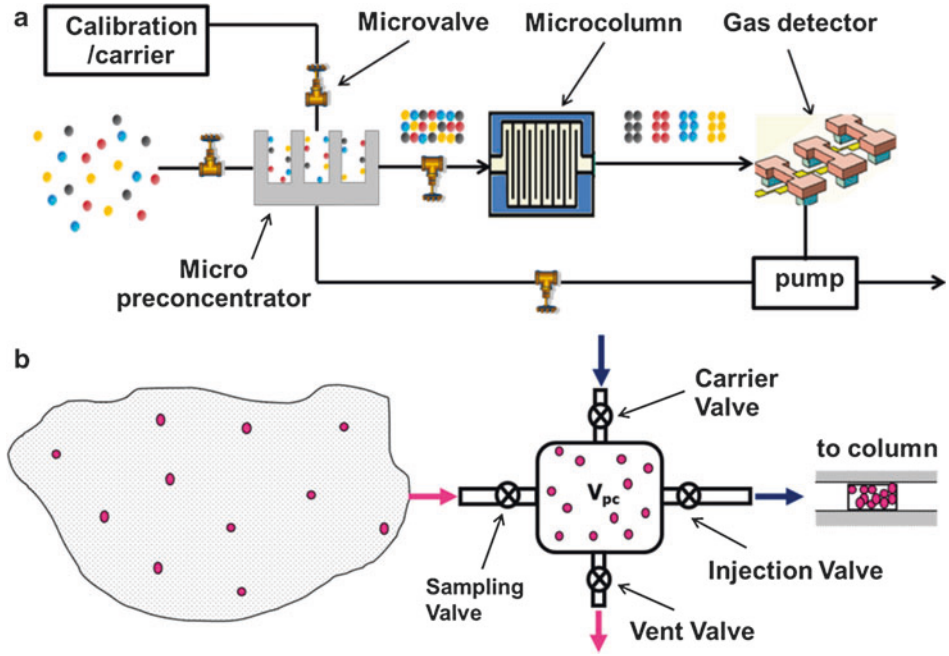
$$N_P = 5.545\left(\frac{t_R}{W_a}\right)^2$$

where t_R is the retention time. Therefore, the main role of the preconcentrator is to decrease W_a, i.e., a narrow bandwidth, for a higher number of available theoretical plate.

Micro Gas Preconcentrator (μGPC)

Many environmentally toxic chemicals and other harmful airborne substances are hazardous to human at concentration below 1 ppb, but many portable gas detectors have difficulties in detecting such a trace concentration of toxic vapors. In line with developing a microscale/portable detector with the highest possible sensitivity for a high performance miniature GC system, μGPCs can provide an alternative means to improve the detection limit of the detection system. The role of the μGPC is to capture the vapors of low concentration (<1 ppb) from ambient and thermally desorb them to a much smaller volume eluted in a separation column, thereby increasing the effective concentration to facilitate detection as well as providing a sharp injection plug to promote high-speed chromatographic separations.

Almost all μGPCs rely on sorption trapping preconcentration. There are two types of preconcentration mode with regard to sorption trapping: exhaustive versus equilibrium preconcentration [5]. In exhaustive preconcentration, all the analytes in a given sample volume are captured for analysis. A typical embodiment of exhaustive preconcentration is a tube packed with a granular adsorbent material, depleting all the sample gases and letting only a carrier gas pass through. While a quantitative trapping can be obtained using this method, it is important not to exceed the predetermined maximum sample volume, *aka* breakthrough volume. In equilibrium preconcentration, the analytes are absorbed into absorbent films until the analytes in stationary and mobile phases reach equilibrium. Because the analytes are not completely trapped by the sorbent system, the captured analyte mass is independent of the sampled volume but rather directly related to the concentration of analytes. The fact that the sample volume does not need to be measured is a key advantage of equilibrium preconcentration approach.



Micropreconcentrator Technology for Portable Gas Chromatography System, Fig. 1 (**a**) A simplified diagram of a micro gas chromatography system including the key components such as a micro-preconcentrator, microvalves, a microcolumn, and a gas detector(s) and (**b**) a diagram of the micro-preconcentrator with the four microvalves (sampling, carrier, vent, and injection) – sampling a trace concentration of target analyte into a small volume reactor packed with adsorbents and later thermally desorbing the analyte into a column

Figure 1a shows how the key components (μGPC, microvalves, microcolumn, sensor array, micropump, and carrier source) are connected in the microGC with a detailed view of the μGPC in Fig. 1b. First, a low concentration of the analyte is pumped into the μGPC chamber and preferentially adsorbed/absorbed onto/into sorbents from a gas stream. In this sampling stage, sampling and vent valves are open, while the other two valves are closed. Once sufficient mass of analytes is gained in the sorbents, the adsorbed/absorbed analytes are thermally desorbed into a smaller volume (corresponding to the chamber volume of the μGPC) and swept into a microcolumn with a carrier gas. In this injection stage, carrier and injection valves are open, while the sampling and vent valves are closed. In the end, the eluted peaks would become sharp and tall with the μGPC, increasing overall sensitivity and detection limit of the system.

Thermal Desorption The trapped analytes need to be recovered for the analysis, and for a microsystem, thermal desorption is considered to be the only viable method due to its simplicity and compatibility with miniaturization. Thermal desorption is carried out by a rapid heating of the sorbents to release the trapped analytes, and this thermally desorbed plug is often coupled to a GC column for efficient separation. The factors that influence the desorption efficiency and the narrowness of the desorption peak are the preconcentrator geometry, sorbent materials, and the operation parameters. These factors need to be considered when designing the μGPC. Especially the geometry of μGPC and sorbent type are the most important design consideration. Some of the operation conditions such as a heating rate, final desorption temperature, desorption duration, and a flow rate will also have an effect on the μGPC design. Desorption performance can be measured by the width and height of desorbed band (*aka* injected plug) – the narrower and taller the desorbed peak, the better the preconcentration device. In general, a fast heating rate and a higher

flow rate produce an arrow bandwidth in the peak. The operation parameters for desorption cannot be however arbitrarily chosen, and a care must be taken such that the desorption temperature is not too high to cause the decomposition of the analytes or sorbent degradation.

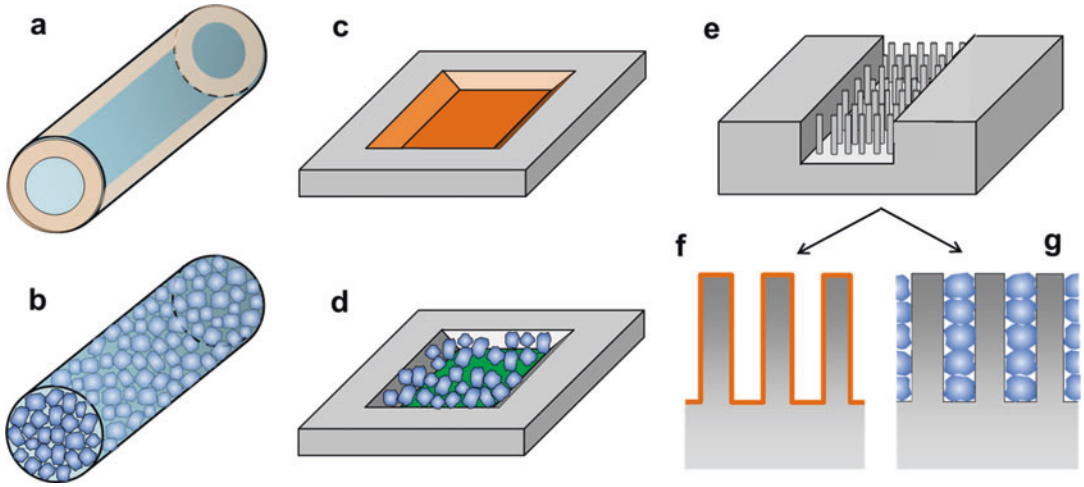
Absorbent/Adsorbent Materials A great deal of absorbent or adsorbent materials has been developed and commercially available for gas preconcentration and sampling applications. While absorbent refers to a process by which a material absorbs analytes into its volume, adsorbent is a process by which analytes are accumulated on the surface of the sorbent material. Therefore, whether absorbent or adsorbent materials are used and what characteristics sorbent materials should have are determined based on the type of target analytes. Some common criteria of the sorbent materials include (i) the specific surface area and the porous structure to trap the target analyte, (ii) ability to completely and quickly desorb the analytes, (iii) reversible absorption (or adsorption) and desorption, (iv) chemical stability especially with humidity, (v) thermal stability, and (vi) durability or reusability. Supelco (subsidiary of Sigma Aldrich) is one of the leading manufacturers/distributors for sorbents and has retention and desorption data available for a wide range of sorbents, facilitating the selection of the appropriate sorbents.

The types of absorbent/adsorbent materials are carbon-based or inorganic granular adsorbents and porous polymers. Porous polymers are typically applied as a thin film onto the structure within the μGPC and have smaller surface area and breakthrough volume in comparison to granular-type sorbents. Instead, the low partition coefficient of the porous polymers allows equilibrium in the analyte concentration between the stationary phase (polymer coating) and the mobile phase (carrier gas) to be reached relatively quickly. One of the main disadvantages of porous polymer-based sorbents is its low thermal stability, making it difficult to use them for high-boiling-point compounds that require higher thermal desorption temperatures. Popular trade names for porous polymer sorbents are Tenex®, Porapak™, Chromosorb®, and HayeSep®.

Conversely, granular adsorbents, in general, possess higher breakthrough/retention volumes than porous polymers, leading to higher adsorption capacity and preconcentration factor. Some of the granular adsorbents (e.g., carbon-based adsorbents) are stable at higher temperatures suitable for analysis of high-boiling-point compounds. But it requires a longer time to reach an equilibrium, and a high adsorption capacity typically results in poor desorption efficiency, meaning that it takes more time and energy to completely desorb the trapped analytes. Common granular adsorbents for the μGPCs include inorganic molecular sieves (silica or alumina zeolite), carbon-based molecular sieves (Carboxen® or Carbosieve®), and graphitized carbon black (Carbopack™ series and Carbotrap®). Note that porous polymer adsorbents can be also available in a granular form.

Design of μGPC Portable air sampling and preconcentration have been conducted using a tube preconcentrator, typically consisting of a stainless steel or glass tube packed with one or multiple adsorbent materials. Camsco Inc. is one of the leading manufacturers to produce thermal desorption preconcentrator tubes, and each tube costs around $100. A personal air sampling pump draws a known volume of air at a set flow rate (e.g., 100 ∼ 500 mL/min) through the tube preconcentrator. After completion of sampling in the field, desorption of analytes from the preconcentrator is performed using the desorption apparatus in the laboratory for further analysis. The size and cost of this desorption instrument are similar to the benchtop GC system, and therefore the miniaturization of the desorption instrument along with the preconcentrator unit has been pursued for the past two decades so that both air sampling and analysis can be done on site.

Like the conventional preconcentrators, μGPCs need to be equipped with three key components – adsorbent materials, a heating element for desorption, and fluidic interconnects. In order to promote miniaturization of the device, the preconcentration chamber which holds adsorbents



Micropreconcentrator Technology for Portable Gas Chromatography System, Fig. 2 Schematic diagrams of various μGPC designs: (**a**) wall-coated open tube, (**b**) a packed-bed tube, (**c**) a microheater membrane coated with a thin adsorbent layer, (**d**) a microheater surface packed with granular adsorbents, and (**e**) a preconcentrator chamber filled with an array of microposts; two possible ways to incorporate adsorbents are the surface of the microposts is coated with a thin adsorbent layer (see **f**) and the empty space packed with granular adsorbents (see **g**)

is generally made of a thermally conducting material and directly integrated with a heating element. Figure 2 shows schematic diagrams of the common μGPC designs that have been explored by various research groups. Capillary preconcentrators coated with a thin layer of the adsorbent layer on the inner wall or packed with granular adsorbents are shown in Fig. 2a, b, respectively. The configuration and operation of the capillary preconcentrators are similar to those of the separation column: porous layer open tubular column for Fig. 2a and packed-bed column for Fig. 2b. While capillary preconcentrators offer an effective test platform to characterize the adsorbents in light of preconcentrator operation, the fabrication process is not compatible with the other components of the micro gas chromatography system. A micromachined chamber atop the membrane heater in Fig. 2c and d coated with a thin layer of the adsorbent layer or packed with granular adsorbents is a popular design. One example of the membrane microheater μGPC is the μChemLab® at Sandia National Laboratories, employing a freestanding silicon nitride microhotplate onto which the porous polymer adsorbents were spray-coated [6]. Higher adsorption capacity and a resultant increase in a preconcentration factor can be achieved by incorporating three-dimensional microstructures into the preconcentration chamber (see Fig. 2e–g). An array of slats or posts is commonly utilized to distribute heat more uniformly.

These μGPC designs can be categorized into either equilibrium-type or exhaustive-type preconcentrators. The definition of each preconcentration mode was discussed in the previous section, and Table 1 summarizes the characteristics of each mode. In essence, the equilibrium-mode preconcentrators employ a thin-layer adsorbent layer, which becomes quickly saturated with the analyte concentration and facilitates fast desorption. The designs shown in Fig. 2a, c, f are considered to be equilibrium-mode preconcentrators. A thin adsorbent layer on the heater surface implicates their very low thermal mass and proximity of the thin adsorbents to the heater surface, permitting fast desorption kinetics and low power consumption for operation. But the main disadvantage is their inherently low adsorption capacity and thus low preconcentration factor. On the other hand, the exhaustive-type preconcentrators are supposed to trap all analytes from the sample stream for a larger preconcentrator factor, and therefore the

Micropreconcentrator Technology for Portable Gas Chromatography System, Table 1 Summary of the two different preconcentration modes

	Equilibrium preconcentration	Exhaustive preconcentration
Adsorption capacity (breakthrough volume)	Low	High
Adsorbent design	Thin layer	Packed-bed
Desorption efficiency	High	Low
Desorption time	Short (<100 ms)	Long (>1 s)
Power consumption	Low (<0.2 W)	High (>1 W)
Eluted peak width	Narrower	Broader

packed-bed designs with the granular adsorbents are widely used as shown in Fig. 2b, d, g. Closeness in the packed-bed adsorbents forces a sample gas to flow by poro-diffusion and induces a higher pressure drop (meaning more pumping power). Moreover, the high surface-area granular adsorbents are a poor thermal conductor, leading to a non-uniform heat distribution in the μGPC. Some adsorbents away from the heater surface in the packed-bed require a longer time for a complete desorption.

Whether or not the equilibrium-type or exhaustive-type preconcentrators should be used for μGPCs is related to two engineering trade-offs:

1. Miniaturization versus sample capacity – reduction in device size results in limited surface sites for adsorption and subsequent decrease in a preconcentration factor.
2. High density versus low density packing of adsorbents – solid granular adsorbents packed in the μGPC increase the surface area and adsorption capacity, which comes at a price of a large pressure drop for a fluid flow, large thermal contact resistance, and poor heat and mass transfer during sampling and desorption.

From Fig. 2, each type of the μGPCs can have a range of geometrical configurations and adsorbent/absorbent materials, and therefore a choice of the type, design, and materials of the μGPCs depends on the desired performance characteristics such as adsorption capacity and desorption efficiency (typically lumped as a preconcentration factor) and operation constraints (flow conductance or power consumption).

Fabrication of μGPCs Most of the μGPCs have been built on the silicon substrate because (i) silicon has a relatively high thermal conductivity (~150 W/m · K) for uniform unit conduction; (ii) doped silicon itself can be used as a microheater; (iii) a wide range of micromachining techniques are available for silicon, facilitating the fabrication of intricate geometries; and (iv) other micro gas chromatography components are also based on silicon – ease of integration. A few examples of silicon-based μGPCs will be discussed to delineate the fabrication sequences. The membrane heater-based μGPC like one from Sandia National Laboratories (Fig. 2c, d) consists of a low-pressure chemical vapor deposition (LPCVD) silicon nitride membrane on which a platinum microheater is photolithographically patterned and a silicon cavity created by KOH anisotropic etching followed by deposition of porous polymer adsorbents. The μGPCs with the 3D Si microstructures have been adopted by more research groups – the Wireless Integrated MicroSensing and Systems team from University of Michigan [7, 8], Shannon's group from University of Illinois [9, 10], and Agah's group from Virginia Tech [11, 12]. Arrays of microslats and posts with various cross-sectional geometries have been proposed to better trap analytes and at the same time reduce a pressure drop across the μGPC. These 3D microstructures often serve as microheaters and have been created using photolithography and deep reactive ion etching (DRIE). The porous polymer adsorbents are coated on the surfaces of microstructures (Fig. 2f) by spray-coating or flowing the polymer solution through the microchannels created after bonding with a cover plate. The granular adsorbents can be integrated into the voids of the 3D microstructures (Fig. 2g) by manual insertion with an aid of a shaking table or sonication.

The micromachining techniques such as photolithography, metal deposition, etching, and bonding are scalable (high throughput and large area) and can be used to build the μGPC structures in a manufacturable fashion. However, a method to integrate adsorbent materials is not scalable and incompatible with the micromachining techniques regardless of the type and design of the μGPC. These fabrication challenges associated with integrating heterogeneous materials into μGPCs are as follows:

1. μGPC packed with adsorbent particles: Packing adsorbent particles whose size is comparable to the channel dimension of the device is not trivial. A good packing density entails the application of high pressure, shaking, or ultrasonication, which can damage the microstructures of μGPC. Thus, far adsorbent packing has been carried out for one device at a time with a low success rate.
2. μGPC coated with a thin polymer absorbent: It is difficult to achieve uniform coating. Significantly thicker films are obtained in the corners or turns of the microchannel walls due to higher surface tension. Nonuniform thicknesses may cause slow or incomplete desorption.

Another common design/fabrication issue is the harsh operation condition (due to large temperature difference between adsorption and desorption states). Cyclic thermal stresses exerted to fluidic and electrical connections in μGPCs may cause the premature device failure [10].

References

1. Voiculescu, I., Zaghloul, M., Narasimhan, N.: Microfabricated chemical preconcentrators for gas-phase microanalytical detection systems. TrAC Trends Anal. Chem. **27**, 327–343 (2008)
2. Jackwerth, E., Mizuike, A., Zolotov, Y.A., Berndt, H., Hohn, R., Kuzmin, N.M.: Separation and preconcentration of trace substances. I. Preconcentration for inorganic trace analysis. Pure Appl. Chem. **51**, 1195–1211 (1979)
3. Demeestere, K., Dewulf, J., De Witte, B., Van Langenhove, H.: Sample preparation for the analysis of volatile organic compounds in air and water matrices. Adv. Sample Prep. Part II **1153**, 130–144 (2007)
4. Yeom, J., Agonafer, D.D., Han, J.-H., Shannon, M.A.: Low Reynolds number flow across an array of cylindrical microposts in a microchannel and figure--of-merit analysis of micropost-filled microreactors. J. Micromech. Microeng. **19**, 065025–065025 (2009)
5. Veeneman, R.A.: Design and Characterization of a Multi-Vapor Preconcentrator for a Micro-Scale Gas Chromatograph, Ph.D. Dissertation, University of Michigan, Ann Arbor, MI (2009)
6. Lewis, P.R., Manginell, R.P., Adkins, D.R., Kottenstette, R.J., Wheeler, D.R., Sokolowski, S.S., Trudell, D.E., Byrnes, J.E., Okandan, M., Bauer, J.M., Manley, R.G., Frye-Mason, G.C.: Recent advancement in the gas-phase MicroChemLab. J. IEEE Sens. **6**, 784–795 (2006)
7. Tian, W.-C., Pang, S.W., Lu, C.-J., Zeller, E.T.: Microfabricated preconcentrator-focuser for a microscale gas chromatograph. J. Microelectromech. Syst. **12**, 264–272 (2003)
8. Tian, W.-C., Chan, H.K.L., Lu, C.-J., Pang, S.W., Zellers, E.T.: Multi-stage microfabricated preconcentrator-focuser for microscale gas chromatography system. J. Microelectromech. Syst. **14**, 498–507 (2005)
9. Tang, Y., Yeom, J., Han, J., Bae, B., Masel, R.I., Shannon, M.A.: A micro-post preconcentrator for a microscale gas chromatography system. Presented at the proceeding of micro total analysis systems workshop, Boston, MA, 9/10/2005
10. Yeom, J., Field, C.R., Bae, B., Masel, R.I., Shannon, M.A.: The design, fabrication and characterization of a silicon microheater for an integrated MEMS gas preconcentrator. J. Micromech. Microeng. **18**, 125001–125001 (2008)
11. Alfeeli, B., Cho, D., Ashraf-Khorassani, M., Taylor, L. T., Agah, M.: MEMS-based multi-inlet/outlet preconcentrator coated by inkjet printing of polymer adsorbents. Sens. Actuators B Chem. **133**, 24–32 (2008)
12. Alfeeli, B., Agah, M.: Micro preconcentrator for handheld diagnostics of cancer biomarkers in breath. In: Sensors, 2010 IEEE, Kona, HI, pp. 2490–2493 (2010)

Micropreconcentrators for Micro Gas Chromatography System

▶ Micropreconcentrator Technology for Portable Gas Chromatography System

Micropumps

Shawn Litster
Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

Synonyms

MEMs pumps; Microfluidic pumps; Mini pumps

Definition

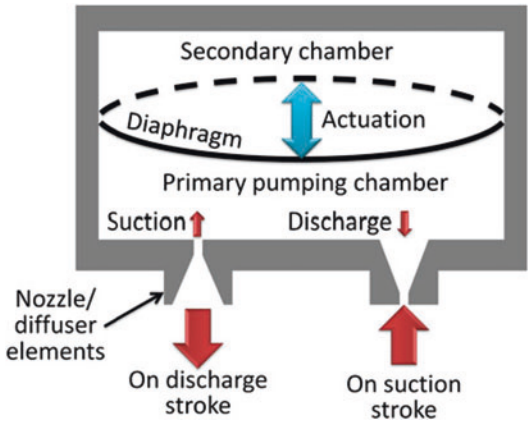
A micropump is a miniature fluid delivery device capable of generating flow at finite pressure loads. Micropumps are smaller than the conventional pumps, which use rotary or sliding motion to generate flow, and are usually made by microscale fabrication methods such as micromachining or microinjection molding. Because of their small size, micropumps are capable of using novel actuation mechanisms that are only effective with small length scales, such as piezoelectric strain and electrokinetic effects.

Overview

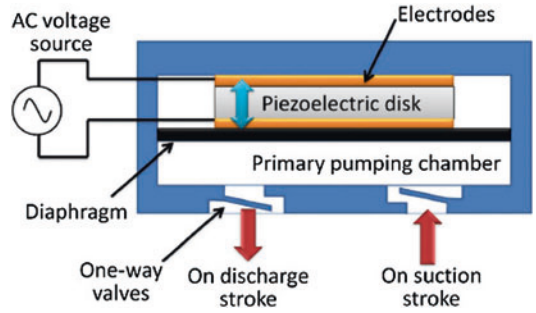
The demand for micropumps arises from a broad spectrum of applications, including drug delivery, coolant pumping for thermal management, fuel delivery for miniature energy conversion devices like micro-fuel cells, and fluid handling for miniature chemical analysis systems. The following is an introductory review of micropump technology that outlines the various classifications of micropumps and pumping mechanisms. The term "micropump" often refers to pumps with length scales below those of conventional fluid pumping devices that rely on rotary or piston-cylinder mechanisms. Thus, the size of micropumps ranges from lengths of several centimeters to tens of micrometers and less. Furthermore, micropumps typically do not feature the sliding contacts and gaps for the clearance of components, such as the vanes in conventional rotary vane pumps, due to fabrication tolerance limits and the dominance of surfaces, high viscous stresses, and low inertia at smaller scales. A common feature of micropumps is fluid conduits that are sufficiently small that viscous forces dominate over inertia and viscosity rapidly diffuses momentum (i.e., inertia). These conditions are characterized by a low Reynolds number, a nondimensional measure of the ratio of inertial forces to viscous forces.

Micropumps can also be identified by the use of micro-fabrication methods for constructing the pumps, often following microelectromechanical systems (MEMS) fabrication techniques. Often this entails micromachining of silicon and glass substrates, but also microinjection molding of polymers, polymer hot pressing, stereolithography, and printed circuit board methods. As advanced micro- and nanoscale fabrication technologies proliferate and become more affordable, the need for compact devices for pumping small volumes will become more pronounced.

Micropumps are often divided into two main groups: mechanical pumps that displace a diaphragm to generate flow and nonmechanical pumps that have no moving parts. Alternatively, micropumps can be classified by the temporal characteristics of their flow, including whether the flow is periodic as in the case of pumps with oscillating diaphragms, continuous as in the case of many dynamic nonmechanical pumps, or aperiodic as in the case of one-time, single-displacement pneumatic pumps. In the following section, the primary mechanical and nonmechanical types of micropumps are reviewed. These descriptions are intended as brief overviews of the more common micropump architectures and actuation mechanisms and this is certainly not an exhaustive review. For detailed reviews of micropump technology, the reader is referred to the more comprehensive reviews by Laser and Santiago [1], Woias [2], and Nisar et al. [3]. These reviews include comparative benchmarking of various micropumps, which is valuable in selecting a micropump for a given application.



Micropumps, Fig. 1 Schematic of mechanical displacement micropump. The pumped fluid is displaced within the primary pumping chamber and the actuation mechanism is located in the secondary chamber. This schematic is shown with nozzle/diffuser elements for directing the flow in preferential direction



Micropumps, Fig. 2 Schematic of a piezoelectric micropump that uses axial strain in a piezoelectric disk to displace the diaphragm. This pump is shown with one-way valves for setting the flow direction

Mechanical Pumps

As Fig. 1 shows, mechanical pumps use the displacement of a diaphragm to pump fluids. The fluid being pumped passes through the primary pumping chamber and the actuation component is typically the secondary chamber on the other side of the diaphragm. The pump operation consists of two strokes; the discharge stroke when the diaphragm pressurize the primary pumping and the suction stroke when fluid is drawn into the chamber as the diaphragm retracts. These pumps can be devised to deliver a one-time, finite volume fluid or to deliver a continuous, albeit pulsatile, flow. To achieve continual flow, mechanical micropump designs need to include valves or nozzles and diffusers for directing the flow in a preferential direction. Figure 1 illustrates a mechanical pump with nozzle/diffuser elements, whereas Fig. 2 shows a pump with a pair of one-way valves. The nozzle/diffuser approach requires a flow of sufficient inertia such that geometry imposes a preferential direction. Within the family of mechanical micropumps, different pumps technologies are distinguished by the origins of the actuation force that displaces the membrane. The various micropumps include those that have piezoelectric, electrostatic, electromagnetic, thermal, and chemical actuation of membrane displacement.

As depicted in Fig. 1, the actuated diaphragm pressurizes and displaces the fluid in the pump chamber. The valve or nozzle/diffuser arrangement directs the flow in the forward pumping direction. Mechanical pumps can also be arranged in series within elongated pump chamber to mimic a peristaltic pump using a traveling wave of actuation that continually pushes the fluids along in a forward direction. Thus, in those cases, the membranes also serve as the valves that direct the forward flow. A benefit of that arrangement is that the flow can be reversed. One useful design specification for mechanical pump is the compression ratio, which is the ratio of the pumps chamber volume without membrane actuation to the maximum value of the displaced volume with actuation. Another useful metric is the stroke length, which represents the maximum displacement of the diaphragm. The following subsections outline the specific methods for membrane actuation.

Piezoelectric Actuation

Flow in piezoelectric micropumps is actuated by the strain of a piezoelectric substrate upon the application of an electric field and the resulting diaphragm deflection. The use of piezoelectrics in micropumps can be traced to piezoelectric inkjet printer heads first developed by IBM in the 1970s. Although many variations exist, a common arrangement, shown in Fig. 2, is the placement

of a piezoelectric disk, with electrodes on each face, between a deformable diaphragm and the rigid housing of the pump. The pumped fluid is located in a chamber on the other side of the diaphragm. In this case, the axial strain in the piezoelectric disk is used for actuation. In another arrangement, the piezoelectric disk is affixed to the diaphragm and the radial strain induces the deflection of the diaphragm. The primary chamber is connected to the inlet and outlet by valves or a nozzle/diffuser pair. Periodic flow is generated by applying an axial AC electric field to the piezoelectric substrate, which generates a cycle of suction and discharge strokes similar to piston-cylinder pumps. These pumps benefit from being operated at high frequencies, often several kHz. For sufficient stroke length of the diaphragm, piezoelectric pumps are typically operated with high voltages, ranging from 50 to 300 V, which can be considered a disadvantage of piezoelectric pumps. Several commercial pumps feature embedded voltage step-up circuits and have a much lower device-level voltage input. Another challenge of piezoelectric pumps is the small stroke length due to strain limitations of the piezoelectric materials.

Electrostatic Actuation

Mechanical electrostatic micropumps use the electrostatic attraction between polarized parallel electrodes to deflect a diaphragm and produce flow. In this case, one of the electrodes is mounted on the diaphragm and the other on the rigid pump housing (top of the secondary chamber). When the voltage is applied, the electrodes experience a mutual attractive electrostatic force, generating the suction stroke. When the voltage is removed, the elastic deflection of the diaphragm is reversed and the diaphragm returns to its original position, generating the discharge stroke. Thus, an AC electric field or voltage pulses can be used to generate periodic flow. Two advantages of electrostatic pumps are their low power consumption and their fast response time. However, they are challenged by small stroke lengths (<5 μm). This small stroke arises because the induced electrostatic force on the diaphragm scales as one over the electrode spacing squared ($1/x^2$). Thus, the electrodes must be closely spaced, limiting the maximum displacement.

Electromagnetic Actuation

In contrast to electrostatic forces, strong electromagnetic forces can be induced over large length scales. In typical arrangements, a permanent magnet is mounted to the surface of the diaphragm and coils are attached to the housing to generate magnetic fields when electrical currents are applied. A benefit of electromagnetic actuation is the large forces with low applied voltages. Moreover, large diaphragm displacements have been reported with electromagnetic actuation, as much as 200 μm [3]. However, electromagnetic actuation is challenged by unfavorable scaling with miniaturization as well as the greater difficulties in fabricating micro-magnetic devices. In addition, these electromagnetic micropumps tend to be very inefficient, requiring large amounts of power and generating significant heat.

Thermal Actuation

Thermal actuation refers to mechanical micropumps that use heat to actuate the flow. There are two main categories of such pumps: (1) thermopneumatic and (2) phase-change. Thermopneumatic micropumps use the expansion of gases upon heating to generate the diaphragm displacement. The heat is generated by resistance heating (Joule heating) in thin-film heaters, such as aluminum traces. During heating, the expansion of the gas in the secondary chamber drives the diaphragm deflection and the discharge stroke. The temporal response of thermopneumatic pumps is slowed by the heat capacitance of the device, which slows the return of the diaphragm during the suction stroke when heat is removed. Some thermopneumatic micropumps vent the gas upon the beginning of the suction stroke to accelerate the diaphragm. Benefits of thermopneumatic pumps include the ability to generate high displacements. The high displacement is partially enabled by the use of low modulus diaphragms, unlike the more rigid diaphragms of the piezoelectric and electrostatic membranes. This difference is allowed because no components are mounted to the diaphragm in thermopneumatic pumps and

thus large strain is permissible. Unfortunately, thermopneumatic pumps consume a large amount of power and are rather inefficient.

Phase-change actuation for micropumps uses heat to generate displacements by the evaporation/condensation or melting/solidification of a secondary pumping fluid. The evaporation/condensation pumps are also sometimes referred to as "bubble" pumps. Ideally, the melting point or boiling point of the secondary fluid is not significantly far from the temperature of device's surroundings in order to reduce parasitic sensible heating. For evaporation/condensation pumps, the secondary fluids used include isopropyl alcohol. Paraffin has been previously used for melting/solidification pumps.

Bimetallic, Shape Memory Alloy, and Chemical Actuation

Other actuation methods for mechanical pumps include bimetallic, shape memory alloy, and chemical types. In general, these mechanisms share the feature of a bending displacement which actuates the flow. These actuators also share the feature that they can be used to form the diaphragm. Bimetallic and shape memory alloy actuation types use the response of these components to heat to actuate the diaphragm. Bimetallic actuation employs differences in the thermal expansion coefficients between layered materials to generate a bending displacement. Shape memory alloys use the change in these metal's properties caused by the transitions from the malleable martensite phase to the more rigid austenite phase with the application of heat. With these materials, the shape memory alloy can be the actuation component as well as the diaphragm [4].

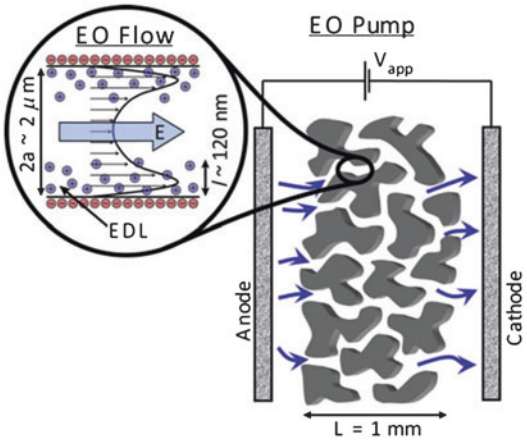
Chemical actuation can be achieved with ion conductive films, which are often a hydrated polymer electrolyte film with surface adhered electrodes. Examples of these films are perfluorosulfonic acid membranes, which are a Teflon-like polymer backbone with sulfonic acid groups attached to the backbone by side chains. When an electric field is applied across this film, sometimes referred to as an ionic polymer–metal composite (IPMC), the mobile ions migrate towards the electrode with opposite polarity. The ions exert a drag force on the surrounding water, which causes swelling on one side of the film and generates a bending moment similar to that in a bimetallic strip. A benefit of IPMCs is the large displacement at low voltages. However, they require humid environments [5].

Nonmechanical Pumps

Nonmechanical pumps, sometimes referred to as continuous or dynamic pumps, rely on applied electric and magnetic fields as well as chemical reactions to drive the flow. Variations of these pumps include electroosmotic (EO), electrohydrodynamic (EHD), magnetohydrodynamic (MHD), and electrolysis pumps. A common characteristic of these pumps is their continuous flow in comparison to the periodic or aperiodic flows of the mechanical pumps with reciprocating diaphragms described in the previous section. Another feature of many nonmechanical pumps is the generation of flow due to body forces generated within the fluid rather than surface forces due to displacements. A benefit of nonmechanical, continuous pumps is that the lack of moving parts can simplify fabrication and enable smaller devices.

Electroosmotic (EO) Micropumps

EO pumps use the electric double layer (EDL) that naturally forms between solid surfaces and liquids to generate flow. EO pumps can be prepared in a number of arrangements, including a single microchannel, arrays of microchannels, and a porous substrate. In the case of pumping aqueous solutions with a porous glass substrate, silanol groups on the surface of the glass spontaneously deprotonate, creating a negative surface charge and a net-positive layer of mobile ions with a generated potential of roughly -60 to -100 mV. The EDL thickness is characterized by the Debye length, λ. The Debye length physically represents the distance from the charged wall where the electric potential energy balances the thermal energy, with lengths ranging between 1 and 100 nm depending on the ionic strength of the fluid being pumped.



Micropumps, Fig. 3 Schematic of EO flow in a porous glass EO micropump. The EO pump is formed by placing electrodes on either side of a porous glass substrate. When saturated with water, the walls of the porous glass generate a negative charge according to the deprotonation of silanol groups on the surface. A diffuse layer of positive charge in the liquid shields the negative wall charge and completes the EDL. When an electric field, *E*, is applied, the diffuse charges experience a columbic force and bulk flow is generated by viscous interaction

As depicted in Fig. 3 (inset), applying an electric field tangential to a porous glass substrate induces a Columbic force on this mobile ion layer. The viscous interaction between ions and water generates a bulk flow. EO pumps with porous substrates can be fabricated from glass-particle-packed fused silica capillaries, porous borosilicate glass, in situ polymerized porous monoliths, and both planar and porous silicon. Ideally, the porous substrate has high porosity for high flow rates and small pores sizes (<2 μm) to achieve reasonable pressure capacities. A thin substrate is desirable so that high electric fields can be generated with low voltages. The maximum flow rate of EO pumps is linearly proportional to the electric field across the substrate, and the maximum pressure is linearly proportional to the applied voltage.

A benefit of EO pumps is their lack of moving parts, high pressure and flow rate capabilities, and low power consumption. Two issues with the common form of EO pumps are the sensitivity of the EO mechanism on the properties of the fluid and the bubbles generated by the electrolysis at the electrodes for applying the electric field. To avoid the issue with the sensitivity to the pumped fluid, pumps have been developed to use the EO flow of a secondary fluid to displace a diaphragm that drives the primary fluid [6]. Other classes of EO pump use high frequency AC electric fields and asymmetric electrodes to induced charge and generate flow [7]. Another form of induced charge EO pumps are those that employ electroosmosis of the second kind (EO-2) arising from ion-permselective porous media [8, 9]. A benefit of EO-2 pumps is that the flow scales quadratically rather than linearly with applied voltage.

Electrolysis Pumps

Electrolysis pumps are a promising class of micropumps because they have a low voltage requirement and high pressure capacity. In these pumps, the flow is actuated by the generation of gas bubbles by electrochemical reactions. These reactions include the hydrogen and oxygen evolution reactions. These pumps typically operate with an applied voltage of 1.5–3 V. The voltage of 1.5 V is a lower bound because it is the thermal neutral voltage for the decomposition of liquid water. Relative to EO pumps, which also employ electrochemical reactions, electrolysis pumps have a lower flow rate per unit of current. The maximum theoretical flow rate per unit of current, which occurs when the hydrogen and oxygen evolution reactions simultaneously generate displacement, is 11 nl/min/μA. Li et al. [10] demonstrated an implantable electrolysis drug delivery pump with a flow rate per current of 6 nl/min/μA. However, electrolysis pumps are less sensitive to changes in the properties of the aqueous solution and the flow rate can be accurately controlled by the current, which dictates the reaction rate and molar gas generation rate.

Electrohydrodynamic (EHD) Pumps

EHD pumps use an electric field to generate charge and then use that applied charge to induce the flow of a dielectric fluid. Several mechanisms exist for generating charge with an electric field. Two reported types of EHD pumps are the induction and injection types. An induction EHD pump generates charge at the interface of immiscible

fluids with different properties, such as different permittivities. Injection EHD pumps use the charge generated electrochemically at the electrode interfaces. Many of the EHD micropumps developed require very high applied voltages (>200 V). A considerable challenge of EHD pumps in practical applications is the need to have liquids with quite low conductivities, restricting them to certain applications.

Magnetohydrodynamic (MHD) Pumps

MHD pumps use the magnetic component of the Lorentz force to generate flow. This Lorentz force is mathematically described as the cross product of the ionic current with the magnetic field. Thus, axial flow can be generated in a rectangular channel by driving ionic current between electrodes on two opposing channel walls while using permanent magnets to generate a magnetic field between the other two opposing walls. Problematic bubbles due to electrolysis of aqueous solutions can arise with DC ionic currents and this has led to the use of alternating currents that suppress gas evolution. Since the force is proportional to current, these pumps perform best with high conductivity fluids, which restricts their application.

Characterizing Micropump Performance

There are several important figures of merit (FOMs) by which to evaluate micropump performance and suitability to a given application. These FOMs are often either related to pumping hydrodynamic capabilities, the pump power consumption, or the dependence of the pump on certain fluid properties (e.g., high conductivity for MHD pumps). The reviews by Laser and Santiago [1] and Nisar et al. [3] provide useful tabulated comparisons of the many FOMs. Here, the FOMs that generally apply to all of the pumps described are reviewed.

The key FOMs for hydrodynamic micropump performance are the device's maximum flow rate and maximum pressure. The maximum flow rate, Q_{max}, is the flow rate measured where there is zero pressure load. In contrast, the maximum pressure, p_{max}, is evaluated when flow is completely prevented and the total flow rate is zero. Such a condition can be achieved by plugging a pump outlet with a pressure transducer. It is important to note that the required maximum flow rate and pressure depend on the application. For example, an important benchmark for the maximum pressure of micropumps is the central vein pressure of approximately 1.4 kPa, which may serve as a minimum pressure for drug delivery pumps.

A metric related to the maximum flow rate, is the self-pumping frequency, which is the ratio of the maximum flow rate to the volume of the pump. This is important when considering the miniaturization of pump technologies, where it is preferred that the pump's flow rate scales proportionally with its size. This scaling depends on the physics of the flow generation and the fabrication technology. High self-pumping frequencies have been reported for piezoelectric pumps, with the highest values being greater than 10 min^{-1}.

Although the maximum pressure and maximum flow rate serve as first as indication of whether a micropump is sufficient for a given application, the pump's performance curve should be evaluated to determine if a pump is truly viable for a given flow rate and pressure load requirement. Micropump performance curves are often presented on a chart of flow rate versus the pressure load. This curve can usually be approximated as linear line between the maximum flow rate and the maximum pressure, entailing that the flow rate is half its maximum value when the pressure load is half its maximum value. These charts usually contain a family of curves that differ by the constant value of voltage, current, or power that is input to the pump. Thus, the required power, voltage, or current can be determined for a given operating condition.

The pump's required power ($P = VI$) and voltage can be equally important when considering the viability of a micropump. For example, with micropumps for implantable drug delivery devices, it is crucial the power is low such that battery replacement is either rarely required

(>10 year) or not necessary. Similar requirements are required for micropumps in portable devices, such as pumps for fluid handling in distributed chemical sensors.

Another important FOM for evaluating pump performance is efficiency, which is a function of flow rate, pressure load, and power; $\eta = Qp = V_{app}I$. Thus, the efficiency value depends on the operating point within the pump performance curve. To avoid ambiguity, Laser and Santiago [1] used an efficiency metric based on linear pump performance curves to compare micropumps, which estimates the maximum efficiency when the pressure load and the flow rate are half of their maximum values. This efficiency metric is defined by the equation $\eta = 0.25Q_{max}p_{max} = V_{app}I$. Due to many of the actuation methods for micropumps being very inefficient, often including thermal actuation by Joule heating, micropumps tend to have very low efficiencies. For example, for most EO pumps, their efficiency is <1 %. However, in many applications, the efficiency is an irrelevant FOM because the power consumption is negligible relative to the power of the total system that includes the micropump. An example is the negligible power draw of an EO pump that used to regulate liquid water within a polymer electrolyte fuel cell [11].

Cross-References

► Micro/Nano Flow Characterization Techniques

References

1. Laser, D.J., Santiago, J.G.: A review of micropumps. J. Micromech. Microeng. **14**, R35 (2004)
2. Woias, P.: Micropumps – past, progress and future prospects. Sens. Actuator B-Chem. **105**, 28 (2005)
3. Nisar, A., Afzulpurkar, N., Mahaisavariya, B., Tuantranont, A.: MEMS-based micropumps in drug delivery and biomedical applications. Sens. Actuator B-Chem. **130**, 917 (2008)
4. Makino, E., Mitsuya, T., Shibata, T.: Fabrication of TiNi shape memory micropump. Sens. Actuator A-Phys. **88**, 256 (2001)
5. Nguyen, T.T., Goo, N.S., Nguyen, V.K., Yoo, Y., Park, S.: Design, fabrication, and experimental characterization of a flap valve IPMC micropump with a flexibly supported diaphragm. Sens. Actuator A-Phys. **141**, 640 (2008)
6. Litster, S., Suss, M.E., Santiago, J.G.: A two-liquid electroosmotic pump using low applied voltage and power. Sens. Actuator A-Phys. **163**, 311 (2010)
7. Squires, T.M.: Induced-charge electrokinetics: fundamental challenges and opportunities. Lab Chip **9**, 2477 (2009)
8. Mishchuk, N.A., Heldal, T., Volden, T., Auerswald, J., Knapp, H.: Micropump based on electroosmosis of the second kind. Electrophoresis **30**, 3499 (2009)
9. Kivanc, F.C., Litster, S.: Pumping with electroosmosis of the second kind in mesoporous skeletons. Sens. Actuator B-Chem. **151**, 394 (2011)
10. Li, P.-Y., Shih, J., Lo, R., Saati, S., Agrawal, R., Humayun, M.S., Tai, Y.-C., Meng, E.: An electrochemical intraocular drug delivery device. Sens. Actuator A-Phys. **143**, 41 (2008)
11. Litster, S., Buie, C.R., Santiago, J.G.: Engineering model for coupling wicks and electroosmotic pumps with proton exchange membrane fuel cells for active water management. Electrochim. Acta **54**, 6223 (2009)

M

Microrobotics

► Micro- and Nanomanipulation for Nanomanufacturing

Microrororobotics

► Nanorobotics

Microscale Fluid Mechanics

► Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering

Microscopic Modeling of Solid-State Quantum Devices

Fabrizio Dolcini[1], Rita Claudia Iotti[2] and Fausto Rossi[1]
[1]Department of Applied Science and Technology, Politecnico di Torino, Torino, Italy
[2]Electrical and Computer Engineering, University of California – Davis Integrated Nanodevices and Nanosystems Lab, Davis, CA, USA

Synonyms

Decoherence and dissipation in nanoelectronic devices; Density-matrix theory; Device modeling; Electronic quantum transport; Nanostructured materials

Definition

Semiconductor-based optoelectronic quantum devices are new-generation nanosystems in which the carrier dynamics shows purely quantum-mechanical or phase-coherence phenomena. The latter result from a proper combination of quantum confinement effects and ultrafast intra- and/or interband optical excitations. A cohesive theoretical treatment of transport versus optical properties is both the key tool for the microscopic investigation of such nanosystems and the essential prerequisite for the modeling and optimization of customized device designs. Diverse simulation strategies, ranging from semiclassical to fully quantum-mechanical approaches, may be devised each with its own validity limits. More specifically, one has to distinguish between ultrafast and steady-state properties of the system, in the linear as well as in the nonlinear response regime, thus identifying the corresponding levels of description. The most severe problem in quantum-device modeling is the microscopic treatment of a quantum system with spatial open boundaries. Any solid state electronic nanodevice is an intrinsically open quantum system, exchanging both energy with the host material and carriers with connected reservoirs. Its out-of-equilibrium behavior is determined by a complex interplay between electronic dissipation and decoherence induced by inelastic processes within the device, as well as by the coupling of the latter to metallic electrodes.

Introduction

Following the path paved by the early proposal by Esaki and Tsu [1], artificially tailored semiconducting heterostructures at the nanoscale now form the leading edge of device technology. The design of modern optoelectronic devices, in fact, heavily exploits the principles of band-gap engineering [2] achieved by confining charge carriers in spatial regions comparable to their de Broglie wavelengths. A proper combination of quantum confinement effects and ultrafast intra- and/or interband optical excitations may give rise to purely quantum-mechanical or phase-coherence phenomena in the carrier dynamics of new-generation optoelectronic devices, and for this reason the latter are often referred to as semiconductor quantum devices [3]. Indeed, present micro/nanoelectronics technology together with modern ultrafast optical spectroscopy push device dimensions and external electrooptical excitations toward space- and timescales where the traditional semiclassical picture mentioned above can no longer be applied and genuine quantum approaches are imperative [3, 4]. However, in spite of the quantum-mechanical nature of electron and photon dynamics in the core region of typical solid-state nanodevices – e.g., superlattices [5] and quantum-dot structures [6] – the overall behavior of such quantum systems is often governed by a highly nontrivial interplay between phase coherence and energy dissipation/decoherence [7, 8], the latter being also strongly influenced by the presence of spatial boundaries [9].

This entry aims to present a cohesive theoretical treatment of transport versus optical properties of semiconductor nanomaterials, essential prerequisite for the microscopic modeling of new-generation optoelectronic quantum devices. We shall discuss different simulation strategies,

ranging from semiclassical to fully quantum-mechanical approaches, together with their validity limits. More specifically, we shall distinguish between ultrafast and steady-state properties of the system, in the linear as well as in the nonlinear response regime, thus identifying the corresponding levels of description. In particular, we shall address the most severe problem in quantum-device modeling: the microscopic treatment of a quantum system with spatial open boundaries. In this context, (i) we shall point out and explain intrinsic limitations of the conventional quantum-device modeling based on the well-known Wigner-function paradigm [9] (providing a definite answer to open questions related to the applicability of the semiclassical spatial-boundary-condition scheme [10]), and (ii) we shall discuss how to overcome such limitations by adopting a nonlocal density-matrix based treatment, in terms of a recently proposed Lindblad-like system-reservoir scattering superoperator [11].

The present essay is organized as follows: In section "Microscopic Modeling of Semiconductor Nanostructures" we shall provide a thorough (intra versus interband) theoretical framework for the microscopic modeling of semiconductor nanostructures based on the density-matrix formalism. In section "From Nanomaterials to Nanodevices" we shall extend such framework to semiconductor nanodevices; in particular, we shall address the most critical task in quantum-device modeling, i.e., the microscopic description of a finite-size quantum nanodevice with open spatial boundaries, corresponding, e.g., to the case of a "small" semiconductor nanodevice embedded into an electric circuit.

Microscopic Modeling of Semiconductor Nanostructures

Physical System

In order to provide a global (intra versus interband) description of the electro-optical properties of semiconductor nanomaterials and nanodevices, we consider a generic carrier gas within a semiconductor nanostructure in the presence of electromagnetic fields. The corresponding Hamiltonian can be schematically written as

$$\hat{H} = \hat{H}^{\circ} + \hat{H}' + \hat{H}_{\mathrm{ern}}. \tag{1}$$

Here, the noninteracting term

$$\hat{H}^{\circ} = \hat{H}_{\mathrm{c}} + \hat{H}_{\mathrm{p}} \tag{2}$$

is the sum of the noninteracting carrier and phonon Hamiltonians, the many-body term

$$\hat{H}' = \sum_{s} \hat{H}'_{s} \tag{3}$$

may include all relevant interaction mechanisms s, namely various carrier-carrier as well as carrier-phonon contributions, and the Hamiltonian $\hat{H}_{\mathrm{em}}$ describes the interaction of the carrier system with the electromagnetic (em) radiation.

More specifically, the noninteracting-carrier Hamiltonian $\hat{H}_{\mathrm{c}}$ describes the scattering-free carrier system within the effective three-dimensional confinement potential of the nanomaterial. By denoting with α the generic noninteracting carrier state and with ϵ_{α} the corresponding energy, the noninteracting-carrier Hamiltonian can be written as

$$\hat{H}_{\mathrm{c}} = \sum_{\alpha} \epsilon_{\alpha} \hat{c}^{\dagger}_{\alpha} \hat{c}_{\alpha}, \tag{4}$$

where the usual second-quantization picture in terms of creation ($\hat{c}^{\dagger}_{\alpha}$) and annihilation ($\hat{c}_{\alpha}$) operators has been employed.

The explicit form of the various many-body contributions $\hat{H}'_{s}$ in Eq. 3 as well as of the electromagnetic-coupling term $\hat{H}_{\mathrm{em}}$ can be found in Rossi and Kuhn [7]; the latter depend on the specific interaction mechanism considered, and typically involve additional degrees of freedom, like phonons, photons, etc. From the point of view of the carrier subsystem, such contributions can be divided into two different classes: single-particle terms – including carrier-phonon and carrier-light interactions – and two-body Coulomb interactions.

Single-particle Description

In order to investigate the electro-optical response of a nanomaterial, it is crucial to study the time evolution of single-particle quantities, such as the total carrier density, mean kinetic energy, charge current, and so on. In general, such quantities are given by a suitable (quantum-plus-statistical) average of a corresponding single-particle operator $\hat{a}$, usually expressed in terms of the global density-matrix operator $\hat{\varrho}$ as [3, 12]

$$\langle a \rangle = \mathrm{tr}\{\hat{a}\hat{\varrho}\}. \tag{5}$$

Recalling that any single-particle electronic operator is of the form

$$\hat{a} = \sum_{\alpha_1 \alpha_2} a_{\alpha_2 \alpha_1} \hat{c}^{\dagger}_{\alpha_2} \hat{c}_{\alpha_1}, \tag{6}$$

the average value in Eq. 5 can simply be written as

$$\langle a \rangle = \sum_{\alpha_1 \alpha_2} a_{\alpha_2 \alpha_1} \rho_{\alpha_1 \alpha_2}, \tag{7}$$

where

$$\rho_{\alpha_1 \alpha_2} = \mathrm{tr}\left\{\hat{c}^{\dagger}_{\alpha_2} \hat{c}_{\alpha_1} \hat{\varrho}\right\} \tag{8}$$

is the single-particle density matrix [7]. Its diagonal elements ($\alpha_1 = \alpha_2$) correspond to the single-particle carrier distribution of the semiclassical Boltzmann theory (see section "Semiclassical Limit") while the off-diagonal contributions ($\alpha_1 \neq \alpha_2$), also referred to as interstate polarizations [7], describe quantum-mechanical phase coherence between the single-particle states α_1 and α_2.

Before continuing, it is worth mentioning that an alternative approach, equivalent to the density-matrix formalism considered here, is given by the nonequilibrium Green's-function technique. An introduction to the theory of nonequilibrium Green's functions with applications to many problems in transport and optics of semiconductors can be found in the book by Hartmut Haug and Antti-Pekka Jauho [13]; by employing – and further developing and extending – such nonequilibrium Green's function formalism, a number of groups have proposed efficient quantum-transport treatments for the study of various meso- and nanoscale structures as well as of corresponding micro- and optoelectronic devices [14].

As discussed below, in addition to this basic subdivision between diagonal (classical) and off-diagonal (quantum) density-matrix elements, the latter are also classified according to their intra- versus interband nature. It is thus clear that the actual set of density-matrix elements $\rho_{\alpha_1 \alpha_2}$ to be considered depends strongly on the particular problem under examination, e.g., ultrafast versus steady-state conditions as well as intra- versus interband carrier dynamics.

In the Schrödinger picture, the time evolution of the average value in Eq. 7 is fully dictated by the time evolution of the single-particle density matrix Eq. 8. Starting from the Heisenberg equations of motion for the creation and destruction operators, it is possible to derive a set of equations of motion for the density-matrix elements $\rho_{\alpha_1 \alpha_2}$ – often referred to as generalized semiconductor Bloch equations [3, 7] – whose general structure is given by:

$$\frac{\mathrm{d}\rho_{\alpha_1 \alpha_2}}{\mathrm{d}t} = \left.\frac{\mathrm{d}\rho_{\alpha_1 \alpha_2}}{\mathrm{d}t}\right|_{\hat{H}^{\circ}} + \left.\frac{\mathrm{d}\rho_{\alpha_1 \alpha_2}}{\mathrm{d}t}\right|_{\hat{H}'} + \left.\frac{\mathrm{d}\rho_{\alpha_1 \alpha_2}}{\mathrm{d}t}\right|_{\hat{H}_{\mathrm{em}}}. \tag{9}$$

The time evolution induced by the noninteracting Hamiltonian $\hat{H}^{\circ}$ in Eq. 1 can be evaluated exactly, yielding

$$\left.\frac{\mathrm{d}\rho_{\alpha_1 \alpha_2}}{\mathrm{d}t}\right|_{\hat{H}^{\circ}} = \frac{\epsilon_{\alpha_1} - \epsilon_{\alpha_2}}{\mathrm{i}\hbar} \rho_{\alpha_1 \alpha_2}. \tag{10}$$

The resulting dynamics is thus diagonal within the $\alpha_1 \alpha_2$ representation, i.e., each density-matrix element evolves independently according to

$$\rho_{\alpha_1 \alpha_2}(t)\big|_{\hat{H}^{\circ}} = \mathrm{e}^{-\mathrm{i}\omega_{\alpha_1 \alpha_2}(t - t_{\mathrm{o}})} \rho_{\alpha_1 \alpha_2}(t_{\mathrm{o}}) \tag{11}$$

with $\hbar\omega_{\alpha_1 \alpha_2} = \epsilon_{\alpha_1} - \epsilon_{\alpha_2}$. In particular, while the diagonal elements are not affected by the free dynamics, the off-diagonal elements (describing

interstate electronic polarization) rotate with a frequency proportional to the energy difference between states α_1 and α_2. As we shall discuss below, this off-diagonal dynamics – absent in the semiclassical (diagonal) limit – gives rise to a variety of phase-coherence phenomena in quantum optoelectronic devices [3, 6].

The time evolution induced by the many-body Hamiltonian in Eq. 3 involves, in general, phonon-assisted as well as higher-order density matrices. Thus, in order to "close" the set of equations of motion (with respect to the single-particle density matrix in Eq. 8) approximations are needed. A detailed discussion of the various approximation schemes – based on a dynamical expansion in powers of the various interaction Hamiltonians $\hat{H}'_s$ – can be found in Rossi and Kuhn [7]. In particular, the "mean-field" approximation together with the Markov limit allows one to derive a set of closed equations of motion still local in time. For a number of interaction mechanisms, including carrier-phonon as well as carrier-carrier scattering, their second-order contributions to the time evolution can be cast into the general form:

$$\left.\frac{\mathrm{d}\rho_{\alpha_1\alpha_2}}{\mathrm{d}t}\right|_{\hat{H}'} = \Gamma(\rho)_{\alpha_1\alpha_2}. \tag{12}$$

Although $\Gamma(\rho)$ is, in general, a nonlinear function of ρ, in the low-carrier-density limit (i.e., $|\rho_{\alpha_1\alpha_2}| \ll 1$), carrier-carrier scattering as well as Pauli-blocking phenomena play a minor role, and the many-body contribution in Eq. 12 can be written as

$$\left.\frac{\mathrm{d}\rho_{\alpha_1\alpha_2}}{\mathrm{d}t}\right|_{\hat{H}'} = \sum_{\alpha'_1\alpha'_2} \Gamma_{\alpha_1\alpha_2,\alpha'_1\alpha'_2}\rho_{\alpha'_1\alpha'_2}, \tag{13}$$

where $\Gamma_{\alpha_1\alpha_2,\alpha'_1\alpha'_2}$ are the matrix elements of the so-called scattering superoperator Γ [3]: their real parts describe incoherent phenomena, like energy-dissipation and decoherence processes, while their imaginary parts lead to modifications of the free coherent evolution in Eq. 11 through nontrivial energy-renormalisation effects.

The explicit form of the time evolution induced by the electromagnetic-radiation coupling Hamiltonian $\hat{H}_{\mathrm{em}}$ heavily depends on the adopted level of description, which, in turn, is dictated by the problem under examination: in order to investigate high-field phenomena and/or ultrafast nonlinear optical excitations, this interaction can be treated exactly in terms of purely classical electromagnetic fields; in contrast, in order to investigate quantum-optics phenomena – like photoluminescence experiments – additional approximation schemes are needed [7].

By combining Eqs. 9, 10 and 13, the explicit form of the generalized semiconductor Bloch equations turns out to be

$$\frac{\mathrm{d}\rho_{\alpha_1\alpha_2}}{\mathrm{d}t} = \sum_{\alpha'_1\alpha'_2} L_{\alpha_1\alpha_2,\alpha'_1\alpha'_2}\rho_{\alpha'_1\alpha'_2} + \left.\frac{\mathrm{d}\rho_{\alpha_1\alpha_2}}{\mathrm{d}t}\right|_{\hat{H}_{\mathrm{em}}}, \tag{14}$$

where the effective Liouville superoperator

$$L_{\alpha_1\alpha_2,\alpha'_1\alpha'_2} = \frac{\epsilon_{\alpha_1} - \epsilon_{\alpha_2}}{\mathrm{i}\hbar}\delta_{\alpha_1\alpha_2,\alpha'_1\alpha'_2} + \Gamma_{\alpha_1\alpha_2,\alpha'_1\alpha'_2} \tag{15}$$

is the sum of two terms: free single-particle evolution plus scattering-induced energy-dissipation/decoherence dynamics.

The above single-particle quantum equation applies to a variety of physical problems [3], ranging from quantum-transport phenomena to ultrafast electro-optical processes. However, it is worth emphasizing that the degree of accuracy of such density-matrix formalism is intimately related to the choice of the scattering superoperator Γ. Indeed, oversimplified phenomenological treatments can lead, for instance, to a violation of the positive-definite character of the density matrix [15], a crucial prerequisite of any quantum-mechanical time evolution. This issue is particularly important for solid-state devices, which are characterized by a complex many-electron quantum evolution. A general prescription ensuring the positive-definite character of the density matrix consists in expressing dissipation versus decoherence phenomena in terms of suitable Lindblad-type superoperators, containing a few crucial system-environment coupling parameters. This problem has been discussed in Taj

et al. [16], where a novel Markov procedure leading to positive-definite superoperators has been proposed. More specifically, thanks to a proper reformulation of the Markov limit, for each single-particle interaction mechanism in the low-density limit one is able to perform a fully microscopic derivation of a corresponding Lindblad scattering superoperator, thereby preserving the positivity of the single-particle density matrix. The Lindblad-type approach in Taj et al. [16] has been recently extended to the high-density regime [17].

Global Versus Partial Treatments

The above-discussed general or global density-matrix formalism does not distinguish between intra- and interband excitations, i.e., it is not grounded on an explicit separation between valence and conduction bands. Indeed, such a formulation applies to bulk as well as to nanostructured materials: for the case of bulk crystals one has $\alpha \equiv \mathbf{k}v$, where $\mathbf{k}$ denotes the three-dimensional carrier wavevector and v is the crystal band index; for the case of a semiconductor nanostructure, by employing the conventional envelope-function picture [3], the generic label α of the noninteracting electronic states is given again by the band index v, whereas the three-dimensional wavevector $\mathbf{k}$ is replaced by a low-dimensional wavevector $\mathbf{k}_{\|}$ plus a subband index n, i.e., $\alpha \equiv \mathbf{k}_{\|}nv$.

Recalling that for a semiconductor nanostructure an optical excitation – according to its spectral range – may induce transitions between different electronic bands (interband transitions) or within the same band (intraband transitions), it follows that the relevant set of density-matrix elements $\rho_{\alpha\alpha'}$ to be considered depends crucially on the excitation spectral region. More specifically, in addition to the state populations corresponding to the diagonal density-matrix elements $\left(\rho_{\mathbf{k}_{\|}nv,\mathbf{k}_{\|}nv}\right)$, for intraband ($v = v'$) excitations the only relevant quantities are given by intraband polarizations corresponding to the off-diagonal density-matrix elements $\rho_{\mathbf{k}_{\|}nv,\mathbf{k}_{\|}n'v}$; the latter are also referred to as intersubband ($n \to n'$) polarizations. In contrast, for interband ($v \neq v'$) excitations the only relevant quantities are given by interband polarizations corresponding to the off-diagonal density-matrix elements $\rho_{\mathbf{k}_{\|}nv,\mathbf{k}_{\|}n'v'}$; the latter describe interband optical transitions connecting the generic subband n in band v to the generic subband n' in band v'. As observed above, moving from bulk to nanostructures the three-dimensional wavevector $\mathbf{k}$ is replaced by the so-called parallel-subspace wavevector $\mathbf{k}_{\|}$, while – in addition to the band index v – the subband index n should be considered as well.

The intra- versus interband classification scheme just introduced suggests to separate the global set of states α according to the well-established electron–hole picture [3, 7]. To this end, we shall denote by $i \equiv \mathbf{k}_{\|}n_e v_e$ and $j = -\mathbf{k}_{\|}n_h v_h$, respectively, electron and hole states, and with $\hat{c}_i^{\dagger}$ $(\hat{c}_i)$ and $\hat{d}_j^{\dagger}$ $(\hat{d}_j)$ the corresponding creation (destruction) operators. Within such electron–hole picture, the original/global single-particle density matrix in Eq. 8 is replaced by the following electron and hole density matrices,

$$\mathrm{f}^{e}_{i_1 i_2} = \mathrm{tr}\left\{\hat{c}^{\dagger}_{i_2}\hat{c}_{i_1}\hat{\varrho}\right\}, \quad \mathrm{f}^{h}_{j_1 j_2} = \mathrm{tr}\left\{\hat{d}^{\dagger}_{j_2}\hat{d}_{j_1}\hat{\varrho}\right\}, \quad (16)$$

together with the following interband-polarization matrix

$$p_{i_1 j_1} = \mathrm{tr}\left\{\hat{d}_{j_1}\hat{c}_{i_1}\hat{\varrho}\right\} \quad (17)$$

and the density-matrix equation in 14 is transformed into the following set of coupled kinetic equations:

$$\begin{aligned}
\frac{\mathrm{d}\mathrm{f}^{e}_{i_1 i_2}}{\mathrm{d}t} &= \sum_{i'_1 i'_2} L^{e}_{i_1 i_2, i'_1 i'_2}\mathrm{f}^{e}_{i'_1 i'_2} + \left.\frac{\mathrm{d}\mathrm{f}^{e}_{i_1 i_2}}{\mathrm{d}t}\right|_{\hat{H}_{\mathrm{em}}} \\
\frac{\mathrm{d}\mathrm{f}^{h}_{j_1 j_2}}{\mathrm{d}t} &= \sum_{j'_1 j'_2} L^{h}_{j_1 j_2, j'_1 j'_2}\mathrm{f}^{h}_{j'_1 j'_2} + \left.\frac{\mathrm{d}\mathrm{f}^{h}_{j'_1 j'_2}}{\mathrm{d}t}\right|_{\hat{H}_{\mathrm{em}}} \\
\frac{\mathrm{d}p_{i_1 j_1}}{\mathrm{d}t} &= \sum_{i'_1 j'_1} L^{p}_{i_1 j_1, i'_1 j'_1} p_{i'_1 j'_1} + \left.\frac{\mathrm{d}p_{i'_1 j'_1}}{\mathrm{d}t}\right|_{\hat{H}_{\mathrm{em}}},
\end{aligned} \quad (18)$$

where L^e, L^h, and L^p are again effective Liouville superoperators acting separately on electrons, holes, and interband polarizations; similar to the general structure in Eq. 15, each of these superoperators is again the sum of two

terms: free single-particle rotation plus scattering-induced energy-dissipation/decoherence dynamics.

As anticipated, in order to describe the ultrafast nonlinear response of bulk as well as nanostructured materials, the nontrivial coupling between electron/hole density matrices and corresponding interband polarizations is expected to play a major role, and for this reason a solution of the above coupled set of kinetic equations (equivalent to the global density-matrix equation (14)) is imperative [3, 7]. In contrast, for the case of an interband optical excitation that varies slowly compared to the typical carrier free-dynamics as well as decoherence time-scales, the interband polarization fields $p_{i_1 j_i}$ may be adiabatically eliminated [7], and the original set in Eq. 18 reduces to:

$$\begin{aligned} \frac{\mathrm{d}f^{e}_{i_1 i_2}}{\mathrm{d}t} &= \sum_{i'_1 i'_2} L^{e}_{i_1 i_2, i'_1 i'_2} f^{e}_{i'_1 i'_2} + G^{e}_{i_1 i_2} \\ \frac{\mathrm{d}f^{h}_{j_1 j_2}}{\mathrm{d}t} &= \sum_{j'_1 j'_2} L^{h}_{j_1 j_2, j'_1 j'_2} f^{h}_{j'_1 j'_2} + G^{h}_{j_1 j_2}. \end{aligned} \tag{19}$$

Equation 19 thus express the effect of the external electromagnetic excitation simply via generalized generation/recombination rates $G^{e/h}$, describing the effect on the electron–hole system of both intra- and interband optical excitations. It follows that within such approximation scheme these generation/recombination rates act as input parameters determined by the shape of the external electromagnetic excitation, and electron and hole subsystems are not coupled anymore. In particular, in the presence of an intraband (i.e., infrared) excitation only, the two density-matrix equations in 19 can be solved independently.

The main conclusion of the analysis presented so far is that the level of description used to simulate the electro-optical response of a semiconductor nanostructure depends strongly on the nature of the external electromagnetic excitation: for the case of ultrafast nonlinear interband excitations one deals with a highly nontrivial coupling between carrier populations and interband polarizations, often dominated by excitonic effects [3, 7]. In contrast, for slowly-varying interband illumination or for intraband excitations only, the electron and hole subsystems can be simulated independently. Moreover, for specific operation conditions of the nanodevice, phase-coherence phenomena – described by off-diagonal density-matrix elements – can be safely neglected, and the semiclassical or Boltzmann picture is recovered (see below).

Semiclassical Limit

For various optoelectronic devices operating in steady-state or quasi-steady-state conditions, electronic phase coherence is expected to play a minor role, and for this reason one is allowed to employ the standard semiclassical picture. A relevant example of such a semiclassical modeling is the microscopic simulation of infrared quantum-cascade lasers [8] and photodetectors.

In any semiclassical treatment the off-diagonal elements of the density matrix are neglected, i.e., the density-matrix elements acquire the form

$$\rho_{\alpha_1 \alpha_2} = f_{\alpha_1} \delta_{\alpha_1 \alpha_2}. \tag{20}$$

In this limit (i) the coherent free evolution in Eq. 10 vanishes, and (ii) the many-body contribution in Eq. 13 reduces to

$$\left.\frac{\mathrm{d}f_\alpha}{\mathrm{d}t}\right|_{\hat{H}'} = \sum_{\alpha'} [P_{\alpha\alpha'} f_{\alpha'} - P_{\alpha'\alpha} f_\alpha], \tag{21}$$

where $P_{\alpha\alpha'}$ is the total (i.e., summed over all relevant interaction mechanisms) scattering rate from state α' to state α. It follows that the semiclassical limit of the density-matrix equation (14) is given by the following Boltzmann-like equation

$$\frac{\mathrm{d}f_\alpha}{\mathrm{d}t} = \sum_{\alpha'} [P_{\alpha\alpha'} f_{\alpha'} - P_{\alpha'\alpha} f_\alpha] + \left.\frac{\mathrm{d}f_\alpha}{\mathrm{d}t}\right|_{\hat{H}_{\mathrm{em}}}. \tag{22}$$

Let us finally reformulate the semiclassical limit in terms of the electron–hole picture previously introduced (see Eqs. 16 and 17). By inserting into the effective set of density-matrix equations (19) the corresponding semiclassical-limit prescription, i.e.,

$$\mathrm{f}^{e}_{i_1 i_2} = f^{e}_{i_1}\delta_{i_1 i_2},\ \mathrm{f}^{h}_{j_1 j_2} = f^{h}_{j_1}\delta_{j_1 j_2}, \tag{23}$$

one obtains the following set of Boltzmann-like equations for electron and hole carrier distributions:

$$\begin{aligned}\frac{\mathrm{d}f^{e}_{i}}{\mathrm{d}t} &= \sum_{i'}\left[P^{e}_{ii'}f^{e}_{i'} - P^{e}_{i'i}f^{e}_{i}\right] + G^{e}_{ii}\\ \frac{\mathrm{d}f^{h}_{j}}{\mathrm{d}t} &= \sum_{j'}\left[P^{h}_{jj'}f^{h}_{j'} - P^{h}_{j'j}f^{h}_{j}\right] + G^{h}_{jj}.\end{aligned} \tag{24}$$

Compared to the quantum-mechanical equation in 19, here the effect of the external electromagnetic excitation is expressed via conventional carrier generation/recombination rates, given by the diagonal elements of the generalized (i.e., off-diagonal) generation/recombination rates $G^{e/h}$ in Eq. 19.

From Nanomaterials to Nanodevices

The density-matrix treatment discussed so far applies to "extended systems," i.e., systems extending over the whole coordinate space; in contrast, a semiconductor device [3, 8] is typically a "localized system," i.e., a portion of material characterized by a well precise volume and spatial boundaries acting as electric contacts.

The impact of such spatially open boundaries depends crucially on the actual size/length of the nanodevice under examination. As we shall see, for periodically repeated nanostructures, finite-size effects can be safely neglected, and one can still employ the density-matrix description of section "Single-particle Description" as well as its semiclassical version in section "Semiclassical Limit". In contrast, when the effective size of the nanodevice is comparable to the relevant system space scales, e.g., scattering free path and/or coherence length, the microscopic modeling of such quantum devices constitutes a formidable task. Indeed, while in the semiclassical limit – typically applicable to steady-state operation regimes in the presence of a strong dissipation/decoherence dynamics – such finite-size effects can be properly incorporated via effective injection/loss models [3], the investigation of genuine quantum phenomena in finite-size nanodevices is still an open problem.

Modeling of Periodically-repeated Nanodevices

State-of-the-art semiconductor-based optoelectronic devices at the nanoscale are often based on periodically repeated core regions, for which spatial-boundary effects may be safely neglected. This strategy is exploited, for example, to increase either the photocurrent in infrared detectors or the emitted power in coherent-light sources like quantum-cascade lasers [2]. This feature turns out to be crucial also in the proper modeling of these optoelectronic devices [8], since it allows one to consistently "close the circuit" without the need for any phenomenological parameter accounting for the coupling of the device with its electric contacts. While for small/moderate applied fields it is possible to exploit the fully three-dimensional superlattice periodicity of the nanostructure under examination, in the presence of a significant applied bias the latter cannot be treated via the conventional drift term of the Boltzmann equation [18]. In this regime, the applied field needs to be added to the nanostructure potential profile; as a result, the total potential acting on the carriers is not periodic anymore, i.e., the usual superlattice periodicity is lost. Nevertheless, by assuming a uniform bias drop, the effective field profile across the nanostructure is still periodic, which implies that also in the presence of a strong homogeneous external field all relevant physical properties (e.g., the semiclassical distribution function f_α) should display the periodicity of the nanostructure under investigation.

More specifically, by neglecting finite-size effects, the theoretical description of the whole (periodically repeated) nanodevice under examination can still be grounded on the global density-matrix equation in 14 as well as on its semiclassical version in Eq. 22, where the generic carrier state is specified again by its low-dimensional wavevector plus its subband and band indices, i.e., $\alpha \equiv \mathbf{k}_{\|}nv$. Since such nanostructures are obtained as periodic repetition of a given building block usually referred to as "device stage," it

follows that in the presence of an applied uniform electric field the Wannier-Stark-like states localized in the different stages are physically equivalent. Based on such translational symmetry, the global subband index n – running over the whole nanostructure – may be conveniently replaced by a stage index ℓ plus an intrastage subband index $\overline{n}$, i.e., $n \equiv \overline{n}\ell$.

Employing the compact notation $\alpha = \overline{\alpha}\ell$ (with $\overline{\alpha} \equiv \mathbf{k}_{||}\overline{n}v$) and recalling that the carrier distribution function is stage-independent, the global Boltzmann-like equation (22) can be transformed into the following single-stage version

$$\frac{df_{\overline{\alpha}}}{dt} = \sum_{\overline{\alpha}'} \left[\overline{P}_{\overline{\alpha}\overline{\alpha}'} f_{\overline{\alpha}'} - \overline{P}_{\overline{\alpha}'\overline{\alpha}} f_{\overline{\alpha}}\right] + \left.\frac{df_{\overline{\alpha}}}{dt}\right|_{\hat{H}_{\mathrm{em}}} \tag{25}$$

with

$$\overline{P}_{\overline{\alpha}\overline{\alpha}'} = \sum_{\ell'} P_{\overline{\alpha}\ell,\overline{\alpha}'\ell'}. \tag{26}$$

This represents a strong simplification of the problem, since within the present single-stage scheme the kinetic description of the device is limited, e.g., to the $\ell = 0$ stage only, and thus to a limited number of subbands $\overline{n}$ (typically of the order of ten [8]); moreover, in view of the spatial periodicity, all inter-stage scattering rates depend on the stage difference $\ell' - \ell$ only, and for this reason the single-stage rates in Eq. 26 are ℓ-independent as well.

Within such single-stage picture each device stage actually behaves as an injector/collector of carriers for the following/previous one; this suggests an alternative rearrangement of the global Boltzmann-like equation (22) by splitting the scattering term into intra-stage ($\ell = \ell'$) and inter-stage ($\ell \neq \ell'$) contributions; as a result one obtains

$$\frac{df_{\overline{\alpha}}}{dt} = \left.\frac{df_{\overline{\alpha}}}{dt}\right|_{\mathrm{intra}} + \left.\frac{df_{\overline{\alpha}}}{dt}\right|_{\mathrm{inter}} \tag{27}$$

with

$$\left.\frac{df_{\overline{\alpha}}}{dt}\right|_{\mathrm{inter}} = \overline{S}_{\overline{\alpha}} - \overline{\Gamma}_{\overline{\alpha}} f_{\overline{\alpha}}. \tag{28}$$

Here, the first term describes intra-stage scattering processes only, while the second contribution has been intentionally expressed in terms of a source/injection term

$$\overline{S}_{\overline{\alpha}} = \sum_{\ell \neq 0} \sum_{\overline{\alpha}'} P_{\overline{\alpha}0,\overline{\alpha}'\ell} f_{\overline{\alpha}'} \tag{29}$$

and of a loss/escape rate

$$\overline{\Gamma}_{\overline{\alpha}} = \sum_{\ell \neq 0} \sum_{\overline{\alpha}'} P_{\overline{\alpha}'\ell,\overline{\alpha}0}. \tag{30}$$

This result may be regarded as a formal derivation/justification of the partially phenomenological injection/escape loss model [3] usually employed in the semiclassical simulation of finite-size quantum devices (see below).

Such single-stage picture – introduced above for the case of the Boltzmann-like Eq. 22 – applies to the density-matrix equation 14 as well. More specifically, by neglecting any inter-stage ($\ell \neq \ell'$) phase coherence (i.e., the electron coherence length is significantly shorter than the period of the nanostructure), the global quantum equation (14) can be cast into the following single-stage version:

$$\frac{\mathrm{d}\rho_{\overline{\alpha}_1\overline{\alpha}_2}}{\mathrm{d}t} = \sum_{\overline{\alpha}_1'\overline{\alpha}_2'} \overline{L}_{\overline{\alpha}_1\overline{\alpha}_2,\overline{\alpha}_1'\overline{\alpha}_2'} \rho_{\overline{\alpha}_1'\overline{\alpha}_2'} + \left.\frac{\mathrm{d}\rho_{\overline{\alpha}_1\overline{\alpha}_2}}{\mathrm{d}t}\right|_{\hat{H}_{\mathrm{em}}}, \tag{31}$$

with

$$\overline{L}_{\overline{\alpha}_1\overline{\alpha}_2,\overline{\alpha}_1'\overline{\alpha}_2'} = \frac{\epsilon_{\overline{\alpha}_1} - \epsilon_{\overline{\alpha}_2}}{\mathrm{i}\hbar} \delta_{\overline{\alpha}_1\overline{\alpha}_2,\overline{\alpha}_1'\overline{\alpha}_2'} + \overline{\Gamma}_{\overline{\alpha}_1\overline{\alpha}_2,\overline{\alpha}_1'\overline{\alpha}_2'}, \tag{32}$$

where

$$\overline{\Gamma}_{\overline{\alpha}_1\overline{\alpha}_2,\overline{\alpha}_1'\overline{\alpha}_2'} = \sum_{\ell} \Gamma_{\overline{\alpha}_1 0,\overline{\alpha}_2 0;\overline{\alpha}_1'\ell,\overline{\alpha}_2'\ell} \tag{33}$$

is the reduced or single-stage version of the scattering superoperator in Eq. 13. The above single-stage density-matrix equation has been successfully employed for the investigation of ultrafast phase-coherence versus energy-dissipation/decoherence in quantum-cascade lasers [8].

Modeling of Finite-size Quantum Nanodevices

The theoretical modeling of finite-size nanodevices is by far the most difficult task. Indeed, in sharp contrast to the case of the periodically repeated nanostructures considered in the above subsection, the description of finite-size electronic and optoelectronic devices cannot be grounded on the extended-system formalism presented in section "Microscopic Modeling of Semiconductor Nanostructures".

Once again, for the case of nanodevices operating in steady-state or quasi-steady-state conditions and in the presence of significant dissipation/decoherence phenomena, a semiclassical treatment of the problem is fully satisfactory. To this end, based on the rigorous injection/escape equation Eq. 28, the effect of the external charge reservoirs on the device active region (see Fig. 1) is usually described via a partially phenomenological injection/escape model of the form:

$$\left.\frac{\mathrm{d}f_\alpha}{\mathrm{d}t}\right|_{\mathrm{res}} = S_\alpha - \Gamma_\alpha f_\alpha = -\Gamma_\alpha\left(f_\alpha - f_\alpha^\circ\right). \qquad (34)$$

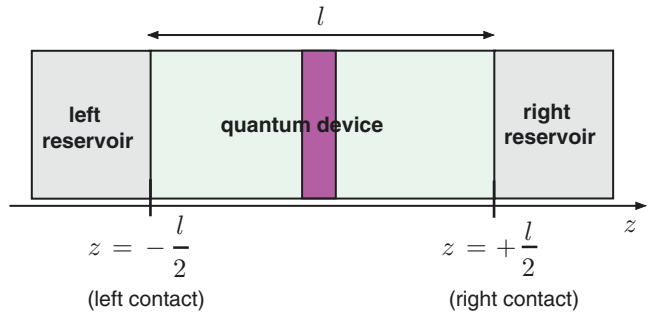
Such effective spatial-boundary term may be regarded again as a source/injection contribution $S_\alpha = \Gamma_\alpha f_\alpha^\circ$ describing a sort of quasithermal injection rate into state α (dictated by the reservoir carrier distribution f_α°) minus a loss term $\Gamma_\alpha f_\alpha$ describing the escape of carriers from state α according to the escape rate Γ_α.

For genuine semiconductor quantum devices such a semiclassical picture is not applicable, and a fully quantum-mechanical treatment is imperative. A finite-size nanodevice is a "localized system," i.e., a portion of material characterized by a well defined volume and by spatial boundaries acting as electric contacts to external charge reservoirs, as sketched in Fig. 1; here, z denotes the transport direction, l is the longitudinal length of the device, and the electric contacts are located at $z = -l/2$ and $z = +l/2$. The modeling of such a nanostructured device thus represents an intrinsically space-dependent problem, so that a real-space quantum treatment accounting for the presence of quite different spatial regions becomes mandatory.

To this purpose, the Wigner-function formalism [9] is a widely adopted framework. Within this formalism, the statistical quantum state of the electronic subsystem is fully described in terms of the Wigner function; for the case of a one-dimensional electronic system (with real-space coordinate z and wavevector k) the latter is defined over the conventional phase-space z, k as the Weyl-Wigner transform of the single-particle density-matrix [3, 12]:

$$f_W(z,k) = \sum_{\alpha_1\alpha_2} W_{\alpha_2\alpha_1}(z,k)\rho_{\alpha_1\alpha_2} \qquad (35)$$

with



Microscopic Modeling of Solid-State Quantum Devices, Fig. 1 Schematic representation of a typical semiconductor-based quantum device as an open system connected to two external charge reservoirs. Here, the distance between the interfaces is l, and z is the longitudinal transport direction

$$W_{\alpha_2\alpha_1}(z,k)=\int dz'\phi_{\alpha_1}\left(z+\frac{z'}{2}\right)e^{-ikz'}\phi^*_{\alpha_2}\left(z-\frac{z'}{2}\right), \tag{36}$$

where $\phi_\alpha(z) = \langle z|\alpha\rangle$ denotes the real-space wavefunction corresponding to the noninteracting carrier state α.

By neglecting all many-body interactions in Eq. 3 as well as the electromagnetic-radiation coupling in Eq. 1, it is possible to derive a closed equation of motion for the Wigner function in Eq. 35. More specifically, describing the electronic band structure via the conventional envelope-function theory [3] in terms of a single effective mass m^* and denoting by $V(z)$ the effective potential profile of the semiconductor nanostructure, the resulting equation of motion – obtained applying the Weyl-Wigner transform Eq. 35 to the noninteracting evolution in Eq. 10 – is given by

$$\begin{aligned}&\frac{\partial f_W(z,k)}{\partial t}+\upsilon(k)\frac{\partial f_W(z,k)}{\partial z}\\&+\int dk'\mathcal{V}(z,k-k')f_W(z,k')\\&=0,\end{aligned} \tag{37}$$

where

$$\mathcal{V}(z,k'')=\frac{i}{\hbar}\int dz'\frac{e^{-ik''z'}}{2\pi}\left[V\left(z+\frac{z'}{2}\right)-V\left(z-\frac{z'}{2}\right)\right] \tag{38}$$

is a nonlocal potential superoperator corresponding to the device profile $V(z)$.

Equation 37, also referred to as the Wigner transport equation, describes the time evolution of the Wigner function in the coherent limit, i.e., in the absence of energy-dissipation/decoherence processes. In steady-state conditions ($\partial f_W(z, k)/\partial t = 0$) the latter reduces to

$$\upsilon(k)\frac{\partial f_W(z,k)}{\partial z}=-\int dk'\mathcal{V}(z,k-k')f_W(z,k'). \tag{39}$$

In order to account for many-body dissipation versus decoherence processes, the simplest model is the well-known relaxation-time approximation; the latter amounts to replace the general scattering-superoperator dynamics in Eq. 13 with a relaxation-time term of the form:

$$\left.\frac{d\rho_{\alpha_1\alpha_2}}{dt}\right|_{\hat{H}'}=-\overline{\Gamma}\left(\rho_{\alpha_1\alpha_2}-\rho^\circ_{\alpha_1\alpha_2}\right), \tag{40}$$

where $\rho^\circ_{\alpha_1\alpha_2}$ are the matrix elements of the equilibrium density-matrix operator

$$\hat{\rho}^\circ=\sum_k|k\rangle f^\circ(\epsilon_k)\langle k| \tag{41}$$

(expressed via the corresponding Fermi-Dirac distribution f°) and $\overline{\Gamma}$ is the relaxation rate. By applying to the relaxation term Eq. 40 the Weyl-Wigner transform in Eq. 35, it is easy to get

$$\begin{aligned}\upsilon(k)\frac{\partial f_W(z,k)}{\partial z}=&\int dk'\mathcal{V}(z,k-k')f_W(z,k')\\&-\overline{\Gamma}\left(f_W(z,k)-f^\circ_W(k)\right),\end{aligned} \tag{42}$$

where the equilibrium Wigner function coincides with the Fermi-Dirac distribution f°, i.e., $f^\circ_w(z,k) = f^\circ(\epsilon_k)$. The relaxation rate $\overline{\Gamma}$ can be regarded as a sort of effective or average inverse coherence time, and is mainly determined by carrier-phonon as well as carrier-carrier scattering.

The inclusion of incoherent processes via the above relaxation-time framework implies the appearance of an additional contribution to the steady-state Wigner equation (39), leading to the generalized Wigner equation

$$\begin{aligned}\upsilon(k)\frac{\partial f_W(z,k)}{\partial z}=&-\int dk'\mathcal{V}(z,k-k')f_W(z,k')\\&-\overline{\Gamma}\left(f_W(z,k)-f^\circ_W(k)\right),\end{aligned} \tag{43}$$

where $f^\circ_w(k) \equiv f^\circ(\epsilon_k)$ denotes the equilibrium Fermi-Dirac distribution induced by the host material.

By applying the Weyl-Wigner transform (35) to the general average-value prescription in Eq. 5, one obtains

$$\langle a \rangle = \frac{1}{2\pi} \int dz \int dk a_W(z,k) f_W(z,k) \tag{44}$$

with

$$a_W(z,k) = \sum_{\alpha_1 \alpha_2} W_{\alpha_1 \alpha_2}(z,k) a_{\alpha_1 \alpha_2}. \tag{45}$$

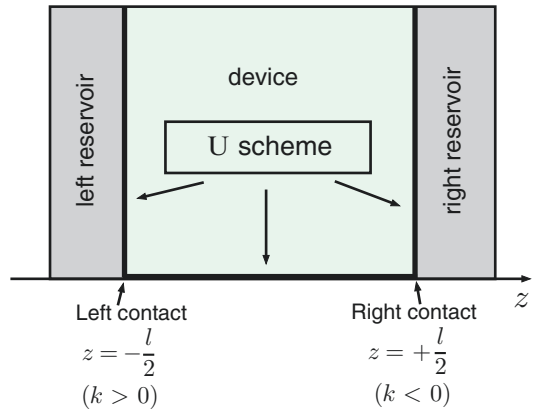
In particular, the average charge and current densities may be expressed via the Wigner function as:

$$n(z) = \frac{1}{2\pi} \int dk f_W(z,k), \tag{46}$$

$$J(z) = \frac{1}{2\pi} \int dk v(k) f_W(z,k). \tag{47}$$

Based on the steady-state Wigner equation (39) as well as on its generalized version in Eq. 43, various approaches for the study of quantum-transport phenomena in semiconductor nanomaterials and nanodevices have been proposed In particular, starting from the pioneering work by Frensley [9], quantum-transport simulations based on a direct numerical solution of the Wigner equation have been performed via finite-difference approaches, by imposing on the Wigner function the standard U boundary-condition scheme (see Fig. 2) of the semiclassical device modeling.

Motivated by a few unphysical results obtained via the generalized semiconductor Bloch equations mentioned above, a more recent study [19] has however shown that the application of the conventional space boundary condition scheme to the Wigner transport equation may lead to negative spatial charge densities, unambiguous proof of the failure of such classical-like Wigner-function treatment. Indeed, the conventional boundary-condition scheme adopted for the solution of the Wigner equation exhibits intrinsic limitations, whose impact may lead to totally unphysical results, especially in the coherent-transport regime. In the following, we shall first review and discuss such limitations [10] and then present a sounder approach based on a nonlocal density-matrix framework [11].

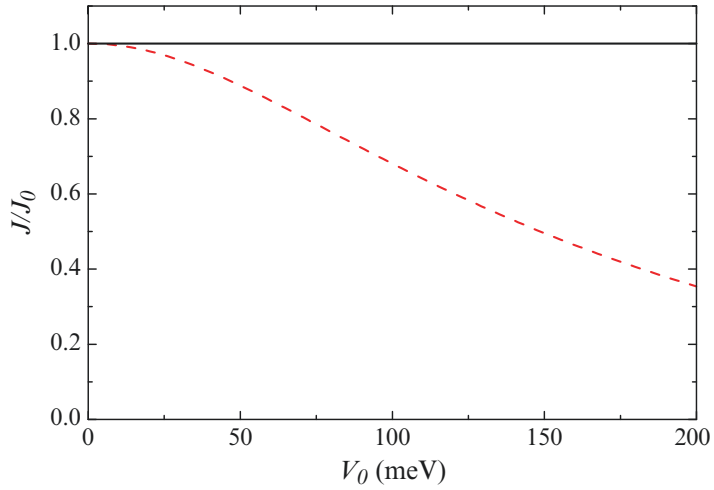


Microscopic Modeling of Solid-State Quantum Devices, Fig. 2 The U boundary condition scheme adopted in semiclassical device modeling [3] for a one-dimensional problem. The value of the Wigner function $f_W(z, k)$ is specified at the boundaries of the active region, i.e., $f_W(-l/2, k > 0)$ and $f_W(+l/2, k < 0)$ are fixed by the incoming carrier distribution function

Failure of the Conventional Boundary-Condition Scheme

In order to point out the most severe limitations of the conventional, i.e., classical-like, boundary-condition scheme applied to the Wigner transport equation, it is worth summarizing its main features.

In terms of the variable z, the Wigner transport equation (39) [as well as its generalized version in Eq. 43] is a first-order differential equation. In this respect, it is thus similar to the steady-state version of the Boltzmann equation [3]. Based on this analogy, outlined in the pioneering work by Frensley [9], in the last decades several quantum-transport problems have been treated by following a semiclassical approach, i.e. by applying to the Wigner transport equation the simulation strategies commonly adopted for the Boltzmann equation. Indeed most of these studies [9] are based on a numerical solution of Eq. 43, where one imposes on $f_W(z,k)$ the inflow or U spatial boundary-condition scheme of the semiclassical device modeling. The latter, depicted in Fig. 2, consists in requiring that the inflowing Wigner function acquires some fixed values at the two contacts $z = \pm l/2$, and that these values are determined by classical-like quasi-thermal distributions of carriers incoming from the two



Microscopic Modeling of Solid-State Quantum Devices, Fig. 3 Steady-state current density J across the rectangular barrier Eq. 50 (device length $l = 40$ nm and barrier width $a = 2$ nm) as a function of the barrier height V_0 (in units of the barrier-free ($V_0 = 0$) value J_0) in the coherent limit $\left(\overline{\Gamma} \to 0\right)$ and in the presence of a room-temperature carrier injection only from the left (chemical potential $\mu = 4k_BT$). The result of the Wigner-function calculation (*solid curve*) is independent of the barrier height V_0. This unphysical behavior is compared with the analytical current density predicted by the Landauer-Büttiker theory (*dashed curve*), which decreases as a function of V_0. Here, the Wigner-function result has been obtained as a numerical solution of the integral Wigner equation in (49) via a standard phase-space discretization scheme (in terms of a 120 × 120 uniform grid) (see text)

reservoirs. Explicitly, the values $f_W(-l/2, k)$ are specified for carriers incoming from the left reservoir ($k > 0$) and the values $f_W(+l/2, k)$ are specified for electrons incoming from the right reservoir ($k < 0$). In a compact notation, introducing $z^b(k) = \text{sign}(k)l/2$ (where sign is the sign function), the Wigner transport equation (39) is thus equipped with the k-dependent boundary condition

$$f^b(k) \equiv f_W\left(z^b(k), k\right). \tag{48}$$

It is thus useful to rewrite the Wigner problem Eqs. 43, 44, 45, 46, 47, and 48 in an equivalent integral form [19]:

$$\upsilon(k) f_W(z,k) = \upsilon(k) f^b(k) - \int_{z^b(k)}^{z} dz' \int dk' \mathcal{V}(z', k-k') f_W(z', k') - \overline{\Gamma} \int_{z^b(k)}^{z} dz' \left(f_W(z', k) - f_W^{\circ}(k)\right). \tag{49}$$

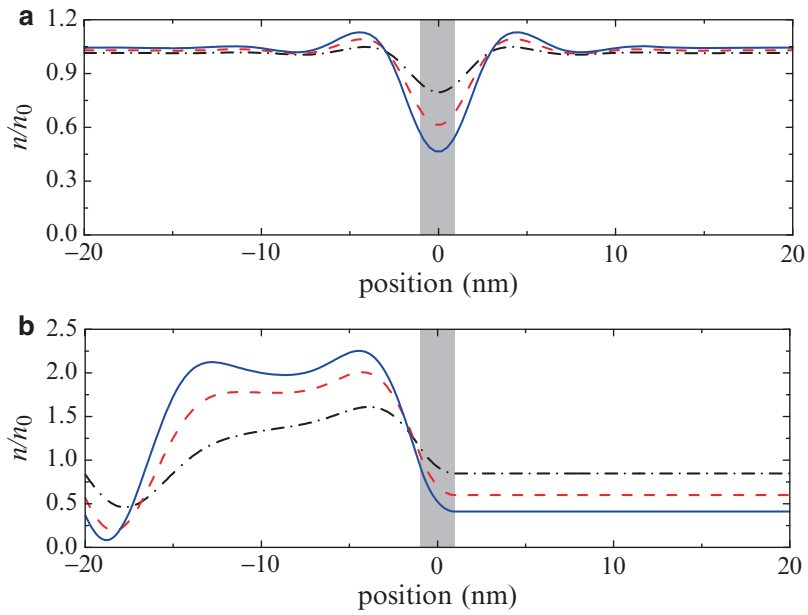
In order to show the intrinsic limitations of such classical-like boundary-condition treatment, let us consider, as prototypical device, a potential profile given by a rectangular barrier of width a and height V_0, i.e.,

$$V(z) = V_0 \theta\left(\frac{a}{2} - |z|\right), \tag{50}$$

where θ denotes the usual Heaviside step function. For this particular device potential profile, the explicit form of the corresponding potential superoperator in Eq. 38 turns out to be

$$\mathcal{V}(z, k'') = -\frac{4}{2\pi\hbar} V_0 \sin\left(2k''z\right) \frac{\sin\left(k''a\right)}{k''}. \tag{51}$$

As a first step, we shall investigate steady-state charge transport across the potential barrier in Eq. 50. In particular, in the coherent-limit regime $\left(\overline{\Gamma} \to 0\right)$ the results of the Wigner-function calculation may be directly compared to those of the well-known Landauer-Büttiker theory [14] based on elementary scattering-state calculations. Such comparison is reported in Fig. 3, where the steady-state current density corresponding to a room-temperature carrier injection from the left contact

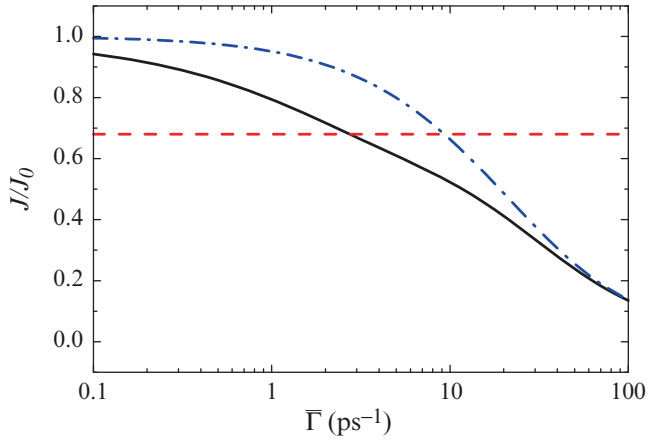


Microscopic Modeling of Solid-State Quantum Devices, Fig. 4 Spatial carrier density n (in units of its barrier-free ($V_0 = 0$) value n_0) in the presence of a rectangular barrier (Eq. 50) for the same device parameters and operation conditions employed in Fig. 3: comparison between the Wigner-function (panel (**a**)) and the Landauer-Büttiker results (panel (**b**)), for three different values of the barrier height: $V_0 = 50$ meV (*dash-dotted curves*), $V_0 = 100$ meV (*dashed curves*), and $V_0 = 150$ meV (*solid curves*) (see text)

only is plotted as a function of the barrier height V_0. The result of the Landauer-Büttiker formalism (dashed curve) describes a well-established behavior, i.e. for increasing values of the barrier height the tunneling process across the barrier is progressively reduced. In contrast, the result of the Wigner-function calculation (solid curve) appears to be unphysical: the value of the current turns out to be *independent* of the barrier height, and coincides with the barrier-free case.

In order to better understand such highly nonphysical behavior, it is useful to analyze the shape of the spatial charge density within the device active region. Figure 4 shows a comparison between the spatial charge density provided by the numerical solution of the Wigner equation (a) and the Landauer-Büttiker theory (b) for three different values of the barrier height: $V_0 = 50$ meV (dash-dotted curves), $V_0 = 100$ meV (dashed curves), and $V_0 = 150$ meV (solid curves). The charge-density profile given by the Landauer-Büttiker approach (b) well reproduces the physically expected scenario mentioned above: for increasing values of the barrier height, an increasing fraction of the injected carriers are reflected, and only a small fraction of them is transmitted into the right part of the device, inducing an asymmetry in the carrier density across the barrier. In contrast, the Wigner-function scenario reported in (a) clearly shows that an increase of the barrier height only results into a more pronounced minimum in the charge density due to the presence of the barrier itself. However, for any value of the barrier height the charge profile is always symmetric, implying that reflection is not accounted for. Indeed in this case all the injected carriers are able to reach the right contact of the device. This result shows that, in the coherent limit $\overline{\Gamma} \to 0$, the conventional Wigner-equation approach is definitely unable to account for key phenomena in mesoscopic transport theory, like quantum-mechanical reflection versus transmission.

In the presence of energy-dissipation/decoherence processes, the anomalous behavior reported in Figs. 3 and 4 is expected to be partially mitigated. Figure 5 shows the steady-state current density as a function of the relaxation rate $\overline{\Gamma}$ obtained from a numerical solution of the Wigner equation 43 for a fixed value of the barrier height ($V_0 = 100$ meV, solid curve) as well as for the barrier-free case ($V_0 = 0$, dash-dotted curve), compared to the (dissipation-free) result of the Landauer-Büttiker theory (dashed curve). As one can see, for a relaxation time $\left(\overline{\Gamma}^{-1}\right)$ of about 400 fs the current provided by the



Microscopic Modeling of Solid-State Quantum Devices, Fig. 5 Steady-state current density J (in units of its barrier-free and coherent-limit value J_0) as a function of the relaxation rate $\overline{\Gamma}$. The result is obtained from a numerical solution of the integral Wigner equation (49) for a fixed value of the barrier height ($V_0 = 100$ meV, *solid curve*) as well as for the barrier-free case ($V_0 = 0$, *dash-dotted curve*), compared to the (dissipation-free) result of the Landauer-Büttiker theory (*dashed curve*). Here, all the device parameters and operation conditions are the same employed in Figs. 3 and 4 (see text)

Wigner-function calculation (solid curve) well reproduces the Landauer-Büttiker one (dashed curve). However, it is crucial to point out the quite different physical origin of such current suppression. Within the Landauer-Büttiker approach this is due to quantum-mechanical reflection versus transmission in a fully coherent system, whereas within the Wigner-function approach this suppression is essentially driven by the presence of dissipation, via the relaxation-time term. The presence of the barrier has nevertheless some influence, as shown by the fact that the barrier-free current (dash-dotted curve) differs, in general, from the potential-barrier one (solid curve). Such difference vanishes in the coherent limit $\left(\overline{\Gamma} \to 0\right)$ as well as in the fully incoherent one $\left(\overline{\Gamma} \to \infty\right)$.

The numerical analysis presented so far clearly shows intrinsic limitations of the conventional Wigner-function treatment. In order to identify the origin of such anomalies, a few remarks are now in order.

Given an electron in the noninteracting state α, the corresponding pure-state density matrix is

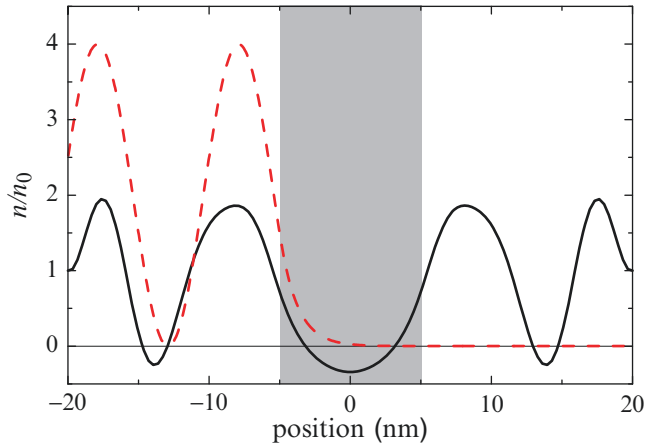
$$\rho_{\alpha_1\alpha_2} = \delta_{\alpha\alpha_1}\delta_{\alpha\alpha_2}, \tag{52}$$

and the corresponding pure-state Wigner function – obtained via the standard Weyl-Wigner transform (35) – is simply given by the diagonal elements $W_{\alpha\alpha}(z, k)$ in Eq. 36:

$$f_W^{\alpha}(z,k) = \int dz' \phi_{\alpha}\left(z + \frac{z'}{2}\right) e^{-ikz'} \phi_{\alpha}^{*}\left(z - \frac{z'}{2}\right). \tag{53}$$

In the coherent limit $\left(\overline{\Gamma} = 0\right)$ it is possible to show that (i) the Wigner function (53) is always a solution of the Wigner equation (39); (ii) by employing as inflow boundary values $f^b(k)$ (see Eq. 48) those given by the Wigner solution itself, the pure-state Wigner function (53) is a solution of the integral equation (49).

This implies that, if one is able to compute the Wigner function from the density matrix (like in Eq. 53 or any thermal average of it), such Wigner function will properly describe all the quantum-mechanical features of the problem – including reflection versus transmission phenomena – and will produce the very same results provided by the Landauer-Büttiker formalism. However, in the conventional Wigner-function approach, the Wigner function is initially unknown, and the spatial boundary condition $f^b(k)$ is treated as an arbitrary



Microscopic Modeling of Solid-State Quantum Devices, Fig. 6 Spatial carrier density n (in units of its barrier-free value n_0) corresponding to the rectangular-barrier profile in Eq. 50 (device length $l = 40$ nm, barrier width $a = 10$ nm, and barrier height $V_0 = 150$ meV): comparison between the Wigner-function result obtained in the coherent limit $\left(\overline{\Gamma} \to 0\right)$ assuming a monoenergetic injection $\left(\overline{\epsilon} = \hbar^2\overline{k}^2/(2m^*) = 50\,\text{meV}\right)$ from the left (*solid curve*) and the analytical result corresponding to its scattering-state calculation (*dashed curve*) (see text)

input function, usually expressed in terms of equilibrium or quasi equilibrium inflowing semiclassical carrier distributions (see Fig. 2). As recently pointed out in Savio and Poncet [20], such a semiclassical treatment/description of the boundary function $f^b(k)$ seems to be not necessarily compatible with the quantum-mechanical nature of a genuine Wigner function.

The above remarks about quantum-versus-classical boundary values lead to a crucial question: for a given semiclassical-like (i.e., positive-definite) boundary condition $f^b(k)$, under which condition is the corresponding solution $f_W(z, k)$ of the Wigner transport equation (39) physically acceptable?

To concretely address this important issue, we have performed a numerical solution of the integral Wigner equation (49) in the coherent limit $\left(\overline{\Gamma} \to 0\right)$ for the case of the rectangular-barrier potential in Eq. 50. In order to definitely point out nonphysical features of the Wigner-function treatment, we have replaced the thermal carrier injection employed so far by a simple monoenergetic carrier injection from left, i.e.,

$$f^b(k) \propto \delta\left(k - \overline{k}\right). \tag{54}$$

The result of such calculation (solid curve), together with the corresponding quantum-mechanical (i.e., scattering-state) one (dashed curve), is reported in Fig. 6. As one can see, in addition to the presence of an anomalous amount of charge on the right-hand side of the device with respect to the scattering-state result (dashed curve), the carrier-density profile corresponding to the mono energetic injection (solid curve) also displays a significant negative-value region and thus comes out to be definitely nonphysical, i.e., no more positive-definite. This result also agrees with the one reported in Taj et al. [19], thereby confirming the strong physical limitations of the conventional Wigner-function modeling. From the numerical analysis performed so far, one is forced to conclude that the set of mathematical solutions $f_W(z, k)$ corresponding to the set of all possible semiclassical (i.e., positive-definite) boundary conditions $f^b(k)$ contains also nonphysical solutions, i.e., solutions not corresponding to the Weyl-Wigner transform of a density matrix. Indeed, this important conclusion is fully supported by the detailed discussion about physical versus nonphysical solutions of the Wigner transport equation recently proposed in Rosati et al. [10], thus confirming the serious limitations of the conventional device modeling based on the Wigner-function formalism originally pointed out in Taj et al. [19].

Density-Matrix Based Treatment

In order to overcome the intrinsic limitations of the conventional Wigner-function modeling just recalled, the crucial step is to replace the local (i.e., classical-like) boundary condition-scheme treatment of the device-reservoir interaction with a fully nonlocal approach; to this end, in order to ensure/maintain the positive-definite character of the electronic density matrix, a possible strategy, recently proposed in Dolcini et al. [11], is to describe the system/device-environment/reservoir interaction via a Lindblad-like coupling term.

As discussed in detail in Dolcini et al. [11] and Rosati et al. [17], the explicit form of the desired device-reservoir coupling superoperator should fulfill the following three basic requirements: (i) it should preserve the positivity of the single-particle density matrix $\rho_{\alpha_1\alpha_2}$; (ii) due to the thermal (i.e., fully incoherent) nature of the external reservoirs, it should not induce phase coherence on the device electronic subsystem; (iii) in the semi-classical limit it should reduce to the standard injection/escape model in Eq. 34.

More specifically, in Dolcini et al. [11] the density-matrix treatment of section "Single-particle Description" is extended to quantum systems with spatially open boundaries, namely to the case of a quantum device electrically connected to one or more external carrier reservoirs; this is performed adding to the density-matrix equation (9) the device-reservoir scattering superoperator

$$\left.\frac{\mathrm{d}\rho_{\alpha_1\alpha_2}}{\mathrm{d}t}\right|_{\mathrm{res}} = -\frac{\Gamma_{\alpha_1}+\Gamma_{\alpha_2}}{2}\left(\rho_{\alpha_1\alpha_2}-\rho^{\mathrm{res}}_{\alpha_1\alpha_2}\right), \quad (55)$$

where

$$\rho^{\mathrm{res}}_{\alpha_1\alpha_2} = f^{\circ}_{\alpha_1}\delta_{\alpha_1\alpha_2} \quad (56)$$

denotes the quasiequilibrium density matrix of the external carrier reservoirs (injected over our single-particle states α) [11].

The above scattering superoperator can be regarded as the quantum-mechanical generalization of the injection/escape model in Eq. 34. Opposite to the genuine Lindblad form usually employed for the description of various carrier-environment coupling mechanisms [16, 17], the device-reservoir superoperator in Eq. 55 involves the reservoir density matrix $\rho^{\mathrm{res}}_{\alpha_1\alpha_2}$ in Eq. 56. This feature makes our density-matrix equation inhomogeneous, implying that the trace of the density matrix is not conserved, as expected in a system where the number of particles may change with time. Nevertheless, as discussed in Dolcini et al. [11]) the positive-definite character of the single-particle density matrix is ensured by the Lindblad-like form of the proposed coupling term in Eq. 55. It is also worth noticing that in this approach the local (i.e., classical-like) boundary-condition treatment of the system-reservoir interaction employed in the Wigner formulation [9] is replaced by a nonlocal quantum-mechanical coupling.

In order to point out the main features of the device-reservoir coupling in Eq. 55, it is useful to analyze the so-called mesoscopic regime, i.e., the case where inelastic scattering mechanism are negligible. In this regime the steady-state (ss) density matrix is given by

$$\rho^{\mathrm{ss}}_{\alpha_1\alpha_2} = \rho^{\mathrm{res}}_{\alpha_1\alpha_2} = f^{\circ}_{\alpha_1}\delta_{\alpha_1\alpha_2}, \quad (57)$$

and the standard Landauer-Büttiker result [14] is recovered: the steady-state single-particle density matrix $\rho^{\mathrm{ss}}_{\alpha_1\alpha_2}$ of the device is diagonal, thus preventing any phase-coherence transfer between reservoirs and device. Its diagonal values coincide with the distribution f°_{α} of the carriers injected from the reservoirs. Notice that such steady-state solution $\rho^{\mathrm{ss}}_{\alpha_1\alpha_2}$ is independent of the value of the device-reservoir coupling constants Γ_{α}.

It is however worth stressing that this density-matrix approach goes far beyond the steady-state regime just recalled, and enables one to address ultrafast-dynamics phenomena, nowadays accessible via time-resolved optical experiments. In particular, if at time $t = 0$ one "prepares" a device electron in a coherent superposition of single-particle states α, characterized by a device density matrix $\overline{\rho}_{\alpha_1\alpha_2}$, in the mesoscopic regime its subsequent time evolution is given by:

$$\rho_{\alpha_1\alpha_2}(t) = \rho^{ss}_{\alpha_1\alpha_2} + \left(\overline{\rho}_{\alpha_1\alpha_2} - \rho^{ss}_{\alpha_1\alpha_2}\right) e^{\frac{\left(\Sigma^*_{\alpha_1} - \Sigma_{\alpha_2}\right)t}{i\hbar}}, \tag{58}$$

where

$$\Sigma_\alpha = \epsilon_\alpha + i\hbar\Gamma_\alpha/2 \tag{59}$$

can be regarded to as the device single-particle self-energy ϵ_α "dressed" by an imaginary life-time contribution induced by the device-reservoir interaction. Thus Eq. 58 enables one to interpret the coupling Γ_α as the inverse time-scale over which the diagonal density-matrix element $\rho_{\alpha\alpha}$ reaches the steady state value $\rho^{ss}_{\alpha\alpha}$; furthermore, the time evolution in Eq. 58 shows the interplay between phase coherence and reservoir-induced dissipation versus decoherence processes. Indeed in Eq. 58 the diagonal contributions ($\alpha_1 = \alpha_2$) describe the population transfer and therefore energy dissipation, whereas the off-diagonal contributions ($\alpha_1 \neq \alpha_2$) exhibit a temporal decay of the inter-state polarizations, leading to decoherence.

In addition to the reservoir-induced dissipation/decoherence just described, in the presence of nonelastic scattering processes one deals with a complex interplay between carrier-reservoir and carrier-environment coupling; as anticipated, the latter may be described adding to the density-matrix equation (9) the device-reservoir superoperator in Eq. 55. The power and flexibility of such microscopic modeling has been recently demonstrated in Dolcini et al. [11] by investigating the highly nontrivial interplay between ultrafast optical-excitation and transport phenomena in a semiconductor-based triple-barrier nanodevice.

Cross-References

- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures
- ▶ Optical and Electronic Properties

References

1. Esaki, L., Tsu, R.: Superlattice and negative differential conductivity in semiconductors. IBM J. Res. Develop **14**, 61 (1970)
2. Capasso, F.: Bandgap Engineering: The Physics of Heterostructure Semiconductor Devices. Wiley, Chichester (2009)
3. Rossi, F.: Theory of Semiconductor Quantum Devices. Springer, Berlin/Heidelberg (2011)
4. Jacoboni, C.: Theory of Electron Transport in Semiconductor Materials and Structures. Springer, Berlin/Heidelberg (2010)
5. Leo, K.: High-Field Transport in Semiconductor Superlattices. Springer, Berlin/Heidelberg (2003)
6. Rossi, F. (ed.): Semiconductor Macroatoms: Basic Physics and Quantum-Device Applications. Imperial College Press, London (2005)
7. Rossi, F., Kuhn, T.: Theory of ultrafast phenomena in photoexcited semiconductors. Rev. Mod. Phys. **74**, 895 (2002)
8. Iotti, R.C., Rossi, F.: Microscopic theory of semiconductor-based optoelectronic devices. Rep. Prog. Phys. **68**, 2533 (2005)
9. Frensley, W.R.: Boundary conditions for open quantum systems driven far from equilibrium. Rev. Mod. Phys. **62**, 745 (1990)
10. Rosati, R., Dolcini, F., Iotti, R.C., Rossi, F.: The Wigner-function formalism applied to semiconductor quantum devices: Failure of the conventional boundary condition scheme. Phys. Rev. B. **88**, 035401 (2013)
11. Dolcini, F., Iotti, R.C., Rossi, F.: Interplay between energy dissipation and reservoir-induced thermalization in nonequilibrium quantum nanodevices. Phys. Rev. B. **88**, 115421 (2013)
12. Toda, M., Kubo, R., Saito, N.: Statistical Physics I. Springer, Berlin/Heidelberg (1983)
13. Haug, H., Jauho, A.-P.: Quantum Kinetics in Transport and Optics of Semiconductors, 2nd edn. Springer, Berlin/Heidelberg (2007)
14. Datta, S.: Quantum Transport: Atom to Transistor. Cambridge University Press, Cambridge (2005)
15. Iotti, R.C., Ciancio, E., Rossi, F.: Quantum transport theory for semiconductor nanostructures: A density-matrix formulation. Phys. Rev. B. **72**, 125347 (2005)
16. Taj, D., Iotti, R.C., Rossi, F.: Microscopic modeling of energy relaxation and decoherence in quantum optoelectronic devices at the nanoscale. Eur. Phys. J. B. **72**, 305 (2009)
17. Rosati, R., Iotti, R.C., Dolcini, F., Rossi, F.: Derivation of nonlinear single-particle equations via many-body Lindblad superoperators: A density-matrix approach. Phys. Rev. B. **90**, 125140 (2014)
18. Ciancio, E., Iotti, R.C., Rossi, F.: Gauge-invariant formulation of high-field transport in semiconductors. Phys. Rev. B. **69**, 165319 (2004)

19. Taj, D., Genovese, L., Rossi, F.: Quantum-transport simulations with the Wigner-function formalism: Failure of conventional boundary-condition schemes. Europhys. Lett. **74**, 1060 (2006)
20. Savio, A., Poncet, A.: Simple solution of the radial schroedinger equation with periodic potential of arbitrary form. J. Appl. Phys. **109**, 033713 (2011)

Microstereolithography

▶ Stereolithography

Microsystem

▶ Lab-on-a-Chip for Studies in *C. elegans*

Microsystems

▶ DNA Manipulation Based on Nanotweezers

Microthermal Analysis

▶ Scanning Thermal Microscopy

Microtrap

▶ Lab-on-a-Chip for Studies in *C. elegans*

μVED (Micro Vacuum Electronic Devices)

▶ MEMS Vacuum Electronics

Mini Pumps

▶ Micropumps

mirasol® Displays

▶ MEMS Display Based on Interferometric Modulation Technology

Modeling Nanoscale Heat Transport

▶ Simulating Nanoscale Heat Transport

Modeling of Batteries and Fuel Cells

▶ Electrochemical Interfaces for Energy Storage and Conversion

Modeling Thermal Properties of Carbon Nanostructure Composites

Clifford W. Padgett[1] and Tad S. Whiteside[2]
[1]Chemistry and Physics, Armstrong Atlantic State University, Savannah, GA, USA
[2]Savannah River National Laboratory, Aiken, SC, USA

Synonyms

Thermal management materials

Definition

Carbon nanostructure composites consist of carbon nanostructures, like carbon nanotubes, diamond nanorods, graphene ribbons, etc., dispersed in a polymer matrix. The thermal properties of interest for carbon nanostructure composites are the thermal conductivity of the matrix, the thermal conductivity of the nanostructure, and the interfacial thermal conductivity between the matrix and the nanostructure. The interfacial

thermal conductivity is the limiting factor on the overall thermal performance of these materials.

Overview

Carbon nanostructures have a relatively high thermal conductivity, which should make them ideal for use as heat sinks, particularly in electronics. At 300 K, isolated single- and multiwalled carbon nanotubes (CNTs) have thermal conductivities that range from 300 to 6,600 $Wm^{-1}\ K^{-1}$ [1], for diamond nanorods (DNRs) the thermal conductivity is near 50 $Wm^{-1}\ K^{-1}$ [2], and vapor grown carbon fibers have a thermal conductivity of 1,950 $Wm^{-1}\ K^{-1}$ [3]. At 300 K, most polymers have a relatively low thermal conductivity, approximately 0.10 $Wm^{-1}\ K^{-1}$ [1, 4, 5]. It has been proposed that combining carbon nanostructures with a polymer matrix will result in an advanced material with improved thermal characteristics compared to the polymer alone, while maintaining or improving other desired properties needed for the application, such as being electrically insulating, having high strength, and being moldable [1, 4, 6, 7]. These carbon nanostructure composites typically consist of single- or multiwall carbon nanotubes (CNTs) or other nanostructures of carbon well dispersed into a polymer matrix at low weight percent concentrations.

The percolation threshold, the minimum concentration where well-dispersed carbon nanotubes become fully interconnected, has been experimentally determined by many researchers [2, 6]. This concentration varies from 0.023 to 11 wt%, depending on the carbon nanostructure, the polymer matrix, and the dispersion method [2, 6]. Once above the percolation threshold, the thermal conductivity of the composite, κ_{comp}, can be estimated from the following equation, assuming the thermal resistance at the nanotube interfaces is negligible [4]:

$$\kappa_{comp} = \left\langle \cos^2\theta \right\rangle \kappa_{nano} V_{nano} \qquad (1)$$

where κ_{nano} is the thermal conductivity of the nanomaterial, V_{nano} is the volume fraction (percent composition/volume concentration) of the nanomaterial, and θ is the angle between a given direction and the nanomaterial axis. The brackets indicate an average over all nanomaterials in thecomposite; for well-aligned nanomaterials, $<\cos^2\theta> = 1$ and for completely random orientation $<\cos^2\theta> = 1/3$. Using this expression, an estimate of the thermal conductivity for a composite consisting of randomly dispersed CNTs with a thermal conductivity of 3,000 $Wm^{-1}\ K^{-1}$, at a volume fraction of 0.01, is 10 $Wm^{-1}\ K^{-1}$. Although experimental studies on composites with CNT concentrations well above the percolation threshold show an increase in thermal conductivity, this increase is limited to less than 1 $Wm^{-1}\ K^{-1}$ at 300 K [5, 6]. It is thought that these lower than expected results are due to limited interfacial thermal conductivity between the CNT and the polymer matrix [5, 7]. The failure of the material to perform as expected has sparked many computational studies. Some have attempted to elucidate the mechanism behind this limited interfacial thermal conductivity while others have worked to find alternative nanostructures with improved performance.

Basic Methodology

Atomic-scale modeling is used to characterize the thermal properties of carbon nanostructure composites. These properties include the thermal conductivity of pristine and functionalized nanostructures and the interfacial thermal conductivity between the carbon nanostructure and the polymer matrix [1, 4–6]. The study of these systems provides insight into the atomic-level physical and chemical interactions that contribute to the thermomechanical properties of carbon nanostructure composites. They also provide important data not readily available from experimental measurements. The results of these studies are being used to develop a complete analytical model of thermal conductivity in nanostructure composites and a more complete theory of how functionalization affects interfacial thermal conductivity.

In a typical study of the thermal conductivity of a carbon nanostructure composite, the isolated

nanostructure is built using standard techniques. This structure is equilibrated at a specified temperature with potential energy functions chosen to describe the bonding and nonbonding interactions on and between the atoms. Commonly used potential energy functions are REBO, ReaxFF, and PCFF; these can be found in the LAMMPS software package [8]. After equilibrating the carbon nanostructure, it is solvated with the matrix material used to create the composite. This is usually done using an optimization program, such as the Packing Optimization for Molecular Dynamics Simulations (Packmol) [9]. The optimization program places the matrix molecules in a defined region of space, ensuring short-range repulsive interactions do not overwhelm the simulation. Once the complete composite has been created, the system is again equilibrated to ensure it begins in a stable state. The thermal conductivity is typically measured using one of the following methods. These methods are Fourier's law, the Müller-Plathe reverse nonequilibrium molecular dynamics, and the Green–Kubo formula.

Fourier's law describes how heat energy moves in a diffuse manner across a temperature gradient

$$J = -\kappa \nabla(T) \tag{2}$$

where the heat energy, J, is related to the spatial gradient of the temperature, $\nabla(T)$ by the thermal conductivity, κ. In this method, the energy added to the hot region equals the energy subtracted from the cold region and is proportional to the heat flux moving between the regions [8].

Another method to determine the thermal conductivity is through the reverse nonequilibrium molecular dynamics method described by Müller-Plathe [8, 10]. In this method, the velocities (kinetic energy) of the hottest atoms in the cold region are switched with the coldest atoms in the hot region. This results in a known quantity of heat flux, J, being transferred against the thermal gradient between the two regions:

$$J = \frac{\sum_{i=1}^{n} \frac{1}{2} m_i \left({v_{hi}}^2 - {v_{ci}}^2 \right)}{2\mathrm{At}} \tag{3}$$

where m is the mass of the atoms, t is the time in the simulation, A is the cross-sectional area, v_h and v_c are the velocities chosen from the cold and hot regions, respectively, and n is the number of transfers. After the system has reached equilibrium the thermal conductivity κ is obtained as:

$$\kappa = -\frac{\langle J(t) \rangle}{\langle \delta T / \delta z \rangle} \tag{4}$$

where $J(t)$ is the total heat flux, $\delta T/\delta z$ is the temperature gradient between the two regions, and the brackets denote the statistical time average.

The third method to determine thermal conductivity is derived from the Green − Kubo formula which relates the ensemble average of the autocorrelation of the heat flux to the thermal conductivity

$$\kappa = \frac{1}{V k_B T^2} \int_0^\infty \langle J(0) J(t) \rangle dt \tag{5}$$

where V is the volume, k_B is the Boltzmann constant, and T is the temperature. The heat flux vector, J, is calculated as

$$J(t) = \sum_i \varepsilon_i v_i + \frac{1}{2} \sum_{ij, i \neq j} r_{ij} \left(f_{ij} \cdot v_i \right) \tag{6}$$

where the fluctuations of the per-atom (i) energy (ε) are added to the per-atom stress tensor in a steady-state equilibrated simulation, v_i is the velocity of the ith atom, r_{ij} is the distantace between atoms i and j, and f_{ij} is the force between atoms i and j [1, 8, 11].

Key Research Findings

The thermal conductivity of a carbon nanotube is influenced by its diameter, chirality, length, temperature, and functionalization. As described above, the thermal conductivity is calculated based on the volume or cross-sectional area of the nanostructure, both of which require the cross-sectional area. There is no standard

definition for measuring the cross-sectional area: one definition uses an annulus centered on the carbon atoms, each with a thickness of 0.34 nm, the van der Walls radius of carbon; another uses the inter-atom separation distance between layers of graphite as the thickness of the annulus; other definitions calculate the area as a circle with a circumference defined by the centers of atoms around the nanotube [1, 10, 11]. Simulations reveal that, for CNTs in a vacuum, the thermal conductivity is minimally affected by the overall diameter or by changes in the chirality [10]. Some simulations indicate for composites, the larger the diameter of the CNT the lower the thermal transfer; this is due to a decrease in the vibrational bending modes in the nanotube [12]. As the length of a carbon nanotube increases, in a vacuum or a composite structure, the thermal conductivity also increases, in a power–law relationship between nanotube length and thermal conductivity, with an exponent value less than one [2, 9, 10]. Simulations indicate that as the temperature of a single carbon nanotube in vacuum increases, the thermal conductivity increases, to a maximum near 100 K, and then decreases exponentially [1, 11]. In carbon nanostructure composite materials, the thermal conductivity of the composite follows the same trend as the isolated carbon nanotube, but with the maximum value found near 300 K [5].

Above 300 K, isolated carbon nanotubes and graphene monolayers have similar thermal conductivities; however, below 300 K, graphene's thermal conductivity is much greater [11]. When simulations of graphene and bulk graphite are compared, it has been shown that the interlayer interactions in the graphite quench the thermal conductivity of this system by an order of magnitude [11]. A similar reduction is seen when two carbon nanotubes come into contact with each other, as in a multiwall carbon nanotube or when a "bundle" or "mat" of single-walled carbon nanotubes are formed [11].

As increasing numbers of functional groups are added to a carbon nanotube, the thermal conductivity of the carbon nanotube decreases [2]. Also, as the carbon nanotube length increases, but the percent functionalization remains constant, the thermal conductivity of the carbon nanotube reaches a maximum with respect to length [2]. Additionally, this maximum occurs at shorter lengths as the percent functionalization increases, implying a short scattering length for phonons in the carbon nanotube. Comparing the phonon power spectra of functionalized carbon nanotubes shows a decrease in spectral density at the high frequency carbon–carbon optical mode, while the low frequency acoustic modes remain nearly constant [2]. These results support the idea that the reduction in thermal conductivity of carbon nanotubes is caused by a reduction in phonon scattering length rather than changes in the vibrational modes caused by functionalization [2].

The thermal conductivity of several carbon nanostructure composite materials has been examined. Composite structures with single-wall carbon nanotubes require only 0.1–0.2 wt% before the onset of percolation, whereas structures composed of vapor grown carbon fibers require a mixture of 1.0–2.0 wt% [5]. In these experiments, the thermal conductivity of the composite structure reached a maximum of approximately 0.5 $Wm^{-1}\ K^{-1}$ while mixing theory predicts a value of 10–20 $Wm^{-1}\ K^{-1}$ [5]. Other studies have shown that aligning the fibers has a dramatic increase in the thermal conductivity; one study showed a composite of aligned vapor grown carbon fiber at 56 vol.% in epoxy had a thermal conductivity of 291 $Wm^{-1}\ K^{-1}$ [3].

Another possible reason for the lower than expected thermal conductivity of composites is the interactions between nanotubes and composite scatter phonons traveling along the tube. In addition, due to the chemical inertness of nanotube walls, only weak dispersion forces exist between the tube and matrix. Atomistic simulations demonstrate that similar dispersion forces acting between graphite planes or between single-wall carbon nanotubes lead to a reduced conductivity but only by a factor of 2 or 3, not by the observed factor of 10 [12]. According to these results, the thermal properties of the composite or nanostructure alone are not enough to explain the experimental observations [12]. Other studies suggest that the low thermal conductivity of composites may be caused by the interfacial thermal resistance [4]. A measure of the relative importance

of the interfacial resistance to the overall heat flow in a composite is obtained from the "equivalent matrix thickness," h. For the planar interface geometry, h is equal to the matrix thickness over which the temperature drop is the same as at the interface, this is given by the ratio of bulk thermal conductivity (κ) to the interfacial conductance (Λ):

$$h = \frac{\kappa}{\Lambda} \tag{7}$$

h has been measured for nanotubes in composite materials [4] and is found to be approximately 5 nm [12]–20 nm [4]. This means that even when touching, two nanotubes appear to be 10–40 nm apart, coupled only by weak dispersion forces and a small contact area [4]. Continuing the discussion of interfacial thermal resistance, Shenogin et al. [12] mention Λ is affected by nanotube length, which is apparent only at small lengths (1–10 nm) because Carlborg et al. [7] show that interfacial conductance (Λ) remains nearly constant for nanotubes at lengths from 1 to 50 nm. For CNTs, this value is approximately 1.6 $MWm^{-2}\,K^{-1}$ in liquid argon [7].

In a single-walled carbon nanotube composite, the interaction between the nanotube and matrix material is weak, and thus requires low frequency vibrational modes to be activated for the transfer of thermal energy between the two materials. When the carbon nanotube is heated, the high-frequency vibrational modes become activated [4, 7]. Energy is transferred between vibrational modes through phonon–phonon coupling. In carbon nanotubes, this intermode transfer is slow, limiting the amount of energy transferred from the high- to low-frequency vibrational modes and then exchanged into the matrix [4]. If the matrix is not bound to the CNT, where the phonon modes become coupled, the matrix material does not greatly effect the Λ [4].

Functionalization of CNT surface before embedding in matrix increases the interfacial thermal conductance by increasing the interfacial adhesion [6]; however, functionalization decreases the intrinsic thermal conductivity of the CNT [2]. Functionalization of the CNT surface at various percent functionalization (0–0.11 %) with hydrocarbon chains of various lengths (8, 12, 18, 50) in a composite has been examined [9]. It was found that as grafting density increases, the interfacial thermal conductivity increases, as the chain length increases, the interfacial thermal conductivity increases to close to 50 $Wm^{-1}\,K^{-1}$, approaching the theoretical limit. Also, the interfacial thermal conductivity is the same if the chains are covalently bound to the nanotube (sp^3 bond) or if there is a defect and then the chain is bound (sp^2 bond) [9].

Future Directions for Research

The current models and results clearly show the potential of carbon nanostructure composite systems for use as thermal management materials. Before their full potential can be realized experimentalists will need to create functionalized carbon nanostructure composites based on the above results and determine their thermal conductivity. Molecular modelers will need to refine their simulations to guide the development of an analytical model that uses the various physical dimensions of the carbon nanostructure, combined with the size of the functional groups and the number of surface atoms functionalized on the carbon nanostructure and the surrounding environment to predict the thermal conductivity of that particular composite.

While an initial model has been developed, as described in reference [5], it does not account for the functionalization effect on the intrinsic thermal conductivity of the carbon nanostructure. With an improved model in hand, and the relationship between functionalization effect and functionalization group determined, modelers can begin to examine the interfacial thermal conductivity of different functional groups in various polymers (i.e., phenyl groups for CNTs embedded in styrene vs. hydrocarbon chains in octane).

In the long term, once the results for functionalized hydrocarbon systems are understood, additional modeling potentials, which can handle halogens or metals, may be used to

develop a more complete theory of more complex composite systems.

Cross-References

- ▶ Active Carbon Nanotube-Polymer Composites
- ▶ Carbon-Nanotubes
- ▶ Thermal Conductivity and Phonon Transport

References

1. Lukes, J.R., Zhong, H.: Thermal conductivity of individual single-wall carbon nanotubes. J. Heat Transf. **129**, 705–716 (2007)
2. Padgett, C.W., Brenner, D.W.: Influence of chemisorption on the thermal conductivity of single-wall carbon nanotubes. Nano Lett. **4**, 1051–1053 (2004)
3. Al-Saleh, M.H., Sundararaj, U.: A review of vapor grown carbon nanofiber/polymer conductive composites. Carbon **47**, 2–22 (2009)
4. Huxtable, S.T., Cahill, D.G., Shenogin, S., Xue, L., Ozisik, R., Barone, P., Usrey, M., Strano, M.S., Siddons, G., Shim, M., Keblinski, P.: Interfacial heat flow in carbon nanotube suspensions. Nat. Mater. **2**, 731–734 (2003)
5. Biercuk, M.J., Llaguno, M.C., Radosavljevic, M., Hyun, J.K., Johnson, A.T.: Carbon nanotube composites for thermal management. Appl. Phys. Lett. **80**, 2767–2769 (2002)
6. Du, J.-H., Bai, J., Cheng, H.-M.: The present status and key problems of carbon nanotube based polymer composites. Express Polym. Lett. **1**, 253–273 (2007)
7. Carlborg, C.F., Shiomi, J., Maruyama, S.: Thermal boundary resistance between single-walled carbon nanotubes and surrounding matrices. Phys. Rev. B **78**, 205406 (2008)
8. Plimpton, S.: Fast parallel algorithms for short-range molecular dynamics. J. Comp. Phys. **117**, 1–19 (1995). http://lammps.sandia.gov
9. Clancy, T.C., Gates, T.S.: Modeling of interfacial modification effects on thermal conductivity of carbon nanotube composites. Polymer **47**, 5990–5996 (2006)
10. Alaghemandi, M., Algaer, E., Böhm, M.C., Müller-Plathe, F.: The thermal conductivity and thermal rectification of carbon nanotubes studied using reverse non-equilibrium molecular dynamics simulations. Nanotechnology **20**, 115704 (2009)
11. Berber, S., Kwon, Y.-K., Tomanek, D.: Unusually high thermal conductivity of carbon nanotubes. Phys. Rev. Lett. **84**, 4613–4616 (2000)
12. Shenogin, S., Xue, L., Ozisik, R., Keblinski, P., Cahill, D.: Role of thermal boundary resistance on the heat flow in carbon-nanotube composites. J. Appl. Phys. **95**, 8136–8144 (2004)

Models for Tumor Growth

Luigi Preziosi
Dipartimento di Matematica, Politecnico di Torino, Torino, Italy

Synonyms

Cancer modeling

Definition

Procedure to transfer biological processes related to tumor growth to a system of equations describing the state and the evolution of the biological system through suitable numerical simulation tools.

Introduction and Historical Landmarks

Cancer is one of the most complex diseases and this by itself represented a strong stimulus for the development of many different models, each one focusing on a specific aspect of its complexity. In fact, in addition to the multiplicity of known tumors, in general one can observe that even focusing on a single type of tumor, there can be several causes of carcinogenesis and several different evolutionary trends. One can then say that tumors are multifactorial and multipath, and mathematical models typically consider only one of the possible starting points and one of the possible evolutions (see [1]). Tumors are also heterogeneous and multicomponent with the presence of several other host cells playing relevant roles in tumor growth and of complex interactions with the environment. Even within the population of tumor cells, multiple clones with different abilities and characteristics can be identified. Every single aspect categorized above can be the object of a mathematical model.

The classification of the type of models to describe tumor evolution is then rather difficult and several criteria can be used. One way is to follow the different phases of growth distinguishing during the evolution of tumors:

- An avascular phase, in which the tumor is not yet surrounded by a capillary network and the nutrients necessary for cell survival reach the cells via a diffusion process
- A crucial phase, called angiogenesis, consisting in the stimulus and the formation of a vascular network surrounding the tumor, which represents the transition from a benign to a malignant growth
- The subsequent vascular growth, when the tumor mass, surrounded by a capillary network, coopts existing and newly formed vessels and grows around them
- The invading phase characterized by a strong tumor–tissue interaction involving both mechanical and chemical processes causing, for instance, the rupture of basement membranes and the degradation of the extracellular matrix
- The diffusion of metastases, including the processes of intravasation and extravasation via the blood or the lymphatic system and the formation of secondary tumors
- The therapeutic phase with the optimization of chemotherapies and radiotherapies, drug delivery, use of nanotechnologies, thermotherapies, and so on

Another way of classifying the mathematical models is related to the scale of interest, in the sense that the model can focus on what happens inside the single cell, on the behavior of the single tumor cell, on the interaction between cells and between tumor and healthy cells, or on the massive growth of the tumor. As a first approximation one can state that in general

- Macroscopic models, aimed at describing the tumor at the tissue level, write as systems of partial differential equations.
- Mesoscopic models, aimed at describing the tumor at the cellular level, develop in the framework of kinetic theories, individual-based models, cellular automata, and similar.
- Microscopic models, aimed at describing the tumor at the subcellular level, are mainly based on systems of ordinary differential equations, or network methods.

The history of mathematical modeling of tumor growth started in the 1950s with the first works by Thomlinson and Gray (1955), Burton (1966), and Greenspan (1972) (see [2] for references), who had mainly the merit of bringing the topic to the attention of the scientific community. The models were mainly based on diffusion equations and mass balance laws. The interest grew considerably in the 1990s with the development of many models which were, however, based on the work just mentioned. The history was then characterized by a burst of attention at the beginning of this century. This was due to several reasons:

- The improved knowledge on the behavior of tumors, both at the tissue and at the cellular level, pushed for the development of more detailed frameworks, such as multiphase models and kinetic models.
- The improvement of computational capacities gave a strong stimulus to the development of individual-based models, cellular automata, cellular Potts models, that could handle an acceptable number of cells.
- The decoding of the human genome and the development of proteomics, with the study of the signaling cascades and of the protein networks governing the behavior of cells, opened up to the burst of subcellular models, to the birth of system biology and to its application in cancer modeling.
- The use of nanotechnologies allowed new experimental measurements performed at the cellular and subcellular level and to the device of possible new therapies to be virtually tested.

Modeling Avascular Tumors

Referring to [2–4] for recent reviews or to the volumes [5–7], the first models dealing with avascular tumor growth worked under the hypotheses that the tumor is made by only one type of cells occupying a time-dependent region of space $\Omega(t)$ keeping a constant density $\overline{\rho}$. In this approach, the tumor cells fill the space as a bunch of rigid spheres in a close-packing configuration, so that the mass balance equation simplifies to

$$\overline{\rho}\nabla \cdot \mathrm{v}_T = \Gamma_T,$$

where Γ_T is a function of the concentration of nutrients and of many fundamental chemical cues, e.g., growth promoting factors and growth inhibitory factors, that are diluted in the extracellular liquid surrounding the cells and that influence all vital functions of the tumor.

Enforcing a symmetry condition, usually spherical symmetry, allows to reduce the number of space variables to one and the velocity vector to a scalar, so that one can directly integrate the equation above to have the velocity in any point and, in particular, the evolution of the free border of the tumor

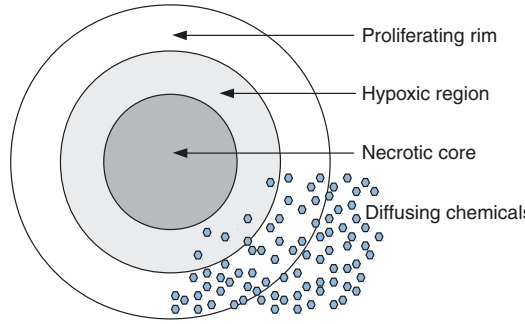
$$\begin{aligned}\frac{dR}{dt}(t) &= v_T(R(t),t)\\ &= \frac{1}{R^2(t)\overline{\rho}}\int_0^{R(t)} \Gamma_T(C_i(r,t))r^2 dr\end{aligned}$$

This equation is equivalent to writing an integral mass balance over the entire tumor volume.

The core of these types of models consisted then in describing how the growth term Γ_T depends on the chemical factors and nutrients c_i diffusing in the environment according to reaction–diffusion equations

$$\frac{\partial c_i}{\partial t} = D_i\nabla^2 c_i + \gamma_i - \partial_i c_i,$$

where D_i is the diffusion coefficient, γ_i is the production term, and δ_i is the absorption/decay term. The problem is then a free boundary problem, being the equation above valid for $r < R(t)$.



Models for Tumor Growth, Fig. 1 Structure of an avascular tumor

In particular, nutrient-limited radius can possibly be obtained if the equation

$$\int_0^{R\infty} \Gamma_T(c_i(r,t))r^2 dr = 0$$

has a solution that can be interpreted as a global balance between proliferation in the outer rim and death mainly occurring in the core (see Fig. 1).

Later on, in order to describe tumors with greater accuracy, it became necessary to consider the fact that, as most tissues, tumors were made of cells (possibily of different types), extracellular liquid, and extracellular matrix. The first multiphase models to appear were, of course, biphasic models mainly deduced with the aim of describing the behavior of (avascular) multicellular spheroids in vitro and to overcome some difficulties encountered in generalizing the constant density models described above either to three-dimensional problems or to problems involving more populations. The tumor mass was then treated as an ensemble of cells wet by an extracellular liquid full of chemicals, such as nutrients and growth factors. Since the size of the proteins dissolved in the extracellular liquid, absorbed or produced by the cells is very small compared to the size of the cells, their evolution can still be described by suitable reaction–-diffusion equations. The ensemble of tumor cells instead needs to be described using mass and momentum balance equations.

The inclusion of the capillary network and the extracellular matrix was a natural step that allowed to consider, still in the multiphase framework, fundamental interactions with the environment and to describe vascular growth, tumor–tissue interaction, and invasion of the surrounding tissue. Though many models consider three phases (cell population, liquid and extracellular matrix), more general multiphase models were also developed to describe the interaction of tumors with external tissues, to include different cell populations, e.g., fibroblasts or possibly engineered macrophages, to consider polyclonality of tumor masses, to distinguish the different constituents making up the extracellular matrix, e.g., collagen and elastin.

Referring to [8], chapter 9 of [6], or chapter 7 of [7] for more details, in addition to the reaction–diffusion equations for the molecules considered essential for the description of the evolution of the system, the structure of the model consists in a set of mass and momentum balance equations for the constituents. Denoting by α the generic constituent, one can then write

$$\frac{\partial \phi_\alpha}{\partial t} + \nabla \cdot (\phi_\alpha \mathbf{v}_\alpha) = \Gamma_\alpha,$$

(where ϕ_α is the volume ratio of the constituent, $\mathbf{v}_\alpha$ its velocity, and Γ_α its growth/death term) and, considering that inertia is negligible in the description of growth phenomena,

$$-\phi_\alpha \nabla p + \nabla \cdot \left(\phi_\alpha \mathbf{T}'_\alpha\right) + \mathbf{m}_\alpha = \mathbf{0}$$

where p is the interstitial pressure of the extracellular fluid, $\mathbf{T}'_\alpha$ is the so-called excess stress tensor, and $\mathbf{m}_\alpha$ is the interaction force accounting for the forces acting on the constituent α due to the interactions with the other components of the mixture.

The model then specializes according to the constituents considered. For instance, in many cases, the tumor mass can be modeled as a porous material, or the mechanical interaction with the liquid can be neglected. In addition, recent nanotechnologies allow to experimentally evaluate the strength of single adhesion bonds between cells and between cells and extracellular matrix and the mechanical properties of the cells. Such techniques include the use of optical tweezers, of AFM cantilevers with suitably functionalized tips to compress the cell body or to stretch the adhesion bonds, of functionalized nanobeads attached to the cell membrane or that, internalized by the cell and dispersed in the cytosol, are then suitably stimulated. This information can then be upscaled to estimate the parameter appearing in the partial differential equations [9].

To be specific as an example it is here reported a tri-phasic model deduced under the assumption that the interactions with the liquid are negligible and that the extracellular matrix is a deformable porous medium in which the cellular population moves and grows like a sticky granular material.

$$\begin{cases} \dfrac{\partial \phi_c}{\partial t} + \nabla \cdot (\phi_c \mathrm{v}_c) = \Gamma_c, \\ \dfrac{\partial \phi_m}{\partial t} + \nabla \cdot (\phi_m \mathrm{v}_m) = \Gamma_m, \\ \mathrm{v}_c - \mathbf{v}_m = K\left(1 - \dfrac{\sigma}{\left|\nabla \cdot (\phi_c \mathrm{T}'_c\right|}\right)_+ \nabla \cdot \left(\phi_c \mathbf{T}'_c\right), \\ \nabla \cdot \left(\phi_c \mathbf{T}'_c + \phi_m \mathbf{T}'_m\right) = \mathbf{0} \end{cases}$$

where c refers to the cellular population, m to the extracellular matrix, $(\cdot)_+$ stands for the positive part of the parenthesis, K is the motility coefficient, and σ measures the adhesive properties of the cell to the extracellular matrix. Of course, the equations above need to be accompanied by the constitutive equations for the excess stress tensors $\mathbf{T}'_c$ and $\mathbf{T}'_m$, describing the response of both cells and ECM to deformation.

Recently, also Lowengrub and coworkers [3] focused their attention on the adhesion properties of cellular populations introducing it as an adhesion energy. Still working in a multiphase framework, the final model consists of a fourth-order nonlinear advection-reaction–diffusion equation of Cahn–Hilliard type characterized by the presence of a term regularizing the backward heat equation term related to adhesion phenomena.

Modeling Angiogenesis

When the tumor grows too much getting larger than few hundreds microns, the amount of nutrients reaching the center of the tumor is not sufficient to feed all cells, so that its central region becomes hypoxic (see Fig. 1). In this situation most cells die, but some tumor cell might have acquired the ability of producing a class of protein generally called tumor angiogenic factors that stimulate the existing capillaries to produce new capillary networks reaching the tumor. This process, called angiogenesis, is a fundamental step in tumor growth because it distinguishes the avascular phase from the vascular phase. The difference between the two phases of growth is dramatic because in the avascular phase the tumor is very small (with a radius that is smaller than a millimeter and limited by the diffusion of nutrients), stationary, and therefore harmless, while in the vascular phase the tumor has all the nutrients to grow in an unlimited way and has a highway to spread metastases throughout the body.

One of the most celebrated model is the one due to Chaplain and Anderson (1998) (see [10] for a review on the topic, and the volume [11]) who had the main merit to relate a continuous model with its discrete version allowing a better visualization of the process of formation of capillary trees starting from existing vessels.

In their model Chaplain and Anderson omit any birth and death process for the cells forming the capillaries focusing on the trajectories formed by the tip cells of the newborn capillaries, assuming that the endothelial cells forming the capillary walls will follow. Other key ingredients are chemotaxis toward higher densities of tumor angiogenic factors produced by the tumor, and haptotaxis toward higher concentration of extracellular matrix that is degraded by the tip cells. The model then writes as

$$\begin{cases} \dfrac{\partial n}{\partial t} + \nabla \cdot \left[\left(w_c \nabla c + w_f \nabla f \right) n \right] = k_n \nabla^2 n, \\ \dfrac{\partial c}{\partial t} = -\delta_c n c, \\ \dfrac{\partial f}{\partial t} = \gamma n - \delta_f n f, \end{cases}$$

where n is the density of tip cells, c is the density of tumor angiogenic factor, and f is the density of fibronectin. In addition, w_c and w_f are motility coefficients referring respectively to chemotaxis and haptotaxis sensitivity, k_n refers to random motility, δ_c and δ_f are the uptake coefficients from the tip cells, and γ is the production rate of fibronectin by the cells.

The aim of the discrete technique then introduced develops in the framework of reinforced random walk. It consists in using the resulting coefficients of the five-point stencil of the standard central finite-different scheme to generate the probabilities of movement of an individual cell in response to the chemoattractant gradients and to diffusion.

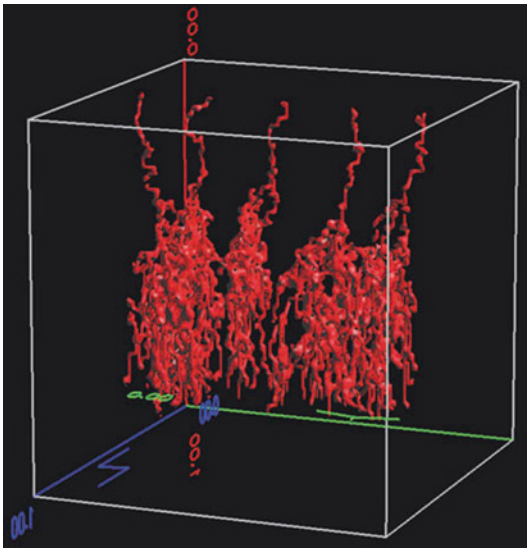
The extraction of a random number will decide whether the tip cell will stay still in the node or will move to a particular neighboring node rather than another. However, with respect to the standard random walk algorithm, in presence of chemotaxis or haptotaxis the motion becomes biased, because the cell has higher probabilities to move up the gradients of chemical factors.

In addition to that, the discretized setup allows to include some phenomena difficult to describe using a model based on partial differential equations, e.g., capillary branching and anastomosis, i.e., the formation of capillary loops.

In fact, when during their motion a capillary tip meets another capillary, then they merge to form a loop. If two sprout tips meet, then only one of the original sprouts continues to grow.

As shown in Fig. 2, the capillary network built in this way looks very realistic and compares well with what is observed experimentally. A three-dimensional animation of the angiogenic process is available at the web site www.maths.dundee.ac.uk/sanderso/3d/index.html.

Historically speaking, the interest in modeling angiogenesis was due to the interest in developing anti-angiogenic therapies aimed at destroying the disordered and leaky tumor-induced vessel network. Unfortunately, at present antiangiogenic therapies don't look so successful as expected. However, there is now a renewed interest due to the concept of vessel remormalization proposed in order to optimize drug delivery [12].



Models for Tumor Growth, Fig. 2 Tumor-induced capillary trees as simulated by Chaplain and Anderson (1998)

Modeling Vascularized Tumors and Invasion

Compared with the models focusing on the avascular phase, relatively fewer models were developed dealing with tumor invasion and vascular growth.

Most of them have a structure similar to the ones mentioned above. For instance, some vascular growth models develop assuming a constant density of tumor cells duplicating along a blood vessel and forming what is called a tumor cord possibly surrounded by a necrotic region farther away from the blood vessel (see chapter 3 of [13] or chapter 7 of [6]). Others develop in the multiphase framework looking more closely at the change in metabolism occurring in some tumor cells that in hypoxic conditions acquire the ability to survive and extract energy also in absence of oxygen. This makes the environment more acidic and unsuitable for the host tissue that dies leaving room for tumor growth (see chapter 4 of [13]). Other models focus on the mechanical interaction with the capillary walls, and on the fact that newborn vessels are immature and therefore subject to mechanical collapse due, for instance, to the pressure exerted by the tumor growing around them. This will cause hypoxia and the generation and remodeling of blood vessels. Others again use the discrete angiogenesis model described above to focus on the perfusion of drugs or on vessel remodeling (see chapter 15 of [6] or [11]).

Another key aspect in tumor invasion is the ability acquired by the tumor cells to produce and secrete certain factors known as matrix degrading enzymes that chop the extracellular matrix and then modify the local tissue environment making it easier for the cells to invade the surrounding tissue finding easier paths to move. This process, called metallo-proteinases, is also used by tumor cells to break through basal membranes, which otherwise would compartimentalize the tumor.

A simple example of models including the action of matrix degrading enzymes is, for instance, the following (see chapter 10 of [5] for more details):

$$\left\{ \frac{\partial n}{\partial t} + \nabla \cdot \left[(\chi(n,f) n \nabla f) = D_n \nabla^2 n + \Gamma, \frac{\partial f}{\partial t} = -\delta m f, \frac{\partial m}{\partial t} = D_m, \nabla^2 m + \mu n - \lambda m, \right. \right.$$

where the first equation describes the diffusion of tumor cells under the action of haptotaxis (second term on the left-hand side) and growth/death (second term on the right-hand side), the second equation describes the degradation of the extracellular matrix, and the third equation the diffusion of matrix degrading enzymes produced by tumor cells.

Models at the Cellular Scale

Looking at the population of tumor cells at the cellular scale, it is possible to realize that the behavior of the cells is not always the same. Their evolution is characterized by the presence of switches and acquired ability that allow them to resist to particularly unfavorable conditions or by

the birth of daughter cells that have a clonal advantage on their mother cells. For instance, in the initial stages of tumor development a corruption in the mechanotransduction pathway leads to the existence of cells that are not so sensitive to the mechanism of contact inhibition of growth, so that they continue activating their duplication program even when they should stop proliferating. This misperception of the compression state by the abnormal cells, called cadherin switch, can generate by itself a clonal advantage on the surrounding cells, leading ultimately to the replacement and the invasion of the healthy tissue by the tumor (see chapter 7 of [7]). After that, other landmarks characterize tumor evolution as the angiogenic switch stimulating the formation of tumor-induced capillary plexus, or the just mentioned metabolic switch from aerobic to anaerobic production of ATP, or the epithelial-mesenchymal transition, that triggers the formation of metastases.

Also from the morphological viewpoint tumor cells present differences that are usually identified by histopathologists to classify tumor malignancy, e.g., presence of particular receptors or organelles, or loss of cell polarity. The mathematical models focusing on the inclusion of the inner evolution of the tumor population briefly described above needs to incorporate a new independent variable which is able to describe the state (or the states) characterizing the cells and how such a state influences their behavior. This state might refer to the aggressiveness, the activity, the maturation state, and so on (see chapter 3 of [6], and [14, 15]).

The key steps in the modeling process are then the selection of the cell populations of interest for the evolution and the identification of their specific activity/ies, their mechanisms of progression, of proliferation, and of interaction.

The structure of the model then develops in the framework of kinetic theories and may be written as follows:

$$\begin{aligned}&\frac{\partial}{\partial t}f_i(t,u)+\frac{\partial}{\partial u}\left[c_i(t,u)f_i(t,u)\right]\\&=S_i+f_i(t,u)\sum_{j=1}^{n}\int_{-1}^{1}p_{ij}(u,w)f_i(t,w)dw\\&\quad+\sum_{j=1}^{n}\left[\int_{-1}^{1}\eta_{ij}(v,w)\Psi_{ij}(v,w;u)\right.\\&\quad\times f_i(t,v)f_j(t,w)dv\,dw-f_i(t,u)\\&\quad\left.\times\int_{-1}^{1}\eta_{ij}(u,w)f_i(t,w)dw\right],\end{aligned}$$

where the second term on the left-hand side represents the intrinsic evolution operator which takes into account of the progression of the cells even in absence of interactions and proliferation/death. The second term on the right-hand side represents the net proliferation operator, i.e., stimulated mitosis minus killing due to interaction with other cells. In it $\eta_{ij}(v, w)$ denotes the number of encounters per unit volume and unit time between cell pairs of the (i, j)-th populations with states v and w, respectively, and $\psi_{ij}(v, w; u)$ denotes the probability of transition of the i-th cell to the state u, given its initial state v and the state w of the encountering cells belonging to the j-th population. The last two terms are related to interactions which only change state in the interacting cells, while the second one refers to proliferative/destructive interactions. Finally S_i is the source/sink term which may be external (e.g., effect of chemotherapy and radiotherapy) or internal, like the net proliferation term (mitosis-apoptosis) $\gamma_i(t, u)\, f_i(t, u)$.

Other popular models focusing on the cellular scale, in general without considering the inner evolution of cells, are based on cellular automata, cellular Potts models, and individual cell-based models [16].

In individual cell-based models, a cell is schematized by its center, then surrounded by a deformable spherical shell or by a Voronoi tessellation (see Part III of [16] for more details). If a spherical cell gets in contact with other cells, it adheres to them and as a result of the contact, the

shape of the cell changes by flattening at the contact area. A similar thing occurs for the interaction with the substrate that is typically a rigid impermeable plane or with other elements of the environment, e.g., bones and vessel walls.

The dynamics of each individual cell is then modeled by Langevin equations in a friction-dominated regime. Thus, in the absence of an external stimulus the cells perform a random movement. For instance, the motion of the i-th cell is modeled by

$$\sum_j \mathrm{C}_{ij}\left(\mathrm{w}_i - \mathrm{w}_j\right) + \mathrm{C}_{is}\mathrm{w}_i = \mathbf{F}_i^{\mathrm{det}} + \mathbf{F}_i^{st},$$

where $\mathbf{w}_i$ are the cell velocities and $C_{i\alpha}$ are related to the friction coefficients, respectively, with between cells ($\alpha = j$) and between cells and the substrate ($\alpha = s$). On the right-hand side, $\mathbf{F}_i^{\mathrm{det}}$ summarizes the deterministic forces related to the total interaction energy between the cell pair (i, j) defined by the sum over all individual energy contributions, that depend on the distance between the cells and the radius of both cells. Thus, cell–cell contacts can equilibrate via cell displacements or changes in the cell radius R. The total interaction energy between a cell and the substrate is defined analogously. The term $\mathbf{F}_i^{st}$ denotes the stochastic force with zero mean and delta-correlated autocorrelation function that models the random component of cell movement.

The cellular Potts model is another modeling framework that has been used to describe several tumor growth problems, such as avascular and vascular tumor growth, angiogenesis, and cell invasion problems (see Part II of [16]). The cellular Potts model is a grid-based, stochastic Monte Carlo method in which each cell is described as an ensemble of nodes. The core driving the evolution of the system is an energy minimization philosophy. Depending on the problem, the energy functionals to be minimized may include adhesion energies between cells or with the substrate, terms mimicking energies, for instance, the energetical response of a cell to a chemical gradient, or persistence of cell motion, and possible constraints, such as on the volume or the surface area of the cells.

The energy functional

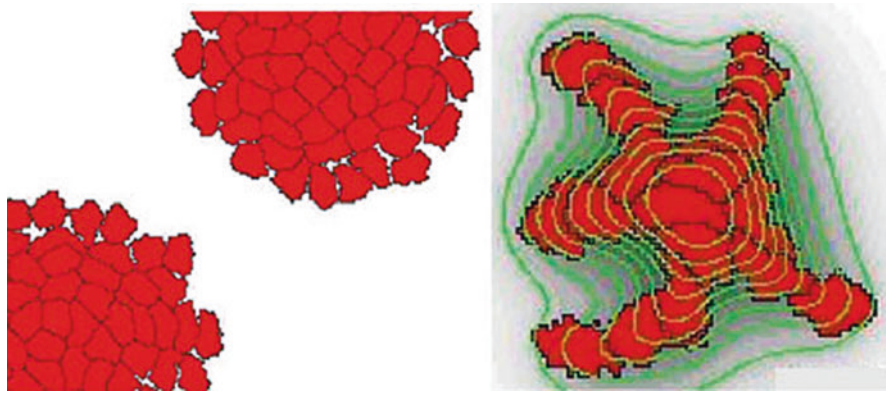
$$H = H_{\mathrm{adhesion}} + H_{\mathrm{shape}} + H_{\mathrm{chemotaxis}} + H_{\mathrm{persistence}} + \ldots\ldots,$$

then describes the state of the system and drives all lattice rearrangements with an algorithm of stochastic minimization (a modified Metropolis procedure).

In this way, it is possible to mimic the natural behavior of migrating cells, via thermal-like membrane fluctuations and extension of pseudopodia to explore the surrounding environment. Procedurally, a lattice site is selected at random and assigned the state of one of its unlike neighbors, which has also been randomly selected. The Hamiltonian of the system is computed before and after the proposed update: If it is reduced as the result of the copy, the change is accepted, otherwise it is accepted with a probability that decreases exponentially with the energy increase that such an unfavorable choice would give. Figure 3 gives an example of a CPM describing the motion of a single cell or of a cell cluster through the mesothelial layer.

Therapy and Multiscale Models

Many models based on systems of ordinary differential equations have been deduced to describe the pharmacokinetics, the pharmacodynamics, or the response of the body, possibly divided in compartments, to cancer treatments. Generally, the aim of these models is the optimization and personalization of therapeutic protocols, identifying optimal doses and optimal schedules. This is usually done using optimal control theory. Similar problems are encountered in testing immunotherapies or combined therapies and in describing the action and the use of radiotherapies [17].



Models for Tumor Growth, Fig. 3 Growing tumor masses (*left*) and invading aggregate of a liver tumor cell line (*right*) as simulated with a cellular Potts model

Focusing here on the recent trends, nowadays there is a big interest in understanding and modeling the action of radiation and drugs at the subcellular level, which means not only how genetic changes or the expression of particular genes or gene networks determine cell behavior, how radiotherapy can induce fatal genetic damages, but also understanding and describing how the protein networks inside the tumor cells work, how the activation of specific signaling cascades govern cell motion, duplication, and death, and therefore what are the main crossroads of the networks to attack or stimulate from the therapeutical viewpoint. Eventually, from the pharmaceutical viewpoint what kind of drug should be developed because it is more likely to be successful.

From the modeling viewpoint, once such networks are understood, the following step consists in inserting them into the macroscopic or the mesoscopic models building what can be called multiscale models, because they take into account of phenomena occurring at different spatial and temporal scales [18, 19]. They, in fact, nest into the macroscopic or mesoscopic representation one or more modules accounting for processes at the microscopic scale. Because of the most recent discoveries in the field of genomics, proteomics, and system biology, these descriptions are now expanding considerably. In this respect, models at the cellular scale, such as individual cell-based models or cellular Potts models, seem more flexible and suited to include subcellular mechanisms, though it is possible to nest protein networks in macroscopic models, so that they influence the coefficients of the partial differential equations.

Coming back to modeling drug delivery, many of the mathematical models dealing with vascular growth were actually developed with the aim of describing the transport of chemotherapeutic drugs, their perfusion through the vessel walls, their effectiveness in killing tumor cells, or as antiangiogenic therapies [11]. This approach has been recently used also to describe the infiltration of engineered macrophages. In fact, it is known that macrophages migrate specifically toward, and localize within, hypoxic tumor regions, that are the hardest region to cure via radiotherapy. So they can be in principle used as vehicles to deliver drugs and perform antitumor activities such as inducing cell lysis or inhibiting angiogenesis. The mathematical models are generally based on advection–diffusion equations with chemotactic terms to describe the directed motion of macrophages toward hypoxic regions.

Very similar models are deduced to describe and optimize drug delivery using other macromolecular carriers like liposomes or suitable nanodevices and nanoparticles. In fact, such macromolecules might more easily reach the tumor sites with respect to other parts of the body exploiting the fact that, because of the action of

tumor angiogenesis factors, tumor-induced capillaries are fenestrated and therefore highly permeable. Moreover, for instance, it is possible to attach an antibody to the liposomes, so that the drug is specifically targeted to the tumor cell. The transport, extravasation, and diffusion of nanoparticle in tumors is also modeled to optimize related therapeutical actions, e.g., magnetic nanoparticles are used so that they can be trapped either to occlude capillaries or to be heated up; gold nanoshells are used to enhance laser thermal therapy [20].

As it might be easily guessed, there is not a single model to approach the same biological phenomenon. Of course, each approach has its advantages and disadvantages. For instance, models based on systems of partial differential equations are more suited and computationally cheaper to describe the behavior of tumor tissues from the macroscopic point of view, but, as discussed above, it is more difficult to account for subcellular mechanisms such as signal transduction, expression, or internalization of receptors, and so on. Conversely, individual cell-based models can focus more closely on the cellular level, but they might become computationally expensive when using a large number of cells, and inefficient in providing a general outlook on the system as a whole. Luckily, it is often not necessary to keep the same level of detail throughout the tissue. The idea behind building what can be called interfacing hybrid models is to use the modeling tools like a microscope, focusing on the cellular or subcellular level where and when needed and blurring when such a detailed description is not needed. For instance, the spatial domain can be partitioned in few time-dependent domains, and in each piece of this moving puzzle a different modeling framework can be used, so that it is possible to take advantage of the positive aspects of both frameworks. Of course, in order to do that one should compare, for the same biological phenomenon, the results obtained via different models and transfer information from one modeling framework to another. Allowing the models to properly cross talk needs the development of suitable mathematical procedures which nowadays represents a challenging research field.

Cross-References

- ▶ Angiogenesis
- ▶ Atomic Force Microscopy
- ▶ ECM
- ▶ Finite Element Methods for Computational Nano-optics

References

1. Wodarz, D., Komarova, N.: Computational Biology of Cancer: Lecture Notes and Mathematical Modeling. World Scientific, Singapore (2005)
2. Araujo, R.P., McElwain, D.L.S.: A history of the study of solid tumor growth: the contribution of mathematical modeling. Bull. Math. Biol. **66**, 1039–1091 (2004)
3. Lowengrub, J.S., Frieboes, H.B., Jin, F., Chuang, Y.L., Li, X., Macklin, P., Wise, S.M., Cristini, V.: Nonlinear modeling of cancer: bridging the gap between cells and tumours. Nonlinearity **23**, R1–R91 (2010)
4. Tracqui, P.: Biophysical models of tumor growth. Rep. Prog. Phys. **72**, 056701 (2009)
5. Preziosi, L. (ed.): Cancer Modeling and Simulation. Chapman & Hall/CRC Press, Boca Raton (2003)
6. Bellomo, N., Chaplain, M.A.J., De Angelis, E. (eds.): Selected Topics on Cancer Modeling: Genesis – Evolution – Immune Competition – Therapy. Birkhäuser, Boston (2007)
7. Mollica, F., Preziosi, L., Rajagopal, K.R.: Modeling of Biological Materials. Birkhäuser, Boston (2007)
8. Preziosi, L., Vitale, G.: New Trends in the Physics and Mechanics of Biological Systems. Lecture Notes of the Les Houches Summer School, vol. 92. Oxford University Press, Oxford (2011)
9. Preziosi, L., Vitale, G.: A multiphase model of tumour and tissue growth including cell adhesion and plastic re-organization. Math. Models Methods Appl. Sci. **21**, 1901–1932 (2011)
10. Mantzaris, N., Webb, S., Othmer, H.G.: Mathematical modeling of tumor-induced angiogenesis. J. Math. Biol. **49**, 111–187 (2004)
11. Chaplain, M.A.J., McDougall, S.: Mathematical Modeling of Angiogenesis. Birkhäuser, Boston (2011)
12. Jain, R.K.: Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy. Science **307**, 58–62 (2005)
13. Quarteroni, A., Formaggia, L., Veneziani, A.: Complex Systems in Biomedicine. Springer, Milano (2006)
14. Bellomo, N.: Modeling Complex Living Systems – Kinetic Theory and Stochastic Game Approach. Birkhäuser, Boston (2008)
15. Bellouquid, A., Delitala, M.: Mathematical Modeling of Complex Biological Systems: A Kinetic Theory Approach. Birkhäuser, Boston (2006)

16. Anderson, A.R.A., Chaplain, M.A.J., Rejniak, K.A.: Single-Cell-Based Models in Biology and Medicine. Springer, Basel (2007)
17. Yakovlev, A.Y., Pavlova, L., Hanin, L.G.: Biomathematical Problems in Optimization of Cancer Radiotherapy. CRC Press, Boca Raton (1993)
18. Cristini, V., Lowengrub, J.S.: Multiscale Modeling of Cancer – An Integrated Experimental and Mathematical Modeling. Cambridge University Press, Cambridge, UK (2010)
19. Preziosi, L., Tosin, A.: Multiphase and multiscale trends in cancer modeling. Math. Model. Nat. Phenom. **4**, 1–11 (2009)
20. Cristini, V., Ferrari, M., Decuzzi, P. (eds.): Nanoparticulate Delivery to Cancerous Lesions: Advances in Mathematical Modeling. Fundamental Biomedical Technologies. Springer, New York (2010)

Modification of Carbon Nanotubes

► Functionalization of Carbon Nanotubes

Modulus

► Interfacial Investigation of Protein Films Using Acoustic Waves

MOEMS/NOEMS

► Nanotechnology

Molding

► Microcontact Printing

Molecular Beam Epitaxy (MBE)

► Physical Vapor Deposition

Molecular Computation

► Molecular Computing

Molecular Computing

Soichiro Tsuda
School of Chemistry, University of Glasgow, Glasgow, UK

Synonyms

DNA computing; Molecular computation; Unconventional computing

Definition

Molecular computing is a novel approach to develop computing systems and devices alternative to conventional silicon-based computing using (biological) molecules.

Overview

Molecular computing research aims at developing computing systems and devices that are fully or partly composed of molecular materials. It typically refers to molecular devices or systems consisting of molecular substrates (mostly biomolecules, such as protein and DNA) for information processing.

Ultimate goals of molecular computing research are not to replace well-established semiconductor materials used for today's computer systems. Molecular computer devices are generally slow, unstable, and hard to connect up together, compared to conventional semiconductor devices. It is less likely that they could be much improved in the future. Instead, a main application domain of molecular computing is the integration with biological systems, such as

biological cells. Molecular computing devices can be directly interfaced with biochemical processes of living cells such as gene regulation network and intracellular signal transduction, whereas semiconductor devices are incompatible with them in terms of size and material. Ultimately, the molecular computing devices would offer more countability to the engineering of biological systems, for example, the controlled release of a therapeutic agent in the cell [1].

In addition to biocompatibility, molecular devices have several other advantages, such as massive parallelism and self-assembly. The former concerns the order in which molecules typically operate. Because molecules in general work in the order of Avogadro number, it has been suggested that theoretically molecular devices which contain a massive amount of molecules can carry out computing in truly parallel fashion, even though each molecule has a little computing power. The latter mainly concerns the free energy minimization process of molecular recognition. If two molecules are complementary, they self-assemble into a supramolecular complex. These advantages lead to DNA computing devices described below, whereas the biocompatibility of molecules bloomed into molecular logic gates, which are also covered here.

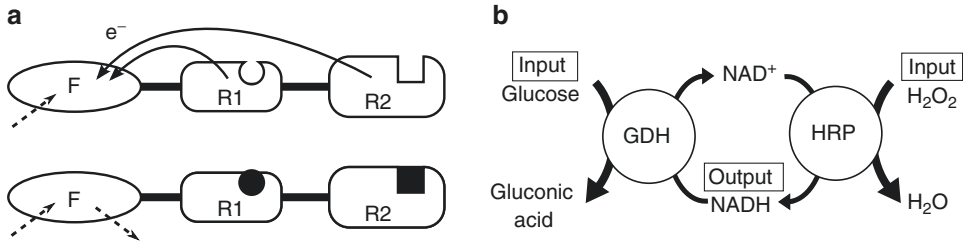
Molecular Logic Gates

In living cells, there are numerous protein "logic gates" operating in order to process external and internal information. An external physical or chemical signal arriving at the cell's surface "turns on" a receptor on the membrane, and the activated receptor releases signals to downstream messenger proteins in a cascade to activate another protein and so forth. In this sense, a living cell can be seen as a system with a vast network of simple switches and logic gates made from molecular materials. Motivated by this view of living cells, research on molecular switch and logic gate aims at designing and developing such molecular computing elements that are controllable from the outside.

A simpler but essential computational element in the electronic devices is the switching device. Tyagi and Kramer implemented a molecular switch in single-stranded DNA using fluorescence resonance energy transfer (FRET). A fluorophore is attached on one end of a single-stranded DNA and a quencher on the other end. The strand folds into a hairpin structure because of complementary sequences in both ends, and the fluorophore and quencher come close together. If a complement single-stranded DNA string is present, they form a double-stranded DNA, which keeps the fluorophore and quencher apart. As a result, the fluorescence is activated. If the fluorescence is considered as output signals, the switch works as a one-input one-output logic gate, NOT gate, which inverts a binary input signal.

Two-input one-output logic gates are also implemented with molecules. De Silva and coworkers developed a molecular AND gate based on photoinduced electron transfer (PET) [2]. The molecule consists of a fluorophore and two receptors (amine and benzocrown ether) connected by spacers (Fig. 1a). Amine binds a proton, and benzocrown ether targets Na^+ as inputs, respectively. If both receptor sites are empty, an electron is transferred to the fluorophore from one of the receptors, and the fluorophore does not emit light. Even when one of the receptor sites is occupied by an input (H^+ or Na^+), an electron is transferred to the fluorophore. Light emission will be observed only when both receptor sites are occupied because electrons are used to bind input H^+ and Na^+ ions. Thus, the AND operation, which outputs "1" only when both inputs are "1," is realized.

Katz and coworkers employed enzyme-catalytic reactions to construct various Boolean logic gates, such as AND, OR, NOR, XOR, YES, and NOT [3]. For instance, the enzymatic XOR gate consists of two enzymes, glucose dehydrogenase (GDH) and horseradish peroxidase (HRP). GDH and HRP take glucose and hydrogen peroxide (H_2O_2) as inputs, respectively, and interact via nicotinamide adenine dinucleotide (NAD), which absorbance changes are used as output (Fig. 1b). A XOR gate outputs "1" when only one input is "1" (i.e., inputs are "0, 1" or "1, 0").



Molecular Computing, Fig. 1 (**a**) A molecular AND gate based on photoinduced electron transfer. F, R1, and R2 represent fluorophore, amine, and benzocrown ether, respectively. *Dotted arrows* indicate excitation and emitted lights. It uses H^+ and Na^+ ions as inputs and fluorescence as outputs (After Ref. [4]). (**b**) An enzymatic XOR gate. If both inputs (glucose and hydrogen peroxide) are either present or absent, the UV absorbance of the system will remain the same as that of before input signal application. Otherwise, it decreases or increases, which is read as output "1" (After Ref. [3])

When inputs are both present or both absent ("0, 0" or "1, 1" case), the small enzyme network keeps balance of NAD, and the ratio of NAD^+ and NADH concentrations remains the same. On the other hand, if the levels of two inputs are not equal, "0, 1" or "1, 0", it loses the balance, and the NADH concentration will decrease or increase. When the optical absorbance of the system with "0, 1" or "1, 0" input is compared to that of a control case (before input signals are applied), the absolute value of absorbance change, particularly in the ultraviolet (UV) range, is quite large, whereas that of "0, 0" or "1, 1" case is almost zero.

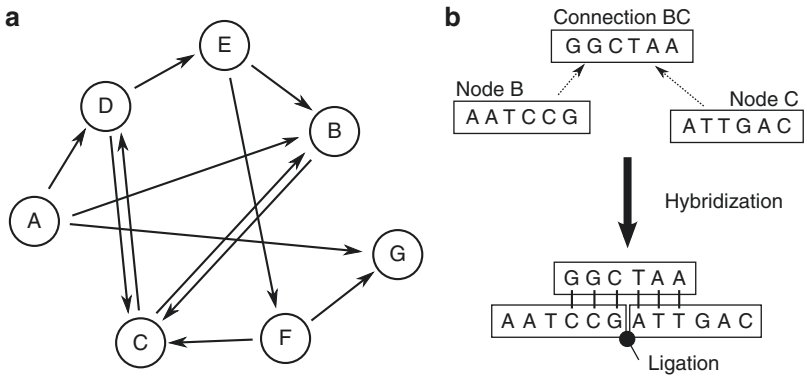
Molecular logic gates can be implemented on various materials (cf. [3, 4]) as well as supermolecules and proteins described above. Those computing elements can be built from supermolecules, macromolecules (e.g., proteins), DNA, and RNA. In some cases, whole cells are also used. Various options are available for inputs and outputs of the systems. As to inputs, molecular logic devices have been developed using light, atomic/molecular ions, magnetic field, redox changes, chelating reactions, oligonucleotides and oligopeptides additions, etc. Output signals can be read by absorbance, fluorescence, electric current, etc. Based on those molecular systems, molecular analogues of digital computing devices can be realized: Boolean logic operations, comparator, flip-flop, arithmetic operations, half adder, full adder, multiplexer, keypad lock, memory unit, and automaton were developed so far.

DNA Computing

DNA computing has become a vast research field which encompasses from formal language theory to practical implementations [5]. This article gives an overview of the first experimental demonstration by Adleman in 1994 [6], which well represents the core ideas of DNA computing.

In the seminal paper, Adleman showed that DNA can be used to solve a type of combinatorial problem called "directed Hamiltonian path problem (directed HPP)." This problem is defined as follows: Given a finite number of nodes and directed connections, determine if there is a path which visits all the nodes only once (Hamiltonian path) in polynomial time. This is known as one of the NP-complete problems, a class of decision problems that can be solved in polynomial time on a nondeterministic Turing machine [7]. In other words, the computational time for solving this type of problems increases exponentially as the size of the problem.

Figure 2a shows a graph Adleman used for DNA computing to solve directed HPP. It has seven nodes labeled as A to G and connections between the nodes. For example, consider if there is a Hamiltonian path starting from node A (in principle, any node can be start of a Hamiltonian path). From A, possible connections are *AB*, *AD*, and *AG*. However, *AG* has no outgoing connections, and it cannot be a Hamiltonian path. Any paths after *AD* are not Hamiltonian paths because they form a loop at some point



Molecular Computing, Fig. 2 (**a**) A Hamiltonian path problem solved by Adleman's DNA computer. (**b**) Encoding of nodes and connections into DNA strands

in the path. By repeating this search process till it finds a path that connects all the nodes without any loop in the path, the path *ABCDEFG* can be determined as only one Hamiltonian path in the graph. The directed HPP can be solved by the brute-force search as above, but with the algorithm, the computational time exponentially increases even when the number of nodes increases linearly.

Adleman tackled this problem by employing the massive parallelism and complementarity of DNA molecules. DNA (deoxyribonucleic acid) consists of nucleotides in polymer form, and each nucleotide consists of a nitrogen-containing base, sugar, and phosphate base. Four types of nucleotides occurring in the DNA are adenine (A), guanine (G), cytosine (C), and thymine (T). Adenine and thymine (A-T) and guanine and cytosine (G-C) can form a base pair with hydrogen bonds, and those pairs are said to be complementary. If a sequence of oligonucleotides (short strand of nucleotides) is complementary with another oligonucleotide, they can form a stable double-stranded DNA.

Path finding of the graph by Adleman's DNA computer was carried out as follows:

1. Encode the graph into the DNA

   Each node and connection was encoded as a single-stranded DNA oligonucleotide. First, each node is represented as a specific length of DNA sequence. In Fig. 2b, nodes B and C are represented as AATCCG and ATTGAC, respectively. Then, if there is a connection between two nodes, for example, connection *BC*, a DNA sequence representing the connection is designed as half of the sequence is complementary to half of the sequence representing node B and another half is to that of the node C sequence (in Fig. 2b, *BC* is encoded as GGCTAA).
2. Generate Random Paths

   Mix all the DNA sequences encoding nodes and connections in a test tube together with DNA ligases. Single-stranded DNA strings self-assemble into double-stranded DNA due to the base pair complementarity. This process is called DNA hybridization. DNA ligases repair breaks in complementary strands of DNA (ligation). At this point, the computation of Hamiltonian path finding has effectively completed. The following steps are required to verify the answer of computation.
3. Extract DNA Sequences Starting with Node a and Ending with Node G

   This step is done using polymerase chain reaction (PCR). Using complementary sequences of node A and node G as primers, PCR amplifies DNA sequences that start with node A and end with G only for easier detection of Hamiltonian path sequences.
4. Extract Sequences with the Length of Exactly Seven Nodes

   Gel electrophoresis separates mixture of DNA fragments with different lengths by their sizes. By comparing with molecular weight markers, an area in the gel which contains DNA sequences with the length of exactly seven nodes can be spotted.
5. Extract Sequences Containing All the Nodes

The last verification process is done by magnetic beads. DNA fragments are first heated and denatured into single-stranded DNA sequences. Then, complementary sequences of a node with magnetic beads attached bind to the single-stranded DNA and picked up by magnets. Repeat this process for each node.

6. Check the Test Tube

   Check if there is any DNA sequence left in the test tube. If yes, it corresponds to a Hamiltonian path of the graph. If not, the graph does not have any Hamiltonian paths starting from node A and ending with node G.

The directed HPP is considered to be a computationally hard problem because a conventional computing system, or conceptually a Turing machine, performs at most one action at a time step, i.e., such a system has only one single processor to process inputs. Thus, if a graph has n nodes and start-end nodes are fixed, potentially it requires $(n-2)!$ computational steps to determine the presence or absence of a Hamiltonian path in the graph. DNA computing, on the other hand, involves a massive amount of simple processors. If a DNA computer has $(n-2)!$ processors, then the computation can be accomplished in a step. It has been suggested that a DNA computer could potentially execute approximately 1.2×10^{18} computations in a second, the energy efficiency could be up to 2×10^{19} computations per joule, and the memory density would be 1 nm^3 per bit, all of which outperform any present-day conventional computing systems.

However, several downsides of DNA computing were also pointed out [1, 8] Error rate. Computations by DNA involve a relatively large amount of errors which inherently come with the use of biological materials. They significantly reduce the reliability of computation, in particular when a number of computations are combined together [2]. Scalability. Provided that Adleman's protocol is employed for problems with larger nodes, the amount of DNA molecules required for encoding exponentially increases, which could be over hundreds of kilograms [3]. Diffusion speed. Even if such a large amount of DNA can be put into place, the computation relies on diffusion of DNA fragments, which does not scale up with the size of DNA. Hence, the reaction speed takes longer for a large-scale problem, but mixing is difficult because DNA molecules are fragile.

Other Molecular Computing Devices

As another approach to exploit computational capabilities of DNA, Winfree proposed an experimentally implementable low-error cellular automata model that can perform Turing-universal computation using the self-assembly capability of DNA tile [9]. When carefully designed DNA strands are mixed in a solution, they self-assemble into two-dimensional shapes according to the designed sequence called DNA tiles. A single-stranded DNA at the edge of a tile ("sticky end") allows the tile to connect only to other tiles which have a complementary sticky end. This "programmability" of DNA tiles allows to self-organize a larger pattern similar to one-dimensional cellular automata (1D-CA), which development is controllable by choosing an initial pattern seed. 1D-CA is proved to have Turing universality, and in this sense, the DNA-based cellular automata system also has the computational power equivalent to a universal Turing machine. Based on this, he and his coworkers have developed tile-based DNA computing devices, such as logic gates [9], automata, gain amplifier, and neural network.

A chemistry-based approach has been pursued by Adamatzky and De Lacy Costello [10]. The main focus in their approach is massively parallel nature of chemical reaction, in particular, Belousov-Zhabotinsky (BZ) reaction. This reaction is known to show excitability: If a BZ solution is stimulated (e.g., by a silver wire), the color of the solution oscillates, and propagating wave patterns are observed. Using the reaction (particularly the propagation waves and their collisions), it has been shown that Boolean logic gates, calculating Voronoi diagram, shortest path planning, robot control can be achieved.

Perspectives

Although there has been a lot of research conducted on molecular computing, this field is still at an early stage of development and is unlikely that this technology would be commonly available in the near future. One of the key issues for molecular computing is the application domain of the technology. It has been claimed that it could be used for life science applications, such as controlled gene regulation. However, there will be many hurdles to develop molecular computing components that work in the complex in vivo environment of living cells. Thus, numerous challenges in the molecular computing research will need to be overcome so that the molecular computing technology will come into wide use.

Another issue for molecular computing devices is the development of computing concepts that make the most of molecular dynamics. Conventional computing concepts are based on the theory of Turing computability, and hence, they are irrelevant to the dynamics of materials on which computing devices are implemented. On the other hand, living organisms arguably employ radically different computing concepts that leverage the physics of biological materials. For example, self-organization, self-repair, and self-reproduction are commonly found in biological systems but are difficult to achieve with conventional computing systems. New computing paradigms, such as cellular automata, neural networks, and evolutionary computing, may provide clues to utilize the complexity of molecular dynamics for information processing.

Molecular computing devices are not likely to compete with and replace conventional computing technology. Rather, they are complementary approaches to silicon-based systems, and a hybrid system which incorporates molecular devices into artificial systems would find potential application domains in, for example, medical and environmental applications. It is, thus, expected that molecular computing will be one of the important technologies, together with other technologies such as synthetic biology, in the age of biomolecular nanotechnology.

Cross-References

- ▶ Computational Systems Bioinformatics for RNAi
- ▶ Nanotechnology Applications in Polymerase Chain Reaction (PCR)
- ▶ Self-Assembly
- ▶ Synthetic Biology

References

1. Machluf, M., Orsola, A., Atala, A.: Controlled release of therapeutic agents: slow delivery and cell encapsulation. World J. Urol. **18**, 80–83 (2000)
2. de Silva, A., Sandanayake, K.R.A.S.: Fluorescent pet (photoinduced electron transfer) sensors for alkali cations: optimization of sensor action by variation of structure and solvent. Tetrahedron Lett. **32**(3), 421–424 (1991)
3. Katz, E., Privman, V.: Enzyme-based logic systems for information processing. Chem. Soc. Rev. **39**, 1835–1857 (2010)
4. de Silva, A.P., Uchiyama, S.: Molecular logic and computing. Nat. Nanotechnol. **2**, 399–410 (2007)
5. Amos, M.: Theoretical and Experimental DNA Computation. Natural computing series. Springer, New York (2003)
6. Adleman, L.M.: Molecular computation of solutions to combinatorial problems. Science **266**(5187), 1021–1024 (1994)
7. Sipser, M.: Introduction to the Theory of Computation. Thomson Course Technology, Boston (2005)
8. Zauner, K.-P.: Molecular information technology. Crit. Rev. Solid State Mater. Sci **30**(1), 33–69 (2005)
9. Winfree, E., Yang, X., Seeman, N.C.: Universal computation via self-assembly of DNA: some theory and experiments. In: Landweber, L.F., Baum, E.B. (eds.) DNA Based Computers II: DIMACS Workshop, pp. 191–213. American Mathematical Society, Providence (1996)
10. Adamatzky, A., Costello, B.D.L., Asai, T.: Reaction–-Diffusion Computers. Elsevier Science, Amsterdam/Boston (2005)

Molecular Dynamics

- ▶ Simulations of Oxide/Polymer Hybrids
- ▶ Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Molecular Dynamics Method

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Molecular Dynamics Simulations of Interactions Between Biological Molecules and Nanomaterials

▶ Molecular Dynamics Simulations of Nano-biomaterials

Molecular Dynamics Simulations of Nano-biomaterials

Melissa A. Pasquinelli[1] and
Yaroslava G. Yingling[2]
[1]Fiber and Polymer Science, Textile Engineering, Chemistry and Science, North Carolina State University, Raleigh, NC, USA
[2]Materials Science and Engineering, North Carolina State University, Raleigh, NC, USA

Synonyms

Molecular dynamics simulations of interactions between biological molecules and nanomaterials

Definitions

Nano-bio materials are materials that are composed of biomolecules (protein, DNA, RNA, lipids) and nanoscale materials (nanoparticles, nanotubes, nanocrystals).

Molecular dynamics simulation is a computer simulation technique that is based on integration of Newton's equations of motion and reflects physical movements of atoms and molecules.

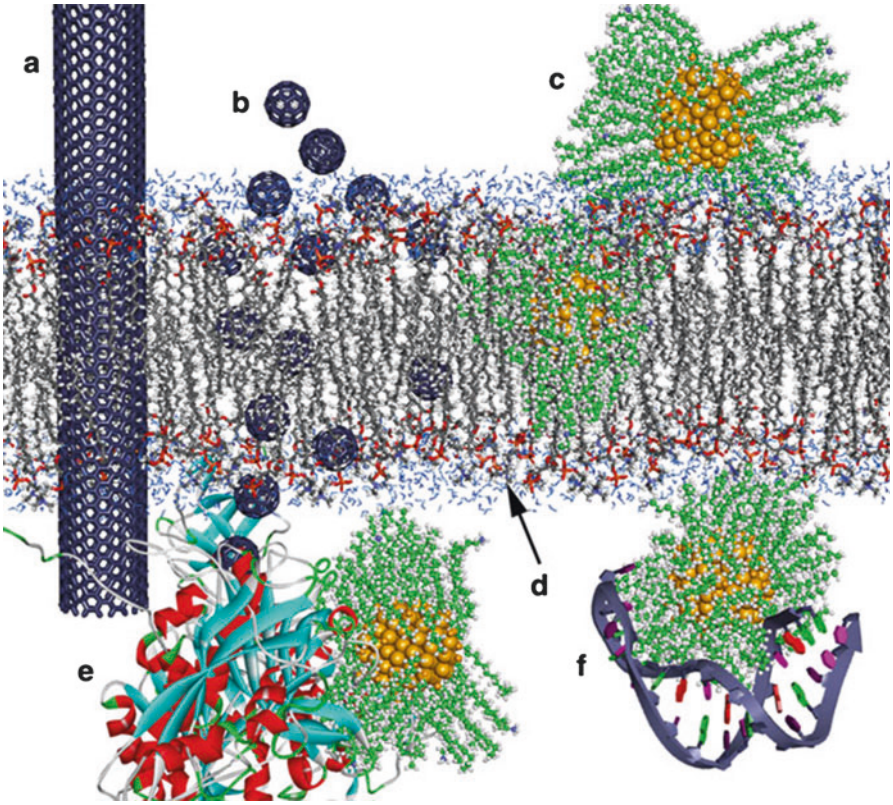
Overview

Introduction

Nanotechnology has received increasing public attention as more and more practical applications of nanoparticles and nanomaterials are emerging, especially in the biomedical field. For example, inorganic nanoparticles (NPs) functionalized with synthetic or biological molecules have been used in a broad range of biomedical applications, including imaging, diagnostics, and drug delivery. The advantage of using NPs in biomedical applications is that a single nanoparticle can simultaneously carry imaging probes, drug molecules, receptor-binding molecules, and antibodies on the surface, which promise the most efficient drug delivery platform and detection scheme. However, interactions of NPs with biological molecules may lead to structural and functional changes, which in turn may result in unexpected biological reactions and toxicity. For example, it has been shown that metallic NPs, quantum dots, and carbon nanotubes can translocate to the brain from different entry points, such as skin, blood, and respiratory pathways.

Moreover, the properties of NPs, such as size, shape, hydrophobicity, charge density, and chemical composition, can directly influence the ability of NPs to enter the cellular environment, bind to proteins, activate receptors, and translocate and accumulate in target tissues, as depicted in Fig. 1. Biological molecules can also be used as scaffolds for precise deposition of NPs for electronic and optical devices. However, the effect of nanomaterials on the biomolecular structure and function is not yet elucidated, which limits advancements within the potential applications and the development of new ones. A better understanding of the biological effects of NPs requires knowledge of the mechanisms of interactions between NPs and biomolecules.

Computer simulations can yield a wealth of insights on interfacial dynamics, structures of ligands, electrostatics, and binding of biomolecules. One of the most promising and rigorous methods for simulating complex and large-scale processes in various materials and biosystems is molecular dynamics (MD) simulations. MD simulations have been used for gaining comprehensive information on "nano-bio" systems, including ligand dynamics



Molecular Dynamics Simulations of Nano-biomaterials, Fig. 1 Atomistic view of the interactions of different types of nanoparticles, (*a*) carbon nanotubes, (*b*) buckyballs (fullerenes), and (*c*) gold nanoparticles, with biological molecules, (*d*) a lipid membrane, (*e*) a protein, and (*f*) nucleic acids

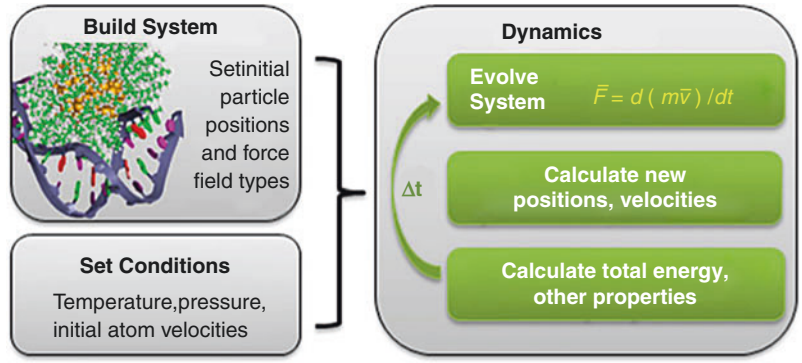
on NPs, conformational transitions of biomolecules, protein and ligand binding, enzymatic catalysis, and the dynamic, thermodynamic, and mechanical properties at different spatial and temporal resolutions [1]. Thus, MD simulations are a natural complement to experimental techniques at the nanoscale to yield insight on molecular interactions involved in nanoparticle–biomolecule binding and mechanisms of molecular recognition.

Overview of Molecular Dynamics Simulations

The steps of an MD simulation are given in Fig. 2. First, the initial system is built; coordinates of some specific biological molecules can be obtained from databases such as the Protein Data Bank [2], while others must be developed using in-house scripts. The initial velocities are assigned from a Boltzmann distribution at the desired temperature. The force field parameters for all of the components are assigned. The system is then evolved with Newton's equations of motion for each Δt, which provides new positions and velocities for each atom.

The potential energy of the system is calculated from force field parameters, which are comprised of both the bonded and nonbonded interactions. The bonded interactions include bond lengths, angles, and dihedrals, whereas the nonbonded terms are mainly electrostatics and van der Waals. The total potential energy can then be defined by the following equation:

$$U = \sum_{\text{bonds}} k1\left(r - r_{eq}\right)^2 + \sum_{\text{angles}} k2\left(\phi - \phi_{eq}\right)^2 + \sum_{\text{dihedrals}} \frac{Vn}{2}\left(1 + \cos\left(n\varphi - \gamma\right)\right) + \sum_{\text{nonbonded}} \left[4\varepsilon_{ij}\left(\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} + \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right) + \frac{q_i q_j}{\varepsilon r_{ij}}\right],$$



Molecular Dynamics Simulations of Nano-biomaterials, Fig. 2 The main components of an MD simulation

where r_{eq} and ϕ_{eq} are equilibration structural parameters, *k1*, *k2*, and *Vn* are force constants, n is dihedral multiplicity, and γ is phase angle for torsional angle parameters.

The choice of a force field is dependent on the chemical composition of the system and the molecules under study, and thus dictate the reliability of the results. A recent review outlines the comparison between most widely used protein force fields [3]. For example, the CHARMM27 [4] force field can be used to simulate DNA, RNA, and lipids, while for CHARMM22 [5] force field is parameterized to simulate proteins. The AMBER program contains force field parameters for DNA, RNA, and proteins (Cornell force field [6]); small molecular ligands (GAFF force field [7]); and carbohydrates (GLYCAM force field [8]). Other all-atom force fields include OPLS [9], which contains parameters for proteins, carbohydrates, and nucleic acids. United atom models such as GROMOS [10] are also available that treat nonpolar hydrogens implicitly, which is commonly done for lipid membranes. The challenge, however, for bio-nano systems is that there is no generalized force field for metallic, carbon, and complex NPs and coarse-grained representations of the molecules, and thus the parameters are calculated on a case-by-case basis. In addition, all-atom MD simulations are not possible for some systems of interest and thus some level of coarse-graining is required, which necessitates a whole new set of force field parameters. An example of a commonly used coarse-grained force field for bio-nano systems is MARTINI [11].

Solvent molecules are often included in the simulation. The most complex solvent is water, and a variety of water models have been developed; the most commonly used models are SPC, TIP3P, and TIP4P. Explicit solvent simulations are the most accurate and can provide information on specific solvent–nanoparticle or solvent–biomolecule interactions; however, they are more computationally extensive since significantly more atoms are being included in the simulations. MD simulations can be significantly accelerated by treating the solvent with continuum dielectric methods. In this case, only the intrasolute electrostatics need to be evaluated. Even though implicit MD simulations are less accurate than simulations with explicit solvent, they permit not only much longer simulations and larger molecules but also provide a variety of sampled conformations.

The time length of simulations is typically based on the type of processes under study and must ensure that a sufficient amount of phase space is sampled. Typically, to observe any significant conformational change of biological molecules, the simulations time needs to be on the order of hundreds of nanoseconds. A useful measure of the accuracy of a simulation is the total energy of the system, which means that typically throughout the simulations, the total energy should not increase. The size of deviations of the total energy gives the measure of precision of the integration. The accuracy of the integration procedure depends on the value of the time step. For the simulations to be as long as possible, the time step needs to be chosen as large as possible. However, the maximum time step is limited by the

Molecular Dynamics Simulations of Nano-biomaterials, Table 1 Summary of the simulations highlighted in this work

	CNT-protein nanobiosensor [12]	C60-potassium channel [13]	NP-membrane [14]	NP-protein [15]	NP-DNA [16]
Level of MD	All-atom	All-atom for C60, proteins; united atoms for lipids	Coarse-grained	Coarse-grained NP; all-atom protein	All-atom
Software	Gromacs	NAMD	Gromacs	CHARMM	NAMD
Force field	Amber99	CHARMM27	Martini	Param22	CHARMM
Electrostatics	Particle mesh Ewald	Particle mesh Ewald	Particle mesh Ewald	Generalized Born and explicit solvent	Particle mesh Ewald
Ensemble	NPT (1 atm., 300 K)	NPT (1 atm., 300 K)	NPT (1 bar, 305 K)	Constant T (300 K, 450 K)	NVT (300 K)
Solvent	TIP3P water + 0.1 M NaCl	TIP3P water	water	TIP3P water	TIP3P + 0.5 M NaCl
Time step (fs)	1.5	2	20	2	2
Simulation time (ns)	10.14–11.25	10–213	60	8	8
Number of atoms	8,200–40,338	90,300–173,000	~65,000	~7,000	~32,000

highest frequency motions in the system, which is typically on the order of a femtosecond.

Advantages and Limitations

The challenges in using an MD description of complex nano-biosystems are the severe limitations of time and length scales, limited to several hundreds of nanoseconds, and moderately sized biomolecules and NPs. However, as computer speed and memory increase, so do the simulation time and size limits. For example, for all-atom simulations the reasonable simulation time 10 years ago was on the order of 1 ns, but now about100 ns time length can be obtained for the same size of a system. The main advantage of an MD simulation is that only details of the microscopic interactions need to be specified and no assumptions are made about the character of the processes under investigation. Moreover, the MD simulation method is capable of providing a complete microscopic description of the dynamical processes involved.

Most importantly, the reliability of MD simulations depends on accurate force fields for all system components. The classical MD simulation method also prevents study of processes like proton transfer, tunneling effects, zero-point motions, behavior of electrons and atomic nuclei, and making/breaking chemical bonds. Problems like these require modeling that applies the laws of quantum mechanics, the fundamental physics equations that describe electrons. However, because of the complex nature of the equations, these models can handle only a few atoms at a time. Some efforts are underway to create force fields that overcome these barriers. Considerable improvements in force fields have also been achieved, making simulations more reliable and accurate.

Application of MD Simulations to Study Nano-Biosystems

In the subsequent sections, an overview of utilizing MD simulations for investigations of a variety of properties of nano-biosystems is provided. Figure 1 summarizes the types of systems that are included in this entry and specific simulation conditions for five simulations highlighted here are summarized in Table 1. The overview begins

with carbon nanoparticles, which are the most actively and successfully used systems for simulations of nanomaterials. Then, the application of MD technique to other nano-biosystems, such as functionalized gold nanoparticles, will be addressed. The focus is on the interactions of these nanoparticles with lipids, proteins, and nucleic acids that have been investigated with MD simulations are then summarized.

Interactions of Carbon NPs with Biological Molecules

The most common type of carbon NPs are fullerenes or buckyballs (Fig. 1b) and carbon nanotubes (Fig. 1a). Both of these carbon NPs are important in many technologically relevant areas due to their unique properties. Buckyballs have potential as drug delivery mechanisms and medical diagnostic devices because of their size, spherical shape, stability, and the ability to be functionalized with other atoms; however, their hydrophobic character is a drawback to some medical applications.

Carbon nanotubes (CNTs) have unique mechanical and electrical properties, as well as strong chemical reactivity due to their high surface area. Their large diameter-to-length ratio provides unique applications, such as microcatheter reinforcements for tissue engineering scaffolds. CNTs can have an impact on biomolecules by enhancing their electrochemical reactivity and promoting protein electron transfer. They are very attractive materials for the design of electrochemical biosensors ranging from an amperometric enzyme electrode to DNA hybridization biosensors. For example, a CNT-based sensor has been developed for detection of hepatitis B virus (HBV), which enables a diagnostic application for nontreated human blood.

A significant amount of MD simulations at various levels of detail have been employed to explore the interactions of CNTs and buckyballs with lipids, proteins, and nucleic acids [17–19]. These simulations have revealed the physicochemical details for a variety of applications. Such focus areas include potential mechanisms for toxicity of carbon nanoparticles and the design of engineering devices such as sensors, drug delivery vehicles, and actuators. Factors for the carbon NPs that were studied include diameter size, structural alignments, aromaticity, nanoparticle curvature, degree of hydrophobicity, and the local environment. Two recent exemplary articles are summarized next.

MD Simulations of a Carbon NP Nanobiosensor

A nanobiosensor has been shown experimentally to be a feasible device for detecting adenovirus proteins, particularly the Knob proteins. This nanodevice is comprised of a single-walled CNT that is functionalized with the coxsackie-adenovirus receptor (CAR) protein. However, many questions remain about the system, including how structural rearrangements and dynamics may impact its efficiency, as well as the specificity of CAR when bound to the CNT outside its native environment. Johnson and coworkers [12] recently performed MD simulations on this nanobiosensor system to provide some molecular insights into these questions. They investigated three different types of systems in its physiological aqueous conditions CAR with a bound Knob protein, and then both systems attached to a CNT. The physiological conditions were created by adding a 100 mM concentration of Na^+ and Cl^- ions to the aqueous system comprised of TIP3P water molecules; an additional three Na^+ ions were added to also neutralize the slightly negative charge on CAR. The CNT was covalently attached to residue Asn98 on CAR through a carboxylic acid defect on the CNT surface, which replicates the experimental conditions that are done via diimide-activated amidation.

The systems were built from a crystal structure of the CAR-Knob complex from the Protein Data Bank. Since CAR is known to only undergo minor structural rearrangements upon binding Knob, this crystal structure was claimed by the authors to be sufficient for both the bound and unbound configurations, after the structural properties were equilibrated from an MD run; structural equilibrium was defined as when the root-mean-square-deviation of the C_α atoms no longer varied with time. The CNT was built the same length as the MD box, connecting terminal atoms across that box boundary, thus creating an infinite tube with the use of periodic boundary conditions.

In addition, the positions of the CNT atoms were constrained with a harmonic potential, which reduces the computational cost of the simulations with minimal effect on the results. The atomistic MD simulations were run with a 1.5 fs time step, for around 11 ns, and with the NPT ensemble at 1 atm and constant temperature of 300 K.

The use of the Amber99 force field was first validated by comparing the simulations on the systems without the CNT in the physiological aqueous environment to known experimental information, such as NMR observations of the structural and dynamical properties of CAR. For the MD simulations of the systems with the CNT present, some interesting observations were made. First, for the systems with the CNT, physisorption of the proteins onto the CNT surface was observed, which still preserved most of the structural features of the unbound CAR and unbound CAR-Knob; the most significant changes occurred in the loop regions and C-terminus since the CNT can restrict the configurational space of the bound proteins. In addition, for the CAR-Knob-CNT complex, a salt bridge and some hydrogen bonds were lost between the proteins, but not enough to impact its overall stability. Therefore, it was concluded that the CNT does not significantly perturb CAR, except for reducing the available configurational space due to confinement. These results indicate that comprised of protein receptors that rely on large conformational changes as part of their mechanisms could be significant.

MD Simulations of C60 Embedded in Potassium Channels

Potassium (K^+) channels are ion channels that span the cell membrane and are selective for K^+. They control a wide variety of cell functions that are involved in biological processes such as nerve impulses, or smooth muscle contraction. The selectivity is due to the so-called selectivity filter, which is a highly conserved sequence. K^+ channels are also targets for a variety of pore-blocking toxins that (1) bind to the extracellular pore entrance and prevent ion conduction; or (2) permeate the pore channel and disrupt binding sites within the pore. Since studies have indicated that fullerenes can cross the lipid membranes of cells, it is expected that they could also impact K^+ channels.

Kraszewski and coworkers [13] recently performed extensive MD simulations to investigate molecular aspects of how K^+ channels are impacted by the presence of the C60 buckyball. They investigated three different types of channels, specifically those comprised of the proteins KcsA, MthK, and Kv1.2 embedded into a fully hydrated DOPC lipid bilayer, which included water molecules and counter ions to neutralize the overall system charge. The total system sizes ranged from 90,300 (KcsA) to about 173,000 atoms (Kv1.2). Up to six C60 molecules were put near the pore region. As a comparison, systems were built of the C60 molecules in a water box and with a fully hydrated DOPC layer without the proteins. The atomistic MD simulations were run with the NAMD software program, using the CHARMM27 force field, where united atoms were used to represent the lipid molecules and chemical bonds between hydrogen and heavy atoms were constrained to their equilibrium values elsewhere.

The C_{60} molecules were first put near the pore on the extracellular region. The MD simulations revealed that no C_{60} molecules anchor or block the selectivity filter of the K^+ channels. In addition, for the voltage-gated K^+ channel, Kv1.2, the C_{60} molecules bound to residues within the voltage sensor domain, which triggered a large conformational movement. When the C_{60} molecules were put near the pore on the intracellular region, a different behavior was observed. Quick binding of the C_{60} molecules was observed, which resulted in conformational changes that plugged the ionic conduction pathway. Finally, when the C_{60} molecules were placed within the membrane core, the C_{60} molecules were observed to migrate laterally toward the membrane proteins, and eventually bind to interhelical domains, which can alter their function. Therefore, these simulations revealed a variety of potential molecular mechanisms for binding of C_{60} nanoparticles, depending on the structure and properties of the channel as well as the initial locations of the nanoparticles.

Interactions of Functionalized Nanoparticles with Biological Molecules

Ligand-protected nanoparticles (Fig. 1c) consists of metallic nanocrystals coated by a monolayer of thiolated molecules. The ligands control the specific properties of the nanoparticle, such as solubility, electro-optical and catalytic behavior. Such nanoparticles can be dispersed in various solvents without irreversible aggregation. Functionalized NPs are a much more challenging case for simulations and modeling than carbon NPs due to their complexity, overall size, and features at multiple length and timescales. In addition, not only are functionalized NPs composed of various organic and inorganic components, but the nanoparticle corona of functionalized ligands is also a dynamic entity by itself. However, MD simulations have been performed to elucidate details about the structure and dynamics of these complex NPs. Recent atomistic MD simulations of NPs functionalized with alkanethiol ligands provided atomistic data on the ligand shell organization and indicated that the ligand's tilt angle is affected by solvent, temperature, nanoparticle size, and the ligand tail length [1, 20].

The complexity of ligand behavior, recognition, dynamic interactions, and assembly between ligand functionalized NPs and biomolecules is not trivial to simulate and challenges the current limits of the MD simulation techniques. To date, most of the simulations that address the interactions between functionalized NP and biological molecules (proteins, ligands, and nucleic acids) have a simplified representation of the nanoparticle and a few simulations have explicit ligands represented via a united-atom approximation, which have provided insights into their biorecognition properties.

A few such studies are summarized next.

Functionalized NPs with Lipids

Properties of functionalized NPs, such as size and type of ligands, can directly control the ability of a NP to enter the cell. The properties that control the NP uptake in cells can cause cellular response and undesirable toxic effects and can also control the performance of NPs in imaging, biosensing, and drug delivery. An understanding of the processes controlling the uptake and accumulations of the NPs by the cell is important for NP design, assessment of nanoparticle toxicity, and improvement of the NP properties for use in imaging and drug delivery applications. However, the rules of uptake kinetics have not yet been determined. In order to gain a deep understanding of NP translocation through a membrane, a complete picture of the processes at the molecular level is needed, which can be delivered through MD simulations. To the best of our knowledge, only coarse-grained MD simulations of NP interactions with the membrane have been used. The advantage of the coarse-grained representation of lipids and NPs is the ability to use larger time and size scales at the expense of observing the intricate behavior of H-bonding network formation and the atomic interactions.

Lin and coworkers [14] recently performed coarse-grained MD simulations of penetration of lipid membrane by gold nanoparticles. They investigated two kinds of lipids, DPPC and DPPG, and NPs with alkyl thiol chains that can contain two different functional group, ammonium and carboxylate, using GROMACS 3.3.2 package [10]. The coarse-grained representation of NPs and lipids was compatible with the MARTINI force field [11]. The gold NP was 2.2 nm in diameter functionalized with 104 alkyl thiol ligands. The NPs were charged with 70 % cationic or anionic coverage represented by the charge on the end functional ligand group and both electronegative and electroneutral bilayers were examined. The coarse-grained force field for the NP was parameterized to represent experimental values of the radius of gyration, the diffusion coefficient, and the average carbon distance. The membrane bilayer was composed of 1,152 coarse-grained lipids in a periodic box filled with explicit water and ions. The initial placement of NPs was chosen to be 7 nm above the bilayer center. The system was carefully equilibrated for a cumulative time of 20 ns before the production run of 40 ns was performed with a 20 fs time step.

The simulations revealed that NP charge and surface change density (electrostatic interactions) control the ability of the NP to penetrate, adhere, or repel from the lipid bilayer. Interestingly, NP penetration into the lipid bilayer causes lipid

disarrangement and formation of a hydrophilic pore that can transport water molecules. The extent of membrane disruption was determined to depend on the NP charge density. Thus, the simulations indicate that control over the NP cellular uptake can be achieved through charge density of the functional groups.

Functionalized NPs with Proteins

The effect of protein interactions with various NPs and nanomaterials can shed light on their possible toxicity and aggregation behavior linked to some neurological disorders. The interface, which is composed of covalent or secondary bonding interactions between NPs and proteins, is highly complex. The properties of this interface may largely affect the protein structure and function. Only a few MD studies were devoted to study protein–NP interactions.

Simulations by Aubin-Tam and coworkers [15] were performed on nanoparticle–protein conjugates comprised of Cyt c from *Saccharomyces cerevisiae* to understand the effect of the NP ligand on protein structure. The NP was modeled as a dummy atom that is 1.5 nm in diameter with Lennard Jones parameters. A total of 21 bathophenanthroline disulfonate (BPS) ligands were uniformly distributed on its surface. The coordinates for the Cyt c protein structure was obtained from the PDB databank (PDB id 2BCN), mutated and covalently linked to the NP surface. MD simulations were performed for 8 ns at 300 and 450 K using CHARMM and param22 all-atom force-fields and the generalized Born implicit solvent. Simulations were also performed in explicit solvent to verify the results in implicit solvent.

MD simulations of Au nanoparticle/Cyt C conjugates indicated that NP labeling induces site-specific, local-to-global structural perturbations to the protein [15]. Protein conformations obtained at 300 and 450 K showed that side chains of positively charged residues interact with BPS ligands, and that perturbation of the secondary structures depends on the NP attachment site, which agrees well with the experimental observations. MD simulations of the semirigid, coarse-grained protein/nanoparticle model revealed the importance of lysine and arginine amino acids on the interactions of Cyt C and nanostructured surfaces, which suggest the design rules for peptides with enhanced surface adsorption [21, 22]. Thus, MD simulations can be successfully employed for identifying mechanisms responsible for fine tuning of the biorecognition properties of the nanocomplex, and more detailed investigations are needed in this area.

Functionalized NPs with Nucleic Acids

MD studies have been successful at corroborating and elucidating experimental findings, or exploring fundamental aspects of DNA dynamics and structure [23]. For example, MD simulations were applied to a 12-base-pair DNA duplex tethered on a neutral silica surface to assess its structural and orientation changes [24]. MD simulations were used to study the conformation of seven DNA strands attached to the (111) gold surface via a six-carbon alkylthiolate linker [25]. These simulations indicated that while the overall structure is disordered, both base stacking and non-Watson-Crick base pairing can be formed between short single-stranded regions of adjacent DNAs. Moreover, the separation distance between DNAs is similar to one calculated by theory and estimated from experiments.

MD simulations of a 2-nm gold nanoparticle functionalized with four single-stranded DNA molecules were used to determine the structure and dynamics of DNA thiolates on a gold surface [16]. The 10-mer of ssDNA polythymines and poly adenines were symmetrically distributed on four (111) gold surfaces of an NP. The interactions between gold atoms were represented by Lennard Jones parameters, and the CHARMM force field was used for the ssDNAs. The resultant DNA-functionalized NP structure was placed in a box with water and 0.5 M NaCl. After equilibration, the NVT Langevin production simulations were carried out using NAMD at 300 K for 8 ns. The results indicated that base–base stacking is better for poly adenines than for poly thymines and the effective radius of DNA-NPs agrees well with experiment. Thus, the MD simulations provided an insight onto the dynamics of DNA strands on the gold NP surface.

In addition, atomistic MD was used to study the behavior of the DNA strands during the assembly of layer-by-layer DNA capsules, which produced a prediction of the optimal length for the NP assembly [26]. Thus, MD simulations were able to provide important information on the behavior of DNA strands on the surface of the NPs; however, the DNA-driven assembly is outside of the reach for atomistic MD simulations. A recent coarse-grained model [27] of DNA-driven NP assembly indicated the possibility to explain the DNA-induced crystallization of NPs and established a general phase diagram.

Summary and Future Directions

The future development of many biomedical devices and the optimization of current systems are dependent on the fundamental knowledge of properties of the interface between NPs and biomolecules. MD simulations are capable of producing such fundamental knowledge, but on limited time and length scales. While the interactions of carbon-nanostructures are significantly explored by MD simulation techniques, the use of MD simulations for other systems such as colloidal gold and magnetic and inorganic nanoparticles is not covered as much. With the increase in computational power and improvements of algorithms, as well as the development of new hardware systems such as GPU systems, it is expected that more MD simulations will be used for detailed studies of interactions between biomolecules and NPs.

Further work is needed on the understanding of interactions between nanomaterials and biomolecules, such as on the surface of inorganic and organic nanoparticles and nanofilms. The investigations at the atomistic and molecular levels will provide information on collective biomolecular behavior and change in properties and will ultimately enable access to new and valuable future applications.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Computational Study of Nanomaterials: from Large-Scale Atomistic Simulations to Mesoscopic Modeling

References

1. Ghorai, P.K., Glotzer, S.C.: Molecular dynamics simulation study of self-assembled monolayers of alkanethiol surfactants on spherical gold nanoparticles. J. Phys. Chem. C **111**, 15857–15862 (2007)
2. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P. E.: The Protein Data Bank. Nucleic Acids Res. **28**, 235–242 (2000)
3. Guvench, O., MacKerell Jr., A.D.: Comparison of protein force fields for molecular dynamics simulations. Methods Mol. Biol. **443**, 63–88 (2008)
4. Case, D.A., Darden, T.A., Cheatham, T.E., Simmerling, C.L., Wang, J., Duke, R.E., Luo, R., Crowley, M., Walker, R.C., Zhang, W., Merz, K.M., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Kolossváry, I., Wong, K.F., Paesani, F., Vanicek, J., Wu, X., Brozell, S.R., Steinbrecher, T., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Mathews, D.H., Seetin, M.G., Sagui, C., Babin, V., Kollman, P.A.: AMBER 10. University of California, San Francisco (2008)
5. MacKerell Jr., A.D., Banavali, N., Foloppe, N.: Development and current status of the CHARMM force field for nucleic acids. Biopolymers **56**, 257–265 (2000)
6. MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiorkiewicz-Kuczera, J., Yin, D., Karplus, M.: All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B **102**, 3586–3616 (1998)
7. Duan, Y., Wu, C., Chowdhury, S., Lee, M.C., Xiong, G.M., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J.M., Kollman, P.: A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. J. Comput. Chem. **24**, 1999–2012 (2003)
8. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., Case, D.A.: Development and testing of a general amber force field. J. Comput. Chem. **25**, 1157–1174 (2004)
9. Damm, W., Halgren, T.A., Murphy, R.B., Smondyrev, A.M., Friesner, R.A., Jorgensen, W.L.: OPLS_2002: a new version of the OPLS-AA force field. Abstr. Pap. Am. Chem. Soc. **224**, U471 (2002)
10. Christen, M., Hunenberger, P.H., Bakowies, D., Baron, R., Burgi, R., Geerke, D.P., Heinz, T.N., Kastenholz, M.A., Krautler, V., Oostenbrink, C., Peter, C., Trzesniak, D., van Gunsteren, W.F.: The GROMOS software for biomolecular simulation: GROMOS05. J. Comput. Chem. **26**, 1719–1751 (2005)

11. Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., de Vries, A.H.: The MARTINI force field: coarse grained model for biomolecular simulations. J. Phys. Chem. B **111**, 7812–7824 (2007)
12. Johnson, R.R., Rego, B.J., Johnson, A.T.C., Klein, M. L.: Computational study of a nanobiosensor: a single-walled carbon nanotube functionalized with the coxsackie-adenovirus receptor. J. Phys. Chem. B **113**, 11589–11593 (2009)
13. Kraszewski, S., Tarek, M., Treptow, W., Ramseyer, C.: Affinity of C-60 neat fullerenes with membrane proteins: a computational study on potassium channels. ACS Nano **4**, 4158–4164 (2010)
14. Lin, J.Q., Zhang, H.W., Chen, Z., Zheng, Y.G.: Penetration of lipid membranes by gold nanoparticles: insights into cellular uptake, cytotoxicity, and their relationship. ACS Nano **4**, 5421–5429 (2010)
15. Aubin-Tam, M.E., Hwang, W., Hamad-Schifferli, K.: Site-directed nanoparticle labeling of cytochrome c. Proc. Natl. Acad. Sci. U. S. A. **106**, 4095–4100 (2009)
16. Lee, O.S., Schatz, G.C.: Molecular dynamics simulation of DNA-functionalized gold nanoparticles. J. Phys. Chem. C **113**, 2316–2321 (2009)
17. Monticelli, L., Salonen, E., Ke, P.C., Vattulainen, I.: Effects of carbon nanoparticles on lipid membranes: a molecular simulation perspective. Soft Matter **5**, 4433–4445 (2009)
18. Makarucha, A.J., Todorova, N., Yarovsky, I.: Nanomaterials in biological environment: a review of computer modelling studies. Eur. Biophys. J. Biophys. Lett. **40**, 103–115 (2011)
19. Ke, P.C., Lamm, M.H.: A biophysical perspective of understanding nanoparticles at large. Phys. Chem. Chem. Phys. **13**, 7273–7283 (2011)
20. Lane, J.M.D., Grest, G.S.: Spontaneous asymmetry of coated spherical nanoparticles in solution and at liquid-vapor interfaces. Phys. Rev. Lett. **10**(4), 235501 (2010)
21. Hung, A., Mwenifumbo, S., Mager, M., Kuna, J.J., Stellacci, F., Yarovsky, I., Stevens, M.M.: Ordering surfaces on the nanoscale: implications for protein adsorption. J. Am. Chem. Soc. **133**, 1438–1450 (2011)
22. Auer, S., Trovato, A., Vendruscolo, M.: A - condensation-ordering mechanism in nanoparticle-catalyzed peptide aggregation. PLoS Comput. Biol. **5**, e1000458 (2009)
23. Mitchell, J.S., Laughton, C.A., Harris, S.A.: Atomistic simulations reveal bubbles, kinks and wrinkles in supercoiled DNA. Nucleic Acids Res. **39**, 3928–3938 (2011)
24. Wong, K.Y., Pettitt, B.M.: A study of DNA tethered to surface by an all-atom molecular dynamics simulation. Theor. Chem. Acc. **106**, 233–235 (2001)
25. Yao, L., Sullivan, J., Hower, J., He, Y., Jiang, S.: Packing structures of single-stranded DNA and double-stranded DNA thiolates on Au(111): a molecular simulation study. J. Chem. Phys. **12**(7), 195101 (2007)
26. Singh, A., Snyder, S., Lee, L., Johnston, A.P.R., Caruso, F., Yingling, Y.G.: Effect of oligonucleotide length on the assembly of dna materials: molecular dynamics simulations of layer-by-layer DNA films. Langmuir **26**, 17339–17347 (2010)
27. Lara, F.V., Starr, F.W.: Stability of DNA-linked nanoparticle crystals I: effect of linker sequence and length. Soft Matter **7**, 2085–2093 (2011)

Molecular Encapsulation

▶ Nanoencapsulation

Molecular Layer Deposition (MLD)

▶ Atomic Layer Deposition

Molecular Modeling

▶ Computational Chemistry for Drug Discovery

Molecular Modeling on Artificial Molecular Motors

Yunfeng Shi
Department of Materials Science and Engineering, MRC RM114, Rensselaer Polytechnic Institute, Troy, NY, USA

Synonyms

Catalytic bimetallic nanorods; Catalytic Janus particle; Nanomotors

Definition

Molecular motors are nanoscale objects capable of converting a certain energy source (chemical energy, photons, etc.) to mechanical work.

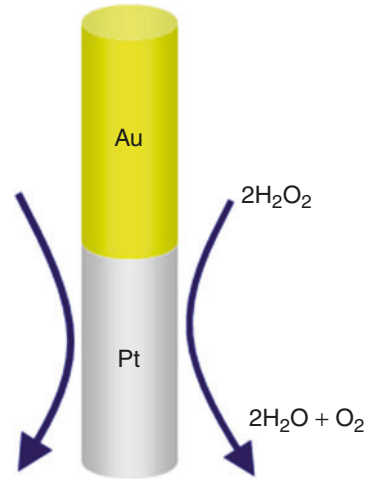
Background

One universal driving force for nanotechnology is miniaturization. The realization of nanometer scaled machineries needs nanometer scaled motors to provide mechanical power for nanofluidic and nanorobotic devices. Another underlying theme for nanotechnology is biomimetic. The fascinating biological world, as optimized by billions of years of evolution, provides answers of science and engineering challenges. In this regard, biomolecular motors play important roles in almost all important physiological processes in a cell, including transportation, actuation, nanostructure assembly, and disassembly. All the aforementioned functionalities are performed at the nanometer scale, thus molecular motors are important enablers in nanotechnologies inspired by biology.

Molecular motors can be classified as active nanostructures (capable of energy conversion or structural transformation upon a local stimuli), as opposed to passive nanostructures (usually operate within a bulk and are not designed to function upon local stimuli). According to Roco's classification, molecular motors belong to the second generation of nanotechnology applications and will play an important role in the third generation (nanosystems) [1].

Survey of Experiments

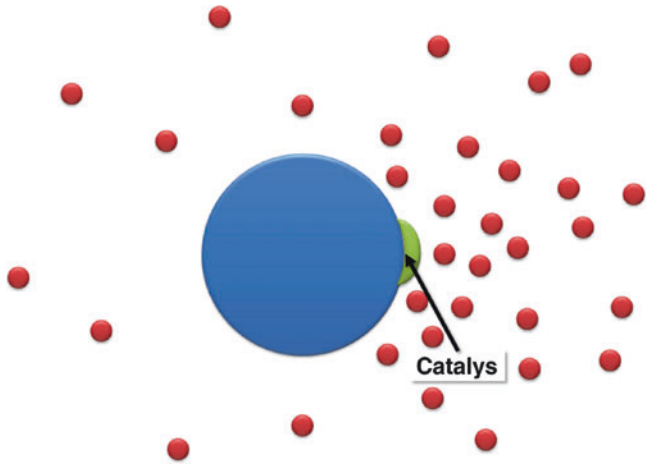
Artificial molecular motors have been fabricated with supramolecules [2], metal nanocrystal, [3] or magnetic colloidal chains [4]. The operation of these motors is mostly powered by an external agitation such as light, an electric/magnetic field, or oscillating chemical concentration. Therefore, unlike biomolecular motors, the motions of the above artificial motors are not strictly autonomous. Recently, it has been shown that catalytic bimetallic nanorods (illustrated in Fig. 1) exhibit interesting self-propelled motion in hydrogen peroxide solutions [5, 6]. Interestingly, these nanorods have comparable size, speed, and energy source (catalytic chemical reactions) to bacteria.



Molecular Modeling on Artificial Molecular Motors, Fig. 1 Illustration of a bimetallic nanorod which relies on decomposition of hydrogen peroxide

Various propulsion mechanisms have been proposed, including bubble recoil [5], interfacial tension [6], self-electrophoresis [7], and self-diffusiophoresis [8]. The driving force for different molecular motors is thought to be system-dependent, and multiple propulsion mechanisms might operate at the same time. For instance, self-electrophoresis seems to be responsible for propulsion in bimetallic nanorod systems, while self-diffusiophoresis is likely to play an important role for systems without electrochemical reactions. An apparent universal feature is to achieve asymmetric reactions by placing the catalysts on one side of the motor or partially blocking the release routes of the products.

The smallest dimension of the catalytic molecular motors investigated in the above studies is about one third of a micron. This is largely due to the technical difficulty in discerning directional motion for objects at submicron sizes [9]. Therefore, it is unclear whether this type of catalytic system can operate when being scaled down to the tens of nanometers (or less) scale. In addition, the thermodynamic efficiency of catalytic molecular motors has not been directly measured. By considering the viscous drag force instead of coupling external loads, Paxton et al. [10] estimated the efficiency of bimetallic nanorods to be on the order of 10^{-9}. The fact that multiple (sometimes



Molecular Modeling on Artificial Molecular Motors, Fig. 2 Illustration of asymmetric product distribution caused by the catalytic patch

competing) propulsion mechanisms operate simultaneously makes it difficult to rationally optimize these nanomotors.

Roles of Simulations

Molecular-level simulations are ideal tool to gain mechanistic details in complex systems, such as the energy conversion process in a molecular motor. In the spirit of taking simulations as "controlled numerical experiments," the goal of molecular simulation is to simplify, instead of reproduce the exact experimental condition, so as to distil the precise mechanism at work. Specific objectives include enhancing the overall energy conversion efficiency and exploring the feasibility of molecular motors that are at least one to two orders of magnitude smaller than current experimental motors. On what follows, three main simulation approaches to model artificial molecular motors based on asymmetric chemistry are described.

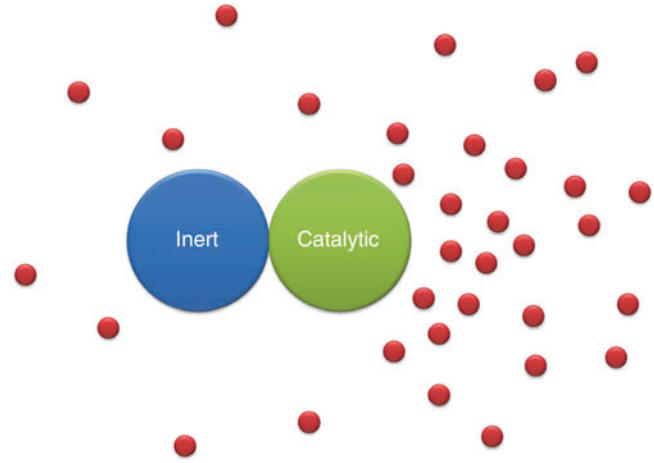
Constitutive Simulation Approach

The first modeling strategy stems from existing treatment of interfacial forces at the continuum level. Although not strictly at the molecular level, this method considers molecular-level events such as reaction and diffusion. Generally, the initial step is to solve the diffusion–reaction equation written below, subjected to the proper boundary conditions such as the motor dimension, catalytic site locations.

$$\partial_t \rho\left(\vec{r},t\right) - D\nabla^2\rho\left(\vec{r},t\right) = \frac{dN_p(t)}{dt}\delta^3\left(\vec{r}-\vec{r}_s\right) \quad (1)$$

ρ is the density of the products, D is the diffusivity, and $\frac{dN_p(t)}{dt}$ is the catalytic reaction rate at site $\vec{r}_s$. As illustrated in Fig. 2, a concentration gradient is established due to the chemical reactions localized near the catalytic site. Thus, products diffuse from high concentration region to low concentration region. Moreover, solvent flow can also occur driven by osmotic forces. Both processes can deform the interfacial layer between the motor–fluid interface and lead to motion of the motor itself.

From the interfacial theory [11], the drift velocity of the motor is proportional to first momentum of the product concentration distribution. In other words, if the product distribution is symmetric, the drift velocity of the motor is zero, and vice versa. As the catalytic site is decorated on the motor, a concentration gradient will be quickly established even though the motor is moving around. Thus, the molecular motor is self-propelled without any external intervention.



Molecular Modeling on Artificial Molecular Motors, Fig. 3 Illustration of a nanodimer consisting of an inert particle and a catalytic particle

Monte Carlo Simulation Approach

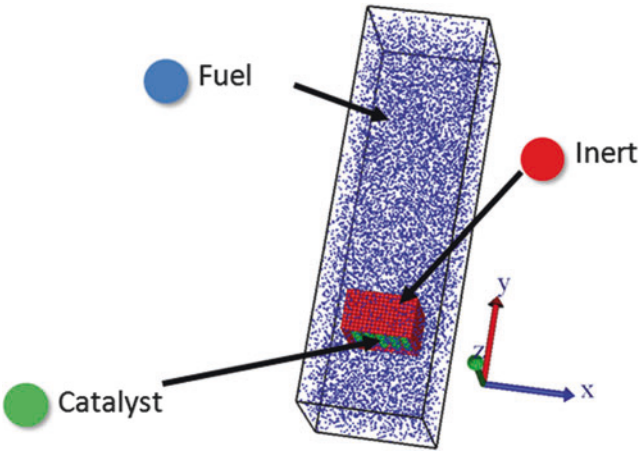
The second approach to treat catalytic chemical reaction is through Monte Carlo (MC) operations, which can be coupled to molecular dynamics (MD) simulations [12]. As illustrated in Fig. 3, a nanodimer consists of a catalytic sphere and a noncatalytic sphere. A collision between the reactant particle and the catalytic sphere will trigger a Monte Carlo operation that will convert it to a product particle at a probability determined by reaction kinetics and temperature. However, neither a collision between the product particle with the catalytic sphere nor collisions between any particles with the noncatalytic spheres will lead to chemical reaction. Thus, the irreversible catalytic chemical reaction can be accomplished.

This method has been used in conjunction with multiparticle collision (MPC) to study the self-motile behavior of nanodimers. MPC is a mesoscopic particle-based simulation method, which is capable of capturing the essential features of MD simulations. As shown in Fig. 3, the product concentration is again distributed asymmetrically around the nanodimer which is a result of the catalytic chemical reaction. By controlling the interparticle interaction, dimer separation, and temperature, the directed motion of the nanodimer can be controlled. The orientational relaxation time for this simulated nanodimer is very long so that the directed motion of the nanodimer is not overwhelmed by the random Brownian motion.

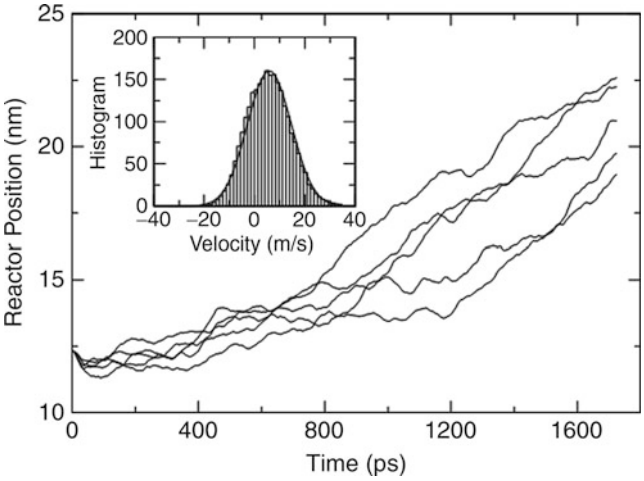
The main disadvantages of this method are the crude representation of chemistry. Particularly, the exothermicity of the chemical reaction is not described in this approach. Moreover, chemical reaction is only possible at the catalytic site. Thus noncatalytic chemical reactions, which are nonetheless present, are not described here.

Reactive Molecular Dynamics Simulation Approach

The third approach is to utilize reactive molecular dynamics simulations to describe the full dynamics of molecular motors. The most important component of this approach is a reactive force field to describe catalytic chemical reaction. Many reactive force fields (REBO [13], ReaxFF [14], etc.) based on the chemistry of bonding have been used to successfully describe hydrocarbon systems, among other systems. However, little work has been done to model catalytic reactions, for instance, the decomposition of hydrogen peroxide with platinum as catalysts. An alternative approach is the recently proposed reactive state summation (RSS) scheme, which does not aim for accuracy but for efficiency and tunability [15]. RSS scheme has been used to describe phase transformations in silica and boron oxide, detonation of nitrogen cubane, growth of nanoporous carbon, and catalytic chemistry in a model fuel-catalyst system [16].



Molecular Modeling on Artificial Molecular Motors, Fig. 4 Illustration of a molecular motor asymmetrically decorated with catalyst



Molecular Modeling on Artificial Molecular Motors, Fig. 5 Displacement of the molecular motor simulated by reactive molecular dynamics. The inset shows the instantaneous velocity distribution which is a shifted Gaussian curve (Adapted from reference [16])

A chemical reaction is accomplished by a transition from one reactive state to another. The reactive state can be characterized conveniently by partial coordination. For instance, oxygen next to just another oxygen is in diatomic oxygen, while oxygen next to just two hydrogen is in a water molecule. The essence of the RSS scheme is to use nonreactive two-body force field to describe individual reactive states, which will be switched on or off by the partial coordination numbers. This switching function can be written as:

$$W_i\left(CN_j\right) = e^{-A_i \sum_{j=1}^{N} \left(CN_j - \overline{CN_j}\right)^2} \qquad (2)$$

Weight function W_i is designed for reactive state i, which is characterized by partial coordination. $\overline{CN_j}$. CN_j is the current partial coordination which can be another reactive state or a transition state in-between. W_i approaches unity if the current reactive state matches reactive state i, while diminishes to zero if it is far away. Thus, the total potential energy of the system can be written as a combination of all nonreactive pair interactions weighted by the above switching function:

$$PE = \sum_{p<q} \sum_{i} W_i\left(CN_j\right) \cdot E\left(r_{p,q}\right) \qquad (3)$$

The first summation loops over all pairs of atoms, while the second summation sums up all reactive state.

Figure 4 shows the model molecular motor inside a fuel environment. This motion can be characterized by a dominant Brownian motion plus a directional motion due to chemical propulsion (Fig. 5). Furthermore, an external force can be coupled to the molecular motor. It was shown that the maximum thermodynamic efficiency of this molecular motor is on the order of 0.0002, which is significantly higher than the experimental bimetallic nanorods [10].

Cross-References

- ▶ Biomimetics
- ▶ Catalyst
- ▶ Nanoparticles
- ▶ Nanorods

References

1. Roco, M.: Nanoscale science and engineering: unifying and transforming tools. AIChE J. **50**(5), 890–897 (2004)
2. Kay, E., Leigh, D., Zerbetto, F.: Synthetic molecular motors and mechanical machines. Angew. Chem. Int. Ed. Engl. **46**(1–2), 72–191 (2007)
3. Regan, B., Aloni, S., Jensen, K., Ritchie, R., Zettl, A.: Nanocrystal-powered nanomotor. Nano Lett. **5**(9), 1730–1733 (2005)
4. Dreyfus, R., Baudry, J., Roper, M., Fermigier, M., Stone, H., Bibette, J.: Microscopic artificial swimmers. Nature **437**(7060), 862–865 (2005)
5. Ismagilov, R., Schwartz, A., Bowden, N., Whitesides, G.: Autonomous movement and self-assembly. Angew. Chem. Int. Ed. Engl. **41**(4), 652 (2002)
6. Paxton, W., Kistler, K., Olmeda, C., Sen, A., St Angelo, S., Cao, Y., Mallouk, T., Lammert, P., Crespi, V.: Catalytic nanomotors: autonomous movement of striped nanorods. J. Am. Chem. Soc. **126**(41), 13424–13431 (2004)
7. Paxton, W., Sundararajan, S., Mallouk, T., Sen, A.: Chemical locomotion. Angew. Chem. Int. Ed. Engl. **45**(33), 5420–5429 (2006)
8. Howse, J., Jones, R., Ryan, A., Gough, T., Vafabakhsh, R., Golestanian, R.: Self-motile colloidal particles: from directed propulsion to random walk. Phys. Rev. Lett. **99**(4), 048102 (2007)
9. Browne, W., Feringa, B.: Making molecular machines work. Nat. Nanotechnol. **1**(1), 25–35 (2006)
10. Paxton, W., Sen, A., Mallouk, T.: Motility of catalytic nanoparticles through self-generated forces. Chem. Eur. J. **11**(22), 6462–6470 (2005)
11. Anderson, J.: Colloid transport by interfacial forces. Annu. Rev. Fluid Mech. **21**, 61–99 (1989)
12. Ruckner, G., Kapral, R.: Chemically powered nanodimers. Phys. Rev. Lett. **98**(15), 150603 (2007)
13. Brenner, D.: Empirical potential for hydrocarbons for use in simulating the chemical vapor-deposition of diamond films. Phys. Rev. B **42**(15), 9458–9471 (1990)
14. van Duin, A., Dasgupta, S., Lorant, F., Goddard, W.: ReaxFF: a reactive force field for hydrocarbons. J. Phys. Chem. A **105**(41), 9396–9409 (2001)
15. Shi, Y.F., Brenner, D.W.: Simulated thermal decomposition and detonation of nitrogen cubane by molecular dynamics. J. Chem. Phys. **127**, 134503 (2007)
16. Shi, Y.F., Huang, L.P., Brenner, D.W.: Computational study of nanometer-scale self-propulsion enabled by asymmetric chemical catalysis. J. Chem. Phys. **131**, 1 (2009)

Molecular Patterning

▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Molecular Plasmonics

▶ Nanopolaritonics

Molecular Printing

▶ Nanoscale Printing

Molecularly Thick Layers

▶ Self-Assembled Monolayers for Nanotribology

Monolayer Lubrication

▶ Self-Assembled Monolayers for Nanotribology

Monolithic Integration of Carbon Nanotubes

▶ CMOS-CNT Integration

Monolithic Total Internal Reflection Biosensor

▶ Whispering Gallery Mode Resonator Biosensors

Mosquito Fascicle Inspired Microneedles

▶ Biomimetic Mosquito-Like Microneedles

Moth-Eye Antireflective Structures

Stuart A. Boden[1] and Darren M. Bagnall[2]
[1]Nanoelectronics and Nanotechnology Research Group, Electronics and Computer Science, University of Southampton, Highfield, Southampton, UK
[2]School of Photovoltaic and Renewable Energy Engineering, The University of New South Wales, Sydney, NSW, Australia

Synonyms

Arrays; Biomimetic; Nanostructured antireflective surfaces; Subwavelength

Definition

Moth-eye antireflective (AR) structures are arrays of protuberances, with dimensions smaller than the wavelength of the light incident upon them. They form a region of graded refractive index at the interface between two media, substantially reducing the amount of light reflected by the interface.

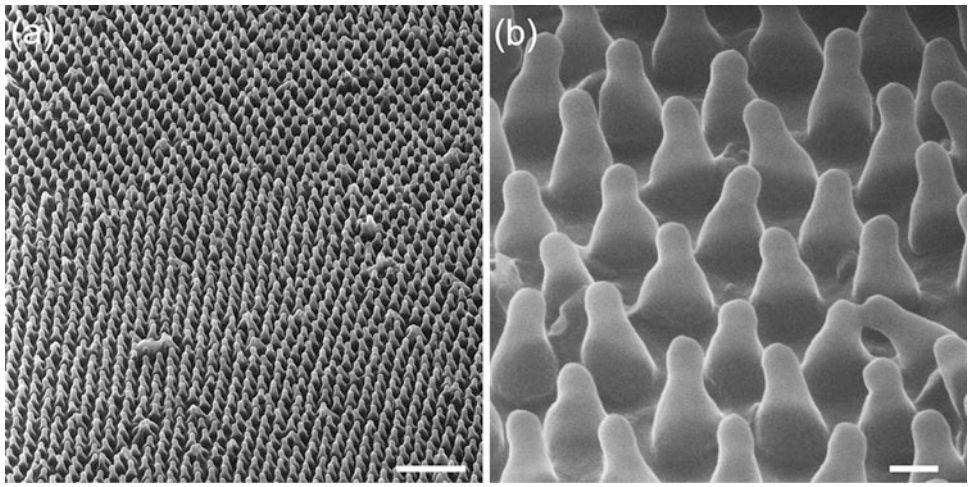
Such protuberance arrays are found on the eyes of some species of moth, butterfly, and fly, where the reduction in reflectance is thought to increase visual system sensitivity, and on transparent sections of insect wing where they have a camouflaging effect. Biomimetic, artificial moth-eye AR structures have been fabricated on a range of materials such as polymers, glass, silicon, and GaAs for applications including photovoltaics and display technologies.

Natural Moth-Eye Structures

From a nanotechnology viewpoint, moth-eye antireflective structures generally refer to artificial structures fabricated to increase the amount of light coupled through the surfaces of optical devices. However, the term originates from the natural world where such structured arrays were first discovered on the corneal surfaces of some species of moth and termed "corneal nipple arrays" [1]. Since then, nipple arrays have also been found on transparent wings of moths and the corneal surfaces of butterflies [2, 3]. Similar antireflective structures have also been identified on the eyes of some species of fly and mosquito [4, 5]. Natural moth-eye structures consist of arrays of protuberances with heights up to 250 nm and spacings of 200–250 nm arranged in domains where, within each domain, the features are arranged in a highly ordered hexagonal array (Fig. 1). The structures reduce the reflectance of the lens surface on which they are formed from approximately 4 % to less than 1 % [3].

Development of nipple arrays on the corneal surfaces of moths for antireflection is thought to serve two purposes:

1. The reduction in reflection is accompanied by an increase in transmission of light into the eye, enhancing the sensitivity of the visual system of the moth.
2. Reflection of sunlight from the eyes could expose the location of the moth to passing



Moth-Eye Antireflective Structures, Fig. 1 Scanning helium ion microscope images of natural moth-eye structures found on the transparent wing of *Cephonodes hylas* (scale bars: (**a**) 1 μm and (**b**) 100 nm)

predators. By reducing the reflectance of light, the moth-eye structures reduce the visibility of the moth, making it less vulnerable to predation.

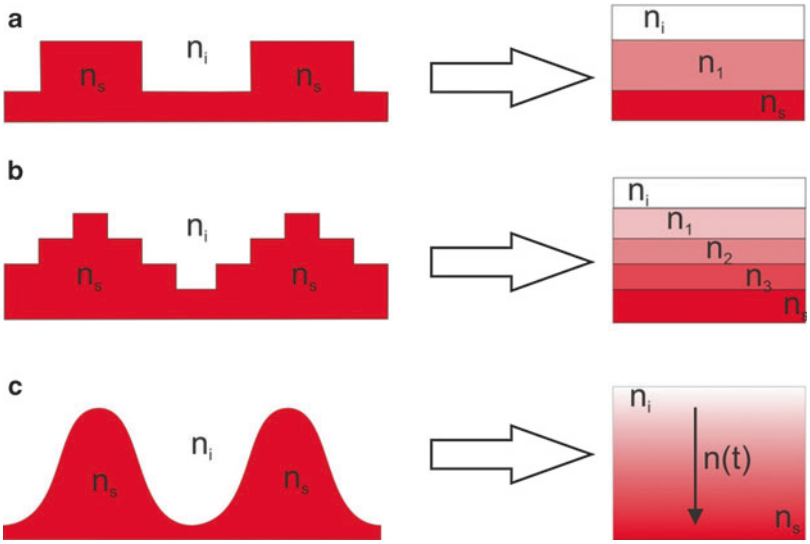
For nipple arrays found on winged surfaces, only the camouflage hypothesis applies. Some species of moth, for example, *Cryptotympana aquila* and *Cephonodes hylas* (Fig. 1), employ nipple arrays on both dorsal and ventral surfaces of transparent sections of the wing, increasing optical transmission to above 90 % over a broadband of wavelengths, with a peak of ~98 % [6], making the creature practically invisible to any passing predators. The protuberance arrays have also been shown to confer superhydrophobicity to a surface, which leads to antifogging, anti-adhesive, and self-cleaning properties, exhibited by a range of species [5, 7].

Inspired by the effectiveness of natural moth-eye structures for reducing surface reflection, researchers have created biomimetic, artificial moth-eye structures in technologically important materials such as silicon and glass, in attempts to lower the amount of light reflected at the surfaces of these materials to a greater extent than is possible with traditional thin-film antireflection coatings (ARCs).

Optical Mechanism

The mechanism through which moth-eye structures reduce reflection relies on their dimensions being less than the wavelength range of incident light. A large proportion of the solar spectrum lies within the visible range (0.4–0.7 μm). When light interacts with structures of dimensions much below this range, it behaves as if it were encountering a homogeneous medium whose optical properties are a weighted spatial average of the profile's optical properties. If the subwavelength features are regular in size and arrangement, the array is known as a zero-order diffraction grating because all the higher orders are evanescent and only the zero-order propagates. Consequently, a surface textured with ridges smaller than the wavelength of light will interact with the light as if it had a single-layer ARC of refractive index governed by the ratio between the ridges and channels (Fig. 2a). Likewise, a stepped profile will act as a multilayer ARC (Fig. 2b), and a tapered profile will behave as an infinite stack of infinitesimally thin layers, introducing a gradual change in refractive index from one medium to the other (Fig. 2c).

The moth-eye array effectively acts as an impedance-matching device smoothing the transition between one medium and another, ensuring



Moth-Eye Antireflective Structures,
Fig. 2 Subwavelength structures and their analogous refractive index profiles, as experienced by incident light: (**a**) ridged profile, (**b**) stepped profile, and (**c**) tapered profile (Adapted from Ref. [8])

that incident light does not encounter a sudden change in refractive index which would cause a proportion to be reflected. Another way to understand this is to imagine that the subwavelength-structured surface is a stack of thin layers, and so the destructive interference between reflections from each layer cancels out all reflected light, maximizing the proportion of light transmitted.

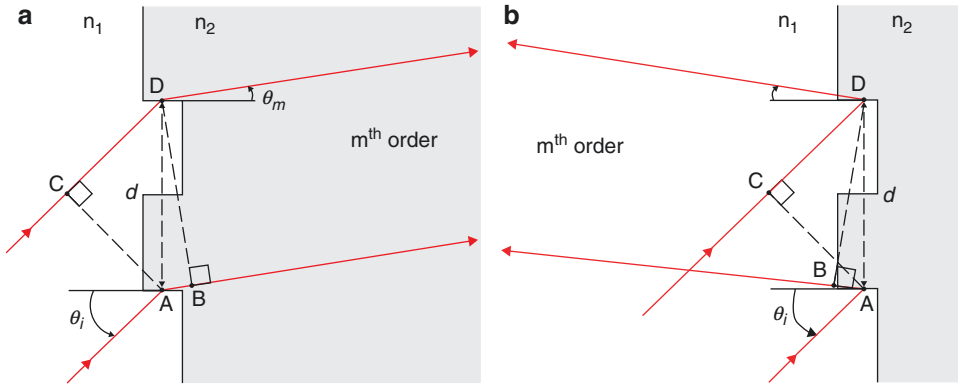
Theoretical Work

The qualitative description given above of the graded refractive index mechanism through which moth-eye structures reduce reflection from a surface suggests that such structures can be modeled by replacing the subwavelength-structured interface region with a stack of many layers, each with an effective refractive index in between those of the media either side of the interface. This is the basis of an approach known as effective medium theory (EMT) or homogenization [9]. The effective refractive index depends on the relative amounts of each medium in each layer which is determined from the profile of the graded index region. Studies of how the reflectance varies with the height-to-wavelength ratio show that if the graded index region is thicker than approximately half the wavelength, reflectance is reduced to below 0.5 % [10]. One way to proceed with this is to establish an optimum refractive index profile of a graded index ARC for minimum reflection over the broadest bandwidth. In one such optimization, this was demonstrated to be a quintic function for all coatings with thicknesses, t, greater than approximately half the wavelength of incident light [11]:

$$n = n_i + (n_s - n_i)\left(10t^3 - 15t^4 + 6t^5\right) \qquad (1)$$

where n_i is the refractive index of the incident medium and n_s is that of the substrate. EMT was then applied to determine the profile of features needed to produce this ideal gradient in various materials.

For EMT to apply however, the features at the interface must be sufficiently small so that they cannot be resolved by the incident light. In other words, the features should not be large enough to scatter light appreciably, either incoherently, as with randomly arranged features, or coherently, as with regular arrays of features such as those found on moth-eye surfaces. To analyze coherent scattering from a 2D moth-eye type array, it is useful to treat it as a diffraction grating with a period small enough so that all diffraction orders other than the zeroth are suppressed. To determine how small the period, d, must be for a zero-order grating, the well-known grating equation for an arbitrary angle of incidence can be used. This is



Moth-Eye Antireflective Structures, Fig. 3 Diagrams for the derivation of the arbitrary-angle grating equation describing diffraction from a grating at the interface between two media of refractive indices n_1 and n_2 (**a**) in transmission and (**b**) in reflection

derived by equating the path-length difference of two rays incident on adjacent features with an integer number of wavelengths so that they constructively interfere. In transmission, this is given by (see Fig. 3a)

$$n_2 AB - n_1 CD = m\lambda \tag{2}$$

$$n_2 \sin\theta_m - n_1 \sin\theta_i = \frac{m\lambda}{d} \tag{3}$$

where m is the order number (an integer), λ is the wavelength, θ_i is the angle of incidence, and θ_m is the outgoing propagation angle for order m. In reflection (Fig. 3b), light is only traveling in one medium so the grating equation becomes

$$\sin\theta_m - \sin\theta_i = \frac{m\lambda}{n_1 d} \tag{4}$$

If the incident medium is air, then $n_1 = 1$. As d is decreased, θ_m increases, with diffraction orders becoming evanescent when $\theta_m = 90°$ for the last remaining order. For all values of d below this, only the zero-order propagates. At normal incidence, $\theta_i = 0°$, and the last order to disappear is the first order, $m = 1$ (or $m = -1$). The zero-order grating condition is then given (for the transmission grating) by

$$d < \frac{\lambda}{n_2} \tag{5}$$

and for the reflection case, it is given by

$$d < \lambda \tag{6}$$

For oblique incidence, if the angle of incidence is positive, the -1 order may still propagate even when d is sufficiently small to suppress the $+1$ order. Therefore, d has to be decreased further to suppress the -1 order. So for non-normal incidence, the zero-order grating condition for reflected orders becomes

$$d < \frac{\lambda}{2} \tag{7}$$

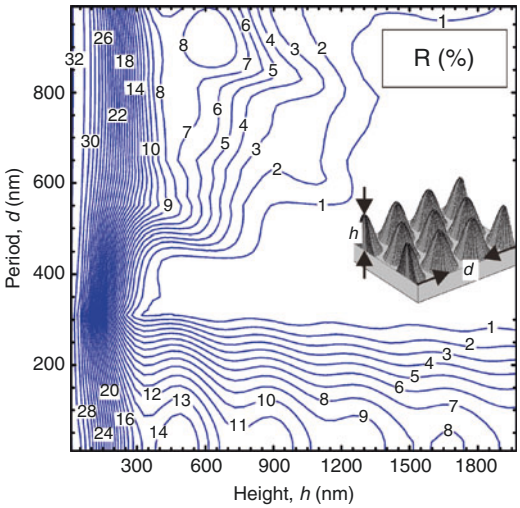
To also suppress *transmitted* orders other than the zeroth, the condition to be satisfied is

$$d < \frac{\lambda}{n_2 + 1} \tag{8}$$

When working in the infrared region, it is relatively easy to fabricate sufficiently small features to satisfy these conditions and produce antireflective, zero-order gratings which can be analyzed using EMT [12]. Fabricating a zero-order grating with $d << \lambda$ to operate in the visible range however is more difficult, and so subwavelength AR schemes for visible wavelength applications tend to have periods closer to the operating wavelengths. EMT breaks down

as the resonance regime is approached (i.e., as $d \rightarrow \lambda$). This, coupled with the increase in computational power available to researchers, led to the rise in popularity of rigorous methods of modeling such structures. Finite-difference time-domain (FDTD) and finite element methods (FEM) have been used, but the most commonly employed modeling method for moth-eye structures is known as rigorous coupled-wave analysis (RCWA) [13]. Developed in the early 1980s, RCWA is a differential-based method to calculate the efficiencies of orders in diffraction gratings. Differential solutions to Maxwell's equations applied to diffraction gratings involve expressing the field in the grating region due to the periodicity present as a Fourier series representation, from which a set of differential equations are developed. RCWA uses an eigenvalue numerical technique to solve these for rectangular grooves. More complicated profiles can be analyzed using a staircase approximation, representing the shape as a series of blocks. By defining the period to be less than the wavelength and representing the feature profile with a staircase approximation, RCWA can be used to model regular subwavelength arrays such as moth-eye structures.

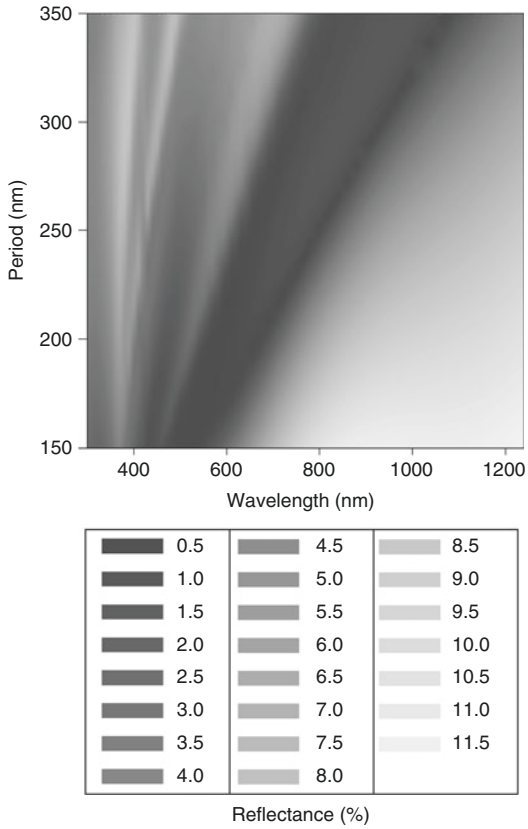
RCWA is widely used to model artificial subwavelength AR structures with good agreement between theory and experiment [14]. The successful experimental verification of RCWA simulation results has meant that the technique can be used to explore the parameter space for artificial moth-eye structure design. It has been shown that pillar height, period (spacing), and shape are all key factors for the AR performance of the moth-eye structure. Parameter maps can be generated of array period versus height or aspect ratio for a single wavelength (Fig. 4) [15, 16]. This type of analysis reveals that for small periods ($d << \lambda$), the structures behave collectively as an effective medium creating a refractive index gradient: Increasing the height in this region initially leads to a sharp drop in reflectance which then remains low but with low-amplitude interference fringes due to discontinuities at the boundaries of the structure, as predicted by EMT. For periods approaching the wavelength, more complicated resonance behavior is observed. Between



Moth-Eye Antireflective Structures, Fig. 4 Parameter map from RCWA simulations of a silicon moth-eye structure showing reflectance (R) versus height and period at a wavelength of 1000 nm [15]

these two regimes exists a low reflectance band at relatively modest feature heights, representing a "sweet spot" for artificial moth-eye designs. In general, for all periods, increasing the height decreases the reflectance, but this comes at the expense of difficulties in the experimental realization of high aspect ratio features. It is a case of diminishing returns for feature heights above approximately half of the wavelength which explains why natural examples of AR nipple arrays generally have heights <250 nm. Although AR with moth-eye structures is a broadband effect, there are local maxima and minima in the reflectance spectra of such arrays, and the position of these can be tuned by varying the period of the array (Fig. 5). In addition, RCWA simulations of moth-eye arrays with different feature profiles indicate that the biomimetic, smoothly tapered design is closer to optimum than arrays comprising sharper or flat-topped features.

Efforts have been made to use RCWA to optimize the design of moth-eye arrays for specific applications, for example, AR for solar cells. One such study showed that a biomimetic moth-eye structure fabricated on a monocrystalline silicon solar cell with a period of 312 nm would be optimal for reducing reflection of the AM 1.5



Moth-Eye Antireflective Structures, Fig. 5 Contour plot, generated by RCWA simulations, of reflectance spectra as the period of the silicon moth-eye array is varied for a feature height of 400 nm [17]

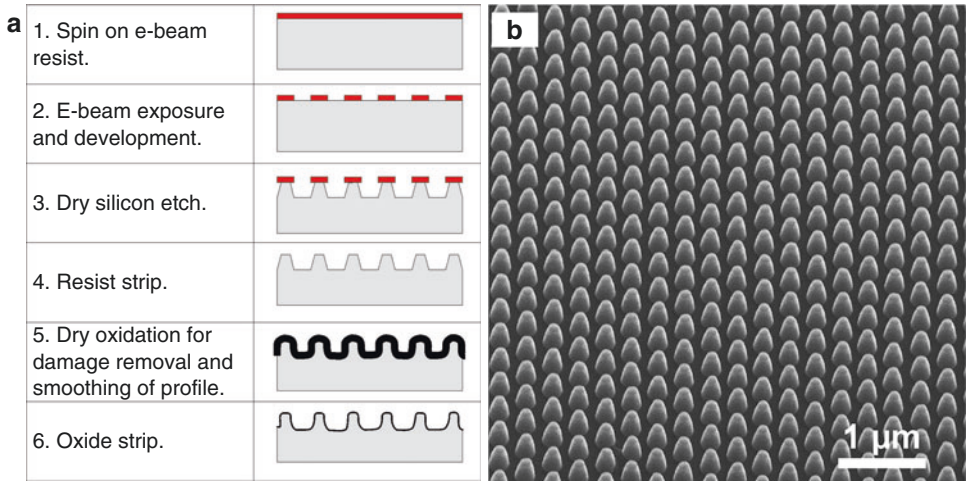
solar spectrum at the silicon-EVA interface (EVA is a typical solar cell encapsulant material) [17]. Calculations on the proposed structure predicted a daily average short-circuit current that was only 0.6 % less than that of a cell with an ideal AR surface (0 % reflectance). This represents around a 3 % improvement over the best double-layer thin-film antireflective coating and demonstrates the potential of moth-eye AR structures for solar cell applications.

Practical Realization

The effectiveness of nature's solution to unwanted reflection in the form of moth-eye structures led to efforts to reproduce such structures for technological applications. As described previously, moth-eye structures designed for visible wavelengths have spacings on the scale of several 100 nm, too small for a suitable mask to be fabricated using standard photolithography techniques. A range of other nanoscale fabrication techniques have been employed to create artificial moth-eye AR structures in a variety of materials. In general, this involves using electron-beam lithography or interference lithography to define a pattern in a polymer film on top of a hard substrate. Etching is then performed to transfer the pattern into the hard substrate. This could be the device itself or could be used as a master to form moth-eye structures in a polymer film by nanoimprinting. Polymer moth-eye structures are often used as an AR surface for glass by careful choice of the polymer so its refractive index matches closely with that of the substrate. However, for other materials such as silicon and GaAs, the lack of polymers with a sufficiently high refractive index means that a further etching step is required to transfer the structure into the underlying substrate to form an effective AR surface.

Electron-Beam Lithography

Although the wavelength of light prevents standard photolithography from being used to form moth-eye arrays for visible operation, sub-100 nm patterning with electron-beam systems is routine. A focused beam of electrons is scanned over a layer of polymeric resist in the pattern required. The resist is developed, leaving behind the exposed resist in the form of arrays of dots. Dry etching is then used to transfer the pattern into the underlying substrate (see Fig. 6). Etch resistance of polymers is relatively low; therefore, in order to achieve sufficient feature height, e-beam sensitive flowable oxides are sometimes used or an intermediate oxide or metal mask is employed. In the latter case, this can be deposited on the substrate before resist spinning and then etched through the resist mask, or alternatively a negative pattern can be formed in the resist and used in a subsequent lift-off procedure to leave a metal mask of the required pattern. Etch chemistries are chosen for the particular material needing to be structured,



Moth-Eye Antireflective Structures, Fig. 6 (**a**) Schematic of silicon moth-eye fabrication process using electron-beam lithography and dry etching and (**b**) scanning helium ion micrograph of silicon moth-eye structure fabricated using electron-beam lithography and dry etching

with a level of anisotropy and selectivity often required in order to achieve sufficient feature height. For patterning silicon, SF_6, HBr, and Cl_2/BCl_2 are often used, followed by an isotropic etch or an oxidation and oxide strip to further shape the features into the required moth-eye profile. In other examples, CHF_3 and $C_4F_8/CH_2/O_2$ mixtures have been used for silica moth-eye structures and $SiCl_4/Cl_2$ for GaAs.

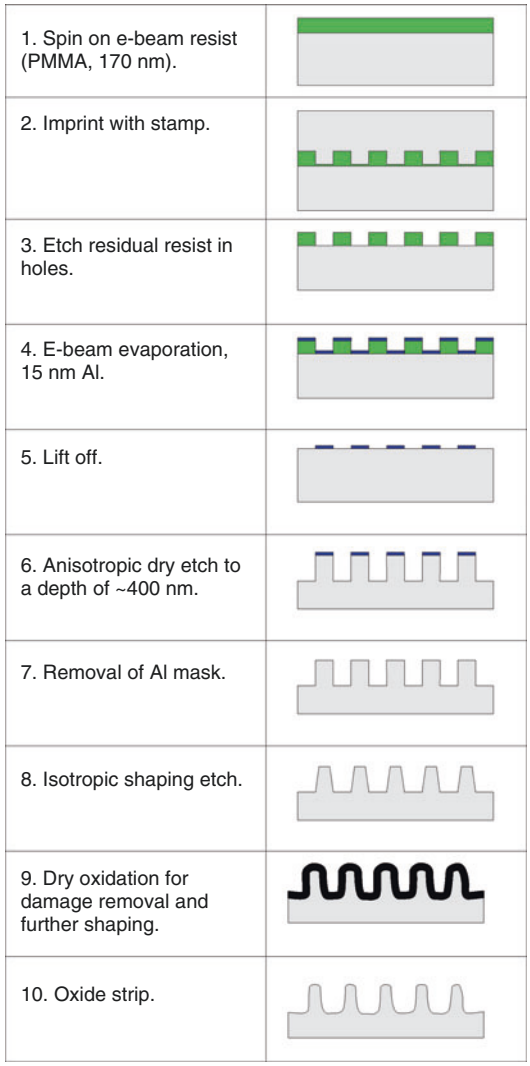
Electron-beam lithography allows precise control over feature size, spacing, and arrangement, but the electron beam has to trace out each individual feature in series. This, combined with the required density of the patterns, means that it is a time-consuming and so expensive fabrication method. Patterning times vary considerably with pattern density and resist sensitivity but are of the order of 1 mm^2 per hour, making this technique unsuitable for commercial, large area moth-eye fabrication.

Interference/Holographic Lithography

The interference pattern formed between two or more laser beams consists of a periodic series of fringes with a spacing less than the wavelength of the laser and so at a scale suitable for moth-eye array fabrication [18]. This interference pattern can be used to expose a layer of photoresist deposited on the surface of a material which is then developed leaving a mask for a subsequent dry etch process. Systems employing a combination of three or four beams can be used to create the two-dimensional patterns required for moth-eye fabrication, but more often, a simpler arrangement of multiple exposures with two beams is used with sample rotation between exposures. Heat treatments can be used between photoresist development and etching to reflow the resist and improve the uniformity and symmetry of the pattern. Careful control of exposure dose and development time is necessary to produce patterns sufficiently dense to form effective AR structures [19]. Interference lithography is a fast and relatively cheap method of fabricating subwavelength structures over large areas and has been successfully implemented to form artificial moth-eye arrays in a range of materials including polymers, glass, silicon, and GaAs.

Nanoimprinting/Nanoimprint Lithography

A more scalable nanopatterning technique for moth-eye structure fabrication has emerged in recent years in the form of nanoimprint technologies [8]. A stamp, fabricated by conventional nanopatterning techniques (e-beam or interference lithography), is pressed into a polymer film



Moth-Eye Antireflective Structures, Fig. 7 Schematic of silicon moth-eye fabrication process using nanoimprint lithography [8]

on a heated substrate. After cooling, the stamp is removed, leaving an imprint of the stamp in the polymer film. For patterning of hard materials, a dry etch can then be performed, using the imprinted polymer film as an etch mask, to transfer the pattern into the substrate. Alternatively, metal is deposited and the imprinted polymer film is used as a lift-off mask, leaving a metal etch mask (Fig. 7). A related technique known as step-and-flash imprint lithography involves the use of a transparent stamp through which UV light is directed to cure the polymer before the stamp is released. The stamp can be reused many times; a protective anti-stiction layer is applied to ensure the easy release of the stamp from the polymer and to reduce wear of the stamp. The scalability of the technique is further enhanced when flexible stamps are used. For example, moth-eye structure fabrication over the entire surface of a silicon wafer has been achieved using nanoimprint lithography with a soft roller-type stamp, followed by a dry etch [20]. With the development of large-scale roll-to-roll processes, this technique has the potential to provide commercially viable moth-eye structures covering large areas required for, for example, solar cells.

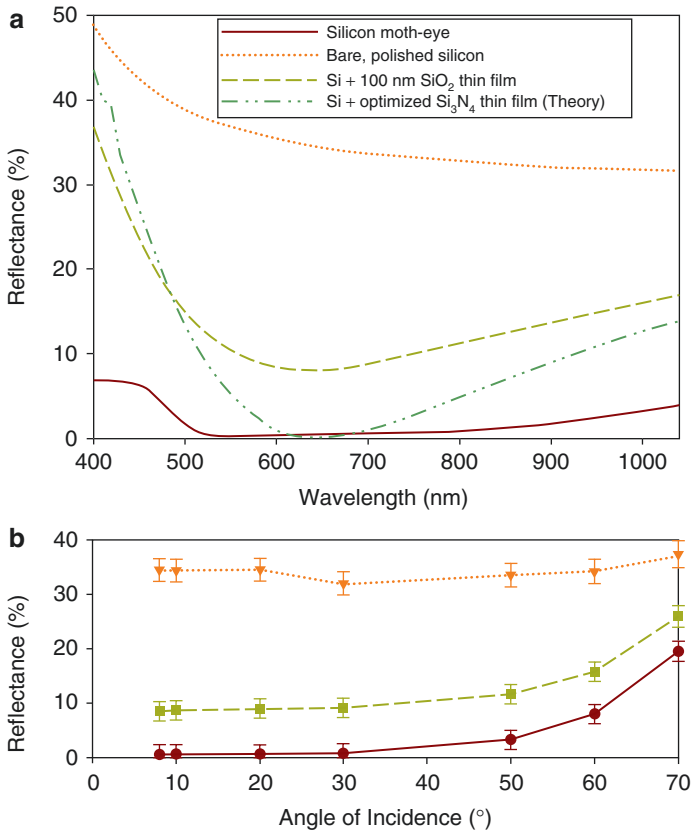
Nanosphere Lithography

An example of a bottom-up, self-assembly approach to artificial moth-eye structure fabrication on the wafer scale involves the formation of a mask with submicron-sized silica spheres deposited from a colloidal suspension. In one example, 360 nm diameter silica spheres dispersed in a polymer were spin coated onto a 4 in. wafer. The polymer was etched away leaving an ordered array of spheres which could then be employed as a mask for an SF_6 etch to form moth-eye structures in the underlying silicon substrate [21]. A similar technique has been used to create moth-eye structures in GaSb, an important material for thermophotovoltaic applications.

Other techniques have been developed that employ the same graded index AR mechanism of moth-eye arrays but with less regular subwavelength structure (and so are not strictly artificial moth eyes). These include the various AR coatings made through sol–gel processes [22], nanophase-separated polymer films [23], and acid etching to obtain nanoscale texturing of a surface [24]. These techniques, which are more accurately described as bio-inspired rather than biomimetic, potentially offer cheaper routes to highly effective AR on a range of materials.

Applications

Applications for moth-eye structures exist wherever suppression of reflection is desired. The



Moth-Eye Antireflective Structures,
Fig. 8 Reflectance versus (**a**) wavelength and (**b**) angle of incidence from measurements on moth-eye structures formed in silicon, compared to bare silicon and thin-film ARCs [26]

broadband and wide-angle nature of their AR properties make moth-eye structures superior to more traditional thin-film AR schemes (Fig. 8). Such applications include reducing reflection from the top surfaces of photovoltaic and thermal solar cells, window panes, lenses, instrument panels, and displays [2, 25]. Moth-eye structures have also been applied to improve the light extraction in light-emitting diodes (LEDs) by reducing the number of photons trapped in the device due to total internal reflection [25]. Thin-film coatings can suffer from cracking and delamination, especially when used in applications involving large changes in temperature because thermal expansion coefficients of the film and substrate may differ. Moth-eye arrays can be formed in the substrate material itself and so do not suffer from these problems, making them particularly suitable for applications involving high light intensities and large thermal ranges such as in laser devices.

The main drawback with the technology is the requirement of dense arrays of submicron scale features, often over large areas, but these issues are being successfully tackled with the development of scalable fabrication techniques such as nanoimprint and interference lithography. There is also concern over the robustness of moth-eye structures and whether they can withstand harsh outdoor conditions. Self-cleaning through the natural superhydrophobicity exhibited by the structures helps reduce contamination buildup, but susceptibility to damage by abrasion remains an issue. The mechanical properties of the substrate material into which the moth-eye structures are formed are therefore important factors in the design of artificial moth eyes for practical purposes. Furthermore, moth-eye arrays may be more suited to application on inner surfaces or at the interfaces between two solids rather than on outer, exposed surfaces.

Most commercial applications of moth-eye structures to date involve the reduction of reflection from a glass or polymer surface because in these cases, an etch step is avoided and interference lithography or nanoimprinting can replicate the design into a soft plastic layer. Molds made by electron-beam or interference lithography for nanoimprinting of moth-eye structures into polymers on the scale of tens of centimeters are available commercially. Nanoimprinting can be used to produce moth-eye AR polymer films that can be applied to displays for mobile telephones, PDAs, e-readers, SatNav devices, etc., and to glass covers of solar cells. By reducing internal reflection, the structures can enhance the perceived brightness of LED/LCD screens and so reduce their power consumption. A combination of moth-eye structures to reduce reflection and larger-scale texturing to reduce the proportion of specularly reflected light creates surfaces with antiglare properties, in addition to low reflectances, for display applications.

The application of moth-eye structures in hard materials (except for when used as a nanoimprint stamp/mold) has seen more limited success, mainly due to the extra cost of etching required to transfer the pattern into the underlying substrate. The main application for moth eyes in semiconductors would be for solar cells, but the cost requirements of patterning large areas with regular arrays of subwavelength scale features mean that thin-film coatings are often used instead. The main trend in this area is a move away from regular, biomimetic arrays and toward more random subwavelength texturing using cheaper, mask less techniques [22–24], which, although employing the same graded index mechanism, no longer resemble the structures found in nature and therefore cannot be classified as artificial moth eyes. Nevertheless, moth-eye structures in semiconductors may still find applications where cheap, large area coverage is not so important, for example, in optical sensors, LEDs, and concentrator solar cells. Furthermore, recent developments in self-assembly techniques may provide a path to continued progress in this area.

Cross-References

- ▶ Antifogging Properties in Mosquito Eyes
- ▶ Biomimetics
- ▶ Electron Beam Lithography (EBL)
- ▶ Nanoimprint Lithography
- ▶ Nanoimprinting
- ▶ Nanostructures for Photonics
- ▶ Organic Photovoltaics: Basic Concepts and Device Physics
- ▶ Scanning Electron Microscopy
- ▶ Self-Assembly of Nanostructures
- ▶ Toward Bioreplicated Texturing of Solar-Cell Surfaces

References

1. Bernhard, C.G.: Structural and functional adaption in a visual system. Endeavour **26**, 79–84 (1967)
2. Parker, A.R., Townley, H.E.: Biomimetics of photonic nanostructures. Nat. Nanotechnol. **2**(6), 347–353 (2007)
3. Stavenga, D.G., Foletti, S., Palasantzas, G., Arikawa, K.: Light on the moth-eye corneal nipple array of butterflies. Proc. R. Soc. B Biol. Sci. **273**(1587), 661–667 (2006)
4. Parker, A.R., Hegedus, Z., Watts, R.A.: Solar-absorber antireflector on the eye of an Eocene fly (45 Ma). Proc. R. Soc. B Biol. Sci. **265**(1398), 811–820 (1998)
5. Gao, X., Yan, X., Yao, X., Xu, L., Zhang, K., Zhang, J., Yang, B., Jiang, L.: The dry-style antifogging properties of mosquito compound eyes and artificial analogues prepared by soft lithography. Adv. Mater. **19**(17), 2213–2217 (2007)
6. Stoddart, P.R., Cadusch, P.J., Boyce, T.M., Erasmus, R.M., Comins, J.D.: Optical properties of chitin: surface-enhanced Raman scattering substrates based on antireflection structures on cicada wings. Nanotechnology **17**(3), 680–686 (2006)
7. Peisker, H., Gorb, S.N.: Always on the bright side of life: anti-adhesive properties of insect ommatidia grating. J. Exp. Biol. **213**(Pt 20), 3457–3462 (2010)
8. Boden, S.A., Bagnall, D.M.: Nanostructured biomimetic moth-eye arrays in silicon by nanoimprint lithography. Proc. SPIE. **7401**, 74010J–74010J–12 (2009)
9. Lalanne, P., Lemercier-lalanne, D.: On the effective medium theory of subwavelength periodic structures. J. Mod. Opt. **43**(10), 2063–2085 (1996)
10. Wilson, S.J., Hutley, M.C.: The optical properties of 'moth eye' antireflection surfaces. Opt. Acta (Lond) **29**(7), 993–1009 (1982)
11. Southwell, W.H., Southwell, W.H.: Pyramid-array surface-relief structures producing antireflection index matching on optical surfaces. J. Opt. Soc. Am. A **8**(3), 549 (1991)

12. Raguin, D.H., Morris, G.M.: Antireflection structured surfaces for the infrared spectral region. Appl. Opt. **32**(7), 1154–1167 (1993)
13. Han, K., Chang, C.-H.: Numerical modeling of sub-wavelength anti-reflective structures for solar module applications. Nanomaterials **4**(1), 87–128 (2014)
14. Toyota, H., Takahara, K., Okano, M., Yotsuya, T., Kikuta, H.: Fabrication of microcone array for antireflection structured surface using metal dotted pattern. Jpn. J. Appl. Phys. **40**(7B), 747–749 (2001)
15. Boden, S.A., Bagnall, D.M.: Tunable reflection minima of nanostructured antireflective surfaces. Appl. Phys. Lett. **93**(13), 133108 (2008)
16. Sai, H., Kanamori, Y., Arafune, K., Ohshita, Y., Yamaguchi, M.: Light trapping effect of submicron surface textures in crystalline Si solar cells. Prog. Photovolt. Res. Appl. **15**, 415–423 (2007)
17. Boden, S.A., Bagnall, D.M.: Optimization of moth-eye antireflection schemes for silicon solar cells. Prog. Photovolt. Res. Appl. **18**(3), 195–203 (2010)
18. Song, Y.M., Bae, S.Y., Yu, J.S., Lee, Y.T.: Closely packed and aspect-ratio-controlled antireflection subwavelength gratings on GaAs using a lenslike shape transfer. Opt. Lett. **34**(11), 1702–1704 (2009)
19. Chen, X., Fan, Z.-C., Zhang, J., Song, G.-F., Chen, L.-H.: Pseudo-rhombus-shaped subwavelength crossed gratings of GaAs for broadband antireflection. Chin. Phys. Lett. **27**(12), 124210 (2010)
20. Chen, Q., Hubbard, G., Shields, P.A., Liu, C., Allsopp, D.W.E., Wang, W.N., Abbott, S.: Broadband moth-eye antireflection coatings fabricated by low-cost nanoimprinting. Appl. Phys. Lett. **94**(26), 263118 (2009)
21. Sun, C.-H., Jiang, P., Jiang, B.: Broadband moth-eye antireflection coatings on silicon. Appl. Phys. Lett. **92**(6), 061112 (2008)
22. Chen, D.: Anti-reflection (AR) coatings made by sol–-gel processes: a review. Sol. Energy Mater. Sol. Cells **68**(3–4), 313–336 (2001)
23. Walheim, S., Schaffer, E., Mlynek, J., Steiner, U.: Nanophase-separated polymer films as high-performance antireflection coatings. Science **283**(5401), 520–522 (1999)
24. Koynov, S., Brandt, M.S., Stutzmann, M.: Black multi-crystalline silicon solar cells. Phys. Status Solidi Rapid Res. Lett. **1**(2), R53–R55 (2007)
25. Brunner, R., Sandfuchs, O., Pacholski, C., Morhard, C., Spatz, J.: Lessons from nature: biomimetic subwavelength structures for high-performance optics. Laser Photonics Rev. **6**(5), 641–659 (2012)
26. Boden S.A.: Biomimetic nanostructured surfaces for antireflection in photovoltaics. Doctoral thesis, University of Southampton (2009)

Motion Sensors

▶ Gyroscopes

Mucoadhesion

▶ Bioadhesion

Multifunctional Particles

▶ Multistage Vectors

Multilamellar Vesicle (MLV)

▶ Liposomes

Multiphoton Fabrication

▶ Laser Beam Machining

Multiscale Fluidics

▶ Integration of Nanostructures Within Microfluidic Devices

Multiscale Modeling

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Multiscale Simulations

▶ Coarse-Grained and Hybrid Simulations of Nanostructures

Multistage Delivery System

▶ Multistage Vectors

Multistage Vectors

Rita E. Serda[1,2] and Mauro Ferrari[2]
[1]Michael E. DeBakey Department of Surgery, Baylor College of Medicine, Houston, TX, USA
[2]Department of NanoMedicine, Houston Methodist Research Institute, Houston, TX, USA

Synonyms

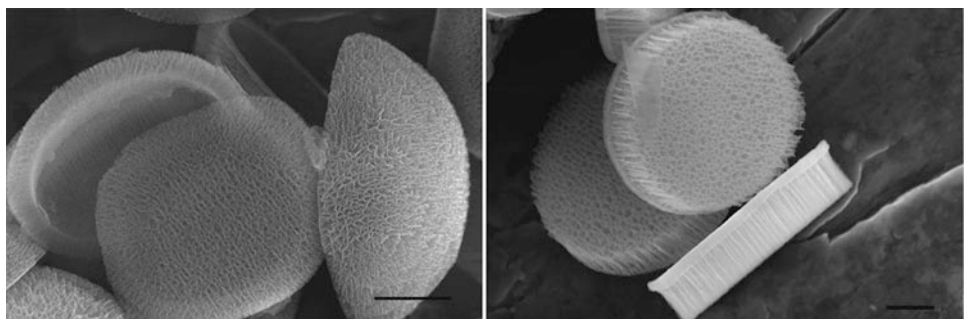
Logic-embedded vectors; Multifunctional particles; Multistage delivery system; Third-generation vectors

Definition

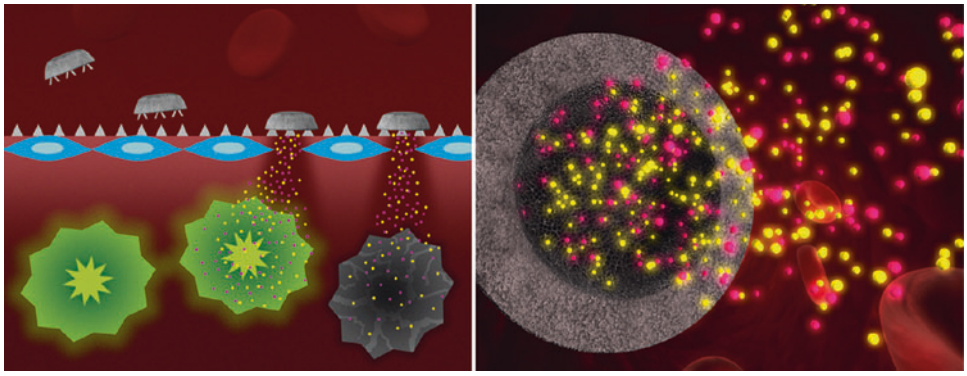
Multistage vectors are multiparticle, multifunctional vectors strategically envisioned to overcome abundant obstacles encountered by drug delivery vectors as they journey from the site of administration to the target site. These vectors are comprised of layers of particles, with secondary nanoparticles and active agents nested into first-stage porous silicon microparticles. The ability to achieve sequential targeting through particle optimization enhances the probability of reaching the appropriate cellular or subcellular target within a lesion with the appropriate drug. Multistage vectors are a subset of Logic-Embedded Vectors (LEVs) [1]. By definition, LEVs are vectors with the ability to perform a time sequence of functions through the use of multiple nano-based components that synergistically provide distinct functionalities.

Porous Silicon Particles as First-Stage Vectors

Well-established semiconductor microfabrication lithography and electrochemical etching techniques are used to fabricate porous silicon microparticles in well-defined, tunable sizes, shapes, and porosities [2]. Photolithographic masks are used to define particle shape, with porosity defined by the silicon-etch protocol. Both the size and shape of the particle affect physiological transport. To date, resulting particles have been in the micrometer size range, with shapes including tubular, discoidal, and hemispherical (Fig. 1). Particle porosities range from approximately 45–80 % [2], with the degree of porosity directly correlated with the rate of particle degradation. Greater particle porosities support faster rates of degradation. Thus, pore size permits controlled release of secondary agents and optimization of the matrix for specific cargo. In addition, chemical modification and the addition of functional groups to the particle surface have been shown to impact particle transport, stability and cargo release kinetics. Particles degrade to orthosilicic acid, a bioavailable form of silicon, when suspended in saline or serum [3].



Multistage Vectors, Fig. 1 Quasihemispherical (*left*; bar 400 nm) and discoidal (*right*, bar 500 nm) porous silicon microparticles. Images were taken using Hitachi S-5500 (*left*) and the Zeiss Neon 40 (*right*) scanning electron microscopes (Images courtesy of Rita Serda (*left*) and Xuewu Liu (*right*))



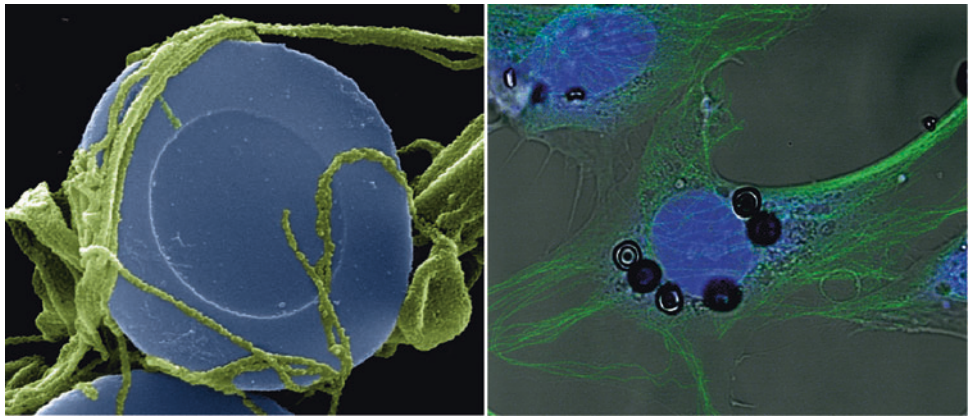
Multistage Vectors, Fig. 2 Schematic illustrating the multistage particle concept, with first-stage porous silicon particles adherent to tumor neovascular endothelia (*left*) or free in blood (*right*) with sustained or triggered release of second-stage nanoparticles. The released nanoparticles enter the tumor interstitial space by mechanisms that include the EPR effect. Molecular recognition of nanoparticles by tumor cells is followed by cellular internalization and trafficking of drugs to the appropriate subcellular organelle or cytosolic complex (Images created by Matthew Landry)

Biobarriers arise as a result of biological compartmentalization and the body's robust defense system. Multiple barriers decrease the probability of reaching the therapeutic target such that only 1 in 10,000 to 1 in 100,000 molecules reach the intended site of action [4]. Sequentially, with respect to intravascular injection site, these barriers include enzymatic degradation; sequestration by phagocytes of the reticuloendothelial system (RES); vascular endothelia; adverse oncotic and interstitial pressure in the tumor; cellular membranes, subcellular organelles such as the nucleus and endosomes; and molecular efflux pumps [5]. Without an effective strategy to negotiate these barriers, new or current therapeutic agents based on enhanced biomolecular selectivity may yield suboptimal therapeutic efficacy. Transport through different compartments and across biological barriers can be enhanced by the presentation of agents, such as permeation enhancers, and nanoparticles that present appropriate molecular moieties to influence biological associations. Sequential presentation of agents housed in multistage vectors is one means of achieving this goal.

When injected intravascularly, multistage vectors travel within the bloodstream, periodically interacting with the vascular endothelium. To enhance accumulation of vectors in the tumor neovasculature, particle parameters and surface chemistry of porous silicon particles are optimally designed with four primary objectives: (1) margination dynamics, (2) firm adhesion at the target site, (3) controlled rate of silicon degradation and/or release of secondary nanoparticles and agents, and (4) transport of particles across the endothelial barrier. The multistage vision is illustrated in Fig. 2. Similar to leukocyte extravasation, first-stage porous silicon particles bounce, roll, and adhere to the vascular wall at sites of molecular recognition.

The final step of extravasation, endothelial transmigration, is accomplished by a variety of mechanisms. With respect to tumors, transmigration may be achieved via sustained release of second-stage nanoparticles and accumulation in the tumor through mechanisms including Enhanced Permeability and Retention (EPR). EPR is based on abnormal blood vessel formation during cancer angiogenesis, leading to poorly aligned endothelial cells with wide fenestrations, as well as ineffective lymphatic drainage. Other mechanisms of transmigration include transcytosis and particle exocytosis.

Rational design is used to guide particle fabrication and produce vectors that can spontaneously move preferentially in close proximity to the blood vessel endothelium. The objective is to



Multistage Vectors, Fig. 3 Cellular associations with porous silicon microparticles. *Left*: a scanning electron micrograph showing endothelial filopodia extending around a silicon microparticle. *Right*: a confocal micrograph showing endothelial cells that have internalized multiple silicon microparticles. Cellular tubulin was labeled with fluorescein isothiocyanate-conjugated antibody and 4′, 6-diamidino-2-phenylindole was used for nuclear labeling (Ref. [10] and [11])

create vectors that exhibit lateral drift from the center of the blood vessel toward the endothelial wall [6, 7]. Unlike spherical particles, nonspherical particles exhibit more complex motions that include tumbling and rolling, with a greater tendency to approach the vessel wall and periodically interact with the endothelial surface.

Targeting the tumor neovasculature is achieved by decorating the porous silicon particle with surface ligands that specifically bind to cell surface receptors over- or uniquely expressed on tumor-associated endothelia. The probability of cellular adhesion increases as the surface density of ligand molecules on the particle surface and receptor molecules on the cell membrane increase. In addition, for all particle shapes, a characteristic size can be identified for which the probability of adhesion has a maximum [8].

Cellular internalization of stage one porous silicon microparticles has been demonstrated in Human Umbilical Vein Endothelial Cells (HUVEC) and Human Microvascular Endothelial Cells (HMVEC) [9]. The process of internalization is actin dependent and includes both phagocytosis and macropinocytosis. Following cellular adhesion, and "cupping" of the particle (i.e., actin cup formation), the plasma membrane extends outward to surround the particle via formation of psuedopodia (Fig. 3, left). Internalized microparticles migrate along microtubules to the perinuclear region of the cell (Fig. 3, right). Macropinocytosis is also involved in particle internalization and is supported by formation of membrane ruffles.

Ligands for targeting specific cell populations include antibodies, peptides, oligonucleotides, small molecules, lipids, and carbohydrates. Optimal ligands include Fab (Fragment antigen binding) fragments, peptides, and aptamers (single-strand oligonucleotides) based on their small size, reduced antigenicity, and low levels of RES recognition. Inflammatory cytokine-driven expression of selectins, adhesion molecules, and integrins on endothelia function as targets for delivery vehicles at sites of pathology which include cancer, atherosclerosis, and rheumatoid arthritis. As an example, E-selectin-specific thioaptamers, which are thiophosphate ester-modified nucleic acids, have been shown to enhance tumor vasculature targeting in a breast cancer xenograft model [12]. Another example of enhanced targeting includes the use of the peptide "RGD" (arginine-glycine-aspartic acid), which is frequently used for integrin-mediated targeting of endothelial cells.

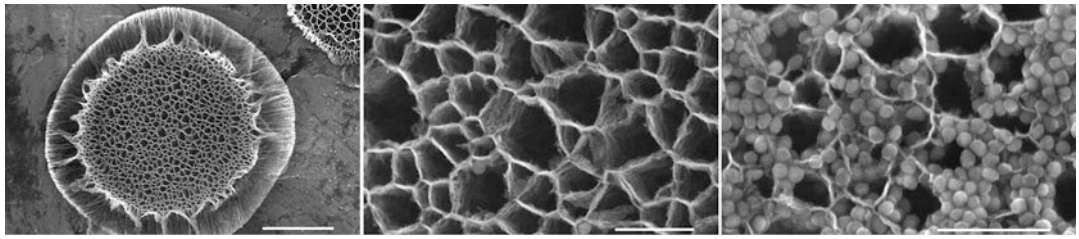
To achieve optimal targeting, a phenomenon known as biological fouling should be minimized through appropriate selection of surface polymers and targeting ligands for modification of each

particle population. Fouling is the process of serum components binding to the particle surface, with particle surface chemistry dictating the array of plasma components that bind to the outer particle creating a corona. Serum components are referred to as either opsonins or dysopsonins, depending on their impact, which either favor or disfavor cellular binding and uptake. Particles with a negative surface potential, including oxidized silicon microparticles and negatively charged polystyrene microparticles, bind to serum dysopsonins with respect to cellular uptake by endothelial cells, preventing adhesion and internalization [9]. Conversely, silicon particles with neutral or positive surface potential dictate adhesion by unique proteins that do not alter cellular uptake by endothelia. One method used to reduce serum protein binding and altered trafficking is functionalization of the particle surface with neutral polymers that reduce cellular uptake, such as poly(ethylene glycol) (PEG) [13–15]. Interactions between the aqueous environment and hydrophilic polymers include extension and mobility of the polymeric chains, such that protein binding is thermodynamically unfavorable [16]. An alternative strategy used to influence cellular interactions is modification of the particle surface to direct intentional binding of specific serum components to dictate cellular interactions. A final strategy to avoid binding of serum components to nanoparticles is to load them into larger carrier particles. Porous silicon particles provide protected transport of secondary and higher-level particles, negating the need for surface PEGylation. As an example, has been demonstrated that transportation of nanoparticles in the silicon matrix stabilizes chitosan-coated nanoparticles, enhancing chitosan-driven endosomal escape of internalized nanoparticles [17]. A potential benefit of cell uptake of carrier microparticles includes transcellular transport of nanoparticles across the endothelial barrier into the targeted tissue.

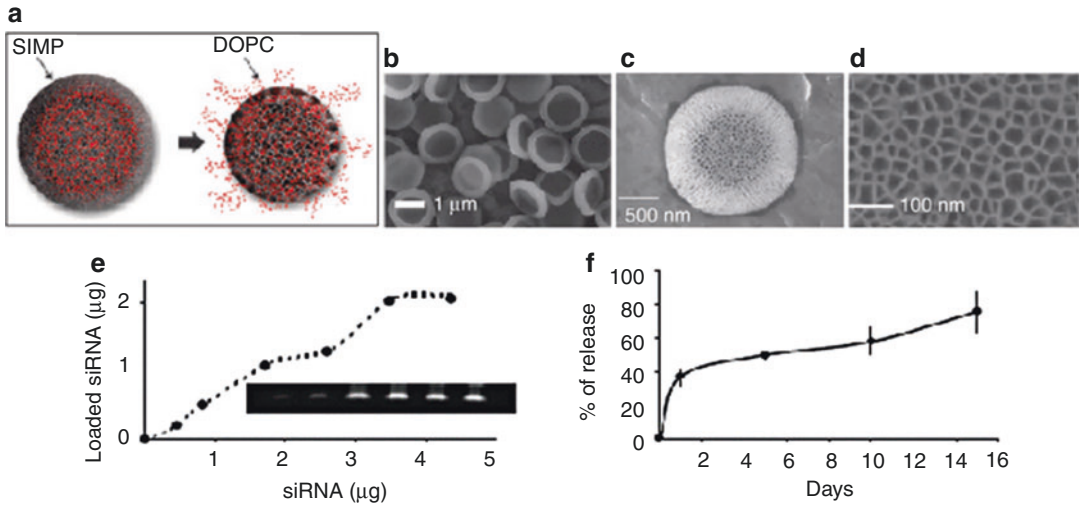
Candidates for Stage Two Vectors

Candidates for stage two vectors include liposomes, polymeric micelles, iron oxide, and gold nanoparticles. Liposomes (e.g., PEGylated liposomal doxorucicin; Doxil) and iron oxide nanoparticles [e.g., AMI-121 (siloxane), OMP (polystyrene-coated), and AMI-25] are two US Food and Drug Administration (FDA)-approved nanoparticles in clinical trials as therapeutic and magnetic resonance imaging agents, respectively. Liposomes are vesicular nanostructures, formed from phospholipid and cholesterol molecules. The inner hydrophilic compartment can encapsulate water-soluble drugs, as well as therapeutic proteins and oligonucleotides. Iron oxide nanoparticles have unpaired electron spins, which makes them superparamagnetic and influences the relaxation properties of protons in surrounding water molecules. This change in relaxation properties creates regions of dark contrast in tissues and cells where the nanoparticles accumulate. Figure 4 contains scanning electron micrographs illustrating the impact of particle surface chemistry on nanoparticle loading in to discoidal porous silicon microparticles [17].

Systemic delivery of therapeutic RNA for RNA interference (RNAi) therapy is challenged by short half-life, lack of plasma membrane penetration, and potential toxicity [18]. Nanoparticle-based delivery has achieved some success, but rapid clearance and instability present limitations. Multistage technology has been used to achieve sustained delivery of siRNA, long-term gene silencing, and reduced toxicity due to lower dose requirements. Neutral liposomes composed of dioleoyl phosphatidylcholine (DOPC) and associated siRNA, loaded into porous silicon microparticles (Fig. 5), exhibited sustained in vivo gene silencing for 3 weeks, significant antitumor effects in two orthotopic mouse models of human ovarian cancer (SKOV3ip1 and HeyA8), and no observable toxicity after a single intravenous injection. A single injection of multistage liposomal EphA2 siRNA resulted in greater than 80 % reduction of EphA2 expression, which lasted for 3 weeks, compared to a 6-day reduction for a single dose of liposomal siRNA. In addition, sustained EphA2 silencing reduced angiogenesis and cell proliferation compared to the control siRNA. Release of 35 % of Alexa555-siRNA-DOPC from porous silicon particles occurred in



Multistage Vectors, Fig. 4 Scanning electron micrographs showing the impact of surface chemistry on loading porous silicon particles. *Left*: a 3 μm discoidal porous silicon microparticle (25 K mag.); *middle*: attempted loading of an oxidized silicon microparticle with 30 nm carboxylated iron oxide nanoparticles (180 K mag.); *right*: successful loading of an oxidized silicon microparticle with 30 nm PEGylated amine iron oxide nanoparticles (200 K mag.). Images were taken using the Hitachi S-5500 high resolution scanning electron microscope at 20 Kv (Images adapted from and courtesy of Small Ref. [17])



Multistage Vectors, Fig. 5 Assembly of a multistage vector containing siRNA-DOPC. (**a**) The multistage concept. (**b**–**d**) Scanning electron micrographs of porous silicon particles at different magnifications. (**e**) Loading of Alexa555-siRNA-DOPC in porous silicon. Loading efficiency was determined by measuring unincorporated fluorescence. (**f**) Release kinetics of Alexa555-DOPC from porous silicon particles in 10 % fetal bovine serum as a measure of fluorescence in the supernatant (Images reproduced from Cancer Research, American Society of Cancer Research Ref. [18])

the first 3 h, followed by sustained release of the remaining amount over 14 days (Fig. 5F). Tumor siRNA was evident in mice treated with a single injection of multistage vectors for longer time periods as compared to a single injection of the same dose of free siRNA. Together these data support improved long-term gene silencing and tumor reduction with multistage vectors as compared to free liposomal siRNA.

Multistage vectors singly or simultaneously loaded with quantum dots (Q-dots) and single-walled carbon nanotubes (SWCNTs) have been demonstrated [3]. 3′-aminopropyltriethoxysilane (APTES)-modified porous silicon particles were loaded with either polyethylene glycol (PEG)-fluorescein isothiocyanate SWCNTs or carboxylated Q-dots. Loading in large-pore silicon particles (20–30 nm) was rapid and abundant; however, loading of small-pore particles (5–10 nm) was minimal (6 and 25 %). The chemical properties of the particle surfaces affected loading, with electrostatic attractions enhancing loading and stability. Release of the nanoparticles at physiological temperature and pH in

Tris-buffered saline was sustained for 6 h. However, Q-dot release kinetics were significantly faster than that observed for SWCNTs, with 80 % of the Q-dots being released in the first 1.5 h, compared to 50 % of the SWCNTs. When the Q-dots and SWCNTs were loaded simultaneously, maximal load was achieved after 60 min, compared to 15 min for single cargo loading. However, release kinetics were unaltered. While SWCNTs were associated with the entire silicon particle, Q-dots were predominantly in the central region of the particle, owing to the presence of larger pores. These early studies have paved the way for current research focusing on codelivery of multiple therapeutic and/or imaging agents.

The biological impact of porous silicon vectors has been evaluated in cell lines representing the immune system (i.e., macrophages) and physiological targets (endothelial cells). Cellular morphology, cell viability, impact on cell cycle, mitotic potential, and proinflammatory responses following cellular engulfment of silicon microparticles have been evaluated [19]. Both cell types internalize microparticles via phagocytosis, with subsequent intracellular transport of vesicles to the perinuclear region of the cell (Fig. 3). During cellular division there is equal partitioning of microparticles between the two daughter cells owing to predominately single particle loading into endosomes during the internalization process. Equal partitioning of endosomes during mitosis, in cells containing microparticles supports undisturbed, healthy cellular function. Normal endothelial and macrophage proliferation and lack of cytotoxicity following internalization of silicon microparticles also support cellular biocompatibility of the first-stage vectors. In vitro studies support that endothelial cells are unaffected by the presence of porous silicon microparticles based on analysis of cell morphology, viability, cell cycle, apoptosis, mitosis, and lack of cytokine release.

Intracellular trafficking of particles is crucial to achieving the desired biological outcome. Intracellular targets include the mitochondria, nucleus, lysosomes, endoplasmic reticulum, golgi apparatus, and cytosolic complexes such as the RNA-induced silencing complex (RISC). Following cellular internalization of multistage vectors, the endosomal-entrapped assembly dissociates, and particles are sorted based on particle properties [17]. The larger silicon particles remain in endosomes, which undergo normal maturation to late endosomes and lysosomes. Released nanoparticles, for example, iron oxide nanoparticles, dissoicate from the silicon carrier and traffic to new destinations driven by the properties of the nanoparticle. For example, as stated previously, chitosan coating of the nanoparticles favors endosomal escape, while a PEGylated amine surface favors continued vesicular enclosure. Membrane-bound vesicles containing iron oxide nanoparticles have been reported to be secreted from both macrophages and endothelial cells. These membrane-enclosed vesicles typically have a diameter near 1 μm. The biological role of these biovesicles includes both proximal and distal communication between cells.

Recent findings challenge the dogma of particles functioning as static depots once they reach their destination. Endothelial cells shuttle particles between cells through cellular connections, known as tunneling nanotube (TNT), and via cellular shuttles [20]. TNTs are channels between cells that allow for the exchange of cellular constituents, including organelles [21, 22]. Cellular shuttles are relay cells. As documented using monolayer cell cultures, endothelial cells are capable of picking up particles from donor cells and transporting them to acceptor cells. Cell-to-cell transport of drug carriers may be one mechanism that can be manipulated to spread drug across specific cell populations.

Conclusions and Future Directions

Multistage vectors have the potential to bypass multiple, sequential barriers, targeting at multiple levels in order to achieve selective accumulation in tumors or pathogenic tissues. The ability to reproducibly and precisely tailor particle size, shape, and porosity creates an abundant pool of first-stage vectors that can be optimized for cargo loading and release, vascular margination, cellular targeting, adhesion, and internalization. The ability to load diverse cargo in the first-stage vector exponentially increases the possible cellular or

structural targets and therapeutic objectives. Therefore, multistage vectors exemplify a new paradigm in therapy that moves away from single-agent therapy toward combinatorial therapy with potential for personalization and real-time monitoring of therapeutic progress.

Cross-References

- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ Liposomes
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Scanning Electron Microscopy

References

1. Ferrari, M.: Frontiers in cancer nanomedicine: transport oncophysics and logic-embedded vectors. Trends Biotechnol. **28**(4), 181–188 (2010)
2. Chiappini, C., Tasciotti, E., Fakhoury, J.R., Fine, D., Pullan, L., Wang, Y.C., Fu, L., Liu, X., Ferrari, M.: Tailored porous silicon microparticles: fabrication and properties. Chemphyschem **11**(5), 1029–1035 (2010)
3. Tasciotti, E., Liu, X., Bhavane, R., Plant, K., Leonard, A.D., Price, B.K., Cheng, M.M., Decuzzi, P., Tour, J.M., Robertson, F., Ferrari, M.: Mesoporous silicon particles as a multistage delivery system for imaging and therapeutic applications. Nat. Nanotechnol. **3**(3), 151–157 (2008)
4. Ferrari, M.: Nanovector therapeutics. Curr. Opin. Chem. Biol. **9**(4), 343–346 (2005)
5. Serda, R.E., Godin, B., Blanco, E., Chiappini, C., Ferrari, M.: Multi-stage delivery nano-particle systems for therapeutic applications. Biochim. Biophys. Acta **1810**, 317–329 (2010)
6. Decuzzi, P., Ferrari, M.: Design maps for nanoparticles targeting the diseased microvasculature. Biomaterials **29**(3), 377–384 (2008)
7. Decuzzi, P., Ferrari, M.: Modulating cellular adhesion through nanotopography. Biomaterials **31**(1), 173–179 (2010)
8. Lee, S.Y., Ferrari, M., Decuzzi, P.: Shaping nano-/micro-particles for enhanced vascular interaction in laminar flows. Nanotechnology **20**(49), 495101 (2009)
9. Serda, R.E., Gu, J., Bhavane, R.C., Liu, X., Chiappini, C., Decuzzi, P., Ferrari, M.: The association of silicon microparticles with endothelial cells in drug delivery to the vasculature. Biomaterials **30**(13), 2440–2448 (2009)
10. Serda, R., Robertson, F., Ferrari, M.: Overcoming nature's fastball – bio-barriers in cancer nanotechnology (not peer reviewed). Nano. Mag. Small Sci. **5**, 10–13 (2008)
11. Ferrati, S., Mack, A., Chiappini, C., Liu, X., Bean, A. J., Ferrari, M., Serda, R.E.: Intracellular trafficking of silicon particles and logic-embedded vectors. Nanoscale **2**(8), 1512–1520 (2010)
12. Mann, A.P., Somasunderam, A., Nieves-Alicea, R., Li, X., Hu, A., Sood, A.K., Ferrari, M., Gorenstein, D.G., Tanaka, T.: Identification of thioaptamer ligand against E-selectin: potential application for inflamed vasculature targeting. PLoS One **5**(9), e13050 (2010)
13. Hashizume, H., Baluk, P., Morikawa, S., McLean, J. W., Thurston, G., Roberge, S., Jain, R.K., McDonald, D.M.: Openings between defective endothelial cells explain tumor vessel leakiness. Am. J. Pathol. **156**(4), 1363–1380 (2000)
14. Maeda, H.: The enhanced permeability and retention (EPR) effect in tumor vasculature: the key role of tumor-selective macromolecular drug targeting. Adv. Enzym Regul. **41**, 189–207 (2001)
15. Torchilin, V.P.: Recent advances with liposomes as pharmaceutical carriers. Nat. Rev. Drug Discov. **4**(2), 145–160 (2005)
16. Hrkach, J.S., Peracchia, M.T., Domb, A., Lotan, N., Langer, R.: Nanotechnology for biomaterials engineering: structural characterization of amphiphilic polymeric nanoparticles by 1H NMR spectroscopy. Biomaterials **18**(1), 27–30 (1997)
17. Serda, R.E., Mack, A., van de Ven, A.L., Ferrati, S., Dunner Jr., K., Godin, B., Chiappini, C., Landry, M., Brousseau, L., Liu, X., Bean, A.J., Ferrari, M.: Logic-embedded vectors for intracellular partitioning, endosomal escape, and exocytosis of nanoparticles. Small **6**, 2691–2700 (2010)
18. Tanaka, T., Mangala, L.S., Vivas-Mejia, P.E., Nieves-Alicea, R., Mann, A.P., Mora, E., Han, H.D., Shahzad, M.M., Liu, X., Bhavane, R., Gu, J., Fakhoury, J.R., Chiappini, C., Lu, C., Matsuo, K., et al.: Sustained small interfering RNA delivery by mesoporous silicon particles. Cancer Res. **70**(9), 3687–3696 (2010)
19. Serda, R.E., Ferrati, S., Godin, B., Tasciotti, E., Liu, X., Ferrari, M.: Mitotic partitioning of silicon microparticles. Nanoscale **2**, 173–288 (2009)
20. Ferrati, S., Shamsudeen, S., Summers, H.D., Rees, P., Abbey, J.V., Schmulen, J., Liu, X., Wong, S.T., Bean, A. J., Ferrari, M., Serda, R.E. Small **8**(20), 3151–60 (2012)
21. Rustom, A., Saffrich, R., Markovic, I., Walther, P., Gerdes, H.H. Science, **303**(5660), 1007–10 (2004)
22. Onfelt, B., Nedvetzki, S., Benninger, R.K., Purbhoo, M. A., Sowinski, S., Hume, A. N., Seabra, M. C., Neil, M. A., French, P. M., Davis, D. M. Journal of immunology **177**(12), 8476–83 (2006)

Multiwalled Carbon Nanotubes (MWCNTs)

- ▶ Chemical Vapor Deposition (CVD)

N

Nano (Evanescent-Wave)-Particle Image Velocimetry

► Micro/Nano Flow Characterization Techniques

Nano Manipulation

► 3D Micro-/Nanomanipulation with Force Spectroscopy
► CNT Handling and Integration

Nanoalloy Simulation

Giovanni Barcaro[1], Luca Sementa[2] and Alessandro Fortunelli[3]
[1]CNR IPCF, Consiglio Nazionale delle Ricerche, Pisa, Tuscany, Italy
[2]CNR-ICCOM, Consiglio Nazionale delle Ricerche, Pisa, Tuscany, Italy
[3]Istituto per la Chimica dei Composti Organometallici (Institute for the Chemistry of Organometallic Compounds, ICCOM), Consiglio Nazionale delle Ricerche (National Research Council, CNR), Pisa, PI, Italy

Synonyms

Binary systems; Metal alloy nanoparticles; Nanoscale metal alloys

Definition

The term "nanoalloy" is employed in recent scientific literature to design materials consisting of metal structures with at least one nanoscale dimension and composed by two or more different metals. Often, the term is restricted to nanoscale alloy particles, i.e., multicomponent metallic aggregates with diameters ranging between 1 and 100 nm [1]. With respect to pure systems, nanoalloys present additional degrees of freedom of composition and chemical ordering, that is, the ratio among the elemental components can be varied as well as how these species are distributed within the nanostructure. By combining these variables with nanoscale confinement effects, it is possible to confer these nanomaterials' unique properties. They are therefore of great interest in basic science as well as in current nanotechnology and find applications in a variety of fields: from chemical sensing to heterogeneous catalysis and from magnetic recording to optoelectronic devices, thermal treatment of cancer, and "cell imaging," i.e., colorimetric probes for DNA detection [2]. Such interest at the experimental and industrial level has stimulated corresponding theoretical efforts, and a field of modern computational materials science is devoted to the prediction of the structure and properties of nanoalloys via simulations. While many theoretical methods here employed are borrowed from or shared with other fields, some are specific. For example, in nanoalloys the chemical ordering or

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

compositional structure, i.e., the way in which the different chemical elements are positioned within a fixed structural framework, gives rise to a huge number of isomers called "homotops" [3]. The computational approaches used to deal with this issue by efficiently sampling the chemical ordering phase space are thus specific to the nanoalloy field.

Introduction

Nanostructured alloy materials, such as alloy particles with nanoscale diameter, extended alloy materials with nanoscale structure, etc., represent an important field of current nanotechnology. Due to the increased freedom associated with composition and chemical ordering, new physical phenomena appear in metal multicomponent nanosystems with respect to their pure counterparts and novel or profoundly modified properties. There are several reasons why a gain can be achieved by employing a multicomponent nanostructured system with respect to pure and/or extended one which justify the importance of nanostructured alloys. First, new structural motifs with respect to pure systems can be created by alloying, thus generating novel structure-property relationships. Second, the chemical and physical properties of, e.g., a nanoparticle can be tuned by varying not only its size and shape but also its composition (i.e., the ratio between the different chemical species within the particle) and the degree of atomic mixing or chemical ordering or compositional structure (i.e., the distribution of the different atomic species in the structural framework). For example, the surface chemistry of a metal particle and its interaction with the environment can be modified by coating with an overlayer of a different metal, or additional functionalities can be added to change its recognition capabilities. The mechanical strength, the plasmonic and nonlinear optical, as well as magnetic response also exhibit features in multicomponent nanostructured systems which are different from their pure counterparts. This allows one to finely modulate the physics of the metal-metal interactions and consequently to tune the properties of these systems nearly at will. The fundamental reasons of this varied behavior ultimately lie in the peculiar features of confined metallic bond [4]. In metallic nanomaterials, in fact, the conduction band electrons, whose mobility enhances transport and response properties, are confined within a finite space, thus producing finite-size effects. Moreover, in transition metals in addition to conduction band electrons, one also finds localized electrons in partially filled d-shells, which assure simultaneously easy formation and breaking of chemical bonds (and thus a rich and varied chemistry) and large magnetic moments. Finally, the sensitivity of all these appealing properties to size, shape, composition, and external fields make that metal nanoparticles and nanoalloys are really unique materials, an ideal topic of research and investigation in many fields of science and technology.

From the point of view of simulating this diverse behavior, aiming at a rational design and a full exploitation of nanoalloy properties, significant challenges must be faced.

For example, chemical ordering can be crucial in determining the system properties. Alloy nanoparticles exhibit a different optical response depending on whether they are core-shell or random solutions; mechanical properties are equally affected by surface segregation, and the precise arrangement of the species in the particle is important in catalysis, dominated by processes occurring at surface or subsurface shells [2]. Describing and predicting the static and dynamical elemental distribution within a nanoalloy system is an extremely difficult task due to the huge number of possible configurations.

As another example, the need of accurately describing the mobility and polarizability of the metallic conduction band electrons entails that sophisticated electronic structure methods must often be employed to describe these systems, methods which are computationally demanding, and are thus unfeasible for the larger nanoalloy systems.

Finally, it should be recalled that metal nanostructures are intrinsically metastable phases, tending toward the thermodynamically stable bulk phases if not hindered by protective layers

which act as a stabilizing barrier or other forms of stabilization, and the interplay between thermodynamic and kinetic effects creates additional complexity to nanoalloy simulations.

These issues will be briefly expounded in this essay.

Simulation Methods

In this section, we briefly present some of the theoretical and computational methods that are employed in the field of nanostructured alloys.

First-Principles Methods

The first observation concerns the fact that – given the relatively small size of nanoalloys in at least one dimension – the description is often atomistic, i.e., the coordinates of each atom constituting the nanostructure are specified. To investigate structural properties, what is then needed is an expression for the energy of the system as a function of coordinates: $E(\mathbf{X})$, where E is the energy and $\mathbf{X}$ is the atomic coordinates. Such an energy expression can be derived using different methods.

At the most fundamental level, one can use first-principles methods (i.e., methods not introducing parameters other than the basic physical constants) in which the energy $E(\mathbf{X})$ is obtained by solving the time-independent many-body Schrödinger equation [5]:

$$H\Psi(\mathbf{r},\mathbf{X}) = E\Psi(\mathbf{r},\mathbf{X}) \tag{1}$$

where H is the Hamiltonian operator and E and Ψ are its lowest eigenvalue (energy) and corresponding eigenvector (ground state wave function), respectively: Ψ depends on both electronic ($\mathbf{r}$) and nuclear ($\mathbf{X}$) coordinates. In practice, it is computationally unfeasible to solve this equation rigorously except for systems composed of few electrons and ions, and approximate approaches have to be used to deal with realistic systems. Among the available approaches, the most widely employed in the nanoalloy field is density-functional theory (DFT), as it realizes the best compromise between chemical accuracy and computational effort. DFT in its most common form implies solving the single-particle Kohn-Sham equations [6]:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r},\mathbf{X}) + \int\frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + V_{xc}[(\mathbf{r})](\mathbf{r})\right)\varphi_i(\mathbf{r}) = \lambda_i\,\varphi_i(\mathbf{r}) \tag{2}$$

where the Kohn-Sham Hamiltonian operator (in parenthesis) is expressed as a functional of the electronic density $\rho(\mathbf{r})$ of the system, while $\varphi_i(\mathbf{r})$ and λ_i are the corresponding single-particle wave functions and energies, respectively. The Kohn-Sham Hamiltonian contains contributions of kinetic energy, external potential, coulombic repulsion, and exchange-correlation potential. Despite being less computationally demanding than the complete Schrödinger equation, DFT is still a first-principles method with an explicit description of the electronic wave function and can only be used in practice when the size of the system is relatively small – say, between subnanometer and 3 nm as typical dimensions, even though advances in software and hardware are continuously extending the scope of systems amenable to accurate DFT modelling.

Analytic Potentials

For larger systems – say, between 2 and 10 nm in size – a common approximation relies on the use of analytic potentials. In these approaches, the energy function $E(\mathbf{X})$ is expressed in terms of analytic expressions of the atomic coordinates. As the number of such coordinates is huge, further simplifications must be sought. A common practice is then first to define $E(\mathbf{X})$ as a sum of site atomic energies $E^{i,atomic}(\mathbf{X})$:

$$E = \sum\nolimits_i E^{i,\ atomic}(\mathbf{X}) \tag{3}$$

where the sum runs over all *i*-atoms, and then to express the dependence of $E^{i,\ atomic}(\mathbf{X})$ upon $\mathbf{X}$ through generalized collective variables (GCV), i.e., a limited number of functions which describe the distribution of neighbors around the given atom and depend in an analytically compact form upon their number, distance, and orientation.

What distinguishes the various analytic potentials is then the choice of the GCV and the analytic form by which $E^{i,\ atomic}$ depends on the GCV. In this respect, it has been shown that, to be accurate, an analytic potential for nanoalloys must satisfy two requirements and reproduce two salient features of the metallic bond:

(i) Its incremental character, i.e., the possibility of spanning a wide interval of coordination numbers (number of close neighbors), which in metallic systems typically range from 1 to 12 (or 13–14 in some cases)
(ii) Its many-body or nonlinear character, that is, the fact that atomic energies are not simply linear or additive, but go to saturation as the number of neighbors increases

These features can be illustrated by considering one of the simplest analytic potentials for metal systems: the second-moment-approximation (SMA) tight-binding potential, in which each atomic energy is written as a sum of a pair repulsive ($E^{i,rep}$) and a many-body attractive ($E^{i,\ att}$) component:

$$E = \sum_{i=1}^{N} \left\{ E^{i,\,rep}(i) + E^{i,\,att}(i) \right\} \tag{4}$$

where

$$E^{i,\,rep}(i) = \sum_{j \neq i}^{N} A \exp\left\{ -p\left(\frac{r_{ij}}{r_0} - 1 \right) \right\} \tag{5}$$

$$E^{i,\,att}(i) = \left[\sum_{j \neq i}^{N} \xi^2 \exp\left\{ -2q\left(\frac{r_{ij}}{r_0} \right) - 1 \right\} \right]^{1/2} \tag{6}$$

and the sums in Eqs. 5 and 6 run over all *j*-atoms different from the *i*-th. The parameters A, r_0, ξ, p, and q are characteristic of the given metal element and are usually fitted to the experimental quantities such as the cohesive energy, the equilibrium lattice parameter, and independent elastic constants for the reference crystal structure. It can be noted that in the SMA analytic potential, the repulsive term in $E^{i,atomic}(\mathbf{X})$ is of the Born-Maier form [7] and is linear, whereas the attractive term describes the many-body character of the metallic bond and is nonlinear as it is proportional to the square root of a sum over neighbors, thus satisfying requirement (ii) above. When working with nanoalloys, the Eqs. 4, 5, and 6 are generalized by defining three types of pair interactions – A-A, B-B, and A-B – and the corresponding A, r_0, ξ, p, and q parameters for each pair. Usually the parameters corresponding to the pure A and B elements are employed for the homo-interactions (A-A and B-B types), while the A-B hetero-interactions are described by a different set that can be derived by averaging the parameters of the pure elements or by fitting to (experimental or computational) data corresponding to bulk alloys. The analytic forms (4, 5, and 6) can be further refined by taking into account directionality effects in the choice of the GCV [8]: these effects are important when dealing with third-row transition metals where, due the relativistic contraction of the *s*-orbitals, *d*-orbitals (which are not isotropically distributed around the nucleus) play a role in metallic binding. As an example, the dipolar contribution of the neighbor distribution around the *i*-th atom is given by the following expression:

$$E^{dip,\,i} = \sum_{\sigma = x,\,y,\,z} \left(\sum_{j \neq i} \rho^{dip}\left(r_{ij}\right) \frac{r_{ij}^{\sigma}}{r_{ij}} \right)^2 \tag{7}$$

where $\rho^{dip}(r_{ij})$ is a function of the distance and r_{ij}^{σ} is the σ-Cartesian component of the interatomic distance r_{ij}.

Effective Hamiltonians

For even larger systems – say, between 5 and 50 nm in size – a third class of methods are often used, named effective Hamiltonian or pair-interaction methods [9]. In these approaches, the energy expression $E(\mathbf{X})$ is further simplified by ignoring the dependence of the atomic site energies upon the atomic coordinates and assuming that they only depend on the number of first neighbors, defined as atoms lying at a distance

smaller than a given threshold from the given *i*-th one. According to such scheme, each atom in the bulk gives the same contribution, whereas the surface energy of the particle can be written in the following way:

$$E_s = \sum_{i=1}^{N} \frac{1}{2}\left\{ \sum_{j \neq i}^{N} \left(\delta w_{ij} - \Delta Z_i w_b \right) \right\} \qquad (8)$$

where w_b is the bulk energy, ΔZ_i is the number of broken bonds of atom *i*, and δw_{ij} is a polynomial function also depending on the number of broken bonds. The functions w_b and δw_{ij} can be fitted to first-principles calculations or experimental values of, e.g., surface energies. These approaches can be applied to larger alloy nanoparticles. In such methods, in fact, an energy evaluation scale linearly with the number of atoms with a small prefactor related to the average number of first neighbors and is thus feasible for systems containing billions of atoms in a cpu time of the order of a second on a current standard processor. A given structural framework for the nanostructure is considered, and the main underlying assumption is that the structural relaxation around each atom only depends upon its coordination number, an assumption which is usually reasonable when the dimensions of the nanostructure exceed 4–5 nm.

As anticipated at the beginning of this section, only atomistic approaches have been discussed so far, which are appropriate due to the limited size of the systems. However, for the larger nanoalloys with sizes in the range of several tens of nanometers, continuum models and associated finite element solvers can in principle be employed [10], although this possibility has been rarely considered until now.

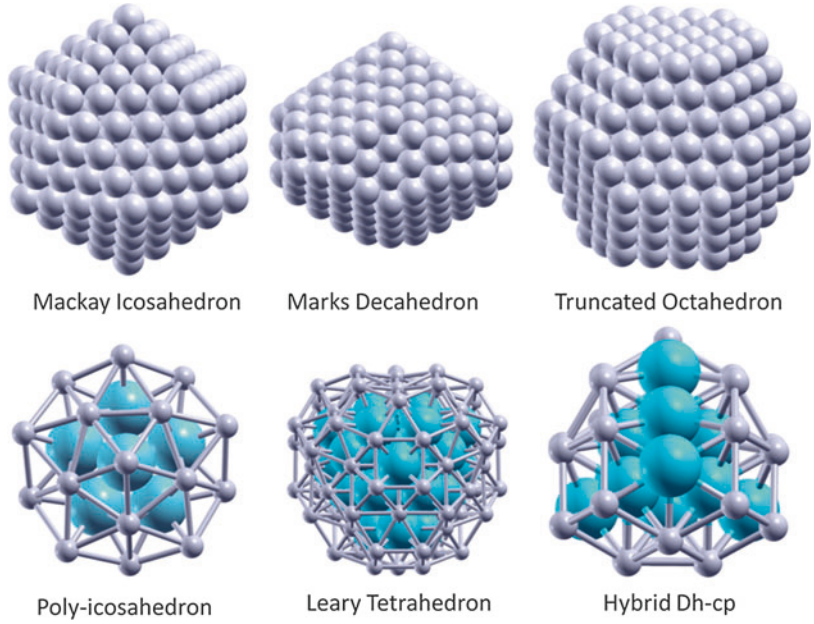
To conclude this section, two points can be highlighted.

First, in nanoalloys a complication exists connected with the *match* among the energy scales of the different elements composing the alloy. The presence of multiple length and energy scales is common to any composite system and must be dealt with care in order to achieve reliable predictions. Let us clarify this point via a specific example. If a given theoretical approach is not capable of predicting the surface energies of a pure element accurately, but it does predict the *ratios* between surface energies correctly, this is sufficient to predict the shape of a pure nanoparticle accurately. The shape of a nanoparticle is determined in fact only by the ratios of the surface energies, not by their absolute values, and multiplying all energies by a common scaling factor does not affect these ratios. However, if the scaling factors for the different elements in a mixed system are different, this will spoil the accuracy of the prediction. In nanoalloys an *absolute accuracy* in the description of the properties of each element (rather than a *relative accuracy*) must therefore be achieved, and this is a much more stringent demand on the theoretical approach.

Second, a problem in the field of nanoalloy simulation is how to compare the stability of nanoalloys with different chemical composition. In this respect, a very useful quantity is the mixing energy. Suppose to consider alloy nanoparticles of a given size (i.e., the same number of atoms) but different chemical composition, and suppose to know the energy of each one of them. The mixing energy is then defined as

$$\Delta[N_A, N_B] = E_{alloy}[N_A, N_B] - N_A E_A[N] / N - N_B E_B[N] / N \qquad (9)$$

where $E_{alloy}[N_A, N_B]$ is the energy of a nanoalloy composed of N_A atoms of the species A and N_B atoms of the species B, $N = N_A + N_B$ is the total number of atoms in the nanoalloy, E_A [N] is the energy of a pure structure of N atoms, and E_B [N] is the corresponding quantity for the B species (it is assumed that these are global minimum energies, i.e., the lowest energies among all possible isomers). $\Delta[N_A, N_B]$ is the mixing energy of a nanoalloy of the given size at the given composition, i.e., the energy released by mixing N_A A-type clusters and N_B B-type clusters all of size N to produce N alloyed clusters of size N with composition (N_A, N_B), divided by N. The quantity $\Delta[N_A, N_B]$ provides a measure of how



Nanoalloy Simulation, Fig. 1 Classical structural motifs (first row), together with structural motifs peculiar to nanoalloyed systems (second row). In particular, in the second row, a poly-icosahedron of 34 atoms (a 27-7 according to the definition given in Ref. [11]), a 98-atom Leary tetrahedron, and a 38-atom hybrid decahedral (Dh)/close-packed (cp) are shown

thermodynamically favorable is mixing the two elements at the given size and composition. Its definition represents the natural generalization of the corresponding quantity used for bulk alloys [2].

Structural Properties

Surfaces and interfaces play a crucial role in nanostructured systems, due to the large ratio between the number of atoms at the surface with respect to those in the bulk or core of the nanostructure. In such systems, the energy associated with surfaces or interfaces thus represents a sizeable component of the total energy. Moreover, some properties, such as catalytic ones, are entirely dominated by local chemical binding at the surface. The particle shape, determining the type of surfaces of the particle, is usually described in terms of model polyhedra from classical geometry, such as octahedra and truncated octahedra (crystalline motifs) or icosahedra and decahedra (noncrystalline motifs); see the first row of Fig. 1. Noncrystalline motifs cannot give rise to periodic patterns because they contain fivefold symmetry axes which are incompatible with translational symmetry, but they can exist in finite systems such as nanoparticles. The Mackay icosahedron is probably the most common one; see Fig. 1: this structure is limited by (111) close-packed facets only, which are usually the surfaces with the lowest energy, thus minimizing the cluster surface energy. However, in icosahedral structure, an internal strain is present due to the fact that radial inter-shell bonds are compressed, whereas intra-shell bonds are expanded. Mackay icosahedra are therefore highly stressed structures and are expected to be the favorable only at small sizes, in which the low surface energy overcomes the internal stress contribution. Another noncrystalline motif is represented by the decahedron: this structure is formed by two pentagonal pyramids sharing a base; its surface has only (111) close-packed facets, but its shape is very far from being spherical. A truncation of the five edges limiting the common basis of the pyramids to expose (100)-like facets is advantageous and produces the so-called Ino decahedron. Other cuts with reentrances exposing further (111) close-packed facets separating neighboring (100)-like facets can also be advantageous and produce

Marks decahedra; see Fig. 1. Decahedra are also strained structures, but the strain is much smaller than for icosahedra, making them favorable in an intermediate range of sizes [12].

In the family of crystalline motifs, the atoms of the clusters are positioned in the positions characterizing the metal in its bulk form (e.g., face-centered cubic, fcc). These motifs optimize volume energy contributions at the expense of a higher surface energy with respect to other motifs, so that they are thus preferred for large particles in which the volume contribution dominates. Other more exotic motifs can be encountered especially in the nanoalloy field, in which the mixture of different elements greatly increases structural degrees of freedom. Among these, a notable class is that of poly-icosahedra (pIh), which are built by packing elementary icosahedra (Ih) of 13 atoms in such a way that interior atoms in the structure are surrounded by a complete Ih_{13} shell (see second row of Fig. 1). Another common occurrence is represented by hybrid motifs, such as hybrid decahedral/close-packed structures, in which a close-packed core in the form of a trigonal pyramid or bipyramid is surrounded by atoms growing on the core faces in hcp(111) stacking plus further atoms arranged along edges between two faces thus creating a local decahedral arrangement [2]. The occurrence of these more exotic (such as hybrid) motifs in alloyed nanoparticles can be rationalized. Suppose, for example, to mix two species with different bulk cohesive energies. If the size mismatch between the two species is small or such that the more cohesive element also presents the smaller size and the two elements do not have a great tendency to mix, it is reasonable to expect a core-shell arrangement with a core segregation of the smaller species. Moreover, as this species achieves a complete coordination shell in the core, it will tend to arrange in a close-packed (cp) configuration. The less cohesive species, instead, being surface segregated and thus low coordinated, will tend to adopt noncrystalline, fivefold arrangements. Hybrid structures which combine a close-packed core with a decahedral or icosahedral surface can therefore be favorable.

Chemical Ordering

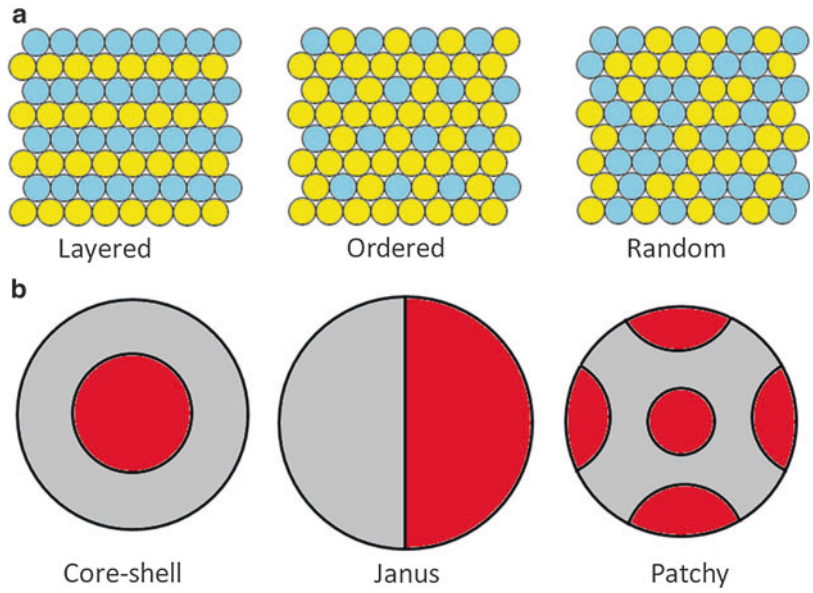
As recalled above, in addition to shape, specific to nanoalloys are the issues related with chemical ordering or compositional structure, i.e., the way in which elements are distributed within the nanostructure.

Let us focus for simplicity on a binary nanoalloy.

In the case of homogeneous systems, when the chemical ordering is uniform throughout the nanostructure, we can distinguish three variants of mixing patterns, depicted in Fig. 2a: (a1) layered, in which atomic planes composed of each component alternate regularly; (a2) ordered mixing, with a regular arrangement and a thorough intermingling of the two species; and (a3) random, i.e., mixing in which no long-range order can be recognized. These variants closely correspond to chemical orderings found in extended multicomponent systems (bulk alloys).

In the case of inhomogeneous or segregated systems, in which the nanostructure can be partitioned into disconnected, individually homogeneous pieces, three main segregation patterns can be defined, illustrated in Fig. 2b: (b1) core-shell, in which one piece (the inner core) is surrounded by a crown (or shell) of a different composition; (b2) Janus, in which the dividing interface corresponds to a plane bisecting the structure; and (b3) patchy, in which the surface of the nanostructure is a patchwork of islands of atoms of the two elements. Of course, there are intermediate cases in between these limiting ones, with variants such as multi-shell, in which the particle can be partitioned in more than one concentric shell, or off-centered core-shell structures, in which the core is asymmetrically centered and in the extreme can give rise to quasi-Janus arrangements, in which the core is so close to the surface that is covered only by a very thin layer with a different composition.

Clearly, each segregated piece depicted in Fig. 2b can be associated with one of the different mixing patterns depicted in Fig. 2a, thus giving rise to a multitude of different chemical ordering



Nanoalloy Simulation, Fig. 2 (**a**) Short-range mixing patterns in homogeneous systems and (**b**) possible realizations of segregated systems. It should be noted that the figures on the left and in the middle of part (**b**) represent a cross-section of the particle, whereas the figure on the right shows a surface view

possibilities. An important topic in nanoalloy simulation is therefore how to single out which chemical ordering is realized in a given nanoalloy (i.e., structural recognition). Order parameters can be defined to achieve this. An order parameter is an analytic function of the system coordinates that describes the structural features of the system. Commonly used order parameters are those describing the coordination environment of the atoms, such as those derived from a common neighbor analysis [13]. Other order parameters associated with chemical order or compositional structure can be defined to distinguish the ordered phases of bulk alloys and nanoalloys [14]. For example, an order parameter useful to characterize the short-range intermixing of two species is the following. Suppose that x_A and x_B are the relative quantities of the two components within an A-B nanoalloy and n_{AA} and n_{AB} the average number of homo- and hetero-bonds, respectively, around the species A. By defining $n_{AM} = n_{AA} + n_{AB}$, the α parameter

$$\alpha = 1 - \frac{n_{AB}/n_{AM}}{x_B} \tag{10}$$

becomes an indication of the tendency of the nanoalloy to segregate (positive values of α) or to mix (negative values of α). More sophisticated order parameters take into account not only the number of nearest neighbors of each species but also their relative orientation: for example, one useful order parameter is the dipolar contribution coming from the distribution of first neighbors around each atom; see its form in (7) [8]. Many other possible order parameters can of course be defined, tailored to the specific system under consideration. The order parameters can be used as averages over the whole nanoalloy or locally for each individual atom. In the case of homogeneous nanoalloys, the order parameters relative to chemical ordering will be independent of the position, whereas in the case of inhomogeneous or segregated systems, the value of such order parameters will vary within the structure.

Given the huge variety of possible structural families and compositional structures, one of the main issues in the theoretical research and simulation on nanoalloys is the accurate prediction of their favorable shape and degree of atomic mixing. As in other field of computational nanoscience, the problem is finding effective ways to systematically explore the potential energy surface (i.e., the $E(\mathbf{X})$ function) of the system [15]. This computational problem can be daunting, given the incremental and scarcely directional character of the metallic bond and the additional complication that chemical ordering

greatly increases the degrees of freedom of the system. In terms of structural framework, it is believed that the number of local minima increases exponentially with the size of the nanoparticle [15]. These different local minima give rise to structural isomers, such as icosahedra, decahedra, etc., as shown in Fig. 1, with all their numerous nonsymmetric variants. Chemical ordering also increases the complexity of the problem in a non-polynomial (exponential) way. In detail, the number of different possible realizations of a nanoalloy of a given size, structural framework, and composition increases in a combinatorial way due to the set of possible chemical orderings. In fact, neglecting the reduction due to point-group symmetry, for a cluster with a given structure and made of N atoms of which N_A of species A and N_B of species B, one has $(N_{tot})!/((N_A)!(N_B)!)$ possible different "homotops," i.e., isomers sharing the same skeletal structure and composition but differing in the mixing pattern. This number increases exponentially with the size of the system, rapidly becoming impossible to thoroughly sample or explore even using computationally inexpensive theoretical approaches such as effective Hamiltonians. In the following, the methods employed to predict the preferred (energetically most favorable) geometric and compositional structures of nanoalloys are briefly discussed.

Global optimization (GO) approaches: a viable procedure to search for the equilibrium configurations of a pure or alloy system is to start from a random arrangement and to consider moves eventually leading to lower-energy configurations, as realized in the basin-hopping (BH) algorithm [15]. The BH algorithm to explore the structural degrees of freedom of the system is realized as follows:

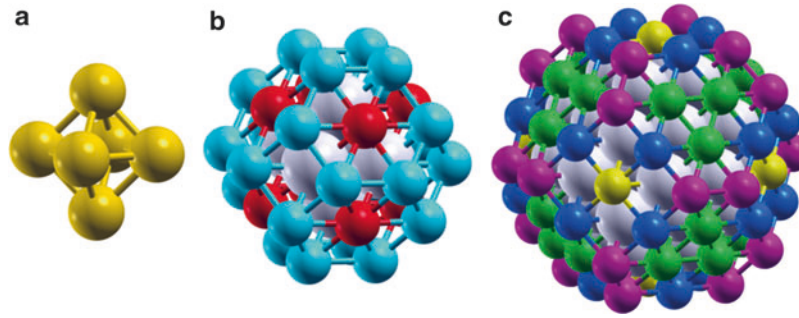
1. An initial random configuration of the system is chosen, a local geometry optimization is performed, and the final energy (the fitness parameter) is registered as E_1.
2. Starting from the relaxed configuration, the atoms of the system are subjected to a random move, a new local geometry optimization is performed, and the final energy is registered as E_2.
3. A random number (rndm) between 0 and 1 is generated, and the movement of step 2 is accepted only if $\exp[(E_1 - E_2)/k_B T] > \text{rndm}$ (Metropolis criterion).
4. Steps 2 and 3 – the Monte Carlo steps – are repeated a given number of times.

Depending on the k_BT parameter, which plays the role of a fictitious temperature, some high-energy configurations are accepted, and the search is able to explore different structural motifs (belonging to different funnels of the PES) of the metal cluster.

A crucial point in the implementation of the BH search is choice of the random move. This can be simply a spatial "kick," realized by randomly displacing the Cartesian coordinates of all atoms up to a certain amount, or more sophisticated choices, such as short runs of unstable molecular dynamics especially effective for larger clusters. The BH approach can be also used to explore the space of chemical ordering. In this context, the BH random move is usually taken as an exchange of the coordinates of two atomic species, as realized in the exchange-only basin-hopping (BH) algorithm.

The thoroughness of the BH search is much increased when the BH protocol is coupled with *structural recognition* algorithms, i.e., algorithms which classify geometrical configurations as belonging to a given structural motif (or structural family). Structural recognition makes use of the order parameters discussed above by distinguishing different structural families on the basis of the corresponding values of a chosen order parameter.

An alternative possibility is to consider structurally magic clusters, i.e., configurations exhibiting structural shell closure and thus as a norm a high point-group symmetry. This high symmetry is exploited to partition the atoms into symmetry "orbits," i.e., group of atoms equivalent by symmetry (transformed into one another by the operation of the symmetry group); see Fig. 3. As a consequence, the degrees of freedom of the system are reduced from the number of atoms N to the number of symmetry-nonequivalent orbits, N_{orb}, and correspondingly the number of distinct



Nanoalloy Simulation, Fig. 3 Orbit structure of selected truncated octahedral (TO) clusters: (**a**) 6-atom TO; (**b**) 38-atom TO, obtained adding 32 atoms to (**a**); (**c**) 116-atom TO, obtained adding 78 atoms to (**b**). Atoms belonging to different symmetry orbits are singled out by different colors

homotops is exponentially decreased, making theoretical simulations feasible even at the first-principles level for medium-sized particles. The use of magic clusters is also very useful for studying properties, both to reduce the computational effort associated with a single computation and to reduce the number of homotops to be considered.

Another efficient tool to drastically reduce the computational effort and to enlarge the scope of systems amenable to simulations is based on the use of periodic models. Instead of studying a 0D system, in fact, one can consider a 1D or "wire" model in which the section of the cluster is replicated ad infinitum and the number of nonequivalent atoms (in this case, those contained in the unit cell) is strongly reduced.

Catalytic Properties

So far static structural properties have been discussed, connected with the energetically most favorable configurations. One of the most important applications of nanoalloys and therefore of the corresponding simulation methods is however in the field of chemical reactions and in particular of catalysis, which involves *dynamic* in addition to *static* structural properties. Nanoalloy systems are indeed very relevant in catalysis and can be very efficient catalysts. A catalyst is a specific agent which is (i) able to increase the rate of a chemical reaction and (ii) which is not consumed by the transformation process. The important catalytic properties of nanoalloys are associated with the following features: (a) the incremental character of metallic bond, spanning a wide interval of coordination numbers (as recalled above), and the great polarizability of metallic electrons make that metal surfaces in general can promote chemical reactions by adsorbing/transforming/desorbing ligand species; (b) the multifunctional nature of nanoalloys as mixtures of different metals greatly expands catalytic degrees of freedom by creating geometrically or electronically novel adsorption sites; and (c) the high surface/volume ratio of nanoalloys as nanoscale systems makes them particularly efficient in catalysis with respect to extended crystals. The much higher selectivity of Cu-Ru particles with respect to pure Ru ones in orienting the dehydrogenation of cyclohexane into benzene rather than its hydrogenolysis into methane [16] and the greater efficiency of Ni_3Fe with respect to pure Ni systems as methanation catalyst [17] are examples of the important of nanoalloys in the catalysis field.

To simulate chemical reactivity, theoretical methods are required to be able to predict not only the minimum energy of a system but also the *minimum energy path* connecting an initial configuration (the reactants) with a final configuration (the products). One can distinguish two cases, according to whether one has a knowledge of the initial and final configurations or local minima to be connected or not. In the former case, the

nudged elastic band (NEB) method [18] is one of the most popular algorithms to find the reaction path between the two configurations and is implemented in many simulation codes. If instead knowledge on the final configuration is lacking, then the search for the local minima connected to the starting structure must use other methods, such as the intrinsic reaction coordinate (IRC) algorithm [18].

Another form of structural dynamics which is specific to nanoalloys is the kinetics of chemical ordering, i.e., the rate at which the different atoms interdiffuse within the alloy nanostructure. Information on interdiffusion kinetics is important to ascertain whether the thermodynamic predictions obtained via global optimization simulations and the favorable arrangements of the elemental species at the surface are then realized in practice. It often happens in fact that after a synthesis process, the as-prepared alloy nanoparticles are in a metastable phase or configuration, for example, with a distribution of the two elements which does not correspond to the lowest-energy or thermodynamically favorable arrangement. In such cases, their chemical ordering will evolve in time according to the kinetics of atomic interdiffusion. This phenomenon from a simulation point of view is akin to catalysis and can be treated using similar theoretical approaches [2].

Even assuming that the potential energy surface of a given system can be explored in an efficient way, a basic issue remains in modelling catalysis by nanoalloys that screening possible nanoalloy catalysts to find the optimal combination and composition via a trial-and-error approach is an extremely demanding task. This is true both at the experimental and theoretical level, even when combinatorial screening can be applied [19]. The number of possible binary pairs of an alloyed catalyst and of the possible different compositions is in fact huge, and some guiding principles are needed to orient the search. In this context, the concept of descriptors of the catalytic activity is commonly used.

The starting point is a basic principle in catalysis, the Sabatier principle [20], which states that, when trying to optimize the performance of any given catalyst, one faces as a norm conflicting requirements. For example, if the adsorption energy of the reactants on the catalyst is too strong (or they dissociate too easily), then the successive reaction energy barriers or desorption energies will also be large, whereas vice versa if the adsorption energy is too low, then the dissociation step necessary for the catalytic activity will be difficult. The Sabatier principle then states that the best compromise lies in between strong and weak energetics, and the performance of a catalyst as a function of any quantity correlated with the reaction energetics will go through a maximum, leading to a "volcano" relationship between catalytic efficiency and this quantity.

Another crucial point is the finding that linear correlations usually exist between different quantities such as adsorption energies and energy barriers, also called Brønsted-Evans-Polanyi (BEP) relations. The presence of such correlations effectively reduces the number of independent variables for a given catalytic process.

The strategy that results from combining the Sabatier principle and the existence of BEP relations is therefore to select a limited number of "descriptors," i.e., energy quantities that monitor the catalytic activity or selectivity of a catalyst [17], and plot catalytic efficiency as a function of such descriptors. These descriptors can be the adsorption energies of intermediates, reactants, and products or the activation energies of the rate-determining elementary steps, decided on the basis of experimental and/or theoretical results. In plotting the catalytic efficiency as a function of descriptors, one then expects to find a volcano curve, i.e., a curve with a maximum. In the search of new catalysts, one can then focus on the vicinity of the maximum of the volcano curve, which is the most interesting region, thus drastically reducing the effort of the simulation screening.

Conclusions

In conclusion, simulations can now be conducted and used as an effective tool to characterize and

predict the structure and properties of nanoalloys, i.e., metal alloy materials consisting of structures with at least one dimension at the nanoscale. The palettes of methods used in this field are mostly atomistic and range from sophisticated first-principles approaches to simpler analytic potentials to effective Hamiltonians. Given the striking structural, catalytic, mechanical, optical, and magnetic properties of these materials and their consequent use in fuel cells, energy storage, nanomedicine, magnetic nanostructures, plasmonics, and therapeutic and diagnostic applications [2], simulation methods are rapidly evolving, trying to keep the pace of experimental and technological developments. In the present essay, attention has been devoted to the simulation of structural behavior of nanoalloy systems, both static (i.e., favored arrangements and configurations) and dynamic (i.e., transformation kinetics, chemical reaction, and catalysis), whereas information on a wider set of properties and corresponding theoretical approaches can be found in the cross-reference essays.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Bond-Order Potential
- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Integrated Approach for the Rational Design of Nanoparticles
- ▶ Magnetic Nanoparticles for biomedical applications
- ▶ Magnetic Nanostructures and Spintronics
- ▶ Mechanical Properties of Nanocrystalline Metals
- ▶ Metal Nanoparticles from First Principles
- ▶ Multiscale Modeling
- ▶ Nanocluster
- ▶ Nanoscale Particle
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Optical Response of Metal Particles
- ▶ Optimization of Nanoparticles
- ▶ Simulation of Supported Metal Clusters

References

1. Mariscal, M.M., Oviedo, O.A., Leiva, E.P.M. (eds.): Metal Clusters and Nanoalloys – From Modeling to Applications. Springer, New York (2012)
2. Barcaro, G., Caro, A., Fortunelli, A.: Alloys on the nanoscale. In: Vajtai, R. (ed.) Springer Handbook: Nanomaterials. Springer, Berlin (2013)
3. Jellinek, J., Krissinel, E.B.: Ni_nAl_m alloy clusters: analysis of structural forms and their energy ordering. Chem Phys Lett **258**, 283 (1996)
4. Pauling, L.: The Nature of the Chemical Bond. Cornell University Press, New York (1960)
5. Mahan, G.D.: Quantum Mechanics in a Nutshell. Princeton University Press, Princeton (2008)
6. Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Phys Rev **140**, 1133A (1965)
7. Finnis, M.: Interatomic Forces in Condensed Matter. Oxford Series on Materials Modelling. Oxford University Press, Oxford (2003)
8. Baskes, M.I.: Application of the embedded-atom method to covalent materials: a semiempirical potential for silicon. Phys Rev Lett **59**, 2666 (1987)
9. Ruban, A.V., Abrikosov, A.I.: Configurational thermodynamics of alloys from first principles: effective cluster interactions. Rep Progr Phys **71**, 046501 (2008)
10. Tadmor, E.B., Miller, R.E.: Modeling Materials: Continuum, Atomistic and Multiscale Techniques. Cambridge University Press, Cambridge (2011)
11. Rossi, G., Rapallo, A., Mottet, C., Fortunelli, A., Baletto, F., Ferrando, R.: Magic polyicosahedral core-shell clusters. Phys Rev Lett **93**, 105503 (2004)
12. Baletto, F., Ferrando, R., Fortunelli, A., Montalenti, F., Mottet, C.: Crossover among structural motifs in transition and noble-metal clusters. J Chem Phys **116**, 3856 (2002)
13. Honeycutt, J.D., Andersen, H.C.: Molecular dynamics study of melting and freezing of small Lennard-Jones clusters. J Phys Chem **91**, 4950 (1987)
14. Atanasov, I.V., Hou, M.: Equilibrium ordering properties of Au-Pd alloys and nanoalloys. Surf Sci **603**, 2639 (2009)
15. Wales, D.J.: Energy Landscapes: Applications to Clusters, Biomolecules and Glasses. Cambridge University Press, Cambridge (2004)
16. Sinfelt, J.H.: Catalysis by alloys and bimetallic clusters. Acc Chem Res **10**, 15 (1977)
17. Nørskov, J.K., Bligaard, T., Rossmeisl, J., Christensen, C.H.: Towards the computational design of solid catalysts. Nat Chem **1**, 37 (2009)
18. Henkelman, G.; Jóhannesson, G.; Jónsson, H.: Methods for finding saddle points and minimum energy paths. In: Schwartz, S.D. (ed) Theoretical Methods in Condensed Phase Chemistry. Springer, Amsterdam (2002)
19. Guerin, S., Hayden, B.E., Lee, C.E., Mormiche, C., Russell, A.E.: High-throughput synthesis and

screening of ternary metal alloys for electrocatalysis. J Phys Chem B **110**, 14355 (2006)
20. Sabatier, P.: Hydrogénations et déshydrogénations par catalyse. Ber Deutsch Chem Gesellshaft **44**, 1984 (1911)

Nanoantennas

▶ Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy

Nanoassembly

▶ Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly
▶ Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes
▶ Ligand-Directed Gold-Phage Nanosystems

Nanobelts

▶ Chemical Vapor Deposition (CVD)

Nanobiosensors

▶ Biosensors

Nanobonding

▶ Nanorobotic Spot Welding

Nanocalorimeter

▶ Nanocalorimetry

Nanocalorimetric Sensors

▶ Nanocalorimetry

Nanocalorimetry

Jean-Luc Garden and Olivier Bourgeois
Institut Néel CNRS-UJF, Grenoble, France

Synonyms

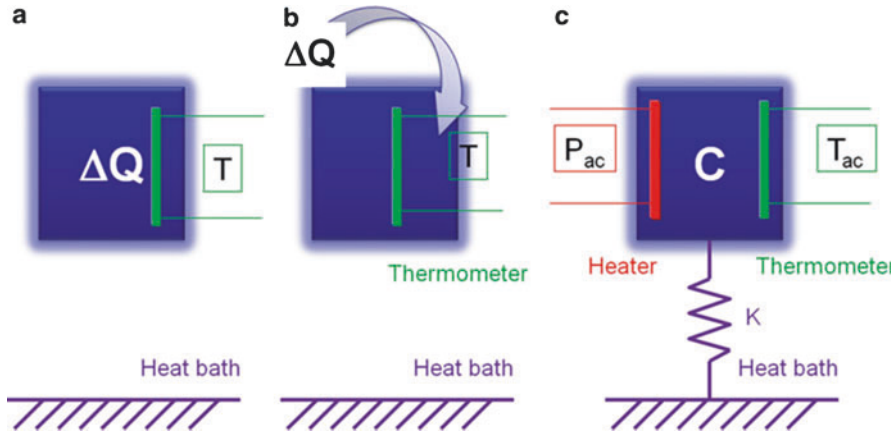
ac-Calorimetry; Attocalorimetry; Biocalorimetry; Differential scanning calorimetry; Femtocalorimetry; Heat capacity; Isothermal calorimetry; Microcalorimetry; Nanocalorimeter; Nanocalorimetric sensors; Nanoscale thermal analysis; Nanotechnology; Relaxation calorimetry; Thermodynamics of small systems; Thermometry

Definition

Calorimetry is the part of thermodynamics which aims to measure any quantity of heat (enthalpy, specific heat, heat release) stored, released or brought into play in any state of matter, in a reaction, or in phase transitions [1].

More precisely, the terminology of "Nanocalorimetry" may cover different concepts depending on the area of science where it is used. It concerns any calorimetric method in which either the samples to be studied have a size in the range of the nanometer scale or the measured energies involved are of the order of the nanojoule or below.

Although these two types of definitions are very different, they both have something in common. They are indeed concerned with the development of miniaturized sensors and more generally with ultrasensitive experiments imposed when small systems are involved. Indeed, most of the modern sensors are built to carry out measurements at the ultimate limit of



Nanocalorimetry, Fig. 1 Thermal schemes for calorimetric experiments (**a**) a quantity of heat ΔQ is released in the sample, an increase of temperature will appear. This system is assumed to be infinitely isolated from the heat bath. (**b**) An external heat is supplied to the sample also infinitely isolated from the heat bath. By measuring the increase of temperature we will get access to the heat capacity of the sample (adiabatic calorimetry). (**c**) The sample is linked to the heat bath through a thermal conductance K; a thermometer and a heater allow the measurement of the thermal properties (C and K)

what can be detected by calorimetry on thin films or very small volume samples.

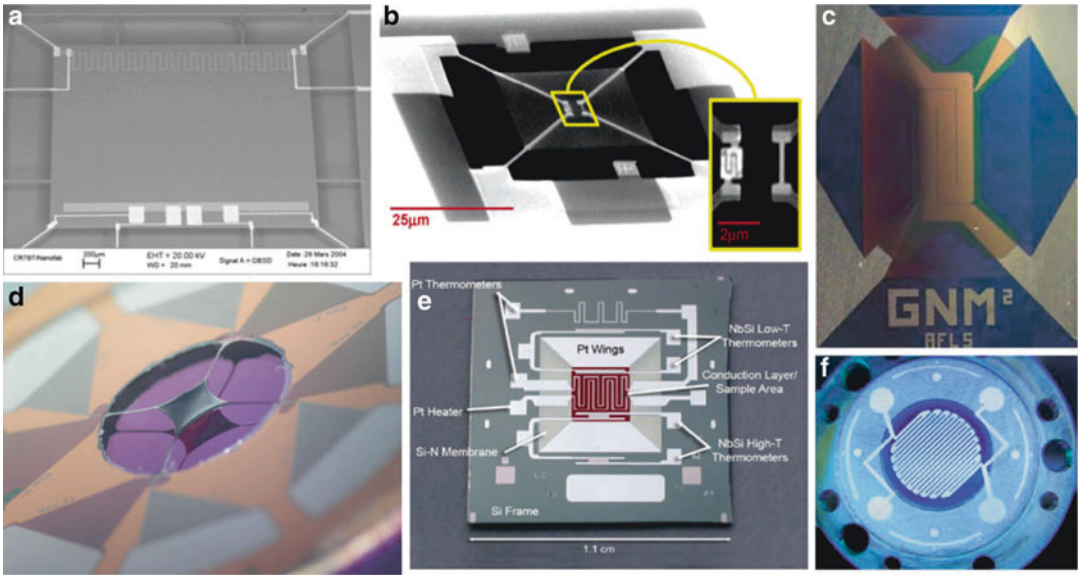
In this entry, the key issues governing an experiment of nanocalorimetry are discussed. In particular, for each point, recent experiments conducted in this area over the world are presented. This will provide a non-exhaustive overview of what is currently done in nanocalorimetry. Particular attention will be given to micro- and nanofabrication technologies as well as highly sensitive thermal technique necessary to achieve an experiment of nanocalorimetry. Here, bolometry is not discussed; despite its thematic proximity; the measurement of radiation and possible refrigeration at the mesoscopic scale has been described extensively in other works [2].

Nanocalorimetry

Introduction

In calorimetry, thermal isolation is the major issue. The heat capacity of a sample under study, or the measurement of exchanges of energy between the sample and its environment, is properly measured only if the sample is correctly isolated from its environment (adiabatic conditions). In practice, the thermal isolation is not perfect, and one overcomes this problem by calibrating correctly the thermal link between the sample and its surroundings. As discussed later in this entry, the thermal relaxation rate, τ, which determines the condition of adiabaticity, is a crucial parameter for the selection of measurement methods adapted to the physics to be studied (see the section "Experimental Techniques"). This problem of thermal isolation is more problematic in nanocalorimetry because the samples are very small in size and therefore have small masses and thus very small heat capacities. Therefore, to fulfill the requirement of adiabaticity, the experimentalist makes thin suspended membranes micrometer thick that will support small objects to be studied (see Fig. 1).

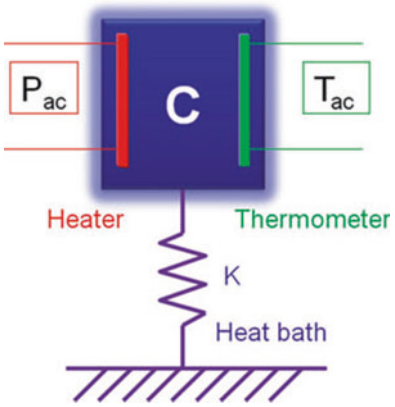
One of the advantage of working with membrane sensor is the reduction of addenda (the sample holder and the sensitive measuring elements are called "addenda"); this is especially important when working with very small systems. The small thickness of the membrane reduces also the thermal coupling to the outside providing then thermal isolation. Some membranes are structured to further limit the exchange of heat between the measurement area and the thermal bath. In this case, they look like micro-trampolines suspended by arms that allow the passage of current leads (see Fig. 2).



Nanocalorimetry, Fig. 2 Various sensors made in silicon, silicon nitride, or polymer. (**a**) Silicon membrane for very low temperature heat capacity measurement by ac calorimetry [3]. (**b**) Silicon nitride sensor made by e-beam lithography used for relaxation calorimetry [4]. (**c**) Silicon nitride sensor used for fast scanning calorimetry [5]. (**d**) Polyparaxylylene membrane for phase transition detection in thin magnetic films around 100 K [6]. (**e**) Silicon nitride sensor used for relaxation calorimetry over a wide range of temperature (Cooke 2010). (**f**) Polyimide membrane for heat capacity measurement of thin polymer films at ambient temperature [7]

Another point is that, since in the case of nanocalorimetry the amount of energy to be measured is small, the resolution of the calorimetric measurement must be sufficiently high to access the expected thermal properties. A first means to increase the signal-to-noise ratio (SNR) of the measurement is to reduce the noise of the detected signals. In other words, one must develop a low noise electronic chain adapted to the chosen experimental technique and also adapted to the sensor converting temperature changes in measurable signals (usually voltage). This aspect, which concerns the sensitive measurements and thermometry, is discussed later on (see the section "Thermometry"). A second way is to reduce significantly the heat capacity of sample holder and measuring devices (thermometers and heaters).

In this case, the use of nanotechnology and microfabrication is indispensable for the realization of sample holder for which the heat capacity is at least equivalent (or smaller) than the sample heat capacity being studied. The use of these technologies is crucial for measuring heat capacities of objects of very small size or for the



Nanocalorimetry, Fig. 3 Scheme of the principle of a nanocalorimetric measurement on membrane

detection of low energy in relevant thermodynamic transformations. Therefore, micro- and nanotechnologies are essential keys to nanocalorimetry. This justifies the use of thin microfabricated membranes for thermal isolation (see Fig. 3). Each membrane contains the sensitive elements. These elements are micromachined in thin layers of sub-micrometer

thicknesses deposited by vacuum evaporation technique or by magnetron sputtering. Microphotolithography techniques are used to shape the geometry of the sensitive elements and define their impedances, and therefore define the sensitivity of the calorimetric measurement.

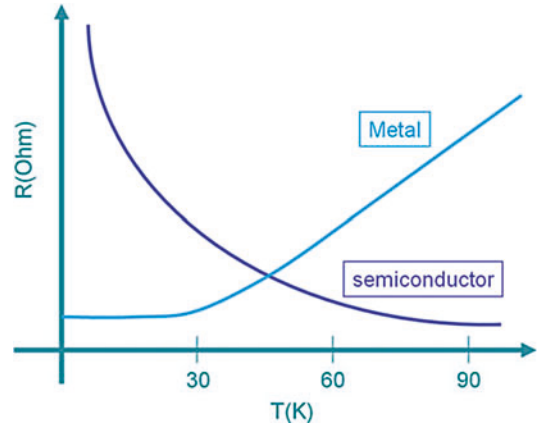
Another major issue in calorimetry is the temperature homogeneity of the sample and the addenda. Indeed, the low thermal diffusivity of some samples limits the dynamic of temperature variations during the measurement. The relaxation time of thermal diffusion in the sample and addenda, τ_{diff}, also depends on the geometry of nanocalorimeters. In general, a thin layer of gold, or highly diffusive material, is deposited by vacuum evaporation technique or sputtering on the measurement area so that the temperature is as uniform as possible in the sample. Finally, the thin thickness of the membranes reduces the thermal coupling to the outside providing then thermal isolation, while the isothermal layer ensures thermal homogeneity of the sensitive area. The experimental timescale Δt_{exp}, the timescale over which the thermal measurement occurs has to be slower than the diffusion time (to ensure a homogeneous temperature) and faster than the thermalization time of the sensing part to the heat bath. This last point can be mathematically summarized by the two following inequalities:

$$\tau_{\text{diff}} << \Delta t_{\text{exp}} << \tau$$

where τ_{diff} is the thermal diffusion time and τ is the thermal relaxation time.

Thermometry

The basis of the measurement of heat is essentially based on the ability of the experimentalist to accurately measure temperature. To obtain a sensitive measurement of temperature, it requires access to a physical quantity X, which varies greatly with temperature. This large variation with temperature is characterized by what is called the temperature coefficient $\frac{1}{X}\frac{dX}{dT}$. A highly sensitive measurement of temperature variation will allow a direct access to very small energy, which is the case generally when nanosystems are involved. The physical quantity X will be anything that can be accurately measured: a volume, a pressure, or an electrical quantity (capacitance, resistance, etc. . .).



Nanocalorimetry, Fig. 4 Variation of resistance versus temperature in the cases of metals and semiconductors

Some of the most used technique to measure temperature in nanocalorimetric experiments are detailed below.

Resistive thermometry: One of the most commonly used physical quantities to measure a temperature is the resistance. The resistivity of pure metals is changing a lot with temperature from very high temperature down to 30 K. Then, these metals will be perfect thermometers in this temperature range. The most common thermometer is platinum; this metal when elaborated in thin films geometry has a positive temperature coefficient of resistance of about $\frac{1}{R}\frac{dR}{dT} = 3 \times 10^{-3}\text{K}^{-1}$; the resistance is decreasing when the temperature is decreased (Fig. 4). Other materials have also a strong change of resistance with temperature like semiconductor or like Anderson or Mott insulator. These materials have a huge increase of resistance as the temperature is lowered because the electron transport is more and more limited as the thermal activation is diminished. These materials (germanium, carbon, niobium silicon, niobium nitride) are widely used as thermometer at low temperature. They are much more efficient than metals below 40 K with temperature coefficient that can be above 1 K^{-1}! Temperatures as low as few millikelvin, close to the absolute zero, may be measured using this type of thermometry. The limitation comes from the applied current

necessary to measure the resistance; the Joule heating creates parasitic power. This dissipated power may induce temperature gradients and then induces error when estimating the real temperature of a nanosystem.

Noise thermometry: Any electrons in a regular resistance are subject to Brownian motion. This Brownian motion takes its origin in the temperature activation. This movement of electrons gives birth to a varying voltage across the resistor which is called the Johnson–Nyquist noise. Due to its origin this noise is a function of the temperature through $V = \sqrt{k_B T R \Delta f}$, where k_B is the Boltzmann constant, T the temperature, R the resistance, and Δf the frequency window where this noise is measured. By measuring the voltage versus time, knowing the resistance, one can deduce the temperature of the resistor R. This thermometry is not very sensitive but it has the major advantage of not dissipating any spurious power because its measurement does not require any electrical current.

Thermocouple thermometry: A thermocouple is composed of a junction between two different materials. A temperature difference between a "hot" and a "cold" junction will be converted into a voltage (Seebeck effect). The measurement of this voltage is a widespread way of measuring a temperature. However, the amplitude in volt per Kelvin produced by a usual thermocouple is small (of the order of magnitude of several hundreds of microvolts per degree), it needs at least a few hundred couples in a system called thermopile to obtain sensitivities equivalent to those of resistive thermometers.

A thermopile is a device composed of plenty of thermocouples connected electrically in series and thermally in parallel. Each couple participates in the total impedance of the sensor and therefore increases the thermal noise of electrons in conductors (Johnson–Nyquist noise). The fact that it is not necessary to polarize and thus generate thermal power in the sample to obtain a measuring signal is a major advantage of this technique.

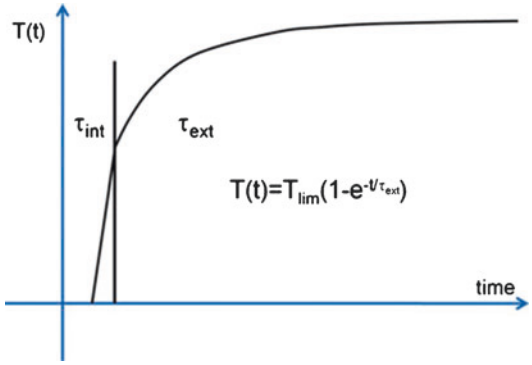
When working close to room temperature, most calorimeters and differential nanocalorimeters use thermopile to measure directly the temperature difference between the sample cell and a reference neutral cell (see DSC). In this case, the "hot" and "cold" junctions are perfectly thermally coupled to the sample and reference, respectively (or vice versa). For zero signal detection devices, two thermopiles connecting sample/thermal bath and reference/thermal bath, respectively, are mounted in opposition in order that same temperature elevations of the sample and references with respect to thermal bath gives approximately zero signal. Knowing the thermal link (thermal exchange coefficient) between the sample and the reference with respect to the surrounding, the voltage from the thermopile is proportional to the difference of heat flows (W) exchanged between the sample and the reference with respect to thermal bath, respectively. One prefers to present the converted signals collected by thermopiles in watts rather than in Kelvin because their direct integration along time gives the heat absorbed or released by the sample during the experiment of calorimetry. In conclusion, the thermopile is the ideal element to detect differential temperature or differential heat flow between two objects without direct generation of power in one or (and) the other of these objects.

N

Experimental Techniques

Principle of Measurement

Calorimetry is the measurement of heat exchanges between a system for which the thermal properties have to be studied and its environment (thermal bath). There are two types of calorimetric measurements. The first is the measurement of the specific heat. In this case, the experimentalist provides a given heat flux to the sample and measures the resulting temperature rise (Fig. 1c). The second concerns the measures of energy released or absorbed by a sample during any transformations or physicochemical interaction at constant temperature; one speaks in this case about isothermal calorimetry. In all cases, the experiment of calorimetry consists in measuring a change in temperature. Thermometry and the measurement electronics are therefore two essential elements for nanocalorimetry. The temperature sensor is then chosen as a function of the particular



Nanocalorimetry, Fig. 5 Illustration of the two significant timescales in a nanocalorimetric experiment

experimental methods used; the latter being in fact adapted to the physical phenomena that the experimentalist wishes to study.

Let us introduce two characteristic times which will be very useful in the rest of this entry. First the internal thermalization time noted τ_{int} is defined; this time is related to the diffusion of heat inside the sample to be studied. Secondly, the external thermalization time noted τ_{ext} is defined. This time is given by the ratio C/K where C is the heat capacity of the system and K the thermal conductance of the link to the bath (see Fig. 1). If a heat power is supplied to the sample at a rate faster than the internal thermal time, then the temperature of the sample is not homogeneous (see Fig. 5). On the other hand, if a power is supplied to a sample over a very long time, then the external thermal time will be dominant. A gradient of temperature will be established following an exponential law between the sample and the heat bath as shown in the Fig. 5.

Experimental Methods

Adiabatic Calorimetry

Regarding the specific heat, various experimental methods allow its measurement. The most traditional is the adiabatic calorimetry in which the sample and its addenda are isolated as much as possible from their environment. A thermal power of known value is sent at a timescale much smaller than the thermal relaxation time τ_{ext}. The measured temperature rise is inversely proportional to the heat capacity of the sample (plus addenda)

$$\Delta T = \frac{\int_0^{\Delta t} P dt}{C}$$

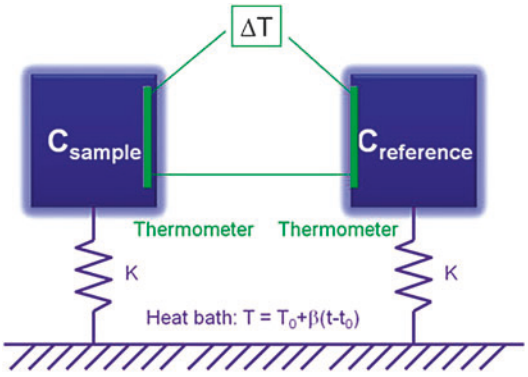
Then, it remains simply to change the temperature of the thermal bath and carry out a new measurement in order to have a final C (T) curve. In practice, adiabaticity is not perfect, and the heat exchange coefficient has to be taken into account, and therefore energy losses between the sample and the thermal bath. This is especially true in nanocalorimetry because objects to be studied are small and therefore their heat capacity is small to. It is therefore very difficult to thermally isolate them, and the adiabatic nanocalorimetry method cannot be implemented. Scientists have developed calorimetric techniques better suited to the measurement of small samples to overcome the problem of heat loss. They are presented in the following sections.

Differential Scanning Calorimetry

A second widespread calorimetric experimental technique is called differential scanning calorimetry (DSC) [8, 9]. It is a continuous method in which the sample temperature follows a temperature ramp imposed by the experimentalist (Fig. 6). The fact that it is differential indicates that what is measured is directly the temperature difference between the sample and a neutral reference. However, one must keep in mind that any calorimetric method can be conceived in differential mode. In the case of DSC, differential heat capacity ΔC and differential heat flux ΔP (between sample and reference) measured during the ramp temperature obey the following equation:

$$\Delta C = \frac{\Delta P + \tau \frac{d(\Delta P)}{dt}}{\beta}$$

where β is the scanning temperature rate and τ the thermal relaxation time. This differential heat flux (or thermal power) is related to the differential



Nanocalorimetry, Fig. 6 Thermal scheme for DSC experiment. One cell contains the sample and the other cell contains a neutral reference. During the scanning temperature rate the temperature difference (or heat flux difference) is recorded as a function of time

temperature by means of the heat exchange coefficient K:

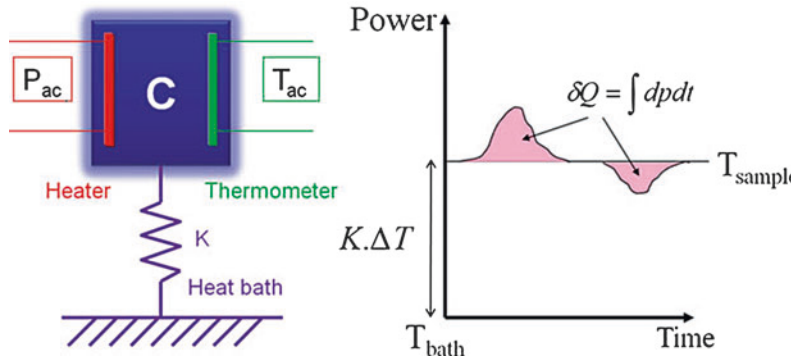
$$\Delta P = K\Delta T$$

Compared with the adiabatic calorimetry, here the heat exchange coefficient K also plays a role at the same level than the heat capacity C. The former is measured by means of a prior calibration in quasi-static mode where the above equation is valid. In DSC, the scanning rate $\beta = dT/dt$, is an important factor since it determines the signal intensity. Indeed, the first equation above shows that for a given differential heat capacity, the differential signal ΔP is for a part proportional to this scanning rate. However, in this case, the power of resolution of the measurement is reduced due to the presence of the second term revealing the dynamic of the DSC. Again, the thermal time constant plays a key role at high scanning rates. It should be noted that the above equations are only valid for a large thermal diffusivity in the sample and the sample holder, so that the time constant of heat diffusion (τ_{int}) does not become the limiting factor. They are also valid under the assumption of perfect thermal symmetry of the calorimetric head (second-order terms are not shown) [8, 9]. Classical DSC based micro- and nanotechnologies are rare. Let us mention SiN membrane DSC for heat capacity detection of nl-range liquid droplets [10] and MEMS-DSC device where protein folding processes are analyzed via the measurement of differential heat capacity [11].

Fast speed DSC: In nanocalorimetry, the samples to be studied are so small that the internal thermalization time is not limiting. In this case, scientists have circumvent the problem of isolation of small samples in using scanning rates with values going from about 10^4 to 10^6 K/s. This new experimental technique, called "Fast Speed DSC," has grown significantly in recent years. They use microsensors, typically SiN thin membranes, produced through micro- and nanotechnology [12–14]). These very high speeds yield to high sensitivities and allow measurements of very thin films from 100 to 1,000 K [5, 15]. A last feature of these rapid calorimetric measurements is that, at such high temperature rates, kinetics of studied thermal event can be observed. This is also true for high-frequency nanocalorimetric methods as seen in the following.

Isothermal Calorimetry

A third widely used measurement method is called Isothermal Calorimetry [16]. It consists of measuring absorption (or removals) of energy in a sample over time at constant temperature (Fig. 7). Titration nanocalorimetry is an experimental method issued from Isothermal Calorimetry in which a compound A in solution (titrant) interacts with compound B in solution (titrate), producing or absorbing energy at constant temperature. If an aliquot (few percent in volume) of the titrant solution containing the molecule A interacts with the solution containing the molecule B, then the experiment can be carried out N times to achieve complete disappearance of B. In this way, through this type of experience, not only the enthalpy of reaction between molecules A and B can be obtained (during the firsts interactions), but also, thanks to the saturation curve $\Delta H = f(N)$, the constant of reaction between A and B is obtainable, like for acid-base chemical titrations. The constant of reaction allows access to the Gibbs free energy at a certain temperature. With calorimetric titration, the thermodynamic parameters ΔH, ΔG, and ΔS *could be accessed* for the



Nanocalorimetry, Fig. 7 Thermal scheme for isothermal calorimetry experiment. The power necessary to maintain the temperature of the sample at a constant value is represented versus time. The integration of this power provides the energy absorbed or dissipated by the sample at such temperature

reaction A/B which provide complete thermodynamic data at determined temperatures. This technique is mainly used in chemistry, biochemistry, and biology [17]. Since in this field, firstly biochemical and biophysical reactions generally have relatively low energy values, and secondly available volumes of biological samples are generally reduced (simply because they are expensive due to the cost of the synthesis) the use of nanocalorimetry is a natural choice. In the past decades, the development of isothermal titration nanocalorimeters has been increasing a lot [18, 19]. Another approach to induce interaction of A with B is to mix the two solutions in the sensitive area via two differentiated liquid inputs. The mixture produces or absorbs energy in the sensitive area which is detected by the sensor and then flows through a third flow output path. One speaks about flow nanocalorimeters [20–24]. An original way of binding has been obtained by means of electrostatic mixing [25].

A major technical difficulty in the development of isothermal titration or flow nanocalorimeters is the need for a coupling of microfluidic techniques with that of the microfabrication of nanosensors. Indeed, one needs to bring the various liquid reactants on the measurement area while maintaining the highest thermal insulation of the same area. Certain isothermal nanocalorimeter measures only the power or heat released or absorbed by a small biological object already positioned (e.g., living cells) along time [26–28]. In the field of biocalorimetry, volumes of analyzed samples comprise between few nanoliter and few microliter, and minimum detectable powers are between few nanoWatt and few hundredth of nanoWatt depending on the nanocalorimeters designs. Integration of power measured versus time yields to minimum recorded energies from few tenth of nanoJoule to few microJoule (see tables showing performances of various isothermal calorimeters in Refs. [29, 21]).

Low Temperature Calorimetry

Traditionally, at very low temperatures, the most used calorimetric method is the adiabatic calorimetry. However, as the adiabatic nanocalorimetry is difficult to carry out at low dimension, scientists have developed new techniques known as dynamic calorimetry techniques, which overcome the problem of heat loss. There are two main different methods: the relaxation calorimetry and ac-calorimetry [30]. Relaxation calorimetry overcomes poor insulation in directly measuring the exponential decay of the temperature of the sample compared to the thermal bath after a pulse of power. Measuring the thermal conductance K is necessary to have the heat capacity. On contrary, ac-calorimetry depends only on the frequency of temperature oscillations to be placed in adiabatic condition or quasi-adiabatic conditions. The sample temperature oscillates at a frequency such that no heat loss holds over one period of oscillation. In this case, it is not necessary to measure K. These different experimental techniques suitable for nanocalorimetry are described below.

Relaxation calorimetry: Relaxation calorimetry is to apply a heat pulse on the sensitive area of a sensor. An increase in the temperature of the

sample appears; this increase will be faster than heat leak because the internal diffusion time is very short. The technique involves recording the decrease of the temperature sensor over time:

$$T = T_{bath} + \frac{P_0}{K}\exp\left(-\frac{t}{\tau}\right)$$

This decay is exponential and depends on the ratio between the heat capacity of the sensor with the sample and the thermal conductance to the bath. The thermal conductance can be estimated by other means, hence from the exponential decay the heat capacity can be deduced. This measurement is usually performed on a membrane where the thermometer and heater are lithographed. It is a sensitive method which may cover a large temperature range [31, 32]. One weak point of this technique is that the measure is not based on an oscillating method, and therefore the SNR is degraded by the presence of thermal drift. This technique has been applied in some recent cases of specific heat measurement on very small sensors at low temperature [4, 33].

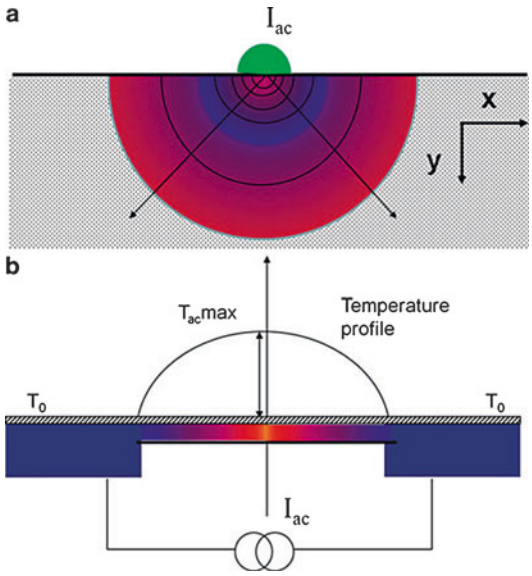
AC-calorimetry: In ac calorimetry, an input thermal power $P(t) = P_{dc} + P_{ac}$ constituted by a dc and a ac term is supplied to a system connected to a thermal bath by means of a thermal conductance of known value [34–37]. In focusing on the ac term only, the corresponding oscillating temperature T_{ac} is measured. At low frequencies the measurement is not adiabatic or quasi-adiabatic, and at high frequencies the temperature of the sample is not homogeneous anymore. This means that the thermal relaxation time and the thermal diffusion time have to be carefully taken into account. However, in choosing the appropriate frequency range, it is possible to circumvent the problem of thermal insulation of small objects to be measured by nanocalorimetry. It remains to design a calorimetric device with an appropriate working frequency range. The frequency of the temperature oscillation has to be faster than the relaxation time to the heat bath τ_{ext} (adiabatic condition) and slower than the diffusion time in the sensor τ_{int}. Under these circumstances, the heat capacity of the sample is simply obtained by the ratio of the ac power on the ac temperature:

$$C = \frac{P_{ac}}{i\omega T_{ac}}$$

The measured heat capacity is actually a complex number with real and imaginary components of different physical meaning. These are obtained by means of the measure of either the amplitude or the phase of the oscillating temperature. This method is particularly sensitive because locking-amplifiers with narrow bandwidth filter can be used for oscillating temperature recording but also because dc or low frequency thermal drifts are less important. Heat capacity resolutions $\Delta C/C$ of the order of 10^{-5} to 10^{-4} are currently obtained while in relaxation calorimetry or other methods resolutions lay between 10^{-3} and 10^{-2}. Thus, ac-calorimetry is a technique particularly adapted to nanocalorimetry because, on one side it allows measures on small objects, and on another side it allows measurement in the nanoJoule/Kelvin or nanoJoule/picoJoule ranges [6, 38]. Although ac calorimetry has been developed in the field of low temperature physics [39]), for few decades it has been adapted to different temperature ranges in several domains of research (magnetic films, biological objects, polymers, etc. . .) where the detection of fine phase transitions or phase transformations were researched [7, 40, 41]. More precisely, the record in sensitivity was obtained in attojoule calorimetry (10^{-18} J) which allowed access to complete new physics [3]. The principal inconvenient of ac-calorimetry is the useable frequency range limited to no more than one or two decades when measurement of the dynamic of the studied system via $C(\omega)$ is wanted. This disadvantage is avoided in another dynamic calorimetric method called 3ω method (Fig. 8).

Other Methods

3ω-method. In the 3ω method, the thermal power that generates temperature oscillations in the sample is provided by the thermometer itself. A current of frequency f passes through the thermometer-like heater resistance (called the transducer), which produces a thermal power oscillating at frequency $2f$ due to Joule effect. As the transducer temperature varies at that



Nanocalorimetry, Fig. 8 Schemes of the 3ω method with the thermal wave in the plane (perpendicular to the transducer) (**a**) or along the sample (parallel to the transducer) (**b**)

frequency, this results in a $3f$ term which contains all the thermal information of the sample [42, 43]:

$$\delta T_{ac} = \frac{P_0}{K\sqrt{1+(\omega\tau)^2}} \text{ or } \delta T_{ac} = \frac{P_0}{\sqrt{2\omega CK}}$$

The two above equations apply depending on the geometrical design of the sensor. ω is the pulsation, related to the frequency through $\omega = 2\pi f$. At very low frequency, the amplitude of the temperature oscillation is only related to the thermal conductance. If the measurement is made as a function of the frequency, the heat capacity can be extracted through the estimation of the thermalization time $\tau = \frac{C}{K}$.

The purely electrical term oscillating at frequency f, which is much larger than the $3f$ component, has generally to be removed using a Wheatstone bridge circuit, otherwise the quality of the measurement will suffer. This method, non-adiabatic by principle, was used to measure the thermal properties of material of very small thermal diffusivity for which a frequency dependence of the specific heat was expected. It has, for example, been successfully used to study the frequency-dependent specific heat at the glass transition of some glass formers over several decades of frequency [42]. Depending on the geometry of the couple transducer with sample, the thermal effusivity $c \times k$ (product of specific heat by thermal conductivity) is measured or, upon particular geometrical conditions, the thermal conductivity k is directly measured at low frequency (see equations above) [44]. In nanocalorimetry, this method has been successfully applied to the measurement of the thermal conductivity of crystalline silicon wires of nanometric sizes for which phonon-blocking effects were observed at very low temperatures [45, 46]. Recent experiments use the 3ω method for thermal properties measurement of nl liquid-like samples [47, 48].

Unconventional methods: measurement of caloric curve. In this technique developed for the in-flight measurement of clusters (hundred of atom), one uses the photofragmentation of the clusters by a laser beam to estimate the internal energy U as a function of temperature. The clusters ions are selected using a mass spectrometer and thermalized in helium gas. Then they are irradiated by a laser, the fragments produced are analyzed by a second mass spectrometer. Their number depends directly on the inner energy of each clusters, hence its measurement gives the internal energy of the cluster. Its derivative $\left(C = \frac{\partial U}{\partial T}\right)$ will give the heat capacity of the clusters ions (see Refs. [49, 50]). Melting point and heat capacity of cluster of 140 atoms or less could be measured using this technique, a performance never equaled with other techniques. Others unconventional methods of nano- and microcalorimetry exist but are not described in this brief entry.

Micro- and Nanosensors

It is the need to do thermodynamic measurements on specific physical or chemical nanosystem that will determine what type of sensor that must be developed for nanocalorimetry. Secondly, the choice of the experimental technique (as detailed

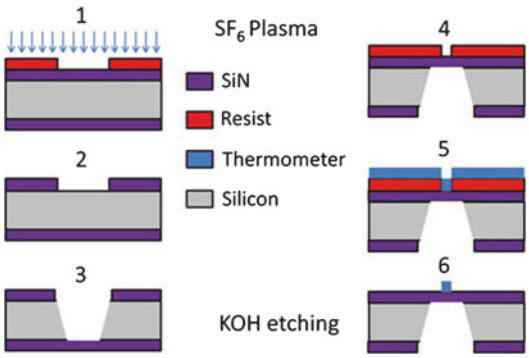
in the previous section) will also influence the conception of the sensing part. Several examples of nanocalorimeter based on micro- or nanofabrication are shown below where the best results have been obtained in terms of sensitivity and/or resolution. Great care must be taken in the choice of materials along with the geometry of the system (thermometry and heater); it will fully determine the thermal performance of the sensor. Two main parameters should orient the choice of materials for the sample holder: small heat capacity and a complete absence of phase change in the working temperature area to avoid spurious signal coming from the holder. The geometry will be chosen to reduce the leak to the heat bath or to control it through a dedicated thermal link. Thermal data of the main materials used for manufacturing sensors are summarized in Table 1. This table will be fruitfully used during the design of a new sensor depending of its future working temperature range or its particular specifications (materials compatibility).

The first attempt to measure the heat capacity of thin films was done by G.D. Zally et al. [51]. He built a calorimeter based on a very thin pyrex membrane where thermometer and heater were attached. Since this attempt, numerous different sensors have been built in various materials from silicon, silicon nitride, glass to polymer. Some of them are detailed below.

Silicon, Silicon Nitride, or Diamond Membranes

As mentioned above, the choice of the materials is given by the specification of the experiment to be performed. The best material for application in a wide temperature range (from 40 K to 1,000 K) is, for sure, silicon nitride (SiN). This amorphous material is totally inert in that temperature window and can be easily manipulated. Fabrication of membrane out of SiN is masterized in all cleanrooms in the world using chemical etching of silicon by KOH (see the sketch in Fig. 9). Thermometer and heater can be lithographed on top, which permits the measurement of thermal properties. The design of the heat link and the choice of materials for the transducers will set the performance of the calorimeter.

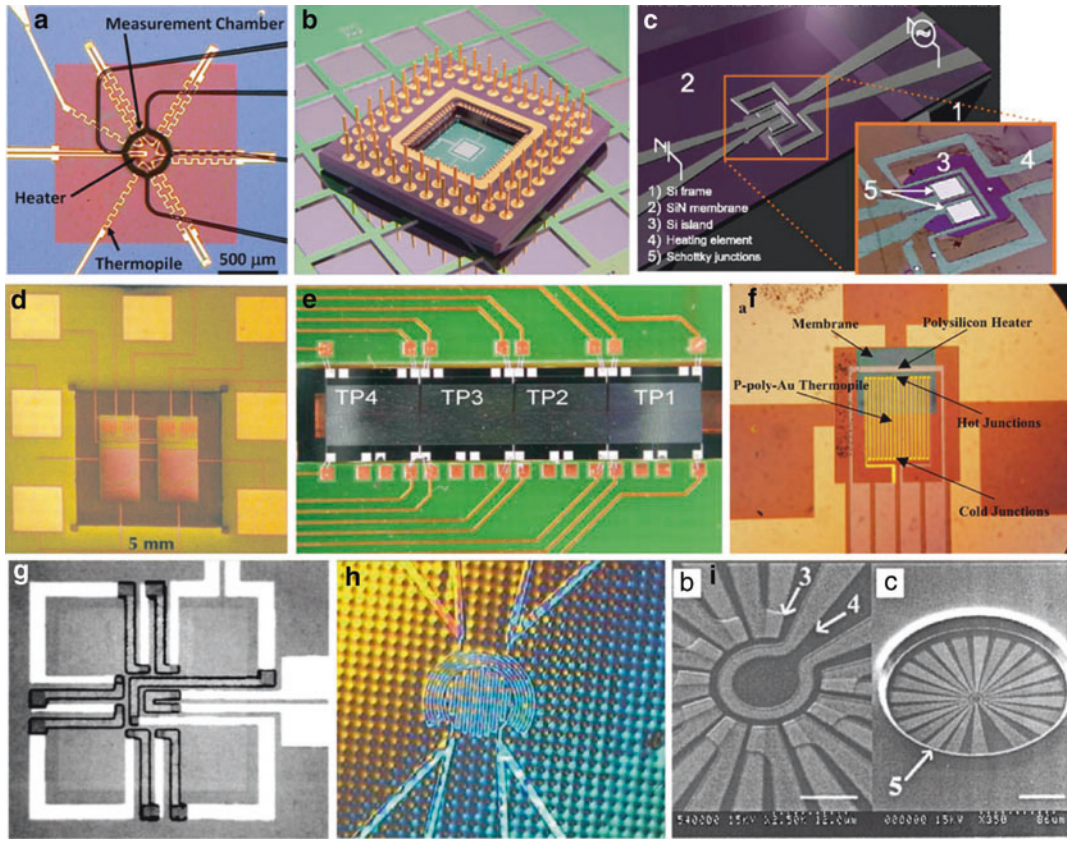
If one wants to apply nanocalorimetry technique at low or very low temperature, SiN is not the best choice. As it can be seen in the Table 1, the heat capacity is high as compared to single crystal silicon. Then the fabrication of silicon membrane is more appropriate and should be preferred. Moreover, the structuring of the membrane



Nanocalorimetry, Fig. 9 Example of microfabrication procedure of a suspended membrane of SiN. *1* Spinning of resist and photolithography, *2* remove the resist, *3* chemical etching of silicon with KOH, *4* photolithography of the transducer, *5* deposition of the transducer, *6* lift-off

Nanocalorimetry, Table 1 Thermal parameters for various materials from high temperature to very low temperature. The thermal conductivity is given in W/cm K and the specific heat in J/g K

	Si		SiN		Pure copper		Polymer (PTFE)		Glass	
	c	k	c	k	c	k	c	k	c	k
1,000 K	8.2	0.3	4	0.2	0.5	4	NA	NA	3	2.10^{-2}
300 K	6.7	1.5	2	0.1	0.3	4	0.9	3.10^{-3}	1	10^{-2}
100 K	2.5	8	1	4.10^{-2}	0.25	5	0.3	2.10^{-3}	0.6	8.10^{-3}
10 K	2.10^{-3}	0.3	10^{-3}	6.10^{-3}	8.10^{-4}	160	0.18	10^{-3}	6.10^{-3}	10^{-3}
1 K	5.10^{-7}	$\sim 10^{-3}$	$\sim 10^{-6}$	5.10^{-4}	$\sim 10^{-5}$	40	$\sim 10^{-3}$	7.10^{-5}	7.10^{-6}	5.10^{-4}
0.1 K	$\sim 10^{-11}$	NA	$\sim 10^{-9}$	NA	$\sim 10^{-6}$	0.4	NA	NA	8.10^{-8}	10^{-5}



Nanocalorimetry, Fig. 10 Various sensors for isothermal or biological liquid-sample nanocalorimetric experiments. (**a**) Parylene membrane for isothermal chemical and biological interactions [21], (**b**) silicon sensor for isothermal chemical and biological interactions (Photograph courtesy www.xensor.nl, [14]), (**c**) silicon nitride sensor for small-volume-liquid calorimetric detection [18], (**d**) polyimide membrane for electrostatic-liquid-mixing calorimetric detection [25], (**e**) silicon microfluidic chamber for flow-through isothermal calorimetry [52], (**f**) glass microfluidic reaction chamber for flow-through calorimetry [24], (**g**) silicon nitride membrane for small-volume-liquid calorimetric detection [26], (**h**) micro-posts sustained polyimide membrane for small-volume-liquid calorimetric detection [7], (**i**) silicon nitride suspended sensor for isothermal detection of living cells [27]

will be absolutely necessary to create a well-defined isotherm. Indeed, the thermal conductance of silicon (or SiN) is still high below 10 K, and then a proper thermal isolation lies in suspending a membrane through isolated arms (see Fig. 2a or b). In Fig. 10, examples (c), (g), and (i) are silicon membrane for isothermal flow-through or small-volume-liquid nanocalorimetric measurements.

Polymer Membranes

In nanocalorimetry, thin polymer membranes are also used to insure good thermal insulation of the measurement area. This type of membrane is mostly used for medium to high temperatures because of the low thermal conductivity of polymers (see Table 1). At very low temperatures, they are less attractive because their heat capacity is very high (see Table 1). To ensure uniformity of temperature in these sensors, the low thermal diffusivity of these materials is generally eliminated through the deposition of a thin layer of metal diffusing heat over the entire surface of the sensitive area (e.g., gold, aluminum, silver). Figures 2d, f and 10a, d, h are examples of polymer membranes for experiments of ac-nanocalorimetry or isothermal nanocalorimetry on thin magnetic films, thin polymer films, and low volumes

chemical or biological objects or reactions. The thickness of these membranes can range from a few hundreds of nanometers to several tens of micrometers. The use of micro- and nanotechnologies is obviously necessary to deposit and lithographically machined sensing elements on the membrane, as well as a possible step of membrane design through dry etching techniques (e.g., reactive ion etching or oxygen plasma).

Conclusion

In that not complete review, various experimental methods named "nanocalorimetry" have been presented. It has been shown that it concerns all calorimetric method for which the dimensions of the objects studied are in the order of magnitude of nanometer, or measured energy or power are in the range of nanojoule or nanowatt. The review has been based on numerous examples of nanocalorimeters existing in the literature from low temperature to room temperature with application in condensed matter, chemistry, biophysics, or biology in general. The aspect of micro- and nanofabrication, essential for the achievement of nanocalorimetric devices, has been emphasized as well as the importance of sensitive instrumentation, electronic conditioning, and the choice of materials for the thermal detectors.

Calorimetry is by essence a universal method of measurement, covering by this way a wide spectrum of different researches. The emergence of nanocalorimetry in recent decades has allowed the exploration of new and original fields of research. With the evolution of modern micro- and nanofabrication technologies, many new developments will appear especially toward more sensitive sensor especially in biophysics. Significant room for improvement still exists in many areas of nanoscience, which makes nanocalorimetry quite an open subject of research.

Cross-References

- ▶ Biosensors
- ▶ Nanotechnology
- ▶ Thermal Actuators
- ▶ Thermal Conductivity and Phonon Transport

References

1. Lavoisier, A., de Laplace, P.S.: Mémoire sur la chaleur, p. 355. Académie des Sciences, Paris (1780)
2. Giazotto, F., Heikkilä, T.T., Luukanen, A., Savin, A.M., Pekola, J.P.: Opportunities for mesoscopics in thermometry and refrigeration: physics and applications. Rev. Mod. Phys. **78**, 217 (2006)
3. Bourgeois, O., Skipetrov, S., Ong, F., Chaussy, J.: Attojoule calorimetry of mesoscopic superconducting loops. Phys. Rev. Lett. **94**, 057007 (2005)
4. Chung Fon, W., Schwab, K.C., Worlock, J.M., Roukes, M.L.: Nanoscale, phonon-coupled calorimetry with sub-attojoule/Kelvin resolution. Nano Lett. **5**, 1968 (2005)
5. Lopeandia, A.F., Valenzuela, J., Rodriguez-Viejo, J.: Power compensated thin film calorimetry at fast heating rates. Sensor. Actuat. A **143**, 256–264 (2008)
6. Lopeandia, A.F., Andre, E., Garden, J.-L., Givord, D., Bourgeois, O.: Highly sensitive parylene membrane-based ac-calorimeter for small mass magnetic samples. Rev. Sci. Instrum. **81**, 053901 (2010)
7. Garden, J.-L., Guillou, H., Lopeandia, A.F., Richard, J., Heron, J.-S., Souche, G.M., Ong, F.R., Vianay, B., Bourgeois, O.: Thermodynamics of small systems by nanocalorimetry: from physical to biological nano-objects. Thermochim. Acta **492**, 16–28 (2009)
8. Claudy, P.: Analyse Calorimétrique Différentielle. Lavoisier, Paris (2005)
9. Höhne, G.W.H., Hemminger, W.F., Flammersheim, H.-J.: Differential Scanning Calorimetry, 2nd edn. Springer, Berlin/Heidelberg (2010)
10. Youssef, S., Podlecki, J., Al Asmar, R., Sorli, B., Cyril, O., Foucaran, A.: MEMS scanning calorimeter with serpentine-shaped platinum resistors for characterizations of microsamples. J. Microelectromech. Syst. **18**, 414–423 (2009)
11. Wang, L., Wang, B., Lin, Q.: Demonstration of MEMS-based differential scanning calorimetry for determining thermodynamic properties of biomolecules. Sensor. Actuat. B **134**, 953–958 (2008)
12. Anahory, Y., Guihard, M., Smeets, D., Karmouch, R., Schiettekatte, F., Vasseur, P., Desjardins, P., Hu, L.A., Allen, L.H., Leon-Gutierrez, E., Rodriguez-Viejo, J.: Fabrication, characterization and modeling of single-crystal thin film calorimeter sensors. Thermochim. Acta **510**, 126–136 (2010)
13. Minakov, A.A., Adamovsky, S.A., Schick, C.: Non-adiabatic thin-film (chip) nanocalorimetry. Thermochim. Acta **432**, 177–185 (2005)
14. van Herwaarden, A.W.: Overview of calorimeter chips for various applications. Thermochim. Acta **432**, 192–201 (2005)
15. Efremov, M.Y., Olson, E.A., Zhang, M., Schiettekatte, F., Zhang, Z.S., Allen, L.H.: Ultrasensitive, fast,

thin-film differential scanning calorimeter. Rev. Sci. Instrum. **75**, 179–191 (2004)
16. Ladbury, J.E., Chowdhry, B.Z. (eds.): Biocalorimetry, Applications of Calorimetry in the Biological Sciences. Wiley, Chichester (1998)
17. Russell, D.J., Hansen, L.D.: Calorimeters for biotechnology. Thermochim. Acta **445**, 151–159 (2006)
18. Hakala, T.K., Toppari, J.J., Torma, P.: A hybrid method for calorimetry with subnanoliter samples using Schottky junctions. J. Appl. Phys. **101**, 034512 (2007)
19. Xu, J., Reiserer, R., Tellinghuisen, J., Wikswo, J.P., Baudenbacher, F.J.: A microfabricated nanocalorimeter: design, characterization, and chemical calibration. Anal. Chem. **80**, 2728–2733 (2008)
20. Köhler, J.M., Zieren, M.: Micro flow calorimeter for thermoelectrical detection of heat of reaction in small volumes. Fresenius J. Anal. Chem. **358**, 683–686 (1997)
21. Lee, W., Fon, W., Axelrod, B.W., Roukes, M.L.: High-sensitivity microfluidic calorimeters for biological and chemical applications. Proc. Nat. Acad. Sci. **106**, 15225–15230 (2009)
22. Lerchner, J., Wolf, A., Schneider, H.-J., Mertens, F., Kessler, E., Baier, V., Funfak, A., Nietzsch, M., Krugel, M.: Nano-calorimetry of small-sized biological samples. Thermochim. Acta **477**, 48–53 (2008)
23. Nam, S.K., Kim, J.K., Cho, S.C., Lee, S.K.: Design and characterization of a high resolution microfluidic heat flux sensor with thermal modulation. Sensors **10**, 6594–6611 (2010)
24. Zhang, Y.Y., Tadigadapa, S.: Calorimetric biosensors with integrated microfluidic channels. Biosens. Bioelectron. **19**, 1733–1743 (2004)
25. Torres, F.E., De Bruyker, D., Bell, A.G., Wolkin, M. V., Peeters, E., Williamson, J.R., Anderson, G.B., Schmitz, G.P., Recht, M.I., Schweizer, S., Scott, L. G., Ho, J.H., Elrod, S.A., Schultz, P.G., Lerner, R.A., Bruce, R.H.: Enthalpy arrays. Proc. Nat. Acad. Sci. **101**, 9517–9522 (2004)
26. Chancellor, E.B., Wikswo, J.P., Baudenbacher, F., Radparvar, M., Osterman, D.: Heat conduction calorimeter for massively parallel high throughput measurements with picoliter sample volumes. Appl. Phys. Lett. **85**, 2408–2410 (2004)
27. Johannessen, E.A., Weaver, J.M.R., Bourova, L., Svoboda, P., Cobbold, P.H., Cooper, J.M.: Micromachined nanocalorimetric sensor for ultra-low-volume cell-based assays. Anal. Chem. **74**, 2190–2197 (2002)
28. Verhaegen, K., Baert, K., Simaels, J., Van Driessche, W.: A high-throughput silicon microphysiometer. Sensor. Actuat. **82**, 186–190 (2000)
29. Braissant, O., Wirz, D., Göpfert, B., Daniels, A.U.: Biomedical use of isothermal microcalorimeters. Sensors **10**, 9369–9383 (2010)
30. Bourgeois, O.: Thermal Nanosystems and Nanomaterials. Topics in Applied Physics, pp. 537–567. Springer, Berlin (2009)
31. Cooke, D.W., Michel, K.J., Hellman, F.: Thermodynamic measurements of submilligram bulk samples using a membrane-based "calorimeter on a chip". Rev. Sci. Instrum. **79**, 053902 (2008)
32. Queen, D.R., Hellman, F.: Thin film nanocalorimeter for heat capacity measurements of 30 nm films. Rev. Sci. Instrum. **80**, 063901 (2009)
33. Suh, K.S., Kim, J.W., Kim, H.J., Park, Y.D., Kim, K. H.: Enhanced accuracy in a silicon-nitride-membrane-based microcalorimeter with variation of lateral layout. Thermochim. Acta **490**, 1–7 (2009)
34. Corbino, O.M.: Thermal oscillations in lamps of thin fibers with alternating current flowing through them and the resulting effect on the rectifier as a result of the presence of even-numbered harmonics. Phys. Z. **11**, 413–417 (1910)
35. Kraftmakher, Y.: Modulation calorimetry and related techniques. Phys. Rep. **356**, 1–117 (2002)
36. Sullivan, P.F., Seidel, G.: Steady-state ac-temperature calorimetry. Phys. Rev. **173**, 679–685 (1968)
37. Tagliati, S., Rydh, A.: Absolute accuracy in membrane-based ac nanocalorimetry. Thermochim. Acta **522**, 66–71 (2011)
38. Fominaya, F., Fournier, T., Gandit, P., Chaussy, J.: Nanocalorimeter for high resolution measurements of low temperature heat capacities of thin films and single crystals. Rev. Sci. Instrum. **68**, 4191–4195 (1997)
39. Minakov, A.A., Roy, S.B., Bugoslavsky, Y.V., Cohen, L.F.: Thin-film alternating current nanocalorimeter for low temperatures and high magnetic fields. Rev. Sci. Instrum. **76**, 043906 (2005)
40. Garden, J.-L., Château, E., Chaussy, J.: Highly sensitive ac nanocalorimeter for microliter-scale liquids or biological samples. Appl. Phys. Lett. **84**, 3597–3599 (2004)
41. Yao, H., Ema, K., Fukada, H., Takahashi, K., Hatta, I.: ac nanocalorimeter for measuring heat capacity of biological macromolecules in solution. Rev. Sci. Instrum. **74**, 4164–4168 (2003)
42. Birge, N.O., Dixon, P.K., Menon, N.: Specific heat spectroscopy: origins, status and applications of the 3 omega method. Thermochim. Acta **304**(305), 51–66 (1997)
43. Corbino, O.M.: Periodic resistance changes of fine metal threads, which are brought together by alternating streams as well as deduction of their thermo characteristics at high temperatures. Phys. Z. **12**, 292–295 (1911)
44. Cahill, D.G.: Thermal-conductivity measurement from 30-K to 750-K - the 3-omega method. Rev. Sci. Instrum. **61**, 802–808 (1990)
45. Bourgeois, O., Fournier, T., Chaussy, J.: Measurement of the thermal conductance of silicon nanowires at low temperature. J. Appl. Phys. **101**, 016104 (2007)
46. Heron, J.-S., Bera, C., Fournier, T., Mingo, N., Bourgeois, O.: Blocking phonons via nanoscale geometrical design. Phys. Rev. B **82**, 155458 (2010)

47. Choi, S.R., Kim, D.: Real-time thermal characterization of 12 nl fluid samples in a microchannel. Rev. Sci. Instrum. **79**, 064901 (2008)
48. Park, B.K., Park, J., Kim, D.: Three-omega method to measure thermal properties of subnanoliter liquid samples. Rev. Sci. Instrum. **81**, 066104 (2010)
49. Breaux, G.A., Benirschke, R.C., Sugai, T., Kinnear, B.S., Jarrold, M.F.: Hot and solid gallium clusters: too small to melt. Phys. Rev. Lett. **91**, 215508 (2003)
50. Schmidt, M., Kusche, R., Kronmuller, W., von Issendorf, B., Haberland, H.: Experimental determination of the melting point and heat capacity for a free cluster of 139 sodium atoms. Phys. Rev. Lett. **79**, 99 (1997)
51. Zally, G.D., Mochel, J.M.: Fluctuation heat capacity in superconducting thin films of amorphous BiSb. Phys. Rev. Lett. **27**, 1710 (1971)
52. Lerchner, J., Wolf, A., Wolf, G., Baier, V., Kessler, E., Nietzsch, M., Krügel, M.: A new micro-fluid chip calorimeter for biochemical applications. Thermochim. Acta **445**, 144–150 (2006)

Nanocarriers

► Acoustic Nanoparticle Synthesis for Applications in Nanomedicine

Nanoceramics by Spark Plasma Sintering

Yan Xiong[1,2] and Zhijian Shen[2]
[1]School of Materials Science and Engineering, Hubei Province Key Laboratory of Green Materials for Light Industry, School of Materials, Hubei University of Technology, Wuhan, China
[2]Department of Materials and Environmental Chemistry, Arrhenius Laboratory, Stockholm University, Stockholm, Sweden

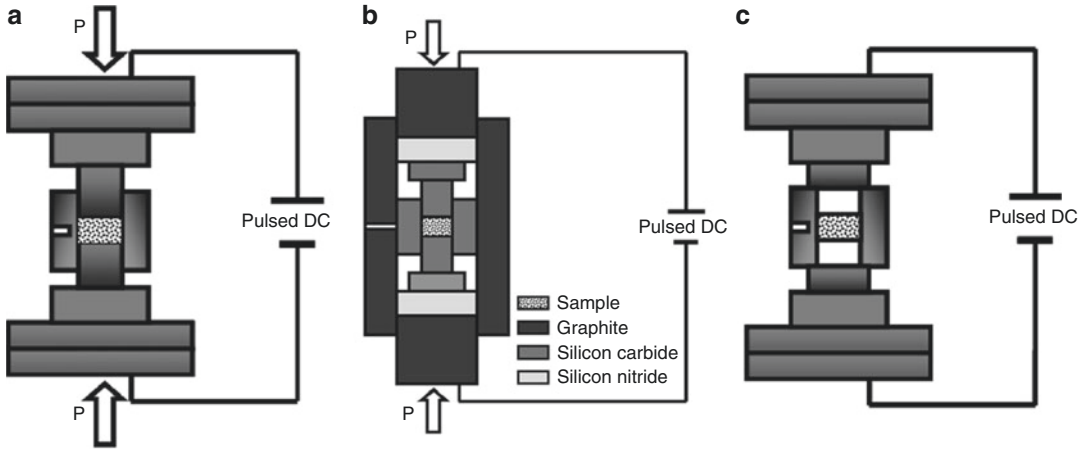
Definition

Nanoceramics	Defined as ceramics with the average grain size at least in one dimension less than 100 nm.
Spark Plasma Sintering	Defined as a sintering process during which the powders or green bodies are loaded in a die and are pressed under a uniaxial pressure. Instead of conventional use of an external heating source, a pulsed DC current is allowed to pass through the electrically conducting pressure die and, in appropriate cases, also through the sample.
Kinetic window	Defined as a temperature interval in which the grain growth and densification process can be separated during sintering.
Grain sliding	Defined as sliding of grains over one another through grain boundaries.

Spark Plasma Sintering

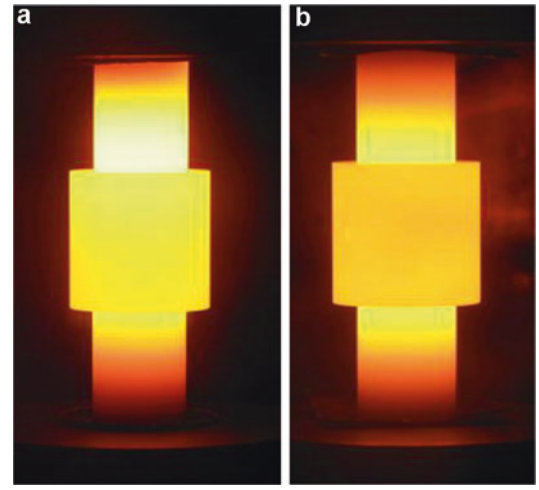
Spark plasma sintering (SPS) is a sintering process similar to conventional hot pressing, in which the samples are loaded in a die and a uniaxial pressure is applied during the sintering. However, instead of using an external heating source, a pulsed direct current is allowed to pass through the electrically conducting pressure die and, in appropriate cases, also through the sample. This implies that the die also acts as a heating source and that the sample is heated from both outside and inside, as outlined in the schematic drawing shown in Fig. 1a. Thus, the unique features of the process are the possibilities of using very fast heating rates and very short holding times to obtain fully dense samples. Three factors that contribute to the rapid densification process can be discerned: (i) the application of a mechanical pressure, (ii) the use of rapid and efficient heating, and (iii) the use of pulsed direct current, implying that the samples are also exposed to an electrical field [1].

During spark plasma sintering the temperature is measured and regulated either by a pyrometer focused on the outer wall of the graphite die or by a thermocouple inserted into the graphite dies. This implies that the real temperature the sample



Nanoceramics by Spark Plasma Sintering, Fig. 1 Schematic figures of the die setup in (**a**) normal spark plasma sintering (*SPS*), (**b**) high-pressure spark plasma sintering (*HP-SPS*), and (**c**) pressureless spark plasma sintering (*PL-SPS*)

experienced during the sintering process is different from what is indicated by the temperature measurement. Such temperature difference is determined by the geometry and size of the die, the thermal and electrical conductivity of the sample, the assembly of the die and punches, the level of applied pressure, and the applied heating rate. Moreover, the contact resistance shows important influence on the temperature distribution during spark plasma sintering. As shown in Fig. 2, the same die set can result in dramatically different temperature distributions when the contacting surfaces of the punches are smooth (a) or rough (b). Since the temperature in most cases is collected from only one spot on the outer wall of the die, it is extremely important to keep the same die configuration for a series of experiments in order to make the results repeatable and comparable [2].



Nanoceramics by Spark Plasma Sintering, Fig. 2 Images of the temperature distributions resulted from (**a**) smooth and (**b**) rough contacting surfaces

The SPS setup can be modified and combined with established sintering technologies to generate further variations. For instance, by using the conception of matryoshka doll, a high uniaxial pressure of up to 1 GPa can be applied with a modified setup as schemed in Fig. 1b [3]. This variation of high-pressure SPS allows further reduction of the sintering temperature. On the other hand, the uniaxial pressure may be totally removed by placing the two oppositely moving punches on the die, with the setup as scheduled in Fig. 1c [4]. In such a case of pressureless SPS, the graphite die acts as a graphite crucible, in which the specimen is not subjected to any external pressure, but only to fast heating. The SPS process may also be used only for gaining close packing of particles at a temperature lower than what was demanded for full densification. Such a process step known as warm pressing in combination with the conventional pressureless sintering and/or hot isostatic pressing for completing the final-stage sintering is beneficial for retaining the nanostructures in densified bulks [5].

The idea to compact metallic materials by an electro-discharge process was originally proposed in the 1960s. Based on this concept, three sintering processes have been developed and commercialized, named as "spark plasma sintering" (SPS), "plasma activated sintering" (PAS), and "electroconsolidation." The characteristics of the three processes are as follows: (i) In the SPS process, a pulsed DC (3.3 ms pulses of 0.5–10 kA intensity) is applied from the beginning to the end of the sintering cycle; (ii) in the PAS process a pulsed DC is normally applied at room temperature for a short period of time and then a continuous DC during the remainder of the sintering process. This procedure is often referred to in the literature as a "single pulse cycle process," with a typical 30–60 ms pulse current of ~1000 A intensity during 60–90 s. In a few cases, the pulsed DC procedure is repeated during the sintering process, and the procedure is then referred to as a "multiple pulse cycle process"; (iii) in the electroconsolidation process, various types of currents (DC, pulsed DC, or AC) can be used, i.e., the system is designed to be flexible in terms of power supply. These processes have now been developed beyond the production of small objects with simple shapes, and a continuous production of compacts of complex geometry and of pieces with diameters larger than 150 mm has been achieved. Despite the fact that a uniaxial pressure is applied, green bodies of complex geometry can be exposed to a "pseudo-isostatic" pressure when being embedded in free-flowing electrically conducting particulates that act as pressure-transmitting medium inside the die. The SPS technology can be combined with various forms of existing production and automatization concepts and technologies. Thus, besides batch-type setups commonly used in laboratory scale, robotic continuous production by SPS is also possible [6–10].

Demand of Separating Grain Growth from Densification for Sintering of Nanoceramics

Nanoceramics, by definition, are ceramics with the average grain size (GS) at least in one dimension being less than 100 nm. A major obstacle for retaining nanostructure in bulks encountered during the bottom-up approaches of sintering of nanoparticles is to separate the grain growth from the densification. It is very challenging to overcome this obstacle, as both the densification and grain growth use the same driving force of reducing the total surface energy of the nanoparticles and the same mass transportation mechanism of interfacial atomic diffusion. Applying high temperatures required for fully densifying ceramic nanoparticles often results in grain growth due to Ostwald ripening when traditional sintering techniques are used. The presence of agglomerates in powders with very fine particle sizes that often leads to an inhomogeneous density distribution in green compacts would contribute further to local grain coarsening during final-stage sintering and would produce nonuniformity in the resulting microstructure.

To minimize grain growth, two strategies are often implemented with a common aim of defining a kinetic window, i.e., a temperature interval in which the grain growth and densification process can be separated: (i) optimizing the packing of nanoparticles and (ii) performing the densification at a temperature as low as possible but just high enough to allow sufficient interfacial atomic diffusion for both the particle sliding and the interfacial atomic diffusion promoting densification. The packing of nanoparticles can be optimized already in the green body stage by involving, for instance, a wet chemical process to increase both the homogeneity and the density of the green body, and/or by applying a pressure to assist the particle rearrangement via particle sliding. Quantitatively, it is well known that the closest sphere packing density is 74.05 %, but, in practice, depending on the particle size and size distribution, presence of agglomerate and applied pressure, etc., dry pressing of nano-sized powders often approaches green body densities around 40–50 % of the theoretical density (TD). Densities up to 60–70 % TD have been achieved by processes such as slip casting, gel casting, and pressure filtration. It was found that the coarsening of particles via surface diffusion is inevitable when pressureless sintering is applied to achieve a

critical density of ~77 % TD from the initial green body with density of 40–50 % TD. To the extent that the starting powder is agglomerated, it is often difficult to de-agglomerate it only by applying high pressure at room temperature. However, when the green body is heated to a temperature that is below the temperature where the grain growth is initiated and is subjected to uniaxial pressures, there it is possible to achieve a packing density in the vicinity of 80 % without any major change of the grain size [11, 12].

Typical Sintering Behaviors Observed During Spark Plasma Sintering

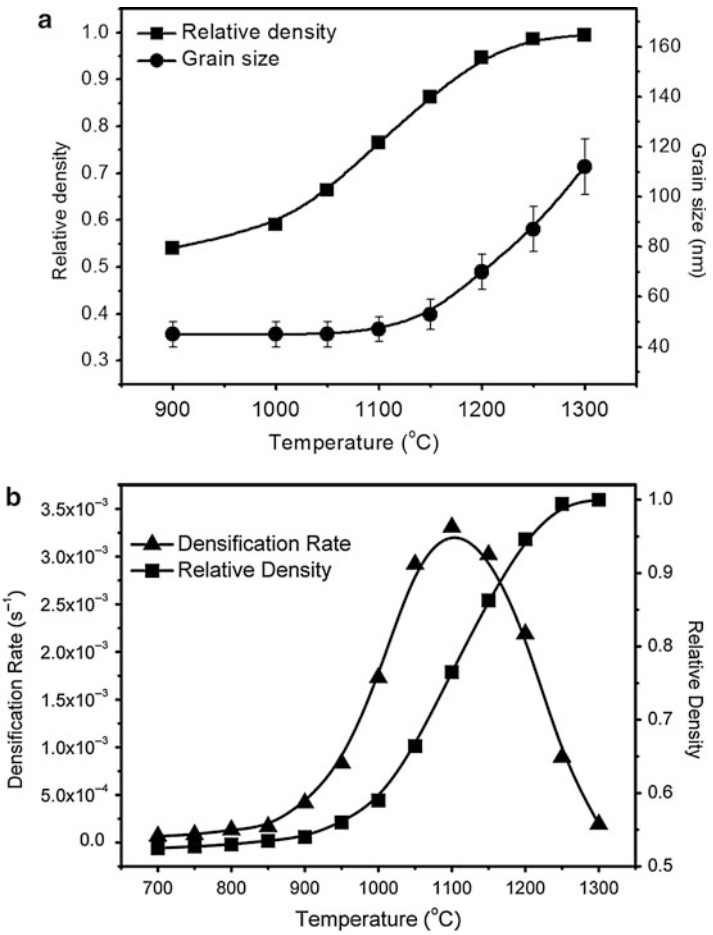
Dramatic Change in Densification Rate

A general observation is the unusual high densification rate achieved in the early and intermediate stage of SPS. The sintering behavior of 3Y-TZP is illustrated as an example presented in Fig. 3. As shown in Fig. 3a, the density rapidly increases from 50 % to ~90 % TD accompanied with rather limited grain growth. The maximum densification rate achievable during this stage of sintering is often in the order of 10^{-3}/s for solid-state sintering and up to 10^{-2}/s in liquid-phase sintering conditions [13]. However, the densification rate dramatically decreases by one order of magnitude during the final stage of sintering. It was reported that the calculated effective activation energies for early stage densification in SPS, based on four classic pressure-assisted models, were considerably lower than those expected by any classical diffusion-related process [14]. This indicates that possible mechanisms beyond the classic interfacial atomic diffusion-related processes may be involved in SPS process. The radical change of the densification rates might reflect the switching of different densification mechanisms that dominate different sintering stages.

Refinement of Pore Structure

When ceramics are viewed as composites of solid and pore phases, densification can be regarded as a process during which pores are gradually replaced by solid grains. It was reported that the enlargement of pores in the intermediate stage of sintering was suppressed during spark plasma sintering which, otherwise, is a phenomenon commonly observed in conventional pressureless sintering. This refinement of pore structure is beneficial for facilitating subsequent densification in the final-stage sintering [15]. Figures 4 and 5 illustrate the evolution of the grain and pore structures against the relative densities of the sintered bodies in the high-density regimes for 3Y-TZP nanoceramics. The samples densified by SPS at 1100 °C for 3 min under 100 MPa achieved a density of 87 % TD which was further densified either by a follow-up SPS at 1100 °C under 100 MPa up to 60 min or by conventional pressureless sintering up to 1350 °C. It can be noticed that the maximum pore size experiences a dynamic enlargement with one order of magnitude increase in size to more than ~480 nm when the sample is pressureless sintered to 1250 °C. Thereafter, the maximum pore size slowly decreases upon further sintering, whereas rapid grain growth was observed resulting in an average grain size larger than 200 nm in the fully densified body. On the contrary, the pore size in samples further SPS processed is suppressed to less than ~150 nm. The easy elimination of small pores during prolonged soaking yielded a final grain size less than 100 nm in spark plasma sintered dense body.

Figure 5a shows an SEM micrograph taken on the partially densified 3Y-TZP samples of 87 % TD, in which both the average grain size (GS) and the maximum pore size measured by linear interception are ~50 nm. Figure 5b shows an SEM micrograph taken on a sample with 94.3 % TD obtained by further pressureless sintering at 1250 °C, revealing the destruction of the microstructural homogeneity due to the local formation of large pores. The removal of those large pores needs much higher temperature which will result in rapid grain growth during the final-stage sintering. On the contrary, the microstructural homogeneity and the fine grain structure can be preserved in a condition as shown in Fig. 5c by SPS at 1100 °C for 10 min under 100 MPa, as the SEM micrograph shown in Fig. 5c revealed. The remaining small pores homogenously distributed at triple junctions in this sample can be eliminated



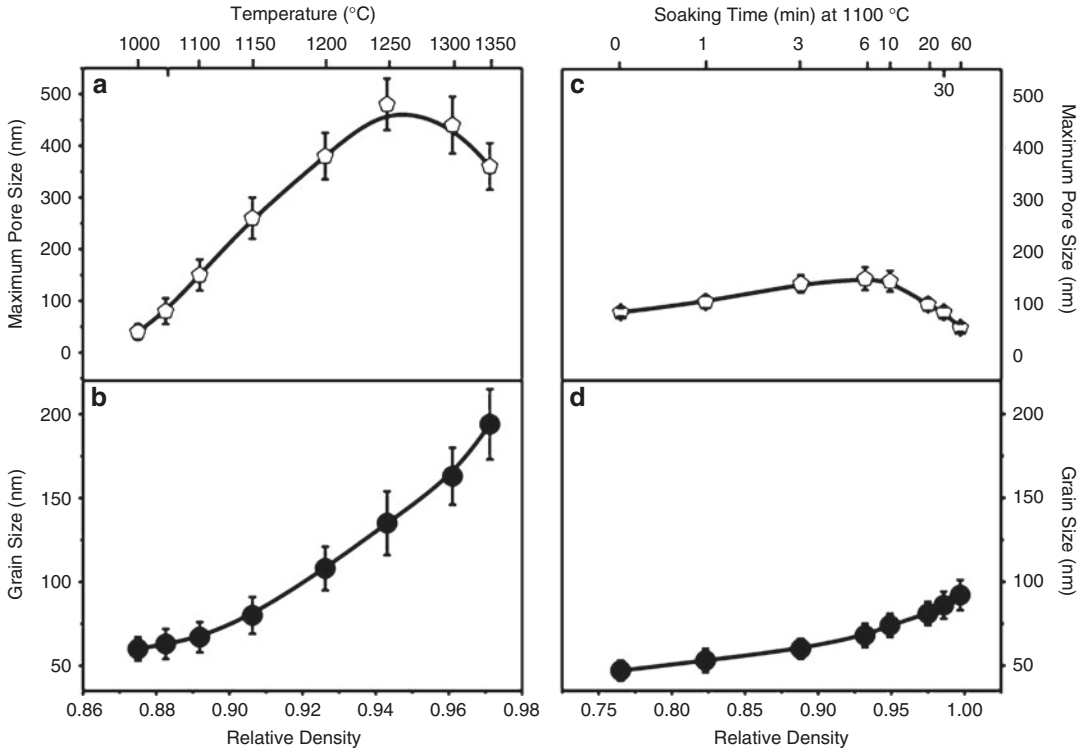
Nanoceramics by Spark Plasma Sintering, Fig. 3 (**a**) The evolution of density and grain size and (**b**) densification rate at elevating temperature of 3Y-TZP nanoceramics by spark plasma sintering under 100 MPa pressure at 100 °C/min heating rate

easily via interfacial atomic diffusion without any extensive grain growth [16].

Grain Sliding as a Dominating Densification Mechanism During the Early and Intermediate Stage of Spark Plasma Sintering

Grain sliding is a phenomenon defined as sliding of grains over one another through grain boundaries which is often observed during pressure-assisted sintering processes. It usually facilitates rapid mass transport and it is responsible for superplastic deformation of nanoceramics. Grain sliding and atomic diffusion-based mass transports are not independent mechanisms. They occur sequentially, implying that the slower mechanism controls the time required for achieving the full densification. Although, it has been reported that during the pressureless sintering of nanocrystalline oxide powders, a sintering mechanism based on particle repacking that occurs concurrently with particle coarsening may be active until ~84 % TD [17]. The active region of particle rearrangement enabled by grain sliding can possibly be further extended to ~93 % TD in SPS conditions [18, 19]. The rapid and efficient heating and the higher pressure that is inherently applicable during SPS appear to be important parameters that bear strong positive influences on the densification of ceramic powders, particularly nanopowders, through grain sliding.

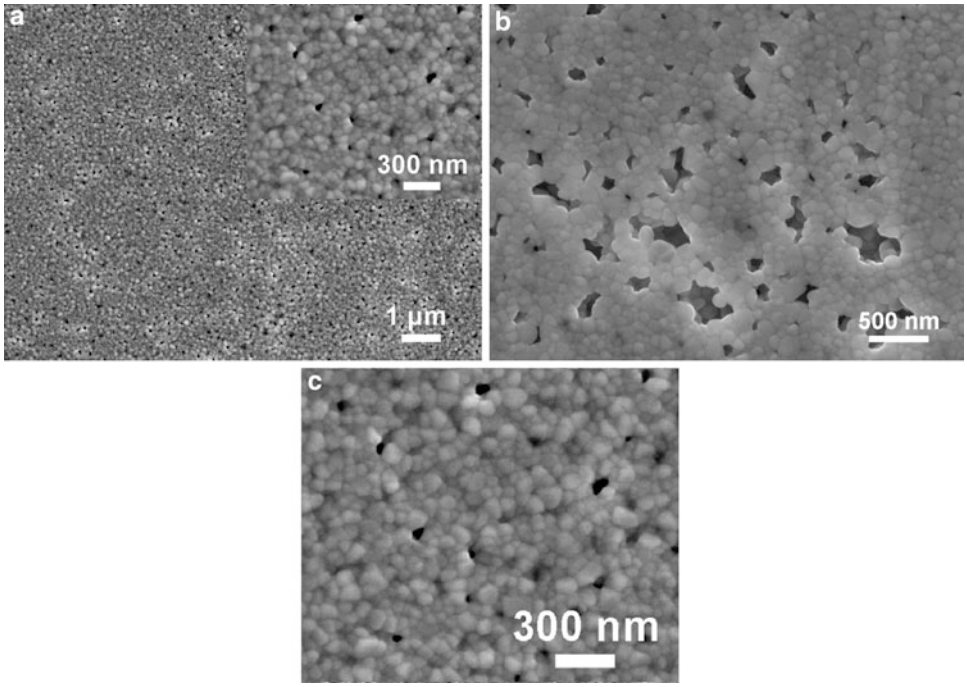
In case of spark plasma sintering of a nano-sized $BaTiO_3$ powder, with the application of a low pressure of 25 MPa, a pressure that normally



Nanoceramics by Spark Plasma Sintering, Fig. 4 Comparisons of the grain size and maximum pore size evolution in two post treatment processes of (**a, b**) by pressureless sintering up to 1350 °C and (**c, d**) by spark plasma sintering soaked at 1100 °C under 100 MPa constant pressure

is used during conventional hot pressing, the maximum shrinkage rate is observed when the sample achieves ~70 % TD. By applying a higher pressure, the shrinkage rate increases and reaches the maximum at higher relative density. Thus, under a pressure of 100 MPa, the maximum shrinkage is achieved at ~90 % TD, implying after that only 10 % of porosity needs to be removed by involving the interfacial atomic diffusion mechanism. The sintering period before the sample reaches the maximum shrinkage can be empirically defined as the initial and intermediate stage during which the densification is enhanced by the smash of agglomerates, the rearrangement of particles, and the close packing of particles, while mechanisms based on power law creep and volume diffusion would probably be excluded. Under high pressures very high strain rates are generated above a critical temperature where an atomic relaxation process in interfacial regions is activated that favor the densification through grain sliding [1, 12].

This shift of the maximum shrinkage rate toward higher relative density minimizes the necessary involvement of the interfacial atomic diffusion mechanisms for densification, thus increases the possibility of reducing the grain coarsening commonly associated with the densification. Accordingly, a large number of nanoceramics and ceramic nanocomposites have been prepared during the last decade [19]. The success in retaining nano-grained microstructure and in reducing the residual pore size down to less than 10 nm has made many dense nanoceramics appearing optically transparent or translucent [20]. It has also drawn attention to spark plasma sintering as a versatile tool for preparing nanoceramics.



Nanoceramics by Spark Plasma Sintering, Fig. 5 (**a**) SEM images of initial 3Y-TZP nanoceramics with 87 % of theoretical density obtained by spark plasma sintering for 3 min at 1100 °C under 100 MPa constant pressure. SEM images of pore structure morphologies in the sample obtained (**b**) by pressureless post treatment at 1250 °C and by (**c**) 10 min soaking at 1100 °C under 100 MPa pressure in SPS conditions

Cross-References

- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Nanocrystalline Functional Materials in Bulk Form with Grain Size below 50 nm
- ▶ Nanocrystalline Materials
- ▶ Nano-Effects
- ▶ Nanomaterials
- ▶ Nano-Sized Particle

References

1. Nygren, M., Shen, Z.J.: Hot pressing and spark plasma synthesis. In: Riedel, R., Chen, I.W. (eds.) Ceramics Science and Technology, vol. 3, pp. 189–214. Wiley-VCH, Weinheim (2010)
2. Raj, R., Rehman, A.: Can die configuration influence field-assisted sintering of oxides in the SPS process? J. Am. Ceram. Soc. **96**, 3697–3700 (2013)
3. Anselmi-Tamburinin, U., Garay, J.E., Munir, Z.A.: Fast low-temperature consolidation of bulk nanometric ceramic materials. Scripta Mater. **54**, 823–828 (2006)
4. Salamon, D., Maca, K., Shen, Z.J.: Rapid sintering of crack-free zirconia ceramics by pressure-less spark plasma sintering. Scripta Mater. **66**, 899–902 (2012)
5. Trunec, M., Maca, K., Shen, Z.J.: Warm pressing of zirconia nanoparticles by the spark plasma sintering technique. Scripta Mater. **59**, 23–26 (2008)
6. Inoue, K.: US Patent No.3,241,956 (1966)
7. Tokita, M.: Mechanism of spark plasma sintering. J. Soc. Powder Technol. Jpn. **30**, 790–804 (1993)
8. Yamazaki, K., Risbud, S.H., Aoyama, H., Shoda, K.: PAS (Plasma activated sintering): transient sintering process control for rapid consolidation of powders. J. Mater. Proc. Tech. **56**, 955–965 (1996)
9. Merkle, B.D., Borton, O.M., Goldberger, W.M.: Development of the electroconsolidation process. Adv. Powder Metall. Part. Mater. **3**, 93–101 (2000)
10. Munir, Z.A., Quach, D.V.: Electric current activation of sintering: a review of the pulsed electric current sintering process. J. Am. Ceram. Soc. **94**, 1–19 (2011)
11. Mayo, M.J.: Processing of nanocrystalline ceramics from ultrafine particles. Inter. Mater. Rev. **41**, 85–115 (1996)
12. Shen, Z.J., Nygren, M.: Microstructural prototyping of ceramics by kinetic engineering: applications of spark plasma sintering. Chem. Rec. **5**, 173–184 (2005)
13. Shen, Z.J., Peng, H., Liu, J., Nygren, M.: Conversion from nano- to micron-sized structures: experimental

observations. J. Eur. Ceram. Soc. **24**, 3447–3452 (2004)
14. Xu, J., Casolco, S.R., Garay, J.E.: Effect of varying displacement rates on the densification of nanostructured zirconia by current activation. J. Am. Ceram. Soc. **92**, 1506–1513 (2009)
15. Xiong, Y., Fu, Z.Y., Pouchly, V., Maca, K., Shen, Z.J.: Preparation of transparent 3Y-TZP nanoceramics with no low-temperature degradation. J. Am. Ceram. Soc. **97**, 1402–1406 (2014)
16. Chen, P.L., Chen, I.W.: Sintering of fine oxide powder. II. Sintering mechanisms. J. Am. Ceram. Soc. **80**, 637–645 (1997)
17. Xiong, Y., Hu, J.F., Shen, Z.J., Pouchly, K., Maca, V.: Preparation of transparent nanoceramics by suppressing pore coalescence. J. Am. Ceram. Soc. **94**, 4269–4273 (2011)
18. Xiong, Y., Hu, J.F., Shen, Z.J.: Dynamic pore coalescence in nanoceramic consolidated by two-step sintering procedure. J. Eur. Ceram. Soc. **33**, 2087–2092 (2013)
19. Munir, Z.A., Quach, D.V.: Electric current activation of sintering: a review of the pulsed electric current sintering process. J. Am. Ceram. Soc. **94**, 1–19 (2011)
20. Xiong, Y., Fu, Z.Y., Pouchly, V., Maca, K., Shen, Z.J.: Preparation of transparent 3Y-TZP nanoceramics with no low-temperature degradation. J. Am. Ceram. Soc. **97**, 1402–1406 (2014)

Nanochannels for Nanofluidics: Fabrication Aspects

Aeraj ul Haque[1] and Aloke Kumar[2]
[1]Biodetection Technologies Section, Energy Systems Division, Argonne National Laboratory, Lemont, IL, USA
[2]Department of Mechanical Engineering, University of Alberta, AB, Canada

Synonyms

Nanofluidic channels; Nanopores; Nanoslits

Definition

Nanochannels or nanoslits are defined as channels with at least one physical dimension equal to or less than 100 nm. The term has, however, been loosely applied to cover channels with a critical dimension as large as 500 nm. Nanopore is a term sometimes used interchangeably with nanochannels. Here, nanochannels refer to channels where fluid flow is on the surface of the substrate, while nanopores are referred to as channels where flow is through the substrate. Another important concept in the field of nanochannels is 1D and 2D nanochannels. A 1D nanochannel can be classified as a channel with both physical dimensions in the nanometer regime, while a 2D nanochannel is defined as one with only one dimension in the nanometer range. These terms, however, have been used interchangeably by some researchers leading to confusion. A better parameter is aspect ratio, that is, the ratio between the height and width or between the diameter and length (e.g., in case of nanotubes). A ratio of more than 100 would represent a high aspect ratio. Nanofluidics refers to study and manipulation of fluid flow in nanochannels. This is the area where nanochannels are finding most promising applications and the following discussion will only focus on applications of nanochannels to nanofluidic technologies.

Overview

Nanochannels are defined as channels that have at least one physical dimension that is less than or equal to 100 nm. The area of research that deals with the behavior of fluids, particularly ionic fluids in nanochannels is called nanofluidics. While nanofluidic encompasses analysis of liquid, gas, or super critical fluid in nanochannels, it is the study of ionic liquids and charged particles in nanochannels that has generated the most interest. This is because of the unique nanoscale electrostatic interactions, entropy, electrodynamics, and steric phenomenon observed in nanochannels that are not observed in microchannels [1]. In nanochannels, the Debye screening length that defines the extent to which surface charge of the nanochannel walls interacts with the bulk fluid is now comparable to the critical dimension of the nanochannel [2]. This proximity allows the wall surface charge to interact with charged particles such as ion, proteins, or DNA in fluid inside the

nanochannel, and surface charge effects can no longer be ignored as is the case in microfluidics. The dominance of ionic interactions leads to unique transport phenomenon caused by complex electrical double layer-electrical double layer (EDL-EDL) interactions, dominant streaming potentials, and particle/biomolecule-EDL interactions [1].

This behavior of fluids in nanochannels is now being exploited for applications targeting biomolecule separation and detection. Some of the techniques used include *co-ion exclusion*. Here the concentration of co-ions is highly depleted in the nanochannel compared to the bulk solution. Electrokinetic flow causes other interesting phenomenon like *concentration polarization* where ions of opposing charges are concentrated at the anode and cathode [3]. Using this approach DNA was shown to be depleted at one end and concentrated at the other end of a nanochannel. These features of nanofluidics have led to extensive fundamental studies and they are now being exploited for biological applications [1]. Biosensing using nanofluidics is beneficial because the small length scales and electrostatic interactions can be leveraged for facile separation of biomolecules, increasing the speed and sensitivity of bioassays. Many applications have targeted DNA separation, concentration, and detection.

Nanochannels are also relevant to cellular organisms. Every living cell is host to a broad collection of nanochannels that traverse the cell membrane and take part in active transport of ions and key nutrients. The most important ones are ion channels; they are central to balancing the intra- and extracellular concentrations of the physiological ions Na^+, K^+, Ca^{2+}, and Cl^- [4] maintaining the cell membrane potential. Also important are aquaporins that nonselectively transport water, ions, and metabolites through the cell. These naturally occurring macromolecular protein-based nanochannels play other significant roles including cell signaling and are the target of many therapeutic drugs. It is no coincidence that the first attempts at nanochannel synthesis involved the naturally occurring protein nanopore – α-hemolysin, a 10-nm long and 1.4-nm internal diameter protein that was involved in lipid bilayer membrane. The setup was used to study DNA translocation through the nanopore in the presence of an electric field, and was promoted as an extremely sensitive single molecule detector. Since then many efforts have focused on fabricating nanopores as small as 2 nm that can be employed for unique applications in single molecule detection and DNA sequencing.

Nanofabrication is a key enabler of nanofluidic research and like microfluidics, the area of nanochannel fabrication has greatly benefitted from the rich knowledge in silicon processing available from the microelectronics industry. Silicon and silicon oxide are, therefore, the most commonly used substrates for nanochannel fabrication. Recently, other cheaper materials such as polydimethylsiloxane (PDMS) and polymethylmethacrylate (PMMA) are becoming popular alternatives for creating nanochannels.

Nanochannels can be fabricated using either top-down approach or bottom-up approach. The top-down approach involves selective deposition and removal of material from substrates generally using classical nanofabrication methods such as photolithography, wet and dry etching, and substrate bonding. This is currently the dominant approach and complex features can be fabricated with relative ease. The equipment and protocols used for top-down fabrication are however expensive, which can drive up the cost of nanochannel devices substantially. In the bottom-up approach, nanochannels are either created through self-assembly or through electrochemical treatment of a suitable substrate such as alumina. This approach can be more affordable and a large array of nanochannels can be created very easily, though spatial control is difficult. Therefore, it is not easy to fabricate complicated geometries and interfacing with the macro-world might also be a challenge.

Historical Perspective

In 1939, bacterial gramicidin, a naturally occurring antibiotic and a nanochannel were isolated by Rene Dubos. The nAChR ion channel was the first naturally occurring ion channel to be isolated. In 1996, α-hemolysin was incorporated into a lipid bilayer and in a first, the possibility of using

nanochannels for performing a practical application – DNA sequencing was demonstrated [5]. Fabrication of tunable nanopores in SiO_2 and SiN with minimum dimensions ~1.8 nm using ion beam sculpting was first reported by Golovchenko and colleagues at Harvard in 2001. Dekker's group at Delft demonstrated the patterning of 20-nm diameter pores in Si, SiO_2, and SiN with electron beam lithography (EBL) and then shrinkage of these pores under tunneling electron microscope (TEM). Nanoimprint lithography was developed in 1996 by Chou and colleagues [6] and in 1997, Stern et al. [2] demonstrated the fabrication of 50-nm high nanochannels by etching a buried interfacial amorphous silicon sacrificial layer.

Basic Methodology and Key Findings

Fabricating Nanochannels with the Top-Down Approach

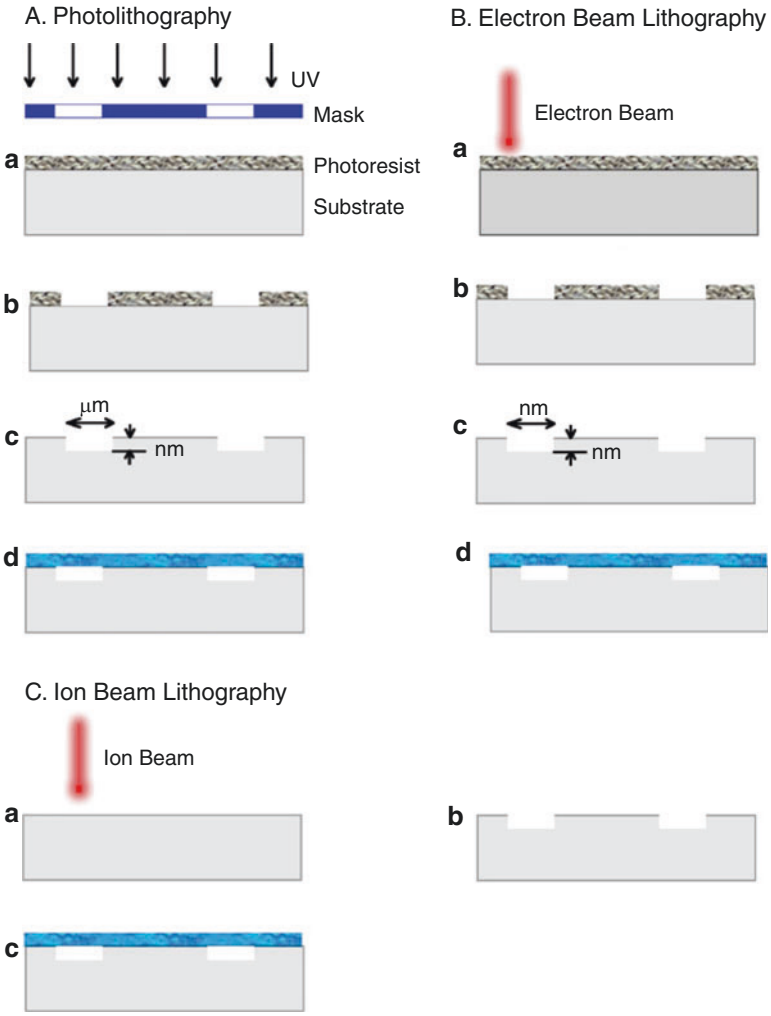
Generally the top-down techniques used for fabricating nanochannels employ fabrication protocols borrowed from the microelectronics industry. These include photolithography, physical deposition, and wet and dry etching, to name a few.

Photolithography

The simplest and most convenient method of fabricating nanochannels is through conventional UV photolithography. A UV-sensitive photopolymer or photoresist is spun coat on silicon, SiO_2, glass, or other appropriate substrate and then exposed to a UV light source through a glass mask. This mask has the nanochannel pattern defined in it. The pattern can also be exposed serially using a UV laser writer. The photopolymer is developed in a chemical developer that removes the unexposed or exposed resist (depending on the resist tone) and transfers the pattern in the photoresist. This is followed with a wet or dry etching step that etches the channel in the substrate. The resolution of conventional UV photolithography used in academia is generally in the micron range (1–3 μm). With advanced photolithography techniques using deep UV, extreme UV light source, or X-rays and expensive diamond masks, resolution below 100 nm can be obtained. These instruments and techniques are, however, very expensive and not readily available in the academia. Therefore, fabrication of nanochannels by photolithography is generally achieved by controlling the etch depth. Nanochannels fabricated using conventional photolithography are, therefore, 2D nanochannels. The channels are closed by bonding silicon, glass, or PDMS cover on the substrate. High aspect ratio nanochannels with depth as small as 2 nm have been fabricated with this method [7].

Electron Beam Lithography (EBL)

In EBL, a high-powered electron beam is used to transfer the pattern into an appropriate e-beam-sensitive resist. This is a maskless process as the pattern is directly written into the resist with the e-beam. The most commonly used e-beam resist is polymethylmethacrylate (PMMA), but now more advanced resists such as ZEP and Hydrogen Silsequioxane (HSQ) that provide better resolution are becoming increasingly popular. Beam diameter smaller than 10 nm are possible; so theoretically line widths of that order can be directly patterned onto the e-beam resist. Practically however, this is a tough task as inelastic scattering often causes over exposure of the resist. Also at very small dimensions over development of the resist is always a concern. This can be overcome using cold development processes at sub-zero temperatures [8]. The patterns can then be transferred in the underlying substrate (Si, glass etc.) with a wet or dry etching process. EBL combined with RIE has also been used to fabricate nanopores that have diameters in the tens of nanometers in Si, SiO_2, and Si_2N_3 membranes. In an interesting discovery made by Dekker's group [9], exposure of nanopores formed in silicon membranes to high-energy electron beams was found to fluidize the silicon and cause shrinkage of the nanopores. Switching the beam off quenches the shrinkage process and the new reduced diameter pore retains its shape. Using this method single-nanometer size pores were fabricated successfully. The central advantage of EBL is the higher degree of control and the possibility of obtaining channels with varying widths in the nanometer



Nanochannels for Nanofluidics: Fabrication Aspects, Fig. 1 (**A**) Nanochannel patterning using photolithography; (*a*) Photoresist spun coat on the nanochannel substrate and exposed to UV light through a photomask that has the nanochannel pattern defined in it. After development of the photoresist (*b*) the substrate is etched with RIE to the desired nm depth. A second substrate is bonded on this substrate to close the nanochannel. (**B**) Nanochannel definition with EBL; (*a*) An electron beam is used to directly write the nanochannel pattern on an e-beam-sensitive resist spun coat on the substrate. After photoresist development (*b*) RIE is used to etch nanochannels in the substrate. (**C**) With IBL an ion beam is used to directly etch the pattern in an appropriate substrate

regime. The key drawback of EBL is that due to its serial patterning nature, it is a slow technique and thus throughput is very low. It also requires highly trained staff and is relatively expensive.

Ion Beam Lithography (IBL)

Similar to EBL, Ion Beam Lithography is a serial patterning technique that uses a focused beam of ions such as protons, or heavy ions such as gallium, to directly write the pattern. The heavier ions scatter less and therefore, IBL is assumed to yield better resolution compared to EBL. It can be used for directly etching and fabricating nanochannels in substrates like silicon and glass without the need of a photoresist, a major advantage over EBL. Nanochannels down to 20 nm were directly milled with IBL in silicon and glass, and proton beams were also used to fabricate nanochannels in polymers such as PMMA. Nanopores in silicon oxide and silicon nitride membranes as small as 5 nm have been reported with IBL [10]. This technique, however, suffers from the same drawbacks as EBL – it has very low throughput, requires highly skilled researchers for operation, and is expensive. The doping of the substrate with the heavy metal ions is also a concern (Fig. 1).

Sacrificial Technique

In sacrificial techniques, an intermediate sacrificial layer is generally used between the two substrates that will be bonded to form the

nanochannels. A less than 100 nm thick layer of a sacrificial material in the pattern of the nanochannel is first deposited on the nanochannel substrate in the desired thickness. This thickness later defines the depth of the nanochannel. For this reason it is important that a sacrificial material is chosen that can be deposited at a thickness below 100 nm with high uniformity. Commonly used sacrificial materials include amorphous silicon, polysilicon, and even metals such as chromium and gold. The deposited sacrificial layer is first patterned generally using photolithography and dry or wet etching into the desired nanochannel shape. The second nanochannel substrate is then bonded with the substrate that was patterned with the sacrificial layer. The two substrates bond only in places where they come in direct contact with each other and do not bond with the sacrificial layer. In another variation, a structural layer can also be deposited on the sacrificial layer. The sacrificial layer is now wet etched in an appropriate etchant that selectively etches the sacrificial layer, but does not or minimally affects the substrates, leaving behind nanochannels. Using this approach it is possible to make high aspect ratio channels, and nanochannels that were 20-nm deep, 0.5 to 200-μm wide and up to 4-mm long were reported [2]. Since wet etching of the channels is solely based on diffusion inside the channel, this is a slow approach. It can take a couple of days to etch through the entire sacrificial layer. Also because selectivity of etchant toward sacrificial layer versus the substrate is not always 100 %, tapering is often observed in nanochannels fabricated using this technique.

Nanomolding Techniques

Micromolding of polymer elastomeric materials provides a cheap and easy way for rapid prototyping of devices and therefore occupies widespread popularity in the microfluidics arena. Polydimethylsiloxane (PDMS) is by far the most popular material used for micromolding of fluidic channels because it can retain the features of the mold with up to 99 % accuracy. Nanochannel fabrication in elastomeric materials is much more difficult mainly because at nanometer dimensions, they tend to deform and collapse [6]. This inhibits direct molding of nanochannels in PDMS using molds of the nanochannels. If designed properly, however, this deformation and collapse can actually be used for fabricating nanochannels. One method is crack-induced generation of nanochannels in PDMS. When a slab of PDMS is surface oxidized, for example, in an oxygen plasma, the surface hardens creating an elastic modulus mismatch between the surface and the bulk PDMS. Application of a uniaxial tensile force generates cracks in the slab forming triangular nanochannels. The nanochannels can be sealed by oxygen plasma bonding with another PDMS slab, glass, silicon or by using this slab as a mold for transferring the nanochannel pattern into another PDMS layer. Nanochannels with heights of 80 nm have been fabricated with this technique. In an alternative approach, thinner and thicker PDMS slabs that have both been surface oxidized in oxygen plasma are bonded together and a uniaxial tensile force is applied. This causes crack generation at the weak oxidized bonding interface producing diamond-shaped nanochannels. In both these cases, tuning of the tensile force controls the depth of the nanochannels. In a similar approach called wrinkle nanochannels (WNC), a stretched slab of PDMS is exposed to an oxygen plasma or UV light source and then relaxed. This creates wrinkle patterns along the PDMS surface. The wavelength of this sinusoidal wave pattern is controlled by the difference in elastic moduli of the PDMS slab and the substrate on which it is bonded, while the amplitude depends on the applied strain. With this technique nanoslits with widths and heights in the tens of nanometers have been fabricated.

Other researchers have used the inherent elasticity of PDMS that allows it to deform around structures to fabricate nanochannels. In the first method, shallow beams were fabricated on a silicon substrate using reactive ion etching (RIE). A thin film of oxygen plasma–treated PDMS is bonded on this silicon substrate. The PDMS film deforms in the crevices of the substrate in a wave like pattern and bonds to the top of the beams and part of the base of the crevice. Nanosized gaps at the edge of the beams and the point where PDMS bonds at the base of the crevice result in the

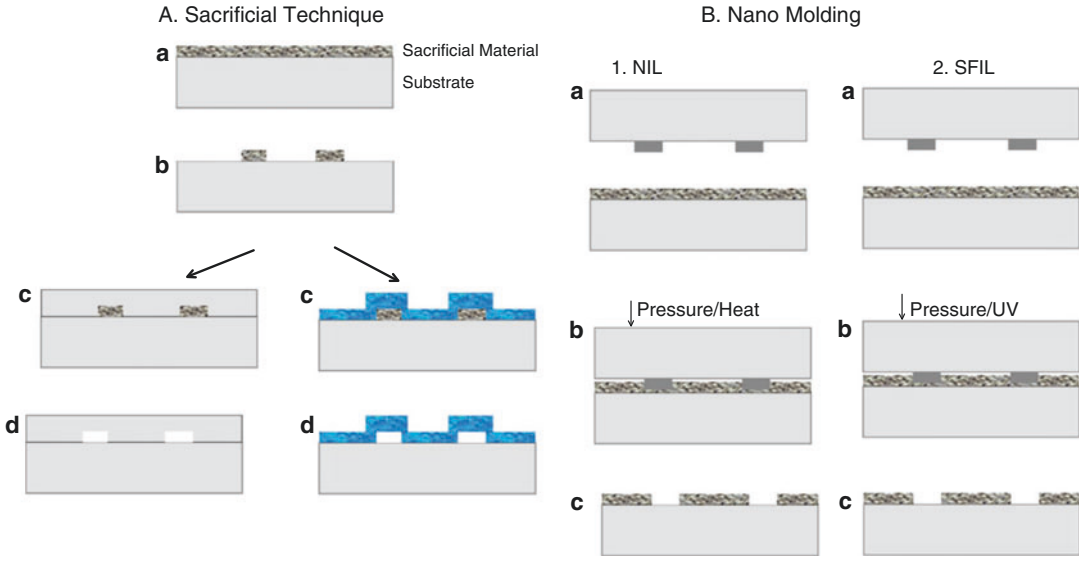
formation of nanochannels. In a similar technique called roof collapse technique, precision steps with strictly controlled height-to-width ratio of 0.2 were fabricated [6]. These are then used to mold microchannels in PDMS. When this PDMS slab is bonded to a substrate, the roof of the microchannel collapses. While the bulk of the channel is sealed nanoslits still remain open at the edges of the microchannels. With this technique nanochannels with heights as small as 60 nm have been fabricated. This technique is particularly applicable for creating variously shaped nanochannels.

Unlike elastomers, thermoplastics such as poly (methyl methacrylate) (PMMA), cyclic olefin co (polymer) (COC), and poly(ethylene terephthalate) (PET) deform negligibly and can retain their shape on direct imprinting with nanoscale features. This is the principle behind nano imprint lithography (NIL) pioneered by Chou et al. [11, 12]. A nanomold or template is first created generally using EBL followed by RIE in silicon or glass, or by direct writing using focused ion beam (FIB). A thermoplastic resist is spin coated on another substrate such as silicon and heated above its glass transition temperature (T_g). The template is now pressed on the spun coat thermoplastic under vacuum and the thermoplastic is allowed to cure. The template is released and a short RIE is performed to remove the residual thermoplastic leaving the nanochannel patterns embossed in it. The thermoplastic can itself be used as the material for the nanochannels or the thermoplastic can be used as a mask for etching the underlying substrate with RIE. With this technique 20-nm deep nanochannels that are centimeters in length were fabricated for nanofluidic applications. In a similar technique called step and flash imprint lithography (SFIL), a UV curable polymer is spun coat on an appropriate substrate. A template of the nanochannel pattern is fabricated in a transparent substrate as discussed before and pressed on the photoresist. The photoresist is exposed to UV light through the transparent template curing it and embossing the pattern in it. The template is removed and a short RIE step is performed to remove the residual photoresist layer. Another embossing technique is called reversal imprint. Here, a liquid polymer or prepolymer is spin coated directly on the template that is previously surface treated with fluorosilanes or other adhesion inhibitors. The template with liquid polymer is now transferred on another substrate and hardened. The template is removed such that the hardened polymer is left on the substrate. In a novel technique, silica nanowires were used as the nanochannel template. Silica nanowires, produced by tapering single-mode optical fibers in an alcohol flame [6], were positioned on a glass substrate using a scanning tunneling microscope (STM). The nanowire was hot embossed in a polycarbonate substrate and then etched away with hydrofluoric acid yielding the nanochannel in polycarbonate. The polycarbonate nanochannel was closed with PDMS. The need for STM position makes this technique very slow. Also this nanowire molding is confined to straight nanochannels.

The greatest advantage of nanomolding techniques is that they can yield high throughput and are relatively inexpensive. Particularly, NIL is very popular since equipment and materials for nanomolding using this technique are now available commercially through NanoNex®. The real expensive part in nanomolding techniques is the template that still has to be fabricated with EBL or FIB, but once the mold is machined it can be reused to imprint multiple devices (Fig. 2).

Miscellaneous Top-Down Techniques

Edge Lithography takes the advantage of asymmetry at the edge of a structure to fabricate nanochannels. In one example, polysilicon was deposited at the edge of a step on SiO_2 film. This was followed with anisotropic etching of polysilicon and the SiO_2 film such that the thickness of the polysilicon is converted into the width of the nanochannel. In another example, a 10-nm thick SiO_2 film was deposited on a Si_2N_3 substrate and then further coated with another Si_2N_3 masking layer. With controlled under etching 20 nm high and 200 nm wide nanochannels were fabricated. *Phase shift lithography* (PSL) is also considered as an edge lithography technique [2] that uses phase-shift photomasks to beat the diffraction limit of UV light. These are transparent

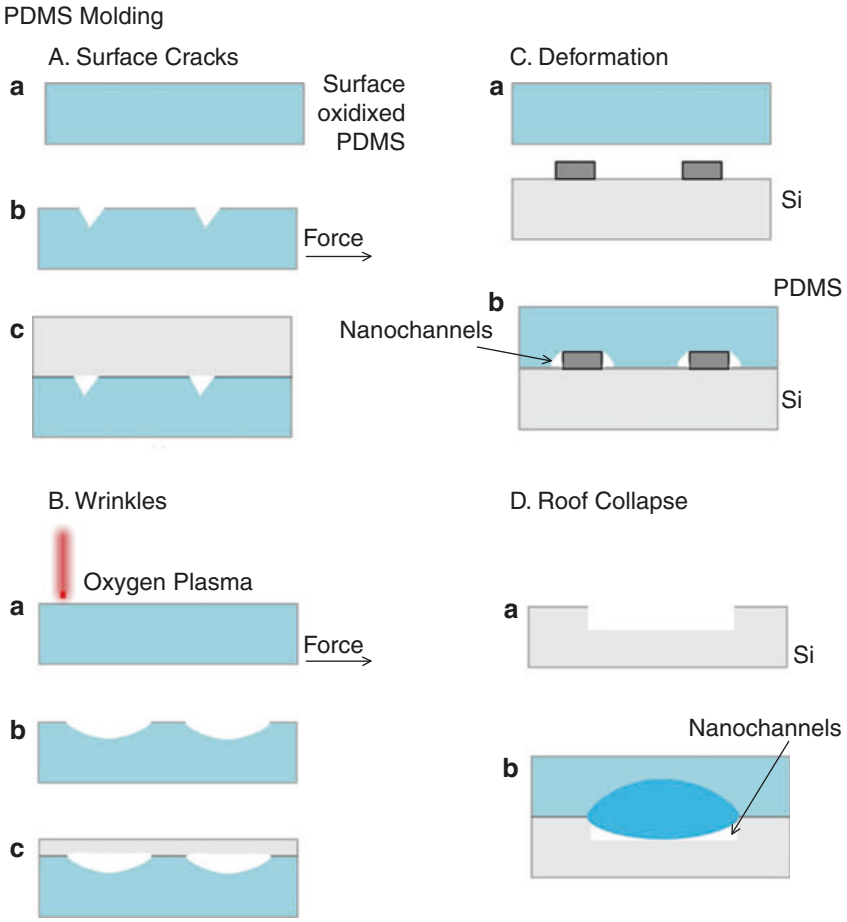


Nanochannels for Nanofluidics: Fabrication Aspects, Fig. 2 (**A**) Nanochannel patterning with sacrificial technique; (*a*) An appropriate sacrificial material is deposited on a substrate and the nanochannel pattern defined with photolithography (*b*). Then either a second substrate is directly bonded on top of this substrate or another structural material is deposited (*c*). The sacrificial material is then etched leaving the nanochannels. (**B**) A thermoplastic (*1a*) or a photosensitive (*2a*) polymer is spun coat on a substrate and embossed with a mold of the nanochannels. The polymer is then applied to heat and pressure (*1b*) or UV light (*2b*) curing the polymer. The mold is removed (*1c*, *2c*) leaving the nanochannels defined in the polymer

masks that include trenches with depth $d = \lambda [2(n - 1)]$ where λ is the wavelength of light and n is an integer. The phase shift in light caused by the trenches result in destructive interference, and interference patterns that are below 100 nm resolution down to 30 nm have been reported with this technique. *Interference lithography* (*IL*) is another interference pattern–based lithography technique. Here, the photoresist is exposed to the interference pattern produced by two coherent laser beams whose fringe period is given by $T = (\lambda/2)\ sin\ (\theta/2)$ where θ is the angle between the two light waves. IL is difficult and the setup to produce the interference patterns is complex but it can yield a dense array of nanochannels at a high rate without a mask. There is, however, limited flexibility in the number of shapes that can be patterned with this technique.

A novel technique was demonstrated by Han's group using *junction gap breakdown* to fabricate nanogap nanochannels in PDMS [13]. Two chevron-shaped microchannels were fabricated on either side of a horizontal microchannel such that there was a 40 μm gap between the tip of the chevron channels and the central horizontal channel. The channels were filled with an ionic buffer and a high voltage was applied with the horizontal channel as the anode while the chevron channels as cathode. This caused a junction gap breakdown at the tip of the chevron channels resulting in nanogap nanochannels. Pulling optical fibers and capillaries that are heated above their glass transition temperature can also be used to fabricate nanochannels. Co-axial fibers that were a mixture of silica sol–gel and motor oil were extruded to yield silica nanochannels with an inner diameter of 20 nm after annealing and removal of the oil. Only vertical nanochannels shaped like venturies can be fabricated with this technique (Fig. 3).

Atomic Force Microscopy (AFM) Nanolithography has also been used to create nanoscale features and depressions on substrates. Using high aspect ratio Si tips 80-nm wide and 30-nm deep channels were fabricated in PMMA [6]. AFM nanolithography like EBL and FIB is a



Nanochannels for Nanofluidics: Fabrication Aspects, Fig. 3 Fabrication of nanochannels in PDMS using (**A**) production of nanochannels through formation of cracks and (**B**) formation of wrinkles. Both techniques use surface oxidation of PDMS and application of a tensile force to create these nanochannels. Bonding with a second PDMS, silicon or glass layer closes the nanochannels. Nanochannels can also be formed in PDMS using deformation around pillars patterned in a substrate (**C**) or collapsing of the roof (**D**) of a PDMS microchannel that yields nanochannels at the bottom corners of the pillar

serial technique and is in fact even more tedious and slower than the beam-based techniques. While throughput can be increased using multiple probe tips such as on Dip Pen Nanolithography systems available from NanoInk®, limitations in scan area and the collateral damage around the indentation site reduce the practicality of this technique.

An attractive method of fabricating nanochannels is the fabrication of microchannels in a substrate followed by anisotropic deposition of another material. Controlled deposition converts the microchannel to a nanochannel. Using this approach parylene was deposited in microchannels etched in silicon. The parylene deposited preferentially at the top of the channel compared to the base. This caused the channel to close at top forming parylene nanotubes. As parylene is biocompatible these nanotubes can find biomedical applications such as in implants or as synthetic arteries or veins. A layer-by-layer polyelectrolyte multilayer deposition approach and non-conformal deposition was also used to convert microchannels to nanochannels [14]. Microchannels 1 μm wide and 15 μm deep were fabricated in silicon with photolithography

and deep RIE. Plasma Enhanced Chemical Vapor Deposition (PECVD) was then used to deposit silica on the device. The silica deposited preferentially on the top of the channels reducing its width at that location. Next conformal Low Pressure Chemical Vapor Deposition (LPCVD) of polysilicon was deposited that closed the microchannel at the top. After thermal oxidation to reduce the channel width to the desired level, layer-by-layer deposition of polyelectrolyte films was performed to reduce the channel width to 200 nm.

Bottom-Up Nanochannel Fabrication Techniques

Bottom-up fabrication of nanochannels involves processes where nanochannels are created electrochemically, through controlled atomic synthesis and assembly of synthetic nanochannels/nanopores and through self-assembly of polymers or naturally occurring nanochannels.

Electrochemical Synthesis

Thin sheets of aluminum can be anodized in acid solutions such as oxalic acid, sulfuric acid, or phosphoric acid with application of voltage/voltage cycles to produce aluminum oxide nanochannel arrays. These alumina nanochannel arrays called anodic aluminum oxide (AAO) or porous anodic alumina (PAA) have been around for a long time and are primarily used as nanofilters. Recently interest in using AAO as nanotemplates for ordered carbon nanotube (CNT) or nanowire growth and in nanofluidic applications has grown considerably. The anodization process involved the growth of an initial oxide layer on the surface of the aluminum film. Oxygen molecules then penetrate and oxidize the aluminum only at sites that provide low electrical resistance. The oxide dissolves in the acid resulting in pore nucleation and growth at these sites. The applied electrical field causes the pores to penetrate perpendicular to the aluminum surface until they traverse the entire alumina film. The pore radius, lengths, and densities can easily be tuned by controlling the number of the anodization voltage cycles and the applied voltage [15]. NIL was also used to produce dimples on the aluminum film that act as anodization nucleation sites [2].

Nanotubes: Carbon Nanotubes and Nanowires

With advances in physical vapor deposition techniques and catalysis it is now possible to synthesize nanotubes and nanowires with excellent control. Among these, carbon nanotubes (CNTs) have enjoyed great attention. CNTs are rolls of graphene, an atomic layer of carbon arranged in classic honey comb pattern. They can be single-walled (SWCNTs) and multi-walled (MWCNTs) and can have diameters from a couple of nanometers to hundreds of nanometers. They have excellent mechanical and electrical properties (tensile strength in the GPa range) and can be conducting or semi-conducting. They can easily be functionalized with biomolecules, leading to their vast use in bio-applications. CNTs can be synthesized using arc discharge, laser ablation, and chemical vapor deposition (CVD and MOCVD) methods. In the arc-discharge process, carbon atoms are vaporized in helium plasma ignited by an electric arc produced by applying high currents between a carbon anode and cathode. This process produces MWCNTs. To produce SWCNTs a metal catalyst such as iron or cobalt is added to the system. Laser ablation is similar to arc discharge, that is, it involves evaporation of a carbon target but now high-intensity laser pulses are employed for ablation. The process is generally performed in a tube furnace heated to 1,200 °C. The furnace is flushed with an inert gas that blows the synthesized CNTs downstream where they are collected. The CVD process involves synthesis of CNTs by the decomposition of a hydrocarbon such as acetylene in the presence of an appropriate catalyst. After nanotube synthesis the reactor is cooled down and the CNTs collected. By controlling the temperature and pressure inside the reactor, careful selection of precursor hydrocarbon and catalyst SWCNTs and MWCNTs can be produced in the same reactor. The CVD process is particularly advantageous for growing ordered arrays of CNTs on substrates that have been appropriately seeded with a catalyst.

An MOCVD process was used to fabricate CNT array in a PAA template that has found applications in CNT transistors and as a glucose biosensor [15]. While CNTs are readymade nanopores, their application is tricky mainly because of integration with the macro-world. CNT nanopores were fabricated in by trapping CNTs with dielectrophoresis and then spin coating SU-8 on top of it that is patterned to produce fluidic reservoirs.

Nanowires can also be used as templates for the fabrication of nanochannels. Nanowires can be synthesized using various techniques such as electroplating (for metallic nanowires), chemical vapor deposition, vapor–liquid–solid method, etc. Bottom-up fabrication of SiO_2 nanochannels was demonstrated with Si nanowires. The Si nanowires were surface oxidized via a timed oxidation process such that the core was still Si. Isotropic etching of the Si core was performed with XeF_2 gas that yielded SiO_2 nanochannels with hollow cores.

Block Copolymers

Block copolymers are soft materials that are composed of two or more immiscible macromolecules. These macromolecules can self-assemble into nanoscopic arrangements with each macromolecule comprising the block polymer maintaining its original properties [16]. Block copolymers can be used to form nanoporous membranes that have ordered nanopores. Polystyrene-block-PDMS (PS-PDMS) was spin coated on a silicon substrate. The block copolymer self-assembled to form ~16 nm diameter vertical cylinders of PS arranged uniformly at a pitch of ~23 nm in a PDMS block. Exposure to oxygen plasma etched the PS leaving nanopores in the unetched PDMS. The PDMS was later used as a mask for etching the underlying substrate. More recently efforts are being made to arrange long channel like arrays of block copolymers. One of the block copolymer phases can then be selectively etched to form long nanochannels. Misalignment over long distances limits the maximum feature length.

Self-Assembly of Naturally Occurring Nanochannels

Naturally occurring protein nanochannels are present in the cell phospholipid bilayer and the intracellular membranes of all eukaryotic and prokaryotic cells. They perform critical cellular functions from maintaining the ion balance, transport of metabolites to cell signaling, communication, and balancing osmotic pressure. Particularly important are ion channels that are also the target of therapeutics. Efforts have been directed toward isolating ion channels and using them as nanochannels for other applications. Porin MspA, a 3.1 nm internal diameter *Mycobacterium smegmatis* channel protein selective for hydrophobic nutrients, was self-assembled inside highly ordered pyrolytic graphite using hydrophobic interactions [17]. In another approach 1.4–4.6 nm nanopores were fabricated by self-assembly of α-hemolysin monomers in a lipid bilayer. Thus far demonstrations of self-assembled isolated naturally occurring nanochannels or their monomers are limited and practical applications are nonexistent. This is mainly because of difficulty of integrating such nanochannels with complex nanofluidic devices. Nonetheless, this represents a very interesting area of research and applications are expected to grow as better ways of integrating the top-down and bottom-up approaches are developed.

Future Directions

Nanofluidics is a frontier area of science today, and ease and affordability of nanofabrication can be key attributes to the broader embrace of the technology. Nanochannel fabrication on substrates like silicon and glass has its advantages such as high fidelity, but fabrication in such cases can suffer from high fabrication costs and low throughput. Advances in polymer-based nanofabrication can offset some of these disadvantages. However, polymer-based fabrication techniques can suffer from the inherent lack of mechanical rigidity of the material, leading to deformation issues. As advances in

nanofabrication are made, the quest for customized surface modification of channel walls and integration of various components into nanochannels will also increase.

Cross-References

- ▶ Applications of Nanofluidics
- ▶ Electron Beam Lithography (EBL)
- ▶ Nanoimprint Lithography
- ▶ Surface-Modified Microfluidics and Nanofluidics

References

1. Napoli, M., Eijkel, J.C.T., Pennathur, S.: Nanofluidic technology for biomolecule applications. Lab Chip **10**, 957–985 (2010)
2. Abgrall, P., Nguyen, N.T.: Nanofluidics devices and their applications. Anal. Chem. **80**, 2326–2341 (2008)
3. Kovarik, M.L., Jacobson, S.C.: Nanofluidics in lab-on-a-chip devices. Anal. Chem. **81**, 7133–7140 (2009)
4. Hille, B.: Ion Channels of Excitable Membranes, 3rd edn. Sinauer, Sunderland (2001)
5. Dekker, C.: Solid-state nanopores. Nat. Nanotechnol. **2**, 209–215 (2007)
6. Chantiwas, R., Park, S., Soper, S.A., Kim, B.C., Takayama, S., Sunkara, V., Hwang, H., Cho, Y.K.: Flexible fabrication and applications of polymeric nanochannels and nanoslits. Chem. Soc. Rev. **40**, 3677–3702 (2011)
7. Duan, C., Majumdar, A.: Anomalous ion transport in 2-nm hydrophilic nanochannels. Nat. Nanotechnol. **5**(12), 848–852 (2010)
8. Ocola, L.E., Stein, A.: Effect of cold development on improvement in electron-beam nanopatterning resolution and line roughness. J. Vac. Sci. Technol. B **24**, 3061–3065 (2006)
9. Storm, A.J., Chen, J.H., Ling, X.S., Zandbergen, H. W., Dekker, C.: Fabrication of solid-state nanopores with single-nanometre precision. Nat. Mater. **2**, 537–540 (2003)
10. Lo, C.J., Aref, T., Bezryadin, A.: Fabrication of symmetric sub-5 nm nanopores using focused ion and electron beams. Nanotechnology **17**, 3264–3267 (2006)
11. Chou, S.Y., Krauss, P.R., Renstrom, P.J.: Imprint lithography with 25-nanometer resolution. Science **272**, 85–87 (1996)
12. Perry, J.L., Kandlikar, S.G.: Review of fabrication of nanochannels for single phase liquid flow. Microfluid. Nanofluid. **2**, 185–193 (2006)
13. Lee, J.H., Chug, S., Kim, S.J., Han, J.: Poly(dimethylsiloxane)-based protein preconcentration using a nanogap generated by junction gap breakdown. Anal. Chem. **79**, 6868–6873 (2007)
14. DeRocher, J.P., Mao, P., Han, J., Rubner, M.F., Cohen, R.E.: Layer-by-layer assembly of polyelectrolytes in nanofluidic devices. Macromolecules **43**, 2430–2437 (2010)
15. Claussen, J.C., Franklin, A.D., Ul Haque, A., Porterfield, D.M., Fisher, T.S.: Electrochemical biosensor of nanocube-augmented carbon nanotube networks. ACS Nano **3**, 37–44 (2009)
16. Jackson, E.A., Hillmyer, M.A.: Nanoporous membranes derived from block co-polymers: from drug delivery to water filtration. ACS Nano **4**, 3548–3553 (2010)
17. Mijatovic, D., Eijkel, J.C.T., Van den Berg, A.: Technologies for nanofluidic systems: top-down vs bottom-up – a review. Lab Chip **5**, 492–500 (2005)

Nanocluster

- ▶ Synthesis of Subnanometric Metal Nanoparticles

Nanocoil

- ▶ Nanorobotics for NEMS Using Helical Nanostructures

Nanocombs

- ▶ Chemical Vapor Deposition (CVD)

Nanocomposite Materials

- ▶ Implementing Alternating Nanolaminates in Trenched Energy Storage Systems

Nanocomposites

- ▶ Implementing Alternating Nanolaminates in Trenched Energy Storage Systems

Nano-concrete

► Nanoengineered Concrete

Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

Umberto Anselmi-Tamburini and
Ilenia G. Tredici
Department of Chemistry, University of Pavia, Pavia, Italy

Synonyms

Crystallite; Densification; Massive; Nanometric; Nano-sized; Nanostructure; Sintered

Definition

Crystalline materials are defined as solid materials in which constituents (atoms, ions, molecules) are organized in space following a regular, symmetric long-range ordered lattice.

Functionality refers to typical chemicophysical properties owned by materials, which provide them the possibility to be exploited in technological applications, as in electrical conductors, thermoelectrics, ferroelectrics, magnets, and transparent optical devices.

Nanocrystals are characterized by at least one dimension under a critical value, generally 100 nm, below which peculiar properties related to the nanostructure tend to emerge.

Bulk form refers to materials presenting a macroscopic massive body, in contrast to loose powders.

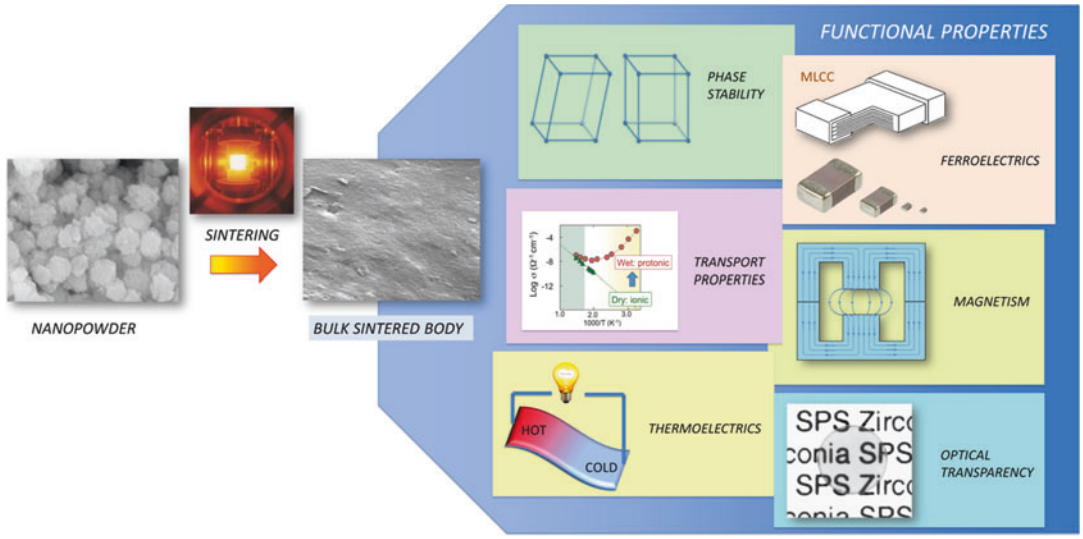
Overview

Nanostructure has a significant influence on the physical properties of functional materials, and the study of this influence has been the focus of a large research effort in the last two decades. Although some effects can be observed in all the range of nanometric dimension (below 100 nm), most of these modifications became evident only when the materials present at least one dimension falling in the low nanometric range, below 50 nm. These extreme dimensions can routinely be obtained for materials in the form of nanoparticles, thin films, or even more complex nanostructures but represent still a serious challenge in the case of materials in bulk form. On the other hand, several technological applications require materials to be in bulk form, and this justifies the interest toward their synthesis [1].

In the case of functional materials the only practical approach toward the synthesis of bulk nanostructured samples is represented by the densification of nanopowders. However, traditional densification techniques, such as pressureless sintering and hotpressing, are generally accompanied by such a significant grain growth that in most cases the nanostructure of the starting powders is lost. For this reason the synthesis of bulk nanometric functional materials has been obtained almost exclusively using nonconventional sintering methods. These methods will be introduced and described briefly in the following section. The influence of nanostructure on some physical properties of functional materials, with particular emphasis on phase stability, electrical and thermal transport, and optical and magnetic properties of nano-oxides, will be discussed in some detail in the subsequent sections (Fig. 1).

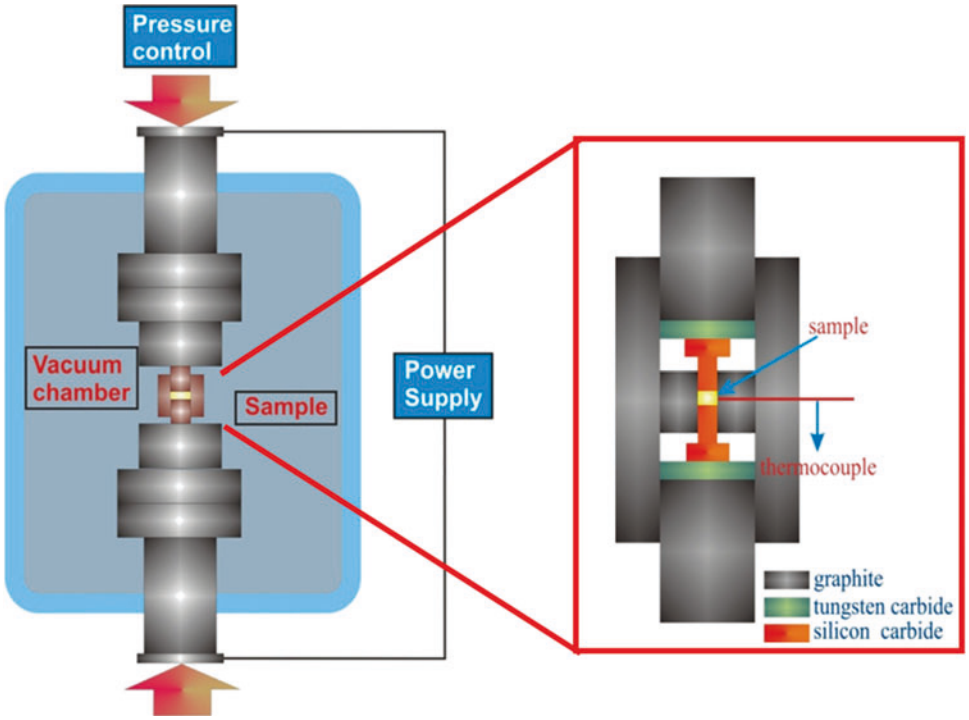
Synthesis Methods

As mentioned before, bulk nanocrystalline functional materials are generally obtained through the sintering of nanopowders. Other methods proposed over the years cannot generally be applied to functional materials. Severe plastic deformation methods, such as Equal-Channel Angular Pressing (ECAP) or High-Pressure Torsion (HPT), are limited to metallic materials, while the controlled crystallization of amorphous precursors can be successfully applied only to a very



Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 1 Nanostructure and functional properties. The schematics depict the main functional properties reviewed in the present essay

limited number of systems. The sintering of nanopowders presents the great advantage, at least in principle, of being applicable to almost any system, as the progress in the synthesis methods experienced in the last two decades allowed to obtain nanopowders of almost all the technologically relevant functional materials. Sintering, however, is always accompanied by a significant grain growth. The two processes, in fact, share the same mechanisms. For this reason the traditional sintering techniques, represented by pressureless sintering and hotpressing, have been largely unsuccessful in producing bulk materials characterized by high values of relative densities and grain size in the low nanometric range. These methods, in fact, involve the use of long annealing times (several hours) at high temperatures (usually around 2/3 of the melting point) and are mostly driven by bulk and grain boundary (GB) diffusion processes. In such conditions the nanostructure of the starting powders is usually lost. A further difficulty derives from the fact that most nanopowders present a relevant degree of agglomeration, particularly when they derive from solution methods. Agglomerates with dimensions in the range of micrometers or even tens of micrometers are not uncommon, resulting in green pellets characterized by the presence of macroporosity that can be eliminated only with great difficulty. To overcome these limitations, in the last decades nonconventional sintering methods have been actively investigated. Fast firing and two-step sintering have been first considered and produced some interesting results, but the real breakthrough has been obtained using a group of techniques that are known in the literature with the names Field Assisted Sintering (FAS), Spark Plasma Sintering (SPS), or Field Assisted Rapid Sintering (FARS) [2]. These techniques share the use of elevated uniaxial pressures combined with extremely rapid sintering cycles, resulting in a marked decrease in the associated grain growth. The reference to field and spark effects appearing in their name derives from the use of pulsed electric fields, although the relevance of this effect is still debated. Pressure, on the other hand, has been demonstrated to enhance drastically the densification of nanopowders. High pressures, in fact, can induce nanoparticle rearrangement, plastic and superplastic deformation, particle sliding, and pore shrinkage, allowing densification to be realized at lower temperatures. Figure 2 presents the schematics of a typical SPS apparatus [2]. The sample is usually contained in a die made out of an electrically conducting material (usually graphite, but high-strength steel and



Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 2 Schematics of the HP-FAST apparatus. Uniaxial pressure and electric current are applied to a graphite die, placed in a vacuum chamber. The inset shows an enlarged detailed view of the die setup: the high-pressure components are made of SiC or WC, while the dies and the low-pressure components are made of graphite. The temperature is recorded through a thermocouple inserted in the wall of the die

tungsten carbide have also been used), which is placed between two hydraulic rams acting also as electrodes. No furnace is present in these systems, but the die is heated up directly by joule heating. This is obtained through a low-voltage, high-intensity current flowing through the ram/electrode assembly. Current intensities of several thousands of amperes are usually employed. The absence of a real furnace, the low thermal inertia of the die, and the presence of an intimate contact between the sample and the die offer the possibility to realize very high heating rates. Values between 100 °C/min and 300 °C/min are commonly used, although values as high as 1000 °C/min have been reported. Commercial SPS apparatuses generally employ pulsed currents, with waveform that in most systems can be controlled over a wide range. In order to reduce even further the grain growth, high-pressure modification of the technique has been proposed. At higher pressures, in fact, densification is obtained at even lower temperatures, resulting in an even further decrease in the grain growth. Since the compressive strength of the graphite does not generally allow pressures higher than 140 MPa, double-stage dies, presenting a high-pressure section made out of a high-strength material, such as SiC or unbound WC, are usually employed (Fig. 2). Pressures as high as 1 GPa can be easily obtained using this setup [2]. Bulk, almost fully dense materials characterized by a grain size just above 10 nm and by a relative density (r.d.) above 98 % have been obtained with this approach, even in the case of ceramic and badly agglomerated nanopowders. This technique represents at the moment the best and more versatile approach toward the synthesis of bulk nanocrystalline functional materials.

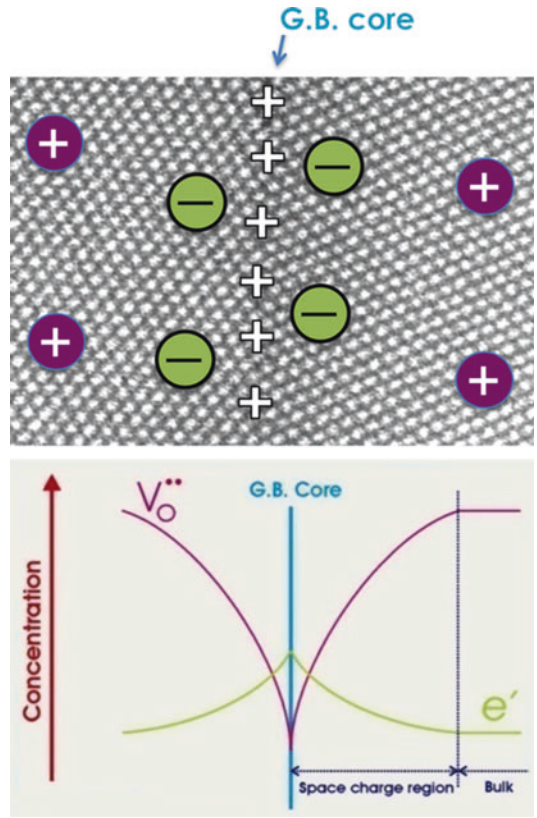
Effect of Nanostructure on Bulk Properties of Functional Materials

As in other class of materials, the nanostructure has a profound influence on the physical properties of bulk nanostructured functional ionic materials. In analogy with the term "nanoelectronics" used for semiconductors, the term "nanoionics" has been proposed for the study of this influence in ionic materials [3]. In some cases, nanostructuring simply enhances features that are already present in the material, while in other cases it introduces new features. Maier introduced the terms "trivial" and "real" (or "true") size effects to describe these two conditions [4]. In ionic materials most of the typical phenomena associated to the nanostructure fall in the first group. "True" size effects that are usually associated to confinement effects [5] are fairly uncommon in these materials.

The first, most obvious effect of nanostructuring is represented by a dramatic increase in the number of interfaces. In anionic materials this might produce some peculiar effects. In these materials, in fact, surfaces and interfaces are associated to a net electrostatic charge (Fig. 3). This charge interacts with the charged point defects present in the bulk of the materials, producing, in proximity to the interface, accumulation or depletion of point defects. The regions in which this accumulation or depletion takes place is generally named "Space Charge Region" (SCR) and typically extends for few nanometers around the GBs. In this region the defect concentration and distribution might vary considerably in respect to the one observed in the bulk of the material (Fig. 3) [4].

While in micrometric materials the effect of SCRs is negligible, in nanostructured materials, where the fraction of atoms located within these regions might become relevant, their influence might become significant and can produce macroscopically evident modification of the physical properties (Fig. 4) [6].

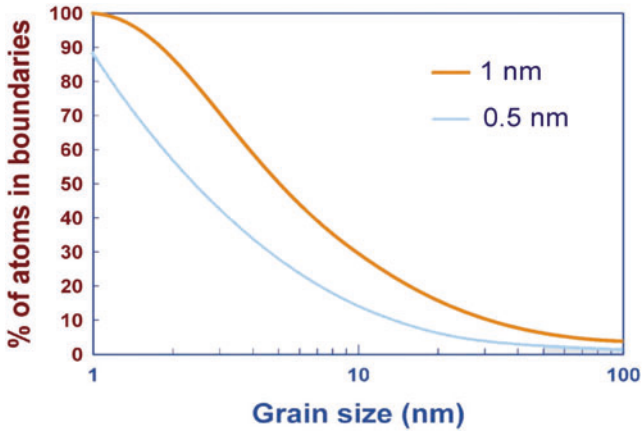
In the following we will discuss some example of effects associated to the nanostructure in functional ionic materials, focusing in particular on the ones related with phase stability, electrical conductivity, dielectric, ferroelectric, piezoelectric, magnetic properties and optical transparency.



Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 3 GB core and defect concentration in the SCR. In ionic oxides such as Ceria and Zirconia the GB core is positively charged and attracts negatively charged defects (such as electrons), while it repels positively charged ones (such as oxygen vacancies), producing an altered point defects distribution

Phase Stability

The modification of phase equilibria in polymorphic materials represents one of the best-known and more interesting effects of the nanostructure. Table 1 reports some examples in the case of functional oxides [7]. In all these cases the nanostructure tends to stabilize the high-temperature form of the material. In coarse-grained materials the same effect can only be obtained through alloying with doping elements. In the early

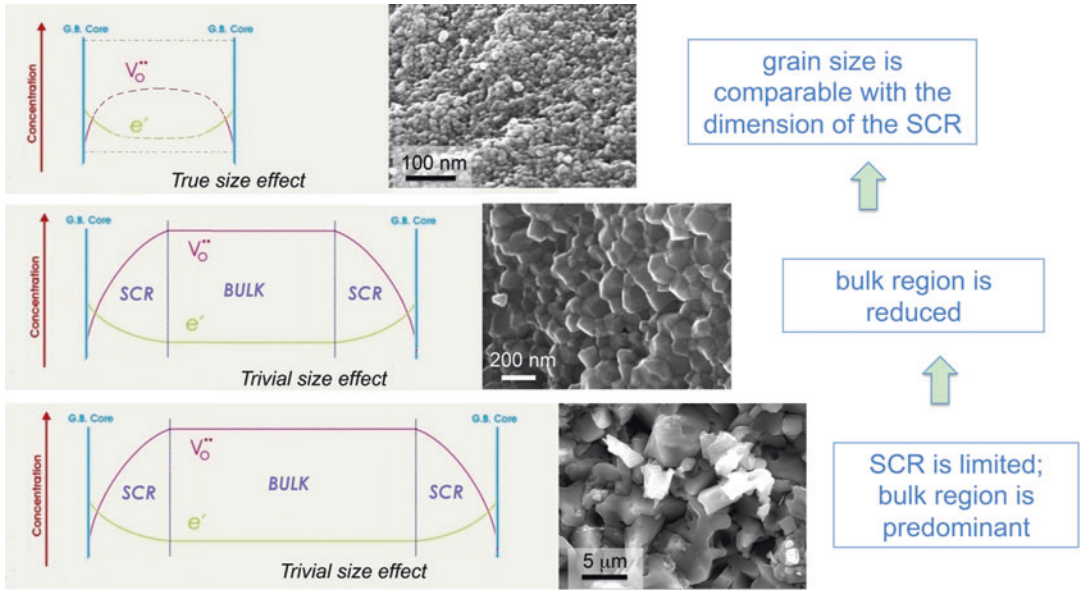


Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 4 Fraction of atoms located in proximity of the GB, as a function of grain size, for two distinct GB widths: 1 nm (*orange*) and 0.5 nm (*blue*) [6]

Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Table 1 Nanomaterials with unusual phase compositions. Crystal lattice structure designations: *Hex* hexagonal, *M* monoclinic, *T* tetragonal, and *C* cubic [7]

Material	Lattice type	
	Standard conditions	Nanomaterials
ZrO_2	M	T (C?)
TiO_2	Rutile	Anatase
Al_2O_3	Hex (α-Al_2O_3)	C (γ-Al_2O_3)
Y_2O_3	C (α-Y_2O_3)	M (γ-Y_2O_3)
$BaTiO_3$	T	C
$PbTiO_3$	T	C

literature it was suggested that the effect of the nanostructure was to stabilize kinetically metastable crystalline forms. There is now a general consensus that in most cases the crystallographic form observed in nanomaterials represents indeed the most thermodynamically stable form of the material in these conditions. The contributions driving toward this inversion in thermodynamic stability are various. The most important is surely represented by the contribution of the surface or interfacial energy. When materials with a grain size of few nm are considered a significant excess of free energy is introduced in the material. As a result, below a critical grain size the difference in free energies between two polymorphs might vanish [8]. Other contributions, however, must be considered, such as presence of strain energy, external and internal hydrostatic pressure, presence of structural similarities between the polymorphs, surface contamination (such as adsorbed water), bulk contaminants (particularly anions), and point defects (oxygen vacancies). As a result, only in few cases a general consensus exists on the conditions required for the observation of phase stabilization effects induced by the nanostructure. In the case of pure Zirconia, for instance, the size stabilization of the metastable tetragonal phase over the monoclinic form is known since early 1960, but it was observed only in nanopowders and thin films. Only recently it has been possible to produce size stabilization also in bulk sintered bodies. On the basis of surface energy alone, the critical grain size can be predicted in this system to be around 10 nm. However, agglomeration, strain energy, and presence of adsorbed molecules (water in particular) modify significantly this value. Agglomeration and densification increase the limit value. In fact, tetragonal Zirconia has been obtained in bulk form even with grain sizes around 30 nm (95 % tetragonal phase), when strongly agglomerated nanopowders were used to obtain fully dense sintered bodies in HP-FAST. On the contrary, when soft, nonagglomerated powders were sintered in the same conditions, a relevant transition to the monoclinic phase was detected [1]. Strain energy, on the other hand, can significantly reduce the critical grain size, justifying the experimental evidences of lack of stabilization even in loose powders characterized by extremely small grain size (even 5–6 nm). Concerning the presence of adsorbed water, in the case of Zirconia



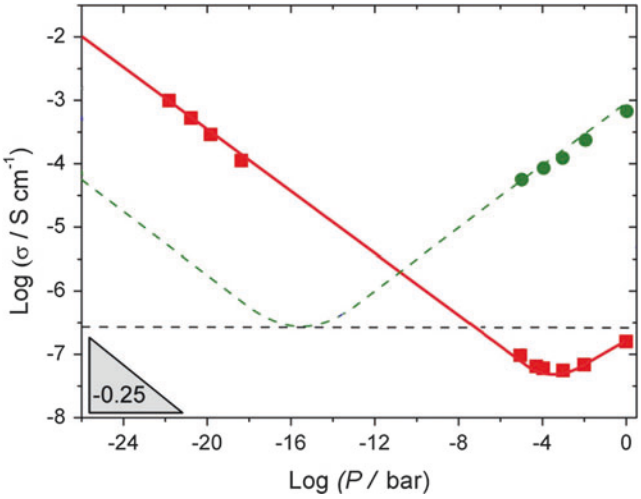
Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 5 Grain size and point defects distribution in ionic oxides. Decreasing the grain size from micrometric to nanometric produces a relevant alteration in the ratio between the extents of the SCR and of the bulk, regular material. As long as the bulk material is predominant, the interfaces and the SCRs are distant enough and the defect concentrations reach equilibrium in the bulk. When the grain size approaches the size of the SCR no more regular bulk material is availabble and significant modification in the electrical properties are expected

the exposure of the nanopowders to a moist environment drastically reduces the stability of the tetragonal phase, even in presence of extremely small grain sizes.

For titanium oxide, the transition from the anatase to the rutile form is generally believed to happen at a critical grain size of about 50 nm. It is generally quite difficult to keep this material in the anatase form during sintering: grain growth occurs at very low temperatures in Titania (between 400 °C and 700 °C), resulting in a transition to rutile structure. Also in this case many experimental parameters can influence the stabilization; one of the most important is the pressure applied during sintering, since the transition from anatase to rutile involves a reduction of the molar volume. Optimization of the sintering conditions allowed the synthesis of Titania in anatase form with high r.d. (>90 vol.%) with an average grain size of about 30 nm. The samples were obtained by hotpressing, at temperatures as low as 490 °C, under a pressure of 450 MPa, applied for 2 h. Even higher r.d. (above 95 vol.%) has been obtained using shorter sintering times with HP-FAST at 700 °C, under a uniaxial pressure of 620 MPa for 5 min. In this case the final samples were characterized by a grain size around 50 nm.

Electrical Conductivity

The bulk conductivity of functional ionic materials is generally interpreted on the basis of the SCR model described previously. In ionic conductors, in fact, a significant fraction of the overall resistivity is associated to the grain boundaries that are characterized by a "blocking" behavior toward the ionic charge carriers that are usually depleted in proximity of the SCR. Nanostructuring greatly increases the number of GBs, resulting in an even further decrease in the conductivity of the material. On the other side, when the scaling down reaches a value for which the grain size is comparable with the size of the SCR, a significant change in the defect equilibria is expected, as the "regular" bulk material tends to disappear (Fig. 5) [4].



Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 6 Effect of nanostructure on the conductivity of $SrTiO_3$ at different oxygen partial pressures. Conductivity was measured at a temperature of 544 °C. *Green dashed line*: microcrystalline $SrTiO_3$; *solid red line*: nanocrystalline $SrTiO_3$ with a grain size of ~30 nm (Adapted from Ref. [11])

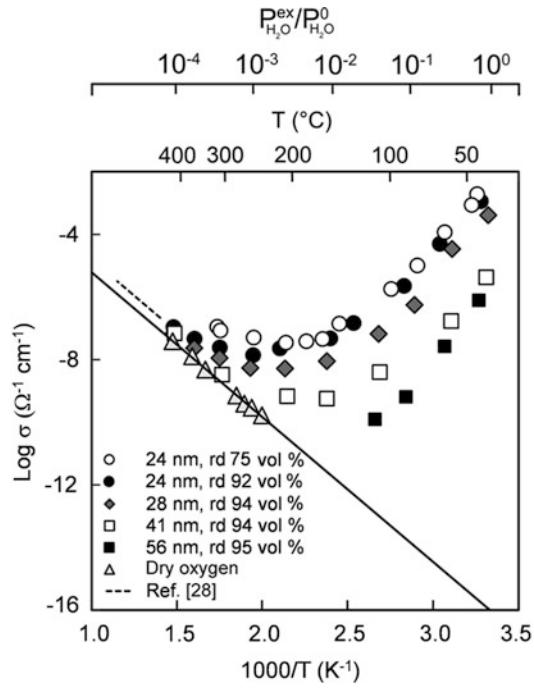
An experimental evidence for these phenomena has been firstly reported in the case of CaF_2/BaF_2 heterolayers [9]. When thickness of the alternating layers of CaF_2 and BaF_2 was decreased from 500 to 16 nm, an increase in the ionic conductivity of 2 orders of magnitude was observed and was interpreted on the basis of the model described in Fig. 5. In polycrystalline materials such relevant effect on the ionic conductivity has not been found even in materials characterized by a grain size well below 50 nm. In the case of Zirconia, reduction of the grain size down to 10 nm resulted only in an overall decrease in conductivity, although a small increase in the specific GB conductivity has been observed. However, this enhancement is not enough to increase the overall conductivity, as the GB conductivity remains more than one order of magnitude lower than the bulk conductivity. Guo argued recently that in order to observe a significant increase in conductivity due to size effects, the grain dimension should be decreased to less than 5 nm, since the typical SCR dimension is, for Zirconia, around 2.5 nm [10]. Such grain size, however, has never been reported for bulk Zirconia.

In the case of pure and doped Ceria the situation is complicated by the possibility of mixed conduction. Ceria, in fact, is not generally a pure ionic conductor and might present some electronic conductivity as well, particularly for low doping contents. Since the GB cores in Ceria are positively charged they tend to attract the available free electrons within the SCR, resulting in a further enhancement of the electronic conductivity in nanometric materials, where the free energy required for the formation of free electrons is even lower than in the micrometric counterpart.

Since the dimension of the SCR depends on the dielectric constant of the material, ionic materials characterized by high values of dielectric constant present a particular interest, as they are characterized by wider SCRs. $SrTiO_3$ represents a typical example. This oxide, in effect, shows striking modification in its conduction mechanism even for grain size in the range of 50 nm [11]. $SrTiO_3$, in fact, is a semiconductor that presents a transition between *p* and *n* conductivity when the oxygen partial pressure is modified. In micrometric $SrTiO_3$ this transition falls at an oxygen partial pressure around 10^{-16} bar (Fig. 6). Samples prepared by SPS with grain size of ca. 30 nm, however, show a dramatic change in the position of this transition, which shifts more than 12 orders of magnitude, up to 10^{-4} bar. In respect to coarse-grained counterparts, nanometric samples showed also an oxygen vacancy conductivity (*p*-type) depressed of 6 orders of magnitude and an *n*-type conductivity increased of 3 orders of magnitude (Fig. 6) [11]. This behavior can be considered as one of the most striking examples of modification in the electrical properties due to the nanostructure in ionic functional materials.

Another phenomenon evidencing the influence of the nanostructure on the mechanism of conduction of functional materials is represented by the dramatic increase in the low-temperature conductivity manifested by oxides such as Yttria-Stabilized Zirconia (YSZ), Gadolinium- or Samarium-doped Ceria and Titania, at temperatures below 300 °C, in presence of a moist environment. These oxides present generally a high-temperature ionic conduction, due to the movement of oxygen. However, when produced in bulk nanometric form they present a low-temperature conductivity that can be associated to the movement of protons. None of these oxides, however, is generally recognized to be a typical protonic conductor, particularly in the low-temperature regime. Surprisingly, relatively high conductivity values have been observed also in almost fully dense materials [12]. The most remarkable feature, however, is represented by the strong dependence of this phenomenon from the grain size of the material: it becomes particularly evident for grain sizes below 50–80 nm and increases drastically when the grain size is further decreased. No significant dependence on the crystal structure and on the degree of doping of the material has been observed. Even the degree of porosity resulted to have a limited influence [12].

The typical characteristics of this phenomenon are summarized in Fig. 7 in the case of Titania. The figure reports the temperature dependence of conductivity for five different samples characterized by grain size between 24 and 56 nm and r.d. ranging between 75 and 95 vol.%, when exposed to a dry ($p_{H_2O} < 3$ ppm) or a moist environment. The solid line here indicates the ionic conductivity of Titania in a dry oxygen atmosphere. It appears evident as in nanocrystalline samples the low-temperature conductivity under humid atmosphere is increased of several orders of magnitude over the extrapolated high-temperature ionic conductivity. The figure also shows the strong dependence of conductivity on grain size: for temperatures $\leq$200 °C the conductivity increases more than 3 orders of magnitude, as the grain size is reduced from 56 to 24 nm.



Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 7 Influence of grain size and porosity on total conductivity of nanostructured bulk Titania. $P^{0}_{H_2O}$ is the equilibrium water partial pressure at a given T; $P^{ex}_{H_2O}$ is the water partial pressure used in in our experiments ($P^{ex}_{H_2O} = 32$ mbar) [12]

The transport mechanism responsible for this phenomenon has been object of some debate. In the earlier investigations the movement of protons along the GB of the material was considered as a possible conduction mechanism. This seemed to fit well with some of the observed characteristics, such as the very limited dependence on the sample porosity and the strong dependence on the grain size. Further subsequent investigations, however, have evidenced the presence of water molecules within the material, even in high-density samples. This suggested the possibility of a conduction mechanism based on hopping of protons within a percolating network of water molecules, located in the residual nanoporosity of the material [12]. This conclusion has been supported by several observations, particularly by the low value of the activation energy, which is incompatible with any kind of solid-state process, and by ^{1}H-NMR and DSC analysis that evidenced the presence of

free water molecules within the sample, confined in narrow pores. The presence of a percolating porosity seems to be in contrast with the high values of r.d. reported in most cases, but this result can probably be justified considering some characteristics of the nonconventional sintering methods used for the densification of the materials. This point is particularly interesting as it suggests the possibility that the morphology of these bulk materials obtained by low-temperature high-pressure sintering techniques might differ significantly from the one based on the classical sintering theory. Since the densification does not derive primarily from bulk or grain boundary diffusion but from plastic deformation, a network of open percolating nanoporosity might remain present ever for very high values of relative density, producing eventually unexpected behavior in the material. This point must be better clarified in the future, particularly in the case of materials that are sensitive to liquid- and gas-phase interactions [12].

Thermal Conductivity

Thermal conductivity is another physical property that is significantly modified by the nanostructure. This effect finds direct application in thermal protection coatings. However, wide attention has been given recently to the possibility to exploit the influence of nanostructure on thermal conductivity to improve the efficiency of thermoelectric materials. Thermoelectric devices allow a direct conversion of stable temperature gradients in electric power. In order to achieve the maximum efficiency, thermoelectric materials must present a high electrical conductivity and a low thermal conductivity, to guarantee the maintenance of the thermal gradient. The efficiency of these materials is generally characterized on the basis of a figure of merit, defined as

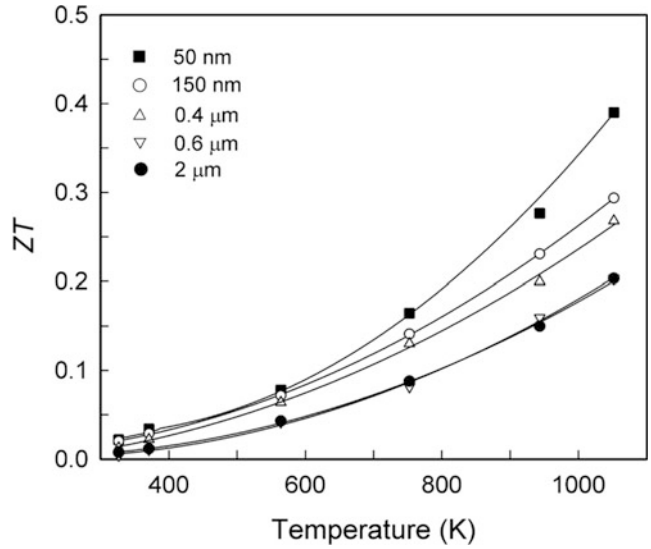
$$ZT = \frac{\alpha^2 \sigma}{k} T$$

where T indicates the temperature, α the Seebeck coefficient, σ the electrical conductivity, and k the thermal conductivity. The term $\alpha^2\sigma$ is called the power factor, PF. It is evident from this relation as an increase in the figure of merit can be achieved not only through an increase in the intrinsic thermoelectric power of the material, α, but also through a decrease of the thermal conductivity. Nanostructure is known to produce a decrease in electrical conductivity, but the decrease of thermal conductivity is expected to be even higher, resulting in an overall increase of ZT.

According to the kinetic theory of heat transfer, the thermal conductivity is proportional to the product of specific heat, sound velocity, and phonon mean free path. While nanostructure has not a relevant effect on the first two quantities, it greatly affects the phonon mean free path, which is shortened when the grain size is reduced. The mean free path is also heavily dependent on the density of lattice defects (foreign atoms, cation and anion vacancies, etc.). As a result, the reduction of k in the nanosized samples is attributed to the combined effect of high concentration of GBs and to the presence of nanopores. The phonon scattering produced by the GB introduces an intrinsic interfacial resistance to heat transfer, generally known as Kapitza resistance. When the grain size is reduced to the low nanometric range this contribution is expected to become particularly relevant, opening the possibility to obtain interesting ZT values even in materials characterized by a modest thermoelectric effect [13, 14].

Common thermoelectric materials are usually represented by semiconducting alloys, based on rare and toxic elements, such as Lead, Cadmium, Tellurium, Antimony, and Germanium, and are generally characterized by fairly high values of thermoelectric coefficient. To allow a widespread application of the thermoelectric conversion, however, other classes of material have been recently investigated. Metal oxides represent a promising alternative, thanks to their high thermal and chemical stability, low cost, and low environmental impact. Their Seebeck coefficient, however, is usually low. For these materials a reduction in the grain size in the low nanometric range would be particularly beneficial.

Most of the nanostructured bulk ceramic oxides investigated for their thermoelectric



Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 8 Temperature dependence of *ZT* for $In_{1.92}(ZnCe)_{0.08}O_3$ bulk samples with different grain sizes (Adapted from Ref. [15])

properties have been realized using FAST techniques, but only in few cases it was possible to obtain grain sizes below 50 nm. In samples of Ga-doped ZnO with grain size reduced to 30 nm, thermal conductivity is indeed reduced, but also electrical conductivity is decreased significantly and the resulting *ZT* is not improved over the analog micrograined material. On the opposite, nanostructuring of In_2O_3, doped with Ce and Zn, shows a noticeable increase in *ZT*, in respect to coarse-grained counterpart. Samples with grain size of 50 nm, obtained by SPS, have a *ZT* of 0.40, i.e., twice the value of the corresponding micrometric material (Fig. 8) [15]. Anatase TiO_2 shows a decrease of 10 % of thermal conductivity when the grain size is reduced to 20 nm. $SrTiO_3$ represents another good candidate for thermoelectric applications, since at room temperature it shows a high PF value. When its grain size is reduced to 50 nm, thermal conductivity at r.t. is halved, in respect to coarse-grained samples, while at 1000 K the reduction is around 24 %. La-doped $SrTiO_3$ shows an even more relevant effect; when the grain size is reduced to 24 nm, *k* is in fact reduced of one order of magnitude in respect to single-crystal analogues, even though electrical conductivity is depressed as well.

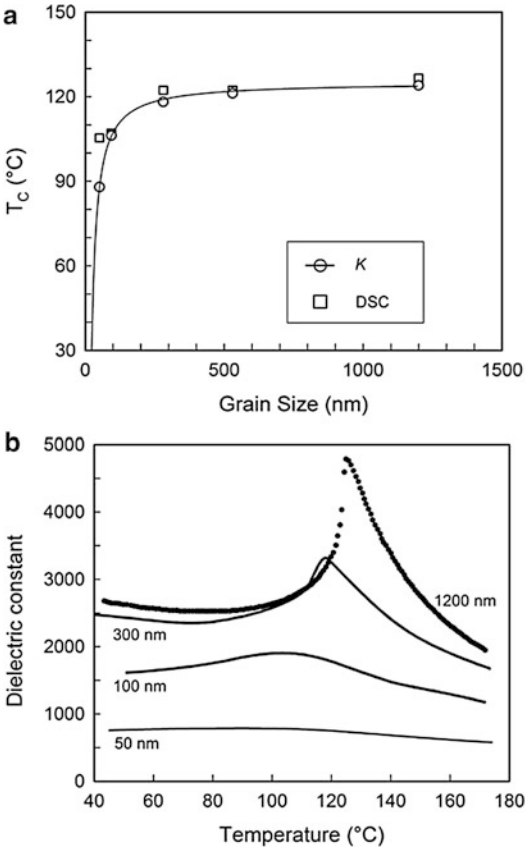
A further decrease of grain size, possibly down to 10 nm, seems to be necessary for a real breakthrough in the reduction of the thermal conductivity. Moreover, a better control over defects and density of scattering sites should help in improving the achievable efficiency even further. However, the results so far obtained confirm that a reduction of the thermal conductivity can indeed be realized in bulk nanometric thermoelectric materials.

Ferroelectric Nanoceramics

Ferroelectric ceramics, such as $BaTiO_3$ and Pb(Zr, Ti)O_3, are largely used in the electronic industry as a key component of ferroelectric memories, multilayer ceramic capacitors (MLCCs), pyroelectric systems, and piezoelectric transducers/actuators. Due to the continuous trend toward the miniaturization of these components, the interest toward the influence of the nanostructure on the ferroelectric properties of these materials has grown considerably. In the case of barium titanate a loss of the ferroelectric properties in extremely fine-grained materials has been repeatedly reported, although the threshold for this phenomenon was difficult to define. In order to clarify this point, bulk nanometric titanate ceramics with extremely small grain size have been produced and investigated. These materials have been generally prepared using high-pressure sintering, as small amounts of residual porosity

produce a decrease of the relative dielectric constant [11, 16].

The loss in the ferroelectric characteristics is strictly associated, in these materials, with the modification in the phase stability produced by the nanostructure. Ferroelectricity in $BaTiO_3$ is produced by the displacement of the Ti ion in its tetragonal structure. The cubic polymorph, on the other hand, does not present ferroelectricity. The transition between the two polymorphs takes place at the Curie temperature, T_c, which falls at 403 K in the single crystal. As mentioned previously, the nanostructure can modify the relative stability of the polymorphs. In the case of $BaTiO_3$, this results in the stabilization of the cubic form, with a consequent decrease of T_c with the grain size (Fig. 9) [16]. The complete stabilization is predicted for particles characterized by a dimension between few nanometers and few tens of nanometers. A further effect is associated to the increase in the GB density, as the regions in proximity of the GB are considered not ferroelectric. Despite these theoretical previsions, a clear and unequivocal threshold for the disappearance of the ferroelectricy has not been identified. Several authors, in fact, have reported the persistence of ferroelectricity, at least at local scale, even in materials characterized by a size reduced to 8 nm [1]. This behavior can probably be explained considering that in the case of extremely small grain sizes the properties of the material tend to depend strongly not only on the crystallographic structure but also on other features, such as the point defect concentration and distribution, incorporation of foreign atoms and hydroxyl groups, aggregation, porosity, residual stresses, etc.



Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 9 Ferroelectric $BaTiO_3$ nanoceramics. (**a**) Dependence of T_C from grain size. Cyrcles: dielectric measurements; squares: DSC measurements. The *solid curves* are a guide for the eyes. (**b**) Relative dielectric constant at 10^4 Hz of $BaTiO_3$ ceramics with different grain sizes as a function of temperature (Adapted from Ref. [16])

Magnetic Properties

Ferrites represent another group of functional oxides presenting important technological applications [17]. They are in fact largely applied in microwave and radio technology and in computer data storage. New fields of applications are also emerging in the area of biomedical technologies and of thermochemical hydrogen production through water splitting. The general formula of spinel ferrites is AB_2O_4, where A is a divalent metal cation (e.g., Nickel, Cadmium, Zinc, Magnesium, Copper, etc.) and B is Fe. The structural, electrical, and magnetic properties of these materials strongly depend not only on their stoichiometry and on the cationic distributions between the available (A) and (B) sites but also on their microstructure, particle size, and surface-to-volume ratio. As a result, nanostructured materials present significant modification in their magnetic properties. They show, in fact, unusual saturation magnetization and high-coercivity field. Ferrites that are magnetically ordered may also exhibit a

superparamagnetic behavior when in nanocrystalline form. However, only a very limited number of nanostructured bulk materials has been described in the literature. Nanometric $CoFe_2O_4$ ferrite characterized by a grain size of 10 nm and prepared by SPS has been shown to present promising mechanical and magnetic properties [18]. Other examples are represented by $Zn_{0.5}$,$Ni_{0.5}Fe_2O_4$, characterized by a grain size of 60 nm, and by a complex $Ni_{0.4}Zn_{0.4}Co_{0.2}Cu_{0.02}Fe_{1.98}O_4$, presenting a grain size between 38 and 54 nm.

Optical Transparency

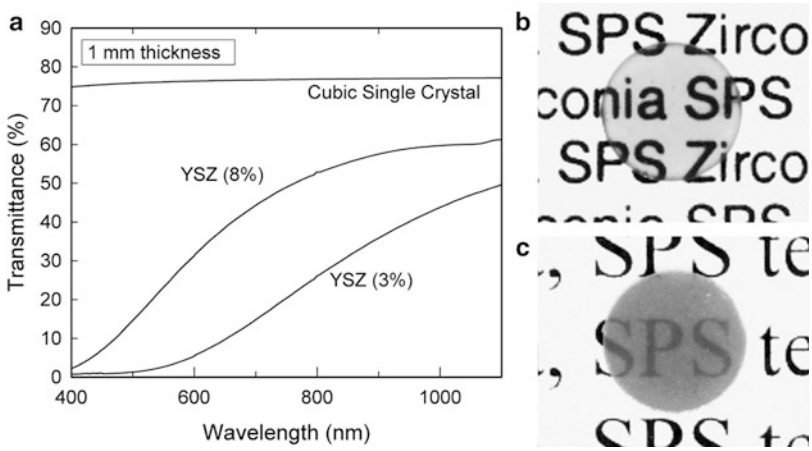
Nanostructure has a strong influence on the optical properties of materials. Some of the most evident optical effects are related to confinement, such as in quantum dots or in plasmonic resonators. These effects are well known and are well described elsewhere, but nanostructure might produce other modifications. Particularly interesting in this regard is the effect of nanostructure on the optical transparency of ceramic materials.

Applications requiring optical transparency in bulk materials usually employ single crystals of oxides, such as sapphire. But single crystals are difficult to make and can be produced only in relatively small dimensions. Polycrystalline ceramics represent an appealing alternative, thanks to their lower costs and improved mechanical properties. The degree of transparency of a ceramic polycrystalline material depends on two main parameters: residual porosity and GB density. The former should be kept in the range between 10^{-2} and 10^{-3} vol.% or less, while the latter is usually kept as low as possible, especially when the material presents an anisotropic refractive index. The scattering efficiency of spherical pores, however, decreases drastically when their dimension becomes smaller than the radiation wavelength [19]. Consequently, it has been suggested that materials presenting residual porosity in the low nanometric range should be transparent, even in presence of a significant amount of residual nanoporosity.

Usually low sintering temperatures and very high pressures, sometimes exceeding 1 GPa, have been employed, in order to obtain transparent materials characterized by a uniform nanoporosity and by a grain size in the low nanometric range [1]. The spinel $MgAl_2O_4$ was successfully sintered, reaching optical transparency, under pressures between 2 and 5 GPa. At 4 GPa, for sintering temperatures as low as 600 °C, the spinel showed a high transmittance level (>70 %) for a large wavelength range (600–2000 nm). The sample had a grain size of 40 nm. The same material, with a grain size reduced to 30 nm, was prepared by a two-step sintering at higher temperatures (795 °C), reaching a higher level of optical transmission (80 % for 1000–2000 nm and 70 % for 375–2500 nm) [20]. An even smaller grain size (ca. 25 nm) was obtained by increasing the pressure up to 5 GPa and with no need of heating: the ambient temperature densification provided a material that was transparent for IR and visible radiation. Transparent Y_2O_3 doped with Eu was obtained under even higher pressures, up to 8 GPa, at relatively low temperatures (200 °C), reaching 60 % of transmission in respect to the analogous single crystal [1]. Other nanocrystalline bulk materials showing a good level of transmission were $Lu_{2-x}Ho_xO_3$, $Y_3Al_5O_{12}$, and $Y_3Fe_5O_{12}$ [1]. To obtain the required level of nanoporosity at lower sintering pressures, the corresponding temperatures applied during the densification process need to be remarkably increased: only at temperatures as high as 1000 °C sufficiently transparent YSZ samples were produced by HP-FAST. The sintering cycles however were very short, reduced to only 5 min. Samples 1 mm thick, characterized by a transmittance between 70 % and 80 % of their maximum theoretical value, have been obtained (Fig. 10) [20].

Future Direction for Research

The investigation on possible use of the nanostructure in controlling the physical properties of functional materials in bulk form is still in its early stage of development. Only recently it has been possible to produce these materials in amounts



Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm, Fig. 10 (**a**) Optical spectra of two samples of 8 mol% YSZ and 3 mol% YSZ, sintered at ~1000 °C under a pressure of ~800 MPa, for 5 min. The theoretical maximum transmittance of a sample of cubic Zirconia single crystal is added as a reference. (**b**) Optical micrograph of the sample of 1 mm thick 8 mol% YSZ. (**c**) Optical micrograph of the sample of 1 mm thick 3 mol% YSZ (Adapted from Ref. [20])

large enough to allow a systematic study of their characteristics. In the area of ceramic functional materials, particularly oxides, high-pressure spark plasma sintering and other high-pressure field assisted sintering techniques have emerged as the only approaches allowing a fairly simple and reproducible way for their synthesis. However, the microscopic mechanisms involved in the densification of the nanopowders using these techniques are still poorly understood, and the determination of the experimental conditions required for each system is still largely based on an empirical approach. Furthermore, very limited information is generally available on the details of the obtained nanostructures, as the characterization of bulk ceramic materials with such a small grain size is quite problematic. Despite all this, it has been possible to evidence as the nanostructure represents a powerful tool for the modification of the bulk characteristics of functional materials. It has been demonstrated, in fact, that through a careful control of the nanostructure, phase stability, electrical, magnetic and optical properties of bulk materials can be altered significantly. In this regard, the nanostructure is emerging as an alternative to the traditional methods, represented by alloying with doping elements and control of point defect equilibria, for determining the bulk properties of functional materials. The traditional methods present limitations generally due to thermodynamic constraints. The use of nanostructured materials can in some cases overcome these limitations, as the prevalence of interfacial processes might overshadow the properties of the bulk. Of course, since the nanomaterials characterized by extreme nanostructures are generally thermodynamically unstable, the stability of the produced materials in the long term represents a point of concern.

Cross-References

► Nanostructured Thermoelectric Materials

References

1. Maglia, F., Tredici, I.G., Anselmi-Tamburini, U.: Densification and properties of bulk nanocrystalline functional ceramics with grain size below 50 nm. J. Eur. Ceram. Soc. **33**, 1045–1066 (2013) and references therein
2. Anselmi-Tamburini, U., Maglia, F., Tredici, I.G.: Field assisted sintering of nanometric ceramic materials. Ceram. Trans. **232**, 133–149 (2012)
3. Schoonman, J.: Nanoionics. Solid State Ion. **157**, 319–326 (2003)

4. Maier, J.: Nanoionics: trivial and non-trivial size effects on ion conduction in solids. Z. Phys. Chem. **217**, 415–436 (2003)
5. Wu, X.L., Fan, J.Y., Qiu, T., Yang, X., Siu, G.G., Chu, P.K.: Experimental evidence for the quantum confinement effect in 3C-SiC nanocrystallites. Phys. Rev. Lett. **94**, 026102 (2005)
6. Chadwick, A.V., Savin, S.L.P.: Structure and dynamics in nanoionic materials. Solid State Ion. **177**, 3001–3008 (2006)
7. Andrievskii, R.A., Khachoyan, A.V.: Role of size-dependent effects and interfaces in physicochemical properties of consolidated nanomaterials. Russ. J. Gen. Chem. **80**, 555–566 (2010)
8. Garvie, R.C., Goss, M.F.: Intrinsic size dependence of the phase transformation temperature in zirconia microcrystals. J. Mater. Sci. **21**, 1253–1257 (1986)
9. Sata, N., Eberman, K., Eberl, K., Maier, J.: Mesoscopic fast ion conduction in nanometre-scale planar heterostructures. Nature **408**, 946–949 (2000)
10. Guo, X.: Can we achieve significantly higher ionic conductivity in nanostructured zirconia? Scripta Mater. **65**, 96–101 (2011)
11. Lupetin, P., Gregori, G., Maier, J.: Mesoscopic charge carriers chemistry in nanocrystalline $SrTiO_3$. Angew. Chem. Int. Ed. **49**, 10123–10126 (2010)
12. Tredici, I.G., Maglia, F., Ferrara, C., Mustarelli, P., Anselmi-Tamburini, U.: Mechanism of low-temperature protonic conductivity in bulk, high-density, nanometric titanium oxide. Adv. Funct. Mater. **24**, 5137–5146 (2014)
13. Yang, H.-Y., Bai, G.-R., Thompson, L.Y., Eastman, J. A.: Interfacial thermal resistance in nanocrystalline yttria-stabilized zirconia. Acta Mater. **50**, 2309–2317 (2002)
14. Dresselhaus, M.S., Chen, G., Tang, M.Y., Yang, R., Lee, H., Wang, D.Z., Ren, Z.F., Fleurial, J.-P., Gogna, P.: New directions for low-dimensional thermoelectric materials. Adv. Mater. **19**, 1043–1053 (2007)
15. Lan, J., Lin, Y.-H., Liu, Y., Xu, S., Nan, C.-W.: High thermoelectric performance of nanostructured In_2O_3-based ceramics. J. Am. Ceram. Soc. **95**, 2465–2469 (2012)
16. Zhao, Z., Buscaglia, V., Viviani, M., Buscaglia, M.T., Mitoseriu, L., Testino, A., Nygren, M., Johnsson, M., Nanni, P.: Grain-size effects on the ferroelectric behavior of dense nanocrystalline $BaTiO_3$ ceramics. Phys. Rev. B **70**, 024107 (2004)
17. Valenzuela, R.: Novel applications of ferrites. Phys. Res. Int. **2012**(591839) (2012)
18. Imine, S., Schoenstein, F., Mercone, S., Zaghrioui, M., Bettahar, N., Jouini, N.: Bottom-up and new compaction processes: a way to tunable properties of nanostructured cobalt ferrite ceramics. J. Eur. Ceram. Soc. **31**, 2943–2955 (2011)
19. Anselmi-Tamburini, U., Woolman, J.N., Munir, Z.A.: Transparent nanometric cubic and tetragonal zirconia obtained by high-pressure pulsed electric current sintering. Adv. Funct. Mater. **17**, 3267–3273 (2007)
20. Wollmershauser, J.A., Feigelson, B.N., Gorzkowski, E.P., Ellis, C.T., Goswami, R., Qadri, S.B., Tischler, J. G., Kub, F.J., Everett, R.K.: An extended hardness limit in bulk nanoceramics. Acta Mater. **69**, 9–16 (2014)

Nanocrystalline Materials

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
▶ Nanosized Nanocrystalline and Nanotwinned Metals

Nanodentistry

Hans Deyhle, Simone Hieber and Bert Müller
Biomaterials Science Center (BMC), University of Basel, University Hospital Basel, Basel, Switzerland

Synonyms

Nanotechnology in dental medicine

Definition

Nanodentistry is defined as the science and technology of diagnosing, treating, and preventing oral and dental diseases, relieving pain, and of preserving and improving dental health, applying materials structured on the nanometer scale [1]. The nanotechnology considers nanostructures between 1 and 100 nm exhibiting properties and functionalities that fundamentally differ from that of other length scales, as the surface of the nanostructures dominates the material properties usually given by the bulk. The nanomaterials are not only promised to improve the properties and functionalities of dental products but also to lead to the development of innovative, novel products for the benefit of patients. In particular,

nanomaterials have been part of engineered products, such as implants, dental and surgical instruments, dental care wares, as well as diagnostics tools.

Overview

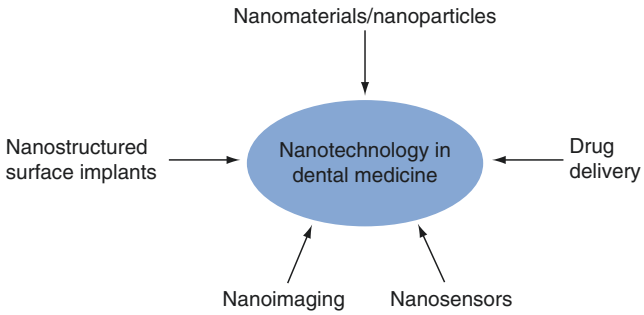
The cover story of R. A. Freitas Jr. in the *Journal of the American Dental Association* more than a decade ago [2] introduced the term *nanodentistry* to a larger community. He developed visions effectively using dental nanorobots for orthodontic realignments, for dentition regeneration procedures and, finally, oral health maintenance. He already elucidated the increasing influence of nanomaterials and the importance of tissue engineering. It has been pointed out that properly configured dentifrobots will recognize and subsequently destroy pathogenic bacteria residing in the plaque and elsewhere.

Although most of Freitas' ideas are still science fiction, today many applications of nanometer-scale components are known in the field of dentistry, which include nanoparticles in sensitive toothpastes, nanostructured surfaces of dental implants to improve osseointegration, and to reduce inflammatory reactions of different kinds of drugs (cp. Fig. 1). Further applications still on the research stage are classified into nanoimaging to reveal the nanometer-sized features in human tissues and man-made implants or nanosensors to quantify physical, chemical, or biological parameters within the oral cavity.

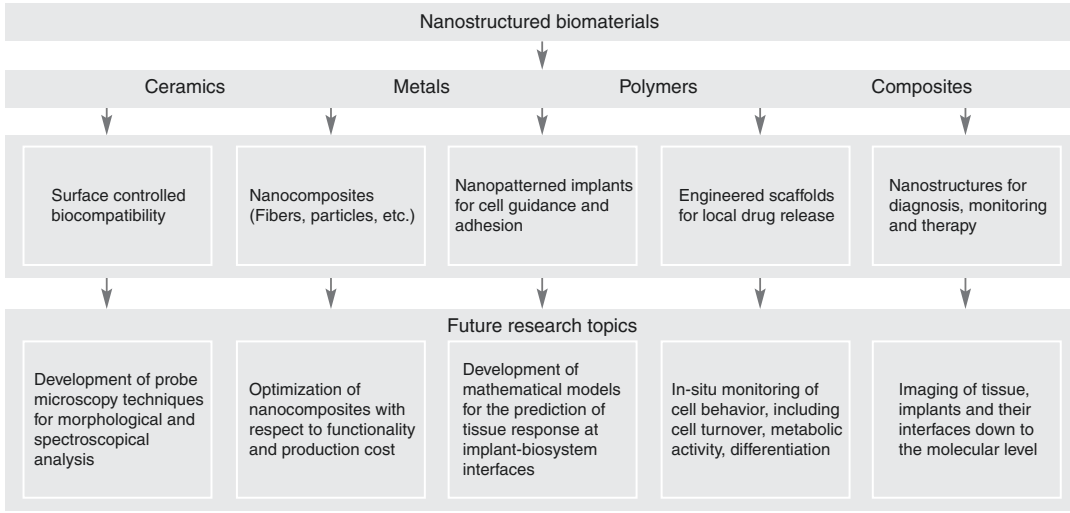
Dental materials are ceramics, metals, and polymers or any combination (cp. Fig. 2). Applying the knowledge and experience of nanotechnology to solve problems in the field of dental medicine, nanometer-sized patterns have been created on surfaces of implants and surgical instruments and incorporated into the bulk of the dental materials to accomplish the dedicated functionality. Nanostructures can improve the biocompatibility of dental implants or carry drugs toward the target within the oral cavity. Reactive nanostructures are going to emerge in diagnosis, monitoring, and therapy to enhance the patient treatments. Diverse activities augment the basic research of bio-nano-materials. Scanning probe microscopy techniques, for instance, enable us to morphologically and spectroscopically analyze hard and soft tissues to understand the structure–-function relationship on the nanometer scale. Sophisticated mathematical simulations allow tailoring the material properties by the combination of nanometer-sized components. The outcome also predicts the tissue response and the self-organization at the implant–biosystem interfaces down to the molecular scale. Highly advanced in vivo measurements support cell biologists to determine the cell behavior with respect to proliferation, differentiation, and metabolism, especially at the interface between the dental material and the surrounding tissues. For this target, high-resolution imaging facilities have become more and more vital to quantify the different kinds of nanostructures with the desired precision.

Biocompatibility

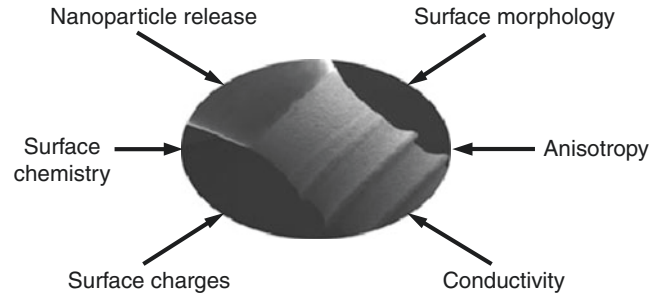
Biocompatibility cannot be simply described as a specific property of the selected biomaterial, as well known from mechanical quantities such as



Nanodentistry, Fig. 1 Key issues currently related to nanotechnology in dental medicine



Nanodentistry, Fig. 2 Materials science for nanodentistry



Nanodentistry, Fig. 3 Several parameters allow tailoring the biocompatibility of dental implants

Young's modulus. It is a peculiar term and defined as ". . . the ability of a material to perform with an appropriate host response in a specific application." [3]. This means it not only depends on the material itself but also from the location within the human body where the material is to be placed. Consequently, biocompatibility is not only related to intrinsic material properties. Its classification relies on standards (mainly ISO 10993 [4]). Therefore, the regular updates of these standards may change the status of a material concerning biocompatibility. Natural scientists and engineers are often not familiar with such a practice and, hence, usually terminate their work on such research areas.

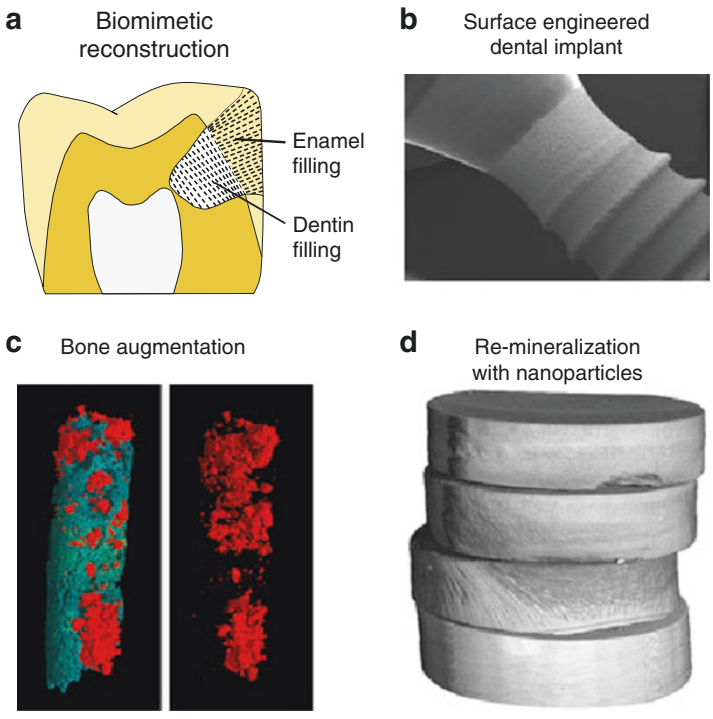
The encouraging argument, however, correlates to the experimental observations that biocompatibility of implants can be intentionally manipulated to optimize the clinical outcome. Figure 3 illustrates how some selected phenomena can influence the biocompatibility. For a dental titanium implant, for example, the surface is made rough down to the molecular level by sandblasting and etching procedures to ensure osseointegration and to reduce the inflammatory reactions [5]. Surface morphology can offer certain angles for dedicated protein absorption and activity. In addition, the surface chemistry is commonly tailored concerning oxide thickness, stoichiometry, and normally functionalized to obtain a hydrophilic surface. Further relevant factors are particle and ion release, surface charges, electrical conductivity, and anisotropies as present on patterned surface micro- and nanostructures. The measurement of the surface roughness on real implants is more than challenging, since there is

no single parameter and no single experimental setup, which allow simultaneously determining roughness on the micro- and nanometer scales.

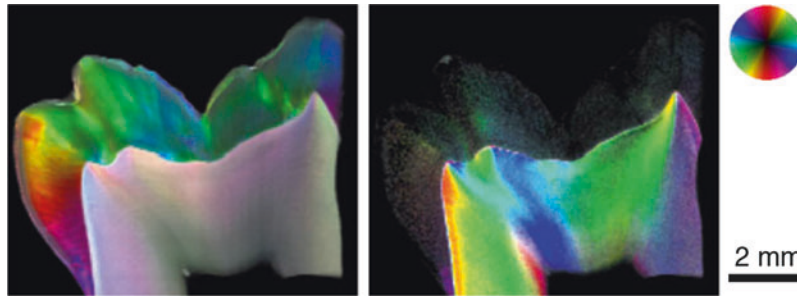
Key Research Findings

Figure 4 shows representative applications of nanotechnology in dental medicine. The fillers of reconstruction materials, the surfaces of dental implants, the different kinds of bone-augmentation ceramics and glass ceramics, as well as the re-mineralization of hard tissues by means of nanoparticles exploit nanotechnology by now and possess enormous growing potential. Nature-analogue, anisotropic restorations will replace today's dental filling and inlay materials as well as artificial teeth [6]. Researchers are still searching how the anisotropy can be introduced to obtain nanostructures similar to the ones present in healthy human teeth.

It belongs to the key impacts to realize that the ceramic components in dentin are orthogonally oriented to the nanostructures of the same size in enamel (see Fig. 5). This finding well observable in scanning small-angle X-ray scattering could explain why the dentin-enamel junction acts as an effective crack barrier. Small-angle X-ray scattering in scanning mode, as performed at the cSAXS beamline (Swiss Light Source, Paul Scherrer Institut, Switzerland) allows uncovering the mean orientation, abundance, and anisotropy of nanometer-sized components in human tissues over macroscopic areas [7]. It is in general difficult to isolate the contribution of specific components to the scattering signal, see entry "► Imaging the Human Body Down to the Molecular Level." The collagen in human dentin, however, yields a characteristic signal related to its 67 nm-wide periodicity. Due to the sensitivity of scattering data to periodic structures, it is possible to extract a signal unequivocally related to collagen, thus allowing gathering component-specific information. Several 400 μm-thin slices from a human third molar were examined. Figure 5 shows the processed scattering signal from the features with sizes between 60 and 70 nm. The mean orientation of the scattering signal is coded according to the color wheel. The brightness relates to the abundance of nanostructures, while



Nanodentistry, Fig. 4 Applications of nanotechnology in dentistry with growing potential



Nanodentistry, Fig. 5 Processed small-angle X-ray scattering data in the range between 60 and 70 nm. The mean orientation of the scattering signal is according to the color wheel. The brightness relates to the nanoparticle abundance, while color saturation codes their degree of anisotropy

color saturation matches the degree of the anisotropy. The orientation of the scattering signal in both enamel and dentin is perpendicular to the crystallite orientation. On the left-hand side, the information from the total scattering signal is shown, which is mainly produced by the ceramic components of the tooth. Enamel and dentin can clearly be distinguished due to their different scattering potential. The strong color in the enamel is related to the high degree of anisotropy, that is, the high degree of orientation of the calcium phosphate crystallites, and their strong scattering potential in this range. In the dentin, the anisotropy is less distinct, and decreases with increasing distance from the dentin-enamel junction (DEJ). Moreover, a perpendicular orientation of the scattering signal, and thus of the nanostructures, between enamel and dentin is observed. While the crystallites are mainly oriented perpendicular to the DEJ in enamel, parallel to the main direction of mechanical load, the dentin structures appear parallel to the DEJ, perpendicular to the mechanical loading. It is well known that the DEJ acts as an effective crack barrier, impeding the propagation of cracks from the enamel into the dentin. The image on the right in Fig. 5 shows only the signal related to the collagen matrix. Due to the distinctive periodicity of 67 nm along the collagen fibrils, the scattering signal is oriented parallel to the fibrils. No significant signal is found in the enamel. The mean scattering orientation of the collagen is found to be mainly perpendicular to that of the total signal on the left-hand image, revealing a parallel arrangement of collagen fibrils with the calcium phosphate crystallites. The collagen abundance is maximal near the DEJ and decreases toward the pulp. Abrupt changes in collagen fibril orientation are found along lines connecting the tooth cusps and the pulp. In this manner, the structure of the human teeth along the entire nanometer range is evaluated.

As pointed out above, the part of dental implants to be fixed in the bone is made rough. It contains nanostructures which resemble the apatite crystallites in human bone and therefore offer nanostructures with angles for suitable protein absorption. The suppliers have their secrets for the surface engineering processing, which includes sandblasting, etching, and electrochemical functionalization. Surface-engineered dental implants are very well established and do have a high success rate. During the next decade, we can expect a significant reduction in price so that a broader community of our aging society can benefit from such surface-engineered implants with dedicated nanostructures.

The calcium phosphate phases for bone augmentation gain more and more importance in our aging population. The resorbable calcium phosphate ceramics or bio-glasses do actively support the bone formation. These bone substitutes are combined with autologous bony tissues to repair larger and larger defects arising from cancer surgery, trauma, or just tooth loss.

There are numerous reasons for the de-mineralization of teeth. The mineral loss,

especially owing to attacks of caries bacteria, has a massive effect on lifetime and function of teeth. Therefore, strategies have been developed to at least partially re-mineralize the enamel and dentin as the robust alternative to mechanical removal and conventional filling the defect. Optimized dental hygiene has reached restricted success in small caries lesions. Researchers all over the world use slices of dentin to explore the re-mineralization capacity of a variety of nanotechnology-based ceramics particles. Using a set of slices, a direct comparison of the different materials and procedures has become possible.

Future Research

The dental nano-biomaterials have to be investigated to judge the toxicity and the biocompatibility and to pursue the personalized medicine with respect to more and more frequent material incompatibilities. The development of nanocontainers for a variety of drugs aims for targeted delivery preventing undesired side effects. For larger caries regions, which grow fast into the dentin, the collagen network might be still available, but the mineral phase is already dissolved. The return of the calcium phosphates and traces of other minerals into the affected region for restoration purposes is challenging but not impossible, as the collagen network could provide nucleation sites for the re-mineralization.

Bio-inspired dental fillings and inlays have to exhibit the anisotropy and orientation of organic and inorganic nanometer-sized components as found in enamel and dentin. Today, the dentists already use different fillings for dentin and enamel but without oriented anisotropic ceramics. Therefore, scientists and engineers have to discover opportunities to incorporate anisotropic nanometer-sized ceramics into the fillers and to carry them into the desired direction. Hence nature-analogue mechanical properties should be obtained. Based on these oriented nanostructures, one should search for self-repairing materials. Here again the human body provides concepts: the remodeling of bone.

Cross-References

- ▶ AFM in Liquids
- ▶ Atomic Force Microscopy
- ▶ Basic MEMS Actuators
- ▶ Bioadhesion
- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Biomimetics
- ▶ Biosensors
- ▶ Chitosan Nanoparticles
- ▶ Confocal Laser Scanning Microscopy
- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ Imaging the Human Body Down to the Molecular Level
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Microfabricated Probe Technology
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Nanostructured Functionalized Surfaces
- ▶ Scanning Electron Microscopy
- ▶ Selected Synchrotron Radiation Techniques
- ▶ Self-Repairing Materials

References

1. Dosch, H., Van de Voorde, M.H. (eds.): Gennesys White Paper: A New European Partnership Between Nanomaterials Science & Nanotechnology and Synchrotron Radiation and Neutron Facilities. Max-Planck-Institut für Metallforschung, Stuttgart (2009)
2. Freitas, R.A.: Nanodentistry. J. Am. Dent. Assoc. **131**, 1559–1565 (2000)
3. Williams, D.F.: Williams Dictionary of Biomaterials. Liverpool University Press, Liverpool (1999)
4. ISO 10993: Biological Evaluation of Medical Devices. Institute ANS. www.iso.org (2009)
5. Müller, B.: Natural formation of nanostructures: from fundamentals in metal heteroepitaxy to applications in optics and biomaterials sciences. Surf. Rev. Lett. **8**, 169–228 (2001)
6. Bar-Cohen, Y. (ed.): Biomimetics: Nature-Based Innovation. CRC Press/Taylor & Francis Group, Abingdon (2011)
7. Müller, B., Deyhle, H., Bradley, D., Farquharson, M., Schulz, G., Müller-Gerbl, M., et al.: Scanning x-ray scattering: evaluating the nanostructure of human tissues. Eur. J. Clin. Nanomed. **3**, 30–33 (2010)

Nanodevices

▶ Tight-Binding Simulations of Nanowires

Nano-ecotoxicology

▶ Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates
▶ Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment

Nano-effects

▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity

Nanoelectrodes

▶ Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes

Nanoelectromechanical Systems

▶ Nonlinear and Parametric NEMS Resonators

Nanoelectromechanical Systems (NEMS)

▶ Nanomechanical Resonators Based on III–V Semiconductors

Nanoencapsulation

Mary Cano-Sarabia[1] and Daniel Maspoch[1,2]
[1]ICN2 – Institut Catala de Nanociencia i Nanotecnologia, Bellaterra, Barcelona, Spain
[2]Institució Catalana de Recerca i Estudis Avançats (ICREA), Bellaterra, Barcelona, Spain

Synonyms

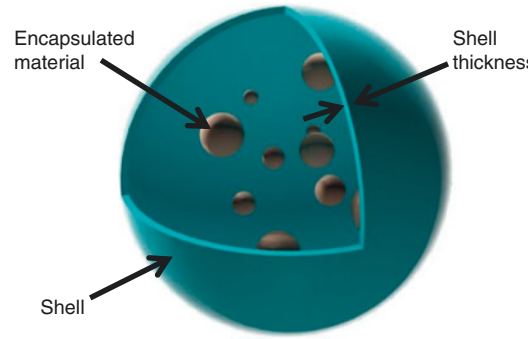
Molecular encapsulation

Definition

Nanoencapsulation is defined as the entrapping of active ingredients in nanometer sized capsules.

Overview

Nanoencapsulation is defined as the technology of packaging nanoparticles of solid, liquid, or gas, also known as the core or active, within a secondary material, named as the matrix or shell, to form nanocapsules (see Fig. 1) [1]. The core contains the active ingredient (e.g., drugs, perfumes, biocides, vitamins, etc., see Table 1) while the shell isolates and protects the core from the surrounding environment. This protection can be



Nanoencapsulation, Fig. 1 Scheme of a nanocapsule

Nanoencapsulation, Table 1 Core and shell materials more commonly used in different nanoencapsulation applications

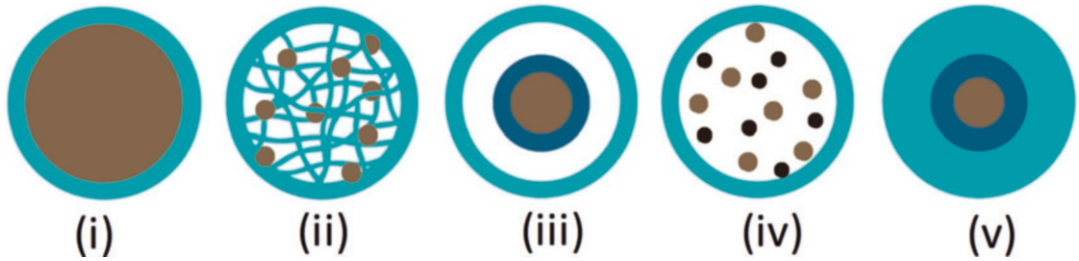
Application	Materials
Food	C: acidulants (acetic, sorbic, lactic), flavoring agents (citrus/mint oils, oleoresin, oregan, menthol), sweeteners (sugar, aspartame), colorants (β-carotene), lipids (fish oil, linoleic oil, palmitic acid), vitamins and minerals (vitamin A, D, E and K, folic acid), salts, preservatives (gallic acid), antimicrobials and antioxidants (terpene, limonene, black pepper, catechins), probiotics (lactobacillus, bifidobacteria) enzymes and microorganisms (lipase, invertase, penicillium roqueforti) S: carbohydrates (sucrose, malto- and cyclo-dextrines, chitosan), gums (agar, arabic, gum acacia, sodium alginate, lipids (oils, paraffin, stearic acid, fats, beeswax, phospholipids, Stealth® liposomes), celluloses, proteins (albumin, casein, gelatin, gluten, peptides, zein), synthethic elatomers (polyacrylamide, polyacrylate, polyethylene, polyvinyl alcohol, polyvinyl acetate), synthetic polymers (acrylonitrile, polybutadiene, poly(lactide-co-glycolide), silicon dioxide
Pharmaceutics	C: hormones (insulin), vaccines, oncologic (doxorubicin, cisplatin) and hematology actives (anthracyclines), proteins (bovine serum albumin, globulin, myoglobin,), enzymes (glucose oxidase), antibiotics (amoxicillin), drugs (felodipine, nicardipine, heparin), bacteria (lactobacillus acidophilus) and hepatocytes S: carbohydrates (chitosan, alginate, dextran), proteins (polylysine, polyglutamic, albumin, peptides), synthethic elatomers (poly(alkyl acrylate), poly(ethyleneglycol)), celluloses (ethylcellulose), synthetic polymers (poly(lactide-co-glycolide, polyethyleneglycol), lipids (phospholipids), carbon nanotubes, erythrocytes, gums (alginate), silicon dioxide, phyllosilicate
Agriculture	C: Insecticides (endosulfan), herbicides (fenamiphos), pheromones, fertilizers, pesticides, microbicides S: Polyurea, polyurethane, gelatin, gum Arabic, chitin, chitosan
Textile	C: PCM, emollients (aloe vera, vitamin E, lanolin), dyes, fabrics softeners, flame retardants S: Poly(styrene), urea-formaldehyde, ethylcellulose
Paintings	C: Biocide (OIT), pigments, anti-foulant agents, anti-corrosion agents S: Melamine formaldehyde, gelatine – acacia, silica, zeolite
Personal care	C: natural ingredients, essential oils (menthol, lime, lavender), preservatives, vitamins (A, E, C), retinol and triclosan S: carbohydrates (sucrose, malto- and cyclo-dextrines, chitosan), cellulose, silicon dioxide

C core, *S* shell

permanent or temporal, in which case the core is generally released by diffusion or in response to a trigger, such as shear, pH, or enzyme action, thus enabling their controlled and timed delivery to a targeted site [2, 3].

Nanocapsules may range from 1 to 1000 nm in size and they have a multitude of different shapes, depending on the materials and methods used to prepare them [2]. The structure of encapsulated ingredients (see Table 1) which largely depend on the selected shell material and nanoencapsulation method can be classified into two main categories (see Fig. 2): capsules with (a) a core that is surrounded by a shell of the matrix material or (b) a core that is entrapped within a continuous network of the matrix material. Variations of these morphologies include capsules with multiple cores or multilayered capsules [1]. However, the most significant feature of nanocapsules is their nanoscopic size that provides a large surface area. The total surface area is inversely proportional to the capsule diameter. This large surface area is appropriate for incorporating recognition species (functionalization with peptides, antibodies, organic polymers, etc.), sites of adsorption and desorption, chemical reactions, light scattering, among others.

Many different materials can be used as encapsulating matrices which must be selected depending on the critical properties needed for each intended application (see Table 1). The majority of these carriers are proteins (lecithin, legumin, gelatin, and albumin), polysaccharides (dextrin, starch, gums, chitosan, and alginates), fats, liposomes [4], biopolymers, co-polymers (poly(lactic-co-glycolic acid), micelles, organogels, dendrimers, solid nanoparticles



Nanoencapsulation, Fig. 2 Morphologies of nanocapsules (from *left* to *right*): (i) single-core capsule, (ii) dispersed core in polymer gel, (iii) multi-layer capsule, (iv) dual-core capsule and (v) single-core-multi-shell capsule

(SLN), polymeric nanoparticles, emulsions-base systems, and metal-organic particles [5].

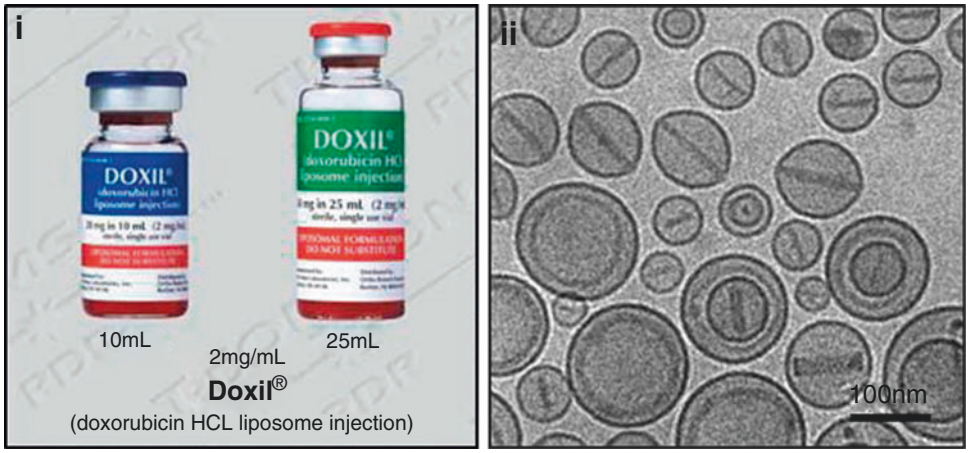
The main reason for nanoencapsulating species is to ensure that the encapsulated material reaches the area of action without being adversely affected by the external environment through which it passes. Thus, the principal reasons for encapsulation are: (a) separation of incompatible materials, (b) conversion of liquids to free flowing solids [6], (c) increased stability (e.g., protection of the encapsulated against oxidation or deactivation due to reaction in the environment), (d) masking organoleptic properties like color, taste, odor of substances, (e) safe handling of toxic materials, and (f) controlled and targeted release of encapsulated active compounds. In addition, nanoencapsulation allows to modify the release of encapsulated active materials, providing sustained release (maintaining the right concentration), long lasting (and therefore improving effects), target release (improving adhesion, penetration, or recognition of tissues and cells), or triggered release (mainly by environmental changes in pH, temperature). Encapsulation and release modification (sustained, triggered, or targeted release) also reduces doses and, therefore, potential toxicity of drugs. This is of the main interest in medical applications. Moreover, since the encapsulated molecule is not in direct contact with the environment, irritation at the site of administration is commonly reduced.

Because of this wide range of advantages and properties, nanoencapsulation technologies have applications in many industrial sectors, including the medical, pharmaceutical, cosmetics, chemical, textile, construction, agricultural and food industries. In fact, this technique is already commonplace within a range of industries, but it is accepted that only around 10 % of potential applications are currently being exploited. For example, in the case of medicine delivery, the core material (e.g., drugs, vaccines, etc.) can be targeted and released in a controlled manner and at a specific location (see Fig. 3). In agrochemistry, the nanocapsules were proposed for controlling the release of insecticides or pesticides and improve their efficiency and decrease the quantity of active substances for a given application. Or, in the textile industry, durable textile fragrances and skin softeners in the form of nanocapsules have been developed [6].

Basis Nanoencapsulation Techniques

Before considering the fabrication of nanoencapsulated products, one should first determine which are all the characteristics, properties, and capabilities that such products must satisfy for each intended application. For this, the following questions must be taken into consideration before choosing which nanoencapsulation methodology is the more appropriate: [2]

1. What functionality should the encapsulated ingredients (core) provide to the final product?
2. What kind of coating material (matrix) should be selected?



Nanoencapsulation, Fig. 3 (i) Nanoliposomes loaded with the chemotherapeutic agent doxorubicin in commercial product Doxil® (Ortho Biotech) and (ii) SEM image of doxorubicin liposomal in Nuclisomes® formulation (Image taken from Ref. [20], With kind permission from Springer Science+Business Media: Ref. [16], 2010)

3. What processing conditions must the encapsulated ingredients survive before releasing its content?
4. What is the optimal concentration of the active ingredient in the nanocapsule?
5. By what mechanism the ingredient must be released from the nanocapsules?
6. What are the particle size, density, and stability requirements for the encapsulated ingredient?
7. What are the cost constraints of the encapsulated ingredient?

Similarly, an understanding of the physicochemical properties of the core and factors that control the interfacial and aggregation behavior of the matrix materials is crucial to choose suitable processes for nanoencapsulation [1]. Today, a multitude of techniques are used for nanoencapsulating active species. The more popular techniques can be classified in three main families: chemical, physicochemical, and physicomechanical encapsulation processes [3]. The chemical and physicochemical processes include those methods in which chemical reactions are involved in the nanocapsule formation. In contrast, in the physicomechanical methods, no chemical reactions are involved in the nanocapsule formation and only shape fabrication usually takes place. Some of the techniques that are commonly used are briefly discussed below (see Table 2).

Chemical Processes

The chemical processes are building-up techniques where the nanocapsules are synthesized through the nucleation and growth processes involving elementary building blocks, such as ions, atoms, and molecules. The most important chemical methods include the emulsions, suspensions, precipitations, sol–gels, and polymerizations. The advantages of these chemical methods are that they provide nanocapsules with high purity, high uniformity, small particle size, narrow size distribution, dispersable, good chemical homogeneity, and more reactivity.

The majority of chemical methods are based on the initial formation of an emulsion. An emulsion is a suspension of small droplets (diameters below 1000 nm), usually stabilized by an emulsifier, of one liquid in a second liquid in which the first will not mix. These emulsions are conventional emulsions, multilayered emulsions, microemulsions, nanoemulsions, solid lipid emulsions, or double emulsions. Generally, the "*emulsification*" methods consist on first creating an emulsion of a liquid core in a continuous phase using shaking, stirring, homogenizing, or spray processes. In a second step, a shell is formed via organic polymerization, or the emulsion homogenized by microfluidisers, or the suspension atomized for solvent evaporation. Control over the properties of emulsion-based nanoencapsulation systems

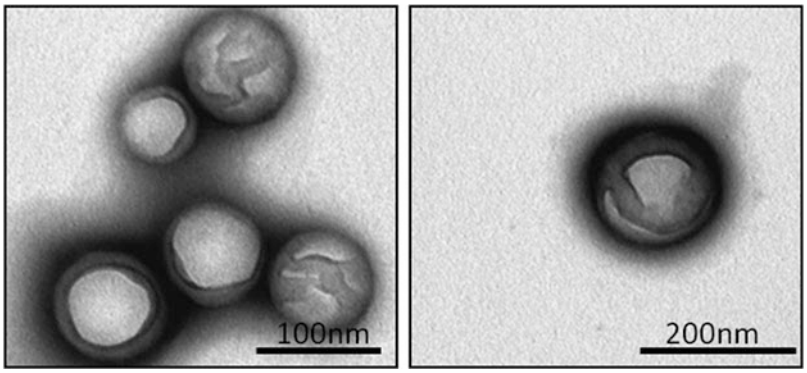
Nanoencapsulation, Table 2 Several nanoencapsulation techniques and the steps involved in each process

Nanoencapsulation technique	Processed involved
Sol–gel	(a) Solution of core and polymer (b) Formation of Sol phase (c) Gelation (d) Solidification
In situ polymerization	(a) Preparation of core solution (b) Addition of droplets of monomer
Coacervation	(a) Formation of a three-immiscible chemical phases (b) Deposition of the coating (c) Solidification of the coating
Rapid expansion of supercritical solution	(a) Preparation of solution of core and shell materials in CO_2 (b) Depressurization through a nozzle
Liposome entrapment	(a) Microfluidization (b) Ultrasonication (c) Reverse-phase evaporation
Inclusion complexes	Preparation of complexes by mixing or grinding
Spray-drying	(a) Preparation of a dispersion (b) Homogenization of the dispersion (c) Atomization of the dispersion
Solvent evaporation	(a) Preparation of solution of polymer and core (b) Solvent evaporation by heating
Electrocoextrusion	(a) Preparation of core solution and wall solution (b) Simultaneous spraying of two solutions from two coaxial capillaries

can be achieved by tailoring the dispersed phase as well as the structure of emulsions [1]. In general, the so-called microemulsions are defined as thermodynamically stable transparent or translucent isotropic dispersions of nanodroplets with a size ranging between 5 and 1000 nm, whereas nanoemulsions are metastable dispersions of nanoscale droplets (<100 nm) and optically transparent. This last property is of interest, for example, to the food and beverage industry as it enables the delivery of flavors and bioactive ingredients in clear emulsions. For example, kinetic stable nanoemulsions of triglyceride oils may be achieved using high-pressure homogenizers as the final step [7].

Emulsion encapsulation methods include the "in situ *polymerization*" and the "*interfacial polymerization*" methodologies, among others. In the "in situ *polymerization,*" the monomer is added dropwise to an emulsion which is formed by the aqueous polymerization solution, the active ingredient and an emulsifier, under continuous stirring. Initially, a low molecular weight polymer is formed in the continuous phase, and as the polymerization proceeds, the polymer gradually grows and simultaneously entraps the core material to form the final nanocapsules. For example, insulin loaded poly(alkyl cyanoacrylate) nanocapsules are prepared using this technique [8]. On the other hand, in the "*interfacial polymerization*" technique, the capsule shell is formed at/on the droplet surface by polymerization of the reactive monomers. First, the core material is dispersed into the continuous phase where the monomer is dissolved, forming an emulsion. A coreactant is then added to the mixture, provoking the polymerization at the core interface, thus generating the capsule shell. As an illustrative example, this method is used for the encapsulation of polyaminoacids in PLGA-polaxamer nanocapsules (see Fig. 4) [9].

The "*sol–gel*" method is a process based on the preparation of a solution, sol and gel, followed by the solidification and heat treatment of the organic and inorganic compounds. "*Sol–gel*" encapsulation involves the evolution of inorganic networks through the formation of a colloidal suspension (sol) and gelation of the sol to form a network in a continuous liquid phase (gel). In this method, the capsules can range between submicron sizes to a few tens of microns [10]. The most popular use of this process is the production of immobilized biomaterials, such as enzymes and microorganisms. Its main advantages are the very mild operating conditions as well as the small particle size achievable and the affordable raw materials. From an industrial point of view, this process is easily scaled up even considering the tight cost-in-use constraints characteristics of the food industry. For example, this method can be used for the immobilization of lipase enzymes within a



Nanoencapsulation, Fig. 4 TEM image of a polyaminoacid loaded in PLGA: poloxamer nanocapsules (Image taken from Ref. [9])

phyllosilicate and vinyltriethoxysilane sol–gel matrices and beta-glucosidase in alginate-silicate sol–gel matrices.

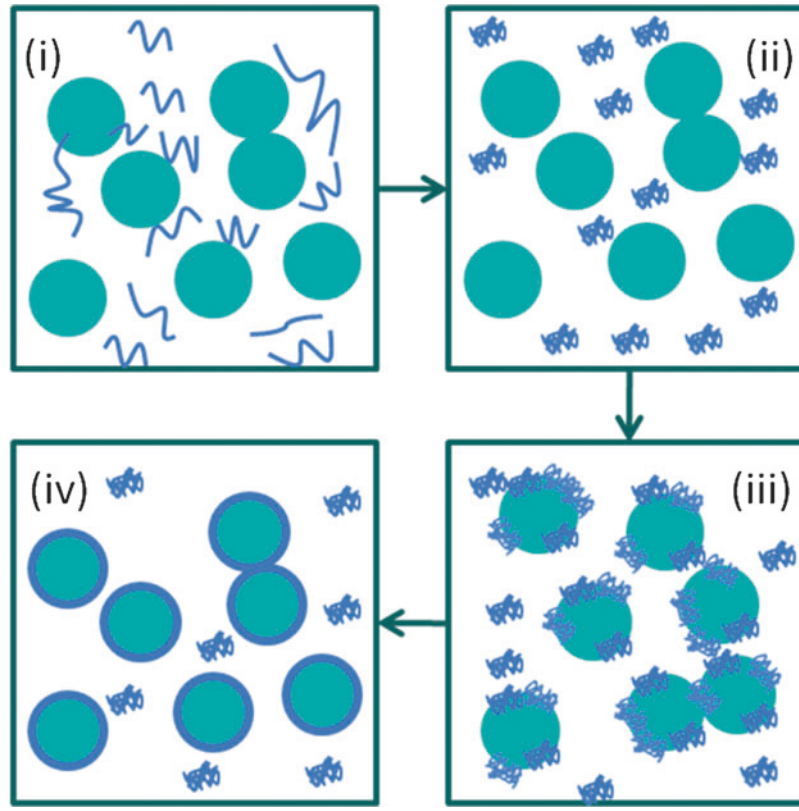
In addition to the chemical nanoencapsulation techniques described above, nanoencapsulation can also be carried out by "*suspension cross-linking.*" This is the method of choice for encapsulating proteins and polysaccharides. This method involves the dispersion of a polymer aqueous solution containing the core material in an immiscible organic solvent to form an emulsion. Droplets are hardened by covalent cross-linking and directly converted to capsules. The cross-linking process is accomplished either thermally or by the use of cross-linking agents. Albumin nanocapsules containing doxorubicin and magnetite particles are prepared by this technique. Salmon calcitonin-loaded chitosan-PEG nanocapsules are prepared with this methodology as well.

Finally, nanoencapsulation can also be carried out via "*hydrothermal methods*" where chemical reactions in water at high temperatures and high pressures are involved for the nanocapsule formation and via "*oxidation processes*" that oxide and deoxidize directly the raw materials in the liquid phase or quasi-liquid phase state to prepare the nanocapsules.

Physicochemical Processes

Physicochemical methods are the processes that combine both the chemical and physical methods during the nanoencapsulation. For example, "*phase inversion nanoencapsulation*" (PIN) is a method for highly efficient encapsulation of drugs, proteins, or plasmid DNA into biodegradable polymer nanospheres. Capsules are produced with a limited size distribution, ranging from 100 nm to 10 μm in diameter, and labile drugs are efficiently encapsulated without denaturation. In this method, the active compound is added to a dilute polymer solution (normally in methylene chloride) which is then poured rapidly into an unstirred bath of nonsolvent (petroleum) at a solvent to nonsolvent ratio of 1:100, causing nano- and microspheres to form spontaneously. The drug dicumarol, proteins such as bovine serum albumin, hormones such as insulin, and plasmid DNA have been successfully encapsulated using the PIN system in combination with biologically erodible, biologically adhesive polymers like fumaric acid, sebacic acid, poly-lactide-co-glycolide (PLGA), polylactic (PLA) acid, and polystyrene.

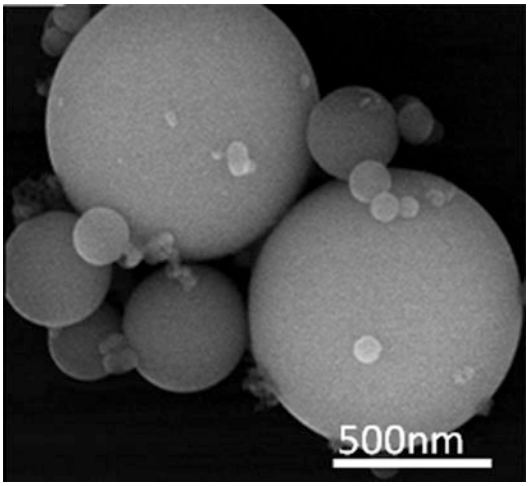
"*Coacervation and phase separation*" is another physicochemical method based on the gradual desolvation of a homogeneous polymer solution into a polymer-rich phase (coacervate) and the poor polymer phase (coacervation medium). Phase separation processes are divided into simple and complex coacervation. The mechanism of capsule formation is identical in both processes, except in the way in which the phase separation is carried out (see Fig. 5). In a simple coacervation, a desolvation agent, such as gelatin or ethyl cellulose, is added to the aqueous or organic media for phase separation, whereas complex coacervation involves complexation between two oppositely charged polymers, both soluble in an aqueous media. Nanoencapsulation by coacervation is carried out by preparing an aqueous



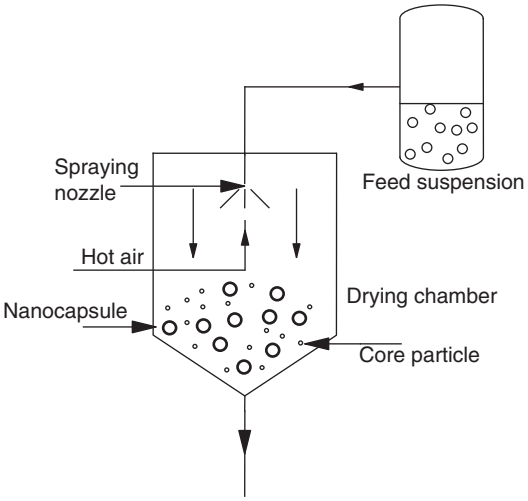
Nanoencapsulation, Fig. 5 Schematic illustration of a typical coacervation process, involving (i) dispersion of the core material in a solution of the shell polymer, (ii) initial separation of the coacervate from the solution, (iii) coating of the core material by microdroplets of the coacervate and (iv) coalescence of the coacervate to form continuous shell by reticulation at the interface

polymer solution (1–10 %) at 40–50 °C where the hydrophobic core material is also dispersed/dissolved. Stabilizers are usually added to the mixture to keep the individuality of the final particles. Then, a suitable desolvating agent is gradually added to the mixture leading to the formation of partially desolvated polymer molecules and, hence, their precipitation on the surface of the core material. Finally, the prepared capsules are stabilized by cross-linking, desolvation, or thermal treatment (e.g., cooling at 5 °C). An illustrative example is the gelatin nanocapsules (840 nm) loaded with the protein/peptide drug bovine serum albumin (BSA) which have been produced by a modified coacervation method free of desolvating agent.

Another common physicochemical approach is the use of molecular systems that self-assemble into capsules at given conditions, such as the aggregation of lipid molecules into spherically closed bilayer structures at low concentrations, the so-called vesicles or liposomes, or molecular inclusion complexes, such as cyclodextrins and dendrimers or hyperbranched polymers. "*Liposomes*" are the smallest artificial vesicles of spherical shape that can be produced from natural nontoxic phospholipids and cholesterol. When phospholipids are dispersed in an aqueous phase, the liposomes are formed spontaneously, thus encapsulating either aqueous or lipid-soluble materials. In general, they are considered very versatile tools in biology, biochemistry, and nanomedicine because they are currently been used for delivery of vaccines, hormones, enzymes, and vitamins (see Fig. 3) [4]. The great advantage of liposomes over other nanoencapsulation techniques is the stability that liposomes impart to water-soluble material in high water activity applications [2]. Recently, it has been published a very interesting nanoencapsulation technology that produced capsules coined "colloidosomes" because of their resemblance to liposomes, while using colloid particles as shell material. Also, the ultimate achievement in enzyme protection and



Nanoencapsulation, Fig. 6 SEM image of PLGA-encapsulated SiO2 powder prepared by rapid expansion of supercritical suspension process (Reprinted with permission from Ref. [11]. Copyright 2009 American Chemical Society)



Nanoencapsulation, Fig. 7 Schematic representation of the spray-drying process

delivery is represented by polymeric vesicles which are called polymersomes.

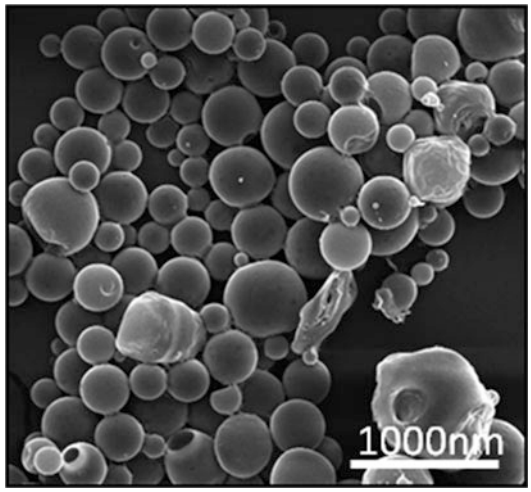
"*Inclusion complexes*" generally refers to the supramolecular association of the encapsulated substance into a cavity-bearing complex or dendrimer (shell). Small organic molecules can displace the water from the inner cavity and form thermodynamically stable complexes. The encapsulated unit is kept within the cavity by hydrogen bonding, van der Waals forces or by the entropy-driven hydrophobic effect. Examples include the encapsulation of hydrophobic vitamins (A, E or K) in cyclic oligosaccharides such as β-cyclodextrins.

Other physico-chemical processes for nanoencapsulation are based on *supercritical fluids*. In these methods, the core ingredient is dispersed in a matrix material solubilized in a supercritical fluid (usually carbon dioxide). Removal of the carbon dioxide results in the core being encapsulated within the matrix material. For example, in *rapid expansion of supercritical solutions* (RESS), the dispersion is sprayed through a nozzle and a particulate material containing the core is formed. Titanium dioxide and silica nanoparticles coated with poly(lactide-co-glycolide) are prepared using the RESS process (see Fig. 6) [11].

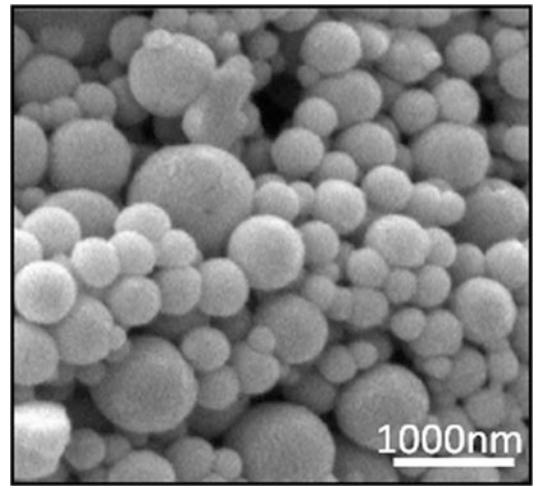
In addition to the nanoencapsulation techniques described above, microfluidic platforms are started to be exploited for the synthesis of nanocapsules because these platforms allow a precise control of their size and shape and, therefore, of their physico-chemical properties [12]. For example, anti-fungal drug itraconazole has been encapsulated in polymethylsilsesquioxane nanospheres with a controlled drug release profile using a V-junction microfluidic device [13]. Also, other physico-chemical processes are "*solid lipid nanoparticles*" (SLN), "*layer-by-layer deposition* (e.g., *encapsulation by polyelectrolyte multilayer*)" and "*controlled precipitation.*"

Physico-Mechanical Processes

The most important physico-mechanical methodologies for encapsulating desired species at the nanometer scale are: (i) spray drying, (ii) solvent evaporation/solvent extraction, and (iii) electroencapsulation-based techniques. Among them, "*spray drying*" is a process widely used in the encapsulation of fragrances, oils and flavors. In a "*spray drying process,*" core particles are dispersed in a polymer solution of wall material and atomized into a hot chamber (see Fig. 7). As the solvent evaporates, the shell material solidifies onto the core particles, obtaining polycore or



Nanoencapsulation, Fig. 8 SEM image of oxaprozin/methylated-β-cyclodextrin-loaded PLGA nanocapsules (Image taken from Ref. [9])

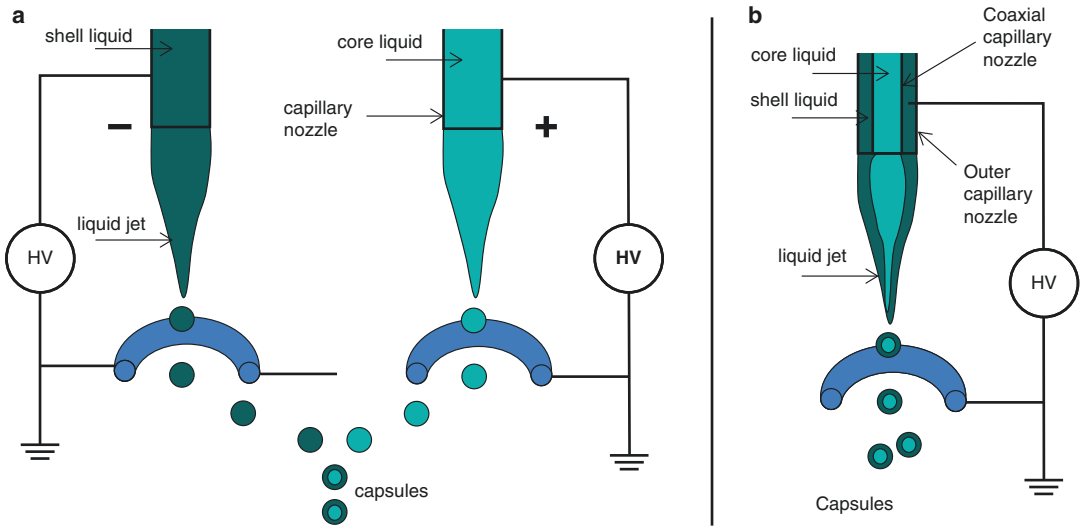


Nanoencapsulation, Fig. 9 SEM image of naproxen-PLGA nanocapsules prepared by solvent evaporation/extraction technique (Reprinted from Ref. [14], Copyright (2010), with permission from Elsevier)

matrix type capsules. The one-droplet-to-one-particle mechanism during the spray drying sets the lower limit of the particle size to 100 nm. For example, this technique is used for preparing 260 nm in diameter amifostine/PLGA nanocapsules.

"*Solvent evaporation/solvent extraction*" is a method based on the dissolution of a polymer in a water immiscible organic solvent, such as dichloromethane or chloroform, where the core material is dissolved or dispersed as well. The mixture is then added dropwise to a stirring aqueous solution to form small polymer droplets containing the encapsulated material, and then, the solvent is eliminated. Here, the coat material shrinks around the core material and encapsulates it. The solvent elimination can be done by simple solvent evaporation (heat or reduced pressure) or by solvent extraction with a third liquid which is precipitant for the polymer and miscible with both water and the solvent. This method is suitable for the preparation of fluorescein-loaded nanocapsules based on PLGA using droplet-based microfluidic platforms. PLGA nanocapsules loaded with naproxen and oxaprozin-cyclodextrin are also prepared using this methodology (see Figs. 8 and 9) [14]. In some cases, the "*solvent evaporation/solvent extraction*" method is also combined with "*freeze drying*" process; for example, in the encapsulation of hypericin in polylactic acid.

Recently, pharmaceuticals, cosmetic or food industries seek for new, more efficient technologies for nanocapsules production which do not degrade the chemical or biological properties of the material. Electroencapsulation-based techniques are excellent candidates to be some of these technologies. "*Electrospraying*" is a straightforward bottom-up method to manufacture capsules of submicrometer diameters. It is a method of liquid atomization by means of electrical forces. In "*electrospraying,*" the liquid flowing out of a capillary nozzle is subjected to an electrical shear stress by maintaining the nozzle at high electric potential which forces the liquid to be dispersed into fine droplets. The nanocapsules are produced by previously dissolving the polymer and the active compound into the solvent(s). With this method, the droplet sizes can range from hundreds micrometers down to several tens of nanometers, and they can be controlled by electrical means, i.e., by adjusting the flow rate and voltage applied to the nozzle. Recently, research in electronanoencapsulation is aimed at developing new drugs delivery systems, medicine



Nanoencapsulation, Fig. 10 Schematic representation of two electro-encapsulation processes: (i) electrospraying of two oppositely charged droplets and (ii) electrocoextrusion

production and ingredients dosage in the cosmetic and food industries since the utilization of electrical forces can increase the effectiveness of nanoencapsulation. As a consequence, some "*electrospraying*" techniques have recently been described. For example," *Impacting of two oppositely charged droplets*" is a method where both of the droplet's streams are emitted from two separated capillary nozzles maintained at opposite potentials, one of the capillaries at positive and the other at negative (see Fig. 10a). The droplets collide due to Coulomb attraction forming a capsule via submerging the droplet of higher surface tension within that of smaller surface tension. Another method is the "*electrospraying/evaporation of colloidal suspension,*" in which a suspension is initially electrosprayed followed by the shell solidification by solvent evaporation. Similarly, in the "*electrospraying/gelatinization of colloidal suspension*" method, a suspension of a core material is electrosprayed into a bath containing a gelatinizing or polymerizing agent. This agent forms a hard envelope on the core material. Among other actives, eletrospraying has been tested for the encapsulation of the drug paclitaxel in PLGA and polycaprolactone. Recently, nanocapsules composed of natural polymers (e.g., chitosan) for ampicillin and elastin-like polypeptides for doxorubicin have also been fabricated using the "*electrospraying*" methods.

The "*electrocoextrusion*" process is another electronanoencapsulation technique based on the simultaneous spraying of two different liquids from two coaxial capillaries that are maintained at the same potential (see Fig. 10b). In this method, the core liquid flows from the central capillary and the envelope liquid flows through the annular nozzle between the capillaries. It is important to note that the core liquid must have a high resistivity and the envelope sufficiently high conductivity in order to be used in this methodology. At the present time, the smallest capsules reported using this methodology has a diameter of 150 nm, and they were produced using a photopolymer (Somos 6120-DuPont) and ethylene glycol.

Besides the formation of spherical nanocapsules, electronanoencapsulation technologies can also allow the encapsulation into nanofibers. Indeed, "*electrospinning*" is a method to produce nanofibers where a suspension of the core material in a polymer solution is supplied from a spinneret and forms a droplet at the spinneret exit. In this methodology, an electric field is

applied to the suspension by immersing an electrode in it and placing the counter-electrode some distance from the spinneret and then jetting sets in. This field provokes an electrically induced bending instability in the solution jet which causes stretching of the bent sections of the jet. The solvent eventually evaporates, the jet dries and solidifies, and the spun nanofibers are deposited on the counter electrode. The nanofibers are collected onto the edge of a grounded collector disc. This method has been used in biotechnology, i.e., for the capturing of living cells (virus and bacteria) in poly(vinyl alcohol) nanofibers.

In addition to the techniques described above, other physicomechanical methods that can be used for encapsulating desired species at the nanoscale are the "*melt-solidification,*" "*centrifugal suspension separation,*" "*vibrating and coaxial nozzle,*" and "*pan-coating*" processes.

Release Mechanisms

Different release mechanisms of encapsulated materials provide controlled, sustained, and targeted release of core materials. There are three main mechanisms by which encapsulated material is delivered from a capsule: (i) mechanical rupture of the capsule wall, (ii) dissolution (impact/shear/pressure) or melting of the wall, and (iii) diffusion through the wall. The desired mechanism depends on the nature of the application. For example, carbonless copy paper and scratch and sniff perfumes rely on mechanical rupture of shell to release the core contents. In these applications, the rupture may be caused by pressure or due to propagation of cracks. In contrast, detergent industry uses dissolution of shell wall of powder detergents for the release of encapsulated protease enzyme in order to remove bloodstains from clothing. Similarly, in pharmaceutical industry, nanoencapsulated products are designed for sustained and controlled release by either degradation or by diffusion through the shell. For example, doxorubicin is encapsulated in liposomes to control its release into the body reducing its cardiotoxicity [15].

Applications

Nanoencapsulation is believed by many scientists and technological fields to be one of the most important scientific topics of the next decade with a huge impact on medicine, pharmacology, construction, cosmetics, chemistry, textile, agriculture and food industry, among others. Nowadays, nanoencapsulation is being used in a diverse range of markets, including, but not limited to agrochemicals, aromatherapy, paints, coatings, colorants, pharmaceuticals, personal care, adhesives, biocides, nutraceuticals, oil and gas, textiles, paper systems, electronics and imaging systems (see Table 3). Depending on the application, the nanoencapsulated product can be delivered as a dry, free flowing powder, as slurry or in the form of a wet filter cake. A brief explanation of some applications in some of the most powerful sectors is given below.

Agriculture

Nanoencapsulation appears to be the most promising technique for short-term application in the field. One of the most important applications of nanoencapsulated products is in the area of crop protection in agriculture. In the case of controlling parasitic plants, encapsulation can be used to solve problems regarding phytotoxicity on the crop of the herbicides used against the parasite. For example, glyphosate, imidazolinones, and sulfonylureas can be encapsulated within a polymeric shell. Germination stimulants can be also delivered into the soil inside nanocapsules protecting them from degradation.

Pharmaceutics and Medicine

Nanoencapsulation currently has the highest potential in the pharmaceutical industry because it allows the possibility to deliver a drug at the right moment, in the right place and at an adequate concentration. One of the simplest examples is the protection of orally applied drugs from the attack of acids in the stomach before they can be adsorbed in the intestine. This problem is solved in many cases by the encapsulation of the active ingredient in chitosan and PLGA (poly(lactic-co-glycolic acid) nanocapsules. In addition, both

Nanoencapsulation, Table 3 Examples of nanoencapsulated materials in the market

Product	Company	Industry
Caelyx® Myocet® LBL-Intra® PharmFilm® Doxil® Lipozyme IM-60 AmBisome® Bioral™ Amphotericin B	Sequus Pharma Cephalon Limited Capsulution MonoSol Rx Alza Corp. NNBN America Inc. Gilead BioDelivery Sciences	Pharmaceutical
MicroMatrix™ VitaDHA™ Pharmasmooth™ Cavamax®, Omegadry® LycoVit® NovaSOL® Nanotea NutraLease Nutri-Nano CoQ_{10} Solgar NanoCeuticals™	ABN Blue California Balchem Corp. Wacker Chemie AG BASF Aquanova Shenzhen Become Ind & Trade Shemen Ind. Leonia RBC Life Science Inc.	Food
Plenitude® Primasys® Tinoderm Cutanova®	L'Oréal Cognis BASF Dr.Rimpler GmbH	Personal care
Combidex® Verigene® SAIVI™ 715 Fluoresbrite® Endorem® Ferucarbotran Resovist Feraheme® IDX FluidMAG-ARA	Advanced Magnetic Nanosphere Triton Biosystem Invitrogen PolySciences Guerbet Bayer GE healthcare Ocean NanoTech Chemicell	Imaging/ Diagnostic

polymers are widely used in many aspects of biomedical research ranging from tissue engineering to drug delivery because they are highly biocompatible, easily available, and provide excellent delivery capabilities. For example, ascorbyl palmitate can be encapsulated in a chitosan derivative with an efficiency of 84 % at 56 % drug loading. The nanocapsules display no short-term cytotoxicity against the human skin melanoma. On the other hand, nanoencapsulation is also used for cancer therapy. Indeed, two nanoencapsulated products named as Doxil® and Abraxane® have recently been approved by the European Medicines Agency (EMA). Doxil® is a liposomal system for doxorubicin delivery and treatment of ovarian carcinoma (see Fig. 3) and Abraxane® is an albumin nanoparticle taxol conjugate for the treatment of metastatic breast cancer [16]. More recently, there is also an increasing interest in the preparation of polymeric nanocapsules incorporating vegetable oils (mainly, from linseed, sunflower, grape seed, almond, Brazil nut, olive, and coconut) as a consequence of their antifungal and antioxidant activities [17]. Another tendency is to use nanoencapsulation to enhance the stability, biocompatibility, or efficacy of other type of nanoparticles, such as quantum dots, widely used in multimodal imaging and imaging-guided therapies [18].

Food Industry

Today's food industry makes use of nanoencapsulation to develop "functional foods" with ingredients that improve the nutritional content of food without affecting the taste, aroma, or texture of food and enhance the shelf-life and stability of the ingredient and the finished food product [1]. Nanoencapsulation is used to protect sensitive food ingredients (i.e., flavors, oils, vitamins) as well as to provide viable texture blending, appealing aroma release and taste, odor and color masking. This technology also enables food companies to incorporate minerals, vitamins, flavors, and essential oils [7]. In addition, nanoencapsulation can simplify the food manufacturing process by converting liquids to solid powder, reducing costs allowing low cost handling equipment. Moreover, nanocapsules also help fragile and sensitive materials (oils, vitamins, enzymes) to survive processing and packaging conditions and to stabilize the shelf life of the active compound. For example, there are several commercial nanoencapsulated products in the market that offer specific supplement of Omega-3 DHA and of natural form of folated for pregnant and lactating women, such as VitaDHA™ Materna.

Personal Care

Nowadays nanoencapsulation technologies play an important and growing role in the health and beauty industry. They offer an ideal and unique carrier system for active ingredients aside from an effective protection against photodegradation. Their striking effects can enhance the tactile and visual appearance of a variety of cosmetic and personal care products, including bath gels, lotions, tanning, makeup, aromatherapy, etc. Furthermore, nanoencapsulation can make significant savings for formulator as it can reduce the amount of active ingredients needed. For example, the retinyl palmitate oil which is the most stable form of vitamin A and plays an important role in cellular differentiation and carcinogenesis prevention has been recently prepared as a novel nanocapsule antiaging formulation. Flexible polymeric nanocapsules with retinyl palmitate core and PLA shell as carrier of drugs or simply as a reservoir of vitamin A were prepared. The main advantage of these capsules is that they can easily pass through biological barriers, due to deformability characteristics, reaching deeper layers of the skin and allowing a uniform permeation in the skin. Another example are liposomes as source of magnesium ascorbyl phosphate, a stable derivative of vitamin C, which allow higher effects of vitamin, such as whitening and antioxidants thanks to the targeted activity provided by the liposome.

Other Applications

Besides many classical applications, nanoencapsulation technologies provide a large number of other applications. One of them is the preparation of high-surface area carbon nanocapsules (100 nm) with hollow cores and porous shells. Such mesoporous carbon nanocapsules are of extensive interest for catalysis, separation, gas storage, and fuel cells, among others applications. Another novel application is the nanoencapsulation of phase change materials for energy storage since this class of materials is able to absorb and release energy during its phase change [19]. Nanoencapsulation also plays an important role in the construction industry by improving traditional materials and facilitating the incorporation of nanomaterials. For example, nanoencapsulated corrosion inhibitors have shown good results in protecting steel, and carbon encapsulated nickel and iron nanoparticles have improved cement mortar resistance to compressive strength [20].

Future Directions for Research

The nanoencapsulation of active compounds for controlled release applications is a promising alternative to solve some of the major problems of active ingredients faced by industries. Today, a wide range of encapsulated products have been developed, manufactured, and successfully marketed in the pharmaceutical, medical, and cosmetic industries. However, nanoencapsulation has still found a small market in many other industries such as food, construction, etc. In these potential fields, nanoencapsulation is still far from being fully developed since the development time is rather long and requires multidisciplinary cooperation. For this reason, great advances in nanoencapsulation are highly expected in the near future. These advances will include the development of novel materials to be used as matrices, more stable, efficient and biocompatible nanocapsules and advanced nanoencapsulation technologies as well as their transferring to industrial-scale mass production. Such improvements and novel capabilities will expand the scope of application for nanoencapsulation technologies in many industries, where such innovations will certainly open new opportunities to develop a novel generation of products.

Cross-References

- ▶ Liposomes
- ▶ Nanomedicine
- ▶ Nanoparticles

References

1. Augustin, M.A., Hemar, Y.: Nano- and micro-structured assemblies for encapsulation of food ingredients. Chem. Soc. Rev. **38**, 902–912 (2009)

2. Desai, K.G.H., Park, H.J.: Recent developments in microencapsulation of food ingredients. Drying Technol. **23**, 1361–1394 (2005)
3. Jyothi, N.V.N., Prasanna, P.M., Sakarkar, S.N., Prabha, K.S., Ramaiah, P.S., Srawan, G.Y.: Microencapsulation techniques, factors influencing encapsulation efficiency. J. Microencapsul. **27**, 187–197 (2010)
4. Ciobanu, M., Heurtault, B., Schultz, P., Ruhlmann, C., Muller, C.D., Frisch, B.: Layersome: development and optimization of stable liposomes as drug delivery system. Int. J. Pharm. **344**, 154–157 (2007)
5. Imaz, I., Rubio-Martinez, M., Garcia-Fernandez, L., Garcia, F., Ruiz-Molina, D., Hernando, J., et al.: Coordination polymer particles as potential drug delivery systems. Chem. Commun. **46**, 4737–4739 (2010)
6. Jaworek, A.: Electrostatic micro- and nanoencapsulation and electroemulsification: a brief review. J. Microencapsul. **25**, 443–468 (2008)
7. Wooster, T.J., Golding, M., Sanguansri, P.: Impact of oil type on nanoemulsion formation and Ostwald ripening stability. Langmuir **24**, 12758–12765 (2008)
8. Damge, C., Vranckx, H., Balschmidt, P., Couvreur, P.: Poly(alkyl cyanoacrylate) nanospheres for oral administration of insulin. J. Pharm. Sci. **86**, 1403–1409 (1997)
9. d'Angelo, I., Parajó, Y., Horváth, A., Kéri, G., La Rotonda, M.I., Alonso, M.J.: Improved delivery of angiogenesis inhibitors from PLGA: poloxamer blend micro- and nanoparticles. J. Microencapsul. **27**, 57–66 (2010)
10. Ciriminna, R., Sciortino, M., Alonzo, G., Schrijver, A., Pagliaro, M.: From molecules to systems: sol–gel microencapsulation in silica-based materials. Chem. Rev. **111**, 765–789 (2011)
11. Kongsombut, B., Tsutsumi, A., Suankaew, N., Charinpanitkul, T.: Encapsulation of SiO_2 and TiO_2 fine powders with poly(DL-lactic-co-glycolic acid) by rapid expansion of supercritical CO_2 incorporated with ethanol cosolvent. Ind. Eng. Chem. Res. **48**, 11230–11235 (2009)
12. Puigmarti-Luis, J.: Microfluidic platforms: a mainstream technology for the preparation of crystals. Chem. Soc. Rev. **43**, 2253–2271 (2014)
13. Kucuk, I., Ahmad, Z., Edirisinghe, M., Orlu-Gul, M.: Utilization of microfluidic V-junction device to prepare surface itraconazole adsorbed nanospheres. Int. J. Pharm. **472**, 339–346 (2014)
14. Javadzadeh, Y., Ahadi, F., Davaran, S., Mohammadi, G., Sabzevari, A., Adibkia, K.: Preparation and physicochemical characterization of naproxen-PLGA nanoparticles. Colloids Surf. B Biointerfaces **81**, 498–502 (2010)
15. Bonacucina, G., Cespi, M., Misici-Falzi, M., Palmieri, G.F.: Colloidal soft matter as drug delivery system. J. Pharm. Sci. **98**, 1–42 (2009)
16. Fondell, A., Edwards, K., Ickenstein, L., Sjöberg, S., Carlsson, J., Gedda, L.: Nuclisome: a novel concept for radionuclide therapy using targeting liposomes. Eur. J. Nucl. Med. Mol. Imaging **37**, 114–123 (2010)
17. Santos, S.S., Lorenzoni, A., Pegoraro, N.S., Denardi, L.B., Alves, S.H., Schaffazick, S.R., et al.: Formulation and in vitro evaluation of coconut oil-core cationic nanocapsules intended for vaginal delivery of clotrimazole. Colloids Surf. B Biointerfaces **116**, 270–276 (2014)
18. Li, K., Liu, B.: Polymer-encapsulated organic nanoparticles for fluorescence and photoacoustic imaging. Chem. Soc. Rev. **43**, 6570–6597 (2014)
19. Fuensanta, M., Paiphansiri, U., Romero-Sánchez, M.D., Guillem, C., López-Buendía, Á.M., Landfester, K.: Thermal properties of a novel nanoencapsulated phase change material for thermal energy storage. Thermochim. Acta **565**, 95–101 (2013)
20. Hanus, M.J., Harris, A.T.: Nanotechnology innovations for the construction industry. Prog. Mater. Sci. **58**, 1056–1102 (2013)

Nanoengineered Concrete

Konstantin Sobolev[1] and Florence Sanchez[2]
[1]Department of Civil Engineering and Mechanics, University of Wisconsin-Milwaukee, Milwaukee, WI, USA
[2]Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA

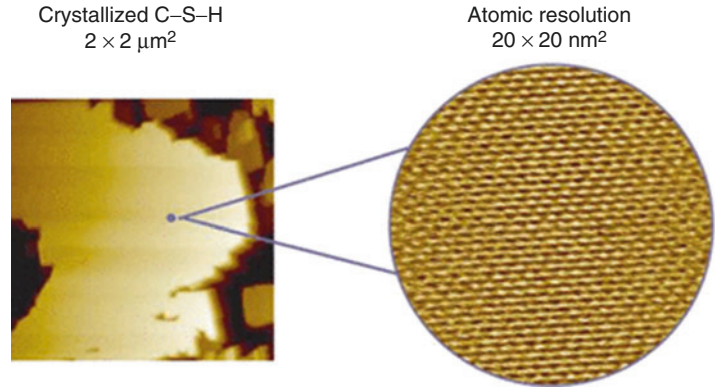
Synonyms

Nano-concrete

Definitions

What Is Nano-concrete?

Portland cement concrete is a complex nanostructured, multiphase, multiscale composite material that evolves over time [1, 2]. The elementary block calcium-silicate-hydrate (C-S-H), which holds the concrete composite together, is also a nanostructured material (Fig. 1) [3]. The properties of concrete exist in multiple length scales (nano to micro to macro). Processes occurring at the nanoscale ultimately affect the engineering properties and long-term performance of concrete [2].



Nanoengineered Concrete,
Fig. 1 Nanoscale structure of C-S-H crystallized on calcite substrate and revealed by AFM (Ca/Si = 0.9) (Reprinted with permission from the American Ceramic Society Bulletin, 2005, vol. 84, N 11)

The nanotechnology of concrete can be represented by two directions: nanoscience and nanoengineering (or nanomodification) [1, 2, 4]. As defined by Sanchez and Sobolev [2] nanoscience "deals with the measurement and characterization of the nano- and microscale structure of cement-based materials to better understand how this structure affects macroscale properties and performance through the use of advanced characterization techniques and atomistic or molecular level modeling." Nanoengineering includes "the techniques of manipulation of the structure at the nanometer scale to develop a new generation of tailored, multifunctional, cementitious composites with superior mechanical performance and durability potentially having a range of novel properties such as: low electrical resistivity, self-sensing capabilities, self-cleaning, self-healing, high ductility, and self-control of cracks [2]." Concrete can be nanoengineered by the incorporation of nanosized (less than 100 nm) building blocks or objects (e.g., nanoparticles and nanotubes) to control material behavior and add novel properties or by the grafting of molecules onto cement particles, cement phases, aggregates, and additives (including nanosized additives) to provide surface functionality, which can be adjusted to promote specific interfacial interactions.

Introduction

Though nanotechnology is a new emergent technology, nanosized objects have existed for millennia, from ancient times when humans began to use nanosized materials in glass [5] to modern times where "classic" photography employs silver nanoparticles sensitive to light. Natural materials have been shown to exhibit exceptional mechanical properties that are in some cases unmatched by man-made materials. For example, the exceptional performance of biomaterials such as mollusk shells has been attributed to the presence of calcium-based nanocrystals [6]. One of the most promising directions in nanotechnology is a better understanding of the relationship between microstructure and strengthening mechanisms employed by nature with the goal to ultimately mimic nature's design principle and self-organization (bottom-up approach) [5].

The nanoparticle is the elementary building block in nanotechnology and is comprised of up to thousands of atoms combined into a cluster of 1–100 nm. The conditions of the bottom-up production of nanoparticles allow for control of the size and shape of the particle. A reduction in size provides an exceptional surface area-to-volume ratio and changes in the surface energy, surface chemistry, and surface morphology of the particle, altering its basic properties and reactivity [1, 2, 5, 6]. Nanoparticles have been shown to significantly enhance the mechanical performance of a variety of materials, including metals, polymers, ceramic, and concrete composites [1, 5–7].

By far, the most significant development has been the synthesis of fullerenes (C_{60}) and fullerene-like structures such as carbon nanotubes [5]. Since their discovery, carbon nanotubes have

been used in numerous applications varying from superconductors, optical and electronic devices, sorbents, drug delivery systems, and nanostructured materials. The exceptional tensile strength of carbon nanotubes (as much as 20 times higher than that of steel) makes them ideal reinforcing components for structural applications, such as bridges and high-rise buildings. Polymer-based nanofibers and nanotubes [5, 8, 9] are promising nanomaterials that could open the door for new developments and innovative applications in building and construction industry.

Understanding the Structure at the Nanolevel and "Bottom-Up" Design

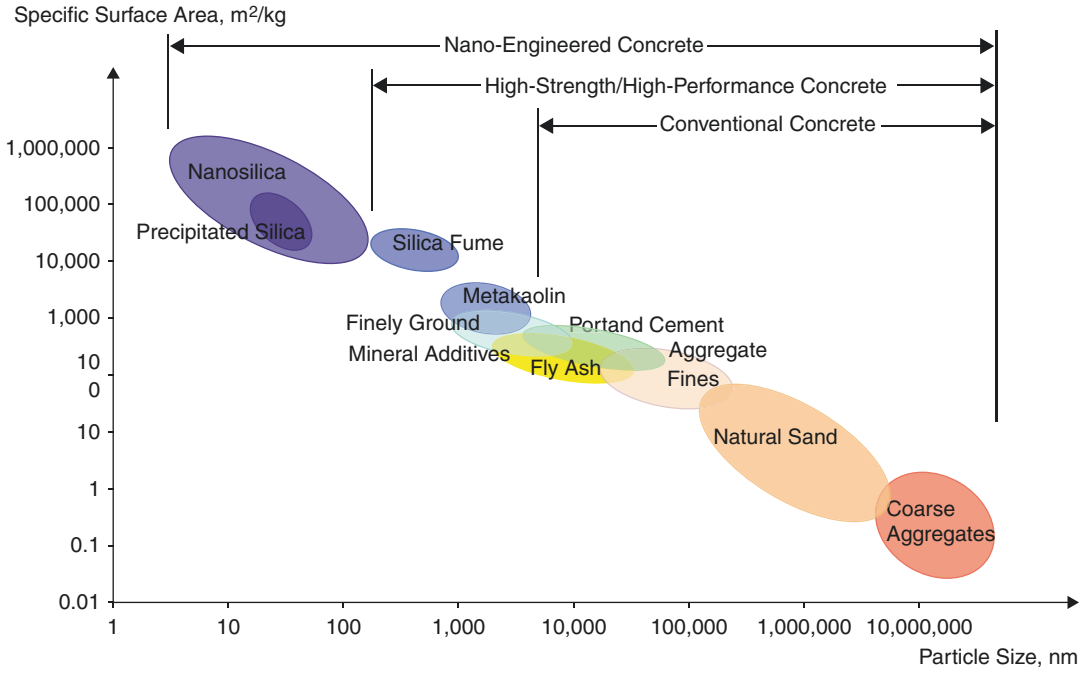
Most research on nanotechnology in concrete has focused to date on the investigation of the structure and mechanical properties of concrete at the nanoscale [1–4, 10, 11]. Recent advances in instrumentations have made it possible to characterize the structure of concrete at the nanoscale and to measure the local mechanical properties of its micro- and nanoscopic phases [10]. Significant progress in understanding nanoscale processes in cementitious materials has been achieved thanks to the use of nanoscale characterization techniques [1, 2, 10, 11]. These advanced techniques include nuclear magnetic resonance, atomic force microscopy, micro-and nanoindentation, neutron scattering, ultrasonic force microscopy, and focus-ion beam (FIB) nanotomography. For example, the use of atomic force microscopy (AFM) has revealed that, contrary to general thought, nanoscale C-S-H has in fact a highly ordered structure (Fig. 1) [3]. A better understanding of the structure of concrete at the nanolevel will allow for a better control of concrete performance and even the tailoring of desired properties and is expected therefore to affect the method of production and use of concrete.

Another application of nanotechnology in concrete has come from the "bottom-up" possibilities of nanochemistry with the development of new products such as novel superplasticizers and new coating materials [12, 13]. One such superplasticizer was recently developed by BASF and was specifically designed for controlling the slump retention characteristic of the concrete mixture [13]. Another example is the development of coating materials with new self-cleaning properties, discoloration resistance, antigraffiti protection, and high scratch-and-wear resistance. These self-cleaning materials are based on photocatalyst technology [12]. Titanium dioxide (TiO_2) is used as a photocatalyst for the decomposition of organic compounds. TiO_2 is active under exposure to UV light, exhibiting self-cleaning and disinfecting properties. Another aspect of self-cleaning is provided by the hydrophilicity of the surface, which helps to prevent dust and dirt from attaching to it.

In the past, major developments in concrete technology have been achieved through the use of superfine particles such as fly ash and silica fume. Recent advances in nanochemistry and the development of new methods for synthesis of nanoparticles are now expected to offer a new range of possibilities for improvement of concrete performance [14, 15] (Fig. 2). Incorporation of nanoparticles into conventional construction materials can provide the materials with advanced or smart properties that are of specific interest for high-rise, long-span, or intelligent infrastructure systems [1, 2, 10, 11, 15].

Performance of Concrete with Nanoparticles

The addition of nanoparticles has been shown to improve various performances of concrete [1, 2, 15–20]. For example, nanosilica (silicon dioxide nanoparticles, nano-SiO_2) has been shown to improve workability and strength in high-performance and self-compacting concrete [1, 2, 14, 15]. Improvement in concrete performance has been attributed to several effects of the nanoparticles. Well-dispersed nanoparticles can act as crystallization centers for cement hydrates, thereby accelerating the hydration reactions, and can act as filler, filling the voids between the cement grains, thus reducing the material porosity [15–17]. Well-dispersed nanoparticles have also been shown to promote the formation of



Nanoengineered Concrete, Fig. 2 Particle size and specific surface area related to concrete materials (Adapted from [1])

smaller-sized crystals (e.g., $Ca[OH]_2$ and AFm), thus densifying the microstructure, and to promote crack arrest and interlocking, improving the material mechanical properties.

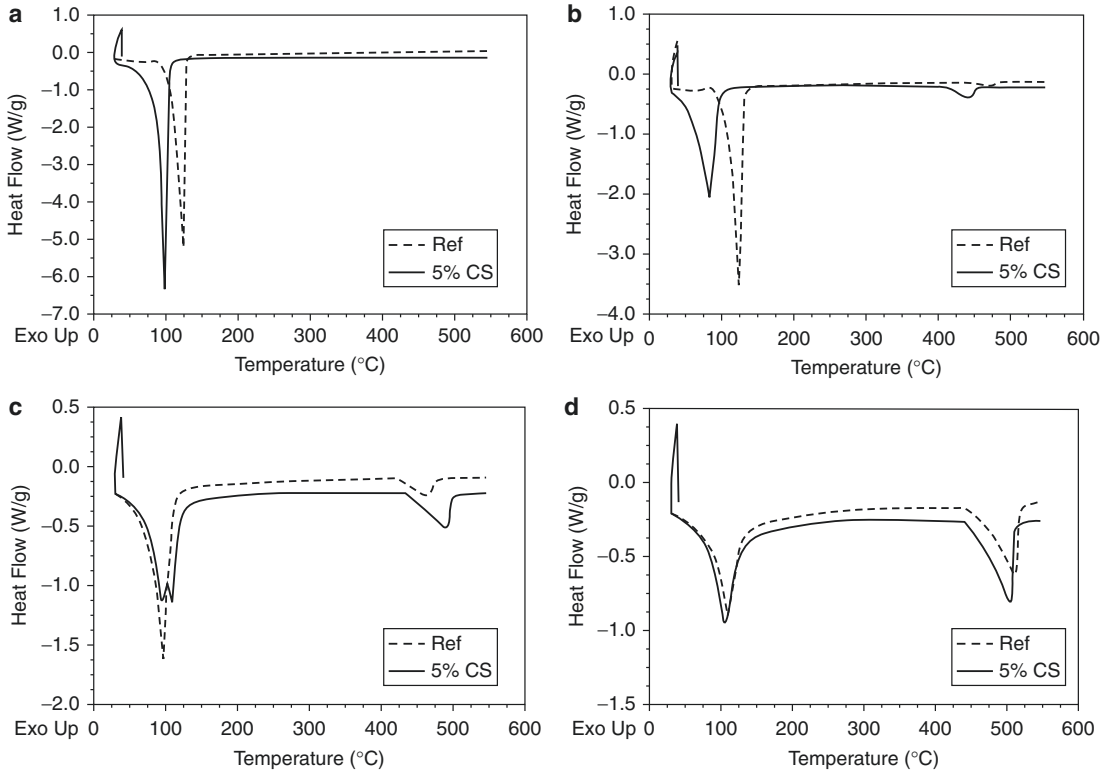
Most of the research on the addition of nanoparticles to cementitious systems has been performed using nano-SiO_2 [1, 2, 14–20]. Nano-SiO_2 has been found to accelerate the hydration reactions of C_3S, improve concrete workability and early age strength, be more efficient in enhancing strength than silica fume, and improve the bond between the aggregates and the cement paste [14–20]. The following provides a few examples of recent studies in this area highlighting some of the effects of nano-SiO_2 addition.

The accelerating effect of nano-SiO_2 (colloidal silica, CS) on the hydration of C_3S (alite, Ca_3SiO_5) has been reported by Björnström et al. [17]. Differential scanning calorimetry (DSC) for samples containing up to 5 % of colloidal silica (Fig. 3) showed the initial dependence of C-S-H formation on the addition of colloidal silica and suggested that colloidal silica mostly affects the initial silica polymerization rates rather than the amount of product ultimately formed [17].

Porro et al. [16] reported that the compressive strength of cement pastes increased with the reduction in particle size of nano-SiO_2. The strength improvement was attributed to the formation of larger silicate chains of the C-S-H gel in cement pastes containing nano-SiO_2.

In contrast to silica fume (SF), the addition of nano-SiO_2 at dosages of 1–5 % has been reported by Qing et al. [18] to increase the thickness and viscosity of superplasticized cement pastes and shorten the initial and final setting times. Nano-SiO_2 significantly improved the compressive strength in the early stages of hardening. After 3 days of curing, the cement pastes with 5 % nano-SiO_2 were 41 % stronger than the reference pastes. At 28 days, the strength was improved by 25 %. Without exception, the strength of the cement pastes containing nano-SiO_2 improved with increasing dosage of the nanoadditive.

The addition of nano-SiO_2 has been shown to significantly improve the pozzolanic activity of



Nanoengineered Concrete, Fig. 3 Differential scanning calorimetry of reference (Ref.) and samples containing 5 % of CS at different hydration times: (**a**) 1 h hydration; (**b**) 4 h hydration; (**c**) 8 h hydration; and (**d**) 24 h hydration [20]

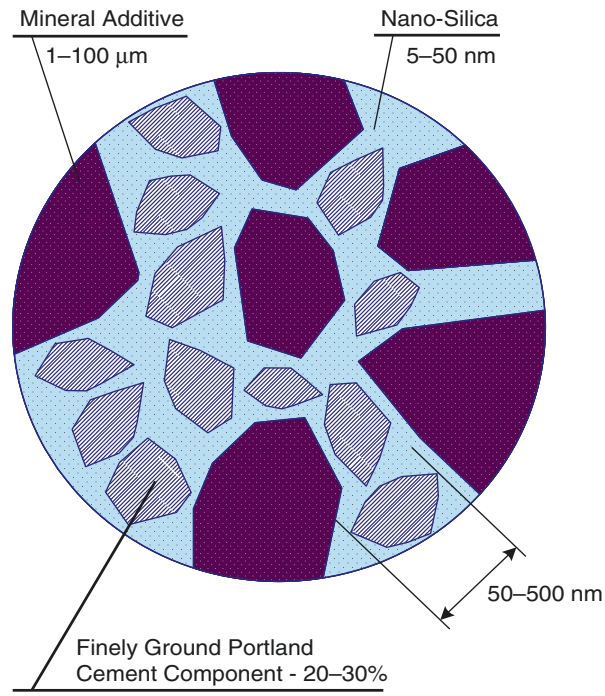
N

fly ash. In a study performed by Li [19], the addition of 4 % nano-SiO_2 to concretes prepared using 50 % type F fly ash, a water-to-cement ratio of 0.28, and a superplasticizer dosage of 2 % (by weight of binder) helped to achieve reaction levels typically obtained at 6 months after only 24 days. The pozzolanic reaction of nano-SiO_2 was shown to be very quick, reaching 70 % of its ultimate value (achieved by "common" pozzolans at the age of 2 years) within 3 days and ~95 % in a 2-week period. In contrast, plain fly ash reacted with lime very slowly, reaching a significant hydration degree only at the age of 2 years.

The addition of nano-SiO_2 at a dosage of 1–2 % has been shown by Collepardi et al. [14] to improve the performance of low-heat self-compacting concrete (SCC). The addition of nano-SiO_2 to blended blast furnace slag concretes made the mixtures more cohesive and reduced bleeding and segregation but had little effect on the slump loss measured within the first 30 min. The best performance was obtained for the blended blast furnace slag concrete prepared using ground fly ash (GFA) as mineral additive, 2 % nanosilica, and 1.5 % superplasticizer with a compressive strength of 55 MPa at the age of 28 days, low bleeding, higher slump flow, and very little slump loss.

The addition of nano-SiO_2 has also been reported by Hosseini et al. [20] to improve the interfacial transition zone (ITZ) by combining with calcium hydroxide crystals to produce C-S-H leading to a denser, more uniform ITZ. This resulted in a significant improvement of the compressive strengths of recycled aggregate concretes.

As an alternative to nanoparticle addition, the idea of nanobinder was recently proposed to improve concrete performance [1]. In nanobinder, the mineral additives act as the primary components, and nanosized cement is used to fill the gaps between the particles (Fig. 4). The nanosized cement thus acts as a glue to bind the less reactive



Nanoengineered Concrete, Fig. 4 The concept of nanobinder (nanoengineered cement) (Adapted from [1])

mineral additives together. The nanosized cement can be obtained either by colloidal milling of conventional (optionally high C_2S) portland cement clinker (top-down approach) or by self-assembly (bottom-up approach) through mechanochemically induced topochemical reactions [1].

Tailoring of Nanoparticles for Optimal Performance

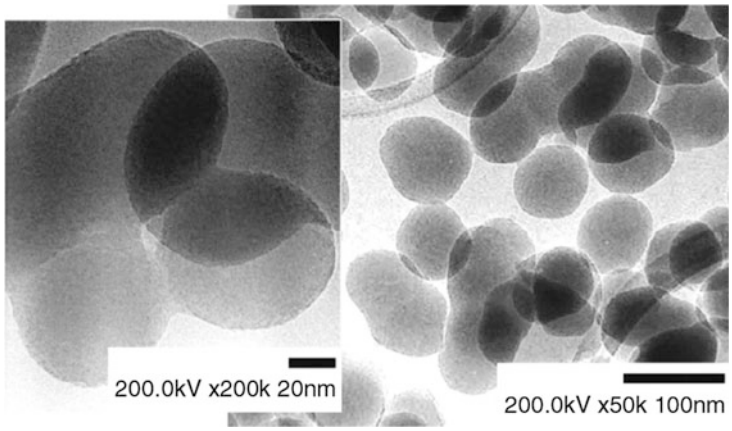
While small quantities of nanoparticles (less than 1 % by weight of cement) are generally sufficient to improve the performance of composites [2, 14–20], practical application in concrete requires manufacturing a considerable amount of these particles [5–7]. Being able to produce nanoparticles in large quantities at a low cost is, therefore, a paramount. Several technologies have been used for the production of nanoparticles [5–7]. The sol–gel process is the most commonly used for the production of nanosilica (Fig. 5). The chemical procedure consists of the formation of a colloidal suspension (sol) and gelation of the sol to form a network in a continuous liquid phase (gel). The procedure can be summarized as follows [7, 15]:

1. Hydrolysis of the precursor
2. Condensation and polymerization of monomers to form the particles
3. Growth of the particles
4. Agglomeration of the particles, followed by the formation of a network and gel structure
5. Drying (optional) to remove the solvents and thermal treatment (optional) to remove the surface functional groups and obtain the desired crystal structure

Trymethylethoxysilane or tetraethoxysilane (TMOS/TEOS) is typically applied as the precursor in the production of nanosilica, and the following reaction takes place [7, 15]:

$$n\mathrm{Si}(\mathrm{OC_2H_5})_4 + 2n\mathrm{H_2O}\xrightarrow[\mathrm{NH_3}]{\mathrm{C_2H_5OH}} n\mathrm{SiO_2} + 4n\mathrm{C_2H_5OH}$$

pH, temperature, concentration of reagents, H_2O/Si molar ratio, and type of catalyst are important



Nanoengineered Concrete, Fig. 5 The morphology and particle size distribution of nano-SiO_2 (TEM) synthesized by the sol–gel method [15]

Nanoengineered Concrete, Table 1 Design and properties of nano-SiO_2 [15]

Specimen type[a]	Molar ratio TEOS/Etanol/H_2O	Reaction time, hours	Particle size (TEM), nm	Surface area (BET), m^2/kg
1B3	1/24/6	3	15–65	116,000
2B3	1/6/6	3	30, 60–70	145,000
3B3	1/6/24	3	15–20	133,000
4B3	1/24/24	3	5	163,000
1A3	1/24/6	3	5	510,100
2A3	1/6/6	3	<10	263,500
3A3	1/6/24	3	<10, 17	337,100
4A3	1/24/24	3	5	382,200

[a]Sample coding

ABC
- First number denotes molar ratio combination as per as experimental matrix
- Last number corresponds to the reaction time in hours
- Letter – denotes reaction media: A-for acid (pH=2) and B-for base (pH=9)

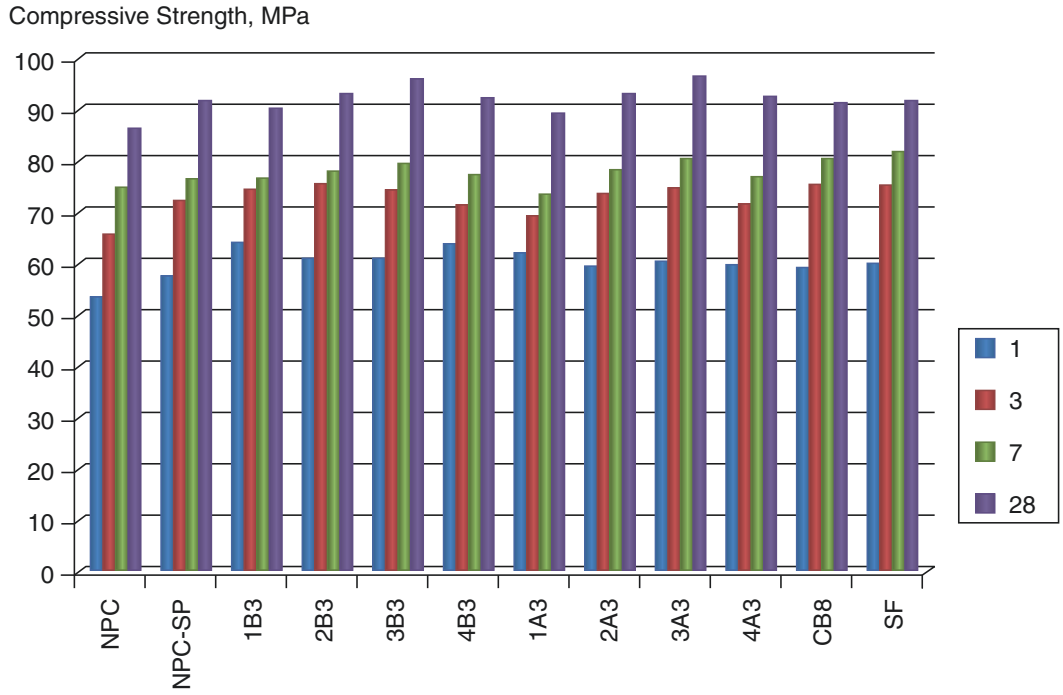
parameters that can affect the production and characteristics of the nanoparticles [7, 15].

Flores et al. [15] used the sol–gel process to synthesize nano-SiO_2 with a size range of 5–100 nm (Fig. 5).

Different experimental conditions were examined as summarized in Table 1. X-ray diffraction study showed that the obtained nano-SiO_2 was a highly amorphous material with predominant crystallite size of 1–2.5 nm [15]. The addition of 0.25 % (by weight of cementitious material) of the synthesized nano-SiO_2 improved the early strength of portland cement mortars with a 17 % strength increase at 1 day. The 28-day compressive strength was improved by as much as 10 % and the flexural strength by as much as 25 % [15]. Proper dispersion of the nano-SiO_2 within the cement paste is critical for improved performance [1, 15]. The use of polycarboxylate acrylic ester (PAE) superplasticizer, ultrasonification, and/or high-speed mixing has proven to be efficient in the dispersion of nano-SiO_2 [15]. The addition of 0.1 % superplasticizer and 0.25 % nano-SiO_2 synthesized using the sol–gel process has shown a 20 % increase in the compressive strength of cement mortars (Fig. 6).

Developing New Functionalities

Photocatalytic TiO_2 has been used in white cement-based concrete, providing the material with self-cleaning and air purification capabilities [12, 21]. It is important for architectural concrete



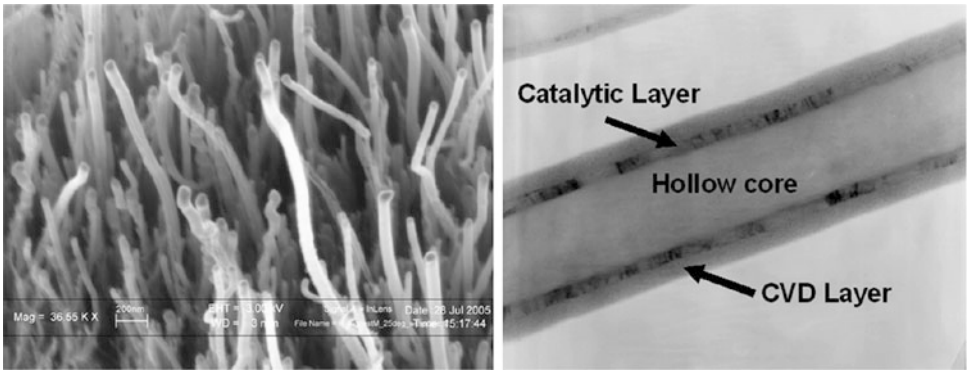
Nanoengineered Concrete, Fig. 6 Compressive strength of superplasticized mortars with nano-SiO_2 at different ages (NPC and NPC-SP are reference portland cement and superplasticized mortars; nanoparticle specification as per as Table 1, Fig. 5; CB-8 and SF are cembinder-8 and silica fume) [15]

to maintain its aesthetic and decorative characteristics such as color over its entire service life, even in highly polluted urban environments. Photocatalytic materials provide a smart solution to effectively minimize contaminant buildup at the surface of architectural concrete through photocatalytic oxidation. One of the most popular photocatalytic materials used in cement is TiO_2. The anatase polymorph of TiO_2 (especially nanosized) is the most effective TiO_2-based photocatalytic material, outperforming rutile in cement-based materials [12, 21]. Photocatalytic nano-TiO_2 embedded into cement matrix has also been reported to be very effective for NO_x abatement [21].

Application of Carbon Nanotubes and Carbon Nanofibers

Due to extraordinary strength and tensile modulus, carbon nanotubes and carbon nanofibers (CNTs/CNFs) are potential candidates for use as nanoreinforcement in cement-based materials [1, 2, 5, 6, 22–26]. Vapor-grown CNTs are concentric cylinders made from rolled-up planes of hexagonally bound carbon atoms. Vapor-grown CNFs (Fig. 7) exhibit a "herringbone" or cup-stacked structure of nested graphene layers with a hollow core. The production of CNTs on an industrial scale remains expensive. In contrast, CNFs have a lower production cost and can be made at a higher production capacity [27]. The use of CNTs/CNFs in cement remains limited [2]. The proper dispersion of CNTs/CNFs within the cement paste is one of the main challenges [2, 22–26] due to their high hydrophobicity and strong self-attraction. A number of methods have been used to improve the dispersion of CNTs/CNFs in cement, including the use of surfactants, surface treatment with acids, high shear mixing, and sonication [22–26].

To date mixed performance results have been achieved ranging from significant improvements for small addition of CNTs/CNFs to minimal or no improvement. Among the promising results,



Nanoengineered Concrete, Fig. 7 SEM (*left*) and TEM (*right*) image of carbon nanofibers [27]

Makar et al. [23] demonstrated that CNTs can affect early age hydration and indicated the potential of CNTs for crack bridging and enhanced stress transfer. Shah et al. [24] showed that, after dispersion in water using surfactant and ultrasonic energy, small amounts of CNTs (0.048 % and 0.08 %) can result in a significant (∼50 %) increase in the Young's modulus of cement pastes.

Smart Stress-Sensing Composites

The incorporation of conductive nanomaterials (CNTs/CNFs, nano-Fe_2O_3, etc.) into cement-based composites has been reported to provide the materials with self-sensing capability [28, 29]. It has been shown that carbon fiber-reinforced cement may function as a piezoresistive strain sensor [28]. Due to the high conductivity of the carbon fibers compared to the cement matrix, the introduction of the fibers into cement decreases the electrical resistivity of the material. Strain sensitivity (i.e., fractional change in electrical resistance per unit strain) of as much as 700 has been obtained with incorporation of 0.2–0.5 % (by volume) carbon microfibers [28]. Recently, a self-sensing cement material with multiwalled carbon nanotube (MWCNT) has been proposed for traffic monitoring [29]. The piezoresistive properties of the MWCNT-cement composite enabled the detection of mechanical stresses induced by traffic flow.

Future Developments

Concrete science is expected to benefit significantly from the rapidly growing field of nanotechnology. Advances in the characterization of the nanoscale structure of concrete are providing new insights about concrete. New developments are underway in the nanoengineering of concrete. Some of the most promising include the development of [1]

- Highly efficient superplasticizers for concrete and high-strength fibers with exceptional energy-absorbing capacity
- Concrete with self-cleaning properties, discoloration resistance, and antigraffiti protection, based on photocatalytic technology
- Reinforced binders with nanoparticles or nanotubes
- Concrete with smart properties such as temperature-, moisture-, or stress-sensing capability; deformative properties; and nonshrinking and low thermal expansion
- Ecobinders modified with nanoparticles and produced with significantly reduced amount of portland cement
- Ecobinders with nanoparticles based on alternative systems (MgO, phosphate, geopolymers, gypsum)
- Self-healing concrete with controlled supply of internal moisture or solution of repair/bonding agent

Conclusions

Nanotechnology in concrete is still in its exploratory phase, and full-scale applications in the field of construction are to date very limited. Nevertheless, the potential for nanotechnology to help improve the performance and durability of one of the oldest known construction materials is most promising.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Carbon Nanotubes
- ▶ Nanoparticles
- ▶ Sol–Gel Method
- ▶ Spider Silk

References

1. Sobolev, K., Ferrada-Gutiérrez, M.: How nanotechnology can change the concrete world: part 2. Am. Ceram. Soc. Bull. **11**, 16–19 (2005)
2. Sanchez, F., Sobolev, K.: Nanotechnology in concrete – a review. Constr. Build. Mater. **24**(11), 2060–2071 (2010)
3. Plassard, C., Lesniewska, E., Pochard, I., Nonat, A.: Investigation of the surface structure and elastic properties of calcium silicate hydrates at the nanoscale. Ultramicroscopy **100**(3–4), 331–338 (2004)
4. Scrivener, K.L.: Nanotechnology and cementitious materials. In: Bittnar, Z., Bartos, P.J.M., Nemecek, J., Smilauer, V., Zeman, J. (eds.) Nanotechnology in Construction: Proceedings of the NICOM3, 3rd International Symposium on Nanotechnology in Construction, pp. 37–42. Prague (2009)
5. Bhushan, B. (ed.): Handbook of Nanotechnology. Springer, Berlin (2004)
6. Atkinson, W.I.: Nanocosm – Nanotechnology and the Big Changes Coming from the Inconceivably Small, pp. 36–39. AMACOM, New York (2003)
7. Wilson, M., Smith, K.K.G., Simmons, M., Raguse, B.: Nanotechnology – Basic Science and Emerging Technologies. Chapman and Hall/CRC, Boca Raton/London (2000)
8. Mart, J., Mijangos, C.: Tailored polymer-based nanofibers and nanotubes by means of different infiltration methods into alumina nanopores. Langmuir **25**(2), 1181–1187 (2009)
9. Dalton, A.B.., Collins, S., Razal, S.J., Munoz, E., Ebron, V.H., Kim, B.G., Coleman, J.N., Ferraris, J.P., Baughman, R.H.: Continuous carbon nanotube composite fibers: properties, potential applications, and problems. J. Mater. Chem. **14**, 1–3 (2004)
10. Trtik, P., Bartos, P.J.M.: Nanotechnology and concrete: what can we utilise from the upcoming technologies? In: Proceeding of the 2nd Anna Maria Workshop: Cement & Concrete: Trends & Challenges, pp. 109–120. Holmes Beach (2001)
11. Beaudoin, J.J.: Why engineers need materials science. Concr. Int. **21**(8), 86–89 (1999)
12. Watanabe, T., Kojima, E., Norimoto, K., Kimura, T., Machida, M., Hayakawa, M., Kitamura, A., Chikuni, M., Saeki, Y., Kuga, T., Nakashima, Y.: Multifunctional material with photocatalytic functions and method of manufacturing same. US Patent US6294247 (2001)
13. Corradi, M., Khurana, R., Magarotto, R.: Controlling performance in ready mixed concrete. Concr. Int. **26**(8), 123–126 (2004)
14. Collepardi, M., Ogoumah-Olagot, J.J., Skarp, U., Troli, R.: Influence of amorphous colloidal silica on the properties of self-compacting concretes. In: Proceedings of the International Conference, Challenges in Concrete Construction – Innovations and Developments in Concrete Materials and Construction, pp. 473–483. Dundee (2002)
15. Flores, I., Sobolev, K., Torres, L.M., Valdez, P.L., Zarazua, E., Cuellar, E.L.: Performance of cement systems with Nano- SiO_2 particles produced using sol-gel method. In: TRB First International Conference in North America on Nanotechnology in Cement and Concrete. Irvine (2010)
16. Porro, A., Dolado, J.S., Campillo, I., Erkizia, E., de Miguel, Y., Sáez de Ibarra, Y., Ayuela, A.: Effects of nanosilica additions on cement pastes. In: Ravindra, R. K., Dhir, K., Newlands, M.D., Csetenyi, L.J. (eds.) Applications of Nanotechnology in Concrete Design. Thomas Telford, London (2005)
17. Bjornstrom, J., Martinelli, A., Matic, A., Borjesson, L., Panas, I.: Accelerating effects of colloidal nano-silica for beneficial calcium-silicate-hydrate formation in cement. Chem. Phys. Lett. **392**(1–3), 242–248 (2004)
18. Qing, Y., Zenan, Z., Deyu, K., Rongshen, C.: Influence of nano-SiO_2 addition on properties of hardened cement paste as compared with silica fume. Constr. Build. Mater. **21**(3), 539–545 (2007)
19. Li, G.: Properties of high-volume fly ash concrete incorporating nano-SiO_2. Cem. Concr. Res. **34**, 1043–1049 (2004)
20. Hosseini, P.B., Booshehrian, A., Delkash, M., Ghavami, S., Zanjani, M.K.: Use of nano-SiO_2 to improve microstructure and compressive strength of recycled aggregate concretes. In: NICOM3, 3rd International Symposium on Nanotechnology in Construction. Prague (2009)
21. Cassar, L., Pepe, C., Tognon, G., Guerrini, G.L., Amadelli, R.: White cement for architectural concrete possessing photocatalytic properties. In: Proceedings of the 11th International Congress on the Chemistry of Cement (ICCC), p. 2012. Durban (2003)

22. Sanchez, F.: Carbon nanofiber/cement composites: challenges and promises as structural materials. Int. J. Mater. Struct. Integr. **3**(2–3), 217–226 (2009). Special issue on nanotechnology for structural materials
23. Makar, J.M., Margeson, J., Luh, J.: Carbon nanotube/cement composites – early results and potential applications. In: Proceedings of 3rd International Conference on Construction Materials: Performance, Innovations and Structural Implications, p. 1–10. Vancouver (2005)
24. Shah, S.P., Konsta-Gdoutos, M.S., Metaxa, Z.S., Mondal, P.: Nanoscale modification of cementitious materials. In: Bittnar, Z., Bartos, P.J.M., Nemecek, J., Smilauer, V., Zeman, J. (eds.) Nanotechnology in Construction: Proceedings of the NICOM3, 3rd International Symposium on Nanotechnology in Construction, pp. 125–130. Prague (2009)
25. Gay, C., Sanchez, F.: Performance of carbon nanofibers/cementitious composites with a high-range water-reducer. J. Transport. Res. Board **2142**, 109–113 (2010). Nanotechnology in Cement and Concrete, volume 2
26. Sanchez, F., Ince, C.: Microstructure and macroscopic properties of hybrid carbon nanofibers/silica fume cement composites. Compos. Sci. Technol. **69**(7–8), 1310–1318 (2009)
27. Pyrograf-III carbon nanofiber. http://pyrografproducts.com/Merchant5/merchant.mvc?Screen=cp_nanofiber. Accessed 25 May 2011
28. Chung, D.: Piezoresistive cement-based materials for strain sensing. J. Intell. Mater. Syst. Struct. **13**(9), 599–609 (2002)
29. Li, H., Xiao, H.G., Ou, J.P.: A study on mechanical and pressure-sensitive properties of cement mortar with nanophase materials. Cem. Concr. Res. **34**, 435–438 (2004)

Nanoengineered Hydrogels for Cell Engineering

Esmaiel Jabbari
Biomimetic Materials and Tissue Engineering Laboratory, Department of Chemical Engineering, Swearingen Engineering Center, University of South Carolina, Columbia, SC, USA

Synonyms

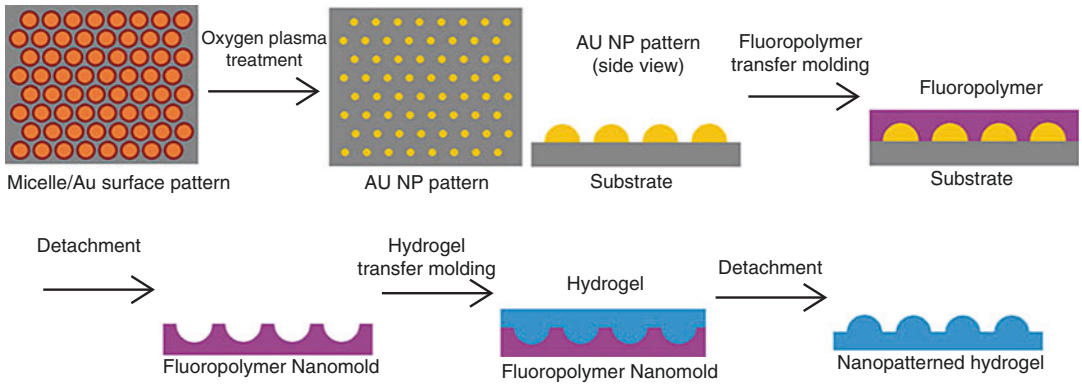
Nanostructured hydrogels; Patterned hydrogels

Definition

Nanoengineered hydrogels are soft, high water content, hydrophilic networks with nanoscale topographical or biomolecular patterns for mimicking the extracellular matrix of tissues in order to control cell fate and function.

Introduction

Hydrogels are three-dimensional (3D) cross-linked polymeric networks that retain a significant fraction of water in their structure without dissolving when placed in physiological solution. Due to their high water content, biomolecules like peptides and proteins and cells immobilized in hydrogels are more likely to retain their biological activity. Poly(ethylene glycol) (PEG)-based hydrogels, due to minimal interaction of the ethylene oxide repeating unit with biomolecules, do not elicit an immune response. Therefore, PEG hydrogels are used extensively for cell immobilization and encapsulation. Furthermore, PEG networks can be conjugated with bioactive molecules to produce cell-responsive hydrogels for applications in tissue engineering and regenerative medicine [1]. The cell fate including shape, morphology, phenotype, and function is determined by interaction with soluble and insoluble factors in the extracellular matrix (ECM) [2]. These interactions are between the functional motifs of the ECM proteins and binding sites of the glycoproteins embedded in the cell's plasma membrane. The functional motifs as well as geometrical and morphological features of the ECM are ordered at the nanoscale. As a result, synthetic hydrogels patterned with biologically active molecules or topographical nanostructures are important in understanding the role of ECM on cell fate, and the engineering of cell environment at the nanoscale. Patterning of hydrogels at the nanoscale includes surface techniques (2D patterning) and bulk/matrix techniques (3D patterning). The cell-nanostructure interaction is two dimensional for surface-patterned hydrogels while the interaction is three dimensional for matrix-patterned hydrogels. Surface patterning techniques include



Nanoengineered Hydrogels for Cell Engineering, Fig. 1 Schematic diagram for fabrication of a surface nanopatterned hydrogel by micelle nanolithography (Refs. [7, 11])

electron beam patterning, micelle nanolithography, UV-assisted nanomolding, and nanostamping. Matrix techniques include polymer phase separation and multiphoton chemical patterning.

Electron Beam Patterning

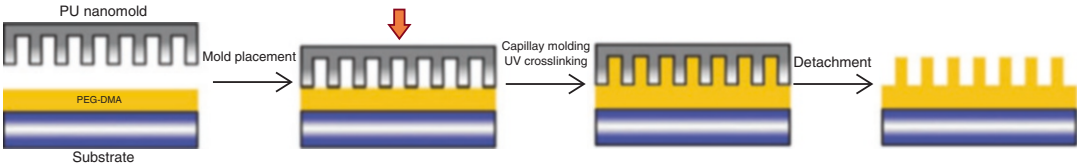
Hydrogels are surface-patterned at the nanoscale using a focused electron beam. In this method, a thin layer of hydrogel precursor solution, spin-coated on silicon wafers, is exposed to a focused electron beam using a modified field-emission scanning electron microscope [3]. The extent of cross-linking and density of the pattern can be controlled with this technique by varying the pixel dwell time to irradiation dose and interpixel spacing. After irradiation, the pattern is exposed by dissolving unexposed areas with organic solvents and rinsing in water. Topographical patterns with sizes ranging 50–1,000 nm can be produced with this technique. As an example, poly(sodium 4-styrenesulfonate-co-poly(ethylene glycol) methacrylate) (pSS-co-pPEGMA) with heparin-mimicking thiol groups was patterned with electron beam and the patterned gel is used for localized attachment of vasculogenic proteins bFGF and VEGF [4]. As another example, Boc-protected aminoxy tetraethylene glycol methacrylate was patterned with electron beam, the Boc-protecting group was removed to expose the aminoxy groups for localized grafting of proteins to the patterned hydrogel via oxime ligation [5].

Micelle Nanolithography

Self-assembled array of metal-impregnated diblock copolymers is used for nanoscale surface patterning of hydrogels. In this technique, a monolayer of polystyrene-block-poly(2-vinylpyridine)(PS-b-P2VP) and $HAuCl_4$ micelles is formed on a substrate (glass or silicon wafers) by dip-coating [6]. Next, the micellar array is treated with oxygen plasma to remove the PS-b-P2VP copolymer and exposing the ordered array of gold nanoparticles (NPs) on the substrate, as shown in Fig. 1. The pattern of gold NPs is moved to a nonadhesive fluorosiloxane layer by transfer-molding, followed by transfer to a hydrogel surface as shown in Fig. 1. In a variation of this technique, polystyrene homopolymer, as an ordering interference reagent, is mixed with PS-b-P2VP and $HAuCl_4$ micelles to produce a disordered array of gold NPs on the substrate [7], leading to a disordered nanopattern on the hydrogel surface. In another variation, $PtCl_4$ is used in the place of $HAuCl_4$ to form a nanoscale pattern of platinum NPs on the substrate [8]. The platinum pattern is moved to a PDMS substrate by transfer-molding, as shown in Fig. 2, to produce a catalytic nanostamp for chemical patterning of



Nanoengineered Hydrogels for Cell Engineering, Fig. 2 Schematic diagram for fabrication of PDMS nanostamp (Ref. [8])



Nanoengineered Hydrogels for Cell Engineering, Fig. 3 Schematic diagram for UV-assisted capillary nanomolding (Ref. [12])

hydrogel surfaces. In another interesting approach to generate micro- nanostructured hydrogels, an array of gold NPs on a glass substrate was coated with a photoresist layer [9]. After photolithography and removal of the uncovered gold NPs and the protecting resist layer, the microstructured nanoarray of gold NPs was covered with acrylated PEG hydrogel precursor solution and cross-linked by ultraviolet (UV) radiation. Upon detaching from the glass substrate, a hydrogel surface with microstructured array of gold NPs was generated which could potentially be used to study the modulation of nano- and microscale patterns of biomolecules on cell function.

Nanomolding and Nanostamping

The block-copolymer-templated micelle lithography is used to fabricate stamps and molds in sub-100 nanoscale for patterning [8]. In this approach, a silicon substrate is patterned with platinum (Pt) NPs by micelle lithography and the bare silicon surface is covalently covered with a fluorosiloxane monolayer. Next, a PDMS layer is spin-coated on the patterned substrate followed by curing. Upon peeling the PDMS layer, a Pt nanopatterned catalytic stamp and a silicon nanopatterned mold are produced. The catalytic stamp can be used to print nanoscale biomolecular patterns on hydrogels. The pattern on silicon/glass substrate can be moved to a polyurethane (PU), PDMS, or other soft hydrophobic substrates [10] for subsequent patterning of hydrogel surfaces [11]. As an example, in UV-assisted capillary lithography [12], a nanopatterned PU mold is placed over a thin layer of methacrylated PEG precursor solution to make conformal contact by capillary forces, as shown in Fig. 3. After cross-linking of the PEG layer by UV-assisted photopolymerization, the PU mold is peeled off to produce a PEG hydrogel with a surface pattern ranging 50–500 nm in size. Direct molding of a hydrogel precursor solution with nanotemplated quartz followed by cross-linking can also be used to form hydrogels with nanopatterned surface topography [13].

Copolymer Phase Separation

Phase separation of amphiphilic di- and triblock copolymers at the nanoscale is used to form hydrogels with less than 50 nm matrix nanopatterns. For example, amphiphilic triblock copolymers of poly(ε-caprolactone) and PEG were allowed to self-assemble in aqueous solution to form 30–40 nm ordered cylindrical structures [14]. Following electron beam cross-linking of the microphase separated mixture, the PCL phase was removed by hydrolysis to form a porous hydrogel, which can potentially be patterned in 3D with

cell-responsive functional groups. In a similar approach, microphase separation of a blend of polystyrene-block-poly(ethylene oxide) (PS-b-PEO) and arginine-glycine-aspartic acid (RGD)-conjugated PS-b-PEO was used to form nanoscale domains of RGD integrin-binding peptide in the hydrogel matrix [15]. By manipulating phase separation of the diblock copolymer, domain sizes of 8–14 nm, lateral spacing of 62–44 nm, and island densities of 450–900 per μm^2 were formed. In a related approach, a blend of methacrylated PEG-b-PLA-b-PEG (PLA denotes poly(lactic acid)), methacrylated biotin-PEG, and a self-assembling lyotropic liquid crystalline mesophase as the polymerization template was used to generate biotinylated hydrogels with lamellar matrix morphology [16]. The biotin-avidin chemistry was used to attach integrin-binding RGD peptide to the nanostructured matrix. In a similar approach, a mixture of blank alginate chains and integrin-binding RGD-coupled alginate chains was cross-linked in aqueous solution to form nanopatterned RGD islands in an alginate matrix [17]. Similarly, a mixture of thiol-derivatized hyaluronic acid (HA) and acrylated nanogels was cross-linked to form a uniformly distributed nanostructured HA hydrogel [18].

Cell Engineering with Nanopatterned Gels

When cardiomyocytes were seeded on PEG hydrogels patterned with 50–150 nm nanopillars, cell adhesion was significantly enhanced and formed 3D aggregates, which retained their conductive and contractile properties [12]. Neonatal rat ventricular myocytes (NRVMs) seeded on anisotropically nanofabricated hydrogel substrates, with ridges/grooves in the 50–800 nm range, had perfect alignment with the ridge/groove direction, the direction of cell contraction was closely aligned with the pattern direction, and the action potential direction was anisotropic, when compared to unpatterned hydrogel substrates [19]. The spreading of NIH-3T3 fibroblasts increased from 1,100 to 1,700 μm^2 as 3D lateral spacing of RGD integrin-binding domains in a PS-b-PEO hydrogel matrix was decreased from 62 to 44 nm [15]. The phosphorylation of Y397 focal adhesion kinase, cell spreading, and osteogenic differentiation of MC3T3 preosteoblasts depended on the spacing of RGD islands in an alginate matrix. Closer spacing of RGD islands favored phosphorylation of Y397 and cell spreading while wider spacing favored differentiation, and proliferation was a function of RGD density in the matrix [17]. Block copolymer micelle lithography followed by ligand conjugation was used to form a gradient in the spacing of cyclic-RGD peptide, specifically recognized by $\alpha_v\beta_3$ integrins, on a substrate [20]. The gradient in cyclic-RGD spacing induced cell orientation toward shorter spacing. In another study, integrin clustering on MC3T3-E1 osteoblasts depended on the local order of RGD ligand arrangement on the substrate for average ligand spacings >70 nm [7]. Cell adhesion was lower on substrates with ordered RGD pattern and higher on disordered patterns. Human embryonic stem cells seeded on polyurethane gels with 350-nm grooved arrays, fabricated by UV capillary force lithography, differentiated directly into the neuronal lineage in the absence of differentiation-inducing agents [21]. Submicron pattern of cell-repulsive PEG hydrogel on an adhesive substrate can discriminate between different cell types. At the highest PEG density with overlapping patterns, neither neurite cells nor astrocytes adhered to the substrate [22]. As the spacing between PEG patterns was increased to 2 μm, neurons were able to grow and form long axons while astrocytes were unable to grow. At higher PEG spacing of 10 μm, both neurons and astrocytes were able to grow on the patterned substrates. In an interesting study, poly (vinyl alcohol) (PVA) nanogels were used to form nanoarrays of tethered lipid bilayer rafts (tLBRMs) on a solid substrate to study the effect of cell membrane's lipid rafts on signal transduction [23]. In this work, PVA hydrogel was nanopatterned on a gold substrate using UV nanolithography to form selective nanowells. After covering the gold nanowells with a monolayer of vesicle rupturing promoter, tLBRMs were selectively immobilized on a gold substrate between the PVA nanogels by monolayer-assisted

vesicle fusion. These studies clearly demonstrate that engineered hydrogels can play a major role in understanding the effect of topography and ECM ligands, patterned at the nanoscale, on cell shape, morphology, phenotype, and fate.

Future Direction

A number of techniques such as electron beam patterning and micelle lithography have been developed for nanoscale patterning of hydrogel surfaces. However, there is a need to develop techniques for 3D patterning of hydrogels at the nanoscale, as the interaction of cells with ligands in the ECM is three dimensional. Related to that, the effect of surface pattern of adhesive ligands on cell shape and morphology has been studied extensively but there is a need to conduct similar studies in 3D within a nanopatterned hydrogel matrix. Those studies should be extended to the effect of 3D pattern of ECM ligands on cell differentiation, lineage selection, and fate.

Cross-References

- ▶ Bioadhesion
- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Biomimetic Synthesis of Nanomaterials
- ▶ Biomimetics
- ▶ BioPatterning
- ▶ Dip-Pen Nanolithography
- ▶ Electron Beam Lithography (EBL)
- ▶ Injectable Hydrogel
- ▶ MEMS on Flexible Substrates
- ▶ Nanoimprint Lithography
- ▶ Nanoimprinting
- ▶ Nanopatterned Surfaces for Exploring and Regulating Cell Behavior
- ▶ Nanoscale Printing
- ▶ Nanotechnology
- ▶ Nanotechnology in Cardiovascular Diseases
- ▶ Self-assembly
- ▶ Self-assembly of Nanostructures
- ▶ Smart Hydrogels
- ▶ Stereolithography

References

1. He, X.Z., Ma, J.Y., Jabbari, E.: Effect of grafting RGD and BMP-2 protein-derived peptides to a hydrogel substrate on osteogenic differentiation of marrow stromal cells. Langmuir **24**, 12508–12516 (2008)
2. Fisher, O.Z., Khademhosseini, A., Langer, R., Peppas, N.A.: Bioinspired materials for controlling stem cell fate. Acc. Chem. Res. **43**, 419–428 (2009)
3. Hong, Y., Krsko, P., Libera, M.: Protein surface patterning using nanoscale PEG hydrogels. Langmuir **20**, 11123–11126 (2004)
4. Christman, K.L., Vazquez-Dorbatt, V., Schopf, E., Kolodziej, C.M., Li, R.C., Broyer, R.M., Chen, Y., Maynard, H.D.: Nanoscale growth factor patterns by immobilization on a heparin-mimicking polymer. J. Am. Chem. Soc. **130**, 16585–16591 (2008)
5. Christman, K.L., Broyer, R.M., Schopf, E., Kolodziej, C.M., Chen, Y., Maynard, H.D.: Protein nanopatterns by oxime bond formation. Langmuir **27**, 1415–1418 (2010)
6. Blummel, J., Perschmann, N., Aydin, D., Drinjakovic, J., Surrey, T., Lopez-Garcia, M., Kessler, H., Spatz, J. P.: Protein repellent properties of covalently attached PEG coatings on nanostructured SiO_2-based interfaces. Biomaterials **28**, 4739–4747 (2007)
7. Huang, J., Grater, S.V., Corbellini, F., Rinck, S., Bock, E., Kemkemer, R., Kessler, H., Ding, J., Spatz, J.P.: Impact of order and disorder in RGD nanopatterns on cell adhesion. Nano Lett. **9**, 1111–1116 (2009)
8. Mizuno, H., Buriak, J.M.: Nanoscale patterning of organic monolayers by catalytic stamp lithography: scope and limitations. ACS Appl. Mater. Interfaces **1**, 2711–2720 (2009)
9. Aydin, D., Louban, I., Perschmann, N., Blümmel, J., Lohmüller, T., Cavalcanti-Adam, E.A., Haas, T.L., Walczak, H., Kessler, H., Fiammengo, R., Spatz, J.P.: Polymeric substrates with tunable elasticity and nanoscopically controlled biomolecule presentation. Langmuir **26**, 15472–15480 (2010)
10. Graeter, S.V., Huang, J., Perschmann, N., López-García, M., Kessler, H., Ding, J., Spatz, J.P.: Mimicking cellular environments by nanostructured soft interfaces. Nano Lett. **7**, 1413–1418 (2007)
11. Diez, M., Mela, P., Seshan, V., Möller, M., Lensen, M. C.: Nanomolding of PEG-based hydrogels with Sub-10-nm resolution. Small **5**, 2756–2760 (2009)
12. Kim, D.-H., Kim, P., Song, I., Cha, J.M., Lee, S.H., Kim, B., Suh, K.Y.: Guided three-dimensional growth of functional cardiomyocytes on polyethylene glycol nanostructures. Langmuir **22**, 5419–5426 (2006)
13. Gaston, A., Khokhar, A.Z., Bilbao, L., Sáez-Martínez, V., Corres, A., Obieta, I., Gadegaard, N., Nanopatterned, U.V.: UV curable hydrogels for biomedical applications. Microelectron. Eng. **87**, 1057–1061 (2009)
14. Kang, J., Beers, K.J.: Synthesis and characterization of PCL-b-PEO-b-PCL-based nanostructured and porous hydrogels. Biomacromolecules **7**, 453–458 (2006)

15. George, P.A., Doran, M.R., Croll, T.I., Munro, T.P., Cooper-White, J.J.: Nanoscale presentation of cell adhesive molecules via block copolymer self-assembly. Biomaterials **30**, 4732–4737 (2009)
16. Clapper, J.D., Pearce, M.E., Guymon, C.A., Salem, A. K.: Biotinylated biodegradable nanotemplated hydrogel networks for cell interactive applications. Biomacromolecules **9**, 1188–1194 (2008)
17. Comisar, W.A., Kazmers, N.H., Mooney, D.J., Linderman, J.J.: Engineering RGD nanopatterned hydrogels to control preosteoblast behavior: a combined computational and experimental approach. Biomaterials **28**, 4409–4417 (2007)
18. Bencherif, S.A., Washburn, N.R., Matyjaszewski, K.: Synthesis by AGET ATRP of degradable nanogel precursors for in situ formation of nanostructured hyaluronic acid hydrogel. Biomacromolecules **10**, 2499–2507 (2009)
19. Kim, D.-H., Lipke, E.A., Kim, P., Cheong, R., Thompson, S., Delannoy, M., Suh, K.-Y., Tung, L., Levchenko, A.: Nanoscale cues regulate the structure and function of macroscopic cardiac tissue constructs. Proc. Natl. Acad. Sci. U. S. A. **107**, 565–570 (2010)
20. Hirschfeld-Warneken, V.C., Arnold, M., Cavalcanti-Adam, A., López-García, M., Kessler, H., Spatz, J.P.: Cell adhesion and polarisation on molecularly defined spacing gradient surfaces of cyclic RGDfK peptide patches. Eur. J. Cell Biol. **87**, 743–750 (2008)
21. Lee, M.R., Kwon, K.W., Jung, H., Kim, H.N., Suh, K. Y., Kim, K., Kim, K.-S.: Direct differentiation of human embryonic stem cells into selective neurons on nanoscale ridge/groove pattern arrays. Biomaterials **31**, 4360–4366 (2010)
22. Krsko, P., McCann, T.E., Thach, T.-T., Laabs, T.L., Geller, H.M., Libera, M.R.: Length-scale mediated adhesion and directed growth of neural cells by surface-patterned poly(ethylene glycol) hydrogels. Biomaterials **30**, 721–729 (2009)
23. Lee, B.K., Lee, H.Y., Kim, P., Suh, K.Y., Kawai, T.: Nanoarrays of tethered lipid bilayer rafts on poly(vinyl alcohol) hydrogels. Lab Chip **9**, 132–139 (2009)

Nanofabrication

▶ Integration of Nanostructures Within Microfluidic Devices
▶ Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly

Nanofatigue Tester

▶ Nanoindentation

Nano-FET

▶ Nanostructure Field Effect Transistor Biosensors

Nanofibrous Materials and Composites

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

Nanofluidic Channels

▶ Nanochannels for Nanofluidics: Fabrication Aspects

Nanofluidics

▶ Applications of Nanofluidics
▶ Electrokinetic Fluid Flow in Nanostructures
▶ Integration of Nanostructures Within Microfluidic Devices

Nanofocusing of Light

▶ Light Localization for Nano-optical Devices

Nanogap Biosensors

J. Tanner Nevill and Daniele Malleo
Fluxion Biosciences, South San Francisco, CA, USA

Definition

A nanogap biosensor is an arrangement of two electrodes separated by no more than 300 nm that is used to electrically detect biologically relevant

materials, reactions, or interactions in solution. The limit of 300 nm is imposed because it represents the practical upper limit of the characteristic thickness of the electrical double layer, which forms at all charged surfaces in aqueous solutions.

Overview

Nanogap biosensors are made using various micro- and nano-fabrication techniques to create a nanometer-scale detection region. This region is interrogated using electrical techniques to sense and measure the presence or activity of a biomolecule. All-electrical measurements offer the promise of low cost, rapid, and label-free detection. Many arguments are made for using nanogap biosensors, including higher sensitivity, reduction of parasitic effects, overlapping double layers, and the use of molecular-sized cavities. Liang and Chou use a nanogap electrode arrangement inside of a nanofluidic channel like a pore for detecting (and eventually sequencing) single strands of DNA [1]. Im et al. exploit the fact that a molecule can occupy a significant portion of the nanogap, thereby making it very sensitive to low numbers of molecules: By functionalizing the surface of the nanogap, analyte binding significantly changes the electrical characteristic of the sensing electrodes [2].

The electrical double layer, a cloud of counterions that balances out the charge present on any surface in solution, is often seen as a parasitic component of electrical measurements: This layer shields the bulk solution from the induced electric field thus decreasing the sensitivity of alternating current detection techniques. Because nanogaps have dimensions similar or smaller than the double layer thickness, the electric potential in a nanogap distributes differently than in a macroscale gap. In fact, nanogap devices typically create environs where the volume of interest consists exclusively of double layer regions, actually eliminating the bulk solution from the measurement. Elimination of the bulk response is desirable in many situations, where measurement and quantification of very few or even single molecule events is of interest, making the concept of overlapping double layers in nanogaps very compelling.

The sensors developed thus far can be classified by the geometry of the detection region. According to the number of dimensions along which the sensing cavity exists, there are three categories as depicted in Fig. 1: 1D nanogaps refer to point gap junctions, 2D devices consist of coplanar electrodes or "fractured planes," and 3D nanogaps are "nanocavities," where the electrode surfaces forming the cavity have a significant area relative to the gap length.

As form follows function, the measurement techniques and applications can be correlated with the geometry. 1D and 2D nanogaps are typically interrogated with DC voltages to produce a resistive measurement, while 3D nanogaps are mostly interrogated using AC measurement techniques, to measure a complex impedance response. Consequently, the discussion of nanogap biosensors is divided below into two sections.

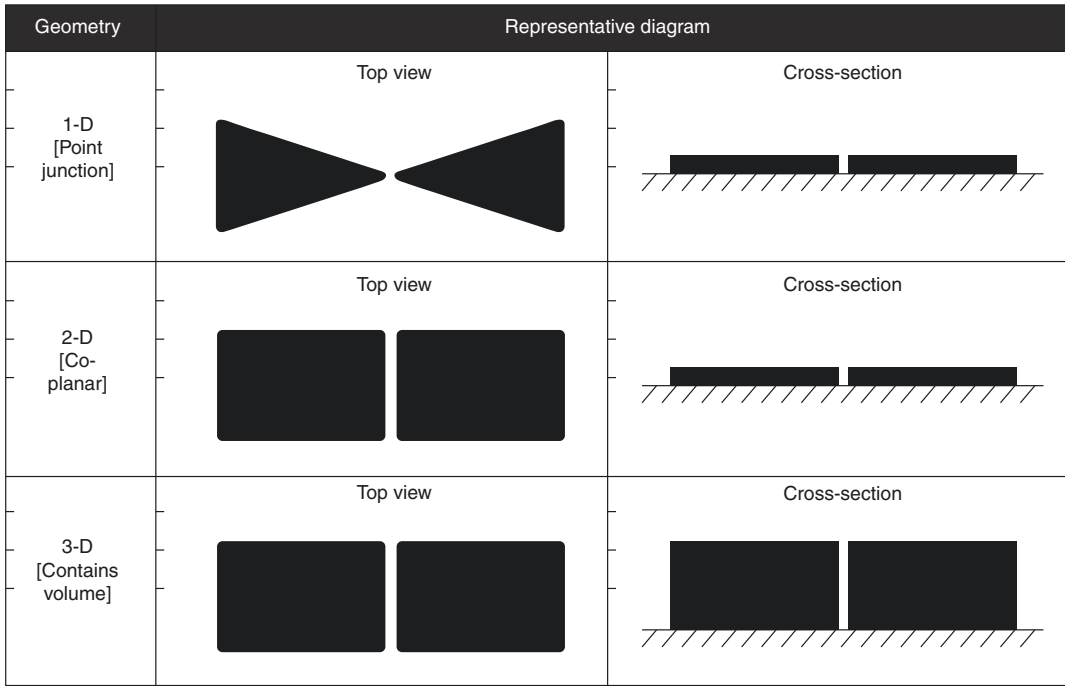
One- and Two-Dimensional Nanogap Biosensors

Advantages and Disadvantages

The key advantage of one- and two-dimensional nanogaps as biosensors is the opportunity for surface-bound molecules to occupy a significant portion of the interelectrode spacing, thereby increasing sensitivity, even enabling single molecule detection [3].

The main drawback of these sensors is the difficulty in ensuring specificity in the measurements. For this reason, the electrodes in the device need to have been functionalized appropriately before sample introduction, to enable the determination of presence or absence of specific molecules quickly and with off-the-shelf electrical meters: IV (current–voltage) meters, electrochemical measurement stations, or patch clamp amplifiers.

From a fabrication point of view, while it is in principle very easy to fabricate one-dimensional nanogap electrodes by electromigration, repeatability of this process can be an issue; conversely, fabrication via electron beam lithography allows for extreme control and repeatability but has slower throughput and higher cost.



Nanogap Biosensors, Fig. 1 Definition of the three categories of nanogap geometry. Form follows function and each geometry is typically tuned for a different function

Methods of Fabrication

One-dimensional nanogap structures are most commonly realized using electromigration, electron-beam lithography, or a combination of the two techniques. In some cases to shrink gaps to nanometer dimensions, methods such as surface-catalyzed metal deposition or dielectrophoretic transport of metal nanoparticles are used. The latter methods are advantageous as they enable fabrication of many nanogap electrodes at once. On the other hand, electron beam lithography offers the elegant advantage of directly writing nanoscale features in photoresist, which, incorporated with ion milling, lift-off, or unique metal deposition techniques, allows for the creation of nanogap electrodes. Electromigration is an intriguing technique, because it was an originally undesirable effect often encountered in the electronics industry. As metal lines got smaller and smaller, the relatively higher current densities would cause breaks to occur. These breaks were a result of a momentum transfer between the electrons comprising current and the atoms in the metal lattice [4]. Breaks tended to happen at grain boundaries or along defects in the lattice, and were highly dependent on temperature. Researchers interested in molectronics (a field of study whereby individual molecules are used to perform electronic functions) recognized that this technique could be used to create nanogap electrode junctions. This method offers the ability to get very small point gaps in an inexpensive manner, but there is a notable lack of control. Numerous groups have made strides in creating feedback control for electromigrated nanogap devices in order to reliably control the gap sizes in addition to offering parallel fabrication methods [5]. Another method is the chemical deposition of metal or electroplating which allows for the creation of nanogap devices using traditional fabrication methods that would not normally be capable of nanoscale dimensions. Lastly, a few groups have used either mechanical manipulation of electrodes to create nanogap devices or have mechanically broken wires to create nanoscale gaps [6].

Two-dimensional nanogap devices are similarly realized using metal deposition over a sacrificial linear feature to create a linear nanogap. Oblique metal deposition is a simple way to create a metal overhang in order to form a nanogap [7]. Similar to one-dimensional nanogap sensors, electron beam lithography is a suitable alternative for 2D nanogap fabrication as well as the use of sacrificial layers [8].

Measurement Methods

Direct Current Measurements

The most elementary way of acquiring data from a nanogap biosensor is to apply a DC (direct current) potential, while measuring current flow through the sensor electrodes: When the gap is "open," no current flows through except for a small tunneling current. When a biomolecule, or other sample, is attached to both the electrodes or otherwise closes the circuit, a larger direct current starts to flow. Specificity in the measurement can be attained by appropriate functionalization of the electrodes' surface.

In the case of measurements carried out in solution a background current is continuously measured due to ion transport in the solution. Sample attachment to the electrodes is not necessary, and just proximity of charged species (such as DNA molecules) can be detected by measuring increases in current.

Ohm's law, $I = \frac{V}{R}$, is in most cases sufficient to interpret the data acquired.

The tunneling current by itself can be used to characterize the device and estimate the distance between two electrodes, if the gap is too small to be detected by optical or electron microscopy.

The equation describing the tunneling current between two electrodes is:

$$I = \frac{K_1 A}{S^2}\left[X^2 \exp(-k_2 X) - Y^2 \exp(-k_2 s Y)\right] \quad (1)$$

where $\sqrt{\phi - \frac{V}{2}}$, $Y = \sqrt{\phi + \frac{V}{2}}$, $k_1 = 6.32 \times 10^{10} V s^{-1}$, $k_2 = 1.025 J^{-\frac{1}{2}}$, A is the emission area in cm^2, s is the electrode separation or distance in Å, ϕ is the barrier height, and V is the voltage in Volts [9].

Current–Voltage (IV) Measurements

To determine the behavior and characteristics of the molecule under study, or the sensor itself, DC potential can be swept in incremental steps from a negative to a positive potential and the resultant current traces drawn: Asymmetries and/or nonlinearities in the measured IV traces can be used to extrapolate characteristics of the sample under study.

Three-Dimensional Nanogap Biosensors

Advantages and Disadvantages

Three-dimensional nanogap devices are significantly different from one- and two-dimensional nanogaps in fabrication methods, measurement techniques, and in many cases, applications. 3D nanogaps contain a significant volume relative to the nanogap width. This nanocavity allows for both surface bound and free molecules to be monitored. Alternating current measurement techniques are typically used (e.g., dielectric or impedance spectroscopy), which provide data from which meaningful biological parameters can be extrapolated (e.g., binding efficiency of biomolecules, hydrodynamic radius of proteins).

The key advantages of nanogap biosensors using impedance detection techniques (e.g., impedance spectroscopy) are a homogeneous interrogation electric field and the opportunity for surface-bound molecules to occupy a significant portion of the detection volume, thereby increasing sensitivity.

The disadvantages of these devices include complex fabrication, difficulty in coupling to external equipment without introducing parasitic components, and the need for on-sensor reference electrodes and controls.

Methods of Fabrication

Three-dimensional nanogap devices are almost exclusively fabricated by provisionally incorporating a sacrificial layer of nanoscale dimensions, which is used to separate two conducting (metal or semiconductor) electrodes. When this sacrificial layer is removed or etched away, in the final steps of the fabrication protocol, a nanocavity is created.

A common technique is to use a thermally grown oxide layer, which creates a precise thickness by tightly controlling growth conditions. This layer can then later be fully or partially removed with a wet etch.

The materials used to form the electrodes vary, but the significant majority of nanogap devices use either gold or semiconductor electrodes. Gold has numerous advantages over semiconductors for many of these devices, including biocompatibility, ease of surface modification (through thiol-based chemistries), stability (resistance to oxidation), and off-the-shelf availability. One disadvantage of gold is the lack of compatibility with some fabrication processes. In such cases semiconductor electrodes have been used, despite their inherent shortcomings: mediocre conductivity and susceptibility to spontaneous surface oxidation [10].

Overlapping Electrical Double Layers

In order to understand why 3D nanogap devices are so interesting as biosensors, the topic of overlapping electrical double layers must be covered. The electrical double layer at a single electrode surface is discussed in depth by Lyklema [11]. At a charged surface, electromigration of charged species balances with diffusion to create a diffuse layer of counterions that shields the bulk solution from the electrode's potential. Inside a nanogap, the situation is much different: The double layers can occupy the entire volume. The presence of the bulk solution, which can be measured in terms of parasitic capacitance and resistance, is virtually nonexistent. Depending on the solution's properties, the double layers can actually "overlap." This term has been used to describe the situation where the Debye length is greater than the distance between the electrodes. The Debye length (λD) is the characteristic length scale of the double layer thickness, and is defined for a symmetric electrolyte:

$$\lambda_D = \sqrt{\frac{\varepsilon_0 \varepsilon RT}{2z^2F^2c_0}} \tag{2}$$

where ε_0 is the permittivity of free space, ε is the relative permittivity of the solution, R is the gas constant, T is the temperature in Kelvin, z is the chemical valence of the electrolyte, F is the Faraday constant, and c_0 is the molar concentration of the electrolyte. Table 1 shows theoretical Debye lengths for a symmetric monovalent electrolyte at room temperature. These lengths can range from less than a nanometer to hundreds of nanometers, so for many of the nanogap devices in use and under development, overlapping double layers are a commonality.

Nanogap Biosensors, Table 1 Debye lengths (λD) as a function of electrolyte concentration expressed as molarity (M)

Debye lengths	
c_0 (M)	λD (nm)
10^0	0.30
$10-^1$	0.96
$10-^2$	3.04
$10-^3$	9.62
$10-^4$	30.4
$10-^5$	96.2
$10-^6$	304

To investigate the electric potential that results from the distribution of ions or charges inside of a nanogap, Poisson's equation is used:

$$\nabla^2\phi = \frac{-\rho E}{\varepsilon} \tag{3}$$

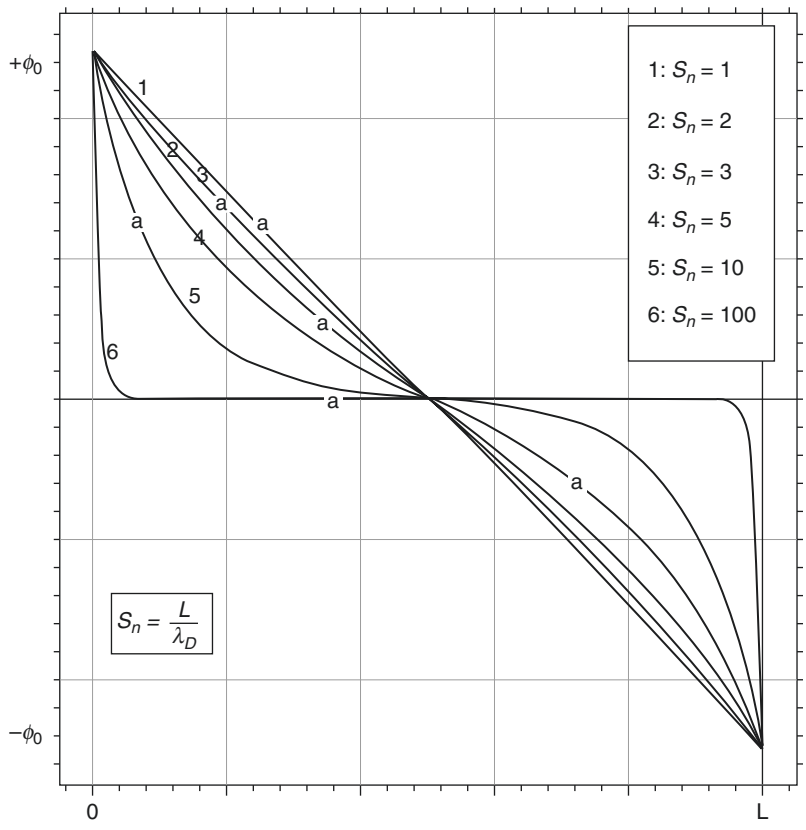
where ϕ is the electric potential and ρE is the charge density:

$$\rho_E = F\sum_i z_i c_i \tag{4}$$

where i represents the ion species. If the nanogap is to be considered a closed system, mass must be conserved, so the following equation is also required:

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot J_i + R_i \tag{5}$$

where J_i is the flux density in mol/cm^2/s and R_i is the production of species i, which is assumed to be zero at the voltages common for impedance spectroscopy. The flux density can be shown as:



Nanogap Biosensors, Fig. 2 Numerical solution of overlapping double layers. The electrode surfaces are represented by the *left* and *right* plot borders. The *y*-axis is the potential, and the *x*-axis is the distance between the electrodes. The potential distribution is plotted as the dimensionless number, S_n

$$J_i = -D_i \nabla c_i - z_i F c_i u_i \nabla \phi \tag{6}$$

where D_i is the diffusion coefficient and u_i is the ion mobility. This represents diffusive flux and electric flux, as represented by the two terms in this equation, respectively. Flux across the boundaries is assumed to be zero, and the potential on the two electrode surfaces are:

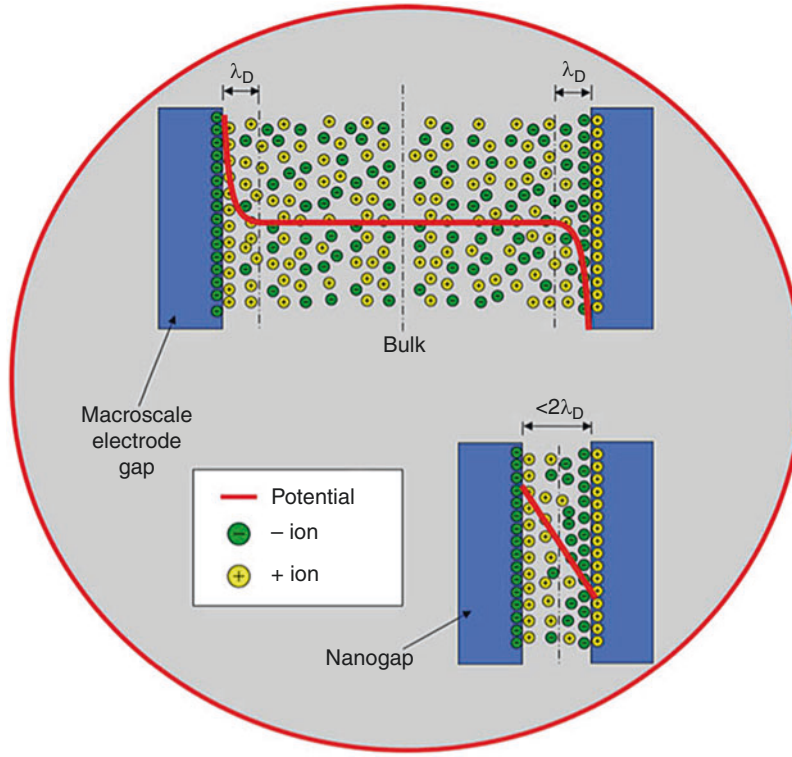
$$\phi(x = 0) = -\phi(x = L) = \phi_0 \tag{7}$$

The distance between the electrodes, L, is varied. A dimensionless ratio of the distance between the electrodes and the Debye length is defined to be S_n:

$$S_n \frac{L}{\lambda_D} \tag{8}$$

The results from this solution are shown in Fig. 2. Although this analysis is based on steady-state assumptions, it is a good approximation for most impedance spectroscopy systems, since the relaxation frequency of the electrical double layer is typically faster than the frequency of the current injected by the analyzer.

It is clear from the numerical solution shown in Fig. 2 that when the double layers overlap, the potential drop across the electrodes approaches a linear regime. This is significantly different from a macroscopic gap, where virtually all of the potential drops across the double layers and the bulk solution in a macro-gap typically exhibits a null electric field. This is an important point for biosensors, since the entire solution should ideally be subjected to the same conditions. For macroscale electrode sensors, molecules in the middle of the gap will behave much differently from molecules found on or near the electrode surfaces. The uniform field inside a nanogap is a significant advantage for these devices, since all molecules inside of the gap will see the same electrical conditions. The bulk solution is essentially removed, thereby providing, at least in theory, a more homogenous



Nanogap Biosensors, Fig. 3 Electrical double layers and overlapping electrical double layers. Schematic of ion distribution inside macroscale gaps and nanogap devices, and resultant potential distribution. λ_D represents the Debye length

response. Figure 3 depicts a simple illustration of the electrical double layers in a macroscale gap and a nanogap with overlapping double layers.

Measurement Methods

The detection method is typically achieved by measuring the complex impedance response to either a single excitation frequency or a full range of frequencies, or spectrum: The impedance spectrum of a system is measured by applying a frequency-dependent excitation voltage and measuring the resulting current. From the impedance spectrum and the geometrical parameters of the system, the complex permittivity or dielectric spectrum is derived.

Kaatze and Feldman have comprehensively reviewed the basic principles of dielectric spectroscopy and the methods in use to measure the dielectric properties of liquid samples over the frequency range from about 10^{-6}–10^{12} Hz [12].

Conventionally, a small AC (alternating current) potential, $V^*(j\omega)$, swept over a range of frequencies is applied. The electrical current response, $I^*(j\omega)$, is measured and the complex impedance $Z^*(j\omega)$, of the system is:

$$|Z^*(j\omega)| = \frac{V^*(j\omega)}{I^*(j\omega)} = Z_{\mathrm{RE}}(j\omega) + jZ_{\mathrm{IM}}(j\omega) \quad (9)$$

Here $Z^*_{\mathrm{RE}}(j\omega)$ and $Z^*_{\mathrm{IM}}(j\omega)$ are the real and imaginary parts of the complex impedance, respectively. The magnitude and phase angle of the complex impedance are

$$|Z^*(j\omega)| = \sqrt{[Z_{\mathrm{RE}}(j\omega)]^2 + [Z_{\mathrm{IM}}(j\omega)]^2} \quad (10)$$

and

$$\angle Z^*(j\omega) = \arctan\frac{Z_{\mathrm{RE}}(j\omega)}{Z_{\mathrm{IM}}(j\omega)} \quad (11)$$

respectively.

Knowing the geometrical and physical parameters of the biosensor, it is possible to model the physical mechanisms of conductivity and

polarizability with equivalent circuit analysis. Using networks of resistors and capacitors, the response of the system under controlled conditions can be predicted, and consequently the detection of biological events that cause deviations from the predicted response.

Determination of an appropriate equivalent model requires some a priori knowledge of the chemistry and physics of the system. Because multiple models can be found to fit any acquired set of impedance data, the choice of the "most appropriate" model is often the result of a complex balance between complexity, in terms of raw number of circuit elements employed, and efficiency, in terms of which chemical and physical phenomena of the system need to be characterized and with what level of sensitivity.

In general, it is of the utmost importance to design a nanogap sensor system where the number of parasitic circuit elements (which tend to increase the complexity of the resultant equivalent circuit model) is minimized, while the impedance response of the system is maximally affected by specific biological events taking place in the nanogap. This is usually achieved by a combination of clever design, appropriate choice of materials, and a reasoned choice of range of frequencies to be measured.

To maximize sensitivity, it is advantageous to have large active areas and small interelectrode distances, maximizing device capacitance. In the particular case where biomolecules are sensed as they bind to the surface and modify the double layer, it is also convenient to minimize the contribution of the bulk solution to the overall impedance, which leads to a sensitivity response that is inversely proportional to the interelectrode distance. Defining ξ as the *quality factor*, the two dependencies can be formulated as follows:

For a generic parallel-plate capacitive sensor, the measured quantity is capacitance, while the sensed parameter is the permittivity ε. The two are related by the well-known expression:

$$C = \varepsilon \frac{A}{d} \tag{12}$$

where A is the area of the electrodes, d the distance separating them, and ε the permittivity of the dielectric between the two plates. To maximize the measured capacitance given a fixed permittivity it is necessary to maximize the area-to-interelectrode distance ratio. Therefore, the first quality factor can be defined as

$$\xi_1 = \frac{C}{\varepsilon} = \frac{A}{d} \tag{13}$$

For the specific case of sensors sensitive to surface binding events where bulk contributions (due to the total volume enclosed by two parallel electrodes) are to be minimized, a second quality factor can be defined as the area-to-volume ratio, i.e.,

$$\xi_2 = \frac{A}{V} = \frac{A}{A \cdot d} \tag{14}$$

The combination of these two factors yields:

$$\xi = \xi_1 \xi_2 = \frac{A}{d^2} \tag{15}$$

This dimensionless parameter can be used to quantitatively compare the efficiency of different nanogap sensors, as well as to optimize the design of new sensors.

Although alternating current methods are the most common methods used with 3D nanogap biosensors, changes in direct currents (and similarly, shifts in IV curves) have been used [13].

Key Research Findings

Some key pieces of literature are briefly listed below. This list is not meant to be exhaustive, but rather illustrative of the most significant research findings achieved using nanogap sensors, and a starting point for further exploration.

One- and Two-Dimensional Nanogap Biosensors

One- and two-dimensional nanogap devices have been used mostly in proof-of-concept studies, to determine the ability to detect and to a limited extent, characterize nucleic acids (DNA, RNA)

binding events, and presence of other biologically relevant biomolecules such as proteins and antibodies. Chen et al. detected specific target DNA concentrations of as low as 10 picomolar identifying single-base-pair mismatches [14]. Liang et al. observed electrical signals caused by 1.1 kilobase-pair (kbp) double-stranded (ds)-DNA passing through the gap in the nanogap detectors with a gap equal to or less than 13 nm [1]. More recently, Cingolani et al. demonstrated that hybridization events are detected by the discrete increase in conductivity when the target AuNC–DNA conjugates bridge the electrode gap [15].

The electrical detection of other biomolecular interactions has also been demonstrated. For example, Haguet et al. have detected the capture of antibodies as well as the specific interaction between the biotin and the streptavidin molecules [16]. Subsequently, the same group has shown that the same method can be extended to the detection of immunoglobulin G (IgGs) in serum using protein probes [17].

Three-Dimensional Nanogap Biosensors

Three-dimensional nanogap resistive sensors with gaps of 5, 10, and 15 nm have been used to detect biotin–streptavidin binding events by measuring peak increases in direct current [13].

Single-frequency AC impedance measurements have been used successfully with three-dimensional nanogap sensors adopting metal electrodes and integrated reference sensing on-chip, for the detection of thrombin, a blood clotting factor, and to investigate its binding properties to a specific antibody and RNA-aptamer. A similar device was used to study the interaction of the Rev peptide with a corresponding RNA anti-Rev aptamer in concentration spanning the range of 100 nM–2 μ-M. The parasitic effect of the double layer was significantly reduced by using measurement frequencies of 800–1,280 MHz [18].

AC impedance spectroscopy has also been used to specifically detect the presence of thrombin in solution by looking at impedance spectra recorded using an 20 mV AC signal swept over a frequency range of 10 Hz–100 kHz [19].

Similarly, hybridization of poly-A nucleotides to immobilized poly-T between nanogap electrodes has been shown to result in a measurable change in capacitance over a frequency range of 100 Hz–10 kHz [20].

Future Directions

At this point in time, the development of nanogap biosensors is still in the research phase, i.e., there are no commercially available nanogap biosensors. However, the promise of sensitive, label-free devices with small footprints is alluring. Great strides have been made in terms of nanogap manufacturing and detection. Future work is needed in scale up of manufacturing, as few groups have achieved repeatable large-scale fabrication. This will be essential in developing sensors in quantities at low cost.

For applications relating to nucleic acid determination and sequencing, increases in sensitivity and specificity are necessary for these biosensors to compete with existing industry standard sensors which offer fluorescence-based readouts.

Specific label-free, all-electrical nucleic acid detection can still be valuable for diagnostic point of care devices, although additional work is needed to enable multiplexed detection which will allow the analysis of a panel of biomarker sequences, rather than a single short target sequence.

Similarly useful will be nanogap devices that enable the detection of specific proteins in serum or blood, forfeiting the cost and complexity associated with traditional antigen–antibody assays such as ELISAs. Devices aimed at point-of-care testing are particularly exciting because this application capitalizes on the inherent advantages of a nanogap: small volumes, small footprint, low power consumption, and label-free, all-electrical detection. The future will likely see many point-of-care biosensors using electrical-based detection, and perhaps nanogap technology can play an important role.

Cross-References

- ▶ Biosensors
- ▶ Dry Etching Processes
- ▶ SU-8 Photoresist

References

1. Liang, X., Chou, S.Y.: Nanogap detector inside nanofluidic channel for fast real-time label-free DNA analysis. Nano Lett. **8**, 1472–1476 (2008)
2. Im, H., et al.: A dielectric-modulated field-effect transistor for biosensing. Nat. Nanotech. **2**(7), 430–434 (2007)
3. Gu, B., et al.: Nanogap field-effect transistor biosensors for electrical detection of avian influenza. Small **5**(21), 2407–2412 (2009)
4. Chiras, S., Clarke, D.R.: Dielectric cracking produced by electromigration in microelectronic interconnects. J. Appl. Phys. **88**, 6302–6312 (2000)
5. Fernandez-Martinez, I., Gonzalez, Y., Briones, F.: Parallel nanogap fabrication with nanometer size control using III-V semiconductor epitaxial technology. Nanotechnology **19**(27), 275302 (2008)
6. Higuchi, Y., et al.: Application of simple mechanical polishing to fabrication of nanogap flat electrodes. Jpn. J. Appl. Phys. **45**, L145–L147 (2006)
7. Li, T., Hu, W., Zhu, D.: Nanogap electrodes. Adv. Mater. **22**(2), 286–300 (2009)
8. Lazzarino, M., et al.: Twin cantilevers with a nanogap for single molecule experimentation. Microelectron. Eng. **83**(4–9), 1309–1311 (2006)
9. Stroscio, J.A., Kaiser, W.J.: Scanning tunneling microscopy. In: Celotta, R., Lucatorto, T. (eds.) Methods of experimental physics, vol. 27. Academic, San Diego (1993)
10. Ionescu-Zanetti, C., et al.: Nanogap capacitors: sensitivity to sample permittivity changes. J. Appl. Phys. **99**(2), 024305 (2006)
11. Lyklema, J.: Fundamentals of Interface and Colloid Science, p. 768. Academic, London (1995)
12. Kaatze, U., Feldman, Y.: Broadband dielectric spectrometry of liquids and biosystems. Meas. Sci. Technol. **17**(2), R17 (2006)
13. Jang, D.Y., et al.: Sublithographic vertical gold nanogap for label-free electrical detection of protein-ligand binding. J. Vac. Sci. Technol. B **25**(2), 443–447 (2007)
14. Chen, F., et al.: Electrochemical approach for fabricating nanogap electrodes with well controllable separation. Appl. Phys. Lett. **86**(12), 123105 (2005)
15. Maruccio, G., et al.: A nanobiosensor to detect single hybridization events. Analyst **134**(12), 2458–2461 (2009)
16. Haguet, V., et al.: Combined nanogap nanoparticles nanosensor for electrical detection of biomolecular interactions between polypeptides. Appl. Phys. Lett. **84**(7), 1213 (2004)
17. Marcon, L., Stievenard, D., Melnyk, O.: Characterization of nanogap chemical reactivity using peptide-capped gold nanoparticles and electrical detection. Bioconjug. Chem. **19**(4), 802–805 (2008)
18. Schlecht, U., et al.: Detection of Rev peptides with impedance-sensors – comparison of device-geometries. Biosens. Bioelectron. **22**(9–10), 2337–2340 (2007)
19. Mannoor, M.S., et al.: Nanogap dielectric spectroscopy for aptamer-based protein detection. Biophys. J. **98**(4), 724–732 (2010)
20. Choi, Y.K., et al.: Sublithographic nanofabrication technology for nanocatalysts and DNA chips. J. Vac. Sci. Technol. B **21**(6), 2951–2955 (2003)

Nanogrippers

Peter Bøggild
DTU Nanotech – Department of micro- and Nanotechnology, Technical University of Denmark, Lyngby, Denmark

Synonyms

Nanotweezers

Definition

Nanogrippers are micro- or nanoactuators equipped with high-precision multiple end effectors, capable of applying controllable, opposing forces on nanoscale objects and thereby allowing these to be lifted free from a surface, held in a well-defined position, moved to another position, and/or placed in a controlled way.

Overview

Nanogrippers allow structures with at least one dimension being of order 100 nm or smaller to be manipulated in a predictable manner. Such grippers consist of one or more mechanically connected actuators, which determine the gap between two or more end effectors. Arrangements where independent macroscopic motors or actuators control multiple end effector individually are not considered to constitute nanotweezers or nanogrippers. The actuators are operated through variations of a physical quantity, which is typically an electrical current or voltage, but might also be the temperature, a magnetic field, or even light. While there is a wide selection of actuators

available for conventional microelectromechanical systems (MEMSs), nearly all nanogrippers which have been demonstrated in practical manipulation of nanostructures so far have either used electrostatic or electrothermal actuation. These actuation principles combine simplicity and compactness of design with predictable actuation behavior and are well suited for downscaling. The nanogripper itself could in principle have any size, as long as this does not prevent it from accessing and manipulating nanostructures in a controlled manner. There are, however, several reasons why grippers shown to be effective in manipulation of nanostructures tend to have an overall size significantly smaller than microgrippers; large grippers tend to block the line of sight of the microscope which is necessary to align the gripper to the object. By using conventional contact photolithography, the lateral resolution with which the end effectors – the parts of the gripper touching the sample – can be shaped is limited to around 1 μm. While high-precision fabrication techniques such as electron beam lithography and focused ion beam lithography can be used to achieve more well-defined end effectors, nanogrippers have been realized with multiwalled carbon nanotubes acting as both electrostatic actuators and end effectors. Nanogrippers can be used to manipulate objects which are difficult or time-consuming to approach, touch, and move or to place them in special positions. While manipulation is inherently slow compared to conventional parallel fabrication techniques in terms of mass fabrication, gripper-based nanoassembly has shown to be a flexible, rapid prototyping technique with a short turnaround time for small-scale production of nanodevices. Robotic manipulation using microgrippers has, for instance, been used to break off individual carbon nanotubes from their initial positions and mount these onto cantilever probes for high-aspect ratio atomic force microscopy. Other applications for nanogrippers include using the multiple tips as probes for fast electrical characterization, for assembly of nanostructures using dielectrophoresis, or to arrange nanoscale objects for subsequent structural analysis with a transmission electron microscope.

Design Considerations

The choice of actuator principle has considerable influence on the size, shape, generated force, and thereby the applicability of a gripper tool, and for nanoscale gripping the options are in practice much fewer as compared to microgrippers. Piezoelectric actuators made from ceramic materials can generate very precise motion with a high force but are notoriously difficult to integrate in a small tool. A large number of actuation principles have been proposed for nanogrippers and nanotweezers; however, nearly all attempts so far resulted in working devices having been made with either electrostatic or electrothermal actuation. Both of these principles provide a reasonable compromise between ease of fabrication and device integration, high reliability, high resolution, a large generated force, and a small overall footprint. The discussion will mainly focus on these two, as a treatment of the numerous other potential actuation principles and their role in hypothetical nanogrippers is beyond the scope of this entry. Emerging actuator concepts that may become important for more complex gripping/grasping schemes are briefly mentioned.

With a very few exceptions, all nanogrippers perform their motion in a single plane. Conventional bulk micro- and nanofabrication is highly suited for making in-plane actuators for electrostatic and electrothermal actuation, which is very convenient for miniaturization. Where electrostatic actuation uses Coulomb attraction between oppositely charged capacitor plates to generate a force, electrothermal actuation arises from exploiting the thermal expansion of materials, when heated up by an electrical current. In a review article, Sahu [22] discusses the most relevant actuator options for micro- and nanomanipulation and lists the key design characteristics, which are adopted for nanogrippers in the following:

Size – small size to allow interaction with nanoscale components. An important trade-off for all types of grippers is to make the actuators, or at least the end effectors, small and precise enough for delicate manipulation with a good

line of sight between the microscope and the manipulated object while at the same time providing a sufficiently solid structure to overcome surface forces between the object and the substrate or in some cases to break off the object from its initially fixed position. Many designs fail to be ever used in practice either because they make it extremely difficult to observe the manipulation process due to their bulkiness or because the force they provide is way too small to detach the objects of interest from a surface in a controlled way. Finally, larger actuators often suffer from out-of-plane bending of up to tens of micrometers, requiring constant realignment and can lead to damage of the samples. This is not necessarily a problem for manipulation of microscale objects but can make nanomanipulation extremely difficult and cumbersome.

Range of motion and applied force – depending on the size of the objects, as well as how much force is required to release them from their initial position, a proper balance between range of motion (actuation range) and gripping force must be chosen – generally, this is a trade-off given by the choice of actuation principle and design.

Actuation resolution – the deflection per input voltage /current should be small enough to accurately administer the force, e.g., in the range 0.1 nm/mV for a voltage source. Since the deflection is basically quadratic in drive signal, both for electrothermal and electrostatic nanogrippers, the resolution depends on the deflection.

Contact and force sensing – to avoid damage to samples and to allow detection of the nanogripper, contact and force sensing are very attractive. This is best done with either capacitive or piezoresistive readout, which unfortunately both add to the complexity of the devices and therefore make design of nanogrippers even more challenging. In practice, analysis of microscope images often serves as a convenient way to both detect contact and infer the applied forces. Knowing the spring constant of a gripper, the gripping force can be estimated from the observed deviation from its known mechanical equilibrium position. It is still very challenging to detect forces in a range that is relevant for manipulation of nanoscale objects (Yamahata et al. 2008a), in particular if such sensitive force-sensing components have to be integrated with a minute nanogripper device. For biological nano- and microscale objects, the relevant force range is pN to nN. Nanonewton force resolution for microgrippers can be achieved by capacitive readout [15]. Manipulation of hard, rigid objects like silicon nanowires or multiwalled carbon nanotubes may involve forces in nN, while for picking up (detachment or breaking) of such structures, forces in the μN range are relevant [3].

Power – a low power dissipation is beneficial for avoiding damage to the samples and in the case of electrothermal grippers also to protect the grippers themselves. High peak temperatures in such grippers may cause deformation or failure not just in polymer but also metallic or silicon actuator structures. Careful topological design of the layout [23] or proper choice or combinations of materials [12] can significantly reduce the peak temperature at the end effector and the overall heat dissipation to the environment.

Fabrication and material selection – both for electrostatic and electrothermal actuators, the choice of material and fabrication strategy can have a great impact. While silicon is often chosen as a default material due to the availability of proven fabrication methods and its excellent mechanical properties, there are sometimes more suited alternatives. For instance, the much larger thermal expansion coefficient of the polymer SU-8 in combination with a metallic heating layer can provide large actuation stroke at far lower temperatures than pure metallic or silicon actuators [12]. With photolithography, it is possible to fabricate end effectors capable of manipulating nanoscale objects with dimensions down to around 100 nm [9, 20, 21]. For entering the sub-100 nm range, electron beam lithography (EBL) is an obvious choice due to its order-of-magnitude improvement of the resolution

compared to photolithography. Alternatively, focused ion beam milling can be used to fine-tune end effectors [8, 10].

Micro- and nanoscale surface interaction – while van der Waals interactions comprise the most consistent and ubiquitous type of surface force, several other types of interactions interfere with nanogripper operation depending on the environment and the involved surfaces. Such effects are mostly but not always unwanted. In humid conditions capillary forces are the most prominent among a wide range of surface/contact interactive effects [6], while electrostatic forces due to charging by the electron beam can lead to strong stiction and unpredictable motion and behavior of nanostructures, for instance, inside a scanning electron or ion microscopes where the constant bombardment with charged particles can lead to strong charging effects. In both focused electron and ion microscopy, the intense beam can also directly lead to "gluing" or fusing of objects, by deposition of material.

Scaling, Surface Forces, and Pick and Place

The concept of nanogrippers is based on the presumption that manipulation using opposing forces is somehow useful similar to human hands, tweezers, and grippers in interacting with the environment. While robotic grippers comprise one cornerstone in the industrial civilization with more than a million industrial robots in operation worldwide, many of which are using gripper-like end effectors, it is not at all obvious that manipulation can ever be as important on a scale that is six orders of magnitude smaller. In the macroscale world, mechanical forces and gravity play a large role, simply due to the volumetric size and mass of the objects compared to the surface area. A stone is picked up using a hand, while sand or dust is far easier to wipe up with a wet cloth. The dominant role of surface forces in nanomanipulation can be illustrated by considering a sphere of radius R in close contact with a flat surface, assuming these to be separated by a few Ångstrøm. Both the van der Waals force [6]

$$F_{vdW} = \frac{A_H R}{6D^2} \tag{1}$$

and the capillary force

$$F_{\text{cap}} = 4\pi\gamma_L R\cos\theta \tag{2}$$

for this system are linearly dependent on R. The Hamaker constant $A_H \approx 10^{-19}J$ represents the combined van der Waals interactions acting between the atoms in the sphere and the surface, $\gamma_L = 0.072\ \text{J/m}^2$ is the surface tension of water, and θ is the contact angle. Comparing the magnitude of these forces to the gravitational force, $F_g = 4\pi R^3\rho g/3$, where ρ is the density and g is the gravitational constant, the surface forces exceed the gravitational force for objects smaller than roughly 10^{-3} m. While for macroscopic objects surface roughness generally decreases the effective contact area and thus the magnitude of the surface forces by orders of magnitude, surfaces and nanostructures encountered in nanomanipulation are ever so often in much more intimate contact – either because the nanostructures are small compared to the scale of the surface features or because the surfaces involved are atomically smooth, such as the surface of a carbon nanotube. Even in the extreme case of the effective contact area being reduced by three orders of magnitude due to surface roughness effects, the surface forces still dominate over the gravitational forces for objects in the submicron scale.

If the object can be picked up by adhesion forces using just a single end effector, it appears that there is not much use of dual end effectors; it is frequently observed in practice that the object remains attached to one of the end effectors after the gap has been opened. The macroscale analogue is picking up sand or salt using a pair of tweezers as compared to using a moist cloth. In this case, the capillary forces are clearly more effective than a mechanical force. However, since the surface forces can keep nanoscale objects strongly fixed to a surface, the possibility

of applying *a larger* mechanical force is required to pick up the objects [21]. For nanowires or carbon nanotubes grown directly on a substrate, the dual end effector is highly suited; epitaxial nanowires and carbon nanotubes typically have a maximal strength of order 10 GPa, with defect-free single-walled carbon nanotubes closer to 60 GPa, significantly larger than, for instance, hardened steel (1–2 GPa). A dual end effector can apply the mechanical force necessary to break off the object, maintain it in a fixed position between the jaws [21], and finally deliver it with the correct orientation at the target device or position. Dual end effectors are also convenient for electrical measurements of an object held between the jaws [14].

Environment and Microscopy

As seen from the above discussion, the performance characteristics of a nanogripper depend on a complex interplay between the material properties, actuation principle, size, and design of the gripper but also the application scenario in which the gripper are to be used. A key aspect of a nanogripper's usability is indeed the manipulation environment. This is determined by the environmental tolerances of the samples to be manipulated, the requirements and limitations of the microscopy technique needed for monitoring the process, and finally the physical space required to host the nanogripper, samples as well as the often quite bulky positioning manipulators. Issues related to environment and microscopy are discussed in the following.

Nanomanipulation may be performed in vacuum, ambient, or liquid conditions. A liquid environment is often chosen for manipulation of biological samples, and in this case the solution is typically a saline solution, which is electrically conducting. The biological objects are typically damaged at elevated temperatures, which nearly rules out electrothermal actuators as these invariably dissipate heat into the surrounding liquid when immersed. Similarly, most electrostatic actuators reported on today require application of voltages larger than a few volts, which leads to hydrolysis and consequently bubble formation in the liquid, which prevents any form of nanomanipulation. Finally, the liquid environment makes optical microscopy the only viable option, which even at the very best conditions makes alignment or even visualization of immersed sub-100 nm structures extremely difficult. The three-dimensional nature of nanogrippers excludes ultrahigh-resolution monitoring possible with scanning probe microscopy, as such techniques can only image and manipulate 2D and quasi-2D surfaces. Although atomic force microscope probes with integrated carbon nanotube tweezers have been fabricated by Akita and coworkers [2], a combination of scanning electron microscopy and a separate scanning probe was used to in a later demonstration [1]. There has thus been little progress in gripper-based manipulation of nanostructures in a liquid environment so far; here, optical tweezers are is a more powerful method along far-field manipulation strategies such as dielectrophoresis (Yamahata et al. 2008).

Ambient or atmospheric conditions allow both electrostatic and electrothermal grippers to operate in a useful way. Just as for a liquid environment, optical microscopy is the only practical option, providing imaging of wire-like structure (such as carbon nanotubes) with sizes down to roughly 100 nm and in some cases tens of nanometers. In ambient conditions, and in particular in very humid conditions [6], water films will reside on most surfaces and lead to strong capillary forces as discussed above. These surface forces can be very powerful and can make picking and placing of objects difficult.

Vacuum conditions allow electron microscopes to be used, and this is a key advantage. The scanning electron microscope allows imaging of nanostructures with lateral dimensions down to 1 nm (such as single-walled carbon nanotubes) while providing sufficient space inside the specimen chamber for the manipulation stages required to move the nanogripper. Capillary forces are largely eliminated, and the absence of air and water makes the behavior of the gripper more predictable. However, the constant and concentrated irradiation with electrons may damage the

samples, just as particularly biological samples cannot stay alive in vacuum. Scanning electron microscopes also tend to deposit thin layers of carbonaceous contamination, which can influence the sample, change the morphology, or even bind the samples in unwanted position. A consequence of this is sometimes that the manipulation has to be conducted with a strict control of the imaging, to avoid excessive contamination. This electron beam-induced deposition (EBiD) can be used to an advantage, providing a handy "glue," which can be applied simply by focusing the electron beam at a junction until a sufficient amount of carbon material has mechanically connected the two objects [9, 13, 23].

In terms of electron microscopes, other types of microscopes are relevant to mention. The transmission electron microscope provides atomic resolution in ultrahigh vacuum conditions, but allows far less than a cubic centimeter of sample space. It is possible to manipulate a single sharp tip inside a TEM using highly specialized nanomanipulators, but so far, nanogrippers as such have not been used in in situ TEM.

The environmental scanning electron microscope (ESEM) allows up to about 10 mbar gas to be inside the sample chamber, with a moderate degradation of the imaging resolution. The ESEM is a very flexible system for nanogripper operation, providing high-resolution monitoring, a certain degree of control of the environment, and sufficient space to operate one or more nanogrippers in situ. The dual-beam FIB/SEM combines a scanning electron microscope with a focused ion beam system, which is highly convenient for combination of nanogripper manipulation with in situ shaping of nanostructures and nanotools, deposition of metal electrodes, or fastening glue by EBiD. To date, most gripper-based manipulation of nanoscale objects has been done inside some type of scanning electron microscope.

Finally, the nanomanipulation or nanorobotic system used to perform the manipulation is a key factor in the success of a nanogripper. A deep discussion of nanomanipulation systems is beyond the scope of this entry; however, a few, highly successful examples of platforms for complex manipulative operations include, for instance, the semiautomatic nanomanipulation system by Eichhorn, Fatikow, and coworkers [13] with integrated depth detection using advanced three-dimensional scanning electron microscopy and image recognition technology and commercial, high-resolution multi-manipulator workstations from the US company Zyvex Corporation and German companies Klocke Nanotechnik GmbH, Kleindiek Nanotechnik GmbH, and SmarAct GmbH.

Electrostatic Nanogrippers

The electrostatic actuator is very widespread in MEMS, due to its convenient simplicity and ease of integration with standard silicon fabrication processes. It has a fast response time, provides nanometer-precision motion, consumes very little power, and thus dissipates virtually no heat to the surroundings. High-precision capacitive position sensors can be made with the same fabrication process and a similar capacitor layout, which means that integration of capacitive position sensors is natural and straightforward using bulk microfabrication. Nevertheless, some of the simplest, smallest, most well-known electrostatic nanogrippers are constructed through manually nanoassembly, rather than using bulk microfabrication.

Electrostatic Actuation

Electrostatic actuators are variable capacitors with air, liquid, or vacuum as a dielectric. A voltage difference V applied to two plates leads to a net electrostatic force acting between these. Depending on the design, the force F_{el} is translated into motion of two or more end effectors which can close around an object and hold it, thus constituting an electrostatic gripper. The electrostatic gripper comes in two varieties, the plate-type or the comb-type actuator. In the generic plate-type actuator, the capacitor gap is variable, while in the generic comb-type actuator, the effective (overlap) area is varied.

For a plate capacitor with area A and gap size g and dielectric absolute permittivity ε, the capacitance is $C = \varepsilon A/g$. The capacitor plates are

assumed to be kept in position by a spring with the spring constant k. With an applied voltage V across the capacitor plates, the change in potential energy ΔU for the spring-capacitor-voltage source system [24] is

$$\Delta U = \frac{1}{2}k(g - g_0)^2 - \frac{1}{2}CV^2, \tag{3}$$

where g_0 is the initial gap size. In equilibrium the total potential energy is minimum, $\nabla(\Delta U) = \mathbf{0}$, and thus for actuation with one degree of freedom actuation, the balance equation becomes

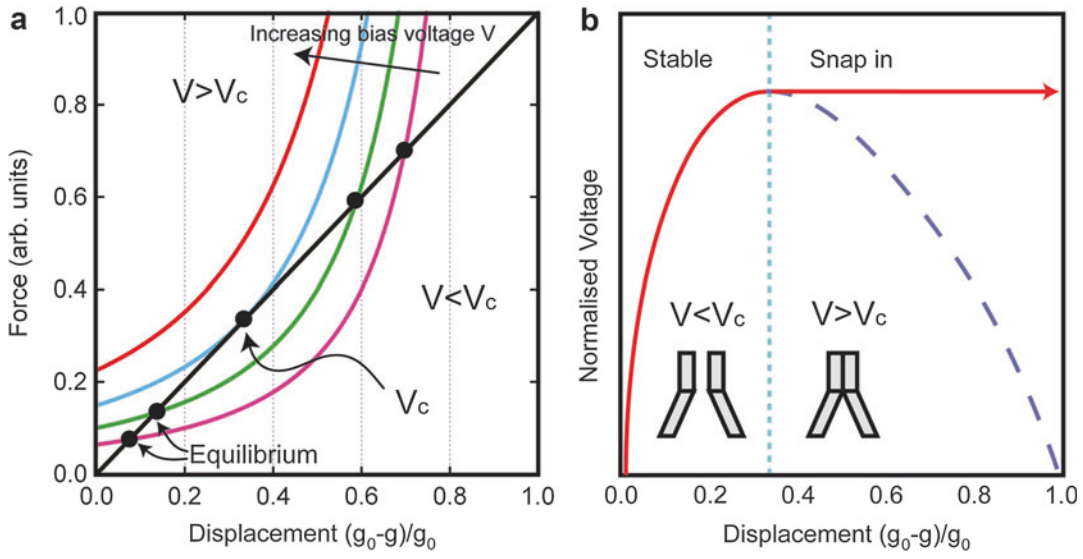
$$g - g_0 = \frac{1}{2}\frac{V^2}{K}\frac{\partial C}{\partial g} = -\frac{\varepsilon A V^2}{2kg^2}. \tag{4}$$

This corresponds to a spring force $F_m = -k(g - g_0)$ balancing the electrostatic force $F_e = \varepsilon A V^2/2g^2$, which is quadratic in V. As the gap is decreased, the magnitude of the capacitance gradient increases. At a critical voltage V_c corresponding to the gap size $g = 2g_0/3$, the spring force can no longer balance the electrostatic force, and the system becomes unstable, which is illustrated in Fig. 1. The actuator will experience a "pull-in," where the gap size is reduced to zero, possibly leaving the system in this state, even after the voltage has been reduced to below V_c (see Fig. 1b).

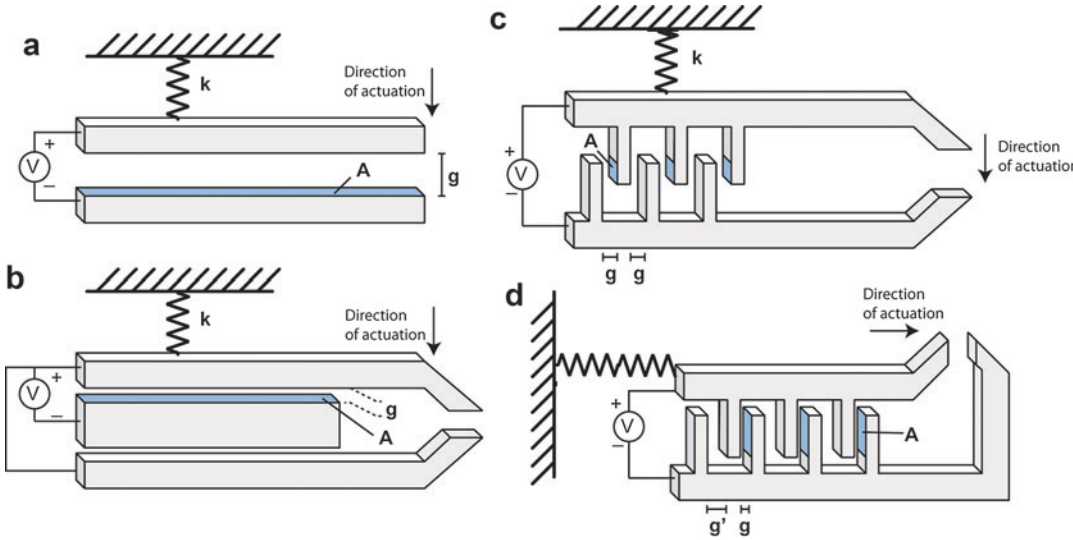
The instability can be avoided with the comb-type design, where the capacitors are equipped with interleaved, sliding pins. With this type of actuator, the capacitance is changed through the area rather than the gap size, as shown in Fig. 2. The force generated by comb-type actuators scales with the number of gaps, so higher forces can be achieved by increasing the area of the actuators. The displacement is thus [22]

$$\Delta x = N\frac{\varepsilon h V^2}{kg^2}, \tag{5}$$

where N is the number of gaps, h is their height, and the area between each capacitor pair is $A = h\Delta x$ (see Fig. 2). Comb-drives are generally larger and more bulky than plate capacitor drives and have only been used in a few instances for nanoscale manipulation. The operating voltage for electrostatic nano- and microgrippers ranges from about 5 V [2] to several hundred volts,



Nanogrippers, Fig. 1 (**a**) Force against displacement showing the mechanical equilibrium points where the linear elastic restoration force is balanced by the electrostatic force. At the critical voltage V_c, there is a singular equilibrium point, beyond which there are no stable solutions; the actuator thus experiences "pull-in." (**b**) The normalized bias voltage versus displacement illustrates the sudden snap-in event at Vc



Nanogrippers, Fig. 2 Different types of electrostatic actuators for nanogrippers. (**a**) Basic plate capacitor, where the gap size is changed upon increasing the bias voltage V, while the area A is constant. The movable part is thought to be connected to the device via an elastic support, here represented by a spring with a spring constant k. (**b**) A fixed center electrode for the two end effectors to remain at the same potential, thus preventing current from running through the gripped object. (**c**) A comb-drive used to close two end effectors. In this configuration the overlap area A between the two arrays of capacitors is varied rather than the gap size g, which avoids the pull-in instability. (**d**) Comb-like drive where the direction of motion of the movable part is lateral rather than vertical, with capacitor gaps g' and g. This configuration suffers from the pull-in instability, since it is essentially of the parallel plate type

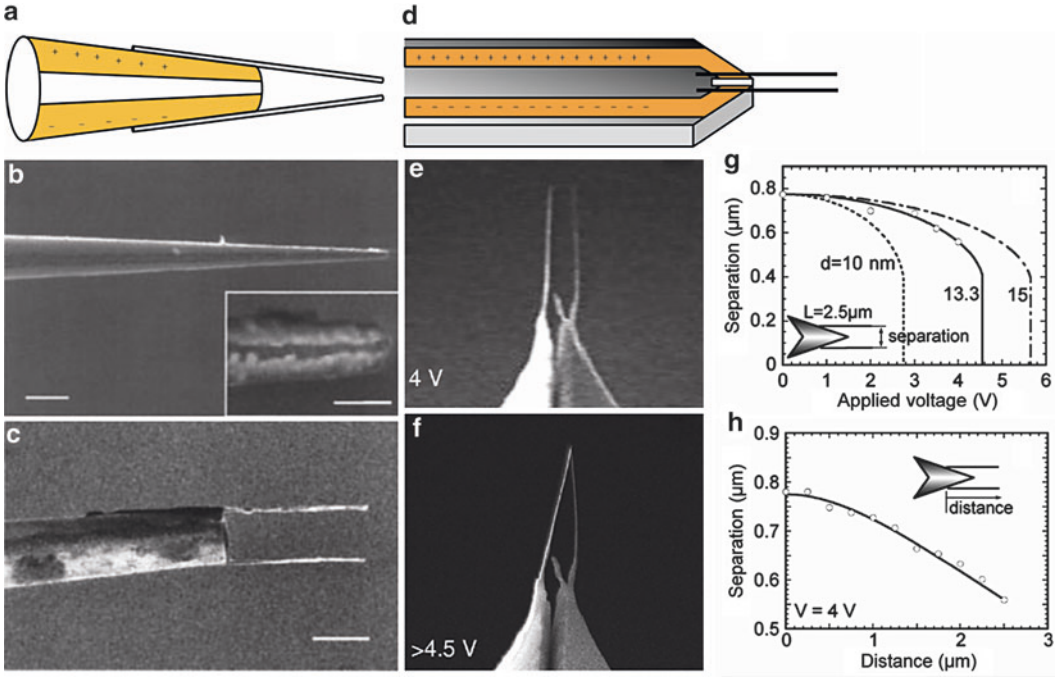
depending on the design, the size, and the requirements, with a resolution of order 1 nm.

Nanotube Nanotweezers

Although microgrippers of different types were developed already in the 1990, the first and most spectacular type of gripper tool capable of manipulating nanostructures is the nanotube nanotweezers [14]. This gripper is comprised of multiwalled carbon nanotubes glued to the sides of a glass microcapillary under an optical microscope. On each side of the glass capillary, a metal electrode was evaporated to provide electrical contact to the carbon nanotubes. By applying a moderate voltage, the nanotubes were attracted toward each other, until pull-in at 8.5 V where the gap g was roughly equal to $2g_0/3$. The authors demonstrated picking up of 500 nm diameter fluorescent nanobeads. As an added feature, the high conductance of the carbon nanotubes allowed electrical characterization of the held object. Following this approach, Akita et al. [2] integrated carbon nanotubes with an atomic force microscope probe, which was split in two electrically separated parts using focused ion beam milling (see Fig. 3).

Batch-Fabricated Grippers

While an impressive range of electrostatic microgrippers has been fabricated using bulk microfabrication, the majority is not designed to engage in manipulation of nanoscale objects, as this require very precisely manufactured end effectors as well as high resolution and precision of motion. Nelsons and coworkers [5] developed a range of advanced, highly integrated, and very well-characterized electrostatic grippers, which have since been commercialized by the Swiss company FemtoTools (see Fig. 4b). These grippers have a well-functioning force feedback sensing; have been tested thoroughly in air, liquid, and SEM; and are fully integrated with measurement and control electronics and computer interfaces. Although primarily designed for microscale manipulation, these are capable of nanoscale manipulation, as shown in the inset in Fig. 4b. Yamahata and coworkers reported a precision-



Nanogrippers, Fig. 3 (**a**) Kim and Lieber's nanotube nanotweezers [14] with multiwalled carbon nanotubes glued to the sides of a glass capillary, where an electrode layer has been deposited on each side. (**b**) Magnified view of the capillary after metal deposition. (**c**) The nanotweezers after mounting of the nanotubes. (**d**) Nanotweezers made by focused ion beam milling a metallized cantilever to separate the electrode into two [2]. (**e**) Two nanotubes attached to the tips at 4 V. (**f**) At a bias voltage of 4.5 V, the nanotubes pull in. (**g**) The separation versus applied voltage, showing pull-in occurring at different voltages depending on the initial separation. (**h**) Variation of gap separation for different initial lengths of the nanotubes, all at a bias voltage of 4V

manufactured electrostatic nanogripper with an integrated displacement sensor intended for biomolecules and demonstrated trapping of bundles of DNA molecules using dielectrophoresis rather than mechanical gripping (Yamahata et al. 2008). Mølhave [21] reported manipulation of silicon nanowires inside a scanning electron microscope using a plate capacitor gripper with five legs, of which two served as thin flexible actuator arms and three as rigid electrodes, allowing the gripper to both open and close compared to its neutral position (see Fig. 4).

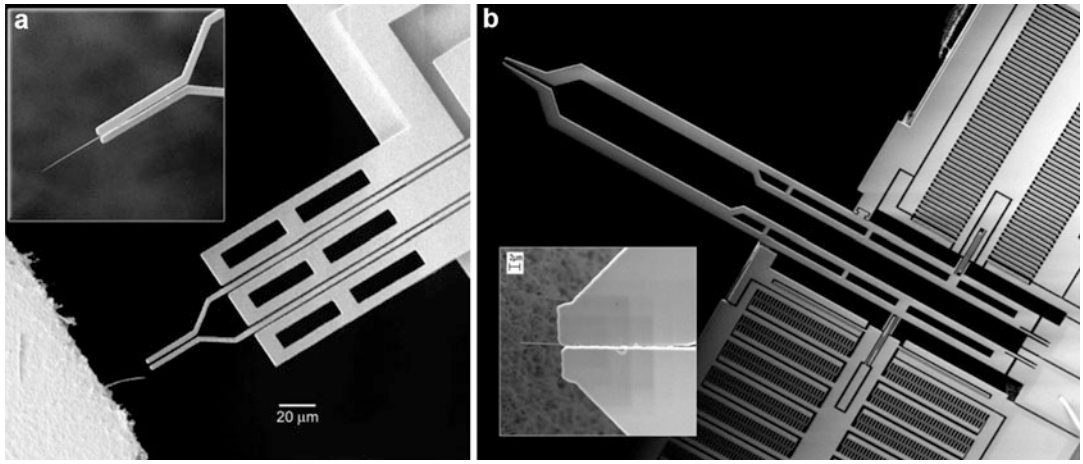
Thermal Grippers

The volume of most solids will increase with temperature, and this effect can be exploited in several ways to provide accurate, yet forceful motion. The thermal expansion of, for instance, a rod with length L as the temperature is changed from T_s to T_h can be written as

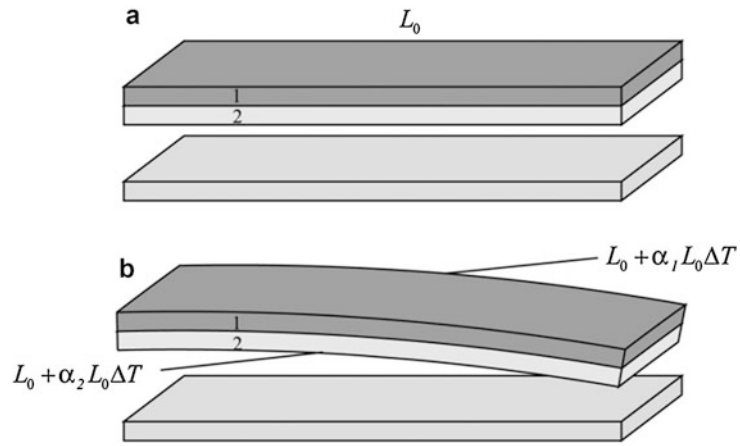
$$L(T_h) = L(T_s)\,(1 + \alpha\Delta T), \Delta T = T_h - T_s \quad (6)$$

where $\propto$ can be regarded as a constant, nearly independent of temperature. The thermal expansion coefficient varies between about 1–5 for Si, Ge, and C, 5–30 ppm/K for clean room relevant metals, and 52 ppm/K for a polymer material such as SU-8. A temperature increase of 100 K thus only leads to a 0.5 % expansion of silicon; hence, a key design feature is amplification of this expansion to a useful actuation range. A basic strategy is to obtain a different thermal expansion for two objects, which are connected mechanically in a way that amplifies the deformation.

In a bimorph actuator this can be done by coupling two layers of material with different thermal expansion coefficients α_1 and α_2 to each



Nanogrippers, Fig. 4 Two examples of electrostatic actuators: (**a**) a plate capacitor-driven microgripper with both open and close capability [21]. The *inset* shows the gripper holding a silicon nanowire with a diameter of 200 nm. (**b**) Gripper with lateral comb-drives and integrated capacitive position readout. The inset shows a similar gripper manipulating a silicon nanowire (Courtesy of V. Eichhorn and FemtoTools)



Nanogrippers, Fig. 5 Bimorph actuator used as a gripper. Two-layered beams with different thermal expansion coefficients α_1 and α_2 and initial length L_0 will deflect sideways upon a uniform temperature change ΔT. This could be used to form a compact nanogripper

other in a single beam as illustrated in Fig. 5. A homogeneous heating of the structure will then cause the beam to deflect toward the side with the smallest expansion. The bimorph actuation principle is well suited for downscaling, as demonstrated in the work by Sul and coworkers [26], which was capable of delivering 1 μN at a temperature of 430 K. The actuator was realized by depositing aluminum on multiwalled carbon nanotubes. Yet, there are no examples in literature of nanomanipulation carried out with a bimorph gripper.

Electrothermal Actuation

Correspondingly, a current I passing through two rods with different cross sections will be led to different temperature increases through Joule heating, as the dissipated power in a volume element of length Δx and cross section A is given by $RI^2 = \rho(dx/A)A^2J^2 = \rho AJ^2dx$, where the current density is $J = I/A$ and the $\rho = RA/dx$ is electrical resistivity. The temperature distribution can be found by solving the heat conduction continuity equation, which in its simplest form reads

$$\frac{d^2T}{dx^2} = -\frac{\rho J^2}{\kappa}, \tag{7}$$

where κ is the thermal conductivity. The heat exchange with the surrounding air/liquid is ignored, which corresponds to an electrothermal device in vacuum. Moreover, energy loss by radiation is omitted. While this can be solved analytically for temperature distribution and actuation response with many basic actuator designs, finite element method is often used for its greater applicability to nontrivial designs [23]. Two often used designs are the Guckel actuator (also called thin-beam thick-beam) and the ribcage actuator, also called the Chevron actuator [22]. The Guckel actuator is a loop where the two conducting arms have different cross sections and the thick beam has a hinge to allow flexible sideways motion. The ribcage actuator consists of opposing rods, which are angled slightly, so that the longitudinal expansion of the rods forces the actuator to move perpendicular to the beam direction, as indicated in Fig. 6, which also amplifies the motion.

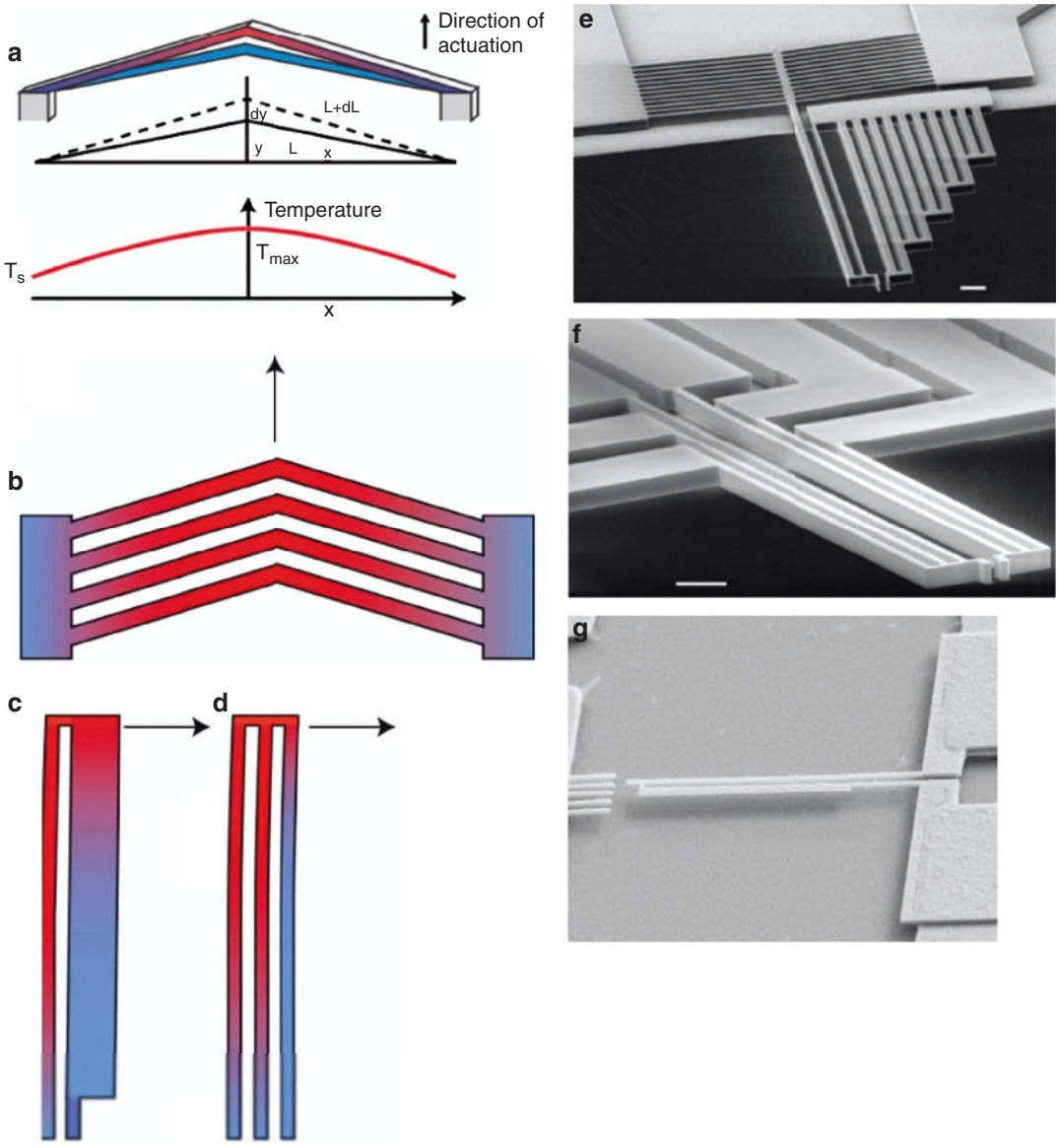
Carlson and Eichhorn [9] used an advanced nanorobotic manipulation system equipped with an electrothermal ribcage gripper to break off a multiwalled carbon nanotube from a substrate, mount it as a tip of an atomic force microscope probe, and finally fasten the nanotube to the probe by electron beam-induced deposition. The US company Zyvex manufactures a line of advanced electrothermal microgrippers. The conversion from electrothermal actuation to a mechanical motion of the end effectors is done through a highly optimized design where a capacitive comb structure provides force readout. While these grippers were clearly designed for micron-sized or larger objects, the precision and overall size of the end effectors should allow manipulation of submicron objects.

Chronis [12] employed a hybrid material approach to achieve a large actuation range at a much lower temperature than what can be achieved in bulk silicon or metal. Thin metallic structures for conducting the electrical current was combined with the high thermal expansion coefficient material SU-8 (52 ppm/K). The gripper has submicron features, can open almost 10 μm at a bias voltage of around 1 V, and most notably imposes a very small heat load onto the surroundings. Finite element calculations showed the temperature increase at the end effector of a fraction of a degree, while the overall temperature increase was around 60 K for the gripper opening of 12 μm. When using such a hybrid with different expansion coefficients, there is a risk of out-of-plane bending due to the bimorph effect. In this case, however, the thickness of the SU-8 layer was larger than the metal film that out-of-plane bending could be avoided.

Topology Optimization

Compared to electrostatic nanotweezers where the basic capacitor structures have proven their worth in millions of commercial MEMS devices, there is more room for improvement for electrothermal actuators. Electrostatic actuation is governed by large thermal gradients and involves the detailed shape of the actuator as well as the material properties in the mechanical, electric, and thermal domain. While the formalism outlined above (Eq. 7) is capable of predicting the behavior of any design, electrothermal actuators are often based on rather simple variations of either the thin-beam thick-beam actuator or the ribcage/V-shaped actuator.

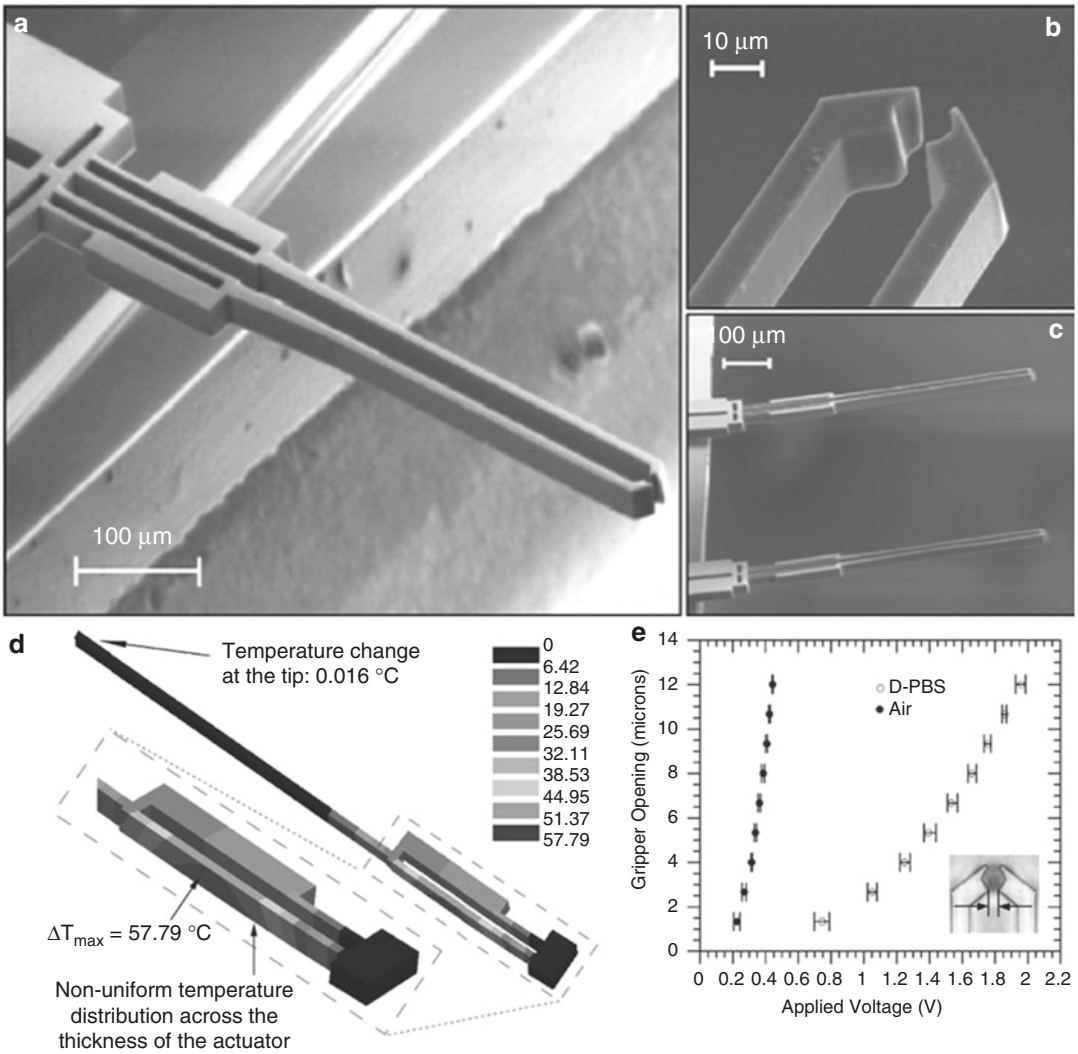
For nanogrippers, however, there are good reasons to push beyond the standard solutions. When downsizing the actuators to make overall smaller and more precise grippers, the available force typically also is decreased. To avoid strong out-of-plane bending, a high aspect ratio of the gripper arms is necessary, making it difficult at the same time to reduce the size and maintain a high lateral spring constant. Just as for macroscopical tweezers, the optimal shape of a gripper in terms of accessibility and actuation stroke is long and slender, while for strength and force, it should be short and wide. Sardan et al. [23] replaced the three-beam actuator with a topology-optimized actuator and achieved two orders of magnitude higher spring constant at the same actuation stroke (1 μm), while the peak temperatures were conveniently moved toward the base rather than



Nanogrippers, Fig. 6 Electrothermal actuators. (**a**) A V-actuator is pushed in the upward direction, as the opposing beams expand upon heating. By passing a current through the structure, it will attain a parabolic temperature distribution with the peak temperature Tmax in the center. (**b**) Ribcage or "Chevron" actuator consisting of many V-actuators in parallel. (**c**) Thin-beam thick-beam or "Guckel" actuator. As current is passed through the loop, the larger heat generated in the narrowest beam will lead to a skewed expansion, i.e., to the wide beam. (**d**) The three-beam actuator operates by passing a current through either the left or right two legs, consequently allowing for actuation in two directions, or open as well as close capability of the gripper. (**e**) Silicon ribcage actuator [9]. (**f**) Silicon three-beam actuator [3]. (**g**) Thin-beam thick-beam actuator with submicron dimensions [17]

end effectors (see Fig. 7). This strategy leads to grippers capable of successful, repetitive mounting of carbon nanotubes onto AFM probes, TEM grids, and fixed electrodes [23].

Cagliani and coworkers [8] used electron beam lithography to batch-fabricate topology-optimized nanogrippers with a length of 35–50 and a gap size of 300 nm, reducing the overall

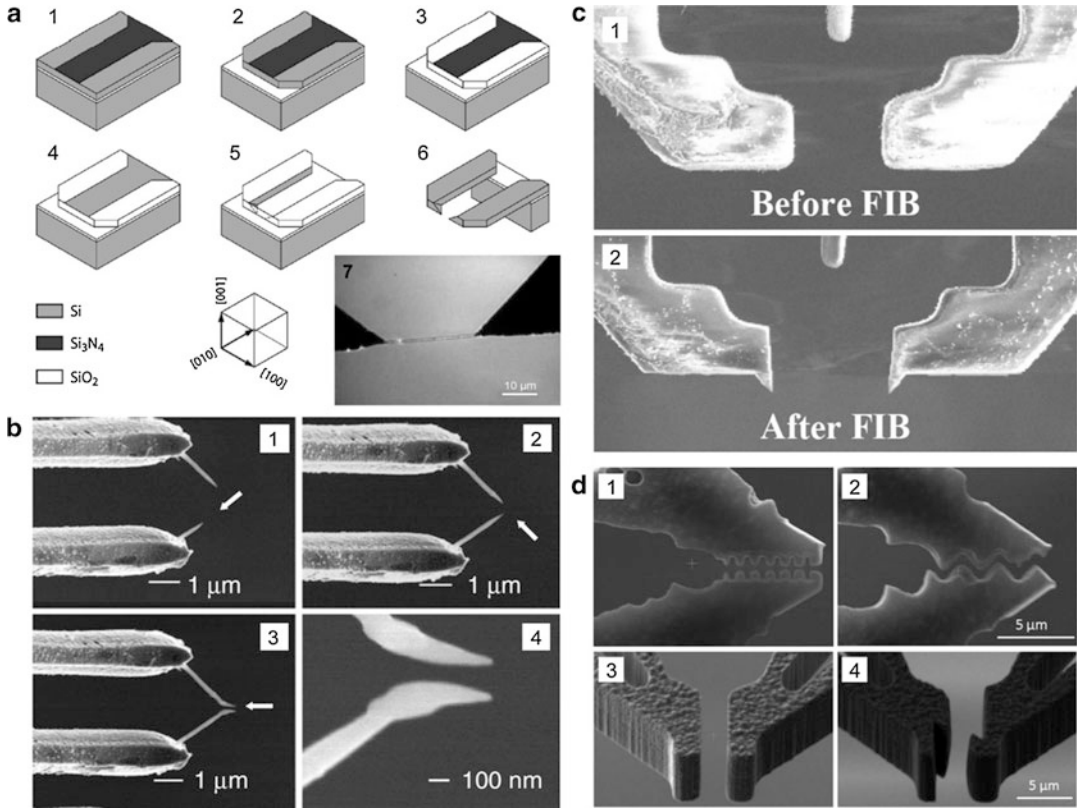


Nanogrippers, Fig. 7 (**a**) Micro-/nanogripper with electrothermal actuators defined in SU-8 with thin chromium-gold electrodes heating the Guckel-type actuators at the base. (**b**) Zoomed in and (**c**) overview micrographs of the gripper. (**d**) Finite element calculations showing negligible temperature increase at the tip. (**e**) Actuation curves for the gripper in both liquid (D-PBS) and air. Due to a different thermal resistance in liquid as compared to air, opening the gripper in liquid requires far higher applied voltage [12]

size by a factor of 2–6 times compared to Ref. [23]. The reduced size and thickness of the end effectors in combination with the enhanced rigidity given by the topology-optimized layout made it possible to break off as-grown gallium phosphide nanowires from the substrate inside a scanning electron microscope. By coating the grippers with a thin layer of gold, bundles of single-walled carbon nanotubes could be lifted up and mounted inside a transmission electron microscope for subsequent electrical and structural characterization. This device is so far the smallest batch-fabricated nanogripper in operation.

Nanoengineering of End Effectors

Just as for macroscale grippers, the size and shape of the end effectors – the critical interfacial part of the gripper where the contact is made to the nanostructures – are vitally important for successful application of the tool. While



Nanogrippers, Fig. 8 Customization of end effectors. (**a**) Anisotropic wet etch with potassium hydroxide (KOH) etch can be used to create sharp edges and tips in silicon. A narrow gap with ultrasharp opposing tips was created in a batch process using KOH etch by Yamahata (Yamahata et al. 2008). (**b**) Electron beam-induced deposition performed in the direction of the white arrow shows gradual build-up of dual supertips on silicon dioxide end effectors with a gap of 100 nm. The method was used to create gaps down to 25 nm. [7]. (**c**) Focused ion beam milling used to reshape end effectors (Chen, Zhang et al.). (**d**.1) Silicon microgripper fabricated with an initial closed gap, which is subsequently opened and further shaped by the focused ion beam milling for reduced contact area. (**d**.2) FIB milling is done so that the gripper will apply a finite force even in the neutral position, i.e., with no applied voltage. (**d**.3–4) show end effectors equipped with a self-locking mechanism, so the gripper when pushed together will lock in this position similar to some macroscopic tweezers (Courtesy of Özlem Sardan and Mikkel Klarskov)

conventional nanofabrication techniques such as electron beam lithography are sometimes used [8], this critical part of the gripper is often engineered using creative and unusual solutions. Elegant use of photolithography combined with an anisotropic potassium hydroxide wet etch allowed Yamahata (Yamahata et al. 2008) to create sharp opposing tips on the end effectors in a batch process (see Fig. 8a). Boggild et al. [7] used angled electron beam-induced deposition (EBiD) to define needle-shaped end effectors with a gap size of just 100 nm inside a scanning electron microscope, as shown in Fig. 8b. A gap size of just 25 nm could be achieved by gradually allowing the deposited material on the two tips to grow until the desired gap size was achieved. Focused ion beam milling is highly convenient for in situ reshaping of nanodevices, since it can both perform milling of material and deposition using precursor gases. FIB milling has been used to define end effectors with a more well-defined or smooth contact plane (Fig. 8c), to achieve a smaller contact area for reduced surface force interactions (Fig. 8d.1), or to engineer the end

effector so that it applies a force in its *neutral* position and has to be opened rather than closed during the actuation phase (Fig. 8d.2). Finally, FIB can be used for more exotic mechanism such as self-locking grippers which can keep a nanostructure mechanically fixed even when "released" (Fig. 8d.4). As an example, Bartenwerfer et al. ([4], #19567) demonstrated that careful FIB milling of the end effectors to become slightly tilted with respect to the vertical direction allowed picking up an object with one angle and mounting it with another, which is otherwise very difficult using linear actuated end effectors. Micrometer-sized flakes were picked up and turned 90° while being held within the gripper, before being stored in a cartridge.

Emerging Concepts for Advanced Gripping

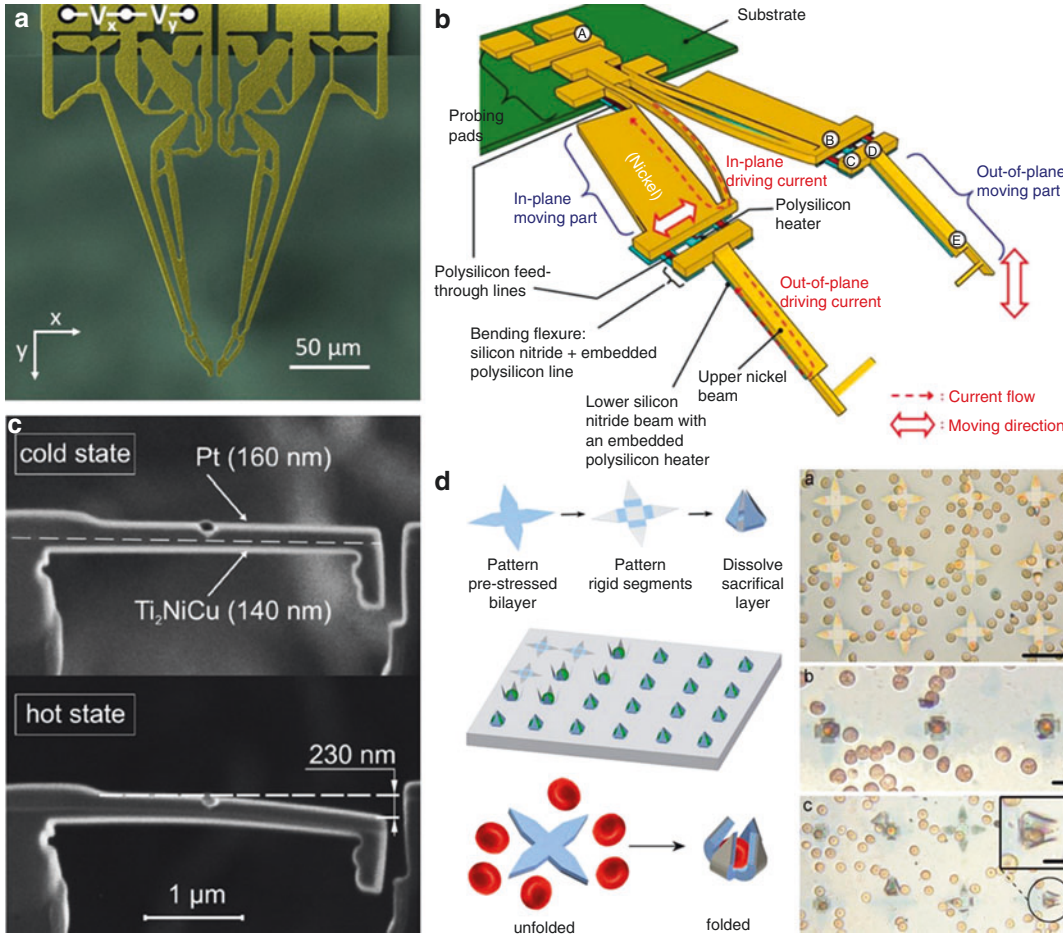
Within the general field of micro- and macroactuation, several emerging approaches have the potential to become relevant for nanogrippers; in some cases proof-of-concept prototypes have already been realized with dimensions, functionality, and actuation precision in the relevant range. A general restriction of nanogripper-based nanomanipulation in comparison with macroscopic robotic manipulation is the lack of non-translational degrees of freedom (DOF), such as rotation, as previously mentioned in section "Design Considerations." While a typical manipulation system features a stage with three linear DOF, as well as in-plane motion of one or two end effectors, the three-dimensional grasping capability of a human hand with multiple joints does not exist in any nanogripper reported in literature. Gripping or grasping is likely to be more flexible and object tolerant if not just the shape of the end effector but also the angle can be changed dynamically. In the following a few promising technologies for enhancing nanogripper functionality are discussed.

Multiple-Axis Grippers

For electrothermal as well as electrostatic actuators, it is relatively straightforward to add an extra in-plane degree of freedom, turning a single arm into a two-axis actuators. This will in many situations be possible without changing materials or fabrication process. The complexity of the gripper will increase, which will typically lead to a less compact design and possibly increase of out-of-plane bending and cross talk between the axes. For electrothermal actuators, cross talk is hard to avoid, as heat may easily dissipate from one actuator part to another in an uncontrollable way. Here, topology optimization could be helpful in achieving multiple-DOF designs with a small footprint and manageable cross talk, as such features can be incorporated explicitly as optimization goals. The review paper [25] describes the general topology optimization design procedure and examples of multiple-DOF electrothermal actuators. Figure 9a shows a two-axis topology-optimized gripper fabricated with the same process as in Ref. [23]. Chen and coworkers [10] included hinges in an actuator fabricated by metal MUMP (multiuser MEMS process), allowing the actuator arms of a prototype gripper to bend controllably in plane as well as out of plane (see Fig. 9b). While this particular design is not easily scaled toward nanoscale manipulation, the use of a standard planar MEMS process to integrate hinges into actuator structures and thereby achieve out-of-plane actuation is relevant. Angular alignment of end effectors for optimal interaction with nonspherical objects is a tedious and difficult task. By placing the hinges closer to the end effectors, the angle of the end effector could be changed without large translational shifts, thus increasing the versatility of the gripper.

Shape Memory Alloys (SMAs)

SMAs are alloys that upon heating recovers the original shape given by annealing at a high temperature, even after severe deformation at low temperature. While SMAs can produce a very high energy density (up to 10 MJ/m^{-3}) compared to other materials ([29], #19552), the memory effect is a one-way actuation process; an actuator needs external forces to deform with respect to the memory state. Also, SMAs are difficult to fabricate with small dimensions. Zakharov



Nanogrippers, Fig. 9 (**a**) Two-axis topology-optimized electrothermal gripper, with Vx and Vy indicating the terminals for actuation in the x and y direction, respectively. The length of the microgripper is 200 µm (Courtesy of Özlem Sardan and Mikkel Klarskov. (**b**) Two-axis microgripper with out-of-plane bending of the individual arms, incorporated by bending flexures (electrothermally actuated) [11]. (**c**) Shape memory alloy actuator demonstration showing significant actuation range compared to the length of the actuator [30]. (**d**) Built-in stress during fabrication allows small gripper structures equipped with hinges to fold and capture living blood cells upon release from surface [19]

et al. constructed a composite bimorph actuator with a $Ti_{50}Ni_{25}Cu_{25}$ SMA for generating closing motion of a dual end effector and a Pt layer for generating the stress that recovers the "open" state, as shown in Fig. 9c. The structure was milled out by focused ion beam milling. Provided the issues concerning machining and shaping SMAs with high precision are addressed, such materials could be used to generate high forces in compact end effectors.

Origami/Folding Grippers

The philosophy of producing independent end effectors with a built-in compact actuation mechanism is taken a step further in nano-origami. Nano- and microgrippers were fabricated by layering metals, silicon, and polymers, where a combination of built-in, temperature-dependent stresses and well-placed hinges allows the structure to fold and capture a target object in a handlike motion [18]. These structures are nearly

mm sized but provide predictable closing of the jaws from thermal or biochemical stimuli. As a nickel layer is integrated in the structure, they could even be manipulated with a magnetic field, allowing tetherless object retrieval and transport. A similar but simpler and much smaller design with 10–70 μm gripper arm [19] could catch individual living blood cells (see Fig. 9d). With this design the built-in stress during deposition causes the grippers to close once released from the substrate, which was done with a timed chemical process.

While these studies are clearly far from high-precision nanomanipulation, the intriguing prospect of a compact, reversible three-dimensional gripping suggests that this may be developed in the future. Applications have so far been within liquid-phase environment (life science); however, it is quite possible that similar principles could become feasible in a nanogripper operating in dry or vacuum environment, provided an approach to reversible open and close operations is developed.

Applications

In terms of nanomanipulation, the objects of interest may be small in either one, two, or three directions, i.e., plates, wires, or particles, and this makes quite a difference in terms of how they can be handled and with which tools. In the following the feasibility and optimal strategy of manipulating such objects will be discussed in the light of possible applications.

Quasi-0D Structures: Particles

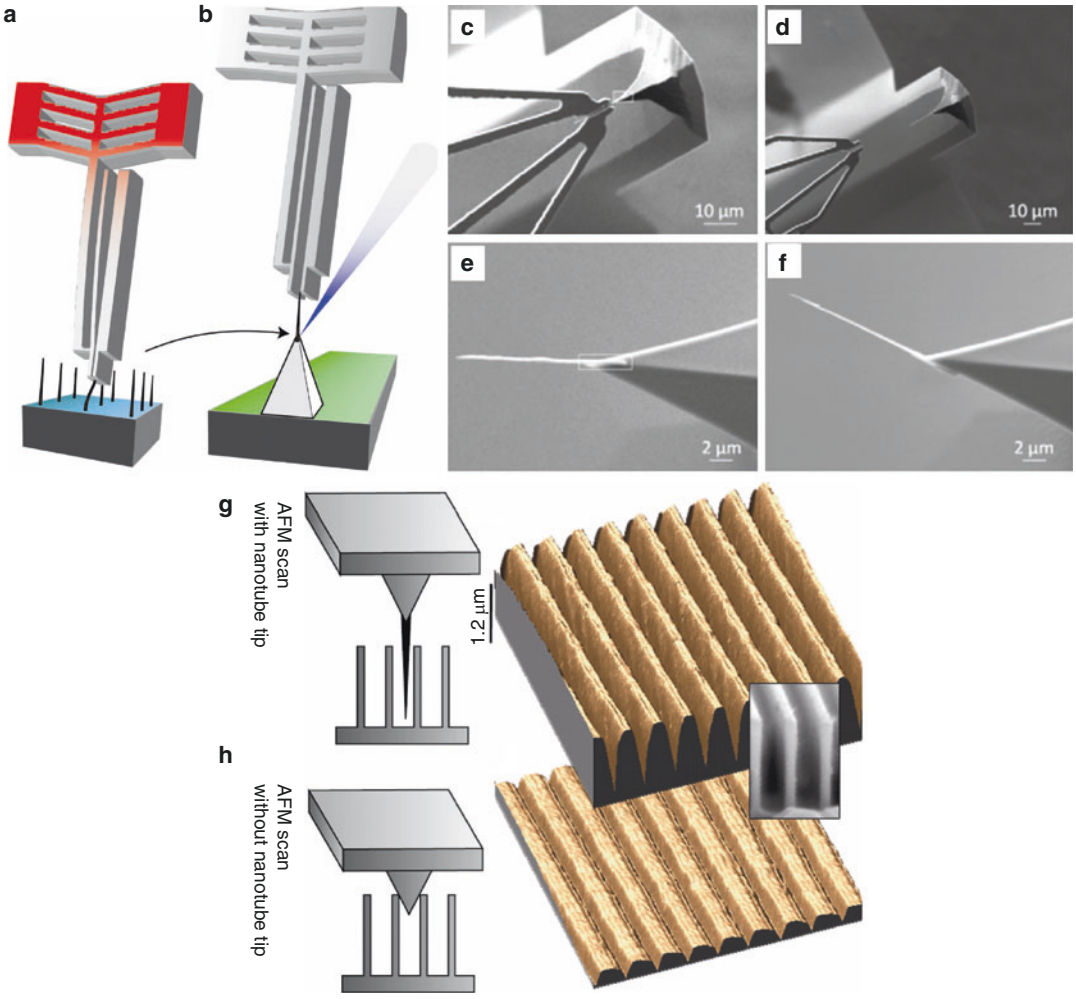
There are a few example pickups of submicron particles to demonstrate the capability of a nanogripper [1, 14]. Such manipulation can also be combined with electrical characterization of an object if the end effectors are conducting, such as carbon nanotubes [14]. For the manipulation process itself, a simple tip will in most cases be effective for particles as numerous pick-and-place experiments using scanning probe tips have. Submicron and nanoscale particles will in the majority of practical scenarios be notoriously difficult to pick up using a gripper, either because the gripper force is too weak to overcome the surface forces or too bulky to access the particle in a predictable way. This situation is comparable to lifting up grains of sand using fingers, which is possible but hardly convenient. While microgrippers [19] and other types of micromanipulators [27] have shown clear potential for handling and analyzing of biological particles down to the size of red blood cells, it is doubtful that mechanical nanogrippers will play an important role in applications involving manipulation of nanoparticles in the foreseeable future.

Quasi-1D Structures: Fibers, Wires, and Tubes

Just as for the macroscale, the most obvious and convenient object to handle with a gripper or a pair of tweezers is a quasi-one-dimensional object, such as a carbon nanotube. The oblong shape of the object makes access far easier than for nanoparticles, with excellent visibility as the nanowires/nanotube does not block the vision. Moreover, the fact that the free end of the nanowires or nanotube can easily be positioned and even by electron beam-induced deposition (EBiD) of focused ion beam deposition provides a straightforward and reliable pick-and-place process, which can even be automatized [13]. Manipulation of nanowires/nanotubes falls in two categories:

(i) Investigation of the mechanical, electrical, or structural properties of individual nanowires/nanotubes. The mechanical properties of nanostructures can be measured by inferring the gripping force from the visible deformation of the nanogripper upon contact, given that the spring constant is known, or by using sensors integrated in the gripping device [15]. The electrical properties can be measured if the end effectors are conducting [8, 14] or if the nanostructure is placed onto electrodes [23]. The structural properties can be measured by transporting the nanostructure to a transmission electron microscope [8] or onto a TEM grid [23].

(ii) Direct application of the nanostructure to perform a function in a device. Microfabricated electrothermal grippers were used in a

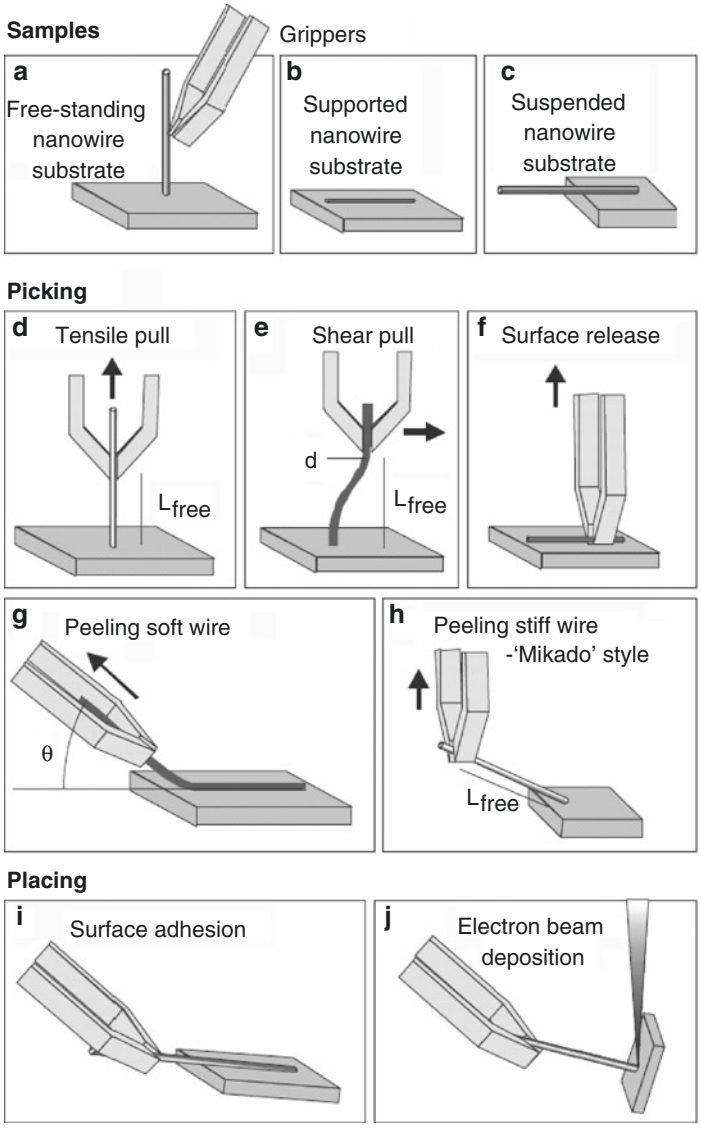


Nanogrippers, Fig. 10 (**a**–**b**) Illustration of pick and place of vertically aligned multiwalled carbon nanotubes from their as-grown position to the tip of a scanning probe tip, where the *blue triangle* illustrates the electron beam-induced deposition used to fasten the nanotube to the probe. (**c**–**d**) The nanogripper during and after mounting of the nanotube tip. (**e**–**f**) Examples of different configurations for the nanotubes on the AFM probe. (**g**–**h**) AFM scanning of deep trenches with and without the supertip, showing a clear improvement of the nanotube-enhanced probe compared to a commercial supersharp AFM probe

semiautomated assembly procedure [23] to first break off multiwalled carbon nanotubes from their as-grown position on a substrate, transport and align them to an atomic force probe, and finally hold them in place, while electron beam-induced deposition was used to fasten the nanotube to the probe.

Figure 10a, b illustrates the basic procedure and the device during (11c) and after (11d) mounting of carbon nanotube is shown as scanning electron images. (11e) and (11f) show two configurations with different angles of the carbon nanotube tip, and (g) and (h) shows AFM scans of 2.5 µm-deep microfabricated trenches with and without the carbon nanotube supertip attached to the AFM probe. The trenches are well resolved by the supertip-enhanced AFM probe.

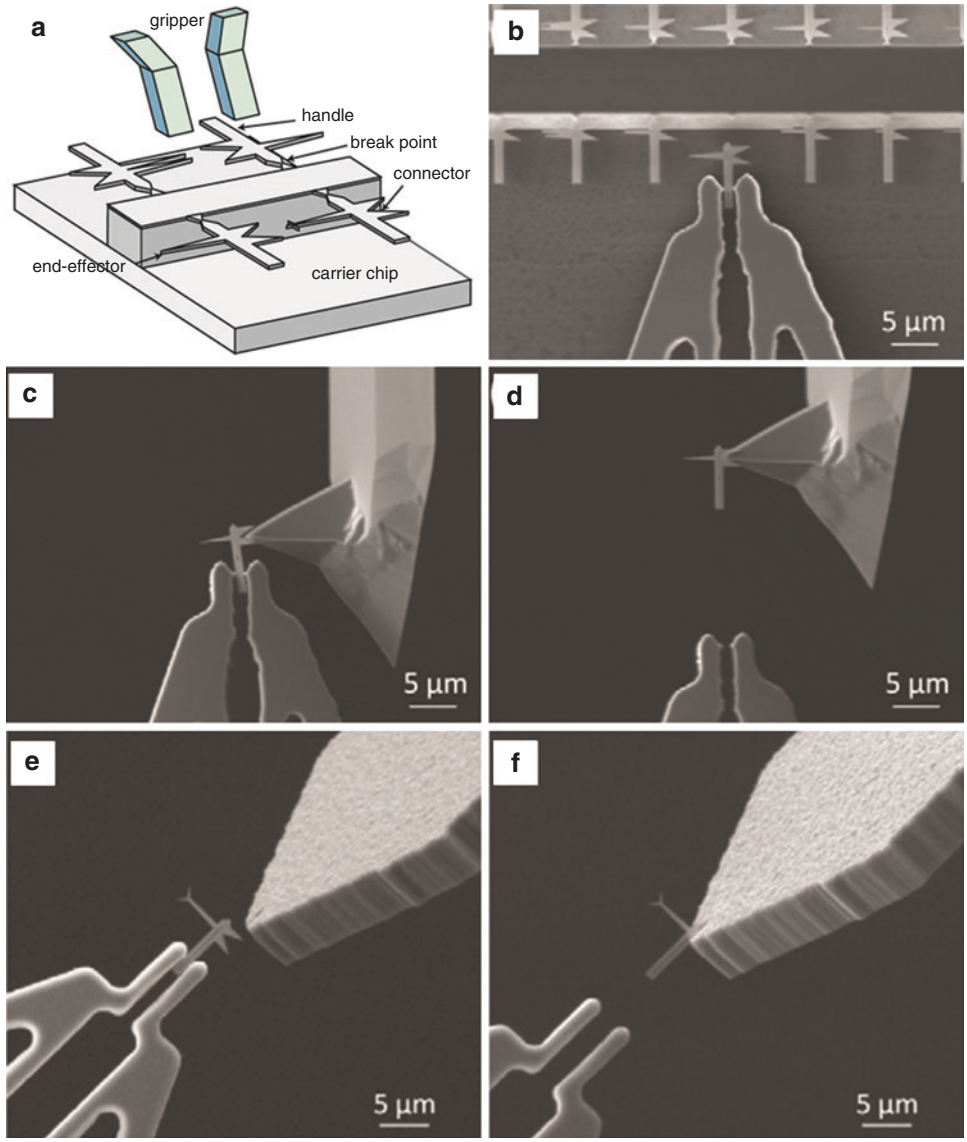
Detaching nanorods from a fixed position using a nanogripper is a nontrivial problem. Mølhave [21] divided the possible arrangements of nanorods on a surface into freestanding,



Nanogrippers, Fig. 11 (**a**–**c**) Initial configurations of rodlike nanostructures. Picking strategies: (**d**) tensile pull, (**e**) shear pull, and (**f**) surface release and (**g**) peeling of soft wire and (**h**) peeling of rigid wire. Strategies for placing using either (**i**) surface adhesion or (**j**) electron beam-induced deposition (EBiD) [21]

supported and (partly) suspended initial states, and compared different pick-and-place strategies for such structures. Cartridges of carbon nanotubes suspended from a razor sharp knife edge were used by [1] to assemble the nanotubes on scanning probe tips to be used as nanotweezers.

Freestanding nanowires and nanotubes are very convenient objects for manipulation: they can be catalytically grown with great uniformity and in large numbers using physical or chemical vapor deposition techniques, with positions and diameter determined by pre-positioned catalytic particles. Thus, it is straightforward to produce large arrays of accessible nanostructures, with well-controlled dimensions, position, and physical properties. The deterministic nature of such a periodic arrangement directly facilitates automation. Detachment of as-grown structures, however, requires a considerably larger force as the often very high maximal strength must be overcome. Mølhave considered mainly two approaches to detach freestanding 1D nanostructures: tensile pull along the longitudinal



Nanogrippers, Fig. 12 (**a**) Illustration of the "nanobit" concept (**b**) arrays of premade scanning probe tips fabricated with electron beam lithography, with a single "nanobit" being detached using a nanogripper. (**c**) Mounting of nanobits onto AFM scanning probe tip. (**d**) The nanobit-enhanced scanning probe ready to scan. (**e**) A sidetip-equipped nanobit tip mounted in a focused ion beam slit in an AFM scanning probe. (**f**) [16]

axis of the nanorod and shear pull in the perpendicular direction with respect to the substrate (sideways). In practical experiments as well as finite element calculations by Andersen et al. [3], it turned out that shear pulling is far the most effective strategy for multiwalled carbon nanotubes. A range of principal sample configurations and gripper operations are shown in Fig. 11.

Quasi-2D Structures: Flakes, Membranes, and Slices

Microgrippers can be used to manipulate thin slices of material, which has been cut out by

focused ion beam milling, onto a transmission electron grid for inspection of the cross-sectional structure. Although the thickness of such slices is often just of order of hundred nanometers to allow transmission of electrons through the slice, the grippers do not have to be particularly small for this task; in fact, a considerable force may be required to detach the TEM samples from their support. Nanogrippers might be envisioned for manipulation of very small flakes of material onto TEM grids or for other plate- or flake-like nanostructures. Kumar and coworkers [16] demonstrated predictable pick and place of nanoscale flake-like scanning probe supertips as shown in Fig. 12. These detachable, 130 nm-thin silicon nitride membrane tips, termed "nanobits," were defined by electron beam lithography and arranged in large reservoirs for convenient access. The most successful strategy for picking up these objects turned out to be grasping at the edges of the flakes with the gripper and breaking the tips off with a sideways motion (see Fig. 12), similar to the strategy used to detach as-grown carbon nanotubes [9]. In these experiments, several nanobits were mounted on AFM tips and used for scanning of deep trenches and sidewalls, showing considerable improvement over conventional AFM probes.

Non-manipulative Applications

Even when not actuated mechanically, nanogrippers provide two individual fingers with a narrow spacing, which can easily be made conductive. This has shown to be convenient for a number of tasks that are not themselves mechanical manipulation including "grasping" nanostructures through dielectrophoresis (Yamahata et al. 2008), electrical characterization ([8, 14, 23]; Yamahata et al. 2008), transporting nanostructures into a TEM for structural analysis, and in situ experiments [8] as well as alternatively placing nanostructures on a TEM grid for analysis. Nanostructures with diameters in the 10 nm range are exceedingly difficult to manipulate through mechanically actuation of dual end effectors.

Conclusion

Despite recent progress in gripper technology as well as robotic manipulation systems, reliable operation of nanogrippers still remains a challenge. The necessity of real-time observation of the mechanical interactions on the nanoscale and surface forces, as well as the difficulty in creating gripper devices small enough to handle sub-100 nm structures, is limiting the transition from literature's numerous proposed nanogripper concepts to devices that sustain practical applications. All in all, state-of-the-art nanogrippers are highly useful tools for basic nanomaterial research, in addition to providing a shortened turnaround time for single device prototyping and perhaps even small-scale production of specialized nanodevices, which involves nanostructures in the 10–1000 nm range.

Cross-References

- ▶ Basic MEMS Actuators
- ▶ Carbon Nanotubes
- ▶ DNA Manipulation Based on Nanotweezers
- ▶ Electron-Beam-Induced Deposition
- ▶ Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)
- ▶ Manipulating
- ▶ Nanorobotic Assembly
- ▶ Nanorobotic Spot Welding
- ▶ Nanorobotics
- ▶ Nanorobotics for MEMS and NEMS
- ▶ Thermal Actuators

References

1. Akita, S., Nakayama, Y.: Manipulation of nanomaterial by carbon nanotube nanotweezers in scanning probe microscope. Jap. J. Appl. Phys. Part 1 Regul. Papers Short Notes Rev. Papers **41**(6B), 4242–4245 (2002)
2. Akita, S., Nakayama, Y., et al.: Nanotweezers consisting of carbon nanotubes operating in an atomic force microscope. Appl. Phys. Lett. **79**(11), 1691–1693 (2001)
3. Andersen, K.N., Petersen, D.H., et al.: Multimodal electrothermal silicon microgrippers for nanotube

manipulation. IEEE Trans. Nanotechnol. **8**(1), 76–85 (2009)
4. Bartenwerfer, M., Eichhorn, V., Fatikow, S., Becker, M., Savenko, A., Yildiz, I., Boggild, P.: Robotics and Automation (ICRA), 2013 IEEE International Conference on, pp. 1730–1735 (2013)
5. Beyeler, F., Neild, A., et al.: Monolithically fabricated microgripper with integrated force sensor for manipulating microobjects and biological cells aligned in an ultrasonic field. J. Microelectromech. Syst. **16**(1), 7–15 (2007)
6. Bhushan, B. (ed.): Handbook of Nanotechnology. Springer□Verlag Berlin, Heidelberg, New York (2004)
7. Boggild, P., Hansen, T.M., et al.: Fabrication and actuation of customized nanotweezers with a 25 nm gap. Nanotechnology **12**(3), 331–335 (2001)
8. Cagliani, A., Wierzbicki, R., et al.: Manipulation and in situ transmission electron microscope characterization of sub-100 nm nanostructures using a microfabricated nanogripper. J. Micromech. Microeng. **20**(3), p. 035009 (2010)
9. Carlson, K., Andersen, K. N., et al.: A carbon nanofibre scanning probe assembled using an electrothermal microgripper. Nanotechnology **18**(34), 345501 (2007)
10. Chen, B.K., Zhang, Y., et al.: From micrograpping to nanogripping. Mems 2010: 23rd IEEE International Conference on Micro Electro Mechanical Systems, Technical Digest, Waichai, pp. 296–299
11. Chen, D.S., Yeh, P.F., et al.: An electrothermal actuator with two degrees of freedom serving as the arm of a MEMS gripper. IEEE Trans. Ind. Electron. **61**(10), 5465–5471 (2014)
12. Chronis, N., Lee, L.P.: Electrothermally activated SU-8 microgripper for single cell manipulation in solution. J. Microelectromech. Syst. **14**(4), 857–863 (2005)
13. Eichhorn, V., Fatikow, S., et al.: Nano Lab: A nanorobotic system for automated pick-and-place handling and characterization of CNTs. ICRA: 2009 I.-E. International Conference on Robotics and Automation, Kobe, vol. 1–7, pp. 1643–1648 (2009)
14. Kim, P., Lieber, C.M.: Nanotube nanotweezers. Science **286**(5447), 2148–2150 (1999)
15. Kim, K., Liu, X.Y., et al.: Nanonewton force-controlled manipulation of biological cells using a monolithic MEMS microgripper with two-axis force feedback. J. Micromech. Microeng. **18**(5) p. 055013 (2008)
16. Kumar, R.T.R., Hassan, S.U., et al.: Nanobits: customizable scanning probe tips. Nanotechnology **20**(39) p. 395703 (2009)
17. Lee, J.S., Park, D.S.W., et al.: Sub-micron metallic electrothermal actuators. J. Micromech. Microeng. **15**(2), 322–327 (2005)
18. Leong, T.G., Randall, C.L., et al.: Tetherless thermobiochemically actuated microgrippers. Proc. Natl. Acad. Sci. U. S. A. **106**(3), 703–708 (2009)
19. Malachowski, K., Jamal, M., et al.: Self-folding single cell grippers. Nano Lett. **14**(7), 4164–4170 (2014)
20. Molhave, K., Hansen, T.M., et al.: Towards pick-and-place assembly of nanostructures. J. Nanosci. Nanotechnol. **4**(3), 279–282 (2004)
21. Molhave, K., Wich, T., et al.: Pick-and-place nanomanipulation using microfabricated grippers. Nanotechnology **17**(10), 2434–2441 (2006)
22. Sahu, B., Taylor, C.R., et al.: Emerging challenges of microactuators for nanoscale positioning, assembly, and manipulation. J. Manufact. Sci. Eng. Trans. Asme. **132**(3), p. 030917 (2010)
23. Sardan, O., Eichhorn, V., et al.: Rapid prototyping of nanotube-based devices using topology-optimized microgrippers. Nanotechnology **19**(49) p. 495503 (2008)
24. Senturia, S.D.: Microsystem Design. Kluwer, Boston (2001)
25. Sigmund, O.: Design of multiphysics actuators using topology optimization – part I: one-material structures. Comput. Methods Appl. Mech. Eng. **190**(49-50), 6577–6604 (2001)
26. Sul, O., Yang E.H.: A multi-walled carbon nanotube-aluminum bimorph nanoactuator. Nanotechnology **20**(9) p. 095502 (2009)
27. Xie, H., Yin, M.N., et al.: In Situ quantification of living cell adhesion forces: single cell force spectroscopy with a nanotweezer. Langmuir **30**(10), 2952–2959 (2014)
28. Yamahata, C., Collard, D., et al.: Silicon nanotweezers: A new biophysical tool for molecular experimentation. Mems 2008: 21st IEEE International Conference on Micro Electro Mechanical Systems, Technical Digest, Tucson, pp. 681–684 (2008a)
29. Yamahata, C., Collard, D., et al.: Silicon nanotweezers with subnanometer resolution for the micromanipulation of biomolecules. J. Microelectromech. Syst. **17**(3), 623–631 (2008b)
30. Zakharov, D., Lebedev, G., et al.: Submicron-sized actuators based on enhanced shape memory composite material fabricated by FIB-CVD. Smart Mater. Struct. **21**(5), 052001 (2012)

Nanohardness

► Nanomechanical Properties of Nanostructures

Nanohardness Tester

► Nanoindentation

Nanohelix

► Nanorobotics for NEMS Using Helical Nanostructures

Nanoimmunology

► Interaction of the Immune System with Nanoparticles

Nanoimprint Lithography

► Nanoimprinting
► Nanoscale Printing

Nanoimprinting

Wei Chen[1] and Chunlei (Peggy) Wang[2]
[1]CAS Key Laboratory of Low-Carbon Conversion Science and Engineering, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China
[2]Department of Mechanical and Materials Engineering, Florida International University, Miami, FL, USA

Synonyms

Nanoimprint lithography

Definition

Nanoimprinting is a type of lithographic technology that provides a simple and efficient approach of patterning various nanostructures with sub-10-nm resolution.

Overview

Nanoimprinting can promise state-of-the-art smaller electronics. Although nanoimprint lithography (NIL) is not yet an industrial process with well-established standard, extensive interests in NIL processes come from a large community of sensor, biochip, and nano-optics manufactures and institutes [1]. Nanoimprinting has received increased research motivation due to its ability to create smaller functional device structures such as nanoscale transistors, single-electron memory, nanofluidic channels, etc., via an accurate and efficient nanolithography process [2].

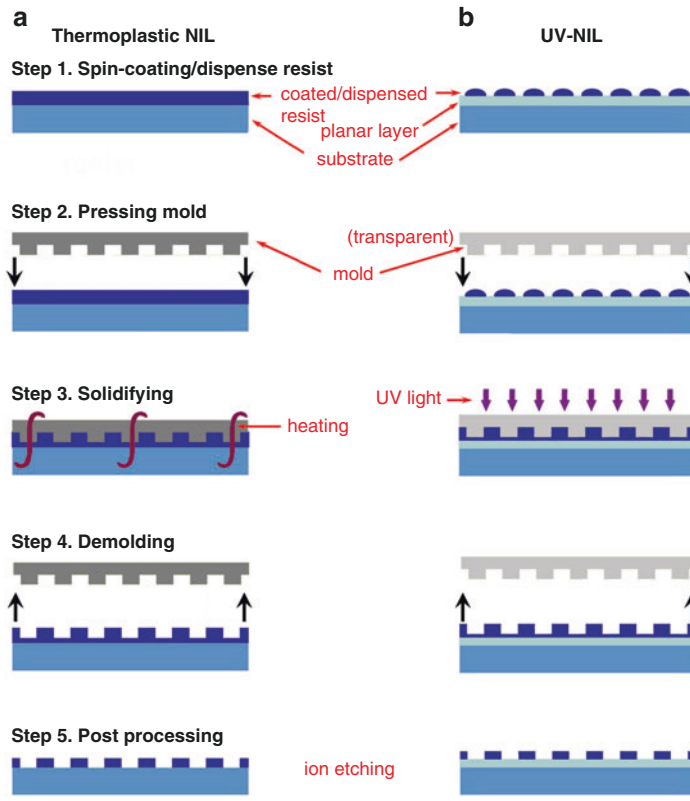
NIL is very similar to conventional hot embossing, which is based on stamping of a pattern into a polymer softened by raising the temperature of the polymer just above its glass transition temperature. The hot embossing stamp can be made in a variety of ways including micromachining from silicon, LIGA (a German acronym for Lithographie (lithography), Galvanoformung (electroplating), Abformung (molding)), and machining using a CNC (computer numerical control) tool (for making large features). A wide variety of polymers have been successfully hot embossed with submicron size features, including PC (polycarbonate) and PMMA (polymethylmethacrylate).

In the current encyclopedia, a general introduction about NIL process, the involved materials such as resist and mold (template), as well as NIL manufacturers are presented. The detailed and intensive discussions about NIL are lucubrated in the related review papers [3, 4].

Key Concepts

Nanoimprinting Process

NIL, i.e., thermal NIL process, was first reported by Chou S. et al. from the University of Minnesota in 1995 [5, 6]. Subsequently, Haisma J. et al. from Philips Research laboratories developed photo-NIL, i.e., UV-NIL process [7]. The processes of NIL mainly involves the following steps: spin-coating or dispensing resist (polymer or monomer) on the substrate, pressing resist layer with



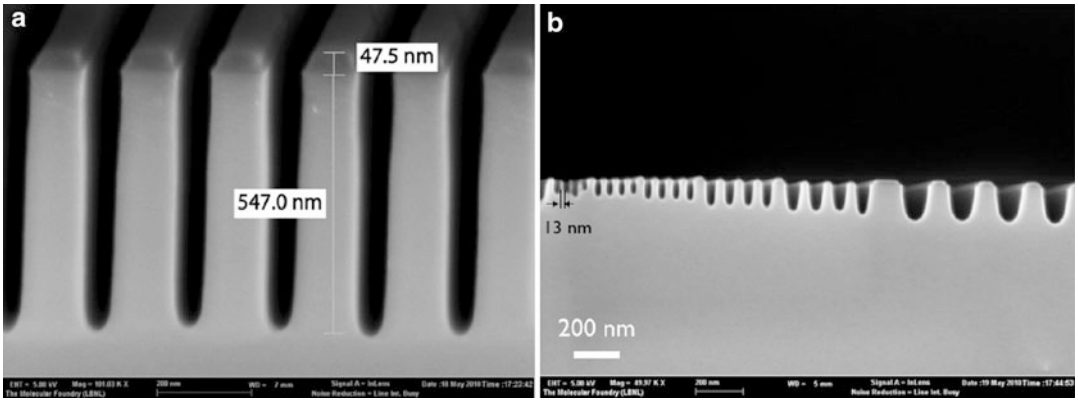
Nanoimprinting, Fig. 1 Schematic of typical (**a**) thermal NIL and (**b**) UV-NIL process

patterned mold, solidifying patterned resist by heating or UV exposure, separating mold (demolding), and post-processing. The schematic drawing of the NIL fabrication process is shown in Fig. 1. Briefly, a resist with thermoplastic or photosensitive characteristic is first spin-coated or dispensed on a wafer substrate. As for photo-NIL process, there is a predeposited planar layer on the substrate. Then, a hard mold containing nanometer-scale surface-relief features is pressed into the above thin resist cast. Note that a transparent mold is requested in the UV-NIL. As shown in Fig. 1 (Step 2 and 3), the mold protrusions are squeezed into the soft resist layer. After the treatment of heating or UV exposure, the patterns of mold are solidified inside the resist. The mold is then removed from the imprinted resist layer, i.e., demolding process. The desired nanostructured patterns in the residual layer are obtained after the appropriate pattern definition process using ion etching posttreatment.

The NIL process is suitable for fabricating various nanodevices, whose basic components can be prepared through traditional UV-NIL process. Figure 2 shows examples of patterns transferred into silicon [8]. The residual nanostructure layer with an aspect ratio greater than six is obtained after reactive ion etching (as depicted in Fig. 2a). Maintaining the fidelity of the resist mask by controlling the residual layer thickness allows transfer features down to 13 nm as shown in Fig. 2b. These various nanostructures offer basic units to fabricate photonics or other functional nanodevices at low cost and high throughput.

Nanoimprinting Resist

The resist materials used in imprinting should be able to deform easily under an applied pressure and have sufficient mechanical strength. The resist should also have good mold-releasing properties to maintain their structural integrity during



Nanoimprinting, Fig. 2 SEM pictures of imprinted gratings after transfer into silicon wafers. (**a**) 90-nm-linewidth gratings with 547-nm-depth into silicon and 47.5-nm-thick UV-NIL resist remaining after the etching steps. (**b**) Grating with variable pitch/linewidth. The minimum feature size is 13 nm with an etching depth of around 50 nm [8]

the demolding process since imprint lithography makes a conformal replica of the surface-relief patterns by mechanical embossing.

Thermoplastic Resist Materials

The thermoplastic polymers such as PMMA and PS (polystyrene) can be easily deposited on the substrate with desired thicknesses by spin-coating process. These polymers are usually used as imprint resists in thermal NIL. The glass transition temperature (T_g) of the polymer is a key parameter for the determination of the solidifying treatment temperature in thermal molding processes. Typically a suitable imprint temperature is chosen 70–90 °C above the resist's T_g, so that the polymer materials reach a viscous flow state. In order to preserve the imprinted pattern, the mold and the imprinted resist must be cooled down to below T_g. In addition, there is a trade-off between the imprinting temperature and the thermal stability, which indicates that although a low-T_g material can be used in NIL for the sake of reducing the processing temperature, the imprinted patterns are also unstable and tend to deform at temperatures close to the imprinting temperature.

In a NIL process, the polymers also need to have a sufficiently low viscosity for efficient squeeze to complete the imprinting process within a practical time frame [9]. The viscosity of a polymer material not only depends on the temperature, but also depends strongly on the polymer's molecular weight (M_w) relative to the so-called critical molecular weight (M_c) of a given polymer. The M_c can be interpreted as the molecular weight at which a temporary network of entanglements spans over macroscopic dimensions. In practice, low-molecular-weight polymers with $M_w < M_c$ can be imprinted at lower temperatures, lower pressures, or within shorter times. Therefore, the choices of T_g and M_w are both important in maintaining the structural stability of the imprinted patterns.

Various alternatives have been exploited to lower the viscosity of thermoplastic materials. These include dissolving the polymer in its monomer [10] or another solvent and using a PDMS (polydimethylsiloxane) stamp for solvent evaporation [11]. It is inferred that thermally curable or thermosetting polymers are excellent resist for NIL because of the possibility of low-pressure imprinting and good mechanical integrity after cross-linking by thermal treatment.

UV-NIL Resist Materials

Different from thermoplastic materials requiring high temperature and pressures, photo-NIL (UV-NIL) materials can be handled at ambient temperatures. UV-curable NIL materials are composed of a mixture of monomers or prepolymers and a suitable photoinitiator, and often chemicals are added which decrease the effect of radical scavengers on photopolymerization [12].

The monomer fluid with low Young's modulus and low viscosity is less sensitive to the effects of pattern density in imprinting process, which allows the use of a small-area mold and the patterning of large-area substrate by a step-and-repeat process. Immediately during contact of the stamp with the liquid mixture, filling of the mold starts by capillary forces, which pulls the stamp toward the substrate. Therefore, the general strategy is as follows: Low viscosities are needed both for rapid dispensing and filling of mold cavities. Thin resin layers on top of a thicker transfer layer are used to achieve a homogeneous film thickness. Cross-linking and photopolymer conversion are adapted to achieve high curing speed and high etch resistance in the following breakthrough plasma etching process. Shrinkage and etching rate need to be controlled for optimum pattern transfer. Often trade-offs are needed to achieve a good balance between good physicochemical properties wetting and curing kinetics and the suitability for the NIL process [13]. For specific applications such as the dual damascene process for the structuring of interconnects on microchips, resists with adapted dielectric properties have been developed [14].

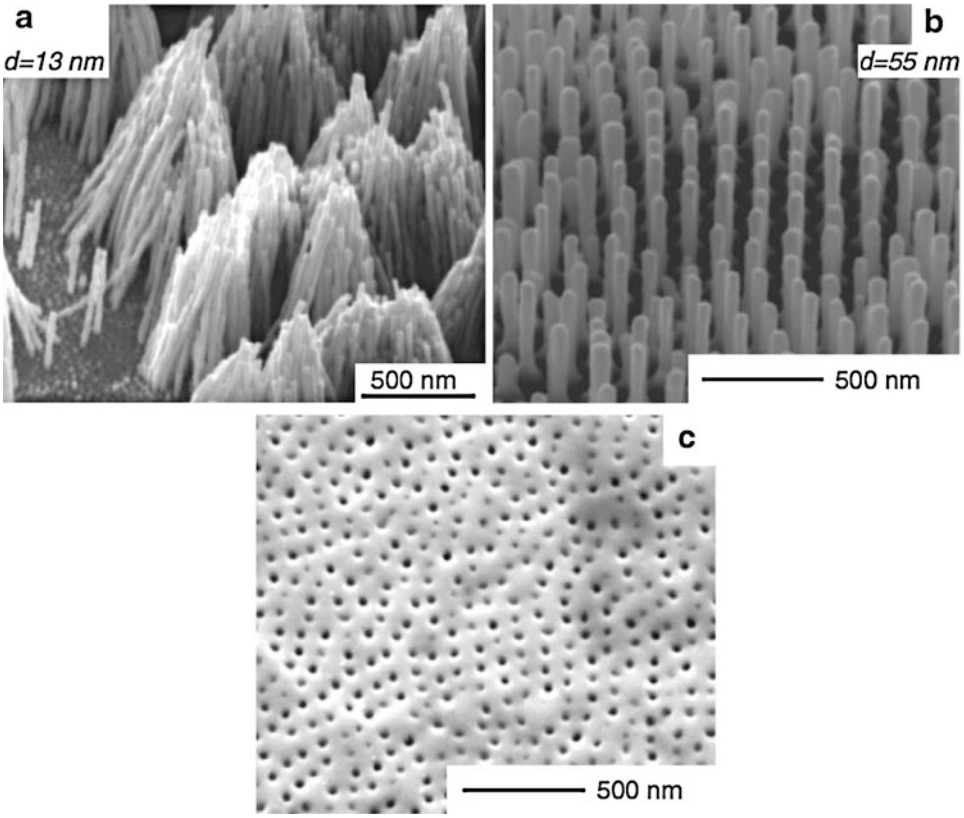
It should be recognized that NIL cannot only be used to form patterns in a polymer resist but can also be extended to create desired structures in many other polymer systems, especially those that have special functionalities, or can be used to form functional polymer device structures directly.

Nanoimprinting Mold

The mold in the nanoimprint technique plays the same role as the photomask in photolithography. The mold used in NIL can essentially be any type of solid material that has a high strength and durability, which can be made of metals, dielectrics, or semiconductors. Various hard materials such as Si, SiO_2, SiC, silicon nitride, and sapphire are suitable for imprinting molds [15]. Diamond is also investigated as a potential mold material for NIL [16]. These studies showed that Si and SiO_2 have sufficient hardness and durability for the nanoimprint application, which make themselves a very good pair for the NIL process. In addition, the UV transparent properties are essential for mold materials of UV-NIL, which render UV light pass through mold itself and irradiate the resist layer.

After spin-coated with a resist, the mold materials are patterned with various features using UV lithography, interference lithography, and electron beam lithography depending on the size requirements. A minimum lateral feature size can be obtained using electron beam lithography; interference lithography is fitted for large-area periodic features while UV lithography for microscale and larger features. Then, a hard masking layer, such as a metal, can be deposited over the patterned resist template, followed by a lift-off process that removes the resist template and the material on top, leaving a patterned mask layer on the substrate. Lastly reactive ion etching process is used to selectively etch away the mold materials in the unmasked area and produce surface-relief features required for NIL.

Silicon-based molds are expensive to make and are not very durable. Schroers J. et al. at Yale University developed the nanopatterning of amorphous metals which can be used as inexpensive templates for nanoimprinting [17]. These bulk metallic glasses (BMGs) could be ideal materials for small-scale applications due to their superior mechanical and homogenous properties. There molds with features as small as 13 nm can be used to imprint onto polymers and other materials. Figure 3 shows SEM images of Pt-BMG nanorods with 13-nm (Fig. 3a) and 55-nm (Fig. 3b) diameters fabricated by embossing on porous alumina. The Pt-BMG substrate with nanorods was pressed into a PMMA sheet at 160 °C. Figure 3c shows an SEM image of 55-nm-diameter holes imprinted in Pt-BMG using the crystallized Pt-BMG mold. The results demonstrate the ability of metallic glasses to precisely replicate mold features ranging from 250 to 13 nm. The two attractive features of metallic glass molding – the ability to replicate features from tens of micrometers down to 13 nm and the possibility of using BMGs as nanotemplates – stem directly from the unique softening behavior and homogeneous structure of metallic glasses.



Nanoimprinting, Fig. 3 SEM images of (**a**) Pt-BMG mold with 13-nm-diameter rods fabricated by embossing on porous alumina; (**b**) Pt-BMG mold with 55-nm-diameter rods; (**c**) The Pt-BMG mold, after crystallization, imprints holes into the same Pt-BMG [17]

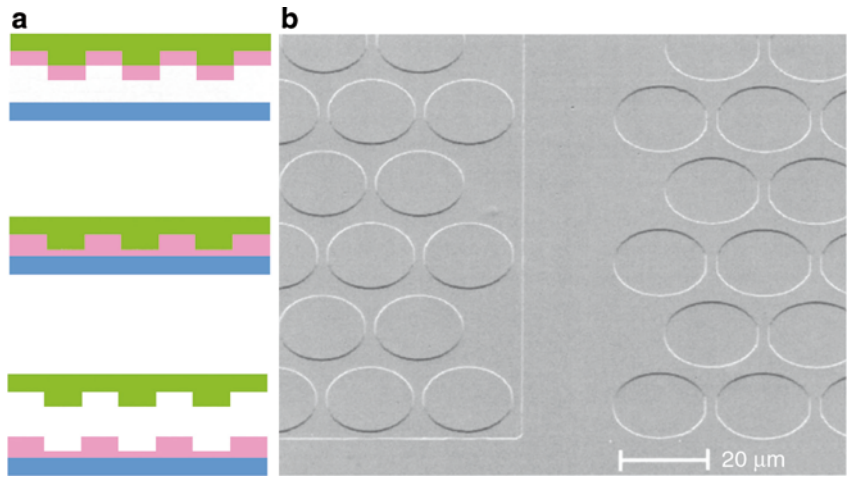
Variation of Nanoimprinting

Some innovative processes such as roller-type NIL, combined thermal/UV-NIL, and reverse tone NIL are developed to meet various requirement, which make conventional NIL more flexible and practicable.

Conventional NIL process cannot significantly improve the throughput in the patterning of large-area product with low cost because it is not a continuous process. To overcome this problem, a continuous roller-type NIL is proposed and developed. In roller-type NIL process, only the adjacency area of contact line is pressed for a given time, significantly reducing the printing force and providing better uniformity [18]. Jacob J. et al. demonstrated that large-area high-speed continuous roll-to-roll nanoimprinting of 1D and 2D micron to sub-100 nm features on flexible substrate using perfluoropolyether composite molds on a custom designed roll-to-roll nanoimprinter [19].

Combined thermal/UV-NIL can address the issue of replicating patterns with various sizes and complexity by the separate lithography steps [20]. Sole NIL process experiences difficulties in replicating larger features or resulting in incomplete pattern.

In order to obtain nanosized patterns on the flexible substrate, such as polymer membrane, a reversal imprinting technique was developed [21]. Different from conventional NIL process, a resist layer is first spin-coated on a patterned hard mold and then transferred to a substrate under an elevated temperature and pressure (as shown in Fig. 4a). The reverse NIL process can be accomplished under different pattern transfer modes by controlling imprinting temperature and degree of surface planarization of the spin-coated mold.



Nanoimprinting, Fig. 4 (**a**) Schematic illustrations of the pattern transfer processes in reversal nanoimprinting at temperatures well above the glass transition temperature; (**b**) patterns in PMMA created by reversal imprinting at 175 °C on a 50-mm-thick Kapton film: 190-nm-deep micrometer-sized mold coated with a 7 % solution [21]

Figure 4b shows PMMA patterns created by reversal imprinting at 175 °C after spin-coating the mold with 190-nm-deep micrometer-sized features.

In addition, recently a die nanoimprinting technique has been developed, which can effectively enhance the stress uniformity within the supercooled liquid of glassy alloys [22]. By employing the technique, large-area amorphous metallic nanowire arrays with nanowire diameter of about 55 nm and uniform length around 650 nm have been prepared.

Main Manufacturers for Nanoimprinting

The NIL machines are very important for the development and applications of NIL technology. In this review, the most famous manufactures in the world are listed. The general information about these enterprises are excerpted from their homepage, and more details can be referred to corresponding website, which are attached in the last sentence of every paragraph.

AMO

AMO, i.e., Gesellschaft für Angewandte Mikro- und Optoelektronik mbH, is an important enterprise in the area of nanofabrication. AMO supports customers with focused research and development, prototyping, and contract manufacturing. AMO offers the entire infrastructure required for nanofabrication for semiconductor-based applications and related technologies. Nanofabrication, nanoelectronics, nanophotonics, and biotechnology are AMO's competences. AMO is a research service provider for nanofabrication, a pathfinder for your exceptional demands. For more information, readers can visit the website of AMO: http://www.amo.de.

EV

EV Group (EVG) is a leading supplier of wafer bonding and lithography equipment for the MEMS, nanotechnology, and semiconductor markets. EVG devotes to be the first in exploring new techniques and serving next-generation applications of micro- and nanofabrication technologies. The basis for EVG's success is their products: lithography, bonding, and imprint systems. EVG holds the dominant share of the market for all types of wafer bonding equipment and is the market and technology leader in lithography and nanoimprinting. The official website of EV is http://www.evgroup.com.

NLS

Nanolithosolutions (NLS) Inc. empowers researchers and technology developers through their easy-to-use nanolithography platform to develop next generation of nanotechnology applications. NLS provides complete turnkey solution for nanolithography applications. The product offering includes nanofabrication equipments, all the necessary chemicals and materials, imprinting mold preparation process and imprint fabrication process, and custom process development services. The website of NLS Inc. is http://www.nanolithosolution.com.

MII

Molecular Imprints Inc. (MII) is the market and technology leader for nanopatterning systems and solutions and uses its innovative Jet and Flash™ Imprint Lithography (J-FIL™) technology to provide high-resolution, low-cost-of-ownership nanopatterning capability for flat panel display, hard disk drive industries, light-emitting diodes, life science, and other high tech industries. J-FIL™ has further application in emerging light-emitting diodes (LED), solar energy, and biotechnology markets. Http://www.molecularimprints.com is the website of MII.

Nanonex

Nanonex was founded in 1999 by Stephen Y. Chou, a Professor of Princeton University and a renowned pioneer in nanotechnology. The mission of Nanonex is to deliver, today, user-friendly NIL tools and solutions to a broad spectrum of markets and to both experts and nonexperts of microfabrication. Nanonex products include NIL tools, resists, masks, and processes for a variety of NIL including thermal and photo-curable NIL, as well as direct imprinting of materials. Nanonex's current team consists of seasoned business and technical members from industry as well as the experts in nanoimprint technology and applications. The Nanonex solution makes a turnkey nanoimprint operation today a reality. The website of Nanonex is http://www.nanonex.com.

OAI

OAI is a Silicon Valley-based manufacturer of advanced precision equipment for the solar, photovoltaic, semiconductor, MEMS, and microfluidics industries. The company offers a broad portfolio of field-proven products that include mask aligners, UV exposure systems, UV light sources, solar simulators, I-V testers, and reference cells. The website of OAI is http://www.oainet.com.

Obducat

Obducat is the world-leading supplier of lithography solutions for manufacturing and replication of advanced micro- and nanoscale structures. Obducat provides viable and cost-effective lithography solutions that will give a competitive edge to the customers, enabling them to deliver breakthrough applications and achieve improved profitability and success. Obducat is the first company to commercialize NIL. Today it is the market leader with the largest installed base worldwide of NIL. The details about Obducat are shown in the website http://www.obducat.com.

SUSS MicroTec

SUSS MicroTec is a supplier of equipment and process solutions for microstructuring applications with more than 60 years of engineering experience. The solution portfolio covers a comprehensive range of products and solutions for backend lithography, wafer bonding, and photomask processing, complemented by micro-optical components. The website of SUSS MicroTec is http://www.suss.com.

Martini Tech

Martini Tech is a company specialized in micro- and nanofabrication based in Japan. Martini Tech offers the following services: NIL sample master mold and custom nanoimprint mold design and manufacturing, UV nanoimprint and hot embossing replica on polymer, MEMS design and development, MEMS foundry, patterned sapphire substrate manufacturing, MOCVD deposition of gallium nitride on sapphire substrates for LED applications, sputtering deposition, and other micro- and nano-manufacturing services.

The official website of Martini Tech is http://www.martini-tech.com.

Future Directions for Nanoimprinting

Nanoimprinting technique offers overwhelming advantages for nanoscale device fabrications, including low cost, resolution scalability, and pattern repeatability. The simplicity of NIL has made it appealing to researchers in various fields, such as next-generation HDD storage and future semiconductor devices. However, standard processes or criterions about NIL are not establish, which affect the NIL industrial application and development. In addition, the fabrications of high-quality templates with high-resolution periodic features are still key issues limiting the popularization of NIL. Despite all difficulties, a large community of researchers and enterprises, such as Molecular Imprints Inc. (MII) and AMO, are trying to establish standard NIL processes and develop significantly advanced materials and molds. Toshiba's most advanced production process for NAND flash memory available is 19 nm and will work with fab-tool maker Canon to develop 15-nm NAND chips with volume production slated for 2015. IBM also focused on understanding the basic physics and chemistry of the nanoimprint process, from initial resist dispense to removal of the template after cure, and use that knowledge to design optimum resist materials and interfacial surface treatments. In the near future, NIL will become a preferred and universal process in nanofabrication technology.

Cross-References

- ▶ DUV Photolithography and Materials
- ▶ Nanoscale Printing

References

1. Ofir, Y., Moran, I.W., Subramani, C., Carter, K.R., Rotello, V.M.: Nanoimprint lithography for functional three-dimensional patterns. Adv. Mater. **22**, 3608–3614 (2008)
2. Glinsner, T., Kreindl, G., Kast, M.: Nanoimprint lithography. Optik & Photonik **2**, 42–45 (2010)
3. Guo, L.J.: Nanoimprint lithography: methods and materials requirements. Adv. Mater. **19**, 495–513 (2007)
4. Schift, H.: Nanoimprint lithography: an old story in modern times? A review. J Val. Sci. Technol. B **26**, 458–480 (2008)
5. Chou, S.Y., Krauss, P.R., Renstrom, P.J.: Imprint of sub-25 nm vias and trenches in polymers. Appl. Phys. Lett. **67**, 3114–3116 (1995)
6. Chou, S.Y., Krauss, P.R., Renstrom, P.J.: Imprint lithography with 25-nanometer resolution. Science **272**, 85–87 (1996)
7. Haisma, J., Verheijen, M., Heuvel, K.V.D.: Mold-assisted nanolithography: a process for reliable pattern replication. J. Vac. Sci. Technol. B **14**, 4124–4128 (1996)
8. Peroz, C., Dhuey, S., Volger, M., Wu, Y., Olynick, D., Cabrini, S.: Step and repeat UV nanoimprint lithography on pre-spin coated resist film: a promising route for fabricating nanodevices. Nanotechnology **21**, 445301 (2010)
9. Hirai, Y.: Polymer science in nanoimprint lithography. J. Photopolym. Sci. Technol. **18**, 551–558 (2005)
10. Yung, G.Y., Ganapathiappan, S., Li, X., Ohlberg, D.A. A., Olynick, D.L., Chen, Y., Tong, W.M., Williams, R. S.: Fabrication of molecular-electronic circuits by nanoimprint lithography at low temperatures and pressures. Appl. Phys. A **78**, 1169–1173 (2004)
11. Khang, D.Y., Lee, H.H.: Room-temperature imprint lithography by solvent vapor treatment. Appl. Phys. A **76**, 870–872 (2000)
12. Long, B.K., Keitz, B.K., Willson, C.G.: Materials for step and flash imprint lithography (S-FIL®). J. Mater. Chem. **17**, 3575–3580 (2007)
13. Colburn, M., Choi, B.J., Sreenivasan, S.V., Bonnecaze, R.T., Willson, C.G.: Ramifications of lubrication theory on imprint lithography. Microelectron. Eng. **75**, 321–329 (2004)
14. Schmid, G.M., Stewart, M.D., Wetzel, J., Palmieri, F., Hao, J.J., Nishimura, Y., Jen, K., Kim, E.K., Resnick, D.J., Liddle, J.A., Willson, C.G.: Implementation of an imprint damascene process for interconnect fabrication. J. Vac. Sci. Technol. B **24**, 1283–1291 (2006)
15. Guo, L.J.: Recent progress in nanoimprint technology and its applications. J. Phys. D **37**, R123–R141 (2004)
16. Taniguchi, J., Tokano, Y., Miyamoto, I., Komuro, M., Hiroshima, H.: Diamond nanoimprint lithography. Nanotechnology **13**, 592–596 (2002)
17. Kumar, G., Tang, H.X., Schroers, J.: Nanomoulding with amorphous metals. Nature **457**, 868–872 (2009)
18. Tan, H., Gilbertson, A., Chou, S.Y.: Roller nanoimprint lithography. J. Vac. Sci. Technol. B **16**, 3926–3928 (1998)
19. Jacob, J., Yuying, T., Jonathan, P.R., James, J.W., Kenneth, R.C.: Large-area, continuous roll-to-roll nanoimprinting with PFPE composite molds. Nanotechnology **24**, 505307 (2013)

20. Pfeiffer, K., Fink, A., Gruetzner, G., Bleidiessel, G., Schulz, H., Scheer, H.: Multistep profiles by mix and match of nanoimprint and UV lithography. Microelectron. Eng. **57–58**, 381–387 (2001)
21. Huang, X.D., Bao, L.R., Cheng, X., Guo, L.J., Pang, S.W., Yee, A.F.: Reversal imprinting by transferring polymer from mold to substrate. J. Vac. Sci. Technol. B **20**, 2872–2876 (2002)
22. Xue, L., Yang, S., Jin-Feng, L., Na, C., Ke-Fu, Y.: Large-area and uniform amorphous metallic nanowire arrays prepared by die nanoimprinting. J. Alloys Compd. **605**, 7–11 (2014)

Nanoindentation

Bharat Bhushan and Manuel L. B. Palacio
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Elastic modulus tester; Nanofatigue tester; Nanohardness tester; Nanoscratch tester

Definition

Nanoindentation refers to the experiment conducted using a depth-sensing indenter at the nanometer or sub-micrometer scale. The nanoindentation apparatus continuously monitors the load and the position of the indenter relative to the surface of the specimen (depth of an indent) during the indentation process. Sharp indenters can be used to measure hardness, elastic modulus, continuous stiffness, scratch resistance, film-substrate adhesion, residual stresses, time-dependent creep and relaxation properties, fracture toughness, and fatigue.

Overview

Mechanical properties of solid surfaces and thin films are of interest as the mechanical properties affect the tribological performance of surfaces [1–5]. Among the mechanical properties of interest, one or more of which can be obtained using commercial and specialized indentation testers, are elastic–plastic deformation behavior, hardness, Young's modulus of elasticity, scratch resistance, film-substrate adhesion, residual stresses, time-dependent creep and relaxation properties, fracture toughness, and fatigue. Indentation measurements can assess structural heterogeneities on and underneath the surface, such as diffusion gradients, precipitates, presence of buried layers, grain boundaries, and modification of surface composition. Physical contacts at sliding interfaces in magnetic storage devices and micro/nanoelectromechanical systems (MEMS/NEMS) occur at very low loads; thus, friction and wear of sliding surfaces is primarily controlled by the physical and chemical properties of a few surface atomic layers. In this entry, the applications of nanoindentation to nanotechnology will be illustrated for the digital micromirror device (DMD), probe-based ferroelectric information storage, and biological MEMS sensor devices.

Nanoindentation Apparatus

In the depth-sensing indentation instruments, the load-indentation depth is continuously monitored during the loading and unloading processes. Commercially available nanoindenters are made by various manufacturers, some of which include Agilent, Oak Ridge, TN; Hysitron, Minneapolis, MN; CSIRO, Lindfield, Australia; CSM Instruments SA, Neuchatel, Switzerland, and Micro Materials, Wrexham, UK.

An example of a common nanoindentation apparatus is the Nanoindenter XP, manufactured by Agilent. The apparatus continuously monitors the load and the position of the indenter relative to the surface of the specimen (depth of an indent) during the indentation process. The area of the indent is then calculated from knowledge of the geometry of the tip of the diamond indenter. Mechanical properties measurements can be made at a minimum penetration depth of about 20 nm (or a plastic depth of about 15 nm) [6]. Specifications and commonly used operating parameters for commercial nanoindenters are given in Table 1.

Nanoindentation, Table 1 Specification and commonly used operating parameters for a commercial NanoIndenter

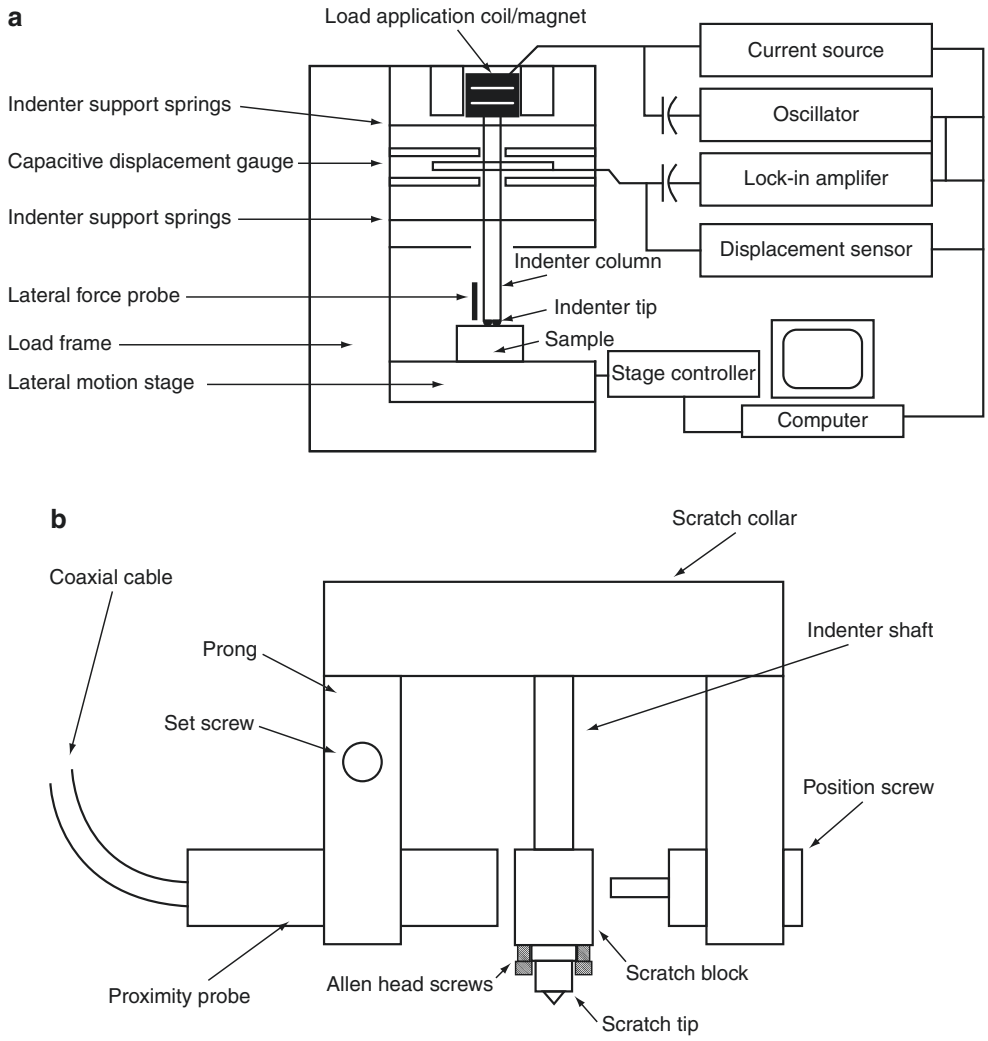
Load range	
Standard heads	0–500 mN
High load head	0–10 N
Load resolution	
Standard head	±50 nN
High load head	±50 nN
Vertical displacement range	0–500 mm
Vertical displacement resolution	±0.05 nm
Typical approach rate	10 nm/s
Typical indentation load rate	10 % of peak load/s
Typical indentation displacement rate	10 % peak displacement/s
Optical microscope magnification	Up to 1,500 *x*
AFM objective magnification	Up to one million *x*
Spatial resolution of the X-Y-Z table	±400 nm in the *X* and *Y* directions
Area examined in a single series of indentations	150 × 150 mm
Minimum penetration depth	~20 nm
Continuous stiffness option	
Frequency range	10–150 Hz
Time constant	0.33 s
Smallest measurable distance	0.1 nm
Scratch and tangential force option	
Scratch velocity	Max. 100 mm/s with 20 points/mm
Tangential displacement range	2 mm
Tangential displacement resolution	100 nm
Tangential load resolution	10 mN
Minimum measurable tangential load	10 mN

The NanoIndenter XP consists of three major components: the indenter head, an optical/atomic force microscope, and an *x*-*y*-*z* motorized precision table for positioning and transporting the sample between the optical microscope and indenter, Fig. 1a. The load on the indenter is generated using a voice coil in permanent magnet assembly, attached to the top of the indenter (loading) column. The generated load is the vector product of the current through the coil and the magnetic field strength of the permanent magnet. (Other methods of load application are found in other instruments, e.g., electrostatic.) Two interchangeable indenter heads are available with a load range of 50–840 mN. The displacement of the indenter is measured using a three-plate capacitive displacement sensor. The indenter column is attached to the moving plate. Two outer plates are maintained at equal and opposite drive voltages. The output voltage of the center-pickup plate is uniquely related to the position of that plate in the capacitive gap. The displacement system is calibrated using a laser-interferometric calibration system. This plate-and-indenter assembly is supported by two leaf springs cut in such a fashion to have very low stiffness. The motion is damped by airflow around the central plate of the capacitor, which is attached to the loading column.

A diamond tip is attached at the bottom of the indenter rod. The indenter head assembly is rigidly attached to the "U" beam below which the *x*-*y*-*z* table rides, Fig. 1a. The optical microscope is also attached to the beam. The position of an indent on a specimen is selected using the optical microscope. The specimens are held on an *x*-*y*-*z* table whose position relative to the microscope or the indenter is controlled with a joystick. The spatial resolution of the position of the table in the *x*-*y* plane is ±250 nm and its position is observed on the computer screen. The nanoindenter is placed on a vibration-isolation table, and the entire apparatus is enclosed in a heavy wooden cabinet to ensure the thermal stability of the samples.

The Indenters

The main requirements for the indenter are high elastic modulus, no plastic deformation, low friction, smooth surface, and a well-defined geometry that is capable of making a well-defined indentation impression. The first four requirements are satisfied by choosing the diamond material for the tip. A well-defined perfect tip shape is difficult to achieve. Berkovich is a three-sided pyramid and provides a sharply pointed tip compared with the Vickers or Knoop indenters, which are four-sided pyramids and have a slight offset (0.5–1 mm) [6]. Because any three nonparallel planes intersect at a single point, it is relatively easy to grind a



Nanoindentation, Fig. 1 Schematics of the nanoindenter XP (**a**) showing the major components – the indenter head, a *X-Y-Z* motorized precision table, and tangential force option, and (**b**) detailed schematic of the tangential force option hardware (not to scale and the front and rear prongs not shown)

sharp tip on an indenter if Berkovich geometry is used. However, an indenter with a sharp tip suffers from a finite but an exceptionally difficult-to-measure tip bluntness. In addition, pointed indenters produce a virtually constant plastic strain impression and there is the additional problem of assessing the elastic modulus from the continuously varying unloading slope. Spherical indentation overcomes many of the problems associated with pointed indenters. With a spherical indenter, one is able to follow the transition from elastic to plastic behavior and thereby define the yield stress [6]. However, a sharper tip is desirable, especially for extremely thin films requiring shallow indentation. Therefore, the Berkovich indenter is most commonly used for measurements of nanomechanical properties.

The Berkovich indenter, directly brazed to a 304 stainless steel holder, is a three-sided (triangular-based) pyramidal diamond, with a nominal angle of 65.3° between the (side) face and the normal to the base at apex, an angle of 76.9° between edge and normal, and with a radius of the tip less than 0.1 μm.

Another three-sided pyramidal indenter, the cube corner indenter, can displace more than three times the volume of the Berkovich indenter at the same load, thereby producing much higher stresses and strains in the vicinity of the contact and reducing the cracking threshold. This makes this indenter ideal for the estimation of fracture toughness at relatively small scales. The spherical indenter initiates elastic contact and then causes elastic–plastic contact at higher loads, and is well suited for the examination of yielding and work hardening.

Nanoscratch and Lateral Force Measurements

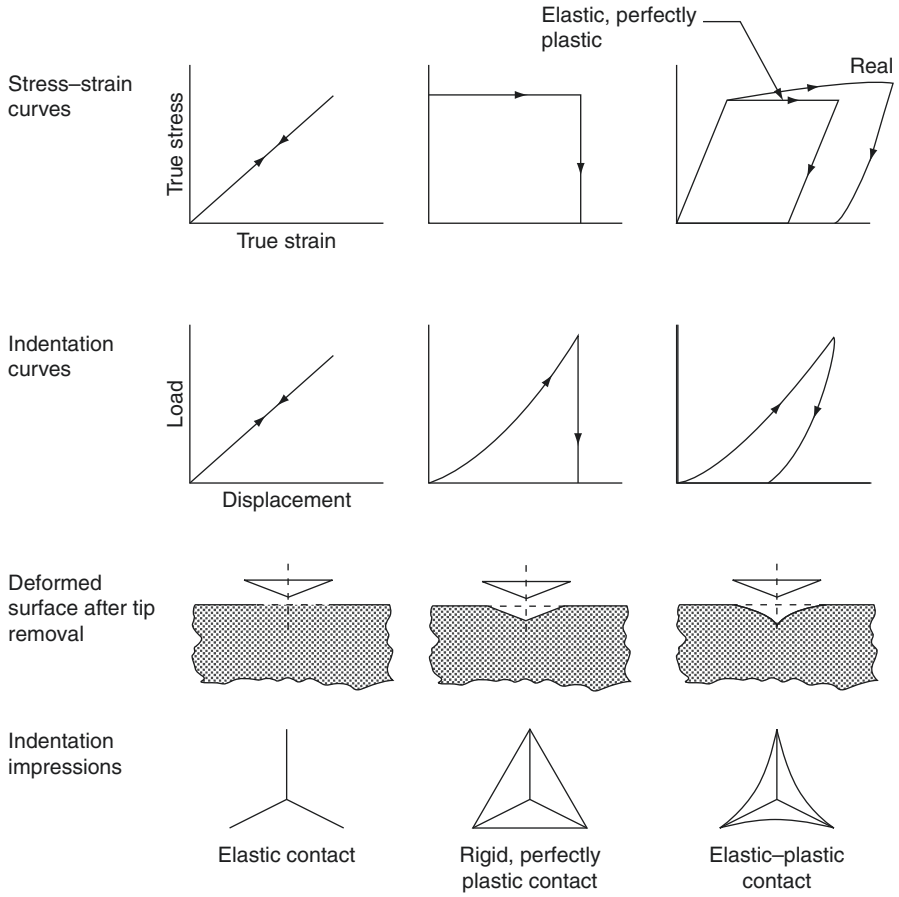
The nanoscratch and lateral force option which allows making of the scratches of various lengths at programmable loads is described here. Lateral (friction) forces can also be measured simultaneously. The additional hardware for the lateral force option includes a set of proximity (capacitance) probes for measurement of lateral displacement or force in the two lateral directions along x and y, and a special "scratch collar" which mounts around the indenter shaft with hardness indenter, Fig. 1b. A scratch block is mounted on the end of the indenter shaft, in line with the proximity probes and the positioning screws. The scratch tip is attached to the scratch block with two Allen head screws. The scratch tip can be a Berkovich indenter or a conventional conical diamond tip with a tip radius of about 1–5 mm and an included angle of 60–90° (typically 1 μm of tip radius with 60° of included angle). A larger included angle of 90° may be desirable for a more durable tip. The tip radius should not be very small because it will get blunt readily [7].

During scratching, a load is applied up to a specified indentation load or up to a specified indentation depth, and the lateral motion of the sample is measured. In addition, of course, load and indentation depth are monitored. Scratches can be made either at the constant load or at ramp-up load. Measurement of lateral force allows the calculations of the coefficient of friction during scratching. As an example, for the Nanoindenter XP, the resolution of the proximity probe provides resolution of lateral force of about 2 mN; therefore, a minimum load of about 20 mN can be measured. Consequently, a minimum normal load of about 0.2 mN should be used for a sample with coefficient of friction of about 0.1 using the Nanoindenter XP. Microscopy of the scratch produced at ramp-up load allows the measurement of critical load required to break up of the film (if any) and scratch width and general observations of scratch morphology. Typically, draw acceleration (mm/s^2), and draw velocity (mm/s) are 10 mm/s^2 and 5 mm/s, respectively.

Analysis of Indentation Data

An indentation curve is the relationship between the load W and the displacement (or indentation depth or penetration depth) h, which is continuously monitored and recorded during indentation. Stress–strain curves, typical indentation curves, the deformed surfaces after tip removal, and residual impressions of indentation for ideal elastic, rigid-perfectly plastic and elastic-perfectly plastic, and real elastic–plastic solids are shown in Fig. 2. For an elastic solid, the sample deforms elastically according to Young's modulus, and the deformation is recovered during unloading. As a result, there is no impression of the indentation after unloading. For a rigid-perfectly plastic solid, no deformation occurs until yield stress is reached, when plastic flow takes place. There is no recovery during unloading and the impression remains unchanged. In the case of elastic–plastic solid, it deforms elastically according to Young's modulus and then it deforms plastically. The elastic deformation is recovered during unloading. In the case of an elastic-perfectly plastic solid, there is no work hardening.

All engineering surfaces follow real elastic–plastic deformation behavior with work hardening [1, 3, 5, 8, 9]. The deformation pattern of a real elastic–plastic sample during and after indentation is shown schematically in Fig. 3a [10]. The contact depth (h_c) is defined as the depth of indenter in contact with the sample under load. The depth measured during the indentation (h) includes the depression of the sample around the indentation in addition to the contact depth. The depression of the sample around the indentation ($h_s = h - h_c$) is caused by elastic displacements and must be subtracted from the data to obtain the actual depth of indentation or actual hardness. At peak load, the load and



Nanoindentation, Fig. 2 Schematics of stress–strain curves, typical indentation curves, deformed surfaces after tip removal, and residual impressions of indentation, for ideal elastic, rigid-perfectly plastic, elastic-perfectly plastic (ideal), and real elastic–plastic solids

displacement are W_{max} and h_{max}, respectively, and the radius of the contact circle is a. Upon unloading, the elastic displacements in the contact region are recovered and when the indenter is fully withdrawn, the final depth of the residual hardness impression is h_f. A load–displacement curve schematic is shown in Fig. 3b.

Oliver and Pharr [10] developed an expression for h_c at the maximum load (required for hardness calculation) from h_{max},

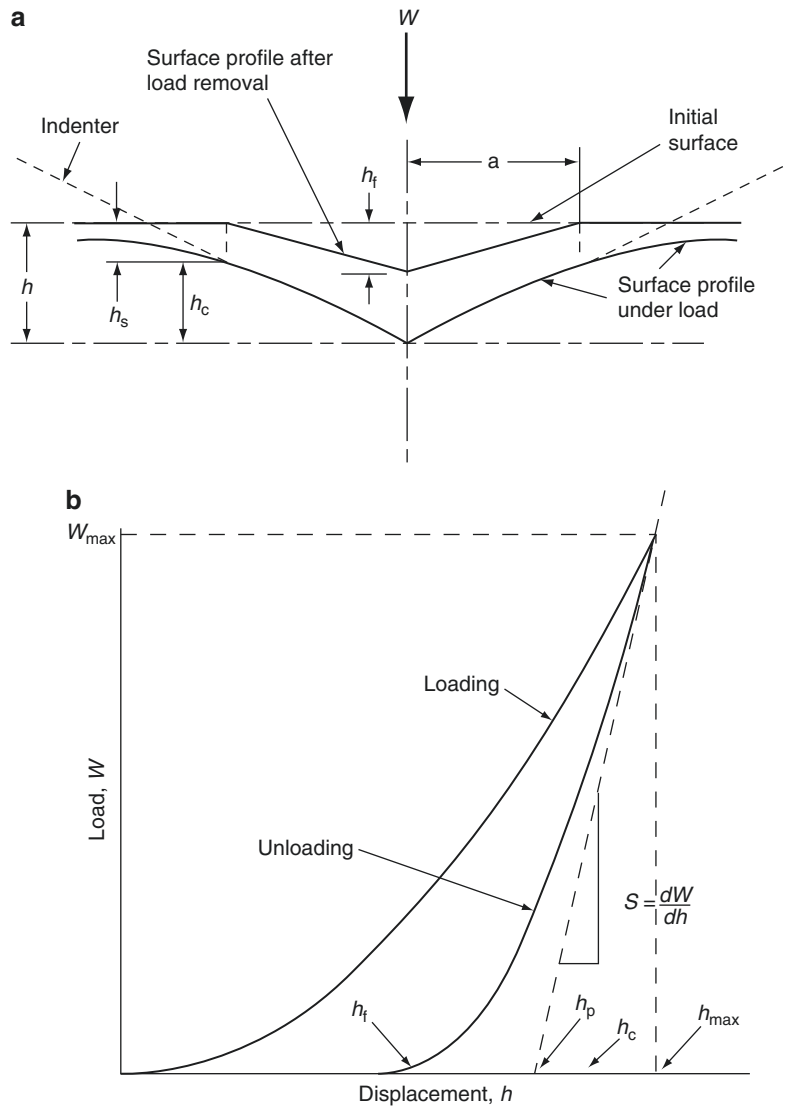
$$h_c = h_{max} - \varepsilon W_{max} = S_{max} \quad (1)$$

where $\varepsilon = 0.72$ for the conical indenter, $\varepsilon = 0.75$ for the paraboloid of revolution, and $\varepsilon = 1$ for the flat punch; and S_{max} is the stiffness (=1/compliance) equal to the slope of unloading curve (dW/dh) at the maximum load. Oliver and Pharr [10] assumed that behavior of the Berkovich indenter is similar to that of the conical indenter, since cross-sectional areas of both types of indenters varies as the square of the contact depth and their geometries are singular at the tip. Therefore, for Berkovich indenter, $\varepsilon \sim 0.72$. Thus, h_c is slightly larger than plastic indentation depth (h_p), given by

$$h_p = h_{max} - W_{max} = S_{max} \quad (2)$$

For a Vickers indenter with ideal pyramidal geometry (ideally sharp tip), projected contact area-to-depth relationship is given as [6, 11]

N



Nanoindentation, Fig. 3 (**a**) Schematic representation of the indentation process illustrating the depression of the sample around the indentation and the decrease in indentation depth upon unloading [10], and (**b**) schematic of load–displacement curve

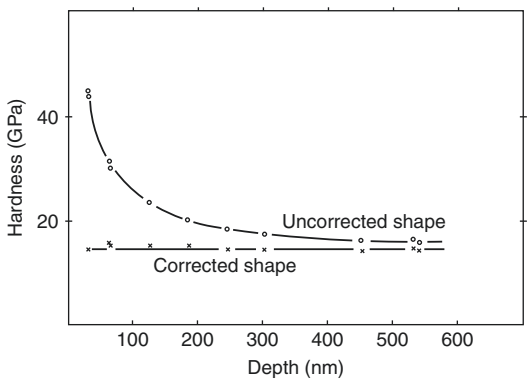
$$A = 24.5h_c^2 \tag{3}$$

Since the area-to-depth relationship is equivalent for both typical Berkovich and Vickers pyramids, Eq. 3 holds for the ideal Berkovich indenter as well. However, the indenter tip is generally rounded so that ideal geometry is not maintained near the tip. The actual indentation depth, h_c, produces a larger contact area than would be expected for an indenter with an ideal shape. For the real indenter used in actual measurements, the nominal shape is characterized by an area function $F(h_c)$, which relates projected contact area of the indenter to the contact depth

$$A^{1/2} = F(h_c) \tag{4}$$

The functional form must be established experimentally prior to the analysis.

Hardness Berkovich hardness *HB* (or H_B) is defined as the load divided by the projected contact area of the indentation. It is the mean pressure that a material will support under load. From the indentation curve, the hardness at maximum load can be obtained as



Nanoindentation, Fig. 4 Hardness as a function of indentation depth for polished single-crystal silicon (111) calculated from the area function with and without tip shape calibration [11]

$$HB = W_{\max} = A \tag{5}$$

where $W_{\max}$ is the maximum indentation load and A is the projected contact area at the peak load. The contact area at the peak load is determined by the geometry of the indenter and the corresponding contact depth h_c using Eq. 1 and 4. A plot of hardness as a function of indentation depth for polished single-crystal silicon (111), with and without tip shape calibration, is shown in Fig. 4. It should be noted that for this example, tip shape calibration is necessary and the hardness is independent of the corrected depth.

It should be pointed out that hardness measured using this definition may be different from that obtained from the more conventional definition in which the area is determined by direct measurement of the size of the residual hardness impression. The reason for the difference is that, in some materials, a small portion of the contact area under load is not plastically deformed, and, as a result, the contact area measured by observation of the residual hardness impression may be less than that at peak load. However, for most materials, measurements using two techniques give similar results.

Stresses during the indentation process can induce phase transformation in some ceramic materials, such as yttria-stabilized ZrO_2. Stress-induced-phase transformation can cause a change in volume of indented material that will in turn affect the load–displacement data. A relationship between indentation pressure and phase transformation and resulting change in volume of indental material during the indentation would need to be studied.

Modulus of Elasticity Even though during loading a sample undergoes elastic–plastic deformation, the initial unloading is an elastic event. Therefore, the Young's modulus of elasticity or, simply, the elastic modulus of the specimen can be inferred from the initial slope of the unloading curve (dW/dh) called stiffness (1/compliance) (at the maximum load). It should be noted that the contact stiffness is measured only at the maximum load, and no restrictions are placed on the unloading data being linear during any portion of the unloading.

It has been found that the compliance for the indenter is approximately *independent of the shape* (with a variation of at most 3 %) if the projected area is fixed [6]. The relationship for the compliance C (inverse of stiffness S) for an (Vickers, Knoop, and Berkovich) indenter is given as

$$C = \frac{1}{S} = \frac{dh}{dW} \approx \frac{1}{2E_r}\left(\frac{\pi}{A}\right)^{1/2} \tag{6}$$

where $\frac{1}{E_r} = \frac{1-\nu_s^2}{E_s} + \frac{1-\nu_i^2}{E_i}$, dW/dh is the slope of the unloading curve at the maximum load, E_r, E_s, and E_i are the reduced modulus and elastic moduli of the specimen and the indenter, and ν_s and ν_i are the Poisson's ratios of the specimen and indenter. C (or S) is the experimentally measured compliance (or stiffness) at the maximum load during unloading, and A is the projected contact area at the maximum load.

The contact depth h_c is related to the projected area of the indentation A for a real indenter by Eq. 4. A plot of the measured compliance (dh/dW) the reciprocal of the corrected indentation depth obtained from various indentation curves (one data point at maximum load for each curve) should yield a straight line with slope proportional to $1/E_r$ [11]. E_s can then be calculated, provided Poisson's ratio with great precision is known to obtain a good value of the modulus. For a diamond indenter, $E_i = 1{,}140$ GPa and $n_i = 0.07$ are taken.

The most frequently used method to measure initial unloading stiffness (S) is by Doerner and Nix [11], who measured S by fitting a straight line to the upper about one third portion of the unloading curve. The problem with this is that for nonlinear loading data, the measured stiffness depends on how much of the data is used in the fit. Oliver and Pharr [10] proposed a new procedure. They found that the entire unloading data are well-described by a simple power law relation:

$$W = B(h - h_f)^m, \tag{7}$$

where the constants B and m are determined by a least-square fit. The initial unloading slope is then found analytically, differentiating this expression and evaluating the derivative at the maximum load and maximum depth.

This analysis is based on an elastic solution that only accounts for sink-in (the indented material around the indenter below its original surface). However, in the more realistic case of elastic–plastic contact, sink-in or pile-up (the indented material around the indenter above its original surface) can occur depending on the specific mechanical properties of the material. For pile-up situations, the just-described method would underestimate the true contact area by as much as 50 %. This in turn leads to overestimation of the hardness and elastic modulus. Based on some modeling, pile-up is significant only when $h_f/h_{max} > 0.7$ and the material does not appreciably work harden. Note that h_f/h_{max} equal to zero corresponds to fully elastic deformation and a value of 1 corresponds to rigid-plastic behavior. Compressive residual stresses result in pile-up whereas tensile stresses result in sink-in. Although some correction procedures have been proposed [12], the real contact area measurement requires imaging of indentation impressions.

Determination of Load Frame Compliance and Indenter Area Function The measured displacements are the sum of the indentation depths in the specimen and the displacements of suspending springs and the displacements associated with the measuring instruments, referred to as load frame compliance. Therefore, to determine accurately the specimen depth, load frame compliance must be known. This is especially important for large indentations made with high modulus for which the load frame displacement can be a significant fraction of the total displacement. The exact shape of the diamond indenter tip needs to be measured because hardness and elastic modulus depend on the contact areas derived from measured depths. The tip gets blunt and its shape significantly affects the prediction of mechanical properties (Fig. 4).

Oliver and Pharr [10] proposed a method for determining area functions. Their method is based only on one assumption that Young's modulus is independent of indentation depth. They modeled the load frame and the specimen as two springs in series; thus,

$$C = C_s + C_f \tag{8}$$

where C, C_s, and C_f are the total measured compliance, specimen compliance, and load frame compliance, respectively. Combining Eqs. 6 and 8 yields

$$C = C_f + \frac{1}{2E_r}\left(\frac{\pi}{A}\right)^{1/2} \tag{9}$$

From Eq. 9, it should be noted that if the modulus of elasticity is constant, a plot of C as a function of $A^{1/2}$ is linear and the vertical intercept gives C_f. Accurate values of C_f are obtained when the specimen compliance is small, i.e., for large indentations.

To determine the area function and the load frame compliance, Oliver and Pharr made relatively large indentations in aluminum because of its low hardness. In addition, for the larger aluminum indentations (typically 700–4,000 nm deep), the area function for a perfect Berkovich indenter (Eq. 3) can be used to provide a first estimate of the contact.

Using the measured C_f value, they calculated contact areas for indentations made at shallow depths on the aluminum with measured E_r and/or on a harder fused silica surface with published values of E_r, by rewriting Eq. 9 as

$$A = \frac{\pi}{4}\frac{1}{E_r^2}\frac{1}{(C - C_f)^2} \tag{10}$$

from which an initial guess at the area function was made by fitting A as a function h_c data to an eighth-order polynomial

$$A = 24.5h_c^2 + C_1h_c + C_2h_c^{1/2} + C_3h_c^{1/4} + \ldots + C_8h_c^{1/128} \tag{11}$$

where C_1 through C_8 are constants. The first term describes the perfect shape of the indenter; the others describe deviations from the Berkovich geometry due to blunting of the tip.

Key Research Findings

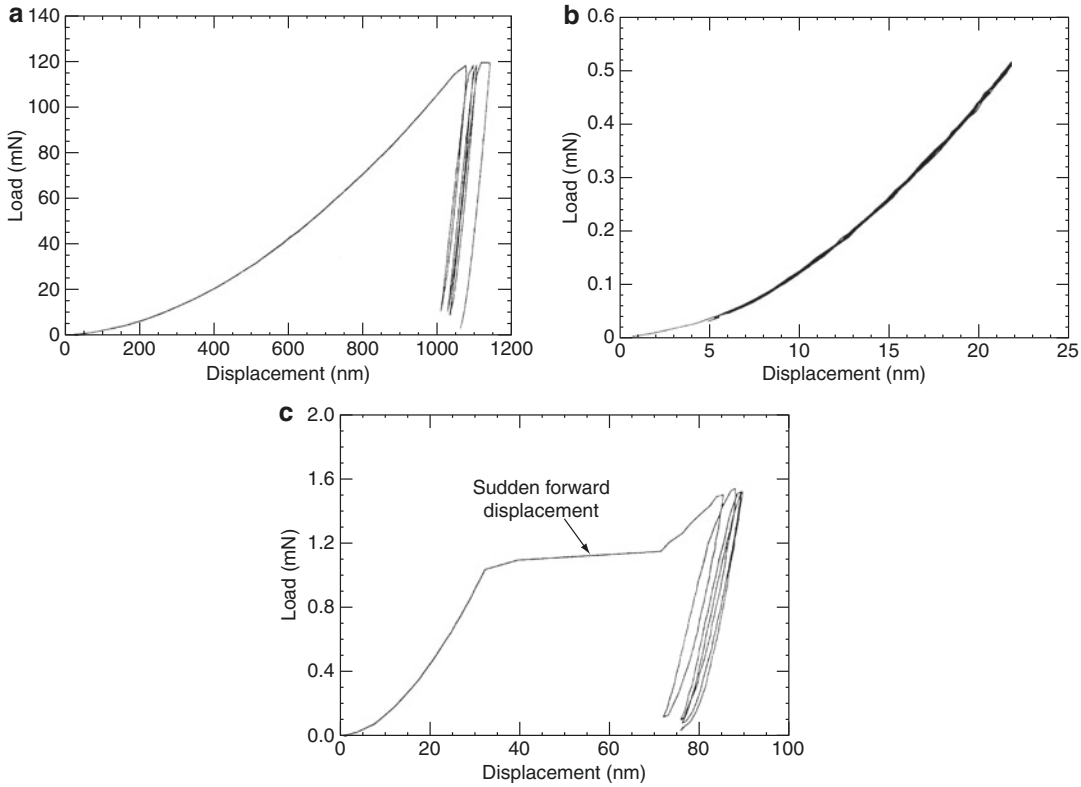
Examples of Measured Mechanical Properties

To illustrate the usage of nanoindentation techniques, typical data obtained on various bulk materials, coatings, and microstructures is presented below.

Load–Displacement Curves A variety of mechanical phenomena, such as the transition from elastic to plastic deformation, creep deformation, formation of subsurface cracks, and crystallographic phase transition, can be studied by the load–displacement curves obtained at different loading conditions [6]. The load-time sequence used by Oliver and Pharr [10] for study of various materials included three loading–unloading cycles, hold for 100 s at 10 % of the peak load, reload, hold for 100 s, and unload. Load–displacement curves for electropolished single-crystal tungsten are shown in Fig. 5. The tungsten data are typical of materials in which the hardness is relatively small compared to the modulus, as is observed in most metals; most of the indenter displacement in these metals is accommodated plastically and only a small portion is recovered on unloading. Figure 5 shows that the peak load displacements shift to higher values in successive loading/unloading cycles. In addition, the relatively large displacement just prior to final unloading is due to creep during the 100-s hold period at peak load. Indentation at a very low load of 0.5 mN caused only elastic displacements (Fig. 5b). At higher peak loads, the indentation is not just elastic (Fig. 5c). When a threshold load of about 1 mN is reached, a sudden jump in displacement corresponding to the onset of plasticity is observed, and a permanent hardness impression is formed. Tungsten at low loads (Fig. 5c) exhibit distinct hysteresis loops, as might be expected if there were a small amount of reverse plasticity upon loading. However, the looping degenerates with cycling after three or four cycles, the load–displacement behavior is largely elastic.

Another set of load–displacement curves is shown in Fig. 6 for glass fibers, which are used as reinforcements in numerous materials including plastic, rubber, cement as well as heat insulation materials [13]. The indents were made on polished cross sections of the fibers. Fibers C, Q, and S are all 16 μm diameter fibers, each with different formulations. The load–displacement plots of indentations made at 0.1 mN peak loads on all fibers exhibit several displacement discontinuities of pop-in marks during the first and second loading steps of the indentation, as denoted by the arrows in Fig. 6. Pop-in marks in the loading curves result from sudden penetrations of the tip into the sample. Fiber S exhibits the smallest pop-in marks, followed by fibers Q and C, which suggests that Fiber S has the highest resistance to plastic deformation as well as crack formation and crack propagation.

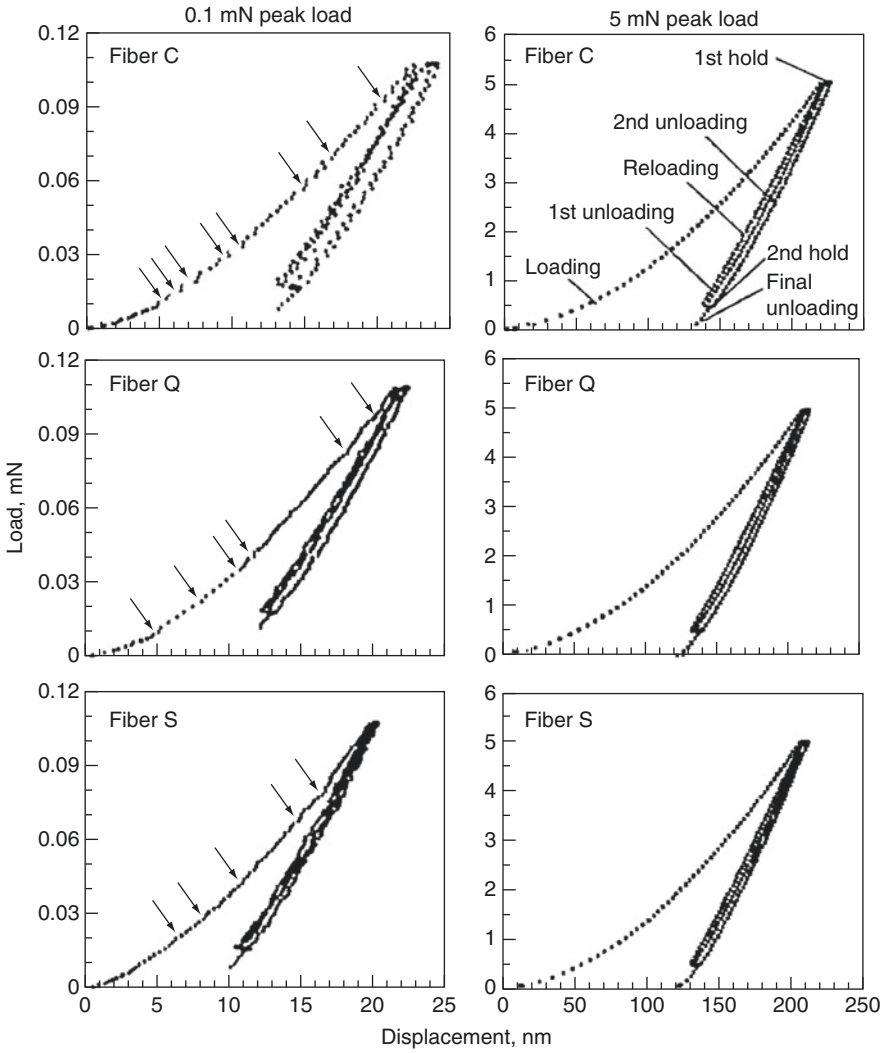
Hardness and Elastic Modulus Measurements The nanoindenter is commonly used to measure surface mechanical properties of bulk materials [6, 10]. Hardness and elastic moduli for six bulk materials and three materials used as a substrate for the construction of magnetic rigid disks and single-crystal silicon are presented in Fig. 7. The data show that there is a very little indentation size effect in several materials on the hardness values. In the case of aluminum and tungsten in Fig. 7, there is a modest increase in hardness at low loads, which could be due to surface-localized cold work resulting from polishing [10]. The modulus data also show that there is very little evidence for an indentation size effect; i.e., the moduli remain more or less constant over the entire range of load.



Nanoindentation, Fig. 5 Load versus displacement plot for an electro-polished single-crystal tungsten (**a**) at a peak load of 120 mN, (**b**) at a peak load of 0.5 mN (elastic contact), and (**c**) at a peak load of 1.5 mN showing the yield point [10]

The nanoindenter is ideal for the measurement of mechanical properties of thin films and composite structures. A number of investigators have reported hardness of composite structures with thin films on a substrate. It is widely accepted that to measure true hardness of the films, the indentation depth should not exceed 10 % of the film thickness [14]. However, based on a finite element analysis of the indentation of thin films of various thicknesses it was found that the true hardness of the films could be obtained if the indentation depth does not exceed about 30 % of the film thickness. At higher indentation depths, the composite hardness changes with the indentation depth [6].

The significance of the composite hardness and modulus on multilayered thin films is illustrated for the components of the digital micromirror device (DMD). The DMD is the main component of commercial DLP^{TM} projectors, invented at Texas Instruments (TI) in 1987. The DMD chip is composed of an array of micromirrors, along with a yoke and hinge array underneath. The contact between the tips at the end of the yoke and the underlying landing site is required for operation. Since the DMD chip is composed of various thin film structures, the mechanical properties of the component materials are important to the DMD performance. Figure 8 shows the hardness and elastic modulus as a function of contact depth for the DMD materials measured using the continuous stiffness method, with data for Si (1 0 0) included for reference. The samples belong to three categories: bulk material, single-layer, and multilayered materials. For bulk material, Si (1 0 0), the hardness and elastic modulus does not change with contact depth, as expected. For a hard film on soft substrate, TiN/Si, the hardness and

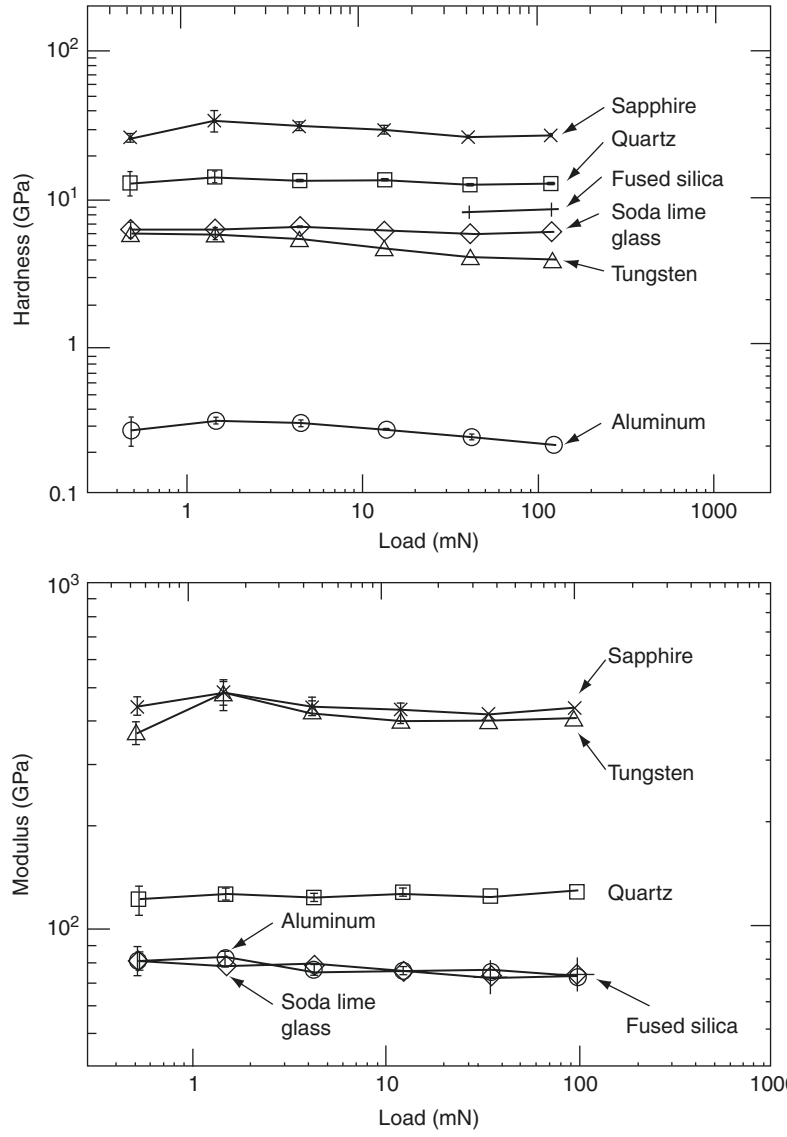


Nanoindentation, Fig. 6 Representative load versus displacement plots of indentations made at 0.1 and 5 mN peak indentation loads on various glass fibers [13]

elastic modulus values decrease with contact depth. For soft films on hard substrates, SiO_2/Si and Al alloy/Si, their hardness and elastic modulus values increase with contact depth. The hardness and elastic modulus of TiN/Si and SiO_2/Si gradually approach the hardness and elastic modulus values of Si (1 0 0) after the contact depths exceed 100 nm, due to the substrate effect [15].

For multilayered materials, TiN/Al alloy/Si and SiO_2/TiN/Al alloy/Si, there are two common observations. For both samples, the lowest hardness values were obtained near the TiN/Al alloy interfaces, which are approached at the contact depth of about 100 and 200 nm, respectively. It is observed that both samples show much lower hardness values (~5 GPa) at the contact depth of about 30 nm as compared to the corresponding single-layer materials. The substrate effect contributions are expected to be small at this contact depth.

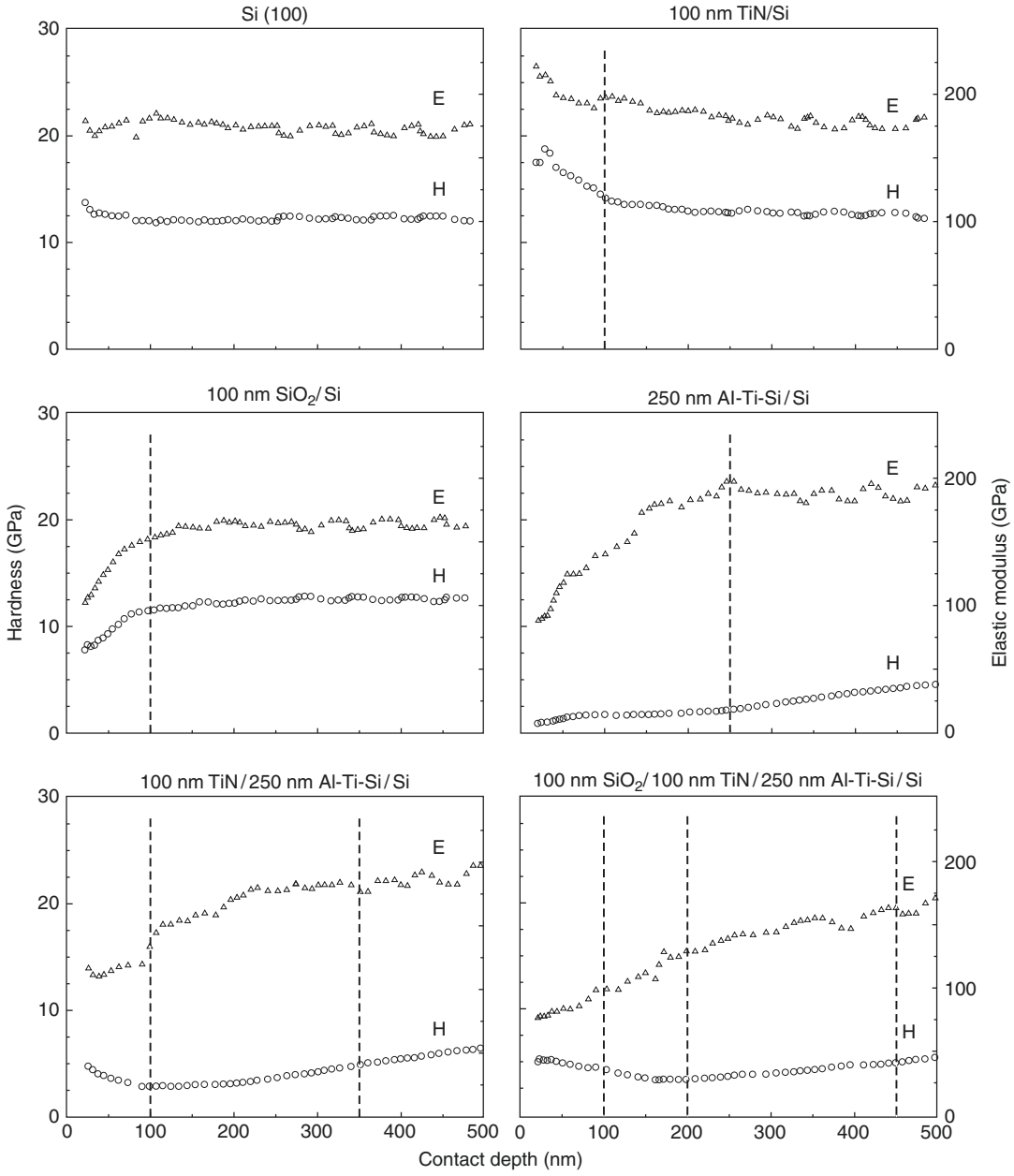
It is known that for a metallic thin film, the measured hardness is sensitive to grain size, film texture, film thickness, level of adhesion of the film to the substrate, and presence of residual



Nanoindentation, Fig. 7 Hardness and elastic modulus as a function of load for six bulk materials – mechanically – polished single-crystal aluminum, electro-polished single-crystal tungsten, soda-lime glass, fused silica, (001) single-crystal sapphire, and (001) single-crystal quartz [10]

stress. When TiN was reactively sputtered onto Al alloy, the stress state and microstructure (grain size, texture) of TiN might be changed as compared to TiN/Si, since the film growth substrate is changed. According to the Hall–Petch relationship, in metal and alloy films, if the grain size of the TiN/Al alloy/Si was greater than that of the TiN/Si, then the hardness would decrease. It is believed that films grown by physical and chemical vapor deposition techniques usually display a preferred orientation, which is dependent upon deposition process variables. Hardness reductions by as much as two orders have been observed in cubic coatings (e.g., TiC, TiN, ZrC) as a function of the preferred orientation plane [e.g., (1 1 1), (1 0 0) and (1 1 0)]. For SiO_2/TiN/Al alloy/Si, since SiO_2 is usually amorphous when deposited as films, the grain size and texture may not be the parameters responsible for the hardness decrease of SiO_2/TiN/Al alloy/Si [15].

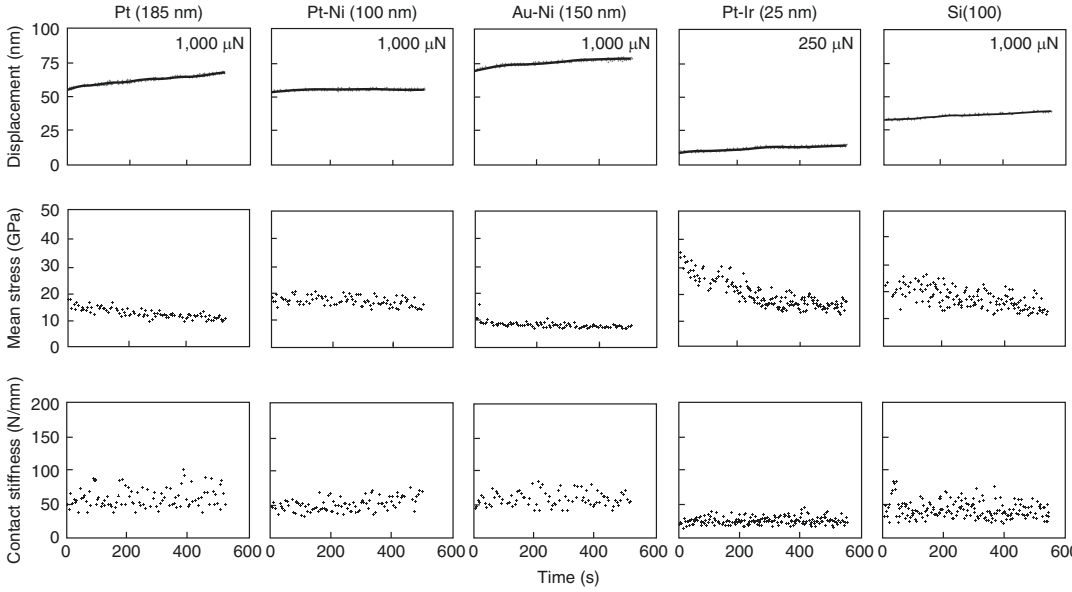
Creep Resistance Most materials, including ceramics (and even diamond) are found to creep at temperatures well below half their melting points, even at room temperature. Indentation



Nanoindentation, Fig. 8 Hardness and elastic modulus as a function of contact depth for the thin films used in the digital micromirror device (*DMD*) [15]

creep and indentation load relaxation (ILR) tests are commonly used to measure the time-dependent flow of materials. These techniques offer an advantage of being able to probe the deformation properties of a thin film as a function of indentation depth and location [6].

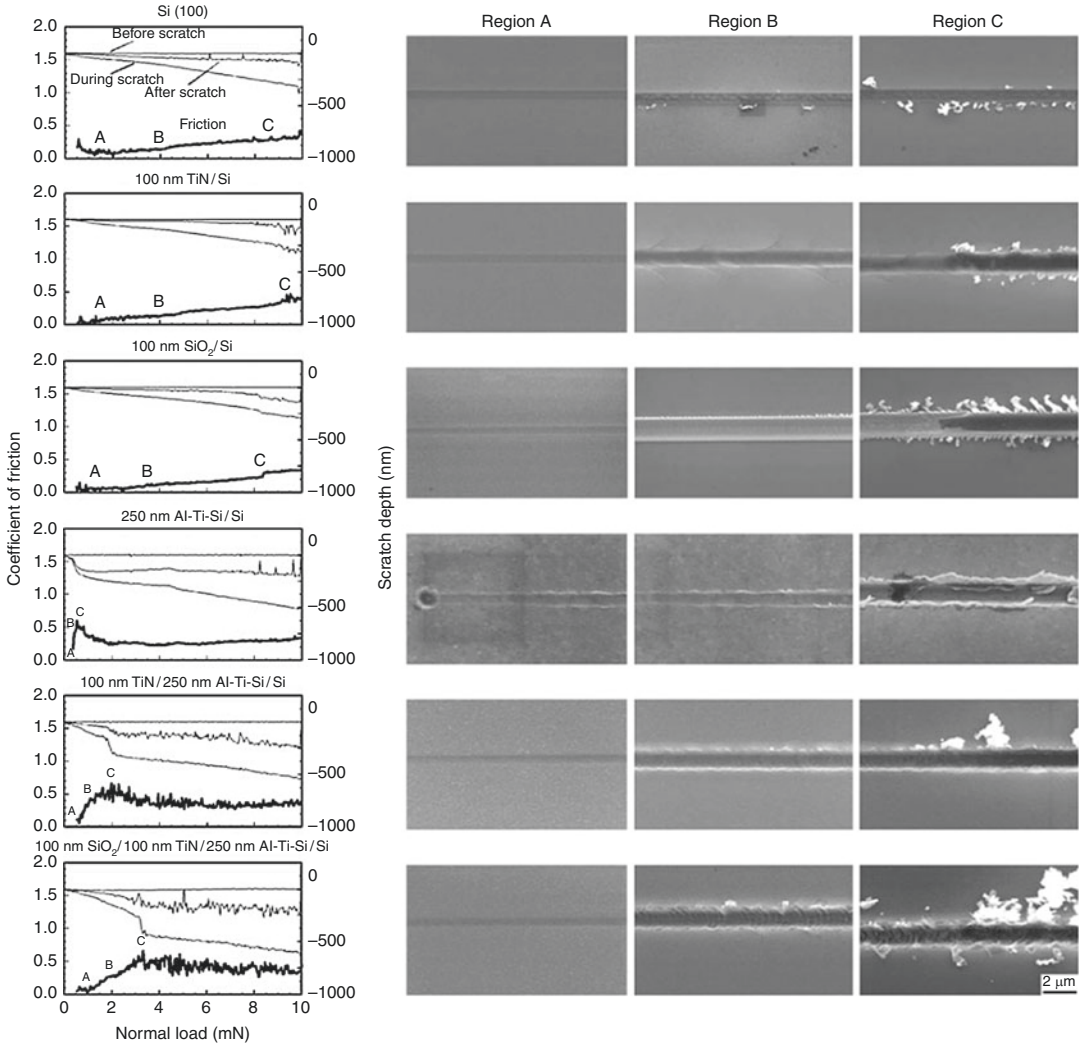
An example of nanoscale creep is shown in Fig. 9 for the noble metal thin films used for probe-based ferroelectric data recording. In this technology, a conductive atomic force microscope (AFM) probe with a noble metal coating is placed in contact with lead zirconate titanate (PZT) film,



Nanoindentation, Fig. 9 Creep displacement, mean stress and contact stiffness as a function of time of the various metal-coated AFM probes used in ferroelectric data recording and Si(100) as reference [16]

which serves as the ferroelectric material. The long-term device performance will be compromised if the metal films exhibit excessive creep [16]. Figure 9 shows creep data for the various metal-coated probes and Si, where the change in displacement, mean stress, and contact stiffness was monitored while holding the tip at the maximum imposed load (P_{max}). A lower P_{max} was applied on Pt–Ir as it is a very thin coating; the selected load of 250 mN ensures that the tip is measuring the creep resistance of the coating and did not penetrate the substrate during the creep experiment. From the displacement plots, it is observed that the three alloyed coatings (Pt–Ni, Pt–Ir, and Au–Ni) exhibited less creep compared to Pt. The Si surface also shows significant creep, which is consistent with previous reports. This creep on silicon is a combination of two factors: initial creep can occur as the tip penetrates the native oxide surface, which is then followed by a dislocation glide plasticity mechanism [16, 17].

Microscratch Resistance of Various Materials Using Micro/Nanoscratch Technique The microscratch technique is commonly used for screening of bulk materials for wear resistance. As stated earlier, normal load applied to the scratch tip is gradually increased during scratching until the material is damaged. Friction force can be measured during the scratch test [7]. After the scratch test, the morphology of the scratch region including debris is observed in an SEM. Based on the combination of the changes in the friction force as a function of normal load and SEM observations, the critical load is determined and the deformation mode is identified. Any damage to the material surface as a result of scratching at a critical ramp-up load results in an abrupt or gradual increase in friction. The material may deform either by plastic deformation or fracture. Ductile materials (all metals) deform primarily by plastic deformation, resulting in significant plowing during scratching. Tracks are produced whose width and depth increase with an increase in the normal load. Plowing results in a continuous increase in the coefficient of friction with an increase in the normal load during scratching. Debris is generally ribbon-like or curly, whereas brittle materials deform primarily by brittle fracture with some plastic deformation. In the brittle fracture mode, the coefficient of friction increases very little until a critical load is reached at which the material fails catastrophically and produces



Nanoindentation, Fig. 10 Coefficient of friction and scratch depth profiles as a function of normal load for DMD thin films and SEM images of three regions over scratches: at the beginning of the scratch (indicated by "A" on the friction profile), at the point of damage initiation (indicated by "B" on the friction profile), and at the point of film delamination (indicated by "C" on the friction profile). For Si substrate, region C is where severe damage began) [15]

fine debris, which is rounded, and the coefficient of friction increases rapidly above the critical load.

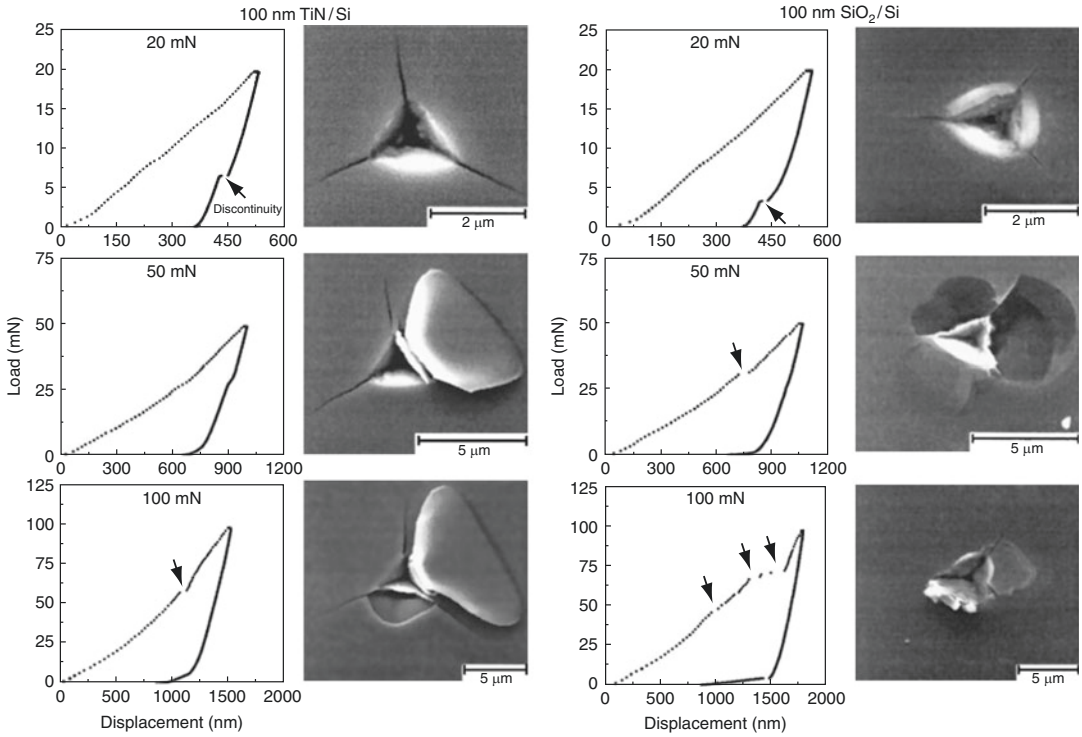
In the example for the digital micromirror device (DMD) mentioned above, the repeated contact between the yoke and landing site components eventually leads to wear over time. The wear resistance of these materials can be evaluated through a nanoscratch experiment. For the components of the DMD, Fig. 10 shows the coefficient of friction and scratch depth profiles as a function of increasing normal load, along with SEM images of three regions over scratches [15]. SEM images were taken at the beginning of the scratch (indicated by "A" on the friction profile), at the point of initiation of damage at which the coefficient of friction increases by a factor of two (indicated by "B" on the friction profile), and at the point of the initiation of the delamination of the films (for Si substrate, this region represents

the position where severe damage began) (indicated by "C" on the friction profile). The normal load corresponding to region B is defined as the "critical load." Before the critical load (from regions A to B), the coefficient of friction of the Si (1 0 0), TiN and SiO_2 films increased at a slower rate, and at the critical load (region B), these three samples exhibit a small increase in the coefficient of friction. However, for the remaining samples, i.e., Al alloy, TiN/Al alloy, and SiO_2/TiN/Al alloy, the coefficient of friction increased sharply from regions A to B. The slopes of the increase of the coefficient of friction in these three samples from regions A to B are inversely proportional to the hardness of the samples. For instance, the Al alloy shows the lowest hardness, but highest slope of the increase of the coefficient of friction. The increase of the coefficient of friction during scratching was attributed to the increasing plowing of the sample by the tip with increasing normal load, and the softer material led to deeper plowing of the tip and faster increase of the coefficient of friction. At region C, all films were delaminated and experienced an abrupt increase in the coefficient of friction, which was associated with catastrophic failure as well as significant plowing of the tip into the sample.

The SEM images show that all the samples experienced plastic deformation at the beginning of the scratch (region A). For Al alloy, because it is very soft, region A is very close to region B. The indent of the conical tip, the plastic deformation associated with the pile-up was observed at the beginning of the scratch for Al alloy. At region C, for Si (1 0 0) substrate, only some debris particles were found without any cracks on the side of the scratch, which is responsible for the small increase in the coefficient of friction. For TiN, SiO_2 and Al alloy films, it is clear that at region C, the scratch tip reached the Si substrate and delaminated the film. The irregular cracks found on the side of the scratch in TiN indicate that TiN is brittle. However, it is interesting to observe that the SiO_2 film shows high ductility according to the SEM images at region C, where regular film pieces were found to attach to the sample surface even after the film was delaminated. As expected, severe plowing and high pile-up were found for Al alloy. For TiN/Al alloy and SiO_2/TiN/Al alloy, a lot of debris were found at region C, indicating that more materials were torn away from the films. By investigating the corresponding scratch depths, it can be seen that at region C, the in situ scratch depth of TiN/Al alloy and SiO_2/TiN/Al alloy are about 350 and 450 nm, respectively, indicating that for both multilayer films, the tip reached the Si substrates. This may explain why more material was torn away from the multilayered films at region C. For the thin film materials used in DMD, the damage initiated at the "critical load" should be prevented in order to make the DMD operate properly. Therefore, the film topography and mechanical property changes due to scratch in the region from A to B, especially the film failure mechanism at region B, are of great interest for DMD [15].

Nanofracture Toughness Nanoindentation can be used to determine fracture toughness [6]. The indentation cracking method is especially useful for measurement of fracture toughness of thin films or small volumes. For fracture toughness measurement of ultrathin films ranging from 100 nm to few micrometers, traditional indentation or four-point flexure techniques cannot be used. Because of shallow indentation depths required in the indentation technique, it is difficult to measure a radial crack length under even SEM. Li et al. [18] developed a novel technique based on nanoindentation in which through-thickness cracking in the coating is detected from a discontinuity in the load–displacement curve and energy released during cracking is obtained from the curve. Based on the energy released, fracture mechanics analysis is used to calculate fracture toughness. A cube corner is preferred because the through-thickness cracking of hard films can be accomplished at lower loads.

An example of the fracture toughness determination for the thin films used in DMD is shown in Fig. 11, which contains load–displacement curves of indentations made at 20, 50, and 100 mN peak indentation loads together with the scanning electron microscopy (SEM) micrographs of indentations on the TiN/Si and SiO_2/Si. Steps are found in the loading curves obtained at 100 mN for TiN/Si, and at 50 mN (and 100 mN) for SiO_2 as shown by



Nanoindentation, Fig. 11 Load–displacement curves of indentations made at 20, 50, and 100 mN peak loads and corresponding SEM micrographs of the indents on two of the films used in the DMD. *Solid arrows* indicate discontinuity during the unloading portion of the curve and the *hollow arrows* indicates a step during the loading portion of the load–displacement curve [19]

N

arrows in Fig. 11. The steps in the loading curves of indentations result from the ring-like through-thickness cracking as shown in the SEM images. These ring-like through-thickness cracking were used to calculate the fracture toughness together with the strain energy release obtained from the steps in the loading curves, which will be described later. At 20 mN, the radial cracks were found in both TiN/Si and SiO_2/Si. The discontinuity in the displacement observed during unloading is believed to be due to the formation of lateral cracks at the base of the median crack which results in the surface of the specimen being thrust upward [19].

Li et al. [18] suggest that the steps in the loading curve on the films result from the film cracking, and that the fracture process progresses in three stages: (1) first ring-like through-thickness cracks form around the indenter by high stresses in the contact area; (2) delamination and buckling occur around the contact area at the film/substrate interface by high lateral pressure; and (3) second ring-like through-thickness cracks and spalling are generated by high bending stresses at the edges of the buckled film. The fracture toughness was calculated from the second ring-like through-thickness cracking generated in the third stage. In this stage, the stress concentration at the end of the interfacial crack cannot be relaxed by the propagation of the interfacial crack. With an increase in indentation depth, the height of the bulged film increases. When the height reaches a critical value, the bending stresses caused by the bulged film around the indenter will result in the second ring-like through-thickness crack formation and spalling at the edge of the buckled film, which leads to a step in the loading curve. The step in the loading curve is totally from the film cracking and not from the interfacial cracking or the substrate cracking [19].

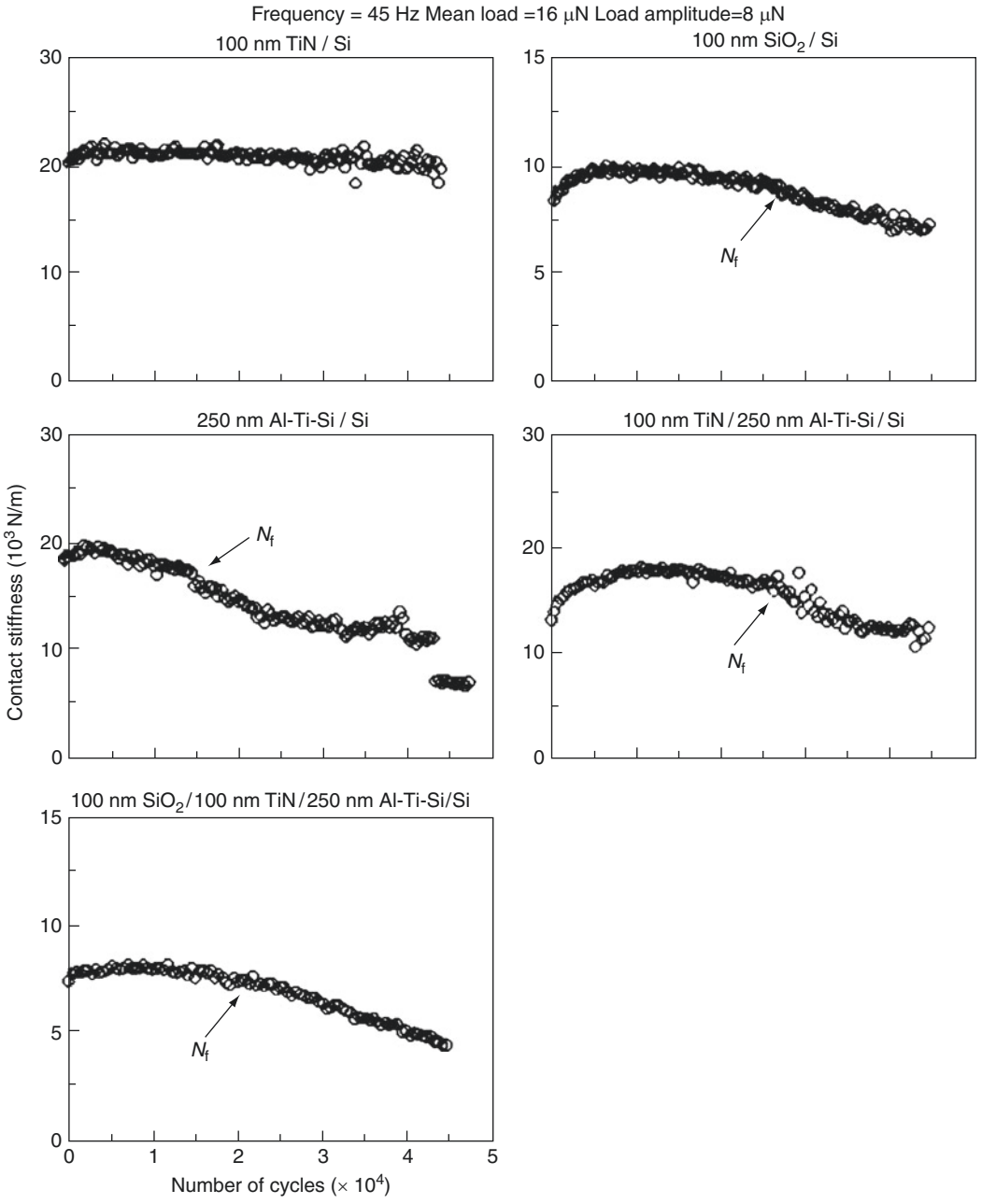
Nanofatigue Fatigue fracturing progresses through a material via changes within the material at the tip of a crack, where there is high stress intensity. There are several situations: cyclic fatigue, stress corrosion, and static fatigue. In this discussion, the focus is on cyclic fatigue, which results from the cyclic loading of machine components; e.g., the stresses cycle from tension and compression occurs in a loaded rotating shaft.

To compare the fatigue lives of the thin film structures used in the DMD, the contact stiffness data as a function of the number of cycles for TiN/Si, SiO_2/Si, Al alloy/Si, TiN/Al alloy/Si, and SiO_2/TiN/Al alloy/Si cyclically deformed at an oscillating load amplitude of 8 mN with a mean load of 16 mN at a frequency of 45 Hz are shown in Fig. 12. For all the samples, the contact stiffness increased to varying degrees at the beginning of the fatigue test as a result of strain hardening. TiN/Si and Al alloy/Si films have the longest and shortest N_f, respectively. By comparing the N_f of multilayered thin film structures with the corresponding single top layer films ∼ TiN and SiO_2), it can be seen that the multilayered thin film structures exhibited shorter N_f. Ranking of the nanoscale fatigue data is similar to results from hardness studies, i.e., the hardness of multilayered thin film structures is lower than that of the corresponding single top layer films [19].

The relationship between fatigue life and hardness is as follows. At the beginning of the fatigue test, the tip penetrates into the sample a certain displacement, depending on the hardness of the sample. Then the tip is held and a cyclic load is applied. As the number of cycles increases, the tip penetrates deeper. It is known that during the indentation fatigue test, even though there may be no cracks, the tip of the sharp indenter can sink deeper into the material and extend the plastic zone. At this stage, the compressive stress and the deformation volume may not be high enough to cause cracks. In this case, the contact stiffness would not decrease abruptly. It is known that voids and impurities act as original cracks. As the number of cycles reaches a critical value N_f, the original cracks beneath and around the tip would propagate due to the accumulated deformation, as well as new cracks that may develop at high deformation volume. This would cause the contact stiffness to drop suddenly. Since TiN/Si is the hardest sample, it resisted the tip to sink and consequently results in lower accumulated deformation volume. So it took a longer time for TiN/Si to fail due to the fatigue-induced crack. For softer materials, SiO_2/Si and Al alloy/Si, the tip sank easily and caused larger deformation volume and earlier fatigue failure. This explanation can also be used to understand the data that multilayered thin film structures with soft interlayer showed shorter fatigue life than the corresponding top hard films. The hardness of multilayered structures is lower than the top hard films responsible for shorter fatigue life [19].

Deformation of Microbeams Microscale structures, such as beams, are common in MEMS/NEMS devices. In this example, the focus is on polymer microbeams designed for the BioMEMS device known as a cell force sensor. In the work by Palacio et al. [20], the materials considered were poly(propyl methacrylate), poly(methyl methacrylate), poly(styrene), and a poly(styrene)-nanoclay composite (abbreviated as PPMA, PMMA, PS and PS/clay, respectively). For this application, nanoindentation can be used for conducting bending experiments in the normal and lateral directions. These experiments provide information on the elastic modulus, yield strength, and breaking strength of these materials. This example illustrates how nanoindentation can be used beyond the more traditional experiments carried out on bulk materials and thin films.

Bending experiments were performed by Palacio et al. [20] with applied loads in the millinewton range in order to investigate the behavior of polymer beams beyond the elastic deformation regime. Figure 13a shows normal load–displacement profiles at maximum imposed loads of 0.5, 1.5 and 3.0 mN applied on three different beams in one wafer sample. For PPMA, PMMA and PS, the loading portion overlaps, indicating uniform beam quality. Slight variability in the loading profile is observed in the PS/clay composite, which is attributed to variable interfacial adhesion between the filler and the matrix and inhomogeneities in the dispersion of the clay particles.

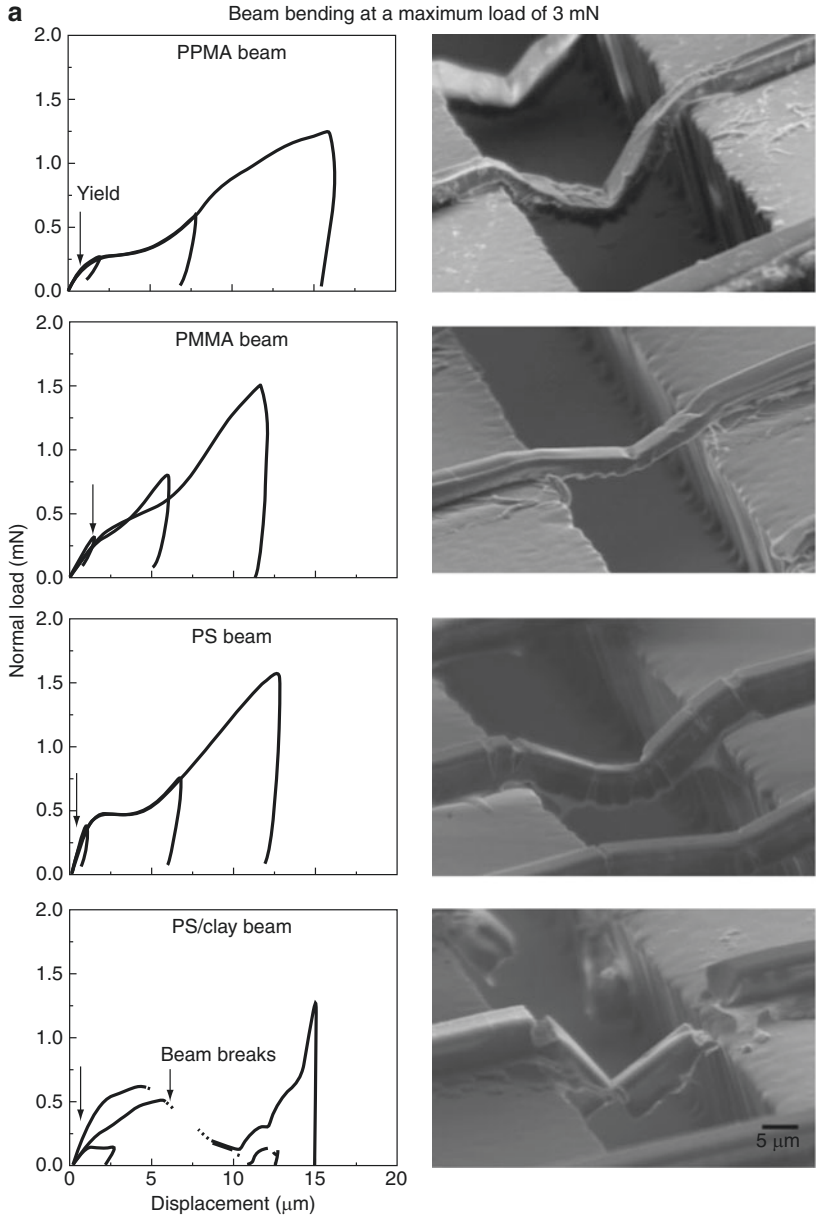


Nanoindentation, Fig. 12 Contact stiffness as a function of the number of cycles for various thin films used in the DMD [19]

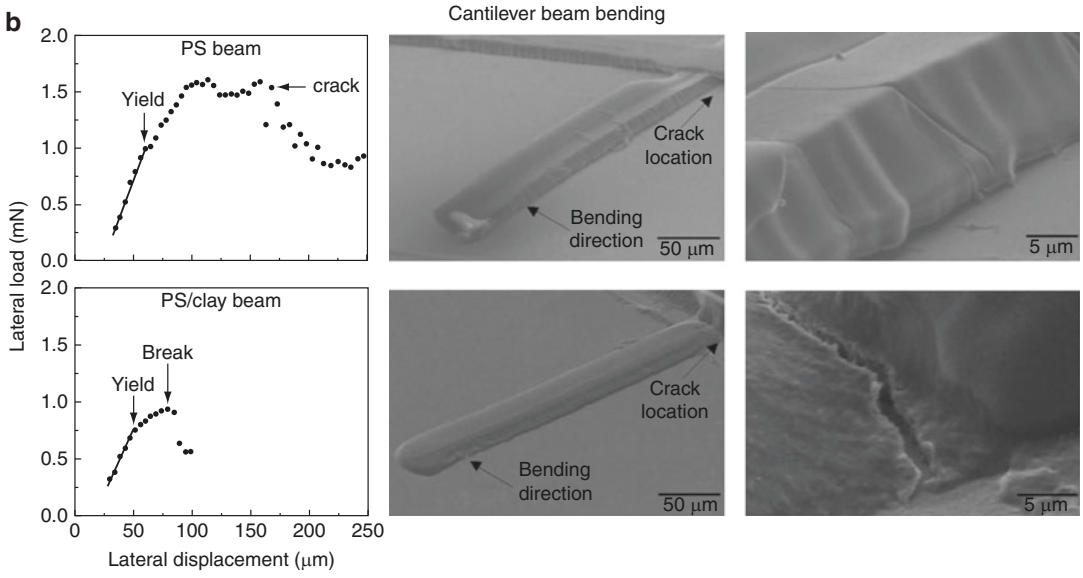
The first inflection in the load–displacement profiles implies the onset of yield and its location is indicated by arrows. The three unfilled polymers (PPMA, PMMA and PS) exhibit ductility, as seen in the SEM images for the beams tested with an applied load of 3 mN. The beams stretch at the ends and underneath the center of the beam. These correspond to regions subjected to maximum tensile stress. The beams deformed symmetrically, indicating that the load was applied equidistant

from the clamping points. The PS/clay nanocomposite beam exhibited slight yielding, followed by breaking, as indicated by an arrow in the load–displacement profile. After breaking, the load continued to increase because a predefined load is imposed by the nanoindenter. The corresponding SEM image for PS/clay shows a flat fracture surface, indicating that the crack propagated perpendicular to the direction of the applied stress [20].

From the load–displacement curves, the yield strength can be evaluated by using [20, 21]:



Nanoindentation, Fig. 13 (continued)



Nanoindentation, Fig. 13 (**a**) Load–displacement curves for the normal beam bending of the PPMA, PMMA, PS and PS/clay polymer beams for use in the cell force BioMEMS sensor device. SEM images correspond to the beams subjected to 3 mN load. (**b**) Lateral bending of PS and PS/clay cantilever beams. The lateral force is plotted as a function of the lateral displacement, and the *arrows* indicate the onset of yield for both PS and PS/clay, cracking for PS and breaking for PS/clay. SEM images indicate the bending direction and the crack locations [20]

$$\sigma = \frac{3}{4}\frac{Fl}{bh^2} \tag{12}$$

where F is the load corresponding to yield of the beam (as indicated by arrows in the load–displacement profile), l, b, and h are the beam's length, width and height, respectively. It is assumed that the beam ends are clamped.

Figure 13b presents the results of the lateral bending experiments on PS and PS/clay cantilever beams. On the left column are the plots for the lateral force recorded as a function of the lateral displacement. From the linear elastic regime (as indicated by the straight line drawn over the data points), the modulus can be evaluated. For a cantilever beam with one end clamped, it is expressed as [20, 21]:

$$E = \frac{l^3}{3I}m \tag{13}$$

where l is the beam length, I is the area moment of inertia for the beam cross section, and m is the slope of the linear region of the force–displacement curve.

From the load–displacement curves, the yield strength can be evaluated by using [20, 21]:

$$\sigma = \frac{6Fl}{bh^2} \tag{14}$$

where F, l, b, and h are defined similarly as in Eq. 12. The force used for calculating the yield strength is taken at the point where a change in slope was observed (as indicated by the arrows on the left in Fig. 13b).

For PS, a steady decrease in the lateral load was observed at 160 mm, while for the PS/clay beam, the drop at 80 mm was abrupt (as indicated by the arrows on the right). SEM images reveal that for PS, this corresponds to crack formation at the base of the beam that did not run throughout the entire thickness of the beam. However for PS/clay, the beam broke, which implies that the addition of the nanoclay filler induces embrittlement of the beam.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Friction Force Microscopy

References

1. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
2. Bhushan, B.: Modern Tribology Handbook, vol. 1 and 2. CRC Press, Boca Raton (2001)
3. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
4. Bhushan, B.: Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
5. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
6. Bhushan, B., Li, X.: Nanomechanical characterization of solid surfaces and thin films. Int. Mater. Rev. **48**, 125–164 (2003)
7. Bhushan, B., Gupta, B.K., Azarian, M.H.: Nanoindentation, microscratch, friction and wear studies of coatings for contact recording applications. Wear **181–183**, 743–758 (1995)
8. Gerberich, W.W., Gao, H., Sundgren, J.-E., Baker, S.P. (eds.): Thin Films: Stresses and Mechanical Properties VI. Symposium proceedings, materials research society, vol. 436. Materials Research Society, Pittsburgh (1996)
9. Johnson, K.L.: Contact Mechanics. Cambridge University Press, Cambridge (1985)
10. Oliver, W.C., Pharr, G.M.: An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments. J. Mater. Res. **7**, 1564–1583 (1992)
11. Doerner, M.F., Nix, W.D.: A method for interpreting the data from depth-sensing indentation instruments. J. Mater. Res. **1**, 601–609 (1986)
12. Tsui, T.Y., Pharr, G.M.: Substrate effects on nanoindentation mechanical property measurement of soft films on hard substrates. J. Mater. Res. **14**, 292–301 (1999)
13. Li, X., Bhushan, B., McGinnis, P.: Nanoscale mechanical characterization of glass fibers. Mater. Lett. **29**, 215–220 (1996)
14. Tabor, D.: The Hardness of Metals. Clarendon, Oxford (1951)
15. Wei, G., Bhushan, B., Jacobs, S.J.: Nanomechanical characterization of multilayered thin film structures for digital micromirror devices. Ultramicroscopy **100**, 375–389 (2004)
16. Palacio, M., Bhushan, B.: Nanotribological and nanomechanical properties of lubricated PZT thin films for ferroelectric data storage applications. J. Vac. Sci. Technol. A **26**, 768–776 (2008)
17. Bhushan, B., Kulkarni, A., Bonin, W., Wyrobek, J.T.: Nanoindentation and picoindentation measurements using a capacitive transducer system in atomic force microscopy. Philos. Mag. A **74**, 1117–1128 (1996)
18. Li, X., Diao, D., Bhushan, B.: Fracture mechanisms of thin amorphous carbon films in nanoindentation. Acta Mater. **45**, 4453–4461 (1997)
19. Wei, G., Bhushan, B., Jacobs, S.J.: Nanoscale fatigue and fracture toughness measurements of multilayered thin film structures for digital micromirror devices. J. Vac. Sci. Technol. A **26**, 768–776 (2004)
20. Palacio, M., Ferrell, N., Hansford, D., Bhushan, B.: Nanomechanical characterization of polymer beam structures for BioMEMS applications. Sens. Actuator A **135**, 637–650 (2007)
21. Young, W.C., Budynas, R.G.: Roark's Formulas for Stress and Strain. McGraw-Hill, New York (2002)

Nanointerconnection

- ▶ Nanorobotic Spot Welding

Nanojoining

- ▶ Nanorobotic Spot Welding

Nanolayered Coatings

- ▶ Implementing Alternating Nanolaminates in Trenched Energy Storage Systems

Nanolayers

- ▶ Implementing Alternating Nanolaminates in Trenched Energy Storage Systems

Nanomachining

- ▶ Ultraprecision Machining (UPM)

Nanomagnetism

► Magnetic Nanostructures and Spintronics

Nanomanipulation for Cell Analysis

► Nanomanipulation of Biocells

Nanomanipulation of Biocells

Yajing Shen[1] and Toshio Fukuda[2]
[1]Department of Mechanical and Biomedical Engineering, College of Science and Engineering, City University of Hong Kong, Kowloon, Hong Kong, China
[2]School of Mechatronic Engineering, Beijing Institute of Technology, Beijing, China

Synonyms

Biologicalcell manipulation; Nanomanipulation for cell analysis

Definition

Nanomanipulation for biocells is the usage of precise manipulation techniques to tackle the fundamental and practical problems in biological cells at small scale, including, but not limited to, cell positioning, cell characterization, cell assembly, and so on.

Overview

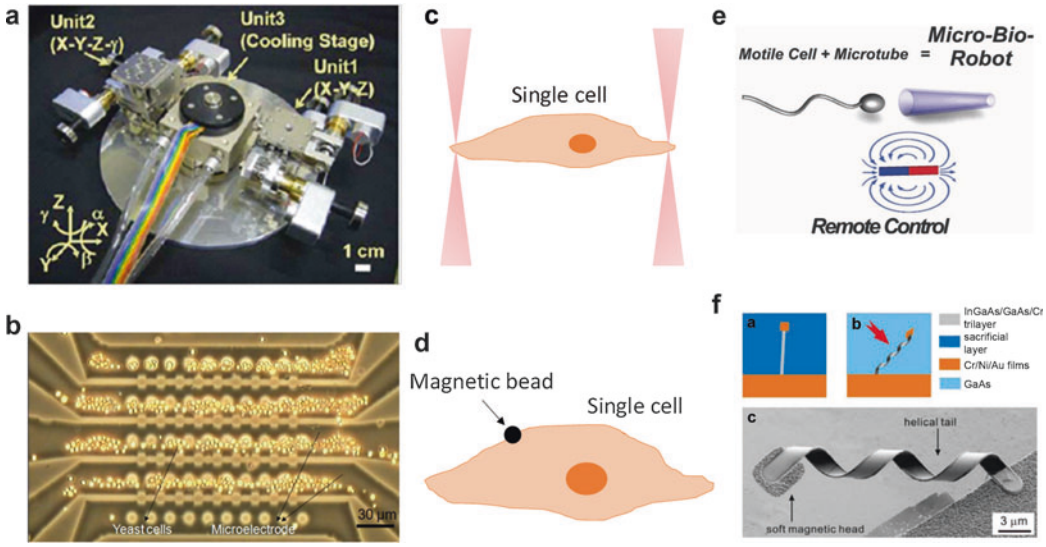
Biological cell is the basic unit of life. This small organizational system maintains a highly complex and hierarchical architecture of interconnected molecular networks. The analysis on biological cell greatly benefits the basic research in life science and also clinical applications. Nanomanipulation enables to control the cell's position precisely; thereby, it provides the possibility to study the cell's properties at very small scale. Nowadays, nanomanipulation has been successfully used in cell positioning, cell characterization, cell assembly, and other biological cell analysis fields.

Key Techniques in Nanomanipulation

Nanomanipulator

A main type of nanomanipulator is based on the piezo effect, which is able to convert the electrical power input to the mechanical movement output. Piezo actuator can provide long movement distance up to several centimeters and high positioning resolution up to a few nanometers. Because of these advantages, this type of actuator has been widely used as the basic units to develop the nanomanipulation system. Figure 1a illustrates an example of this kind of nanomanipulation system. It has seven degrees of freedom in total and can provide positioning resolution at nanometer scale, which has been successfully used to study the single cell's property [1]. One potential drawback for this type of nanomanipulator is the existence of wires, which makes it not applicable for the cell manipulation in narrow closed environments, such as inside a microfluidic chip.

Wireless manipulators, including dielectrophoresis (DEP), optical tweezers, magnetic twisting cytometry, and others, avoid the usage of electrical wires (Fig. 1b–d). Thus, they are able to manipulate the sample remotely without physical contact, which provides the possibility for the biological cell analysis in a closed environment. Nowadays, these manipulation systems have been widely used for the cell analysis in the microchip, including for cell array, cell characterization, and so on [2]. One potential problem of the wireless manipulation is that they have special requirements on the manipulation environment, such as light transmission solution, magnetic objects, and so on. Another problem is force-producing ability of this kind of system is relative low, i.e., usually less than micro-Newton. Thus, the application area of the wireless manipulators is not as wide as the piezo actuator.



Nanomanipulation of Biocells, Fig. 1 Examples of the manipulators at small scale. (**a**) Piezo-based nanomanipulator [1]. (**b**) Dielectrophoresis [5]. (**c**) Optical tweezers. (**d**) Magnetic twisting cytometry. (**e**) Sperm-flagella-driven micro-bio-robot [3]. (**f**) Artificial bacterial flagella [4] (Reproduced with permission from Refs. [1, 3–5])

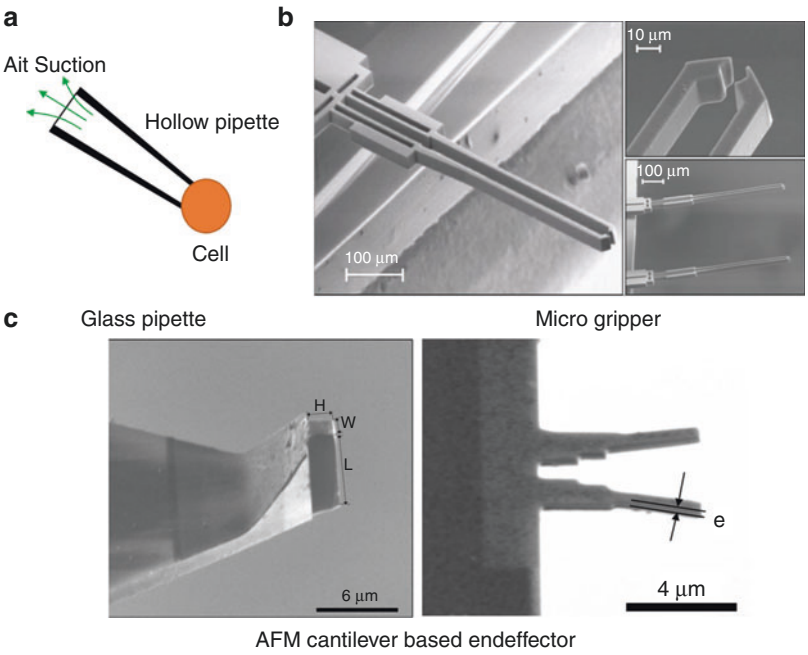
In addition to the above methods, the bio-inspired manipulator also has fast-rising growth these days. Creatures or the actuation part of creatures have been taken as the manipulator directly, such as sperm, bacteria, flagella, protein, DNA, and so on (Fig. 1e) [3]. These bionanomanipulators have many advantages, such as tiny size, simple structure, self-assembly, and biocompatibility. Inspired by the actuation mechanism of bacteria, an artificial flagella is developed (Fig. 1f) [4]. Although the research in this field is still at the preliminary stage, it is thought that the biochemical-inspired nanomanipulator would play more and more important roles in the future.

End Effector

End effector is the "hand" to manipulate the biological cells. The nanomanipulation system, especially for the manipulator with wire, is hard to manipulate the biological cell directly. Thus, the design of the end effector regulates the working efficiency of the nanomanipulation system directly. Generally, the end effector should be of a small tip to fit the size of the cell. The simple endeffectors include tungsten probe, glass needle, micropipette, and so on (Fig. 2a). To enhance the manipulation ability, many functional end effectors are also developed based on micro–nano-fabricated techniques, such as MEMS (Micro-electromechanical Systems)-based and AFM (atomic force microscopy) cantilever-based end effectors, which are illustrated in Fig. 2b, c respectively.

MEMS device is one type of successfully applicable end effector for the manipulation of cells at small scale. It is usually designed based on the electrostatic mechanism. Controlling by the input voltage, the different parts of the MEMS device can move relatively. Therefore, the MEMS device enables to do the open and close motion if well designed with two probes, such as microgripper [6]. In addition, some unique materials, such as shape memory alloy and thermal expansion polymer, are also used to improve the movement ability of the MEMS device.

AFM cantilever is another type of widely used end effector for cell manipulation. AFM cantilever can be integrated with the nanomanipulation system easily owing to its small size. More important, AFM cantilever can provide force information during the manipulation process based on the



Nanomanipulation of Biocells, Fig. 2 Examples for the end effectors. (**a**) Micropipette. (**b**) Microgripper [6]. (**c**) AFM cantilever–based end effector [7, 8] (Reproduced with permission from Refs. [6–8])

deflection of the cantilever beam. Regarding these advantages, AFM cantilever provides a great opportunity for the cell characterization, including the mechanical, electrical, and other properties. To improve the function of AFM cantilever further, various shapes of end effectors are developed based on AFM cantilever. These end effectors not only keep the force measurement ability but also have specific function according to the design. For instance, by modifying the tip of AFM cantilever via focused ion beam (FIB) etching, different functional end effectors are developed to study the properties of single cell, such as the nanopicker for cell-cell adhesion study, dual probe for cell electricity study, microprobe for cell adhesion strength study, and so on [7–9].

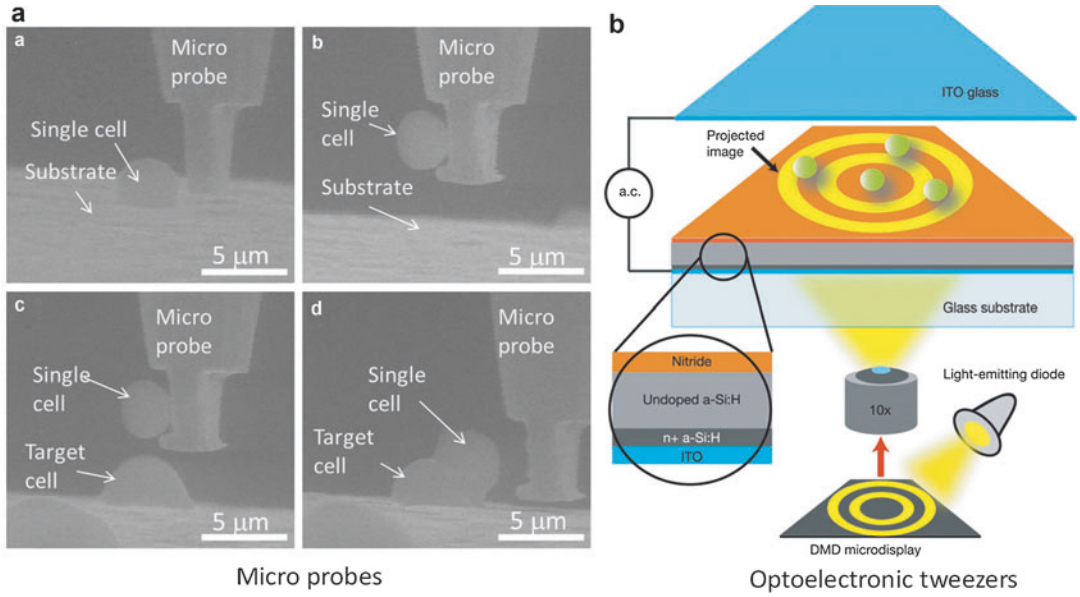
Examples of Applications

Cell Positioning

Cell positioning is the basic technique for cell array and single-cell analysis, in which nanomanipulation techniques provide a great support. The techniques for cell positioning can be classified into direct and indirect cell positioning according to the physical contact and noncontact manipulation, as illustrated in Fig. 3.

In the direct cell positioning, the end effector, such as micropipette and microgripper, are assembled on the nanomanipulator firstly. Driven by the nanomanipulator, the position of the end effector is able to be precisely controlled. Since the end effector is able to handle and release the cell, this integrated system enables the pick-transport-place of a specific biological cell precisely and effectively. The issue for the direct manipulation is that the end effector has to contact with the cell during the manipulation process. Thus, it is difficult to implement this platform in the closed environment, like inside the microfluidic channel.

Indirect cell positioning method is based on the wireless actuation techniques, which can manipulate the sample remotely without physical contact. Optical tweezers are instruments that use a highly focused laser beam to hold and manipulate small objects. In the late 1980s, Arthur and Joseph demonstrated the first application of optical tweezers to the biological sciences, using it to trap an individual tobacco mosaic virus and Escherichia coli bacterium [10]. Today, optical tweezer has been a

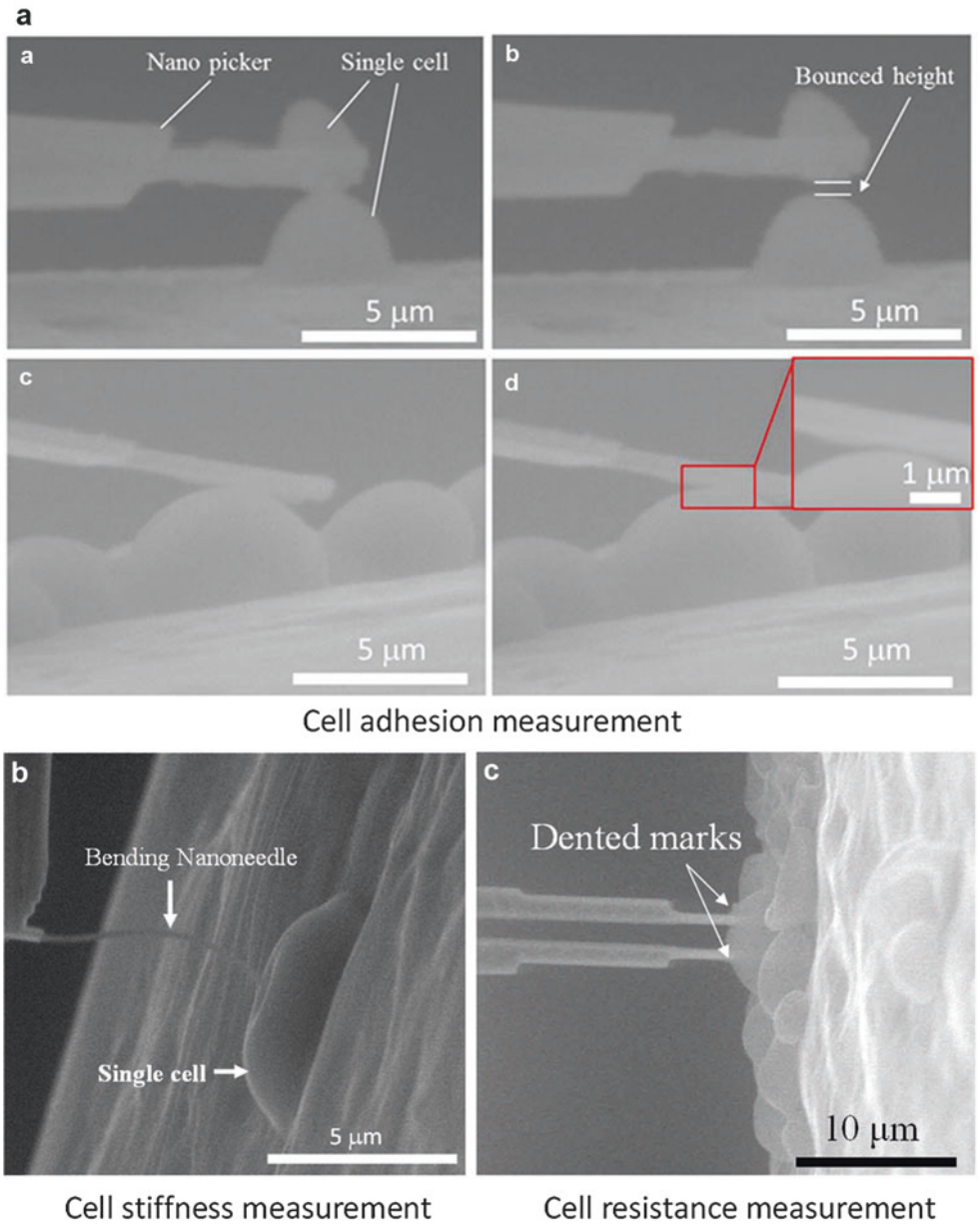


Nanomanipulation of Biocells, Fig. 3 Application examples of the nanomanipulation for cell positioning. (**a**) Manipulation with physical contact: cell transfer by microprobe [8]. (**b**) Manipulation without physical contact: cell array by optoelectronic tweezers [11] (Reproduced with permission from Refs. [8, 11])

common scientific instrument in trapping cells to study a variety of biological problems. Because optical tweezers can manipulate the object precisely without physical contact, it is very useful for the cell positioning inside the microchip. The main issue of the optical tweezers is that the generated force by laser is usually limited to nano-Newton, which is not large enough to drive the nonsuspending cells some time. Dielectrophoresis (DEP) is the phenomenon in which a force is exerted on a dielectric particle when it is subjected to a nonuniform electric field. DEP has the advantages of noncontact manipulation, good controllability, and high integration ability with the microchip. Therefore, it is another powerful technique for the cell trapping and sorting, especially for the cell analysis in the microfluidic chip. The drawback of DEP is that the strength of the force depends strongly on the medium and particles' electrical properties, on the particles' shape and size, as well as on the frequency of the electric field. Another problem is that it is difficult to make the size of the electrode to nanoscale, which affects the manipulation accuracy. To overcome these drawbacks, a new concept called optoelectronic tweezers (OETs) is proposed [11]. OETs use projected optical images to grab and corral tiny particles with sizes ranging from hundreds of micrometers to tens of nanometers. In this method, light first creates "virtual electrodes" on the substrate. Then, the image in conjunction with an externally applied electrical bias creates the localized DEP traps in the illuminated areas. This method combines the advantages of optical tweezers and electrode-based DEP, which is expected to have a bright future in the cell positioning.

Cell Characterization

Cell characterization provides a quantitative method to study cell behavior, which greatly benefits the basic research and the clinical application. In the characterization process, the cell has to be handled and manipulated flexibly to allow the measurement of different properties, such as strength, stiffness, electrical conductivities, and so on. As the key technique to provide precise positioning, nanomanipulation involves deeply in cell characterization. Figure 4 demonstrates three application examples of cell characterization, i.e., cell adhesion, stiffness, and electrical resistance.



Nanomanipulation of Biocells, Fig. 4 Application examples of the nanomanipulation for cell characterization. (**a**) Adhesion study [8]. (**b**) Stiffness characterization [21]. (**c**) Electrical properties study [9] (Reproduced with permission from Refs. [8, 9, 21])

Cell adhesion is a very important research topic in biological field, since many cell activities depend on the attachment of cells to the extracellular matrix or neighboring cells. One traditional nanomanipulation technique for cell adhesion study is micropipette aspiration [12]. In this method, two aspiration pipette end effectors are assembled on two micromanipulators, respectively. The negative pressure from the pipette allows the cell to be grasped on the tip.

Driven by the nanomanipulator, the two pipettes pull the cells from opposite direction to separate them. With the increasing of the suction force gradually, the cell would be separated finally. In this process, the cell-cell adhesion strength is able to be calculated based on the suction pressure. AFM system is another powerful nanomanipulation system to study the cell's adhesion properties, including both the cell-substrate and the cell-cell adhesion [13, 14]. Here, AFM cantilever is not only the end effector to manipulate the cell but also the sensor to measure the adhesion force. One common setup is to use the AFM cantilever to detach the cell on the sample stage, in which the adhesion strength is measured based on the deflection of the AFM cantilever. If one cell is immobilized on a modified AFM cantilever tip, the adhesion cell-cell force can be measured, in which process the specific tip is used to press on and leave from another cell. The adhesion between these two cells can be calculated based on the deflection of the AFM cantilever. In addition, AFM nanomanipulation system also allows the adhesion strength measurement between the cell and the proteins. In this process, the AFM cantilever tip is modified with protein first. Then, this cantilever is used to study the cell-protein adhesion following the same process of cell-cell adhesion measurement. AFM system has the advantages of higher positioning accuracy and higher force resolution. One main issue is that the AFM system cannot scan the sample to get high-resolution image during the adhesion characterization process. Environmental scanning electron microscope (ESEM) is one unique type of SEM. It can not only keep the real-time observation ability of SEM but also allow the nonconductive and aqueous sample observation. Integrating with the nanomanipulation technique, ESEM provides a novel approach for single-cell analysis [1, 15]. AFM cantilever–based end effector is usually used in this characterization system. By modifying the AFM cantilever with nanofabrication technique, various end effectors can be developed for different characterization tasks, such as the microputter and nanopicker for the cell-substrate and cell-cell adhesion measurement respectively [7, 8]. This system allows the real-time imaging at nanometer resolution during the manipulation, which provides more information to benefit the single-cell characterization.

Cell stiffness and elastic characterization is another application example of the nanomanipulation system. Recently, various nanomanipulation methods have been reported to study single cell stiffness, including MEMS tool, optical tweezers, magnetic beads, AFM system, ESEM nanotool, and so on. MEMS tool is an effective end effector for cell characterization owing to its force measurement ability. For example, an integrated manipulator and MEMS probe system is developed to measure the stiffness of the pollen tube cell [16]. This characterization system allows measuring the local region of the cell, but the probe size of the MEMS tool is difficult to be at nanoscale. Optical tweezers enable high positioning accuracy and high force resolution for the stiffness characterization. It has been successfully used to study the mechanics of the human red blood cell [17]. The relative low force output ability ($\sim$pN) and light-induced damage are thought as the challenges for optical tweezer manipulation. Magnetic manipulation drive small beads to deform the cell based on the controlling of the magnetic field, which avoids potential light-induced damage as in laser trapping. The local viscoelastic value of the cell can be measured based on the applied magnetic strength on the beads. In this method, the modeling and controlling of the nonhomogeneous magnetic field are considered as the challenge. The high force sensitivity of the AFM system makes it particularly suitable for cell stiffness study [18]. Here, the AFM cantilever is used as the nanoindentor to deform the cell. Then, the stiffness value is calculated based on the cell deformation and the applied force. One potential problem in this method is the pyramid structure of AFM tip, which causes difficulty to the mechanical modeling when the cell has a large deformation. To overcome this drawback, a nanoneedle with a long slim probe is designed based on FIB etching technique [19]. Driven by the nanomanipulator, the nanoneedle allows a large deformation on the cell surface, thereby to provide more accurate mechanical models for the cell.

Nowadays, the characterization of a cell's electrical property has attracted increasing interest, because it reflects the physiological state of the cell and has wide application in medical field. To study the electrical properties of the cell accurately, a precise nanomanipulation system and a tiny conductive probe are required. An integrated system, consisting of robot-assisted AFM nanomanipulation system and the traditional patch-clamp system, is able to measure the electrical response of the human cell corresponding to external mechanical stimulation [20]. This system expects to benefit the drug test in the future. More than the single probe, an end effector with dual probes is also developed to study the electrical resistance of the cell [9]. Driven by the nanomanipulation system, the dual probes penetrate the cell wall first. Then, the current between the two probes is measured by applying a low voltage. This experiment setup can detect the resistance difference of living and dead cells, which expects to have wide applications in cell biology study.

Cell Assembly

As the basic unit, a cell is always considered as the "block" of life. How these blocks build life is one of the biggest questions in the world. Therefore, increasing interests have been focused on the assembly of 3D biostructure from single cells.

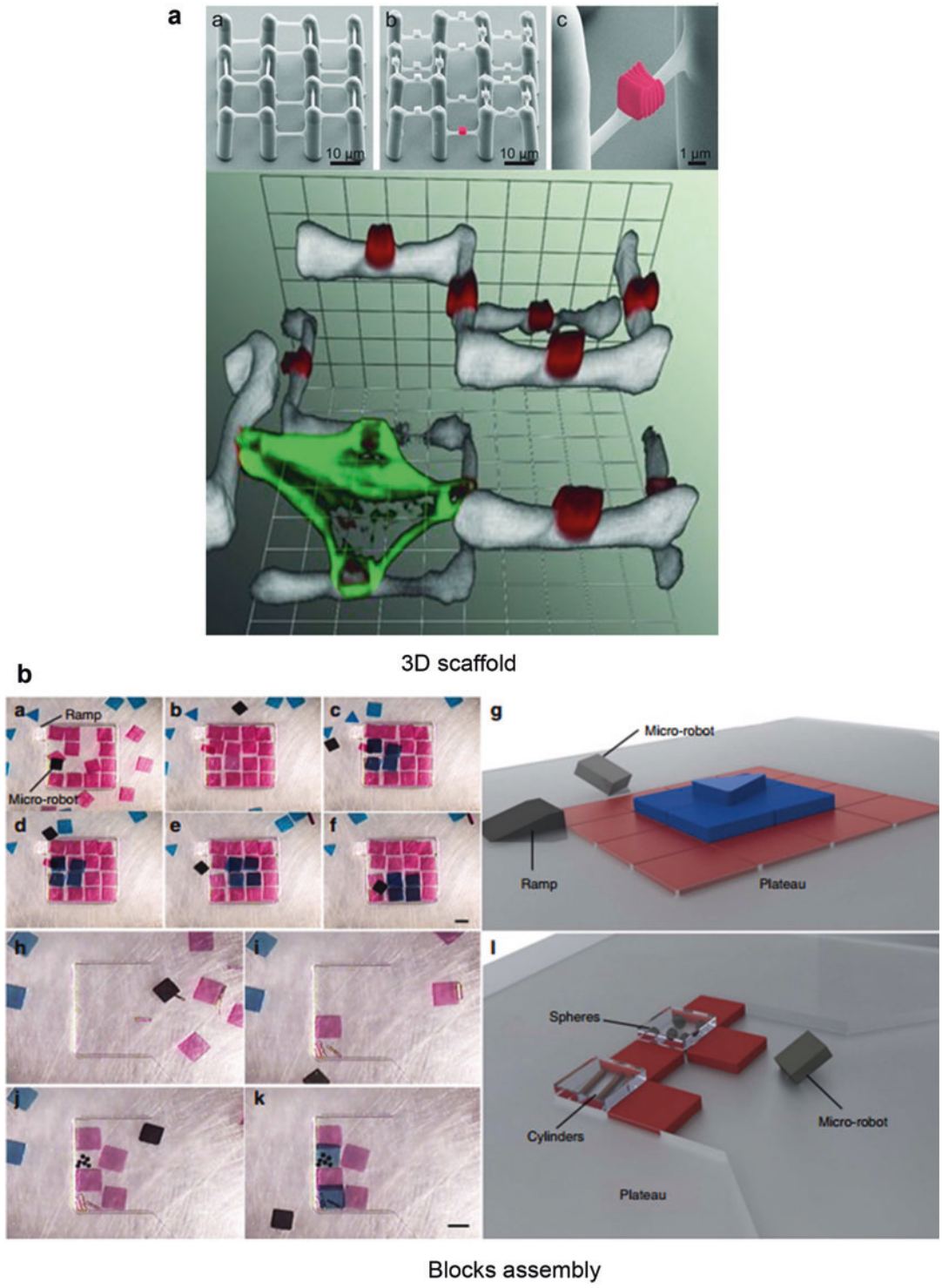
Generally, a cell requires a relative strong "scaffold" to support its growth in 3D space. Thus, the preparation of 3D scaffold is one key step in cell assembly. 3D printing is a novel technique that combines robotic manipulation and material science together. It is able to print the scaffold or components with single cells to 3D structure directly. With the support of nanomanipulation, 3D printing can also work at small scale. For instance, a scaffold with structure at nanoscale is fabricated based on nanoscribe technique, which is successfully used to study the growth of single cells in 3D space [22]. In addition to 3D printing, other fabrication techniques are also applicable for 3D cell assembly, such as the controllable optical induced method. Because some gels are sensitive to light, the formation of the gel can be controlled by light. On the basis of this principle, 3D gel has been generated successfully with different shape and height by controlling the light illumination area and the exposure time [23].

Another assembly method is to directly build the blocks with cells to 3D structure, as house building. First, the biological cells are capsulated in series gel blocks by biochemical process. Then, small robots actuated by magnetic fields are used to manipulate these blocks to build complex 3D structures [24]. To improve the manipulation ability, the biocompatible magnetic nanoparticles can also be mixed in the gel, so the position of the gel can be manipulated by magnetic fields more flexibly.

Nanomanipulation shows its high applicability in 3D cell assembly owing to its controllability and flexibility. One main goal of 3D cell assembly is to construct artificial functional units that can take place in our natural organ one day. Thus, there are still many challenges for the manipulation technique, such as the precise position control of the cell in 3D space, the assembly of different types of cells, and so on (Fig. 5).

Future Research

Nanomanipulation technique has greatly benefited the biological cell research from many aspects. However, there are still lots of challenges for the manipulation system. The first one is the actuation technique for the nanomanipulator, which covers power supply, energy transmission, driven approach, and so on. To address the problems in practice, the actuation system for the nanomanipulator should be much smaller, higher output capability, and higher efficiency. Biological creature is the "nanorobot" in nature, which has high working efficiency and exactly compatible for bioapplication. "Learn from nature" will drive new inspiration to the nanomanipulation, which is expected to benefit the biomedical field in the future. Secondly, the data with high spatial and time resolution would help us to understand the cell activities better. Nowadays, people have shown increasing interest to cell research at the subcellular or molecule level. Thus, the nanomanipulation system should be much more accurate and intelligent to meet the above



Nanomanipulation of Biocells, Fig. 5 Application examples of the nanomanipulation for cell assembly. (**a**) Fabrication of 3D scaffold for cell growth in space [22]. (**b**) Assembly of blocks containing cells [24] (Reproduced with permission from Refs. [22, 24])

requirement. Last, the commercialization of the nanomanipulation system should be considered. Although many exciting nanomanipulation techniques have been reported, most of them are still in the laboratory. How to bring these techniques to market to benefit our society will be a significant task in the future.

Cross-References

- ▶ 3D Micro-/Nanomanipulation with Force Spectroscopy
- ▶ Nanorobotic Assembly
- ▶ Nanorobotic Manipulation of Biological Cells
- ▶ Nanorobotic Spot Welding
- ▶ Nanorobotics
- ▶ Nanorobotics for Bioengineering
- ▶ Nanorobotics for NEMS Using Helical Nanostructures
- ▶ Robot-Based Automation on the Nanoscale

References

1. Shen, Y., Ahmad, M.R., Nakajima, M., Kojima, S., Homma, M., Fukuda, T.: Evaluation of the single yeast cell's adhesion to ITO substrates with various surface energies via ESEM nanorobotic manipulation system. IEEE Trans. NanoBiosci. **10**, 217–224 (2011)
2. Yi, C., Li, C.-W., Ji, S., Yang, M.: Microfluidics technology for manipulation and analysis of biological cells. Anal. Chim. Acta **560**, 1–23 (2006)
3. Magdanz, V., Sanchez, S., Schmidt, O.G.: Development of a sperm-flagella driven micro-bio-robot. Adv. Mater. **25**, 6581–6588 (2013)
4. Zhang, L., Abbott, J.J., Dong, L., Peyer, K.E., Kratochvil, B.E., Zhang, H., et al.: Characterizing the swimming properties of artificial bacterial flagella. Nano Lett. **9**, 3663–3667 (2009)
5. Yue, T., Nakajima, M., Tajima, H., Fukuda, T.: Fabrication of microstructures embedding controllable particles inside dielectrophoretic microfluidic devices. Int. J. Adv. Robot. Syst. **10**, 1–9 (2013)
6. Chronis, N., Lee, L.P.: Electrothermally activated SU-8 microgripper for single cell manipulation in solution. J. Microelectromech. Syst. **14**, 857–863 (2005)
7. Shen, Y., Nakajima, M., Ridzuan Ahmad, M., Kojima, S., Homma, M., Fukuda, T.: Effect of ambient humidity on the strength of the adhesion force of single yeast cell inside environmental-SEM. Ultramicroscopy **111**, 1176–1183 (2011)
8. Shen, Y., Nakajima, M., Kojima, S., Homma, M., Fukuda, T.: Study of the time effect on the strength of cell–cell adhesion force by a novel nano-picker. Biochem. Biophys. Res. Commun. **409**, 160–165 (2011)
9. Ahmad, M.R., Nakajima, M., Kojima, M., Kojima, S., Homma, M., Fukuda, T.: Instantaneous and quantitative single cells viability determination using dual nanoprobe inside ESEM. IEEE Trans. Nanotechnol. **11**, 298–306 (2012)
10. Ashkin, A., Dziedzic, J.: Optical trapping and manipulation of viruses and bacteria. Science **235**, 1517–1520 (1987)
11. Chiou, P.Y., Ohta, A.T., Wu, M.C.: Massively parallel manipulation of single cells and microparticles using optical images. Nature **436**, 370–372 (2005)
12. Chu, Y.S., Dufour, S., Thiery, J.P., Perez, E., Pincet, F.: Johnson-Kendall-Roberts theory applied to living cells. Phys. Rev. Lett. **94**, 28102 (2005)
13. Puech, P.-H., Poole, K., Knebel, D., Muller, D.J.: A new technical approach to quantify cell–cell adhesion forces by AFM. Ultramicroscopy **106**, 637–644 (2006)
14. Franz, C.M., Taubenberger, A., Puech, P.-H., Muller, D.J.: Studying integrin-mediated cell adhesion at the single-molecule level using AFM force spectroscopy. Sci. Signal. **2007**, pl5-pl5 (2007)
15. Shen, Y., Nakajima, M., Kojima, S., Homma, M., Kojima, M., Fukuda, T.: Single cell adhesion force measurement for cell viability identification using an AFM cantilever-based micro putter. Meas. Sci. Technol. **22**, 115802 (2011)
16. Vogler, H., Draeger, C., Weber, A., Felekis, D., Eichenberger, C., Routier-Kierzkowska, A.L., et al.: The pollen tube: a soft shell with a hard core. Plant J. **73**, 617–627 (2013)
17. Dao, M., Lim, C., Suresh, S.: Mechanics of the human red blood cell deformed by optical tweezers. J. Mech. Phys. Solids **51**, 2259–2280 (2003)
18. Haga, H., Sasaki, S., Kawabata, K., Ito, E., Ushiki, T., Sambongi, T.: Elasticity mapping of living fibroblasts by AFM and immunofluorescence observation of the cytoskeleton. Ultramicroscopy **82**, 253–258 (2000)
19. Shen, Y., Nakajima, M., Yang, Z., Tajima, H., Najdovski, Z., Homma, M., et al.: Single cell stiffness measurement at various humidity conditions by nanomanipulation of a nano-needle. Nanotechnology **24**, 145703 (2013)
20. Zhang, C., Li, P., Liu, L., Wang, Y., Gao, Z., Li, G.: Development of mechanostimulated patch-clamp system for cellular physiological study. IEEE/ASME Trans. Mechatron. **19**, 1138–1147 (2014)
21. Ahmad, M.R., Nakajima, M., Kojima, S., Homma, M., Fukuda, T.: In situ single cell mechanics characterization of yeast cells using nanoneedles inside environmental SEM. IEEE Trans. Nanotechnol. **7**, 607–616 (2008)
22. Klein, F., Richter, B., Striebel, T., Franz, C.M., Freymann, G.V., Wegener, M., et al.:

Two-component polymer scaffolds for controlled three-dimensional cell sulture. Adv. Mater. **23**, 1341–1345 (2011)
23. Liu, N., Liang, W., Liu, L., Wang, Y., Mai, J.D., Lee, G.-B., et al.: Extracellular-controlled breast cancer cell formation and growth using non-UV patterned hydrogels via optically-induced electrokinetics. Lab Chip **14**, 1367–1376 (2014)
24. Tasoglu, S., Diller, E., Guven, S., Sitti, M., Demirci, U.: Untethered micro-robotic coding of three-dimensional material composition. Nat. Commun. **5**, 3124 (2014)

Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly

Thomas M. Crawford[1] and Olin T. Mefford IV[2]
[1]Department of Physics and Astronomy, Smart State Center for Experimental Nanoscale Physics, University South Carolina, Columbia, SC, USA
[2]Department of Materials Science and Engineering, Clemson University, Clemson, SC, USA

Synonyms

Self-assembly; nanofabrication; bottom-up processing

Definition

Nanomanufacturing originally refers to alternative approaches to achieving nanoscale resolution in manufactured devices by means other than integrated circuit fabrication (i.e., photolithography performed in a large-scale cleanroom on large area silicon wafers). One example of so-called bottom-up processing includes the concept of self-assembly [1], where nanoscale material components interact with each other or with external forces to build themselves into complex structures at larger length scales. Examples of external forces include thermal, electrical, magnetic, etc. Here, a magnetic field-driven nanomanufacturing technology is discussed that uses the disk from a commercial disk drive as a programmable template to direct self-assembly and nanomanufacture at cm length scales with nanometer resolution.

Commercial magnetic recording uses a nanometer-scale electromagnet to magnetically align <5 nm diameter Co grains that are embedded in a Cr matrix along one of two directions to create a binary pattern of 1s and 0s for the storage of information [2]. These patterns are then arrayed over a two-dimensional circular disk into concentric rings, known as tracks, with individual bits oriented along the circumference of the tracks. The electromagnet is fabricated as a part of a recording head that is on a movable arm that reaches individual tracks, while the spinning of the disk allows the head to reach the entire circumference of a given track. Typical patterns recorded onto a disk are as small as 12 nm by 30 nm, with transitions between regions a mere 1 nm in size. This size continues to decrease as the recording industry scales the component materials and technologies to increase the number of bits that can be stored on a single disk. To recover the stored 1s and 0s, a separate sensor fabricated on the same recording head uses a magnetically induced change in electrical resistance (magnetoresistance) to sense the spatially varying magnetic fields emitted from the patterned magnetic regions, which change sign depending on whether the recorded bit is a 1 or a 0.

Because the magnetic fields change sign over a spatial region that is only a few nm across, the spatial gradient in magnetic field located at these transitions can exceed 10 MT/m. In addition to enabling magnetic information storage, these field gradients can be used to nucleate and direct the self-assembly of nano-sized magnetic particles into complex patterns over the entire disk surface (approaching 50 cm^2). Nanoparticles as small as 5 nm in diameter can thus be assembled into cm-sized patterns with single nm resolution, spanning seven orders of magnitude in size with a single assembly step that is effectively passive.

The particles to be assembled are colloidally stabilized in a dilute fluid (ferrofluid). This fluid is

then introduced in the space above the disk medium, and the magnetic field gradients attract the particles to the surface through a combination of Newtonian mechanics and diffusion. As they arrive at the surface, they cluster around the recorded transitions, replicating the pattern that has been recorded into the medium. Once the pattern has been assembled on the template surface, the fluid and remaining particles are removed, and the surface is rinsed. Next, an air-curable, water-based polymer is spin coated onto the nanoparticle/template surface. After curing, the polymer is peeled from the template surface, transferring the patterned nanoparticle assembly to a free-standing polymer film, yielding an engineered polymer nanocomposite material. The recorded template can then be reexposed to new nanoparticle solution and the manufacturing process is repeated. Templates can be reused over and over to manufacture large quantities of material. The number and size of templates can be scaled to increase manufacturing output while maintaining single nanometer resolution. Once the template is recorded, the assembly process is passive. The fluid is placed on the surface, and the nanomanufacturing proceeds via the magnetic and other forces present at the disk surface.

Overview: Magnetic Field-Directed Self-Assembly

This entry will describe a method for nanomanufacturing using the patterned, nanoscale magnetic field gradients found at the surface of a standard, commercial, hard disk drive. The disk in the disk drive is coated with a thin film containing billions of tiny nanomagnets that can be magnetized by an external field. Once magnetized, these grains create their own magnetic field, and, at nanoscale transitions where these grains switch their magnetic direction, extremely large magnetic field gradients arise due to the degree of spatial confinement. This highly engineered nanotechnology can be adapted to enable manufacturing of macroscopic composites containing patterned nanomaterials, or hierarchical nanomanufacturing.

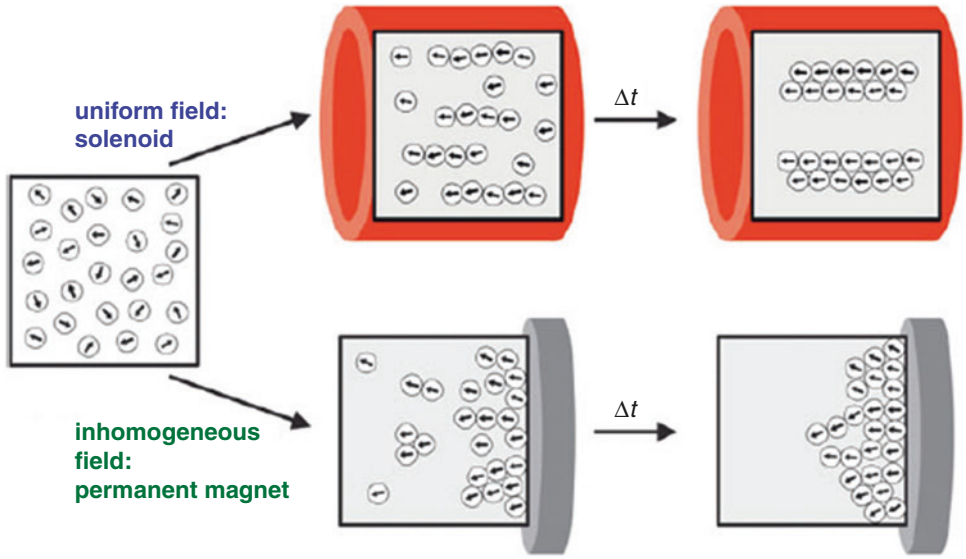
Francis Bitter first detailed his work observing magnetic inhomogeneities, i.e., magnetic domains, using a fluid suspension of maghemite γ-(Fe_2O_3) particles in 1931 [3]. Known thereafter as the "Bitter technique," the use of iron filings, micro- and nanoparticle ferrofluids, and other materials have long been staples of physical science demonstrations from elementary school through college. Broadly, the reason magnetic particles are attracted to these inhomogeneities is because a spatially changing magnetic field is present, i.e., a magnetic field gradient (Fig. 1). One can picture the process as follows. A local magnetic field magnetizes a particle that can be paramagnetic, ferromagnetic, superparamagnetic, etc., and then the magnetized particle is attracted toward the direction of increasing gradient, which is opposite to the direction of the aligned moment.

This magnetic force, which is why magnets stick to refrigerators, why giant electromagnets can pick up entire automobiles, and why entire trains can be levitated, allows assembly of magnetized objects into structures. In addition to these macroscopic manifestations of this force, it also works at the nanoscale, i.e., nine orders of magnitude smaller. While the concept of self-assembly is well established, as discussed in recent review articles [1, 4], magnetic field-directed self-assembly is now arguably a separate subfield within self-assembly [5]. Combined with recent interest in magnetic nanoparticles [6, 7], magnetic field gradient-driven assembly has seen strong interest over the past decade across multiple fields.

The motion of magnetized objects caused by magnetic fields is known as magnetophoresis. The force on a magnetized object can be calculated as the gradient of a potential energy; here the energy of a magnetized object in a magnetic field is as follows [8]:

$$\vec{F}_{\mathrm{mag}} = \mu_0 \vec{\nabla} \left(\vec{m} \cdot \vec{H} \right)$$

where in SI units, μ_0 is the permeability of free space, **m** is the magnetic moment of the magnetized objects in Am^2, and **H** is the magnetic field in A/m. By applying standard vector identities and



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 1 Schematic comparing magnetic field-directed self-assembly in uniform magnetic fields, i.e., produced by a solenoid, and in field gradients (i.e., spatially varying fields), which can be generated by permanent magnets. The uniform field aligns the nanoparticle moments, and it is the nonuniform field emitted by the particles themselves that self-assembles them into chains. For the permanent magnets, its inhomogeneous field generates attractive forces that pull the nanoparticles toward the region of the strongest field. *Arrows* indicate the magnetization directions of single-domain magnetic nanoparticles, and Δt indicates the passage of time in the applied field. Used with permission and reprinted from Ref. [5]

the definition of the vector product, this expression can also be written as

$$\vec{F}_{\text{mag}} = \mu_0 \left(\vec{m} \cdot \vec{\nabla} \vec{H} \right)$$

In the rewritten expression, the force is more easily seen as arising from the component of the magnetic moment along the magnetic field gradient scaled by the size of the gradient itself. Importantly, for this force to be maximized, the magnetic field must be sufficiently strong to magnetize the object, i.e., align its moment, in the direction of the local field. Thus, the magnitude of the force depends on both the size of the object's moment and the size of the gradient, as well as the degree to which that moment is aligned with the local direction of the field gradient. This force is the basic driver of magnetic self-assembly. In the next section, magnetic recording, i.e., the disk drive, is described with the goal to identify why it makes a potentially transformative template for nano assembly.

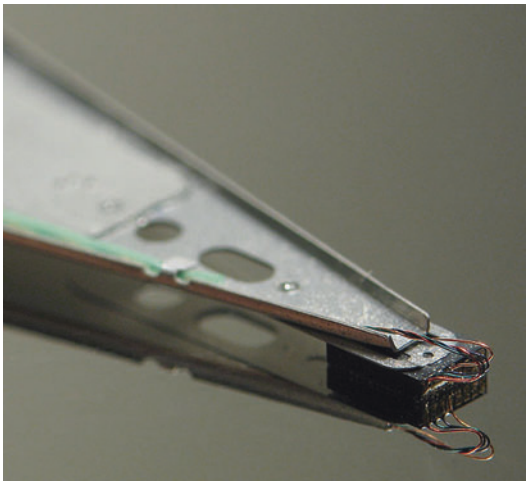
A Primer on Magnetic Recording

Figure 2 shows a cutaway view of the dominant magnetic recording technology, the disk drive. In addition to motors for spinning the disks and actuating the heads to sweep radially across the data tracks, the thin films that coat the disks, called the recording medium, together with the layers used to create the write electromagnet and the readback sensor, are the magnetic subsystems. Figure 3 shows a close-up of a recording head, while Fig. 4 shows a schematic of the recording process [2].

The medium is deposited onto an aluminum or glass disk and consists of multiple underlayers to establish the correct crystal texture for growing the recording layer itself, which typically is a thin



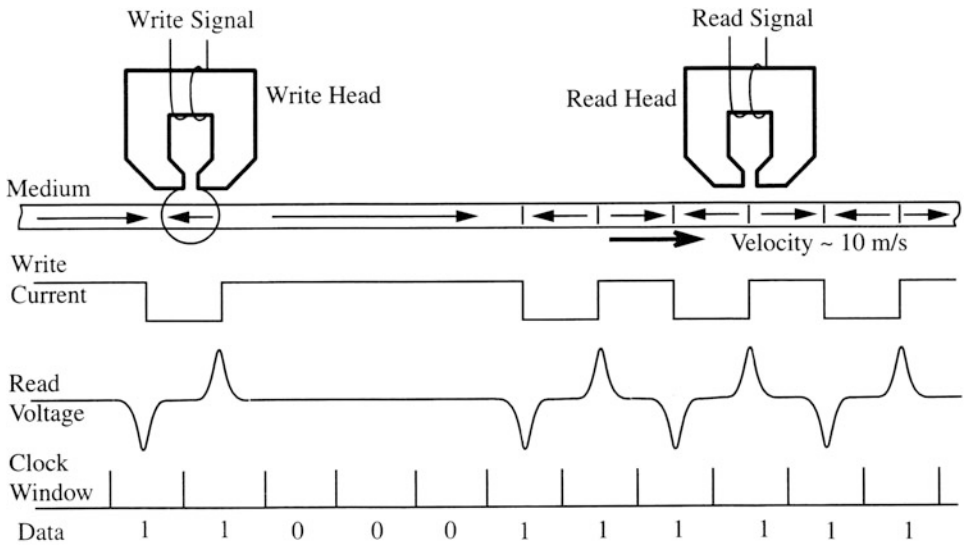
Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 2 The inside of a disk drive. The circular reflective surface is the recording medium, which is where the nanoscale magnetic field patterns are created by a magnet in the recording head, which is located at the end of the armature to the upper right of the platter. The heads pivot using a voice coil (far right corner) to access circumferential rings at different radii from the disk center



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 3 Close-up of recording head (and its mirror image in the reflective surface of the recording medium). The wires bring current to a >1 T electromagnet that writes, i.e., magnetizes the nanometer-scale Co grains that comprise the thin-film media. It also contains a magnetoresistive sensor that can detect the spatially changing magnetic fields that indicate the state of a data bit

film that contains Co, Cr, and Pt. The Pt acts as a catalyst to enable nanometer-sized Co grains to be separated by Cr at the grain boundaries, allowing for the proper magnetic anisotropy to control the grain-to-grain coupling. On top of the recording media layer is a diamond-like carbon layer followed by a monolayer hydrocarbon lubricant. The diamond-like carbon serves as wear protector, while the lubricant helps prevent the head from sticking to the media.

The recording head is fabricated using cleanroom-based nanoelectronic processing to build the nanometer-scale electromagnet on a highly wear-resistant ceramic of Al, Ti, and C. First, the magnetoresistive readback sensor is fabricated using high-resolution lithography. This sensor consists of several magnetic films together with biasing magnets to help orient the magnetic moments in the sensor to yield a linear resistance variation in response to an applied magnetic field. This readback sensor is deposited between two thicker, magnetically soft films that shield the sensor layer from stray magnetic fields, i.e., from adjacent magnetic bits. Often the top shield is connected with the bottom pole of the write electromagnet. After this first pole, a coil is deposited to energize the electromagnet. Finally, a top pole is deposited and high-resolution lithography is used to ensure that the field of the electromagnet is confined to the ~30 nm wide region of a single track.



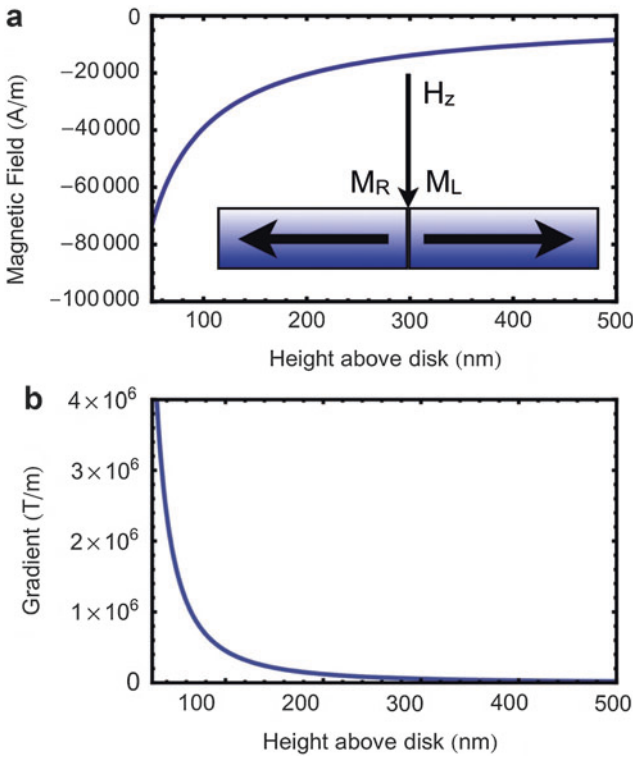
Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 4 Schematic of magnetic recording process (longitudinal recording). The write head magnetizes the medium. Where the current energizes the electromagnetic in the write head, a transition between directions (shown as opposing arrows in the medium) is created. At these locations a spatially confined magnetic field is emitted from the medium (here labeled as "read voltage"). In magnetic recording, these spatially confined magnetic fields indicate a "1" in a bit sequence; they create enormous magnetic field gradients that can be used to direct assembly of magnetic nanoparticles for nanomanufacturing. Used with permission and reprinted from Ref. [2]

The recording head is mounted on a suspension that includes a gimbal to allow the head to fly over the surface of the disk that spins beneath it. This gimbal allows the AlTiC substrate, called a slider, to both pitch and roll. The part of the head that is exposed to the disk, called the air bearing surface, is in reality the thickness of the substrate, while the layers that comprise the read and write head are oriented 90° to the plane of the disk. This surface is further textured with lithography and etching to provide a negative pressure cavity that counteracts the lift generated by the air dragged under the head by the spinning disk, in order to keep the head ~3–5 nm away from the disk surface during recording.

To illustrate the enormity of the field gradients present at the recorded "bits" in disk drive media, Fig. 5 plots the perpendicular magnetic field (Fig. 5a) and the perpendicular component of the magnetic field gradient (Fig. 5b) as a function of height above the disk medium surface (centered on the transition shown in the inset) [2].

As discussed above, the field gradient near the surface is three orders of magnitude larger than the gradients traditionally used for magnetophoretic nanoparticle transport. The magnetic force on a nanoparticle can be integrated over different heights to obtain the work done by the magnetic field [2, 9]. In moving a 10 nm particle from 50 to 40 nm, the magnetic force in Fig. 5 does $\sim k_B T$ of work on the nanoparticle at 295 K. Thus, at five nanoparticle diameters above the surface or closer, the work done by the magnetic field rapidly grows beyond the thermal energy that drives Brownian motion. Even though the work done by the field is less than $k_B T$ for $z > 50$ nm, it is always directed toward the surface, while Brownian forces are randomly oriented. This force anisotropy assists the recorded transitions to capture nanoparticles and assemble them on the medium surface.

Regardless of whether one considers 1.5 T electromagnets that produce 30 nm confined fields, recorded bit areas that are 10×30 nm^2, or a 3 nm separation between the disk and a head flying at 30 m/s, the disk drive is a nanotechnology in almost every facet of its operation. Before discussing how to nanomanufacture composites



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 5 (**a**) Magnetic field in the z-direction (z = height above the disk) of a longitudinally recorded transition as a function of height. (**b**) Z-component of the gradient of the z-magnetic field as a function of height above the disk. Note that the gradient changes rapidly with height, falling from a maximum of 10–20 MT/m at 10–20 nm above the disk to kT/m in a mere 500 nm. Used with permission and reprinted from Ref. [9]

and devices using magnetic field-directed self-assembly with a disk drive, it is important to consider the raw material that is used in the manufacturing process, a material that comes with its own set of built in complexities, the chemically-synthesized magnetic nanoparticle.

Synthesis of Magnetic Nanoparticles

The production of magnetic nanomaterials can be accomplished either through top-down or bottom-up techniques. Regardless of production technique, the vast majority of these colloidal suspensions consist of magnetic particles suspended in some sort of carrier media. The composition of the magnetic particles is typically magnetite (Fe_3O_4) or maghemite ($\gamma-Fe_2O_3$) because they are biocompatible and have relative oxidative stability. To maintain stable suspensions of these colloids, attractive forces including van der Waals and magnetic interactions need to be overcome through electrostatic, osmotic, and/or repulsive forces [10, 11]. Typically the surface of the inorganic particle is modified with an organic material to alter the surface charge or provide a steric layer to reduce interparticle interactions. Concentration of these materials is also important as rates of aggregation have been tied to the average interparticle distance, which is derived from concentration [12, 13].

To produce the magnetic nanoparticles, several different methods have been developed. A brief overview of topics is described here. Nonetheless, the reader is encouraged to seek more detailed reviews available in the literature [10, 14–17].

One of the earliest methods to produce these materials was the top-down technique of grinding these materials to sizes suitable for a colloidal suspension. For example, Papell first described the preparation of oleic acid-stabilized magnetite particles in heptane following grinding for 19 days [18]. Later, Rosensweig and Kaiser demonstrated that ferrofluids could be prepared by grinding for a variety of media beyond hydrocarbons, including fluorocarbons, water, and

esters [19]. While this method has some advantages, including low cost, versatility, and the ability to create large batches of materials, there are distinct limitations in size control of the particles produced.

One of the simplest bottom-up technique is the coprecipitation of metal salts to form particles. This technique was pioneered by Khalafalla and Reimers [20]. Within the nanoparticle community, this is sometimes referred to as the "Massart technique" as this methodology is the one most commonly used for the formation of aqueous-based fluids [21]. To date, this is an extremely popular method for those new to the field. In addition, it makes a great demonstration for students of any age [22]. Typical sizes of magnetite nanoparticles produced by coprecipitation are around 10 nm in diameter. Unfortunately broad polydispersities and aggregation are concerns with this process.

In an effort to control particle size, reverse micelles have been used. For example, Pileni demonstrated the production of cobalt ferrite nanoparticles with relatively uniform size using dodecyl sulfate inverse micelles as templates [23]. Alternatively, by using dioctyl sodium sulfosuccinate (AOT) to form micelles, chemistries similar to coprecipitation can yield improved size control [24]. While there are some improvements, this technique has yet to provide the versatility promised. Also, the presence of the charged surfactants used to create these materials may present challenges for additional modification.

In a similar vein, hydrothermal reactions can be employed with aqueous-based reactants in high-pressure autoclaves, where typical reaction temperatures are 200-250 °C and pressures are ~ thousands of psi. An example synthesis includes the production of iron oxide nanorods via the reaction of iron sulfate, iron (III) chloride, and ethylenediamine in water at 120 °C for 20 h in a sealed reactor [25]. Similar methods include use of iron (II) and iron (III) chlorides and ammonium hydroxide and heating at 250 °C for 24 h [26]. Again, this method does help with size distributions, but the resulting particles are typically heavily aggregated. Benefits do include the use of aqueous methods that are more environmentally benign than organic routes.

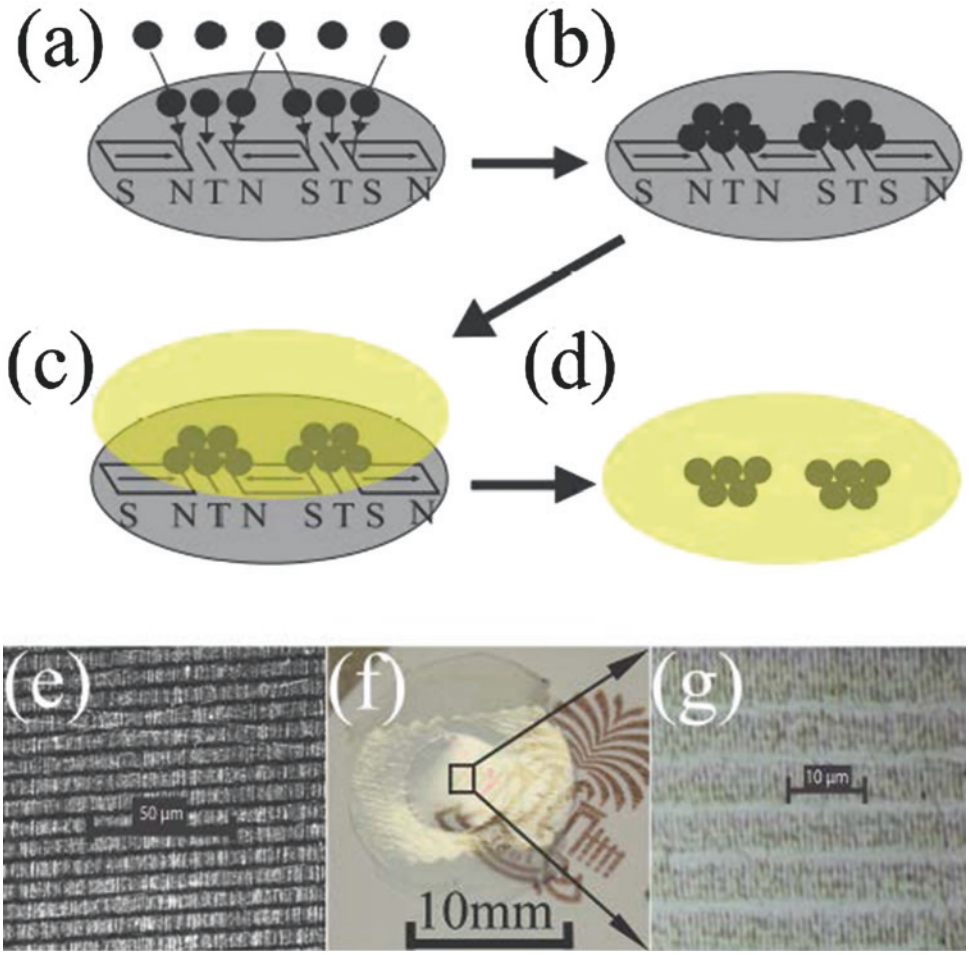
Thermal decomposition of chelates of metal precursors provides the best route to date of creating magnetic particles with uniform particle size distributions. With careful selection of stabilizing ligands, these materials can be further modified by removal and replacement of the sacrificial ligand, which is often oleic acid and oleylamine. The seminal papers in this area include the thermal decomposition of iron (III) acetylacetonate [27, 28], iron pentacarbonyl [29], and iron oleate [30, 31]. Typical procedures involve controlled reaction temperatures typically in the range of 300–350 °C, where the organic metal chelate is decomposed and off gassed.

More recently alternatives to traditional chemical processes are being explored. For example, flame spray pyrolysis is another method to produce particles but has limitations associated with addition modification. The composition and oxidative state can be controlled through the gas input of the flame [32]. Another alternative method of note is electrochemical or electrodedeposition, where a sacrificial metal anode(s) is charged in solution media containing surfactant. The size of the resulting particles can be controlled through the current applied to the system [33].

Nanomanufacturing 2D Structures

To use the spatially confined fields at the surface of the disk drive to direct self-assembly of magnetic nanoparticles, the magnetic recording medium must be programmed (encoded) with a desired pattern to be manufactured. After the template is programmed, the recording medium is coated with dilute colloidal fluid containing the magnetic nanoparticles discussed in the previous section. These nanoparticles can be thought of as raw materials or feedstocks for the manufacturing process. Figure 6 graphically illustrates the process schematically and with images [34].

The original studies of self-assembly for nanomanufacturing with recording media templates were performed on prerecorded, 1 cm diameter pieces of longitudinal magnetic recording media. These media have magnetization ~450 kA/m as determined using a vibrating sample



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 6 Schematic of nanomanufacturing with magnetically recorded nanotemplates. (**a**) Magnetic nanoparticles are attracted to the magnetic field gradients that are emitted from longitudinally recorded transitions on the surface of recorded magnetic media. (**b**) The nanoparticles self-assemble onto the medium surface at the recorded template locations. (**c**) A liquid polymer film is spin coated onto the disk surface. (**d**) After curing, the film is peeled from the disk medium. (**e**–**g**) Images of the assembled nanoparticles on the medium, the peeled film, and the nanoparticles in the film after transfer, respectively. Used with permission and reprinted from Ref. [34]

magnetometer (VSM). They are patterned into sectors along the circumference of the disk prior to distribution, through a process called servo writing, which creates regions known as servo sectors between the data sectors [2]. These regions contain complex magnetic field patterns which are used with a servo algorithm to keep the recording heads accurately positioned over the data tracks during read and write operations. It is these regions which are easiest to use for nanoparticle self-assembly [9, 35].

To create custom magnetic field templates to assemble user-customized patterns, it is critical to be able to record an arbitrary pattern into the medium. Fortunately, magnetic recording has been performed by raster scanning the head in contact with the disk for a number of years, using an instrument called a contact recording tester [36, 37]. The contact tester is similar to a scanning probe microscope, only it uses a recording head instead of a sharp tip. Either the head or the media is mounted on a precision XY scanning

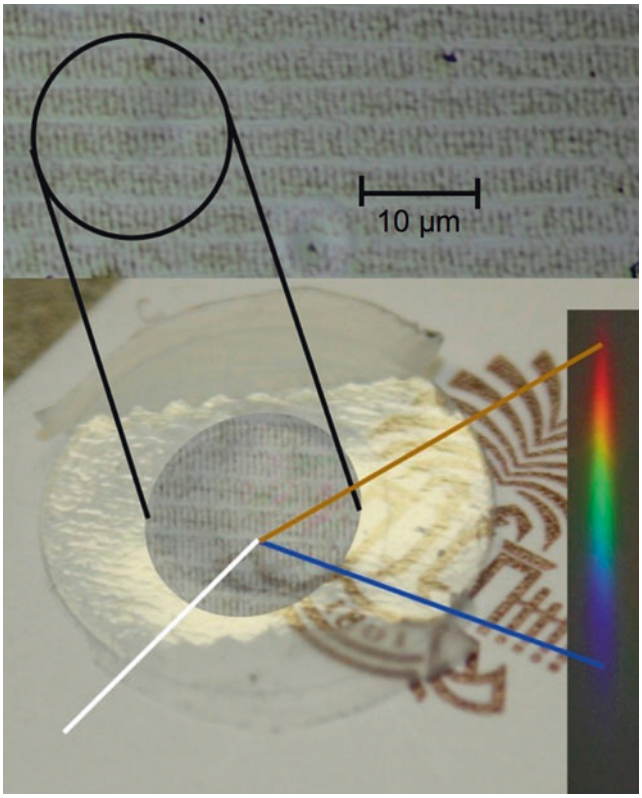
stage, and electrical connections are made to the recording head to allow for both programmable magnetic recording and pattern imaging using the readback sensor.

A variety of nanoparticle species can be assembled; however, they have to contain sufficient magnetizable matter so that the magnetic field gradient can exert a force on the particle sufficient to cause it to move through the fluid. The original experiments employed both commercial ferrofluid and in-house colloidal nanoparticles synthesized using standard coprecipitation recipes. Monodisperse thermal decomposition particles (as discussed above) can also be used, and assembly can be done in water or a nonpolar organic solvent such as hexane. The ferrofluid is diluted significantly so the nanoparticles are well separated and is then pipetted onto the surface of the medium. Alternatively, the media coupon can be dipped into a beaker filled with the dilute ferrofluid. The superparamagnetic nanoparticles then assemble on the patterned magnetic transitions in a passive process that can take anywhere from 1 min to an hour, depending on the properties of the ferrofluid. After assembly, the ferrofluid is removed from the coupon surface. This removal can be done with a pipet, with a pump, or by simply tilting the coupon on its side and allowing the fluid to drain off. The coupon surface can be dried with nitrogen, or it can be rinsed with freshwater or hexane and then dried. After assembly and fluid removal, the coupon surface contains a patterned array of self-assembled nanoparticles that are stable and easily imaged with an optical microscope or a scanning electron microscope (for nanometer-scale features). Typically dark-field optical microscopy is used together with a CCD camera and imaging software to acquire images of the patterns for analysis. The diamond-like carbon (DLC) coating on the surface of the medium is sufficiently conductive, so charging in the SEM is not an issue, and high-quality images of the assembled nanoparticles can be obtained with nanometer resolution. To obtain quantitative information about the height of the assembled nanoparticles, atomic force microscopy (AFM) can be used. However, care must be taken to ensure that the AFM tip does not become contaminated with nanoparticles that can reduce the lateral resolution or corrupt the height measurement.

Transferring the Assembled Nanoparticles to a Standalone Film

After metrology, to confirm that the nanoparticles are assembled on the disk drive surface using optical, electron, or atomic force microscopy and to verify that the pattern is visible and properly defined, the next step in the nanomanufacturing process is to remove these nanoparticles from the template surface and embed them in a separate film for application. Once the nanoparticle assembly is transferred, the disk template can be cleaned and the assembly process repeated. Because of its wear-resistant diamond-like carbon coating, the media template can be used repeatedly to manufacture many standalone assembled nanoparticle films. Obviously, many templates could be coated simultaneously to achieve manufacturing scale, and in principle, each template could have a different pattern.

To transfer the nanoparticle array, a liquid polymer is spin coated onto the medium surface. One example of an air-curable water-based polymer is Diskcoat 4220 from General Chemical Co., Brighton, MI, which is designed to be cured and then peeled from a surface. After spin coating, the polymer films are cured in an airstream for $\sim$20 min at room temperature. After curing, the polymer–nanoparticle assembly is peeled from the medium with adhesive tape. Using circular tape with a central hole yields a suspended film with a 5 mm diameter circular window. The Diskcoat can be diluted in DI water, and by varying dilution, spin speed, and spin time, a variety of polymer thicknesses can be obtained, ranging from 100 nm to several micrometers. Typically it is difficult to peel films thicker than 200 nm and still yield an intact film, and peeled film thicknesses range from 200 to 500 nm. With peeling, these polymer films remove the nanoparticle assembly from the medium surface, i.e., the peeling causes pattern transfer, leaving the medium clean and the nanoparticle array embedded at the



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 7 Upper panel: Optical microscope image of a diffraction grating embedded in polymer film nanomanufactured using pattern transfer nanomanufacturing. The grating period was 750 nm as shown with the 10 μm scale bar. The lower panel shows the adhesive tape (with a cartoon of the grating) showing white light diffraction from the grating. The rainbow spectrum at right was obtained from an all-nanoparticle diffraction grating nanomanufactured PTNM™

surface of the cured polymer. These nanomanufactured materials are flexible, transparent, and contain the programmed arrays of nanoparticles [9, 34].

Figure 7 shows a reflection optical diffraction spectrum created by sending collimated white light through the assembled nanoparticle layer in the central 5 mm region in the lower panel. The peeled films are mechanically stable but are relaxed, meaning they can have wrinkles. These wrinkles can be made to repeatedly change shape in response to a temporary increase in humidity, temperature, or magnetic field. For example, breathing on the polymer film, induces a shape distortion sufficient to cause the diffracted spot to vanish. However, the diffraction returns as soon as the humidity is reduced to ambient.

These polymer/nanoparticle diffraction gratings have been used in a spectrograph to demonstrate the viability of these nanomanufactured materials. Using standard diffraction grating analysis, the angle of diffraction was measured for different optical wavelengths, verifying that the line spacing was close to what was measured with the optical microscope. Importantly, the analysis of the diffracted signal from the nanomanufactured grating was based on a 1–2 mm diameter region of the film, whereas the optical microscope measurement was done on a limited sample size, say 50–75 lines. The optical microscope and diffraction measurements yielded 742 $\pm$ 12 nm and 770 $\pm$ 10 nm, respectively [34]. Thus, while the agreement was not perfect, the diffraction measurement was over a much larger sample of nanoparticle patterns than could be counted from a single (or even multiple) microscope images. Given the quality of the diffracted spectra (which were limited by the optical components of the spectrometer and not by the grating), these nanomanufactured materials could find application for building wavelength-sized structures for photonics applications.

To prove that the pattern transfer process completely transfers the nanoparticles to the polymer film, magnetometry is performed on the first

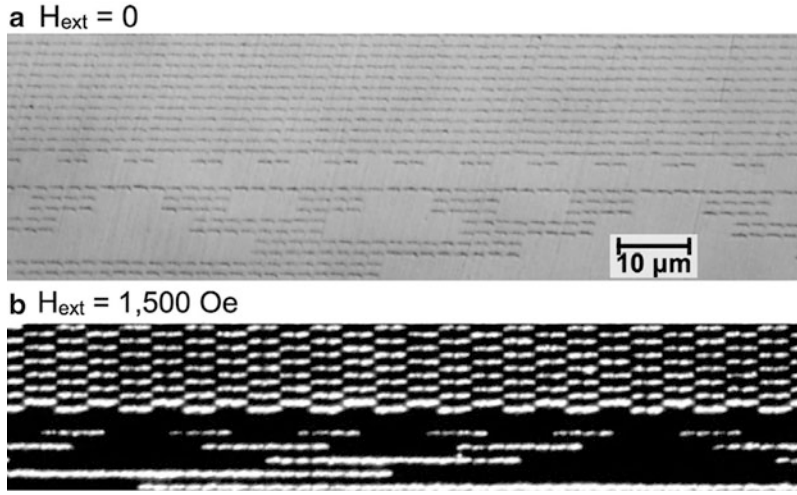
peeled film, yielding a magnetic moment versus field curve that would be expected for superparamagnetic nanoparticles. After the first transfer, second and third polymer coat, cure, and peels were performed on the same template without cleaning the surface. The second and third peeled films were measured with the VSM, and the resulting measurements are indistinguishable from measurements made on films without any nanoparticles. Based on the VSM's rms sensitivity of 1×10^{-6} emu, which is 2 % of the total moment measured from first peeled films that contain the patterned, transferred nanoparticles, at minimum, 98 % of the nanoparticles are transferred to the film during the initial peeling step. Optical microscopy images obtained of the coupon surface after peeling show no remaining nanoparticles, and AFM measurements of coupons after peeling likewise show no particles in several locally imaged regions.

Manufacturing Process Extensions

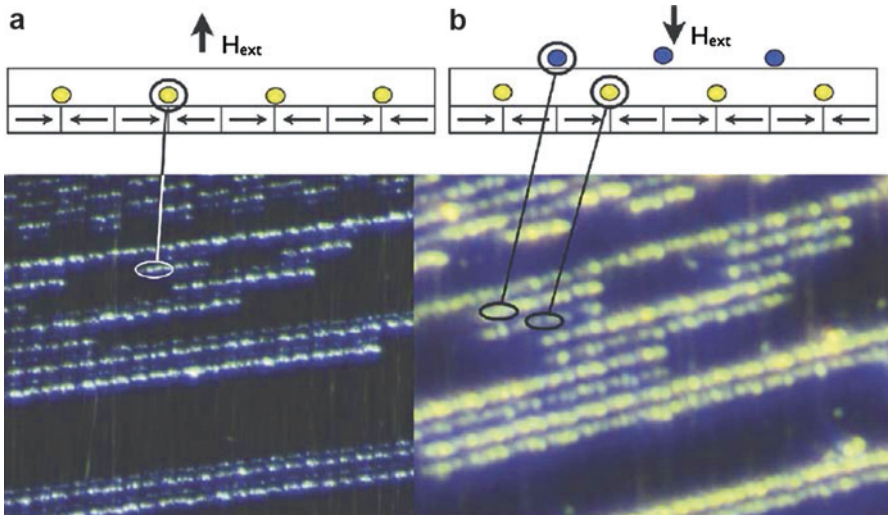
With the pattern creation, assembly and spin/cure/peel polymer processes established, there are several extensions that can be made to gain further control over the assembly process. For example, we extend the nanoparticle assembly process by applying a uniform external field over the fluid cell during the nanoparticle self-assembly step [9].

Figure 8a, b shows optical images of the same region of media coated in zero external-field (Fig. 8a) and with an external field, $H_{ext} = 1.5$ kOe, directed toward the medium surface (Fig. 8b). In contrast to the zero-field assembly process, here the external field superposes with the field emitted from the patterned transitions and changes the assembly process. Superposition of a 1.5 kOe field with the transition field means H_{net} goes to 0 at a height of $z = 20$ nm above the surface. Therefore, the external field will exceed the medium field for $z > 20$ nm. The larger external field will cause nanoparticles further than 20 nm from the surface to be magnetized along the external field direction.

These pre-magnetized nanoparticles will now be repelled from transitions with oppositely directed local fields (with $H = 0$ at $z = 20$ nm) while still being attracted to transitions whose local field constructively superposes with the external field. Thus, in Fig. 8a every transition is coated, while in Fig. 8b only every other transition is coated. Close inspection of Fig. 8 shows



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 8 (**a**) Optical image of servo patterns assembled in zero external field. (**b**) Same pattern showing nanoparticles assembled in a 1.5 kOe magnetic field. Note that every feature is no longer coated, with gaps opening up in the structures that are closed in panel (**a**). Used with permission and reprinted from Ref. [9]



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 9 A bilayer, interleaved composite. (**a**) Image of first layer assembled in an external field. Note that the double features are visible, but one is much weaker. (**b**) Image after a second round of coating was performed only this time in an oppositely directed magnetic field. The features that were dim in panel (**a**) are now bright. The schematic above the images represents which layers and which locations correspond to particular features in the images (Used with permission and reprinted from Ref. [9])

similarities between the two images, with spatially modulated features in the upper half of Fig. 8b replacing horizontal lines in the upper half of Fig. 8a, and gaps opening in the lower half of Fig. 8b compared with the lower half of Fig. 8a. Generally, there are half as many coated features in Fig. 8b as in Fig. 8a. These changes from Fig. 8a to b are all caused by the repulsion of pre-magnetized particles from opposing, spatially alternating transitions.

While $H_{ext} = 1.5$ kOe is sufficient to prevent nanoparticle coating on opposing transitions, in weaker external fields, for example, $H_{ext} = 900$ Oe, the opposing transitions can be partially coated with nanoparticles. Since the external field magnitude controls the height where $H_{net} = 0$ above opposing transitions, as well as enhances the magnetization of nanoparticles above aligned transitions, transitions with constructive field superposition can yield thicker nanoparticle layers than the zero external field case, while transitions which oppose the external field produce a thinner layer than the zero-field case. Thus, manufacturing process control over topography within the assembly can be achieved, or to create assemblies with varying nanoparticle layer thicknesses within the assembly simply by applying a varying external magnetic field.

The direction of the external field can be used to control the pattern across multiple layers to form a vertically separated, horizontally interleaved nanoparticle composite, with the layers assembled in oppositely directed fields and separated by a polymer layer. Figure 9a, b shows this process, with the left panel (Fig. 9a) demonstrating external field assembly at "positive" field transitions. After removing the nanoparticle fluid, the layer is spin coated with the polymer. Here instead of peeling the polymer, the polymer surface is exposed to fresh magnetic fluid, now with a "negative" magnetic field (right panel (Fig. 9b)), and thus the particles assemble on "negative" field transitions. After this second nanoparticle assembly step, the film is coated with a second layer of polymer, which is cured and peeled from the medium. The dark-field images show alternating bright and dim features. The dimmer interleaved features are the "negative" field transitions, which are still slightly coated at the intermediate external field used

here, $H_{ext} = 900$ Oe. The dark-field image in panel (Fig. 9b) (two layers of nanoparticles) shows bright nanoparticle lines at the locations in panel A that were dim and weakly coated. Thus, Fig. 9a, b demonstrates the potential for nanomanufacturing 3D composites with both inter-layer and intra-layer control over the assembled topography and feature spacing.

Metrology for Pattern Transfer Nanomanufacturing™

Measuring the assembled nanoparticles to determine pattern quality and repeatability, is a critical step in establishing whether this approach can lead to manufacturable materials. As discussed above, a variety of microscopy techniques have been employed to image these patterns. One difficulty with large area patterns is not having the ability to measure the same region repeatedly. To determine repeatability of the assembly process over multiple assembly steps, we employed optical photolithography to place a series of numbered frames on top of the disk coupon [35]. These 30 nm thick Cu frames were deposited onto patterned photoresist using sputtering, and the photoresist was then lifted off to permanently label particular regions of the disk media surface. These frames enable one to find the exact same region of the disk with AFM or SEM so that the assembled pattern can be (re)imaged as a function of coating process, nanoparticle species, colloidal stability, etc., to determine process repeatability and optimize the assembly process. Figure 10a–d shows a series of images of disk drive servo patterns: Coupon images obtained with (Fig. 10a) MFM, (Fig. 10b) dark-field microscopy, and (Fig. 10c) AFM of the coupon surface after assembly. Figure 10d shows an AFM of the peeled film itself. While Fig. 10a–c demonstrate that the nanoparticle assembly (imaged both optically and with AFM) closely matches the magnetic force gradients imaged with MFM, Fig. 10d shows that the nanoparticles maintain their pattern after transfer to the polymer film [9].

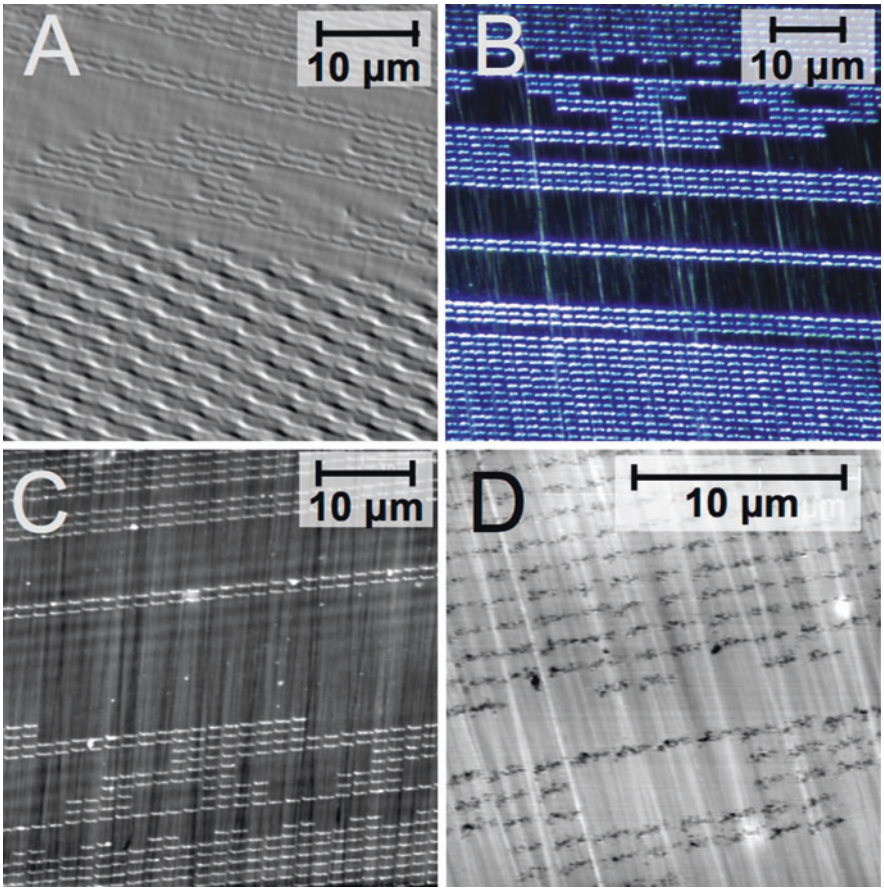
Note the striations observed in Fig. 10c, d. These marks are ~10 nm high and are likely polishing marks left from the disk manufacturing process. These surface features do not change from coating to coating and therefore make an excellent reference that can be used for image correlation to determine the repeatability of our assembly process. Cross-correlation of these marks obtained by aligning multiple images and correlating the same line scan from two images (Fig. 11) suggests that the spatial precision of the nanoparticle assembly on a particular region of the disk surface is $\sim 25 \pm 11$ nm [35].

Real-Time Diffraction Metrology

Early experiments showed that the longer the colloidal fluid remains on the media coupon, the more nanoparticles are assembled. Smaller concentrations and shorter times gave fainter patterns, while larger concentration and longer times gave darker patterns. A key need, therefore, is to understand the dynamics of the nanoparticle assembly process, as well what happens over a large patterned area beyond simply imaging it. A difference of a few seconds is hard to observe from an image.

Similarly, it takes a significant quantity of time to obtain AFM images from multiple repeated coatings, and thus a technique to monitor the assembly process as a function of time and assess quality over large areas was needed. Since diffraction grating patterns had been assembled and used in a spectrometer as discussed above [9, 34], a real-time dynamic measurement of optical diffraction was attempted [38]. Note that without the nanoparticles, the medium surface is not topographically patterned, merely magnetically patterned. Thus, without imaging the disk using a magnetically sensitive probe, an unpatterned disk appears visually flat. However, as soon as nanoparticles begin to assemble onto the medium, they are directed by the recorded template. At the earliest stage of assembly, these patterns appear to the eye, in an electron or optical microscope image, to be random. However, the diffracted signal exists almost immediately because the scattered light from these patterned nanoparticles is in phase on the diffraction detector.

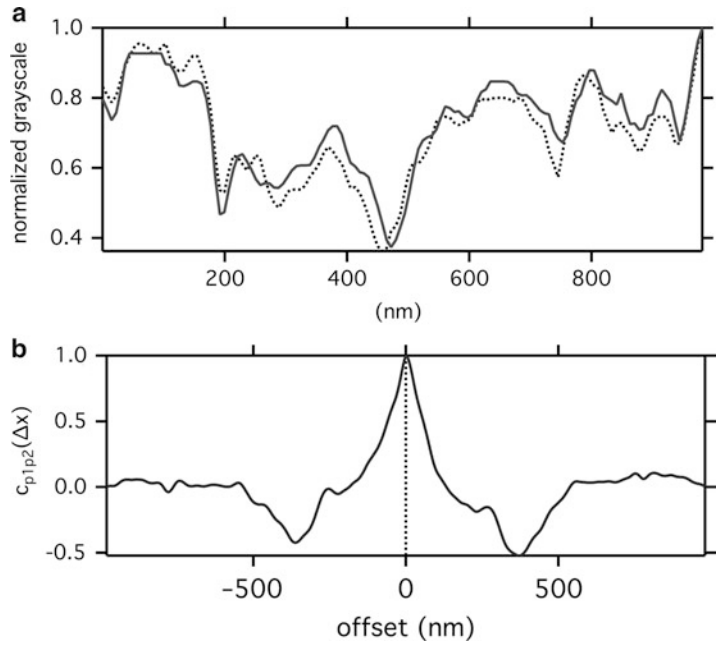
To perform this real-time study of nanoparticle self-assembly, the disk media coupon is placed in



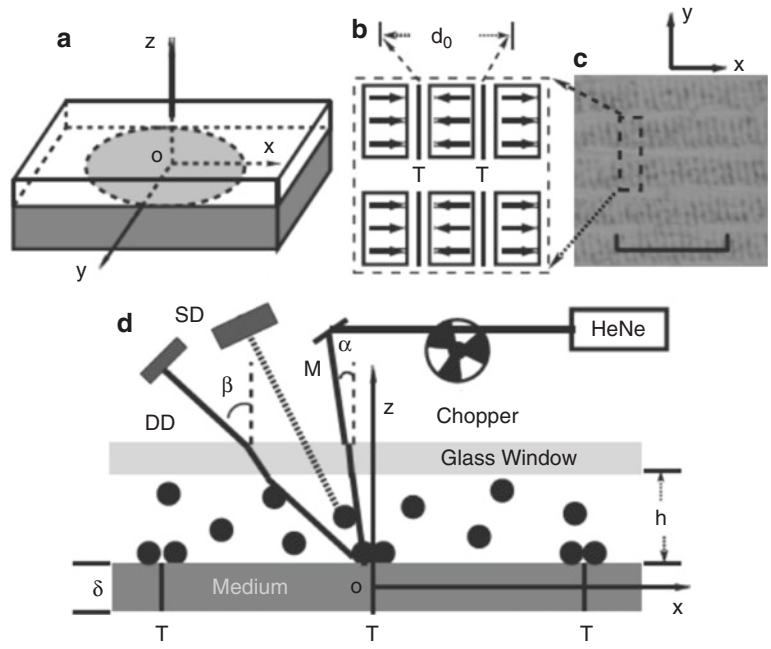
Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 10 Metrology verifying pattern transfer nanomanufacturing. (**a**) Magnetic force microscopy (MFM) image of disk media without nanoparticles showing the underlying magnetic field template. (**b**) Optical image of assembled nanoparticles. (**c**) Atomic force microscopy (AFM) image of the same region shown in (**a**–**b**). (**d**) AFM image of the peeled film showing the transferred nanoparticles as well as the molding of the polishing striation topography into the surface of the polymer film (Used with permission and reprinted from Ref. [9])

a transparent fluid cell that has a glass lid. Figure 12 shows a schematic of the apparatus used for real-time monitoring of diffraction during assembly. The medium coupon is located at the bottom of the cell, 0.5 mm vertical distance (h) from a glass window held onto the cell by a vacuum pump. The nanoparticle suspension is injected into the fluid cell by a syringe pump, and the fluid flow is stopped immediately after the fluid is pumped into the cell. A HeNe laser (wavelength 632.8 nm) is incident onto the coupon with a 5° angle to avoid multiple light reflections between the glass window and the coupon surface. The diffraction angle is related to the grating spacing and the laser wavelength by the standard diffraction grating equation [39]. A photodetector (DD) is used to measure first-order diffraction, while a second photodetector (SD) is used to measure the laser scattering from the nanoparticles in the fluid as well as on the surface.

Both light scattering and diffraction efficiency (SE and DE) are measured and are given by the ratio of the scattered and diffracted intensity divided by incident power, respectively. Of course the diffraction detector includes intensity both from the grating as well as background scattering. The diffraction plus scattered signal changes



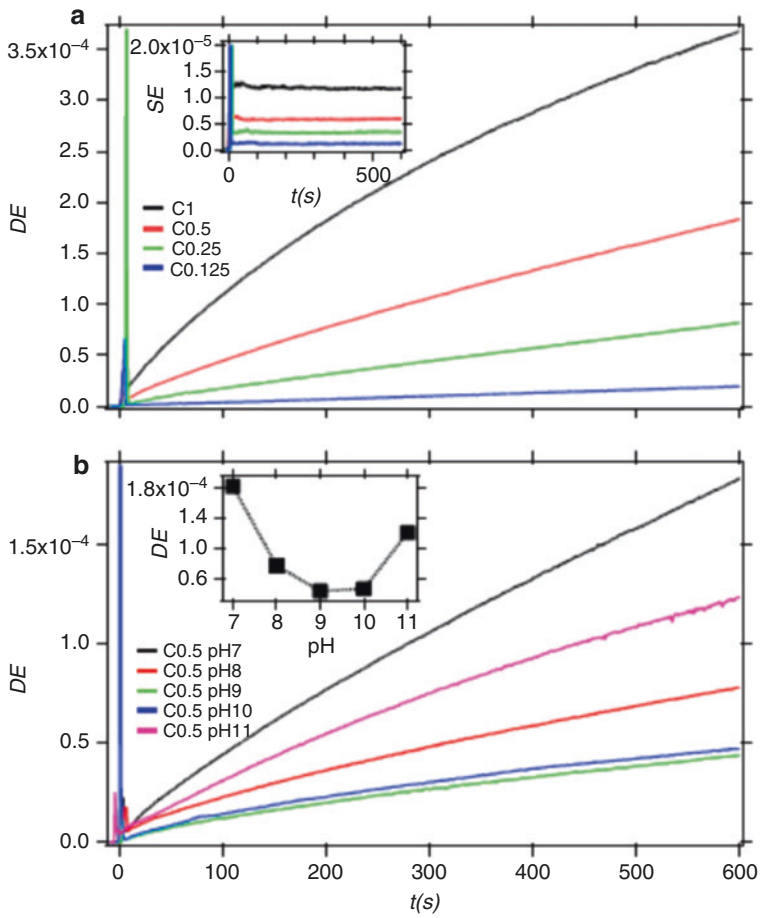
Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 11 Cross-correlation of polishing striations obtained from two separate AFM images of the assembled nanoparticles. The resolution of the cross-correlation is ~1 nm and was used to obtain an assembly-to-assembly repeatability of 25 ± 12 nm (i.e., within 1 single 12 nm nanoparticle). Used with permission and reprinted from Ref. [35]



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 12 Apparatus for measuring diffraction in real time during the nanoparticle self-assembly process. (**a**) Schematic of fluid cell with outline of media coupon. (**b**) Orientation of magnetic transitions and (**c**) corresponding optical image. (**d**) Experimental schematic showing optical detection scheme at an arbitrary point in time during the self-assembly process. Used with permission and reprinted from Ref. [38]

linearly with the scattered intensity, and the diffraction intensity is obtained by subtracting the scattered intensity from the diffracted intensity.

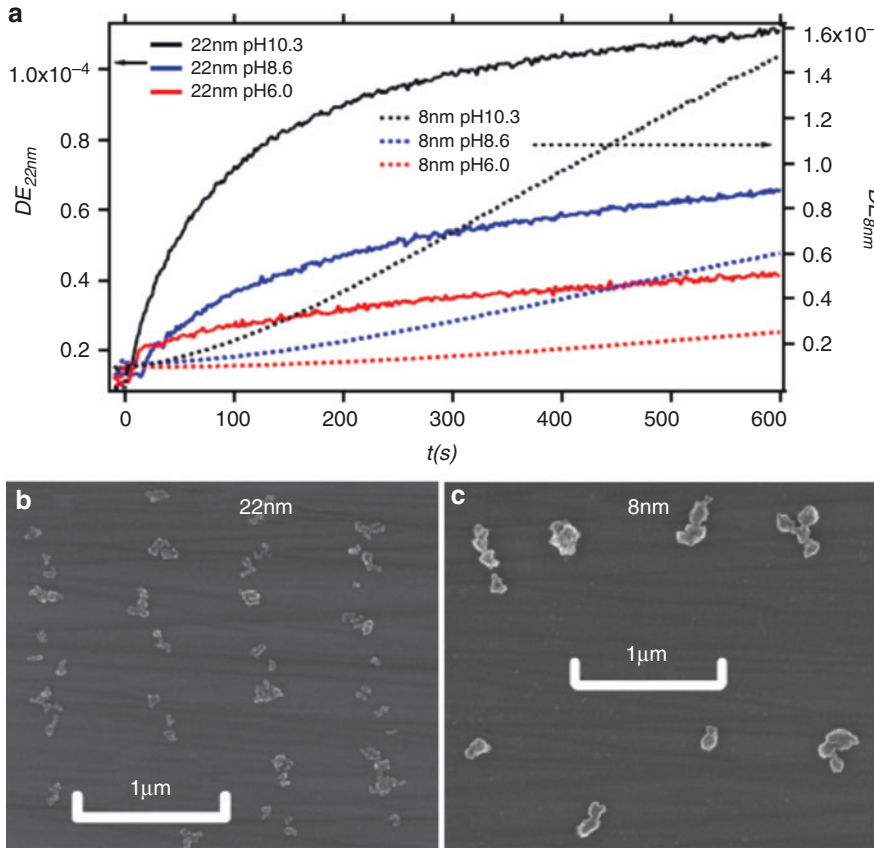
Figure 13 shows the measured first-order diffraction as a function of time from zero to 10 min for four particle suspensions with different concentrations 1, 0.5, 0.25, and 0.125, respectively (curves C1, C0.5, C0.25, C0.125 in Fig. 13a). Here, C1 has a 0.002 % volume concentration of nanoparticles. Both signals show spikes that



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 13 Real-time measurement of nanoparticle nanomanufacturing process. (**a**) Diffraction efficiency as a function of time for different colloidal nanoparticle concentrations (inset shows scattering intensity). (**b**) Diffraction efficiency versus time for different solution pH values. (Inset shows dependence of diffraction efficiency on pH.) Used with permission and reprinted from Ref. [38]

remain for 3–5 s, caused by light scattering from the wavefront of flowing fluid. After the spike, the SE remains constant as a function of time [inset in Fig. 13a], and the average SE is a linear function of the suspension concentration. However, the diffracted intensity is not linear, but increases rapidly at a rate that slows as time increases. For the same instant of time, larger concentrations produce larger diffraction (doubling concentration produces almost doubles the DE for $t > 100$ s). Figure 13b shows first-order DE as a function of t for C0.5 as a function of pH, for values of 7, 8, 9, 10, and 11. The DE decreases as the pH value increases but then begins to increase again [inset in Fig. 13b]. At $t = 600$ s, the pH7 suspension produces a DE three times larger than the pH9 suspension. Interestingly, the average hydrodynamic diameter obtained using dynamic light scattering does not change with pH for these nanoparticles. This result is in strong contrast to how the diffraction intensity changes with suspension pH.

The DE also varies for nanoparticle suspensions with different core particle sizes and colloidal stabilities. Figure 14a shows first-order diffraction efficiencies (DE22nm and DE8nm) versus time for equal concentration suspensions with pH 6.0, 8.6, and 10.3 that were made from nanoparticle colloids with 8 nm and 22 nm diameter nanoparticle species. DLS measurements show that the 8 nm nanoparticles have significantly weaker colloidal stability than the 22 nm nanoparticles, but both types of particles remained colloidally suspended over the measurement timescale. Interestingly, for the same pH and times greater than 100 s, the 8 nm particles produce diffraction efficiencies that are ten times larger than the 22 nm particles. Further, the smaller



Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly, Fig. 14 Demonstration of the ability to distinguish between aggregated and non-aggregated nanoparticle species. (**a**) Non-aggregated 22 nm Fe_3O_4 nanoparticles and aggregated 8 nm Fe_3O_4 nanoparticles assembling versus time for 3 different pH values. Note that the aggregated particles have intensities nearly 12 times larger than the non-aggregated particles. (**c**–**d**) SEM images of assembled 22 and 8 nm gratings after 15 min. For the higher diffraction efficiency 8 nm particles, the SEM image hardly appears to suggest a diffraction grating (Used with permission and reproduced from Ref. [38])

particles display dramatically different time dependence than the other curve shapes. These 8 nm particles show "inverted" curve shapes, i.e., the rate of DE increase accelerates with time as opposed to slowing.

The 8 nm suspension appears to contain large aggregates of particles with significantly weaker colloidal stabilities. Figure 14b, c shows representative scanning electron microscopy (SEM) images of the 8 nm and 22 nm particles that are coated on the medium surface, respectively. A statistical analysis of SEM images shows that the nanoparticle aggregates attracted from the 8 nm particle suspensions are more periodically distributed than those from the 22 nm particles. In addition, the 8 nm particles show significantly larger aggregates than the 22 nm particles. This further suggests the aggregation of particles causes the inverted shape. While larger aggregates explain the large difference in DE between the two particle species, since particles scatter light with the intensity proportional to the sixth power of diameter, it is still not clear how the aggregation changes the curve shape.

While development of this real-time measurement is ongoing, it appears to have significantly

greater sensitivity to subtle changes in colloidal properties relative to what can be measured with light scattering alone. The scattered signal correlates well with what is observed with DLS, suggesting that the assembly process is very sensitive to the colloidal stability

Conclusions and Outlook

This entry has provided an overview of how to combine magnetic field-directed self-assembly, magnetic recording, and magnetic nanoparticles to manufacture patterned composite materials with 25 nm repeatability over length scales as large as 10 cm using a standard 3.5 in. hard disk drive platter. A number of measurement techniques necessary to verify the quality and repeatability of nanomanufactured structures were discussed, including a novel metrology that uses a diffraction grating template to monitor self-assembly of nanoparticles in real time. This metrology appears to be more sensitive to colloidal stability, likely because the large field gradients at the media surface aggregate the nanoparticles on the surface in a manner that depends strongly on the surface stabilization of the nanoparticles.

While not yet able to make patterns with single nanometer feature control, this approach to nanomanufacturing appears to have repeatability and precision sufficient to build materials for manipulating light at wavelengths from 400 nm to several micrometers. Importantly, since the assembly process requires primarily a disk drive and magnetic nanoparticles, there is potential for patterning photonic structures at lower cost than wafer-based processing.

It also has strong future potential. To date, the only media and heads studied are those that are commercially available. Such media and heads are highly engineered for storing, reproducing, and maintaining the fidelity of digital data. If a purpose-designed media/head combination can be engineered specifically for nanoparticle self-assembly, the potential exists to improve pattern quality such that it is limited only by the diameter of the nanoparticle feedstocks used as raw materials. Such purpose-designed recording components could allow further improvements in manufacturing repeatability and user customizability. Given the low cost and ability to rapidly reprogram the template, this nanomanufacturing technology could find application in rapid prototyping or rapid on-site assembly of custom optical patterns for fiber optics, custom diffractive optics, and spectrometers/optical sensors.

References

1. Whitesides, G.M., Grzybowski, B.: Self-assembly at all scales. Science **295**, 2418–2421 (2002)
2. Wang, S.X., Taratorin, A.M.: Magnetic Information Storage Technology. Academic, San Diego (1999)
3. Bitter, F.: On inhomogeneities in the magnetization of ferromagnetic materials. Phys. Rev. **38**, 1903–1905 (1931)
4. Bishop, K.J.M., Wilmer, C.E., Soh, S., Grzybowski, B.A.: Nanoscale forces and their uses in self-assembly. Small **5**, 1600–1630 (2009)
5. Tracy, J.B., Crawford, T.M.: Magnetic field directed self-assembly of magnetic nanoparticles. MRS Bull. **38**, 915–920 (2013)
6. Gubin, S.: Magnetic Nanoparticles. Wiley-VCH, Weinheim (2009)
7. Kumar, C.: Magnetic Nanomaterials, vol. 4. Wiley, Weinheim (2009)
8. John R.R., Frederick J.M., Robert W.C.: Foundations of Electromagnetic Theory. Addison Wesley, Reading (1960)
9. Henderson, J., Shi, S., Cakmaktepe, S., Crawford, T. M.: Pattern transfer nanomanufacturing using magnetic recording for programmed nanoparticle assembly. Nanotechnology **23**, 185304 (2012)
10. Willard, M.A., Kurihara, L.K., Carpenter, E.E., Calvin, S., Harris, V.G.: Chemically prepared magnetic nanoparticles. Int. Mater. Rev. **49**, 125–170 (2004)
11. Saville, S.L., Stone, R.C., Qi, B., Mefford, O.T.: Investigation of the stability of magnetite nanoparticles functionalized with catechol based ligands in biological media. J. Mater. Chem. **22**, 24909 (2012)
12. Saville, S.L., Woodward, R.C., House, M.J., Tokarev, A., Hammers, J., Bin, Q., Shaw, J., Saunders, M., Varsani, R.R., Pierre, T.G.S., et al.: The effect of magnetically induced linear aggregates on proton transverse relaxation rates of aqueous suspensions of polymer coated magnetic nanoparticles. Nanoscale **5**, 2152–2163 (2013)
13. Faraudo, J., Andreu, J.S., Camacho, J.: Understanding diluted dispersions of superparamagnetic particles

under strong magnetic fields: a review of concepts, theory and simulations. Soft Matter **9**, 6654 (2013)
14. Tartaj, P., del Puerto Morales, M., Veintemillas-Verdaguer, S., González-Carreno, T., Serna, C.J.: The preparation of magnetic nanoparticles for applications in biomedicine. J. Phys. D Appl. Phys. **36**, R182 (2003)
15. Roca, A.G., Costo, R., Rebolledo, A.F., Veintemillas-Verdaguer, S., Tartaj, P., González-Carreño, T., Morales, M.P., Serna, C.J.: Progress in the preparation of magnetic nanoparticles for applications in biomedicine. J. Phys. D Appl. Phys. **42**, 224002 (2009)
16. Huber, D.L.: Synthesis, properties, and applications of iron nanoparticles. Small **1**, 482–501 (2005)
17. Laurent, S., Forge, D., Port, M., Roch, A., Robic, C., Vander Elst, L., Muller, R.N.: Magnetic iron oxide nanoparticles: synthesis, stabilization, vectorization, physicochemical characterizations, and biological applications. Chem. Rev. **108**, 2064–2110 (2008)
18. Papell, S.S.: Low viscosity magnetic fluid obtained by the colloidal suspension of magnetic particles. US Patent Office, 1965
19. Rosensweig, R.E., Kaiser, R.: Study of ferromagnetic liquid, phase I. NTIS Report No. NASW-1219, NASA Report (1967)
20. Khalafalla, S., Reimers, G.: Magnetofluids and their manufacture. US Patent Office, 1973
21. Massart, R.: Preparation of aqueous magnetic liquids in alkaline and acidic media. IEEE Trans. Magn. **17**, 1247–1248 (1981)
22. Berger, P., Adelman, N., Beckman, K., Campbell, D., Ellis, A., Lisensky, G.: Preparation and properties of an aqueous ferrofluid. J. Chem. Edu. **76**, 943–948 (1999)
23. Pileni, M.P.: Nanosized particles made in colloidal assemblies. Langmuir **13**, 3266–3276 (1997)
24. López-Quintela, M.A., Rivas, J.: Chemical reactions in microemulsions: a powerful method to obtain ultrafine particles. J. Colloid Interface Sci. **158**, 446–451 (1993)
25. Wan, J., Chen, X., Wang, Z., Yang, X., Qian, Y.: A soft-template-assisted hydrothermal approach to single-crystal Fe_3O_4 nanorods. J. Cryst. Growth **276**, 571–576 (2005)
26. Daou, T.J., Pourroy, G., Begin-Colin, S., Grenèche, J. M., Ulhaq-Bouillet, C., Legaré, P., Bernhardt, P., Leuvrey, C., Rogez, G.: Hydrothermal synthesis of monodisperse magnetite nanoparticles. Chem./ Mater. **18**, 4399–4404 (2006)
27. Sun, S., Zeng, H.: Size-controlled synthesis of magnetite nanoparticles. J. Am. Chem. Soc. **124**, 8204–8205 (2002)
28. Sun, S., Zeng, H., Robinson, D.B., Raoux, S., Rice, P.M., Wang, S.X., Lis, G.: Monodisperse MFe2O4 (M = Fe, Co, Mn) nanoparticles. J. Am. Chem. Soc. **126**, 273–279 (2004)
29. Hyeon, T., Lee, S.S., Park, J., Chung, Y., Na, H.B.: Synthesis of highly crystalline and monodisperse maghemite nanocrystallites without a size-selection process. J. Am. Chem. Soc. **123**, 12798–12801 (2001)
30. Park, J.; An, K.; Hwang, Y.; Park, J.-G.; Noh, H.-J.; Kim, J.-Y.; Park, J.-H.; Hwang, N.-M.; Hyeon, T. Ultra-large-scale syntheses of monodisperse nanocrystals. Nat. Publ. Group **3**, 891–895 (2004)
31. Yu, W.W., Falkner, J.C., Yavuz, C.T., Colvin, V.L.: Synthesis of monodisperse iron oxide nanocrystals by thermal decomposition of iron carboxylate salts. Chem. Commun. 2306 (2004)
32. Strobel, R., Pratsinis, S.E.: Direct synthesis of maghemite, magnetite and Wustite nanoparticles by flame spray pyrolysis. Adv. Powder Technol. **20**, 190–194 (2009)
33. Pascal, C., Pascal, J.L., Favier, F., Elidrissi Moubtassim, M.L., Payen, C.: Electrochemical synthesis for the control of Γ-Fe_2O_3 nanoparticle size. Morphology, microstructure, and magnetic behavior. Chem. Mater. **11**, 141–147 (1999)
34. Ye, L., Terry, B., Mefford, O.T., Rinaldi, C.: All-nanoparticle concave diffraction grating fabricated by self-assembly onto magnetically-recorded templates. Opt. Express **21**, 1066 (2013)
35. Henderson, J.R., Crawford, T.M.: Repeatability of magnetic-field driven self-assembly of magnetic nanoparticles. J. Appl. Phys. **109**, 07D329 (2011)
36. Moser, A., Weller, D., Best, M.E., Doerner, M.F.: Dynamic coercivity measurements in thin film recording media using a contact write/read tester. J. Appl. Phys. **85**, 5018 (1999)
37. Leonhardt, T.D., Van de Veerdonk, R., Van Der Heijden, P., Clinton, T.W., Crawford, T.M.: Comparison of perpendicular and longitudinal magnetic recording using a contact write/read tester. IEEE Trans. Magn. **37**, 1580–1582 (2001)
38. Ye, L., Qi, B., Pearson, T., Cordeau, Y.: Real time monitoring of superparamagnetic nanoparticle self-assembly on surfaces of magnetic recording media. J. Appl. Phys. **115**, 17B513 (2014)
39. Hecht, E.: Optics, 4th edn. Addison Wesley, San Francisco (2002)

Nanomaterials

► Influence of Defects in Photocatalysis
► Integration of Nanostructures Within Microfluidic Devices
► Nanostructures for Energy

Nanomaterials and Nanoproducts

► Physicochemical Properties of Nanoparticles in Relation with Toxicity

Nanomaterials for Electrical Energy Storage Devices

Shrikant C. Nagpure and Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Accumulator; Batteries; Electric double layer capacitor; Fuel cell; Nanostructured materials

Definition

An electrical energy storage device stores the input energy and delivers it in the form of electrical energy when needed.

A material that is engineered with morphological features on the order of less than 100 nm and exhibits different characteristics compared to its bulk counterpart is defined as a nanomaterial.

Overview

Need for Energy Storage Devices

Storage of electrical energy is one of the major research focuses of this century. Energy storage devices have already helped revolutionize the electronic gadget industry, but apart from this, energy storage devices of higher capacity and power rating can prove to be very beneficial in other stationary applications such as load-leveling in existing power plants and nonstationary applications such as electric vehicles.

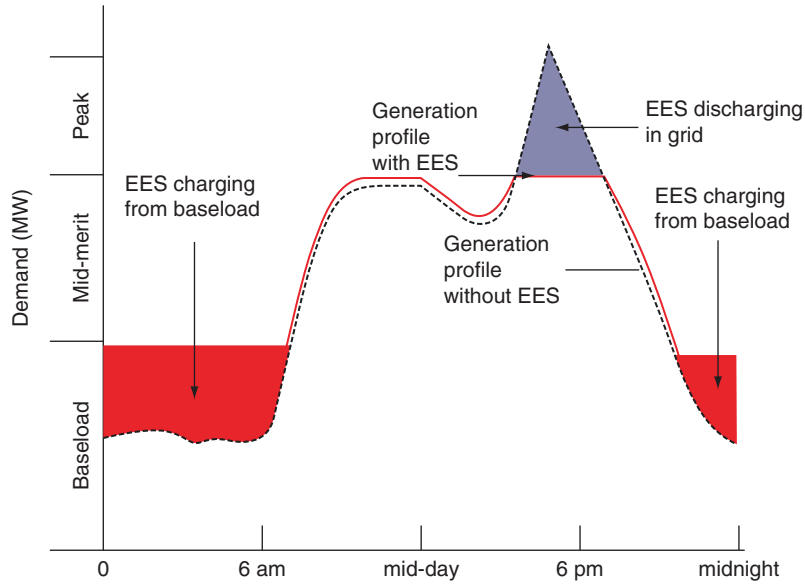
The demand for electric power from an electric grid varies based on the season, day of the week, hour, and also due to transients. Figure 1 shows a typical electrical power supply and demand curve over a 24-h period [53]. The power plants are always designed to meet the peak power demand. However, in a 24-h cycle, there are instances when the demand for the power supply is less than the peak power available. Thus, at times, power plants are forced to operate at lower output levels which prove to be uneconomical as the power plants are most efficient at their rated capacity. Energy storage devices can help balance this demand by shifting the load from peak to off-peak hours. During the off-peak hours the electrical energy can be stored in an energy storage device, and then this stored energy would be used during the peak hours. This load-leveling would help in designing power plants with lower rated capacity, thus reducing the cost of installation. The operating cost of a power plant can be reduced by operating the plants at the rated capacity for most of the time during its life. Another advantage is that the utility company can manage the demand with the existing power plants and defer the installation of the higher capacity power plants and thus reduce the overheads.

Electrical energy storage devices also play a vital role in the design and development of power plants based on other unconventional renewable energy sources, such as solar or wind energy [53]. The energy harvested in such power plants varies largely due to operating conditions such as weather, wind speed, and temperature. In this case, electrical energy storage devices can be used to store the energy during the smooth operation of the plant and overcome the intermittent nature of it. Thus, harvesting energy from unconventional sources would be more feasible.

Apart from these stationary applications energy storage devices are also necessary for certain nonstationary applications. Since about 2000, batteries have been seriously considered in the development of electric vehicles. The automotive industry has developed hybrid electric vehicles, plug-in hybrid electric vehicles, and electric vehicles, which would reduce dependence on oil and reduce greenhouse gas emissions.

Need for Nanomaterials in Energy Storage

Nanomaterials, due to their unique characteristics, are very instrumental in developing energy storage devices with high energy and power density. Energy conversion in energy storage devices takes place with a chemical reaction at the surface, charge transfer, etc. These processes occur at the nanoscale. Nanomaterials provide a large surface



Nanomaterials for Electrical Energy Storage Devices, Fig. 1 A typical electrical power demand profile for a 24-h period. During early morning hours the demand is below the capacity of the power plant while the demand is at peak in the evening around 6 PM. During off-peak hours an electrical energy storage (*EES*) device would store the energy and during the peak hours the EES will be discharged. Thus EES would level the load and the actual demand cycle will be as shown by the solid line (Adapted from Ref. [53])

area to volume ratio for such processes to occur. A high surface area permits a high contact area with the electrolyte and a larger number of reaction sites. In certain cases they enable electrode reactions to occur that cannot take place for materials composed of micrometer-sized particles.

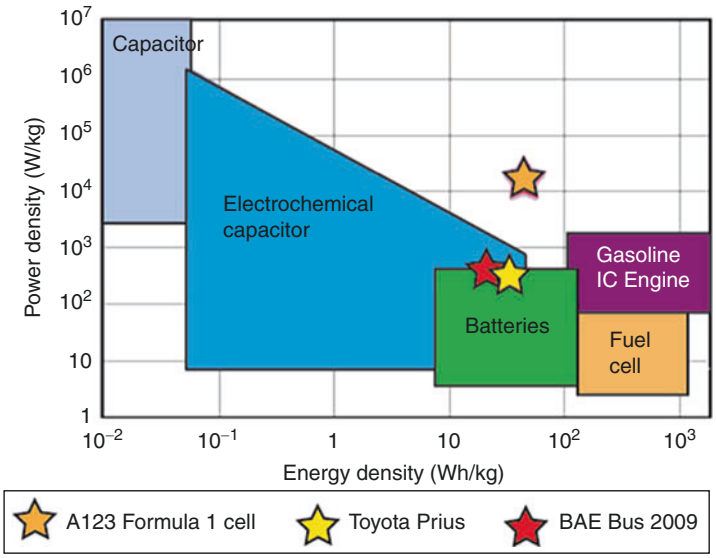
The kinetics of the energy storage device are governed by the kinetics of the diffusion process within the device elements. Nanomaterials help in reducing the diffusion length and hence improve the kinetics of the operation of an energy storage device. Thus, nanomaterials can improve the charge–discharge rates. The electron transport within the electrode is also improved by nanometer-sized particles.

Since the charge transfer reactions occur mostly at the surface of the nanometer-sized particles, the induced stress and crystal lattice breakdown is less during the repeated charge–discharge cycles. This helps in improving the cycle life of the device.

Recent Trends in Electrical Energy Storage Devices

A Ragone plot, shown in Fig. 2, is used to compare the energy density and power density of various electrical energy storage (EES) devices [53, 54]. Energy is defined as the capacity to do work, so energy density represents the capacity of the EES, i.e., the amount of energy stored in the device. Power is defined as the rate of doing work, so power density represents the rate at which the EES is charged or discharged.

As seen in the Ragone plot (Fig. 2), a fuel cell has the highest energy density but poor power density, so fuel cells are suitable for applications where high energy storage is necessary. They also tend to operate at optimum efficiency only at a constant output. They have a certain warm-up time necessary to bring them to operating conditions, and as such they are often coupled with other EES. As opposed to fuel cells, an electrochemical capacitor has high power density but low energy density, so they are suitable for applications where rapid charge/discharge in necessary. However the output of the electrochemical capacitor depends largely on its state of charge, and thus it fails to deliver energy at constant voltage [53]. Batteries lie in between the fuel cell and the electrochemical capacitor on the Ragone plot. They have higher power density than fuel cells and do not require any warm-up time. They also have higher energy density than an electrochemical capacitor and have a flat voltage curve during discharge. Thus, batteries have been actively pursued as the EES for immediate future applications.



Nanomaterials for Electrical Energy Storage Devices, Fig. 2 Ragone chart showing the comparison of the power density and energy density for batteries, capacitors, and fuel cells and internal combustion engine. Energy density represents the capacity while power density represents the rate of the charging/discharging of EES device (Adapted from Ref. [54])

Li-Ion Batteries

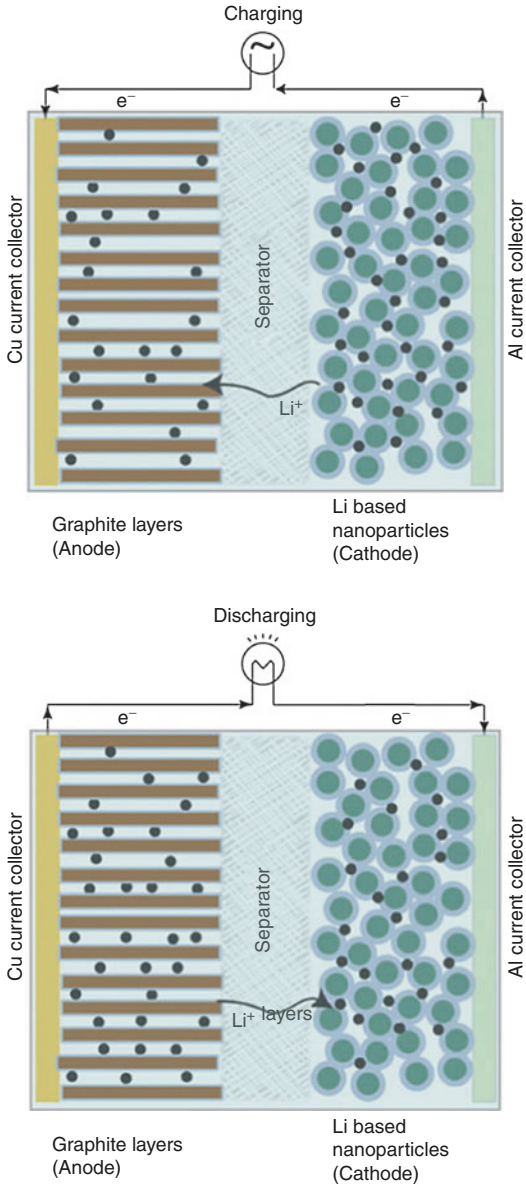
In 1912, G. N. Lewis started developing lithium batteries. Lithium (atomic number 3, group 1, period 2) metal was used as an anode in these batteries. Lithium is an alkali metal, and being the lightest of the metals with the greatest electrochemical potential, it has the largest energy density for its weight. However, lithium is very unstable and highly reactive, as it has only one valence electron, which makes it unsafe as the electrode material. As such, research shifted from a lithium battery to a lithium-ion battery, which is a much safer option. In 1991, Sony commercialized Li-ion batteries and started using them in its electronic products. Since then the importance and market for Li-ion batteries has been ever increasing.

Figure 3 shows a schematic of a typical lithium-ion battery. In a typical cylindrical lithium-ion battery separator, anode, cathode and separator are stacked alternately then rolled and packed in a cylindrical can filled with electrolyte. The anode is the electropositive electrode which holds the lithium in the charged state. The cathode is the electronegative electrode which serves as the source for the lithium [3]. The electrodes are immersed in an electrolyte solution which is mostly a lithium hexafluorophosphate ($LiPF_6$) salt dissolved in organic solvent. A separator is used in between the anode and the cathode to avoid short-circuiting between the electrodes. The electrochemical reaction in the case of the lithium-ion battery is based on the intercalation of Li^+ ions in the electrodes. During charging the Li^+ ions are transferred from the cathode and intercalated into the anode. During discharging, a deintercalation reaction occurs, and Li^+ ions are transferred from the anode to the cathode while the electrons flow through the external circuit. The cell voltage is determined by the energy of the chemical reaction, such as given by Eq. 1, occurring in the cell [37].

$$Li_{1-x}FePO_4 + Li_xC_6 \rightleftarrows Li_{1-x}C_6 + Li_xF_ePO_4 \quad (1)$$

The anode and cathode electrodes used in Li-ion batteries are complex composites with active material bonded to the current collector with polymeric binders. The anode material is bonded to the copper current collector and the cathode material is bonded onto an aluminum current collector. The cathode material is also coated with carbon black to improve the electronic conductivity.

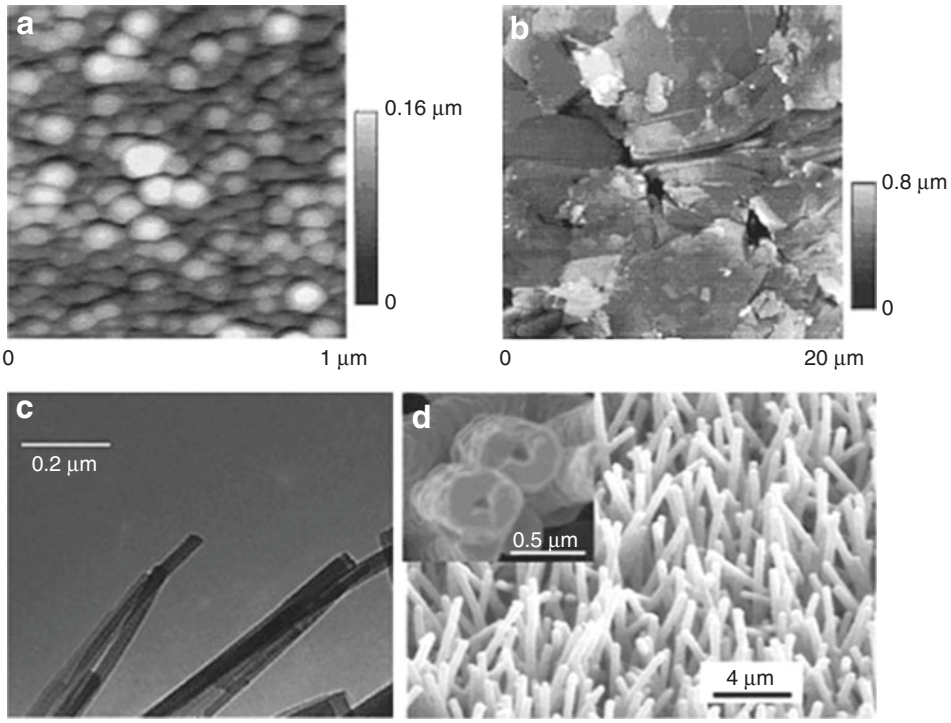
Anode Lithium is an ideal anode material for Li-ion batteries, but because of the plating out of lithium during the charge/discharge cycle and subsequent formation of dendrites, there is a risk of the cell being short-circuited. Instead,



Nanomaterials for Electrical Energy Storage Devices, Fig. 3 Schematic showing operation of a Li-ion battery. In a cylindrical battery separator, anode, cathode and separator are stacked alternately then rolled and packed in a cylindrical can filled with electrolyte solvent containing Li salt. During charging Li^+ ions are inserted in the anode structure by virtue of intercalation and during discharging, these Li^+ ions are removed from the anode structure by virtue of deintercalation and transferred to the cathode

carbonaceous materials are preferred for anodes in most commercially available Li-ion batteries [22]. Studies have been conducted on graphite [6], C-C composite, mesocarbon microbeads (MCMB) [18], carbon nanotubes [13, 52], and carbon films. Evaporated carbon film and a natural graphite anode is shown in Fig. 4a, b respectively [22]. These anodes have a very good rate of lithium insertion/removal and thus improve the charge/discharge rate (power rating) of the battery, but in any Li-ion cell with a carbonaceous anode a solid electrolyte interphase (SEI) is formed during the cycling of the cell [21, 22, 43]. The SEI is formed from the electrolyte decomposition products [50]. This SEI layer prevents the graphite surface from further exfoliation and also prevents further reduction of the electrolyte and consumption of active lithium. In the case of nanoparticulate graphite, the consumption of active lithium would be still higher, and the excessive charge developed between the graphite surface and the SEI would result in a loss of overall cell voltage. Also, it is important to note that the lithium is intercalated into graphite at potentials less than 100 mV versus Li^+/Li. There is always a risk of lithium depositing on the graphite surface resulting in dendrite formation and fatal short-circuiting of the cell [6]. Carbon nanotubes exhibit twice the lithium storage compared with graphite but have similar problems of surface-layer formation and safety [6]. To overcome some disadvantages of graphite, the research has been focused on the titanium oxide.

The nanoparticles of non-toxic $Li_4Ti_5O_{12}$ ($Li[Li_{1/3}Ti_{5/3}]O_4$) with spinel structure have high diffusion rates due to short diffusion distance. In this spinel structured oxide, lithium can be intercalated to give a composition range of $Li_{4+x}Ti_5O_{12}$, $0 < x < 3$ [14, 20], but the capacity of the $Li_4Ti_5O_{12}$ is only 150 mAh/g compared to 300 mAh/g capacity of graphite. Instead nanotubes/nanowires composed of TiO_2-(B), as shown in Fig. 4c, are actively studied [6]. This fifth polymorph of titanium dioxide has all the advantages of $Li_4Ti_5O_{12}$ but also has a capacity of 300 mAh/g. TiO_2-(B) is also safer as it eliminates the risk of lithium plating on the electrode and can be cycled with higher rates as in $Li_4Ti_5O_{12}$. The TiO_2-(B) nanowires exhibit better performance than TiO_2-(B) nanoparticles as they have higher reversibility of intercalation



Nanomaterials for Electrical Energy Storage Devices, Fig. 4 Different types of anode materials used in the Li-ion battery. (**a**) AFM image of a carbon film on glass prepared by electron-beam evaporation technique [22]; (**b**) AFM image of a natural graphite powder pressed together in a hydraulic press [22]; (**c**) TEM image of TiO_2-B nanowires [6]; and (**d**) SEM image of mesoporous Co_3O_4 nanowire arrays grown on Ti foil. The *inset* shows the open tips of the nanowire [28]

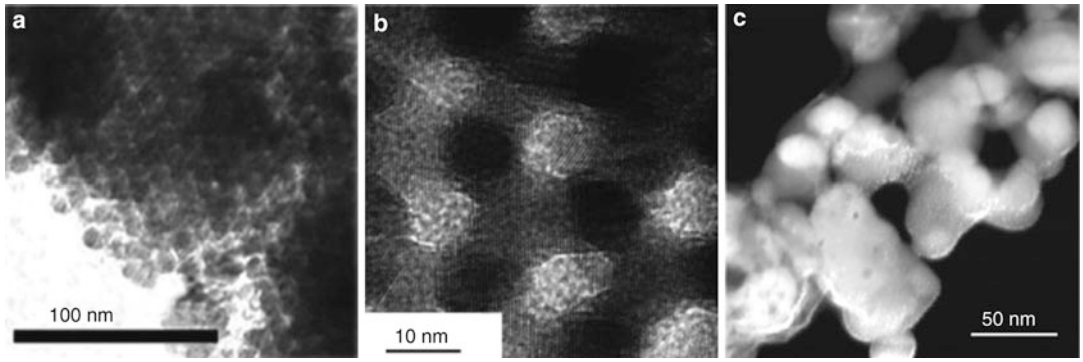
N

(>99.9 % per cycle, after the first cycle). Other similar oxide nanowire electrode materials of interest include Sn, Co, and V oxides [6]. Lithium metal alloys such as $Li_{4.4}Sn$ (993 mAh/g) and $Li_{4.4}Si$ (4,200 mAh/g) with high specific capacity are also being considered as anode materials in Li-ion batteries. The major drawback of these materials is the large volume change during the cycling of the cell to accommodate the large amounts of lithium [6].

Lithium's ability to react reversibly with the binary transition metal oxides is exploited in developing anode materials such as mesoporous Co_3O_4, shown in Fig. 4d [28]. This reaction is not limited by the one or two lithium atoms per host as in intercalation, and so such materials can have a higher capacity than conventional anode materials. The poor kinetics of these types of materials can be improved by using nanowires or nanoparticulates as the anode material.

Cathode The potential area of greatest opportunity for nanomaterials is the cathode of the Li-ion battery [53]. The most common cathode materials are layered oxides $LiMO_2$, the spinels $Li[M_2]O_4$ and olivines $LiMPO_4$ where M is a transition metal atom [16]. The nanoparticulates of these materials reduce the diffusion length, provide a higher electrolyte/electrode contact area, and thus improve the charge/discharge kinetics of the Li-ion batteries. The nanoparticles of these lithium compounds are made by grinding, synthesis from solution, or by solgel approaches [6].

Initially a $LiCoO_2$, as shown in Fig. 5a, was developed as the cathode material for Li-ion batteries [6, 16], but due to expensive and toxic cobalt they were later replaced by oxides such as $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$ $LiNi_{0.8}Co_{0.2}O_2$, $Li_1-xNi_1-yCoyO_2$, and $LiMn_{0.5}Ni_{0.5}O_2$ in commercial batteries [16, 23, 24]. These oxides have a



Nanomaterials for Electrical Energy Storage Devices, Fig. 5 Different types of cathode materials used in the Li-ion battery. (**a**) TEM image of mesoporous $LiCoO_2$ transition metal oxide [6], (**b**) TEM image of nanostructured $LiMn_2O_4$ spinel [6], and (**c**) TEM image of olivine structured $LiFePO_4$ nanoparticles embedded in carbon matrix [38]

high operating voltage, in the range of 2.75–4.3 V [16]. They have very high energy density and power density but they lack the necessary structural stability for deep discharge cycles, during which the host oxide structure collapses upon removal of more than 50 % of the Li. Li spinel compounds, such as $LiMn_2O_4$ shown in Fig. 5b, has good structural stability [6]. They also have higher operating voltage and high charge/discharge rate but low energy density. The most recent among these lithium compounds has been the olivine structured $LiFePO_4$ shown in Fig. 5c [38]. It has a lower operating voltage of ~3.3 V but demonstrates higher power and energy density along with a good structural stability.

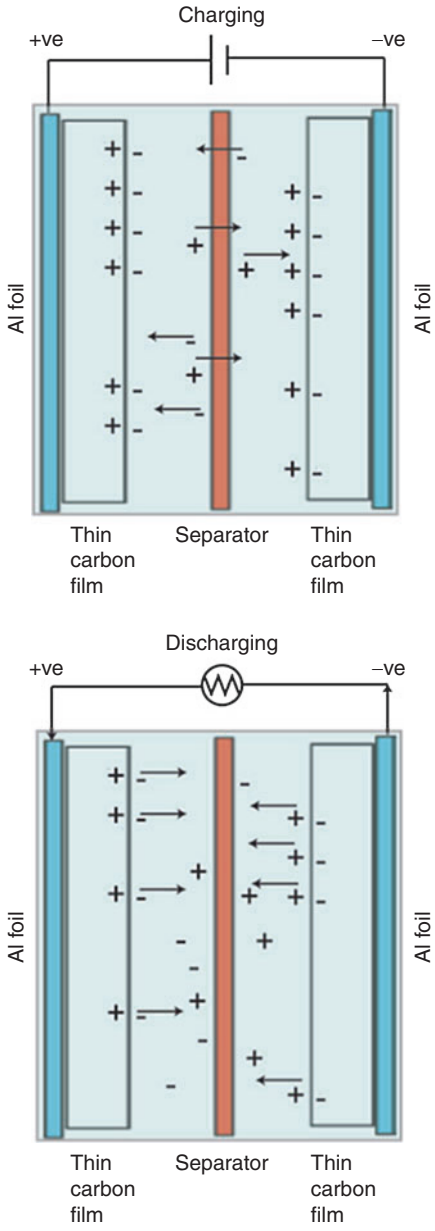
Electrochemical Capacitors

Capacitors are commonly used in power and consumer electronic circuits. However they have very low capacitance in the range of few millifarads which make them unsuitable for large storage devices. In 1957, Becker filed the first patent describing the concept of electrochemical capacitors (EC) which have capacitance in the range of a few hundred farads [4, 46]. It was then commercialized by Nippon Electric Company (NEC), Japan, under a license from Standard Oil of Ohio (SOHIO) [46]. EC are also called supercapacitors or ultracapacitors. The two different types of EC are the electric double layer capacitor (EDLC) and the pseudocapacitor.

A schematic showing the operation of an EDLC is shown in Fig. 6. EDLC consists of two electrodes with a thin coating of active material having very high surface area. These electrodes are immersed in an electrolyte, mostly sulfuric acid or acetonitrile, containing both positive and negative ions. When the capacitor is connected to the external power source, the electric field attracts the opposite charged ions toward the surface of the electrode. The charge separation occurs near the surface of the electrode, and energy is stored in the EDLC. During discharging, as the electrons flow through the load the ions disperse in the electrolyte solution and the electrode-electrolyte interface fades [25, 35]. A separator is used to prevent the tunneling of the electrons between the electrodes through the electrolyte. The other type of EC capacitor, the pseudocapacitor, has essentially the properties of the EDLC as well as a battery. Thus a charge is stored by the surface as well as subsurface activities. In EDLC there is no physical or chemical change in the electrode, and the ion adsorption process is highly reversible. Thus, theoretically EDLC can be charged/discharged without any damage to its components [34].

The charge separation at the electrode-electrolyte interface is given by Helmholtz's double layer capacitance C:

$$C = \frac{\varepsilon_r \varepsilon_o A}{d} \tag{2}$$



Nanomaterials for Electrical Energy Storage Devices, Fig. 6 Schematic showing operation of an electrochemical capacitor (*EC*). During charging the ions in the electrolyte are attracted to the nanoporous electrodes with opposite charge. This creates the charge separation at the electrode-electrolyte interface and the energy is stored in EC. During discharging, as the electrons flow through the load the ions disperse in the electrolyte solution and the electrode-electrolyte interface fades

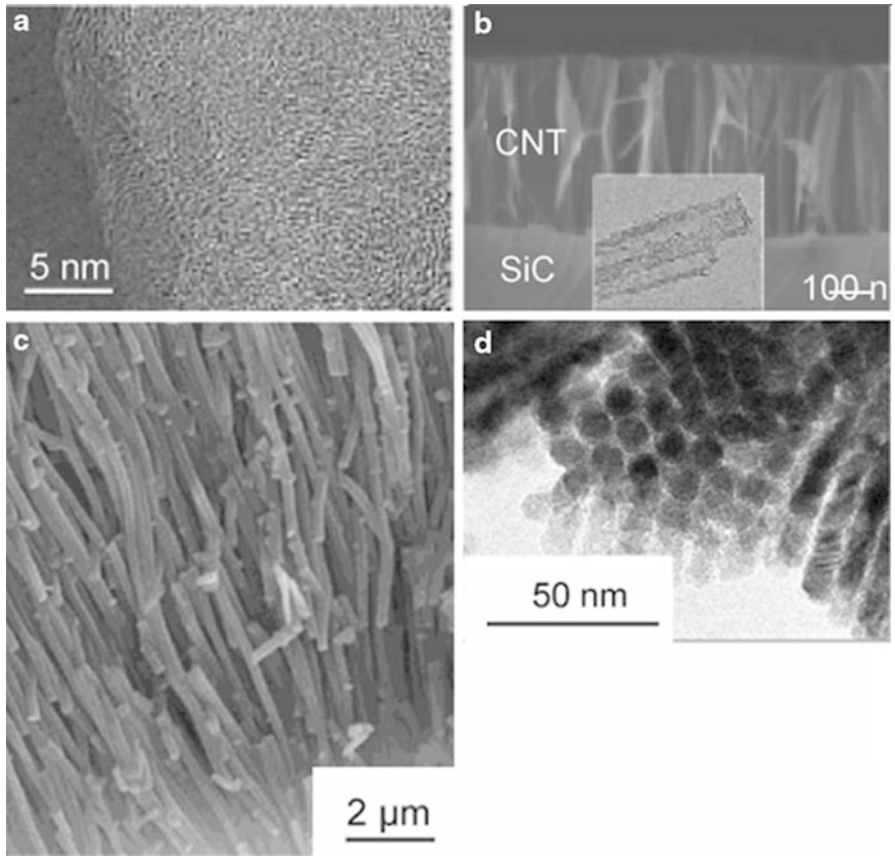
where ε_r is the electrolyte dielectric constant, ε_o is the dielectric constant of the vacuum, d is the effective thickness of the double layer (charge separation distance), and A is the electrode surface area [46].

Many materials used as an anode in Li-ion batteries are generally considered suitable for EC. In general, to achieve high capacitance between the charged double layer it is necessary to have a high specific surface area (SSA) of the double layer material and electrodes with high electronic conductivity. Carbonaceous material has high SSA, high electronic conductivity, electrochemical stability, and structural stability and is suitable for the EC. The various carbon-based materials such as carbon fabrics, carbide-derived carbons, carbon nanotubes, onion-like carbon, and nanohorns are studied as potential EC electrode materials. Disordered microporous carbon (SiC-derived carbon), shown in Fig. 7a, and carbon nanotubes, shown in Fig. 7c, are among the various carbonaceous ideal EC electrode materials due to their high SSA [46].

Several other materials, such as V_2O_5 nanowires shown in Fig. 7b, and ruthenium oxides are actively being studied as EC electrode materials due to their broad range of morphologies [12, 49]. Hydrated ruthenium oxide has the highest capacity, 200 Wh/kg [53]. The other promising EC material is nickel oxide due to its high specific capacitance value compared to carbon materials [1, 10, 27, 33]. Thin films of nickel oxide can be easily synthesized through electrochemical route or solgel method. The nanostructured NiO thin films shown in Fig. 7d can be produced with specific microstructure, high surface area, and reduced crystallite size [39, 51].

Fuel Cell

A fuel cell converts the energy stored in the source fuel to usable electrical energy. It is different from other EES in that the reactants flow through the cell rather than being sealed within the cell. There are several types of fuel cells, but the most common are the solid oxide fuel cell (SOFC) and the proton exchange membrane fuel cell (PEMFC) [7]. A schematic of SOFC is shown in Fig. 8a



Nanomaterials for Electrical Energy Storage Devices, Fig. 7 Different types of electrode materials used in EC. (**a**) TEM image of a disordered microporous carbon (SiC-derived carbon) [46], (**b**) SEM image of an array of carbon nanotubes (CNT) on Si-C (*inset* shows the TEM image of the same CNT) [46], (**c**) TEM image of V_20_5 nanowires [49], and (**d**) TEM image of ordered mesoporous NiO [51]

and schematic of PEMFC is shown in Fig. 8b. SOFC is the most fuel flexible among all fuel cells. Other than pure hydrogen, it can run on hydrocarbon, methanol, or even gasoline, but in the case of PEMFC pure hydrogen is the only choice of fuel. Essentially both fuel cells have similar principles of operation except that in SOFC, O^{2-} ions are conducted through the solid oxide membrane (SOM) while in PEMFC H^+ ions are conducted through the proton exchange membrane (PEM) [17, 19].

In SOFC, the oxygen is oxidized to O^{2-} as per the following half-cell reaction at the cathode:

$$\frac{1}{2}O_2 + 2e^- \rightarrow O^{2-} \tag{3}$$

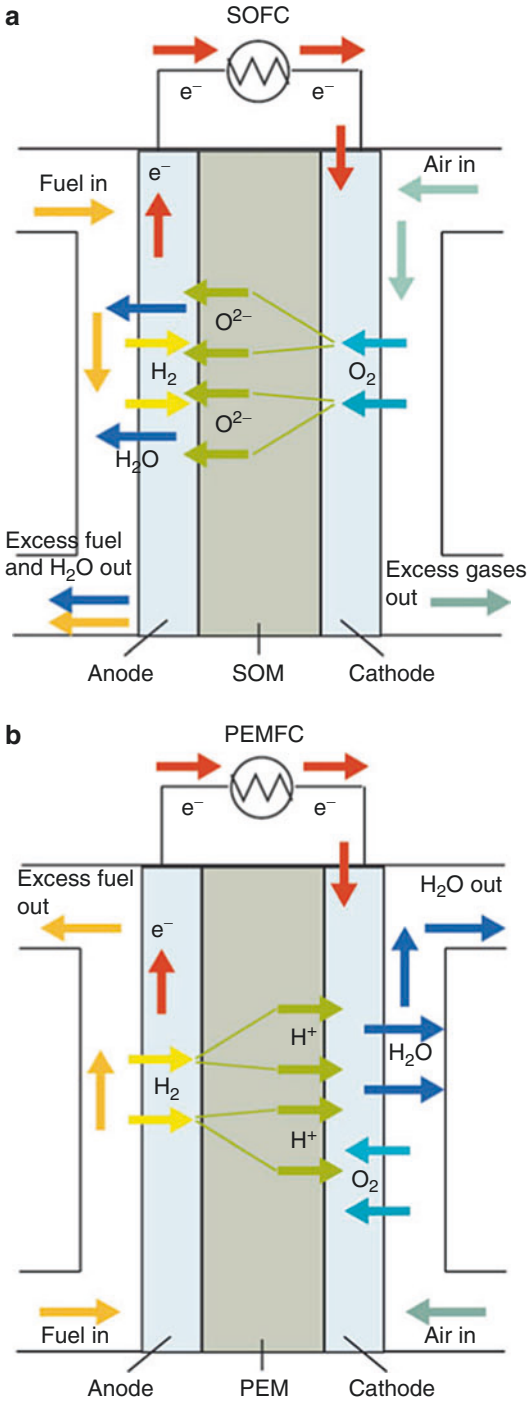
These O^{2-} ions are conducted through the solid oxide membrane to the anode where they are reduced by hydrogen with water as a by-product. The anode half-cell reaction is:

$$H_2 + O^{2-} \rightarrow H_2O + 2e^- \tag{4}$$

In the case of PEMFC, hydrogen is reduced at the anode, and the H^+ ions are formed as per the following half-cell reaction:

$$H_2 + 2H^+ + 2e^- \tag{5}$$

These H^+ ions are conducted through the proton exchange membrane, and they are combined with oxygen at the cathode to form water as the



Nanomaterials for Electrical Energy Storage Devices, Fig. 8 (**a**) Schematic showing operation of a solid oxide fuel cell (*SOFC*). Oxygen is reduced at cathode and the resulting O^{2-} ions are conducted through the solid oxide electrolyte membrane (*SOM*) to the anode where they are oxidized with the hydrogen to form water. (**b**) Schematic showing operation of a proton exchange membrane fuel cell (*PEMFC*). Hydrogen is oxidized at the anode and the resulting H^{+} ions are conducted through the proton exchange membrane (*PEM*) to the cathode where they are reduced with O to form water

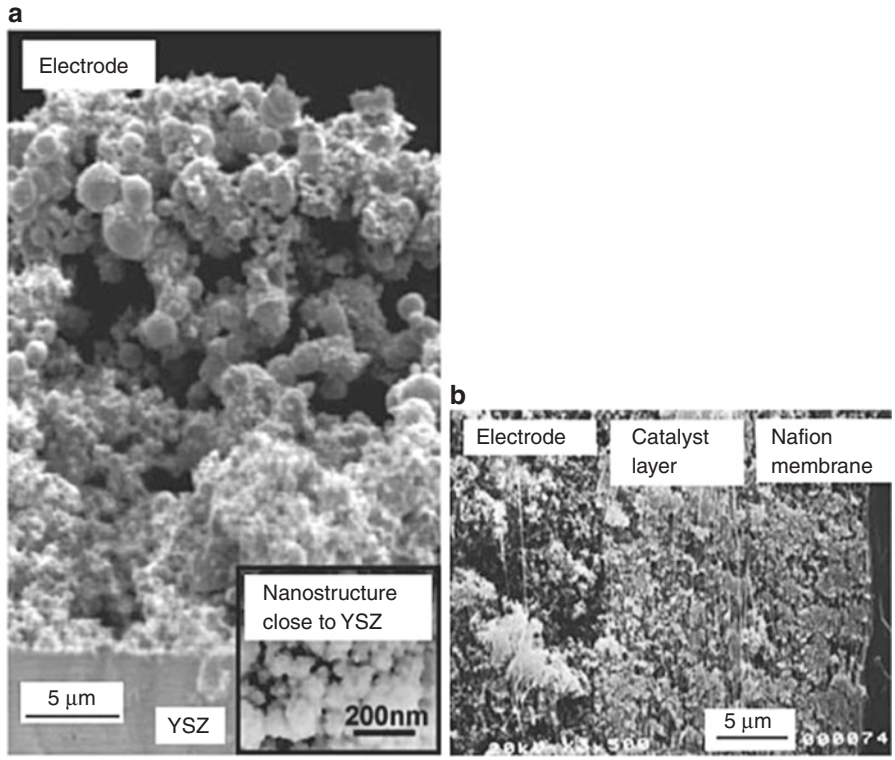
by-product of the reactions. The half-cell reaction at the cathode is:

$$\frac{1}{2}O_2 + 2H^+ + 2e^- \rightarrow H_2O \qquad (6)$$

Due to the high operating temperatures, in the range of 500–1,000 °C, SOFC is more cost effective as it does not need an expensive catalyst such as Pt, used in PEMFC, but the disadvantage of high operating temperatures is that the components of SOFC should demonstrate higher thermal and thermomechanical stability. In the case of the PEMFC the operating temperature is lower (70–100 °C), but the proton exchange membrane needs to be hydrated to optimum levels to maintain the proton conductivity [17]. Too much wetting of the membrane causes flooding of the cell. If the membrane is dry then it can also lead to cracking of the membrane and overheating of the cell [17].

A good electrolyte membrane for a fuel cell should have high ionic conductivity, low electronic conductivity, should be stable in both oxidizing and reducing environments, good mechanical properties, and good thermal stability. The electrodes should have high electronic conductivity and high (electro)catalytic activity. The anode catalyzes the oxidation of the fuel components by the ions transported through the electrolyte membrane and the cathode catalyzes the reduction of oxygen [19].

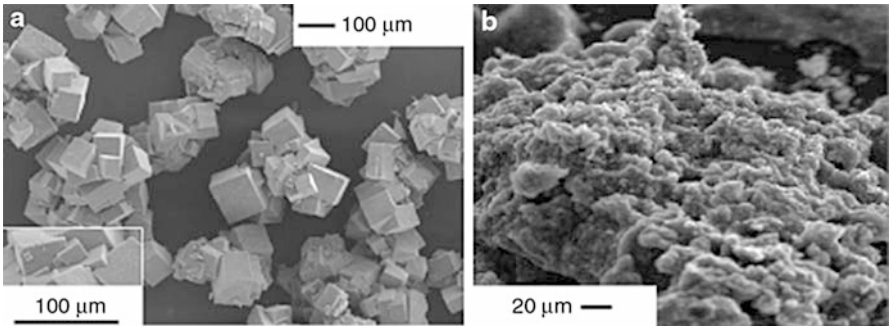
SOFC An extensive review of SOFC has been presented by Minh [36], Singhal [47], and Ormerod [41]. The most widely accepted electrolyte for SOFC is yttria-stabilized zirconia (YSZ) [39]. Figure 9a shows a complex structure of the composite cathode supported by a 240 μm thick dense YSZ electrolyte. The cathode fabricated by combustion chemical vapor deposition (CVD) consists of three porous layer structures, with about 5 μm thick 60 wt% lanthanum strontium



Nanomaterials for Electrical Energy Storage Devices, Fig. 9 Different types of electrode-electrolyte materials used in SOFC and PEMFC. (**a**) SEM image of strontium doped lanthanum manganite (*LSM*)–strontium doped lanthanum cobaltite (*LSC*)–gadolinia doped ceria (*GDC*) composite cathode electrode fabricated on yttria-stabilized zirconia (*YSZ*) electrolyte membrane used in SOFC. The *inset* shows the high resolution imaged of two nanostructured bottom layers close to the YSZ (Adapted from Ref. [30]). (**b**) PEM membrane-electrode assembly used in PEMFC. SEM image of the cross-section of the Nafion® 117 membrane and electrode impregnated with the Nafion® 5 % solution [5]

manganite (LSM)-40 wt% gadolinia doped ceria (GDC) fine agglomerates (0.5 µm diameter) at the bottom (close to YSZ electrolyte), followed by 5 µm thick 30 wt% LSM-30 wt% lanthanum strontium cobaltite (LSC)-40 wt% GDC fine agglomerates (0.5 µm diameter), and 15 µm thick 60 wt% LSC- 40 wt% GDC coarse agglomerates (2–3 µm diameter) on the top (air side) [30]. The inset shows the nanostructure present in the bottom two layers. This nanostructure offers extremely high surface area for oxygen reduction. The anode is mostly a ceramic metal composite (cermet) comprised of Ni + YSZ [17]. The other electrolyte systems investigated for SOFC are strontium, magnesium-doped lanthanum gallate (LSGM), and gadolinium- or samarium-doped ceria (CGO or CSO).

PEMFC Operation of PEMFC and its materials has been covered extensively in Costamagna and Srinivas [8, 9]. In the case of PEMFC the most commonly used electrolyte is Nafion® (sulfonated polymer), manufactured by DuPont. Nafion is comprised of perfluorinated backbones, which provide chemical stability, and of sulfonated side-groups which aggregate and facilitate hydration, which is necessary for proton conductivity. Figure 9b shows a typical membrane-electrode assembly of PEMFC [5]. In this assembly, the Nafion® membrane is hot-pressed onto the electrode impregnated with Nafion® 5 % solution by



Nanomaterials for Electrical Energy Storage Devices, Fig. 10 Different types of materials used to store hydrogen. (**a**) SEM image of metal-organic framework – 5 (*MOF-5*) which stores hydrogen by physisorption on its surface [42], and (**b**) SEM image of the complex borohydride ($LiBH_4$ + 1/2 $ZnCl_2$) doped with nano-Ni which stores hydrogen by chemisorptions

brushing technique. Other sulfonated polymers have been discussed by Kreuer [26]. For hydrogen/air fuel cells, Pt nanoparticles supported on carbon are utilized for both the anode and cathode.

Hydrogen storage Hydrogen is the optimal fuel for all types of fuel cells. Storage of hydrogen on-board a transportation vehicle continues to be one of the most challenging topics for the commercialization of hydrogen-fueled vehicles. Hydrogen has very high energy density by mass but very poor volumetric energy density. The higher volumetric energy density can be achieved for smaller but not lighter containers by compressing pure hydrogen to higher pressures, but compressing the pure hydrogen requires a high-powered compressor. Similarly, higher volumetric density can be achieved by liquid hydrogen, but it requires cryogenic storage systems as liquid hydrogen boils at 20.268°K, increasing the cost of insulation. To overcome these difficulties, storage of hydrogen by adsorption (physisorption) on materials with high specific surface area or chemical bonding (chemisorption) in covalent and ionic compounds is becoming more attractive [55]. Nanostructured materials improve the thermodynamics and kinetics of hydrogen adsorption and dissociation by decreasing the required diffusion length and increasing the diffusion rate [40]. Carbonaceous nanomaterials such as activated carbon, graphene nanofibers, carbon nanotubes, or a group of synthetic materials called as metal oxide frameworks (MOF) are considered for storage by physisorption [29]. The hydrogen is adsorbed onto the surface by weak van der Walls interaction between the hydrogen molecules and the surface of the adsorbent material [15, 39]. MOF-5, shown in Fig. 10a, with a formula unit of $Zn_4O(C_{24}H_{12})_{12}$ or Zn_4O (benzene decarboxylate)$_3$, is one such extensively studied MOF material [42]. A hydrogen storage capacity of 4 wt% at 78 K (−195 °C) and 1 wt% at room temperature has been reported for different MOFs [40]. However, due to the highly porous nature of these materials, volumetric capacity may still be a significant issue [40]. The other frequently studied physisorption-based materials include zeolites [11], clathrate hydrates [45, 48], and hollow glass microspheres (HGM) [2]. Usually hydrogen storage by physisorption-based materials, such as MOF, HGM, CNT, etc., has poor capacity at normal operating conditions [29]. Figure 10b shows another class of materials – a complex hydride, borohydride ($LiBH_4$ + 1/2 $ZnCl_2$) doped with nano-Ni which stores hydrogen by chemisorptions – used in storage of hydrogen [40]. These hydrides along with other chemisorption materials, e.g., magnesium-based hydride, metal nitrides, and amides imides, are very expensive, and in some cases the hydrogen absorption/desorption process is irreversible [29, 31, 32, 44]. The huge capacity values of various materials published in respected technical and

scientific reports should be accepted with caution as many values were wrongfully reported due to gross experimental measurement errors [29].

Challenges

There is a profound effect of spatial boundaries and contribution of surfaces due to small particle size on many of the properties of materials; this challenges researchers to develop new theories or at least adapt and develop theories that have been established for bulk materials [46]. Understanding how materials store and transport charges at electrode-electrolyte interfaces is critically important and will require a fundamental understanding of charge transfer and transport mechanisms. High electrolyte/electrode surface areas may lead to more significant side reactions with the electrolyte, and more difficulty maintaining interparticle contact. There is also a challenge to synthesize nanoparticles having controlled dimensions with a cost effective process. The characterization of functional and structural properties of these nanomaterials also challenges the scientific community in modifying the existing tools and/or developing the new tools.

Cross-References

► Aging

References

1. Ahlberg, E., Palmquist, U., Simic, N., Sjovall, R.: Capacity loss in Ni–Cd pocket plate batteries. The origin of the second voltage plateau. J. Power Sources **85**(2), 245–253 (2000)
2. Akunets, A.A., Basov, N.G., Bushuev, V.S., Dorogotovtsev, V.M., Gromov, A.I., Isakov, A.I., Kovylnikov, V.N., Merkul'ev, Y.A., Nikitenko, A.I., Tolokonnikov, S.M.: Super-high-strength microballoons for hydrogen storage. Int. J. Hydrog. Energy **19**, 697–700 (1994)
3. Anonymous: Basic Research Needs for Electrical Energy Storage. Office of Basic Energy Sciences, U.S. Department of Energy, Washington, DC (2007)
4. Becker, H.E.: Low voltage electrolytic capacitor. US Patent 2,800,616 (1957)
5. Broka, K., Ekdunge, P.: Oxygen and hydrogen permeation properties and water uptake of Nafion® 117 membrane and recast film for PEM fuel cell. J. Appl. Electrochem. **27**, 117–123 (1997)
6. Bruce, P.G., Scrosati, B., Tarascon, J.-M.: Nanomaterials for rechargeable lithium batteries. Angew. Chem. Int. Ed. **47**, 2930–2946 (2008)
7. Carrette, L., Friedrich, K.A., Stimming, U.: Fuel cells: principles, types, fuels and applications. ChemPhysChem **1**, 162–193 (2000)
8. Costamagna, P., Srinivasan, S.: Quantum jumps in the PEMFC science and technology from the 1960s to the year 2000. Part I. Fundamental scientific aspects. J. Power Sources **102**, 242–252 (2001)
9. Costamagna, P., Srinivasan, S.: Quantum jumps in the PEMFC science and technology from the 1960s to the year 2000. Part II. Engineering, technology development and application aspects. J. Power Sources **102**, 253–269 (2001)
10. Deabate, S., Fourgeot, F., Henn, F.: Electrochemical behaviour of the β(II)-Ni(OH)2/β(III)-NiOOH redox couple upon potentiodynamic cycling conditions. Electrochim. Acta **51**(25), 5430–5437 (2006)
11. Dong, J., Wang, X., Xu, H., Zhao, Q., Li, J.: Hydrogen storage in several microporous zeolites. Int. J. Hydrog. Energy **32**, 4998–5004 (2007)
12. Drezen, T., Kwon, N.-H., Bowenb, P., Teerlinck, I., Isono, M., Exnar, I.: Effect of particle size on LiMnPO4 cathodes. J. Power Sources **174**, 949–953 (2007)
13. Fan, Q., Chupas, P.J., Whittingham, M.S.: Characterization of amorphous and crystalline tin-cobalt anodes. Electrochem. Solid-State Lett. **10**, A274–A278 (2007)
14. Ferg, E., Gummow, R.J., de Kock, A., Thackeray, M. M.: Spinel anodes for lithium-ion batteries. J. Electrochem. Soc. **141**, L147–L150 (1994)
15. Felderhoff, M., Weidenthaler, C., Helmolt, R., von Eberle, U.: Hydrogen storage: the remaining scientific and technological challenges. Phys. Chem. Chem. Phys. **9**, 2643–2653 (2007)
16. Goodenough, J.B.: Cathode materials: a personal perspective. J. Power Sources **174**, 996–1000 (2007)
17. Haile, S.M.: Fuel cell materials and components. Acta Mater. **51**, 5981–6000 (2003)
18. Hossain, S., Kim, Y.-K., Saleh, Y., Loutfy, R.: Comparative studies of MCMB and C-C composite as anodes for lithium-ion battery systems. J. Power Sources **114**, 264–276 (2003)
19. Jacobson, A.J.: Materials for solid oxide fuel cells. Chem. Mater. **22**, 660–674 (2010)
20. Jansen, A.N., Kahalan, A.J., Kepler, K.D., Nelson, P. A., Amine, K., Dees, D.W., Vissers, D.R., Thackeray, M.M.: Development of a high-power lithium-ion battery. J. Power Sources **81**, 902–905 (1999)
21. Kanno, R., Kawamoto, Y., Takeda, Y., Ohashi, S., Imanishi, N., Yamamoto, O.: Carbon fiber as a negative electrode in lithium secondary cells. J. Electrochem. Soc. **139**, 3397–3404 (1992)
22. Kong, F., Kostecki, R., Nadeau, G., Song, X., Zaghib, K., Kinoshita, K., McLarnon, F.: In-situ studies of SEI formation. J. Power Sources **97–98**, 58–66 (2001)

23. Kostecki, R., McLarnon, F.: Degradation of $LiNi_{0.8}Co_{0.2}O_2$ cathode surfaces in highpower lithium-ion batteries. Electrochem. Solid-State Lett. **5**, A164–A166 (2002)
24. Kostecki, R., McLarnon, F.: Local probe studies of degradation of composite $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$ cathodes in high-power lithium-ion batteries. Electrochem. Solid-State Lett. **7**, A380–A383 (2004)
25. Kötz, R., Carlen, M.: Principles and application of electrochemical capacitors. Electrochim. Acta **45**, 2483–2498 (2000)
26. Kreuer, K.D.: On the development of proton conducting polymer membranes for hydrogen and methanol fuel cells. J. Membr. Sci. **185**, 29–39 (2001)
27. Léger, C., Tessier, C., Ménétrier, M., Denage, C., Delmas, C.: Investigation of the second discharge plateau of the β(*III*)-NiOOH/β (*II*)-$Ni(OH)_2$ system. J. Electrochem. Soc. **146**, 924–932 (1999)
28. Li, Y., Tan, B., Wu, Y.: Mesoporous Co_3O_4 nanowire arrays for lithium ion batteries with high capacity and rate capability. Nano Lett. **8**, 265–270 (2008)
29. Lim, K.L., Kazemian, H., Yaakob, Z., Daud, W.R.W.: Solid-state materials and methods for hydrogen storage: a critical review. Chem. Eng. Technol. **33**, 213–226 (2010)
30. Liu, Y., Compson, C., Liu, M.: Nanostructured and functionally graded cathodes for intermediate temperature solid oxide fuel cells. J. Power Sources **138**, 194–198 (2004)
31. Liu, Y., Xiong, Z., Hu, J., Wu, G., Chen, P., Murata, K., Sakata, K.: Hydrogen absorption/desorption behaviors over a quaternary Mg–Ca–Li–N–H system. J. Power Sources **159**, 135–138 (2006)
32. Luo, K., Liu, Y., Wang, F., Gao, M.I., Pan, H.: Hydrogen storage in a Li–Al–N ternary system. Int. J. Hydrog. Energy **34**, 8101–8107 (2009)
33. Mancier, V., Willmann, P., Metrot, A.: A semi theoretical approach of the second plateau appearing during the discharge of aged nickel oxyhydroxide electrodes. J. Power Sources **85**, 181–185 (2000)
34. Miller, J.R., Burke, A.F.: Electrochemical capacitors: challenges and opportunities for real-world applications. Electrochem. Soc. Interface. **17**, 53–57 (2008)
35. Miller, J.R., Simon, P.: Electrochemical capacitors for energy management. Science **321**, 651–652 (2008)
36. Minh, N.Q.: Ceramic fuel cells. J. Am. Ceram. Soc. **78**, 563–588 (1993)
37. Nagpure, S.C., Bhushan, B.: In: Bhushan, B., Fuchs, H. (eds.) Applied Scanning Probe Methods – Biomimetics and Industrial Applications, vol. 13, pp. 203–233. Springer, Heidelberg (2009)
38. Nagpure, S.C., Babu, S.S., Bhushan, B., Kumar, A., Mishra, R., Windl, W., Kovarik, L., Mills, M.: Local electronic structure of $LiFePO_4$ nanoparticles in aged Li-ion batteries (2011, submitted)
39. Nalwa, H.S.: Nanomaterials for Energy Storage Application, 1st edn. American Scientific Publishers, Stevenson Ranch (2009)
40. Niemann, M.U., Srinivasan, S.S., Phani, A.R., Kumar, A., Goswami, D.Y., Stefanakos, E.K.: Nanomaterials for hydrogen storage applications: a review. J. Nanomater. **2008**, 1–9 (2008)
41. Ormerod, R.M.: Solid oxide fuel cells. Chem. Soc. Rev. **32**, 17–28 (2003)
42. Panella, B., Hirscher, M., Pütter, H., Müller, U.: Hydrogen adsorption in metal–organic frameworks: Cu-MOFs and Zn-MOFs compared. Adv. Funct. Mater. **16**, 520–524 (2006)
43. Peled, E.: The electrochemical behavior of alkali and alkaline earth metals in nonaqueous battery systems-The solid electrolyte interphase model. J. Electrochem. Soc. **126**, 2047–2051 (1979)
44. Reule, H., Hirscher, M., Weißhardt, A., Kronmüller, H.: Hydrogen desorption properties of mechanically alloyed MgH_2 composite materials. J. Alloys Compd. **305**, 246–252 (2005)
45. Rovetto, L.J., Strobel, T.A., Hester, K.C., Dec, S.F., Koh, C.A., Miller, K.T., Sloan, E.D.: Molecular Hydrogen Storage in Novel Binary Clathrate Hydrates at Near-Ambient Temperatures and Pressures, pp. 629–632. Basic Energy Sciences, U.S. Department of Energy, Washington, DC (2006)
46. Simon, P., Gogotsi, Y.: Materials for electrochemical capacitors. Nat. Mater. **7**, 845–854 (2008)
47. Singhal, S.C.: Solid oxide fuel cells for stationary, mobile, and military applications. Solid State Ionics **152–153**, 405–410 (2002)
48. Sloan, E.D.: Clathrate Hydrates of Natural Gases, 2nd edn. CRC Press, Boca Raton (1998)
49. Takahashi, K., Limmer, S.J., Wang, Y., Cao, G.: Synthesis and electrochemical properties of single-crystal V_2O_5 nanorod arrays by template-based electrodeposition. J. Phys. Chem. B **108**, 9795–9800 (2004)
50. Vetter, J., Nováak, P., Wagner, M.R., Veit, C., Möller, K.–.C., Besenhard, J.O., Winter, M., Wohlfahrt-Mehrens, M., Vogler, C., Hammouche, A.: Ageing mechanisms in lithium-ion batteries. J. Power Sources **147**, 269–281 (2005)
51. Wang, Y., Xia, Y.: Electrochemical capacitance characterization of NiO with ordered mesoporous structure synthesized by template SBA-15. Electrochim. Acta **51**, 3223–3227 (2006)
52. Whittingham, M.S.: Electrical energy storage and intercalation chemistry. Science **192**, 1126–1127 (1976)
53. Whittingham, M.S.: Materials challenges facing electrical energy storage. MRS Bull. **33**, 411–419 (2008)
54. Whittingham, M.S.: Materials challenges facing electrochemical energy storage: batteries and capacitors. In: Ginley, D.S., Cohen, D. (eds.) Fundamentals of Materials for Energy and Environmental Sustainability. Materials Research Society, Warrendale (2011), (in press)
55. Züttel, A.: Materials for hydrogen storage. Mater. Today **6**, 24–33 (2003)

Nanomaterials for Excitonic Solar Cells

Mónica Lira-Cantú and Irene González-Valls
Laboratory of Nanostructured Materials for Photovoltaic Energy, Escola Tecnica Superior d Enginyeria (ETSE), Centre d'Investigació en Nanociència i Nanotecnología (CIN2, CSIC), Bellaterra (Barcelona), Spain

Synonyms

Dye sensitized solar cells; Excitonic solar cell; Hybrid solar cells; Nanostructured solar cells; Organic solar cells; Quantum dot solar cells

Definition

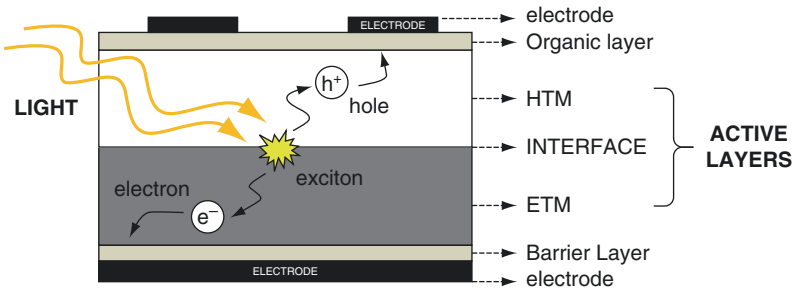
An excitonic solar cell (XSC) is characterized by the formation of an *exciton* that is produced from the absorbed photons from sunlight. An exciton is a tightly bound electron–hole pair which must be split for charge generation. The exciton dissociation takes place at the interface between the constituent semiconductors. Nanostructuration of interfaces is a crucial approach toward highly efficient devices.

Overview

Among the many energy resources, the conversion of sunlight into electricity has emerged as one of the most promising alternatives. The evolution of photovoltaic systems initiated several years ago with the silicon-based solar cells (SiSCs). The need to lower energy cost and to improve device versatility (e.g., flexible, lightweight devices) has resulted on the maturity of the technology toward inorganic thin film photovoltaics (TFSCs) and, more recently, into a new generation called excitonic solar cells (XSCs) [1–3]. The most representative examples of XSCs are dye-sensitized solar cells [1, 4], hybrid solar cells [3, 5–9], and organic polymer solar cells [2, 3, 10–13]. The main difference between inorganic solar cells (SiSCs and TFSCs) and XSCs is their photoconversion mechanism. In an inorganic solar cell, the absorption of light leads to the direct formation of free electron–hole pairs and the direct production of charge carriers. In XSCs, the absorption of photons from sunlight produces an exciton, a tightly electron–hole pair which must be separated for charge generation. The exciton dissociation takes place at the interface of the constituent semiconductors. While inorganic solar cells relay only on inorganic semiconductors, XSCs apply organic semiconductor materials as part of the device active layer. The presence of organic semiconductors (e.g., dyes, small molecules, organic polymers) implies weak intermolecular interactions and low dielectric constants which leads to greater localization of the photoexcited states. The dissociation of the exciton can only be possible if the band offset energy at the junction is greater than the exciton binding energy in the material from which it was generated. In comparison, excitons in SiSCs have binding energies of about 20 meV, and, as a consequence, the absorption of photons generates free charge carriers in the bulk. The electron and hole formed segregates to the electrodes due to the build-up potential generated at the p-n junction. These different mechanisms indicate that each type of solar cell shows a high level of dependence on the crystallinity of the materials applied. Therefore, inorganic solar cells require highly pure and highly crystalline inorganic semiconductors (e.g., Silicon), while XSCs can utilize less crystalline and less pure compounds (e.g., TiO_2). The application of less crystalline materials together with organic semiconductors indicates that the interface between constituents must be carefully controlled, but also permits the fabrication of XSCs by low-cost and scalable printing techniques. The key for highly efficient XSCs is the control of the interfaces through materials nanostructuration.

Classification of Excitonic Solar Cells (XSCs)

XSCs can be classified by the type of semiconductor material applied or by the type of p-n junction used in the final device [7]. Both types of classification are strongly interconnected and



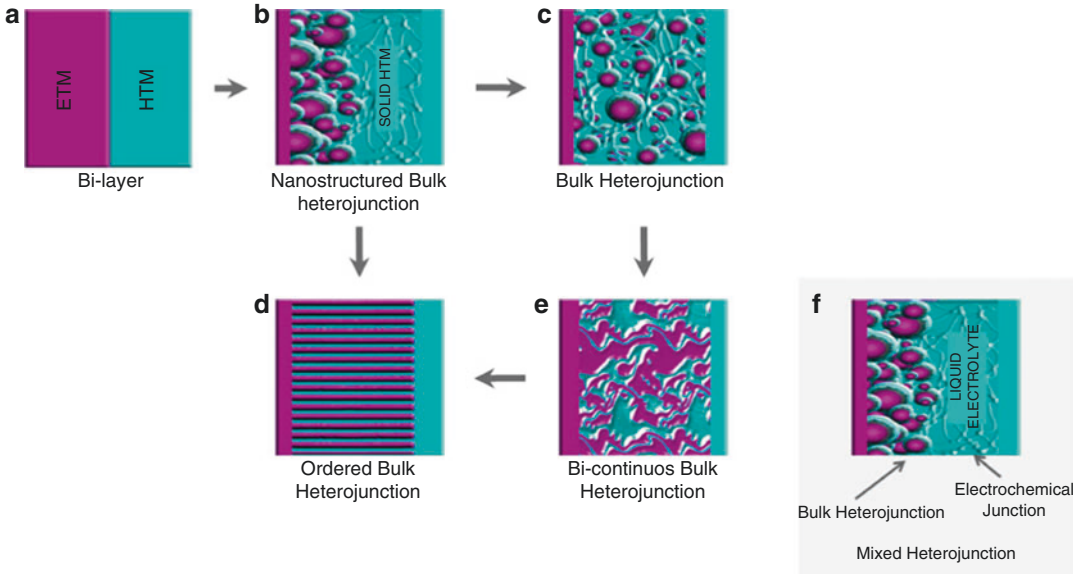
Nanomaterials for Excitonic Solar Cells, Fig. 1 A general schematic representation of an excitonic solar cell (*XSC*) showing the active layers made of the hole transport material (*HTM*) and the electron transport material (*ETM*). The possible extra layers that can be found in XSCs are also indicated. The interface plays a crucial role in exciton dissociation

sometimes generate confusion. A good example is given by the terms *bulk heterojunction solar cell* and *organic solar cell*, which are usually employed when a blend between a polymer and an electron transport material (e.g., P3HT and PCBM) is used as the active layer. In this case, the term *organic solar cell* indicates that the device is being labeled depending on the materials employed (two organic semiconductors). In the case of the term *bulk heterojunction*, the solar cell is defined by the type of p-n junction, since the organic semiconductors employed for the fabrication of the device (donor and acceptor) are mixed heterogeneously in the bulk.

Before a general classification can be defined, the identification of the basic components of XSCs must be made. The main parts of these systems are the electron transport material (ETM) and the hole transport material (HTM) that are in contact through an *interface* (Fig. 1). In some cases, the ETM and the HTM are separated by a thin layer of a light-harvesting molecule (LHM), like the dye in a dye-sensitized solar cell, or inorganic quantum dots (QD) in some hybrid solar cells. In other cases, the light-harvesting material has a dual purpose acting also as the HTM, for example, in an oxide/polymer hybrid solar cell, where the polymer is the responsible for light harvesting and hole transport at the same time. The ETM, LHM, and HTM are "sandwiched" between two electrically conductive electrodes, accountable for charge collection. Thus, in a working device, the LHM is the responsible for photon absorption, the exciton that is formed is then dissociated at the *interface* between the HTM and ETM. Both the ETM and the HTM are responsible for charge transport (electrons and holes, respectively) through the current collectors. *Thus, the LHM, the ETM, and the HTM are the main components of XSCs* for which the device is being classified. Other components of the solar cell offer benefits to the whole device, for example, oxide buffer layers in DSCs or PEDOT and oxide layers in OSCs, among others. These extra layers are used for multiple purposes, especially to facilitate electron or hole transport or to improve interface connection, but are not the main active layer of the solar cell. Thus, in order to identify the type of XSCs under study, a clear differentiation of these extra layers from the main active layer of the XSCs should be made. An easy rule of tomb is that the XSCs will never function without the main active layers, but they can still show photovoltaic response (perhaps not optimal) when the "extra" layers are not present. Figure 1 shows a schematic representation of a XSCs indicating each layer within the device.

Classification by the Type of p-n Junction

XSCs can be classified by the type of junction used [3]. However, it has to be clear that the p-n junction is composed by the ETM and the HTM. Examples of ETM are semiconductor oxides like TiO_2 or ZnO, inorganic semiconductors like Sb_2S_3 or CdS, or C_{60} or C_{70} derivatives. Conducting organic polymers usually act as HTM, but some of them can also act as



Nanomaterials for Excitonic Solar Cells, Fig. 2 Schematic representation of p-n junctions found in XSCs: (**a**) bilayer, (**b**) nanostructured bilayer, (**c**) nanostructured bulk heterojunction, (**d**) ordered bulk heterojunction, (**e**) bicontinous bulk heterojunction, and (**f**) mixed bulk heterojunction and electrochemical junction

the light-harvesting material (LHM) like polythiophenes. There are also HTMs made of small molecules like the *spiro*-OMeTAD. In the case of LHM, a dye molecule is an excellent example. These dyes are usually applied in DSC, but some conducting polymers and inorganic quantum dots can also act as the LHM.

In an XSC, the p-n junction is a critical parameter, since the exciton diffusion length is limited to 10–20 nm in polymers and up to 1 μm in high-quality small molecule films. Thus, in the device structure, the p-n junction must maximize the volume of the light-harvesting material at the interface of the constituent semiconductors, contributing in this way to an optimum charge generation and providing mobility pathways to enhance charge collection. A schematic representation of some of the possible p-n junctions is illustrated in Fig. 2. The most basic p-n junction is the bilayer (Fig. 2a) where a smooth and flat interface is made between the HTM and the ETM. The example in Fig. 2b represents a nanostructured bulk heterojunction where the surface area of the p-n junction has been enhanced by the application of nanoparticles increasing the amount of LHM present at the interface, if compared to bilayer flat devices represented in Fig. 2a. This type of heterojunction can be well ordered, as represented in Fig. 2d, by the application of vertically aligned nanowires, nanorods, nanotubes, or similar. If the nanoparticles in the electrode shown in Fig. 2b, are distributed heterogeneously within a polymer matrix, the final electrode will look like as in Fig. 2c. The latter configuration is now called a "bulk heterojunction." In this junction, the electrons and holes must be transported through the bulk and reach the current collector; thus an interconnection between materials constituents (especially between molecules, nanoparticles, or similar) must be made. The latter is achievable usually by the application of thermal treatments (or similar), and the resulting p-n junction is a bicontinuous network represented in Fig. 2e. A bicontinuous bulk heterojunction is usually observed in organic solar cells (OSC) where the ETM, a C_{60} derivative like the PCBM, is well dispersed within a polymer matrix, like the P3HT. The same heterojunction can also be found in, for example, an HSC of the type polymer/oxide where different nanoforms of ZnO, can dispersed on a polymer matrix like P3HT polymers. The bicontinuous bulk heterojunction is

characterized by the randomly dispersion of the active constituents. A drawback is observed if some of the materials used stay isolated after thermal treatment, forming islands that cannot account for exciton dissociation, and an unoptimized charge generation is observable form the solar cell. Thus, optimization of the bulk heterojunction to enhance charge separation at the interface is possible if these interfaces can be arranged, for example, by the application of well-ordered nanostructured electrodes. Nanostructuration provides order, and order means control over the electron transport and charge collection. Figure 2d represents an example of such electrode: a well-ordered nanostructured bulk heterojunction made of vertically aligned nanorods, nanowires, or similar (other different nanoforms are also possible).

Two or more heterojunctions can be part of a single device. For example, Fig. 2f represents a device where a bulk heterojunction and an electrochemical junction are used. This example is representative of a DSC where the bulk heterojunction is made between the interaction of the TiO_2 electrode and the light-harvesting dye, and the electrochemical junction is made by the TiO_2/dye electrode and the liquid electrolyte. In a solid-state dye-sensitized solar cell (ss-DSC), where a solid organic semiconductor replaces the liquid electrolyte, both junctions are bulk heterojunctions [3].

Classification by the Type of Material

Classification of XSCs can be made by the type of semiconductor applied: organic (polymers, dyes, small organic molecules, etc.) or inorganic (e.g., ZnO, TiO_2, CdSe, etc.). The outcome is three basic solar cell types: organic solar cells (OSCs), hybrid solar cells (HSCs), and dye-sensitized solar cells (DSCs). In OSCs, both materials are purely organic semiconductors, for example, the blend between two conducting organic polymers or a mixture between a C_{60} derivative and a conjugated polymer. In the case of HSC, the device is characterized for the presence of, at least, one phase of an inorganic semiconductor (e.g., Sb_2S_3, TiO_2) and a second phase of an organic semiconductor, like a conducting organic polymer. Although DSC can be classified also as HSC (due to the application of organic and inorganic materials), it is considered a different type of XSC, due to the presence of an electrolyte (semiliquid, gel, solid, etc.).

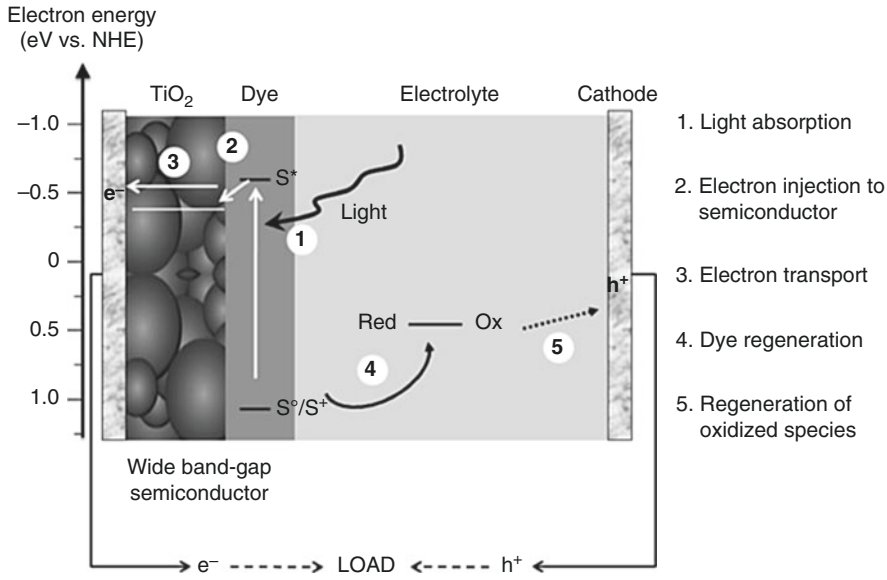
Dye-Sensitized Solar Cell (DSC)

A DSC is characterized by an electrode made of a semiconductor oxide, usually TiO_2 but others like ZnO can also be employed. A charge transfer dye is anchored to the surface of the semiconductor oxide, and finally, a liquid electrolyte containing a redox couple, usually iodide/triiodide in a solvent, completes the device (Fig. 3). The photovoltaic process initiates when photons from sunlight are absorbed by the dye. The exciton is formed and dissociated at the interface between the dye molecule and the semiconductor oxide. The corresponding electron travels through the semiconductor oxide until reaching the current collector (a transparent electrode), and the hole travels through the liquid electrolyte. The regeneration of the dye is possible when it accepts an electron from the electrolyte. The iodine is regenerated by the reduction of the triiodide at the counter electrode (Pt), and the circuit is complete by the migration of the electron through an external load. The voltage generated from the DSC under illumination corresponds to the difference between the Fermi level of the electron in the oxide and the redox potential of the electrolyte [1, 4].

A variation of the DSC is the solid-state dye-sensitized solar cell (ss-DSC) that rose in order to overcome the problems observed in DSC due to the application of liquid electrolyte (solvent evaporation, electrode corrosion, among others). Thus solid, semisolid, gel-type, or nanoclay-based hole conductors have been employed to replace it.

Organic Solar Cells (OSC)

Organic solar cells are characterized for the presence of organic semiconductors in a bulk heterojunction configuration as the main active

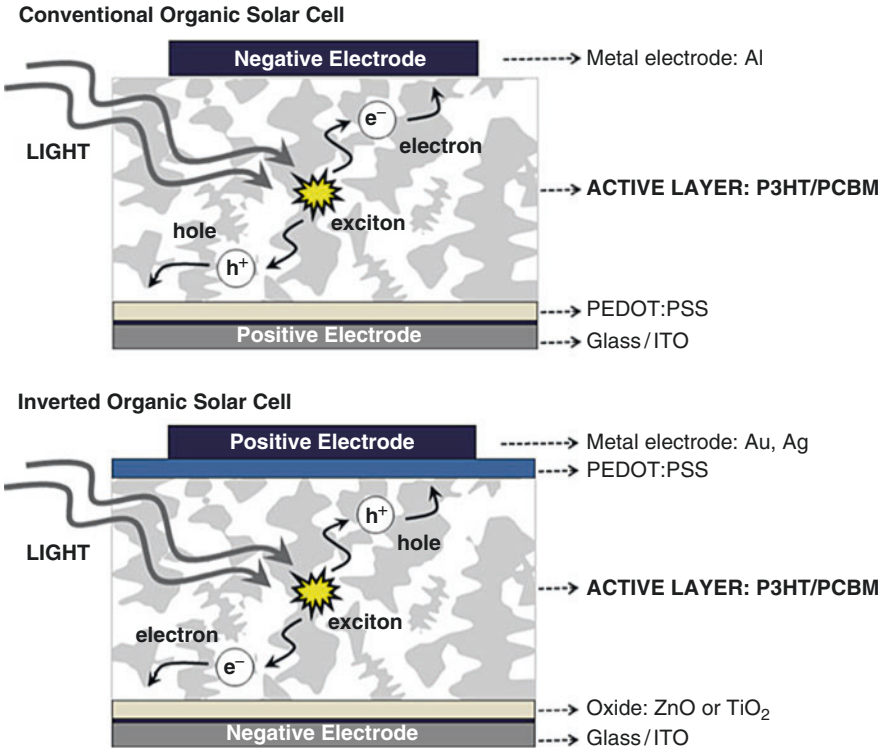


Nanomaterials for Excitonic Solar Cells, Fig. 3 Schematic representation of a DSC. The photovoltaic process initiates with the absorption of photons from sunlight on the light-harvesting dye (S). The different processes taking place are light absorption (*1*), electron injection into the TiO_2 (*2*), electron transport through the TiO_2 (*3*), dye regeneration (*4*), and electrolyte regeneration (*5*) (Adapted from Ref. [1])

layer [2, 10–13]. The best combination is the blend between a C60 derivative, PCBM, acting as the ETM and a conducting organic polymer as the HTM, yet other combinations of organic semiconductors are possible (e.g., the blend between two polymers) [14, 15]. In general, the photovoltaic energy conversion in OSCs is shown schematically in Fig. 4. Photons from sunlight are absorbed by the light-harvesting component, usually a polymer. An exciton is formed at the interface between the HTM and the ETM (or donor/acceptor interface) and dissociates in electron and hole. The electron travels through the ETM and the hole through the HTM until reaching the electrodes and generate charge. Since the active materials that form the bulk heterojunction are distributed on the whole active layer, thin films used, made of organic PEDOT:PSS or inorganic semiconductor oxides, determine the direction at which the electrons and holes migrate, as shown in Fig. 4. Thus, two types of configurations can be found in OSC: the conventional and the inverted. In a conventional OSC, a thin film of PEDOT:PSS is placed between the transparent conducting oxide (TOC) and the active material. Thus, the holes generated at the interface will be transported toward the PEDOT:PSS electrode and the electrons toward the opposite direction. In an inverted device, a thin-film layer of a semiconductor oxide is placed between the TOC and the active layer, while in the opposite side of the solar cell, a layer of the PEDOT:PSS polymer is used. This configuration allows for the electrons to travel toward the side of the oxide thin film, while holes are transported in the opposite direction toward the PEDOT:PSS layer. Although lower photovoltaic properties have been found for the inverted solar cell configuration, the lifetime of the device is being enhanced.

Hybrid Solar Cells (HSC)

Hybrid solar cells present very similar characteristics and mechanisms with OSC; the main difference is that at one of the main components is an inorganic semiconductor [5–9, 16, 17]. The ETMs in an HSC are, for example, semiconductor oxides like TiO_2, ZnO, etc., or inorganic semiconductors like CdS, CdSe, and Sb_2S_3. In both cases, the



Nanomaterials for Excitonic Solar Cells, Fig. 4 Schematic representation of organic solar cells (*OSC*) showing the two configurations: conventional and inverted. Charge generation is possible though three general steps: photon absorption, charge separation, and charge extraction

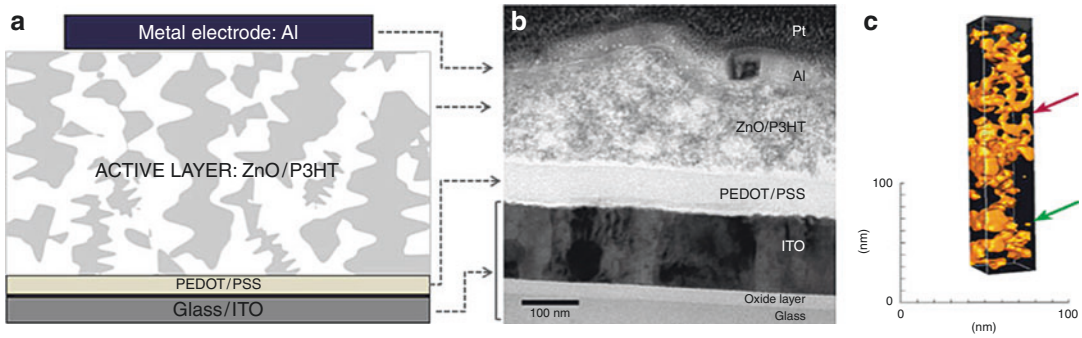
N

inorganic semiconductor can be used in any form: as a dense layers, nanoparticles, nanorods, nanowires, etc. Except for the case of dense layers, all other nanoforms can be mixed within a conducting organic polymer to form a bulk heterojunction configuration. Figure 5 shows an example of an HSC. The active layer is composed of nanoparticles of ZnO (the ETM) in direct contact with a conducting organic polymer like P3HT (as the HTM) acting as the main active layer. The presence of a PEDOT:PSS layer on the ITO side is used as the hole collector, and the Al metal electrode as the electron collector. The position of these electrodes indicates that the device is in its conventional configuration.

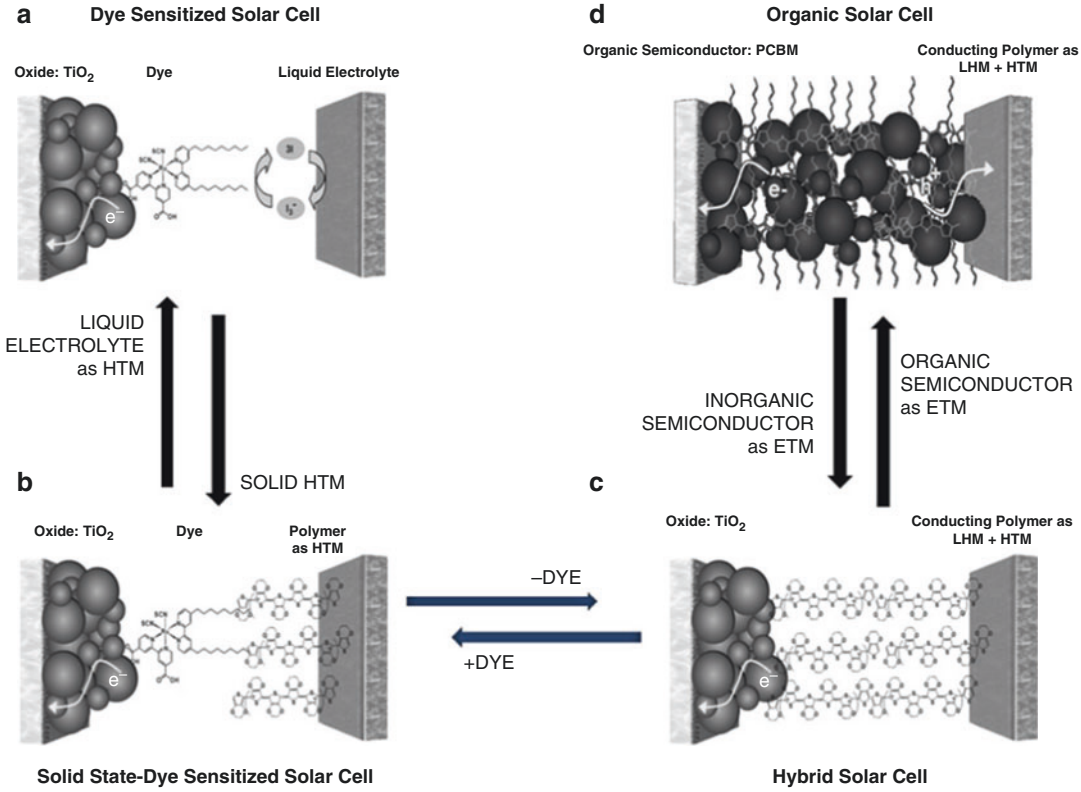
Figure 5c shows a SEM-3D tomography of a section of the active layer of an HSC made of nanoparticles of ZnO and a conducting polymer, P3HT. In the image, the polymer P3HT is represented as transparent and the ZnO nanoparticles in gray. The arrows show two of the main problems observed in bulk heterojunctions: the presence of islands or isolated material (useless as charge carriers) and the connection of some ZnO domains connected only by the top but not through a correct path for charge transport. The latter reduces the efficient electron transport lowering overall device efficiency.

Device Configuration

All XSCs share similar device configuration, active materials, and p-n junctions. To illustrate the latter, in Fig. 6 representation of different XSCs is shown, where the same device configuration is maintained though all the examples and only the materials employed have been changed depending on the type of XSC. Figure 6a represents the DSC device made with a nanoparticulated electrode (usually TiO_2) with a Ru-based dye anchored to it. A redox liquid



Nanomaterials for Excitonic Solar Cells, Fig. 5 (**a**) Schematic representation of an HSC made of ZnO/P3HT bulk heterojunction. (**b**) SEM image of transmission electron micrograph of a cross section of the photovoltaic cell. The platinum layer is deposited later to allow preparation of the TEM sample. (**c**) The *lower arrow* indicates an isolated ZnO domain. The *upper arrow* indicates a ZnO domain that is connected to the top, but not through a strictly rising path (Copyright Nature Publishing Group for images (**b**) and (**c**) [16])



Nanomaterials for Excitonic Solar Cells, Fig. 6 Schematic representation of XSCs by the type of material applied and the relation between solar cell types: (**a**) dye-sensitized solar cell (*DSC*), (**b**) solid-state dye-sensitized solar cell (*ss-DSC*), (**c**) hybrid solar cell (*HSC*), and (**d**) organic polymer solar cell (*OSC*)

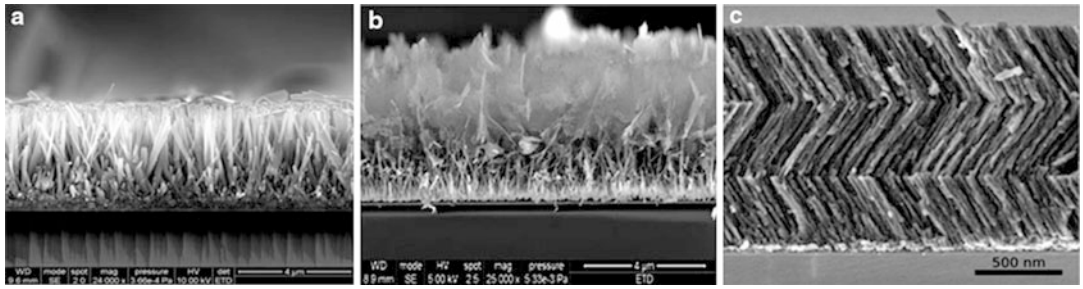
electrolyte completes the device. All the components are "sandwiched" between two electrically conductive electrodes, FTO and Pt respectively, where one of these electrodes must be transparent for light irradiation. A modification of the DSC has been made by the replacement of the liquid electrolyte, with the purpose of minimizing its disadvantages like solvent evaporation, electrode corrosion, etc. Thus, the replacement of the liquid electrolyte in a DSC by a solid HTM, like a conducting organic polymer or organic molecule like *spiro*-OMeTAD, results in what is called a solid-state dye-sensitized solar cell, ss-DSC (Fig. 6b). The light-harvesting dye and the solid hole conductor can be replaced by a single material, for example, a polymer like P3HT which is characterized by a dual function: to collect photons from sunlight and to work as a hole transport. In this simplified version of the ss-DSC, where the ETM is an inorganic semiconductor (e.g., TiO_2, CdS, or similar) and the HTM is an organic polymer (P3HT, MEH-PPV, etc.), the solar cell is known as a hybrid solar cell, HSC (Fig. 6c). The term hybrid is used since the device is made of both organic and inorganic compounds. In this HSC, replacing the *inorganic* ETM semiconductor (e.g., TiO_2, CdS, or similar) by an *organic* ETM semiconductor (e.g., C_{60}, C_{70}, or similar) without modifying the HTM (same organic semiconductor) would result in an organic solar cell (Fig. 6d).

The Benefit of Nanostructuration in XSCs

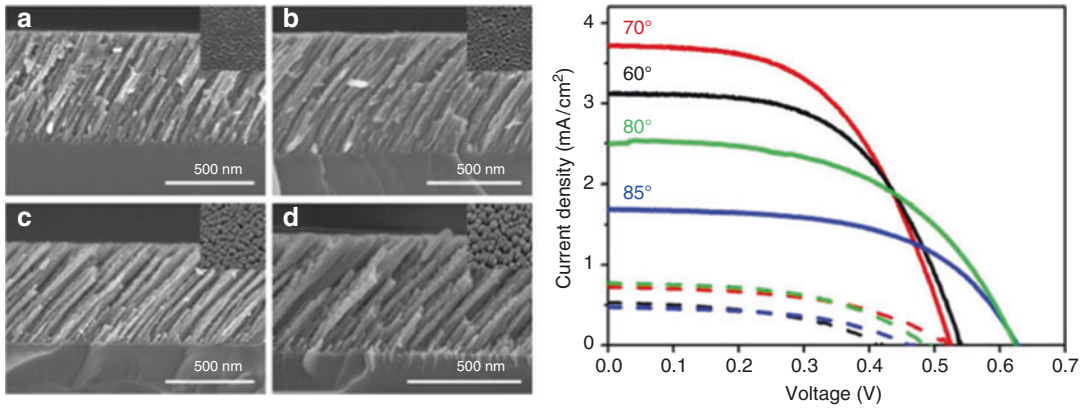
The development of new materials tailored at the nanoscale has become a major research area due to the unique and novel properties that can be obtained. Nanostructured excitonic solar cells have shown the advantages of controlling the interfaces at the nanoscale level. For example, in a bulk heterojunction made of inorganic nanoparticles and a conducting organic polymer, the inorganic semiconductor oxide nanoparticle is dispersed within the conjugated polymer matrix. The resulting oxide nanoparticle/polymer blend must be annealed to improve nanoparticle connection and to favor electron pathway through the electrode. Nevertheless, these devices are still limited by incomplete exciton dissociation due to the isolation of some nanoparticles from each other after thermal treatment.

A nanostructured thin film, for example, well-crystalline and vertically aligned nanowires can provide higher interfacial area between the donor and the acceptor material (polymer/oxide, respectively) with highly efficient electron transport pathways. In the case of DSCs, the replacement of the nanoparticles electrode by vertically aligned nanostructures emerged also as a possibility to obtain faster electron transport thus improving solar cell efficiency [7]. Moreover, the need to replace problematic liquid electrolytes in DSC by solid hole conductors with slower kinetics impels the application of faster electron transport materials like in vertically aligned nanostructures. Figure 7 shows SEM images of different oxide nanostructures, from nanorods (a), nanotrees (b), nanocolumns (c). The electron transport in an electrode made of nanoparticles is limited by the grain boundaries encountered between each nanoparticle. The application of nanorods, especially if well ordered, paves the way for an efficient electron transport through the crystalline structure. The nanostructure can also be tailored to enhance light absorption without eliminating the advantage of the nanorods. The nanotrees (Fig. 7b) were synthesized with a dual purpose: to enhance the light-harvesting properties of the nanorods by increasing the surface area and by maintaining the good electron transport properties that characterize nanorods.

A practical approach of electrode nanostructuration where high transparency is needed has been recently reported. In this case, increasing the NR length will increase light scattering and decrease the transparency of the DSSC device. A singular example to overcome these problems is the application of nanostructured thin films [7, 18]. Transparent TiO_2 thin films were grown by the physical vapor deposition method at glancing incidences (PVD-GLAD). This method of synthesis furnishes aligned nanocolumnar layers at well-controlled oblique angles. This nanostructure is advantageous for solar collection, as it has already been proved that tilting the nanocolumns improves light entrapment efficiency. Zig-zag and



Nanomaterials for Excitonic Solar Cells, Fig. 7 Well-ordered semiconductor oxide (ZnO and TiO_2)-based electrodes applied in XSCs: (**a**) vertically aligned ZnO NANORODS [7] (**b**) vertically aligned ZnO nanotrees, and (**c**) well-ordered TiO_2 nanocolumns with tailored inclination angle (From Ref. [18], copyright 2011, Royal Society of Chemistry)



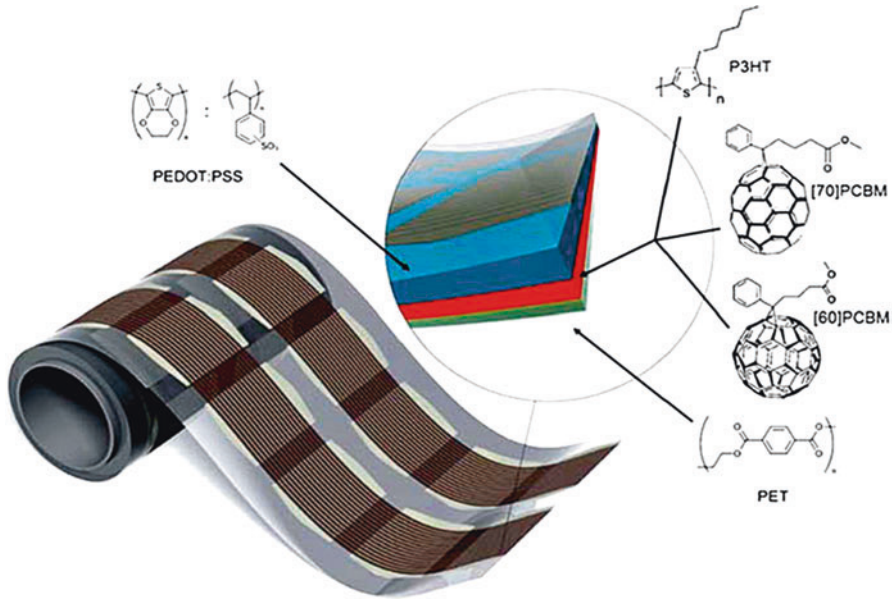
Nanomaterials for Excitonic Solar Cells, Fig. 8 *Left*: Ordered nanocolumns of TiO_2 deposited by the GLAD technique at different angles (From Ref. [18], copyright 2011, Royal Society of Chemistry), and *Right*: IV curves of the DSC applying the TiO_2 electrodes

other complex morphologies can be also prepared by this method. The resulting thin-film electrodes present very open and porous microstructure where nanometer size columns with a high internal porosity are separated by wide pores that extend from the electrode substrate up to the film surface. Many are the factors to be considered when synthesizing these nanostructures: the effect of electrode thickness, TiO_2 crystallinity, topology when arranged into multilayers, the pore structure, or the distribution and concentration of dye molecules within the film thickness. Figure 8 shows SEM images (left) of ordered nanocolumns of TiO_2 deposited by the GLAD technique at different angles. Each sample is 500 nm thick and the angles at which were deposited vary between 60° and 85°. Analyses of the samples as electrode in DSC demonstrated that an optimal deposition angle exists at 70°.

The enhancement observed has been attributed to the increase in the dye-loading capacity of the thin films. Setting the deposition angle at around 70°, at which a 500-nm-thick electrode solar cell presents a maximum efficiency of 1.04 %. A zig-zag nanostructure of the TiO_2 thin films made of stacks of 500-nm layers deposited at the optimum 70° angles, was applied to grow thicker films. This configuration ensured good mechanical stability and high transparency for the films. Power conversion efficiency was observed to increase with electrode thickness, reaching a conversion efficiency of 2.78 % for 3-μm

electrodes of quite good transparency [18]. The high transparency of the obtained cells supports its implementation onto colored glass and other related applications.

There are many possible methods to fabricate nanomaterials, and the observed properties depend a great extent on the preparation conditions. Nanostructure synthesis can be found as simple nanoparticles or as nanorods, nanobelts, branched nanorods, nanowires, ultranarrow nanobelts, hierarchical nanostructures, nanocombs, nanosprings, nanospirals, searnless nanorings, core-shell nanostructures, nanocages, nanoflowers, among many others. It is, however, also difficult to establish the optimal synthesis methodology for solar cell applications because the efficiency of an XSC depends on many aspects and not only on the material nanostructure, like thin film thickness, nanostructure dimension or surface area, but also on overall device preparation like polymer or dye applied, back contacts, active area, etc. However, for XSCs, the most attractive syntheses techniques due to the ease of its process, the low temperature employed, and the possibility to obtain the oxide nanostructures on a large area by continuous methods and in different nanostructures are the hydrothermal methods, among them the chemical bath deposition. The SILAR method is also an alternative which interest is growing due to the ease of its application. These methods permit the synthesis of materials not only as nanoparticles but also as well-aligned nanostructures or core-shell nanostructures. These synthesis methods have been applied for the synthesis of many different inorganic semiconductors like TiO_2, ZnO, CdSe, etc. The synthesis involves precursor baths with the ions of the materials to be synthesized and the technique can be applied for the synthesis of isolated materials (e.g., nanoparticles), the fabrication of nanostructured electrodes, or the sensitization of molecules on different substrates. One of the most attractive examples applying the hydrothermal method is the synthesis of vertically aligned nanostructures. The reason behind the interest in these materials is due to the excellent transport properties if compared to nanoparticles. An example is the synthesis of ZnO nanorods that can be made at low temperature, on almost any substrate, and in a continuous mode [19]. The latter are attractive conditions for large scale production. In general, the method to obtain vertically aligned ZnO nanorods is based on a two-step technique involving the coating of a substrate with ZnO seeds of less than 10 nm in diameter. Every ZnO seed works as a nucleation site for the formation of ZnO nanowires under mid-wet conditions. The synthesis conditions, like zinc concentration in solution, growth time, or growth temperature, determine the final nanorod dimensions and quality. Also the stability of the precursor ZnO seeds (nanoparticles) or the substrates to be applied are important factors affecting the final ZnO nanostructures with reproducible results. As an example, the variation of ZnO nanorod length and diameter can be adjusted depending on the additives added during synthesis. Polyethylenimine (PEI) has been used for large surface area electrodes, so higher amount of dye absorption can be obtained when applied in a DSC. Nanorods as long as 30 μm were observed when PEI additive was used; without additive the length of the nanorods was not more than 5 μm, and their diameters increased also with time. Yet, the longest nanorods, 20–25 μm, show only one fifth of the surface area of an electrode made of nanoparticles [7]. Vertically aligned nanostructures of ZnO have been extensively applied in XSCs. Nevertheless, in terms of upstanding and well-ordered alignment, the current synthesis methodologies applied for their fabrication produce irreproducible nanostructures with dimensions that are too far away from the ideal electrode characteristics. Synthesis condition must be controlled carefully; the minimal change in synthesis condition or even the roughness of the underlying substrate can affect nanorod dimension. Thus, a careful design of the nanostructured electrode made of vertically aligned nanorods should be made in order to fulfill the best requirements for the type of solar cell where it will be applied. For example, one of the most problematic issues observed in XSCs is dye or polymer infiltration (wettability); a solvent-soluble dye could reach the whole surface area of the nanostructured electrode more easily than a



Nanomaterials for Excitonic Solar Cells, Fig. 9 A roll of the printed modules showing how two parallel sets of modules were prepared simultaneously. The zoom-in shows the layer stack which is PET-ITO-ZnO-P3HT: PCBM-PEDOT:PSS-printed silver. The device stack was encapsulated using roll-to-roll lamination post production (From Ref. [20], copyright 2011 Royal Society of Chemistry)

viscous polymer solution which has to penetrate into dense packed nanowire network.

Vertically aligned nanocolumns (nanorods, nanowires, nanotips, etc.) made of semiconductor oxides have been proved to enhance electron transport and reduce the recombination probability in different excitonic solar cells (organic, dye sensitized, or hybrid). Nevertheless, the power conversion efficiencies of these modified solar cells are still far below that obtained with nanoparticle-based electrodes. Their lower efficiency has been attributed to the relatively low surface area of nanocolumnar electrodes and some inefficiency of light absorption since some photons fall on the gap between adjacent vertical nanorods. Moreover, the possibility to obtain highly porous ZnO nanomaterials with good crystalline structure is possible and it suggests that there is room for improvement on solar cell efficiency. In principle, the efficiency of these type of solar cells could be enhanced by increasing the thickness of the aligned nanocolumnar thin films. The solar cells can be nanostructured to get higher surface areas and longer nanocolumns. Nevertheless, recent results show that even NR lengths as large as 40 μm result in low power conversion efficiency. Most recent results show that the modification of the synthesis methodology by the application of additives or the tuning of the synthesis conditions (temperature, time, etc.) reduces traps found on the semiconductor surface. The challenge nowadays resides in the interplay between large-area printing technology and high-quality nanostructuration. The control of both parameters will balance final device cost and functionality.

Future Trends: From Nanomaterials to Large-Scale Printed Devices

XSCs are promising devices due to the possibility to be fabricated by scalable and low-cost printing techniques (as example see Fig. 9), unachievable nowadays for photovoltaic systems based on crystalline silicon. Application of printing methods, like screen printing, roll-to-roll, gravure, offset, lithography, inject, etc., will also permit the integration of other printing electronic devices (lighting, batteries, etc.) to constitute a complete system [10, 12, 20–22]. A challenge in printing

technology is to achieve well-ordered nanostructuration to the main active layer. For the latter, friendly and low-temperature processes (synthesis, pre-/post treatments, etc.) are required.

The development of liquid precursors or "inks," is the key for the scale-up of laboratory devices to commercialization of large-scale printed photovoltaics. Inks can be made from solutions or suspensions. Solution-based inks present advantages like uniform composition and phase, and the possibility of modification of the material (e.g., doping). Nevertheless, the ink solubility may depend on factors such as additives, solvents, humidity, and temperature, among others. All these parameters must be carefully controlled in order to obtain a final useful ink. The preparation of suspensions is another alternative which allows the application of nanostructured solids designed and synthesized at the nanoscale level, with good results [20–22]. Moreover, some recent synthesis methods allow for certain degree of control of the final nanoforms [22]. In the suspension, the problem of the elimination of organic precursors or additives after treatments is a challenge, but the benefits range from tunable electrical and optical properties to the enhancement of mechanical characteristics.

Cross-References

- ▶ Hybrid Solar Cells
- ▶ Organic Photovoltaics: Basic Concepts and Device Physics
- ▶ Plasmonic Structures for Solar Energy Harvesting
- ▶ Toward Bioreplicated Texturing of Solar-Cell Surfaces

References

1. O'Regan, B., Grätzel, M.: A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. Nature **353**(6346), 737–740 (1991)
2. Sun, S.-S., Sariciftci, N.S.: Organic Photovoltaics: Mechanisms, Materials, and Devices (Optical Science and Engineering). CRC Press, Boca Raton (2005). ISBN 0-8247-5963-X
3. Nelson, J.: The Physics of Solar Cells, p. 360. Imperial College Press, London (2003)
4. Hagfeldt, A., Boschloo, G., Sun, L., Kloo, L., Pettersson, H.: Dye-sensitized solar cells. Chem. Rev. **110**(11), 6595–6663 (2010)
5. Beek, W.J.E., Janssen, R.A.J.: Hybrid polymer-inorganic photovoltaic cells. In: Merhari, L. (ed.) Hybrid Nanocomposites for Nanotechnology: Electronic, Optical, Magnetic and Biomedical Applications, pp. 321–386. Springer, New York (2009)
6. Lira-Cantu, M., Gomez-Romero, P.: Multifunctional hybrid materials based on conducting organic polymers. Nanocomposite systems with photo-electro-ionic properties and applications. In: Gomez-Romero, P., Sanchez, C. (eds.) Functional Hybrid Materials, pp. 210–269. Wiley, Weinheim (2004)
7. Gonzalez-Valls, I., Lira-Cantu, M.: Vertically-aligned nanostructures of ZnO for excitonic solar cells: a review. Energy Environ. Sci. **2**(1), 19–34 (2009)
8. Lira-Cantu, M., Gomez-Romero, P.: Next-generation hybrid nanocomposite materials based on conducting organic polymers: energy storage and conversion devices. In: Merhari, L. (ed.) Hybrid Nanocomposites for Nanotechnology: Electronic, Optical, Magnetic and Biomedical Applications, pp. 289–320. Springer, New York (2009)
9. Nozik, A.J., Beard, M.C., Luther, J.M., Law, M., Ellingson, R.J., Johnson, J.C.: Semiconductor quantum dots and quantum dot arrays and applications of multiple exciton generation to third-generation photovoltaic solar cells. Chem. Rev. **110**(11), 6873–6890 (2010)
10. Krebs, F.C.: Polymeric Solar Cells: Materials, Design, Manufacture. DEStech, Lancaster (2010)
11. Günes, S., Neugebauer, H., Sariciftci, N.S.: Conjugated polymer-based organic solar cells. Chem. Rev. **107**(4), 1324–1338 (2007)
12. Brabec, C., Dyakonov, V., Scherf, U.: Organic Photovoltaics: Materials, Device Physics and Manufacturing Technologies. Wiley, Weinhaim (2009)
13. Clarke, T.M., Durrant, J.R.: Charge photogeneration in organic solar cells. Chem. Rev. **110**(11), 6736–6767 (2010)
14. Segura, J.L., Martín, N., Guldi, D.M.: Materials for organic solar cells: the C60/π-conjugated oligomer approach. Chem. Soc. Rev. **34**(1), 31–47 (2005)
15. Martín, N., Giacalone, F., Prato, M.: Fullerene Polymers: Synthesis, Properties and Applications, p. 305. Wiley, Weinheim (2009)
16. Oosterhout, S.D., Wienk, M.M., van Bavel, S.S., Thiedmann, R., Koster, L.J.A., Gilot, J., Loos, J., Schmidt, V., Janssen, R.A.J.: The effect of three-dimensional morphology on the efficiency of hybrid polymer solar cells. Nat. Mater. **8**, 818–824 (2009)
17. Kamat, P.V., Tvrdy, K., Baker, D.R., Radich, J.G.: Beyond photovoltaics: semiconductor nanoarchitectures for liquid-junction solar cells. Chem. Rev. **110**(11), 6664–6688 (2010)
18. González-García, L., González-Valls, I., Lira-Cantu, M., Barranco, A., Gonzalez-Elipe, A.G.: Aligned TiO_2

nanocolumnar layers prepared by PVD-GLAD for transparent dye sensitized solar cells. Energy Environ. Sci. (2001). doi:10.1039/C0EE00489H
19. Morkoc, H., Ozgur, U.: Zinc Oxide: Fundamentals, Materials and Device Technology. Wiley, Weinheim (2009)
20. Krebs, F.C., Fyenbo, J., Jorgenssen, M.: Product integration of compact roll-to-roll processed polymer solar cell modules: methods and manufacture using flexographic printing, slot-die coating and rotary screen printing. J. Mater. Chem. **20**, 8994–9001 (2010)
21. Arias, A.C., MacKenzie, J.D., McCulloch, I., Rivnay, J., Saello, A.: Materials and applications for large area electronics: solution-based approaches. Chem. Rev. **110**(1), 3–24 (2010)
22. Habas, S.E., Platt, H.A.S., van Hest, F.A.M., Ginley, D. S.: Low-cost inorganic solar cells: from ink to printed device. Chem. Rev. **110**(11), 6571–6594 (2010)

Nanomaterials for Sensors

► Nanostructured Materials for Sensing

Nanomechanical

► Nonlinear and Parametric NEMS Resonators

Nano-mechanical Machining

► Ultraprecision Machining (UPM)

Nanomechanical Properties of Nanostructures

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Bending strength; Creep; Fatigue strength; Fracture toughness; Hardness; Indentation; Nanohardness; Relaxation; Young's modulus

Definition

Structural integrity is of paramount importance in all devices. Load applied during the use of devices can result in component failure. Cracks can develop and propagate under tensile stresses, leading to failure. Knowledge of the mechanical properties of nanostructures is necessary for designing realistic micro-/nanoelectromechanical systems (MEMS/NEMS) and biological micro-/nanoelectromechanical systems (BioMEMS/BioNEMS) devices. Elastic and inelastic properties are needed to predict deformation from an applied load in the elastic and inelastic regimes, respectively. The strength property is needed to predict the allowable operating limit. Some of the properties of interest are hardness, elastic modulus, bending strength, fracture toughness, and fatigue strength. Many of the mechanical properties are scale dependent; therefore, these should be measured at relevant scales. Atomic force microscopy and nanoindenters can be used satisfactorily to evaluate the mechanical properties of micro/nanoscale structures.

Overview

A wide variety of MEMS, including Si-based devices, chemical and biological sensors and actuators, and miniature non-silicon structures (e.g., devices made from plastics or ceramics) have been fabricated with dimensions in the range of a couple to a few thousand microns (see e.g., [4, 5, 9, 10, 15–17, 21–23, 29, 32, 40]). A variety of NEMS have also been produced (see e.g., [13, 14, 18, 33, 38, 43]). MEMS/NEMS technology and fabrication processes have found a variety of applications in biology and biomedicine, leading to the establishment of an entirely new field known as BioMEMS/BioNEMS [2, 3, 9, 12, 19, 24, 30, 31, 37, 44]. The ability to use micro/nanofabrication processes to develop precision devices that can interface with biological environments at the cellular and molecular level has led to advances in the fields of biosensor technology [42, 44], drug delivery [34, 36], and tissue engineering [25]. The miniaturization of fluidic

systems using micro/nanofabrication techniques has led to new and more efficient devices for medical diagnostics and biochemical analysis [39]. The largest "killer" industrial applications of MEMS include accelerometers (over a billion US dollars a year in 2004), pressure sensors for manifold absolute pressure sensing for engines (more than 30 million units in 2004), and tire pressure measurements, inkjet printer heads (more than 500 million units in 2004), and digital micromirror devices (about US $700 million revenues in 2004). BioMEMS and BioNEMS are increasingly used in commercial applications. Largest applications of BioMEMS include silicon-based disposable blood pressure sensor chips for blood pressure monitoring (more than 25 million units in 2004), and a variety of biosensors.

Structural integrity is of paramount importance in all devices. Load applied during the use of devices can result in component failure. Cracks can develop and propagate under tensile stresses leading to failure [6, 8]. Friction/stiction and wear limit the lifetimes and compromise the performance and reliability of devices involving relative motion [4, 5, 7, 9, 10]. Most MEMS/NEMS applications demand extreme reliability. Stress and deformation analyses are carried out for an optimal design. MEMS/NEMS designers require mechanical properties on the nanoscale. Elastic and inelastic properties are needed to predict deformation from an applied load in the elastic and inelastic regimes, respectively. The strength property is needed to predict the allowable operating limit. Some of the properties of interest are hardness, elastic modulus, creep, bending strength (fracture stress), fracture toughness, and fatigue strength. Micro/nanostructures have some surface topography and local scratches dependent upon the manufacturing process. Surface roughness and local scratches may compromise the reliability of the devices, and their effect needs to be studied.

Most mechanical properties are scale dependent [5, 6, 8–10]. Several researchers have measured the mechanical properties of silicon and silicon-based milli- to microscale structures including tensile tests and bending tests, resonant structure tests for the measurement of elastic properties, fracture toughness tests, and fatigue tests. A few researchers have measured the mechanical properties of nanoscale structures using atomic force microscopy (AFM) and nanoindentation. For stress and deformation analyses of simple geometries and boundary conditions, analytical models can be used. For analysis of complex geometries, numerical models are needed. Conventional finite element method (FEM) can be used down to a few tens of nanometer scale, although its applicability is questionable at nanoscale. FEM has been used for simulation and prediction of residual stresses and strains induced in MEMS devices during fabrication, to perform fault analysis in order to study MEMS faulty behavior, to compute mechanical strain resulting from doping of silicon, analyze micromechanical experimental data, and nanomechanical experimental data. FEM analysis of nanostructures has been performed to analyze the effect of types of surface roughness and scratches on stresses in nanostructures.

Commonly used materials for MEMS/NEMS are single-crystal silicon and silicon-based materials (e.g., SiO_2 and polysilicon films deposited by low pressure chemical vapor deposition (LPCVD) process) [9, 10]. An early study showed silicon to be a mechanically resilient material in addition to its favorable electronic properties. Single-crystal 3C-SiC (cubic or β-SiC) films, deposited by atmospheric pressure chemical vapor deposition (APCVD) process on large-area silicon substrates, are produced for high-temperature micro/nanosensor and actuator applications. Amorphous alloys can be formed on both metal and silicon substrates by sputtering and plating techniques, providing more flexibilities in surface-integration. Electroless deposited Ni-P amorphous thin films have been used to construct microdevices, especially using the so-called LIGA techniques. Micro/nanodevices need conductors to provide power as well as electrical/magnetic signals to make them functional. Electroplated gold films have found wide applications in electronic devices because of their ability to make thin films and process simplicity.

As the field of MEMS/NEMS has progressed, alternative materials, especially polymers, have

established an important role in their construction [9, 10]. This trend has been driven by the reduced cost associated with polymer materials. Polymer microfabrication processes, including micromolding and hot embossing techniques, can be orders of magnitude less expensive than traditional silicon photolithography processes. Many polymers are biocompatible, so these can be integrated into biomedical devices with minimal detrimental effects to the host or on the biofluids. An improvement in device functionality (relative to silicon) is also possible due to the mechanical properties of the polymer used, which have properties much closer to biological tissues. Polymer BioMEMS structures involving microbeams have been designed to measure cellular forces. Polymer materials most commonly used for biomedical applications include poly (methyl methacrylate) (PMMA), poly(dimethylsiloxane) (PDMS), and polystyrene. Another material of interest due to ease of fabrication is poly(propyl methacrylate) (PPMA), which has lower glass transition temperature (Tg) (35–43 °C) than PMMA (104–106 °C), which permits low temperature thermal processing.

Experimental Techniques for Measurement of Mechanical Properties of Nanostructures

Indentation and Scratch Tests Using Micro/Nanoindenters

A nanoindenter is commonly used to measure hardness, elastic modulus, and fracture toughness, and to perform micro/nanoscratch studies to get a measure of scratch/wear resistance of materials [5, 9–11].

Hardness and Elastic Modulus

The nanoindenter monitors and records the dynamic load and displacement of the three-sided pyramidal diamond (Berkovich) indenter during indentation with a force resolution of about 75 nN and displacement resolution of about 0.1 nm. Hardness and elastic modulus are calculated from the load–displacement data [5, 10, 11]. The peak indentation load depends on the mechanical properties of the specimen; a harder material requires a higher load for reasonable indentation depth.

Fracture Toughness

The indentation technique for fracture toughness measurement of brittle samples on the microscale is based on the measurement of the lengths of median-radial cracks produced by indentation. A Vickers indenter (a four-sided diamond pyramid) is used in a microhardness tester. A load on the order of 0.5 N is typically used in making the Vickers indentations. The indentation impressions are examined using an optical microscope with Nomarski interference contrast to measure the length of median-radial cracks, c. The fracture toughness (K_{IC}) is calculated by the following relation [26]:

$$K_{IC} = \alpha \left(\frac{E}{H}\right)^{1/2} \left(\frac{P}{c^{3/2}}\right) \quad (1)$$

where α is an empirical constant depending on the geometry of the indenter, H and E are hardness and elastic moduli, and P is the peak indentation load. For Vickers indenters, α has been empirically found based on experimental data and is equal to 0.016 [5]. Both E and H values are obtained from the nanoindentation data. The crack length is measured from the center of the indent to the end of crack using an optical microscope. For one indent, all crack lengths are measured. The crack length c is obtained from the average value of several indents.

Indentation Creep

For indentation creep test of polymer samples, the test is performed using a continuous stiffness measurements (CSM) technique [11]. In a study by Wei et al. [45], the indentation load was typically 30 μN, and the loading rate was 3 μN/s. The tip was held typically for 600 s after the indentation load reached 30 μN. To measure the mean stress and contact stiffness, during the hold segment the indenter was oscillated at a peak-to-peak load amplitude of 1.2 μN and a frequency of 45 Hz.

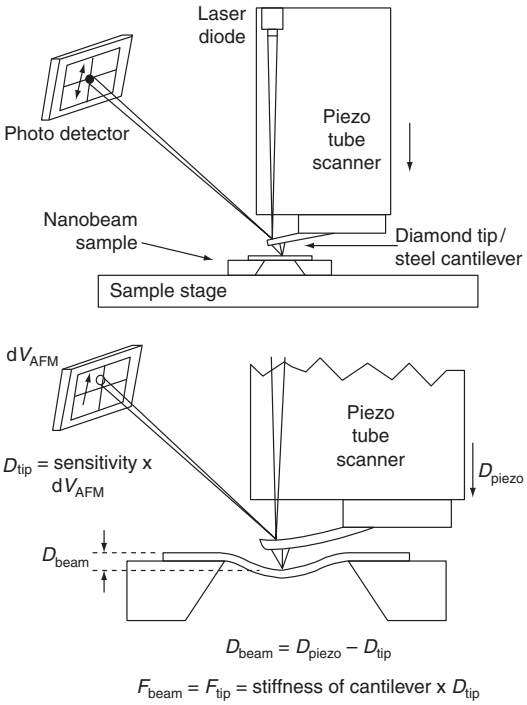
Scratch Resistance

In micro/nanoscratch studies, in a nanoindenter, a conical diamond indenter having a tip radius of about 1 μm and an included angle of 60° is drawn over the sample surface, and the load is ramped up until substantial damage occurs [9–11]. The coefficient of friction is monitored during scratching. In order to obtain scratch depths during scratching, the surface profile of the sample surface is first obtained by translating the sample at a low load of about 0.2 mN, which is insufficient to damage a hard sample surface. The 500 μm long scratches are made by translating the sample while ramping the loads on the conical tip over different loads dependent upon the material hardness. The actual depth during scratching is obtained by subtracting the initial profile from the scratch depth measured during scratching. In order to measure the scratch depth after the scratch, the scratched surface is profiled at a low load of 0.2 mN and is subtracted from the actual surface profile before scratching.

Bending Tests of Nanostructures Using an AFM

Quasi-static bending tests of fixed nanobeam arrays in the normal direction are carried out using an AFM [41]. A three-sided pyramidal diamond tip (with a radius of about 200 nm) mounted on a rectangular stainless steel cantilever is used for the bending tests. The beam stiffness is selected based on the desired load range. The stiffness of the cantilever beams for application of a normal load up to 100 μN is about 150–200 N/m.

For the bending test, the tip is brought over the nanobeam array with the help of the sample stage of the AFM and a built-in high magnification optical microscope (Fig. 1) [41]. For the fine positioning of the tip over a chosen beam, the array is scanned in contact mode at a contact load of about 2–4 μN, which results in negligible damage to the sample. After scanning, the tip is located at one end of a chosen beam. To position the tip at the center of the beam span, the tip is moved to the other end of the beam by giving the *X*-piezo an offset voltage. The value of this offset is determined after several such attempts have been made



Nanomechanical Properties of Nanostructures, Fig. 1 Schematic showing the details of a nanoscale bending test using an AFM. The AFM tip is brought to the center of the nanobeam and the piezo is extended over a known distance. By measuring the tip displacement, a load–displacement curve of the nanobeam can be obtained [41]

in order to minimize the effects of piezo drift. Half of this offset is then applied to the *X*-piezo after the tip is positioned at one end of the beam, which usually results in the tip being moved to the center of the span. Once the tip is positioned over the center of the beam span, the tip is held stationary without scanning, and the *Z*-piezo is extended by a known distance, typically about 2.5 μm, at a rate of 10 nm/s, as shown in Fig. 1. During this time, the vertical deflection signal (dV_{AFM}), which is proportional to the deflection of the cantilever (D_{tip}), is monitored. The displacement of the piezo is equal to the sum of the displacements of the cantilever and the nanobeam. Hence the displacement of the nanobeam (D_{beam}) under the point of load can be determined as

$$D_{beam} - D_{piezo} - D_{tip} \quad (2)$$

The load (F_{beam}) on the nanobeam is the same as the load on the tip/cantilever (F_{tip}) and is given by

$$F_{\text{beam}} = F_{\text{tip}} = D_{\text{tip}} \times k \tag{3}$$

where k is the stiffness of the tip/cantilever. In this manner, a load–displacement curve for each nanobeam can be obtained.

The photodetector sensitivity of the cantilever needs to be calibrated to obtain D_{tip} in nm. For this calibration, the tip is pushed against a smooth diamond sample by moving the Z-piezo over a known distance. For the hard diamond material, the actual deflection of the tip can be assumed to be the same as the Z-piezo travel (D_{piezo}), and the photodetector sensitivity (S) for the cantilever setup is determined as

$$S = D_{\text{piezo}}/\mathrm{d}V_{\text{AFM}}\text{nm}/V \tag{4}$$

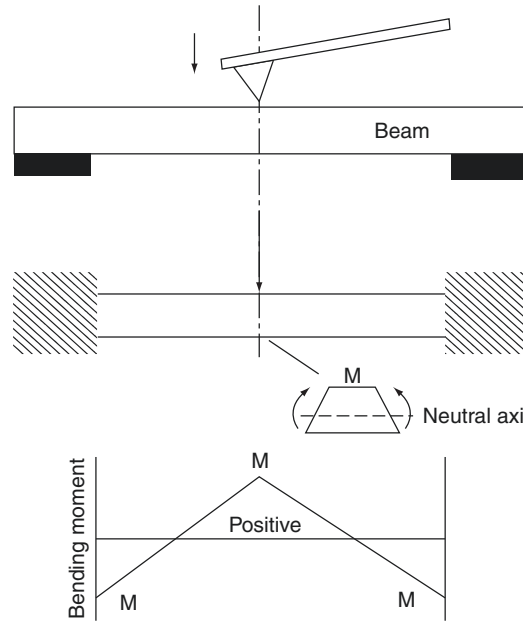
In the measurements, D_{tip} is given as $\mathrm{d}V_{\text{AFM}} \times S$.

Since a sharp tip would result in an undesirable large local indentation, Sundararajan and Bhushan [41] used a diamond tip which was worn (blunt). Indentation experiments using this tip on a silicon substrate yielded a residual depth of less than 8 nm at a maximum load of 120 μN, which is negligible compared to displacements of the beams (several hundred nanometers). Hence one can assume that negligible local indentation or damage is created during the bending process of the beams and that the displacement calculated from Eq. 2 is from the beam structure.

Elastic Modulus and Bending Strength

Elastic modulus and bending strength (fracture stress) of the beams can be estimated by equations based on the assumption that the beams follow the linear elastic theory of an isotropic material. This is probably valid since the beams have high length-to-width ℓ/w and length-to-thickness ℓ/t ratios and also since the length direction is along the principal stress direction during the test. For a fixed elastic beam loaded at the center of the span, the elastic modulus is expressed as

$$E = \frac{\ell^3}{192I}\text{m} \tag{5}$$



Nanomechanical Properties of Nanostructures, Fig. 2 A schematic of the bending moments generated in the beam during a quasi-static bending experiment, with the load at the center of the span. The maximum moments occur under the load and at the fixed ends. Due to the trapezoidal cross section, the maximum tensile bending stresses occur at the top surfaces at the fixed ends

where ℓ is the beam length, I is the area moment of inertia for the beam cross section, and m is the slope of the linear region of the load–displacement curve during bending [46]. The area moment of inertia for a beam with a trapezoidal cross section is calculated from the following equation:

$$I = \frac{4w_1^2 w_1 w_2 + w_2^2}{36(w_1 + w_2)} t^3 \tag{6}$$

where w_1 and w_2 are the upper and lower widths, respectively, and t is the thickness of the beam. According to linear elastic theory, for a centrally loaded beam, the moment diagram is shown in Fig. 2. The maximum moments are generated at the ends (negative moment) and under the loading point (positive moment) as shown in Fig. 2. The bending stresses generated in the beam are proportional to the moments and are compressive or tensile about the neutral axis (line of zero stress).

The maximum tensile stress (σ_b, which is the bending strength or fracture stress) is produced on the top surface at both the ends and is given by [46]

$$\sigma_b = \frac{F_{\max} \ell e_1}{8I} \tag{7}$$

where $F_{\max}$ is the applied load at failure, and e_1 is the distance of the top surface from the neutral plane of the beam cross section and is given by [46]

$$e_1 = \frac{t(w_1 + 2w_2)}{3(w_1 + w_2)} \tag{8}$$

Although the moment value at the center of the beam is the same as at the ends, the tensile stresses at the center (generated on the bottom surface) are less than those generated at the ends (per Eq. 7) because the distance from the neutral axis to the bottom surface is less than e_1. This is because of the trapezoidal cross section of the beam, which results in the neutral axis being closer to the bottom surface than the top (Fig. 2).

In the preceding analysis, the beams were assumed to have fixed ends. However, in the nanobeams used by Sundararajan and Bhushan [41], the underside of the beams was pinned over some distance on either side of the span. Hence a finite element model of the beams was created to see if the difference in the boundary conditions affected the stresses and displacements of the beams. It was found that the difference in the stresses was less than 1 %. This indicates that the boundary conditions near the ends of the actual beams are not that different from that of fixed ends. Therefore the bending strength values can be calculated from Eq. 7.

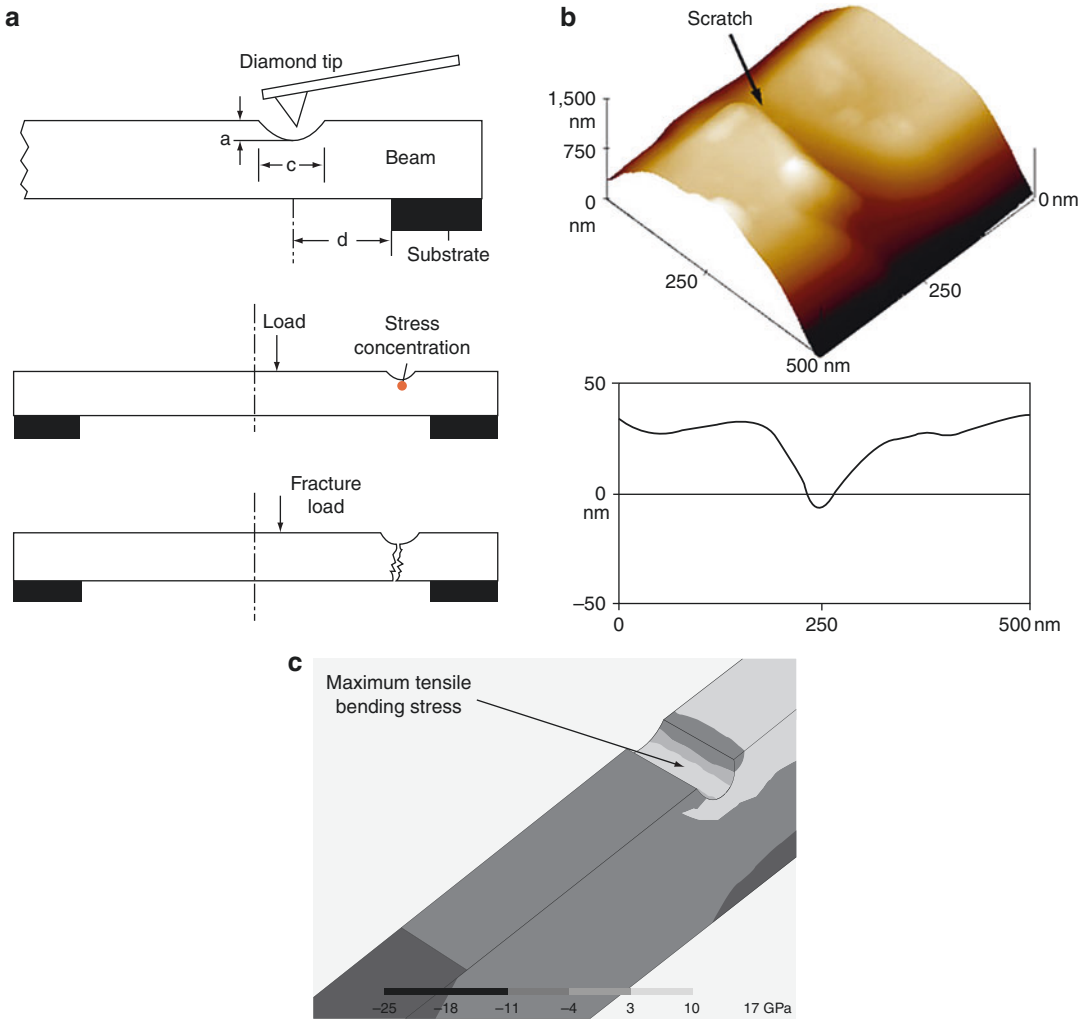
Fracture Toughness

Fracture toughness is another important parameter for brittle materials such as silicon. In the case of the nanobeam arrays, these are not best suited for fracture toughness measurements because they do not possess regions of uniform stress during bending. Sundararajan and Bhushan [41] developed a methodology, and its steps are outlined schematically in Fig. 3a. First, a crack of known geometry is introduced in the region of maximum tensile bending stress, i.e., on the top surface near the ends of the beam. This is achieved by generating a scratch at high normal load across the width (w_1) of the beam using a sharp diamond tip (radius < 100 nm). A typical scratch thus generated is shown in Fig. 3b. By bending the beam as shown, a stress concentration will be formed under the scratch. This will lead to failure of the beam under the scratch once a critical load (fracture load) is attained. The fracture load and relevant dimensions of the scratch are input into the FEM model, which is used to generate the fracture stress plots. Figure 3c shows an FEM simulation of one such experiment, which reveals that the maximum stress does occur under the scratch.

If one assumes that the scratch tip acts as a crack tip, a bending stress will tend to open the crack in Mode *I*. In this case, the stress field around the crack tip can be described by the stress intensity parameter K_I (for Mode *I*) for linear elastic materials [20]. In particular the stresses corresponding to the bending stresses are described by

$$\sigma = \frac{K_1}{\sqrt{2\pi r}} \cos\left(\frac{\theta}{2}\right)\left[1 + \sin\left(\frac{\theta}{2}\right)\sin\frac{3\theta}{2}\right] \tag{9}$$

for every point $p(r,\theta)$ around the crack tip as shown in Fig. 4. If the fracture stress (σ_f) is substituted into the left hand side of Eq. 9, then the K_I value can be substituted by its critical value, which is the fracture toughness K_{IC}. Now, the fracture stress can be determined for the point ($r = 0$, $\theta = 0$), i.e., right under the crack tip as explained above. However, $r = 0$ cannot be substituted in Eq. 9. The alternative is to substitute a value for r which is as close to zero as possible. For silicon, a reasonable number is the distance between neighboring atoms in the (111) plane, the plane along which silicon exhibits the lowest fracture energy. This value was calculated from silicon unit cell dimensions of 0.5431 nm [1] to be 0.4 nm (half of the face diagonal). This assumes that Si displays no plastic zone around the crack tip, which is reasonable since in tension, silicon is not known to display much plastic deformation at room temperature. Sundararajan and Bhushan [41] used values $r = 0.4$–1.6 nm (i.e., distances up to four times the distance between the nearest neighboring



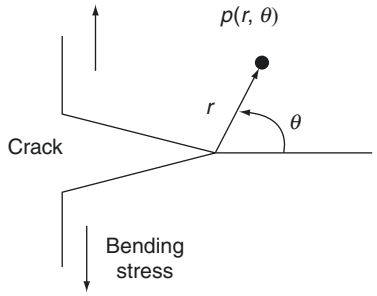
Nanomechanical Properties of Nanostructures, Fig. 3 (**a**) Schematic of technique to generate a defect (crack) of known dimensions in order to estimate fracture toughness. A diamond tip is used to generate a scratch across the width of the beam. When the beam is loaded as shown, a stress concentration is formed at the bottom of the scratch. The fracture load is then used to evaluate the stresses using FEM. (**b**) AFM 3-D image and 2-D profile of a typical scratch. (**c**) Finite element model results verifying that the maximum bending stress occurs at the bottom of the scratch [41]

atoms) to estimate the fracture toughness for both Si and SiO_2 according to the following equation:

$$K_{IC} = \sigma_f \sqrt{2\pi r} \quad r = 0.4 \text{ to } 1.6\,\text{nm} \qquad (10)$$

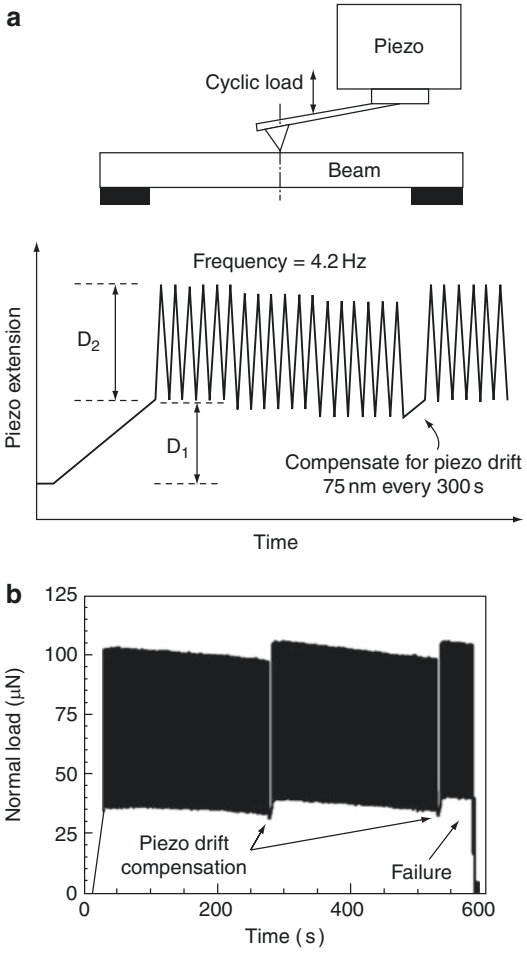
Fatigue Strength

In addition to the properties mentioned so far that can be evaluated from quasi-static bending tests, the fatigue properties of nanostructures are also of interest. This is especially true for MEMS/NEMS involving vibrating structures such as oscillators and comb drives and hinges in digital micromirror devices. To study the fatigue properties of the nanobeams, Sundararajan and Bhushan [41] applied monotonic cyclic stresses using an AFM, Fig. 5a. Similar to the bending test, the diamond tip is first positioned at the center of the beam span. In order to ensure that the tip is always in contact with the beam (as opposed to impacting it), the piezo is first extended by a distance D_1,



Nanomechanical Properties of Nanostructures, Fig. 4 Schematic of crack tip and coordinate systems used in Eq. 9 to describe a stress field around the crack tip in terms of the stress intensity parameter, K_I [41]

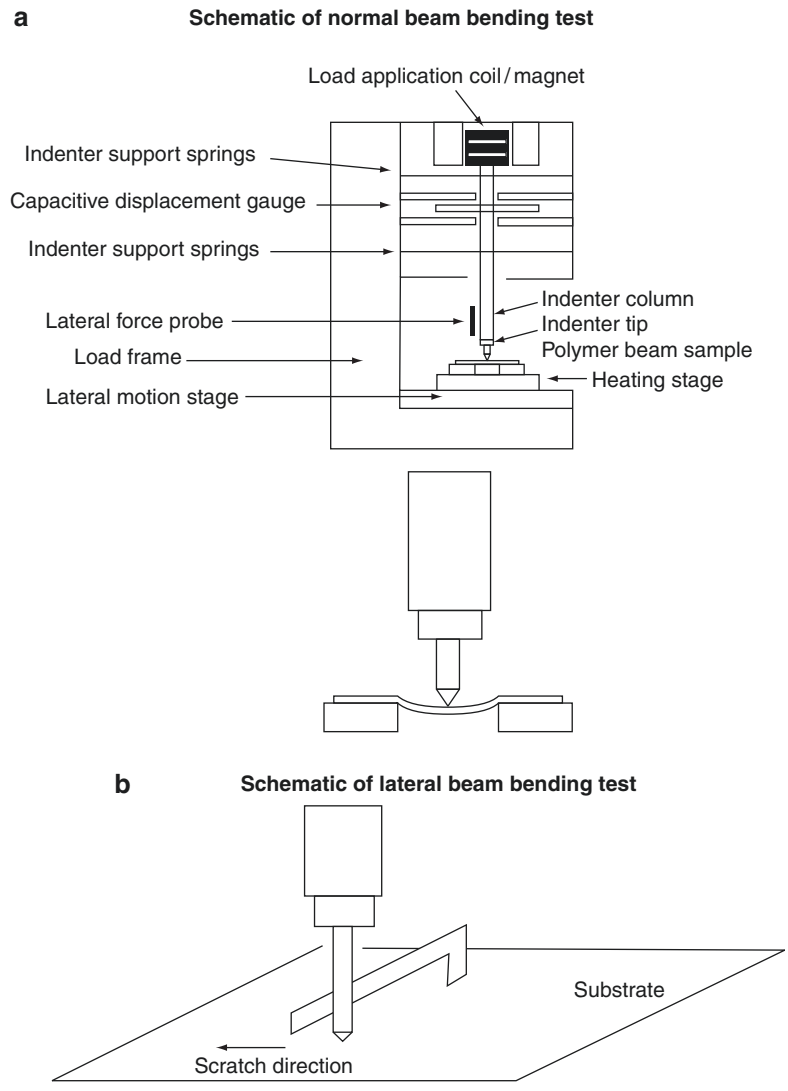
which ensures a minimum stress on the beam. After this extension, a cyclic displacement of amplitude D_2 is applied continuously until failure of the beam occurs. This results in the application of a cyclic load to the beam. The maximum frequency of the cyclic load that could be attained using the AFM by Sundararajan and Bhushan [41] was 4.2 Hz. The vertical deflection signal of the tip is monitored throughout the experiment. The signal follows the pattern of the piezo input up to failure, which is indicated by a sudden drop in the signal. During initial runs, piezo drift was observed that caused the piezo to gradually move away from the beam (i.e., to retract), resulting in a continuous decrease in the applied normal load. In order to compensate for this, the piezo is given a finite extension of 75 nm every 300 s as shown in Fig. 5a. This results in keeping the applied loads fairly constant. The normal load variation (calculated from the vertical deflection signal) from a fatigue test is shown in Fig. 5b. The cyclic stress amplitudes (corresponding to D_2) and fatigue lives are recorded for every sample tested. Values for D_1 are set such that minimum stress levels are about 20 % of the bending strengths.



Nanomechanical Properties of Nanostructures, Fig. 5 (**a**) Schematic showing the details of the technique to study fatigue behavior of the nanobeams. The diamond tip is located at the middle of the span and a cyclic load at 4.2 Hz is applied to the beam by forcing the piezo to move in the pattern shown. An extension is made every 300 s to compensate for the piezo drift to ensure that the load on the beam is kept fairly constant. (**b**) Data from a fatigue experiment on a nanobeam until failure. The normal load is computed from the raw vertical deflection signal. The compensations for piezo drift keep the load fairly constant [41]

Bending Tests of Micro/Nanostructures Using a Nanoindenter

Quasi-static bending tests of micro/nanostructures in both normal and lateral directions are carried out using a nanoindenter (Fig. 6) [27, 28, 35, 45]. The advantage of a nanoindenter is that loads up to about 400 mN, higher than that in AFM (up to about 100 μN), can be used for structures requiring high loads for experiments. Bending experiments in the normal direction have been carried out on the suspended beams using the nanoindenter fitted with a conical tip that has a 1 μm radius of curvature and a 60° included angle. Figure 6a shows the schematic of the beam bending experimental setup. To avoid the indenter tip



Nanomechanical Properties of Nanostructures, Fig. 6 Schematic of micro/nanoscale bending experimental setup using a nanoindenter (**a**) in the normal direction, and (**b**) in the lateral direction

pushing into the specimen, a blunt tip is used in the bending and fatigue tests. For ceramic and metallic beam samples, Li et al. [28] used a diamond conical indenter with a radius of 1 μm and an included angle of 60°. For polymer beam samples, Wei et al. [45] and Palacio et al. [35] reported that the diamond tip penetrated the polymer beams easily and caused considerable plastic deformation during the bending test, which led to significant errors in the measurements. To avoid this issue, the diamond tip was dip-coated with PMMA (about 1–2 μm thick) by dipping the tip in the 2 % PMMA (wt/wt) solution for about 5 s. Load position used was at the center of the span for the bridge beams and at 10 μm off from the free end of cantilever beams. An optical microscope with a magnification of 1500× or an in situ AFM is used to locate the loading position. Then the specimen is moved under the indenter location with a resolution of about 200 nm in the longitudinal direction and less than 100 nm in the lateral direction. Using the analysis presented earlier, elastic modulus and bending strength of the beams can be obtained from the load–displacement curves [28, 35, 45]. For fatigue tests, an oscillating load is applied and contact stiffness is measured during the tests. A significant drop in the contact stiffness during

the test is a measure of the number of cycles to failure [27].

A schematic of the lateral bending experimental setup in shown in Fig. 6b. The nanoindenter was fitted with a conical tip which has a 1 μm radius of curvature and a 60° included angle. The tip was positioned at 200–300 μm away from the beam anchor and 20 μm from the edge of the beam length. The indenter was programmed to perform a scratch at a constant load of 400 μN, and the stage was moved at a rate of 10 μm/s. Since the lateral force contribution of the substrate surface is constant, any additional force recorded comes from the cantilever beam response. From the linear elastic regime in the measured lateral force as a function of lateral displacement, the elastic modulus can be evaluated. For a cantilever beam with one end clamped, it is expressed as [46]

$$E = \frac{\ell^3}{3I}\mathrm{m} \tag{11}$$

where, as described earlier, ℓ is the beam length, I the area moment of inertia for the beam cross section, and m is the slope of the linear region of the force–displacement curve.

Closure

Mechanical properties of nanostructures are necessary in designing realistic MEMS/NEMS and BioMEMS/BioNEMS devices. These can be measured using an AFM or nanoindenter. Most mechanical properties are scale dependent.

Cross-References

► Nanotechnology

References

1. Anonymous.: Properties of silicon. In: EMIS Datareviews Series No. 4. INSPEC, Institution of Electrical Engineers, London
2. Becker, H., Lacascio, L.E.: Polymer microfluidic devices. Talanta **56**, 267–287 (2002)
3. Beebe, D.J., Mensing, G.A., Walker, G.M.: Physics and applications of microfluidics in biology. Annu. Rev. Biomed. Eng. **4**, 261–286 (2002)
4. Bhushan, B.: Tribology Issues and Opportunities in MEMS. Academic, Dordrecht (1998)
5. Bhushan, B.: Handbook of Micro/Nanotribology, 2nd edn. CRC Press, Boca Raton (1999)
6. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
7. Bhushan, B.: Macro- and microtribology of MEMS materials. In: Bhushan, B. (ed.) Modern Tribology Handbook, pp. 1515–1548. CRC Press, Boca Raton (2001)
8. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
9. Bhushan, B.: Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
10. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
11. Bhushan, B., Li, X.: Nanomechanical characterization of solid surfaces and thin films. Int. Mater. Rev. **48**, 125–164 (2003)
12. Cheng, J., Kricka, L.J. (eds.): Biochip Technology. Harwood Academic, Philadelphia (2001)
13. Dresselhaus, M.S., Dresselhaus, G., Avouris, P.H.: Carbon Nanotubes – Synthesis, Structure, Properties and Applications. Springer, Berlin (2001)
14. Drexler, K.E.: Nanosystems: Molecular Machinery, Manufacturing and Computation. Wiley, New York (1992)
15. Elwenspoek, M., Wiegerink, R.: Mechanical Microsensors. Springer, Berlin (2001)
16. Fujimasa, I.: Micromachines: A New Era in Mechanical Engineering. Oxford University Press, Oxford, UK (1996)
17. Gad-el-Hak, M.: The MEMS Handbook. CRC Press, Boca Raton (2002)
18. Goddard, W.A., Brenner, D.W., Lyshevski, S.E., Iafrate, G.J. (eds.): Handbook of Nanoscience, Engineering, and Technology. CRC Press, Boca Raton (2002)
19. Heller, M.J., Guttman, A. (eds.): Integrated Microfabricated Biodevices. Marcel Dekker, New York (2001)
20. Hertzberg, R.W.: Deformation and Fracture Mechanics of Engineering Materials, 3rd edn, pp. 277–278. Wiley, New York (1989)
21. Hierlemann, A.: Integrated Chemical Microsensor Systems in CMOS Technology. Springer, Berlin (2005)
22. Hsu, T.R.: MEMS and Microsystems: Design and Manufacture. McGraw-Hill, Boston (2002)
23. Kovacs, G.T.A.: Micromachined Transducers Sourcebook. WCB McGraw-Hill, Boston (1998)
24. Lai Poh San, C., Yap, E.P.H. (eds.): Frontiers in Human Genetics. World Scientific, Singapore (2001)
25. Lanza, R.P., Langer, R., Vacanti, J. (eds.): Principles of Tissue Engineering. Academic, San Diego (2000)
26. Lawn, B.R., Evans, A.G., Marshall, D.B.: Elastic/Plastic indentation damage in ceramics: the median/radial system. J. Am. Ceram. Soc. **63**, 574 (1980)

27. Li, X., Bhushan, B.: Fatigue studies of nanoscale structures for MEMS/NEMS applications using nanoindentation techniques. Surf. Coat. Technol. **163–164**, 521–526 (2003)
28. Li, X., Bhushan, B., Takashima, K., Baek, C.W., Kim, Y.K.: Mechanical characterization of micro/nanoscale structures for MEMS/NEMS applications using nanoindentation techniques. Ultramicroscopy **97**, 481–494 (2003)
29. Madou, M.: Fundamentals of Microfabrication: The Science of Miniaturization, 2nd edn. CRC Press, Boca Raton (2002)
30. Manz, A., Becker, H. (eds.): Microsystem Technology in Chemistry and Life Sciences, Topics in Current Chemistry, vol. 194. Springer, Heidelberg (1998)
31. Mastrangelo, C.H., Becker, H. (eds.).: Microfluidics and BioMEMS. In: Proceedings of SPIE, vol. 4560. SPIE, Bellingham (2001)
32. Muller, R.S., Howe, R.T., Senturia, S.D., Smith, R.L., White, R.M.: Microsensors. IEEE, New York (1990)
33. Nalwa, H.S. (ed.): Nanostructures Materials and Nanotechnology. Academic, San Diego (2002)
34. Oeberg, P.A., Togawa, T., Spelman, F.A.: Sensors in Medicine and Health Care. Wiley, New York (2004)
35. Palacio, M., Bhushan, B., Ferrell, N., Hansford, D.: Nanomechanical characterization of polymer beam structures for BioMEMS applications. Sensors Actuators A **135**, 637–650 (2007)
36. Park, K. (ed.): Controlled Drug Delivery: Challenges and Strategies. American Chemical Society, Washington, DC (1997)
37. Poole, C.P., Owens, F.J.: Introduction to Nanotechnology. Wiley, Hoboken (2003)
38. Rietman, E.A.: Molecular Engineering of Nanosystems. Springer, New York (2001)
39. Schulte, T.H., Bardell, R.L., Weigl, B.H.: Microfluidic technologies in clinical diagnostics. Clin. Chim. Acta **321**, 1–10 (2002)
40. Senturia, S.D.: Microsystem Design. Academic, Boston (2000)
41. Sundararajan, S., Bhushan, B.: Development of AFM-based techniques to measure mechanical properties of nanoscale structures. Sensors Actuators A **101**, 338–351 (2002)
42. Tang, W.C., Lee, A.P.: Defense applications of MEMS. MRS Bull. **26**, 318–319 (2001)
43. Timp, G. (ed.): Nanotechnology. Springer, New York (1999)
44. van der Berg, A. (ed.): Lab-on-a-Chip: Chemistry in Miniaturized Synthesis and Analysis Systems. Elsevier, Amsterdam (2003)
45. Wei, G., Bhushan, B., Ferrell, N., Hansford, D.: Microfabrication and nanomechanical characterization of polymer MEMS for biological applications. J. Vac. Sci. Technol. A **23**, 811–819 (2005)
46. Young, W.C., Budynas, R.G.: Roark's Formulas for Stress and Strain. McGraw-Hill, New York (2002)

Nanomechanical Resonant Sensors and Fluid Interactions

Rustom B. Bhiladvala
Department of Mechanical Engineering,
University of Victoria, Victoria, BC, Canada

Definition

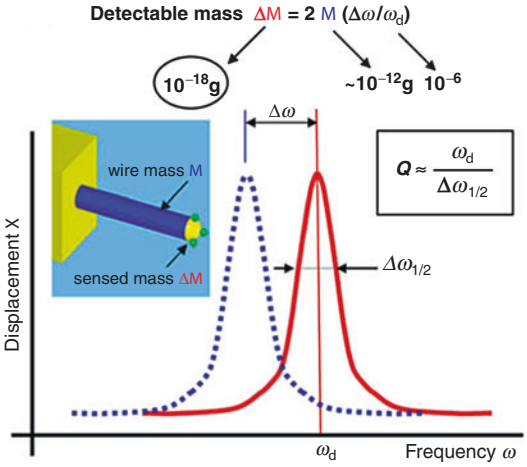
Nanomechanical resonators (or nanoresonators, for this entry) are typically simple nanoscale objects such as rods, beams of rectangular cross-section, grids, or hollow tubes, set into mechanical vibration around a resonance frequency. This entry focuses on ambient fluid (gas or liquid) interaction with nanoresonator-based sensors.

Introduction

Nanoresonators, often referred to as examples of NEMS (Nano Electro Mechanical Systems) devices, are of technological interest because of their potential for extraordinary sensitivity as mass and force measurement devices. To convey this sensitivity to the nonspecialist, a relevant unit for measurable added mass is a zeptogram ($1\ \text{zg} = 10^{-21}$ g), the mass of about 33 water molecules, or a billionth part of one trillionth of a typical 0.1 ml water droplet seen in a household faucet; a relevant unit for measurable force is an attonewton ($1\ \text{aN} = 10^{-18}$ N), equivalent to the weight of one millionth part of one billionth of such a droplet. In the last half-decade, published reports of zeptogram mass resolution [1] and attonewton force resolution [2], using nanomechanical structures of submicron dimensions, now allow us to anticipate new roles for nanoresonators in healthcare, biological research, surface science, and several interdisciplinary areas across physics, chemistry, and engineering.

How has the scientific community moved to use the extraordinary capabilities of nanomechanical resonators to measure such small forces and mass? Perhaps the most widespread interest expressed has

been in molecular diagnosis – screening for, or quantifying, small concentrations of biomarker molecules of devastating disease states such as cancer and neurodegeneration, with the goal of early intervention to improve prognosis – well before disease symptoms appear. Cantilevered tips of less than 100 nm width have been batch fabricated for metrology of biomolecules and Magnetic Resonance Force Microscopy (MRFM) with 4 nm resolution for imaging single viruses [2]. The role of mechanics in the intriguing and complex world of living cells and tissues is being investigated by cell stiffness and tissue elasticity measurements using small-scale resonators for force microscopy in liquids [3, 4]. In the physical sciences, the quest to correlate the structure and behavior of matter at small scales, occupies a significant fraction of research with nanoresonators. Examples are the quantum behavior of mesoscopic mechanical systems, the understanding of elastic and dissipative processes in solids with a large surface to volume ratio, the description of fluid–solid interaction at small scales, high frequencies and in particular, in gases at moderate rarefaction.



Nanomechanical Resonant Sensors and Fluid Interactions, Fig. 1 Schematic of a cylindrical nanoresonator made by nanowire assembly. The left-shift $\Delta\omega$ of the damped resonance peak frequency ω_d, upon binding of mass, allows attogram-level mass resolution. The quality factor, Q, as defined, may be obtained from the spectrum by measuring the full width at half-peak power, $\Delta\omega_{1/2}$

Two measured quantities of interest in nanoresonator-fluid interaction are the resonance peak frequency ω_d, and the quality factor or Q-factor, (symbol Q), as shown in Fig. 1. Nanowire (NW) resonators, tenths of a micron in diameter and a few microns long, typically have resonance frequencies in the MHz-range and small masses (pg), which allow attogram-level resolution of added mass. Measurement of damping force uses the quality factor, Q, as defined in Fig. 1. The Q-factor is the most relevant quantity for fluid-nanoresonator interaction studies. Why is nanoresonator interaction with fluids – liquids or gases – of interest? There are two points of view. The first view to emerge was that interaction with fluids is a nuisance in many nanoresonator applications and it is necessary to learn enough about it to reduce or correct for its effects. Resonance experiments showing great promise for sensor performance in a high-vacuum environment, quickly encountered the sobering reality of high fluid damping in air and water. Measurement attempts showed that high fluid damping could reduce accuracy of frequency peak location, decrease signal-to-noise ratio, and, in liquids, make resonance spectra difficult or impossible to detect. Efforts aimed at addressing these problems constitute the bulk of published research in fluid-nanoresonator interaction. The second and more recent view to emerge is that fluid interaction with nanoresonators is worthy of study in its own right. Such studies aim to answer questions of broad applicability in fluid mechanics of external oscillatory flows, in which viscous forces dominate over inertial forces (i.e., flows at low Reynolds numbers, Re). Examples are the material behavior of fluids at very high forcing frequencies, where the timescales can be small compared to relaxation time of the fluid, and flow in moderately rarefied gases, where the derivation of governing equations and boundary conditions to describe the flow are a subject of active research. The fluid mechanics of internal (channel) flows of liquids at micro and nanoscale is acquiring some maturity, with several books now available in microfluidics and nanofluidics. The study of oscillatory flows around nanostuctures is presently seeing a period of vigorous early growth in research publication.

This entry is organized in four sections. The section "Nanoresonator Applications Involving Fluid Interaction", divides nanoresonator-fluid interaction applications into those in which fluid interaction is primarily a performance concern to be reckoned with or is proposed as a sensor principle and those applications where research toward formulating and answering questions in fluid mechanics is the primary goal. As the scope and limitations for both sets of applications is coupled strongly to choices of materials and transduction methods, this coupling is first introduced in the section "Interdisciplinary Development of Nanoresonator Sensor Platforms", but plays a significant role throughout the article. The section "Measurement Relations" deals with the development of the simplest relations required to interpret measurements. The role of geometry, materials, and fabrication schemes in nanoresonator-fluid interaction measurement is discussed in detail in the section "Fabrication and Materials", and a summary and a few reviews with complementary material are presented in the section "Summary and Related Reviews".

Nanoresonator Applications Involving Fluid Interaction

While the potential field of nanoresonator applications is large, this entry is restricted to applications of nanomechanical resonators that involve fluid interaction. It is useful to divide these into two cases: (1) Nanoresonators for technological applications involving fluid interaction and (2) Nanoresonators as fluid damping sensors for research in fluid mechanics. In case (1), an accurate value of the Q-factor is not important, provided it is large enough ($Q >> 1$) and does not change significantly in the course of measurement. A significant decrease in Q-factor as a result of increasing ambient gas pressure can reduce accuracy and signal-to-noise ratio and constitutes a decrease in performance in this case. In case (2), an accurate value of the Q-factor is required, as it constitues the raw data for the measurement of fluid damping. In this case, a larger change in Q-factor for a given change in fluid damping is desirable, as this increases the dynamic range for fluid damping measurement.

Technological Applications Involving Fluid Interaction

The extraordinary mass and force sensitivity of nanoresonators and their size, commensurate with length scales of large macromolecules in living cells, makes them well suited to cellular and molecular studies. For microcantilevers, Lavrik et al. [5] provide a comprehensive review of studies using resonance mode detection, static deflection, or a combination of both for gas and vapor phase analyte sensing. Molecular diagnosis/screening can be enabled by functionalizing nanoresonators for specific binding of target biomarker molecules and using the frequency shift following target binding. Accurate measurements of small mass require a resonant spectrum with adequate signal-to-noise ratio and a sharp resonant peak (high quality factor) for the resonance curve. Since the aqueous environment is native to most cells and biomolecules, this would be most preferred for operating nanoresonator biosensors – but it is also the environment with highest fluid damping, which degrades the overall performance of the resonators. Demonstration of ideas such as the use of higher harmonics of the fundamental resonance frequency [6] or high pretension in nanobeams [7] for viable resonance in fluids are an established stream of research literature focused mainly on reducing the deleterious effects of fluid damping on sensor performance in measurement schemes. Use of the change in damping of a single nanoresonator or the resonance properties of a group of fluid-coupled nanoresonators by the binding of biomolecular analytes have also been proposed [8]. The development of scanned probe microscopies for mapping and understanding the structure of biomolecules is an important research area. Q-control techniques have been developed to offset the deleterious effects of fluid damping in scanned force microscopies.

Sensing for Fluid Mechanics Research

Nanoresonator interaction with gases of varying density can span rarefaction regimes from

continuum flow to highly rarefied (free-molecular) flow. While both these limits and quantitative approaches to them are familiar, neither free-molecular nor standard continuum calculation methods are valid [9] in the intermediate transition regime of rarefied gas flow. NWs can serve as measuring instruments of gas damping forces from atmospheric pressure to a modest vacuum ~10^{-3} atm and spanning Knudsen numbers (gas molecule mean free path/wire diameter) in the range of $0.1 < \mathrm{Kn} < 100$ which includes the transition and free-molecular flow regimes of rarefied gas dynamics. Neither the continuum fluid approach nor simple models for molecular momentum transport can be trivially extended to work in the transition regime – the development of theoretical and computational models in this regime is currently a research area of significant interest [9–11]. Gas damping force measurements with nanowires can provide data for these models.

A theoretical quantitative description allowing accurate calculation in the transition regime must blend smoothly into known solutions for the free-molecular and continuum regimes. For fluid which may be regarded as a continuum, what is the effect of forcing at high frequencies, or times small compared to any relaxation (i.e., adjustment) time within the fluid? For a smooth blend at the continuum end, the large span in frequencies (and Reynolds numbers, Re) can require either steady or unsteady forms of the Stokes equation, and for a few exceptional high-frequency cases, the usually appropriate neglect of inertial effects represented by the nonlinear term in the Navier–Stokes equations is not valid (i.e., $\mathrm{Re} > 1$). How should the physics of unsteadiness and inertia be modeled in the transition regime of rarefaction? How should transition regime calculations for geometries more complex than beams be approached? What are the appropriate length scales to be used in determining the boundaries of rarefaction regimes? How do the results scale with Mach number (nanoresonator velocity/gas molecule velocity)? Fundamental insights into fluid mechanics at small scales and high frequencies can be gained by comparing results from proposed models with those from tabletop experiments using oscillatory nanoresonator flows.

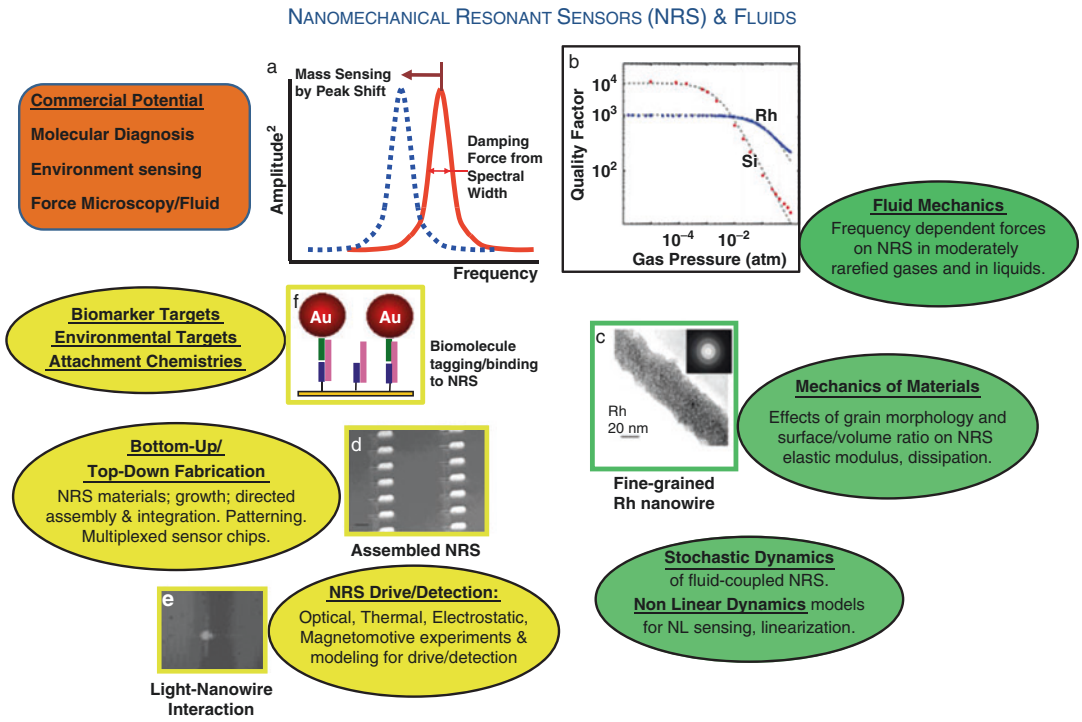
While early experimental demonstrations of viable micro or nanoresonator operation in fluids [5, 6] have provided some data suitable for a study of resonator-fluid interaction, experimental studies focused solely on understanding fluid interaction are now beginning to emerge [12, 13]. Better answers to some of the questions above are being proposed, as members of computational, analytical, and experimental subcommunities undertake collaborative work [10, 14] to improve understanding of fluid mechanics at small scales, high frequencies, and in regimes of gas rarefaction that are not yet well described. These fundamental studies aim to create the analytical and computational framework, which in turn will enable a diverse range of useful applications. Such applications range from the very small – for example, sensitive nanoscale instruments for understanding biology and the physical sciences, building sensors for healthcare, engineering and technology – to the macroscale – for example, drag force and trajectory calculations of space vehicle or satellite fragments following atmospheric entry.

The potential to create extraordinarily sensitive instruments from nanoresonators is easy to see in principle, but the path to realization of such sensor platforms is paved with challenges. Some of these challenges are introduced and explained with an example in the following section.

Interdisciplinary Development of Nanoresonator Sensor Platforms

The challenge in measurements is that the high surface to volume ratio of nanostructures can cause an unusual and strong coupling of physical effects. Sensor platform development must typically reckon with such coupling between transduction (vibration detection and driving) schemes, intrinsic stresses from fabrication, and nonlinear geometric effects [15]. While the focus here is on nanoresonators and fluid mechanics, note that the actual conduct of work for technological applications requires making several choices in other areas that are strongly coupled and cannot be ignored.

A representative view of these areas is shown in Fig. 2. A few partially realized examples of



Nanomechanical Resonant Sensors and Fluid Interactions, Fig. 2 One view of the scope of research for nanomechanical resonant sensors (*NRS*) involving fluid interaction. Commercial potential (*top left*) and experimental methods development (three ellipses at *left, yellow*) have served as drivers. Fields of engineering science which support and are enriched by such research, are shown on the *right* (three ellipses, *green*) ((**c**, **d**) reproduced with permission, from Refs. [16, 17])

commercial potential are listed at the top left of the figure. Experimental research areas for engineering and technological development are shown in the three ellipses on the left. What are the sensor targets – for example, which gases, biological macromolecules, or cell components? How can attachment chemistry be developed to help functionalize a nanomechanical resonator (NR) – that is, to give it surface properties that will enable specific capture of only the selected target(s)? The attachment chemistry is dependent on the choice of NR material. The material choice may be limited by the fabrication approach used. Choice of the fabrication approach, in turn, may alter the ability for creating multiplexed sensors (capable of sensing different targets on the same chip). What schemes are suitable for driving and detecting resonance for these small high-frequency devices? Each method listed in Fig. 2 is associated with advantages, and frequently, with measurement artifacts. Experiments and theoretical or computational analysis are often required to quantify the artifacts. Fundamental understanding in the disciplines of Fluid Mechanics, Mechanics of Materials, and Nonlinear Dynamical Systems (right half of figure) guides this engineering and technological work. Many aspects of this understanding need to be improved. A strong application for nanoresonators in fluids is thus to help improve the framework of areas of Fluid Mechanics that are not well understood. For example, what happens when a shear disturbance (vorticity) is introduced in a fluid at time scales far smaller than the time taken for viscosity to smoothen it out (viscous relaxation time)? Nanoresonator-fluid experiments are especially suitable for this question – they can provide very small forcing timescales (very high frequencies) and are typically associated with small fluid inertia (low

Reynolds number) so that the nonlinear inertial effects are associated negligible.

As an example of iterative work due to interdisciplinary coupling shown in Fig. 2, consider a long, thin, doubly-clamped resonant nanobeam made from a metal film, with high intrinsic stress from a top-down fabrication process. Suppose the metal was chosen because the attachment chemistry for molecules enabling specific capture of a desired detection target molecule is well understood. During optical measurement of the motion of the doubly-clamped nanobeam in a high damping gas environment, one may turn up the power of the measuring laser to improve signal strength at low vibration amplitude. Experiments and modeling (nonlinear dynamics, mechanics of materials) have shown this can cause periodic heating of the vibrating beam and a thermomechanical coupling which has been shown to introduce softening and hardening nonlinearity as a function of laser power [15]. To circumvent this problem arising from intrinsic axial stress in the metal film, a similar long, thin, cantilevered metal beam (no axial stress) will be a necessary replacement. Such beams are difficult to make with a top-down technique, spurring efforts in the use of alternate fabrication techniques to make such structures.

This example shows that the conduct of work as well as the correct interpretation of data needs continual interaction between different disciplines to sort out artifacts arising from the coupling of physical effects, materials, fabrication schemes, and transduction methods. It is important to be mindful of this coupling while approaching the following section of relations used in the design of experiments and the interpretation of data.

Measurement Relations

This section presents different definitions of the Q-factor (Fig. 1) based on material properties, energy, and on spectral or time-domain measurements of resonator amplitude. The distinctions between definitions are usually and justifiably ignored if the Q-factor is very large compared to unity, but can be significant for fluid-nanoresonator interaction with high damping. These relations and the relations for the resonance frequency also provide a scaling guide for geometry and material choice to tune the range of Q-factor for a given range of fluid damping.

When the modes of vibration are sufficiently separated, a single degree of freedom, harmonically driven, linear resonator model is reasonable for understanding key features of fluid-nanoresonator interaction. The damped resonance peak frequency ω_d and the dimensionless quality factor (Q-factor, or Q), which allows the measurement of dissipation through the sharpness of the resonance peak, are the most widely used quantities. Most instruments based on resonators avail of simplifications that are possible when the quality factor is large ($Q >> 1$). High fluid damping produces systems with low Q-factor, leading to errors from the use of these simplifications. It is helpful to gather the concepts and relations necessary to maintain the accuracy of calculations for resonators where Q may not be much larger than 1. A summary is included later in this section.

The displacement with time, $\{x(t)\}$, of a resonator with mass M, damping C, and stiffness K, driven harmonically with a forcing amplitude F_0 at a frequency ω, as seen in Fig. 3, is

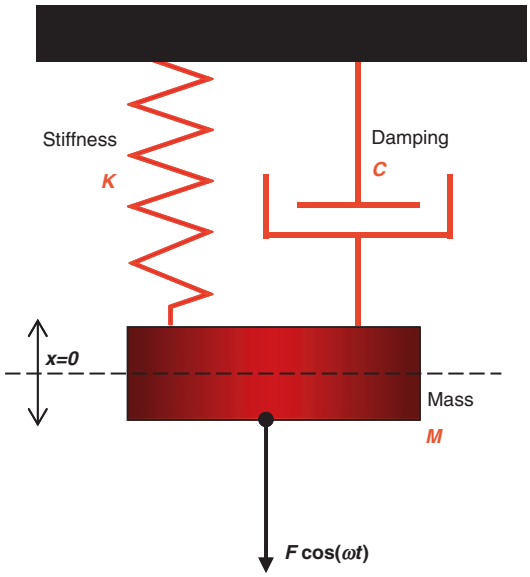
$$M\ddot{x} + C\dot{x} + Kx = F_0 \cos(\omega t) \tag{1}$$

A surrogate form of the Q-factor, defined using material properties alone, is the dimensionless ratio Q_p, given by $Q_p = \sqrt{MK}/C = M\omega_0/C$ where $\omega_0 = \sqrt{K/M}$ is the undamped resonance frequency. Q_p is related to the commonly used damping ratio ζ through $\zeta = C/C_c = 1/(2Q_p)$, where $C_c = 2\sqrt{MK}$ is the critical value of damping. Damping decreases the resonance frequency ω_d, from the undamped resonance value ω_0, as given by the relation,

$$\omega_d = \omega_0\left[1 - 1/\left(2Q_p^2\right)\right]^{\frac{1}{2}} = \omega_0\left[1 - 2\zeta^2\right]^{\frac{1}{2}} \tag{2}$$

For steady-state resonance, the mean power supplied by the driving force balances the energy dissipation rate, and the total stored kinetic energy and spring potential energy, averaged over many

N



Nanomechanical Resonant Sensors and Fluid Interactions, Fig. 3 Single degree of freedom model for resonator

cycles, does not change with time. With displacement amplitude x_0, the cycle-averaged value for the displacement-squared, $\langle x^2 \rangle = (x_0^2/2)$ and velocity-squared $\langle \dot{x}^2 \rangle = (\omega^2 x_0^2/2)$. The mean power dissipated, $<P>$, the average energy dissipated per cycle $< E_{dissip} >$ and the stored energy $< E_{stored} >$ are related [18] by

$$\langle E_{dissip} \rangle = \langle P \rangle (2\pi/\omega) = C < \dot{x}^2 > \left(\frac{2\pi}{\omega}\right) = \pi C \omega x_0^2 \tag{3a}$$

$$\langle E_{stored} \rangle = M(\omega^2 + \omega_0^2) x_0^2/4. \tag{3b}$$

If the average energy stored in the resonator is very large compared to the energy dissipated (= energy supplied, for steady-state) per cycle, one may anticipate lower noise operation when driven, and a slower decay of the oscillation amplitude envelope when the driving force is switched off, allowing for a time-domain measurement of damping. These are measures of the quality of a resonator, expressed by the magnitude of the quality factor, defined [18] as

$$Q = 2\pi \frac{\langle E_{stored} \rangle}{\langle E_{dissip} \rangle} \tag{4}$$

Using Eqs. 3a and 3b, this yields

$$Q = \frac{M(\omega_0^2 + \omega^2)}{2C\omega} = \frac{K + M\omega^2}{2C\omega}. \tag{5}$$

For forcing at the damped resonance frequency ω_d, using Eq. 5 with the definition for Q_p, $\omega = \omega_d$, with ω_d given by Eq. 2, the expression for the quality factor reduces to

$$Q = \frac{Q_p^2 - (1/4)}{\sqrt{Q_p^2 - (1/2)}}. \tag{6}$$

Most applications aim for quality factors $Q >> 1$, for which ω_d reduces to ω_0 and Q to Q_p, as seen from Eqs. 2 and 6. It is also easy to show [18] that for high Q-factors, the measured resonance spectrum, expressed as a plot of amplitude-squared against forcing frequency, has a full width at half maximum (FWHM) height, as seen in Fig. 1, equal to damping coefficient divided by mass, C/M, which allows for a simple measurement of damping.

The ringdown method is an alternative to the spectral technique above. It is a time-domain technique for measuring damping. In this method, the forcing for a driven nanoresonator is turned off, and the amplitude of the oscillation begins to decay with time t as $\exp(-t/\tau)$, where the time constant $\tau = C/M = \omega_0/Q_p$ [18]. Q_p can be obtained from τ, the time taken for the measured amplitude envelope to reduce to 1/e of the previous value.

Nanoresonator operation in a fluid environment is typically associated with high damping (C). Some consequences of low Q_p due to high damping may now be examined. Low Q-factor spectra, $Q \sim 1$, have been encountered in measurement attempts in liquids. For low Q-spectra, the spectral width (FWHM) may be hard to obtain and the spectra are noisier – the signal-to-noise ratio, SNR $\sim Q^{1/2}$. Using the best-fit Lorentzian curve solution of Eq. 1 to low-Q experimental spectra is preferable for the evaluation of the

Nanomechanical Resonant Sensors and Fluid Interactions, Table 1 Significant deviations at low Q_p

Q_p	ζ	$(Q - Q_p)/Q_p$	$(\omega_d - \omega_0)/\omega_0$
1,000	0.0005	3.12E-14	−2.50E-07
100	0.0050	3.13E-10	−2.50E-05
10	0.0500	3.14E-06	−2.50E-03
5	0.1000	5.10E-05	−1.01E-02
2	0.2500	2.22E-03	−6.46E-02
1	0.5000	6.07E-02	−2.93E-01

damping. As seen in Table 1, obtained using Eqs. 2 and 6, the deviations of Q from Q_p and ωd from ω_0, become significant and increase rapidly as Q_p approaches a value of 1. As an example, the fractional shift of the resonance peak frequency due to damping, changes from 0.25 ppm (parts per million) at Q_p of 1,000 to 65,000 ppm at Q_p of 2. Sensitive mass detection aims for resolution of frequency shifts of the order of 10 ppm or less, between mass-loaded and unloaded measurement states. For $Q_p \sim 1$, small variations in fluid damping between these states could themselves cause frequency shifts that will exceed the frequency shift from small added mass. The experimental evaluation of the Q-factor (sketch in Fig. 1) shows that the accuracy of any sensing scheme dependent on location of the peak frequency will also degrade significantly for broad peaks (low Q). In liquids and dense gases, there are cases where fluid damping can be high enough to prevent the detection of the resonance peak altogether. Ringdown measurements also become less accurate or impossible, if Q is low enough to reduce the displacement amplitude to zero in a time comparable to one oscillation time period. Finally, when an electrostatic or piezoelectric drive is used for exciting and driving the nanostructure through resonance, the large range of swept frequency required to capture low Q-factor spectra, has the potential to create nonconstant forcing F_0. This arises due to changes in the amplitude of the driving signal, due to significant frequency-dependent changes in the impedance of the driving circuit (parasitics).

Techniques for enhancing the Q-factor Q_p of nanoscale resonators in fluids could include the use of resonance at higher frequency harmonics (higher ω_0), and an increase of effective string-like stiffness (higher K, ω_0), by using tensioned beams [7]. These have demonstrated viable resonance with high quality factor in some fluid measurements, and should prove useful for many applications. However, their reduced sensitivity to fluid damping makes them less suitable as measurement devices for well-resolved and accurate measurement of the value of fluid damping, when that is the object of the study.

To complete this section, the experimental steps used for the evaluation of gas damping force on nanomechanical resonators are reviewed. A few results from mechanics of materials and vibration of continuous beams will be useful to see the interplay of geometry, material characteristics, and measurement artifacts related to nanoresonator fabrication and transduction schemes that are characteristic of nanoresonator measurements.

Consider fluid interactions with the nanoresonator that are linear – that is, they are not functions of products of the displacement x or its time derivatives. The nanoresonator motion may then be represented by the equation governing the displacement x,

$$\left(M_s + M_f\right)\ddot{x} + \left(C_s + C_f\right)\dot{x} + \left(K_s + K_f\right)x = F_0 \cos\left(\omega t\right)\cdot \tag{7}$$

In Eq. 7, subscripts s, f denote structural intrinsic (solid material) and fluid quantities respectively. The fluid force on the beam in phase with the velocity is represented using fluid damping Cf; the force on the beam due to the fluid in phase with the acceleration can be significant in liquids, and is represented by the fluid added mass term M_f; a restoring spring force component is represented by a fluid stiffness term K_f. A measurement at high vacuum, where fluid molecules have no appreciable effect, will involve only the intrinsic quantities. In nanoresonators, Eq. 2, with typical $Q_{p,s} >> 1$, based on intrinsic damping C_s, shows that C_s is not large enough to cause an appreciable shift of the resonance peak frequency from the undamped value.

The undamped resonance frequency for transverse flexural vibrations of a solid beam having l, M_s, K_s, ρ_s, E, and I as the length, mass, stiffness,

density, Young's modulus, and cross-sectional moment of inertia, is [19]

$$\omega_s \approx \omega_0 = \sqrt{K_s/M_s} = \beta_n^2\sqrt{EI/M_s l^3} \tag{8}$$

$$= \beta_n^2\sqrt{(1/12)^*(E/\rho_s)^*}\left(t/l^2\right) \tag{8a}$$

$$= \beta_n^2\sqrt{(1/16)^*(E/\rho_s)^*}\left(d/l^2\right) \tag{8b}$$

where Eq. 8a is for a rectangular beam of width w, and thickness t parallel to the direction of motion and Eq. 8b is for a circular section beam (nanowire) of diameter d.

For rigid clamping, β_0 is 1.88 for cantilever beams and 4.73 for doubly-clamped beams vibrating in the fundamental mode ($n = 0$). The surrogate Q-factor for the beam in vacuum is given by,

$$Q_{p,s} = \sqrt{M_s K_s/C_s} = M_s\omega_0/C_s$$
$$= (1/C_s)^*\sqrt{M_s EI/l^3} \tag{9}$$

$$= (1/C_s)^*\beta_n^2\sqrt{(E\rho_s/12^*}\left(wl^2/l\right) \tag{9a}$$

$$= (1/C_s)^*\beta_n^2\sqrt{(E\rho_s/16^*}\left(\pi d^3/4l\right) \tag{9b}$$

with Eq. 9a for the rectangular beam and (Eq. 9b) for the circular section nanowire. If squeeze-film damping is negligible, the effective stiffness in the fluid remains unchanged from its intrinsic value in vacuum, so that $K_{s+f} = K_s = M_s\,\omega_0{}^2$. Substituting for effective mass and damping, $M_{s+f} = M_s + M_f$ and $C_{s+f} = C_s + C_f$, yields the damped resonance frequency ω_{s+f} and quality factor $Q_{p,s+f}$ for the beam in fluid,

$$\omega_{s+f} = \omega_0\sqrt{\left(1 - 1/(2Q_{p,s+f}^2\right)} \tag{10}$$

$$Q_{p,s+f} = \sqrt{\left(M_{s+f}K_s\right)}/C_{s+f}$$
$$= \mathrm{Q}_{p,s}\left[\frac{\left\{1 + \left(M_f/M_s\right)\right\}^{0.5}}{1 + \left(C_f/C_s\right)}\right] \tag{11}$$

For rarefied gas damping experiments with nanoresonators, the situation is usually simplified with fluid added mass M_f being negligible. A first measurement may be made at high vacuum, with no gas damping ($C_f = 0$), for which the Q-factor value, Q_s, depends only on the structural properties of the solid nanowire, M_s, C_s, and K_s. Following this, the ambient gas pressure in the chamber is increased in steps, and the frequency ω_{s+f} and quality factor Q_{s+f} are measured at each step. In experiments, we assume the surrogate property-related Q-factor (Q_p) relations hold true for the measured Q-factor values. This yields

$$Q_s = \frac{M_s\omega_0}{C_s} \tag{12a}$$

$$Q_{s+f} = \frac{M_s\omega_0}{C_s + C_f} \tag{12b}$$

Using Eq. 12a and mass M_s from known density and geometry of nanoresonator, the value of intrinsic damping C_s may be obtained from measurement at high vacuum. Subsequent measurements of the Q-factor Q_{s+f}, at each value of ambient gas pressure, are related using Eqs. 12a, b to obtain

$$Q_{s+f} = Q_s\left(1 + \frac{C_f}{C_s}\right)^{-1}, \tag{13}$$

from which the gas damping C_f can be calculated. For liquids or dense gases, the added fluid mass M_f is not negligible, and the full Eq. 11 must be considered (see [9]), in place of its subset Eq. 13.

These relations provide a summary for the conduct of experimental work and are useful in a number of ways. For measurements in vacuum, they establish the expected scaling of frequency and Q-factor with parameters that depend on geometry, material, mode shape, and type of clamping. They allow evaluation of the rigidity and repeatability of the clamps (through the value of mode constant β_n in Eq. 9) if the geometric parameters and material properties are accurately known. Most importantly, they allow experimental testing of models for fluid damping C_f and fluid added mass M_f. These relations also show how nanoresonator materials may be chosen to

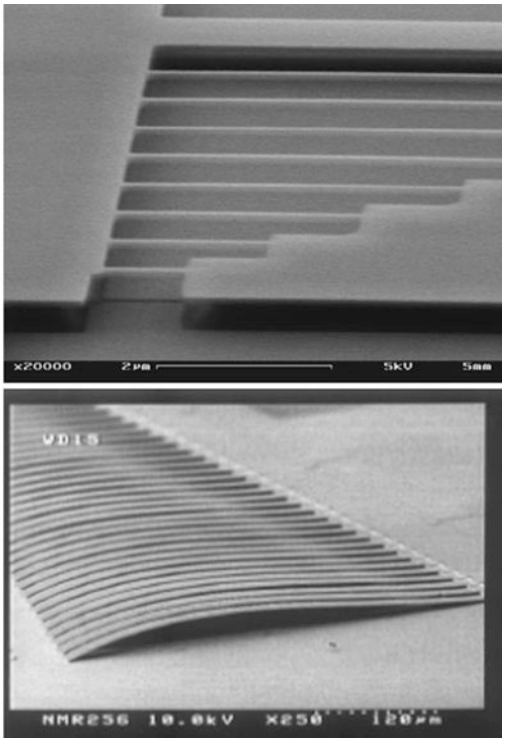
increase or decrease the Q-factor range for a given range of fluid damping. This is discussed in the following section detailing the coupling of fabrication, materials, and transduction choices.

Fabrication and Materials

Nanomechanical resonators are suspended structures – most often long and prismatic, clamped at one end (cantilevered) or both ends (doubly-clamped). How does the choice of geometry, material, fabrication, and transduction schemes influence the operation of nanomechanical resonant sensors in fluids? Earlier discussion of Fig. 2 showed the strong coupling between these choices and the necessity of using the framework of mechanics of fluids, mechanics of materials, and nonlinear/stochastic dynamics as guides to make sensor design choices. Fabrication choice plays a dominant role and it is worth briefly reviewing two fabrication paradigms – termed top-down and bottom-up fabrication and an example of the synergy that may be obtained by hybrid methods combining features of each.

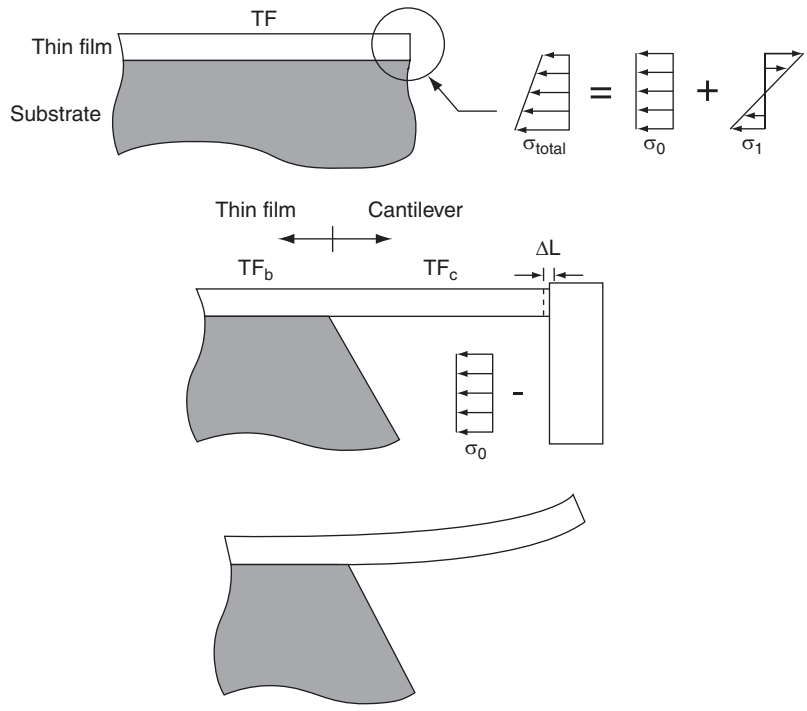
Top-down Methods

Top-down methods are named for the processes of patterning and etching from larger planar sheets of material (top-level) to obtain smaller desired structures (bottom-level), typically as large arrays. No assembly is required as the structures are built in the desired locations. Typically, a sacrificial layer is sandwiched between the substrate and a film of the intended nanoresonator material (device layer). An etch mask, usually made from a photosensitive chemical that resists etching (photoresist) is used while etching through the device layer, leaving it in the shape of the desired resonator array. A second etch targeted at the sacrificial layer material is then used to remove material beneath the devices and create suspended resonators. Doubly-clamped resonators created by this method are shown in Fig. 4. Optical lithography allows transferring micrometer scale features on to large (upto centimeter-scale) areas of photoresist for patterning thousands of devices



Nanomechanical Resonant Sensors and Fluid Interactions, Fig. 4 (**a**) Doubly-clamped nanoresonator beams of different lengths with the same undercut (Figure reprinted from Ref. [20], with permission). (**b**) Distortion of long thin metal cantilevers attempted using a top-down process (Figure reprinted from Ref. [21], with permission)

using a single exposure through a patterned mask. This well-established technique, long used in microelectronics to make integrated circuits (ICs), is easy to implement, and is almost always preferred at microscale. Minimum feature sizes for ICs, limited by the wavelength of light used, have been extended to submicron scale with deep-UV light sources and phase-shift masks. Nanoresonators, involving smaller features, from tens to hundreds of nanometers, typically have been made by electron-beam (e-beam) lithography. Here, no mask is used, but an electron beam is rastered across a film of e-beam resist to create the pattern for nanoscale structures. Time-consuming coverage of large areas required to make commercial sensor arrays, the cost of procuring and operating e-beam equipment, and the fabrication sensitivity to small perturbations do



Nanomechanical Resonant Sensors and Fluid Interactions, Fig. 5 Film stress and stress gradients create residual axial stresses in doubly-clamped beams and distort cantilevered beams, during their release from the substrate, by etching (Figure reprinted from Ref. [22], with permission)

not allow inexpensive production of large-area sensor arrays by e-beam lithography. An alternative pattern transfer technique of imprinting, molding, and stamping known as microcontact printing, or soft lithography, is similar to the use of an inked rubber stamp. This technique, somewhat surprisingly, has been used to demonstrate the ability to transfer pattern features below 100 nm in size. The master pattern in this case may be made by e-beam lithography and transferred to one or more soft "stamp pads," thus reducing the requirements for e-beam lithography, with the promise of lowered costs. Mature techniques, relatively well-controlled shape, size and position, the ability to make devices with well-ordered internal structure through the use of single-crystal layers, are important strengths of top-down fabrication [20].

There are limitations of top-down fabrication for nanoresonator sensor arrays. Besides the cost of large-area coverage, the limited choice of local material selection (e.g., difficult to have resonator material different from clamp material) can restrict functionality. Geometries such as smooth rods or hollow tubes, with a high cross-sectional moment of inertia allowing for a higher stiffness per unit mass, are prohibitively difficult to produce in large numbers by top-down techniques. Undercuts, or significant underetching of supports during the removal of the sacrificial layer [20] are seen as a lighter-gray region around the periphery of the harp-like structure of Fig. 4a. Undercuts depend on the sacrificial layer etch time and are of the same depth for all structures released on a chip. They can cause significant increase in the compliance, particularly for short, stiff, nanoresonators seen in Fig. 4a. Energy losses and cross talk between devices through undercuts can cause significant changes in frequencies and Q-factors. During release by etching of the sacrificial layer, shape distortion (see Fig. 4b) in the case of cantilevered structures [21] or residual axial stress in doubly-clamped nanoresonators (see Fig. 5) typically arises due to stress gradients in the device layer [22]. For long thin metal film structures, these effects are particularly difficult to control. Higher residual axial stress adds an increasing string-like component to the restoring force (normally flexural), changing the simple frequency scaling presented in Eq. 9 and makes the nanoresonator

more prone to geometric stiffness nonlinearities. High axial stress and string stiffness can, however, be used to increase the effective stiffness K and the peak frequency, to enable a higher Q-factor value for given damping, provided a sufficient linear range exists. Distorted cantilevers also cannot be described by Eq. 9 – furthermore, it may be difficult or impossible to drive and detect resonance in such structures. Film stress gradient effects can be controlled in device materials such as single-crystal silicon, silicon nitride or carbide, using silicon oxide as a sacrificial layer. These materials are frequently used for laboratory experiments. However, stress gradient effects (Fig. 5) are very pronounced for metal films. It is therefore difficult for top-down methods to yield long, thin metal cantilevers which are free of distortion – a "convex-up" example [21] of such distortion for titanium cantilevers is seen in Fig. 4b. Though long thin cantilevered resonators made of metals may have desirable surface functionalization attributes, they are difficult to achieve using top-down methods. Finally, for multiplexing, or the sensing of multiple analytes on the same chip, functionalization of different devices to capture different targets is tedious, and one may anticipate the value of nanostructures that can be made and functionalized off-chip, then directed to predetermined areas of a chip and integrated to form nanoresonators.

While top-down methods remain the primary choice for most microcantilever fabrication, in spite of their obvious strengths, many of the shortcomings of established top-down techniques discussed above have led to an interest in alternatives that bottom-up fabrication could provide. This is particularly true for nanoscale resonators, when long thin cantilevers made of metals are required for their attachment chemistries, or when undercut reduction or multiplexed sensor arrays are important.

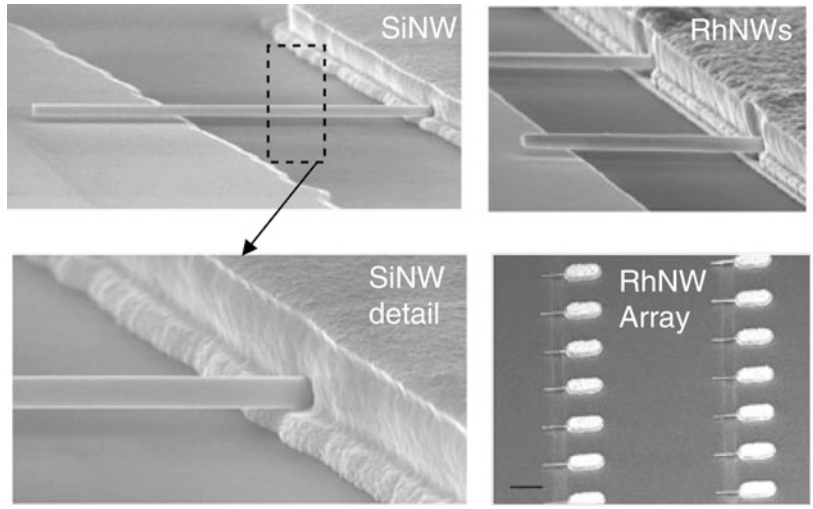
Bottom-Up Methods

Bottom-up methods are identified by their use of assembly of smaller units into larger structures, for example, single nanoresonator elements assembled and integrated into arrays for sensing. While the possibilities for improved control of individual element geometry and clamping are attractive, making arrays of a few tens of thousands of elements by an individual pick and place technique is unfeasible. Bottom-up methods are thus defined by the use of physical or chemical forces for self-assembly or directed assembly of nanostructures. A recent review of the directed assembly of nanowires [23] discusses assembly by molecular forces, electrostatic interactions, shear forces, magnetic fields, and dielectrophoresis. A significant advantage over top-down methods is that special vapor or solution phase synthetic chemistry can be used to tailor the shape, composition, and crystallinity of nanoscale elements, before their assembly and integration on a sensor chip. This greatly widens the scope of material properties (e.g., nanowires made of magnetic, piezoelectric, semiconductor, or multiple-metal segments) and available geometries (e.g., cylindrical or hexagonal section rods, hollow tubes, single and multi-walled carbon nanotubes) that can be put to use. The elements may also be functionalized off-chip to capture different sensing targets on the same chip (multiplexing), with each target-type assembled in a predetermined area on the chip. This requires the development of assembly and integration techniques that do not destroy the capability of molecules used for functionalization.

N

Results Using a Hybrid Fabrication Strategy for Nanoresonator Arrays

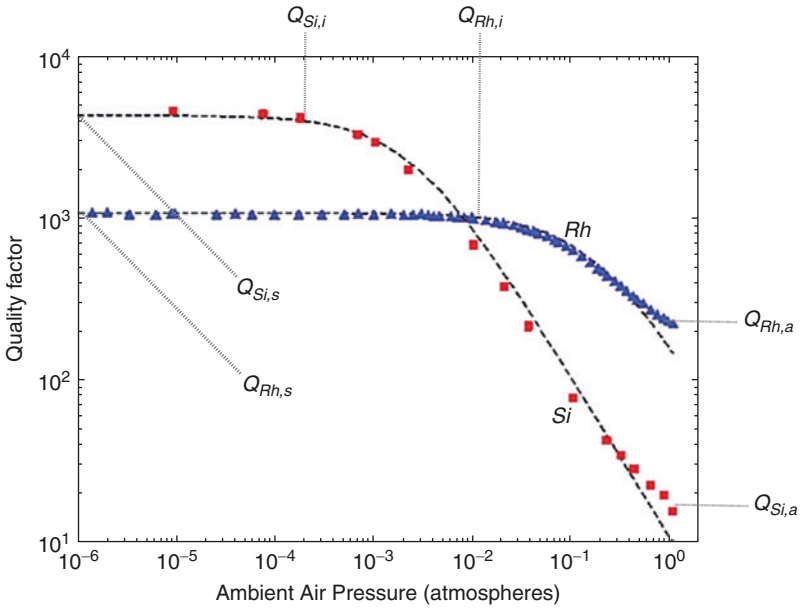
Hybrid strategies combine the ease of patterning simple planar structures, borrowed from top-down methods, with the flexibility of tailoring individual nanostructures provided by bottom-up methods, while partially overcoming the difficulties of each method. As an example of this synergy, consider the hybrid bottom-up method reported for nanoresonator array assembly and integration, and designed to accommodate a multiplexed biosensor chip [16]. In this study, both metallic rhodium and semiconducting silicon were synthesized off-chip. Polycrystalline rhodium nanowires were grown by electrodeposition in nanoporous aluminum oxide membranes, while predominately single-crystal silicon nanowires were grown by the Au-catalyzed vapor–liquid–solid (VLS) technique. Bottom-up directed assembly was used to position single nanowires at lithographically defined locations



Nanomechanical Resonant Sensors and Fluid Interactions, Fig. 6 Silicon and rhodium nanowire resonators made by a hybrid scheme using top-down patterning of electrodes and a bottom-up directed assembly scheme. Scale bar is 10 μm, for picture at *lower right* (Figure reprinted from Ref. [16], with permission)

on a silicon chip for fabricating multiplexed arrays. To do so, patterned arrays of wells in a sacrificial insulating photoresist layer covering metal guiding electrodes were defined on the chip surface. Dielectrophoretic forces were used to align the nanowires within the wells. Thick metal clamps were then built over the guiding electrodes by electrodeposition. The photoresist was dissolved to leave arrays of freestanding single-NW cantilevers as shown in scanning electron micrograph images of Fig. 6. The figures show the electrodeposited metal in intimate contact with the nanowire at each nanowire clamp. The clamps were shown to be of repeatable rigidity, with no undercuts. Besides securing the nanowire to the chip, the choice of gold for the clamp material, irrespective of the nanowire material, enables a good electrical connection between the chip and conductive nanowire, which can be used for electrical drive and detection of single nanowires in the array. Figure 6 shows that the distortion found in metal cantilevered nanowires from thin film stress evolution in the top-down process is absent. The photoresist used as a sacrificial layer was removed without the need for etchants that could damage the surface functionalization for biosensing. The example hybrid method above eliminates expensive e-beam lithography (reduces cost), shows the viability of long cantilevered metal nanoresonators. It uses lithographic patterning to enable directed assembly and nanoresonator clamping, allows choice of different materials for clamps and nanoresonators, and shows the potential to use different nanoresonator cross-section shapes, materials, and sensor functionalization chemistries for multiplexing on a single chip.

As seen above, the choice of fabrication technique plays a strong role in determining the nanoresonator properties available for sensing. In particular, flexible methods such as the hybrid technique described above, allow a wider choice of nanoresonator materials and cross sections. The density, Young's modulus and intrinsic damping within the nanowire material (C_s) can serve to provide some measure of tuning of the Q-factor for a given fluid damping, depending on the application desired. Figure 7 provides an example of this. It shows the Q-factor change from moderate vacuum to atmospheric pressure for a rhodium and a silicon nanowire of roughly comparable dimensions and fluid damping C_f. Although fine-grained, polycrystalline Rh should be expected to have a much higher intrinsic damping Cs compared to the single-crystal silicon nanowire, its higher elastic modulus and density compared to silicon compensate (Eq. 9), giving it a vacuum Q-value ($Q_{Rh,s}$) as high as a fourth of that of the silicon nanoresonator ($Q_{Si,s}$). The experimental data [16] of Fig. 7 further agrees with a prediction that may be made from Eq. 13 – the Q-factor for the silicon nanowire, with smaller intrinsic



Nanomechanical Resonant Sensors and Fluid Interactions, Fig. 7 Measured quality factor variation with ambient gas pressure for silicon and rhodium nanowires of comparable size (Figure reprinted from Ref. [16], with permission)

damping C_s, should begin to deviate from its vacuum value at a lower pressure value (point $Q_{Si,i}$ vs. $Q_{Rh,i}$) and show a sharper fall-off with pressure, than the rhodium nanowire. The Q-factor of the rhodium nanowire at atmospheric pressure ($Q_{Rh,a}$) is seen to be higher than that of the silicon nanowire ($Q_{Si,a}$). Gas damping of the nanoresonator motion increases as ambient gas pressure is raised from high vacuum to atmospheric pressure. Equation 13 shows that the high intrinsic damping in the rhodium nanowire serves as a buffer to delay the onset of Q-factor reduction and to limit its rate of change. This makes it more suitable for applications such as small mass sensing where reduction in Q-factor would adversely affect performance. However, the value of the Q-factor for the silicon nanowire begins to decrease at a lower pressure and then shows a larger change in measured Q-factor over the same pressure range, compared to the rhodium nanowire. As an instrument for measuring gas damping force, the silicon wire begins to show response at lower pressure and has a greater dynamic range than the rhodium nanowire. Silicon nanowires would thus serve as better instruments for studying gas damping as a function of pressure than rhodium nanowires.

In this section, the link between fabrication technique and nanoresonator material choice has been reviewed. Well-established top-down methods remain a strong choice for most batch-fabricated resonant sensors at microscale. For nanoresonators with dimensions far smaller than the wavelength of UV light sources, the costs of patterning nanostructures by rastering an electron beam, increase with chip area. Bottom-up methods, which rely on assembly of nanostructures made off-chip, begin to look more favorable for such structures. A larger variety of materials and shapes may be synthesized and assembled by these techniques, but precise location and individual addressing of assembled structures can be a challenge. An example of a hybrid fabrication scheme that borrows simple top-down photolithography steps for patterning relatively large microscale structures to enable spatially directed assembly and clamping to create nanoresonator arrays was reviewed. Nanoresonators of two different materials –polycrystalline rhodium and single-crystal silicon, illustrated how this method may be used to match material with different performance requirements for two different nanoresonator applications. The data show rhodium to be a more favorable nanoresonator material for a sensitive mass detector operating in air at near atmospheric pressure. The material properties of

silicon make it a better choice for a nanoresonant sensor of gas damping force over a large range of pressure, spanning the transition and free molecular flow regimes of rarefied gas dynamics.

Summary and Related Reviews

This section provides a summary of the choice of topics presented in this entry and points to kindred reviews of complementary material, which could not be accommodated here. Nanoscale resonant sensors were chosen due to their potential for creating new instrument systems of unprecedented sensitivity – the promise of their application to biosensing has been included in a recent comprehensive review of mechanical biosensors [24]. Table 1 of Ref. [24] indicates detection limits of 300 fM (femto-molar) and 1.5 fM concentration have been achieved with mechanical microcantilever resonators with and without mass-amplifying labels, respectively, compared to ~1,000 fM for the nonmechanical standard test ELISA (enzyme-linked immunosorbent assay). Biomolecular target capture kinetics, target depletion upon capture from small analyte volumes, target sequestration of low-abundance molecules by abundant nontarget molecules are a few other areas where a research effort is focused on extending the current boundaries of knowledge for molecular diagnosis with resonant cantilevers. Related work may be found in the entries on "Cantilever Biosensors" and "Nanowire Biosensors" in this Encyclopedia.

Work in nanoresonators constitutes a large field and the focus in this entry has been restricted to nanoresonators interacting with fluids – either gases or liquids. The goal has been to establish notions and the simplest useful quantitative framework to interrogate fluid interaction with nanoresonators. Damping in fluid environments can be strong and lead to low Q-factor operation, associated with a divergence in values calculated from different definitions of the Q-factor. Low Q-factor operation demands large frequency sweeps, which can bring about nonconstant forcing due to frequency-dependent circuit parasitics. Figure 2 provides an example summary of the coupled interdisciplinary work required to deal with the interdependencies between material properties desired for an application, fabrication schemes available, and transduction artifacts. Such work both requires and can extend knowledge in mechanics of materials, and fluid mechanics of continuum and rarefied gas flows at small scales [9, 14]. Details of transduction artifacts in nanoresonator measurements are worthy of a separate review and have not been included here. Ekinci [25] provides a useful review of nanoresonator transduction schemes. While the field of micro/nanofluidics, dealing with small-scale internal flows has acquired some maturity, small-scale external flows at high frequencies realized with nanoresonators hold promise of both technological application and an improved understanding of fluid mechanics.

Cross-References

- ▶ AFM in Liquids
- ▶ Electric-Field-Assisted Deterministic Nanowire Assembly
- ▶ Nanomechanical Properties of Nanostructures

References

1. Jensen, K., Kim, K., Zettl, A.: An atomic-resolution nanomechanical mass sensor. Nat. Nanotechnol. **3**, 533–537 (2008)
2. Hickman, S.A., et al.: Batch-fabrication of cantilevered magnets on attonewton-sensitivity mechanical oscillators for scanned-probe nanoscale magnetic resonance imaging. ACS Nano **4**, 7141–7150 (2010)
3. Radmacher, M.: Studying the mechanics of cellular processes by atomic force microscopy. In: Wang, Y.L., Discher, D.E. (eds.) Cell Mechanics, pp. 347–372. Elsevier, Amsterdam (2007)
4. Engler, A.J., Rehfeldt, F., Sen, S., Discher, D.E.: Microtissue elasticity: measurements by atomic force microscopy and its influence on cell differentiation. In: Wang, Y.L., Discher, D.E. (eds.) Cell Mechanics, pp. 521–545. Elsevier, Amsterdam (2007)
5. Lavrik, N.V., Sepaniak, M.J., Datskos, P.G.: Cantilever transducers as a platform for chemical and biological sensors. Rev. Sci. Instrum. **75**, 2229–2253 (2004)
6. Naik, T., Longmire, E.K., Mantell, S.C.: Dynamic response of a cantilever in liquid near a solid wall. Sens. Actuators A **102**, 240–254 (2003)

7. Verbridge, S.S., Craighead, H.G., Parpia, J.M.: A megahertz nanomechanical resonator with room temperature quality factor over a million. Appl. Phys. Lett. **92**, 013112 (2008)
8. Arlett, J.L., et al.: BioNEMS: nanomechanical systems for single-molecule biophysics. In: Nobel Symposium 131, Bäckaskog Slott, 31 Aug 2005
9. Bhiladvala, R.B., Wang, Z.J.: Effect of fluids on the Q factor and resonance frequency of oscillating micrometer and nanometer scale beams. Phys. Rev. E **69**, 036307 (2004)
10. Ramanathan, S., Koch, D.L., Bhiladvala, R.B.: Non-continuum drag force on a nanowire vibrating normal to a wall: simulations and theory. Phys. Fluids **22**, 103101 (2010)
11. Zheng, Y.S., Reese, J.M., Struchtrup, H.: Comparing macroscopic continuum models for rarefied gas dynamics: a new test method. J. Comput. Phys. **218**, 748–769 (2006)
12. Karabacak, D.M., Yakhot, V., Ekinci, K.L.: High-frequency nanofluidics: an experimental study using nanomechanical resonators. Phys. Rev. Lett. **98**, 254505 (2007)
13. Svitelskiy, O., et al.: Pressurized fluid damping of nanoelectromechanical systems. Phys. Rev. Lett. **103**, 244501 (2009)
14. Ekinci, K.L., et al.: High frequency nanofluidics: a universal formulation of the fluid dynamics of MEMS and NEMS. Lab Chip **10**, 3013–3025 (2010)
15. Sahai, T., Bhiladvala, R.B., Zehnder, A.T.: Thermomechanical transitions in doubly-clamped micro-oscillators. Int. J. Non Linear Mech. **42**, 596–607 (2007)
16. Li, M.W., et al.: Bottom-up assembly of large-area nanowire resonator arrays. Nat. Nanotechnol. **3**, 88–92 (2008)
17. Tian, M., Wang, J., Kurtz, J., Mallouk, T.E., Chan, M. H.W.: Electrochemical growth of single-crystal metal nanowires via a two-dimensional nucleation and growth mechanism. Nano Lett. **7**, 919–923 (2003)
18. Feynman, R.P., Leighton, R.B., Sands, M.: The Feynman Lectures on Physics. Addison Wesley Longman, Reading (1970)
19. Weaver, W., Timoshenko, S.P., Young, D.H.: Vibration Problems in Engineering. Wiley, New York (1990)
20. Carr, D.W., et al.: Measurement of mechanical resonance and losses in nanometer scale silicon wires. Appl. Phys. Lett. **75**, 920–922 (1999)
21. O'Mahony, C., et al.: Titanium as a micromechanical material. J. Micromech. Microeng. **12**, 438–443 (2002)
22. Fang, W., Wickert, J.A.: Determining mean and gradient residual stresses in thin films using micromachined cantilevers. J. Micromech. Microeng. **6**, 301–309 (1996)
23. Wang, M.C.P., Gates, B.D.: Directed assembly of nanowires. Mater. Today **12**, 34–43 (2009)
24. Arlett, J.L., Myers, E.B., Roukes, M.L.: Comparative advantages of mechanical biosensors. Nat. Nanotechnol. **6**, 213–215 (2011)
25. Ekinci, K.L.: Electromechanical transducers at the nanoscale: actuation and sensing of motion in nanoelectromechanical systems (NEMS). Small **1**, 786–797 (2005)

Nanomechanical Resonators Based on III–V Semiconductors

Imran Mahboob and Hiroshi Yamaguchi
NTT Basic Research Laboratories, NTT Corporation, Atsugi-shi, Kanagawa, Japan

Synonyms

Microelectromechanical systems (MEMS); Nanoelectromechanical systems (NEMS); Optomechanical resonators

Definition

Mechanical resonators synthesized from III-V semiconductors are micro- to nanoscale mechanically compliable devices with a single degree of freedom, usually in the form of beams and cantilevers, whose mechanical motion is excited and detected via the electronic and optical properties inherent to the underlying compound semiconductor.

Overview

In early reports, nanomechanical resonators were invariably synthesized from silicon which was metalized to form electrical leads that enabled mechanical motion to be excited via electrostatic or magnetomotive techniques and detected by capacitive or electromotive force methods, respectively [1]. The role played by silicon was to form the mechanical resonator and to sustain the electrical circuitry, enabling the motion

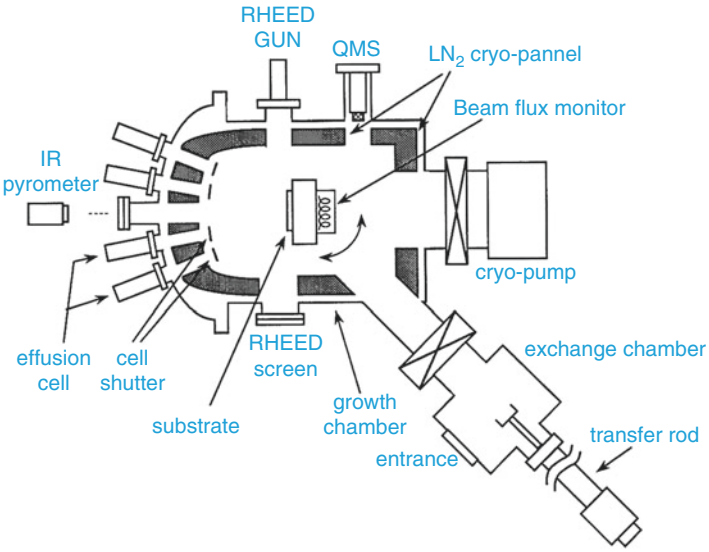
transduction. The only relevant intrinsic material requirement related to the purity and crystallinity of the silicon which enabled the realization of high-quality mechanical resonators. The quest for mechanical performance initially lead to the substitution of silicon with tensile-strained silicon nitride which resulted in quality factors of millions at room temperature [2]. The ultimate evolution of this approach led to the abandonment of silicon all together, and instead the mechanical resonator was formed from an amorphous metal which could readily be coupled to the electrical transduction circuit and simultaneously be tensile strained via thermal annealing for unprecedented mechanical performance [3].

Although the approaches outlined above have been incredibly successful, they have utilized silicon purely for its mechanical performance, and the electrical and optical properties inherent to it have played little or no role in either transducing the mechanical motion or controlling its vibration dynamics (note the optical properties referred to here relate to the underlying semiconductor band structure rather than radiation pressure, gradient optical and photothermal forces, all of which have been demonstrated with silicon-based optomechanical resonators). This state of affairs has primarily arisen from the fact that silicon is an indirect bandgap semiconductor and consequently it has limited optical functionality. The low mobility of its carries arising from the large electron/hole effective masses has also resulted in the intrinsic electronic properties of silicon, proving challenging to exploit, for instance, the piezoresistance arising from sub-Angstrom level mechanical vibrations at high frequency is almost undetectable. In contrast, direct bandgap III–V semiconductors with highly mobile carriers are capable of hosting optical and electronic phenomenon which can be modulated by mechanical fields and be readily measured, thus bringing a range of functions unavailable to silicon mechanics to the fore.

As an illustrative example, the optical emission wavelength in III–V compound semiconductors can be tuned via strain which shifts the conduction and valence band edges, thus modifying the bandgap energy. Consequently, the strain field arising in a III–V semiconductor-based mechanical resonator naturally leads to the optical emission being modulated on resonance, thus enabling the realization of a unique motion transducer that is unavailable to silicon mechanics [4]. The electronic properties in III–V semiconductors can also be far superior to those in silicon, for example, their electron mobility can be orders of magnitude greater, particularly in a two-dimensional electron gas (2DEG), thus making III–V semiconductors highly amenable to transducing the high-frequency mechanical vibrations in nanoscale mechanical resonators [5, 6]. On the other hand, the compound nature of III–V semiconductors lends itself to hosting piezoelectric fields, thus enabling transducers to be engineered directly into the mechanically compliable element which not only permits highly efficient displacement transduction but also renders the mechanical properties to be controlled, for example, the resonance frequency [7, 8].

In spite of the potential offered by III–V semiconductor-based mechanics, silicon mechanics (including SiC and SiN) have been ubiquitously embraced primarily due to the widespread availability of silicon stemming from the microelectronics industry. However, as research avenues with silicon have become exhausted, compound III–V semiconductors have emerged as natural arena in which optoelectromechanics exploiting their intrinsic material properties can be developed offering physics, functions, and applications beyond silicon mechanics.

To that end, this entry is organized as follows: The first section details crystal growth where electronic and optical properties of choice can be selectively incorporated into the semiconductor. The second section examines the optomechanical properties of III–V resonators by exploiting their direct bandgaps without requiring cavities to mediate this coupling. The third section describes the transport properties of III–V semiconductors which can both be modulated by the mechanics, and their subsequent back action can even control the mechanical vibration characteristics. The fourth section deals with the piezoelectric properties of III–V semiconductors which can be exploited to dynamically engineer the mechanical



Nanomechanical Resonators Based on III–V Semiconductors, Fig. 1 A simplified schematic of an MBE system showing all the major components needed to synthesize high-quality III–V semiconductor crystals

resonator's harmonic potential. Finally, the last entry will briefly discuss future trends and prospects of III–V compound semiconductor-based mechanical resonators.

MBE Crystal Growth

III–V compound semiconductors have a key advantage for application to mechanical devices due to their high-quality and function-enabled single crystal heterostructures being readily available via state-of-the-art growth techniques, for example, molecular-beam epitaxy (MBE) and metalorganic vapor phase epitaxy (MOVPE). These crystal growth techniques were originally pioneered to synthesize perfect materials that could be applied to developing high-speed electronics and high-performance optoelectronic devices which required defect-free interfaces at their heterojunctions. These growth techniques enabled crystals to be grown layer by layer with different chemical compositions to form artificial heterostructures with atomically flat interfaces while preserving their crystallinity. One of the most successful material systems developed by these growth techniques are the GaAs/AlGaAs heterostructures which exhibit a lattice mismatch of less than 0.3 % resulting in dislocation-free crystals. Other material systems developed using the techniques pioneered for GaAs/AlGaAs include InGaAs/InAlAs/InP, GaN/InGaN/AlGaN, and InAs/GaSb/AlGsSb which are employed to access different parts of the electromagnetic spectrum in optoelectronic devices. For the purpose of this chapter, the MBE growth technique will be briefly reviewed as MBE-grown GaAs/AlGaAs crystals are used to develop nanomechanical resonators described below.

Fundamentally MBE is an advanced evaporation technique in which precisely controlled substrate and source temperatures in an ultrahigh vacuum environment enable extremely stable deposition which permits high-quality single crystal films to be grown. Figure 1 shows a two-chamber configuration which is the minimum requirement of a practical MBE system.

The larger chamber is used for crystal growth in which a stable standby vacuum pressure typically on the order of 10^{-10} Torr is maintained. The smaller chamber is for loading or unloading the substrate or sample from and to the atmospheric environment, respectively, and it is operated at pressures below 10^{-7} Torr after evacuation. In order to maintain the ultrahigh vacuum in the growth chamber, it is baked for several days to desorb all atmospheric contamination and is subsequently not directly exposed to the local atmospheric environment via the smaller load-lock chamber. The source materials for crystal growth

are stored in effusion cells from where they are evaporated. The evaporation rate can be controlled by the cell temperature and its duration by the shutter located in front of the effusion cell. The evaporated molecular beams are directed toward the substrate where a seed crystal with a defined crystal orientation is located which gestates synthesis of a crystalline film. Crucially the substrate temperature is precisely controlled during this process which facilitates equilibrium growth conditions yielding almost perfect crystals. To that end, MBE systems are usually equipped with in situ characterization systems, typically reflection high-energy electron diffraction (RHEED), which can monitor the growth rate and surface quality during the growth process.

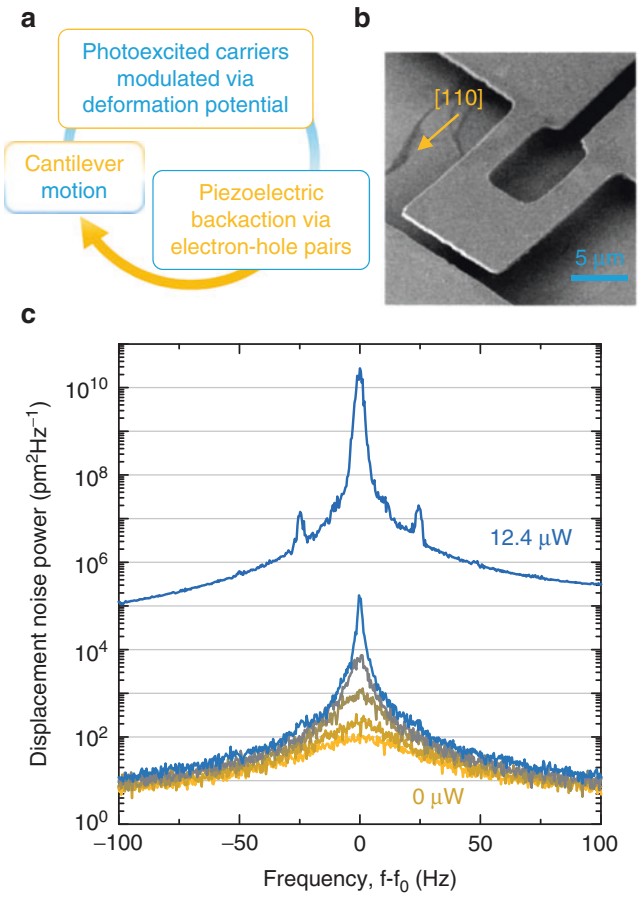
A major success of MBE is the growth of high-mobility modulation-doped heterostructures in which the carrier transport layer is isolated from the doping layer leading to the scattering of carriers by ionized donors being strongly suppressed. Indeed low-temperature electron mobilities of more than 10^6 $cmV^{-1}s^{-1}$ are now routinely accessed in such heterostructures. The use of such perfectly crystalline layers in the fabrication of mechanical devices not only yields high-quality factor resonators, but it also makes it possible to study the coupling between low-dimensional electronic systems and mechanical degrees of freedoms. Moreover, in optically active heterostructures in which high luminescence efficiencies are available allows for new flavors of optomechanical coupling to be explored, and ultimately these systems will pave the way toward a new class of optoelectromechanical platform.

Optomechanics with Direct Bandgap III–V Semiconductors

Invariably almost all optomechanical systems utilize a cavity to mediate the interaction between optics and mechanics, where the cavity's wavelength is modulated by the motion of the mechanical element, leading to parametric coupling which enables the mechanical oscillation dynamics to be optically controlled and engineered [9]. Remarkably a III–V semiconductor-based optomechanical system integrated with a laser actuator and a Fabry–Perot cavity-based displacement detector, where the cantilever formed one end of the cavity, was pioneered as long ago as 1993 [10].

On the other hand, the availability of direct bandgaps in III–V semiconductors, for example, in the GaAs/AlGaAs system, offers a different route to optomechanical coupling where crucially the optical cavity can be eliminated. Specifically, this can be realized via carriers excited at wavelengths selectively determined by the semiconductor bandgap which plays a similar role as an optical cavity. Particularly, the two major strain-stimulated phenomena in III–V semiconductors, namely, piezoelectricity and the deformation potential, can mediate this cavity-free optomechanical coupling. The former arises between the optically excited electron–hole pairs which generate stress in the resonator and leading to a bending moment which develops into a mechanical resonance. The latter induces a strain-dependent bandgap energy which results in the mechanical motion modulating the optical absorption wavelength. These processes in sequence lead to the establishment of a feedback loop (see Fig. 2a) which enables the mechanical resonance characteristics to be dynamically engineered via carrier-mediated optomechanical coupling.

The mechanical cantilever shown in Fig. 2b was the test bed system in which this cavity-free optomechanical coupling was explored, and it consists of a bilayer structure composed of Si-doped GaAs (100 nm) above undoped GaAs (200 nm) [11]. Excitation of this system with a TiSa laser, focused on one of the cantilever's clamping points, i.e., the position of maximal strain, with a wavelength around the GaAs bandgap (~840 nm at 50 K), resulted in the cantilever's thermal motion being amplified to the point where it self-oscillated above a threshold laser power of 10 μW as shown in Fig. 2c. In this configuration, the cantilever's motion modulates absorption of the laser light via the bandgap through the deformation potential which in turn transmits this modulation to the generated carriers



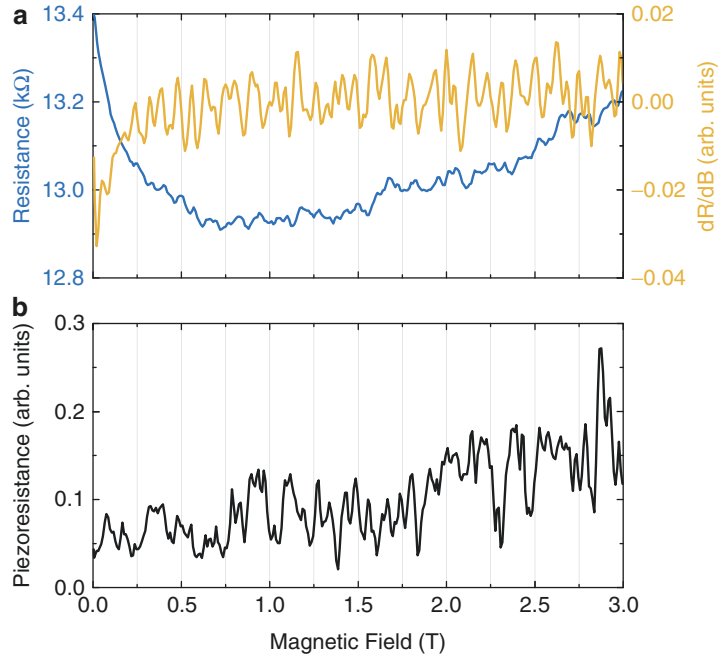
Nanomechanical Resonators Based on III–V Semiconductors, Fig. 2 (**a**) A schematic of the feedback loop at the heart of the carrier-mediated optomechanical coupling established by the deformation potential and the piezoelectric effect. (**b**) A scanning electron micrograph of the cantilever used to explore the cavity-free optomechanical coupling. (**c**) The noise power spectra from the cantilever's thermal motion acquired at 50 K in high vacuum via optical interferometry. As the laser's power, at 840 nm, is increased from 0 to 12.4 μW, the cantilever's motion is amplified to the point where it begins to self-oscillate, resulting in a dramatic increase in the corresponding displacement noise power (Figure 2c is adapted with permission from Okamoto et al. [11]. Copyrighted by the American Physical Society)

that underpin the resultant piezoelectric back action force, and it leads to a positive feedback loop that can amplify the mechanical motion. On the other hand, a negative feedback loop can also be activated with the inversion of the piezoelectric back action force's polarity by simply selecting a perpendicular crystal orientation for the cantilever, i.e., [$\bar{1}$10]. Indeed this polarity selection via the crystal orientation clearly demonstrates the central role played by the photoexcited carriers in this optomechanical coupling. Ultimately if the piezoelectric back action force can be made sufficiently strong in this configuration, it could even lead to mode cooling [11, 12].

Most tantalizingly this carrier-mediated optomechanical coupling opens up new directions for the use of III–V semiconductor-based optomechanical resonators, for example, it provides a unique probe for studying the optical and electrical properties of semiconductors such as carrier dynamics and strain effects [13].

Electromechanics with Quantum Transport

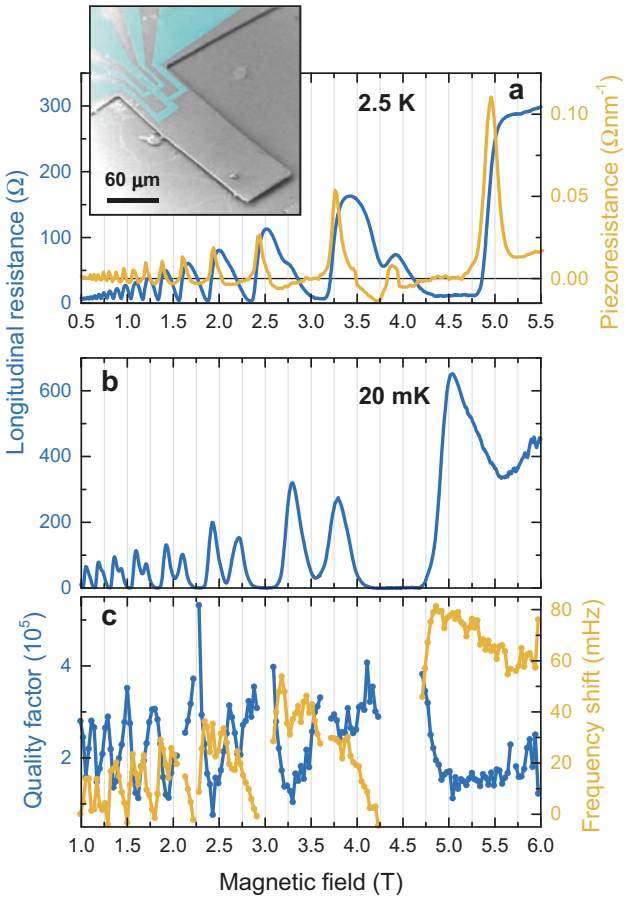
Nanomechanical resonators fabricated from III–V semiconductors can uniquely be integrated with low-dimensional electron systems which can sustain a range of quantum transport phenomena, for example, the quantum Hall effect. Since the resonance characteristics of the mechanics are fundamentally determined by their intrinsic system parameters, for example, the mass, any change in these critical parameters can be readily detected by the corresponding change in the resonance. Consequently, the dynamics of the integrated electron system can also be probed via the



Nanomechanical Resonators Based on III–V Semiconductors, Fig. 3 (**a**) The resistance across a cantilever, of a similar design to that shown in Fig. 2b, measured as a function of magnetic field at 2.5 K and the corresponding derivative. (**b**) The piezoresistance measured across the cantilever on resonance, with a frequency of 283.16 kHz and a quality factor of approximately 12,000, exhibits fluctuations that are reminiscent to *dR/dB* which indicates their similar origin, namely, from electron interference in the quasi-one-dimensional channels (Adapted with permission from Yamaguchi et al. [14]. Copyrighted by the American Physical Society)

mechanical resonator, and uniquely in equilibrium, which can enable the intrinsic properties of the electron system to be determined in contrast to transport measurements. Moreover, this platform can also be used to impart strain fields, akin to electric and magnetic fields, to the low-dimensional structures integrated into the mechanics in order to study their effect on and control of the underlying electron system.

As an example of this approach, the coupling between mechanics and a quasi-one-dimensional electron system in a cantilever of similar design to the one shown in Fig. 2a is reviewed. In this case, the cantilever consisted of a bilayer composed of InAs (15 nm) above AlGaSb (285 nm) where the legs of the cantilever of 4 μm width sustained the quasi-one-dimensional channels in the InAs layer which confined the electrons two dimensionally in-plane [14]. Transport measurements across the cantilever, where the two legs acted as the source and drain, enabled the resistance (R) to be extracted as a function of magnetic field (B) in a high vacuum at 2.5 K. This measurement, shown in Fig. 3a, exhibits the well-known universal conductance fluctuations, which can more clearly be seen in the derivative *dR/dB*, and they arise from interference between the different electron trajectories confined in the quasi-one-dimensional channels. On the other hand, when the piezoresistance is measured with the cantilever excited on resonance, it also exhibits aperiodic oscillations, similar to *dR/dB*, as shown in Fig. 2b. These observations can be understood as follows: in the case of the former, the electron interference is driven by the magnetic field which modifies the phase of the different electron trajectories, resulting in the aperiodic conductance fluctuations. In the case of the latter, the mechanical motion of the cantilever directly modifies the wavelength of the electron trajectories in addition



Nanomechanical Resonators Based on III–V Semiconductors, Fig. 4 (**a**) The longitudinal resistance and the on-resonance piezoresistance as a function of magnetic field measured at 2.5 K in a 2DEG confined to a Hall bar that is integrated onto the cantilever, with a resonance frequency of 20.19 kHz and a quality factor of 108,000, which is shown in the scanning electron micrograph in the inset where the Hall bar is artificially colored. (**b**) The longitudinal resistance as a function of magnetic field measured in the same device at 20 mK. (**c**) The cantilever's resonance exhibits changes in quality factor and shifts in frequency, both of which correlate with the quantum Hall effect in the 2DEG integrated into the cantilever (Figures 4b and 4c have been reproduced with permission from Yamaguchi et al. [17]. Copyright 2012, AIP Publishing LLC)

to the magnetic field modifying their phase, resulting in interference that manifests itself in the aperiodic oscillations in the piezoresistance. Consequently, the mechanical motion in principle can be detected with high sensitivity by exploiting this electron interference where it is maximized from the constructive interference [15].

Another example of this approach is the integration of a high-mobility 2DEG, in the form of a Hall bar, into the clamping point of the cantilever shown in Fig. 4a which consists of the layer structure detailed in Fig. 6. The integrated 2DEG exhibits an extremely high mobility of 10^6 $cm^2V^{-1}s^{-1}$ at liquid helium temperatures which enables the quantum Hall effect to be accessed, as evidenced when the 2DEG's resistance is measured with the application of a magnetic field, and it reveals zero resistance minima corresponding to the localization of electrons as shown in Fig. 4a. Surprisingly, when the piezoresistance in the 2DEG is measured with the cantilever on resonance, it correlates with quantum Hall effect as shown in Fig. 4a. Indeed the piezoresistance exhibits a dramatic enhancement at magnetic fields corresponding to the transition

between localized and extended states, i.e., where *dR/dB* is maximized at which point it yields a piezoresistance gauge factor that is more than two orders of magnitude greater than silicon [16]. This enhancement arises from motional-induced strain at the position of the 2DEG which modulates the position of the Fermi level with respect to the conduction band minimum. The resultant change in the carrier density shifts the piezoresistance response with respect to the magnetic field. Consequently, if the magnetic field is critically biased between the localized and extended states, the motional strain can drive transitions between the zero and high resistance states, resulting in the piezoresistance being dramatically enhanced.

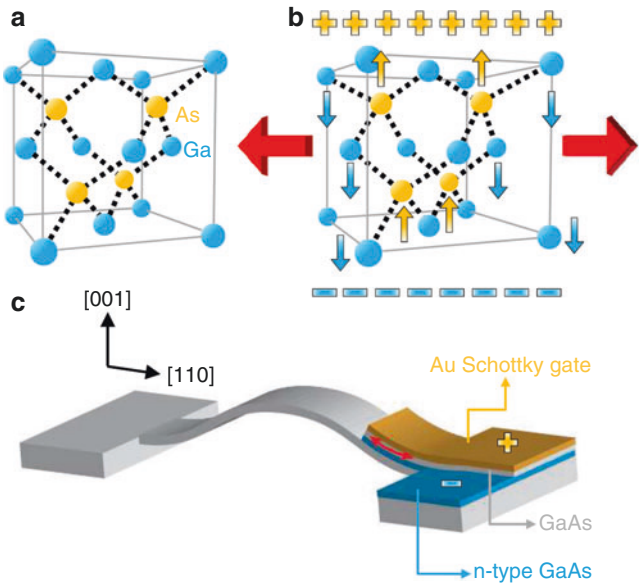
As the cantilever's motion can modify the electronic states of the 2DEG, analogously the 2DEG can also exert a back action onto the mechanics. Indeed if the cantilever's resonance frequency is monitored as a function of the magnetic field, it exhibits shifts which correlate with the magnetoresistance in the 2DEG as shown in Fig. 4b, c. The amplitude of the sawtooth-like frequency shifts at the edge of the localized/extended states in the 2DEG increases in proportion to the applied magnetic field. This behavior originates from the torque generated on the cantilever from the magnetic moment in the 2DEG which tries to align itself with the external magnetic field and is known as the de Haas-van Alphen effect [17]. Since the magnetic moment in the 2DEG oscillates between the extended and localized electron states, it results in the torque and hence the cantilever's resonance exhibiting frequency shifts with the same oscillatory period.

However, more surprisingly, this measurement also reveals that the cantilever's quality factor also correlates with the electron localization/delocalization observed in the magnetoresistance as shown in Fig. 4c. This effect arises from the strain-induced modulation of the Fermi level, resulting in current flow that undergoes ohmic losses, and the resultant energy dissipation manifests itself as a reduction in the cantilever's quality factor at magnetic fields corresponding to the extended electron states in the 2DEG. On the other hand, at magnetic fields where the localized electron states emerge, strain-induced current flow stimulated by the cantilever's motion is unavailable, and consequently ohmic losses and the corresponding energy dissipation are suppressed, resulting in an enhancement in the quality factor [17]. Consequently, the energy loss measured by the cantilever offers a unique probe to study the equilibrium electron dynamics in the 2DEG.

The above results vividly illustrate that III–V semiconductor-based low-dimensional electron systems can be readily incorporated into mechanical resonators which can be used to not only transduce the mechanical motion, but the mechanics can also provide a unique probe to study the electron dynamics of these systems.

Piezoelectric Mechanical Resonators

III–V semiconductors offer enormous potential for developing highly integrated electromechanical resonators due to the availability of the piezoelectric effect. The word piezoelectricity originates from the Greek word *piezien* which means to press and was discovered by the Curie brothers in 1880 who showed that certain crystals develop opposite charges on their surface with the application of pressure. Nowadays, the piezoelectric effect can be defined as the electric polarization produced with the application of mechanical strain in certain classes of crystal and can be expressed as $P_i = e_{ijk}S_{jk}$ where P is the polarization, S is the strain tensor, e is the piezoelectric coefficient, and i, j, and k are the Cartesian indices. This effect is reversible, namely, an electric field can also strain the crystal and the strength of the electromechanical coupling is determined by the piezoelectric coefficient which varies from crystal to crystal. At the microscopic level, this effect arises from a crystal lacking inversion symmetry along a particular crystal orientation, for example, in GaAs, the prototypical III–V semiconductor possesses a zinc blende crystal structure with a center of inversion along the [110] crystal orientation as shown in Fig. 5a. Application of an electric field along the [001] crystal orientation as shown in Fig. 5b results in the cation and anion sublattices from the gallium and arsenic atoms, respectively, being displaced due to their different electronegativity and thus being



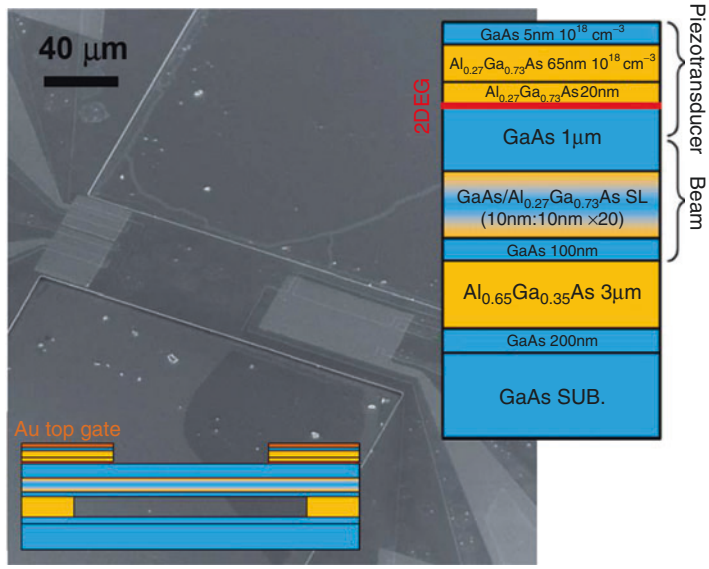
Nanomechanical Resonators Based on III–V Semiconductors, Fig. 5 (**a**) A schematic of the zinc blende arrangement of atoms in a GaAs unit cell. (**b**) Application of an electric field along the [001] crystal orientation causes the gallium and arsenic atoms to be displaced and polarized, resulting in strain being generated along the [110] crystal orientation via the piezoelectric effect. (**c**) Sandwiching GaAs between an n-doped GaAs layer and Au electrode enables a piezotransducer to be formed over a clamping point on the mechanical resonator which can activate mechanical vibrations via the piezoelectrically generated strain (*red arrow*). This transducer can also detect the mechanical vibration via the motional strain which induces a measureable voltage

polarized. The creation of dipoles in response to the electric field results in strain being generated along the [110] crystal axis and is at the heart of the piezoelectric effect in III–V semiconductors.

A scheme to exploit this elemental electromechanical effect as a motion transducer in a mechanical resonator is depicted in Fig. 5c. Here a hypothetical piezotransducer is created over one of the clamping points of a GaAs mechanical resonator, which is a bridge-like structure that is designed to only move up and down, and it consists of a Au Schottky gate and an n-type GaAs layer sandwiching a thin layer of undoped GaAs. As described in Fig. 5b, application of an electric field between the Au electrode and the n-type GaAs layer results in an elongation along the [110] crystal orientation, leading to a bending moment emerging in the mechanical element as a consequence of the doubly clamped geometry. If the electric field is modulated at the natural frequency of the mechanical element, this bending moment evolves into a mechanical resonance.

Although this concept had originally been developed for lead zirconate titanate piezoelectric transducers deposited onto Si-based micromechanical resonators [18], Roukes et al. extended this to GaAs nanomechanical cantilevers and beams integrated with p-type GaAs/intrinsic-GaAs/n-type (PIN-diode)-based piezotransducers which enabled large electric fields to be confined over nanometer length scales. Consequently, the application of even a single-electron charge could activate nanomechanical resonances with frequencies of 10s of MHz, demonstrating the efficiency of this highly integrated motion transducer [7]. However, detection of the corresponding mechanical resonance via the motion-induced voltage in the PIN-piezotransducer was unavailable.

To enable both mechanical resonances to be actuated and detected via the piezoelectric effect, the device shown in Fig. 6 was fabricated via conventional lithography, lift-off, and wet etching processes in a GaAs/AlGaAs modulation-doped

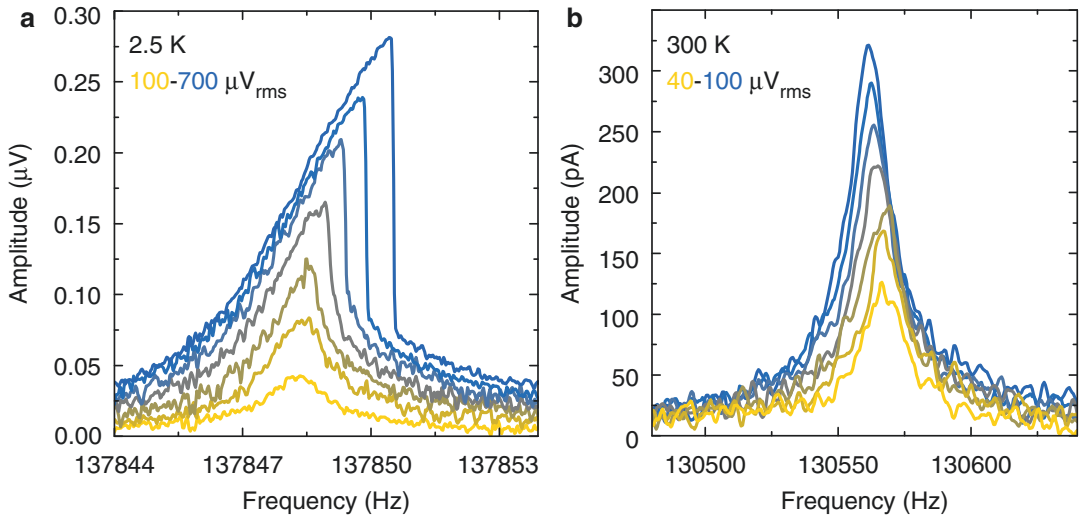


Nanomechanical Resonators Based on III–V Semiconductors, Fig. 6 A scanning electron micrograph of the electromechanical resonator formed from the GaAs/AlGaAs modulation-doped heterostructure, sustaining a 2DEG as detailed in the upper inset. The electromechanical resonator is integrated with piezoelectric transducers at both clamping points via the 2DEG and Au Schottky gates as detailed in the lower inset. The mechanically compliant element is a bridge-like structure which is suspended above the GaAs substrate, and when activated it vibrates with a sub-nanometer amplitude in the out-of-plane direction

heterostructure, sustaining a 2DEG located 90 nm below the surface. The 2DEG enabled the formation of a thin and highly integrated AlGaAs piezotransducer sandwiched between the Au Schottky gate and located at both clamping points of a micromechanical resonator as shown in Fig. 6. The corresponding device when measured in high vacuum and liquid helium temperatures enabled not only highly efficient activation of mechanical motion but also detection of that motion via the generated piezovoltage as shown in Fig. 7a [8]. Further, by incorporating a Si-transistor on-chip to amplify the motion-induced piezovoltage even enabled room temperature operation as shown in Fig. 7b [20]. Indeed operation frequencies of ~10 Mhz with displacement responsivities as high as 4 Am^{-1} yielding displacement sensitivities approaching 10^{-15} $mHz^{-1/2}$ have also been demonstrated in this system [21, 22].

A key feature of the piezoelectric transducers is their capacity to control the mechanical resonance frequency with the application of DC voltage. This results in the generation of strain via the piezoelectric effect which is converted into stress in the doubly clamped geometry shown in Fig. 6. This effect is analogous to tuning the frequency of a stringed instrument via the tension in the strings. For the device shown in Fig. 6 and measured at cryogenic temperatures (cf. Fig 7a), application of voltage to the largest gate electrode could tune the resonance frequency by ~30 HzV^{-1} which corresponds to more than 30 times the bandwidth of the resonance as shown in Fig. 8. For nanoscale resonators, this effect can be further enhanced, as consequence of their smaller masses and spring constants, which has yielded values as high as 80 $kHzV^{-1}$ [23].

At a more fundamental level, the ability to tune the tension and hence the frequency of the mechanical resonator with the application of voltage corresponds to controlling the harmonic potential underpinning the mechanical resonator. Specifically, the equation of motion of a mechanical resonator can be expressed as $m\ddot{x} + m\omega_0\dot{x}/Q + m\omega_0^2 x(t) = F(t)$ where m is the effective mass of the resonator with a natural frequency ω_0 and a spring constant $K = \omega_0^2 m$

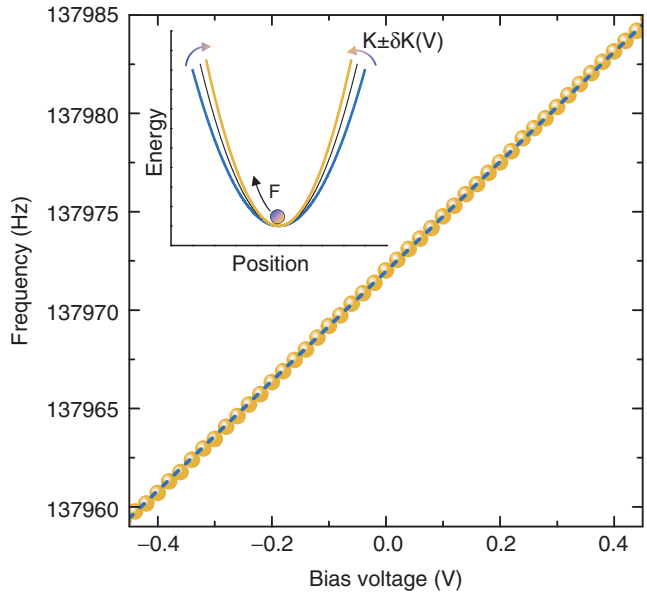


Nanomechanical Resonators Based on III–V Semiconductors, Fig. 7 (**a**) The frequency response of the mechanical oscillator shown in Fig. 6 measured at 10^{-7} mbar and 2.5 K when piezoelectrically excited via the largest piezotransducer in increments of 100 μV_{rms} reveals the fundamental mode at 137,848 Hz with $Q = 115{,}000$. By increasing the driving amplitude, the resonance develops the well-known Duffing nonlinearity [19]. (**b**) The frequency response of an identical device measured at room temperature and in a high vacuum with actuation amplitudes in increments of 10 μV_{rms}, where the motion-induced piezovoltage is amplified on-chip via a Si field-effect transistor, reveals the fundamental mode at 130,570 Hz with $Q = 10{,}000$. The piezotransducer can even activate the mechanical resonance with an amplitude as small as 10 μV_{rms} which corresponds to just 100 electrons being introduced onto the gate electrode. The high-quality factor resonance is also replicated at room temperature as a consequence of the excellent crystalline quality of the MBE-grown III–V semiconductor (Figure 7a is adapted from Mahboob and Yamaguchi [8]. Figure 7b has been reproduced with permission from Mahboob et al. [20]. Copyright 2009, AIP Publishing LLC)

N

that is driven by a force $F(t)$ yielding a displacement $x(t)$. This equation can be imagined as depicted in the inset to Fig. 8 where the mechanical motion corresponds to the ball moving in a harmonic potential in response to $F(t)$, and the energy loss rate during this motion is given by ω_{-0}/Q. Application of DC bias to the piezotransducer modifies the spring constant to $K \pm \delta K(V)$ which corresponds to the curvature of the harmonic potential changing where an increase in curvature with positive bias results in the ball oscillating more quickly, or in other words the resonance frequency increases and vice versa. Consequently, the piezotransducer not only enables a mechanical resonance to be activated and measured, but it also permits the harmonic potential which defines the dynamics of the mechanical resonator to be controlled.

The ability to control a fundamental parameter of the harmonic oscillator opens up the possibility of parametric resonance [24, 25]. Specifically modulating the spring constant of the mechanical resonator, or in other words its harmonic potential as depicted in the inset to Fig. 9, via application of voltage V to the piezotransducer at twice its natural frequency, namely, $\delta K = cV\cos(2\omega_0 t)$ where c is the coupling coefficient, can activate a parametric resonance. This effect is analogous to a child periodically standing and squatting on a swing during its arc to modulate its length in order to swing. Experimentally the mechanical resonator shown in Fig. 6 is parametrically excited by applying an AC bias to the largest piezotransducer at twice its natural frequency while simultaneously measuring its frequency response at half of the actuation frequency via another piezotransducer. The parametric resonance line shape shown in Fig. 9 is highly asymmetric, and it does not become damped as the actuation frequency is swept forward across the

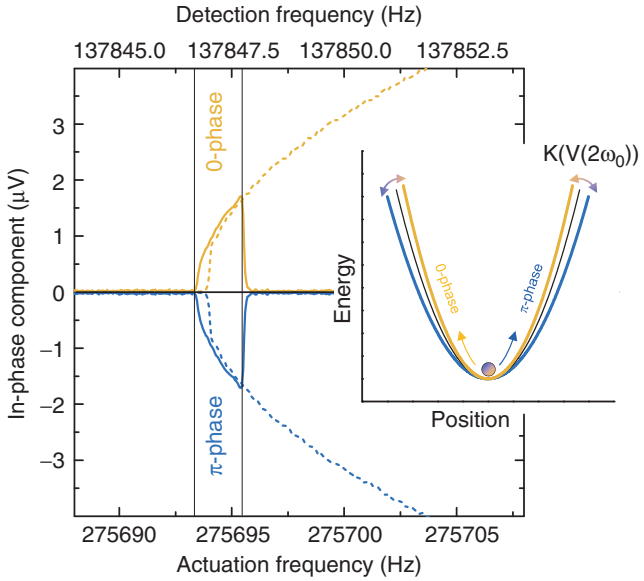


Nanomechanical Resonators Based on III–V Semiconductors, Fig. 8 The resonance shown in Fig. 7a is activated by one of the small piezotransducers shown in Fig. 6 and is measured in the adjacent piezotransducer, as a function of DC bias applied to the largest piezotransducer, and it reveals a linear frequency shift (points) corresponding to ~30 HzV^{-1} extracted from the fit shown in the dashed line. The shift in the resonance frequency via piezoelectric control of the spring constant can be envisaged as controlling the underlying harmonic potential of the mechanical resonator which determines its oscillation dynamics as depicted in the inset (Adapted from Mahboob and Yamaguchi [8])

resonance, but it does become damped if the actuation frequency is swept back across the resonance.

A unique feature of the parametric resonance is that it has only two stable oscillation phases which can be experimentally observed by decomposing the mechanical displacement into $x(t) = P\ \cos(\omega_0 t) + Q\sin(\omega_0 t)$ in a phase-sensitive detector where P and Q are slowly varying in-phase and quadrature components. In Fig. 9, the in-phase component of the parametric resonance is measured for increasing and decreasing frequencies multiple times. In the spectral region around the harmonic resonance, the parametric resonance can *only* oscillate with either a positive or a negative in-phase component of the same magnitude corresponding to a π radian phase difference between them. Physically this phase bistability can be understood by examining the dynamics of the ball in the harmonic potential as it is modulated at twice its natural frequency. In this regime, the ball has the option to move either toward the left or the right where the initial trajectory is selected by its random fluctuation driven by the thermal Langevin force. Consequently, if the initial movement of the ball is toward the left, this then corresponds to the 0 oscillation phase, for example, and by moving toward the right first corresponds to the π oscillation phase. By repeatedly parametrically exciting across the resonance, both oscillation phases can be accessed randomly which are identical in magnitude as shown in Fig. 9.

The availability of bistable oscillation phases to the parametric resonance provides a platform within which a mechanical memory can be executed [19, 26]. Indeed if a weak harmonic trigger is applied via $F(t)$ with a specific phase, the parametric resonance can be selectively switched between the 0 and π oscillation phases, enabling binary logicals 0 and 1 to be stored and switched between [8]. This concept can even be extended to build logic gates by utilizing multiple parametric

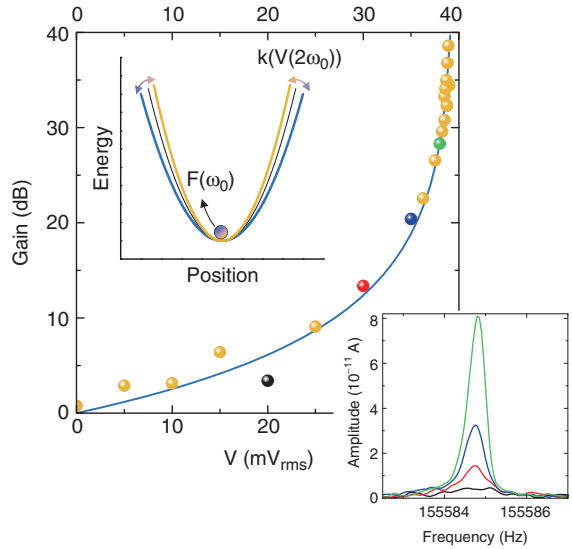


Nanomechanical Resonators Based on III–V Semiconductors, Fig. 9 A parametric resonance activated in the device shown in Fig. 6 at 10^{-7} mbar and 2.5 K via the largest piezotransducer with an amplitude of 80 mV_{rms} and measured in one of the smaller piezotransducers at half the actuation frequency. The parametric response is decomposed in phase space where the dashed and solid traces correspond to increasing and decreasing the actuation frequency across the resonance, respectively, and the two colors correspond to 0 and π oscillation phases. The two oscillation phases can be understood as the motion of the mechanical oscillator, as depicted by the ball in the harmonic potential in the inset, moving either first to the left or to the right in response to the parametric modulation of the harmonic potential (Adapted from Mahboob and Yamaguchi [8])

N

resonances [27], and it paves the way toward highly efficient mechanical computing [28]. Although parametric resonances in mechanical systems have previously been accessed via electrostatic transducers, they have invariably required 10s of volts [25]. In stark contrast, the highly integrated nature of the piezotransducers enables parametric resonances to be activated with amplitudes as small as 10 mV_{rms}, making this phenomenon more accessible for the development of technological applications [27].

The ability to modulate the spring constant of the mechanical resonator via the piezotransducer also opens up the possibility of degenerate parametric amplification. This effect is comparable to the child being *harmonically* pushed on the swing by an adult while the child alternatively raises and lowers their legs on the swing arc to *parametrically amplify* the swing's motion. Analogously this corresponds to the harmonic potential being modulated at twice the frequency of the harmonic force which is simultaneously driving the motion of the ball as depicted in the upper inset to Fig. 10. Clearly this effect is maximized when the harmonic actuation and the parametric modulation are in phase. In Fig. 10, the mechanical resonator is harmonically actuated with an amplitude of 1 μV_{rms}, and simultaneously it is parametrically pumped at exactly twice the harmonic actuation frequency. As the amplitude of the pump is increased, the harmonic motion is amplified as explicitly shown in the lower inset to Fig. 10, and gains of 40 dB could be achieved for pump amplitudes below 40 mV_{rms} beyond which parametric resonance occurred as shown earlier in Fig. 9 [29]. The degenerate parametric amplifier increases the mechanical resonator's vibration amplitude from a small harmonic driving force which in turn enhances its force responsivity, yielding an unprecedented force sensitivity of 200×10^{-18} $NHz^{-1/2}$ in a stiff system with $K > 100$ Nm^{-1} [30]. In contrast, earlier attempts to parametrically amplify the harmonic motion of the mechanical resonator utilized off-board electrostatic transducers which not

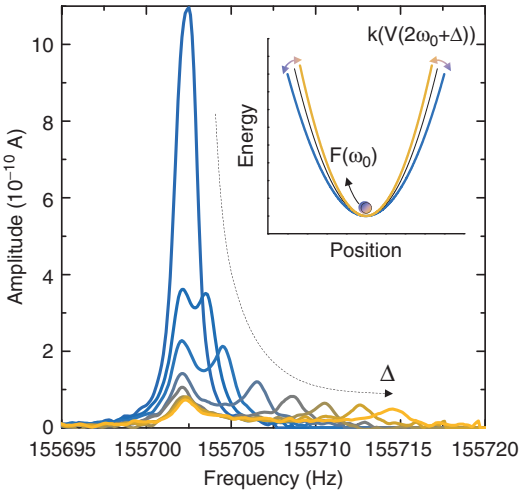


Nanomechanical Resonators Based on III–V Semiconductors, Fig. 10 A weak harmonic input can be amplified by parametrically modulating the harmonic potential at twice its frequency below the parametric resonance threshold as depicted in the upper inset. Experimentally, the resonator shown in Fig. 6 was harmonically activated with a 1 μV_{rms} input, while simultaneously it was parametrically pumped at exactly twice the frequency of the harmonic input, i.e., degenerately. As the pump amplitude is increased, the weak harmonic input can be amplified (points), as detailed in the lower inset where the color of the traces corresponds to the colored points, with a gain of 40 dB being possible before the onset of parametric resonance. The solid line corresponds to the theoretical response of a degenerate parametric amplifier indicating the almost perfect performance of the electromechanical variant [29] (Reproduced with permission from Mahboob et al. [29]. Copyright 2010, AIP Publishing LLC)

only required large voltages but also gave only modest gains of ~20 dB [31]. Again the integrated nature of the III–V semiconductor-based piezotransducer facilitates highly efficient degenerate parametric amplification.

A final example of the functionality accessed in the mechanical domain via the piezotransducers' ability to modulate the spring constant is nondegenerate parametric amplification [32]. In this configuration, the modulation of the harmonic potential is detuned by Δ from the degenerate case outlined above as depicted in the inset to Fig. 11. The nondegenerate parametric amplifier is routinely employed in optics to tune the frequency of a laser by exploiting the interaction of a high-frequency and high-intensity pump beam (ω_p) with a lower-frequency and lower-intensity signal beam (ω_s) in a crystal whose nonlinearity is activated by the pump, for example, the Kerr effect. This interaction amplifies the signal beam as well as generating an *idler* beam (ω_i) where this process conserves energy, i.e., $\hbar\omega_p = \hbar\omega_s + \hbar\omega_i$ and $\hbar$ is the reduced Planck constant.

This phenomenon can be replicated in a mechanical resonator with a periodic nonlinearity in its harmonic potential activated by the detuned piezoelectric pump. When the mechanical resonator is harmonically excited with a signal excitation at the fundamental mechanical mode ω_0 and simultaneously parametrically pumped at $2\omega_0 + \Delta$, an idler is created at their frequency difference $\omega_0 + \Delta$. By increasing the detuning Δ, the idler can be observed at the same time as the signal excitation within the bandwidth of the fundamental mode as shown in Fig. 11. Further if the mechanical resonator is excited with multiple detuned pumps, it results in the generation of multiple idlers which can further intermix via the periodic nonlinearity in the harmonic potential, and it results in a rich spectrum of idlers being generated. Indeed these idlers have even been exploited to construct



Nanomechanical Resonators Based on III–V Semiconductors, Fig. 11 The inset shows the interaction of a harmonic input and the parametric modulation of the harmonic potential at almost twice its frequency which can generate a new vibration, the idler, at their frequency difference. Experimentally the electromechanical oscillator shown in Fig. 6 is harmonically excited at 50 μV_{rms}, while simultaneously it is parametrically pumped with an amplitude of 40 mV_{rms} at frequency that is slightly detuned from twice the harmonic input. As the pump is detuned from $\Delta/2\pi = 0$ to 15 Hz, the idler emerges and becomes increasingly visible, whereas the harmonic input remains fixed. Note that when the detuning is zero, the idler and the signal input overlap, leading to degenerate parametric amplification (Adapted from Mahboob et al. [32])

all-mechanical logic gates and circuits that can be executed in parallel [32]. This concept can be further extended to multiple vibration modes, enabling the generation of combs and correlated mechanical states [33, 34]. Although nondegenerate parametric amplification has been widely exploited in optics, the extension of this concept to mechanics opens up the tantalizing prospect of phonons being utilized in the same vein as photons for range of applications.

III–V semiconductor-based mechanical resonators enable piezoelectric transducers to be synthesized directly within the mechanical element during crystal growth and subsequent fabrication. The piezotransducer can not only harmonically activate mechanical motion with unrivaled efficiency, but it can also detect motion with femtometer sensitivity. Moreover, it can also induce tension in the mechanical element to tune the resonance frequency which provides the key to dynamically engineer a range of nonlinear phenomenon into the mechanical domain, thus dramatically expanding the functionality portfolio of mechanical resonators.

Outlook and Prospects

Although the focus of this chapter has been on nanomechanical resonators fabricated from III–V compounds based on aluminum, gallium, indium, and antimony, arsenic atoms, respectively, many other compounds have remained largely unexplored. Of particular interest are the III-nitride semiconductors which sustain piezoelectric constants that are an order of magnitude larger than the more conventional III–V compounds detailed above as a consequence of the large size mismatch between the group III and nitrogen atoms [35]. Indeed given the promise of enhanced electro-to-mechanics coupling mediated via the piezoelectric effect has led to III-nitride semiconductors being explored for mechanical resonators [36, 37]. However, invariably the III-nitride films used to fabricate the mechanical resonators have been prepared via sputtering, resulting in poor crystalline quality which inevitably not only compromises the piezoelectric effect, but it also results in low-quality factor resonators. To that end, MOVPE-based crystalline III-nitride resonators offer the prospect of bypassing the above limitations, leading to new opportunities; for instance, the enhanced electric-to-mechanics coupling could give rise to mechanical single-electron detectors [38, 39]. In addition to offering access to more energetic parts of the electromagnetic spectrum to explore optomechanical coupling, III-nitride resonators also hold the promise of quality factor engineering as the nitrogen films can be readily grown with tensile strain [40]. Ultimately, and in contrast to the more conventional silicon-based mechanical resonators, the choice of atoms and the design of the layer structure can enable III–V semiconductor-based mechanical resonators to be readily functionalized for electrical, optical, or mechanical applications.

Cross-References

- ▶ Magnetic Resonance Force Microscopy
- ▶ Nanomechanical Resonant Sensors and Fluid Interactions
- ▶ Nems Piezoelectric Switches
- ▶ Optomechanical Resonators

References

1. Ekinci, K.: Electromechanical transducers at the nanoscale: Actuation and sensing of motion in nanoelectromechanical systems (nems). Small **1**, 786 (2005)
2. Verbridge, S.S., Craighead, H.G., Parpia, J.M.: A megahertz nanomechanical resonator with room temperature quality factor over a million. Appl. Phys. Lett. **92**, 013112 (2008)
3. Regal, C.A., Teufel, J.D., Lehnert, K.W.: Measuring nanomechanical motion with a microwave cavity interferometer. Nat. Phys. **4**, 555 (2008)
4. Montinaro, M., Wüst, G., Munsch, M., Fontana, Y., Russo-Averchi, E., Heiss, M., Morral, A.F.I., Warburton, R.J., Poggio, M.: Quantum dot optomechanics in a fully self-assembled nanowire. Nano Lett. **14**, 4454 (2014)
5. Tang, H.X., Huang, X.M.H., Roukes, M.L., Bichler, M., Wegscheider, W.: Two-dimensional electron-gas actuation and transduction for gaas nanoelectromechanical systems. Appl. Phys. Lett. **81**, 3879 (2002)
6. Cleland, A., Aldridge, J., Driscoll, D., Gossard, A.: Nanomechanical displacement sensing using a quantum point contact. Appl. Phys. Lett. **81**, 1699 (2002)
7. Masmanidis, S.C., Karabalin, R.B., Vlaminck, I., Borghs, G., Freeman, M.R., Roukes, M.L.: Multifunctional nanomechanical systems via tunably coupled piezoelectric actuation. Science **317**, 780 (2007)
8. Mahboob, I., Yamaguchi, H.: Bit storage and bit flip operations in an electromechanical oscillator. Nat. Nanotechnol. **3**, 275 (2008)
9. Aspelmeyer, M., Kippenberg, T. J., Marquardt, F.: Cavity optomechanics. Rev. Mod. Phys. **86**, 1391 (2014).
10. Ukita, H., Uenishi, Y., Tanaka, H.: A photomicrodynamic system with a mechanical resonator monolithically integrated with laser diodes on gallium arsenide. Science **260**, 786 (1993)
11. Okamoto, H., Ito, D., Onomitsu, K., Sanada, H., Gotoh, H., Sogawa, T., Yamaguchi, H.: Vibration amplification, damping, and self-oscillations in micromechanical resonators induced by optomechanical coupling through carrier excitation. Phys. Rev. Lett. **106**, 036801 (2011)
12. Okamoto, H., Watanabe, T., Ito, D., Onomitsu, K., Sanada, H., Gotoh, H., Sogawa, T., Yamaguchi, H.: Carrier-mediated optomechanical coupling in gaas cantilevers. Phys. Rev. B **84**, 014305 (2011)
13. Watanabe, T., Okamoto, H., Onomitsu, K., Gotoh, H., Sogawa, T., Yamaguchi, H.: Optomechanical photoabsorption spectroscopy of exciton states in gaas. Appl. Phys. Lett. **101**, 082107 (2012)
14. Yamaguchi, H., Tokura, Y., Miyashia, S., Hirayama, Y.: Quantum interference effects in the magnetopiezoresistance of inas/algasb quasi-one-dimensional electron systems. Phys. Rev. Lett. **93**, 036603 (2004)
15. Mahboob, I., Okamoto, H., Ueki, M., Yamaguchi, H.: Electron phase modulation in a suspended inasalgasb nanomechanical beam. Appl. Phys. Lett. **89**, 192106 (2006)
16. Yamaguchi, H., Okamoto, H., Maruta, Y., Ishihara, S., Miyashita, S., Hirayama, Y.: Giant magneto-piezoresistance and internal friction in a two-dimensional electron system. Jpn. J. Appl. Phys. **46**, L658 (2007)
17. Yamaguchi, H., Okamoto, H., Ishihara, S., Hirayama, Y.: Motion detection of a micromechanical cantilever through magneto-piezovoltage in two-dimensional electron systems. Appl. Phys. Lett. **100**, 012106 (2012)
18. Piekarski, B., Devoe, D., Dubey, M., Kaul, R., Conrad, J.: Surface micromachined piezoelectric resonant beam filters. Sens. Actuat. A **91**, 313 (2001)
19. Badzey, R.L., Zolfagharkhani, G., Gaidarzhy, A., Mohanty, P.: A controllable nanomechanical memory element. Appl. Phys. Lett. **85**, 3587 (2004)
20. Mahboob, I., Nishiguchi, K., Fujiwara, A., Yamaguchi, H.: Room temperature piezoelectric displacement detection via a silicon field effect transistor. Appl. Phys. Lett. **95**, 233102 (2009)
21. Mahboob, I., Nier, V., Nishiguchi, K., Fujiwara, A., Yamaguchi, H.: Multi-mode parametric coupling in an electromechanical resonator. Appl. Phys. Lett. **103**, 153105 (2013)
22. Okazaki, Y., Mahboob, I., Onomitsu, K., Sasaki, S., Yamaguchi, H.: Quantum point contact displacement transducer for a mechanical resonator at sub-kelvin temperatures. Appl. Phys. Lett. **103**, 192105 (2013)
23. Hatanaka, D., Mahboob, I., Okamoto, H., Onomitsu, K., Yamaguchii, H.: An electromechanical membrane resonator. Appl. Phys. Lett. **101**, 063102 (2012)
24. Nayfeh, A.H., Mook, D.T.: Nonlinear Oscillations. Wiley Classics Library New York (1995)
25. Turner, K.L., Miller, S.A., Hartwell, P.G., MacDonald, N.C., Strogatz, S.H., Adams, S.G.: Five parametric resonances in a microelectromechanical system. Nature **396**, 149 (1998)
26. Goto, E.: The parametron, a digital computing element which utilises parametric oscillation. Proc. IRE **47**, 1304 (1959)
27. Mahboob, I., Mounaix, M., Nishiguchi, K., Fujiwara, A., Yamaguchi, H.: A multimode electromechanical parametric resonator array. Sci. Rep. **4**, 4448 (2014)
28. Roukes, M. L.: Mechanical computation, redux? IEEE IEDM Tech. Dig. 539–542 (2004)
29. Mahboob, I., Flurin, E., Nishiguchi, K., Fujiwara, A., Yamaguchi, H.: Enhanced force sensitivity and noise

squeezing in an electromechanical resonator coupled to a nanotransistor. Appl. Phys. Lett. **97**, 253105 (2010)
30. Mamin, H.J., Rugar, D.: Sub-attonewton force detection at millikelvin temperatures. Appl. Phys. Lett. **79**, 3358 (2001)
31. Rugar, D., Grütter, P.: Mechanical parametric amplification and thermomechanical noise squeezing. Phys. Rev. Lett. **67**, 699–702 (1991)
32. Mahboob, I., Flurin, E., Nishiguchi, K., Fujiwara, A., Yamaguchi, H.: Interconnect-free parallel logic circuits in a single mechanical resonator. Nat. Commun. **2**, 198 (2011)
33. Mahboob, I., Wilmart, Q., Nishiguchi, K., Fujiwara, A., Yamaguchi, H.: Wide-band idler generation in a gaas electromechanical resonator. Phys. Rev. B **84**, 113411 (2011)
34. Mahboob, I., Wilmart, Q., Nishiguchi, K., Fujiwara, A., Yamaguchi, H.: Tuneable electromechanical comb generation. Appl. Phys. Lett. **100**, 113109 (2012)
35. Bernardini, F., Fiorentini, V., Vanderbilt, D.: Spontaneous polarization and piezoelectric constants of iii-v nitrides. Phys. Rev. B **56**, R10024(R) (1997)
36. Karabalin, R.B., Matheny, M.H., Feng, X.L., Defaÿ, E., Rhun, G.L., Marcoux, C., Hentz, S., Andreucci, P., Roukes, M.L.: Piezoelectric nanoelectromechanical resonators based on aluminum nitride thin films. Appl. Phys. Lett. **95**, 103111 (2009)
37. O'Connell, A.D., Hofheinz, M., Ansmann, M., Bialczak, R.C., Lenander, M., Lucero, E., Neeley, M., Sank, D., Wang, H., Weides, M., et al.: Quantum ground state and singlephonon control of a mechanical resonator. Nature **464**, 697 (2010)
38. Taniyasu, Y., Kasu, M., Makimoto, T.: An aluminium nitride light-emitting diode with a wavelength of 210 nanometres. Nature **441**, 325 (2006)
39. Faucher, M., Grimbert, B., Cordier, Y., Baron, N., Wilk, A., Lahreche, H., Bove, P., Francois, M., Tilmant, P., Gehin, T., et al.: Amplified piezoelectric transduction of nanoscale motion in gallium nitride electromechanical resonators. Appl. Phys. Lett. **94**, 233506 (2009)
40. Onomitsu, K., Mitsuhara, M., Yamamoto, H., Yamaguchi, H.: Ultrahigh-q micromechanical resonators by using epitaxially induced tensile strain in ganas. Appl. Phys. Express **6**, 111201 (2013)

Nanomechanics

► Reliability of Nanostructures

Nanomedicine

► Acoustic Nanoparticle Synthesis for Applications in Nanomedicine

Nanometric

► Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

Nanomotors

► Molecular Modeling on Artificial Molecular Motors

Nano-optical Biosensors

► Nanophotonic Structures for Biosensing

Nano-optics Biosensing

► Nanophotonic Structures for Biosensing

Nano-optics, Nanoplasmonics

► Optical Properties of Metal Nanoparticles

Nano-optomechanical Systems (NOMS)

Huan Li and Mo Li
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN, USA

Synonyms

Cavity optomechanics; Micro-optomechanical systems (MOMS); Optomechanical resonators; Optomechanics

Definition

Nano- and micro-optomechanical systems (NOMS and MOMS) are nanoscale and microscale mechanical devices that are driven by optical forces and sensed with optical detection methods.

Chemical and Physical Principles

Light can generate forces on matter. The nature of these forces is electromagnetic force, or Lorentz force. Fundamentally they originate from the forces exerted on the charges (electrons and nuclei) that constitute the matter by the electric and magnetic fields in light. The optical forces are generally considered too weak to have any significance at macroscale. However, the emergence and rapid progress of nanotechnology provided an unprecedented platform where the very feeble optical forces can play significant roles and lead to various optomechanical phenomena. On the one hand, nanoscale structures can be engineered to tightly confine light near or below the diffraction limit, which greatly enhances the light field strength, hence the optical force. On the other hand, nanoscale objects possess small mass and are generally compliant, therefore can be easily moved or deformed by the optical forces. Conversely, the motion and deformation of these nanoscale objects can change the propagation and frequency of light. In NOMS devices, this new form of light-matter interactions can be rationally engineered and lead to many unprecedented physical phenomena in both classical and quantum regimes. Also, NOMS devices have enabled and opened up possibilities of novel functionalities for practical applications.

The electrical and comparably conventional counterpart of NOMS and MOMS devices are ▶ NEMS and ▶ MEMS devices, which employ the interactions between electricity (as the letter "E" in their acronyms implies) and mechanical structures via electrical forces and mechanical motions (and/or deformations).

Optical Force

Optical forces may manifest themselves in drastically different ways, including (but not limited to) pushing, pulling, squeezing or expanding, determined by the particular configuration of mechanical structure and light field confinement in the device. Consequently, in different contexts and situations, different terminologies for optical forces are used, including gradient force, optical dipole force, optical electrostriction force, and, perhaps the most commonly known, radiation pressure.

To calculate optical force, there are in general three methods – calculating the Lorentz force directly, calculating the time derivative of the momentum of the light wave, and calculating the space derivative of energy (or work) of the optical field. Within each method, light can be modeled quantum-mechanically as photons or classically as electromagnetic waves. In certain cases semiclassical approaches may also be used.

Before the methods of optical force calculations are briefly reviewed, it is worth emphasizing that, to obtain correct and meaningful results, caution must be exercised when using any of the following methods. The applicability and subtle implications of the methods must be carefully checked. Microscopic and macroscopic quantities such as charge, current, and fields sometimes need to be clearly differentiated. Any medium that is anisotropic, inhomogeneous, lossy, or dispersive may need special treatment. Calculations of the electromagnetic field momentum inside a medium, the optical force distribution inside a medium, or the total optical force exerted on a partial volume of a medium may lead to ambiguous results, sometimes involving the Abraham–Minkowski dilemma [1]. The readers will be referred to dedicated literature for more details.

When calculating optical forces directly, the expression for Lorentz force density [2] should be used, and light is modeled as electromagnetic waves.

$$\mathbf{f}^{\mathrm{L}} = \rho\mathbf{E} + \mathbf{J} \times \mathbf{B} \tag{1}$$

Oftentimes the charge density ρ and current density $\mathbf{J}$ are further expressed in terms of electric and magnetic fields according to Maxwell's equations, leading to the following two distinct expressions for the force density [2, 3]:

$$\mathbf{f}^{c} = -(\nabla \cdot \mathbf{P})\mathbf{E} + \dot{\mathbf{P}} \times \mathbf{B} \tag{2}$$

$$\mathbf{f}^{d} = (\mathbf{P} \cdot \nabla)\mathbf{E} + \dot{\mathbf{P}} \times \mathbf{B} \tag{3}$$

Generally, the electric and magnetic fields of the light inside the medium volume of interest should be calculated first before Eqs. 1, 2 or 3 is used to calculate the optical force density. The total optical force is the integration of the force density over the medium volume of interest. For more details and subtleties the readers are referred to the following literature [2–5].

When calculating optical forces as the time derivative of light momentum, light can be modeled as photons or electromagnetic waves. In either case, the total optical force should be calculated from the net momentum flow into a closed surface that surrounds the medium volume of interest. Generally, for simple cases concerning only monochromatic plane waves in vacuum, reflected or absorbed by a flat surface, it is very convenient and intuitive to model light as photons, the momentum carried by each of which is h/λ, according to the de Broglie relations. For more complicated cases, it is more convenient to model light as electromagnetic waves that carry momentum. In practice, Maxwell's stress tensor [6] $\mathbf{T}$ defined in Eq. 4 is often used to quantify the linear momentum flux (rate of flow of momentum per unit area) of electromagnetic waves.

$$T_{ij} = \varepsilon_0 \left[E_i E_j + c^2 B_i B_j - \frac{1}{2}\delta_{ij}\left(\mathbf{E} \cdot \mathbf{E} + c^2 \mathbf{B} \cdot \mathbf{B}\right) \right] \tag{4}$$

The total optical force can be found by the integration of the normal component of $\mathbf{T}$ over the closed surface that surrounds the medium volume of interest.

$$\mathbf{F} = \oint_S \mathbf{n} \cdot \mathbf{T} da \tag{5}$$

For more details and subtleties, especially about the momentum of light in medium, the readers are referred to the following literature [6–9].

When calculating optical forces as the space derivative of light energy (or the work done by the light), closed systems and open systems need to be treated differently. The simpler case is the closed systems, which in this context means a perfect cavity where light is forever trapped inside. Light is conveniently modeled as photons. Assume that the cavity, whose angular resonance frequency is ω_0, very slowly undergoes an infinitesimal mechanical displacement along a generalized coordinate q. Due to the optical force exerted on the cavity, photons inside the cavity do work and their energy (hence frequency) changes according to the shift of the cavity angular resonance frequency. Therefore the optical force along the same generalized coordinate q satisfies Eq. 6, where U is the total photon energy and N is the total photon number.

$$F_q(q) = -\frac{dU}{dq} = -N\hbar \frac{d\omega_0(q)}{dq} \tag{6}$$

This conclusion is generally a very good approximation for any real-world high-quality cavities excited on or near resonance. However, many real-world NOMS devices cannot be approximated as closed systems and have to be treated as open systems, which light can flow into and out of. In this case, the Response Theory of Optical Forces (RTOF) [10] should be used. RTOF is a semiclassical approach that relates the work done by the photons transiently stored in the system to the phase changes of the light waves, when the system very slowly undergoes an infinitesimal mechanical displacement along a generalized coordinate q. For a lossless one-port system driven at a single frequency, the optical force along the generalized coordinate q satisfies Eq. 7, where Φ is the photon flux (rate of flow of photon number) and ϕ is the phase of the output light wave. For more details and subtleties the readers are referred to the following literature [10, 11].

$$F_q(q) = \Phi\hbar \frac{d\phi(q)}{dq} \tag{7}$$

Finally, a completely different category of optical force is photothermal force [12]. Since optical

absorption and heat generation are inevitable in all realistic materials, thermal effects due to optical heating, i.e., photothermal effects, are always present. On devices with multilayers of different materials with differential thermal expansion coefficients, optical heating can induce mechanical motion which is equivalent to a thermal force applied on the device. A significant drawback of photothermal actuation, however, is that its effective bandwidth is inherently limited by the cutoff frequency determined by the device's thermal response time constant. Beyond this cutoff frequency, which is often much lower than the device's mechanical resonance frequency, the actuation efficiency rolls off at 20 dB per frequency decade. Therefore, due to its low efficiency as well as its dissipative nature, photothermal effects are less desirable in NOMS and often need to be minimized by reducing optical absorption.

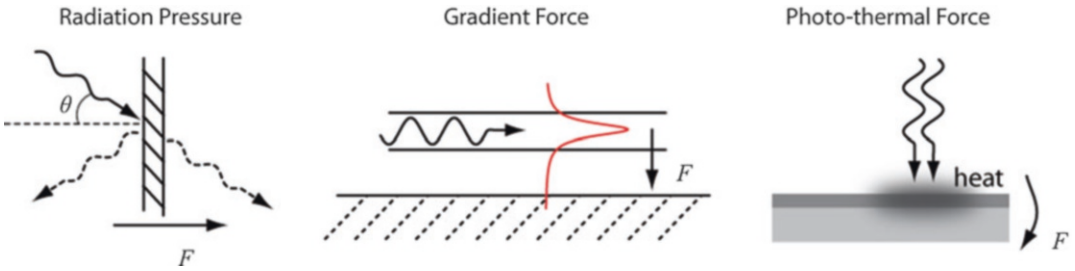
Optical Transduction of Mechanical Motion and Deformation

Various optical transduction techniques have been applied to measure mechanical motion and deformation. These techniques can be divided into two general categories: those based on optical interference and those based on light deflection. The first category includes all kinds of laser interferometers such as Fabry-Perot, Michelson, Mach-Zehnder, and Sagnac interferometers. The laser Doppler vibrometer also belongs to this category. In a recent magnetic resonance force microscopy (MRFM) experiment, an optical fiber-based Fabry-Perot interferometer is used to read out a microcantilever's motion with attonewton force resolution [13]. The second category includes the optical lever technique, which is widely used in atomic force microscopy (▶ AFM) to detect the motion of microcantilever probes. The above methods are all implemented with free-space optics whereby a laser beam has to be focused on the mechanical device's surface and the reflected optical intensity or interference is measured. Therefore, these techniques are subject to the diffraction limit of conventional optics: when the device's dimensional scales are significantly smaller than the optical wavelength as in the situation of nanomechanical devices, these methods will have significantly reduced sensitivity if they are not completely inapplicable.

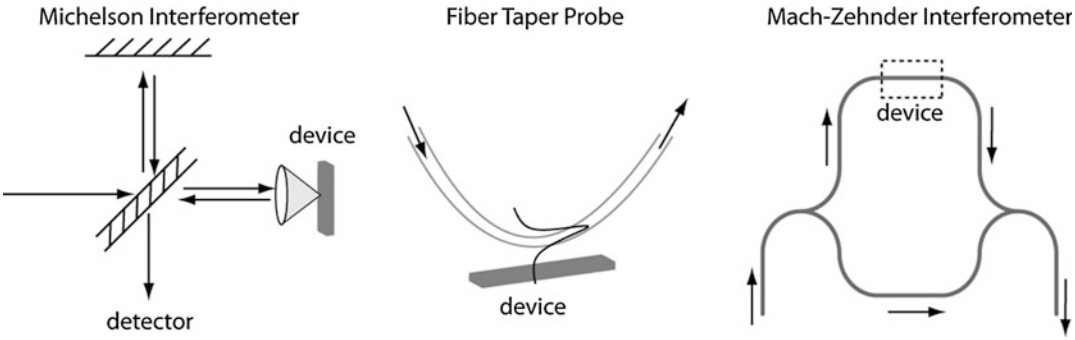
Recent development of near-field optics and integrated photonics can extend optical transduction beyond the diffraction limit. In near-field schemes, a tapered fiber or a microscale optical resonator, for example, are employed as optical probes when they are brought to the proximity of the mechanical devices [14, 15]. At a distance less than the wavelength, the optical field in the fiber or the resonator can evanescently couple with the mechanical devices. The motion of the mechanical devices will either cause additional scattering loss or shift the resonance frequency of the resonator so it can be measured in the transmitted optical signals with high precision. In integrated photonic schemes, optical interferometers can be implemented on a chip using dielectric waveguides. For example, a Mach-Zehnder interferometer built with silicon waveguides is demonstrated in reference [16] to measure the vibration of a nanomechanical beam that is embedded in the sensing "arm" of the interferometer. The motion of the beam induces a phase shift to the optical signal propagating through it. Interference between the optical signals from the sensing and the reference arms of the interferometer converts the phase shift to an intensity variation at its output where the signal is measured with a photodetector. Other than interferometers, various integrated optical resonators or cavities including microrings, microdisks, and photonic crystals have been utilized to measure nanomechanical devices with even better resolution [17, 18]. These devices of optical cavities with mechanical elements embedded inside are among the representative systems of the emerging field of "cavity optomechanics" [19, 20].

Key Research Findings

Although the first conception of radiation pressure dates back to as early as 1619, when Johannes Kepler attempted to explain the observation that a tail of a comet always points away from the sun, it was not until late 1980s that optical force



Nano-optomechanical Systems (NOMS), Fig. 1 Optical forces that are available for actuating mechanical devices



Nano-optomechanical Systems (NOMS), Fig. 2 Examples of free-space interferometer, near-field fiber taper probe, and integrated waveguide interferometer

N

found its first remarkable and widespread application – ▶ optical tweezers [21]. Experimentally demonstrated and reported by A. Ashkin et al. in 1986 for the first time [22], optical tweezers use optical forces to trap and manipulate biological cells or mesoscale particles micron- or nanometer-sized.

In 2005, the interactions between the light field inside a Fabry-Perot (F-P) cavity and a movable cavity mirror predicted in V. B. Braginsky et al.'s theoretical work [23, 24] were experimentally observed and reported for the first time by T. J. Kippenberg et al. [25, 26]. Instead of an F-P cavity, this experiment was conducted with a microtoroid ultrahigh-quality-factor (Q) optical cavity accessed by a fiber taper, which turned out to be an excellent experimental platform for cavity optomechanics [19, 20, 27]. The circulating light field inside the microtoroid cavity exerts distributed optical force that tends to expand the toroid. The resonance frequency of the microtoroid cavity red-shifts when the circumference of the toroid increases, leading to change of the intracavity light field. Due to the ultrahigh resonator Q, the change of the intracavity light field, hence that of the optical force, is delayed from the mechanical deformation. These dynamic interactions [19] modify the effective spring constant and damping coefficient of the mechanical mode, resulting in intriguing phenomena such as optical spring and optomechanical heating (also known as amplification) [26] and cooling [28, 29] of the Brownian motion.

The aforementioned breakthrough was followed immediately by almost a decade of active theoretical and experimental investigations up to now, which are still ongoing and constitute a whole new burgeoning branch of nanotechnology, nano-optomechanical systems (NOMS). Although the research scope of NOMS is now much wider than that of cavity optomechanics, the latter has always been the most important part of NOMS so far. The major experimental

milestones in the development of NOMS should include the following, sorted by the time of their publication in peer-reviewed journals.

In 2005 and 2006, cavity optomechanical heating (or amplification) and cooling were demonstrated, respectively, in silica microtoroid optical cavities [26, 28]. Also in 2006, the same phenomena were demonstrated in an F-P cavity with a microscale movable mirror [30, 31].

In 2008, two more NOMS experimental platforms were proposed and implemented. One is a high-finesse optical F-P cavity with a silicon nitride micromechanical membrane inside, which is parallel to the cavity mirrors and of nanoscale thickness [32]. The other is nanoscale suspended silicon photonic waveguides fabricated on silicon on insulator (SOI) wafers [16]. The former platform is implemented with high-quality free-space optics, while the latter is completely integrated on an SOI chip.

In 2009, three more NOMS experimental platforms were proposed and implemented. One is the "zipper" cavities, which consist of a pair of specially patterned nanoscale silicon beams to support colocalized photon and phonon modes that realize large per-photon optical forces [17]. Another is optomechanical crystals, which are suspended one-dimensional (1D) nanoscale periodic silicon structures, whose photonic and phononic bandgaps and defects are engineered to support colocalized photon and phonon modes that strongly interact [33]. The third is vertically stacked silicon nitride double disk resonators, whose optical response can be efficiently statically controlled using attractive and repulsive optical forces [34].

In 2009, tunable attractive and repulsive optical forces exerted on two coupled parallel nanoscale optical waveguides were demonstrated on an integrated silicon photonics platform [35].

In 2010, the tunable nonlinear optomechanical coupling in a low-loss system, which consists of an F-P cavity with a membrane of nanoscale thickness inside, was systematically explored [36].

In 2010 and 2011, optomechanically induced transparency (OMIT) was demonstrated with a microtoroid [37] and an optomechanical crystal [38], respectively.

In 2011, laser cooling of a nanomechanical oscillator into its quantum ground state was achieved in an optomechanical crystal device [29].

In 2011, a nonvolatile mechanical memory element fully integrated on an SOI chip was demonstrated, in which bits are written and reset by using optomechanical back-action [39]. The bistability (buckling up or buckling down) of a suspended silicon nanobeam with compressive stress is used to represent "0" or "1."

In 2012, optomechanical synchronization of two dissimilar vertically stacked silicon nitride double disk oscillators was demonstrated. The two oscillators are coupled through optical cavity field [40].

In 2012, quantum-coherent coupling of a mechanical oscillator to an optical cavity mode was achieved in a microtoroid cavity optomechanical system [41].

In 2013, squeezed light from a silicon micromechanical resonator was demonstrated with the "zipper" cavity [42].

In 2014, optically mediated hybridization between two mechanical modes was demonstrated in an F-P cavity with a membrane of nanoscale thickness inside [43].

In 2014, optomechanical photon shuttling between photonic cavities was demonstrated with a "photon see-saw" device fully integrated on an SOI chip [44].

Examples of Application

Nano-optomechanical systems have demonstrated both exceedingly high sensitivity and large dynamic range in measuring their mechanical motion. Therefore, when optimized for specific applications, they can be highly sensitive mechanical transducers that are capable of measuring weak forces in attonewton range, minuscule masses below a zeptogram and ultimately at the single molecule level, magnetic fields below a picotesla, and mechanical effects induced by single electrons or spins. The intrinsic high speed, efficiency, and sensitivity of optical actuation and sensing also enable the operation of mechanical resonators at very high frequencies, which

makes them very promising for applications in RF/microwave communication and signal processing.

The optomechanical effect also provides an all-optical approach to tune the functionalities of integrated photonic devices adaptively. Many reconfigurable photonic devices, including tunable optical filters, couplers, and switches, have been proposed and demonstrated. For example, in reference [45], the frequency of a microring optical filter is tuned by the application of optical force by more than 4 nm, a remarkably large spectral range that is hardly attainable with other methods.

Cross-References

- ▶ AFM
- ▶ MEMS
- ▶ NEMS
- ▶ Optical Tweezers

References

1. Pfeifer, R.N.C., Nieminen, T.A., Heckenberg, N.R., Rubinsztein-Dunlop H.: Colloquium: Momentum of an electromagnetic wave in dielectric media. Rev. Mod. Phys. **79**(4), 1197–1216 (2007)
2. Barnett, S.M., Loudon, R.: On the electromagnetic force on a dielectric medium. J. Phys. B. At. Mol. Opt. Phys. **39**(15), S671–S684 (2006)
3. Zakharian, A.R., Polynkin, P., Mansuripur, M., Moloney, J.V.: Single-beam trapping of micro-beads in polarized light: numerical simulations. Opt. Express **14**(8), 3660 (2006)
4. Loudon, R., Barnett, S.M.: Theory of the radiation pressure on dielectric slabs, prisms and single surfaces. Opt. Express **14**(24), 11855 (2006)
5. Mansuripur, M., Zakharian, A.R., Wright, E.M.: Electromagnetic-force distribution inside matter. Phys. Rev. A **88**(2), 023826 (2013)
6. Jackson, J.D.: Classical Electrodynamics, 3rd edn. Wiley, New York (1999)
7. Barnett, S.M., Loudon, R.: The enigma of optical momentum in a medium. Philos. Transact. A. Math. Phys. Eng. Sci. **368**(1914), 927–939 (2010)
8. Baxter, C., Loudon, R.: Radiation pressure and the photon momentum in dielectrics. J. Mod. Opt. **57**(10), 830–842 (2010)
9. Landau, L.D., Lifshitz, E.M., Pitaevskii, L.P.: Electrodynamics of Continuous Media, 2nd edn. Pergamon Press, Oxford, New York (1984)
10. Rakich, P.T., Popović, M.A., Wang, Z.: General treatment of optical forces and potentials in mechanically variable photonic systems. Opt. Express **17**(20), 18116–18135 (2009)
11. Rakich, P.T., Wang, Z., Davids, P.: Scaling of optical forces in dielectric waveguides: rigorous connection between radiation pressure and dispersion. Opt. Lett. **36**(2), 217–219 (2011)
12. Metzger, C.H., Karrai, K.: Cavity cooling of a microlever. Nature **432**(7020), 1002–1005 (2004)
13. Rugar, D., Budakian, R., Mamin, H.J., Chui, B.W.: Single spin detection by magnetic resonance force microscopy. Nature **430**(6997), 329–332 (2004)
14. Anetsberger, G., Arcizet, O., Unterreithmeier, Q.P., Rivière, R., Schliesser, A., Weig, E.M., Kotthaus, J. P., Kippenberg, T.J.: Near-field cavity optomechanics with nanomechanical oscillators. Nat. Phys. **5**(12), 909–914 (2009)
15. Basarir, O., Bramhavar, S., Ekinci, K.L.: Near-field optical transducer for nanomechanical resonators. Appl. Phys. Lett. **97**(25), 253114 (2010)
16. Li, M., Pernice, W.H.P., Xiong, C., Baehr-Jones, T., Hochberg, M., Tang, H.X.: Harnessing optical forces in integrated photonic circuits. Nature **456**(7221), 480–484 (2008)
17. Eichenfield, M., Camacho, R., Chan, J., Vahala, K.J., Painter, O.: A picogram- and nanometre-scale photonic-crystal optomechanical cavity. Nature **459**(7246), 550–555 (2009)
18. Li, M., Pernice, W.H.P., Tang, H.X.: Ultrahigh-frequency nano-optomechanical resonators in slot waveguide ring cavities. Appl. Phys. Lett. **97**(18), 183110 (2010)
19. Kippenberg, T.J., Vahala, K.J.: Cavity optomechanics: back-action at the mesoscale. Science **321**(5893), 1172–1176 (2008)
20. Aspelmeyer, M., Kippenberg, T. J., Marquardt, F.: Cavity optomechanics. Rev. Mod. Phys. 86(4), 1391–1452 (2014)
21. Moffitt, J.R., Chemla, Y.R., Smith, S.B., Bustamante, C.: Recent advances in optical tweezers. Annu. Rev. Biochem. **77**, 205–228 (2008)
22. Ashkin, A., Dziedzic, J.M., Bjorkholm, J.E., Chu, S.: Observation of a single-beam gradient force optical trap for dielectric particles. Opt. Lett. **11**(5), 288 (1986)
23. Braginsky, V.B., Strigin, S.E., Vyatchanin, S.P.: Parametric oscillatory instability in Fabry–Perot interferometer. Phys. Lett. A **287**(5–6), 331–338 (2001)
24. Braginsky, V.B., Strigin, S.E., Vyatchanin, S.P.: Analysis of parametric oscillatory instability in power recycled LIGO interferometer. Phys. Lett. A **305** (3–4), 111–124 (2002)
25. Rokhsari, H., Kippenberg, T.J., Carmon, T., Vahala, K. J.: Radiation-pressure-driven micro-mechanical oscillator. Opt. Express **13**(14), 5293 (2005)
26. Kippenberg, T., Rokhsari, H., Carmon, T., Scherer, A., Vahala, K.: Analysis of radiation-pressure induced mechanical oscillation of an optical microcavity. Phys. Rev. Lett. **95**(3), 033901 (2005)

27. Kippenberg, T.J., Vahala, K.J.: Cavity optomechanics. Opt. Express **15**(25), 17172 (2007)
28. Schliesser, A., Del'Haye, P., Nooshi, N., Vahala, K., Kippenberg, T.: Radiation pressure cooling of a micromechanical oscillator using dynamical backaction. Phys. Rev. Lett. **97**(24), 243905 (2006)
29. Chan, J., Alegre, T.P.M., Safavi-Naeini, A.H., Hill, J. T., Krause, A., Gröblacher, S., Aspelmeyer, M., Painter, O.: Laser cooling of a nanomechanical oscillator into its quantum ground state. Nature **478**(7367), 89–92 (2011)
30. Gigan, S., Böhm, H.R., Paternostro, M., Blaser, F., Langer, G., Hertzberg, J.B., Schwab, K.C., Bäuerle, D., Aspelmeyer, M., Zeilinger, A.: Self-cooling of a micromirror by radiation pressure. Nature **444**(7115), 67–70 (2006)
31. Arcizet, O., Cohadon, P.-F., Briant, T., Pinard, M., Heidmann, A.: Radiation-pressure cooling and optomechanical instability of a micromirror. Nature **444**(7115), 71–74 (2006)
32. Thompson, J.D., Zwickl, B.M., Jayich, A.M., Marquardt, F., Girvin, S.M., Harris, J.G.E.: Strong dispersive coupling of a high-finesse cavity to a micromechanical membrane. Nature **452**(7183), 72–75 (2008)
33. Eichenfield, M., Chan, J., Camacho, R.M., Vahala, K. J., Painter, O.: Optomechanical crystals. Nature **462**(7269), 78–82 (2009)
34. Wiederhecker, G.S., Chen, L., Gondarenko, A., Lipson, M.: Controlling photonic structures using optical forces. Nature **462**(7273), 633–636 (2009)
35. Li, M., Pernice, W.H.P., Tang, H.X.: Tunable bipolar optical interactions between guided lightwaves. Nat. Photonics **3**(8), 464–468 (2009)
36. Sankey, J.C., Yang, C., Zwickl, B.M., Jayich, A.M., Harris, J.G.E.: Strong and tunable nonlinear optomechanical coupling in a low-loss system. Nat. Phys. **6**(9), 707–712 (2010)
37. Weis, S., Rivière, R., Deléglise, S., Gavartin, E., Arcizet, O., Schliesser, A., Kippenberg, T.J.: Optomechanically induced transparency. Science **330**(6010), 1520–1523 (2010)
38. Safavi-Naeini, A.H., Mayer Alegre, T.P., Chan, J., Eichenfield, M., Winger, M., Lin, Q., Hill, J.T., Chang, D.E., Painter, O.: Electromagnetically induced transparency and slow light with optomechanics. Nature **472**(7341), 69–73 (2011)
39. Bagheri, M., Poot, M., Li, M., Pernice, W.P.H., Tang, H.X.: Dynamic manipulation of nanomechanical resonators in the high-amplitude regime and non-volatile mechanical memory operation. Nat. Nanotechnol. **6**(11), 726–732 (2011)
40. Zhang, M., Wiederhecker, G.S., Manipatruni, S., Barnard, A., McEuen, P., Lipson, M.: Synchronization of micromechanical oscillators using light. Phys. Rev. Lett. **109**(23), 233906 (2012)
41. Verhagen, E., Deléglise, S., Weis, S., Schliesser, A., Kippenberg, T.J.: Quantum-coherent coupling of a mechanical oscillator to an optical cavity mode. Nature **482**(7383), 63–67 (2012)
42. Safavi-Naeini, A.H., Gröblacher, S., Hill, J.T., Chan, J., Aspelmeyer, M., Painter, O.: Squeezed light from a silicon micromechanical resonator. Nature **500**(7461), 185–189 (2013)
43. Shkarin, A.B.., Flowers-Jacobs, N.E., Hoch, S.W., Kashkanova, A.D., Deutsch, C., Reichel, J., Harris, J. G.E.: Optically mediated hybridization between two mechanical modes. Phys. Rev. Lett. **112**(1), 013602 (2014)
44. Li, H., Li, M.: Optomechanical photon shuttling between photonic cavities. Nat. Nanotechnol. **9**(11), 913–919 (2014)
45. Rosenberg, J., Lin, Q., Painter, O.: Static and dynamic wavelength routing via the gradient optical force. Nat. Photonics **3**(8), 478–483 (2009)

Nanoparticle

► Toxicology: Plants and Nanoparticles

Nanoparticle Cytotoxicity

Nastassja A. Lewinski
Department of Bioengineering, Rice University, Houston, TX, USA

Synonyms

Cellular toxicity; In vitro toxicity

Definition

Nanoparticle cytotoxicity is defined as the extent to which the interaction of nanoparticles with cells disrupts cellular structures and/or processes essential for cell survival and proliferation. Cytotoxicity assays are a quick and simple way to perform initial acute toxicity assessments. Coupling nanoparticle cytotoxicity data with other safety testing data can help predict the biocompatibility of nanoparticles.

Overview

When evaluating the biocompatibility of a chemical, often the first step is to screen for adverse effects in cultured cells before using animal models. Compared to animal studies, in vitro testing is less ethically ambiguous, easier to control and reproduce, and less expensive. Conducting preliminary evaluations in vitro also provides a basis for prioritizing chemical testing and reducing the number of animals used for in vivo toxicity testing. Generally, the first endpoint tested in chemical biocompatibility assessments is cell lethality. Cytotoxicity assays are used to measure the extent of cell death and inhibition of cell proliferation after exposure to a chemical. Since toxicity is often dose-dependent, a range of concentrations is tested to determine the LC_{50} or the concentration at which 50 % cell death occurs. The LC_{50} serves as a point of comparison of the toxicity of chemicals and is used to group chemicals into toxicity classes ranging from extremely toxic (e.g., botulinum toxin) to relatively harmless (e.g., water). Knowledge of dose is essential as many chemicals that are beneficial at low doses can be toxic at high doses.

When evaluating the toxicity of a new chemical, characterizing the composition, purity, stability, and solubility is necessary before testing. In the case of nanoparticles, more extensive characterization, including particle size and size distribution, shape, crystal structure, surface charge, and surface chemistry, is warranted since these factors may also contribute to the observed effects. Generally size, shape, and agglomeration state are thought to be governing properties; however, surface composition and structure also influence the way nanoparticles interact with biological environments. Careful attention should be paid to how the various physiochemical properties of nanoparticles modify their toxicity. Although a detailed assessment of the various particle characterization techniques is beyond the scope of this definition, the reader is referred to review articles on this topic [1, 2]. Given the combinatorial nature of nanoparticle testing, in vitro assays facilitate quick screening of nanoparticle formulations to identify which nanoparticles elicit adverse effects. Depending on their intended use, the results from initial in vitro tests can be used to prioritize them for further development and/or in vivo testing. The following sections will describe some of the most common cytotoxicity assays, the potential mechanisms of nanoparticle toxicity that have been determined to date, and some challenges for future nanoparticle cytotoxicity research.

Common Cytotoxicity Assays

Cytotoxicity assays can be categorized by the endpoints they evaluate. Markers of cell injury and death include changes in membrane integrity, mitochondrial function, and cell proliferation. While some of these endpoints can be identified by microscopy, visual inspection of cellular changes provides only qualitative information. Instead, a majority of published nanoparticle cytotoxicity studies present measurements of cell death using colorimetric methods. A compilation of the assays discussed is presented in Table 1 toward the end of section.

Membrane Integrity

The cell membrane serves as a physical barrier separating the cell from its surrounding environment, with traffic into and out of the cell regulated by transporters, receptors, and secretion pathways. Exposure to nanoparticles could compromise the cell membrane, making it "leaky" which is the basis for the following assays. Membrane integrity is typically assessed by measuring either (1) the uptake of a cell impermeable dye, (2) the retention of a cell permeable dye, or (3) the release of a cell biomolecule into the extracellular media.

Cell impermeable dyes can only enter cells with compromised membranes; therefore, injured or dead cells are stained while intact or live cells remain unstained. Examples of dyes normally excluded from intact cells that are utilized for cytotoxicity assays include trypan blue, ethidium bromide (EtBr), and propidium iodide (PI). Trypan blue is an absorbing dye, staining dead cells blue which can be manually counted using a hemocytometer. Although trypan blue assay

Nanoparticle Cytotoxicity, Table 1 Summary of common cytotoxicity assays

Assay	Mechanism of action	Indication	Limitations in testing NPs
Trypan blue	Dye exclusion	Dead cells stained blue	
Neutral red	Lysosomal retention	Live cells stained red, abs = 540 nm	Dye adsorption to NPs in the absence of cells results in absorbance readings
Calcein AM	Dye retention	Live cells fluoresce green, em = 520 nm	
Fluorecein diacetate	Dye retention	Live cells fluoresce green, em = 520 nm	
Ethidium homodimer	DNA marker	Dead cells fluoresce red, em = 620 nm	
Propidium iodide	DNA marker	Dead cells fluoresce red, em = 620 nm	Dye adsorption to NPs quenches fluorescence
LDH	LDH release	LDH presence converts dye, abs = 490 nm	
MTT	Mitochondrial reduction	Live cells stained purple, abs = 570 nm	Substrate adsorption to NPs reduces absorbance
Alamar blue	Mitochondrial reduction	Live cells fluoresce red, em = 590 nm	Substrate adsorption to NPs quenches fluorescence
Luciferin/luciferase	ATP-dependent oxidation	Live cells luminesce, em = 560 nm	
Annexin V-FITC	Phosphatidylserine binding	Apoptotic cells fluoresce green, em = 520 nm	

NPs nanoparticles, *abs* absorbance, *em* emission

results have been published by a few groups, it is generally considered a qualitative assay as it is less sensitive and less reliable than other assays. EtBr and PI are red fluorescent dyes, whose peak emission intensities at 620 nm are greatly enhanced upon binding to DNA and can be measured using a flow cytometer or spectrofluorometer. Since quantitative measurements can be derived from these fluorescent dye based assays, they are more widely used. In addition, EtBr and PI can be used to identify necrotic cells and when co-stained with annexin V-FITC differentiate them from apoptotic cells. Refer to the *Apoptosis* Versus *Necrosis* section for further discussion.

A complementary technique to dye exclusion uses cell permeable dyes, which tend to accumulate within cells, to assess membrane integrity. If the cell membrane is damaged, the dye can leak out of cells or its uptake is decreased, with fluorescence intensity differentiating live and dead cells. Examples of commonly used dyes include neutral red, calcein acetoxymethyl (calcein AM), and fluorescein diacetate (FDA). Neutral red absorbs at 540 nm and can be measured with a spectrophotometer. Calcein AM and FDA are nonfluorescent dyes, which are converted by intracellular esterases to their fluorescent counterpart once within the cell. Cytotoxicity can be quantified by measuring the fluorescence intensity at 520 nm using a flow cytometer or spectrofluorometer. While the fluorescent dyes can be used individually, they are more commonly used in a double "live/dead" assay. The cell sample is exposed to fluorescent dye pairs (e.g., calcein AM/EtBr, FDA/PI) where live cells fluoresce green and dead cells fluoresce red.

Another membrane integrity assay detects lactate dehydrogenase release. Lactate dehyrogenase (LDH) is an enzyme normally present in the cytosol and can only be detected outside the cell after cell damage has occurred. In this assay, LDH released from damaged cells oxidizes lactate to pyruvate which promotes the conversion of tetrazolium salt INT to formazan. The amount of LDH

released, determined by measuring the absorption of formazan at 490 nm, is proportional to the number of cells damaged or lysed. Significant LDH release from the cytosol is associated with necrosis.

Mitochondrial Function

Changes in cell function due to toxicity can occur before serious cell damage, such as compromised membrane integrity, are detected. Indicators of metabolic activity can be used to detect early cell injury. Considered the "gold standard" for cytotoxicity, the MTT assay is a colorimetric assay that measures the activity of mitochondrial enzymes. MTT, or 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyl tetrazolium bromide, is yellow in solution but is converted into an insoluble purple formazan product in living cells. Absorbance measurements at 570 nm can be made after solublizing the formazan with DMSO or isopropanol, and the absorbance intensities can be correlated to cell viability. In addition, the absorbance is also representative of cell number and can be used to measure cell proliferation. While MTT is less expensive and more common, other tetrazolium dyes (MTS, XTT, or WST) similarly produce a formazan that absorbs around 570 nm in live cells but have greater sensitivity with less reaction time. Similar to the MTT assay, the alamar blue assay is another colorimetric assay that measures metabolic enzyme activity. The assay is based on the reduction resazurin, a nonfluorescent dye, to resorufin, a pink fluorescent dye, mainly by acting as an electron acceptor for enzymes such as NADP and FADH during oxygen consumption. The converted dye fluoresces at 590 nm and can be detected using a fluorometer. In addition, alamar blue shows little to no cytotoxic effects in tested cells, and thus has been used to perform several tests or kinetic measurements on the same set of cells.

Measurement of ATP levels and oxygen consumption rate are less commonly used methods of assessing nanoparticle cytotoxicity. This could due to the assays being more expensive or requiring specialized equipment. Briefly, ATP bioluminescence assay measures the concentration of ATP, which is proportional to the number of metabolically functional cells, using a luciferin-luciferase system. Healthy cells maintain ATP concentrations through a balance of production and consumption pathways. Since the oxygen consumption rate is disrupted upon cell death, cells lose the ability to produce ATP, and since ATPases can continue to consume ATP after cell death, the concentration of ATP decreases. The decrease in ATP concentration corresponds to a loss of luciferin–luciferase luminescence. The oxygen consumption rate of cells can be measured using electron paramagnetic resonance oximetry, Clark-type polarographic oxygen electrode, and fluorescent oxygen sensors. However, to date, little or no nanoparticle cytotoxicity studies use these approaches.

Proliferation

If nanoparticle exposure induces cell senescence but not cell death, this could be considered a cytotoxic effect as cell proliferation has been disturbed. Comparing the number of cells present at different exposure time points can indicate effects on cell proliferation. Cell number can be measured by direct cell counting, nuclei counting, or by the measurement of total cell protein or DNA content, which are proportional to the number of cells. Fluorescence from dyes that label DNA, such as Hoechst or PI, is directly proportional to the sample DNA content. Compared to the control, higher fluorescence indicates proliferation and lower fluorescence signals cytotoxicity. This method is relatively cheap and fast. In comparison, markers such as ^{3}H-thymidine and 5-bromo-2′-deoxyruidine (BrdU) incorporation into DNA can also be used to measure cell proliferation; however, both methods require significant sample processing. As mentioned previously, MTT can also be used to measure cell proliferation since the dye absorbance is proportional to cell number.

Apoptosis Versus Necrosis

Differentiating the type of cell death can elucidate information about the mechanism of toxicity. For example, the plasma membrane of apoptotic cells remains largely intact, and therefore cell death via

may not be accounted for when using assays based on compromised membrane integrity, such as PI staining. Apoptosis involves a series of signaling cascades that are initiated by the cell after disruption of intracellular homeostasis. During apoptosis, phosphatidylserine is transported from the inner surface of the cell membrane to the outer surface. Annexin V binds specifically to phosphatidylserine and therefore can be used to label apoptotic cells. In contrast, necrosis is passive and results from irreversible cell damage, such as disruption of the plasma membrane. As mentioned previously, PI is a nuclear stain that only fluoresces in cells with compromised plasma membranes. Combined cell staining with annexin V–FITC (fluorescein isothiocyanate) and PI allows for the discrimination of intact cells (no staining), early apoptotic (FITC staining only), and late apoptotic or necrotic cells (both PI and FITC staining). Staining can be observed using fluorescence microscopy or quantified using flow cytometry.

Additional Study Design Considerations

Choice of Cell Type

Most nanoparticle cytotoxicity data is generated from monocultures of cells representative of specific organs of the body. Recently co-culture in vitro systems have been developed to better simulate the cellular environment in vivo [3]. Cell lines recommended in the ASTM E2526-08 and ISO 10993-5 cytotoxicity testing standard protocols include LLC-PK3 porcine kidney cells, HepG2 liver cancer cells, and 3 T3 mouse embryo fibroblasts. Other cell lines often used in nanoparticle cytotoxicity studies include A549 lung cancer cells, HEK293 human embryonic kidney cells, human epidermal keratinocytes, human dermal fibroblasts, and HeLa human epithelial cancer cells. Epithelial cells are often used because they form the outer barrier layer of organs and are usually the first cells to come into contact with nanoparticles during exposure. Regardless of which cell lines chosen, it is important to recognize that there may be differences in sensitivity between cell types [4]. In addition, if primary cells are chosen, these cells may have greater sensitivity than established cell lines.

Suitability for High-Throughput Screening

A considerable amount of time and money can be lost on nanoparticle development when they fail late in the process due to toxicity issues. The growing number of nanoparticles needing to be screened places new demands on cytotoxicity tests, with more attention being drawn to the use of high throughput methods. High-throughput screening (HTS) utilizes automated systems to rapidly conduct in vitro tests on microtiter plates. If multiple measurements are obtained from each well, then the process is referred to as high content screening. Several of the established cytotoxicity assays mentioned have been successfully adapted to 96-well and 384-well plates for use in high-throughput plate reader and flow cytometer setups [5]. Since the advantage of in vitro cytotoxicity screening is the ability to evaluate the toxicity of a large number of compounds, testing nanoparticle libraries with HTS is one strategy toward generating the volume of data needed to identify toxic nanoparticle formulations, elucidate nanoparticle structure-activity relationships, and better prioritize nanoparticles for further in vivo testing.

Potential Mechanisms of Nanoparticle Toxicity

Critical determinants of nanoparticle-induced cytotoxicity include exposed cell type, composition, and size of nanoparticles. In addition, through one-by-one evaluation of nanoparticles, structure-activity relationships specific to certain nanoparticle formulations have been observed. Several reviews provide detailed discussions on general and nanoparticle-type specific cytotoxicity, and recent examples are included in the References section. A summary of the current understanding on the mechanisms behind toxicity observed from six widely used nanoparticles with reference to recent reviews is presented in Table 2. The following section elaborates on some of the effects that are common across nanoparticle types.

Aggregation and Agglomeration

Nanoparticle stability can vary in different dispersion solutions, such as deionized water and cell

Nanoparticle Cytotoxicity, Table 2 Toxicological effects of some widely used NPs

NP	Toxicity observations	Reference
C_{60}	Unmodified C_{60} more toxic than functionalized (hydroxylated and carboxylated) C_{60}, residual THF solvent can contribute to toxicity observed from unmodified C_{60} cluster (nC_{60}) toxicity, photocatalytic behavior contributes to toxicity	[6]
CNTs	Unpurified as-prepared CNTs exhibit toxicity possibly due to presence of catalytic iron, single-walled CNTs elicit more toxic response than multiwalled CNTs	[7]
Au	Relatively "nontoxic," cationic AuNPs more toxic than anionic AuNPs, CTAB and citrate coatings contribute to toxicity but can be mitigated with PEG	[8]
Ag	Silver ion release contributes to toxicity, ROS generation and oxidative stress are two major mechanisms, different surface chemistries induce different DNA damage responses	[9]
QDs	Heavy-metal (e.g., Cd) leaching contributes to toxicity, shell and surface coating can reduce toxicity by containing core and modulating surface charge, toxicity also influenced by oxidative & photolytic stability	[10]
TiO_2	Anatase form ~10 times more cytotoxic than rutile form, photoactivity enhances toxicity, oxidant driven inflammation and protein nitration also associated with TiO_2 NP exposure	[11]

CNTs carbon nanotubes, *NPs* nanoparticles, *THF* tetrahydrofuran, *CTAB* cetyl trimethylammonium bromide, *PEG* polyethylene glycol

culture media. Aggregation, or the strong binding of nanoparticles together, and agglomeration, or the clustering of nanoparticles held together by relatively weak forces, can change the concentration and physicochemical properties of nanoparticles in solution. The properties affected are their size, size distribution, surface area, and therefore their surface reactivity. Since these parameters play a role in the toxicity of nanoparticles, it is important to keep in mind the potential contribution of aggregation and agglomeration during study design and analysis. Methods of reducing nanoparticle aggregation and agglomeration as well as homogeneously dispersing them in liquids include surface modifications and sonication before exposure. While surface modifications may improve nanoparticle stability, they may also influence their bioactivity. The durability of the surface coatings once inside biological environments is another issue, and release of the surface coatings may also contribute to the toxicity of nanoparticles.

Cellular Uptake

Uptake of nanoparticles into cells through endocytosis or phagocytosis processes is size-dependent and can vary greatly among cell types. The amount of nanoparticle internalization by cells can contribute to the observed toxicity as internalized nanoparticles may interact with cell organelles and disrupt cellular equilibrium or homeostasis. Transmission electron microscopy has been the preferred method of studying the cellular uptake of nanoparticles due to its high resolution, resolving the intracellular location of nanoparticles. Most studies have found internalized nanoparticles localize in endosomes or lysosomes. The low pH in endosomes and lysosomes has been suggested to dissolve or degrade nanoparticles releasing their core atoms. This can result in cytotoxicity, especially with metal nanoparticles, since metal ions can disrupt calcium homeostasis and/or generate reactive oxygen species.

Oxidative Stress and Inflammation

In vitro studies have shown that nanoparticles of various compositions generate free radicals and reactive oxygen species (ROS) more than larger particles, due to their higher surface area. ROS can damage cells by inducing oxidative stress, which involves a redox imbalance within cells often due to increased intracellular ROS and decreased antioxidants. The oxidative stress associated with nanoparticle exposure can result from:

- The presence of transition metals (e.g., iron, copper, nickel) catalyzing Fenton reactions.

- Generation of ROS at the particle surface.
- Phagocytosis of nanoparticles by inflammatory cells, such as alveolar macrophages and neutrophils.
- Physical damage due to nanoparticles interacting and/or entering mitochondria, altering mitochondrial function.

The cellular damage resulting from ROS generation and activation of inflammatory cells can also lead to inflammation. Markers of inflammation include the pro-inflammatory cytokines IL-1b, IL-6, and TNF-α plus the chemokine IL-8. Cytokines can be quantified using the enzyme-linked immunosorbant assay (ELISA), and ELISA results have been reported for metal oxides, carbon-based nanoparticle, silica, and quantum dots. However, carbon nanoparticles and metal oxides have also been reported to adsorb cytokines IL-8 and IL-6, interfering with the ELISA assay. These observations should be considered when interpreting results from nanoparticle inflammation testing.

DNA Damage

Of the cellular events that can occur in the absence of cell death, DNA damage is of critical concern as it can initiate disease development, such as cancer, and genetic abnormalities in reproductive cells. Both the nanoparticles themselves as well as nanoparticle-elicited events, such as overload and persistent inflammation, have been suggested to play a role in nanoparticle-induced genotoxicity. However, to date, limited genotoxicity testing of nanoparticles has been conducted. Genotoxicity may result from nanoparticle exposure due to their (a) small size which may facilitate their transport across cell and nuclear membranes where they may interact directly with DNA, (b) high surface area and/or chemical components (e.g., heavy metals) could generate ROS, which damage DNA. The comet assay has been used to detect DNA damage in individual cells using gel electrophoresis. Cells with damaged DNA appear as "comets" with intact DNA residing in the head portion and broken DNA pieces migrating away, forming the tail. A DNA-specific dye such as PI is used to read the gel, and the amount of DNA found in the tail is proportional to the amount of DNA damage. The comet assay is the most widely used genotoxicity test in published nanoparticle studies; however, it only nonspecifically indicates the presence of DNA strand breaks. The most investigated oxidative lesion caused by hydroxyl radical attack, 8-OHdG, has been shown to be induced in vitro by nanoparticles, including carbon nanotubes, CeO_2, gold, and TiO_2 nanoparticles [12]. Isotope-dilution gas chromatography–mass spectrometry (GC-MS) can be used to detect and quantify DNA damage mediated by ROS produced from nanoparticle interactions in cells. This method is more robust compared to the commonly used comet assay since it has been shown to differentiate among more than 20 DNA base lesions. Since oxidatively modified DNA lesions have different mutagenic potential, it is important to identify the types of DNA modifications that occur due to nanoparticle exposure.

Challenges for Future Nanoparticle Cytotoxicity Testing

It is important to consider that in vitro dose–response data serves as a starting point for evaluating the cytotoxicity of nanoparticles. Often in vitro data cannot be extrapolated to nor is representative of what would occur in vivo. For example, cells exposed to nanoparticles suspended in cell culture media poorly reflects inhalation exposure conditions, where lung cells at the air–liquid interface are exposed to aerosols of nanoparticles. In addition, in vitro cell culture systems are controlled environments, often containing a single cell line at low cell density, which may not accurately model normal cell–cell interactions or tissue injury signaling cascades. However, these shortcomings arise from the inherent differences between in vitro and in vivo systems. Cytotoxicity assays also presents unique challenges for nanoparticle testing since particulates differ from soluble chemicals. The following section describes areas for improvement in nanoparticle cytotoxicity testing.

Dose Metric

Compared to soluble chemicals, nanoparticles can agglomerate, aggregate and, as a consequence, settle out of solution depending on their size, concentration, and surface chemistry. These processes can affect the cellular dose, making quantification of the exposure concentration more complicated and less comparable across nanoparticle types. As a result, questions have been raised on the applicability of standard cytotoxicity assay protocols, developed to test soluble chemicals, for nanoparticles. Nanoparticle–cell interactions may be enhanced due to nanoparticles settling to the cell surface, leading to potential nanoparticle overloading of the cells. Resulting cytotoxicity may therefore be due to the higher exposure level from poor stability. It is important to account for this phenomenon when conducting nanoparticle in vitro dose–response assessment. In addition, the exposure dose is often inconsistently reported. There are several dose metrics that can be used, such as mass per volume, particles per volume, mass per surface area, or mass per cell number. To aid study interpretation, enough data should be reported so that the reported values can be converted to different dose metric units.

Most in vitro studies assess dosimetry by observing the dose–response relationship after external introduction of different concentrations of nanoparticles. Yet the internalized concentration of nanoparticles may better correlate to cytotoxicity [13]. A more informative representation of exposure dose would be to quantify the amount of nanoparticles externally attached or internalized by the cells (delivered dose) in addition to the concentration of nanoparticles added to the media (administered dose). Teeguarden et al. and more recently Wittmaack provide thoughtful discussions on how the particulate nature of nanoparticles introduces issues such as gravitational settling that impact dosimetry [14, 15]. Despite these examples, few studies make a distinction between administered and delivered dose. Knowing the external and internal fractions of the delivered dose could also give insights into the mechanisms of nanoparticle toxicity; however, experimental methods to resolve these concentrations need to be developed. Future research should work to correlate the deposited and internalized nanoparticle concentrations to observed cytotoxicity.

Nanoparticle Interference

Due to their high surface area, and in some cases optical properties, the presence of nanoparticles in the test solutions can interfere with cytotoxicity assay results. The increased adsorption capacity and increased catalytic activity resulting from their high surface area can either bind or interact with assay dyes leading to false readings. For example, carbon nanomaterials, especially single-walled carbon nanotubes (SWCNTs), have been reported to interact with cytotoxicity indicator dyes, including MTT, WST-1, Coomassie blue, alamar blue, and neutral red. Adsorption of the dyes to the carbon nanotubes results in either false negative data for the absorbing dyes or false positive data for the fluorescent dyes. Fluorescence quenching by carbon nanomaterials has also been reported with adsorption of the dye suggested as a reason for the quenching response. Other ways nanoparticles can contribute optical interference include physically blocking the emitted light, scattering the excitation light, and adding their own fluorescence signal (e.g., quantum dots or polystyrene beads). The reader is also referred to a review that describes in more detail these and other assay interferences derived from nanoparticles [16, 17]. Future developments could include designing label-free cytotoxicity assays, which would eliminate the issue of nanoparticle–dye interactions. As a safeguard, conducting multiple tests is advantageous to ensure valid conclusions are drawn.

Inter-Laboratory Consistency

While many studies using two or more different cytotoxicity assays have reported reasonable agreement in the results, large discrepancies have also been reported. Lanone et al. showed that for highly toxic or not toxic nanoparticles, inter-laboratory reproducibility was good, with LC_{50} values very similar for toxic nanoparticles [4].

However, their data also indicated that the reproducibility for nanoparticles with intermediate toxicity was lower in comparison. Since the cell lines and methodology were the same between labs, the variation between labs found suggests that there may be other factors unaccounted for in protocol development to ensure reliability in performance. These could include differences in nanoparticle storage methods, age of nanoparticle solution at testing, instrument calibration, and user performance may have contributed to the variability observed. Until data generated in different labs come within an accepted margin of error, inconsistencies in study findings and questions on methodology appropriateness during nanoparticle cytotoxicity assessment will remain.

Conclusion

Cytotoxicity testing constitutes just one of many steps required for nanoparticle biocompatibility and human health risk assessment. Traditionally, in vitro toxicity testing focuses on whether or not exposure to a potentially toxic agent results in cell death. However, changes in cellular function may serve as early signs of cytotoxicity when no cell damage or death may be apparent. As not all disruptive effects result in cell death, more extensive toxicity studies have attempted to determine the sublethal effects of nanoparticles. Given the variety of assays developed for cytotoxicity testing, it is important to verify that the endpoints of the assay used are appropriate for the hypothesis being tested. In addition, many of these assays were developed for testing soluble chemicals and may not be suitable for nanoparticle testing. Evidence of nanoparticle interferences with the indicator dyes has raised questions on the accuracy of currently available nanoparticle cytotoxicity data. Researchers should take care when interpreting the available literature as well as when conducting their own nanoparticle cytotoxicity testing. Ultimately, the goal of nanoparticle cytotoxicity assessments is to provide initial toxicity data to guide further risk assessment, which will inform decision making to protect human health.

Cross-References

- ▶ Genotoxicity of Nanoparticles
- ▶ Quantum Dot Toxicity
- ▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles

References

1. Card, J.W., Magnuson, B.A.: A method to assess the quality of studies that examine the toxicity of engineered nanomaterials. Int. J. Toxicol. **29**, 402–410 (2010)
2. Powers, K.W., Palazuelos, M., Moudgil, B.M., Roberts, S.M.: Characterization of the size, shape, and state of dispersion of nanoparticles for toxicological studies. Nanotoxicology **1**, 42–51 (2007)
3. Stone, V., Johnston, H., Schins, R.P.F.: Development of in vitro systems for nanotoxicology: methodological considerations. Crit. Rev. Toxicol. **39**, 613–626 (2009)
4. Lanone, S., Rogerieux, F., Geys, J., Dupont, A., Maillot-Marechal, E., Boczkowski, J., Lacroix, G., Hoet, P.: Comparative toxicity of 24 manufactured nanoparticles in human alveolar epithelial and macrophage cell lines. Part. Fibre Toxicol. **6**, 14 (2009)
5. Riss, T., Moravec, R., Niles, A.: Assay development for cell viability and apoptosis for high-throughput screening. In: Chen, T. (ed.) A Practical Guide to Assay Development and High-Throughput Screening in Drug Discovery. CRC Press, Hoboken (2009)
6. Johnston, H.J., Hutchison, G.R., Christensen, F.M., Aschberger, K., Stone, V.: The biological mechanisms and physicochemical characteristics responsible for driving fullerene toxicity. Toxicol. Sci. **114**, 162–182 (2010)
7. Cui, H.-F., Vashist, S.K., Al-Rubeaan, K., Luong, J.H. T., Sheu, F.-S.: Interfacing carbon nanotubes with living mammalian cells and cytotoxicity issues. Chem. Res. Toxicol. **23**, 1131–1147 (2010)
8. Alkilany, A.M., Murphy, C.J.: Toxicity and cellular uptake of gold nanoparticles: what we have learned so far? J. Nanopart. Res. **12**, 2313–2333 (2010)
9. Ahamed, M., AlSalhi, M.S., Siddiqui, M.K.J.: Silver nanoparticle applications and human health. Clin. Chim. Acta **411**, 1841–1848 (2010)
10. Pelley, J.L., Daar, A.S., Saner, M.A.: State of academic knowledge on toxicity and biological fate of quantum dots. Toxicol. Sci. **112**, 276–296 (2009)
11. Johnston, H.J., Hutchison, G.R., Christensen, F.M., Peters, S., Hankin, S., Stone, V.: Identification of the mechanisms that drive the toxicity of TiO(2)particulates: the contribution of physicochemical characteristics. Part. Fibre Toxicol. **6**, 33 (2009)
12. Petersen, E., Nelson, B.: Mechanisms and measurements of nanomaterial-induced oxidative damage to DNA. Anal. Bioanal. Chem. **398**, 613–650 (2010)
13. Lewinski, N., Colvin, V., Drezek, R.: Cytotoxicity of nanoparticles. Small **4**, 26–49 (2008)

14. Teeguarden, J.G., Hinderliter, P.M., Orr, G., Thrall, B.D., Pounds, J.G.: Particokinetics in vitro: dosimetry considerations for in vitro nanoparticle toxicity assessments. Toxicol. Sci. **95**, 300–312 (2007)
15. Wittmaack, K.: Novel dose metric for apparent cytotoxicity effects generated by in vitro cell exposure to silica nanoparticles. Chem. Res. Toxicol. **24**(2), 150–158 (2011)
16. Doak, S.H., Griffiths, S.M., Manshian, B., Singh, N., Williams, P.M., Brown, A.P., Jenkins, G.J.: Confounding experimental considerations in nanogenotoxicology. Mutagenesis **24**, 285–293 (2009)
17. Kroll, A., Pillukat, M.H., Hahn, D., Schnekenburger, J.: Current in vitro methods in nanoparticle risk assessment: limitations and challenges. Eur. J. Pharm. Biopharm. **72**, 370–377 (2009)

Nano-particle Image Velocimetry

► Evanescent-Wave Particle Visualization in Nano- and Microchannels

Nanoparticle Self-Assembly

► Electric Field–Directed Assembly of Bioderivatized Nanoparticles

Nanoparticle Tracking Analysis

Matthew Wright
Malvern Instruments, Malvern, Worcestershire County, UK

Synonyms

LM10; NS300; NS500; NTA

Definition

Nanoparticle Tracking Analysis (NTA) is an analytical methodology which facilitates the direct and real-time visualization and characterization of nanomaterials sized between 10 and 1,500 nm in liquid suspension.

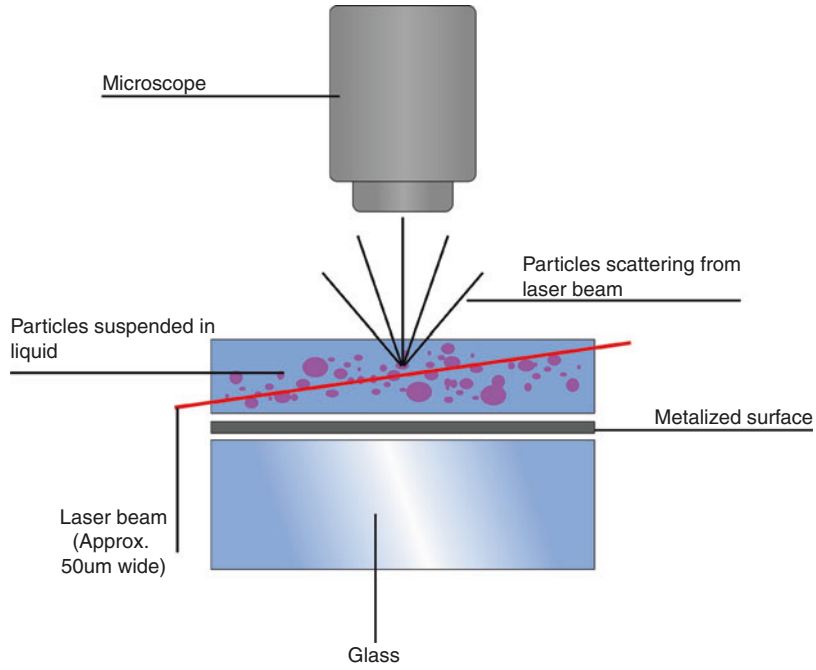
The aim of this entry is to address the need to characterize different properties of particles in the nanoscale, explain the principles of operation of Nanoparticle Tracking Analysis, explore the concentrations and size ranges measurable, look at other parameters characterized by Nanoparticle Tracking Analysis (NTA) as well as functions and applications of the technique.

The use of nanoscale materials is widespread throughout many different industries. Whether this is in pigments for the paint industry, or viruses used in the development and manufacture of vaccines, materials in this size range play a large part in the effectiveness and constancy of many products and processes. As differences in not only the size and concentration, but also shape and both magnetic and electrical charge of these materials can make a substantial difference to the end product, it is clear there is a need for processes and techniques that have the ability to characterize the different properties of these materials. There are numerous techniques that allow different aspects of this combination of properties to be determined, each with their own sets of benefits and limitations. These techniques include dynamic light scattering, also known as photon correlation spectroscopy, analytical disk centrifuge, flow cytometry, transmission electron microscopy, and scanning electron microscopy.

NTA is a multiparameter methodology that allows the direct and real-time visualization, sizing, and concentration determination of materials between 10 nm and 1.5 μm in liquid suspension. As populations of particles can be simultaneously analyzed on an individual basis, it is ideally suited for analysis of polydisperse systems ranging from 10–20 nm to 1–1.5 μm (depending on particle type). Concurrent measurements also allow users to acquire information on concentration, zeta potential, and relative intensity of light scattered and also to analyze fluorescently labeled particles as discussed by Carr et al. [1].

Principles of Operation and Analysis

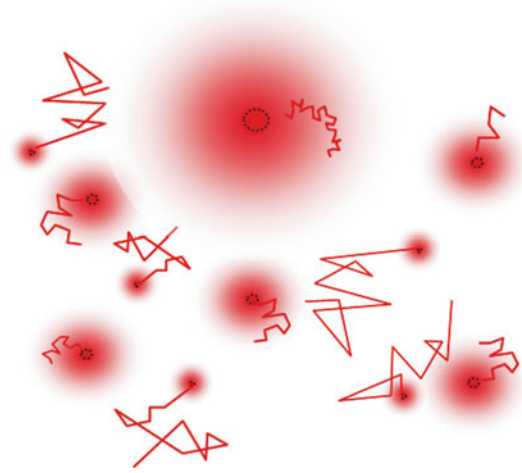
NTA obtains particle size distributions of samples by utilizing properties of both light scattering and Brownian motion.



Nanoparticle Tracking Analysis, Fig. 1 Configuration of components used by NTA in order to visualize nanoparticles (Figure courtesy: NanoSight Ltd.)

A finely focused laser beam (usually from laser diodes operating at 642, 532, 488, or 405 nm) is passed through an optical flat with a beveled edge within a sample chamber. The angle that the laser is introduced to the flat is such that when the beam reaches the interface between the flat and the liquid layer of solvent containing the sample above it, it is refracted almost flat. This compression of the laser to such a low profile gives a very high power density. As such, a region of intense illumination is created within which particles can be visualized by utilizing the light scattered from them by Rayleigh scattering. Figure 1 shows a simplified version of the configuration of laser (running from left to right), glass flat, and objective lens used by systems using NTA. This allows particles to be visualized using light scattered.

A conventional microscope arrangement with a × 20 magnification objective fitted is used in order to visualize the particles in suspension. Coupled with this microscope arrangement, is a video camera (typically a CCD, emCCD, or CMOS). This camera, operating at 30 frames a second, is used to capture a video field of approximately 106 × 80 × 6 μm, within which particles in the sample can be seen moving under Brownian motion. Figure 2 shows the movement of a system of polydisperse nanoparticles under Brownian motion. On average, smaller particles move at a greater rate than larger particles, as can be seen by the relative step lengths of the trails following the particles (each change in direction or "step" indicating one video frame).



Nanoparticle Tracking Analysis, Fig. 2 Showing the movement of polydisperse nanoparticles under Brownian motion (Figure courtesy: NanoSight Ltd.)

Once a field of view containing particles in suspension moving under Brownian motion has been established, the NTA software is used to record a video file of the movement, typically 30–60 s in duration. This video file is then analyzed by the software which automatically finds the center of each scattering point using a proprietary version of a Centroid tracking algorithm to give x- and y-coordinates for each particle.

Equation 1: Centroid tracking algorithm

$$C_X = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\left(x_i \cdot \mathbf{I}_{ij}\right)}{\sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{I}_{ij}}$$

where

- x_i = *Coordinate of pixel on the x-axis*
- $\boldsymbol{I}_{ij}$ = *Pixel intensity [2]* The center of the same scattering point is then found in the next frame, and the distance that the particle has moved in that time is calculated.

Equation 2: Finding displacement of particles between frames (r)

$$r^2 = \Delta x^2 + \Delta y^2$$

By tracking this movement over several consecutive frames, the software determines the mean displacement and hence the diffusion coefficient of each particle within the field of view. Provided that the temperature (T) and the viscosity (η) of the system are known, the Stokes-Einstein equation (Eq. 3) is then used to relate the diffusion coefficient to the hydrodynamic sphere equivalent particle diameter of the material. The Stokes-Einstein equation is as follows:

Equation 3: The Stokes-Einstein equation

$$Dt = \frac{K_B T}{3\pi\eta d_h}$$

where

- Dt = *Diffusion coefficient*
- K_B = *Boltzmann constant*
- T = *Temperature*
- η = Viscosity
- d_h = *Sphere equivalent hydrodynamic diameter particle size*

As Brownian motion occurs in three dimensions, a modified version of the Stokes-Einstein equation is needed for NTA as this technique only observes the motion in two dimensions. However, it is possible to determine Dt from the mean squared displacement of a particle in one, two, or three dimensions (Eq. 4a–c, respectively).

Equation 4: (a, b, and c): Tracking Brownian motion in different numbers of dimensions

$$\overline{(x)^2} = \frac{2TK_B T}{3\pi\eta d} \quad (a)$$

$$\overline{(x, y)^2} = \frac{4TK_B T}{3\pi\eta d} \quad (b)$$

$$\overline{(x, y, z)^2} = \frac{2TK_B T}{\pi\eta d} \quad (c)$$

So in the case of Nanoparticle Tracking Analysis where measurements are made in two dimensions:

Equation 5: Tracking Brownian motion in two dimensions to determine particle size

$$\frac{\overline{(x, y)^2}}{4} = Dt = \frac{K_B T}{3\pi\eta d_h}$$

Equation 5 is applied to the trajectories of each of the particles within the field of view allowing the user to obtain a reading of particle size (hydrodynamic diameter, in nm) against concentration (absolute in particles per ml).

Other Measurable Parameters

Concentration

The field of view interrogated by NTA systems is of known dimensions and therefore known

volume which means that determining the concentration of particles in terms of particles per ml can easily be achieved using the NTA software. As the software not only tracks and sizes every particle within the field of view but also counts the number of particles identified, a concentration can be achieved by extrapolation of this number.

The field of view measures 106 × 80 × 6 μm. This leads to a volume of 50,880 μm^3. Equation 6 shows how this is used in order to determine particle concentration.

Equation 6: Extrapolation of particle count to concentration in particles per ml

$$\text{Concentration} = P_{count} \times \left(\frac{10^{12} \mu m^3}{50,880 \mu m^3} \right)$$

$$\text{Concentration} = P_{count} \times 1.97 \times 10^7$$

where P_{count} = Number of particles within the field of view.

In a 2010 study Filipe et al. investigated the ability for NTA to measure concentration and concluded that the capability of NTA to determine approximate concentration ranges for submicron particles offers an advantage over other techniques for particle sizing [3]. A study conducted by Du et al. found that concentrations obtained using NTA correlated very well with those they calculated using other methodologies [4].

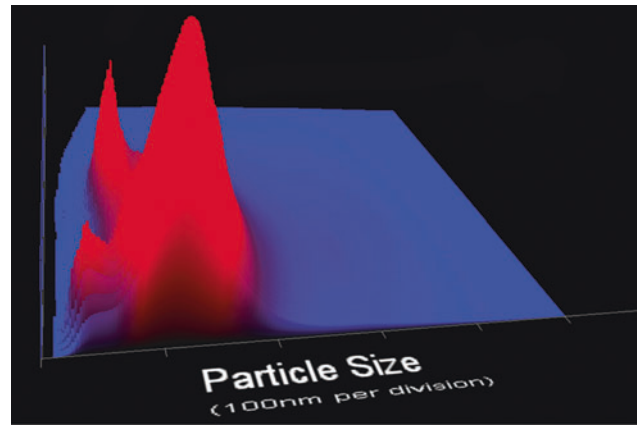
Scatter Intensity

While the size of particles is determined by the NTA software using Brownian motion, the system also measures the relative amount of light scattered by each particle in the scattering volume. These two properties can be plotted as a function of one another allowing differentiation of particles not only by size, but also by composition. For a mixture of similar-sized particles with significantly different refractive indices (e.g., gold particles of 100 nm and polystyrene particles of 100 nm), both would appear in the same size distribution, but can be separated when size is plotted against relative refractive index. Figure 3 shows a distribution of 30-nm and 60-nm gold nanoparticles mixed with a population of 100-nm polystyrene particles, separated by both particle size and the relative amount of light scattered. Particle size is given on the x-axis, with relative intensity on the y-axis and number of particle on the z-axis.

These populations may have merged with one another on a simple particle size versus concentration graph. However, by employing the third dimension of relative intensity of light scattered, it is easy to define the three discrete populations of materials.

Fluorescence

Being a direct visualization and tracking method, NTA does not rely on interference from coherent light sources. This means that fluorescently labeled or inherently fluorescent materials can be



Nanoparticle Tracking Analysis, Fig. 3 Illustrating differentiation of particle material by comparing particle size (x-axis), relative intensity (y-axis), and number of particles (z-axis) (Figure courtesy: NanoSight Ltd.)

analyzed without affecting the results that are obtained. In order to do this, typically a green laser (532 nm), a blue laser (488 nm), or a violet laser (405 nm) is used in place of the usual red laser used within the system allowing the fluorescent excitation of particles with absorption spectra that span these wavelengths. By comparing profiles generated with and without a filter removing the excitation wavelength present in the optical path of the system, the user is able to determine the effectiveness of their labeling process, or the proportion of labeled to unlabelled material within their sample, and whether or not the presence or absence of fluorescence is in any way related to the hydrodynamic sphere equivalent particle size.

Suspension Viscosity

NTA can also be used to determine unknown viscosities of fluids. The system uses the Stokes-Einstein equation in order to relate the diffusion coefficient of the material to the sphere equivalent particle size. However, if the size of the particles in question is already well characterized, then it is possible to rearrange the Stokes-Einstein equation as shown in Eq. 7 and use the known parameters of diffusion coefficient, particle size, and temperature in order to use the collected data to determine the viscosity of the system. This will be briefly discussed in the section on drug delivery studies.

Equation 7: Stokes-Einstein equation to find viscosity

$$\eta = \frac{K_B T}{3\pi D_t d_h}$$

where

- D_t = *Diffusion coefficient*
- K_B = *Boltzmann constant*
- T = *Temperature*
- η = Viscosity
- d_h = *Sphere equivalent hydrodynamic diameter particle size*

Zeta Potential

The zeta potential of a system is a measure of charge stability and has an effect on all particle-particle interactions within a system. Understanding zeta potential is of critical importance in controlling dispersion and determining the stability of a nanoparticle suspension. The zeta potential is the electric potential at the slip plane between the bound layer of solvent molecules surrounding the particle and the overall solution. This can be closely linked to the particle's surface charge in simple systems but is also heavily dependent on the properties of the solvent in which the particles are suspended. A higher level of zeta potential results in greater electrostatic repulsion between the particles, which results in less aggregation and flocculation.

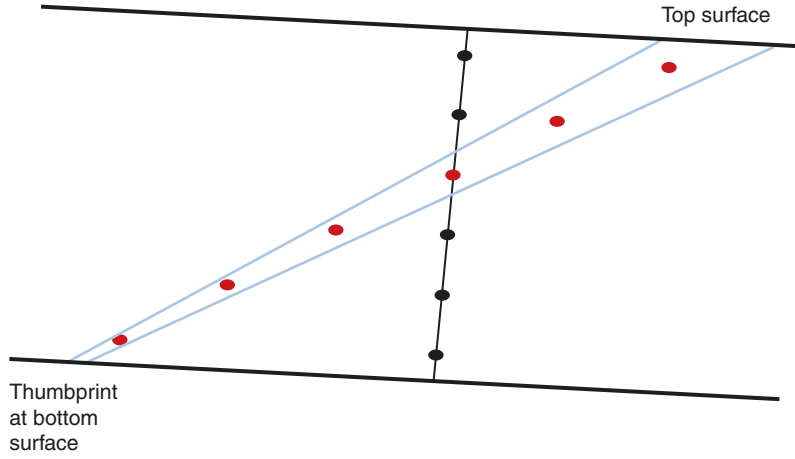
In order to carry out investigations into the zeta potential of a system, the hardware used for NTA is adapted so that the sample chamber is fitted with platinum electrodes through which an electric field can be applied to the cell.

This field causes motion of both the sample particles (electrophoresis) and the diluent (electroosmosis). The NTA software records the apparent drift velocity for each tracked particle, which will be a superposition of both electrophoresis and electroosmosis. This total velocity is measured at different depths within the sample chamber (as shown in Fig. 4, each dot corresponding to a depth of measurement), making it possible to separate these components and obtain a measurement of the electrophoretic velocity (due to the force impinged directly on the particles), and hence the zeta potential of the particles.

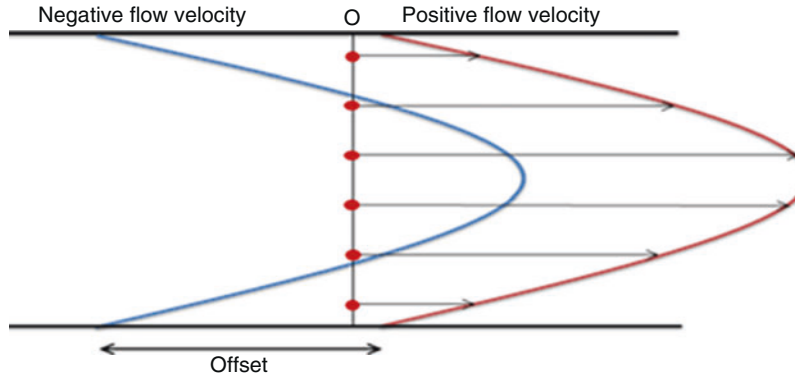
By employing the assumption that the net electroosmotic flow summed over the chamber depth will be zero, as should be the case for a closed system, the electroosmotic profile can be inferred, as shown in Fig. 5.

Using the calculated flow profile, the velocity component caused by electroosmosis is removed from the tracked velocities for each particle. The corrected drift velocities then give a measure of the electrophoretic mobility of individual particles.

The electric field strength (E) within the flow cell is determined using the voltage (V) applied through the sample by the electrodes and the distance between the electrode surfaces (d) (Eq. 8).



Nanoparticle Tracking Analysis, Fig. 4 Diagram showing profile map positions where six standardized depth locations intersect the laser beam (Figure courtesy: NanoSight Ltd.)



Nanoparticle Tracking Analysis, Fig. 5 Total velocity profile measured using NTA (*red*) and electroosmotic velocity profile (*blue*) inferred by subtraction of a contact offset to obtain a velocity profile that sums to zero over the channel depth (Figure courtesy: NanoSight Ltd.)

Equation 8: Determining electric field strength

$$\mathrm{E} = \frac{\mathrm{V}}{d}$$

The particle velocities can then be converted into electrophoretic mobilities (velocity divided by electric field strength).

By application of the Henry equation using the Smoluchowski approximation (appropriate for aqueous diluent media with moderate electrolyte concentration, seen in Eq. 9), the zeta potential (ZP) for each particle can be calculated:

Equation 9: Calculation of zeta potential

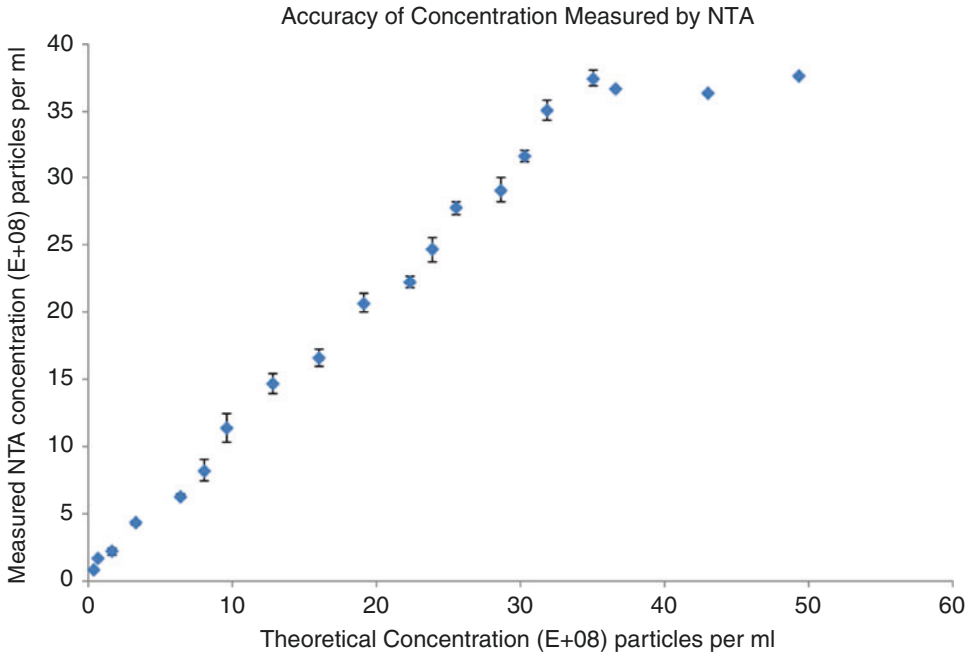
$$ZP = \frac{\mu\eta}{\epsilon_0\epsilon_r}$$

where μ is the electrophoretic mobility of the particle, ϵ_0 is the permittivity of free space, ϵ_r is the relative sample solution permittivity, and η is the sample solution viscosity.

This particle-by-particle approach allows different charge populations within a sample to be determined rather than being lost in the "noise" of an ensemble technique.

Concentration and Size Ranges Measurable

The ability of NTA to analyze polydisperse samples comes from the ability to track and analyze each particle in the system individually, irrespective of the others around it. It is therefore important that a sufficient number of particles are analyzed during analysis to ensure that the data produced is statistically viable. Repeatable and accurate results can be obtained when analyzing



Nanoparticle Tracking Analysis, Fig. 6 Relationship between absolute concentration (x) versus that measured by NTA(y) (Figure courtesy: NanoSight Ltd.)

particles in concentrations between 10^7 and 10^9 particles per ml. Below this concentration, the number of particles within the field of view is so low that the typical analysis time of 30–60 s will only allow a very small number of particles to be analyzed, leading to inaccuracies of a statistical nature. Above 10^9 particles per ml, the field of view becomes very crowded. As Brownian motion is a random process, when two particles cross over, the tracking software is unable to determine which particle is which and must therefore stop tracking both to avoid erroneous tracking. When the field of view is crowded, the rate at which crossover of particles occurs is so great that there is insufficient time to determine diffusion coefficient and therefore size.

A high concentration of particles also leads to erroneous results when trying to identify solution concentration. Figure 6 shows the relationship between the concentration of particles in a solution and the concentration as measured using NTA.

It is clear that the correlation between the actual and measured concentration remains stable until a concentration of around 3.5×10^9 particles per ml whereupon the accuracy of the NTA concentration reading begins to deteriorate.

The typical size range between which NTA operates is from 10 to 20 nm up to 1.5 μm. The lower limit of this figure is mainly dependent upon the amount of light scattered by the particles. This value in turn varies in relation to particle size, shape, refractive index (real and imaginary), illumination power, wavelength, power, and polarization as well as solvent refractive index. The amount of light falling on the detector also fluctuates according to many variables including efficiency of optics, sensitivity, and spectral response of the camera.

In their 1983 work, Bohren et al. comprehensively describe the theory of light scatter and the formula for Rayleigh scattering of small particles [5]:

Equation 10: Rayleigh scattering of light by small particles

$$\frac{\mathbf{I}}{\mathbf{I}_{in}} = \frac{16\pi^4 a^6}{r^2\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)\sin^2\psi$$

where

- a = *Particle radius*
- n_1 = *Refractive index of particle*
- n_2 = *Refractive index of solvent*
- λ = Wavelength of light
- I_{in} = *Incident power per unit area*
- I = *The scattered power per unit area*
- r = *Distance from the scattering region and*
- ψ = Angle between input polarization and scattering direction

The total scattering (P_{scat}) into an aperture of collection angle θ (numerical aperture NA = sin θ) is then:

Equation 11: Total light scattered into an aperture

$$\mathbf{P}_{\text{scat}} = \frac{64\pi^4 a^6}{\lambda^4}\left(\frac{n^2-1}{n^2+2}\right)\eta_0 \mathbf{I}_{\text{in}}$$

where $\eta_0 = \frac{(1-\cos\theta)}{4} + \frac{1-\cos^3\theta}{12}$

As the system has fixed laser wavelength and power as well as a fixed aperture and detection angle, the variables that affect the lower limit of detection consist primarily of the particle size and the relative refractive index difference between the particles and the solvent in which they are suspended.

As light scattering at these scales varies as a function of radius [6], as particles reduce in size, they scatter considerably less light. The functionality of NTA relies on the ability to visualize the particles, and so if the amount of light scattered is not sufficient that the particles can be seen, then it is not possible to size them. For strongly scattering materials such as colloidal gold, the lower limit of detection is around 10 nm. For those materials with a weaker scattering potential, such as biological materials, the lower limit of detection is more likely to be 20–40 nm. The upper size limit of the system is determined by the rate of Brownian motion. When particles reach a certain size, the rate of Brownian motion becomes so low that inherent centering errors in the software start to negate the tracking of the movement leading to inaccuracies in sizing. For dense, high Ri particles such as gold, their rapid sedimentation rate and excessive scattering restrict the upper size at which they can be analyzed to approximately <500 nm. This limit is typically 1–2 μm.

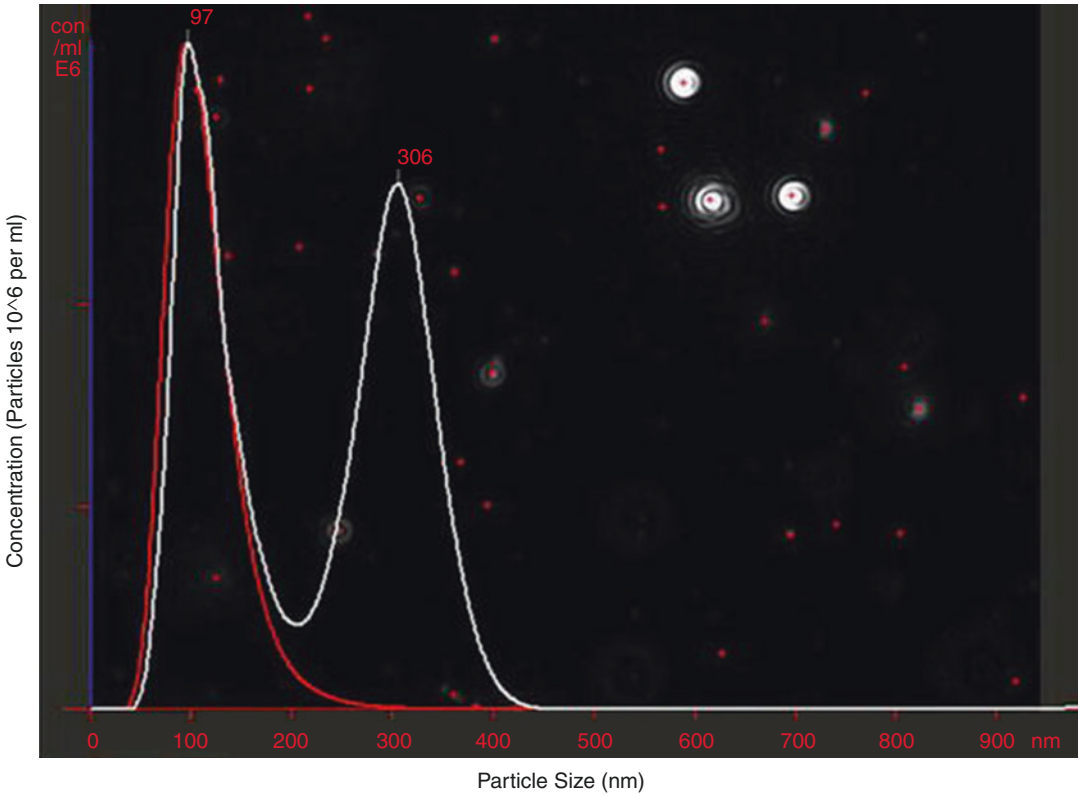
Applications of Nanoparticle Tracking Analysis

Owing to the varied nature of work at the nanoscale, NTA has been found to be of use across a variety of different applications.

Virology and Vaccine Production. Due to the nature in which NTA tracks, sizes, and counts particles on an individual basis, the methodology allows users to obtain an immediate and direct estimation of product purity and concentration. This also allows researchers and manufacturers to monitor the degree and rate of formation of aggregates within a virus preparation, complementing the process of product optimization and purification. An example of the use of NTA for virus purification studies can be seen in Fig. 7. This shows the difference between a partially purified virus preparation (white line) and the same sample having undergone a successful purification protocol, removing all the aggregated material or contaminates (red line). In this figure, the x-axis represents the particle size and the y-axis represents the particle concentration in virus particles per ml.

Exosomes. Originally, it was thought that microvesicles (typically in the 100 nm to 1 μm range) and nanovesicles (also known as exosomes, with a size range of around 30–100 nm) were cellular debris. However, it is now appreciated that these cell-derived particles are of significant importance and play an important role in cell-to-cell communication and cell signaling.

Recent research suggests that exosomes are present in significantly elevated concentrations in samples taken from patients with a number of diseases including preeclampsia, diabetes, coronary artery disease, and cancer. NTA is not only able to visualize, size, and determine concentration of these materials in liquid suspension, but when used with specific fluorescent labels, for example, antibody-mediated fluorescent labeling,



Nanoparticle Tracking Analysis, Fig. 7 Illustrating the use of NTA in virus purification processes with purified (*red line*) against non-purified (*white line*) product shown (Figure courtesy: NanoSight Ltd.)

it can be used to selectively analyze specific populations within a complex sample.

In their evaluation of systems for the analysis of exosomes in liquid suspension, Harrison et al. concluded that NTA's single-particle tracking and sizing offers advantages over bulk scattering measurements found in other light scattering methodologies [6].

Nanoparticle Toxicology and Ecotoxicology. Current research suggests that the biological response to nanoparticles in living organisms, and thus any potential health implications, can be affected by particle size, composition, shape, aggregation state, and exposure route. Before any studies into these effects can be carried out, it is imperative that the particles used are as well characterized as possible, in particular their size distribution and concentration in appropriate test media. Researchers are also interested in what happens when engineered nanoparticles are released into the natural environment.

Protein Aggregation. Biological activity and immunogenicity of a biopharmaceutical product are both highly susceptible to the state of protein aggregation. These aggregates span a broad range of sizes, from the small oligomers (a few nanometers) to insoluble micron-sized aggregates that can contain millions of monomer units. Determining and understanding the state of aggregation in these systems can not only help in improving product stability, but can also lead to efficiency gains and aid in process optimization. Individual protein monomers are far smaller in size than the lower resolving limit of the Nanoparticle Tracking Analysis methodology. When a solution containing only protein monomers is viewed using this technique, a background "haze" is seen within which individual monomers cannot be resolved. However, when aggregates start to form within the system, and reach around 25–30 nm in size, it is possible to differentiate them from the background signal and count and

size them as the technique normally would. This, combined with the possibility of controlling the temperature within the sample cell, from 10° below ambient to 60 °C, can allow characterization of the parameters that cause aggregation within the product. A full critical evaluation of the NTA technique by Filipe et al. concluded that the technique was suitable for the analysis of protein aggregates; however, care should be taken that sample preparation does not affect the aggregate distribution.

Drug Delivery. By utilizing Nanotechnology, it is possible to deliver drugs to a specific area of the body. Localized drug delivery allows the required drug volume to be reduced, along with any associated negative health implications. Both poorly soluble and fragile drugs can be carried using drug delivery vectors such as degradable polymeric nanoparticles or liposomes. It is clear that both the particle size and concentration will affect the payload, and so Nanoparticle Tracking Analysis provides a unique way to monitor these properties in liquid suspension. The location of drug delivery can also have an impact on the effectiveness of the drug. If particles are unable to pass through biological barriers, then these can act as natural filters, preventing the drug from reaching the desired location. By using Nanoparticle Tracking Analysis to compare concentrations of particles before and after passing through membranes and barriers, researchers are able to study the ability for the drug delivery vector to reach the desired location.

Nanobubbles. Nanobubbles are stable pockets of gas in liquid suspension which can be applied to drug delivery, cleaning, and sterilization. As Nanoparticle Tracking Analysis allows direct visualization, sizing, and concentration determination of the nanobubbles in liquid suspension, users are able to characterize their product in the final liquid state. This could prove to be difficult when using other techniques such as electron microscopy, due to the preparation required in order for electron microscopy to function.

Inks and Pigments. In the field of inkjet applications, the particle size and concentrations of pigment particles are very important. The particle size has a great effect on the strength of the color generated, and aggregation of the particles could lead to the blockage of nozzles. The high-resolution size distribution that Nanoparticle Tracking Analysis can provide allows both studies into the relationship between size and color to be analyzed, as well as the conditions that lead to aggregation of pigment particles.

In conclusion, Nanoparticle Tracking Analysis is an analytical technique which allows the direct visualization, sizing, and concentration determination of materials between 10 and 1,500 nm in liquid suspension. The technique works by tracking the Brownian motion of each particle within the field of view, irrespective of the others, and relating this movement to the sphere equivalent hydrodynamic diameter of the particle. This parameter is calculated for each particle within the field of view, and a particle size distribution is produced. In addition, particle concentration, light scattering intensity, zeta potential, and fluorescent properties can also be measured. The technique is of use in a variety of different applications, both as a competing and complementary technique to other particle sizing methodologies.

Cross-References

- ▶ Chitosan Nanoparticles
- ▶ Detection of Nanoparticulate Biomarkers in Blood
- ▶ Dye-Doped Nanoparticles in Biomedical Diagnostics
- ▶ Liposomes
- ▶ Nanoencapsulation
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Perfluorocarbon Nanoparticles
- ▶ Toxicology: Plants and Nanoparticles

References

1. Carr, B., Hole, P., Malloy, A., Nelson, P., Wright, M., Smith, J.: Applications of nanoparticle tracking analysis in nanoparticle research – a mini review. Eur. J. Parenter. Pharm. Sci. **14**(2), 45–50 (2009)

2. Cheezum, M.K., Walker, W.F., Guilford, W.H.: Quantitative comparison of algorithms for tracking single fluorescent particles. Biophys. J. **81**, 2378–2388 (2001)
3. Filipe, V., Hawe, A., Jiskoot, W.: Critical evaluation of nanoparticle tracking analysis (NTA) by nanosight for the measurement of nanoparticles and protein aggregates. Pharm. Res. (2010). doi:10.1007/s11095-010-0073-2
4. Du, S., Kendall, K., Morris, S., Sweet, C.: Measuring number-concentrations of nanoparticles and viruses in liquids on-line. J. Chem. Technol. Biotechnol. (2010). doi:10.1002/jctb.2459
5. Bohren, C., Huffman, D.: Absorption and Scattering of Light by Small Particles. Wiley, New York (1983)
6. Harrison, P., Dragovic, R., Albanyan, A., Lawrie, A., Murphy, MF and Sargent, I.: Application of dynamic light scattering to the measurement of microparticles. In: 55th Annual Meeting of the Scientific and Standardization Committee of the ISTH, The XXII Congress of The International Society on Thrombosis and Haemostasis, Boston, 11–18 July 2009 (in press)

Nanoparticle–Adsorbome

► Nanoparticle–Biocorona

Nanoparticle–Biocorona

Jonathan H. Shannahan
Department of Pharmaceutical Sciences, University of Colorado – Anschutz Medical Campus, Aurora, CO, USA

Synonyms

Nanoparticle–adsorbome; Nanoparticle–biomolecular corona; Nanoparticle–protein corona

Definition

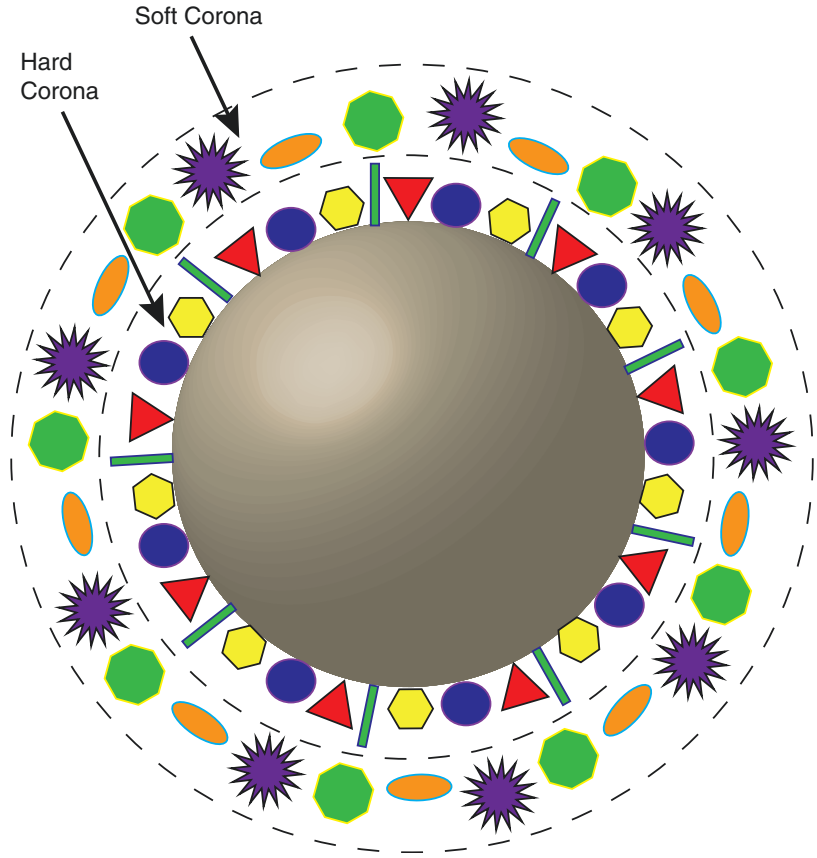
Nanoparticle–biocorona: a layer(s) of biomolecules that adsorbs to the surface of a nanoparticle in a biological environment.

Introduction

Following their introduction into a physiological environment, the surface of nanoparticles readily associates a variety of biomolecules including proteins, peptides, and lipids forming a nanoparticle–biocorona. This biocorona can be subdivided into two portions, the "hard" which consists of tightly bound biomolecules and "soft" that consists of loosely associated biomolecules (Fig. 1). The "soft" biocorona sits on top of the "hard" biocorona and is more dynamic in nature with biomolecules rapidly being exchanged. The composition of the biocorona is dependent on the physicochemical properties of the nanoparticle as well as the identity of the biological environment. The addition of the biocorona can modify numerous nanoparticle characteristics including hydrodynamic size, charge, and state of aggregation; and most importantly it modifies the nanoparticle biological interface. The addition of the biocorona, and the new interface it imparts, mediates subsequent physiological interactions with additional biomolecules and cellular membranes and receptors. Ultimately, the addition of the biocorona on the surface of nanoparticles can alter their activity, biodistribution, pharmacokinetics, cellular uptake, toxicity, and clearance. Understanding and controlling the biocorona is necessary to enhance the therapeutic efficacy of nanomedicines while also mitigating unintended toxicity.

Methods to Evaluate Coronal Structure and Composition

The evaluation of the nanoparticle–biocorona typically is performed by introducing a nanoparticle into a biological environment or by collecting samples such as serum, plasma, gastrointestinal fluid, or alveolar lavage fluid, which represent the in vivo environment. Following introduction into a biological environment, nanoparticles are often isolated for analysis through centrifugation or size exclusion chromatography. The structure and properties of the biocorona are evaluated by assessments of biocoronal thickness, charge,



Nanoparticle–Biocorona, Fig. 1 Graphical representation of the nanoparticle–biocorona. The "hard" biocorona is composed of tightly bound biomolecules, which do not readily dissociate from the nanoparticle. The "soft" biocorona is composed of loosely bound biomolecules and is dynamic in nature with biomolecules constantly being exchanged

strength of protein interactions, and protein conformational changes. The thickness of the biocorona on the surface of the nanoparticle can be studied through the use of dynamic light scattering, size exclusion chromatography, differential centrifugal sedimentation, and/or transmission electron microscopy. Alterations in surface charge or zeta potential can be evaluated by microelectrophoresis. The strength of protein interactions can be studied through the use of isothermal titration calorimetry and surface plasmon resonance, whereas circular dichroism and Fourier-transform spectroscopy and computational modeling are often used to determine changes in protein structure due to association with the nanoparticle. Evaluation of the compositional biomolecular content of the biocorona is performed by isolating biomolecules from the surface of the nanoparticle and utilizing SDS-PAGE and/or liquid chromatography–tandem mass spectrometry for the identification and quantification of individual components. Removal of the "hard" corona for biomolecule identification is difficult due to the strong association between the biomolecules and the nanoparticle, whereas the "soft" corona can easily be lost during processing due to its dynamic nature and weak interactions.

Directing Protein Adsorption for Nanomedicine Targeting

The field of nanomedicine is beginning to utilize the nanoparticle–biocorona to increase the targeting capability of nanoparticles for drug delivery. Typically, the formation of the biocorona has been associated with a decrease in cellular uptake; however, biocoronas formed of specific

ligands to direct nanoparticles to particular receptors have shown promise under in vitro experimentation. In contrast, the biocorona may have unintended consequences on nanoparticle delivery. For example, the enhanced targeting ability of nanoparticles with a transferrin biocorona was lost in vivo due to the addition of the endogenous biocorona [1]. The complexity of the biocorona as it pertains to nanoparticle cellular uptake will influence the targeting of nanomedicines. It has been demonstrated that the association of a specific protein with nanoparticles of differing size and surface coatings can have vastly different effects on cellular uptake, thereby modifying the capacity to effectively target nanotherapeutics. For example, the formulation of an IgG biocorona on 20 nm silica-coated silver nanoparticles increases cellular uptake in human epidermal keratinocytes. In contrast, an IgG biocorona on 110 nm silica-coated silver nanoparticles had no effect on uptake, whereas cellular uptake was reduced by changing the surface coating to citrate on the 110 nm silver nanoparticle with an IgG biocorona [2]. Instead of preloading the biocorona onto nanoparticles, it may be more feasible to synthesize nanoparticles with specific physicochemical characteristics, which allow for the enrichment of certain biomolecules, resulting in a biocorona that facilitates targeted uptake. For example, PEG-coated polyhexadecyl cyanoacrylate nanoparticles have been shown to translocate from the circulation into the brain, whereas non-PEG-coated polyhexadecyl cyanoacrylate did not readily cross the blood–brain barrier [3]. Upon investigation, this differential uptake was related to the identity of the biocorona, specifically the preferential abundance of apolipoproteins in the biocorona of PEG-coated nanoparticles.

Adverse Biological Implications of the Nanoparticle–Biocorona

The association of the biocorona with nanoparticles may result in unintended adverse biological consequences. Binding of biomolecules to the surface of nanoparticles can result in protein modifications and conformational changes, which may facilitate toxicity through activation of the immune system. Many nanoparticles have been shown to associate with biocoronas that are rich in immunoglobulins and complement proteins, which act as opsonins, facilitating uptake of nanoparticles by phagocytic cells (initiating an inflammatory response) as well as resulting in increased biodistribution to the liver and spleen. Albumin biocoronas however have been shown to increase retention time in the circulation, whereas apolipoprotein-rich biocoronas cause increased distribution into the brain. These effects demonstrate biocorona-dependent differences in the biodistribution of nanoparticles, which need to be taken into consideration for their effective utilization in biomedical applications to reduce off target effects. The addition of the biocorona can also have unintended consequences on the release of drugs from nanoparticles by not only interfering with biodistribution but also possibly causing premature or delayed release of drugs from the nanoparticle carrier [4].

Future Directions

Recent research efforts have begun evaluating how to regulate the nanoparticle–biocorona for applications in nanomedicine. However, most of the work has been done to identify the protein content of the biocorona, while less emphasis has been placed on other biomolecules such as lipids and other macromolecules. Many studies have also evaluated the influence of diverse nanoparticle physicochemical characteristics on the composition of the nanoparticle–biocorona; however, modifications in the physiological environment, which may influence the nanoparticle's pharmacokinetic profile, still require extensive study. Further, methods and investigation of the more dynamic "soft" corona need to be performed. Most studies also focus on an individual time point of biocorona formation; however, the kinetics of the biocorona lacks sufficient

analysis. Furthermore, since the "soft" corona is more dynamic with almost a constant exchange of biomolecules, studies need to be performed, addressing the impact of possible biomolecule conformational changes and oxidative modifications and the likely resulting biological consequences. Currently, the evaluation of the biocorona is limited to in vitro assessment; however, extensive in vivo analysis is needed, particularly in human populations of various underlying disease states. Ultimately, for the future development of effective nanomedicines, it will be necessary to predict the identity and the biological impact of the biocorona based on the physicochemical properties of the nanomaterial.

Cross-References

- ▶ Cellular and Molecular Toxicity of Nanoparticles
- ▶ Drug Delivery System
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Transmission Electron Microscopy
- ▶ Uptake/Internalization/Sequestration/Biodistribution

References

1. Salvati, A., Pitek, A.S., Monopoli, M.P., Prapainop, K., Bombelli, F.B., Hristov, D.R., Kelly, P.M., Aberg, C., Mahon, E., Dawson, K.A.: Transferrin-functionalized nanoparticles lose their targeting capabilities when a biomolecule corona adsorbs on the surface. Nat. Nanotechnol. **8**(2), 137–143 (2013)
2. Monteiro-Riviere, N.A., Samberg, M.E., Oldenburg, S. J., Riviere, J.E.: Protein binding modulates the cellular uptake of silver nanoparticles into human cells: implications for in vitro to in vivo extrapolations? Toxicol. Lett. **220**(3), 286–293 (2013)
3. Kim, H.R., Andrieux, K., Gil, S., Taverna, M., Chacun, H., Desmaele, D., Taran, F., Georgin, D., Couvreur, P.: Translocation of poly(ethylene glycol-co-hexadecyl) cyanoacrylate nanoparticles into rat brain endothelial cells: role of apolipoproteins in receptor-mediated endocytosis. Biomacromolecules **8**(3), 793–799 (2007)
4. Paula, A.J., Araujo Junior, R.T., Martinez, D.S., Paredes-Gamero, E.J., Nader, H.B., Duran, N., Justo, G.Z., Alves, O.L.: Influence of protein corona on the transport of molecules into cells by mesoporous silica nanoparticles. ACS Appl. Mater. Interfaces **5**(17), 8387–8393 (2013)

Nanoparticle–Biomolecular Corona

▶ Nanoparticle–Biocorona

Nanoparticle–Protein Corona

▶ Nanoparticle–Biocorona

Nanoparticles

Paola Martino[1], Paolo Allia[2] and Alessandro Chiolerio[3]
[1]Politronica Inkjet Printing Technologies S.r.l, Torino, Italy
[2]Materials Science and Chemical Engineering Department, Politecnico di Torino, Torino, Italy
[3]Physics Department, Politecnico di Torino, Torino, Italy

Synonyms

Nanopowders; Synthesis of nanoparticles

Definition

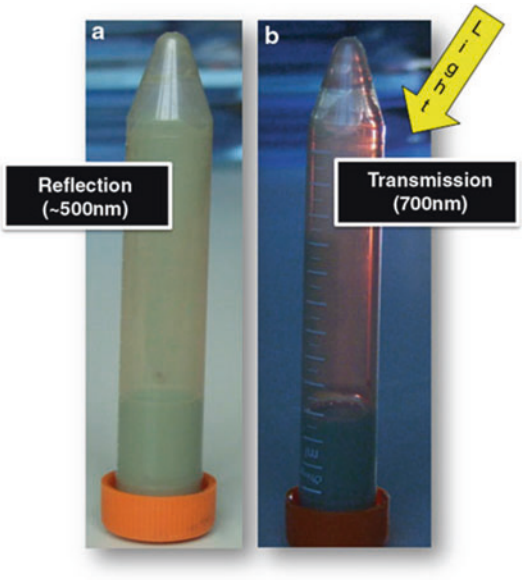
Ultrafine particles featuring a size distribution in the nanometer range.

Introduction

Studies of nanostructured materials have been steadily developing in recent years; in particular, metallic, magnetic, and oxide nanoparticles (NPs) have become a subject of increasing scientific interest. At the present time they hold a place of great importance that covers a wide range of potential applications including biomedical, optical, and electronics.

Although NPs are generally considered an invention of modern science, they actually have a very long history: centuries ago, NPs of silver, gold, and copper were used by artisans to decorate pottery; the first examples date from the ninth century A.D. and were found in Mesopotamia [1]. An important archeological finding is a Roman cup, currently at British Museum in London, called the Lycurgus cup (fourth century A.D.). It contains tiny amounts of colloidal gold and silver, which give it unusual optical properties: when illuminated from outside, it appears green; when retro-illuminated by an inner source, it glows red. An example of this property is well represented in Fig. 1a, b that shows a special silver colloidal NPs ink (by Politronica® Inkjet Printing Technologies S.r.l.) in emitted (green) and transmitted (red) light.

In addition, NPs exist widely in the natural world, for example, as photochemical by-products generated by plants and algae. They have also been created for thousands of years as by-products of combustion and cooking, and more recently from industrial and vehicle exhausts [2].



Nanoparticles, Fig. 1 Silver colloidal NPs ink (by Politronica® Inkjet Printing Technologies S.r.l.) in reflected light (**a** – *green*) and in transmitted light (**b** – *red*)

Definition of Nanoparticles

The term "nanoparticles" (NPs) is a very general one and mainly identifies ultrafine particles in the size of a nanometer. The definition of NPs differs depending upon the materials, fields, and application one refers to. In a narrow sense, they are regarded as the particles smaller than 10–20 nm but, generally speaking, all particles in the 1 nm–1 μm range may still be defined as NPs [3].

The principal parameters of NPs are their shape (including aspect ratios where appropriate), size, and the morphological substructure of the substance.

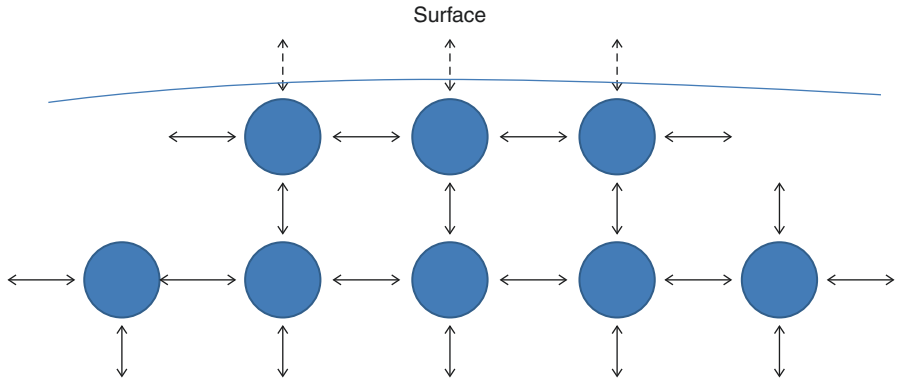
"Size Effect" and Size-Dependent Properties

NPs represent a bridge between bulk materials and atomic, ionic, or molecular structures. Bulk materials should have constant physical properties regardless of its size, whereas at the nanoscale, size-dependent properties are often observed.

The changes in the physical properties with particle size (from bulk to NPs) are called "size effects": by decreasing particle size, the solid particles generally tend to show different properties from the bulk material due to the increased importance of surface upon bulk. Even physical properties like melting point and dielectric constant may change. In addition, it is important to underline that the "size effect" not only allows one to discriminate between physical properties of the bulk when compared to NPs, but also to distinguish among NPs having different size distributions.

For bulk materials larger than one micrometer, the fraction of atoms, ions, or molecules at the surface is insignificant in relation to the fraction corresponding to its volume. In NPs the fraction of atoms, ions, molecules located at the surface becomes greater and, therefore, more active because it is subject to asymmetric interaction forces due to a different bonding state, the bonds at the interface between solid and vacuum (or liquid or atmosphere) are broken (Fig. 2). This affects either the properties of the particle itself or its interaction with other materials.

One of the most important size-dependent consequences of the "size effect" of NPs is the drastic reduction of the *melting point* of ultrafine particles



Nanoparticles, Fig. 2 Schematic representation of surface of a NP. The bonds on the solid surface are broken up

if compared to that of bulk material: the high surface area to volume ratio of NPs provides a big driving force for diffusion. NPs, in fact, tend to be able to move easier at lower temperatures. For example, the melting point of gold is 1,336 K, but if the size of material decreases to 2.5 nm, it becomes about 570 K [3].

A second size-dependent feature of NPs is the *dielectric constant*. The use of NPs embedded in electronic devices allows one to improve the product performance making smaller devices with high dielectric constants. $PbTiO_3$ represent a significant example of this feature because its dielectric constant tends to increase considerably as the particles become smaller than about 20 nm [3].

NPs often possess unexpected *optical properties* as they are small enough to absorb the light with a specific wavelength: a coherent oscillation of the conduction band electrons is induced by interaction with the optical electromagnetic field. Consequentially, a transmitted light with different color depending upon the kind of metal and particle size is obtained. This resonance is called Surface Plasmon Resonance (SPR) (Fig. 3) [4].

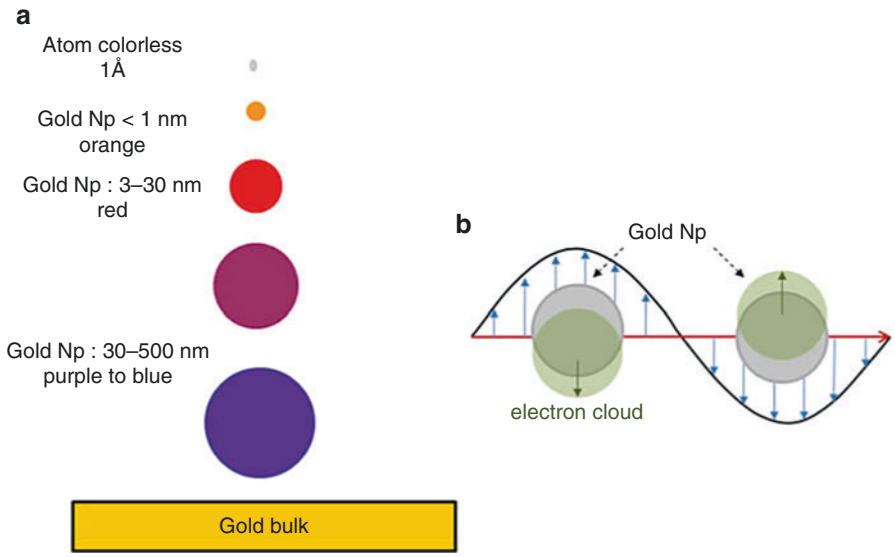
Another size-dependent feature of NPs are *mechanical properties*. Nanocomposites with nanoscale dispersed phases and nanocrystalline materials provide quite different mechanical properties when compared to conventional microstructures. The hardness of the crystalline materials generally increases with the decreasing of crystallite size and the mechanical strength of the metal and ceramic materials considerably increases by decreasing grain size down to the nanoscale or by making composites at the nanoscale [5].

The NPs are seldom used by themselves; they are dispersed instead in other materials or combined with them. The dispersion process is a crucial point for the NPs technology as well as their preparation methods, because the performance of final products is affected by their dispersion conditions. In this way, it is expected with great possibility to develop various new materials and applications by producing and processing the NPs, which have different properties from the bulk material due to size effects as mentioned above [3].

Synthesis Methods

Synthesis methods for NPs are typically grouped into two categories: *top-down* and *bottom-up* approaches (Fig. 4). The first approach involves the breaking down of the bulk material into nano-sized structures or particles. This approach may address milling or attrition, chemical methods, and volatilization of a solid followed by condensation of the volatilized components. Metal NPs so produced are usually large in size (close to 1 μm) and have wide size distribution. High-energy ball milling is shown below as an example of such a method.

The alternative approach, *bottom-up*, refers to the buildup of a material from the bottom: atom-by-atom, molecule-by-molecule, or cluster-by-cluster in a gas phase or in solution. The latter approach is far more popular and less expensive in the synthesis of NPs and can be identified in two



Nanoparticles, Fig. 3 (**a**) Size-dependent changes in color of gold, from gold bulk to atomic size; (**b**) schematic drawing of plasmon oscillation for a gold NP, showing the displacement of the conduction electron charge cloud relative to the nuclei

sub-categories: (1) Vapor (Gas) Phase Methods (e.g., pyrolysis, inert gas condensation); (2) Liquid Phase Methods (e.g., solvothermal reaction, sol–gel, microemulsion method).

Top-Down: High-Energy Ball Milling

Within top-down technologies, one of the most representative techniques is High-Energy Ball Milling (HEBM). This technique is well known as a simple and inexpensive method for the production of nanostructured materials, in particular it allows the preparation of alloys and composites, which cannot be synthesized via conventional routes. Different types of milling devices have been developed for different purposes including tumbler mills, attrition mills, shaker mills, vibratory mills, planetary mills, etc. In HEBM, particles of the starting material are subject to heavy deformation, cold work hardening, and subsequent fragmentation [6].

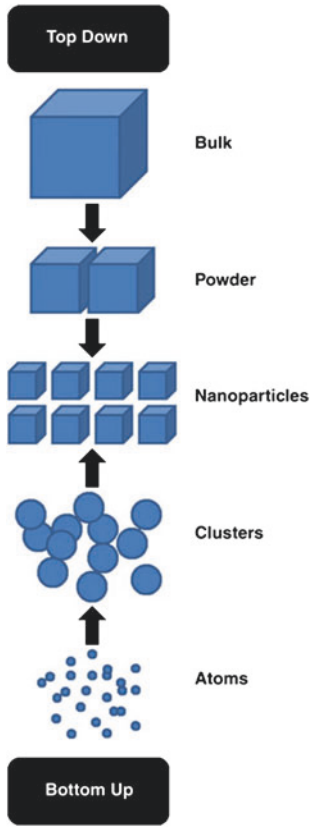
In a HEBM system, the grinding balls and the material in the grinding bowl are thus acted upon by centrifugal forces, which constantly change in direction and intensity. The grinding bowl and the supporting disc rotate in opposite directions, so that the centrifugal forces alternatively act in the same and opposite directions (Fig. 5). The process requires a long milling time to activate and complete the structural changes and chemical reactions, as well as a control of milling atmosphere and temperature which are crucial to create the desired structural changes or chemical reactions [7].

Bottom-Up: Vapor (Gas) Phase Methods

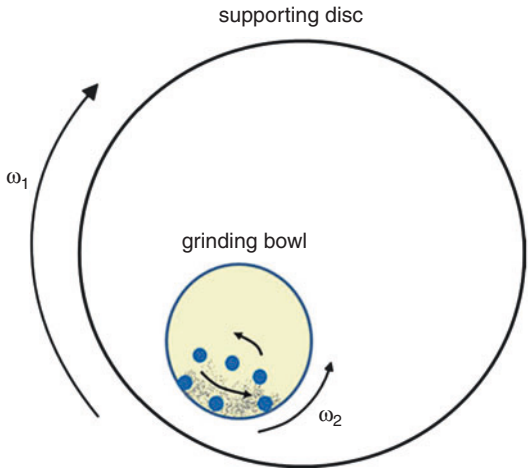
Vapor (Gas) phase synthesis is a well-known chemical manufacturing technique for an extensive variety of nano-sized particles. The transformation of a gaseous precursor to the final particulate is a complex physical process involving nucleation of the particulate phase, condensation of gasses onto particles, and coagulation and coalescence between particles.

Vapor phase fabrication processes are typically conducted in vacuum under elevated temperatures. The mechanism for particle formation starts with the precursor vaporization (typically involves a catalyst) followed by the nucleation and growth stage.

Synthesis of NPs via high-temperature processing is a promising technique that offers a good route to production of high-purity NPs with specifically tailored chemical and physical



Nanoparticles, Fig. 4 Schematic representation of the *bottom-up* and *top-down* synthesis processes of nanomaterials



Nanoparticles, Fig. 5 Schematic drawing of a high-energy ball mills

properties. However, NPs generated by gas phase processes are usually in the form of aggregates, due to their Brownian coagulation at high temperatures in the reactor: coagulation is an intrinsic mechanism which inevitably occurs at high particle concentrations and therefore in all industrial aerosol processes [8].

Bottom-Up: Spray Pyrolysis Methods

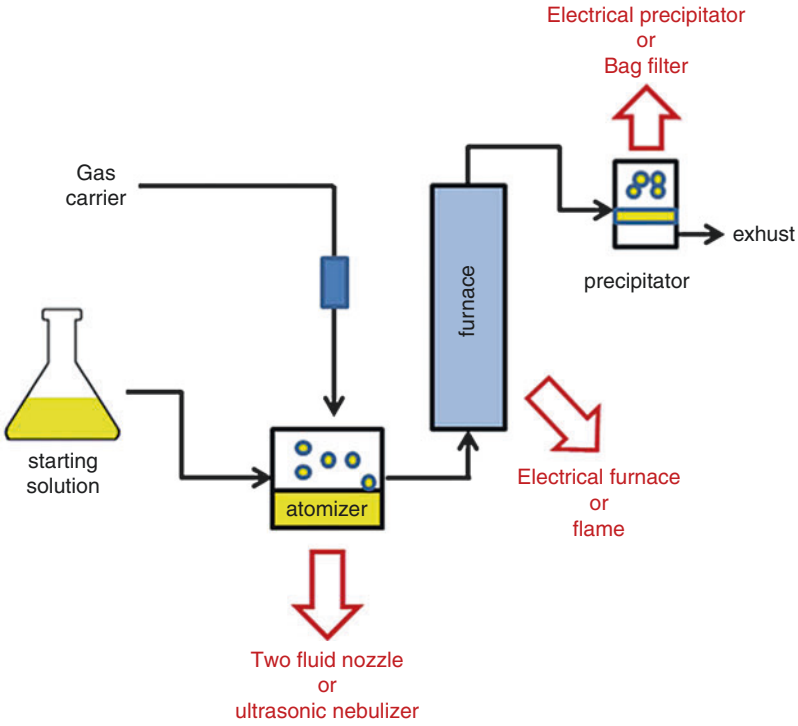
Spray methods, such as spray pyrolysis, spray drying, and the flame spray method, have the advantage of being continuous, low-cost, and single-step methods for the preparation of NPs and nanocomposites (which includes metals, metal oxides, non-oxide ceramics, superconducting materials, and nanophase materials) that are spherical, chemically homogeneous, and comprised of multiple materials. A representative spray system is shown schematically in Fig. 6.

In spray pyrolysis, the starting liquid solution is dispersed as droplets into a gas carrier and then sprayed into a drying chamber (furnace) that is heated above the solvent vaporization temperature. The remaining solutes undergo precipitation, thermal decomposition, and intraparticle reactions to form NPs. The solid particles are then collected by an electric field precipitator or a bag filter.

This technique is flexible as it allows the use of a wide range of precursors, solvents, and process conditions and thus provides control over particle size and composition [9].

Bottom-Up: Inert Gas Condensation

Inert Gas Condensation (IGC) process entails the evaporation of a source substance in an inert gas atmosphere to generate the nanopowder which is convectively transported and collected on a cold substrate. The NPs develop in a zone just above the evaporative source, due to interactions between the hot vapor species and the much colder inert gas atoms in the chamber. A feature of the process is the ability to generate non-agglomerated nanopowders, which are sinterable at relatively low temperatures. IGC process is generally classified as a Physical Vapor Deposition (PVD) method, but if a catalytic process is involved in addition, reference is made to it as *Chemical Vapor Deposition* (CVD) method.



Nanoparticles,
Fig. 6 Schematic drawing of a spray pyrolysis system

CVD is one of the most important synthesis processes for fabricating metal, semiconductor, and magnetic NPs. In the CVD process, the reaction or the thermal decomposition of gas phase takes place in a reaction chamber maintained under vacuum at high temperature (>1,173 K).

In common CVD reactors, source molecules react and produce both molecular and cluster intermediates in the gas phase. The intermediates become supersaturated and cause the nucleation in the gas phase due to the low vapor pressure. Nucleated NPs collide vigorously with each other and form agglomerated particles [10].

Growth and agglomeration of the particles are mitigated via rapid expansion of the two-phase gas stream at the outlet of the reaction chamber. Subsequent heat treatment of the synthesized nanopowders in various high-purity gas streams allows compositional and structural modifications, including particle purification and crystallization, as well as transformation to a desirable size, composition, and morphology. CVD is widely used to produce carbon nanotubes [11].

Bottom-Up: Liquid Phase Methods

Liquid phase methods include the set of processes that involve a wet chemistry route such as *Hydrothermal (solvothermal) synthesis, sol–gel synthesis*, and *microemulsion synthesis*. The particle formation mechanism is the same as in the vapor phase process. While in gas condensation it is not easy to control grain size and crystal shape, and furthermore its product yield is low, the wet chemistry route provides reasonable control over growth by utilizing organic ligands that act as growth confining agents.

Bottom-Up: Hydrothermal Synthesis

Hydrothermal synthesis of NPs provides precise control over particle size, shape distribution, and crystalline fraction.

This method takes advantage of the solubility of almost all inorganic substances in water at elevated temperatures and pressures and subsequent crystallization of the dissolved material from the fluid.

Water at high temperatures plays an essential role in the precursor material transformation

because the vapor pressure is much higher and the structure of water at elevated temperatures is different from that at room temperature. The properties of reactants, including their solubility and reactivity, also change at high temperatures. The solvent is not limited to water but includes also other polar or nonpolar solvents, such as benzene, and the process is more appropriately called *solvothermal synthesis* in different solvents [12].

Bottom-Up: Sol–Gel Synthesis

The sol–gel process is a recent technique widely used in the field of material science and ceramic engineering for the generation of colloidal NPs from liquid phase. Sol–gel method is well adapted for oxide NPs and composite nanopowder synthesis starting from a chemical solution (*sol*) which is followed by the formation of a *gel*. It is based on inorganic polymerization reaction including hydrolysis, polycondensation, drying and thermal decomposition.

Typical starting materials are inorganic precursors: metal or nonmetal alkoxides (MOR) dispersed in an appropriate solvent. Precursors hydrolyze with water or alcohols, according to the reaction scheme:

$$-\mathrm{MOR} + \mathrm{H_2O} \rightarrow -\mathrm{MOH} + \mathrm{ROH}\,(\mathrm{Hydrolysis})$$

$$-\mathrm{MOR} + \mathrm{MOR} \rightarrow -\mathrm{MOR} + \mathrm{MOR}$$

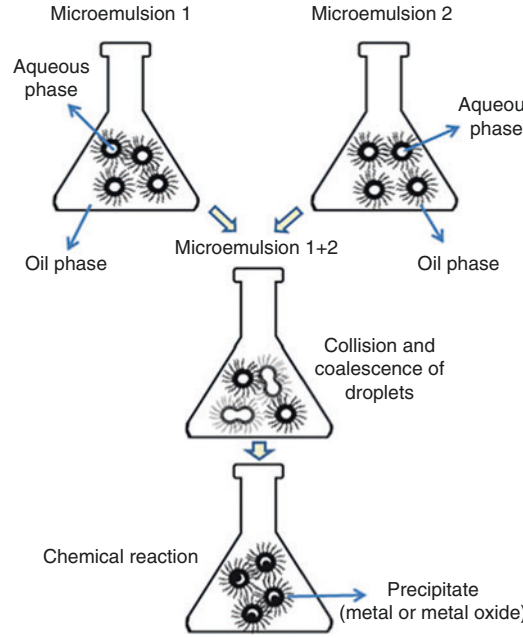
or

$$-\mathrm{MOR} + \mathrm{MOR} \rightarrow -\mathrm{MOM} + \mathrm{H_2O}(\mathrm{Condensation})$$

The rates of hydrolysis and condensation are important parameters that affect the properties of final products. Slower and more controlled hydrolysis typically leads to smaller particle size and more unique properties. After solution condensation to gel, the solvent should be removed. This requires a drying process, which is typically accompanied by a significant amount of shrinkage and densification. Generally, higher calcination temperatures are needed to decompose the organic precursors [11].

Bottom-Up: Microemulsion Method

A microemulsion is a thermodynamically stable dispersion of at least three components: two immiscible fluids and a surfactant with amphiphilic properties (stabilizer).



Nanoparticles, Fig. 7 Proposed mechanism for the formation of metal particles within micelle "nanochamber"

A surfactant is a molecule that possesses both polar and nonpolar moieties. In very diluted water (or oil) solutions, it dissolves and exists as monomer, but when its concentration exceeds a certain minimum critical micelle concentration (CMC), the stabilizer molecules associate spontaneously to form aggregates-micelles [13]. The formation of microdroplets can be in the form of oil-swollen micelles dispersed in the aqueous phase as the oil/water (O/W) microemulsion or water-swollen micelles dispersed in oil as the water/oil (W/O) microemulsion (reverse microemulsion) [12]. The formation of O/W or W/O micelles is driven by strong hydrophobic interactions of the hydrophobic tail of the surfactant molecule (O/W micelle) or by hydrophilic interactions of the polar head of the surfactant molecule (W/O micelle).

Micelles in these systems can be described as "nanoreactors" providing a suitable environment for controlled nucleation and growth. Figure 7 illustrates the mixing of two microemulsions containing the appropriate reactants (metal salt and reducing agent) during the collision of the

Nanoparticles, Table 1 Parameters versus principal characterization methods for Nps

Parameters	Characterization methods
Particle size, size distribution, morphology	Scanning Electron Microscopy (SEM), Transmission Electron Microscopy (TEM), Atomic Force Microscopy (AFM), X-ray Diffraction (XRD), Laser diffractometry
Charge determination	Zeta potentiometer
Surface hydrophobicity	Water contact angle measurements, hydrophobic interaction chromatography
Magnetic features	Vibrating Sample Magnetometry (VSM), Alternating gradient force magnetometry (AGFM), Magnetic Force Microscopy (MFM)
Electronic properties	Percolation threshold measure (both current–voltage DC and AC), dielectric characterization

water droplets. This intermicellar interchange process is very fast (from 10 μs to 1 ms) and it is the rate-limiting step for particle growth, but it is slow when compared to diffusion of reagents inside the nanoconfinement ("nanoreactor"). The attractive interaction between the droplets, which is responsible for the percolation process, is of great importance for the kinetics [13].

Characterization of Nanoparticles

NPs are generally characterized in particle size distribution, morphology, charge determination, and surface hydrophobicity. More specific characterizations depend on the properties of NPs synthesized such as magnetic characterizations for NPs featuring magnetic properties (e.g., Fe, Fe_3O_4) and electric characterizations for conductive or dielectric NPs (e.g., Ag, Cu, Au, SiO_2, $BaTiO_3$, etc.). Different parameters versus principal characterization methods used for NPs are given in Table 1.

Magnetic Nanoparticles

Magnetic NPs containing pure 3D transition metals – or some mixture of their oxides – are one of the most studied nanomaterials in view of the prospective, ubiquitous applications in quite different areas, the most notable being biomedicine, sensor technology, and magnetic recording [14]. According to their usage, magnetic NPs are often either embedded in a (typically diamagnetic) solid [15], or dispersed in a fluid; in some cases they are surrounded by an outer shell of a diamagnetic material.

This entry deals with particles having sizes not exceeding a few tens of nanometers. The lower limit is taken to be a few (>2) nanometers. As a result, the number of magnetic atoms per NP typically ranges between hundreds and a few millions. Smaller aggregates should be more properly termed magnetic clusters; their magnetic properties are closer to those of magnetic surfaces and are described in a quite different way. On an even smaller scale, magnetic molecules (mainly organic molecules containing a small number, about 10, of transition metal ions) are the subject of much study even if their applicability is still controversial.

Nowadays, magnetic NPs are routinely prepared by a variety of techniques which can be termed as physical (e.g., laser-assisted deposition, sputtering, etc.) or chemical (e.g., wet chemical routes) [14, 16].

The main feature of bona fide NPs made of a particular element or compound is that their magnetic properties are quite different from those of the corresponding bulk material. In a simplified picture, an ideal magnetic NP can be figured out as provided of a highly symmetrical body (e.g., spherical, ellipsoid, acicular, etc.) and characterized by a homogeneous magnetization provided by the ferro- (or ferri-) magnetic alignment of elementary magnetic moments, giving rise to a mesoscopic permanent magnetic moment associated to the particle. Of course this picture can be improved to account for the role played by the NP surface, whose chemical and magnetic properties can be quite different from those at the particle core. Strictly speaking, the locution "core-shell nanoparticle" [17] means a particle whose chemical or stoichiometric properties are substantially different when moving from the center to the surface (e.g., a metallic core surrounded by an

oxide shell). In a sense, however, all magnetic NPs fall in this category at least when their magnetic properties are considered, even when surface and core do not substantially differ in their chemistry.

The simplified approach is instrumental in understanding the reason why magnetic properties of NPs change with respect to bulk counterparts. Any macroscopic magnetic material is usually divided into magnetic domains in order to minimize the magnetostatic energy; the number of domains, as well as their shape, is dictated by the interplay and balance between magnetic energies, such as exchange, anisotropy, and domain-wall energy. When the size of a piece of magnetic material is ideally shrunk down to the sub-micrometer range without changing its aspect ratio, all the associated magnetic energies decrease accordingly; however, they decrease following different laws. Both magnetostatic energy and anisotropy energy are volume energies; however, domain wall energy being a surface energy, it has a slower reduction rate with reducing particle size. This makes the nucleation of magnetic domains unfavorable when the size of a particle becomes smaller. Below a critical size which strongly depends on the material's magnetic properties and morphology, the total energy at remanence is no longer decreased by magnetic domains, and the NP remains in the single-domain state. The magnetic moment points along one of the magnetically "easy" directions corresponding to the minima of the anisotropy energy. In order to have it switched away from the easy direction, one must apply a magnetic field; on removing it, the magnetic moment will be again aligned along any one of the easy directions. This is the so-called single-domain (magnetically) blocked state [18]. However, on further decreasing the particle size, the anisotropy energy keeps decreasing as the particle volume, and eventually becomes so small that it falls below k_BT, where k_B is the Boltzmann's constant and T the absolute temperature, that is, the typical thermal energy steadily exchanged at equilibrium by the NP with the thermal bath. As a consequence, the magnetization is no longer "blocked" along an easy direction, being continuously affected by thermal (random) torques of thermal origin, and the magnetic moment vector behaves in a sense as a free rotator which can be aligned by an external magnetic field, very much as paramagnetic substances do. However, the NP's magnetic moment is usually much larger than that of a single atom or ion, so that it can be termed as a "supermoment," and the particular magnetic behavior of such a NP is known as "superparamagnetism" [19]. The size below which a NP ceases to be in the magnetically blocked state at a given temperature to enter the superparamagnetic (SP) region, and the temperature above which a NP of given size begins to follow the SP behavior are termed critical size (temperature) for the onset of superparamagnetism, respectively. Both depend on NP size, shape, and the leading anisotropy term.

The transition from blocked to SP state is not a thermodynamic phase transition, because the SP behavior emerges only when the magnetic moment's direction randomly fluctuates at a rate high enough that during the time needed to perform a single measurement it virtually takes all directions in space, so that its time average is zero at zero applied field. The observation time plays a crucial role in determining whether a NP at a given temperature is in the SP state or not. From this viewpoint, the blocked-SP transition resembles, in a sense, that between an undercooled liquid and a glass.

SP particles exhibit a reversible magnetization versus field behavior, as paramagnetic substances do. Therefore, no hysteresis loop appears when an alternating magnetic field is applied to a system of genuine SP particles. The material's coercivity and remanence disappear. On the other hand, blocked single-domain NPs usually exhibit coercivity and remanence values which can be much higher than the ones measured in a macroscopic specimen of the same material. This means that a system containing blocked NPs made of a magnetically soft material acts as a magnetically hard material [18]. Therefore, two quite opposite magnetic behaviors are expected at the same temperature, strictly depending on NP size (and shape) alone.

Magnetic NPs in the SP phase are attractive for biomedical applications, because they can be

easily driven by a magnetic field (an oscillating field makes them their "free" moments to oscillate; an inhomogeneous field exerts a force on them, displacing them toward a target point). Oscillating magnetic moments dissipate energy as heat, and may effectively contribute in the local heating of organic tissue, for example, malignant tissue. Magnetic NPs may be properly functionalized with diamagnetic, organic shells in order to act as the portable inner core of a larger, multi-layered particle aimed to a specific intra-body target where it can release specific drugs carried by the organic shell (drug delivery). Other applications of functionalized particles include cell separation and DNA reconnaissance.

Blocked NPs are interesting because they can be exploited as low-cost substitutes of more expensive high-coercivity materials in permanent magnet industry and in magnetic data storage [16]. Hard-disk magnetic memories are often based upon nanoparticulate media where the information can be safely stored owing to the high coercivity of the medium. In this case, superparamagnetism is detrimental because it causes a loss of information. The SP limit introduces a lower boundary to the size of the NPs which a magnetic memory can be composed of, and poses a major, still partially unsolved technological problem in the development of higher-density magnetic memories.

A major drawback of almost all techniques proposed so far to produce magnetic NPs (at least, all of the low-cost ones) is the lack of full control on their size, which results in a (more or less) broad size distribution. Considering how size matters in determining the magnetic properties of these magnetic nanomaterials, it is apparent why many efforts of both fundamental and applied research are now aimed to develop preparation techniques resulting in monodisperse nanoparticle systems.

Application of Nanoparticles

Here it is proposed a selection of the application fields of NPs, in addition to those of magnetic NPs which were summarized in the previous paragraph. As references, the same literature previously cited was kept.

Drinking water purification is a new and promising application, performed making use of Fe, Ni, Au, Ag, or Cu NPs, for example for pesticide removal and organic contaminant detection. The NPs feature an enhanced reactivity and selectivity toward some dangerous elements (such as heavy metals) and organic substances (pesticides); when combined with a MEMS structure, this property is used to detect the chemicals by monitoring certain physical properties of the system (resonance frequency, impedance, etc.).

The enhanced chemical activity of metal NPs is also well known in catalysis, where transition metal NPs are used in the synthesis of organic compounds. Catalytic NPs may be surrounded by a dendritic polymeric structure (dendrimer) in order to allow catalyst recovery at the end of the chemical reaction. A peculiar form of catalysis, called water splitting, is performed using semiconductor NPs (Cu_2O, Fe_3O_4), where sunlight is directly converted in a current (either electrons or holes, depending on the semiconductor properties) and used to split the water molecule into oxygen and hydrogen, to produce an environmentally friendly fuel.

Cellular imaging, intended as optical microscopy performed on cells using fluorescence, Raman spectroscopy and SERS, is normally enhanced by uptake of NPs into the cell body. For this purpose are often used Au and, TiO_2 NPs, and so-called quantum-dots, where this definition includes semiconductor NPs, as CdSe and ZnS.

Photoluminescence of semiconductor NPs (quantum dots) is well known and studied in CdTe, Au/CdTe NPs assemblies, PbS, CdS, etc.

Biomedical applications of noble metal NPs include targeted drug delivery and targeted heat release; similar applications are more commonly approached by using ferromagnetic NPs (e.g., hyperthermia treatment on malignant cells). Also polymeric NPs, which are not a matter of discussion here, are currently used as vectors for drug delivery.

Inhibition of microorganisms and antimicrobial effects are obtained using Ag and/or TiO_2 NPs.

For solar cells optimization noble metal (Au, Ag) NPs have been proposed as electromagnetic

field enhancement media due to their SPR, even though TiO_2 and SiO_2 NPs possess more interesting antireflection properties.

Flexible electronics: Metallic NP tracks (Au, Ag, Cu, Sn), deposited by means of inkjet printing technology, are sintered toward electrical percolation at very low temperatures, allowing the realization of conductive paths on flexible polymeric substrates [20]. Currently, research is going on, considering possible solutions to assemble NPs and realize functional devices.

Besides the direct application of NPs, they are currently added to either metallic or ceramic or polymeric matrices to realize nanocomposite materials featuring enhanced mechanical, electronic, thermal, magnetic, and optical properties, which are not a matter of discussion in this entry.

Cross-References

- ▶ Chemical Vapor Deposition (CVD)
- ▶ Chitosan Nanoparticles
- ▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles
- ▶ Nanostructures for Energy
- ▶ Physical Vapor Deposition
- ▶ Simulation of Supported Metal Clusters
- ▶ Use of Nanotechnology in Pregnancy

References

1. ESRF newsletter n°38. December 2003 and references therein
2. Patel, J.K., Patel, D.J., Pandya, V.M.: An overview: nanoparticles. Int. J. Pharm. Sci. Nanotechnol. **1**(3), 215–220 (2008), and references therein
3. Hosokawa, M., Nogi, K., Naito, M., Yokoyama, T.: Nanoparticle Technology Handbook, pp. 5–48. Elsevier, Oxford (2007), and references therein
4. Ung, T., Liz-Marzánand, L.M., Mulvaney, P.: Redox catalysis using $Ag@SiO_2$ colloids. J. Phys. Chem. B **103**(32), 6770–6773 (1999), and references therein
5. Scientific Committee on Emerging and Newly Identified Health Risks (SCENIHR): The appropriateness of existing methodologies to assess the potential risks associated with engineered and adventitious products of nanotechnologies (2007) and references therein
6. Munõz, J.E., Cervantes, J., Esparza, R., Rosas, G.: Iron nanoparticles produced by high-energy ball milling. J. Nanopart. Res. **9**, 945–950 (2007), and references therein
7. Chen, Y., Pui Li, C., Chen, H., Chen, Y.: One-dimensional nanomaterials synthesized using high-energy ball milling and annealing process. Sci. Technol. Adv. Mater. **7**, 839–846 (2006), and references therein
8. Gutsch, A., Krämer, M., Michael, G., Mühlenweg, H., Pridöhl, M., Zimmermann, G.: Gas-phase production of nanoparticles. KONA **20**, 24–37 (2002), and references therein
9. Iskandar, F.: Nanoparticle processing for optical applications – a review. Adv. Powder Technol. **20**, 283–292 (2009), and references therein
10. Adachi, M., Tsukui, S., Okuyama, K.: Nanoparticle formation mechanism in CVD reactor with ionization of source vapor. J. Nanopart. Res. **5**, 31–37 (2003), and references therein
11. Kumar, M., Ando, Y.: Chemical vapor deposition of carbon nanotubes: a review on growth mechanism and mass production. J. Nanosci. Nanotechnol. **10**(6), 3739–3758 (2010), and references therein
12. Tavakoli, A., Sohrabi, M., Kargari, A.: A review of methods for synthesis of nanostructured metals with emphasis on iron compounds. Chem. Pap. **61**(3), 151–170 (2007), and references therein
13. Capek, I.: Preparation of metal nanoparticles in water-in-oil (w/o) microemulsions. Adv. Colloid Interface Sci. **110**, 49 (2004), and references therein
14. Reiss, G., Hütten, A.: Magnetic nanoparticles: applications beyond data storage. Nat. Mater. **4**, 725–726 (2005)
15. Allia, P., Tiberto, P., Coisson, M., Chiolerio, A., Celegato, F., Vinai, F., Sangermano, M., Suber, L., Marchegiani, G.: Evidence for magnetic interactions among magnetite nanoparticles dispersed in photoreticulated PEGDA-600 matrix. J. Nanopart. Res. (2011). doi:10.1007/s11051-011-0249-7, in press
16. Gubin, S.P., Koksharov, Y.A., Khomutov, G.B., Yurkov, G.Y.: Magnetic nanoparticles: preparation, structure and properties. Russ. Chem. Rev. **74**(6), 489–520 (2005)
17. Iglesias, Ò., Labarta, A., Batlle, X.: Exchange bias phenomenology and models of core/shell nanoparticles. J. Nanosci. Nanotechnol. **8**, 2761–2780 (2008)
18. Tannous, C., Gieraltowski, J.: The Stoner–Wohlfarth model of ferromagnetism. Eur. J. Phys. **29**, 475–487 (2008)
19. Knobel, M., Nunes, W.C., Socolovsky, L.M., De Biasi, E., Vargas, J.M., Denardin, J.C.: Superparamagnetism and other magnetic features in granular materials: a review on ideal and real systems. J. Nanosci. Nanotechnol. **8**, 2836–2857 (2008)
20. Chiolerio, A., Maccioni, G., Martino, P., Cotto, M., Pandolfi, P., Rivolo, P., Ferrero, S., Scaltrito, L.: Inkjet printing and low power laser annealing of silver nanoparticle traces for the realization of low resistivity lines for flexible electronics. Microelectron. Microeng. **88**, 2481–2483 (2011), and references therein

Nanoparticles Toxicity Examined In Vivo in Animals

► In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles

Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods

Caihong Liu[1,2] and Leon Shaw[1,2]
[1]Wanger Institute for Sustainable Energy Research, Illinois Institute of Technology, Chicago, IL, USA
[2]Department of Mechanical, Materials and Aerospace Engineering, Illinois Institute of Technology, Chicago, IL, USA

Synonyms

Core/shell nanostructures; Synthesis of nanoparticles; Wet chemical synthesis

Definition

Nanoparticulate materials are defined as the materials with particle sizes smaller than 100 nm and exhibit different characteristics compared to their bulk counterpart.

Core/shell structures are a class of engineered nanostructures which have core and shell layers with different properties, compositions, crystal structures, etc.

Wet chemistry methods is a term used to refer to chemical approaches generally done in the liquid phase.

Introduction

Nanoparticulate materials with particle sizes smaller than 100 nm have attracted enormous attention over the past three decades because of their unusual properties that cannot be derived from conventional micrometer-sized particles. Due to their unique properties, nanoparticles have found wide applications in energy, electronics, photonics, magnetics, biomedical engineering, catalysis, transportation, health care, communication, and medicine. This entry focuses on synthesis of nanoparticles. Generally speaking, synthesis methods of nanoparticles can be classified into two categories, one being wet chemistry methods and the other non-wet chemistry methods. The latter could include inert gas condensation, chemical vapor synthesis, laser ablation, high-energy ball milling, laser-assisted synthesis, and plasma-assisted synthesis among others. Wet chemistry methods may include sol–gel process, precipitation, hydrothermal synthesis, reduction, solvothermal synthesis, cryochemical synthesis, and spray pyrolysis. This entry will describe the key principles and major parameters of the first four wet chemistry methods, while the last three methods, requiring large deviation from the ambient temperature during synthesis, either significantly below room temperature (i.e., cryochemical synthesis) or significantly above room temperature (e.g., solvothermal synthesis and spray pyrolysis), will not be described in this entry. Nanoparticles with core/shell structures will also be addressed in this entry because they often offer special functionalities that cannot be obtained even from nanoparticles. The methods for synthesis of core/shell structures are typically derived from a combination of two different wet chemistry methods or from one wet chemistry method with two steps of different chemistries. Therefore, the basic principles of synthesis science for core/shell structures are in many cases the same as individual wet chemistry methods except the additional requirement of interface control without which core/shell structures are difficult to synthesize.

Sol–Gel Process

Key Principles of Sol–Gel Process

This process involves the hydrolysis and condensation of molecular precursors to form a sol which

can then precipitate out as nanoparticles or change to a gel, a semirigid network, by evaporation of the solvent or by continuous inorganic polymerization. Whether nanoparticles or gels are formed depends on the nature of the precursor(s) and the synthesis conditions. The precursors for the sol–gel process can be inorganic or organic salts (e.g., nitrates, chlorides, acetates, etc.) as well as alkoxide precursors [1, 2]. Metal alkoxides, $M(OR)_n$ where M is a metallic element (such as Si, Al, Fe, Cr, Mn, Ce, Y, W, etc.) and OR is an alkoxide ligand with R being an alkyl group, are most widely used precursors. The sol–gel reaction using metal alkoxides can be considered as a two-step inorganic polymerization [1–5]. The first step is the hydrolysis of alkoxide ligands to yield hydroxylated metal centers, M-OH:

$$M(OR)_n + H_2O = M(OR)_{n-1}(OH) + ROH \quad (1)$$

The second step is the condensation of the hydroxylated species to form oxypolymers, involving an oxylation reaction which creates oxygen bridges and expels XOH species as follows:

$$M - OH + M - OX = M - O - M + XOH(X = H \text{ or } R) \quad (2)$$

Transition metal alkoxides has a strong electrophilic character and their ability to increase the coordination number is high [1–3]. As a result, their hydrolysis and condensation rates are very fast, which leads to rapid growth of oxypolymers and thus precipitation of nanoparticles. To avoid fast precipitation and to form a gel, metal alkoxides need to be modified at the molecular level by chelating ligands such as acetylacetone which is nucleophilic. The chelated complex has a reduced electrophilic character of the transition metal and lowers hydrolysis and condensation rates substantially, and thus gels rather than nanoparticles can be formed [2].

In contrast, nontransition metal alkoxides such as tetraethyl orthosilicate ($Si(OC_2H_5)_4$, TEOS) have a weak electrophilic character because transition metal elements are more electropositive than silicon [1]. As a result, their hydrolysis and condensation rates are very slow, about five orders of magnitude slower than that of transition metal alkoxides [1, 3]. Therefore, the products of TEOS can be adjusted to be nanoparticles or gels relatively easier by controlling the pH value and hydrolysis ratio during sol–gel reactions [5].

There are multiple major parameters during the sol–gel process that can alter the reactions and final products. These major parameters and their effects are discussed below.

Major Parameters of Sol–Gel Process

Chemical Compositions of the Precursors

As pointed out above, the nature of the precursor can significantly affects the sol–gel reaction kinetics and final products. For example, cerium alkoxides, acetylacetone-modified cerium alkoxides, and cerium acetylacetonates can all be used as sol–gel precursors to synthesize CeO_2. However, cerium alkoxide precursors often lead to the formation of nanoparticles due to the strong electrophilic character of their cerium metal and thus rapid hydrolysis and condensation rates [3]. In contrast, acetylacetone-modified cerium alkoxides exhibit preferential hydrolysis of less electronegative alkoxide ligands, while strongly bonded complexing groups are difficult to be hydrolyzed [1]. Similarly, cerium acetylacetonates with the absence of alkoxide ligands have been found to be hydrolyzed very slowly [4]. Specifically, the detailed FTIR analysis reveals that the hydrolysis rate of cerium acetylacetonates in methanol is relatively fast at the early stage of hydrolysis (<6 h at room temperature), but becomes very slow beyond 6-h of hydrolysis, and hydrolysis cannot be completed even after 7 days of hydrolysis at room temperature [4]. The slow hydrolysis and condensation rates of both cerium acetylacetonates and acetylacetone-modified cerium alkoxides facilitate the formation of gels [1–3].

Hydrolysis Ratio

The hydrolysis ratio, as defined by Eq. 3, is the main external parameter and has been shown to be

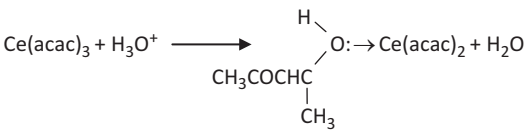
important in controlling the hydrolysis and condensation of metal alkoxides [3]:

$$h = \frac{[H_2O]}{[M(OR)_z]} \quad (3)$$

where $[H_2O]$ and $[M(OR)_z]$ are the molar numbers for the water and metal alkoxide, respectively. When $h < 1$, water molecules are not sufficient for complete hydrolysis, and thus gelation or precipitation cannot occur. When $1 < h < z$ (where z is the valance charge of M metal as shown in Eq. 3), chain oxypolymers can form. When $h > z$, excess water is present and cross-linked oxypolymers can be obtained, leading to the formation of nanoparticles or gels. The hydrolysis ratio also has significant impact on the hydrolysis and condensation rates [5]. When h is small ($< \sim 1$), the hydrolysis rate is low, and thus the sol–gel reaction, which includes both hydrolysis and condensation, is slow. When h is high ($> \sim 7$), the hydrolysis rate is high, but the condensation ratio is low. As a result, the sol–gel reaction is slow. When h is intermediate ($1 < h < 7$), the sol–gel reaction is the fastest because both hydrolysis and condensation rates are high.

Catalysts

Both acids and bases can catalyze sol–gel reactions. For example, acids and bases have been used to catalyze the sol–gel formation of SiO_2 using TEOS as the precursor [5]. Acids can catalyze hydrolysis of metal alkoxides and acetylacetonates because negatively charged alkoxide groups and acetylacetonate (acac) anions can be protonated by H_3O^+ ions, as shown below [2].



(4)

Under such conditions, the prototropic transfer and the departure of the leaving group from the transition state can no longer be the rate-limiting step, and thus hydrolysis rates are increased substantially. As a result, nanoparticles can be formed from acetylacetonate precursors in spite of the nucleophilic nature of the acetylacetonate anions [2]. However, not all acids can catalyze hydrolysis. For example, carboxylic acids can serve as a complexing agent to react with metal alkoxides via a nucleophilic substitution to slow down hydrolysis and condensation of metal alkoxides [1, 3].

Bases catalyze hydrolysis of transition metal alkoxides by nucleophilic addition of OH^- ions, which decreases the positive charge of the transition metal atom [3]. Condensation is accelerated through the formation of highly nucleophilic species such as $M\text{-}O^-$, as shown below [3]:

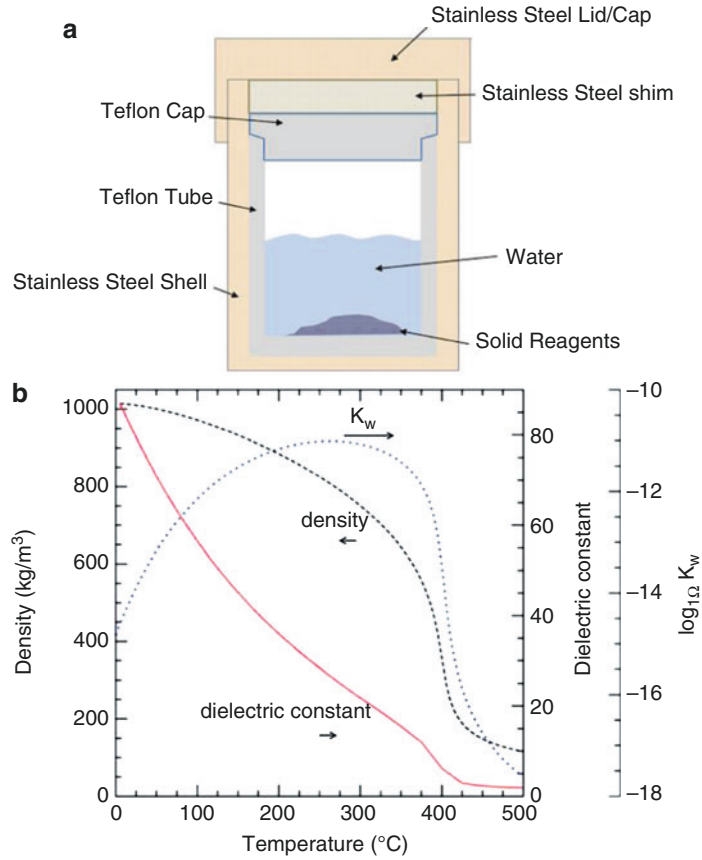
$$M - OH + OH^- = M - O^- + H_2O \quad (5)$$

The reactive condensation intermediate shown above will attack the more positively charged metal atom to complete the condensation reaction and form oxypolymers.

Combination of the pH Value and Hydrolysis Ratio

The pH value of the sol–gel solution will change when acids or bases are added to catalyze the reactions. It has been shown that the pH value and hydrolysis ratio have a combined effect on the morphologies of the final products, particularly for the SiO_2 system [5]. Under base-catalyzed conditions ($pH > 9$) and high hydrolysis ratios (h = 7–25), the hydrolysis rate of silicon alkoxides is high and condensation is rate limiting. As a result, smooth, uniform nanoparticles are formed. Under these conditions, depolymerization of clusters to form monomers is facile. This leads to particle coarsening if the reaction time is long, a phenomenon called Ostwald ripening in which many smaller particles are depolymerized into monomers which then deposit on larger particles via repolymerization.

Under acid-catalyzed conditions ($pH < 3$) and high hydrolysis ratios, the hydrolysis rate of silicon alkoxides is high, but condensation and depolymerization rates are low. With these conditions, monomers are consumed by condensation but not produced by depolymerization, leaving only



Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Fig. 1 (**a**) Schematic of a Teflon-lined stainless steel autoclave typically used for hydrothermal synthesis. (**b**) Density, dielectric constant and ionic product (K_w) of water at 30 MPa as a function of temperature (Reproduced from Einarsrud and Grande [6])

clusters which are unable to interpenetrate and condense to fill the entire space. As a result, randomly branched oxypolymer structures are formed [5].

Under low pH and low hydrolysis ratios, linear oxypolymer chains are formed. This is the condition used to produce gels for SiO_2 fiber production. However, the mechanism of forming linear oxypolymer chains under the low pH and low hydrolysis ratio condition is not fully understood yet [5].

Reaction Temperature

Increasing temperature typically results in increased hydrolysis and condensation rates. Thus, poorly reactive precursors, such as Si $(OR)_4$, often require relatively high temperature (40–80 °C) to activate the sol–gel reaction. In contrast, for highly reactive precursors, such as transition metal alkoxides, the reaction temperature must be lowered in order to slow down hydrolysis and condensation [3].

Hydrothermal Synthesis

Key Principles of Hydrothermal Synthesis

The hydrothermal synthetic process generally involves heterogeneous chemical reaction in aqueous media above room temperature ($T > 25$ °C, normally above 100 °C) and at a pressure greater than 1 atm. The chemical solutions are heated in a sealed vessel (autoclave, Fig. 1a); thereby the autogenous pressure increases inside the vessel beyond the atmospheric pressure, especially when temperature exceeds the boiling point of the solvent. This

will provide increased solubility and reactivity of the precursors used. The process is called hydrothermal when water is used as the solvent, while the term solvothermal is used in the case of nonaqueous solvent. During the hydrothermal synthesis, the density, ionic product, viscosity, and dielectric constant of water change with temperature and pressure.

The most accepted mechanism of hydrothermal synthetic process is dissolution–reaction–reprecipitation, which includes dissolution of precursors, diffusion of dissolved precursor species, and reaction/precipitation to give the target compound. Once a nucleus forms, the particle grows depending on the crystallite- and morphology-control experimental conditions. The dissolution and diffusivity of precursors are increased by the high pressure and reduced viscosity of the water under hydrothermal conditions.

Hydrothermal method has been so generally used to synthesize almost every kind of nanostructures, such as ceramics, metals, and semiconductors. All forms of nanostructures can be prepared with hydrothermal synthesis, namely, nanoparticles, nanowires, nanorods, core–shell nanostructures, porous nanostructures, etc. Hydrothermal synthesis offers many advantages over conventional and nonconventional synthetic methods, especially for ceramics. This process is simple, low cost, easy to scale up (continuous fabrication, etc.), and environmentally benign. The ability to precipitate already crystallized powders directly from solution regulates the rate and uniformity of nucleation, growth, and aging, which results in improved control of size distribution and morphology of crystallites and significantly reduced aggregation levels. Another important advantage of the hydrothermal synthesis is that the purity of hydrothermally synthesized nanostructures can exceed the purity of the starting materials. It is because the hydrothermal crystallization is a self-purifying process, during which the growing crystals/crystallites tend to reject impurities present in the growth environment. In other synthesis routes such as high temperature calcination, the self-purifying process is less likely to occur. The low reaction temperatures also avoid other problems encountered with high temperature processes, for example, poor stoichiometry control due to volatilization of components (e.g., Pb and Cd volatilization in Pb-/Cd-based ceramics).

Major Parameters of Hydrothermal Synthesis

Temperature and Pressure

The autoclave (Fig. 1a) is heated either by placing it in a water bath, oil bath, or oven or by using microwaves. During hydrothermal synthesis, the density, ionic product, viscosity, and dielectric constant of water change with temperature and pressure [6], as illustrated in Fig. 1b. At and above the critical point (374 °C, 218 atm), the densities of water gas and liquid phases are equal and the difference between liquid and gas disappears, becoming supercritical fluid. Most hydrothermal syntheses are conducted below the critical point, which is under mild hydrothermal conditions. The filling factor (or the percentage of free volume) of the autoclave and temperature determine the prevailing pressure in the vessel. On using a filling factor lower than 32 %, the liquid water–gas interface inside the autoclave will be lowered, and at a certain temperature, the autoclave will be completely filled with only the gas phase [6]. Typically, precursor aqueous solution or suspension fills the autoclave up to 70–80 % of its capacity. It should be noted that the precursors used will influence the composition of both the liquid and gas phase, hence change the pressure–temperature profile of the liquid–gas equilibrium in the autoclave during the hydrothermal synthesis. In addition, the pressure will be affected if gas product/side-product is generated during the process at the given temperature, e.g., nitrates and carbonates decomposing to nitrous gases or CO_2.

In general, an increase in temperature and/or decrease in pressure thermodynamically favors the less hydrated solid oxide phases. Temperature, pressure, and pH which are most important thermodynamic parameters determine which substance can be produced.

Precursors, Mineralizer, pH, and Pourbaix Diagrams

Precursors can be oxides, hydroxides, or salts of the elements to build up the material. The reaction formula can be described as follows:

$$MA + OH^-/\text{additives} \rightarrow M_xO_y \quad (6)$$

or

$$MA + M'B + OH^-/\text{additives} \rightarrow M_xM'_y\ O_z \text{ or } M_xM'_y \quad (7)$$

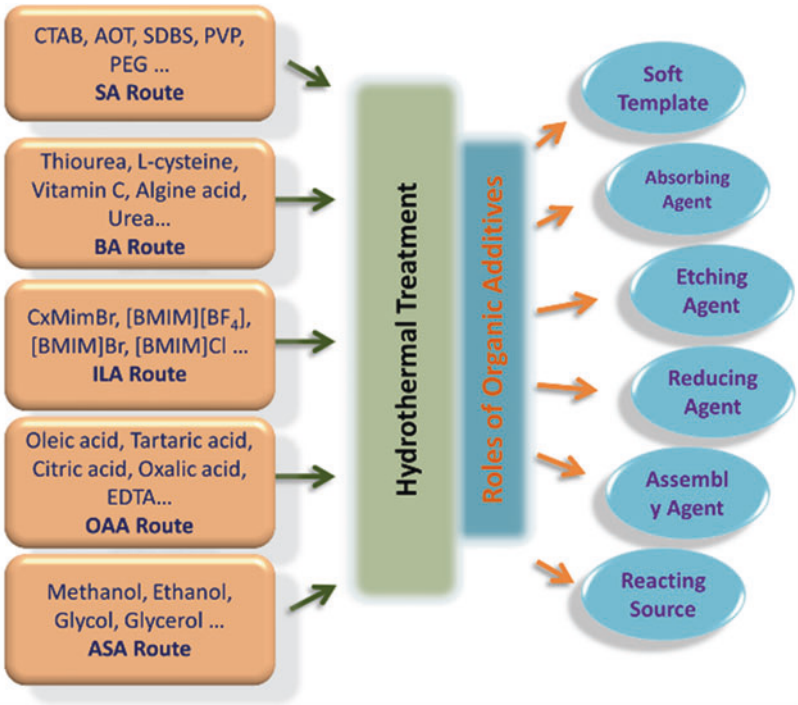
Small changes in the chemical parameters such as the type of precursors, concentration, and pH can dramatically affect the phase composition and morphology of the product. For achieving metal oxides or complex oxides, NaOH, KOH, or some other basic reagents are used as **mineralizer** to further increase the solubility of precursors by adjusting the pH value of the solution system. At progressively higher pH, the solubility of some metal oxide precursors is governed by stepwise hydrolysis species, such as GaOOH, as shown below.

$GaOOH\ (s) + 3H^+ + 4H_2O \rightarrow Ga(H_2O)_6^{3+}\ (aq)$	Acidic	(8)
$GaOOH\ (s) + 2H^+ + 4H_2O \rightarrow Ga(OH)(H_2O)_5^{2+}\ (aq)$		(9)
$GaOOH\ (s) + H^+ + 4H_2O \rightarrow Ga(OH)_2(H_2O)_4^{+}\ (aq)$		(10)
$GaOOH\ (s) + 4H_2O \rightarrow Ga(OH)_3(H_2O)_3\ (aq)$		(11)
$GaOOH\ (s) + 2H_2O \rightarrow Ga(OH)_4^-\ (aq) + H^+$	Very basic	(12)

In addition, different salts and organic additives like complexing agents are commonly used as mineralizers, especially for larger metal cations to enhance solubility. For instance, EDTA and citric acid were used for the formation of Pb $[Zr_xTi_{1-x}]O_3$ ($0 \leq x \leq 1$) to overcome the problem of differing solubilities between different cations. The organic additive-assisted hydrothermal synthetic strategies could realize further control of the morphology, uniformity, and dispersibility of inorganic nanostructures via the introduction of various organic additives compared with the organic additive-free hydrothermal routes. According to the properties of organic additives, studies based on these strategies may be mainly classified into five kinds of basic routes: surfactant-assisted (SA) route, biomolecule-assisted (BA) route, ionic liquid-assisted (ILA) route, organic acid-assisted (OAA) route, and alcoholic solvent-assisted (ASA) route [7]. Figure 2 shows the basic states of these five routes and the effects of organic additives in them. These organic additive or organic mineralizers can function as soft template, assembly agent, etching agent, reducing agent, reacting source, or absorbing agent. The function of these organic additives will be further mentioned in the hereinafter morphology control and template-assisted hydrothermal synthesis.

These mineralizers not only facilitate the solubility and hydrolysis of precursors but also alter/adjust the pH value of the solution system which is critical factor in most case for the final product phase. The solution pH determines different hydrolysis species of a metal cation and their abundance. Reactions in which elements possess various oxidation states are also dependent on the pH value, as well as oxidation potential of the solution. The Pourbaix diagram (also known as potential/pH diagram, EH/pH diagram, or pE/pH diagram) maps out possible stable phases of metal species, or the reaction equilibria and dominant metal species in relation to the redox potential and pH of an aqueous system. It summarizes thermodynamic possibilities, redox properties, and chemical speciation for reactions in which metals capable of existing in different oxidation states participate. Therefore, it is very important to hydrothermal synthesis.

However, most Pourbaix diagrams and formed species are constructed from experimental



Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Fig. 2 Five basic routes of the organic additive-assisted hydrothermal synthetic strategies, namely, SA route, BA route, ILA route, OAA route, and ASA route (Reproduced from Shi et al. [7])

information and collected under ambient conditions. Higher temperatures and pressures toward preparative hydrothermal chemistry are necessary. The constructions of elevated temperature Pourbaix diagrams, which are incomplete for many metal ion species, require experimental data (e.g., Gibbs free energy, standard molar entropy, and specific heat capacity) under hydrothermal conditions. To overcome this, numerous high-temperature metal ion-H_2O systems have been extrapolated from room temperature data although limited by the uncertainty of extrapolated data above 150 °C [6]. Calculated Pourbaix diagrams can aid experimental formulations by illustrating the trends in species forming as the temperature increases.

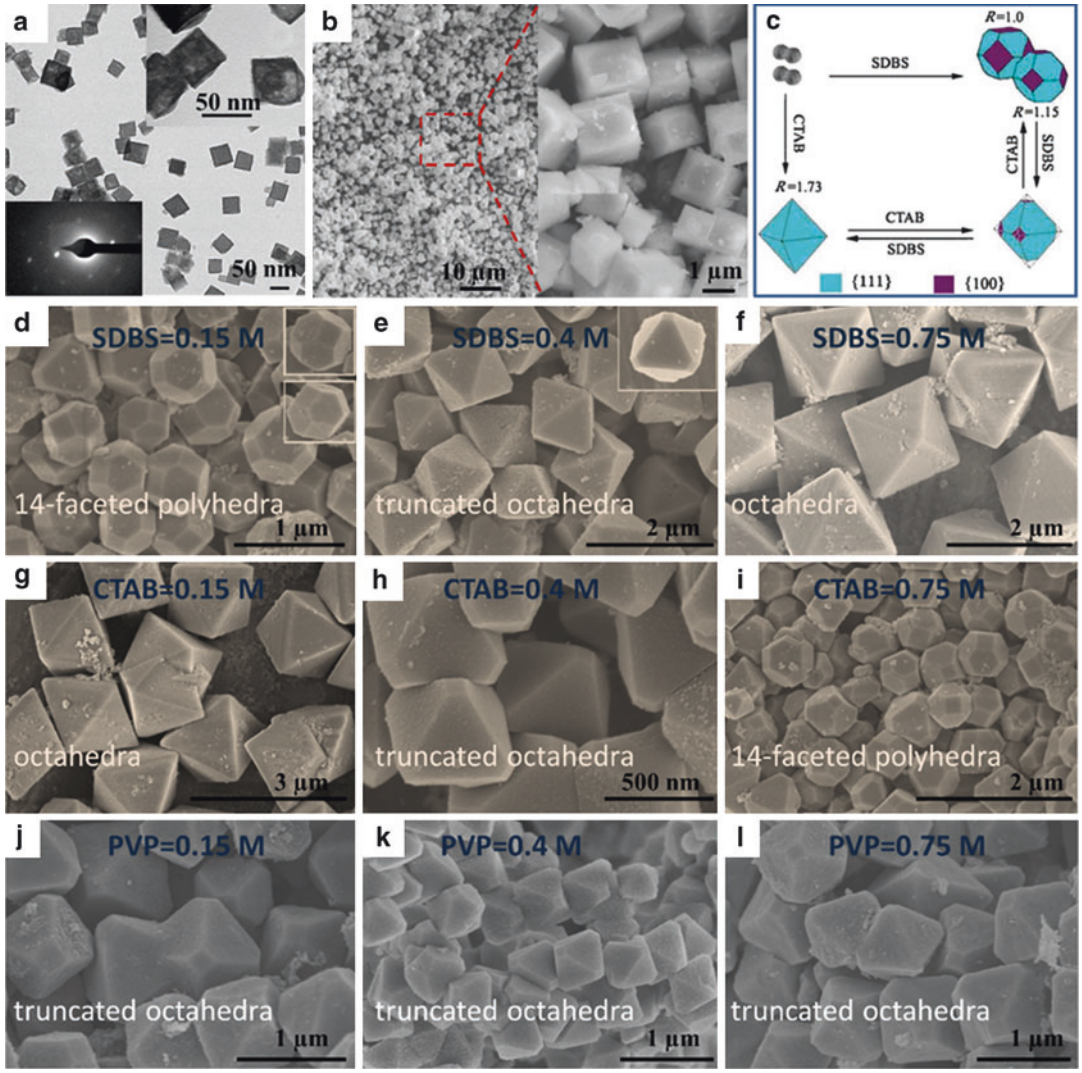
Morphology and Polymorph Control

Application-Driven Nanostructure Control

Crystalline products can be in a variety of forms, such as equiaxed (e.g., cubes, spherical), elongated (fibers, whiskers, nanorods, nanotubes), plates, nanoribbons, nanobelts, etc. Core–shell particles (semiconductor/semiconductor, semiconductor/metal, metal/metal, ceramic/semiconductor, ceramic/conductor, etc.) and composite nanostructures consisting of a mix of at least two different materials (SnO_2/CdS, Li_2CoSiO_4/C, TiO_2/SiO_2, Au/C) can be also prepared. Generally, the size and morphology can be engineered over a wide range and even adopt nonequilibrium morphologies. The size and morphology control are the key factors to many applications, such as nanostructures for catalysis, battery active material, optoelectronics, etc. This topic is one of the most exciting subfields in both nanoscience and catalysis and aims to allow the control of chemical reactions through changes in the size, shape, and interface of solid catalysts at the nanoscale.

Size Control

The particle size can be controlled by reaction time and temperature, precursor concentrations, and mineralizer species. If the outcome nanoparticles grow according to Ostwald-ripening process, the particle size will be



Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Fig. 3 Microscopic structures of hydrothermally synthesized perovskite $ZnSnO_3 \cdot 3H_2O$ nanoparticles under surfactant-free condition with different temperatures: (**a**) room temperature (RT, ~25°C), (**b**) 60°C and/or anionic (SDBS), cationic (CTAB), and nonionic surfactants (PVP) with increased surfactant concentration. (**c**) schematically illustrated the transformation in crystal shapes as a function of the ratio between the growths rates along the <100> and <111> directions (R) (Figures (**a**–**b**) adopted from Liu et al. [8] and Fig. (**c**–**l**) reproduced from Geng et al. [9])

uniformed and increased with the increase of reaction time or temperature. Typical examples can be found in synthesizing many kinds of metal oxide or metal oxide complex nanostructures. Figure 3a, b showed $ZnSnO_3 \cdot 3H_2O$ (perovskite, ABO_3) nano-/micro-cubes fabricated under different temperatures and the particle size increased with temperature [8]. When the surfactants, such as sodium dodecylbenzenesulfonate (SDBS), cetyl trimethylammonium bromide (CTAB), and polyvinylpyrrolidone (PVP), are added in the solution, the particle size and morphology of perovskite $ZnSnO_3 \cdot 3H_2O$ can be further finely tuned (Fig. 3c–l) [9].

Morphology and Growth Direction Control (Additives and Anions)

As previously mentioned, controlling the shape of oxide particles on the nanometer scale enriches their active sites (i.e., their electronic structure and surface coordination environment) through the selective exposure of reactive facets on their surface, which improves the catalytic performance. In principle, the shape of an oxide nanoparticle is governed through thermodynamics and kinetics. Thermodynamic equilibrium states minimize particle surface energy and preferentially expose the less active or inactive crystal facets. Morphology control to expose a large fraction of reactive facets generally can only occur via kinetic mediation of the growth of different crystal facets. The shape of the particles is determined primarily by the intrinsic energy balance of the crystallographic features and growth regime between the thermodynamic and kinetic processes. The systematic manipulation of the synthesis parameters, such as the precursor composition, pH, temperature, and surfactant, can feasibly adjust the relative growth rates of the different crystal facets and thus control particle morphology. Mostly used strategies are involved with the use of surfactants such as CTAB and SDBS or adjusting anion species in metal salt as structure-directing agents to induce the anisotropic nanocrystal growth by selectively capping specific facets.

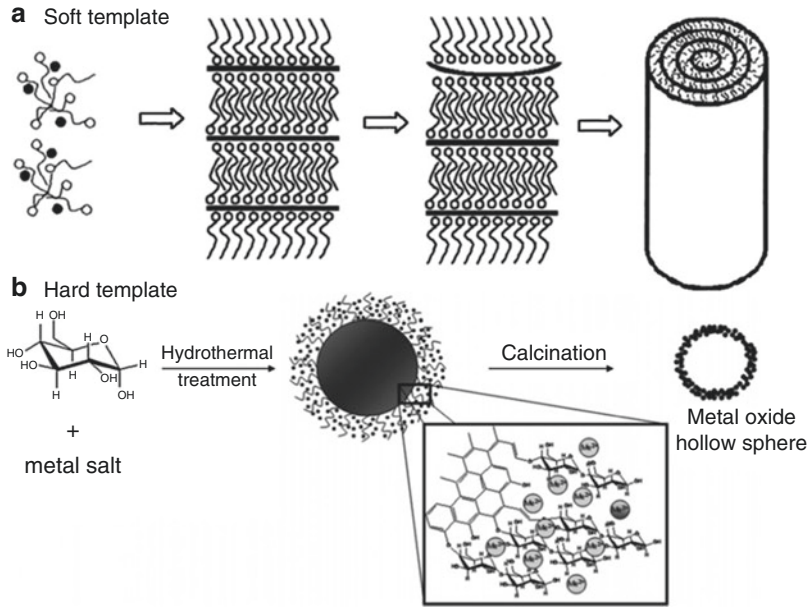
One example has been shown in Fig. 3. In the hydrothermal synthesis of perovskite $ZnSnO_3 \cdot 3H_2O$ nanostructures, anionic (SDBS), cationic (CTAB), and nonionic surfactants (PVP) with increased surfactant concentration were used. The effects of these surfactants on the specific facet growth have been summarized in Fig. 3c. It is found the CTAB facilitates the growth of {111} facets of $ZnSnO_3$, while SDBS favors the formation of {100} facets. No effect of PVP on facet selectivity was observed [9].

Except surfactants and other organic additives, the type of anion species in the metal salt affected the shape of the nanoparticles. For instance, the hydrothermal treatment of $Ce(NO_3)_3 \cdot 6H_2O$ in ammonia at 180 °C yielded spherical nanoparticles with 12 nm in size. The use of $CeCl_3 \cdot 7H_2O$ as the cerium source resulted in ceria nanowires of ~40 nm in diameter and 3–10 mm in length; replacement of the salt anions with HPO_4^{2-} obtained CeO_2 nanorods having diameters of 5–10 nm and lengths of 7.5–50 nm [10]. The chemical role of these anions was ascribed to their selective interaction with the specific CeO_2 facets. Chloride ions preferentially adsorbed onto the (111) facets and led to the growth of CeO_2 nanowires with polygonal cross-sections. This selectivity was further confirmed by the morphology evolution from nanoparticles to nanowires via the hydrothermal treatment of a mixture of $CeCl_3 \cdot 7H_2O$ and $Ce(NO_3)_3 \cdot 6H_2O$. An increase in the Cl^-/NO_3^- molar ratio from 1/3 to 3/1 changed the products from spheres to rods [10].

Hydrothermal Synthesis with Templates

Soft Template

Surfactants, biomolecules, and organic acids have been systematically studied as "soft templates" in the hydrothermal synthesis of hollow inorganic nanostructures. Numerous examples for each of the above can be found in literatures. Figure 4a gives a typical surfactant-templated hydrothermal synthesis. The cationic surfactant CTAB was used to aggregate VO_3^- nucleus to form lamellar structures. Under hydrothermal conditions, these lamellar sheets became loose at the edges and then roll into themselves to form the VO_x nanotubes [7]. Biomolecules used include macromolecules, e.g., proteins, polysaccharides, lipids, and nucleic acids, and small molecules, e.g., vitamins, glycolipids, glycerolipids, phospholipids, hormones, and so on. Due to their special structures, these biomolecules tend to self-assemble into certain patterns which can be used as template for the formation of inorganic nanostructures. The organic acids such as ethylenediaminetetraacetic acid (EDTA), oleic acid, tartaric acid, citric acid, oxalic acid, salicylic acid, and malic acid have been found to play four important roles in the growth process of nanostructures, namely, as a structure-directing agent, assembly agent, reducing agent, and gas bubble release agent. Due to strong chelating and capping capabilities, organic acids display well balancing the growth rate of



Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Fig. 4 Schematic illustration of soft (**a**) and hard (**b**) template in hydrothermal synthesis. (**a**) mechanism for the formation of vanadium oxide nanotubes: the mixture of NH_4VO_3 and the template molecules; layered structures formed through the hydrothermal treatment; the beginning stage of the rolling process; and the formed nanotubes. (**b**) Synthesis of metal oxide hollow spheres from hydrothermally treated carbohydrate–metal salt mixtures with carbon nanosphere template (Figures reproduced from Shi et al. [7])

different facets. Interestingly, same organic acid in a different reaction environment often exhibits different or even opposite structure-directing effects.

Hard Template

Polystyrene (PS), silica (SiO_2), and carbon nano-/microspheres are widely used in hydrothermal synthesis as "hard templates" for hollow or porous nanostructures. After hydrothermal synthesis, these hard templates can be removed either by chemical etching or thermal annealing. Among them, PS and SiO_2 nanospheres are normally added with metal salt precursor. But carbon nanosphere template can be formed in the hydrothermal process. Hence, the template precursor and metal oxides precursors are mixed in one pot. Figure 4b shows the hydrothermal synthesis method for the formation of hollow metal oxide (Fe_2O_3, NiO, $7Co_3O_4$, CeO_2, MgO, and CuO) nanospheres [7]. In this process, various metal salts were dissolved together with glucose in water. During the hydrothermal treatment, carbon spheres are formed with metal ions incorporated into their hydrophilic shell. Under high-temperature calcination, hollow metal oxide spheres could be obtained. Changing the ratio between the metal salt and glucose can realize the control of the surface area and shell thickness of hollow oxides spheres.

Design of Hydrothermal Experiments

At the present stage of understanding the hydrothermal synthesis of nanostructures, it is difficult to predict the effect of the chosen reaction parameters. A majority of the hydrothermal synthesis work that has been done in the past has used trial-and-error methods for process development. This type of experimental approach suffers from the inability to clearly discern between processes that are controlled by either thermodynamics or kinetics. Thermodynamic processing variables, such as temperature, pH, and pressure, determine not only the processing space for a given material

but also influence both reaction and crystallization kinetics. Hence, to approach a rational materials design for hydrothermal synthesis (especially for new phases), thermodynamic calculations of phase equilibria to aid in the determination of the parameter space which will give the targeted phase need to be conducted. For a given precursor system, the effects of concentration, temperature, and pressure can be modeled to define the processing variable space over which the phase (s) of interest is stable. The application of in situ characterization techniques to study the reaction intermediates and reaction mechanisms also can significantly facilitate a rational design for hydrothermal synthesis for nanostructures with desired composition, crystal phase, size, and morphology.

Reduction

Key Principle of Reduction Synthesis

This process involves the reduction of an inorganic salt in liquid phase by reducers. This wet chemical reduction method is applied to synthesize metal nanoparticles as well as metal oxide nanostructures with decreased oxidation number. During metal nanoparticle synthesis, metal salt reduction can be expressed as

$$M^{n+} + ne^- \rightarrow M \quad (13)$$

Whether the redox reaction takes place or not relies on whether the standard electrode potential of the metals/ions is higher than that of the other redox pair.

The widely investigated metal nanomaterials include plasmonic Ag, Au, and Cu nanoparticles which can be used as biosensor imaging and optoelectronic devices; Pt, Pd, Ir, and Ru nanostructures as well as Ag and Au which have outstanding catalytic activity; and transition ferromagnetic metals, such as Fe, Co, and Ni nanoparticles which have been widely used in different magnetic and spintronic devices. A variety of bimetallic nanostructures have interested research communities because their performance often surpasses the enhancement of the properties associated with their single counterparts and exhibits many interesting and useful properties with broadened application fields often explained by synergistic effects of nanomaterials.

Metal salt reduction is well suited for metal and metal alloy nanoparticle synthesis. This facilitates pure, well-controlled nanoparticle formation with desired crystallinity. However, the reaction process can be complicated and the reaction by-products can be toxic. The discovery of various reducing agents and stabilizers has made possible the development of mild conditions that can be employed for the preparation of large amounts of nanoparticles of reproducible quality.

Major Parameters for Reduction Process

Same as other wet chemical synthetic methods, the solvent and temperature of the liquid system play important roles in reduction synthesis. It relates with the dissolution of precursor and stability of the target products/phases. High-temperature reduction in metal salts in the presence of stabilizing agents can be employed to produce transition metal nanoparticles (cobalt, nickel) that do not crystallize well at room temperature. Generally, the reaction temperature for reduction synthesis range from 60 °C or room temperature to 300 °C. And solvents can be water and/or organic compounds, such as polyol. Here, the effects on above two parameters won't be discussed in detail since pretty similar effects have been described in other wet chemical methods. Instead, the chemical factors, such as chemical composition of precursors, reducers, and stabilizers, will be focused in the following discussion.

Chemical Compositions of the Precursors

The species present in the solution have an impact on the resulting nanoparticles. Metal carbonyls are reported precursors for Fe, Co, and Ni nanosized colloids. To fabricate Au, Pt, Pd nanoparticles, metal halide salts (such as $AuCl_4^-$, $PtCl_6^{2-}$, $PdCl_2$) have been commonly used as precursors. Metal precursor choice has a major effect on the resulting nanoparticles and their behaviors. For example, the usual synthetic route to prepare Ag nanoparticles involves the

reduction of $AgNO_3$ or Ag organic compounds. Highly monodisperse myristate-capped silver nanoparticles can be obtained by suspending silver myristate (14 C in the alkyl chain) in triethylamine (reducer) followed by gentle heating at 80 °C for 2 h. This process gradually produces a solution of uniformly spherical silver nanoparticles that can be precipitated by addition of acetone. Thus, the nanoparticles can be isolated as solid materials and redispersed in nonpolar solvents where they are relatively stable. The particle size and size distribution are affected by the alkyl chain length of the carboxylate ligand in Ag organic compounds. With silver stearate (18 C in the alkyl chain) instead, the particles are highly monodispersed with sizes between 1.9 and 3.5 nm, whereas with silver octanoate (6 C in the main alkyl chain), the particle distribution is broad and the sizes are larger (between 5.5 nm and 31.5 nm) [11]. Here, the silver organic compounds played both as precursor and stabilizer or capping reagents. Many other metal organic compounds have also been developed based on that.

Reducers

As reported, the Ag particle size decreases, but the size distribution is also broader when octylamine is in the place of triethylamine as reducing agent [11]. *The size and morphology of the metal nanoparticles can be tuned by choosing a suitable reducing agent.* A weak reducing agent, such as citrate or tartrate, will favor a slow reaction allowing particle growth over a long period to yield faceted and small (<5 nm) nanoparticles, whereas with relatively strong reducing agents, such as $NaBH_4$, formamide, or o-anisidine, bigger and generally spherical nanoparticles will form. For achieving desired Au nanostructures, many variations of reduction method have been developed in the past decade. $NaBH_4$, sodium citrate, potassium bitartrate, dimethyl formamide, ascorbic acid, and alcohols are some of the reducing agents that have been successfully used. $NaBH_4$, potassium bitartrate, superhydrides, amines, and alcohols are widely reported as reducing agents for fabrication of Pd and Pt nanoparticles.

In some case, the solvent can be the reducing agent. One example is the synthesis of monodisperse silver nanocubes (100 nm) by reduction of silver nitrate with ethylene glycol. In this preparation, PVP stabilizer was used to form a monolayer on the surface of Ag nanoparticles to prevent fast nanoparticle growth and agglomeration. Ethylene glycol serves both as reducing agent and solvent. Moreover, polyol synthesis was a simple and versatile route to many other colloidal metals and/or alloys such as Au, Cu, Co, Ir, Ni, Pd, Pt, Ru, CoNi, and FeNi. Reasons for the popularity and versatility of this synthesis include the ability for polyols to dissolve many precursor salts and their highly temperature-dependent reducing power [12].

Stabilizers

Stabilizer, sometimes called protective agent or capping agent, can be used to form a monolayer on the surface of nanoparticles to prevent fast nanoparticle growth, Ostwald ripening, or agglomeration. Moreover, the concentration of stabilizer, or more effectively the ratio of (surface group of stabilizer)/(metal salt), is one of the critical factors to metal particle size and size distribution. And the ratio varies for different metal nanoparticles. Stabilizer can be polymer, aromatic amine, and alkanethiols, etc., depending on the weak bonding between metal and organic compounds.

Thiols are excellent stabilizers for capping gold nanoparticles owing to strong Au-S bonding, which enables the isolation of the particles as solid materials following solvent evaporation. The isolated particles can be readily redissolved in nonpolar solvents where they are indefinitely stable. Other good protective agents, such as phosphanes, phosphines, amines, chalcogenides, carboxylates, and polymers, have been successfully used. Long-chain n-alkanethiols are the most common protective agents employed to stabilize silver colloids, even though aromatic amines (such as aniline), carboxylic acids, and polymers have been also employed. The most common capping agents are n-alkanethiols and polymers, but a few examples of Pt and Pd nanoparticles capped with amines, alkyl isocyanides, and cyclodextrins have also

Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Table 1 Precipitation diagram of aqueous solutions

	NO_3^-	Cl^-	SO_4^-	OH^-	CO_3^{2-}	PO_4^{3-}
Group 1 Cations (Na^+, K^+) and NH_4^+						
Group 2 Cations (Mg^{2+}, Ca^{2+}, Ba^{2+})			**$BaSO_4$**	**$Mg(OH)_2$**		
Transition metal cations		**AgCl**				

been reported. Co, Ni, and Ni-Co highly monodispersed nanoparticles have been achieved in the presence of stabilizers including oleic acid, TOP, tributylamine, and hexadecylamine via reduction of metal acetates in solution by polyalcohols or $NaBH_4$.

Precipitation

Key Principles of Precipitation Reaction

This is a process in which insoluble solid particles separate out of a solution. To obtain nanoparticles from this process in a controlled manner, it normally requires mixing of a cation-containing solution with an anion-containing solution. Insoluble nanoparticles are formed upon mixing of the two clear solutions. The following reaction illustrates how $Ni(OH)_2$ nanoparticles are synthesized via precipitation reactions:

$$\left(Ni^{2+} + 2Cl^-\right) + \left(2Na^+ + 2OH^-\right) = Ni(OH)_2 \downarrow + 2Na^+ + 2Cl^- \quad (14)$$

For the reaction above, one aqueous solution containing nickel chloride ($NiCl_2$) and the other with sodium hydroxide (NaOH) are mixed together to produce the $Ni(OH)_2$ precipitate. To predict this type of reaction, one needs to know whether the product is insoluble in the solvent (e.g., water) or not. Table 1 shows the precipitation diagram for aqueous solutions [13]. The compounds inside the hatched areas are not soluble, while the compounds inside the blank areas are soluble. For example, Na_2CO_3 is soluble in an aqueous solution, while $MnCO_3$ is not. Thus, when a $Ni(NO_3)_2$ solution is mixed with Na_2CO_3, $NiCO_3$ will precipitate out of the solution. Table 1 also reveals that AgCl, $BaSO_4$, and $Mg(OH)_2$ are insoluble, while all other chlorides and sulfates are soluble. Based on Table 1, many nanoparticles can be synthesized via precipitation reactions.

Major Parameters of Precipitation Reactions

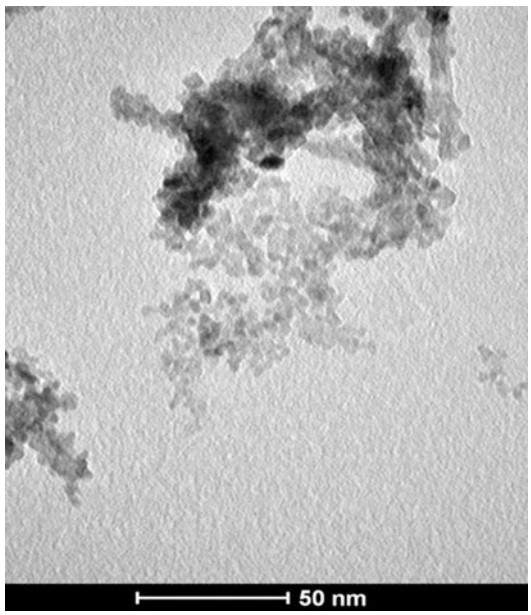
Precipitants

Precipitation reactions are often induced by addition of a precipitant, without which precipitation will not occur. Ammonium hydroxide (NH_4OH) is often used as a precipitant to synthesize hydroxides and oxides [14]. For example, to synthesize gadolinium-doped ceria ($Ce_{0.8}Gd_{0.2}O_{1.9}$), particles of 10 nm in diameter (Fig. 5), cerium nitrate, $Ce(NO_3)_3$, and gadolinium nitrate, $Gd(NO_3)_3$, are dissolved in deionized water first in a right proportion. The nitrate solution is then added at room temperature dropwise to the NH_4OH precipitant, while stirred using a magnetic bar, and maintained at a pH of 10. The solution is then heated to 75 °C while being stirred continuously and homogenized at that temperature for 1 h [14].

Selection of a proper precipitant is essential for synthesizing nanoparticles with the desired composition. To synthesize nanoparticles of $Ce_{0.8}Gd_{0.2}O_{1.9}$ plus (Ni,Cu)O, one can start with an aqueous solution containing proper amounts of $Ce(NO_3)_3$, $Gd(NO_3)_3$, $Ni(NO_3)_2$, and $Cu(NO_3)_2$. However, using NH_4OH as the precipitant results in nanoparticles of $Ce_{0.8}Gd_{0.2}O_{1.9}$ + NiO. No CuO or (Ni,Cu)O is produced, indicating that Cu^{2+} ions do not precipitate out as hydroxides or oxides from the solution. In contrast, using NaOH as the precipitant leads to the formation of $Ce_{0.8}Gd_{0.2}O_{1.9}$ + (Ni,Cu)O nanoparticles because NaOH is a stronger precipitant than NH_4OH.

Oxygen Dissolved in Water

The presence of oxygen in aqueous solutions could oxidize the precipitate, leading to a product with the composition different from that predicted from Table 1. The effect of oxygen dissolved in water can be deduced based on electrochemical potential of the element, as illustrated below. Adding NH_4OH to each of the $Ni(NO_3)_2$, $Co(NO_3)_2$, $Fe(NO_3)_2$, and $Mn(NO_3)_2$ solutions can result in hydroxides and oxides with different oxidation states. As indicated by the reactions below, $Ni(OH)_2$ will be the final product of the precipitation reaction of Ni^{2+} ions even in the presence of oxygen dissolved in water. However, this is not the case for other metallic elements. For Fe^{2+} and Mn^{2+} ions, their precipitation products, $Fe(OH)_2$ and $Mn(OH)_2$, will be oxidized by the oxygen in the solution to form Fe_2O_3 and Mn_3O_4, respectively. For Co^{2+} ions, its precipitation product, $Co(OH)_2$, will also be oxidized by the oxygen in the solution, but not to form oxides. Instead, an oxyhydroxide, CoO(OH), is formed, suggesting the Co atoms have a better resistance to oxidation than Fe and Mn. By comparing the final products of these metallic ions with their standard electrochemical potentials in aqueous solution (Table 2), it can be seen that the propensity for each metallic hydroxide to be oxidized by the oxygen in the aqueous solution is inversely proportional to its standard electrochemical potential. The lower the potential is, the easier its hydroxide is to be oxidized by the oxygen in the aqueous solution to form its respective oxide. Ni has the highest resistance to oxidation among these elements, and thus its $Ni(OH)_2$ is stable even in the presence of oxygen in the aqueous solution. Co has an intermediate resistance to oxidation, and thus its $Co(OH)_2$ is oxidized to an oxyhydroxide, a compound between a hydroxide and an oxide. Fe and Mn have the lowest resistance to oxidation among these elements, and thus its hydroxides are oxidized all the way to form oxides.



Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Fig. 5 Transmission electron microscopy (TEM) image of $Ce_{0.8}Gd_{0.2}O_{1.9}$ nanoparticles synthesized via precipitation. Note the uniform size of particles of ~10 nm in diameter

$$Ni^{2+} + 2\,OH^- = Ni(OH)_2 \quad (15)$$

$$Co^{2+} + 2\,OH^- = Co(OH)_2 \quad (16)$$

$$Co(OH)_2 + ¼\,O_2 = CoO(OH) + ½\,H_2O \quad (17)$$

$$Fe^{2+} + 2\,OH^- = Fe(OH)_2 \quad (18)$$

$$2\ Fe(OH)_2 + ½\,O_2 = Fe_2O_3 + 2\,H_2O \quad (19)$$

$$Mn^{2+} + 2\ OH^- = Mn(OH)_2 \quad (20)$$

$$3\ Mn(OH)_2 + ½\,O_2 = Mn_3O_4 + 3\,H_2O \quad (21)$$

Multicomponent Compounds via Coprecipitation

Multicomponent hydroxides and oxides can be formed via coprecipitation of insoluble solid

Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Table 2 Standard electrochemical potentials of various species in aqueous solution at 25 °C

Species	Mn^{2+}/Mn	Fe^{2+}/Fe	Co^{2+}/Co	Ni^{2+}/Ni
E^o	−1.182 V	−0.409 V	−0.282 V	−0.236 V

with multicomponents. The synthesis of $Ce_{0.8}Gd_{0.2}O_{1.9}$ discussed in section "Precipitants" is a good example of coprecipitation. In the case of $Ce_{0.8}Gd_{0.2}O_{1.9}$, the two cations have similar properties (e.g., similar propensity for oxidation), and thus the molar ratio of Ce to Gd in the precipitated $Ce_{0.8}Gd_{0.2}O_{1.9}$ nanoparticles is identical to that of the Ce ions to Gd ions in the solution. However, this may not be the case if the cations have dissimilar properties. A good example for this case is the synthesis of $Ni_{0.5}Mn_{0.5}(OH)_2$ via coprecipitation [15]. When the nitrate solution containing equal molar ratios of Ni and Mn cations is bubbled with N_2, the precipitate is $Ni_{0.5}Mn_{0.5}(OH)_2$. However, if the nitrate solution is not bubbled with N_2, the precipitate is composed of $Ni_{5/6}Mn_{1/6}(OH)_2(NO_3)_{1/6} + Mn_3O_4$. Such a dramatic change in the final products is due to the fact that Mn cations are easier to get oxidized than Ni cations (see Table 2). As a result, Mn cations are oxidized to become Mn_3O_4 by the oxygen in the solution if the oxygen is present in the solution.

Solvents for Precipitation

For precursors that are unstable in an aqueous solution, precipitation in a nonaqueous solvent is often utilized. Synthesis of semiconductor PdSe nanoparticles via precipitation is a good example of this kind [15]. In this process, lead-cyclohexanebutyrate (Pb-cHxBu) and selenium precursors are dissolved in a tri-butyl-phosphine (TBP) solution at room temperature. Another solution containing the tri-octyl-phosphine-oxide (TOPO) surfactant is prepared at 150 °C. The TBP solution is then injected rapidly into the TOPO solution, followed by immediate drop of the temperature to 118 °C. The reaction between Pb-cHxBu and Se at 150 °C leads to the formation of PdSe, and the rapid cooling to 118 °C results in precipitation, while TOPO serves as a capping agent to prevent the growth of PdSe nanoparticles [15]. Nonaqueous precipitation often offers excellent crystallinity, monodispersity, and high-purity nanoparticles. However, disadvantages may include use of expensive precursors, sensitivity to moisture, and complicated experimental setups.

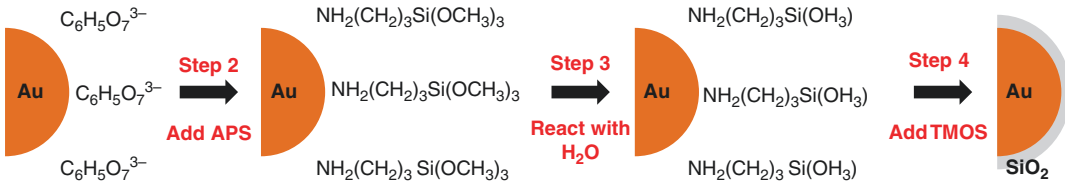
Core/Shell Structures

Key Principles in Synthesizing Core/Shell Structures

As pointed out before, the methods for synthesis of core/shell structures are typically derived from a combination of two different wet chemistry methods or from one wet chemistry method with two steps of different chemistries. Therefore, the basic principles of synthesis science for core/shell structures are in many cases the same as individual wet chemistry methods except the additional requirement of interfacial control. The purpose of interface control is to create a preferential environment for the shell material to deposit on the surface of the core and prevent homogeneous nucleation of the shell material. If the latter occurs, one ends up with a mixture of two types of nanoparticles of different compositions instead of the desired core/shell structure. Thus, our discussion below will focus on interfacial control in synthesizing different core/shell systems, including metal–polymer, metal–ceramic, polymer–metal, polymer–ceramic, ceramic–polymer, ceramic–metal, and semiconductor–semiconductor core/shell structures. The principles that guide the interfacial control for each of these systems will be illustrated by examples below.

It should be noted that not all core/shell structures are produced through "two-step" methods mentioned above. Some core/shell structures are synthesized via "one-pot" methods. However, one-pot methods are chemistry specific and thus only work for certain systems. In contrast, two-step methods are very general and can be used to synthesize almost any core/shell structures. Since one-pot methods are system specific, the coverage of this topic would require the discussion of system by system. The space here does not permit such coverage, and thus this topic will not be addressed further.

To simplify the discussion below, core/shell structures will be expressed using a shorthand defined here. For instance, CdSe/ZnS will stand for a CdSe core with a ZnS shell. Ag/SiO_2 will represent an Ag core with a SiO_2 shell.



Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Fig. 6 Schematic of the procedure in synthesizing the Au/SiO_2 core/shell structure. Step 1 – reduction reaction to synthesize Au nanoparticles (not shown); Step 2 – adding APS; Step 3 – hydrolysis of the silane group of APS; and Step 4 – hydrolysis and condensation of TMOS to form the SiO_2 shell. See the text for more description of the procedure and reactions

Semiconductor–Semiconductor Core/Shell Structures

Semiconductor core–shell structures, such as CdSe/ZnS, CdSe/CdS, and PdSe/CdSe, are normally synthesized by two-step precipitations. The first precipitation reaction leads to the formation of the core, and the second precipitation reaction forms the shell. For most of these core and shell structures, the lattice mismatch between the core and shell is very small (~4 %) and thus allow hetero-epitaxial growth of the shell directly on the surface of the core. Under these conditions, no surfactants are needed, and the core/shell formation is achieved by preventing homogeneous nucleation of the shell material in the solution. Only growth on the surface of the core is promoted. This can be done by controlling the concentration of the growth species so that the supersaturation is not high enough for nucleation, but high enough for growth. There are normally two approaches to control the supersaturation of growth species. One is by the dropwise addition of the growth precursor solution into the reaction mixture which contains the cores. Another approach is to vary the growth temperature because the supersaturation is a function of temperature. An example synthesizing the CdSe/ZnS core–shell structure using both approaches coupled with two-step precipitations can be found in Hines and Guyot-Sionnest [16].

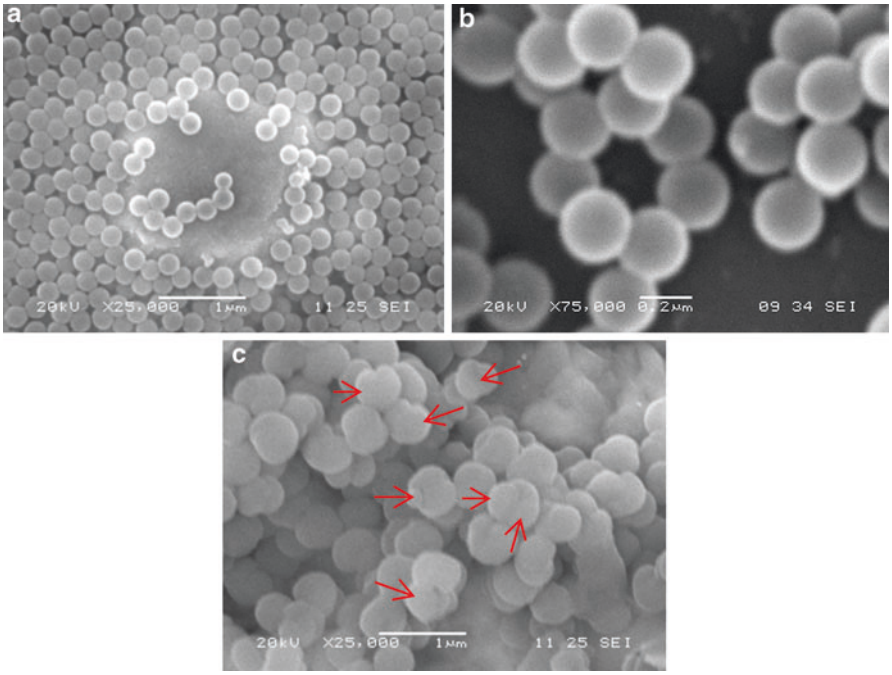
Metal–Ceramic Core/Shell Structures

Such core/shell structures often require the use of coupling agents to allow the ceramic shell to interact with the metal core. An example of this kind is the Au/SiO_2 core/shell structure. Gold surface has very little electrostatic affinity for silica because gold does not form a passivation oxide layer. Thus, to ensure that SiO_2 can grow on the surface of Au, (3-aminopropyl) trimethoxysilane (APS) can be used which links to Au particle surface at one end and to SiO_2 at the other end [17]. Figure 6 shows schematically how the Au/SiO_2 core/shell structure is synthesized with the aid of APS.

As shown in Fig. 6, the Au nanoparticles are first synthesized via reduction of chloroauric acid ($HAuCl_4$) by sodium citrate ($Na_3C_6H_5O_7$). At the as-synthesized condition, the surface of Au nanoparticles is negatively charged due to the absorption of negatively charged citrate groups. In Step 2, an APS aqueous solution is added to the Au sol to allow the complexation of the amine groups with the Au surface, leading to the replacement of the citrate group by the amine group, with the silane group of APS pointing into the solution. The silane group then undergoes hydrolysis to form silanol groups (Step 3), which can then participate in the condensation reaction with tetramethyl orthosilicate (TMOS), the SiO_2 precursor, to form the SiO_2 shell in Step 4.

Polymer–Ceramic Core/Shell Structures

Interfacial control is also necessary for polymer–ceramic core/shell formation because chemical properties of polymers and ceramics are dramatically different. A good example to illustrate the interfacial control in synthesizing polymer–ceramic core/shell structures is the polystyrene (PS)/SiO_2 core/shell structure. After formation of PS nanoparticles via emulsion polymerization, PS nanoparticles can be dispersed in water with the aid of sulfates or amines. Sulfates



Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Fig. 7 SEM images of polystyrene nanoparticles before and after subjecting to the sol–gel reaction to form the SiO_2 shell under different conditions: (**a**) before subjecting to the sol–gel reaction, (**b**) after subjecting to the sol–gel reaction using the positively charged PS nanoparticles, and (**c**) after subjecting to the sol–gel reaction using the negatively charged PS nanoparticles. *Arrows* in (**c**) point out several nonspherical particles and particles with different sizes

result in PS nanoparticles with negatively charged surface, whereas amines give rise to PS nanoparticles with positively charged surface. In order to form the SiO_2 shell on the surface of PS nanoparticles via the sol–gel process, PS nanoparticles with proper surface charges are necessary. Using tetraethyl orthosilicate (TEOS) as the SiO_2 precursor and conducting sol–gel reactions at pH > 7, PS nanoparticles with positively charged surface (amine-stabilized) should be used to form the core/shell structure. This is because of the negative charge of the intermediate $\equiv Si-O^-$ formed during hydrolysis of the sol–gel reaction, as show in Eqs. 22 and 23.

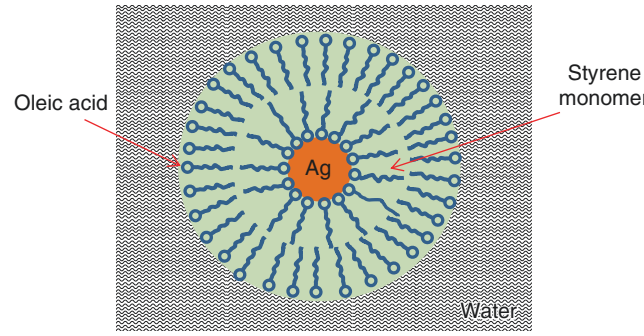
$$\equiv Si - OH + OH^- \rightarrow \equiv Si - O^- + H_2O \quad (22)$$

$$\equiv Si - O^- + \equiv Si - OH \rightarrow \equiv Si - O - Si \equiv + OH^- \quad (23)$$

A positively charged PS surface allows the condensation reaction to take place at the surface of PS nanoparticles and thus the formation of the core/shell structure. If PS nanoparticles are negatively charged at their surface, the intermediate $\equiv Si-O^-$ cannot approach the PS surface due to charge repulsion. As a result, homogeneous nucleation of SiO_2 takes place in the solution and no core/shell structure is formed. Figure 7 compares the scanning electron microscopy (SEM) images of nanoparticles after sol–gel formation of SiO_2 in a PS nanoparticle sol. It can be seen that when PS is positively charged at the surface, the resulting particle sizes are very uniform, indicating the formation of a uniform SiO_2 shell on the PS surface. In contrast, when PS is negatively charged at the surface, different sizes of SiO_2 particles are present, and some particles have irregular shapes, indicating no formation of the core/shell structure.

Metal–Polymer Core/Shell Structures

Interfacial control is required for metal–polymer core/shell formation because chemical properties



Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods, Fig. 8 Schematic of formation of the Ag/PS core/shell structure in the presence of oleic acid

of metals and polymers are dramatically different. The importance of the interfacial control for this group of core/shell structures can be well illustrated by the Ag/PS core/shell structure. To synthesize the Ag/PS core/shell structure, one can start with reduction of $AgNO_3$ to form Ag sol, followed by polymerization of styrene in emulsion of oleic acid [18]. The thickness of the PS shell can be readily controlled by changing the concentration of monomers in emulsion. The presence of oleic acid ($C_{17}H_{33}COOH$) is required to form the polymer shell. Without oleic acid, only polymer particles are formed and Ag particles are not coated with the polymer. Figure 8 illustrates the principle of forming the PS shell on the surface of an Ag particle. Oleic acid has a high affinity for the oxide layer on the surface of Ag. Thus, it absorbs on the surface of Ag nanoparticles to form a micro-reactor within which styrene monomers are present, and polymerization of styrene can proceed when water-soluble initiator is added. Clearly, oleic acid is essential for the formation of the PS shell.

Ceramic–Polymer Core/Shell Structures

In general, this group of the core/shell structure also requires interfacial control. This can be illustrated well using the formation of the SiO_2/poly (divinylbenzene) (PDVB) core/shell structure. Monosized SiO_2 nanoparticles can be synthesized first via the sol–gel process. After generation of SiO_2 nanoparticles, they are pretreated with a coupling agent, e.g., 4-vinylpyridine, and then mixed with divinylbenzene monomers and a radical polymerization initiator [19]. The pretreatment with 4-vinylpyridine imparts the surface of SiO_2 particles with the vinyl functional group which can subsequently participate in polymerization with divinylbenzene monomers and thus promote the formation of the PDVB shell.

Metal–Metal Core/Shell Structures

Similar to semiconductor–semiconductor core/shell structures, metal–metal core/shell structures may not need surfactants because of the similar properties between the core and shell materials. The Pd/Ag core/shell structure synthesized via co-reduction can serve as a good example for this synthesis method. A $Pd(NO_3)_2$ and $AgNO_3$ mixture in an aqueous solution can be reduced by $NaBH_4$ readily to form the Pd/Ag core/shell structure. During the co-reduction process, Pd will be reduced first because it has a higher redox potential than Ag. Thus, Pd nanoparticles grow first, which is then followed by reduction of Ag ions and the direct growth of Ag on the surface of the Pd nanoparticles.

Concluding Remarks

In all, the wet chemical synthetic methods have been predominant in synthesis numerous nanostructures over decades. The synthetic methods discussed above all possess virtues over physical deposition methods or top-down methods, including that it is very simple, is less costly, can easily be scaled-up, and easily be tuned by varying

the synthesis parameters like temperature, precursors, and concentration of the solution, etc. Core–shell nanostructures are just one of the typical examples for the ability of these wet chemical methods toward hierarchical nanostructures [20]. By understanding the mechanism and effective factors, diversity of advanced nanostructures can be synthesized in large volume in a controlled manner. These nanomaterials and synthetic technologies have been applied in so many devices and appliances (from military to domestic) and affected our daily life deeper than we noticed. They are the foundation of our nanotechnologies and certainly will affect our life more in every aspect in the future.

Cross-References

- ▶ Bottom-Up Nanofabrication
- ▶ Gel Chemical Synthesis
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Sol–Gel Method
- ▶ Synthesis of Nanoparticles
- ▶ Wet Chemical Processing

References

1. Livage, J., Sanchez, C.: Sol–gel chemistry. J. Non-Cryst. Solids **145**, 11–19 (1992)
2. Shaw, L., Shen, C., Thomas, E.L.: Synthesis of gadolinia-doped ceria gels and powders from acetylacetonate precursors. J. Sol–Gel Sci. Technol. **53**, 1–11 (2010)
3. Livage, J., Henry, M., Sanchez, C.: Sol–gel chemistry of transition metal oxides. Prog. Solid State Chem. **18**, 259–341 (1988)
4. Shen, C., Shaw, L.: FTIR analysis of the hydrolysis rate in the sol–gel formation of gadolinia-doped ceria with acetylacetonate precursors. J. Sol–Gel Sci. Technol. **53**, 571–577 (2010)
5. Brinker, C.J.: Hydrolysis and condensation of silicates: effects on structure. J. Non-Cryst. Solids **100**, 31–50 (1988)
6. Einarsrud, M.-A., Grande, T.: 1D oxide nanostructures from chemical solutions. Chem. Soc. Rev. **43**, 2187–2199 (2014)
7. Shi, W., Song, S., Zhang, H.: Hydrothermal synthetic strategies of inorganic semiconducting nanostructures. Chem. Soc. Rev. **42**, 5714–5743 (2013)
8. Liu, C., Chen, H., Ren, Z., Dardona, S., Piech, M., Gao, H., Gao, P.-X.: Controlled synthesis and structure tunability of photocatalytically active mesoporous metal-based stannate nanostructures. Appl. Surf. Sci. **296**, 53–60 (2014)
9. Geng, B.Y., Fang, C.H., Zhan, F.M., Yu, N.: Synthesis of polyhedral $ZnSnO_3$ microcrystals with controlled exposed facets and their selective gas-sensing properties. Small **4**, 1337–1343 (2008)
10. Li, Y., Shen, W.: Morphology-dependent nanocatalysts: Rod-shaped oxides. Chem. Soc. Rev. **43**, 1543–1574 (2014)
11. Masala, O., Seshadri, M.: Synthesis routes for large volumes of nanoparticles. Annu. Rev. Mater. Res. **34**, 41–81 (2004)
12. Chen, H.M., Liu, R.-S.: Architecture of metallic nanostructures: synthesis strategy and specific applications. J. Phys. Chem. C **115**, 3513–3527 (2011)
13. Masterton, W.L., Hurley, C.N.: Chemistry Principles and Reactions, 5th edn. Brooks/Cole – Thomson Learning, Belmont, MA/CA (2004)
14. Lavorato, P., Shaw, L.: Low temperature sintering of gadolinium-doped ceria for solid oxide fuel cells. Ceram. Trans. **238**, 65–75 (2012)
15. Sashchiuk, A., Amirav, L., Bashouti, M., Krueger, M., Sivan, U., Lifshitz, E.: PbSe nanocrystal assemblies: synthesis and structural, optical, and electrical characterization. Nano Lett. **4**, 159–165 (2004)
16. Hines, M.A., Guyot-Sionnest, P.: Synthesis and characterization of strongly luminescing ZnS-capped CdSe nanocrystals. J. Phys. Chem. **100**, 468–471 (1996)
17. Liz-Marzan, L.M., Giersig, M., Mulvaney, P.: Synthesis of nanosized gold – silica core – shell particle. Langmuir **12**, 4329–4335 (1996)
18. Quaroni, L., Chumanov, G.: Preparation of polymer-coated functionalized silver nanoparticles. J. Am. Chem. Soc. **121**, 10642–10643 (1999)
19. Oyama, H.T., Sprycha, R., Xie, Y., Partch, R.E., Matijevic, E.: Coating of uniform inorganic particles with polymers. Coll. Polym. Sci. Interf. Sci. **160**, 298–303 (1993)
20. Gao, P.-X., Shimpi, P., Gao, H., Liu, C., Guo, Y., Cai, W., Liao, K.-T., Wrobel, G., Zhang, Z., Ren, Z., Lin, H.-J.: Hierarchical assembly of multifunctional oxide-based composite nanostructures for energy and environmental applications. Int. J. Mol. Sci. **13**, 7393–7423 (2012)

Nanopatterned Substrata

▶ Nanopatterned Surfaces for Exploring and Regulating Cell Behavior

Nanopatterned Surfaces for Exploring and Regulating Cell Behavior

Jason P. Gleghorn[1] and Celeste M. Nelson[1,2]
[1]Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, USA
[2]Department of Molecular Biology, Princeton University, Princeton, NJ, USA

Abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
AFM	Atomic force microscopy
ECM	Extracellular matrix
ES	Embryonic stem cell
MSC	Mesenchymal stem cell
RGD	Argine-glycine-aspartic acid
SAM	Self-assembled monolayer

Synonyms

Nanopatterned substrata

Definition

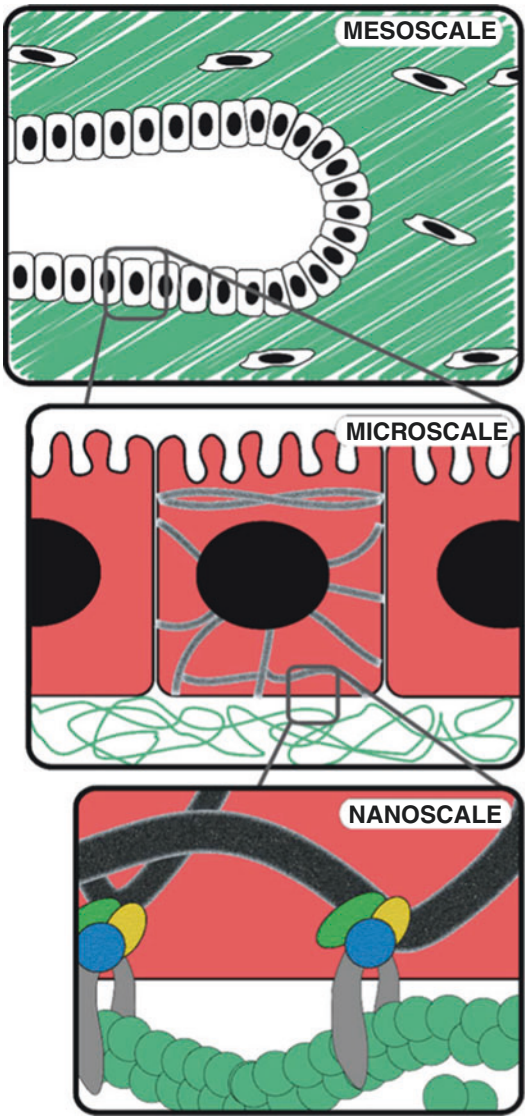
Surfaces (substrata) that are fabricated to have chemically or topographically nanometer-sized features, which are used to investigate cell morphology, phenotype, and/or behavior.

Overview

Cells Sense and Respond to Their Environment

Cellular environments consist of complex mechanical, chemical, electrical, and topological gradients ranging in size from the nanoscale (1–100 nm) to the macroscale (>1 mm) (Fig. 1). Cell behavior is governed by these multiscale environmental properties, which, when disrupted, lead to aberrant tissue function and disease. At the mesoscale (100–1,000 μm), aggregate cell behavior resulting from cell-cell interactions, coupled with properties of the extracellular matrix (ECM), create the tissue architecture and sculpt the mechanical and chemical microenvironment. Heterogeneity of the tissue architecture at the mesoscale produces unique microenvironments. Cell-cell and integrin-dependent cell-ECM interactions at the microscale (1–100 μm) regulate cell shape and adhesion, and consequently affect cell survival, growth, differentiation, and migration. At the nanoscale, molecular and subcellular processes such as integrin activation, focal adhesion formation, actin polymerization, and cytoskeletal organization enable the cell to be physically coupled to its ECM. Additionally, membrane-bound receptors signal through intracellular pathways to modulate gene expression and protein synthesis in response to chemical signals such as morphogens.

Cells interact with and regulate their surroundings in a relationship referred to as dynamic reciprocity [1]. Through dynamic and reciprocal crosstalk, cell behavior is modulated by cell-cell and cell-ECM interactions, and that behavior sculpts and defines the microenvironment. One of the many ways in which the cell senses and reacts to its environment is through integrin-dependent signaling. Transmembrane integrins attach and mechanically couple to several ECM ligands, including the prototypic sequence arginine-glycine-aspartic acid (RGD). Integrin engagement leads to recruitment and clustering to form focal adhesions. Intracellularly, integrins within the focal adhesion associate with several proteins including vinculin, talin, tensin, α-actinin, paxillin, Src, and focal adhesion kinase. These proteins couple the focal adhesion to the actin cytoskeleton. This physical linkage enables the cell to adhere to, move along, and interact mechanically with the ECM. In addition, the focal adhesion activates several kinase cascades that biochemically transduce information about the microenvironment into the nucleus, which leads to changes in gene expression. These gene expression changes can result in the production and secretion of soluble signaling molecules, enzymes, and ECM components that



Nanopatterned Surfaces for Exploring and Regulating Cell Behavior, Fig. 1 Cell interaction and behavior are governed over multiple length scales. Cell-cell and cell-matrix interactions collectively form tissue architecture and function at the mesoscale. Microscale interactions guide cell shape and connectivity, whereas nanoscale interactions influence cell processes such as adhesion, motility, and gene expression

shape and maintain the extracellular environment. The formation, maturation, and disassembly of focal adhesions are not only key for cell spreading and migration, but also appear to be central modulators of many cellular functions including proliferation and differentiation.

Engineering Model Substrata

The tissue microenvironment is complex and the responses of cells to microenvironmental interactions can occur through individual and synergistic mechanisms. As a result, model substrata are often employed to decouple and isolate the effects of various microenvironmental factors. These model surfaces typically fit into one of two categories: chemically or topologically patterned. Chemically patterned surfaces contain regions of active and inactive (inert) biomolecules and thus present spatial patterns of chemical ligands to the cells of interest. ECM proteins such as fibronectin, vitronectin, laminin, and collagen or their active binding motifs are often used. In contrast, topologically patterned surfaces alter the surface topology that is presented to the cells. In these systems, etched pits, posts, grooves, and gratings are used to characterize cell behavior. Regardless of the type of model substratum used, advances in micro- and nanotechnology have enabled the creation of such tailored surfaces to investigate the interactions between cells and their microenvironment at multiple length scales.

Because of the existence of several simple, well-established techniques, micropatterned surfaces have become a popular tool to interrogate cell behavior. Micrometer-scale topology can be created using standard photolithography coupled with traditional etching processes; however, at the micrometer scale most studies have focused on chemically patterned substrata. For chemically patterned surfaces, photolithographic techniques can be used to spatially deposit or remove different materials and thus define regions for chemical functionalization. Microcontact printing is the most common method of microscale chemical functionalization. This technique employs the use of micropatterned elastomeric stamps that are "painted" with the ECM ligand of interest. The coated stamps are then pressed onto a treated surface, which deposits the cell-adhesive ECM ligands onto the surface in the same spatial pattern as that of the stamp.

Micrometer-sized features can be used to control cell position and connectivity; and thus micropatterning is often employed to investigate both mesoscale and microscale cell behavior.

Micropatterned surfaces have been used to coculture fibroblasts and hepatocytes to maintain normal hepatocyte phenotype [2]. Additional mesoscale behavior has been investigated with epithelial sheets grown on micropatterned islands of varying size and shape. Patterns of cell proliferation emerge that result from cell-cell and cell-ECM interactions within the epithelial sheet [3]. Chemical micropatterning has also been used with single cells to identify the role of cell shape in determining cell fate [4]. Geometric control of cell growth and viability was demonstrated by varying the size of chemically patterned islands of ECM.

Although micropatterning is a powerful tool to investigate cell behavior, particularly at the mesoscale, this approach has its limitations. Micropatterning techniques do not afford the ability to control the surface density of the patterned ligand. Additionally, the minimum feature sizes achieved with these techniques do not permit the exposure of multiple different ligands to a single cell. These drawbacks make it difficult to probe cell-ECM interactions at the scale of integrin engagement or focal adhesion formation. Additionally, cells in vivo are not exposed to flat two-dimensional (2D) surfaces but rather to a complex topological microenvironment. The ECM consists primarily of interwoven protein fibers ranging from 10 to 300 nm in diameter. Basement membranes consist of complex mixtures of nanoscale pits, pores, and fibers ranging in size from 5 to 200 nm with peaks and valleys of approximately 100 nm in height [5]. The investigation of cell behavior in response to these native nanotopologies cannot be explored using micropatterned surfaces.

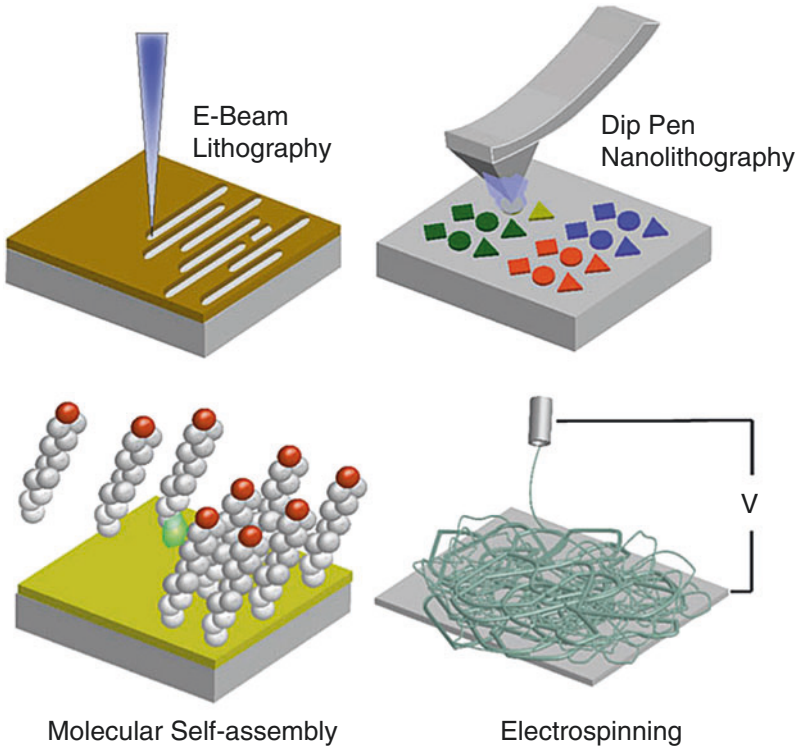
The advent of new tools and fabrication techniques has enabled the creation of spatial patterns in chemistry and topology that can vary greatly over the length scale of a cell. Nanoscale differences in surface roughness, fiber diameter, and ligand spacing can be produced using micropatterned substrata to more closely recapitulate the native tissue microenvironment. Thus, the goal of nanoscale patterning is to investigate molecular mechanisms and subcellular processes that determine fundamental cell behaviors including adhesion, proliferation, gene expression, and differentiation.

Fabricating Nanopatterned Substrata

Unlike micropatterned surfaces, fabrication of nanopatterned surfaces large enough for multicellular studies is technically challenging as well as time and labor intensive. As a result, studies to date have focused primarily on the behavior of single cells. The types of fabrication methodologies discussed herein are by no means comprehensive, but rather a sampling of commonly used techniques to create chemically and topologically defined nanopatterned surfaces. Fabrication technologies adapted from the semiconductor industry as well as novel nanobiotechnology methods are rapidly changing, enabling faster and simpler creation of nanopatterned features. For a more comprehensive review of fabrication techniques, the reader is directed to specialized reviews [6].

Topologically nanopatterned substrata are created using several techniques. These surfaces can be categorized as unordered or ordered topologies. Unordered topologies typically arise spontaneously as a result of processing methods including chemical etching and polymer demixing. Surfaces with unordered topologies have randomly patterned features and lack orientation and geometrical control, but the fabrication processes are relatively simple, fast, and inexpensive. Ordered topologies, on the other hand, are created with precise control of feature pattern, orientation, and geometry. Electron-beam (e-beam) lithography (Fig. 2) is the most commonly used approach to create nanoscale ordered topologies. E-beam lithography uses an electron beam to develop a layer of resist coated on the substratum. Following e-beam lithography, further processing of the substratum using standard etching techniques results in nanoscale topologies including nanoscale grooves, gratings, or pits.

Dip-pen nanolithography (Fig. 2) is a method to create chemically nanopatterned surfaces using an atomic force microscopy (AFM) tip [7]. The AFM tip is “dipped” into a reservoir, coating the tip with the ligand of interest, and is then used to repeatedly deposit small amounts of the ligand

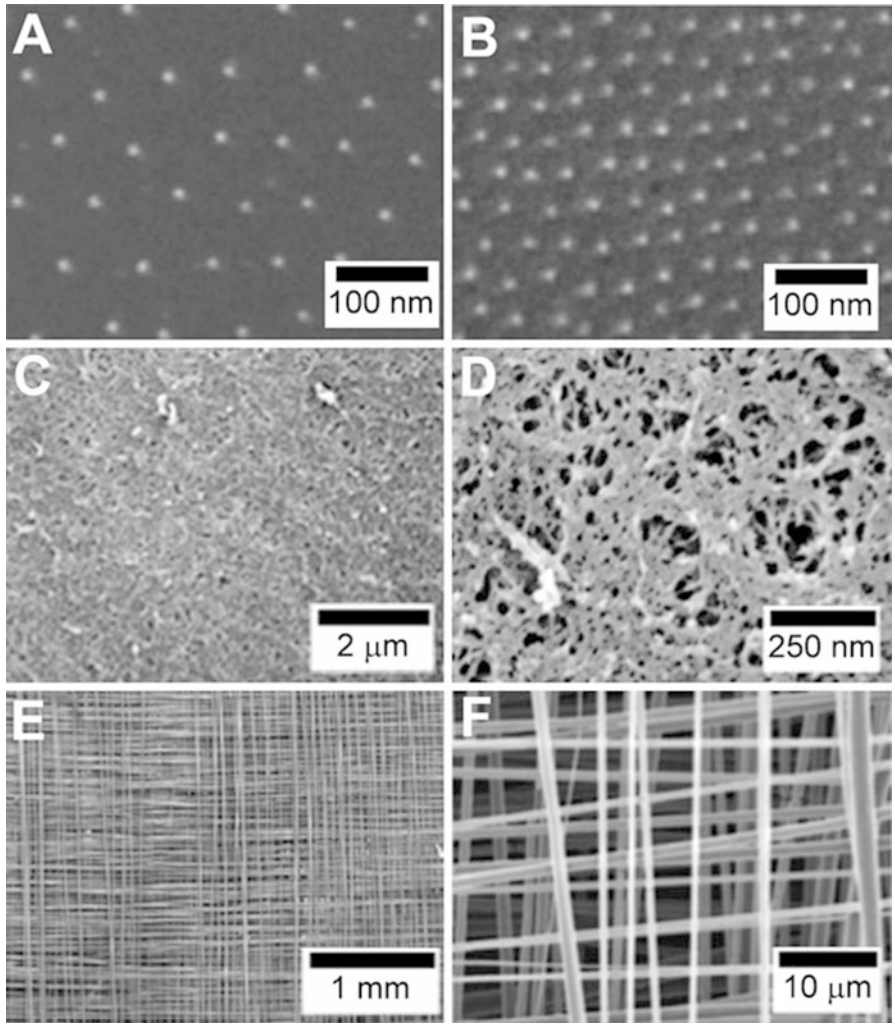


Nanopatterned Surfaces for Exploring and Regulating Cell Behavior, Fig. 2 Cartoon representations of various fabrication modalities used to create nanopatterned surfaces

onto the substratum to form chemically patterned regions with an approximate spatial resolution of 5 nm. This technology enables the placement of different adhesive ligands in close proximity in defined shapes with nanoscale precision and dimensions. Additionally, this technology can be multiplexed to create a higher throughput chemical patterning system by combining multiple AFM tips in parallel.

In addition to the "top-down" fabrication strategies used to deposit or remove material from a surface as described above, newer "bottom-up" technologies of molecular self- or templated-assembly (Fig. 2) have been used to create larger chemically nanopatterned surfaces. Molecular self-assembly is a broad category of techniques wherein (supra)molecular building blocks or colloids are spontaneously self-organized as a result of their interactions to form nanoscale topology and/or spatial patterns of cell-adhesive ligands [8]. Similarly, templated-assembly creates the same types of surfaces with the use of a reusable template to define an organization for the molecular building blocks. These newer fabrication techniques (e.g., self-assembled monolayers (SAMs) and phase-separated block copolymers) have enabled the reliable creation of large patterned areas.

Whereas all of the fabrication techniques described above can be combined to create surfaces with heterogeneously patterned features that vary across length scales, other techniques such as electrospinning [9] do exist to create patterned surfaces that simultaneously expose cells to topologies with widely different dimensions. In electrospinning, an electric field is used to produce polymer fibers from a liquid solution. As the fiber orientation and diameter are easily controlled with simple experimental parameters, fibrous mats containing both nanoscale and microscale fibers can be created with aligned or randomly oriented topologies. Additionally, multiple fibers and complex fiber geometries can be spun simultaneously from different materials or conjugated with different cell-adhesive ligands. This produces substrata that more closely mimic the topology of tissues in vivo: complex fibrous meshes with nano- to micro-sized fibers and randomly



Nanopatterned Surfaces for Exploring and Regulating Cell Behavior, Fig. 3 Example substrata. (**a**, **b**) Ligand-coated gold nanodots can be spaced at varying distances to form chemically nanopatterned substrata (Reproduced with permission from Selhuber-Unkel et al. [12]). (**c**, **d**) The basement membrane of the cornea has topographical variations with both micrometer- and nanometer-sized features (Reproduced with permission from Abrams et al. [5]). Electrospun meshes can be formed to create both unordered and ordered topographies (**e**) and fiber diameters can vary within individual samples (**f**) (Reproduced with permission from Carnell et al. [13])

sized micropores/pits. Thus, electrospun substrata are widely used to determine how cell behavior is regulated by fiber alignment and multiscale distributions of fiber size and surface ligands.

Key Research Findings

Chemically nanopatterned substrata have been primarily used to investigate integrin engagement and focal adhesion formation in an effort to dissect out cell adhesion properties. The majority of these studies have been carried out on self-assembled surfaces containing RGD-functionalized gold nanoparticles [10] (Fig. 3). Integrin receptors are approximately 10 nm wide and these "nanodots" can be fabricated such that they bind to single integrins. As a result, spacing of bound integrins can be altered by changing the spacing of the nanodots on the substrata.

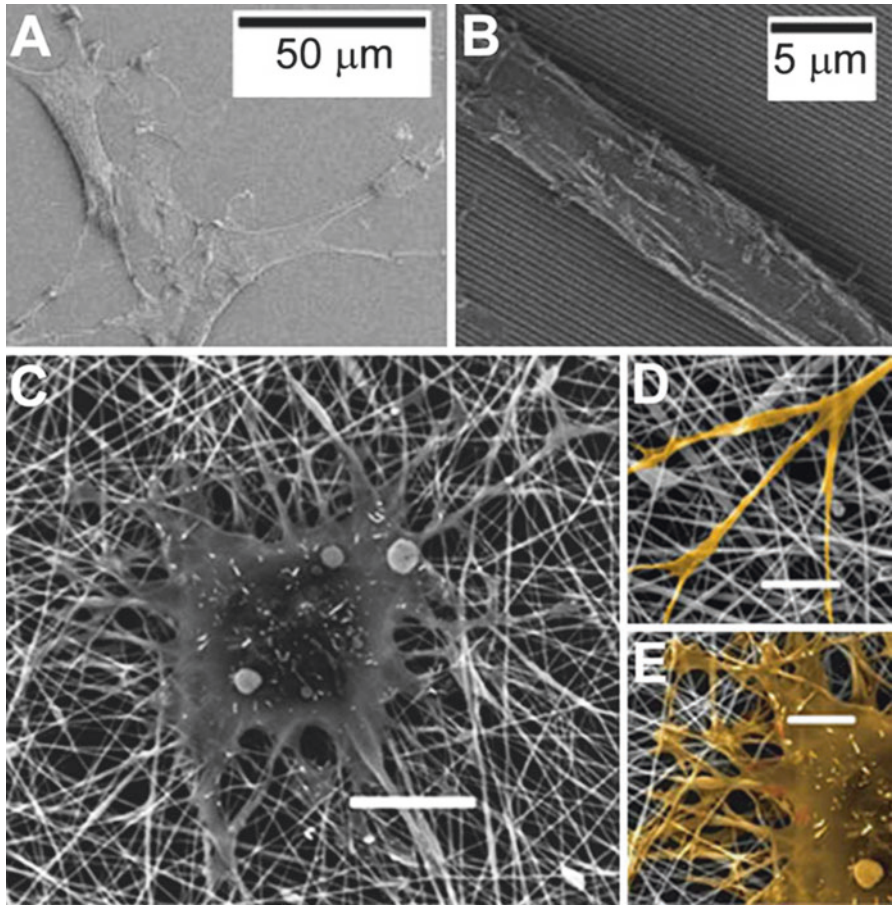
These nanopatterned surfaces have revealed that cell behavior is regulated by the spacing of adhesive ligands. Results from early studies showed that cell motility was affected on surfaces with RGD ligand spacings in the range of 6–300 nm and that spacings from 14 to 25 nm influenced cell-adhesion strength. Subsequent studies explored nanodot spacings between 28 and 85 nm [11] and found that RGD ligand spacings from 58 to 73 nm are optimal for integrin clustering and activation, whereas ligand separation greater than 73 nm causes significantly less cell adhesion, spreading, and integrin-mediated focal contact formation in osteoblasts. Additional studies have measured cell detachment forces from these substrata and demonstrated that spacings greater than 90 nm inhibited focal adhesion formation and decreased detachment forces compared to spacings less than 50 nm [12]. Overall, the data obtained using these chemically nanopatterned surfaces indicate that RGD ligand spacings should be less than approximately 70 nm to stimulate collective cell functions. Additionally, these stimulatory effects tend to increase as the ligand spacing decreases down to approximately 10 nm.

Studies using topologically nanopatterned substrata can be grouped into two categories, ordered or unordered, based on the layout of the surface features. Ordered surfaces consist of repeated features that are arranged with consistent height, spacing, and orientation. Such surfaces include those with uniform posts or pits and gratings formed by regularly spaced groves. Conversely, surfaces with randomly oriented, varying sized features, such as electrospun meshes, are classified as unordered. The mechanisms by which nanotopographic cues regulate cell behavior are not well understood. The emerging picture is that changes in cytoskeletal organization and structure may drive some of these cell behaviors, potentially in response to the underlying geometry or feature size of the surface. Regardless, it is clear that there is a range of different responses, or differing levels of response that vary by cell type.

Morphology and proliferation are altered when cells are cultured on surfaces with ordered nanotopography [14] (Fig. 4). Human mesenchymal stem cells (MSCs) grown on 350 nm-wide grooves have an aligned cytoskeleton along the direction of the features. Likewise, stem-cell-derived osteoblasts not only aligned and spread in response to nanogrooves of polystyrene, but also showed alignment in actin and mineralized matrix. Additional studies with human embryonic stem cells (ES) cultured on 600 nm polydimethylsiloxane (PDMS) ridges have demonstrated an altered organization of cytoskeletal components including F-actin, vimentin, γ-tubulin, and α-tubulin. The alterations to morphology and proliferation were eliminated when the ES cells were exposed to actin-disrupting drugs [15]. Generally, cells seem to be more sensitive to the depth of the groove as opposed to the groove pitch or spacing. Additionally, cell orientation increases with increasing groove depth on grated surfaces but decreases as the groove width or pitch increases.

Whereas the phenomenon of cell and cytoskeletal alignment is fairly conserved over a wide range of nanogratings of varying dimensions, cell behavior is more varied in response to ordered surfaces with pits and posts. The adhesion of osteoblasts was significantly increased when cultured on 11 nm-high posts compared to either 85 nm-high posts or flat polystyrene [16]. Similarly, fibroblasts also demonstrated increased adhesion and spreading on 13 nm posts compared with 95 nm-high posts. However, endothelial cells displayed less spreading and lower cytokine production when cultured on 100 nm-high posts as compared to flat control surfaces. In general, cells cultured on surfaces that are uniformly or randomly textured with nanometer-sized pits and pillars do not display aligned or preferential orientations.

Although comparing different unordered topologies is difficult due to differences in feature size, shape, uniformity, and chemistry, a trend does seem to emerge. In electrospun fiber meshes ranging from 283 to 1,425 nm in diameter, rat hippocampus-derived adult neural stem cells differentiate and proliferate in response to fiber diameter [17]. For fiber meshes smaller than 283 nm, cells did not demonstrate preferential patterns of spreading; however, as the fiber



Nanopatterned Surfaces for Exploring and Regulating Cell Behavior, Fig. 4 Examples of cell responses to nanotopography. Uniform cell spreading is observed on flat substrata (**a**), whereas cells spread and align along the direction of nanograted substrata (**b**) (Reproduced with permission from Yim et al. [14]). Cells spread, attaching to and around fibers on an unordered electrospun mesh (**c**–**e**). The cell is highlighted in *yellow* and scale bars are 10 μm (**c**) and 5 μm (**d**, **e**) (Reproduced with permission from Christopherson et al. [17])

diameter increased, cells extended along the length of the fiber. In contrast, proliferation increased as fiber diameter decreased. In another study investigating cell adhesion, fibroblasts demonstrated a higher level of adhesion to surfaces with unordered topology as compared to either flat surfaces or those with ordered topology [18]. In summary, the general trend is that cell behaviors such as proliferation, adhesion, and spreading are positively stimulated as the feature size decreases from ~100 to 10 nm.

While almost all of the studies to date have focused exclusively on either chemically or topologically patterned substrata, these parameters are difficult to decouple. Chemically nanopatterned surfaces inevitably have 5–10 nm variations in height across the substratum due to the islands of conjugated ligands. In topographically nanopatterned systems, differences in protein transport and adsorption due to structural features may result in differences in the presentation of adhesive areas and thereby alter cell behavior. Few studies have investigated the competitive or synergistic effects of both chemical and topographic cues [16]. When continuous chemically patterned strips of cell-adhesive ligand were orthogonally overlaid onto nanogrooves, fibroblasts preferentially oriented along the chemical

patterns. However, when fibronectin was orthogonally overlaid discontinuously along the top of a nanogroove pattern, osteoblasts preferentially aligned along the topographic pattern. From these limited data it is clear that a variety of systems involving both chemical and topological patterning over varying nanometer-sized scales need to be employed to fully understand interaction effects.

Applications

The basic science studies highlighted herein that seek to understand how nanoscale environmental cues influence cell behavior are fundamental to a range of applications. In particular, these data can be used to engineer biomaterial surface characteristics for improved performance. Materials that interact with the circulatory system, such as those used for stents or cardiac catheters, could be designed with a nanopatterned surface to prevent cell adhesion and clotting. Conversely, orthopedic implants, which need to integrate strongly with the surrounding tissue, could be designed with a nanopatterned bioactive surface to encourage cell adhesion, proliferation, and motility into the surface. Thus, understanding the nano-environmental cues will enable engineers to recreate physiologically desirable microenvironments for various biomaterial-tissue applications.

Nanopatterning also has applications for diagnostics and cell biology studies such as cell-based biosensors and cell sorting and culture systems. Integration of micro- and nanotechnologies with cell biology has yielded many new and exciting approaches. Use of microfluidics has created cells-on-a-chip or lab-on-a-chip designs where differential cell adhesion based on cell type is critical. Introduction of nanopatterned surfaces can enable a passive physical modification to these systems that would save fabrication time and complexity. Nanopatterned surfaces can be used as passive cell sorting systems that leverage the fact that different cell types respond differently to the same nanotopology. Thus a heterogeneous mixture of cells could be exposed to a nanopatterned surface and only the desired cell population would adhere. Additionally, as understanding of stem cell behavior on nanopatterned substrata grows, one could envision designing substrata that control stem cell adhesion, proliferation, and differentiation – cell behaviors that are difficult to control and that are currently modulated using complex media formulations and cell handling procedures.

Finally, studies with nanopatterned substrata can be integrated into the fields of tissue engineering and regenerative medicine. It is clear that the native in vivo environments and the chemical and physical cues to which cells are exposed are vastly different from those presented by traditional in vitro surfaces such as tissue culture plastic. Examination of signaling cascades within cells on nanopatterned surfaces can reveal how cells detect and respond to chemical and topographic signals in terms of short- and long-term cell functions, such as changes in gene expression and protein synthesis. These details of scaffold architecture and ligand presentation can improve tissue engineering technologies, and provide insight on disease processes where microenvironmental regulation is disrupted, such as in cancer. While some examples of nanopatterned tissue scaffolds exist, particularly with nanoporous materials for bone tissue engineering, the link between cell behavior in 2D and 3D is not straightforward. Cells in 3D environments adhere differently than those in 2D environments and contain different cell-ECM adhesions [19]. A recent study has even demonstrated that cells cultured on 1D patterns (lines) of adhesive ligands behave more similarly to cells in 3D matrices than to those on 2D chemically patterned islands [20]. Thus, while the link between cell behaviors in 2D and 3D is not certain, it is clear that further understanding gained using well-controlled nanopatterned surfaces can contribute to improved 3D tissue models and translational tissue engineering technologies.

Cross-References

▶ Dip-Pen Nanolithography
▶ Electron-Beam-Induced Deposition

- ▶ Electrospinning
- ▶ Nanoimprint Lithography
- ▶ Self-assembly

References

1. Bissell, M.J., Hall, H.G., Parry, G.: How does the extracellular matrix direct gene expression? J. Theor. Biol. **99**(1), 31–68 (1982)
2. Bhatia, S.N., Balis, U.J., Yarmush, M.L., Toner, M.: Probing heterotypic cell interactions: hepatocyte function in microfabricated co-cultures. J. Biomater. Sci. Polym. Ed. **9**(11), 1137–1160 (1998)
3. Nelson, C.M., Jean, R.P., Tan, J.L., Liu, W.F., Sniadecki, N.J., Spector, A.A., Chen, C.S.: Emergent patterns of growth controlled by multicellular form and mechanics. Proc. Natl. Acad. Sci. U. S. A. **102**(33), 11594–11599 (2005)
4. Chen, C.S., Mrksich, M., Huang, S., Whitesides, G. M., Ingber, D.E.: Geometric control of cell life and death. Science **276**(5317), 1425–1428 (1997)
5. Abrams, G.A., Goodman, S.L., Nealey, P.F., Franco, M., Murphy, C.J.: Nanoscale topography of the basement membrane underlying the corneal epithelium of the rhesus macaque. Cell Tissue Res. **299**(1), 39–46 (2000)
6. Norman, J.J., Desai, T.A.: Methods for fabrication of nanoscale topography for tissue engineering scaffolds. Ann. Biomed. Eng. **34**(1), 89–101 (2006)
7. Lee, K.B., Park, S.J., Mirkin, C.A., Smith, J.C., Mrksich, M.: Protein nanoarrays generated by dip-pen nanolithography. Science **295**(5560), 1702–1705 (2002)
8. Zhang, S.: Fabrication of novel biomaterials through molecular self-assembly. Nat. Biotechnol. **21**(10), 1171–1178 (2003)
9. Pham, Q.P., Sharma, U., Mikos, A.G.: Electrospinning of polymeric nanofibers for tissue engineering applications: a review. Tissue Eng. **12**(5), 1197–1211 (2006)
10. Lim, J.Y., Donahue, H.J.: Cell sensing and response to micro- and nanostructured surfaces produced by chemical and topographic patterning. Tissue Eng. **13**(8), 1879–1891 (2007)
11. Arnold, M., Cavalcanti-Adam, E.A., Glass, R., Blummel, J., Eck, W., Kantlehner, M., Kessler, H., Spatz, J.P.: Activation of integrin function by nanopatterned adhesive interfaces. Chemphyschem **5**(3), 383–388 (2004)
12. Selhuber-Unkel, C., Erdmann, T., Lopez-Garcia, M., Kessler, H., Schwarz, U.S., Spatz, J.P.: Cell adhesion strength is controlled by intermolecular spacing of adhesion receptors. Biophys. J. **98**(4), 543–551 (2010)
13. Carnell, L.S., Siochi, E.J., Holloway, N.M., Stephens, R.M., Rhim, C., Niklason, L.E., Clark, R.L.: Aligned mats from electrospun single fibers. Macromolecules **41**(14), 5345–5349 (2008)
14. Yim, E.K., Pang, S.W., Leong, K.W.: Synthetic nanostructures inducing differentiation of human mesenchymal stem cells into neuronal lineage. Exp. Cell Res. **313**(9), 1820–1829 (2007)
15. Gerecht, S., Bettinger, C.J., Zhang, Z., Borenstein, J. T., Vunjak-Novakovic, G., Langer, R.: The effect of actin disrupting agents on contact guidance of human embryonic stem cells. Biomaterials **28**(28), 4068–4077 (2007)
16. Lim, J.Y., Hansen, J.C., Siedlecki, C.A., Runt, J., Donahue, H.J.: Human foetal osteoblastic cell response to polymer-demixed nanotopographic interfaces. J. R. Soc. Interface **2**(2), 97–108 (2005)
17. Christopherson, G.T., Song, H., Mao, H.Q.: The influence of fiber diameter of electrospun substrates on neural stem cell differentiation and proliferation. Biomaterials **30**(4), 556–564 (2009)
18. Curtis, A.S., Casey, B., Gallagher, J.O., Pasqui, D., Wood, M.A., Wilkinson, C.D.: Substratum nanotopography and the adhesion of biological cells. Are symmetry or regularity of nanotopography important? Biophys. Chem. **94**(3), 275–283 (2001)
19. Cukierman, E., Pankov, R., Stevens, D.R., Yamada, K. M.: Taking cell-matrix adhesions to the third dimension. Science **294**(5547), 1708–1712 (2001)
20. Doyle, A.D., Wang, F.W., Matsumoto, K., Yamada, K. M.: One-dimensional topography underlies three-dimensional fibrillar cell migration. J. Cell Biol. **184**(4), 481–490 (2009)

Nano-patterning

▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Nanophotonic Fabrication

Takashi Yatsui
Department of Electrical Engineering and Information Systems School of Engineering, University of Tokyo, Tokyo, Japan

Definition

The word nanophotonic means the use of an optical near field. By using the optical near field, the spatial resolution of the fabricated structure is not limited by the diffraction limit of the light. In

addition, the optical near field can couple with material excitations including electron, hole, and phonon in the nanometer-scale range and result in the energy up-conversion with respect to the incident light energy. The coupled state of photon and material excitations is a quasiparticle and called a dressed photon. The true meaning of the nanophotonic fabrication requires the use of the dressed photon. Therefore, the unique fabrication process can be realized in the nanometer scale using the dressed photon, such as sub-nanometer-scale surface flattening [1], visible water splitting, and light-emitting diode with an indirect bandgap semiconductor [2].

Cross-References

► Nanofabrication

References

1. Yatsui, T.: Nanophotonic Fabrication : Self-Assembly and Deposition Techniques, Springer-Verlag, Berlin, 124 (2012)
2. Ohtsu, M.: Dressed Photons, Springer-Verlag, Berlin, 324 (2013)

Nanophotonic Structures for Biosensing

Emiliano Descrovi and Francesca Frascella
DISAT – Dipartimento di Scienza Applicata e Tecnologia, Politecnico di Torino, Torino, Italy

Synonyms

Nano-optical biosensors; Nano-optics biosensing

Definition

One-dimensional, two-dimensional, or three-dimensional structures made of suitable materials, having dimensions ranging from few nanometers to few hundreds of nanometers in size, and said structures showing a well-defined interaction with light providing an optical transduction mechanism for detecting/revealing specific biomolecules in close proximity to the structure itself.

Overview

Optical biosensors [1] constitute powerful detection and analysis tools with wide applications in the biomedical domain. They can provide parallel detection within a single device. Generally speaking, there are two detection methods that are implemented in optical biosensing: fluorescence-based detection and label-free detection. Thanks to the recent advances of technological capabilities for the fabrication of nanometer-sized structures, most of the conventional techniques can be improved, and new original techniques have become feasible.

In fluorescence-based detection, target molecules are labeled with fluorescent tags, typically organic dyes and/or chemically functionalized/decorated quantum dots; the overall intensity of fluorescence indicates the presence of the target molecules. Some quantification can also be inferred, but in general quantitative analysis is challenging due to the fluorescence signal bias, as the number of fluorophores on each molecule cannot be precisely controlled. While fluorescence-based detection is extremely sensitive, with the detection limit down to a single molecule, it suffers from laborious labeling processes that may also interfere with the biological/chemical function of the biomolecule and/or its interaction with the environment.

In label-free techniques, target molecules are not labeled or altered and are detected in their natural forms. This type of detection is relatively easy and cheap to perform and allows for kinetic measurement of molecular interaction.

Despite all these differences between fluorescence-based and label-free detection, both protocols can benefit from the use of proper nanostructures and provide complementary information regarding interactions among biomolecules. In this framework, the use of dielectric,

metallic, and metallo-dielectric structures obtained by means of a number of methods, e.g., optical, electronic, or ionic lithographic processes, physical/chemical deposition/sputtering, chemical synthesis, or self-assembling, can improve both fluorescence-based and label-free biosensing. In the following, a broad variety of structures are described with particular focus to their respective main applications, but additional original combinations are still under investigation.

Label-Free Detection

Label-free techniques include refractive index (RI) detection, optical absorption detection, and Raman spectroscopic detection. In a typical label-free biosensing application, a means should be provided in order to uniquely identify the desired target biomolecules. There are basically two strategies that are implemented: a chemical functionalization of the optically sensitive area of the photonic nanostructure and a spectroscopic measurement. In the first case, the specificity of the photonic biosensor is demanded to the chemical functionalization, therefore allowing the specific detection of desired target molecules on the unique molecular recognition mechanism embodied by the chemical functionalization. In the second case, the recognition of the target molecules is performed by detecting the corresponding spectroscopic fingerprint (either the IR absorption or the Raman emission spectra). Unfortunately, the low Raman activity of the majority of biomolecules makes them little suitable for Raman analysis. Furthermore, a detectable Raman spectrum cannot be attained by simply increasing the power of the laser excitation radiation, because irreversible damages to the biological matter can be induced. Photonic metallic nanostructures producing strong field enhancements by means of localized surface plasmons (LSPs) can overcome this limitation, thus allowing a surface-enhanced Raman scattering (SERS) or a tip-enhanced Raman scattering spectroscopic detection. Nanostructures sustaining LSP can also provide for surface-enhanced infrared absorption (SEIRA) spectroscopic detection.

Surface-enhanced Raman spectroscopy (SERS) was originally discovered in the 1970s where it was found that submonolayers of small organic molecules, when adsorbed onto the surface of silver nanoparticles (and a few other metals such as copper and gold), would exhibit greatly enhanced Raman intensities (enhancements of 10^6) [2]. These results have always held the promise for using this technique to observe very low concentrations of molecules on nanoparticles and nanostructured surfaces, but only recently has this promise started to be fulfilled in a predictable way.

Thanks to exciting advances in techniques for making nanoparticles (such as nanosphere e-beam lithography and focused ion beam), to characterizing the surfaces using electron and scanning probe microscopies, to functionalizing the surfaces of the particles using self-assembled monolayers with attached chemical receptors, and to the laser and optics technology associated with measuring the Raman spectra, a number of important applications have been reported recently. SERS has almost exclusively been associated with three metals, silver (by far the most important), gold, and copper. The generalization to other metals and other materials has been explored since the discovery of SERS, but only in the last few years have the tools been available to generate well-characterized experiments where the reported enhancements are reliable.

Light incident on the nanoparticles induces the conduction electrons in them to oscillate collectively with a resonant frequency that depends on the nanoparticles' size, shape, and composition. As a result of these LSPR modes, the nanoparticles absorb and scatter light so intensely that single nanoparticles are easily observed by the eye using dark-field (optical scattering) microscopy. The shape of the nanoparticle extinction and scattering spectra, and in particular the peak wavelength λ_{max}, depends on nanoparticle composition, size, shape, orientation, and local dielectric environment [3]. The LSPR can be tuned during fabrication by controlling these parameters with a variety of chemical syntheses and lithographic techniques. Although silver and gold are the most commonly used materials, LSPR is theoretically possible in any metal, alloy, or semiconductor with a large negative

real dielectric constant and small imaginary dielectric constant.

When molecules adsorb onto a plasmonic nanoparticle or move to within a few nanometers of its surface, the local electromagnetic fields around the nanoparticle can enhance the Raman scattering by a factor of 10^6–10^8 for an ensemble of molecules and by as much as 10^{14}–10^{15} for single molecules. Two principal enhancement mechanisms are generally thought to explain the large SERS signals. First, an electromagnetic enhancement factor arises because LSPR modes in the metal nanoparticles focus the incident light energy at the nanoparticle surface and also increase the density of states at Stokes-shifted wavelengths. The electromagnetic enhancement factor is typically 10^5–10^8 for nanoprisms, but simulations indicate that it can be as large as 10^{13} for structures such as arrays of nanoprism dimers. Second, chemical enhancement factors arise from changes in the molecular electronic state or resonant enhancements from either existing molecular excitations or newly formed charge transfer states.

Wet chemical synthesis methods have now made it possible to fabricate plasmonic nanoparticles having a variety of shapes (e.g., spheres, triangles, prisms, rods, and cubes) with controllable sizes and narrow size distributions. Furthermore, metallic–dielectric and metallic–metallic core–shell nanoparticles and mixed metallic–alloy nanoparticles with different shapes (e.g.,, nanoshells and nanoprism) have been prepared. Fabrication of this large variety of different structures makes a variety of applications possible, as the spectral position of the surface plasmon resonance depends on both the shape and the size of the nanoparticle [4].

An alternative possibility of exploiting the localized near field generated in close proximity to metallic nanostructures is to employ a metallic nano-probe or tip where LSPs are coupled. Such a structure allows a well-localized source of highly confined electromagnetic field to be raster scanned on the surface of a given sample, in such a way that a high spatial resolution Raman mapping of the surface can be performed [5]. Tip-enhanced Raman scattering (TERS) has been observed for a number of molecules adsorbed at various substrates, including single-crystalline metal surfaces, showing thereby more than a millionfold enhancement of the Raman scattering [6].

In order to excite LSPs at the apex of a metallic nano-probe, a further structuring can be performed. For example, gratings can be lithographed on the side surface of a gold cantilevered tip to ease the excitation of plasmon polaritons upon external lateral illumination of the tip. Alternatively, resonant nano-antennas can be fabricated directly on the top of the tip.

An effect quite similar to SERS occurs in the mid-infrared region: molecules on metal surfaces show infrared absorption 10–1000 times more intense than would be expected from conventional measurements without the metal. This effect is referred to as surface-enhanced infrared absorption (SEIRA) to emphasize the analogy to SERS [7, 8]. Vacuum-evaporated thin metal films are used most frequently for SEIRA experiments. Thin metal films that show strong SEIRA are not continuous but consist of metal islands smaller than the wavelength of the incident light. The average dimension of the islands is few tens of nanometers. The density, shape, and size of the islands depend on the mass thickness, the evaporation conditions, and the chemical nature of the substrate. As the mass thickness increases, the islands grow in size, contact each other, and eventually form a continuous film. Connection of the islands significantly reduces the enhancement, suggesting that the small metal islands play an important role in the enhancement. The metal islands are polarized by the incident photon field through the excitation of collective electron resonance, or localized plasmon, modes, and the dipole p induced in an island generate a local electromagnetic field stronger than the incident photon field around the island. Experimental results suggest that the collective electron resonance is excited even in the mid-infrared region and that the enhanced electromagnetic field enhances the infrared absorption of adsorbed molecules.

Unfortunately, SERS-based biosensing shows low reliable quantifiable signals, because it

usually suffers from a wide distribution of enhancement factors. Recently, metamaterials have emerged as an excellent sensing technology that complements traditional plasmonic biosensors based on metallic colloidal nanoparticles or nanostructures [9].

Metamaterials are materials engineered to show properties that you cannot find in nature. They are conventional materials as metals or polymers densely packed in a periodic way. Their engineered structure (shape, geometry, size, orientation) and arrangement give them their peculiar properties. Those that exhibit a negative index of refraction for particular wavelengths have attracted significant research. Especially, the plasmonic responses of metamaterials can be tuned to the visible near-infrared region of the electromagnetic spectrum, which is in resonance with available laser frequencies widely used in spectroscopy, enabling a strong signal enhancement for both quantitative measurement and molecular identification [10].

As previously stated beside those label-free techniques based on spectroscopic measurements, there are quantitative methods based on the measurements of refractive index changes close to the nanostructured surface of the photonic biosensor. Among these, the surface plasmon resonance (SPR) technique and the waveguide-based (or optical fiber-based) technique play a major role.

SPR is based on the coupling of propagating or stationary plasmons on silver/gold/copper thin films. The potential of surface plasmon resonance (SPR) for characterization of thin films and monitoring processes at metal interfaces was recognized in the late 1970s. Generally, an SPR optical sensor comprises an optical system, a transducing medium which interrelates the optical and (bio) chemical domains, and an electronic system supporting the optoelectronic components of the sensor and allowing data processing. The transducing medium transforms changes in the quantity of interest into changes in the refractive index, which may be determined by optically interrogating the SPR. The sensor sensitivity, stability, and resolution depend upon properties of both the optical system and the transducing medium [11].

SPR can be used in a prism coupler configuration or with a grating coupler configuration. If a metallic–dielectric interface is periodically distorted, the incident optical wave is diffracted forming a series of beams directed away from the surface at a variety of angles. If the total component of momentum along the interface of a diffracted order is equal to that of the plasmon, the optical wave may couple to the plasmon. Grating-based optical SPR sensors have been demonstrated which use the measurement of the light intensity variations at SPR and the wavelength interrogation. A high-sensitivity SPR grating-based gas sensor using silver as an SPR active metal has attained sensitivity of 1000 nm RIU^{-1} in wavelength interrogation mode; in angular interrogation mode, the system's sensitivity would be about 100 deg RIU^{-1}. Gold-based SPR grating sensors have been used for monitoring biomolecular interactions in aqueous environments, with estimated refractive index sensitivity of 30 deg RIU^{-1} and 900 % RIU^{-1} in the angular interrogation and intensity measurement modes, respectively. The grating coupler has been also used for the excitation of an SPW in an optoelectronic SPR-sensing device based on a Schottky barrier semiconductor structure. Periodic and quasiperiodic structuring of smooth metal films sustaining plasmons can be used to create Bragg mirrors (stationary plasmons) and Fabry–Pérot cavities.

Electromagnetic modes propagating at the interface between a homogeneous medium and a truncated periodic structure like a dielectric monodimensional photonic crystal (1DPC), also referred to as Bloch surface waves (BSWs), have been proposed as an alternative to surface plasmon polaritons (SPP). The possibility to excite simultaneously several different BSW modes in the same 1DPC allows for self-referencing sensing [12]. In analogy to surface plasmons on smooth thin films, a surface periodic patterning of 1DPC sustaining BSWs has been performed in order to fold BSW dispersion curves, open photonic band gaps in the dispersions, and further enhance the sensitivity of BSW upon surface perturbations.

Dielectric micro- and nanostructures can be used for biosensing based on either refractometric or absorption measurements. Such sub-wavelength structures include Bragg mirrors, gratings, microcavities, micro-rings, and optical fibers and waveguides based on photonic crystals (PCs). Provided that some of the photonic structures mentioned above produce strong field localization that can be exploited for field enhancement effects, a large portion of their application consists in measuring shifts of the resonance conditions (light incident angle and/or wavelength) upon a perturbation of the refractive index of the external medium, containing the target biomolecules [13, 14].

Bragg mirrors consist of a periodic stack of dielectric materials acting like a lossless mirror for a given wavelength band. Periodic multilayers can be fabricated, e.g., by sputtering, thermal evaporation, and chemical vapor deposition (CVD) techniques. When one or more defect layers (spacers) are introduced therein, a defect (cavity) mode is obtained. In such a simple case, the cavity is a one-dimensional Fabry–Pérot cavity that has been widely used for refractometric measurements. If the material is porous in nature, such as electrochemically etched porous silicon, target molecules can penetrate inside the porous photonic structure (pore sizes are from few nanometers to tens of nanometers), thus providing a much stronger perturbation and lowering the detection threshold. If fluorescent labels are used, one-dimensional microcavities can also be used to improve the detection of fluorescence radiation, as described in the following.

Photonic crystal surfaces can be designed to provide a wide range of functions that are used to perform biochemical and cell-based assays. Detection of the optical resonant reflections from photonic crystal surfaces enables high-sensitivity label-free biosensing, while the enhanced electromagnetic fields that occur at resonant wavelengths can be used to enhance the detection sensitivity of any surface-based fluorescence assay. Fabrication of photonic crystals from inexpensive plastic materials over large surface areas enables them to be incorporated into standard formats that include microplates, microarrays, and microfluidic channels. The periodic modulation of the refractive index within a PC, along with incorporation of intentional line or point "defects" in the PC, can be used to concentrate and direct the electromagnetic fields associated with light to produce efficient wavelength-selective reflectors, waveguides, optical circuits, beam-steering devices, and optical multiplexors. PCs can be designed to interact strongly with particular optical wavelengths through selection of their materials and the period of their modulation. At the wavelengths of "optical resonance" (also known as "guided-mode resonance"), light will couple strongly to the PC structure for a particular incident angle, resulting in electric fields inside the PC that can be many times higher than the electric field of incident radiation. These optical resonances can be observed by an external observer simply by illuminating the PC at normal incidence with a broadband of and observing a narrowband of wavelengths that are reflected back with nearly 100 % efficiency.

Two-dimensional photonic crystals used for sensing applications are typically microfabricated periodic nanostructures on a two-dimensional planar silicon on insulator (SOI) chip. Such structures can be designed to exhibit a two-dimensional photonic band gap (PBG) that can be tuned by varying the structural parameters. Photonic cavity states within this PBG are formed by systematically introducing defects into the perfect hexagonal lattice during the fabrication process. In the case of nanocavities, the modal volumes are very small, of the order of $V < 1\ \mu m^3$, and, furthermore, the strongly localized electric field in the cavity is highly sensitive to local changes of the refractive index of the environment.

The photonic crystal concept is fruitfully applied to optical fibers, in which the transverse cross section of the fiber is patterned with channels specifically arranged, wherein the refractive index contrast provided by the nanostructuration results in multiple effects such as optical field localization, multimodal guiding, etc.

By combining the technology available for waveguide fabrication and the working principle of Fourier transform-based spectrometers, a suitable micro-device based on a high-density sample of the standing-wave interferogram produced inside a single waveguide has been obtained [15].

All the abovementioned techniques can be combined for obtaining devices and structures aimed at biosensing based on one or more of the described principles. For example, optical fibers presenting metal coatings or hosting metallic nanostructures can combine guiding and plasmonic effects, and fiber grating can ease light coupling/decoupling, resonance conditions, etc.

Fluorescence-Based Detection

As already mentioned, dielectric micro- and nanostructures can provide a means to improve the detection of labeled target molecules in close contact or impregnated within the photonic nanostructure. An example of such a structure is represented by one-dimensional porous silicon microcavities [16], in which the fluorescence emitted by labeled molecules (e.g., proteins) is enhanced upon diffusion of said molecules inside the porous structures of the cavity. Additional one-dimensional, two-dimensional, and three-dimensional (e.g., artificial opals) dielectric structures can be exploited for such enhancement effects, reaching very high Q factor of the order of 10^3 or more [17–19].

In recent years there has been a growing interest in the interactions of fluorophores with metallic surfaces or particles. Near-field interactions are those occurring within a wavelength distance of an excited fluorophore. The spectral properties of fluorophores can be dramatically altered by near-field interactions with the electron clouds present in metals. These interactions modify the emission in ways not seen in classical fluorescence experiments and can be controlled by properly structuring the surface of plasmonic substrates [20, 21].

While all metal–fluorophore interactions are based on the same physics [22], the effects can be different based on the geometry of the metal structure. Three possibilities are considered, in which a fluorophore is interacting with silver particles, a smooth metal surface, and a metal surface with a regular pattern. Metal particles can be typically used to increase the fluorescence intensities. This increase occurs by a combination of enhanced fields around the metal and rapid and efficient plasmophore emission. These effects are usually called metal-enhanced fluorescence (MEF) and typically result in increased intensities and decreased lifetimes.

In the case of a fluorophore interacting with a smooth metal film, it is typically about 40 nm-thick silver or gold. In this case the fluorophore creates plasmons which radiate at a defined angle into the substrate. Typically the intensities and lifetimes are not dramatically changed. This phenomenon is called surface plasmon-coupled emission (SPCE). Here the emission spectrum is the same as the fluorophore but the polarization properties indicate that the plasmon is radiating. Finally, an exemplary fluorophore can be placed over regular patterns (nano-rings, grooves, gratings, etc.). In this case the emission at certain wavelengths is expected to show well-defined beaming into the substrate (or air), while other wavelengths are deflected from the normal. These effects are due to a combination of interactions with a smooth surface and with the sub-wavelength features (according to a diffraction phenomenon). This more general effect is called plasmon-controlled fluorescence (PCF). Plasmon-controlled fluorescence provides an opportunity to create ultra-bright probes based on complexes of metal particles with fluorophores. The basic idea is to create metal–fluorophore complexes in which the metal particle increases the brightness of the bound fluorophores. This can be accomplished by trapping the fluorophores in metal shells or by coating metal particles with fluorophores [23]. Theory has predicted that fluorophores in metal shells can be 100-fold brighter than the isolated fluorophore, even after consideration of the transfer efficiency of incident light into the shell and radiation out of the shell. Such probes may have the advantage over quantum dots because of the low toxicity of silver particles, even if the silver surface is completely exposed.

Instead of organic fluorophores such as dyes, solid-state semiconductor nanoparticles (quantum dots) can be used. The luminescent nanostructures are becoming more and more popular in labeling bioassay [24].

Cross-References

- Active Plasmonic Devices
- Biosensors
- Chemical Vapor Deposition (CVD)
- Detection of Nanoparticulate Biomarkers in Blood
- Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy
- Microfluidic Whole-Cell Biosensor
- Nanostructured Materials for Sensing
- Nanostructures for Photonics
- Optical Properties of Metal Nanoparticles
- Optical Techniques for Nanostructure Characterization
- Photonic Crystal Nanobeam Cavities
- Polymer Coatings
- Scanning Near-Field Optical Microscopy
- Self-Assembled Monolayers
- Theory of Optical Metamaterials

References

1. Fan, X., White, I.M., Shopova, S.I., Zhu, H., Suter, J.D., Sun, Y.: Sensitive optical biosensors for unlabeled targets: A review. Anal. Chim. Acta **620**, 8–26 (2008)
2. Kneipp, K., Moskovits, M., Kneipp, H.: Surface Enhanced Raman Scattering. Springer, Heidelberg (2006)
3. Anker, J.N., Hall, W.P., Lyandres, O., Shah, N.C., Zhao, J., Van Duyne, R.P.: Biosensing with plasmonic nanosensors. Nat. Mater. **7**, 442–453 (2008)
4. Lal, S., Link, S., Halas, N.J.: Nano-optics from sensing to waveguiding. Nano Lett. **1**, 641–648 (2007)
5. Steidtner, J., Pettinger, B.: Tip-enhanced raman spectroscopy and microscopy on single dye molecules with 15 nm resolution. Phys. Rev. Lett. **100**, 236101 (2008)
6. Sonntag, M.D., Pozzi, E.A., Jiang, N., Hersam, M.C., Van Duyne, R.P.: Recent advances in tip-enhanced raman spectroscopy. J. Phys. Chem. Lett. **5**, 3125–3130 (2014)
7. Osawa, M.: Surface-enhanced infrared absorption. In: Kawata, S. (ed.) Near-Field Optics and Surface Plasmon Polaritons. Springer, Heidelberg (2001)
8. Le Ru, E.C., Etchegoin, P.G.: Surface-Enhanced Raman Spectroscopy and Related Plasmonic Effects. Elsevier, Amsterdam (2009)
9. Yoa, K., Liu, Y.: Plasmonic metamaterials. Nanotechnol Rev **3**(2), 177–210 (2014)
10. Zhang, C., Wen, X., Dodson, S.L., Dao, N.T., Wong, L.M., Wang, S., Li, S., Tu, A., Xiong, Q.: Metamaterials-based label-free nanosensor for conformation and affinity biosensing. ACS Nano **7**(10), 7583–7591 (2013)
11. Homola, J.: Surface Plasmon Resonance Based Sensors. Springer, Berlin (2006)
12. Konopsky, V.N., Alieva, E.V.: A biosensor based on photonic crystal surface waves with an independent registration of the liquid refractive index. Biosens. Bioelectron. **25**, 1212–1216 (2010)
13. Zourob, M., Lakhatakia, A.: Optical Guided-Wave Chemical and Biosensor II. Springer, Heidelberg (2010)
14. Wang, X.D., Wolfbeis, O.: Fiber optic chemical sensors and biosensors. Anal. Chem. **85**(2), 487–508 (2013)
15. le Coarer, E., et al.: Wavelength-scale stationary-wave integrated fourier-transform spectrometry. Nat. Photonics. **1**, 473–478 (2007)
16. Sciacca, B., Frascella, F., Venturello, A., Rivolo, P., Descrovi, E., Giorgis, F., Geobaldo, F.: Doubly resonant porous silicon microcavities for enhanced detection of fluorescent organic molecules. Sens. Actuators. B. **137**, 467–470 (2009)
17. Dorfner, D., Zabel, T., Hürlimann, T., Hauke, N., Frandsen, L., Rant, U., Abstreiter, G., Finley, J.: Photonic crystal nanostructures for optical biosensing applications. Biosens. Bioelectron. **24**, 3688–3692 (2009)
18. Frascella, F., Ricciardi, S., Pasquardini, L., Potrich, C., Angelini, A., Chiadò, A., Pederzolli, C., De Leo, N., Rivolo, P., Pirri, C.F., Descrovi, E.: Enhanced fluorescence detection of miRNA-16 on a photonic crystal. Analyst **140**, 5459–5463 (2015)
19. Shafiee, H., Lidstone, E.A., Jahangir, M., Inci, F., Hanhauser, E., Henrich, T.J., Kuritzkes, D.R., Cunningham, B.T., Demirci, U.: Nanostructured optical photonic crystal biosensor for HIV viral load measurement. Sci. Rep. **4**, 4116 (2014). 1–7
20. Lakowicz, J.R., et al.: Plasmon-controlled fluorescence: a new paradigm in fluorescence spectroscopy. Analyst **133**, 1308–1346 (2008)
21. Ming, T., Chen, H., Jiang, R., Li, Q., Wang, J.: Plasmon-controlled fluorescence: Beyond the intensity enhancement. J. Phys. Chem. Lett. **3**(2), 191–202 (2012)
22. Lakowicz, J.R.: Principles of Fluorescence Spectroscopy, 3rd edn. Springer, New York (2006)
23. Burns, A., Ow, H., Wiesner, U.: Fluorescent core–shell silica nanoparticles: towards "Lab on a Particle" architectures for nanobiotechnology. Chem. Soc. Rev. **35**, 1028–1042 (2006)
24. Viste, P., Plain, J., Jaffiol, R., Vial, A., Adam, P.M., Royer, P.: Enhancement and quenching regimes in metal–semiconductor hybrid optical nanosources. ACS Nano **4**, 759–764 (2010)

Nano-piezoelectricity

▶ Piezoelectric Effect at Nanoscale

Nano-pillars

▶ Nanosized Nanocrystalline and Nanotwinned Metals

Nanoplasmonic Enhancement

▶ Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics

Nanoplasmonic Sensing

▶ Refractometric Sensing Using Plasmonic Nanoparticles

Nanoplasmonics Involved in Localized Photopolymerization

Kyoko Masui[1] and Satoru Shoji[2]
[1]Applied Physics of Engineering, Osaka University, Osaka, Japan
[2]Department of Engineering Science, The University of Electro- Communications, Chofu, Tokyo, Japan

Synonyms

Plasmon-Assisted Two-Photon Polymerization; Photopolymerization Induced by Localized Surface Plasmon Resonance

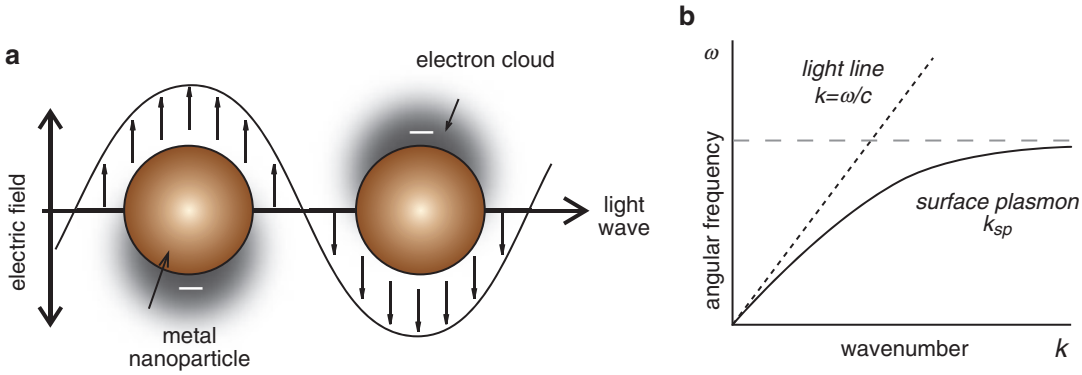
Definition

Nanoplasmonics is an optical phenomenon occurring at the nanoscale vicinity confined to the surface of metal nanoparticles, and the electromagnetic field is generated only at their surface by the interaction with incident light. This effect induces a localized photopolymerization only at the restricted field near the nanostructures.

Nanoplasmonics on Metal Nanoparticles

Localized surface plasmon resonance (LSPR) is an interaction between free electrons of metal nanoparticles and incident light confined with three dimensions [1–5]. When light is incident to a metal nanoparticle with wavelength matched with the electron frequency, the collective oscillations of free electrons occur, and a localized electromagnetic field is formed at the surface. The intensity of the local electromagnetic field around the surface of nanoparticles is strongly enhanced, compared with the field of the incident light. Figure 1a shows a schematic image of LSPR excitation that forms an electron dipole on the metal nanoparticles.

The peak of the LSPR mode depends on material, size, shape, and aspect ratio (the ratio of long-to-short length) of nanoparticles, as well as dielectric properties of their surrounding environment. Spherical nanoparticles have only one LSPR mode. For example, gold spherical particles with diameter of less than 100 nm in aqueous solution show a strong extinction peak at 500–600 nm. On the other hand, nonspherical nanoparticles exhibit anisotropic optical and electronic responses. In the case of rodlike particles, namely, nanorods, two distinct peaks appear. One is the transverse LSPR mode excited by the light polarized along the short axis of nanorods. The other is the longitudinal LSPR mode excited along the long axis. The resonant wavelength of the longitudinal mode shifts to shorter/longer wavelength decreasing/increasing the aspect ratio of the nanorod. The resonant wavelength of the longitudinal mode is tuned across a broad spectral range, covering the



Nanoplasmonics Involved in Localized Photopolymerization, Fig. 1 (**a**) Schematic image of localized surface plasmon resonance that forms electron dipole on metal nanoparticles, (**b**) the dispersion relation of surface plasmon polariton at the interface between a metal and a dielectric medium

visible and near-infrared regions according to their aspect ratio.

Metal colloidal nanoparticles have been traditionally utilized as coloring materials, such as stained glass. Michael Faraday found that the color of ruby stained glass was caused by colloidal gold nanoparticles. In 1857, he had a Bakerian Lecture of the Royal Society entitled "Experimental Relations of Gold (and other Metals) to Light" [6]. He showed a ruby aqueous solution synthesized with gold salt and reducing agent. In his day, it was impossible to observe such small particles. In the twentieth century, by development of the electron microscope, it was revealed that the ruby red solution made by Faraday's method contained particles with average diameters of $3 \sim 30$ nm.

Dispersion Relation of LSPR

LSPR is three-dimensionally confined surface plasmon polariton (SPP), which propagates at the interface between metal and dielectric surface. The propagation mode of SPP can be explained with Maxwell's equation. Figure 1b shows the dispersion relation of SPP at the interface between a metal and a dielectric medium represented by the following equation:

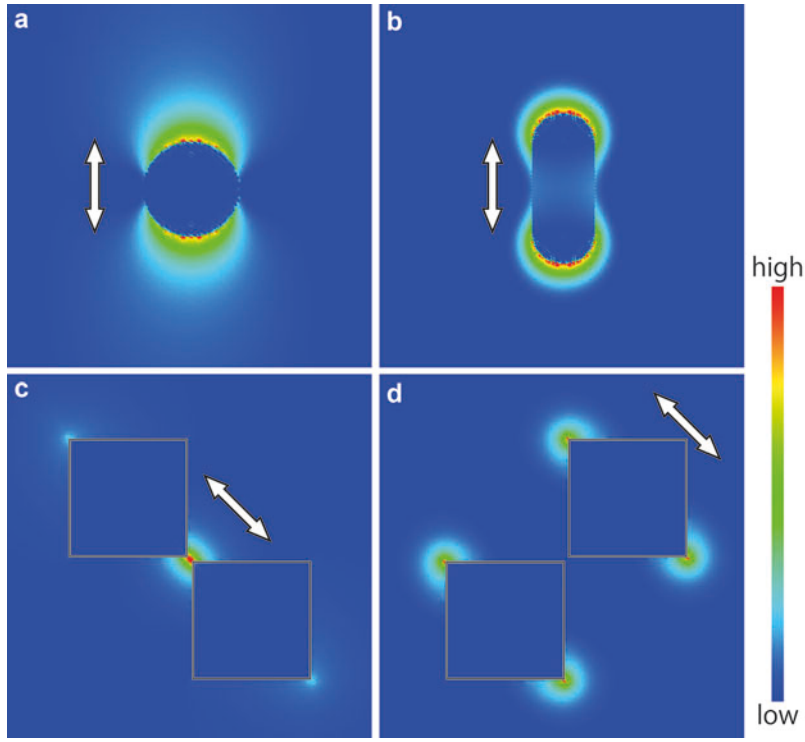
$$K_{sp} = \frac{\omega}{c}\sqrt{\frac{\varepsilon_m \cdot \varepsilon_d}{\varepsilon_m \cdot \varepsilon_d}}$$

where K_{sp} is the wave vector of surface plasmon, ω is the angular frequency, c is speed of light in vacuum, and ε_m and ε_d are the dielectric constants of the metal and the surrounding dielectric medium, respectively.

When wave vector of light at frequency ω propagates through the dielectric medium ε_d, its dispersion curve is straight. The wavelength of SPP is shorter than the wavelength of light line in vacuum. The curve of the dispersion relation of SPP does not interact with the straight light line. However, the evanescent light with transverse magnetic polarization can react with the electric field on the metal surface. Coupling with evanescent wave, the dispersion relation of light inclines to the new line and makes a cross point with the dispersion curve of SPP. Therefore, the electromagnetic field on the metal surface is enhanced, whose intensity is larger than the incident light.

Theoretical Calculation of LSPR

The intensity distribution of enhanced field has been theoretically investigated with calculation methods, such as the finite difference time domain method (FDTD) and discrete dipole approximation. Extension of Mie scattering theory is utilized to analyze the mode of plasmon resonance. FDTD is a calculation method based on Maxwell's equation. Figure 2 shows the field intensity patterns by using calculation software of Poynting for Optics (Fujitsu Limited). Gold nanoparticle (Fig. 2a), nanorod (Fig. 2b), and nanoblocks with a nanogap (Fig. 2c, d) were set as model structures. A nanoparticle had a diameter of 50 nm. A gold nanorod



Nanoplasmonics Involved in Localized Photopolymerization, Fig. 2 FDTD calculation result of the intensity distribution of local enhanced field on (**a**) gold nanoparticle, (**b**) gold nanorod, (**c**, **d**) gold nanoblocks with nano gap

was designed as a cylinder capped with a half sphere at both ends. The total length of the nanorod was 60 nm, and the diameter was 25 nm. Each size of gold nanoblocks was X = Y = 80 nm and Z = 40 nm and was set with a gap distance of 4 nm. All components were set in vacuum. The dielectric function of gold was represented with a combination of the Drude and Lorentz model. Each incident light was set to the resonant wavelength of each structure polarized in the arrow direction. For example, in the case of nanorod, the polarization direction was set parallel to the long axis. The entire field was discretized with a spatial mesh with variable size, which reached down to 0.5 nm in the vicinity of the structure. Far-field and near-field intensity patterns were monitored on a range of x-y planes cutting through the structures. As a result, the enhanced field was localized at both corners of nanoparticles along the incident polarization direction. At the gap between nanoblocks, a strong electromagnetic field was excited as a near-field enhancement on their surface. The strong enhancement was up to 3 orders larger than the incident light. When the incident light was polarized perpendicular to the axis of the pair, the enhanced field is concentrated on the left and right tips of the nanoblocks.

Visualization of Local Enhanced Field

Experimental observation of the intensity distribution caused by LSPR on the nanostructures as mentioned above gives us valuable information to utilize the optical field enhancement in optical applications. There are roughly two types of direct observation methods to visualize the optical field and to understand the basic features. One is to use a microscopic technique [1, 7] and the other to induce a photochemical reaction at the local enhanced field.

Local Enhanced Field Observation by Microscopic Technique

When the diameter of nanoparticles is much smaller than the wavelength of incident light, observation methods beyond the diffraction limit

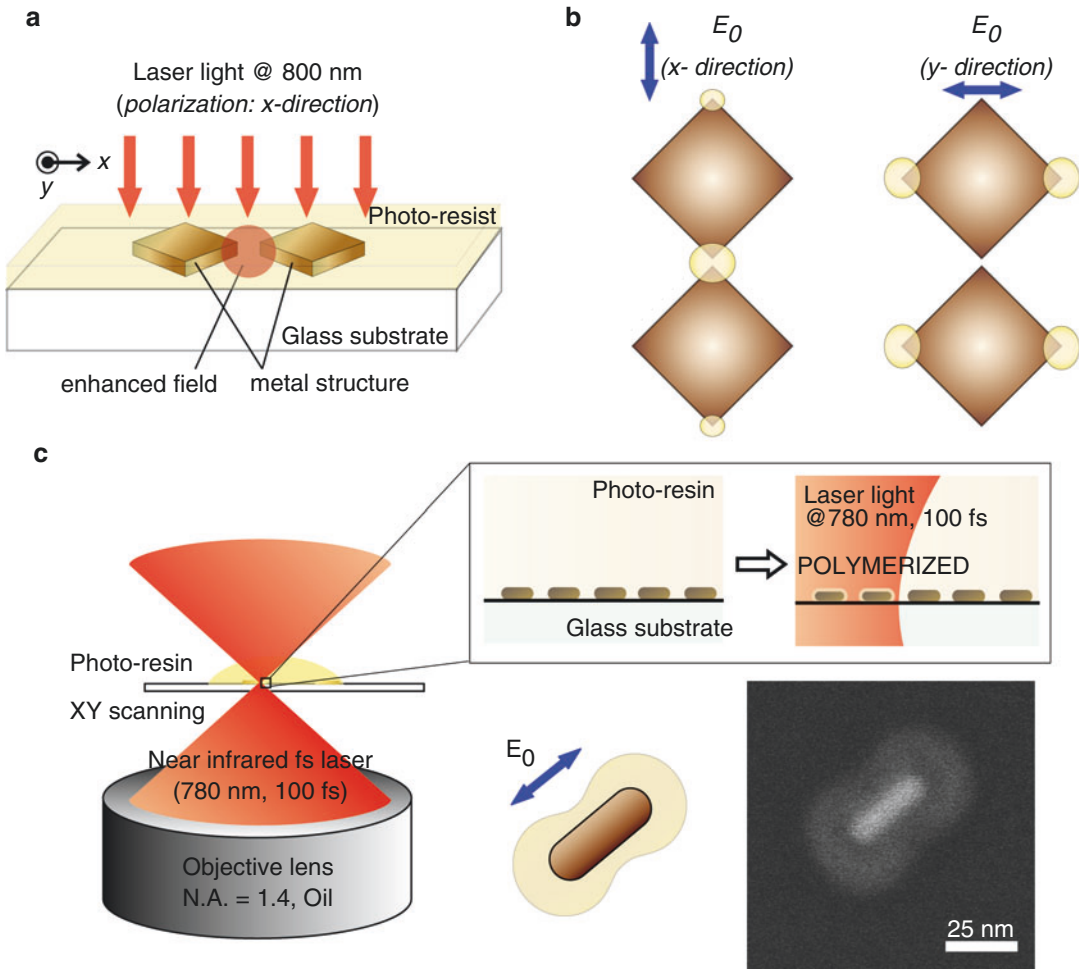
of light are needed to visualize the optically enhanced field on the small nanostructures. The spatial resolution of a conventional optical microscope is restricted to the wavelength of incident light, because of the diffraction limit of light. For observation with high spatial resolution, several types of optical microscopic techniques have been studied, that is classified with super-resolution far-field optical microscopy with nonlinear optical responses, and aperture-type and scattering-type near-field optical microscopy. Electron microscopy with cathodoluminescent detection and with electron energy loss spectroscopy has also been used for observation of LSPR modes [1, 7]. The spatial resolution is similar to that of optical microscopy.

These microscopic techniques show the result of localized electromagnetic density on metal nanostructures. The resulting images correspond to the wave functions of the excited state on the nanostructures. The wave functions are indicated as amplitude of the collective oscillation of plasmon as a function of the position.

Photochemical Reaction Induced by Local Enhanced Field

Photochemical reaction is utilized for visualization of the intensity distribution of local enhanced field. For example, photopolymerization induced by a local enhanced field has been demonstrated with several types of nanostructures, such as spherical nanoparticles, nanorods, nanobowties, and nanoblocks with a nanogap. These kinds of metallic nanostructures are produced by lithography technique. Various types of lithography techniques have been proposed to produce arbitrary nanostructures with the size, shape, configuration, and gap size between nanostructures, not only the polymeric structures but also the metallic structures [8]. The nanostructures are fabricated on a substrate and are covered with negative or positive photoresist. The absorption peak appears in the ultraviolet (UV) region, because of the photoinitiator contained in the photoresist. The wavelength of incident light is selected as a resonant wavelength of nanostructures in the visible or near-infrared regions for generating one-photon or multiphoton absorption of photoresist. The intensity of light is set weaker than that for usual two-photon polymerization. When the light is incident on the nanostructures, LSPR occurs on their surface. Only at the enhanced field with high intensity, cross-linking reaction of photoresist is induced via absorption of photoinitiator in photoresist and photopolymerization occurs. Figure 3a shows the schematics of predictive photopolymerization patterns on nanoblocks with a nanogap. The nanoblocks are set on a substrate and covered with photoresist, and then, polarized light at a wavelength of 800 nm is incident to them. When the light is polarized along x-direction and the absorption peak of photo resist exists around 400 nm, the enhanced field is generated and is localized at the nanogap between nanoblocks. Therefore, photopolymerization occurs at the local position. The enhanced field corresponds to the FDTD calculation result shown in Fig. 2c. Depending on the polarization direction of light, the position of the enhanced field changes. The arrows shown in Fig. 3b indicate the polarization direction of incident light. If the light is incident to the pair of nanoblocks parallel to the long axis of the pair, photopolymerization occurs at the gap. When the incident polarization is perpendicular to the gap, the enhanced field is excited at each corner of nanoblocks, not at the nanogap.

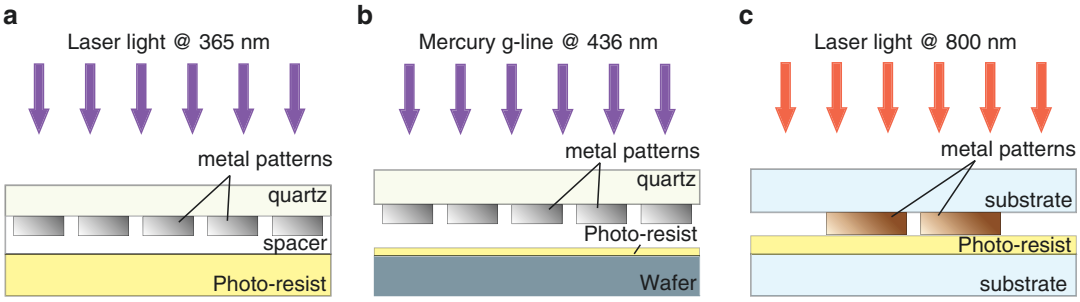
Not only the lithography techniques but also the chemical synthesis methods are useful for producing nanoparticles, which have been established as a simple mass production method of uniform metal nanoparticles [9]. A number of methods are reported to synthesize nanoparticles with various sizes and shapes. Figure 3c shows the schematic image and experimental result of photopolymerization at the local enhanced field at a single gold nanoparticle [10]. Controlled by additive chemicals during synthesis, the resonant wavelength of nanorod was tuned to fit the incident laser light. Synthesized nanoparticles in aqueous solution are dried on a substrate and then are covered with UV photosensitive resin composed of methyl methacrylate (MMA) monomer. A femtosecond laser with a wavelength of 780 nm was irradiated onto the nanorod, which was polarized parallel to the long axis, and then



Nanoplasmonics Involved in Localized Photopolymerization, Fig. 3 Schematic images of (**a**) Schematic image of setup for photo polymerization by local enhanced field at the nanogap induced by LSPR of metal nanoblocks, (**b**) predictive photopolymerization part, (**c**) experimental condition and the experimental result of LSPR-induced photopolymerization at the gold nanorods

the uncured resin was washed away. By scanning electron microscope (SEM) observation, photopolymerization was obtained only at both edges of the nanorod. The thickness of the polymerized layer covered the nanorod with length of 62 nm and diameter of 22 nm were characterized as 21 and 14 nm at the end and the middle of the gold nanorod, respectively. The shape and thickness of the polymerized part was in proportion to the intensity of laser light. The shape was changed from the "dumbbell" to the "ellipsoid" with increasing laser intensity. The thickness of obtained photo-polymerized layer was less than 30 nm from the surface. This result indicates that the effective intensity distribution of enhanced field of plasmonic resonance is less than 30 nm. It is clear that the shape of the polymerized area provides a snapshot of the near-field intensity distribution with high spatial resolution.

Furthermore, the local enhanced field induces a local movement of the polymer film caused by a photochromic reaction via two-photon absorption [11]. Polymer film mixed with photochromic dye molecule, such as azo-benzene and diarylethene, is put on the metal nanostructures. Since the dye molecules give rise to trans-cis isomerization after



Nanoplasmonics Involved in Localized Photopolymerization, Fig. 4 Schematic images of (**a**) plasmonic nanolithography with metal photomask, (**b**) surface plasmon resonant interference lithography (SPRINT), (**c**) nanogap-assisted surface plasmon nanolithography

photo-excitation, the dye along the local enhanced field changes its orientation in the polymer matrix, resulting in forming the surface relief of polymer film. In the case of nanoparticles, the local enhanced field is generated at their corner or edge. Two holes are observed on the polymer film close to the particles along the direction of the polarization of incident light.

Plasmonic Nanolithography

Three-dimensional (3D) micro/nanofabrication techniques have been established to produce micro/nanoscale devices. By reducing the wavelength of incident light and utilizing a nonlinear response of photo resists, the fabrication resolution has been improved beyond the diffraction limit of light. Two-photo polymerization technique is also well known as a fabrication method of 3D arbitrary micro/nanostructures [8, 12–14]. When a femto second near infrared laser light with high intensity is focused in UV-photosensitive resin, the intensity of photons becomes high only at the focus position and the opportunity of two-photon absorption of the resin becomes high. Then, two-photon polymerization occurs. Scanning the focus spot of laser light can produce arbitrary micro/nano structures. The finest structure by using this technique is reported as less than 20 nm.

On the other hand, plasmonic nanolithography is one high-resolution lithography technique with SPP, for achieving fabrication resolution beyond the diffraction limit of light. SPP generates a strong enhancement of electromagnetic field at the surface of metal nanostructures.

Several methods have been reported, such as plasmonic nanolithography with photomask [11, 15], surface plasmon interference nanolithography [11, 15], and nanogap-assisted surface plasmon nanolithography [11, 16, 17]. Furthermore, there is a method that a local photopolymerization excited at a surface of metal nanoparticles by enhanced field is used as a glue to fabricate 3D structure of metal/polymer composite. [18]

Plasmonic Nanolithography with Photomask

Experimental possibility of plasmonic nanolithography has been studied with a photomask that has metal patterns on quartz substrate, as shown in Fig. 4a [11, 15]. Aluminum or titanium exhibits strong plasmon resonances in UV region. The light at a wavelength of 365 nm is blocked at the aluminum or titanium patterns and transmitted through the photomask. 2D periodic structures on aluminum or titanium film are fabricated by using Focused Ion Beam lithography and Electron Beam Lithography. Then, the film was sandwiched by quartz substrate and PMMA spacer to work as a photomask. UV light was irradiated through an array of 40 nm holes with 170 nm periods on the mask and the pattern was transferred to a negative photo resist of SU-8. The size of the resulting structure was 90 nm with the same period. However, small holes on the aluminum mask led to a strong attenuation of light

transmission. Silver exhibits plasmon resonance with UV light, so serves as a material for photomask As a result, 60 nm resolution patterns were obtained. The size was less than 1/6 of the wavelength of incident light. The features of the obtained patterns were not perfect, but almost equal to those of the photomask.

Surface Plasmon Resonant Interference Lithography

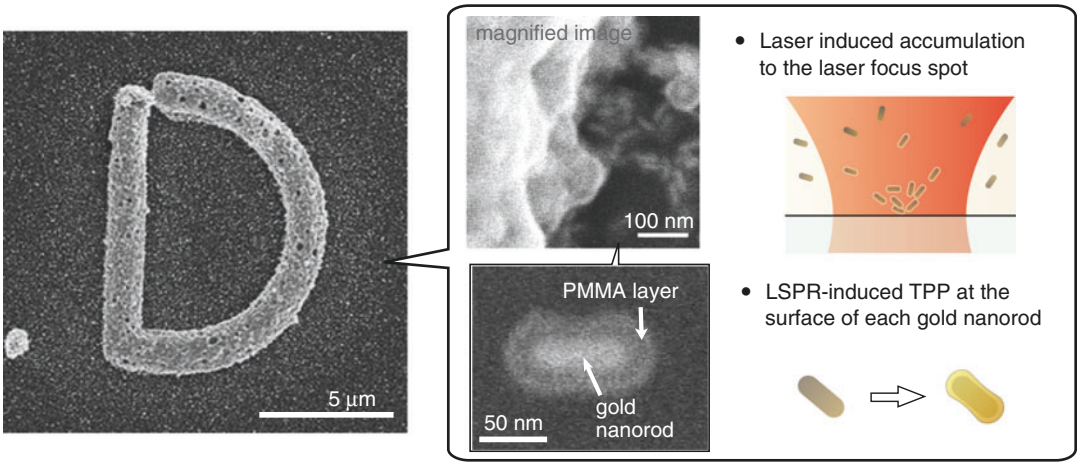
Surface plasmon resonant interference lithography (SPRINT) is a method to fabricate the periodic grating patterns with sub-100 nm lines with high efficiency using photo-resin and UV or visible light [11, 15]. The point of SPRINT is to use interference of SPP for producing nanoscale spatial resolution and to suppress the exposure time by using the enhancement of SPP. Figure 4b shows the schematic setup. When the light (Mercury g-line at the wavelength of 436 nm) is illuminated from the top, the light transmits through the mask via SPP. The SPP propagates and interferes with each other and then, induces the strong field with spatial distribution in nanoscale on the metal surface. The resulting enhanced field is irradiated to the photo resist directly below the mask. As an experiment, a silver mask was fabricated on a 2-mm thick quartz plate with a periodicity of 300 nm and the opening of 1/5 of the periodicity. The resulting structure was 50 nm line width of grating pattern with a period of 300 nm. The other experiment using silver and magnesium fluoride multilayer achieved a feature size of 20 nm of periodic patterns under the wavelength of 193 nm. Changing the pattern of photomask with various size, shape and layout easily controls the fabricated patterns.

Nanogap-Assisted Surface Plasmon Nanolithography

The hot spot of metal nanostructures with nanogap is useful to fabricate a nanometer-size polymer structure by transferring it to photo resist [11, 16, 17]. This technique is similar with the method to visualize the local enhanced field on nanoparticles or structures. Figure 4c indicates the geometry of a pair of gold nanoblocks separated by a nano gap and the setup for nanogap-assisted surface plasmon nanolithography with positive photo resist. An array of gold nanoblocks (80, 80, 35 nm) with a gap width of 4 nm was prepared on a glass substrate as a photomask. A pair of nanoblocks resonated to the light of wavelength around 800 nm, which polarization direction was set parallel to the long axis of the pair. The photomask was put on a thin layer of positive photo resist coated on substrate. This photo resist has absorption around 500 nm. A femto second laser light with a wavelength of 800 nm is illuminated through the photomask to the photo resist. Through the two-photon absorption process, the photo resist was polymerized only at the nano gap position of the photomask. The calculated result of the enhanced field was 5 orders larger than the incident light. As an experimental result, a polymerized spot with 5 nm in a diameter was obtained with a periodicity of 400 nm. This corresponded to the hot spot of the nano gap between nanostructures. This method can be utilized as a lithography technique to produce small nano patterns with the diameter in sub-nanometer scale.

3D Micro/Nanofabrication of Metal Nanoparticle Aggregated Structures Based on LSPR-Induced Photo Polymerization

LSPR at the metal nanoparticles has potential to produce local photopolymerization at its surface. Not only to the nanoparticles fixed on a substrate, LSPR on the nanoparticles floating in photosensitive resin is also excited by light irradiation of their resonant wavelength [18]. By means of LSPR at meal nanoparticles, 3D micro/nanofabrication technique was proposed to fabricate metal nanoparticles/polymer composite with arbitrary structures. The key of this technique is LSPR induced two-photon polymerization on metal nanoparticles and optical gradient force to the nanoparticles into the focus position. The resulting composite structures are formed with aggregation of metal nanoparticles glued by polymer. Figure 5 shows the experimental result and the schematic image of experiment. Gold nanorods was chemically synthesized, which absorption peak was around 520 nm and 780 nm corresponding to their anisotropic shape.



Nanoplasmonics Involved in Localized Photopolymerization, Fig. 5 Schematic image and the resulting SEM image of 3D micro/nanofabrication technique for making metal nanoparticles aggregated structures base on LSPR-induced photopolymerization

The nanorods were mono-dispersed in UV photosensitive resin composed of MMA monomer. Illuminated by femto second laser light at a wavelength of 780 nm, LSPR on the nanorods was excited and photo-polymerization was observed locally at both their edges. Because the absorption peak of photosensitive resin appears at the wavelength shorter than 400 nm, this polymerization is caused by two-photon absorption of photosensitive resin and two-photon polymerization was induced only at the enhanced field. Simultaneously, the floating nanorods accumulated at the focus spot by optical gradient force generated by tightly focusing the laser light. Gold nanorods and the polymerized layer at their surface worked as building blocks and glue, respectively, to make aggregation of nanorods. SEM image shows the resulting structure by scanning the laser spot. Each nanorods were covered by polymer layer and stacked together. This method can create metal nanorod aggregates with desired micro/nanostructures.

Summary

Nanoplasmonics is an optical phenomenon on metal nanostructures. LSPR is an interaction with metal nanoparticles and incident light, which causes a collective oscillation of free electrons at the surface of metal nanoparticles. It is known that the intensity of LSPR enhanced field is very high only at the interface between metal and dielectric materials and localized at the edge of nanoparticles or nanostructures. The visualization of the enhanced field is an important role for development of plasmonic nanolithography. Not only the observation by using optical microscopy but also the photochemical reactions induced by the enhanced field on metal nanostructures have been used for visualization, such as the local photopolymerization at the surface of metal nanoparticles, and the movement of polymer films. Furthermore, this restricted enhanced field can produce nanostructures with a size beyond the diffraction limit of light. Nanopatterns of metal nanostructures can be reflected to the photoresist with photo-polymerization induced by local enhanced field of LSPR. The local photopolymerization occurs not only at the surface of nanostructures fixed on substrates, but also that of floating nanoparticles in photosensitive resin. A fabrication technique with floating meral nanoparticles with LSPR induced photo polymerization was proposed for fabrication of 3D metal nanoparticle aggregated structures in micro/nano scale. By the illumination of incident light, floating nanoparticles are coated by photo-polymerized layer induced by

LSPR of nanoparticles. Occurring simultaneously with their accumulation by optical gradient force, the polymer layer works as a glue to fix them together. These techniques would pave the way for novel micro/nanofabrication methods for plasmonic materials and devices.

Cross-References

- ▶ Finite-Difference Time-Domain Technique
- ▶ Local Surface Plasmon Resonance (LSPR)
- ▶ Nanophotonic Fabrication

References

1. Masuhara, H., Kawata, S.: Nanoplasmonics, From Fundamentals to Applications. Elsevier, Amsterdam (2006)
2. Chon, J.W.M., Iniewski, K.: Nanoplasmonics: Advanced Device Applications. CRC Press, London (2013)
3. Kawata, S., Ohtsu, M., Irie, M.: Near-Field Optics and Surface Plasmon Polaritons. Springer, New York (2001)
4. Klimov, V.: Nanoplasmonics. CRC Press, London (2014)
5. Alexander, M.S.: Plasmonics: Fundamentals and Applications. Springer, Heidelberg (2007)
6. Edwards, P.P., Thomas, J.M.: Gold in a metallic divided state – from Faraday to present-day nanoscience. Angew. Chem. Int. Ed. **46**, 5480–5486 (2007)
7. Okamoto, H., Imura, K.: Visualizing the optical field structures in metal nanostructures. J. Phys. Chem. Lett. **4**, 2230–2241 (2013)
8. Shoji, S., Masui, K.: Nano-/Microfabrication. In: Kobayashi, S., Müllen, K. (eds.) Encyclopedia of Plymeric Nanomaterials. Springer, Heidelberg (2013)
9. Huang, X., Neretina, S., El-Sayed, M.A.: Gold Nanorods: from synthesis and properties to biological and biomedical applications. Adv. Mater. **21**, 4880–4910 (2009)
10. Masui, K., Shoji, S., Jin, F., Duan, X.-M., Kawata, S.: Plasmonic resonance enhancement of single gold nanorod in two-photon photopolymerization for fabrication of polymer/metal nanocomposites. Appl. Phys. A **106**, 773–778 (2012)
11. Ueno, K., Misawa, H.: Surface plasmon-enhanced photochemical reactions. J. Photochem. Photobiol. C Photochem. Rev. **15**, 31–52 (2013)
12. Kawata, S., Sun, H.-B., Tanaka, T., Takada, K.: Finer features for functional microdevices. Nature **412**, 697–698 (2001)
13. Shoji, S., Kawata, S.: Two-photon lithography. In: Cabrini, S., Kawata, S. (eds.) Nanofabrication Handbook, pp. 229–250. Taylor & Francis, London (2012)
14. Bachelot, R., Consonni, M.: Lithography and manipulation based on the optical properties of metal nanostructures. In: Landis, S. (ed.) Nano- Lithography, pp. 207–230. Wiley, New York (2013)
15. Liu, H., Teng, J.: Plasmonic nanolothography: towards next generation nanopatterning. J. Mol. Eng. Mater. **1**(1), 1250005 1–19 (2013)
16. Ueno, K., Takabatake, S., Nishijima, Y., Mizeikis, V., Yokata, Y., Misawa, H.: Nanogap-assisted surface plasmon nanolithography. J. Phys. Chem. Lett. **1**, 657–662 (2010)
17. Gao, S., Ueno, K., Misawa, H.: Plasmonic antenna effects on photochemical reactions. Acc. Chem. Res. **44**(4), 251–260 (2011)
18. Masui, K., Shoji, S., Asaba, K., Rodgers, T.C., Jin, F., Duan, X.-M., Kawata, S.: Laser fabrication of Au nanorod aggregates microstructures assisted by two-photon polymerization. Opt. Express **19**(23), 22786–22796 (2011)

Nanopolaritonics

Daniel Neuhauser[1], Christopher Arntsen[1] and Kenneth A. Lopata[2]

[1]Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA, USA

[2]William R. Wiley Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, WA, USA

Synonyms

FDTD; Maxwell–Bloch simulations or equations or studies; Maxwell–Schrödinger simulations (or equations); Molecular plasmonics; Near-field molecular optics

Definition

Molecular nanopolaritonics is the study of nanoscale control of near-field energy transport via combined plasmon–molecule excitations.

Introduction

Absorption and emission of light (from near-IR to UV) on the nanoscale is qualitatively associated with one of two types of entities: excitons and plasmons. Excitons are associated with transition of electrons between occupied and unoccupied orbitals, i.e., a particle-hole excitation. Plasmons, on the other hand, are classical-like collective excitation of electrons in metals that are analogous to tidal waves. Specifically, when an electric field excites a plasmon, all the valence electrons in the metal shift, in tandem, creating a surface dipole; the remaining background exerts a restoring force and the resulting oscillation is in the IR to UV range, depending on geometry.

The coupling of an exciton and a plasmon is known as a polariton. Polaritons are usually associated with surface excitons. They have been applied in various fields, and at low temperatures even exhibit Bose–Einstein condensates [1]. Here, the combination of excitons and plasmons on the nanoscale is discussed, where the plasmon is part of a relatively small metal structure and even more importantly, the exciton is associated with a molecule or a cluster. Such a combination is labeled as nanopolariton [2], and its properties and approaches for simulation are discussed here. The greatest emphasis here is on the manipulation of transport of nanopolaritons by varying the molecular properties, as detailed below.

Broadly speaking plasmon–molecule interactions come in two forms: the influence of plasmonic fields on molecules, and the influence of molecules on plasmonic fields. The first, which is the typical focus of plasmon–molecule studies, is where plasmons focus light into molecules driving them into highly excited states, such as in surface-enhanced Raman spectroscopy (see ▶ Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy). SERS relies on the fact that classically, an electric field near a metal surface is very strong, especially near the plasmon resonance. In SERS, the massive electric fields due to plasmons on a metal lead to huge enhancements in Raman scattering, allowing for the detection of single molecules [3]. Other examples of plasmons acting on molecules include plasmon-enhanced fluorescence, near-field optical microscopy, and enhancements in FRET [4].

The second direction, where single molecules have a significant effect on plasmons, is more subtle and less understood. Molecules have a much smaller transition dipole moment than metal nanostructures (as they have few electrons, versus hundreds to million and more electrons in metal structures) and therefore molecules will generally have a negligible effect on overall light absorption of a polaritonic system, except possibly at very low temperatures. However there is one place where plasmonic interactions are weak: in the transport of plasmons from one structure to another. As this transport involves comparatively weaker coupling between plasmon modes, under certain circumstances it can be influenced by nearby molecules. While this article focuses primarily on molecule-mediated plasmon transport, linked DNA molecules have been shown to alter the dielectric properties of arrays of metal nanoparticles [5].

The length scales relevant for nanopolaritonics are significantly smaller than even UV wavelengths. At such scales, the electromagnetic fields are essentially "near fields," i.e., the field due to a dipole is closer to its electrostatic form, so it is strong and decays rapidly with distance; the far-field-radiated part, proportional to the acceleration of the dipole, is much weaker over these scales (the ratio between the far and near fields for a distance scale R is approximately R^2/λ^2, where λ is the wavelength). The simulation of the near-field combined plasmon–molecule excitations (polaritons) involve a complicated interplay of distinct mechanisms, including quantum excitations, classical electrodynamics, and charge transfer, which involve a range of length and energy scales. This makes concise treatment of these systems very difficult: molecules on the order of a nanometer must be treated quantum mechanically, whereas plasmonic materials with sizes of up to hundreds of nanometers are usually too large to treat quantum mechanically in a computationally efficient way, and thus must be

treated classically in most cases. Therefore, a primary thrust of theoretical molecular nanopolaritonics is and will be the development of methods which can effectively treat and provide an interface between these classical (metal) and quantum (molecular) regions.

Nanopolaritonics: Methodology

Overview

There are several techniques used to model behavior in nanopolaritonic systems, but in general they involve coupling classical fields to a quantum mechanical molecule. Some examples are described in greater detail below. Briefly, the simplest method evolves the spatial charge density of the metal in time and couples plasmons to a molecular density matrix via dipole interactions. A more sophisticated treatment models classical regions with the finite-difference time-domain (FDTD) method (see the entry in this volume on ▶ Finite-Difference Time-Domain Technique), and couples the classical electric field to the molecular density matrix again via dipole interactions. This particular technique has the advantage that the classical description can easily be interfaced with essentially any type of quantum mechanical treatment of a molecule, including time-dependent density functional theory (TDDFT), time-dependent Hartree–Fock (TDHF), or semiempirical methods, as described later.

Plasmon Surface Modes Coupled to Electronic Density Matrices

In the simplest treatment, appropriate for molecules near metallic structures, a set of modes is first used to represent the plasmons via surface modes on the metals [6]. The charge density on the metals is written as

$$\sigma(\mathbf{r},t) = \sum_J C_J(t) S_J(\mathbf{r}) \tag{1}$$

where $S_J(\mathbf{r})$ is a surface-localized charge mode on a metal particle (e.g., dipole, quadrupole, etc., for a spherical metal particle, or a more general function in other cases), and the time-dependent coefficients which determine the evolution of the surface modes were introduced. The dynamics of these surface modes is associated with plasmons; the collective oscillation of the charges causes a time-dependent surface charge.

The surface charges are then coupled via dipole interaction to specific molecules; the resulting equation for the motion of the modes, including interaction with other modes and with specific molecules is then [2]:

$$\sum_K M_{JK}\ddot{C}_K(t) = -\sum_K V_{JK}C_K(t) - \sum_K M_{JK}\frac{\dot{C}_K(t)}{\tau_K} - \sum_{ij}\rho_{ij}(t)q_{ij,J} - v_J(t). \tag{2}$$

The first three terms are purely plasmon related, including mass terms (M_{JK}) associated with the kinetic energy of the surface modes, potential terms (V_{JK}) associated with coulomb interactions between the surface modes, and Drude damping terms ($1/\tau_K$) responsible for the dephasing and damping of the surface modes. The next term contains a dipole integral $q_{ij,J}$ coupling the plasmon modes to the molecule; the molecular density is represented as a density matrix with a constant (pre-interaction) part, ρ^0_{ij}, and a time-dependent part induced by interaction with the plasmons, $\rho_{ij}(t)$. Finally, the last term is the interaction with the external field, $V_J(t)$.

The other relevant equation is the time-dependence of the density matrix for the molecules. For each molecule, schematically, apply the Liouville equation:

$$i\frac{d\rho_{ij}(t)}{dt} = [F(\rho),\rho]_{ij} - i\gamma_{ij}\left(\rho_{ij}(t) - \rho^0_{ij}\right) + \left[\sum_J C_J(t)q_J,\rho\right]_{ij}, \tag{3}$$

where the following variables are introduced: the Fock matrix ($F(\rho)$), which depends itself on the density matrix; the damping coefficients (γ_{ij}) which determine the return of ρ to the equilibrium density matrix, ρ_0; and the coupling to the

classical fields of the plasmons via the dipole coefficients $q_{ij,J}$. Details of the different possibilities for a Fock operator are presented later.

The simplest way to view Eqs. 2 and 3 is when the excitation is not very large, as then they can be linearized in the density matrix. Then, the linearized equations have a simple form:

$$i\frac{d\mathbf{f}}{dt} = \mathbf{L}\mathbf{f}, \tag{4}$$

where $\mathbf{f}$ is the nanopolariton, i.e., the combined vector describing both the plasmon modes and the density matrix. Schematically, $\mathbf{f}$ is made from four parts: the surface charge coefficients and their time derivatives, as well as the real and imaginary part of the excitation of the density matrix away from the ground state:

$$\mathbf{f}^T = \left(C_J, \frac{dC}{dt}, \mathrm{Re}\left(\rho - \rho^0\right), \mathrm{Im}\left(\rho - \rho^0\right)\right). \tag{5}$$

In the generalized Liouville coupling matrix, L is straightforwardly shown to be

$$L = \begin{pmatrix} 0 & -\mathbf{M}^{-1} & 0 & 0 \\ -\mathbf{V} & -\tau^{-1} & -q & 0 \\ 0 & 0 & -\gamma & \Delta \\ -q & 0 & -\Delta - Z^{2B} & -\gamma \end{pmatrix}$$

The specific terms in the Liouville matrix result from a linear-response analysis of Eqs. 2 and 3.

Maxwell–Schrödinger Simulations

The plasmonic-mode model is appropriate for qualitative studies, especially because of its very modest computational requirements. However, a quantitative investigation requires a more realistic treatment of the electric field distribution. The best-known approach for simulating the electromagnetic fields is FDTD, where one propagates the Maxwell equations [7]. Assuming for simplicity no magnetic susceptibility effects, these read:

$$\begin{aligned} \frac{\partial \mathbf{E}(\mathbf{r},t)}{\partial} &= \frac{1}{\varepsilon_{\mathrm{eff}}(\mathbf{r})}\left(\nabla \times \mathbf{H}(\mathbf{r},t) - \mathbf{J}\,(\mathbf{r},t)\right) \\ \frac{\partial \mathbf{H}(\mathbf{r},t)}{\partial t} &= -\frac{1}{\mu_0}\nabla \times \mathbf{E}(\mathbf{r},t), \end{aligned} \tag{6}$$

where the electric and magnetic fields and the frequency-independent part of the susceptibility are introduced.

The current, J, is made from two parts, $\mathbf{J} = \mathbf{J}_p + \mathbf{J}_m$; the first term is the usual plasmonic contribution, which in FDTD is simulated as a sum of terms,

$$\mathbf{J}_p(\mathbf{r},t) = \sum_n \mathbf{J}_{p,n}(\mathbf{r},t), \tag{7}$$

where the individual terms fulfill

$$\frac{d\mathbf{J}_{p,n}(\mathbf{r},t)}{dt} = \alpha_n(\mathbf{r})\mathbf{J}_{p,n}(\mathbf{r},t) + \beta_n(\mathbf{r})\mathbf{E}(\mathbf{r},t). \tag{8}$$

The sum typically includes 1–5 terms, and the individual terms ($\alpha_n(\mathbf{r})$ and $\beta_n(\mathbf{r})$) in the sum are usually fitted to give the correct dielectric function in the frequency range of interest. Other, somewhat more sophisticated time evolution techniques for Eq. 8 have also been suggested.

The second part of the current is the molecular contribution, which for a single molecule will be

$$\begin{aligned} \mathbf{J}_M(\mathbf{r},t) &= \delta\left((\mathbf{r} - \mathbf{r}_M)\frac{d}{dt}\mathrm{Tr}(\rho\mathbf{x})\right) \\ &= -i\delta(\mathbf{r} - \mathbf{r}_M)\mathrm{Tr}([F,\rho]\mathbf{x}) \end{aligned} \tag{9}$$

where $\mathbf{x}$ is the electronic 3-D coordinate, and the delta function in the position of the molecule is represented numerically on the FDTD grid. The molecular current is obtained from a time-dependent study of the density matrix (or alternatively from propagating the molecular orbitals). In the simplest version, where only a dipole electric field is assumed, these take a form similar to Eq. 3 but now the coupling is directly due to the electric field:

$$i\frac{d\rho_{ij}(t)}{dt} = \left[F_{\mathrm{tot}}(\rho),\rho\right]_{ij} - i\gamma_{ij}\left(\rho_{ij}(t) - \rho_{ij}^0\right), \tag{10}$$

where

$$F_{\text{tot}}(\rho) = F(\rho) - \mu \cdot \mathbf{E}. \tag{11}$$

Molecular Simulation Methodologies

Throughout, the molecules are simulated in real time. For the present purposes, the molecules will be treated by a Liouville equation or a time-dependent orbital equation, where the initially occupied orbitals are propagated as:

$$i\frac{\partial \psi_n}{\partial t} = F(\rho)\psi_n. \tag{12}$$

There are different possible Fock operators; the most established are TDDFT and TDHF. Both methods, however, could be numerically expensive, so a simple alternate approach is to propagate in time a semiempirical approach; an example is TD-PM3 (time-dependent Parameter-Method 3) where the Fock operator has the generic TDHF form:

$$F_{ij}H_{0,ij} + \sum_{kl} V_{ijkl\rho kl}, \tag{13}$$

but with a very sparse set of matrix elements, V_{ijkl}. TDPM3 generally gives excitation energies with a $\sim$ 15 % accuracy [8], so is very suitable for the qualitative questions which arise in the study of nanopolaritonics. As the method uses a minimal basis set and scales well with system size, it is ideal for simulating large dye molecules.

Splitting Fields

It turns out that when there is one molecule, the delta function associated with having a localized charge is problematic and leads to numerical singularities. The reason is that this strong delta function source due to the molecule acts back on the molecule, creating an artificial self-energy problem. More precisely, for a single electron this is a self-energy problem as the electron interacts with itself; for a general molecule, the interaction with the molecule-induced current induces a double-counting problem, since the Fock operator already includes the electron–electron interaction and the interaction with the molecular current has essentially the same role.

This difficulty was alleviated through the development of a new formalism, in which the electric field is divided into two parts, a molecular and a plasmonic part [9]. The molecule is only affected by the external part, while the total field acts on the plasmon. Specifically, for the electric field the resulting equations are:

$$\frac{\partial \mathbf{E}_p}{\partial t} = \frac{1}{\varepsilon_{\text{eff}}}\nabla \times \mathbf{H} + \left(\frac{1}{\varepsilon_{\text{eff}}} - \frac{1}{\varepsilon_0}\right)\nabla \times \mathbf{H}_m - \frac{\mathbf{J}_p}{\varepsilon_{\text{eff}}}, \tag{14}$$

$$\frac{\partial \mathbf{E}_m}{\partial t} = \frac{1}{\varepsilon_0}(\nabla \times \mathbf{H}_m - \mathbf{J}_m) \tag{15}$$

while the magnetic fields equations are simpler

$$\frac{\partial \mathbf{H}_p}{\partial t} = \frac{1}{\mu_0}\nabla \times \mathbf{E}_p, \tag{16}$$

$$\frac{\partial \mathbf{H}_m}{\partial t} = \frac{1}{\mu_0}\nabla \times \mathbf{E}_m. \tag{17}$$

As mentioned, only the electric field that acts on the molecule enters into the evolution equation for the density matrix. Only the plasmonic field affects now the molecule, so that the Fock operator for the evolution of the density matrix is now:

$$F_{\text{tot}}(\rho) = F(\rho) - \mu \cdot \mathbf{E}_p. \tag{18}$$

The advantage in this approach is that the decoupling prevents numerical singularities. The disadvantage is that a separate electric field is required for each molecule. Therefore, future approaches will mix the FDTD treatment of the plasmonic field with a dipole and quadrupole treatment of the molecules, i.e., write the total electric field as

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_p(\mathbf{r},t) + \sum_j \mu(\mathbf{r},t')\nabla\frac{1}{|\mathbf{r}-\mathbf{r}_j|} + \text{Quad.}, \tag{19}$$

where "Quad." refers to quadrupolar interaction with nearest neighbors, and the molecular dipole is introduced. The magnetic field is written

similarly in terms of the molecular current and the plasmonic field, and the equation of evolution of the total field can be derived directly from Eq. 14. This mixed field-dipole approach will allow the treatment of a host of molecules interacting with an external field.

In the extreme limit that a continuum of molecules are required, the FDTD equations will be solved directly; then, the effect of each molecule will be relatively small so that the separation between plasmonic and molecular fields will not be needed, and the resulting, single electric field equations are the familiar Maxwell–Bloch equations, i.e., there will be no need to split the electric fields.

Key Findings

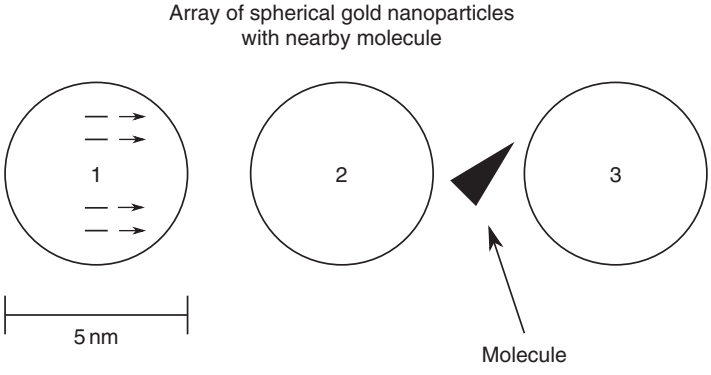
The fundamental conclusion from plasmon mode and Maxwell–Schrödinger (section Nanopolaritonics: Methodology) simulations is that under certain circumstances single molecules have a marked effect on the near-field transport of plasmons, as detailed below.

Molecule-Mediated Plasmon Transfer in Waveguides

An illustrative example of the type of system studied is shown in Fig. 1, where a single two-level dipolar molecule is located near a linear array of gold nanoparticles. In the absence of a molecule, an initial *x*-polarized plasmon excitation on the leftmost (input) nanoparticle will propagate down the array, which after a few femtoseconds results in an *x*-polarized plasmon on the rightmost (output) nanoparticle. Since the *x*, *y*, and *z* plasmon modes are decoupled in this linear configuration, no *y* or *z* excitation is induced on the output nanoparticle. Noble metal plasmonic waveguides of this type have been extensively studied both theoretically and experimentally [10].

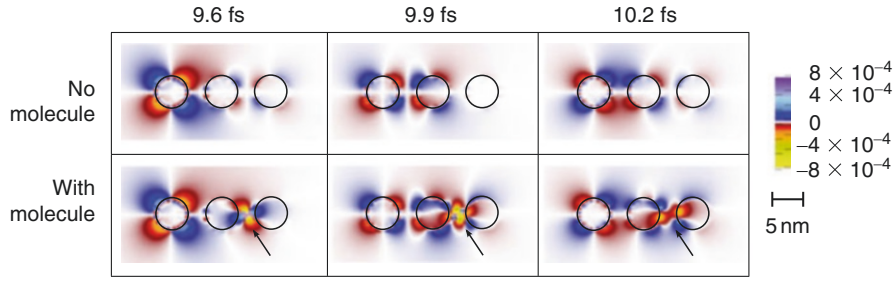
Perhaps surprisingly, the addition of a single molecule to the waveguide can have a drastic effect. Take, as a simple example, an *xy*-oriented two-level molecule with a strong resonance at $\omega_m = 2.55$ eV, which is close to the plasmon frequency of the array (2.60 eV). Figures 2 and 3 show results of a FDTD/linear-response TDDFT simulation. Figure 2 shows the clear transmission effect of a molecule in a time-domain simulation; the presence of the molecule causes significant enhancement in the amount of transmitted light in the nonnative initial polarization.

The frequency-domain picture is even more revealing. Figure 3 shows that the Fourier transform of the total current on the output nanoparticle displays a Fano-like resonance. As a reminder, Fano resonance are generally the result of coupling of single (or a few, in the generalized case) "dark" quantum level(s) to a continuum of states.



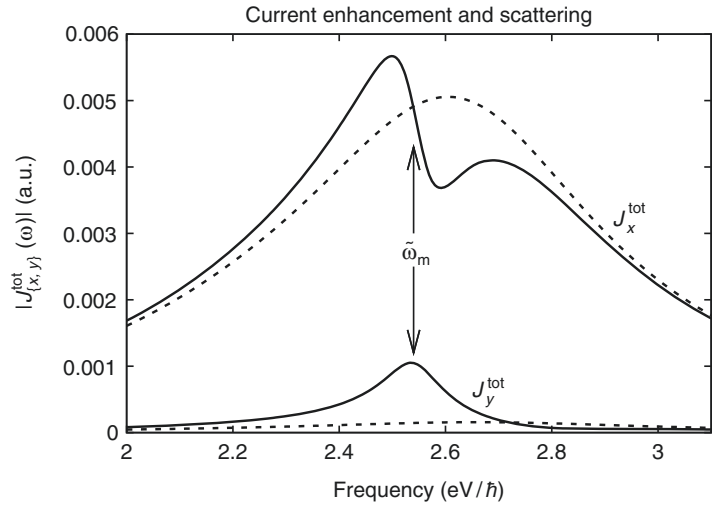
Nanopolaritonics, Fig. 1 Schematic of the computed setup. Three small metal spheres are placed close to each other. The *leftmost* sphere is excited; the *middle* sphere mediates this transfer. The molecule is placed between the middle and the final, *rightmost* sphere, and mediates this transport. The direction of the molecule influences the final polarization of the molecule

N



Nanopolaritonics, Fig. 2 Three snapshots in time of the *y*-component of the electric field with (*bottom*) and without (*top*) a molecule, for an excitation starting on the left sphere. The dipolar molecular field serves to scatter *x*-oriented energy (not shown) into *y*-oriented energy (From Ref. [9])

Nanopolaritonics, Fig. 3 Fourier transform of the currents on the output nanoparticle of a linear waveguide, showing that a single two-level *xy*-oriented molecule rotates incident plasmon polarization from $x \rightarrow y$ (From Ref. [9])



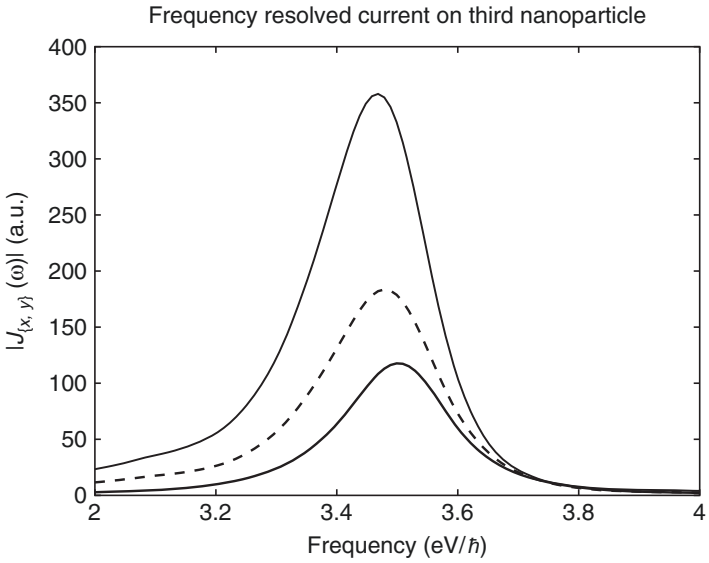
Their most obvious mark is a sine-like modification of the observable (spectrum or transmission profile) which in the original Fano resonance case will actually lead to a dip of zero absorption due to destructive interference caused by the presence of the quantum level.

In the present simulation case, a clear Fano-like transmission profile results due to the presence of the molecule; the excitation of the *x*-polarization is drastically modified. Specifically, the molecule enhances the *x*-polarized plasmon transmission over a range of frequencies below ω_m and decreases the transport for a range above. Further, the molecule rotates the excitation from the *x*-polarization to the *y*-polarization, resulting in a significant *y*-current (approximately 20 % of the *x*-current). Physically, the molecule, which has a transition dipole moment aligned along the *xy* axis, scatters *x*-polarized near-field light and reemits *y*-polarized light.

Despite the fact that the molecule does not affect the overall absorption of the array, the intense molecule–plasmon resonance in the gap between nanoparticles strongly influences plasmon transport, to the point of significantly scattering plasmons between otherwise orthogonal modes of excitation.

Realistic Simulations

To study this problem in a more rigorous manner, i.e., by using a more realistic quantum-chemistry treatment, the FDTD approach has also been combined with a time-dependent semiempirical method, TDPM3, mentioned above, to show rotation of current for large molecules with a significant number of possible excitations [11].



Nanopolaritonics, Fig. 4 Frequency-resolved current on the third silver nanoparticle of the system shown in Fig. 1 with 2-nm nanoparticles with (*bold lines*) and without (*thin line*) a molecule in the *x*-direction (*solid*) and *y*-direction (*dashed*) (From Ref. [11])

The system examined used silver nanoparticles, and thus tartrazine, a yellow dye molecule having a strong transition dipole moment near the plasmon resonance of silver, was chosen. The realistic simulations verify that molecules with an excitation energy near the plasmon resonance of the metal scatter the *x*-oriented current into the *y*-direction, as observed with the two-level molecule. The results are shown in Fig. 4. These results are encouraging as they confirm that even a molecule with a broad range of excitations can influence the transport of plasmons.

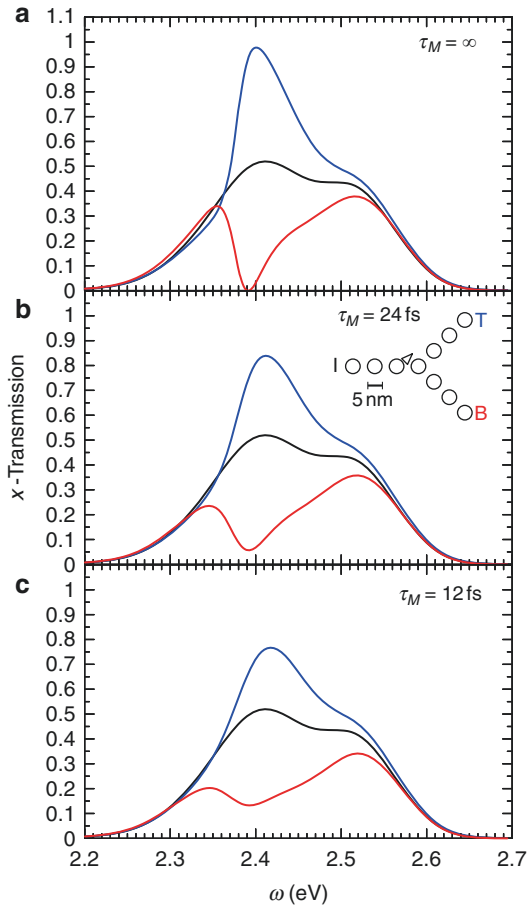
These studies offer insight into near-field light transport in heterogeneous nanosystems. Single molecules, although too weak to play a role in overall light absorption, can have drastic effects on transport if strategically placed in "hot spots," where they can modulate the comparatively lower energy coupling between nearby metal structures.

Since this is a resonance effect, it is highly dependent on the materials involved. The molecule, in order to scatter current, must have a strong optical transition at or near the coupling (i.e., plasmon resonance) of the waveguide. On one hand, this limits the diversity of the systems that can be used; however, this specificity is actually advantageous for sensing applications and tunable frequency-dependent wave guiding.

Controlling Plasmon Transport with Single Molecules

The strong effect of a single molecule on nanopolariton transport can be exploited to selectively control plasmon transport on the nanoscale. This has implications in the field of plasmonics, where molecule–plasmon interactions have promise as device components such as switches or higher harmonic generators. A molecule placed near a fork-like arrangement of gold nanoparticles, e.g., can selectively divert incoming plasmon waves into either the top or bottom paths, as demonstrated in Fig. 5, which shows transmission through the fork calculated using the plasmon mode approach. Beyond being sensitive to the molecule's location, this effect is also highly dependent on its orientation: When the molecule points upward, the plasmons are directed into the upper branch of the fork (*blue curve* in Fig. 5) much more than to the lower branch (*red curve*). In the absence of a molecule (*black curve*), the transition probabilities of the two branches are equal.

The controllability can be extended to use in molecular switches: Molecules can potentially be used to gate the transfer of near-fields in a specific and deliberate manner. Likewise, a similar configuration can act as a type of highly specific



Nanopolaritonics, Fig. 5 Simplified-model study of control of radiation transfer through a fork by placing a molecule aligned parallel to the top branch of the fork. *Blue top* branch, *red bottom*, *black* either branch without a molecule. The frames differ by the amount of dephasing on the excited state of the molecule. The gold dots have the experimentally fitted dephasing time, ~6 fs (From Ref. [2])

molecular sensor, sensitive to a molecule's identity, orientation, and position.

Limitations

There are several important caveats to these findings. First, compared to metals, molecules have a much smaller capacity to absorb and radiate light, as they are unable to further absorb light (of a given frequency) once already in their excited state. FDTD/TDDFT simulations of nanopolaritonic systems which treat the molecule beyond linear response have shown that under increasing field intensities the polarization rotation effects saturate rapidly; this transition region is especially interesting, since it can be used as a nonlinear gating mechanism.

Future Directions

Nanopolaritonics is still in its infancy. Although a variety of simulations have demonstrated proof of concept for controlling near-field light with molecules, there is much untapped potential for novel plasmon–molecule applications, as well as many outstanding theoretical and experimental questions. This last section briefly overviews future directions for both experiments and theory.

Transport-Based Sensing

Among the most promising potential applications is transport-based sensing, where extremely dilute concentrations of molecules are detected and uniquely identified based on how they alter the plasmon transport properties of a nanostructure system. In contrast to a SERS-like device, which is used to enhance and probe the response of a molecule, here the plasmon transport is measured, which is predicted to be extremely sensitive to the presence of select molecules. A pair of nanoscale metal tips, e.g., can act as a sensor for molecules which drift between them. Alternatively, one can build a parallel detector using a chip containing a series of different plasmonic junctions, each with a slightly different plasmon resonance, which are tuned to different target analyte molecules and excited by different frequencies of light.

Plasmonic Metamaterials

Another potential application of nanopolaritonics is in the development of more sophisticated metamaterials. Beyond simply mediating wave guiding, molecules could also mediate coupling between elements of a plasmonic metamaterial. This could lead to, e.g., tunable metamaterials where the presence of different molecules shifts the resonant frequency of the device.

Need for Experimental Studies

There is a pressing need for high-quality experimental measurements of plasmon transport in the

presence of strongly interacting molecules. Whereas there have been a wealth of experiments concerning the effects of intense fields on single molecules, studies on the effects of molecules on plasmons are lagging.

Improved Multiscale Simulation Methodologies

From a theoretical standpoint, progress needs to be made on efficiently and accurately capturing plasmon–molecule interactions, especially in the "chemical" regime where molecules are bound to the surface of a metal and share electrons. In reality both charge and energy are shared on a surface, and developing real-time embedding approaches capable of describing this is necessary to drive the field forward. This is a daunting task as it necessitates treating the quantum mechanical and classical regions as open systems. Moreover, seamless integration of the two regions requires careful choice of either boundary conditions or embedding potentials.

Cross-References

► Finite-Difference Time-Domain Technique
► Nanoparticles

References

1. Kasprzak, J., Richard, M., Kundermann, S., Baas, A., Jeambrun, P., Keeling, J.M.J., Marchetti, F.M., Szymanska, M.H., Andre, R., Staehli, J.L., Savona, V., Littlewood, P.B., Deveaud, B., Dang, L.S.: Bose-Einstein condensation of exciton polaritons. Nature **443**, 409–414 (2006)
2. Neuhauser, D., Lopata, K.: Molecular nanopolaritonics: cross manipulation of near-field plasmons and molecules. I. Theory and application to junction control. J. Chem. Phys. **127**(154715) (2007)
3. Nie, S., Emory, S.R.: Probing single molecules and single nanoparticles by surface-enhanced Raman scattering. Science **275**, 1102–1106 (1997)
4. Govorov, A.O., Lee, J., Kotov, N.A.: Theory of plasmon-enhanced Förster energy transfer in optically excited semiconductor and metal nanoparticles. Phys. Rev. B. **76**, 125308 (2007)
5. Lazarides, A.A., Schatz, G.C.: DNA-linked metal nanosphere materials: Fourier-transform solutions for the optical response. J. Chem. Phys. **112**, 2987–2993 (2000)
6. Prodan, E., Nordlander, P.: Plasmon hybridization in spherical nanoparticles. J. Chem. Phys. **120**, 5444–5454 (2004)
7. Gray, S.K., Kupka, T.: Propagation of light in metallic nanowire arrays: finite-difference time-domain studies of silver cylinders. Phys. Rev. B. **68**, 045415 (2003)
8. Bartell, L.A., Wall, M.R., Neuhauser, D.: A time-dependent semiempirical approach to determining excited states. J. Chem. Phys. **132**, 234106 (2010)
9. Lopata, K., Neuhauser, D.: Multiscale Maxwell–Schrödinger modeling: a split field finite-difference time-domain approach to molecular nanopolaritonics. J. Chem. Phys. **130**, 104707 (2009)
10. Maier, S.A., Atwater, H.A.: Plasmonics: localization and guiding of electromagnetic energy in metal/dielectric structures. J. Appl. Phys. **98**, 011101 (2005)
11. Arntsen, C., Lopata, K., Bartell, L.A., Wall, M.R., Neuhauser, D.: Modeling molecular effects on plasmonic transport: silver nanoparticles with tartrazine. J. Chem. Phys. **134**, 084101 (2011)

Nanopores

► Nanochannels for Nanofluidics: Fabrication Aspects

Nanopowders

► High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles
► Nanoparticles

Nanoprinting

► Micro- and Nanomanipulation for Nanomanufacturing

Nanoribbons

► Graphene

Nanorobotic Assembly

► Robot-Based Automation on the Nanoscale

Nanorobotic Manipulation of Biological Cells

► Robot-Based Automation on the Nanoscale

Nanorobotic Platforms for Biological Studies

► Nanorobotics for Bioengineering and Medical Interventions

Nanorobotic Spot Welding

Lixin Dong[1], Xinyong Tao[2], Zheng Fan[1], Zhang Li[3], Xiaobin Zhang[4] and Bradley J. Nelson[5]
[1]Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA
[2]College of Chemical Engineering and Materials Science, Zhejiang University of Technology, Hangzhou, China
[3]Institute of Robotics and Intelligent Systems, ETH Zurich, Zurich, Switzerland
[4]Department of Materials Science and Engineering, Zhejiang University, Hangzhou, China
[5]Institute of Robotics and Intelligent Systems, ETH Zurich, Zurich, Switzerland

Synonyms

Nanobonding; Nanointerconnection; Nanojoining; Nanosoldering

Definition

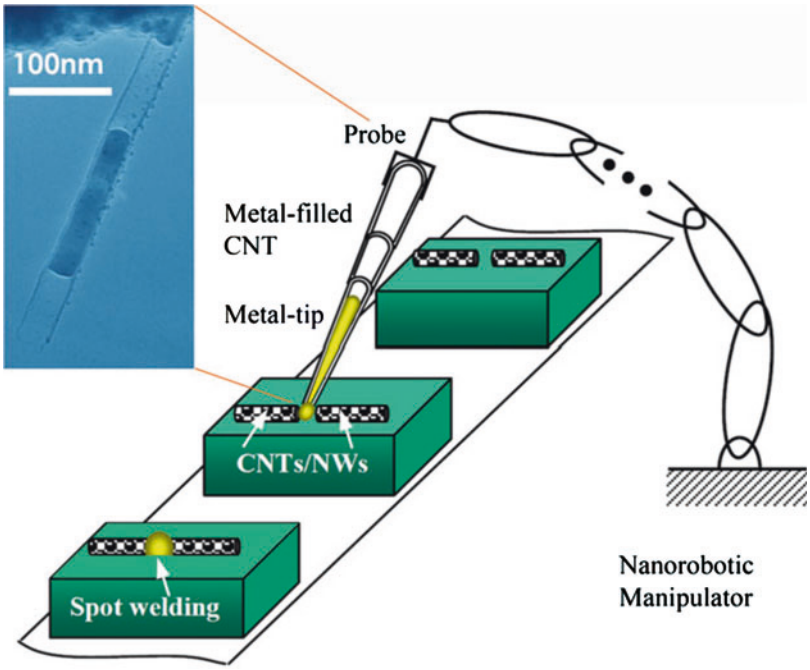
Spot welding, a widely used resistance welding process, was originally developed by Elihu Thompson in the early 1900s [1]. The most common application of spot welding is in the automobile industry, where it is widely used and most often performed automatically by preprogrammed industrial robots on assembly lines. With the continuing development of bottom-up nanotechnology fabrication processes, the nanoscale version of robotic spot welding, or nanorobotic spot welding [2], may likewise play an important role in interconnecting carbon nanotubes (CNTs), nanowires, nanobelts, nanohelixes, and other nanomaterials and structures for the assembly of nanocircuits and nanoelectromechanical systems (NEMS).

Overview

Van der Waals forces [3], electron-beam-induced deposition (EBID) [4], focused-ion-beam chemical vapor deposition (FIB-CVD), high-intensity electron-beam welding [5], Joule-heating-induced joining [6], and nanomechanochemical bonding [7] are experimentally demonstrated interconnection strategies, although all have limitations. Van der Waals forces are generally very weak, Joule-heating-induced joining and nanomechanochemical bonding are promising but not yet mature, and the other methods involve high-energy electron or ion beams, which significantly limits their applications. Another interconnection approach is to use CNTs, with their hollow cores and large aspect ratios, as possible conduits for nanoscale amounts of various materials that can be used to fuse CNTs together.

The concept of nanorobotic spot welding is schematically shown in Fig. 1. A metal-filled carbon nanotube is positioned with a nanorobotic manipulator to the interconnection site and the metal is then deposited to solder the nano-building blocks, such as CNTs and nanowires, together, or to weld them onto electrodes.

To facilitate the deposition, the encapsulated metallic materials may need to be melted or



Nanorobotic Spot Welding, Fig. 1 Concept of nanorobotic spot welding

evaporated, which means only materials with a lower melting point than that of CNTs can be applied. Fortunately, most pure metals have a lower melting point than that of CNTs in vacuum and a large portion of them have a lower melting point than that of CNTs in air (Table 1). Alloy solders routinely used in the larger scale for electronics and silversmithing may also serve as excellent candidates of nanosolders and they typically have lower melting points than the pure metals. These include soft solders (melting range: 90–450 °C, typically tin-lead (Sn-Pb) alloys), lead-free solders (melting point 5–20 °C higher than soft solders, alloys contain tin, copper (Cu), silver (Ag), bismuth (Bi), indium (In), zinc (Zn), antimony (Sb), and traces of other metals), and hard solders (melting point >450 °C, alloys of copper with either zinc or silver).

Figure 2 shows CNT samples synthesized using an alkali-doped Cu catalyst by a thermal CVD method [11]. Figure 2a–d show their structures imaged by field emission scanning electron microscopy (FESEM, Sirion, FEI), transmission electron microscope (TEM, JEM-2010, 200 kV), selected area electron diffraction (SAED), and high-resolution TEM (HRTEM). It can be seen from these observations that the yield of the Cu-filled CNTs was high. The CNTs were up to 5 μm long with outer diameters in a range of 40–80 nm. The single-crystalline Cu nanoneedles were encapsulated in graphite walls approximately 4–6 nm thick in the first sections of the bamboo-structured CNTs. The graphite layers were not parallel to the tube axis.

Key Research Findings

An experimental investigation of controlled melting and flowing of single crystalline copper from CNTs has been performed using nanorobotic manipulation [3], and its application in spot welding of nanotubes has been demonstrated using this copper. Because copper is a good conductor of heat and electricity and has a very low binding energy (0.1–0.144 eV/atom) when bound to carbon, copper-filled CNTs were used in the investigation.

The experiments were performed in a CM30 TEM equipped with a scanning tunneling microscope (STM) built in a TEM holder (Nanofactory Instruments AB, ST-1000) serving as a

Nanorobotic Spot Welding, Table 1 Melting point of pure metals and alloy solders

Materials	Melting point (°C)[a]	Possibility to be filled into CNTs
W	3410	
CNTs in vacuum	2800 [8]	
Fe	1535	Produced by pyrolysis of ferrocene powder [9]
Y	1523	Yttrium carbide synthesized from yttrium oxide powder mixed with isopropanol [10]
Cu	1083	Synthesized using an alkali doped Cu catalyst by a thermal CVD method [11]
Au	1064.43	Capillary filling [12]
Ag	961.93	Capillary filling [13]
CNTs in air	750 [8]	
Pb	327.5	Capillary filling [14]
Bi	271.3	Capillary filling
Sn	231.9	Synthesized by catalytic deposition of acetylene using nanocrystalline SnO_2 as a catalyst [11]
$Sn_{50}Pb_{50}$	183–212	
Ga	29.78	Synthesized in a vertical radiofrequency furnace from a homogeneous mixture of Ga_2O_3 and pure, amorphous, active carbon under a flow of pure N_2 gas [15]

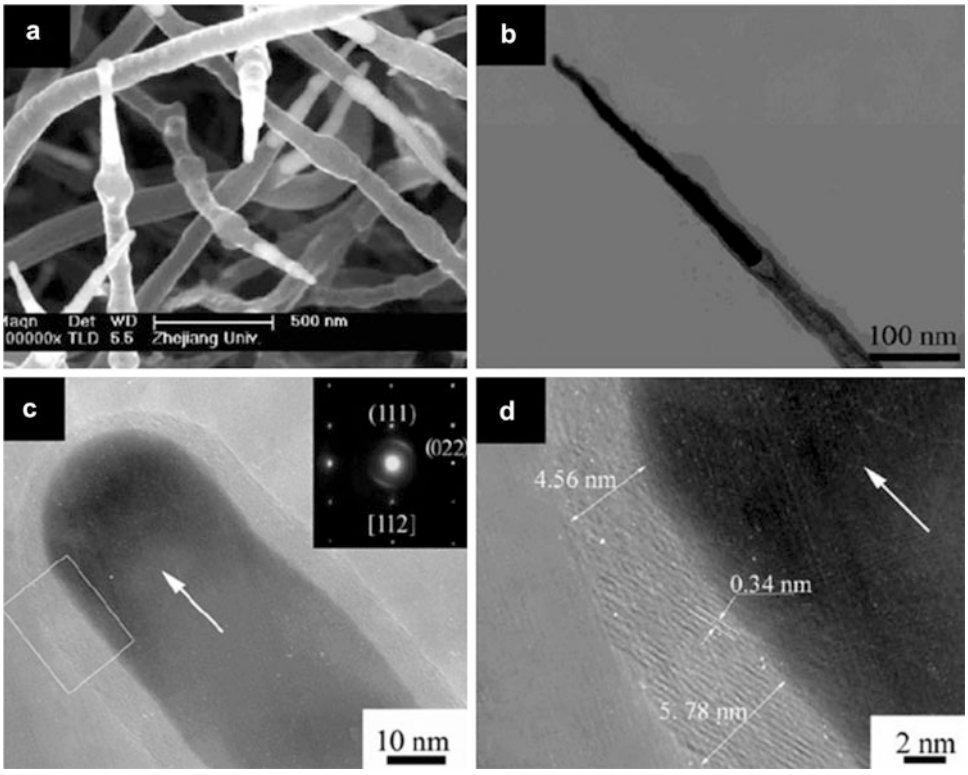
[a]Data source for melting points not specified: http://www.chemicalelements.com/show/meltingpoint.html

manipulator, as shown in Fig. 3a or schematically in Fig. 3b. The material consisting of a CNT bundle was attached to a 0.35 mm thick Au wire using silver paint, and the wire was held in the specimen holder. The probe was an etched 10 μm thick tungsten wire with a tip radius of approximately 100 nm (Picoprobe, T-4-10-1 mm). The probe can be positioned in a millimeter-scale workspace with subnanometer resolution with the STM unit actuated by a three-degree-of-freedom piezo-tube, making it possible to select a specific CNT. Physical contact can be made between the probe and the tip of a nanotube. Applying a voltage between the probe and the sample holder establishes an electrical circuit through a CNT and injects thermal energy into the system via Joule heating. By increasing the applied voltage, the local temperature can be increased past the melting point of the copper encapsulated in a tube. The process was recorded by TEM images and real-time video.

In the experiment, the tip of a Cu-filled CNT was first brought into contact with the tungsten probe. Then a bias voltage was applied on the two ends of the CNT with the tungsten probe serving as the anode. The voltage was slowly increased from 0 mV with 100 mV steps. When the voltage reached 1500 mV, a vacant section inside the carbon shells appeared, indicating that the copper has begun melting. Transportation of the copper to the probe-tip contact moved the vacant section to the root of the first bamboo section of the Cu-filled CNT. After the vacant section reached the root, the copper core started to flow to the tip of the CNT when bias was increased to 2500 mV.

Figure 4a is a series of time-resolved TEM images taken from video frames showing the flowing process. The copper core started to flow inside the carbon shell from the bottom to the tip of the first bamboo section as the bias voltage reached 2.5 V. The entire process continued for about 70 s. The flow rate was found to be 11.6 nm/s according to the change of apparent length of the copper core (Fig. 4b). Accordingly, the mass change was calculated as shown in Fig. 4c, and the mass flow rate can then be determined by fitting the data to the curve, yielding approximately 120 ag/s, which is strikingly slow and well controllable, allowing precise delivery of mass at the attogram scale for time-based control that can readily reach sub-second precision.

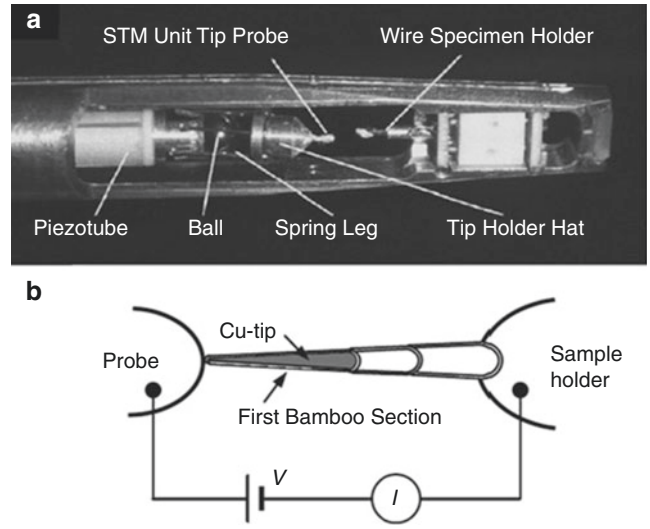
Figure 4d shows time-resolved current versus voltage characteristics under a constant positive bias of 2.5 V. The current density under 2.5 V when flowing occurred was then calculated according to the cross-sectional area as $2.60–3.07 \times 10^6$ A/cm^2. This is comparable to the observed value for electromigration of iron in CNTs (ca. 7×10^6 A/cm^2) [9]. The difference can be a result of the lower binding energy of copper to the carbon shells (0.1–0.144 eV/atom) than that of irons to carbon shells (0.3 eV/atom). The high current



Nanorobotic Spot Welding, Fig. 2 (**a**–**d**) Copper-filled CNTs. (**a**) FESEM images of Cu-filled CNTs. Observation shows that all the CNTs have sharp tips filled with metal nanoneedles. These CNTs are up to 5 μm long with outer diameters in a range of 40–80 nm. (**b**) A typical copper-filled CNT synthesized for 30 min. (**c**) HRTEM image reveals that the Cu nanoneedles are encapsulated in graphite walls approximately 4–6 nm thick. The inset is the corresponding SAED pattern of the Cu nanoneedle along the zone axis, showing that the Cu nanoneedle is single crystalline. The appearance of a pair of arcs in the SAED pattern indicates some orientation of the (002) planes in the carbon tubes. (**d**) The magnified image of the rectangular region in (**c**). The *large arrows* indicate the growth direction of the CNTs. The interlayer spacing of carbon nanotube is about 0.34 nm, consistent with the (002) plane lattice parameter of graphite. It can also be seen that the graphite layers are not parallel to the tube axis

densities employed here will lead to resistive heating. Temperatures as high as 2000–3000 °C have been estimated according to the lattice spacing change in electric breakdown experiments on non-filled multiwalled nanotubes (MWNTs) [16] at a slightly higher bias (3 V) than those used here. The current density J and the mass flow rate $\dot{m}$ were then correlated as shownine. The relation $\dot{m} = 0.3135J^2 - 1.6206J + 2.1373$ suggests that a real positive value of $\dot{m}$ ($\geq$42.9 ag/s) can only be given when the current density J surpasses 2.5847×10^6 A/cm^2. The existence of this threshold also implies that the mechanism of the observed flowing is most possibly electromigration [9]. Under a negative bias, that is, when the tungsten probe served as a cathode, the flow was observed in the opposite direction.

Other possible mechanisms for flow can be excluded. Capillary force can induce filling/flowing, but the direction should be opposite to the observed flow, that is, from the tip to the bottom of the carbon shells. Thermal expansion can enable flow, but the flow should be isotropic, heading toward both the tip and the bottom. A recent investigation showed that the irradiation of MWNTs can cause a large pressure buildup within the nanotube core that can plastically deform, extrude, and break encapsulated solid



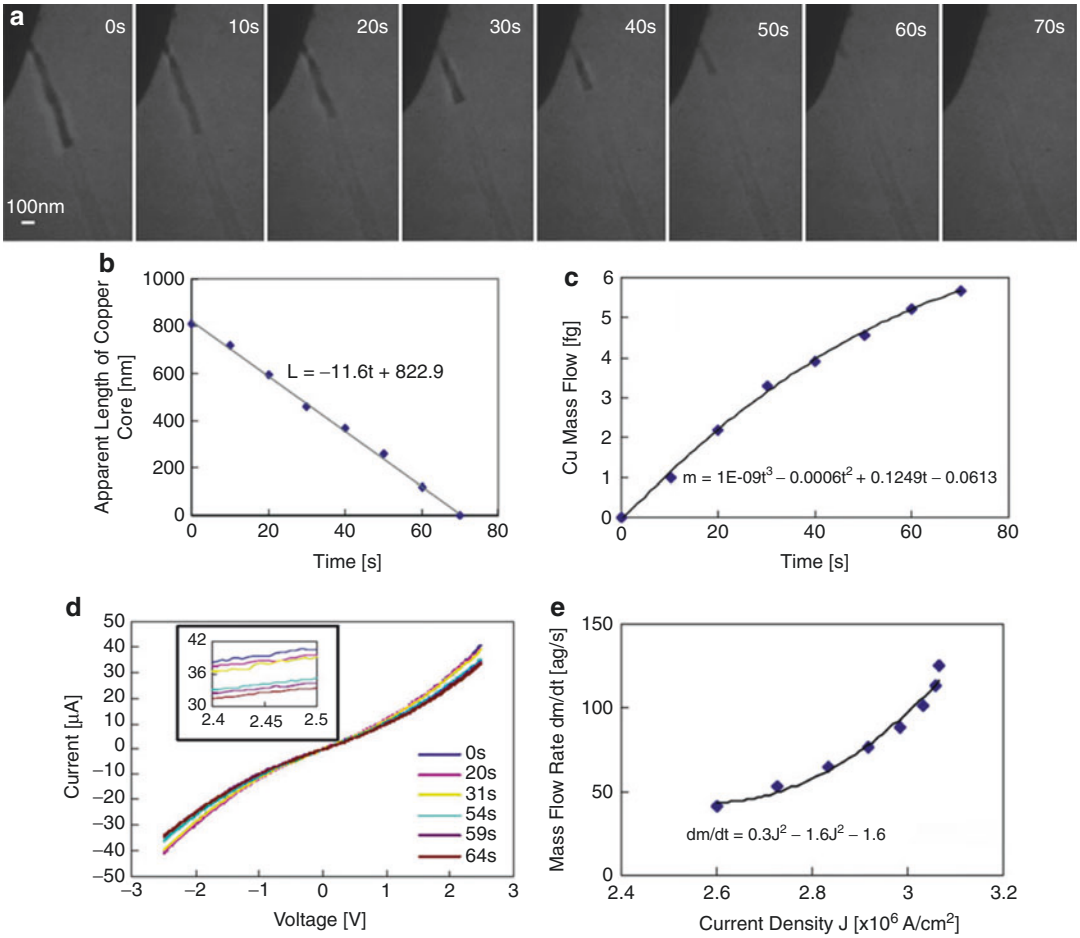
Nanorobotic Spot Welding, Fig. 3 (**a**, **b**) Nanorobotic manipulation system in a TEM

material [17]. In the experiments, however, no contraction of the carbon shells was observed.

Although there is no way of measuring the actual temperature of the nanotubes or the copper core, this process provides the possibility for estimating the temperature according to the resistivity change if Matthissen's rule is applied on temperature dependence of resistivity, provided that the resistivity was known at a certain temperature. Unfortunately, the latter is unknown. However, it has been noted that the temperature for CNT growth from 100-nm scale Cu particles in the synthesis (700 °C) is far lower than the melting point of bulk Cu (1083 °C) [11]. It is well known that the surface-to-volume ratio with respect to a nanosized particle can affect the melting point. Although the exact reason for melting at such a low temperature is not yet known, a rough estimation of the melting point of the copper core can be made, that is, around 700 °C. This is comparable to the in situ electron microscope observations of the melting point of the encapsulated 20–60 nm diameter Cu nanocrystals in multilayer graphitic carbon spheres, which was reported at 802 °C [18]. It is also known according to the experiment that the melting point of the carbon shells is much higher than that of the copper core. This provides further evidence of molten copper inside carbon shells.

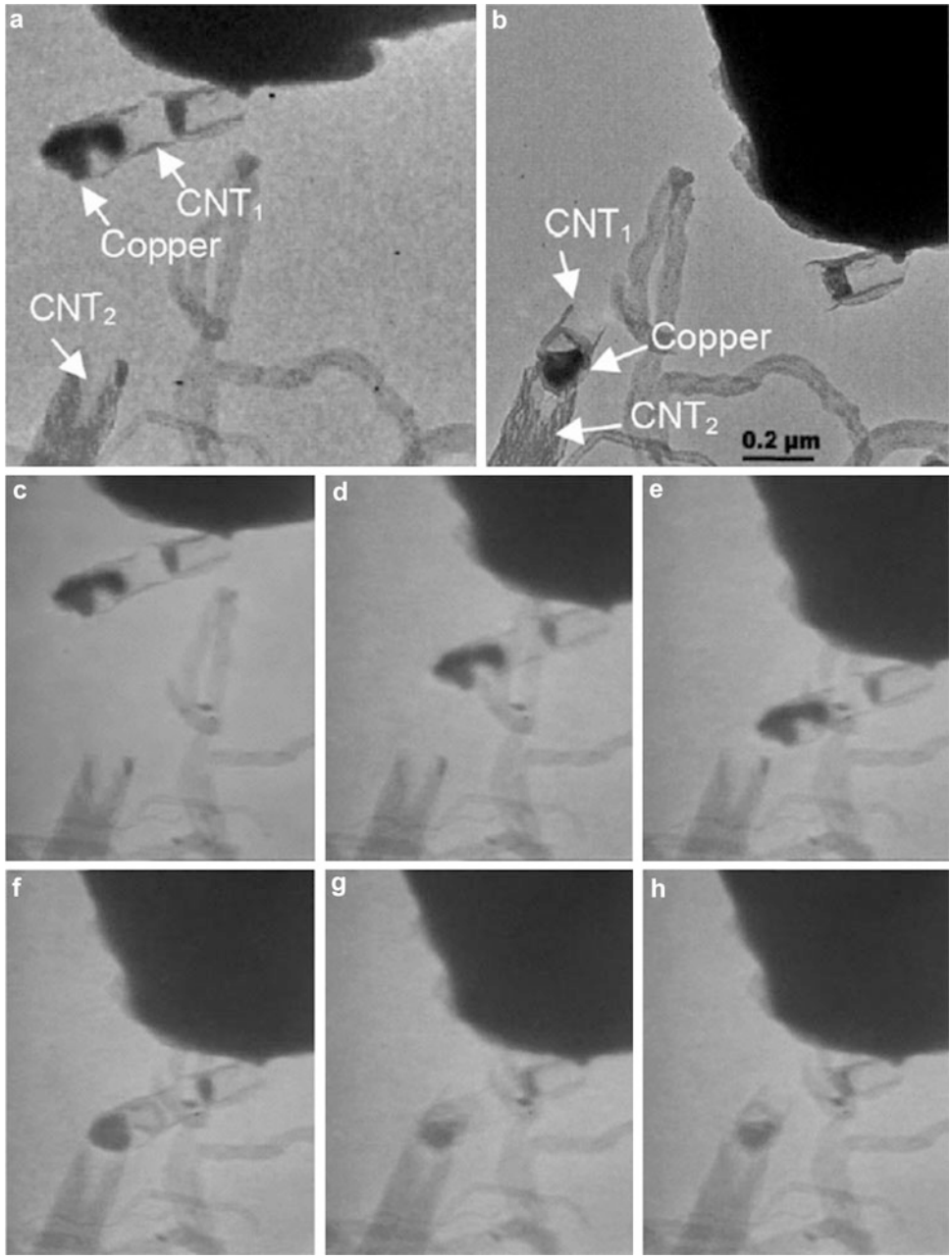
Examples of Application

The application of such controlled transportation of the copper core was then investigated. Self-soldering of CNTs with copper encapsulated in a CNT is shown in Fig. 5. A copper-filled tube, CNT_1, was first attached to a tungsten probe (Fig. 5a). A section of CNT_1 was then attached to and soldered onto CNT_2 by the melted copper (Fig. 5b). Figure 5c–h are video frames showing the soldering process. A copper-filled tube, CNT_1, was attached to a probe, and brought into contact with another tube, CNT_2. The probe has a −10 V bias. Parts c–e of Fig. 5 show three different positions as the probe was approaching CNT_2, and Fig. 5f shows how the contact has been made. The shape change of the copper suggests a melting process has occurred. It has been found from the video frames (25 fps) that the melting of the copper happened in a very short interval (<70 ms). With a higher bias (−15 V), CNT_1 was broken (Fig. 5g) and its end section remained soldered to the tip of CNT_2 (Fig. 5h).



Nanorobotic Spot Welding, Fig. 4 (**a**) Time-resolved TEM images from video frames showing the flowing process. The copper core started to flow inside the carbon shells from the root to the tip as the bias voltage reached up to 2.5 V. The whole process was continued for about 70 s. (**b**) The flow rate has been found to be 11.6 nm/s according to the change of apparent length of the copper core. The tungsten probe has been positively biased. (**c**) The mass changes along the time. The mass flow rate can be then drawn out from the fitting curve as approximately 0.12 fg/s. (**d**) Time-resolved current versus voltage characteristics under a constant bias of 2.5 V. The current density under 2.5 V as flowing occurred is about 2.60–3.07 × 106 A/cm^2. (**e**) Correlation of the current density J and the mass flow rate dm/dt

Compared with the other interconnection processes previously investigated, electrically driven spot welding has several interesting aspects: (1) A very low current can induce the melting and drive the flow; which is much more efficient that irradiation-based techniques involving high-energy electron beams [4, 5, 17, 19, 20], FIB, or lasers. Combined with dielectrophoretic assembly, it is possible to solder the tubes onto electrodes for batch fabrication of NEMS. (2) The welding site can be readily selected using nanorobotic manipulation, which enables 3D position and orientation control for continuous mass delivery and will potentially enable 3D prototyping and assembly. (3) The melting occurs rapidly (at least at the millisecond level), several orders of magnitude faster than that using high-intensity electron-beam or FIB, which is generally



Nanorobotic Spot Welding, Fig. 5 Self-soldering of CNTs with copper encapsulated in a CNT. (**a**) A copper-filled tube, CNT_1, was attached to a tungsten probe. (**b**) A section of CNT_1 was soldered onto CNT_2 by the melted copper. (**c**–**h**) Video frames showing the soldering process. A copper-filled tube, CNT1, was first attached to a probe, and brought into contact with another tube, CNT_2. The probe (cathode) has a −10 V bias. (**c**–**e**) Three different positions as the probe is approaching CNT_2. (**f**) Showing how contact has been made. The shape change of the copper suggested a melting process has occurred. With a higher bias (−15 V), CNT_1 was broken (**g**) and its end section remains soldered to the tip of CNT_2 (**h**)

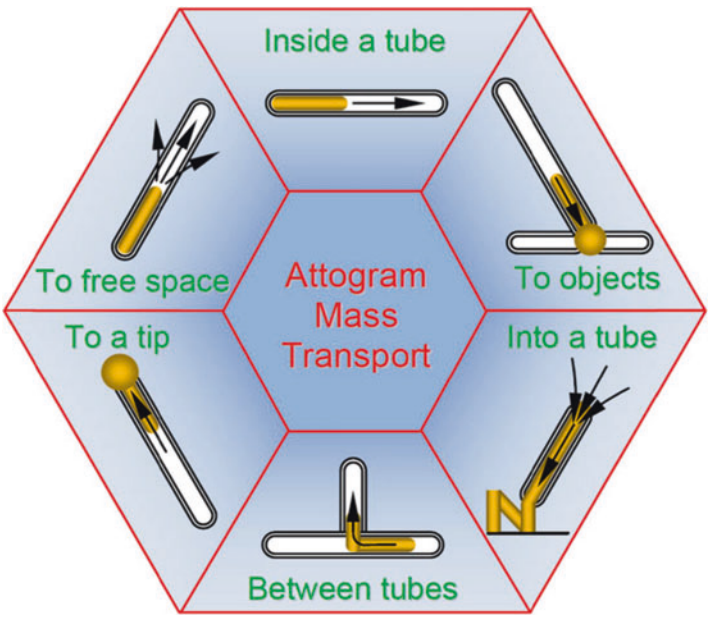
on the order of a minute [4, 5, 17, 20, 21]. (4) Because both the rate and direction of mass transport depends on the external electrical drive, precise control of an ultrasmall mass delivery is possible. Time-based control will allow the delivery of attograms of mass. (5) Copper has good compatibility in the conventional semiconductor industry. The experiments show that it will likewise play an important role for scaled-down systems. Carbon shells provide an effective barrier against oxidation and consequently ensure a long-term stability of the copper core, which also facilitates the conservation of the material than conveying mass on the external surface of nanotubes.

Future Directions for Research

The investigation on nanorobotic spot welding has so far shown that controlled melting and flowing of copper inside nanotube shells can be realized by applying a low bias voltage (~1.5–2.5 V). The melting can facilitate the flowing and it is a result of Joule heating, whereas the flowing is caused by electromigration. The mass flow rate can be controlled to the scale of attogram per second based on time: a typical value is around 1 atom/μs.

As an emerging technique, there are still a lot of unsolved problems with nanorobotic spot welding. (1) The materials investigated are limited to copper and tin. More efforts need to be put into the filling of conventional alloy solders and silver-based solders to broaden the adaptability of this technique. Universal and high yield filling techniques need to be developed. (2) The controllability of melting and flowing is related to the bias profile, which is determined mainly by the diversity of the resistivity of CNTs. Uniform CNTs in a batch are strongly required. (3) TEM provided an ideal environment for the in situ investigations. To lower the cost for practical applications, it would be a better solution to combine the spot welder with an AFM. The impact of the change of the environment from a high-vacuum chamber to the air needs to be evaluated. Obviously, solders with lower melting points will be required. (4) For automatic batch soldering, continuous feeding, arrays of solders (can be individually filled with heteromaterials), automatic detection of the soldering sites, and path planning are some examples of critical techniques.

Nanorobotic spot welding is a fundamental technology related to the mass transport inside, between, and into nanochannels (Fig. 6). Mass transport systems at this scale provide a platform



Nanorobotic Spot Welding, Fig. 6 Related investigations and applications

for the investigation of nanofluidics and can serve as components for prototyping nanostructures. Mass flowing out from a nanotube can generate a variety of results based on the targets. Evaporation to a free space is a result of discharging, ionization, or heating. Deposition on a surface can be obtained if placing the nanotube tip against such a target. Nanorobotic spot welding is based on mass transport to the interface between two objects. Free-standing nanostructure can be formed on the tip of the injector nanotube if moving it back from the target. As an emerging additive nanolithography technique, electromigration-based deposition (EMBD) is a more general term representing the mass transport of metallic materials against a surface. The investigations of all these systems are correlated and can benefit from the development of each other.

Cross-References

► Electron-Beam-Induced Deposition

References

1. Cary, H.B., Helzer, S.C.: Modern Welding Technology, 6th edn. Prentice-Hall, Upper Saddle River (2005)
2. Dong, L.X., Tao, X.Y., Zhang, L., Zhang, X.B., Nelson, B.J.: Nanorobotic spot welding: controlled metal deposition with attogram precision from copper-filled carbon nanotubes. Nano Lett. **7**(1), 58–63 (2007)
3. Dong, L.X., Arai, F., Fukuda, T.: Destructive constructions of nanostructures with carbon nanotubes through nanorobotic manipulation. IEEE-ASME Trans. Mechatron. **9**(2), 350–357 (2004)
4. Dong, L.X., Arai, F., Fukuda, T.: Electron-beam-induced deposition with carbon nanotube emitters. Appl. Phys. Lett. **81**(10), 1919–1921 (2002)
5. Xu, S.Y., Tian, M.L., Wang, J.G., Xu, J., Redwing, J. M., Chan, M.H.W.: Nanometer-scale modification and welding of silicon and metallic nanowires with a high-intensity electron beam. Small **1**(12), 1221–1229 (2005)
6. Hirayama, H., Kawamoto, Y., Ohshima, Y., Takayanagi, K.: Nanospot welding of carbon nanotubes. Appl. Phys. Lett. **79**(8), 1169–1171 (2001)
7. Dong, L.X., Arai, F., Fukuda, T.: Nanoassembly of carbon nanotubes through mechanochemical nanorobotic manipulations. Jpn. J. Appl. Phys. Part 1 – Regul. Pap. Short Notes Rev. Pap. **42**(1), 295–298 (2003)
8. Collins, P.G., Avouris, P.: Nanotubes for electronics. Sci. Am. **283**(6), 62–69 (2000)
9. Svensson, K., Olin, H., Olsson, E.: Nanopipettes for metal transport. Phys. Rev. Lett. **93**(14), 145901 (2004)
10. Seraphin, S., Zhou, D., Jiao, J., Withers, J.C., Loutfy, R.: Yttrium carbide in nanotubes. Nature **362**(6420), 6503 (1993)
11. Tao, X.Y., Zhang, X.B., Cheng, J.P., Luo, Z.Q., Zhou, S.M., Liu, F.: Thermal cvd synthesis of carbon nanotubes filled with single-crystalline Cu nanoneedles at tips. Diamond Relat. Mater. **15**(9), 1271–1275 (2006)
12. Chu, A., Cook, J., Heesom, R.J.R., Hutchison, J.L., Green, M.L.H., Sloan, J.: Filling of carbon nanotubes with silver, gold, and gold chloride. Chem. Mater. **8**(12), 2751–2754 (1996)
13. Ugarte, D., Chatelain, A., de Heer, W.A.: Nanocapillarity and chemistry in carbon nanotubes. Science **274**(5294), 1897–1899 (1996)
14. Ajayan, P.M., Iijima, S.: Capillarity-induced filling of carbon nanotubes. Nature **361**(6410), 333–334 (1993)
15. Gao, Y.H., Bando, Y.: Carbon nanothermometer containing gallium – gallium's macroscopic properties are retained on a miniature scale in this nanodevice. Nature **415**(6872), 599 (2002)
16. Huang, J.Y., Chen, S., Jo, S.H., Wang, Z., Han, D.X., Chen, G., Dresselhaus, M.S., Ren, Z.F.: Atomic-scale imaging of wall-by-wall breakdown and concurrent transport measurements in multiwall carbon nanotubes. Phys. Rev. Lett. **94**(23), 236802 (2005)
17. Sun, L., Banhart, F., Krasheninnikov, A.V., Rodrıguez-Manzo, J.A., Terrones, M., Ajayan, P.M.: Carbon nanotubes as high-pressure cylinders and nanoextruders. Science **312**, 1199–1202 (2006)
18. Schaper, A.K., Phillipp, F., Hou, H.: Melting behavior of copper nanocrystals encapsulated in onion-like carbon cages. J. Mater. Res. **20**(7), 1844–1850 (2005)
19. Madsen, D.N., Molhave, K., Mateiu, R., Rasmussen, A.M., Brorson, M., Jacobsen, C.J.H., Boggild, P.: Soldering of nanotubes onto microelectrodes. Nano Lett. **3**(1), 47–49 (2003)
20. Yokota, T., Murayama, M., Howe, J.M.: In situ transmission-electron-microscopy investigation of melting in submicron Al-Si alloy particles under electron-beam irradiation. Phys. Rev. Lett. **91**, 265504 (2003)
21. Matsui, S., Kaito, T., Fujita, J., Komuro, M., Kanda, K., Haruyama, Y.: Three-dimensional nanostructure fabrication by focused-ion-beam chemical vapor deposition. J. Vac. Sci. Technol. B **18**(6), 3181–3184 (2000)

Nanorobotics

Soichiro Tsuda
School of Chemistry, University of Glasgow, Glasgow, UK

Synonyms

Microrobotics

Definition

Nanorobots are robotic devices operating at micro- and nanoscale.

Overview

Nanorobotics is an emerging subfield of nanotechnology that aims to build robotic devices operating at micro- and nanoscale. The envisioned applications of nanorobotics are quite wide, ranging from medical and environmental sensing to space and military applications [1, 2]. Devices at micron scale that handle micron objects (e.g., bacterial robots), which are sometimes termed as "microrobots," are also referred to as nanorobots here. Nano manipulators that control nanoscale objects using microscope devices are also referred to as nanorobots, but it is not covered in this article.

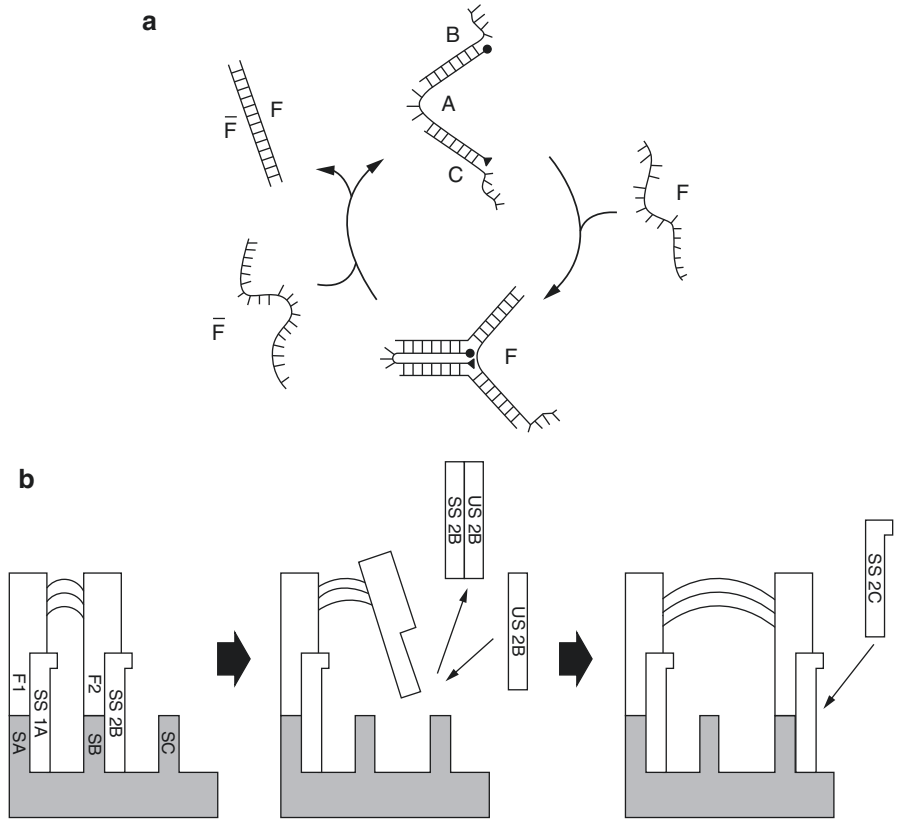
While nanotechnology developed so far is concerned with the construction of static nanoscale structures and manipulation of nanoscale objects, the field of nanorobotics studies more dynamic and active structures. It includes the design and development of nanoscale machines that are capable of actuation, sensing, signaling, information processing, and intelligent behavior at nanoscale [1]. Being extremely small (potentially at atomic scale), nanorobots are unlikely to be equipped with many functions as in their macroscopic counterparts. A possible solution to implement more complex functionality would be to use swarm intelligence and cooperative behavior, in which a swarm of nanorobots communicate each other and accomplish tasks collectively. Another issue in regard to nanorobotics is the basic law that governs the dynamics of nanorobots. Conventional robots at macroscopic scale are governed by Newtonian mechanics, whereas nanorobots are governed mostly by quantum mechanics. The design principle of nanorobots is thus bound to be different from macroscale robots. For example, Brownian motion would need to be taken into serious consideration at such scale. Molecules constituting nanoscale devices are constantly exposed to collisions with other smaller molecules (e.g., water molecules) and not in a rigid, static structure [3]. Therefore, novel schemes of design, manufacturing, programming, and control for nanorobots are needed because schemes for macroscale robots cannot be applied.

Although nanorobots in practical use are yet to be developed, several prototypes of nanorobots have been built in recent years. In particular, they have been explored in the context of biomolecular machines inspired by "nature's way of doing things" [1, 2]. Biomolecules are of particular interest in the field of nanorobotics because they operate at extraordinarily high efficiencies and readily available in nature [4]. In addition, they have several functions which are difficult to achieve with artificial molecular machines, for example, self-assembly and self-replication. As it is impossible to precisely fabricate nanomachines atom by atom, nature employs self-organization to build natural molecular machines. Proteins, the most widely used nature's nanomachines, are constructed using DNA as blueprint. A DNA sequence is translated into a chain of amino acids via messenger RNA, and the amino acid chain will self-assemble into a three-dimensional structure through free-energy minimization. By separating replication process from the system itself (i.e., protein), self-replication becomes easier, and over one billion possible types of proteins are possible out of 20 amino acids and four nucleic acids. Potentially, nanorobots could be fabricated in a similar way that biological devices are assembled. In fact, several prototypes of nanorobots, such as

DNA-based robots, rely on self-assembling nature of biomolecules for their operation. The following section gives a brief overview of nanorobotic devices implemented on various materials.

DNA-Based Nanorobots

Yurke and coworkers developed a molecular machine that takes DNA strands as fuels, called DNA tweezers [5]. The machine itself is also consists of three single-strands of DNA (Fig. 1a). Two of these stands (Strand B and C) have complementary sequence with another strand (Strand A). In particular, half of Strand B is complementary to half of Strand A and half of Strand C is to another half of Strand A. Thus, when they are bound together, the DNA tweezers has a single-stranded DNA sequence on each end, which are parts of Strand B and C. Strand A is labeled with fluorescent dye TET (5′ tetrachloro-fluorescein phosphoramidite) and TAMRA (carboxy-tetramethylrhodamine) on each end. Fluorescence emission from TET reports the positions of DNA tweezers. When TET is excited with a 514.5 nm argon ion laser, it emits fluorescent light with a peak wavelength of 536 nm. TAMRA works as a quencher of TET due to resonant intramolecular energy transfer when they come close together. When no fuel DNA (Strand F) is present, the DNA tweezers is "open" and TET fluoresces. If Strand F is added, it binds to the



Nanorobotics, Fig. 1 (**a**) A schematic diagram of DNA tweezers. Open-close operations are controlled by additional fuel strands F and $\overline{F}$. *Black circle* and *triangle* indicate TET (fluorescent dye) and TAMRA (quencher), respectively. After Yurke et al. [5]. (**b**) DNA walker walks on the steps constructed with single-stranded DNA. F1 and F2 indicate feet of the DNA walker, SA, SB, and SC are steps for DNA walker. Gaits of the walker are precisely controlled by external DNA strands, SS (Set Strand) and US (Unset Strand). See text for details. After Sherman and Seeman [6]

sticky ends of DNA tweezers because it has complementary sequences to single-stranded ends of the tweezers (parts of Strand B and C). As a result, they take a Y shape and the DNA tweezers are in "closed" position. This brings together TET and TAMRA attached to Strand A, and the light emission from TET is quenched. The tweezers can be reopened by adding Strand $\overline{F}$, a complementary DNA to Strand F. It cleaves the hybridization between the DNA tweezers and Strand F and forms a double-stranded DNA with Strand F. This open-close cycle can be repeated several times by removing fuel DNA added in the previous cycle. They have estimated the force exerted by the closing as 15 pN, which is at the upper range of measured forces exerted by biomolecular motors, such as kinesin and myosin.

Sherman and Seeman developed a simple mobile robot based on DNA strands, called DNA walker [6]. The walker has two feet (Foot 1 and 2 in Fig. 1b), each of which is comprised of double-stranded and single-stranded DNA sequences. Two DNA feet are connected with three single-stranded DNA. The DNA walker walks on a line of "stepping stones" that are also made of single-stranded DNA sequences. The DNA walker is set to "stand" on stepping stones by another DNA strands (Set Strand) that are complementary to a Foot and a Step. For example, to set Foot 1 on Step A, a Set Strand 1A, which is complementary to the sequences of Foot 1 and Step A, is added to a solution, and it hybridizes with the foot and step to connect them together. The molecular machine walks on the stepping stones by addition and removal of Set Strands: If feet are set on neighboring stones Step A and B - (linked by Set Strand 1A and 2B), the DNA walker lifts up one of legs by adding a complementary DNA strand to Set Strand 2B (called Unset Strand 2B). The strand cleaves the binding of Set Strand 2B with the foot and step to form a double-stranded DNA, and thus the Foot 2 is released from Step B. Foot 2 can be set on Step C by adding new strand Set Strand 2C. The same process can be applied to Foot 1 in order to step forward from Step A to B. Hence, the DNA walker can walk on the steps by repeating the process.

Based on the DNA walker, Lund and colleagues developed a DNA-based molecular robot that walks on two-dimensional prescribed landscapes [7]. The robot, called molecular spider, consists of a head and three legs, all of which are made of DNA. The way the molecular spider walks is basically similar to the DNA walker: It proceeds on a "track" because each leg (a single-stranded DNA) hybridizes and dehybridizes with another single-stranded DNA with complementary sequences to the leg. A big difference with the spider is autonomous behavior. It carries out sequences of actions, such as "start," "follow," "turn," and "stop," without any external controls. This autonomous action relies on catalytic activity of the spider legs. A DNA sequence with catalytic action, called DNA enzyme, constitutes each leg, and a track on the 2D DNA origami landscape [8] is constructed with three lines of oligodeoxynucleotide substrates (i.e., stepping stones). The locomotion of DNA spider is carried out as follows: First, a leg binds to a substrate by hybridization. DNA enzyme cleaves the single ribonucleotide linkage in the substrate strand into two shorter products. Due to a lower affinity of the products for the DNA enzyme, the leg will be dissociated from the site. It will quickly find a nearby substrate to reattach because multiple legs binding to substrates prevent the molecular spider from being completely dissociated from the track. Although the leg could attach to a site visited before (i.e., stepping back), it is likely to stay longer on new sites because the substrate has higher affinity for the enzyme than the products. Once a molecular spider is set on the starting point of a track, three legs bind and cleave the substrates in parallel, and the robot moves forward as a result. It has been shown that the spider can not only walk on a straight track, but also turns the direction (30° and 90°). After it reaches the end of the track, it would perform random Brownian walks because all the substrates are converted to the product and, therefore, any sites on the track have same affinity for the legs. To prevent this, "stop" sites, which are non-chimeric and uncleavable analogues of the substrate, can be introduced at the end of a track. When a leg steps on the site, cleavage of DNA strand does

not occur, and, therefore, the robot stays in the stop sites and never goes back.

Cell-Based Nanorobots

Another approach towards nanorobots is to employ whole living cells as machines. As mentioned above, natural biomolecular machines are quite efficient in many aspects, and it is hard to fabricate competing artificial robotic elements from scratch. Cell-based nanorobotics aims at making use of biological functionality of living organisms for man-made tasks, such as cargo transportation.

Behkam and Sitti developed a bacterial actuator for propulsion of polystyrene microbeads [9]. The cell-based actuator employs flagella motors of *Serratia marcescens*. Using a plasma-based bacteria patterning technique, the flagellated chemotactic bacteria have been randomly attached to half of a 10 μm microbead. This configuration of bacteria attachment propels the microbead, although the direction of motion is stochastic. It has been shown that the more bacteria are attached, the faster the bead is propelled. The on-off motion of the flagella motors can be controlled by introducing copper ions (Cu^{2+}) and ethylenediaminetetraacetic acid (EDTA). Steager and colleagues [10] showed the bacteria can thrust SU-8 microfabricated structures (larger than the microbeads) and the movement of the structures can be halted and resumed by UV exposure.

Martel and colleagues has been taking a different approach to the development of bacterial microrobots [11]. They point out several drawbacks of bacteria actuators using *Serratia marcescens*: First, the directional control is difficult. Second, the chemical control of bacterial motion is not appropriate for applications in many environments, especially when applications for operations in the human body are in focus. Third, chemical control is not compatible with electronic circuits, i.e., digital computers. To tackle these problems, they use magnetotactic bacteria (MTB). The bacteria, marine magnetotactic coccus strain MC-1, incorporate a chain of nanoparticles called magnetosomes which are used to orient the bacteria in geomagnetic fields. On the other hand, the direction of magnetotactic bacteria can be controlled by external magnetic field by inducing a torque on a magnetosome. A bacterium can produce a thrust force exceeding 4 pN, and a collective of MC-1 cells assembled by computer-controlled magnetic field can generate larger force. It has been shown that the collective of approximately 5,000 flagellated magnetotactic bacteria can not only propel microstructures, but also complex microassembly tasks, such as building a pyramid-like structure [12].

Other Types of Nanorobots

There are several approaches toward nanorobots apart from the DNA and cell-based implementations. Another example is protein-based nanorobotic system. Protein is the most common natural molecular robotic devices and abundant in living cells. Central to the protein-based nanorobotics is the conformational change of three-dimensional structure because proteins are quite sensitive to the environmental changes. It has been proposed to use viral proteins as motors (viral protein linear (VPL) motor). This has been inspired by the fact that the families of retroviruses, such as influenza virus and HIV-1, are able to sensitively detect a drop in pH as a signal that its future host is close. The drop of pH is detected by a protein on the viral membrane, which changes the conformation from loosely connected structure to a distinctive α-helical structure. Artificially synthesized protein-like viral proteins could potentially used to produce forces of the order of μN.

Inorganic nanorobots are also pursued (cf. [13]). What has been proposed is controllable nanomotor devices using catalytic reaction of bimetal. When cylindrical asymmetrical alloys (e.g., gold and platinum) are suspended in a solution with hydrogen peroxide as fuel, the platinum segments catalyze the oxidation of hydrogen peroxide (i.e., $H_2O_2 \rightarrow O_2 + 2H^+ + 2e^-$) whereas the gold segments enhance the opposite reaction. As a result, electrons in the nanomotor flows from the

platinum to gold side along with migration protons in the double layer on the motor, and the nanomotor shows random but non-Brownian movements. Precise directional control for performing complex tasks is also possible by applying magnetic field. Ferromagnetic nickel incorporated in the nanomotors allows the aligning of the motor and modulation of magnetic field strength enables start and stop control of the motion.

Cross-References

- ▶ 3D Micro-/Nanomanipulation with Force Spectroscopy
- ▶ CNT Handling and Integration
- ▶ Micro- and Nanomanipulation for Nanomanufacturing
- ▶ Nanorobotic Spot Welding
- ▶ Nanorobotics for Bioengineering
- ▶ Nanorobotics for NEMS Using Helical Nanostructures
- ▶ Robot-Based Automation on the Nanoscale

References

1. Ummat, A., Dubey, A., Mavroidis, C.: Bionanorobotics: a field inspired by nature, Chapter 7. In: Bar-Cohen, Y. (ed.) Biomimetics: Biologically Inspired Technologies, pp. 1–227. CRC, Boca Raton (2006)
2. Ummat, A., Dubey, A., Sharma, G., Mavroidis, C.: Bio-nano-robotics: state of the art and future challenges, Chapter 19. In: Yarmush, M.L. (ed.) The Biomedical Engineering Handbook, vol. 3, 3rd edn. CRC, Boca Raton (2006)
3. Yanagida, T.: Fluctuation as a tool of biological molecular machines. Biosystems **93**(1–2), 3–7 (2008)
4. Weir, N.A., Sierra, D.P., Jones, J.F.: A review of research in the field of nanorobotics. In: Technical Report SAND 2005-6808. Sandia National Laboratories (2005)
5. Yurke, B., Turber Jr., A.J., Mills, A.P., Simmel, F.C., Neumann, J.L.: A DNA-fuelled molecular machine made of DNA. Nature **406**(August), 605–608 (2000)
6. Sherman, W.B., Seeman, N.C.: A precisely controlled DNA biped walking device. Nano Lett. **4**(7), 1203–1207 (2004)
7. Lund, K., Manzo, A.J., Dabby, N., Michelotti, N., Johnson-Buck, A., Nangreave, J., Taylor, S., Pei, R., Stojanovic, M.N., Walter, N.G., Winfree, E., Yan, H.: Molecular robots guided by prescriptive landscapes. Nature **465**(7295), 206–210 (2010)
8. Rothemund, P.W.: Folding DNA to create nanoscale shapes and patterns. Nature **440**(7082), 297–302 (2006)
9. Behkam, B., Sitti, M.: Effect of quantity and configuration of attached bacteria on bacterial propulsion of microbeads. Appl. Phys. Lett. **93**(22), 223901 (2008)
10. Steager, E., Kim, C., Patel, J., Bith, S., Naik, C., Reber, L., Kim, M.: Control of microfabricated structures powered by flagellated bacteria using phototaxis. Appl. Phys. Lett. **90**(26), 263901 (2009)
11. Martel, S., Mohammadi, M., Felfoul, O., Lu, Z., Pouponneau, P.: Flagellated magnetotactic bacteria as controlled mri-trackable propulsion and steering systems for medical nanorobots operating in the human microvasculature. Int. J. Rob. Res. **28**(4), 571–582 (2009)
12. Martel, S., Mohammadi, M.: Using a swarm of self-propelled natural microrobots in the form of flagellated bacteria to perform complex micro-assembly tasks. In: 2010 I.E. International Conference on Robotics and Automation (ICRA), Anchorage, May 2010, pp. 500–505 (2010)
13. Wang, J.: Can man-made nanomachines compete with nature biomotors? ACS Nano **3**(1), 4–9 (2009)

Nanorobotics for Bioengineering

- ▶ Robot-Based Automation on the Nanoscale

Nanorobotics for Bioengineering and Medical Interventions

Sylvain Martel
NanoRobotics Laboratory, Department of Computer and Software Engineering, and Institute of Biomedical Engineering, École Polytechnique de Montréal (EPM), Montréal, QC, Canada

Synonyms

Bioinspired and bacterial nanorobotics; Medical nanorobotics; Nanorobotic platforms for biological studies; Robotic platforms for nanomanipulations

Definition

Nanorobotics for bioengineering is the use of robotic platforms capable of nanometer resolution, nanoscale robots, or microscale robots relying on nanoscale components and phenomena being applied to living systems.

Overview

Robotics

The field of nanorobotics bridges nanotechnology with robotics. As such, to understand the fundamentals of nanorobotics, one must first know the basic elements being associated with robotics and nanotechnology.

While robotics would be defined as the science or study of the technology associated with the design, fabrication, theory, and application of robots, a robot would typically be referred to as a machine or device that operates automatically or by remote control. As such, a robot must have an appropriate structure suitable for its working environment while being able to hold all embedded components and will include a power source (embedded, wired, or wireless), actuation (e.g., for displacement, grasping, etc.), sensing (vision, pressure, temperature, etc.), and control or computation (embedded or remote). Depending on the tasks and requirements, communication may also be involved.

Nanotechnology

Nanotechnology on the other hand is often defined as the understanding and control of matter at dimensions of roughly 1–100 nanometers (nm), where unique phenomena enable novel applications. Nanotechnology involves imaging, measuring, modeling, and manipulating matter at this length scale. It is important to note that at the nanoscale, the physical, chemical, and biological properties of materials differ from the properties of matter at the macroscale, and this fact has motivated research and development in nanotechnology to be directed toward understanding and creating improved materials, devices, and systems that exploit these new properties.

Nanorobotics

So far, researches in nanorobotics are being done within a range delimitated by two extremes. The bottom one is devoted to the design of robots with overall dimensions in the nanometer scale, while the other involves interactions with nanoscale objects using macroscopic instruments such as scanning probe microscopes (SPMs) used for automated nanomanipulation. Indeed, nanotechnology and nanoscience began in the early 1980s in great part due to the invention of the scanning tunneling microscope (STM) [1] that allowed scientists to observe atomic structures. Later, other forms of SPMs such as the well-known atomic force microscope (AFM) [2] appeared allowing more tasks to be performed at the atomic and molecular scales. Soon, engineers put efforts at making SPM-based tasks remotely [3] and later automatically [4] with more precision by integrating robotic principles, and, hence, these efforts have led among the first known applications of nanorobotics.

Although intensive research efforts were done in the various disciplines required to enhance the functionality and performance of these macroscale nanorobotic platforms, other efforts aimed at designing robots with overall dimensions in the nanometer scale. But although some experimental principles in the form of primitive molecular machines have been shown, the field of creating real practical machines or robots with overall dimensions in the nanometer range (e.g., 1–100 nm) is still beyond present technological limits.

As such, in an attempt to design the smallest practical robots, some research efforts concentrated on nanorobots with overall dimensions in the micrometer (μm) scale. These microscale nanorobots typically differ from the more classic microrobots by the fact that unlike the latter, the formers exploit the unique properties of nanotechnology by embedding the appropriate nanoscale components to integrate within the robots, the required functionalities to allow them to accomplish their assigned tasks.

Previous experimental results have also shown that the implementation of these so-called microscale nanorobots can be entirely artificial

(synthetic) including biologically inspired nanorobots, natural including bacterial nanorobots (e.g., bacteria controlled by the computer to perform the same tasks as futuristic artificial microscale nanorobots), or hybrid (also sometimes referred to as bio-nanorobots), i.e., built with a mix of artificial or synthetic and biological components.

Bioengineering

Although several applications for nanorobotics have been envisioned including but not limited to nano-assembly tasks within nanofactories, by observing the more recent experimental results it becomes obvious that the field of nanorobotics has great potential in providing significant contributions for instrumentation in bioengineering.

Bioengineering [5] can be defined as the application of engineering design and technology to living systems. As such, bioengineers are concerned with the application of engineering sciences, methods, and techniques to problems mainly encountered in medicine and biology. Although macroscale nanorobotics as defined earlier can play a great role for biology, microscale nanorobots due to their overall dimensions could be particularly suited for in vivo applications in medicine. These two main approaches in nanorobotics applied to bioengineering are described in more detail in the following sections.

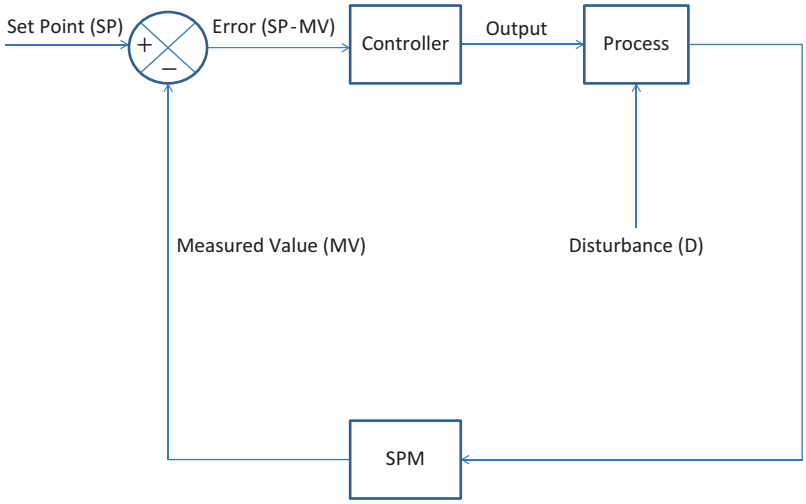
Macroscale Nanorobotics in Bioengineering

Macroscale nanorobotic platforms can be generally categorized as contact-mode nanomanipulation systems using SPM and non-contact-mode systems such as the ones using techniques such as optical trapping.

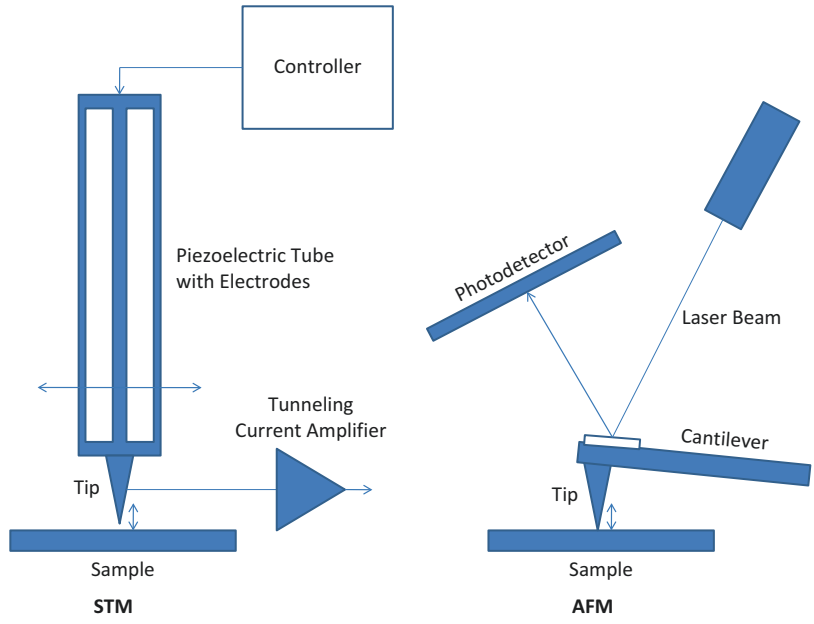
SPM-Based Macroscale Nanorobotics

In recent years, SPMs were used to image and manipulate biological samples, and as such, it had a tremendous impact on the field of biology. Indeed, SPMs can be used to gather intrinsic properties of individual molecules as opposed to just bulk properties of larger samples. Therefore, since SPM technology has become an invaluable tool for the understanding of biological structures and processes at the nanoscale, it was adopted by the nanorobotics community for providing nanoscale sensing and manipulation in their platforms. The basic principle of these macroscale nanorobotic platforms is depicted in Fig. 1.

Figure 1 shows a closed-loop control system based on negative feedback. Here, the controller implementing a given control algorithm computes an output signal that will typically be used to actuate or move a SPM tip (i.e., the interfacing point between the SPM and the sample) at a desired location or set point (SP). As such, the controller corrects for any drifts resulting from the



Nanorobotics for Bioengineering and Medical Interventions, Fig. 1 Block diagram showing the basic principle of SPM-based macroscale nanorobotic platforms



Nanorobotics for Bioengineering and Medical Interventions, Fig. 2 Basic functional principles of STM and AFM

difference between the measured value (MV) from the SPM and the set point (SP). These drifts or errors may be caused by the SPM itself (e.g., hysteresis and nonlinear characteristics of the piezoelectric (PZT) actuators used to move the SPM tip) and by various disturbances in the process. Such closed-loop control process is repeated at a rate sufficiently high to avoid excessive drifts beyond an acceptable error margin. The output computed by the digital controller is converted to an analog form using a digital-to-analog converter (DAC) and amplified to drive the actuator acting on the SPM which in most cases will be piezoelectric (PZT) actuators which are capable of atomic-scale resolution.

Although there is a huge variety of SPMs, so far most development projects of macroscale SPM-based nanorobotic systems used STM or AFM, and as such, they will be emphasized here. The type of SPM and the mode used for biological samples are of critical importance. For instance, AFM has been particularly popular in biology since it has the unique capability of imaging biological samples with molecular resolution in buffer solution. Not only topographical images of surfaces with nanometer- to angstrom-scale resolution can be achieved, but the forces between single molecules and mechanical properties of biological samples can be investigated as well. Since the measurements can be made in buffer solutions, it allows researchers to maintain biological samples alive within a physiological-like environment. This allows temporal changes in nanostructures to be measured such as comparing structural changes before and after the addition of chemical reagents. These advantages distinguish AFM from conventional imaging techniques of comparable resolution, including electron microscopy (EM) which will be discussed later. Thus, the attractive potential of SPMs and particularly AFM for biologists is the ability to visualize molecular processes under natural or physiological conditions. As such, AFM offers the resolution of most commercial electron microscopes but under the experimental conditions familiar with the much lower resolution of optical microscopes. For these reasons, AFM has been the most popular SPM in nanorobotics for bioengineering applications. The basic principle of AFM compared to STM is depicted in Fig. 2.

As shown in Fig. 2, the method used for AFM differs from STM. For STM, a controller generates voltage outputs to electrodes on a piezo-tube in order to deflect the tip in any directions. When

the tip is positioned close enough to a sufficiently conductive sample, electrons from the atoms on the surface of the sample are attracted to the positively charged tip resulting into a tunneling current which is then amplified to create an output analog signal that will be converted to digital form by an analog-to-digital converter (ADC). This digital signal is then fed back to the digital controller that will compute the voltage outputs to move the tip in order to perform a scan of the surface of the sample using closed-loop control as depicted in Fig. 1. The changes in the magnitude of the tunneling current can be used to create an image of the atomic structure, or the tip can be positioned to maintain the tunneling current constant. In the latter case, the position of the tip or nanorobotics is controlled as depicted in Fig. 1 and is used to create the image of the atomic structure.

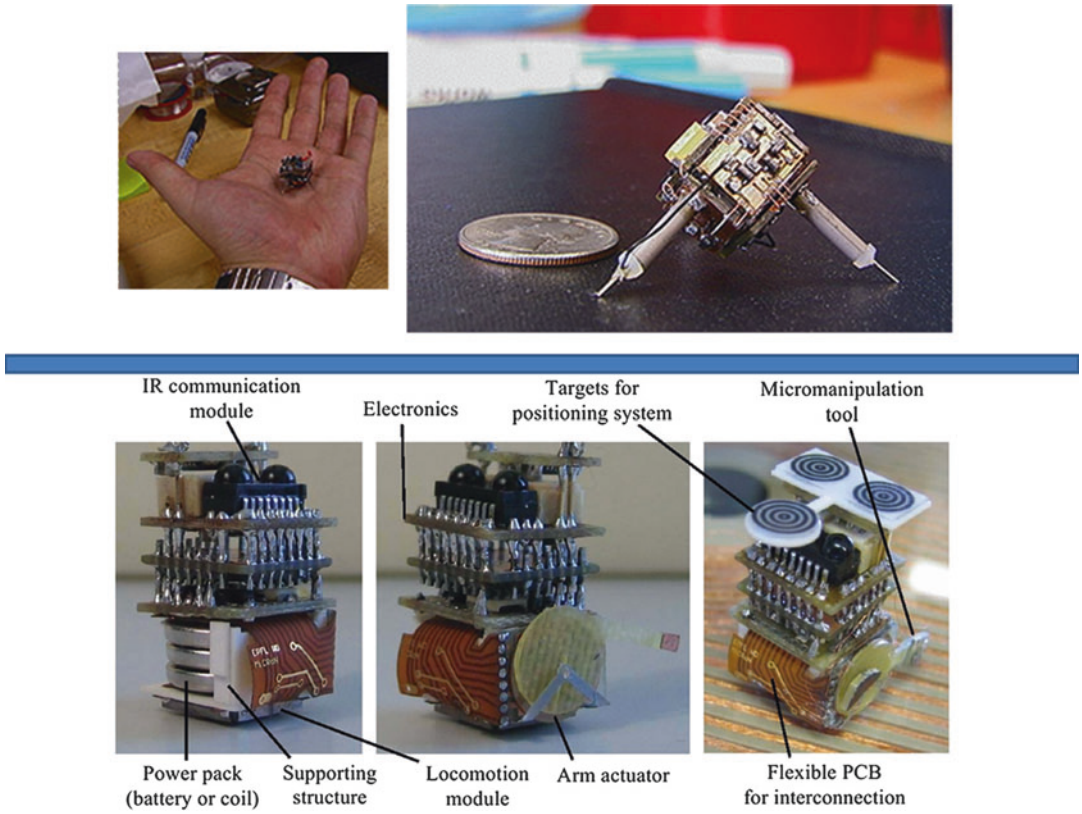
For AFM, a cantilever with a tip is used instead. When the cantilever is deflected due to the interaction between the tip and the sample, a laser beam is also deflected along a photodetector. The position of the laser beam on the photodetector provides information about the interaction between the tip and the sample that can be used by a controller to perform closed-loop control in a macroscale nanorobotic platform. Such information can not only be used for generating an image point by point, but it also provides information such as the force involved, etc. The latter would prove to be useful to characterize the mechanical properties of biological samples and in performing nanomanipulation tasks under robotic closed-loop control such as the nanomanipulation of single DNA molecules [6].

Research and development of macroscale nanorobotic systems has taken two main approaches that could be referred to as desktop SPM-based platforms and platforms based on miniature robots carrying a SPM. Instead of bringing the samples to the instrument as the former suggests, the latter proposes to bring the instruments to the samples, allowing the potential for higher throughput through parallelism and enhanced scheduling or dispatching of each instrumented miniature robots. The concept of a nanorobotic platform relying on the dispatching of SPM-based miniature robots to the samples was first proposed and introduced with a robot known as the NanoWalker [7, 8]. Photographs of the first prototypes of the NanoWalker are being shown in the top section of Fig. 3.

The NanoWalker project began in the late 1990s. Equipped with three piezo-legs mounted in a pyramidal fashion with the apex pointing upward, the miniature robot could move with nanometer precision using approximately 4000 steps per second (being equivalent to the resonant frequency of the robot). The maximum recording displacement speed was approximately 200 mm per second. The final design of the NanoWalker robot was equipped with an onboard computer with infrared (IR) wireless communication. It was powered through the legs when in contact with a special power floor. Although it was designed to have several types of SPMs, the initial design included only an embedded STM. With an STM, the NanoWalker robot, although it could later be upgraded with an AFM, was not initially suited for bioengineering applications.

This issue was corrected later with the MiCRoN robot [9] (see bottom of Fig. 3) capable of maximum displacement speed of 0.4 mm per second while being powered with batteries or wirelessly using inductive coupling. The goal of this project was to make automated manipulation with nanometer precision of micrometer-scale objects possible by these robots. Bioengineering applications included closed-loop force control for an embedded AFM tool in cell manipulation [10].

But to date, demonstrations of bioengineering applications for both types of macroscale nanorobotic platforms are still relatively few. On the other hand, efficient automatic micrometer-scale manipulations in bioengineering have been demonstrated for desktop versions especially in autonomous cell injection [10] where micromanipulation strategies required for manipulating individual biological cells with a microrobotic system capable of conducting automatic embryo pronuclear DNA injection (cell injection) were described. One of the main reasons explaining that micromanipulation instead of nanomanipulation has generally been applied more



Nanorobotics for Bioengineering and Medical Interventions, Fig. 3 *Top*: Photographs showing two different prototypes of the NanoWalker robot. *Bottom*: Photographs of the MiCRoN robot

successfully in robotics is the fact that autonomous nanoscale operations is much more challenging due in great part to the accuracy required for successful operations coped with the difficulty in gathering real-time visual feedback information for the essential closed-loop control used in robotics.

To improve performance in gathering real-time visual feedback information, SPM-based microscale mobile nanorobots were placed in the vacuum chamber of a scanning electron microscope (SEM) [11]. The SEM electron beam was used to provide real-time 3D positional information about the tool used for nanomanipulation. Applications in bioengineering were limited since the samples needed to be electrically conductive and placed in a vacuum environment. Living cells, tissues, and soft-bodied organisms usually required chemical fixation to preserve and stabilize their structure.

Non-contact-Mode Macroscale Nanorobotics

Non-contact nanomanipulation is presently dominated by optical trapping using optical tweezers. Optical tweezers use the forces exerted by a strongly focused beam of light to trap and move objects ranging in size from tens of nanometers to tens of micrometers [12]. Since its introduction in 1986, the optical tweezer has become an important tool for research in the fields of physical chemistry, soft condensed matter physics, and biology. Since optical tweezers apply loads in the order of piconewtons (pN) in a non-contact mode, it is widely used in bio-nanomanipulation and mechanical studies of very fragile biological samples. For instance, single molecule manipulation combined with interaction study of double-stranded DNA [13] is only one example. Nanorobotic platforms relying on optical tweezers generally operate in a manner to assist a human

operator. Furthermore, unlike SPMs, optical tweezers are generally more difficult to integrate in miniature robots, and, hence, they are typically more appropriate for desktop macroscale nanorobotic platforms.

Microscale Nanorobotics in Medical Interventions

Because of their overall dimensions, research and development efforts for microscale nanorobotics applied to medical interventions have been focused primarily toward in vivo target therapeutic interventions and diagnostics. Nanorobotic R&D efforts for in vivo applications were initiated in great part following the first demonstration of the MR-Sub (Magnetic Resonance Submarine) project where an untethered object was successfully navigated automatically for the first time in the blood vessels of a living animal [14] with the experimental data published in 2007. Although this field is relatively new, substantial advanced experimental results have been recently published, providing important insights about the future possibilities for medical interventions in the human body. But as shown in the following sections, several types of microscale nanorobots have been and still been investigated for such bioengineering applications.

Artificial and Synthetic Microscale Nanorobotics: An Introduction

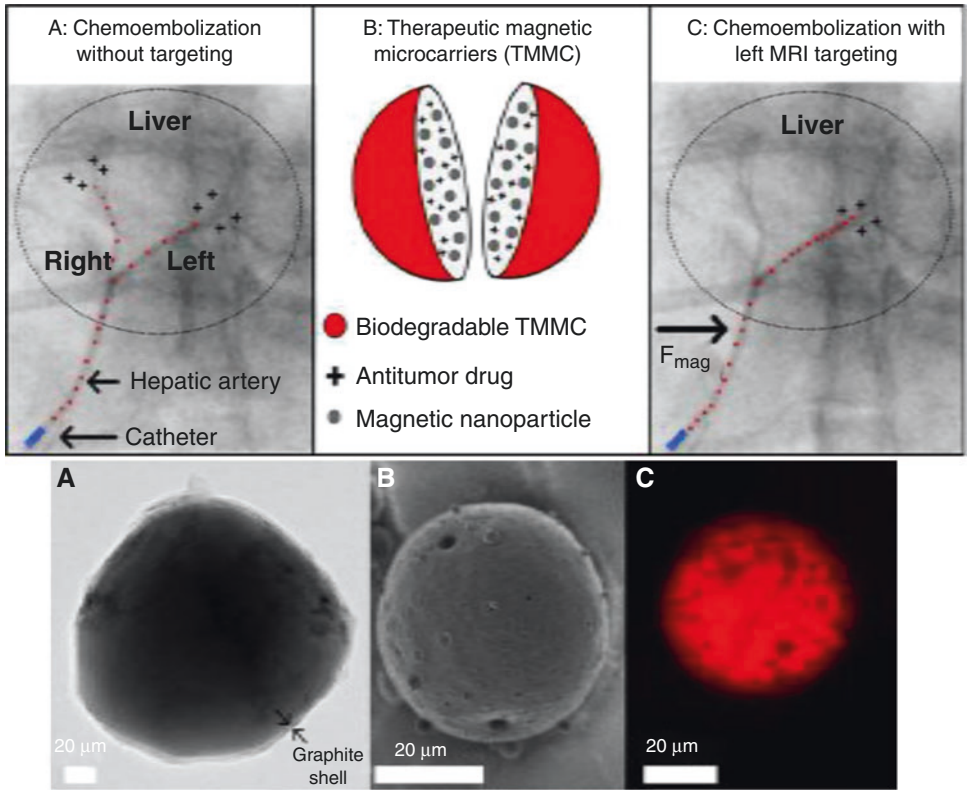
Although both artificial and synthetic robots refer to man-made robots, synthetic robots, although artificial, often imply unlike non-synthetic artificial robots the use of a chemical process to produce such robots. Hence, synthetic microscale nanorobots can typically be produced in large quantities at a relatively lower cost compared to the production of non-synthetic artificial (later simply referred to as artificial) microscale nanorobots.

The concept of integrating processes based on self-assembly for the synthesis of synthetic microscale nanorobots is an important yet critical aspect since for many targeted nano-medical interventions, large quantities of the so-called synthetic microscale nanorobots produced at a relatively low cost would typically be required, e.g., in targeted drug delivery through the vasculature in cancer therapy. On the other hand, one or a relatively small number of artificial microscale nanorobots could be useful in other particular medical interventions such as in surgical tasks.

But at these scales, one of the primary challenges shared by both artificial and synthetic microscale nanorobotics is power. Traditional onboard power supplies such as batteries cannot be scaled at these levels, and, therefore, two main approaches have been considered, namely, relying on an external source of energy or harvesting energy from the environment. For artificial and synthetic microscale robots, the former has been preferred where in most cases a magnetic field gradient generated by electric coils and induced on the microscale robots was used to create motion.

Synthetic Microscale Nanorobotics: Therapeutic Magnetic Micro-carriers (TMMCs)

The first untethered examples based on nanometer-scale components and depicted in Fig. 4 that were navigated through the blood vessels toward a predefined target in a living animal using the same method previously reported in Ref. [14] with a larger untethered device being propelled by magnetic gradients have been recently described in Ref. [15]. Referred to as therapeutic magnetic micro-carriers (TMMCs), these untethered microscale agents were synthesized to act like synthetic microscale nanorobots as defined previously. Indeed, each TMMC had an appropriate structure with a diameter of approximately 50 micrometers which was the most suitable dimensions for targeted liver chemoembolization in live rabbits in these particular experiments. For humans, the diameter would typically be expanded between 100 and 300 micrometers. The structure was built with a biocompatible and biodegradable polymer that was able to hold all embedded components, namely, the magnetic nanoparticles (MNPs) and the therapeutic agent. The embedded MNPs allowed propulsion and navigation of the TMMCs in the blood vessels by inducing

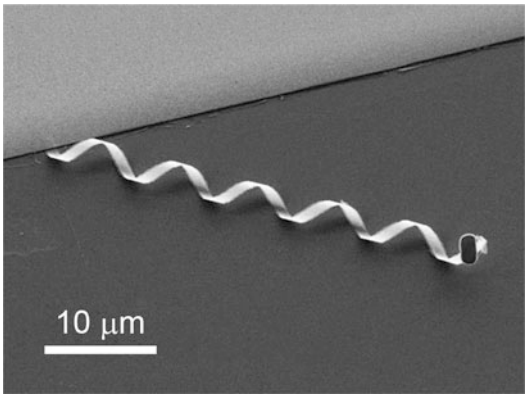


Nanorobotics for Bioengineering and Medical Interventions, Fig. 4 Unlike simple passive micro-entities that would be injected in the bloodstreams without control, TMMCs (B-upper) are synthesized to be compatible and integrated within a robotic environment allowing them to be controlled precisely as required to perform a given task such as the precise controlled navigation and delivery of therapeutic agent to a specific targeted location (C-upper vs. A-upper). In the lower section, photographs of the TMMC (B), with image of the loaded drug (C), are shown with one of the MNPs (A) allowing closed-loop MRN to be performed by inducing the correct directional propelling force on the embedded MNPs from MR-tracking information gathered from distortions in the MRI homogeneous magnetic field caused by the same MNPs (Reprinted from Pouponneau et al. [15] with permission)

propelling force using magnetic gradients generated by an upgraded clinical magnetic resonance imaging (MRI) system, a method referred to as magnetic resonance navigation (MRN). The same MNPs were also acting as MRI contrast agents, and as such, they enabled the TMMCs to be tracked by MRI, providing the possibility for feedback information required to execute closed-loop navigation control. Like other forms of robotic systems such as the ones relying on a central computer to remotely control a robotic arm or manipulator, this approach also relies on computation to remotely control special agents designed to perform a particular task. While the larger and more familiar robots often use mechanical means to embed functionalities (e.g., mechanical grippers), microscale nanorobots for bioengineering would not typically use mechanical parts, and as such, they would not appear as conventional robots as we know since they rely more on less familiar fields such as nanotechnology and biochemistry to implement embedded functionalities (e.g., with the use of antibodies instead of mechanical grippers).

But with MRN, the propelling force induced decreases with the quantity of magnetic material embedded in the microscale nanorobots, and, hence, larger magnetic gradients must be generated to navigate smaller nanorobots efficiently in tinier blood vessels. In an upgraded clinical MRI

platform, an approximate tenfold increase in magnetic gradients is technologically feasible from conventional scanners, increasing the typical 40 milli-teslas (mT) per meter (m) 3D gradients of many conventional MRI systems up to approximately 400 $mT \cdot m^{-1}$ in upgraded MRI scanners (referred to as MRN systems, i.e., upgraded MRI scanners by software only or a combination of software and hardware for MRN operations). Despite the increase in magnetic gradients, miniaturization of the microscale nanorobots conceived for MRN-based interventions is still limited to a few tens of micrometers in diameters. To navigate in the microvasculature and in the tiniest blood vessels found in humans and required to target deeper inside tumors, microscale nanorobots with an overall diameter of approximately 2 micrometers are required. Therefore, in order to compensate for the limitation of generating larger magnetic gradients for interventions in humans, other approaches have been investigated.



Nanorobotics for Bioengineering and Medical Interventions, Fig. 5 Photograph of an artificial bacterial flagellum about half as long as the thickness of a human hair. This one can swim at a speed of up to one body length per second (Image: Institute of Robotics and Intelligent Systems/ETH Zurich. Image reprinted with permission)

Biologically Inspired Artificial Microscale Nanorobotics: Artificial Bacterial Flagella (ABF)

One of the best known approaches to compensate for the limitation in generating larger magnetic gradients is referred to as artificial bacterial flagella (ABF) where a good up-to-date review is given in Ref. [16]. ABF (one example is depicted in Fig. 5) are good examples of biologically inspired artificial devices that can act under computer closed-loop control. Indeed, in nature, microorganisms swim in a low Reynolds number regime using a variety of techniques that differ from macroscale swimmers. The most studied experimental technique in micro- and nanorobotics due in great part to the possibility in micro- and nanofabrication technologies has been flagellated propulsion systems.

ABF typically rely on a rotating magnetic field creating a torque instead of a pulling force as for MRN to move the robot forward. Since creating a torque requires much lower magnitudes of magnetic gradients compared to pulling as done during MRN operations, the gradients required could be decreased to a few $mT \cdot m^{-1}$ instead of a few hundred $mT \cdot m^{-1}$. But to date, no known experiments have been done yet in the blood vessels using ABF. Although the approach has advantages, several issues would have to be addressed before being adopted for targeted interventions through the human's vasculature. Two critical aspects are the biodegradation and response to the immune system since once located deep in the human's body, possibility for recovery will be unlikely. For the TMMCs described earlier, these issues have already been resolved by using materials that are biodegradable and biocompatible and MNPs with dimensions that already prove to be compatible with the human's immune system.

Another consideration for robotic operations is the possibility of gathering positional feedback data for closed-loop navigational control. The spatial resolution of all medical imaging modalities is too low for detecting a robot or an aggregation of robots with overall diameters of only a few micrometers or a few tens of micrometers, respectively, due to the space available in the microvasculature. X-ray systems have among the best spatial resolution to image inside humans (approximately 200–300 micrometers), but they will fail to detect and hence track microscale nanorobots operating in the microvasculature such as in capillaries. Although MRI has typically

a slightly lower spatial resolution than X-ray, the microscale nanorobots can be made such that the embedded magnetic particles would create a magnetic field distortion much larger than the robot itself and large enough to be visible with MRI. Past studies have shown that a single magnetic object as small as 10 micrometers could be detected inside the human's body using MRI. But because MRI scanners have a high homogeneous magnetic field referred to as the B_0 field (1.5 T or 3 T in most clinical MRI scanners) along a given direction, the shape of the robots must ideally be isotropic such as the spherical shapes of the TMMCs depicted in Fig. 4. For instance, the B_0 field would induce a directional torque on ABF if operated inside a MRI scanner such that they will always point to the same direction during the entire intervention and, hence, making directional control or steering for targeting purpose impossible. On the other hand, because of the spherical shape of each TMMC, steering in all directions is not an issue. But one fact remains true that up to now, TMMCs are still too large to be navigated in the microvasculature and, as such, the idea of using a propulsion method suited for low Reynolds hydrodynamic conditions such as flagella-based propulsion allowing for the implementation of smaller untethered devices is still a valid and promising approach.

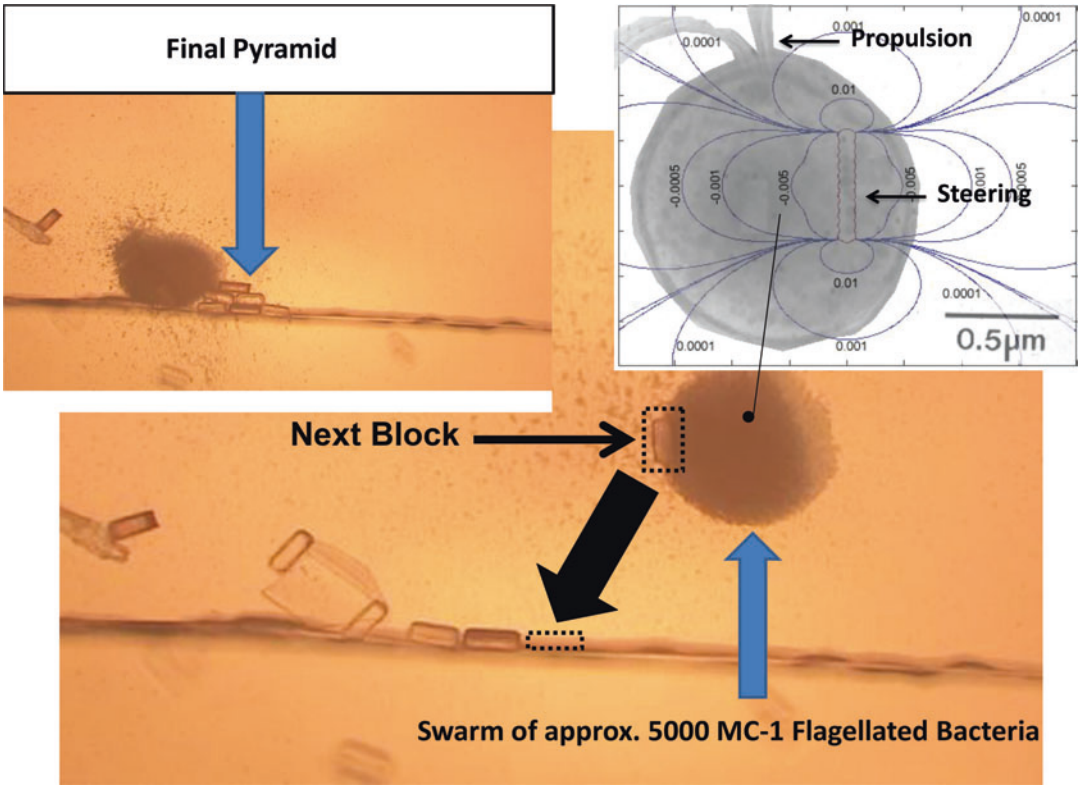
Natural Microscale Nanorobotics: Bacterial Nanorobots

One strategy that was recently considered in nanorobotics was to harness what nature provides instead of mimicking it using modern technologies. An example that showed its potential through experimentations conducted in the blood vessels of living mice is the use of MC-1 flagellated magnetotactic bacteria (MTB) acting as bacterial nanorobots or natural microscale nanorobots [17]. Although most people would see this bacterium as a simple living microorganism, in an engineering point of view, this microorganism can be considered as a sophisticated robot with a very effective embedded propulsion system and control interface. The possibilities of computer closed-loop control of these MTB compared to potential far-futuristic artificial nanorobots can be appreciated by observing the example of a tiny pyramid being built one component at a time by a swarm of approximately 5000 MC-1 flagellated bacteria (see Fig. 6).

As depicted in the upper right corner of Fig. 6, the cell of the MC-1 bacterium is spherical in shape and measures between 1 and 2 micrometers in diameter, making it ideal for navigation in the tiniest blood vessels found in humans. Each cell has two bundles of flagella providing a thrust force exceeding 4pN, making it a serious candidate for operations in the human's microvasculature. This value is also relatively high compared to other known species of flagellated bacteria characterized with a typical thrust force in the range of 0.3–0.5pN. This allows the MC-1 bacteria to swim in water and human's blood at speeds often exceeding 200 μm/s. This is a very high speed when we know that the swimming speeds of other well-known species of flagellated bacteria in the same conditions are ~30 μm/s. Maximum displacement speeds of ~300 μm/s and corresponding to ~150 times its cell's length for a relatively large proportion of the MC-1 bacterium samples have also been recorded experimentally.

Unlike most bacteria that are based on chemotaxis to detect nutrient gradients and hence influence their motility, the direction of displacement of MTB with their chain of magnetosomes which are membrane-based MNPs that acts like an embedded nano-compass can be influenced by a directional magnetic field, hence allowing for magnetotaxis-based control by inducing a directional torque on the same chain acting as a nano-steering system. Since this magnetic nano-compass is extremely sensitive, an extremely low-magnitude magnetic field is required to influence their directional motions.

Since the same MNPs used for directional control also act as MRI contrast agents, it was shown that a sufficiently large aggregate of MTB can be detected using a clinical MRI scanner. But as for the ABF, MTB cannot change direction while operating in the MRI scanner. But unlike artificial or synthetic microscale nanorobots, these microorganisms do not need real-time closed-loop navigation control but rather computer guidance for determining their positions using MRI at relatively



Nanorobotics for Bioengineering and Medical Interventions, Fig. 6 Computer control of a swarm of approximately 5000 magnetotactic bacteria is demonstrated by the coordinated construction of a tiny pyramidal structure using blocks with length approximately equivalent to a human's hair thickness

long time intervals in order to guide them toward the targeted site. Indeed, MTB when encountering an obstacle will not require real-time closed-loop navigation control to direct them around the obstacle as an artificial robot would, but they will typically swim without any external control around obstacles and toward the direction indicated through magnetotaxis using a special platform named a magnetotaxis system, a platform located outside and sufficiently far from the MRI scanner. This self-contact avoidance capability is a significant advantage compared to artificial versions since real-time closed-loop navigation control is not possible in these regions considering that the microvasculature cannot be imaged due to a lack of spatial resolution in existing medical imaging modalities. Nonetheless, although these bacteria can swim efficiently in the microvasculature, they are ineffective in larger blood vessels such as the arteries and the arterioles where blood flow velocities are too high. Hence, a means of transport is needed to bring such bacterial nanorobots from the injection location in an artery to the entrance of the microvascular network.

Hybrid Microscale Nanorobotics: Bio-nanorobots

One solution that is under study for the transport of bacterial nanorobots toward the microvasculature is the use of microscale liposomal transporters. Already, such transporters with diameters of a few tens of micrometers have been synthesized with free swimming MC-1 bacteria inside. Although not tested yet in a living animal, these liposomal microscale transporters can be synthesized with MNPs acting as propulsion engines and for MRI-based tracking for MRN operations. This is a perfect example of

hybrid microscale nanorobots where biochemistry and biology are used in their implementations. Since biological components are also used, they could also be referred to as bio-nanorobots. Bio-nanorobots could also refer to artificial or synthetic robots relying on nanostructures such as MNPs being functionalized with biomolecules. Indeed, often in targeted therapeutic or diagnostic applications, microscale nanorobots would be functionalized, i.e., that they will be coated with molecular components, e.g., antibodies that have specificity for the targeted cells such as specific cancerous cells. Functionalized hybrid or bio-nanorobots could not only help in achieving more retention when coming in contact with the targeted cells for targeted therapies, but functionalized microscale nanorobots could also be potentially useful in targeted diagnostics as well.

Magnetic Field Actuation Approaches in Medical Nanorobotics

Magnetism is presently the most popular and suitable means of actuation for medical nanorobotics. With magnetic actuation, it is important to understand that both the magnetic field strength which is the amplitude of the magnetic field at the location of the nanorobots being actuated and the magnetic gradient which is the variation of the magnetic field at the location of the nanorobots that must be moved are important parameters to consider. Increasing the field strength results in an increase in the magnetization level of the nanorobots up to a maximum level known as the saturation magnetization. The magnetic gradient is used to move the nanorobots in a given direction. The higher is the magnetization of the nanorobots, the higher is the force induced on the nanorobots from the magnetic gradient. Therefore, ideally both the magnetic field strength and the magnetic gradient must be increased. But there are technological limitations that lead to compromises in the form of various magnetic field actuation approaches where each one would be better suited for particular physiological conditions and/or types of medical interventions.

When using one or several permanent or electromagnets around the patient, the magnetic field strength decays rapidly as the distance for the magnetic source increases. For whole-body interventions where nanorobots must operate in deep tissues and especially in constrained physiological environments where the size of the nanorobots becomes very small, the magnetic field strength becomes very small and insufficient to induce sufficient directional actuation forces from the magnetic field gradient. As such, increase in the magnetic field strength has been provided with MRN as specified earlier by exploiting the high uniform magnetic field of a clinical MRI scanner. In modern MRI scanners such as the 3 T-platforms, the magnetic field strength is sufficient to bring the best magnetic nanorobots at saturation magnetization at any depth in the body of a patient. But the main drawback is that clinical MRI scanners offer limited magnetic directional gradients. Such gradients have been increased in MRI scanners by inserting a gradient coil assembly inside the tunnel of the scanner, but by doing so, the reduced inner diameter prevents whole-body interventions when using conventional clinical scanners and becomes more appropriate for parts of the body that can fit in such restricted space such as the head.

To maintain the advantage of a high magnetic field strength while providing much higher magnetic gradients, another approach known as dipole field navigation (DFN) is presently under investigation [18]. The basic principle is that when a magnetic object such as a sphere is placed in the tunnel of a clinical MRI scanner, it creates a dipole field that distorts the high uniform magnetic field inside the tunnel of the scanner. Such distortion results in a variation of the magnetic field that can be used as directional magnetic gradients. With DFN, nanorobots are submitted to both high magnetic gradients and high magnetic field strength. The drawback of this approach is that because the uniform field is distorted, real-time navigation control based on feedback positional data during travel is not presently possible as for MRN. There are two main approaches in DFN, namely, static and dynamic DFN. For static DFN, one or more magnetic spheres are positioned prior to the navigation phase and remain statically positioned at specific locations necessary to create the required set of directional gradients based on previous

acquired images of the vascular network. For dynamic DFN, the set of magnetic gradients is achieved by dynamically moving a magnetic sphere along a specific path outside the patient typically by inducing a pulling force using the scanner's imaging gradient coils. A hybrid version is also possible where a magnetic sphere is moved in the neighborhood of statically positioned magnetic spheres.

Another approach provides a compromise in magnetic field strength between the use of permanent or electromagnets alone and MRN or DFN, but with much higher magnetic gradients. To achieve such higher magnetic gradients, superconducting magnets as opposed to permanent or electromagnets must be considered. In a clinical setting, such superconducting magnet is in the form of a clinical MR scanner where superconductivity is used to generate the high uniform field inside the tunnel of the scanner. But such magnetic field cannot be modulated and must be maintained constant. This is the reason that imaging gradients in a MRI scanner are generated by coils that are not superconducting. In proximity of the scanner, the high magnetic field generated by the superconducting magnet decays rapidly which results in a very high magnetic gradient and a relatively high field strength when placed in proximity of the scanner. But since such high magnet cannot be moved and the field cannot be varied for navigation purpose, the patient is robotically moved instead outside and in proximity of the scanner based on the computed directional gradients required. This is the fundamental idea behind an approach that has been referred to as fringe field navigation (FFN) [19]. The actual main drawbacks of this approach are the relatively slow directional changes of the gradients and a lack of an effective real-time tracking modality.

Medical Nanorobots with Additional Functionalities

Embedding more functionalities in each nanorobotic agent would lead to higher impacts on a wider range of medical interventions. But embedding additional functions in such small entities becomes a real challenge due to the fact that adding more components is often not an option. One strategy has been to integrate components with characteristics suitable for the implementation of several functions. Magnetic nanoparticles, for instance, synthesized adequately can not only be used for inducing displacement forces and acting as a contrast agent for tracking the position or location of the nanorobots by MRI but can also create a localized elevation of the temperature (known as hyperthermia) to enhance treatment [20] or for other functions such as temporarily opening the blood-brain barrier (BBB) [21] or computer-triggered drug releases by changing the shape of the nanorobotic agents [22]. These are just some examples of several functions that can be implemented in such nanorobotic agents.

Conclusions

The two main forms of nanorobotic platforms for practical applications in bioengineering are the macroscale nanorobotic platforms and the microscale nanorobotic platforms which are mainly dedicated to biology and in vivo medical interventions, respectively. Macroscale nanorobotic systems rely mainly on AFM for contact-mode nanomanipulation and studies of biological samples, while optical tweezers are mostly used for non-contact applications. Microscale nanorobots can be artificial, synthetic, or natural. Synthetic versions called TMMCs based on biodegradable polymer and MNPs and capable of releasing therapeutic agents have already been navigated to target sites in live animals using an upgraded clinical MRI scanner, but the miniaturization of these synthetic carriers is limited to a few tens of micrometers in diameter due to limitation in increasing magnetic gradients beyond approximately half a tesla for human subjects. To compensate for such limitation, artificial versions called ABF using a rotary magnetic field to induce a torque to mimic the flagellum of bacteria have been proposed. Although the magnitude of the magnetic gradients could be reduced significantly, several issues would need to be resolved before ABF can be used for in vivo interventions in the blood vessels. Instead of a biomimetic artificial version, another strategy relies on harnessing

natural bacteria and, more specifically, the MC-1 flagellated magnetotactic bacteria. But the latter would need to be encapsulated in special microscale carriers for many target medical interventions. As such, hybrid microscale nanorobots are likely to play an important role in medical interventions conducted deeply in the vasculature, especially hybrid functionalized nanorobots.

This entry only provided a quick overview of the present state of the art for nanorobotics in bioengineering, and for more details, the readers should consult the adequate literature. Research and development in the field of nanorobotics is progressing at a very fast pace where significant progresses have been observed particularly in medical applications where efforts have been oriented toward the developments of medical nanorobots with enhanced capabilities.

Cross-References

- ▶ Biomimetics
- ▶ Biosensors
- ▶ CMOS MEMS Biosensors
- ▶ Dielectrophoresis
- ▶ DNA Manipulation Based on Nanotweezers
- ▶ Liposomes
- ▶ Magnetic Nanoparticles for Biomedical Applications
- ▶ Nanogrippers
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Nanorobotic Assembly
- ▶ Nanorobotic Manipulation of Biological Cells
- ▶ Nanorobotics
- ▶ Nanotechnology
- ▶ Optical Tweezers
- ▶ Robot-Based Automation on the Nanoscale
- ▶ Scanning Tunneling Microscopy

References

1. Binnig, G., Rohrer, H.: Scanning tunneling microscopy. IBM J. Res. Dev. **30**, 4 (1986)
2. Binnig, G., Quate, C.F., Gerber, C.: Atomic force microscope. Phys. Rev. Lett. **56**, 930–933 (1986)
3. Sitti, M., Hashimoto, H.: Tele-nanorobotics using atomic force microscope. In: Proceedings of the International Conference Intelligent Robots and Systems, Victoria, BC, Canada, pp. 1739–1746 (1998)
4. Mokaberi, B., Yun, J., Wang, M., Requicha, A. A. G.: Automated nanomanipulation with atomic force microscopes. In: Proceedings of IEEE International Conference on Robotics and Automation, Rome, Italy, pp. 1406–1412 (2007)
5. Rege, K.: Methods in Bioengineering: Nanoscale Bioengineering and Nanomedicine, p. 332. ISBN: 978-1-59693-410-8
6. Hu, J., Zhang, Y., Li, B., Gao, H.B., Hartmann, U., Li, M.Q.: Nanomanipulation of single DNA molecules and its applications. Surf. Interface Anal. **36**(2), 124–126 (2004)
7. Martel, S., Madden, P., Sosnowski, L., Hunter, I., Lafontaine, S.: NanoWalker: a fully autonomous highly integrated miniature robot for nano-scale measurements. In: Proceedings of the European Optical Society (EOS) and SPIE International Symposium on Envirosense, Microsystems Metrology and Inspection, vol. 382, Munich, Germany, p. 12 (1999)
8. Martel, S., Hunter, I.: Nanofactories based on a fleet of scientific instruments configured as miniature autonomous robots. J. Micromechatronics **2**(3–4), 201–214 (2004)
9. Brufau, J., Puig-Vidal, M., López-Sánchez, J., Samitier, J., Driesen, W., Breguet, J.-M., Snis, N., Simu, U., Johansson, S., Gao, J., Velten, T., Seyfried, J., Estaña, R., Woern, H.: MICRON: Small autonomous robot for cell manipulation applications. In: Proceedings of the IEEE International Conference on Robotics and Automation, Barcelona, Spain (2005)
10. Sun, Y., Nelson, B.: Biological cell injection using an autonomous microrobotic system. Int. J. Robot. Res. **21**(10–11), 861–868 (2002)
11. Fatikow, S.: Automated Nanohandling by Microrobots. Springer Series in Advanced Manufacturing, p. 346. Springer, London (2008)
12. Grier, D.G.: A revolution in optical manipulation. Nature **424**, 810–816 (2003)
13. Bennink, M., Schaver, O., Kanaar, R., et al.: Single molecule manipulation of double-stranded DNA with Reca and Yoyo-1. Cytemetry **36**, 200–208 (1999)
14. Martel, S., Mathieu, J.-B., Felfoul, O., Chanu, A., Aboussouan, E., Tamz, S., Pouponneau, P.: Automatic navigation of an untethered device in the artery of a living animal using a conventional clinical magnetic imaging system. Appl. Phys. Lett. **90**, 114105 (2007)
15. Pouponneau, P., Leroux, J.-C., Soulez, G., Gaboury, L., Martel, S.: Coencapsulation of magnetic nanoparticles and doxorubicin into biodegradable microcarriers for deep tissue targeting by vascular MRI navigation. Biomaterials **32**, 3481–3486 (2011)
16. Zhang, L., Peyer, K.E., Nelson, B.J.: Artificial bacterial flagella for micromanipulation. Lab Chip **10**, 2203–2215 (2010)

17. Martel, S., Mohammadi, M., Felfoul, O., Lu, Z., Pouponneau, P.: Flagellated magnetotactic bacteria as controlled MRI-trackable propulsion and steering systems for medical nanorobots operating in the human microvasculature. Int. J. Robot. Res. **28**(4), 571–582 (2009)
18. Latullipe, M., Martel, S.: Dipole field navigation for targeted drug delivery. In: IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob), Sao Paulo (2013)
19. Tremblay, C., Conan, B., Loghin, D., Bigot, A., Martel, S.: Fringe field navigation for catheterization. In: 6th European Conference of the International Federation for Medical and Biological Engineering. (MBEC), Dubrovnik (2013)
20. Mornet, S., Vasseur, S., Grasset, F., Duguet, E.: Magnetic nanoparticle design for medical diagnosis and therapy. J. Mater. Chem. **14**(14), 2161–2175 (2004)
21. Tabatabaei, S.N., Duchemin, S., Girouard, H., Martel, S.: Towards MR-navigable nanorobotic carriers for drug delivery into the brain. IEEE Conf. Robot. Autom. **14**, 727–732 (2012)
22. Tabatabaei, S.N., Lapointe, J., Martel, S.: Shrinkable hydrogel-based magnetic microrobots for interventions in the vascular network. Adv. Robot. **25**, 1049–1067 (2011)

Nanorobotics for MEMS and NEMS

► Robot-Based Automation on the Nanoscale

Nanorobotics for NEMS Using Helical Nanostructures

Didi Xu[1], Zhang Li[1], Lixin Dong[2] and Bradley J. Nelson[1]
[1]Institute of Robotics and Intelligent Systems, ETH Zurich, Zurich, Switzerland
[2]Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA

Synonyms

Helical nanobelt; Nanocoil; Nanohelix; Nanospring; Rolled-up nanostructure; Scrolled nanostructure

Definition

Three-dimensional helical nanostructures, such as carbon nanotube coils (CNTs) [1], zinc oxide helical nanobelts [2], and rolled-up helical nanostructures [3–7] (Fig. 1 [5, 8–10]) have attracted intensive research interest because of their potential applications in nanoelectromechanical systems (NEMS) as springs, electromagnets, inductors, resonators, sensors, and actuators [11–14]. These NEMS will serve as the tools for fabricating future nanorobots, and in the meantime basic components constructing those nanorobots. Three-dimensional helical nanostructures could be manipulated and characterized by the nanorobotic tools, which could be a good candidate for fascinating applications, such as, mass sensors in femtogram ranges, force sensors at pico-Newton scales, and resonators at GHz ranges.

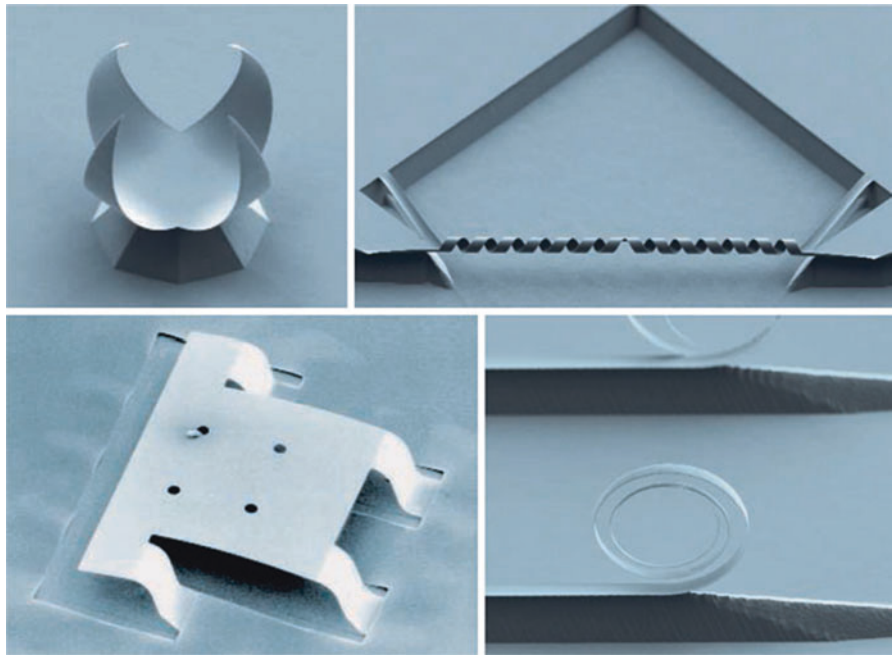
Physical and Chemical Properties

Bottom-up Fabrication of Helical Nanostructures

Top-down and bottom-up nanofabrication strategies are being independently investigated by various researchers to realize helical nanostructures. Bottom-up strategies are assembly-based techniques, resulting in controllable nanopatterns at large scales, such as carbon nanotube coils [1], silica nanosprings [15], and SiC nanohelices [16]. For example, helical nanostructures of semiconducting zinc oxide (ZnO) were grown using a vapor-solid process. A rigid structural alteration, as a result of superlattice formation [17], gives rise to the formation of a helical structure, either left-handed or right-handed, as shown in Fig. 2. The diameters, widths, and pitch distances of the nanohelix range between 300 and 700, 100–500, and 500–2,500 nm, respectively. The length of the as-fabricated ZnO nanohelix can be as long as 100 μm and the thickness of the helical ribbon is less than 20 nm [17].

Top-down Fabrication of Helical Nanostructures

Top-down approaches are based on conventional fabrication process, including nanolithography,

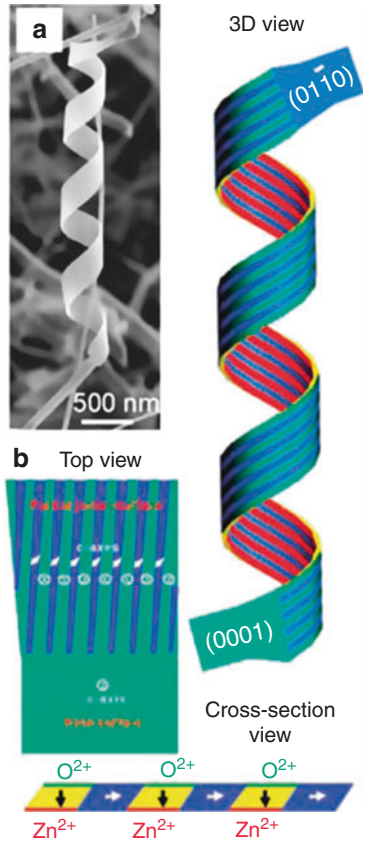


Nanorobotics for NEMS Using Helical Nanostructures, Fig. 1 Various helical nanostructures. *Bottom left* is reproduced with permission from Ref. [8], (Copyright 2003 APEX/JJAP. The others are in courtesy of L. Zhang and D. Grützmacher)

nanoimprinting, and chemical etching. Nanocoils are created through a top-down fabrication process in which a strained nanometer-thick heteroepitaxial layer of semiconductor compounds curls up to form 3-D structures [4]. Using GaAs-InAs bilayer as an example, when the basic bilayer film is released from an InP substrate by selective etching, the elastic forces of two layers are oppositely directed, generating a nonzero moment of forces, therefore its top layer contracts and its bottom layer expands, causing the film to curl and roll up [4]. The preferential direction for the film to roll up depending on the orientation of a mesa stripe on the substrate, either tube-like or helical nanobelts, can be formed [5]. A helical geometry based on the bilayer SiGe/Si can be achieved through this process (Fig. 3) [9]. The gray arrows indicate the <100> directions, that is, the smallest Young modulus direction on a <001> substrate, in which the bilayer will start to roll up when it has been freed from the substrate. Several techniques have been proposed to improve the control over the fabrication process in terms of the resulting length, shape, and orientation of structures. It has been demonstrated that the misaligned angle between the stripe and etching direction and the edge effect of the helical-shaped ribbon are crucial to determine the chirality and pitch of a helical nanobelt [5, 9]. Moreover, various materials, such as metal, dielectrics, and polymers, can be integrated into the helical nanobelt by thin film deposition and lithographic patterning [9]. It is worthy of note that nanohelices with a minimum diameter of 7 nm were obtained using 6-monolayer-thick InGaAs/GaAs bilayers [4].

Key Research Findings

Nanorobotic manipulation [18] enables property characterization to be performed after intermediate processes, and in situ characterization can be performed using manipulation rather than conventional static observations. Nanorobotic approach extends the lower limit of robotic manipulation to the dimension of nanometers, and it paves the way for the design and fabrication

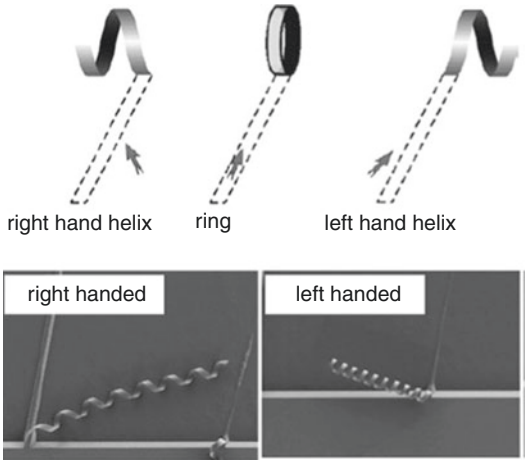


Nanorobotics for NEMS Using Helical Nanostructures, Fig. 2 Initiation and formation process of a nanohelix. (**a**) Low magnification SEM image of a ZnO nanohelix showing its starting point and finishing end. (**b**) A three-dimensional (3-D) schematic model of a nanohelix, showing its initiating point and finishing end. The periodicity of the superlattices may result in the formation of periodic piezoelectric domains (Reproduced with permission from Ref. [17], Copyright 2005 AAAS)

of nanocomponents having applications in NEMS.

Mechanical Properties of Helical Structures

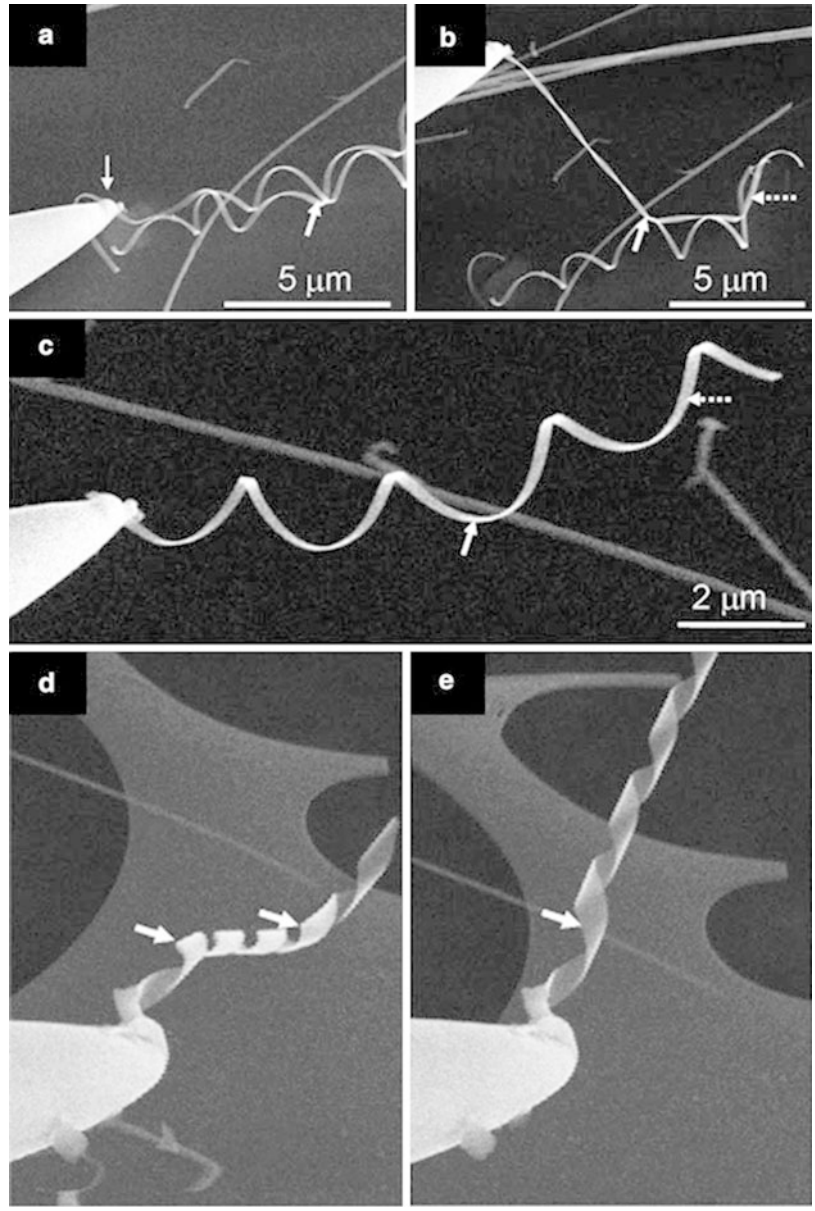
By in situ manipulation using a nanoprobe, mechanical properties of various helical structures have been investigated, such as superlattice-structured ZnO nanohelices, carbon nanocoils, and rolled-up nanosprings [11, 12, 14]. It is found that the superelasticity (shape memory) of superlattice-structured ZnO nanohelices enables the nanohelix to recover its shape after an extremely large axial stretching and compressing,



Nanorobotics for NEMS Using Helical Nanostructures, Fig. 3 Illustration of the effect of the angle between the mesa stripe and preferred <100> etch direction (*gray arrows*) on the formation of rings or helices with *right*- and *left*-handed chirality. The thickness of the SiGe/Si/Cr ribbons and the diameter of the helices are 40 nm and 3 μm, respectively (Reproduced with permission from Ref. [9], Copyright 2006 ACS. Reproduced with permission from Ref. [5], Copyright 2005 IOP)

which could be of great interest for fabricating nanoscale elastic energy storage in microelectromechanical systems (MEMS) and NEMS. The manipulation experiments of rolled up helical nanostructures can be performed by a nanomanipulator and an atomic force microscope (AFM) cantilever inside a scanning electron microscope (SEM). The nanohelix was picked up by welding its one end using platinum (Pt) deposition onto the tip of a tungsten nanoprobe, and the mechanical deformation was recorded via extending or compressing the nanohelix using the nanoprobe. As a result, its spring constant can be increased continuously for up to 300–800 %.

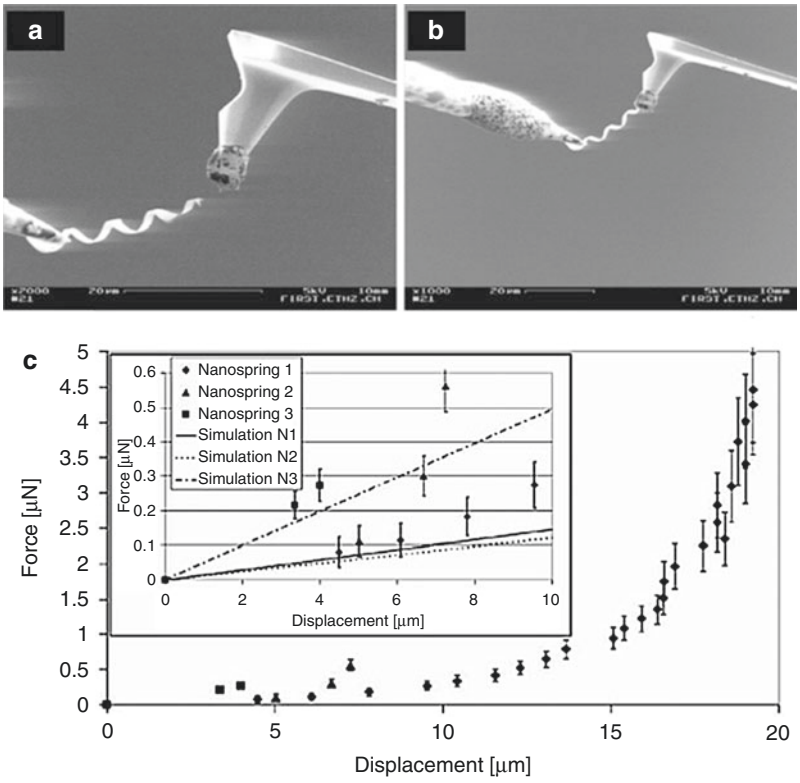
In Fig. 4a, b, it is evident that the welded nanohelix can be pulled to an almost straightened shape, with its twist cycles preserved. Next, the welded nanohelix is released and started to restore its original shape, as shown in Fig. 4c. With comparison to the initial nanohelix dimensions in Fig. 4a, it indicates that the nanohelix has an almost identical dimensionality including pitch and radius, suggesting a complete elastic recovery



Nanorobotics for NEMS Using Helical Nanostructures, Fig. 4 Manipulation process of a nanohelix during axial stretching and compressing. (**a**) One end of a nanohelix was welded with Pt onto a tungsten nanoprobe. (**b**) A SEM image shows the extremely stretched nanohelix. (**c**) A complete restoring of the nanohelix shape after releasing. (**d**, **e**) Compressed deformation process of a nanohelix induced by the nanoprobe. Both stretching and compressing suggest the superelasticity behavior (Reproduced with permission from Ref. [12], Copyright 2006 ACS)

from stretching. Similar mechanical behavior, that is, the superelasticity (shape memory) behavior, has also been observed when applying a compression load on the welded nanohelix (Fig. 4d, e). The spring constant of ZnO nanohelices is in a range of 0.07–4.2 N/m, and the transverse fracture force is in the micro-Newton range [12, 17].

Nanomanipulators can also be used to carry out the mechanical characterization of rolled-up helical structure. A probe picked up an InGaAs/GaAs nanospring and attached it to the AFM tip (Fig. 5a, b). Then, a tensile force was applied to the nanospring, while continuous SEM images were taken to detect the AFM tip displacement and the nanospring deformation. When the tensile force was increased dramatically, the attachment between the nanospring and the AFM cantilever broke, and the nanosprings returned to their initial shape at the zero-displacement position. From the displacement data and the known stiffness of the



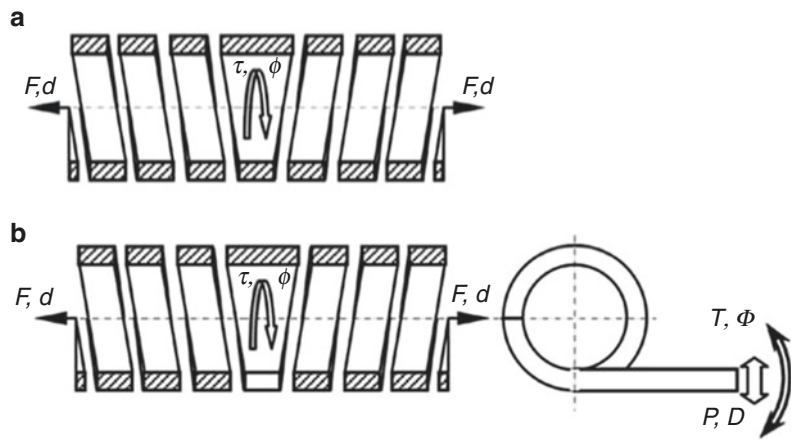
Nanorobotics for NEMS Using Helical Nanostructures, Fig. 5 Mechanical characterization. (**a**, **b**) Break and pick up nanospring and attach it to AFM tip. (**c**) Simulation and experimental results of axial force versus nanospring displacement (Reproduced with permission from Ref. [14], Copyright 2006 ACS)

AFM cantilever, the tensile force acting on the nanosprings versus the nanospring displacement was plotted (Fig. 5c). For the nanospring, an overall elastic strain of 48 % was measured when the attachment to the AFM tip broke. Ultra-flexible rolled-up Si-based nanospring with a spring constant of 0.003 N/m was also reported which exhibits long-range linear elasticity (with ca. 189 % relative elongation) and excellent mechanical stability [19]. The rolled-up structure is promising for use in electromechanical sensors, such as force or pressure sensors, because of intrinsic piezoresistive and piezoelectric properties of these materials.

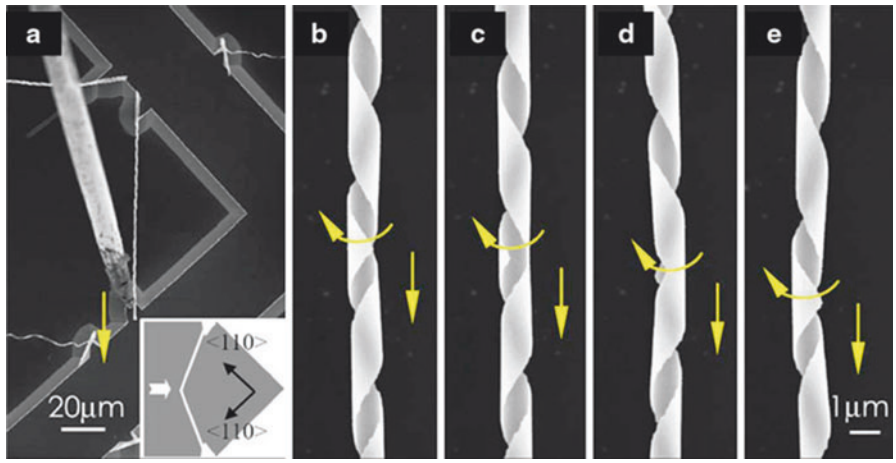
Linear-to-Rotary Motion Converter

The conversion between various forms of motions will play an important role in future NEMS applications. Among various nano building blocks, 3-D helical nanostructures, for example, dual-chirality helical nanobelts (DCHNBs), could realize conversion between linear and rotary motion [20]. As schematically shown in Fig. 6a, the motion converter consists of a DCHNB with a left-handed and a right-handed part. By linearly stretching both ends of the DCHNB, the central part was mobilized in rotary motion; allowing for linear-to-rotary motion conversion, which has not been previously demonstrated in other nanostructures. With an extended arm, the output can be linear (small displacement) or rotary (large displacement) motion, as seen in Fig. 6b. When a tensile force F is applied to the DCHNB, it elongates and rotates about the unwinding direction.

To implement and characterize motion converters, experimental investigations have been performed in the SEM using a nanomanipulator equipped with a tungsten probe and an AFM cantilever. Manipulation of an as-fabricated DCHNB was performed by cutting the lower end of the DCHNB shown in Fig. 7a. A "sticky" probe was then prepared by dipping a tungsten probe (Picoprobe, T-4-10-1 mm) into a silver tape and attached on the lower end of the DCHNB. Rotation was then generated in the central part by moving the probe downward (Fig. 7b–e).



Nanorobotics for NEMS Using Helical Nanostructures, Fig. 6 Motion converter based on the dual-chirality helical nanobelts (*DCHNB*). (**a**) Linear-to-rotary motion converter using a DCHNB. (**b**) Transmission converting linear motion to extended linear (small displacement)/rotary (large displacement) motion (Reproduced with permission from Ref. [20], Copyright 2009 IEEE)

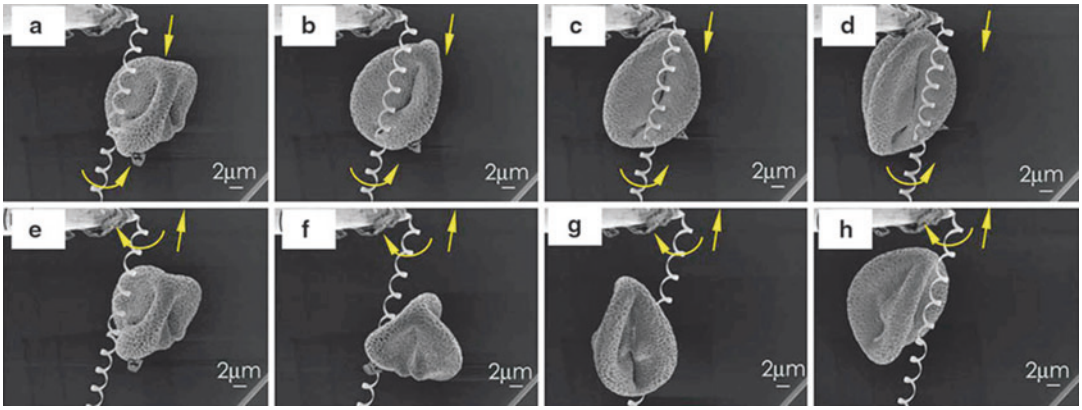


Nanorobotics for NEMS Using Helical Nanostructures, Fig. 7 In situ characterization of motion conversion using nanorobotic manipulation. (**a**) One end of an as-fabricated DCHNB is cut and attached to a "sticky" probe. (**b**, **e**) Rotation of the central part while moving the probe downward (Reproduced with permission from Ref. [20], Copyright 2009 IEEE)

Examples of Application

A fundamental application of the rolled up nanohelix is the 3-D microscopy based on the linear-to-rotary motion conversion [20], in which samples are rotated so that different views are exposed to light, electron beams, or focused-ion beams (FIBs). Due to their extremely compact size, the dual-chirality helical nanobelts are well suited to serve as rotary scanners for creating 3-D scanning probe microscope (SPMs), rotary stages for a microgoniometer for observing an object from different crystal surfaces, and components of nanomachines.

Figure 8 shows 3-D imaging of a pollen grain. The pollen grain is picked up by a Picoprobe and attached to the DCHNB. The motion converter has been characterized before loading the sample. Figure 8a–d shows that the sample rotates counterclockwise (top view) for approximately 180° when releasing the extended DCHNB. Figure 8e–h shows further stretching of the DCHNB, and the sample rotates clockwise (top view) for another 180°. It can be seen that the



Nanorobotics for NEMS Using Helical Nanostructures, Fig. 8 Three-dimensional microscopy of a pollen grain. (**a**–**d**) When releasing the extended DCHNB, the sample rotates counterclockwise (*top view*). (**e**–**h**) When stretching further the DCHNB, the sample rotates clockwise (*top view*) (Reproduced with permission from Ref. [20], Copyright 2009 IEEE)

different aspects of the sample can be exposed to the electron beam simply by extending or releasing the DCHNB with a small displacement. Further possibilities for this converter include goniometry for a transmission electron microscope (TEM), tomography for a SEM or a TEM, and 3-D SPM. By integrating in the fabrication processes of DCHNBs, rolled-up spirals can serve as claws for holding relatively large samples [6].

References

1. Amelinckx, S., Zhang, X.B., Bernaerts, D., Zhang, X. F., Ivanov, V., Nagy, J.B.: A formation mechanism for catalytically grown helix-shaped graphite nanotubes. Science **265**(5172), 635–639 (1994)
2. Kong, X.Y., Wang, Z.L.: Spontaneous polarization-induced nanohelixes, nanosprings, and nanorings of piezoelectric nanobelts. Nano Lett. **3**(12), 1625–1631 (2003)
3. Schmidt, O.G., Eberl, K.: Nanotechnology – thin solid films roll up into nanotubes. Nature **410**(6825), 168 (2001)
4. Prinz, V.Y., Seleznev, V.A., Gutakovsky, A.K., Chehovskiy, A.V., Preobrazhenskii, V.V., Putyato, M.A., Gavrilova, T.A.: Free-standing and overgrown InGaAs/GaAs nanotubes, nanohelices and their arrays. Physica E **6**(1–4), 828–831 (2000)
5. Zhang, L., Deckhardt, E., Weber, A., Schonenberger, C., Grutzmacher, D.: Controllable fabrication of SiGe/Si and SiGe/Si/Cr helical nanobelts. Nanotechnology **16**(6), 655–663 (2005)
6. Zhang, L., Dong, L.X., Bell, D.J., Nelson, B.J., Schönenberger, C., Grützmacher, D.: Fabrication and characterization of freestanding Si/Cr micro- and nanospirals. Microelectron. Eng. **83**(4–9), 1237–1240 (2006)
7. Schmidt, O.G., Schmarje, N., Deneke, C., Muller, C., Jin-Phillipp, N.Y.: Three-dimensional nano-objects evolving from a two-dimensional layer technology. Adv. Mater. **13**(10), 756–759 (2001)
8. Kubota, K., Fleischmann, T., Saravanan, S., Vaccaro, P.O., Aida, T.: Self-assembly of microstage using micro-origami technique on GaAs. Jpn. J. Appl. Phys **42**(6B), 4079–4083 (2003). Part 1 – Regular Papers Short Notes & Review Papers
9. Zhang, L., Ruh, E., Grutzmacher, D., Dong, L.X., Bell, D.J., Nelson, B.J., Schonenberger, C.: Anomalous coiling of SiGe/Si and SiGe/Si/Cr helical nanobelts. Nano Lett. **6**(7), 1311–1317 (2006)
10. Cho, A.: News focus: nanotechnology: pretty as you please, curling films turn themselves into nanodevices. Science **313**, 164–165 (2006)
11. Chen, X.Q., Zhang, S.L., Dikin, D.A., Ding, W.Q., Ruoff, R.S., Pan, L.J., Nakayama, Y.: Mechanics of a carbon nanocoil. Nano Lett. **3**(9), 1299–1304 (2003)
12. Gao, P.X., Mai, W.J., Wang, Z.L.: Superelasticity and nanofracture mechanics of ZnO nanohelices. Nano Lett. **6**(11), 2536–2543 (2006)
13. Bell, D.J., Sun, Y., Zhang, L., Dong, L.X., Nelson, B. J., Grützmacher, D.: Three-dimensional nanosprings for electromechanical sensors. Sens. Actuator A-Phys. **130–131**, 54–61 (2006)
14. Bell, D.J., Dong, L.X., Nelson, B.J., Golling, M., Zhang, L., Grutzmacher, D.: Fabrication and characterization of three-dimensional InGaAs/GaAs nanosprings. Nano Lett. **6**(4), 725–729 (2006)
15. Zhang, H.F., Wang, C.M., Buck, E.C., Wang, L.S.: Synthesis, characterization, and manipulation of helical SiO_2 nanosprings. Nano Lett. **3**(5), 577–580 (2003)

16. Zhang, D.Q., Alkhateeb, A., Han, H.M., Mahmood, H., McIlroy, D.N., Norton, M.G.: Silicon carbide nanosprings. Nano Lett. **3**(7), 983–987 (2003)
17. Gao, P.X., Ding, Y., Mai, W.J., Hughes, W.L., Lao, C. S., Wang, Z.L.: Conversion of zinc oxide nanobelts into superlattice-structured nanohelices. Science **309**(5741), 1700–1704 (2005)
18. Dong, L.X., Nelson, B.J.: Robotics in the small, part II: nanorobotics. IEEE Robot. Autom. Mag. **14**(3), 111–121 (2007)
19. Dai, L., Zhang, L., Dong, L.X., Shen, W.Z., Zhang, X. B., Ye, Z.Z., Nelson, B.J.: Long-range linear elasticity and mechanical instability of self-scrolling binormal nanohelices under a uniaxial load. Nanoscale **3**, 4301–4306 (2011)
20. Dong, L.X., Zhang, L., Kratochvil, B.E., Shou, K.Y., Nelson, B.J.: Dual-chirality helical nanobelts: linear-to-rotary motion converters for three-dimensional microscopy. J. Microelectromech. Syst. **18**(5), 1047–1053 (2009)

Nanorobotics for Synchrotron Radiation Applications

Nabil Amari, David Folio and Antoine Ferreira
INSA Centre Val de Loire, Université d'Orléans, Bourges, France

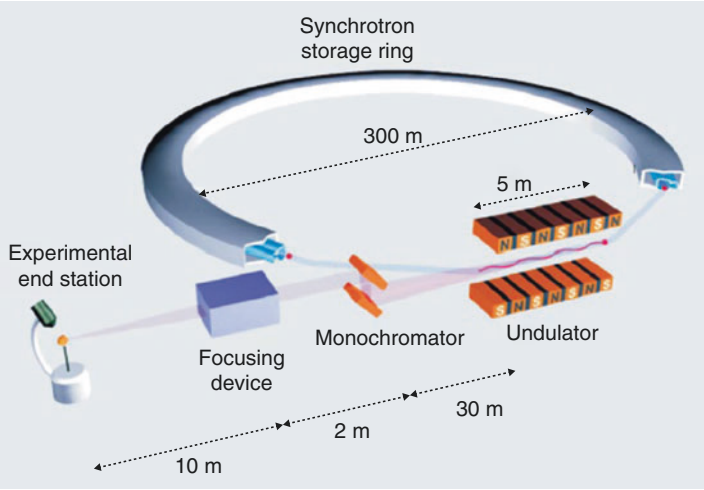
Definition

By nanorobotic manipulation, it is meant that nano-objects are localized, positioned, and placed by controlling external forces under the X-ray beam of a synchrotron.

Introduction

Nanorobotics for Synchrotron Radiation Facilities

In nanosciences and nanotechnology, the frontiers between chemistry, physics, biology, and engineering become blurred, and more and more an integrated, multimodal approach to the characterization of the intrinsic properties of individual nano-objects becomes essential. The goal is to distinguish the properties of the object itself free from object–surface or object–environment interactions. Indeed in the continuous reduction of the device size where smaller and smaller material quantity is involved, the individual properties of each nano-object will tend to modify the collective properties. In this frame, the development of reliable applications, often based on auto-organization process, required the knowledge of the effects that are coming from the intrinsic properties of the nano-object compared to the ones that are coming from its interaction with the environment as well as what is due to a collective behavior of several nano-objects in interaction. The key point is to study these differences on objects ranging from few nanometers to 1 μm, where specific and size-dependent properties find their expression. Since long, several spectroscopic techniques as X-ray (for inorganic sample) or laser radiation (for inorganic and biological sample) have been used to characterize structural and electronic properties of materials and objects. Laser spectroscopies as fluorescence or Raman have improved their efficiency and can also be used to analyze very small quantity of materials down to a single nano-object [1]. In parallel, the development of synchrotron radiation has increased the energy span of radiation available for characterizing materials. As the focalization of X-ray beam improves regularly (presently less than 100×100 nm^2), it becomes possible to analyze smaller and smaller sample [2] (10^8 atoms which corresponds to a 100 nm radius object). Based on this X-ray optical progress and to address the needs of the nano-community, specific beamlines dedicated to nanosciences are currently developed or will be developed in the frame of the ESRF (European Synchrotron Radiation Facility located in Grenoble, France) Upgrade Programme, for example, UPBL1, UPBL4, and UPBL5 [3]. This work is carried out within the PIANHO project n°ANR-NANO-042 funded by the French National Agency (ANR) in the frame of its 2009 Program in Nanosciences, Nanotechnologies, and Nanosystems (P3N2009). As illustration, Fig. 1 shows the general scheme of the robotic platform handling the nano-sample directly installed on synchrotron radiation end station for nano-object characterization experiments.



Nanorobotics for Synchrotron Radiation Applications, Fig. 1 Scheme of the synchrotron radiation facility with AFM-based robotic system (NanoWorker) directly installed on synchrotron radiation end station for nano-object characterization experiments

Usually, the robotic platform is composed of a multidegree-of-freedom robotic manipulator (Fig. 2a) handling a vacuum chamber with inert gas (Fig. 2b) that holds the nanomanipulation system inside. Different optical access are possible for X-ray radiation, fluorescence detector, alignment on the beam, and alignment microscope.

Application Fields of the X-Ray NanoWorker

Nanorobotics systems have the possibility to measure and characterize intrinsic properties of an individual nano-object by picking up and holding a free nano-object in the probe beam (synchrotron or laser beam) for several minutes and releasing it without damage. To reach these goals, we propose in this entry to develop an advanced nanomanipulation platform composed of a dual scanning probe system called X-ray NanoWorker (see Fig. 2). It is composed of two atomic force microscope probes and associated electronics for morphological characterization of objects in the nanoregime under synchrotron radiation X-ray beams (XAFM or X-Ray AFM). The potential impacts of this X-ray NanoWorker will touch all the scientific areas that already use the synchrotron radiation. Indeed, it will provide a unique way to access the intrinsic properties of inorganic, organic, and biological nano-objects as modified carbon nanotube, proteins, or macromolecules [4]. In the middle term, with few modifications, the nanorobotic platform could be used in different environments that could be liquid or in controlled atmosphere and with all sources of excitation (lasers, protons, electrons, etc.). It appears highly promising the possibility of mechanically modifying in situ both inorganic and biological materials. In solid-state physics, it would offer the unique opportunity of submitting to stress single nanocrystals and observe the formation in situ of dislocations [5]. In biology it would offer literally a hand in modifying the structure of molecules the same way scanning probe microscopes (SPMs) do with atoms in 2D systems.

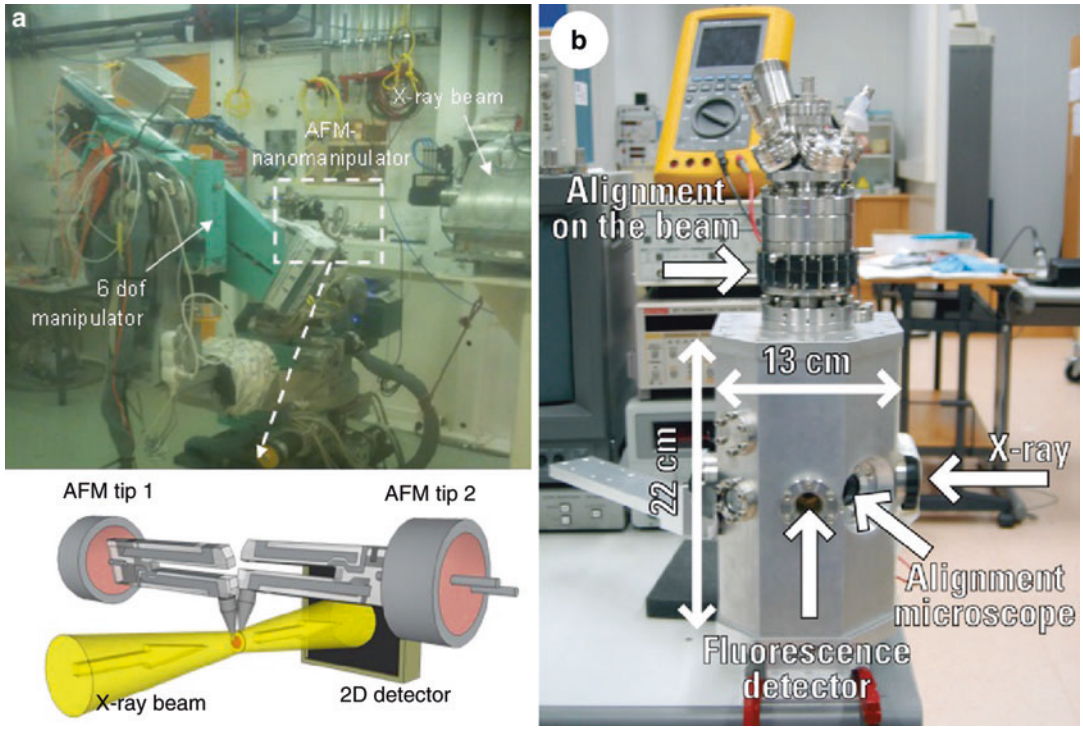
Actual Limitations in X-Ray Nanomanipulation

In this section, we analyze the different scientific and technological limitations associated with the nanomanipulation under an X-ray or laser beamline.

Physical Aspects of the Existing Techniques of Nanomanipulation

At the nanometer scale the physical effects that dominate the interactions between the objects are completely different from the micrometric or larger scale.

Indeed, friction as well as adhesive, magnetic, or electric forces largely overcomes gravity. This



Nanorobotics for Synchrotron Radiation Applications, Fig. 2 The experimental setup for the combined planning and tracking experiments at the European Synchrotron Radiation Facility located in Grenoble, France. (**a**) Experimental end-station setup composed of a 6-DoF serial PUMA manipulator as a beamline sample holder (NanoWorker). (**b**) General view of the vacuum chamber that holds the nanomanipulation system X-ray NanoWorker (courtesy of Prof. Joël Chevrier, ESRF Grenoble, France).

is due to the very low mass of the nano-object on one side as well as to the strongly increased surface–volume ratio. In addition, as the distances are small, quantum phenomena as tunneling effects or Casimir forces become not negligible. These quantum effects increase the complexity of the interrelation between the objects and their environment. For all of these reasons, object manipulation at nanometer scale is very different from micromanipulation, and new strategies far from our classical human perception and habits have to be elaborated. Nevertheless, the need of nanomanipulation is not new, and specific tools have been developed to interact and manipulate at nanoscale. Three main techniques are currently employed: optical tweezers, scanning probes, and mechanical nano-grippers:

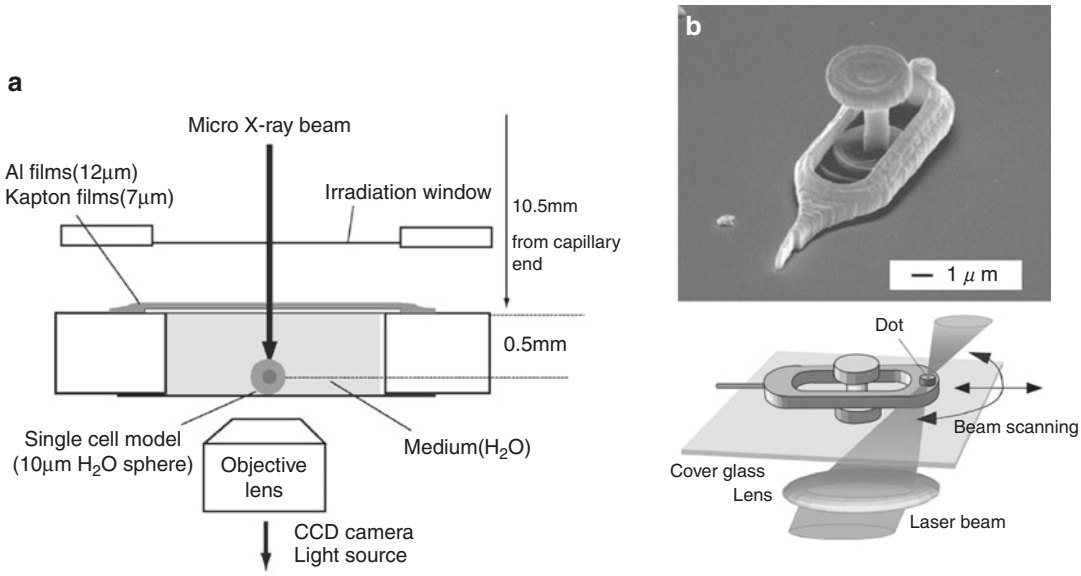
– Optical tweezers: Optical tweezers have been developed 20 years ago [6] and are currently used to manipulate microspheres or biological entities in liquid environment. Optical tweezers trap objects near the waist of strongly focused laser beams and position them in 3D with nanometer resolution. Nevertheless, optical trapping in air or in vacuum remains difficult, because van der Waals or capillary forces are several orders of magnitude larger than the optical ones, even if lately they succeeded in trapping droplets in air [7].
– Scanning probe technique: Scanning probe microscopy (SPM) techniques are well known in surface science to study surface morphology and properties from micron down to atomic scale. In addition to scanning probe, local probes have demonstrated their capability to manipulate single nano-objects, molecules, and atoms on surfaces in different environments (from ultrahigh vacuum to liquids, see, e.g., http://www.almaden.ibm.com/vis/stm/

corral.html). However, this manipulation stays mainly in 2D; it confines this technique to support object on a surface: until now, this technique has not been successful to pick up and place a nano-object from a given support to another one. More recently, a SPM developed for operating on an X-ray beamline has demonstrated the compatibility of the local probe technique with the noisy X-ray working environment of the experimental end stations [8]. This beamline AFM can image with a nanometer resolution, and the tip has been used as a nano-indenter on a germanium pyramidal dot or metal surfaces in order to characterize in situ with the X-ray the induced crystal.

– Mechanical nano-gripper: In the past year, mechanical nano-grippers have been fabricated through MEMS/NEMS (micro-/nano-electromechanical system) technology process or by the use of carbon nanotubes [9]. These devices, when coupled to a SEM [10] (scanning electron microscope), a fluorescent microscope [11], or an AFM [12], are capable of manipulating individual inorganic or biological sub-micro-/nano-objects in different environments. However, all these prototypes present several common weak points. Firstly, an additional online technique is required to control in real time the object manipulation (SEM, optical fluorescent microscope, AFM), which limits the type of nanomanipulation and work environments. Secondly, the determination of the more reliable and effective manipulation strategy is based on trial–error method that is time consuming and not fully efficient because the gripping and holding methods are based on short-range interactions (strong contact). These methods can induce modifications of geometrical, electrical, or mechanical object properties; thus, inherent information about individual and intrinsic properties of nano-object are hardly obtained.

Limitations on Nanomanipulation Under a Beamline

Actually, manipulation at nanoscale still presents some significant weakness, such as low accuracy or low repeatability. Our goal is to combine the high-lateral resolution imaging capability of atomic force microscopy (AFM) with the high penetration power and high element sensitivity of X-rays. We will show the ability of our home-built microscope to perform AFM imaging on a beamline sample holder, align the AFM tip with the X-ray beam, and perform total electron yield (TEY) detection with the AFM tip. The present entry addresses the possibility to measure and characterize intrinsic properties of an individual biological cell by picking up and holding a "free" micro-/nano-object in the probe beam (synchrotron or laser beam) for several minutes and releasing it without damage. To reach these goals, we propose to develop an advanced nanomanipulation platform: a dual scanning probe's coupling with a haptic system through specific and novel control feedback loops. The whole system is called an X-ray NanoWorker. High-precision position measurement systems based on laser beam reflection and/or transmission are commonly used in nanorobotics applications. In particular, laser beam nanomanipulation is a useful tool to either characterize material properties of organic cells [13], to study biological radiation effects on single cellular [13], or to interact with cells even in liquid environment [2, 27]. Classically, a laser beam nanomanipulation platform is composed of the optical device, including the laser beam and sensor (e.g., a camera or position-sensitive detector (PSD)), and alignment mechanisms. The frame structure for maintaining the optical configuration allows precise control positioning of the cells or microrobots. For instance, Fig. 3 illustrates the general microbeam system for radiation and transport tasks. The main problem is then to focus the beam in a few micrometer-size spots and to control actively the beam direction to stabilize the beam at a desired location with a nanometer resolution. Such issues must be addressed when focusing a near-infrared laser beam at a nerve cell's leading edge [14, 15], when the laser beam perfectly tracks the moving AFM probes [16] during manipulation tasks, or when the laser beam illuminates a micro-object handled by a nano-gripper for material characterization [17]. In the abovementioned cases, precise laser



Nanorobotics for Synchrotron Radiation Applications, Fig. 3 The microbeam system for radiation and transport: (**a**) model of single cell irradiation with micro X-ray beam [13] and (**b**) optically driven micro tools in liquid for cellular manipulation [2] (Courtesy from the author of [2])

beam tracking of dynamic position with high-bandwidth rejection of disturbances produced by nanomanipulator platform vibration, piezoelectric actuator thermal drifts, photodetector noises, Brownian motion of laser beam, and atmospheric turbulence is critical for the success of micro- and nanomanipulation radiation tasks.

The single photodiode sensor is currently being introduced in the nanomanipulation tasks and used only for position measurement. Especially, the PSD is advantageous in terms of high-speed response compared to the camera processing. In this entry, we propose to use the PSD at the same time as indicator and feedback control in order to increase the overall performance and reliability of microbeam irradiation systems. The literature provides mainly two laser beam tracking configurations, that is, steering (i) the laser beam or (ii) the photodetector. In the first case, some works propose to use fast tilt two-axis steering mirrors based on electrostatic MEMS actuators [18] or piezoelectric actuators with a fixed four-quadrant PSD. In the second case, the PSD is driven by a dual actuation system with robot micro-/nanomanipulators [19] or x-y linear positioning stages [20]. Whatever the technology involved, robust control of the laser beam tracking system is needed.

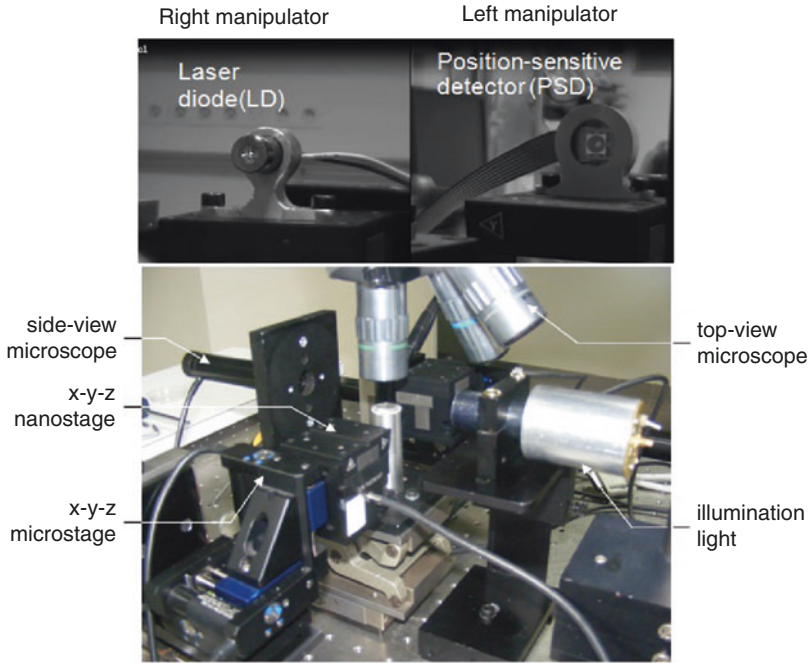
Control Systems Dedicated to Nanoscale Instrument

Classically, most mechanical micro- or nano-grippers used piezoelectric actuators to move in 2D or 3D in the nano-space. Indeed, piezoelectric components allow continuous motion with high resolution (about 1 nm). Moreover, they are easy to implement against other positioning systems as capacitive or inductive one. Nevertheless, piezoelectric devices present several drawbacks: (i) hysteresis behavior, (ii) creep, and (iii) thermal drift. These weaknesses lead to nonlinear effects of piezo-actuation which are difficult to model and to correct. As nanopositioning is a strong prerequisite for nanomanipulation, the design of reliable and precise nanopositioning control systems is a serious bottleneck in terms of speed and accuracy for nano-investigation. In addition, many applications impose some important

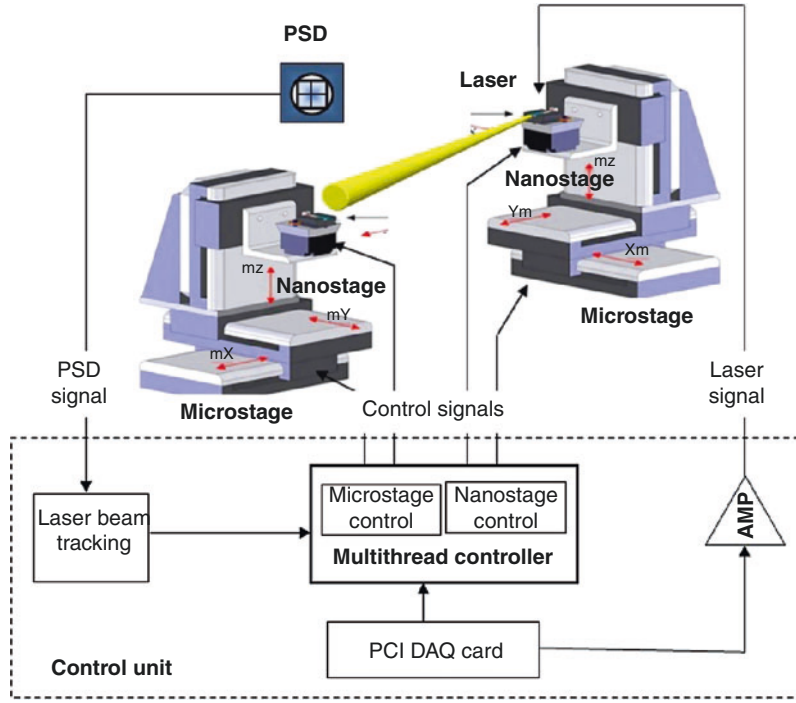
constraint on the tracking bandwidth and reliability in term of repeatability of experiments. Since a few decades, there are further and further scientific studies on the design of new control scheme dedicated to nanopositioning. Hence, the main difficulty to overcome is to synthesize control strategies robust to piezoelectric nonlinearities. The different control approaches are broadly classified into feedback, feedforward, iterative, and sensorless control [21]. One most considerate approach relies on feedforward control scheme. Indeed, the use of feedforward where energy-based model of the piezoelectric behavior (for instance, by integrating the dependence of hysteresis loops) is applied can lead to improved output tracking performance. More recently, some control law which combines feedforward and feedback strategies has been reported [22]. Another solution to design fine control law with high-repeatability movement is to use, for instance, a Kalman estimator to compensate the nonlinear effect and the piezoelectric hysteresis behavior [23]. Nevertheless, all of these new nanopositioning approaches are mainly dedicated to scanning application and not yet to manipulation aim. The following sections are based on previous works published by the authors in [28].

Experimental Setup

The experimental setup of the beam pointing and tracking is shown in Fig. 4. Two controllable micro-/nanomanipulators face each other, composed of 3-DoF high-precision dual stages, i.e., magnetic *X-Y-Z* closed-loop microstage (MCL Nano-Bio2M, from Physics Instruments) and piezoelectric *x-y-z* closed-loop nanostage (P-611.3S NanoCube, from Physics Instruments), respectively. Each pair of micro-/nanomanipulators constitutes a compact computer-controllable system for *X-Y-Z* alignment (scanning) and positioning. In particular, the travel range of the microstage in the *X-Y-Z* directions is about few centimeters, while the fine motion of the nanostage is about 100 μm. Thus, this dual-stage manipulator combines the advantages of ultrahigh resolution of the piezo-stage together with the long travel range of the



Nanorobotics for Synchrotron Radiation Applications, Fig. 4 Experimental setup



Nanorobotics for Synchrotron Radiation Applications, Fig. 5 Schematic diagram of the architecture of the laser beam tracking control system

micropositioning stages. The laser source is mounted on top of the nanostage (right manipulator) producing the laser beam. The main components of the beam steering experiment are a 635 nm laser. A four-quadrant position-sensitive detector (PSD) is mounted on top of the nanostage (left manipulator). The PSD measures the position of the image that the laser beam forms on a fixed plane. On the side view, a white light illuminates the workspace for top-view (optical microscope: Mitutoyo × 50) imaging for observing the laser beam when subjected to the micrometer variation and side-view (digital microscope: TIMM × 150) imaging used to observe the image of the laser beam in the operating region of the PSD. The sample platform is fixed on the system base.

Figure 5 shows the overall control scheme for power, laser beam tracking, and micro-/nanomanipulator control. The laser beam motion control (Brownian or stochastic trajectory) and measurement sequences are processed in real time using MATLAB™ xPC software with a stand-alone target machine operating at a sample and hold rate of 2 kHz. A data acquisition (DAQ) (NI 6289) card is used for high-speed capturing of photodiode voltage output from a lock-in to detect maximum laser beam intensity and beam tracking. A multi-thread planning and control system is developed to independently manage the coordination during parallel laser beam motion and tracking, respectively.

Dynamics Modeling

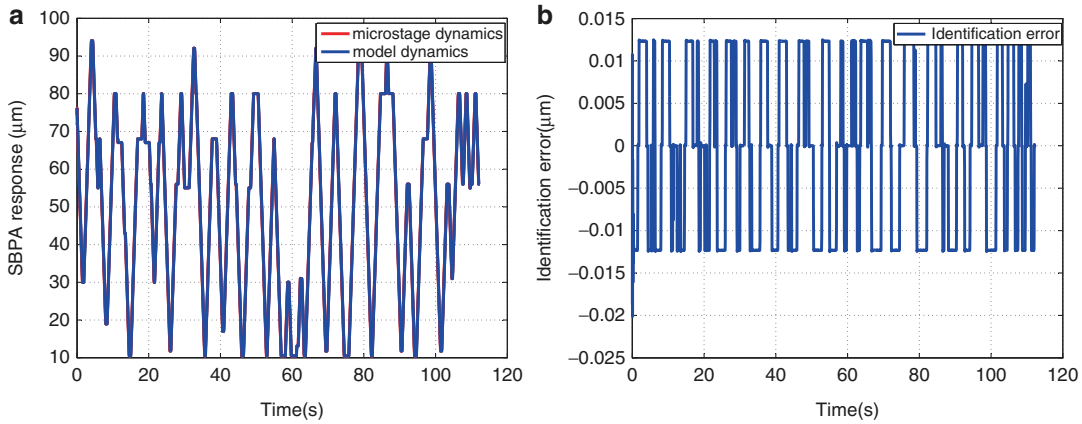
This section reviews the different model dynamics of the different system components.

Dynamics of Piezoelectric and Magnetic Actuators

The first step for controller synthesis is to identify the model dynamics of the actuation platform. As no parametric information on drivers are available, an identification phase is needed to set up dynamics modeling of dual micro-/nanostages. The piezoelectric 3-DoF nanostage and the magnetic 3-DoF microstage are deemed as three-input and two-output (MIMO) system.

Nanorobotics for Synchrotron Radiation Applications, Table 1 Identified parameters of the dynamic models of the micro (G_m)- and nano (G_n)-stages for each x-, y-, and z-axis

Parameters	G_{mx}	G_{my}	G_{mz}	G_{nx}	G_{ny}	G_{nz}
a_1	0.4336	0.4363	0.4335	−0.00557	−0.00579	−0.000556
a_2	−0.36	−0.3599	−0.36	0.0196	0.025	0.0222
a_3	−0.162	−0.158	−0.171	−0.93	−0.997	−0.99999
b_0	0.9755	0.9755	0.9755	0	0	0
b_1	0.4769	0.48	0.46	0.00535	0.003	0.000556
b_2	−0.3779	−0.3559	−0.389	0.0271	0.00741	0.00535
b_3	−0.168	−0.201	−0.156	0.0507	0.0117	0.00761



Nanorobotics for Synchrotron Radiation Applications, Fig. 6 Identification results along the x-axis (**a**) microstage dynamics when excited by a PRBS signal and (**b**) the identification error

Thus, the modeling approach is based on system identification by experimental responses of micro-/nanostages using a pseudo-random binary sequence (PRBS) input signal. This input signal is implemented as a set of digital step signals with several operating points ranging from 0 to 100 µm with an elementary step of about 10 µm. The driving frequencies are lower than the first resonant mode of the nanostage system. The identification procedure is then realized using a closed-loop response in order to protect the system submitted to large domain frequency and subtract the model dynamic in open loop. To reduce the system order, the dynamic models of the micro (G_m)- and nano (G_n)-stages are chosen as a third-order approximation for each x-, y-, and z-axis, respectively:

$$G_{m(x,y,z)}(\mathsf{z}) = \frac{b_0 + b_1\mathsf{z}^{-1} + b_2\mathsf{z}^{-2} + b_3\mathsf{z}^{-3}}{1 + a_1\mathsf{z}^{-1} + a_2\mathsf{z}^{-2} + a_3\mathsf{z}^{-3}} \quad (1)$$

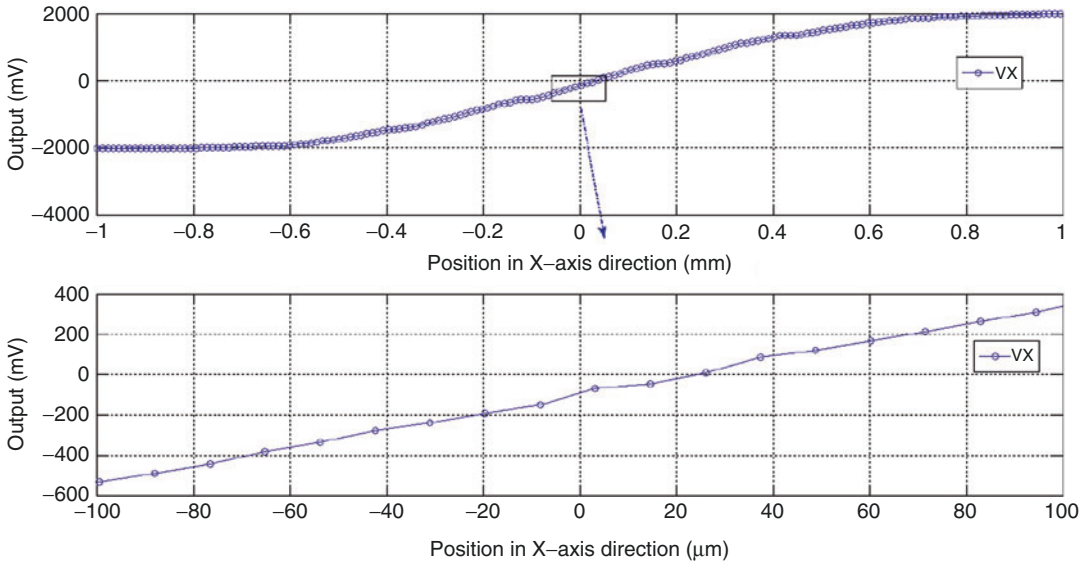
$$G_{n(x,y,z)}(\mathsf{z}) = \frac{b_1\mathsf{z}^{-1} + b_2\mathsf{z}^{-2} + b_3\mathsf{z}^{-3}}{1 + a_1\mathsf{z}^{-1} + a_2\mathsf{z}^{-2} + a_3\mathsf{z}^{-3}} \quad (2)$$

This choice was based on the correlation factor rate between the response of the model dynamics and the manipulators. The proposed optimized structure gives very high correlation factor values close to 99.5 %. The obtained parameters values are listed in the Table 1.

Figure 6 depicts the identification results of the dynamic responses and the identified model to the PRBS excitation signal along the x-axis of the microstage. Let us notice that a same behavior could be reported for the other axes. As one can see the identified model approximates suitably the system dynamics with an error close to 0.01 µm.

Dynamics of Four-Quadrant Detector

A four-quadrant position-sensitive detector (PSD) has four photosensing parts arranged in four



Nanorobotics for Synchrotron Radiation Applications, Fig. 7 Output voltage curve V_x with a zoom in the block area near zero on a four-quadrant PSD

quadrants, respectively. When the elements are lighted by a beam of laser, they will generate currents according to the light intensity and then amplified into voltage signals. The combinations of voltages V_1 to V_4 can be used to indicate the offsets of the spot in relation to the center of the PSD as follows:

$$\begin{aligned} V_x &= (V_1 + V_4) - (V_2 + V_3) \\ V_y &= (V_1 + V_2) - (V_3 + V_4) \\ V_s &= V_1 + V_2 + V_3 + V_4. \end{aligned} \tag{3}$$

The V_x and V_y channel outputs are directly related to the energy of the laser beam that falls in each quadrant, while V_s is the sum voltage. It is assumed that the light intensity on the laser beam's cross section obeys Gaussian distribution. The current generated by each sensing element can be described as given in:

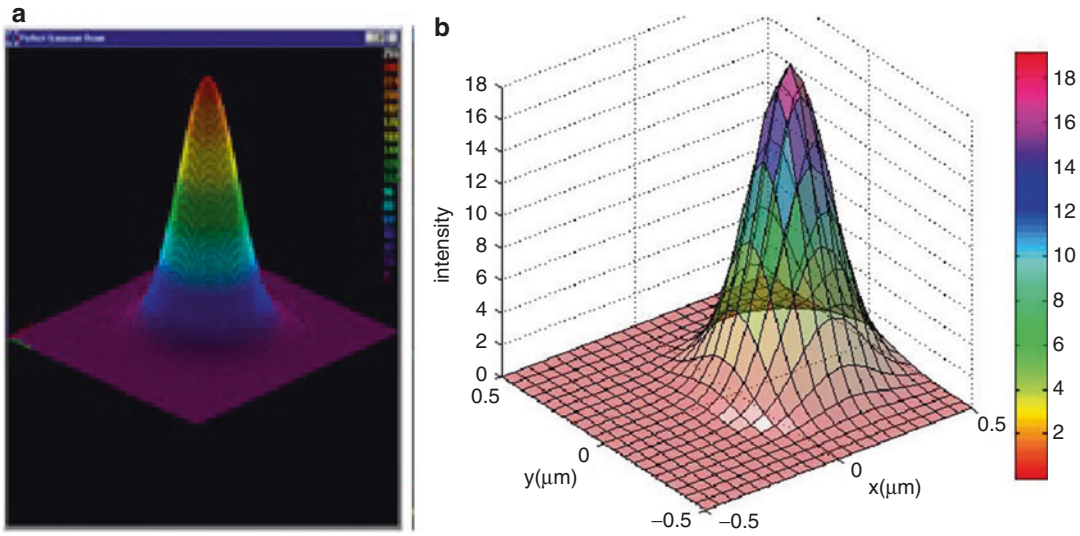
$$I = k_1 \iint \frac{2E_l}{\pi^2 r} \cdot \mathrm{e}^{\frac{2\left(x_1^2 + y_1^2\right)}{r^2}} \cdot dx_1 dy_1 \tag{4}$$

where I is the current, r is the radius of the laser light spot, E_t is the energy of the laser beam, (x_1, y_1) is the coordinate of a point on the light spot in a coordinate system located at the center of the light spot, and k_1 is a coefficient. The calibration of the PSD is performed by moving the laser beam in the operating region of the fixed PSD. Figure 7 presents an output voltage V_x measured by the PSD when the laser beam moves in the x-axis direction, with respect to the manipulator measured position. As one can see, in the operation region (small neighborhood of the aligned location), the photodiode voltage output V_x is approximately linearly related to light intensity units, with a negative slope. As the curve V_y is similar to that, it is omitted here. Obviously, when the spot is located in the sensing surface $V_x \neq 0$, $V_y \neq 0$ while if the spot is located in the center $V_x = 0$, $V_y = 0$. Moreover, we notice that there is a nonzero output at $x = 0$ due to the amplification of the noisy signals provided by the PSD sensor, which will be processed by the Kalman filter (KF).

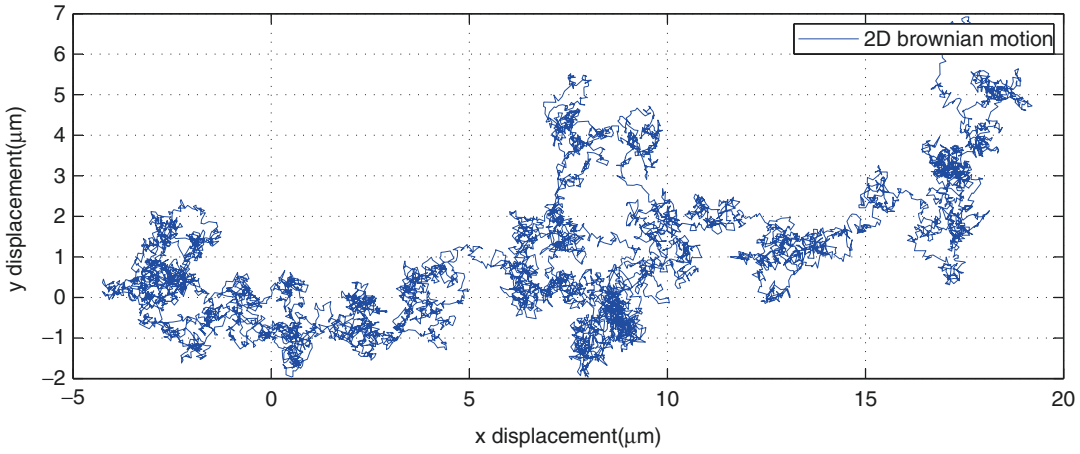
Furthermore, as one can see in Fig. 8, the experimental intensity sensed by the PSD can be fitted with a Gaussian distribution as calculated by the theoretical Eq. 4.

Dynamics of Laser Beam Position

In this study, a dynamic disturbance is considered where the laser beam motion variation is assumed similar to the Brownian motion (represented in



Nanorobotics for Synchrotron Radiation Applications, Fig. 8 Light intensity on the laser beam's cross section: (**a**) theoretical and (**b**) experimental intensity obeying to Gaussian distribution



Nanorobotics for Synchrotron Radiation Applications, Fig. 9 Particle Brownian motion

Fig. 9) of a particle subjected to excitation and frictional forces. The Brownian motion is applied as an input disturbance to the position of the laser beam carrier and to the expected displacement measured by the PSD sensor. The Brownian motion is given by the following generalized differential equation:

$$\frac{d^2x(t)}{dt^2} + \beta_x \frac{dx(t)}{dt} = W_x \qquad (5)$$

where β_x is the coefficient of friction and $W_x \sim \dot{N}\left(0, \delta_x^2\right)$. To estimate with a discrete filter the laser beam positions at each sampling time t_k, a discrete model of the continuous dynamics (Eq. 5) is necessary. In the x-axis the discretized equations of motion using a zero-order hold (ZOH) are given by:

$$\dot{x}_k = \frac{x_k - x_{k-1}}{\Delta T}; \quad \ddot{x}_k = \frac{\dot{x}_k - \dot{x}_{k-1}}{\Delta T} \qquad (6)$$

From Eqs. 5 and 6, we obtain:

$$\dot{x}_k = \frac{\dot{x}_{k-1} + \Delta T W_{x_k}}{1 + \beta_x \Delta T} = a_x \dot{x}_{k-1} + b_x W_{x_k} \quad (7)$$

where ΔT is the discretization time step and the statistic properties of the excitation W_{x_k} force are assumed to be a zero-mean Gaussian random variable with variance δ_x^2. The y-axis can be modeled in the same manner as the x-axis, though with different dynamics. For 2D representation, the source state at discrete time k is defined as:

$$X_k = [\, x_k \quad y_k \quad \dot{x}_k \quad \dot{y}_k \,]^T \quad (8)$$

where (x_k, y_k) and $(\dot{x}_k, \dot{y}_k)$ are the source potion in the plane x-y and velocity, respectively. The discrete state space of the Brownian laser beam is then represented by:

$$X_k = AX_{k-1} + BW_k \quad (9)$$

$$Y_k = CX_{k-1} \quad (10)$$

The state representation matrices (A, B) are derived from the particle dynamics defined in Eqs. 6 and 7, and $W_k \sim N(0, Q)$ is a zero-mean Gaussian random variable with matrix variance Q. It comes from Eq. 9 that:

$$X_k = \sum_{i=1}^{k} A^{k-i} B W_i + A^k X_0 \quad (11)$$

Because successive random variables W_i form a priori discrete zero-mean white Gaussian process, X_k from Eq. 11 is Gaussian if the knowledge on X_0 is assumed Gaussian or equal to some fixed value. Its a priori variance at each step k can be calculated, thanks to:

$$\sigma^2(X_k) = \sum_{i=1}^{k} A^{k-i} B \sigma^2(W_i) + A^k \sigma^2(X_0) \quad (12)$$

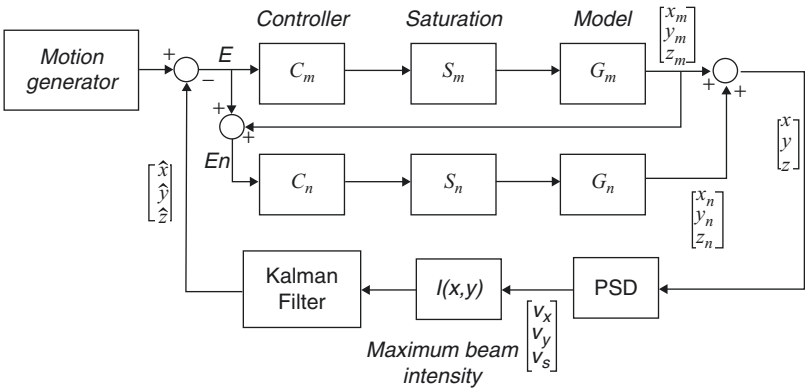
Eq. 12 shows that the bigger is the variance of W_k to set, the bigger is the a priori uncertainty variance on the possible values of the modeled unknown position of the beam laser at t_k. Finally, the measurement Y_k of position takes into account the discrete-time zero-mean white Gaussian noise V_k and variance R added by the four-quadrant position-sensitive detector, that is:

$$Y_k = CX_{k-1} + V_k \quad (13)$$

Control Scheme of Beam Pointing and Tracking

Dual Micro-/Nanomanipulator Controller

The problem considered here is to track the position variation of the laser beam into the x-y plane of the PSD by robust control commands sent to the dual micro-/nanomanipulator's motions. The localization of the current laser beam position depends on the maximum beam intensity detection. Hence, a monitoring of the maximum intensity of the laser beam is needed [24]. It implies to integrate a prediction model that anticipates the a priori laser beam motion, taking into account both dynamics, i.e., the laser beam and manipulator models. However, the different dynamics of the manipulators should be considered. The first actuator stage (microstage) has a large moving range but a low bandwidth while the second actuator stage (nanostage) has a high bandwidth but small moving range. The dual problem considered here is to know how to coordinate both micro- and nanostages in order to track the laser beam motions using only signal position feedback delivered by the PSD. We adopted the master–slave control design to transform the dual-stage control design problem into decoupled or sequential multiple independent controller designed separately. Figure 10 presents the master–slave control scheme to control two independent outputs of the micromanipulator G_m (1) and nanomanipulator G_n (2) systems by only one position feedback signal. This approach allows to combine the optical device with the micro-alignment mechanisms for an autonomous procedure especially when it comes to illuminating a micro-object under the focus of the laser beam for several minutes where the laser beam position is required or limiting the gripper deflection in interaction with a micro-object to not damage a flexible micro-object and the gripper tool.



Nanorobotics for Synchrotron Radiation Applications, Fig. 10 Master–slave controller with decoupling structure for maximum light tracking

Thanks to the master–slave controller based on the mixed sensitivity, it is possible to fix the adapted dynamic performances of microstage using a simple proportional–integral–derivative (PID) controller in the C_m controller. To ensure robustness of the piezoelectric nanopositioning stages, a H_∞ control scheme is implemented.

Kalman Filter Estimator

In robotics, the Kalman filter (KF) is most suited to problems in tracking, localization, and navigation and less so to problems in mapping [25, 26]. This is because the algorithm works best with well-defined state descriptions (e.g., positions, velocities) and for states where observation and time-propagation models are also well understood. The prediction–estimation stages of the Kalman filter are derived from Eqs. 9 and 13:

Prediction. A prediction $\hat{X}_{k|k-1}$ of the state at time k and its covariance $P_{k|k-1}$ is computed according to:

$$\hat{X}_{k|k-1} = A\hat{X}_{k-1|k-1} + BU_k \tag{14}$$

$$P_{k|k-1} = AP_{k-1|k-1}A^T + Q(k) \tag{15}$$

Update. At time k an observation $y(k)$ is made, and the updated estimate $\hat{X}_{k|k-1}$ of the state X_k, together with the updated estimate covariance $P_{k|k}$, is computed from the state prediction and observation according to:

$$\hat{X}_{k|k} = \hat{X}_{k|k-1} + K_k\left(Y_k - C_k\hat{X}_{k|k-1}\right) \tag{16}$$

$$P_{k|k} = P_{k|k-1} - K_kS_kK_k^T \tag{17}$$

where the gain matrix K_k is given by:

$$K_k = P_{k|k-1}C_kS_k^{-1} \tag{18}$$

where:

$$S_k = C_kP_{k|k-1}C_k + R_k \tag{19}$$

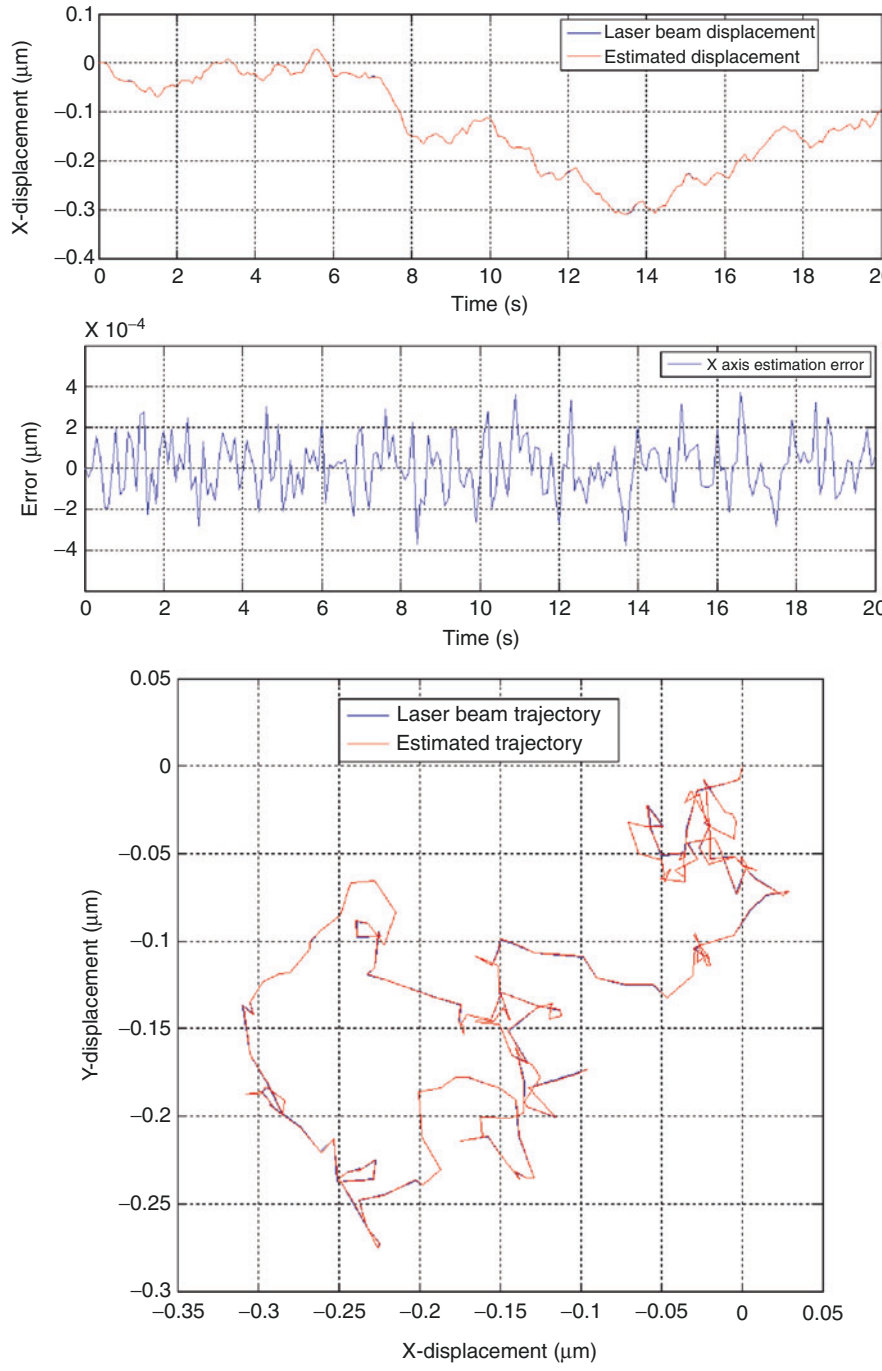
is the innovation covariance. The difference between the observation Y_k and the prediction observation $C_k\hat{X}_{k|k-1}$ is termed the innovation or residual $r(k)$. Thus, the input of the KF is the noisy measurement of the laser beam displacement in the x-y direction delivered by the photodiode detector and $\hat{X}_k$ is the output of the filter representing the estimation of the displacement at time t_k.

Figure 11 illustrates a simulation result of the laser beam tracking using a Kalman filter. Here, the laser beam is animated with a synthetic Brownian motion. As one can see, the KF allows to estimate efficiently the laser beam motion.

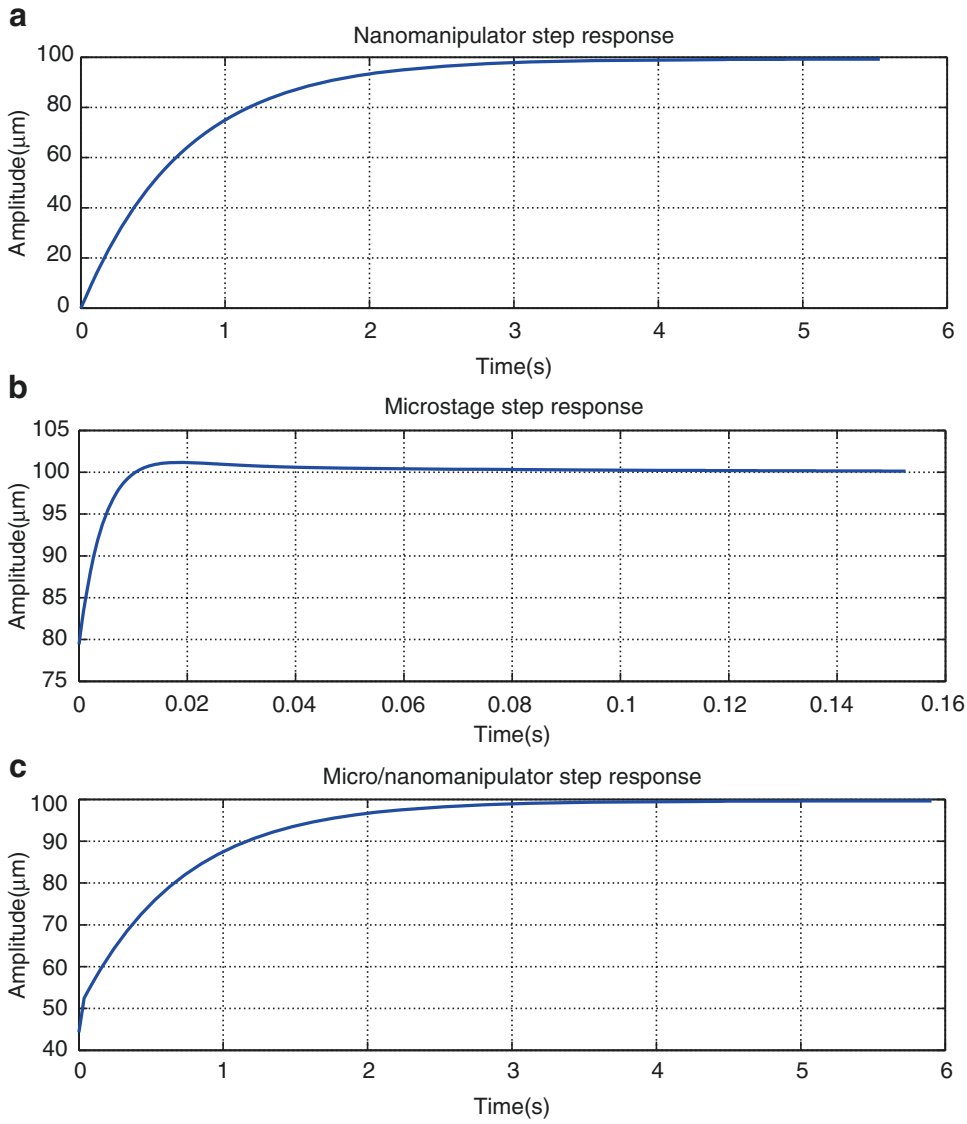
Experiments

Dual Micro-/Nanomanipulator Control Evaluation

First, the performance of the dual micro-/nanomanipulator is evaluated using step response experiments with a reference set to 100 µm. Figure 12 illustrates the obtained step response



Nanorobotics for Synchrotron Radiation Applications, Fig. 11 Simulation of the laser beam tracking with the Kalman filter: motion estimation along the *x*-axis with its corresponding errors and the 2D displacement



Nanorobotics for Synchrotron Radiation Applications, Fig. 12 Step responses of the (**a**) nanomanipulator stage, (**b**) the micromanipulator stage, and (**c**) the dual micro-/nanomanipulator system

of the x-axis of the (a) nanostage with the H_∞ controller, (b) microstage with the proportional–integral–derivative (PID) controller, and (c) master–slave controller where the decision of switching between the microstage and the nanostage depends on the error resolution. At first glance, the dual micro-/nanomanipulator controller converges with no error to the reference, with no overshoot, and with a settling time of about 4 s leading to a velocity of 25 μm/s.

The closed-loop sensitivity function S_T defines the performance of the dual micro-/nanomanipulator controller. Classically, in the frequency domain, sensitivity S_T is equal to the sum of the closed-loop sensitivity of the micromanipulator S_M and the nanomanipulator S_N (Fig. 10), that is:

$$S_T = \frac{1}{1 + G_T} = S_N S_M \tag{20}$$

with

$$S_N = \frac{1}{1 + C_N G_N}, \quad S_M = \frac{1}{1 + C_M G_M} \tag{21}$$

The sensitivity of the dual micro-/nanomanipulator demonstrates the possibility to design and combine the performances of each stage defined by their sensitivity function for control synthesis. Firstly, the microstage is governed by a serial proportional–integral–derivative (PID) controller formulated for each axis as:

$$C_M = K_p \frac{\tau_I s + 1}{\tau_I s} (\tau_D s + 1) \tag{22}$$

For the step response of the x-axis, the PID parameters are set to $K_{Px} = 140$, $\tau_{Ix} = 180$, and $\tau_{Dx} = 120$. Secondly, the nanostage uses a H_∞ robust controller. Classically, the H_∞ synthesis problem is expressed as a matrix optimization problem, that is:

$$\min_k \|N(k)\|, \quad N = [W_1 KS \;\; W_2 T \;\; W_3 S]^T \tag{23}$$

For a displacement in the x-direction of the nanomanipulator, the H_∞ synthesis allows to obtain a robust controller by using the following weighting functions:

$$\begin{aligned} \mathbf{W}_{1x} &= \frac{s+1}{s+0.9258}, \quad \mathbf{W}_{2x} = 0.7097, \\ \mathbf{W}_{3x} &= \frac{s+0.005935}{0.9613s+0.9032} \end{aligned} \tag{24}$$

The optimized discrete transfer function of the x-axis nanomanipulator controller is then given by:

$$\mathbf{C}_{Nx}(z) = \frac{1.186z^4 - 4.01z^3 + 5.623z^2 - 3.953z + 1.154}{z^4 - 3.25z^3 + 4.405z^2 - 3.06z + 0.9044} \tag{25}$$

Laser Beam Motion Prediction

Secondly, the laser beam motion is estimated using the presented Kalman filter (KF) based on the identified Brownian model. To evaluate the performance of the KF, the laser beam is mounted on the right manipulator, while the PSD sensor is mounted on the left manipulator, as depicted in Fig. 4. Then, a random trajectory is generated to the laser beam before to initiate the laser beam tracking by the PSD sensor. The data rate of the sensor is $\Delta T = 2$. The KF parameters, i.e., measurement noise matrix R, process noise matrix Q, and initial state error matrix P_0, are defined as:

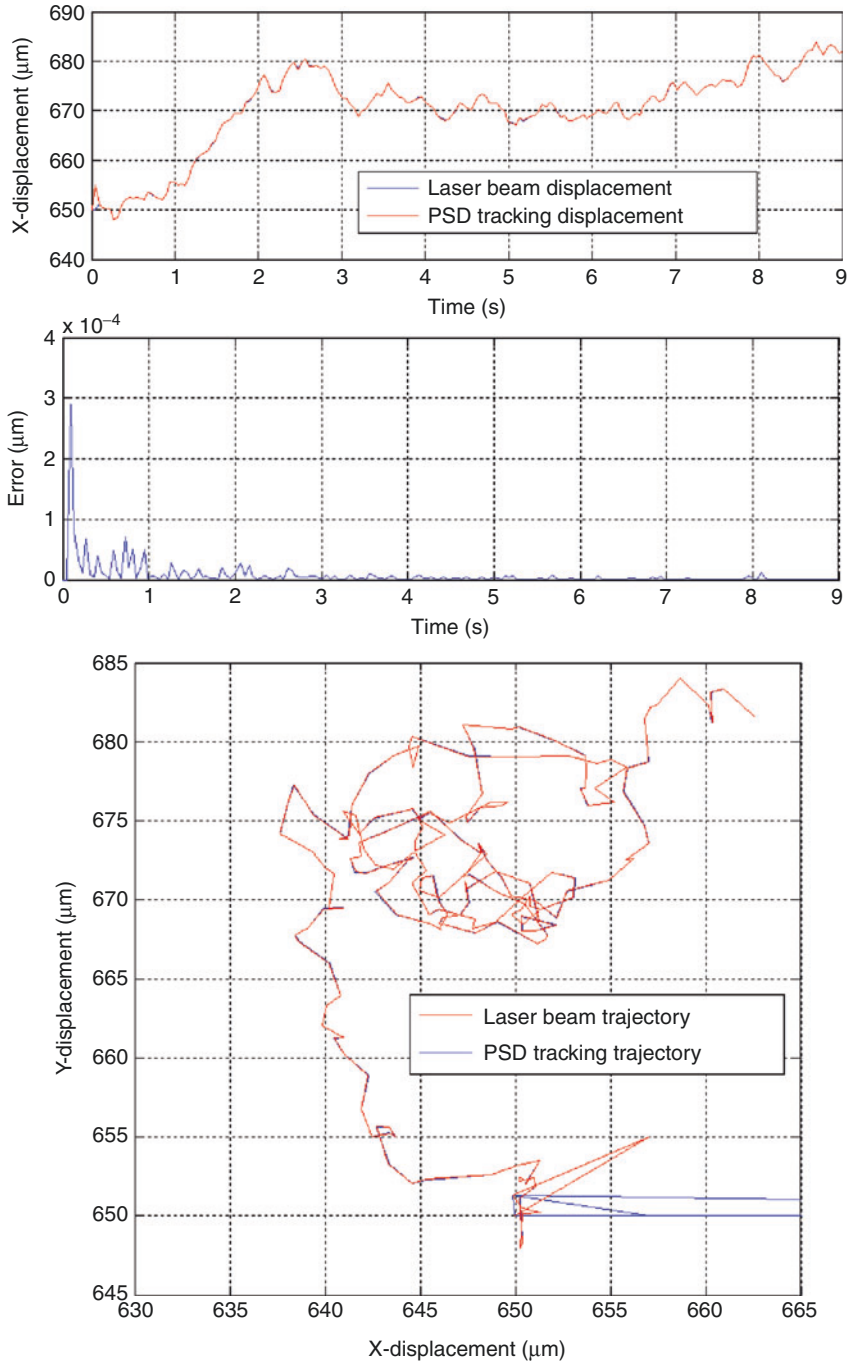
$$\begin{aligned} \mathrm{Q} &= 10^{-2}\mathbb{I}_{4\times4} \\ &\quad + \begin{bmatrix} 0 & 2b_x^2 & b_x & b_x \\ 2b_x^2 & 0 & b_x^2 & b_x^2 \\ b_x & b_x & 0 & 2b^3 \\ b_x & b_x^2 & 2b_x^3 & 0 \end{bmatrix}, \quad \text{and } R \\ &= 10^{-4}\mathbb{I}_{2\times2}. \end{aligned} \tag{26}$$

where $\mathbb{I}_{n\times n}$ is the $n \times n$ identity matrix. These noise matrices were chosen empirically in order to achieve the best performance of the filter.

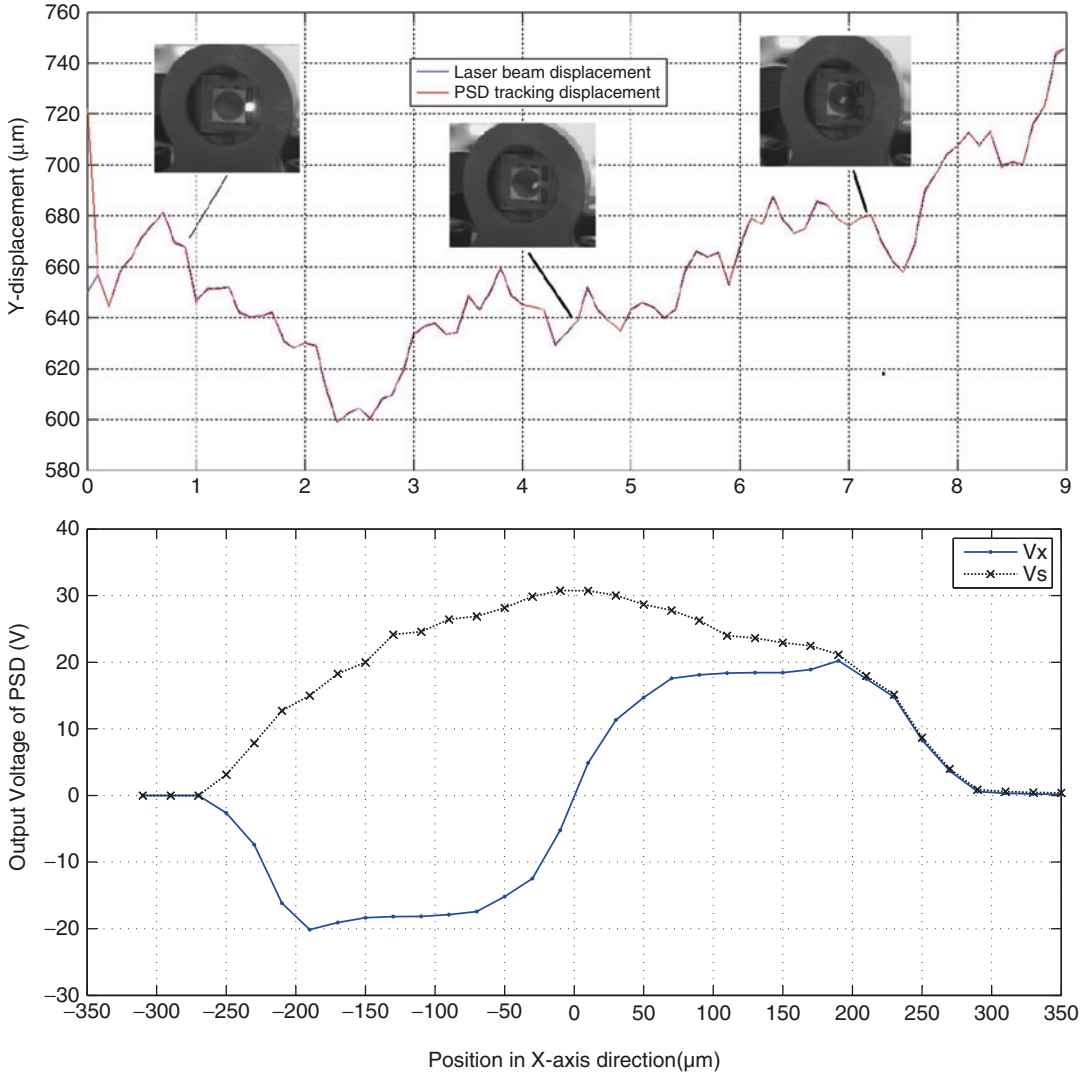
To evaluate the performance of Kalman filter, we settled the maximum intensity of the laser beam in the PSD center. The resulting estimation of laser beam motion prediction using KF is presented in Fig. 13. The blue color represents the PSD trajectory resulting from the KF and red color the given variation in the position of the laser beam. At first glance, the KF succeeds to reach quickly the real position and to follow the laser beam motion variation very closely. As one can see in Fig. 11c, the PSD trajectory presents a convergence phase (bottom right part of the plot). This convergence step is due to the KF that has to reach its optimum prediction along with tracking the laser beam motions. As illustrated, the performances of the KF, in terms of precision, converge to the real position of the laser beam with a minimal estimation error around 0.01 nm. These results demonstrate clearly the sensitivity performances of the KF algorithm to detect and to estimate small variations of the laser beam motions.

Laser Beam Maximum Intensity Tracking

In this scenario, the PSD has to track the laser beam motion and find its maximum intensity.



Nanorobotics for Synchrotron Radiation Applications, Fig. 13 Experimental tracking with Kalman filter: motion estimation along *x*-*y* axis with corresponding errors and 2D displacement



Nanorobotics for Synchrotron Radiation Applications, Fig. 14 Laser beam detection of maximum intensity

In addition here, the right manipulator moves the laser beam with a composite motion defined by a constant displacement together with a Brownian motion. The resulting motion of the PSD along the *y*-axis is shown in Fig. 14. These results illustrate a typical tracking task of a composite motion with a fast convergence time of the dual micro-/nanomanipulator controller despite the constant bias displacement. Furthermore, the dual controller is able to find and to maintain the maximum beam light intensity in the center of the PSD against external perturbations. Let us also notice that this approach assumes variations of laser light intensity during motion, measurement noise, and high motion dynamics. Moreover, these experiments demonstrate the robust estimation of the laser beam motion in real time. As expected, the filtered estimate exhibits smaller variations. Finally, the master–slave controller with decoupled sensitivity is optimized in terms of tracking error.

Summary

In this entry, the authors discussed the major applications of nanorobotics systems in synchrotron radiation facility centers for the characterization of nanomaterials. As the focalization of X-ray beam improves regularly (presently less than 100×100 nm^2), it becomes possible to analyze smaller and smaller sample [2] (10^8 atoms which corresponds to a 100 nm radius object). Based on this X-ray optical progress and to address the needs of the nano-community, specific beamlines dedicated to nanosciences are currently developed or will be developed in the frame of the ESRF (European Synchrotron Radiation Facility located in Grenoble, France) Upgrade Programme, for example, UPBL1, UPBL4, and UPBL5 [3]. This entry has presented a study on the control problem of a laser beam illuminating a target when subjected to dynamic disturbances using light intensity feedback. The main idea is to guide and track the beam with a hybrid micro-/nanomanipulator which is driven by a control signal generated by processing the beam intensity sensed by a four-quadrant photodiode. The simulations and experiments demonstrated the efficiency of the approach when submitted to external disturbances, such as platform vibration, piezoelectric actuator thermal drifts, photodetector noises, Brownian motion of laser beam, and atmospheric turbulence. The use of a Kalman filter algorithm to estimate the laser beam motion has proven to be efficient despite the encountered high dynamics at such nanometer scale. Different experiments of single biological cell irradiation with a micro X-ray beam subjected to environmental perturbations are actually investigated with the proposed beam tracking system.

References

1. Duesberg, G.S.: Polarized Raman spectroscopy on isolated single-wall carbon nanotubes. Phys. Rev. Lett. **85**, 5436 (2000)
2. Schroer, C.G., et al.: Coherent X-ray diffraction imaging with nanofocused illumination. Phys. Rev. Lett. **101**, 090801 (2008)
3. The ESRF Purple book. http://www.esrf.eu/Upgrade/documentation/purple-book/
4. Rodrigues, M. S., et al.: Local detection of X-ray spectroscopies with an in-situ atomic force microscope. J. Instrum. **3**(12), 12004 (2008)
5. Scheler, T., et al.: Probing the elastic properties of individual nanostructures by combining in-situ AFM and micro X-ray diffraction. Appl. Phys. Lett. **94**, 023109 (2009)
6. Ashkin, A., et al.: Observation of a single-beam gradient force optical trap for dielectric particles. Opt. Lett. **11**, 288 (1986)
7. Guidon, M.: Longitudinal optical binding of high optical contrast microdroplets in air. Phys. Rev. Lett. **96**, 143902 (2006)
8. Dhez, O., et al.: X-tip: a new tool for nanoscience or how to combine X-ray spectroscopies to local probe analysis. AIP Conf. Proc. **879**, 1391–1394 (2007)
9. Kim, P., et al.: Nanotube nanotweezers. Science **286**, 2148–2150 (1999)
10. Molhave, K., et al.: Pick-and place nanomanipulation using microfabricated grippers. Nanotechnology **17**, 2434–2441 (2006)
11. Hashiguchi, G.: Micromachined nanotweezer for DNA manipulation in aqueous solution. Anal. Chem. **75**, 4347–4350 (2003)
12. Watanabe, H., et al.: Single molecules DNA devices measured with triple-probe atomic force microscope. Appl. Phys. Lett. **79**, 2462–2464 (2001)
13. Maruo, S., Ikuta, K., Korogi, H.: Submicron manipulation tools driven by light in a liquid. Appl. Phys. Lett. **82**, 133 (2003)
14. Ehrlicher, A., Betz, T., Stuhrmann, B., Koch, D., Milner, V., Raizen, M.G., Kas, J.: Guiding neuronal growth with light. Proc. Natl. Acad. Sci. U. S. A. **99**, 16024 (2002)
15. Stuhrmann, B., Gogler, M., Betz, T., Ehrlicher, A., Koch, D., Kas, J.: Automated tracking and laser micromanipulation of motile cells. Rev. Sci. Instrum. **76**, 035105 (2005)
16. Xie, H., Régnier, S.: High-efficiency automated nanomanipulation with parallel imaging/manipulation force microscopy. IEEE Trans. Nanotechnol. **11**(1), 21–33 (2012)
17. Pawel K. Orzechowski, P.K., Gibson, J.S., Tsao, T-C: Optimal disturbance rejection by LTI feedback control in a laser beam steering system. In: 43rd IEEE Conference on Decision and Control, December 14–17, 2004, pp. 2143–2148, Atlantis (2004)
18. Lee, C.C., Park, J.: Temperature measurement of visible light-emitting diodes using nematic liquid crystal thermography with laser illumination. IEEE Photon. Technol. Lett. **16**(7), 1706–1708 (2004)
19. Evans, R.B., Griesbach, J.S., Messner, W.C.: Piezoelectric microactuator for dual stage control. IEEE Trans. Mag. **35**, 977–981 (1999)
20. Perez-Arancibia, N.O., Gibson, J.S., Tsao, T.-C.: Observer-based intensity-feedback control for laser

beam pointing and tracking. IEEE Trans. Control Syst. Technol. **20**(1), 31–47 (2012)
21. Devasia, S., Eleftheriou, E., Moheimani, S.: A survey of control issues in nanopositioning. IEEE Trans. Control Syst. Technol. **15**, 802–823 (2007)
22. Lee, C., Salapaka, S.M.: Robust broadband nanopositioning: fundamental tradeoffs, analysis, and design in a two-degree-of-freedom control framework. Nanotechnology **20**, 03501 (2009)
23. Mokaberi, B., Requicha, A.: Drift compensation for automatic nanomanipulation with scanning probe microscopes. IEEE Trans. Autom. Sci. Eng. **3**, 199–207 (2006)
24. Nguyen, T.T., Amthor, A., Ament, C.: High precision laser tracker system for contactless position measurement, 2011. In: IEEE International Conference on Control System, November 25–27, 2011, pp. 97–102, Penang, Malaysia (2011)
25. Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Bas. Engr. **82**, 35–45 (1960)
26. Durrant-Whyte, H., Henderson, T.C.: Multisensor data fusion. In: Springer Handbook of Robotics, pp. 585–610. Springer, Berlin/Heidelberg (2008)
27. Kuchimaru, T., Sato, F., Higashino, Y., Shimizu, K., Kato, Y., Iida, T.: Microdosimetric characteristics of micro X-ray beam for single cell irradiation. IEEE Trans. Nucl. Sci. **53**(3), 1363 (2006)
28. Amari, N., Folio, D., Ferreira, A.: Motion of a micro/nanomanipulator using a laser beam tracking system. IEEE Int. Conf. Control Syst. Int. J. Optomechatron. **8**(1), 30–46 (2011)

Nanorods

► Nanostructures for Energy
► Physical Vapor Deposition

Nanoscaffold

► Ligand-Directed Gold-Phage Nanosystems

Nanoscale Drug Vector

► Fullerenes for Drug Delivery

Nanoscale Fluid Dynamics

► Applications of Nanofluidics

Nanoscale Fluid Mechanics

► Applications of Nanofluidics
► Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering

Nanoscale Heat Transport

► Thermal Conductivity and Phonon Transport

Nanoscale Mass Transport

► Applications of Nanofluidics

Nanoscale Metal Alloys

► Nanoalloy Simulation

Nanoscale Metals

► Simulation of Supported Metal Clusters

Nanoscale Particle

► Gas-Phase Nanoparticle Formation

Nanoscale Printing

Timothy J. Merkel[1] and
Joseph M. DeSimone[1,2,3,4,5,6,7,8]
[1]Department of Chemistry, University of North Carolina, Chapel Hill, NC, USA
[2]Department of Pharmacology, Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA
[3]Carolina Center of Cancer Nanotechnology Excellence, University of North Carolina, Chapel Hill, NC, USA
[4]Institute for Advanced Materials, University of North Carolina, Chapel Hill, NC, USA
[5]Institute for Nanomedicine, University of North Carolina, Chapel Hill, NC, USA
[6]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA
[7]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC, USA
[8]Sloan–Kettering Institute for Cancer Research, Memorial Sloan–Kettering Cancer Center, New York, NY, USA

Synonyms

Dip-pen nanolithography; Imprint lithography; Microcontact printing; Molecular printing; Nanoimprint lithography; Nanotransfer printing; PRINT®; Replica molding; Soft lithography

Definition

Nanoscale printing refers to the patterning of surfaces in two or three dimensions with at least one feature on the submicron length scale. The printing techniques may resemble macroscopic printing, in which a material, or ink, is transferred to a substrate by a stamp or a stylus. However, nanoscale printing may also refer to embossing or molding processes, in which topographical features are transferred from a three-dimensional mold to a surface by either elevating or depressing areas or by the addition of material. For the case of patterning features in three dimensions, two distinct outcomes are possible: the features generated can remain connected by an underlying layer of material, yielding an embossed film or, alternately, the molded features can be free of the connecting layer, yielding discrete particles. Arrays of features have been used to direct the assembly of materials for applications in molecular electronics, genetic analysis, and disease detection. Particles generated by these methods have found a myriad of uses, including those in drug delivery and biomedical imaging.

Overview

Methods used to generate nanoscale structures are generally divided into "top-down" and "bottom-up" approaches. In bottom-up approaches, features are generated from the atomic or molecular level to the supermolecular or macroscopic level and utilize interactions between molecules or particles to achieve two- or three-dimensional structures [1]. For top-down methods, materials are processed at the desired size scale, with external forces directing the material to form the desired architecture. The development of nanoscale printing methods for drug delivery and imaging applications have been enabled by the remarkable developments in the semiconductor industry over the last 50 years to reduce feature sizes in order to increase the number of transistors in a microchip [2]. Now these top-down technologies are poised to be exploited for benefit in a variety of new industries, especially in the life sciences and energy sciences fields.

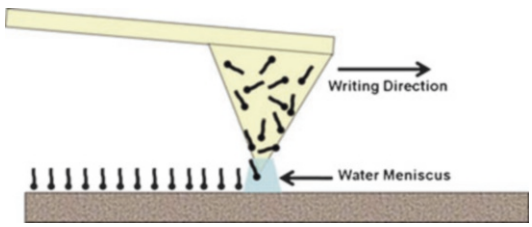
Patterning Surfaces

Two complementary nanoscale printing techniques have emerged for the transfer of reagents, both chemical and biological to a surface: techniques based on scanning probe microscope-based lithography (SPL) and soft lithography-based

techniques. Based on the development of scanning probe microscopy in the 1980s, SPL nanofabrication techniques have emerged that exploit the use of the tip of the scanning probe to manipulate matter on a surface. These methods offer the advantage of extremely high resolution as these probes are capable of manipulating individual atoms or molecules. However, this typically requires high vacuum and temperatures which approach absolute zero, and is not readily scalable for patterning of large areas [2]. While these methods represent powerful tools for high-resolution patterning over a small area, probe tips were not used in a printing method until the advent of dip-pen nanolithography (DPN) in 1999 [3].

Patterning with Dip-Pen Nanolithography

Using an "ink"-coated atomic force microscope (AFM) tip as a stylus, DPN is a direct-write method in which soft or hard materials can be deposited to a surface with sub 50 nm resolution [3]. Developed in the Mirkin laboratory at Northwestern University in 1999, DPN relies on humidity to cause the spontaneous formation of a meniscus between the AFM tip and the surface (Fig. 1). The ink is deposited along this meniscus, with the deposition rate dependent upon factors such as the nature of the ink, the contact time between the tip and the substrate, and the coverage of the ink on the AFM tip. The first example of DPN deposited an alkanethiol onto a gold substrate, with the strong interactions between the gold substrate and the thiol overcoming the weaker interaction between the tip and the thiol to drive the formation of patterns on the surface. A variety of metallic, semiconducting and insulating substrates have been used with DPN [3]. The technique has proven to be quite versatile with respect to the composition of the ink; structures have been formed from organometallics, polymers, colloidal nanoparticles, metal ions, and a range of biologicals including DNA, proteins, peptides, and affinity templates which allowed for the subsequent immobilization of viruses [2].



Nanoscale Printing, Fig. 1 A schematic showing the dip-pen nanolithography (*DPN*) technique. In DPN, ink molecules are transferred from the tip of an atomic force microscope stylus to a surface through a water meniscus

A primary application of DPN techniques is the generation of biological nanoarrays for the study of cellular genetics and rapid screening of diseases. DPN-generated nanoarrays of an antibody have been used to screen for the human immunodeficiency HIV-1 virus antigen in serum samples. This study demonstrated an increase in detection limits of orders of magnitude over conventional enzyme-linked immunosorbent assays [3]. Nanoarrays generated in this way have advantages of holding 10^4–10^5 more features than conventional microarrays, offering the opportunity for either decreased sample volume or for the screening of a larger number of targets per experiment.

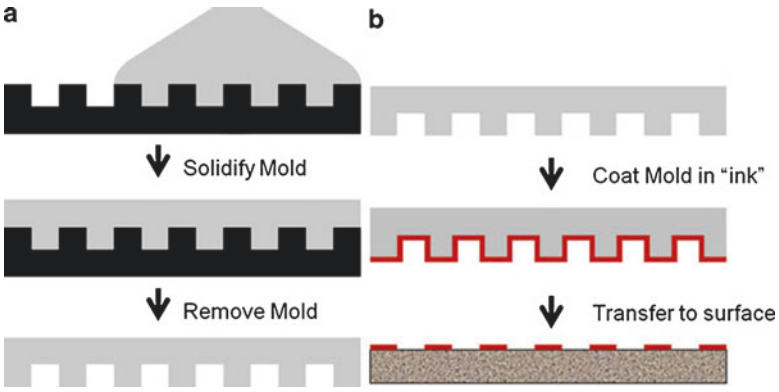
While printing with a single AFM tip over large areas would be impractically slow, massively parallel arrays of cantilevers have been fabricated to increase the throughput of the technique. Arrays with as many as 1.3 million cantilevers have been microfabricated toward this end, increasing the throughput by four orders of magnitude. A number of variations to DPN have emerged to broaden the applicability of the technique, with modification of the tips a common theme. Coating the AFM tip with a silicone polymer increases the amount of ink that can be loaded to the tip, allowing for printing of larger areas. The inclusion of microfluidic channels to deliver ink to the tips, termed nanofountain pen (NFP) probes, ensures a constant supply of ink to the surface. Features with greater than 50 nm resolution have been printed with NFP probes. The technique also serves to keep the ink molecules solvated, an aspect which can be critical for biological applications. Heating of the AFM tip (thermal DPN) can allow for deposition of insoluble ink molecules that are solids at room temperatures.

Patterning with Soft Lithography Techniques

Pioneered in the 1990s in the Whitesides laboratory at Harvard University, soft lithography begins with the generation of a rigid template which is used to generate an elastomeric mold or stamp which replicates the three-dimensional shape of the original template. Master templates are typically prepared by photolithographic techniques, which were pioneered in the electronics industry. The molds of these master templates are typically generated from poly (dimethyl siloxane) (PDMS), although polyurethanes, polyamides, perfluoropolyethers, and natural polymers such as agarose have been used for this purpose. To fabricate the mold, the desired polymer is poured onto the master template and cured, generating a relief structure with the raised portions of the stamp defining the pattern to be transferred.

Microcontact printing (μCP) is the process by which the mold is used as a stamp after it is coated with an ink, and subsequently brought into contact with a substrate to transfer the ink (Fig. 2). The ink transfers from the stamp by chemisorption or physisorption to the surface. While alkanethiol inks transferred to a gold substrate are, perhaps, the most common usage of μCP, the technique has also been used to pattern proteins, DNA, cells, colloids, and polymers [4] onto a variety of surfaces. The flexibility of the elastomeric stamp allows for conformal atomic level contact to be achieved on flat and curved surfaces, and over large stamp areas. The μCP technique has advantages as it is low-cost, is high-throughput and is easy to use which has earned it popularity as a tool among researchers. While the low cost of the elastomeric molds has rendered μCP accessible to researchers in many fields, there are some limitations to the technique. Patterns are determined by the original master template, so that each stamp can only produce one pattern. The mechanical properties of the stamp can limit the flexibility of the design; features spaced too widely (>20 times the feature height for typical PDMS) or with aspect ratios outside of the 0.2–2 range can cause defects from deformation of the stamp [2]. Stamping of features smaller than 150 nm is a challenge using conventional methods, though features as small as 80 nm have been prepared by using DPN to modify the surface of a flat stamp to passivate surrounding areas with a perfluorinated monolayer.

Nanotransfer printing (nTP) is the process of transferring a thin solid film from a nanoscale-featured stamp to a substrate. In this case, the stamp can be either rigid or flexible and the material transferred is typically a metal. In a typical process, the stamp is coated with a thin film (e.g., 20 nm of gold) and brought into contact with the desired surface. The film can transfer from the stamp by a number of methods, including binding



Nanoscale Printing, Fig. 2 A schematic for microcontact printing (*μCP*). (**a**) The fabrication of a mold (*top* to *bottom*). The mold material (*gray*) is poured onto a master template with topographical features (*black*). The liquid mold materials solidifies and is removed from the master template, producing a mold with the inverse of the features found on the master template. (**b**) The μCP process (*top* to *bottom*). The elastomeric mold is coated with a chemical ink (*red*). Ink is transferred to the surface by contact to the raised features of the mold

to reactive self-assembled monolayers (SAMs) like alkanethiols, cold welding between two metal layers, or condensation between silanol or titanol groups [5]. Applications for nTP-derived structures are typically electronic in nature, including organic thin film transistors, capacitors, and electrostatic lenses. Features as small as 70 nm have been fabricated with nTP.

The recent innovation of polymer pen lithography (PPL) in 2008 combines many of the advantages of μCP and DPN by using an elastomeric array of tips to print a pattern. With PPL, a PDMS stamp consisting of an array of tips is affixed to a piezo scanner and delivers an ink with spot sizes from 65 nm to over 10 μm determined by varying the force and time over which the ink is delivered [6]. The elastomeric arrays of PPM pens, which can contain as many as 10 [8] pyramid-shaped tips, are made from a master prepared by traditional photolithographic techniques. The cost of preparing new arrays of tips is greatly reduced for PPL compared to DPN, with the cost of a new PPL array estimated at less than $1 after the fabrication of the initial master template [2]. The resulting arrays are transparent, allowing for registration of the cantilevers with previously printed features. The master template used to generate the arrays can be filled by injet printing, allowing it to be used as an array of inkwells with which the array of tips is perfectly aligned. The entire elastomeric array absorbs ink, acting as a reservoir and allowing for printing over large areas without reinking.

Molding and Embossing Methods

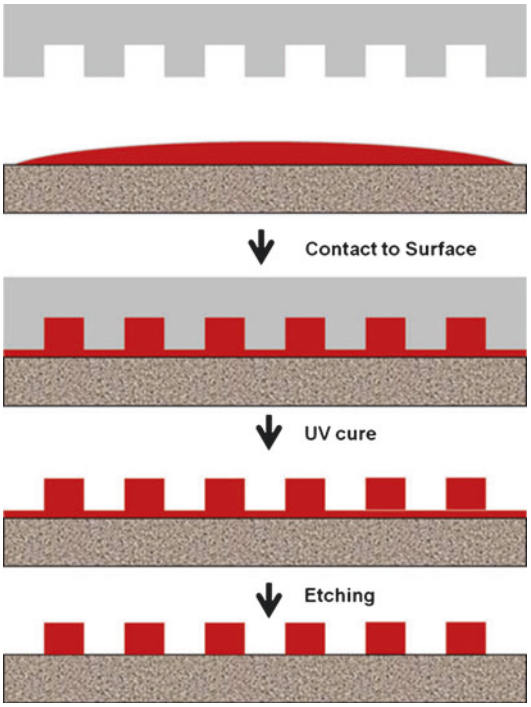
The molding process involves pouring a liquid precursor onto a nanopatterned template and curing or solidifying it into a solid which replicates the original features. Molding features on the nanoscale can be performed with both hard and soft molds. This is the basis for techniques such as step-and-flash imprint lithography (SFIL), replica molding (RM), and pattern replication in non-wetting templates (PRINT). Embossing or imprinting methods use a mold to transfer features to an initially flat polymer film. Embossing methods include nanoimprint lithography (NIL), solvent-assisted micromolding (SAMIM), and PRINT.

Hard molds have been fabricated out of quartz, silicon, and metals, though quartz and silicon are the most commonly used. Features as small as 10 nm have been transferred into silicon, while quartz molds have a minimum size of 20 nm [4]. Though soft molds are less costly and easier to fabricate, rigid molds offer several advantages over softer molds. Rigid molds will show little deformation over large areas with the pressures required for embossing. Hard molds are stable at the temperatures required for cross-linking polymer precursors and are chemically inert to them. Release of the mold from the patterned substrate can be difficult, depending on the surface properties of the mold and the molded polymer, and release coatings of fluoroalkyl silanes are typically linked to the mold surface to facilitate release. Quartz substrates are transparent to UV light and visible light, while silicon is not, allowing for photochemical cross-linking of prepolymer with quartz molds.

Step-and-Flash Imprint Lithography

Step-and-flash imprint lithography replicates the topography of a quartz mold using a photocurable prepolymer solution as the molded material at room temperature using low applied pressure. The prepolymer solution consists of a monomer and a photoinitiator with hardening of the material proceeding by free-radical polymerization upon exposure to UV light. An important consideration for complete filling of the mold is the viscosity of the prepolymer, or monomer, solution; low molecular weight monomers are used due to their low viscosity ($<$5cPs) which allows for complete filling of the mold features. The quartz mold is pressed down onto a substrate which is coated with a liquid prepolymer and cured with UV light. While the mold features fill completely, they remain connected by a thin residual layer of cured monomer. This layer can be removed by an etching process (such as reactive ion etching), yielding discrete features on the substrate (Fig. 3).

The quartz molds used in SFIL are typically coated with a release layer to facilitate release of the mold from the cured polymer and to increase the lifetime of the mold. Failure of the mold to release cleanly can result in broken mold features,



Nanoscale Printing, Fig. 3 A schematic for step-and-flash imprint lithography (*SFIL*). From *top* to *bottom*: A UV transparent quartz mold (*gray*) is brought into contact with a thin coating of a prepolymer solution (*red*) on a surface. The prepolymer is exposed to UV light through the quartz mold, leaving a patterned polymer surface after removal of the mold. An etching step removes the thin layer of polymer which connected the raised features

which adds to the cost of this technique. The first reported coatings were fluorinated silanes, which provided a mold lifetime of less than 100 uses. Newer coatings have extended this lifetime to the patterning of over 1,500 substrates [4].

In addition to the generation of patterned surfaces, SFIL can be used to generate polymeric particles. In this case, after the etching step removes the connecting layer from between the particles, a subsequent step provides release of the cured polymer from the substrate. Particles can be removed by physical agitation or by dissolution of a sacrificial surface coating (i.e., poly vinyl alcohol) on the initial substrate. While the etching process, which involves exposure to an oxygen plasma, could be considered quite harsh, the biological activity of a protein encapsulated in SFIL particles was maintained after 20 s of exposure to plasma [7]. SFIL particles have potential for use in drug delivery, as particles containing an enzymatically degradable peptide sequence and carrying a DNA plasmid cargo were fabricated and demonstrated successful cargo release in an in vitro experiment [8].

Nanoimprint Lithography

Nanoimprint lithography (NIL) involves the transfer of a pattern from a rigid stamp to a polymer film via application of pressure at temperatures above the glass transition temperature of the polymer. The typical mold material is silicon with features as small as 5 nm patterned. While NIL is typically used to pattern polymers, biomolecules have also been successfully patterned. High pressures and temperatures are required with NIL, putting stress on the molds and limiting their lifetime to $\sim$ 50 imprints. The high viscosity of many of the polymers selected for patterning is another issue with this technique. Low molecular weight polymers may be desirable due to their lower viscosities and relative ease of processing [9].

Imprint Lithography

Soft, or imprint, lithography generates elastomeric molds from rigid substrates and uses the molds to replicate the original topographical features. Imprint lithography methods were first seen as a low-cost, high-resolution alternative to photolithographic techniques for the replication of patterns in the micro- and nanoscales. The typical mold materials have included cross-linked poly (dimethyl siloxane) (PDMS or silicone) or perfluorinated polyethers which have low toxicity, are extremely flexible, and have low surface energies, all of which make them excellent materials for molding applications.

Replica Molding

The replica molding (RM) process consists of transferring the topographical pattern of a master template to an elastomeric mold (typically made from poly (dimethyl siloxane) (PDMS)) which is in turn used to transfer the pattern to a liquid which is cured, or solidified, in contact with the mold. Molded materials can be photo or thermally curable, though there are limitations. Many low

molecular weight monomers display some solubility in PDMS, causing the mold to swell and limiting the molding process. Molded materials include polymers, gels, precursors to ceramics and carbons, luminescent phosphors, salts, and colloids [5]. The use of the soft PDMS mold allows for molding of non-flat surfaces, an advantage over hard template molding techniques. While molds may only remain viable for approximately 20 replications, the mold material is inexpensive and many molds can be prepared from a single master template. Replication of the high-cost master templates in a low-cost material makes this technique accessible to a broad range of researchers.

Replica molding in PDMS has transferred features as small as 3 nm wide structures with 0.5 nm vertical deflections. Based on the physics of van der Waals interactions, the theoretical ultimate size limit for molding with PDMS is less than 0.5 nm. Another advantage of PDMS is the low surface energy of the material, which promotes facile release from the master template as well as the patterned surface. The PDMS mold surface can be modified to tailor the hydrophilicity; treatment with oxygen plasma and with fluorosilane have been explored to increase or decrease hydrophilicity, respectively. Changing the surface properties of the mold has implications in wetting and filling of the mold as well as mold release [8].

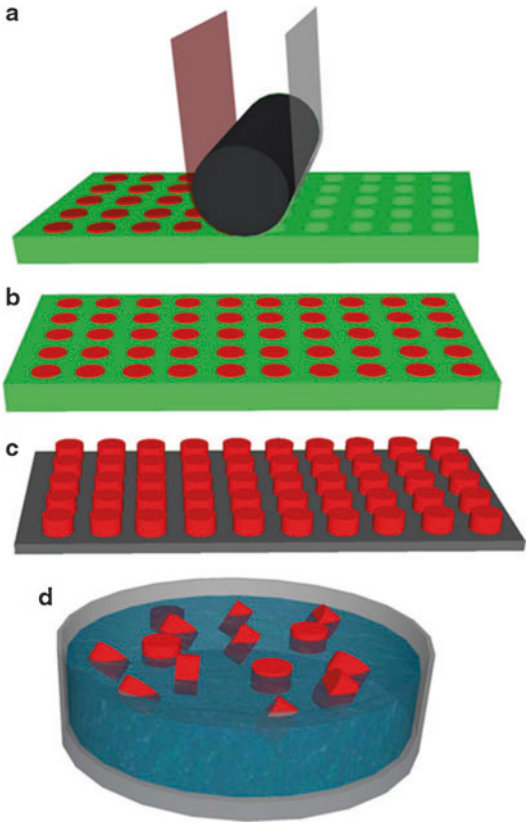
One limitation of RM techniques is difficulty in eliminating the thin layer of material which connects the patterned features, called a flash layer. While this layer can be etched away generating isolated structures or free particles, as in the case of SFIL, exposure of biological or otherwise sensitive materials to this harsh condition is thought to be undesirable due to the risk of degradation. However, there are many examples of particle replication using PDMS molds. Particles were generated from thermoplastics, thermosets, and hyperbranched polymers, though these were all on the micron size range and the generation of nanoparticles remains a challenge.

PRINT

The particle (or pattern) replication in non-wetting templates (PRINT®, Liquidia Technologies) technique, developed in the DeSimone laboratory at the University of North Carolina at Chapel Hill and commercialized by Liquidia Technologies, has extensively focused on the use of a perfluoro polyether (PFPE) elastomer as the mold material to replicate the features of a rigid master template. Though similar to other soft lithography techniques, the use of PFPE offers several advantages over PDMS. First, the PFPE is not as responsive to solvents as PDMS, which tends to swell when exposed to organics. This allows for the patterning of a wide variety of substrates with varied properties. Second, PFPE has a lower surface energy than PDMS which is advantageous for particle making applications. Due to the non-wetting nature of the PFPE mold, the depressions in the mold can be filled without wetting the landmass in between, yielding distinct particles without the troublesome flash layer. Thirdly, the properties of the PFPE allow for facile release of the particles from the mold without the modification of the mold surface as is required with PDMS molds. Lastly, PFPE forms a less flexible mold than PDMS, which is advantageous for high fidelity molding since the PFPE molds are less prone to unwanted deformation than their PDMS counterparts [10].

A drawback of the PRINT technique is the cost associated with PFPE when compared to the relatively inexpensive PDMS. The amount of PFPE can be minimized with the use of an inexpensive backing material coated with a very thin film of the PFPE molding material. Molds produced in this way have facilitated scale-up and commercialization of the PRINT technique, with a rate of 1,000 ft of mold per day [7]. Scalability is crucial to the application of PRINT, and other particle fabrication techniques, in nanomedicine due to the high material demands of clinical trials.

PFPE has been used to replicate the topography of a variety of nanoscale objects, including carbon nanotubes with diameters of 2 nm and soft objects such as block copolymer micelles and viruses [8], which were then replicated on a substrate by an embossing process. To generate particles, the PRINT mold is filled with a liquid preparticle solution, with the excess solution removed with a high surface-energy polymer



Nanoscale Printing, Fig. 4 A schematic for particle replication in non-wetting templates (*PRINT*). (**a**) A preparticle solution (*red*) is distributed evenly into an elastomeric mold made from perfluoro polyether (*green*) by a roller (*black*) covered by a high-surface-energy sheet (*gray*) which removes excess solution. (**b**) The particle solution is solidified in the mold. (**c**) Particles are removed from the mold with a harvesting film (*gray*). (**d**) The film is dissolved, freeing the particles

sheet via a wicking process. The preparticle solution is solidified and particles removed from the mold by physical agitation or by transfer to a surface through use of an adhesive which is subsequently dissolved, yielding free particles (Fig. 4). Particles have been generated from a variety of materials, including cross-linked PEG hydrogels, linear biocompatible polyesters, proteins, therapeutics, and block copolymer micelles. A variety of cargos have been encapsulated in PRINT particles, including proteins, anticancer drugs, nucleic acids, as well as a variety of imaging agents such as iron oxides, gadolinium chelates, and ^{64}Cu chelates [10]. PRINT has evolved to a roll-to-roll process and includes the use of a very important film splitting technique that allows the flash-free filling of the cavities in the mold to generate the individual particles. Such a roll-to-roll process is amenable to a wide range of materials to be processed using PRINT including the use of an increasing variety of mold materials that are being employed which are tailored to the chemistries being used to fill the mold [7].

Particles generated by nanoscale printing methods like PRINT and SFIL are of growing interest to researchers in the areas of drug delivery and medical imaging. These processes offer precise control over the size, shape, and composition of the particle, with broad capabilities with respect to the loading of imaging agents and/or therapeutics. PRINT particles of various shapes and sizes were used to probe cellular uptake kinetics and mechanisms with respect to particle size and shape.

Statement of disclosure: Joseph DeSimone is a founder, member of the board of directors, and maintains a financial interest in Liquidia Technologies. Liquidia was founded in 2004 to commercialize PRINT technology and other discoveries of Professor Joseph DeSimone and colleagues at the University of North Carolina at Chapel Hill.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Dip-Pen Nanolithography
- ▶ Microcontact Printing
- ▶ Nanoimprint Lithography
- ▶ Nanoimprinting

References

1. Claridge, S.A., Castleman, A.W., Khanna, S.N., Murray, C.B., Sen, A., Weiss, P.S.: Cluster-assembled materials. ACS Nano **3**, 244 (2009)
2. Braunschweig, A.B.., Huo, F.W., Mirkin, C.A.: Molecular printing. Nat. Chem. **1**, 353 (2009)
3. Salaita, K., Wang, Y.H., Mirkin, C.A.: Applications of dip-pen nanolithography. Nat. Nanotechnol. **2**, 145 (2007)
4. Gates, B.D., Xu, Q., Stewart, M., Ryan, D., Willson, C.G., Whitesides, G.M.: New approaches to

nanofabrication: molding, printing, and other techniques. Chem. Rev. **105**, 1171 (2005)
5. Gates, B.D.: Nanofabrication with molds and stamps. Mater. Today **8**, 44 (2005)
6. Zheng, Z., Daniel, W.L., Giam, L.R., Huo, F., Senesi, A.J., Zheng, G., Mirkin, C.A.: Multiplexed protein arrays enabled by polymer pen lithography: addressing the inking challenge. Angew. Chem. **121**, 7762 (2009)
7. Merkel, T.J., Herlihy, K.P., Nunes, J., Orgel, R.M., Rolland, J.P., DeSimone, J.M.: Scalable, shape-specific, top-down fabrication methods for the synthesis of engineered colloidal particles. Langmuir **26**, 13086 (2010)
8. Canelas, D.A., Herlihy, K.P., DeSimone, J.M.: Top-down particle fabrication: control of size and shape for diagnostic imaging and drug delivery. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. **1**, 391 (2009)
9. Zankovych, S., Hoffmann, T., Seekamp, J., Bruch, J. U., Torres, C.M.S.: Nanoimprint lithography: challenges and prospects. Nanotechnology **12**, 91 (2001)
10. Jeong, W., Napier, M.E., DeSimone, J.M.: Challenging nature's monopoly on the creation of well-defined nanoparticles. Nanomedicine-Uk **5**, 633 (2010)

Nanoscale Properties of Solid–Liquid Interfaces

Lucio Colombi Ciacchi and Susan Köppen
Hybrid Materials Interfaces Group, Faculty of Production Engineering and Bremen Center for Computational Materials Science, University of Bremen, Bremen, Germany

Synonyms

Solid–liquid interfaces

Definition

A solid–liquid interface is the boundary region separating a solid phase from a liquid phase. While its extension is strongly dependent on the particular phases, for poorly soluble solids and small-molecule liquids such as water, it spreads over a thickness of the order of a few nm. The properties of solid–liquid interfaces can be described at a macroscopic, continuum level, but are ultimately governed by atomistic details. These are the subject of the present entry, which is written with a particular emphasis on results stemming from molecular dynamics simulations and focuses especially on interfaces between inorganic solids and water.

Macroscopic Description of Solid–Liquid Interfaces

The macroscopic interaction between a liquid phase (l) in contact with a solid surface (s) is described by the *interface energy* γ_{sl}, defined as the (positive) change in free energy associated with a unitary change of the interface area. The spreading of a liquid droplet over a solid surface can be quantified by the equilibrium *contact angle* θ at the droplet's edge, which is related to γ_{sl} and the surface energies of the two separated phases, γ_s and γ_l, via the Young equation

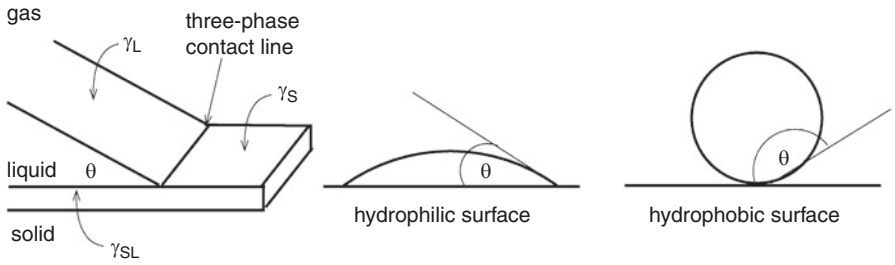
$$\gamma_{s1} + \gamma_1 \cos\theta = \gamma_s \tag{1}$$

Small values of the contact angle indicate easy spreading of the liquid phase and thus wetting of the solid surface, while large values indicate only partial spreading, up to the limit of complete dewetting for $\theta = 180°$. In the case of water, the solid is considered *hydrophilic* or *hydrophobic* depending on whether θ is smaller or larger than 90° (Fig. 1).

Chemical Details of Solid–Water Interfaces

The interface energy and the contact angle defined above capture effects at the macroscopic level that are ultimately driven by the microscopic interactions of liquid molecules with the superficial atomic layers of the underlying solid substrate.

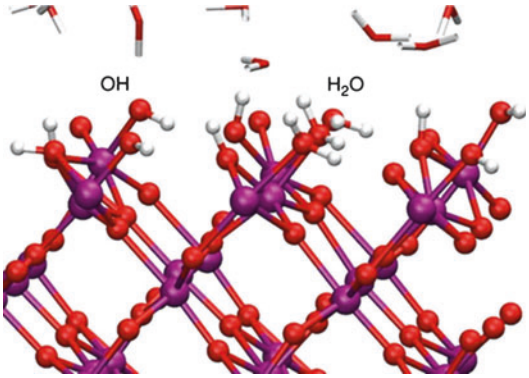
Water molecules interact with hydrophobic surfaces via weak *van der Waals forces*. These include a repulsive contribution at very small distances arising from the non-allowed mixing of molecular orbitals with electronic surface states,



Nanoscale Properties of Solid–Liquid Interfaces, Fig. 1 Scheme of the three-phase contact of a liquid droplet with a contact angle θ on a solid surface and typical water droplet shapes on a hydrophilic and a hydrophobic surface

and attractive contributions due to *London dispersion* and *induced polarization* effects. Polarization is especially important in the case of polar liquids with a high dielectric constant, such as water, close to non-reactive metallic surfaces, such as gold or defect-free graphene sheets. It is important to note that all other metal surfaces are reactive in a water environment and spontaneously form ultrathin *native oxide* layers (with a thickness of the order of 1 nm), which confer them a hydrophilic character. Border-line cases are noble metals such as platinum, to which water is able to chemisorb through relatively weak chemical bonds, but not to trigger the formation of superficial oxides under normal conditions. Non-metallic hydrophobic surfaces of technological relevance include for instance hydrogen-terminated silicon or carbon (typically produced via etching of the natively oxidized surfaces with fluoridic acid), hydrocarbon or fluorocarbon polymers such as polyethylene and Teflon, and, to a lesser extent, silicones.

The surfaces of ceramic oxides, natively oxidized metals, and polar polymers such as polyethylene oxide are hydrophilic. Water molecules bind to them strongly via hydrogen bonds, with typical bond distances in the range 1.5–2.0 Å, which are smaller than typical van der Waals distances by about 1.5 Å. Covalent bonds are readily formed upon reaction between water molecules and anhydrous inorganic solid surfaces. Depending on the bulk composition and the surface crystallographic direction, the reaction can lead either to molecularly chemisorbed water, or to dissociative chemisorption with formation of terminal OH groups (Fig. 2). Mixed molecular and dissociative



Nanoscale Properties of Solid–Liquid Interfaces, Fig. 2 Chemisorbed water layer on a rutile $TiO_2(100)$ surface, as obtained in first-principles molecular dynamics simulations. Note the mixed dissociative adsorption mode with both terminal OH and H_2O groups

adsorption is common and the resulting hydroxylation equilibrium on the surface can be determined with the help of the so-called *Multi Site Complexation Model* [1]. In analogy with the standard nomenclature of organic alcohols, terminal hydroxyl groups can be classified as: *geminal* (two groups bound to the same site), *vicinal* (two groups bound to nearest-neighbor sites), or *isolated* (otherwise).

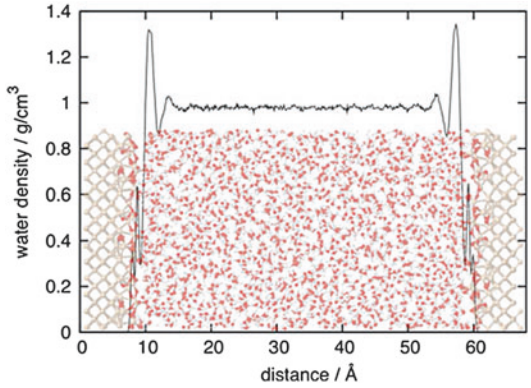
After chemisorption, the terminal groups may either donate or accept protons to or from the water solution, depending on their equilibrium dissociation constants and the solution pH. This results in the formation of charged surface sites, such as $O-$, $OH-$, H_2O^+, and an average surface charge density σ. At the macroscopic level, charged surfaces in contact with electrolyte solutions can be described by well-known theories

such as the *Gouy-Chapman electrical double layer model* [2]. The protonation/deprotonation equilibrium at the interface can be shifted by varying the pH of the solution. The pH value at which the surface is neutral is defined as the *point of zero charge* (pzc), an intrinsic property of the solid surface. This is conceptually different, but technically often difficult to distinguish, from the *isoelectric point* (IEP), defined as the pH for which the electrokinetic *ζ-potential* is zero. The latter is defined as the electrostatic potential at the edge of the layer of fluid molecules that remain immobile (either chemisorbed or otherwise strongly adsorbed) over the surface when subjected to a hydrodynamic flow. The ζ-potential at a given pH, and thus the IEP upon continuous pH variation, can be determined, e.g., from electrophoretic measurements of solid particle suspensions.

Molecular Ordering in Surface Proximity

Liquid molecules close to a solid–liquid interface experience a substantial reordering with respect to their arrangement in bulk, induced by the interaction with the surface atoms. This manifests itself with a typical increase of the liquid density in proximity of the surface, followed by density oscillations with amplitude of the order of the molecular diameter, which may reach up to several nm within the liquid phase. The density profile of water at a natively oxidized Si(100) surface, as obtained in molecular dynamics (MD) simulations [3], is reported in Fig. 3, clearly showing the oscillatory behavior described above. Noteworthy is the presence of water molecules deep inside the crevices of the atomic-scale roughness intrinsic of the native oxide layer, as indicated by the small features preceding the first main density peak. These features are characteristic of amorphous oxide phases such as glasses or superficial native oxides.

In contrast, hydrophobic surfaces are separated from the liquid phase by a vacuum gap with thickness of the order of a van der Waals distance, known as *depletion layer*. Knowing the density profiles $\rho_s(z)$ and $\rho_l(z)$ and the corresponding bulk



Nanoscale Properties of Solid–Liquid Interfaces, Fig. 3 Density profile of water in contact with a natively oxidized Si(100) surface, schematically shown in the background, as obtained in molecular dynamics simulation [3]

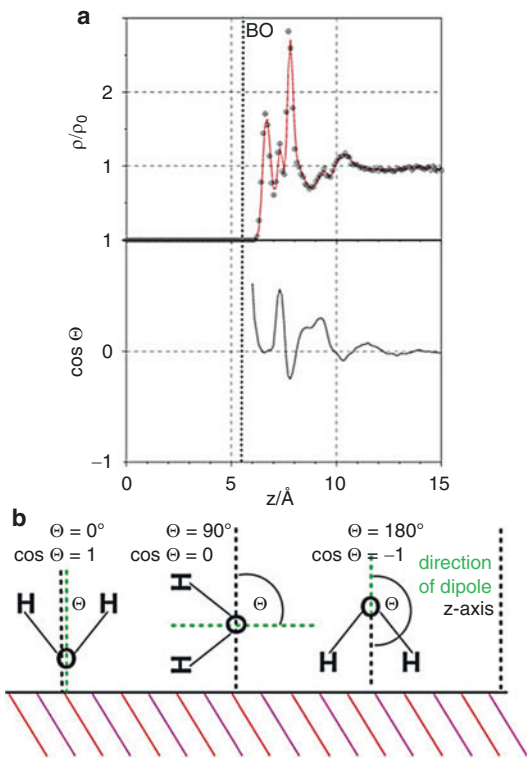
values ρ_s^B and ρ_l^B of the solid and liquid phases, the thickness δ of the depletion layer can be calculated as

$$\delta = \int_{-\infty}^{+\infty} \left(1 - \rho_s(z)/\rho_s^B\right) - \left(\rho_l(z)/\rho_l^B\right) \mathrm{d}z \quad (2)$$

Note that this definition is general for any interface between two phases, in particular also for hydrophilic, rough, or porous surfaces in contact with water, where negative values of δ can be expected. For ideally flat surfaces (where δ cannot be negative), MD simulations have shown that the degree of hydrophobicity, as measured by the contact angle θ, is related to the thickness of the depletion layer according to [4]

$$\delta^{-2} \propto (1 + \cos\theta) \quad (3)$$

In the water layers closer to the interface, water molecules assume, on average, a preferential orientation that strongly depends on the surface charge, the available terminal groups, the presence of adsorbed counterions, and the surface polarizability. This effect can be quantified by an analysis of the average angle between the water molecular axis (i.e., the molecular dipole) and the surface plane, as obtained in MD simulations. An example is shown in Fig. 4 for the case of a TiO_2 surface in contact with a saline water solution.



Nanoscale Properties of Solid–Liquid Interfaces, Fig. 4 Profiles of the density and molecular dipole orientation of liquid water at a rutile TiO_2(100) surface in contact with a 0.15 M NaCl solution at pH 9. Note the non-trivial structuring induced both by negatively charged surface sites and by adsorbed positive counterions. The angle Θ is defined as in the scheme here above [5]. The label BO indicates the position of the outermost O atoms (bridging oxygen) of the surface

Consistently with the variation of preferential water orientation, also the profile of the electrostatic potential in the liquid phase shows a characteristic oscillatory behavior [5]. Moreover, the higher molecular density, preferential orientation, and reduced mobility (see section "Self-Diffusion at Solid–Liquid Interfaces" below) of the water layers in surface proximity are at the origin of a considerably reduced *relative dielectric constant* with respect to bulk liquid water. While the latter is about 80, values around 10 describe correctly the dielectric properties of the water (and ion) layers in the interface region [6].

Slip Length at Solid–Liquid Interfaces

The presence of a depletion layer between hydrophobic surfaces and liquid water is connected with the existence of a positive *slip length* characterizing the flow of liquid at the interface. The slip length b is defined as the ratio between the relative velocity of liquid molecules at the interface with respect to the surface and the *shear rate* at the surface, or

$$v(0) - v_{surf} = b\dot{\gamma}(0) \tag{4}$$

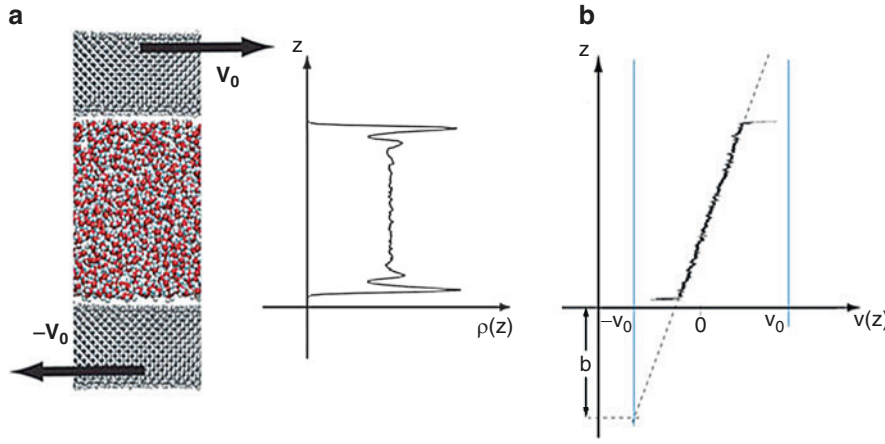
The slip length can be computed theoretically in MD simulations under shear boundary conditions, in which a bulk liquid is confined between two surfaces moving at equal and opposite velocities v_0 (Fig. 5). At the solid–liquid interface, the molecules slip over the depletion layer and move with a velocity different from v_0. The slip length is obtained by extrapolation of the linear velocity profile (see Fig. 5b).

By varying the strength of the molecule-surface interactions (i.e., the degree of hydrophobicity), Netz and collaborators have demonstrated that the slip length is related to the contact angle θ and the depletion layer thickness δ by the following relationships:

$$b \propto (1 + \cos\theta)^{-2}; \quad b \propto \delta^4 \tag{5}$$

These, combined, lead to the relationship between δ and the contact angle described in the previous section.

In the case of hydrophilic surfaces, the water-surface hydrogen bonds effectively anchor the interfacial molecules to the surfaces, and no-slip conditions are attained. Simulations under shear boundary conditions have revealed that in this case, the shear rate at the interface is smaller close to the surface than in the bulk, in a region defined roughly by the first main density peak of the layered water structure. This smaller shear rate corresponds to an increase of water viscosity, the extent of which is about half an order of magnitude and is strongly dependent on the degree of surface hydrophilicity. The increased velocity gradient in the bulk region leads, by extrapolation, to negative slip lengths for hydrophilic surfaces.



Nanoscale Properties of Solid–Liquid Interfaces, Fig. 5 Water confined between two H-terminated C (100) surfaces under shear boundary conditions, the surfaces moving at equal and opposite velocities v_0. Density profile of the water molecules (cf. Fig. 3) between the surfaces (**a**). Profile of the average molecular velocity between the surfaces (**b**). Note the non-zero relative velocity at the solid–liquid interface, defining a slip length b (Reproduced with permission from Ref. [4])

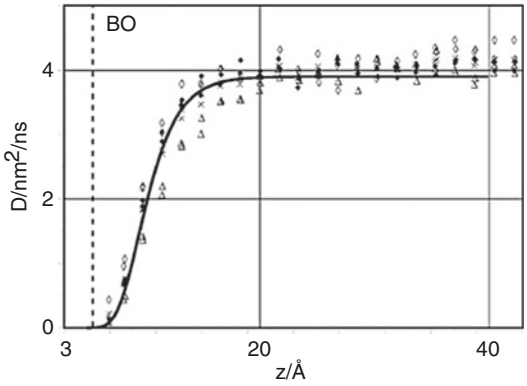
Self-Diffusion at Solid–Liquid Interfaces

The layering of the molecular structure of liquids in proximity of a solid surface and the corresponding increase of viscosity are at the origin of a reduced mobility of the molecules close to the interface. In particular, in the case of anhydrous oxide surfaces in contact with water, the first water layer is chemisorbed to reactive surface sites and thus almost completely immobile, except for the intrinsic equilibrium exchange between adsorbed and desorbed molecules. Correspondingly, the self-diffusion coefficient of water is vanishingly small at hydrophilic surfaces and increases at increasing distances from the surface. (Note that in this discussion the chemisorbed water is considered to be part of the liquid phase, but it could be considered equally well as belonging to the solid.) Quantitative investigations of the self-diffusion coefficient in MD simulations are difficult due to a number of reasons (see e.g., Ref. [7] and references therein). In particular, the long residence time of water molecules close to the interface makes difficult the convergence of the velocity autocorrelation function from which the diffusion coefficient is typically calculated.

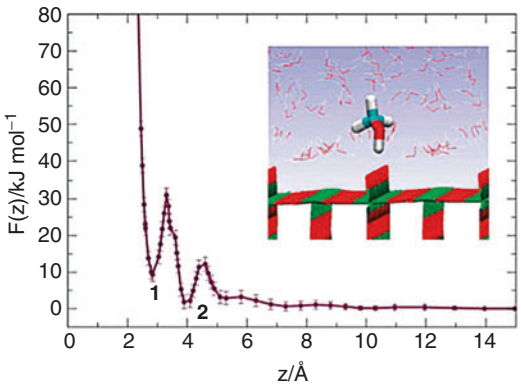
Nevertheless, the available literature results show that the diffusion coefficient of water at polar surfaces qualitatively follows a sigmoidal function, indicative of strong water trapping in the near-surface region. Simulations of a 0.15 M saline water solution in contact with the TiO_2(100) surface (Fig. 6) reveal that the bulk diffusive behavior of the water solvent is reached for distances of about 1.5 nm from the solid–liquid interface.

Molecular Adsorption at Solid–Liquid Interfaces

The adsorption of molecules, in particular organic residues and peptides, at solid–liquid interfaces has gained large attention in the past few years, due to its importance in a variety of applications such as biomedical implants, biosensors, pharmaceutical packaging, and many others. Atomistic details of the adsorption can be successfully investigated in MD simulations, provided that the force fields employed in the simulations realistically capture the chemical details of the interface. It is now well established that the water layering at the interface has an important influence both on the structural and energetic details of the adsorption. This holds for small molecules as well as for larger peptides such as collagen fragments



Nanoscale Properties of Solid–Liquid Interfaces, Fig. 6 Variation of the self-diffusion coefficient of water near a hydrophilic surface, calculated in an MD simulation for a 0.15 M NaCl solution in contact with a TiO_2(100) surface terminated according to the hydroxylation equilibrium at pH 7.4. The different symbols correspond to different directions used to compute the velocity autocorrelation function, indicating a roughly isotropic diffusive behavior [5]. The label BO is defined as in Fig. 4



Nanoscale Properties of Solid–Liquid Interfaces, Fig. 7 Free energy profile of a methanol molecule interacting with a rutile TiO_2(110)–water interface. The oscillatory behavior reflects the water layering in surface proximity (cf. Fig. 3). Note the lack of a net driving force for adsorption but the presence of relatively deep local minima that can kinetically trap the molecule close to the surface (Adapted and reproduced with permission from Ref. [9])

[8]. In the latter case, for instance, a helical molecular configuration has been observed to be stabilized at hydrophobic surfaces in part by a striking matching between the water density oscillations in surface proximity and the ordered hydrophilicity/hydrophobicity pattern of the adsorbed peptide.

Correspondingly, the free energy profile relative to the approach or retraction of a molecule from solid–liquid interfaces often presents a complex structure as a direct consequence of the oscillatory behavior of the solvent layers. This can be observed in MD simulations that sample the molecular *potential of mean force* near the interface by means of constrained methods such as *umbrella sampling* (Fig. 7).

It is interesting to note that an influence of the interfacial water structuring on free energy adsorption profiles has been observed both for hydrophilic surfaces such as TiO_2 [9] and for rather hydrophobic surfaces such as gold [10]. As mentioned above, in the latter case, an important contribution of the interaction across the solid–liquid interface is the electronic charge redistribution in the metal surface when polar molecules approach it (induced polarization).

Cross-References

- Applications of Nanofluidics
- Surface Energy and Chemical Potential at Nanoscale
- Wetting Transitions

References

1. van Hiemstra, T., Riemsdijk, W.H.: A surface structural approach to ion adsorption: the charge distribution (CD) model. J. Colloid Interface Sci. **179**, 488–508 (1996)
2. Israelachvili, J.: Intermolecular and Surface Forces. Academic, London (1991)
3. Cole, D.J., Csányi, G., Payne, M.C., Spearing, S.M., Colombi Ciacchi, L.: Development of a classical force field for the oxidised Si surface: application to hydrophilic wafer bonding. J. Chem. Phys. **127**, 204704 (2007)
4. Huang, D.M., Sendner, C., Horinek, D., Netz, R.R., Bocquet, L.: Water slippage versus contact angle: a quasiuniversal relationship. Phys. Rev. Lett. **101**, 226101 (2008)
5. Köppen, S., Langel, W.: Simulation of the interface of (100) rutile with aqueous ionic solution. Surf. Sci. **600**, 2040–2050 (2006)
6. Butt, H.-J., Graf, K., Kappl, M.: Physics and Chemistry of Interfaces. Wiley-VCH, Weinheim (2006)

7. Sendner, C., Horinek, D., Bocquet, L., Netz, R.R.: Interfacial water at hydrophobic and hydrophilic surfaces: slip, viscosity, and diffusion. Langmuir **25**, 10768–10781 (2009)
8. Cole, D.J., Payne, M.C., Colombi Ciacchi, L.: Water structuring and collagen adsorption at hydrophilic and hydrophobic silicon surfaces. Phys. Chem. Chem. Phys. **11**, 11395 (2009)
9. Monti, S., Walsh, T.R.: Free energy calculations of the adsorption of amino acid analogues at the aqueous titania interface. J. Phys. Chem. C **114**, 22197–22206 (2010)
10. Hoefling, M., Iori, F., Corni, S., Gottschalk, K.-E.: Interaction of amino acids with the Au(111) surface: adsorption free energies from molecular dynamics simulations. Langmuir **26**, 8347–8351 (2010)

Nano-scale Structuring

▶ Optical and Mechanical Nanostructured Coatings for Future Large-Scale Manufacturing

Nanoscale Technologies

▶ Use of Nanotechnology in Pregnancy

Nanoscale Thermal Analysis

▶ Nanocalorimetry

Nanoscale Water Phase Diagram

Michael Nosonovsky
Department of Mechanical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Definition

The phase diagram of water shows conditions (temperature and pressure) at which thermodynamically distinct phases can occur at equilibrium at the nanoscale.

Theoretical Background

Physical properties of many materials at the nanoscale are different from their properties at the macroscale due to the so-called scale effect. The causes of the scale effect include geometrical reasons (such as high surface-to-volume ratios at the nanoscale) and physical reasons (different physical mechanisms acting at the nanoscale compared to those at the macroscale).

Capillary Bridges and the Kelvin Equation

Besides high surface-to-volume ratios, an important feature of a small mechanical system is that capillary forces are often present. When contact of two solid bodies occurs in air, water tends to condense near the contact spots. This is because a certain amount of vapor is always present in air. Even under low relative humidity (*RH*), it is not possible to completely eliminate condensation in the form of water capillary bridges forming menisci near the contact spots, such as the tips of the asperities of rough solid surfaces. The menisci are usually concave and, therefore, the pressure inside them is reduced compared with the pressure outside in accordance with the Laplace theory. This leads to the attractive capillary force between the contacting bodies that is proportional to the foundation area of the meniscus. For small systems, this capillary force may become very significant, dominating over other forces such as the van der Waals adhesion or electrostatic forces. Therefore, understanding the properties of water in capillary bridges (menisci) is very important.

A flat water–air interface is in thermodynamic equilibrium with a certain amount of vapor in air, the partial pressure of which is referred to as "the saturated vapor pressure," P_{sat}, so that evaporation and condensation between the flat interface and vapor occurs at the same rate. If the partial pressure of vapor in air, P_V, is lower than P_{sat}, evaporation prevails over condensation, whereas in the opposite case ($P_V > P_{sat}$) condensation prevails over the evaporation. The ratio of the two values is the relative humidity, $RH = P_V/P_{sat}$. For a concave interface, the equilibrium occurs at $P_V/P_{sat} < 1$, whereas for a convex interface is occurs at $P_V/P_{sat} > 1$. The exact value of P_V/P_{sat}

for a meniscus of a given curvature is given by the Kelvin equation. The pressure drop of water with density ρ risen for the height h in a capillary tube is $\Delta P = \rho g h$. Using the Laplace equation and the hydrostatic formula for vapor pressure $P = P_0 \exp(-gh/RT)$ – where R is the gas constant, P_0 is the pressure at the surface ($h = 0$), and T is the temperature – yields the Kelvin equation:

$$\gamma_{LV}\left(\frac{1}{R_1}+\frac{1}{R_2}\right) = \rho RT \ln\frac{P_V}{P_{sat}} \qquad (1)$$

where R_1 and R_2 are the principle radii of curvature of the meniscus ($R_k = (1/R_1 + 1/R_2)^{-1}$ is referred to as the Kelvin radius). Equation 1 relates the interface curvature at the thermodynamic equilibrium with the ratio of actual and saturated vapor pressure, P_V/P_{sat} (relative humidity). According to the Kelvin theory, a concave meniscus with a negative curvature given by Eq. 3 may form at any relative humidity [1].

Water at the Nanoscale

Normally, the phase state of water is uniquely characterized by its pressure and temperature. The phase diagram of water shows whether water is in its solid, liquid, or vapor state at a given temperature and pressure. However, at the nanoscale, the situation may be different. Both molecular dynamics (MD) simulations and experimental studies with the atomic force microscope (AFM) and the surface force apparatus (SFA) show that the state of nanoscale volumes of water at a given pressure and temperature is not always the same as prescribed by the macroscale water phase diagram. In particular, confined small water volumes demonstrate quite unusual properties, such as the melting point depression. Several effects that can be detected by AFM are related to this unusual phase behavior of the nanoscale capillary bridges.

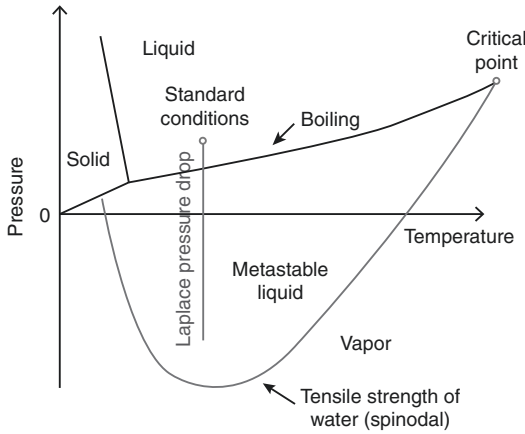
First is the metastability of small volumes of water. When pressure is below the liquid–vapor transition line of the water phase diagram, vapor is the most stable state. However, the liquid–vapor transition involves energy barriers, and thus a metastable liquid state is possible [1]. At the macroscale, random fluctuations are normally large enough to overcome the barriers. The metastable states are very fragile, and thus they are not normally observed, except for special circumstances (e.g., superheated water). However, at the nanoscale, the barriers are large compared with the scale of the system, and metastable states can exist for long intervals of time [1].

Second, it was argued that ice could form at room temperature inside the capillary bridges. The water in such a bridge can be neither liquid, nor solid, but form a liquid–ice condensate. This explains the slow rearrangement of the bridges between tungsten AFM tips and graphite surfaces reported in recent experiments [2]. The presence of ice-like structured water adsorbed at the silicon oxide surface was suggested to be responsible for the large RH dependence of the adhesion force in the single-asperity contact between silicon oxide surfaces in AFM experiments.

Third, pressure inside the capillary bridge is lower than the pressure outside. At the reduced pressure, the water boiling point is lower than 100 °C. There is experimental evidence that boiling in the capillary bridges indeed happens in concave water menisci. Investigating nanoscale water phase transitions using AFM data can lead to modifications of the conventional water phase diagram by introducing a scale dependence. Such a diagram should depend on the characteristic size of the system, in addition to the pressure and temperature dependencies [3].

Stretched Water

When a phase transition line in the phase diagram is crossed, it is normally expected that the substance would change its phase state. However, such a change would require additional energy input for nucleation of seeds of the new phase. For example, the liquid–vapor transition requires nucleation of vapor bubbles (this process is called cavitation), while the liquid–solid transition requires nucleation of ice crystals. If special measures are taken to prevent nucleation of the seeds of the new phase, it is possible to postpone the transition to the equilibrium state phase leading to supercooled, superheated, or stretched water (Fig. 1).



Nanoscale Water Phase Diagram, Fig. 1 Schematic of a phase diagram of water showing solid, liquid, vapor, and metastable liquid phases. While the energy barriers separating the metastable states are small, at the nanoscale these barriers are very significant

Two factors affect a vapor bubble inside liquid, needed as a nucleation center for vapor in superheated or stretched water: the pressure and the interfacial tension. While the negative pressure (tensile stress) is acting to expand the size of the bubble, the interfacial tension is acting to collapse it. For a small bubble, the interfacial stress dominates; however, for a large bubble, the pressure dominates. Therefore, there is a critical radius of the bubble, and a bubble with a radius larger than the critical radius would grow, whereas one with a smaller radius would collapse. The value of the critical radius is given by:

$$R_C = \frac{2\gamma_{LV}}{P_{sat} - P} \tag{2}$$

where γ_{LV} is the liquid–vapor surface tension, P_{sat} is the saturated vapor pressure, and P is the actual liquid pressure [1].

With the further decrease of pressure in the negative region, the so-called spinodal limit can be reached. At that pressure, the critical cavitation radius becomes of the same order as the thickness of the liquid–vapor interface. In this case, there is a lower energy path of nucleation connecting the liquid to the gas phase by expansion of a smoothly varying density profile. In the phase diagram, the line that corresponds to the spinodal limit is expected to go all the way to the critical point. This critical spinodal pressure at a given temperature effectively constitutes the tensile strength of liquid water. Various theoretical considerations, experimental observations, and MD simulation results have been used to determine the value of the spinodal limit. At room temperature, the spinodal pressure is between −150 and −250 MPa [1].

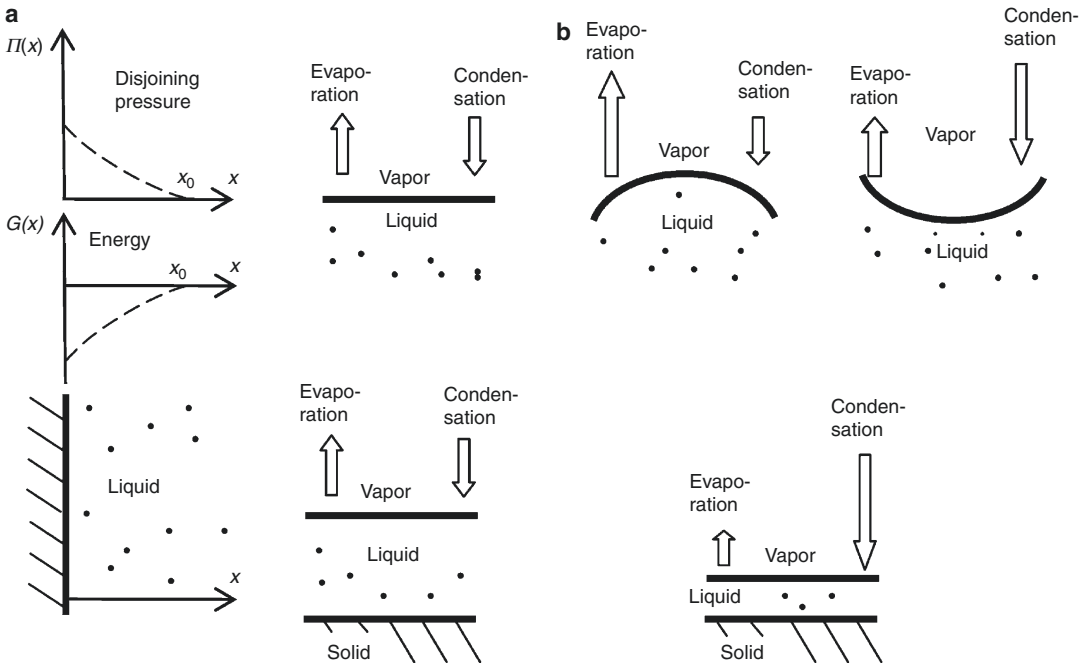
Disjoining Pressure

Another important effect is the so-called disjoining pressure. The adhesion force between the solid and water has a certain range x_0 and decreases with the distance x from the interface. As a result, water in the layer next to the interface is subjected to higher pressure $P = P_0 + \Pi_d(x)$ than the bulk pressure, where $\Pi_d(x)$ is the so-called disjoining pressure. For a thin liquid layer of thickness $H < x_0$, the evaporation/condensation equilibrium will be shifted toward condensation, since an additional attractive force acts upon water molecules from the solid surface. As a result, a thin water layer can be in equilibrium with undersaturated vapor, $P_V/P_{sat} < 1$, so the effect of the disjoining pressure on the thermodynamic equilibrium is similar to the effect of concave meniscus [4]. The influence of the disjoining pressure and the related wetting films in narrow confinement have to be considered as they may significantly change the meniscus curvature (Fig. 2). A typical effect of nanoscale confinement is the appearance of capillarity-induced negative pressure [3].

Key Findings

Negative Pressure in Water Capillary Bridges

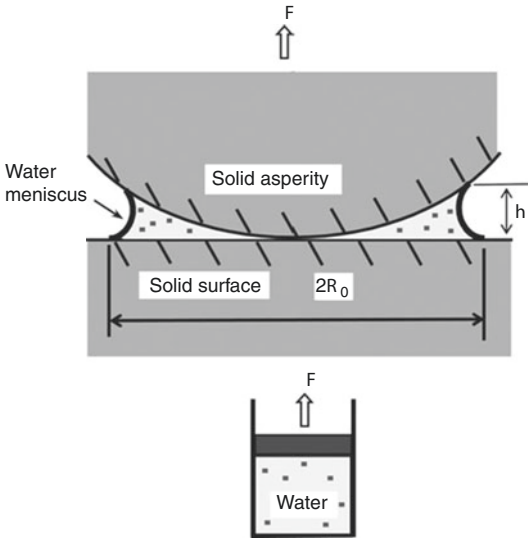
It is very difficult to obtain deeply negative pressure at the macroscale because of bubble nucleation. The pressure values that have been reported constitute −19 MPa with the isochoric cooling method, −17.6 MPa with the modified centrifugal method, and −25 MPa with the acoustic method. The situation is different at the micro- and nanoscale. The pressure of −140 MPa in the microscopic aqueous inclusions in quartz crystals



Nanoscale Water Phase Diagram, Fig. 2 (**a**) Disjoining pressure near a solid surface (**b**) The effect of meniscus curvature (*flat*, *concave*, and *convex*) and liquid layer thickness on the evaporation–condensation equilibrium

was reported. Water plugs in hydrophilic nanochannels can be at a significant absolute negative pressure due to tensile capillary forces [1].

Negative pressure of −160 MPa was reported in nanoscale liquid capillary bridges. The capillary component of the adhesion force was determined by the comparison of the adhesion force in air (at different *RH* levels) and in ultra-high vacuum (UHV), where capillary bridges are absent. Very small values of the Atomic force microscope (AFM) tip radius (on the order of 10 nm) correspond to a small size of meniscus (meniscus foundation area on the order of 10^{-16} m^2) and thus even a small force of 10 nN corresponds to huge pressure differences inside and outside the meniscus on the order of 100 MPa or 1,000 atm (Fig. 3). The reason for the existence of such large tension limits lies in the strength of the hydrogen bonds which must be broken in order to form a cavity that is large enough to act as a stable nucleus for cavitation. The metastable state of the capillary bridges is observed because the size of the bridges



Nanoscale Water Phase Diagram, Fig. 3 A water capillary bridge between an AFM tip and a flat sample. The Laplace pressure inside the bridge is reduced because the bridge is concave. The bridge is under tension, similarly to water in a sealed tube with a piston with the applied tensile force *F*

is smaller than the critical cavitation radius given by Eq. 2 and, therefore, no cavitation occurs [1].

Room Temperature Ice in Capillary Bridges

Most AFM results show that the capillary force has a strong dependence on *RH*: first it increases with increasing relative humidity up to approximately 30 % and then decreases with a further increase of the relative humidity. For low humidity, the pressure difference inside and outside the bridge is large; however, the size of the bridge is small due to high curvature of the meniscus. For very low humidity, the size of the bridge may be insufficient to generate a capillary force. On the other hand, for high humidity, the pressure difference is small; however, the meniscus is large (but, of course, its size is limited by the radius of the tip). Thus, it is expected that in the limiting cases of $RH = 0$ % and $RH = 100$ % the capillary force would vanish, since at $RH = 0$ % there is no capillary bridge and at $RH = 100$ % the Laplace pressure drop is zero. The capillary force achieves its maximum at a certain value of *RH* [3].

It has been argued that the capillary force alone cannot explain the observed magnitude of the humidity dependence. The origin of the large *RH* dependence is due to the presence of ice-like structured water adsorbed at the silicon oxide surface at room temperature. Thus, the force is due to the rupture of an ice–ice bridge at the center of the contact region with the structure, thickness, and viscoelastic behavior of the adsorbed water layer influencing the adhesion force [2].

Water often forms adsorbed layers at the surface of many materials, and these layers may have an ice-like structure and be connected to the capillary bridge, as indicated by the study of the molecular configuration of adsorbed water as a function of relative humidity using attenuated total reflection (ATR) infrared spectroscopy [4]. A completely hydrogen-bonded ice-like network of water grows up to three layers as the relative humidity increases from 0 % to 30 %. In the relative humidity range of 30–60 %, the liquid water structure starts appearing, while the ice-like structure continues growing to saturation. Above 60 % relative humidity, the liquid water configuration grows on top of the ice-like layer.

During the contact between a tungsten tip and a graphite surface it was found that water in the nanoscale capillary bridges may include ice or be a mixture of liquid water with ice [2]. Such a mixture would have high viscosity and may act like glue, demonstrating elastic response.

Low-Temperature Water Boiling

Another important consequence of the water pressure reduction in the capillary bridges is the possibility of boiling at temperatures lower than 100 °C. If a capillary bridge is large enough, so that cavitation occurs, the pressure inside the bridge is reduced compared with the ambient, and the boiling point is depressed by the amount ΔT. The Clausius–Clapeyron equation provides the change of pressure $\Delta P = P_2 - P_1$ as related to the change of the boiling temperature $\Delta T = T_2 - T_1$ as

$$\ln(P_2/P_1) = \Delta h/R(1/T_1 - 1/T_2) \qquad (3)$$

where Δh is the change of specific enthalpy [54]. Note that ΔP is given by Eq. 1.

For example, for a meniscus between an AFM cantilever and a flat surface with a gap between them of 3 μm and the principal radii of curvature of the meniscus $R_1 = R_2 = -2$ μm, the Kelvin radius is $R_k = (1/R_1 + 1/R_2)^{-1} = -1$ μm and the pressure drop, based on (6) is $\Delta P = 72.9$ kPa (or 0.72 atm). Thus, the pressure inside the meniscus is $P = 28.1$ kPa (or 0.28 atm), which corresponds to the boiling point of about 333 K [3].

Nanobubbles

There is no adequate theoretical explanation why ice can be present in the capillary bridges at room temperature; however, it is clear that this effect is related to the adsorption of water to solid surfaces. The ability of water to form thin layers on hydrophobic surfaces is associated with the disjoining pressure caused by the spatial dependence of the van der Waals adhesion force. Due to the dependence of the adhesion force of the distance (on the scale of nanometers), the energy of the water

molecules in the vicinity of the solid surface is lower than the energy of those far away. Thus, it is often not only energetically profitable for water to cover the solid, but also to create a thin layer.

It is noted that not only the solid phase can be adsorbed at a surface. Very thin (5–80 nm) gas films could exist for a long time at the interface between a hydrophobic solid and water. Despite certain thermodynamic objections for nanobubble existence (such as expected high pressures in bubbles leading to their dissolution), some experimental studies with the AFM and other equipment have demonstrated the existence of nanobubbles at the solid/water interface [5, 6]. Nanobubbles can play a significant role in a number of processes, such as the long-range hydrophobic attractive force.

Water in Confinement

A number of studies are devoted to water phase behavior in confinement. MD simulations show that water confined to carbon nanotubes of a critical size under ambient conditions (1 bar, 300 K) can undergo a transition into a state having ice-like mobility with an amount of hydrogen bonding similar to that in liquid water. Maeda [7] investigated experimentally phase transitions of capillary-held liquids in a slit-like pore. When the surfaces were placed at a separation very close to the Kelvin length, it was possible to detect a stage in which the system was in an apparent kinetic equilibrium between two physical states – with and without the liquid connecting the two surfaces. Giovambattista et al. [8] used MD simulation to investigate the phase behavior of water confined between two nanoscale hydrophobic or hydrophilic plates at room temperature, varying the pressure ($-0.15\,\text{GPa} \leq P \leq 0.2\,\text{GPa}$) and plate separation ($0.4\,\text{nm} \leq d \leq 1.6\,\text{nm}$). With hydrophobic plates, capillary evaporation occurred between the plates at low P. The threshold value of d at which this transition occurs decreases with P (e.g., 1.6 nm at $P = 0.05$ GPa, 0.5 nm at $P = 0.1$ GPa), until, at high P, no capillary evaporation occurred. For $d = 0.6$ nm and $P > 0.1$ GPa, the system crystallizes into a bilayer ice. They developed a P–d phase diagram showing the vapor, liquid, and bilayer ice. With hydrophilic plates, water always remained liquid even at the P at which bulk water cavitates. This indicates that size is important for a phase diagram due to the finite range of the adhesion force in addition to the critical cavitation radius and curvature of the meniscus, which were discussed in the preceding sections.

Cross-References

- ▶ AFM in Liquids
- ▶ Disjoining Pressure and Capillary Adhesion
- ▶ Wetting Transitions

References

1. Yang, S.-H., Nosonovsky, M., Zhang, H., Chung, K.-H.: Negative pressure in water capillary bridges at nanocontacts. Chem. Phys. Lett. **451**, 88–92 (2008)
2. Jinesh, K.B., Frenken, J.W.M.: Capillary condensation in atomic scale friction: how water acts like a glue. Phys. Rev. Lett. **96**, 166103 (2006)
3. Nosonovsky, M., Bhushan, B.: Phase behavior of capillary bridges: towards nanoscale water phase diagram. Phys. Chem. Chem. Phys. **10**, 2137–2144 (2008)
4. Asay, D.B., Kim, S.H.: Effects of adsorbed water layer structure on adhesion force of silicon oxide nanoasperity contact with humid ambient. J. Chem. Phys. **124**, 174712 (2006)
5. Steitz, R., Gutberlet Hauss, T., Klosgen, B., Krastev, R., Schemmel, S., Simonsen, A.C., Findenegg, G.H.: Nanobubbles and their precursor layer at the interface of water against a hydrophobic substrate. Langmuir **19**, 2409–2418 (2003)
6. Mashl, R.J., Joseph, S., Aluru, N.R., Jakobsson, E.: Anomalously immobilized water: a new water phase induced by confinement in nanotubes. Nano Lett. **3**, 589–592 (2003)
7. Maeda, N.: Phase transitions of capillary-held liquids in a slit-like pore. J. Phys. Chem. B **110**, 25982–25993 (2006)
8. Giovambattista, N., Rossky, P.J., Debenedetti, P.G.: Effect of pressure on the phase behavior and structure of water confined between nanoscale hydrophobic and hydrophilic plates. Phys. Rev. E **73**, 1–14 (2006)

Nanoscratch Tester

- ▶ Nanoindentation

Nano-sized

▶ Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

Nanosized Nanocrystalline and Nanotwinned Metals

Dongchan Jang
Department of Applied Physics and Materials Science, 230 Steele Laboratory, California Institute of Technology, Pasadena, CA, USA

Synonyms

Electroplating; Nanocrystalline materials; Nano-pillars; Nano-twinned materials

Definition

- Nano-pillars are pillar-shaped specimens at the nanoscale, whose aspect ratio (height/diameter) is generally within $3 \sim 7$.
- Nanocrystalline material refers to a class of polycrystalline materials whose grain size is smaller than $\sim$100 nm.
- Nano-twin refers to a class of crystalline materials that contain a number of parallel coherent twin boundaries with the inter-boundary distance at the nanoscale.

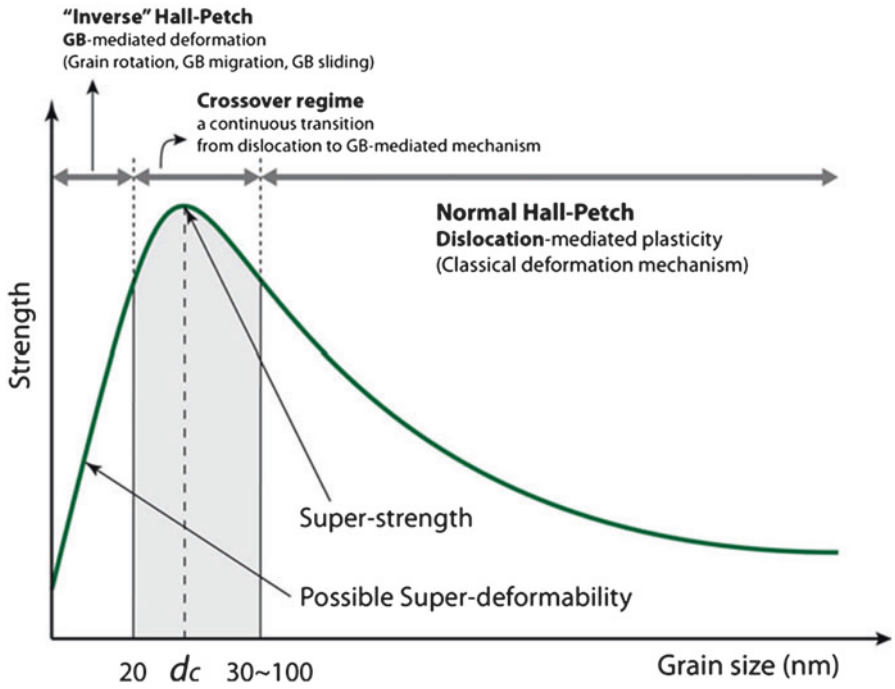
Overview

Traditionally, the most fundamental principle in materials science is that the properties of materials are strongly influenced by their internal microstructure [1]. During the plastic deformation of many metallic systems, dislocations carry most of the plasticity and hence their interactions with the internal microstructure, such as grain boundaries, solute atoms, precipitation or even other dislocations, are the most influential in determining mechanical responses of those materials [2]. However, recent studies found that the microstructure is not the only factor that affects the mechanical properties, but the external dimension of the specimen also significantly changes the mechanical behavior when it is reduced to the submicron and nanometer scale [3]. Therefore, when the effects from those two factors, microstructure and specimen size at the nanoscale, are combined, the materials often exhibit very unique properties distinguished from both of the bulk samples containing same kind of microstructure and nano-samples not containing the corresponding microstructure [3]. Especially, the homogeneous internal interfaces, such as ordinary grain boundary (GB) and coherent twin boundary (TB), are one of the most interesting microstructures as their interaction to dislocations is a key parameter in understanding plasticity of metallic materials. In this entry, the review of the recent studies investigating the mechanical properties of nano-sized metallic samples containing ordinary GBs and coherent TBs is presented. The main focus is placed on the results from uniaxial tension/compression experiments on the nanometer-sized pillar-shaped specimen (nano-pillars). Fabrication techniques to produce the nano-pillars with various internal interfaces are also introduced.

Mechanical Properties

Nanocrystalline Nano-pillars

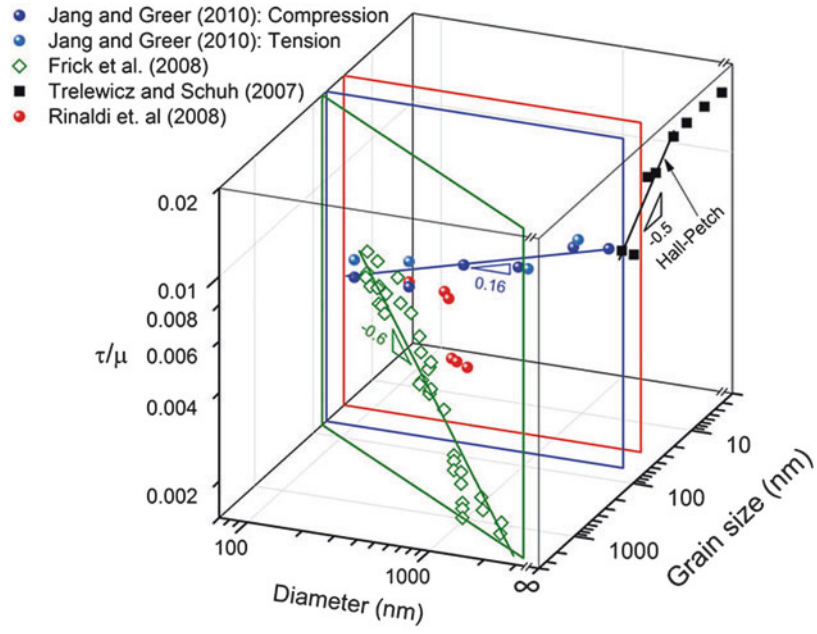
GB is the interface between differently oriented grains in the polycrystalline materials. Generally, there is atomic mismatch across the GB as the crystallographic orientation of one grain abruptly changes to that of the adjacent one [1]. In conventional bulk polycrystalline materials, GBs usually act as a barrier to the dislocation motion across the boundary because of the discontinuity of atomic planes at the GB [1]. As a result, the strength of the bulk polycrystalline materials is inversely correlated with the grain size because total GB area increases with decreasing grain size, so dislocations are impeded more often with smaller grain size. For many conventional polycrystalline



Nanosized Nanocrystalline and Nanotwinned Metals, Fig. 1 Strength of polycrystalline materials as a function of grain size: Hall–Petch relation and transition to inverse Hall–Petch (Reprint with permission from [3])

metals, the strength is inversely proportional to the square root of grain size, i.e., $\sigma_y = \sigma_o + k_y d^{-1/2}$, where σ_y and d are the yield strength and grain size, respectively, and σ_o and k_y are constants for a particular material. This expression is called Hall–Petch relation [1, 4, 5]. However, this Hall–Petch relation cannot be extrapolated down to infinitesimally small grain sizes. Instead, it is reported to break down when grain size is reduced below a critical value (usually $10 \sim 20$ nm). Below this critical size, the strength is often saturated or decreases with further reduction of grain size. This reversed relation between the strength and grain size is called inverse Hall–Petch equation. The deviation from the Hall–Petch relation is likely due to the transition in plastic deformation mechanism from dislocation-driven in Hall–Petch to GB-mediated in inverse Hall–Petch. The dependence of strength on the grain size and the corresponding deformation mechanism change are schematically summarized in Fig. 1. Further details of bulk nanocrystalline materials are available in the entry "▶ Mechanical Properties of Nanocrystalline Metals."

As briefly stated in the Overview section, the specimen with external dimensions at the nanoscale shows different mechanical responses from those of their bulk counterpart. For example, the strengths of single-crystalline nano-pillars drastically increase when their diameters decrease below submicron scale, showing the well-known "Smaller is stronger" trend (See the entry "▶ Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures" for further details). In nanocrystalline nano-pillars, where both of the external specimen and internal grain sizes are in the nanometer range, the mechanical properties are clearly distinguishable from those of bulk nanocrystalline materials and single-crystalline nano-pillars. Figure 2 shows the 3D plot, where the strengths for nearly the same materials, Ni, are plotted as functions of nano-pillar diameter (D) and grain size (d). Quantities τ and μ in this plot are the shear strength and shear modulus, respectively. When $D \gg d$, the sample dimension approaches



Nanosized Nanocrystalline and Nanotwinned Metals, Fig. 2 3D representation of strength of Ni as functions of nano-pillar diameter (D) and grain size (d). The quantities τ and μ are the shear strength and shear modulus, respectively (Reprint with permission from [3])

to the bulk poly-/nanocrystalline materials, and the strength becomes independent of the sample size, D. Therefore, the strength is the function of grain size, d, only and follows the ordinary Hall–Petch and inverse Hall–Petch relations of the bulk nanocrystalline Ni. In Fig. 2, the data points (solid squares) on the vertical plane perpendicular to the D axis at $D \sim \infty$ show these relations. When $D = d$, there is only one grain across the diameter, and the resultant microstructure is single crystal (another possibility is bamboo structure, where the GBs are orthogonal to the pillar axis, but this structure is not discussed in this entry). The single-crystalline nano-pillars display "Smaller is stronger" trend, where the strength increases as decreasing the pillar diameter, D. In Fig. 2, the green plane satisfies the condition of $D = d$, and the data points on that plane (open diamonds) represent the strength of single-crystalline Ni nano-pillars as a function of pillar diameter, D (or equivalently as a function of d). The linear regression line in this log-log scale plot shows the power-law slope of ~0.6. Further details on the single-crystalline nano-pillars are available in Refs. [3, 6]. While both of bulk nanocrystalline Ni and single-crystalline Ni nano-pillars exhibit "Smaller is stronger" trend, where the strength increases as decreasing either d or D (in bulk nanocrystalline samples, this statement is valid only down to the critical grain size of ~20 nm), the strength of nanocrystalline nano-pillars shows different aspects from both of them. In Fig. 2, the vertical plane perpendicular to the d axis at $d = 60$ nm displays the strength of nanocrystalline Ni nano-pillars with fixed grain size of 60 nm as a function of D ranged from 1,600 to 100 nm. Unlike bulk nanocrystalline Ni or single-crystalline Ni nano-pillars, this nanocrystalline Ni nano-pillars show "Smaller is weaker" trend, where the strength decreases with decreasing D. Possible explanation to this trend can be due to the contribution of the free surface, whose relative fraction increases with decreasing the pillar diameter. In general, the strength of an individual grain embedded within the surrounding ones (i.e., a grain in bulk polycrystalline materials) is higher than that of the same but completely freed grain. It is because for plastic deformation in bulk polycrystalline materials, all grains must deform cooperatively by matching the displacement across the GB, and it applies additional constraints to the plastic deformation. However, when the adjacent grains are removed, those constraints are released, and the grain can deform

more easily. In nanocrystalline nano-pillars, the grains located at the free surface (surface grains) have less constraint than those in the pillar (inner grains) since the surface grains face the empty space at least to one direction. Therefore, the strength of the surfaces grains is lower than that of the inner grains, dividing the whole nano-pillars into hard inner core and soft outer shell. The thickness of the outer shell is approximately same with the average grain size of the nano-pillar. Therefore, when they have a fixed average grain size, the relative fraction of the soft outer shell increases as the diameter of the nanocrystalline nano-pillars decreases. As a result, the strength of the nanocrystalline nano-pillar lowers [3].

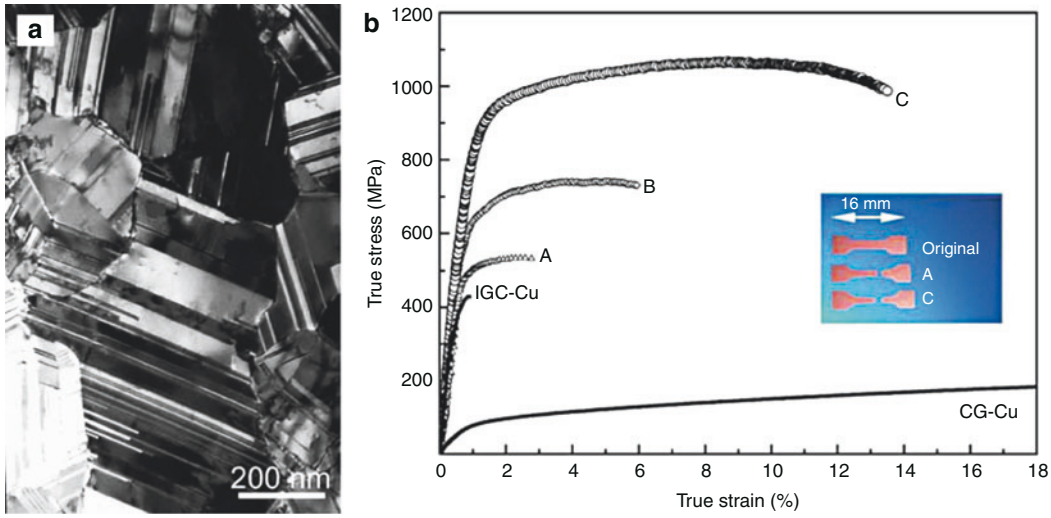
Another important observation on the nanocrystalline nano-pillars is the onset of GB-mediated plasticity at larger grain size than that of bulk nanocrystalline samples. In bulk nanocrystalline metals, activation of GB-mediated plasticity is usually correlated with transition from the Hall–Petch to inverse Hall–Petch relation, and this transition grain size for Ni is around 20 nm, as can be seen in Fig. 2. However, recent study reveals that reduction of the nano-pillar diameter to ~100 nm facilitates the activation of GB-mediated process, such as grain boundary sliding and grain rotation, at 60-nm grain size, which is significantly larger than the transition size of bulk samples, 20 nm [3].

Nano-twinned Nano-pillars

In general, the misorientation angles between the crystallographic planes across the ordinary GBs described in the previous section are randomly distributed, and atomic structure of those GBs is incoherent in that the lattice planes lose the crystallographic registry across the interface. However, there is a different type of internal coherent and homogeneous interface, called twin boundary (TB). Twinned microstructure usually refers to a special type of relative adjacent grain orientations, where one of the grains has the mirror-symmetric crystallographic structure of the other, and the mirror plane is called a twin boundary [1]. In this structure, one of the grains is called matrix and the other twin. Particularly, when the lattice planes do not lose the crystallographic registry across the boundary, the TB is called coherent. Because the TBs are also a special type of the GB, their strengthening mechanism shares many common factors with the ordinary GBs, i.e., TBs also interfere with the glide of dislocations. However, their influence on strengthening is known to be less pronounced because they are less effective at resisting the penetration of the moving dislocations, especially when the distance between the neighboring TBs is in micron range [7].

Over the last decade, the so-called nano-twinned materials (usually metals) have attracted a lot of scientific and industrial interest as promising candidates for the next generation structural and/or electrical material because of their simultaneous attainment of high strength, ductility, and electrical conductivity. In these nano-twinned metals, the nano-twinned area consists of alternative pileup of nanometer-thick planar matrix and twin grains (nano-twin lamellas). The transmission electron microscope (TEM) image of a typical nano-twinned Cu and the corresponding stress–strain curve is presented in Fig. 3 [8].

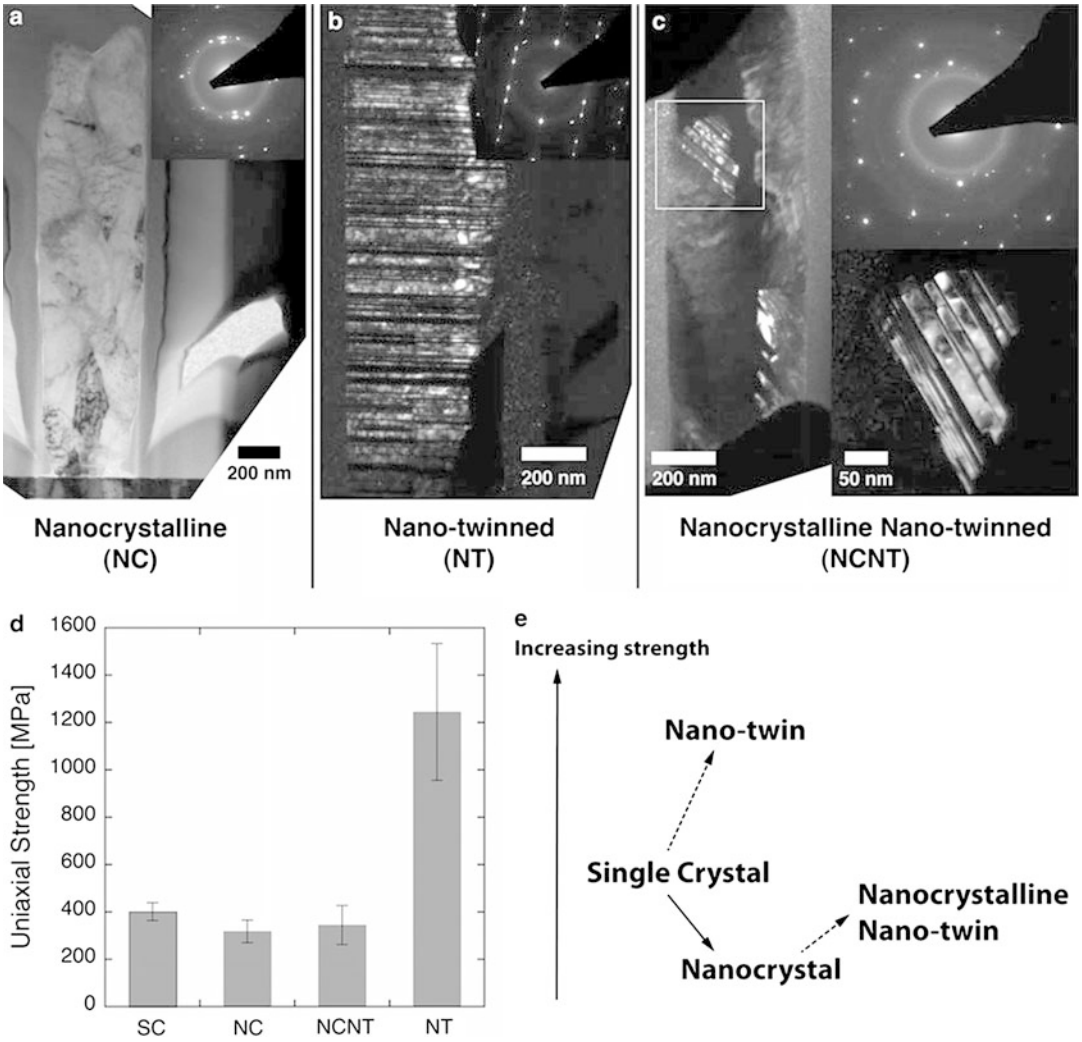
An important difference between ordinary GB and coherent TB in their roles during plastic deformation is that while the dislocation can hardly move along or across the ordinary GBs because of their disordered or discontinued atomic structure, the coherent TBs relatively easily allow dislocations to glide along and across the boundary [7]. The consequence of this difference is that dislocation mobility is highly limited in the presence of GBs, resulting in increased strength with sacrificing the ductility in regular nanocrystalline metals. On the other hand, because of its coherent nature, the TBs are capable of accommodating more various interactions with moving dislocations other than simply blocking their glide. For example, a moving dislocation can be absorbed to the TB, leaving a sessile dislocation on the interface plane, or can be dissociated into two gliding partial dislocations on the TB plane, leaving the stacking fault, or can transmit across the TB, leaving a step at the TB [7]. All of these interactions gradually degrade coherency of the TBs, hinder subsequent dislocations in interacting in the similar manners, and hence change the



Nanosized Nanocrystalline and Nanotwinned Metals, Fig. 3 (**a**) TEM observations of typical microstructure in nano-twinned Cu sample. (**b**) Typical tensile stress–strain curves for the as-deposited Cu samples with average nano-twin thickness of 93, 30, and 15 nm for samples A, B, and C, respectively. For comparison, the stress–strain curves for coarse-grained polycrystalline Cu sample (CG-Cu) and nanocrystalline Cu (IGC-Cu) are also displayed (Reprint with permission from [8])

characteristic of coherent TB close to that of ordinary incoherent GBs. As a result, it gradually requires higher applied stress to keep the dislocations moving, and the material is work-hardened [7]. Usually, the increase of work-hardening coefficient correlates with higher resistance to instability, such as necking [4, 9], so it can contribute to the effective increase of ductility.

Most of nano-twinned samples fabricated to date are polycrystalline in the sense that the nano-twin lamellas are embedded within polycrystalline grains as sub-grain structures as can be seen in Fig. 3. Properties measured in the polycrystalline nano-twinned samples are convoluted by the effects from many different microstructures like GBs, and differently orientated TBs, etc. Therefore, the specific role of TBs and nano-twinned lamellae in the enhanced mechanical properties is far from being understood. However, recent studies on the nano-twinned nano-pillars, where the specimen contains well-aligned TBs only, offered an insight into understanding the specific role of TB during plastic deformation. The TEM images in Fig. 4a–c show the Cu nano-pillars with 500-nm diameters, whose microstructures are nanocrystalline (NC), purely nano-twinned (NT), and nanocrystalline nano-twinned (NCNT), respectively [10]. NC nano-pillars only contain the randomly distributed ordinary GB with average grain size of ~160 nm, while NT nano-pillars have only coherent TBs orthogonal to the pillar axis. Average twin thickness of NT nano-pillars is ~8.8 nm. NCNT nano-pillars contain both of GBs and TBs, where nano-twin lamellae are embedded within the regular grains. Average grain size and twin thickness of NCNT nano-pillars are 250 and 6.7 nm, respectively. The strengths of those Cu nano-pillars measured by uniaxial nano-compression experiments are shown in Fig. 4d, where SC stands for single-crystalline Cu nano-pillars. These results suggest that in Cu nano-pillars with 500-nm diameter, the two different interfaces, ordinary GB and coherent TB, appear to affect the strengths differently. Given that strength of NC is lower than that of SC by ~20 %, the ordinary GBs seem to lower the strength of the nano-pillars. On the other hand, strength of NT is higher than that of SC by a factor of ~3, implying that orthogonally aligned TBs drastically increase the strength of nano-pillars. Comparing the strengths of NC and NCNT, TBs also increase the strength even in the



Nanosized Nanocrystalline and Nanotwinned Metals, Fig. 4 (**a**) TEM bright-field image and diffraction pattern (*inset*) of a typical nanocrystalline (*NC*) Cu nano-pillar. TEM dark-field images and diffraction patterns (*upper-right insets*) of typical (**b**) nano-twinned (*NT*) and (**c**) nanocrystalline nano-twinned (*NCNT*) Cu nano-pillars. The *lower-right inset* in (**c**) is the zoomed-up image of the region within the *white box*. (**d**) The average yield strength of each microstructure. (**e**) Illustration showing the effect of grain and nano-twin boundaries on the strength of 500-nm Cu nano-pillars (Reprint with permission from [10])

polycrystalline case. The influence of GBs and TBs on the strength of 500-nm Cu nano-pillars is summarized in the diagram in Fig. 4.

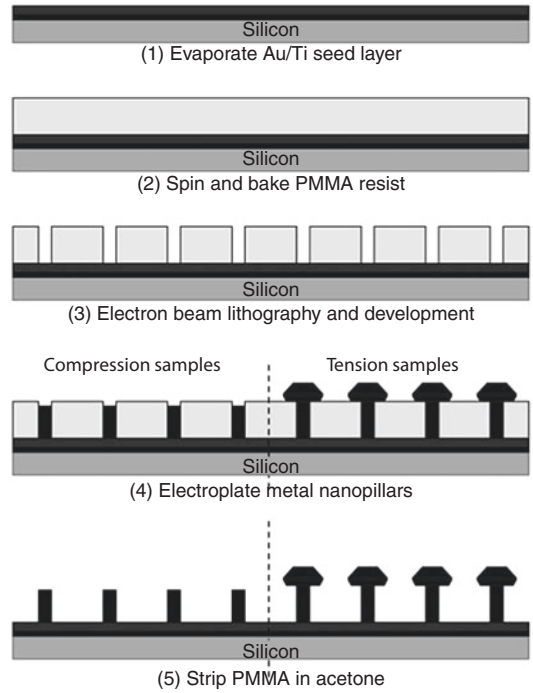
Fabrication of Nano-pillars via Electroplating

Traditionally, electroplating is a widely used metallic surface coating technique on a conductive surface. During electroplating, the metallic ions in the electrolyte migrate to the surface of the cathode under externally applied electrical voltage between the cathode and anode, and deposit a thin metallic layer on the cathode. Basically, the fabrication of nano-pillars shares the same physics and chemistry with this traditional electroplating technique. However, rather than forming a metallic layer of uniform thickness, it requires the ions to deposit only onto

prescribed locations. This can be achieved by using the pre-patterned template as the cathode. The template consists of the chemically etchable (e.g., polymethylmethacrylate (PMMA)) patterned nonconductive sacrificial layer on top of a conductive metallic thin film deposited on an insulating substrate (Si, SiO_2 or sapphire). A pattern of desired shapes can be written via e-beam lithography, and the resist is subsequently developed to generate arrays of through-holes to the underlying conductive metallic film. The underlying thin metallic layer provides an electric-conducting path for the electroplating, and the ions deposit onto the metallic layer through the open spaces in the patterned template. After electroplating, the PMMA sacrificial layer can be stripped off in acetone. Therefore, the resulting electroplated structure becomes a negative of the pattern. The nonconductivity of the substrate prevents the ions from depositing to the backside of the template. The schematic illustration of this process is presented in Fig. 5.

The specific details of the template fabrication are available in [11]. For example, a 100-nm thick Au seed layer on top of the 20-nm thick Ti adhesion layer is deposited on the Si wafer by electron beam evaporation. The Au layer provides the electric-conducting path for the electroplating and Ti layer enhances the adhesion of the Au layer to the Si substrate. The PMMA resist diluted in anisole is spin-coated on top of the Au seed layer. The thickness of the PMMA resist determines the maximum height of the nano-pillars, as illustrated in Fig. 5. Table 1 shows the list of some examples of the resist parameters and corresponding layer thicknesses [11].

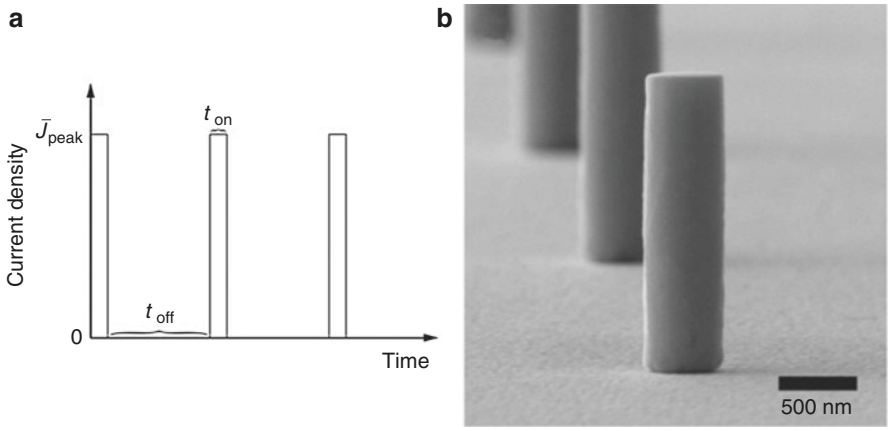
The fabricated template can be used as the part of the cathode during the electroplating. The cathode is split into two parts: the template and a Cu-coated dummy chip. The purpose of the dummy chip is to precisely control the current density at the cathode ($\overline{J}$), which is defined as the total current divided by the cross-sectional area of the cathode. With the template alone, the cross-sectional area, which is the sum of the area of the patterns on the template, is extremely small (on the order of nm^2 or μm^2) so that the current density can sharply change by a small error in the



Nanosized Nanocrystalline and Nanotwinned Metals, Fig. 5 Schematic representation of template fabrication and electroplating (Reprint with permission from [11])

Nanosized Nanocrystalline and Nanotwinned Metals, Table 1 Polymethylmethacrylate (PMMA) resist and electron beam exposure parameters for template fabrication [11]. The quantity d is the diameter of the circular patterns, which is effectively the same as the diameter of the electroplated nano-pillars, and L is the thickness of PMMA (Reprint with permission from [11])

Pillar dimensions		Resist parameter		Exposure parameters
d (nm)	L (nm)	Dilution (%)	Speed (rpm)	Dose ($\mu C/cm^2$)
25	100	4	2,500	2,600
50	200	4	2,500	1,600
100	400	7	3,000	1,450
150	600	7	1,500	1,600
200	800	9	4,500	1,550
250	1,000	9	2,000	1,500
500	2,000	11	2,500	1,600
750	3,000	11	1,500	1,750



Nanosized Nanocrystalline and Nanotwinned Metals, Fig. 6 (**a**) The waveform of the pulsed electroplating current. (**b**) SEM image of a typical as-fabricated Cu nano-pillar (Reprint with permission from [10])

Nanosized Nanocrystalline and Nanotwinned Metals, Table 2 The electroplating condition for the fabrication of 500-nm Cu nano-pillars for each microstructure (Reprint with permission from [10])

Microstructure	Electrolyte composition	J_{on} (A/cm^2)	t_{on} (ms)	t_{off}(ms)	Feature sizes	
					Grain size (nm)	Twin thickness (nm)
Nano-twinned (NT)	125 g/L of $CuSO_4{}^{\circ}5H_2O$ + 50 g/L of H_2SO_4	0.4	1	100	N/A	8.8
Nanocrystalline (NC)	100 g/L of $C_6H_8O_7{\cdot}H_2O$ + 28 g/L of $CuSO_4{}^{\circ}5H_2O$ + 50 g/L of $(NH_4)_2SO_4$	1.2	1	100	160	N/A
Nanocrystalline Nano-twinned (NCNT)	50 g/L of $C_6H_8O_7{\cdot}H_2O$ + 28 g/L of $CuSO_4{}^{\circ}5H_2O$ + 50 g/L of $(NH_4)_2SO_4$	0.72	1	100	250	6.7

area measurement. By adding the dummy chip whose cross-sectional area is a few orders of magnitude larger than the patterns, the total cross-sectional area of the anode becomes insensitive to the small error in the area measurement, so that the current density at the cathode can be reliably controlled. As the anode material, either insoluble platinized titanium mesh or soluble metallic plate can be used, depending on the material to be plated. Microstructure of the electroplated nano-pillars is highly dependent on the plating condition, including composition of electrolyte, waveform of the applied current, and current density at the cathode. For example, in order to fabricate the Cu nano-pillars, the platinized titanium mesh was used as the anode material, and rectangular pulsed current was applied between the cathode and anode, as shown in Fig. 6a, where the current density is maintained at $\bar{J}_{peak}$ during on time, t_{on} and becomes zero during off time, t_{off} [10]. Electrolyte composition is also an important factor to control the microstructure of the Cu nano-pillars. In Table 2, the electroplating parameters to create the NT, NC, and NCNT Cu nano-pillars are summarized, including the electrolyte composition, $\bar{J}_{peak}$, t_{on}, and t_{off} [10]. The Scanning Electron Microscope (SEM) image of a typical as-fabricated pillar is shown in Fig. 6b.

Fundamentally, all conventional electroplatable metals can be also fabricated into the nano-pillar forms utilizing the method in the previous paragraph. However, determining electroplating parameters for specific microstructure,

especially the waveform of the applied current, can be challenging. While specific details have to be developed for each material, the general guideline is that (1) the grain size inversely scales with the plating overpotential, and hence reducing forward current density often yields the larger grain size. (2) Sometimes, using pulsed current is efficient in producing larger grain size as it periodically reduces the depletion region near the cathode, and consequently lowers the overpotential. The recipes to create single crystalline Cu [11], polycrystalline Au [11], and polycrystalline In [12] nano-pillars are available elsewhere.

Cross-References

- ▶ Mechanical Properties of Nanocrystalline Metals
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Plasticity Theory at Small Scales
- ▶ Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures

References

1. Callister, W.D.: Materials Science and Engineering: An Introduction. Wiley, New York (1997)
2. Hirth, J.P., Lothe, J.: Theory of Dislocations. Wiley, New York (1982)
3. Greer, J.R., De Hosson, J.T.M.: Plasticity in small-sized metallic systems: intrinsic versus extrinsic size effect. Prog. Mater. Sci. **56**, 654–724 (2011)
4. Courtney, T.H.: Mechanical Behavior of Materials. McGraw-Hill, Boston (2000)
5. Meyers, M.A., Mishin, Y., Benson, D.J.: Mechanical properties of nanocrystalline materials. Prog. Mater. Sci. **51**, 427–556 (2006)
6. Zhu, T., Li, J.: Ultra-strength materials. Prog. Mater. Sci. **55**, 710–757 (2010)
7. Lu, K., Lu, L., Suresh, S.: Strengthening materials by engineering coherent internal boundaries at the nano-scale. Science **324**, 349–352 (2009)
8. Shen, Y.F., Lu, L., Lu, Q.H., Jin, Z.H., Lu, K.: Tensile properties of copper with nano-scale twins. Scr. Mater. **52**, 989–994 (2005)
9. Dieter, G.E.: Mechanical Metallurgy. McGraw-Hill, New York (1988)
10. Jang, D., Cai, C., Greer, J.R.: Influence of homogeneous interfaces on the strength of 500 nm diameter Cu nanopillars. Nano Lett. **11**, 1743–1746 (2011)
11. Burek, M.J., Greer, J.R.: Fabrication and microstructure control of nanoscale mechanical testing specimens via electron beam lithography and electroplating. Nano Lett. **10**, 69–76 (2010)
12. Lee, G., Kim, J., Budiman, A.S., Tamura, N., Kunz, M., Chen, K., Burek, M.J., Greer, J.R., Tsui, T.Y.: Fabrication, structure and mechanical properties of indium nanopillars. Acta Mater. **58**, 1361–1368 (2010)

Nano-sized Particle

▶ Gas-Phase Nanoparticle Formation

Nanoslits

▶ Nanochannels for Nanofluidics: Fabrication Aspects

Nanosoldering

▶ Nanorobotic Spot Welding

Nanospring

▶ Nanorobotics for NEMS Using Helical Nanostructures

Nanostructure

▶ Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm
▶ Structure and Stability of Protein Materials

Nanostructure Field Effect Transistor Biosensors

Jason Li[1], Steve To[2], Lidan You[4] and Yu Sun[3]
[1]Advanced Pharmaceutics and Drug Delivery Laboratory, Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada
[2]Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada
[3]Department of Mechanical and Industrial Engineering and Institute of Biomaterials and Biomedical Engineering and Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada
[4]Department of Mechanical and Industrial Engineering and Institute of Biomaterials and Biomedical Engineering, University of Toronto, Toronto, ON, Canada

Synonyms

Bio-FET; Chem-FET; CNT biosensor; CNT-FET; DNA FET; Nano-FET; Nanowire FET biosensor

Definition

Nanostructure field effect transistor (nano-FET) biosensors are nanoscale electronic devices used to detect and measure the presence and concentration of specific biological molecules within a given sample solution. These sensors utilize a quasi-one-dimensional semiconducting nanostructure as the active sensing element to detect the presence of local electric fields produced by charged biological molecules.

Overview

Biomolecular Detection Technologies

Methods for detecting and measuring the presence, abundance, and activity of biological molecules, such as proteins, DNA, and RNA, enable the investigation of basic biological function, molecular diagnosis of diseases, and development of therapeutic treatments. The combination of polyacrylamide gel–based electrophoresis separation (and related blotting techniques) and absorption-based chromogenic dye staining (e.g., Coomassie brilliant blue or silver staining) has traditionally formed the core technologies for protein and nucleic acid detection. Subsequent development of fluorescence- and luminescence-based labeling stains offered the opportunity for multicolor labeling, making multiplexed analysis possible. However, these methods were limited to the detection of biological molecules on gels and blots. Immuno-labeling of fluorescent and luminescent probes allowed specific protein detection and facilitated the emergence of immunoassays able to detect and quantify the abundance of a target protein molecule in solution. These optical detection methods suffer from large sample volume consumption, high reagent costs, and long assay times which, when combined with the drawback of using labeling molecules, preclude the application of these technologies to many laboratory and clinical applications (e.g., real-time protein quantification and point-of-care clinical diagnostics).

A number of novel label-free protein-detection strategies have since been developed in an effort to overcome some of the aforementioned limitations. Many of these strategies leverage nonoptical transduction modalities to circumvent the limitations imposed by optical probes. Examples include micro-/nano-surface plasmon resonance sensors [1], micro- and nano-cantilevers that translate biomolecular interactions into mechanical deformations [2], nanostructured field effect transistors (FETs) that measure intrinsic biomolecular charge [3], and electrochemical sensors that translate biomolecular adsorption to changes in redox current [4]. Of these technologies, nano-FET biosensors offer unique device characteristics including ultrasensitive, real-time, and label-free measurement capability, compact physical form, CMOS-compatible and simple on-chip integration with circuitry and microfluidic systems, direct electrical readout, and multiplexed detection capability.

Structure

Nano-FET biosensors are composed of a semiconducting quasi-one-dimensional nanostructure (e.g., nanowire or nanotube) bridging the source and drain electrodes (Fig. 1). The nanostructure serves as the transistor's channel and functions as the sensing element of the device. The biosensor structure deviates slightly from that of the transistor in the placement and structure of the gating electrode. While the gate electrode in a nano-FET is typically located directly above the channel, separated by a thin dielectric layer, the nano-FET biosensor uses a solution gate electrode that is typically composed of platinum or Ag/AgCl and is immersed within the aqueous solution being measured. Furthermore, the nanostructure surface is functionalized with a layer of analyte-specific receptors to impart a degree of specificity to the sensor. This may be accomplished via chemical bonding using bifunctional linker molecules, or through physical interaction due to adsorption, van der Waals forces, and weak electrostatic forces.

While the specific nanostructure employed may vary in shape, size, and composition, it is important that these structures be quasi-one-dimensional and exhibit semiconductive material properties. Typical nanostructures used in the construction of nano-FETs include single and multi-walled carbon nanotubes, graphene, and silicon nanowires. Devices constructed from other semiconductive materials such as metal-oxides and nanostructures of various shapes including nanotubes, nanobelts, nanorods, nanoribbons, and nanobars have also been reported.

Physics

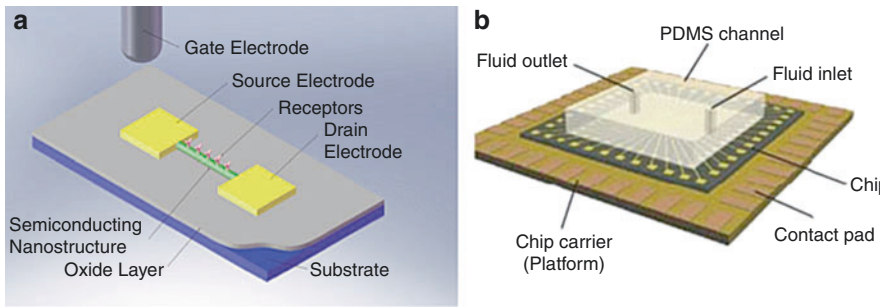
Nano-FET biosensors utilize semiconductive nanostructures to detect the presence of charged molecules in the immediate vicinity around the nanostructure surface. The sensing mechanism is based on the observation that an externally applied electric potential is capable of modulating the number of free charge carriers within a semiconductive material, thus leading to a measurable change in electrical conductance. For a cylindrical semiconducting nanostructure with radius R and length L, the charge carrier concentration may be described as a function of the applied potential:

$$n_0 = \frac{C(V_G - V_t)}{q\pi R^2 L} \tag{1}$$

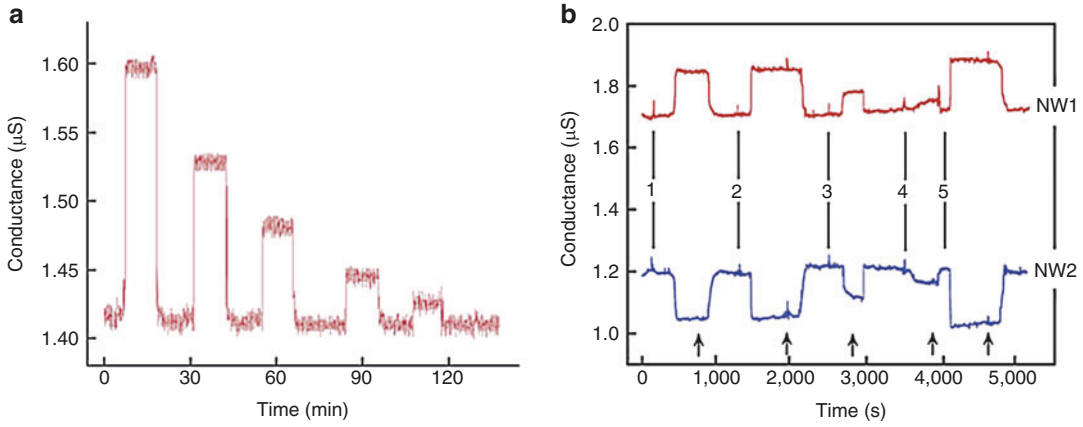
where C is the gate capacitance, q is the elementary charge, V_G is the gate voltage, and V_t is the threshold voltage [6]. As the conductance of the semiconducting nanowire is described by:

$$G = nq\mu\left(\frac{\pi R^2}{L}\right) \tag{2}$$

the current passing through the nanowire (I_D) for a given source-drain voltage (V_{DS}) is therefore:



Nanostructure Field Effect Transistor Biosensors, Fig. 1 (**a**) Schematic of a single nano-FET biosensor. (**b**) Schematic of a nano-FET biosensor chip with integrated microfluidic sample delivery system. The nano-FET biosensor array is housed within the microfluidic system (Adapted with permission from Macmillan Publishers Ltd. [5], copyright 006)



Nanostructure Field Effect Transistor Biosensors, Fig. 2 Real-time nanowire FET sensing results. (**a**) Conductance versus time data recorded following alternate delivery of prostate-specific antigen (*PSA*) and pure buffer solution (1 μM phosphate (potassium salt) containing 2 μM KCl, ph 7.4). Subsequent PSA concentrations were 5 ng ml^{-1}, 0.9 ng ml^{-1}, 9 pg ml^{-1}, 0.9 pg ml^{-1}, and 90 fg ml^{-1}, respectively. (**b**) Complementary sensing of PSA using p-type (NW1) and n-type (NW2) nanowire FET devices. *Vertical lines* correspond to addition of PSA solutions of (*1*, *2*) 0.9 ng ml^{-1}, (*3*) 9 pg ml^{-1}, (*4*) 0.9 pg ml^{-1}, and (*5*) 5 ng ml^{-1}. *Arrows* on the *bottom* represent the injections of sensing buffer solution (Adapted with permission from Macmillan Publishers Ltd. [5], copyright 2006)

$$I_D = \frac{\pi R^2}{L} n_0 q \mu V_{DS} \tag{3}$$

where μ is the carrier mobility. Charged functional groups present on the surface of bound biological molecules such as proteins and nucleic acids provide the necessary electric potential required to induce a conductance change in the semiconductive nanostructure, thus enabling the direct electrical detection of such molecules using nano-FET devices (Fig. 2). The strength of a molecule's surface charge is a function of electrolyte pH and is characterized by the molecule's isoelectric point.

The simplified analysis qualitatively illustrates the physical sensing mechanism behind nano-FET biosensors. However, this treatment does not accurately describe the complexities of the nano-FET biosensor system. Equations 1, 2, and 3 describe the electrical behavior of a cylindrical semiconducting crystal subjected to a uniform charge density over the entire nanostructure surface. This assumption fails to take into consideration the nonuniform nanostructure surface charge distribution resulting from discrete biomolecular binding events. The discrete gating mechanism by individual molecules in solution introduces several additional phenomena that significantly influence the sensor's response including depletion of analyte from solution, analyte transport mechanisms to the nanostructure surface, and analyte-receptor binding and unbinding kinetics. These effects become particularly important in the detection of rare analytes.

Equation 3 also neglects the influence of electrostatic screening – the spatial arrangement of mobile ions around a charged molecule to effectively neutralize that charge. Importantly, target analytes must be sufficiently close to the sensor surface in order to elicit a conductance change due to electrostatic screening by mobile charges around the analyte molecule (e.g., ions in solution and charge carriers in the semiconducting crystal). The Debye screening length, the distance away from a charged molecule at which the electrostatic potential is diminished to zero due to screening, defines the required proximity.

Debye length (aqueous electrolyte):

$$k^{-1} = \frac{1}{\sqrt{8\pi\lambda_B N_A I}} \tag{4}$$

where I is the ionic strength of the electrolyte (mole/m^3), N_A is the Avogadro number, and λ_B is

the Bjerrum length of the medium (0.7 nm for water).

Debye length (silicon):

$$L_D = \sqrt{\frac{\varepsilon_0 \varepsilon_r k_B T}{q^2 N_D}} \quad (5)$$

where ε_r is the dielectric constant of the semiconducting material, k_B is the Boltzmann's constant, T is the absolute temperature in Kelvin, q is the elementary charge, and N_d is the density of donors in a substrate. The implications of charge screening and a finite Debye length make several important predictions on nano-FET biosensing performance. Namely, these relations suggest a conductance dependence on charge carrier concentrations (ionic strength of an electrolyte solution and doping density of a semiconductor), size of the semiconducting crystal, strength of the molecule's net surface charge, and the nanostructure-analyte separation distance. Inclusion of these phenomena into a unified analytical model that accurately describes experimental nano-FET biosensor response is an active area of research.

Performance Parameters

Nano-FET Sensitivity

Nano-FET sensitivity is the induced change in device conductance (ΔG_{sd}), upon exposure to a certain biomolecular stimulus for a constant source-drain voltage (V_{ds}). Sensitivity is defined as:

$$\text{Sensitivity} = \frac{\Delta G_{sd}}{G_0} \quad (6)$$

where ΔG_{sd} is the direct conductance change observed in sensing experiments and G_0 is the initial device conductance [7]. Because ΔG_{sd} depends on the specific nanowire parameters (e.g., diameter, mobility, etc.) this value does not reflect the intrinsic sensitivity. Instead, it is more meaningful to characterize the sensitivity using a dimensionless parameter (6). Normalization enables comparison across devices with different physical dimensions.

Sensitivity is closely related to three additional parameters: *detection limit* (the smallest analyte concentration that can be measured, which is dictated by the system signal-to-noise ratio); *measurement resolution* (the smallest detectable change in analyte concentration, which is dictated by noise levels); and *measurement range* (the measurable analyte concentrations, with the upper and lower bounds defined by nanostructure saturation and the detection limit, respectively). A higher sensitivity directly translates into lower detection limit, finer measurement resolution, and a broader measurement range. Typical protein and DNA detection limits reported in literature are in the pM (picomolar) to fM (femtomolar) range [8, 9]. Single virus detection has also been achieved [10].

Specificity

Specificity is a measure of a biosensors relative responsiveness to the target analyte as compared to all other molecules present within the sample solution. Very specific sensors should possess high sensitivity toward the target analyte and low sensitivity toward all other biomolecules. This attribute is imparted onto a nano-FET biosensor by functionalizing the nanowire surface (either covalently or noncovalently) with a confluent monolayer of analyte-specific receptors/ligands that bind specifically to the molecule of interest.

Response Time

Response time is the time it takes for the biosensor system to obtain a stable biomolecular concentration measurement upon introduction of the sample solution to the sensor surface. As with any dynamic system, this characteristic is captured in the step response of the system by the *settling time*. For a nano-FET biosensor, this parameter is influenced by the analyte-receptor binding and unbinding kinetics, transport of analyte molecules to the sensor surface, concentration and degree of mixing of the analyte solution, and the electrical response behavior of the measurement system. The typical response time of nano-FET biosensors is on the order of minutes.

Key Research Findings

Active areas of research in the field of nano-FET biosensing include: (1) optimization of biosensor performance based on the systematic examination of how fundamental device parameters affect sensitivity, specificity, and response time; (2) development of robust large-scale fabrication strategies for the manufacturing of biosensor arrays; and (3) development of an analytical model capable of accurately describing and predicting nano-FET biosensor performance.

Performance Optimization

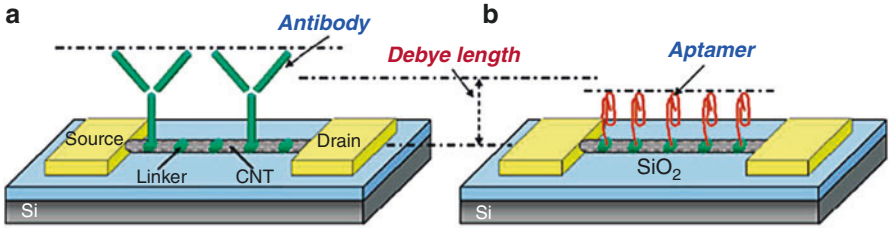
Minimizing the Effects of Charge Screening

As the effect of charge screening acts to minimize the total nanostructure volume gated by surface charges, thus reducing device sensitivity, strategies that increase screening length are expected to improve device sensitivity. The maximum device sensitivity may be obtained in a situation where the effective screening length is much larger than the radius of the nanowire. Improved device sensitivity can thus be obtained through careful design and optimization of several device parameters that influence the effective screening length:

Diameter Reduction of the nanostructure diameter will dramatically increase the surface-to-volume ratio, thus improving device sensitivity. This is the primary advantage of decreasing the channel size and explains why nanostructure FET sensors exhibit significantly higher performance as compared to traditional planar FET devices.

Dielectric Thickness Silicon nanowires form a native oxide layer (~1–2 nm) on their surface that serves as a dielectric layer reducing the electric field strength acting on the nanowire itself. This layer also increases the separation distance between surface-bound molecules and the semiconductive core of the nanowire. Thus, a thinner oxide layer is expected to improve sensitivity. Complete removal of the native oxide layer surrounding silicon nanowires has been shown to increase device sensitivity [11, 12].

Functionalization Scheme and Receptor Size Decreasing the separation distance between the receptor-bound analyte molecule and the nanostructure surface minimizes the effect of electrolyte screening. While antibodies are commonly used to selectively bind analyte molecules to the biosensor surface, the use of smaller capture probes such as antibody fragments or aptamers in place of antibodies can increase device sensitivity without the loss of selectivity (Fig. 3) [13, 14]. Similarly, utilizing shorter bifunctional linker molecules to attach the capture probes to the nanostructure surface may also increase device sensitivity. The choice of functionalization strategy may also dictate the capture probe density and coverage on the nanostructure surface. A poor capture probe density can result in the nonspecific adsorption of molecules onto unmodified areas of the nanowire surface and may also limit the upper



Nanostructure Field Effect Transistor Biosensors, Fig. 3 Schematic representation of nano-FET biosensors functionalized with (**a**) antibodies and (**b**) aptamers. Aptamers (and antibody fragments) are smaller than whole antibodies. The use of these capture probes thus results in increased nano-FET sensitivity. For certain experimental conditions, aptamers may serve as the ideal probe candidate as they are shorter than the system Debye length while antibodies are not (Reprinted with permission from [13]. Copyright 2007 American Chemical Society)

detection limit of the sensor due to capture probe saturation.

Electrolyte Ion Concentration and pH The ionic strength of the electrolyte solution influences the efficiency of charge screening as the density of free ions is directly related to size of the electric double layer formed around a charged molecule. Higher ionic strength electrolytes are thus able to screen a charged molecule over a shorter distance. It is therefore advantageous to perform biosensing experiments in low-ion conditions to maximize sensitivity [15]. However, this is not always feasible as many sensing applications are performed with physiological samples (e.g., serum, whole blood, etc.).

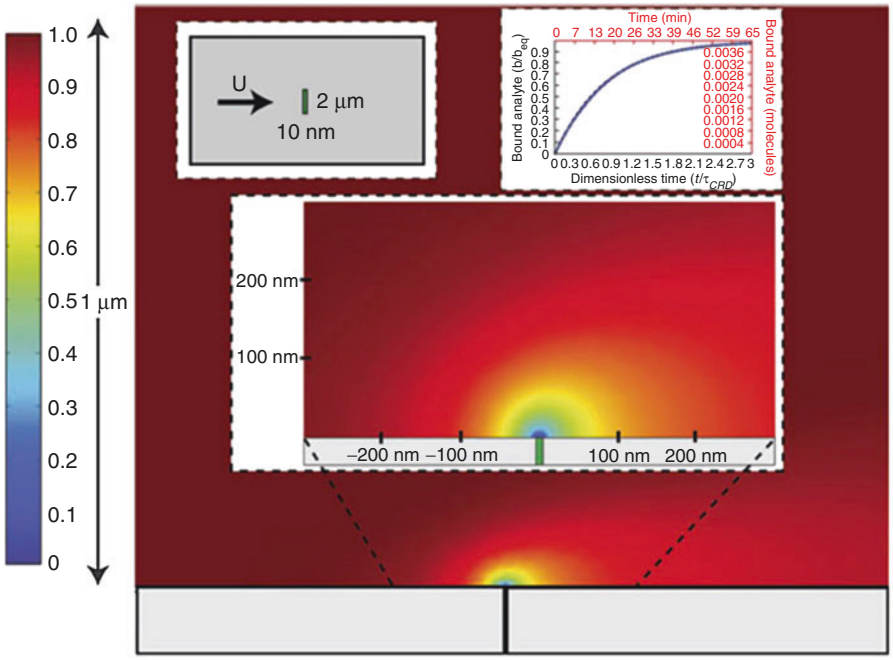
The pH of a solution also influences device sensitivity by changing the effective electrical charge on a biological molecule. Biological molecules such as proteins and nucleic acids contain both acidic and basic functional groups that may be positively or negatively charged depending on the availability of protons (H^+) present in solution. The pH for which the negative and positive charges on the molecule are balanced, resulting in zero net charge, is defined as the isoelectric point (*pI*). The net charge on a molecule is therefore a function of the electrolyte pH and can become increasingly more positively or negatively charged as the term |pI - pH| increases. While tuning this parameter may be useful in specific applications, the pH of a sample is usually fixed at physiological levels (pH = 7.4).

Carrier Concentration The screening length within a semiconductive nanowire dictates the depth of penetration into the nanostructure for which surface-bound charged molecules are able to gate. For example, the screening length in a silicon nanowire operating with a typical carrier concentration of 10^{18}–10^{19} cm^{-3} is $\sim$ 1–2 nm, suggesting that only a small portion of the total nanowire volume (a $\sim$ 1–2-nm thick layer at the nanowire surface) is affected by surface charges. Decreasing the charge carrier concentration will therefore increase device sensitivity by increasing the screening length and the total gated volume. This can be achieved by varying either of two device parameters: the nanostructure doping density or the bias voltage applied through a gate electrode on the nano-FET biosensor. In terms of semiconductors, lower doping densities result in longer screening lengths and therefore produce higher sensitivity devices [16]. It is instructive to note that the optimum biosensor sensitivity is obtained when the device is operating in the subthreshold regime, as under these conditions, the transconductance is at a maximum [7, 17, 18].

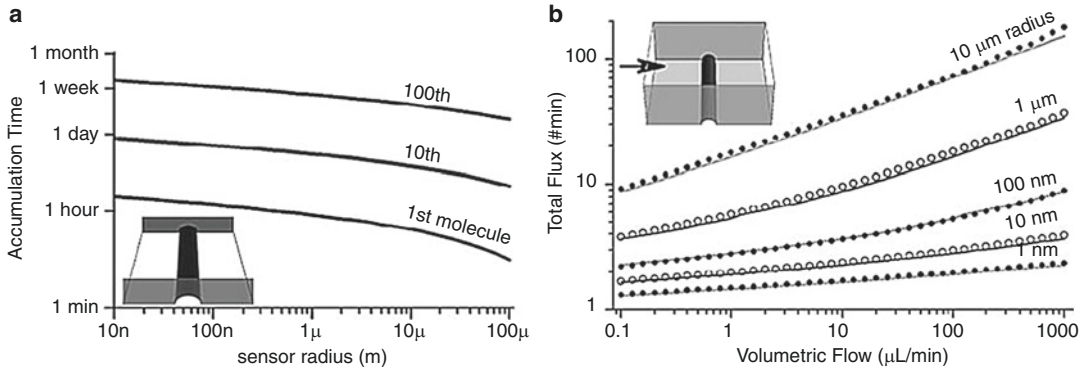
Optimization of Analyte Delivery Efficiency

Biomolecular detection using nano-FET sensors requires the delivery of analyte to the sensor surface. The efficiency of this process can therefore limit the performance of nano-FET biosensors. Several phenomena that influence this process include diffusive and convective molecular transport mechanisms of analyte to the biosensor surface, depletion of free analyte from solution, and nonspecific analyte adsorption and binding.

Molecular Transport As with many surface-based biosensor systems, the transport of analyte in solution to the sensor surface plays a crucial role in governing binding kinetics and ultimately sensor performance. This is especially true when analyzing samples with dilute concentrations of target molecules and/or employing microfluidic systems for efficient and automated handling of small sample solution volumes. A number of competing physical processes influence target transport to the sensor surface. Analyte molecules suspended in solution may diffuse randomly within the solution, be convected along with flowing fluid, bind to adjacent surface-bound receptors, or subsequently unbind to reenter solution. The binding (or collection) of biomolecules onto the sensor surface simultaneously depletes the surrounding solution to form a so-called depletion region around the sensor (Fig. 4). As the depletion region grows in size, the analyte diffusive flux decreases and thus the collection rate becomes slower. Through finite element analysis, Sheehan and Whitman determined that the size and shape of the sensor profoundly affects the total analyte flux to the sensor surface (Fig. 5a)



Nanostructure Field Effect Transistor Biosensors, Fig. 4 The steady-state depletion zone around a nanowire sensor from a pure mass-transport problem. The depletion zone is thick compared to the nanowire, but is substantially thinner than the channel itself ($\delta_s \sim L\mathrm{Pe}_s^{-1/2} = 100$ nm). Reaction-limited binding onto the nanowire follows the Langmuir binding curve (inset *top right*) (Adapted with permission from Macmillan Publishers Ltd. Nature Biotechnology [19], copyright 2006)



Nanostructure Field Effect Transistor Biosensors, Fig. 5 (**a**) Theoretical calculation of the time required for a 10-μm-long hemicylindrical sensor to accumulate 1, 10, and 100 molecules due to diffusive molecular transport. The sensor lies at the *bottom* of a channel whose width is equal to the sensor's length and which is filled with a 1 fM analyte solution. (**b**) Finite element analysis (*points*) and theoretical calculation (*lines*) of the total flux of molecules to a hemicylindrical sensor in a microchannel (800 μm wide, 100 μm high). The total flux plotted is the steady-state value at the given volumetric flow rate for a 1 fM analyte concentration (Adapted with permission from [20]. Copyright 2005 American Chemical Society)

and further argues that without directed transport of biomolecules, individual nanoscale sensors would be limited to picomolar order sensitivity for practical time scales (hours to days) [20].

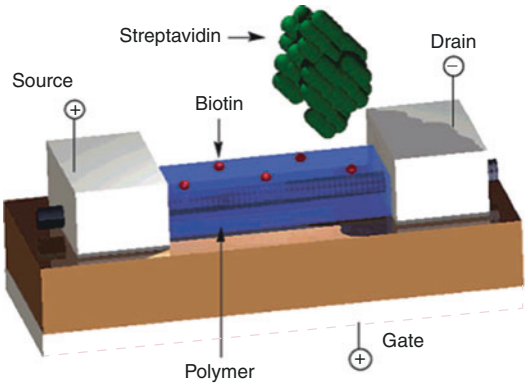
While the diffusive depletion zone grows indefinitely, albeit at an ever-decreasing rate, introduction of convective flow (assumed to be laminar) halts this growth resulting in a steady-state depletion zone with a length scale defined by the balance between the convective analyte flux delivered to the depletion zone boundary and the diffusive flux out of the depletion zone. By adjusting the convective flow rate through the channel, it is possible to tune the size of the depletion layer and thus control the rate at which molecules are delivered to the sensor surface (Fig. 5b). However, Squires et al. argue that the total mass transport varies only weakly with the flow rate as flow rates must be increased 1,000-fold to enhance flux by a factor of 10 [19]. The treatment thus far has assumed instantaneous analyte-receptor coupling upon transport of target molecules to the sensor surface. This represents one of two sensing regimes that can be described as being "transport-limited" as the availability of analyte adjacent to the sensor surface limits sensor response.

A second biosensing regime arises when considering the kinetics of analyte-receptor association and dissociation (assumed to be first-order Langmuir kinetics) [19]. If delivery of target molecules to the sensor surface occurs at a faster rate than the net analyte-receptor association rate, then the transport is said to be "reaction limited." In general, the reaction kinetics is determined by the fidelity of the immobilized reagents. While each of the two regimes imposes an independent maximum target collection rate, the "reaction-limited" condition is most desirable for nano-FET operation as, for a given device, biomolecular sensing is purely a function of analyte concentration. The sensor size and flow rate should thus be adjusted to ensure nano-FET operation within this regime. Alternatively, introduction of turbulent flow or fluid mixing through innovative microfluidic design may also be employed to overcome the limitations imposed by convective and diffusive molecular transport [8].

Depletion of Free Analyte As silicon and metal-oxide nanostructures share similar surface chemistries with several popular microfabrication substrates (e.g., silicon wafers, glass, quartz), surface modification techniques used to attach receptor molecules to the nanostructure surface can also functionalize the device substrate. In this situation, the total device surface area exposed to the sample solution during biomolecular detection experiments is proportional to the number of receptors available for reaction. Analyte binding events with substrate-bound receptors do not induce a change in nanostructure conductance. Rather, these reactions serve only to deplete free analyte molecules from solution, thus reducing the number of analyte molecules available for reaction with nanostructure-bound receptors. Minimizing the total surface area exposed to sample solutions is, therefore, important for maximizing device sensitivity. This effect may be circumvented using nanostructure-specific surface modification chemistries.

Nonspecific Adsorption Random adsorption of biological molecules onto the nano-FET sensor produces a false-positive measurement signal. This effect may be minimized by grafting a dense layer of long nonreactive polymer chains (e.g., polyethylene glycol) onto the nanostructure surface alongside receptor molecules (Fig. 6). The polymer chains prevent molecular adsorption directly onto the nanostructure surface by acting as a spacer while the co-functionalized receptors are able to bring analyte molecules into intimate contact with the nanostructure through molecular binding. The length of the polymer chains used for this application should exceed the screening length of the system.

Nonspecific adsorption of random molecules is especially problematic when working with samples containing numerous types of biological molecules such as whole blood. One strategy that enables detection of a specific analyte within a whole blood sample involves the use of a sample pre-purification procedure before nano-FET measurement [22]. In this method, the whole blood is introduced into a chamber in which analyte-specific receptors have been immobilized on to



Nanostructure Field Effect Transistor Biosensors, Fig. 6 Schematic of a nanotube field effect transistor coated with a polymeric functional layer. Co-functionalization with a molecular receptor (biotin) and a nonreactive polymer (e.g., polyethylene glycol or polyethylene oxide) prevents nonspecific adsorption of molecules onto the nanotube surface (Adapted with permission from [21]. Copyright 2003 American Chemical Society)

the chamber walls. Upon binding of free analytes to the surface-bound receptors, the whole blood solution is flushed out of the chamber and discarded. The receptors are then detached from the chamber surfaces and transported to the nano-FET biosensor for analysis.

Large-Scale Methods of Fabrication

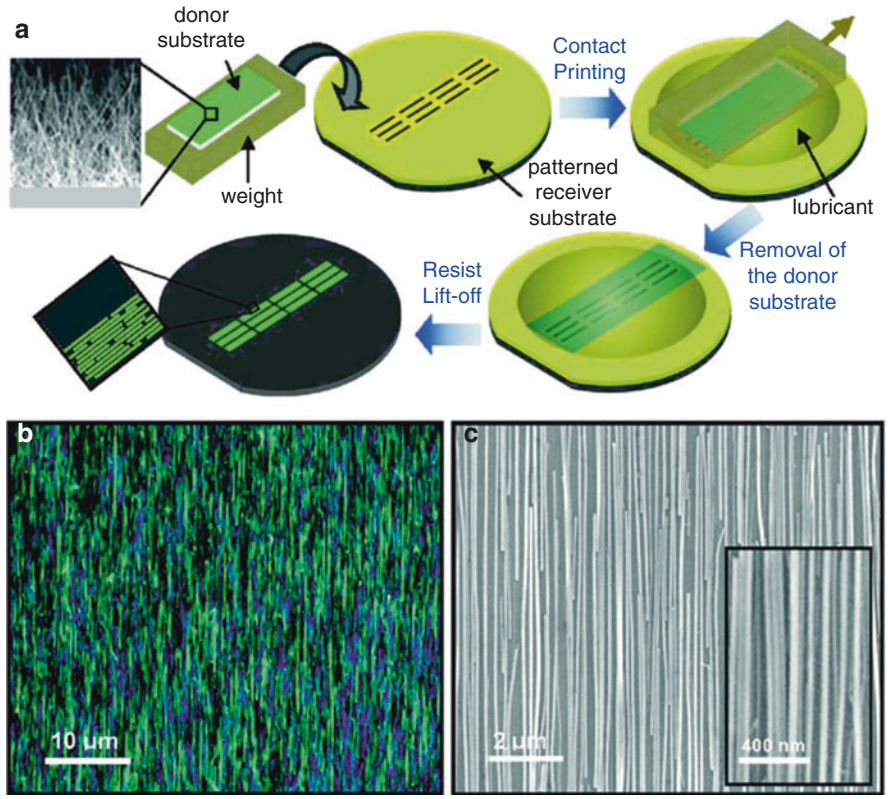
Reliable large-scale fabrication of nano-FET biosensors remains to be a significant technological challenge, impeding the adoption of this technology in medical diagnostic and biomolecular detection applications. Unlike traditional microfabrication methods for producing micrometer-level structures, the difficulties encountered in nano-FET biosensor fabrication arise from the process of establishing electrical contact with nano-sized structures in a high-throughput and reproducible manner.

Early methods for manufacturing nano-FETs involve (1) the growth of individual nanostructures using a chemical vapor deposition process, (2) transfer of as-grown nanostructures from the growth substrate onto a device substrate (a process which results in the random placement and orientation of the deposited nanostructures), and (3) identification of a suitable nanostructure on the device substrate followed by the patterning of metallic electrodes onto the ends of that nanostructure using a mask-less nano-lithographic tool (e.g., electron beam lithography or focused ion beam lithography), which has a low throughput and is only suitable for proof-of-principle use. Recently devised batch-fabrication strategies for the wafer-scale production of nano-FET arrays can be broadly categorized into "bottom-up" or "top-down" fabrication methods.

Bottom-Up Fabrication Methods

One strategy for bottom-up fabrication is to make use of pre-synthesized freestanding nanostructures using epitaxial growth methods such as chemical vapor deposition. This strategy affords several advantages over top-down and other bottom-up approaches including (1) strict control over nanostructure dimensions and electrical properties, (2) the ability to assemble nano-FET devices onto flexible and transparent substrates, and (3) the ability to incorporate nanostructures with unique chemical compositions and architecture into nano-FET devices. Successful incorporation of these nanostructures into nano-FET devices relies primarily on the ability to transfer, align, and position the as-grown nanostructures onto a device substrate. Contacting electrodes may then be patterned using conventional photolithography to form individual sensors or sensor arrays.

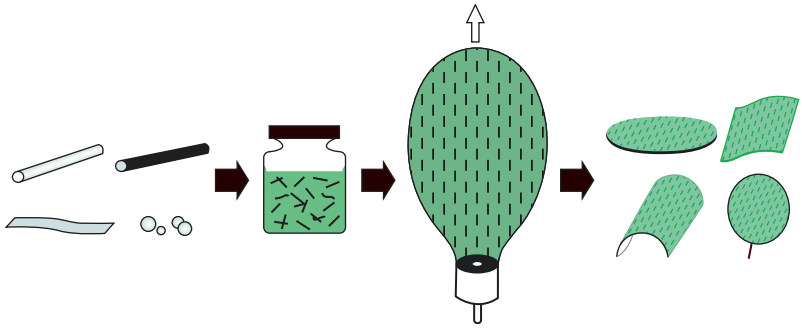
Flow Alignment Flow alignment is a method that makes use of flowing fluid to align suspended nanostructures along a single orientation in the direction of flow [23]. By passing a nanostructure suspension through a microfluidic structure formed between a poly(dimethylsiloxane) mold and the device substrate, free-flowing nanostructures assemble onto the device substrate due to surface forces and remain relatively aligned parallel to the direction of flow. The resulting degree of alignment is a function of the flow rate as increased flow velocities result in a narrower angular distribution among deposited nanostructures. Furthermore, varying the total flow duration dictates the nanostructure deposition density, with longer flow durations favoring



Nanostructure Field Effect Transistor Biosensors, Fig. 7 Contact printing of nanowires. (**a**) Schematic of the contact printing process for producing well-aligned nanowire arrays. (**b**) Dark-field optical and (**c**) SEM images of 30-nm nanowires printed on a Si/SiO2 substrate showing highly dense and aligned monolayer of nanowires. The self-limiting process limits the transfer of nanowires to a single layer, without significant nanowire bundling (Adapted with permission from [24]. Copyright 2008 American Chemical Society)

higher nanostructure densities. This method is useful in controlling the degree of nanostructure alignment and the deposition density, thus allowing subsequent electrode formation. Nano-FET fabrication can therefore be achieved by patterning a pair of electrodes parallel to the flow direction using conventional microfabrication processes. For this physical arrangement, minimizing the electrode pair separation distance increases the likelihood of a bridging nanostructure between the electrode pair; however, the exact number of bridging nanostructures resulting from this process is probabilistic.

Nanostructure Contact Printing Nanostructure contact printing is a method that provides similar results to the flow alignment strategy. This method, however, relies on mechanical shear forces to detach freestanding nanostructures from the growth substrate, and friction to align the deposited the nanowires onto the device substrate [24]. In this method, the growth substrate is inverted and placed on top of the device substrate such that the nanostructures are sandwiched in between (Fig. 7). Weights are added on top of the growth substrate to control the applied pressure between the two substrates in order to optimize the deposited nanostructure density. The growth substrate is then pulled over the device substrate at a continuous velocity in order to deposit and align the nanostructures onto the device substrate. The shear velocity and the amount of friction between the two substrates influence the resulting degree of nanostructure alignment. The latter may be adjusted with the use of a lubricant.



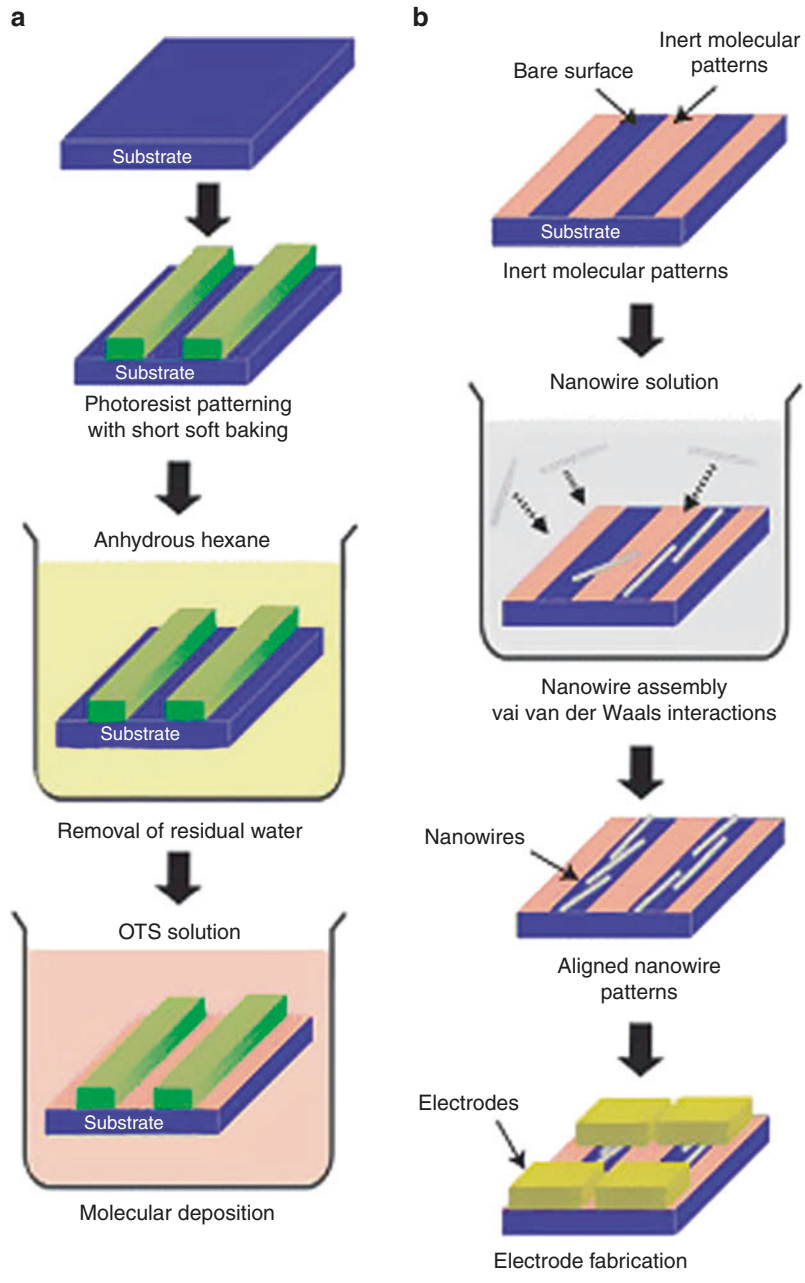
Nanostructure Field Effect Transistor Biosensors, Fig. 8 Blown bubble film approach. Nanostructures are dispersed in a polymer solution. A volume of solution is expanded as a bubble using a die to produce well-aligned nanostructures suspended within a thin film. The film is then contacted with a substrate to deposit well-aligned nanostructures onto the substrate ([25] Reproduced with permission of The Royal Society of Chemistry)

Thin-Film Nanowire Suspension Thin-film nanowire suspension is a method that makes use of the observation that nanostructures suspended within a liquid thin film will align along a single direction when that film is elongated. Therefore, this method involves the formation of a large balloon from a nanostructure suspension (Fig. 8) [25]. The aligned nanostructures within the thin film can then be transferred to a substrate by placing the substrate in contact with the balloon. Nano-FET fabrication would subsequently proceed as in the case of flow alignment.

Directed Self-Assembly Directed self-assembly relies on the minimization of surface energy to assemble suspended nanostructures into predefined locations on a substrate [26]. In this method, surface modification strategies are used to render certain portions of the device substrate either hydrophobic or hydrophilic (Fig. 9). Upon immersion of the substrate into a nanostructure suspension, individual nanostructures will self-assemble onto the hydrophobic portions of the substrate. This strategy can be used to deposit individual nanostructures at specific locations over a large area on a substrate. A similar strategy using analyte-ligand interactions has also been employed in which suspended nanostructures are functionalized with biotin molecules. Upon immersion of a substrate with bound streptavidin molecules at predefined locations, biotin-streptavidin conjugation forces the nanostructures to assemble onto the device substrate at predefined locations.

Dielectrophoresis Dielectrophoresis (DEP) utilizes electrical fields to manipulate nanostructures. In this method, metallic contact electrodes are patterned on the device substrate prior to the assembly of nanostructures onto the device. By applying a biased alternating voltage across the electrodes, a local nonuniform electric field is produced, exerting a force on suspended semiconducting nanostructures. This force causes the nanostructures to assemble across the two electrodes. This process is easily amenable to the large-scale production of nano-FET arrays and can be used to fabricate single nanostructure FET devices; however, typical DEP trapping processes yield networked multi-nanostructure FET devices [27]. This process is highly susceptible to the experimental conditions including the size, shape, and properties of the nanostructure to be manipulated, and parameters of the electrical signals, as well as the electrical properties of the surrounding medium.

In Situ Growth of Nanostructures In situ growth of nanostructures is a method that utilizes patterned catalysts to selectively synthesize nanostructures at certain regions on the device substrate. These catalysts (e.g., iron or gold particles) are required to initiate and/or propagate nanostructure growth. In some cases, the direction

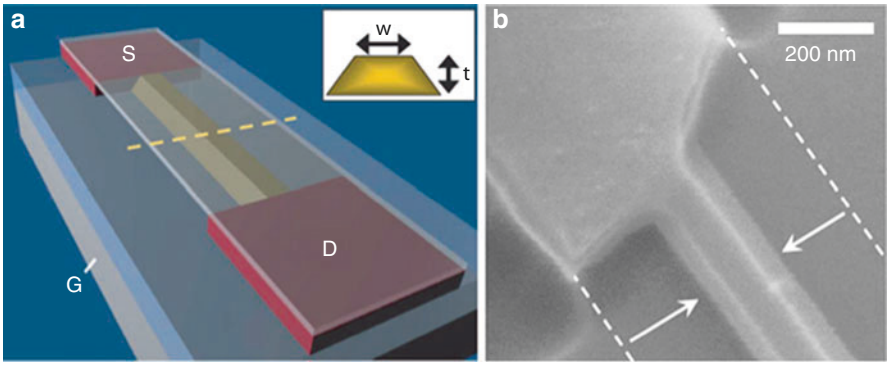


Nanostructure Field Effect Transistor Biosensors, Fig. 9 Linker-free directed self-assembly. (**a**) The patterning process of octadecyltrichlorosilane self-assembled monolayer on a solid substrate via photolithography. (**b**) Self-assembly suspended nanostructures onto the bare surface regions on the substrate. Subsequent photolithography and metallization can be used to pattern electrodes onto the ends of the deposited nanostructures (Adapted with permission from Macmillan Publishers Ltd. Nature Nanotechnology [26]: copyright 2006)

of nanostructure growth can be influenced with the application of an electric field, therefore enabling control of the resulting nanostructure orientation. This method usually produces numerous nanostructures and is, therefore, useful in fabricating nano-FET sensors that utilize a network of nanostructures.

Top-Down Fabrication Method

Anisotropic Lateral Wet Etching Top-down methods are typically based on the anisotropic lateral wet etching of nanometer-thin SOI (silicon-on-insulator) wafers to produce well-defined nanostructures from the device layer



Nanostructure Field Effect Transistor Biosensors, Fig. 10 (**a**) Schematic of nano-FET device after anisotropic etching. The silicon-on-insulator active channel (*yellow*, width w and thickness *t*) is undercut etched, whereas degenerate leads (*red*) are etch-resistant. The source (*S*), drain (*D*), and underlying backgate (*G*) are labeled. (**b**) Scanning electron micrograph of a complete device (Adapted with permission from Macmillan Publishers Ltd: Nature [26], copyright 2007)

(Fig. 10) [8]. Micro- or nano-sized etch masks are patterned via conventional or e-beam photolithography and anisotropically time-etched to produce nanometer-wide nanostructures. Source and drain electrodes are degenerately doped, rendering them conductive and unaffected by the etchant. This approach enables wafer-scale formation of semiconducting nanostructures at precise locations on a wafer-scale, making subsequent electrode fabrication relatively straightforward using conventional photolithography. As this method relies on time-controlled anisotropic etching for the removal of material, it is highly susceptible to small changes in processing conditions such as etch time, processing temperature and mixing, and device-layer doping density and crystal orientation.

Future Directions

At this point in time, the development of nano-FET biosensors is still in the proof-of-concept phase. The promise of highly sensitive, label-free electrical sensors with real-time measurement capability remains attractive for numerous medical and basic science research applications. While great progress has been made in nano-FET fabrication and detection, continued research into large-scale fabrication methods for batch-manufacturing nano-FET sensors will be essential for the ultimate commercial success of this technology and its application to clinical and research applications. In addition to low cost and high yield, these fabrication methods must produce devices with consistent sensing performance across the nano-FET sensors such that measurements made with these devices can be standardized. Another critical hurdle that must be overcome is the issue of poor analyte specificity and nonspecific binding. While some strategies have been put forth to remedy these issues, these solutions must be further improved upon before nano-FET biosensors are able to reliably analyze whole blood, serum, and other specimens. Lastly, efforts in the area of nano-FET biosensor integration with microfluidic and lab-on-chip devices will facilitate point-of-care diagnostic applications and real-time closed-loop drug delivery systems. Successful lab-on-chip integration will also enable the creation of multiplexed nano-FET biosensor arrays, which could hold great promise in the areas of basic chemical and biological research, high-throughput screening systems for drug development, and novel in vitro biology experimentation. The future will likely see many point-of-care biosensors using electrical-based detection, and nano-FET biosensing technologies can possibly play an important role.

Cross-References

- ▶ Biosensors
- ▶ Carbon Nanotubes
- ▶ Chemical Vapor Deposition (CVD)
- ▶ Dielectrophoresis
- ▶ Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes
- ▶ Electric-Field-Assisted Deterministic Nanowire Assembly
- ▶ Functionalization of Carbon Nanotubes
- ▶ Nanophotonic Structures for Biosensing
- ▶ Nanostructured Materials for Sensing
- ▶ Nanotechnology
- ▶ Self-assembly of Nanostructures

References

1. Lee, S.W., et al.: Highly sensitive biosensing using arrays of plasmonic au nanodisks realized by nanoimprint lithography. ACS Nano **5**, 897–904 (2011)
2. Fritz, J.: Cantilever biosensors. Analyst **133**, 855–863 (2008)
3. Cui, Y., Wei, Q., Park, H., Lieber, C.M.: Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. Science **293**, 1289–1292 (2001)
4. Soleymani, L., Fang, Z., Sargent, E.H., Kelley, S.O.: Programming the detection limits of biosensors through controlled nanostructuring. Nat. Nanotechnol. **4**, 844–848 (2009)
5. Patolsky, F., Zheng, G., Lieber, C.M.: Fabrication of silicon nanowire devices for ultrasensitive, label-free, real-time detection of biological and chemical species. Nat. Protoc. **1**, 1711–1724 (2006)
6. Fan, Z., Lu, J.G.: Gate-refreshable nanowire chemical sensors. Appl. Phys. Lett. **86**, 123510–123513 (2005)
7. Gao, X.P., Zheng, G., Lieber, C.M.: Subthreshold regime has the optimal sensitivity for nanowire FET biosensors. Nano Lett. **10**, 547–552 (2010)
8. Stern, E., et al.: Label-free immunodetection with CMOS-compatible semiconducting nanowires. Nature **445**, 519–522 (2007)
9. Zheng, G., Patolsky, F., Cui, Y., Wang, W.U., Lieber, C.M.: Multiplexed electrical detection of cancer markers with nanowire sensor arrays. Nat. Biotechnol. **23**, 1294–1301 (2005)
10. Patolsky, F., et al.: Electrical detection of single viruses. Proc. Natl. Acad. Sci. U. S. A. **101**, 14017–14022 (2004)
11. Bunimovich, Y.L., et al.: Quantitative real-time measurements of DNA hybridization with alkylated nonoxidized silicon nanowires in electrolyte solution. J. Am. Chem. Soc. **128**, 16323–16331 (2006)
12. Zhang, G.J., et al.: Highly sensitive measurements of PNA-DNA hybridization using oxide-etched silicon nanowire biosensors. Biosens. Bioelectron. **23**, 1701–1707 (2008)
13. Maehashi, K., et al.: Label-free protein biosensor based on aptamer-modified carbon nanotube field-effect transistors. Anal. Chem. **79**, 782–787 (2007)
14. Zhang, G.J., et al.: DNA sensing by silicon nanowire: charge layer distance dependence. Nano Lett. **8**, 1066–1070 (2008)
15. Stern, E., et al.: Importance of the Debye screening length on nanowire field effect transistor sensors. Nano Lett. **7**, 3405–3409 (2007)
16. Nair, P.R., Alam, M.A.: Design considerations of silicon nanowire biosensors. Electron Devices IEEE Trans. **54**, 3400–3408 (2007)
17. Heller, I., Mannik, J., Lemay, S.G., Dekker, C.: Optimizing the signal-to-noise ratio for biosensing with carbon nanotube transistors. Nano Lett. **9**, 377–382 (2009)
18. Lu, M.P., Hsiao, C.Y., Lai, W.T., Yang, Y.S.: Probing the sensitivity of nanowire-based biosensors using liquid-gating. Nanotechnology **21**, 425505 (2010)
19. Squires, T.M., Messinger, R.J., Manalis, S.R.: Making it stick: convection, reaction and diffusion in surface-based biosensors. Nat. Biotechnol. **26**, 417–426 (2008)
20. Sheehan, P.E., Whitman, L.J.: Detection limits for nanoscale biosensors. Nano Lett. **5**, 803–807 (2005)
21. Star, A., Garbiel, J.P., Bradley, K., Gruner, G.: Electronic detection of specific protein binding using nanotube FET devices. Nano Lett. **3**, 459–463 (2003)
22. Stern, E., et al.: Label-free biomarker detection from whole blood. Nat. Nanotechnol. **5**, 138–142 (2010)
23. Huang, Y., Duan, X., Wei, Q., Lieber, C.M.: Directed assembly of one-dimensional nanostructures into functional networks. Science **291**, 630–633 (2001)
24. Fan, Z., et al.: Wafer-scale assembly of highly ordered semiconductor nanowire arrays by contact printing. Nano Lett. **8**, 20–25 (2008)
25. Yu, G., Li, X., Lieber, C.M., Cao, A.: Nanomaterial-incorporated blown bubble film for large-area aligned nanostructures. J. Mater. Chem. **18**, 728–734 (2008)
26. Lee, M., et al.: Linker-free directed assembly of high-performance integrated devices based on nanotubes and nanowires. Nat. Nanotechnol. **1**, 66–71 (2006)
27. Vijayaraghavan, A., et al.: Ultra-large-scale directed assembly of single-walled carbon nanotube devices. Nano Lett. **7**, 1556–1560 (2007)

Nanostructured Antireflective Surfaces

- ▶ Moth-Eye Antireflective Structures

Nanostructured Functionalized Surfaces

Lorenzo Lunello[1], Cristina Potrich[1], Laura Pasquardini[2] and Cecilia Pederzolli[2]
[1]Biofunctional Surfaces and Interfaces, FBK-CMM Bruno Kessler Foundation and CNR-IBF, Povo, TN, Italy
[2]Biofunctional Surfaces and Interfaces, FBK-CMM Bruno Kessler Foundation, Povo, TN, Italy

Synonyms

Functionalized nanomaterials

Definition

Nanostructured functionalized surfaces are materials that contain microscale and nanoscale features produced in a well-controlled manner with nanocoating, film deposition, or in situ surface modifications at nanoscale level, with the aim to improve the material performances. The potential applications of nanostructured functionalized surfaces are wide; some examples are in vivo imaging and diagnostics (nanosized contrast agents for medical imaging, miniaturized devices); in vitro miniaturized diagnostics (lab-on-a-chip, point-of-care measurements); drug development, delivery, and therapy (nanoparticles, nanobody, affibodies); tissue engineering; and implants of drug dispenser/factory.

Overview

Nanostructures or nanomaterials have peculiar physical and/or chemical properties, different from their bulk material. Main differences rely on topography or chemistry or on both aspects of the outer layer of the material which becomes "nanostructured" when these topographical and/or chemical modifications are made at a nanometric level. Nanostructured surfaces can be produced by surface modification of materials conventionally interacting with biological systems such as human body (biomaterials). Nanofunctionalization techniques can be divided in two main categories: (a) nanocoating and film deposition and (b) in situ surface nanofunctionalization. These two types of techniques are often combined to produce surfaces with hybrid nanostructures such as a coating/film and a nanostructured zone. Typical techniques belonging to the first category are, for example, plasma spraying, plasma immersion ion implantation and deposition, chemical or physical vapor deposition, cold spraying, and self-assembly, whereas the second category is represented by methods like laser etching, shot blasting, acid and alkali treatments, anodic oxidation, micro-arc oxidation, ion implantation, and similar [1].

The nanofunctionalization of the surface can improve the performances of biomaterials, helping their use in clinical applications. In fact, cells life in complex tissues is controlled through modulation of cell binding to the extracellular matrix, which is composed by nanostructures. Cells respond to micro- and nanofeatures with different chemistries and topographies, resulting in changes in cell alignment, polarization, elongation, migration, proliferation, and gene expression. Through micro- and nanopatterning techniques it is possible to precisely position selected biomolecules on a substrate for driving cell growth and, possibly, regulating cell functions [2].

While flat nanostructured surfaces are mainly suitable for cells growing in different conditions depending on chemical and topographical patterning of substrates, other nanostructures are also deeply studied for various applications. Nanostructures such as metal or metal oxide nanoparticles, quantum dots, silica, carbon nanotubes, C_{60}, hydrogel, block copolymer, and dendrimer are reported [3]. Main applications of these different types of nanostructures are imaging (both biolabeling and biosensing, comprising diagnostics purposes) and drug or gene delivery (for instance nanobodies or affibodies against cancer cells). Nanostructured surfaces, such as possibly functionalized nanoparticles or even nanotubes, are also included in micro- and

nanodevices aimed at diagnostics and/or medical treatments.

Various methods for the fabrication of nanostructured functionalized surfaces are available, either adapted from macro-scale technologies or ad hoc developed, while fewer are the techniques suitable for the production of nanoarrays containing multiple features such as a range of different proteins or small ligands. Topographical and chemical nanopatterning of surfaces by methodologies substantially developed for fabrication in the microelectronic field are not reported in this entry, whereas scanning probe–based techniques utilized to obtain functional structures at nanoscale (such as nanoarray) with different features are described in the following paragraphs. They are based on, for example, dip pen nanolithography, SNOM lithography, and nanografting.

Flat Nanostructured Surfaces

Flat nanostructured functionalized surfaces are surfaces with defined regions of different chemical functionality to achieve site-specific attachment of one species in some areas while minimizing unwanted surface interactions in other areas. Two surface patterning types can be distinguished, a structural topography and a chemical patterning, but at nanometric scale these features can be mixed. Moreover, a topographical pattern can be performed maintaining the surface chemistry, but the majority of chemical techniques induce a change in topography. Different strategies to achieve such patterned surfaces have been proposed, mainly using techniques developed in the microelectronics field to chemically modify some parts of the surface, for example, via electron irradiation or extreme UV interference lithography or via microcontact printing. Alternatively, the combination of different substrate materials for the creation of patterned surfaces, commonly via photolithography, is also reported [4]. The production of patterns of low cost and high complexity (three or more addressable chemistries) over large areas was recently described [5]. The method proposes the creation of a long-range-ordered array of microscale "cups" with multiple material regions by a combination of colloidal lithography and the bottom-up deposition techniques of plasma polymerization and evaporation as well as sputtering via physical vapor deposition. Plasma polymerization possesses indeed the unique practical advantages that offer single-step coating with nanometer-scale control over thickness and a wide range of functional groups can be presented with control over density.

Patterned surfaces gained great attention since cells respond to micro- and nanofeatures with different chemistries and topographies. The influence of various topographic features such as grooves, ridges, stops, pores, wells, and nodes in microscale or nanoscale has been analyzed by culturing a wide variety of cells in the presence of such patterned nanosurfaces. Fibroblasts are greatly studied with nanotopography, showing as a general behavior that the adhesion increases with decreasing feature size or height. Morphology and cytoskeleton arrangement also change. The details of the nanotopography, and hence randomness and pattern, play an important role in eliciting different cell responses [6]. Muscle cells, endothelial cells, blood cells, hepatocytes, and many other cell lines were also studied, as revised by Yim et al. [6], concluding that cells vary their adhesion, proliferation, migration, and gene expression properties if grown on nanostructured substrates. This behavior could be ascribed to the fact that cells in their natural environment interact with extracellular matrix components which are in the nanometer scale. For example, the basement membrane of many tissues displays features of pores, fibers, and ridges in the nanometer range.

Moreover, surface nanofunctionalization was demonstrated to significantly affect cellular and subcellular functions, as reviewed by Liu et al. [1], stating that clinical applications of biomaterials can be improved by producing a nanostructured surface. Apart from physical or chemical treatments aimed at the creation of nanostructures on a material surface, the modification of a substrate with specific ligand which improves cell attachment and differentiation is also well documented.

Bell et al. [7], for example, reported the use of oligopeptides to coat titanium surfaces for osteointegration of orthopedic implants by favoring attachment of osteoprogenitor cells and promoting osteoblastic differentiation. Stem cell differentiation to desired phenotypes was also deeply studied in response to nanostructured growing substrates [2]. In particular, it was demonstrated that it is possible to drive cell adhesion, migration, proliferation, and even differentiation of stem cells, by the design and fabrication of growing surfaces or scaffolds with controlled topography and chemistry at the micro- and nanoscale. Topography acts on cell shape and alignment through mechanical interaction of the plasma membrane and the specific ligands present on it with the nanostructures on the substrate surface. However, the most important parameter at micro- and nanometer level seems the order or randomness characteristic of the surface. Chemical patterning regulates cell functions by mechanical interaction (e.g., adhesive micro-islets) or by direct receptor-ligand interaction with cell-membrane proteins. Topography and chemistry can also be combined to have synergistic effects and mimic the natural cell micro-environment.

Apart from eukaryotic cells, bacteria can also colonize surfaces possibly responding to nanotopography. Main difference is however the fact that bacteria live on surfaces as a community within a species- and strain-specific extracellular polymer, forming the so-called biofilm. Little is known about the response of bacteria toward nanostructures and only indirect experimental results support the idea that bacteria are able to use molecular features of their cell membrane as sensors and implement intracellular signaling pathways to sense the surface and to react to the stimuli created by the surface [8].

Functionalized Nanostructures

Various nanostructures or nanomaterials sensitive to environmental or biological parameters have been reported. Usually they are made of metals, metal oxides, semiconductors, silica, carbon nanotubes, C_{60}, hydrogel, block copolymer, and dendrimer [3]. Nanometer-sized metallic structures, such as platinum, gold, and silver nanoparticles, exhibit unique optical properties. In fact, the characteristics of the plasmons excited by the light on their surface are depending on the dielectric constant of the media that constitute the interface. The adsorption of molecules to the nanoparticle changes this parameter, leading to detectable optical changes of the plasmon resonance. Moreover, these nanostructures can be functionalized to specifically interact, for example, with cells for site-specific drug delivery or imaging purposes. It was clearly demonstrated that differently decorated nanoparticles interact in different extent with cell membranes and cell uptake is indeed greatly influenced [9]. Iron oxide magnetic nanoparticles have been used widely in biological detection because of their good biocompatibility and ability to be manipulated under an external magnetic field. Quantum dots are also extensively used for biolabeling and biosensing, since they exhibit advantages in signal brightness, photostability, multicolor-light emission, and size-tunable properties.

Carbon-based nanomaterials are recently gaining interest for biomedical applications. Carbon nanotubes have been used as drug or gene carriers and for biosensing, whereas certain fullerene C_{60} derivatives and fluorescent carbon dots have been developed as nontoxic alternatives to semiconductor quantum dots for biomedical imaging [10].

Self-assembly has been also extensively explored to functionalize the material surface and build nanostructures or nanomaterials, including ordered mono- and multilayers, liposomes or vesicles, nanoparticles, nanotubes, and nanofibers. Moreover, self-assembled nanostructures from block copolymers have also frequently been reported. They consist of two or more covalently bounded blocks with different physical and chemical properties, resulting in a supramolecular assembly that can respond to changes in temperature, pH, and redox potential. In fact, they can generate various microdomain morphologies by means of intramolecular forces.

Since active selectivity is often desirable, targeted drug delivery carriers are functionalized

with antibodies or antibody fragments to provide such active localization. The conjugation of multiple antibodies to each nanocarrier enhances their specificity and ability to detect their targets; nanocarriers can thus be surface-functionalized with multiple distinct antibodies to overcome tumor heterogeneity. Functional, single-domain heavy-chain antibodies, referred to as nanobodies, have been raised against cancer targets with the purpose to either antagonize receptor function or deliver an enzyme for prodrug activation. Affibodies against a variety of cancer-related targets have been developed and are now commercially available, including: EGFR, HER2, and transferrin [11]. Nanoparticles have also been capped with various biocompatible polymers such as PVA, dextran, silica, and inorganic capping agents like biotin, citric acid, avidin, carbodiimide, SiO_2, and chitosan-PEG. These molecules then act as linkers for the coupling of cytotoxic drugs or target antibodies. A step forward is, for example, the design of luminomagnetic nanocarriers functionalized with the ligand folic acid [12]. This approach combines the luminescent and magnetic properties of the luminomagnets with the specificity of folic acid making them feasible for folate receptor-mediated molecular imaging and targeted drug delivery. Besides these, the small size of these luminomagnetic particles (about 8 nm) can be exploited for their easy elimination from the body after delivery of the drugs.

Furthermore, micro- and nanostructures are included in the development of smaller medical devices or systems aimed to drug and gene delivery, tissue engineering, biosensors, and diagnostic systems. The final goal relies in developing implantable sensing and/or treatment devices to collect information for diagnosing disease and administrating treatment in vivo [3].

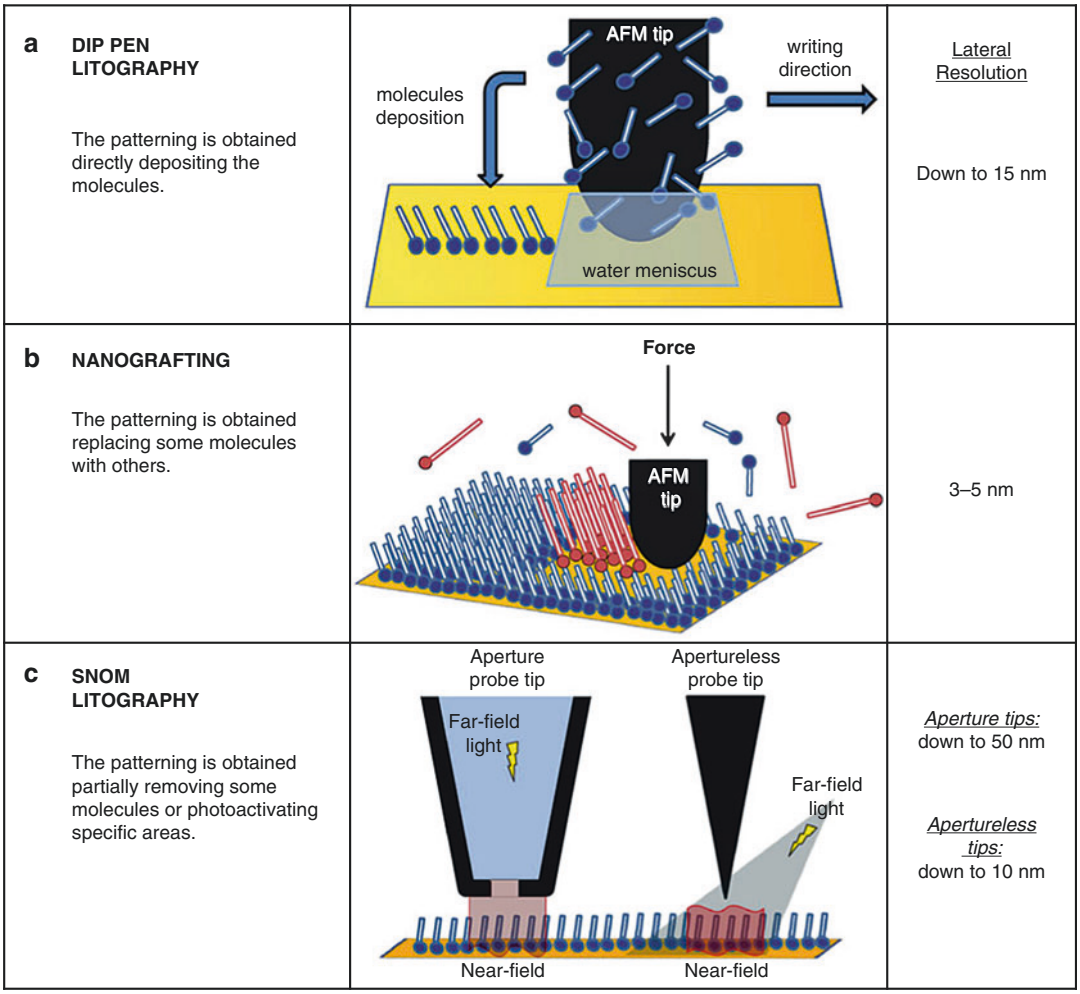
Finally, another class of nanostructured materials is often mentioned as potentially suitable for bio-applications. Hydrogels and molecular imprinted polymers can be considered either as flat nanofunctionalized surfaces or nanosystems similar to those described in this paragraph. Hydrogels are 3D high-molecular-weight networks composed of a polymer backbone, water, and a cross-linking agent. They can be stimuli sensitive, thus being of great interest in drug delivery, cell encapsulation, and tissue engineering [3]. Molecular imprinted polymers are also described as powerful nanotools for specific applications. They are tailor-made biomimetic receptors often used for functionalizing the surface of nanomaterials with sensing or separation purposes. Molecular imprinted polymers are produced by polymerization in the presence of molecular templates that are removed afterward, thus creating complementary nanostructured surfaces. These new surfaces can bind selectively the target (or template) used to prepare them [3].

Nanoscale Patterns Fabrication

Nanoscale patterns are generally quite difficult to work with, both in terms of preparation and characterization. Besides downscaling the fabrication methods originally developed for microconditions to nanoscale (such as photolithography, nanoimprint lithography), dedicated fabrication methods have been proposed for these applications. The following paragraphs describe the scanning probe–based techniques most commonly utilized to obtain functional structures at nanoscale such as nanoarrays containing different features.

Dip Pen Nanolithography

A scanning probe technique that has been developed for the production of nanoarrays is the ▶ Dip-Pen Nanolithography (DPN), first reported by Mirkin's group in 1999 [13]. DPN makes use of one or several AFM tips that are dipped into an ink and subsequently used to write a nanoscale pattern on a surface. The method was presented as a powerful technique to deposit molecules on a substrate from AFM tips at resolutions comparable to those achieved with much more expensive lithographic methods. The principle takes advantage of the water meniscus that naturally forms between the ink-coated probe tip and the substrate when the scanning was done in contact mode under ambient laboratory conditions. The ink moves on the substrate by capillary transport through the meniscus, as shown in Fig. 1, Panel A. A key point is the correct choice of the ink and the



Nanostructured Functionalized Surfaces, Fig. 1 Summary of scanning probe–based techniques for functional surface nanopatterning. A brief description of the principle, an illustrative cartoon, and the lateral resolution of each technique are reported

substrate, which have to possess reciprocal chemical affinities, thus favoring the chemical adsorption of the ink molecules onto the substrate. An useful application of this technique is related to the precise functionalization of nanoscale devices previously prepared by more conventional lithographic methods. The resolution of DPN depends on several parameters such as the grain size of the substrate, the molecule diffusion rate, the tip-substrate contact time and the scan speed, and finally on the environmental conditions such as humidity. The minimum feature sizes producible by DPN are in the range of 15 nm [8].

Nanografting

An effective way to introduce chemical and functional features with nanoscale lateral resolution down to few nanometers has been developed by Liu and coworkers starting from 1997 [14, 15] using an AFM (▶ AFM) tip. This method is based on the local modification of a self-assembled monolayer (▶ Self-Assembled Monolayers) surface, previously deposited starting from an alkanethiol solution, on a flat gold substrate. The surface is then mounted in a liquid environment AFM cell, allowing its observation and nanomanipulation with the AFM tip.

While scanning the surface in a low load regime (<1 nN) results in the three-dimensional imaging of the surface, increasing the load over a threshold value one obtains the removal of the previously deposited thiolated molecules under the tip, locally disrupting the monolayer and exposing the gold surface. If thiolated molecules (usually different from that used in the initial SAM deposition) are present in solution, a newly deposited SAM is obtained in areas where the tip is removing the old one, locally changing morphology and chemistry of the surface. Patches with different height with respect to the surrounding surface can be deposited and detected in topography AFM images. Nanografted structures can also introduce characteristic groups in specific areas that show a chemical contrast in AFM friction images. The principle is illustrated in Fig. 1, Panel B.

Nanografting can be applied even with a solution of the same molecule previously used for the initial SAM formation (homogeneous nanografting), obtaining in this case the removal of defects that were possibly present in the SAM [15]. Having obtained a nanopatterned surface endowed with specific chemical groups in well-defined patches, one can then exploit this method to create nanoarrays of biomolecules. Liu and coworkers [16] showed how to take advantage of both non-covalent and covalent attachments to form such protein nanoarrays, down to a lateral scale of tens of nanometers. This method generally leads to non-oriented protein patches, as demonstrated by the AFM analysis, that shows the presence of proteins with different heights with respect to the surface, corresponding to the different protein orientation. The nanografting technology can be also successfully applied to the realization of DNA nanoarrays, when thiolated single-stranded DNAs are used in the solution during the nanografting process. DNA lines as narrow as 10 nm can be indeed fabricated. Moreover, nanografting allows a controlled density deposition of DNA strands, depending on the number of tip passages over the same area [17]. Successful DNA hybridization has been demonstrated on such nanoarrays.

It is also possible to extend such a technique to realize protein nanoarrays: Patches of single-stranded DNAs of different sequences are deposited using nanografting in specific places of a surface previously covered with a SAM able to discourage aspecific protein adhesion. Then, proteins linked with DNA sequences that are complementary to that of the patch where the protein deposition is desired are incubated over the surface. Thanks to the DNA complementarity, a specific protein localization is then obtained. The effectiveness of this methodology for the realization of nanoscale protein sensors has been demonstrated using human serum enriched with antibodies against specific proteins that were previously immobilized on nanografted patches, showing that antibodies bind to the patches where their targets are present [18].

A method that in similar fashion takes advantage of the DNA hybridization properties to direct the localization of nanosized objects with the help of an AFM probe (although not based on nanografting) has been demonstrated by Kufer and coworkers in 2008 [19]. Their method is based on the fabrication of specific areas on a glass surface, namely, depot areas – where DNA-coupled nanoobjects are picked up by a single-stranded functionalized AFM tip, and target areas – where the nanoobjects are delivered with a precision of some nm to form the desired pattern. More than 5,000 object transfers with the same tip have been demonstrated.

SNOM Lithography

In 1983, it was proved the Scanning Near-field Optical Microscopy (SNOM), shortly after the invention of the first proximal probe technology, the Scanning Tunneling Microscope (STM). These two techniques are based on a similar principle, but in the case of STM the detection is performed on the electron tunneling currents, while in the case of SNOM the detection is performed on the tunneling of photons. This characteristic makes the SNOM suitable for the analysis of nonconductive samples [20].

By illuminating the sample with near-field light, the probe tip is held very close to the sample without touching. An evanescent light field is created in very close proximity of surface. The evanescent light through the aperture is incident on the sample, exciting the atoms in the surface to reradiate propagating waves. The same probe can

then collect the propagating light and transmit it to an appropriate detector. The probe tip can be designed in two different ways, an aperture-based tip and an apertureless-based tip (as illustrated in Fig. 1, Panel C). In an aperture-based SNOM, however, the nanometer-scale aperture limits the light throughput and the resolution. To overcome these limitations an apertureless SNOM was developed that uses an apertureless probe, similar to an STM or AFM tip, to scan within the near field of the sample. In this case the sample is first irradiated from above by far-field light in order to produce the evanescent field on the surface of the sample. The evanescent photons then excite the atoms in the probe tip, which reradiate the propagating photons.

An interesting application of this technique is related to the lithography-like process. The SNOM probe is used as an energy source for modifying a variety of materials at the nanoscale level, including conventional photoresists and non-conventional electro-magneto-optical materials as well as ferroelectric and SAM. The photoresists layer deposited on a substrate may be changed for instance by laser pulses emitted from the SNOM probe, offering resolutions unachievable by conventional optical and laser systems (well beyond the diffraction limit). Patterns are generated by scanning the probe over these target surfaces and both aperture and apertureless probe tips are used, obtaining different resolution limits (50 nm for aperture probe and down to 10 nm for the apertureless probe).

Cross-References

- ▶ AFM
- ▶ Chitosan Nanoparticles
- ▶ Dip-Pen Nanolithography
- ▶ Functionalization of Carbon Nanotubes
- ▶ Microfabricated Probe Technology
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Self-Assembled Monolayers
- ▶ Self-Assembly of Nanostructures

References

1. Liu, X., Chu, P.K., Ding, C.: Surface nano-functionalization of biomaterials. Mater. Sci. Eng. R **70**, 275–302 (2010)
2. Martínez, E., Lagunas, A., Mills, C.A., Rodríguez-Seguí, S., Estévez, M., Oberhansl, S., Comelles, J., Samitier, J.: Stem cell differentiation by functionalized micro- and nanostructured surfaces. Nanomedicine (Lond.) **4**(1), 65–82 (2009)
3. He, J., Qi, X., Miao, Y., Wu, H.L., He, N., Zhu, J.J.: Application of smart nanostructures in medicine. Nanomedicine (Lond.) **5**(7), 1129–1138 (2010)
4. Briand, E., Humblot, V., Landoulsi, J., Petronis, S., Pradier, C.M., Kasemo, B., Svedhem, S.: Chemical modifications of Au/SiO_2 template substrates for patterned biofunctional surfaces. Langmuir **27**(2), 678–685 (2011)
5. Ogaki, R., Cole, M.A., Sutherland, D.S., Kingshott, P.: Microcup arrays featuring multiple chemical regions patterned with nanoscale precision. Adv. Mater. **23**(16), 1876–1881 (2011)
6. Yim, E.K.F., Leong, K.W.: Significance of synthetic nanostructures in dictating cellular response. Nanomed. Nanotechnol. Biol. Med. **1**, 10–21 (2005)
7. Bell, B.F., Schuler, M., Tosatti, S., Textor, M., Schwartz, Z., Boyan, B.D.: Osteoblast response to titanium surfaces functionalized with extracellular matrix peptide biomimetics. Clin. Oral Impl.ants Res. **22**(8), 865–872 (2011)
8. Anselme, K., Davidson, P., Popa, A.M., Giazzon, M., Liley, M., Ploux, L.: The interaction of cells and bacteria with surfaces structured at the nanometre scale. Acta Biomater. **6**, 3824–3846 (2010)
9. Verma, A., Uzun, O., Hu, Y., Hu, Y., Han, H.S., Watson, N., Chen, S., Irvine, D.J., Stellacci, F.: Surface-structure-regulated cell-membrane penetration by monolayer-protected nanoparticles. Nat. Mater. **7**, 588–595 (2008)
10. Xing, Y., Dai, L.: Nanodiamonds for nanomedicine. Nanomedicine **4**(2), 207–218 (2009)
11. Murday, J.S., Siegel, R.W., Stein, J., Wright, J.F.: Translational nanomedicine: status assessment and opportunities. Nanomedicine **5**(3), 251–273 (2009)
12. Dutta, R.K., Sharma, P.K., Pandey, A.C.: Design and surface modification of potential luminomagnetic nanocarriers for biomedical applications. J. Nanopart. Res. **12**, 1211–1219 (2010)
13. Piner, R.D., Zhu, J., Xu, F., Hong, S., Mirkin, C.A.: "Dip-pen" nanolithography. Science **283**, 661–663 (1999)
14. Xu, S., Liu, G.: Nanometer-scale fabrication by simultaneous nanoshaving and molecular self-assembly. Langmuir **13**(2), 127–129 (1997)
15. Xu, S., Miller, S., Laibinis, P.E., Liu, G.: Fabrication of nanometer scale patterns within self-assembled monolayers by nanografting. Langmuir **15**(21), 7244–7251 (1999)

16. Wadu-Mesthrige, K., Xu, S., Amro, N.A., Liu, G.: Fabrication and imaging of nanometer-sized protein patterns. Langmuir **15**(25), 8580–8583 (1999)
17. Mirmomtaz, E., Castronovo, M., Grunwald, C., Bano, F., Scaini, D., Ensafi, A.A., Scoles, G., Casalis, L.: Quantitative study of the effect of coverage on the hybridization efficiency of surface-bound DNA nanostructures. Nanoletters **8**(12), 4134–4139 (2008)
18. Bano, F., Fruk, L., Sanavio, B., Glettenberg, M., Casalis, L., Niemeyer, C.M., Scoles, G.: Toward multiprotein nanoarrays using nanografting and DNA directed immobilization of proteins. Nanoletters **9**(7), 2614–2618 (2009)
19. Kufer, S.K., Puchner, E.M., Gumpp, H., Liedl, T., Gaub, H.E.: Single-molecule cut-and-paste surface assembly. Science **319**, 594–596 (2008)
20. Tseng, A.A.: Recent developments in nanofabrication using scanning near-field optical microscope lithography. Opt. Laser Technol. **39**, 514–526 (2007)

Nanostructured Hydrogels

▶ Nanoengineered Hydrogels for Cell Engineering

Nanostructured Materials

▶ Microscopic Modeling of Solid-State Quantum Devices
▶ Nanomaterials for Electrical Energy Storage Devices

Nanostructured Materials for Sensing

Mariangela Lombardi
IIT – Italian Institute of Technology @ POLITO – Centre for Space Human Robotics, Torino, Italy

Synonyms

Nanomaterials for sensors

Definition

Materials having nanometric dimensions are able to improve their functional properties, which allow their exploitation in highly performing sensors.

Overview

Materials with at least one dimension smaller than 100 nm are classified as nanostructured. Compared to micro-sized ones, nanomaterials are characterized by a higher surface to bulk ratio, usually leading to an enhancement of optical, mechanical, electrical, structural, and magnetic properties [1].

In the last few years, research focused its efforts on the development of multifunctional composite materials, able to perform multiple non-structural functions, sometimes combined with a suitable mechanical response. If nanostructured, they can also imply additional functional properties with respect to conventional composites. Multifunctional materials can be used for developing highly performing sensors, particularly in the case of single-phased and composite materials containing carbon nanotubes and piezoelectric phases [2].

Piezoelectric Materials

Piezoelectric materials are able to directly convert a mechanical signal into an electrical one and vice versa. For this reason piezoelectric materials produce a charge when subjected to a stress (direct effect) or a strain when an electric field is applied (converse effect). These materials can then be used as actuators or sensors, by exploiting their converse or direct piezoelectric effect, respectively. Generally piezoelectric sensors are used for measuring several parameters, like acceleration, angular rate, pressure, force, temperature, mass, magnetic field, and amount of several chemical substances [3].

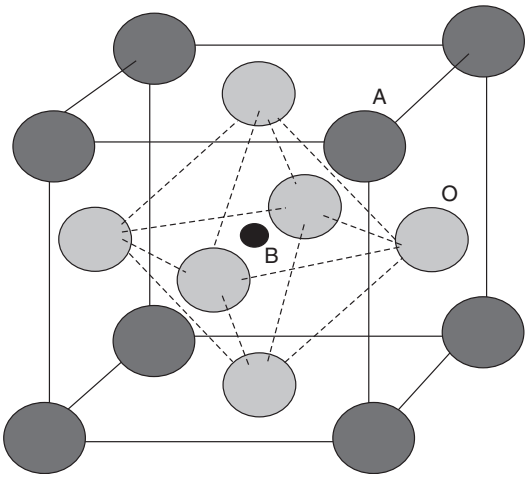
Piezoelectricity can be expressed by several materials, such as single crystals, ceramics, polymers, and composites [4, 5]. Even if they naturally

present this behavior, piezoelectric single crystals, as quartz and Rochelle salt, cannot be easily used being isotropic, e.g., they have different properties depending on the cut and the orientation directions. Most piezoelectric ceramics applied in commercial sensing devices are made of polycrystalline ferroelectric ceramics [6], which present a higher strength and an easier fabrication with respect to single crystals [4]. Ferroelectrics are characterized by a spontaneous polarization which can be orientated under a realizable electric field. In fact, they initially contain randomly oriented, polarized regions, named domains. These polarized regions form upon cooling at a temperature called the Curie temperature to minimize the total elastic energy of the ceramic structure. In this state, they do not present piezoelectric properties. When an electric field is applied, the domains align; this process is known as poling. When the electric field is removed, most domains remain aligned and piezoelectricity is induced.

Generally the piezoelectric ceramics present a composition that implies a structural instability such as polymorphic phase transition (PPT) or morphotropic phase boundary (MBP) [7]. PPT-based systems are characterized by a large temperature dependence of piezoelectric properties and a rapid depoling, due to temperature changes close to PPT temperature. In the case of MBP-based materials, piezoelectric features are slightly influenced by temperature, but their working temperatures approach a half of Curie temperature.

Among polycrystalline ferroelectric ceramics, the first system studied for its piezoelectric properties was barium titanate ($BaTiO_3$) in 1943 and it remained the primary electroceramic material until the discovery of the lead zirconate titanate or PZT ($PbZr_xTi_{1-x}O_3$) in 1954 [4]. The former is a PPT-based, high-performance piezoelectric ceramic, the latter presents a MBP from tetragonal to rhombohedral phase at about 48 mol% $PbTiO_3$ content.

For a long time PZT was the most used piezoceramic for its very attractive behavior in sensor applications [5, 7]. Notwithstanding this, in 2003 the European Union claimed PZT as a hazardous substance and research efforts were



Nanostructured Materials for Sensing, Fig. 1 The perovskite structure common in most ceramic piezoelectric materials (With permission from Ref. [5])

then focused on lead-free piezoceramics having high performances. For this reason, in the last decade, alternative systems were deeply studied modifying the composition of the traditional ferroelectric ceramics. In particular, this *compositional engineering approach* involved perovskite-based materials, exploiting their solid solutions with other ferroelectric systems or their doping, e.g., the introduction of different types of cations within their structure. The perovskite structure ABO_3 consists in a face-centered cubic array, in which the large cations A are placed at the corners of the unit cell, the smaller cations B in the body center, and the oxygen atoms in the center of each face, as illustrated in Fig. 1 [5–7].

Even if the high symmetry of the perovskite structure could lead to improved poling and consequently higher piezoelectric performances, currently no innovative lead-free ceramics are able to compare the piezoelectric and thermal behavior of PZT-based materials.

On the other hand, a *structural engineering approach*, based on template grain growth, grain size optimization, and domain engineering, is able to enhance the electromechanical performances of existing piezoelectric materials. For instance, in the case of $BaTiO_3$, a microstructural refinement can double the piezoelectric coefficient. In this case, the finer grain size implies a decrease of

the domain size and an increase in density of domain walls, which are able to influence the piezoelectric properties [5, 8]. Moreover, a fourfold increase of the piezoelectric coefficient can be reached synergically combining the effect of submicrometric domain sizes and grain orientation obtained by template grain growth technique. The positive effect of the grain orientation was also demonstrated in the binary system $(Bi_{1/2}K_{1/2})TiO_3$–$BaTiO_3$ (BKT–BT) [5].

Generally piezoelectric ceramics can be applied in sensors as bulk components or thin films, by using micromachining techniques [2, 3]. Thin films from one tenth to several tens of microns in thickness can present different properties with respect to the corresponding bulk material. This is due to the dependence of the piezoelectric features to the morphology and density of the film, as well as to the compositional aspects related to the stoichiometry and presence of impurities. Moreover, the piezoelectric behavior can also be influenced by the orientation and nature of the substrate material. All the deposition methods are based on three fundamental steps; firstly, the atomic or molecular species are yielded by chemical or physical processes. In the former case chemical vapor deposition or sol–gel techniques can be used; in the latter one, sputtering or pulsed laser deposition can be employed. After this step, deposition on the substrate is achieved through transport and condensation mechanisms and finally, the film is obtained by crystallization and annealing [3].

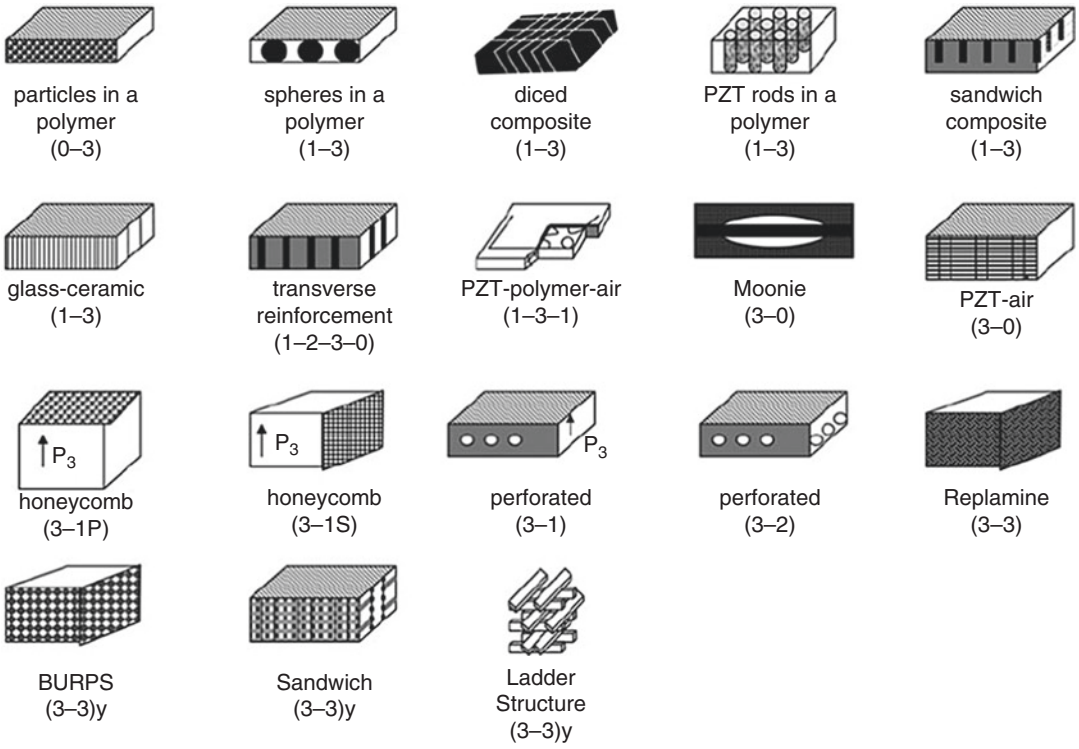
In order to achieve a long-term stability of piezoelectricity at room temperature and to develop flexible devices, piezoelectric polymers can be employed [9]. Generally ferroelectric materials based on polyvinylidene fluoride (PVDF), odd nylon, vinylidene cyanide (VDCN), and aromatic and aliphatic polyurea families are exploited [9, 10]. In the case of PVDF-based polymers, the residual polarization after poling is due to the alignment of CF_2 dipoles present in their structure $(CH_2–CF_2)_n$ and trapped charges. Dipoles are oriented in the crystalline phases inside lamellae and in the interfaces between crystalline and noncrystalline phases, while the charge traps are present at the surface of crystallites [10]. The electromechanical performances of these polymeric materials can be improved by the polymerization of PVDF with other fluorinated polymers, which implies an increase of the crystallinity [9]. In the case of odd-nylon and polyurea families the field-induced orientation of N–H and C–O dipoles is stabilized by the hydrogen bonding present between adjacent polymer chains. On the other hand, in the case of amorphous polymers such as VDCN-based materials, an easier orientation of the C–CN dipoles is obtained by poling at temperatures near to the glass transition temperature and freezing the residual polarization, through the cooling down of the system under the electric field [9, 10].

Among piezoelectric polymers, PVDF presents the best electromechanical behavior at room temperature, but its performances decrease at higher temperature [9]. This drawback can be partially limited through the copolymerization of PVDF [11]. Other ferroelectric polymers can be used, but they are characterized by less piezoelectric properties [9, 11].

Compared to the pure polymer performances, piezoelectric composites offer significantly higher functional properties. It is in fact possible to synergically couple the best features of the ceramic and polymeric phases while minimizing the poorest properties [4, 11]. Composites can be classified on the basis of the connectivity between the constituent phases according to Newnham notation [4]. This notation describes the number of dimensions each phase is physically in contact with itself. When just two components are considered, ten configurations are possible: 0-0, 1-0, 2-0, 3-0, 1-1, 2-1, 3-1, 2-2, 3-2, or 3-3, illustrated in Fig. 2 [5].

Among them, 0-3, 1-3, and 2-2 connectivity have been recently developed, in order to tailor the piezoelectric response through the exploitation of anisotropic composites having different ceramic phase contents [11].

In the last few years, miniaturization of devices based on piezoelectric behavior involves the exploitation of carbon nanotubes as functional materials [12].



Nanostructured Materials for Sensing, Fig. 2 Possible connectivity in bicomponent piezoelectric composites (With permission of Ref. [5])

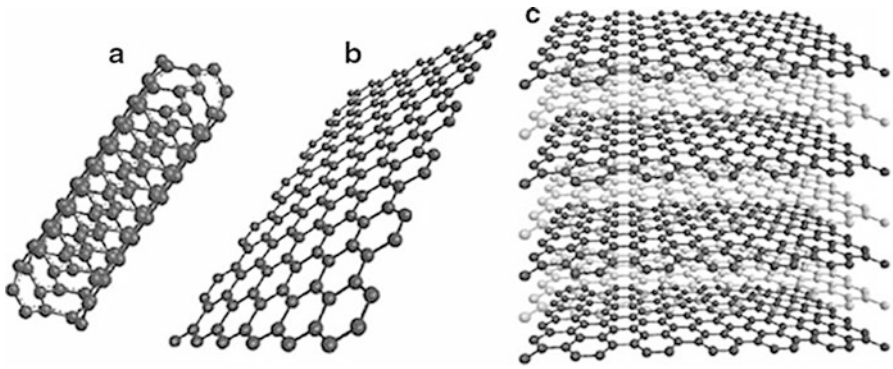
N

Carbon Nanomaterials

Multidimensional carbon nanomaterials include graphite, graphene, and carbon nanotubes. They are made of carbon atoms in an exagonal configuration, bonded each other with sp^2 bonds [13]. Graphene is a two-dimensional sheet of carbon atoms and represents the repetitive unit of the graphite. Carbon nanotubes (CNT) consist in graphene sheets rolled to form cilynders, as shown in Fig. 3. In particular, carbon nanotubes can be classified as single-walled (SWCNT) or multi-walled (MWCNT); the former, consisting in a single graphene cylinder, the latter, being made of concentric cylinders with an interlayer spacing of 0.34 nm. CNT nanostructure, atomic arrangement, diameter, and length strongly affect their mechanical, electric, and thermal properties [12].

Carbon nanotubes show an electromechanical behavior that, unlike to piezoelectric ceramics, is due to the phonon frequency shifts as a consequence of a charge injection [12]. This feature is mainly exploited in carbon nanotubes composites, which are characterized by a change in resistance when subjected to a chemical, thermal, or mechanical loading. Among them, the most promising sensing devices based on carbon nanotubes are the chemiresistive-type ones. In fact, their high sensitivity, small size, high surface area, and high aspect ratio imply the ability to detect smaller concentrations of analyte than traditional sensors.

In comparison to carbon nanotubes, the homogenous distribution of electrochemically active sites at a nanometer scale and the higher purity could make graphene an ideal material for sensing and detection. Notwithstanding this, no consolidated studies have been carried out about its application in sensing devices to date [14].



Nanostructured Materials for Sensing, Fig. 3 Structure of the multidimensional carbon nanomaterials: (**a**) CNT, (**b**) graphene, and (**c**) graphite (With permission of Ref. [13])

Cross-References

- ▶ Active Carbon Nanotube-Polymer Composites
- ▶ Carbon Nanotubes
- ▶ Graphene
- ▶ NEMS Piezoelectric Switches
- ▶ Piezoelectric Effect at Nanoscale
- ▶ Piezoelectric MEMS Switch
- ▶ Sol–gel Method
- ▶ Synthesis of Carbon Nanotubes
- ▶ Synthesis of Graphene

References

1. Fauchais, P., Montavon, G., Lima, R.S., Marple, B.R.: Engineering a new class of thermal spray nano-based microstructures from agglomerated nanostructured particles, suspensions and solutions: an invited review. J. Phys. D Appl. Phys. **44**(53), 093001 (2011)
2. Gibson, R.F.: A review of recent research on mechanics of multifunctional composite materials and structures. Compos. Struct. **92**, 2793–2810 (2010)
3. Tadigadapa, M., Mateti, K.: Piezoelectric MEMS sensors: state-of-the-art and perspectives. Meas. Sci. Technol. **20**(30), 092001 (2009)
4. Tressler, J.F., Alkoy, S., Newnham, R.E.: Piezoelectric sensors and sensor materials. J. Electroceram. **2**, 257–272 (1998)
5. Cook-Chennault, K.A., Thambi, N., Sastry, A.M.: Powering MEMS portable devices-a review of non-regenerative and regenerative power supply systems with special emphasis on piezoelectric energy harvesting systems. Smart Mater. Struct. **17**(33), 043001 (2008)
6. Rodel, J., Jo, W., Seifert, K.T.P., Anton, E.M., Granzow, T., Damjanovic, D.: Perspective on the development of lead-free piezoceramics. J. Am. Ceram. Soc. **92**, 1153–1177 (2009)
7. Leontsev, S.O., Eitel, R.E.: Progress in engineering high strain lead-free piezoelectric ceramics. Sci. Technol. Adv. Mater. **11**(13), 044302 (2010)
8. Setter, N., Damjanovic, D., Eng, L., Fox, G., Gevorgian, S., Hong, S., Kingon, A., Kohlstedt, H., Park, N.Y., Stephenson, G.B., Stolitchnov, I., Taganstev, A.K., Taylor, D.V., Yamada, T., Streiffer, S.: Ferroelectric thin films: review of materials, properties, and applications. J. Appl. Phys. **100**(46), 051606 (2006)
9. Eberle, G., Schmidt, H., Eisenmenger, W.: Piezoelectric polymer electrets. IEEE Trans. Dielect. Elect. Insul. **3**, 624–646 (1996)
10. Fukada, E.: History and recent progress in piezoelectric polymers. IEEE T. Ultrason. Ferr. **47**, 1277–1290 (2000)
11. Lushcheikin, G.A.: New polymer-containing piezoelectric materials. Phys. Solid State **48**, 1023–1025 (2006)
12. Li, C., Thostenson, E.T., Chou, T.W.: Sensors and actuators based on carbon nanotubes and their composites: a review. Compos. Sci. Technol. **68**, 1227–1249 (2008)
13. Pumera, M., Ambrosi, A., Bonanni, A., Chng, E.L.K., Poh, H.L.: Graphene for electrochemical sensing and biosensing. Trac-Trend. Anal. Chem. **29**, 954–965 (2010)
14. Lü, P., Feng, Y.Y., Zhang, X.Q., Li, Y., Feng, W.: Recent progresses in application of functionalized graphene sheets. Sci. China Tech. Sci. **53**, 2311–2319 (2010)

Nanostructured SERS-Active Metal Oxides

▶ Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy

Nanostructured Solar Cells

▶ Nanomaterials for Excitonic Solar Cells

Nanostructured Thermoelectric Materials

Yufei Liu and Jian He
Department of Physics and Astronomy, Clemson University, Clemson, SC, USA

Synonyms

Thermoelectric nanomaterials and nanostructured thermoelectrics

Definition

Nanostructured thermoelectric materials are materials that contain nanoscale constituents and exhibit enhanced thermoelectric performance due to nanoscale phenomena.

Introduction

Electricity and heat are two important forms of energy; electricity remains the most convenient form of energy, while heat is an abundant but low-grade form of energy. As the simplest technology applicable to direct heat–electricity energy conversion, *thermoelectricity* is secured a position in the solution package of sustainable and alternative energy for mankind in the twenty-first century [1]. The study of thermoelectrics is an application-driven material-oriented fundamental research; the ultimate goal is to develop higher performance thermoelectric material for applications in power generation and heat management. The thermoelectric performance of a material is gauged by its figure of merit, ZT with $ZT = \frac{\sigma\alpha^2 T}{\kappa} = \frac{PF}{\kappa} = \frac{PF}{(\kappa_{ph}+\kappa_e)}$, σ is the electrical conductivity, α is the thermopower (i.e., Seebeck coefficient), κ is the total thermal conductivity, PF is the power factor, and T is the temperature in Kelvin. The κ is often expressed as the sum of the lattice thermal conductivity, κ_{ph}, and the carrier thermal conductivity, κ_e. ZT is the material parameter of prime importance in thermoelectrics research. A high ZT requires a high PF and a low κ_{ph}, succinctly phrased as "electron-crystal phonon-glass" (ECPG).

To date, degenerate semiconductors such as BiSb, Bi_2Te_3, SiGe, PbTe, $CoSb_3$-based skutterudites, and half-Heusler (HH) compounds remain the cornerstone for the state-of-the-art thermoelectric materials. These materials possess several common features [2]: (i) the bandgap size *on the order of* $10\,k_B T_o$, where k_B is the Boltzmann constant and T_o is the operation temperature; (ii) the carrier concentration *on the order of* 10^{18}–10^{20} cm^{-3}; (iii) high-symmetry crystal structure, either cubic or hexagonal (rhombohedral); (iv) small electronegativity difference among constituent elements; (v) low Debye temperature and/or strong anharmonicity. The first three features aim at a subtle balance among the carrier concentration, mobility, effective mass, and the bipolar effect, in line with "electron crystal." The last feature aims at a low κ_{ph}, in line with "phonon glass." Though initially developed for bulk thermoelectric materials, these criteria, *in principle*, hold for nanostructured thermoelectric materials.

In the past decade, the synergy of nanotechnology and traditional bulk thermoelectric material has culminated into a new paradigm, "nanostructured thermoelectric material" [3]. The major rationale behind this new paradigm lies in that *thermoelectricity* arises from *two* fundamental flows, namely, the charge flow and the entropy flow. The characteristic length scales of these two flows, namely, the mean free path and the wavelength of electrons and phonons, are on the order of *nanometers*. Hence, the nanostructuring process provides an extra layer of control of the charge and entropy flows beyond the atomic scale; one can resort to the classical size effects to optimize the mean free path of electrons and phonons while utilizing the quantum size effects

to optimize the electron band structure and phonon dispersion relations.

Recent Advances in Nanostructured Thermoelectric Materials

The main theme of this entry is on the latest advances in nanostructured thermoelectric material research. In general, advance can be achieved via (i) suppressing the lattice thermal conductivity in line with the "phonon-glass" approach and/or (ii) enhancing the *PF* in line with the "electron-crystal" approach. To date, most advances are achieved in line with the "phonon-glass" approach; *ZT* is almost doubled in several nanostructured thermoelectric materials due to significantly reduced κ_{ph} [3]. In contrast, less progress was made in enhancing the *PF*. We are now at a *crossroad*: the room for further reduction of κ_{ph} is getting marginal. Though it was proved that the κ_{ph} could be reduced below the *alloying limit* in specific materials [4], the procedure is hard to scale up. As such, the future of nanostructured thermoelectric material research hinges upon the successful focus shift from "phonon glass" to "electron crystal."

Next, we will briefly discuss two promising mechanisms that enhance the *PF* in nanostructured thermoelectric material, namely, (i) the electron energy-filtering mechanism and (ii) the "invisible dopants" mechanism. Despite different nomenclatures, these mechanisms are inherently related: the electron energy-filtering mechanism encompasses both *resonant* scattering and *nonresonant* scattering; the *invisible dopants* mechanism is a unique type of *resonant* scattering (namely, *antiresonant scattering*). In practice, the electron energy-filtering scheme is often implemented at interfaces; this is how "filtering" initially got its name.

Electron Energy-Filtering Mechanism

In this section, we use the Drude model and the Mott relation to help illustrate the rationales behind the electron energy-filtering scheme. Despite the assumptions and oversimplifications used, the conclusion drawn is general. The electron energy-filtering scheme was initially proposed for thermionics, the other direct heat–electricity conversion technology. The major difference between thermionics and thermoelectrics is that the electron transport in a thermionic process is ballistic across the energy barrier, while the electron transport in a thermoelectric process is quasi-equilibrium and diffusive [5].

In a quasi-equilibrium and diffusive electron transport process, it is instructive to relate the energy-specific electrical conductivity $\sigma(E)$ and the thermopower α in the Drude model (Eq. 1) and also in the differential form of the Mott relation (Eq. 2) [6]:

$$\sigma(E) = n(E)e\mu(E) = n(E)e^2\frac{\tau(E)}{m^*}\ , \qquad (1)$$

where the energy-specific carrier concentration $n(E)$ is defined as the number of carriers between energy E *and* $E + \Delta E$, $\mu(E)$ is the energy-specific mobility, $\tau(E)$ is the relaxation time, and m^* is the effective mass.

$$\alpha = \frac{k_B}{q}\int_0^\infty \sigma(E)\left(\frac{E-E_F}{k_BT}\right)\left(\frac{\partial f_0(E)}{\partial E}\right)dE \Big/ \int_0^\infty \sigma(E)\left(-\frac{\partial f_0(E)}{\partial E}\right)dE, \qquad (2)$$

where q is the elementary charge, k_B is the Boltzmann constant, E_F is the Fermi energy, and $f_0(E)$ is the Fermi–Dirac distribution function. The "Fermi window factor" $\left(\frac{\partial f_0(E)}{\partial E}\right)$ is, *to the first order*, a bell curve centered at $E = E_F$ and with a narrow width on the order of k_BT. Therefore, α is, *to the first order*, proportional to the mean excess energy $<E-E_F>$. At a given carrier concentration, the higher the $<E-E_F>$, the higher the α. In total, a high *PF*, thus, necessitates a *large* and *energy-asymmetric* $\sigma(E)$ within the Fermi window (i.e., a large and asymmetric electron density of states $g(E)$ near E_F).

Furthermore, one may rewrite Eq. 2 for a degenerate single-band electron system [7]:

$$\alpha = \frac{\pi^2}{3}\frac{k_B}{q}(k_B T)\left[\frac{g(E)}{n(E)} + \frac{1}{\mu(E)}\frac{d\mu(E)}{dE}\right] \quad (3)$$

Note that each of the *two* terms in the square bracket of Eq. 3 corresponds to a mechanism that governs α, the central component of the *PF*. While the electron band structure governs the $g(E)/n(E)$ term, the $\mu(E)$-containing term embodies the detailed form of energy-dependent scattering. The electron energy-filtering scheme aims at the $\mu(E)$-containing term.

The electron energy-filtering scheme ensures a higher α via filtering out those electrons with lower mean excess energy $<E-E_F>$; moreover, the reduction in σ is compensated by the increase in α, leading to a higher *PF*. More importantly, the stronger the energy dependence of scattering, the higher the *PF*, provided the mobility degradation is not large. With the relaxation time approximation [8], one may invoke the Matthiessen's rule to take into account the contribution from various energy-dependent scattering mechanisms:

$$\frac{1}{\tau} = \sum \frac{1}{\tau_i} \quad (4)$$

where τ_i is the relaxation time for the *ith* scattering channels and τ is the total relaxation time.

For *nonresonant* scattering mechanisms, it is customary to assume a power law $\tau(E) \sim E^r$; r is called the scattering parameter or the scattering exponent [8]. Specifically, $r = -1/2$ for electrons scattered by acoustic phonons, $r = 3/2$ for charged impurities, and $r = 0$ for charge neutral imperfections. These characteristic *nonresonant* scattering exponents are widely used in analyzing the mobility data to extract the dominant scattering mechanism. In view of Eq. 3, a larger magnitude of r, i.e., a stronger energy dependence, is thermoelectrically beneficial as long as μ is not degraded too much. In Pb or Ag nanoparticles embedded in a PbTe bulk host matrix, Nernst coefficient measurements showed that the scattering parameter r changed from $0.2 < r < 0.7$ to $r > 3$ upon nanocompositing, indicating an electron energy-filtering effect, though $r > 3$ doesn't correspond to any known scattering mechanism [9].

The scattering mechanism can be *resonant*. By *resonant* we mean the scattering is so strongly energy dependent and even diverging at specific energies that the aforementioned power law expression doesn't hold any longer. Though *resonant* scattering has several key advantages over *nonresonant* scattering, there are more restrictions and challenges in material chemistry to implement the *resonant scattering* mechanism than the *nonresonant* scattering mechanism. More discussion on the *resonant scattering* mechanism will be provided in section "Invisible Dopants" Mechanism.

The electron energy-filtering mechanism is often implemented on the interfaces. Nanostructured thermoelectric material is featured by a multitude of interfaces (grain boundaries); the interfacial effects will be more pronounced than conventional polycrystalline materials. To enhance the *PF*, interfaces must undertake multi-tier preferential scattering:

(i) Tier I: Interfaces should scatter phonons more effectively *than* electrons.
(ii) Tier II: Interfaces should scatter those carriers with lower mean excess energy more effectively *than* those with higher mean energy.
(iii) Tier III: Interfacial charged defects should scatter minority carriers more effectively *than* majority carriers.

"Invisible Dopants" Mechanism

The above discussion on the electron-filtering mechanism highlights the significance of energy-dependent scattering in thermoelectricity. In general, a stronger energy dependence tends to give rise to a higher α (cf. Eq. 3) and often a higher *PF*, provided the μ is not much degraded. Among the *nonresonant* scattering mechanisms, charged impurity scattering provides the strongest energy dependence $r = 3/2$; therefore, hetero-electron doping has been a prominent approach in thermoelectric material research for decades. However, one should note the *dual* role of such a hetero-electron dopant: it is an effective charge carrier donor and also a strong carrier scatterer. To mitigate this dilemma, people have developed

"modulation-doping" or "zonal-doping" method [10], in which the charge carriers are "separated" from the donors via, say, embedding heavily doped nanoparticles in a high-mobility host matrix and properly adjusting their band offset.

In this vein, attaining a high *PF* is thus broken down into specific tasks as (i) to introduce charge carriers, (ii) to maintain a strong energy dependence of scattering, and (iii) to minimize the mobility degradation. To this end, the resonance energy levels mechanism [11] has proved very effective in several bulk thermoelectric materials. The resonance energy levels induce narrow peak in $g(E)$ that centers at the resonant energy E_D and often adopts a Lorentzian line shape:

$$\tau_r(E) = \tau_0\left(1 + \left(\frac{E - E_D}{\Gamma/2}\right)^2\right) \qquad (5)$$

where τ_0 is the minimum value of $\tau_r(E)$ reached at E_D. Equation 5 can be incorporated into Eqs. 1 and 4 for in-depth data analysis.

Importantly, the resonant energy levels *not only* scatter charge carriers *but also* conduct current. This unique feature arises from the wave nature of electron. In quantum mechanical partial-wave analysis, the resonant scattering mechanism is described in terms of the phase angle shift δ_l of lth partial wave caused by a potential, where l is the orbital angular momentum quantum number [12]. Hence, *resonant scattering* mechanism is a "dephasing process."

The "invisible dopants" mechanism is also known as *antiresonant scattering* [13]. In an *antiresonant scattering* process, the phase angle shift is minimized by properly designing the potential profile. This interesting proposition by Zabarjadi et al. combines the *resonant scattering* mechanism and the modulation-doping method [10], thus easing the interdependence of the carrier concentration, the energy dependence of scattering. and the mobility.

The *invisible dopant* scheme is viable to be implemented in host matrix–nanoparticle composites. Specifically, the carrier-nanoparticle scattering cross section may exhibit a sharp dip (minimum) versus energy near E_F, making the nanoparticle *invisible* to conduction electrons. The most prominent consequence is that the mobility is retained, despite the numerous grain boundaries in nanostructured thermoelectric material. Second, the fast-varying scattering cross section with energy in the vicinity of E_F is favorable for a high α. One example of implementing the "invisible dopants" mechanism is to embed spherically symmetric core–shell particles of appropriate size, effective mass, and band offset inside a semiconductor host matrix (Fig. 1).

Several Examples

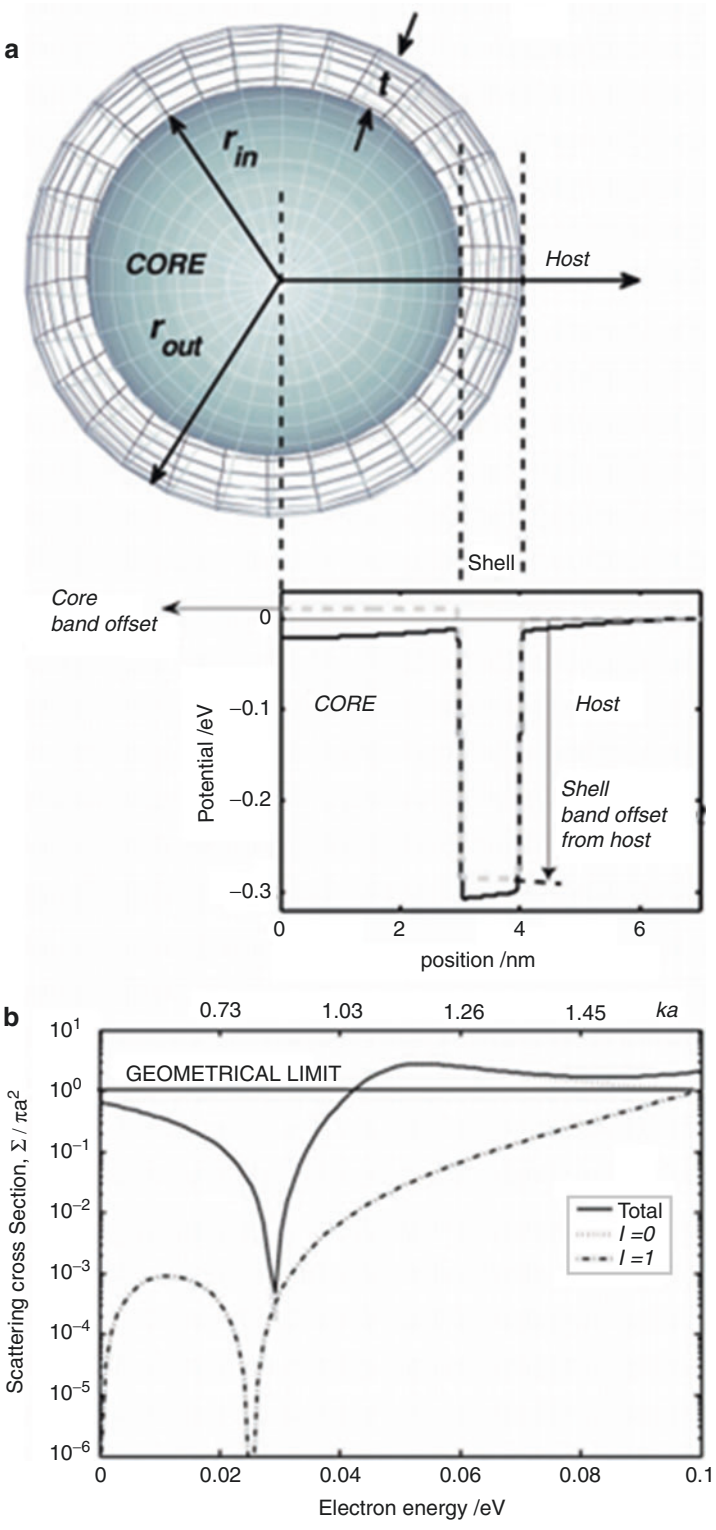
In this section, we will briefly introduce several state-of-the-art nanostructured thermoelectric materials, in which the *PF* is enhanced upon nanostructuring.

Half-Heusler-Based Nanostructured Thermoelectric Materials

Makongo et al. observed the simultaneous increases in the Seebeck coefficient and electrical conductivity of (Zr,Hf)NiSn half-Heusler compounds upon (Zr,Hf)Ni_2Sn full-Heusler nanoinclusions [14]. The 120 % enhancement of *PF* has been attributed to electron energy-filtering effect (Fig. 2). Xie et al. attributed an observed simultaneous improvement of *all* three thermoelectric properties of *p*-type Ti(Co,Fe)Sb-inSb nanocomposites to the combined interfacial effects [15].

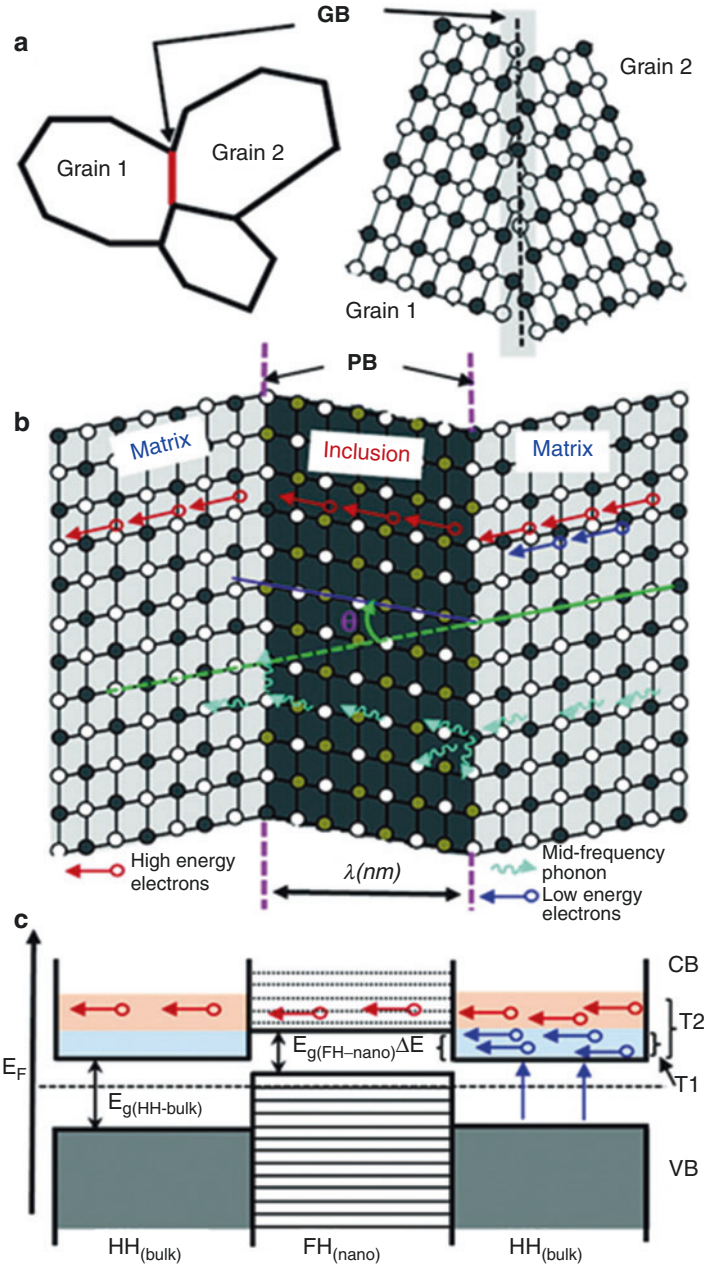
$CoSb_3$-Based Nanostructured Thermoelectric Materials

Tan et al. reported an improved *PF* and *ZT* in $CeFe_{4-x}Zn_xSb_{12}$-ZnSb nanocomposites [16]. Interestingly, the presence of ZnSb nanophase affects the crystal chemistry of the host matrix; Zn substitutionally dopes the Fe site, thus increasing the effective mass and the Seebeck coefficient. These results revealed an important aspect of those in situ formed nanostructured thermoelectric materials in which the crystal chemistry of the constituent phase is interdependent. While this constitutes a challenge in phase component control in sample synthesis, it provides a new layer of tunability in thermoelectric research.



Nanostructured Thermoelectric Materials, Fig. 1 (**a**) A core–shell structured spherical nanoparticle. The radial potential profile of the nanoparticle is plotted. The band offset profile across the core–shell nanoparticle is plotted with the dashed line and the screened bent potential with the *solid line*. (**b**) The total conduction electron–nanoparticle scattering cross section versus energy (*bottom scale*) or ka (*top scale*) is depicted with a *solid line. k* is the wave vector; a is the particle radius. The contribution of the zeroth order (*dashed line*) and the first order (*dotted line*) partial waves are displayed. The volume fraction of the nanoparticles is 1 %; the doping level is 5×10^{16} cm^{-3}, and this doping level corresponds to about 1 electron per nanoparticle (Taken from Ref. [13])

N



Nanostructured Thermoelectric Materials,
Fig. 2 Schematic illustration of atomic scale structural engineering that improves the thermoelectric performance of bulk HH matrices. (**a**) Incoherent grain boundaries (*GB*); (**b**) coherent phase boundaries (*PB*) between matrix and inclusion phases with similar structure, which allow close crystallographic registry between both regions; (**c**) heterojunction between bulk HH matrix and nm FH inclusion highlighting the electron energy-filtering mechanism. The conduction band offset, ΔE, is a critical parameter for simultaneous large increase in both Seebeck coefficient and electrical conductivity through filtering of low energy electrons. *Cyan* and *orange* color represents the distribution of low and high energy electrons at temperatures T1 and T2 (T2 > T1), respectively (Modified from Ref. [14])

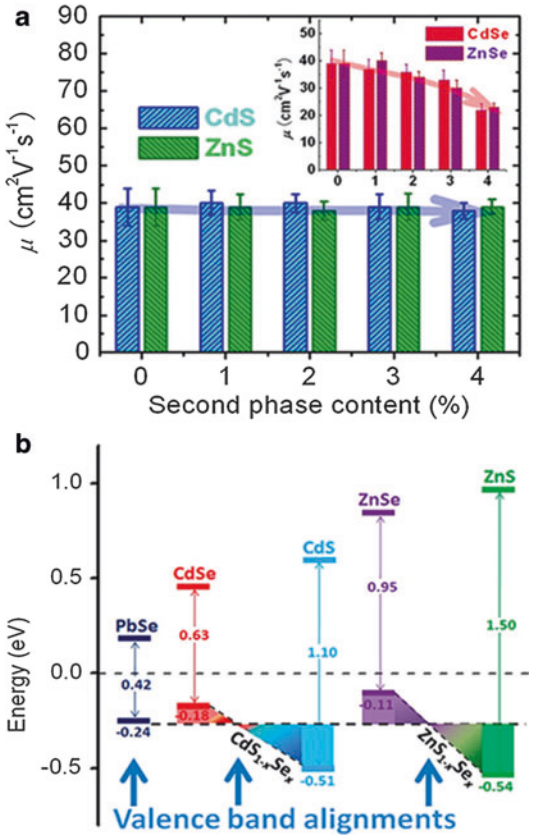
Bi_2Te_3-Based Nanostructured Thermoelectric Materials

Recent studies suggested that the interfacial charged defects *not only* tune the carrier concentration within the grain (and thus suppress the detrimental *bipolar* effect) *but also* act as potential barriers that scatter minority carriers more effectively than majority carriers. This was *inferred* from the observations in the ball-milled hot-pressed *p*-type Bi_2Te [17] and the chemically exfoliated spark-plasma-sintered *n*-type Bi_2Te_3 [18]. Concerning the effect on the electrical transport, interfacial charge defects are *not* a simple superposition of a charged point ("zero-

dimensional") defect and a charge neutral ("two-dimensional") grain boundary.

PbTe/PbSe-Based Nanostructured Thermoelectric Materials

Zhao et al. reported that the valence band alignment between PbSe and compositionally alloyed nanostructures of $CdS_{1-x}Se_x/ZnS_{1-x}Se_x$ plays an important role in the enhancement of *PF* (Fig. 3) [19]. A coherent or semi-coherent (epitaxial like, cf. Fig. 2) grain boundary is preferred to retain electrical properties, especially the mobility, while the total interfacial area per volume governs the lattice thermal conductivity [20].



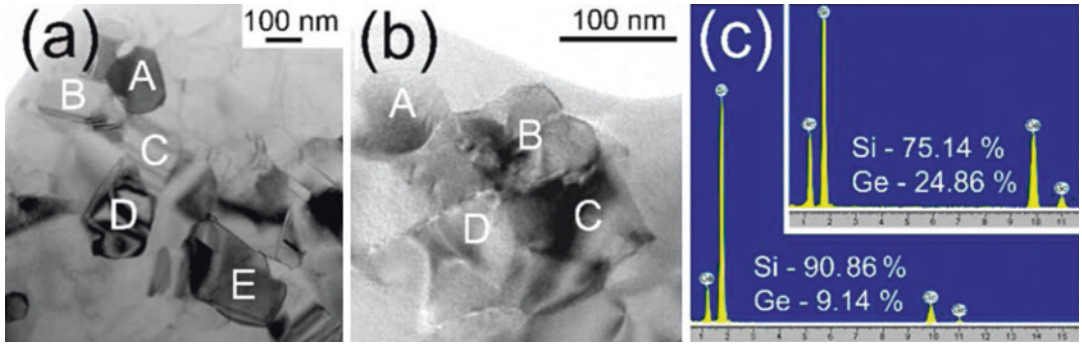
Nanostructured Thermoelectric Materials, Fig. 3 Carrier mobility and valence band alignments of $Pb_{0.98}Na_{0.02}Se$ nanocomposites. (**a**) Carrier mobility as a function of nanophase content with a similar carrier concentration of $\sim 3 \times 10^{20}$ cm^{-3}, the impact of CdS and ZnS on the carrier mobility is negligible. (**b**) The density functional theory calculations show the band alignments of the conduction band (*CB*) and valence band (*VB*) energy levels for PbSe, CdS, CdSe, ZnS, and ZnSe. Notably, CdS_{1-x} Se_x and $ZnS_{1-x}Se_x$ solid solutions show an energy level "bridging" (Modified from Ref. [19])

SiGe-Based Nanostructured Thermoelectric Materials

Yu et al. implemented modulation doping to balance the carrier concentration and the mobility in SiGe-based nanocomposites; the resulting *PF* and *ZT* are significantly improved [10]. Figure 4 shows the TEM images and EDS spectra of modulation-doped sample with a 35 % nanoparticle molar fraction. Proper band alignment between the bulk matrix and the nanoparticles is crucial in implementing the modulation-doping scheme in a three-dimensional thermoelectric material as it ensures the carriers irreversibly flow from the nanoparticles into the bulk matrix. Importantly, the interfacial effects, say, dangling bonds, impurities, and trapped charges may lead to undesired effects on proper band alignment. The charge transfer between the bulk matrix and the nanophase makes them not the same phase as they stay separate.

Perspective Remarks

The recent progress in nanostructured thermoelectric materials has significantly increased the figure of merit. In a broader view and in the long run, the synergy of nanotechnology and traditional bulk thermoelectric materials will lead to more novel ideas and new material preparation/fabrication methods regarding the power factor enhancement. In the mean time, the progress in characterization techniques will allow for a better understanding of the charge and entropy flow and their interplay on the nanometer scale. In conjunction with the arsenal, we have accumulated over the past decade for suppressing the lattice thermal conductivity, this new knowledge will put thermoelectrics research in an unprecedented position to bring the performance of thermoelectric material and the range of thermoelectric application to the next level.



Nanostructured Thermoelectric Materials, Fig. 4 (**a**, **b**) TEM images and (**c**) EDS spectra [the spectrum in (**c**) is for the grain E in (**a**), the inset in (**c**) is for the grain A in (**b**)] of modulation-doped sample with a 35 % nanoparticle molar fraction (Taken from Ref. [10])

Cross-References

- ▶ Nanostructured Thermoelectric Materials
- ▶ Nanostructures for Energy
- ▶ Thermal Conductivity and Phonon Transport

References

1. Nolas, G.S., Sharp, J., Goldsmid, H.J.: Thermoelectrics Basic Principles and New Materials Developments. Springer, Berlin/Heidelberg (2001)
2. He, J., Tritt, T.M.: Thermal-to-electrical energy conversion from the nanotechnology perspective. In: Garcia-Martinez, J. (ed.) Nanotechnology for the Energy Challenge, pp. 47–78. Wiley-VCH, Weinheim (2010)
3. Vineis, C.J., Shakouri, A., Majumdar, A., Kanatzidis, M.G.: Nanostructured thermoelectrics: big efficiency gains from small features. Adv. Mater. **22**, 3970–3980 (2010)
4. Chiritescu, C., Cahill, D.G., Nguyen, N., Johnson, D., Bodapati, A., Keblinski, P., Zschack, P.: Ultralow thermal conductivity in disordered, layered WSe_2 crystals. Science **315**, 351–353 (2007)
5. Mahan, G.D.: Thermionic refrigeration. J. Appl. Phys. **76**, 4362–4366 (1994)
6. Cutler, M., Mott, N.F.: Observation of Anderson localization in an electron gas. Phys. Rev. **181**, 1336–1340 (1969)
7. Irkhin, V.Y., Irkhin, Y.P.: Electronic Structure, Correlated Effects and Physical Properties of *d*- and *f*-Metals and Their Compounds. Cambridge International Science Publishing, Cambridge, UK (2007)
8. Nag, B.R.: Electron Transport in Compound Semiconductors. Springer, Berlin/Heidelberg (1980)
9. Heremans, J.P., Thrush, C.M., Morelli, D.T.: Thermopower enhancement in lead telluride nanostructures. Phys. Rev. B **70**, 115334 (2004)
10. (a) Zabarjadi, M., Joshi, G., Zhu, G.H., Yu, B., Minnich, A., Lan, Y.C., Wang, X.W., Dresselhaus, M.S., Ren, Z.F., Chen, G.: Power factor enhancement by modulation doping in bulk nanocomposites. Nano Lett. **11**, 2225–2230 (2011); (b) Yu, B., Zebarjadi, M., Wang, H., Lukas, K., Wang, H.Z., Wang, D.Z., Opeil, C., Dresselhaus, M.S., Chen, G., Ren, Z.F.: Enhancement of thermoelectric properties by modulation-doping in silicon germanium alloy nanocomposites. Nano Lett. **12**, 2077–2082 (2012)
11. Heremans, J.P.: The effect of resonant energy levels on the thermoelectric power and thermoelectric power factor. In: Rowe, D.M. (ed.) CRC Handbook of Thermoelectrics, pp. 12–1. CRC Press, Boca Raton (2012)
12. Sakurai, J.J.: Modern Quantum Mechanics, Section 7.8, revised edition. Prentice Hall. Upper Saddle River, New Jersey, USA (1994)
13. Zebarjadi, M., Liao, B., Esfarjani, K., Dresselhaus, M., Chen, G.: Enhancing the thermoelectric power factor by using invisible dopants. Adv. Mater. **25**, 1577–1582 (2013)
14. Makongo, J.P.A., Misra, D.K., Zhou, X., Pant, A., Shabetai, M.R., Su, X., Uher, C., Stokes, K.L., Poudeu, P.F.P.: Simultaneous large enhancements in thermopower and electrical conductivity of bulk nanostructured half-Heusler alloys. J. Am. Chem. Soc. **133**, 18843–18852 (2011)
15. Xie, W., Yan, Y., Zhu, S., Zhou, M., Populoh, S., Gałązka, K., Poon, S.J., Weidenkaff, A., He, J., Tang, X., Tritt, T.M.: Significant ZT enhancement in p-type Ti(Co, Fe)Sb-InSb nanocomposites via a synergistic high mobility electron injection, energy filtering and boundary scattering approach. Acta Mater. **61**, 2087–2094 (2013)
16. Tan, G.J., Wang, S.Y., Li, H., Yan, Y.G., Tang, X.F.: Enhanced thermoelectric performance in zinc substituted p-type filled skutterudites $CeFe_{4-x}Zn_xSb_{12}$. J. Solid State Chem. **187**, 316–322 (2012)
17. Poudel, B., Hao, Q., Ma, Y., Lan, Y., Minnich, A., Yu, B., Yan, X., Wang, D., Muto, A., Vashaee, D., Chen, X., Liu, J., Dresselhaus, M.S., Chen, G., Ren, Z.:

High-Thermoelectric Performance of Nanostructured Bismuth Antimony Telluride Bulk Alloys. Science **320**, 634–638 (2008)
18. Puneet, P., Podila, R., Karakaya, M., Zhu, S., He, J., Tritt, T.M., Dresselhaus, M.S., Rao, A.M.: Preferential scattering by interfacial charged defects for enhanced thermoelectric performance in few-layered *n*-type Bi_2Te_3. Sci. Rep. **3**, 3212 (2013)
19. Zhao, L., Hao, S., Lo, S.H., Wu, C.I., Zhou, X., Lee, Y., Li, H., Biswas, K., Hogan, T.P., Uher, C., Wolverton, C., Draivd, V.P., Kanatzidis, M.G.: High thermoelectric performance via hierarchical compositionally alloyed nanostructures. J. Am. Chem. Soc. **135**, 7364–7370 (2013)
20. Dresselhaus, M.S., Chen, G., Tang, M.Y., Yang, R.G., Lee, H., Wang, D., Ren, Z.F., Fleurial, J.P., Gogna, P.: New directions for low-dimensional thermoelectric materials. Adv. Mater. **19**, 1043–1053 (2007)

Nanostructures

► Tight-Binding Simulations of Nanowires

Nanostructures and Characteristics of Carbon Nanofibers

Anupama B. Kaul[1], Jaesung Lee[2] and Philip Feng[2]
[1]Department of Metallurgical and Materials Engineering, Department of Electrical and Computer Engineering (Joint), University of Texas, El Paso, College of Engineering, El Paso, TX, USA
[2]Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA

Introduction

Research on nanocarbons has been at the forefront over the past several decades and has unarguably contributed tremendously to the birth of the field of nanotechnology. These efforts include providing insights into the novel properties of carbon nanomaterials, development of large-area, wafer-scale synthesis techniques, design of new characterization tools for understanding the role of imperfections in these materials, and the development of processing methods to integrate these materials into novel device architectures. In particular, in the context of applications of interest to the electronics industry, carbon nanomaterials have been proposed as viable alternatives to traditional materials used in Si integrated-circuits (ICs), which includes their exploration in nanoscale transistors [1–6], interconnects [7, 8], field-emission displays [9], biosensors [10], heat-transporting assemblies [11], thermo-electric [12], photo-voltaic [13], and optical materials [14] platforms, as well as nano-electro-mechanical-systems (NEMS) [15–17] given their remarkable mechanical properties [18].

The origin of the extraordinary materials properties [19] prevalent in carbon-based nanomaterials arises from the structural arrangement of carbon atoms in the crystalline lattice. A diverse spectrum of allotropes is possible with these materials, which display a rich variety of physical properties. Even for the commonly known three-dimensional (3D) form of carbon, namely diamond and graphite, properties differ from the sp^3-bonded diamond to the sp^2-bonded graphene from which graphite is derived. A single layer of these carbon atoms in graphite results in the two-dimensional (2D) form of carbon, namely graphene, which is a honeycomb arrangement of carbon atoms. The lowest dimensional form of carbon, namely zero-dimensional (0D) buckyballs or buckminsterfullerene C_{60} spheres were discovered in 1985 by Smalley and coworkers [20], for which they received a Nobel Prize in chemistry in 1997.

When a 2D graphene sheet is rolled into a cylinder, a one-dimensional (1D) or quasi-1D form of carbon results, namely carbon nanotubes (CNTs). This 1D form of carbon has captured the interest of scientists and engineers for nearly two decades since its discovery in 1991 [21]. This first form of CNTs synthesized using an arc-discharge process was for multi-walled carbon nanotubes (MWCNTs). Two years later, another form of CNTs, namely single-walled carbon nanotubes (SWCNTs) [22], was discovered which is a single

rolled-up sheet of graphene having a typical diameter of 1–2 nm. MWCNTs consist of concentric cylinders that have an interlayer spacing of 0.3–0.4 nm and diameters that are at least an order of magnitude larger than SWCNTs, between 10 nm and 30 nm. When a 2D graphene sheet is patterned into strips, another 1D or quasi-1D form of carbon results, namely graphene nanoribbons (GNRs) [23]. An additional quasi-1D structure, carbon nanofibers (CNFs), comprise of graphene layers inclined to the central axis, commonly referred to as the fishbone or herringbone crystal structure.

Graphitic carbon nanomaterials, such as 0D fullerenes, 1D SWCNTs, quasi-1D MWCNTs and CNFs, and 3D graphite, are thus all derivatives of graphene's 2D honeycomb lattice. The sp^2-bonding of carbon atoms in graphitic carbon nanomaterials generally affords them exceptional materials properties, such as a high-charge-carrier mobility, high-thermal conductivity, and excellent mechanical flexibility, which make them attractive for a range of applications [24–27]. It is interesting to note that fullerenes – the lowest dimensional form of carbon nanomaterials – were the first to be discovered amongst the graphitic carbon nanomaterial family by Kroto et al. [20] in 1985, followed by carbon nanotubes in 1991 by Iijima [21], while graphene remained elusive until recently despite being the basis for the other forms of carbon nanomaterials.

Since the mechanical exfoliation of graphene in 2004 [28], there has been a surge in research for unveiling the remarkable properties of this single sheet of carbon atoms and applying these properties to a wide range of devices and technological applications. As early as 1962, Boehm et al. [29] reported on the synthesis of single and multilayer graphene through the reduction of graphene oxide, and this technique has gained renewed interest recently. In addition, graphene was also suggested to form on transition metals such as Pt (100) and Ni (111) by hydrocarbon dissociation [30–32] and by other techniques such as metal intercalation [33]. While these surface science investigations of graphene on metals [34] occurred several decades ago, the exploration of fundamental properties, such as the quantum hall effect and other interesting phenomena [35–37], as well as the ensuing applications of graphene, did not occur until after 2004, following the seminal experiments of Novoselov and Geim on mechanically exfoliated graphene that earned them the 2010 Nobel Prize in Physics.

Up until recently, two-dimensional (2D) atomic monolayers were presumed to generally exist only as parts of larger three-dimensional (3D) structures, such as epitaxially grown crystals on lattice-matched substrates [38]. The existence of free-standing graphene was dismissed for some time as not being thermodynamically stable by Landau [39] and then several decades later by Mermin [40]. At the same time, the electronic structure and properties of graphene were analyzed theoretically by Wallace as early as 1947 [41] who showed that a single sheet of sp^2-hybridized carbon would have a linear energy dispersion relation at the *K-point* of the Brillouin zone.

Due to its flexibility, strength, high conductivity, transparency, and low cost, graphene has been proposed for a number of applications reviewed recently [42], such as a replacement for indium tin oxide (ITO) [43]. It also appears attractive for organic light-emitting diodes (OLEDs) as well as displays, touch screens [44], and solar cells [45]. The large surface-area-to-volume ratio of graphene suggests it also has promise in ultracapacitor applications [46] or chemical sensors [47, 48], and coupled with its remarkable electronic properties it is a strong contender for next-generation electronic devices in nanoelectronics [49, 50]. Graphene and graphene-like materials have also been applied to composite materials applications [15, 51, 52].

In this entry, we present an overview of nanocarbon materials, focusing specifically on 1D CNTs, CNFs, and 2D graphene. In section "Nanocarbons for Optical Absorbers" of this entry, we will discuss the application of 1D CNTs for optical absorbers, and in section "Nanocarbons for NEMS,", the use of 1D CNFs and 2D graphene for NEMS is highlighted, specifically for resonator applications for mass and strain sensing.

Nanocarbons for Optical Absorbers

Nanomaterials for Optical Absorbers

The ability of nanomaterials to trap light effectively has important implications for their use in energy harnessing, optical blacks for radiometry, as well as detectors. A survey of a host of nanomaterials, such as CdSe nanocrystals [53], graphene [54], surface plasmon modes in metallic nanoparticles [55], reveals the promise such materials have in a wide range of optical applications. In this section, we describe another type of nanomaterial which is exceptional at trapping incoming light as a result of its unique physical structure. Such a structure is comprised of porous arrays of thin (10–15 nm diameter), vertically oriented MWCNTs. Interestingly, porosity in ordered nanostructures has been theoretically [56] shown to enhance optical absorption properties, which has also been verified experimentally in Si-based nanomaterials [57, 58], and similar conclusions have also been drawn from theoretical calculations made on low-density arrays of CNTs [59].

The CNTs synthesized here for the optical absorber applications suggest that the mechanism of transducing optical energy into thermal energy may be an attractive platform for using such absorbers in energy harnessing, high sensitivity thermal detectors, radiative cooling, thermography, antireflection coatings, and optical baffles to reduce scattering. In section "Synthesis and Optical Characterization," the synthesis and optical characterization of the absorbers are discussed, followed in section "High-Temperature Optical Performance" by the ruggedness of these absorbers when they are exposed to high temperatures.

Synthesis and Optical Characterization

Previously synthesized MWCNTs [60] and SWCNTs [61] for optical absorber applications used water-assisted thermal CVD [62], which yields exceptionally high growth rates with CNT lengths >100 μm; here alignment occurs primarily via the crowding effect [63, 64]. While thermal CVD is generally considered ineffective in aligning short CNTs (<10 μm), here we show that a glow discharge leads to vertically aligned CNTs with lengths more than an order of magnitude lower, which nonetheless still yield broadband, high-efficiency optical absorption characteristics in the ultraviolet (UV)-to-IR range. This vastly extends the previously reported measurements on MWCNTs that were conducted in the visible regime [61] to well into the IR regime where it is increasingly difficult to find suitable black and opaque coatings. A thin and yet highly absorbing coating with absorptance A is valuable for thermal detector applications in the IR, for radiometry in order to enhance sensitivity, since the detectivity $D^* \propto A$ [65], and there is also a concomitant reduction in the response time.

Besides the importance of vertical alignment in yielding highly absorbing materials, the other structural trait for enhancing optical absorption efficiency is a high site density. Unlike earlier reports where the CNTs were synthesized directly on Si or SiO_2 [60, 61], we demonstrate growth of high-efficiency MWCNT absorbers directly on metallic substrates with site densities as high as $\sim 4 \times 10^{11}/cm^2$. In many other applications, it is desirable to grow CNTs directly on metals for lowering contact resistance, but the challenges in stabilizing catalyst particles on metallic surfaces at high temperatures have generally reduced site densities of CNTs to manyfold (up to 100 X). In addition, prior attempts at growing MWCNTs for optical absorber applications on substrates other than Si, such as $LiNbO_3$, yielded an absorption efficiency of ~85 % from $\lambda \sim 600$ nm to 1800 nm [66], whereas the CNT absorbers synthesized here on metallic substrates are shown to have an absorption efficiency of >99.98 % from $\lambda \sim$ 350 nm to 2500 nm [67]. Even cermet-based materials, currently used for solar selective coatings on metallic substrates such as Cu and Al [68], have absorption efficiencies that are several orders of magnitude (up to 10^4X) lower than that reported in this entry. A plasma-based process increases the potential of forming these absorbers at lower synthesis temperatures compared to thermal CVD, increasing future prospects of integrating such absorbers with a wider range of materials such as low-cost flexible substrates for solar cells or with thermoelectrics, as well as fragile,

temperature-sensitive micro-machined structures for IR sensing.

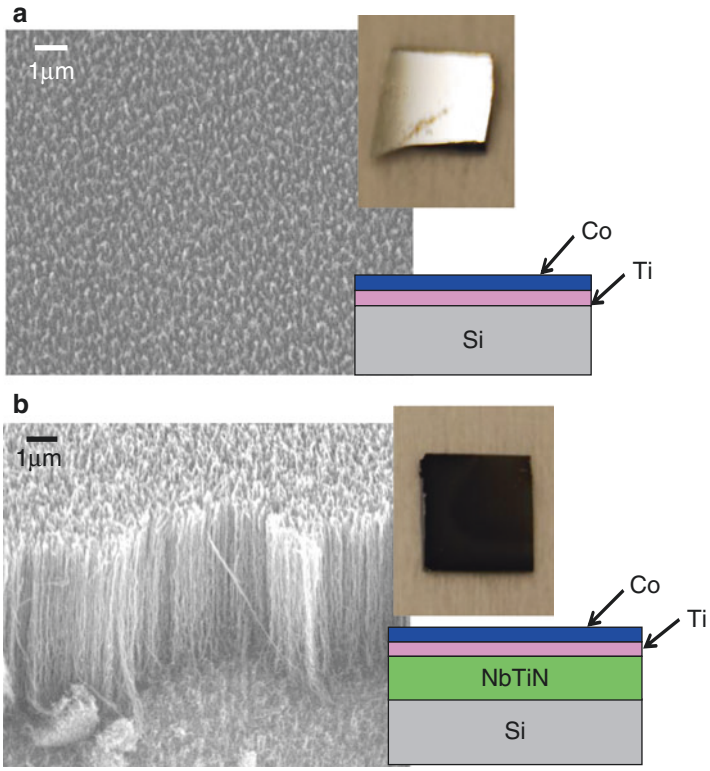
The initial substrate for the synthesis of the MWCNTs was a <100> oriented Si wafer on which a layer of 100–200 nm thick NbTiN – a refractory, high temperature conducting nitride – was deposited reactively in a N_2 and Ar ambient using dc magnetron sputtering at a power of ∼220 W and 5 mTorr. Bi-metallic layers of Co (thickness range, 0.6–6 nm) and 2.5 nm thick Ti were e-beam evaporated and served as the catalyst. Beside the Co/Ti/NbTiN/Si templates, control samples of Co/Ti/Si, Co/NbTiN/Si, and Co/Si were also prepared. Multiple samples (area ∼4 cm^2) were placed on a wafer during PECVD growth so that comparative analysis could be performed for different combinations of templates under similar synthesis conditions. At temperatures in the range of 550–750 °C, H_2 was flowed into the chamber for several minutes, and the growth gases acetylene (C_2H_2) and ammonia (NH_3) were then introduced to a typical pressure of ∼5 Torr, and the discharge was then ignited.

The choice of the template for PECVD synthesis of our MWCNTs was vital in synthesizing a high-density array of CNTs and directly impacts the optical absorption characteristics. For example, the scanning electron microscope (SEM) image in Fig. 1a shows amorphous carbon deposits when Co/Ti was placed directly on Si at 750 °C, exhibiting a largely reflective surface (inset of Fig. 1a). On the other hand, using a Co/Ti/NbTiN template yielded a visually black sample to the naked eye (inset of Fig. 1b), and the SEM image (Fig. 1b) depicts a high-density array of MWCNTs which traps incoming light and suppresses reflection. The lack of growth of MWCNTs on Co/Ti/Si templates (Fig. 1a) suggests that the presence of a refractory metallic nitride, such as NbTiN, is important in stabilizing the catalyst nanoparticle to prevent diffusion and alloying of the catalyst with the underlying Si at high temperatures. In addition, the density of MWCNTs in the absence of the Ti layer on the Co/NbTiN templates was low. It is speculated that the Ti may enable the Co to fragment into nanoparticles, similar to the role of Mo in the Co-Mo bi-metallic catalyst system [69]. The Ti-Co system also appears to incorporate a larger fraction of C compared to Co alone, enhancing CNT growth [70]. Besides being of interest as absorbers in solar photo-thermal applications, the high areal density of MWCNTs on reflective, low resistivity (∼110 μΩ-cm) metallic substrates may substantially reduce the CNT-to-substrate contact resistance.

The high magnification image in Fig. 2a shows that the surface of the MWCNT arrays is rough, a factor which also contributes to scattering the incoming light diffusively. Shown in Fig. 2b is the SEM image of our benchmark, a Au-black absorber, which was synthesized using approaches similar to prior reports [71], where the percolated random network of such a diffuse metal-black should be apparent.

The optical measurements on the samples were conducted from $\lambda \sim 350$ nm to 2500 nm using a high resolution, fiber coupled, spectroradiometer (ASD inc, Fieldspec Pro) where a standard white light beam was focused at normal incidence to the sample, as shown by the schematic in the inset of Fig. 2c. The bare fiber connector of the spectroradiometer was oriented at ∼40° from the normal. Relative reflectance spectra were obtained by first white referencing the spectroradiometer to a 99.99 % reflective spectralon panel. The reflected light intensity from the sample was then measured, and the spectra compared for samples synthesized under different growth conditions.

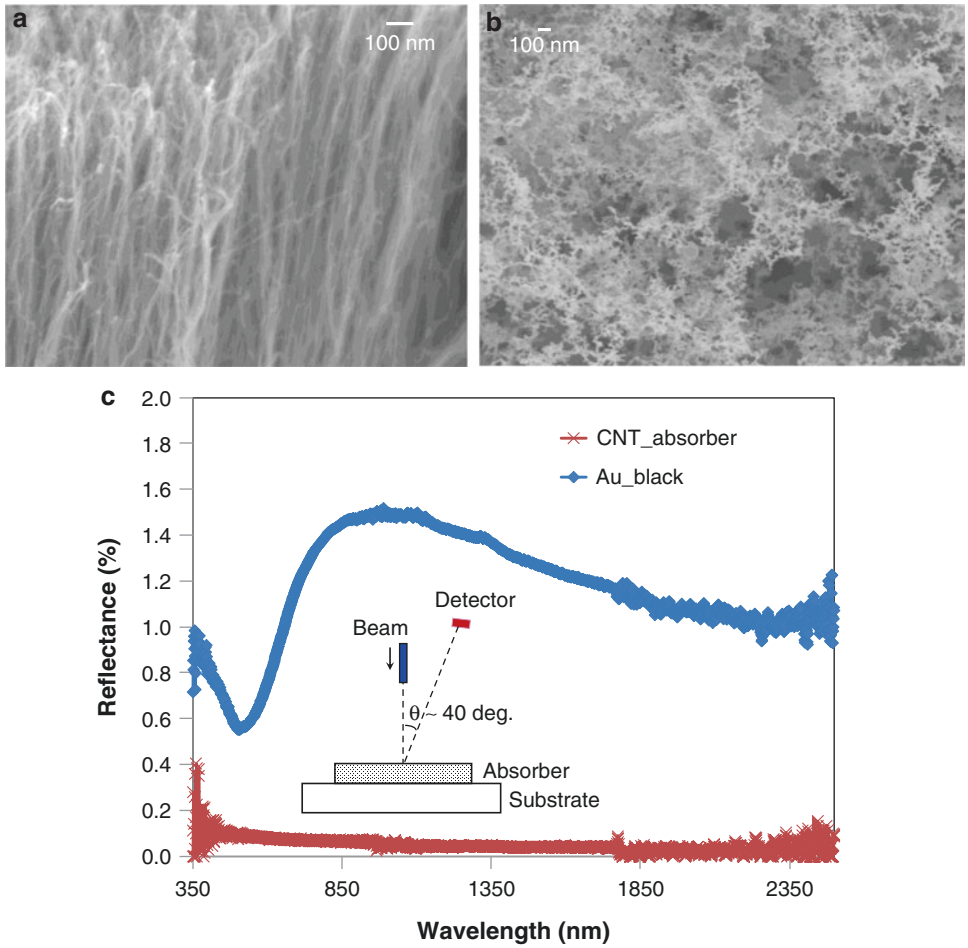
The optical reflectance response of the CNT absorber is shown in Fig. 2c, where the spectra are compared to the reference Au-black absorber. The reflectance R of the CNT absorber is nearly two orders of magnitude lower than that of the Au-black, e.g., ∼0.02 % at $\lambda \sim 2000$ nm compared to 1.1 % for Au-black. Other commonly used absorbers, such as NiP, have higher $R \sim 0.5$–1 % for $\lambda \sim 320$–2140 nm [72] while ultra-black NiP alloy has $R \sim 0.16$–0.18 % from $\lambda \sim 488$–1500 nm [73] and black paint has $R > 2.5$ % from $\lambda \sim 600$–1600 nm. Top-down synthesized Si nanotips exhibit $R \sim 0.09$ % at $\lambda \sim 1000$ nm [74], while bottom-up synthesized nanocone arrays have been reported to have an absorption efficiency of ∼ 93 % between $\lambda \sim 400$–650 nm [75].



Nanostructures and Characteristics of Carbon Nanofibers, Fig. 1 (**a**) SEM micrograph of a Co/Ti/Si sample after dc PECVD growth. *Top right inset* shows an optical image of the sample depicting a reflective surface. (**b**) SEM micrograph of a Co/Ti/NbTiN sample after growth, depicting a high density carpet of MWCNTs. *Top right inset* shows an optical image of the sample depicting a visually black sample to the naked eye. The spatial uniformity of the MWCNT ensembles is high over large length scales

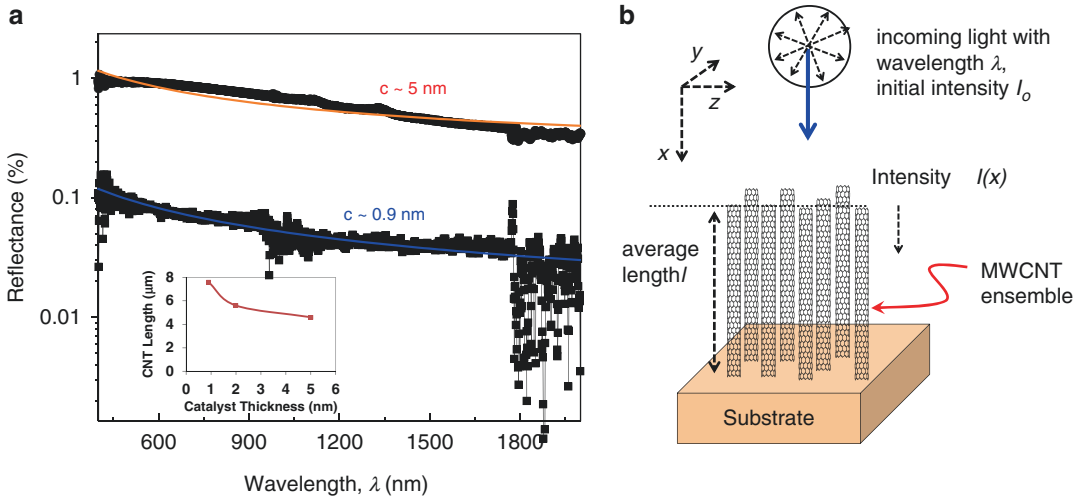
The catalyst thickness appears to be an important synthesis parameter that impacted the optical absorption efficiency in these carbon-based nanoabsorbers. Shown in Fig. 3a are reflectance spectra taken for two samples synthesized at Co catalyst thicknesses $c \sim 5$ nm and 0.9 nm (Ti thickness fixed at 2.5 nm). The sample with $c \sim 0.9$ nm has a wavelength independent response from $\lambda \sim 350$ nm – 2500 nm with R in the 0.02–0.03 % range. The sample with $c \sim 5$ nm, synthesized at identical conditions, has a wavelength dependent R which decreased from 0.94 % at $\lambda \sim 400$ nm to ~ 0.33 % at $\lambda \sim 2000$ nm.

Tentatively, the decreased reflectance/increased absorption may be expressed through an exponential decrease of the transmitted intensity $I(x)$ following a simple Lambert–Beer law formulation, i.e., $I(x) = I_o \, exp\,(-\alpha x)$, where I_o is the initial intensity of the incoming light and α is the absorption coefficient. The schematic in Fig. 3b shows the geometry of the optical interrogation, as the incoming light traverses through the sparse aerogel-like structure of CNTs with intensity $I(x)$ at any vertical location inside the CNT array. The typical absorption coefficient with $\alpha \sim 10^4$ cm^{-1} for $I(x = 8\ \mu m)$ is ~ 30 times the $I(x = 4.5\ \mu m)$. Now, the reflectance seems to decrease to the same order, i.e., on the average R drops from ~0.94 to ~0.03, ~30 times as well. The SEM images of the samples with $c \sim 0.9$ nm, 2 nm, and 5 nm are shown in Fig. 4a, b, and c, respectively. This yielded a MWCNT site density of $\sim 4 \times 10^{11}$/cm^2 with MWCNT diameters $d \sim 10$–15 nm for $c \sim 0.9$, and a site density of $\sim 6 \times 10^9$/cm^2 with $d \sim 80$-100 nm for $c \sim 5$ nm. Although the length l of the MWCNTs decreased as c increased (inset of Fig. 3a), with $c \sim 5$ nm, l was still >5 μm, well above λ in these measurements, suggesting that the reduced absorption from the thicker catalyst is likely a result of changes in the fill fraction. The ability to engineer optical absorption efficiency by controlling the catalyst thickness is an attractive feature in tuning the optical absorption properties of the MWCNT ensembles.

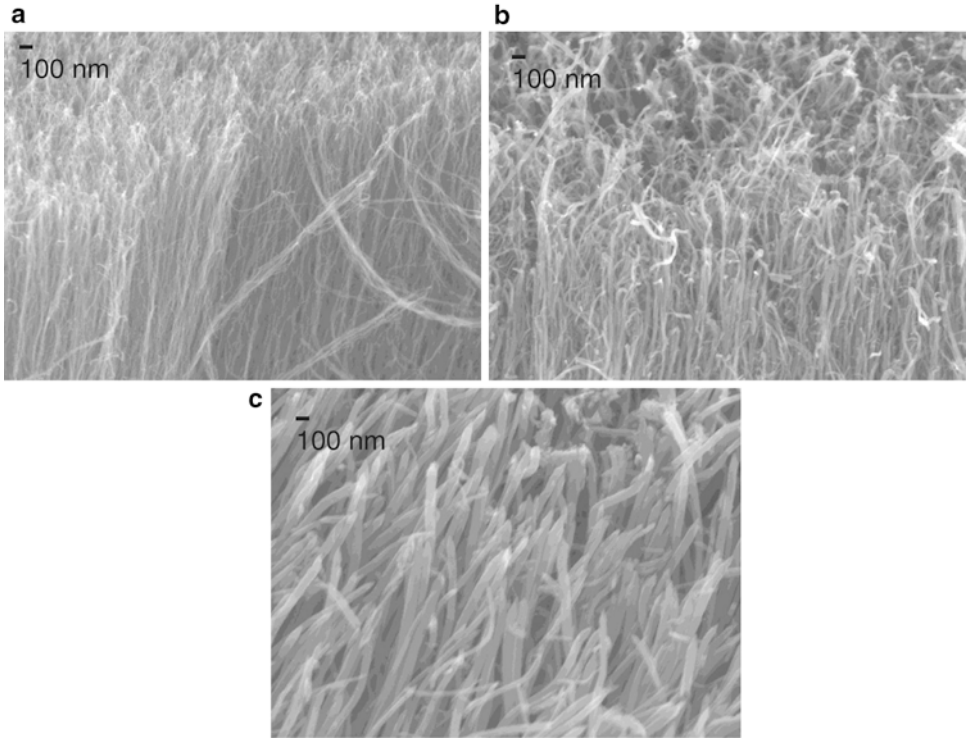


Nanostructures and Characteristics of Carbon Nanofibers, Fig. 2 (**a**) High magnification SEM image shows the porous, vertically aligned morphology of the CNT absorbers, in contrast to the reference benchmark Au-black absorber sample in (**b**) which depicts a percolated, randomly aligned network of fibers. SEMs taken at 30° viewing angle. (**c**) Reflectance measurement from $\lambda \sim$ 350–2500 nm for the MWCNT absorber and a Au-black absorber sample. The measurement setup is illustrated in the *bottom inset*. The Au-black reference sample has $R \sim$100 X larger where $R \sim 0.02$ % at $\lambda \sim 2000$ nm for the CNT sample compared to 1.1 % for the Au-black

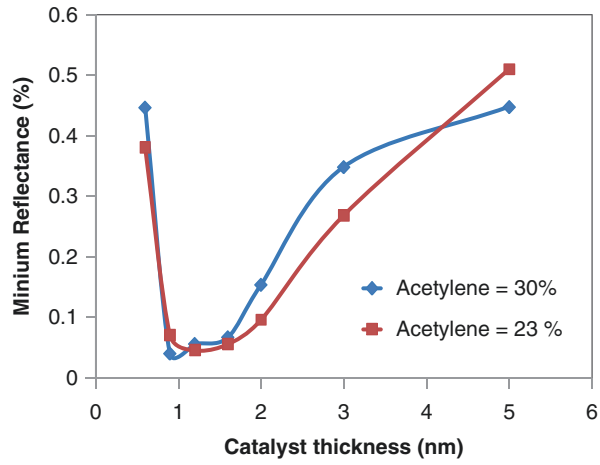
A mechanism by which porous objects suppress reflection is through a reduction in the effective refractive index, n. However, porosity alone may not necessarily be the primary factor involved since the Au-black absorber samples, a largely porous structure (SEM in Fig. 2b), had higher reflectance compared to the MWCNT samples. This enhanced absorption may arise from the weak coupling of electrons in the vertically oriented CNTs to the incoming normally incident radiation, with minimal back scattering and enables light to propagate into the long pores within the arrays until it is finally absorbed. A phenomenological model for absorption was developed using a formulation where the ensembles are treated as a composite medium consisting of nanostructures and air. The intensity at any given point x in Fig. 2b is given by $I(x) = I_o \exp(-\alpha x)$, where $\alpha = \frac{4\pi\kappa}{\lambda}$ and κ is the extinction coefficient. Assuming that there is no effective transmission through the substrate, $R(x) \sim (I_o - I(x))$. The corresponding variation of R with λ was then fit to $\sim a_1 e^{a_2/\lambda} + a_3$ where a_1 is related to the incident intensity I_o, a_2 is a



Nanostructures and Characteristics of Carbon Nanofibers, Fig. 3 (**a**) Optical reflectance spectra taken for two samples with $c \sim 0.9$ nm and 5 nm. The sample with $c \sim 5$ nm shows a wavelength dependent R. Inset shows the variation of l with c; superimposed in the reflectance versus wavelength data are theoretical fits from which the ratio of κ at $c \sim 0.9$ nm and 0.5 nm was determined. (**b**) The geometry used for the modeling analysis



Nanostructures and Characteristics of Carbon Nanofibers, Fig. 4 The morphology of the MWCNTs is depicted in (**a**), (**b**), and (**c**) for $c \sim 0.9$ nm, 2 nm, and 5 nm, respectively. In (**a**), thin, vertically aligned tubes are depicted that have a high fill fraction, with a site density of $\sim 4 \times 10^{11}/\text{cm}^2$ and MWCNT diameters ~10–15 nm, whereas in (**c**) the site density decreases to $\sim 6 \times 10^{9}/\text{cm}^2$ with MWCNT diameters ~80–100 nm



Nanostructures and Characteristics of Carbon Nanofibers, Fig. 5 The reflectance measurement as a function of c (taken at $\lambda \sim 1.5$ μm) for two acetylene gas ratios (30 % and 23 %). Growth conditions were: 750 ºC, 172 W of plasma power, 30 % C_2H_2, 5 Torr

measure of the optical absorption length ($=\kappa \cdot l$), and a_3 is a constant. The fit to the data is shown in Fig. 3a for $c \sim 0.9$ nm and 5 nm. From the fits, the value of a_2 was determined to be $\sim$ 0.025 and $\sim$0.026 for $c \sim 0.9$ nm and 5 nm, respectively, and given that the ratio, $\frac{(a_2)_{0.9}}{(a_2)_5} = \frac{(\kappa \cdot l)_{0.9}}{(\kappa \cdot l)_5}$ and that l is 8 μm and 5 μm, respectively, we obtain a ratio of the extinction coefficients, $\frac{\kappa_{0.9}}{\kappa_5}$ of $\sim$0.6.

It is interesting that the extinction ratio is smaller for the MWCNTs grown with $c \sim 0.9$ nm compared to the CNTs grown with $c \sim 5$ nm. This observation can be rationalized on the basis of a smaller area fraction in the former case, i.e., 0.31 versus 0.47. While such a rationalization does not explicitly consider the volume absorption due to a larger l in the former case (i.e., 8 μm vs. 5 μm), it is justified since it has previously been shown that for the case of absorption in Si nanowires [56], the absorption in a thin film over a wide energy range, comparable to the one used in the present work, is on the average equivalent to the absorption in the nanowires. The larger absolute magnitude of R for the sample with $c \sim 5$ nm compared to the sample with $c \sim 0.9$ nm may indicate an influence of the substrate in the latter, the effect of which is more pronounced due to a shorter l for $c \sim 5$ nm.

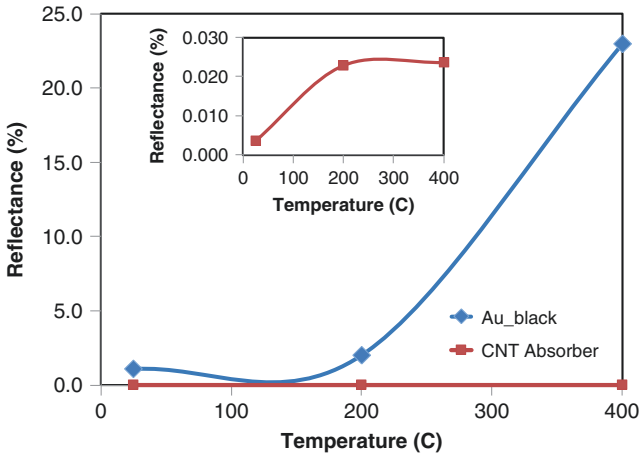
A more detailed analysis of the impact of catalyst thickness on the optical reflectance properties of the MWCNT absorbers was conducted for a wide range of catalyst thicknesses (Fig. 5). This data (at $\lambda \sim 1500$ nm) shows a minimum of R at $c \sim 1$ nm. However, R increases when $c \sim 0.6$ nm due to the inability to nucleate a high enough areal density of MWCNTs; such behavior was consistent for two different acetylene gas concentrations, as indicated in Fig. 5.

High-Temperature Optical Performance

We now present data which demonstrates the exceptionally low R of the MWCNT absorbers even after they were exposed to temperatures as high as 400 °C in air under an oxidizing environment, as might be expected with incident solar radiation. In comparison, the structural characteristics of the Au-black absorber reference gradually deteriorate with increasing temperature. However, the structural characteristics of the MWCNT absorber samples are largely unchanged when heated from 25 °C to 400 °C, as can be inferred from the data in Fig. 6.

From the optical spectra (Fig. 6), it is apparent that R of the Au-black absorber sample increases as it is heated from 25 °C to 200 °C (2 %) and is $\sim$23 % at 400 °C (at $\lambda \sim 2000$ nm). On the other hand, the R of the CNT absorbers is still very low ($\sim$0.022 %) after heating to 200 °C (inset Fig. 6) and remains unchanged after exposure to temperatures as high as 400 °C which can be correlated to the structural integrity of the CNT absorbers to temperatures as high as 400 °C.

In conclusion, we have successfully shown that, through catalyst engineering, PECVD synthesized MWCNTs yield a high site density on



Nanostructures and Characteristics of Carbon Nanofibers, Fig. 6 Optical reflectance measurements were made in the λ ∼ 350 nm–2500 nm range for the Au-black and CNT absorber samples at 25 ºC and after subjecting them to temperatures up to 400 ºC. Shown is *R* of the Au-black and CNT absorber samples as a function of temperature. The Au-black absorber shows *R* increases up to 23 % after heating to 400 ºC. The inset shows *R* of the CNT sample increases slightly after exposure to 200 ºC but it is still very low, ∼ 0.022 % at λ ∼ 2000 nm and remains unchanged after exposure to temperatures as high as ∼400 ºC

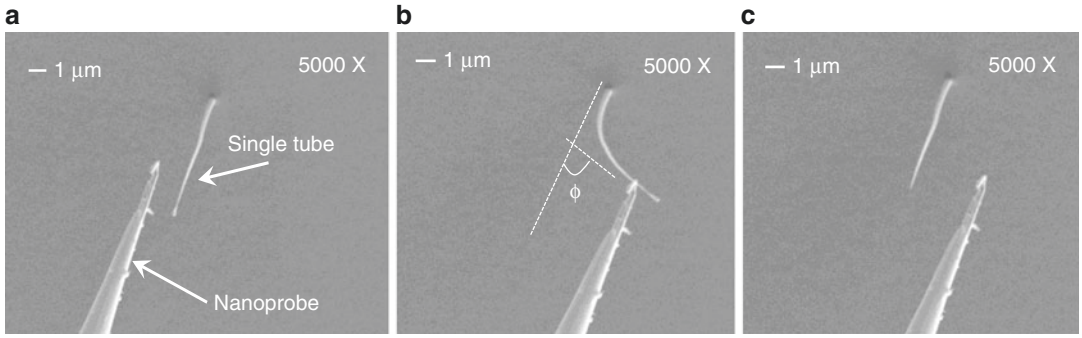
metallic substrates which exhibit ultra-low reflectance (∼0.02 %) over a wide spectral range from UV-to-IR for relatively thin (< 10 μm) absorber ensembles. Their highly efficient optical absorption properties and exceptional ruggedness at high temperatures suggest their promise in solar photothermal applications and IR thermal detectors for radiometry applications. In addition, the use of a plasma-based process increases the potential for synthesizing the absorbers at lower temperatures in the future, increasing the likelihood of integrating the absorbers with low-cost flexible substrates, potentially for solar-cell applications, as well as thermoelectrics and micro-machined structures for enabling new classes of IR sensors, particularly for rugged environments.

Besides the use of nanocarbons for optical absorbers, we now turn to their use as mechanical resonators for NEMS applications, as described in section "Nanocarbons for NEMS."

Nanocarbons for NEMS

Nano-electro-mechanical-systems (NEMS) are gaining increasing attention as viable alternatives for beyond-CMOS architectures, for enabling zero-leakage, ultralow-power, abrupt switching devices due to their inherently mechanical construction [76]. Such nanomechanical structures also offer advantages as electronic components operational in harsh environments, such as high radiation and high temperatures, in contrast to conventional solid-state devices which are more vulnerable to failure in hostile environments. In addition to their utility as potential beyond-CMOS switches, dynamical and resonant properties of NEMS have allowed these characteristics to be exploited for various sensing applications, such as mass sensing [77], force detection [78], and measurements in the quantum-mechanical regime [79].

While many materials have been explored for NEMS devices including conventional materials such as Si [80] and SiC [81], carbon-based materials have also been noted as another promising candidate for NEMS. Nanocarbons, such as graphene [82], CNTs [15, 83, 84], and CNFs [17, 85], exhibit ultralow density and an exceptionally high Young's modulus. Coupled with their high elasticity and mechanical resilience, nanocarbon materials appear to be ideally suited for enabling high-performance, high-longevity NEMS resonators. In section "Carbon Nanofiber



Nanostructures and Characteristics of Carbon Nanofibers, Fig. 7 (**a**) Bending test on a single CNF with a nanoprobe reveals that the CNF is extremely resilient and able to tolerate bending angles ϕ as large as 70° as shown in (**b**). The CNF returns elastically to its initial position as shown in (**c**)

(CNF)-Based NEMS Resonators," we discuss the application of quasi-1D CNFs for NEMS resonator applications, followed in section "Graphene-Based NEMS Resonators" with a discussion on the use in 2D graphene for NEMS.

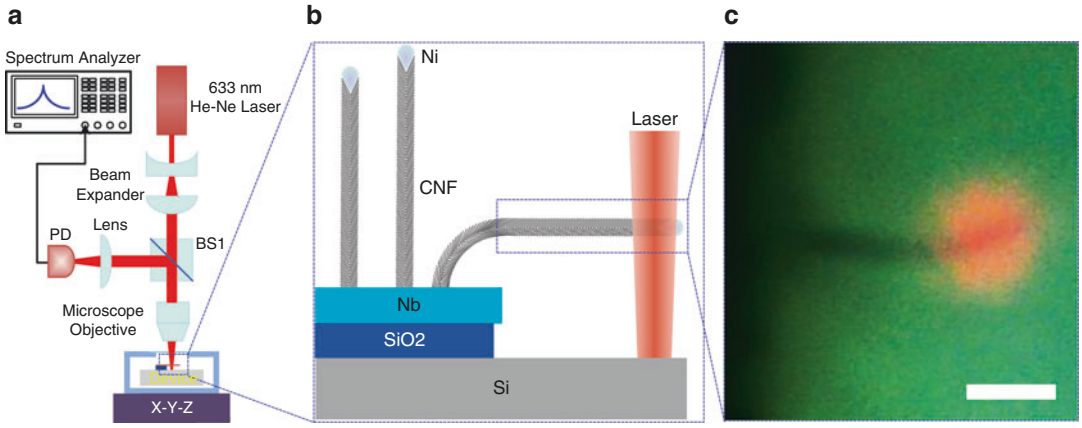
Carbon Nanofiber (CNF)-Based NEMS Resonators

Our carbon nanofibers were synthesized using electric field assisted growth. Electric fields applied during CVD growth of nanotubes serve to preferentially orient them in the direction of the field [86] which has also been observed in the PECVD synthesis of MWCNTs and fibers [87, 88]. We have also applied the inherent E-fields present in plasma growth to develop processes for forming vertically aligned MWCNTs using dc PECVD, where a gas mixture of NH_3 and C_2H_2 were used at ~ 700 °C. A stable and uniform dc glow discharge results over an area as large as ~ 75 mm. The plasma is more intensely concentrated toward the base electrode as pressure is increased, and the self-charging voltage also saturates after the first few minutes of growth under constant current control.

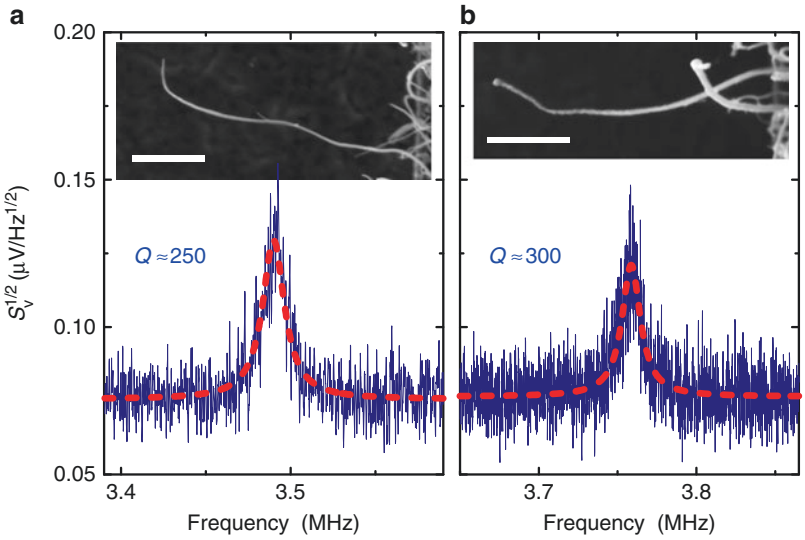
The vertically oriented CNFs were synthesized with pre-patterned Ni catalyst islands. Catalyst islands in the 300 nm range were formed in between the electrodes using top-down wafer scale approaches. After the liftoff process of the e-beam evaporated catalyst islands, high purity acetylene (C_2H_2) and ammonia (NH_3) were introduced at 700 °C, which served as the carbon feedstock and diluent gas, respectively, for the bottom-up synthesis of the CNFs using dc PECVD. When the desired growth pressure had been attained (~ 5 Torr), the dc discharge was ignited at a power of ~200 W, and growth was carried out for a fixed duration to the desired length of the CNFs. The CNFs ranged in length from 1 μm to 10 μm.

Given that the CNFs synthesized here using PECVD are targeted for NEMS applications, it is important to gain some insights into their mechanical properties. In this regard, bending tests were conducted on individual CNFs using a nanoprobe inside an SEM. The nanoprobe was mechanically manipulated so that it physically deflected the CNF to the right, as shown in Fig. 7a. The CNF sustained bending angles ϕ as large as ϕ ~ 70° (Fig. 7b), and it then returned elastically to its initial position (Fig. 7c). The CNFs were able to tolerate such severe strains over tens of cycles without detachment from the substrate or fracture within the tube body, suggesting their mechanical resilience and elasticity are extremely attractive for NEMS applications.

The resonance measurements are conducted using an optical interferometry setup as shown in Fig. 8, where a 633 nm He-Ne laser with a spot size of ~1 μm is focused on a single CNF inside a vacuum chamber equipped with an X-Y-Z stage. The X-Y-Z stage is manipulated so that the laser spot is placed precisely on the tip of the CNF, as shown by the schematic in Fig. 8b.



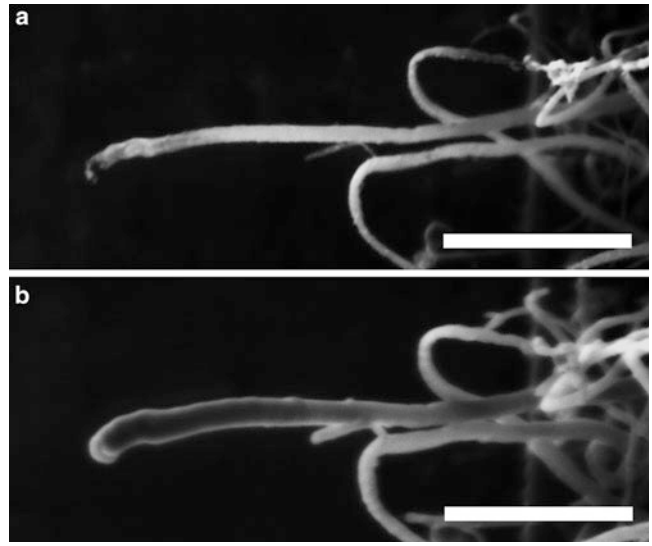
Nanostructures and Characteristics of Carbon Nanofibers, Fig. 8 Implementing laser optical interferometry on individual CNFs. (**a**) An optical interferometry displacement detection system. (**b**) A schematic illustration of a focused laser beam interrogating a single CNF. (**c**) An optical micrograph of a CNF under a laser spot focused near the tip of the CNF. Scale bar is 1 μm



Nanostructures and Characteristics of Carbon Nanofibers, Fig. 9 Resonances measured from two individual CNFs. Red dashed line shows fitting results for determining Q factors of the resonances. Insets are SEM images of the CNFs. Scale bars are 1 μm

Figure 8c depicts the optical microscope image of a single CNF captured within the focused laser spot. In prior measurements, high laser power was shown to melt and oxidize the Ni catalyst, leading to structural deformations in the CNFs [89]. It appears that at 500 μW, the laser power is low enough to prevent damage to the Ni catalyst and yet high enough to enable resonance measurements with high precision. As the reflected light from the CNF and the substrate interferes, the Brownian motion arising from fundamental thermal fluctuations in the CNF is detected by a photo detector, and the readout is captured by a spectrum analyzer that displays the fundamental thermo mechanical noise spectrum.

The undriven Brownian motion of a number of CNFs is recorded, with representative examples shown in Figs. 9a, b, where the top insets depict the corresponding scanning SEM images of the CNFs under consideration. The Q is extracted by fitting the measured thermo mechanical noise spectrum to a damped harmonic resonator model. Figure 9a exhibits a measured resonance at 3.490 MHz with Q of 250, while Fig. 9b shows



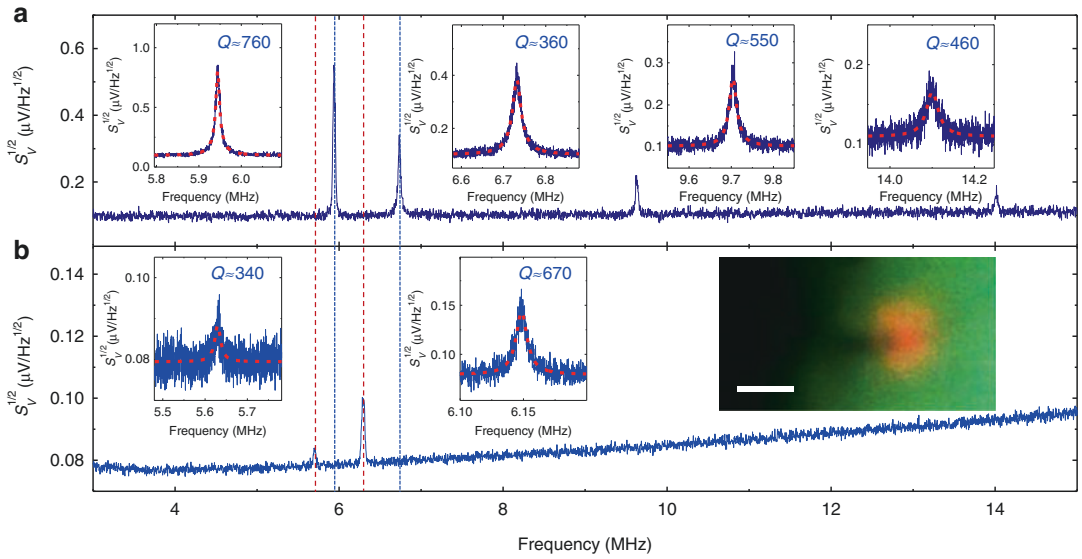
Nanostructures and Characteristics of Carbon Nanofibers, Fig. 10 SEM images of a CNF cantilever resonator showing the effects of electron beam induced carbon deposition on surface. (**a**) *Top view* SEM image before electron beam shower. (**b**) After ~15 min of 3 kV, 1.9 nA electron beam irradiation. Scale bar: 1 μm

a resonance frequency of 3.789 MHz with Q of 300. Most of the devices tested had CNF lengths and diameters in the range of $L \approx 2 - 6$ μm and $d \approx 40 - 100$ nm, respectively, and the corresponding fundamental mode of the resonance frequency was within $\sim 3 - 7$ MHz.

The resonance characteristics of the CNFs is examined before and after exposing the CNFs to electron radiation inside an SEM operated at relatively low acceleration voltage (3 kV), high current (1.9 nA), and long irradiation time. The electron beam inside an SEM can easily react with residual organic material in the SEM chamber and deposit amorphous carbon on the surface of the CNFs. This is clearly visible from the SEM images in Fig. 10, which shows the CNF to be apparently thicker after electron beam exposure (Fig. 10b) compared to before exposure (Fig. 10a).

Figure 11 shows the measured multimode resonance response from this single CNF before any electron beam treatment in the SEM, with four distinct resonance modes detected in the $5 - 15$ MHz range, with Qs of $360 - 760$. After irradiation of the CNF with the electron beam, the resonance response is immediately measured again. As shown in Fig. 11b, the first 2 major resonance peaks, originally at 5.94 MHz and 6.73 MHz, shift down to lower frequencies at 5.63 MHz and 6.15 MHz, respectively, with noticeable changes in Q factors. In addition, the signal amplitudes are lowered, and the two initially higher frequency resonances at 9.70 MHz and 14.10 MHz are no longer present after the electron beam shower. The disappearance of the 3rd and 4th modes can also be explained on the basis of carbon deposition which makes those peaks less responsive to optical detection.

The above results indicate that our 1D CNF NEMS devices encounter a reduction in resonance frequency due to electron beam irradiation induced amorphous carbon deposition on the surface. Based on measured frequency shift of the lowest resonance mode, the added mass can be calculated by using $\delta M/M \approx 2\delta f_0/f_0$, where, δM, M, δf_0, and f_0 are the added mass, total mass of resonator, added mass induced frequency shift, and initial resonance frequency of the device, respectively. Given the measured resonance frequency shift of the fundamental mode in Fig. 11b, we estimate the deposited mass on the CNF surface to be $\delta M \approx 4.4$ fg, while the CNF's own mass is $M \approx 41.5$ fg. Being able to have response to this ~5.2 % frequency shift due to ~10.6 % loaded mass suggests that the CNFs have exciting prospects for enabling robust adsorption-based mass sensors with very wide dynamic range.

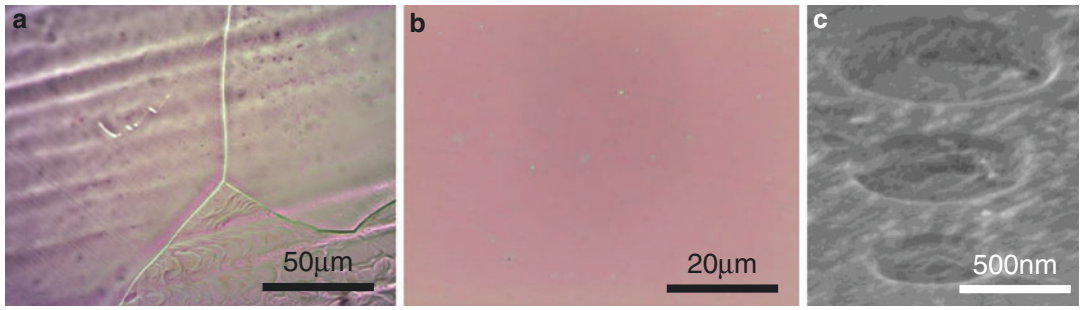


Nanostructures and Characteristics of Carbon Nanofibers, Fig. 11 Multimode resonance responses measured before and after electron beam radiation. (**a**) A wide range resonance response of an individual CNF before electron exposure. (**b**) Response spectrum after an electron beam shower. Insets show details of individual resonances. Inset image shows a CNF with laser spot (scale bar: 1 μm)

Graphene-Based NEMS Resonators

Just like carbon's other low dimensionality – 0D and 1D structures – 2D graphene exhibits very interesting properties. Electrons in graphene interact with the honeycomb lattice in a remarkable way which leads to unique electron transport properties. Graphene's charge carriers are referred to as massless Dirac fermions, where the electrons move relativistically at speeds of $\sim 10^6$ ms^{-1} instead of 3×10^8 ms^{-1}. These unique physical attributes make graphene a remarkable platform in which to also examine relativistic and interesting low-dimensional physics effects, besides the obvious potential it has for technological applications. The massless carriers and minimal scattering over micron size crystallites result in quantum mechanical effects that can survive even at room temperature. Graphene also exhibits ballistic room-temperature electron mobility of 2.5×10^5 $cm^2/V\text{-}s$ [90], the ability to carry ultra-high current densities $\sim 10^9$ A/cm^2 [91], an ultra-high Young's modulus of 1 TPa with a breaking strength of ~40 N/m [92] and excellent elasticity for accommodating strains of up to 20 % without breaking, a very high thermal conductivity of ~5000 W/m-K [93, 94], and optical absorption of 2.3 % [54]. Graphene also appears to be an excellent barrier for gases [95].

In work conducted here, we have explored the potential of graphene for enabling 2D NEMS resonators. The 2D graphene resonators were fabricated by transferring monolayer CVD grown graphene to lithographically pre-patterned microtrenches. While the transfer of graphene from metal growth substrates has become common practice for laboratory device fabrication, more work is necessary to achieve reproducibility and address issues of defect introduction. As an example, when clean interfaces are achieved between two transferred graphene layers, hybridized properties emerge that open new scientific and technological opportunities. It is also critical to grow graphene directly on high-quality dielectrics such as hexagonal boron nitride (h-BN), which is free of dangling bonds and charge in homogeneities, unlike SiO_2. This is crucial in order to maintain graphene's high charge carrier mobility and, in this regard, all future heterostructure architectures with active graphene elements.

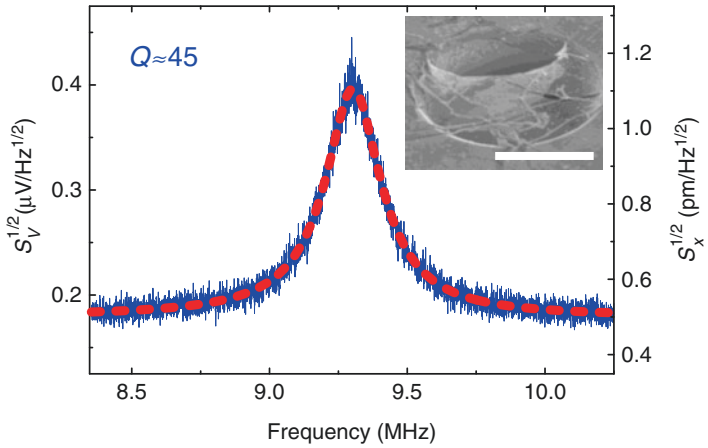


Nanostructures and Characteristics of Carbon Nanofibers, Fig. 12 CVD graphene on different substrates. (**a**) An optical microscope image of CVD graphene on Cu foil. Different sizes of Cu grains exhibit different surface morphologies. (**b**) Transferred CVD graphene on 290 nm-thick SiO_2 on Si substrate. Graphene is continuous over entire optical images. (**c**) A prospective view SEM image of suspended graphene on circular microtrenches, demonstrating drumhead graphene resonators

Nonetheless, although polycrystalline CVD graphene consists of various defects such as grains boundaries arising from initial growth and through transfer to other substrates, the mechanical properties of CVD grown graphene are comparable to that of mechanically exfoliated monocrystalline graphene [96]. In addition to the excellent mechanical properties, polycrystalline CVD graphene can be grown over large areas on copper (Cu) foil, facilitating the fabrication of large suspended structures through transfer techniques to pre-patterned microtrenches. In contrast, achieving such large suspended structures with mechanically exfoliated, monocrystalline graphene has been a significant challenge. CVD growth of graphene is performed by using a one-inch (1″-diameter) quartz tube furnace. Before growth, Cu foil is carefully cleaned with acetic acid to remove the native Cu oxide and the foil is then annealed at 1000 °C for 3 h in hydrogen (H_2) to remove any residual oxide; this anneal also allows for the small Cu grains to merge to form larger grains that serve as growth templates for graphene. Figure 12a shows an optical microscope image of graphene on Cu foil, showing the smooth surface of merged large Cu grains during the annealing process. Graphene is then synthesized using H_2 and methane (CH_4) gases at 1000 °C for 5 min. During the growth, we flow 100 sccm of CH_4 with 20 sccm of H_2 and at a pressure of 10 Torr. After growth, we perform Raman spectroscopy to characterize the CVD grown graphene. From Raman results, our CVD graphene has very high uniformity and high quality [97]. CVD graphene is then transferred to pre-patterned microtrenches to fabricate graphene resonators by using conventional wet transfer processes. Figure 12 demonstrates examples of transferred graphene onto (b) SiO2 substrates and (b) suspended over an array of microtrenches to yield drumhead resonators.

The resonance characteristics of circular drumhead CVD graphene resonators are measured by using an optical interferometry system, similar to the schematic shown in Fig. 8a. Figure 13 shows an example of the measured thermo mechanical noise spectrum from a drumhead graphene resonator. Thermo mechanical noise at resonance is given by $S_{x,th}{}^{1/2}(\omega_m) = (4k_BTQ/\omega_m{}^3M_{eff})^{1/2}$, where T, ω_m, k_B, Q, and M_{eff} are temperature, angular resonance frequency, the Boltzmann's constant, Q factor, and the device effective mass, respectively. From this equation, the calculated Brownian motion of the device at room temperature is $S_{x,th}{}^{1/2}(\omega_m) = 0.98$ pm/Hz$^{1/2}$ and responsivity and sensitivity of interferometry system with the graphene resonator are $\Re = S_{v,th}{}^{1/2}/S_{x,th}{}^{1/2} = 0.358$ μV/pm and $S_{x,sys}{}^{1/2} = 0.49$ pm/Hz$^{1/2}$, respectively.

The measured resonance characteristics provide precise mechanical properties of graphene. By using the drumhead membrane resonator model, $f_0 = 2.4048(\gamma/\rho t)^{1/2}/\pi d$, where, d, γ, ρ, and t are diameter, tension (N/m), density, and



Nanostructures and Characteristics of Carbon Nanofibers, Fig. 13 Measured thermo mechanical resonance response of a monolayer CVD graphene resonator. *Inset* shows an SEM image. Scale bar is 5 μm

thickness of the resonators, respectively, and the estimated tension in the resonator is $\gamma \sim 0.01$ N/m (strain of 3×10^{-5}), demonstrating the unique capability of measuring ultra small strain levels in a 2D NEMS platform.

Conclusions

In summary, we have successfully shown that dc PECVD synthesized MWCNTs exhibit ultra-low reflectance properties over a wide spectral range from UV-to-IR for relatively thin (<10 μm) absorber ensembles. The structural characteristics of the MWCNT absorbers were engineered by controlling the bottom-up synthesis parameters during PECVD which enabled optimization of the optical properties of the nanoabsorbers. Such vertically oriented nanocarbon aero-gel structures were also found to be extremely rugged, withstanding temperatures of up to 400C, without showing signs of any significant degradation in optical properties.

We also demonstrate 1D CNFs and 2D resonant NEMS derived from nanocarbon materials, where the resonance frequency and the quality (Q) factor of the devices are measured experimentally using ultrasensitive optical interferometry. The 1D nanocarbon resonators are formed using CNFs which are prototyped into few-μm-long cantilever-shaped 1D resonators, where the resonance frequency and Qs are extracted from measurements of the undriven thermo mechanical noise spectrum. The thermo mechanical noise measurements yield resonances in the $\sim 3 - 15$ MHz range, with Q of $\sim 200 - 800$. Significant changes in resonance characteristics were observed as a result of electron beam induced amorphous carbon deposition on the CNFs, which suggests that 1D CNF resonators have strong prospects for ultrasensitive mass detection. We also presented results on NEMS resonators based on 2D graphene nanomembranes, which exhibit robust undriven thermo mechanical resonances for the extraction of ultra small strain levels.

References

1. Avouris, P., Chen, Z., Perebeinos, V.: Nat. Nanotechnol. **2**, 605 (2007)
2. Cao, Q., Rogers, J.: Adv. Mater. **21**, 29 (2009)
3. Bachtold, A., Hadley, P., Nakanishi, T., Dekker, C.: Science **294**, 1317 (2001)
4. Zhang, Z., Liang, X., Wang, S., Yao, K., Hu, Y., Zhu, Y., Chen, Q., Zhou, W., Li, Y., Yao, Y., Zhang, J., Peng, L.M.: Nano Lett. **7**, 3603 (2007)
5. Appenzeller, J., Lin, Y., Knoch, J., Chen, Z., Avouris, P.: IEEE Trans. Electron Devices **52**, 2568 (2005)
6. Bandaru, P.R., Daraio, C., Jin, S., Rao, A.M.: Nat. Mater. **4**, 663 (2005)
7. Li, J., Cassell, A., Ng, H.T., Stevens, R., Han, J., Meyyappan, M.: Appl. Phys. Lett. **82**, 2491 (2003)
8. Li, H., Xu, C., Srivastava, N., Banerjee, K.: IEEE Trans. Elect. Dev. **56**, 1799 (2009)
9. Choi, W.B., Chung, D.S., Kang, J.H., Kim, H.Y., Jin, Y.W., Han, I.T., Lee, Y.H., Jung, J.E., Lee, N.S., Park, G.S., Kim, J.M.: Appl. Phys. Lett. **75**, 3129 (1999)

10. Lu, F., Gu, L., Meziani, M.J., Wang, X., Luo, P.G., Veca, L.M., Cao, L., Sun, Y.P.: Adv. Mater. **21**, 139 (2009)
11. Yu, C., Saha, S., Zhou, J., Shi, K., Cassel, A.M., Cruden, B.A., Ngo, Q., Li, J.: J. Heat Transfer-Trans. ASME **128**, 234 (2006)
12. Wei, P., Bao, W., Pu, Y., Lau, C.N., Shi, J.: Phys. Rev. Lett. **102**, 166808 (2009)
13. Ago, H., Petritsch, K., Shaffer, M.S.P., Windle, A.H., Friend, R.H.: Adv. Mater. **11**, 1281 (1999)
14. Homma, Y., Chiashi, S., Kobayashi, Y.: Rep. Prog. Phys. **72**, 066502 (2009)
15. Rueckes, T., Kim, K., Joselevich, E., Tseng, G.Y., Cheung, C.L., Lieber, C.M.: Carbon nanotube-based nonvolatile random access memory for molecular computing. Science **289**, 94–97 (2000)
16. Jang, J.E., Cha, S.N., Choi, Y.J., Kang, D.J., Butler, T. P., Hasko, D.G., Jung, J.E., Kim, J.M., Amaratunga, G.A.J.: Nat. Nanotech. **3**, 26 (2008)
17. Kaul, A.B.., Wong, E.W., Epp, L., Hunt, B.D.: Electromechanical carbon nanotube switches for high-frequency applications. Nano Lett. **6**, 942–947 (2006)
18. Yu, M.F., Lourie, O., Dyer, M.J., Moloni, K., Kelly, T. F., Ruoff, R.S.: Science **287**, 637 (2000)
19. Dresselhaus, M.S., Dresselhaus, G., Avouris, P. (eds.): Carbon Nanotubes. Springer, Berlin (2001)
20. Kroto, H.W., Heath, J.R., O'Brien, S.C., Curl, R.F., Smalley, R.E.: Nature **318**, 162 (1985)
21. Iijima, S.: Nature **354**, 56 (1991)
22. Iijima, S., Ichihashi, T.: Nature **363**, 603 (1993)
23. Han, M., Ozyilmaz, B., Zhang, Y., Kim, P.: Phys. Rev. Lett. **98**, 206805 (2007)
24. Geim, A.K., Novoselov, K.S.: Nat. Mater. **6**, 183 (2007)
25. Dai, L. (ed): In Carbon Nanotechnology: Recent Developments in Chemistry, Physics, Materials Science and Device Applications. Elsevier, Oxford (2006)
26. Pierson, H.O.: Handbook of Carbon, Graphite, Diamond, and Fullerenes: Properties, Processing and Applications. Noyes Publications, Park Ridge (1993)
27. Krüger, A.: Carbon Materials and Nanotechnology. Wiley, Weinheim (2010)
28. Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Zhang, Y., Dubonos, S.V., Grigorieva, I.V., Firsov, A.A.: Science **306**, 666 (2004)
29. Boehm, H.P., Clauss, A., Fischer, G.O., Hoffmann, U.: Z. Naturforsch. B **17**, 150 (1962)
30. Morgan, A.E., Somarjai, G.A.: Surf. Sci. **12**, 405 (1968)
31. May, J.W.: Surf. Sci. **17**, 267 (1969)
32. Gamo, Y., Nagashima, A., Wakabayashi, M., Terai, M., Oshima, C.: Surf. Sci. **374**, 61 (1997)
33. Dedkov, Y.S., Shikin, A.M., Adamchuk, V.K., Molodtsov, S.L., Laubschat, C., Bauer, A., Kaindl, G.: Phys. Rev. B **64**, 035405 (2001)
34. Batzill, M.: Surf. Sci. Rep. **67**, 83 (2012)
35. Zhang, Y.B., Tan, Y.W., Stormer, H.L., Kim, P.: Nature **438**, 201 (2005)
36. Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Katsnelson, M.I., Grigorieva, I.V., Dubonos, S.V., Firsov, A.A.: Nature **438**, 197 (2005)
37. Zhan, D., Yan, J., Lai, L., Ni, Z., Liu, L., Shen, Z.: Adv. Mat. **24**, 4055 (2012)
38. Evans, J.W., Thiel, P.A., Bartelt, M.C.: Sur. Sci. Rep. **61**, 1 (2006)
39. Landau, L.D.: Phys. Z. Sowjetunion **11**, 26 (1937)
40. Mermin, N.D.: Phys. Rev. **176**, 250 (1968)
41. Wallace, P.R.: Phys. Rev. **71**, 622 (1947)
42. Novoselov, K.S., Fal'ko, V.I., Colombo, L., Gellert, P. R., Schwab, M.G., Kim, K.: Nature **490**, 192 (2012)
43. Wang, X., Zhi, Z., Mullen, K.: Nano Lett. **8**, 323 (2009)
44. Matyba, P., Yamaguchi, H., Eda, G., Chhowalla, M., Edman, L., Robinson, N.D.: ACS Nano **4**, 637 (2010)
45. Miao, X., Tongay, S., Petterson, M.K., Berke, K., Rinzler, A.G., Appleton, B.R., Hebard, A.F.: Nano Lett. **12**, 2745 (2012)
46. Stoller, M.D., Park, S., Zhu, Y., An, J., Ruoff, R.S.: Nano Lett. **8**, 3498 (2008)
47. Schedin, F., Geim, A.K., Morozov, S.V., Hill, E.W., Blake, P., Katsnelson, M.I., Novoselov, K.S.: Nat. Mater. **6**, 652 (2007)
48. Robinson, J.T., Perkins, F.K., Snow, E.S., Wei, Z., Sheehan, P.E.: Nano Lett. **8**, 3137 (2008)
49. Hill, E.W., Geim, A.K., Novoselov, K., Schedin, F., Blake, P.: IEEE Trans. Mag. **42**, 2694 (2006)
50. Heersche, H.B., Jarillo-Herrero, P., Oostinga, J.B., Vandersypen, L.M.K., Morpurgo, A.F.: Nature **446**, 56 (2007)
51. Stankovich, S., Piner, R.D., Chen, X., Wu, N., Nguyen, S.T., Ruoff, R.S.: J. Mater. Chem. **16**, 155 (2006)
52. Stankovich, S., Dikin, D.A., Dommett, G.H.B., Kohlhaas, K.M., Zimney, E.J., Stach, E.A., Piner, R.D., Nguyen, S.T., Ruoff, R.S.: Nature **442**, 282 (2006)
53. Huynh, W.U., Dittmer, J.J., Alivisatos, A.P.: Hybrid nanorod-polymer solar cells. Science **295**(5564), 2425–2427 (2002)
54. Nair, R.R., Blake, P., Grigorenko, A.N., Novoselov, K. S., Booth, T.J., Stauber, T., Peres, N.M.R., Geim, A. K.: Fine structure constant defines visual transparency of graphene. Science **320**(5881), 1308–1308 (2008)
55. Derkacs, D., Lim, S.H., Matheu, P., Mar, W., Yu, E.T.: Appl. Phys. Lett. **89**, 093103 (2006)
56. Hu, L., Chen, G.: Nano Lett. **7**(11), 3249–3252 (2007)
57. Tsakalakos, L., Balch, J., Fronheiser, J., Shih, M.Y., LeBoeuf, S.F., Pietrozykowski, M., Codella, P.J., Sulima, O., Rand, J., Kumar, A.D., Korevaar, B.A.: Strong broadband optical absorption in silicon nanowire films. J. Nanophotonics **1**, 013552 (2007)
58. Peng, K., Wu, Y., Fang, H., Zhong, X., Xu, Y., Zhu, J.: Uniform, axial-orientation alignment of one-dimensional single-crystal silicon nanostructure arrays. Angew. Chem. Int. Ed. **44**, 2737–2742 (2005)

59. Garcia-Vidal, F.J., Pitarke, J.M., Pendry, J.B.: Effective medium theory of the optical properties of aligned carbon nanotubes. Phys. Rev. Lett. **78**, 4289–4292 (1997)
60. Yang, Z.-P., Ci, L., Bur, J.A., Lin, S.-Y., Ajayan, P.M.: Nano Lett. **8**, 446 (2008)
61. Mizuno, K., Ishii, J., Kishida, H., Hayamizu, Y., Yasuda, S., Futaba, D.N., Yumura, M., Hata, K.: Proc. Natl. Acad. Sci. U. S. A. **106**, 6044 (2009)
62. Hata, K., Futaba, D.N., Mizuno, K., Namai, T., Yumura, M., Iijima, S.: Science **306**, 1362 (2004)
63. Fan, S., Chapline, M.G., Franklin, N.R., Tombler, T. W., Cassell, A.M., Dai, H.: Science **283**, 512 (1999)
64. Andrews, R., Jacques, D., Rao, A.M., Derbyshire, F., Qian, D., Fan, X., Dickey, E.C., Chen, J.: Chem. Phys. Lett. **303**, 467 (1999)
65. Eriksson, P., Andersson, J.Y., Stemme, G.: Phys. Scr. **T54**, 165 (1994)
66. Lehman, J.H., Deshpande, R., Rice, P., To, B., Dillon, A.C.: Infrared Phys. Tech. **47**, 246 (2006)
67. Kaul, A.B.., Coles, J.B., Megerian, K.G., Eastwood, M., Green, R.O., Bandaru, P.R.: Small **9**, 1058 (2013)
68. Nunes, C., Teixeira, V., Collares-Pereira, M., Monteiro, A., Roman, E., Martin-Gago, J.: Vacuum **67**, 623 (2002)
69. Murakami, Y., Chiashi, S., Miyauchi, Y., Maruyama, S.: Jpn. J. Appl. Phys. **43**, 1221 (2004)
70. Sato, S., Kawabata, A., Kondo, D., Nihei, M., Awano, Y.: Chem. Phys. Lett. **402**, 149 (2005)
71. Advena, D.J., Bly, V.T., Cox, J.T.: Appl. Opt. **32**, 1136 (1993)
72. Johnson, C.E.: Metal Finish. **78**, 21 (1980)
73. Kodama, S., Horiuchi, M., Kuni, T., Kuroda, K.: IEEE Trans. Instrum. Meas. **39**, 230 (1990)
74. Lee, C., Bae, S., Mobasser, S., Manohara, H.: Nano Lett. **5**, 2438 (2005)
75. Zhu, J., Yu, Z., Burkhard, G.F., Hsu, C.-M., Connor, S. T., Xu, Y., Wang, Q., McGehee, M., Fan, S., Cui, Y.: Nano Lett. **9**, 279 (2009)
76. Feng, P.X.-L.: Nanoelectromechanical switching devices: scaling toward ultimate energy efficiency and longevity (invited talk). The 3rd Berkeley symposium on energy efficient electronic systems (E3S), Berkeley, 28–29 Oct 2013, pp. 1–2, doi:10.1109/E3S.2013.6705881
77. Naik, A.K., Hanay, M.S., Hiebert, W.K., Feng, X.L., Roukes, M.L.: Towards single-molecule nanomechanical mass spectrometry. Nature Nanotech. **4**, 445–450 (2009)
78. Rugar, D., Budakian, R., Mamin, H.J., Chui, B.W.: Single spin detection by magnetic resonance force microscopy. Nature **430**, 329–332 (2004)
79. Rocheleau, T., Ndukum, T., Macklin, C., Hertzberg, J. B., Clerk, A.A., Schwab, K.C.: Preparation and detection of a mechanical resonator near the ground state of motion. Nature **463**, 72–75 (2010)
80. Feng, X.L., He, R., Yang, P., Roukes, M.L.: Very high frequency silicon nanowire electromechanical resonators. Nano Lett. **7**, 1953–1959 (2007)
81. Huang, X.M.H., Feng, X.L., Zorman, C.A., Mehregany, M., Roukes, M.L.: VHF, UHF and microwave frequency nanomechanical resonators. New J. Phys. **7**, 247 (2005)
82. Bunch, J.S., van der Zande, A.M., Verbridge, S.S., Frank, I.W., Tanenbaum, D.M., Parpia, J.M., Craighead, H.G., McEuen, P.L.: Electromechanical resonators from graphene sheets. Science **315**, 490–493 (2007)
83. Moser, J., Güttinger, J., Eichler, A., Esplandiu, M.J., Liu, D.E., Dykman, M.I., Bachtold, A.: Ultrasensitive force detection with a nanotube mechanical resonator. Nature Nanotech. **8**, 493–496 (2013)
84. Eriksson, A., Lee, S., Sourab, A., Issacsson, A., Kaunisto, R., Kinaret, J.M., Campbell, E.E.B.: Direct transmission detection of tunable mechanical resonance in an individual carbon nanofiber relay. Nano Lett. **8**, 1224–1228 (2008)
85. Kaul, A.B.., Megerian, K.G., Jennings, A.T., Greer, J. R.: In situ characterization of vertically oriented carbon nanofibers for three-dimensional nano-electromechanical device applications. Nanotechnology **21**, 315501 (2010)
86. Zhang, Y., Chang, A., Cao, J., Wang, Q., Kim, W., Li, Y.M., Morros, N., Yenilmez, E., Kong, J., Dai, H.J.: Electric-field directed growth of aligned single-walled carbon nanotubes. Appl. Phys. Lett. **79**, 3155–3157 (2001)
87. Teo, K.B.K., Chhowalla, M., Amaratunga, G.A., Milne, W.I., Pirio, G., Legagneux, P., Wyczisk, F., Olivier, J., Pribat, D.: Characterization of plasma-enhanced chemical vapor deposition carbon nanotubes by Auger electron spectroscopy. J. Vac. Sci. Technol. B **20**, 116–121 (2002)
88. Melechko, A.V., Merkulov, V.I., McKnight, T.E., Guillorn, M.A., Klein, K.L., Lowndes, D.H., Simpson, M.L.: J. Appl. Phys. **97**, 041301 (2005)
89. Lee, J., Feng, P.X.-L., Kaul, A.B..: Characterization of plasma synthesized vertical carbon nanofibers for nanoelectronics applications. MRS Proceedings **1451** (2012 MRS Spring Meeting, Symposium EE – Nano Carbon Materials & Devices, San Francisco, 9–13 April 2012), pp. 117–122, doi:10.1557/opl.2012.922
90. Mayorov, A.S., Gorbachev, R.V., Morozov, S.V., Britnell, L., Jalil, R., Ponomarenko, L.A., Blake, P., Novoselov, K.S., Watanabe, K., Taniguchi, T., Geim, A.K.: Nano Lett. **11**, 2396 (2011)
91. Moser, J., Barreiro, A., Bachtold, A.: Appl. Phys. Lett. **91**, 163513 (2007)
92. Lee, C., Wei, X.D., Kysar, J.W., Hone, J.: Science **321**, 385 (2008)
93. Balandin, A.A., Ghosh, S., Bao, W.Z., Calizo, I., Teweldebrhan, D., Miao, F., Lau, C.N.: Nano Lett. **8**, 902 (2008)
94. Balandin, A.A.: Nature Mater. **10**, 569 (2011)
95. Bunch, J.S., Verbridge, S.S., Alden, J.S., van der Zande, A.M., Parpia, J.M., Craighead, H.G., McEuen, P.L.: Nano Lett. **8**, 2458 (2008)

96. Lee, G.-H., Cooper, R.C., An, S.J., Lee, S., van der Zande, A., Petrone, N., Hammerberg, A.G., Lee, C., Crawford, B., Oliver, W., Kysar, J.W., Hone, J.: High-strength chemical-vapor-deposited Graphene and grain boundaries. Science **340**, 1073–1076 (2013)
97. Lee, J., Feng, P.X.-L.: High frequency graphene nanomechanical resonators and transducers. In: *Proceedings of the IEEE international frequency control symposium* (*IFCS 2012*), Baltimore, 21–24 May 2012, pp. 1–7, doi:10.1109/FCS.2012.6243742

Nanostructures and Characteristics of Graphene and Graphene Oxides

Tian-Ling Ren, He Tian, Wen-Tian Mi and Yi Yang
Institute of Microelectronics, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, Haidian, China

Synonyms

Graphene laser writing; Laser-reduced graphene oxide; Laser-scribed graphene

Definition

The laser scribing technology is the method via which the graphene oxide could be converted into reduced graphene oxide by the laser from the LightScribe DVD driver on specific pattern precisely.

Background

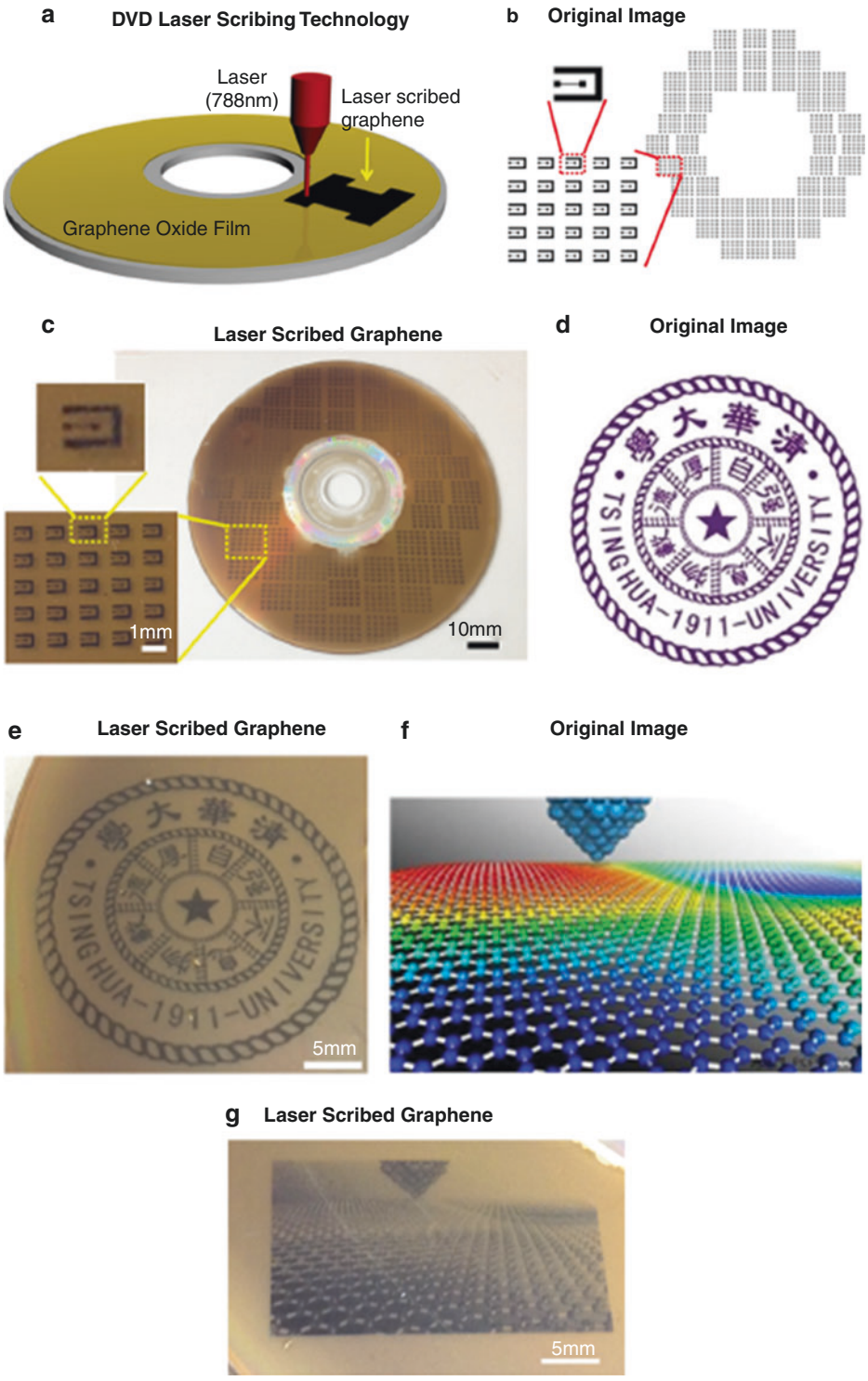
Large-scale graphene film has been prepared by chemical method at low cost. The graphene oxide (GO) dispersion could be obtained by the Hummers method [1], and it could be used to prepare the graphene oxide film via various ways [2] such as spin coating, spray coating, and vacuum filtration. Ultimately, the graphene film could be obtained by reduction. Since there would be some oxygen functional groups left in the film after reduction, it is also called the reduced graphene oxide (rGO) film. The large-scale GO film prepared via vacuum filtration method was firstly reported by Eda, G., et al. (Fig. 1), which revealed the application potential of rGO in transparent electrodes and transistors [3].

Though large-scale graphene film could be obtained by various methods, the laser scribe method [4, 5] was raised to achieve the precise pattern of graphene, which had never been realized before. The graphene pattern could be obtained by single-step laser scribing simply and cost-effectively, even on flexible substrates.

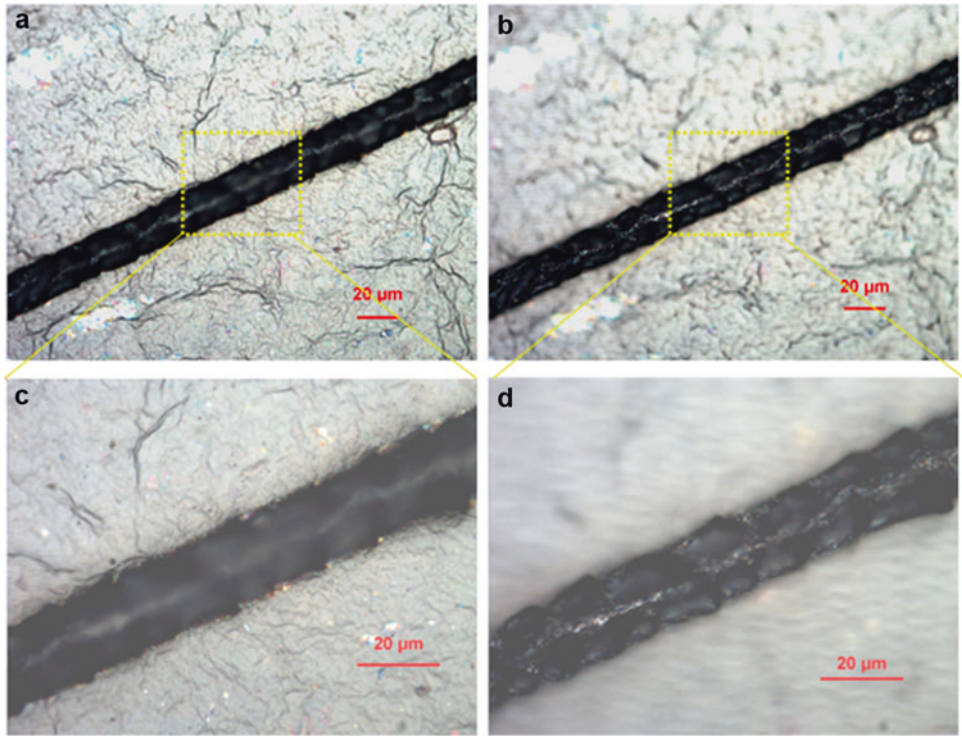
The laser (wavelength 788 nm, power 5 mW) in the LightScribe DVD driver could reduce the GO to rGO (Fig. 2a). And the laser could be controlled by software and transfer the input



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 1 The GO film via vacuum filtration method (Figure reproduced with permission from Ref. [3], Nature Publishing Group)



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 2 Samples of laser-scribed graphene (Figure reproduced with permission from Ref. [6], Nature Publishing Group)

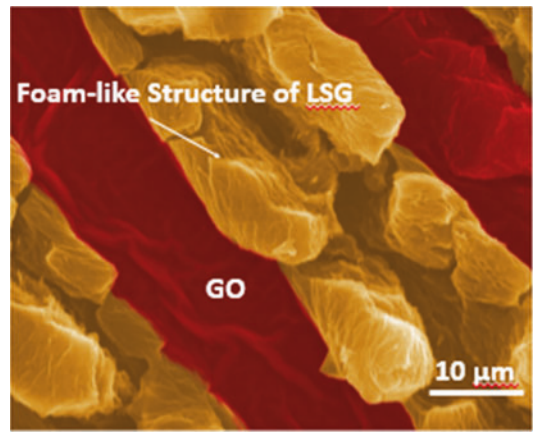


Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 3 The optical image of laser-scribed graphene micro-ribbon

image to graphene pattern. The in-plane transistor design is shown in Fig. 2b, and it could be revealed on the graphene film in ~25 min by laser scribing. The golden film in Fig. 2c is graphene oxide and the black part is multilayer graphene. Complicated graphs could be also converted into graphene patterns, such as the Tsinghua logo shown in Fig. 2d–e. And a colored picture could be converted into grayscale graphene pattern (Fig. 2f–g), in which the grayscale corresponds with the degree of reduction.

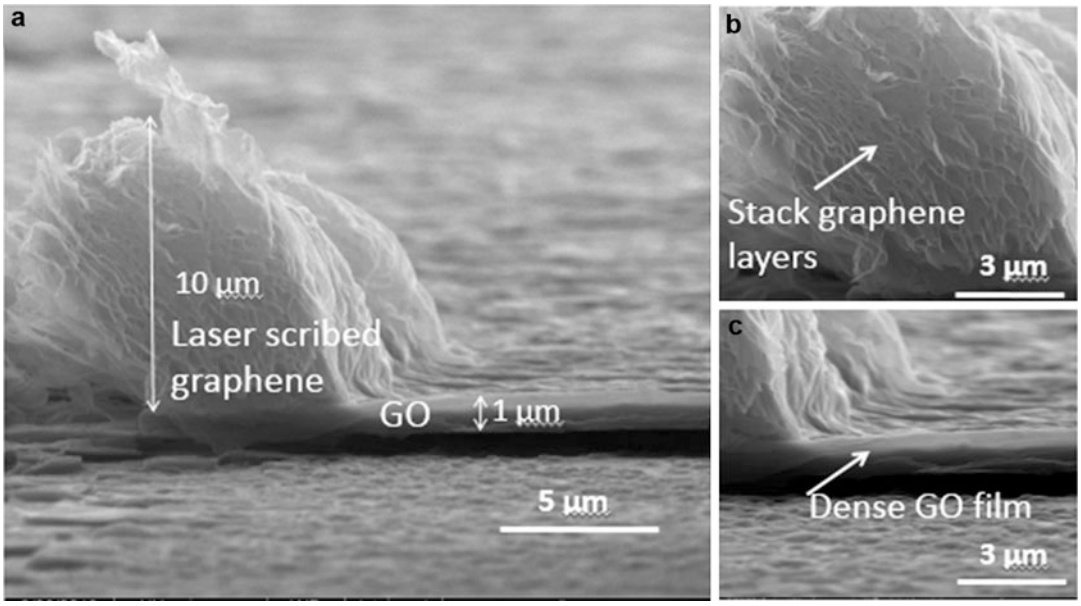
Characterization of Laser-Scribed Graphene

The spot size of the laser, which we used here is 20 μm, determines the precision of the laser-scribed graphene. And it could be observed under a microscope that the laser-scribed graphene consists of long micro-ribbon (Fig. 3). Figure 3a–b is the optical image of laser-scribed

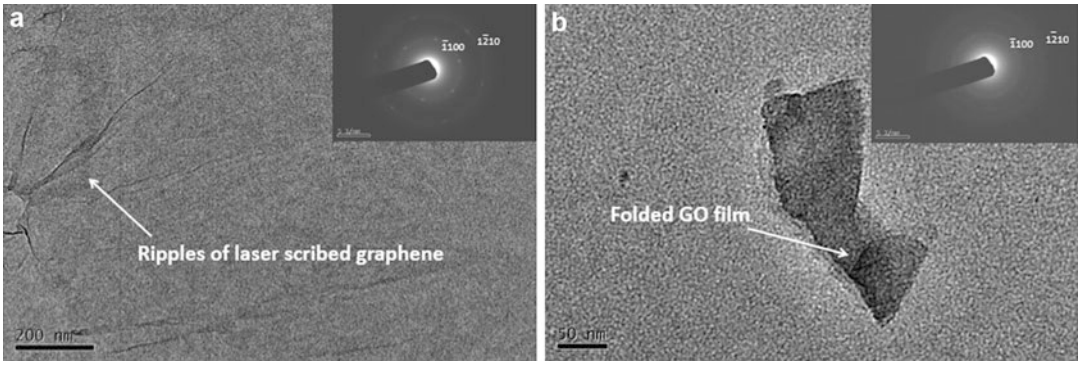


Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 4 The colored SEM image of laser-scribed graphene micro-ribbon (Figure reproduced with permission from Ref. [7], Nature Publishing Group)

graphene micro-ribbon under low magnification. In high-light conditions, the GO is white and the rGO is black. Increasing the magnification, a white line in the center of rGO micro-ribbon



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 5 The lateral view of the laser-scribed graphene 3D structure (Figure reproduced with permission from Ref. [6], Nature Publishing Group)

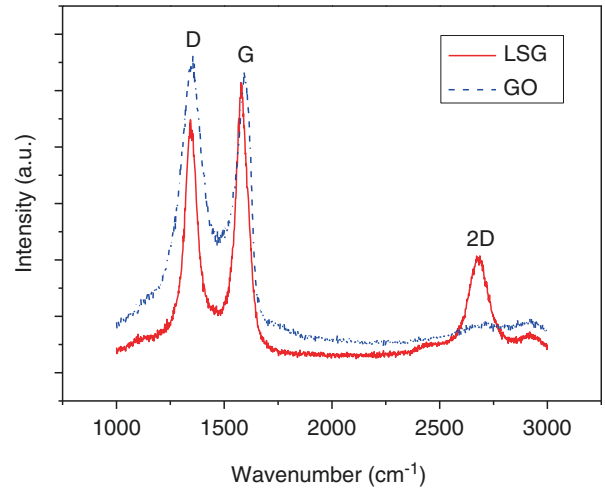


Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 6 The TEM image of (**a**) the laser-scribed graphene and (**b**) GO (Figure reproduced with permission from Ref. [6], Nature Publishing Group)

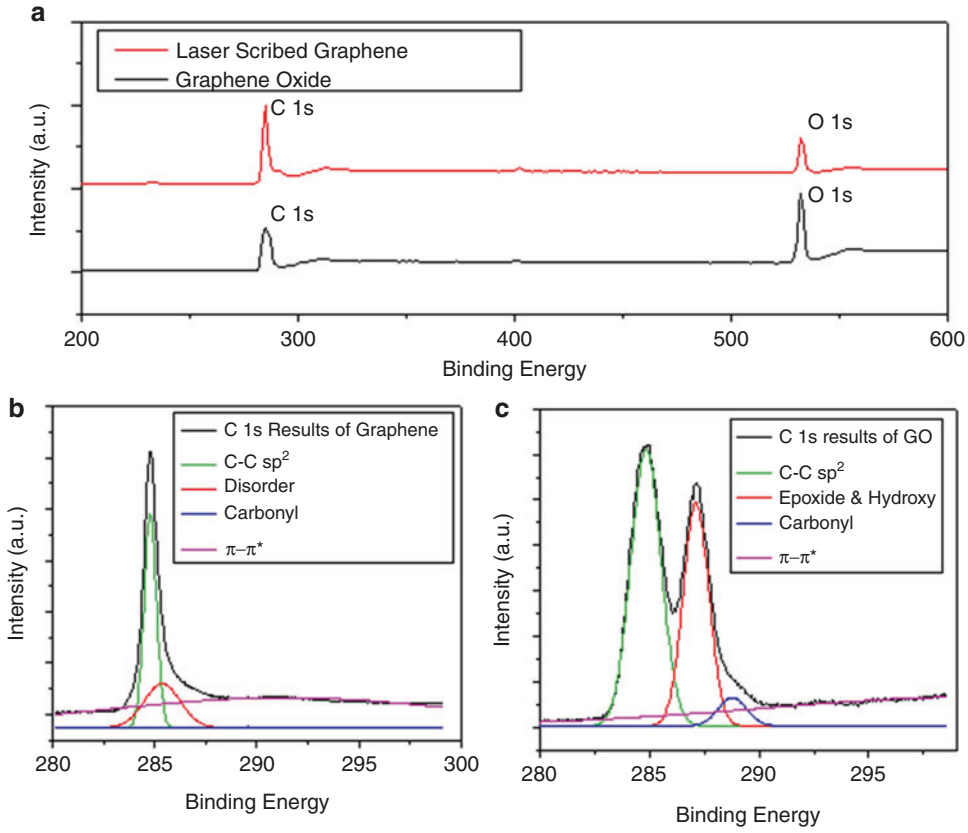
could be observed. Remarkably, different focus planes of GO and rGO, which are clearly shown in Fig. 3c–d, suggest the 3D structure of laser-scribed graphene micro-ribbon.

The colored SEM image of laser-scribed graphene reveals the 3D structure more clearly. The red part in the image is GO and the yellow one is rGO. The contrast of light and shade in the image of rGO ribbon indicates the valley in the center and the elevation on the two sides, which resemble the foam structure (Fig. 4).

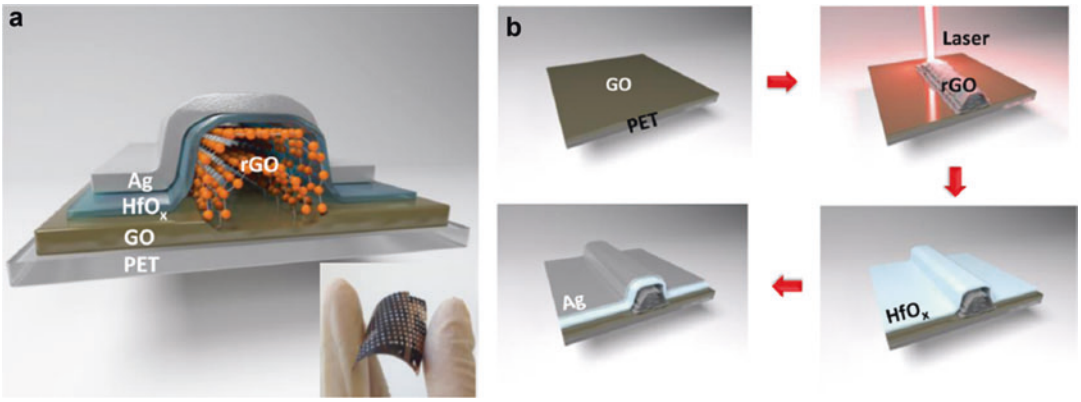
The structural change of graphene before and after the laser scribing is shown in Fig. 5. The thickness of the laser-scribed graphene is 10 μm, which is 10 times bigger than that of the original GO film. It could be observed in the magnified image that the rGO is comprised of loose stacked graphene layers and the GO film is made of dense



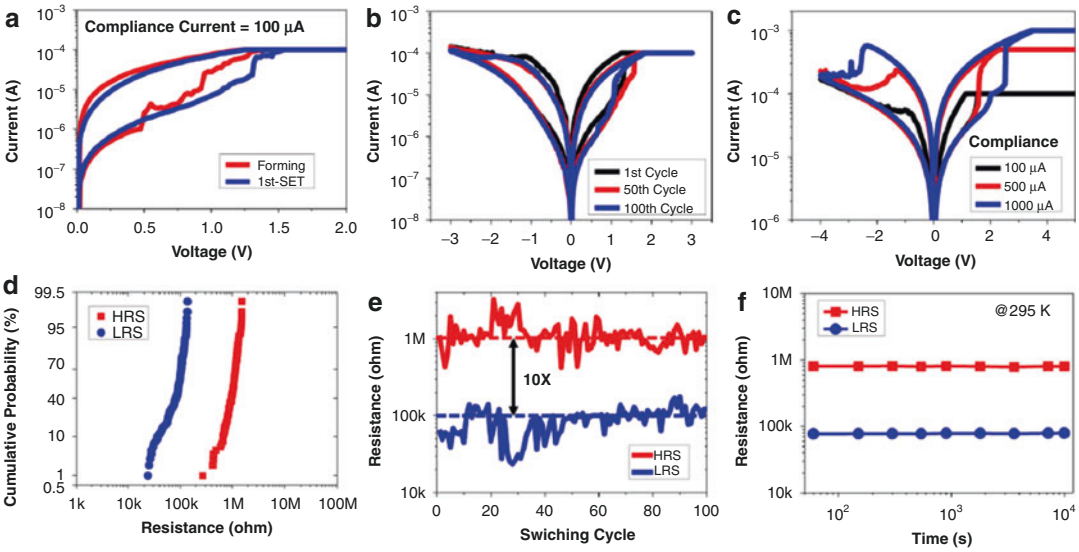
Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 7 The Raman spectrum of GO and rGO



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 8 The XPS spectrum of GO and rGO



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 9 Flexible graphene RRAM (Figure reproduced with permission from Ref. [8], American Chemical Society)

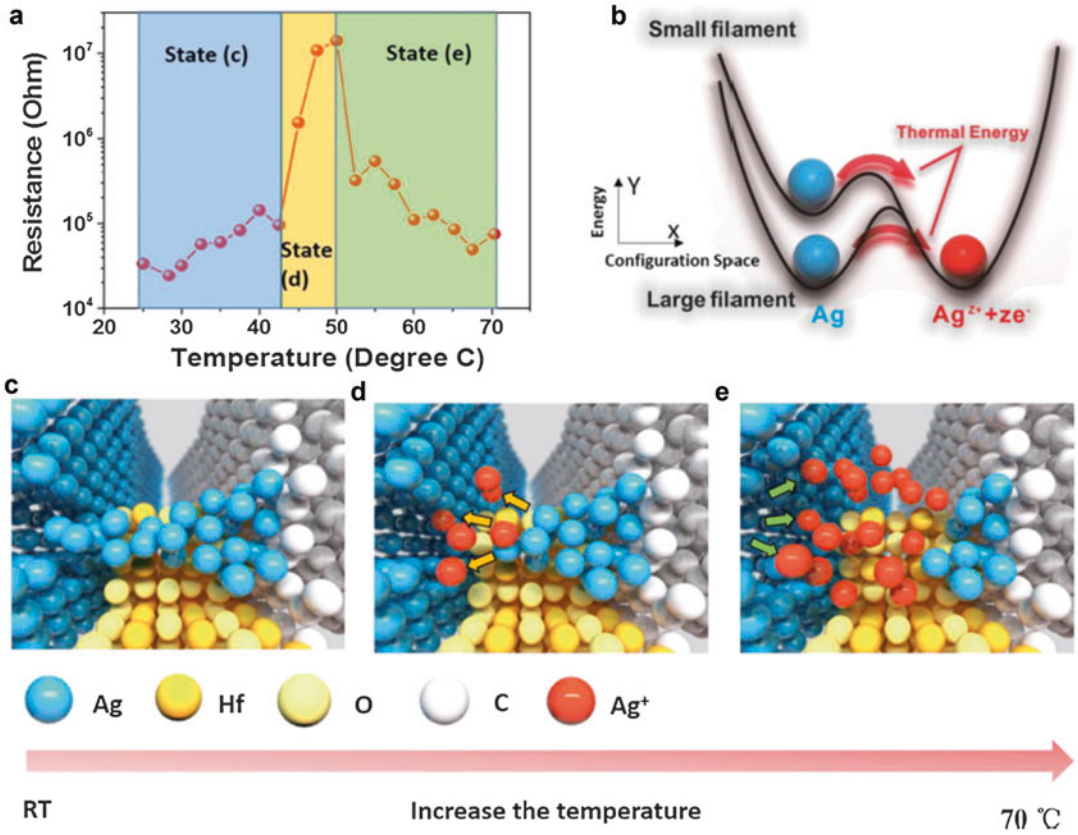


Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 10 Measurement results of the flexible graphene RRAM (Figure reproduced with permission from Ref. [8], American Chemical Society)

layers. On the contrary, the thickness of the graphene film would decrease after traditional thermal reduction, which results from the loss of oxygen functional groups. However, the high-energy laser beam would convert the oxygen functional groups into the gas containing oxygen, which finally leads to the loose structure of rGO.

The diffraction pattern of the laser-scribed graphene (Fig. 6a) vindicates the hexagonal cellular lattice of graphene. On the contrary, the absence of crystal lattice in the diffraction pattern of the stacked GO film (Fig. 6b) indicates the broken lattice structure of GO, which results from the lattice defects due to the existence of oxygen functional groups.

The change of the lattice structure could be also revealed in the Raman spectrum (Fig. 7). The decrease of the D peak intensity suggests that part of the C-C bonds were repaired after the laser



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 11 The working principle of the graphene RRAM (Figure reproduced with permission from Ref. [8], American Chemical Society)

scribing. However, since there are still some defects left in the laser-scribed graphene, the D peak remains obvious. Besides, the left shift of the G peak indicates that the doping level of graphene decreased after the reduction, resulting from the removal of oxygen. And the appearance of obvious 2D peak after the laser scribing indicates the stacked multilayer graphene generated after reduction.

The component of the laser-scribed graphene is shown in the XPS spectrum (Fig. 8). The oxygen peak decreased obviously after the reduction. And the peaks of C-C sp^2 bond and π bond are enhanced.

Application

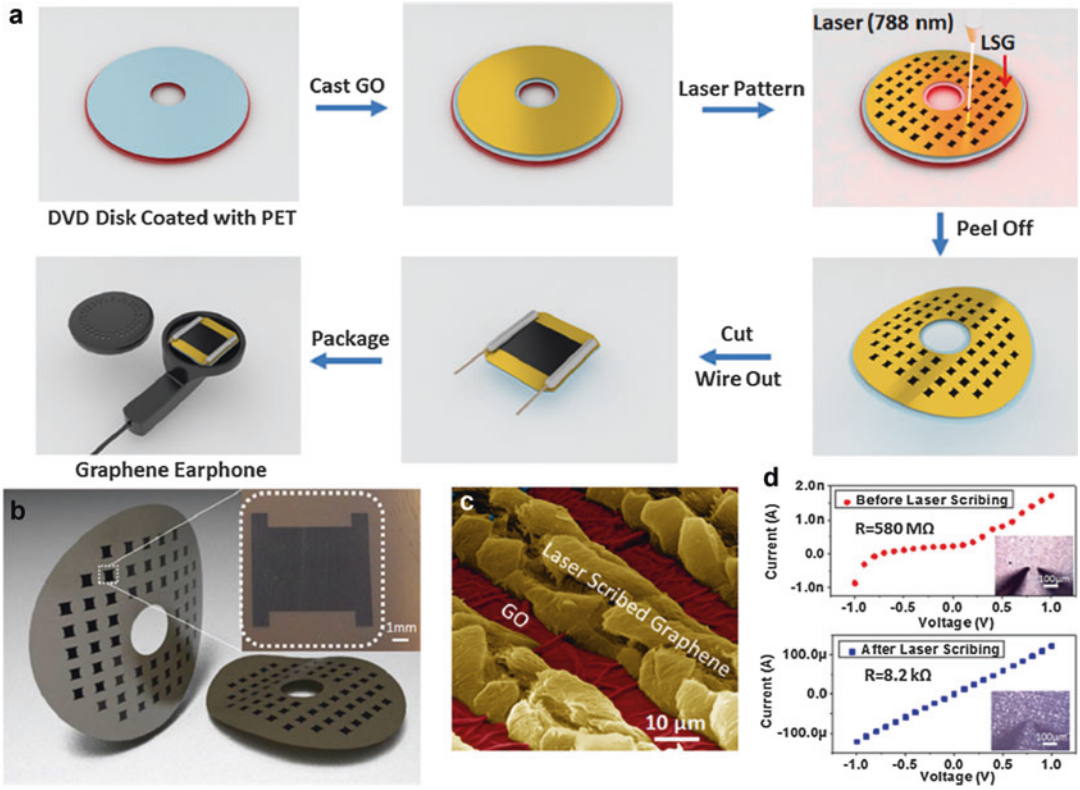
Graphene Resistive Random Access Memory

Using the laser scribing technology, the flexible graphene RRAM [8] with the structure like FinFET (Fig. 9a) could increase the robustness of the device.

As shown in Fig. 9b, the GO dispersion was firstly spin coated on the flexible PET substrates, and the GO could be converted into rGO in the specific location. Then, the HfO_x and the top electrode are deposited consequently.

The excellent performance of the flexible graphene RRAM is clearly shown in Fig. 10, including the forming-free feature (Fig. 10a), repeatability (Fig. 10b), multivalued feature (Fig. 10c), uniformity (Fig. 10d), feasible resistance window (Fig. 10e), and stability (Fig. 10f).

The working principle of the graphene RRAM could be revealed by the temperature characteristic analysis. When temperature rises, the increase of the resistance indicates that the conductive filaments are made of silver. Keeping an increasing temperature, some silver filament would

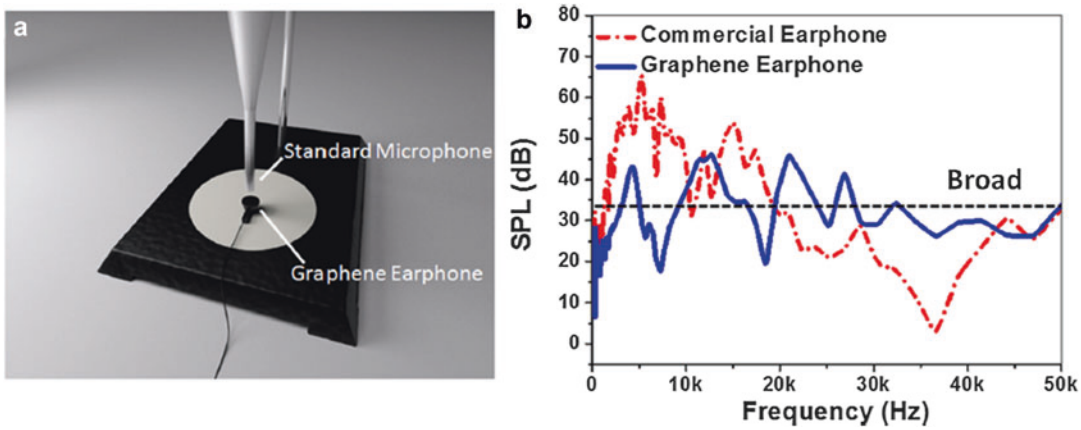


Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 12 Graphene earphone (Figure reproduced with permission from Ref. [9], American Chemical Society)

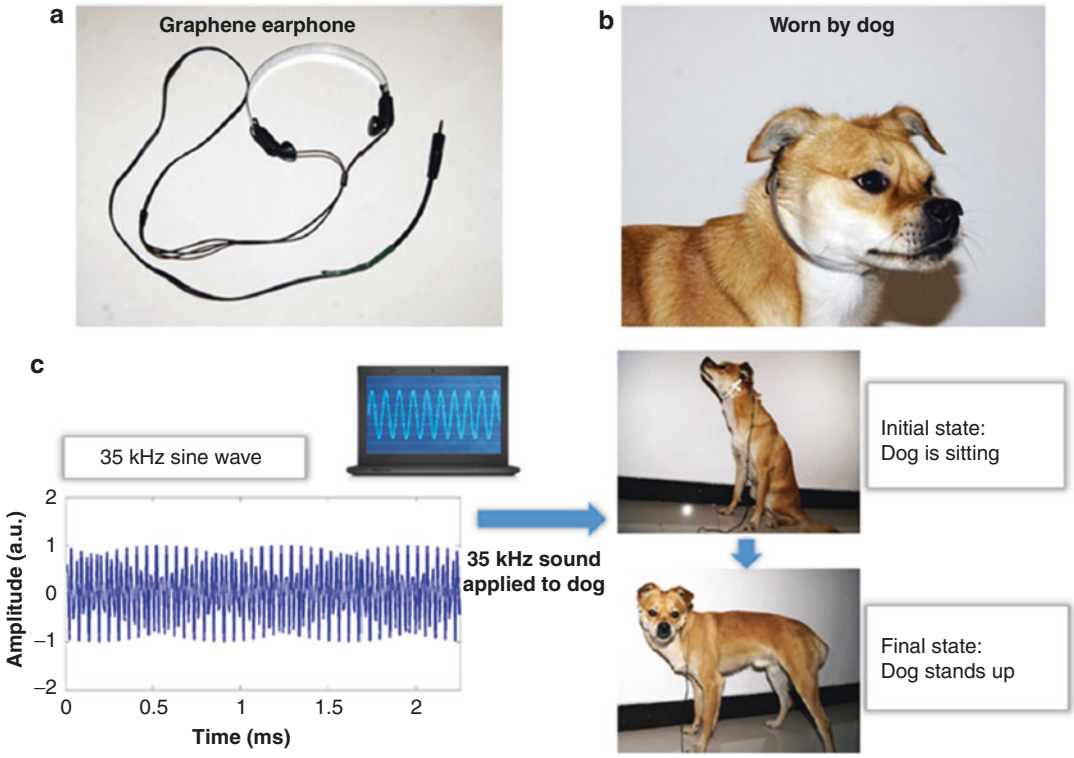
N

Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 13 Packaged graphene earphone (Figure reproduced with permission from Ref. [9], American Chemical Society)





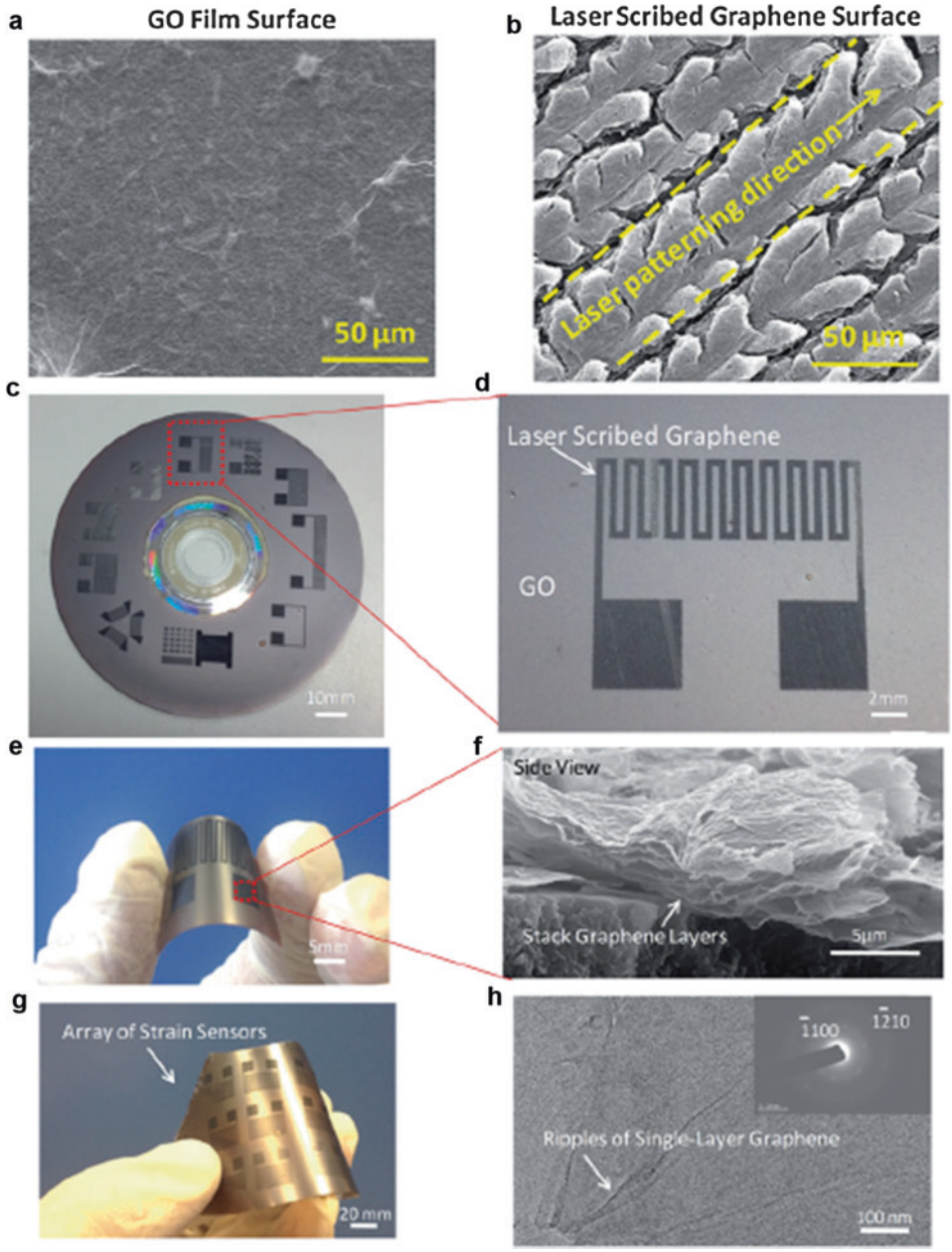
Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 14 The performance test of graphene earphone (Figure reproduced with permission from Ref. [9], American Chemical Society)



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 15 The graphene earphone as a tool for interspecies communication (Figure reproduced with permission from Ref. [9], American Chemical Society)

transfer to silver ions, leading to the increase of the resistance. As the silver ions increase, which are also conductive, the resistance would fall again (Fig. 11).

The working principle could be confirmed by the electrical measurement. The threshold voltage is only 0.5 V when the top electrode is silver. But for the device with Pt top

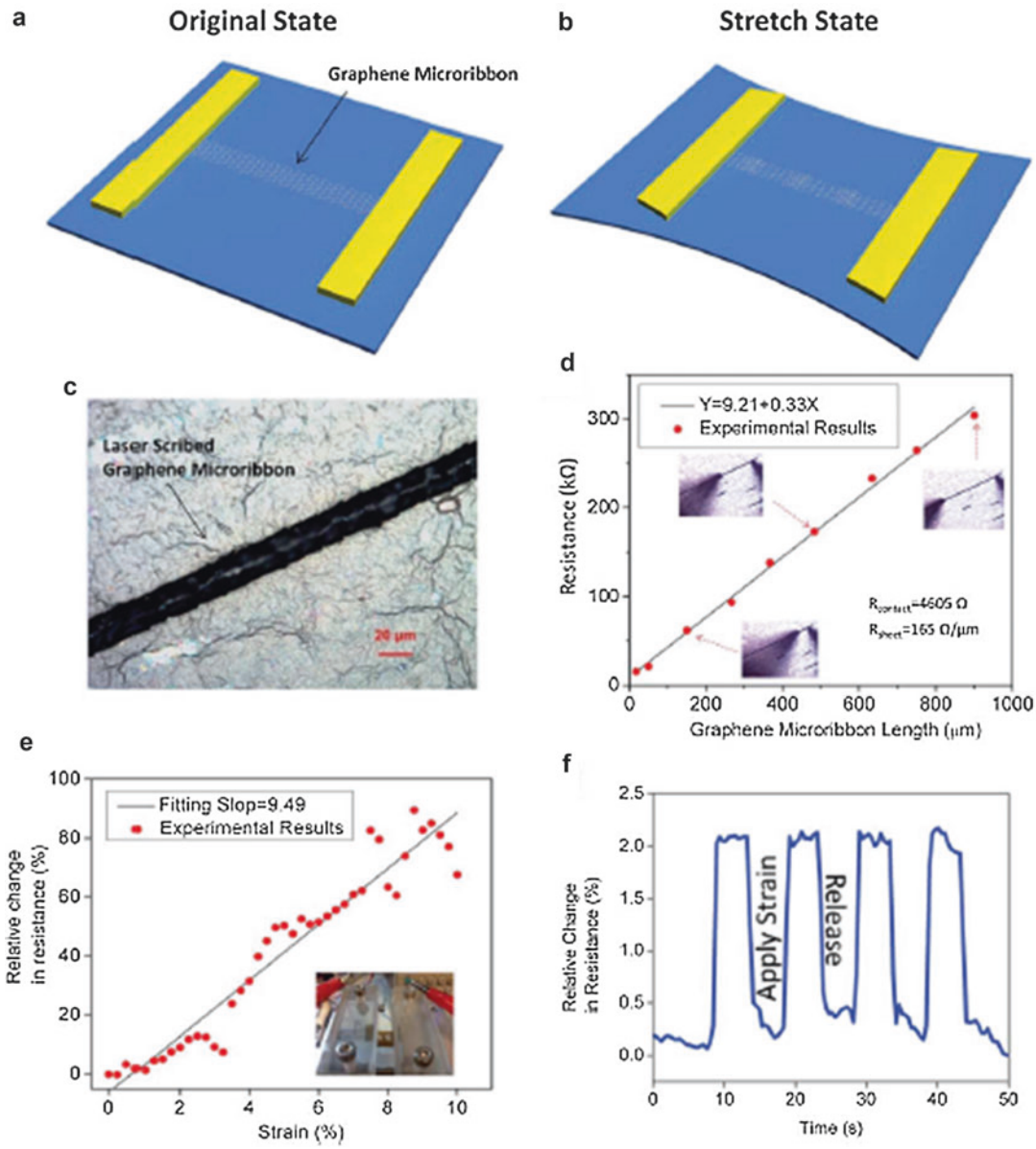


Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 16 Graphene strain sensor (Figure reproduced with permission from Ref. [10], The Royal Society of Chemistry)

electrodes, the threshold voltage is 2 V, which suggests that the breakdown voltage of the HfO_x is 2 V. Therefore, the HfO_x could not be broken through under 0.5 V to form the oxygen filament, which vindicates that the filament consists of silver.

Graphene Earphone

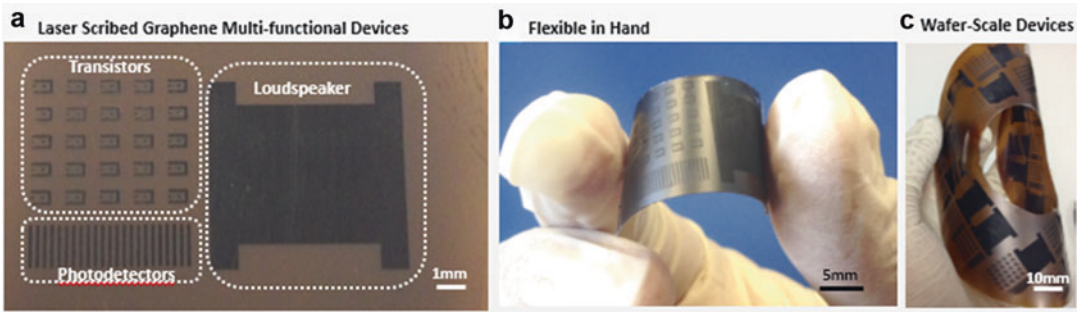
The flexible graphene earphone [9] could be fabricated based on the laser scribing technology. Compared to traditional earphones, the graphene earphone has the characteristics of flexibility and ultralow thickness.



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 17 Test results of the graphene strain sensor (Figure reproduced with permission from Ref. [10], The Royal Society of Chemistry)

With the laser scribing technology, the graphene sound source array could be fabricated on the PET substrate (Fig. 12a–b). The sound source was wired out and packaged into commercial earphone case (Fig. 13). The I-V curve indicates the good conductivity of the graphene device.

The acoustic properties of the graphene earphone are tested on the measurement system shown in Fig. 14a. And the test result, the sound press level frequency spectrum, indicates that the response of the graphene earphone is flatter than that of commercial earphones, especially in the ultrasonic band. Moreover, the frequency



Nanostructures and Characteristics of Graphene and Graphene Oxides, Fig. 18 Wafer-scale graphene device integration (Figure reproduced with permission from Ref. [6], Nature Publishing Group)

spectrum of the graphene earphone could cover not only the audibility band (from 20 Hz to 20 kHz for human being) but also the ultrasonic band (from 20 kHz to 50 kHz), which is the audibility zone for some animals. It means that the graphene earphone could be the tool for inter-species communication, of which the example is shown in Fig. 14. The 35 kHz acoustic signal is conveyed to the dog via graphene earphones, and the dog stood up as soon as the signal is delivered (Fig. 15).

Graphene Strain Sensor

Strain sensors are widely used in various areas, but most of them are based on the rigid substrates, which limit the sensitivity of the sensor. Based on the laser scribing technology, the flexible graphene strain sensor [10] is achieved with the greatly enhanced sensitivity.

The pattern of the graphene strain sensor array could be fabricated on the GO film via the laser scribing technology on the flexible PET substrates with low cost and high speed (Fig. 16). The laser-scribed graphene is made of loose stacked graphene layers, which could greatly enhance the sensitivity of the graphene strain sensor.

The test results of the graphene strain sensor are shown in Fig. 17. The sheet resistance was observed while stretching the sensor. The sensitivity calculated from the data is 9.49, and Fig. 17f indicated the rapid response of the graphene strain sensor.

Wafer-Scale Graphene Devices Integration

Since a variety of graphene devices could be easily patterned on the GO film via the laser scribing technology, it is possible to integrate multifunctional graphene devices [6] on a whole wafer (Fig. 18). As shown in Fig. 18a, the graphene transistor, photodetector, and loudspeaker were fabricated on the same chip. And the substrates could be even flexible (Fig. 18b). The fabrication process of a whole wafer could be finished in 25 min.

N

References

1. Hummers, W., Offeman, R.: Preparation of graphitic oxide. J. Am. Chem. Soc. **80**, 1339–1339 (1958)
2. Dreyer, D.R., Park, S., Bielawski, C.W., Ruoff, R.S.: The chemistry of graphene oxide. Chem. Soc. Rev. **39**, 228–240 (2010)
3. Eda, G., Fanchini, G., Chhowalla, M.: Large-area ultrathin films of reduced graphene oxide as a transparent and flexible electronic material. Nat. Nanotechnol. **3**, 270–274 (2008)
4. El-Kady, M.F., Strong, V., Dubin, S., Kaner, R.B.: Laser scribing of high-performance and flexible graphene-based electrochemical capacitors. Science **335**, 1326–1330 (2012)
5. El-Kady, M.F., Kaner, R.B.: Scalable fabrication of high-power graphene micro-supercapacitors for flexible and on-chip energy storage. Nat. Commun. **4**, 1475 (2013)
6. Tian, H., Yang, Y., Xie, D., Cui, Y.-L., Mi, W.-T., Zhang, Y., Ren, T.-L.: Wafer-scale integration of graphene-based electronic, optoelectronic and electroacoustic devices. Sci. Rep. **4**, 3598 (2014)
7. Tian, H., Shu, Y., Wang, X.-F., Mohammad, A.M., Bie, Z., Xie, Q.-Y., Li, C., Mi, W.-T., Yang, Y., Ren,

T.-L.: A graphene-based resistive pressure sensor with record-high sensitivity in a wide pressure range. Sci. Rep. **5**, 8603 (2015)
8. Tian, H., Chen, H.-Y., Ren, T.-L., Li, C., Xue, Q.-T., Mohammad, A.M., Wu, C., Yang, Y., Wong, H.-S.P.: Cost-effective, transfer-free, flexible resistive random access memory using laser-scribed reduced graphene oxide patterning technology. Nano Lett. **14**, 3214–3219 (2014)
9. Tian, H., Li, C., Mohammad, A.M., Cui, Y.-L., Mi, W.-T., Yang, Y., Xie, D., Ren, T.-L.: Graphene earphones: entertainment for both humans and animals. ACS Nano **8**, 5883–5890 (2014)
10. Tian, H., Shu, Y., Cui, Y.-L., Mi, W.-T., Yang, Y., Xie, D., Ren, T.-L.: Scalable fabrication of high-performance and flexible graphene strain sensors. Nanoscale **6**, 699–705 (2014)

Nanostructures Based on Porous Silicon

Luca Boarino[1] and Giampiero Amato[2]
[1]Istituto Nazionale di Ricerca Metrologica, Nanofacility Piemonte, Torino, Italy
[2]Quantum Research Laboratory, Electromagnetism Division, Istituto Nazionale di Ricerca Metrologica, Torino, Italy

Definition

The term porous silicon (or, the more general, porous semiconductor) has been and is largely used to define a class of semiconducting material in which pores are opened, basically by means of electrochemical or chemical processes. According to the present definition (and to IUPAC), three different material types are identified, say, nanoporous, mesoporous, or macroporous. The important point is that these types are defined according to the dimension of the pores, instead of the residual structures.

In practice, a correlation between pore diameters and structure sizes is found, e.g., ensembles of macropores and nanostructures are found very seldom, and are the product of more complicated technological steps, not the simple (electro)chemical etching.

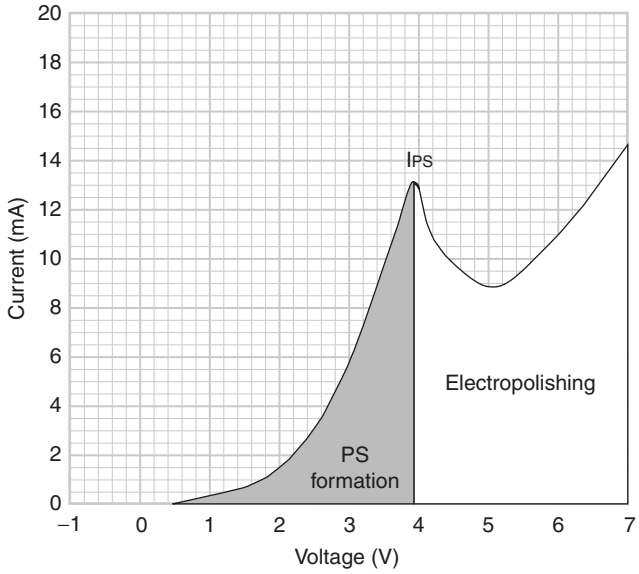
The term porous silicon has been extended in the past to define some kinds of disordered arrays of silicon nanowires. This extended definition does not keep into account for some regular macroporous structures obtained by means of electrochemical etching coupled with photolithography. Moreover, disorder is a typical consequence for the (electro) chemical processes employed, and cannot be considered as distinctive for the material. In other words, nanostructured arrays can be produced in many different ways, not only (electro) chemical ones, and ordered (electro) chemical etching is in principle possible.

Overview

Historical Notes

Porous silicon (PS) was accidentally discovered by Arthur and Ingenborg Uhlir at the Bell Labs, in 1956, in the framework of a research work on silicon and germanium electropolishing. If the electropolishing current was not reached for the electrochemical etching of the silicon wafer, instead of a clean and polished surface, a brown or black layer appeared, due to a nonuniform dissolution of silicon atom (see Fig. 1). The material is porous, full of fine holes, and the electropolishing regime is reached only at the pore tips in these conditions. This former evidence of PS was published in the Bell Lab's technical notes [1] and the study of this system abandoned until the 1970s–1980s, when the huge specific surface of the material, of the order of hundreds of square meters per square centimeter, gave origin to spectroscopic studies, as a precursor for thick oxide growth on silicon, and to the first applications to gas sensing. In 1977, the first patent of a humidity sensor based on porous silicon was deposited [2]. PS was also applied to silicon on insulator (SOI) technology [3].

At the end of the 80s, Leigh T. Canham working at the British Defense Research Agency obtained the experimental confirmation of his intuition that the nanofilaments of porous nature could exhibit good luminescence properties, due to quantum confinement effects. Although other authors were investigating on the phenomenon,



Nanostructures Based on Porous Silicon, Fig. 1 Porous silicon formation voltammogram. Over the I_{PS}, electropolishing regime occurs. At lower currents, porous silicon is formed

like V. Lehmann and U. Gösele, Canham was the first to publish, in 1990, the experimental evidence of PS room temperature photoluminescence [4, 5]. This experiment stimulated the work of many research groups all over the world in the nonlinear optical and electrical properties, and the number of published papers arose exponentially from 1991 to 1995. The premises of a brilliant future for silicon-based lasers were not maintained, due to the low efficiency of electroluminescence of this material, slow decay times, and to the chemical and mechanical stability.

In the mid-1990s, the main applications of porous silicon were micromachining, gas sensing, and with 1995 the medical and biological applications received a big boost, due to the demonstration of hydroxyapatite growth on porous silicon end to the proven biocompatibility and hemocompatibility of nanostructure silicon, again by L. T. Canham [6].

Other studies showed that B50 rat hippocampal cells have a better adhesion onto porous silicon with respect to untreated surfaces, so indicating suitability of porous silicon for cell-culturing purposes.

In 2001, a group of researchers at the Technical University of Munich [7] announced that hydrogenated porous silicon, in contact with liquid oxygen, reacts explosively, releasing more energy than an equivalent amount of TNT with higher speed, and with no necessity of a detonating material also for small quantities. The application of this discovery, substituting the liquid oxygen requiring very low temperatures with salts rich in oxygen, is in the air bag igniters in the automotive field, in the satellite propulsion, and in atomic emission spectroscopy [8].

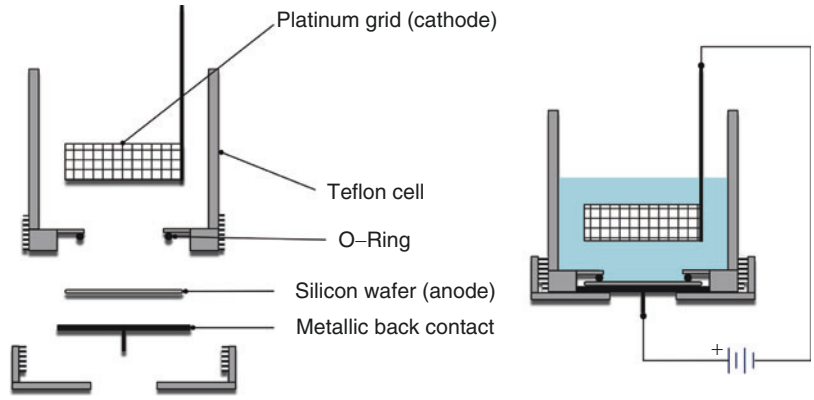
Fabrication of Porous Silicon

Anodization

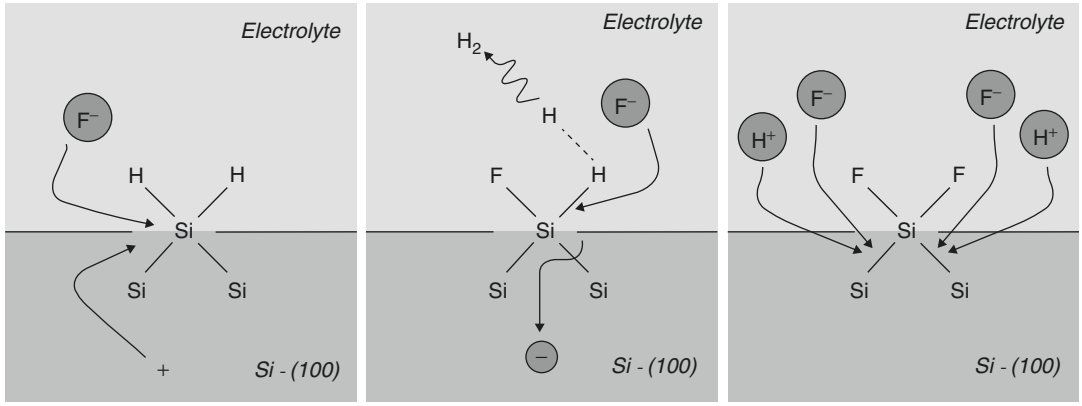
Porous silicon is realized by electrochemical anodization of a silicon substrate, in a solution of hydrofluoric acid (HF) and ethanol in different ratio, depending on the silicon wafer resistivity on the desired etch rate and thickness and characteristics of the porous layer. The cathode is a platinum wire or a grid; the beaker can be Teflon or PVC (Fig. 2). To obtain homogeneous and thick layers, a pulsed bias is preferable, so giving time to the solution at the pore tips to recover its title by fluorine ions diffusion from the bath.

The ethanol is added to the hydrofluoric acid to favor the hydrogen bubbles removal from the porous silicon layer during formation.

Formation Mechanisms The selective removal of the silicon atoms (local electropolishing at the



Nanostructures Based on Porous Silicon, Fig. 2 Electrochemical cell scheme for porous silicon fabrication, (exploded, *left*, and assembled, *right*)



Nanostructures Based on Porous Silicon, Fig. 3 Dissolution sequence of silicon in presence of HF and holes

pore tips) or complete electropolishing is performed by the fluorine ions in presence of two or four holes in the bulk of the semiconductor, allowing the breaking of the Si-Si backbonding (Fig. 3). After the initial work of Turner, Memming, and Schwandt proposed the overall dissolution reaction:

$$\mathrm{Si} + 2\mathrm{HF} + \lambda h^{+} \rightarrow \mathrm{SiF_2} + 2\mathrm{H}^{+} + (2-\lambda)\mathrm{e}^{-} \quad (1)$$

$$\mathrm{SiF_2} + 2\mathrm{HF} \rightarrow \mathrm{SiF_4} + \mathrm{H_2} \quad (2)$$

$$\mathrm{SiF_4} + 2\,\mathrm{HF} \rightarrow \mathrm{H_2SiF_6} \quad (3)$$

Where h^{+} is the hole, e^{-} the electron, and λ is the number of charges exchanged.

Similar mechanisms have been proposed in literature, including silicon oxide formation in presence of H_2O and dissolution in presence of HF, like in the case of the electropolishing regime:

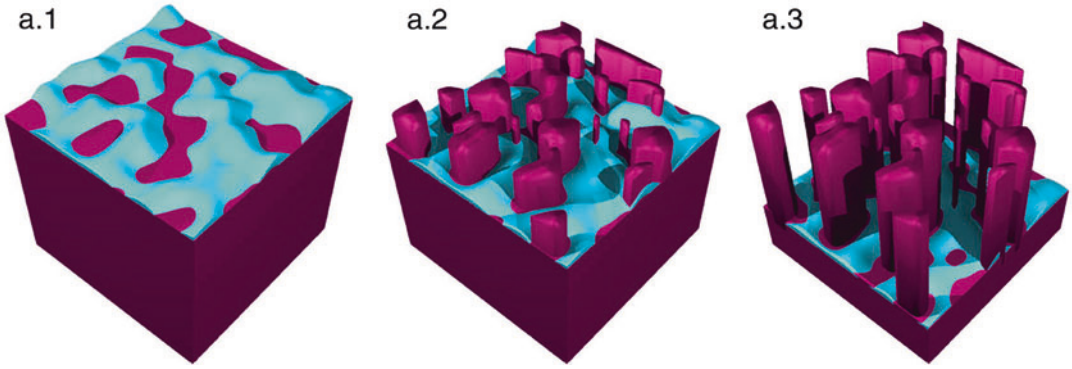
$$\mathrm{Si} + 2\mathrm{H_2O} \rightarrow \mathrm{SiO_2} + 4\mathrm{H}^{+} + 4\mathrm{e}^{-}$$
$$\mathrm{E} = -0.84\mathrm{V/NHE(Normal\ Hydrogen\ Electrode)} \quad (4)$$

$$\mathrm{SiO_2} + 2\mathrm{HF_2}^{-} + 2\mathrm{HF} \rightarrow \mathrm{SiF_6}^{2-} + 2\mathrm{H_2O} \quad (5)$$

Stain Etching

Porous silicon formation occurs also in the absence of an external biasing, by stain etching with hydrofluoric acid, nitric acid, and water.

The vacancies necessary to silicon dissolution in this case are provided by HNO_3, the etchant



Nanostructures Based on Porous Silicon, Fig. 4 Dissolution sequence of silicon metal-assisted etching. In *light blue* the silver clusters are depicted, for different etching times (from a.1 to a.3)

concentrations are generally higher, and the etch rates are lower than in the electrochemical process. The first studies were carried on during the 1960s to probe P-N junction, then the discovery of room temperature photoluminescence of porous silicon obtained electrochemically renewed the interest for stain etching, and a systematic study regarding the structural and optical properties of porous films have been exploited [9].

The reactions are the following:

- Anode:

$$\mathrm{Si} + 2\mathrm{H_2O} + \mathrm{m\,h^+} \rightarrow \mathrm{SiO_2} + 4\mathrm{H^+} + (4 - \mathrm{m})\mathrm{e^-} \tag{6}$$

$$\mathrm{SiO_2} + 6\mathrm{HF} \rightarrow \mathrm{H_2SiF_6} + 2\mathrm{H_2O} \tag{7}$$

- Cathode:

$$\mathrm{HNO_3} + 3\mathrm{H^+} \rightarrow \mathrm{NO} + 2\mathrm{H_2O} + 3\mathrm{h^+} \tag{8}$$

- Overall:

$$3\mathrm{Si} + 4\mathrm{HNO_3} + 18\mathrm{HF} \rightarrow 3\mathrm{H_2SiF_6} + 4\mathrm{NO} + 8\mathrm{H_2O} + 3(4 - \mathrm{m})\mathrm{h^+} + 3(4 - \mathrm{m})\mathrm{e^-} \tag{9}$$

Recently, K. Kolasinsky [10] developed a new type of PS stain etching oxidants like, Fe(III), Mn (VII), and Cr(VI) drastically reducing hydrogen bubbling so obtaining very homogeneous and highly luminescent materials.

Metal-Assisted Etching

Metal-assisted etching (MAE) has been recently reported by Li, Bohn, and coworkers [11]; it is a sort of self-catalytic etching of silicon, with no need of polarization, since the formation mechanism is ruled by local redox potential of the metal with respect to silicon. This drives the holes necessary to dissolution at the metal–semiconductor interface. Noble metals like gold, silver, or platinum can be used and obtained by precipitation starting from liquid metallic salts like $AgNO_3$ and others but also deposited in form of thin solid films by sputtering or evaporation.

The metal clusters at the surface of the silicon wafer, rapidly sink following the crystallographic directions and, define silicon walls or wires (Fig. 4), depending on the metal particle distribution at surface, on the temperature, and concentration of the etchants. When the metal is patterned on the surface, MAE leads to regular morphologies.

The dissolution of silicon in HF and an oxidizing agent is a combination of chemical and electroless process, having strong similarities with porous silicon formation or electropolishing by stain etching in HF-HNO_3 [12].

In electroless metal-assisted etching, holes are injected in silicon by H_2O_2, independently from the original substrate doping. In this case, the oxidizing agent concentration can be considered

equivalent to current density in electrochemical process. HF concentration rules dissolution and surface chemistry at the same way for both the processes. The ratio between oxidizing agent and HF deeply affects the etching regimes and the morphologies.

The reaction mechanisms involved in the electroless dissolution of silicon in presence of metal particles are rather complex and will be given in summary following the work of Chartier and coworkers [13] for the case of Ag on Si.

The Ag metallic nuclei at the silicon surface are charged of electrons, and for charge compensation, localize holes at silicon–silver interface, in majority coming from H_2O_2 dissociation at cathode in presence of catalyzing Ag, as shown in Eq. 10 and forming SiO_2 or removing Si at anode, as in Eq. 11.

- Cathodic reaction

  The Ag particles at cathode catalyze the reduction of H_2O_2 as in Eq. 3, taking advantage from a favorable redox potential, and enhancing the etch rates by supplying free holes in proximity of Si-Ag interface.

$$H_2O_2 + 2H^+ \rightarrow 2H_2O + 2h^+$$
$$E_0 = +1.76\ \text{V/NHE(Normal Hydrogen Electrode)} \quad (10)$$

- Anodic reaction

  In analogy with the well-known chemical dissolution of Si in HF-HNO_3 [14], summarizing the different mechanisms proposed in literature and following the work of Chartier, the following anodic reaction is proposed:

$$Si + 6HF + nh^+ \rightarrow H_2SiF_6 + nH^+ + [(4-n)/2]H_2 \quad (11)$$

The general form is valid for the two already cited electrochemical regimes, porous silicon formation and electropolishing, where n, ranging from 2 to 4, is the number of holes per dissolved Si atom.

In the case of $n = 2$, porous silicon formation occurs, with the following overall dissolution anodic reaction:

$$Si + 4HF_2^- \rightarrow SiF_6^{2-} + 2HF + H_2 + 2e^- E_0 = -1.2\ \text{V/NHE} \quad (12)$$

also valid for HF/HNO_3 stain etching well below the critical current density of electropolishing (J_{PS}).

In the case of $n = 3$, ½ H_2 per dissolved silicon atom, as in the case of 6 M HF – 6 M HNO_3 solutions or in stable electrochemical dissolution at J_{PS} [15].

When $n = 4$, the electropolishing regime occurs, as in the case of low HF/HNO_3 ratio or for current densities higher than J_{PS}, with a different mechanism involving the formation and dissolution of intermediate species like SiO_2 and without H_2 formation (Eqs. 13a and 13b).

$$Si + 2H_2O \rightarrow SiO_2 + 4H^+ + 4e^- \quad E_0 = -0.84\ \text{V/NHE} \quad (13a)$$

$$SiO_2 + 2HF_2^- + 2HF \rightarrow SiF_6^{2-} + 2H_2O \quad (13b)$$

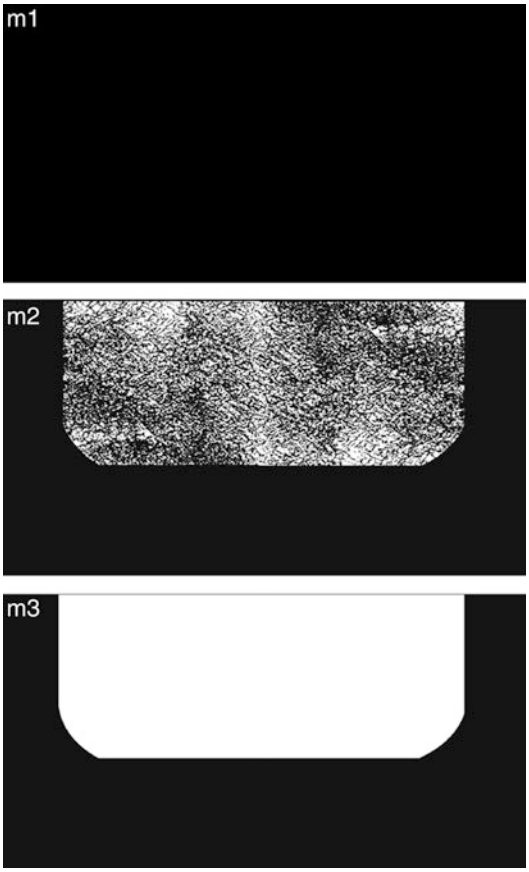
The local mechanism of silicon dissolution in proximity of Ag particle is still unclear, but the pore morphology indicates hole consumption close to the silicon/metal clusters interface, also thanks to the catalytic reduction of H_2O_2.

This phenomenon can be ascribed to a faster surface reaction with respect to holes diffusion due to coupling between silicon valence bonds and energy levels of silver or a depletion layer around Ag nanoparticles allowing an easier access of holes in proximity of Ag/Si interface with respect to HF/Si interface to be produced.

Classification of Porous Silicon

Porosity

Porosity is defined as the volume fraction between silicon structures and voids in the porous layer, and it is determined by gravimetry by weighting the silicon wafer before the etching (m1), after the etching and drying (m2), and after PS removal by alkaline solution (m3) as in Fig. 5. Porosity is a number between 0 and 1 and is obtained by the formula:



Nanostructures Based on Porous Silicon, Fig. 5 The three weights used in gravimetry for porosity evaluation. The first, m1, is the weight of the silicon sample before anodization;; the second, m2, is the weight of the sample after PS formation and at least 20 min of drying;; the third, m3, is the weight of the sample after alkaline removal of the PS layer

$$P(\%) = (m1 - m2)/(m1 - m3)$$

The most critical parameters affecting PS structure and porosity are current density and HF concentration. Porosity can range from 4 % for macroporous silicon to 95 % for mesoporous layers. Porosity is only an indicative parameter, and it does not yield information regarding the microstructure of the PS layer, structure size, and relative size distributions that are much more significative for the prediction and interpretation of physical phenomena like optical, thermal, and electrical responses of the system.

From gravimetry measurements, the thickness of the porous layer can also be determined using the following formula:

$$W = (m1 - m3)/S \times d$$

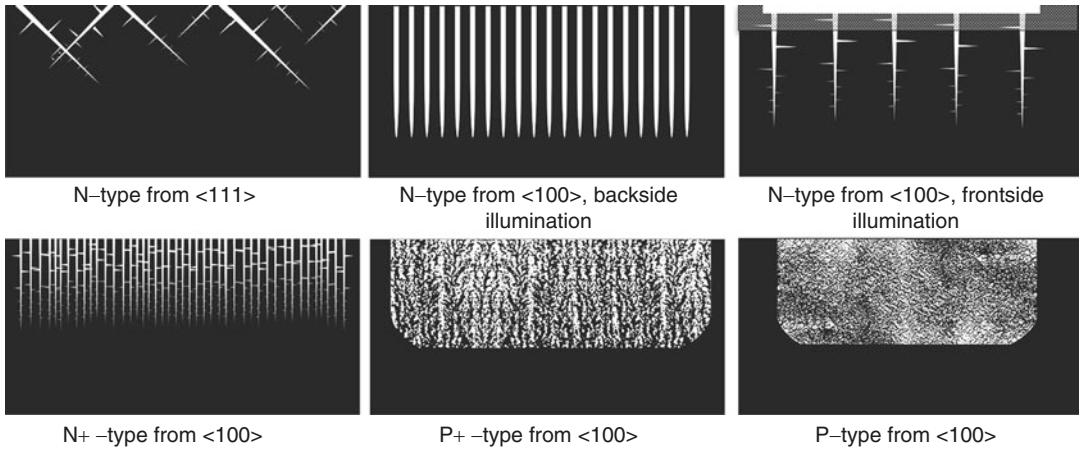
Where, d is the silicon density and S is the silicon surface exposed during etching. Other destructive methods to evaluate thickness are SEM sections, profiling after the porous layer removal, and among the nondestructive methods is ellipsometry, even if the final result is dependent on the adopted model.

Pore Size, Morphology, and Orientation

According to an IUPAC convention [16], porous silicon has been divided into three categories based on the size of its pores: macroporous, mesoporous, and microporous.

Type	Microporous	Mesoporous	Macroporous
Pore width (nm)	Less than 2	Between 2 and 50	Larger than 50

The pores are strongly dependent on the substrate characteristics like crystalline orientation and impurity concentration. In general, pores initiate at the wafer surface and tend to follow the crystalline directions, an example for different orientations is shown in Fig. 6. Since the electrochemical etching is dependent on the number of holes available in the substrate, PS morphology is deeply affected by the original impurity content of the donors or acceptors in the substrate. In low resistivity N-type silicon, no etching occurs in the absence of light or high bias. In the presence of light, if the energy hν of the radiation is greater or equal to the silicon band gap, 1.12 eV at room temperature, the energy excess is transferred to the crystal lattice by rapid thermal processes on the timescale of picoseconds, and the amount equal to the silicon band gap creates an electron–hole couple in conduction band and valence band, respectively. The holes can thus contribute to silicon dissolution, in presence of fluoride ions of the solution, removing the silicon–silicon bonding. If the illumination comes from the top of the silicon surface, from the same side of the etching



Nanostructures Based on Porous Silicon, Fig. 6 Some PS morphologies: from *left* to *right*, first row: N-type PS silicon from (111) substrate orientation, N-type PS silicon from (100) substrate orientation and *backside* illumination, N-type PS silicon from (100) substrate orientation and *front* illumination;; second row: N + type PS silicon from (100), P + type PS silicon from (100), P-type PS silicon from (100)

solution, the pores will initiate from the top and will follow the 100 direction, forming regular or branched macropores, depending on parameters like electric field and free carrier concentration available, but also a microporous phase, due to the quantity of holes generated by light. The depth of this layer depends on the spectral component of the lamp. If the light reaches the substrate from the bottom, the holes will drift through the silicon substrates, reaching the pore pits at the upper interface with HF and ethanol. There is a high probability that the holes will reach the pore tips, so contributing to maintain a stable and regular pore growth.

In N-type macroporous silicon pore etching, there are several critical parameters to be monitored for a regular and sufficiently fast etching: temperature, current density, electrode potential, and illumination intensity.

For the current density, galvanostatic or potentiostatic conditions can be used. Typically constant current conditions (i.e., galvanostatic) are adopted, so allowing the formation of pores under constant dissolution regime.

So, regular macropores on N-type silicon with backside illumination can be controlled in two ways: (1) by using galvanostatic/potentiostatic control maintaining the illumination intensity constant, (2) fixing current and potential and adjusting the illumination intensity via feedback loop.

The highly doped N-type and P-type silicon give origin to mesoporous structures. In the case of P-type, the great availability of free holes due to doping and the low resistivity of the material gives origin to high etch rates, of the order of hundreds of nanometers per second, depending on the current density. A typical value can be 300 nm/s with 300 mA/cm^2 in HF:Eth 1:1 starting from HF 49 %. Mesoporous silicon has an intrinsic anisotropy. The material shows an interconnected columnar structure, with structure size distribution extending from 1.5 to 15 nm [17]. This material is scarcely luminescent;; it is mechanically robust and its morphology is determined by the presence of impurities, since the etching does not remove them selectively (e.g., Boron), but the anodization deactivates the trivalent atoms leaving them in subsurface position [18].

For the highly doped N-type silicon, the mechanism is rather different. The right branch of the curve is determined by the applied current density, and the regime is electrochemical;; the left branch of the curve is more a chemical dissolution, with slow etch rates, and so reaching high porosities.

Low resistivity P-type silicon gives origin to a tiny coral structure, with typical sizes around

1–2 nm. These PS layers are fragile and show room temperature photoluminescence.

Optical Properties and Photoluminescence

Generally, nano-PS emits bright visible light at room temperature in spite of the fact that the pristine c-Si is a well-known indirect-gap semiconductor, with a fair emission in the near-IR, at cryogenic T.

The application of photoluminescence (PL) technique in PS involves several different physical processes, which make the final product (the PL spectrum) difficult to interpret. Absorption of the incident photon creates an electron–hole pair. If nanocrystallites are somehow interconnected, or imbedded in a medium, these particles can escape through different mechanisms. If not, they can decay to lower states *in the same nanocrystal*, or could recombine radiatively, directly from the excited state. If localized states are present, trapping into these shallow states can occur. On the contrary, deep states, e.g., unpassivated dangling bonds, can enable the competing nonradiative recombination of electron–hole pairs.

After several years of investigation, there is a general consensus about the origin of such strong light emission. Resonant PL technique proves that emission arises in the crystalline core of the material (instead of other Si-O, Si-H, or siloxene compounds) [19] where the energy gap is broadened, thanks to quantum confinement (QC) effects. On the other hand, the efficient passivation of the nonradiative pathways (due to H-, or O- termination of the Si nanostructures) plays an important role.

PL decay times are in the order from μs to ms, which makes practical application of PS in integrated optoelectronics unlikely. In particular, the red PL shows nonexponential decay, with typical times ranging in the μs and ms range at room T and at low T, respectively. The competition between radiative and nonradiative recombination processes depends on T. Generally, the PL intensity reaches a maximum around 50–200 K (depending upon the sample), the simple meaning being that one of the competing processes, say, the radiative one, prevails at those T values. The T dependence of the PL lifetime is the consequence of exchange splitting of excitons. Indeed, the electron and hole, with spin 1/2, can be combined into a singlet ($S = 0$) and a triplet ($S = 1$) state, the exchange interaction stabilizing the triplet. While at high T both of them are populated, the PL arising from the "fast" decay of the singlet state, at low T the PL becomes much slower because the triplet state is virtually stable. In the T range at which the PL emission is maximum, the radiative process is still related to population of singlet states, whereas the nonradiative one in quenched. When T is further decreased, radiative processes are now related to the slow decay of triplet states, and nonradiative ones become competing again.

PS Structures

PS has been fruitfully employed as a sacrificial layer since it is a strongly reactive material and easy to etch [20].

Generally, a wide class of materials, from metals to semiconductors and insulators, can be deposited onto PS. Due to the intrinsic reactivity of PS, it can be easily etched, giving rise to large undercutting of the layers deposited on it. This method is proved to be very efficient for realizing suspended membranes, bridges, or cantilevers.

PS has been also employed for the epitaxial deposition of c-Si because it conserves the lattice of the pristine material and can be used as a seed for further growth. Thanks to its reactivity and fragility, it allows transfer of the epitaxial film onto foreign substrates.

Macroporous Silicon

The former studies of N-type silicon anodization in the dark started in the early 1970s. Etch pits and silicon dissolution in dark is due to and breakdown between surface states tunneling in conduction band. From these studies, some evidences were clear: the channels formation is not due to crystal defects, the pores direction follows the orientation of the substrate, the growth rate is in general constant with anodization time, so the diffusion within the pores is not a rate-limiting process, and the pores' walls are passivated.

Beale proposed a depletion of porous layer, explaining the high resistivity of the material

after anodization, and localizing carriers at the pore tips. Several concepts introduced by this physical model were then adopted in successive works, but no chemical mechanisms were included, so providing poor comprehension of pore size variations and morphologies. Smith and coworkers, at the end of 1908s, postulated that the rate of pore growth is limited by the diffusion of the holes randomly walking in the bulk of silicon. This model was mainly ruled by the diffusion length L, a quantity that is function of dopant concentration and potential. A sticking factor varying from point to point of the partially oxidized pore tip was necessary to explain different morphologies and pore diameter variations. This model was too general to explain the conduction mechanisms and morphologies in different types of substrates and like the previous was not taking into account the electrochemical reactions at the interface. Zhang and coworkers [21] in the same period, thanks to a systematic study of the i-V characteristics, clarified that the PS formation or electropolishing is not dependent on the electronic properties of the substrate, but on the result of the two competing reactions occurring at the pore tips: direct dissolution by fluoride species and indirect removal of silicon through silicon dioxide formation and dissolution in HF. In the following years, Zhang also proposed that surface curvature of the pore tips could reduce the effective width of the space charge layer greatly increasing the interface tunneling current and the dissolution rates of silicon dissolution and oxide formation and removal. This model is also applicable to other substrates and PS morphologies.

Lehmann and Föll in 1990 described the formation of macropores from low-doped N-type silicon in aqueous HF-electrolyte, backside illumination, and pore nucleation predefined lithographically. They experimentally verified the hypothesis that the space charge region (SCR) around the pores can focus the holes diffusing in the silicon bulk on the pore tips, so enabling the growth of regular macropores. In further studies, Lehmann demonstrated that the equilibrium of the carrier transport in the semiconductor and the mass transfer in the electrolyte determines the etch rates and the morphologies in N-type silicon. At low current densities, there is no accumulation of holes and dissolution takes place only at the pore tips, leaving the walls unaffected thanks to carrier depletion. At high current densities, the reaction is mass transport limited and holes accumulate at the pore tips, dissolving walls and enlarging the pores.

Lehmann tried to unify the models describing the formation mechanisms of all types of PS, using the assumption that all the porous structures except the pore tips are passivated by carrier depletion. Depletion is due to quantum confinement for microporous silicon and to space charge layer for porous structures larger than few nanometers.

Föll et al. developed the current burst theory, one of the most recent and comprehensive models of PS formation;; it is so far the only model considering the electrochemical reactions under formation conditions in the different PS morphologies. The main assumptions are that the electrochemical reaction of silicon dissolution operates in microscopic units with a definite spatial and temporal distribution. Other important points are the following:

Current flow is inhomogeneous in space and time, from here the local current burst.
Current flow induces direct silicon dissolution and oxidation and removal following current burst.

As a result of hydrogen termination and desorption, the position of Fermi level and so of the space charge layer oscillates with the time.

The rate of hydrogen termination is fastest on the (111) surface, determining the probability of current burst on surfaces of different orientations.

This model is rather powerful in explaining many aspects involved in the PS formation, but it is extremely difficult to be expressed in mathematical form. [21].

Porous Silicon-Based Photonics

In nano- and mesoporous cases, application in 1D photonic structures deserves a mention. From an optical point of view, PS is seen from visible light

as a homogeneous material. In fact, the sizes of the pores and the Si nanocrystals are well below the light wavelength. When porosity is varied, the macroscopic dielectric constant (and the refractive index) is varied, too. In principle, it is possible to "choose" the refractive index of the layer between the limits of $n = 3.4$ (porosity $= 0$) to $n = 1$ (porosity $= 1$). The range of variation for n is smaller, but n can be varied in a continuous way in this wide range. Being the dependence of n from porosity relatively simple and considering that porosity monotonically increases with the anodization current, by changing the current during anodization in the right way, one is able to obtain optical devices like Bragg reflectors or Fabry-Perot microcavities done by PS. The advantage of this technology is evident: it allows obtaining optical devices directly integrated on Si [22].

A Bragg reflector is a periodical structure of layers with high and low refractive index. The optical thickness of both is chosen to be $\lambda/4$, λ being the light wavelength at which maximum rejection is wanted. The Bragg reflector is then a sort of 1D photonic crystal, there the rejection band stands for photons like energy gap stands for electrons in semiconductors. Correspondingly, one can choose to insert defect states within such gap: this can be accomplished by means of a Fabry-Perot microcavity, where a layer with optical thickness equal to $\lambda/2$ is sandwiched between two Bragg reflectors. The PS anodization technology allows for realizing Bragg reflectors with >98 % reflectivity at the rejection band and Fabry-Perot microcavities with peak FWHM <10 nm. This method could in principle be used to confine light arising from PL emission, but, unfortunately, the best optical 1D structures are obtained in highly doped material, in which the PL is generally poor. The roughness between layers of different porosity is dramatic in lightly doped PS and greatly affects the optical performances of the 1D devices. A possible solution has been found in freezing the anodization etchant;; roughness is greatly reduced in this way, but the quality is still far from devices realized by standard vacuum deposition methods.

Porous Silicon-Based Gas Sensors

The first patent related to porous silicon regards the anodization of silicon for the fabrication of a humidity sensor, US patent no. US4057823, by P. J. Burkhardt and M. R. Poponiak from IBM, in 1976 [2]. The high specific surface of oxidized mesoporous silicon was used here to detect moisture by capacitance variations in a capacitor in which oxidized porous silicon was the dielectric.

From then, a number of works investigated different readout approaches regarding the sensing properties of porous silicon in simple architectures and surface chemistry mechanisms. More recently, advanced architectures (rugate filters, FET transistors) and complex surface chemistry (metallization, silanization, carbonization, and other surface derivatizations) have been proposed trying to increase sensitivity, selectivity and reliability.

From this early application to 1996, several works were published on the sensitivity of porous silicon to different vapors and polar gases in saturation conditions, (i.e., parts per thousands level) [23–27]. In this year, Harper and Sailor observed PL quenching of N-type nanoporous silicon in contact with NO_2. The sensitivity was at ppb (parts per billion) level, but the effect appeared to be irreversible [28].

In 1998 and in the following years, Boarino et al. [29] reported a reversible effect of conductivity variation of p+ type mesoporous silicon in presence of NO_2 and successively with NH_3. The response in presence of these probe gases (the first attracting and the second donating an electron) was also confirmed by insitu FTIR spectroscopy. After investigations of the phenomenon by Electron Spin resonance and NMR, the interaction mechanisms were clarified as a chemisorption mechanism of the gases onto porous silicon surface, reactivating the boron impurities deactivated by surface dangling bond after electrochemical etching. The NO_2 has the higher electron affinity (2.2 eV), and the NH_3 easily tends to donate its outer electron. The material, apparently depleted from free carriers after etching, is reactivating its impurities in interaction with these gases restoring the original conductivity in presence of these

gases [18]. Since those years, a number of applications and investigations regarding NO_2 and gas sensing devices based on porous silicon have been exploited [30–34].

Good sensitivity to other gases and compounds has been obtained by porous silicon derivatization and functionalization,[35–37], a number of processes in which different chemistries are built on the porous silicon surface by proper immobilization of molecules and biomolecules, reducing aging, increasing sensitivity, selectivity, and reliability.

The methods of detection are several, mainly optical, electrical, and electrochemical, based on variations of refractive index, radiative recombination processes, dielectric constant, conductivity, and others in interaction with gas species (chemisorption, physisorption, or condensation processes). Thanks to the high specific surface of porous silicon, around 10^7 times greater than the bulk materials, the variation of a number of physical–chemical parameters have been demonstrated as a function of gas and vapor concentrations, like photoluminescence spectrum, optical reflection, transmission, diffracted optical power, capacitance, current, and resistance [38]. Several readout architectures have been proposed for the electrical parameters capacitors, resistors, diodes, and CMOS transistors [34]]; for the optical monolayers, waveguides, multilayers [39], Bragg mirrors [40], Fabry-Perot resonant cavities [41], rugate filters [42]; but also spectrometry, interferometry [43], ellipsometry, birifringence, conductometry, impedance spectroscopy, and others.

Cross-References

- Dry Etching
- Microfabricated Probe Technology
- Nanostructured Functionalized Surfaces
- Nanostructured Materials for Sensing
- Nanostructures for Photonics
- Nanostructures for Surface Functionalization and Surface Properties

References

1. Uhlir, A.: Electrolytic shaping of germanium and silicon. Bell Syst. Tech. J. **35**, 333–347 (1956)
2. Johannes, P., (Poughkeepsie Burkhardt, NY), Poponiak, M. R. (Newburgh, NY).: Porous silicon dioxide moisture sensor and method for manufacture of a moisture sensor United States Patent 4057823 (1977)
3. Yonehara, T.: BESOI with porous silicon: ELTRAN®. In: Canham, L.T. (ed.) Properties of Porous Silicon, pp. 23–28. Institution of Engineering and Technology, London (1997). ISBN 0852969325
4. Canham, L.T.: Silicon quantum wire array fabrication by electrochemical and chemical dissolution of wafers. Appl. Phys. Lett. **57**(10), 1046–1048 (1990)
5. Lehmann, V., Gosele, U.: Porous silicon formation: a quantum wire effect. Appl. Phys. Lett. **58**(8), 856–858 (1991)
6. Canham, L.T.: Bioactive silicon structure fabrication through nanoetching techniques. Adv. Mater. **7**, 1033–1037 (1995)
7. Kovalev, D., Timoshenko, V.Y., Künzner, N., Gross, E., Koch, F.: Strong explosive interaction of hydrogenated porous silicon with oxygen at cryogenic temperatures. Phys. Rev. Lett. **87**, 068301–068303 (2001)
8. Mikulec, F.V., Kirtland, J.D., Sailor, M.J.: Explosive nanocrystalline porous silicon and its use in atomic emission spectroscopy. Adv. Mater. **14**, 38–41 (2002)
9. Coffer, J.L.: Porous silicon formation by stain etching. In: Canham, L.T. (ed.) Properties of Porous Silicon, pp. 23–28. Institution of Engineering and Technology, London (1997). ISBN 0852969325
10. Kolasinski, K.W.: New approaches to the production of porous silicon by stain etching. In: Granitzer, P., Rumpf, K. (eds.) Nanostructured Semiconductors: From Basic Research to Applications. Pan Stanford, Singapore (2011)
11. Li, X., Bohn, P.W.: Metal-assisted chemical etching in HF/H_2O_2 produces porous silicon. Appl. Phys. Lett. **77**(16), 2572 (2000)
12. Amato, G.: Optical and morphological properties of light-emitting porous silicon prepared by chemical dissolution of silicon wafers. Jpn. J. Appl. Phys. **34**, 1716 (1995)
13. Chartier, C., Bastide, S., Levy-Clement, C.: Metal-assisted chemical etching of silicon in $HF-H_2O_2$. Electrochim. Acta **53**(17), 5509–5516 (2008)
14. Turner, D.R.: On the mechanism of chemically etching germanium and silicon. J. Electrochem. Soc. **107**, 810–816 (1960)
15. Lehmann, V.: Electrochemistry of Silicon, p. 58. Wiley, Weinheim (2002)
16. Rouquerol, J., Avnir, D., Fairbridge, C.W., Everett, D. H., Haynes, J.H., Pernicone, N., Ramsay, J.D.F., Sing, K.S.W., Unger, K.K.: Recommendations for the

characterization of porous solids. Pure Appl. Chem. **66**(8), 1739 (1994)
17. Canham, L.T.: Skeleton size distribution in porous silicon. In: Canham, L.T. (ed.) Properties of Porous Silicon, pp. 23–28. Institution of Engineering and Technology, London (1997). ISBN 0852969325
18. Garrone, E., Geobaldo, F., Rivolo, P., Amato, G., Boarino, L., Chiesa, M., Giamello, E., Gobetto, R., Ugliengo, P., Viale, A.: A nanostructured porous silicon near insulator becomes either a p- or an n-type semiconductor upon gas adsorption. Adv. Mater. **17**(5), 528–531 (2005)
19. Amato, G., Rosenbauer, M.: Absorption and photoluminescence in porous silicon. In: Amato, G., Delerue, C., von Bardeleben, H. (eds.) Optical and Structural Properties of Porous Silicon Nanostructures, pp. 3–52. Gordon and Breach, Amsterdam (1997)
20. Lang, W.: Micromachining applications of porous silicon. In: Amato, G., Delerue, C., von Bardeleben, H.J. (eds.) Optical and Structural Properties of Porous Silicon Nanostructures, pp. 597–620. Gordon and Breach, Amsterdam (1997)
21. Zhang, X.G.: Macropores in n-type silicon and breakdown model. In: Zhang, X.G. (ed.) Electrochemistry of Silicon and its Oxide, pp. 410–418. Kluwer/Plenum, New York (2001)
22. Berger, M.G., Thönissen, M., Theiß, W., Münder, H.: Microoptical devices based on porous silicon. In: Amato, G., Delerue, C., von Bardeleben, H.J. (eds.) Optical and Structural Properties of Porous Silicon Nanostructures, pp. 557–595. Gordon and Breach, Amsterdam (1997)
23. Halimaoui, A.: Porous silicon: material processing, properties and applications. In: Vial, J.C., Derrien, J. (eds.) Porous Silicon Science and Technology, pp. 33–52. Les Editions De Physique, Springer, Les Ulis, Berlin (1995)
24. Anderson, R.C., Muller, R.S., Tobias, C.W.: Investigations of porous silicon for vapor sensing. Sens. Actuators A **23**, 835–839 (1990)
25. Schechter, I., Ben-Chorin, M., Kux, A.: Gas sensing properties of porous silicon. Anal. Chem **67**, 3727–3732 (1995)
26. Motohashi, A., Kawakami, M., Aoyagi, H., Kinoshita, A., Satou, A.: Gas identification by a single gas sensor using porous silicon as the sensitive material. Jpn. J. Appl. Phys. **34**, 5840–5852 (1995)
27. Stievenard, D., Deresmes, D.: Are electrical properties of an aluminum–porous silicon junction governed by dangling bonds? Appl. Phys. Lett. **67**(11), 1570–1574 (1995)
28. Harper, J., Sailor, M.: Detection of nitric oxide and nitrogen dioxide with photoluminescent porous silicon. Anal. Chem. **68**, 3713–3717 (1996)
29. Boarino, L., Baratto, C., Geobaldo, F., Amato, G., Comini, E., Rossi, A.M., Faglia, G., Leŕondel, G., Sberveglieri, G.: NO_2 monitoring at room temperature by a porous silicon gas sensor. Mater. Sci. Eng. B **69–70**, 210–214 (2000)
30. Timoshenko, V.Y., Dittrich, T., Lysenko, V., Lisachenko, M.G., Koch, F.: Free charge carriers in mesoporous silicon. Phys. Rev. B **64**(8), 085314 (2001)
31. Pancheri, L., Oton, C.J., Gaburro, Z., Soncini, G., Pavesi, L.: Very sensitive porous silicon NO_2 sensor. Sens. Actuators B **89**, 237–239 (2003)
32. Borini, S., Boarino, L., Amato, G.: Coulomb blockade tuned by NO_2 molecules in nanostructured silicon. Adv. Mater. **18**, 2422–2426 (2006)
33. Boarino, L., Borini, S., Amato, G.: Electrical properties of mesoporous silicon: from a surface effect to coulomb blockade and more. J. Electrochem. Soc. **156**(12), K223–K226 (2009)
34. Barillaro, G., Strambini, L.M.: An integrated CMOS sensing chip for NO_2 detection. Sens. Actuators B **134**, 585–590 (2008)
35. Stewart, M.P., Buriak, J.M.: Chemical and biological applications of porous silicon technology. Adv. Mater. **12**, 859–869 (2000)
36. Belhousse, S., Cheraga, H., Gabouze, N., Outamzabet, R.: Fabrication and characterisation of a new sensing device based on hydrocarbon groups (CHx) coated porous silicon. Sens. Actuators B **100**, 250–255 (2004)
37. Björkqvist, M., Salonen, J., Paski, J., Laine, E.: Characterization of thermally carbonized porous silicon humidity sensor. Sens. Actuators A **112**, 244–247 (2004)
38. Archera, M., Christophersen, M., Fauchet, P.M.: Electrical porous silicon chemical sensor for detection of organic solvents. Sens. Actuators B **106**, 347–357 (2005)
39. Ruminski, A.M., Moore, M.M., Sailor, M.J.: Humidity-compensating sensor for volatile organic compounds using stacked porous silicon photonic crystals. Adv. Funct. Mater. **18**, 3418–3426 (2008)
40. Snow, P.A., Squire, E.K., Russell, P.S.J.: Vapor sensing using the optical properties of porous silicon Bragg mirrors. J. Appl. Phys. **86**(4), 1781–1784 (1999)
41. Zangooie, S., Jansson, R., Arwin, H.: Ellipsometric characterization of anisotropic porous silicon Fabry–Pérot filters and investigation of temperature effects on capillary condensation efficiency. J. Appl. Phys. **86**(2), 850–858 (1999)
42. Chapron, J., Alekseev, S.A., Lysenko, V., Zaitsev, V. N., Barbier, D.: Analysis of interaction between chemical agents and porous Si nanostructures using optical sensing properties of infra-red Rugate filters. Sens. Actuators B **120**, 706–711 (2007)
43. Letant, S.E., Sailor, M.J.: Detection of HF gas with a porous silicon interferometer. Adv. Mater. **12**(5), 355–359 (2000)

Nanostructures by Electrospinning

Daoheng Sun[1], Gaofeng Zheng[2] and Dezhi Wu[3]
[1]School of Aerospace Engineering, Xiamen University, Fujian, China
[2]Department of Mechanical and Electrical Engineering, Xiamen University, Fujian, People's Republic of China
[3]Department of Aeronautics, Xiamen University, Fujian, People's Republic of China

Synonyms

Electrohydrodynamic forming

Definition

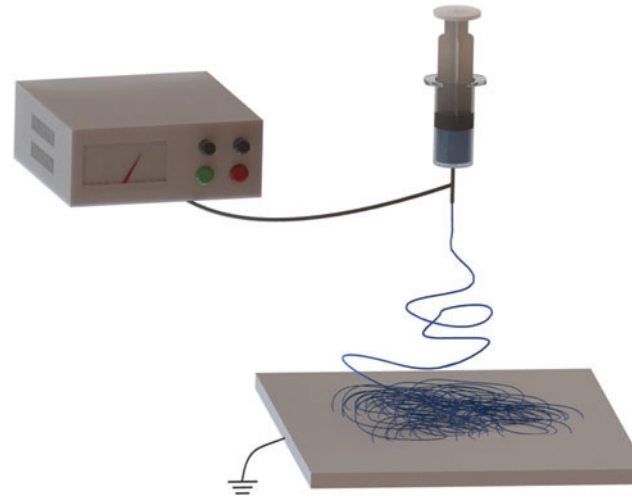
Electrospinning is a method to fabricate nanofibers via coupled effects of electric field on viscous fluids, e.g., polymeric, ceramic, or metallic solutions/dispersions/melts. In electrospinning process, fine charged jet/jets issue from liquid surface, following elongation, evaporation, and solidification. Then solidified nanofibers can be collected.

Introduction

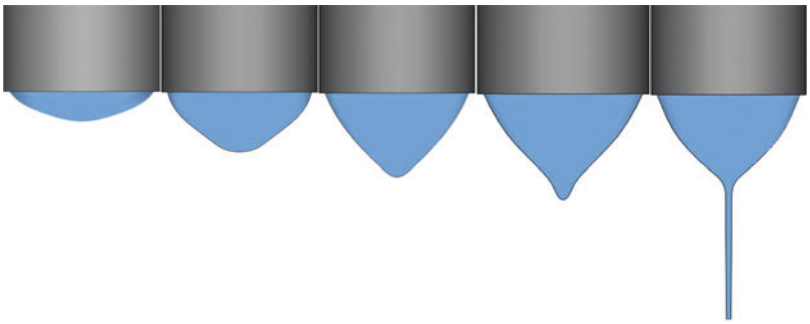
In the electrospinning, a micro/nano jet is generated from the solutions/dispersions/melts under electrostatic field force (EF). As shown in Fig. 1, a typical electrospinning setup consists of a high voltage source, a spinneret, a grounded collector, and a pump. The pendant droplet at the nozzle tip is deformed into a Taylor cone under the EF. As the EF overcomes the surface tension, a fine jet can be ejected from the tip of Taylor cone as illustrated in Fig. 2. After solvent volatilization, solidification, and bending instability, the charged jet can be turned into a solid nanofiber on the grounded collector in Fig. 3.

The base diameter of Taylor cone is smaller than the inner diameter of the nozzle; therefore, electrospinning provides a simple way to decrease the size of spun-structure. Till now, hundreds of materials, such as polymer, ceramic, metallic, and bio materials, have been fabricated into nano/micro fibrous structure by using electrospinning.

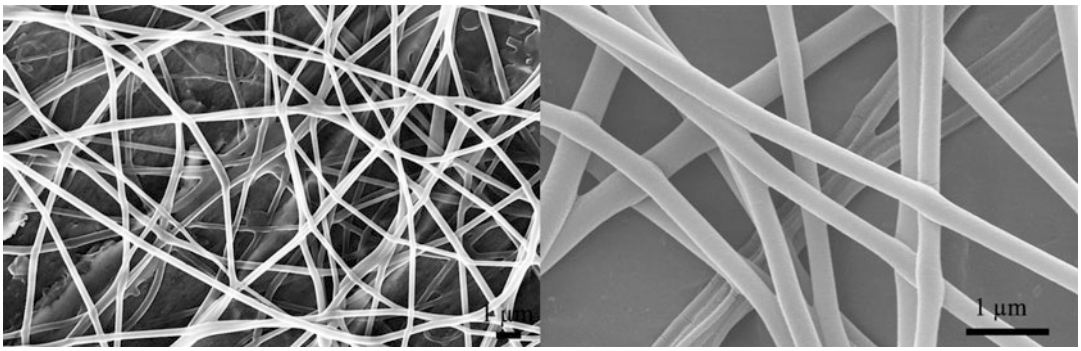
Due to the unique features of low cost, easy process, small line width, and good material compatibility, nanostructures by electrospinning have showed potential applications in filtration, biomedical, energy materials, catalysis, etc. Since 1995, and more than 10,000 scientific papers



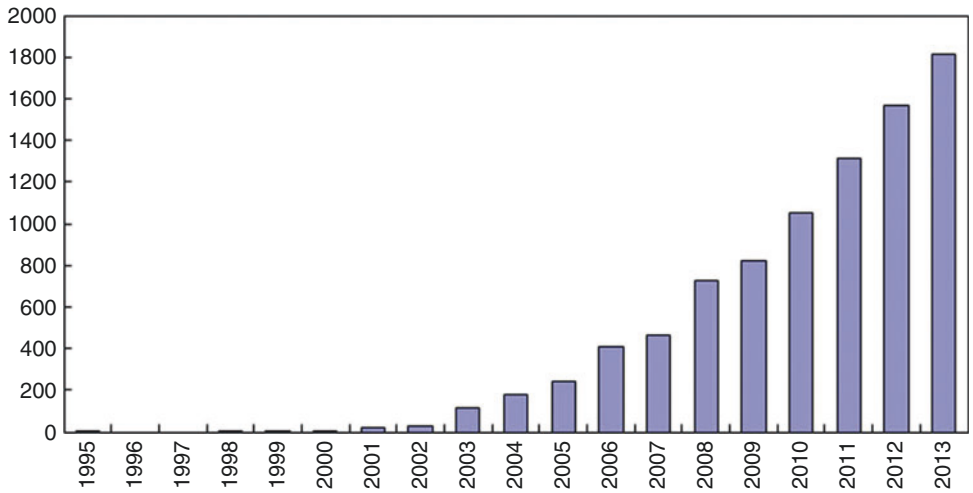
Nanostructures by Electrospinning, Fig. 1 Typical electrospinning setup



Nanostructures by Electrospinning, Fig. 2 Schematic diagram of charged jet ejection process



Nanostructures by Electrospinning, Fig. 3 SEM image of electrospun nanofiber



Nanostructures by Electrospinning, Fig. 4 Publications number on electrospinning

concerning electrospinning have been published (source: Web of Science). The statistical data chart of published papers is shown in Fig. 4 (The keyword for the researching is "electrospinning" in the topic.).

Concepts

As EF overcomes the surface tension, a charged fluid jet emanates from the apex of Taylor cone according to the Rayleigh instability [1].

G. I. Taylor [2] demonstrated the theoretical formula to calculate the critical voltage V_c for forming the charged jet, which can be expressed as:

$$V_c^2 = \frac{4H^2}{L^2}\left[\ln\frac{2L}{R_0} - \frac{3}{2}\right](0.117\pi\gamma R_0)$$

Where, H is the distance between spinneret and collector, L and R_0 represent the length and the radius of the spinneret, respectively. According to the experiment data and theoretical analysis, electric field strength higher than $1 \sim 10$ kV/m is required to generate a jet.

The charged jet that is ejected from spinneret is accelerated by FE and moves toward the collector directly. During the electrospinning process, the Coulomb Repulsion Force (CRF) and EF are the dominant factors that determine the ejection and motion behaviors of charged jet. There are two identified charge transfer mechanisms of free electron flow: Ohmic conduction and convective flow in the jet ejection process. The charged jet is elongated by EF and the stretching ratio from Taylor cone to solid nanofiber is about $1000 \sim 2000$. On the other hand, the solvent evaporation also decreases the diameter of charged jet further. The interaction between solvent and polymer determines the surface morphologies of electrospun nanofiber.

The motion behaviors of the as-spun nanofiber can be divided into two stages. In the first stage, the charged jet moves in a straight track. The competition between surface tension and EF leads to symmetrical instability, which affects the uniformity of electrospun nanofibers. In the second stage, CRF, nonuniform electric field, fluctuation instability makes the charged jet step into whipping and splitting, resulting in the random deposition of electrospun nanofiber.

Electrospinning jet displays viscous and elastic characteristics under stress. Viscoelasticity is used to describe the behaviors of electrospinning jet, which comprises the properties of solid and liquid phase. The axial stress inside the jet blocks the asymmetric instability and decreases the area of deposited zone of electrospun nanofiber.

The ejection behavior depends on process parameters of solution properties, applied voltage, polymer molecular weight, spinneret to collector distance, and ambient conditions. The solution properties of viscosity, surface tension, solution concentration, viscoelasticity, electric conductivity, and dielectric constant are the main factors that determine the ejection phenomena of electrospinning.

High viscosity solution produces continuous charged jet. Nanofiber is fabricated through the continuous jet ejection mode. Smooth, porous, beaded, or ribbon nanofiber can be generated by adjusting the experiment conditions. Due to its high surface tension, the jet of low viscosity solution is broken into nanoparticles. Electrospray and drop on demand (DoD) are two typical modes of discrete jet ejection. The jet is broken into nanoparticles with diameter ranging from nanometers to microns by the CRF. After that, electrosprayed particles are deposited on the substrate to form micro/nano membrane. On the other discrete ejection mode, jet is shrunk into micro droplets along the motion track. The droplets can be printed according to the requirements, called drop on demand (DoD). Orderly droplet with diameter ranging from 100 nm to 100 μm can be printed on the substrate to construct micro/nano patterns.

Direct current (DC) high voltage is often used to extend the solution and produce electrospun jet with positive charges. The jet and nanofiber carries positive charges from the spinneret under DC electric field. Currently, alternating current (AC) high voltage and pulse high voltage are also used to improve the controlling performance on the morphologies and deposition behaviors of electrospun nanofiber.

Historical Developments

The theory of electrospinning can be dated back to 1600s. At that time William Gilbert firstly discussed the deformation of liquid droplets in an electric field. In 1882, Lord Rayleigh studied the instability of a charged droplet. Fine charged jet can be ejected from the solution surface when

the charge CRF is larger than the surface tension. In 1900, Cooley J. F. [3] published the first patent on the formation and ejection process of charged droplet in the electric field. Then, Anton Formals [4] described a patent for the generation of textile yarns in 1934 and Norton C. L. [5] invented the technique of melt electrospinning ejection in 1936. In 1964, Geoffrey Taylor [2] discussed in detail the deformation process of liquid in electric field. As voltage is increased to destroy the equilibrium between the EF and the surface tension, a liquid drop under the nozzle of spinneret would be deformed into conical shape (with a half angle of 49.3°) called Taylor cone. And then, a theoretical shape of liquid drop is also discussed by Geoffrey Taylor in his work. The word "electrospinning" first appeared in the published literature in 1995.

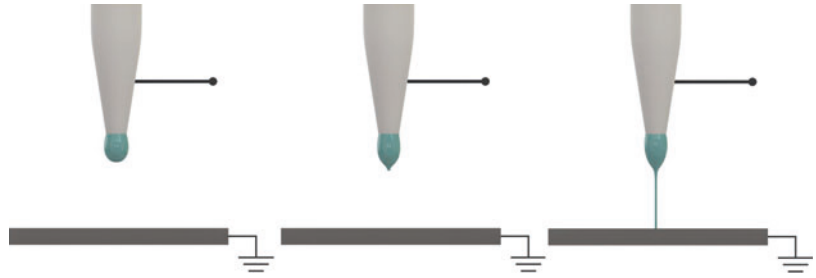
From 1990s, a lot of scientists focused on the different aspects of electrospinning. Firstly, the bending instability and electrically forced jet-stability theories have been studied further to understand the mechanism of electrospinning process better, which provide good ways to improve the deposition and morphology control. Secondly, more than 200 types of materials can be electrospun into nanofibers, including polymer, ceramic, metal, carbon, composite materials, and inorganic materials. Multifunctional nanostructure of piezoelectric, degradable, fluorescent, catalysis, photovoltaic, magnetic sensing, and optical nanofiber have been electrospun for the specific applications. At present, wide applications of electrospun nanostructures have been displayed in the fields of tissue engineering, drug delivery, filters, fiber reinforcement, sensor, optoelectronics, field-effect transistor (FET), nanogenerator, microelectromechanical systems (MEMS) device, etc. Diverse electrospinning setups have been built up to realize high efficiency and orderly deposition of electrospun nanofibers.

Charged Jet Ejection

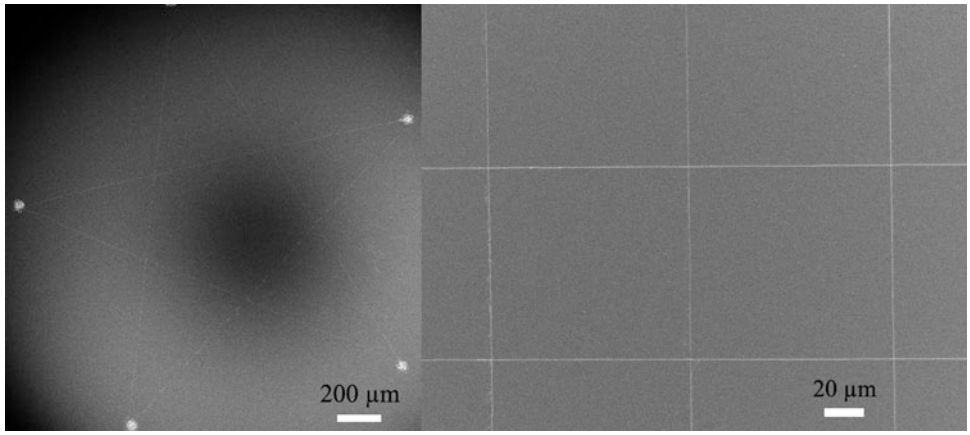
Hollow Spinneret. Most conventional electrospinning processes use a hollow nozzle as the spinneret. During the electrospinning process, the solution is transferred to the nozzle of the hollow spinneret by a pump. Then, a high voltage (typically 5–20 kV) is applied on the spinneret; the distance between spinneret and collector is usually 50 ∼ 150 mm. As applied voltage increases to critical value, the electrostatic force overcomes the surface tension of the solution and a jet is ejected from the tip of the Taylor cone. Undergoing the instability process of whipping and splitting instability, nanofibers can be gained on the collector as nonwoven. As there is only one jet ejected from one hollow spinneret. The throughput of electrospun nanofiber via single hollow spinneret is only ranging from 0.01 to 0.05 g/h. The low throughput of nanofiber is a great obstacle for the hollow spinneret. Multi spinneret array is utilized to increase the throughput of electrospun nanofiber.

Core-Shell Spinneret. There are two liquid supply passages coaxially assembled in a spinneret, and two materials can be ejected synchronously into a charged jet. Thus, different types of nanofibers can be fabricated through such spinneret. For example, oil material and polymer material can be used in the core-shell spinneret synchronously to construct hollow fibers. When the oil material and polymer are transferred through inner and outer passages respectively, double-layer nanofibers with oil material and polymer can be generated through the spinneret. The oil material is removed further and hollow fibers are gained. The assisted gas is also introduced into the outer passages and acted as the focused sheath gas that can increase the stretching force on the jet. The focused sheath gas would decrease the threshold voltage for the ejection of jet and the diameter of electrospun nanofiber. Assisted sheath gas provides a novel way to promote the control of the jet.

Solid Tip Spinneret. A solid tip is also utilized as spinneret, which is immersed into and pulled out of solution during electrospinning. Sun D. H. and coworkers invented a novel electrospinning method called "near-field electrospinning (NFES)," in which a solid tungsten spinneret with tip diameter of 25 μm is used as spinneret [6], as shown in Fig. 5. Then, a liquid droplet with diameter of 50 μm is adhered to the



Nanostructures by Electrospinning, Fig. 5 Near-Field Electrospinning process

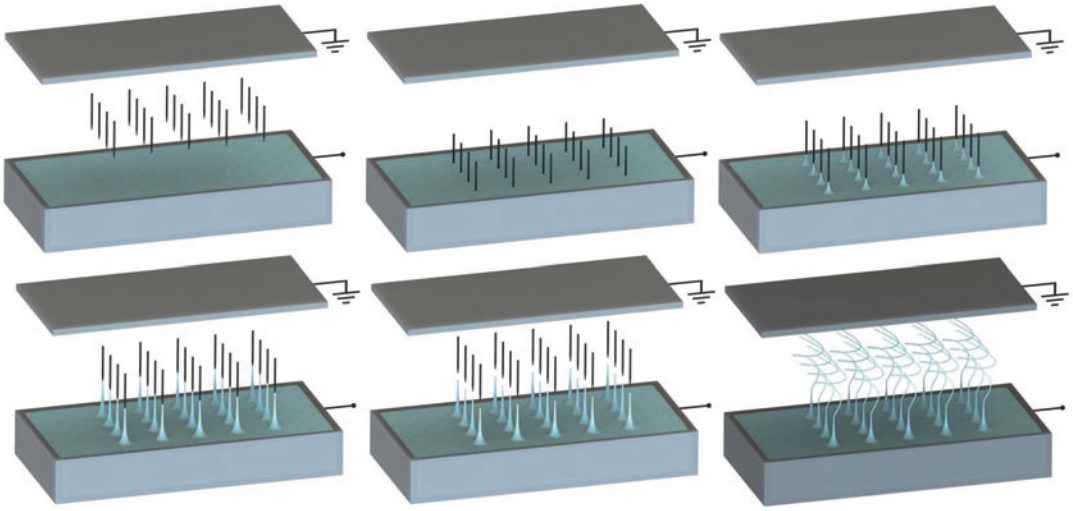


Nanostructures by Electrospinning, Fig. 6 Direct-written orderly nanofibrous pattern through near-field electrospinning

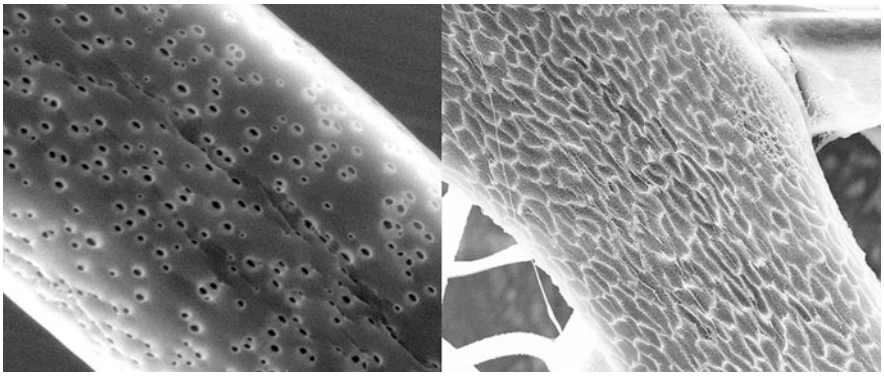
tip of the solid spinneret and straight jet under the tip of solid spinneret can be used to direct-write orderly nanofiber. The distance between spinneret and collector is shortened to 0.5 ~ 3 mm, and the jet deposits on the collector before the bending instability stage appears. Such method decreases the voltage required for the jet ejection from 10 kV to 0.6 kV. The oriental nanofibrous pattern can be direct-written as shown in Fig. 6.

Free Surface Spinneret. Electrohydrodynamic fluctuation instability induced by the electric field is used to generate multi-charged jets from the solution surface. The production efficiency of electrospun nanofibers can be increased greatly. There are several methods to realize the multi-jet ejection. For example, magnetic powder was added into the solution to enhance the fluctuation instability on the solution surface and increase the production efficiency of electrospun nanofibers [7]. Another good method is that cylinder and wire electrode are used as the spinneret and the multi-jet can be ejected from the solution surface that stems from the fluctuation instability on the surface of electrode [8, 9].

Inducing Spinneret. An outward drawing force is introduced to promote the easiness of ejection. Lin L. W. and coworkers [10] used a tungsten probe with tip diameter of 1 μm to draw fine jet from the droplet under the nozzle of spinneret, which decreases the voltage required for the ejection of charged jet and the diameter of electrospun nanofiber. The number and ejection position of charged jets can be controlled and adjusted by the inducing tip. Besides the single inducing tip, the probe array can be used to induce multi-jet from solution surface. The multi-jet induced by the probe array can sustain long and stable ejection. Moreover, the ejection efficiency of electrospun nanofiber is significantly improved, as shown in Fig. 7.



Nanostructures by Electrospinning, Fig. 7 Multi-jet induced by the probe array



Nanostructures by Electrospinning, Fig. 8 Porous morphologies

Morphology and Nanofiber Controlling of Electrospun Nanofiber

By adjusting the electrospinning parameters, electrospun nanofiber with different morphologies and diameter can be fabricated to meet the requirements of different applications.

Smooth Morphology. With the abundant elongation and low evaporation rate, charged jet can be transformed into smooth and uniform nanofiber. Smooth nanofiber is the usual type gained from electrospinning, which can be used as the connection in micro/nano electronic device.

Porous Morphology. High evaporation rate is the main factor that produces the porous structure in the nanofiber. The porous structure increases the specific surface area in the nanostructure by electrospinning. Porous nanostructure improves the performance of sensing and adsorption.

Beaded Structure. The competition between surface tension and EF is the most important role that defines the morphology of electrospun nanostructure. The fluid jet shrinks into beaded structure under the surface tension (Fig. 8).

Diameter Controlling. The effects of process parameters on the diameter of nanofiber have been studied deeply. Attributed to the random

disturbances in the ejection process, electrospun nanofibers have large diameter distribution range. Decreasing the diameter and the diameter distribution range is one of the great difficulties for the industrial application of electrospinning technology. Several works have been done to predict and control the diameter of electrospun nanofiber [11, 12]. Constant diameter electrospinning system has become the research focus for the application of nanostructure by electrospinning.

Deposition Controlling of Electrospun Nanofiber

Cylinder Collector. Rotating mandrel applies excess stretching force on the charged jet. As the rotation speed of the cylinder collector is higher than the travel speed of jet, the electrospun nanofibers can be taken up on the surface of the mandrel and wounded around the cylinder collector. Then, the oriented nanofibers can be deposited on the surface of the collector along rotation direction of the cylinder. As more nanofibers deposit on the cylinder collector, the CRF stemming from the accumulated charges causes subsequent nanofibers to be randomly deposited on the surface of collector. A modified cylinder collector of rotating wire drum was used to gain aligned nanofibers with slower rotation speed.

Knife-Edge Disk Collector. The electric field vector is guided directly to the knife-edge of the disk [13]. Coupled with the rotation of disk, oriented nanofibers are deposited along the edge of disk collector. The rotation speed of disk required to gain aligned nanofibers is about 1000 rpm. The angle of knife-edge can be changed to adjust the deposition pattern of nanofibers.

Parallel Electrode Collector. A collector composed of parallel electrodes can change the electric field distribution above the collector compared with that of traditional collector. With the guidance of uniform electric field perpendicular to the edge of electrode, the jet and nanofiber would be deposited orderly across the gap between parallel electrodes. The accumulated charges on the surface of deposited nanofiber would increase the mutual repulsion among charged nanofibers, and they also promote the parallel distribution of nanofibers. The aligned nanofiber deposited across the gap between parallel electrodes can be removed easily from the collector.

Inducing Tip Collector. Electric field distribution is an important role in determining the motion and deposition behaviors of the jet or nanofibers. The centralized and restriction electric field induced by the tip guides the nanofiber deposition on the predesigned point and decreases the deposition zone of charged nanofibers. The restriction electric field of inducing tip can also guide nanofiber deposited construct 3D structure layer by layer [14].

Near-Field Electrospinning. A solid tungsten electrode is utilized as spinneret, and the solution is supplied by immersing mode. During electrospinning process, the solid spinneret is inserted into the solution and carries out a droplet on the tip. The diameter of droplet adsorbed to the tip of spinneret is about 50 μm, which turns into Taylor cone and a jet. The tip restricts the position of the charged jet, and then the positioning accuracy of electrospun nanofiber is improved. Thanks to the $0.5 \sim 3$ mm distance between spinneret and collector, the straight stable jet is used to direct-write orderly nanofibers. The solid spinneret increases the electric field strength nearby, and the minimal voltage required for the jet ejection is decreased to only 0.6 kV in near-field electrospinning.

Applications of Electrospun Nanofiber

Reinforcement. Reinforcement material in composite system is the most important application for electrospun nanofiber. The modulus and strength to weight ratio of composite system can be increased by the addition of electrospun nanofiber. The Young's modulus of composite system is twice larger than that of bulk materials. Nanofiber with finer diameter plays better role to increase the mechanical properties of composite system, which can be applied in tissue engineering and structural elements, etc [15].

Filtration. Thanks to the good characteristics of high filtration efficiencies and low flow resistance, filtration is an important engineering field

for the application of electrospun nanofiber. During the filtration process, the size and morphology characteristics of electrospun nanofiber should match the scale of particles or droplets to increase the filtration performance. With the advantages of good porous interconnectivity, larger surface area, controllable pore sizes, high permeability, and low basis weight, nanofibers have demonstrated greater potential in the application of filtration. At present, many high-efficiency nanofibrous filters made of electrospun nanofibers have been manufactured and put into market.

Catalysis. With the advantages of larger surface area, electrospun nanofiber also provides an excellent way to promote the catalytic process. The catalytic actions occur on the surface of catalysis. Then, the active sites after catalytic actions prevent catalytic actions further. Nanofibers used as catalysts or catalyst supports can be beneficial to enhance the catalytic actions. Taking the advantages of high porosities and large surface area, electrospun nanofibers have been widely applied in the catalytic fields of chemical catalysts, electrochemical catalysts and photocatalysts, etc.

Micro/Nano Electronic System. The reduction of dimensionality of semiconductor or conductive electrospun nanostructures favors promoting the electrical and thermal transport properties. Electrospun nanostructures have displayed unique advantages in fabrication fields of micro/nano electronic, sensing devices. Continuous nanostructure, which is used as interconnects and functional units for the integration fabrication of electronic, optoelectronic, electrochemical, and electromechanical devices, can be electrospun orderly on the collector directly without photolithography and mask. The controllable deposition position of single nanofiber can be used in the fabrication process of micro/nano devices. Conductive, semiconductor, and sensing nanostructure can be fabricated through the electrospinning process, which have been used in the various electronic fields, such as field-effect transistor, chemical sensors, etc. The specific surface characteristics make electrospun nanofibers good candidates for sensing applications, such as humidity sensors and gas sensors.

Biomedical. Drug delivery system and tissue engineering are the main applications of electrospun nanofiber in the field of biomedical engineering. With the help of electrospun nanofiber, drugs can be delivered to the site of action at a controllable rate by adjusting the morphology, porosity, and composition of electrospun nanofiber. And then, the therapeutic safety and efficacy can be improved by the drug delivery system of antibiotic drug, anticancer drug, gene, and biomolecule, etc. For example, the postsurgical adhesions and infections can be prevented by the drug delivery system of electrospun nanofibers. The electrospun nanofibrous scaffolds can promote the growth of cell and improve the biocompatibility of material by decreasing the diameter of nanofiber. The porosity on the surface of electrospun nanofiber increases the surface area of nanostructure and plays a favorable influence on the cell growth. The nanofibrous scaffolds with large surface area provide good support for the growth of cell. The damages of cells or tissue occurred by the injury or disease can be repaired or regenerated well by the nanofibrous scaffolds. The three-dimensional nanofibrous structure is a new and good candidate for the human cell systems.

Energy Materials. Electrospun nanofibers can be used in high-efficiency energy conservation of lithium-ion battery and fuel cell [16]. With the advantages of short diffusion path, fast intercalation kinetics, large specific surface areas, thermal stability, wettability, and large porosity, nanofibrous membrane can be used as separator in lithium-ion battery with high energy density. Due to the large surface to volume rate and controlled diversity in compositions, proton exchange membranes based on electrospun nanofiber provide a good opportunity for the high efficient fuel cells. Lin L. W. and coworkers [17] utilized near-field electrospinning to direct-write piezoelectric nanofiber from poly(vinylidene fluoride) (PVDF) solution directly, which has high conversion efficiency and displays great advantages in the fields of sensing and actuation.

Micro/Nano Integration System. Electrospun nanostructures are also used as sacrificial layer to fabrication micro/nano integration

system [18, 19]. Firstly, orderly pattern nanofiber is printed on the substrate through the electrospinning process. Secondly, the structural material is deposited over the pattern nanofiber. Thirdly, the structural material layer is solidified and the sacrificial layer of nanofiber is removed. And then, the 3D micro/nano channel can be fabricated, which displays application potential in the fields of micro/nano fluidic chip, sensor, or biological diagnosis.

Outlook and Trend

Electrospinning is a versatile method to fabricate various nanomaterials through a simple and cost-effective process. Various materials can be used to produce nanofiber through electrospinning process. The multifunctional nanofibers have been applied in many fields of membrane reinforcement, biomedical, optoelectronics, flexible electronics, wound healing, micro/nano sensor, etc. The highly efficient production, direct three-dimensional structures forming, and precise integration deposition of functional nanofibers are the main development trends for the nanostructure by electrospinning. The ejection, motion, rheology, and deposition behaviors of the charged jet would be studied further in the future, which can improve the controlling level of electrospinning technology to meet the requirements of various applications.

Cross-References

- ▶ Electric-Field-Assisted Deterministic Nanowire Assembly
- ▶ Electrokinetic Manipulation
- ▶ Electrostatic Actuation of Droplets
- ▶ Flexible Electronics
- ▶ Lotus Effect
- ▶ Nanomaterials for Electrical Energy Storage Devices
- ▶ Nanomaterials for Sensors
- ▶ Nanoscale Printing
- ▶ Nanotransfer Printing
- ▶ Sol–gel Method

References

1. Rayleigh, L.: XX. On the equilibrium of liquid conducting masses charged with electricity. Philos. Mag. **14**, 184–186 (1882). doi:10.1080/14786448208628425
2. Taylor, G.: Disintegration of water drops in an electric field. Proc. R. Soc. Lond. A Math. Phys. Sci. **280**, 383–397 (1964). doi:10.1098/rspa.1964.0151
3. Cooley J.: Improved methods of and apparatus for electrically separating the relatively volatile liquid component from the component of relatively fixed substances of composite fluids. 19 May 1900
4. Formhals A.: Process and apparatus for preparing artificial threads. 2 Oct 1934
5. Norton C. L.: Method of and apparatus for producing fibrous or filamentary material. 21 July 1936
6. Sun, D.H., Chang, C., Li, S., Lin, L.W.: Near-field electrospinning. Nano Lett. **6**, 839–842 (2006). doi:10.1021/nl0602701
7. Yarin, A., Zussman, E.: Upward needleless electrospinning of multiple nanofibers. Polymer **45**, 2977–2980 (2004). doi:10.1016/j.polymer.2004.02.066
8. Wu, D.Z., Huang, X.P., Lai, X.T., Sun, D.H., Lin, L.W.: High throughput tip-less electrospinning via a circular cylindrical electrode. J. Nanosci. Nanotechnol. **10**, 4221–4226 (2010). doi:10.1166/jnn.2010.2194
9. Wang, X., Niu, H.T., Lin, T., Wang, X.G.: Needleless electrospinning of nanofibers with a conical wire coil. Polym. Eng. Sci. **49**, 1582–1586 (2009). doi:10.1002/pen.21377
10. Chang, C., Limkrailassiri, K., Lin, L.W.: Continuous near-field electrospinning for large area deposition of orderly nanofiber patterns. Appl. Phys. Lett. **93**, 123111 (2008). doi:10.1063/1.2975834
11. Munir, M.M., Iskandar, F., Khairurrijal, Okuyama, A.: A constant-current electrospinning system for production of high quality nanofibers. Rev. Sci. Instrum. **79**, 093904 (2008). doi:10.1063/1.2981699
12. Thompson, C.J., Chase, G.G., Yarin, A.L., Reneker, D.H.: Effects of parameters on nanofiber diameter determined from electrospinning model. Polymer **48**, 6913–6922 (2007). doi:10.1016/j.polymer.2007.09.017
13. Theron, A., Zussman, E., Yarin, A.L.: Electrostatic field-assisted alignment of electrospun nanofibres. Nanotechnology **12**, 384–390 (2001). doi:10.1088/0957-4484/12/3/329
14. Kim, H.-Y., Lee, M., Park, K.J., Kim, S., Mahadevan, L.: Nanopottery: coiling of electrospun polymer nanofibers. Nano Lett. **10**, 2138–2140 (2010). doi:10.1021/nl100824d
15. Li, Z.L., Wang, C.: One-Dimensional Nanostructures Electrospinning Technique and Unique Nanofibers. Springer, Heidelberg/New York/Dordrecht/London (2013)

16. Ding, B., Yu, J.: Electrospun Nanofibers for Energy and Environmental Applications. Springer, Heidelberg/New York/Dordrecht/London (2014)
17. Chang, C., Tran, V.H., Wang, J., Fuh, Y.-K., Lin, L.: Direct-write piezoelectric polymeric nanogenerator with high energy conversion efficiency. Nano Lett. **10**, 726–731 (2010). doi:10.1021/nl9040719
18. Wang, X., Zheng, G., Xu, L., Cheng, W., Xu, B., Huang, Y., Sun, D.: Fabrication of nanochannels via near-field electrospinning. Appl. Phys. A Mater. Sci. Process. **108**, 825–828 (2012). doi:10.1007/s00339-012-6975-6
19. Srivastava Y., Marquez M., Thorsen T.: Microfluidic electrospinning of biphasic nanofibers with Janus morphology. Biomicrofluidics **3** (2009). doi:10.1063/1.3009288

Nanostructures for Coloration (Organisms Other Than Animals)

Ille C. Gebeshuber[1] and David W. Lee[2]
[1]Institute of Applied Physics, Vienna University of Technology, Wien, Austria
[2]Department of Biological Sciences, Florida International University Modesto Maidique Campus, Miami, FL, USA

Synonyms

Iridescent colors (organisms other than animals); Physical colors; Structural colors

Definition

Structural colors refer to colors generated by minuscule structures, with the characteristic dimension of the structures on the order of the wavelength of the visible light (i.e., some tens up to hundreds of nanometers). Examples for structural colors are the colors of CDs and DVDs, the colors of soap bubbles or oil films on water (thin films), or the colors of certain butterfly wings (e.g., photonic crystals) and even plants. Tiny wax crystals in the blue spruce scatter the light (Tyndall scattering), resulting in the blue hue. Thin films in tropical understory plants and diffraction gratings in hibiscus and tulip flowers are just some more examples of the amazing variety of natural nanostructures that are the basis for coloration in some plants. This entry reviews the physics behind structural colors; lists plants, microorganisms, and virus species with nanostructures responsible for their coloration; and also touches upon the multifunctionality of materials in organisms, nanobioconvergence as an emergent science, and biomimetics as a promising field for knowledge transfer from biology to engineering and the arts.

Introduction

Iridescent, metallic, or grayish coloration in some plants is not caused by pigments but rather by physical structures with characteristic sizes on the order of some tens up to several hundreds of nanometers [1]. The name-giver for the term iridescent, that is, color change with the viewing angle, is Iris (ιρισ), the Greek goddess and personification of the rainbow.

According to previous studies about various kinds of colors and their origins, colors are divided into two different categories: chemical colors and structural colors. Chemical colors are caused by pigments, whereas structural colors are caused by structures. The interaction of visible light with structures causes shiny, bright colors that might also show iridescence and/or metallic appearance. Vertical spatial structuring may limit the change in color bandwidth with angle – in this way, structural colors of specific hues such as the brilliant blue of the Morpho butterfly can be generated [2, 3]. In some of these structural colors with small bandwidth at ambient conditions, viewing at small angles and high light intensities reveals a wide spectrum of colors (e.g., in the national Malaysian butterfly Raja Brooke's Birdwing, *Trogonoptera brookiana*).

In chemical colors, light is selectively reflected, absorbed, and transmitted. Pigments reflect the wavelengths of light that produce a color and absorb the other wavelengths. Ultramarine blue, for example, reflects blue light and absorbs other colors. Light absorbed by the

pigment yields altered chemical bonds of conjugated systems and other components of the pigment and is finally released as heat, as reemission of light via fluorescence or by passing the energy on to another molecule. Plant pigments are largely soluble in aqueous solutions or lipids, and they actually work more like dyes in the sense that they do not actually directly reflect but allow the passage of wavelengths that are then reflected by structures with different refractive indices, such as cell walls. Plant pigments are therefore more like watercolors on white paper than the colors used in an oil painting.

In structural colors, on the other hand, the incident light is reflected, scattered, and deflected on structures, with negligible energy exchange between the material and the light, resulting in strong, shiny coloration.

The objective of this entry is to give a review of the physical mechanisms leading to structural colors, introduce plant species as well as microorganisms and viruses with the respective nanostructures thought to be responsible for or contribute to the coloration, give some examples for structural colors in current technology, and introduce nanobioconvergence as an emergent science and biomimetics as a promising field for knowledge transfer from biology to engineering and the arts.

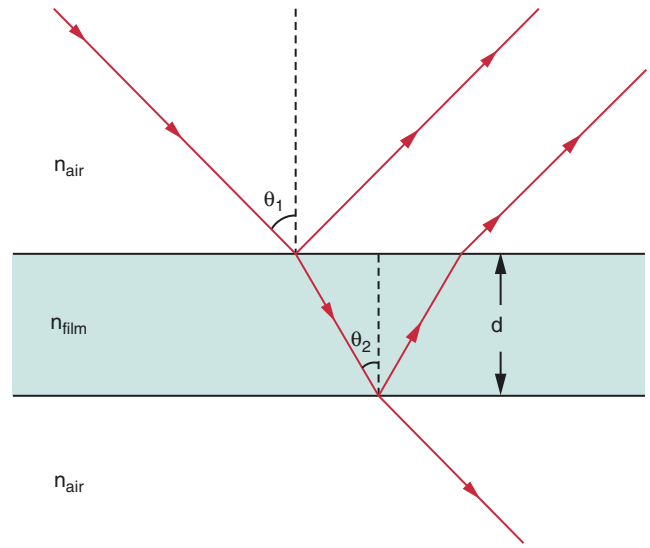
The Physics of Structural Colors

Five physical phenomena lead to structural coloration: thin-film interference, multilayer interference, scattering, diffraction, and photonic crystals.

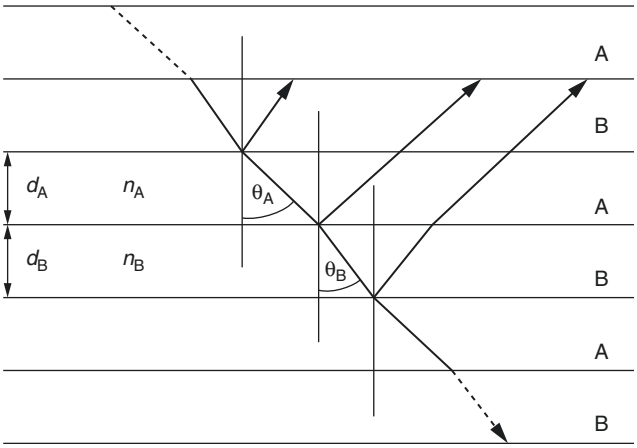
Thin-Film Interference

Thin-film interference occurs when an incident light wave is partly reflected by the upper and lower boundaries of a thin film (Fig. 1). It is one of the simplest phenomena in structural coloration and can be found in various instances in nature, for example, in the wings of houseflies, in oil films on water, or in soap bubbles.

Light that strikes a thin film surface can be either transmitted or reflected from upper or lower boundary, respectively. This can be described by the Fresnel equation. Interference between two reflected light waves can be constructive or destructive, depending on their phase difference. Of relevance are the thickness of the film layer (d), the refractive index of the film (n), and the angle of incidence of the original wave on the film (Θ_1). In addition, if the refractive index n_2 of the thin film is larger than the refractive index of the outside medium n_1, there is a 180° phase shift in the reflected wave. This is, for example, the case in a soap bubble, with $n_1 = n_{\text{air}} = 1$ and $n_2 = n_{\text{film}}$, with $n_{\text{air}} < n_{\text{film}}$.



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 1 Thin-film interference in the case of, for example, soap bubbles



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 2 Multilayer interference. $n_B > n_A$. A change in viewing angle corresponds to a change in the perceived color, due to changes in the apparent film thickness

The reflection on the upper boundary of the film (the air-film boundary) causes a 180° phase shift in the reflected wave (since $n_{air} < n_{film}$). Light that is transmitted at this first boundary continues to the lower film-air interface, where it is either reflected or transmitted. The reflection at this boundary has no phase change because $n_{film} > n_{air}$.

The condition for constructive interference for a soap bubble in air is given in Eq. 1; the condition for destructive interference is given in Eq. 2:

$$2n_{film}d\cos(\theta_2) = (m - 1/2)\lambda \quad (1)$$

$$2n_{film}d\cos(\theta_2) = m\lambda \quad (2)$$

where d equals the film thickness, n_{film} equals the refractive index of the film, θ_2 equals the angle of incidence of the wave on the lower film-air boundary, m is an integer, and λ is the wavelength of the light.

Summing up, there are two conditions that should be satisfied for constructive interference: firstly, the thin film should be thin enough to crest the reflected waves, and secondly, the two reflected waves should be in one phase. The color seen depends also upon the viewing angle, which changes the apparent film thickness.

Multilayer Interference

As described above, a light wave can be reflected from both boundaries of a thin film layer and dependent on the phase of the two reflected light waves either destructive or constructive interference occurs. The same phenomenon occurs in a series of thin films, a multilayer. A schematic of a multilayer is given in Fig. 2. With $n_B > n_A$, between each A-B interface, a 180° change in phase takes place, while at the B-A interface, there is no phase change. Reflection or refraction at the surface of a medium with a lower refractive index causes no phase shift. Reflection at the surface of a medium with a higher refractive index causes a phase shift of half a wavelength.

The condition for constructive interference for a multilayer is given in Eq. 3; the condition for maximum reflection is given in Eq. 4:

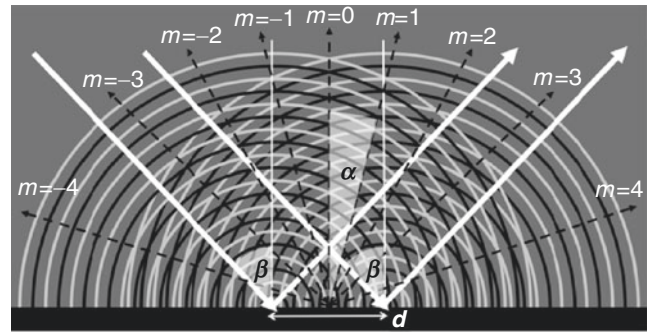
$$2(n_A d_A \cos\theta_A + n_B d_B \cos\theta_B) = m\lambda \quad (3)$$

$$2n_A d_A \cos\theta_A = (m - 1/2)\lambda \quad (4)$$

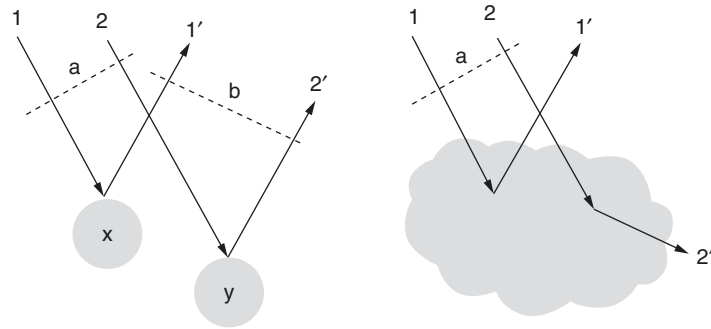
In ideal multilayers, all waves interfere constructively, resulting in colorful reflections, whereas in nonideal multilayers, some waves interfere destructively and the reflection is less colorful [4].

Diffraction Gratings

The grating equation (Eq. 5) correlates the spacing of the grating (d) with the angles of the diffracted and incident beams (α, β) and the wavelength of illumining light (λ) (Fig. 3). The coloration in a particular direction (viewing angle) is generated by the interfering components from each slit of the grating. The diffracted light has maxima at angles



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 3 Schematic of the color generating mechanism by the diffraction of light by a diffraction grating (*black bar* on *bottom*)



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 4 Coherent (*left*) and incoherent scattering (*right*)

where there is no phase difference between the waves. The 0th order reflection corresponds to direct transmission and is denoted $m = 0$. Other maxima occur at angles with $m = \pm 1, \pm 2, \pm 3, \ldots$.

$$d(\sin\alpha - \sin\beta) = m\lambda \qquad (5)$$

CDs and DVDs are examples for this phenomenon. Recently, diffraction gratings that produce colors have been described in flowers [5].

Scattering

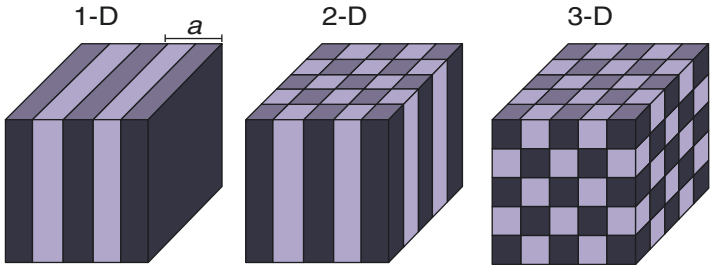
The term scattering is a more general denomination for the interference of light waves with different wavelengths reflected from scattering objects either in a constructive or destructive way (Fig. 4). In terms of coloration, scattering can yield either strong or weak colors. Examples of scattering, and more precisely coherent scattering, are the thin-film interference and the multilayer film interference described above.

In coherent scattering, there is a definite phase relationship between incoming and scattered waves, whereas in incoherent scattering, this is not the case.

The two main types of scattering are termed Rayleigh scattering and Tyndall scattering. In both types of scattering, the intensity of the scattered light depends on the fourth power of the frequency. In Rayleigh scattering, the responsible particles are much smaller than the wavelength of the incident light, whereas in Tyndall scattering, the particles are macroscopic (e.g., dust particles in air or small fat particles in water, as in milk). Since blue light (i.e., shorter wavelengths) is scattered more than red light (i.e., long wavelengths), milk, tobacco smoke, and the sky are blue or have a blue hue. In plants, for example, the blue of the blue spruce can be explained by scattering [6]. Multiple scattering within regular structures produces highly reflective bands in the reflection spectrum and leads to the brilliant colors observed in, for example, iridescent fruits ([1] and references therein).

Photonic Crystals

Photonic crystals exhibit a periodicity a in the refractive index (Fig. 5). There are one-, two-,



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 5 Schematics of one-, two-, and three-dimensional photonic crystals (1-D, 2-D, 3-D). The colors represent materials with different refractive indices. The spatial period of the material is called the lattice constant, *a*

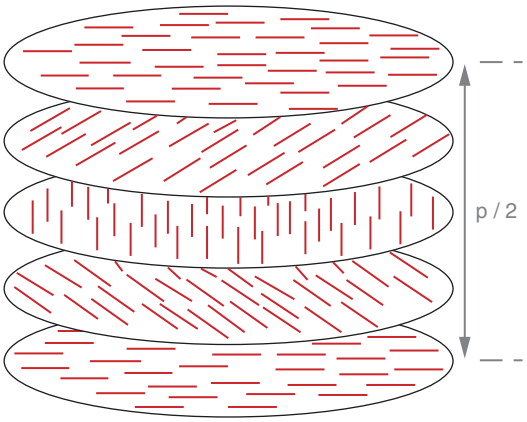
and three-dimensional photonic crystals. Depending on their wavelength, photons either can be transmitted through the crystal or not (allowed and forbidden energy bands). For effects in the visible range, the periodicity of the photonic crystal has to be between about 200 nm (blue) and 350 nm (red).

Vegetable opal is an example for a photonic crystal produced by plants. Opal is the only gemstone that is known to be formed also as a plant-produced product. There are three different types of opal produced by photosynthetic organisms: minuscule iridescent diatoms, phytoliths, and tabasheer. Tabasheer (bamboo opal, plant opal) is a hard, translucent to opaque whitish substance extracted from the joints of bamboo [7]. It can have a beautiful blue hue and was – cut as cabochons – used in China for fashion jewelry.

Cholesteric Liquid Crystals

Cholesteric liquid crystals are layered helicoidal structures, also known as chiral nematic liquid crystals, within which molecules take a preferred direction that gradually changes from layer to layer (Fig. 6). The variation of the director axis tends to be periodic in nature. The period of this variation (the distance over which a full rotation of 360° is completed) is known as the pitch, *p*.

Certain iridescent tropical understory ferns contain a helicoidal layering of cellulose microfibrils that produces multiple interference layers ([1] and references therein), similar to the helicoidal microstructure in a cholesteric liquid crystal and the characteristic liquid crystals in beetles [2].



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 6 The chiral nematic phase in a cholesteric liquid crystal. *p* refers to the chiral pitch (Image reproduced with permission under the terms of the GNU Free Documentation License)

The peak wavelength of the reflected light, λ, is determined by the helicoidal pitch, *p*, where

$$\lambda = pn \tag{6}$$

and *n* is the average refractive index of the anisotropic material [8].

Plants (and Other Nonanimals) with Structural Colors

Structural colors in nature are found in various animals, plants, microorganisms, and even viruses [1–3, 9]. Studies of structural colors date back to

the seventeenth century when the earliest scientific description of structural colors appeared in *Micrographia*, written by Robert Hooke in 1665. In his book *Opticks*, Sir Isaac Newton already in 1704 related iridescence to optical interference:

> The finely colour'd feathers of some birds, and particularly those of the peacocks' tail, do in the very same part of the feather appear of several colours in several positions of the eye, after the very same manner that thin plates were found to do.

Since then, various organisms with structural colors (almost all in animals) have been reported. This chapter focuses on organisms outside of the animal kingdom, primarily plants, but also macroalgae, slime molds, diatoms, and viral particles. For plants, the discussion is organized along the physical phenomena as well as the various organs where such color is found, that is, leaves, flowers, fruits, and seeds. Other organisms with simpler organization are classified by the major groups, for example, the red algae (Rhodophyta).

Gentner (1909) [10] performed various experiments on the origin of the structural colors generating the blue hue on leaves and fruits. He acknowledges Mohl (1870) [11] as the first one to describe this phenomenon (that time wrongly attributed to fluorescence), in the evergreen shrub Laurestine (*Viburnum tinus*).

The first English report on structural colors in plants was by Lee and Lowry [12], who described the physical basis and ecological significance of iridescence in blue plants. The most common mechanisms in plants leading to structural coloration are multilayer interference and diffraction gratings. Multilayer interference is found predominantly in shade-plant leaves, suggesting a role either in photoprotection or in optimizing capture of photosynthetically active light ([1] and references therein). Surprisingly, diffraction gratings may be a common feature of petals, and recent work has shown that bees use them as cues to identify rewarding flowers. Structural colors may be surprisingly frequent in the plant kingdom and still much remains to be discovered about their distribution, development, and function [13].

Nonanimals with Coloration Caused by Thin-Film Interference

True Slime Molds

Slime molds are organisms that normally live as single cells and that can aggregate to a multicelled organism and form fruiting bodies that produce spores. They have the characteristics of single-celled microorganisms as well as of fungi and occur worldwide on decaying plant material. Inchaussandague and coworkers [14] described structural colors produced by a single thin layer in the slime mold *Diachea leucopoda*. Interference within the peridium, a 200 nm transparent layer invariant along the different cells of the organism, is the basis for this color.

Plants with Coloration Caused by Multilayer Interference

Tropical Understory Ferns

The iridescent blue peacock fern Willdenow's Spikemoss, *Selaginella willdenowii* (actually a lycophyte or fern ally), is a common plant in the Malaysian rain forests [15]. It exhibits blue iridescence; investigation with TEM reveals various layers (with less than 100 nm thickness each) on the surface. The photo shown in Fig. 7 was taken in the Bukit Wang Recreational Forest in Malaysia, with very long exposure time as to better reveal the blue coloration of the leaves. Seen with the naked eye, the fern is bluish green and seems to glow in semidarkness. The blue coloration is iridescent and changes with the viewing angle. Holding and tilting the leaves results in color change from green to nearly total blue, although not as strong as seen in the long exposure time photograph shown in Fig. 7.

Also another species of *Selaginella* shows iridescence: *Selaginella uncinata*. As in *S. willdenowii*, a multilayer system where each layer is thinner than 100 nm is responsible for the structural coloration. When the leaves of these two *Selaginella* species are dried, the coloration disappears but comes back again when the plants are hydrated again. When a droplet of water is put on a *Selaginella* leaf, the blue coloration



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 7 The iridescent blue peacock fern *Selaginella willdenowii*, a common plant in the Malaysian rainforest. © Mr. Foozi Saad, IPGM, Malaysia (Image reproduced with permission)

disappears. Other tropical understory ferns where coloration is caused by multilayer interference are *Diplazium crenatoserratum*, *Lindsaea lucida*, *Danaea nodosa*, and *Trichomanes elegans* (the neotropical Bristle Fern, also called the dime store plant because it looks like it is made from plastic and is shade tolerant, is native to the American tropics) ([1] and references therein).

Tropical Understory Begonias

Gould and Lee (1996) reported structurally modified chloroplasts (termed "iridoplasts") in the Malaysian tropical understory plant *Begonia pavonina* (the Peacock Begonia) [16]. The structural coloration is generated by many layers, each less than 100 nm thick. In addition to *B. pavonina*, the authors have observed several other iridescent blue species on the Malaysian peninsula. Modified plastids produce blue colors in *Begonia*, *Phyllagathis* (see section "Other Tropical Understory Plants"), and *Trichomanes* (see section "Tropical Understory Ferns").

Other Tropical Understory Plants

A traditional Malaysian medicinal plant, the melastome *Phyllagathis rotundifolia* (in Malay known as *tapak sulaiman*) and also the Malaysian understory species *P. griffithii* exhibit iridoplasts, similar to the ones in *Begonia pavonina* described above.

Iridescent Macroalgae

Among the red algae (Rhodophyta), several species of *Mazzaella* (=*Iridaea*) are strikingly iridescent [17]. Green and brown algae have been recorded as iridescent, but the phenomenon is most commonly found in red algae, where blue and green colors appear on the surface at certain stages of the life cycle. Iridescent red algae (e.g., *Mazzaella flaccida*, *Mazzaella cordata*) exhibit shiny blue, emerald-green, and deep red colors. The blades are relatively thin and vary in shape and size (up to 3 ft long and 10 in. wide) depending on the habitat and the amount of wave exposure. *Mazzaella flaccida* is iridescent yellow green with purple or brown near the base of the blade, and the iridescent blade of *Mazzaella splendens* is dark purple with a blue iridescent sheen. It is also known as the rainbow-leaf seaweed. The coloration disappears when the algae are dried and is restored when the algae become wet again. Wet *Iridaea* look like they have been coated with oil (Fig. 8), because of their rainbow sheen; iridescent spots, shaped like a teardrop, with a multilayer system of 17 electron opaque and translucent layers, some tens to a few hundred of nanometers thick, are responsible for the coloration. The iridescence might be a by-product of a wear protection mechanism of the algae.

The edible seaweed *Chondrus crispus* (carragheen, Irish moss) is a red algal species



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 8 *Mazzaella* sp., a highly iridescent red alga. © 2003 by J. Harvey, http://www.JohnHarveyPhoto.com (Image reproduced with permission)

that shows blue iridescence due to multilayer interference effects on the tips [18].

The function of algal iridescence is still unclear. Suggestions include a role in camouflage or a role in optimizing photosynthesis by enhancing the absorption of useful wavelengths of light at the expense of increased reflection of other wavelengths.

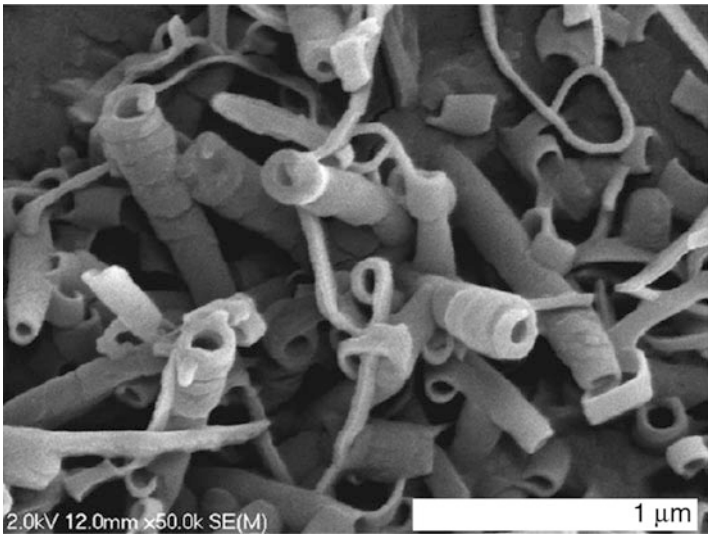
Plants with Coloration Caused by Diffraction Gratings

Glover and Whitney [13] reviewed the original research on iridescence in plants and presented very recent results on floral iridescence produced by diffractive optics. They identified iridescence in flowers of *Hibiscus trionum* (the inner part of its petals has an oily iridescence overlying red pigmentation) and tulip species (e.g., *Tulipa kolpakowskiana*) and demonstrated that iridescence was generated through diffraction gratings. The iridescence in the hibiscus is obvious to human eyes (appearing blue, green, and yellow depending on the angle from which it is viewed), whereas the iridescence in the tulip is in the ultraviolet spectrum, which bees can see, but people cannot. The surface striations in the diffraction gratings observed in the tulips are about 1 μm apart. Whitney and her coworkers investigated 22 tulip species from around the world and found ordered striations in 18 of them. They report iridescence generated by diffraction gratings in 10 families of angiosperms, tulips and hibiscus being just two of them: from the mallow family, the species example they give is *Hibiscus trionum*; from the lily family, *Tulipa* sp.; from the aster, daisy, or sunflower family, *Gazania klebsiana*; from the pea family, *Ulex europaeus*; from the Loasaceae family (of bristly hairy and often climbing plants), *Mentzelia lindleyi*; from the buttercup or crowfoot family, *Adonis aestivalis*; from the willow herb or evening-primrose family, *Oenothera biennis*; from the nightshades, *Nolana paradoxa*; from the iris family, *Ixia viridiflora*; and from the peony family, *Paeonia lactiflora*. According to this extensive list, it might well be that many more flowers are iridescent than previously thought – since they are iridescent in parts of the spectrum that people cannot see. The principal reason why not more species with diffraction-based coloration are known might simply be that until now nobody has looked.

Plants with Coloration Caused by Scattering

Some green leaves look white or silvery because microscopic air spaces in surface hairs reflect the light. Other leaves (such as the wax palm) and some fruits (such as plums and grapes) have a white "bloom" which is a surface deposit of wax.

Whitish, silvery, and other metallic finishes of leaves are generally produced by reflection of the light off microstructures, nonliving plant hairs called trichomes. In deserts, this increased reflection, especially of infrared radiation, off the



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 9 Monohelical tubular epicuticular wax structures in *Wollemia nobilis*. Scale bar 1 µm (From Ref. [19], reproduced with permission)

microstructures is important since it reduces heating of the leaf. However, a reflecting leaf has a much lower photosynthetic capacity than would an equal leaf without the reflectant trichomes. The Desert Brittlebush *Encelia farinosa* produces nonreflectant green leaves during the cooler springtime, when it is beneficial to have warmer leaves.

Silvery, gray, white, or blue coloration in plants might arise due to light scattering on three-dimensional epicuticular wax structures (structural wax). Leaves with a bluish gray waxy surface are called glaucous.

Generally, radiation is scattered across the spectrum, with increased reflectance in the ultraviolet, visible, and infrared windows. The size, distribution, and orientation of wax crystals and other surface features determine the extent to which light is scattered at the plant surface. In some plant species such as the Blue Spruce (*Picea pungens*), preferential scattering of shorter wavelengths and enhanced reflectance of UV radiation occur.

Epicuticular wax is a complex mixture of long-chain aliphatic and cyclic compounds. The bluish leaf hues of the evergreen Desert Mahonia *Berberis trifoliolata*, for example, arise from wax structures on nipple-like projections from the epidermal cells, called epidermal papillae. In some cases, for example, in the Blue Finger, *Kleinia mandraliscae*, and in cabbage, *Brassica oleracea*, powdery wax can be rubbed from the surface, causing the gray or bluish hue to disappear and revealing the green leaf color beneath. The Wollemia Pine *Wollemia nobilis*, an evergreen tree with bluish green mature foliage, has monohelical tubular epicuticular wax structures with a diameter of about 100 nm (Fig. 9) [19]. Depending on the wax compound present, epicuticular wax crystals can have various shapes [20]: The glaucous eucalyptus tree (*Eucalyptus gunnii*) has nanoscale β-diketone epicuticular wax tubules; the blue-gray foliage of the Dusty Meadow Rue (*Thalictrum flavum glaucum*) has nanoscale nonacosanol tubules. The Black Locust Tree (*Robinia pseudoacacia*) has dark, dull, blue-green summer foliage with nanoscale epicuticular wax platelets that are arranged in rosettes. The leaves of Wild Cabbage (*Brassica oleracea*) show simple nanoscale rodlets, whereas the nanoscale rodlets on the Silky Sassafras (*Sassafras albidum*) leaves are transversely ridged.

The epicuticular wax in the skin of plums (*Prunus domestica* L.) consists of an underlying amorphous layer adjacent to the cuticle proper together with crystalline granules of wax protruding from the surface. The epicuticular waxes in olive leaves serve as radiation inceptors.

Coloration Caused by Photonic Crystals: Plants, Diatoms, and Viruses

Iridescent Blue Fruits

The fruits of the Blue Quandong (*Elaeocarpus angustifolius*, syn. *E. grandis*) and the Blue Nun (*Delarbrea michiana*) have brilliant blue coloration ([1] and references therein). The iridescent blue color of the quandong fruit is actually enhanced by wetting, not like in *Selaginella*, mentioned above, where it disappears when wet. Iridosomes, multilayered systems arranged in 3-D structures, similar to the structure yielding coloration in Morpho butterflies, are responsible for the coloration of these fruits. Iridosomes are different from the iridoplasts described above for leaves: They are secreted by the epidermal cells of the fruit and are located outside the cell membrane but inside the cell wall. The structures of the iridosomes are at least partly cellulosic. The stone of the silver quandong fruit is termed *rudraksha*. It is hard and highly ornamented and is used throughout India and Southeast Asia in religious jewelry.

Edelweiss

Edelweiss (*Leontopodium nivale* subsp. *alpinum*) is a European alpine flower that grows at high elevations up to 3400 m. Its body is covered with white hairs. Vigneron and coworkers showed in 2005 [21] that the internal structure of these hairs acts as a two-dimensional phonic crystal. The hairs are hollow tubes, about 10 μm in diameter, with an array of parallel striations with a lateral diameter of 176 nm around the external surface. Through diffraction effects determined by the optical fiber with photonic crystal cladding, the hairs absorb the majority of the UV light, thereby acting as an efficient sun block for the structures beneath.

Spores

Some ferns produce two types of spores: microspores and megaspores (size: some hundreds of micrometers). Hemsley et al. [22] reported iridescent megaspores in *Selaginella* (fossil and modern). They report that spore iridescence is not generally visible until the outer layers of the spore wall (with different organizations to the iridescent layer) are removed. The three-dimensional pattern yielding the iridescence, for example, in megaspores of *Selaginella exaltata*, is still unresolved. According to SEM images, an ordered structure of 240 nm diameter particles, perhaps an opal-like colloidal crystal that is generated via self-assembly, is the reason for the coloration.

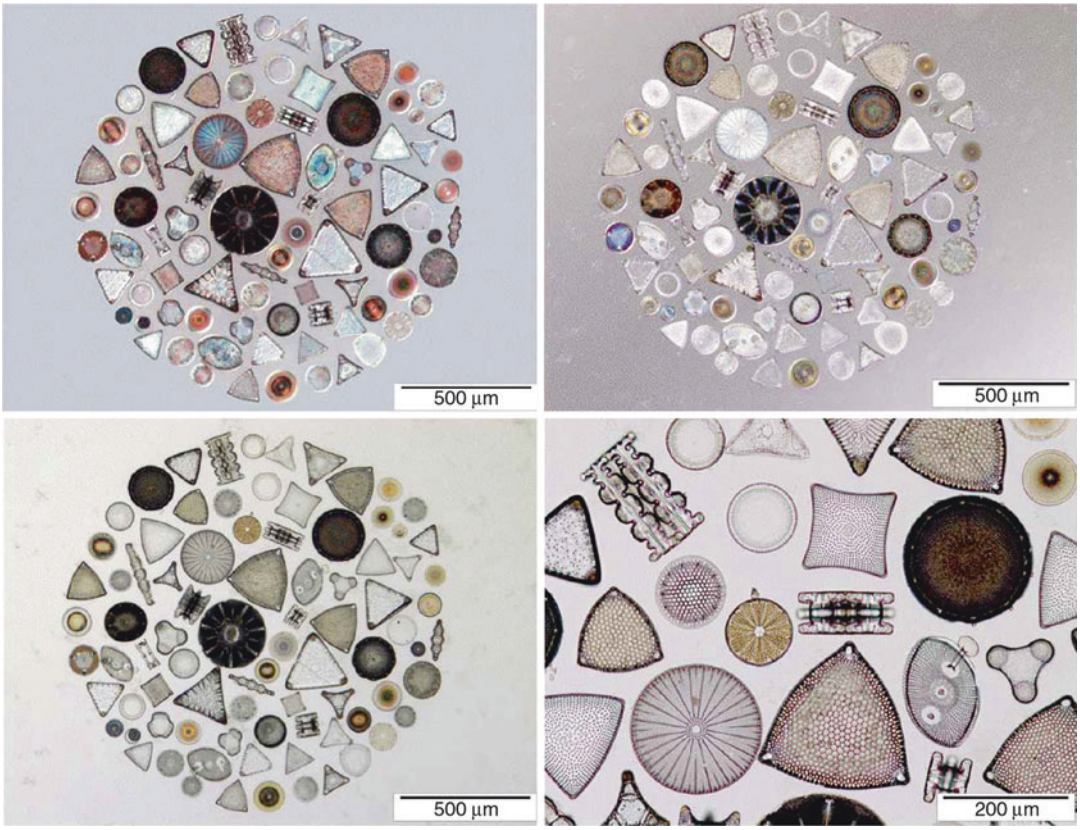
Tabasheer

Opal is the only gemstone that can be produced in biotic and abiotic ways. Tabasheer (vegetable opal, pearl opal) consists of hydrated silica and is produced when bamboo is hurt; the plant sap comes to the surface and dries in small nodules. These little nodules are opaque or translucent white, sometimes with a faint hint of blue. They can be polished as cabochons and are often used in the orient as jewelry. Tabasheer is highly porous and hygroscopic and has been used in ancient times to remove snake poison from the body – therefore, it is also known as snake stone [7].

Diatoms

Diatoms (Bacillariophyta) are found in both freshwater and marine environments, as well as in damp soils and on moist surfaces. They are unicellular microalgae with a cell wall consisting of a siliceous skeleton enveloped by a thin organic case. The cell walls of each diatom form a pillbox-like shell consisting of two parts that fit within each other.

Diatoms exhibit an amazing diversity of nanostructured frameworks (Fig. 10), including two-dimensional inverse photonic crystals. The cell wall exhibits periodic arrangement of pores in the micrometer to nanometer range – each diatom species has its own specific morphology. Individual diatoms range in size from few micrometers up to several millimeters, although only a few species are larger than 200 μm. A patch of diatoms can look like iridescent scum to the naked eye. Under the microscope, beautiful iridescence can be seen on the single-cell level, and a microscope slide covered with a monolayer of diatoms also exhibits iridescence to the naked eye (with



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 10 Sample with various diatoms from New Zealand, laid by hand by Elger in 1925. The slide was photographed by Friedel Hinz under the optical microscope with slightly varying angles of view, revealing color change in some of the diatoms (*top* images and *bottom left*). *Bottom right*: Zoom. Marine fossil diatoms from Gave Valley, New Zealand, Sample T1/21 prepared by Elger, 1925. Image (c) 2011, F. Hinz, AWI Bremerhaven, Germany (Image reproduced with permission)

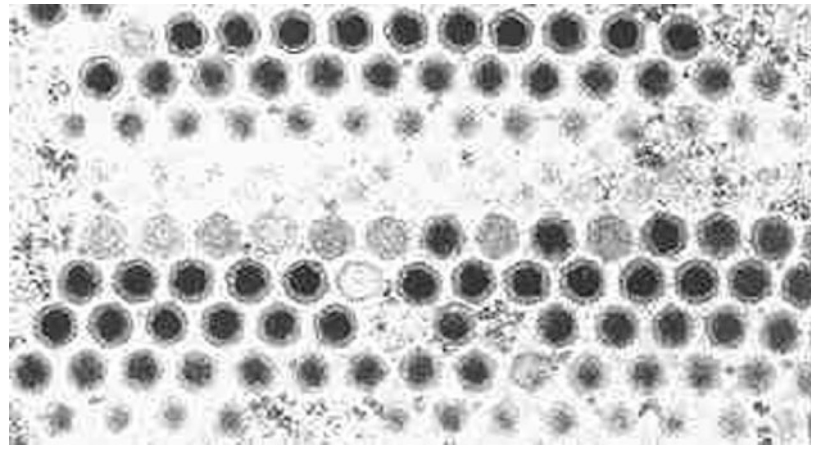
good illumination or in sunlight). Iridescent effects are used widely in color cosmetic products and personal care packaging, and there is great potential for using diatoms in this industry.

Viruses

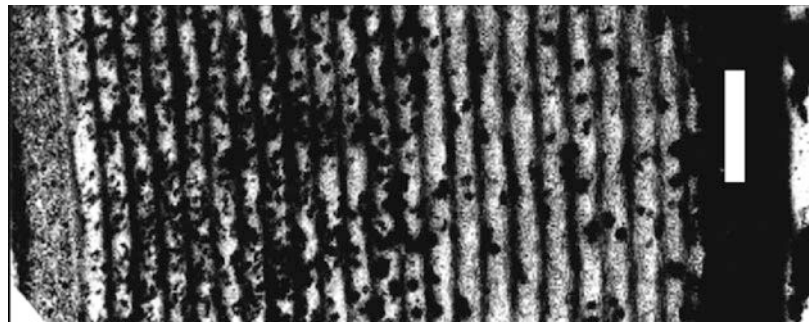
Iridescent viruses (Iridoviridae) are between 120 and 140 nm in diameter and infect insects, fishes, and frogs. Iridescent viruses produce systemic infections, with the highest concentrations in the outer layer of the skin (the epidermis) and the fat bodies. Insect larvae infected with this virus develop iridescent lavender-blue, blue, and blue-green coloration. In older infections, the iridescent color is generally more vivid, and frequently, there are also small, extremely brilliant islands. Electron microscope examination reveals numerous viruses in paracrystalline arrays (see Fig. 11) [23].

Plants with Coloration Caused by Cholesteric Liquid Crystals

In three iridescent tropical understory fern species, structurally modified chloroplasts with helicoidal structures, similar to the characteristic liquid crystals in beetles [2], might be the reason for the blue coloration. These ferns are *Danaea nodosa* (with many multilayers, each less than 100 nm thin), the necklace fern *Lindsea lucida* (17 layers, 192 nm spacing), and *Diplazium tomentosum* (20–23 layers, 141 nm spacing; see Fig. 12) [1].



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 11 Paracrystalline array of virus particles within an infected cell. This array gives rise to the iridescent phenomenon (Image Source: http://www.microbiologybytes.com/virology/kalmakoff/Iridoviruses.html. Permission pending)



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 12 Ultrastructure of a *Diplazium tomentosum* leaf. Scale bar 0.5 μm (Image © with one of the authors (DWL))

Structural Coloration Caused by Not Yet Described or Not Yet Identified Mechanisms

Macroalgae

Various red algae (Rhodophyta) and brown algae (Phaeophyta) exhibit iridescence and are completely unstudied. For example, of yet unknown origin is the bright blue, purple, or red iridescence in the red alga *Ochtodes secundiramea*.

Dictyota are widespread brown algae (Phaeophyta) along the Atlantic coast and very well known. *Dictyota mertensii* has blue/green iridescence; *Dictyota dichotoma* is also known as the purple peacock algae; *Dictyota barayresii* shows a translucent, iridescent blue; *Dictyota sandvicensis* exhibits a yellow-green iridescence; and *Dictyota humifusa* is light brown, often with brilliant blue iridescence. Iridescent bodies in the brown algae *Cystoseira stricta* are specialized vacuoles (vesicles) that contain numerous dense globules inside. The physics of their iridescence has not yet been clarified. The Hawaiian brown alga *Stypopodium hawaiiensis* has a beautiful blue or green color, and *Dictyopteris zonarioides* sometimes is iridescent blue [17].

Additional to the iridescence in red algae with multilayers responsible for the coloration (see section "Iridescent Macroalgae" above), the following red algae exhibit structural coloration: The Blue Branching Seaweed *Fauchea laciniata* shows deep red with iridescent blue; thin layer interference (whether from one layer or a multilayer system still needs to be clarified) is the reason for its coloration. *Fryeella gardneri* exhibits iridescent purple; *Cryptopleura ruprechtiana* is mildly iridescent. In *Chondracanthus corymbiferus*, the younger



Nanostructures for Coloration (Organisms Other Than Animals), Fig. 13 The iridescent Blue Strap Fern *Elaphoglossum herminieri*, at Sarapiqui, Heredia, near Puerto Viejo, La Selva Biological Station, Costa Rica. © 2008 R.C. Moran, The New York Botanical Garden (rmoran@nybg.org) [ref. DOL23374] (Image reproduced with permission)

blades are smooth and iridescent. Other iridescent red algae are *Chondria coerulescens* and *Drachiella spectabilis*. Also *Maripelta rotata* and *Botryoglossum farlowianum* are iridescent.

Slime Molds

As described above in section "True Slime Molds," the physical reason for the iridescent color in at least one type of slime mold is a single transparent thin layer. Slime mold iridescence is also described for some other species; however, its physical basis still needs to be determined. Spores of *Diachea subsessilis* exhibit dull greenish gray varying to blue iridescence; the peridium in *Diachea deviata* is iridescent and persistent. Beautiful bronze iridescent colors, sometimes tinged with blue, can be observed in the peridium of *D. subsessilis* (with the membranous peridium being colorless in water mounts, suggesting that pigments are not involved in the color production).

Plants

Bryophytes Bryophytes include the mosses, liverworts, and hornworts. The moss *Schistostega pennata* exhibits golden-green iridescence when it grows in caves where the light always comes from only one direction. This moss (also known as Dragon's Gold) shines like emerald jewels from the darkness of a rock crevice or cave. This unusual property is the result of lens-shaped cells with curved upper surface that focus the light on one point in the interior of the cell, where the chloroplasts aggregate. It remains to be determined if nanostructures are responsible for the coloration of this moss (Lee [1] and references therein).

Ferns *Elaphoglossum herminieri*, a strap fern that is native to tropical rain forests of the neotropics, shows iridescent blue color (Fig. 13) that is not removed by wetting. The reason of the coloration is unknown, as it is in *Elaphoglossum wurdackii* and *E. metallicum*. The oil fern *Microsorum thailandicum* (*Microsorum steerei*) from Southeast Asia has blue-green iridescent leaves. Also ferns of the family Vittariaceae can have iridescent scales. In this family, especially ferns of the genus *Antrophyum* (e.g., *Antrophyum formosanum*) and *Haplopteris* can be densely covered with iridescent scales. *Antrophyum mannianum*, a mountain fern liking full shade, has more or less iridescent leaves. There are also reports on an *Antrophyum* species found in Borneo (Mt. Mulu or Gunung Mulu) with an

iridescence that makes the plant look like it is covered with metallic eye shadow. *Haplopteris* ferns live in tropical and subtropical Asia. Obscure iridescence is exhibited by *Haplopteris amboinensis. H. doniana*, *H. taeniophylla*, *H. himalayensis*, *H. mediosora*, *H. fudzinoi*, *H. linearifolia*, *H. hainanensis*, *H. sikkimensis*, *H. elongata*, and *H. anguste-elongata* show bright iridescence. Also, *H. flexuosa* is iridescent, but not as bright. Malaysian individuals of the fern *Didymochlaena truncatula* are reported to be iridescent. The Venus Hairfern *Adiantum capillus-veneris* has iridescent stems, and the Aleutian Maidenhair *Adiantum aleuticum* has iridescent blackish foliage.

Grasses (Poaceae) Iridescence can be observed in the rosy inflorescences of the ornamental grass *Miscanthus sinensis*. Further grasses with startling iridescent inflorescences are the small grass *Pennisetum alopecuroides*, 'Little Bunny', and the even smaller variety 'Little Honey'. In late summer, these grasses are covered with blooms that look like hairy caterpillars.

Sedges (Cyperaceae) A sedge is a grass-like or rush-like plant growing in wet places having solid stems, narrow grass-like leaves, and spikelets of inconspicuous flowers. Leaves and seedlings of sedges can exhibit iridescence. The understory sedge *Mapania caudata* from Malaysia and also *M. graminea* exhibit iridescence in their leaves. The Smooth Black Sedge, *Carex nigra*, shows whitish iridescence, and the Autumn Sedge *Carex dipsacea* has iridescent olive green leaves. *Carex flagellifera* 'Toffee Twist' has iridescent, slender leaves.

The seedlings of the Pond Flatsedge *Cyperus ochraceus* are dark iridescent gray because of an outer one-cell thick covering of translucent cells, whereas the seedlings of the Marsh Flatsedge *Cyperus pseudovegetus* are brown with a very thin translucent-iridescent layer of cells. Also *Carex squarrosa* has blackish seedlings with iridescent superficial cells (when fully mature). In *Fimbristylis annua* and *F. vahlii*, the seedlings are often iridescent. Harper's Fimbry *F. perpusilla* has pale brown seedlings with iridescent tints, and the Southern Fimbry *F. decipiens* has seeds of whitened iridescent to brown.

Hypoxidaceae: Hypoxis The Star Grass *Hypoxis sessilis* has seedlings that are black but with iridescent membranous coats.

Orchids (Orchidaceae) Iridescent foliage in orchids occurs, for example, in *Macodes petola*, *Masdevallia* (*Byrsella*) *caesia*, and *Aulosepalum pyramidale*, where it is bright green with an iridescent shine. Flowers of many species of orchid (such as *Ophrys apifera* and the Mirror of Venus *Ophrys speculum*) have been reported to have an iridescent patch, the speculum. The shape and iridescence of this iridescent patch are thought to mimic the closed wings of female wasps or bees and sexually deceive the respective males, thereby pollinating the orchid (see [13] and references therein).

Others There are various further plant species that exhibit iridescence. Some of them are mentioned below. The main point is that there is a lot to study, and potentially several novel mechanisms of color production in organisms are yet to be identified.

- Purslanes (Portulacaceae): *Portulaca*. Seeds mostly glossy black or iridescent gray have been reported for various species of *Portulaca*. The reason of their iridescence is still unknown. In the Moss-rose Purslane *Portulaca grandiflora* Hooker, the mature seeds are steely gray and often iridescent; in the Pink Purslane *Portulaca pilosa*, the seeds are black, with very slight purplish iridescence when mature; in *Portulaca psammotropha*, the seeds are black, turning iridescent gray when fully mature; and in the Shrubby Purslane (Copper Purslane) *Portulaca suffrutescens* Engelmann and *P. halimoides*, the seeds are leaden and slightly iridescent.
- Grapes (Vitaceae): *Ampelopsis*. The Porcelainberry, *Ampelopsis brevipedunculata*,

has slightly iridescent fruits showing white, green, turquoise, blue, brown, and violet coloration.
- Lardizabalaceae (no common name): *Decaisnia*. The blue beans from the Blue Bean Shrub (*Decaisnea fargesii*) from Bhutan have iridescent, 4 in. large blue fruits.
- Pinks (Caryophyllaceae): *Spergularia*. Other plants with iridescent seeds are the Blackseed Sandspurry, *Spergularia atrosperma*, whose black seeds are often iridescent.
- Ice plants (Aizoaceae): *Sesuvium*. Another plant with iridescent seeds is the Slender Seapurslane, *Sesuvium maritimum*, with brownish-black, smooth, and somewhat iridescent seeds.
- Plantaginaceae (no common name): *Bacopa*. The Blue Waterhyssop, *Bacopa caroliniana*, has iridescent seeds.
- Goosefoot (Chenopodiaceae): *Chenopodium*. The White Goosefoot, *Chenopodium giganteum*, is an annual plant in which the young shoots are covered with a fine iridescent magenta powder.
- Rapataceae: *Stegolepis*. The leaves of *Stegolepis ligulata* have blue-green iridescence; even more impressive is the larger *S. hitchcockii*.
- Mallows (Malvaceae): *Hibiscus*. The African Rosemallow (*H. acetosella*) produces iridescent lobed maroon leaves on woody stalks.
- Palms (Arecaceae): *Chamaedorea*. Very impressive are the leaves of the Metallic Palm *Chamaedorea metallica*.
- Hydrangea (Hydrangeaceae): The United States patent PP18294 refers to the invention of a *Hydrangea macrophylla* plant named "HYMMAD I" whose inflorescences mature to an iridescent lime green.
- Starflowers (Hypoxidaceae): The flowering plant *Spiloxene capensis* can have white, cream yellow, or pink flowers with a dark center. In the pink form, the center is iridescent blue green, whereas in the white form, the center is iridescent bronze green.
- Ice plants (Aizoaceae): The inflorescences of the Hardy Pink Ice Plant *Delosperma cooperi* 'Mesa Verde™' have iridescent hues of salmon-pink.
- Gesneriads (Gesneriaceae): *Rhynchoglossum obliquum* is a highly shade-resistant flowering plant with deep blue iridescent inflorescences.
- Sea hollies (Apiaceae): *Eryngium maritimum*, Seaside Eryngo, and *E.* x *tripartitum*, Sea Holly 'Big Blue', have surprisingly iridescent blue flowers.
- Spikemoss (Selaginellaceae): *Selaginella kraussiana* 'Gold Tips', with the common names Golden Clubmoss and Krauss's Spikemoss, is a spikemoss found naturally in the Canary Islands, the Azores, and parts of mainland Africa. It has dark green foliage with golden tips.
- Begoniaceae: *Begonia*. Many begonia species have iridescent leaves; however, the coloration of the leaves is strongly dependent on the environment the plants are grown in. In *Begonia Rex*, the leaves can be of shiny metallic red. A begonia plant named 'Lady Rose' with iridescent foliage, flower bud, petals, and tepals was patented by Mr. Terry McCullough in December 1998 (patent number plant 10,736). There are no reproductive organs formed in this invention. Other iridescent begonias are *B. burkillii*, *B. limprichtii*, *B. congesta*, *B. hahiepiana*, and *B. sizemoreae* and the hybrids *Begonia* bandit, Bethlehem star begonia, *B*. comedian, *B*. Palomar prince, *B*. wild pony, *B*. his majesty, *B*. merry Christmas, *B*. little darling, *B*. max gold, *B*. stained glass, *B*. regal minuet, *B*. shirtsleeves, *B*. Venetian red, and *B*. hocking tutu terror. The authors have observed several iridescent members of this genus in Malaysia, besides the *B. pavonina* described previously.

Conclusion and Outlook

In this entry, the authors give a review of the physical mechanisms leading to structural colors, introduce plant species as well as microorganisms and viruses with the respective nanostructures

thought to be responsible for or contribute to the coloration, and present quite a number or organisms where the physical reason for the structural coloration has not yet been determined. Many animals have senses that either work in different bandwidths compared to the ones in humans, or they sense completely different properties (such as with echolocation or the magnetic sense). Since plants and animals interact on various levels and for various reasons, plants have evolved various properties tuned to animals.

Structural colors are increasingly being used in current and emerging technology as well as in the arts. Some examples: The Austrian company Attophotonics develops structural colors as sensors for lab-on-a-chip applications. The Japanese company Teijin Fibers Limited produced one blue dress made from Morphotex® structurally colored fiber, a flattened polyester fiber mimicking Morpho butterfly structures. Qualcomm's Mirasol™ display technology was inspired by butterfly wings. ChromaFlair® and SpectraFlair® color-shifting paints from the company JDSU create color-shifting effects in cars. The emerging contemporary design practice Biornametics explores a new methodology to interconnect scientific evidence with creative design in the field of architecture, with role models from nature (e.g., structural colors in organisms) being investigated and the findings applied to design strategies.

More often than not, natural structural colors are the inspiration for such technical products and applications. The field of biomimetics is dealing with the identification of deep principles in living nature and their transfer to humankind [24]. General biological principles identified by the German biologist Werner Nachtigall that can be applied by engineers who are not at all involved in biology are, for example, integration instead of additive construction, optimization of the whole instead of maximization of a single component feature, multifunctionality instead of mono-functionality, energy efficiency, and development via trial-and-error processes. We are hopeful that biomimetic colors will not only be tailored to show the bright and shiny properties as the natural examples do but will also exhibit one of the most important properties ensuring continuity of the biosphere: sustainability. Currently, with nanotechnology as booming new promising field, the sciences have started to converge, with nanobioconvergence as one of the most promising examples. New materials, structures, and processes can arise from this novel approach, perhaps exhibiting some of the multifunctional properties of the inspiring organisms, where various levels of hierarchy (with functionality on each and every level) are so refined. The convergence of the ways of thinking and novel approaches promises an interesting future.

Cross-References

- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Biomimetics
- ▶ Biornametics: Architecture Defined by Natural Patterns
- ▶ Lotus Effect
- ▶ Nanoimprinting
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanostructured Functionalized Surfaces
- ▶ Nanostructures for Photonics
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Nanotechnology
- ▶ Optical and Electronic Properties
- ▶ Rose Petal Effect
- ▶ Scanning Electron Microscopy
- ▶ Self-Assembly of Nanostructures
- ▶ Self-Repairing Materials
- ▶ Transmission Electron Microscopy

References

1. Lee, D.: Nature's Palette: The Science of Plant Color. University of Chicago Press, Chicago (2007)
2. Berthier, S.: Iridescences: The Physical Colors of Insects. Springer, New York (2007)
3. Kinoshita, S.: Structural Colors in the Realm of Nature. World Scientific Publishing Company, Singapore (2008)
4. Vukusic, P., Sambles, J.R.: Photonic structures in biology. Nature **424**, 852–855 (2003)
5. Whitney, H.M., Kolle, M., Andrew, P., Chittka, L., Steiner, U., Glover, B.J.: Floral iridescence, produced by diffractive optics, acts as a cue for animal pollinators. Science **323**(1), 130–133 (2009)

6. Martin, J.T., Juniper, B.E.: The Cuticles of Plants, p. 109. Edward Arnold, Edinburgh (1970)
7. Eckert, A.W.: The World of Opals, pp. 16 and 183. Wiley, New York (1997)
8. de Gennes, P.G., Prost, J.: The Physics of Liquid Crystals. The International Series of Monographs on Physics. Oxford University Press, Oxford (1995)
9. Williams, T.: The Iridoviruses. Adv. Vir. Res. **46**, 345–412 (1996)
10. Gentner, G.: Über den Blauglanz auf Blättern und Früchten. Flora Bot Zeitung **99**(4), 337–354 (1909) (in German)
11. Mohl, H. v: Über die blaue Färbung der Früchte von *Viburnum tinus*. Flora Bot. Zeitung **27** (1870) (in German)
12. Lee, D.W., Lowry, J.B.: Physical basis and ecological significance of iridescence in blue plants. Nature **254**, 50–51 (1975)
13. Glover, B.J., Whitney, H.M.: Structural colour and iridescence in plants: the poorly studied relations of pigment colour. Ann. Bot. **105**(4), 505–511 (2010)
14. Inchaussandague, M., Skigin, D., Carmaran, C., Rosenfeldt, S.: Structural color in myxomycetes. Opt. Express **18**(15), 16055–16063 (2010)
15. Diah, S.Z.M., Karman, S.B., Gebeshuber, I.C.: Nanostructural colouration in Malaysian plants: lessons for biomimetics and biomaterials. J. Nanomater. **2014**, Article ID 878409(15p) (2014)
16. Gould, K.S., Lee D.W.: Physical and ultrastructural basis of blue leaf iridescence in four Malaysian understory plants. Am. J. Bot. **83**(1), 45–50 (1996)
17. Gerwick, W.H., Lang, N.J.: Structural, chemical and ecological studies on iridescence in *Iridaea* (Rhodophyta). J. Phycol. **13**(2), 121–127 (1977)
18. Pedersén, M., Roomans, G.M., Hosten, A.V.: Blue iridescence and bromine in the cuticle of the red alga *Chondrus crispus* Stackh. Bot. Mar. **23**, 193–196 (1980)
19. Dragota, S., Riederer, M.: Epicuticular wax crystals of *Wollemia nobilis*: morphology and chemical composition. Ann. Bot. **100**, 225–231 (2007)
20. Koch, K., Barthlott, W.: Superhydrophobic and superhydrophilic plant surfaces: an inspiration for biomimetic materials. Phil. Trans. R. Soc. A **367**(1893), 1487–1509 (2009)
21. Vigneron, J.P., Rassart, M., Vértesy, Z., Kertész, K., Sarrazin, M., Biró, L.P., Ertz, D., Lousse, V.: Optical structure and function of the white filamentary hair covering the edelweiss bracts. Phys. Rev. E **71**(8p), 011906 (2005)
22. Hemsley, A.R., Collinson, M.E., Kovach, W.L., Vincent, B., Williams, T.: The role of self-assembly in biological systems: evidence from iridescent colloidal sporopollenin in *Selaginella* megaspore walls. Phil. Trans. R. Soc. Lond. B **345**, 163–173 (1994)
23. Willis, D.B. (ed.): Iridoviridae. Current Topics in Microbiology and Immunology, vol. 116. Springer, Berlin/Heidelberg/New York/Tokyo (1985)
24. Gruber, P., Bruckner, D., Hellmich, C., Schmiedmayer, H.-B., Stachelberger, H., Gebeshuber, I.C. (eds.): Biomimetics – Materials, Structures and Processes. Examples, Ideas and Case Studies. Springer, Berlin (2011)

Nanostructures for Energy

Stefano Bianco[1], Angelica Chiodoni[1], Jijeesh R. Nair[2], Claudio Gerbaldi[2] and Marzia Quaglio[1]
[1]Center for Space Human Robotics, Fondazione Istituto Italiano di Tecnologia, Torino, Italy
[2]Department of Applied Science and Technology – DISAT, Institute of Chemistry Politecnico di Torino, Group for Applied Materials and Electrochemistry – GAMELab, Torino, Italy

Synonyms

Nanomaterials; Nanoparticles; Nanorods; Nanostructures: Nanodevices; Nanotubes; Nanowires

Definition

Objects whose dimensions range between 100 and few nanometers can be classified as nanostructures. Since the physical-chemical properties of matter change dramatically at the nanometer scale, the introduction of this classification results in opening a particularly important new field of investigation for next-generation materials and devices.

Introduction

The worldwide need for power is mainly satisfied by the use of well-known energy resources such as oil, coal, hydropower, natural gas, and nuclear power. The environmental impact of most of those energy sources is tremendous, mainly

because of the emission of greenhouse gases that causes the increase in the globally averaged temperature. In the next decades, a rise of the global energy demand is expected, mainly because of the growth of world population especially in the developing countries. It is interesting to notice how the preindustrial world's energy consumption grew about 0.3 terawatts (TW) – year per year – and this value dramatically increased to 14 TW in years 2000s, and it is expected to double to 28 TW to year 2050. In the same period, the world population increased only five times, clearly showing how the energy consumption is not uniformly distributed. An improvement in life quality in developing countries will cause a further tremendous increase in energy demand in the next future [1, 2]. A fast-growing country needs far more energy than a developed country, essentially because a mature industrialized economy has learned how to produce and use energy more efficiently [2].

In this scenario, exploring renewable, efficient, and green energy sources is the greatest challenge to be won for sustainable human progress. Innovation is one of the key strategies to address this challenge, and some of the most promising solutions are related to nanotechnologies. Materials at the nanometer scale are able to show completely different properties than those we are used to in our everyday lives. Understanding the huge potentialities of materials at the nanoscale and learning how to drive them in real devices open new extraordinary opportunities. The field of alternative energy is certainly the most challenging platform for some of nanotechnology's most exciting contributions [3]. New nanomaterials and nanostructures showed recently their potentiality in the development of devices for both energy production and storage like third-generation photovoltaic, fuel cells, and lithium-based batteries.

Solar energy conversion systems are limited mainly by the poor properties of current materials, showing limited charge-carrier mobility and significantly high charge recombination rates for materials with sustainable costs. Moreover, theoretical limitations (Shockley/Queisser limit) are well established and fix a superior insurmountable efficiency threshold. Nanotechnology could help in the design of innovative architectures for new-generation solar cells, actually opening to a possible real alternative to traditional silicon solar cells. Significant examples are dye-sensitized solar cells (DSCs). In the last few years, significant efforts were spent in designing new dye molecules with an improved optical absorption coefficient as well as optimized mesoporous semiconducting layers with lower charge recombination rates. Devices based on photosensitized nanostructured TiO_2 films and on liquid electrolytes have demonstrated light-to-current conversion efficiency greater than 12 % [4]. Other approaches, like hybrid organic solar cells based on conjugated polymers or quantum dot solar cells with a light harvesting performed by semiconductor nanocrystals, are showing promising laboratory results, even if the reaching of satisfactory efficiency values still remains a challenge, mainly because of consistent losses during photoinduced charge separation and charge transport across the nanoassemblies.

Fuel cells represent another technological field in which new concepts from nanotechnology can be beneficial. The development of this kind of devices for large-scale use is still limited by the lack of adequate systems for hydrogen storage. At the same time, nanotechnologies offer new interesting opportunities, directly storing hydrogen into the fuel cells by nanostructures of carbon, zeolites, stacked clays, or light metals hydrides [5]. The introduction of nanoengineering also opens to the design of new electrodes. Nanotechnologies can lead a more reduced use of precious metals (e.g., platinum nanoparticles for high surface area and low volume) along with improved membrane function and durability [6].

Used batteries and accumulators represent a tremendous environmental problem in current days. Nanomaterials could help not only in developing batteries showing higher energy densities but also devices that will reduce the use of disposal energy storage systems [7]. For example, it was demonstrated that silicon-based nanowires could significantly improve the specific capacity of Li-ion accumulators, currently powering up electronic devices such as laptops, video cameras, cell phones, and many others [8]. In fact, such a

new technology can grant a ten times increase of the capacity of traditional Li-ion batteries. Also, nanotechnology can reduce the possibility of batteries catching fire by providing less flammable electrode materials.

New applications of nanotechnology are underway in the energy field. Their use will undoubtedly impact our society, hopefully changing and improving the way we are currently thinking to energy production and storage, leading to a more sustainable human growth.

Energy Supply Systems

The traditional approach to energy supply is based on the production of huge amounts (from MW to GW) of energy in localized production sites and its following delivery for minute use, causing significant losses across the distribution net. Petroleum and coal industry, nuclear power stations, hydroelectric dams, wind installations, and geothermal power installations are common examples of this approach (concerning both nonrenewable and renewable energy sources).

Nanotechnologies are expected to dramatically increase the efficiency of alternative energy devices in all the different fields [3], creating the starting point for a new approach to a more sustainable energy management. In fact, it is reasonable to believe that the current scenario will be slightly modified in the next future. The centralized production of high quantities of energy will be more intensively joined by a distributed system of energetic hot spots, able to produce very low amounts of energy, with high spatial density. By this way, the production of energy occurs in direct connection or in close proximity with its use in devices (often miniaturized or microdevices), thus avoiding losses. In this framework, the efficiency of the single energy harvester could be no more a crucial point, while other aspects (like flexibility, integration in complex systems, and even aesthetics) gain importance.

This kind of approaches covers a broad interest in different fields, starting from new-generation buildings to new smart portable devices. The firsts are designed and fabricated according to a sustainable construction approach and are able not only to reduce the total energy consumption but also to produce a large part of the energy required inside the building [9]. The next future devices (as household electrical appliances, cellular phones, computers, and biomedical devices) will be able to manage energy more efficiently, thanks to the introduction of new power supply systems. These systems will present an energy source, hopefully a renewable one as solar, harvesting, or thermoelectric, connected to a storage system as lithium-ion batteries, both managed by proper electronic power management circuits [10]. In this new futuristic scenario, the role of nanomaterials and nanostructures in devices for energy production and storage will be more and more crucial.

Two complementary strategies can be considered to manage energy in energy supply systems. A first class of devices, such as batteries, fuel cells, and supercapacitors, needs a finite quantity or a continuous supply of an active material or fuel to properly work. Devices of this kind can be classified as non-regenerative power supplies. A second class of tools, such as photovoltaic cells, harvesters, and thermoelectric devices, do not need any material or fuel to work and can be classified as regenerative devices. In the next sections, some valuable examples of both categories will be briefly introduced.

Nanostructures for Energy Production

Dye-Sensitized Solar Cells (DSSCs)

Dye-sensitized solar cells (DSSCs) were invented in the early 1990s by Michael Grätzel and Brian O'Regan at the École Polytechnique Fédérale de Lausanne (CH) [11], following extensive research in semiconductor photoelectrochemistry that started in the 1970s. These cells have immediately drawn a huge interest in the scientific community, because of their potentiality for the fabrication of cells with relatively high efficiency and potential low cost. In fact, these kinds of solar devices can be produced with abundantly available organic materials and widely diffused inexpensive semiconductors, using simple and scalable

technologies and avoiding costly and energy-intensive high-vacuum treatments and material purification steps.

The working principle of the cell [4] is based on the presence of a molecular light absorber, which is able to harvest photons from sunlight and convert them to electrons, pumped up from a lower to a higher energetic level creating an electric potential that can be used to produce electric work. The light-absorbing dye is attached as a monolayer to the surface of a nanocrystalline film of wide band gap semiconductor (TiO_2 in a pure anatase form in the most common DSSC architecture). The semiconductor works as electron collector, since the photogenerated charges are fast injected in the conduction band of the nanostructured solid. A careful design of the molecular energy levels of the dye makes the charge transfer toward the oxide strongly favored from an energetic point of view, thus drastically reducing the electron-hole recombination process. The dye is electrically regenerated by charge donations from an electrolyte, which penetrates the pores of the nanofilm to have a close proximity with the molecules of the sensitizer, creating a 3D interpenetrating network junction. The electrolyte is an ionic conductor able to accept hole from (or donate electrons to) the oxidized dye, and the best performance is noticed for liquid electrolytes containing a I^-/I_3^- redox system with a ruthenium-based sensitizer or a cobalt redox mediator with a donor-pi-acceptor porphyrin dye. The electrons percolate through the nanoparticles, which are interconnected because of a short sintering process, and are collected in an external circuit, thus generating electric power without causing any permanent chemical transformation inside the cell.

The role of the nanostructures in designing suitable photoelectrodes for high-performance cells is crucial. In fact, as reported from the first experiments of Grätzel, it emerges that it is possible to move from a sunlight-to-current conversion efficiency of 1.2 % using a polycrystalline anatase film [12] to the 7.9 % using a nanostructured mesoporous TiO_2 layer [11], just varying the active exposed surface area of the electrode. In fact, on a flat surface a monolayer of dye is able to absorb only a little percentage of the impinging photons, since its optical cross section is sensibly smaller with respect to the geometrical area it occupies. Thus, for good collection efficiency, it is mandatory that a great amount of area is exposed to the incident flux of photons, and this is possible only with a nanostructuring of the photoanode. By carefully controlling the porosity of the film, it is possible to find the optimal equilibrium between the thickness of the film and the path that photogenerated charges are able to cover (e.g., the diffusion length) before recombination.

Harvesting

Energy harvesting is the direct extraction of background energy from the environment and its conversion in electrical power [13]. Harvestable ambient energies can be classified in three main classes, namely, radiations (see paragraph 3.1), kinetic/vibration energy, and thermal energy. Energy recovery is possible through properly designed devices. The heart of energy harvesting devices is an active material that is actually able to change/modify at least one of its properties as a response to the stimulus of ambient energy. The change of the material behavior, induced by environmental energy, is usually turned into electrical power. The diffusion of harvesting devices has been very limited till now by their low efficiency in electrical energy production if compared to traditional energy sources, mainly due to the lack of well-performing materials. The development of new nanostructures and nanomaterials is promising to fasten up the development of a new generation of highly efficient devices for energy harvesting [3].

Energy from Vibrations

Several sources of environmental vibrations can be observed and used to harvest energy, such as vibrations from household appliances, moving equipment, trains, helicopters flying overhead, or human movement. Micro-generators based on microelectromechanical system (MEMS) technology have a leading role to extract and recover that energy from the ambient.

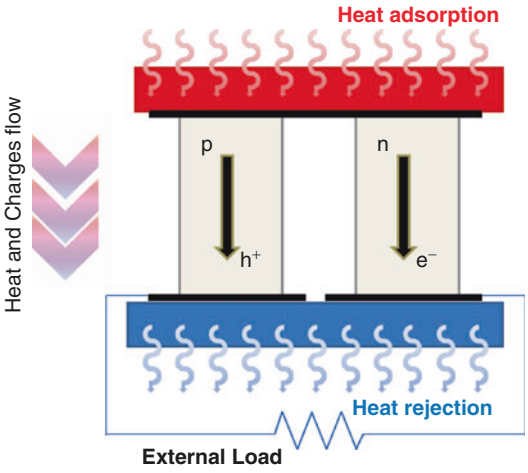
In piezoelectric-based energy harvesting, a piezoelectric material is needed as the active

material. Piezoelectric materials are able to directly convert the mechanical strain induced by vibrations into an electrical energy. The conversion is possible through the structural deformation of the material that induces a local spatial charge separation through an electric dipole. Piezoelectricity can occur in polycrystalline ferroelectric ceramic materials such as barium titanate ($BaTiO_3$) and lead zirconate titanate (PZT) and polymeric materials as poly(vinylidene fluoride) PVDF.

Traditional piezoelectric ceramics are very brittle, needing to be handled carefully to avoid breakage, and complicate bonding procedures. Their usability is also affected by their hard and fragile behavior since they have limited ability in following complex curved surfaces. This aspect is particularly limiting when sensors and harvesters are considered, since the ability of devices as harvesters to capture and extract energy is also related to the ability of the material (and so of the device) to well couple to the surface of the object generating the vibrations. Traditional polymeric piezoelectric materials can overcome this limitations being flexible and compliant, but unfortunately their electromechanical coupling coefficient is not comparable to traditional ceramics. These limitations have increased the attention to other approaches, and hybrid nanocomposite materials are one of the most promising alternatives. The final properties of nanocomposite materials are related to the connectivity between phases [10]. Nanocomposite materials are so promising for energy harvesting device because of high coupling factors, to which they add low acoustic impedance, mechanical flexibility, and broad bandwidth coupled with low mechanical quality factor that make them interesting candidates also for transducer and acoustic applications.

Energy from Temperature Gradients

A thermoelectric material is able to generate a voltage when a temperature gradient is applied to it (Seebeck effect) and vice versa (Peltier effect). The thermoelectric effect is related to the ability of charge carriers to move freely in metals and semiconductors (like molecules in a gas), and



Nanostructures for Energy, Fig. 1 Schematic of a thermoelectric device based on a p-n junction between semiconducting thermoelectric materials

while moving, they can conduct charge and heat. If a temperature gradient is applied to the material, charge carriers tend to move from the hot part to the cold region of the material through a diffusion process. While moving to the cold side, charges are accumulated, creating a voltage that can be externally exploited [14]. The conversion of thermal energy in electrical energy is possible through p-n junctions made of thermoelectric materials as described in Fig. 1.

Thermoelectric generators are characterized by no moving parts that make them silent, they are reliable and scalable, and so they are considered quite interesting as power generators. Efficiency for thermoelectric materials and devices can be described by a nondimensional figure of merit that can be evaluated as $ZT = \sigma S^2 T/\kappa$, where S is the Seebeck coefficient, σ is the electrical conductivity, k is the thermal conductivity, and T is the absolute temperature.

For a long time, thermoelectric materials have been considered too inefficient to be cost-effective for applications. The current state-of-the-art material in this field is a Bi_2T_3-based alloy showing a ZT close to one. New interest in thermoelectrics began in the mid of the 1990s, since theory suggested that thermoelectric efficiency could be significantly enhanced by the use of nanostructures. A well-performing thermoelectric

material needs a variety of conflicting properties to be well optimized to work properly. It needs a large thermopower, which means a large value of the Seebeck coefficient; it must be a good electrical conductor but showing at the same time a very low thermal conductivity. In traditional materials, these transport characteristics are dependent from one to the other, being deeply linked to interrelated material properties.

Based on both theoretical and experimental predictions, the current approach to increase ZT is to use nanostructured materials, both in form of films and for quasi-1D structures, like nanowires, nanobelts, and nanotubes. Nanostructured films are quite interesting because they can use the same fabrication routes of traditional materials, while showing improved thermal behavior, thanks to phonon scattering at the interfaces among nanograins. The main issue in nanostructured films is to control electron scattering at grain boundaries between randomly oriented grains that could decrease also electrical conductivity [14]. In one-dimensional (1D) materials as nanowires, the thermoelectric parameters (S, σ, and k) can be optimized independently. A high value of ZT is predicted for 1D nanostructures, as a result of an increase of charge mobility due to the quantum confinement of the density of the states, together with a reduction in thermal diffusivity because of scattering at the boundaries [3].

Fuel Cells

A fuel cell (FC) is an electrochemical cell that converts directly a chemical energy into electricity. This characteristic makes it an appealing system to provide a relatively cheap power source, by using hydrogen and ethanol as fuels and by producing water as waste.

Over the past few years, fuel cells have demonstrated increased reliability and lower costs because of the use of nanomaterials. Use of nanotechnology in the manufacturing process results in increased surface area and high aspect ratio, which originates large power and energy densities, a long shelf life, and ease of miniaturization; these features, combined with MEMS technology [10], are important for the development of a more powerful fuel cell for portable electronic devices. Nanomaterials are also increasingly used in the production, purification, and storage of hydrogen. In fact, at present, the ability to store hydrogen at high volumetric and gravimetric densities, as well as the ability to extract/insert it at sufficiently rapid rates, is still a hot topic that could condition the further improvements of the fuel cell technology. In particular, a big effort has been done by the research community in the past decades to identify candidates that could fit such strict characteristics. As an example of investigated nanostructures, there are CNTs, considered a promising solution, even if recently they were partially abandoned, light metal hydrides [5], graphene.

A fuel cell is mainly constituted by an anode, a cathode, and an electrolyte. At the anode, a catalyst oxidizes the fuel, usually hydrogen, turning it into a positively charged ion and a negatively charged electron. The electrons travel through a wire creating the electric current, while the ions travel through the electrolyte to the cathode. Once reaching the cathode, the ions recombine with the electrons and with another chemical, usually oxygen, to form water or carbon dioxide. The need to increase the performance of the fuel cells has nowadays motivated the researches to develop new materials and new synthesis approaches in order to gain the mentioned increased surface area and high aspect ratio of the different catalysts used for the different reactions at the anode and at the cathode.

There are many types of fuel cells, depending on the type of electrolyte used and on their working temperature.

In the low-temperature range, there are the proton exchange membrane fuel cells (PEMFCs), which use a polymer electrolyte membrane as electrolyte and work with above ambient temperatures up to 100 °C; phosphoric acid fuel cells (PAFCs), with phosphoric acid used as the electrolyte and working in the range 150–200 °C; and alkaline fuel cells (AFCs) that use a solution of potassium hydroxide in water as their electrolyte and operate in the range 150–200 °C.

In the high-temperature range, there are the solid oxide fuel cells (SOFCs) that use a hard ceramic compound of metal (like calcium or

zirconium) oxides as electrolyte and work in the range 800–1000 °C and molten carbonate fuel cells (MCFCs) that use high-temperature compounds of salt (like sodium or magnesium) carbonates as the electrolyte and work at 650 °C. SOFCs can run on hydrogen, butane, methanol, and other petroleum products.

Among the various types of fuel cells, PEMFC technology has become the dominant choice for the portable electronics market, a market that is rapidly increasing. Nowadays, nanotechnology is deeply embedded in the design of fuel cell components, and therefore, the development of nanostructured catalysts at the electrodes or optimized proton exchange membranes (PEMs) has become an active area of research for fuel cell applications. The challenging possibility to tailor the properties of new materials to dedicated designs had speed up the development of portable and more efficient devices.

At present, many efforts were devoted to obtain an optimization of the PEM characteristics, but despite great advancement in the development of nanostructured membranes, some restrictions and limitations remain. One of the challenges in current PEM fuel cell development is the lack of appropriate membrane materials to be used as proton-conductive membranes for improving the performance of the fuel cell.

The high cost of the commercially available Nafion® membrane, the most diffused polymeric electrolyte in FC, has induced the development of new proton-conducting membrane materials, including polybenzimidazoles, polyamides, polyether imides, polysulfones, polyphenylene sulfides, polyetheretherketones, and polyphenyquinoxalines. However, the Nafion® membranes show superior performance because of their high ionic conductivity compared to other non-fluorinated membranes. Proton conductivity is the most important property of membrane efficiency in fuel cells.

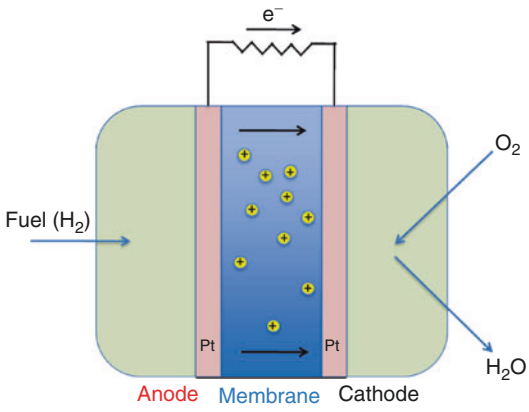
Membranes with acceptable conductivity at high temperatures are desirable for fuel cell applications, but there is always the issue of decreased performance at high temperatures due to dehydration of the membrane and a resulting reduction in ionic conductivity. To overcome the problems faced by proton exchange membranes, numerous studies were conducted to develop alternative materials for membrane preparation. One of the approaches to improve water retention in the membrane is to incorporate hydrophilic, inorganic materials [6]. With this method, it is difficult to avoid nanoparticle self-agglomeration in the polymeric matrix, and the method therefore induces nonhomogeneous dispersions and poor electrochemical performance. Because good inorganic nanoparticle-doped membranes require the dopants to be distributed evenly in the membrane and able to fill hydrophilic clusters in the Nafion® membrane, addition of nanosized inorganic particles is carried out by various methods, such as the sol-gel method, the solution casting method, and PECVD to obtain high-performance nanocomposite membranes. The use of surfactants can assist in creating a uniform distribution of nanoparticles as well.

A second challenge for the ultimate commercialization of fuel cells is the preparation of active, robust, and low-cost catalysts. Emphasis is placed on nanoengineering-based fabrication, processing, and characterization of multimetallic nanoparticles (transition metals) with controllable size (1–10 nm), shape, composition, and morphology (Fig. 2).

Another type of fuel cell that is drumming up an increasing interest as a potential candidate as portable device, and that differentiates it from the previously discussed fuel cells because it uses living organisms, is the microbial fuel cell (MFC).

MFCs are bioelectrochemical systems that can generate electricity by directly bioconverting the chemical energy of organic substrates into electrical energy through catalytic reactions of microorganisms under anaerobic conditions. In general, MFCs are able to oxidize biodegradable fuels, such as organic waste, to generate electrical power. The typical electrolyte is generally an aqueous solution or wastewater, acting as a feed source.

Some advantages of MFC over other types of fuel cells are that the reactions proceed at ambient temperature and ambient pressure, that they exploit waste rather than producing it, and that water is the only solvent used.



Nanostructures for Energy, Fig. 2 Basic operation of a PEM fuel cell. Hydrogen enters at the anode, where it is catalytically dissociated and ionized. The protons migrate by ionic conduction to the cathode through the polymer electrolyte membrane. Electrons travel through the external circuit where they do electrical work before arriving at the cathode. At the cathode, electrons, protons, and oxygen molecules form water molecules

Since MFCs satisfy at least 7 of the 12 principles needed to identify a chemical process as green, they belong to the class of "green chemistry."

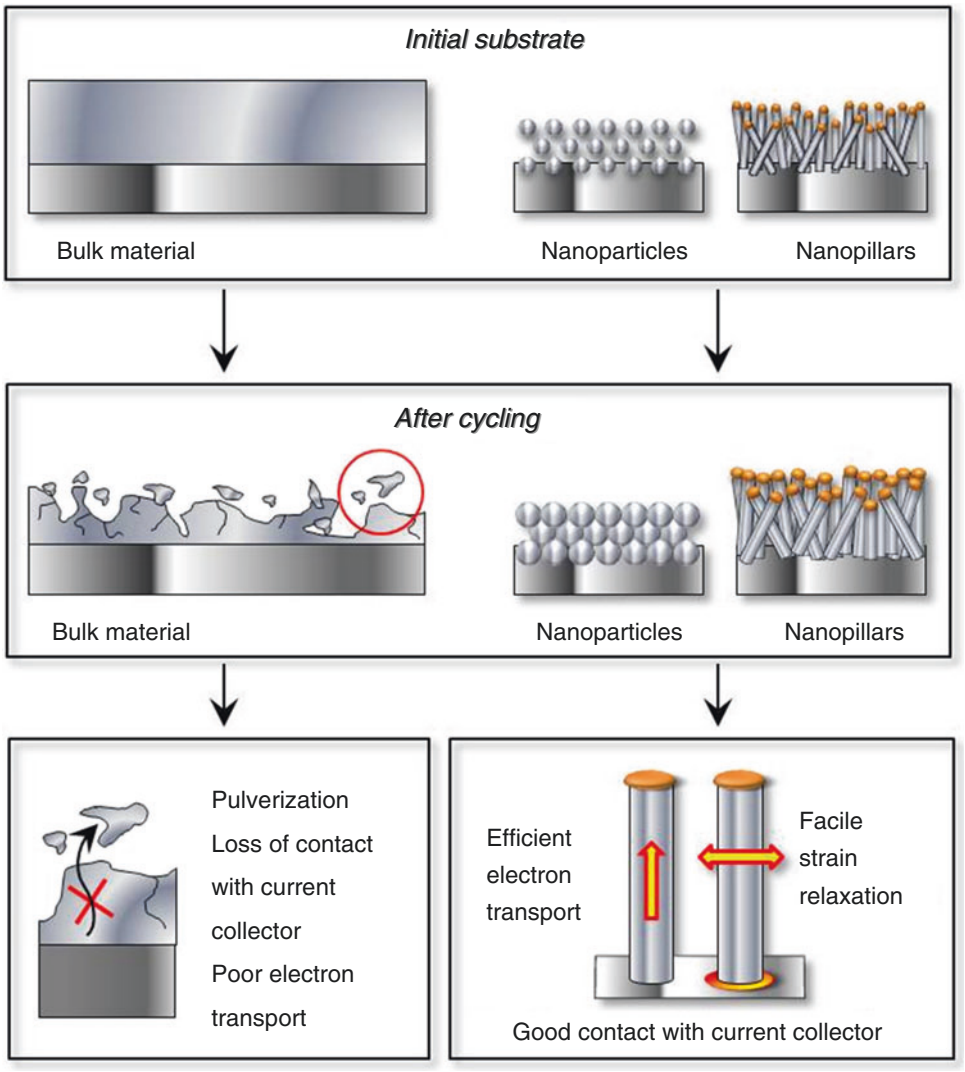
Nanostructures for Energy Storage Systems

Many of the sustainable energy alternatives described produce or require electricity. Therefore, novel more efficient ways to store electricity are very much needed in the way to a more sustainable production, transformation, and use of energy. In this respect, electrochemical systems, such as batteries and supercapacitors, which can efficiently store and deliver energy on demand in stand-alone power plants, as well as provide power quality and load leveling of the electrical grid in integrated systems, are playing a crucial role in the present energy economy. In addition, they are also envisaged as the power sources of choice for the progressive diffusion of sustainable hybrid/electric vehicles at high levels.

Lithium-Ion Secondary Batteries

Lithium-ion batteries (LiB) are one of the greatest successes of modern electrochemistry of materials. Their science and technology have been extensively reported in reviews [7] and dedicated books [8] to which the reader is referred for more details. A commercial LiB does not contain lithium metal; it comprises a negative electrode (generally, graphitic carbon) and a positive electrode (generally, layered lithium metal oxides such as $LiCoO_2$), both capable of reversibly intercalate Li^+ ions, these being separated by a nonaqueous lithium-ion conducting electrolyte (generally, a separator diaphragm soaked in a mixed ethylene-diethyl carbonate solution of $LiPF_6$). During discharge, Li^+ ions carry the current from the negative to the positive electrode, through the nonaqueous electrolyte [8]. During charge, an external electrical power source (the charging circuit) applies a higher voltage (but of the same polarity) than that produced by the battery, forcing the current to pass in the reverse direction. Lithium ions then migrate from the positive to the negative electrode, where they become embedded in the porous electrode material in a process known as "intercalation."

The current generation LiBs are increasingly being used by virtue of their high (volumetric and gravimetric) energy density, but although commercially successful, some disadvantages arise, related to relatively low power density and large volume change on reaction, safety, and costs. In fact, we are reaching the limits in performance using the current electrodes/electrolytes, and further breakthroughs in material chemistry are essential, especially modifying and improving already known materials. The aforesaid shortcomings can be reduced by the application of nanotechnology. It is important to design and fabricate nanostructured electrode materials that provide high surface area and short diffusion paths for ionic transport and electronic conduction. This gives rise to a larger electrode/electrolyte surface and enables Li-ion diffusion through the electrode material to be limited as much as possible. Moreover, the volume changes and structural evolution that occur as the electrodes accept and release ions



Nanostructures for Energy, Fig. 3 Example of how nanostructuring can help in improving the performances of Li-ion battery electrodes by reducing the strain associated with lithium ion insertion/removal. Readapted from Chan et al. [15]

can be quite different for nanoscale materials, thus greatly affecting the cycle/service life.

The potential advantages and disadvantages associated with the development of nanomaterials applied to LiBs can be summarized as follows [7]. The advantages include: (a) The strain associated with lithium-ion insertion/removal is better accommodated (along with a wider solid-solution region), thus improving service life as shown in Fig. 3. (b) New electrode reactions occur, which are not possible for materials composed of micrometer-sized particles. (c) A high surface area permits a more extensive electrode/electrolyte contact area, leading to higher charge/discharge rates. (d) The reduced dimensions significantly reduce the path lengths for electronic and Li^+ transport, thus permitting operation with low electronic conductivity and/or at higher power. As for disadvantages, the following have been identified: (a) An increase in undesirable

(side) reactions with the electrolyte due to the high electrode/electrolyte contact area, possibly leading to self-discharge, poor cycling, and calendar life. (b) The volume of the electrode increases because of poor particle packing, leading to lower volumetric energy densities unless special compaction methods are developed. (c) Potentially, nanoparticles may be more complex to synthesize especially that their dimensions may be difficult to control.

There are numerous examples in which nanostructured materials as both electrodes and electrolytes have been used in Li-ion batteries. In meso- and/or nanoporous architectures, for instance, the nanoscale domains of the material are accessible to molecular reactants because of rapid mass transport via diffusion through the continuous volume of mesopores. Very recently, B. Scrosati, Y.-K. Sun, and colleagues developed a high-capacity, nanostructured tin-carbon anode and a high-voltage, high-rate Li $[Ni_{0.45}Co_{0.1}Mn_{1.45}]O_4$ spinel cathode. When the two parts are put together, the result is a high-performance battery with a high energy density and rate capacity, highly suitable for powering low or zero emission HEV or EV vehicles [16]. The ultimate expression of the nanoscale in rechargeable lithium batteries is the formation of 3D nanoarchitectured cells, in which pillared anodes and cathodes are interdigitated [17]. The prospect of having active and passive multifunctional components interconnected within a 3D architecture is envisioned to lead to superior energy storage capacity, high-volume synthesis, and improved safety. A nanowire battery has been invented by a team led by Dr. Y. Cui at Stanford University in 2007 [15], which produces ten times the amount of electricity of existing LiBs. The team's invention consisted of a stainless steel anode covered with silicon nanowires, to replace the traditional graphite anode. Silicon, which stores ten times more lithium than graphite, allows a far greater energy density on the anode, thus reducing the mass of the battery. The large surface area further allows for fast charging and discharging and efficient electron transport down the length of each nanowire.

Lithium-Sulfur Secondary Batteries

The high energy density battery system development is highly lucrative to fulfill the mounting demand from personal devices and electric vehicles. LiBs have dominated the rechargeable high energy battery market until now, in fast-moving electronics and to an extent in HEV and EV, but future enhancements in energy density require a different approach that can overrun the existing battery chemistries and mechanisms [18].

The 1960s saw an initial step in the development of lithium-sulfur (LiS) batteries, which in recent years gain an absolute momentum due to the high theoretical capacity (1672 mAh g^{-1}) they may offer. LiS have a theoretical energy density of 2600 Wh kg^{-1}, originating from the electrochemical reactions between lithium metal and elemental sulfur [19]:

$$S_8 + Li^+ + e^- \rightarrow Li_2S_x \quad (1)$$

$$Li_2S_x + Li^+ + e^- \rightarrow Li_2S_2 \text{and/or } Li_2S \quad (2)$$

The reaction (Eq. 1) produces soluble lithium polysulfides, and the reaction (Eq. 2) leads to the formation of insoluble lithium sulfide (Li_2S_2 and/or Li_2S).

The overall electrochemical reactions comprise of solid-liquid-solid transformations that cause great complexity in LiS systems upon charge/discharge. Plenty of problems persist, restricting further development. Smartly designed sulfur-carbon nanomaterials (e.g., core-shell-structured composites) may overcome the obstacles; moreover, efficient trapping configurations for polysulfides will improve the utilization of sulfur during cycling. Recently, research efforts are focused on the cathode side of the cell where nanomaterials have been widely applied as conducting frameworks. In this respect, carbon nanotubes (CNTs), mesoporous carbons, yolk-shell TiO_2 spheres, carbon fibers, carbon spheres, graphene, carbide-derived carbon, graphene oxides, CNT/graphene hybrids, etc. have been implemented that demonstrated to greatly improve the electrode capacity. Several important strategies were evaluated including the

confinement of the sulfur and related reaction products within the micropores, hollow sphere structures, conducting polymer coatings, polysulfide reservoirs, carbon interlayers, and even the bags created by folding the graphene sheets to limit the diffusion of polysulfides to the electrolyte and subsequent decay of the cell performance.

Lithium-Air Secondary Batteries

One of the most vibrating upcoming terms in the energy sector is the report on the development of new energy storage system: the lithium-air battery, for short Li-air, the so-called breathing battery. The concept is attracting great attention due to the possibility of achieving high energy density comparable to internal combustion engine; it possesses the potential to one day replace the LiB technology of today's EVs.

The main difference between LiB and Li-air is the absence of the physical existence of an insertion cathode material (e.g., transition metal oxide), a key component active in the electric current flow, which is replaced by air [20]. Thus, the rechargeable Li-air battery is lighter with the potential to stack more energy than its commercial counterparts do. Li-air batteries can be architectured either in aqueous or nonaqueous configurations. The basics of nonaqueous rechargeable Li-air cell lie in the reversible formation and decomposition of lithium peroxide (Li_2O_2) at the cathodic side during the discharge and the recharge processes, respectively. The aprotic version was first investigated by Abraham et al. [21]. The reaction product Li_2O_2 deposits at the surface of the air electrode (cathode, positive). On the contrary, in an aqueous-based Li-air system, the reaction product ($LiOH•H_2O$) is soluble into the electrolyte solution.

One of the key challenges of the aqueous system is the low output power of the protected lithium electrode. In this case, protection of the anode is necessary, and it is achieved by laminating the lithium metal using NASICON-type lithium conducting solid electrolyte (LATP); such thin layers can isolate the lithium metal from directly encountering the water molecules. Such mixed metallic oxides are also proposed as electrolyte systems as well, and they can be developed and deposited using technologies familiar to nanomaterial preparation.

In the case of aprotic Li-air systems, one of the key points is the lowering of the Li_2O_2 oxidation overpotential at the cathode. It is widely accepted that electrocatalysts can play a vital role in improving the cycling ability of such cells; they can be placed in the cathodic part of the cell in very less quantity, and serious efforts are focused on their development. Thus, the electrocatalyst in the cathode of the Li-air cell can significantly influence the oxygen reduction reaction (ORR) and oxygen evolution reaction (OER). Nanostructured MnO_2 with different structures, along with noble metals, supported on various allotropes of carbon, is widely pursued due to its low cost, low toxicity, and good structural stability. There are other metals and metal oxides including Pd, Pt, Ru, PdO, and RuO_2 being studied in various shapes and forms that may influence the overall reaction kinetics. There are other areas where nanomaterials have been widely used in Li-air systems such as membrane to improve the ionic conductivity and, moreover, to prepare oxygen permeable membranes so that unpurified atmospheric air can be used as the source of O_2 for the formation of Li_2O_2. Thus, in every aspect of the lithium-based energy storage systems, nanomaterials are fundamental.

Electrochemical Capacitors

Electrochemical capacitors (ECs), also named "supercapacitors," are capacitors with relatively high energy density (typically on the order of thousands of times greater with respect to conventional electrolytic capacitors) [22]. The term supercapacitor is usually used to describe an energy storage device based on the charge storage in the electrical double layer (EDL) of a high surface area carbon in aqueous electrolytes. In fact, they store electrical energy, like batteries, but using a different mechanism: while batteries do it chemically, supercapacitors store electricity physically, by separating the positive and negative charges. Their lifetime is in principle infinite, as they operate solely on electrostatic surface charge accumulation. In comparison with conventional batteries or fuel cells, ECs also have a much

higher power density. They can be charged or discharged at a rate that is typically limited by current heating of the electrodes. So while existing ECs have energy densities that are perhaps 1/10 that of a conventional battery, their power density is generally 10–100 times greater. Supercapacitors can be used as uninterruptable power sources (UPSs), can be coupled with batteries to provide peak power, and can replace batteries for memory backup. The supercapacitor is ideal for energy storage that undergoes frequent charge and discharge cycles at high current and short duration. They have applications such as energy storage devices used in vehicles and for smaller applications like home solar energy systems where extremely fast charging is a valuable feature.

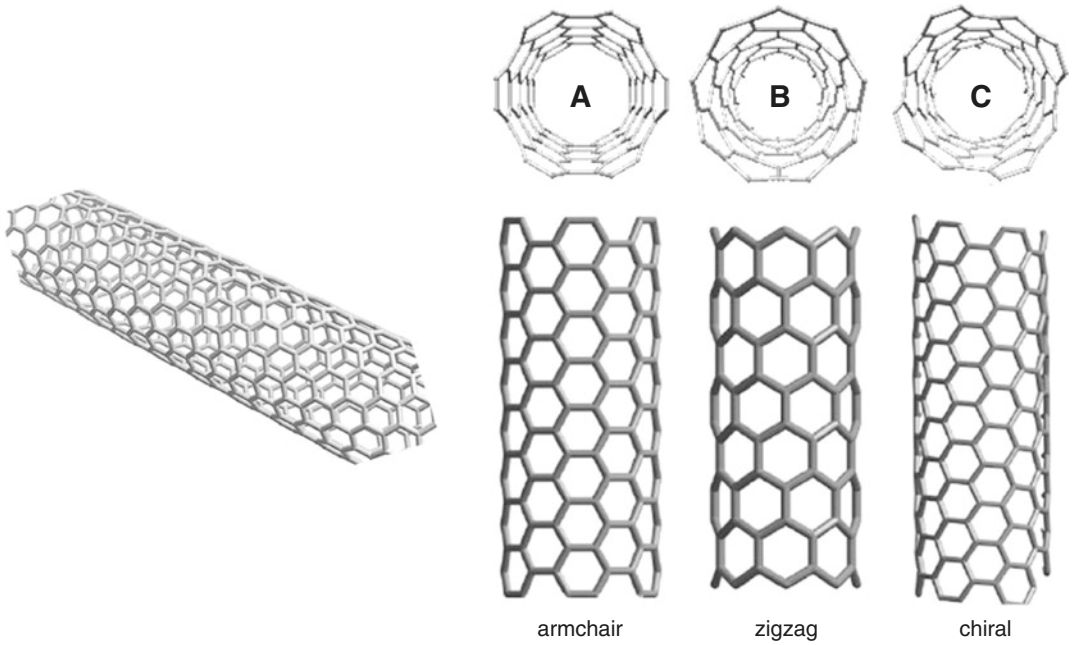
Since their discovery, these devices have attracted considerably less attention with respect to batteries as energy storage devices. Nevertheless, thanks to both the contribution of nanotechnology and the better understanding of charge storage mechanisms (ion behavior in small pores) the interest on ECs has noticeably increased very recently [22]. Nanotechnology is important because it enables large surface areas to be obtained to optimize performance in terms of capacitance and ionic and electronic resistance. There are numerous examples in the literature that show that nanosized crystallites and/or mesoporous materials exhibit significantly higher specific capacitance as compared to nonporous materials or micron-sized materials. The assembly of nanoscale materials is also important.

Recent trends in supercapacitors involve the development of high-surface-area-activated carbon electrodes to optimize the performance in terms of capacitance and overall conductivity. Attention has been focused on nanostructured carbons, such as aerogels, nanotubes, and nanotemplates [23]. One structure envisioned to be of interest is an array of vertically aligned carbon nanotubes where the spacing between the tubes is matched to the diameters of the solvated electrolyte ions. The main focus has been the optimization of interparticle contact resistance and electrolyte wettability of the pores. The capacitance increase is almost 50 % compared with the best performing commercially available activated carbons. Nanostructured metal oxide materials have also been investigated for supercapacitor applications, especially RuO_2 and $RuO_2 \cdot xH_2O$ nanotube composites [24].

The Impact of Nanostructures on Traditional Energy Sources: An Overview

As the worldwide demand for energy rises at an ever-increasing rate, there is a new urgency to improve the efficiency and sustainability of actual power generation technologies. One of the keys to addressing this challenge is innovation, and some of the most promising solutions are occurring at the nanoscale. Perhaps the most mainstream acceptance within renewable technology has come in the form of wind energy, which is approaching cost competitiveness with traditional energy sources. Countries such as Germany, Spain, and Denmark are already beginning to utilize substantial amounts of wind energy to meet their growing electricity needs. Nanotechnology impacts the wind industry by improving turbine performance and reliability to allow for longer lifetime, less fatigue failure, and lower costs of generation. New lubricants containing nanoparticles act like mini ball bearings, thus reducing the friction generated from turbine rotation, consequently improving its life cycle. Advancements in nanocoatings, such as de-icing and self-cleaning technologies, also help improving efficiencies, thus rendering ice and dirt buildup on the turbines virtually nonexistent.

In spite of the optimistic predictions of its proponents, alternative energy sources such as wind, hydro, and biomass are unlikely to seriously challenge coal, oil, natural gas, or nuclear power for a meaningful share of the worldwide energy needs in the near future. However, how these traditional energy sources are produced will rapidly change. In fact, nanotechnology is already at work



Nanostructures for Energy, Fig. 4 Structure of single-walled carbon nanotubes (SWCNTs). Carbon nanotubes can be metallic or semiconducting, depending on how their chicken wire-like lattice rolls into a tube

modifying how coal, oil, and gas are being produced. One of the more prominent examples can be found in the application of new nanoscale catalysts to the production and refining processes of our fossil fuel resources. Moreover, modern fossil fuel power plants, including coal-fired systems, require components that are able to withstand higher temperatures and pressures to improve efficiency, lower maintenance costs, and decrease emissions levels. In this field, nanotechnology-based innovation is being credited with improvements in efficiencies and life-spans of existing technologies, as well as the introduction of novel and disruptive power-generating components.

Nanotechnology has the potential to revolutionize the electric utility industry as well. Many of the components used in the electric utility industry can be modified to yield incremental improvements in the overall product performance. The end benefit is that electric utility providers can improve their operating margins by making existing equipment both last longer and more efficient. Nanotechnology can provide immediate improvements to the electrical utility system of moving power by efficiently improving the transmission of electrons, as well as creating newer and even better superconducting materials. The real payoff may be carbon nanotubes (CNTs), which are unbelievably thin, possess 100 times the strength of steel, have only 1/6 the density of aluminum, and efficiently conduct electricity. The most promising are the single-walled carbon nanotubes (SWCNTs). The broad range of uses for SWNTs is connected to the different nanotube properties according to how the hexagonal lattice is oriented (see Fig. 4), which together with the nanotube diameter determines the chirality. This combination of strength and conductivity suggests that in the future, energy providers should not have to dig up city streets to lay new wires. The existing lines could be simply replaced with super strong, highly conductive carbon nanotube wires.

Reduction of energy consumption is another active area in nanotechnology, targets being the

development of more efficient lighting, better combustion systems, and/or use of lighter and stronger materials in the transportation sector. Light-emitting diodes (LEDs) are interesting examples of how energy consumption can be reduced or "smart" windows, which can change reflective properties to attract or deflect sunlight depending on the customer's needs. Similarly, improved insulation materials could have a noticeable impact on overall energy demand. All of these products are made possible by advancements in nanotechnology. Nevertheless, the greatest long-term cost savings will probably come from the creation of "intelligent energy networks," that is, integrated systems of more energy-efficient computers and sensors that gauge and regulate customer energy use. Such networks already exist today, but as nanotechnology enables computers and sensors to become smaller, but correspondingly more powerful, the cost effectiveness of deploying such devices will increase, along with the number of possible applications.

Conclusions

New materials hold the key to fundamental advances in energy production and storage, both of which are vital in order to meet the challenge of global warming and the finite nature of fossil fuels. Nanomaterials in particular offer unique properties or combinations of properties as electrodes and electrolytes in a range of energy devices. Through innovation, we can improve the efficiencies of the technologies we have and discover new ways by which we can prosper. Nanotechnology provides us with the opportunity to attain sustainable development, thus overcoming one of the greatest challenges of our time, by using some of the simplest and smallest means at our disposal.

In conclusion, nanotechnology, known as paying attention to structures less than 100 nm and technology of twentieth century, had better be preferred owing to benefits such as being potentially inexpensive, effective, and promising for time to come. Hence, many companies and researchers are getting curious day by day in order to explore nanotechnology especially its energy application. Indeed, to minimize energy waste, it should become common and applicable as soon as possible in spite of the fact that it is a quite new technology.

Cross-References

- Carbon-Nanotubes
- Hybrid Solar Cells
- Nanomaterials for Electrical Energy Storage Devices
- Nanomaterials for Excitonic Solar Cells
- Nanostructured Materials for Sensing
- Piezoresistivity

References

1. US DOE Office of Basic Energy Sciences (Alivisatos, P., Cummings, P., De Yoreo, J., Fichthorn, K., Gates, B., Hwang, R., Lowndes, D., Majumdar, A., Makowski, L., Michalske, T., Misewich, J., Murray, C., Sibener, S., Teague, S., and Williams, E.): Nanoscience Research for Energy Needs. Report of the March 2004 National Nanotechnology Initiative Grand Challenge Workshop, 2nd edn, June 2005
2. Arunachalam, V.S., Fleisher, E.L.: Harnessing materials for energy. MRS Bull. **33**, 61 (2008)
3. Liu, J., Cao, G., Yang, Z., Wang, D., Dubois, D., Zhou, X., Graff, G.L., Pederson, L.R., Zhang, J.G.: Oriented nanostructures for energy conversion and storage. ChemSusChem **1**, 676–697 (2008)
4. Hardin, B.E., Snaith, H.J., McGehee, M.D.: The renaissance of dye-sensitized solar cells. Nat. Photonics **6**, 162–169 (2012)
5. Crabtree, G.W., Dresselhaus, M.S.: Harnessing of materials for energy – the hydrogen fuel alternative. MRS Bull. **33** (2008)
6. Thiam, H.S., Daud, W.R.W., Kamarudin, S.K., Mohammad, A.B.., Kadhum, A.A.H., Loh, K.S., Majlan, E.H.: Overview on nanostructured membrane in fuel cell applications. Int. J. Hydrog. Energy **36**, 3187–3205 (2011)
7. Aricò, A.S., Bruce, P., Scrosati, B., Tarascon, J.M., van Schalkwijk, W.: Nanostructured materials for advanced energy conversion and storage devices. Nat. Mater. **4**, 366–377 (2005)

8. van Schalkwijk, W., Scrosati, B. (eds.): Advances in Lithium-Ion Batteries. Kluwer/Plenum, New York (2002)
9. Sartori, I., Hestnes, A.G.: Energy use in the life cycle of conventional and low-energy buildings: a review article. Energy Build. **39**, 249–257 (2007)
10. Cook-Chennault, K.A., Thambi, N., Sastry, A. M.: Powering MEMS portable devices – a review of non-regenerative and regenerative power supply systems with special emphasis on piezoelectric energy harvesting systems. Smart Mater. Struct. **17** (2008)
11. O'Regan, B., Grätzel, M.: A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO_2 films. Nature **353**, 737–740 (1991)
12. Vlachopoulos, N., Liska, P., Augustynski, J., Grätzel, M.: Very efficient visible light energy harvesting and conversion by spectral sensitization of high surface area polycrystalline titanium dioxide films. J. Am. Chem. Soc. **110**, 1216–1220 (1988)
13. Angrist, S.: Direct Energy Conversion. Allyn & Bacon, Boston (1982)
14. Snyder, G.J., Toberer, E.S.: Complex thermoelectric materials. Nat. Mater. **7**, 105–114 (2008)
15. Chan, C.K., Peng, H., Liu, G., McIlwrath, K., Zhang, X.F., Huggins, R.A., Cui, Y.: High-performance lithium battery anodes using silicon nanowires. Nat. Nanotechnol. **3**, 31–35 (2008)
16. Hassoun, J., Lee, K.S., Sun, Y.K., Scrosati, B.: An advanced lithium ion battery based on high performance electrode materials. J. Am. Chem. Soc. **133**, 3139–3143 (2011)
17. Long, J., Dunn, B., Rolison, D., White, H.: Three-dimensional battery architectures. Chem. Rev. **104**, 4463–4492 (2004)
18. Bruce, P.G., Freunberger, S.A., Hardwick, L.J., Tarascon, J.-M.: Li–O_2 and Li–S batteries with high energy storage. Nat. Mater. **11**, 19–29 (2011)
19. Huang, J.-Q., Zhang, Q., Peng, H.-J., Liu, X.-Y., Qian, W.-Z., Wei, F.: Ionic shield for polysulfides towards highly-stable lithium–sulfur batteries. Energy Environ. Sci. **7**, 347–353 (2014)
20. Girishkumar, G., McCloskey, B., Luntz, A.C., Swanson, S., Wilcke, W.: Lithium-air battery: promise and challenges. J. Phys. Chem. Lett. **1**, 2193–2203 (2010)
21. Abraham, K.M., Jiang, Z.: A polymer electrolyte-based rechargeable lithium/oxygen battery. J. Electrochem. Soc. **143**, 1–5 (1996)
22. Conway, B.E.: Electrochemical Supercapacitors. Kluwer Academic/Plenum, New York (1999)
23. Frackowiak, E., Béguin, F.: Carbon materials for the electrochemical storage of energy in capacitors. Carbon **39**, 937–950 (2001)
24. Kim, I.H., Kim, J.H., Kim, K.B.: Electrochemical characterization of electrochemically prepared ruthenium oxide/carbon nanotube electrode for supercapacitor application. Electrochem. Solid State Lett. **8**, A369–A372 (2005)

Nanostructures for Photonics

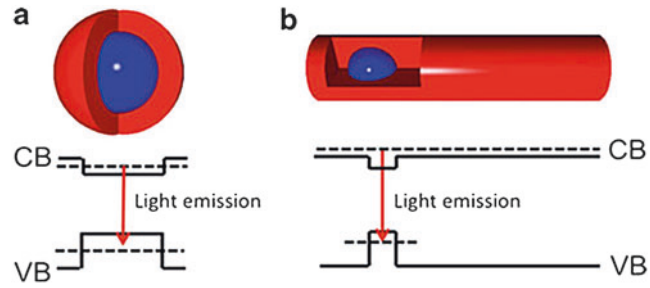
Andrea Toma[1], Remo Proietti Zaccaria[1], Roman Krahne[1], Alessandro Alabastri[1], Maria Laura Coluccio[1,2], Gobind Das[1], Carlo Liberale[1], Francesco De Angelis[1], Marco Francardi[2,3], Federico Mecarini[1], Francesco Gentile[1,2], Angelo Accardo[2,4], Liberato Manna[1] and Enzo Di Fabrizio[1,2]
[1]Nanobiotech Facility, Istituto Italiano di Tecnologia, Genoa, Italy
[2]Lab. BIONEM, Dipartimento di Medicina Sperimentale e Clinica, Università "Magna Grecia" di Catanzaro, Catanzaro, Italy
[3]International School for Advanced Studies (SISSA), Trieste, Italy
[4]Soft Matter Structures Group ID13 – MICROFOCUS Beamline, European Synchrotron Radiation Facility, Grenoble, France

Definition

"Nanostructures for Photonics" represents a collection of devices whose working principles are based on photons at the nanoscale level.

Overview

In the last decade, new disciplines emerged where the terms *nano* and *photon* happen to play a fundamental descriptive role. Their origin dates back to the ancient Greece, with *nano* (=νανοζ) indicating something of small dimensions, and *photon* (=φϖΨψρζ) indicating the light. In Physics, but also in Chemistry, Biology, Material Science, and many more fields these terms have been used for indicating the studies of extremely small objects by means of light. This new technological frontier has already produced some amazing results: LCD based mobile phones, nanoparticles for optical tagging of biological molecules and hyperthermia, or devices for single-molecule detection are just few of many examples where nanophotonics plays a crucial role. In terms of optical devices,



Nanostructures for Photonics, Fig. 1 Schemes illustrating the core-shell architectures and the related band structure of spherical (**a**) and dot/rod nanocrystal structures (**b**)

their functionality can be divided in three general key parts: *light generation, light guiding* and *light harvesting*. Here will be introduced some approaches describing these aspects, starting from colloidal nanocrystals for photonics to end, through the introduction of photonic crystals and metamaterials, with plasmonic devices for Raman spectroscopy. The aim is to prove that photonics at nanoscale level is not just an exotic research field but a mature discipline capable of offering devices for a multitude of uses.

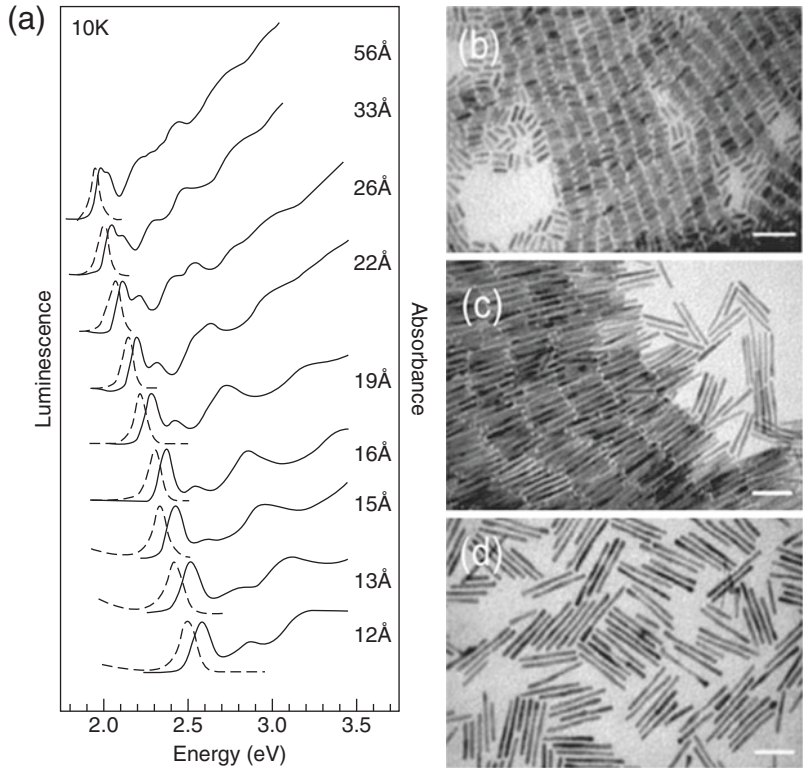
Light Generation: Semiconductor Nanocrystals for Lasing Applications

Colloidal semiconductor nanocrystals are a very attractive alternative as light emitting materials to more conventional nanostructures, because they can be fabricated in a cost-effective bottom-up approach by wet chemical synthesis. In the last decade, great progress has been achieved regarding the control on nanorcrystal shape and composition, leading to size dispersions of less than 5 % in spherical particle samples, and to rod- and multibranched shapes. One particular property of colloidal nanocrystals is their exposed surface that needs to be stabilized via surfactant molecules. The related obstacle that has to be overcome toward highly efficient light emission is the nonradiative recombination of the electrons and holes at surface states. The most promising solution to this problem are core-shell architectures, where a low band gap material nanocrystal is capped by a second material with higher band gap. In such structures, the photo-excited charges are confined by the nearly defect-free internal interface between the two materials (see band alignment illustration in Fig. 1).

Here, the emission wavelength is determined on one hand by the band gap of the core material, and on the other hand by the quantum confinement effects, since typical diameters of nanocrystals are in the range of 1–8 nm. The most popular systems for light emission in the visible consist of CdSe, CdS, ZnSe, and ZnS materials. For example, the emission of CdSe based systems can be tuned by the nanocrystal size in the range from 1.8 to 2.5 eV, as demonstrated in Fig. 2a.

Two very successful core-shell architectures are sketched in Fig. 1, the spherically symmetric core-shell structure, and the dot-in-a-rod (dot/rod) structure that can be fabricated by the seeded growth technique [1]. The dot/rod architecture in Fig. 1b combines the advantageous optical properties of rod-shaped nanocrystals, such as strongly linearly polarized emission and oriented assembly, with a high quantum yield (ratio photons in/photons out) comparable to that of spherical nanocrystals (Fig. 2).

For lasing devices two factors are important: the optical gain of the light emitting material and the quality of the resonant cavity that provides the optical feedback. The optical gain can be improved by the composition and architecture of the nanocrystals, and by the homogeneity and density of their assembly. The optical feedback can be obtained from semitransparent mirrors or distributed feedback resonators that consist of Bragg reflectors. The optical gain of the material is typically characterized by transient (time-resolved) optical absorption spectroscopy, which reveals the population dynamics of the energetic

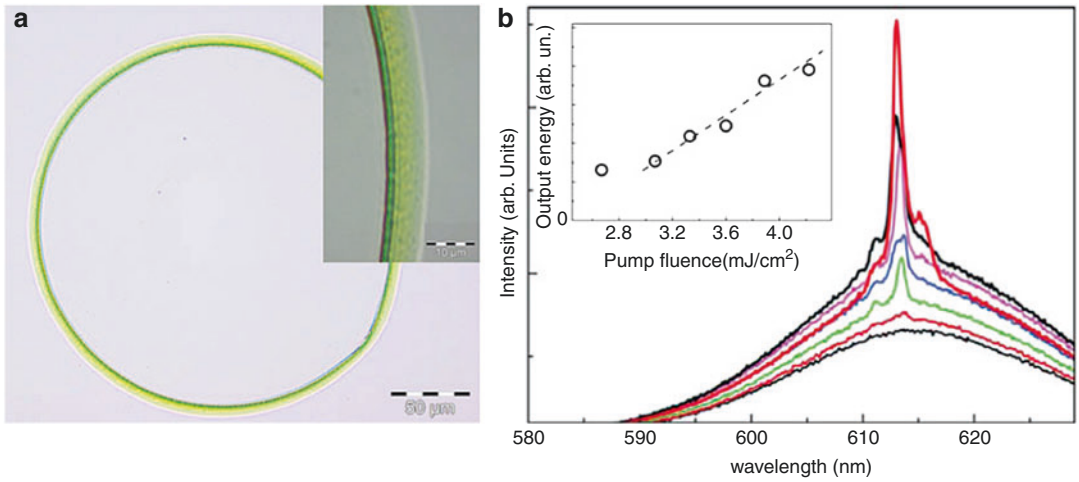


Nanostructures for Photonics, Fig. 2 (**a**) Emission and absorption spectra of spherical CdSe nanocrystals (With permission from Ref. [2]). (**b**) Transmission electron microscopy images of dot-in-a-rod core-shell nanorods with similar core size and different rod length. The scale bar corresponds to 50 nm (With permission from Ref. [1])

levels of the nanocrystals [3]. If the emitting ground state population maintains an occupation larger than unity, optical gain is achieved. One peculiar property of the dot/rod core-shell nanocrystals is that optical gain can be obtained from the optical transitions related to the core and from the transitions in the shell material. This effect extends the spectral range significantly, for example in the CdSe/CdS system from 570 to 650 nm (core emission) to 490 nm emission wavelength of the shell [4]. So far, the fabrication of efficient optical cavities for nanocrystal assemblies remain challenging since, for example, conventional sequential evaporation of Bragg reflectors does not work due to the incorporation of the inhomogeneous nanocrystal layer. Hence, novel approaches to obtain cavities have been sought, like the loading of cylindrical microcavities with nanocrystal solutions via the capillary effect [5], or the hot embossing of two pre-fabricated Bragg reflectors which embed the nanocrystal layer in a sandwich structure [6]. Another innovative approach consists in the laser emission from core-shell nanorod assemblies where the layer of nanorods itself formed the resonant cavity [3]. Here, circular rings with micron scale width and sharply defined lateral edges were formed via the coffee stain effect by simple deposition of nanocrystals dissolved in toluene (see Fig. 3a). The lateral edges of the rings themselves functioned as optical interfaces and were suitable for achieving feedback in a Fabry–Perot resonator scheme (Fig. 3b). Such microlasers could be very interesting for a new generation of photonic devices integrated onto substrates for lab-on-a-chip or for other compact, integrated applications.

Light Guiding: From Photonic Crystals to Metamaterials

Transportation of light to a detector is one of the fundamental aspects of any optical device. The light, during its path, might be affected by the surrounding world thus undergoing a



Nanostructures for Photonics, Fig. 3 (**a**) Optical microscope images of a self-assembled *circular* ring formed via coffee stain evaporation dynamics. The ring consists of densely packed core-shell nanorods and acts as a microcavity for laser feedback. (**b**) Emission spectra from a Fabry–Perot self-assembled microcavity made of nanorods with 25 nm length recorded at different pump fluences. Inset: emission intensity as a function of the pump fluence (With permission from Ref. [3])

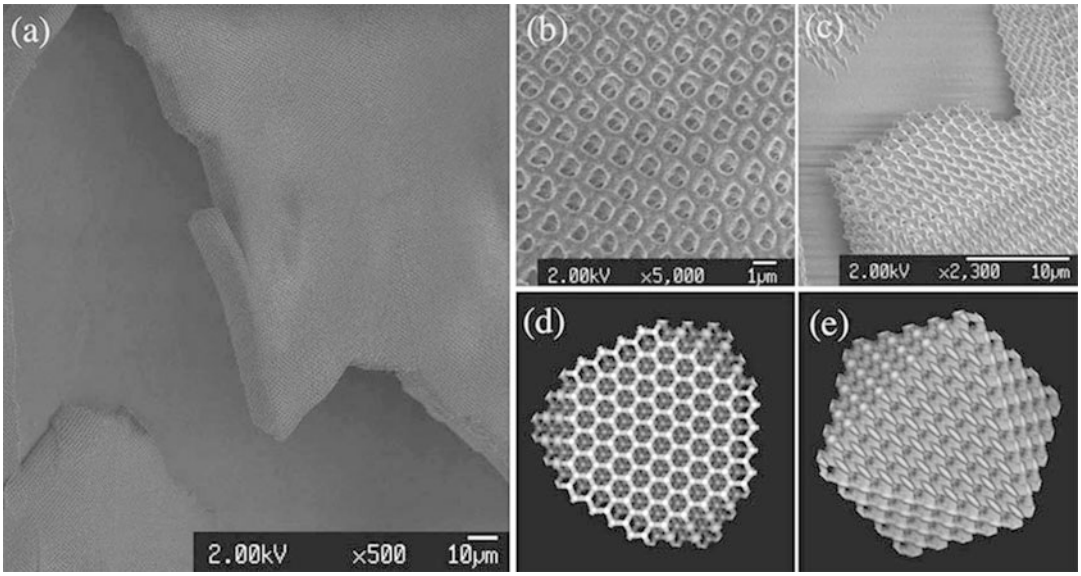
modification of its properties. This change can be temporary, namely existing only during the path toward the detector or even permanent. An example of the first kind is given by the effective wavelength of surface plasmon polaritons on a metallic interface whereas permanent modification can be well described by white light entering a fiber. In fact, in this case the light will loose most of its spectral components once the detector will be reached. A further example is given by metamaterials, usually defined as exotic devices capable of strongly modifying the standard properties of the light. These different situations indicate a strong relation between the intended application and the method of transportation of light which must be properly chosen.

Periodic and Quasiperiodic Fibers

A standard and well-established method for the transportation of light is by means of optical fibers. These devices are nowadays commonly used for applications such as Internet, allowing a high-speed data transfer (Fig. 4).

A different kind of fiber known as Photonic Crystal Fiber (PCF) was introduced in [8] with the idea of improving the classical fibers in many aspects, such as single-mode operation regime in a wide range of wavelengths, air-core guidance or reduced cross-sectional dimensions. In fact, while standard fibers have diameters of some hundreds of micrometers, PCFs can be reduced down to few microns [9–11].

The physics behind the working principle of PCFs fibers lies in their geometrical characteristics. Photonic crystals (PCs) are structures showing a *translational* periodicity of the medium permittivity [12, 13]. A three dimensional example of a face centered cubic PC is given in Fig. 4. The main characteristic of a PC is the capability of controlling the interaction between radiation and matter in order to obtain new physics. For example, they can induce a local modification of the field, which in turn will affect the optical characteristics of samples embedded in the crystal. For the present purposes, these devices will be seen either as filters or waveguides. In fact, according to their geometrical and material properties, PCs can realize optical gaps in specific spectral regions. Consequences are the impossibility of the light to propagate inside the crystal. Figure 5a is an example of a square-like photonic crystal waveguide where light at 2.85 μm is obliged in following the free path of the crystal. The structure is formed by Silicon cylinders of 200 nm

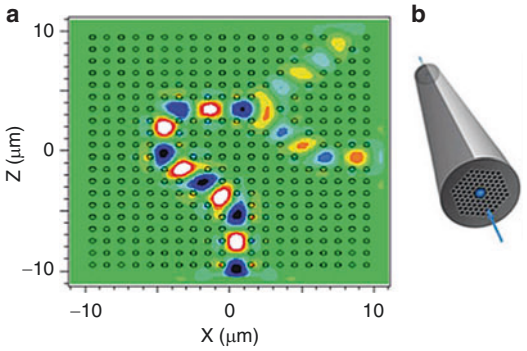


Nanostructures for Photonics, Fig. 4 The scanning electron microscope image of the scaffold type a fcc optical lattice. The lattice constant is about 1.5 μm. (**a**) A bird's eye view. (**b**, **c**) Magnified images of (**b**) *top* surface of the structure, and (**c**) cross section of a crack in the rim of the structure. (**d**, **e**) Simulated cross section of fcc lattice in (**a**) $< 1,1,1 >$ and (**b**) $< 1,1,-1 >$ (With permission from Ref. [7])

radius in air. The periodicity (lattice constant) of the crystal is 1 μm. When the optical gap concept is associated to the fiber world, Photonic Crystal Fibers are obtained. In this case, light propagates along the axis of the cylinders as in Fig. 5b.

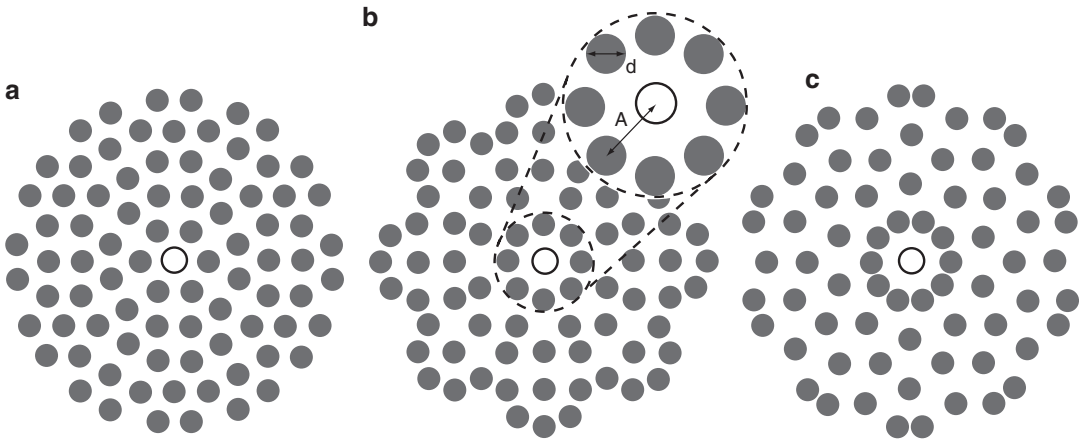
Among the PCFs special attention deserves the Photonic Quasicrystal Fibers (PQFs). Their geometry is similar to photonic crystals with the fundamental characteristic of loosing the translational periodicity. Only rotational and mirror symmetry are proper of a quasicrystal [14, 15]. Figure 6 gives some examples of quasicrystal. One of the advantages of considering PQFs versus PCFs is in a wider single-mode operation regime than the PCFs, maintaining similar cross-sectional dimensions.

The typical approach for determining the optical characteristic of a fiber is the calculation of its V parameter. This value is usually considered as the quantity capable of identifying the single-mode operation regime of a fiber. In fact, in the case of single-mode fiber, for a given wavelength and polarization, only a particular mode can propagate. This avoids the problem of the destructive interference that occurs in a multimode fiber,

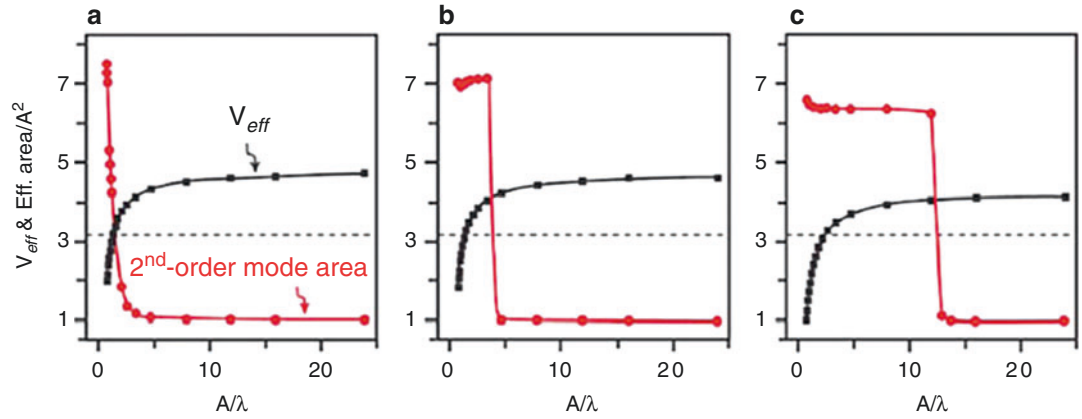


Nanostructures for Photonics, Fig. 5 (**a**) Two-dimensional photonic crystal waveguide. Transverse magnetic polarized light at 2.85 μm coming from the *bottom* of the figure travels along the free path. The period (distance between two Silicon cylinders) is 1 μm, the radius is 200 nm. (**b**) Photonic crystal fiber. The *blue arrow* indicates the direction of propagation of the light

which makes single-mode fibers the best choice for applications related to long distance communications. This explains why fibers working as single mode in a wide range of frequencies are very desirable for a number of applications. However, in case of PQFs the V parameter method has been demonstrated not to be working in many



Nanostructures for Photonics, Fig. 6 Cross section of three kinds of PQFs: (**a**) sixfold, (**b**) eightfold, and (**c**) tenfold. *Gray filled circles* denote air holes drilled in silica fiber. The central hole (empty *circle*) is absent in the fibers. The quantities *A* and *d* represent the lattice constant and the hole diameter, respectively (With permission from Ref. [15])



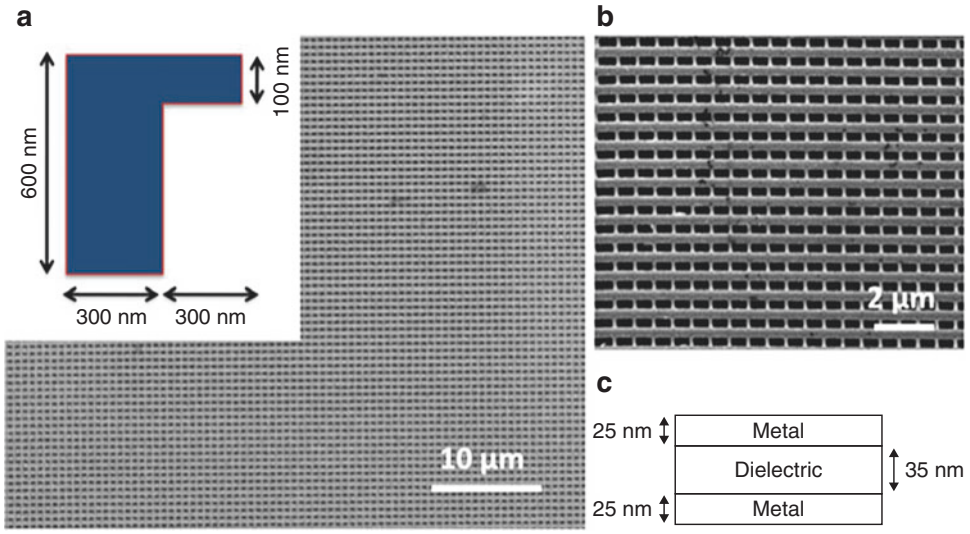
Nanostructures for Photonics, Fig. 7 Effective V parameter (*black squares and curves*) and the second-order mode normalized area (*red circles and curves*) as a function of A/λ for (**a**) sixfold, (**b**) eightfold, and (**c**) tenfold quasicrystal fibers. The dashed lines are drawn at π. A is the lattice constant (With permission from Ref. [15])

cases [15]. For this reason, a different approach consisting in the direct calculation of the effective area of the second-order mode is preferable. Figure 7 shows a comparison between these two approaches for three kinds of fibers.

Besides a disagreement between the V parameter and the effective area approach as in Fig. 7b, c, it is possible to notice that the single-mode operation regime for both the eightfold and tenfold fibers is much more extended than the sixfold one. This makes these kinds of fiber more robust for transporting signal at long distances.

Fishnet Structures: A First Step Toward Metamaterials

Metamaterials are artificial man-made materials exhibiting particular properties such as negative refraction [16–19], which can lead to several applications like the realization of superlenses or

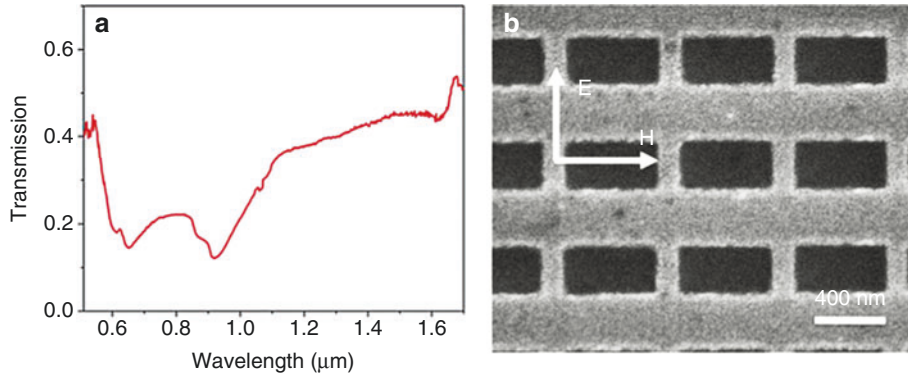


Nanostructures for Photonics, Fig. 8 (**a**, **b**) SEM images of a fabricated fishnet structure. (**c**) Sketch of the *vertical* structure. The *inset* in (**a**) shows the cell unit of the fishnet structure

invisibility cloak. In order to reach a negative refractive index, simultaneously negative electric permittivity and magnetic permeability need to be reached as pointed out by Veselago in 1968 [16]. A negative permittivity is a relatively easy task to obtain, since all metals below their plasma frequency exhibit it. On the other hand, a negative permeability is more difficult to achieve; the lack of magnetic charges and the weak interaction between matter and the magnetic part of light, make scientists capable to attain it only in certain frequency regions. Combining thin metallic wires and rods in a three layers Metal/Dielectric/Metal structure (as shown in Fig. 8), constitutes the so called *fishnet structure* which has been proved to exhibit peculiar characteristics both in metamaterials and plasmonics fields.

Fishnet structures have been widely investigated in metamaterials research field both from an experimental and a theoretical point of view [20]. The reason is that their geometry is simple and its fabrication is relatively easy. However, the mechanism behind the negative refractive index behavior is nowadays not completely understood. The main idea is that properly polarized light passing through the structure can experience at the same time both negative permittivity and permeability. Considering, for instance, an impinging electromagnetic field polarized as shown in Fig. 9b: the electric field oscillates along thin wires allowing a negative electric permittivity and the magnetic field can induce a resonance through the aligned sets of rods, just like in a LC circuit. In fact, whereas the capacitor-like response is directly given by the vertical structure of the fishnet (two metallic plates separated by a dielectric, see Fig. 8c), the inductive response is expressed by the electric currents, generated by the oscillating magnetic field, that circularly flow in the plates (as regular currents) and across the dielectric (as displacement current). The resonant behavior can be observed, for example, in the transmission spectrum (Fig. 9a): if the magnetic resonance occurs, a dip in the transmission coefficient can be noticed.

Before performing the experiments, also according to previous studies where similar structure were investigated, a magnetic resonance between 1,300 and 1,500 nm was expected [20]. In Fig. 9b, this kind of resonance is not clearly visible, probably because of high losses in that spectral range which damp transmission or reflection features. Nevertheless, it is possible to notice a resonant behavior at shorter wavelengths.



Nanostructures for Photonics, Fig. 9 (**a**) Transmissivity spectrum for the presented fishnet structure. (**b**) SEM micrograph of measured fishnet structure and fields direction of the incident light

Indeed current loops, which are supposed to induce the magnetic resonances, are generated by internal gap surface plasmons polaritons (SPPs). This means that, more than one resonance exist, actually several of them may occur as far as the matching condition subsists:

$$k_{SPP} = k_{wave} + mG_x + nG_y \text{ with } : G_x = \frac{2\pi}{a_x} \text{ and } G_y = \frac{2\pi}{a_y}$$

In this relation, k_{SPP} and k_{wave} are, respectively, the wave vector of the excited SPP and the one of the incident light. a_x, a_y are the lattice periodicities along the x and y fishnets axis. Therefore, higher frequency resonances can be reached as far as the lattice provides the missing momentum through G_x and/or G_y. However, a negative refractive index is not guaranteed in the entire spectrum since the equivalent plasma frequency of the medium is dominated by the cutoff wavelength of the holes of the fishnet.

To conclude, fishnet structures represent a good example on the concept of metamaterials and how their fascinating properties are related to plasmonic effects. A deep study of these structures can be considered a first step to understand the underlying physics of negative refraction enabling, eventually, the realization of more complex devices exploiting the effects of negative index materials.

Light Harvesting: Supported Metal Nanoarrays for Plasmonic Devices

Light harvesting plays an important role for the operation of many kinds of devices, such as biosensors or photovoltaic cells. In fact, besides the generation and transportation of light, an efficient coupling mechanism between the guiding device and the investigated bio-sample in order to produce a readable signal is a fundamental step of a generic photonic device. For example, metallic nanostructures have the property of realizing a high localized electric field which can be efficiently coupled to a bio-sample. This, in turn, means high signal-to-noise ratio from the sample, namely a good signal for the detector can be produced.

Metallic nanoparticles exhibit specific optical properties in the visible/near-infrared range due to the collective excitation of the conduction electrons (Localized Surface Plasmon Resonances – LSPRs). The field enhancement provided by metallic nanostructures has been used to manipulate light-matter interactions and enhance nonlinear phenomena. The strong effects associated to the LSPR excitation on metal nanoparticles are the core of many interesting perspective applications such as single- molecule detection and photovoltaic conversion. Here two specific examples are introduced, that is, nanoaggragate and nanoantenna arrays where light harvesting and field enhancement will be

exploited with respect to high sensitivity Raman spectroscopy.

The use of advanced lithographic techniques allows precise control of particle shape and size, as well as interparticle separation [21, 22]. Electron beam lithography (EBL) has become the method of choice for fabricating single or periodic arrays of metal nanostructures endowed with the desired morphological features [23]. It is versatile and provides high resolution and precise control over the geometry and separation of nanostructures, guaranteeing a fabrication reproducibility, and precision down to the nanometer scale.

Nanoaggregate Array

The fabrication of an active SERS device was achieved by means of site selective deposition of Ag nanoparticles, exploiting electroless technique on Si pre-patterned substrates [23]. A large array of nanostructures with high degree of reproducibility at the nanometer scale was realized using electron beam lithography onto a Si substrate spin-coated with 50 nm thick ZEP layer.

The Ag deposition was realized by means of a redox reaction between the metal and the Si surface remained uncovered after the lithographic process. The reaction occurs in a dilute (0.1–0.5 M) HF solution in which is dissolved a metal salt, that is, $AgNO_3$, at a concentration of about 1 mM. The global chemical reaction is:

$$4Ag^{+} + Si(s) + 6HF \rightarrow 4Ag^{0} + H_2SiF_6 + 4H^{+} \quad (1)$$

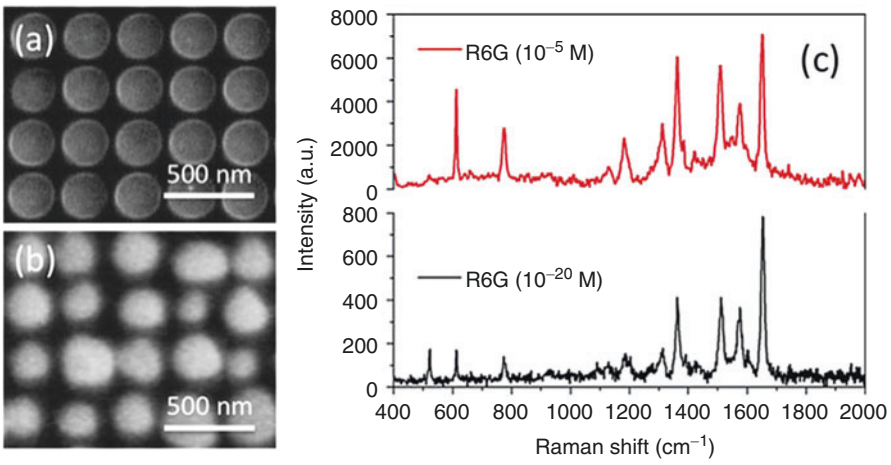
which in its turn can be separated into two half-cell reactions:

$$Si + 2H_2O \rightarrow SiO_2 + 4H^{+} + 4e^{-}(\text{Anode} - \text{Si oxidation}) \quad (2)$$

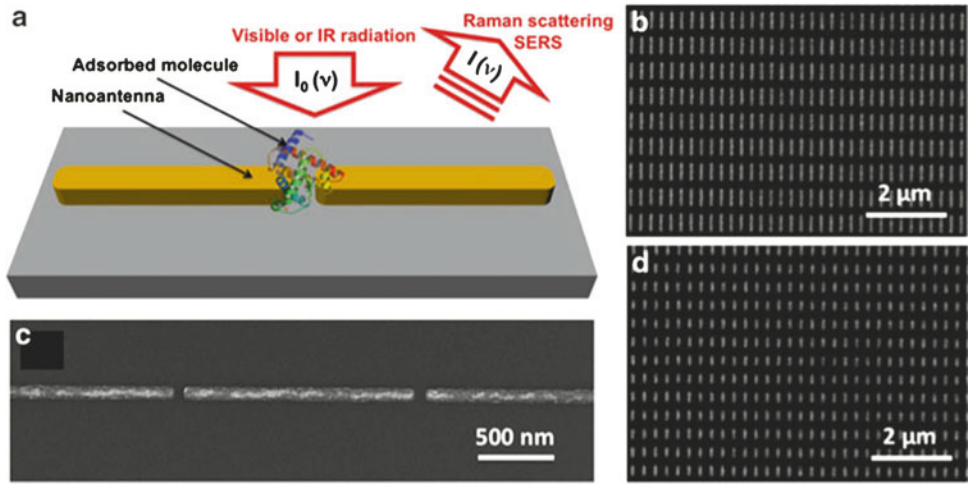
$$Ag^{+} + e^{-} \rightarrow Ag^{0}(\text{Catode} - \text{Ag reduction}) \quad (3)$$

The driving force of the whole process is coming from the potential difference in the two half-reactions, thus fostering the growth of Ag grains and the formation of well-organized metallic nanostructures (see Fig. 10a, b).

In order to probe the enhancement behavior in Raman spectroscopy related to the plasmon excitation, the sample was immersed in Rhodamine 6G (R6G) solutions with different concentration (10^{-5} M and 10^{-20} M in H_2O) for 30 min. The sample was gently rinsed in DI-water to remove molecules in excess not chemisorbed on the metal surface and then dried in N_2 flow.



Nanostructures for Photonics, Fig. 10 SEM micrographs of (**a**) array of empty holes fabricated by EBL technique and (**b**) Ag nanoaggregates selectively grown inside the nanohole pattern. (**c**) SERS spectra acquired from 10^{-5} to 10^{-20} M R6G, absorbed on the silver structures



Nanostructures for Photonics, Fig. 11 (**a**) The nanoantenna concept: fabrication of a highly sensitive and specific nanobiosensor based on extraordinary optical signal enhancement. (**b**, **d**) Representative SEM images of two different nanoantenna arrays fabricated by EBL technique. The dimensions of the NAs are of length 410 nm (**b**), 200 nm (**d**) while width and height both are 60 nm for all NA. (**c**) SEM of coupled nanoantenna array

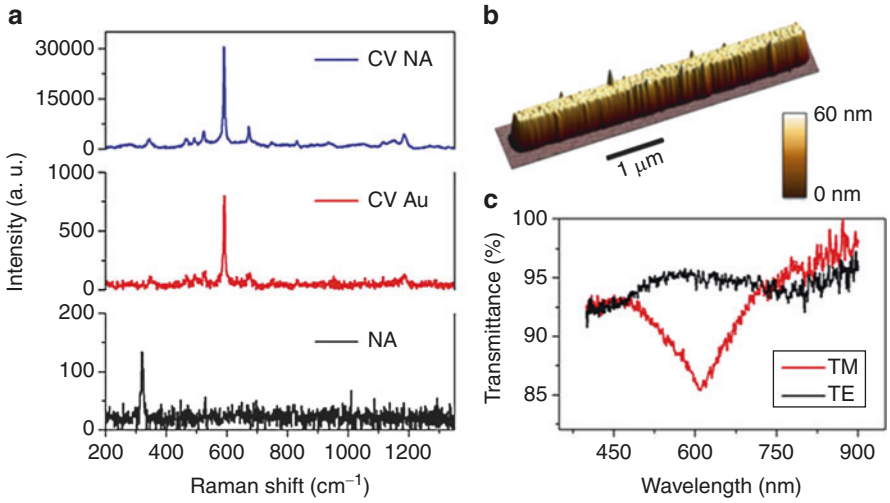
In Fig. 10c, Raman spectra of R6G molecules, adsorbed on a Si substrate with a 60 s Ag deposition, are shown. There are various bands for R6G compound at around 1,650, 1,509, and 1,361 cm^{-1}, attributed to the xanthenes ring stretching of C–C vibrations, 1,575 cm^{-1} attributed to C–O stretching; while the band at 1,183 cm^{-1} is related to the C–H bending and N–H bending vibration of xanthenes ring. As can be clearly observed in Fig. 10c, the Raman intensity decreases with the decreasing concentration of R6G, as expected. Though, the concentration is very low, various bands are clearly visible even for R6G molecule with concentration of 10^{-20} M, keeping accumulation time 60 s and laser power 0.0018 mW.

Nanoantenna Array

The ability to tailor the optical properties of structured plasmonic materials has proven to be crucial for a wide variety of applications, using them for surface-enhanced spectroscopy and biosensing. While a symmetrical geometry allows polarization independent plasmon excitation, a dichroic absorption is well expected for nanoantenna (NA) fabrication. Large scan overview of different nanoantenna patterns are reported in Fig. 11. The sample topography has been characterized recurring to scanning electron microscopy (SEM) and atomic force microscopy (AFM) operated in tapping mode.

The optical properties of the nanoantenna array, shown in Fig. 12c, were investigated by means of spectroscopic transmission of polarized light in the range between 400 and 900 nm. The polarization of the incident light was varied from transverse magnetic TM (electric field perpendicular to the short NA axis) to transverse electric TE (electric field parallel to the long NA axis). The optical transmittance spectra present evidence of a clear anisotropic behavior; for TM polarization a localized minimum around 620 nm is found. This is the typical behavior exhibited by subwavelength metal nanoparticles sustaining a localized surface plasmon resonance.

Evidence of plasmon-induced SERS effects is given in Fig. 12a in which the different Raman emission lines from cresyl violet (CV) molecules (with a concentration of 3.46 μM) deposited on the gold nanoantenna arrays are shown. The sample was excited by 633 nm laser wavelength (laser power = 0.14 mW and accumulation time = 50 s) through a 150X objective. The measurements were performed on bare nanoantenna arrays



Nanostructures for Photonics, Fig. 12 (**a**) SERS spectra of cresyl *violet* molecules deposited on NA sample (*blue line*) and flat Au film (*red line*). Background measurement performed on bare NA structure is also reported (*black line*). (**b**) AFM 3D image. (**c**) Transmission optical spectra of the NA array of Fig. 1.11b. *Red/black lines* are for TM/TE polarization respectively

(sample NA), on cresyl violet molecules chemisorbed on flat Au film (sample CV Au) and on CV deposited on the nanoantenna arrays (sample CV NA). To ensure the quality of the sample surface, the background measurements were performed at different positions of the nanoantenna SERS device. As shown in Fig. 12a (NA line) no characteristic Raman band, except at around 320 cm^{-1} related to the Ca-F vibration from the substrate, is observed. The characteristic vibrational bands of CV are observed in the SERS spectrum (CV NA trace). Intense Raman bands centered at around 591, 882, 927, and 1,189 cm^{-1} can be attributed to the N–H_2 rocking vibration, two benzene group bending, out-of phase N–H_2 rocking vibration, and combination of N–H_2 rocking and C–H_X rocking, respectively.

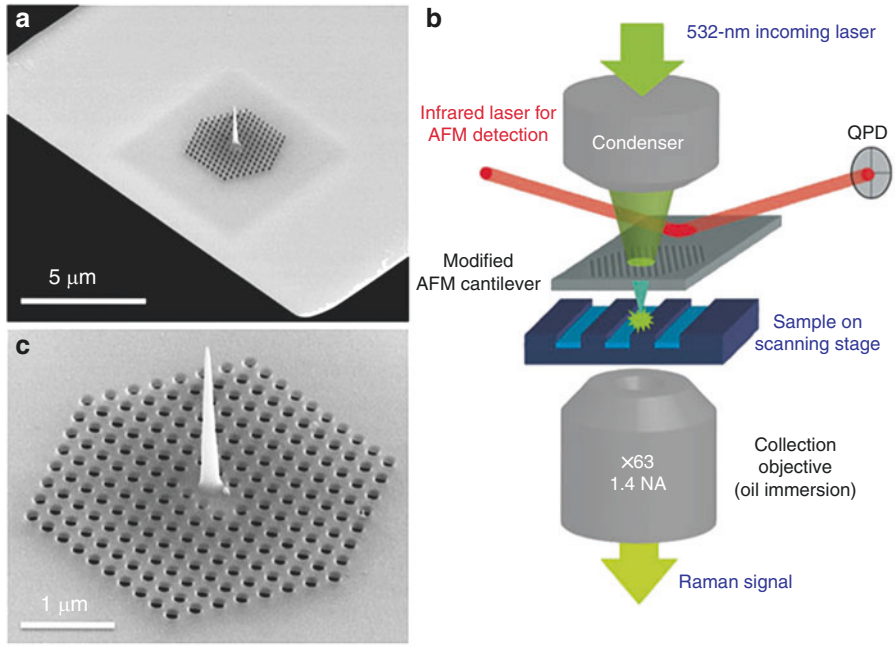
CV molecules chemisorbed over flat Au film are also shown in Fig. 12a. A giant enhancement in Raman signal for CV NA sample with respect to the CV Au one can be clearly observed. The maximum enhancement of the Raman emission is obtained when the excitation is polarized along to the nanoantenna short axis, in correspondence to the resonant excitation of the LSP resonance. The evaluated SERS enhancement for the fabricated nanoantenna device is 10^7.

Novel Nanophotonic Devices

The three standard aspects of a photonic device, namely light generation, guiding, and harvesting can be combined in a number of ways to obtain many kinds of functionalities. For example, metallic-nanorods can be used for shrinking the wavelength of the light to improve the coupling with detectors such as nanoantenna arrays, or optical fiber can be properly modified for three-dimensional optical trapping [24]. Here will be introduced two specific examples: photonic crystals combined to plasmonic devices for the realization of background-free Raman probes, and superhydrophobic devices for single-molecule detection.

Integration of Photonic Crystals and Plasmonic Devices for Label-Free Chemical and Structural Detection

A combination of different nanostructures for photonics give rise to a novel photonic-plasmonic device that allows label-free detection, through Raman spectroscopy, of single inorganic nanoparticles and of monolayers of organic compounds [25] and that allows, integrated on an atomic force microscope (AFM) cantilever, to

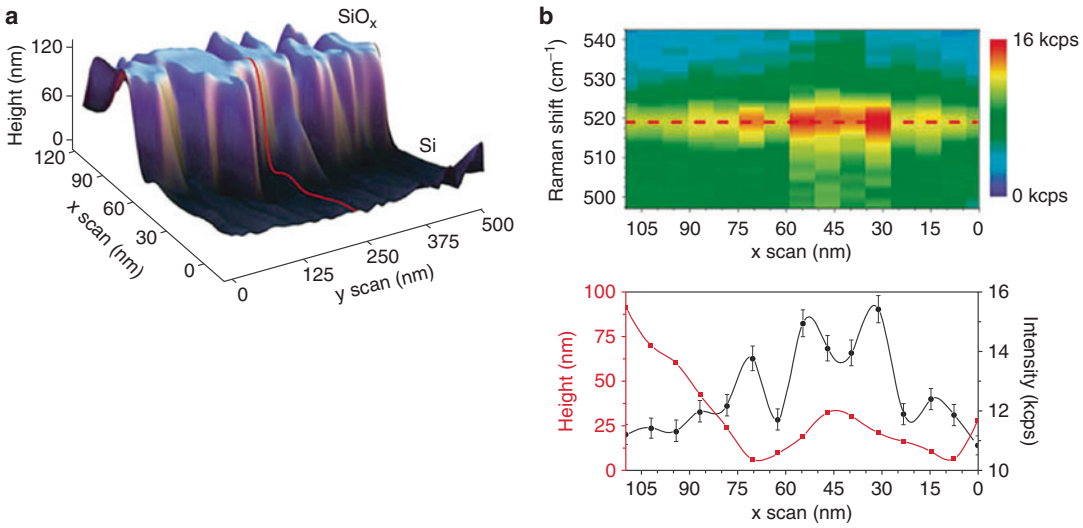


Nanostructures for Photonics, Fig. 13 (**a**) SEM image of the device fabricated on a silicon nitride AFM cantilever. (**c**) Zoom on the photonic crystal cavity and the tapered plasmonic waveguide with radius of curvature at the apex of 5 nm. (**b**) Sketch of the experimental setup, showing the integration of the photonic–plasmonic device into an AFM–Raman microscope

obtain topographic, chemical, and structural information at an unprecedently high spatial resolution (below 10 nm) [26].

The device consists of a two-dimensional dielectric photonic crystal cavity, which can be patterned on a Si_3N_4 AFM cantilever, together with a tapered silver plasmonic waveguide placed at the center of the cavity (Fig. 13a). The photonic crystal consisted of a triangular lattice of air holes (lattice constant = 250 nm, hole diameter = 160 nm) patterned on a 100-nm-thick Si_3N_4 membrane. Three missing holes in the center generated a photonic crystal cavity, termed L3, tuned at $\lambda = 532$ nm. The plasmonic waveguide is a 2.5-μm-tall metallic cone, with a base diameter of 300 nm, and the minimum radius of curvature at the apex is controllable in the range 2.5–5 nm (Fig. 1c). The entire device is fabricated by means of Focused Ion Beam milling and electron-beam induced deposition [27]. Ion beam milling was used to define the photonic crystal whereas the silver tapered waveguide was grown in the center of the cavity using electron-beam induced deposition from a gas precursor containing a platinum–carbon polymer $(CH_3)_3Pt(C_pCH_3)$.

The photonic crystal cavity enables an efficient coupling between the external laser source (used for Raman excitation) and the tapered waveguide, together with optimal sample illumination. In fact, the focal plane from which the SPPs are launched is 2.5 μm away from the cone apex where the Raman excitation is generated. This drastically lowers the background intensity at the sample plane. Thus, the spatial separation between SPP generation and Raman excitation efficiently increases the signal-to-noise ratio in the Raman spectrum, giving rise to a new optical configuration, namely Surface Plasmon Polariton Enhanced Spectroscopy (SPPERS). The focusing of SPPs follows an adiabatic compression mechanism, which causes a strong localization of the electrical field in a region with a size comparable to the apex radius of curvature [28]. This device can be fully integrated with an inverted Raman microscope combined with an AFM stage, as shown in Fig. 13b.



Nanostructures for Photonics, Fig. 14 (**a**) SEM image of the device fabricated on a silicon nitride AFM cantilever. (**c**) Zoom on the photonic crystal cavity and the tapered plasmonic waveguide with radius of curvature at the apex of 5 nm. (**b**) Sketch of the experimental setup, showing the integration of the photonic–plasmonic device into an AFM–Raman microscope (With permission from Ref. [26])

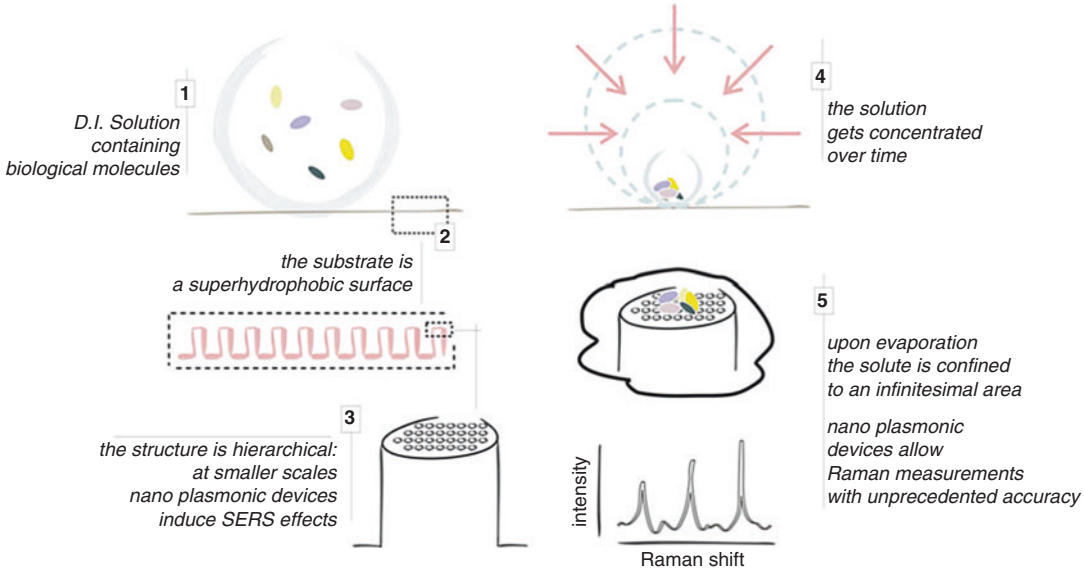
The SPP spatial confinement in the near field is comparable to the radius of curvature of the tapered waveguide apex, which lies between 2.5 and 5 nm. The expected electric field enhancement E due to adiabatic compression is defined as $E = E_{tip}/E_{base}$, where E_{tip} and E_{base} are the maximum electrical fields at the apex and bottom of the waveguide, respectively, and reaches values of practical interest in the range of 100 when the radius of curvature of the tip is below 5 nm. For successful spectroscopy applications, this necessitates state-of-the-art fabrication quality.

The architecture of the device and the mechanical stiffness of the tapered waveguide allow using it both as an AFM tip and as a nanometer light source for near-field Raman excitation. In other words, it is possible to obtain topographic and chemical mapping to yield structural and stress information with the same spatial resolution as is imposed by the geometry of the tip, that is, limited only by its radius of curvature.

Device performances have been assessed by measuring a test sample for which the chemical composition had periodic spatial variations in the range of few nanometers. The test sample consisted of patches and gratings of silicon nanocrystals having different pitches (from several micrometers to a submicrometer scale) obtained by laser melting of a silica substrate previously fabricated by plasma-enhanced chemical vapor deposition.

The AFM–Raman setup was operated in tapping mode and wet conditions (in distilled water) to measure simultaneously the topography and Raman intensity map by scanning the sample and acquiring Raman data across the lithographic structures, point by point.

A severe test on the described sample was performed by scanning the sample in a patterned region in which the crystalline and amorphous topography forms a grating with submicrometer pitch. From the AFM measurements, it could be seen that silicon, upon laser-induced crystallization, has its height reduced by 100 nm, which results in the formation of a trench on the sample surface, as shown in the three-dimensional AFM map in Fig. 14a. The AFM scan step size was 7 nm. The red line in Fig. 14a is a representative AFM line scan which provides simultaneously the Raman intensity map shown in Fig. 2b (upper panel), in which the Raman spectrum extends from 504 to 540 cm^{-1}. The acquisition time of each Raman map was 5 s. Figure 14b (lower panel) shows the topography and Raman intensity



Nanostructures for Photonics, Fig. 15 Biophotonic nanostructures integrated with superhydrophobic surfaces can lead to the fabrication of novel devices with advanced sensing capabilities

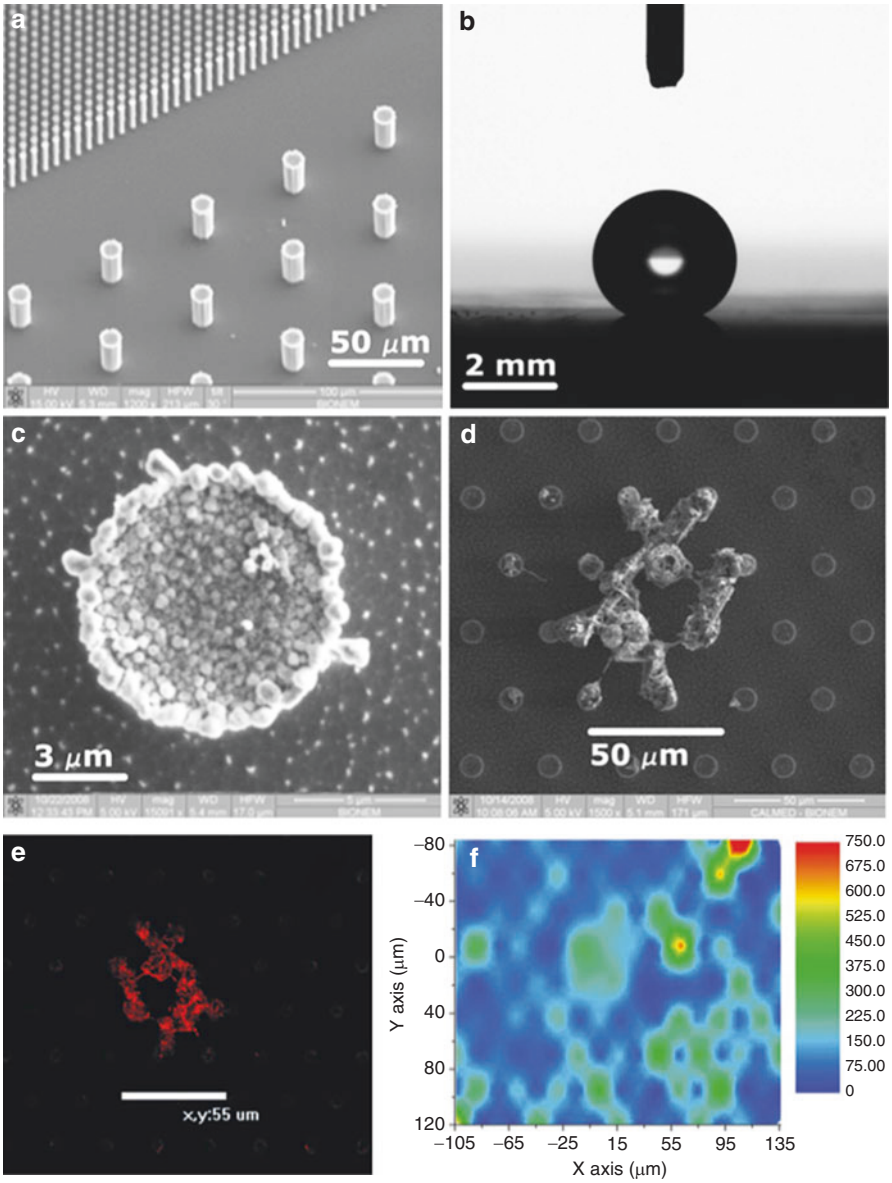
for the same line scan. The Raman intensity is taken at the maximum of the silicon phonon peak at 520 cm^{-1}. The Raman intensity and relative peak amplitudes change from point to point, indicating an optical resolution of better than 7 nm.

In conclusion, the development of a device based on the combination of photonic–plasmonic nanostructures that is fully compatible with AFM and Raman spectroscopy opens up a number of significant opportunities in the nanoscale chemical mapping of materials. The approach could be of particular benefit in the nanoscale analysis of biological matter, such as in the investigation of cell membrane protein spatial distributions, in which chemical/physical bond strengths and amino acid composition in vivo conditions could be revealed. Device architectures based on the adiabatic generation of SPPs have considerable potential for applications in which a reliable nanometer-sized light source is of utmost importance.

Integration of Biophotonic and Superhydrophobic Devices for the Detection of Low-Concentrated Biomolecules

Here, will be investigated how biophotonic nanostructures can be integrated to superhydrophobic surfaces (SHSs) to obtain devices with advanced sensing capabilities, thus offering new possibilities in the field of clinical medicine, especially toward the early diagnosis.

Superhydrophobicity is a phenomenon whereby a drop post upon a surface would preserve its original spherical shape rather than spreading or wetting indefinitely the plane of contact (Fig. 15). In the celebrated model of Cassie [29], a surface would be superhydrophobic on account of the pockets of air that remain trapped between the liquid and the substrate, and the smaller the fraction of solid in contact with the drop the larger the apparent contact angle. SHSs retain unique properties in terms of wettability that can be reviewed as follows: (a) SHSs have superior adhesive properties, in the sense that they exhibit vanishing friction coefficients; (b) a droplet, deposited upon these surfaces, would accordingly preserve a quasi-spherical shape while evaporates, and the contact area at the interface would thus progressively reduce; (c) SHSs can be artificially reproduced using micro and nanofabrication techniques. Using the properties above, micro textured surfaces may be successfully exploited to concentrate tiny amounts of moieties over micrometric areas, and



Nanostructures for Photonics, Fig. 16 (**a**) SEM micrograph of periodic hexagonal lattice of cylindrical Si micro pillars used as SHS. (**b**) The *spherical* shape of a drop positioned upon the SHS is preserved. (**c**) Arrays of Ag nanograin aggregates positioned upon the pillars. (**d**) SEM image of the residual solute of R6G at the end of process of evaporation. The intensity of fluorescence is shown in (**e**), while Micro-Raman mapping measurements are reported in (**f**)

consequently measure these moieties with unprecedented accuracy [30, 31].

SHSs typically comprise a periodic hexagonal lattice of cylindrical Si micro pillars with a certain diameter and pitch (Fig. 16a). The micro pillars are realized combining optical lithography, electroless growth, and Bosch Reactive Ion Etching (RIE) techniques. Nonconventional biophotonic nanostructures such as arrays of silver nanograin aggregates, conveniently positioned upon the pillars, would complete a hierarchical structure thus permitting the

identification of substances in the single-molecule regime. These "two-stages" micro and nanostructures act as superhydrophobic surfaces with an increased contact angle ranging from 155° to about 175°.

A particular kind of nanogeometry based plasmonic device is here reported as example, which is an electroless grown random assembly of silver nanograins [23]. Nevertheless, the method can be extended to a number of different plasmonic nanostructures, as those described in the sections above.

Imagine to deposit a drop of an extremely diluted solution upon a textured, superhydrophobic substrate. The drop would evaporate over time and thus the solution would get more and more concentrated (Fig. 15). At the end of the process, the residual solute would be confined within an exceptionally small region of the plane. With a suitable design, few molecules may be conveniently enforced to confine into the smallest area conceivable, at the limit upon a sole pillar. Nanogeometry based biophotonic devices, conveniently tiling these surfaces, would probe/detect the moieties with extremely small resolution limits. Here, the beneficial effects of superhydrophobicity and nanogeometry based spectroscopy are combined and conveyed into a unique platform, and from the combination of the two novel properties arise permitting the identification of proteins or analytes with unprecedented accuracy.

In order to clarify the concept, an experiment performed using R6G molecules, is reported. Evaporation processes of drops of D.I. water were followed over time. Few molecules were conveniently enforced to confine into a small area. A solution was investigated with initial concentration as small as 10^{-18} M. Figure 16d shows a SEM image of the residual solute of R6G at the end of process of evaporation. The intensity of fluorescence as in Fig. 16e is correctly proportional to the quantity of substance deposited upon the pillars, and thus the intensity signal of fluorescence follows the solute distribution upon the substrate. Micro-Raman mapping measurements were also performed (Fig. 16f). The mapping analysis was performed by referring the band centered at 1,650 cm^{-1}. The results demonstrate that the method is effective in measuring extremely small diluted solutions.

Cross-References

- Biosensors
- Electron Beam Lithography (EBL)
- Light Localization for Nano-optical Devices
- Nanoparticles
- Nanophotonic Structures for Biosensing
- Optical Properties of Metal Nanoparticles
- Surface Plasmon-Polariton-Based Detectors

References

1. Carbone, L., Nobile, C., De Giorgi, M., Sala, F.D., Morello, G., Pompa, P., Hytch, M., Snoeck, E., Fiore, A., Franchini, I.R., Nadasan, M., Silvestre, A.F., Chiodo, L., Kudera, S., Cingolani, R., Krahne, R., Manna, L.: Synthesis and micrometer-scale assembly of colloidal Cdse/Cds nanorods prepared by a seeded growth approach. Nano Lett. **7**, 2942 (2007)
2. Efros, A.L., Rosen, M., Kuno, M., Nirmal, M., Norris, D.J., Bawendi, M.: Band-edge exciton in quantum dots of semiconductors with a degenerate valence band: dark and bright exciton states. Phys. Rev. B-Condens. Mat. **54**, 4843 (1996)
3. Zavelani-Rossi, M., Lupo, M.G., Krahne, R., Manna, L., Lanzani, G.: Lasing in self-assembled microcavities of Cdse/Cds Core/Shell colloidal quantum rods. Nanoscale **2**, 931 (2010)
4. Krahne, R., Zavelani-Rossi, M., Lupo, M.G., Manna, L., Lanzani, G.: Amplified spontaneous emission from core and shell transitions in Cdse/Cds nanorods fabricated by seeded growth. Appl. Phys. Lett. **98**, 063105 (2011)
5. Kazes, M., Lewis, D.Y., Ebenstein, Y., Mokari, T., Banin, U.: Lasing from semiconductor quantum rods in a cylindrical microcavity. Adv. Mater. **14**, 317 (2002)
6. Martiradonna, L., Carbone, L., De Giorgi, M., Manna, L., Gigli, G., Cingolani, R., De Vittorio, M.: High Q-factor colloidal nanocrystal-based vertical microcavity by hot embossing technology. Appl. Phys. Lett. **88**, 181108 (2006)

7. Shoji, S., Proietti Zaccaria, R., Sun, H., Kawata, S.: Multi-step multi-beam laser interference patterning of three-dimensional photonic lattices. Opt. Express **14**, 2039 (2006)
8. Knight, J.C., Birks, T.A., Russell, P.S.J., Atkin, D.M.: All-silica single-mode optical fiber with photonic crystal cladding. Opt. Lett. **21**, 1547 (1996)
9. Galli, M., Agio, M., Andreani, L.C., Atzeni, L., Bajoni, D., Guizzetti, G., Businaro, L., Di Fabrizio, E., Romanato, F., Passaseo, A.: Optical properties and photonic bands of GaAs photonic crystal waveguides with tilted square lattice. Eur. Phys. J. B **27**, 79 (2002)
10. Malvezzi, A.M., Vecchi, G., Patrini, M., Guizzetti, G., Andreani, L.C., Romanato, F., Businaro, L., Di Fabrizio, E., Passaseo, A., De Vittorio, M.: Resonant second-harmonic generation in a GaAs photonic crystal waveguide. Phys. Rev. B **68**, 1613061 (2003)
11. De Vittorio, M., Todaro, M.T., Stomeo, T., Cingolani, R., Cojoc, D., Di Fabrizio, E.: Two-dimensional photonic crystal waveguide obtained by e-beam direct writing of SU8-2000 photoresist. Microelectron. Eng. **73–74**, 388 (2004)
12. Proietti Zaccaria, R., Verma, P., Kawaguchi, S., Shoji, S., Kawata, S.: Manipulating full photonic band gaps in two dimensional birefringent photonic crystals. Opt. Express **16**, 14812 (2008)
13. Zhao, H., Proietti Zaccaria, R., Verma, P., Song, J., Sun, H.: Single-mode operation regime for 12-fold index-guiding quasicrystal optical fibers. Appl. Phys. B **100**, 499 (2010)
14. Zhao, H., Proietti Zaccaria, R., Song, J., Kawata, S., Sun, H.: Photonic quasicrystals exhibit zero-transmission regions due to translational arrangement of constituent parts. Phys. Rev. B **79**, 115118 (2009)
15. Zhao, H., Proietti Zaccaria, R., Verma, P., Song, J., Sun, H.: Validity of the V parameter for photonic quasi-crystal fibers. Opt. Lett. **35**, 071064 (2010)
16. Veselago, V.: The electrodynamics of substances with simultaneously negative values of ε and μ. Sov. Phys. Uspekhi **10**, 509 (1968)
17. Papasimakis, N., Luo, Z., Shen, Z.X., De Angelis, F., Di Fabrizio, E., Nikolaenko, A.E., Zheludev, N.I.: Graphene in a photonic metamaterial. Opt. Express **18**, 8353 (2010)
18. Nikolaenko, A.E., De Angelis, F., Boden, S.A., Papasimakis, N., Ashburn, P., Di Fabrizio, E., Zheludev, N.I.: Carbon nanotubes in a photonic metamaterial. Phys. Rev. Lett. **104**, 153902 (2010)
19. Sámson, Z.L., MacDonald, K.F., De Angelis, F., Gholipour, B., Knight, K., Huang, C.C., Di Fabrizio, E., Hewak, D.W., Zheludev, N.I.: Metamaterial electro-optic switch of nanoscale thickness. Appl. Phys. Lett. **96**, 143105 (2010)
20. Mary, A., Rodrigo, S.G., Garcia-Vidal, F.J., Martin-Moreno, L.: Theory of negative-refractive-index response of double-fishnet structures. Phys. Rev. Lett. **101**, 103902 (2008)
21. Pérennès, F., Marmiroli, B., Matteucci, M., Tormen, M., Vaccari, L., Di Fabrizio, E.: Sharp beveled tip hollow microneedle arrays fabricated by LIGA and 3D soft lithography with polyvinyl alcohol. J. Micromech. Microeng. **16**, 473 (2006)
22. Tormen, M., Businaro, L., Altissimo, M., Romanato, F., Cabrini, S., Perennes, F., Proietti, R., Sun, H.-B., Kawata, S., Di Fabrizio, E.: 3D patterning by means of nanoimprinting. X-ray and two-photon lithography. Microelectron. Eng. **73–74**, 535 (2004)
23. Das, G., Mecarini, F., Gentile, F., De Angelis, F., Kumar, M.H.G., Candeloro, P., Liberale, C., Cuda, G., Di Fabrizio, E.: Nano-patterned SERS substrate: application for protein analysis vs. temperature. Biosens. Bioelectron. **24**, 1693 (2009)
24. Liberale, C., Minzioni, P., Bragheri, F., De Angelis, F., Di Fabrizio, E., Cristiani, I.: Miniaturized all-fibre probe for three dimensional optical trapping and manipulation. Nat. Photonics **1**, 723 (2007)
25. De Angelis, F., Patrini, M., Das, G., Maksymov, I., Galli, M., Businaro, L., Andreani, L.C., Di Fabrizio, E.: A hybrid plasmonic-photonic nanodevice for label-free detection of a few molecules. Nano Lett. **8**, 2321 (2008)
26. De Angelis, F., Das, G., Candeloro, P., Patrini, M., Galli, M., Bek, A., Lazzarino, M., Maksymov, I., Liberale, C., Andreani, L.C., Di Fabrizio, E.: Nanoscale chemical mapping using three-dimensional adiabatic compression of surface plasmon polaritons. Nat. Nanotechnol. **5**, 67 (2010)
27. Cabrini, S., Carpentiero, A., Kumar, R., Businaro, L., Candeloro, P., Prasciolu, M., Gosparini, A., Andreani, C., De Vittorio, M., Stomeo, T., Di Fabrizio, E.: Focused ion beam lithography for two dimensional array structures for photonic applications. Microelectron. Eng. **2005**(78–79), 11 (2005)
28. Stockman, M.I.: Nanofocusing of optical energy in tapered plasmonic waveguides. Phys. Rev. Lett. **93**, 137404 (2004)
29. Lafuma, A., Quéré, D.: Superhydrophobic states. Nat. Mater. **2**, 457 (2003)
30. Accardo, A., Gentile, F., Mecarini, F., De Angelis, F., Burghammer, M., Di Fabrizio, E., Riekel, C.: In situ X-ray scattering studies of protein solution droplets drying on micro- and nanopatterned superhydrophobic PMMA surfaces. Langmuir **26**, 15057 (2010)
31. De Angelis, F., Gentile, F., Mecarini, F., Das, G., Moretti, M., Candeloro, P., Coluccio, M.L., Cojoc, G., Accardo, A., Liberale, C., Zaccaria, R.P., Perozziello, G., Tirinato, L., Toma, A., Cuda, G., Cingolani, R., Di Fabrizio, E.: Breaking the diffusion limit with super hydrophobic delivery of few molecules to plasmonic nanofocusing structures. Nat. Photonics **5**, 682 (2011)

Nanostructures for Surface Functionalization and Surface Properties

Paola Rivolo
Dipartimento di Scienza dei Materiali e Ingegneria Chimica, Politecnico di Torino, Torino, Italy

Synonyms

Chemical modification; Functionalization; Organosilanes; Patterning; Physical modification; Plasma grafting; Plasma polymerization; Self-assembled monolayers; Surface properties; Thiols

Definition

Modification and functionalization of substrates can lead to the presence at materials surfaces of molecular nanostructures. Their 2D packing and chemical functionalities may impart to surfaces particular properties that can be very different from the pristine ones.

Modification is a very general term concerning every kind of treatment leading to change the physical morphology or surface chemistry of a given material in order to obtain tailored properties with dependence on the material application field.

Functionalization is mainly related to the introduction at a surface of chemical groups with a variable surface density which are requested to subsequently react with other species.

When a process or a group of processes are applied to the surface of an inorganic or organic material, affecting both topography/morphology and surface chemistry at the micro- and/or nanoscale level, the material surface results physically and chemically patterned.

In the following the main modification, functionalization, and patterning surface techniques will be described according to process methods, substrate chemistry, and nanotechnological application fields.

Surface Modification and Functionalization

In general, some materials which have excellent bulk physical and chemical properties often do not possess suitable surface properties required for specific applications.

For this reason, surface-modification techniques, that can transform these materials into valuable finished products, become an important part of surface science and technology.

Processes, developed in the last decade, are able to simply modify surface morphology of materials or give rise to nanostructured coatings aimed to simply affect the adhesion or anti-adhesion properties, mainly hydrophilicity and hydrophobicity, or to promote specific bonding with other species. These features can be a crucial advantage in several applications in the fields of protective, anti-adhesion, adhesion coatings, friction and wear, composites, microelectronics and thin-film technology, molding and lithography, technical fabrics, biomaterials, sensing and in particular biosensing.

Generally, these surface-modification methods can be divided into two classes: physical modifications and chemical modifications [1].

In this entry will be reported the main topics regarding surface chemical modifications, both the surface derivatization or grafting method which depend on the chemistry of the substrate surface and the functional organic coating deposition which can be performed on the surface of very different materials without the need of a specific chemical binding with the surface (no-dependence on chemistry).

Finally, some surface patterning procedures devoted to get functional structures will be described as following process of the previously reported methods or based on other modification techniques including physical processes.

Chemical Derivatization or Grafting

The key advantage of these techniques is that the surface of the materials can be modified or tailored to acquire very distinctive properties through the choice of different grafting precursors, while maintaining the substrate properties.

The crucial point is to single out the suitable molecule to be grafted according to the chemistry of the material substrate, that is, the kind of atoms/groups with which the substrate is truncated with.

Before the grafting step, commonly a surface activation is needed in order to create reactive sites on the substrate surface or enhance the amount of the ones naturally already present.

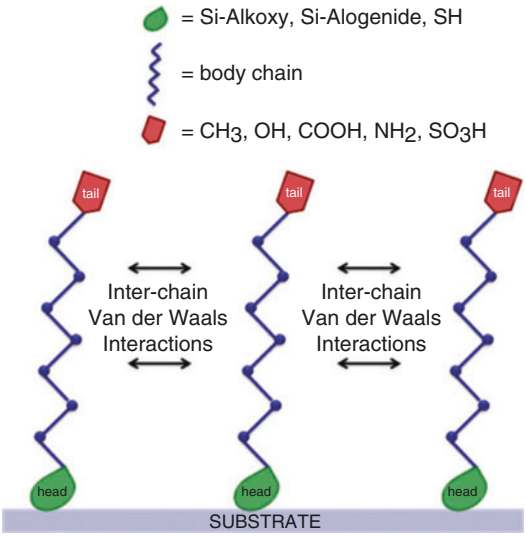
Practically, one can generate reactive groups through UV, high-energy electrons, γ irradiation, plasma treatment, ozone exposure [1] and chemical reactions performed in liquid and vapor phase. The kind of activation treatment heavily depends on materials.

For sake of simplicity, the different types of derivatizing treatments and the related materials activation procedures will be considered according to substrate chemistry.

Generally, among the surface-modification techniques, the formation of a Self-Assembled Monolayer (SAM) has been intensively studied. SAM is a kind of "bottom-up" technique which provides advantages regarding the surface modification. It leads to a structure which is at, or close to, thermodynamic equilibrium, and thus tends to self-healing/defect rejection and leads to a closely packed, well-ordered, and stable configuration on the surface [1]. SAMs offer unique opportunities to increase fundamental understanding of self-organization, structure–property relationships, and interfacial phenomena. The ability to tailor both head and tail groups of the constituent molecules makes SAMs excellent systems for a more fundamental understanding of phenomena (e.g., ordering and growth, wetting, adhesion, lubrication, corrosion, etc.) affected by competing intermolecular, molecule-substrate and molecule-solvent interactions.

SAMs provide the needed design flexibility, both at the individual molecular and at the material levels, and offer a vehicle for investigation of specific interactions at interfaces and of the effect of increasing molecular complexity on the structure and stability of two-dimensional assemblies.

The key factors characterizing the SAM precursor molecules consist in (a) the "head," (b) the "tail," and (c) the body chain length of the molecule (Fig. 1). The "head" is one of the terminal



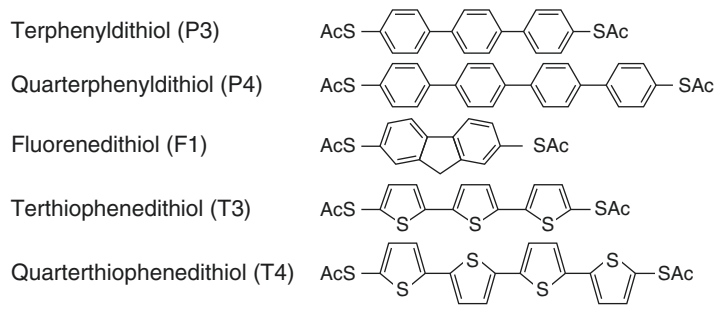
Nanostructures for Surface Functionalization and Surface Properties, Fig. 1 Scheme of a general SAM formed on the surface of a substrate

group of these normally linear species and has the role to anchor the surface by a favored and fast reaction producing the formation of a usually stable (covalent) bond linking the surface. The choice of this termination is driven by the chemistry of the material substrate.

The "tail" of the molecule, that is, the termination opposite to the "head," is deputed to perform the desired effect at the SAM surface. So as a rule of thumb, it can be a methyl termination ($–CH_3$), a perfluoroalkyl termination that generate at the most a hydrophobic or super-hydrophobic and repulsive behavior or a functional group such as –SH, –OH, $–NH_2$, –COOH able to carry out a subsequent reaction with other molecule (for example, biomolecules, polymers, suitable linker molecules) or simply express hydrophilic and adhesion properties.

The molecule length (the hydrocarbons chains can range, for instance, from C_3 to C_{24}) will strongly depend on the application for which SAM structures are needed.

Some reviews on the self-assembled monolayers have covered the SAM formation and structure their characterization by electrochemical and scanning tunneling microscopy [1], by PM-IRRAS vibrational spectroscopy techniques



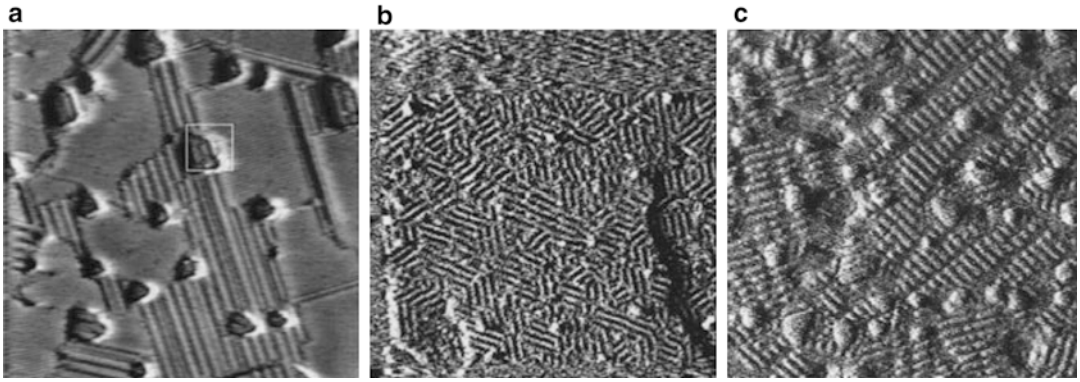
Nanostructures for Surface Functionalization and Surface Properties, Fig. 2 Schematic structures of conjugated molecules studied in [7]

and XPS analysis [2], their application to biosensors [1], anti-stiction in MEMS devices and in nano-imprinting lithography techniques [3], and their advantages in providing controlled surface properties and unique reactions [1].

In particular, such monolayers have been generated mostly on metal surfaces, such as Au, Ag, and Cu, using the thiol chemistry [1]. Nuzzo and Allara, reported in 1983, [4] the procedure of bifunctional organic sulfides species self-assembly on the surface of Au (111) freshly evaporated substrates in the liquid phase by simply putting in contact the samples with a diluted (milliMolar range) solution of the reagent, for some hours. Successively many authors observed with other n-alkyl thiolates that in the first few seconds the species are able to bind the gold surface and in more than 1 h, they self-assemble into a compact, well-ordered flat intercrossed molecular monolayer. The –SH head is able to establish a strong Au–S bond and the chain-to-chain interaction performed by the adjacent grafted molecules is dominated by weak dispersion forces if nonfunctional interacting groups are present in the middle of the aliphatic chain. In this case, H-bond or other stronger van der Waals forces can be exerted. The interchain forces are able to produce, as calculated by Godin [5] an equilibrium separation between two alkyl chains of 0.44 nm. The alkanethiol SAM is formed as a densely packed structure showing $(\sqrt{3} \times \sqrt{3})R30^\circ$ overlayer on gold surface where the molecule is tilted about 30° from the surface normal with the sulfur atom chemically bonded to gold atoms. This implies that a compressive surface stress has to be generated during the self-assembly, and this stress will increase with increasing alkyl chain length.

Patel et al. [6] studied the effect of chain length on the reactivity of SAMs toward protein immobilization generating SAMs from short- and long-chain carboxylic acid terminating alkyl-thiols, 11-MercaptoUndecanoic Acid (11-MUA), 3-MercaptoPropanoic Acid (3-MPA) and from the mixture, 1:10, of the both. The accessibility of the carboxylic groups by the amino group of a lysine termination exposed by the protein, *catalase*, to be anchored, decreases in the order Mixed > 11-MUA > 3-MPA. The experiment was performed by previously activating the carboxylic groups by a water soluble 1-ethyl-3-[3-(dimethylamino) propyl]carbodiimide hydrochloride (EDC) and *N*-hydroxysuccinimide (NHS). Successful formation of the NHS ester intermediate is reliant on the accessibility of the terminal carboxylate groups. It was assessed that a steric packing of these acid groups can limit the rate of intermediate formation with full conversion of accessible acid groups only occurring after several repeated reaction cycles. The co-adsorption of different alkyl-thiols provides the system a degree of disorder which is strategic for an efficient protein immobilization.

SAMs of conjugate alkyl-dithiols are one of the more interesting solutions focused by research in molecular electronics which is open to the possibility of substituting inorganic materials with organic molecules in the core regions of devices. Jiang [7] reported STM characterization of the electronic transport properties of SAM of containing double bonds (insaturation) in the body chain such as the aromatic dithiols (Fig. 2).



Nanostructures for Surface Functionalization and Surface Properties, Fig. 3 (**a**) STM image (95 × 95 nm) of octanethiol SAMs on Au(111); (**b**) STM image (70 × 70 nm) of decanedithiol SAMs on Au(111); (**c**) STM image (75 × 75 nm) of the mixture SAMs on Au(111)

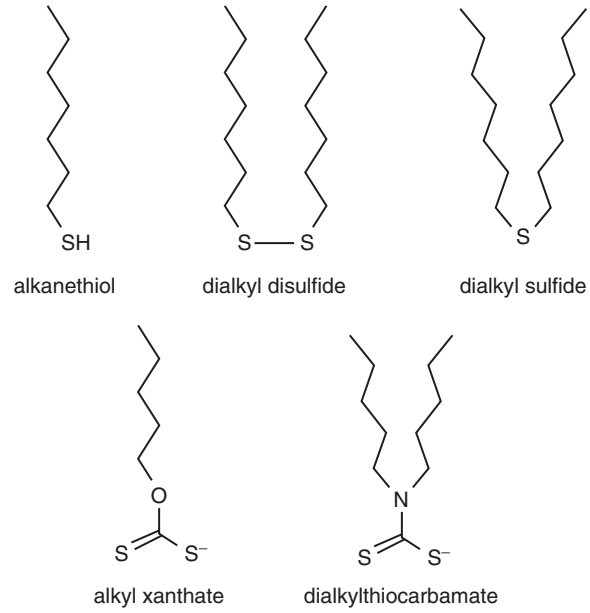
Co-assembly of conjugated dithiols and alkanemonothiols is suggested as a route to obtain dithiols that are denser packed and more vertical (with respect to what is obtained from pure dithiol routes). This surface condition is suitable to the evaporation on the top formed SAM of an Au layer and to the formation of other Au–S bonds with the free thiol end of the dithiol molecules.

Kobayashi [8] reported the study by Ultra High Vacuum Scanning Tunneling Microscopy (UHV-STM) of alkyl-thiol and alkyl dithiol SAMs on Au surfaces, characterizing the morphology of resulting nanostructures and the strength of molecules interaction with the surface. Figure 3 shows the different appearance of the surface according to the nature mono- or dithiolic of the molecule used to form the SAMs both by a liquid process in ethanolic solution (24 h) and by a vapor functionalization process (from a 10 mM ethanolic solution, for 2 h in a closed vessel). In (a) the 1,8-octanethiol SAM results in nanometric islands where molecules assume a surface-normal arrangements. In (b) the nanometric stripe structures are evident due to the molecules of dithiols (1,6-hexanedithiol, 1,10-decanedithiol) lying with their axes parallel to the surface (the bright spots are assigned as sulfur atoms). The different arrangement between (a) and (b) is probably ascribed to the stronger interaction between both thiol groups at the molecular ends with the gold atoms with respect to the case of alkanethiol SAMs. The parallel arrangement is more preferred than the normal arrangement in the case of alkanedithiol SAMs. In (c) the SAM is obtained by a mixture of the vapor of the ethanol solution which contained octanethiol and octanedithiol molecules at 1 mM each. The UHV-STM image shows stripe phase regions corresponding to the areas where alkanedithiol SAMs are formed. On the other hand, the island regions may be attributed to alkanethiol SAMs.

Ag (Silver) and other materials such as GaAs (Gallium Arsenide) can be functionalized by alkyl-thiols. Dubowski et al. [9] recently studied the formation mechanism of T-Alkanethiols (where T = NH_2, COOH, CH_3) SAMs on III–V semiconductors surfaces, in particular GaAs (001). They proposed the formation at the interface of a S–As or S–Ga bond according to which of the two elements is the richer in the exposed surface. They demonstrated the dependence of PhotoLuminescence signal on the terminal T group and the possibility to use these functionalized semiconducting materials in biosensing by PL technique.

In 1996 Ulman [10] reported other organosulfur molecules (Fig. 4) in addition to alkane thiols able to strongly graft and self-assemble on semiconductors surface, such as GaAs and InP, and other metals such as copper, platinum, mercury, iron, nanosize γ-Fe_2O_3 particles.

In Fig. 5a are displayed the diphosphates molecules used for Zirconium (Zr^{4+}) ions exploiting



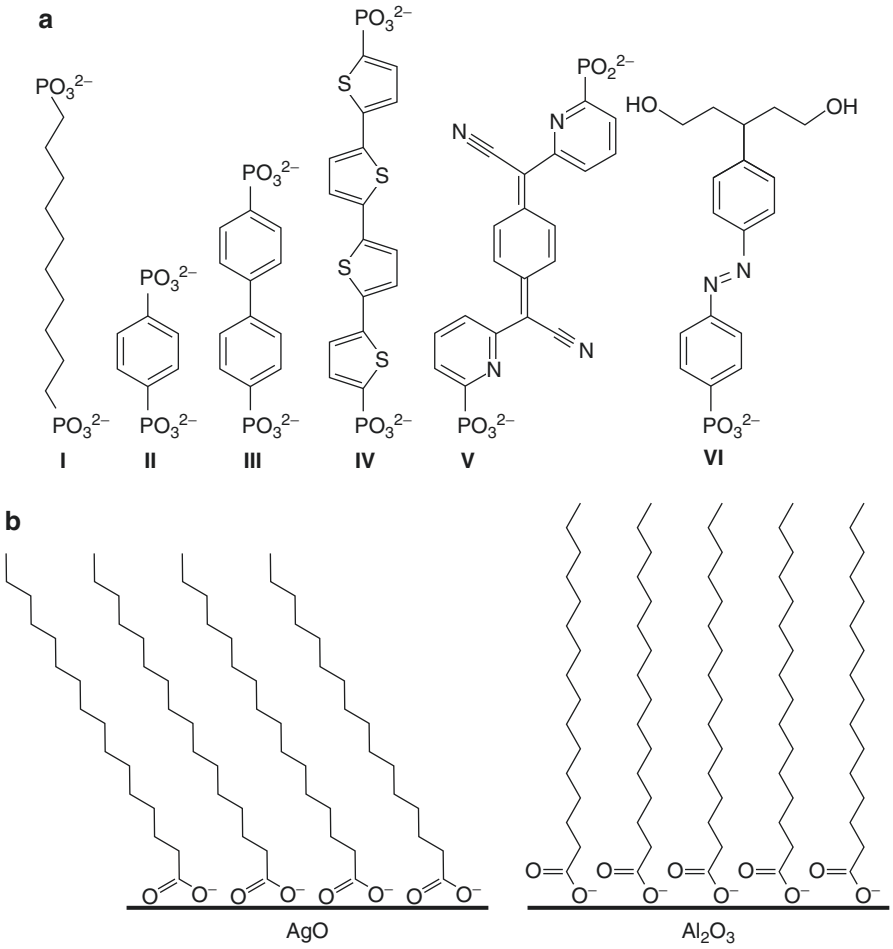
Nanostructures for Surface Functionalization and Surface Properties, Fig. 4 Surface-active organosulfur compounds that form monolayers on metals and semiconductors [10]

their ability to form highly insoluble salts with tetravalent transition metal ions. At last, fatty acids, which are long-chain n-alkanoic acids ($C_nH_{2n+1}COOH$), are known, by an acid–base reaction, to form a surface ordered salt between the carboxylate anion and a surface metal cation, for example, with AgO and Al_2O_3 (Fig. 5b).

For inorganic oxides, the typical molecules used to form SAMs monolayers are organosilanes. They are characterized by a variability of possible structures and by the easiness of reaction performed both in the liquid and in the vapor phase.

In particular, Silicon/Silica (Si/SiO_2), glass and quartz surfaces, Aluminum/Alumina (Al/Al_2O_3), phyllosilicates such as mica, magnetic materials such as maghemite (γ-Fe_2O_3) nanoparticles, germanium oxide [1], and Zinc oxide (ZnO) [11] exposing naturally hydroxyl-terminated (–OH) surfaces or zinc selenide [1] and Gallium Nitride (GaN) Aluminum Nitride (AlN) and Silicon Nitride (SiN) [12–14], where the –OH species can be easily created by a pretreatment, react with alkylchlorosilanes and alkylalkoxysilanes. Stable siloxy linkages [2, 3] are formed between hydroxylated Silica and 3-aminopropyltriethoxysilane (3-APTES). The reaction proceeds through the hydrolysis of silane alkoxy terminations and the condensation with surface hydroxyls and elimination of an alcohol molecule. Further condensation between adjacent molecules already grafted to the surfaces take part stabilizing the SAM through the formation of siloxane bridges (Si–O–Si).

Glass and Si/SiO_2 are the most common inorganic materials used in many technological applications as optical and electrical materials and as MEMS (MicroElectroMechanical Systems). One major factor that limits the widespread use and reliability of MEMS is adhesion. Adhesion is a result of the dominance of surface forces, such as capillary, hydrogen bonding, electrostatic, and van der Waals forces, over body forces at the microscale. When internal restoring forces of microstructures cannot overcome surface adhesive forces, the devices are said to suffer from stiction. Many procedures based on organosilanes SAMs formation are reported to impart to MEMSanti-stiction properties. Comparing the SAM formation from dimethyldichlorosilane ($(CH_3)_2SiCl_2$, DDMS) and tridecafluoro-1,1,2,2-tetra-hydrooctyltrichlorosilane ($CF_3(CF_2)_5(CH_2)_2SiCl_3$, FOTS) [3], by a liquid (thermal annealing about 65 °C in anhydrous toluene solution) and a vapor procedure (dosage of organosilane vapor pressure



Nanostructures for Surface Functionalization and Surface Properties, Fig. 5 (**a**) Different diphosphonic acids used in self-assembled multilayer preparation. (**b**) A schematic description of fatty acid monolayers on AgO and on Al_2O_3

on an Oxygen plasma pre-activated surface in a low pressure CVD style reactor), it results that the vapor phase method avoid vertical polymerization of precursor molecule producing a flatter SAM with the same efficiency in anti-stiction properties.

3-APTES has been shown to be an efficient SAM precursor [2, 15] for biosensing of protein tumoral markers based on micro-cantilever devices.

Both the functionalization routes described above are efficient in reducing the mass effect of the functional coating which is detrimental for the device sensitivity and in providing a maximum coverage in the order of 10^{14} $-NH_2$ groups/cm^2 [2].

In the field of magneto-chemotherapeutics, where spatial control of conjugated active biomolecules under magnetic field and Magnetic Resonance Imaging (MRI) contrast can be achieved simultaneously are becoming always more relevant functionalized magnetic nanoparticles (NP) able to conjugate specific biomolecules, for instance, Matrix MetalloProteinase (MMP), zinc-containing endopeptidases, which are responsible for the regulated degradation and processing of extracellular matrices (ECM), with physiological functions such as wound healing, tissue remodeling, tumor

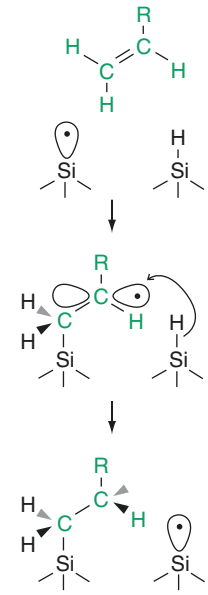
growth, and metastasis [16]. The silanization of the NP was conducted under anoxic (N_2 purging) suspension consisting of glycerol, aqueous methanol, and acetic acid. Mercaptopropyltrimethoxysilane (MPTMS) was added slowly to the suspension to ensure homogeneous hydrolysis of the silane agent, followed by reflux heating for 16 h. It is demonstrated that by this procedure it is possible not only to get at the NP surface a –SH terminated SAM able to react with MMP by the cross-linker Sulfosuccinimidyl-4--(*N*-maleimidomethyl)cyclohexane-1-carboxylate (Sulfo-SMCC), but also to reduce the iron oxide phase from γ-Fe_2O_3 to the higher magnetic moment phase of Fe_3O_4.

Other methods to functionalized Iron Oxide NP are reported by Wu et al. in [17]. For instance, alkyl phosphonates and phosphates are used which bind efficiently to iron oxide particle surfaces, and favor, due to their good biocompatibility, the utilization of encapsulated magnetic NPs in nanomedicines.

The no-oxide surfaces, such as nitrides, can be easily hydroxylated in order to exploit organosilanes chemistry in order to achieve surface functionalization. The well-assessed procedure consists in the substrate immersion in H_2SO_4:H_2O_2 (3:1) solution, also called Piranha solution, for 20 min, rinsed with deionized water and dried under Nitrogen flux prior to silanization. The highly oxidizing reagents produce, for GaN (deposited by MetalOrganic Chemical Vapor Deposition – MOCVD), an oxide layer exposing hydroxyl species. For AlN (also obtained by MOCVD) which is already covered by a native oxide layer, the only effect is surface hydroxylation [12].

The Piranha solution is used also on crystalline Silicon and on SiO_2 thermally grown from crystalline Silicon. On SiN (deposited by Plasma Enhanced Chemical Vapor Deposition, PECVD), a solution with 10 % nitric acid (HNO_3) at 80 °C for 20 min [13] or a mixture of CH_3COOH and H_2O_2 for 3 h [14] is used to activate the silanols (Si–OH); for the both water is used for rinsing and N_2 for drying.

Also polymers can be activated in order to react with organosilane "head" (trialkoxy- or trichloro-terminations) as suggested by Kühn et al. in [18]. The double procedure consists in an Oxygen plasma step in order to produce mixed oxygen containing species OH, C=O, C–O–C, CHO, COOH, C–O–OH, CO_3, C=C, etc., on a polyolefin polymer such as PolyEthylene (PE) or PolyPropylene (PP) and then a reduction step in diborane (gaseous procedure) or $LiAlH_4$ solution in order to obtain a monotype homofunctional surface exposing hydroxyls.



Nanostructures for Surface Functionalization and Surface Properties, Fig. 6 Proposed mechanism for surface hydrosilylation. Initial loss of silicon hydride generates a silicon dangling bond. Reaction between the silicon and an alkene molecule leads to an attached alkyl radical, which may abstract a hydrogen atom from a neighboring silicon

Since long time, Silicon surface reconstruction has been studied in order to understand the most suitable Silicon face (100) and (111) to get an organic functionalization avoiding detrimental side oxidation effects. Bent [19] reported some example of Silicon hydrosilylation by alkenes or alkines, molecules bearing C=C terminal groups that react with the Si–H of the Si surface under different activation methods.

A radical initiator, a species used to create free radicals in reactions, starts the reaction by abstracting H from a surface Si–H bond, producing a silicon dangling bond (which is also referred to as a silicon radical). The mechanism is illustrated in Fig. 6 on a (111) silicon surface.

The dangling bond is represented as an orbital with a dot, where the dot represents a single, unpaired electron. The silicon dangling bond

then reacts with the alkene molecule, forming a new Si–C bond at the surface and leading to the attachment of the organic molecule, which itself is left with a carbon radical. This surface bound organic radical is thought to abstract a hydrogen atom either from an unreacted alkene molecule or from a neighboring Si–H group on the surface. The net reaction sequence produces a stable, closed-shell organic group bound directly to the Si (111) surface. Hydrosilylation can be achieved also under UV irradiation in the range 200–400 nm, thermal activation at $T > 150$ °C (but a low quality of resulting organic monolayers has been noticed), introducing in the reactant mixture acidic molecules (called Lewis acids) to catalyze the reaction.

The main applications of organic layers on Silicon are related to protection of the underlying substrate from attack by etchants, attribution of chiral properties to the surface, addition of new optical properties to silicon and recently the substitution of organosilanes in sensing technology when the presence of oxide has to be avoided, also if the organic monolayers lacks in terms of order and cohesion, typical of SAMs.

Silicon/Silica and Diamond surface can be also directly derivatized through polymers grafting. On their own, polymers expose functionalities which are able to give further reactions.

Surface graft polymerization is a chemical modification method. In surface graft polymerization, the modification is achieved by grafting suitable macromolecular chains on the surface of materials through covalent bonding. The key advantage of these techniques is that the surface of the materials can be modified or tailored to acquire very distinctive properties through the choice of different grafting monomers, while maintaining the substrate properties. It also ensures an easy and controllable introduction of graft chains with a high density and exact localization onto the surface. Compared with the physically coated polymer chains, the covalent attachment of the grafted chains onto a material surface avoids their desorption and maintains a long-term chemical stability of the introduced chains.

Chiari et al. [20] used a *ter*-polymer of *N*,*N*-dimethylacrylamide, *N*-acryloyloxysuccinimide, and 3-(trimethoxysilyl)propyl methacrylate. The copolymer (DMA-NAS-MAPS) graft the SiO_2 surface of oxidized silicon or glass by a simple and robust procedure consisting in the dip coating in an aqueous solution, of substrates of different shapes in a short time span (less than 1 h) and in mild conditions without harmful treatments with harsh washing solutions. The film formed on the surface is thin, on the order of a few nanometers. For fluorescent microarray chips, this is a suitable coating as it does not affect the fluorescence intensification provided by the silicon/silicon oxide substrate. The trimethoxysilyl terminations react with surface silanols, as well as an organosilane, the acryloyloxysuccinimidyl ends give rise to covalent bonds with amino groups of biomolecules and the DMA units form the structural backbone of the polymer.

Recently, Oxidized Ultrananocrystalline diamond (UNCD) has been successfully coated with poly(styrene) (PS) using the self-initiated photografting and photopolymerization (SIPGP) approach with styrene as the monomer, within 16 h, under constant irradiation with UV light ($\lambda_{max} = 350$ nm) [21]. The obtained coating stable grafted to surface is then easily derivatized by sulphonation, nitration, and amido-alkylation carried out according to well-known liquid chemistry routes and so allowing the incorporation of additional functional molecules, such as fluorescent labels. This will open the way for designing more advanced biosensing schemes, incorporating multifunctional elements and with a higher loading capacity for biomolecules.

Ruckenstein et al. reported the graft polymerization reaction on already SAMs coated surfaces of glass and silicon wafers in order to introduce additional reactive sites. The method involves the aniline substitution at Br-terminated SAM through the aniline amino termination and further the surface oxidative graft polymerization of aniline on the modified glass surface via the covalently immobilized aniline sites. A PolyAniline (PANI) coating is so obtained, performing the reaction in the polymerization solution containing aniline, and due to its well-known conductivity

(conductive polymer) this can be a suitable substrate to be used as biosensor and biomaterial.

Another route to further derivatize an organosilane surface for biosensing purposes consists in the grafting, from a 8.5 pH solution, of homofunctional linkers such as Glutaraldehyde ($OHC–CH_2–CH_2–CH_2–CHO$) which is able through an aldehyde end to react with NH_2 exposed at amino-SAM modified surface and, then, through the second aldehyde termination to immobilize biomolecules exposing amino terminations.

This method has been successfully applied by Ricciardi et al. [15] for functionalization of cantilever based biosensors aimed to the detection of tumoral markers (Angiopoietin).

Surface Modification: No-Dependence on Materials Chemistry

Among the surface-modification techniques, deposition of coating obtained by plasma polymerization presents, respect to surface derivatization discussed in the previous paragraph, the advantage to allow the tailoring of properties of any kind of surface without involving the substrate surface chemistry in the functionalization mechanism.

Macromolecular plasma chemistry [22] has developed in the last two decades in the following directions: plasma-enhanced synthesis (deposition and/or grafting) of thin layer macromolecular structures; surface functionalization of polymeric materials; and etching of inorganic or polymeric substrate surfaces. Plasma-enhanced synthesis involves the dissociation of starting materials and reorganization of the resulting neutral and charged molecular fragments into macromolecular structures on the surfaces located inside or outside of the plasma zone. Outside the plasma zone, recombination mechanisms usually result in the incorporation of molecular fragments with lower degrees of dissociation into the nascent macromolecular structures. When the starting materials are common monomers the recombination processes are more complex due to the development of simultaneous conventional polymerization reactions along with the fragment-recombination mechanisms initiated by the plasma-created and surface-attached active species (e.g., ions, free radicals). In these cases, complex structures result and structural identification is difficult.

However, an advantage of conventional monomer-based processes is that specific functionalities can be retained in the polymer structures and consequently some characteristics can be more easily predicted and designed into the final product.

Macromolecular thin layers can also be synthesized by the plasma generation of active sites (e.g., ions, free radicals) or reactive functionalities (primary amine groups, etc.) on substrate surfaces, followed by the development of graft-polymerization reactions in the absence of the plasma in the presence of "true monomers." Remote (the recombination processes are developed outside the plasma zone) and pulsed (plasma sustained according to preselected duty cycles) plasma processes, allow recombination mechanisms in the absence of the plasma state in time or in space, and are characterized by significantly minimized dissociation processes. These approaches do not change dramatically the nature of molecular fragmentation processes (the nature and relative ratios of charged and neutral active species); however, they limit significantly the intensities of plasma-induced fragmentation mechanisms, and allow the development of recombination processes in the partial or total absence of plasma state. As a result, molecular precursors can be physically or through covalent bonding incorporated through surface-located charged and neutral (e.g., free radicals) active species, into the nascent macromolecular layers.

Owing to the high reactivity of plasma species, surface functionalization reactions of even the most inert polymeric substrates can be conveniently achieved. These mechanisms usually involve non-polymerizing gas plasmas, which generate active molecular fragments that covalently attach during the plasma reactions to the activated substrate surfaces. Recently, it has been demonstrated that even neutral molecules can be connected to active sites with covalent bonds on plasma-exposed substrates under in situ conditions [22]. At high particle energies, etching

processes dominate both the deposition and functionalization mechanisms.

Depending on the nature of the plasma gases, inert (sputtering) or reactive etching mechanisms can be developed. It should, however, be noted that these mechanisms can have practical importance; they allow the creation of specific surface morphologies by selective ablation, for instance, of amorphous and crystalline surface regions.

The nature of plasmas, the modalities of transferring electric or electromagnetic field intensities to the reaction systems, the geometry of reactor (shape and volumes of the reaction vessel, geometrical location of electrodes and substrates, etc.) and the selected experimental conditions (pressure, power, flow rates of gases, temperature of the substrates, etc.) crucially influence the gas-phase and surface-related plasma chemistry. Consequently, comparing experimental results from a large variety of plasma systems, even for identical starting components, is extremely difficult.

Cold-plasma-mediated processes involve both gas-phase and surface reaction mechanisms. The gas-phase reaction mechanisms involve the interaction of ab initio existing and nascent neutral and charged plasma-created species, including atoms, molecules, free radicals, ions of either polarity, excited species, electrons, and photons. Besides the recombination mechanisms developed on the surfaces which confine the plasma, the active species of the discharge interact and continuously tailor the artificially exposed (reactor walls, various substrates, etc.) and self-generated (e.g., plasma-synthesized macromolecular structures) surface layers. The competition between the recombination-deposition (deposition, grafting, functionalization) processes and "destructive interaction" of plasma species (etching) with the nascent macromolecular structures will control the intensities and the predominance of ablation, surface functionalization and macromolecular-film-formation reactions.

The mechanisms of surface functionalization of polymeric substrates are different from the gas-phase processes. While electrons play the most important role in the plasma state, positive ions also play a significant role in the surface chemistry during the interactions of plasma species with polymers. The resulting valence-ionized polymer chains undergo neutralization reactions leading to sufficiently intense localized internal energy concentrations (electronically excited states), which can induce homolytic bond cleavages. Free-radical and unsaturated-bond development can result in cross-linking of polymeric layers. Free radicals can further induce chemical reactions involving plasma species in situ and affect reactions controlled by specific chemical environments (gas-phase or condensed-phase compounds, including, oxygen, monomer molecules, etc.) in the absence of plasma.

Qualitative and quantitative evaluation of charged and neutral species of the gas phase and identification of charged and neutral surface functionalities on substrate surfaces is essential for understanding the plasma-induced reaction mechanisms and physical processes. Adaptation of in situ diagnostic techniques (including real time measurements) for relatively high and low pressure (100–1000 mTorr) plasma environments is a research area where many questions still remain unanswered. Continued research will promote the development of novel and/or improved processes for tailoring materials surface characteristics and the development of advanced laboratory plasma installations, required for scaling up plasma processes to industrial levels.

Recently, surface treatments and modification based on vapor phase plasma-assisted techniques have been widely applied to several biomedical fields. These processes are able to provide specific monotype chemical functionalities, exposed at materials surfaces. Respect to traditional liquid phase procedures, advantages such as the high process control and the use of small quantities of reagents, are crucial for the enhancement of efficiency of the interactions that materials must perform according to the application.

In particular, procedures for in situ plasma polymerization, starting from vapor released by liquid monomers, are presently used to synthesize innovative polymeric thin films applicable as substrates for cell culture [23], as adhesion promoter in prosthetic implants and for molecular recognition in sensors and biosensors.

The low pressure plasma polymerization, starting from the same reagent and only varying the process parameters (modulation of the plasma discharge, monomer or mixture of monomers/process gas, reactant vapor partial pressure, etc.) leads to the deposition of polymeric materials with different features in terms of density of functional groups, wettability (hydrophilicity/hydrophobicity), and chemical stability [24].

The most studied polymers in biomedical and sensing fields, display as functional groups mainly carboxyls (–COOH) and amines ($–NH_2$), hydroxyls (–OH), and aldehydes (–CHO). Fewer reports exist on hydroxy and aldehyde surfaces prepared by plasma methods. Hydroxy surfaces can be prepared by water plasma treatment or the plasma polymerization of alkyl alcohol vapors. Water plasma treatment on many polymer substrates suffers from aging, with surface adaptation leading to the movement of surface-modification effects into the polymer. Both hydroxy and aldehyde surfaces have been used for the covalent immobilization of biologically active molecules. Aging effects are less well documented than for amine surfaces.

Aminated surfaces have been fabricated using various plasma vapors or mixtures (allylamines, ethylenediamine, diaminopropane, *n*-heptylamine [24]) and have found wide use for biointerface applications. However, in many cases the amine surfaces have a rather limited shelf life, with post-plasma oxidation reactions and surface adaptation leading to the disappearance of amine groups from the surface. Aging is a widespread phenomenon that often has not been recognized, particularly in some of the earlier studies on the use of plasma-fabricated surfaces for biointerfacial applications, and can markedly alter the surface chemistry [23].

Plasma-fabricated surfaces that contain carboxyl groups have also been well documented and are obtained, for example, by acrylic acid [23] and vinyl acetic acid or applying CO_2 plasma on olefin polymers [24].

In order to better tune the distribution and density of functional groups, are also widely known processes of plasma copolymerization, where the monomer containing functional groups of interest is mixed with a precursor which does not contain specific functionalities (e.g., styrene or octadiene [25]) and which has the function to dilute the reactive groups at the surface. For example, carboxylate copolymer surfaces have shown excellent ability to support the colonization of some human cell lines of clinical interest. Immobilization of proteins onto plasma-carboxylated surfaces is also well established. Generally, these surface chemistries show good ability to support cell colonization, though the effectiveness seems to depend on the process vapor and the plasma conditions.

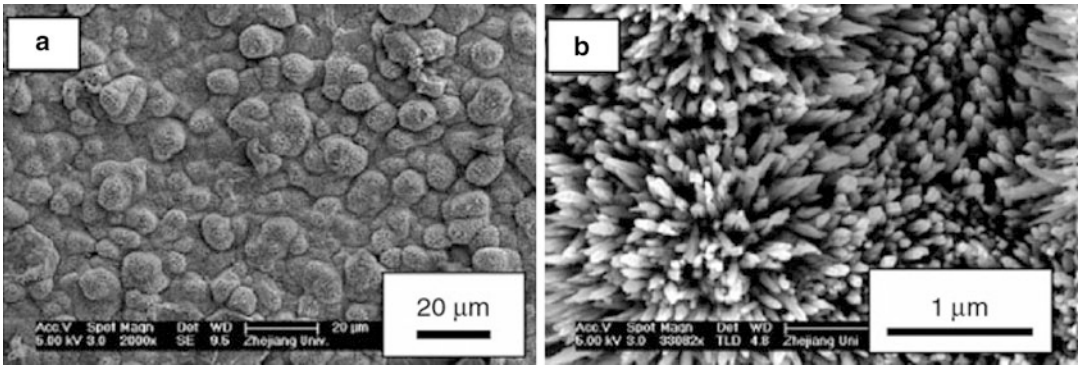
An example in which process parameters affect the final properties of the plasma-polymer film is reported by Detomaso et al. [23]. Starting from Acrylic Acid vapor, processes obtained by Continuous Wave (CW) plasma discharge and by Modulated Wave (MW), that is pulsed, plasma discharge are compared respect to the efficiency of the final polymer coating toward cells growth. The –COOH terminated polymer obtained by MW, with a Duty Cycle (D.C. = $(T_{on}/T_{on} + T_{off})$ of 5, where $T_{on} = 5$ ms and $T_{off} = 95$ ms, is less useful to murine fibroblast growth than the CW plasma film, because more soluble in water due to the higher density of –COOH groups and the higher retention of the monomer (acrylic acid) structure (low cross-linking and backbone chain propagation).

Materials Surface Modification: Morphology Effects

Fabrication of micro- and nano-patterned surfaces is a fundamental step for the development of engineered materials.

For biometerials, it is well known that a series of interactions occur between their surface and the biological environment after they have been implanted into the human body. Therefore, the biomaterials surface plays an extremely important role in the response of artificial medical devices to the biological environment.

The efficacy of artificial implants is determined mainly by their surface characteristics such as surface morphology, microstructure, composition, and properties. Liu et al. [26] reported many surface-modification methods mainly based on inorganic



Nanostructures for Surface Functionalization and Surface Properties, Fig. 7 SEM images of a titania nanorod formed on titanium substrate: (**a**) low magnification, (**b**) high magnification

nano-coatings deposition and functionalization at the nanoscale of obtained structures. Typical coating and film deposition techniques are plasma spraying, plasma immersion ion implantation and deposition (PIII&D), solgel, chemical vapor deposition (CVD), physical vapor deposition (PVD), cold spraying, self-assembly, and so on, whereas in situ surface-modification techniques include laser etching, shot blasting, acid and alkali treatments, anodic oxidation, micro-arc oxidation, ion implantation, etc.

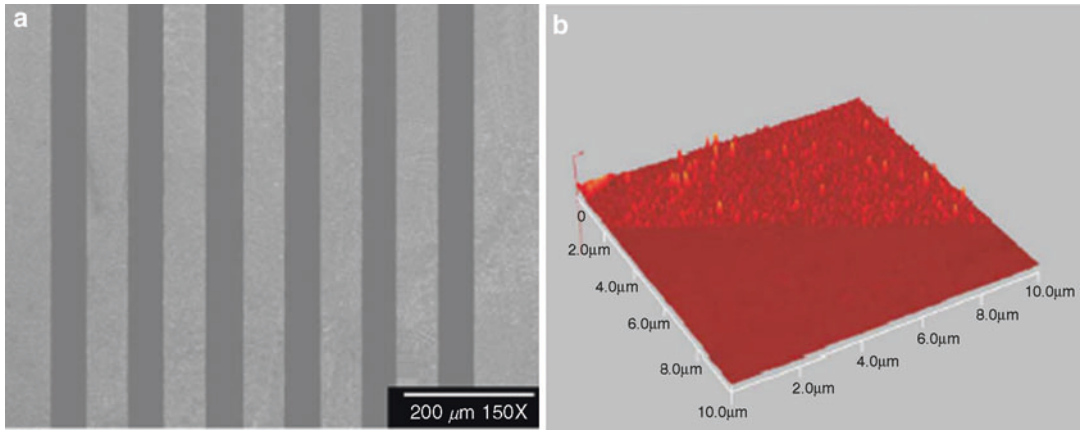
In bioceramics field, for example, a critical grain size for osteoblast adhesion between 49 and 67 nm for alumina and 32 and 56 nm for titania, provides evidence about the ability of nanophase alumina and titania to simulate the desirable materials characteristics of bones in a physiological environment, particularly those associated with the enhancement of protein interactions such as adsorption, configuration, and bioactivity as well as subsequent osteoblast adhesion.

Metal surfaces with low micrometer to nanophase topography enhance adhesion of osteoblasts. A surface with various nanostructures can be obtained by chemically treating titanium and its alloys. Chemical treatments are suitable for the creation of micrometer-scale and nanometer-scale textures on large-area surfaces and multifaceted devices with complex 3D shape due to the non-line-of-sight nature. Titania nanorods can be produced at low temperature on a large scale via direct oxidation of titanium with hydrogen peroxide. The ordered nanorod layer with a thickness of approximately 1 mm is deposited on a 2 mm thick condensed anatase layer as illustrated by Fig. 7.

Concerning chemical patterning, microstructured surfaces containing opposite bio-properties such as cell and/or protein adhesive features in a non-bio-adhesive matrix provide valuable tools in the field of biotechnology and numerous methods have been successfully developed for the production of such surfaces. Besides, the development of chemical nanopatterning technology is great of interest due to its ability to provide powerful tools for studying fundamental aspects of cell adhesion at the level of single protein/receptor molecules or for the implementation of new generations of miniaturized biochips requiring smaller reagent quantities with lower detection limits [27].

In general [28], the chemical patterning techniques commonly used in bioengineering can be divided into four categories: (a) lithographic patterning (photo-assisted and electron-beam assisted), (b) scanning probe microscopies (SPM); (c) transfer via polymeric stamps (soft lithography), and (d) physical patterning. The challenge with plasma polymerization lies in developing patterning techniques by which, reproducible spatial resolution and distribution of chemistries is achieved and the specific functional properties of the thin films are retained.

The simplest patterning approach reported by literature is the one based on the use of physical mask (solid materials deposited on defined areas



Nanostructures for Surface Functionalization and Surface Properties, Fig. 8 (**a**) SEM image of the patterned PANI film on a silicon substrate; and (**b**) selected edge area of (**a**) under AFM

of the substrate), e.g., polymeric masks and TEM grids.

Sometimes, the substrates are previously coated with another plasma polymer obtained starting from a precursor which does not contain specific functionalities (e.g., styrene or octadiene [25]) with the role of non-bio-adhesive matrix.

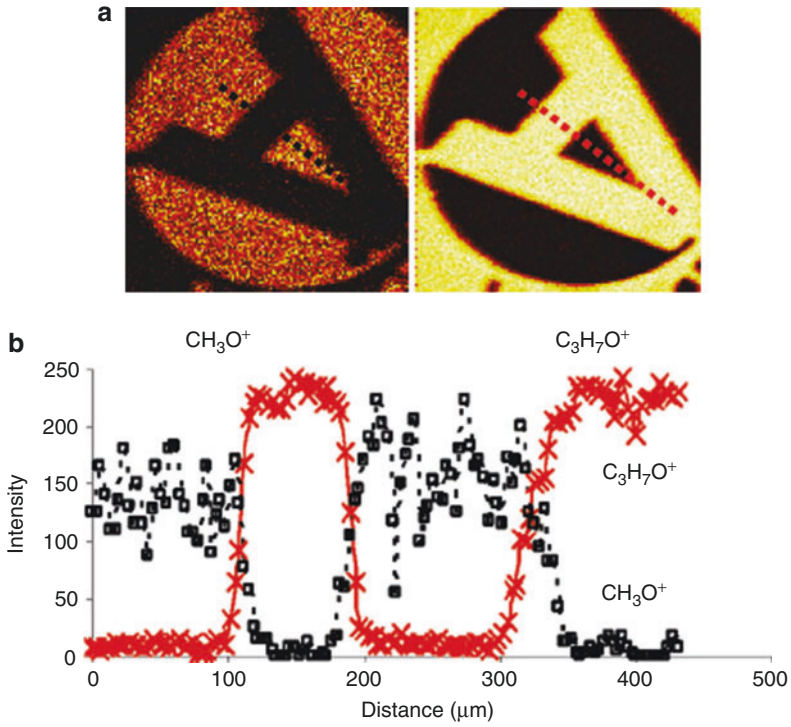
Ruckenstein et al. [1] reported a method for graft polymerization patterning of polyaniline on Si surface previously coated by SAM monolayers from phenyltrichlorosilanes and further subjected to a laser photolithographic process. The subsequent treatment with trifluoromethanesulfonic acid (triflic acid; HOTf) results in the generation of reactive silyl cation sites on the backbones, which can react with suitable monomers to form functionalized polysilanes.

The aniline monomer undergoes to nucleophilic substitution reaction followed by the surface oxidative graft polymerization of aniline on the covalently immobilized aniline sites. The grafted PANI thus formed provided a negative of the lithographic mask. Figure 8a presents the scanning electron microscopy (SEM) image of the patterned PANI after the graft polymerization. The white strips represent the grafted PANI areas, whereas the black strips, the nongrafted ones. A relatively clear boundary between the patterned and nonpatterned areas could be observed in the SEM image. A closer look at the boundaries between the two can be achieved under a higher magnification by AFM (Fig. 8b), which demonstrated that this method provides a high line edge acuity of the patterned PANI.

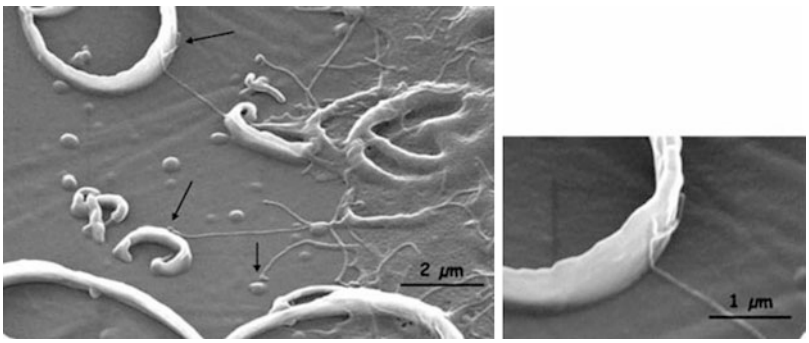
Mishra et al. [28] studied, by using a TEM grid placed onto an already plasma polymer coated silicon wafer, a wide range of standard size patterns in order to vary the resultant pattern features from approximately 5–200 μm. After the plasma polymerization process starting from, for example, glycidyl methacrylate (ppGMA), maleic anhydride (ppMA), acrylic acid (ppAAc), allylamine (ppAAm), tetraglyme (ppTG), and grid remotion, AFM characterization highlighted the presence, at the physical boundaries, of a diffused interface rather than a sharp step (Fig. 9) and Time-of-Flight Secondary Ion Mass Spectrometry (ToF-SIMS) analysis showed that there was diffusion of the plasma deposit below the physical mask during the second polymer layer deposition.

On the other side, ion beam etching and plasma etching through masks as well as electron beam lithography (EBL) [27], have been shown to be able to produce chemically resolved microscale and in some cases nanoscale plasma polymer patterns.

Bretagnol et al. [27] showed that chemically active features of polyacrylic acid (pAA),with lateral size down to 300 nm, exposing –COOH groups, obtained by EBL and the use of a photoresist mask, were able to graft a model protein



Nanostructures for Surface Functionalization and Surface Properties, Fig. 9 Positive single ion ToF-SSIMS images of ppMA patterned photolithographically on Si wafer precoated with ppTTg (500 × 500 μm^2). Unique secondary molecular fragments are presented in *inset* (**a**). The pixel intensity plot for secondary ions $C_3H_7O^+$ (ppTTg) and CH_3O^+ (ppMA) shown in *inset* (**b**). The image resolution was calculated to be ~10 μm, based on 80:20 definition



Nanostructures for Surface Functionalization and Surface Properties, Fig. 10 SEM images of MG 63 osteoblasts on hydrophobic structured samples showing in detail how filopodia could interact with nanostructures on the surface

(bovine serum albumin), even if a wet chemical step in acetone was performed.

Photolithographic patterning for plasma polymers is a particularly attractive approach for the fabrication of complex chemical arrays and the opportunity presented by plasma polymerization to produce thin films from a very wide range of different monomers with several properties such as surface charge, hydrophilicity, biospecificity, and adhesion properties alternated to the opposite such as hydrophobicity and, generally, non-fouling characteristics. Recently, it has been demonstrated that specific properties including protein resistance are maintained after performing the photolithographic patterning technique, with both spatial and chemical resolution down to 1 μm [27].

Finally, an exotic method for surface nanopatterning, based on plasma polymerization has been reported by Gristina et al. [29]. Teflon ribbon-like nanostructures (Fig. 10) have been obtained by C_2F_4 MW (low D.C.) plasma polymerization and it has been demonstrated that these structures promote fibroblast cell adhesion

through proteins interaction of cell filopodia with the nanostructures surface notwithstanding the well-known teflon super-hydrophobicity.

Cross-References

- Applications of Nanofluidics
- Bioadhesion
- BioPatterning
- Biosensors
- Charge Transport in Self-Assembled Monolayers
- Microfabricated Probe Technology
- Nanophotonic Structures for Biosensing
- Nanostructured Functionalized Surfaces
- Self-Assembled Monolayers

References

1. Ruckenstein, E., Li, Z.F.: Surface modification and functionalization through the self-assembled monolayer and graft polymerization. Adv. Colloid Interface Sci. **113**, 43–63 (2005)
2. Fiorilli, S., Rivolo, P., Descrovi, E., Ricciardi, C., Pasquardini, L., Lunelli, L., et al.: Vapor-phase self-assembled monolayers of aminosilane on plasma-activated silicon substrates. J. Colloid Interface Sci. **321**, 235–241 (2008)
3. Ashurst, W.A., Carraro, C., Maboudian, R., Frey, W.: Wafer level anti-stiction coatings for MEMS. Sens.. Actuator A Phys. **104**, 213–221 (2003)
4. Nuzzo, R.G., Allara, D.L.: Adsorption of bifunctional organic disulfides on gold surfaces. J. Am. Chem. Soc. **105**, 4481–4483 (1983)
5. Godin, M.: Surface stress, kinetics and structure of alkanethiol self-assembled monolayers. Ph.-D. dissertation, McGill University (2004)
6. Patel, N., Davies, M.C., Heaton, R.J., Roberts, C.J., Tendler, S.J.B., Williams, P.M.: A scanning probemicroscopy study of the physisorption and chemisorption of protein molecules onto carboxylate terminated self-assembledmonolayers. Appl.. Phys.. A **66**, S569–S574 (1998)
7. Jiang, W., Zhitenev, N., Bao, Z., Meng, H., Abusch-Magder, D., Tennant, D., et al.: Structure and bonding issues at the interface between gold and self-assembled conjugated dithiol monolayers. Langmuir **21**, 8751–8757 (2005)
8. Kobayashi, K., Horiuchi, T., Yamada, H., Matsushige, K.: STM studies on nanoscopic structures and electric characteristics of alkanethiol and alkanedithiol self-assembled monolayers. Thin Solid Films **331**, 210–215 (1998)
9. Dubowski, J.J., Voznyy, O., Marshall, G.M.: Molecular self-assembly and passivation of GaAs (0 0 1) with alkanethiol monolayers: a view towards bio-functionalization. Appl. Surf. Sci. **256**, 5714–5721 (2010)
10. Ulman, A.: Formation and structure of self-assembled monolayers. Chem. Rev. **96**, 1533–1554 (1996)
11. SelegÅrd, L., Khranovskyy, V., Söderlind, F., Vahlberg, C., Ahrén, M., Käll, P.-O., Yakimova, R., Uvdal, K.: Biotinylation of ZnO nanoparticles and thin films: a two-step surface functionalization study. ACS Appl. Mater. Interfaces **2**, 2128–2135 (2010)
12. Baur, B., Steinhoff, G., Hernando, J., Purrucker, O., Tanaka, M., Nickel, B., Stutzmann, M., Eickhoff, M.: Chemical functionalization of GaN and AlN surfaces. Appl. Phys. Lett. **87**, 263901-1–263901-3 (2005)
13. Tlili, A., Jarboui, M.A., Abdelghani, A., Fathallah, D.M., Maaref, M.A.: A novel silicon nitride biosensor for specific antibody–antigen interaction. Mater.. Sci. Eng.. C Biomimetic. Supramol.. Syst. **25**, 490–495 (2005)
14. Lee, S.-H., Lee, C.-S., Shin, D.-S., Kim, B.-G., Lee, Y.-S., Kim, Y.-K.: Micro protein patterning using a lift-off process with fluorocarbon thin film. Sens.. Actuator B-Chem. **99**, 623–632 (2004)
15. Ricciardi, C., Fiorilli, S., Bianco, S., Canavese, G., Castagna, R., Ferrante, I., et al.: Development of microcantilever-based biosensor array to detect Angiopoietin-1, a marker of tumor angiogenesis. Biosens. Bioelectron. **25**, 1193–1198 (2010)
16. Li, D., Teoh, W.Y., Djunaedi, D., Gooding, J.J., Selomulya, C., Amal, R.: Facile functionalization and phase reduction route of magnetic iron oxide nanoparticles for conjugation of matrix metalloproteinase. Adv. Eng. Mater. **12**, B210–B214 (2010)
17. Wu, W., He, Q., Jiang, C.: Magnetic iron oxide nanoparticles: synthesis and surface functionalization strategies. Nanoscale Res. Lett. **3**, 397–415 (2008)
18. Kühn, G., Weidner, S., Decker, R., Ghode, A., Friedrich, J.: Selective surface functionalization of polyolefins by plasma treatment followed by chemical reduction. Surf. Coat. Technol. **116–119**, 796–801 (1999)
19. Bent, S.F.: Organic functionalization of group IV semiconductor surfaces: principles, examples, applications, and prospects. Surf. Sci. **500**, 879–903 (2002)
20. Cretich, M., Di Carlo, G., Longhi, R., Gotti, C., Spinella, N., Coffa, S., et al.: High sensitivity protein assays on microarray silicon slides. Anal. Chem. **81**, 5197–5203 (2009)
21. Steenackers, M., Lud, S.Q., Niedermeier, M., Bruno, P., Gruen, D.M., Feulner, P., et al.: Structured polymer grafts on diamond. J. Am. Chem. Soc. **129**, 15655–15661 (2007)
22. Denes, F.S., Manolache, S.: Macromolecular plasma-chemistry: an emerging field of polymer science. Prog. Polym. Sci. **29**, 815–885 (2004)

23. Detomaso, L., Gristina, R., D'Agostino, R., Senesi, G. S., Favia, P.: Plasma deposited acrylic acid coatings: surface characterization and attachment of 3 T3 murine fibroblast cell lines. Surf. Coat. Technol. **200**, 1022–1025 (2005)
24. Siow, K.S., Britcher, L., Kumar, S., Griesser, H.J.: Plasma methods for the generation of chemically reactive surfaces for biomolecule immobilization and cell colonization – a review. Plasma Process. Polym. **3**, 392–418 (2006)
25. Swaraj, S., Oran, U., Friedrich, J.F., Lippitz, A., Unger, W.E.S.: Surface chemical analysis of plasma-deposited copolymer films prepared from feed gas mixtures of ethylene or styrene with allyl alcohol. Plasma Process. Polym. **4**, 376–389 (2007)
26. Liu, X., Chu, P.K., Ding, C.: Surface nano-functionalization of biomaterials. Mater.. Sci. Eng.. R Rep. **70**, 275–302 (2010)
27. Brétagnol, F., Ceriotti, L., Valsesia, A., Sasaki, T., Ceccone, G., Gilliland, D., et al.: Fabrication of functional nano-patterned surfaces by a combination of plasma processes and electron-beam lithography. Nanotechnology **18**, 135303–135307 (2007)
28. Mishra, G., Easton, C.D., McArthur, S.L.: Physical vs. photolithographic patterning of plasma polymers: an investigation by ToF-SSIMS and multivariate analysis. Langmuir **26**, 3720–3730 (2010)
29. Gristina, R., D'Aloia, E., Senesi, G.S., Milella, A., Nardulli, M., Sardella, E., et al.: Increasing cell adhesion on plasma deposited fluorocarbon coatings by changing the surface topography. J. Biomed.. Mater.. Res.. B Appl.. Biomater. **88B**, 139–149 (2009)

Nanostructures: Nanodevices

► Nanostructures for Energy

Nanotechnology

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

BioMEMS/NEMS; MEMS; Micro-/Nanodevices; Micro-/Nanomachines; Micro-/Nanosystems; MOEMS/NOEMS; NEMS; RF-MEMS/NEMS

Definition

Nanotechnology literally means any technology done on a nanoscale that has applications in the real world. Nanotechnology encompasses the production and application of physical, chemical, and biological systems at scales ranging from individual atoms or molecules to submicron dimensions, as well as the integration of the resulting nanostructures into larger systems.

Overview

Nanotechnology is likely to have a profound impact on world economy and society in the early twenty-first century, comparable to that of semiconductor technology, information technology, or cellular and molecular biology. Science and technology research in nanotechnology promises breakthroughs in areas such as materials and manufacturing, nanoelectronics, medicine and healthcare, energy, biotechnology, information technology, and national security. It is widely felt that nanotechnology will be the next industrial revolution.

Nanometer-scale features are mainly built up from their elemental constituents. Examples include chemical synthesis, the spontaneous self-assembly of molecular clusters (molecular self-assembly) from simple reagents in solution, biological molecules (e.g., DNA) used as building blocks for the production of three-dimensional nanostructures, or quantum dots (nanocrystals) of an arbitrary diameter (about 10–105 atoms). The definition of a nanoparticle is an aggregate of atoms bonded together with a radius between 1 and 100 nm. It typically consists of 10–105 atoms. A variety of vacuum deposition and nonequilibrium plasma chemistry techniques are used to produce layered nanocomposites and nanotubes. Atomically controlled structures are produced using molecular beam epitaxy and organo-metallic vapor phase epitaxy. Micro- and nanosystem components are fabricated using top-down lithographic and nonlithographic fabrication techniques and range in size from micro- to nanometers. Continued improvements in

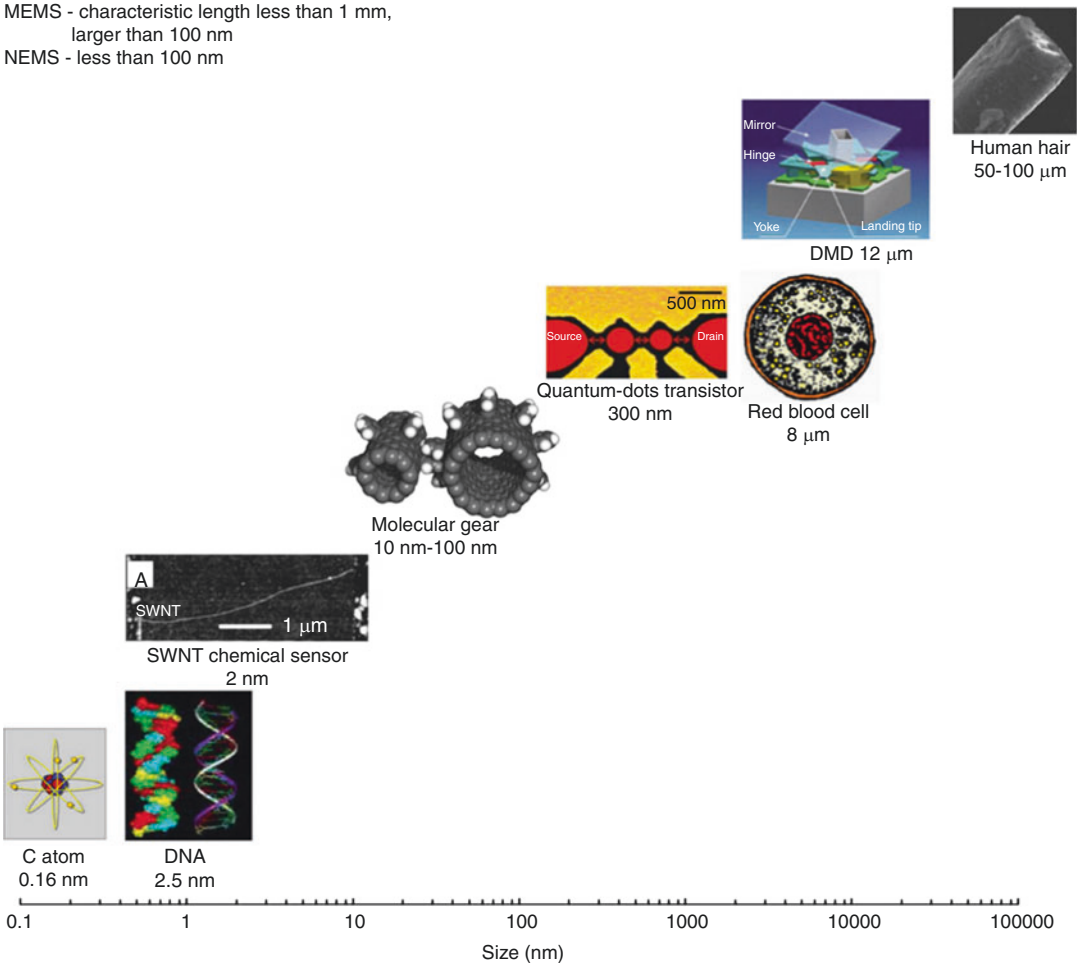
lithography for use in the production of nanocomponents have resulted in line widths as small as 10 nm in experimental prototypes. The nanotechnology field, in addition to the fabrication of nanosystems, provides impetus to the development of experimental and computational tools.

The discovery of novel materials, processes, and phenomena at the nanoscale and the development of new experimental and theoretical techniques for research provide fresh opportunities for the development of innovative nanosystems and nanostructured materials. The properties of materials at the nanoscale can be very different from those at a larger scale. When the dimension of a material is reduced from a large size, the properties remain the same at first; then small changes occur, until finally when the size drops below 100 nm, and dramatic changes in properties can occur. If only one length of a three-dimensional nanostructure is of the nanodimension, the structure is referred to as a quantum well; if two sides are of nanometer length, the structure is referred to as a quantum wire. A quantum dot has all three dimensions in the nanorange. The word quantum is associated with these three types of nanostructures because the changes in properties arise from the quantum-mechanical nature of physics in the domain of the ultra-small. Materials can be nanostructured for new properties and novel performance. This field is opening new venues in science and technology.

Micro- and nanosystems include Micro/NanoElectroMechanical Systems. MEMS refers to microscopic devices that have a characteristic length of less than 1 mm but more than 100 nm and combine electrical and mechanical components. NEMS refers to nanoscopic devices that have a characteristic length of less than 100 nm and combine electrical and mechanical components. In mesoscale devices, if the functional components are on the micro- or nanoscale, they may be referred to as MEMS or NEMS, respectively. These are referred to as an intelligent miniaturized system comprising of sensing, processing, and/or actuating functions and combine electrical and mechanical components. The acronym MEMS originated in the USA. The term commonly used in Europe is microsystem technology (MST), and in Japan it is micromachines. Another term generally used is micro/nanodevices. MEMS/NEMS terms are also now used in a broad sense and include electrical, mechanical, fluidic, optical, and/or biological function. MEMS/NEMS for optical applications are referred to as micro/nanooptoelectromechanical systems (MOEMS/NOEMS). MEMS/NEMS for electronic applications are referred to as radio-frequency-MEMS/NEMS or RF-MEMS/RF-NEMS. MEMS/NEMS for biological applications are referred to as BioMEMS/BioNEMS.

To put the dimensions of MEMS/NEMS and BioNEMS in perspective, see Fig. 1 and Table 1 [4]. Individual atoms are typically a fraction of a nanometer in diameter, DNA molecules are about 2.5 nm wide, biological cells are in the range of thousands of nm in diameter, and human hair is about 75 μm in diameter. Smaller length of BioNEMS shown in the figure is about 2 nm, NEMS ranges in size from 10 to 300 nm, and the size of MEMS is 12,000 nm. The mass of a micromachined silicon structure can be as low as 1 nN, and NEMS can be built with mass as low as 10–20 N with cross-sections of about 10 nm. In comparison, the mass of a drop of water is about 10 μN, and the mass of an eyelash is about 100 nN.

MEMS/NEMS and BioMEMS/BioNEMS are expected to have a major impact on people's lives, comparable to that of semiconductor technology, information technology, or cellular and molecular biology [5, 6]. MEMS/NEMS and BioMEMS/BioNEMS are used in electromechanical, electronics, information/communication, chemical, and biological applications. The MEMS industry in 2004 was worth about $4.5 billion with a projected annual growth rate of 17 %, Fig. 2 [7]. The NEMS industry was worth about $10 billion in 2004, mostly in nanomaterials, Fig. 2 [8]. Growth of Si-based MEMS/NEMS may slow down, and nonsilicon MEMS may pick up during the next decade. It is expected to expand in this decade, in nanomaterials and biomedical



Nanotechnology, Fig. 1 Dimensions of MEMS/NEMS and BioNEMS in perspective. Examples shown are of a single-walled carbon nanotube (SWNT) chemical sensor [1], of molecular dynamic simulations of carbon-nanotube–based gears [2], quantum-dot transistor obtained from van der Wiel et al. [3], and DMD obtained from www.dlp.com. For comparison, dimensions and weights of various biological objects found in nature are also presented

applications as well as in nanoelectronics or molecular electronics. For example, miniaturized diagnostics could be implanted for the early diagnosis of illness. Targeted drug delivery devices are under development. Due to the enabling nature of these systems and because of the significant impact they can have on both commercial and defense applications, industry as well as federal governments have taken special interest in seeing growth nurtured in this field. MEMS/NEMS and BioMEMS/BioNEMS are the next logical step in the "silicon revolution."

Background and Research Expenditures

On December 29, 1959 at the California Institute of Technology, Nobel Laureate Richard P. Feynman gave a talk at the Annual Meeting of the American Physical Society that has become one of the twentieth century classic science lectures, titled "There's Plenty of Room at the Bottom" [9]. He presented a technological vision of extreme miniaturization in 1959, several years before the word "chip" became part of the lexicon. He talked about the problem of manipulating and

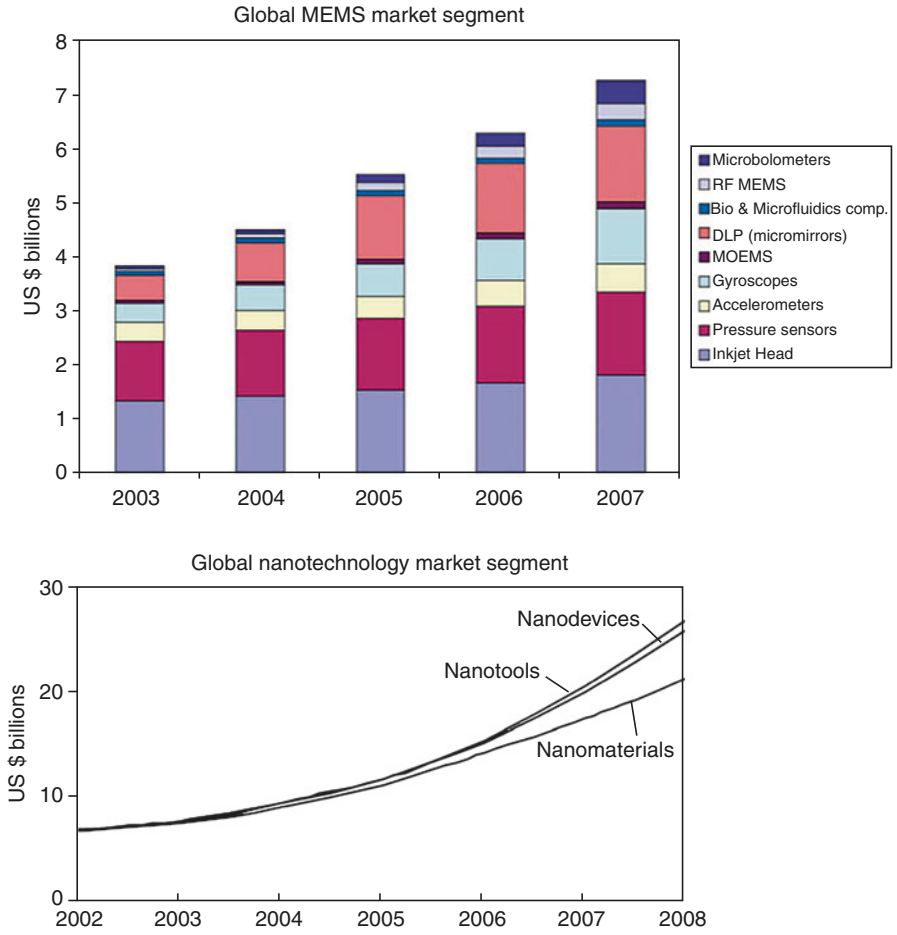
Nanotechnology, Table 1 Characteristic dimensions and weights in perspective

(a) Characteristic dimensions in perspective	
NEMS characteristic length	<100 nm
MEMS characteristic length	<1 mm and > 100 nm
SWCNT Chemical sensor	~2 nm
Molecular gear	~10 nm
Quantum-dots transistor	300 nm
Digital micromirror	12,000 nm
Individual atoms	Typically fraction of a nm in diameter
DNA molecules	~2.5 nm wide
Biological cells	In the range of thousands of nm in diameter
Human hair	~75,000 nm in diameter
(b) Weight in perspective	
NEMS built with cross-sections of about 10 nm	As low as 10–20 N
Micromachines silicon structure	As low as 1 nN
Eyelash	~100 nN
Water droplet	~10 μN

controlling things on a small scale. Extrapolating from known physical laws, Feynman envisioned a technology using the ultimate toolbox of nature, building nanoobjects atom by atom or molecule by molecule. Since the 1980s, many inventions and discoveries in the fabrication of nanoobjects have been testaments to his vision. In recognition of this reality, the National Science and Technology Council (NSTC) of the White House created the Interagency Working Group on Nanoscience, Engineering and Technology (IWGN) in 1998. In a January 2000 speech at the same institute, former President W. J. Clinton talked about the exciting promise of "nanotechnology" and the importance of expanding research in nanoscale science and technology more broadly. Later that month, he announced in his State of the Union Address an ambitious $497 million federal, multi-agency national nanotechnology initiative (NNI) in the fiscal year 2001 budget, and made the NNI a top science and technology priority [10, 11]. The objective of this initiative was to form a broad-based coalition in which the academe, the private sector, and local, state, and federal governments work together to push the envelope of nanoscience and nanoengineering to reap nanotechnology's potential social and economic benefits.

The funding in the USA has continued to increase. In January 2003, the US senate introduced a bill to establish a National Nanotechnology Program. On December 3, 2003, former President George W. Bush signed into law the twenty-first century Nanotechnology Research and Development Act. The legislation put into law programs and activities supported by the National Nanotechnology Initiative. The bill gave nanotechnology a permanent home in the federal government and authorized $3.7 billion to be spent in the 4-year period beginning in October 2005 for nanotechnology initiatives at five federal agencies. The funds would provide grants to researchers, coordinate R&D across five federal agencies (National Science Foundation (NSF), Department of Energy (DOE), NASA, National Institute of Standards and Technology (NIST), and Environmental Protection Agency (EPA)), establish interdisciplinary research centers, and accelerate technology transfer into the private sector. In addition, the Departments of Defense (DOD), Homeland Security, Agriculture, and Justice as well as the National Institutes of Health (NIH) also fund large R&D activities. They currently account for more than one-third of the federal budget for nanotechnology.

The European Union (EU) made nanosciences and nanotechnologies a priority in the Sixth Framework Program (FP6) in 2002 for a period of 2003–2006. They had dedicated small funds in FP4 and FP5 before. FP6 was tailored to help better structure European research and to cope with the strategic objectives set out in Lisbon in 2000. Japan identified nanotechnology as one of its main research priorities in 2001. The funding levels increased sharply from $400 million in 2001 to around $950 million in 2004. In 2003, South Korea embarked upon a 10-year program with around $2 billion of public funding, and Taiwan has committed around $600 million of public funding over 6 years. Singapore and China are also investing on a large scale. Russia is well funded as well.

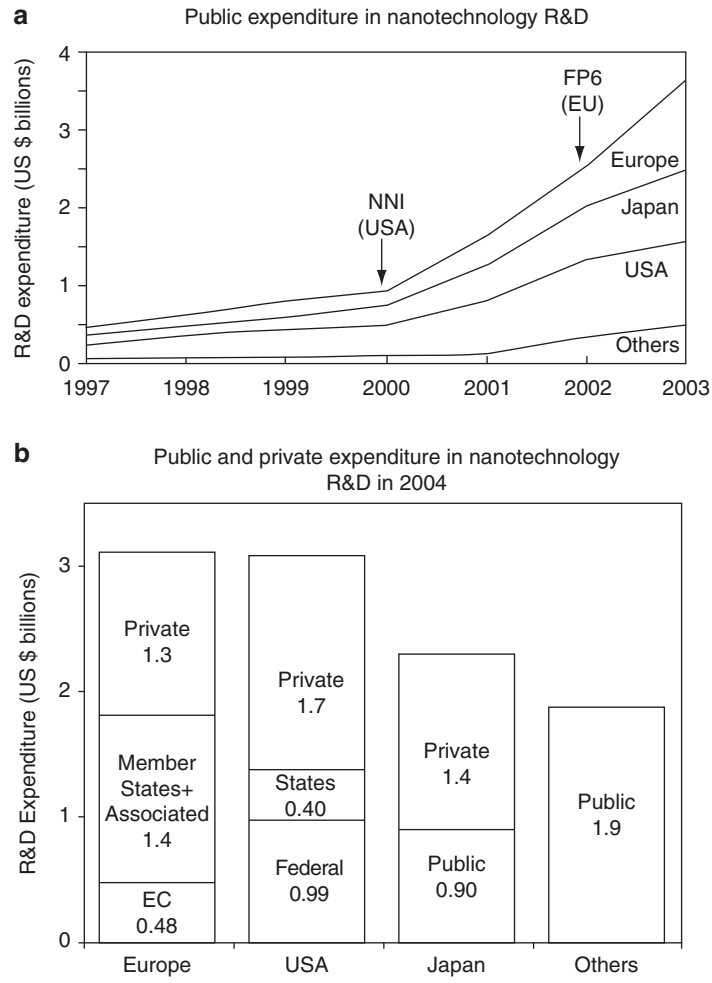


Nanotechnology, Fig. 2 Global MEMS and nanotechnology market segments

Figure 3a shows the public expenditure breakdown in nanotechnology R&D around the world, with about $5 billion in 2004, being about equal by USA, Japan, and Europe. Next public expenditure on a per capita basis is compared. The average expenditures per capita for the USA, EU-25, and Japan are about $3.7 billion, $2.4 billion, and $6.2 billion, respectively [12]. Figure 3b shows the breakdown of expenditures in 2004 by public and private sources with more than $10 billion spent in nanotechnology research. Two thirds of this came from corporate and private funding. The private expenditure in USA and Japan was slightly larger than that of the public, whereas in Europe it was about one-third. Figure 4 shows the public and private expenditure breakdown in 2004 in various countries. Japan and USA had the largest expenditure, followed by Germany, Taiwan, South Korea, UK, Australia, China, France, and Italy. Figure 5 shows a breakdown of worldwide publications and patents. USA and Canada led, followed by Europe and Asia. Figure 6 shows the breakdown in start-up companies around the world (1997–2002). Entrepreneurship in USA is clearly evident followed by Europe.

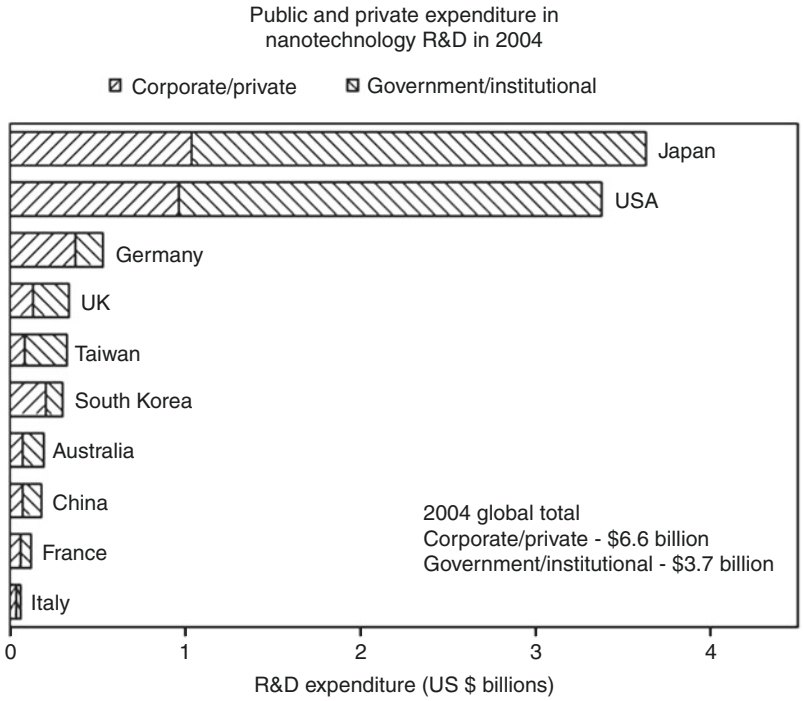
Applications in Different Fields

Science and technology continue to move forward in masking the fabrication of micro/nanodevices



Nanotechnology, Fig. 3 Breakdown of public expenditure in nanotechnology R&D (**a**) around the world (European Commission 2003), and (**b**) by public and private resources in 2004 (European Commission 2005; figures from private sources based upon Lux Research)

and systems possible for a variety of industrial, consumer, and biomedical applications (e.g., [4, 13, 14]). A variety of MEMS devices have been produced, and some are commercially used [4, 15–24]. A variety of sensors are used in industrial, consumer, defense, and biomedical applications. Various micro/nanostructures or micro/nanocomponents are used in micro-instruments and other industrial applications such as micromirror arrays. The largest "killer" MEMS applications include accelerometers (some 90 million units installed in vehicles in 2004), silicon-based piezoresistive pressure sensors for manifold absolute pressure sensing for engines and for disposable blood pressure sensors (about 30 million units and about 25 million units, respectively), capacitive pressure sensors for tire pressure measurements (about 37 million units in 2005), thermal inkjet printheads (about 500 million units in 2004), and digital micromirror arrays for digital projection display (about $700 million revenue in 2004), and optical cross-connections in telecommunications. Other applications of MEMS devices include chemical/biosensors and gas sensors, microresonators, infrared detectors and focal plane arrays for earth observations, space science and missile defense applications, pico-satellites for space applications, fuel cells, and many hydraulic, pneumatic, and other consumer products. MEMS devices are also being pursued in magnetic storage systems [25], where they are being developed for super compact and ultrahigh-recording-density magnetic disk drives.

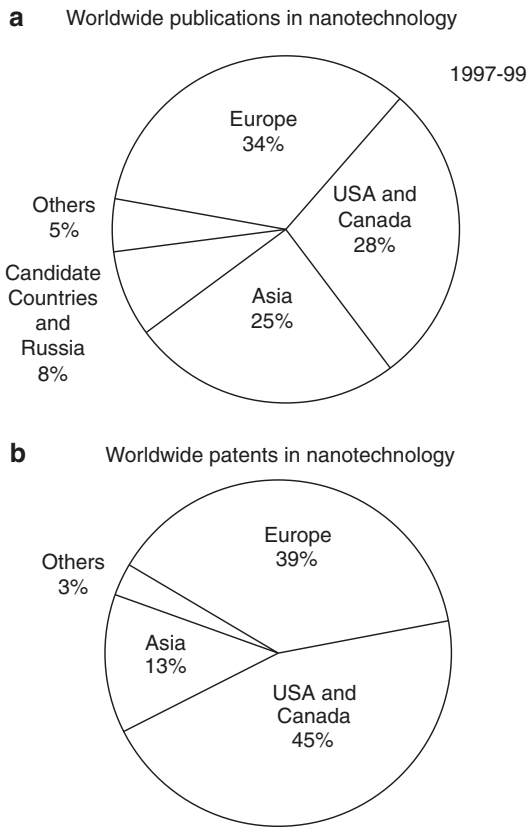


Nanotechnology, Fig. 4 Breakdown of public and private expenditures in nanotechnology R&D in 2004 in various countries [8]

NEMS are produced by nanomachining in a typical top-down approach and bottom-up approach, largely relying on nanochemistry [4, 26–32]. Examples of NEMS include microcantilevers with integrated sharp nanotips for scanning tunneling microscopy (STM) and atomic force microscopy (AFM), quantum corral formed using STM by placing atoms one by one, AFM cantilever array for data storage, AFM tips for nanolithography, dip-pen lithography for printing molecules, nanowires, carbon nanotubes, quantum wires (QWRs), quantum boxes (QBs), quantum-dot transistors, nanotube-based sensors, biological (DNA) motors, molecular gears by attaching benzene molecules to the outer walls of carbon nanotubes, devices incorporating nm-thick films (e.g., in giant magnetoresistive or GMR read/write magnetic heads and magnetic media) for magnetic rigid disk drives and magnetic tape drives, nanopatterned magnetic rigid disks, and nanoparticles (e.g., nanoparticles in magnetic tape substrates and magnetic particles in magnetic tape coatings).

Nanoelectronics can be used to build computer memory using individual molecules or nanotubes to store bits of information, molecular switches, molecular or nanotube transistors, nanotube flat-panel displays, nanotube integrated circuits, fast logic gates, switches, nanoscopic lasers, and nanotubes as electrodes in fuel cells.

BioMEMS/BioNEMS are increasingly used in commercial and defense applications (see, e.g., [4, 33–39]. They are used for chemical and biochemical analyses (biosensors) in medical diagnostics (e.g., DNA, RNA, proteins, cells, blood pressure and assays, and toxin identification) [39], tissue engineering [40], and implantable pharmaceutical drug delivery [41, 42]. Biosensors, also referred to as biochips, deal with liquids and gases. There are two types of biosensors. A large variety of biosensors are based on micro/nanofluidics. Micro/nanofluidic devices offer the ability to work with smaller reagent volumes and shorter reaction times, and perform analyses multiple times at once. The second type of biosensors includes micro/nanoarrays which perform one type of analysis thousands of times. Micro/nanoarrays are a tool used in biotechnology research to analyze DNA or proteins to diagnose diseases or discover new drugs. Also called DNA arrays, they can identify thousands of genes simultaneously



Nanotechnology, Fig. 5 Breakdown of (**a**) worldwide publications and (**b**) worldwide patents (European Commission 2003)

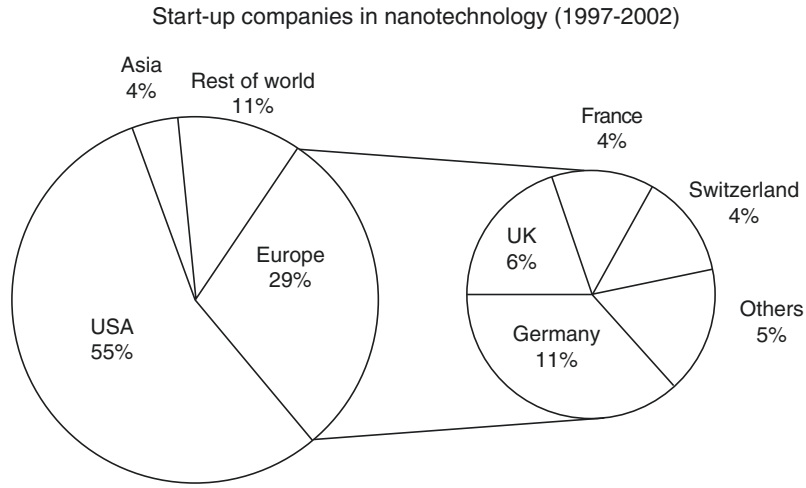
[35]. They include a microarray of silicon nanowires, roughly a few nm in size, to selectively bind and detect even a single biological molecule, such as DNA or protein, by using nanoelectronics to detect the slight electrical charge caused by such binding, or a microarray of carbon nanotubes to electrically detect glucose.

After the tragedy of September 11, 2001, concern of biological and chemical warfare has led to the development of handheld units with bio- and chemical sensors for detection of biological germs, chemical or nerve agents, and mustard agents, and chemical precursors to protect subways, airports, water supply, and population at large.

BioMEMS/BioNEMS are also being developed for minimal invasive surgery including endoscopic surgery, laser angioplasty, and microscopic surgery. Other applications include implantable drug-delivery devices – micro/nanoparticles with drug molecules encapsulated in functionalized shells for site-specific targeting applications, and a silicon capsule with a nanoporous membrane filled with drugs for long-term delivery.

Various Issues

There is an increasing need for a multidisciplinary, system-oriented approach to the manufacturing of micro/nanodevices which



Nanotechnology, Fig. 6 Breakdown of start-up companies around the world (1997–2002) (CEA, Bureau d'Etude Marketing)

function reliably. This can only be achieved through the cross-fertilization of ideas from different disciplines and the systematic flow of information and people among research groups. Common potential failure mechanisms for MEMS/NEMS requiring relative motion that need to be addressed in order to increase their reliability are: adhesion, friction, wear, fracture, fatigue, and contamination [4, 43–46]. Surface micro/nanomachined structures often include smooth and chemically active surfaces. Due to large surface area to volume ratio in MEMS/NEMS, they are particularly prone to stiction (high static friction) as part of normal operation. Fracture occurs when the load on a microdevice is greater than the strength of the material. Fracture is a serious reliability concern, particularly for brittle materials used in the construction of these components, since it can immediately or would eventually lead to catastrophic failures. Additionally, debris can be formed from the fracturing of microstructures, leading to other failure processes. For less brittle materials, repeated loading over a long period of time causes fatigue that would also lead to the breaking and fracturing of the device. In principle, this failure mode is relatively easy to observe and simple to predict. However, the material properties of thin films are often not known, making fatigue predictions error-prone.

Many MEMS/NEMS devices operate near their thermal dissipation limit. They may encounter hot spots that may cause failures, particularly in weak structures such as diaphragms or cantilevers. Thermal stressing and relaxation caused by thermal variations can create material delamination and fatigue in cantilevers. In large temperature changes, as experienced in the space environment, bimetallic beams will also experience warping due to mismatched coefficients of thermal expansion. Packaging has been a big problem. The contamination that probably happens in packaging and during storage also can strongly influence the reliability of MEMS/NEMS. For example, a particulate dust landed on one of the electrodes of a comb drive can cause catastrophic failure. There are no MEMS/NEMS fabrications standards, which make it difficult to transfer fabrication steps in MEMS/NEMS between boundaries.

Obviously, studies of determination and suppression of active failure mechanisms affecting this new and promising technology are critical to high reliability of MEMS/NEMS and are determining factors in successful practical application.

Adhesion between a biological molecular layer and the substrate, referred to as "bioadhesion," and reduction of friction and wear of biological layers, biocompatibility, and biofouling for BioMEMS/BioNEMS are important.

Mechanical properties are known to exhibit a dependence on specimen size. Mechanical property evaluation of nanometer-scaled structures is carried out to help design reliable systems since good mechanical properties are of critical importance in such applications. Some of the properties of interest are: Young's modulus of elasticity, hardness, bending strength, fracture toughness, and fatigue life. Finite element modeling is carried out to study the effects of surface roughness and scratches on stresses in nanostructures. When nanostructures are smaller than a fundamental physical length scale, conventional theory may no longer apply, and new phenomena emerges. Molecular mechanics is used to simulate the behavior of a nano-object.

The societal, ethical, political, and health/safety implications are getting major attention [12]. One of the prime reasons is to avoid some of the public skepticism that surrounded the debate over biotech advances such as genetically modified foods, while at the same time dispelling some of the misconceptions the public may already have about nanotechnology. Health/safety issues need to be addressed as well. For example, one key question is what happens to nanoparticles (such as buckyballs or nanotubes) in the environment and are they toxic in the human body, if digested.

Cross-References

- ▶ Biomimetics
- ▶ MEMS
- ▶ Micromachines

- ▶ Nanocalorimetry
- ▶ Nanofabrication
- ▶ NEMS
- ▶ Use of Nanotechnology in Pregnancy

References

1. Chen, R.J., Choi, H.C., Bangsaruntip, S., Yenilmex, E., Tang, X., Wang, Q., Chang, Y.L., Dai, H.: An investigation of the mechansims of electrode sensing of protein adsorption on carbon nanotube devices. J. Am. Chem. Soc. **126**, 1563–1568 (2004)
2. Srivastava, D.: Computational nanotechnology of carbon nanotubes. In: Meyyappan, M. (ed.) Carbon Nanotubes: Science and Applications, pp. 25–36. CRC Press, Boca Raton (2004)
3. van der Wiel, W.G., De Franceschi, S., Elzerman, J. M., Fujisawa, T., Tarucha, S., Kauwenhoven, L.P.: Electron transport through double quantum dots. Rev. Mod. Phys. **75**, 1–22 (2003)
4. Bhushan, B.: Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
5. Anonymous: Microelectromechanical Systems: Advanced Materials and Fabrication Methods, NMAB-483. National Academy, Washington, DC (1997)
6. Roukes, M.: Nanoelectromechanical systems face the future. Phys. World **14**, 25–31 (2001)
7. Eloy, J.C.: Status of the MEMS industry 2005. Report Yole Developpement, France. Presented at SPIE Photonics West, San Jose (2005)
8. Lawrence, S.: Nanotech grows up. Technol. Rev. **108**(6), 31 (2005)
9. Feynman, R.P.: There's plenty of room at the bottom. Eng. Sci. **23**, 22–36 (1960). www.zyvex.com/nanotech/feynman.html
10. Amato, I.: Nanotechnology. www.ostp.gov/nstc/html/iwgn/iwgn.public.brochure/welcome.htm or www.nsf.gov/home/crssprgm/nano/nsfnnireports.htm (2000)
11. Anonymous: National nanotechnology initiative. www.ostp.gov/nstc/html/iwgn.fy01budsuppl/nni.pdf or www.nsf.gov/home/crssprgm/nano/nsfnnireports.htm (2000)
12. Anonymous: Towards a European Strategy for Nanotechnology, European Commission, Research Directorate, Brussels (2004)
13. Anonymous: Small Tech 101 – An Introduction to Micro and Nanotechnology. Small Times (2003)
14. Schulenburg, M.: Nanotechnology – Innovation for Tomorrow's World. European Commission Research DG, Brussels (2004)
15. Bhushan, B.: Tribology Issues and Opportunities in MEMS. Kluwer, Dordrecht (1998)
16. Elwenspoek, M., Wiegerink, R.: Mechanical Microsensors. Springer, Berlin (2001)
17. Fujimasa, I.: Micromachines: A New Era in Mechanical Engineering. Oxford University Press, Oxford (1996)
18. Hierlemann, A.: Integrated Chemical Microsensor Systems in CMOS Technology. Springer, Berlin (2005)
19. Hsu, T.R.: MEMS and Microsystems: Design and Manufacture. McGraw-Hill, Boston (2002)
20. Kovacs, G.T.A.: Micromachined Transducers Sourcebook. WCB McGraw-Hill, Boston (1998)
21. Madou, M.: Fundamentals of Microfabrication: The Science of Miniaturization, 2nd edn. CRC Press, Boca Raton (2002)
22. Senturia, S.D.: Microsystem Design. Kluwer, Boston (2000)
23. Trimmer, W.S. (ed.): Micromachines and MEMS, Classical and Seminal Papers to 1990. IEEE Press, New York (1997)
24. Muller, R.S., Howe, R.T., Senturia, S.D., Smith, R.L., White, R.M.: Microsensors. IEEE Press, New York (1991)
25. Bhushan, B.: Tribology and Mechanics of Magnetic Storage Devices, 2nd edn. Springer, New York (1996)
26. Dresselhaus, M.S., Dresselhaus, G., Avouris, P.: Carbon Nanotubes – Synthesis, Structure, Properties, and Applications. Springer, Berlin (2001)
27. Drexler, K.E.: Nanosystems: Molecular Machinery, Manufacturing and Computation. Wiley, New York (1992)
28. Timp, G. (ed.): Nanotechnology. Springer, New York (1999)
29. Goddard, W.A., Brenner, D.W., Lyshevski, S.E., Iafrate, G.J. (eds.): Handbook of Nanoscience, Engineering, and Technology. CRC Press, Boca Raton (2002)
30. Nalwa, H.S. (ed.): Nanostructured Materials and Nanotechnology. Academic, San Diego (2002)
31. Poole, C.P., Owens, F.J.: Introduction to Nanotechnology. Wiley, New York (2003)
32. Rietman, E.A.: Molecular Engineering of Nanosystems. Springer, New York (2001)
33. Cheng, J., Kricka, L.J. (eds.): Biochip Technology. Harwood Academic, Philadephia (2001)
34. Heller, M.J., Guttman, A. (eds.): Integrated Microfabricated Biodevices. Marcel Dekker, New York (2001)
35. Lai Poh San, C., Yap, E.P.H. (eds.): Frontiers in Human Genetics. World Scientific, Singapore (2001)
36. Manz, A., Becker, H. (eds.): Microsystem Technology in Chemistry and Life Sciences. Topics in Current Chemistry, vol. 194. Springer, Heidelberg (1998)
37. Mastrangelo, C.H., Becker, H. (ed.): Microfluidics and BioMEMS. Proceedings of SPIE, vol. 4560. SPIE, Bellingham (2001)
38. Becker, H., Locascio, L.E.: Polymer microfluidic devices. Talanta **56**, 267–287 (2002)
39. van der Berg, A. (ed.): Lab-on-a-Chip: Chemistry in Miniaturized Synthesis and Analysis Systems. Elsevier, Amsterdam (2003)
40. Lanza, R.P., Langer, R., Vacanti, J. (eds.): Principles of Tissue Engineering, 2nd edn. Academic, San Diego (2000)

41. Oeberg, P.A., Togawa, T., Spelman, F.A.: Sensors in Medicine and Health Care. Wiley, New York (2004)
42. Park, K. (ed.): Controlled Drug Delivery: Challenges and Strategies. American Chemical Society, Washington, DC (1997)
43. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
44. Bhushan, B.: Handbook of Micro/Nanotribology. CRC Press, Boca Raton (1999)
45. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
46. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)

Nanotechnology Applications in Polymerase Chain Reaction (PCR)

Kuo-Sheng Ma[1], Yingnan Ma[2] and Fred Chiou[3]
[1]Kennesaw State University, Marietta, GA, USA
[2]Electrical Mechatronics Engineering, Southern Polytechnic State University, Marietta, GA, USA
[3]Department of Engineering Technology, Weber State University, Ogden, UT, USA

Synonyms

Nucleic acid amplification; Oligonucleotide amplification

Definition

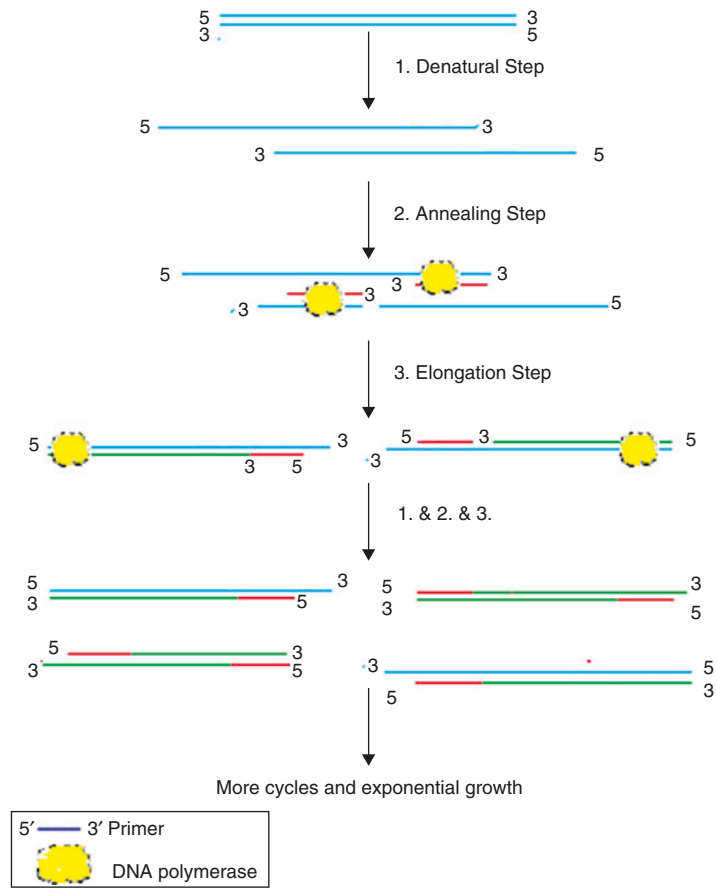
Polymerase chain reaction (PCR) is the most prevalent technology used in modern molecular biology research. Generally, it is a technique to amplify the amount of a piece of DNA with a specific sequence. Although the PCR technique is mature, improvement of its efficiency is still an emerging area of research. Recently, nanotechnology is getting more attention in this application. Several nanometer-sized materials such as carbon nanotubes, gold nanoparticles, quantum dots, and metal oxide nanoparticles have been employed.

Introduction

The polymerase chain reaction (PCR) technique is a series of chemical processes to exponentially amplify the number of a specific DNA sequence, producing hundreds of thousands of copies of DNA from a low original sample concentration. This revolutionary technique was first developed by Dr. Kary Mullis in 1983 [1]. Since then, it has been broadly utilized in medical, biomedical, and biological research fields for a variety of applications [2, 3]. In year 2013, the market of global polymerase chain reaction (PCR) with related products had reached US $6 billion with 6–10 % compound annual growth rate (CAGR) [4]. The growth include DNA sequencing, DNA cloning, identification of functional genes, the disease diagnosis, and the identification of genetic fingerprints associated with the PCR technique.

A basic PCR reaction needs several components in the mixture reagent. These include:

(a) A small amount of original DNA (mRNA in some applications) as a template which contains the oligonucleotide segment to be amplified.
(b) Primers. These are short sequences of oligonucleotides in which their sequences are designed complementary to the specific site of the template DNA. Two primers are used for one amplification process and each primer complementary binds to one strand of the paired DNA template at the opposite ends of the segment.
(c) DNA polymerases. These are the enzymes that perform DNA synthesis in the amplification process. The Taq polymerase is most prevalently used in the conventional PCR process and other polymerases are also used on different applications.
(d) Deoxynucleotide triphosphates (dNTPs). These are the building blocks of DNAs which are polymerized by the enzyme (polymerase) complementary to the sequence of DNA templates.
(e) Buffer solution. This is the salt solution to provide an optimum condition for DNA synthesis. Divalent cations such as Mg2+ or Mn2+ are normally used in the buffer solution



Nanotechnology Applications in Polymerase Chain Reaction (PCR), Fig. 1 Schematic illustration of the PCR amplification processes

to mediate DNA mutagenesis or decrease the error rate of the DNA polymerase.

The technique is based on the cycles of repeated temperature changes of the reactant in reaction tubes or plates. The thermal cycling is normally carried out in a thermal cycler which performs heating and cooling on the reactants. A typical PCR process consists of 20–40 cycles, with each cycle consisting of three basic heating and cooling steps on discrete temperatures. The temperatures employed and the duration of specific temperature in a cycle depend on the reactant used in the mixture for each specific process. In a conventional PCR process, as the first step (or called as denature step), the mixture is heated up to 94–98 °C for 20–30 s to break hydrogen bonds between complementary double-stranded DNA (templates) and yield single-stranded DNA molecules. In the second step (or called as annealing step), the temperature of the solution is lowered to 50–65 °C for 20–40 s allowing the primers to bind to the separated strands of DNA template. Sequentially, the enzyme (DNA polymerase) binds to the hybridization sites of primer and template to initiate DNA amplifications. In the next step (also called the extension/elongation step), the temperature of solution is risen to a temperature at which the enzyme has its optimum activity. A temperature of 72 °C is used for the commonly used Taq polymerase. At this step, the amount of DNA is doubled due to the enzyme synthesizing a new complementary strand to each DNA template by polymerizing dNTPs in 5′–3′ direction. These three basic steps form a cycle, and repetitions of this cycle lead to exponential amplification of the amount of DNA fragments. Figure 1 schematically shows a conventional PCR process.

Variations on the Basic PCR Technique

PCR is such a powerful technique to provide a relatively easy, efficient method to amplify the amount of desired DNA sequence. Since it was invented, many modifications to tailor the basic PCR thermal cycle approach to fit specific applications have been made. The features of these modified techniques are summarized in Table 1.

Nanotechnology in PCR

Although the PCR technique is mature and has been broadly employed in various applications, there is still an effort to improve the specificity and efficiency of it by using different advanced technologies. In the nanotechnology, the development and discovery of more applications of nanoparticles is an emerging area of research. Nanoparticle-assisted PCR has attracted more attentions, and many nanomaterials have been studied for their potential in developing better PCR technique. Various types of nanoparticles such as carbon nanotubes, metal nanoparticles (especially gold nanoparticles), semiconductor quantum dots, and metal oxide nanoparticles have been introduced into PCR, and their effect on reaction efficiency is studied. Some studies showed these nanoparticles dramatically improve PCR in efficiency and specificity, but some studies showed contrary results. In this entry, the development of the technique which employed the nanoparticles in PCR would be introduced. The different types of nanoparticles utilized in the studies and their effect in assisting PCR as well as mechanisms will be discussed.

Carbon Nanotubes

Gao et al. [14] investigated the effects of single-walled carbon nanotubes (SWCNTs) on PCR. Their results showed the amount of PCR product increased by adding single-walled carbon nanotubes into the PCR reactant mixture. They further discovered that adding higher concentrations of SWCNTs (>3 ug/ul) reversed the effect (amount of PCR product decreased). Using SEM and HRTEM, they showed that DNA templates and Taq enzymes were attached to bundles of SWCNTs. XPS results further suggested that there might be a chemical reaction between SWCNTs and components in PCR solution. The authors concluded that the SWCNTs have the potential to act as catalysts in the solution, which increased the yield of PCR products.

Zhang et al. [15] had done the similar study and reported the beneficial effect of using single-walled carbon nanotubes (SWCNTs) and multiwalled carbon nanotubes (MWCNTs) to enhance the amplification efficiency of long PCR up to 14 kb. Rather than the pristine carbon nanotubes, the authors employed the CNTs with surface functionalization. Both hydroxylic and carboxylic functionalized CNTs had been studied and had similar enhancing effects of long PCR.

Shi et al. [16] demonstrated using polyethyleneimine (PEI)-modified multiwalled carbon nanotubes (MWCNTs) to improve the efficiency of PCR. The authors tried three different polarities of PEI on MWCNTs and examined their influences on PCR. They discovered that only the MWCNTs with positive and negative PEI have effects on PCR efficiency but not the neutral PEI sample. In addition, positive PEI-coated MWCNTs had stronger effects than the negative PEI-coated MWCNTs. The authors speculated the effects might be from thermal conductivity of MWCNT and electrostatic interactions between PCR components and positive-charged molecules on MWCNTs.

The opposite effect on CNTs enhancing PCR had also been reported. Yang et al. [17] studied the effects of multiwalled carbon nanotubes and single-walled carbon nanotubes functionalized with and without carboxylic groups on PCR. Their results showed that CNTs reduced the efficiency of PCR. From their report, the inhibition effect on PCR was increased in the order of CNT-COOH > pristine CNT and SWCNT > MWCNT. The authors attributed the inhibition effect of nanomaterials on PCR to the decrease of the DNA polymerase activity, which they believed to be a significant effect of how nanomaterials impact biological systems.

Nanotechnology Applications in Polymerase Chain Reaction (PCR), Table 1 Variation techniques based on PCR process

Name	Functions	References
1. Quantitative PCR (qPCR)	A technique used to quantify the PCR products during the amplification cycles. Because the measurement is in "real time" of amplification, it is also called quantitative real-time PCR (qRT-PCR). The technique employed the fluorescent dyes in the PCR solution to measure the amount of amplified products in the log phase of the reaction before the plateau. The commonly used dyes are SYBR Green or TaqMan which is a DNA probe containing a fluorophore	[5]
2. Reverse transcription PCR (RT-PCR)	A preferred technique used to detect and quantify mRNA. In mRNA detection applications, the target RNA needs to be amplified by PCR. Thus, RNA samples need to be first reverse transcribed into cDNA by reverse transcriptase, and cDNA sequences will be amplified by PCR. This technique is widely used in profiling of gene expressions	[6]
3. Polymerase cycling assembly (PCA)	A technique to synthesize long DNA sequences by a PCR. It employed a pool of oligonucleotides with designed short overlapping sequences in each piece of oligonucleotide. Rather than using two primers, this technique uses multiple short oligonucleotides which are part of the single strand of the target sequences. During the elongation cycle, the rest of sequences will be filled in by polymerase. The technique is popular for the production of synthetic genes and even synthetic genomes	[7]
4. Helicase-dependent amplification	This technique is similar to traditional PCR but uses a constant temperature instead of temperature cycling in denaturation and annealing steps. Rather than thermal denaturation, a DNA helicase is used to separate the double strand of DNA which provides single-stranded DNA template for primers and polymerase binding	[8]
5. Hot start PCR	A modified process to reduce nonspecific amplification of DNA by inactivating the Taq polymerase at a lower temperature in the initial stages of the conventional PCR. It can be performed manually by adding a primer to start amplification after heating the reaction components to the melting temperature. Common methods for this technique include chemical modifications, wax-barrier methods, inhibition by a Taq-directed antibody, and by the presence of covalently bound inhibitors. Due to lack of nonspecific hybridization of primers to template, the amplified DNA bands tend to be cleaner	[9]
6. Methylation-specific PCR (MSP)	A technique is used to detect methylation of CpG islands in the genomic DNA. DNA is first treated with sodium bisulfite, which converts unmethylated cytosine bases to uracil. A special primer is designed to distinguish methylated from unmethylated DNA. The process takes advantage of the sequence differences resulting from bisulfite modification. MSP using qPCR can also be performed to obtain quantitative rather than qualitative information about methylation	[10]
7. Allele-specific PCR	It is a variation of the PCR which is used as a diagnostic or cloning technique to identify single-nucleotide polymorphism (SNPs) (single-base differences in DNA). It requires prior knowledge of a DNA sequence and differences between alleles. The allele-specific PCR uses primers whose 3′ ends encompass the SNP	[11]
8. Solid-phase PCR	In a solid-phase PCR (SP-PCR), the amplification takes place directly on a solid substrate. The technique encompasses multiple modifications, including polony amplification where PCR colonies are derived in a gel matrix, bridge PCR in which primers are covalently linked to a solid-support surface, and conventional solid-phase PCR in which an asymmetric PCR is employed. A primer is immobilized on a solid support which has the sequence matching one of the aqueous primers	[12, 13]

Gold Nanoparticles

Nanometer-sized gold particles (Au-NPs) and its counterparts with different surface modifications had also been studied of their influences on PCR. Hu et al. [18] reported that nanogold significantly improved the specificity and yield of PCR. The author attributed the enhancement to the greater affinity of Au-NPs to single-stranded DNA (ssDNA) than to double-stranded DNA (dsDNA) which reduced the mispairing of nucleotides to the template sequence in the elongation step. In their other work [19], the authors did the multiple rounds of PCR with 10 nm Au-NPs. The result showed that the Au-NPs effectively enhanced the specificity and yield of error-prone amplification. As the result showed, after six rounds of amplification, the target band was still manifested by a single bright band in gel electrophoresis. However, the target band disappeared after four rounds of amplification of the sample without adding Au-NPs. This study has shown that with Au-NP enhancement, a large amount of target DNA can be produced even with very low copies of original DNA samples. Liu et al. [20] had done the similar study using 13 nm Au-NPs in the PCR reagent and discovered the efficiency was increased. In this study, the authors also compared different PCR systems by using different DNA polymerases, various DNA sizes for amplification, and complex samples. Their results demonstrated that Au-NPs increased the efficiency from fivefold to tenfold in a conventional PCR and by 104-fold in a real-time PCR. The authors hypothesized that the enhancement is due to the excellent heat transfer property of the Au-NPs in PCR solution.

In addition to the bare Au-NPs, the surface-modified Au-NPs had also been studied with their influences on PCR. He et al. [21] reported using poly(diallyldimethylammonium) chloride (PDDA) molecule on Au-NPs to improve the efficiency and specificity of PCR. The authors used positive-charged PDDA, neutral PDDA, and negative-charged PDDA on Au-NP and discovered all these three modifications will enhance the efficiency of the PCR. In the mechanism study, the authors discovered that the positive-charged PDDA-modified Au-NP had the highest efficiency on PCR. It was speculated that the surface modifications with PDDA molecule would cause interactions between Au-NPs and components in PCR, which enhanced the specificity of molecule recognitions and further increased the efficiency. In addition, positive PDDA molecules caused stronger electrostatic attractions on negative-charged PCR components and thus induced higher efficiency enhancement than the other counterparts did. Shi et al. [22, 23] also reported the similar discovery of enhancement of PCR efficiency with surface-modified Au-NPs. In their study, the authors used poly(amidoamine) (PAMAM) dendrimer-grafted Au-NPs as the efficiency enhancer on PCR. The authors drew the same conclusion as He's group that the modified molecules on Au-NPs induced more interactions between PCR components and as the matter of fact enhancing the efficiency. The authors also provided a sophisticated model that PCR efficiency enhancement was mainly beneficially by the good thermal conductivity of Au-NPs, which enhanced local heat transfer rate and reduced mispairing between primers and templates. Further study will be conducted by the authors.

Although Au-NP-enhanced PCR had been investigated, different results had been reported. Emslie et al. [24] reported the contrary results from their study. The authors investigated the efficiency of real-time quantitative polymerase chain reaction (qPCR) by adding Au-NPs into the PCR solution. It was found that Au-NPs affect the qPCR amplification due to fluorescence quenching by Au-NPs. Rather than the quenching effect, it was discovered that Au-NPs destabilized the double-stranded PCR products which further reduced the efficiency of qPCR. In contrast to those previous reports, Willson et al. [25] discovered that Au-NPs suppressed the amplification of longer products and rather enhanced the amplification of shorter products. In their discovery, when Au-NPs were added in the PCR solution, some of the DNA polymerases were nonspecifically absorbed on nanoparticles and thus reduced the amount of polymerase and the efficiency of amplification. In their recent report, Hu et al. [26] also discovered that the Au-NPs can

enhance the PCR efficiency only by employing native Taq polymerase. No improvement was found in a PCR based on the recombinant Taq polymerase. However, the underlying mechanism was still unclear.

Yeow et al. [27] had done the study of the proposed mechanism for Au-NPs affecting the PCR. The authors studied the PCR efficiency with different Au-NP sizes (5 nm, 10 nm, and 20 nm) and different concentrations. Evidence from their study showed that the possible mechanism for Au-NPs affecting the PCR is through DNA polymerase-Au-NP binding which modulated the enzyme activity for the amplification. Besides, the binding strongly related to the total surface area of the nanoparticles in the reaction solution.

Semiconductor Quantum Dots

Another type of nanoparticles such as semiconductor quantum dots (QDs) has been studied in affecting the PCR efficiency. Li et al. [28] reported that the QDs dramatically improved the yield and specificity of PCR. In their study, by employing mercaptoacetic acid (MAA)-coated QDs, yield and specificity of PCR enhanced even at lower annealing temperature. Rather than MAA surface-coated QDs, streptavidin surface-modified QDs also effectively improved the specificity of PCR which demonstrated that the effect was not due to the molecules of surface modification but instead due to the QDs itself. Further study is needed to discover the underlying mechanism. Another study using QDs in PCR was done by Xu et al. [29]. In this study, the authors also discovered that the QDs could increase the specificity of the PCR at different annealing temperatures. Thus, as the authors concluded, this method could make the PCR widely applicable, especially in the multi-round PCR with different annealing temperatures.

Guan et al. [30] discovered different effects of QDs using different polymerase systems. The PCR efficiency was reduced by introducing high-concentration QDs in reagents, but this effect was reversed in Ex PCR systems using Ex Taq polymerase. The mechanism was unknown. The same group also discovered the QDs were able to enhance fluorescence signals without interrupting the DNA amplification in real-time PCR. Zheng et al. [31, 32] studied the influences of the water-soluble CdTe quantum dots in the hot start (HS) PCR reaction. Their results showed the higher amplicon yields and reduced nonspecific amplification on hot start PCR with premix PCR reagents and QDs. The study also discovered the effects are reversible by increasing the polymerase concentration in the PCR reagent, which demonstrated the interactions between QDs and the polymerase. Further investigation of the working mechanism was conducted.

Other Nanoparticles

Several different types of nanoparticles have also been studied of their effects in PCR. Mahapatra et al. [33] recently reported that the TiO_2 nanoparticles with the size of 25 nm diameter significantly enhanced the PCR efficiency even with various types of templates (i.e., plasmid DNA, genomic DNA, and complementary DNA). Besides, the yield was also enhanced by utilizing TiO_2 nanoparticles in the PCR. Investigations of the mechanism by simulations showed evidences of fast and efficient heat transfer in the PCR solution in the presence of TiO_2 nanoparticles to enhance the PCR efficiency. Consistent with the simulation results, experimentally, the authors observed the increase of denaturation of genomic DNA in the presence of TiO_2 nanoparticles in PCR solution which indicated higher thermal conductivity through the reaction buffer. This technique may be useful for reduction of the overall PCR reaction time and further enhanced the efficiency and specificity of PCR in a variety of samples.

The platinum nanoparticles have also been evaluated to optimize DNA amplification on PCR. Ventimiglia et al. [34] demonstrated that the (beta-cyclodextrins) capped platinum NPs were able to increase the PCR efficiency. The authors discovered the enhancement was directly caused by the metal thermal efficiency and was

not influenced by the properties of capping agents. In addition, the reduction of the reagent's concentration was another benefit by applying metal NPs on PCR.

The effect of C-60 on PCR was studied by Jiang et al. [35] using a qPCR system. The authors reported a dramatical enhancement of the qPCR at the beginning of the exponential phase but a significant inhibition at the plateau region. In addition, the resulting melting curve demonstrated that C-60 decreased the melting temperatures (Tm) of the template DNA fragments. The authors hypothesized this is due to the enhancement of qPCR efficiency by C-60 in the initial phase of PCR. However, when there is more DNA amplicon in solution, C-60 can disrupt DNA replication by binding to DNA to decrease the access of templates and changing the conformation of DNA templates which reduced the amplification efficiency at the later cycles of the reaction.

Conclusion

Although the utilization of nanoparticles in conventional PCR is just emerging, the potentials to enhance the process have been demonstrated by some of the works. Due to its excellent thermal conductivity, high surface-to-volume ratio, and high affinity to bind molecules on surface, nanoparticle-assisted PCR can greatly reduce the total process time and specificity of the process. This is critical for the technique to be used in medical diagnosis and infield detection applications. In addition, due to the high efficiency of process with NP enhancement, less amount of expensive reagents is needed which significantly reduce the experimental costs. However, until now, the NP-assisted technology has not yet been revolutionized in practice. To enable users to embrace the technique, future development in identifying type of NPs, size of NPs, concentration of NPs in PCR reagent, and specific PCR components are all essential. In addition, optimization of process design to maximize the enhancements and further reduce the time and cost of the process is also important to evolve this technology out of research and to be used routinely.

Cross-References

- ▶ Biosensing
- ▶ Carbon Nanotubes
- ▶ Double-Walled Carbon Nanotubes (DWCNTs)
- ▶ Gold Nanorods
- ▶ Nanomaterials
- ▶ Nanoparticle
- ▶ Nanoscale Particle
- ▶ Nanowires
- ▶ Nucleic Acid Amplification
- ▶ Quantum Dot
- ▶ Semiconductor Nanocrystals

References

1. Mullis, K., et al.: Specific enzymatic amplification of DNA in vitro – the polymerase chain-reaction. Cold Spring Harb. Symp. Quant. Biol. **51**, 263–273 (1986)
2. Saiki, R.K., et al.: Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle-cell anemia. Science **230**(4732), 1350–1354 (1985)
3. Saiki, R.K., et al.: Primer-directed enzymatic amplification of DNA with a thermostable DNA-polymerase. Science **239**(4839), 487–491 (1988)
4. http://www.transparencymarketresearch.com/press release/pcr-technologies.htm
5. Kellogg, D.E., Sninsky, J.J., Kwok, S.: Quantitation of HIV-1 proviral DNA relative to cellular DNA by the polymerase chain-reaction. Anal. Biochem. **189**(2), 202–208 (1990)
6. Mackay, I.M., Arden, K.E., Nitsche, A.: Real-time PCR in virology. Nucleic Acids Res. **30**(6), 1292–1305 (2002)
7. Stemmer, W.P.C., et al.: Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. Gene **164**(1), 49–53 (1995)
8. Vincent, M., Xu, Y., Kong, H.M.: Helicase-dependent isothermal DNA amplification. EMBO Rep. **5**(8), 795–800 (2004)
9. Chou, Q., et al.: Prevention of pre-PCR mis-priming and primer dimerization improves low-copy-number amplifications. Nucleic Acids Res. **20**(7), 1717–1723 (1992)
10. Herman, J.G., et al.: Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands. Proc. Natl. Acad. Sci. U. S. A. **93**(18), 9821–9826 (1996)
11. Newton, C.R., et al.: Analysis of any point mutation in DNA – the amplification refractory mutation system (arms). Nucleic Acids Res. **17**(7), 2503–2516 (1989)
12. Bing, D.H., Sawosik, T.M., Word, C.J.: Assay performance results with the AmpliType(R) PM PCR

amplification and typing kit. Crime Lab. Dig. **23**(2), 27–45 (1996)
13. Khan, Z., Poetter, K., Park, D.J.: Enhanced solid phase PCR: mechanisms to increase priming by solid support primers. Anal. Biochem. **375**(2), 391–393 (2008)
14. Cui, D.X., et al.: Effects of single-walled carbon nanotubes on the polymerase chain reaction. Nanotechnology **15**(1), 154–157 (2004)
15. Zhang, Z.Z., et al.: Aqueous suspension of carbon nanotubes enhances the specificity of long PCR. Biotechniques **44**(4), 537–545 (2008)
16. Shi, X.Y., et al.: Effect of surface charge of polyethyleneimine-modified multiwalled carbon nanotubes on the improvement of polymerase chain reaction. Nanoscale **3**(4), 1741–1747 (2011)
17. Yi, C.Q., et al.: Interactions between carbon nanotubes and DNA polymerase and restriction endonucleases. Nanotechnology **18**(2), 6 (2007)
18. Li, H.K., et al.: Nanoparticle PCR: nanogold-assisted PCR with enhanced specificity. Angew. Chem. Int. Ed. **44**(32), 5100–5103 (2005)
19. Pan, J.K., et al.: Nanogold-assisted multi-round polymerase to chain reaction (PCR). J. Nanosci. Nanotechnol. **7**(12), 4428–4433 (2007)
20. Li, M., et al.: Enhancing the efficiency of a PCR using gold nanoparticles. Nucleic Acids Res. **33**(21), 10 (2005)
21. Yuan, L.F., He, Y.J.: Effect of surface charge of PDDA-protected gold nanoparticles on the specificity and efficiency of DNA polymerase chain reaction. Analyst **138**(2), 539–545 (2013)
22. Shi, X.Y., et al.: Effect of the surface functional groups of dendrimer-entrapped gold nanoparticles on the improvement of PCR. Electrophoresis **33**(16), 2598–2603 (2012)
23. Shi, X.Y., et al.: A highly effective polymerase chain reaction enhancer based on dendrimer-entrapped gold nanoparticles. Analyst **137**(1), 223–228 (2012)
24. Haber, A.L., et al.: Addition of gold nanoparticles to real-time PCR: effect on PCR profile and SYBR Green I fluorescence. Anal. Bioanal. Chem. **392**(5), 887–896 (2008)
25. Vu, B.V., Litvinov, D., Willson, R.C.: Gold nanoparticle effects in polymerase chain reaction: favoring of smaller products by polymerase adsorption. Anal. Chem. **80**(14), 5462–5467 (2008)
26. Yang, W.C., et al.: Evaluation of gold nanoparticles as the additive in real-time polymerase chain reaction with SYBR green I dye. Nanotechnology **19**(25), 9 (2008)
27. Wan, W.J., Yeow, J.T.W.: The effects of gold nanoparticles with different sizes on polymerase chain reaction efficiency. Nanotechnology **20**(32), 5 (2009)
28. Ma, L., et al.: Maximizing specificity and yield of PCR by the quantum dot itself rather than property of the quantum dot surface. Biochimie **91**(8), 969–973 (2009)
29. Wang, L.B., et al.: Effects of quantum dots in polymerase chain reaction. J. Phys. Chem. B **113**(21), 7637–7641 (2009)
30. Xun, Z., Zhao, X.Y., Guan, Y.F.: Improved thermal cycling durability and PCR compatibility of polymer coated quantum dot. Nanotechnology **24**(35), 355504 (2013)
31. Zhang, Z.Z., et al.: Quantum dots trigger hot-start effects for pfu-based polymerase chain reaction. J. Exp. Nanosci. **9**(10), 1051–1063 (2014)
32. Zhang, Z.Z., et al.: A hot start alternative for high-fidelity DNA polymerase amplification mediated by quantum dots. Acta Biochim. Biophys. Sin. **46**(6), 502–511 (2014)
33. Khaliq, A., et al.: Enhancement in the efficiency of polymerase chain reaction by TiO_2 nanoparticles: crucial role of enhanced thermal conductivity. Nanotechnology **21**(25), 11 (2010)
34. Ventimiglia, G., Petralia, S., Barbuzzi, T.: Polymerase chain reaction efficiency improved by water soluble beta-cyclodextrins capped platinum nanoparticles. Mater. Sci. Eng. C Mater. Biol. Appl. **32**(4), 848–850 (2012)
35. Liang, Y., et al.: C-60 affects DNA replication in vitro by decreasing the melting temperature of DNA templates. Carbon **47**(6), 1457–1465 (2009)

Nanotechnology in Cardiovascular Diseases

Lu Dai[1], Yongfen Qi[1], Lixin Jia[1], Wei Yu[1] and Jie Du[2]
[1]The Key Laboratory of Remodeling-related Cardiovascular Diseases, Ministry of Education, Capital Medical University, Beijing, China
[2]Beijing Institute of Heart Lung and Blood Vessel Diseases, Beijing Anzhen Hospital, Beijing, China

Definition

Nanomolecular imaging is a new type of imaging with a combination of nanomaterial and molecular biology, using nanoparticles to visualize, characterize, and measure biological processes at molecular and cellular levels in living systems, including radiotracer imaging/nuclear medicine, magnetic resonance (MR) imaging, MR spectroscopy, optical imaging, ultrasound, and other nanomolecular imaging.

Overview

Recently, the emergence of nanomolecular imaging has set the stage for an evolutionary leap in the diagnosis and therapeutics of diseases such as cancer and neurological and cardiovascular diseases (CVDs) [1–3]. CVD is the leading cause of morbidity and mortality in the world. The pathology underpinning CVD is chronic inflammation involving the injury of arterial wall, the infiltration of leukocytes, and the remodeling of vasculature. The main benefits of nanomolecular imaging are derived from the combination of cell type-specific molecular-based, molecular-molecular recognition with nanoparticle-based probes, leading to accurate and high resolution of diseased sites and to the early diagnosis and treatment for CVDs, including atherosclerosis, thrombosis, and myocardial infarction [4–6]. The progress in nanomolecular imaging will play an important role not only in molecular imaging, but also in the development of biosensors, biomarkers, and new drug delivery system.

Nanoparticles are defined as ranging from 5 to 700 nm in diameter and can enter the cells and function inside the cells (such as generation of intracellular imaging and targeting-specific molecular pathways) [4] (see Nanoparticles). For example, new types of nanoparticles have been recently introduced to generate contrast enhancement and improved sensitivity, such as infrared dyes and quantum dot (QD) nanocrystals for optical imaging [7], gold colloids or radiotracer-based nanoassemblies for X-ray contrast [4], gas-filled microbubbles for contrast using ultrasound [8], and nanoparticles of gadolinium or iron oxide for magnetic resonance contrast [4]. These abilities open up a large number of exciting possibilities for early detection, the potential treatment of disease, and a novel type of drug delivery system. Nanomolecular imaging has also been applied to imaging techniques including luminescence imaging and spectroscopy. Indeed, current imaging device can be applied for nanomolecular imaging, such as positron emission tomography (quantitative PET) and magnetic resonance spectroscopy (MRS), magnetic resonance spectroscopic imaging (MRSI), diffusion magnetic resonance imaging (dMRI), and functional magnetic resonance imaging (fMRI).

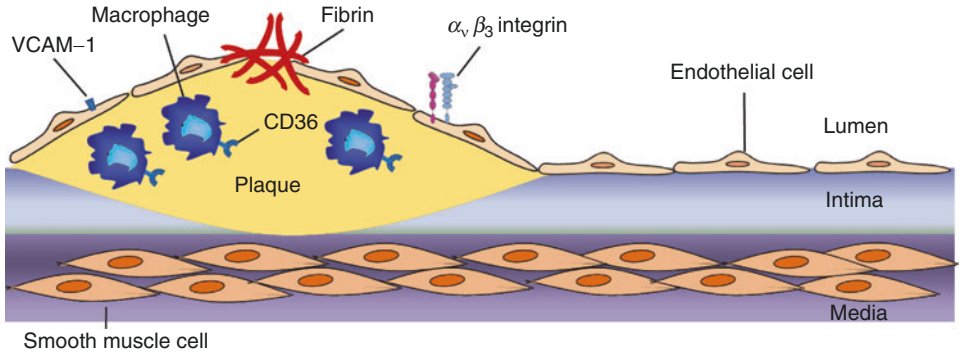
Key Research Findings

Nanomolecular Imaging and Potential Therapy in Cardiovascular Disease

CVD is the largest cause of mortality in the world, accounting for about 40 % of deaths. CVD occurs at a higher rate in diabetes with mortality exceeding 50 %. Risk factors, such as visceral obesity, hypertension, and dyslipidemia, are all associated with increased risk of CVD. The major underlying disease of most CVDs is atherosclerosis, characterized by a complex multifactorial pathophysiology such as the slow accumulation of lipid and inflammatory products in the vessel wall that further leads to calcification and fibrosis in atherosclerotic plaques. Inflammation in the vessel wall is now considered to play an essential role in the initiation, progression, and the final steps of atherosclerosis, namely, plaque destabilization and eventually plaque rupture (Fig. 1) [9]. Atherosclerotic plaques can be divided into two types: stable plaque and unstable plaque (a vulnerable plaque, causing myocardial infarction and stroke). Unstable plaques are at higher risk for plaque disruption, characterized by fibrous cap rupture, ulceration following spontaneous denudation of the endothelium, or intraplaque hemorrhage. This plaque disruption leads to local thrombus formation and vessel occlusion, recognized as the proximate cause of the majority of clinical CVD events such as a heart attack or stroke. Although there have been significant advances in vascular imaging (such as intravascular ultrasound and MRI), current imaging methods are still insufficiently capable for elucidating plaque composition. Therefore, it is critical to develop novel techniques to identify vulnerable plaques. In recent years, nanomolecular imaging provides exciting opportunities to identify vulnerable plaque to reduce the risk of acute coronary syndromes.

Endothelial Cells

The initiating event for the development of atherosclerosis is the endothelial layer inflammation



Nanotechnology in Cardiovascular Diseases, Fig. 1 The structure of plaque in atherosclerosis. *VCAM-1* vascular cell adhesion molecule-1

or dysfunction. Increased adhesiveness of the endothelium leads to the infiltration of leukocytes which could engulf oxidized lipid particles and develop into inflammatory foam cells [10]. These processes can be determined by measuring inflammatory markers, such as vascular cell adhesion molecule-1 (VCAM-1), a molecule found in blood vessel walls at the onset of inflammation – a first stage in the development of atherosclerotic plaques. Kelly et al. identified VCAM-1 internalizing peptide-28 nanoparticles (VINP-28), a novel VCAM-1 peptide affinity ligand from phage display that undergoes internalization and trapping in cells that express VCAM-1, resulting in higher signal amplification ion [11]. That offers an opportunity for detecting early atheromata (fatty streak-like lesions) and targeting VCAM-1-expressing cells of human carotid atherosclerotic plaques.

Macrophages

Macrophages have emerged as a key component of atherosclerosis, which accumulate cholesterol esters to form lipid-laden foam cells in the arterial intima and generate proteinases that lead to plaque destabilization and rupture. The accumulation of macrophage in the subendothelial space could be observed through all stages of vascular lesion development, and the increased macrophage density is associated with lesion progression and rupture [9]. Additionally, the macrophage may serve as an ideal target for the detection and potential therapy of vulnerable inflamed lesions. Recently, macrophage-specific lipid-based nanoparticles were created which improved cardiac magnetic resonance detection and characterization of human atherosclerosis [12]. These nanoparticles were made from phospholipids, a surfactant (Tween 80), and an aliphatic gadolinium complex, which targets CD36, a macrophage class B scavenger receptor. The macrophage-specific (CD36) nanoparticles bind human macrophages and improve cardiac magnetic resonance (CMR) detection. A light-activated theranostic nanoagent for targeted macrophage ablation was invented [13]. This nanoparticle was based upon cross-linked dextran-coated iron oxide (CLIO). It was modified with near-infrared fluorophores and light-activated therapeutic moieties. The therapeutic component was activated at 650 nm and could result in eradication of inflammatory macrophages. Moreover, other potential targets, such as scavenger receptor class B type 1 (SRB1), lectin-like oxidized LDL receptor-1 (LOX-1), receptor for advanced glycation end products (RAGE), and urokinase-type plasminogen activator (uPA) receptor, could be more efficacious and provide an integrated imaging and therapeutic nanoplatform for atherosclerosis [14–17].

Atherosclerotic Plaque

A major component of atherosclerosis plaques is fibrin, which forms the fibrils that entangle the blood platelets and seal up the rent in the artery wall. Fibrin deposition is one of the earliest signs of plaque rupture or erosion and intraplaque

hemorrhage. Densely packed fibrin-targeted nanoparticles can be detected by scanning electron micrographs (SEM) [18]. These nanoparticles were formulated with gadolinium DTPA-bis-oleate (Gd-DTPA-BOA) and presented a highly detectable, homogeneous T1-weighted contrast enhancement that improved with increasing gadolinium level. Higher-resolution scans and scanning electron microscopy revealed that the nanoparticles were present as a thin layer over the clot surface. Then, the development of other matrix-targeted nanoparticles could provide accurate detection and localization of matrix deposits and may allow early, direct identification of vulnerable plaques.

Angiogenesis

In atherosclerosis, angiogenesis is not only related to plaque growth, but also related to plaque instability and rupture. Integrins, cell surface receptors involved in cell-cell and cell-matrix interaction, such as $\alpha_v\beta_3$, are associated with angiogenesis and have been extensively used as a target for nanomolecular imaging [19]. Patrick Winter et al. developed a paramagnetic nanoparticle contrast agent targeted specifically to $\alpha_v\beta_3$ integrins to permit noninvasive molecular imaging of plaque-associated angiogenesis. Moreover, $\alpha_v\beta_3$ integrin-targeted paramagnetic nanoparticles can deliver fumagillin and elicit a marked antiangiogenic response with a minimal drug dosage [19].

Future Directions for Research

Nanomolecular imaging is a new evolving field, which represents the beginning of a revolutionary new age for the clinicians to enhance the diagnosis and treatment of the cardiovascular diseases relying on the powerful imaging probes and hardware. While it is still in the developing phase, the nanomolecular imaging has resulted in lots of nanoparticles that vary in size, shape, change, chemistry, coating, and solubility. More importantly, nanomolecular imaging must utilize the progress of molecular and cellular biology in pathological conditions. At the same time, the potential toxicity of these nanoparticles should also be paid more attention to [20] (see "▶ Cellular Mechanisms of Nanoparticle Toxicity" and "▶ Nanoparticle Cytotoxicity"). The nanoimaging agent synthesis requires the efforts on their subsequent translation to clinical practice. The rapid growth of nanomolecular imaging will change the field of imaging and imaging-guided interventions for cardiovascular diseases.

Cross-References

▶ Cellular Mechanisms of Nanoparticle Toxicity
▶ Nanoparticle Cytotoxicity
▶ Nanoparticles

References

1. Ferrari, M.: Cancer nanotechnology: opportunities and challenges. Nat. Rev. Cancer **5**, 161–171 (2005)
2. Bhaskar, S., et al.: Multifunctional nanocarriers for diagnostics, drug delivery and targeted treatment across blood-brain barrier: perspectives on tracking and neuroimaging. Part. Fibre Toxicol. **7**, 3 (2010)
3. Yang, X.: Nano- and microparticle-based imaging of cardiovascular interventions: overview. Radiology **243**, 340–347 (2007)
4. Wickline, S.A., Neubauer, A.M., Winter, P., Caruthers, S., Lanza, G.: Applications of nanotechnology to atherosclerosis, thrombosis, and vascular biology. Arterioscler. Thromb. Vasc. Biol. **26**, 435–441 (2006)
5. Godin, B., et al.: Emerging applications of nanomedicine for the diagnosis and treatment of cardiovascular diseases. Trends Pharmacol. Sci. **31**, 199–205 (2010)
6. Douma, K., et al.: Nanoparticles for optical molecular imaging of atherosclerosis. Small **5**, 544–557 (2009)
7. Stinaff, E.A., et al.: Optical signatures of coupled quantum dots. Science **311**, 636–639 (2006)
8. Anderson, D.R., Tsutsui, J.M., Xie, F., Radio, S.J., Porter, T.R.: The role of complement in the adherence of microbubbles to dysfunctional arterial endothelium and atherosclerotic plaque. Cardiovasc. Res. **73**, 597–606 (2007)
9. Tabas, I.: Macrophage death and defective inflammation resolution in atherosclerosis. Nat. Rev. Immunol. **10**, 36–46 (2010)
10. Sitia, S., et al.: From endothelial dysfunction to atherosclerosis. Autoimmun. Rev. **9**, 830–834 (2010)
11. Nahrendorf, M., et al.: Noninvasive vascular cell adhesion molecule-1 imaging identifies inflammatory activation of cells in atherosclerosis. Circulation **114**, 1504–1511 (2006)
12. Lipinski, M.J., et al.: Macrophage-specific lipid-based nanoparticles improve cardiac magnetic resonance

detection and characterization of human atherosclerosis. JACC Cardiovasc. Imaging **2**, 637–647 (2009)
13. McCarthy, J.R., Korngold, E., Weissleder, R., Jaffer, F. A.: A light-activated theranostic nanoagent for targeted macrophage ablation in inflammatory atherosclerosis. Small **6**, 2041–2049 (2010)
14. Stephen, S.L., et al.: Scavenger receptors and their potential as therapeutic targets in the treatment of cardiovascular disease. Int J Hypertens. **2010**, 646929 (2010)
15. Farmer, D.G., Kennedy, S.: RAGE, vascular tone and vascular disease. Pharmacol. Ther. **124**, 185–194 (2009)
16. Oka, H., et al.: Lysophosphatidylcholine induces urokinase-type plasminogen activator and its receptor in human macrophages partly through redox-sensitive pathway. Arterioscler. Thromb. Vasc. Biol. **20**, 244–250 (2000)
17. White, S.J., Sala-Newby, G.B., Newby, A.C.: Overexpression of scavenger receptor LOX-1 in endothelial cells promotes atherogenesis in the ApoE(−/−) mouse model. Cardiovasc. Pathol. **6**, 369–373 (2011)
18. Flacke, S., et al.: Novel MRI contrast agent for molecular imaging of fibrin: implications for detecting vulnerable plaques. Circulation **104**, 1280–1285 (2001)
19. Winter, P.M., et al.: Endothelial alpha(v)beta3 integrin-targeted fumagillin nanoparticles inhibit angiogenesis in atherosclerosis. Arterioscler. Thromb. Vasc. Biol. **26**, 2103–2109 (2006)
20. Stone, V., Donaldson, K.: Nanotoxicology: signs of stress. Nat. Nanotechnol. **1**, 23–24 (2006)

Nanotechnology in Dental Medicine

► Nanodentistry

Nanotechnology in Pulmonary Disease

James C. Bonner
Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA

Definition

Nanotechnology in pulmonary disease encompasses the potential human health risks from occupational, environmental, or consumer exposure to engineered nanomaterials as well as the therapeutic benefits of nanomaterials for drug delivery to treat pulmonary diseases, enhance medical imaging of lung diseases, or provide tissue engineering for regenerative medicine.

Introduction

Engineered nanomaterials (ENMs) are intentionally manufactured nanoparticles with at least one dimension <100 nm that can be mass produced with a high degree of conformity in size and shape. The risk of ENMs to human health, specifically pulmonary disease, is not yet known due to the relative infancy of the nanotechnology revolution. However, studies with rodents that show adverse pro-inflammatory, pro-fibrogenic, and carcinogenic effects suggest that at least some ENMs will pose a human health hazard. Also, the utility of ENMs for therapies aimed at the treatment of pulmonary diseases in an early stage of development. Naturally occurring nanoparticles (e.g., nanoscale volcanic dust) or nanoparticles from anthropogenic sources but not intentionally manufactured (e.g., diesel exhaust ultrafine particles) are not ENMs as defined in the context of this essay and are therefore not discussed.

Engineered Nanomaterials Causing Pulmonary Disease

Potential adverse effects of ENMs on the respiratory tract and lungs may occur in occupational settings from airborne nanoparticles released during the manufacturing process, during exposure of nanoparticles released from consumer products containing ENMs, or from exposure to ENMs released into the environment. The most likely route of exposure to ENMs resulting in pulmonary disease is inhalation, but other routes of exposure (dermal, ingestion) should not be excluded.

Pulmonary Fibrosis

Pulmonary fibrosis is a fatal disease characterized by the progressive accumulation of extracellular

matrix proteins produced by activated lung myofibroblasts [1]. A variety of environmental toxic agents such as asbestos and certain metals (Ni, Co) cause pulmonary fibrosis in humans. While pulmonary fibrosis has yet to be associated with ENMs, many ENMs have the potential to cause pulmonary fibrosis, including carbon-based ENMs (carbon nanotubes, carbon nanofibers, and graphene) and metal nanoparticles (Co, Ni, Fe). A rapidly growing body of evidence indicates that carbon nanotubes (CNTs), either single-walled CNTs (SWCNTs) or multiwalled CNTs (MWCNTs), cause pulmonary fibrosis in rodents when delivered to the lungs by inhalation exposure, intratracheal instillation, or pharyngeal aspiration [2]. The underlying mechanisms of fibrogenesis induced by CNTs and other pro-fibrogenic ENMs likely involve several factors. First, the generation of reactive oxygen species (ROS) is important to tissue injury and the activation of resident cells (e.g., macrophages and epithelial cells) during pulmonary fibrosis. Second, the activation of lung cells by ROS drives the production of pro-fibrogenic soluble mediators including platelet-derived growth factor (PDGF), osteopontin (OPN), and transforming growth factor (TGF)-β1 which then drive myofibroblast growth, migration, and collagen deposition, respectively [1]. Certain physiochemical features also determine the fibrogenic potential of CNTs. For example, like asbestos fibers, the length of CNTs has been proposed as an important determinant of fibrogenic potential, and rigid MWCNT tubes longer than 15 or 20 μm are difficult to clear from lung tissues via macrophage-mediated mechanisms [2]. However, long SWNCT can fold and might be cleared more effectively by macrophages. Second, the composition of CNTs is an important determinant of fibrogenicity. Metal catalysts (e.g., nickel, cobalt, iron) are used in the manufacture of CNTs, remain within the tube structure post-synthesis, and have the potential to generate reactive oxygen species (ROS).

A wide spectrum of metal nanoparticles (Fe, Co, Ni, Zn, TiO_2) pose a significant risk for pulmonary fibrosis, especially in occupational settings where relatively high concentrations of these ENMs might be inhaled during the manufacturing process. Also, as mentioned above, some metal nanoparticles (Ni, Co, Fe) are used in chemical vapor deposition reactions to grow CNTs. A large body of literature already exists on metal-induced pulmonary diseases, including pulmonary fibrosis, in humans and rodents [3]. For example, Co and Ni are known causes of "hard metal disease" and "metal fume fever" in workers exposed during welding and other industrial applications. It is also known that Fe mediates asbestos-induced pulmonary fibrosis (i.e., asbestosis). Therefore, it is not surprising that ENMs composed of these same metals would also have fibrogenic potential. In particular, nanometals could be more fibrogenic based on the fact that they have greater surface area per unit mass and therefore have the capacity to generate higher levels of ROS than larger metal particles.

Obstructive Airway Diseases: Asthma and COPD

Obstructive airway diseases include asthma and chronic obstructive pulmonary disease (COPD) [4]. Both diseases have genetic and environmental components. Asthma is an allergen-induced Th2 immune-mediated chronic airway disease characterized by eosinophilic inflammation, mucous cell metaplasia, smooth muscle thickening, and airway fibrosis. Chronic airway remodeling in asthma is accompanied by acute episodic bronchospasm in response to allergens or other environmental factors that narrows airways and limits pulmonary function. COPD, caused primarily by cigarette smoking or inhalation of smoke from burned biomass, is a Th1 immune-mediated lung disease that includes chronic bronchitis and emphysema (progressive degradation of the alveolar walls of the lung).

Some ENMs have been reported to either cause or exacerbate allergic airway disease in rodents, suggesting that they may pose a risk for asthma [5]. Altered airway flow in patients with asthma or COPD may increase airway deposition and

relative deposited concentration of inhaled ENMs as compared to healthy individuals. Also, the physicochemical characteristics of ENMs (e.g., aggregation status, surface charge) could be affected by the presence of mucus and cellular debris in diseased airways. Also, individuals with asthma or COPD have impaired mucociliary clearance and therefore might not be expected to remove inhaled ENMs as effectively as healthy individuals. The presence of activated inflammatory cells and structural airway cells may alter the response to ENMs. Several studies indicate that ENMs exacerbate or cause allergic airway inflammation and remodeling in the lungs of mice [6]. For example, airway fibrosis in mice is increased by the combination of ovalbumin allergen and MWCNTs, whereas ovalbumin or MWCNTs alone do not significantly increase airway fibrosis, suggesting that MWCNTs pose a hazard to individuals with allergic lung inflammatory diseases such as asthma. Intratracheally administered MWCNTs significantly increased ovalbumin-induced T-lymphocyte proliferation and amplified lung Th2 cytokines and chemokines compared to ovalbumin exposure alone. SWCNTs also exacerbate murine allergic airway inflammation via enhanced activation of Th2 immunity, increased oxidative stress, and inappropriate activation of antigen-presenting cells such as dendritic cells. MWCNTs delivered to the lungs of mice may also induce allergic responses through B-cell activation and production of IgE in the absence of any allergen preexposure.

In addition to CNTs, certain metal nanoparticles could pose a risk for the development of asthma or COPD. For example, Ni nanoparticles delivered to mice cause airway mucous cell metaplasia and mucus overproduction; this effect is greatly enhanced in mice lacking the transcription factor T-bet, a spontaneous mouse model of asthma [7]. While the potential human health impact of metal ENMs has yet to be established, the known adverse effects of a variety of metals on asthma suggest that some metal ENMs will either exacerbate or cause asthma.

Lung Cancer and Mesothelioma

Some ENMs show carcinogenicity in rodents and have the potential to cause either lung cancer or mesothelioma (cancer of the pleural lining surrounding the lungs) in humans. However, as with other pulmonary diseases, no ENM has yet to be linked to lung cancer or mesothelioma in humans. A number of studies have focused on the genotoxicity and potential carcinogenicity of CNTs and these have been reviewed [8, 9]. Both SWCNTs and MWCNTs cause genotoxic effects in certain in vitro cell culture systems. SWCNTs cause fragmented centrosomes, multiple mitotic spindle poles, anaphase bridges, and aneuploid chromosome number in cultured primary or immortalized human airway epithelial cell types. MWCNTs induce apoptosis in mouse embryonic stem cells and activate the tumor suppressor protein p53, a marker of DNA damage. CNTs also cause genotoxic effects in vivo. Long and thick MWCNTs cause more DNA damage in the lungs of mice compared to short and thin MWCNTs. Intraperitoneal injections of COOH-functionalized MWCNTs in mice cause structural chromosome aberrations, DNA damage, and micronuclei formulation compared to non-functionalized "pristine" MWCNTs. On the other hand, polyethylene glycol (PEG)-coated SWCNTs reduce genotoxicity in neuronal PC12 cells. Therefore, the degree of genotoxicity caused by modification of CNTs depends on the specific type of functionalization. Since genotoxicity plays a role in the development of cancer, genotoxicity screening-level assays have been developed to predict the potential carcinogenicity of a wide variety of ENMs.

Nanotechnology for Therapeutic Treatment of Pulmonary Diseases

While ENMs generated by nanotechnology will pose some risk to human health as described above through occupational or environmental exposures, other types of ENMs intentionally designed for nanomedicine platforms will offer

significant benefits for improving drug delivery and therapy of respiratory diseases. ENMs have been of great interest for some time as they can be designed to simultaneously carry a drug payload, specifically target features of diseased tissues, and carry an imaging molecule to track drug accumulation and clearance in tissues. Moreover, they can be engineered for more sustained drug delivery to improve pharmacokinetics and reduce the overall amount of drug administered during treatment of disease. A variety of ENMs have been investigated in animal models as tools to improve the therapeutic efficacy of drugs or genes delivered to the lung or other organ systems [10].

The nanotechnology platform for drug delivery contains a number of very different types of nanoscale structures with widely varying properties, including dendrimers, fullerenes, polymeric NPs, nanometals, and carbon nanotubes. Nanoparticles that have been investigated for drug delivery fall into two broad categories: organic "soft" nanoparticles (e.g., dendrimers and liposomes) and "hard" nanoparticles that are produced from carbon, silver, gold, or zinc in a variety of shapes (e.g., spheres, tubes, rods, wires) [11]. Nano-based drug delivery systems offer the potential for improved efficacy of treatment. ENMs are desirable because they easily enter cells and they can be designed to interact with specific cellular structures (e.g., receptors) to allow for accumulation within specific regions of the cell. ENMs can also be designed as pH-labile structures so that they degrade within the more acidic microenvironment of the cell to release drug payloads. Biocompatible ENMs that are used for drug delivery, such as poly-(ethylene glycol)-*poly*lactide co-glycolide copolymer (PEG-PLGA) nanoparticles, generally have low toxicity but often do not persist in tissues long enough for sustained drug or gene payload delivery. Nevertheless, PEG-PLGA NPs might be useful for delivering metabolic inhibitors of inflammation for the treatment of pulmonary diseases such as COPD or hypertension [12].

CNTs are durable and persist in tissues after administration by aerosol inhalation. An attractive property of CNTs, particularly SWCNTs, is that the tube- or fiber-like structure allows for extensive functionalization and loading of cargo. The versatile physicochemical features of CNTs allow for covalent and noncovalent functionalization to simultaneously carry several agents: (1) a drug, (2) an imaging agent (to track the course of delivery), and (3) a specific targeting agent (e.g., antibody selective for diseased tissue) [13]. Despite the versatility of CNTs as delivery platforms, some of the same unique properties that make CNTs desirable for therapeutic applications also make them potentially toxic. Thus, it is currently unclear whether the toxic effects of CNTs might ultimately outweigh any beneficial effects as drug delivery systems [14]. If CNTs are to be ultimately developed as a viable drug delivery platform, functionalization will be required to ameliorate toxicity, minimize the risk of side effects, and target therapeutic agents to the desired organs, tissues, or cells. A better understanding of how these nanomaterials interact at the nanoscale level with biological systems (i.e., the subcellular architecture) will be essential, along with adequate in vitro screening techniques to predict toxicity and thereby allow for safe design of ENMs in drug delivery for lung diseases [15].

For effective therapy, it is desirable to have a nanoparticle delivery system that remains in the target tissues long enough to achieve sustained and targeted release of drug, but nanoparticle persistence in tissues for too long will increase the possibility of toxic side effects and pathogenesis. Biodegradable nanoparticles are currently considered to be the best choice for targeted drug delivery, but biopersistent nanoparticles might also be considered valuable if they are cleared from tissues quickly enough to minimize exposure and retention [16]. Durability and degradation half-life must therefore be carefully considered in the intelligent design of ENMs to maximize beneficial pharmacologic effects while minimizing pathogenic side effects. Also, increasing solubility and decreasing aggregation status through functionalization will likely be important for developing ENMs as delivery platforms.

A variety of types of nanoparticles (lipid, cerium oxide, gold, iron oxide, super

paramagnetic iron oxide, lipid-polycation-DNA, silica-overcoated magnetic cores, and polyethylene glycol phosphatidylethanolamine) have been investigated for the treatment of lung cancer [17]. Nanoparticles can enhance drug delivery by encapsulation against immune response, increasing cellular uptake and tissue penetration, target selectivity, and specificity, promoting apoptosis, and blocking pathways for cancer initiation and progression. Nanoparticles have the ability to selectively accumulate in tumor tissue. The vasculature supplying tumors is often leaky due to loss of tight cell–cell contact between endothelial cells lining the capillaries, and these endothelial cell gaps (0.5–2.5 μm) allow circulating nanoparticles to exit the capillary and selectively accumulate in the tumor tissue [11]. In addition, nanoparticles remain in tumor tissue longer since there is poor lymphatic drainage surrounding tumors, which impedes nanoparticle clearance.

In addition to the ENMs mentioned above, CNTs could also offer potential for the treatment of lung cancer [18]. Tumor-targeted SWCNT constructs have been synthesized from water-soluble CNTs by covalent attachment of tumor-specific monoclonal antibodies [19]. Moreover, this strategy may be used along with thermal ablation of cancer cells by selectively heating CNTs taken up by cancer cells with infrared radiation [20]. While fibrotic reactions may still be a risk of these functionalized CNT platforms in some patients, overall the benefits could outweigh the risks. Other issues will also ultimately determine the potential success of treatment strategies that might utilize carbon nanotubes, such as costs of CNTs versus current biopolymers or other types of nanoparticles. In the words of Paracelsus (1493–1541), "the right dose differentiates a poison and a remedy." The success of CNTs in drug delivery will be achieved by using the lowest possible dose combined with single tube dispersibility, functionalization to target and/or carry a therapeutic agent, and biodegradability to avoid long-term chronic side effects.

Individuals with preexisting pulmonary diseases such as asthma and COPD represent susceptible populations to ENM exposures. However, these are also some of the same individuals that might benefit from improved drug delivery strategies or novel imaging techniques using nanotechnology. A major challenge for drug delivery in chronic airway diseases such as asthma and COPD is penetration through a thick layer of dense mucus that lines the airways. The hydrophobic and electrostatic forces within the mucus layer are highly effective in trapping particles and preventing access to the airway epithelium. Penetration through mucus can be achieved by properly modifying the size and charge of the nanoparticle. Coating particles with a high density of low molecular weight polyethylene glycol (PEG) renders the surface of the particle neutrally charged and improves transport of particles across the mucus layer [11].

ENMs may be effective in the treatment of infectious diseases that cause lung injury. For example, tuberculosis is a fatal respiratory infectious disease that represents a global threat and current oral antituberculosis therapy is limited by systemic toxicity [21]. Nanotechnology represents a promising strategy in the development of inhaled drug delivery systems for the treatment of TB. Nanocarriers (polymeric nanoparticles, solid lipid nanoparticles, nanoliposomes, and nanomicelles) can improve the drug effectiveness and decrease systemic side effects by targeting the drug to the infected area and sustaining release in a prolonged manner.

Conclusions

Nanotechnology is a "double-edged sword" for pulmonary diseases since some ENMs will likely cause pulmonary diseases or exacerbate preexisting disease when encountered in occupational or environmental exposure settings, while others will be highly beneficial for therapeutic applications in nanomedicine. The potential risks to human health and the environment should be carefully considered when designing ENMs for industrial purposes, consumer products, or nanomedicine. There are numerous physicochemical factors (e.g., aspect ratio, surface charge, biodegradability, etc.) that must be considered in the design of functional and safe ENMs for a wide

variety of applications. Because of the infancy of the field of nanotechnology, there are no epidemiologic data to indicate health hazards for ENMs.

Cross-References

- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Genotoxicity of Nanoparticles
- ▶ In Vivo Toxicity of Carbon Nanotubes
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Physicochemical Properties of Nanoparticles in Relation with Toxicity

References

1. Bonner, J.C.: Mesenchymal cell survival in airway and interstitial pulmonary fibrosis. Fibrogenesis Tissue Repair **3**, 15 (2010)
2. Donaldson, K., Murphy, F., Schinwald, A., Duffin, R., Poland, C.A.: Identifying the pulmonary hazard of high aspect ratio nanoparticles to enable their safety-by-design. Nanomedicine (Lond.) **6**(1), 143–156 (2011)
3. Kelleher, P., Pacheco, K., Newman, L.S.: Inorganic dust pneumonias: the metal-related parenchymal disorders. Environ. Health Perspect. **108**(Suppl 4), 685–696 (2000)
4. Bonner, J.C.: Respiratory toxicity. In: Smart, R.C., Hodgson, E. (eds.) Molecular and Biochemical Toxicology, 4th edn, pp. 639–670. Wiley, Hoboken (2008)
5. Thompson, E.A., Sayers, B.C., Glista-Baker, E.E., Shipkowski, K.A., Taylor, A.J., Bonner, J.C.: Innate immune responses to nanoparticle exposure in the lung. J. Environ. Immunol. Toxicol. **1**(3), 150–156 (2013)
6. Ryman-Rasmussen, J.P., Andersen, M.E., Bonner, J. C.: Fate and effects of carbon nanotubes following inhalation. In: Donaldson, K., Poland, C.A., Duffin, R., Bonner, J. (eds.) The Toxicology of Carbon Nanotubes, pp. 118–133. Cambridge University Press, Cambridge (2012)
7. Glista-Baker, E.E., Taylor, A.J., Sayers, B.C., Thompson, E.A., Bonner, J.C.: Nickel nanoparticles cause exaggerated lung and airway remodeling in mice lacking the T-box transcription factor, TBX21 (T-bet). Part. Fibre Toxicol. **11**, 7 (2014)
8. Schins, R.P.F., Albrecht, C., Gerloff, K.: Genotoxicity of carbon nanotubes. In: Donaldson, K., Poland, C.A., Duffin, R., Bonner, J. (eds.) The Toxicology of Carbon Nanotubes, pp. 150–173. Cambridge University Press, Cambridge (2012)
9. Sargent, L.M., Reynolds, S.H., Castranova, V.: Potential pulmonary effects of engineered carbon nanotubes: in vitro genotoxic effects. Nanotoxicology **4**, 396–408 (2010)
10. De Jong, W.H., Borm, P.H.: Drug delivery and nanoparticles: applications and hazards. Int. J. Nanomedicine **3**(2), 133–149 (2008)
11. Thorley, A.J., Tetley, T.D.: New perspectives in nanomedicine. Pharmacol. Ther. **140**(2), 176–185 (2013)
12. Vij, N., Gorde, A.: Theranostic applications of nanotechnology in chronic obstructive lung diseases. In: Vij, N. (ed.) Pulmonary Nanomedicine: Diagnostics, Imaging, and Therapeutics, pp. 1–14. Pan Stanford Publishing, Boca Raton (2012)
13. Prato, M., Kostarelos, K., Bianco, A.: Functionalized carbon nanotubes in drug design and discovery. Acc. Chem. Res. **41**(1), 60–68 (2008)
14. Bonner, J.C.: Carbon nanotubes as delivery systems for respiratory disease: do the dangers outweigh the potential benefits? Expert Rev. Respir. Med. **5**(6), 779–787 (2011)
15. Hirsch, C., Roesslein, M., Krug, H.F., Wick, P.: Nanomaterial cell interactions: are current in vitro tests reliable? Nanomedicine (Lond.) **6**(5), 837–847 (2011)
16. Kreyling, W.G., Hirn, S., Schleh, C.: Nanoparticles in the lung. Nat. Biotechnol. **28**(12), 1275–1276 (2010)
17. Chandolu, V., Dass, C.R.: Treatment of lung cancer using nanoparticle drug delivery systems. Curr. Drug Discov. Technol. **10**(2), 170–176 (2013)
18. Kuroda, S., Yokoyama, T., Tam, J.O., Scott, A.W., Ma, L.L., Shanker, M., Jin, J., Goerlich, C., Willcutts, D., Roth, J.A., Sokolov, K., Johnston, K.P., Ramesh, R.: Multifunctional tumor-targeted nanoparticles for lung cancer. In: Vij, N. (ed.) Pulmonary Nanomedicine: Diagnostics, Imaging, and Therapeutics, pp. 15–44. Pan Stanford Publishing, Boca Raton (2012)
19. Mulvey, J.J., Villa, C.H., McDevitt, M.R., Escorcia, F. E., Casey, E., Scheinberg, D.A.: Self-assembly of carbon nanotubes and antibodies on tumours for targeted amplified delivery. Nat. Nanotechnol. **8**(10), 763–771 (2013)
20. Chakravarty, P., Marches, R., Zimmerman, N.S., Swafford, A.D., Bajaj, P., Musselman, I.H., Pantano, P., Draper, R.K., Vitteta, P.S.: Thermal ablation of tumor cells with antibody-functionalized single-walled carbon nanotubes. Proc. Natl. Acad. Sci. U. S. A. **105**(25), 8697–8702 (2008)
21. Mehanna, M.M., Mohyeldin, S.M., Elgindy, N.A.: Respirable nanocarriers as a promising strategy for antitubercular drug delivery. J. Control. Release **187C**, 183–197 (2014)

Nanothermodynamics

- ▶ Surface Energy and Chemical Potential at Nanoscale

Nano-TiO_2 Life-Cycle Assessment Perspective

Gaetana Quaranta and Véronique Adam
Laboratoire d'Hydrologie et de Géochimie de la Surface, EOST, UDS (Université de Strasbourg), Strasbourg, France

Synonyms

Environmental impact assessment

Definition

Life cycle assessment (LCA) is a standard tool/method (ISO14040, 14044 -2006-) which aims at addressing environmental problems of a product over its life cycle (from cradle to grave).

In the context of nano-TiO_2, the term "life cycle" can be applied to the step of the NP production on its own, which includes an industrial process phase and a waste treatment phase, or it can be applied to the step of nanoparticles incorporation.

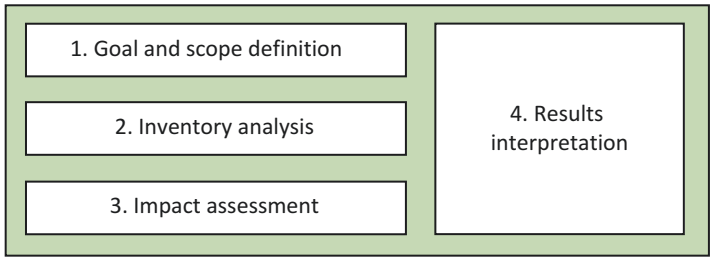
Introduction: Framework and Concept

Nano-TiO_2 represents the second most important part of ENM production worldwide (550–5,500 t/year) and in Europe (55–3,000 t/year) and the most important part in the USA (7,800–38,000 t/year). In the environment, TiO_2 is a mineral which can be found in the state of ilmenite ($FeTiO_3$), rutile (TiO_2), sphene ($CaSiTiO_5$), etc., ilmenite and rutile being the predominant forms of TiO_2, which occurs as three different crystalline phases: anatase, rutile, and brookite. Anatase and rutile are the most commonly encountered ones. At the macroscale [250–350 nm], TiO_2 is used as a pigment for surface coatings, for example, but also in non-pigment applications as a fluxing agent, slag former, etc. At the nanoscale [$<$100 nm], TiO_2 is not inert and finds its applications as a UV light attenuator and as a catalyst. Emerging applications of nano-TiO_2 use its conductive, catalytic, and photocatalytic properties for applications in environmental remediation, including degradation or adsorption of contaminants from surface water, groundwater, and air. However, even if more and more studies were performed on nano-TiO_2 in the last decade, high uncertainties remain regarding their fate and behavior in the environment and their toxicological effects. High uncertainties also remain regarding the methodological approaches to model these parameters in the aim of assessing their environmental impacts and their risks. Consequently, a framework able to deal with these uncertainties is needed for the environmental assessment of nano-TiO_2.

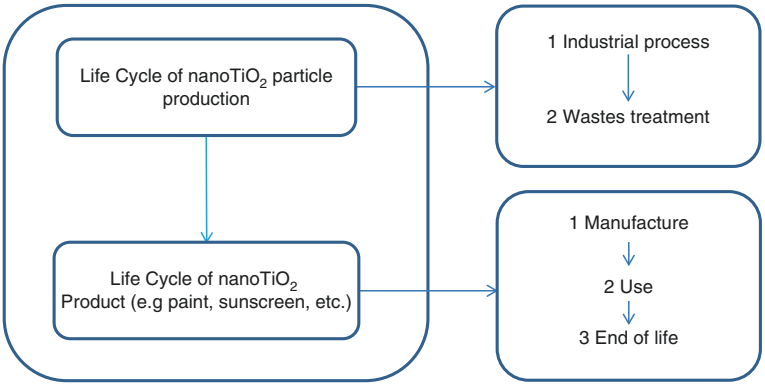
Several tools are used to understand the environmental impacts of products or substances, such as substance flow analysis (SFA), chemical risk assessment, and life-cycle assessment. These tools include the whole life cycle of the product or substance in their analysis. LCA, described as an environmental impact assessment tool for products, processes, or services over their whole life cycles [8], and risk assessment (RA), defined as the process of assessing magnitudes and probabilities of human activities or natural catastrophes adverse effects, are two methodologies used to assess the impacts and risks of substances in the environment. Today, there is consensus among scientists on combining these two tools in order to overcome their respective limitations. LCA is conducted in four stages following the requirements of the ISO 14040: the determination of the goal, scope, and system boundaries, the inventory analysis of inputs and outputs, the assessment of environmental impacts, and the interpretation of results with proposals for improvement (Fig. 1).

The term "life cycle" refers to the extraction and transformation of raw materials, as well as the manufacture, the transformation, and the distribution, use, and end-of-life steps of products.

In the context of nano-TiO_2 (Fig. 2), the term "life cycle" can be applied to the step of the NP production on its own, which includes an industrial process phase and a waste treatment phase, or it can be applied to the step of nanoparticle incorporation into different nanomaterials which are



Nano-TiO_2 Life-Cycle Assessment Perspective, Fig. 1 Life-cycle assessment framework [8]



Nano-TiO_2 Life-Cycle Assessment Perspective, Fig. 2 LCA methodological approach of nano-TiO_2

then used in different products. In this case the cycle includes the manufacture, the use, and the end of life of the nanoproduct. The term "life cycle" can also consider both nanoparticle and nanoproduct LCAs, which are then the two subsystems of the global life cycle.

This entry presents in the first section the life-cycle assessment methodological approach of the nano-TiO_2 particle production and in the second section the current research on the parameters needed for determining the nano-TiO_2 characterization factor in their life-cycle impact assessment.

Life-Cycle Assessment Applied to a TiO_2 NP Production Site

Goal and Scope of the Study

The LCA was conducted following the requirements of the ISO 14040 [8] with this aim of quantifying the environmental impacts related to an industrial TiO_2 production site. In this case the system could be illustrated as in Fig. 3, taking into account both the production process and the treatment of acid effluents.

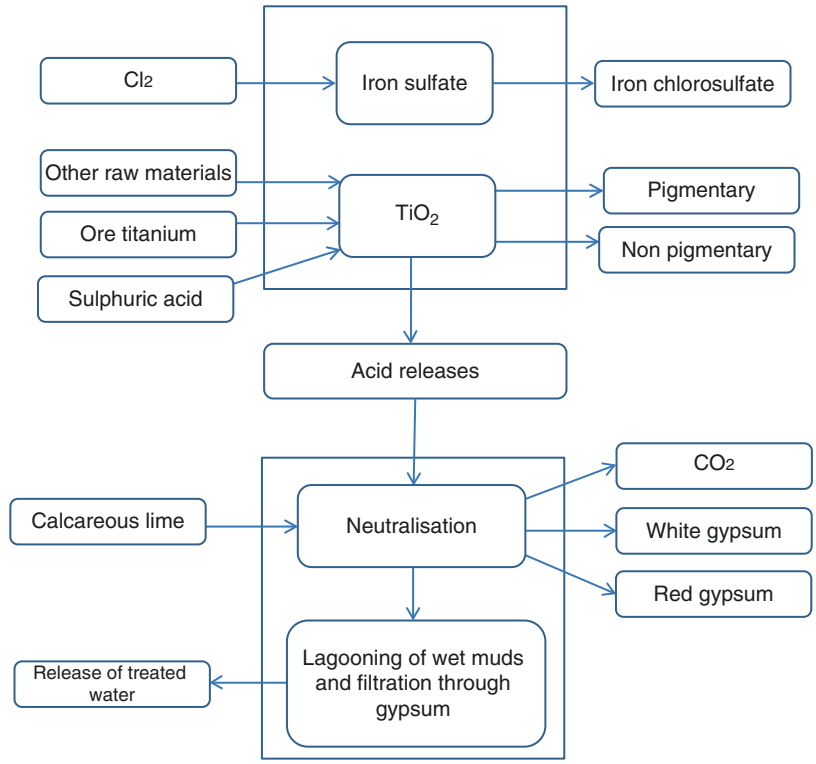
Focusing on the production workshop, the limits of the studied system are represented in Fig. 4. Two sectors are under study: the first one called "black part" includes titanium ore (as input) and its successive treatments, excluding fabrication of iron chlorosulfate from the scope of the study. At the exit of this first sector, a solution of titanyl sulfate has been obtained and can then be treated in a second sector called "white part," including the hydrolysis of the solution leading to the crystallization of titanium dioxide and different washings removing impurities such as iron, chromium, or vanadium.

Then the titanium gel is moved towards different workshops according to the wanted final product. The "anatase" shop produces microparticles of titanium, whereas the "ultrafine" workshop produces nanoparticles of titanium, either on the GT or the DT production lines (Fig. 4).

The selected functional unit (FU), quantifying the function of the system and being the unit to which all impact factors are related, is 1 t of produced nanoparticles.

Data Inventory

A data inventory was obtained from the existing titanium production plant, from a functioning balance of a site among a 10-year period. These data



Nano-TiO_2 Life-Cycle Assessment Perspective, Fig. 3 Simplified scheme of TiO_2 industrial process and materials

came from a reliable source but are global and nonspecific to the nano-titanium workshop. However, it has been possible to distinguish some air emissions related to the "black sector," "white sector," and "ultrafine sector" and other air emissions (called global air sector) related to the other sectors of the plant. Concerning the wastewaters, the data are global (for the whole plant).

The data inventory is reported in Tables 1 and 2.

Impact Assessment

In the impact assessment stage (LCIA), inventory data are assigned to different impact categories. The calculation of category indicators (CIs) is achieved using characterization factors (CFs).

The category indicator is usually expressed with the following equation, which is the numerical expression of the life-cycle inventory result conversion (Eq. 1) [7]:

$$C_i = \sum_{e=1}^{e=m} \sum_{x=1}^{x=n} CF_{i,x,e} \cdot M_{x,e} \qquad (1)$$

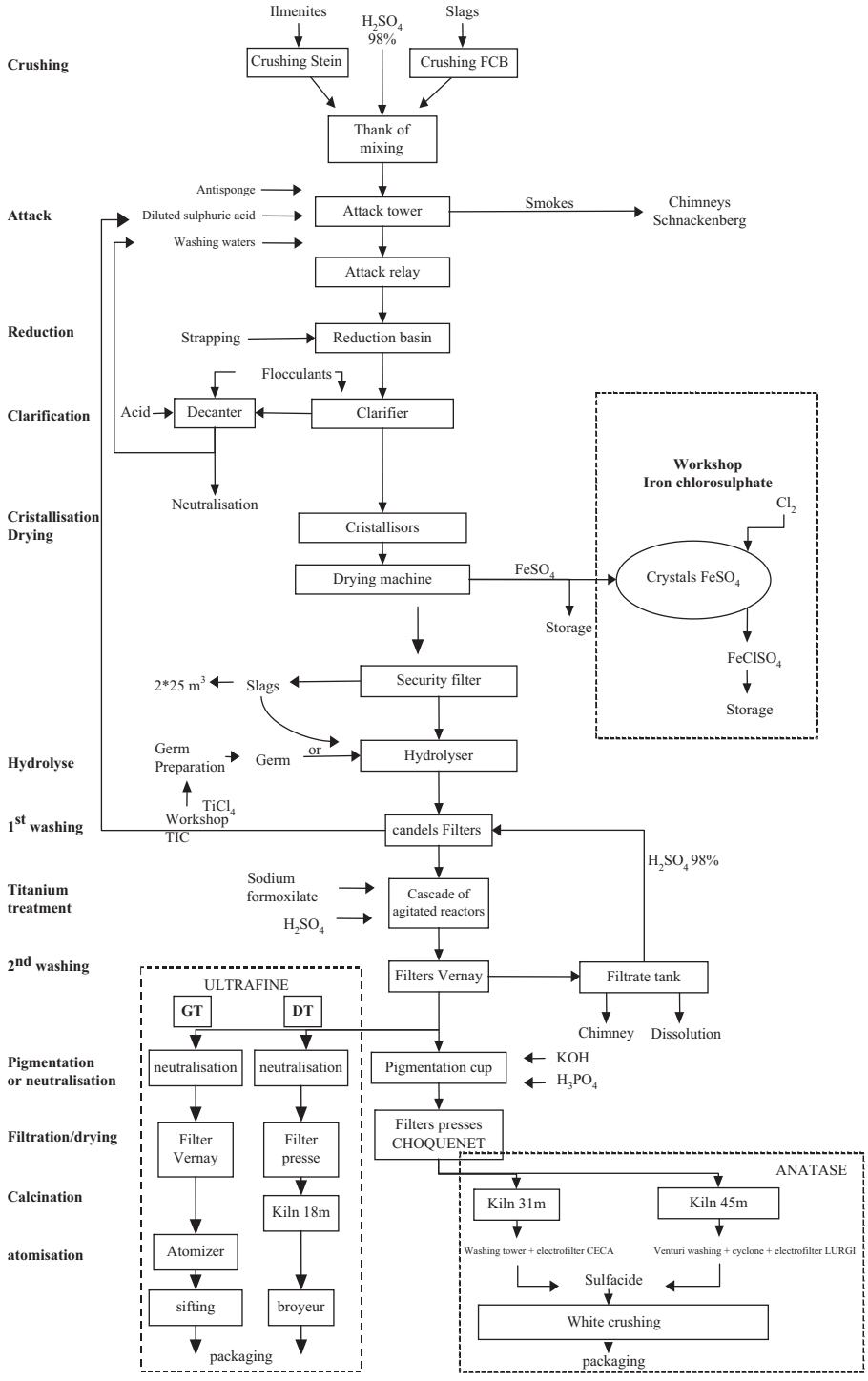
where C_i is the impact score for category i, $M_{x,e}$ the substance mass (LCI results), and $CF_{i,x,e}$ the characterization factor of impact category i for substance x due to an emission to compartment e (dimensionless).

and

$$CF = EF \times FF \qquad (2)$$

where CF is the characterization factor of the impact category, EF is the effect factor of a substance x, and FF is the fate factor of the substance x.

Many LCIA calculation methods/models are available, each providing different CFs. In this study the LCA SIMAPRO software v.7.3.3 (Pre Consultants B.V) has been used with the Ecoinvent database within the program. The study performed an environmental evaluation according to the Impact 2002+ assessment method which offers an attractive implementation of a combined midpoint/damage approach via 14 midpoint categories to four damage categories



Nano-TiO_2 Life-Cycle Assessment Perspective, Fig. 4 Synoptic scheme of the TiO_2 industrial process

Nano-TiO_2 Life-Cycle Assessment Perspective, Table 1 Global emissions

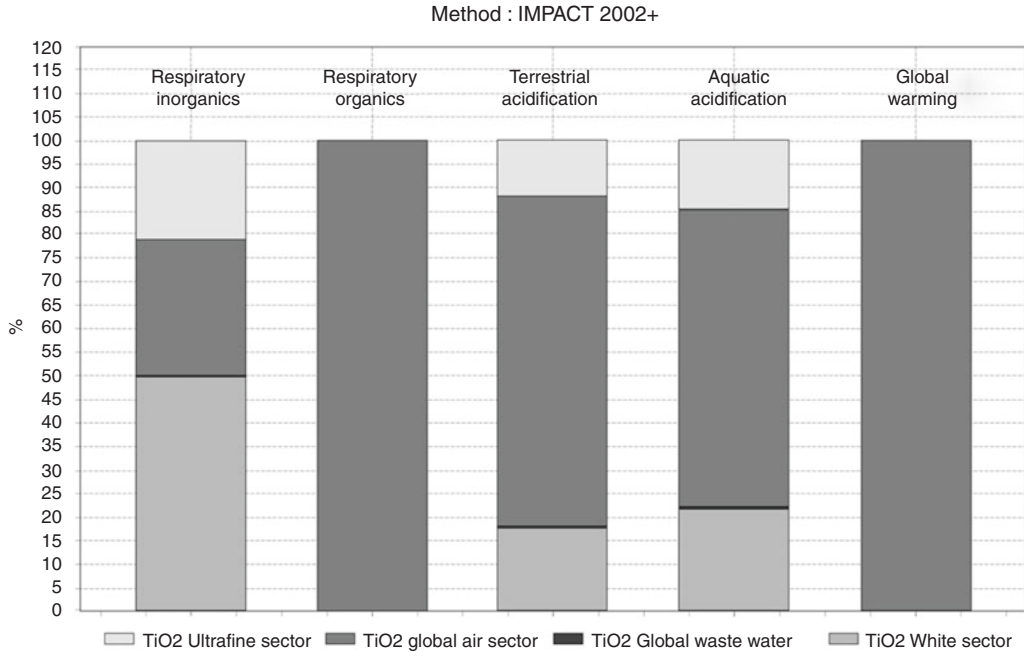
		Year			Average value
		X	Y	Z	
Global water emissions kg/T of TiO_2/year	Total nitrogen	0.1589	0.1606	0.2141	**0.178**
	Total phosphorus	0.0005	0.0026	0.011	**0.0047**
	Total chloride	65.3813	43.4232	54.4593	**54.42**
	Sulfate	47.1293	39.6254	50.7791	**45.9**
Others air emissions (global air sector) kg/T of TiO_2/year	SO_2	9.7872	7.9442	9.395	**9.04**
	NO_x	0.7948	0.884	0.9277	**0.87**
	CH_4	0.0019	0.0022	0.0021	**0.002**
	N_2O	0.0229	0.0271	0.0315	**0.027**
	CO_2	769.8463	820.3313	908.2478	**832.8**
	CO	41.6776	40.8375	47.7222	**43.4**
	NMVOC	0.023	0.0257	0.0293	**0.026**

Nano-TiO_2 Life-Cycle Assessment Perspective, Table 2 Emissions by sector

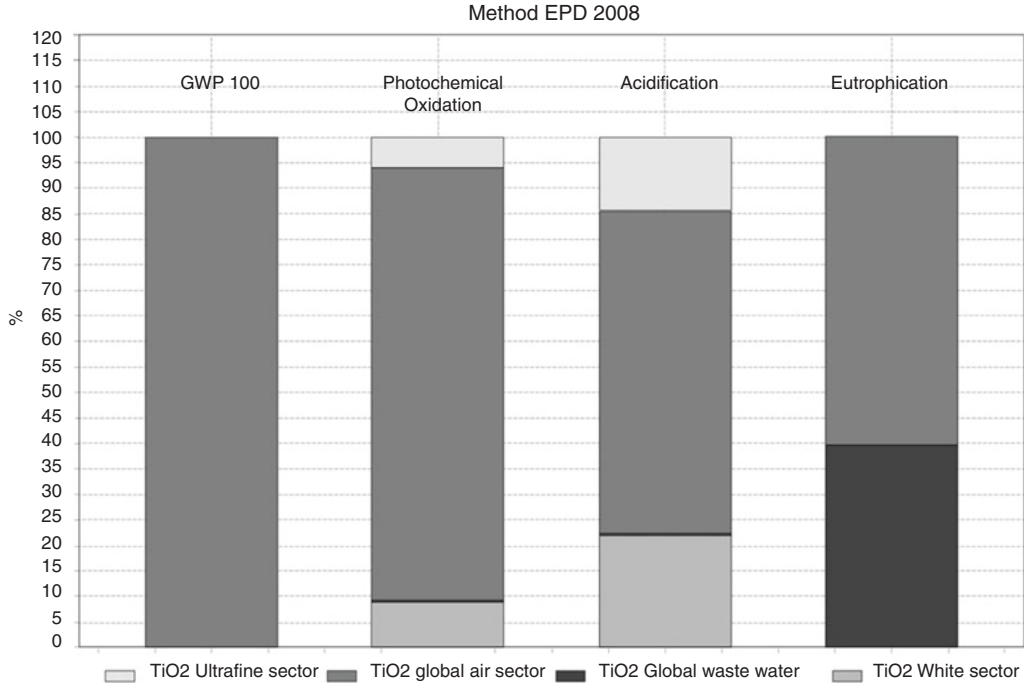
	Measured parameters	Year			Average value
		X	Y	Z	
Black sector	Ore particle kg/h: crushing FCB	0.29	0.01	0.01	0.1
	Ore particle kg/h: crushing Stein	0.05	0.94	0.04	0.3
	Pollutant flows SO_2+SO_3+H_2SO_4, kg SO_2/t of TiO_2: attack tower	0.05	0.104	0.037	0.06
	Pollutant flows SO_2+SO_3+H_2SO_4, kg SO_2/t of TiO_2: oven 31 M and oven 35 M	4.88	2.04	3.79	3.57
White sector	Pollutant flows TiO_2 kg/h: crushing Neumann	0.03	0.04	0.03	0.03
	Pollutant flows of TiO_2 particles kg/h: downstream silo Neumann	0.01	0.01	0.02	0.01
	Pollutant flows SO_2+SO_3+H_2SO_4, kg SO_2/t of TiO_2: four 18 M	3.522	3.491	1.412	2.81
Ultrafine sector	Pollutant flows of TiO_2 particles kg/h: oven 18 M	1.262	0.248	0.235	0.58
	Pollutant flows of TiO_2 particles kg/h: denox atomizer	0.026	0.133	0.687	0.28

but also according to EPD 2008 v1.03 which describes the environmentally relevant physical flows to and from a product system and its subsystems.

Thus, the environmental impact contributions of the **TiO_2 white sector**, the **TiO_2 black sector**, the **TiO_2 ultrafine sector**, the **TiO_2 global air**, and the **TiO_2 global wastewater** have been assessed (Fig. 5). This industrial system shows that impacts are related to the TiO_2 global air that is to say to the atmospheric emissions not assigned to white, black, or ultrafine sector but depending on the other workshops of the plant. The TiO_2 black sector induces no impact; this is due to low quantities of pollutants emitted that are negligible compared to other workshops. Concerning the "respiratory inorganic" effects, the TiO_2 white sector contributes to 50 % of the observed impacts, followed by TiO_2 global air (29 %) and the ultrafine workshop. The emissions of the TiO_2 white sector depend on the SO_2 emissions and on the ore dusts. The ultrafine shop contributes to 21 % of this impact, but high uncertainty remains regarding the nano-TiO_2 particle fate and behavior in the environment and their toxicological effects. For the "respiratory organic" effects, which include all the carbon-based molecules, the TiO_2 global air sector



Analysis 1p "Global Industrial process" Method Impact 2002+/Characterization



Analysis 1p "Global Industrial process" Method EPD 2008/Characterization

Nano-TiO_2 Life-Cycle Assessment Perspective, Fig. 5 Impact assessment results

represents 100 % of this impact category. The different workshops of the industrial process should contribute to this impact category, but due to the lack of data, it was impossible to quantify their emissions and the different contributing molecules. The terrestrial and aquatic acidification depends on SO_2, NO_x, and chlorine emissions which have effects in series, the TiO_2 "global air" subsystem generating the highest impact. In the other workshops, SO_2 alone contributes to these categories, implying a moderated influence. Regarding the aquatic acidification, a question remains about the non-contribution of the TiO_2 global wastewater to this impact, as it has been noted that chlorine was released into water. The global warming is due to the emissions of the TiO_2 global air sector: CO_2, CO, CH_4, and NMVOC (non-methane volatile organic compounds). The data allowing the assessment of the global warming by sector (black and white sectors) being unavailable, only the global warming applied to the whole process has been appreciated.

The results obtained with the EPD 2008 method are illustrated for four impact categories: global warming, photochemical oxidation, aquatic acidification, and eutrophication (Fig. 5). The TiO_2 global air part induces most of the calculated impact. Concerning the global warming and the aquatic acidification, the results are similar to those of the Impact 2002+ method. The contribution to the eutrophication impact category is of 40 % for the TiO_2 global wastewater sector and 60 % for the TiO_2 global air sector, NO_x, N_2O, and Ptotal being the main contributors. Other molecules have only a limited influence in this process. These are releases induced by the whole site, so the analysis is not limited to the production of nano-TiO_2.

Conclusion of the Study

The LCA approach can provide useful information about the main environmental impacts, but the main problem encountered is the lack of data. During the study, the difficulties are of different orders, linked to databases or physicochemical characterization of nanoparticles. The lack of data for each manufacturing process has failed to allow a cross-process comparison: apart from a few sectoral data, most of the data was aggregated over the whole production site. This study focuses on a specific TiO_2 life-cycle stage excluding the effect and the fate of nano-TiO_2 because of the lack of information. The impact calculations are based only on the inputs and outputs of the process (excluding nanoparticles). LCA needs experimental data on characteristics and toxicity of nano-TiO_2 to address all life-cycle stages. It seems necessary to determine which of their properties need to be characterized in priority.

Overview of Necessary Data to Complete a Nano-TiO_2 LCA

The behavior of TiO_2 NMs in environmental compartments relies on their specific properties. Among them, seven have been identified as primordial: crystal phase, shape, primary (as received) size and size distribution, subsequent aggregation state, number concentration, surface area, and surface reactivity (charge and chemistry). This section shows how these characteristics are interrelated and why they are important in the modeling of TiO_2 NP behavior in biotic and abiotic media.

Crystal Phase

TiO_2 is naturally present in three different crystalline phases; the different lattice structures induce differences in mass densities and in reactivity. These differences in reactivity lead to different modes of toxicity. For example, amorphous TiO_2 can produce more reactive oxygen species (ROS) and oxidative stress to organisms than anatase. However, it has been pointed out that the media in which NPs are placed play a role in the generation of ROS.

Shape

In water, the shape of NMs, and the volume they occupy, will mostly influence their gravitational settling, their collision frequencies, and their aggregation. This will affect their sedimentation and finally their final concentrations in the water column and in the sediment. On a toxicity viewpoint, it has been proposed that "fibers" (particles with a diameter-to-length ratio higher than 1:3) have "extended" dimensions and generally lead to

inflammatory responses of the immune system. Thin fibers have a small aerodynamic diameter and can be deposited beyond the ciliated airways in the respiratory tract, but they cannot be entirely processed by immune processes such as phagocytosis [2, 3]. It has also been found that TiO_2 NMs with a length higher than 15 μm cannot be processed by lung macrophages, leading to inflammation [6].

Size and Size Distribution

Smaller particles present higher surface/mass ratio, also called specific surface area. So by definition, the variation of the size of a particle affects its specific surface area. Higher specific surface area leads to higher reactivity, although this phenomenon can be altered by coatings. Thus, a diminution in size has been found to enhance aggregation of spherical anatase TiO_2 NPs and sorption of organic matter. Size also influences the NP dissolution rates, through surface area, surface curvature, and roughness, and alters their transport. A decrease in size will lead to higher diffusion velocity, while increasing size induces higher gravitational settling.

In soils, larger aggregates are more influenced by gravity forces and are more likely to attach to soil particles. These large aggregates are prone to pore straining and may lead to soil clogging.

The point of zero charge (PZC) can be assimilated to the surface charge of nanoparticles (see section "Surface Charge and Chemistry"). It also appears to be affected by the size of the NPs. It has been shown that the PZC of rutile and anatase TiO_2 NPs increases towards higher pH values with decreasing size in water, whatever the shape of the NPs. Nano-TiO_2 aggregates when the solution pH is close to the PZC, whereas nano-TiO_2 remains free at very basic and very acidic pH values.

In organisms, the size of NMs affects their toxicity firstly because of their different transport abilities. Smaller particles will be deposited further in the respiratory tract, passing through the ciliated airways, and will have more ability to pass cell membranes and enter the blood circulation. In the same way, smaller TiO_2 NPs (smaller primary diameter) penetrate deeper into the wheat root and are more readily translocated to the shoot of the plant [15]. Decreasing particle size also induces more structural defects and more reactive groups on the surface, leading to prooxidant effects. An increased toxicity of TiO_2 NPs has been found to *Escherichia coli* and *Pseudomonas aeruginosa* with size decreasing from 30 to 10 nm [9]. Similarly, an increased cytotoxicity of TiO_2 NPs has been showed with decreasing size in murine macrophages, which was related to increasing surface area [20].

However, several studies demonstrated that the assumption that smaller particles induce higher toxicity is not true in all cases [11, 17, 19].

Specific Surface Area

The specific surface area of NPs does not depend only on their size but also on their porosity and surface roughness. Higher specific surface area (surface over mass ratio) leads to enhanced reactivity (e.g., dissolution and aggregation), and it can induce higher deposition on soil particles. Higher specific surface area also induces higher biologic activity per given mass of NP.

Thus, numerous links have been found between specific surface area and NP toxicity, and it is considered as a critical driver of NP toxicity. However, current sampling and analysis methodologies rely on the mass of substances.

Surface Charge and Chemistry

The surface charge of NPs is affected by the pH and the ionic strength of the surrounding medium. It is measured as the zeta potential, which is "the potential at the slipping plane that separates the stationary and mobile phases in tangential flow of the liquid with respect to the surface" [12]. The z-potential is one of the factors controlling electrostatic repulsion: increasing net charge of the surface leads to increasing repulsion between NPs and to lower aggregation rates. A null z-potential is referred to as the point of zero charge (PZC) and favors aggregation; suspensions are said to be stable when the zeta potential of the NPs is higher than 30 mV in absolute value. The charge of NPs also controls their transport in porous media, and it has been shown that positively charged NPs were more readily retained into a sand or soil column than negatively charged particles.

The z-potential of NPs influences their oxidative potential and the production of ROS, inducing increased cytotoxicity and inflammatory effects with increasing charge. However, it has been pointed out that even if increasing charge usually leads to easier uptake of NPs by lysosomes and increased cytotoxicity, z-potential is not a predictor of nanotoxicity, because the charge at specific spots of the NP surface is also a significant contributor to toxicity, as they will determine precisely their penetration into cells.

The solubility of NPs is also an important parameter to take into account when assessing their toxicity. Firstly, it conditions their persistence in the media and in the organism. Secondly, the release of ions increases the oxidative potential and prooxidant effects. Finally, the ions themselves can be toxic, depending on the composition of the NP and on the target organism. TiO_2 NP solubility is low, so that they tend to accumulate in soils and water, leading to high potential bioaccumulation in organisms.

Concentration

Concentration is usually expressed as a mass/volume ratio. However, in the case of NMs, the mass might not be the best metric to use in fate modeling and toxicity assays. Indeed, it has been showed that for the same mass dose, NPs induce higher toxicity than their bulk counterparts, due to higher specific surface area. Debate is still ongoing about the best metric to use, but particle number concentration and available surface area appear as best appropriate. However, Pauluhn [16] showed that the mass-based exposure intensity (actual concentration × duration of exposure) of NPs is a better dose metric for lung burden than surface area (BET surface area × duration of exposure). No correlation between the pulmonary toxicity of TiO_2 NPs to rats and the surface area of the NPs has been found. This could be due to the high aggregation state that was reported.

On the one hand, higher number concentration of particles in a same volume leads to higher collision rates and to higher aggregation rates. The size of NM aggregates is then increased, and deposition is enhanced. Higher number concentration also induces higher chances of contact with organisms and higher potential toxicity. On the other hand, if the NP aggregates are large enough, their entrance in organisms, e.g., via dermal penetration, may be less likely.

Agglomeration and Aggregation State

Larger particles and aggregates tend to bind more weakly than smaller ones. The type of forces entailed in aggregation is determined by the size of the NPs. If the forces between particles are weak, then the attraction process is called agglomeration; in contrary when the forces are stronger, the process is called aggregation. It was found that aggregates of 21 nm NPs present weaker bounds than 5 nm NPs, thus deagglomerating more readily and resulting in a higher final pulmonary toxicity than that of 21 nm NPs [10].

NP aggregation state and their size distribution are important parameters controlling their behavior. Larger aggregates have a higher mass and higher volume than the NPs they are composed of, so they are more affected by the gravitational and buoyant forces acting on them, leading to enhanced settling rate. On the opposite, their diffusion is decreased with increasing size. The aggregation state also affects the surface area of the NM: large and dense aggregates have smaller specific surface area than porous and small aggregates, leading to lower reactivity. Moreover, larger aggregates are more readily filtered in porous media and show higher retention in sand and soil.

The aggregation state of NPs also affects their biouptake. It is generally considered that larger aggregates penetrate less easily into organisms. Furthermore, Li et al. [13] showed that heteroaggregation of TiO_2 NPs with algal cells lead to decreasing food availability and quality for *Ceriodaphnia dubia*.

This section shows that there is no absolute trend determining what property explains best the TiO_2 NM behavior. It is clear that the chemical composition and crystal phase should be determined in priority, but other properties such as size, particle number, and surface area are important in the understanding of the NM behavior, and many of them are intermingled. Adapted exposure and TiO_2 fate modeling are necessary to assess the environmental impact of TiO_2 NM during all life-cycle stages.

Conclusion

Today there is consensus among authors that large benefits arise from the combination of LCA and RA approaches. This combination could help at identifying the fate and effect factors which determine the characterization factor of an impact category and at calculating exposure levels and the potential damage of NPs on environment and human health along the life cycle.

Combining RA and LCA would allow overcoming their respective limitations. Indeed, environmental risk assessment (ERA) generally considers ecotoxicity only, so that negative impacts on other categories (such as climate change) that could arise from the mitigation of ecotoxicity are not acknowledged. The wide range of impact categories in LCA allows stakeholders to have a more comprehensive view of the system under study [1]. In the same way, the consideration of all emissions over the whole life cycle of the substance, as it is the case in LCA, would make RA a more powerful decision support [4, 5, 14] and would also allow improving the life-cycle processes [18]. The time- and site-specific design of ERA would also benefit to LCA, reducing result uncertainties.

References

1. Christensen, F.M., Olsen, S.I.: The potential role of life cycle assessment in regulation of chemicals in the European Union. Int. J. LCA **9**(5), 327–332 (2004)
2. Donaldson, K., Murphy, F.A., Duffin, R., Poland, C.: Asbestos, carbon nanotubes and the pleural mesothelium: a review of the hypothesis regarding the role of long fibre retention in the parietal pleura, inflammation and mesothelioma. Part. Fibre Toxicol. **7**, 5 (2010)
3. Donaldson, K., Schinwald, A., Murphy, F., Cho, W.S., Duffin, R., Tran, L., Poland, C.: The biologically effective dose in inhalation nanotoxicology. Acc. Chem. Res. **46**, 723–732 (2013)
4. Dusinska, M., Fjellsbo, L.M., Olsen, S., Baun, A., Dekkers, S., Nowack, B., et al.: Final Report with Recommendations for Impact Assessment of Engineered Nanomaterials, p. 22. Institute of Work and Health, Lausanne (2012)
5. Eason, T., Meyer, D.E., Curran, M.A., Upadhyayula, V.K.K.: Guidance to Facilitate Decisions for Sustainable Nanotechnology. US Environmental Protection Agency, Washington, DC (2011). 75 pp
6. Hamilton, R.F., Wu, N., Porter, D., Buford, M., Wolfarth, M., Holian, A.: Particle length-dependent titanium dioxide nanomaterials toxicity and bioactivity. Part. Fibre Toxicol. **6**, 35 (2009)
7. Huijbregts, M., Hauschild, M., Jolliet, O., Margni, M., McKone, T., Rosenbaum, R.K., van de Meent, D.: USEtox™ User Manual, 23 pp. (2010). Accessed at: http://www.usetox.org/sites/default/files/support-tutorials/user_manual_usetox.pdf on 18th June 2013. ISO, ISO 14040:2006 environmental management – life cycle assessment – principles and framework, ISOTC 207SC 5 23, 2006
8. ISO: ISO 14044:2006 Environmental Management – Life Cycle Assessment – Requirements and Guidelines, ISOTC 207SC 5 46, 2006
9. Jang, H., Kim, S., Kim, S.: Effect of particle size and phase composition of titanium dioxide nanoparticles on the photocatalytic properties. J. Nanoparticle Res. **3**, 141–147 (2007)
10. Johnston, H.J., Hutchison, G.R., Christensen, F.M., Peters, S., Hankin, S., Stone, V.: Identification of the mechanisms that drive the toxicity of TiO_2 particulates: the contribution of physicochemical characteristics. Part. Fibre Toxicol. **6**, 33 (2009)
11. Kobayashi, N., Naya, M., Endoh, S., Maru, J., Yamamoto, K., Nakanishi, J.: Comparative pulmonary toxicity study of nano-TiO_2 particles of different sizes and agglomerations in rats: different short- and long-term post-instillation results. Toxicology **264**, 110–118 (2009)
12. Leroy, P., Tournassat, C., Bizi, M.: Influence of surface conductivity on the apparent zeta potential of TiO_2 nanoparticles. J. Colloid Interface Sci. **356**, 442–453 (2011)
13. Li, M., Czymmek, K.J., Huang, C.P.: Responses of Ceriodaphnia dubia to TiO_2 and Al_2O_3 nanoparticles: a dynamic nano-toxicity assessment of energy budget distribution. J. Hazard. Mater. **187**, 502–508 (2011)
14. Linkov, I., Seager, T.P.: Coupling multi-criteria decision analysis, life cycle assessment, and risk assessment for emerging threats. Environ. Sci. Technol. **45**, 5068–5074 (2011)
15. Manke, A., Wang, L., Rojanasakul, Y.: Mechanisms of nanoparticle-induced oxidative stress and toxicity. Bio. Med. Res. Int. **2013**, 1–15 (2013)
16. Pauluhn, J.: Retrospective analysis of 4-week inhalation studies in rats with focus on fate and pulmonary toxicity of two nanosized aluminum oxyhydroxides (boehmite) and pigment-grade iron oxide (magnetite): the key metric of dose is particle mass and not particle surface area. Toxicology **259**, 140–148 (2009)
17. Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T.P., Michalkova, A., Hwang, H.M., Toropov, A., Lesczynska, D., Lesczynski, J.: Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. Nat. Nanotechnol. **6**, 175–178 (2011)
18. Sweet, L., Strohm, B.: Nanotechnology – life-cycle risk management. Hum. Ecol. Risk Assess. **12**, 528–551 (2006)

19. Wang, J., Zhou, G., Chen, C., Yu, H., Wang, T., Ma, Y., Jia, G., Gao, Y., Li, B., Sun, J., Li, Y., Jiao, F., Zhao, Y., Chai, Z.: Acute toxicity and biodistribution of different sized titanium dioxide particles in mice after oral administration. Toxicol. Lett. **168**, 176–185 (2007)
20. Xiong, S., George, S., Yu, H., Damoiseaux, R., France, B., Loo, K.W., Ng, J.S.C.: Size influences the cytotoxicity of poly (lactic-co-glycolic acid) (PLGA) and titanium dioxide (TiO_2) nanoparticles. Arch. Toxicol. **87**, 1075–1086 (2013)

Nanotoxicity

► Use of Nanotechnology in Pregnancy

Nanotoxicology

► Physicochemical Properties of Nanoparticles in Relation with Toxicity

Nanotransfer Printing

► Nanoscale Printing

Nanotribology

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Adhesion; Friction; Lubrication; Wear

Definition

The word "Tribology" was coined in 1966. It is derived from the Greek word *tribos*, meaning rubbing so the literal translation would be the science of rubbing. However, it encompasses studies of surface characterization, adhesion, friction, wear, and lubrication. According to Webster dictionary, Tribology is defined as the science and technology of interacting surfaces in relative motion and of the related subjects and practices. Tribology is an interdisciplinary field. Rather complex surface interactions in a tribological interface require knowledge of various disciplines including physics, chemistry, mechanical engineering, solid mechanics, materials science, rheology, applied mathematics, and reliability. At most interfaces of technological relevance, contact occurs at numerous asperities. It is of importance to investigate a single asperity contact in the fundamental tribological studies and thus emphasizing the importance of nanotribology.

Introduction

The mechanisms and dynamics of the interactions of two contacting solids during relative motion, ranging from atomic- to microscale, need to be understood in order to develop fundamental understanding of adhesion, friction, wear, indentation, and lubrication processes. For most solid–solid interfaces of technological relevance, contact occurs at multiple asperities. Consequently, the importance of investigating single asperity contacts in studies of the fundamental micro/nanomechanical and micro/nanotribological properties of surfaces and interfaces has long been recognized. The recent emergence and proliferation of proximal probes, in particular scanning probe microscopes (the scanning tunneling microscope and the atomic force microscope), the surface force apparatus, and of computational techniques for simulating tip-surface interactions and interfacial properties, has allowed systematic investigations of interfacial problems with high resolution as well as ways and means for modifying and manipulating nanoscale structures. These advances have led to the appearance of the new field of nanotribology, which pertains to experimental and theoretical investigations of interfacial processes on scales ranging from the atomic- and molecular-scale to the microscale, occurring during adhesion, friction, scratching, wear,

Nanotribology, Table 1 Comparison of typical operating parameters in SFA, STM and AFM/FFM used for micro/nanotribological studies

Operating parameter	SFA	STM[a]	AFM/FFM
Radius of mating surface/tip	~10 mm	5–100 nm	5–100 nm
Radius of contact area	10–40 μm	N/A	0.05–0.5 nm
Normal load	10–100 mN	N/A	<0.1–500 nN
Sliding velocity	0.001–100 μm/s	0.02–200 μm/s (scan size ~ 1 × 1 nm to 125 × 125 μm; scan rate <1–122 Hz)	0.02–200 μm/s (scan size ~ 1 × 1 nm to 125 × 125 μm; scan rate <1–122 Hz)
Sample limitations	Typically atomically–smooth, optically transparent mica; opaque ceramic, smooth surfaces can also be used	Electrically-conducting samples	None of the above

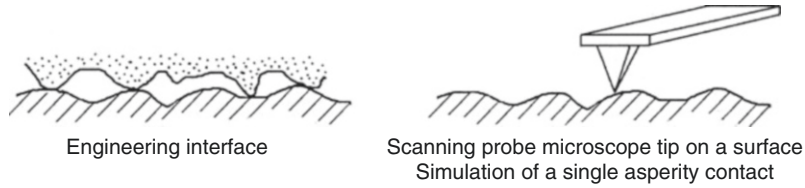
[a]Can only be used for atomic-scale imaging

indentation, and thin-film lubrication at sliding surfaces [2, 4–8, 10–12, 15, 18, 22, 25]. Proximal probes have also been used for mechanical and electrical characterization, in situ characterization of local deformation, and other nanomechanics studies.

Nanotribological and nanomechanics studies are needed to develop fundamental understanding of interfacial phenomena on a small scale and to study interfacial phenomena in micro/nanostructures used in magnetic storage devices, nanotechnology, and other applications [1–12]. Friction and wear of lightly loaded micro/nanocomponents are highly dependent on the surface interactions (few atomic layers). These structures are generally coated with molecularly thin films. Nanotribological and nanomechanics studies are also valuable in the fundamental understanding of interfacial phenomena in macrostructures, and provide a bridge between science and engineering.

The surface force apparatus (SFA), the scanning tunneling microscopes (STM), atomic force and friction force microscopes (AFM and FFM) are widely used in nanotribological and nanomechanics studies. Typical operating parameters are compared in Table 1. The SFA was developed in 1968 and is commonly employed to study both static and dynamic properties of molecularly thin films sandwiched between two molecularly smooth surfaces. The STM, developed in 1981, allows imaging of electrically conducting surfaces with atomic resolution, and has been used for imaging of clean surfaces as well as of lubricant molecules. The introduction of the AFM in 1985 provided a method for measuring ultra-small forces between a probe tip and an engineering (electrically conducting or insulating) surface, and has been used for morphological and surface roughness measurements of surfaces on the nanoscale, as well as for adhesion measurements. Subsequent modifications of the AFM led to the development of the FFM, designed for atomic-scale and microscale studies of friction. This instrument measures forces in the scanning direction. The AFM is also being used for various investigations including scratching, wear, indentation, detection of transfer of material, boundary lubrication, and fabrication and machining. Meanwhile, significant progress in understanding the fundamental nature of bonding and interactions in materials, combined with advances in computer-based modeling and simulation methods, has allowed theoretical studies of complex interfacial phenomena with high resolution in space and time. Such simulations provide insights into atomic-scale energetics, structure, dynamics, thermodynamics, transport, and rheological aspects of tribological processes.

The nature of interactions between two surfaces brought close together, and those between



Nanotribology, Fig. 1 Schematics of an engineering interface and scanning probe microscope tip in contact with an engineering interface

two surfaces in contact as they are separated, have been studied experimentally with the surface force apparatus. This has led to a basic understanding of the normal forces between surfaces, and the way in which these are modified by the presence of a thin liquid or a polymer film. The frictional properties of such systems have been studied by moving the surfaces laterally, and such experiments have provided insights into the molecular-scale operation of lubricants such as thin liquid or polymer films. Complementary to these studies are those in which the AFM tip is used to simulate a single asperity contact with a solid or lubricated surface, Fig. 1. These experiments have demonstrated that the relationship between friction and surface roughness is not always simple or obvious. AFM studies have also revealed much about the nanoscale nature of intimate contact during wear, indentation, and lubrication.

Overview

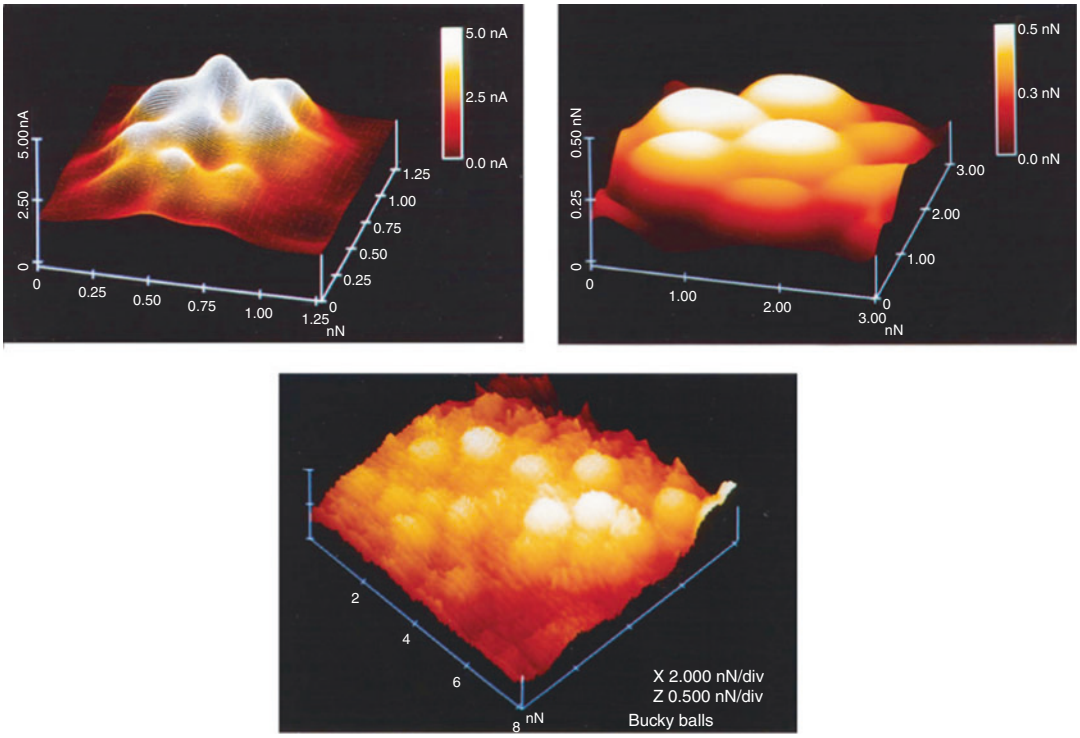
Surface Imaging, Friction, and Adhesion

Surface height imaging down to atomic resolution of electrically conducing surfaces can be carried out using an STM. An AFM can also be used for surface height imaging and roughness characterization down to nanoscale. Figure 2 shows a sequence of STM images at various scan sizes of solvent-deposited C60 film on a 200-nm thick gold-coated freshly cleaved mica [13]. The film consists of clusters of C60 molecules with 8 nm in diameter. The C60 molecules within a cluster appear to pack in a hexagonal array with a spacing of about 1 nm; however, they do not follow any long range order. The measured cage diameter of the C60 molecule is about 0.7 nm, very close to the projected diameter of 0.71 nm.

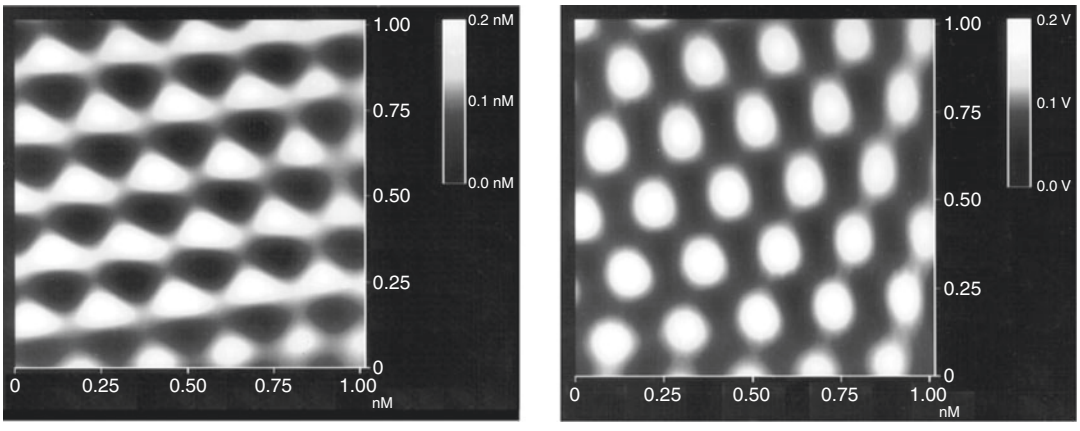
In an AFM measurement during surface imaging, the tip comes in intimate contact with the sample surface and leads to surface deformation with finite tip-sample contact area (typically few atoms). The finite size of contact area prevents the imaging of individual point defects, and only the periodicity of the atomic lattice can be imaged. Figure 3 shows the topography image of freshly cleaved surface of highly oriented pyrolytic graphite (HOPG) [24]. The periodicity of the graphite is clearly observed.

To study friction mechanisms on an atomic scale, a freshly cleaved HOPG has been studied by Mate et al. [21] and Ruan and Bhushan [24]. Figure 4a shows the atomic-scale friction force map (raw data), and Fig. 3 shows the friction force maps (after 2D spectrum filtering with high frequency noise truncated) [24]. Figure 4a also shows line plot of friction force profiles along some crystallographic direction. The actual shape of the friction profile depends upon the spatial location of axis of tip motion. Note that a portion of atomic-scale lateral force is conservative. Mate et al. [21] and Ruan and Bhushan [24] reported that the average friction force linearly increased with normal load and was reversible with load. Friction profiles were similar during sliding of the tip in either direction.

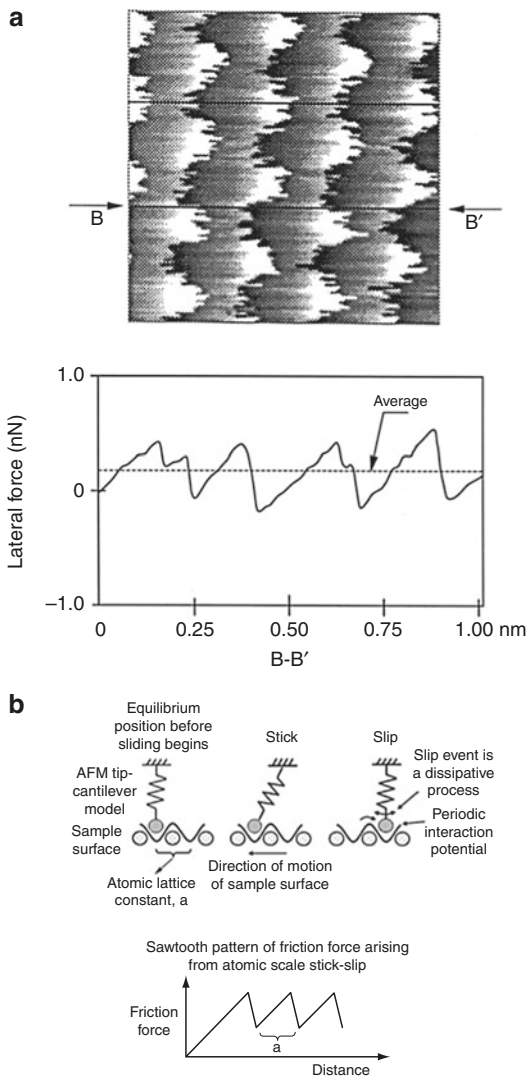
During scanning, the tip moves discontinuously over the sample surface and jumps with discrete steps from one potential minimum (well) to the next. This leads to a saw-tooth-like pattern for the lateral motion (force) with a periodicity of the lattice constant. This motion is called stick–slip movement of the tip [5, 8, 12, 21, 23]. The observed friction force includes two components – conservative and periodic, and nonconservative and constant. If the relative motion of the sample and tip were simply that of two rigid collections of atoms, the effective force would be a conservative force oscillating about zero. Slow reversible elastic deformation would



Nanotribology, Fig. 2 STM images of solvent deposited C_{60} film on a *gold-coated* freshly-cleaved mica at various scan sizes



Nanotribology, Fig. 3 *Gray*-scale plots of surface topography and friction force maps (2D spectrum filtered), measured simultaneously, of a 1 × 1 nm area of freshly cleaved HOPG, showing the atomic-scale variation of topography and friction



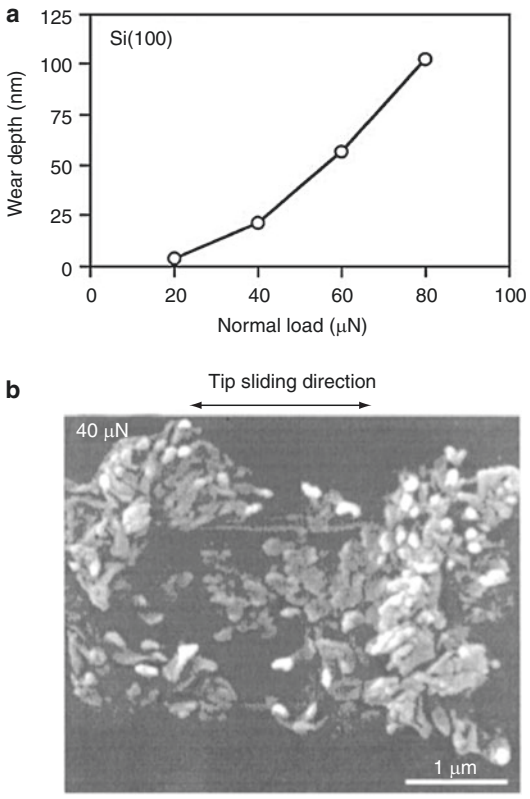
Nanotribology, Fig. 4 (**a**) *Gray* scale plot of friction force map (raw data) of a 1×1 nm^2 area of freshly cleaved HOPG, showing atomic-scale variation of friction force. High points are shows by *lighter color*. Also shown is line plot of friction force profile along the line indicated by *arrows*. The normal load was 25 nN and the cantilever normal stiffness was 0.4 N/m [23], and (**b**) schematic of a model for a tip atom sliding on an atomically flat periodic surface. The schematic shows the tip jumping from one potential minimum to another, resulting in stick–slip behavior

also contribute to conservative force. The origin of the nonconservative direction-dependent force component would be phonon generation, viscous dissipation, or plastic deformation.

The stick–slip on the atomic scale discussed above is the result of the energy barrier required to be overcome for jumping over the atomic corrugations on the sample surface. It corresponds to the energy required for the jump of the tip from a stable equilibrium position on the surface into a neighboring position. The perfect atomic regularity of the surface guarantees the periodicity of the lateral force signal, independent of actual atomic structure of tip apex. Few atoms (based on the magnitude of the friction force, less than 10), on a tip sliding over an array of atoms on the sample, are expected to go through the stick–slip. For simplicity, Fig. 4b shows a simplified model for one atom on a tip with a one-dimensional spring mass system. As the sample surface slides against the AFM tip, the tip remains "stuck" initially until it can overcome the energy (potential) barrier, which is illustrated by a sinusoidal interaction potential as experienced by the tip. After some motion, there is enough energy stored in the spring which leads to "slip" into the neighboring stable equilibrium position. During the slip and before attaining stable equilibrium, stored energy is converted into vibrational energy of the surface atoms in the range of 1,013 Hz (phonon generation) and decays within the range of 10–11 s into heat. (A wave of atoms vibrating in concert is termed a phonon.) The stick–slip phenomenon, resulting from irreversible atomic jumps, can be theoretically modeled with classical mechanical models [27, 28]. The Tomanek-Zhong-Thomas model [27] is the starting point for determining friction force during atomic scale stick–slip. The AFM model describes the total potential as the sum of the potential acting on the tip due to interaction with the sample and the elastic energy stored in the cantilever. Thermally activated stick–slip behavior can explain the velocity effects on friction.

Wear

By scanning the sample in two dimensions with the AFM, wear scars are generated on the surface. Figure 5a shows the effect of normal load on wear depth on Si(100). Note that wear depth is very small below 20 μN of normal load [29]. A normal load of 20 μN corresponds to contact stresses



Nanotribology, Fig. 5 (**a**) Wear depth as a function of normal load for Si(100) after one cycle [29], and (**b**) secondary electron image of wear mark and debris for Si(100) produced at a normal load of 40 μN and one scan cycle

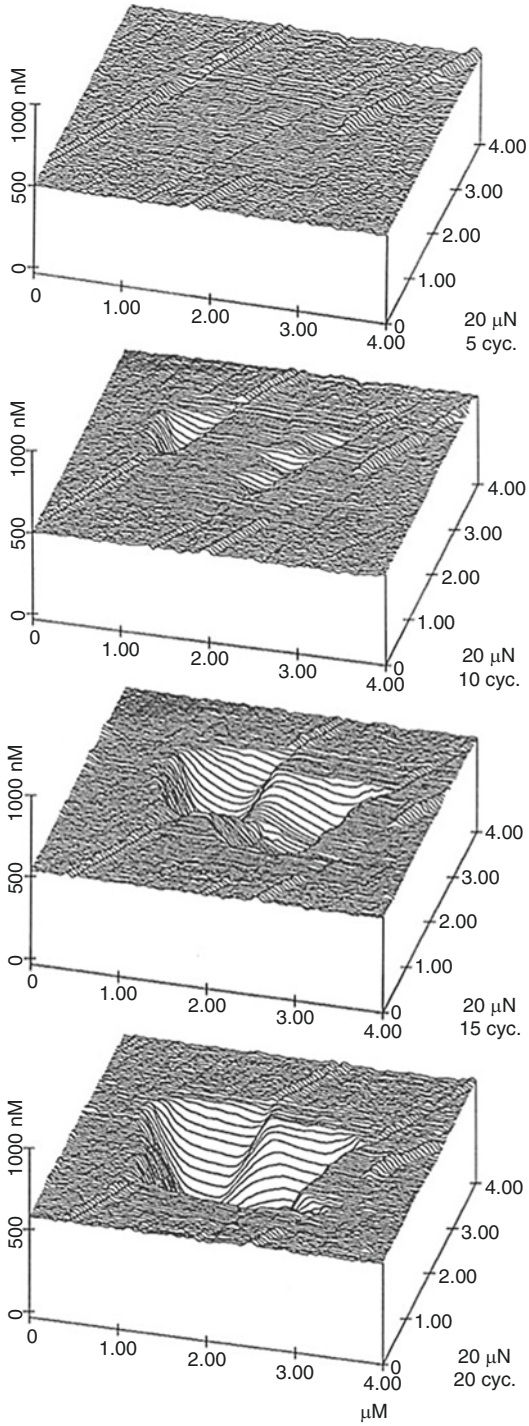
comparable to the hardness of silicon. Primarily, elastic deformation at loads below 20 μN is responsible for low wear. Uniform material removal at the bottom of the wear mark has been reported. An AFM image of the wear mark shows small debris at the edges, probably swiped during AFM scanning. This indicates that the debris is loose (not sticky) and can be removed during the AFM scanning.

Next, the mechanisms of material removal were examined on a microscale in AFM wear experiments [29]. Figure 5b shows a secondary electron image of the wear mark and associated wear particles. The specimen used for the scanning electron microscope (SEM) was not scanned with the AFM after initial wear, in order to retain wear debris in the wear region. Wear debris is clearly observed. In the SEM micrographs, the wear debris appears agglomerated because of high surface energy of the fine particles. Particles appear to be a mixture of rounded and so-called cutting type (feather-like or ribbon-like material). Zhao and Bhushan [29] reported an increase in the number and size of cutting type particles with the normal load. The presence of cutting type particles indicates that the material is removed primarily by plastic deformation.

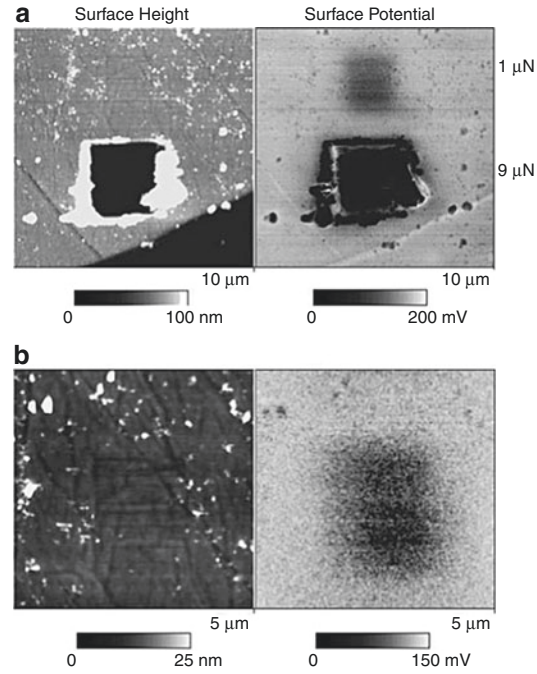
To better understand the material removal mechanisms, Zhao and Bhushan [29] used transmission electron microscopy (TEM). The existence of dislocation arrays observed by them confirms that material removal occurs by plastic deformation. It is concluded that the material on microscale at high loads is removed by plastic deformation with a small contribution from elastic fracture.

To understand wear mechanisms, evolution of wear can be studied using AFM. Figure 6 shows evolution of wear marks of a DLC-coated disk sample. The data illustrate how the microwear profile for a load of 20 μN develops as a function of the number of scanning cycles [14]. Wear is not uniform, but is initiated at the nanoscratches. Surface defects (with high surface energy) present at nanoscratches act as initiation sites for wear. Coating deposition also may not be uniform on and near nanoscratches which may lead to coating delamination. Thus, scratch-free surfaces will be relatively resistant to wear.

Wear precursors (precursors to measurable wear) can be studied by making surface potential measurements [17]. The contact potential difference, or simply the surface potential between two surfaces depends on a variety of parameters such as electronic work function, adsorption, and oxide layers. The surface potential map of an interface gives a measure of changes in the work function which is sensitive to both physical and chemical conditions of the surfaces including structural and chemical changes. Before material is actually removed in a wear process, the surface experiences stresses that result in surface and subsurface changes of structure and/or chemistry. These can cause changes in the measured potential of a surface. An AFM tip allows mapping of surface potential with nanoscale resolution.



Nanotribology, Fig. 6 Surface plots of diamond-like carbon-coated thin-film disk showing the worn region; the normal load and number of test cycles are indicated [14]



Nanotribology, Fig. 7 (**a**) Surface height and change in surface potential maps of wear regions generated at 1 μN (*top*) and 9 μN (*bottom*) on a single crystal aluminum sample showing bright contrast in the surface potential map on the worn regions. (**b**) Close up of *upper* (low load) wear region [17]

Surface height and change in surface potential maps of a polished single-crystal aluminum (100) sample, abraded using a diamond tip at loads of 1 and 9 μN, are shown in Fig. 7a (Note that the sign of the change in surface potential is reversed here from that in DeVecchio and Bhushan [17]). It is evident that both abraded regions show a large potential contrast (∼0.17 V), with respect to the non-abraded area. The black region in the lower right-hand part of the topography scan shows a step that was created during the polishing phase. There is no potential contrast between the high region and the low region of the sample, indicating that the technique is independent of surface height. Figure 7b shows a close up scan of the upper (low load) wear region in Fig. 7a. Notice that while there is no detectable change in the surface topography, there is nonetheless, a large change in the potential of the surface in the worn region. Indeed, the wear

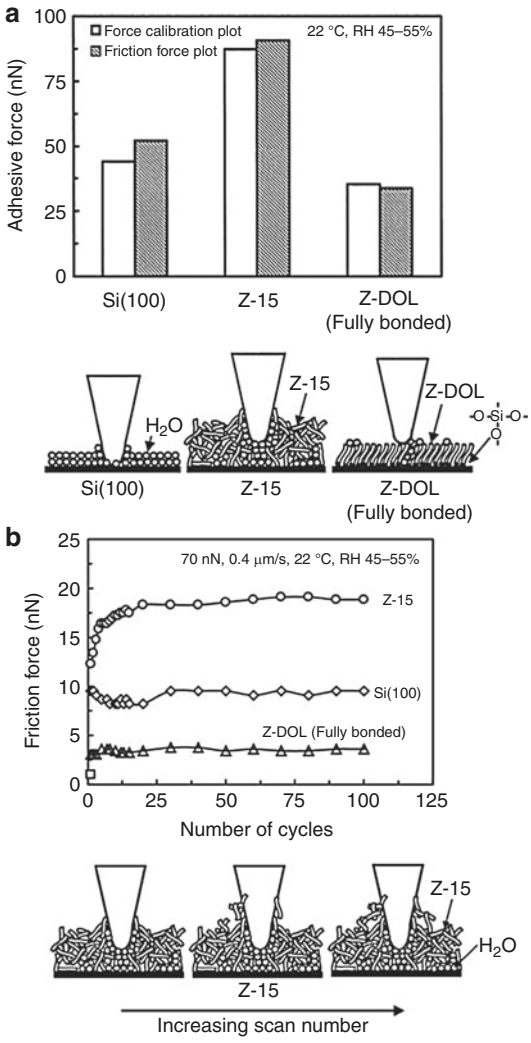
mark of Fig. 7b might not be visible at all in the topography map were it not for the noted absence of wear debris generated nearby and then swept off during the low load scan. Thus, even in the case of zero wear (no measurable deformation of the surface using AFM), there can be a significant change in the surface potential inside the wear mark which is useful for the study of wear precursors. It is believed that the removal of the thin contaminant layer including the natural oxide layer gives rise to the initial change in surface potential. The structural changes, which precede generation of wear debris and/or measurable wear scars, occur under ultra-low loads in the top few nanometers of the sample, and are primarily responsible for the subsequent changes in surface potential.

Boundary Lubrication

The classical approach to lubrication uses freely supported multimolecular layers of liquid lubricants [1, 5, 8, 16]. The liquid lubricants are sometimes chemically bonded to improve their wear resistance [1, 5, 8]. Partially chemically-bonded, molecularly-thick perfluoropolyether (PFPE) films are used for lubrication of magnetic storage media because of their thermal stability and extremely low vapor pressure [1]. Chemically-bonded lubricants are considered as potential candidate lubricants for MEMS/NEMS. Molecularly-thick PFPEs are well suited because of the following properties: low surface tension and low contact angle which allow easy spreading on surfaces and provide hydrophobic properties; chemical and thermal stability which minimize degradation under use; low vapor pressure which provides low out-gassing; high adhesion to substrate via organic functional bonds; and good lubricity which reduces contact surface wear.

For boundary lubrication studies, friction, adhesion, and durability experiments have been performed on virgin Si (100) surfaces and silicon surfaces lubricated with various PFPE lubricants [19, 26]. Results of two of the PFPE lubricants will be presented here, which are –Z-15 (with -CF_3 nonpolar end groups), CF_3–O–(CF_2–CF_2–O)$_m$–(CF_2–O)$_n$–CF_3 (m/n ∼ 2/3) and Z-DOL (with –OH polar end groups), HO–CH_2–CF_2–O–(CF_2–CF_2–O)$_m$–(CF_2–O)$_n$–CF_2–CH_2–OH (m/n ∼ 2/3). Z-DOL film was thermally bonded at 150 °C for 30 min and unbonded fraction was removed by a solvent (BW) [1]. The thicknesses of Z-15 and Z-DOL films were 2.8 and 2.3 nm, respectively. Lubricant chain diameters of these molecules are about 0.6 nm and molecularly thick film generally lie flat on surfaces with high coverage.

The adhesive forces of Si(100), Z-15, and Z-DOL (BW) measured by force calibration plot and friction force versus normal load plot are summarized in Fig. 8a [20]. The data obtained by these two methods are in good agreement. Figure 8(a) shows that the presence of mobile Z-15 lubricant film increases the adhesive force as compared to that of Si(100) by meniscus formation. Whereas, the presence of solid-like phase of the Z-DOL (BW) film reduces the adhesive force as compared to that of Si(100) because of the absence of mobile liquid. The schematic (bottom) in Fig. 8a shows relative size and sources of meniscus. It is well known that the native oxide layer (SiO_2) on the top of Si(100) wafer exhibits hydrophilic properties, and some water molecules can be adsorbed on this surface. The condensed water will form meniscus as the tip approaches the sample surface. The larger adhesive force in Z-15 is not only caused by the Z-15 meniscus alone, the non-polarized Z-15 liquid does not have good wettability and strong bonding with Si(100). Consequently, in the ambient environment, the condensed water molecules from the environment will permeate through the liquid Z-15 lubricant film and compete with the lubricant molecules present on the substrate. The interaction of the liquid lubricant with the substrate is weakened, and a boundary layer of the liquid lubricant forms puddles. This dewetting allows water molecules to be adsorbed on the Si(100) surface along with Z-15 molecules. And both of them can form meniscus while the tip approaches to the surface. Thus the dewetting of liquid Z-15 film results in higher adhesive force and poorer lubrication performance. In addition, as the Z-15 film is pretty soft as compared to the solid Si(100) surface, and penetration of the tip in the film occurs while pushing tip down. This results in the large area of the tip wetted by the liquid to form the meniscus at the tip-liquid (mixture of Z-15 and water)



Nanotribology, Fig. 8 (**a**) Summary of the adhesive forces of Si(100) and Z-15 and Z-DOL (BW) films measured by force calibration plots and friction force versus normal load plots in ambient air. The schematic (*bottom*) showing the effect of meniscus, formed between AFM tip and the surface sample, on the adhesive and friction forces [19], and (**b**) Friction force versus number of sliding cycles for Si(100) and Z-15 and Z-DOL (BW) films at 70 nN, 0.4 μm/s, and in ambient air. Schematic (*bottom*) shows that some liquid Z-15 molecules can be attached onto the tip. The molecular interaction between the attached molecules onto the tip with the Z-15 molecules in the film results in an increase of the friction force with multi scanning [20]

interface. It should also be noted that Z-15 has a higher viscosity compared to water, therefore Z-15 film provides higher resistance to motion and coefficient of friction. In the case of Z-DOL (BW) film, both of the active groups of Z-DOL molecules are mostly bonded on Si(100) substrate, thus the Z-DOL (BW) film has low free surface energy and cannot be displaced readily by water molecules or readily adsorb water molecules. Thus, the use of Z-DOL (BW) can reduce the adhesive force.

To study the durability of lubricant films at nanoscale, the friction of Si(100), Z-15, and Z-DOL (BW) as a function of the number of scanning cycles are shown in Fig. 8b [19]. As observed earlier for adhesive force, friction force of Z-15 is higher than that of Si(100) with the lowest values for Z-DOL(BW). During cycling, friction force of Si(100) shows a slight decrease during initial few cycles then remains constant. This is related to the removal of the native oxide layer. In the case of Z-15 film, the friction force shows an increase during the initial few cycles and then approaches to higher and stable values. This is believed to be caused by the attachment of the Z-15 molecules onto the tip. After several scans, the molecular interaction reaches an equilibrium and after that the friction force remains constant. In the case of Z-DOL (BW) film, the friction force starts out to be low and remains low during the entire test for 100 cycles. It suggests that Z-DOL (BW) molecules do not get attached or displaced as readily as Z-15.

Closure

For most solid-solid interfaces of technological relevance, contact occurs at multiple asperities. A sharp AFM/FFM tip sliding on a surface simulates just one such contact. However, asperities come in all shapes and sizes. AFM/FFM are used to study various tribological phenomena, which include surface roughness, adhesion, friction, scratching, wear, indentation, detection of material transfer, and boundary lubrication. Measurement of atomic-scale friction of a freshly-cleaved highly-oriented pyrolytic graphite exhibits the same periodicity as that of the corresponding topography. Relevant friction mechanism is atomic-scale stick–slip.

Wear rate on the microscale for single-crystal silicon is negligible below 20 μN and is much higher and remains approximately constant at

higher loads. Elastic deformation at low loads is responsible for negligible wear. Most of the wear debris is loose. SEM and TEM studies of the wear region suggest that the material on the microscale is removed by plastic deformation with a small contribution from elastic fracture; this observation corroborates with the scratch data. Evolution of wear has also been studied using AFM. Wear is found to be initiated at nanoscratches. For a sliding interface requiring near-zero friction and wear, contact stresses should be below the hardness of the softer material to minimize plastic deformation and surfaces should be free of nanoscratches. Further, wear precursors can be detected at early stages of wear by using surface potential measurements. It is found that even in the case of zero wear (no measurable deformation of the surface using AFM), there can be a significant change in the surface potential inside the wear mark which is useful for study of wear precursors.

Boundary lubrication studies and measurement of lubricant-film thickness with a lateral resolution on a nanoscale can be conducted using AFM. For chemically bonded lubricant films, the adsorption of water, the formation of meniscus and its change during sliding, and surface properties play an important role on the adhesion, friction, and durability of these films.

Investigations of adhesion, friction, wear, scratching and indentation on nanoscales using the AFM can provide insights into failure mechanisms of materials. Coefficients of friction, wear rates and mechanical properties such as hardness have been found to be different on the nanoscale than on the macroscale; generally, coefficients of friction and wear rates on micro- and nanoscales are smaller, whereas hardness is greater. Therefore, micro/nanotribological studies may help define the regimes for ultra-low friction and near zero wear. These studies also provide insight to atomic origins of adhesion, friction, wear, and lubrication mechanisms.

Cross-References

► Atomic Force Microscopy

References

1. Bhushan, B.: Tribology and Mechanics of Magnetic Storage Devices, 2nd edn. Springer, New York (1996)
2. Bhushan, B.: Micro/Nanotribology and Its Applications, vol. E330. Kluwer, Dordrecht (1997)
3. Bhushan, B.: Tribology Issues and Opportunities in MEMS. Kluwer, Dordrecht (1998)
4. Bhushan, B.: Handbook of Micro/Nanotribology, 2nd edn. CRC Press, Boca Raton (1999)
5. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
6. Bhushan, B.: Modern Tribology Handbook. Principles of Tribology, vol. 1. CRC Press, Boca Raton (2001)
7. Bhushan, B.: Fundamentals of Tribology and Bridging the Gap Between the Macro- and Micro/Nanoscales. NATO Science Series II, vol. 10. Kluwer, Dordrecht (2001)
8. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
9. Bhushan, B.: Adhesion and stiction: mechanisms, measurement techniques, and methods for reduction. J. Vac. Sci. Technol. B **21**, 2262–2296 (2003). invited
10. Bhushan, B.: Nanotribology, nanomechanics and nanomaterials characterization. Philos. Trans. R. Soc. A **366**, 1351–1381 (2008)
11. Bhushan, B.: Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
12. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
13. Bhushan, B., Ruan, J., Gupta, B.K.: A scanning tunnelling microscopy study of fullerene films. J. Phys. D Appl. Phys. **26**, 1319–1322 (1993)
14. Bhushan, B., Koinkar, V.N., Ruan, J.: Microtribology of magnetic media. P. I. Mech. Eng. J.-J. Eng. Tribol. **208**, 17–29 (1994)
15. Bhushan, B., Israelachvili, J.N., Landman, U.: Nanotribology: friction, wear and lubrication at the atomic scale. Nature **374**, 607–616 (1995)
16. Bowden, F.P., Tabor, D.: The Friction and Lubrication of Solids, Part 1. Clarendon, Oxford (1950)
17. DeVecchio, D., Bhushan, B.: Use of a nanoscale Kelvin probe for detecting wear precursors. Rev. Sci. Instrum. **69**, 3618–3624 (1998)
18. Guntherodt, H.J., Anselmetti, D., Meyer, E.: Forces in Scanning Probe Methods, vol. E286. Kluwer, Dordrecht (1995)
19. Liu, H., Bhushan, B.: Adhesion and friction studies of microelectromechanical systems/nanoelectromechanical systems materials using a novel microtriboapparatus. J. Vac. Sci. Technol. A **21**, 1528–1538 (2003)
20. Liu, H., Bhushan, B.: Nanotribological characterization of molecularly-thick lubricant films for applications to MEMS/NEMS by AFM. Ultramicroscopy **97**, 321–340 (2003)

21. Mate, C.M., McClelland, G.M., Erlandsson, R., Chiang, S.: Atomic-scale friction of a tungsten tip on a graphite surface. Phys. Rev. Lett. **59**, 1942–1945 (1987)
22. Persson, B.N.J., Tosatti, E.: Physics of Sliding Friction, vol. E311. Kluwer, Dordrecht (1996)
23. Ruan, J., Bhushan, B.: Atomic-scale and microscale friction of graphite and diamond using friction force microscopy. J. Appl. Phys. **76**, 5022–5035 (1994)
24. Ruan, J., Bhushan, B.: Frictional behavior of highly oriented pyrolytic graphite. J. Appl. Phys. **76**, 8117–8120 (1994)
25. Singer, I.L., Pollock, H.M.: Fundamentals of Friction: Macroscopic and Microscopic Processes, vol. E220. Kluwer, Dordrecht (1992)
26. Tao, Z., Bhushan, B.: Bonding, degradation, and environmental effects on novel perfluoropolyether lubricants. Wear **259**, 1352–1361 (2005)
27. Tomanek, D., Zhong, W., Thomas, H.: Calculation of an atomically modulated friction force in atomic force microscopy. Europhys. Lett. **15**, 887–892 (1991)
28. Tomlinson, G.A.: A molecular theory of friction. Phil. Mag. Ser. **7**, 905–939 (1929)
29. Zhao, X., Bhushan, B.: Material removal mechanism of single-crystal silicon on nanoscale and at ultralow loads. Wear **223**, 66–78 (1998)

Nanotribology Applications in Microprojector Technology

Satish C. Chaparala and Vikram Bhatia
Science and Technology, Corning Incorporated
SP-PR-02-1, Corning, NY, USA

Definition

Microprojectors are being developed to project a larger image from a smaller display in handheld electronic devices such as cell phone, handheld PDA, etc. The technology associated with microprojectors has several applications of nanotribology and those are described in this entry.

Introduction to Microprojector Technology

The capacity of high-speed wireless networks has enabled carriers to provide multimedia services to consumers such as web browsing, live television and gaming, etc. At the same time, the electronics industry is driving toward miniaturization which in turn led to the smaller size of displays on the electronic devices. The smaller display limits the full utilization and enjoyment of these services. Microprojectors have been proposed to address this issue. These devices can either be as an accessory to any electronic device or an embedded component within a mobile device. These are expected to provide images of the size of A4 paper and larger in indoor lighting conditions. The challenge is to develop the correct microprojector configuration to balance size, cost, performance, and efficiency [1]. For instance, currently, the thickness of the projector could be in the range of 7–14 mm that results in overall projector volume between 10 and 20 cm^3. Brightness is obviously a very important attribute of a microprojector. While the current range of products offers 5–10 lm of brightness, 50 lm is required to provide enough brightness in all lighting conditions. Focus-free operation is another desirable attribute and this is possible with the projectors that have lasers as light sources. Resolution is expected to continue to grow as product technology improves. The wide-screen format is typically acceptable for viewing video content. Similar to brightness, contrast is also an important attribute that makes the difference between washed-out images and crisp, dramatic-looking images [2]. These trade-offs are primarily driven by the choices of light source and imaging mechanism.

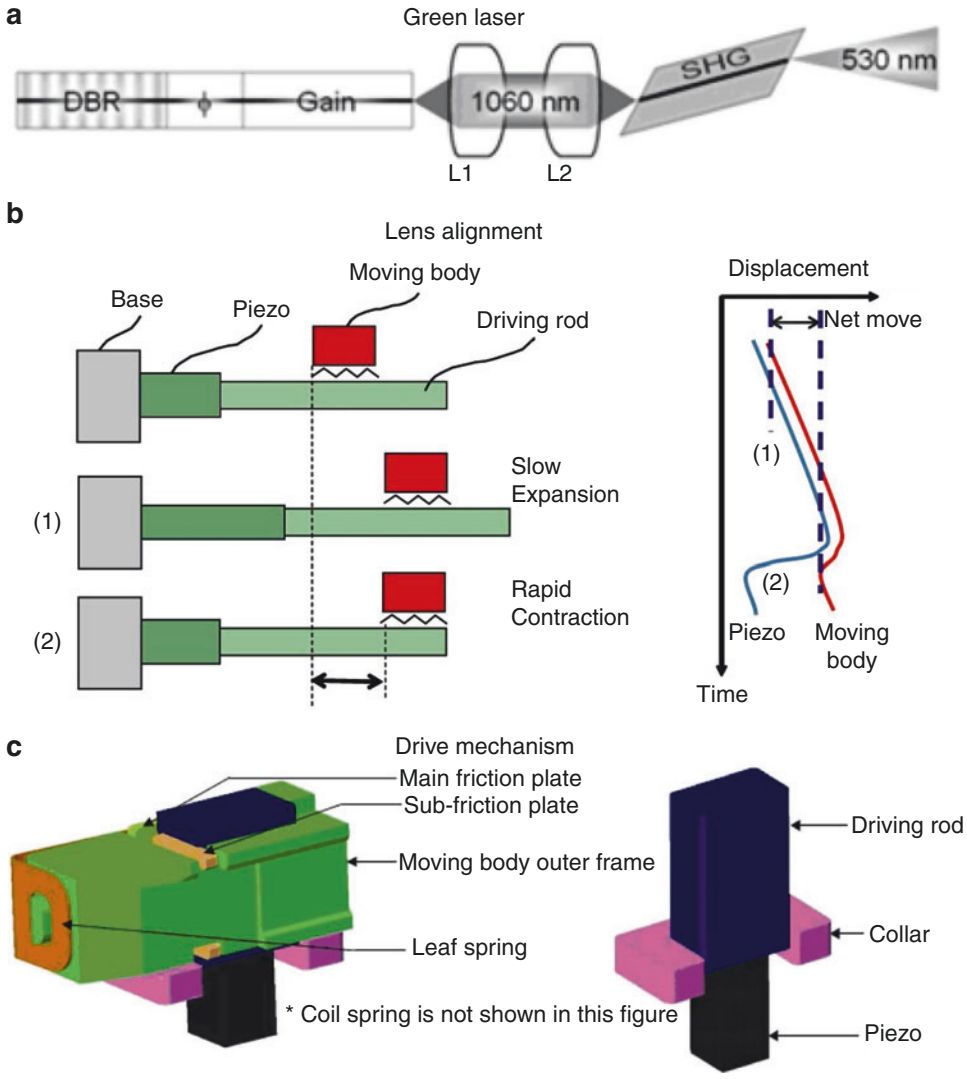
The microprojector technology has two major components in the optical engine: (1) light source and (2) image-producing technology. Optical engines are modules that combine the red-green-blue (RGB) light sources and an imaging element or elements. Optical engines accept an electronic input video signal and project an optical image. Optical engines do not contain power sources or any other functional elements such as wireless communication chips. Three primary colors, namely, red, blue, and green are required to create full color images. The light sources in the projection technology would be semiconductor devices that emit these colors. These devices could be either light-emitting diodes (LEDs) or lasers.

Like a normal diode, when an LED is forward biased or, in other words, switched on, electron combines with a hole, falls into a lower energy level, thereby releasing the energy in the form of photon. The color of the emitted light depends on the energy gap of the semiconductor. Red, blue, and green LEDs are commercially available. LEDs have several advantages that include low energy consumption, longer lifetime, smaller size, faster switching, and good reliability, typically lasting a few years. However, LEDs have higher etendue. Etendue is a property of an optical system that describes how "diverged" the light is in area and angle. Because of high etendue of LEDs, large optical elements (e.g., lenses, etc.) have to be used in the package to collect all the light that is emitted from LED, which in turn increases the size of the projector. If the projector volume is designed to be small using smaller optical elements, some of the optical power (larger angles) will be lost, which in turn decreases the conversion efficiency of the package. Lasers can be used to overcome this problem. A laser is a device that emits light through a process of optical amplification based on the stimulated emission of photons. Laser light is spatially and temporally coherent. Red and blue lasers are commercially available for applications to the optical read/write products such as DVD players, whereas green has been a challenge to produce through currently available processes and materials. A green laser is often referred to as the "keystone" component to enable laser-based mobile projection. Extensive research has been conducted to develop native green lasers using either III-V or II-VI compound semiconductors. Recent technological advancements have led to creation of native green laser light by Sumitomo, Osram, and others. For instance, Sumitomo succeeded in producing 2.4 mW of optical power at 520 nm at room temperature continuous wave lasing of InGaN-based green LEDs on semipolar GaN substrates [3]. Native green lasers are easier to control and could potentially demonstrate greater temperature stability, a smaller form factor, and higher modulation capability at several 100 MHz. However, there has not been sufficient reliability data of these devices under operating conditions. While the native green lasers are under development, a synthetic green laser has already been developed based on frequency doubling of 1,060 nm IR light using second harmonic generation crystal by various companies such as Corning [1, 4–6].

The second critical component of a microprojector is image-producing technology. There are three main image-producing technologies: (1) scanners, (2) digital light processing (DLP), and (3) liquid crystal on silicon (LCOS). Scanners are based on two-dimensional pixel-by-pixel scanning, typically using a single MEMS mirror and uses modulated laser beams as the light source. The image remains in focus at any projection distance and the projector consumes relatively low power [2, 7]. DLP uses millions of tiny mirrors, one for each pixel, to reflect light. This technology involves modulating the image by tilting the mirrors either into or away from the lens path. It is a "reflective" technology and can use either lasers or LEDs as light sources. DLP projection technology using LED as light source is already commercialized. LCOS combines the idea of a DLP which is a reflective technology but uses liquid crystals instead of individual mirrors. In LCOS, liquid crystals are applied to a reflective mirror substrate. As the liquid crystals are modulated based on the incoming video signal, the light is either reflected from the mirror below or blocked. This modulates the light from either laser or LED sources and creates the image. Any feasible combination of these three technologies with either lasers or LEDs could project an image. Finally, all these components can be either wrapped in a case (skin) and sold as a companion projector or embedded in a handset. Projector contains an optical engine, power source, and "industrial design or skin."

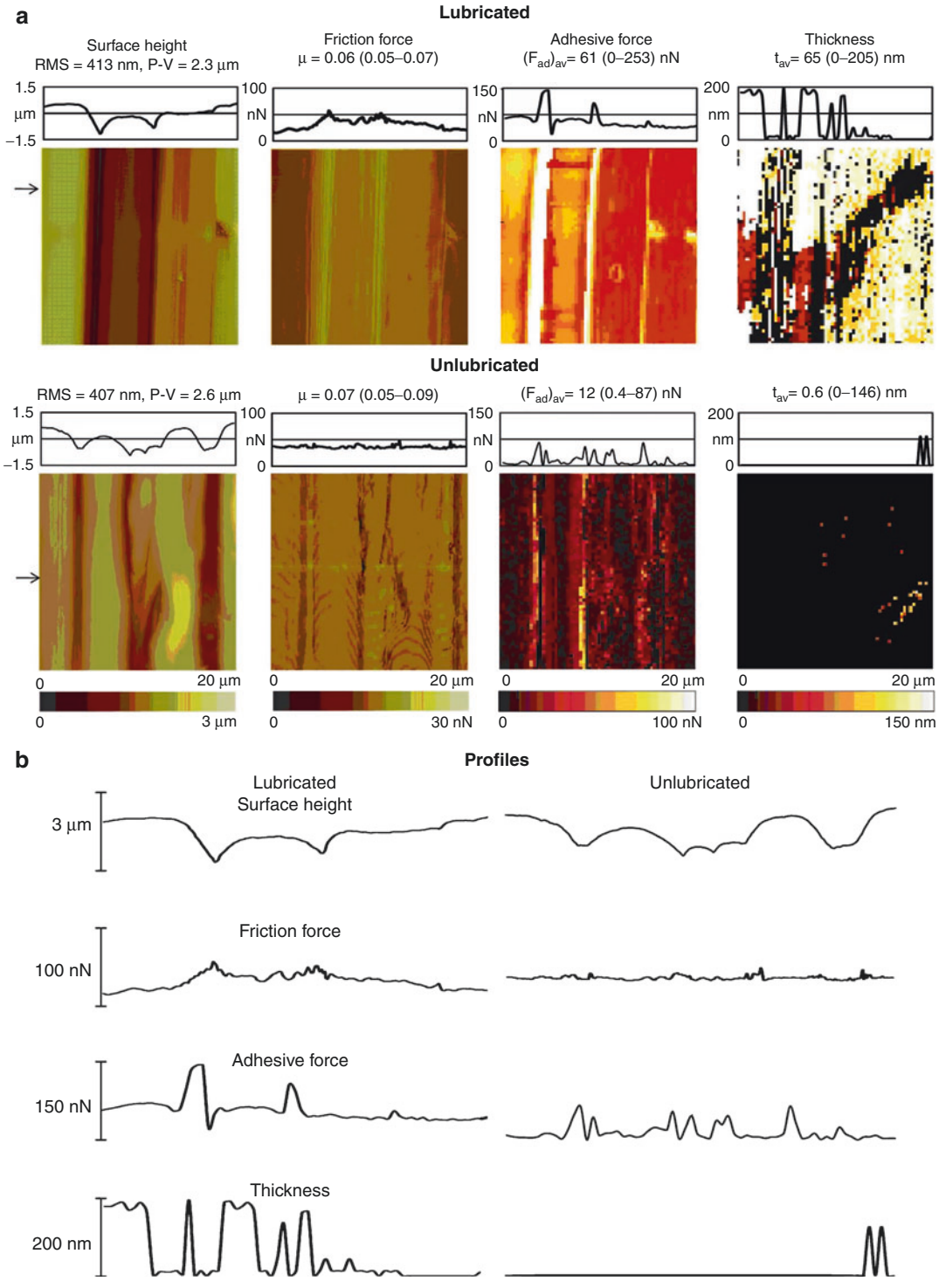
Synthetic Green Laser Package

Until now, a detailed introduction to various projection technologies and components involved in the microprojector product has been provided. The next focus of this entry would be on the keystone component of a laser-based microprojector, which is a green laser optoelectronic



Nanotribology Applications in Microprojector Technology, Fig. 1 Schematic of (**a**) the green laser architecture, (**b**) the operation, and (**c**) the construction of the drive mechanism used in the green laser

package developed by Corning Incorporated and specifically, the role of nano-tribology in its development. Several key tribology design studies will be presented. A frequency doubling technique is used to produce a compact green laser. A distributed bragg reflector (DBR) laser diode generated 1,060 nm wavelength light is passed through second harmonic generating (SHG) crystal resulting in 530 nm green light. Corning's green laser architecture is shown in the Fig. 1 [1, 4, 8]. As can be seen from the figure, there are two lenses in this configuration. The first lens that is in front of the IR laser diode is used to collimate the light beam that emits from the laser diode. The second lens converges the collimated beam onto the waveguide in SHG. Exposure to the operating environment, which can be at a temperature as high as 60 °C and at 75 % relative humidity (RH), can cause lens misalignment and affect optical output. To correct for this, the green laser module uses an adaptive optics component with a drive mechanism to align the optics and



Nanotribology Applications in Microprojector Technology, Fig. 2 (continued)

N

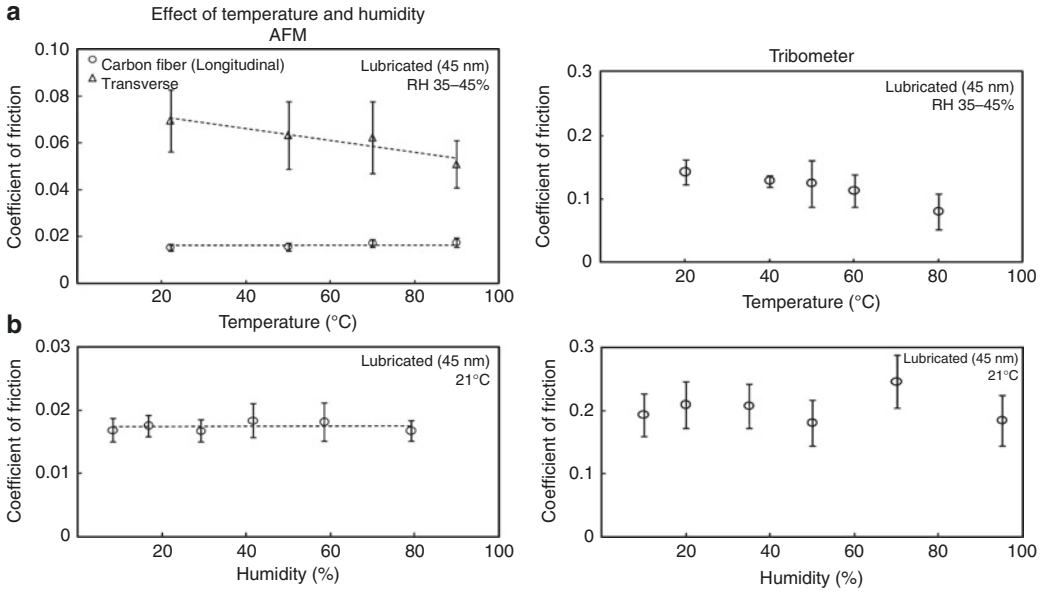
maintain constant power output with time and temperature. Lens alignment along the x- and y-axes is made by using adaptive optics, represented by L_1 and L_2 in the figure below. This adaptive optics component is referred to as smooth impact drive mechanism (SIDM) developed by Konica Minolta. Schematics of the operation and construction of this drive mechanism are shown in Fig. 1. The main components of the drive mechanism are the piezoelectric element, driving rod, moving body, and the friction plate. During operation, a rectangular waveform with a given voltage, frequency, and number of cycles is applied to the piezo element. The rectangular waveform is then converted to a sawtooth motion by the mechanical transfer function of the piezo. On the gentle upward slope, the rod carries the moving body due to friction ("stick" operation), while on the rapid falling slope, the rod "slips" due to inertia. The net displacement is the minimum movement of the drive mechanism. This is repeated until the desired amount of displacement is achieved [4].

Carbon fiber reinforced plastic (CFRP) is used as a component for a wide variety of engineering applications, such as for aircraft structures, automobile, and solar cells, among many others [8–12]. In the drive mechanism, CFRP is the material used as the driving rod. For the friction plate, stainless steel is used. For the moving body, Zn alloy is used as the die-cast part, which is commonly used in automotive applications. Adaptive optics sliding components in microprojectors should be stable at an operating environment up to 60 °C and 75 % relative humidity. To improve tribological performance of the sliding components, perfluoropolyether (PFPE) lubricants are applied to the surfaces of the components. The ideal lubricant should be molecularly thick to protect the surface from wear, easily applicable, able to chemically bond to the surface, and insensitive to environment. PFPE lubricants are known to be most desirable as they have low surface tension, high contact angle, and high adhesion to the substrate allowing easy application and spreading onto the surface as well as providing hydrophobicity [13–19]. Their chemical and thermal stability and low vapor pressure provide low degradation and low outgassing. PFPEs have been extensively investigated, especially in the magnetic disk drive industry. It has been shown that they reduce friction and wear, resulting in lower disk drive failure. It is, therefore, very critical to conduct a comprehensive tribology study to understand the friction mechanisms and effect of temperature, humidity, plasma, etc., on the friction.

Applications of Nano-tribology

The following section presents several key findings from various nano-tribological studies made on the lubricant used in the adaptive optics component. A methodology to measure lubricant film thickness with a nanoscale resolution was developed. Atomic force microscopy (AFM) allows the simulation of a single asperity contact. The morphology measured by AFM is shown in Fig. 2. On the height images of the lubricated and unlubricated samples, the bright long columns are for the individual carbon fibers, and the dark ones for epoxy. The average surface roughness (RMS) of the lubricated CFRP is about 328 nm and that of the unlubricated is 406 nm. Even though both lubricated and unlubricated CFRPs have high average surface roughness, the lubricated sample surface has lower roughness than that of the unlubricated surface. The roughness difference between the lubricated and the unlubricated samples can be attributed to lubricant thickness.

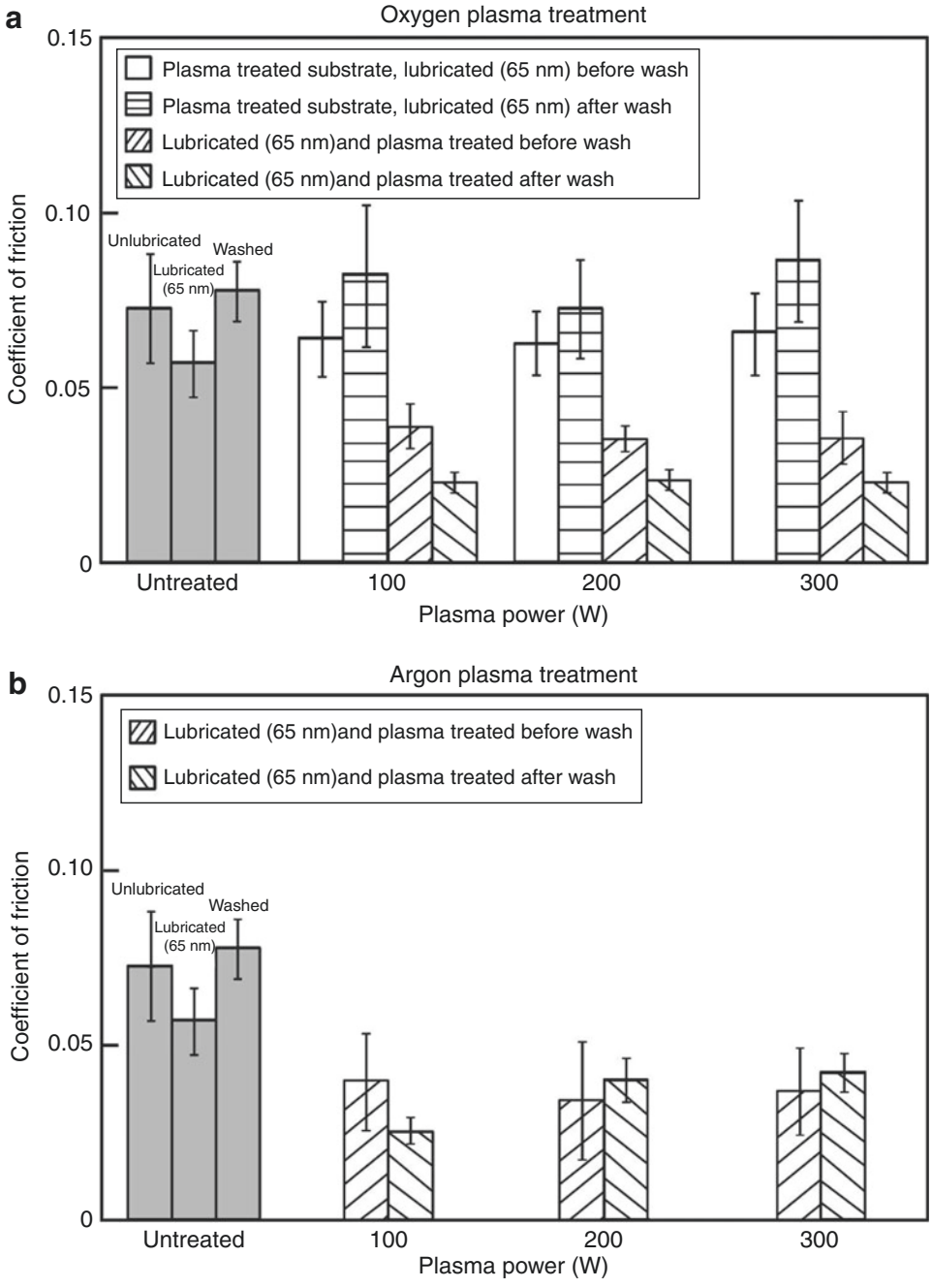
Nanotribology Applications in Microprojector Technology, Fig. 2 (**a**) AFM maps for surface height, friction force, adhesive force, and film thickness distribution for lubricated and unlubricated CFRPs. Shown above each image is a cross section taken at a position denoted by the corresponding *arrows*. (**b**) 2-D cross-sectional profiles to correlate the surface morphology with friction force, adhesive force, and film thickness



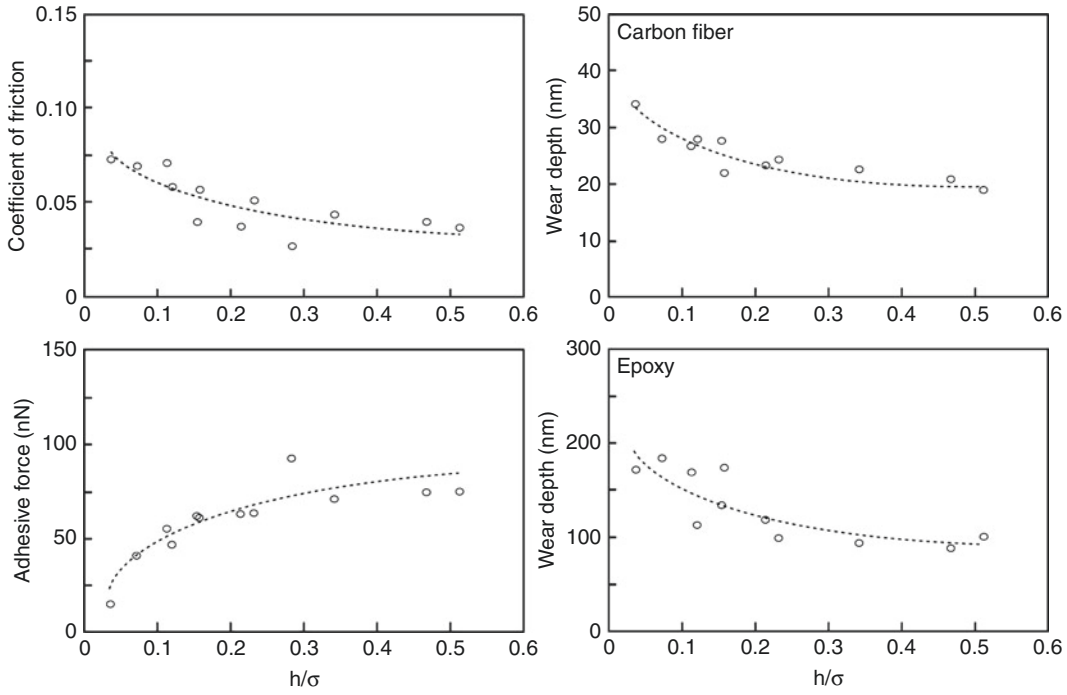
Nanotribology Applications in Microprojector Technology, Fig. 3 Effect of temperature and humidity on lubricated CFRP (average film thickness = 45 nm) using (**a**) AFM on carbon fiber in the longitudinal and transverse directions, and (**b**) tribometer

The degree of bonding was measured by measuring the film thickness before and after washing off the lubricant film. It was found that the lubricant was unbonded to the sample surfaces. Friction force, adhesive force, and lubricant thickness distribution correlate with surface morphology. Lubricant resides in the recessed epoxy regions which act as a lubricant reservoir. It is expected that lubricant will wick up during the operation and get transferred. Coefficient of friction on lubricated samples is lower than that on the unlubricated samples. The epoxy has a higher friction force than carbon fiber because of the presence of the lubricant. Wear depth of epoxy is found to be higher than on the carbon fiber. Effects of composite components, scan direction, length scale, temperature, and humidity were studied. Scanning in the transverse direction includes sliding on both the carbon fiber and the epoxy, and the friction data is close to that of epoxy. Regarding length scale effects, the coefficient of friction on larger length scales is higher due to larger contact area. Figure 3 shows the effect of humidity and temperature. Humidity does not affect the coefficient of friction due to the condensed water layer from the environment which is expected to be thin compared to lubricant thickness. However, increasing temperature leads to viscosity drop of the lubricant, resulting in the decrease in the coefficient of friction.

Thermal, UV, and oxygen and argon plasma treatments in order to allow chemical bonding of lubricant were studied. It was found that the effect of thermal treatment on the lubricant chemical bonding was not obvious. However, the thermal treatment reduced adhesive force and coefficient of friction because of solvent evaporation. For UV treatment, it was found that coefficient of friction, adhesive force, and lubricant thickness are unchanged. UV treatment does not affect chemical bonding of lubricant. For plasma treatment, it is observed that lubricant on the lubricated and subsequently plasma-treated samples is bonded to the surfaces. The plasma treatment also improves coefficient of friction and adhesive force. Lubricant film thickness was reduced by the plasma treatment. However, no difference is found between oxygen and argon plasma treatments. There is no difference between various power levels in plasma treatments either. It is found that



Nanotribology Applications in Microprojector Technology, Fig. 4 Effect of plasma power on coefficient of friction. (**a**) Oxygen and (**b**) argon plasma treatments were carried out. The first three *dark columns* are coefficient of friction on untreated samples. The next groups of columns are for plasma-treated samples. The *white blank* and *horizontally lined columns* in (**a**), which are not in (**b**), are for sample type #1. *Scatter bars* represent $\pm \sigma$ limits

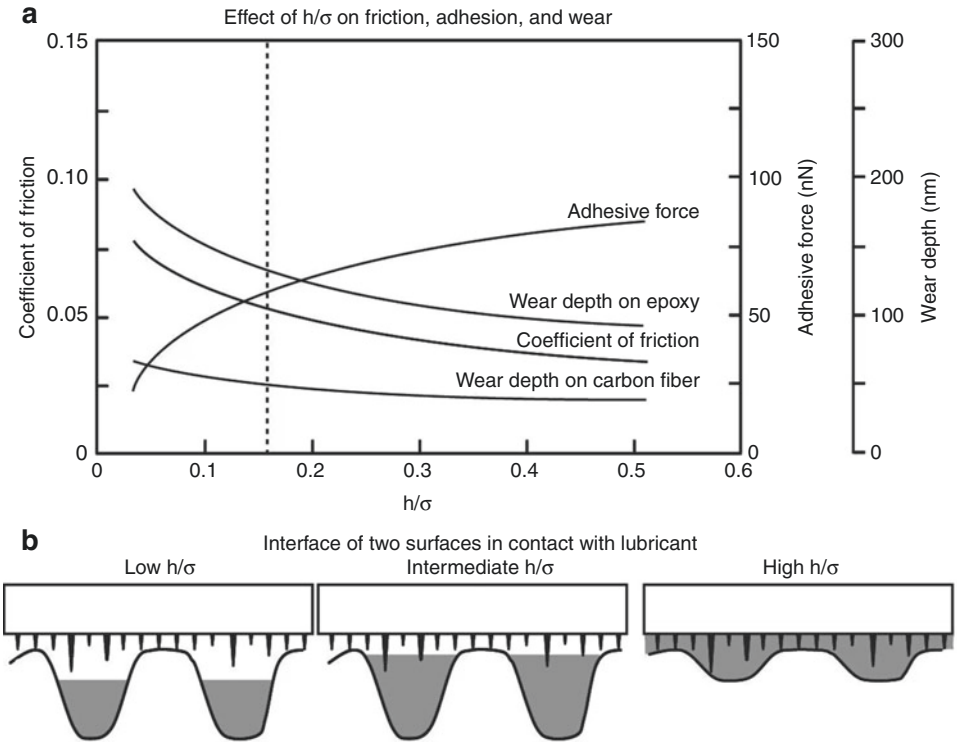


Nanotribology Applications in Microprojector Technology, Fig. 5 The effect of h/σ on coefficient of friction, adhesive force, and wear depth on carbon fibers and epoxy regions at 30 μN of normal load. The *dotted lines* are the trend lines for the data

chemical bonding of lubricant to the sample surfaces due to plasma treatment helps to protect the carbon fiber areas of the CFRP rods from wear but not epoxy areas (Fig. 4).

Further studies revealed that the thermal and UV treatments do not improve wear for the CFRP rods. There is no measurable change in friction, adhesive force, and lubricant thickness after laser exposure to the samples. Therefore, it is believed that lubricant degradation by green and IR laser exposure does not occur or is negligible. Finally, wettability by contact angle measurement for unlubricated and lubricated CFRP rods was studied to simulate the effect of atmospheric humidity. It is found that, in advancing mode, unlubricated CFRP rods are hydrophilic, but the lubricated rods are hydrophobic. Contact angles increase with lubricant thickness. Therefore, humidity should have less effect on coefficient of friction on the lubricated samples. However, contact angle hysteresis is observed due to the large difference of advancing and receding contact angles. It could be explained by chemical heterogeneity caused by the high roughness and the composite effect of the CFRP rods. Finally, the interplay of surface roughness and film thickness of lubricant is optimized for low adhesion, friction, and wear. The effect of h/σ on friction, adhesion, and durability is studied. With an increase in h/σ, the uniformity of the lubricant distribution on the CFRP rods is improved, resulting in stable and reliable operation of the actual devices. Coefficient of friction and wear depth decrease with a decrease in h/σ, but adhesive force increases due to an increase in the amount of lubricant involved in surface interface and in the contact area between the two mating materials, and vice versa (Figs. 5 and 6).

Therefore, there is a trade-off in lubricant film thickness and surface roughness for optimum values of coefficient of friction, adhesive force, and durability. The ratio of the mobile fraction of lubricant film thickness to sample roughness should be decided by the requirement and functionality of the actual devices.



Nanotribology Applications in Microprojector Technology, Fig. 6 (**a**) The trend of the effect of h/σ on coefficient of friction, adhesive force, and wear depth. The *dotted vertical line* is the h/σ value (0.16) of the samples with 410 nm of RMS and 65 nm of lubricant thickness used in some commercial microprojectors. (**b**) A schematic showing the rough interface of CFRP rods and a moving part of microprojector. The *gray colored* area represents lubricant

Conclusion

An introduction to microprojection technology and the components involved is provided. Specifically, a keystone component in laser-based microprojector, green laser package, developed by Corning Incorporated is considered in this study. Applications of nano-tribology in the development of adaptive optics component used in this package are detailed. The adaptive optics component involves sliding mechanical elements, and lubricant was used in between these elements. The tribological design has to be efficient in order for the adaptive optics component to function consistently. Therefore, a comprehensive study was done on tribological design of the lubricant and the surface characteristics of the sliding components. The design specifications obtained from this study were implemented in the final package design.

Cross-References

- ▶ AFM
- ▶ Nanotribology

References

1. Bhatia, V., Gregorski, S.J., Pikula, D., Chaparala, S.C., Loeber, D.A.S., Gollier, J., Gregorski, J.D., Hempstead, M., Ozeki, Y., Hata, Y., Shibatani, K., Nagai, F., Mori, N., Nakabayashi, Y., Mitsugi, N., Nakano, S.: Efficient and compact green laser for micro-projector applications. J. Soc. Inf. Disp. **17**(3), 1–6 (2009)
2. Freeman, M., Champion, M., Madhavan, S.: Scanned laser pico-projectors. Opt. Photonics News **20**, 29–34 (2009)
3. Yoshizumi, Y., Adachi, M., Enya, Y., Kyono, T., Tokuyama, S., Sumitomo, T., Akita, K., Ikegami, T., Ueno, M., Katayama, K., Nakamura, T.: Continuous-

wave operation of 520 nm green InGaN-based laser diodes on semi-polar {2021} GaN substrates. Appl. Phys. Express **2**, 092101-1–092101-3 (2009)
4. Bhatia, V., Gregorski, S.J., Pikula, D., Chaparala, S.C., Loeber, D.A.S., Gollier, J., Ozeki, Y., Hata, Y., Shibatani, K., Nagai, F., Mori, N., Nakabayashi, Y., Mitsugi, N., Nakano, S.: Efficient and compact green laser incorporating adaptive optics for wide operating temperature range. SID Int. Symp. Digest Tech. Pap. **39**, 962–965 (2008)
5. Bhushan, B., Zhao, Z.: Macroscale and microscale tribological studies of molecularly thick boundary layers of perfluoropolyether lubricants for magnetic thin-film rigid disks. J. Info. Storage Proc. Syst. **1**, 1–21 (1999)
6. Bhushan, B., Lee, H., Chaparala, S.C., Bhatia, V.: Nanolubrication of sliding components in adaptive optics used in microprojectors. Appl. Surf. Sci. **256**, 7545–7558 (2010)
7. Graham-Rowe, D.: Projectors get personal. Nat. Photonics **1**, 677–679 (2007)
8. Bhatia, V., Hempstead, M., Grochosinski, J., Sekiguchi, N., Okada, A., Loeber, D.: Compact and efficient green lasers for mobile projector applications. J. Soc. Inf. Disp. **17**(1), 1–6 (2009)
9. Hancox, N.L.: The use of a torsion machine to measure the shear strength and modulus of unidirectional carbon fiber reinforced plastic composites. J. Mater. Sci. **7**, 1030–1036 (1972)
10. Ogisu, T., Shimanuki, M., Kiyoshima, S., Takeda, N.: A basic study of CFRP laminates with embedded prestrained SMA foils for aircraft structures. J. Intell. Mater. Syst. Struct. **16**, 175–185 (2005)
11. HeroldSchmidt, U., Floeth, E., Last, B., Schafer, W., Zaglauer, H.W.: Quantification of smart composites for the use in aerospace applications. Proc. Soc. Photo-Opt. Instrum. Eng. **3044**, 168–175 (1997)
12. Lee, C.K.: Structure, electrochemical and wear-corrosion properties of electroless nickel-phosphorus deposition on CFRP composites. Mater. Chem. Phys. **114**, 125–133 (2009)
13. Rion, J., Leterrier, Y., Manson, J.A.E., Blairon, J.M.: Ultra-light asymmetric photovoltaic sandwich structures. Compos. A: Appl. Sci. Manuf. **40**, 1167–1173 (2009)
14. Palacio, M., Bhushan, B.: Nanotribological properties of novel lubrications for magnetic tapes. Ultramicroscopy **109**, 980–990 (2009)
15. Tao, Z., Bhushan, B.: Bonding, degradation, and environmental effects on novel perfluoropolyether lubricants. Wear **259**, 1352–1361 (2005)
16. Liu, H., Bhushan, B.: Nanotribological characterization of molecularly thick lubricant films for applications to MEMS/NEMS by AFM. Ultramicroscopy **97**, 321–340 (2003)
17. Koinkar, V.N., Bhushan, B.: Microtribological studies of unlubricated and lubricated surfaces using atomic force/friction force microscopy. J. Vac. Sci. Technol. A **14**, 2378–2391 (1996)
18. Zhao, X., Bhushan, B.: Study on degradation mechanisms of lubricants for magnetic thin film rigid disks. Proc. Inst. Mech. Eng. Part J: J. Eng. Tribol. **215**, 173–181 (2001)
19. Bhushan, B., Lee, H., Chaparala, S.C., Bhatia, V.: Chemical bonding and degradation of ultrathin liquid films for nanolubrication of sliding components in adaptive optics. Colloids Surf. A **369**, 39–52 (2010)

Nanotube Adsorption Sensor

▶ Carbon Nanotube NEMS

Nanotube Force Sensor

▶ Carbon Nanotube NEMS

Nanotube Mass Sensor

▶ Carbon Nanotube NEMS

Nanotube Resonant Sensor

▶ Carbon Nanotube NEMS

Nanotube Resonator

▶ Carbon Nanotube NEMS

Nanotubes

▶ Chemical Vapor Deposition (CVD)
▶ Nanostructures for Energy
▶ Physical Vapor Deposition

Nanotweezers

▶ Nanogrippers

Nano-twinned Materials

▶ Nanosized Nanocrystalline and Nanotwinned Metals

Nanovaccinology

▶ Interaction of the Immune System with Nanoparticles

Nanowire FET Biosensor

▶ Nanostructure Field Effect Transistor Biosensors

Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism

Klaus Michael Indlekofer[1] and Joachim Knoch[2]
[1]Hochschule RheinMain University of Applied Sciences, Institute of Microtechnologies (IMtech), Rüsselsheim, Germany
[2]Institute of Semiconductor Electronics, RWTH Aachen University, Aachen, Germany

Synonyms

Nanowire-based field effect transistor (NWFET); Quantum kinetics; Realtime Green's function approach

Definition

The nonequilibrium Green's function formalism (NEGF) is a quantum kinetic approach to nonequilibrium electronic transport in nanoelectronic devices. In principle, it is an exact many-body quantum mechanical approach which is able to describe Coulomb interaction and scattering effects in particular. In numerical implementations for the simulation of realistic device structures, however, the Coulomb interaction is often taken into account in terms of a mean field approximation, combined with a decoupled self-energy approximation.

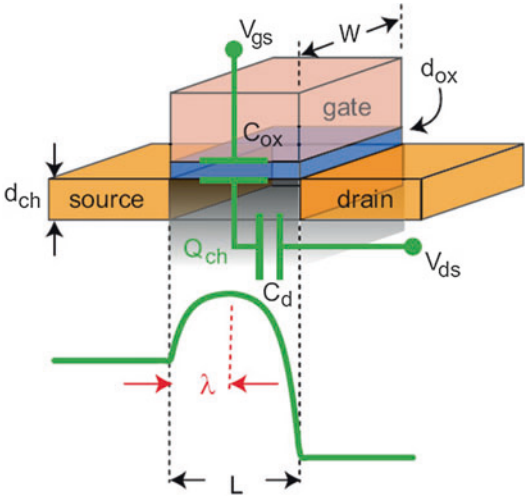
Nanowire Field Effect Transistor

The continued downscaling of the geometric dimensions of metal-oxide-semiconductor field-effect transistors (MOSFET) [1] has been propelled by Moore's law over the last 40 years, i.e., by the promise of an exponential growth of the number of MOSFETs that fit on a single chip, by a strongly increasing circuit performance and reduced costs at the same time. However, when scaling down the channel length of a MOSFET, so-called short channel effects (SCEs) appear that lead to a loss of electrostatic integrity of the device resulting in an exponential increase of the static power consumption due to leakage (see, e.g., [2]).

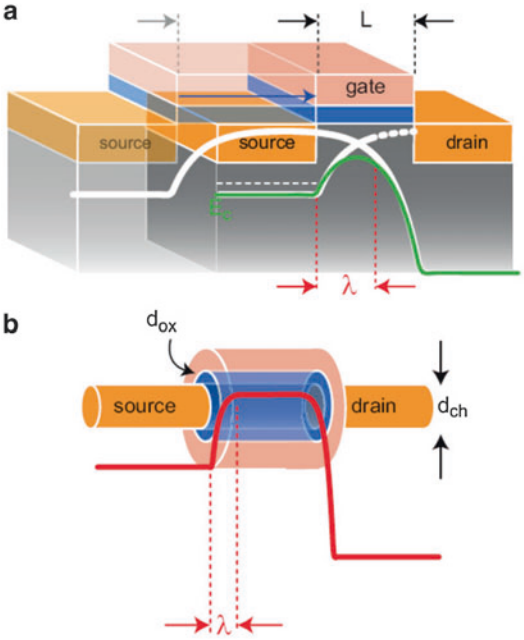
Consider Fig. 1, which shows a schematic of a planar n-type MOSFET. The charge injected into the channel Q_{ch} is determined by the terminal voltages V_{gs} and V_{ds} that are capacitively coupled to the channel via the oxide and drain capacitances C_{ox} and C_{d}, respectively. Short channel effects appear the more Q_{ch} is influenced by C_{d} instead of C_{ox}, i.e., if both capacitances are of the same order $C_{\mathrm{ox}} \approx C_{\mathrm{d}}$. Approximating C_{ox} and C_{d} with the expression for a parallel-plate capacitor, one obtains

$$C_{\mathrm{ox}} = \epsilon_{\mathrm{ox}}\epsilon_0 \frac{W \cdot L}{d_{\mathrm{ox}}} \quad \text{and} \quad C_{\mathrm{d}} = \epsilon_{\mathrm{ch}}\epsilon_0 \frac{W \cdot d_{\mathrm{ch}}}{L} \tag{1}$$

where d_{ox}, d_{ch} denote the oxide and channel thicknesses and L, W are the channel length and width,



Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism, Fig. 1 Planar n-type MOSFET with potential profile within the channel



Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism, Fig. 2 (**a**) Reduction of the channel length in a MOSFET. (**b**) Nanowire FET

respectively (see Fig. 1). Then, from $C_{ox} \approx C_d$, one obtains $L \approx \lambda$, where the length scale

$$\lambda = \sqrt{\frac{\epsilon_{ch}}{\epsilon_{ox}} d_{ox} d_{ch}} \tag{2}$$

represents a screening length of potential variations within the channel (see Fig. 1). In fact, the spatial extent of the source-channel and channel-drain $p-n$ junctions in a MOSFET is on the order of λ, and thus, SCEs disappear if $L \gg \lambda$. For a given channel length L, this can be achieved by using high-k gate dielectrics with $\epsilon_{high-k} \gg \epsilon_{ox}$ as well as ultrathin gate dielectric and channel thicknesses. However, reducing the thicknesses of d_{ox} and d_{ch} is limited by gate leakage due to direct tunneling and threshold voltage variations due to quantization (see e.g., Ref. [3–5]). Therefore, for ultimately scaled MOSFETs, a multigate, nonplanar device architecture has to be employed such as Intel's trigate MOSFETs at the 22 nm node [6]. In particular, a cylindrical gate-all-around (GAA) transistor based on a nanowire with ultrathin diameter provides the best electrostatic integrity, i.e., the smallest values for λ. Comparing the capacitance of a coaxial system

$$C_{ox} = 2\pi\epsilon_0\epsilon_{ox} \frac{L}{\ln(1 + 2d_{ox}/d_{ch})} \tag{3}$$

with the approximation for

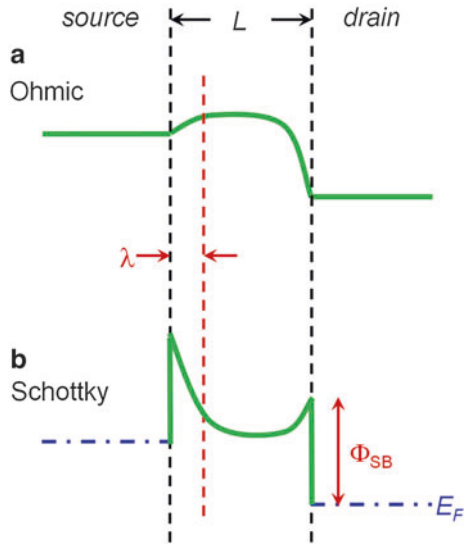
$$C_d = \epsilon_0\epsilon_{ch} \frac{\pi(d_{ch}/2)^2}{L} \tag{4}$$

one obtains [7]

$$\lambda = \sqrt{\frac{d_{ch}^2}{8} \frac{\epsilon_{ch}}{\epsilon_{ox}} \ln\left(1 + 2\frac{d_{ox}}{d_{ch}}\right)}, \tag{5}$$

where d_{ch} is the diameter of the nanowire (see Fig. 2).

In ultimately scaled FETs, the nanowire diameter needs to be on the order of a few nanometers only, which inevitably leads to one-dimensional (1D) electronic transport in the nanowires. On the one hand, 1D transport is beneficial for the performance of FETs since carrier scattering is



Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism, Fig. 3 Potential profile within the NWFET channel for (**a**) doped Ohmic contacts, (**b**) Schottky contacts

suppressed and the density-of-states limit can be approached when scaling the transistor yields performance benefits in terms of reduced dynamic power consumption of integrated circuits [8, 9]. On the other hand, in nanowires with small diameters, dopants become deactivated resulting in increased parasitic resistances of the source/drain extensions in a nanowire FET (NWFET) as has been shown experimentally by Bjoerk et al. [10]. In addition, the small volume of such nanowires leads to device-to-device variability due to random dopant distribution. As a result, metallic source/drain contact electrodes are being considered that allow circumventing all dopant-related issues. However, the Schottky barrier present at the metal–semiconductor interface yields deteriorated on- and off-state device behavior [11]. Fortunately, the tunneling of charge carriers through the Schottky barrier is increased as λ is decreased, and as a result, in ultimately scaled devices that require ultrathin d_{ch} and d_{ox}, the device performance is only slightly lower compared to FETs with (experimentally unrealizable) doped Ohmic source/drain contacts. Figure 3 shows the typical potential profiles within the channel of a NWFET with doped Ohmic source/drain contacts and Schottky contacts (with Schottky barrier height Φ_{SB}), respectively.

The small size of the nanowire devices with 1D electronic transport necessitates a full quantum kinetic simulation approach such as the nonequilibrium Green's function formalism (NEGF) as described below. In the presence of Schottky barriers, few-electron charging effects of quasibound electrons may emerge which are beyond the scope of the commonly employed mean field approximation of the Coulomb interaction (e.g., self-consistent Hartree potential), requiring a many-body (e.g., multiconfigurational) approach.

Channel Electrostatics

Within the channel region of the NWFET, the Coulomb interaction between charges (such as electrons or ionized dopants) is screened due to image charge effects within the surrounding metallic gate electrode. The resulting electrostatic potential $V(x)$ (along the nanowire channel direction) can be modeled by a one-dimensional (1D) modified Poisson equation [7, 12]

$$\frac{\partial^2}{\partial x^2}V(x) - \frac{1}{\lambda^2}V(x) = -\frac{1}{\epsilon_0\epsilon_{ch}S}\rho(x) - \frac{1}{\lambda^2}V_{gs} \tag{6}$$

where λ is the effective screening length (see chapter above), V_{gs} denotes the gate-source voltage, $\rho(x)$ is the 1D total charge density (electrons, holes, and ionized dopants), and $S = \pi d_{ch}^2/4$ is the channel cross-sectional area.

As an analytical solution of the modified Poisson equation for fixed potential boundary conditions, one obtains

$$V(x) = \frac{1}{\epsilon_0\epsilon_{ch}S}\int dx' v(x,x')\ \rho(x') + V_{ext}(x) \tag{7}$$

with the so-called Coulomb Green's function

$$v(x,x') = \frac{\lambda}{2}e^{-\frac{|x-x'|}{\lambda}} + v_{bc}(x,x') \tag{8}$$

which contains a fixed-potential boundary condition term

$$v_{\mathrm{bc}}(x,x') = -\frac{\lambda}{2}e^{-\frac{x+x'}{\lambda}} + \frac{\lambda}{2}e^{-\frac{L}{\lambda}}\left(\frac{\cosh\left(\frac{x-x'}{\lambda}\right)}{\sinh\left(\frac{L}{\lambda}\right)} - \frac{\cosh\left(\frac{x+x'}{\lambda}\right)}{\sinh\left(\frac{L}{\lambda}\right)}\right) \tag{9}$$

stemming from screening effects due to metallic source and drain contacts at the outer ends of the channel (where v must vanish).

The electrostatic control of the FET as a function of the contact potentials V_{gs}, V_{ds} is solely contained within the potential term

$$V_{\mathrm{ext}}(x) = \frac{\sinh\left(\frac{x}{\lambda}\right)}{\sinh\left(\frac{L}{\lambda}\right)}V_{\mathrm{ds}} + \frac{1}{\lambda^2}\int dx' v(x,x')\,V_{\mathrm{gs}} \tag{10}$$

Quantum Kinetic Transport Simulation

On a fundamental level, the nonequlilbrium quantum mechanical state (preparation) of the whole interacting many-electron system of the nanodevice is completely described by its many-body statistical operator. For numerical simulation approaches of realistic nanodevices with a large number of basis states (as opposed to overidealized few-site models), however, this quantity exhibits an expontially scaling, numerically unfeasible number of matrix elements. Therefore, it is reasonable to consider temporal and spatial correlation functions of ordered field operators of the electronic system, so-called real-time Green's functions $G(x_1, t_1; x_2, t_2; \ldots)$, since those quantities can be employed directly for the calculation of relevant physical observables such as charge density and electrical current. It turns out that one can derive a closed set of integrodifferential equations (Dyson equation) for these Green's functions, without the necessity of a detailed knowledge about the full many-body statistical operator. For the numerical evaluation of these equations, approximation and reduction schemes are still required, assuming typical application-relevant conditions for device operation. This concept constitutes the basis for the nonequilibrium Green's function formalism (NEGF) for the quantum kinetic simulation of realistic nanoelectronic devices [13–17].

Within a numerical description of the nanoelectronic system, a suitable single-particle basis has to be chosen for a numerical representation of operators in terms of matrix elements. For the description of heterostructures of nanodevices layer by layer in real space (as opposed to approaches in momentum space), a tight binding representation of spatially localized basis wave functions is advantageous [18, 19]. Here, the single-particle ON basis functions ϕ_j (x, σ) are characterized by a multi-index j, which includes the unit cell site (localized position), orbital, and spin. Within the resulting Hamiltonian, transitions between states j, k are described by so-called hopping matrix elements t_{jk}. Typically, in the case of a thin nanowire, only the lowest radial sub-band within the nanowire needs to be considered, and thus, the NWFET channel can be treated as a 1D chain of localized states.

Within the NEGF approach for the simulation of stationary electronic transport in NWFETs, the two-point real-time Green's functions G^r ("retarded") and $G^<$ ("lesser") are typically considered, defined as

$$\begin{aligned} G^r_{jk}(t) &= -i\Theta(t)\left\langle\left\{c_j(t), c_k^\dagger(0)\right\}\right\rangle \\ G^<_{jk}(t) &= i\left\langle c_k^\dagger(0)c_j(t)\right\rangle \end{aligned} \tag{11}$$

where $c_k^\dagger$ and c_j denote the creation and annihilation operator for state j, respectively. Under nonequilibrium conditions, G^r and $G^<$ are independent, containing information about the electron dynamics (energy spectra, etc.) and the electron kinetics (level occupation, particle transport, etc.), respectively. For a stationary case, G depends on the time difference t only, and in turn, the time variable t can be Fourier transformed into an energy variable E. Consequently, G^r and $G^<$ are numerically represented as energy-dependent single-particle matrices.

G obeys the Dyson equation, which can be written as a set of coupled matrix equations (with E variable omitted for brevity) for G^r and $G^<$:

$$\begin{aligned} G^r &= G^{r0} + G^{r0}\sum{}^r G^r \\ G^< &= (1 + G^r\sum{}^r)G^{<0}(\sum{}^a G^a + 1) + G^r\sum{}^< G^a \end{aligned} \tag{12}$$

where $\sum$ is the so-called self-energy, which is an energy-dependent matrix, and $G^a = G^{r\dagger}$, $\sum{}^a = \sum{}^{r\dagger}$. $\sum$ incorporates all aspects of the system that are not contained within the unperturbed Green's function G^0. Typically, $\sum$ accounts for the Coulomb interaction, contact coupling (to source and drain for particle injection), scattering, etc. Consequently, G^0 has a very simple analytical form of a noninteracting, decoupled single-particle problem:

$$G^{r0}(E) = \frac{1}{E - H_{10} + i\epsilon} \quad (\epsilon \to 0+) \tag{13}$$

whereas $\sum$ becomes very complex and requires suitable approximation schemes for numerically feasible implementations.

From $G^<$, the single-particle density matrix

$$\rho_1 = \frac{1}{2\pi}\int dE \frac{1}{i} G^<(E) \tag{14}$$

can be calculated easily. In turn, all required single-particle expectation values, such as the electron charge density ρ_e and the drain current I_d, can be computed directly from ρ_1:

$$\rho_e(x,\sigma) = -e\sum_{j,k} \phi_j^*(x,\sigma)\ \ \phi_k(x,\sigma)\ \ \rho_{1kj} \tag{15}$$

$$I_d = -\frac{e}{\hbar} \sum_{\substack{j,k \\ x_j \le x_d, \\ x_k > x_d}} 2Im\Big(t_{jk}\ \ \rho_{1kj}\Big) \tag{16}$$

As an approximation under application-relevant conditions (e.g., temperatures well above the relevant Kondo temperature of the nanosystem), the self-energy $\sum$ can be assumed to be of a decoupled form, written as a sum of independent contributions due to contact coupling, Coulomb interaction etc. Then, the retarded self-energy term $\sum_c$ due to source and drain contact coupling, for example, reads as

$$\sum{}^r_{cjk}(E) = \sum_{p,q\in c} t_{cjp} G^{r0}_{cpq}(E) t_{cqk} \tag{17}$$

and the self-energy term $\sum_{ee}$ for the Coulomb interaction in static (i.e., E-independent) mean field approximation (here Hartree-Fock), for example, reads as

$$\sum{}^r_{eeml}[\rho_1] = \sum_{j,k} \big(V_{mjkl} - V_{jmkl}\big)\rho_{1_{kj}} \tag{18}$$

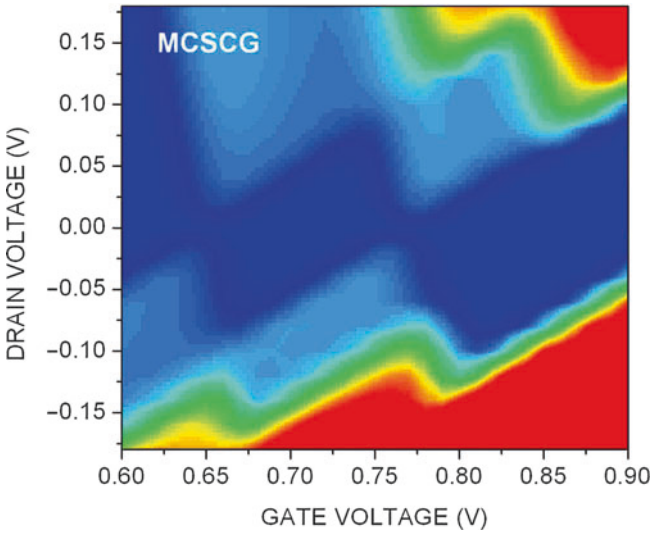
with Coulomb matrix elements V_{mjkl} of the screened Coulomb interaction within the channel obtained from the Coulomb Green's function $v(x,x')$ (see chapter above):

$$\begin{aligned} V_{mjkl} = \ & \frac{e^2}{\epsilon_0\epsilon_{ch}S}\sum_{\sigma,\sigma'}\int dx \int dx'\ \ v(x,x') \\ & \times \phi_m^*(x,\sigma)\,\phi_j^*(x',\sigma')\phi_k(x',\sigma')\phi_l(x,\sigma) \end{aligned} \tag{19}$$

Within the considered approximation scheme, analytical expressions for $\sum^<$ can derived from the $\sum^r$ expressions above. One has to note that $\sum_{ee}$ above depends on ρ_1 which, in turn, results from $G^<$. In combination with the Dyson equation, therefore, a self-consistency problem results that has to be solved iteratively.

Few-Electron Charging Effects and Many-Body Aspects

Under certain conditions, charging effects of individual electrons (or holes) may become visible in the IV characteristics of the nanodevice. For example, so-called Coulomb diamonds (see e.g., Fig. 4) can emerge within the $I_d(V_{gs}, V_{ds})$ diagram, representing regions of suppressed current I_d due to a finite energy ΔE which is required to add a single additional electron to the system.



Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism,
Fig. 4 Simulated drain current with Coulomb diamonds for an InGaAs NWFET with Schottky contacts, $L = 20$ nm, $d_{ch} = 4$ nm, $d_{ox} = 10$ nm at $T = 77$ K [21]

Such few-electron charging effects require the existence of resonantly trapped electronic states which are quasibound due to a weak coupling to the external contacts (e.g., due to Schottky barriers or potential wells in heterostructures). The energy gap due to Coulomb repulsion reads as $\Delta E_C = e^2/C^*$, where C^* denotes the effective capacitance of the system. Analogous, the single-particle quantization energy gap reads as $\Delta E_L \propto \hbar^2/\left(m^* L^2\right)$, which is due to the spatial confinement ($\propto L$) of quasibound electronic states. In general, the observed voltage regions of suppressed current (diamond-like shaped) are a result of the combination of both effects, Coulomb repulsion and level quantization due to confinement, resembling a finite total energy gap ΔE that has to be overcome to add an additional electron to the system. As a criterion for the visibility of this energy gap, the thermal energy $k_B T$ and the level broadening $\Gamma_c = i\left(\sum_c^r - \sum_c^{r\dagger}\right)$ (as a measure of the coupling strength to the contacts) must be small compared to ΔE, leading to the transport regime of so-called single-electron tunneling (SET). One has to note that under application-relevant conditions (above the Kondo temperature of the coupled system of contacts and trapped, interacting states), correlation effects associated with the contacts can be neglected (justifying the approximation of a decoupled self-energy form, see chapter above).

Experimentally as well as numerically, Coulomb blockade effects can be employed to determine effective capacitances C^* from the sizes of the observed Coulomb diamonds, providing valuable information about the nanowire channel.

Unfortunately, the simulation of few-electron Coulomb charging is beyond the scope of the commonly employed static mean field approximation (e.g., Hartree or Hartree-Fock), making a many-body approach to the interacting few-electron system mandatory. Due to the numerically unfeasible Fock space dimension of a realistic NWFET device (with a large number of single-particle basis states), numerical approximation schemes are required. In order to include few-electron Coulomb charging effects of resonantly trapped electrons within the framework of NEGF in a numerically feasible manner, for example, a multiconfigurational ansatz can be employed [20, 21]. The main idea is to write the Green's functions as a weighted mixture over a reduced set of relevant many-body states (so-called configurations) k of relevant quasibound electronic states:

$$\overline{G} = \sum_{\kappa} w_\kappa G[\hat{\rho}_\kappa] \tag{20}$$

where $\hat{\rho}_\kappa$ is the many-body statistical operator corresponding to the system being in

configuration κ. Note that the number of configurations grows exponentially with the number of relevant single-particle states. Relevant single-particle states are defined as those natural orbitals (i.e., eigenstates of ρ_1) that exhibit occupation fluctuations and are quasibound (i.e., weakly coupled), thus being responsible for few-electron charging effects (leading to Coulomb diamonds in the IV characteristics, etc.). This approach, hence, considers a voltage- and parameter-adaptive relevant Fock subspace which can be handled numerically (as compared to the full Fock space of the whole nanowire system). The determination of the relevant natural orbitals and the many-body configuration weights w_κ can be embedded self-consistently within the NEGF algorithm, leading to the so-called multiconfigurational self-consistent Green's function (MCSCG) approach for the inclusion of few-electron Coulomb charging effects (see, e.g., Fig. 4) within the numerical framework of NEGF-based device simulators [20, 21].

Cross-References

- Ab Initio DFT Simulations of Nanostructures
- Bio-FET
- CMOS (Complementary Metal-Oxide-Semiconductor)
- CNT-FET
- Green's Function Integral Equation Method
- Mean Field Approaches to Many-Body Systems
- Nano-FET
- Nanostructure Field Effect Transistor Biosensors
- Nanowire FET Biosensor
- Quantum Dot

References

1. Sze, S.M., Lee, M.-K.: Semiconductor Devices: Physics and Technologies. Wiley, New York (2012)
2. Roy, K., Mukhopadhyay, S., Mahmoodi-Meimand, H.: Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits. Proc. IEEE **91**, 305 (2003)
3. Lo, S.-H., Buchanan, D.A., Taur, Y., Wang, W.: Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide nMOSFETs. IEEE Trans. Electron. Dev. **18**, 209 (1997)
4. Tsutsui, G., Saitoh, M., Nagumo, T., Hiramoto, T.: Impact of SOI thickness fluctuation on threshold voltage variation in ultra-thin body SOI MOSFETs. IEEE Trans. Nanotechnol. **4**, 369 (2005)
5. Knoch, J., Lengeler, B., Appenzeller, J.: Quantum simulations of an ultrashort channel single-gated n-MOSFET on SOI. IEEE Trans. Electron. Dev. **49**, 1212 (2002)
6. Auth, C., et al.: A 22 nm high performance and low-power CMOS technology featuring fully-depleted TriGate transistors, self-aligned contacts and high density MIM capacitors. 2012 VLSI Symp., 131 (2012)
7. Auth, C.P., Plummer, J.D.: Scaling theory for cylindrical, fully-depleted, surrounding-gate MOSFET's. IEEE Electron. Dev. Lett. **18**, 74 (1997)
8. Knoch, J., Riess, W., Appenzeller, J.: Outperforming the conventional scaling rules in the quantum-capacitance limit. IEEE Electron. Dev. Lett. **29**, 372 (2008)
9. Knoch, J.: One-dimensional field-effect transistors. In: Jiannian, Y., Zhai, T. (eds.) One-Dimensional Nanostructures: Principles and Applications. Wiley, Hoboken (2012)
10. Bjoerk, M.T., Schmid, H., Knoch, J., Riel, H., Riess, W.: Dopant deactivation in silicon nanostructures. Nature Nanotechnol. **4**, 103 (2009)
11. Knoch, J., Zhang, M., Mantl, S., Appenzeller, J.: On the performance of single-gated ultrathin body SOI Schottky-barrier MOSFETs. IEEE Trans. Electron. Dev. **53**, 1669 (2006)
12. Pikus, F.G., Likharev, K.K.: Nanoscale field-effect transistor: an ultimate size analysis. Appl. Phys. Lett. **71**, 3661 (1997)
13. Datta, S.: Electronic Transport in Mesoscopic Systems. Cambridge University Press, Cambridge (1998)
14. Ferry, D.K., Goodnick, S.M.: Transport in Nanostructures. Cambridge University Press, Cambridge (1997)
15. Lake, R., Klimeck, G., Bowen, R.C., Jovanovic, D.: Single and multiband modeling of quantum electron transport through layered semiconductor devices. J. Appl. Phys. **81**, 7845 (1997)
16. Schäfer, W., Wegener, M.: Semiconductor Optics and Transport Phenomena. Springer, Berlin (2002)
17. Haug, H., Jauho, A.-P.: Quantum Kinetics in Transport and Optics of Semiconductors. Springer, New York (1998)
18. Vogl, P., Hjalmarson, H.P., Dow, J.D.: A Semi-empirical tight-binding theory of the electronic structure of semiconductors. J. Phys. Chem. Solids **44**, 365 (1983)
19. Støvneng, J.A., Lipavský, P.: Multiband tight-binding approach to tunneling in semiconductor heterostructures – application to gamma-X transfer in GaAs. Phys. Rev. B **49**, 16494 (1994)

20. Indlekofer, K.M., Knoch, J., Appenzeller, J.: Quantum kinetic description of Coulomb effects in one-dimensional nanoscale transistors. Phys. Rev. B **72**, 125308 (2005)
21. Indlekofer, K.M., Knoch, J., Appenzeller, J.: Understanding Coulomb effects in nanoscale Schottky-barrier FETs. IEEE Trans. Electron. Dev. **54**, 1502 (2007)

Nanowire-Based Field Effect Transistor (NWFET)

▶ Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism

Nanowires

▶ Chemical Vapor Deposition (CVD)
▶ Heat Transfer in Semiconductor Nanostructures
▶ Nanostructures for Energy
▶ Physical Vapor Deposition
▶ Tight-Binding Simulations of Nanowires

Nature Inspired Microneedle

▶ Biomimetic Mosquito-Like Microneedles

Near Infrared

▶ Thermal Cancer Ablation Therapies Using Nanoparticles

Near-Field Focusing and Heating

▶ Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing

Near-Field Molecular Optics

▶ Nanopolaritonics

Near-Field Optics

▶ Light Localization for Nano-optical Devices

Near-Field Scanning Optical Microscopy (NSOM)

▶ Scanning Near-Field Optical Microscopy

NEMS

▶ Graphene NEMS
▶ Nanotechnology
▶ Nonlinear and Parametric NEMS Resonators

NEMS Gravimetric Chemical Sensors

▶ NEMS Resonant Chemical Sensors

NEMS Gravimetric Sensors

▶ NEMS Resonant Chemical Sensors

NEMS Mass Sensors

▶ Graphene NEMS
▶ NEMS Resonant Chemical Sensors

NEMS Piezoelectric Switches

Nipun Sinha
Department of Mechanical Engineering and Applied Mechanics, Penn Micro and Nano Systems (PMaNS) Lab, University of Pennsylvania, Philadelphia, PA, USA

Synonyms

Dual-beam piezoelectric switches; Piezoelectric switches

Definition

NEMS Piezoelectric Switches are defined as mechanical switches that use piezoelectric actuation for switching and employ ultrathin or nanopiezoelectric films for realizing these switches.

Background and Motivation

MEMS switches were first demonstrated in 1979 [1]. Since then lot of research activities have focused on their development and optimization. MEMS switches have traditionally provided a multitude of advantages like higher isolation, lower insertion loss, very low power consumption, and better linearity when compared to traditional transistor/FET and pin diodes [2]. Most of the switch development till date has been done for applications in the Radio Frequency MicroElectroMechanical Systems (RF MEMS) domain. The main aim in this case is to use mechanical relays for switching on and off various components in RF architectures and also to facilitate the functioning of reconfigurable systems. More recently, there has been growing interest in the development of low voltage mechanical computing platforms using MEMS switches and reconfigurable mechanical RF front-ends. RF MEMS switches with their favorable on and off characteristics are ideal for such applications.

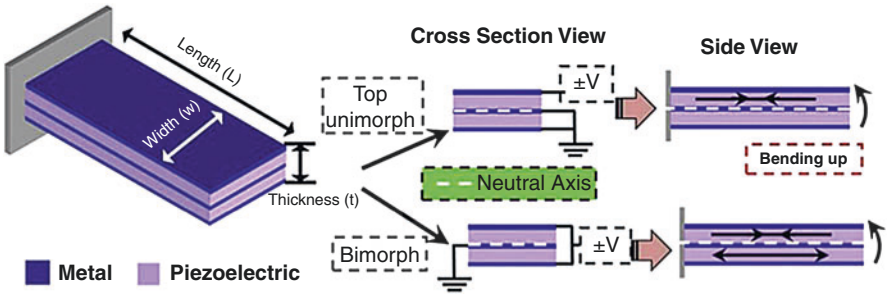
Various actuation methods can be employed for achieving switching, such as electrostatic [3–12], piezoelectric [13–23], electromagnetic [24], thermal [25], and electrothermal [26]. Out of these methods, electrostatic actuation has been the most commonly and widely used actuation method. However, it suffers from inherent disadvantages due to its nonlinear mechanism, which transduction methods like piezoelectric actuation have the ability to overcome.

Piezoelectricity, as a phenomenon, was discovered in 1880 by Pierre Curie and Jacques Curie (the Curie brothers). It is simply characterized by the accumulation of charge in certain materials upon the application of mechanical stress. The converse piezoelectric effect also exists, which makes it a reversible process. In the converse piezoelectric effect, an applied electric field can lead to the generation of a mechanical force, thereby causing the film to strain. This converse effect was not predicted by the Curie brothers initially and was predicted by Gabriel Lippmann in 1881. It was later confirmed experimentally by the Curie brothers. The piezoelectric effect is a linear electromechanical interaction between the electrical and mechanical variables in crystalline and polycrystalline materials. For a material to exhibit this phenomenon its crystal structure needs to possess some asymmetry, so materials with inversion symmetry do not exhibit this property.

This linear interaction can be defined using a set of coupled equations between the strain (S) (6×1 matrix), compliance (s) (6×6 matrix), stress (T) (6×1 matrix), electric field $(\vec{E})$ (3×1 matrix), piezoelectric coefficients (d) (3×6 matrix), permittivity (ε) (3×3 matrix), and electric charge density displacement (D) (3×1 matrix).

$$\begin{aligned}[S] &= \left[s^E\right][T] + \left[d^t\right]\left[\vec{E}\right][D] \\ &= [d][T] + \left[\varepsilon^T\right]\left[\vec{E}\right] \end{aligned} \quad (1)$$

Implementation of piezoelectric actuation using the converse piezoelectric effect necessitates the presence of two electrodes that sandwich



NEMS Piezoelectric Switches, Fig. 1 Schematic representation of the piezoelectric actuator (showing the length (L), width (w), and thickness (t) of the actuator) made of two piezoelectric layers and three metal layers illustrating the two common topologies used for actuation, i.e., the unimorph (using one layer of piezoelectric material for actuation) and bimorph (using two layer of piezoelectric material for actuation)

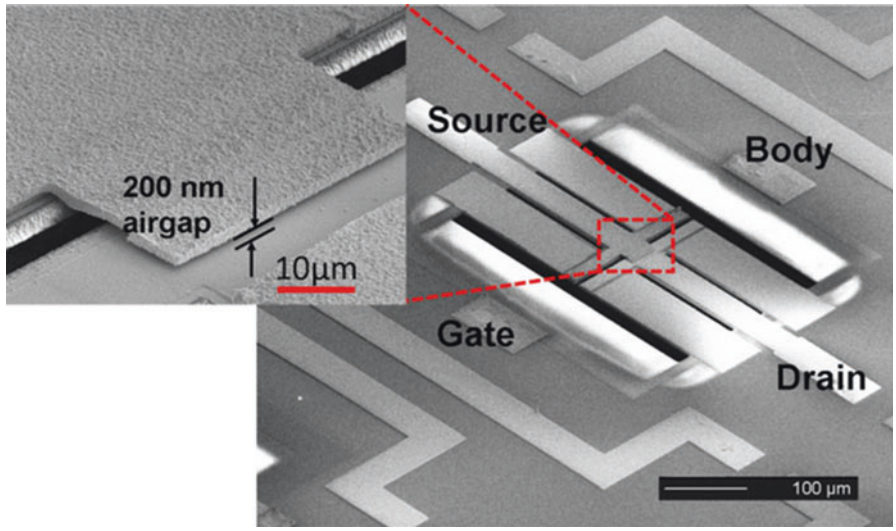
the piezoelectric material to apply an electric field across it. For in plane stretching and consequently bending operation, the d_{31} piezo-coefficient is conventionally used in c-axis oriented, hexagonal materials such as AlN. In this case, an electric field applied across the film thickness generates an in-plane strain through the equivalent d_{31}piezo-coefficient. In a multilayered beam, this strain can be tailored to generate a bending moment about the neutral axis of the composite beam. This moment deflects the beam out-of-plane, i.e., vertically, and this motion is used to open and close an air gap in a mechanical switch. There are two commonly used configurations for implementing piezoelectric actuation, namely, unimorph and bimorph, shown in Fig. 1. In the unimorph configuration [27], there is only one operational piezoelectric layer. A unimorph actuator has both active and passive layers. The passive layers are used for offsetting the neutral axis so that the piezoelectric force can generate a moment about it. In the bimorph configuration [28] there are two active piezoelectric layers, which are simultaneously actuated. The electrodes sandwiching the actuator are properly designed and routed so that the two layers provide for additive bending moments and generate a greater deflection per volt than in the unimorph case, for a given piezoelectric layer thickness.

Majority of the piezoelectric switches made till date have employed either Lead Zirconate Titanate (PZT) or Aluminum Nitride (AlN) as their functional material. When compared to PZT, AlN has the advantage of being compatible with CMOS processes and foundries (by being lead free), exhibits a lower leakage current for comparable film thicknesses and has a lower dielectric constant, which ultimately translates to reduced power consumption for a given actuator dimension. PZT does exhibit a higher piezoelectric coefficient than AlN, but repeatability of high-quality PZT films has been known to be a challenge. This, along with the need to pole PZT post deposition (in the sol–gel process) has been the major reason why none of the prominent examples of PZT switches employ bimorph actuation, whereas this can be easily achieved when using AlN. Unlike PZT, AlN is easy to fabricate and has shown preservation of piezoelectric coefficients when scaled from thick (1 μm) to thin (100 nm) films [29]. Another unique advantage that AlN-based devices have over PZT-based devices is the ability of easy integration with the already well-established AlN-based contour-mode resonator technology [30–33]. With this integration new classes of transceivers that will use single-chip multifrequency and highly reconfigurable systems capable of low power and low loss operation can be realized.

AlN-Based Dual-Beam Piezoelectric RF MEMS Switch

RF MEMS Switch

The AlN-based MEMS switch described herein utilizes a novel dual-beam design, i.e., it consists of two symmetric, released beams that face each



NEMS Piezoelectric Switches, Fig. 2 A SEM of a fabricated three-finger AlN MEMS switch with a zoomed-in view of the planarized nano-air gap (~200 nm). Single gate and body terminals are created by appropriate electrode routing

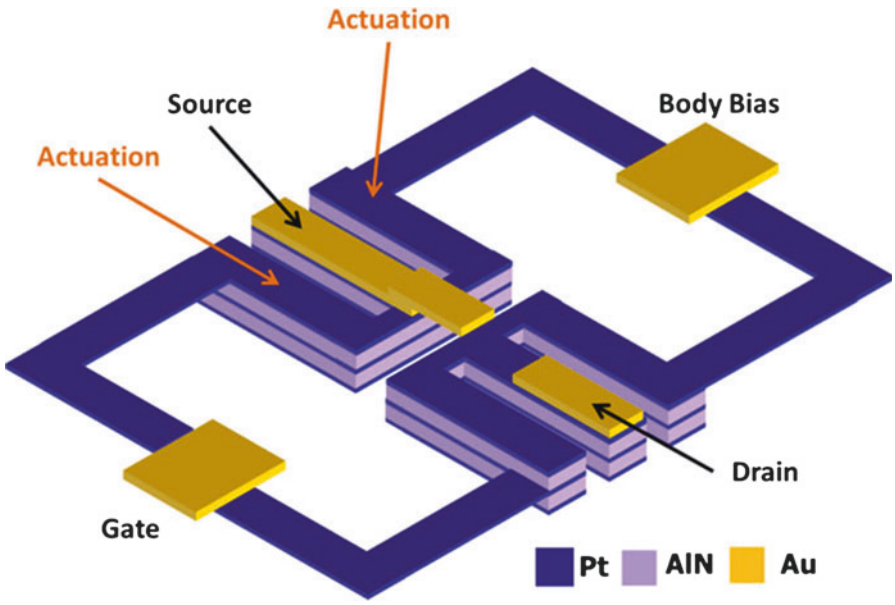
other and move in opposing directions (with respect to each other) with the same two inputs. Figure 2 shows an SEM of a dual-beam switch and the two inputs (i.e., the body and the gate) that are required for the switch to function. The reason for implementing such a design is the presence of residual stress in released multilayered structures. Residual stresses in MEMS and NEMS devices are unwarranted effects that result from the fabrication processes that occur at elevated temperatures. The residual stresses affect the structural reliability and the performance of the system. The deflection induced by residual stresses can cause significant variation in threshold voltages and can lead to a normally closed switch, and, in extreme cases, the release can lead to large deflections rendering the switch inoperable. Therefore, it is essential to overcome this effect in MEMS switches. The dual-beam design helps reducing the impact of this and is introduced primarily because of the need to fabricate switches immune to these residual stresses.

In the dual-beam mechanism employed, both beams can be actuated simultaneously, thus giving: (1) low actuation voltage (halved, as compared to just one beam moving the whole gap), (2) high contact forces (3) active pull-off to open the switch, and (4) faster switching times (as each beam has to move only half the gap). The nano-gap, shown in the zoomed-in view in Fig. 2, is what enables the switch to achieve both good contact forces and isolation.

The three-finger dual-beam switch design used for implementing piezoelectric devices was inherited from the initial RF switch development [21]. This topology comprises of using two symmetric released beams facing each other. Each of these two beams is made of three separate fingers (hence the three-finger design), the two outer fingers are used for actuation and the central finger is employed solely for carrying the signal from the source to the drain as to avoid any gate to source/drain overlap (Fig. 3). The actuators are made up of two layers of AlN, so this design helps in realizing bimorph actuation in piezoelectric switches.

Integration with Contour-Mode Technology

As mentioned earlier, one of the biggest advantages of developing an AlN-based RF MEMS switch was the ability to integrate it with the already proven contour-mode technology to



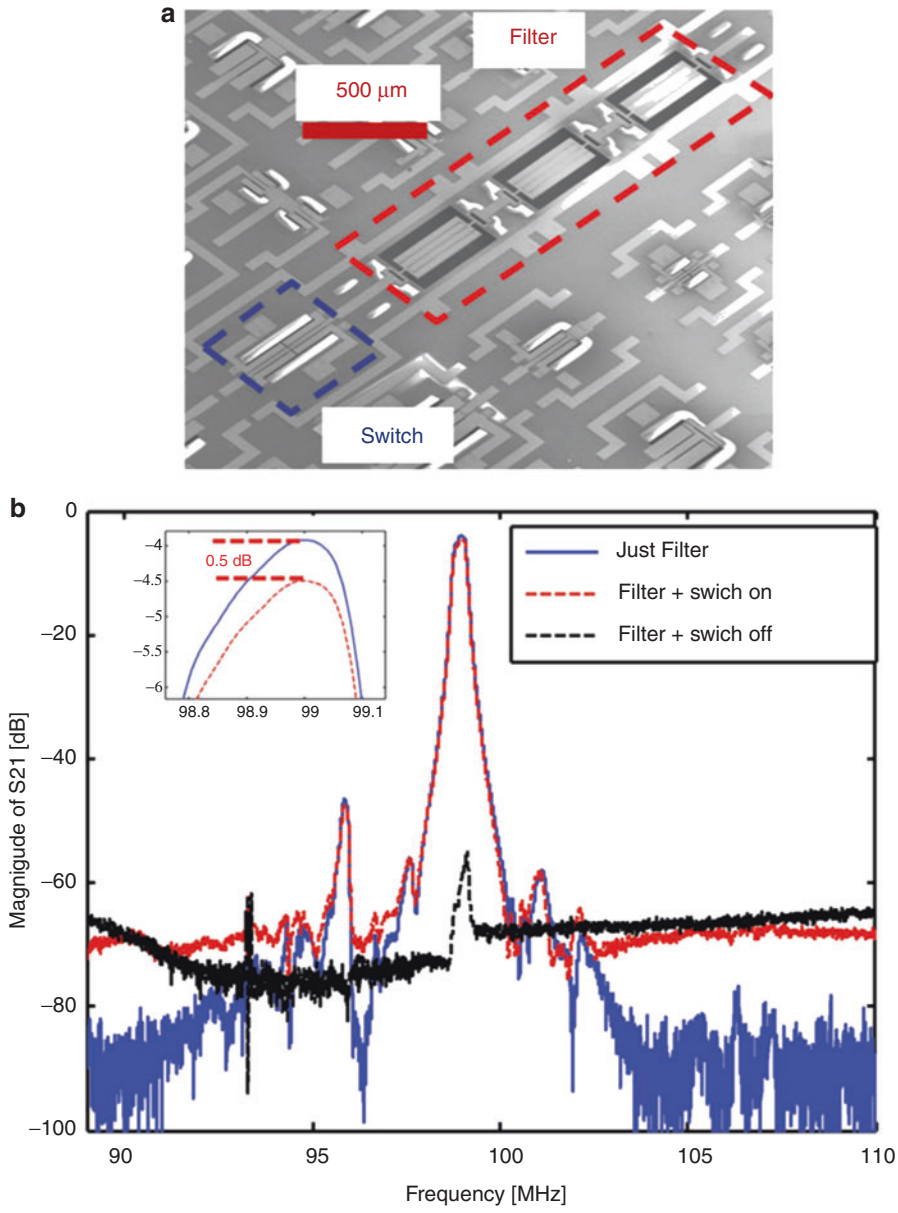
NEMS Piezoelectric Switches, Fig. 3 Schematic representation of the piezoelectric switch, showing the three-finger dual-beam design. The two pads are used to apply the gate and body-bias voltages and are properly connected to the layers of either actuator so that they actuate both the beams simultaneously

form switchable frequency control elements. This monolithic integration has the potential to tremendously impact the size and power efficiency of RF front-ends by drastically reducing parasitics and off-state leakage. Figure 4a shows an SEM of the monolithically integrated third order AlN contour-mode technology filter along with an AlN switch. Figure 4b shows the frequency response of the filter when it has been switched on and off. An attenuation of >50 dB was measured between the on and off states of the filter. The measurements also show the performance of the filter without the switch. The inset in Fig. 4b shows a zoomed-in view of the difference in insertion loss of the filter with the switch and without the switch, at the filter's center frequency (~99 MHz). The presence of the switch in the measuring path introduces a low loss in the performance of the filter, by increasing the insertion loss by ~0.5 dB.

Such integration of switches and frequency control components has the potential to enable new reconfigurable RF architectures based on massive arrays of multifrequency AlN contour-mode resonators and filters.

MEMS/NEMS Switch-Based Logic Using Body-Biasing

MEMS switches can also be easily employed for computing applications. Scaling of CMOS technology to the nano-regime is facing issues with power consumption and reduction of threshold voltages. For overcoming these challenges the International Technology Roadmap for Semiconductors (ITRS) has recognized the development of technologies, like nanomechanical computing based on NEMS switches, as a possible solution for future electronic transistor replacement. Unlike transistors, MEMS/NEMS switches do not suffer from the drawback of passive power loss as they usually have an air gap instead of the physical channel between the source and the drain. Therefore, they can be used for developing efficient low power computing devices. AlN-based switches have been used to implement a unique MEMS switch actuation methodology, i.e., the body-biased actuation, which has enabled low voltage demonstration of basic logic-building blocks. These switches can be referred to as



NEMS Piezoelectric Switches, Fig. 4 (**a**) SEM of a monolithically integrated RF MEMS switch with a contour-mode filter. (**b**) Shows the experimental results of a switchable AlN filter effectively being turned on and off by an AlN RF MEMS switch. The switch introduces low loss and enables a single-chip switchable filter bank solution

piezoelectric mechanical transistors (MTs) as they use mechanical transduction to achieve the function that is traditionally carried out by electronic transistors.

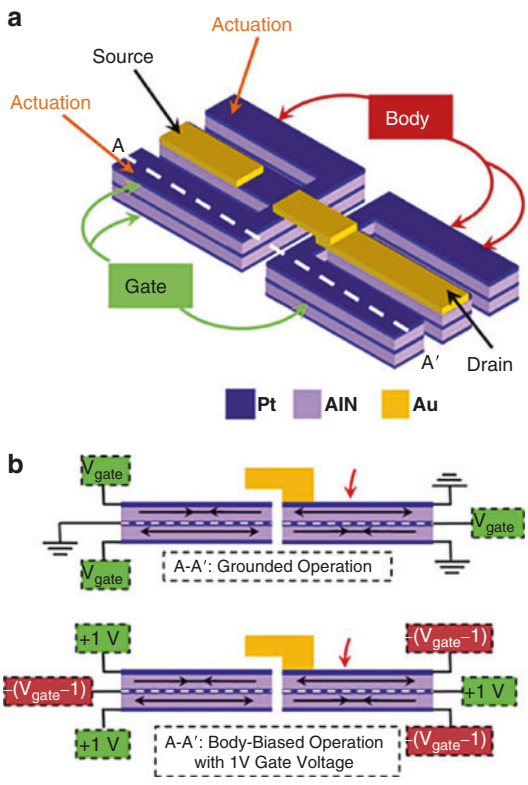
Body-Biased Actuation

As mentioned earlier, piezoelectric actuation necessitates the presence of two actuation electrodes that sandwich the piezoelectric material

and create an electric field across it. It is this electric field across the piezoelectric film, which generates the strain that leads to actuation. A body bias, in the present context, stands for a constant and floating voltage applied to one of the actuation electrodes. Traditionally, the actuators are operated by applying a voltage (V_{gate}) to the gate terminal, which is greater in magnitude than the threshold voltage (V_{th}) of the device, i.e., $V_{gate} > V_{th}$. This mode of actuation is referred to as the "grounded" mode of actuation, since the body terminal of the device is grounded, shown in Fig. 5b. Instead, in the "body-biased" mode of operation, the previously grounded electrode is biased to a constant voltage (V_{bias}) close in magnitude (but opposite in polarity) to the threshold voltage so that a lower gate voltage can be used to open and close the switch. In Fig. 5, the body-biased mode of actuation is explained schematically using a cross-section of a three-finger dual-beam switch [19, 20]. Figure 5b presents a scenario where the gate voltage of the MT is reduced from V_{gate} to 1 V by using a body-bias voltage of $-(V_{gate}-1)$. This new and lower threshold voltage leads to a smaller swing voltage, which effectively lowers the energy required for actuation. From a mechanical perspective, body-biasing can be explained as preloading (prebending) of the beams that form the actuators. Effectively, body-biasing corresponds to the reduction of the air gap between the two contacting surfaces and translates into the ability of tuning the device threshold voltage by modulating the size of the initial contact gap.

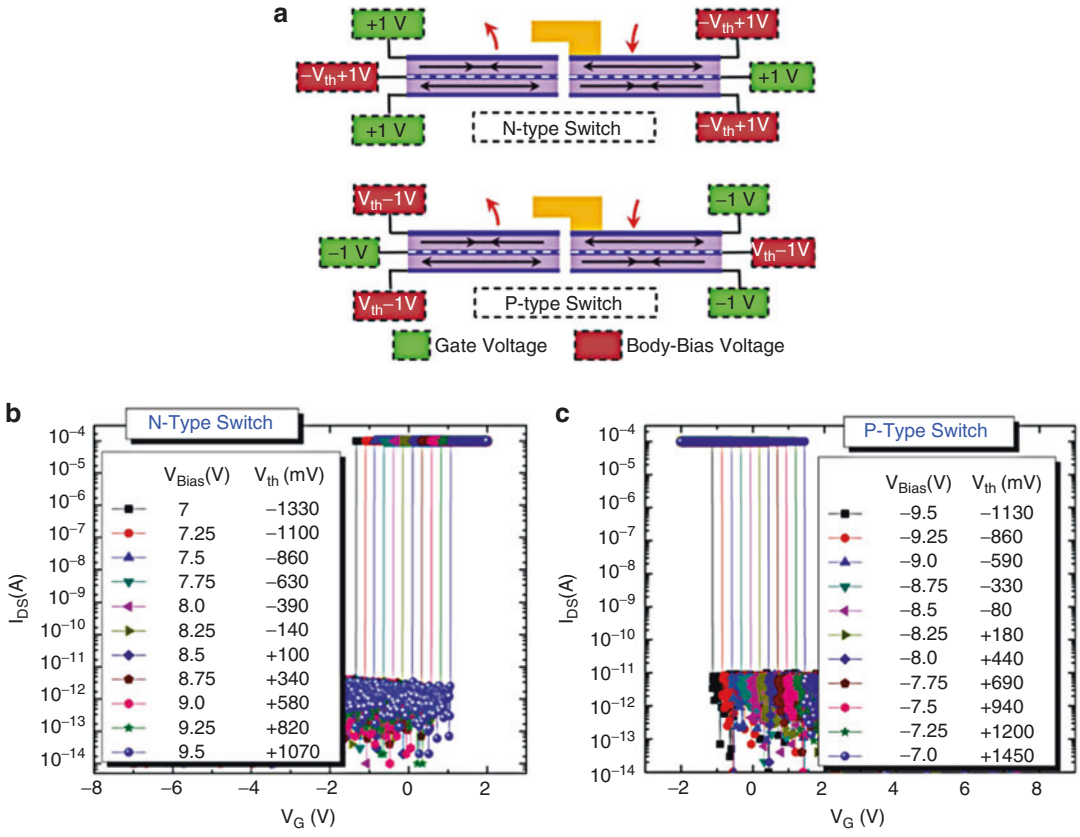
By using the body-biased mode of actuation, the threshold voltage of a piezoelectric switch can be tailored to be any value. The same switch can have a negative or positive threshold voltage, as shown in Fig. 6. The major aim of this technique is to reduce the threshold voltage in magnitude (as low as 1–20 mV), which can be easily achieved by body-baising.

The absence of a doped channel in the MT renders the classification of n- or p-type devices ambiguous. Moreover, this provides the MT with an unparalleled ability to choose or vary the nature of a switch on the basis of need, i.e., the same switch can be made into an n- or p-like device



NEMS Piezoelectric Switches, Fig. 5 (**a**) 3D schematic of a three-finger dual-beam AlN switch highlighting the source, drain, gate, and body terminals. It also shows the A–A′ cross-section used to illustrate the body-biasing method. (**b**) Schematic representation of the principle of operation of an AlN MEMS switch in grounded and body-biased mode of operation. By applying a bias to the body terminal $[-(V_{gate}-1)]$ V, the same effective potential required to close the switch is applied across the AlN layer with a lowered gate voltage of 1 V (1 V is used here as an example, but, theoretically, any value is possible thanks to the extreme linearity of the body-bias technique in a piezoelectric AlN film). This effectively reduces the dynamic energy required for every switching operation. The gold tip is not present along the A–A′ cross-section, but is drawn here to aid the understanding of the mechanism of device operation

(in analogy to p-type and n-type CMOS) without any change in design or fabrication by simply switching the polarity of the body-bias and applying it to the opposing actuating electrode set (as shown in Fig. 6a). For the switch to be actuated, the magnitude and direction of the electric field in the piezoelectric material has to be preserved. The polarity and value of the potential on any particular electrode is inconsequential.



NEMS Piezoelectric Switches, Fig. 6 (**a**) Schematic representation of the complementary operation made possible in piezoelectric switches by using the body-biasing method. By changing the polarity of the body-bias and the electrodes to which it is applied, the nature of the switch (p- or n-like) can be modified without any alteration to its layout design or fabrication. (**b**, **c**) Experimental data verifying that the switch threshold voltage (V_{th}) can be tuned by varying the value of the body bias. This data show that the body-bias technique enables the implementation of variable threshold voltages on the same die and in the same process. The current measurement was limited to a maximum value of 100 μA

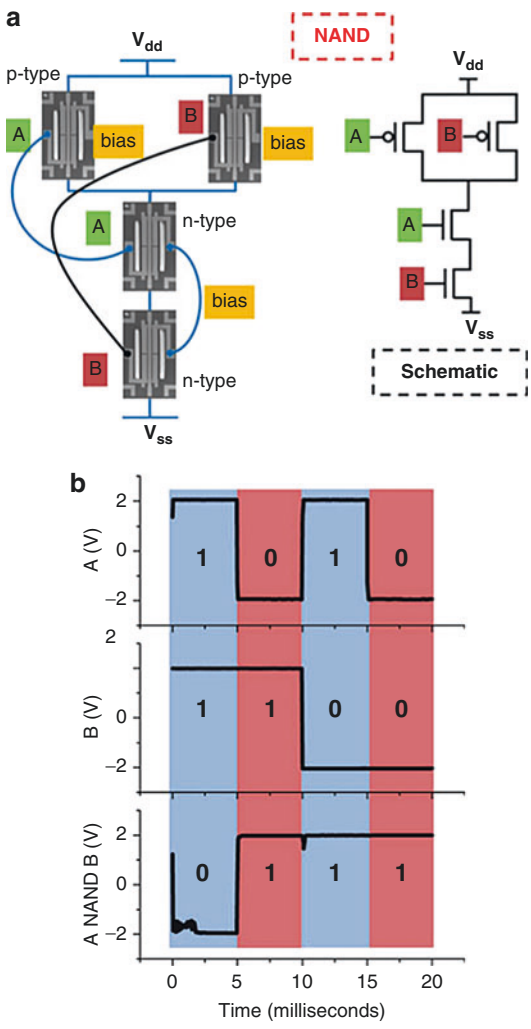
The subthreshold slope has been measured to be <1 mV/dec as the actuation of these structures is not limited by thermal noise. This gives these switches with body-bias the ability to lose nearly no power when they are in standby.

An Agilent B1500A Semiconductor Parameter Analyzer was used for measuring the switch threshold voltage as a function of body bias applied to the MT. Figure 6b, c illustrate the results of 11 successive scans, each, on a 200 × 200 μm^2 device, showing the change in the threshold voltage (V_{th}) as the body-bias (V_{bias}) is increased with a 250 mV step per scan, both for an n-like and a p-like MT. Equivalent to an n-type CMOS transistor, the MT can be made to operate as a depletion or enhancement mode device depending on the value of the body-bias. The absolute value of the body-bias voltages were the same for both the scan sets. Good agreement between threshold values for corresponding positive and negative body-bias values was observed, but small differences in the magnitude of the threshold voltage (10–120 mV) between the n-and p-like cases were recorded and can be attributed to subtle changes in the adhesive forces due to the evolution of the contact. These data conclusively show that, by using body-biasing the same switch can be made to behave either as an n-like switch or as a p-like switch.

Logic-Building Blocks Using MEMS/NEMS Switches

For initial demonstration of mechanical logic achieved by AlN MTs the logic circuit layout was adopted directly from traditional CMOS design and mechanical transistors were used to simply replace the CMOS transistors. This approach was used to implement and validate the ability of realizing low swing voltage dynamically operated complementary piezoelectric logic circuits via body-biasing.

These three-finger dual-beam AlN MEMS switches have been able to demonstrate logic elements like inverters, NAND, and NOR gates. The key factors that have enabled this low voltage demonstration are the use of thin AlN films (250 nm) and of the body-biasing technique. Two AlN MEMS switches were used to demonstrate a low voltage digital inverter using body-biased AlN switches capable of fast switching (~220 ns). The mechanical inverter formed was tested at 100 Hz with a ± 1.5 V (square input wave) voltage swing. The NAND and NOR structures were formed by wire-bonding four separate switches in the desired configuration. The mechanical NAND was formed by using AlN mechanical switches with one input at 100 Hz and the other at 50 Hz with a ± 2 V voltage swing, as shown in Fig. 7a. The experimental results and superimposed equivalent truth table for the dynamic testing of a single cycle of a mechanical NAND gate is shown in Fig. 7b.



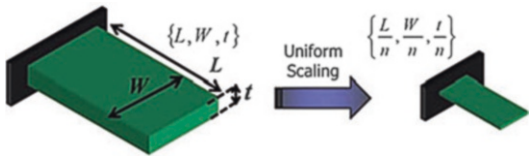
NEMS Piezoelectric Switches, Fig. 7 (**a**) Schematic representation of the switch layout showing the wiring that was required to form the NAND gate. The electronic schematic of a conventional NAND is also shown to illustrate the similarities between the two implementations. (**b**) Experimental data for one cycle of operation of a NAND gate made using mechanical switches and its corresponding truth table. This implementation was achieved with a ± 2 V swing

AlN Device Scaling and Nano-Actuator Development

The necessity to miniaturize the MT and improve device density, scaling of AlN MTs has to be investigated and projected. The results of an analysis of how dimensional scaling affects various important actuator parameters such as displacement, δ_E, resonance frequency, f_n, contact force, Fc, gap size, g, as well as the energy consumption, $(En)_E$, and power consumption, P_E, of a simple logic gate is show in Table 1. The analysis was carried out assuming uniform scaling of the actuator (Fig. 8), i.e., scaling of all the dimensions of the actuator together by the same factor. The other assumption of this study was the use of constant electric field, i.e., the magnitude of the electric field in the piezoelectric was kept constant throughout the scaling process. From the prospective of energy and power consumed the advantages of scaling are amply visible in the table. It is interesting to note the similarities to CMOS

NEMS Piezoelectric Switches, Table 1 Scaling factors for constant-field scaling of piezoelectric MTs. Scaling factor: k

Property	Relation to independent variables (L, w, t, V)	Constant-field scaling
Displacement (∂_E)	$\propto \frac{L^2}{t}$	$\propto \frac{1}{k^1}$
Contact force (F_C)	$\propto \frac{wt^2}{L}$	$\propto \frac{t}{k^2}$
Resonance frequency (f_n)	$\propto \frac{t}{L^2}$	$\propto k^1$
Gap (g)	$\propto \frac{L^2}{t}$	$\propto \frac{1}{k^1}$
Body-bias voltage	$\propto \frac{L^2}{t}$	$\propto \frac{1}{k_1}$
Energy $(En)_E$	$\propto wV$	$\propto \frac{1}{k^3}$
Power (P_E)	$\propto \frac{w_W}{L}$	$\propto \frac{1}{k^2}$

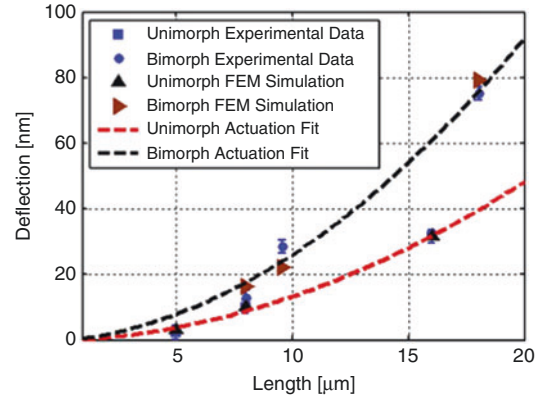


NEMS Piezoelectric Switches, Fig. 8 3D schematic showing uniform scaling as has been proposed for the mechanical transistors. All the dimensions of the actuator should be scaled by the same factor

scaling. The scaling of the channel length and the thickness of the gate oxide are effectively analogous to the scaling of the actuator length and the piezoelectric film thickness. As for CMOS scaling, MT scaling also leads to a linear increase in speed and quadratic reduction in power.

In the past, the biggest hurdle in the scaling of piezoelectric transduction to the nanoscale has been the degradation of piezoelectric properties due to limited orientation in thin films and increases in internal stresses (cracking and excessive deformations in released structures). With better control over the deposition parameters and extensive process development ultrathin AlN films have been successfully developed and demonstrated [34–36].

Nanoactuators have been developed to study the effect of scaling (to the nano-regime) on the piezoelectric properties of AlN. The films were successfully scaled down to the thickness of



NEMS Piezoelectric Switches, Fig. 9 Results of deflection testing of beams of different lengths with unimorph actuation (single layer) and bimorph actuation (double layer) at 4 V and their agreement with FEM analysis. The *dashed line* shows the square dependence that exists analytically between tip displacement and beam length

100 nm, and actuators as well as switches were fabricated using them. DC voltage-induced nanobeam deflections of six beams of different lengths (5–18 μm) under unimorph and bimorph actuation at 4 V are shown in Fig. 9. The measurements were made by a Zygo optical profilometer. Both layers of AlN have been individually actuated and the deflection measurement for each individual layer and both the layers actuated together closely match the FEM predictions (using COMSOL). This matching with FEM analysis verifies the dependence of vertical displacement on the square of the beam length as predicted by conventional Euler beam-bending equations. These deflection measurements along with analytical solutions for the deflection of piezoelectric unimorph and bimorph beams [37] were used for extracting the d_{31} piezoelectric coefficient of the ultrathin AlN film. The extracted experimental d_{31} is within 5 % of the theoretical value of the piezoelectric coefficient.

Finally, the successful scaling of the AlN film led to the demonstration of an AlN piezoelectric nanoswitch that employed 100 nm thick AlN films. These switches validate the use of the body-biasing technique at the nanoscale. As the thickness of the films has been scaled to the nano-dimensions the body-bias voltages have been lowered to <6 V and a better control over the

actuator motion has been demonstrated. This enhanced control over switching translates into a threshold voltage of ~1 mV and the ability to potentially operate with supply voltages of few 10 s of mV.

These demonstrations, both in the RF domain as well as in the field of mechanical computing, clearly show the scalability and the potential that this piezoelectric mechanical switch/MT technology possesses. Given continued effort this can lead to development of low power highly integrated RF and computing systems.

Cross-References

- ▶ Piezoelectric Effect at Nanoscale
- ▶ Piezoelectric MEMS Switch

References

1. Petersen, K.E.: Micromechanical membrane switches on silicon. IBM J. Res. Dev. **23**, 376–385 (1979)
2. Rebeiz, G.M.: RF MEMS Theory, Design and Technology, 1st edn. Wiley, Hoboken (2003)
3. Zavracky, P.M., Majumder, S., McGruer, N.E.: Micromechanical switches fabricated using nickel surface micromachining. J. Microelectromech. Syst. **6**(1), 3–9 (1997)
4. Bozler C., Drangmeister, R., Duffy, S., Gouker, M., Knecht, J., Kushner, L., Parr, R., Rabe, S., Travis, L.: MEMS microswitch arrays for reconfigurable distributed microwave components. In: IEEE MTT-S International Microwave Symposium Digest, Boston, pp. 153–156 (2000)
5. Weon Wi, J., JeongOen, L., Hyun-Ho, Y., Jun-Bo, Y.: Mechanically operated random access memory (MORAM) based on an electrostatic microswitch for nonvolatile memory applications. IEEE Trans. Electron Dev. **55**(10), 2785–2789 (2008)
6. Chong, S., Akarvardar, K., Parsa, R., Yoon, J.-B., Howe, R.T., Mitra, S., Wong, H.-S.P.: Nanoelectromechanical (NEM) relays integrated with SRAM for improved stability and low leakage. In: IEEE/ACM International Conference on Computer-Aided Design, San Jose, 2–5 Nov 2009
7. Parsa, R., Akarvardar, K., Provine, J., Lee, D., Elata, D., Mitra, S., Wong, H.-S.P., Howe, R.T.: Composite polysilicon-platinum lateral nanoelectromechanical relays. In: 14th Solid-State Sensors, Actuators, and Microsystems Workshop, Hilton Head, 6–10 June 2010, pp. 7–10
8. Nathanael, R., Pott, V., Kam, H., Jeon, J., King Liu, T.-J.: 4-terminal relay technology for complementary logic. In: IEDM Technical Digest, pp. 223–226 (2009)
9. Spencer, M., Chen, F., Wang, C.C., Nathanael, R., Fariborzi, H., Gupta, A., Kam, H., Pott, V., Jeon, J., Liu, T.-J.K., Marković, D., Alon, E., Stojanović, V.: Demonstration of integrated micro-electro-mechanical relay circuits for VLSI applications. IEEE J. Solid State Circuits **46**(1), 308–320 (2011)
10. Kam, H., Pott, V., Nathanael, R., Jeon, J., Alon, E., King Liu, T.-J.: Design and reliability of a micro-relay technology for zero-standby-power digital logic applications. In: IEDM Technical Digest, pp. 809–812 (2009)
11. Abele, N., Fritschi, R., Boucart, K., Casset, F., Ancey, P., Ionescu, A.M.: Suspended-gate MOSFET: bringing new MEMS functionality into solid-state MOS transistor. In: IEEE International Electron Devices Meeting. IEDM Technical Digest, pp. 479–481, 5–7 Dec 2005
12. Duffy, S., Bozler, C., Rabe, S., Knecht, J., Travis, L., Wyatt, P., Keast, C., Gouker, M.: MEMS microswitches for reconfigurable microwave circuitry. IEEE Microwave Wireless Compon. Lett. **11**(3), 106–108 (2001)
13. Lee, H.C., Park, J.H., Park, J.Y., Nam, H.J., Bu, J.U.: Design, fabrication and RF performances of two different types of piezoelectrically actuated Ohmic MEMS switches. J. Micromech. Microeng. **15**, 2098–2104 (2004)
14. Park, J.H., Lee, H.C., Park, Y.H., Kim, Y.D., Ji, C.H., Bu, J., Nam, H.J.: A fully wafer-level packaged RF MEMS switch with low actuation voltage using a piezoelectric actuator. J. Micromech. Microeng. **16**, 2281–2286 (2006)
15. Gross, S.J., Tadigadapa, S., Jackson, T.N., Trolier-McKinstry, S., Zhang, Q.Q.: Lead-zirconate-titanate-based piezoelectric micromachined switch. Appl. Phys. Lett. **83**(1), 174–176 (2003)
16. Judy, D.C., Pulskamp, J.S., Polcawich, R.G., Currano, L.: Piezoelectric Nanoswitch. In: 22nd IEEE International Conference on Micro Electro Mechanical Systems, pp. 591–594, 25–29 Jan 2009
17. Proie, R.M., Polcawich, R.G., Pulskamp, J.S., Ivanov, T., Zaghloul, M.E.: Development of a PZT MEMS switch architecture for low-power digital applications. J. Microelectromech. Syst. **20**(4), 1–11 (2011)
18. Judy, D., Polcawich, R.G., Pulskamp, J.: Low voltage complementary MEMS logic using piezoelectric actuators. In: 2008 Solid State Sensor, Actuator and Microsystems Workshop, Hilton Head Island, pp. 328–331 (2008)
19. Sinha, N., Jones, T.S., Guo, Z., Piazza, G.: Demonstration of low voltage and functionally complete logic operations using body-biased complimentary and ultra-thin AlN piezoelectric mechanical switches. In: The 23rd IEEE Conference on Micro Electro Mechanical Systems (MEMS 2010), Hong Kong, pp. 751–754 (2010)

20. Sinha, N., Jones, T.S., Guo, Z., Piazza, G.: Body-biased complementary logic implemented using AlN piezoelectric MEMS switches. In: 2009 I.-E. International Electron Devices Meeting (IEDM 2009), Baltimore, pp. 813–816 (2009)
21. Mahamameed, R., Sinha, N., Pisani, M.B., Piazza, G.: Dual beam actuation of piezoelectric AlN RF mems switches monolithically integrated with AlN contour-mode resonators. J. Micromech. Microeng. **18**, 105011 (2008)
22. Sinha, N., Mahameed, R., Zuo, C., Piazza, G.: Integration of AlN micromechanical contour-mode technology filters with three-finger dual beam AlN MEMS switches. In: Frequency Control Symposium, 2009 Joint with the 22nd European Frequency and Time forum. IEEE International, pp. 1–4, 20–24 Apr 2009
23. Polcawich, R.G., Judy, D., Pulskamp, J.S., Trolier-McKinstry, S., Dubey, M.: Advances in piezoelectrically actuated RF MEMS switches and phase shifters. In: Microwave Symposium, IEEE/MTT-S International, pp. 2083–2086 (2007)
24. Ruan, M., Shen, J., Wheeler, C.B.: Latching micromagnetic relays. J. Microelectromech. Syst. **10**, 511–517 (2001)
25. Streeter, R.D., Hall, C.A., Wood, R., Madadevan, R.: VHF high-power tunable RF bandpass filter using microelectromechanical (MEM) microrelays. Int. J. RF Microwave CAE **11**(5), 261–275 (2001)
26. Blondy, P., Cros, D., Rey, P., Charyet, P., Diem, B., Zanchi, C., Quorin, J.B.: Low voltage high isolation MEMS switches. In: Silicon Monolithic Integrated Circuits RF Systems, pp. 47–49 (2001)
27. Zuo, C., Sinha, N., Piazza, G.: Novel electrode configurations in dual-layer stacked and switchable AlN contour-mode resonators for low impedance filter termination and reduced insertion loss. In: 23rd IEEE International Conference on Micro Electro Mechanical Systems (MEMS 2010), pp. 719–722 (2010)
28. Piazza, G., Stephanou, P.J., Pisano, A.P.: Single-chip multiple-frequency AlN MEMS filters based on contour-mode piezoelectric resonators. J. Microelectromech. Syst. **16**(2), 319–328 (2007)
29. Smits, J.G., Choi, W.-S.: The constituent equations of piezoelectric heterogeneous bimorphs. IEEE Trans. Ultrason. Ferroelect. Freq. Contr. **38**(3), 256–270 (1991)
30. Smits, J.G., Dalke, S.I., Cooney, T.K.: The constituent equations of piezoelectric bimorphs. Sensor. Actuator. A. **28**(1), 41–61 (1991)
31. Sinha, N., Wabiszewski, G.E., Mahameed, R., Felmetsger, V.V., Tanner, S.M., Carpick, R.W., Piazza, G.: Piezoelectric aluminum nitride nanoelectromechanical actuators. Appl. Phys. Lett. **95**, 053106 (2009)
32. Zuo, C., Sinha, N., Pisani, M.B., Perez, C.R., Mahameed, R., Piazza, G.: Channel-select RF MEMS filters based on self-coupled AlN contour-mode piezoelectric resonators. In: IEEE Ultrasonics Symposium, pp. 1156–1159, 28–31 Oct 2007
33. Piazza, G., Stephanou, P.J., Pisano, A.P.: One and two port piezoelectric higher order contour-mode MEMS resonators for mechanical signal processing. Solid State Electron. **51**(11–12), 1596–1608 (2007)
34. Felmetsger, V., Laptev, P., Tanner, S.: Crystal orientation and stess in AC reactively sputtered AlN films on Mo electrodes for electro-acoustic devices. In: Proceedings IEEE IUS 2009, Beijing, pp. 2146–2149 (2008)
35. Felmetsger, V.V., Laptev, P.N., Graham, R.J.: Deposition of ultrathin AlN films for high frequency electro-acoustic devices. J. Vac. Sci. Technol. A. **29**(2), 021014–021014-7 (2011)
36. Sinha, N., Guo, Z., Felmetsger, V.V., Piazza, G.: 100 nm thick aluminum nitride based piezoelectric nano switches exhibiting 1 mV threshold voltage via body-biasing. In: Proceedings of 2010 Solid-State Sensors, Actuators, and Microsystems Workshop, Hilton Head 2010, pp. 352–355 (2010)
37. DeVoe, D.L., Pisano, A.P.: Modeling and optimal design of piezoelectric cantilever microactuators. J. MEMS **6**(3), 266–270 (1997)

NEMS Resonant Chemical Sensors

Matteo Rinaldi
Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA

Synonyms

NEMS gravimetric chemical sensors; NEMS gravimetric sensors; NEMS mass sensors; NEMS resonant mass sensors

Definition

Nanoelectromechanical System (NEMS) Chemical Sensors are devices that combine electrical and mechanical functionalities at the nanoscale for the detection of minute concentrations of target gaseous compounds in the environment. A NEMS Resonant Chemical Sensor is composed of three fundamental elements: an electrically driven nanomechanical resonator (NEMS resonator), a chemical interactive material (CIM) deposited on

the surface of the NEMS resonator, and a readout system. The natural resonance frequency of the nanomechanical resonator is highly sensitive to variations in the device mass per unit area. Therefore, any small amount of mass per unit area loaded on the nanomechanical resonator will produce a relatively large shift in the natural resonance frequency of the device. The smaller is the mass per unit area of the mechanical resonator itself, the larger is the shift in resonance frequency for a given amount of mass per unit area loaded on the device surface. Therefore, an ultraminiaturized (nano) mechanical resonator, with an extremely low mass per unit area, is ultrasensitive to mass loading per unit area. When the sensor is exposed to a certain concentration of a target gaseous compound, a number of gas molecules per unit area (mass per unit area), depending on the affinity between the CIM and the target gas, are absorbed by the CIM on the NEMS resonator surface causing a shift in its natural resonance frequency. Since a specific amount of mass per unit area is absorbed by the CIM for a given concentration of gas in the environment and a specific shift in resonance frequency is caused by a given amount of mass per unit area loaded on the NEMS resonator, it is possible to quantitatively detect the concentration of the target gas by monitoring the natural resonance frequency of the NEMS resonator. The resonance frequency of the NEMS resonator is monitored over time and converted into an electrical signal by the readout system. The fundamental performances of a NEMS resonant chemical sensor are expressed in terms of limit of detection (LOD), response time, and selectivity. The LOD is the smallest amount of adsorbed mass per unit area (hence minimum gas concentration given a certain CIM) that can be resolved by the sensor. This quantity is inversely proportional to the resonator sensitivity to mass loading per unit area and directly proportional to the frequency fluctuations induced by the noise generated in the resonator and in the readout system. Therefore, a nanomechanical resonator characterized by a very large sensitivity to mass loading per unit area and a very low noise is the best candidate for the implementation of a high performance NEMS resonant chemical sensor. Response time and selectivity are instead determined by the chemical and physical properties of the CIM employed.

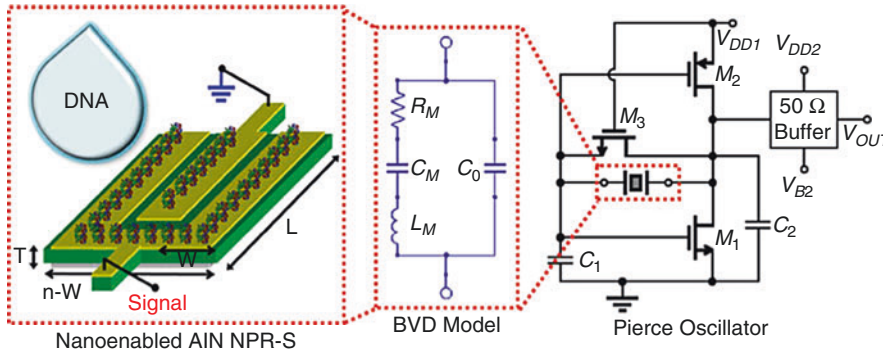
Main Text

In recent years, the demand for highly miniaturized sensor arrays capable of selectively detecting extremely small concentrations (part per trillion, ppt) of multiple gaseous analytes has steadily grown. The necessity to detect such small concentrations requires reliably measuring extremely small variations in the sensor output signal. In this perspective, gravimetric sensors, such as Quartz Crystal Microbalances (QCMs), have a significant advantage over other sensor technologies (conductance-based sensors, Chem-FET, or optical sensors) since they use frequency as the output variable, which is one of the physical quantities that can be monitored with the highest accuracy [1].

Nevertheless, smart, compact, portable, and low cost electronic noses for multiple gas detection cannot be implemented with bulky and unintegrable QCMs. Even if QCMs have been successfully employed as gravimetric sensors, thanks to their extremely high quality factors, which permit to obtain limits of mass detection in the order of few nanograms, their relatively large volume and their inability to be directly integrated on silicon render them unattractive for the fabrication of sensor arrays composed of a large number of mechanical elements.

In response to this growing demand for miniaturization, the use of microelectromechanical (MEMS) and nanoelectromechanical (NEMS) resonators for sensing applications has been extensively explored. The scaling of the device dimensions enables the fabrication of ultrasensitive nanomechanical gravimetric sensors that can be arranged in single-chip arrays capable of selectively detecting extremely small concentrations (ppb-ppt) of multiple gaseous analytes.

Different MEMS/NEMS resonator technologies based on electrostatic or piezoelectric



NEMS Resonant Chemical Sensors, Fig. 1 Schematic representation of the fabricated nanoenabled AlN NPR-S connected in a self-sustained oscillator loop (Pierce oscillator). The butterworth van dyke (*BVD*) model is used to describe the resonator behavior

transduction [2, 3] or piezoresistive displacement transduction [4] have been proposed. In particular, NEMS resonators have been exploited as transducers for the realization of extremely sensitive gravimetric sensors. Sub-attogram mass resolution has been demonstrated in NEMS cantilevers [4], thanks to the minuscule mass and the relatively high quality factor (Q). Nevertheless the greatly reduced dimensions of these devices render their transduction extremely difficult and require the use of cumbersome, complex, and power-inefficient readout techniques.

In this perspective, optimal sensor performance is attained by synthesizing a transducer that occupies a large area (which facilitates efficient transduction) and is very thin (which allows fabricating low mass devices with ultrahigh sensitivity). Suspended membranes with thickness in the nanometer range are therefore desirable. Flexural plates (instead of beams) are a good compromise [5], but higher frequencies and power handling are achieved in bulk mode acoustic resonators [3, 6].

A particularly representative example of high performance bulk mode acoustic NEMS resonant sensors that will be described here involves exciting high frequencies bulk acoustic waves in a piezoelectric nano-plate (thickness 250 nm) made of Aluminum Nitride (AlN). Such AlN Piezoelectric Nano-Plate Resonant Sensor (NPR-S) technology is not only characterized by high values of sensitivity, due to the reduced mass and high frequency of operation of the nanomechanical resonant element, but it is also associated with low noise performance, due to the combination of high quality factor, Q, and power handling capability of the bulk acoustic wave NEMS resonators. In addition, the AlN nanoplate composing the NPR-S can be efficiently actuated and sensed piezoelectrically on-chip solving the fundamental transduction issues associated with electrostatically transduced NEMS resonators and enabling the use of compact and low power CMOS circuits for electronic readout (Fig. 1). The sensor affinity for the adsorption of volatile organic chemicals such as 2,6 dinitrotoluene is enhanced by functionalizing the top gold electrode of the device with a thiol-terminated single stranded DNA sequence (Thiol – 5′ CTT CTG TCT TGA TGT TTG TCA AAC 3′) and the ability to detect concentrations in the ppt range is demonstrated. Thanks to the unique combination of extremely important features such as high sensitivity, low noise, efficient transduction, small form factor, and CMOS integration capability, the NPR-S represents one of the most suitable technologies to devise high performance portable gas sensors arrays that can be easily incorporated in wireless devices and sensor networks for distributed environmental or personal health monitoring.

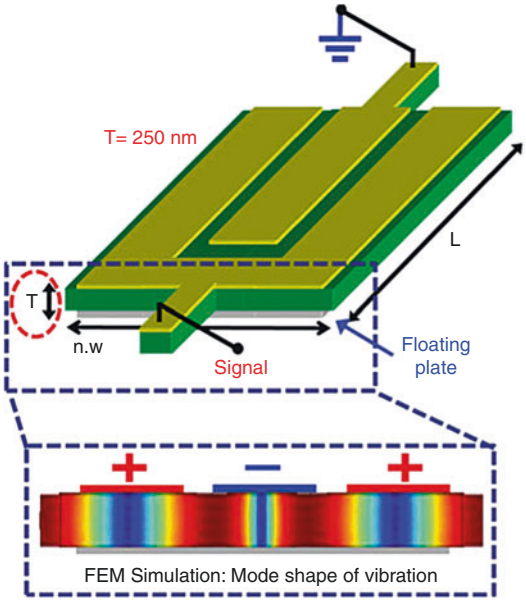
NEMS Resonant Chemical Sensor Design

The fundamental metrics that need to be considered for the design of high performance NEMS

resonant chemical sensors are expressed in terms of limit of detection (LOD), response time, and selectivity. The LOD is the smallest amount of adsorbed mass per unit area (hence minimum gas concentration given a certain chemical interactive material, CIM) that can be resolved by the sensor. This quantity is inversely proportional to the resonator sensitivity to mass loading per unit area and directly proportional to the frequency fluctuations induced by the noise generated in the resonator and in the readout system. Therefore, a nanomechanical resonator characterized by a very large sensitivity to mass loading per unit area and a very low noise is the best candidate for the implementation of a high performance NEMS resonant chemical sensor. Response time and selectivity are instead determined by the chemical and physical properties of the CIM employed. A particularly representative example of high performance NEMS resonant chemical sensor design is given by the Aluminum Nitride (AlN) Nano-Plate Resonant Sensor (NPR-S) technology, which will be described in details herein.

The Aluminum Nitride (AlN) Nano-Plate Resonant Sensor (NPR-S) technology involves exciting a high frequency contour-extensional mode of vibration (Fig. 2) in piezoelectric nano-plates (thickness 250 nm) made out of AlN. The capability to excite higher order contour-extensional mode of vibration in micro-scale AlN plates has been demonstrated by the AlN Contour-Mode resonator (CMR) technology [3, 7, 8]. High performance AlN Contour-Mode resonators in the very high frequency (VHF), ultrahigh frequency (UHF), and superhigh frequency (SHF) range (20–10 GHz) and quality factors, *Qs*, between 1,000 and 3,000 have been reported [7–9]. Piezoelectric sensing and actuation of these devices have been successfully employed in the MEMS domain. While the scaling into the nano-domain of the AlN CMR core technology has been exploited for the implementation of AlN Nano-Plate Resonators with unique performances in sensing applications [10–15].

The reduced mass of the mechanical resonator along with the high quality factor values and the power handling capabilities make the AlN NPR-S technology capable of achieving unprecedented values of limit of detection to mass per unit area in the order of zg/μm^2.



NEMS Resonant Chemical Sensors, Fig. 2 Schematic representation of an AlN Nano-Plate Resonator. The *inset* shows a 2-D FEM simulation of the device mode of vibration

An AlN NPR-S is composed of an AlN nano-plate sandwiched between two metal electrodes (Fig. 2). When an AC voltage is applied across the thickness (*T*) of the device, a contour-extensional mode of vibration is excited through the equivalent d_{31} piezoelectric coefficient of AlN. Given the equivalent mass density, ρ_{eq}, and Young's modulus, E_{eq}, of the material stack (AlN and metal electrodes) that forms the resonator, the center frequency, f_0, of this laterally vibrating mechanical structure is univocally set by the period, *W*, of the metal electrode patterned on the AlN plate and can be approximately expressed as in Eq. 1.

$$f_0 = \frac{1}{W}\sqrt{\frac{E_{eq}}{\rho_{eq}}} \tag{1}$$

The other two geometrical dimensions, thickness, *T*, and length, *L*, set the equivalent electrical impedance of the resonator [3] and can be

designed independently of the desired resonance frequency.

The sensitivity to mass per unit area of an AlN NPR-S loaded on its top surface can be expressed as in Eq. 2 [10, 16].

$$S = -\frac{f_0}{2\rho_{eq}T} \tag{2}$$

It is worth noting that the resonance frequency, f_0, and the thickness, T, of the AlN NPR-S are two independent variables while for QCMs they are intrinsically coupled and cannot be set independently from one another. This is an important and unique advantage of AlN NPR-S. In fact, it permits to set the frequency of operation of these devices according to the specifications of the desired application and independently obtain the required value of mass sensitivity by scaling the device thickness.

Nevertheless, the mass sensitivity of the device cannot be considered the only important parameter for the design of a high performance gravimetric sensor. In fact, the limit of detection, *LOD*, of the sensor (smallest amount of adsorbed mass per unit area that can be resolved) is defined as in Eq. 3:

$$LOD = \frac{\Delta f_{\min}}{S} \tag{3}$$

where $\Delta f_{\min}$ indicates the minimum frequency shift detectable by the sensor readout. Thus, the performance of the sensor can be improved by reducing the thickness, T, of the AlN layer up to the limit at which the smallest value of the ratio, $\Delta f_{\min}/f_0$, is preserved. For example, $\Delta f_{\min}/f_0$ comparable to thicker films and as low as 9×10^{-8} has been attained for a nanoscaled device.

Another crucial aspect directly related to the scaling of the device is the complexity, size, and power efficiency of the required sensor readout. In particular, the ability to connect the nanoelectromechanical resonator to a self-sustained oscillator loop for direct frequency readout greatly simplifies the design of multiplexed sensor platforms involving large arrays of resonant devices. Self-sustained, high performance, Pierce-like oscillator circuits based on thick film (2–3 μm) AlN Contour-Mode MEMS Resonator (CMR) have been demonstrated [17]. The primary power loss in these oscillator circuits is due to the motional resistance, *Rm*, of the resonator [17], whose value is inversely proportional to the device figure of merit, $k_t^2 \cdot Q$ [17], where k_t^2 represents the device electromechanical coupling and Q its quality factor. Therefore, the $k_t^2 \cdot Q$ product of the AlN NPR-S needs to be maintained sufficiently large (>10) to be able to connect it to a compact and low power self-sustained oscillator circuit for direct frequency readout.

A particular device design that proves the scaling advantages of the AlN NPR-S technology consists of a 220 MHz ($W = 15$ μm) AlN resonant nano-plate whose thickness was set to be, $T = 250$ nm, in order to improve the sensitivity of the resonant sensor.

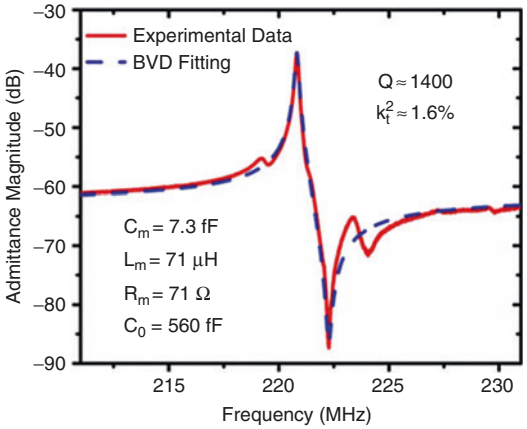
The oscillator circuit topology used as readout for this 250-nm-thick AlN NPR-S is described in more detail in [17]. Briefly, the circuit (Fig. 1) consists of a Pierce oscillator implemented by means of a CMOS inverter biased in its active region. Transistors M1 and M2 form the CMOS inverter while transistor M3 acts as a large resistor to provide biasing in the active region of M1 and M2.

Being functionalization another important aspect in the design of a resonant sensor, an interface layer between the CIM and the VOC has to be selected. In the case of this 250-nm-thick ALN NPR-s, the top gold (Au) electrode was chosen and used to attach thiolated single stranded DNA (HS-ss-DNA) sequences [18]. The capability of single stranded DNA to selectively enhance the adsorption of volatile organic compounds such as dimethyl-methylphosphonate (DMMP, a stimulant for nerve agent sarin) and 2,6 dinitrotoluene (DNT, a stimulant for explosive vapors) has been recently demonstrated [16]. Therefore, in order to enhance the sensor affinity for the adsorption of volatile organic chemicals, the top gold electrode of the AlN NPR-S device was directly functionalized with a thiol-terminated ss-DNA sequence. The capability of employing CIMs with increased effective adsorbing area based on single wall carbon nanotubes (SWNTs) directly

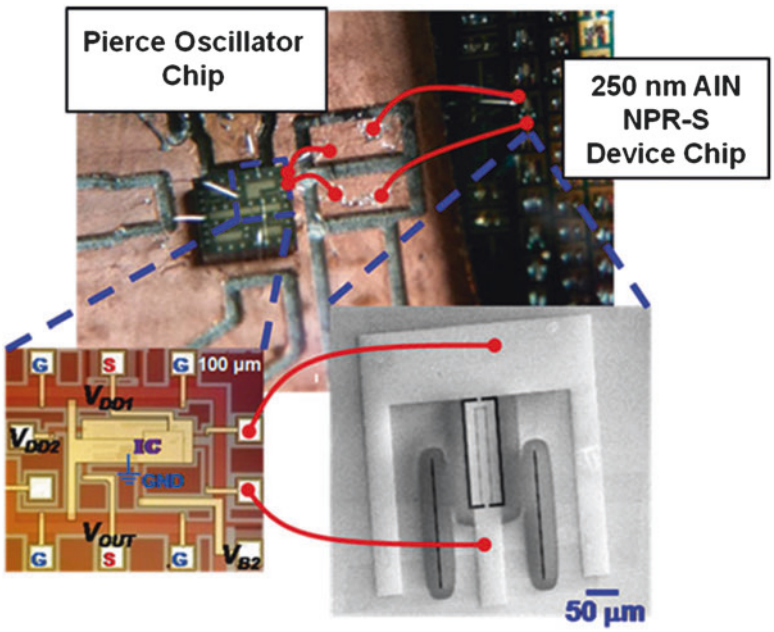
grown on the top surface of a micro-scale CMR and functionalized by DNA strands has also been demonstrated [16].

NEMS Resonant Chemical Sensor Performance

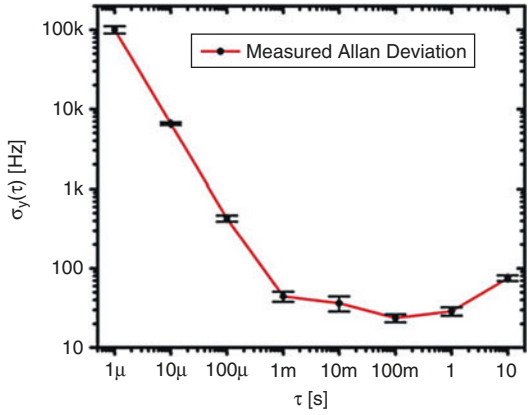
The performances of NEMS resonant chemicals sensor can be fully characterized by performing three sets of experiments: (1) Open loop measurements (i.e., frequency response of the nanomechanical resonator in Fig. 3) are performed to extract the device sensitivity to mass loading and extract the electromechanical parameters of the NEMS resonator [10]. (2) The NEMS resonator is then connected to a self-sustained oscillator circuit (Fig. 4) for direct frequency readout and closed-loop characterization of the resonant sensor performance in terms of noise and power consumption is performed. (3) Finally the sensor is exposed to different concentrations of the target gaseous compounds to verify detection capability. For the ss-DNA-functionalized 250-nm AlN NPR-S described herein, an open loop measurement (frequency response) was first performed and the sensitivity to mass loading was extracted to be approximately 41.6 KHz·μm^2/fg [10]. Despite the device volume reduction, high Q (~1,400 in air) and high k_t^2 (~1.6 %) were attained (Fig. 3) enabling the direct wire bonding of the AlN NPR-S to an oscillator circuit chip fabricated in the ON Semiconductor 0.5 μm CMOS process. The 220 MHz oscillator used a 1.67 V supply voltage and 300 μA of biasing current, which give a total power consumption of 501 μW.



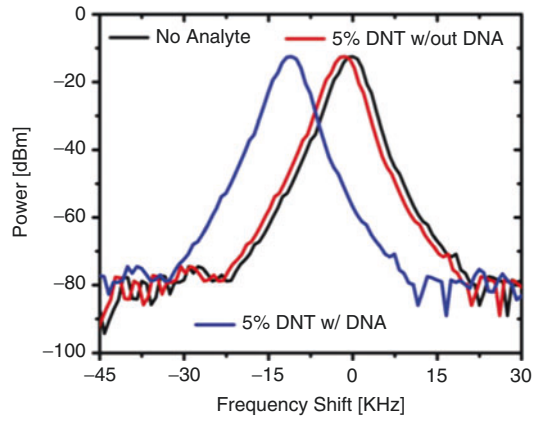
NEMS Resonant Chemical Sensors, Fig. 3 Experimental and fitted (BVD model) admittance curves of the fabricated AlN NPR-S



NEMS Resonant Chemical Sensors, Fig. 4 Picture of the AlN NPR-S wire-bonded to the Pierce oscillator chip fabricated in the ON Semiconductor 0.5 μm CMOS process



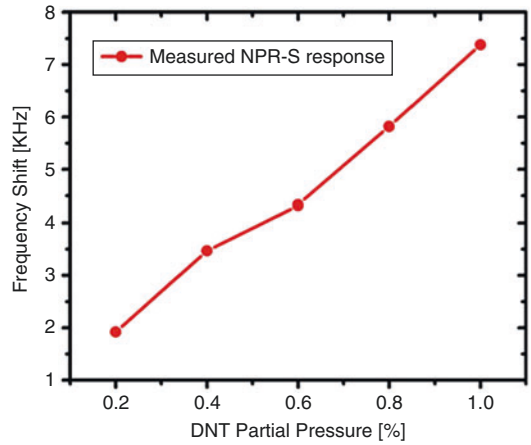
NEMS Resonant Chemical Sensors, Fig. 5 Measured Allan deviation of the 220 MHz oscillator



NEMS Resonant Chemical Sensors, Fig. 6 Sensor response to 5 % DNT vapor pressure before and after DNA functionalization

In order to estimate the noise limited minimum frequency shift (Δf_{min}) detectable by the readout of a NEMS resonant chemical sensor, the short-term frequency stability of the oscillator output signal is characterized by measuring its Allan deviation, $\sigma_Y(\tau)$, [19]. The measured values of Allan deviation for the 220 MHz AlN NPR-S described here are reported in Fig. 5. The minimum value of Allan deviation recorded is approximately 20 Hz and it is achieved for a gate time, τ, of 100 ms. Therefore, considering a minimum detectable frequency shift, Δf_{min}, of 20 Hz, a limit of detection of mass per unit area of about 481 zg/μm^2 can be estimated. This translates to an extrapolated limit of detection for DNT concentrations of about 15 part per trillion (ppt) [20].

The detection capability of a NEMS resonant chemical sensor is finally tested by exposing the device to known concentrations of a target analyte and monitoring the respective frequency shifts of the oscillator output signal. The ss-DNA-functionalized AlN NPR-S in example was exposed to DNT vapor (generated by bubbling method [10]). The improved affinity of the AlN NPR-S to the DNT vapor, due to ss-DNA functionalization (Fig. 6), enabled the detection of very minute analyte concentrations. The measured sensor response for different concentrations of DNT, ranging between 0.2 % and 1 % of the saturated vapor pressure, is reported in Fig. 7. Considering that the DNT saturated vapor pressure at room temperature is equal to 5.67×10^{-4} Torr [20], a concentration of DNT as low as 1.5 ppb was detected by the fabricated nanoenabled AlN NPR-S prototype.



NEMS Resonant Chemical Sensors, Fig. 7 Response of the nanoenabled AlN NPR-S for different concentrations of DNT ranging between 0.2 % and 1 % of the saturated vapor pressure

Cross-References

- ▶ Laterally Vibrating Piezoelectric Resonators
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Piezoelectric Effect at Nanoscale
- ▶ Self-Assembled Monolayers

References

1. Vig, J.R., Filler, R.L., Kim, Y.: Uncooled IR imaging array based on quartz microresonators. J. Microelectromech. Syst. **5**, 131–137 (1996)
2. Wang, K., Wong, A.-C., Nguyen, C.T.C.: VHF free-free beam high-Q micromechanical resonators. J. Microelectromech. Syst. **9**, 347–360 (2000)
3. Piazza, G., Stephanou, P.J., Pisano, A.P.: Piezoelectric aluminum nitride vibrating contour-mode MEMS resonators. J. Microelectromech. Syst. **15**, 1406–1418 (2006)
4. Li, M., Tang, H.X., Roukes, M.L.: Ultra-sensitive NEMS-based cantilevers for sensing, scanned probe and very high-frequency applications. Nat. Nanotechnol. **2**, 114–120 (2007)
5. Lee, H.J., Park, K.K., Oralkan, O., Kupnik, M., Khuri-Yakub, B.T.: CMUT as chemical sensor for DMMP detection. In: 2008 I.E. International Frequency Control Symposium, pp. 434–439 (2008)
6. Ruby, R., Merchant, P.: Micromachined thin film bulk acoustic resonators. In: Proceedings of the 48th IEEE International Frequency Control Symposium, pp. 135–138 (1994)
7. Rinaldi, M., Zuniga, C., Chengjie, Z., Piazza, G.: Super-high-frequency two-port AlN contour-mode resonators for RF applications. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **57**, 38–45 (2010)
8. Rinaldi, M., Zuo, C., Van der Spiegel, J., Piazza, G.: Reconfigurable CMOS oscillator based on multifrequency AlN contour-mode MEMS resonators. IEEE Trans. Electron Devices **58**, 1281–1286 (2011)
9. Rinaldi, M., Zuniga, C., Piazza, G.: 5–10 GHz AlN contour-mode nanoelectromechanical resonators. In: IEEE 22nd International Conference on Micro Electro Mechanical Systems, 2009. MEMS 2009, Sorrento, pp. 916–919 (2009)
10. Rinaldi, M., Zuniga, C., Piazza, G.: Ultra-thin-film AlN contour-mode resonators for sensing applications. In: Ultrasonics Symposium (IUS), 2009 I.E. International, pp. 714–717 (2009)
11. Rinaldi, M., Zuniga, C., Duick, B., Piazza, G.: Use of a single multiplexed CMOS oscillator as direct frequency read-out for an array of eight AlN contour-mode NEMS resonant sensors. In: Proceedings of IEEE Sensors 2010, pp. 2666–2670 (2010)
12. Rinaldi, M., Piazza, G.: Effects of volume and frequency scaling in AlN contour mode NEMS resonators on oscillator phase noise. In: 2011 Joint Conference of the IEEE International Frequency Control and the European Frequency and Time Forum (FCS), pp. 1–5 (2011)
13. Rinaldi, M., Zuniga, C., Piazza, G.: ss-DNA functionalized array of ALN contour-mode NEMS resonant sensors with single CMOS multiplexed oscillator for sub-ppb detection of volatile organic chemicals. In: IEEE 24th International Conference on Micro Electro Mechanical Systems (MEMS 2011), pp. 976–979 (2011)
14. Zuniga, C., Rinaldi, M., Piazza, G.: High frequency piezoelectric resonant nanochannel for bio-sensing applications in liquid environment. In: 2010 I.E. Sensors, pp. 52–55 (2010)
15. Zuniga, C., Rinaldi, M., Piazza, G.: Reduced viscous damping in high frequency Piezoelectric resonant nanochannels for sensing in fluids. In: 2011 I.E. 24th International Conference on Micro Electro Mechanical Systems (MEMS), pp. 960–963 (2011)
16. Zuniga, C., Rinaldi, M., Khamis, S.M., Johnson, A.T., Piazza, G.: Nanoenabled microelectromechanical sensor for volatile organic chemical detection. Appl. Phys. Lett. **94**, 223122–223122-3 (2009)
17. Zuo, C., Van der Spiegel, J., Piazza, G.: 1.05 GHz MEMS oscillator based on lateral-field-excited piezoelectric AlN resonators. In: Frequency Control Symposium, 2009 Joint with the 22nd European Frequency and Time forum. IEEE International, pp. 381–384 (2009)
18. Herne, T.M., Tarlov, M.J.: Characterization of DNA probes immobilized on gold surfaces. J. Am. Chem. Soc. **119**, 8916–8920 (1997)
19. Allan, D.W.: Statistics of atomic frequency standards. Proc. IEEE **54**, 221–230 (1966)
20. Regulatory Determinations Support Document for Selected Contaminants from the Second Drinking Water Contaminant Candidate List (CCL 2), Chapter 7 (2008)

NEMS Resonant Mass Sensors

- ▶ Graphene NEMS
- ▶ NEMS Resonant Chemical Sensors

Neural Activity-Dependent Closed Loop

- ▶ Dynamic Clamp

Neural Interfaces

- ▶ MEMS Neural Probes

N-Methyl-2-pyrrolidone (NMP)

- ▶ SU-8 Photoresist

Nonadiabatic dynamics

► Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Non-conventional Machining

► Ultrasonic Machining

Nonlinear and Parametric NEMS Resonators

Rassul Karabalin
TowerJazz Semiconductors, Newport Beach, CA, USA

Synonyms

Chaos; Coupling; Duffing; Nanoelectromechanical systems; Nanomechanical; NEMS; Nonlinearity; Oscillator; Parametric; Synchronization

Definition

Nonlinear NEMS resonators are nanoelectromechanical system resonators that operate at excitation levels exceeding their linear range. Parametric NEMS resonators are NEMS resonators that are actuated parametrically, i.e., by means of modulating a system parameter at a rate proportional to the resonance frequency.

Introduction: Nonlinearity in NEMS

One of the astonishing examples of miniaturization in recent decades involves the development of micro- and nanoelectromechanical systems (MEMS and NEMS) – tiny moving structures with the sizes down to few atoms across [1]. An important subclass of already diverse field of MEMS and NEMS involves resonant electromechanical systems, which have already shown an unprecedented performance in their ability to sense important physical parameters, probe or manipulate the surfaces of materials, and provide accurate timing reference.

Much of the work in this area has revolved around characteristics of single devices operating in their linear regime. Not only has there been very little progress made on sensor arrays, despite their continuously expressed outstanding potential, but also individual device limitations have been poorly understood, due to the unique nonlinear properties at the nanometer scale. An extensive effort to understand and harness rich nonlinear dynamics of NEMS and MEMS resonators has recently been undertaken by a number of researchers with remarkable results. Thorough understanding of the nonlinear properties is crucial to evaluating the limitations of any practical device's performance. Moreover, the detailed knowledge of nonlinearities led to the development of methods for reducing its detrimental effects and enabled performance enhancement beyond what was considered possible. Besides practical benefits, unique nonlinear properties of NEMS provided an outstanding platform for analyzing complex nonlinear dynamical systems. This entry gives a brief overview of notable findings stemmed from the exploration of nonlinear dynamics in NEMS resonators.

Geometrical shapes, vibration modes, and source materials of miniaturized mechanical resonators vary greatly across the growing diverse field. Each example can be characterized by one or multiple potential sources of nonlinearity. The physical origins of nonlinearity can be divided into three classes:

1. *Material nonlinearity.* The Hooke's law is no longer linear if material stress exceeds linear elasticity range of the structural material. This can be the dominant source in bulk mode resonators [2].
2. *Transduction nonlinearity.* There are several ways transducer can cause nonlinearity. They include readout signal being nonlinear

function of the displacement [3] and the physical nature of the transducer exerting nonlinear force on the resonator [4].

3. *Geometric nonlinearity.* The large vibration displacement distorts mode shape or significant mass displacement occurs in multiple dimensions (*kinetic nonlinearity*), so that the approximation of linear elastic restoring force is invalid. This source is typically dominant in standard NEMS resonators such as beams [5] and cantilevers [6].

Each of the listed sources of nonlinearity is important in certain applications. This entry primarily focuses on nanomechanical systems. Their sizes are typically so small that practical displacements are comparable with dimensional scale; therefore, geometric nonlinearity usually prevails. The transduction nonlinearity is also occasionally important depending on chosen readout mechanism. Thus material nonlinearities rarely play a significant role.

A widespread example of nonlinear nanomechanical system is a doubly clamped beam – a small rectangular mechanical bar with rigid supports on both sides. The nonlinear properties of its flexural out of plane resonant modes are well understood. An example of such system, described in [7], is shown in Fig. 1a. The linear ranges for each relevant physical mechanism are illustrated in Fig. 1b.

Regardless of the origin of nonlinearity, the first correction manifests itself as a next-order nonlinear term in the equation of motion of conventional resonator. The resonator's dynamics is then described by the following expression:

$$m\ddot{x} + \frac{m\omega_0}{Q}\dot{x} + kx + \alpha x^3 = F(t), \qquad (1)$$

where the first three terms on the left-hand side represent regular linear resonator and the fourth term is effectively an elastic restoring force that is proportional to the cube of the displacement (also known as the *Duffing nonlinearity*).

The solution for Eq. 1 can be expressed in an implicit form as a modified Lorentzian:

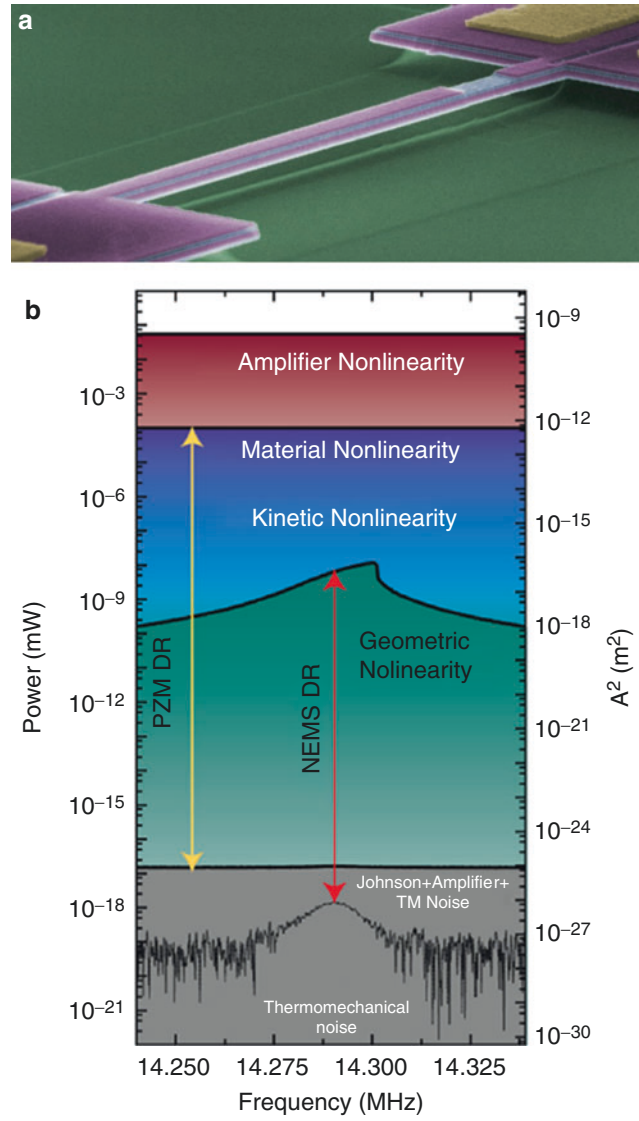
$$x_0 = \frac{F_0}{2m\omega_0^2} \frac{1}{\sqrt{\left(\frac{\omega-\omega_0}{\omega_0} - \frac{3}{8}\frac{\alpha}{m\omega_0^2}x_0^2\right) + \frac{1}{4Q^2}}}. \qquad (2)$$

One obvious consequence of the Duffing nonlinearity, evident from this expression, is frequency pulling. The additional term in the brackets in denominator implies that the effective resonance frequency shifts as the amplitude of motion increases. Another important outcome of nonlinear term is the presence of bistability – indeed Eq. 2 is implicit allowing for several amplitudes of motion to exist for specific drive parameters (amplitude F_0 and frequency ω_0 of the driving force).

Nonlinear Dynamics Experiments: Single Device

NEMS nonlinear properties enabled large number of researchers to perform a variety of experiments, resulting in expanded knowledge of nonlinear systems in general. Their findings illustrated the potential of NEMS resonators as tools for studying discrete nonlinear systems.

As the magnitude of the driving force is increased, bistability manifests itself in the appearance of hysteretic behavior [8]. For certain drive conditions, there exist two stable resonator states represented by two different amplitudes of vibration. The stable state of resonator is then determined by how current drive conditions were attained. If on the other hand the system drive condition is set quickly compared to a period of oscillation, then time dependent dynamical equation has to be considered in order to determine the final state. If the new drive allows for two solutions, then the initial condition of the resonator will determine which stable state to settle. Indeed if the initial amplitude and frequency are close to one of the possible final solutions, then it is natural to settle there. This situation is typically described as the initial condition of the resonator being located within the basin of attraction for this particular stable solution. Nonlinear



Nonlinear and Parametric NEMS Resonators, Fig. 1 (**a**) An SEM image of a doubly clamped beam described in [7]. (**b**) An illustration for the linear ranges of relevant nonlinear mechanisms (Reprinted with a permission from Nano Letters 2013 13 (4), 1622–1626. Copyright 2013. American Chemical Society)

nanomechanical resonator serves as a great platform to conduct the experimental investigation of basins of attraction by perturbing the system and observing the final stable state.

One such experiment, performed by Kozinsky et al. [9], employed nanowire-based nonlinear mechanical resonator to systematically probe the basins of attraction with a very high precision. The experiment benefited from the fact that the relevant time scale per data point, $\sim Q/f_0$, was very short for high-frequency nanoscale devices. This enabled taking many data points before the parameters of the system significantly drifted. The authors observed the separatrix, a border between the two basins of attraction, change shape for varying drive strength, so that one of the basins became progressively smaller and eventually disappeared. The mapped basins of attraction showed a good agreement with theory. However, the observed separatrix was blurred due to ambient fluctuations, including residual noise in the drive system, which caused uncertainty in the preparation of an initial state close to the separatrix.

Such uncertainty gives rise to important physical phenomenon – noise-induced transitions. Ever present noise in any nonlinear system can

cause a spontaneous jump between states. A number of researchers investigated this phenomenon in detail in nonlinear NEMS resonators. In the experiments by Aldridge and Cleland [10], the authors introduced external noise source and varied its intensity. As a result they observed significant modification of hysteretic behavior. Particularly, spontaneous transitions occurred in the vicinity of bifurcation points. Subsequent increase in noise intensity eventually led to blurring of hysteresis width and ultimately to a complete its complete disappearance. In another work by Unterreithmeier et al. [11], the researchers were able to map basins of attraction and observed spontaneous transitions between two stable solutions by means of applying pulsed drive of a given duration, properly chosen to alter the state of the nonlinear resonator.

Since the nonlinear regime is readily accessible in nanoscale devices, the details of dynamical behavior are now increasingly important for proper engineering of any NEMS-based application. Traditionally the understanding of nonlinear properties was used to estimate the dynamic range in order to determine the limits of NEMS performance in linear regime. Recent research, however, has demonstrated that thorough understanding of nonlinear dynamics and incorporating it into a NEMS-based system enables performance enhancement beyond what previously considered possible.

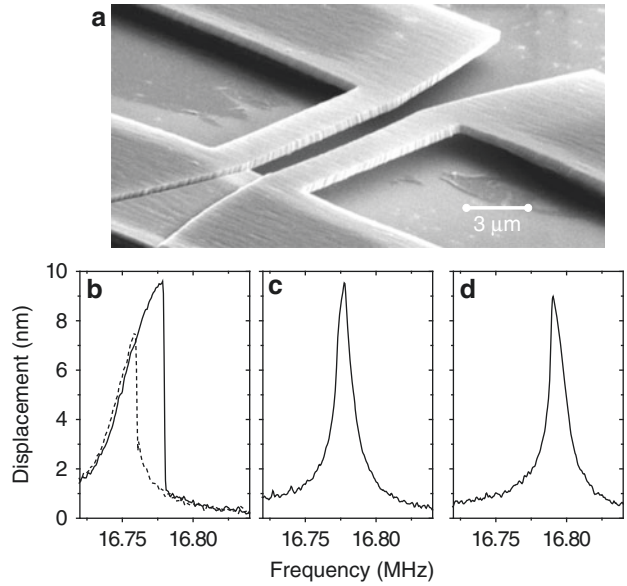
One example of such application involves development of a self-sustained oscillators for time keeping or frequency source applications – a vital component of virtually any application. **In its most basic form, an oscillator consists of a resonator driven on resonance, through feedback, to create a periodic signal sustained by a static energy source**. Since the appearance of quartz-crystal-based oscillators, very few additional conceptual innovations have been introduced: quartz crystal resonators (their frequency-determining elements) operate at the highest possible signal to noise ratio in order to minimize the phase noise – which represents the sideband power spectral density at a given offset frequency, normalized by the oscillator's signal power – an important figure of merit characterizing the performance of an oscillator. The resonator's vibration is kept within *linear* regime; this results in the phase noise that is inversely proportional to the oscillator carrier power. An effort to replace ubiquitous quartz-crystal-based resonator with microfabricated NEMS or MEMS device will potentially lead to a dramatic reduction in size, cost, and power consumption. However, the size miniaturization necessarily leads to a decrease in maximum energy the nanoscale mechanical oscillator can store while remaining in linear harmonic regime. Smaller output power leads to performance degradation. Villanueva et al. [12] showed that these conventional performance limitations could be surpassed if nonlinear dynamics is considered.

The authors analyzed all the contributions to the phase noise in an oscillator based on a *nonlinear* resonator. They discovered the existence of a special region in the parameter space, above the nonlinear threshold, where the dominant contributions to the phase noise are suppressed. As a result the authors demonstrated a phase noise performance beyond the conventional limitations of linear regime. This experiment demonstrated that nonlinear dynamics can be used to improve performance of a practical application, and NEMS-based resonator is a perfect platform to realize that.

Nonlinear Dynamics Experiments: Multiple Devices

A simple increase in complexity of the dynamical system is achieved by considering multiple interacting nonlinear resonators. A relatively simple system of two nonlinear nanomechanical resonators strongly coupled by linear elastic force was theoretically modeled and experimentally studied by the authors of Karabalin et al. [13].

Despite its simplicity this two-resonator system demonstrated inherently rich dynamics, including the tuning of the nonlinearity of one mode by the other, the onset of spontaneous amplitude modulation oscillations, and the first strong evidence for deterministic chaos in nanoscale mechanical systems.



Nonlinear and Parametric NEMS Resonators, Fig. 2 (**a**) An SEM micrograph of two strongly coupled doubly clamped beams, which shared a ledge etched into an insulating layer. Individual electrical connections to the beams were provided. Frequency sweeps of the first mode response for increasing amplitudes of the second mode: the light line in (**b**) is a downward sweep, showing the hysteresis and amplitude jumps. There are no hysteretic jumps for (**c**) and (**d**), so the downward sweeps are not shown (Reprinted with a permission from Phys. Rev. B 79, 165309, 2009. Copyright 2009. American Physical Society)

Two nanomechanical resonators shown in Fig. 2a were linked via a ledge which acted as a strong elastic spring. Two coupled resonators could be reduced to two independent resonant modes, each being a mixture of vibrations from each beam. The authors of [13] have shown that a weak Duffing nonlinearity characteristic of each doubly clamped beam introduced nonlinear coupling terms for the two resonant modes. The multitude of remarkable physical phenomena arising from these simple terms were experimentally investigated and theoretically modeled by the authors.

In the case of a single resonator with a cubic nonlinearity, the peak frequency would effectively shift proportionally to the displacement squared. In two nonlinear coupled resonators, the frequency of one mode changed proportionally to the displacement of the other mode. Such system could therefore be useful as a displacement detector for quantum nondemolition measurements – an important application for nanoelectromechanical systems [14].

The authors further demonstrated that not only resonance frequency but also the nonlinear coefficient of each resonance could be controlled by the vibration state of the other resonance. When the coupling was strong enough, the minimum value of the effective nonlinear coefficient could be negative. In this case the resonance curve tilted to the left – a softening nonlinearity, as opposed to the usual case for hardening nonlinearity typically exhibited by doubly clamped beams, where the peak leaned to the right. It also meant that at some critical value (involving the drive strength and frequency of the second mode), the nonlinear coefficient vanished. Experimental results illustrating this effect are shown in Fig. 2b–d; here the first mode was driven at twice the critical strength. The plots show the shape of the resonance peak for three values of the drive frequency of the second mode. In panel (b), the amplitude A_{II} of the second mode was low, and the first mode spectral response had regular (hardening) nonlinear Duffing shape that leans to the right. For larger A_{II} as in (c), the first mode resonance

peak shape assumed a form close to a Lorentzian – that is, with little apparent nonlinearity. For even larger A_{II} as in (d), the sign of the effective Duffing coefficient became slightly negative, causing the spectral response peak to lean to the left, characteristic of a softening nonlinearity. This quenching of the nonlinearity can be used to enhance dynamic range of nanomechanical devices.

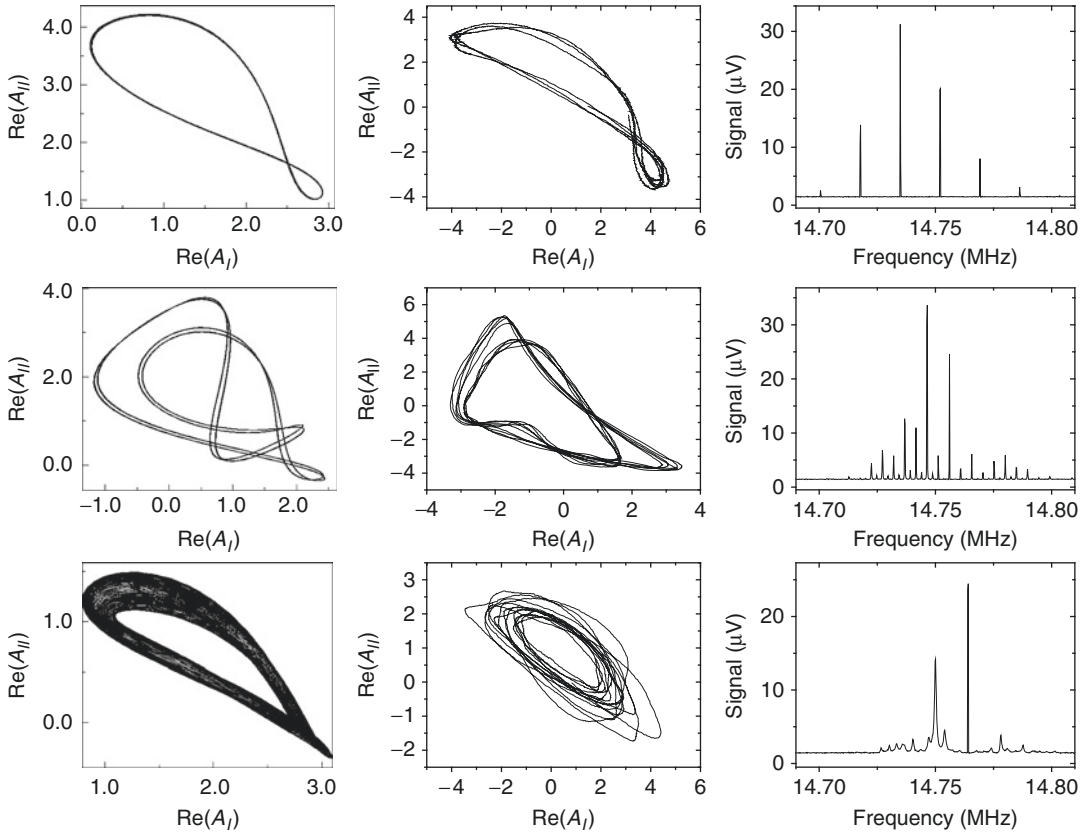
When the drive level of the first mode was increased to an extent comparable to the second mode, a variety of new effects were observed. The dynamical behavior in this case is more complex, as the response of the first mode becomes large enough to cause transitions in the second mode response through the nonlinear coupling. It turned out that for the drive frequencies near the transition points, the fast (about 17 MHz) oscillating response becomes amplitude modulated with a slow frequency of about 10–20 kHz. The theoretical simulations show that the frequency for the amplitude modulation was determined by the linewidth. The dynamics could be intuitively understood as follows: when the first mode was in its lower state and the second mode in its upper state, the first mode switched to the upper state; this caused the second mode to drop to the lower mode. In this condition the second mode became unstable in its lower solution branch and jumped to the upper state, kicking the first mode down to its lower state – thus restarting an oscillatory cycle. As a result the system reached a dynamical equilibrium where the two modes periodically transfer energy between one another.

Examples of simulated and measured amplitude modulation dynamics for three different sets of drive parameters are shown in Fig. 3. The first column shows examples of numerically calculated phase portraits of $\mathrm{Re}A_{\mathrm{II}}$ (real part of the amplitude of second mode) versus $\mathrm{Re}A_{\mathrm{I}}$ (real part of the amplitude of second mode). Phase portraits measured experimentally are shown in the second column. Frequency domain spectrum analysis is shown in the third column. The top row of Fig. 3 shows a relatively simple example where the motion is periodic but the trajectory nonetheless forms a small figure-eight loop. The spectrum of the measured mechanical signal also shows satellite peaks reflecting the anharmonic amplitude modulation.

As the parameters of the system were changed, period doubling or quadrupling in the amplitude modulation was observed. In this case the phase trajectory took two or four revolutions to complete the cycle [15]. An example of period quadrupling is shown in the middle row of Fig. 3. The complicated loop structures were visible in both theoretical and experimental trajectories, and the spectrum revealed amplitude modulation peaks at frequencies that corresponded to double and quadruple periods. Period-doubling transitions are often associated with chaotic dynamics, and indeed chaos in the coupled nanomechanical systems is shown in bottom row of Fig. 3. The broadband components of the spectrum (evident in the shoulders present in amplitude modulation peaks) provided clear evidence for the presence of chaotic dynamics in the experiment. Further illustration was the fact that the phase portrait trajectory did not form a closed loop. The theoretical model showed a similar behavior.

The realization of a large number of resonators performing useful application has long been one of the goals of NEMS research effort. One obvious benefit from employment of multiple NEMS resonators is hypothetical possibility to increase the signal to noise ratio, since the coherently summed signal grows faster than incoherently added noise. Thus the system could achieve not only high sensitivity and small form factor of NEMS, but also stability strong signal level of MEMS and larger mechanical devices. In order to achieve this goal, the resonators have to be synchronized, which turns out to be a challenging obstacle. An important work by Matheny et al. [16] demonstrated that nonlinear dynamics can be used as a tool to synchronize coupled oscillators and thus improve frequency stability of the system.

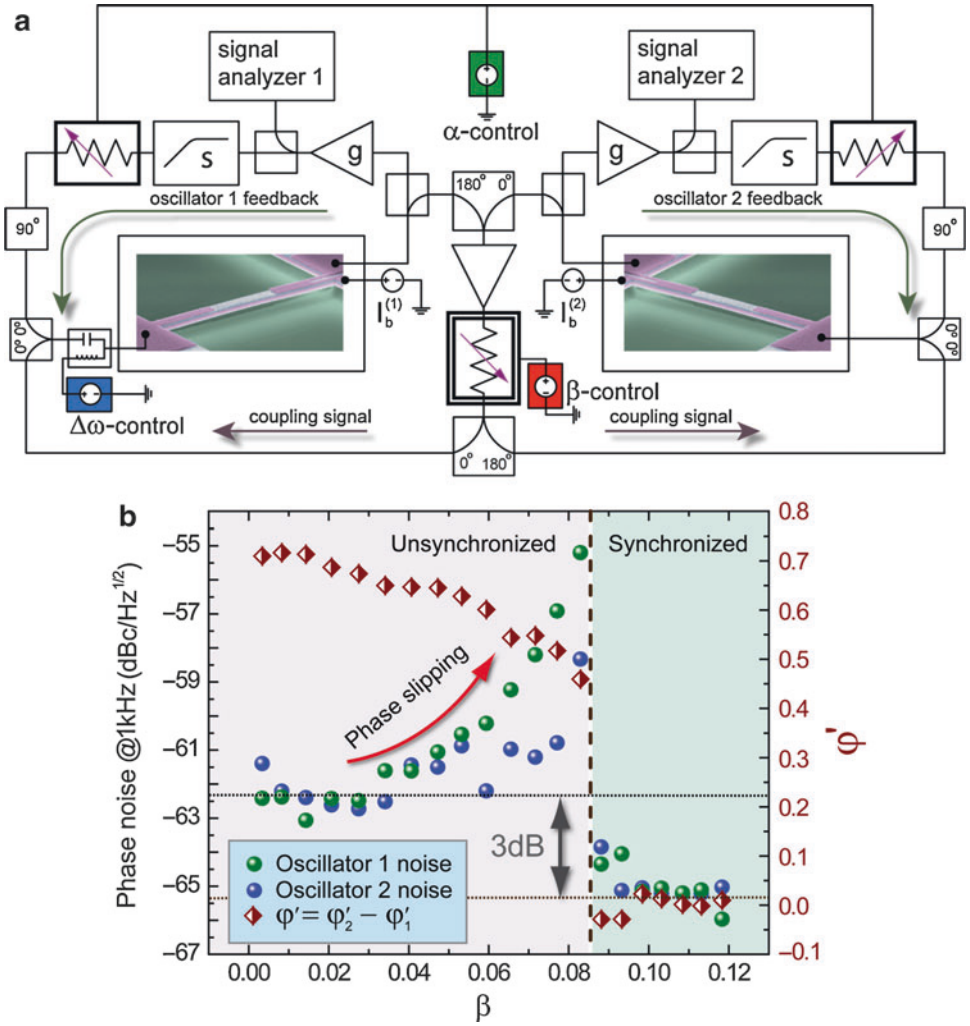
In this work the authors built two self-sustained oscillators out of two mechanically isolated NEMS resonators. The oscillators' natural linear frequencies did not exactly match due to fabrication uncertainties unavoidable at miniscule device dimensions. In this case the synchronization could occur if there existed a frequency



Nonlinear and Parametric NEMS Resonators, Fig. 3 Complex dynamics of two strongly coupled nanomechanical resonators. The three rows correspond to different input parameters (drive frequencies and amplitudes). The first column shows the theoretical calculation of the phase portrait, the second shows its experimental measurement, and the third column shows the corresponding experimental power spectrum of the optical measurement near one of the drive frequencies (Reprinted with a permission from Phys. Rev. B 79 165309, 2009. Copyright 2009. American Physical Society)

tuning mechanism that could be controlled by oscillator motion parameters and interaction. The anharmonicity arising from mechanical nonlinearity provided convenient method able to achieve synchronization under certain conditions [17]. The authors theoretically modeled this system and carefully mapped the parameter space where synchronization occurs. They calculated that stronger coupling strengths were needed for further separation in original frequencies. The authors devised a highly controllable coupling where the electric readout signal from one of the oscillators was directly fed into feedback loop of another oscillator and vice versa. The schematic of their setup is shown in Fig. 4a. The introduction of a cascade of amplifiers and phase shifters enabled the authors to control not only the strength but also the physical nature of interaction.

The synchronization was experimentally observed and carefully studied. A number of important dependencies characteristic of synchronization were measured, thus unequivocally demonstrating a true synchronization in nanomechanical systems for the first time. Thorough experimental work was performed by sweeping control parameters such as initial frequency separation and coupling strength. The experimental results were in remarkable agreement with theoretical predictions.



Nonlinear and Parametric NEMS Resonators, Fig. 4 (**a**) Simplified circuit schematic of the experiment. The NEMS resonators employed were selected to be as identical as possible. Each NEMS resonator (colored SEM micrograph) was embedded in two feedback loops: one was used for creating self-sustained oscillations in each resonator and the other for implementing coupling between the two oscillators. Three control knobs: α for nonlinear coefficient, β for coupling strength, and Δω for frequency offset are shown in the schematic. (**b**) Oscillator phase noise at 1 kHz offset from carrier frequency and oscillator frequency difference (right axis) as coupling is increased. At the value of coupling β = 0.086, the oscillator frequency difference goes to zero, and the phase noise for both oscillators decreases by 3 dB, i.e., corresponding to reducing the phase noise by half (Reprinted with a permission from Phys. Rev. Lett. 112, 014101, 2014. Copyright 2014. American Physical Society)

Finally, the authors explored the behavior of phase noise of the oscillators as it approached synchronization. Figure 4b shows the phase noise of both oscillators as a function of coupling strength. A theoretically predicted drop by 3 dB corresponding to a phase noise reduction by half is evident from the plot. This work showed the potential of nonlinear dynamics at the nanoscale scale. All of the components in the experiment could be realized in CMOS, implying that large-scale networks could be built using the standard semiconductor manufacturing techniques and electronically tested with state-of-the-art signal-processing capabilities. The flexibility of this system permitted high level of coupling control in completely random networks. The demonstrated

phase noise reduction in synchronized oscillators can pave the way toward enabling sensors and frequency sources based on arrays of nanomechanical devices.

Parametric Resonance

Another interesting fundamentally nonlinear effect that can be observed and exploited with nanoelectromechanical systems is parametric resonance. This phenomenon is useful for sensing applications because it provides an opportunity to amplify the signal in mechanical domain prior to applying electrical transduction. This reduces the often difficult challenge of optimally transducing signals from nanoscale mechanical devices into the realm of macroscopic electrical circuits. Parametric resonance is also valuable as a tool to study interesting dynamics in variety of systems ranging from quantum optics to plasma physics. In nano- and microscale mechanical systems, it has attracted significant attention for uses in applications and fundamental research. This section introduces the basic physics of a NEMS parametric resonance system using a specific realization with piezoelectric nanoelectromechanical system and describes an important experiment which explores parametric resonance as a means for building a self-sustained oscillator.

The simplest parametric amplification scheme involves periodic modulation of the spring constant of the system at twice its fundamental resonance frequency. Modulation of the tensile stress of a nanomechanical doubly clamped beam results in tuning of its resonance frequency. Calculating the displacement of the central point of the beam as a single degree-of-freedom harmonic oscillator of effective mass m, spring constant k, and quality factor Q with Duffing nonlinearity, we obtain the following equation of motion:

$$m\,\ddot{x} + \frac{m\omega_0}{Q}\,\dot{x} + \left(k + k_p \cos\left(\omega_p t\right)\right)x + \alpha x^3 = F_0 \cos\left(\omega t + \phi\right), \tag{3}$$

where k_p is a "parametric drive strength," i.e., amplitude of the spring constant modulation. The parametric modulation is performed at twice the resonance frequency, leading to a vibrational amplitude gain given by [8]:

$$G = \frac{A_{\text{pump on}}}{A_{\text{pump off}}} = \left[\frac{\cos^2\left(\phi + \frac{\pi}{4}\right)}{\left(1 + \frac{k_p}{k_t}\right)^2} + \frac{\sin^2\left(\phi + \frac{\pi}{4}\right)}{\left(1 - \frac{k_p}{k_t}\right)^2}\right]^{\frac{1}{2}}, \tag{4}$$

where k_t is the threshold parametric pumping force $2k/Q$. The second term in this equation diverges as k_p approaches k_t; thus significant vibrational amplitude gain can be obtained, provided the appropriate phase relationship between the parametric modulation and driving force.

Piezoelectricity is a well-known phenomenon that directly converts electric fields into mechanical strain. Therefore it presents a convenient method to induce internal stress and thus change effective stiffness. In the experiment by Karabalin et al. [18], a set of doubly clamped beams were fabricated on the same chip. An active piezoelectric layer was sandwiched between the two conducting layers. When a voltage was applied between top and bottom conducting layers, the anisotropic piezoelectric coupling of the material caused longitudinal tensile strain to develop in the intermediate insulating layer. Doubly clamped boundary conditions converted the strain to stress, which either stiffened or softened the beam, thereby changing its resonance frequency.

Due to the remarkable efficiency of the piezoelectric stress tuning mechanism, the realization of parametric resonance could be implemented at very high frequencies with the NEMS devices. When the resonator is parametrically pumped at twice the characteristic frequency, a significant amplitude enhancement was observed consistent with theoretical prediction for parametric gain given in Eq. 4.

When the pumping force exceeded the threshold, the amount of energy delivered to the system became larger than the amount lost to dissipative phenomena. In this situation, in the absence of any

coherent driving signal, the resonator executed spontaneous oscillations close to its characteristic resonance frequency. The behavior of such dynamical system was described in detail in [8]. This interesting physical phenomenon was used to build a novel self-sustained oscillator paradigm in the work by Villanueva et al. [19]. The authors described and demonstrated an alternative to the canonical oscillator topology, namely, a new architecture based on nonresonant *parametric* feedback. A feedback loop in this work possessed a quadratic transfer function to apply parametric excitation at twice the resonant frequency (as opposed to "direct drive" at its resonant frequency).

The authors discovered many advantages of this generalized *parametric feedback oscillator* (PFO) technique, both from fundamental and technological points of view. Among them were (a) circumventing the need to satisfy the conventional "Barkhausen" oscillation conditions, thus enabling a wider variety of resonators to be used for oscillator circuits; (b) obtaining an unprecedented level of control of the resonator's nonlinear characteristics, enabling access to higher amplitudes of operation; (c) realizing a wider range of frequency tunability; and (d) obtaining substantial improvement in frequency stability of the oscillator, compared with that of conventional, direct-drive implementations.

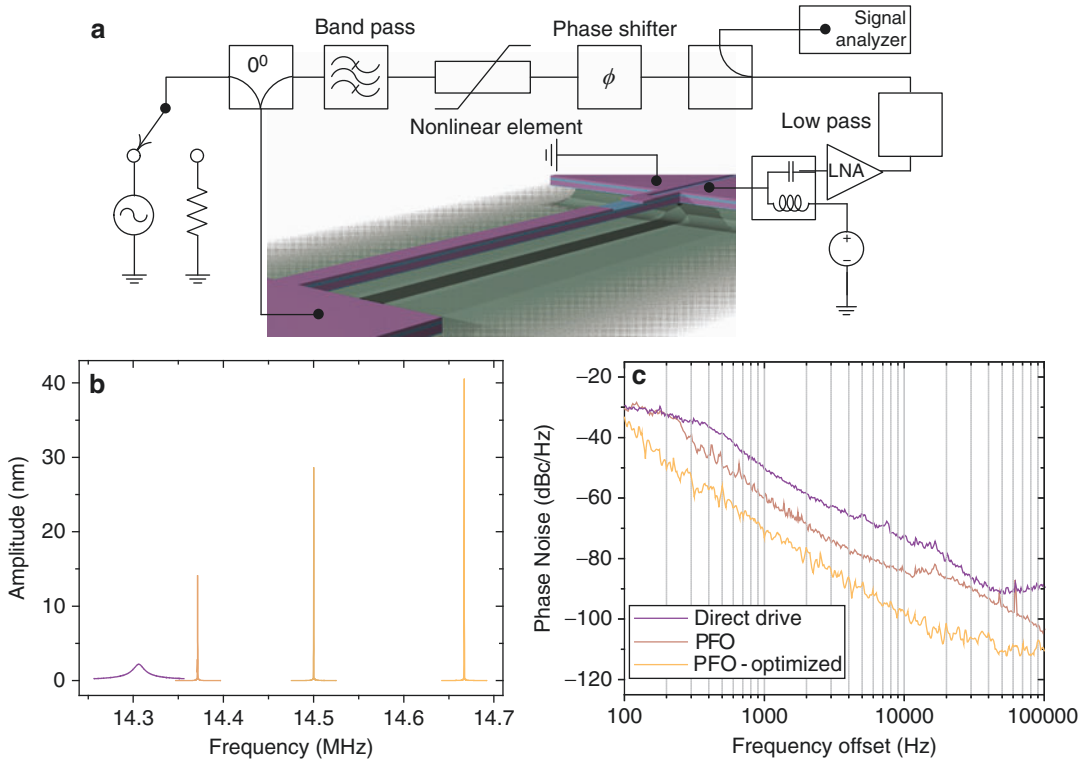
The PFO concept was demonstrated using piezoelectric NEMS, where active piezoelectric layer (AlN) was sandwiched between two conducting layers (molybdenum). The out-of-plane resonator motion was detected using the time-varying, strain-induced resistance changes in a piezometallic (Mo) loop patterned at one end of the beam. Actuation was obtained by applying an AC voltage to an electrode that covers most of the beam's length; this induces longitudinal strain by means of the inverse piezoelectric effect. This time-varying strain could be used to actuate the beam either directly or parametrically.

The schematic of the setup is shown in Fig. 5a. The motional signal was first amplified and filtered to suppress high-frequency noise and higher harmonics ($>f$). Subsequently, the signal was delayed by a voltage-controlled phase shifter (ϕ) and then passed through a nonlinear element optimized to generate a *2f* signal with amplitude proportional to the square of the resonator motion. This frequency-doubled waveform was properly filtered and applied to a resonator actuation port thus inducing a resonator response at f – that is, parametric oscillation – provided the *2f* drive level exceeds the parametric threshold. Unlike traditional oscillator behavior, zero amplitude is a stable solution for our feedback system. This makes it necessary to initiate resonator motion by an external "start-up" source. However, after oscillations commence, the start-up drive can be removed and stable oscillations at f will persist, sustained only by the parametric feedback at *2f*.

Given that the frequency of the feedback (*2f*) and output signal (f) were well separated, cross talk was minimal for the operation of the parametric feedback oscillator; this eliminated a traditional obstacle for oscillators based on small-scale mechanical devices: for miniaturized electromechanical resonators, the output electrical signal was usually dominated by the feedthrough of the actuation voltage. In those conditions it was very challenging to attain an oscillation that used the mechanical resonator as the frequency-determining element.

When the feedback gain was sufficient to parametrically excite the resonator, then aforementioned parametric oscillation criterion was satisfied. The authors observed the inception of oscillation and carefully characterized it by capturing the power spectrum and comparing this to the open-loop resonator frequency response. Figure 5b shows the spectral response of the oscillator for three different values of phase delay, each incremented by 10°. Figure 5b also shows, for comparison, the open-loop resonator response at typical drive in linear regime. The extended tuning range was a direct consequence of parametric feedback, consistent with the theoretical model developed by the authors. This wide frequency range could be useful for applications requiring voltage-controlled oscillators and for the potential synchronization of coupled oscillators described in earlier section.

To characterize and quantify the performance of the parametric feedback oscillator, the authors



Nonlinear and Parametric NEMS Resonators, Fig. 5 (**a**) Schematic diagram of the generalized feedback system. The signal from the resonator was amplified and filtered at high frequencies to eliminate higher harmonics and noise. After an externally controlled (ϕ) phase delay was applied, the signal was passed through a nonlinear element followed by a bandpass filter to ultimately generate a signal at *2f*. This signal was applied to the actuation port of the beam through a power combiner that allowed simultaneous feedback and a direct drive with an external source, which was necessary to initiate the oscillations. Once the self-sustaining state had been reached, the source could be disconnected and the motion persisted. (**b**) Comparison between the open-loop response (*purple*) and the PFO power spectrum (*orange*). Three PFO traces are shown, each for a different value of ϕ (separated by 10°). (**c**) Phase noise measurements of the PFO in both a standard case and at the optimum phase value, showing a reduction of the noise. For comparison, the phase noise of a standard direct-drive oscillator is shown for the same oscillator energy, indicating higher phase noise than PFO over most of the frequency range (Reprinted with a permission from Nano Lett., 2011, 11 (11), pp 5054–5059. Copyright 2011. American Chemical Society)

measured its phase noise. A baseline comparison using separately constructed conventional feedback loop oscillator with direct drive at *f*, using the same active components, was provided. Frequency stability comparisons between this direct-drive oscillator and the prototype PFO, for operation at identical energies, were shown in Fig. 5c. The authors observed that the frequency stability of the PFO was significantly improved compared to that of the traditional oscillator topology. This provided direct evidence of suppression of the effects of phase noise in the feedback electronics. An additional enhancement was observed within the special region close to the mode coupling, where the phase noise was reduced by an additional 15 dB. This suppression was consistent with the theoretical model developed by the authors.

This section clearly demonstrates that parametric resonance is a unique dynamical system that could provide not only interesting insights into the deeper understanding in the field of nonlinear dynamics but also used to enhance performance of long-sought applications of nanoscale electromechanical systems.

Conclusions

A rapidly growing field of NEMS resonator has already shown remarkable progress in developing methodologies in the realm of sensing. NEMS resonators are increasingly employed for a variety of sensing applications including detection of mass, gas, biomolecules, and force – and they have demonstrated unprecedented resolution. The most responsive of these devices are based on frequency-shift sensing, where a stimulus to be measured induces a change in the resonant frequency of the NEMS. The development of NEMS-based self-sustained oscillators is not only promising for frequency control applications but also for miniaturizing the footprint of the whole sensor system. A stable NEMS oscillator element, co-integrated with a few active transistors, would represent the most compact sensing "pixel" achievable.

Today, an appreciable gap remains between many research demonstrations obtained in the laboratory and their translation to realization of reliable commercial applications – however, this gulf is closing quickly. Nonlinear processes are just beginning to be explored to assist in this process. There are enormous potential advantages that can be realized by nonlinear dynamics in nanoelectromechanical systems; among these are intrinsically noiseless amplification, suppression of noise mechanisms, synchronization of arrays of oscillators, and creation of novel types of nonlinear oscillators.

Meanwhile, NEMS resonators are increasingly providing an excellent testing bed for fundamental research that involves the detailed study of nonlinear dynamics; such efforts have potential applications to, and benefits for, many fields of endeavor. These nonlinear dynamical systems provide an opportunity to observe many interesting physical phenomena – among which are multiple, coexisting mechanical states, parametric resonance, noise-induced transition, chaos, and synchronization. It is anticipated that this research could have broad impact on fields as diverse as neuroscience, network theory, and frequency control.

References

1. Roukes, M.L.: Nanoelectromechanical systems face the future. Phys. World **14**(2), 25–31 (2001)
2. Yang, Y., Ng, E.J., Hong, V.A., Ahn, C.H., Chen, Y., Ahadi, E., Dykman, M., Kenny, T.W.: Measurement of the nonlinear elasticity of doped bulk-mode mems resonators. Solid-State Sensors, Actuators, and Microsystems Workshop, pp. 285–288. Hilton Head (2014)
3. Hiebert, W.K., Vick, D., Sauer, V., Freeman, M.R.: Optical interferometric displacement calibration and thermomechanical noise detection in bulk focused ion beam-fabricated nanoelectromechanical systems. J. Micromech. Microeng. **20**, 115038 (2010)
4. Bargatin, I., Kozinsky, I., Roukes, M.L.: Efficient electrothermal actuation of multiple modes of high-frequency nanoelectromechanical resonators. Appl. Phys. Lett. **90**, 093116 (2007)
5. Kozinsky, I., Postma, H.W.C., Bargatin, I., Roukes, M. L.: Tuning nonlinearity, dynamic range, and frequency of nanomechanical resonators. Appl. Phys. Lett. **88**, 253101 (2006)
6. Villanueva, L.G., Karabalin, R.B., Matheny, M.H., Chi, D., Sader, J.E., Roukes, M.L.: Nonlinearity in nanomechanical cantilevers. Phys. Rev. B **87**, 024304 (2013)
7. Matheny, M., Villanueva, L.G., Karabalin, R.B., Sader, J.E., Roukes, M.L.: Nonlinear mode-coupling in nanomechanical systems. Nano Lett. **13**, 1622–1626 (2013)
8. Lifshitz, R., Cross, M.C.: Nonlinear Dynamics of Nanomechanical and Micromechanical Resonators, in Reviews of Nonlinear Dynamics and Complexity. Wiley-VCH, Weinheim (2008)
9. Kozinsky, I., Postma, H.W.C., Kogan, O., Husain, A., Roukes, M.L.: Basins of attraction of a nonlinear nanomechanical resonator. Phys. Rev. Lett. **99**, 207201 (2007)
10. Aldridge, J.S., Cleland, A.N.: Noise-enabled precision measurements of a duffing nanomechanical resonator. Phys. Rev. Lett. **94**, 156403 (2005)
11. Unterreithmeier, Q.P., Faust, T., Kotthaus, J.P.: Nonlinear switching dynamics in a nanomechanical resonator. Phys. Rev. B **81**, 241405 (2010)
12. Villanueva, L.G., Kenig, E., Karabalin, R.B., Matheny, M.H., Lifshitz, R., Cross, M.C., Roukes, M.L.: Surpassing fundamental limits of oscillators using nonlinear resonators. Phys. Rev. Lett. **110**, 177208 (2013)
13. Karabalin, R.B., Cross, M.C., Roukes, M.L.: Nonlinear dynamics and chaos in two coupled nanomechanical resonators. Phys. Rev. B. **79**(16), 165309 (2009)
14. O'Connel, A.D., Hofheinz, M., Ansmann, M., Bialczak, R.C., Lenander, M., Lucero, E., Neeley, M., Sank, D., Wang, H., Weides, M., Wenner, J., Martinis, J.M., Cleland, A.M.: Quantum ground state and single-phonon control of a mechanical resonator. Nature **464**, 697–703 (2010)

15. Strogatz, S.H.: Nonlinear Dynamics and Chaos. Addison Wesley, New York (1994)
16. Matheny, M.H., Grau, M., Villanueva, L.G., Karabalin, R.B., Cross, M.C., Roukes, M.L.: Phase synchronization of two anharmonic nanomechanical oscillators. Phys. Rev. Lett. **112**, 014101 (2014)
17. Cross, M.C., Rogers, J.L., Lifshitz, R., Zumdieck, A.: Synchronization by reactive coupling and nonlinear frequency pulling. Phys. Rev. E **73**, 036205 (2006)
18. Karabalin, R.B., Masmanidis, S.C., Roukes, M.L.: Parametric amplification in high frequency piezoelectric nanomechanical systems. Appl. Phys. Lett. **97**, 183101 (2010)
19. Villanueva, L.G., Karabalin, R.B., Matheny, M.H., Kenig, E., Cross, M.C., Roukes, M.L.: A nanoscale parametric feedback oscillator. Nano Lett. **11**, 5054–5059 (2011)

Nonlinear Electrokinetic Transport

▶ Concentration Polarization at Micro-/Nanofluidic Interfaces

Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures

V. Sai Muthukumar[1], Ramakrishna Podila[2], Benoy Anand[1], S. Siva Sankara Sai[1], K. Venkataramaniah[1], Reji Philip[3] and Apparao M. Rao[2,4]
[1]Department of Physics, Sri Sathya Sai Institute of Higher Learning, Vidyagiri, Puttaparthi, Andhra Pradesh, India
[2]Department of Physics and Astronomy, Clemson University, Clemson, SC, USA
[3]Light and Matter Physics Group, Raman Research Institute, Bangalore, India
[4]Center for Optical Materials Science and Engineering Technologies, Clemson University, Clemson, SC, USA

Definition

Nonlinear optics (NLO) studies the behavior of light in *nonlinear media*, whose dielectric polarization (***P***) responds nonlinearly to the electric field (***E***) of the light. In general, such nonlinear behavior is observed under highly intense light beams. Several fundamental properties of materials such as nonlinear absorption, nonlinear refraction, harmonic generation, etc., can be probed using NLO spectroscopy.

The field of nonlinear optics (NLO) emerged into prominence more than three decades ago with the development of the laser. These milestone discoveries have not only generated much interest in laser science but also set the stage for future work in nonlinear optics. In recent years, the discovery of various nanoscale materials such as fullerenes, carbon nanotubes, and metallic and other semiconducting nanostructures has allowed nonlinear spectroscopy to progress towards technological applications, and the focus has been on the discovery of materials with higher photonic functionalities and better performance [1, 2]. Knowledge of electronic transitions is essential in understanding the linear and nonlinear optical properties of nanomaterials. Many spectroscopic techniques such as optical absorption, photoluminescence, photoacoustics, and x-ray photoelectron spectroscopy are often employed for the investigation of energy levels in nanostructures. More significantly, estimation of the nonlinear interaction of light with nanostructures helps one to probe the higher electronic levels, which are not easily accessible otherwise.

In nanostructures, quantum confinement effects and surface plasmon resonance (SPR), in tandem with other crucial factors like morphology, particle size, and chemical environment, can lead to interesting changes in nonlinear properties with many promising applications such as ultrafast optical switching, optical limiting, passive mode locking, frequency up-conversion lasing, optical data storage, multiphoton-based laser microscopy, etc. Interestingly, various nanomaterials exhibit distinct intensity-dependent nonlinear responses for laser excitation that depends on parameters such as laser wavelength, pulse duration, and repetition rate [2]. The physical basis of observed nonlinear optical responses in nanostructures is briefly elucidated in the

following sections along with an outline for various experimental techniques used for probing their photophysical properties with special reference to the optical limiting application. Optical limiters are materials which are transparent at lower incident fluences, but attenuate light considerably at higher fluences; thus protecting eye and photosensors from damage due to intense optical sources. Various nanostructures are promising as efficient and broadband optical limiters when compared to conventional materials like organic chromophores and dyes [1, 2]

Nonlinear Optical Processes

In general, the response of a material medium (which is lossless and dispersionless) to input light can be described in terms of the electric polarization $\boldsymbol{P}$ [1], which is usually expanded in a Taylor series in the electric field amplitude $\boldsymbol{E}$ as

$$\boldsymbol{P}=\chi^{(1)}\boldsymbol{E}+\chi^{(2)}\boldsymbol{EE}+\chi^{(3)}\boldsymbol{EEE} \tag{1}$$

where $\boldsymbol{\chi^{(n)}}$ are the susceptibility tensors of order n. Far away from a material resonance, the susceptibilities are predominantly real and the material response is instantaneous, but as resonances are approached, the susceptibilities will become complex, with relatively slower responses. The first term, $\boldsymbol{\chi^{(1)}}$, is responsible for linear absorption and refraction, while the remaining terms are associated with different light-induced nonlinear effects. Second-order nonlinear effects due to $\boldsymbol{\chi^{(2)}}$ are present only in noncentrosymmetric materials, and these include second harmonic generation and other three-wave mixing processes, optical rectification, and the electro-optic Pockels effect. Nonlinear optical effects involving the third-order susceptibility $\boldsymbol{\chi^{(3)}}$ include phenomena such as third harmonic generation and four-wave mixing processes, the electro-optic Kerr effect, two-photon absorption, and Raman amplification.

For a nonlinear optical medium, the refractive index and absorption coefficient will be functions of the incident light intensity. For an isotropic centrosymmetric material (where the lowest order nonlinearity is cubic), the intensity-dependent nonlinear refractive index is given by [1–4]

$$n = n_0 + n_2 I \tag{2}$$

and the intensity-dependent nonlinear absorption coefficient is given by

$$\alpha = \alpha_0 + \beta I \tag{3}$$

where n_2 is the nonlinear refractive index coefficient, and β is the nonlinear absorption coefficient. n_0 and α_0 are the linear refractive index and linear absorption coefficient, respectively. I is the intensity of the incident light which is proportional to $\boldsymbol{E}^2$. The third-order nonlinear susceptibility can be very generally written as

$$\chi^{(3)} = \chi^{(3)}_{(R)} + \chi^{(3)}_{I} \tag{4}$$

where the real part $\chi^{(3)}_{(R)}$ accounts for effects that are transient and consume no energy (like the optical Kerr effect, molecular orientation, electrostriction, etc.). The imaginary part $\chi^{(3)}_{I}$ accounts for processes that involve losses such as two-photon absorption. Thus, $\chi^{(3)}$ (both the real and imaginary parts) serves as a figure of merit for investigating potential novel materials for photonic applications.

It may be noted that ultrafast laser pulse excitation (typically in the femtoseconds regime) provides a way to measure pure-electronic nonlinearities exclusively. To observe longer timescale processes such as excited state absorption and thermal lensing, longer laser pulses (typically in the nanosecond regime) are used.

Nonlinear Refraction (NLR)

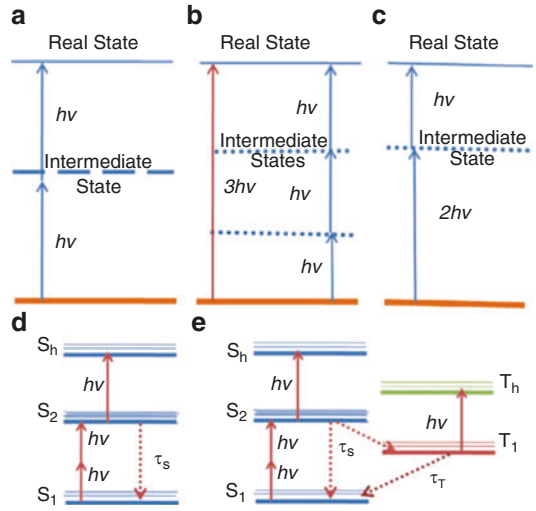
The dominant physical mechanism driving nonlinear refraction in a given medium can vary depending on the material, optical wavelength, irradiance, timescale of the light-matter interaction, and/or geometry. In addition to bound electronic nonlinearities, which are instantaneous, effects like intramolecular motion, molecular reorientation, electrostriction, induced population

change, and photothermal effects may become responsible for nonlinear refractive index in a medium. Depending on the sign (positive or negative) and magnitude of the value of "n_2" in Eq. 2, a laser beam passing through the nonlinear medium may be self-focused or defocused. Such focusing (or defocusing) arising from the real part of the third-order nonlinear susceptibility ($\chi^{(3)}_{(R)}$) is known as the optical Kerr effect (Kerr lens effect). The instantaneous nature of Kerr lensing is used in several applications including optical switching and laser mode-locking, especially in the ultrafast regime. At very high intensities, the optical Kerr effect may tend to saturate. For relatively shorter optical pulses with very high intensity, electronic nonlinearities may, on the other hand, lead to self-steepening and supercontinuum generation effects.

Nonlinear refraction can also occur due to thermal effects (usually referred to as thermal lensing), which typically build up with a relatively long response time of a few milliseconds. Thermal nonlinearity can be dominant for continuous-wave (CW) laser excitation and significant for nanosecond laser pulse excitation. The variation of the refractive index with temperature can be described mathematically as $\tilde{n} = n_o + (dn/dT)\tilde{T}_0$, where n_o is the linear refractive index, $\tilde{T}_0$ is the laser-induced change in temperature, and dn/dT, the temperature-dependent refractive index. The thermal nonlinear index of refraction can be determined using a closed aperture Z-scan measurement, employing either the thermal lens model (TLM) or Sheik-Bahae's formalism [3] for analysis.

Nonlinear Absorption

Nonlinear absorption refers to the change in the transmittance of a material as a function of input light fluence (given in units of J/m^2) [1]. As given above, the nonlinear absorption coefficient of a third-order nonlinear medium can be written as $\alpha = \alpha_0 + \beta I$, where β is the nonlinear coefficient of absorption. Far away from a material resonance, β usually reduces to the two-photon absorption coefficient. However, near a material resonance, two-step phenomena like excited state absorption (ESA) and free carrier absorption (FCA) may also contribute to nonlinear absorption. Such phenomena are collectively referred to as reverse saturable absorption (RSA). An overview of these phenomena is given below.



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 1 Various nonlinear absorption schemes involving 2PA, 3PA and ESA processes. (**a**) depicts an ESA process while (**b**) shows 3PA and (**c**) shows 2PA. The dotted lines represent virtual state and *solid lines* represent real state. (**d**) and (**e**) show three- and five- level ESA models respectively, which have been extensively used in literature (see Ref. [1] for details). (**a**) 2PA virtual intermediate state (*dashed line*). (**b**) 3PA virtual intermediate state (*dotted line*). (**c**) ESA real intermediate state (*dotted line*)

Multiphoton Absorption

Many semiconducting and metallic nanomaterials exhibit multiphoton absorption (MPA). In this discussion, MPA is limited to two- and three-photon absorption (2PA and 3PA, respectively) processes. 2PA involves the simultaneous absorption of two photons to promote an electron from an initial state to a final state through a virtual intermediate level (Fig. 1). 3PA has a lower transition probability and involves three photons and two intermediate states. Since the intermediate level is virtual, energy need be conserved only in the final state. In contrast to reverse saturable absorption, MPA is an instantaneous process occurring in femtosecond timescales. The mechanism of 2PA can be viewed as analogous to a

three-level RSA process where the lifetime of the intermediate state approaches zero and the ground state absorption is extremely low (high linear transparency). Two- and three-photon-absorption-based nanomaterials can be used for applications like frequency conversion lasing, optical limiting, pulse stabilization, and reshaping [1].

Excited State Absorption

In systems like C_{60}, the excited state absorption (ESA) cross section is greater than the ground state absorption cross section, which can lead to an enhanced absorption at higher input intensities. Typically, the ESA process is much stronger than 2PA, and to indicate the collective contribution of 2PA and ESA, the term "effective 2PA" is sometimes used in literature. When excited with nanosecond laser pulses, which are relatively long compared to typical excited state lifetimes of nonfluorescent materials, ESA will contribute significantly to nonlinear absorption. For instance, the effective 2PA cross section of a medium measured under nanosecond (ns) pulse excitation has been found to be about two orders of magnitude larger than the intrinsic 2PA cross section measured under femtosecond (fs) excitation [5]. In the case of intrinsic 2PA, the nonlinear absorption coefficient (β) is independent of the input intensity, but in the case of effective 2PA, the nonlinear absorption coefficient (β_{eff}) varies with input intensity. Relevant examples are discussed in section "Experimental Methods". Another interesting case is that of TPA-induced ESA, where a strong 2PA process considerably increases molecular populations in an excited state, which then absorbs another photon by an ESA process.

Free Carrier Absorption

In semiconducting materials, charge carriers are generated by one-photon as well as two-photon excitations. These electrons or holes can be excited to still higher or lower states in the conduction or valence bands by absorbing additional photons. Such carrier-assisted nonlinear absorption processes are known as "free carrier absorption," which is analogous to excited state absorption in molecular systems [2]. One difference is that free carrier absorption processes are usually independent of the input laser pulse width (assuming shorter timescales for carrier diffusion and recombination). In addition, they are insensitive to particle size and geometry and can be found in both solids and nanoparticle dispersions covering a broad range of temporal and wavelength ranges.

Induced Thermal Scattering

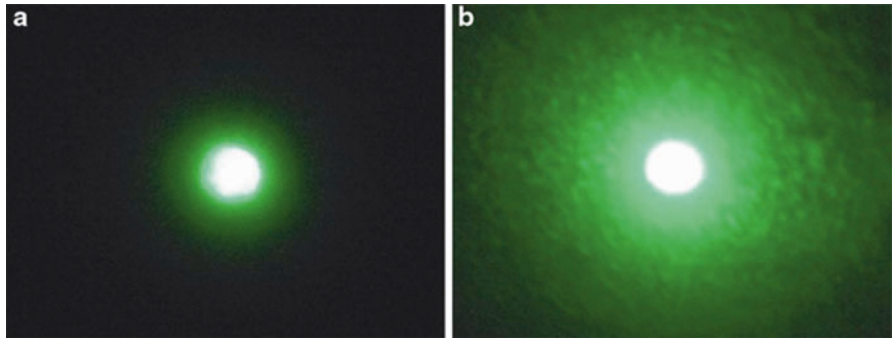
An important mechanism by which laser light transmitted through a medium suffers losses, particularly in the nanosecond excitation regime, is induced thermal scattering (ITS). Processes including the formation of microplasma and solvent microbubbles, and thermally induced local refractive index gradients lead to ITS [4]. Since the formation of thermal scattering centers takes several nanoseconds, these phenomena generally do not affect the transmission of ultrafast laser pulses through the medium. Mechanistically, ITS leads to a spatial redistribution of the incident beam (as shown in Fig. 2), resulting in spatial intensity modulation, which can be efficiently used for optical limiting if appropriate geometries are employed [2, 4].

Experimental Methods

In order to evaluate the contribution of various intensity-dependent nonlinear processes in any material, different methods such as second harmonic generation, four-wave mixing, ellipse rotation, nonlinear interferometry, and Z-scan are often employed [1, 2]. In recent years, Z-scan and four-wave mixing methods have been extensively adopted and applied by various research groups across the world to evaluate the nonlinear optical properties of various bulk and nanostructured materials.

Four-Wave Mixing

Four-wave mixing (FWM) is a phase-sensitive nonlinear effect arising from third-order optical nonlinearity. The concept of three optical fields interacting to produce a fourth optical field is central to the description of all four-wave mixing processes. Along with Z-scan, this is one of the



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 2 Panel (**a**) shows the transverse profile of the incident laser beam, while panel (**b**) shows the beam profile of the scattered beam after traversing through the sample [6]

most widely used techniques to measure the third-order susceptibility $\chi^{(3)}$ of materials [2, 3]. In a dielectric medium, when two light beams are present, each beam independently polarizes the dielectric. The interference between polarized fields then results in harmonics at the sum and difference frequencies. When a third field is incident on the dielectric, it too drives the polarization and will beat with both the other input fields as well as the sum and difference frequencies. This beating with the sum and difference frequencies gives rise to the fourth field in four-wave mixing.

The FWM process is essentially of two types: nondegenerate and degenerate (DFWM). In the former, the optical excitation sources are of different wavelengths, while in the latter, their wavelengths are identical. The strength of the fourth beam is dependent on a coupling constant that is proportional to the effective $\chi^{(3)}$, and hence, measurement of the observed signal will yield information about the $\chi^{(3)}$ tensor components of the medium. Furthermore, DFWM can be employed in the backward (phase conjugate) and BOXCARS configurations, with the choice being made through consideration of the experimental conditions and requirements. In the backward geometry, two waves travel in the backward direction and two travel in the forward direction, whereas in the BOXCARS geometry, all the four waves travel in the forward direction. Using different polarizations for the three beams, all the components of the $\chi^{(3)}$ tensor can be measured. In terms of the intensity-dependent refractive index, the DFWM process is considered to be the interference of two input beams leading to a spatial periodicity resulting in the formation of a grating from which the third beam scatters, generating the fourth wave.

FWM can be used to probe either one-photon resonances or two-photon resonances in a material by measuring the resonant enhancement as one or more of the frequencies are tuned. By tuning the frequencies to multiple resonances in the material, excited state cross sections, lifetimes, and line widths can be measured. One of the main disadvantages of FWM is its sensitivity to alignment, requiring all three coherent beams to simultaneously overlap within the sample in space and time.

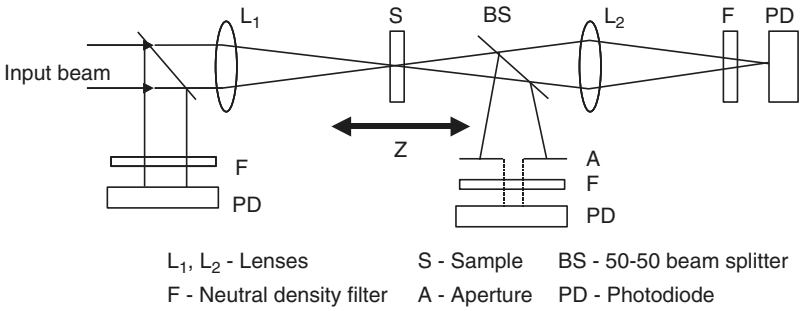
Z-Scan

The Z-scan technique is a sensitive and simple method employed to measure nonlinear absorption and nonlinear refraction. It has been used extensively to study the optical nonlinearity of various materials like semiconductors, nanocrystals, semiconductor-doped glasses, liquid crystals, organic materials, and biomaterials. Introduced by Sheik-Bahae et al. [3] in the early 1990s, this widely used technique has many advantages over other nonlinear spectroscopic methods. In a typical Z-scan, the light-induced change in transmittance (ΔT) of a medium due to optical nonlinearity is measured as a function of input light energy density (fluence) or intensity. A continuous variation of the input fluence is

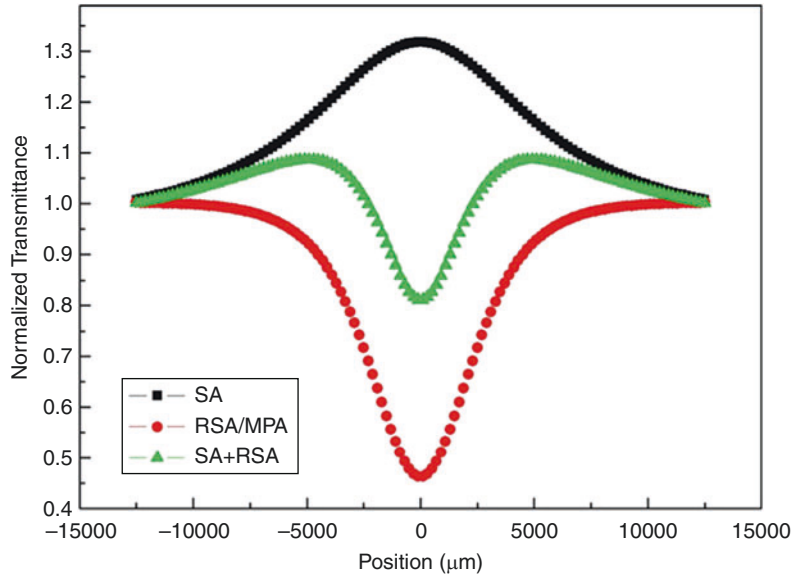
achieved by translating the sample under study through the focal region of a focused laser beam (the sample position is taken as z with $z = 0$ being the focal point for the incident laser beam, hence the name "*Z*-scan"). Consequent increase and decrease of the intensity incident at the sample position lead to wave front distortions (created by nonlinear optical effects in the sample). There are two types of Z-scan techniques, namely, the "closed aperture" Z-scan and the "open aperture" Z-scan.

In the closed aperture Z-scan, which is used for studying nonlinear refraction (NLR), the transmitted beam is passed through an aperture placed in the far field and then measured by a detector, for different values of sample position z. When the medium is far before the focal plane, no self-lensing occurs. As the medium approaches the focal plane, the high intensity begins to induce a lensing effect in the medium. For a negative nonlinearity, this lens tends to collimate the beam, thereby increasing the transmittance through the aperture. At the focal plane, even though the intensity is highest, the influence of the induced lens is minimized, resulting in a transmittance comparable to the linear transmittance. This is similar to placing a thin lens at the focus of a beam; this results in a minimal effect on the far field beam pattern. As the sample is translated beyond the focal plane, the negative lens tends to increase the beam divergence, resulting in a decrease in the aperture transmittance. As the medium is translated still farther from focus, the intensity again becomes weak enough that the induced lensing is negligible. This sequence results in a change in transmittance with a characteristic peak followed by a valley, which is symmetric about the focal plane ($z = 0$). For a positive nonlinearity, the pattern consists of a valley followed by a peak. Thus, a simple Z-scan experiment can give both the sign and magnitude of the optical nonlinearity in a material. In both these cases, a purely refractive nonlinearity is considered, assuming that no absorptive nonlinearities are present. The sensitivity to nonlinear refraction results from the vital role played by aperture A (Fig. 3). If aperture A is removed, then the *Z*-scan is sensitive only to nonlinear absorption (NLA), and this configuration is commonly referred to as the open aperture Z-scan. The absorptive nonlinearity will be a maximum at the focal plane, where the intensity is highest. Thus, the open aperture scheme is employed to characterize materials that exhibit NLA, such as excited state absorption (ESA), two-photon absorption (TPA), saturable absorption (SA), etc.

Even though closed aperture Z-scan is used to study only NLR, in practice, the closed aperture Z-scan curve may contain a contribution from NLA also, depending on the sample. To obtain the NLR coefficient in such a case, it is sufficient to normalize the closed aperture data using the open aperture data before analysis. Although the open aperture Z-scan technique is adequate to detect the presence of nonlinear absorptive effects, it cannot singularly identify the nature of an absorptive nonlinearity. For example, at the outset, it cannot differentiate a 2PA signature from a two-step excitation signature. Figure 4 shows typical Z-scan curves simulated for the MPA, SA, and RSA processes. The Z-scan curves for SA (peak) and RSA (valley) have shapes, which are complementary to each other. It should be noted that the Z-scan curves for RSA and MPA



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 3 A conventional Z-scan layout for simultaneous open aperture and closed aperture measurement



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures,
Fig. 4 Simulated Z-scan curves for the saturable absorption (*SA*), reverse saturable absorption (*RSA*), and multiphoton absorption (*MPA*) processes

look very similar to each other. To confirm the exact nature of the process, it is often necessary to know the linear absorption properties of the sample from a linear absorption spectrum taken using a spectrophotometer. Z-scans performed using ultrafast laser pulses (where RSA effects are a minimum) will also be instructive in this regard.

A recent variation in the Z-scan technique is the white-light continuum Z-scan, where a supercontinuum is used as the light source for doing the Z-scan experiment. The "supercontinuum" is an intense, ultrafast, broadband light pulse, with wavelengths spanning the UV to the near IR region, generated when an ultrafast laser pulse is passed through certain transparent NLO media like heavy water. Conversion of the spectrally narrow pump pulse into the resultant broad spectrum involves the interplay of self-phase modulation and self-focusing of the pulse, due to an intensity-dependent refractive index of the medium. Since the typical supercontinuum spectrum covers the whole visible range, it appears white, and hence it is also known as a "white-light continuum." In the conventional Z-scan technique, nonlinearity can be measured only for one wavelength in one measurement. To determine the spectral dispersion of the nonlinearity, a tunable laser source has to be used, and separate Z-scans need to be done at all the required wavelengths. Obviously, this approach is time consuming and tedious. The idea of using a strong white-light continuum (WLC) as the light source to measure the dispersion of the nonlinearity in a single Z-scan experiment was therefore proposed by Van Stryland's group [7]. Their "WLC Z-scan" allows for rapid, broadband characterization of degenerate NLA and NLR.

Nonlinear Optical Properties of Nanomaterials

Nonlinear optical measurements in low-dimensional materials continue to draw considerable attention since the nonlinear behavior of the complex refractive indices lead to unique optical phenomena [1–3, 7]. Furthermore, the quantum confined nature of the charge carriers, surface effects, and sharp density of states in low-dimensional materials such as nanoparticles, nanowires, nanotubes, and nanorods introduces various new fundamental physical phenomena that can influence the NLO properties. For example, surface effects can lead to ferromagnetic behavior in nanostructured ZnO whereas bulk ZnO is diamagnetic [8], and a semimetal to semiconductor transition occurs in the low dimensional

Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Table 1 Different nonlinear optical processes observed in different nanostructured materials (see text for details)

Type of nanostructures	Cause for NLO properties	Technique used for probing	Reference
C60 Dispersions	NLA (2PA)	Pump-probe techniques	Tutt et al. [4]
	NLR, RSA, NLS, thermal NLO	Z-scan	
C60 Films	NLR, NLA(SA)	Z-scan	
SWNT films	NLR, NLA (FCA)	Z-scan, pump-probe techniques	Chin et al. [9]
SWNTs suspensions	NLR, NLS	Z-scan	
MWNT suspension	NLS	Z-scan and optical limiting measurements	
Metal nanowires	NLS	Optical limiting measurement	Pan et al. [10]
Gold nanoparticles dispersions	NLA, NLS	Z-scan and optical limiting measurements	Reji Philip et al. [11]
CdSe-based quantum dots suspensions	NLR, NLA	Z-scan and optical limiting measurements	Ganeev et al. [12]

forms of bismuth due to quantum confinement effects. Hence, one may utilize these interesting properties of nanostructures to tailor new materials for NLO applications. For example, optical switching, saturable absorption, and optical limiting properties of one-dimensional nanostructures have shown considerable promise compared to bulk materials. In this section, several interesting NLO phenomena which can be observed in various nanostructures using the *Z*-scan technique are discussed. Although the discussion is limited to the *Z*-scan technique, various experimental techniques used for probing the NLO properties of popular nanostructures are summarized in Table 1.

Carbon Nanotubes

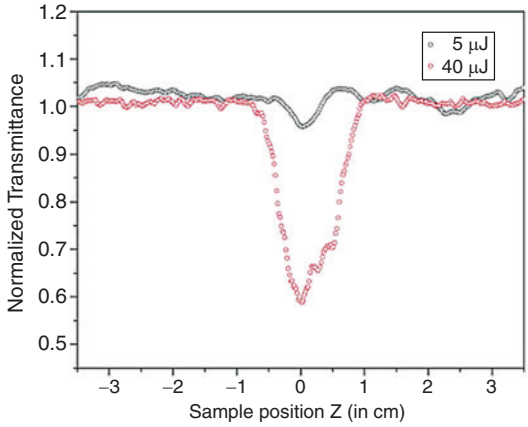
One-dimensional carbon nanotubes (CNTs) have revolutionized various fields of science and technology since their optical and electrical properties allow one to selectively modify and tailor nanotube properties for specific applications. Initial reports on the nonlinear optical limiting performance of multiwalled carbon nanotubes (MWNTs) highlighted the superior performance of nanotubes compared to corresponding properties of C_{60} and carbon black dispersions. Subsequently, it spurred interest in investigating NLO properties of functionalized CNT-based novel nanomaterials. Moreover, the broadband optical absorption profile of carbon nanotubes provides a unique advantage over other optical materials including fullerenes. Theoretical predictions by Margulis et al. claimed that the third-order nonlinear optical susceptibility for MWNTs exceeds that of fullerene molecules by two orders of magnitude [13]. Such enhancement in third-order NLO values arises from the combined effect of π-electron transitions, which are dominant over interband/intraband transitions. Thus, the modification of the π-electron cloud via functionalization or doping provides an efficient means for engineering novel materials with high nonlinear optical susceptibility.

The optical limiting performance of MWNTs at 532-nm excitation wavelength with a 7-ns pulsewidth from a second harmonic generator of a high-power Nd:YAG system (quanta systems) is depicted in Fig. 5. The laser excitation was used in the conventional open aperture *Z*-scan setup (as described in section "Experimental Methods"). Optical limiting response was studied by varying parameters such as the solvent used for nanotube suspension and the incident energy. The MWNTs were suspended in solvents such as deionized water, 2-propanol, and ethanol using sonication (for about ~10 min).

The transmittance curves of MWNT suspensions, as shown in Fig. 5, at different energies indicated improved limiting at higher energies. As expected, in the case of higher energies, early onset of limiting was observed (greater width of

transmittance curve at 40 μJ). The main mechanism responsible for the observed optical response is nonlinear-induced scattering, which was evident as the MWNTs suspension translated across the focal plane in the Z-scan experiment. The observed NLO response of MWNTs arises due to "microbubble" or "microplasma" formation that leads to the observed nonlinear scattering [4].

The leading edge of the pulse is absorbed by the MWNTs and is nonradiatively transferred to the surrounding medium within a few nanoseconds, which causes thermal inhomogeneities in the suspension centered around the nanotubes. Depending on the fluence (F_o) and thermo-optic coefficient (dn/dT), a local change in the temperature results in a local refractive index gradient (Δn) since $\Delta n = (dn/dT)(F0\ \alpha/2\rho C_v)$, where α is absorption coefficient, ρ is the density, and C_v is the specific heat of the sample [14]. The trailing edge of the optical pulse, therefore, "sees" a nonuniform optical medium of scatterers created by the leading edge of the pulse. Thus, the trailing edge gets scattered leading to a spatial redistribution of pulse energy (as shown in Fig. 6)

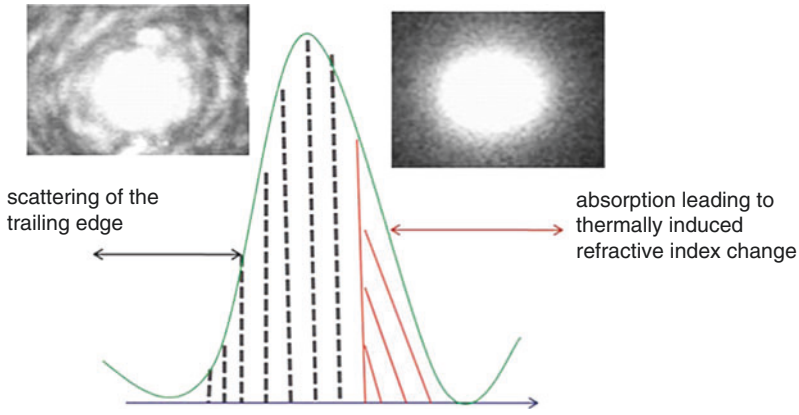


Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 5 Open aperture Z-scan signal for MWNTs suspensions dispersed in water and excited at two different incident energies, 5 μJ and 40 μJ; $\lambda_{excitation} = 532$ nm

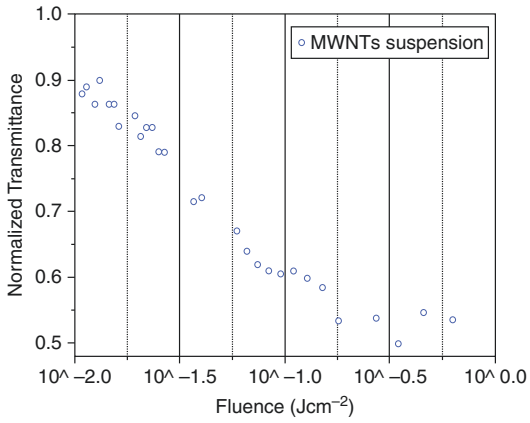
Since the thermodynamical parameters of the solvent media surrounding the nanotubes play a crucial role in the formation of a scattering center, the optical limiting response was also found to be solvent dependent [9]. The threshold fluence for optical limiting (at which the transmittance falls to 50 % its initial value) for MWNTs was determined to be ~0.5 J cm^{-2} (Fig. 7).

Bi Nanorods

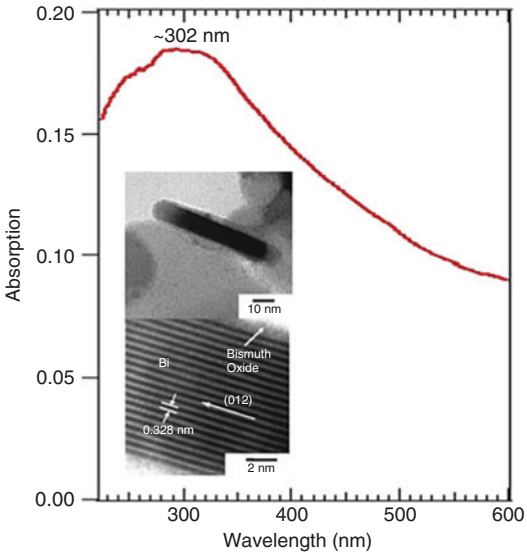
Bismuth is a semimetal in its bulk form and can be prepared in several different nanostructured forms, including thin films, nanowires, nanorods, and quantum dots [15]. Bi is an attractive material for studying how nanostructuring can alter the optical properties of metals. For example, because



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 6 Scheme of nonlinear scattering or induced scattering leading to spatial redistribution of the beam as the intense laser pulse interacts with the MWNTs. The *right inset* shows the mm size laser spot before incidence, and the *left inset* shows the visible scattering after incidence



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 7 Nonlinear optical response of MWNTs suspension at 532-nm excitation



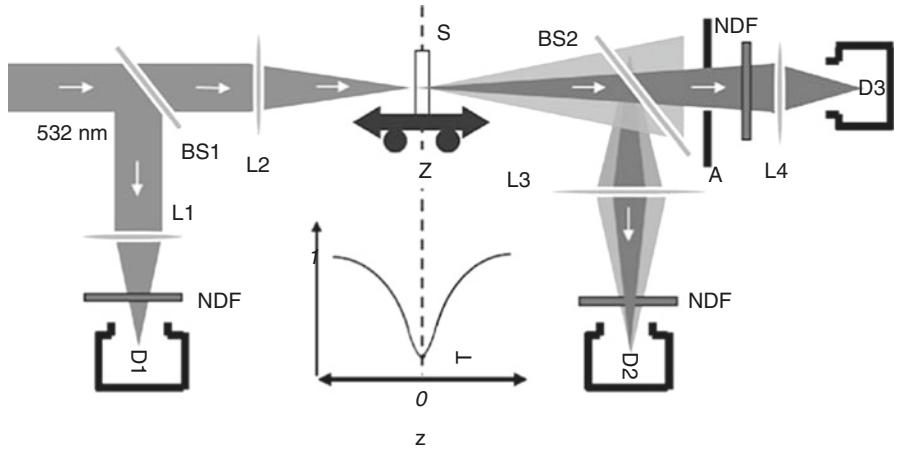
Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 8 UV-Visible absorption of Bi nanorod suspension in the range of 200–600 nm. The broad peak at 302 nm is due to surface plasmons. TEM images of isolated Bi nanorods encased inside a thin oxide layer (*top inset*) and the crystalline planes (012) of trigonal Bi (*bottom inset*)

of the small effective mass components (in the anisotropic effective mass tensor) of Bi, quantum confinement effects are observed in nanowires with diameters as large as 40 nm at low temperature (77 K) [15]. In addition, since the mean free path is long in bismuth (∼250 nm at room temperature), the finite size effects in a Bi crystal are more pronounced. These effects, which are observed in the tens of nanometers range in Bi, occur in other metals only at much smaller length scales and, thus, are likely to be more difficult to probe experimentally. Therefore, understanding these effects in bismuth can assist in the prediction of the optical properties of other nanostructured materials.

A semimetal to semiconductor transition was predicted in the low-dimensional forms of bismuth due to quantum confinement effects. When such a transition occurs, it is expected that several material properties undergo a radical change. For example, the thermoelectric figure of merit is poor in bulk Bi due to nearly equal and opposite contributions from electrons and holes, but is high, ∼2, in small-diameter Bi nanowires. Similar enhancements in NLO properties were also observed for Bi nanorods [16].

Figure 8 shows the transmission electron microscope (TEM) images for Bi NRs synthesized using pulsed vaporization techniques [16]. Unlike bulk Bi, bismuth nanoparticles and nanorods exhibit a broad surface plasmon (SP) peak at ∼281 and ∼302 nm (4.4 and 4.1 eV), respectively. When irradiated by 532 nm (2.33 eV) laser pulses (pulse width ∼7 ns), Bi NRs can absorb energy via interband and intraband transitions, free carrier absorption, and two-photon excitation. The presence of the SP peak makes the absorption a "near-resonant" process. When Bi NRs are suspended in any solvent, the absorbed energy is transferred to the surrounding medium through nonradiative de-excitation, resulting in localized heating and thus scattering centers. In the case of a Z-scan experiment, the leading edge of the pulse creates these scattering centers and the trailing edge of the pulse gets scattered (see Fig. 6). Such a mechanism has been observed for other nanostructured systems such as SWNTs and MWNTs (see section "Carbon Nanotubes").

Previously, Sivaramakrishnan et al. used visible and IR laser pulses from a Nd:YAG laser (quanta system, 7 ns, 10 Hz), at both 532 and

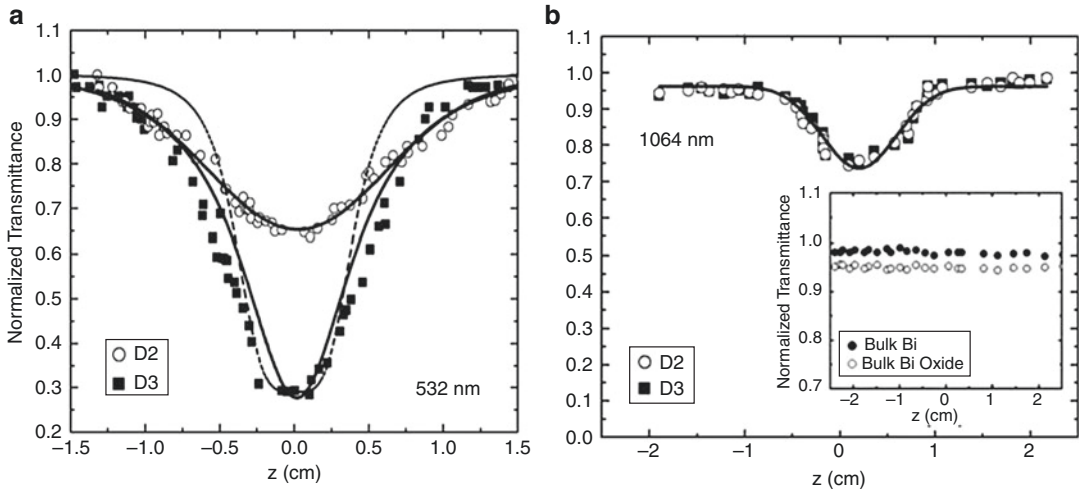


Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 9 Schematic of the experimental setup used for the NLO studies of Bi nanorods. A combination of lenses (*L1–L4*), beam splitters (*BS1*) and (*BS2*), neutral density filters (*NDFs*), aperture A, and detectors (*D1–D3*) are used to measure contributions to optical limiting from the scattered and unscattered components. Bi nanorod suspension is positioned at S and translated through $z = 0$

1,064 nm, to measure the optical limiting performance of Bi nanorod suspensions by the Z-scan technique at room temperature [17]. In order to separate the contributions of NLA and NLS, a modified Z-scan setup (Fig. 9) proposed by Jourdrier et al. was employed. In the modified setup, the beam is split 50:50 (using BS2) to reach two detectors D2 and D3 in order to measure the transmittance T of the sample with detector D1 as a reference. A "2" wide lens (L4) collects light from both the scattered and unscattered parts making D2 insensitive to transmittance changes due to spatial redistribution. This ensures that the collection efficiency is ~90 %. Thus, D2 records changes in transmittance solely due to absorptive effects, and the nonlinear extinction coefficient measured by the Z-scan curve is only due to contributions from NLA. On the other hand, D3 detects only the unscattered part since aperture (A) can be adjusted to prevent the scattered part from reaching D3. Therefore, D3 records a reduced transmittance due to NLS and NLA, unlike the case of D2. The Z-scan curves for the Bi nanorods suspended in chloroform (linear optical transmittance of 60 %) in a 1-mm cuvette made of fused silica (under a 55-μJ, 532-nm, 7-ns excitation) are shown in Fig. 10.

Figure 10a shows the Z-scan curves recorded by detectors D2 (NLA alone) and D3 (NLA and NLS). The depth of the valley in the Z-scan curve is a direct measure of the extent of optical limiting and is different in the D2 and D3 curves with dips of 0.65 and 0.25, respectively. Since D2 records the contribution due to NLA alone and the D3 curve also has contributions from NLS, it is clear that NLS induced by NLA makes a significant contribution to the total extinction. In fact, the contribution of NLS to optical limiting is greater than that due to NLA. Interestingly, in the case of 1,064-nm excitation (55 μJ, 7 ns), the D2 and D3 curves overlap (Fig. 10b), and the minimum transmittance is ~0.75 in both the cases. Thus, NLS is absent at 1,064 nm, and only NLA makes a contribution to the total extinction observed at this wavelength. At 532 nm, however, both NLA and NLS are responsible for the observed optical limiting unlike the 1,064-nm case. This is very unlike the behavior of carbon black and SWNT/MWNT suspensions, where NLS dominates at both 532- and 1,064-nm excitations (see section "Carbon Nanotubes").

As shown in Fig. 8, Bi nanorods used in this study contain a native oxide layer. Z-scan experiments performed using 532-nm excitation on



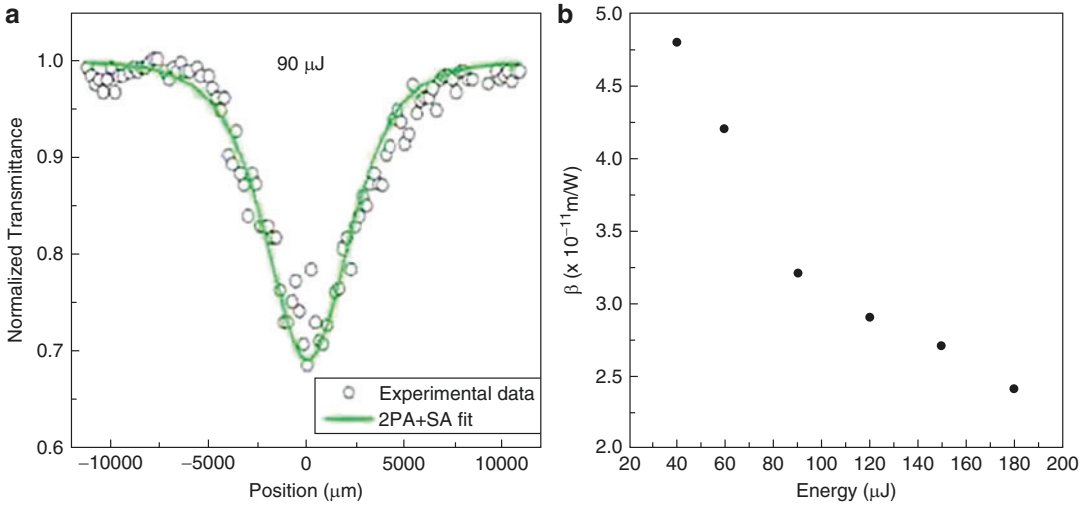
Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 10 (**a**) Normalized transmittance measured at 532-nm excitation by detectors D3 (nonlinear absorption and scattering) and D2 (nonlinear absorption only) for an input energy of 55 μJ. The minimum transmittance recorded by D3 is less than that of D2 as it has contributions due to both scattering and absorption. (**b**) Same as in (**a**) using the 1,064-nm excitation for an input energy of 55 μJ. The *overlapping curves* suggest that nonlinear scattering is absent when excited with 1,064 nm in contrast to the 532-nm case

bulk Bi and bulk Bi_2O_3 did not display any nonlinearity (inset in Fig. 10b).

However, ball-milled Bi_2O_3 powder exhibited strong nonlinearity when excited at a higher fluence (90 μJ; Fig. 11a). Similar to the case of Bi nanorods, the predominant cause for such nonlinearity was found to be two-photon absorption (2PA). Additionally, the saturation of the ground state absorption was also observed for ball-milled Bi_2O_3 samples. Thus an effective nonlinear absorption coefficient $\alpha(I)$, given by $\propto (I) = \frac{\alpha_0}{1 + \frac{I}{I_S}} + \beta I$ was considered for the fitting of the data shown in Fig. 11a, where I_S is the saturation intensity (intensity at which the linear absorption drops to half its original value) and β is the 2PA coefficient. Since Bi nanorods posses a high surface area (hence more of a surface oxide layer), it is possible that the nonlinear response discussed in Fig. 10 arises from the native surface oxide layer rather than core Bi nanorods. Furthermore, in Bi_2O_3, the numerically calculated β's were found to be decreasing with increasing energy as shown in Fig. 11b. The falloff of β with increasing intensity is a consequence of reverse saturation absorption arising from a sequential 2PA [2]. Thus, the contribution of the native Bi_2O_3 layer to the nonlinear properties of Bi nanorods cannot be ruled out despite the absence of any nonlinearity in bulk forms of Bi and Bi_2O_3.

Surface-Modified NiS_2 Nanoparticles

Photonic nanoparticle systems have attracted considerable attention due to a wide variety of NLO applications such as optical limiting, saturable absorption, second harmonic generation, two-photon absorption, all-optical control, and holographic storage control. Typically, such systems are designed with a view of enhancing one of the nonlinear optical properties over their bulk counterparts. Previously, it was shown that nanomaterials (such as surface-modified NiS_2) can also be engineered to deliver systems displaying more complex behavior involving the interplay of two or more nonlinear phenomena [17]. If in such systems, two competing mechanisms operate simultaneously, however, the stronger dominates the weaker. Decoupling them in a single experiment is not straightforward. This section discusses the case of complete quantitative characterization in a single experiment of two *simultaneous* but *contrary* optical transmission



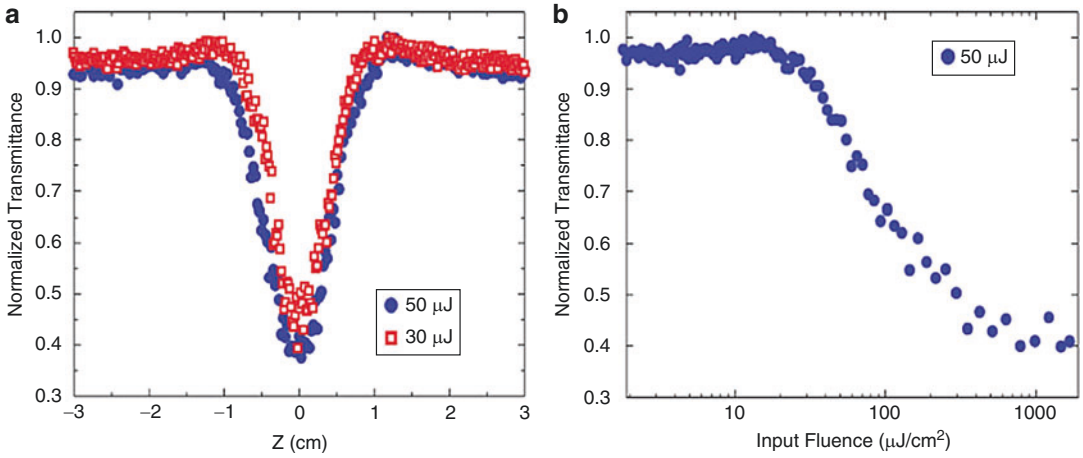
Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 11 (**a**) *Z*-scan data from ball-milled Bi_2O_3 samples obtained with 532-nm, 7-ns pulses. (**b**) The variation of 2 PA coefficients as a function of incident laser energy

phenomena – optical limiting and saturable absorption in surface-modified NiS_2 nanoparticle suspensions by employing a modified *Z*-scan method as in the case of Bi nanorods (Fig. 9). Capping with dimethylglyoxime (DMG) results in the surface modification of the nanoparticles and exhibits a spectral feature essential for the saturable absorption.

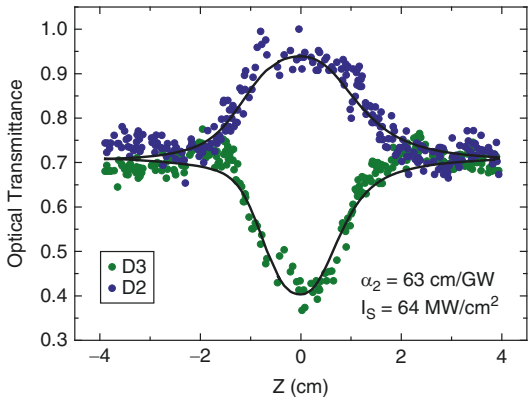
NiS_2 nanoparticles were suspended in an appropriate amount of 2-propanol so that the sample held in a 1-mm-wide fused silica cuvette would have a linear transmittance of 80 % at 532-nm wavelength. The second harmonic of a Q-switched Nd: YAG laser (532 nm) of pulse width 7 ns at 10-Hz repetition was used in the *Z*-scan studies employing a converging lens with a focal length of 20 cm. The *Z-scan curve* and transmittance as a function of input fluence are shown in Fig. 12a. It is clearly evident from Fig. 12b that the optical transmittance falls as intensity increases, exhibiting an *optical limiting* behavior. A transmission of 50 % (which is the threshold fluence for optical limiting) is reached at 300 μJ/cm^2. Such optical limiting in NiS_2 occurs due to the intensity-dependent scattering that is also visible to the naked eye.

Figure 13 shows the *Z*-scan curves (with the set up shown in Fig. 9) from detectors D2 (•) and D3 (•) for NiS_2 nanoparticles suspended in 2-propanol. The *Z*-scan curve due to detector D2 (•) shows an upward *peak* indicative of the saturation of absorption. The transmittance of the sample increases as it is brought closer to the focus of the *Z*-scan lens. This is in sharp contrast to the studies on Bi nanorods discussed earlier where the corresponding curve is a *valley* indicating that nonlinear absorption increases with intensity and transmission falls with intensity (Fig. 10). Saturable absorption observed in this case is similar to the behavior of rhodamine B under similar experimental conditions [6].

Furthermore, in Fig. 13, the Z-scan curve of detector D3 (•) which records intensity-dependent transmittance due to both absorption and scattering shows a typical optical limiting valley. Closer observation reveals that both the fall in transmittance due to optical limiting (•) and the rise in transmittance (•) due to saturable absorption are at a z-position of 1.75 cm with respect to the focal point ($z = 0$). Thus, the two effects are simultaneous and not sequential as has been observed in other experiments on nanoparticles systems [18]. The sequential occurrence of saturable absorption followed by an optical limiting behavior is indicative of excited state absorption.



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 12 (**a**) Z-scan curves for nickel sulfide nanoparticles in 2-propanol and (**b**) the corresponding plot of optical transmission as a function of input fluence



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 13 Z-scan curves of detector D3 (•) and D2 (•). While D3 measures the contribution of both nonlinear scattering and nonlinear absorption, D2 gives a measure of the intensity-dependent absorption alone

The UV–vis absorption spectrum for NiS_2 nanoparticles shows a strong absorption at ~554 nm (Fig. 14). Under 532-nm excitation, the nanoparticles are excited from the ground state (labeled "0") to a higher state (labeled "1") as shown in the *inset* in Fig. 14. As the intensity of the exciting laser pulse is increased to a maximum of 200 MW/cm^2, most of the nanoparticles occupy the excited state. As the ground state is bleached, the system becomes increasingly transparent to the incident laser pulse at 532 nm, resulting in a saturation of absorption and the consequent *Z*-scan curve.

The intensity-dependent absorption coefficient for a two-level system is given by [2]:

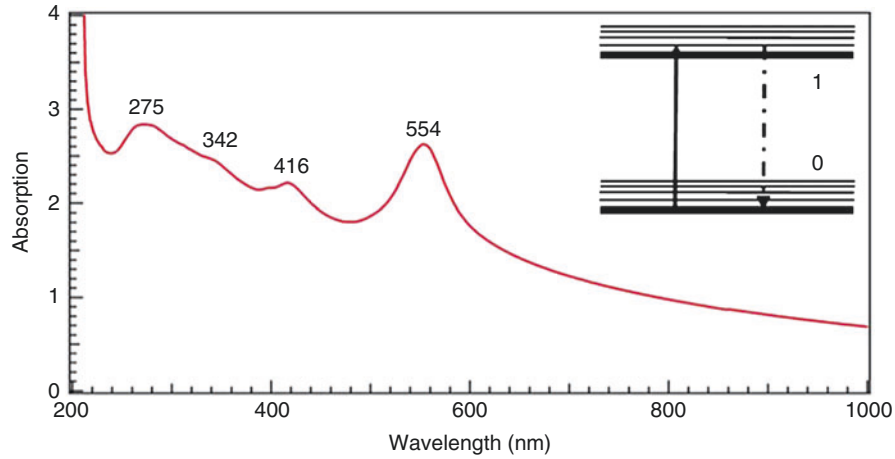
$$\alpha_{SA}(I) = \frac{\alpha_0}{1 + \frac{I}{I_s}} \tag{5}$$

where α_0 is the absorption coefficient, I is the intensity of the laser beam, and I_S the saturation intensity. This explains the upper curve in Fig. 13. The bottom curve can be understood in terms of simultaneous nonlinear intensity-dependent scattering that occurs due to energy transfer from nonradiative decay of excited states, as explained earlier in discussions of MWNT and Bi nanorod systems.

A complete description of the *Z*-scan experiment requires that both the saturable absorption and optical extinction due to nonlinear scattering be taken into account. Thus, in the similar lines of Venkatram et al. [18], we use the following equation describing these two concomitant processes:

$$\frac{dI}{dz} - \left[\alpha_0 + \alpha_2 I + \frac{\alpha_0}{1 + \frac{I}{I_s}}\right] I \tag{6}$$

N



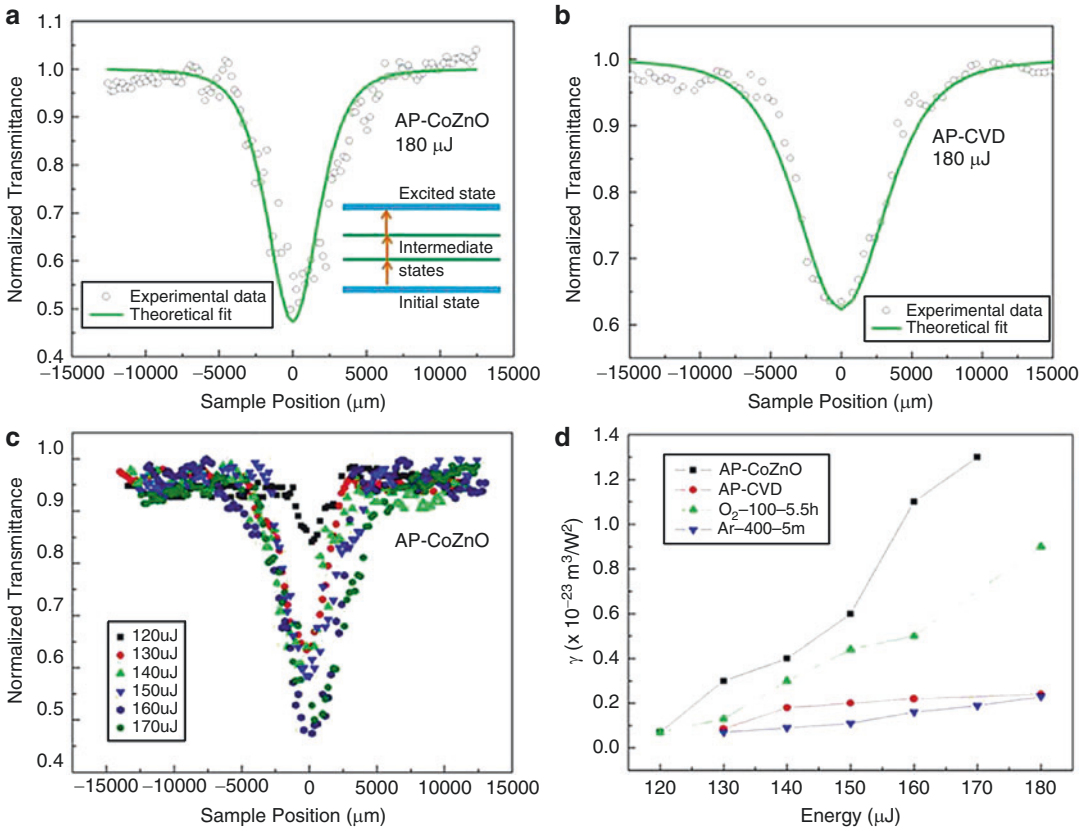
Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 14 UV-Visible spectrum of NiS_2 nanoparticles in 2-propanol. *Inset* is a schematic of the excited manifold for saturable absorption as explained in the text

In the above equation, α_2 is the nonlinear scattering coefficient. The value of I_S and α_2 can be extracted from numerical fitting of data to this propagation. Here, the fourth-order Runge–Kutta method is used to solve the equation while finding the best fit for both parameters [6]. As mentioned earlier, the *Z*-scan curve of D2 is purely due to absorptive effects. The equation is first solved setting $\alpha_2 = 0$ so as to obtain a value for *IS* which in this case is 64 MW/cm^2. This value of *Is* is much smaller than the incident intensity at the focal point (at both 30 μJ and 50 μJ), eventually favoring the domination of nonlinear scattering following the absorption of a faction of the leading part of the incident pulse. Thereafter, the value of α_2 was obtained by using this value of I_S in the complete equation and finding the best fit to the Z-scan curve from detector D3. The nonlinear scattering coefficient, α_2, was found to be 63 cm/GW.

ZnO Nanostructures

Among the metal oxides, ZnO is a remarkable wide bandgap (3.37 eV in bulk) II–VI semiconductor. Its high exciton binding energy (60 meV) allows efficient exciton emission even at room temperature. As a consequence of its noncentrosymmetric crystal structure, ZnO is expected to have nonzero second-order susceptibility. Although many properties of bulk and nanostructured ZnO have been well studied, there is still a lack of understanding in the relationship between the density of defects, dopant level, synthesis conditions, and the observed optical and magnetic properties. In nanostructures, a fairly high surface state density significantly modifies or alters the chemical and physical properties, which can be reflected in optical measurements. NLO properties of different forms of pristine ZnO, such as single crystals, thin films, nanolayers, and nanowires, have been previously reported. It is also very important to understand the NLO properties of transition metal (TM)-doped ZnO nanostructures due to their ferromagnetic ordering at room temperature and importance in spintronic applications. An extensive study of NLO properties of TM-doped ZnO nanostructures can lead to a better understanding of the impact of surface states.

In earlier studies of the origin of ferromagnetism (FM) in pristine micro- and nanostructured ZnO, Podila et al. observed the changes in magnetic properties with varying crystallinity. They attributed the FM to the presence of surface states [8]. The presence of such surface states in pristine and Co-doped ZnO nanostructures (denoted as AP-CVD and AP-CoZnO) can be confirmed using the open aperture Z-scan. It is worth noting



Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures, Fig. 15 (**a**) Z-scan data for AP-CoZnO samples fitted to a three-photon absorption (*3PA*) equation. *Inset* is a schematic depicting effective 3PA. When the intermediate states are virtual, the effect is genuine 3PA, but when they are real, it is an effective 3PA arising mostly from excited state absorption. (**b**) Z-Scan data for pristine ZnO samples fitted to a 3PA equation. (**c**) Z-scan data for AP-CoZnO samples at different incident laser energies. (**d**) Variation of the 3PA coefficient (γ) as a function of the incident laser pulse energy for AP-CVD, AP-CoZnO, O_2-100-5.5 h and Ar-400-5 m samples

that, except when excited by ultrafast laser pulses, ΔT strongly depends on the excited state population density. In view of this, the nonlinear transmission of the ZnO samples was measured in the nanosecond excitation regime using linearly polarized 5-ns optical pulses from a Q-switched frequency-doubled Nd:YAG laser operating at 532 nm. The Z-scan data obtained for AP-CoZnO and AP-CVD samples (Fig. 15a, b, respectively) were found to be best-fit numerically by a three-photon absorption (3PA) process model. It is instructive to note that, compared to the 2PA process described in Eq. 7, the nonlinear absorption coefficient α in the presence of 3PA is given by

$$\alpha(I) = \alpha_o + \gamma I^2 \tag{7}$$

In the above equation, I is the intensity of the laser, α_o the unsaturated linear absorption coefficient, and γ the 3PA coefficient. The value of γ was obtained by fitting the Z-scan curve to the 3PA propagation equation, given by

$$\frac{dI}{dz} = -\alpha_o I - \gamma I^3 \tag{8}$$

with z' being the propagation distance within the sample.

The values of γ numerically obtained from the fits indicated that nonlinear absorption in the

present case arises not only from genuine 3PA but also from "effective" three-photon nonlinearity [1]. Such effective 3PA originates from sequential excited state absorption as shown in the inset of Fig. 15a. In the case of genuine three-photon absorption (3PA), where the transition states involved are virtual, the 3PA coefficient (γ) is a constant and is independent of the incident laser fluence. All genuine ground state absorption coefficients (such as 2PA or 3PA) are dependent on the ground state population N since the absorption coefficient is obtained by multiplying the corresponding absorption cross section (σ) by N. The absorption cross section is a microscopic parameter, independent of N that characterizes the average two-photon/three-photon absorbability per molecule. However, the change in γ will be evident only when there is a substantial change in the ground state population due to absorption. Therefore, if the absorptions are weak, then the coefficients can be considered to be almost constant, since there is a negligible change in N. Genuine 2PA and 3PA are usually very weak phenomena, and, hence, the corresponding coefficients can be considered as material constants at a given wavelength and concentration. But strong excited state absorptions (ESA) during an effective 3PA deplete the ground state population significantly so that the absorption coefficient is no longer a constant. Hence a change in γ with respect to the incident laser intensity can help confirm the presence of surface states in the forbidden gap of ZnO.

Both two-photon absorption and/or sequential absorption of two photons at 532 nm (2.33 eV; inset Fig. 15a) will result in a real terminal level which is above the lowest excitonic state that lies around 363 nm (3.4 eV) for ZnO, and another one-photon absorption to go from this level to the band edge. Figure 15c, d show the variation of γ for AP-CoZnO, AP-CVD, and AP-CVD samples annealed in O_2 (Ar) for 5.5 h (5 min) at 100 °C (400 °C)−O_2-100-5.5 h and Ar-400-5 m, respectively – as a function of incident laser energy. While γ remains almost constant with input energy in AP-CVD, there is a pronounced increase in γ with incident energy in AP-CoZnO and O_2-100-5.5 h samples. Moreover, the 3PA coefficient (and, hence, the optical limiting efficiency) for AP-CoZnO samples is obviously higher than that of pristine ZnO. These observations concur with the enhancement seen in the magnetic moment of AP-CoZnO and O_2-100-5.5 h samples and indicate the presence of enhanced surface state density due to Co atoms in the ZnO lattice. Thus, the increase of γ with the incident laser energy confirms effective 3 PA due to the existence of a wide range of surface states in AP-CoZnO and O_2-100-5.5 h samples. Furthermore, the lowest γ values and a constant trend with input fluence were observed for Ar-400-5 m samples indicating the absence of any surface states. An enhanced surface state density in AP-CoZnO due to Co doping results in stronger FM and better optical limiting compared to pristine ZnO nanostructures [19].

Conclusion

In the preceding sections, the reader is introduced to optical nonlinearities and given an overview of various methods used for nonlinear optical characterization with emphasis on the *Z*-scan technique. The chapter discusses in detail the NLO properties of various nanostructures, such as carbon nanotubes, bismuth nanorods, NiS_2 nanoparticles, and ZnO nanowires, studied using the Z-scan. In the case of MWNTs and Bi nanorods, nonlinear scattering is found to play a major role in determining the nonlinear transmission. The simultaneous onset of two different nonlinear processes (saturable absorption and nonlinear scattering) was observed in NiS_2 nanoparticles. In pristine and doped ZnO nanostructures, the intensity dependence of the 3PA coefficient could be used to confirm the presence of intermediate states. These studies indicate that organic as well as inorganic nanostructures are excellent candidates for nonlinear optical transmission with potential strategic applications, including optical power limiting.

References

1. Sutherland, R.L.: Handbook of Nonlinear Optics. Marcel Dekker, New York (1996)
2. Nalwa, H.S., Miyata, S.: Nonlinear Optics of Organic Molecules and Polymers, p. 841. CRC Press, BacoRato (1997)
3. Sheik-Bahae, M., Said, A.A., Wei, T.H., Hagan, D.J., Van Stryland, E.W.: Sensitive measurement of optical nonlinearities using a single beam. IEEE J. Quantum Electron. **26**, 760 (1990)
4. Tutt, L.W., Boggess, T.F.: A review of optical limiting mechanism and devices using organics, fullerenes, semiconductors and other materials. Prog. Quantum Electron **17**, 299 (1993)
5. Gu, B., Lou, K., Wang, H.T., Ji, W.: Dynamics of two-photon induces three-photon absorption in nanosecond, picosecond and femtosecond regimes. Opt. Lett. **35**, 3 (2010)
6. Rao, S.V., Srinivas, N.K.M.N., Rao, D.N.: Nonlinear absorption and excited state dynamics in Rhodamine B studies using Z-scan and degenerate four wave mixing techniques. Chem. Phys. Lett. **361**, 439 (2002)
7. Balu, M., Joel, H., David, H., Van Stryland, E.: White-light continuum Z-scan technique for nonlinear materials characterization. Opt. Express **12**(16), 3820–3826 (2004)
8. Podila, R., Queen, W., Nath, A., Fazzio, A., Schonalez, A., He, J., Dalpian, G., Skove, M.J., Hwu, S.J., Rao, A.M.: Origin of FM ordering in pristine micro- & nanostructured ZnO. Nano Lett. **10**, 1380 (2010)
9. Chin, K.C., Gohel, A., Elim, H.I., Chen, W., Ji, W., Chong, G.L., Sow, C.H., Wee, A.T.S.: Modified carbon nanotubes as broadband optical limiting nanomaterials. J. Mater. Res. **21**, 2758 (2006)
10. Pan, H., Chen, W., Feng, Y.P., Ji, W.: Optical limiting properties of metal nanowires. Appl. Phys. Lett. **88**, 223106 (2006)
11. Philip, R., Kumar, G.R., Sandhyarani, N., Pradeep, T.: Picosecond optical nonlinearity in monolayer protected gold, silver, and golf-silver alloy nanoclusters. Phys. Rev. B **62**, 13160 (2000)
12. Ganeev, R.A., Ryasnyansky, A.I., Redkorechev, V.I., Fostiropoulos, K., Prieb, G., Usmanov, T.: Variations of nonlinear optical characteristics of C60 thin films at 532 nm. **225**, 131 (2003)
13. Margulis, A.: Theoretical estimations of third-order optical nonlinearities for semiconducting carbon nanotubes. J. Phys. Condens. Matter **11**, 3065 (1999)
14. Joudrier, V., Bourdon, P., Hache, F., Flytzanis, C.: Nonlinear light scattering in a two-component medium: optical limiting application. Appl. Phys. B Lasers Opt. **67**, 627 (1998)
15. Black, M.R., Lin, Y.-M., Cronin, S.B., Rabin, O., Dresselhaus, M.S.: Infrared absorption in Bi nanowires resulting from quantum confinement. Phys. Rev. B **65**, 195417 (2002)
16. Sivaramakrishnan, S., Muthukumar, V.S., Reppert, J., Anija, M., Sivasankara Sai, S., Philip, R., Venkataramaniah, K., Kuthirummal, N., Rao, A.M.: Nonlinear optical scattering and absorption in Bi nanorod suspensions. Appl. Phys. Lett. **91**, 093104 (2007)
17. Muthukumar, V.S., Kiran, J.K., Reppert, J., Satyajit, R., Krishna, V., Nageshwar Rao, G., Siva Rama Krishnan, S., Siva Sankara Sai, S., Venkataramaniah, K., Rao, A.M.: Nonlinear optical transmission of surface modified NiS nanoparticles: saturation of absorption and optical limiting. NANO: Brief Rep. Rev. **3**, 161 (2008)
18. Venkatram, N., Sai Santosh Kumar, R., Narayana Rao, D.: Nonlinear absorption and scattering properties of CdS nanocrystals with its application as a potential optical limiter. J. Appl. Phys. **100**, 074309 (2006)
19. Podila, R., Anand, B., West, J.P., Philip, R., He, J., Skove, M., Hwu, S.J., Rao, A.M.: Evidence for surface states in pristine and Co-doped ZnO nanostructures: magnetization and nonlinear optical studies. Nanotechnology **22**, 095703 (2011)

Nonlinear Parametric MEMS

Congzhong Guo
ECE Department, Carnegie Mellon University, Pittsburgh, PA, USA

N

Synonyms

Parametrically excited nonlinear MEMS resonators

Definition

Nonlinear MEMS resonators excited by parametric excitation so its effective impedance is periodically modified by time-varying drive parameters

Background, Benefits, and Applications

In the rich context of physical and engineering disciplines, the first decade of systematic study on the nonlinear problems and parametrically excited systems dates back to the mid-nineteenth

century, with the work of Mathieu [1], Rayleigh [2], Stoker [3], and Nayfeh [4] as important landmarks. Owing to the use of a considerable amount of exemplary nonlinear problems, the forerunners provided intuitive concepts, such as self-excited systems, forced oscillations, parametric pump, and nonlinear restoring force, and backed them up with rigorous theorems. Recent work on the parametric resonance, especially in the context of resonant microelectromechanical systems (MEMSs), investigates the utility and advantage of exploiting parametric excitation in the areas of gravimetric [5] and chemical [6] sensing, inertial sensing [7], scanning tunneling microscopy [8], closed-loop servo [9], and scanning micromirrors [10].

A second-order mechanical system is typically excited into resonance using harmonic excitation – a time-varying excitation force whose frequency tracks the natural frequency of the system being driven. A system is excited parametrically when its effective impedance is periodically modified by time-varying parameters, such as a periodic driving force. For example, a pendulum can be "pumped" into parametric resonance when its rotational moment of inertia changes twice as fast as the pendulum's natural frequency. One could more intuitively picture a kid on a swing where periodic variation in the center of mass could lead to larger-amplitude rotational motion, if the variation in length is at a frequency that is twice of the swing's natural frequency.

One of the well-known benefits of the nonlinear parametric resonance is the bifurcation phenomenon, in which a directional sweep of drive frequency or phase creates amplitude "jump" in the region around twice the natural frequency. The bifurcation occurs in a system that exhibits bistability. A bistable system has two degenerated steady states. Perhaps the simplest bistable system is a binary bit which has "0" and "1" states. A state transition between "0" and "1" is analogous to the bifurcation phenomenon in MEMS. In the context of the MEMS parametric resonator, a sharp jump in the mechanical displacement amplitude from zero to a nonzero value could take place at a so-called "bifurcation frequency." The frequency at which the jump happens usually depends upon the direction of the parameter sweep, providing a hysteretic effect. Detecting this sharp jump therefore enables sensitive detection of changes in the resonant frequency, which can either be triggered by a change in mass [5] or stress that modifies the suspension spring constant.

Great resolution can be achieved given the bistable nature at the bifurcation point. With this capability, the class of nonlinear resonator becomes a good candidate for highly sensitive mass (i.e., gravimetric) and stress detector. Typically, the parametric resonance frequency links directly to the system's natural frequency, with a proportional constant that can be calculated through the perturbation theory introduced later in the text. Correlating the bifurcation frequency with the vibratory mass [5], one could measure the frequency shift on the sharp boundary of the bifurcation to determine the ultrasmall mass change. Unlike a harmonic-mode resonant mass sensor, the parametric-mode gravimetric detector is more immune to the amplitude noise [11] because the bifurcation response retains its sharpness in the presence of the intentionally injected detection noise. The prior demonstration of attogram mass sensitivity [12] through frequency resolution enhancement in a parametrically excited mass sensor offers motivation for parametric control [9], which is potentially capable of providing extremely sensitive detection of mass change.

Given ample design consideration, certain kinds of mechanical nonlinearity combined with electrostatic/piezoelectric/electromagnetic parametric drive give rise to flat and large-amplitude response over relatively wide-frequency ranges that are potentially useful in making manufacturable resonant devices. For devices with mismatched resonant modes, wideband frequency response makes mode matching possible and avoids the need for frequency tuning, example of which includes matched-mode gyroscope design in the presence of process mismatch [7]. Parametric resonance in MEMS finds further applications in effective Q enhancement through noise squeezing [13] and signal parametric amplification [14], parasitic signal reduction, and mechanical

modulation [15]. Certain classes of nonlinear MEMSs can induce chaos behavior [16] with a particular set of oscillator parameters and could be potentially used for random number generation and signal encryption.

Parametric Excitation

Interdisciplinary nature of the MEMS inherently adopts various distinctive transduction mechanisms for parametric excitation, such as electrostatic, electromagnetic, piezoelectric, and optical actuations. In electrostatically driven resonant MEMS, the most effective time-varying parameter arises from tailored electrostatic actuators that introduce an equivalent electrostatic stiffness that varies with the excitation voltage. Turner and collaborators [8] experimentally measured five parametric resonant modes in a torsional oscillator excited through comb-drive levitation and observed that the parametric resonance occurs near the drive frequencies of $2\omega_r/n$, where ω_r is the natural resonant frequency and n is an integer larger or equal to 1. In most parametrically excited engineering electromechanical systems, the principal parametric resonance ($n = 1$) occurring at twice the natural frequency ($2\omega_r$) can be experimentally observed.

Fundamental difference between the parametric and harmonic resonators is the way they are excited. The harmonic resonator usually has a time-invariant mechanical spring constant, with an external drive that is sinusoidal in time t and independent of the displacement x. At harmonic resonance, the drive frequency tracks the natural frequency of the system. Parametric excitation, on the other hand, arises when the force applied on a MEMS resonator is not only a function of the time but also a function of the displacement of the resonator [17], creating time-varying impedance for the nonlinear system. The combined effective stiffness of the MEMS is periodically modulated by the parametric drive. Research literature is ripe with examples of exciting parametric resonance with a drive whose frequency doubles the natural frequency of the system. With proper designs, a parametric MEMS resonator can be excited into either harmonic or parametric mode with an appropriate selection of the driving electrode and the actuation condition – drive frequency and amplitude.

Electrostatic Parametric Excitation

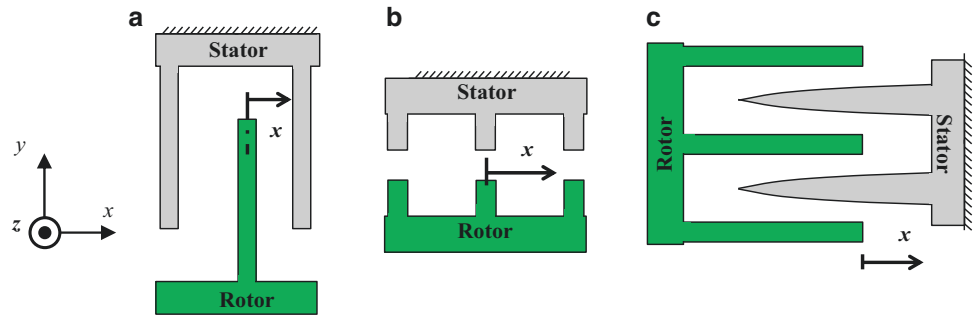
In an electrostatically excited parametric MEMS, the electrostatic force drives the parametric resonator at an electrical frequency, ω_d, which creates a time-varying displacement-dependent spring constant, k_{eff}. Parametric resonance conditions are satisfied when the periodic variation of k_{eff} is at twice the natural frequency, i.e., $\omega_d = 2\omega_r$. A typical sinusoidally varying electrostatic parametric force, F_p, can be expressed as

$$F_p(t,x) = -\left(\gamma_1 x + \gamma_3 x^3\right)\ \left[V_{dc}^2 + V_{ac}^2 \cos\left(2\omega_r t\right)\right] \tag{1}$$

where t is time, x is the displacement of the mass, and V_{dc} and V_{ac} are the DC and AC components of the voltage applied on an electrostatic drive comb that is designed to create a voltage-controlled linear spring constant, γ_1, along with a cubic-displacement nonlinearity, γ_3.

Electrostatic parametric excitation in Eq. 1 is generated with several tailored electromechanical drive topologies. Combined direct and parametric drive can be generated through variable-gap electrostatic transduction. A simple example is a parallel-plate capacitive voltage-force transducer (Fig. 1a), which provides a direct excitation (i.e., constant force) and a linear softening electrostatic spring constant (i.e., γ_1 in Eq. 1), for sufficiently small displacement relative to the gap. Large displacement motion introduces higher-order forcing terms including the cubic spring constant (i.e., γ_3 in Eq. 1) that are usually avoided in this type of transducer design.

In parametric resonant MEMS, it is necessary to identify a parametric drive that provides pure parametric excitation, examples of which include tailored electrostatic comb drives. A well-known example is the non-interdigitated finger (NIF) comb drive [18, 19] (Fig. 1b), which is widely used to provide a well-defined voltage-controlled spring constant. The NIF comb drive is



Nonlinear Parametric MEMS, Fig. 1 Representative examples of electrostatic parametric transduction: (**a**) parallel-plate drive, (**b**) NIF, and (**c**) shaped-finger comb. *x* indicates the direction of motion

constrained to move in a direction orthogonal to the finger length (labeled in *x*) and makes use of the fringing electric field generated between nonoverlapping comb fingers. As the rotor moves away from the stator in $+x$ direction, electric field distribution is modified, and a retention force applies to the rotor in $-x$ direction whose magnitude is proportional to *x*. Both the linear and cubic stiffness can be independently tuned by the ratio of the finger width to its pitch [20] so the parametric response can be adjusted accordingly. The maximum displacement in this NIF design is limited to $\pm 1/2$ of the comb pitch (usually on the order of 1 1 μm mu;m for a typical MEMS NIF resonator). Necessary nonlinearity required to excite the parametric resonance imposes this constraint.

In the interest of extending the displacement range, an electrostatic "shaped-finger" comb is synthesized [21] to provide a large stroke with a small design footprint. It utilizes a nonlinearly varying electrode gap (Fig. 1c) between rotor and stator to provide the necessary electrostatic stiffness for exciting larger-stroke parametric resonance (usually on the order of 10 μm), which is a constraint partly by the length of the comb fingers and partly by the fabrication tolerance. Similar to the constant-gap parallel-moving straight-finger capacitive comb (the same comb as Fig. 1a but moves in *y*), the shaped-finger comb moves laterally in the plane of the wafer and parallel to the finger length, allowing sufficient clearance for large displacement motion. The shaped-finger comb concept was originally introduced by Hirano [22] to achieve a smaller gap between electrodes beyond the fabrication constraints. Later analytic study and experimental validation of the shaped-finger focus on its tuning dynamics, in which the electrode shape is tailored to provide specific electrostatic linear stiffness [23] and to further generate parametric drive [21]. For a resonator that utilizes the shaped-finger comb, separate linear displacement sensing combs can be included to electronically measure the parametric resonance in these large-stroke systems.

Other Parametric Excitation Mechanisms

Other parametric excitation mechanisms have also been adopted as alternative approaches to introduce time-varying excitation via parametric drives. In piezoelectric parametric excitation [6, 24], applying a periodic voltage to an external shear piezoelectric element modulates the total effective stiffness, which is equivalent to applying a periodic base excitation to a resonant device. In electromagnetic parametric excitation [14, 25, 26], the alternation of the polarity of the current and the magnetic field could generate direct or parametric Lorentz force excitation to a MEMS device with current-carrying metal traces that are exposed in a magnetic field. Thermal effect introduced by a laser beam [27] can also modify the time-varying impedance of the silicon disk oscillator, creating parametric amplification effect with proper selection of the optical pumped drive phase.

Parametric Resonance via Electrostatic Excitation

The second-order ordinary differential equation describing a nonlinear mass-spring-damper system is

$$m\ \ddot{x} + b\ \dot{x} + k_1 x + k_3 x^3 = F_p(t, x) \qquad (2)$$

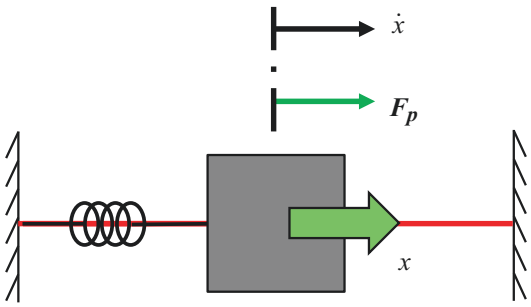
where m is mass, b is the damping coefficient, $\dot{x}$ is the first derivative of x with respect to time t, and k_1 and k_3 are linear and cubic mechanical spring constants (stiffness), respectively. Spring topology and composing beam dimensions set the mechanical stiffness and intentionally introduce essential nonlinearity to the nonlinear system. Substituting the electrostatic parametric force in Eq. 1 into Eq. 2 and rearranging terms yield a nonlinear Mathieu equation:

$$m\ \ddot{x} + b\ \dot{x} + \left(k_1 + \gamma_1 V_{dc}^2\right)x + \left(k_3 + \gamma_3 V_{dc}^2\right)x^3 = -\left(\gamma_1 x + \gamma_3 x^3\right)\ V_{ac}^2 \cos\left(2\omega_r t\right) \qquad (3)$$

Conceptual Example

A conceptual example of parametric excitation is shown in Fig. 2. A block material suspended by a mechanical spring experiences an electrostatic parametric excitation in the form of Eq. 1. Simplifying Eq. 1 by enforcing $V_{dc} = 0$ and $\gamma_3 = 0$, Eq. 1 can be rewritten as

$$F_p = -\gamma_1 x\ V_{ac}^2 \cos\left(2\omega_r t\right) \qquad (4)$$



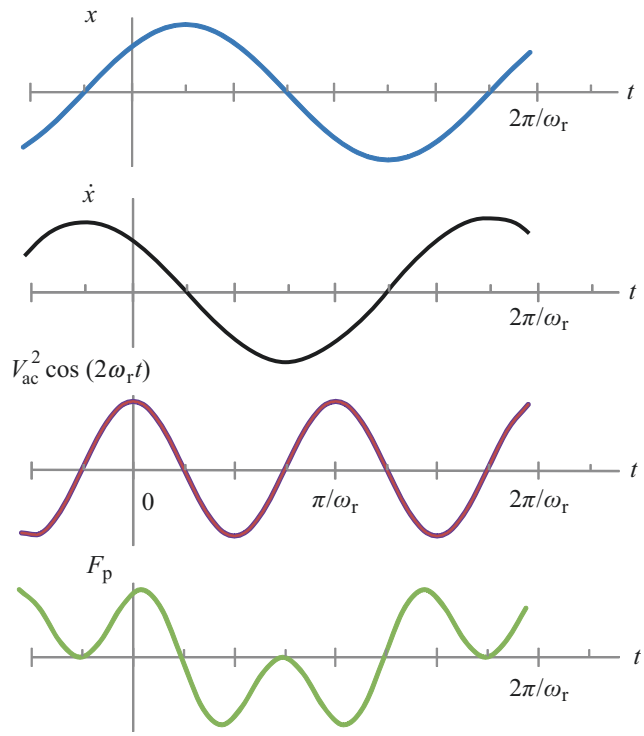
Nonlinear Parametric MEMS, Fig. 2 A block material suspended by a mechanical spring experiences a parametric excitation in the form of Eq. 1

The waveforms of the displacement, x, velocity, $\dot{x}$, the square of the parametric voltage, $V_{ac}^2 \cos(2\omega_r t)$, and parametric force, F_p, are plotted in Fig. 3. The applied parametric force is always in the same direction of the velocity, pumping energy of electric origin into the system at a rate equal to twice the system's natural frequency.

Quantitative Description

A trivial solution to the nonlinear Mathieu Eq. 3 is $x = 0$. At $x = 0$, the damping force, $b\dot{x}$, and the inertial force, $m\ddot{x}$, are both zero valued. Thus the spring force, $k_1 x + k_3 x^3$, is only driven by the parametric force F_p Eq. 1, which is generated by a symmetric drive design with no net force applied when the displacement is zero (as illustrated in Fig. 3). However, with air damping and the associated thermomechanical noise, a very small but nonzero displacement on the order of 10 fm/rtHz is present, and as a result, the electrostatically excited parametric force Eq. 1 is no longer zero and is proportional to the displacement. In a second-order mechanical system, this force exerted on a moving mass acts like a spring. And the applied voltage controls the associated "electrostatic spring constant." The total effective spring constant is a combination of both the mechanical spring constant and the electrostatic spring constant.

Upon application of the parametric drive voltage, the total effective spring constant of the mass-spring-damper system periodically increases (hardens) or decreases (softens). Parametric resonance commences when the parametric excitation and the velocity of the resonator are in phase. For principal parametric resonance at $2\omega_r$, the effective spring constant is modified twice within one period of the oscillation by the time-varying pump voltage. The system establishes parametric resonant amplification and its oscillation amplitude grows exponentially. The exponential growth is limited by the mechanical and electrostatic nonlinearities, controllable by the parametric drive voltage. The pumped displacement amplitude and phase are strong functions [19] of the system parameters such as the vibrating mass, mechanical spring constant, and electrostatic spring constant of the parametric drive.



Nonlinear Parametric MEMS, Fig. 3 The waveforms of the displacement, x, velocity, $\dot{x}$, parametric voltage square, $V_{ac}^2 \cos(2\omega_r t)$, and parametric force, F_p, for the conceptual example in Fig. 2

Nonlinear Mathieu Equation and Analytic Solutions

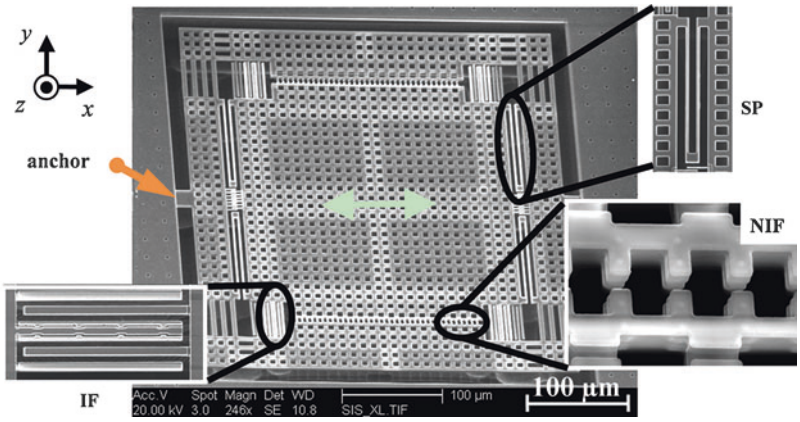
The governing nonlinear Mathieu Eq. 3 is in the form of a second-order differential equation with a cubic-displacement term and an oscillating excitation, whose exact solution is difficult to solve analytically. Perturbation techniques [17] provide a mathematical approach to find an estimated solution. Quantitative descriptions of MEMS parametric resonance have been extensively studied in [18] and [19] by solving Eq. 3 with perturbation techniques. The perturbation approach starts with normalizing Eq. 3 into a general, nondimensionalized form. It then finds the steady-state solutions, i.e., equilibria, using the various mathematical approaches such as the method of averaging and the method of multiple scales [17]. To determine the stability of steady-state solutions, the normalized equation is linearized locally around these equilibria. The system state is allowed to move along these linearized approximations by a small amount to determine the stability of each solution. Stability is often an interesting topic to study as it reveals if the associated steady-state solution is physically realizable.

Nonlinear Mathieu Equation

In electrostatically excited MEMS applications, the Mathieu equation describes a mass-spring-damper system with linear and cubic mechanical stiffness coefficients, driven by an electrostatic force that is not only a function of time but also of the displacement. The nonlinear Mathieu Eq. 3 may be written into a more general form:

$$m\ \ddot{x} + b\ \dot{x} + k_1 x + k_3 x^3 = F_p(t,x) = -\left(\gamma_1 x + \gamma_3 x^3\right)\ V^2. \tag{5}$$

Taking a CMOS-MEMS non-interdigitated comb finger (NIF) resonator described in Fig. 4 as an example, the perforated mass is suspended by four sets of mechanical springs. The sources of damping are usually dominated by the Couette damping of the suspended mass interacting with the underlying substrate, as well as Stokes



Nonlinear Parametric MEMS, Fig. 4 SEM of a CMOS-MEMS nonlinear parametric resonator that moves in x, with zoom-in view of NIF for parametric excitation, interdigitated finger (*IF*) for sensing and self-testing, and the spring (*SP*) for mechanical suspension

damping of the plate, and the squeeze-film damping between the gaps of any interdigitated comb fingers that are used for the capacitive velocity sensing and self-testing. The electrostatic parametric force, F_p, is an odd function of the displacement x. And this force is balanced when the displacement is zero (i.e., $F_p = 0$ when $x = 0$), usually attributed to a symmetric electrode design (i.e., drive combs located on both sides of the resonating mass, as shown on Fig. 4).

Applying a voltage in the form of

$$V(t) = \sqrt{V_{dc}^2 + V_{ac}^2 \cos(\omega_d t)}, \tag{6}$$

across the parametric excitation comb fingers provides a DC forcing term and an AC forcing term with a single frequency component ω_d [28] that decouples the parametric and direct harmonic excitation (i.e., AC term of the parametric force has only one frequency component, which can be used to excite pure parametric resonance).

Applying Eq. 6 on a parametric drive, (e.g., a NIF comb (Fig. 1b) or a shaped-finger comb (Fig. 1c)) generates a nonlinear parametric force in the form of Eq. 1. For the special case of a linear parametric force (i.e., $\gamma_3 = 0$),

$$F_p(t,x) = -\gamma_1 \left[V_{dc}^2 + V_{ac}^2 \cos(\omega_d t)\right] x. \tag{7}$$

Neglecting the mechanical cubic nonlinearity, $k_3 x$, and substituting Eq. 7 into Eq. 5 yield the linear Mathieu equation:

$$m\ \ddot{x} + b\ \dot{x} + \left(k_1 + \gamma_1 V_{dc}^2\right)x = -\gamma_1 V_{ac}^2 \cos(\omega_d t)x. \tag{8}$$

The resonant frequency of the linear Mathieu equation is thus

$$f_r = \frac{1}{2\pi}\sqrt{\frac{k_1 + \gamma_1 V_{dc}^2}{m}}. \tag{9}$$

With the inclusion of the nonlinear terms, γ_3 and k_3, a generalized form of the nonlinear Mathieu equation in Eq. 5 can be derived, which requires special treatments described in the following text.

N

Normalization

In general, perturbation approach starts with a nondimensionalization process, which is used to rewrite the nonlinear Mathieu Eq. 3 into a normalized form so its solutions can be universally applied given the variation in the design parameters:

$$\ddot{s} + \varepsilon\mu\dot{s} + s(\delta + 2\varepsilon\cos 2T) + \varepsilon s^3(\delta_3 + r_3\cos 2T) = 0 \tag{10}$$

where the dimensionless displacement is

$$s = \frac{x}{x_0} \tag{11}$$

and x_0 is the "characteristic displacement" chosen for a specific comb design. For example, x_0 may

be the comb finger pitch (center-to-center distance between electrodes in Fig. 1b) for a NIF resonator or the maximum stroke of a lateral comb resonator (Fig. 1c). The dimensionless time is

$$T = \frac{\omega_d t}{2}, \tag{12}$$

and for principle parametric resonance, $\omega_d = 2\omega_r$. The dimensionless linear parametric electromechanical stiffness, ε, is

$$\varepsilon = \frac{2\gamma_1 {V_{ac}}^2}{m\omega_d^2}. \tag{13a}$$

The dimensionless damping, μ, is

$$\mu = \frac{2b}{\varepsilon m \omega_d}. \tag{13b}$$

The dimensionless linear mechanical stiffness, δ, is

$$\delta = \frac{4\left(k_1 + \gamma_1 {V_{dc}}^2\right)}{m\omega_d^2}. \tag{13c}$$

The dimensionless cubic mechanical stiffness, δ_3, is

$$\delta_3 = \frac{2\left(k_3 + \gamma_3 V_{dc}^2\right)x_0^2}{\gamma_1 V_{ac}^2}. \tag{13d}$$

The dimensionless cubic parametric electromechanical stiffness, r_3, is

$$r_3 = \frac{2\gamma_3 x_0^2}{\gamma_1}. \tag{13e}$$

The relative magnitude of these dimensionless parameters represents their contribution to the parametric response.

Steady-State Solutions

The perturbation method typically starts with finding the steady-state solutions and linearizes the normalized equation about its solutions. An asymptotic expansion deriving the Jacobian matrix of the system is often used to determine the stability of each solution. To arrive at the steady-state solution, an assumed normalized solution

$$s(T) = A(T)\cos\left[\Omega T + B(T)\right] \tag{14}$$

and its normalized time derivative

$$\dot{s}(T) = -A(T)\Omega\sin[\Omega T + B(T)] \tag{15}$$

are imposed to rewrite the normalized Mathieu equation into two coupled equations in the amplitude, A, and phase, B, state-space model. The dimensionless mechanical frequency in Eq. 14 and Eq. 15 is defined as

$$\Omega \triangleq \frac{\omega_d}{2\omega_r}. \tag{16}$$

A combination of single harmonics are derived by plugging Eqs. 14 and 15 into Eq. 10 and rewriting the equation. These harmonics are integrated to keep only the time-invariant components and to discard the fast-varying transients [17–19]. This averaging method yields the first derivative of amplitude and phase:

$$\dot{A} = \frac{1}{8}x_0\,\varepsilon A\left[-4\mu + \left(4 + r_3 A^2\right)\sin\left(2B\right)\right] \tag{17}$$

$$\dot{B} = \frac{1}{8}\varepsilon\left[3A^2\delta_3 - 8\omega_1 + \left(4 + 2r_3 A^2\right)\cos\left(2B\right)\right] \tag{18}$$

where a "detuning parameter," ω_1, is defined such that

$$\Omega = 1 + \varepsilon\omega_1. \tag{19}$$

The quantity $\varepsilon\omega_1$ measures the closeness of the parametric excitation frequency, ω_d, to the predicted parametric resonance occurring around twice the natural frequency $2\omega_r$. The averaged model Eqs. 17 and 18 is used to obtain the steady-state response and the corresponding stability.

In weakly damped systems, one steady-state trivial (zero) and three steady-state nontrivial responses [19] are derived by setting $\dot{A} = 0$ and $\dot{B} = 0$:

$$\alpha_1 = \pm 2x_0\sqrt{\frac{1+2\omega_1}{3\delta_3 - 2r_3}} \tag{20a}$$

$$\beta_1 = \pm\pi\left(n+\frac{1}{2}\right), \text{ where } n = 0,\ 1,\ 2,\ \ldots \tag{20b}$$

$$\alpha_2 = \pm 2x_0\sqrt{\frac{-1+2\omega_1}{3\delta_3 + 2r_3}} \tag{20c}$$

$$\beta_2 = \pm n\pi, \text{ where } n = 0,\ 1,\ 2,\ \ldots \tag{20d}$$

$$\alpha_3 = \pm x_0\sqrt{-\frac{4}{r_3}} \tag{20e}$$

$$\beta_3 = \pm\frac{1}{2}\arccos\left(-\frac{3\delta_3}{r_3} - 2\omega_1\right) \tag{20f}$$

$$\alpha_4 = 0 \tag{20g}$$

$$\beta_4 = \pm\frac{1}{2}\arccos(2\omega_1) \tag{20h}$$

where α_i and β_i are the dimensional steady-state amplitude and phase solutions, respectively. Dimensional steady-state solutions are sometimes written in the amplitude and phase pair $[\alpha_i\ \beta_i]^T$, where $i = 1, 2, 3$, and 4.

Frequency Response

Electrically driven parametric resonance commences around the excitation frequency of $\omega_d \cong 2\omega_r$. With Eqs. 16, 19, and 20a-h, four denormalized (dimensional) mechanical responses are

$$x(t) = \alpha_i \cos\left[\left(1+\varepsilon\omega_{1,\ ij}\right)\omega_r t + \beta_i(t)\right] \tag{21}$$

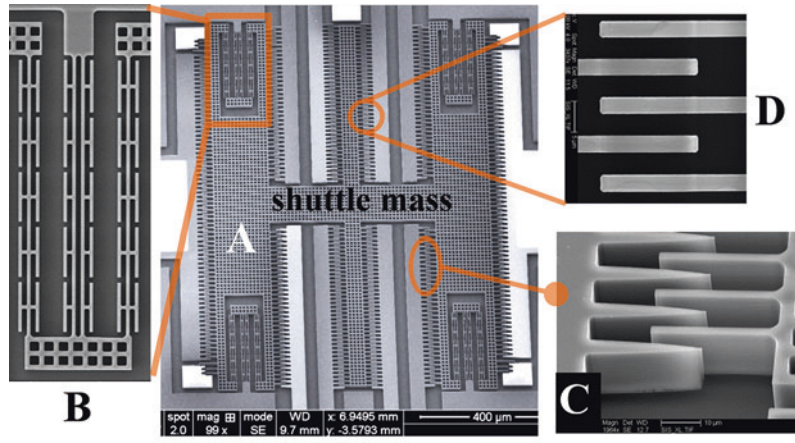
where $i = \{1, 2, 3, 4\}$ denotes the i^{th} steady-state solution and $j = \{1, 2\}$ for $i = \{1, 2, 3\}$ denotes the j^{th} bifurcation frequency, determined by the stability analysis. The stability of the steady-state amplitude in Eq. 20a, c, e, g was extensively studied in [18], which determines the combination of frequency and phase that trigger the bistable jump. The generalized stability analysis uses the determinant of the Jacobian matrix of Eqs. 17 and 18 to derive the "boundaries" between the stable and unstable solution branches. The resulting boundaries are expressed in terms of the "detuning parameter" $\omega_{1,ij}$. In electrostatically driven parametric MEMS, the dimensional electrical excitation frequencies that give rise to bifurcation are

$$f_{\text{elec}} = f_{ij} = 2f_r\left(1+\varepsilon\omega_{1,\ ij}\right). \tag{22}$$

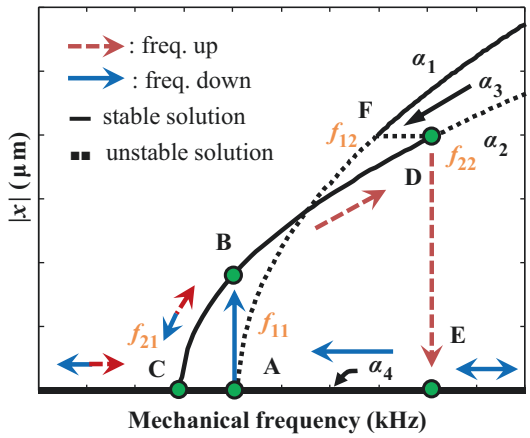
The bifurcation frequencies link to the observable system parameters such as the parametric excitation voltages and frequencies.

Using a fabricated shaped-finger comb parametric resonator [21] shown in Fig. 5 as an example, the shuttle mass (A) suspended by four sets of folded springs (B) is driven parametrically by the shaped-finger comb (C), and its motion is sensed with the straight-finger combs (D). Four denormalized steady-state amplitudes α_i and the corresponding denormalized bifurcation frequencies f_{ij} are plotted in Fig. 6 against the mechanical frequency (the frequency of the displacement).

The dimensional bifurcation frequencies are summarized in Table 1. Each of the solutions α_1 and α_2 has three branches of different stability properties, indicated on the figure with solid (stable branch) and dash (unstable branch) lines. Specifically, the steady-state solutions a_1 and a_2 are separated by the bifurcation frequencies (i.e., points A, F, C, and D) into different regions with well-defined stability. These bifurcation frequencies are labeled f_{ij}. The third solution α_3 is an unstable constant solution, and the fourth solution (zero displacement solution) α_4 is always stable in steady state, but it is not practical between points A and C, because noise will trigger a jump to a_2. The frequencies f_{11} and f_{21} are called "pitchfork" bifurcation frequencies, and f_{12} and f_{22} are called the "saddle-node" bifurcation frequencies. Two "pitchfork" bifurcation frequencies spread out symmetrically on the frequency axis corresponding to $\omega_{1,i1} = \pm 0.5$, where $i = 1,2$.



Nonlinear Parametric MEMS, Fig. 5 SEM of a silicon-on-insulator-MEMS shaped-finger comb parametric resonator suspended by folded springs



Nonlinear Parametric MEMS, Fig. 6 Steady-state frequency response of the folded-flexure shaped-finger comb parametric resonator demonstrating overall hardening nonlinearity

Nonlinear Parametric MEMS, Table 1 The dimensional bifurcation frequencies of the parametric MEMS resonator

Symbol	Physical interpretation
f_{11}	Pitchfork bifurcation from zero solution to the unstable region of α_1
f_{12}	Saddle-node bifurcation of α_1 to α_3
f_{21}	Pitchfork bifurcation where stable branch of α_2 annihilates into the zero solution
f_{22}	Saddle-node bifurcation of α_2 to α_3

In Fig. 6, the nonzero solutions emanating from the bifurcation points f_{11} and f_{21} emerge when the amplitude solutions are real valued.

The sign of the combined nonlinearity $(3\delta_3 \pm 2r_3)$ in Eq. 20a, c determines the direction of the bending. For the shaped-finger comb resonator in Fig. 5, the amplitude solutions a_1 and a_2 bend to the right, indicative of "cubic hardening" nonlinearity. The resonant amplitude is limited by the cubic stiffness nonlinearity. The resonator maintains resonating state along the nonzero solution curves. The type of nonlinearity can vary from spring hardening as in Fig. 6 to mixed nonlinearity [18] and further to spring softening [19].

The parametric frequency response may be measured by chirping (i.e., sweeping) the excitation frequency with time in the region near twice the natural frequency. For most values of the drive frequency, the system is not stimulated and the motion is near $x = 0$. After passing the bifurcation frequency, the system establishes automatic parametric resonance, which builds oscillation amplitude that would grow without bound if not for damping and nonlinearity in the system.

Hysteresis and Bifurcation Behavior

Figure 6 also reveals another aspect of parametric response – hysteresis effect. To the right of the "pitchfork" bifurcation point f_{21}, the "bending" of

the curve results in multivalued solutions at the same excitation frequency. Depending on the initial condition, the response amplitude could home on one of the three solution branches α_1, α_2, and α_4. The system always lands on the stable branches, whereas the unstable branches are not physically realizable in steady state.

In an open loop system, for a decreasing frequency sweep, the resonator displacement follows zero solution with a very small motion created by the thermomechanical noise, until reaching the bifurcation frequency f_{11} of solution α_1 where the stable zero solution intersects with the parametric unstable region of α_1 (point A). At point A, solution α_1 is unstable, while α_2 is stable. With the presence of small Brownian motion, the system jumps from the zero solution to the stable region of α_2 (blue line from A to B) as soon as its trajectory passes through the bifurcation frequency f_{11}. The jump amplitude a_{tran} is defined as the amplitude difference between B and A. At the "pitchfork" bifurcation A, $\omega_{1,11} = -0.5$. Corresponding to the frequency down-sweep, the bifurcation frequency and the jump amplitude are

$$f_{11} = 2f_r\left(1+\varepsilon\omega_{1,11}\right) \cong 2f_{r0}\sqrt{1+\frac{\gamma_1 V_{\text{dc}}^2}{k_1}}\left(1-\frac{\gamma_1 V_{\text{ac}}^2}{4k_1}\right) \tag{23}$$

$$a_{\text{tran}} = a_2|_{f=f_{11}} = 2\sqrt{\frac{\gamma_1 V_{\text{ac}}^2}{-\left(3k_3+3\gamma_3 V_{\text{dc}}^2+2\gamma_3 V_{\text{ac}}^2\right)}} \tag{24}$$

where $f_{r0} = (1/2\pi)\sqrt{k_1/m}$ is the natural frequency and the rest of the terms are independent of mass m. In a practical nonlinear system, the magnitude of the perturbation parameter ε is very small. Thus the exact bifurcation frequency Eq. 23 is very close to twice the natural frequency, providing guidance to experimental search of the bifurcation set points. In Eq. 23, V_{dc} provides an electrostatic tuning effect that shifts the location of the bifurcation. Both the location of the bifurcation and the width of the nonzero solution region can be modified by V_{ac}. In the case of extremely small V_{ac}, solutions α_1 and α_2 pinch and converge, making it difficult to search for the bifurcation because of the very small jump step. After the sharp jump from A to B, the amplitude gradually decreases to zero by following the stable branch of the solution a_2 from B to the "pitchfork" bifurcation point C.

On the other hand, for a slowly increasing frequency sweep, parametric resonance builds up after C. The displacement follows the stable solution branch of α_2 and keeps increasing until meeting a "saddle-node bifurcation" point D, where the stable and unstable branches of solution α_2 meet. The amplitude instantaneously drops from point D to the zero solution (point E). Corresponding to the frequency up-sweep, the bifurcation frequency is

$$f_{22} = 2f_{r0}\sqrt{1+\frac{\gamma_1 V_{\text{dc}}^2}{k_1}}\left[1-\frac{\gamma_1 V_{\text{ac}}^2}{4k_1}-\frac{3\left(k_3+\gamma_3 V_{\text{dc}}^2\right)\gamma_1}{4k_1\gamma_3}\right]. \tag{25}$$

Other than the natural frequency, f_{r0}, all other terms in Eq. 25 are independent of mass. The bifurcation frequency and the resonance amplitude depend heavily on the combined linearity and nonlinearity inherent in the nonlinear MEMS and are both tunable with applied parametric excitation voltages. Measuring shifts in the bifurcation frequencies Eq. 23 and Eq. 25 detects very small changes in resonant mass. Instantaneous "jump" transition from near-zero to high displacement (or the other way around) at the bifurcation frequencies improves the sensitivity of the associated parametric mass detector.

Cross-References

- ▶ CMOS MEMS Fabrication Technologies
- ▶ MEMS (Micro-electro-Mechanical Systems)
- ▶ Nonlinear and Parametric NEMS Resonators
- ▶ Piezoresistivity
- ▶ Sensors

References

1. Mathieu, E.: Memoire sur le mouvement vibratoire d'une membrane de forme elliptique. J. Math. Pures et Appl. **13**, 137–203 (1868)
2. Rayleigh, L.: On the maintenance of vibrations by forces of double frequency, and on the propagation of waves through a medium endowed with a periodic structure. Lond. Edinb. Dublin Philos. Mag. J. Sci. (Fifth Series) **24**(147), 145–159 (1887)
3. Stoker, J.J.: Nonlinear Vibrations in Mechanical and Electrical Systems. Wiley, New York (1950)
4. Nayfeh, A.H., Mook, D.T.: Nonlinear Oscillations. John Wiley & sons, New York (1979)
5. Zhang, W., Turner, K.: Application of parametric resonance amplification in a single-crystal silicon micro-oscillator based mass sensor. Sensors Actuators A Phys. **122**(1), 23–30 (2005)
6. Li, L., Holthoff, E., Shaw, L., Burgner, C., Turner, K.: Noise squeezing controlled parametric bifurcation tracking of MIP-coated microbeam MEMS sensor for TNT explosive gas sensing. J. Microelectromech. Syst. (2014). doi:10.1109/JMEMS.2014.2310206
7. Kacem, N., Hentz, S., Baguet, S., Dufour, R.: Forced large amplitude periodic vibrations of non-linear Mathieu resonators for microgyroscope applications. Int. J. Non Linear Mech. **46**(10), 1347–1355 (2011)
8. Turner, K., Miller, S., Hartwell, P., MacDonald, N., Strogatz, S., Adams, S.: Five parametric resonances in a microelectromechanical system. Nature **396**(6707), 149–152 (1998)
9. Guo, C., Fedder, G.: Bi-state control of parametric resonance. Appl. Phys. Lett. **103**(18), 183512 (2013)
10. Ataman, C., Urey, H.: Modeling and characterization of comb-actuated resonant microscanners. J. Micromech. Microeng. **16**(1), 9–16 (2006)
11. Yie, Z., Zielke, M., Burgner, C., Turner, K.: Comparison of parametric and linear mass detection in the presence of detection noise. J. Micromech. Microeng. **21**, 025027 (2011)
12. Requa, M.: Parametric resonance in microcantilevers for applications in mass sensing. PhD dissertation, University of California, Santa Barbara (2006)
13. Prakash, G., Raman, A., Rhoads, J., Reifenberger, R.: Parametric noise squeezing and parametric resonance of microcantilevers in air and liquid environments. Rev. Sci. Instrum. **83**(6), 065109 (2012)
14. Thompson, M., Horsley, D.: Lorentz force MEMS magnetometer. Hilton Head Workshop 2010: A Solid-State Sensors, Actuators and Microsystems Workshop, pp. 45–48 (2010)
15. Koskenvuori, M., Tittonen, I.: GHz-range FSK-reception with microelectromechanical resonators. Sensors Actuators A Phys. **142**(1), 346–351 (2008)
16. DeMartini, B., Butterfield, H., Moehlis, J., Turner, K.: Chaos for a microelectromechanical oscillator governed by the nonlinear Mathieu equation. J. Microelectromech. Syst. **16**(6), 1314–1323 (2007)
17. Nayfeh, A.: Introduction to Perturbation Technique, pp. 234–256. Wiley, New York (1981)
18. Rhoads, J., Shaw, S., Turner, K., Moehlis, J., DeMartini, B., Zhang, W.: Generalized parametric resonance in electrostatically actuated microelectromechanical oscillators. J. Sound Vib. **296**(4–5), 797–829 (2006)
19. Guo, C., Fedder, G.: Behavioral modeling of a CMOS-MEMS nonlinear parametric resonator. J. Microelectromech. Syst. **22**(6), 1447–1457 (2013)
20. Adams, S., Bertsch, F., MacDonald, N.: Independent tuning of linear and nonlinear stiffness coefficients. J. Microelectromech. Syst. **7**(2), 172–180 (1998)
21. Guo, C., Fedder, G.: A quadratic-shaped-finger comb parametric resonator. J. Micromech. Microeng. **23**(9), 095007 (2013)
22. Hirano, T., Furuhata, T., Gabriel, K., Fujita, H.: Design, fabrication, and operation of submicron gap comb-drive microactuators. J. Microelectromech. Syst. **1**(1), 52–59 (1992)
23. Jensen, B., Mutlu, S., Miller, S., Kurabayashi, K., Allen, J.: Shaped comb fingers for tailored electromechanical restoring force. J. Microelectromech. Syst. **12**(3), 373–383 (2003)
24. Kaajakari, V., Lal, A.: Parametric excitation of circular micromachined polycrystalline silicon disks. Appl. Phys. Lett. **85**(17), 3923–3925 (2004)
25. Requa, M., Turner, K.: Electromechanically driven and sensed parametric resonance in silicon microcantilevers. Appl. Phys. Lett. **88**(26), 263508 (2006)
26. Rhoads, J., Kumar, V., Shaw, S., Turner, K.: The non-linear dynamics of electromagnetically actuated microbeam resonators with purely parametric excitations. Int. J. Non Linear Mech. **55**, 79–89 (2013)
27. Zalalutdinov, M., Olkhovets, A., Zehnder, A., Ilic, B., Czaplewski, D., Craighead, H., Parpia, J.: Optically pumped parametric amplification for micromechanical oscillators. Appl. Phys. Lett. **78**(20), 3142 (2001)
28. Rhoads, J., Shaw, S., Turner, K.: Nonlinear dynamics and its applications in micro- and nanoresonators. J. Dyn. Syst. Meas. Control. **132**(3), 034001 (2010)

Nonlinearity

▶ Nonlinear and Parametric NEMS Resonators

NS300

▶ Nanoparticle Tracking Analysis

NS500

► Nanoparticle Tracking Analysis

NTA

► Nanoparticle Tracking Analysis

Nucleation of Silica Nanowires

► Growth of Silica Nanowires

Nucleic Acid Amplification

► Microarray
► Nanotechnology Applications in Polymerase Chain Reaction (PCR)

O

Officinal Skink

► Friction-Reducing Sandfish Skin

Oil-Repellency of Fish Scales

► Superoleophobicity of Fish Scales

Oligonucleotide Amplification

► Nanotechnology Applications in Polymerase Chain Reaction (PCR)

Oligonucleotide Synthesis

► Microarray

OMFR

► Microfluidic Optomechanics

On-Chip Flow Cytometry

Shady Gawad[1], Ana Valero[2], Thomas Braschler[3], David Holmes[4] and Philippe Renaud[2]
[1]MEAS Switzerland, Bevaix, Switzerland
[2]Microsystems Laboratory, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland
[3]Laboratory of Stem Cell Dynamics, SV-EPFL, Lausanne, Switzerland
[4]London Centre for Nanotechnology, University College London, London, UK

Synonyms

Micro coulter; Micro FACS; Micro flow cytometry

On-Chip Flow Cytometry

Flow cytometry is an analysis technique used to discriminate cells and other microscopic particles suspended in a liquid. Cells pass in rapid succession through a detection region where their properties are measured individually. Embedding such a system on the surface of a chip using microfluidics allows one to handle smaller sample volumes and less reagents (typically picoliter to nanoliter). The sensing is generally performed by optical or electrical detection means, preferably with micro-optic elements or microelectrodes placed around or within the fluidic channels.

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

On-chip flow cytometers may also be co-integrated with a cell-sorting module to obtain purified sample subpopulations, or simply be used to enumerate cell types.

Overview

Coulter counters and fluorescence-activated cell sorters provide rapid high-quality cell analysis and sorting, but are traditionally relatively expensive laboratory-scale machines, requiring corresponding operator expertise and real estate. In view of the emergence of an ever greater number of point-of-care (POC) applications and integration of complicated protocols into easy-to-handle devices, a considerable effort has been made to reduce the footprint of cell analyzers and sorters to chip size. This has given rise to the field of micro-flow cytometry.

Introduction

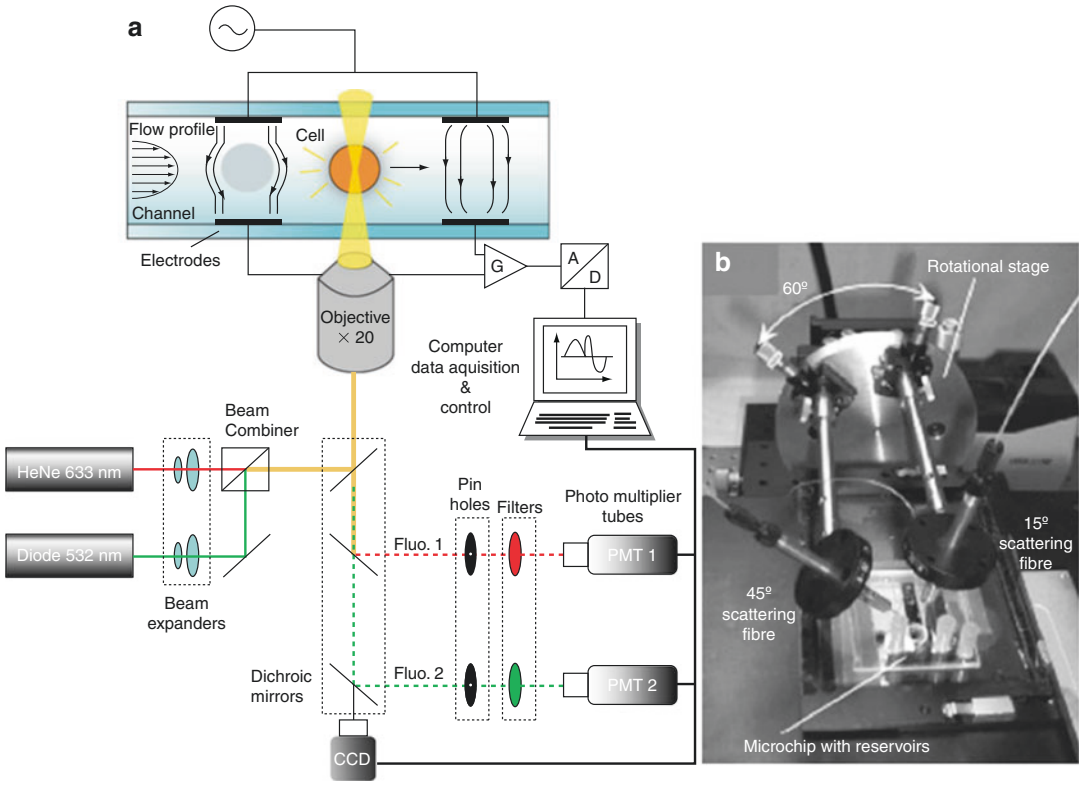
A number of on-chip impedance and dielectrophoretic force spectroscopy as well as optical techniques are presented here. Some of these techniques are currently being developed for use in point-of-care (POC) diagnostic applications and cell counting and separation devices for use in hematology, oncology, and toxicology. Cellular properties such as cell antigen expression, cell size, cell-cycle progression, and apoptosis can all be measured. To detect these properties, flow cytometry can either make use of specific labels (e.g., fluorescently conjugated antibodies) or assess intrinsic cell properties such as light scatter or impedance. This latter case has become an area of intense research in recent years with a number of label-free techniques being realized through the use of novel microfluidic technologies. Label-based analysis is currently the gold standard due to the high specificity and sensitivity achieved through the use of fluorescent probes coupled with molecular recognition elements (monoclonal antibodies, DNA intercalators). For POC applications, the major drawbacks of the cell labeling procedures are the requirement for multiple off-chip preprocessing steps and the extra cost of the label.

Electrical techniques, such as dielectrophoresis or impedance analysis, are generally used in settings where cell labeling is either not needed, or incompatible with the application. Major examples include cell sizing and discrimination, viability or osmotic response measurement in toxicity, and drug response screening as well as detection of intracellular infectious agents such as malaria.

Electrical and optical measurements can be combined onto a single platform thus extending the multiparameter analysis capabilities of such devices and allow powerful new analysis paradigms. Such combined measurements have been used for many years in hematology analyzers but have only recently been demonstrated in microfluidic cytometers, where they are expected to lead to the development of highly refined analysis and sorting instruments. Validation that the impedance and dielectric properties of cells correlate with cell type has been demonstrated through the use of such simultaneous electro-optical measurements.

Optical Detection

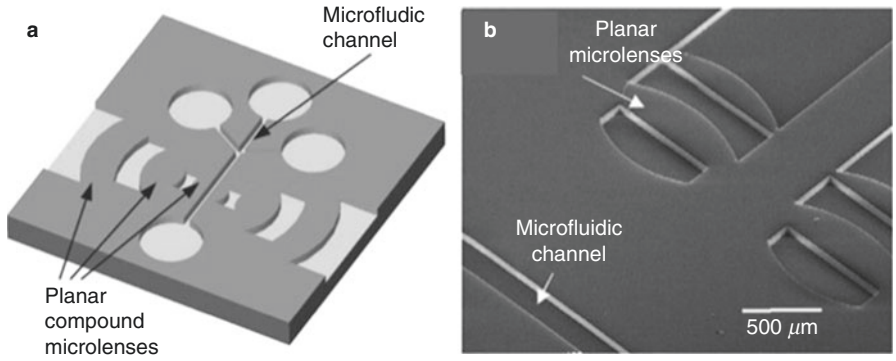
Commercial flow cytometers utilize bulky free-space optical components and allow interrogation, detection, and identification of cells and other particles (e.g., polymer beads, viruses) based on both fluorescence and light-scattering criteria. Such "traditional" optical detection techniques have been widely used with microfluidic devices [1, 2]. However, optical alignment is a problem, with such systems requiring mechanically stable mounting of the microfluidic device and the optical components. Optical benches are typically used to allow and maintain accurate alignment of the light paths. Figure 1a shows schematically the optical elements (e.g., lasers, filters, mirrors, lenses, and photo detectors) in a three-color fluoresence confocal detection setup, as used for the simultaneous optical and impedance detection of cells flowing in a microfluidic impedance cytometer. Figure 1b shows a photograph of a multi-angle light scatter detection setup, using optical fibers mounted at a number of fixed angles,



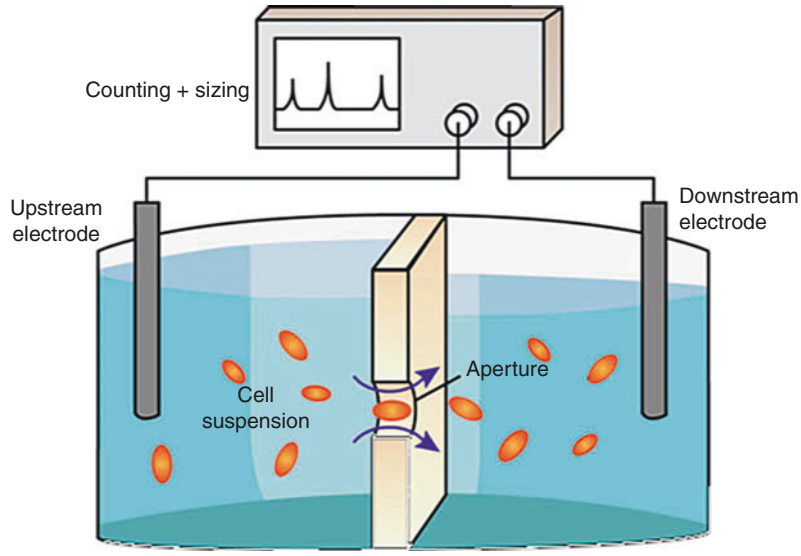
On-Chip Flow Cytometry, Fig. 1 Examples of external optical detection setups used with microfluidic flow cytometers. (**a**) Schematic of a combined impedance and optical detection system as used for blood analysis. (**b**) The use of optical fibers for multi-angle light scatter measurements from a microfluidic chip [2]

allowing analysis of the scatter properties of individual particles flowing in liquid suspension along a glass capillary [2].

The introduction of integrated optical elements into microfluidic chips is essential for the realization of robust, portable, and cheap handheld diagnostic devices. Integrated optics allows a reduction in the size, complexity, and cost of microfluidic devices. Optical alignment issues are avoided as these are now carried out at the design stage. The optical components are produced using the same microfabrication and lithographic techniques used to form the microfluidic channels. Waveguides can be used to precisely deliver excitation light and collect both scattering and fluorescence. One common method of incorporating integrated waveguides into microfabricated flow cytometers is to insert optical fibers into prefabricated aligned groves [3]. One end of the fiber is connected to the light source or detector, and the other end can be inserted directly into the device. Such setups are efficient at guiding light with little attenuation; however, they typically use commercially available optical fibers that impose geometrical limitations on the device and also limit their focusing and light collection capabilities. Photolithography has been used to make well-defined waveguides and lenses out of optically transparent photoresists such as SU-8 [1]. Buried waveguides have been fabricated directly onto silicon wafers in a variety of materials, such as silicon dioxide [2] and doped glassess [3]. These integrated waveguides typically have an attenuation at least five orders of magnitude greater than that of optical fibers necessitating their use in short segments. Figure 2 shows an example of a planar integrated optical system formed in SU-8. The optical elements and microfluidic channel were produced using photolithography.



On-Chip Flow Cytometry, Fig. 2 Planar integrated optics incorporating waveguides and micro-lenses have been used to allow detection of samples flowing in a microfluidic channel. (**a**) shows a schematic of one such device and (**b**) shows a fabricated structure prior to bonding of the lid (Used with permission from Seo et al. [4])



On-Chip Flow Cytometry, Fig. 3 Coulter counter principle. Cells are drawn through a small aperture, modifying the current as they pass, allowing them to be accurately counted and sized

Electrical Detection

The Coulter method of sizing and counting particles is based on measurable changes in electrical resistance produced by the cells or particles passing through a small aperture (see Fig. 3). The current, induced through two large electrodes placed upstream and downstream, is concentrated within the aperture representing the sensing area. The measured voltage pulse height is proportional to the volume of the particle. Several thousand particles per second can be individually counted and sized. In addition, a metering device is used to draw a known volume of the particle suspension through the aperture; a count of the number of pulses can then yield the concentration of particles in the sample.

On-Chip Examples

The first microCoulter devices were presented by Larsen et al. [5] and Koch et al. [6, 7]. These systems were microfabricated in silicon and use microfluidic sheath flows and filter structures. The first reports of on-chip cell measurements were published later by Fuller et al. on granulocytes [8], Larsen et al. on somatic cells using a simple silicon aperture [9], and Gawad et al. on erythrocytes using a coplanar electrode geometry [10]. Historically, the microCoulter technique

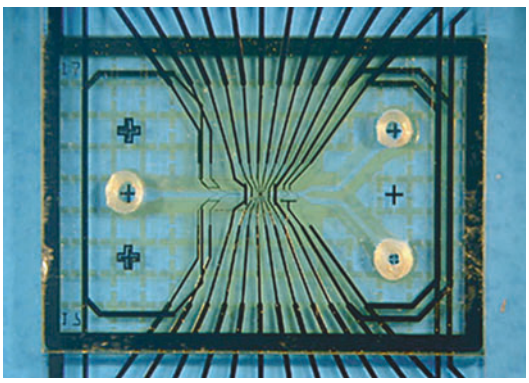
carries out cell sizing at a single measurement frequency, generally at low frequency or DC. Measurements at multiple frequencies in the kHz and MHz range enhance detection of additional attributes of the cell, such as membrane capacitance and intracellular conductivity.

Chip Fabrication

On-chip integration of microfluidics and microelectrode patterns can be produced using relatively simple lithography techniques on a glass wafer. For example, the chip shown in Fig. 4 is fabricated as follows: a lift-off process is used to define metal electrodes, the microfluidic network is then produced on top of the electrodes in polymer [11]. Individual chips are released from the wafer by dicing and reversibly sealed by a molded PDMS block containing fluidic ports or embedded liquid reservoirs. More complex techniques to achieve electrodes patterned on top and bottom of a microchannel have also been presented in order to produce localized electric field (Fig. 4) [12, 13]. The full device is mounted into a custom-made fluidic and electrical interface. The thickness of the polymer layer defines the height the channels and can be adjusted to according to the necessities of each application; for cell work channel dimensions are typically 20–40 μm.

Cell Impedance Spectroscopy

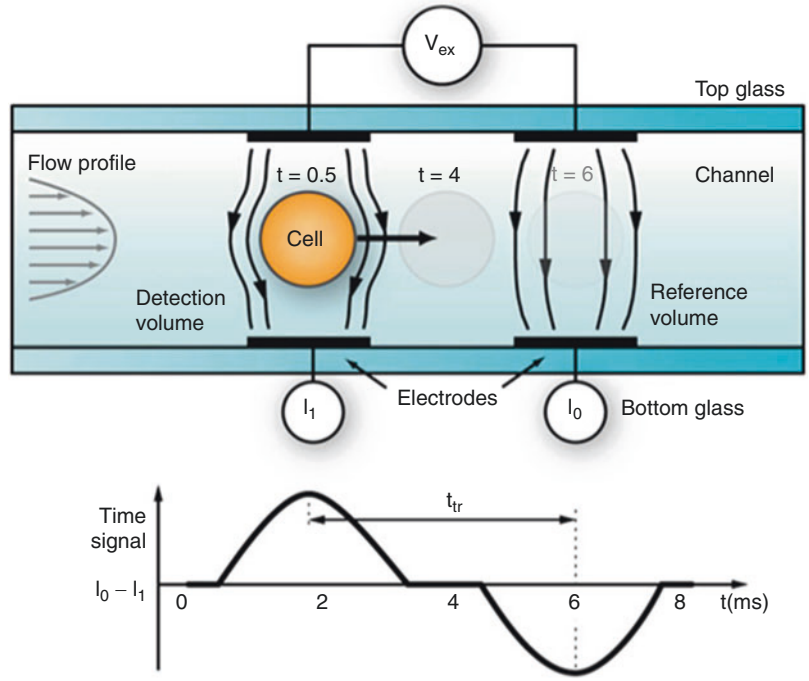
For cell impedance spectroscopy, the excitation frequency range is typically from DC to about 50 MHz. At DC or low frequency the induced ionic current flows around the cell giving a measure of cell size; at higher frequencies (1–10 MHz) the membrane polarizes in response to the applied field giving a measure of the cell membrane properties, at yet higher frequencies (>30 MHz) the current crosses the membrane such that the intracellular properties of the cell are reflected in the current measurement.



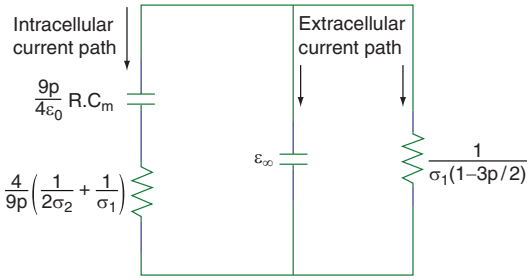
On-Chip Flow Cytometry, Fig. 4 Microfabricated fluidic chip with embedded platinum microelectrodes for dielectric spectroscopy and cell sorting

Collecting significant information using this technique in a microfludic channel has been performed by using micrometer-sized electrodes, slightly larger than the cell itself, patterned on the channel walls. One set of electrodes is used to apply an electric field in a section of the channel while a separate set of electrodes, generally placed on the opposite side of the channel measure the current continuously. As the cell passes between the electrodes a current pulse is measured (Fig. 5). A differential measurement is typically used with one pair of electrodes measuring the current change due to the passing cell and the second pair of electrodes measure the suspending media. Subtracting these two measurements considerably eases further current amplification and allows for rejection of common mode interference. Furthermore, shielding, grounding, and band pass filtering using lock-in techniques are necessary to yield high sensitivity. For impedance spectroscopy, two or more excitation frequencies may be applied simultaneously in order to sense each cell at the different frequencies.

An electrical model of the cell impedance is very useful for understanding the impedance response function. A simple equivalent electric circuit was proposed by Schwan [14] for a single shell model representing the cell membrane capacitance and cytoplasm using a parallel and series arrangement of capacitors and resistors (Fig. 6). This circuit gives a simple representation of the behavior of this type of particle: For excitation frequencies lower than the membrane characteristic frequency, the current lines are surrounding an essentially insulating particle; at higher frequencies, the current gradually enters the cytoplasm as the membrane capacitance is considered as shorted. The left path in the figure represents the current path through the cell



On-Chip Flow Cytometry, Fig. 5 Impedance detection principle and the resulting differential current signal from a passing cell. The detection and reference volumes are inherently switched as the cell passes through each. The cell speed in the measurement channel is computed by measuring the transit time t_{tr} separating the two peaks. The bottom electrode is kept at a virtual ground potential



On-Chip Flow Cytometry, Fig. 6 Equivalent circuit for the Maxwell–Wagner dispersion due to RC charging of the cell membrane. The parameters σ_2 conductivity of the cytoplasm (Sm^{-1}), σ_1 conductivity of the extracellular fluid, Cm specific membrane capacitance (Fm^{-2}), *R* cell radius (m), p = volume fraction of cells in suspension [10]

interior. A more detailed explanation on the circuit models has been published by the authors [15].

This model was originally developed for analysis of bulk cell suspensions but is also valid in the case of a single cell measurement, with the added advantage that the measured particle volume is directly expressed by the single particle volume to detection volume fraction rather than simply representing an average volume of an unknown number of particles.

Cell position within the electric field does also have an important influence on the measured pulse shape and magnitude. By comparing impedance values at different frequencies, an opacity value independent of cell position can be obtained, giving access to some cell properties. Accurate cell sizing requires alignment of the cells on a given flow line, which can be achieved through the use of sheath flows. Microfluidics also allow the use of novel particle focusing modalities such as inertial forces (arising from Dean flows), and dielectrophoretic, ultrasonic, optical, and magnetic forces, which have all been demonstrated.

Cell Dielectrophoresis

For on-chip flow cytometry, cell manipulation in liquids is of primary importance. Cell alignment and cell sorting are preeminent examples. A versatile and established on-chip cell manipulation method is dielectrophoresis.

A dielectric particle such as a cell gets polarized when subjected to an electric field due to an induced dipole moment. When the induced dipole moment of a cell interacts with a nonuniform

electric field, a net movement of the cell is produced. If at a specific frequency, the cell is more polarizable than the surrounding medium, the cell will be attracted toward regions of high field strength (positive dielectrophoresis, pDEP). Conversely, the cell will be repelled from the high electric field region if it is less polarizable than the medium (negative dielectrophoresis, nDEP).

The mathematical expression for the dielectric force for a spherical particle is given by:

$$F_{\mathrm{DEP}} = \langle (P \cdot \nabla) E \rangle = \frac{3\pi}{2} \varepsilon_m V \cdot \mathrm{Re}(f_{\mathrm{CM}}) \nabla |E_{\mathrm{rms}}|^2 \quad (1)$$

where P is the dipole moment, V is the volume of the cell with a radius r, ε_m is the permittivity of the medium, f_{CM} is the Clausius–Mossotti factor. The Clausius–Mossotti factor quantifies the dipole moment induced in a suspended particle relative to the dipole moment that would be induced in an equivalent volume of suspension medium:

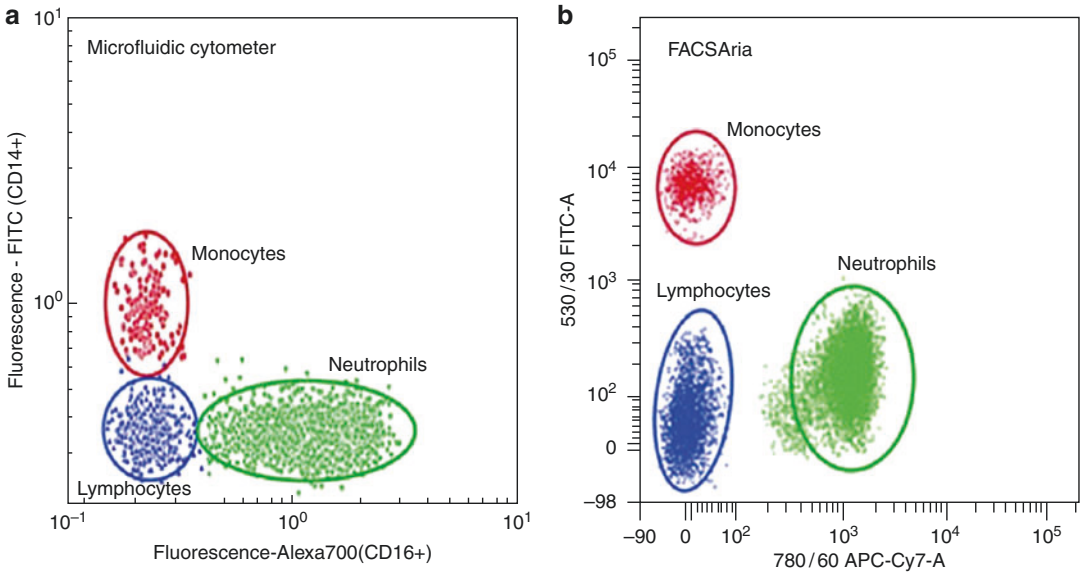
$$f_{\mathrm{CM}}(\omega) = \frac{\tilde{\varepsilon}_p - \tilde{\varepsilon}_m}{\tilde{\varepsilon}_p + 2 \cdot \tilde{\varepsilon}_m} \quad (2)$$

where $\tilde{\varepsilon}_p$ and $\tilde{\varepsilon}_m$ are the complex permittivity of the particles or cells and the medium respectively. Positive DEP will occur for positive values of $\mathrm{Re}(f_{\mathrm{CM}})$, nDEP for negative values, whereas a zero value of (f_{CM}) characterizes the crossover frequency.

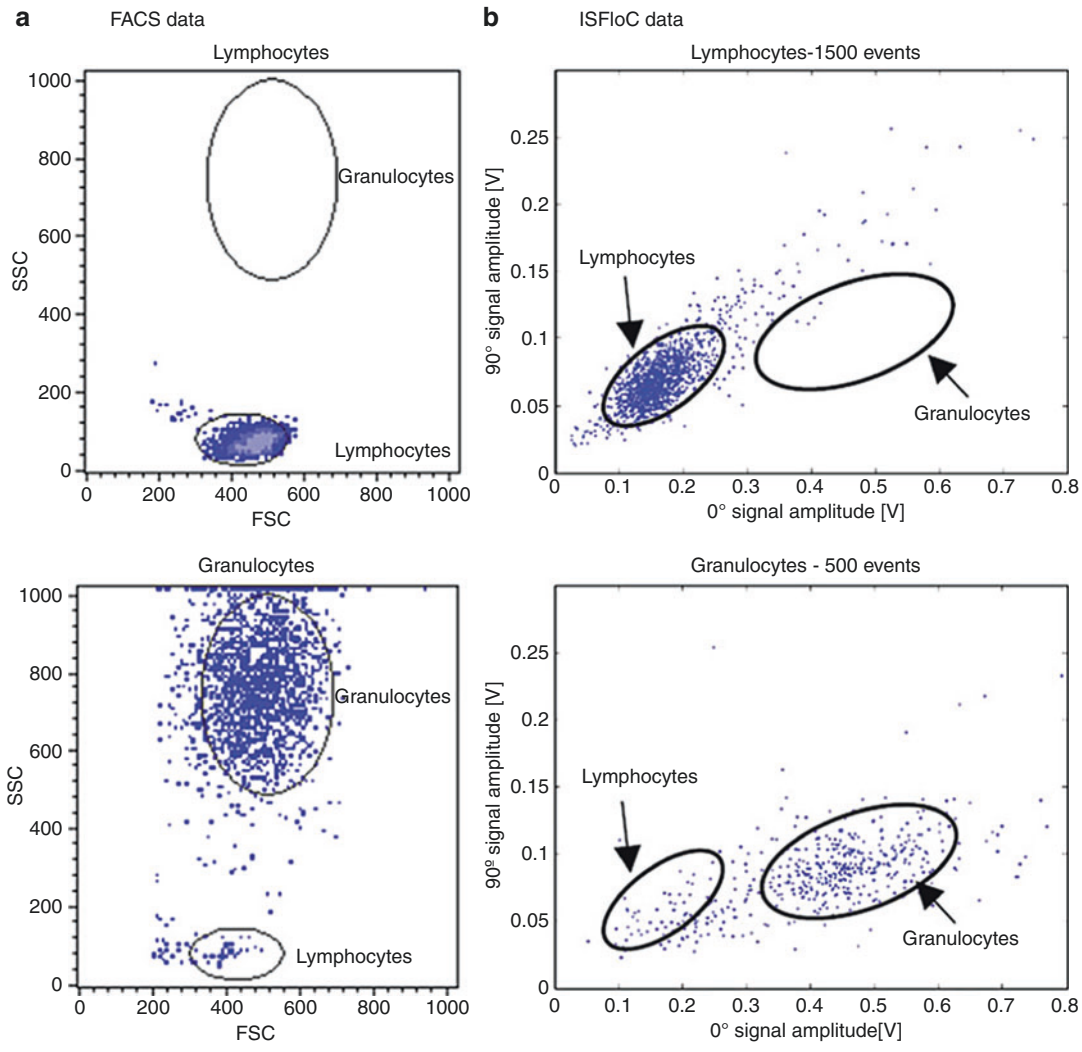
Applications

On-Chip Optical Cell Analysis

Optical detection has been used widely for detection and analysis in lab-on-a-chip devices. A number of highly sensitive flow cytometer devices have been demonstrated in the literature. Figure 7 shows typical data from one such device and compares the data obtained from the microfluidic device with that from a commercial flow cytometer (BD FACSAria). Whole blood was fluorescently labeled (anti-CD14-FITC and anti-CD16-Alexa700) and treated with saponin/formic acid to lyse erythrocytes, this sample was then analyzed using both the microfluidic chip device and a commercial flow cytometer. All the events were triggered on the forward scatter signal, cell populations identified and colored



On-Chip Flow Cytometry, Fig. 7 Scatter plots showing data obtained from blood samples measured using (**a**) microfluidic chip flow cytometer and (**b**) commercial flow cytometer FACSAria

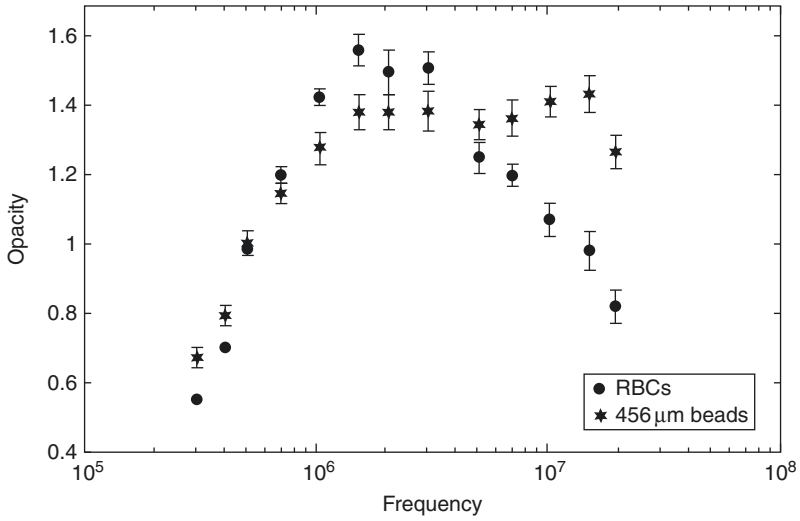


On-Chip Flow Cytometry, Fig. 8 Scatter plots of FACS sorted lymphocytes (*top row*) and granulocytes (*bottom row*). (**a**) Forward (*FSC*) vs. side scatter (*SSC*) data as measured by a FACS instrument. The population of granulocytes is relatively smaller and contaminated with lymphocytes. (FACS measurement courtesy of P. Batard) (**b**) The same populations measured by on-chip impedance cytometry at 1.7 MHz showing the in-phase and out-of-phase signal

according to their fluorescence signal. The data show that both systems can clearly resolve the three major cell populations found in human blood (i.e., neutrophils, monocytes, and lymphocytes).

On-Chip Cell Impedance Spectroscopy

One of the most common applications of cytometry is in hematology analysis for the counting of white blood cells is of particular interest. Blood counts are routinely used to inform in clinical diagnostics and are performed by a range of diagnostic instruments. The example, shown in Fig. 8, makes use of the impedance cytometry chip shown in Fig. 4; human white blood cells are presorted with a FACS instrument (BD, FACSVantage SE) based on fluorescence and scattering. The sorted lymphocyte and granulocyte populations were measured a second time both in the FACS and in the microfabricated



On-Chip Flow Cytometry, Fig. 9 Opacity measurement of RBCs and polystyrene beads in the 300 kHz–20 MHz range. The RBCs show a ®-dispersion near 3 MHz, while the beads have a flat spectrum

cytometer. The granulocyte sample still contains a small amount of the more numerous lymphocytes. Differences in the measured electrical properties between those two cell populations are quite significant and allow clear discrimination using amplitude and phase information retrieved at a single frequency (see Fig. 8).

The above measurement is relatively easy due to the fact that the two cell populations are quite different in size. In order to differentiate between different types of particles of the similar size it is necessary to use several frequencies simultaneously. An example of this is shown below where red blood cells 4–6 µm in diameter are measured along with polystyrene beads of similar dimensions.

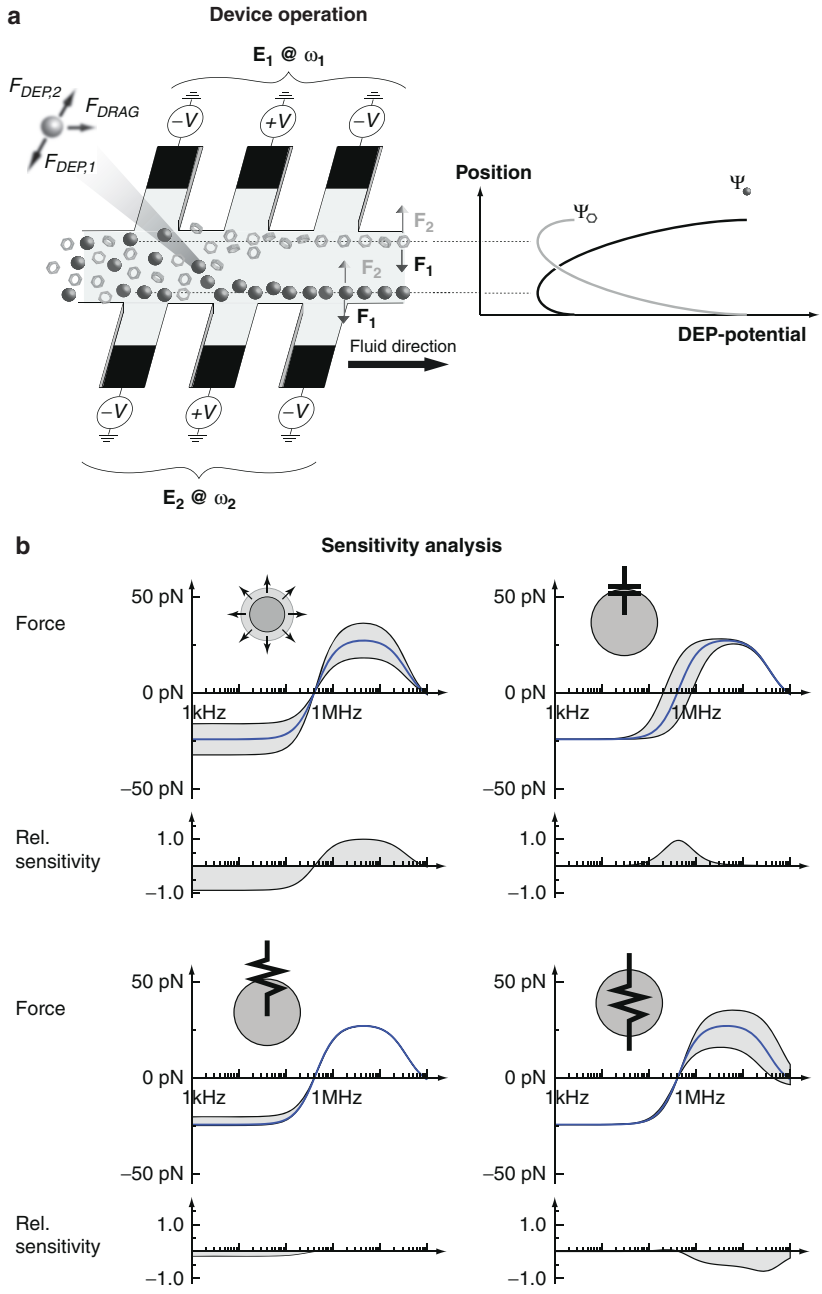
The measured opacity (ratio of high-frequency signal to a fixed low-frequency signal) of a specific cell population is determined using a population averaged frequency sweep technique where two discrete frequencies are employed simultaneously. The low reference frequency is kept constant (500 kHz) while the second frequency is used for the sweep. The ratio of the acquired data pulse height is taken for those two frequencies. This method efficiently suppresses size and position dependence of the measured data. The data shown in Fig. 9 shows a clear difference in the mid- to high-frequency range between the polystyrene beads and the red blood cells (RBCs).

Dielectrophoretic Sorting Device

A sorting device co-integrated on the same chip as one of the above examples could use the acquired real-time data that could be used to sort selected cells to a separated outlet port on the chip. Simple dielectrophoretic sorting devices can be used to modify the trajectory of a cell or particle in flow.

The following application example shows how dielectrophoretic forces may also be used to sort and characterize cells based on cell-cycle stage. The continuous cell-sorting microdevice is based on the opposition of a combination of several dielectrophoretic forces at multiple frequencies that discriminates between cell types according to their dielectric properties, such as the membrane permittivity and the cytoplasm conductivity.

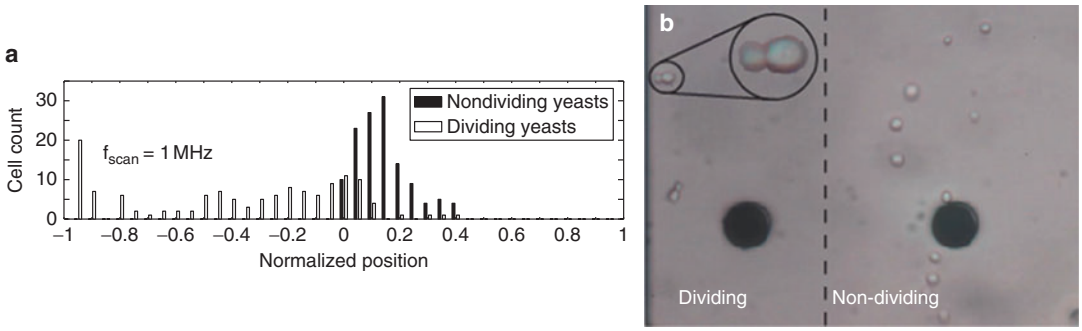
Electric fields are generated by an array of electrodes located in both sidewalls of a main flow channel (Fig. 10a). Cells with different dielectric response experience different force magnitudes and are therefore continuously focused to different equilibrium positions in the flow channel, thus allowing label-free sorting. For cell-sorting purposes, two low-frequency signals (in the range 50–150 kHz), one from each side of the flow channel, focus the cells toward the channel midline, where the force potential minimum is located. For the actual separation, a high-frequency signal (typically in the MHz range) is



On-Chip Flow Cytometry, Fig. 10 (**a**) Principle of the cell-sorting device using multifrequency dielectrophoresis; the illustration shows a case where signals at two frequencies applied from the two liquid electrode arrays are sufficient to differentiate the cells. In many practical cases, including sorting of infected and noninfected red blood cells, three different frequencies are used so that on one side a linear combination of signals at different frequencies is acting on the particles. This configuration of signals takes advantage of the sensitivity of the high-frequency signals while ensuring a total nDEP force which stops cells from being drawn into the electrodes. (**b**) Influence on the different model parameters (1. volume fraction, 2. membrane conductivity, 3. membrane capacitance, 4. cytosol resistivity) on the dielectrophoretic force

superimposed from one of these electrode arrays. The strategy uses the same inherent electrical properties as those that give rise to the impedance differences measured in cells as described above: At low frequency, all cells experience a similar, repulsive force, so that they can be focused toward the channel midline. The high-frequency signal probes cell membrane capacitance and cell interior, and therefore shifts the equilibrium position for cells of different dielectric responses.

Like impedance spectroscopy, the dielectrophoretic force is sensitive to different cellular changes in different frequency windows (Fig. 10b). Since both techniques are



On-Chip Flow Cytometry, Fig. 11 Separation of yeast cells in division. The 1 MHz signal exerts relatively stronger pDEP on cells in division compared to normal yeast cells due to the lower crossover frequency. A subpopulation of cells in division is more attracted to the *left side* than the average, probably related to the stage in the cell division cycle

fundamentally dependent on the Clausius–Mossotti factor, they are generally sensitive to identical changes in similar frequency windows.

Sorting of biological samples using dielectrophoresis is generally based on changes in cell physiology due to the conductivity gradient between intracellular and extracellular space. However, cell morphology also has an influence on the dielectrophoretic properties. A good example where cell physiology and cell morphology are tightly coupled is cell division in budding yeast. When a suspension of yeast cells undergoing active division is injected into a miniaturized dielectrophoretic sorter; low-frequency focusing signals (100 kHz) in combination with a high-frequency discriminating signal at 1 MHz force dividing yeast cells to equilibrium positions on flow lines substantially shifted toward the high-frequency side of the flow channel (Fig. 11). This sorting effect is consistent with known dielectric properties of yeast cells in division [16]: The globally elongated shape due to bud growth increases cell polarizability and thus enhances pDEP as the cells progress through mitoses. However, when the neck closes after anaphase, the effect disappears by a progressive shift of the associated dispersion to lower and lower frequency, causing the pair to behave electrically like two individual single yeast cells. Therefore, the dielectrophoretic cell sorter using multiple-frequency dielectrophoresis is able to characterize and synchronize a yeast cell culture based on the different polarization mechanisms of cells that are associated to their cell-cycle phases [17].

Future Directions

Point-of-care devices based on on-chip flow cytometry are currently emerging. Miniaturization and integration is paramount to lowering the costs and skill levels required to carry out a given analysis. Small chip-based cell analysis devices are expected to be increasingly used to monitor different disease states and physiological functions. Potential applications with large markets include detection of circulating tumor cells (responsible for metastasis), blood analysis for quantification of physiological and immunological function. Such medical devices need to provide an extremely simple user interface and will contain all the necessary chemicals such as labels and buffers in an integrated form. Microfluidics is the key enabling technology, which will allow such devices to be developed.

Aside from the medical diagnostic applications, the small size of the on-chip format will enable continuous monitoring of cells in bioreactors or Petri dishes allowing cell population analysis in addition to the physical and chemical parameters usually measured (e.g., pH, dissolved oxygen, temperature, or glucose concentration). These technologies will lead to advances in the control and monitoring for industrial and laboratory-scale applications.

At a more fundamental level, optical measurements and electrical characterization and actuation are the main physical principles employed in on-chip flow cytometry. However, a series of other physical and chemical phenomena may be used for characterization and sorting, and potentially be integrated along with existing on-chip flow cytometric methods: acoustic deflection and acoustic impedance measurement, deterministic size-dependent displacement by micro- and nanostructured obstacles, cell-surface affinity, gravitational sedimentation rates, magnetic properties or probes, diffusion, etc.

Cross-References

- ▶ CMOS MEMS Biosensors
- ▶ Dielectrophoresis
- ▶ Electrical Impedance Cytometry
- ▶ Integrated Micro-Acoustic Devices
- ▶ Maxwell–Wagner Effect
- ▶ Microfluidic Whole-Cell Biosensor
- ▶ Single Cell Impedance Spectroscopy

References

1. Holmes, D., Pettigrew, D., Reccius, C.H., Gwyer, J.D., van Berkel, C., Holloway, J., Davies, D.E., Morgan, H.: Leukocyte analysis and differentiation using high speed microfluidic single cell impedance cytometry. Lab Chip **9**, 2881–2889 (2009)
2. Pamme, N., Koyama, R., Manz, A.: Counting and sizing of particles and particle agglomerates in a microfluidic device using laser light scattering: application to a particle-enhanced immunoassay. Lab Chip **3**, 187–192 (2003)
3. Cui, L., Zhang, T., Morgan, H.J.: Optical particle detection integrated in a dielectrophoretic lab-on-a-chip. Micromech. Microeng. **12**, 7–12 (2002)
4. Seo, J., Lee, L.P.: Microchip coulter particle counter. Sens. Actuators B Chem. **99**, 615–662 (2004)
5. Larsen, U.D., Blankenstein, G., Branebjerg, J.: In: Proceedings of Transducers, Chicago, pp. 1319–1322 (1997)
6. Koch, M., Evans, A.G.R., Brunnschweiler, A.: In: Proceedings Micromechanics Europe, Ulvik, pp. 155–158 (1998)
7. Koch, M., Evans, A.G.R., Brunnschweiler, A.: In Design and Fabrication of a Micromachined Coulter Counter, Proceedings of Micromechanics Europe, Ulvik, pp. 155–158 (1998)
8. Fuller, C., Hamilton, K., Ackler, J., Gascoyne, H.: In: Micro Total Analysis Systems 2000, Enschede, pp. 265–268 (2000)
9. Larsen, U.D., Norring, H., Telleman, P.: In: 4th International Symposium on Micro Total Analysis Systems, Enschede, pp. 103–106 (2000)
10. Gawad, S., Heuschkel, M., Leung-Ki, Y., Iuzzolino, R., Schild, L., Lerch, P., Renaud, P.: In: 1st Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine & Biology, Lyon, pp. 297–301 (2000)
11. Madou, M.J.: Fundamentals of Microfabrication: The Science of Miniaturization. CRC Press, Boca Raton (2002)
12. Gawad, S., Batard, P., Seger, U., Metz, S., Renaud, P.: In: Proceedings of MicroTAS 2002, Nara, pp. 649–651 (2002)
13. Gawad, S., Schild, L., Renaud, P.: Micromachined impedance spectroscopy flow cytometer for cell analysis and particle sizing. Lab Chip **1**, 76–82 (2001)
14. Pauly, H., Schwan, H.P.: Impendance of a suspension of ball-shaped particles with a shell; a model for the dielectric behavior of cell suspensions and protein solutions. Z. Naturforsch. Teil B: Chem. Biochem. Biophys. Biol. **14**(2), 125–131 (1959)
15. Valero, A., Braschler, T., Renaud, P.: A unified approach to dielectric single cell analysis: impedance and dielectrophoretic force spectroscopy. Lab Chip **10**, 2216–2225 (2010)
16. Asami, K., Sekine, K.: Dielectric modelling of cell division for budding and fission yeast. J. Phys. D Appl. Phys. **40**, 1128–1133 (2007)
17. Demierre, N., Braschler, T., Valero, A., Renaud, P.: In: The 12th International Conference on Miniaturized Systems for Chemistry and Life Sciences, San Diego (2008)

193-nm Lithography

- ▶ DUV Photolithography and Materials

Optical and Electronic Properties

Vanni Lughi
DI3 – Department of Industrial Engineering and Information Technology, University of Trieste, Trieste, Italy

Synonyms

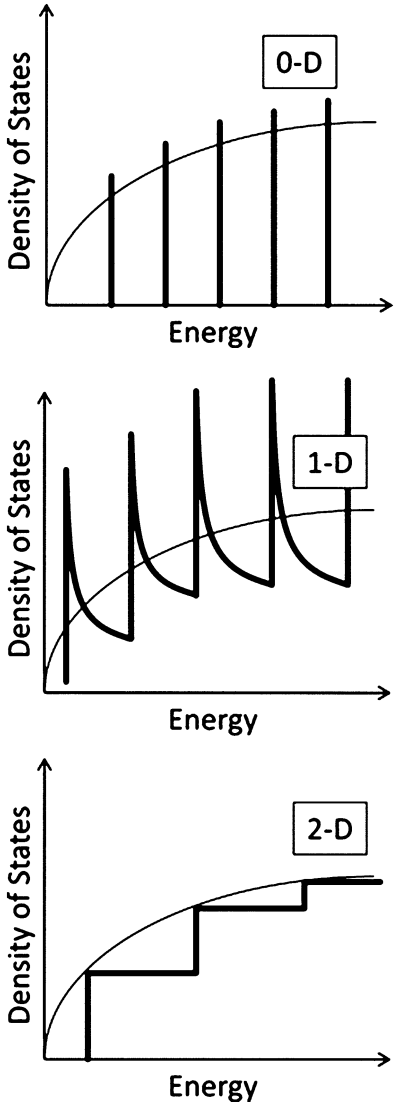
Optoelectronic properties

Definition

The *electronic properties* are a set of parameters and representations that fully describe the state and behavior of electrons in the material. For example, the electronic band structure [1], which describes the state of the electrons in terms of their energy, E, and momentum, $\boldsymbol{k}$, is such a representation. The electronic properties are, in turn, strongly coupled with the electric properties such as electrical conductivity and dielectric response, and with the *optical properties* such as refractive index, damping constant, absorbance, etc. – which describe the response of the material to the electromagnetic radiation.

Nanostructures are portions of a material where at least one spatial dimension is at the nanoscale – as defined below. Nanoparticles, nanowires, and thin films are nanostructures. In nanoparticles, all three spatial directions are at the nanoscale, and therefore electrons are confined to a zero-dimensional space. Analogously, nanowires have two nanoscale dimensions, leading to a confinement of the electrons to a one-dimensional space. Thin films have one nanoscale dimension and electrons are confined to a two-dimensional space. *Nanostructured materials*, i.e., bulk solids presenting a morphology at the nanoscale, shall also be considered here. Electronic, electric, and optical properties in nanostructures and nanostructured can be markedly different than in bulk materials, mainly due to the effect of quantum confinement of electrons in one or more spatial dimensions.

It is important to note that there is not a uniquely defined size threshold below which a material can be considered nanometric: Nanoscale effects emerge at sometimes markedly different scales for different properties. Therefore the "nanoscale" is well defined only once the property under investigation has been specified.



Optical and Electronic Properties, Fig. 1 Density of states for 0-D, 1-D, and 2-D nanostructures. All show a degree of discretization. The density of states 0-D is essentially a set of Dirac delta functions. The density of states of 1-D structures reflects the presence of a number of discrete sub-bands, each following the $E^{-1/2}$ behavior typical of the one-dimensional density of states. Analogously, the density of states of 2-D structures shows sub-bands with the constant behavior typical of the two-dimensional density of states. The *thinner line* shows the $E^{1/2}$ functional form typical of bulk materials

Electronic Structure in Nanostructures

Quantum confinement of the electrons, associated to a size reduction to the nanometric scale in one or more spatial directions, leads to a dramatic change in the electron density of states [2], as sketched in Fig. 1. Qualitative understanding can be obtained from the classic electron-in-a-box model: The main features are the formation of a



Optical and Electronic Properties, Fig. 2 Colloidal suspensions of CdSe nanocrystals showing a change of the absorption properties associated to a change of the nanocrystal size (Photo credit Michele Pianigiani)

set of electronic states at discrete energy values, and the fact that the spacing between such discrete states widens as the quantum confinement increases (i.e., the size of the box reduces). The details of such discretization depend on the number of spatial directions that have been confined. This behavior has a number of consequences on the properties of the nanostructures. For instance, concerning the optical properties, the size dependence on the separation of the electronic states leads to size-dependent absorption and emission properties. Concerning the electronic properties, widening of the separation between electronic states can lead to metal-to-semiconductor or to semiconductor-to-insulator transitions, such as in the classic cases of Bi and Si nanowires, respectively. Along with quantum confinement, another important general phenomenon is s*cattering of electrons at surfaces and interfaces*, which is most relevant for the electrical and thermal conductivity in nanowires and thin films. This shall be discussed in the relevant sections below.

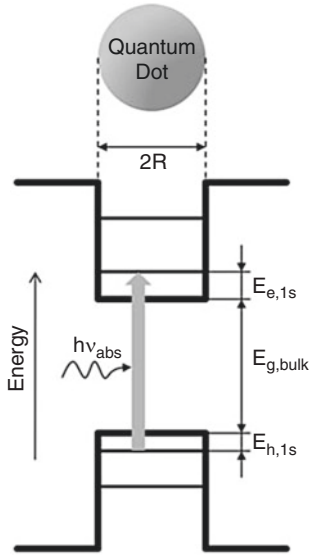
Key Nanostructures and Relevant Optical and Electronic Properties

Most of the scientific and technological interest for the optical and electronic properties of nanostructures is associated to semiconductors and metals, which will be the focus of this entry. Molecules and molecular optoelectronics shall not be considered here. All optical and electronic properties could in principle be analyzed for each kind of nanostructure. However, only a selection of the most relevant ones, from the scientific or technological point of view, is presented here for each class of structures.

Zero-Dimensional Nanostructures: Nanoparticles

A number of techniques have been developed to form nanoparticles of all kinds of materials both in the amorphous and the crystalline phase, including self-assembly from deposition of highly mismatched epitaxial films, aerosol methods, and colloidal syntheses. One of the main reasons for the strong interest that this class of nanostructures has generated over the past two decades is associated to the peculiar optical properties shown in particular by crystalline nanoparticles (nanocrystals, NCs) of metals and semiconductors. In particular, great interest arises because of the fine tunability of such properties in the visible and near-infrared range of the spectrum (Fig. 2).

Semiconductor NCs, also known as *quantum dots (QDs)* or *artificial atoms* [3], owe their characteristic optical properties to the peculiar, atom-like electronic structure illustrated in Fig. 1. As mentioned above, a limited number of electrons are strongly confined within the QD, leading to the emergence of quantum effects. For semiconductors, a good rule of thumb is that the quantum effects typical of NCs start emerging when the



Optical and Electronic Properties, Fig. 3 Schematic of a spherical quantum dot with radius R and of the associated potential profile along any diameter of the sphere. The energy states for electrons and holes are discrete and can be described by the particle-in-a-box model, where the effective masses are used. The *gray arrow* represents the first available exciton state, which in first approximation corresponds to the onset of optical absorption

crystal is smaller than the Bohr radius, aB, of the bulk exciton:

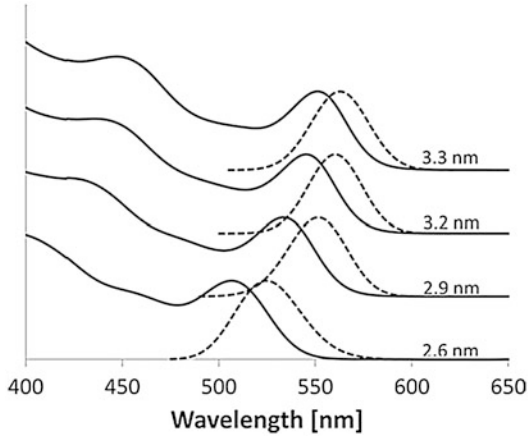
$$a_B = \frac{\hbar^2 \varepsilon}{e^2}\left(\frac{1}{m_e^*} + \frac{1}{m_e^*}\right)$$

Here ħ is the reduced Plank's constant, e the charge of the electron, ε the dielectric constant of the nanoparticle, m^* the effective mass of the electrons and of the holes. In first approximation, the energy structure of a QD and its surrounding environment can be described as sketched in Fig. 3, so that electronic states in the QD are essentially discrete and are quite well described by a simple approach based on the classic analysis of the electron-in-a-box analysis. Therefore, the separation between the highest occupied state and the lowest unoccupied state (also known as the *apparent bandgap* of the quantum dot or the *first exciton state*) is simply given by the following equation [4]:

$$E_{g,\mathrm{dot}} = E_{g,\mathrm{bulk}} + \frac{h^2}{8R^2}\left(\frac{1}{m_e^*} + \frac{1}{m_e^*}\right) - 1.8\frac{e^2}{\varepsilon R}$$

Here h is Plank's constant and R the radius of the nanoparticle. The second and third terms are the energies $E_{e,1s}$ and $E_{h,1s}$ associated to the confinement of the electrons and holes, respectively, calculated as particles in a box. The fourth term accounts for the Coulomb interaction. The *absorption* spectrum of QDs (Fig. 4) is characterized by a sharp absorption edge corresponding to $E_{g,\mathrm{dot}}$ and by a series of less evident peaks reflecting the higher energy exciton states available in the dot. Perhaps the feature of more practical interest in this equation is that it establishes a correlation between the size and the absorption spectrum of NCs. This correlation is found to be quite accurate in some size ranges, so that in principle, it could be used to determine the size of NCs by simply measuring the absorption spectrum. In practice, due to additional effects that are not captured in this simple model, empirical correlations are normally established for different materials by comparing the position of the first absorption peak and the size of nanocrystals as measured from a large number of TEM images. More accurate models based on the effective mass approximation account for the detailed electronic structure [5] and can predict experimental data to excellent accuracy.

Emission of QDs is shown in the photoluminescence (PL) spectra in Fig. 4, characterized by the strong band corresponding to the lowest available energy transition. The photoluminescence is strongly affected by the quality of the NC's surface: Passivation of surface defects by capping the NC with organic ligands (typically aliphatic monocarboxylates, phosphines, thiols) or better with a shell of a larger bandgap semiconductor can lead to quantum yields exceeding 0.9 in the best scenarios. A reduction of the quantum yield of an ensemble of quantum dots is also due to the phenomenon of *blinking*: Emission of a single QDs is intermittent, with dark-bright periods oscillating over a wide range of timescales, up to minutes. The mechanisms underlying blinking are still matter of debate in the scientific community,



Optical and Electronic Properties, Fig. 4 Typical absorption (*solid line*) and photoluminescence (*dashed line*) spectra of quantum dots. This example shows four CdSe nanocrystals with a radius ranging from 2.6 to 3.3 nm (Data courtesy of Luca Cozzarini)

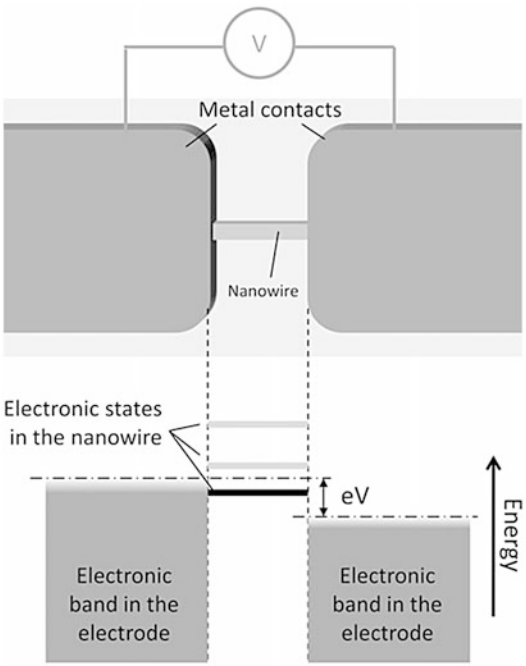
but seem to be associated to the formation of a single delocalized charge carrier within the NC, which can be excited via a non-radiative Auger process (dark phase). This is then followed by a radiative recombination of the carrier (bright phase).

Interestingly, the emission peak occurs at a markedly lower energy than the first absorption peak (Fig. 4). This phenomenon, known as *Stokes shift*, can be explained by observing that in nanocrystals, the valence band – which is normally degenerated in the bulk – splits into a number of states due to strong electron–hole exchange coupling and internal crystal field. The consequence is that the single transition predicted by the electron-in-a-box model is in fact a set of exciton states. Absorption can only occur at exciton states higher than the lowest (dark exciton) because of selection rules associated to the angular momentum, while emission from the lowest state can occur after the state becomes populated via thermalization from higher energy states. The Stokes shift enables a number of interesting applications, such as the possibility of efficiently tailoring the spectrum of a given source of illumination by absorption of unwanted spectrum ranges and re-emission within selected ranges – without incurring in the phenomenon of self-absorption.

Metal NCs are, too, characterized by a strong correlation between size and maximum of the absorption spectrum: Just as for the case of semiconductors, a size reduction leads to a blue shift of the absorption peak. However, in the case of metallic NCs, the absorption is dominated by an entirely different mechanism than in semiconductors, as it is associated to a surface plasmon resonance [4] (due to the interaction between the mobile electron cloud and the fixed atom nuclei in the nanocrystal). For metallic particles that are not too much smaller than the wavelength of the incident radiation, the dependence of the plasmonic resonance on the particle size is governed by Mie's theory [4], which relies on the solution of Maxwell's equations for small metallic spheres interacting with an electromagnetic wave. For particles smaller than about a tenth of the wavelength, the blue shift observed experimentally when decreasing the particle size can only be explained by taking into account quantum effects. In nanocrystalline metals, very high absorption coefficients ($\sim 10^9\,M^{-1}\,cm^{-1}$) are observed, mainly due to the enhancement of the absorption cross-section by the coupling between the plasmon and the polarization modes of the nanoparticle (plasmon-polariton resonance). Moreover, for metal NCs such as gold, silver, and copper, the surface plasmonic resonance occurs in the visible range, leading to interesting applications of such NCs for example as very effective dyes. It is worth mentioning that metal NCs have a wealth of applications as chemical and biological catalysts; however, as this is not the subject of this entry, these applications shall not be discussed here.

One-Dimensional Nanostructures: Nanowires

One of the most interesting phenomena occurring in one-dimensional nanostructures is associated with charge transport. A short nanowire (shorter than the typical mean free path of the material) connecting two bulk electrodes subject to an electrical bias does not behave as it would be expected by just applying the electrical properties of the corresponding material; rather, a quantized conduction behavior independent of the material is observed. A qualitative but powerful argument to



Optical and Electronic Properties, Fig. 5 (*Top*) Schematic of a possible arrangement for assessing the conductance of a short nanowire. (*Bottom*) Energy band diagram of the system. Due to the bias *V* applied to the electrodes, the Fermi levels are offset by *eV*. Only the electron states of the nanowire that fall within this range can contribute to electron transport from one electrode to the other. In the sketch, only one electronic state is available (shown in *black*), and the conductance is $G0 = 2e_2/h$. As the thickness of the nanowire increases, more modes might be included in the useful range *eV*, and the overall conductance will be enhanced by a quantum G_0 for each additional mode

demonstrate this is based on the idea that – as shown above for any nanostructure – the states of the nanowire are discrete, and only a limited number of such states will be comprised within the energy interval between the Fermi levels of the two biased electrodes and therefore available to carry charge carriers (Fig. 5). By carrying out this argument quantitatively, one obtains, for N available quantum states [4] within the nanochannel, a conductance of $G = N\,G_0$, where G_0 is the quantum conductance:

$$G_0 = \frac{2e^2}{h}$$

The inverse of G_0 is $R_L = h/2e_2 = 12.9\ k\Omega$ and is known as the Landauer resistance. Here e is the charge of the electron and h is Plank's constant. There are two remarkable observations: First, the Landauer resistance does not depend on the material and is in principle an universal constant (although a more detailed analysis would reveal a weak dependence on it [6]). Second, as the length of the channel is less than the mean free path of the electrons, no scattering is involved, so no energy is dissipated as a result of the transport mechanism illustrated here. In other words, the Landauer resistance is not to be intended as a classic dissipative resistance, rather as the ratio between the bias applied to the bulk electrodes and the current intended as probability of charge transport. The quantum conductance has been confirmed by quite striking experiments on carbon nanotubes and on short metal nanowires. In the latter, for example, the width of the nanochannel was thinned continuously while the current was measured. The conductance showed a perfectly discrete behavior, changing by multiples of G_0, thus reflecting the progressive reduction of the number of available states in the channel.

When the nanowires are longer than the mean free path of the charge carriers, transport is not controlled by quantum effects. However, for thin nanowires, conductivity can be governed by another important nanoscopic effect associated to surface scattering. As this mechanism has been studied first and most extensively for the case of thin films, the reader is referred to the following section for more detail.

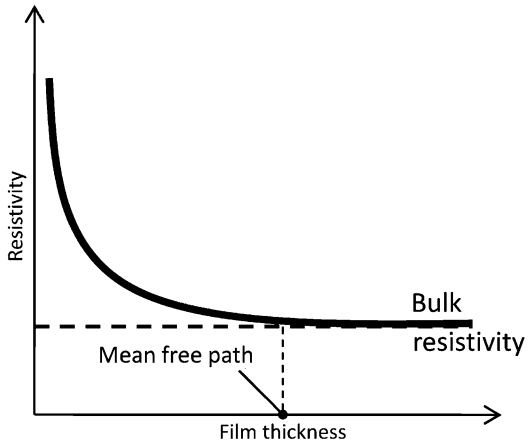
The optical behavior of nanowires is qualitatively analogous to the case of nanocrystals as discussed in the previous section: The main considerations on the size dependence of the discrete electronic states and therefore of the optical transitions still hold, and will not be discussed further.

Carbon nanotubes (CNTs) are another important class of one-dimensional nanostructures [7]. The optoelectronic properties of CNTs are governed by their chirality: One can imagine to obtain a CNT by folding a graphene sheet (a single graphite layer), and the direction of wrapping determines whether the structure has a semiconducting or metallic behavior. For

semiconductor CNTs, the bandgap is also determined by the chirality and the diameter of the tube. Due to the nearly perfect structure over great lengths, metallic CNTs can carry extremely high current densities (exceeding 10^9 A/cm^2). Large-scale applications, however, are currently limited by the lack of industrial techniques for producing CNTs of the desired kind in a controlled way.

Two-Dimensional Nanostructures: Thin Films

Electron transport in thin films and, as mentioned in the previous section, in nanowires can be affected by *surface scattering* [8]. This phenomenon arises when the film thickness is smaller than the carrier mean free path, and leads to a sharp resistivity increase with respect to the bulk value, as sketched in Fig. 6. This behavior can be modeled in principle by assuming that there is a certain degree of inelastic scattering of the charge carriers at the surface, i.e., the carrier velocity after a collision at the interface is partially or entirely uncorrelated with the velocity prior to the collision. This essentially classical model, first proposed by Thompson and later modified to include a degree of elastic scattering, predicts that the ratio between the bulk and the film resistivity (ρ_0 and ρ_f, respectively) is:



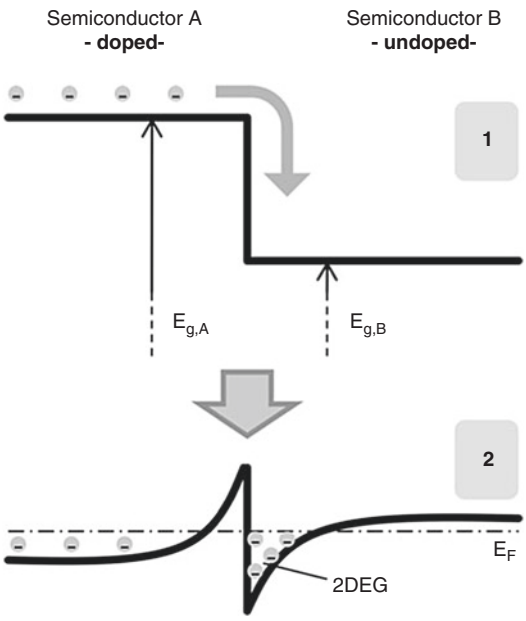
Optical and Electronic Properties, Fig. 6 Sketch of a film's resistivity as a function of the film thickness. Once the thickness of the film becomes comparable with the mean free path of the charge carriers, the resistivity departs from the constant bulk value and rapidly increases

$$\frac{\rho_0}{\rho_f} = \frac{3d}{4\lambda_0}(1+2P)\left(1\mathrm{n}\frac{\lambda_0}{d} + 0.423\right)$$

where P is the fraction of elastic scattering at the surface, d is the film thickness, and $\lambda 0$ is the mean free path. Fuchs and Sondheimer later included quantum effects. A number of difficulties, however, prevent these simple models to be of use in many practical situations, as P is a rather difficult parameter to quantify a priori, grain boundary scattering is not taken into account, and the film microstructure changes with the thickness.

Modern techniques such as Atomic Layer Deposition (ALD), Metal-Organic Chemical Vapor Deposition (MO-CVD), and mainly Molecular Beam Epitaxy (MBE) enable deposition of high-quality nanoscale films with atomic-scale precision. Heterostructures formed by layers of different materials can be deposited, making it possible to freely design novel electronic structures. Known as "bandgap engineering"; this approach has led, over the past 30 years, to remarkable scientific discoveries and technical achievements both in the field of electronics and optics. Only two of the fundamental building blocks that have enabled such tremendous progress in these fields are described here: the quantum well for optics, and the 2-D electron gas for electronics.

A *quantum well* can be fabricated by sandwiching a low-bandgap thin film between two layers of a high-bandgap "barrier" material [2]. The valence and conduction band profiles along the through-thickness direction of the heterostructure have essentially the same features shown in Fig. 3 for a quantum dot (of course, in accord to Fig. 1, the overall density of states is different for the two cases). Along this direction the electrons can be confined within the well, and once again the optical emission can be predicted by assuming that the electronic states are described by the electron-in-a-box model – as long as the confinement potential provided by the barrier is strong enough to resemble an infinite potential. Modern LEDs and semiconductor lasers are based on quantum well structures.



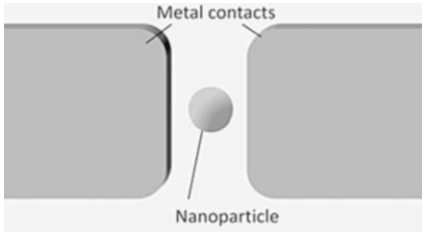
Optical and Electronic Properties, Fig. 7 Conduction band profile of a modulation doped heterostructure: A highly doped semiconductor *A* is placed in contact with an undoped semiconductor *B* whose conduction band lies at a lower energy than the conduction band of semiconductor *A*. (*Top*) An imaginary instant before electrons reach equilibrium. Some electrons statistically migrate from A to B. (*Bottom*) While electrons migrate from *A* to *B*, a potential builds up that would tend to push the electrons back to *A*. However, the conduction band offset determined by the choice of the two materials prevents the electrons to migrate back: The result is that excess electrons are trapped within a thin, two-dimensional layer near the interface (two-dimensional electron gas, 2DEG)

Another important thin film system is the *modulation doped heterostructure* [2]. The associated band profile is depicted in Fig. 7. Because of the initial band alignment, electrons from the highly doped region migrate to the undoped layer, and are confined in proximity of the interface due to the local electric field that naturally arises as the consequence of the electron migration. A *two-dimensional electron gas* (*2DEG*) is thus formed, i.e., a thin layer where a high number of electrons are confined. Most importantly, extremely high electron mobility is preserved thanks to the low doping level in this region. This peculiar 2D nanostructure is widely exploited in the fabrication of most modern high performance electronic devices. From a more fundamental perspective, it is interesting to report that in 2DEGs a *quantized Hall resistance* can be observed: $R_H = h/n\ e_2$, where $n = 1, 2, 3, \ldots$ Note that the quantum of Hall resistance is exactly twice the Landauer resistance observed in short nanowires.

Three-Dimensional Nanostructured Materials

Nanostructured materials are bulk materials characterized by having a relevant structure at the nanoscale, and are often obtained by combining nanostructures such as nanoparticles, nanowires, and thin films. *Large-scale assemblies of 1-D nanostructures*, such as self-assembled quantum dot arrays, colloidal close-packed solids, or assemblies of molecules, fall under this category. A number of mechanisms are relevant for understanding the optical and electronic properties of such assemblies, and will be described in the following.

Concerning the *optical properties* of assemblies of *weakly interacting* 1-D nanostructures [9] such as nanocrystals (NCs), dipolar interaction has an important role, as it can induce permanent polarization of the NCs and lead to ferroelectric and antiferroelectric phases in the assembly. Moreover, interaction between dipoles leads to enhanced absorption in the far IR range. Another relevant mechanism in arrays of weakly interacting NCs is Long Range Resonance Transfer (LRRT), where a transition dipole is transferred from an excited NC to one in its ground state. Due to the weak coupling, the transition rate is much slower than the absorption and relaxation processes. As a result, the absorption spectra of the close-packed NC solids in the optical region are essentially identical to those of the dispersed NCs. On the other hand, the photoluminescence (PL) spectrum is affected by LRRT: The energy transfer always occurs from smaller to larger NCs (the inverse process is not allowed as the smaller particle is transparent to the low-energy transitions that the larger NC can produce), so the emission of larger particles is enhanced. The result is that in weakly coupled NC solids, where there always is a certain spread of the size distribution, the PL is red-shifted with respect to the corresponding PL of dispersed NCs. As the coupling strength increases, e.g., due to a reduced



Optical and Electronic Properties, Fig. 8 Sketch of a system for observing and quantifying the Coulomb blockade mechanism

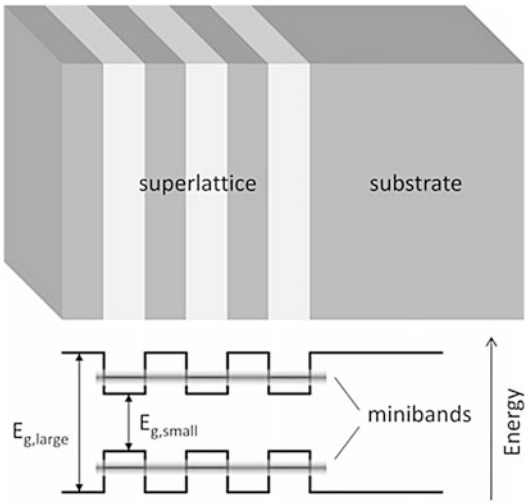
interparticle distance, both PL and absorption redshift as they converge to the optical behavior of the bulk.

Two peculiar mechanisms that are relevant for *charge transport* in large-scale assemblies of 1-D nanostructures are the Coulomb blockade and resonant tunneling [4]. These mechanisms, however, are best understood by first considering one single nanoparticle placed within a narrow gap between two electrodes, as in Fig. 8. The *Coulomb blockade* occurs when the electrical resistance between one electrode and the nanoparticle exceeds h/e^2 (two times the Landauer resistance): In this case, charge tends to accumulate on the particle when bias is applied to the electrodes, and no current flows in the structure until enough voltage has been applied to add one additional charge. By considering simple electrostatic arguments, it can be shown that the voltage required to add a single electron to a spherical particle with a radius of few nanometers is of the order of tens or hundreds of millivolts. Therefore, even at room temperature, the I-V characteristic of such a system shows a staircase-like behavior. Single Electron Transistors, SET, exploit this peculiar mechanism by using a third electrode to independently control the nanoparticle potential, thus enabling tuning or sensing the current at the finest, electron-level scale. If the coupling between the electrodes and the nanoparticle is stronger (i.e., the resistance between the electrodes and the nanoparticle is less than h/e^2), there is no charge accumulation on the particle and *resonant tunneling* occurs, where it can be shown [4] that transmission between the two electrodes is greatly enhanced with respect to normal vacuum tunneling.

Coulomb blockade and resonant tunneling are important elements in determining the electrical conduction regimes in arrays of NCs [9]. For example, in weak coupling conditions, the electric behavior of a metal NC array can been shown to be nonlinear and characterized by a conduction gap, i.e., a voltage threshold below which no current flows – in analogy with the Coulomb blockade. The strength of the coupling can be modified by simply engineering the distance between NCs (for example, in colloidal solids, the distance can be controlled by selecting the length of the organic molecules that serve as capping of the NCs). By using this approach, it is possible to control a transition between an insulator-like phase (weak coupling) to a highly conductive phase (strong coupling). Finally, high conductivity in weakly coupled NC assemblies can also be obtained by optical excitation (*photoconductivity*). For a given bias, the photocurrent increases linearly with the optical excitation intensity.

Three-dimensional arrays of small particles (mainly colloidal crystals, i.e., assemblies obtained from colloidal suspensions) are the subject of a number of studies due to their peculiar optical properties. Particularly interesting is the possibility of creating materials with *optical bandgaps* – i.e., materials that do not propagate light within a certain wavelength range – or, by exploiting higher-order effects, even materials with a negative refractive index [10]. A number of applications can be devised for these structures in the field of nonlinear optics; however at this time, no relevant devices have been realized.

Superlattices are another class of nanostructured materials, typically formed by an alternating sequence of large-bandgap and small-bandgap semiconducting thin films [2]. Such a structure and the associated electronic band profile are illustrated in Fig. 9. If the layer thickness is smaller than the mean free path, carriers can travel – without scattering – between one well to the other. The carrier mobility is very high, such that these structures are currently exploited in a number of high-frequency electronic devices. The periodic structure at the nanoscale produces electronic "minibands" (the simplest cases can be



Optical and Electronic Properties, Fig. 9 (*Top*) Schematic of a thin film heterostructure forming a superlattice. The layers are only a few nanometers thick. (*Bottom*) Energy band structure showing a series of periodically spaced potential wells. A set of *minibands* arises

analyzed by the Kronig-Penney model n), leading to finely tunable optical properties that are commonly exploited in optoelectronic devices such as semiconductor lasers. Finally, superlattices make it possible to observe the *Bloch oscillations*, a fundamental phenomenon predicted for any ideal periodic structure. This leads to a nonlinear, oscillating I-V characteristic with regions of negative differential resistance and can be exploited for fabricating solid-state switches and oscillators.

Cross-References

- ▶ Carbon-Nanotubes
- ▶ Charge Transport in Self-Assembled Monolayers
- ▶ Nanoparticles
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Surface Electronic Structure

References

1. Ashcroft, N.W., Mermin, N.D.: Solid State Physics. Holt-Saunders International Editions, London (1976)
2. Davies, J.H.: The Physics of Low-Dimensional Semiconductors: An Introduction. Cambridge University Press, Cambridge (1997)
3. Rogach, A. (ed.): Semiconductor Nanocrystal Quantum Dots: Synthesis, Assembly, Spectroscopy and Applications. Springer, Wien (2008)
4. Lindsay, S.: Introduction to Nanoscience. Oxford University Press, New York (2009)
5. Efros, A.L., Rosen, M.: The electronic structure of semiconductor nanocrystals. Annu. Rev. Mater. Sci. **30**, 475–521 (2000)
6. van Ruitenbeek, J.M.: Quantum point contacts between metals. In: Sohn, L.L., Kouwenhoven, L.P., Shoen, G. (eds.) Mesoscopic Electron Transport, pp. 549–579. Kluwer, Amsterdam (1997)
7. Dresselhaus, M.S., Dresselhaus, G., Avouris, P. (eds.): Carbon Nanotubes: Synthesis, Structure, Properties, and Applications. Springer, New York (2001)
8. Cao, G.: Nanostructures & Nanomaterials: Synthesis, Properties & Applications. Imperial College Press, London (2004)
9. Murray, C.B., Kagan, C.R., Bawendi, M.G.: Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. Annu. Rev. Mater. Sci. **30**, 545–610 (2000)
10. Sakoda, K.: Optical Properties of Photonic Crystals. Springer, New York (2005)

Optical and Mechanical Nanostructured Coatings for Future Large-Scale Manufacturing

Tomasz Stapinski[1] and
Katarzyna Tkacz–Smiech[2]
[1]Department of Electronics, AGH University of Science and Technology, Krakow, Poland
[2]Faculty of Materials Science and Ceramics, AGH University of Science and Technology, Krakow, Poland

O

Synonyms

Nano-scale structuring

Definition

Nanostructured coatings are functional materials with nanoscale structuring, that is, having a characteristic dimension, in at least one direction, lower than 100 nm. Their family comprises coatings with mean size of the grains and foreign-phase

inclusions smaller than 100 nm, multilayer systems composed of the layers thinner than 100 nm, and composites with dispersed nanoparticles.

The size effect in mechanical nanostructured coatings allows obtaining hardness, toughness, friction and wear, and corrosion resistance enhanced compared to traditional hard coatings.

Optical nanostructured coatings are functional materials which allow the modification of optical properties of bilayer or multilayer structures. The classical optics effects and quantum effects, for example, electron confinement in low-dimensional systems, are present in such structures.

Overview

The first decade of the twenty-first century passed under the keywords like nanoscience, nanotechnology, and nanomaterials. Some nanomaterials were, however, known to people much earlier. The examples are: soot which is nothing but nanocrystalline graphite powder and the films of gold obtained from Faraday's "ruby fluids" (colloidal gold). It was R. F. Feynman, who first in 1959 drew the world's attention to the nanoscale (*nano* in Greek means a dwarf). In half a century since Feynman claimed "There's plenty of room at the bottom" nanoscience has evolved into a revolutionary area of research and technology, and started playing a magnificent role in rapidly growing industry.

The very concept of *nanotechnology* comes from Norio Taniguchi who, in the paper published in 1974, defined it as "the processing of separation, consolidation and deformation of materials by one atom or by one molecule." This definition was spread by K. E. Drexler in his books "Engines of Creation: The Coming Era of Nanotechnology" (1986) and "Nanosystems" (1992). Drexler described nanotechnology as technology based on manipulation of single atoms and molecules. At present a very general definition is accepted according to which nanotechnology is all what one does in nanoscale [1]. The description as above was given by R. E. Smalley awarded the 1996 Nobel Prize for the discovery of fullerenes.

The term *nanomaterials* refers to a particular family of "materials" comprising clusters of three to about thousand of atoms, single stable molecules like fullerenes or nanotubes, nanosize monocrystals and their superlattices, phases composed of nanosize crystallites (grains) and thin films of the thicknesses not exceeding 100 nm. The scale at which nanomaterials are characterized lies between that of interest to condensed matter physicists and chemists.

Specifying, the nanomaterials may be defined as those complying the *dimensional condition* according to which one of the characteristic dimensions (at least in one direction, for example, layer thickness or grain diameter or the dimension of the constituent phase) is lower than 100 nm. This characteristic size-parameter of the nanomaterials is somewhat arbitrary (some sources give 1,000 nm instead of 100 nm and some also give 1 nm as lower limit) but that is the size which determines the unusual properties of nanomaterials. The material properties can change totally if the size-parameter goes below characteristic critical value.

A significant improvement in the properties of nanomaterials compared to bulk materials of the same chemical composition (by factors about 3–5 times but in some cases even 100 up to 1,000 times) is associated just with the small dimension. The nanosize grains have much greater surface to volume ratio than micro-grains and large fraction of their volume should be regarded as surface phase of the properties drastically different from those of the solid. For example, the 1 nm thick interfacial boundary of the 5 nm (diameter) grain takes about 50 % by volume. The surface and interfacial phases have also significant impact on remarkable behavior of mono- and multilayers.

Besides the surface effects, also the phenomena occurring inside the nanosize particles and within the thin layers may be governed by different rules than those formulated for bulk materials. For example, lower friction of nanoparticles may be connected with discrete vibrational modes of the spacing increasing with decreasing particle size.

High efficiency of the material may be reached not only thanks to its low external dimension but

also by tailoring the nano-level internal shape (nano-scale structuring). Such materials are called *nanostructured* ones. They satisfy *structural condition* according to which a way of joining of nanoparticles (or nanolayers in multilayer system) into the bulk must preserve their nanostructured nature and characteristic nanometer dimension of the building blocks (lower than 100 nm). The family of nanostructured materials comprises: the solids and coatings with mean size of the grains and foreign inclusions smaller than 100 nm, the multilayer systems composed of the layers thinner than 100 nm as well as composites (layers and solids) with dispersed nanoparticles. All they exhibit unique properties unachievable in traditional bulk solid.

A technologically important part of nanostructured materials encompasses functional *nanostructured coatings* which may be metal, ceramic or polymer ones or their combinations. The potential benefits of such coatings include mechanical, optical, magnetic, thermal, and electrical properties. Thanks to the present-day deposition techniques (like PVD, CVD, ALD, or electrodeposition), which permit refining the coatings' internal structure at the nanometer scale with simultaneous precise control of chemical composition and impurities level, the nanostructured coatings have become of increasing interest to researchers involved in materials science and surface engineering.

By engineering the coatings at the nanometer level, it is possible to obtain materials of particular optical absorption spectrum. Because linear and nonlinear optical properties are determined by particle–particle interactions some additional low energy bands may appear in the absorption spectrum of nanostructured coatings. This is why they are currently present in optical telecommunication displays and various illumination technologies. They can also be used as components of laser diodes and plasma displays. The technologies based on optical nanostructured coatings have also a great impact on development of conventional energy, especially photovoltaics.

A strong interest is also directed to nanostructured coatings for mechanical applications, where high hardness, toughness, wear and corrosion resistance and low friction are required. Although conventional layers provide some improvement of mechanical properties of the substrate they do not meet all high requirements of modern technologies (including machinery, car and aircraft industry, space exploration or medical tools).

Below, some key topics related to the nanostructured coatings for optical and mechanical applications and to the methods of formation of such coatings are provided. A special stress is given to the coatings which are good candidates as the commercial materials for future large-scale manufacturing.

Optical Coatings

The main stream of the nanostructured coatings technology concerns the production of optical coatings for high-quality products such as *mirrors* and *filters* for high-power *lasers*, and narrow band filters for analysis and communication. The technologies related with optical nanostructured layers have a great impact on development of *nonconventional energy. Solar cell* industry requires permanent increase in modules efficiency, so very important are thin film technologies dedicated to *photovoltaics*.

Photovoltaic Applications

One of the main reasons of huge development of plasma assisted CVD systems is the thin film photovoltaic industry. Commercial production of thin film devices is related with amorphous silicon technology started in the second half of the 1970s [2]. Several features make amorphous silicon–based alloys attractive PV materials. These are: high optical absorption coefficient, low-temperature-deposition process, tunable optical gap in the region where high energy conversion efficiencies are expected. CVD *amorphous silicon*–based materials can be easily doped to get both p- and n-type conductivity. The present single junction a-Si:H commercial solar cells are based on p-i-n technology. The production of cells on glass involves *transparent conducting oxide* (TCO) technologies and Plasma Assisted Chemical Vapor Deposition (PACVD) formation

methods (deposition of p-type layer of boron doped silicon/silicon carbide alloy, intrinsic a-Si:H, phosphorus doped n-type layer, and metal back electrode). CVD methods open the possibility of a tandem a-Si:H cell with a-SiGe:H layer production.

PACVD techniques implemented in thin film solar cells: Silicon due to its availability and environmental sustainability is the main photovoltaic material. Wafer-based silicon technology has had about 70 % of market shares. In 2005, the annual production of thin-film modules reached 100 MW. Typical deposition process for large area substrates mass production is PACVD.

The most promising are microcrystalline/amorphous "micromorph" silicon solar cells (Neuchatel) of record stable efficiency 10 % where the nanotechnology is implemented [3]. The authors proposed the use of textured front and back contacts and intermediate reflectivity layer (IRL). "Micromorph" modules efficiency is about 9 %. The commercialization of such concept was done by Kaneka Corporation giving PV modules of superior light absorption per nominal Watt power. Future increase in conversion efficiency requires proper light scattering at textured interfaces and IRL between the individual cells. High-frequency PACVD enables the preparation of SiO_x IRL of refractive index lower than 1.7 from CO_2 + SiH_4 gas mixture. Flexible tandem n-i-p micromorph solar cell architecture consists of front low pressure CVD (LPCVD) ZnO, a-SiH (PACVD) top cell, asymmetric intermediate reflector (LPCVD-ZnO), μ-Si:H bottom cell, and ZnO/Ag back reflector.

Cadmium telluride solar cells show high performance and relatively low production costs [4]. CdTe is a direct gap semiconductor (1.45 eV). It converts solar energy to electricity more efficiently than indirect band gap materials. The CdTe/CdS solar cells are present in PV market and the production leaders (e.g., First Solar) observed rapidly increasing market. Key technologies for large area CdTe thin film solar cells are CBD/CSS chemical bath deposition and close space sublimation. The application of high-resistivity SnO_2 layer between conducting SnO_2/CdS films in the cell architecture causes the increase in solar cell efficiency. There are some reports on production of resistive (10^3 Ω cm) SnO_2 films by the use of PACVD from $SnCl_4$ and O_2 reagents. The reaction

$$SnCl_4 + O_2 \rightarrow SnO_2 + 2Cl_2$$

takes place at 120–200 °C and SnO_2 amorphous film grows. In next solar cell manufacturing CSS process steps at high temperature (500 °C) amorphous SnO_2 structure changes its form to crystalline. The temperature of solar cell processing is higher than PACVD, SnO_2 layers exhibit crystalline structure, and optical absorption edge moves into higher energy values.

The use of CVD antireflection coatings for improvement of bulk solar cells: PACVD is the method commonly used in fabrication of *antireflective* and *passivating* coatings in crystalline and polycrystalline silicon solar cell technology [5]. One of the steps in manufacturing crystalline silicon solar cells is low-cost deposition of an antireflection coating. The role of such layer is the reduction of reflection and surface/bulk defects passivation. The promising candidate for this aim is hydrogenated silicon nitride produced at low temperatures from SiH_4 + NH_3 gas mixture. Two other materials SiO_2 and TiO_x can be produced by thermal CVD but the former has too low refractive index and latter does not passivate the silicon. The formation of silicon nitride by atmospheric pressure CVD (APCVD) from silane and ammonia mixture is possible at 700–1,000 °C. The use of low pressure LPCVD (0.1 mbar) process from dichlorosilane and ammonia can be performed at 750 °C. Only PACVD technology of silicon nitride is attractive for electronic industry due to low-temperature coating preparation.

In 1981, a-SiN_x:H PACVD technology was transferred from microelectronics industry to photovoltaics (ASE GmbH) and Kyocera used commercially PACVD silicon nitride in crystalline silicon solar cells. Nowadays direct PACVD methods of high frequency or low frequency are widely used in industrial applications. Usually the deposition of antireflective coatings is processed in batch reactors, where achieving of hundreds of

silicon wafers with a-SiN_x:H layers per hour is possible. PACVD systems, the inline technology developed by the Japanese manufacturer Shimadzu, are applied in which the antireflection coating based on silicon nitride guaranteeing excellent surface passivation is produced at 1,147 wafers per hour.

Other amorphous silicon–based alloys like a-Si:C:H [6] and a-Si:C:N:H can be used as antireflective coatings for silicon solar cells. On the basis of the spectral dependence of the reflectivity $R(\lambda)$ (total or specular), the effective reflectivity (R_{eff}) could be calculated from:

$$R_{eff} = \frac{\int R(\lambda)N(\lambda)d\lambda}{\int N(\lambda)d\lambda} \quad (1)$$

where $N(\lambda)$ presents the *AM1.5* standard solar spectrum. The integration was performed over the spectral range from 400 to 1,100 nm. For a-Si:C:H films obtained by PACVD from the gas mixture of the composition $[CH_4]/[CH_4 + SiH_4] = 0.23$ on multicrystalline silicon Si substrates $R_{eff} = 5.9$ % was reached.

For reduction of the front surface reflectance several researchers have used porous silicon layers as antireflective coatings. The refractive index depends on its porosity and morphology. Strong surface recombination due to roughness occurs after the formation of porous silicon onto the emitter in silicon solar cell. PACVD deposited SiO_x films on porous silicon surface provides chemical stability, passivation, low reflectivity, and improvement of electrical properties of solar cell [7].

Solar Cell Nanotechnology: The improvement in solar cell efficiency requires the implementation of new nanostructured materials. Quantum dot–based solar cells could be effective and nonexpensive alternative to bulk silicon concept. The absorption spectrum may be larger and better explored by size control of QDs. Second generation solar cells with QDs have been based on single junction concept. In third generation photovoltaics the cells architectures with QD absorbers in tandem multijunction configuration are used [8].

Semiconductor Nanocrystals

Semiconductor nanocrystals are the zero-dimensional objects. Due to electron confinement nanocrystals have discrete energy spectra dependent on their size. The electronic and optical properties may be tailored. Nanocrystals of energy gap of 1.8 eV may be tuned to 3 eV by decreasing the dimensions from 11.5 to 1.2 nm (CdSe nanocrystals). MBE is a widely used technique for producing two-dimensional quantum wells, superlattices, and three dimensionally confined islands with dimensions below 10 nm. Another method is chemical synthesis of controlled homogeneous nucleation, high-temperature inorganic precipitation in molten glasses, and moderate temperature metal-organic reactions. Numerous theoretical descriptions of electronic structure of nanocrystals are given: electron in the box model, finite well effects, Coulomb interactions, confinement-induced mixing of valence subbands, etc. The applications of sized controlled spectral properties find a place in lasers, emitting light devices, solar cells, and switching devices.

MOCVD (metal-organic CVD) can be applied to grow ZnO quantum dots (QDs). ZnO due to good optical and electrical properties is one of the attractive materials [9]. The band gap is 3.36 at room temperature and it makes the material transparent. After the doping, ZnO reveals higher conductivity and is suitable for electrodes applications. ZnO QDs indicate exciton effect and are the candidates for construction of blue and UV devices like LED and lasers. Among the chemical methods MOCVD conducted at elevated temperatures 350–500 °C in diethylzinc (DEZn), oxygen and argon carrier gas enables the preparation of ZnO QDs. The dimensions of QDs of 10 nm increase and the density decreases with increasing process temperature.

PACVD deposited nanocrystalline silicon (nc-Si) is the material for electronics applications in solar cells, TFT transistors. nc-Si-based solar cells show higher stability and higher efficiencies than a-Si:H ones. TFTs with nc-Si active layer revealed high electron mobility and high performance. Technology conditions influenced the optical properties and grain structure of this

material. SiO_x films grown by PACVD after annealing at temperature above 1,200 °C reveal nc-Si embedded in SiO_2. This material exhibits nonlinear optical properties. Real and imaginary parts of third order susceptibility $\chi^{(3)}$ exhibit nonlinearity. The nonlinear absorption depends on light intensity and $\chi^{(3)}$ decreases when the size of nanocrystals increases [10].

Nanoscaled Si_3N_4 may be obtained by means of various techniques: nitridation process, a-SiN_x:H nanopowder nitridation, CVD, and the most advantageous PACVD. Low temperature PACVD process and possibility of fast deposition are useful for nanocrystals and nanowires synthesis on large area substrates.

Photorefractive semiconductor structures exhibit nonlinear optical properties. Nonlinear optical devices in nanometer scale change their optical properties in response to light intensity and can be used to perform dynamic holography under low light intensity. Low-dimensional systems could enhance optical properties and electronic transport in comparison to bulk materials. High density of deep levels provides semi-insulating conductive character. Defect traps could store photogenerated carriers. The traps produce electric fields and modified optical properties. Low light intensity information in photorefractive materials is stored in the form of trapped charges. The potential application of photorefractive materials is the storage capacity. Due to progress in magnetic and semiconductor technologies no storage system based on holography was developed [11].

Nanostructured Coatings for Mechanical Applications

Nanostructured coatings provide a new generation of functional materials with high potential for significant improvement of the properties of mechanical parts. They include various types of coatings: nano-grain phases, nano-multilayers, and nanocomposites [12] with high operational parameters reached thanks to the size effect. The unusual mechanical properties make the nanostructured coatings a perspective choice instead of traditional ways to enhance the strength, wear and corrosion resistance and service life of any mechanical component. The coatings with tailored internal shape may soon replace currently used protective hard layers and may be competitive to other methods allowing the modification of the existing surface (like carburizing, nitriding, oxidation, or hybrid combination of various techniques).

The application of nanostructured coatings to protect mechanical components is also an attractive alternative to bulk nanostructured materials. The main idea is to bridge the benefits of unique mechanical properties of nanostructured bulk materials (like hardness and strength) and the synergism of the properties which may be obtained in the layer–substrate system.

Mechanical Behavior

The complexity of mechanical behavior of the functional coating, strictly coating–substrate system, depends on many factors. Firstly the bond between the coating and the substrate must be strong enough and free of internal stresses. Secondly, the coating processing or any other surface treatment cannot lower the substrate operational parameters. Finally the coating should obviously have the properties for which it was designed, including resistance to wear and corrosion, low friction, high hardness, and enhanced chemical and thermal stabilities.

A generally accepted model explaining why the hardness, ductility, and strength of nanostructured coatings differ from those of conventional ones is often reduced to small slip distance in nanosize grains and attributed to dislocation activity and enhanced stress relaxation at the grain boundaries. The grain boundary is more disordered than the grain inside and act as an aggregation of pinning points which control dislocation propagation. It consequently prevents the dislocation movement in a continuous slip plane increasing material strength and rising elasticity limit.

The description of increased strength of nanostructured materials and coatings is usually based on the empirical Hall–Petch equation according to which a *yield strength* σ_y of the grained material increases when the grain size goes down:

$$\sigma_y = \sigma_0 + \frac{k}{\sqrt{d}} \quad (2)$$

σ_0 in the above equation is a stress above which dislocation movement starts and *ky* is the strengthening coefficient – both are constant to each material.

Strength testing of bulk and layer materials, including nanostructured ones, is usually confined to hardness measurements and the dependence between grain size and the hardness is described with the Hall–Petch formula. Unfortunately it does not comply all cases. Many results show that the strength and hardness of nano-grain materials and thin coatings fall down below the value predicted by *Carlson and Wronski* [2]. The grain of the diameter lower than 10 nm houses one or two dislocations and its plastic deformation is mostly due to grain boundary processes. Because of hindered dislocation pile-up the grain boundary diffusion does not occur. In such case a release of the stresses takes place via grain boundary sliding which effects to a decrease of the strength. There is a limit of strengthening (not provided by Carlson and Wronski [2]) – infinitely strong materials do not exist.

In the nanostructured materials of broad grain size dispersion (or different sizes of the inclusions), various phenomena may occur side by side. This means that when the mechanisms responsible for the deformation or even destruction of the coating are analyzed not only the mean grain size but also the related distribution should be taken into account. The nanostructured coatings (and in general nanomaterials) of similar size grains will exhibit greater flow stress than the coating containing the grains having the same mean size but more dispersed.

Various approaches have been proposed, next to the Hall–Petch approach, to predict a dependence between the strength (or hardness) and grain size and calculate the lower *d* limit. Some of them are based on the concept of dislocations generated during deformation at the grain boundaries. The others introduce multi-phase model with hard grain boundary and soft grain interior. Those and some other models applicable to nanostructured metals have been reviewed in Weertman [13]. Our knowledge concerning more complex systems like covalent-ionic ceramic materials remains very limited.

Yet another case is represented by amorphous coatings in which a nucleation of cracks is suppressed. It may seem that such coatings should be the best candidates for tough protection. It is not so because amorphous materials are brittle – the crack propagation throughout amorphous structure is easy. The best way to prevent crack propagation is to introduce reinforcing particles into amorphous matrix. Their size should be equal to the width of shear stresses at the localized structure defects. As far as the toughness is concerned, nanostructured coatings designed as amorphous composites with embedded nanosize crystallites are very attractive ones.

Hardness of the coating is not only an indicator of its strength but is also considered as a consequential factor in the designing the coatings of high *work and corrosion resistance*. Strictly, an interplay of hardness and ductility has crucial meaning in this case. Hard coatings improve wear resistance but they lower ductility and increase a probability of brittle failure. Otherwise, the high ductility allows greater plastic deformation at minimized risk of the particles breaking and delamination. Too high ductility causes abrasion danger. As a result, hardness and strength from one side and ductility from the other are antagonists and the well-designed coating should have high hardness (*H*) and low Young (elastic) modulus (*E*) (high *H/E* ratio).

The reports concerning elastic behavior (and ductility) of nanostructured materials are somewhat ambiguous. Most of them, however, confirm a substantial decrease in *Young modulus* of the materials of small grains and attribute it with a presence of the pores (like in foams). In the case of dense materials, only fine nanostructures with the grains of the diameter below 100 nm show clear reducing of Young modulus.

Wear resistance and tribological properties of the coatings may not be considered beyond friction. As a rule, in friction description, adhesive interactions between working surfaces should be considered. Referring to the structure at atomic level one can say that the best way of friction

reduction is to terminate dangling bonds. Such effect is observed when comparing friction of diamond-like carbon (DLC) with much lower friction of hydrogenated DLC. More generally, hydrogen passivation of the surface layer allows friction reduction. The friction may also be regarded via a conversion of the energy of reciprocal (translational) motion of two working surfaces into vibrational energy. Such approach allows modeling friction of nanocrystalline and composite coatings. Vibrational modes of small systems are discrete which impedes phonon damping. As a consequence one observes lower friction of the particles of smaller sizes (e.g., grains or reinforcements in composites). Friction occurs when the inverse of phonon-life time is higher than the vibrational mode spacing. The last one increases with decreasing size. In very low size limit there is no friction.

The size effect in thin layers and multilayer systems amounts to the problem of layer thickness. Like in nanosize grain materials, it is a presence of linear defects (dislocations) which determines the mechanical properties of the single layer and the layers in multilayer coatings. A critical thickness of such layers can be determined. Below this thickness misfitting dislocations may develop causing a degradation of the mechanical characteristic. According to Freud-Nix theory the critical thickness may be evaluated from the condition that the work of dislocation creation be equal to the strain energy recovered from the system.

Regardless of the coating properties (not only mechanical ones but also optical, electrical, etc.) a fundamental problem which has to be overcome in the coating engineering refers to the strength of the bond between the coating and the substrate. Only good adhesion allows rising the synergism of the properties in the coating–substrate system. The strength of the bond of the coating to the substrate is given by a stress needed to delaminate the layer from the substrate and is proportional to intrinsic toughness: $\gamma_{SV} + \gamma_{CV} - \gamma_{SC}$. Its value equals the change in the surface energy γ when the substrate–coatings interface (SC) separates into the substrate–vapor (SV) and coating–vapor (CV) surfaces. The highest values of intrinsic toughness may be achieved by a formation of diffusive layers.

Examples

The nanostructured coatings are good candidates for machining and wear-resistant applications and may effectively compete with the traditional protective coatings like oxides, carbides, and nitrides. The benefits in general include higher strength and hardness, higher ductility and toughness, improved wear and corrosion resistance, and improved tribological properties but lower elastic modulus and friction. The nanostructured coatings have relatively low thermal expansion coefficient followed by high thermal and chemical stability. Operational parameters of any nanostructured coating depend on the type of the structure (nanocrystalline, composite, or multilayer) and on the characteristic size (average grain size, grain-size distribution, layer thickness, and sizes of reinforcing particles). Furthermore, they clearly depend on the chemical composition of the phases present in the structure as well as on the impurities and segregation – both influencing the phenomena at the interfaces and grain boundaries.

Below, some representative examples of nanostructured coatings which can find large-scale applications in mechanical components are reviewed. The review is based mainly on Refs. [14–19].

Nanostructured ceramic alumina Al_2O_3 and alumina-titania Al_2O_3-13 % TiO_2 ceramic coatings exhibit micro-hardness about 15–30 % higher than that of the conventional ones. Their wear resistance is significantly improved, especially under a high wear load. The particularly increased wear resistance has dense coatings free of pores and microstructure defects. The addition of nanosize alumina dispersions results in an increase in the fracture toughness and thermal shock. The alumina coatings may be deposited with application of plasma spray technique. The structure contains partially melted nanosize Al_2O_3 particles trapped between the fully melted coarser Al_2O_3 grains of micrometer size. The dense coatings can be formed at high powers. An additional

enhancement of the properties is achieved by addition of TiO_2. The structure of sprayed Al_2O_3–TiO_2 coatings typically contains the regions of two kinds. The first ones may be described as fully melted regions formed mainly by nano-size grains of c-Al_2O_3 with dissolved Ti^{+4}. The second are unmelted/partially melted nanostructured regions – different than the conventional coating of lamellar structure. The nanostructured Al_2O_3–TiO_2 coatings have been used on the main propulsion shaft in ships.

Titanium dioxide TiO_2 is a chemically stable, relatively hard, and nontoxic material. It exhibits good biocompatibility, raised mechanical performance, high stability, and is nonabsorbable by the human body. The TiO_2 coatings may be effectively produced with variety of methods, like metal-organic CVD, ALD, or sputtering. The stable layers may even be 4 nm thick. The coatings deposited with thermal spray process exhibit nano-size structure with characteristic semi-molten nanostructured TiO_2 particles. The embedded nanoparticles act as crack arresters, improving the crack resistance and hardness of the conventional coating. The titania coatings are seen as a new generation of the coatings for enhancing the performance of long-lasting implants and other biomedical devices.

The nanostructured TiN coatings may be among other fabricated by means of reactive plasma spraying of micrometer titanium powders in the atmosphere. Such coatings contain not only TiN but also Ti_3N phases. They are of lamellar structure – typical for spraying. The grains have columnar shape of the sizes ranging from 60 to 120 nm. The nanostructured TiN coatings are characterized by higher bonding strength and better fracture toughness than those with greater grains. Both TiN and TiO_2 coatings are known for their low chemical reactivity, high hardness and wear and corrosion resistance. Thanks to this they are used for protecting various surfaces. Similar properties to TiN have been found in chromium nitride CrN. It can be obtained in the form of extremely hard, inert, thin layer applicable to precision metal parts.

The fusion of the properties of TiN and TiO_2 may be combined by formatting multilayer nitrided/oxided coatings. The surface layer modified with glow discharge oxynitriding is composed from nanocrystalline and columnar grains of the TiN phase. Addition of oxygen at the end of the process allows formation of a thin layer with inclusions of TiO_2 nanograins and small amount of amorphous phase. Such layers exhibit high corrosion resistance and can be successfully used as coatings on NiTi alloys in biomaterial for orthopedic and dental applications.

Very promising operational parameters have been found in nanostructured coatings of the chemical composition based on Ti, Cu, and N and deposited via reactive cathodic arc ion plating. The received structure consists of one hard and one soft phases. The film hardness, elastic modulus, and residual stress depend on the copper content. For example, the hardness slightly increases when the copper content rises until approximately 1.0 at. %. Further increase of the Cu content (up to about 1.5 at.%) causes rapid increase of the hardness up to a maximum value of approximately 42 GPa. This is approximately 1.5 times higher than the value for the pure TiN film. The highest elastic modulus (of the order of 400 GPa) has been measured for the films with 1.5 at.% Cu. The enhancement of the properties is explained by the reduction of the grain size due to Cu addition.

Titanium Carbide (TiC) is a hard ceramic material of excellent tribological wear properties. The CVD coating based on TiC provides very promising properties making them outstanding candidates for mechanical applications. Titanium carbo-nitride (TiCN) coatings offer slightly higher hardness compared to TiN and slightly lower friction coefficient. It may be used to enhance abrasive wear resistance. It exhibits a complex structure consisting of TiN and TiC bonds in a crystalline matrix – no TiCN molecule exists. TiCN may be deposited in multilayer or graded form by various PVD methods. TiC may also be applied in the form of nanocomposite nc-TiC/a-C(H) coatings (obtained for example by magnetron sputtering). The typical structure of such composite consists of TiC nanocrystallites (2 nm) with volumetric fraction of about 0.3. These nanocrystallites are embedded in

amorphous carbon matrix. Small size of the TiC particles and their relatively low fraction ensure highly homogeneous nanolevel structure. The coatings exhibit low friction, high toughness and wear resistance, and characteristic self-lubrication.

WC-Co coatings are of the great importance with respect to their extra high wear resistance. Nanostructured WC-Co coatings are much denser than conventional ones. In these coatings, nanometer size particles of WC are distributed evenly in the cobalt-rich phase. Though partial decarburization of WC may occur, nanostructured WC-Co has friction coefficient by about 20 % lower than traditional coating. It is followed by increased hardness, fracture toughness, and wear resistance.

Nanocrystalline and nanostructured carbon coatings may be produced in various forms mainly with a use of CVD techniques. Nanocrystalline diamond films can provide very smooth coatings having high hardness and other properties of polycrystalline diamond films. The other structure forms include simple shapes or fibbers, powders, whiskers, and components with complex shapes and surface morphologies coated on various substrates. Ordered and disordered graphite and nano-porous carbon may be formed depending on the processing conditions. These carbon forms can serve as good tribological coatings of low friction coefficient for preservation of various tools or as intermediate thin films for further CVD deposition of diamond. Carbon coatings may be modified by addition of other elements. Addition of hydrogen and formation of hydrogenated structure dramatically lowers friction (thanks to H-terminated bonds). Silicon or nitrogen addition allows obtaining nanocrystalline silicon carbide or nonstoichiometric amorphous carbon nitride layers (with nitrogen content up to 30 at.%) of high hardness and wear resistance. When both are in the structure one obtains $SiC_xN_y(H)$ layers which join the properties of hard SiC and more ductile Si_3N_4. They may be produced with PACVD technique as the uniform ones (e.g., on metals in medical applications) or in graded form as $C:N/SiC_xN_y$ (to enhance properties of polymer surfaces).

Technologies

The properties of the nanostructured coatings may be manipulated at the processing stage on condition that the technology and processing conditions be carefully designed. It is obvious that a spectrum of the properties and hence applications of nanostructured coatings and materials may be enlarged thanks to recent advances in the respective technologies [20]. The techniques applied in the processing of nanostructured coatings which allow fabrication on the commercial scale include:

- Physical Vapor Deposition (PVD) and Chemical Vapor Deposition (CVD)
- Mechanical alloying/milling
- Crystallization of amorphous alloys
- Glow discharge
- Electrodeposition
- Sol–gel
- Hybrid methods applying multidisciplinary techniques

The above methods comprise both top-down techniques, from macroscale to nanoscale, and conversely bottom-up by assembly of atoms.

The CVD methods allow obtaining covalent and ionic-covalent layers of complex composition both for optical and mechanical applications. Actually the metal coatings are seldom produced in this way. CVD makes use of homogeneous (in gaseous phase) and heterogeneous (at the solid–gas interface) chemical reactions. The gaseous precursor is introduced to the reactor by convection and then it diffuses perpendicularly to the substrate surface. Although CVD may be developed in various solutions, Plasma Assisted CVD (PACVD) and Remote Plasma-Assisted CVD (RPACVD) in which plasma is a medium enhancing chemical synthesis seem to be the most promising ones. The use of plasma allows reducing processing temperature which is often critical for some applications (e.g., when polymer is used as a substrate or in the production of semiconductors). In RPACVD, the substrate is removed from the plasma region and the temperature of the substrate may not be higher than room temperature. Plasma may be generated by radio (13.56 MHz;

RFCVD) or microwaves (2.45 GHz; MWCVD) or with application of DC voltage (DC-plasma CVD). PACVD is widely used in nanofabrication of epitaxial, amorphous, monocrystalline, and nanocrystalline thin layers and multilayer and composite coatings. It also allows obtaining gradient layers with the chemical composition changing perpendicularly to the surface. The materials include various forms of carbon (diamond like carbon DLC, hydrogenated DLC, carbon fibers, nanofibers, and nanotubes), the compounds like silicon carbide SiC, carbon nitride CN(H) and oxynitride CNO, titanium nitride TiN, and titanium dioxide TiO_2, or nanocomposites like Mo/Al_2O_3 – mostly for mechanical (and biomedical) applications. A variety of optical coatings may also be deposited with CVD, for example: high refractive silicon nitride, amorphous silicon, or hydrogenated silicon.

Chemical composition may be particularly well tailored when metal-organic CVD (MOCVD) is applied. High-quality thin films may be deposited with flame-based Combustion Chemical Vapor Deposition (CCVD). This technique allows formation of the nanocrystalline thin coatings with given orientation (including epitaxial coatings) as well as nonporous layers.

Atomic Layer Deposition (ALD) is a CVD technique in which the precursor materials are separated during reaction. The precursors react with a surface in a sequential manner, one-at-a-time. This method is likely to have the greatest potential to produce very thin dense layers with precisely controlled chemical composition and atomic structure, for example: oxides (Al_2O_3, TiO_2, ZnO), metal nitrides (TiN, NbN, WN), and pure metals (Ru, Pt).

Like CVD also PVD methods offer great possibilities to produce nanostructured coatings, especially for cutting tools and parts of machinery. The PVD methods may be developed as magnetron sputtering, arc evaporation, and pulsed laser ablation. Magnetron sputtering is the most common PVD technique. It makes use of atomization of the precursor material by energetic bombardment of its surface by ions or neutral particles. There are almost no restrictions in the target materials. Single or multi-elemental targets may be applied, including pure metals and ceramics (semiconductors and isolators). In arc evaporation, a material is evaporated from the cathode target with electric field. The process begins with high-current striking followed by an increase of a localized temperature at the cathode spot to very high values (above 10,000 °C) and vaporization.

The removal (ablation) of atoms from the bulk material in Pulsed Laser Deposition (PLD) is done by vaporization upon laser irradiation. Pulsed Laser Deposition (PLD) is a reliable but rather expensive coating technique that is applied to deposit defined films with complex stoichiometry. The most common hard coatings available by PVD are nitrides: TiN, CrN, TiCN, and TiAlN. The properties of these coatings can be much improved not only by the addition of other elements (like V, Y, or Cr) but also by the development of nanostructure. Multilayers (TiN/C–N, MoS_2 with Cr) and nanocomposite coatings may be formed. Disadvantage of PVD methods, limiting applications, may be relatively low (physical) adsorption of the coatings to the substrate. Thus an effort is directed to the formation of transition diffusion layer. Some solution is supplied by Hybrid Physical-Chemical Vapor Deposition (HPCVD).

Besides CVD and PVD methods there are developed some other techniques applicable in the manufacturing of nanostructured coatings. Ceramic coatings (and particularly thin films) can be deposited with solgel method in which a sol of the precursor is subjected to gelation and densification. The method is particularly useful in fabrication of homogeneous multicomponent coatings. Hybrid coatings can be obtained via introducing a different phase into the sol. With application of post-synthesis treatment it is possible to obtain nanostructured oxides, carbides, or nitrides. Solgel deposition ensures high structure control, also at nanolevel scale. In particular, the nanocrystalline layers of known porosity may be prepared (e.g., TiO_2 thin films). Nanostructured metal coatings can be deposited by electrolytic and electroless plating methods. In electrodeposition porosity-free metallic (pure metals and alloys) amorphous of nano-grain coatings may be produced (e.g., Pd, Ni). The deposition from

an electrolyte solution takes place on conductive substrate. Unlike electrodeposition, electrical conductors are not required as substrates in electroless deposition. Depending on processing conditions and composition the grain sizes may range between 2 and 100 nm.

Mechanical alloying and milling are the methods most often used for commercial scale. Attritor milling is used to manufacture large quantities (in kilogram per batch) of micrometer-size particles with nanostructure grains. Mechanical alloying is a ball-milling process in which blended powders are welded and fractured to achieve alloying at the atomic level. One of the purposes of these techniques is to produce powders suitable for thermal spraying: of given composition, tailored grain size, and distribution.

Thermal spraying methods (like atmospheric plasma spraying APS, high-velocity oxy-fuel thermal spraying HVOF, detonation gun (D-gun) spraying arc spraying AS and flame spraying FS) allow producing metallic, nonmetallic, and ceramic coatings via spraying molten solid particles from a thermal source onto a substrate. The structure of the formed coating depends on the shape of the sprayed particles. The method has been successfully applied in the deposition of 316-stainless steel nanostructured coating or nanocomposite and Cr_3C_2–NiCr and WC-Co cermet coatings.

Summary

Nanostructured coatings offer a great potential to serve as functional materials in many applications. In particular, they may be used in optical technologies, like laser or photovoltaic devices, and to protect components of machinery and tools. These two applications may be introduced on a large scale which make nanostructured coatings manufacturing an important industry with significant economical impact. Reliable coatings are needed to prevent wear and corrosion, to improve durability and frictions of the working component parts used both in commercial and advanced applications (like biomedical ones). They can also enhance yielding of optical devices, like solar cells. Functional nanostructured coatings may be produced thanks to the present-day development of modern technologies applicable on industrial scale. Further development of industry with the implementation of nanostructures becomes one of the key problems in the twenty-first century.

Unresolved problems comprise our still limited understanding of the mechanisms responsible for the size effect. The accepted theories, like the Hall–Petch approach, do not give full explanation of the enhancement of the properties as the characteristic size goes below 100 nm.

Cross-References

- ▶ Chemical Vapor Deposition (CVD)
- ▶ Hybrid Solar Cells
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanotribology
- ▶ Optical and Electronic Properties
- ▶ Physical Vapor Deposition
- ▶ Plasticity Theory at Small Scales
- ▶ Sol–Gel Method

References

1. Bueno, O.: The Drexler-Smalley debate on nanotechnology. Incommensurability at work? HYLE Int. J. Philos. Chem. **10**, 83–98 (2004)
2. Carlson, D.E., Wronski, C.R.: Amorphous silicon solar cell. Appl. Phys. Lett. **28**, 671–783 (1976)
3. Meillaud, F., Feltrin, A., Despeisse, M., Haug, F.-J., Domine, D., Python, M., Soderstrom, T., Cuony, P., Boccard, M., Nicolay, S., Ballif, C.: Realization of high efficiency micromorph tandem silicon solar cells on glass and plastic substrates: issues and potential. Solar Energy Mater. Solar Cells **95**, 127–130 (2012)
4. Feng, L., Wu, L., Lei, Z., Li, W., Cai, Y., Cai, W., Zhang, J., Luo, Q., Li, B., Zheng, J.: Studies of key technologies for large area CdTe thin film solar cells. Thin Solid Films **515**, 5792–5797 (2007)
5. Lee, J.Y., Glunz, S.W.: Investigation of various surface passivation schemes for silicon solar cells. Solar Energy Mater. Solar Cells **90**, 82–92 (2006)
6. Stapinski, T., Swatowska, B.: Amorphous hydrogenated silicon-carbon as new antireflective coating for solar cells. J. Non-Cryst. Solids **352**, 1406–1409 (2006)

7. Remache, L., Mahdjoub, A., Fourmond, E., Dupuis, J., Lemiti, M.: Influence of PECVD SiOx and SiNx:H films on optical and passivation properties of antireflective porous silicon coatings for silicon solar cells. Phys. Status Solidi C **8**(6), 1893–1897 (2011)
8. Ruhle, S., Shalom, M., Zaban, A.: Quantum-dot-sensitized solar cells. Chem. Phys. Chem. **11**, 2290–2304 (2010)
9. Yang, L.M., Ye, Z.Z., Zeng, Y.J., Xu, W.Z., Zhu, L.P., Zhao, B.H.: Density controllable growth of ZnO quantum dots by MOCVD. Solid State Commun. **138**, 577–579 (2006)
10. Vijaya Prakash, G., Cazzanelli, M., Gaburro, Z., Pavesi, L.: Nonlinear optical properties of silicon nanocrystals grown by plasma-enhanced chemical vapor deposition. J. Appl. Phys. **91**, 4607–4910 (2002)
11. Gunter, P., Huignard, J.-P.: Photorefractive Materials and Their Applications: Applications. Springer, New York (2007)
12. Malshe, A.P., Jiang, W., Dhamdhere, A.R.: Nanostructured coatings for machining and wear-resistant applications. J. Min. Metals Mater. Soc. **54**, 28–30 (2002)
13. Weertman, J.R.: Mechanical behavior of nanocrystalline metals. In: Koch, C.C. (ed.) Nanostructured Materials: Processing, Properties and Potential Applications, pp. 397–421. Noyes/William Andrew, Norwich/New York (2002)
14. Zhang, J., He, J., Dong, Y., Li, X., Yan, D.: Microstructure and properties of Al_2O_3-13%TiO_2 coatings sprayed using nanostructured powders. Rare Metals **26**, 391–397 (2007)
15. Xiao, L., Dong, Y.: Characterization of nanostructured TiN coatings fabricated by reactive plasma spraying. Surf. Coat. Technol. **202**, 1928–1934 (2008)
16. Lelatko, J., Goryczka, T., Paczkowski, P., Wierzchon, T., Morawiec, H.: TEM studies of the nitrided/oxided Ni–Ti surface layer. J. Microsc. **237**, 435–438 (2010)
17. Myung, H.S., Han, J.G., Boo, J.H.: A study on the synthesis and formation behavior of nanostructured TiN films by copper doping. Surf. Coat. Technol. **177–178**, 404–408 (2004)
18. Pei, Y.T., Galvan, D., De Hosson, J.T.M., Cavaleiro, A.: Nanostructured TiC/a-C coatings for low friction and wear resistant applications. Surf. Coat. Technol. **198**, 44–50 (2005)
19. He, J., Schoenung, J.M.: A review on nanostructured WC–Co coatings. Surf. Coat. Technol. **157**, 72–79 (2002)
20. Glocker, D.A., Shah, S.I. (eds.): Handbook of Thin Film Process Technology, vol. 1. IoP, Bristol/Philadelphia (2002)

Optical Cavity Biosensor

▶ Whispering Gallery Mode Resonator Biosensors

Optical Frequency Magnetic Dipole Transitions

Sinan Karaveli and Rashid Zia
School of Engineering, Brown University, Providence, RI, USA

Synonyms

Magnetic dipole emitters; Magnetic dipole transitions

Definition

Optical frequency magnetic dipole transitions are processes in which the magnetic field of electromagnetic radiation mediates a change in electronic states via the absorption or emission of a photon.

Overview

Nanoscale fabrication has enabled dramatic developments in the field of optical metamaterials, which in turn have expanded both the range of materials and phenomena available for future photonic devices. For example, researchers have used patterned metal nanostructures, including split-ring resonators and paired nanorods, as building blocks to engineer metamaterials with significant magnetic permeabilities at optical frequencies [1]. Whereas metamaterials research explores the artificial magnetic resonances supported by man-made structures, this entry describes the naturally occurring magnetic dipole transitions in atomic systems.

Based on a comparison of dipole moments in the Bohr model of hydrogen, it is often assumed that magnetic dipole transitions can be neglected at optical frequencies. Arguments in the literature contrast the approximate electric dipole moment ea_0 (where e is the electron charge and a_0 is the Bohr radius) with the Bohr magneton $\mu_B = \alpha e a_0/2$ (where α is the fine-structure constant). In CGS

units, the Bohr magneton is smaller than the approximate electric dipole moment by half the fine-structure constant α, which itself is approximately 1/137. As the transition strength depends on the square of the dipole moment, it would appear that magnetic dipole transitions should be $\sim 10^5$ times weaker than electric dipole transitions.

However, the precise quantum mechanical transition strength depends on the specific electronic structure. In systems where electric dipole transitions are weak or not allowed, magnetic dipole transitions can be significant. In fact, magnetic dipole emission is observed everyday from Lanthanide ions such as trivalent Europium (Eu^{3+}), a common phosphor in fluorescent lighting. Magnetic dipole emission is also significant in the 1,550 nm line of trivalent Erbium (Er^{3+}), the gain source for Erbium-doped fiber amplifiers in telecommunication networks. With the growing interest in optical frequency magnetic resonances and magnetic light-matter interactions, magnetic dipole transitions are also finding new applications in experimental nano-optics and active nanophotonic devices.

Optical Transitions and the Multipole Expansion

In describing the way that light is absorbed or emitted by an atom or molecule, it is often convenient to expand light-matter interactions into multipolar transitions, e.g., electric dipole, magnetic dipole, electric quadrupole, etc. These transitions can be formally derived by a multipolar expansion of the interaction Hamiltonian [2] as follows:

$$\hat{H}_{\text{int}} = -\vec{p}\,.\,\vec{E} - \vec{m}\cdot\vec{B} - \left[\overleftrightarrow{Q}\,\nabla\right]\cdot\vec{E} - \ldots$$

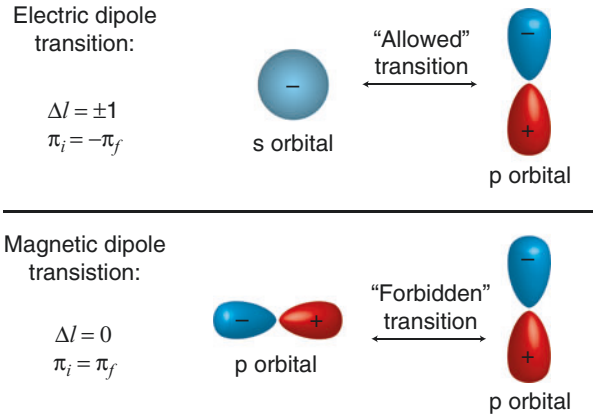
where each term is associated with a different multipolar moment of the electronic charge distribution, e.g., the electric dipole moment $\vec{p}$, the magnetic dipole moment $\vec{m}$, and the electric quadrupole moment Q. This equation shows that the electric dipole moment interacts with the electric field $\vec{E}$, while the magnetic dipole moment interacts with the magnetic field $\vec{B}$. Higher-order terms, such as the electric quadrupole, interact with spatial derivatives of the electric and magnetic fields. Since the magnetic field is related to the curl of the electric field, the magnetic dipole and electric quadrupole terms are often grouped together as terms that depend on the spatial variation of $\vec{E}$.

Electric Dipole Approximation and the Parity Selection Rule

In terms of the electronic states which they connect, electric and magnetic dipole transitions are readily distinguished by the parity selection rule [3, 4]. Since the size of the electron wave function (on the order of angstroms) is much smaller than the wavelength of visible light (λ = 300–700 nm), it is generally fair to assume that the electron sees electromagnetic radiation as a time-varying but spatially uniform field. This is known as the electric dipole approximation, where the spatial variation of the electric field is neglected such that $\vec{E}\,(r,t) \equiv \vec{E}_0 \exp\left(\left(\vec{k}\cdot\vec{r} - \omega t\right)\right) \simeq \vec{E}_0 \exp(-i\omega t)$ and is equivalent to retaining only the first constant term in the Taylor expansion of $\exp\left(i\,\vec{k}\cdot\vec{r}\right) = 1 + i\,\vec{k}\cdot\vec{r} + \frac{\left(\vec{k}\cdot\vec{r}\right)^2}{2!} + \frac{\left(i\vec{k}\cdot\vec{r}\right)^3}{3!} + \ldots.$

In such a spatially uniform electric field, the electron will experience a time-varying force $\vec{F} = -e\vec{E}_0 \exp(-i\omega t)$ that can push or pull it along one axis. This interaction can thus mediate the interorbital transition from a symmetric $1s$ orbital to an antisymmetric $2p$ orbital, as shown in Fig. 1. However, it cannot mediate intra-orbital transitions between electronic wave functions with the same symmetry, e.g., $1s \rightarrow 2s$ or $2p_x \rightarrow 2p_z$. In the electric dipole approximation, the initial and final states of a transition must have different parities. This is known as the parity selection rule for electric dipole allowed transitions. For single-electron wave functions, this rule implies that there must be a change in the orbital quantum number l [3, 4]: $\Delta l = l_{\text{final}} - l_{\text{initial}} = \pm 1$.

When electric dipole transitions are weak or not allowed, higher-order "forbidden" processes such as magnetic dipole and electric quadrupole transitions must be considered. These processes



Optical Frequency Magnetic Dipole Transitions,
Fig. 1 Examples of single-electron dipole transitions. Unlike interorbital electric dipole transitions that occur between electronic states of differing parity, magnetic dipole transitions are intra-orbital transitions between electronic states of the same parity

Optical Frequency Magnetic Dipole Transitions, Table 1 Selection rules for pure LS coupling

Electric dipole transition	Magnetic dipole transition	Electric quadrupole transition
$\Delta L = 0, \pm 1 (0 \nleftrightarrow 0)$	$\Delta L = 0$	$\Delta L = 0, \pm 1, \pm 2 (0 \nleftrightarrow 0, 0 \nleftrightarrow 1)$
$\Delta S = 0$	$\Delta S = 0$	$\Delta S = 0$
$\Delta J = 0, \pm 1 (0 \nleftrightarrow 0)$	$\Delta J = 0, \pm 1 (0 \nleftrightarrow 0)$	$\Delta J = 0, \pm 1, \pm 2 \left(0 \nleftrightarrow 0, 0 \nleftrightarrow 1, \frac{1}{2} \nleftrightarrow \frac{1}{2}\right)$
Change of parity $\Pi_f = -\Pi_i$	*Parity conserved* $\Pi_f = \Pi_i$	*Parity conserved* $\Pi_f = \Pi_i$

are mediated by spatial variations in the electric field. Through the Lorentz Force $\vec{F} = -e\left(\vec{v} \times \vec{B}\right)$ and gradients to the Coulomb Force $\vec{F} = -e\,\vec{E}$, spatial variations in $\vec{E}$ can produce torque and thus drive rotational transitions such as $2px \rightarrow 2pz$, as shown in Fig. 1. For these transitions, the initial and final states have the same parity. For single-electron wave functions, this implies that the orbital quantum number l remains constant [4]: $\Delta l = 0$.

Table 1 shows the selection rules for multielectron configurations in pure *LS* coupling, where the electron wave functions are described in Russell–Saunders notation by their total orbital angular momentum L, total spin angular momentum S, and total angular momentum J [3].

Calculating Line Strengths, Emission Rates, and Oscillator Strengths

To determine the likelihood of magnetic dipole absorption or emission, one must first model the electronic states and then calculate the quantum mechanical matrix element for the magnetic dipole operator $\hat{H}_{MD} = \mu_B\left(\hat{L} + 2|\hat{S}\right)$ [3, 4]. The square of this matrix element is called the magnetic dipole line strength S_{MD}:

$$S_{\mathrm{MD}}(\psi_1, \psi_2) \equiv \left|\left\langle \psi_1 \middle| \hat{H}_{\mathrm{MD}} \middle| \psi_2 \right\rangle\right|^2$$

In SI units,the Bohr magneton is defined as $\mu_B = e\hbar/(2m_e)$ in units of A $\mathrm{m}^2 = (\mathrm{J/T})$. Thus, the line strength for a magnetic dipole transition is in units of $(\mathrm{J/T})^2$, whereas the line strength for an electric dipole transition is in units of $(\mathrm{C\ m})^2$. To make a direct comparison between electric and magnetic transitions, the line strength must be converted into a transition probability or oscillator strength.

The transition probability for spontaneous emission (also known as the Einstein A coefficient) from the upper state ψ_2 to the lower state ψ_1 may be calculated, in units of inverse seconds, from the magnetic dipole line strength as follows [4, 5]:

$$A_{\mathrm{MD},2\rightarrow 1} = \frac{16\pi^3\mu_0}{3h}\left(\frac{n}{\lambda}\right)^3 \frac{S_{\mathrm{MD}}(\psi_1, \psi_2)}{g^2}$$

O

where all terms are in SI units, μ_0 is the permeability of vacuum, n is the refractive index of the host medium, λ is the free-space wavelength, and g_2 is the degeneracy of the initial upper state. (Note that this equation explicitly includes the n^3 dependence of magnetic dipole emission in an isotropic dielectric medium [6].)

The line strength may also be related to the oscillator strength f, a dimensionless quantity often used to compare absorption lines. (The oscillator strength is defined as the ratio of the absorption likelihood for a given transition to that of a single-electron classical harmonic oscillator.) The oscillator strength for the magnetic dipole transition from the lower state ψ_1 to the upper state ψ_2 can be calculated as follows [6, 7]:

$$f_{\mathrm{MD},\,1\to2} = \frac{8\pi^2 m_e}{3he^2c}\left(\frac{n}{\lambda}\right)\frac{S_{\mathrm{MD}}(\psi_1,\psi_2)}{g_1}$$

where all terms are in SI units, and where g_1 is the degeneracy of the initial lower state.

As an overview of such calculations, the following provides a brief example for the 588 nm magnetic dipole transition in Eu^{3+} between the $^5D_0$ ($L = 2$, $S = 2$, $J = 0$) and $^7F_1$ ($L = 3$, $S = 3$, $J = 1$) levels. (This magnetic dipole line is clearly observable in the emission of fluorescent lights.) Based on the selection rules for pure LS coupling in Table 1, it would appear that this transition ($\Delta L = 1$, $\Delta S = 1$, $\Delta J = 1$) should not be allowed. However, it is important to note that the electronic states of most Lanthanide ions are generally described by the *intermediate coupling* configuration of N valence electrons in the $4f$ orbital [5–8]. (Since neither the spin-orbit nor Coulombic interactions dominate, the electronic eigenvectors must be found by diagonalizing the sum of the spin-orbit and Coulombic Hamiltonian matrices. For Lanthanide ions, this is often done in the LS basis by mixing terms of the same J value, with the mixed levels named according to the dominant LSJ component.) Analysis of experimental spectra can be combined with group theory calculations (Racah algebra) to expand the free-ion wave function $|4f^N\psi_J\rangle$ into a basis set of LS coupling levels $|4f^N\alpha SLJ\rangle$ as follows [5–8]:

$$\left|4f^N\psi_J\right\rangle = \sum_{\alpha,S,L} C(\alpha,S,L)\left|4f^N\alpha SLJ\right\rangle$$

where $C(\alpha, S, L)$ is the intermediate coupling coefficient, S is the total spin angular momentum, L is the total orbital angular momentum, J is the total angular momentum, and the term α includes all other quantum numbers required to describe the levels. To emphasize the mixed nature of these levels, brackets are sometimes used around the LS term in naming the dominant component; this notation also emphasizes that the J value outside the bracket is still a good quantum number. For example, the $[^5D]_0$ level in the intermediate coupling of the $4f^6$ configuration of Eu^{3+} includes significant ^5D and ^7F terms, while the $[^7F]_1$ level includes ^7F, ^5P, ^5D, and ^5F terms [9]:

$$\begin{aligned}\left|\left[^5\mathrm{D}\right]_0\right\rangle = {} & -0.2381\left|^7\mathrm{F}_0\right\rangle \\ & -0.1969\left|^5\mathrm{D}_0\right\rangle + 0.6893 \\ & \times\left|^5\mathrm{D}'_0\right\rangle - 0.539\left|^5\mathrm{D}''_0\right\rangle\end{aligned}$$

$$\begin{aligned}\left|\left[^7\mathrm{D}\right]_1\right\rangle = {} & 0.9742\left|^7\mathrm{F}_1\right\rangle - 0.0027\left|^5\mathrm{P}_1\right\rangle + 0.0052 \\ & \times\left|^5\mathrm{D}_1\right\rangle + 0.1472\left|^5\mathrm{D}'_1\right\rangle - 0.1645\left|^5\mathrm{D}''_1\right\rangle \\ & + 0.0263\left|^5\mathrm{F}_1\right\rangle + 0.0162\left|^5F'_1\right\rangle\end{aligned}$$

where the prime terms on the right-hand side are used to denote levels with different Racah parameters (W, U).

To calculate the line strength between these levels, the contributions from each LS basis term to the transition matrix element must be summed and then squared [5, 8]:

$$\begin{aligned}S_{\mathrm{MD}} &\equiv \left|\left\langle f^N\psi_{J'}\middle|\hat{H}_{\mathrm{MD}}\middle|f^N\psi_J\right\rangle\right|^2 \\ &= \left|\sum_{\alpha'S'L',\,\alpha SL} C(\alpha',S',L')C(\alpha,S,L)\left\langle 4f^N\alpha',S',L'J'\middle|\hat{H}_{\mathrm{MD}}\middle|4f^N\alpha SLJ\right\rangle\right|^2\end{aligned}$$

In the *LS* basis, the only nonzero matrix elements in the summation will be those transitions which obey the magnetic dipole selection rules ($\Delta L = 0$, $\Delta S = 0$, $\Delta J = 0, \pm 1$). For the example case of Eu^{3+}, this means that the line strength can be expressed in terms of just two matrix elements [8]:

$$S_{MD}\left(Eu^{3+}:{}^7F_1 \leftrightarrow {}^5D_0\right) \equiv \left|\left\langle [{}^7F]_1 \middle| \hat{H}_{MD} \middle| [{}^5D]_0 \right\rangle\right|^2 = \left|-0.2320\langle {}^7F_1|\hat{H}_{MD}|{}^7F_0\rangle + 0.1891\langle {}^5D_1|\hat{H}_{MD}|{}^5D_0\rangle\right|$$

where the ^5D contribution in the above expression has been simplified to include all Racah terms, (i.e., $0.1891 = -0.1969 \times 0.0052 + 0.6893 \times 0.1472 + 0.539 \times 0.1645$).

The magnetic dipole matrix element between pure *LS* terms, $^{2S+1}L_J$ and $^{2S+1}L_{J'}$, can be explicitly calculated for the three allowed cases ($J' = J - 1$, $J' = J$, and $J' = J + 1$) as follows [5, 7, 8]:

$$\left\langle {}^{2S+1}L_J \middle| \hat{H}_{MD} \middle| {}^{2S+1}L_{J-1} \right\rangle = (-\mu_B) \times \sqrt{\frac{(S+L+J+1)(S+L+1-J)(S-L+J)(L-S+J)}{4J}}$$

$$\left\langle {}^{2S+1}L_J \middle| \hat{H}_{MD} \middle| {}^{2S+1}L_J \right\rangle = (-\mu_B) \times \sqrt{J(J+1)(2J+1)} \left(1 + \frac{J(J+1) + S(S+1) - L(L+1)}{2J(J+1)}\right)$$

$$\left\langle {}^{2S+1}L_J \middle| \hat{H}_{MD} \middle| {}^{2S+1}L_{J+1} \right\rangle = (-\mu_B) \times \sqrt{\frac{(S+L+J+1)(S+L-J)(S-L+J+1)(L-S+J+1)}{4(J+1)}}$$

Using the above expressions, the magnetic dipole line strength for the $^7F_1 \leftrightarrow {}^5D_0$ transition in Eu^{3+} can be calculated as follows:

$$\left\langle {}^7F_1 \middle| \hat{H}_{MD} \middle| {}^7F_0 \right\rangle = -\mu_B 2\sqrt{3}$$

$$\left\langle {}^5D_1 \middle| \hat{H}_{MD} \middle| {}^5D_0 \right\rangle = -\mu_B \sqrt{6}$$

$$S_{MD}\left(Eu^{3+}:{}^7F_1 \leftrightarrow {}^5D_0\right) = 0.1159\mu_B^2$$

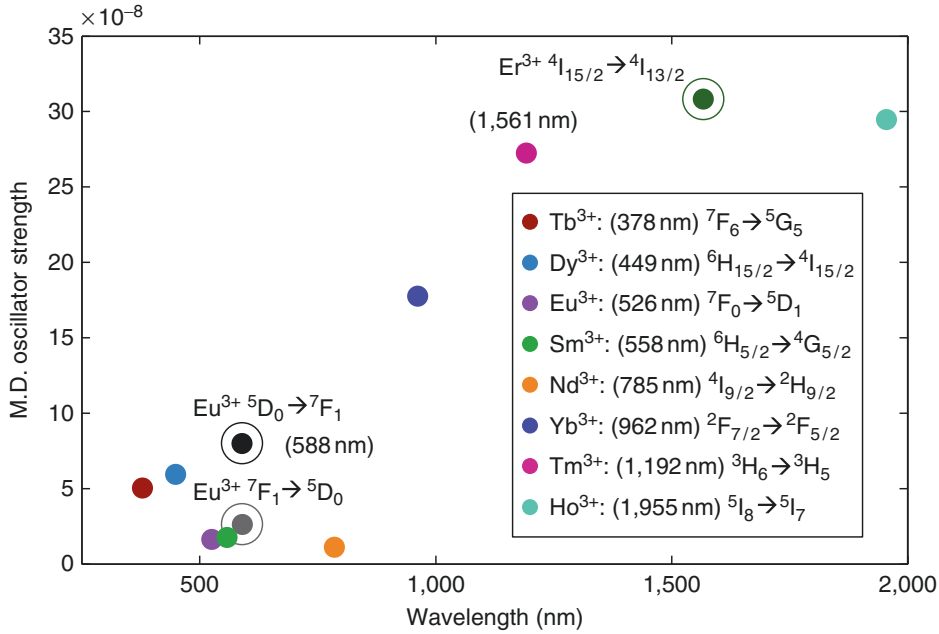
with which the spontaneous emission rate from $^5D_0 \rightarrow {}^7F_1$ and the oscillator strength for absorption from $^7F_1 \rightarrow {}^5D_0$ can be calculated based on the degeneracy $g = 2J + 1$ of the initial state:

$$A_{MD}\left(Eu^{3+}:{}^5D_0 \rightarrow {}^7F_1\right) = \frac{16\pi^3\mu_0}{3h}\left(\frac{n}{\lambda}\right)^3 \frac{0.1159\mu_B^2}{2(0)+1} = 15.4n^3 s^{-1}$$

$$f_{MD}\left(Eu^{3+}:{}^7F_1 \rightarrow {}^5D_0\right) = \frac{8\pi^3 m_e}{3he^2c}\left(\frac{n}{\lambda}\right)\frac{0.1159\mu_B^2}{2(1)+1} = 2.66 \times 10^{-8} n$$

As a reference, Fig. 2 presents the calculated magnetic dipole oscillator strength for a number of Lanthanide ion transitions throughout the visible and near-infrared portion of the electromagnetic spectrum. Although these values may appear small, there are three important considerations that enhance the significance of these terms. First, most observable quantities scale with the product of the oscillator strength and oscillator density, and since these are single ion oscillator strengths, it is possible to pack many ions into a small volume. Second, these values are comparable to the electric dipole oscillator strengths in Lanthanide ions. For example, the electric dipole oscillator strength for the Er^{3+}: $^4I_{15/2} \rightarrow {}^4I_{13/2} \sim 1{,}550$ nm transition is on the order of 2×10^{-6} in vacuum [5]. Third, the magnetic dipole transition probability increases

O



Optical Frequency Magnetic Dipole Transitions, Fig. 2 Lanthanide Magnetic Dipole Transitions in the Visible and Near-Infrared Regime. Absorption line oscillator strength values labeled in the legend box are taken from Carnall et al. [10]. $^7F_1 \rightarrow {}^5D_0$ absorption and $^5D_0 \rightarrow {}^7F_1$ emission lines in Eu^{3+} calculated using eigenvectors from Ofelt [9]. $^4I_{15/2} \rightarrow {}^4I_{13/2}$ absorption line in Er^{3+} calculated using eigenvectors from Weber [5] (all values are based on free-ion calculations in vacuum; within a dielectric host, the oscillator strengths would increase by a factor of n)

significantly with host refractive index. This means that light emission from the 1,550 nm line ($^4I_{13/2} \rightarrow {}^4I_{15/2}$) in Er^{3+} doped LaF_3 is calculated to be 41 % magnetic dipole emission ($A_{MD} = 37.6/s$, $A_{ED} = 54/s$) [5].

Radiation from Electric and Magnetic Dipole Emission

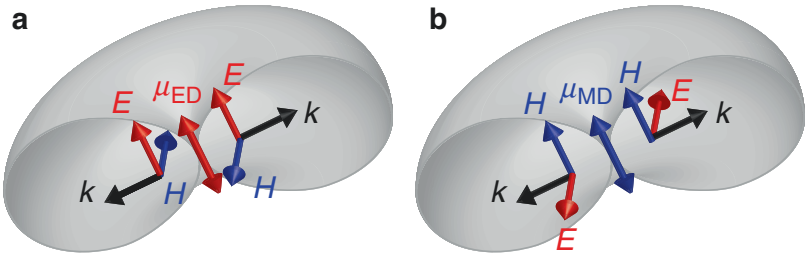
Light emitted by an electric dipole transition can be modeled by the radiation from an oscillating classical electric dipole, i.e., a time-varying infinitesimal current source $\vec{j}\left(\vec{r}, t\right) = \mathrm{Re}\left\{-i\omega\mu_{ED}\exp(-i\omega t)\delta\left(\vec{r}\right)\hat{n}\right\}$. The direction of the oscillation defines the dipole moment orientation ($\hat{n}$), while its frequency defines the frequency of emitted electromagnetic radiation. For an electric dipole moment located at the origin and oriented along the z-axis $\vec{\mu}_{ED}(t) = \mathrm{Re}\{\mu_{ED}\exp(-i\omega t)\hat{z}\}$, the electromagnetic fields in the radiation zone ($kr \gg 1$) are given in spherical coordinates as [2]:

$$\vec{E}_{ED} = E_\theta\hat{\theta} = -\frac{|\mu_{ED}(t)|\sin\theta}{4\pi\varepsilon_r\varepsilon_0}\frac{k^2}{r}e^{ikr}\hat{\theta}$$

$$\vec{H}_{ED} = H_\phi\hat{\phi} = -\sqrt{\frac{\varepsilon_r\varepsilon_0}{\mu_r\mu_0}}\frac{|\mu_{ED}(t)|\sin\theta}{4\pi\varepsilon_r\varepsilon_0}\frac{k^2}{r}e^{ikr}\hat{\phi}$$

where θ is the polar angle measured from the z-axis, ϕ is the azimuthal angle, and $k = (\omega/c)n$ is the wave number in a homogenous dielectric medium with refractive index $n = \sqrt{\varepsilon_r\mu_r}$.

Light emitted from a magnetic dipole transition, on the other hand, resembles the radiation from an oscillating classical magnetic dipole, i.e., a time-varying infinitesimal current loop. The current loop creates a magnetic dipole moment in the direction normal to its surface. For a magnetic dipole moment located at the origin and oriented along the z-axis $\vec{\mu}_{MD}(t) = \mathrm{Re}\{\mu_{MD}\exp(-i\omega t)\hat{z}\}$,



Optical Frequency Magnetic Dipole Transitions, Fig. 3 Schematic representation of fields and radiation pattern emitted by an electric dipole (**a**) and a magnetic dipole (**b**)

the electromagnetic fields in the radiation zone ($kr \gg 1$) are [2]:

$$\vec{E}_{\mathrm{ED}} = E_\phi \hat{\phi} = -\sqrt{\frac{\mu_r \mu_0}{\varepsilon_r \varepsilon_0}} \frac{|\mu_{\mathrm{ED}}(t)| \sin\theta}{4\pi} \frac{k^2}{r} e^{ikr} \hat{\phi}$$

$$\vec{H}_{\mathrm{ED}} = H_\phi \hat{\phi} = -\frac{|\mu_{\mathrm{ED}}(t)| \sin\theta}{4\pi} \frac{k^2}{r} e^{ikr} \hat{\theta}$$

The radiation pattern for electric and magnetic dipoles in the far-field can be obtained by calculating the time-averaged Poynting vector, $\left\langle \vec{S} \right\rangle = \frac{1}{2}\mathrm{Re}\left\{ \vec{E} \times \vec{H}^* \right\}$:

$$\left\langle \vec{S}_{\mathrm{ED}} \right\rangle = \frac{c}{n} \frac{|\mu_{\mathrm{ED}}|^2 \sin^2\theta}{32\pi\varepsilon_r\varepsilon_0} \frac{k^4}{r^2} \hat{r}$$

$$\left\langle \vec{S}_{\mathrm{ED}} \right\rangle = \frac{n}{c} \frac{|\mu_{\mathrm{MD}}|^2 \sin^2\theta}{32\pi\varepsilon_r\varepsilon_0} \frac{k^4}{r^2} \hat{r}$$

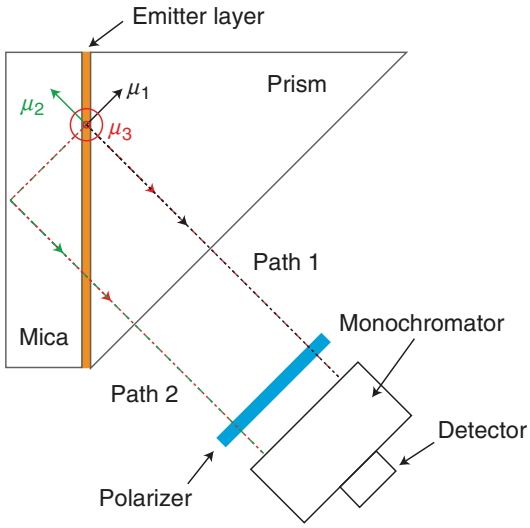
The fields and radiation patterns of electric and magnetic dipole emitters are shown in Fig. 3. It can be seen that both the electric and magnetic dipoles have the same $\sin^2(\theta)$ radiation pattern (shown in gray), but differing field symmetries. While the electric field radiated by an electric dipole has even symmetry with respect to the dipole location, the electric field radiated by a magnetic dipole has odd symmetry. In fact, all the magnetic dipole relations can be obtained from the electric dipole relations by the following transformations [2]:

$$\vec{E}_{\mathrm{ED}} \rightarrow \vec{H}_{\mathrm{MD}} \sqrt{(\mu_r\mu_0)(\varepsilon_r\varepsilon_0)},$$
$$\vec{E}_{\mathrm{ED}} \sqrt{(\mu_r\mu_0)(\varepsilon_r\varepsilon_0)} \rightarrow$$
$$-\vec{E}_{\mathrm{MD}}, \text{and } \mu_{\mathrm{ED}} \rightarrow \mu_{\mathrm{ED}}(n/c).$$

Identify Magnetic Transitions

Despite similar radiation patterns, magnetic and electric dipole emitters have different symmetries with respect to polarization as shown in Fig. 3. Measuring this polarization difference has been the prevalent method to experimentally distinguish electric and magnetic dipole radiation. In atomic gases, the Zeeman effect has been used to lift the degeneracy of the energy states and map the polarization of radiative transitions between different energy states with respect to the applied magnetic field direction [11]. In condensed matter systems, polarization differences combined with interference effects have been used to identify electric and magnetic dipole transitions [11–13].

Figure 4 shows the wide-angle interference experiment first used by Selenyi to confirm the dipole nature of light emission [12]. A very thin layer of emitters (thickness < wavelength of emission) is placed between a prism and thin (~10 μm think) sheet of mica. Assuming that light is emitted by either electric or magnetic dipole transitions, one can decompose all emission into the three orthogonal dipole orientations shown in Fig. 4 as μ_1, μ_2 and μ_3. Light emitted from these dipolar emitters can reach the detector (originally a photographic film) either directly (via path 1) or after reflection from the air–mica interface (via path 2). Since a dipole emitter does not emit in the direction of the dipole orientation, a dipole oriented as μ_1 (μ_2) will only emit via path 1 (path 2). However, a dipole oriented along μ_3 will emit via both paths and interfere at the detector. Thus, any observed interference is due solely to the μ_3 oriented dipole component of emission, and the introduction of a linear polarizer into the detection path allows one to identify whether the dipole is electric or magnetic in nature. For light emitted by an electric dipole oriented along μ_3, the electric field is polarized out-of-plane, whereas for a magnetic dipole oriented along μ_3, the electric



Optical Frequency Magnetic Dipole Transitions, Fig. 4 Wide-angle interference setup used to identify dipolar emission lines

field is polarized in-plane. Thus, when light is emitted by a pure electric dipole, one sees interference fringes for an out-of-plane oriented polarizer, and no fringes for an in-plane oriented polarizer. This was the original experiment performed by Seleyni [12]. On the other hand, when light is emitted by a pure magnetic dipole, one sees interference fringes for an in-plane-oriented polarizer and no fringes for an out-of-plane-oriented polarizer. This was the modified experiment used by Freed and Weissmann to demonstrate that the 588 nm emission in Eu^{3+} (from the $^5D_0 \rightarrow {}^7F_1$ line) is predominantly a magnetic dipole transition [13].

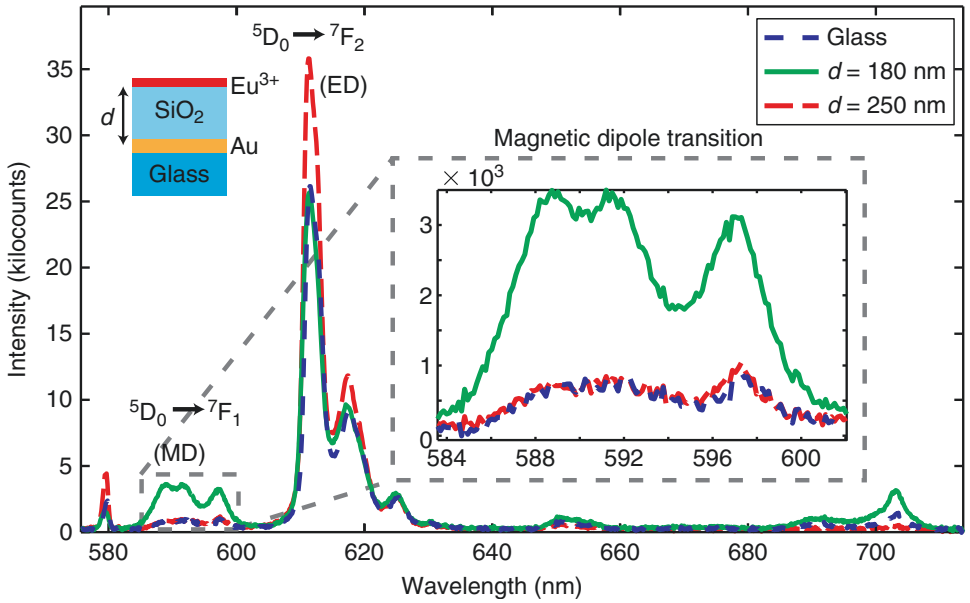
Applications

For decades, Lanthanide ions have served as important light emitters in a broad range of optical technologies from lighting and color displays to solid-state lasers and fiber-optic amplifiers. For example, Eu^{3+} and Tb^{3+} are common red and green phosphors in fluorescent light bulbs and phosphorescent white-light LEDs. Er^{3+} sensitized with Yb^{3+} is the gain media for 1,550 nm fiber amplifiers used in telecommunication networks.

Interestingly, the same shielded electronic structure that makes Lanthanide ions highly efficient (high quantum-yield) light sources also makes them ideal magnetic dipole emitters. Due to the filling rules for electron energy levels, the valence 4*f* electrons in Lanthanide ions are shielded by complete 5 *s* and 5*p* orbitals [6]. These closed shells protect excited electrons from non-radiative scattering events, thereby increasing the likelihood of radiative decay. This shielding also provides the time necessary to observe "forbidden" intra-4*f* magnetic dipole transitions. Although they are not always discussed in the literature, magnetic dipole transitions make significant contributions to the 543 nm line in Tb^{3+}, 980 nm line in Yb^{3+}, and 1,550 nm line in Er^{3+}.

Lanthanide emitters and their magnetic dipole transitions have also played a major role in the study of fundamental light-matter interactions. Drexhage used Eu^{3+} in his pioneering research on modified spontaneous emission near planar interfaces [14]. These studies were among the first detailed demonstrations of the Purcell effect at optical frequencies. Building on this work, Kunz and Lukosz observed clear distinctions in the modified emission rates and radiation patterns of the electric and magnetic dipole transitions in Eu^{3+} near interfaces [15]. These differences, based on self-interference effects, result from the different field symmetries of electric and magnetic dipole emitters. More recently, Eu^{3+} has been used to study local field effects and test the validity of different cavity models for electric and magnetic dipole transitions [16].

The electric and magnetic dipole transitions in Eu^{3+} have also been used extensively by physical chemists to investigate material properties. For example, the $^5D_0 \rightarrow {}^7F_1$ magnetic dipole transition has been used to probe crystal- and ligand-field effects in different host matrices [6]. In addition, the ratio between electric and magnetic dipole emission has been used to gauge the local symmetry. Unlike the $^5D_0 \rightarrow {}^7F_1$ magnetic dipole transition which is only weakly dependent on the host matrix, the $^5D_0 \rightarrow {}^7F_2$ is a hypersensitive electric dipole transition that requires a noncentrosymmetric ion site. Similarly, due to the relative



Optical Frequency Magnetic Dipole Transitions, Fig. 5 Modified emission spectra of Eu^{3+}chelate thin films at varying distances from a planar gold mirror. *Dashed blue line* shows reference spectra on a glass substrate without the gold mirror. Magnetic dipole emission is enhanced when Eu^{3+} ions are separated from the mirror by 180 nm spacer layer of SiO_2 (*green line*), whereas electric dipole emission is enhanced by a 250 nm spacer layer (*dashed red line*)

independence of the $^5D_0 \rightarrow {}^7F_1$ magnetic dipole transitions rate on the local chemical environment, it has been used as an internal reference to infer the quantum efficiency of Eu^{3+} in different materials [17].

Future Directions for Research

With the advent of metamaterials that support artificial magnetic resonances, there has been renewed interest in the natural magnetic resonances of Lanthanide ions. For example, Thommen and Mandel recently proposed that the electric and magnetic dipole transitions in Lanthanides could be used to realize homogeneous materials with simultaneously negative permittivity and permeability [18]. Their calculations suggest that an Er^{3+} doped crystal (i.e., a simple homogenous medium without the need for nanofabrication) could exhibit tunable, negative index response by using a weak magnetic field to split the degenerate electric and magnetic dipole transitions in the $\sim$ 1,550 nm $^4I_{15/2} \rightarrow {}^4I_{13/2}$ line. They also suggested other mixed transitions, including the 5 μm $^4I_{9/2} \rightarrow {}^2I_{11/2}$ transition in Nd^{3+} and the 470 nm $^5I_8 \rightarrow {}^3K_8$ transition in Ho^{3+}, that could support similar effects. Although their proposal has several practical limitations in terms of operating temperature, ion concentrations, and bandwidth, it nevertheless highlights one potential application of natural magnetic dipole transitions in the field of metamaterials. Another proposed application is the use of magnetic dipole transitions as local probes for optical frequency magnetic fields. Noginova and colleagues have suggested that the modified spontaneous emission from the $^5D_0 \rightarrow {}^7F_1$ magnetic dipole transition in Eu^{3+} could be used to study metamaterials and their optical frequency artificial magnetic resonances [19].

Modifying the spontaneous emission of magnetic dipole transitions also holds promise for new ways to tune and modulate Lanthanide-based light sources. By exploiting the coupled nature and differing field symmetries of electric and magnetic dipole transitions in Eu^{3+}, recent work has shown how self-interference effects in this

multilevel system can be used to selectively enhance electric and magnetic dipole emission [20]. By controlling the distance between a thin film of Eu^{3+} chelates and a planar gold mirror, these experiments demonstrated a greater than fourfold enhancement in the emission intensity of the $^5D_0 \rightarrow {}^7F_1$ magnetic dipole transition. This strong magnetic dipole enhancement, as shown in Fig. 5, results from self-interference effects that simultaneously inhibit the competing $^5D_0 \rightarrow {}^7F_2$ electric dipole transition while enhancing the magnetic dipole transition. Karaveli and Zia [20] show how simple optical structures can be used to tune emission spectra in multilevel ions with distinct electric and magnetic transitions, such as Eu^{3+}, and also suggest that similar structures could be used to tune emission polarization in multilevel ions with degenerate electric and magnetic transitions, such as the 1,550 nm line in Er^{3+}.

Cross-References

- ▶ Active Plasmonic Devices
- ▶ Light Localization for Nano-Optical Devices
- ▶ Nanostructures for Photonics
- ▶ Theory of Optical Metamaterials

References

1. Shalaev, V.M.: Optical negative-index metamaterials. Nat. Photonics **1**, 41–48 (2007)
2. Novotny, L., Hecht, B.: Principles of Nano-Optics. Cambridge University Press, Cambridge (2007)
3. Cowan, R.D.: The Theory of Atomic Structure and Spectra. University of California Press, Berkeley/Los Angeles/London (1981)
4. Thorne, A.P., Litzen, U., Johansson, S.: Spectrophysics: Principles and Applications. Springer, Berlin/Heidelberg/New York (1999)
5. Weber, M.J.: Probabilities for radiative and nonradiative decay of Er^{3+} in LaF_3. Phys. Rev. **157**, 262–272 (1967)
6. Görller-Walrand, C., Binnemans, K.: Spectral intensities of f-f transitions. In: Gschneidner Jr., K.A., Eyring, L. (eds.) Handbook on the Physics and Chemistry of Rare Earths, vol. 25, pp. 101–264. North-Holland, Amsterdam (1998)
7. Carnall, W.T., Fields, P.R., Wybourne, B.G.: Spectral intensities of the trivalent lanthanides and actinides in solution. I. Pr^{3+}, Nd^{3+}, Er^{3+}, Tm^{3+} and Yb^{3+}. J. Chem. Phys. **42**, 3797–3806 (1965)
8. Wybourne, B.G., Smentek, L.: Optical Spectroscopy of Lanthanides: Magnetic and Hyperfine Interactions. CRS Press/Taylor & Francis, Boca Raton (2007)
9. Ofelt, G.S.: Structure of the f^6 configuration with application to rare-earth ions. J. Chem. Phys. **38**, 2171–2180 (1963)
10. Carnall, W.T., Fields, P.R., Rajnak, K.: Spectral intensities of the trivalent lanthanides and actinides in solution. II. Pm^{3+}, Sm^{3+}, Eu^{3+}, Gd^{3+}, Tb^{3+}, Dy^{3+}, and Ho^{3+}. J. Chem. Phys. **49**, 4412–4423 (1968)
11. Mrozowski, S.: Forbidden lines in the laboratory. Rev. Mod. Phys. **16**, 153–174 (1944)
12. Selenyi, P.: Wide-angle interference and the nature of elementary light sources. Phys. Rev. **56**, 477–479 (1939)
13. Freed, S., Weissman, S.I.: Multipole nature of elementary sources of radiation – wide-angle interference. Phys. Rev. **60**, 440–442 (1941)
14. Drexhage, K.H.: Interaction of light with monomolecular dye layers. Prog. Opt. **12**, 162–231 (1974)
15. Kunz, R.E., Lukosz, W.: Changes in fluorescence lifetimes induced by variable optical environments. Phys. Rev. B **21**, 4814–4828 (1980) and references therein
16. Rikken, G.L.J.A., Kessener, Y.A.R.R.: Local field effects and electric and magnetic dipole transitions in dielectrics. Phys. Rev. Lett. **74**, 880–883 (1995)
17. Eliseeva, S.V., Bünzli, J.-C.G.: Lanthanide luminescence for functional materials and bio-sciences. Chem. Soc. Rev. **39**, 189–227 (2010)
18. Thommen, Q., Mandel, P.: Left-handed properties of erbium-doped crystals. Opt. Lett. **31**, 1803–1805 (2006)
19. Noginova, N., Zhu, G., Mavy, M., Noginov, M.A.: Magnetic dipole based systems for probing optical magnetism. J. Appl. Phys. **103**, 07E901 (2008)
20. Karaveli, S., Zia, R.: Strong enhancement of magnetic dipole emission in a multilevel electronic system. Opt. Lett. **35**, 3318–3320 (2010)

Optical Lithography

- ▶ Photolithography

Optical Lithography by Liquid Immersion

- ▶ Immersion Lithography

Optical Properties

► Tight-Binding Simulations of Nanowires

Optical Properties of Metal Nanoparticles

Nadine Harris, Martin G. Blaber and George C. Schatz
Department of Chemistry and International Institute for Nanotechnology, Northwestern University, Evanston, IL, USA

Synonyms

Local surface plasmon resonance (LSPR); Nanooptics, nanoplasmonics; Optical response of metal particles; Surface plasmon nanophotonics

Definition

The optical properties of metal nanoparticles are governed by collective excitation of the conduction electrons. The electrons oscillate coherently under irradiation by light in the visible and near-IR regions of the electromagnetic spectrum. These coherent oscillations are known as local surface plasmon resonances. The wavelength of excitation of the plasmon resonances shows strong variation with the size, shape, and dielectric environment of the particles. Modeling this variation by solving Maxwell's equations for the interaction of light with the nanoparticles provides a useful tool for the design and interpretation of experiments. The entry describes traditional analytical and numerical methods used to solve Maxwell's equations, and it presents representative results of calculations using these methods. Firstly, electrostatic methodologies are discussed and applied to the modeling of spheres, shells, and ellipsoids with dimensions considered in the quasistatic (small particle limit) regime. Mie theory is then introduced and used to model the full electromagnetic optical solution for a sphere with dimensions considered outside of the quasistatic regime. Finally, the numerical methodology of the discrete dipole approximation technique is introduced and implemented to model the optical properties of gold rods and prisms with varying aspect ratios.

Overview

The field of plasmonics aims to exploit the unique optical properties of metallic nanostructures to control and manipulate light at the nanoscale. The optical response of these structures is unique because, unlike dielectrics, metals have a negative dielectric function, which upon irradiation with light, enables them to support collective electron excitations known as localized surface plasmon resonances (LSPR) [1, 2]. As the name suggests, these resonances exist between the metal surface and the surrounding dielectric, and not only produce extinction cross sections that exceed their geometric cross sections, but also produce extremely high local electric field intensities. One of the enormous advantages of plasmonic nanoparticles lies in the ability to tune the position of their LSPR from the visible (in the case of gold spheres) to the infrared (IR) part of the spectrum (in the case of anisotropic particles such as gold rods and prisms). Moreover, changing the geometry of the nanostructure allows unprecedented control over the location and magnitude of the concentrated electric fields on the surface of the nanoparticle itself. Further control can be achieved by varying the polarization of the incident light with respect to nanostructure geometry. This has motivated abundant interest in the synthesis of geometrically varying nanometallic structures such as spheres, rods, triangular prisms, and a plethora of other anisotropic particles where the aim is to produce monodisperse particles with precise control over the wavelength of their highly absorbing or scattering LSPR. This capability to control LSPR wavelength is important to a variety of technologies for chemical and biological sensing, where the change of wavelength in the presence of analytes provides the sensing modality. In addition, control of LSPR wavelength is also useful in surface-enhanced Raman scattering (SERS)

studies, where enhanced fields at the surfaces of the nanoparticles leads to enhanced Raman scattering that can be used to identify adsorbed molecules.

Electrodynamic modeling, i.e., solving Maxwell's equations [3] for the interaction of light with the nanoparticles, contributes significantly to the development of this research because the optical response of these particles is well described with classical (continuum) theory. This allows the calculation of the particle extinction cross sections, in other words how much light is absorbed and scattered by the particle, as well as the plasmonically enhanced electric field that is important in SERS.

This entry will focus on both analytical and numerical methods that are used to model the electrodynamic properties of noble metal nanoparticles. Only isolated nanoparticles are considered in this discussion, but it should be noted that arrays of nanoparticles, as well as other nanostructures such as nanostructured metal films are also of great interest to the plasmonics field.

The choice of using either an analytical or a numerical method to solve Maxwell's equations depends on the geometry of the particle to be considered. Spherically symmetric particles such as spheres can be modeled using an analytical solution based on Mie [4] theory. On the other hand, anisotropic particles such as rods and prisms need to be modeled using numerical, grid-based methods such as the discrete dipole approximation (DDA) [5] or the finite-difference time-domain (FDTD) [6] method. This pedagogical entry will focus specifically on the standard Mie and DDA methods because they are both extremely versatile and very well known within the electrodynamics community. The Mie, DDA, and FDTD (and other) methods have advantages and disadvantages over one another, and for further details on these techniques as well as applications that extend beyond what one could describe as the "standard" use of the techniques, the authors refer the reader to the following citations [7, 8].

Section "The Dielectric Function" details the Drude model, which is used to describe the dielectric function of metals, particularly the dependence of the dielectric function on frequency, leading to a simple analytical description for the origin of plasmon resonances, and the dependence of resonance frequency on particle structure. Section "Analytical Methods to Solve Maxwell's Equations" describes the mathematical basis behind the electrostatic analytical techniques used to model the extinction spectra of spheres, shells (coated spheres), and ellipsoids. The full electromagnetic solution (Mie theory) for a sphere is also described in this section. Finally, in section "Numerical Methods to Solve Maxwell's Equations," the mathematical formalism of the numerical DDA method is detailed and applied to calculate the optical properties of gold rods and triangular prisms.

The Dielectric Function

The electromagnetic response of metal particles is governed by the dielectric function of the metal from which they are made. A commonly used model for the complex dielectric function, ($\varepsilon = \varepsilon_{\text{real}}^{\text{Drude}} + i\varepsilon_{\text{imaginary}}^{\text{Drude}}$) of a free-electron gas was described by Drude [1] using:

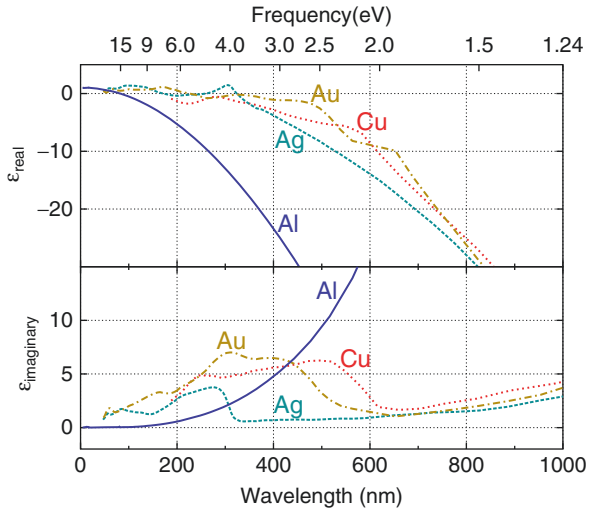
$$\varepsilon_{\text{real}}^{\text{Drude}}(\omega) = 1 - \frac{\omega_{\text{p}}^2}{\omega^2 + \gamma^2} \tag{1a}$$

$$\varepsilon_{\text{imaginary}}^{\text{Drude}}(\omega) = 1 - \frac{\omega_{\text{p}}^2 \gamma}{\omega(\omega^2 + \gamma^2)} \tag{1b}$$

Here the plasma frequency, ω_{p}, determines the dispersion characteristics of the metal and can be determined from the density of free electrons, N, using

$$\omega_{\text{p}} = \sqrt{\frac{Ne^2}{\varepsilon_0 m_{\text{e}}}} \tag{2}$$

where e is the charge on the electron, m_{e} is electron mass, and ε_0 the permittivity, or dielectric function, of free space. The damping constant, γ, describes the way electrons scatter off ions, other electrons, defects, grain boundaries, and interfaces in the system. The real part of the dielectric function describes the motion of the free and



Optical Properties of Metal Nanoparticles, Fig. 1 Experimentally measured dielectric functions of aluminum, copper, silver, and gold. Aluminum exhibits pure Drude behavior, albeit with a large electron–phonon scattering component (large γ). Copper and gold have interband transitions in the visible (which appear as deviation from Drude behavior in $\varepsilon_{imaginary}$) giving them their distinct bulk colors. Silver has interband transitions in the ultraviolet

bound charges in a material. The imaginary part describes the energy *loss* mechanisms in the sample. The loss mechanisms can be broadly cast into two categories: the first category is frequency independent terms, which include electron–phonon, electron–defect, and electron-surface scattering. The electron-surface scattering term can be included as an addition to the experimentally measured dielectric function to account for the finite size of the nanoparticle. This is important when the mean free path of the electrons is approximately greater than the size of the nanoparticle. These frequency independent terms can be included in the Drude phenomenological scattering rate [2]:

$$\gamma = \gamma_{\text{phonon}} + \gamma_{\text{defect}} + \gamma_{\text{surface}}(a) \qquad (3)$$

where a is the size of the particle. The frequency dependent mechanisms include electron–electron scattering (which for metals is quite small compared to the electron–phonon scattering term) and interband transitions. Interband transitions occur when the incident photon excites an electron from an occupied to unoccupied state in the metal. Since interband transitions are not well modeled using a Drude function, the total experimental dielectric function is usually decomposed into a Drude contribution and an interband contribution:

$$\varepsilon^{\text{experimental}}(\omega) = \varepsilon^{\text{Drude}}(\omega) + \varepsilon^{\text{interband}}(\omega) \qquad (4)$$

Figure 1 shows the experimental dielectric functions for aluminum, copper, silver, and gold. In the UV and visible, aluminum exhibits pure Drude character with no interband transitions. Gold and copper have a large interband contribution in the visible, giving them their distinct bulk colors. The interband transitions in silver occur in the UV, and hence bulk silver is highly reflective across the visible.

Properties of Resonances in Terms of the Dielectric Function

When a photon in vacuum impinges on a flat surface with a negative permittivity and low loss (imaginary permittivity), it is reflected unless certain boundary conditions are met. By introducing surface roughness, a grating, or a prism onto the surface, the boundary conditions are modified and allow for the excitation of surface plasmon polaritons (SPP). The SPP manifests as a free-electron charge oscillation that propagates along the surface of the metal.

For metal nanoparticles, the geometry of the particle defines boundary conditions for Maxwell's equations that allow for excitation of local surface plasmon resonances (LSPRs). In this case, the free charges in the particle are displaced from equilibrium by the incident photon. A restoring force arises from the attraction of the electrons to the now positively charged ions. The extent of the displacement, and hence the restoring force, depends on the geometry of the particle and the real part of the dielectric function of the metal. When the motion of the electrons in the particle and the applied field are out of phase by $\pi/2$, all of the energy contained in the applied field can be used to accelerate the electrons, creating a resonance. To maintain causality, the imaginary part of the dielectric function of the metal must always be positive (i.e., it takes finite time for the electrons to respond to the incident field). This positive value introduces a phase lag between the oscillating electrons and the applied field. This limits the rate at which energy can be coupled into the nanoparticle, determining the maximum field strength and the maximum extinction cross section.

It is common to describe the portion of light that interacts with a particle by calculating the extinction *efficiency*. This is the extinction cross section [1] ($C_{\text{ext}} = C_{\text{abs}} + C_{\text{sca}}$, with C_{abs} and C_{sca} the absorption and scattering cross sections, respectively) normalized to the cross-sectional area, A, of the nanoparticle ($C_{\text{ext}}/A = Q_{\text{ext}} = Q_{\text{abs}} + Q_{\text{sca}}$). The importance of scattering versus absorption depends heavily on the application of interest. For plasmonic heating applications, such as photothermal therapy [9], the absorption efficiency determines all of the important properties of the system, and scattering should be minimized to avoid wasting energy. However, for SERS and surface-enhanced fluorescence, the total extinction should be maximized. This is because the total field in the vicinity of the particle is comprised of the absorption limited evanescent field which decays exponentially with distance from the surface, and the radiating or scattered field, which couples energy out of the system.

Analytical Methods to Solve Maxwell's Equations

The Quasistatic Approximation

For particles very much smaller than the wavelength of light, the applied field is effectively constant over the surface of the particle. Because the field is assumed to be spatially uniform but oscillatory in time, this regime is called "quasi"-static. This neglect of retardation not only simplifies Maxwell's equations, but also simplifies the analysis of the general properties of LSPRs for a variety of particle shapes. In particular, for each shape and polarization direction, there is a value of the dielectric function that allows for the excitation of a LSPR. In the quasistatic approximation, this resonance condition only depends on the wavelength through the dielectric function.

Spheres in the Quasistatic Limit: A Derivation

As an example of the simplicity of the quasistatic approximation, a summary is provided of the derivation of the polarizability of a small sphere of radius a with a dielectric function of the form with a dielectric function of the form $\varepsilon = \varepsilon_1 + i\varepsilon_2$, embedded in a non-absorbing medium with dielectric constant ε_{m}. Such a derivation can be found in many electrodynamics text books; here the derivation follows Ohanian [10]. The fundamental equation of quasistatics is Laplace's equation, $\nabla^2\Phi = 0$, where Φ is the scalar potential. For the sphere in a medium, it is appropriate to introduce two scalar potentials, one inside the particle, $r < a$, and the other outside the particle, $r > a$. Omitting the azimuthal angle ϕ (for symmetry reasons), the fields associated with these potentials are:

$$\mathbf{E}_1 = -\nabla\Phi_1(r,\theta), \qquad \mathbf{E}_2 = -\nabla\Phi_2(r,\theta) \quad (5)$$

where the subscript 1 indicates inside and 2 indicates outside the sphere. These potentials obey Laplace's equation in the region in which they are defined:

$$\nabla^2\Phi_1(r, < a, \theta) = 0, \quad \nabla^2\Phi_2(r > a, \theta) = 0 \quad (6)$$

In order to extract useful information from these equations, the scalar potential is required

to satisfy several boundary conditions. The normal component of the electric displacement field ($\mathbf{D} = \varepsilon\mathbf{E}$) must be continuous across the boundary of the particle, and so the following boundary condition is introduced:

$$\varepsilon \frac{\partial \Phi_1}{\partial r}\bigg|_{r=a} = \varepsilon_{\mathrm{m}} \frac{\partial \Phi_2}{\partial r}\bigg|_{r=a} \tag{7}$$

Analogously, the parallel component of the electric field must be continuous at the boundary. This boundary condition is included in the more general expression:

$$\Phi_1(a,\theta) = \Phi_2(a,\theta) \tag{8}$$

The final boundary condition is that the potential outside the particle must reduce to the potential of the incident field at an infinite distance from the sphere (here assuming that the polarization is along the z axis):

$$\lim_{r\to\infty} \Phi_2(r,\theta) = -E_0 r\cos(\theta) \tag{9}$$

Solutions of Laplace's equation in spherical coordinates take the form of the Legendre Polynomials, P_n, of order n:

$$\begin{aligned}\Phi_1(r,\theta) &= \sum_{n=1}^{\infty} B_n r^n P_n(\cos(\theta)),\\ \Phi_2(r,\theta) &= -E_0 r\cos(\theta) + \sum_{n=1}^{\infty} A_n \frac{P_n(\cos(\theta))}{r^{n+1}}\end{aligned} \tag{10}$$

where the coefficients A and B are unknown. Substituting Eq. 10 into the boundary conditions (7) and (8) provides two sets of linear equations:

$$\begin{aligned}&\varepsilon_{\mathrm{m}}\left(-E_0\cos(\theta) + \sum_{n=1}^{\infty}(n+1)A_n\frac{P_n(\cos(\theta))}{a^{n+2}}\right)\\ &\quad = \varepsilon\sum_{n+1}^{\infty} nB_n a^{n-1}P_n(\cos(\theta))\end{aligned} \tag{11}$$

and

$$\begin{aligned}&-E_0 a\cos(\theta) + \sum_{n=1}^{\infty} A_n \frac{P_n(\cos(\theta))}{a^{n+1}}\\ &\quad = \sum_{n=1}^{\infty} B_n a^n P_n(\cos(\theta))\end{aligned} \tag{12}$$

Solving A_1 in terms of the incident field $(-E_o\cos(\theta))$ forces all the higher-order terms to be zero, and gives the following potentials:

$$\Phi_1(r<a,\theta) = -E_0 r\cos(\theta)\frac{3\varepsilon_{\mathrm{m}}}{\varepsilon + 2\varepsilon_{\mathrm{m}}} \tag{13a}$$

$$\Phi_2(r>a,\theta) = -E_0 r\cos(\theta)\left(1 - \frac{a^3}{r^3}\frac{\varepsilon - \varepsilon_{\mathrm{m}}}{\varepsilon + 2\varepsilon_{\mathrm{m}}}\right) \tag{13b}$$

The second term in Eq. 13b is the potential associated with a point dipole. Combining this with Eq. 5, the field outside the particle can be written:

$$\mathbf{E}_2 = \mathbf{E}_{\mathrm{inc}} + \mathbf{p} = \mathbf{E}_{\mathrm{inc}} + \alpha\mathbf{E}_{\mathrm{inc}} \tag{14}$$

This gives the polarizability of the sphere in the background medium:

$$\alpha = 4\pi a^3 \frac{\varepsilon - \varepsilon_m}{\varepsilon + 2\varepsilon_m} \tag{15}$$

There are also higher-order solutions that arise when terms in the incident field higher than the constant term are included. By solving the A_2 term for the incident field, the so-called quadrupole mode is excited. These higher-order resonances are discussed in the following section.

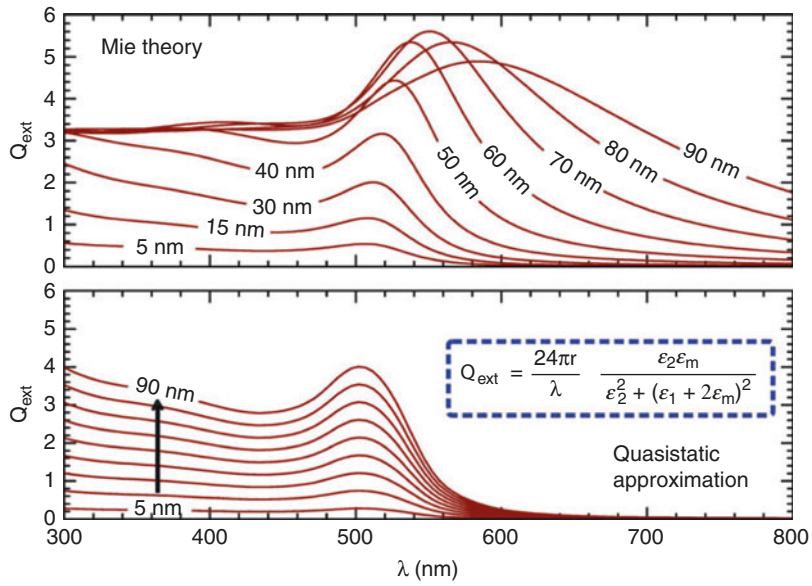
To determine the extinction and scattering cross sections of the sphere described by Eq. 15, the following relations are introduced:

$$C_{\mathrm{ext}} = k\mathrm{Im}(\alpha) \tag{16a}$$

$$C_{\mathrm{sca}} = \frac{k^4}{6\pi}|\alpha|^2 \tag{16b}$$

where $k = 2\pi/\lambda$ is the wavenumber in the medium and λ is the wavelength. Equation 16a is only valid in the limit of small scattering, and so it is sometimes written $C_{\mathrm{abs}} = k\mathrm{Im}(\alpha)$ [1]. To

O



Optical Properties of Metal Nanoparticles, Fig. 2 A comparison of the extinction efficiency of gold nanospheres in vacuum using (*top panel*) Mie's fully retarded solution to Maxwell's equations and (*bottom panel*) the quasistatic approximation. The spectra are labeled with the particle radius

determine the extinction efficiency of the nanosphere, the real and imaginary parts of the dielectric function are substituted into Eq. 15, and then the imaginary part is selected:

$$\alpha = 4\pi a^3 i \frac{3\varepsilon_2 \varepsilon_m}{\varepsilon \frac{2}{2} + (\varepsilon_1 + 2\varepsilon_m)^2} \tag{17}$$

Substituting this into Eq. 16 and dividing by the cross section (πa^2) gives:

$$Q_{\text{ext}} = \frac{24\pi a}{\lambda} \frac{\varepsilon_2 \varepsilon_m}{\varepsilon_2^2 + (\varepsilon_1 + 2\varepsilon_m)^2} \tag{18}$$

This expression has several interesting properties. The extinction efficiency grows larger for increasing sphere radius and contains no information about the range of particle radii for which it is valid. On resonance, the real part of the dielectric function is just $\varepsilon_1 = -2\varepsilon_m$. At the resonance wavelength, λ_p, the strength of the resonance is limited by how small the imaginary part of the dielectric function is, and how large the medium index is. This can be seen by substituting $\varepsilon_1 = -2\varepsilon_m$ into Eq. 18, which, after simplifying, gives:

$$Q_{\text{ext}} = (24\pi a/\lambda)(\varepsilon_m/\varepsilon_2) \tag{19}$$

Figure 2 shows the extinction efficiency of gold nanospheres of increasing radius in vacuum ($\varepsilon_m = 1$). To assess the size regime where the quasistatic approximation is valid, the fully retarded Mie solution is also included (Mie theory is discussed in the next section). There is a strong resonance at around 500 nm and a tail that extends into the UV. The resonance at 500 nm is the dipolar LSPR, which occurs when $\varepsilon_1 = -2$. The UV tail is caused by interband transitions in this region of the spectrum (see Fig. 1). As predicted by Eq. 18, increasing the particle size in the quasistatic approximation proportionally increases the extinction efficiency.

Nanospheres: A Fully Retarded Solution to Maxwell's Equations

Figure 2 shows that the quasistatic approximation predicts the resonance position to within 20 nm for particles with radii up to about 40 nm. Beyond $a = 40$ nm, retardation of the field plays a role in the resonance position and causes a red-shift of the LSPR. To fix this, Mie theory is used: An analytical solution to Maxwell's equations for spheres that was developed by Gustav Mie in 1908 [4]. It fully describes the effects of retardation, including radiative damping and higher-order multipole resonances.

In a generalization of Mie's solution, the extinction and scattering efficiencies are described in terms of a size parameter $x = 2\pi a/\lambda$ and two coefficients a_n and b_n, which cover electric and magnetic partial waves of multipole order n: [1]

$$Q_{\text{ext}} = \frac{2}{x^2}\sum_{n=1}^{\infty}(2n+1)\text{Re}[a_n + b_n] \tag{20a}$$

$$Q_{\text{sca}} = \frac{2}{x^2}\sum_{n=1}^{\infty}(2n+1)\left(|a_n|^2 + |b_n|^2\right) \tag{20b}$$

The coefficients are given by:

$$a_n = \frac{m\psi_n(mx)\psi_n'(x) - \psi_n(x)\psi_n'(mx)}{m\psi_n(mx)\xi_n'(x) - \xi_n(x)\psi_n'(mx)} \tag{21a}$$

$$b_n = \frac{\psi_n(mx)\psi_n'(x) - m\psi_n(x)\psi_n'(mx)}{\psi_n(mx)\xi_n'(x) - m\xi_n(x)\psi_n'(mx)} \tag{21b}$$

where ψ_n and ξ_n are spherical Ricatti–Bessel functions and primes indicate taking a derivative with respect to the argument. The resonance condition is now much more complex in the sense that it is an explicitly size-dependent quantity, even for the dipole mode ($n = 1$). Bohren and Huffman [1] developed a "finite size correction" to adjust the position of the resonance given by the quasistatic approximation. The general principle is that when the denominator of Eq. 21a goes to zero, a surface plasmon is excited. So, by expanding the Ricatti–Bessel functions to order x^2, an additional term appears that can easily be added into Eq. 18. The result is:

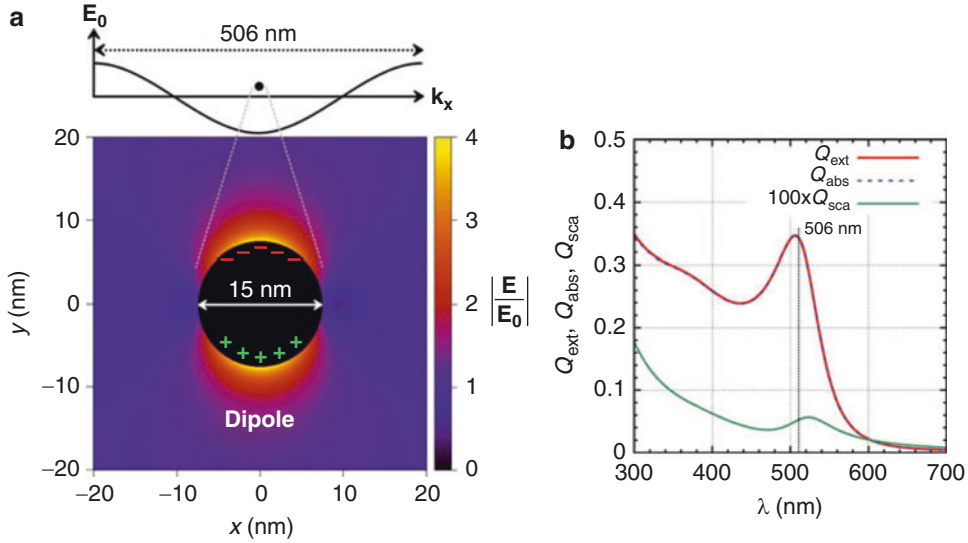
$$Q_{\text{ext}} = \frac{24\pi a}{\lambda}\frac{\varepsilon_2\varepsilon_m}{\varepsilon_2^2 + (\varepsilon_1 + 2\varepsilon_m + 12x^2/5)^2} \tag{22}$$

It is clear that this additional term slowly red-shifts the resonance as the particle size is increased. Including a term proportional to x^3 in the denominator in Eq. 22 leads to a theory called the Modified Long Wavelength approximation [11] in which the plasmon is both shifted and broadened with increasing particle size. These are general features of all nanoparticle plasmonics: Larger particles have red-shifted and broadened resonances. This is evident in Fig. 2, and the broadening is usually ascribed as due to radiative effects where the energy incident on the particle is re-emitted (i.e., scattered) with a high enough rate that plasmon excitation is partially quenched.

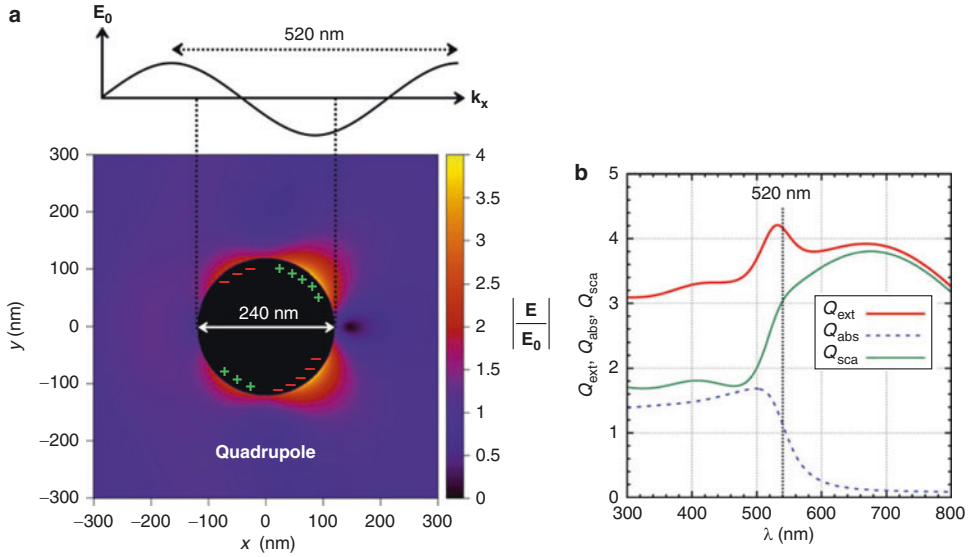
So far the quadrupole ($n = 2$) and higher-order resonances that appear for larger particles have not been discussed. Figures 3 and 4 show the fields around a small (7.5-nm radius) and large (120-nm radius) gold sphere. For the small sphere, the only mode that appears is the dipole mode, as discussed in the quasistatic regime. In this case, the field on the surface of the sphere is almost 90° out of phase with the incident wave, and a large portion of the photon energy is used to accelerate the electrons (this can be seen in Fig. 3b, by comparing the absorption and scattering efficiencies). The large sphere exhibits two modes: a red-shifted, broadened dipole mode at a wavelength of 670 nm and a quadrupole mode at around 520 nm. The dipole mode is broadened due to radiative effects as noted above. For the quadrupole mode, half of the particle is polarized parallel to the incident photon, and half is polarized antiparallel. Both include a phase factor that is not quite 90°. It is easy to assume that this might only occur when the diameter of the particle is comparable to half the wavelength. However, this is not the case and multipole solutions exist even for small particles, they are just dwarfed by the strength of the dominant dipole mode and do not appear in spectra.

Shifting the Surface Plasmon Resonance Wavelength

In the previous section, it was noted that increasing the size of a nanoparticle red-shifts the surface plasmon resonance. As the radius of a sphere is increased, the extinction efficiency passes through a maximum for some finite radius (compare the extinction for 70 and 80 nm spheres in Fig. 2). Beyond this critical radius, radiative depolarization reduces the resonance strength. However, there are a number of other methods for shifting the surface plasmon resonance to the red that do



Optical Properties of Metal Nanoparticles, Fig. 3 (**a**) The strength of the electric field around a 7.5-nm-radius gold particle in vacuum. The relative size of the particle is given compared to the wavelength in the upper panel. At this size, the incident electric field is effectively constant across the particle. (**b**) The extinction, scattering, and absorption efficiencies of the sphere in (**a**). The extinction spectrum is dominated by absorption. The scattering efficiency has been scaled by a factor of 100 to aid visibility; it contributes very little to the total extinction



Optical Properties of Metal Nanoparticles, Fig. 4 (**a**) The electric field around a 120-nm-radius gold particle in vacuum. The particle size is compared to the incident wavelength in the upper panel. Because there is significant variation of the incident field across the surface of the sphere, it is possible to excite a quadrupole mode. (**b**) The extinction, scattering, and absorption efficiencies of the sphere in (**a**). The extinction is dominated by scattering in this size regime

not involve radiative effects. For spheres, the only alternative to increasing the radius is to embed the sphere in a background medium of high refractive index. Alternative, nonspherical particle geometries can be used to red-shift the plasmon wavelength, including nanoshells, nanorods, nanocubes, and nanotriangular prisms. In the following sections, the effects of increasing the refractive index of the medium into which a sphere is embedded are discussed, followed by an examination of quasistatic results for nanoshells and nano-ellipsoids. From there, more complicated structures are introduced; those where analytic solutions are not available. In some cases, the quasistatic limit is not applicable to the particle size regime of interest, and so numerical solutions to Maxwell's equations are investigated.

Firstly, returning to the polarizability of a sphere in the quasistatic limit. There are two equivalent resonance conditions for Eq. 17. The first is the usual way, the frequency is scanned across the spectrum and a resonance occurs when:

$$\varepsilon_1(\omega) = -2\varepsilon_m \tag{23}$$

Alternatively, the background medium index can be varied to coincide with the real part of the dielectric function at a given frequency:

$$\varepsilon_m = -\frac{\varepsilon_1(\omega)}{2} \tag{24}$$

This second resonance condition allows one to determine the general properties of the resonance. By substituting Eq. 22 into Eq. 17, the polarizability reduces to:

$$\alpha = 4\pi a^3 i \frac{-3\varepsilon_1}{2\varepsilon_2} \tag{25}$$

This expression determines the height of the resonance assuming that the medium index can be matched to the permittivity. To explain this further, Fig. 5 shows expression (25) as well as the traditional expression (18) and the full Mie solution for 7.5 nm gold spheres embedded in materials with increasing medium index. The trend shows that for gold, the extinction efficiency is maximized when the medium refractive index is $m = 3$ (materials such as mercury sulfide have a refractive index this high).

Shells in the Quasistatic Limit

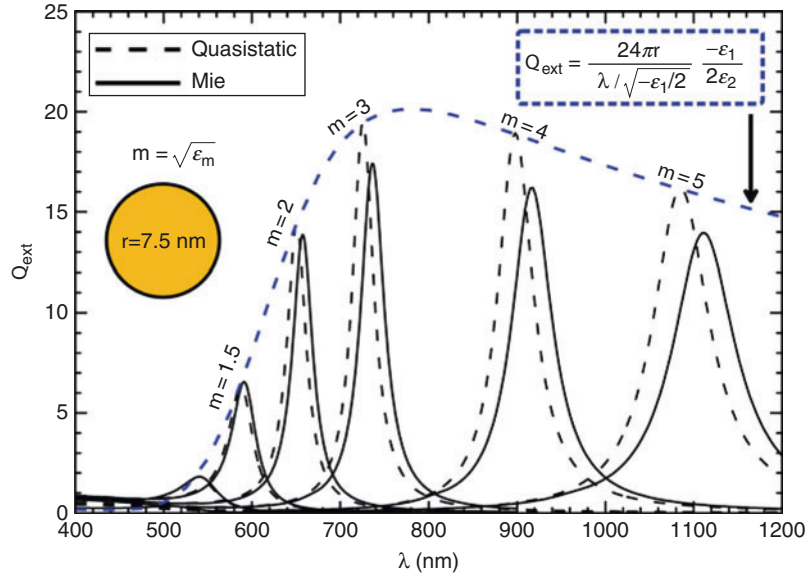
Unlike spheres, nanoshells exhibit two fundamental resonances in analogy to the bonding and antibonding orbitals in molecules [12]. The symmetric dipole like mode shifts from the sphere resonance condition at $\varepsilon_1 = -2\varepsilon_m$ toward the red as the shell thickness decreases. The second mode is an antisymmetric mode that is blue shifted from the sphere resonance. The polarizability of a nanoshell with a core with a dielectric constant ε_c is given by:

$$\alpha = V\frac{(\varepsilon_s - \varepsilon_m)(\varepsilon_c + 2\varepsilon_s) + f(\varepsilon_c - \varepsilon_s)(\varepsilon_m + 2\varepsilon_s)}{(\varepsilon_s + 2\varepsilon_m)(\varepsilon_c + 2\varepsilon_s) + 2f(\varepsilon_s - \varepsilon_m)(\varepsilon_c - 2\varepsilon_s)} \tag{26}$$

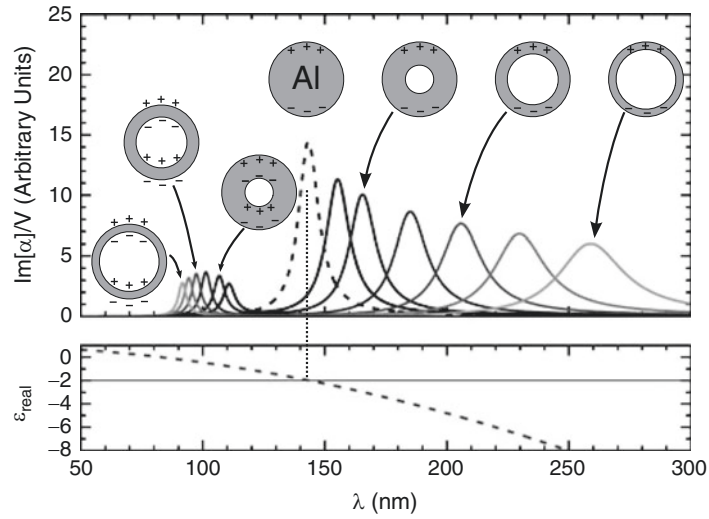
where $f = r_1^3/r_2^3$ is the fraction of the volume occupied by the core (r_1). A schematic of the charge distributions and spectra including both symmetric and antisymmetric resonances for aluminum nanoshells is displayed in Fig. 5. Aluminum was chosen in this case because metals such as gold do not exhibit the blue-shifted resonance due to interband transitions in the region of wavelengths for which $\varepsilon_1 > -2\varepsilon_m$. The red-shifted, symmetric resonance is of much more practical significance, as it allows the plasmon resonance of gold and silver nanoshells to be tuned across the visible spectrum. By changing the shell thickness only marginally, the resonance position can be shifted drastically over a large range of wavelengths. In Fig. 6, the plasmon resonance for gold nanoshells is tuned across the visible by increasing the radius of the core with respect to the shell. Following a similar method that was used for the spheres, an expression can be derived for the general scaling of the plasmon resonance. First, note that a resonance occurs when the denominator of Eq. 24 goes to 0. This can be done by changing the fill factor

$$f = -\frac{(2\varepsilon_m + \varepsilon_1)(\varepsilon_c + 2\varepsilon_1)}{2(\varepsilon_c - \varepsilon_1)(\varepsilon_1 - \varepsilon_m)} \tag{27}$$

O

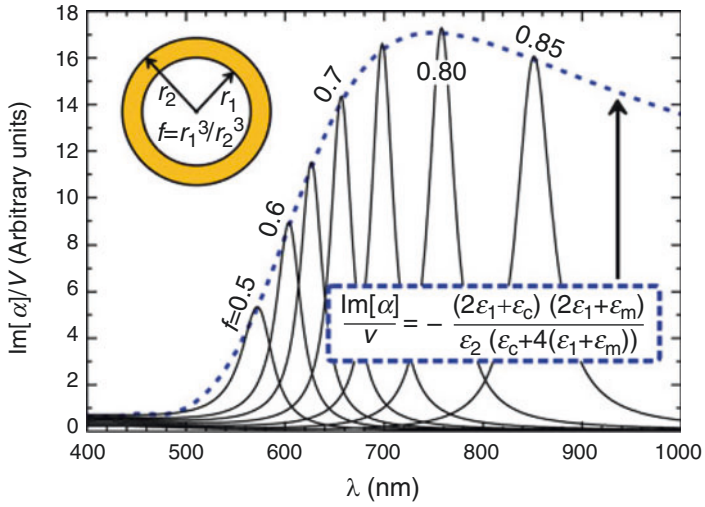


Optical Properties of Metal Nanoparticles, Fig. 5 The extinction efficiency of gold spheres embedded in media with increasing refractive index. The radius of the particle is fixed at 7.5 nm. The quasistatic solutions (*dashed curves*) are given by Eq. 18, and the general trend (*bold dashed curve*) is given by substituting Eq. 25 into Eq. 16a.; in this expression, the wavelength is considered to be in the high refractive index medium, and hence must be divided by Eq. 24. The Mie solutions (*solid curves*) are red-shifted from the quasistatic solution. The shift is exacerbated in high refractive index media because the wavelength in these media is effectively shorter (increasing radiative effects)



Optical Properties of Metal Nanoparticles, Fig. 6 Hybridization of surface plasmon resonances on aluminum nanoshells. *Top panel*: The polarizability normalized to volume for nanoshells with increasing core radius. Data for a nanosphere is included (*dashed curve*). *Bottom panel*: the real part of the dielectric function of aluminum

until a resonance occurs for a given value of the dielectric function of the metal. The fill factor is a real quantity, and so it cannot contribute to the matching of any term in the resonance condition that is imaginary. As such, the imaginary parts of the dielectric function have been removed. Substituting this solution back into Eq. 24, selecting the imaginary part of the resulting



Optical Properties of Metal Nanoparticles, Fig. 7 The normalized polarizability of gold nanoshells with vacuum cores embedded in vacuum (calculated using Eq. 26). The curves are labeled with their respective fill factors. The dashed curve describes the general trend of the resonance on the nanoshell; however, it is only valid for small values of the core and medium permittivity

expression, and assuming that the background medium and the core have small dielectric constants compared to the real part of the dielectric function results in:

$$\frac{\mathrm{Im}[\alpha]}{V} = \frac{(2\varepsilon_1 + \varepsilon_c)(2\varepsilon_1 + \varepsilon_\mathrm{m})}{\varepsilon_2(\varepsilon_\mathrm{c} + 4[\varepsilon_1 + \varepsilon_\mathrm{m}])} \tag{28}$$

which provides a general trend of how the polarizability (and hence the extinction efficiency) varies as the resonance position shifts to the red. This is also shown in Fig. 7. It is interesting to point out that this expression scales with wavelength in the same way as the expression for a sphere (i.e., compare Figs. 5 and 7).

Ellipsoids in the Quasistatic Limit

When the geometry of a nanoparticle is distorted from that of a sphere, additional resonances appear for each axis of the particle which is asymmetric. To simplify the discussion, only prolate, rod-like, spheroids are discussed. A spheroid is an ellipsoid having two equal axes, and in the prolate case, the long axis is the unique axis. For spheroids in the quasistatic limit, two of the three possible plasmon modes are degenerate because the field is assumed constant over the particle. As such, there are two nondegenerate resonances: a red-shifted "longitudinal" mode that occurs along the long axis of the ellipsoid and a blue-shifted "transverse" mode that occurs along the short axis, as first described by Gans [13] in 1912. The polarizability for both axes of the ellipsoid is given by:

$$\alpha_\mathrm{ellipsoid} = V\frac{\varepsilon - \varepsilon_\mathrm{m}}{\varepsilon_\mathrm{m} + L(\varepsilon - \varepsilon_\mathrm{m})} \tag{29}$$

where V is the ellipsoid volume and L is a depolarization factor that depends on the aspect ratio of the ellipsoid (length/width) and the dimension of interest (i.e., $L = L_\mathrm{long}$ or $L = L_\mathrm{trans}$). The depolarization factor for the longitudinal mode is described by: [1]

$$L = \frac{1-e^2}{e^2}\left(-1 + \frac{1}{2e}\ln\frac{1+e}{1-e}\right) \tag{30}$$

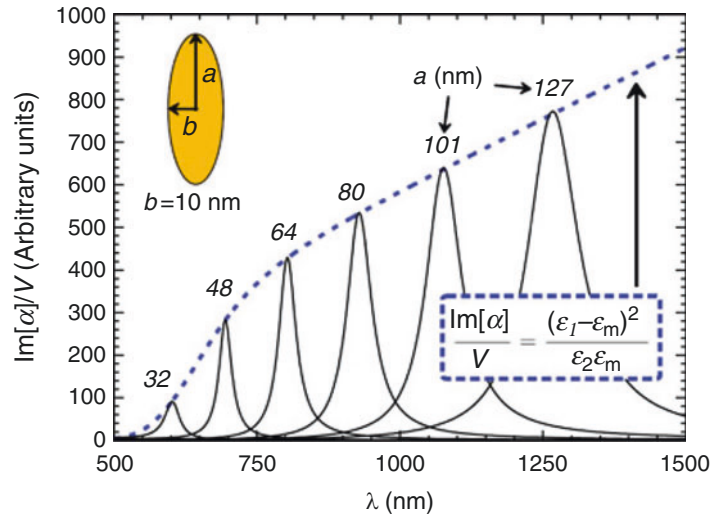
where e is the eccentricity of the particle (a is the radius of the long axis, b is the radius of the short axis:

$$e = \sqrt{1 - \left(b^2/a^2\right)} \tag{31}$$

By noting that the sum of the depolarizations on each of the axes must equal 1, and recalling that two of the modes are degenerate, the transverse depolarization can be expressed as:

$$L_\mathrm{trans} = \frac{1 - L_\mathrm{long}}{2} \tag{32}$$

O



Optical Properties of Metal Nanoparticles, Fig. 8 The longitudinal resonance on a gold ellipsoid as a function of length. The short radius is fixed at 10 nm, and the background dielectric constant is fixed at 1. The strength of the resonance grows continuously as it is shifted into the infrared

For a purely ellipsoidal structure, the transverse resonance blue shifts away from the sphere resonance as the length of the rod is increased, or as the width is decreased. So it would seem that for metals such as gold, the transverse resonance would be heavily damped by interband transitions. In reality, the most common experimental structures are actually capped or flat-ended cylinders, and the transverse resonance appears at approximately the same wavelength as the resonance for a sphere. The longitudinal resonance wavelength depends not specifically on the length of the ellipsoid, but instead on the ratio of the length to the width (aspect ratio). This is evident from the relationship in Eq. 29. Figure 8 shows the effect of increasing the aspect ratio of a gold ellipsoid in vacuum, that is, by fixing the width of the ellipsoid and increasing the length. Just as in previous cases, the imaginary part of Eq. 28 has been selected:

$$\frac{\alpha_{\text{ellipsoid}}}{V} = i\frac{\varepsilon_2\varepsilon_{\text{m}}}{\varepsilon_2^2 L_{\text{long}}^2 + \left(\varepsilon_{\text{m}} + L_{\text{long}}(\varepsilon - \varepsilon_{\text{m}})\right)^2}, \tag{33}$$

which is proportional to the extinction efficiency. The longitudinal resonance on the ellipsoid scales in a fundamentally different way to that of a sphere or a nanoshell. The resonance strength does not peak around 750 nm, but instead continues to climb as the resonance is shifted into the infrared. To determine the general properties of the ellipsoid resonance, the denominator of Eq. 29 is set to zero, and solved for L_{long}:

$$L_{\text{long}} = -\frac{\varepsilon_{\text{m}}}{\varepsilon_1 + \varepsilon_{\text{m}}} \tag{34}$$

Substituting this back into Eq. 33 gives the general scaling of the resonance:

$$\frac{\alpha_{\text{ellipsoid}}}{V} = i\frac{(\varepsilon_1 - \varepsilon_m)^2}{\varepsilon_2\varepsilon_{\text{m}}} \tag{35}$$

In this case, the polarizability is proportional to the real part of the dielectric function *squared*. Because $|\varepsilon_1| > \varepsilon_2$ for wavelengths up to about 10 μm, the resonance strength will continue to grow rapidly.

Numerical Methods to Solve Maxwell's Equations

The first half of this entry described a variety of analytical expressions which can be used to solve Maxwell's equations and calculate the extinction efficiency and electric fields in the vicinity of spherically or axially symmetric particles. Here the discrete dipole approximation (DDA) is discussed. The DDA is one of the numerical techniques that can be used to model the extinction spectra and electric fields of arbitrarily shaped

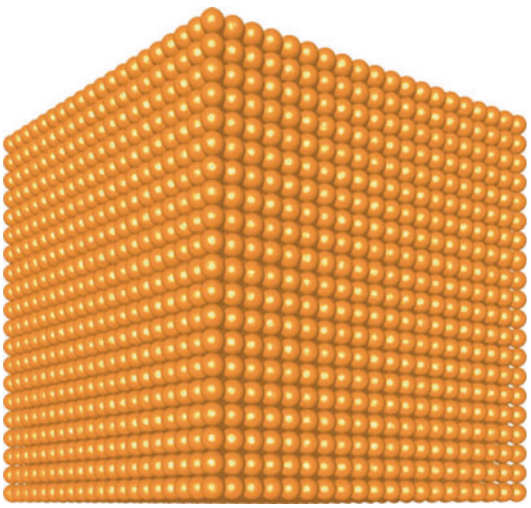
particles which are difficult or impossible to determine analytically.

DDA: The Discrete Dipole Approximation Method

The discrete dipole approximation was first developed in 1973 by Purcell and Pennypacker [14] as a numerical technique to calculate the absorption and scattering of electromagnetic radiation by arbitrary-shaped targets. This technique is very popular in the electrodynamics community due to its implementation by Draine et al. into the freely available software package DDSCAT [5].

A benefit of the DDA is that it can be used to model targets that are not spherically symmetric. It does this by representing an arbitrary-shaped target by an array of N dipoles within a cubic grid, an example of which is shown in Fig. 9. Here the target is a 20-nm-edge-length cube with a dipole size of 1 nm and the edges of the cube lie along the x, y, z Cartesian axes.

The polarizability, α_i, of each dipole is then calculated based upon the dielectric function and position within the target r_i. After the dipole polarizabilities have been calculated the target absorption C_{abs} and scattering, C_{sca} cross sections can be calculated as described below [15].



Optical Properties of Metal Nanoparticles, Fig. 9 A 20-nm-edge-length cube target composed of individual 1-nm dipoles, generated by the DDSCAT target generation function

If the index $i = 1, 2, \ldots N$ represents each dipole within a cubic lattice, then upon application of a plane wave field, the induced dipole moment, P_i is given by [16]:

$$P_i = \alpha_i . E_{loc,i} \tag{36}$$

where α_i is the dipole polarizability tensor and the electric field $E_{loc,i}$ is the sum of the incident electric field $E_{inc(r_i,t)}$ and the radiated electric field $E_{dip,i}$ from the surrounding $N-1$ dipoles [16]

$$E_{loc,i} = E_{inc}(r_i, t) + E_{dip,i} \tag{37}$$

with

$$E_{inc}(r_i, t) = E_0 e^{i(kr_i - \omega t)} \tag{38}$$

where E_0 is the amplitude of the electric field, k is the wave vector, r_i is the position vector of the dipole, ω is the angular frequency, and t is time.

$E_{dip,i}$ is the electric field radiated from all dipoles which are not located at index r_i, as given by [16]

$$E_{dip,i} = -\sum_{j \neq i} A_{ij} . P_j \tag{39}$$

This implies

$$E_{loc,i} = E_0 e^{i(kr_i - \omega t)} - \sum_{j \neq i} A_{ij} . Pj \tag{40}$$

Each element A_{ij} is defined by the following equation for an electric field from a radiating or oscillating dipole [16]

$$A_{ij} . P_j = \frac{e^{ikr_{ij}}}{r_{ij}^3} \left\{ k^2 r_{ij} \times (r_{ij} \times P_j) + \frac{1 - ikr_{ij}}{r_{ij}^2} \left[r_{ij}^2 P_j - 3 r_{ij} (r_{ij} . P_j) \right] \right\} \tag{41}$$

where $i \neq j$, and r_{ij} is a vector that points from dipole j to dipole i (and the scalar magnitude of r_{ij} is used where appropriate in Eq. 41).

Substituting Eq. 41 into Eq. 38 and simplifying to find $E_{inc}(r_i, t)$ gives a set of $3N$ complex linear equations [16]

O

$$E_{inc}(r_i,t) = \frac{P_i}{\alpha_i} + \sum_{j\neq i} A_{ij}.P_j \qquad (42)$$

Iterative methods are used to solve this set of linear equations, and the DDSCAT authors provide a choice of three algorithms. For all of the calculations presented here the Biconjugate Gradient with stabilization algorithm has been implemented to solve Eq. 42 for P_i because it is one of the faster and better-behaved algorithms. Once P_i is known, the cross sections C_{abs} and C_{sca} can be calculated from Eqs. 43 and 44, respectively [16].

$$C_{abs} = \frac{4\pi k}{|E_{inc}|^2}.\sum_{i=1}^{N}\left\{\mathrm{Im}\left[P_i.\left(\alpha_i^{-1}\right)^* P_i^*\right] - \frac{2}{3}k^3|P_i|^2\right\} \qquad (43)$$

$$C_{ext} = \frac{4\pi k}{|E_{inc}|^2}.\sum_{i=1}^{N}\mathrm{Im}\left(E_{inc,i}^*.P_i\right) \qquad (44)$$

In the first implementation of DDA, the Clausius–Mossotti Relation (CMR) [15] was used to calculate the dipole polarizabilities a_i based upon the complex refractive index m (of the material associated with the ith dipole), such that

$$\alpha_{CMR} = \frac{3d^3}{4\pi}\left(\frac{m^2-1}{m^2+2}\right) \qquad (45)$$

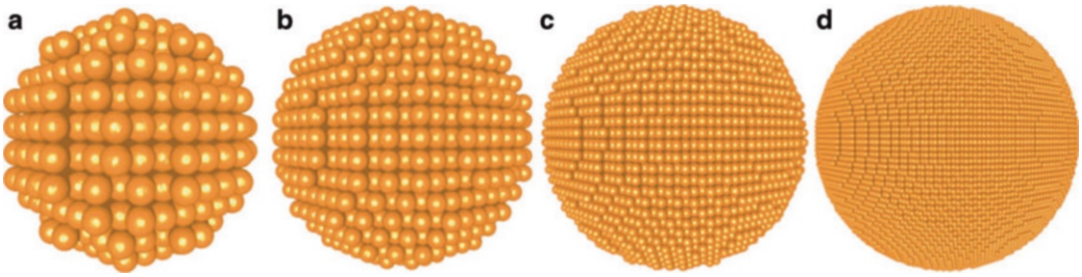
where d is the inter-dipole spacing. This approach is exact in the quasistatic regime; however, it is not exact at finite wavelengths because radiative damping is not taken into consideration in Eq. 45. To overcome this problem, Draine [15] introduced the Lattice-Dispersion-Relation (LDR) which required that the polarizability of dipoles on an infinite lattice give the same dispersion relation as a bulk medium [15]. In the LDR approach, the radiative reaction correction occurs naturally and the polarizability is given by

$$\alpha_{LDR} = \frac{\alpha_{CMR}}{1+\left(\frac{\alpha_{CMR}}{d^3}\right)\left[(b_1+m^2b_2+m^2b_3S)^2(kd)^2 - \left(\frac{2}{3}\right)i(kd)^3\right]} \qquad (46)$$

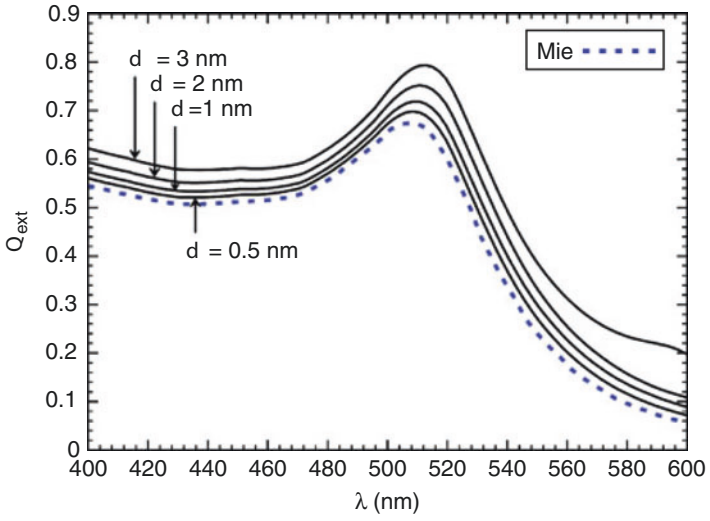
where $b_1 = -1.8915316$, $b_2 = -0.1648469$ and $b_3 = -1.7700004$. S depends on the wave-vector direction and incident polarization and given as $S = \sum_j \left(a_j e_j\right)^2$. a and e are the unit propagation and polarization vectors, respectively, and the directions $j = 1, 2, 3$ correspond to the dipole lattice axes.

The LDR guarantees convergence of the DDA for a homogeneous infinite material; however, its applicability to light interacting with nanoparticles is based on experience and testing of convergence. An example of convergence testing can be seen below, where both Mie theory (exact solution to Maxwell's equations) and DDA (with varying dipole spacings) are used to calculate the extinction spectrum of a 30-nm-diameter gold sphere. Dielectric data from Johnson and Christy [17] for gold in vacuum was used for both sets of calculations. For the DDA method, the 30-nm-diameter sphere was represented with four different dipole spacings, namely, 3 nm (sphere A, 552 dipoles), 2 nm (sphere B, 1,792 dipoles), 1 nm (sphere C, 14,328 dipoles), and 0.5 nm (sphere D, 113,104 dipoles). Schematic figures showing the increased dipole numbers within the sphere, as generated by the DDSCAT target generator, can be seen in Fig. 10.

Comparison of the extinction efficiencies calculated using both Mie theory and DDA are shown in Fig. 11. It can be seen that the spectrum for the sphere with a dipole spacing of 3 nm is not converged with the exact Mie spectrum as only 552 dipoles are used to represent this target structure. As the dipole spacing is reduced from 3 to 1 nm, providing a concomitant increase in the number of dipoles within the sphere target, the spectra start to show convergence with that calculated using Mie theory. Decreasing the dipole spacing to 0.5 nm increases the number of dipoles within the target sphere over 200 times (113, 104); however, convergence of the spectrum to the same peak position and magnitude to the exact Mie spectrum is never quite achieved. Indeed, DDA is not guaranteed to be exact as noted above, but, in practice, the quality of comparisons between



Optical Properties of Metal Nanoparticles, Fig. 10 DDSCAT target generated representation of a 30-nm-diameter gold sphere with decreasing dipole sizes of 3 nm (sphere A), 2 nm (sphere B), 1 nm (sphere C), and 0.5 nm (sphere D)

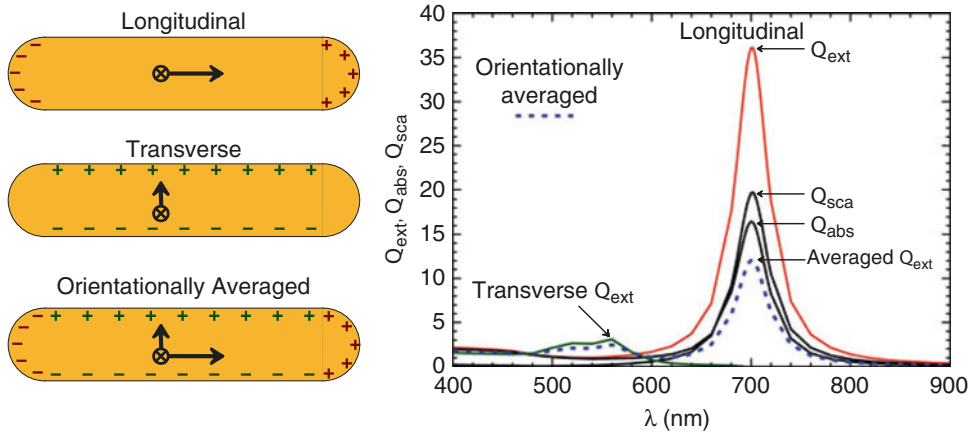


Optical Properties of Metal Nanoparticles, Fig. 11 Calculated extinction spectra of a 30-nm-diameter gold particle in vacuum. Results using both Mie theory and the DDA method with decreasing dipole spacings are shown here with the aim of showing convergence of the spectra with Mie theory as the dipole numbers are increased

DDA and exact results is comparable to that seen in Fig. 11.

Moreover, although not shown here, achieving convergence of near-field properties such as electric fields is more difficult when compared with far-field properties such as extinction efficiencies due to the rapidly changing nature of the contribution of nearby dipoles to the electric field at the particle surface. Decreasing the dipole spacing does help convergence; however, there is a limit to how small the dipoles can be made because as the number of dipoles within the target increases, so does the computational time (approximately as NlnN), and so a compromise must be met between calculation accuracy and computational expense.

One of the benefits of the DDA method is that the dipoles need only be used to represent the target structure and not the surrounding dielectric medium (no grid points are required to describe the empty space surrounding the target) as this is taken into account by dividing the target dielectric data and the incident wavelength by the medium dielectric data to produce a *relative dielectric function*. This reduces the overall number of dipoles required to describe the target/medium system and helps to reduce computational time. A further benefit is that the DDA method is capable of modeling particles with dimensions where the quasistatic regime does not apply, allowing the calculation of higher-order modes. Specific



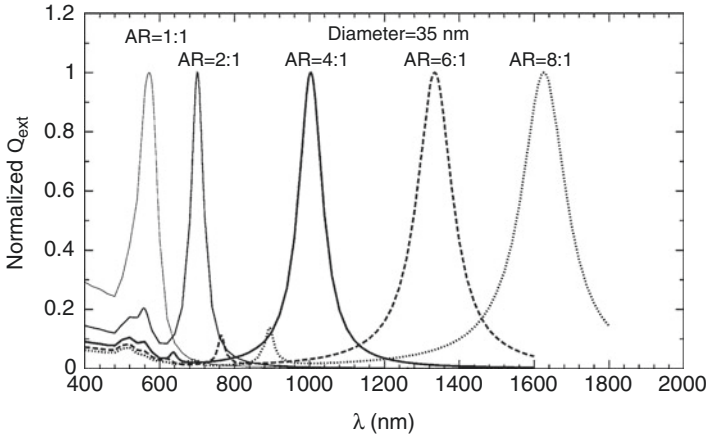
Optical Properties of Metal Nanoparticles, Fig. 12 (*Left*) Schematic depiction of the charge distributions produced when the incident wave vector is perpendicular to the long rod axis with longitudinal (*top rod*), transverse (*center rod*), and orientationally averaged (*bottom rod*) polarizations. (*Right*) The calculated extinction spectra for a 35-nm-diameter, 2:1 aspect ratio rod with different polarizations with the longitudinal mode showing the absorption and scattering contributions

examples will be discussed in the following sections where rods and prisms are modeled.

Implementation of DDA to Calculate the Optical Response of Nanorods

To illustrate the capabilities of the DDA method, here it is used to calculate the optical response of cylindrical gold rods with a constant diameter of 35 nm and increasing aspect ratio. The aspect ratio is defined as the ratio of the rod length to the rod diameter. A dipole spacing of 1 nm was used for all calculations shown here. These rods were modeled to be immersed in water by dividing the incident wavelength and the refractive index of gold, from Frederikse and Weaver [18], by the refractive index of water (1.331) in the DDA calculation. Orientationally averaged calculations were performed to enable all of the resonances (transverse, dipole, and higher order) to be shown. To perform these calculations, the cylindrical rod target was initially assumed to have its long axis along the *y*-axis and its short axis along the *x*-axis of a space-fixed coordinate system. The incident wave vector **k** propagates along the *x*-axis, and polarization was considered along both the *y*- and *z*-axes [19]. The rod was then rotated through 90° in several steps with respect to **k**, and the extinction spectra calculated at each step were combined to generate an orientationally averaged result. Figure 12 (left) shows a schematic depiction of the charge distributions produced within the rod for different polarizations, namely longitudinal; polarized along the long axis (top), transverse; polarized along the short axis (middle) and orientationally averaged; where a combination of polarizations are considered (bottom). The calculated extinction spectra of a 35-nm-diameter, 70-nm-length rod are shown in Fig. 12 (right) with the aim of depicting the variation in extinction spectra produced based on the polarization incident upon the target. The transverse mode is evident between the wavelengths of 500 and 580 nm, while the longitudinal mode is evident at 700 nm. Only the orientationally averaged extinction spectrum shows a combination of both the transverse and longitudinal modes. The scattering efficiency of the longitudinal mode is the dominant component of the extinction efficiency. As the length and diameter of the rod are increased, radiative effects will increase the amount of scattering, and the absorption efficiency will reduce.

Figure 13 shows the calculated orientationally averaged extinction spectra for 35-nm-diameter rods with increasing aspect ratio. The maximum dipole ($n = 1$) extinction value for each rod



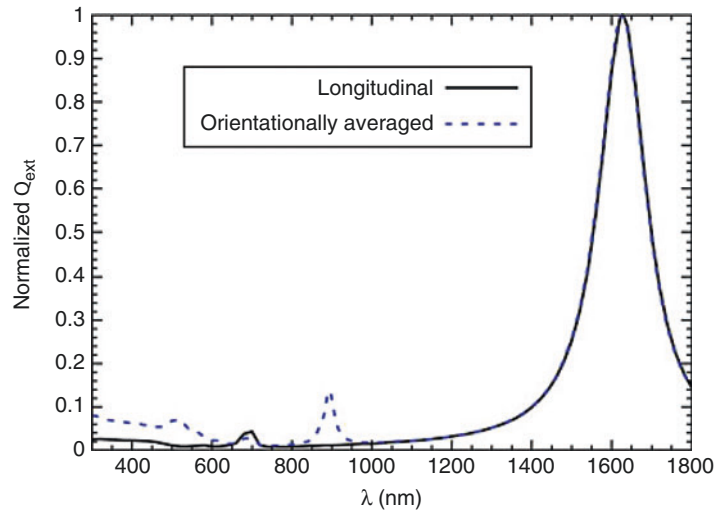
Optical Properties of Metal Nanoparticles, Fig. 13 Orientationally averaged extinction spectra of 35-nm-diameter rods with increasing aspect ratios ranging from 1:1 through to 8:1

spectrum has been normalized to 1 purely for clarity purposes. This is necessary as the charge separation along the rod increases as the length of the rod increases allowing the magnitude of the dipole extinction to increase and dwarfing the quadrupole and higher-order resonances of the smaller aspect ratio rods [19]. For example, the 1:1 aspect ratio rod produces a maximum dipole extinction of ~14.5, while the 8:1 aspect ratio rod produces a maximum dipole extinction value of ~51.5.

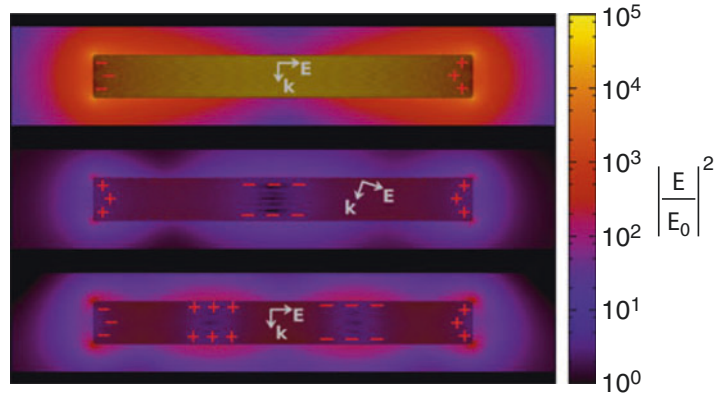
Transverse resonances are evident for the rods with aspect ratios greater than 1:1. These resonances lie in the region between the wavelengths of 500 and 600 nm and are not greatly affected as the aspect ratio of the rod increases as will be described in more detail below. Conversely, it can be seen that the longitudinal dipole resonance red-shifts as the aspect ratio of the rod increases, similar to that described by Gans theory when applied to ellipsoids (see previous section). For example, the rod with an aspect ratio of 2:1 produces a dipole resonance at 581 nm, and increasing the length of the rod to an aspect ratio of 4:1 shifts the dipole resonance to 1,034 nm. A further increase in aspect ratio to 6:1 shifts the resonance to 1,401 nm. These results demonstrate the attractive nature of rods in potential applications that require precise control over the position of the dipole resonance wavelength.

Also evident in Fig. 13 is the presence of higher-order resonances for rods with aspect ratios of 4:1, 6:1 and 8:1. A quadrupole ($n = 2$) resonance is evident at 636 nm for the 4:1 rod and 764 nm for the 6:1 rod. In the case of the 8:1 aspect ratio rod, the quadrupole resonance is evident at 894 nm as well as an octopole ($n = 3$) resonance at 692 nm (more evident in Fig. 14). One of the advantages of using orientationally averaged calculated spectra is that all of the even ($n = 2, 4, \ldots$) and odd ($n = 1, 3, \ldots$) higher-order modes are present because there are no selection rules governing plasmon excitation in this case [19]. In general, the longitudinal plasmons have induced polarizations that are either symmetric or antisymmetric with respect to a 180° rotation of the rod, and only the antisymmetrical resonances can couple if the electric vector of the light (i.e., the polarization vector) is along the rod axis. This leads to the even modes being forbidden for longitudinal polarization, as can be seen more clearly in Fig. 14, where both fixed (longitudinal) and orientationally averaged polarizations were used to model the extinction spectra for the 35-nm-diameter and 280-nm-length (8:1) rod. The quadrupole mode is not evident in the spectrum for fixed longitudinal polarization due to the symmetry of the rod, but the octopole is. Conversely, the orientationally averaged polarization spectrum shows both the quadrupole and octopole higher-order modes.

A more detailed interpretation of the induced charge distributions within the rod for the dipole, quadrupole, and octopole modes can be seen in Fig. 15. Here the $|E|^2$ fields have been calculated for the same 35-nm-diameter, 8:1 aspect ratio rod



Optical Properties of Metal Nanoparticles, Fig. 14 Calculated extinction spectra for a gold 35-nm-diameter, 280-nm-length rod (8:1) with both fixed and orientationally averaged polarizations



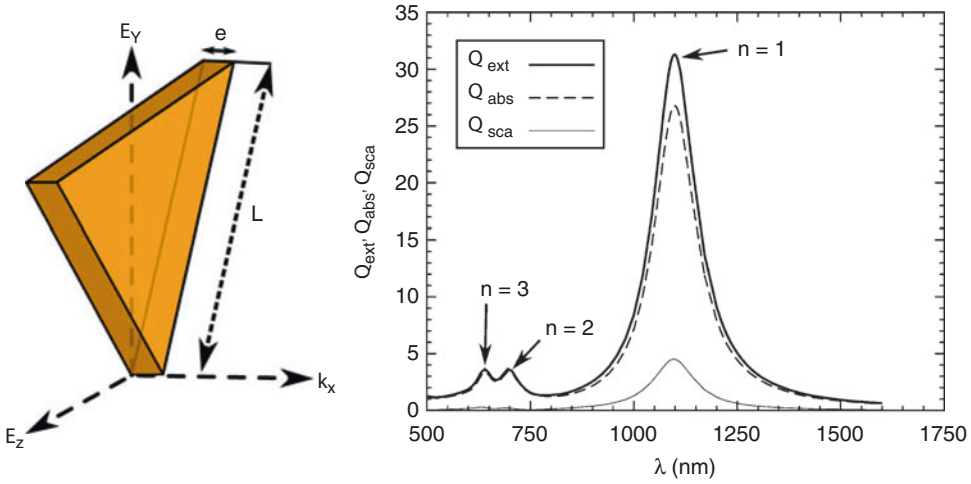
Optical Properties of Metal Nanoparticles, Fig. 15 Calculated $|E|^2$ fields for a 35-nm-diameter, 8:1 aspect ratio rod showing the dipole at 1,628 nm (*top*), quadrupole at 894 nm (*middle*), and octopole at 692 nm (*bottom*)

as described in Fig. 14 at the wavelength corresponding to each of the modes with the dipole at 1,628 nm (top), quadrupole at 894 nm (middle), and the octopole at 692 nm (bottom). Although $|E|^2$ does not show the direction of the induced polarization, the figure shows the expected signatures of odd and even symmetry modes, with a nodal surface in the middle of the rod for the 894 nm (even symmetry) mode.

From this it is clear that variation of the incident electric field across the length *and* the width of the particle is necessary to break the selection rules and excite the 894-nm mode. This also follows a similar vein to the production of the quadrupole resonance in a sphere in Fig. 3b.

Implementation of DDA to Calculate the Optical Response of Triangular Prisms

Triangular prisms have been modeled at great length by the electrodynamics research community because, like rods, they allow for the tunability of plasmon resonances as a function of their aspect ratio. Here gold triangular prisms with a thickness of 7 nm and edge lengths of 60, 90, 110, 120, 130, 150, 170, 190, and 210 nm were modeled by DDA (dipole spacing of 1 nm) with the aim of demonstrating how the resonance modes of these particles can be tuned from the visible to the IR by varying the aspect ratio as well as showing the criteria required to produce higher-order resonances. In this case, the aspect ratio of a



Optical Properties of Metal Nanoparticles, Fig. 16 Orientation of the prism with respect to the incident wave vector and polarization (*left*). Calculated spectra of a 120-nm-edge-length, 7-nm-thick (aspect ratio ~17) prism, showing dipole ($n = 1$) and higher-order modes ($n = 2, 3$)

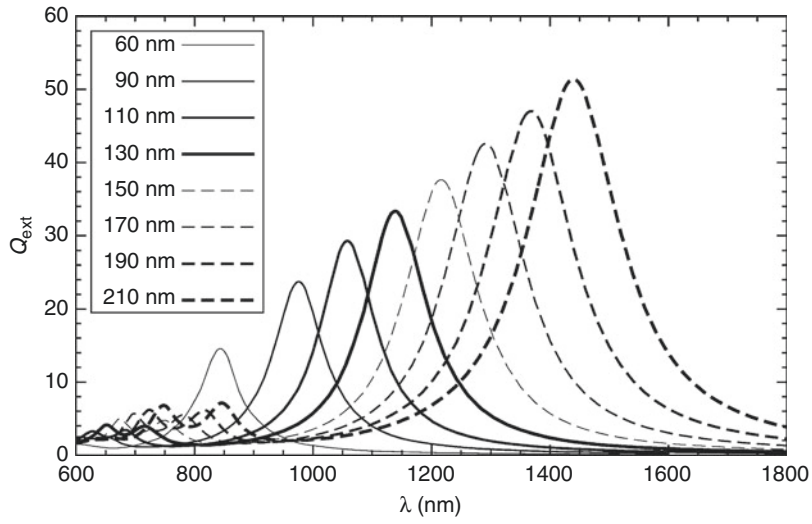
prism is the ratio of the prism length to prism thickness.

All of the prisms were modeled as gold in water using the *relative* dielectric data of Johnson and Christy [17]. The face of the prism was fixed along the *yz*-plane, with the wave vector incident along the *x*-axis and polarization along the *y*-axis, as can be seen in Fig. 16 (left).

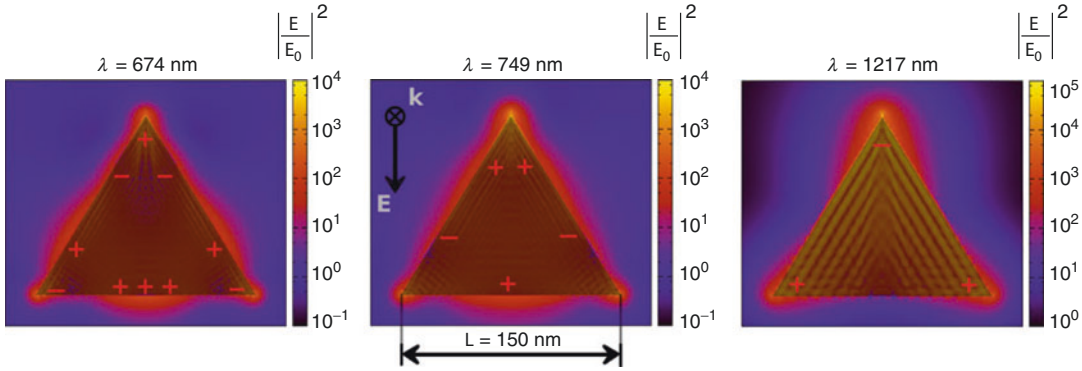
The calculated spectrum for a prism of 7-nm thickness and 120-nm edge length (aspect ratio of ~17) is shown in Fig. 15 (right). Absorption is the dominant contribution to the extinction efficiency due to the size of this particle. If the size of the prism were allowed to increase, scattering would become the dominant contribution to the extinction efficiency, as was also the case for spheres and rods. The dipole mode is excited at 1,099 nm, and higher-order modes are also excited at wavelengths of 696 and 640 nm. These higher-order resonances will be described in more detail below. The focus henceforth will center on the dipole resonance.

Figure 17 shows extinction spectra calculated with polarization along the y-axis for 7-nm-thick prisms with edge lengths of 60 nm (aspect ratio ~8.5) through to 210 nm (aspect ratio = 30). Increasing the aspect ratio of the prism causes the dipole plasmon resonance to red-shift due to shape effects. For example, the 60-nm-edge-length prism excites a dipole resonance at 843 nm. Increasing the edge length of this prism to 210 nm allows the production of a dipole resonance in the near-IR part of the spectrum at 1,082 nm. It is not necessary to normalize the maximum extinction values for each spectrum to 1, as was the case for the rod spectra, because the higher-order modes are not dwarfed by the magnitude of the dipole resonances of the larger aspect ratio prisms. This makes it easy to demonstrate nicely how the extinction cross section at the dipole plasmon resonance maximum exceeds the geometrical extinction cross section by approximately 50 times in the case of the 210-nm-edge-length prism. The magnitude of the dipole resonance wavelength increases because the charge separation along the face of the prism also increases as a function of prism edge length (as was the case for the rods).

It can also be seen from Fig. 17 that structures with edge lengths of 90 nm and larger produce higher-order resonances to the blue of the dipole resonance. Higher-order modes are generally produced because of the variation of the incident field within the particle geometry, as described above for the higher-order resonances in rods. Here, in particular, the higher-order modes are excited



Optical Properties of Metal Nanoparticles, Fig. 17 Extinction spectra for prisms oriented as per Fig. 14 (*left*) with a thickness of 7 nm and edge lengths ranging from 60 to 210 nm



Optical Properties of Metal Nanoparticles, Fig. 18 Calculated $|E|^2$ fields for the 150-nm-edge-length prism with a thickness of 7 nm, showing the dipole ($n = 1$) at 1,217 nm (*right*), quadrupole ($n = 2$) at 749 nm (*middle*), and octopole ($n = 3$) at 674 nm (*left*)

with relative ease. However, this is because the prisms have been modeled with sharp edges, as opposed to the slightly rounded edges of synthesized prisms. Sharp prism edges are more conducive to the excitation of higher-order resonances, and it has been shown that rounding of the prism tips lends itself to a reduction of the excitation of higher-order modes as well as a blue-shifting of the dipole resonance.

From the spectra in Fig. 17, the 150-nm-edge-length prism was selected to provide more detailed information on the higher-order modes. The dipole resonance is excited at 1,217 nm, the quadrupole at 749 nm, and the octopole at 674 nm. The calculated $|E|^2$ electric fields were calculated at the wavelength of each of these modes and can be seen in Fig. 18 for polarization along the *y*-axis.

The presence of the + and − symbols on the figures helps indicate the charge distribution excited for each resonance mode. The production of these resonance modes follows a similar vein to those produced in spheres and rods, as described previously. Essentially, the incident electric field could be considered constant across the thickness of the particle due to the prism dimensions, which

allows for excitation of the dipolar mode at 1,217 nm. At shorter wavelengths, the contribution of the electric field of the incident light varies along the edges of the prism. This allows for quadrupole and octopole modes at 749 and 674 nm, even though the incident field does not change sign across the depth of the particle (as is the case for a sphere). Interestingly, the tips of the prisms are associated with the greatest plasmonic electric field enhancement greatest plasmonic electric field enhancement. This has led to these regions on the prisms to be known as "hot-spots." Near these locations, the sensitivity to the dielectric environment, for example, the surrounding solvent, is greater than the "colder-spots" on the prism. These interactions are short-ranged, which means that only molecules within <10 nm of the hot-spot provide high index of refraction sensitivity. On the other hand, the colder regions of the prism, such as the edges and the faces, are less sensitive to the dielectric environment but are longer ranged (up to 30 nm) [20].

Conclusion

This entry focused on understanding plasmon resonance excitation for noble metal nanoparticles, and using analytical and numerical computational techniques to solve Maxwell's equations for the interaction of light with these nanoparticles so as to predict or interpret the properties of the plasmon resonances. The plasmon resonance properties of the nanoparticles were found to be very sensitive to the size, shape, and dielectric environment of the nanoparticles, but electrodynamic modeling is capable of describing these properties accurately. The choice of electromagnetic modeling technique depends strongly on particle geometry and particle size. Quasistatic analytical solutions were presented for particles that were considered small compared to the wavelength of the incident light, in particular spheres, shells and ellipsoids. Quasistatic theory provides a useful approach for relating plasmon resonance wavelength to nanoparticle structure; however, it is only applicable to small particles. For spheres, it is possible to obtain a full electromagnetic solution as described by Mie theory, and this shows that the LSPR wavelength is strongly dependent upon the diameter of the spherical particle, with a broadening of the plasmon resonance width and the production of higher-order modes evident as the particle diameter is increased. The position of the dipole plasmon resonance for both shells and ellipsoids in the quasistatic regime was dependent upon their aspect ratio, with larger aspect ratios pushing the LSPR further to the red.

Numerical solutions using the DDA were introduced to model the interaction of light with nanostructures which could not be solved by analytical means. At specific applications, the DDA was used to calculate the optical spectra and electric fields for gold rods and triangular prisms. It was shown that the dipolar resonance red-shifts as the aspect ratio of both of these structures is increased, much as is predicted by quasistatic theory for spheroidal particles. However, since the DDA method provides a full electromagnetic solution, the higher-order modes within these structures are also correctly calculated.

Acknowledgment NH and GCS acknowledge the DOE NERC EFRC (DE-SC0000989) for generous support. MGB was supported by AFOSR/DARPA Project BAA07-61 (Grant No. FA9550-08-1-0221) and NSF MRSEC (Grant No. DMR-0520513) at the Materials Research Center of Northwestern University.

Cross-References

- ▶ Active Plasmonic Devices
- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Finite-Difference Frequency-Domain Technique
- ▶ Finite-Difference Time-Domain Technique
- ▶ Gold Nanorods
- ▶ Hollow Gold Nanospheres
- ▶ Light Localization for Nano-optical Devices
- ▶ Nanoparticles
- ▶ Nanostructures for Photonics
- ▶ Plasmon Resonance Energy Transfer Nanospectroscopy
- ▶ Plasmonic Structures for Solar Energy Harvesting

▶ Surface Plasmon Enhanced Optical Bistability and Optical Switching
▶ Synthesis of Gold Nanoparticles

References

1. Bohren, C.F., Huffman, D.R.: Absorption and Scattering of Light by Small Particles. Wiley, Weinheim (2004)
2. Kreibig, U., Vollmer, M.: Optical Properties of Metal Clusters. Springer, Berlin/Heidelberg (1995)
3. Maxwell, J.C.: A dynamical theory of the electromagnetic field. Philos. Trans. R. Soc. Lond. **155**, 459–512 (1865)
4. Mie, G.: Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen. Ann. Phys. **25**, 377–445 (1908)
5. Draine, B.T., Flatau, P.J.: Discrete-dipole approximation for scattering calculations. J. Opt. Soc. Am. A **11**(4), 1491–1499 (1994)
6. Taflove, A., Hagness, S.C.: Computational Electrodynamics: The Finite-Difference Time-Domain Method. Artech House, Boston (2000)
7. Masiello, D., Schatz, G.C.: On the linear response and scattering of an interacting molecule-metal system. J. Chem. Phys. **132**, 064102–064018 (2010)
8. McMahon, J.M., Gray, S.K., Schatz, G.C.: Nonlocal optical response of metal nanostructures with arbitrary shape. Phys. Rev. Lett. **103**, 097403–097404 (2009)
9. Loo, C., Lin, A., Hirsch, L., Lee, M., Barton, J., Halas, N.J., West, J.L., Drezek, R.: Nanoshell-enabled photonics-based imaging and therapy of cancer. Technol. Cancer Res. Treat. **3**(1), 33–40 (2004)
10. Ohanian, H.C.: Classical Electrodynamics. Allyn and Bacon, Boston (1988)
11. Kelly, K.L., Coronado, E., Zhao, L.L., Schatz, G.C.: The optical properties of metal nanoparticles: the influence of size, shape, and dielectric environment. J. Phys. Chem. B **107**(3), 668–677 (2003)
12. Prodan, E., Radloff, C., Halas, N.J., Nordlander, P.: A hybridization model for the plasmon response of complex nanostructures. Science **302**(5644), 419–422 (2003)
13. Gans, G.: Über die Form ultramikroskopischer Goldteilchen. Ann. Phys. **342**(5), 881–900 (1912)
14. Purcell, E.M., Pennypacker, C.R.: Scattering and absorption of light by nonspherical dielectric grains. Astrophys. J. **186**, 705–714 (1973)
15. Collinge, M.J., Draine, B.T.: Discrete-dipole approximation with polarizabilities that account for both finite wavelength and target geometry. J. Opt. Soc. Am. A **21**(10), 2023–2028 (2004)
16. Brioude, A., Jiang, X.C., Pileni, M.P.: Optical properties of gold nanorods: DDA simulations supported by experiments. J. Phys. Chem. B **109**(27), 13138–13142 (2005)
17. Johnson, P.B., Christy, R.W.: Optical constants of the noble metals. Phys. Rev. B **6**(12), 4370–4379 (1972)
18. Weaver, J.H., Frederikse, H.P.R.: Optical Properties of Selected Elements, 82nd edn, p. 12. CRC Press, Boca Raton (2001)
19. Schmucker, A.L., Harris, N., Banholzer, M.J., Blaber, M.G., Osberg, K.D., Schatz, G.C., Mirkin, C.A.: Correlating nanorod structure with experimentally measured and theoretically predicted surface plasmon resonance. ACS Nano **4**, 5453–5463 (2010)
20. Haes, A.J., Zou, S., Schatz, G.C., Duyne, R.P.V.: Nanoscale optical biosensor: short range distance dependence of the localized surface plasmon resonance of noble metal nanoparticles. J. Phys. Chem. B **108**, 6961–6968 (2004)

Optical Reflectance Spectroscopy of Ancient Paper

▶ Spectroscopy of Ancient Documents

Optical Resonance Biosensor

▶ Whispering Gallery Mode Resonator Biosensors

Optical Response of Metal Particles

▶ Optical Properties of Metal Nanoparticles

Optical Ring Resonator Biosensor

▶ Whispering Gallery Mode Resonator Biosensors

Optical Spectroscopy for Ancient Paper Diagnostic

▶ Spectroscopy of Ancient Documents

Optical Techniques for Nanostructure Characterization

Andrés Cantarero
Materials Science Institute, University of Valencia, Valencia, Spain

Synonyms

Spectroscopic techniques

Definition

Optical techniques are that using light to inspect or characterize the materials. Nanostructures are a material or a combination of materials limited in size at least in one direction of the space. It is generally accepted that the size of a single nanostructure must be below 100 nm.

Overview

This entry is devoted to the description of the main optical techniques used in the characterization of nanostructures in general, and, in particular, of semiconductor nanostructures. Most of these techniques have been used in the characterization of bulk materials over the years and the basic principles are widely described in many textbooks. The main variation when they are applied to the characterization of nanostructures is the insertion of a confocal microscope or the combination of optical microscopy with atomic force microscopy (AFM). The wavelength of the light in the so-called optical region ranges from the near ultraviolet (UV), 180–400 nm (3.1–6.9 eV), to the near infrared (IR), 700–5 μm (0.25–1.8 eV). It includes, obviously, the visible region, from 400 to 700 nm (1.8–3.1 eV). The most important techniques in the optical region are optical imaging, absorption and emission spectroscopies, and Raman and infrared spectroscopies. The basic principle of the different techniques, the experimental setups, and the particularities when they are applied to the characterization of nanostructures will be described.

The prefix nano can be applied when one of the dimensions of the material is smaller than 100 nm. This includes two-dimensional structures like quantum wells or superlattices, or more recently graphene or a few layers of transition metal dichalcogenides (WS_2, $MoSe_2$, for instance), quasi-one-dimensional nanostructures called nanowires, and quasi-zero-dimensional nanostructures like self-assembled quantum dots, nanowire heterostructures, or metallic and semiconducting nanoparticles.

Microscopy

Optical microscopy is one of the first steps in the characterization of materials. It allows the visualization and selection of samples prior to manipulating or processing in order to perform more complicated analyses. The microscope was one of the main instruments used during many years in material science and biology. Actually, biology was one of the fields pushing the development of optical microscopy toward a significant improvement of the resolution in the last decade.

Following the Rayleigh criteria, the lateral or in-plane resolution of a microscope is given by the expression [1]:

$$d_{x,y} = \frac{0.61\lambda}{NA},$$

where λ is the wavelength of the light and NA the numerical aperture of the lens or objective, given by:

$$NA = n \sin \alpha,$$

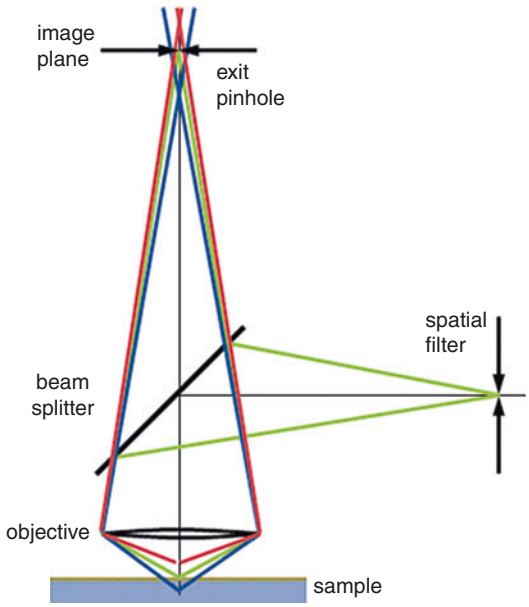
n being the refractive index of the medium (usually air) over the focusing point and 2α the aperture angle. The axial resolution or depth focus is:

$$d_z = \frac{2\lambda}{NA^2}.$$

Even if an objective has a *NA* close to 1 (a very good 100× objective can have a $NA = 0.9$), the resolution limit is $\sim 300 - 500$ nm. Unfortunately, the resolution limit cannot be directly reached with a conventional microscope due to many factors like scattering in the sample surface, optical aberration in the components of the optical system, etc. Further image processing and deconvolution methods can help to improve the spatial resolution [2]. Some of the inherent problems of conventional microscopy have been solved with the help of confocal microscopy.

Confocal Microscopy

The difference between a conventional and a confocal microscope is the illumination (see Fig. 1). In a confocal microscope, two spatial filters are used, one at the entrance, where the light is filtered from the source through a pinhole, and an exit pinhole where the visualization region of the sample is selected (if it is a transparent sample or a living cell, it can be any point inside the field of view of the microscope). The scattered light or the light reflected from any other points in the field of view is rejected by the exit pinhole. With a 100× objective, using a 100 μm entrance pinhole, the image of the pinhole on the sample can be reduced to 1 μm. The field of view of a microscope with a 100× objective is ∼200 μm. In order to obtain an image of the surface of the sample, the exit pinhole must be scanned over the *xy* plane. With a CCD and the appropriate software, it is possible to obtain a 3D image scanning the exit pinhole in the *xy* plane for different depths.



Optical Techniques for Nanostructure Characterization, Fig. 1 Schematic diagram of a confocal microscope. The light is filtered through an entrance pinhole and focused onto the sample with an objective. The collected light is filtered by an exit pinhole. By moving *up* and *down* the exit pinhole, we can collect the light from a different plane in the sample. Usually, either the exit pinhole or the sample is scanned in the *xy* plane in order to obtain an image of the sample on a CCD

Near-Field Scanning Optical Microscopy

The resolution in conventional microscopy cannot be better than $\sim 300 - 500$ nm, since it is limited by the diffraction of the light. There is, however, an ever-increasing demand for optical imaging techniques in the length scale of 1–500 nm, beyond the diffraction limit of the light. On one hand, in materials science, the dimension of many semiconductor devices or structures has been reduced in the last decades to the range of 10–100 nm. There are a number of increasing applications based on semiconductor quantum dots or quantum wires and an increasing demand on optical characterization techniques able to supplying information at these scales. On the other hand, in biology, the interest is focused nowadays in the study of living cells or complex molecules, in the range of a few nanometers. During the last decades, an important amount of imaging techniques has been developed, like scanning electron microscopy (SEM), atomic force microscopy (AFM), or near-field scanning optical microscopy (NSOM or SNOM). The advantage of AFM or NSOM, as compared to SEM, is that the samples under study are not necessarily in vacuum. There are several entries in this book on AFM, SEM, and NSOM. We will give in this section a few basic concepts of NSOM, which will be useful in the next sections.

The basic idea is to perform the measurements with the optical source or detector much closer to the sample than the wavelength of the light. In this "near-field" regime, the resolution is given by the size of the source or detector instead of the wavelength of the light. The advantage of NSOM as compared to AFM is that the interaction of the sample with the light can be analyzed as a function of the wavelength (spectroscopy), or the dynamical behavior of this interaction by means of time-resolved measurements. In near field there are basically two modes of operation; the first one is called aperture NSOM and the second apertureless NSOM or ANSOM [3]. In the first technique, a metalized optical fiber with a very small aperture is utilized to shine light on the sample which is collected from below with a high *NA* objective. The sample is mounted on an *xy* plate in order to map over a finite region and produce an image. The optical resolution is given by the size of the aperture in the optical fiber, which is typically of the order of 50–100 nm. There is an alternative option, which uses the metal tip of an AFM resonating at high frequency. The optical image is collected also with a high *NA* objective, but the signal is filtered using a lock-in amplifier working at the modulation frequency of the tip. Since there is no aperture, the technique is called ANSOM. In ANSOM the signal collected depends on the interaction of the dipole produced by the tip with the sample and the spatial resolution is of the order of the size of the tip, which can be of a few nanometers.

There are other optical techniques allowing the increase of the resolution using far-field optics. A summary of these techniques can be found in Ref. [2]. Some of them (4Π confocal microscopy, image interference microscopy, or harmonic excitation light microscopy) are based on the interference of two or more laser beams and a further image treatment. There are also nonlinear methods and a combination of far-field and near-field optics. The resolution limit in most of the methods is of the order of 100 nm. However, most of these techniques do not supply a direct image of the sample, but a fluorescence image, and they are mainly applied in biology.

High-Resolution Interference Microscopy

The drawback of NSOM is its scanning character; it avoids the collection of information over the entire sample in parallel, and fluctuation on the image can appear. Also, the distance between the tip and the sample has to be kept as constant. Usually in an optical image, only the intensity is taken into account. There is a technique, called phase contrast microscopy, due to Fritz Zernike, where the phase is also considered. This technique is used in biology to increase the contrast in biological systems. It corresponds to the case where the scattered light is small compared to the directly transmitted light. In the case where the scattered light completely interferes with the direct light, there are spatial points in the wave field where the amplitude is zero. Associated to these points, there is a phase singularity. There are two cases of phase singularities, called screw and edge singularities (also called dislocations or optical vortices). In order to obtain singularities in the far field, the object must have a certain width compared to the Abbe resolution limit. Once the singularities appear, their position can be obtained theoretically with an unlimited resolution, high-resolution interference microscope (HRIM). In Ref. [4] a HRIM is described with a lateral spatial accuracy of 10 and 1 nm in depth.

Absorption and Reflectance Techniques

For many years, the standard procedure to characterize the fundamental electronic transition of a semiconductor was the measurement of the absorption coefficient via transmission experiments. In semiconductors, the first optical transition is actually called absorption edge. Transmission cannot be easily measured in the case of semiconductor nanostructures or nanoparticles and alternative techniques must be found out to analyze the absorption edge in that case. In addition, nanostructures present particularities in the absorption as compared to bulk

semiconductors due to the reduction of the dimensionality.

In this section the basic principles giving rise to the absorption in a solid will be given, and some of the techniques commonly used to obtain the optical constants of semiconductor nanostructures will be described.

Optical Constants and Dielectric Function

The response of a material to an electromagnetic wave, in particular in the range from ultraviolet (UV) to near infrared (NIR), satisfies the Maxwell equations. In the most general case, in the ω domain, the electric displacement field can be written in terms of the electric field by means of the dielectric permittivity, $\vec{D} = \tilde{\varepsilon}(\omega)\,\vec{E}$, where $\tilde{\varepsilon} = \varepsilon_1 + i\varepsilon_2$ is a complex quantity ($\vec{E}$ and $\vec{D}$, in general, are not in phase). Usually, $\tilde{\varepsilon}(\omega)$ is named dielectric function, since it is a function of the frequency ω. In an absorbing material, the refractive index is in general complex, $\tilde{n} = n_1 + in_2$, and it is related to the dielectric function through the expression $\tilde{\varepsilon} = \tilde{n}^2$. Both the real and imaginary parts of either $\tilde{\varepsilon}$ or $\tilde{n}$ are named "optical constants" and from its knowledge it is possible to identify the material and to provide useful information on the electronic transitions, impurities, phase transitions, etc. Instead of the electric displacement field, the current density $\vec{J} = \tilde{\sigma}(\omega)\,\vec{E}$, where $\tilde{\sigma} = \sigma_1 + i\sigma_2$ is the complex conductivity $\left(\frac{\tilde{\varepsilon}=\varepsilon_\infty+i\tilde{\sigma}}{\omega}\right)$, can be used in an equivalent formulation. The reflectivity at the surface of the material can be written in terms of the refractive index:

$$\Re = \left|\frac{\tilde{n}-1}{\tilde{n}+1}\right|^2.$$

In semiconductors, instead of the imaginary part of the refractive index (also called extinction coefficient), it is more common to use the absorption coefficient:

$$\alpha(\omega) = \frac{2\omega n_2(\omega)}{c}.$$

Using samples with appropriate thicknesses $d(\alpha d \approx 1)$, it is possible to obtain $\alpha(\omega)$ in a wide spectral range through the Fresnel equations [1], neglecting the dependence of $\Re$ on n_2, which is a very good approximation close to the fundamental electronic transition. Unfortunately, in most of the nanostructure materials, this method is not suitable to measure the absorption coefficient. For metal or semiconducting nanoparticles, an effective medium theory is needed; in the case of semiconductor quantum wells or superlattices, there is always a substrate which most of the times is nontransparent. In other kinds of nanostructures, it is even more difficult to obtain α from transmission experiments. The absorption coefficient α must be obtained from reflectivity or emission experiments in an indirect way.

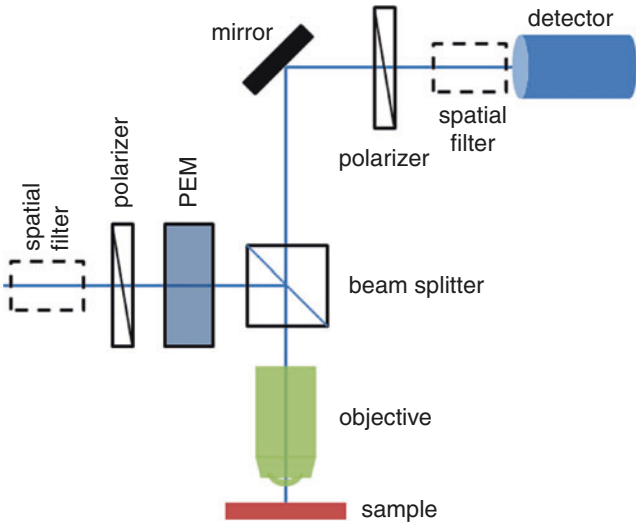
Spectroscopic Ellipsometry

Ellipsometry is an optical technique that analyzes the polarization ellipsis of the light reflected in the sample surface. In the simplest scheme an ellipsometer consists of a light source, a polarizer, an analyzer, and a detector. In spectroscopic ellipsometry usually a Xe lamp and a monochromator is used to supply monochromatic light. In order to obtain elliptical polarization after reflection, the polarization of the source cannot be parallel or perpendicular to the incidence plane (usually, the polarization is rotated 45°). The angle of incidence on the other hand is typically 70° for semiconductors, but it must be chosen carefully, depending on the Brewster angle of the material [5].

The incident and reflected electric field components, parallel (p) and perpendicular (s, from the German *senkrecht*) to the incidence plane, are related through the Jones matrix [6]:

$$\begin{pmatrix} E_p^r \\ E_s^r \end{pmatrix} = \begin{pmatrix} \tilde{R}_{pp} & \tilde{R}_{sp} \\ \tilde{R}_{ps} & \tilde{R}_{ss} \end{pmatrix} \begin{pmatrix} E_p^i \\ E_s^i \end{pmatrix},$$

where $\tilde{R}_{pp}$ and $\tilde{R}_{ss}$ are the complex Fresnel reflection coefficients of the sample for parallel and perpendicularly polarized light. The crossed terms describe the coupling between parallel and perpendicular light and are zero for an isotropic material. From the analysis of the ellipsis of



Optical Techniques for Nanostructure Characterization, Fig. 2 Micro-RDS setup. The polarized light is modulated with a PEM and the collected light is analyzed with a polarizer before detection. The PEM can be located after the polarizer or before the analyzer

polarization, a modulus and a phase can be obtained. If we define:

$$\tilde{p} = \frac{\tilde{R}_{pp}}{\tilde{R}_{ss}} = \tan\psi e^{i\Delta},$$

Δ is the phase difference between the parallel and perpendicular reflection coefficients and the modulus $|\tilde{p}|$ is written as $\tan\psi$ for convenience. Ellipsometry, however, can be used only around or above the fundamental electronic transition, since the imaginary part of the refractive index cannot be negligible as compared to the real part. Furthermore, to interpret the ellipsometry measurements, a realistic model is needed for the dielectric function of the material. These models were developed for anisotropic materials (high- T_c superconductors or semiconductor superlattices) in the 1980s [7]. For nanostructures with two-dimensional or three-dimensional confinement, new models have to be developed based on the Stokes–Müller formalism [5].

Although the measurement spot size can only be reduced to 25 μm, there are many characteristic properties of nanostructures that can be measured with larger spots, whenever the property we are interested in is given by the bundle (in the case of metal nanoparticles with similar sizes or uniform self-assembled quantum dots or quantum wires) or individual access to the nanostructure can be provided (for instance, by isolating a single quantum wire). Quantum confinement can drastically modify the optical properties of nanostructures, in particular, the fundamental electronic transition. Shape effects are also important, since the boundary conditions of the electric field can influence the optical response of nanostructured material. Structural changes in the nanostructure also affect the polarization properties of the reflected light. All these properties can be analyzed with a standard ellipsometer.

Ellipsometry on individual nanostructures has been recently performed by combining a NSOM (detecting through the optical fiber) or combining the ellipsometer with an AFM, enhancing the optical signal with a metallic tip [5]. Future developments go in this direction.

Reflectance Difference Spectroscopy

In reflectance difference spectroscopy (RDS), also termed reflectance anisotropy spectroscopy (RAS), the light incidences perpendicular to the sample surface. The technique measures the difference in the reflected signal between two perpendicular polarizations. A schematic of a RDS setup using a confocal microscope is shown in Fig. 2. The light, polarized at 45°, passes through a photoelastic modulator (PME) with its birefringence axis at 0°. The beam is reflected by a polarization-preserved 50/50 beam splitter down

to the objective lens, being focused on the sample surface. The reflected beam is directed to a second polarizer, set at 0°, and finally detected. The time-independent V_{dc} and the time-dependent V_{2f}, where f is the modulation frequency of the PEM, are simultaneously measured. If the sample is mounted in a xy plate, a RDS image of the sample surface can be obtained. If the light is varied in a spectral range, the spectral variation of the reflectance difference can be analyzed.

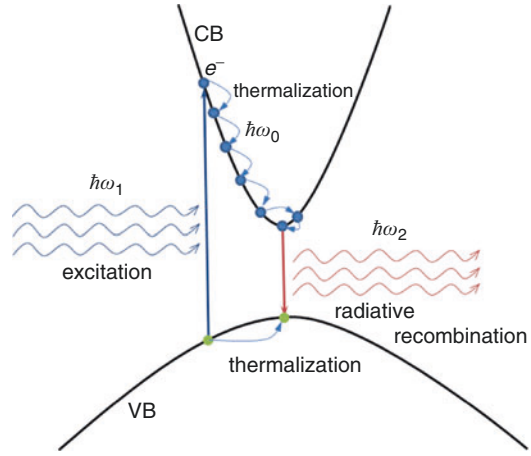
This technique is very sensitive in the analysis of surfaces and interfaces. The interpretation of the RDS data is not straightforward. The anisotropy can be due to spatial dispersion, surface reconstruction, chemical composition (oxidization) of the sample surface, stress, etc. [7].

Emission Techniques

When an electron is excited into the conduction band of a solid, the electron finally recombines with a hole in the valence band. Depending on the manner the electron is excited, the technique acquires a different denomination. If the electron is photoexcited, the recombination is called photoluminescence; if it is excited with an electron beam, it is called cathodoluminescence. In the case of excitation by an electric field, the corresponding process is electroluminescence. In principle, the luminescence decays are exponential or sometimes hyperbolic. But, it is well known that, for some materials, the luminescence glows for hours. This is the case of phosphorescence or thermoluminescence, processes dominated by traps. Only the most important emission processes observed in nanostructures will be discussed here: photoluminescence and cathodoluminescence and in particular their time evolution.

Photoluminescence

In an absorption process, an electron–hole pair is generated by means of a photon. The electron and the hole lose energy by emitting optical phonons, as sketched in Fig. 3, until the electron (hole) reaches the bottom (top) of the conduction (valence) band and finally a photon is radiated. This process is called photoluminescence (PL).



Optical Techniques for Nanostructure Characterization, Fig. 3 Scheme of a typical band-to-band emission process in a semiconductor: a photon of energy $\hbar\omega_1$ reaches a solid creating an electron–hole pair. The electron and hole thermalize emitting optical phonons, and finally the electron–hole pair recombines emitting a photon of energy $\hbar\omega_2$

The PL is, in some sense, the inverse of the absorption, but it is produced at a different energy due to the electron thermalization due to phonon emission. Many times, mostly in biological systems, the PL is called fluorescence, but the physical origin and properties of PL and fluorescence (emission after photon absorption in short decay times) is the same.

In the case of band-to-band transitions, the connection between the emission rate R_{cv} and the absorption coefficient α is given by the Roosbroeck–Shockley relation [7]:

$$R_{cv} = \frac{\alpha(\omega)2\omega^2 n(\omega)^2}{\pi c^2 [e^{\hbar\omega/kT} - 1]},$$

where $\hbar\omega$ is the photon energy, $n(\omega)$ the refractive index, and c the speed of light in vacuum.

There are several phenomena modifying the energy position of the emission or the profile in the particular case of nanostructures. Size effects or quantum confinement can produce a blueshift in the emission. This is a typical effect observed in semiconductor nanoparticles of different diameters, as soon as the diameter decreases. The effect of internal electric fields can shift down in energy

(redshift) the emission peak of the nanostructure. This is the case of the emission in nitride semiconductor quantum dots, for instance (quantum confined Stark effect). The existence of an accumulation or depletion layer in the surface of semiconductor nanowires can also modify the emission profile. Another important issue is the effect of alloying composition in ternary semiconductor nanostructures. Exciton localization due to alloy fluctuations produces an important Stokes shift in the PL.

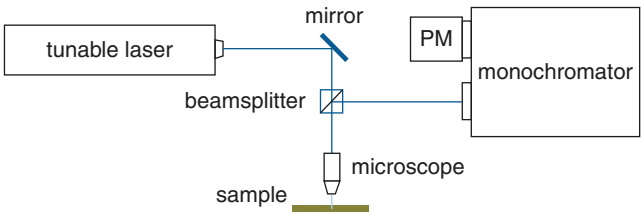
The width of the emission peak of a bundle of nanostructures (like self-assembled quantum dots or nanowires) depends on the dissimilitude of the nanostructures composing the bundle. This width, due to electron–phonon interaction, is called homogeneous broadening. The increase of the broadening due to other factors like alloying fluctuations, defects, etc., is named inhomogeneous broadening. The homogeneous broadening has a Lorentzian line shape while the inhomogeneous broadening is Gaussian.

The emission properties of single nanowires can be measured using several techniques. The most typical method is to separate the nanowires from the substrate by sonicating them in a methanol solution, although they can also be separated mechanically. Large nanowires (from ZnO, Si, GaN, and InP) can be dispersed onto a metallic substrate and directly mapped using optical microscopy. The advantage of a metallic substrate is that the laser heat is dissipated through the substrate and the background does not contain sharp phonon lines that could mask the signal from the nanowires. If the nanowires are small, they can be dispersed on a silicon (or any other semiconductor) substrate marked with a grid by electron beam lithography (EBL). The nanowires can be positioned in relation to the marks by SEM, and from the SEM images, the marks can be located by means of optical microscopy. By studying the PL of a single nanowire, information on the strain distribution, composition, etc., along the nanowire can be obtained. The PL spectra will consist of narrow peaks corresponding to the homogeneous broadening of the emission of individual nanostructures.

EBL has also been used to make submicron holes in an optically opaque metal mask (Cr, for instance) on the surface of material containing self-assembled quantum dots. This enables spatially resolved photoluminescence spectroscopy to be selectively performed on only those dots that lie directly under the aperture. Provided the dot density is small, a single quantum dot can be studied even with a laser spot of 1 μm. Since the structure is just below the aperture, this is equivalent to using a NSOM, i.e., we are analyzing the near field.

Photoluminescence Excitation

The direct measurement of the absorption coefficient in semiconductor nanostructures is not possible in most of the cases. The reason is that the nanostructures are grown on a substrate or they are embedded in a different semiconductor, or simply it is not possible to perform a transmission measurement in order to extract the absorption coefficient. The alternative is to use a technique known as photoluminescence excitation (PLE), assuming that the obtained spectroscopic profile is equivalent, or proportional, to the absorption coefficient. The basic setup for a PLE experiment is shown in Fig. 4.



Optical Techniques for Nanostructure Characterization, Fig. 4 Photoluminescence excitation setup. The laser beam is focused on the sample using a confocal microscope and the emitted light is analyzed with a monochromator

The detection wavelength is selected at the queue of the PL peak using a monochromator and a single channel detector. Then, the excitation wavelength is tuned covering the fundamental transition of the nanostructure. When the absorption coefficient $\alpha(\omega)$ increases, the absorption at the nanostructure increases and consequently the emission intensity also increases. The relation between the emission intensity I_{em} and the excitation intensity I_{exc} is [7]:

$$I_{em} = P_{abs} P_{relax} P_{rad} I_{exc},$$

where P_{abs}, P_{relax}, and P_{rad} are the probabilities of photon absorption, electron–hole relaxation, and radiative recombination, respectively. If electrons and holes relax via phonon scattering, they will lose energy as shown schematically in Fig. 3 until the radiative recombination is produced. In that case, I_{em} can be considered proportional to P_{abs}. If there are non-radiative channels of relaxation, PLE is not proportional to the absorption.

The excitation wavelength can be varied from the UV spectral range to approximately 1 μm in the IR using a dye or a Ti-Sapphire laser, but when an appropriate tunable laser to study the region of interest is not available, it is always possible to use an arc lamp or a halogen lamp in combination with a monochromator as excitation source.

Cathodoluminescence

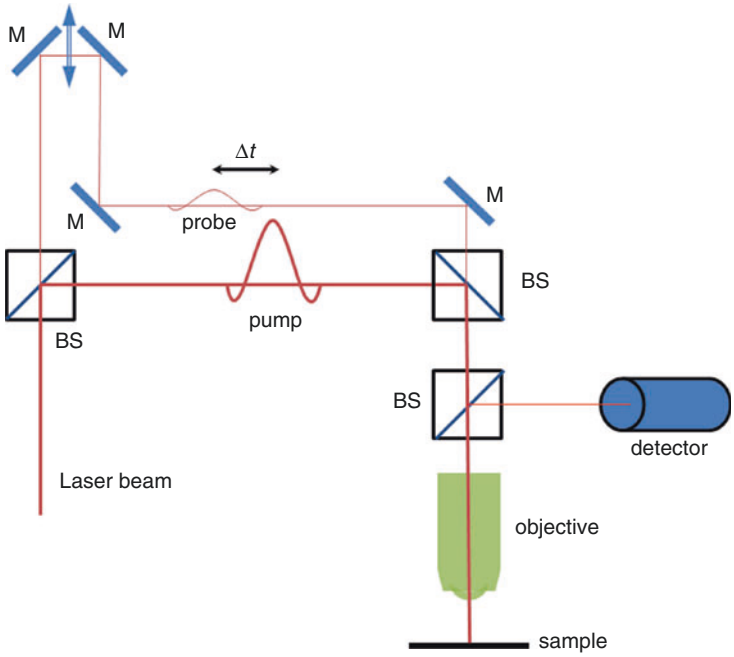
The cathodoluminescence (CL) consists of the analysis of the emission signal in the optical range when excited with an electron beam. The generation of electron–hole pairs when irradiated with energetic electrons occurs through the mechanism of impact ionization. Scanning the electron beam through the sample provides mappings of the sample under study or spectral images point by point if a monochromator is coupled to the system. The spatial resolution in CL is not limited by the diffraction limit of the emitted light, but by the size of the spot of the e-beam. Actually, in bulk materials although the size of the beam can be small (it depends on the voltage applied and the beam current), the diffusion length of the carriers is the dominant limiting factor. In nanostructures, like semiconductor self-assembled quantum dots or nanowires laid on a metallic substrate, the luminescence of individual quantum dots or nanowires can be mapped at the nanometer scale, providing information on the composition, strain, etc. One disadvantage of the technique is that the sample must be under vacuum conditions, thus living cells cannot be studied.

The standard equipment consists of a scanning electron microscope (SEM), either with the incorporation of an optical fiber assembly to extract the optical signal or an external optical microscope with a long working distance objective to collect the light. When the wavelength of the light is below 180 nm, a special setup in vacuum, as that shown in the review article of B. Dierre et al. [8], must be built.

Time-Resolved Photoluminescence

The development of new materials for electronics and optoelectronic applications requires a detailed knowledge of the electron dynamics on very short timescales. In a semiconductor nanostructure, an electron may move ballistically at a velocity of 10^7 cm/s, hence traveling 10 nm in 100 fs. These length and timescales have become of great importance in real devices.

Femtosecond spectroscopy is the most powerful technique to obtain information on electron dynamics on ultrashort timescales, since the electron–phonon interaction, the most important scattering mechanism of electrons in semiconductors, is in the range of femtoseconds. The dynamics of surface plasmon polariton (SPP) excitations in metallic and hybrid nanostructures, which is in the order of 1–100 fs, is also attracting increasing interest. Microscopic resolution can be achieved with confocal microscopy; if an improvement in the resolution is needed, we can use a NSOM or ANSOM. One of the most standard time-resolved techniques used in the study of semiconductor nanostructures and nanoparticles is the pump-probe experiment. A scheme of a pump-probe setup is shown in Fig. 5. A laser beam is divided with a beam splitter into the pump beam, with most of the power, and the probe beam, which arrives to the sample with a temporal delay Δt given by a translation stage. A second beam splitter is used to align the two beams. The response of



Optical Techniques for Nanostructure Characterization, Fig. 5 Pump-probe experiment setup. The laser pulse is divided into two by a beam splitter. The "pump" beam is focused into the sample using a microscope objective, while the "probe" beam is delayed appropriately using a translation stage. The "probe" signal is measured with a detector at the repetition rate of the pulsed laser (in the range of MHz) as indicated in the figure

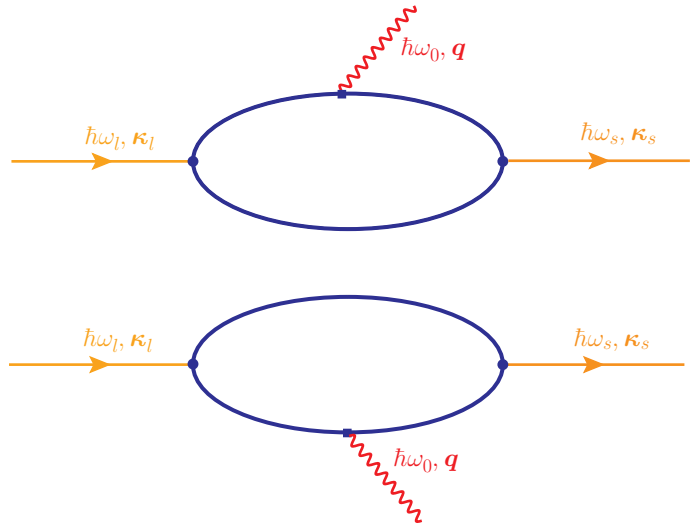
the probe is synchronously detected. Changing the time delay snapshots of the electron dynamics on a femtosecond timescale can be obtained. A review of femtosecond spectroscopy in nanostructures can be found in Ref. [9].

Raman Spectroscopy

Micro-Raman scattering is a nondestructive characterization tool which allows to extract local and valuable information of semiconductor nanostructures. The technique can be applied to analyze first- and second-order phase transitions, phonon anharmonicities, strain states of the nanostructures, etc. Tuning the wavelength, resonance Raman scattering can provide information of single quantum objects. After a brief introduction of Raman scattering, the main techniques based on Raman scattering applied to nanostructures, giving some experimental examples, will be described.

Raman Scattering

Raman scattering (RS) is the inelastic scattering of the light by any kind of quasiparticles: phonons or lattice vibrations, magnon or magnetic oscillations, or electronic excitations (electronic Raman scattering). In semiconductor physics Raman scattering is commonly used for the study of the scattering of light by optical phonons. The direct interaction of light with phonons is produced in the infrared and it is measured by infrared reflectance (IR) spectroscopy. In the Raman effect, the conduction and valence band states participate as virtual intermediate states of the process. Thus, in a normal Raman process, as drawn in the Feynman diagram shown in Fig. 6, a photon produces an electron–hole pair or an exciton. In a second step, the electron or the hole emits or absorbs a phonon, and finally there is a recombination process and the electronic system is reestablished into its fundamental state. Since these are virtual processes, energy is not conserved; only the wave vector is conserved. The only conservation law is given by the Feynman golden rule: the energy of the outgoing photon, $\hbar\omega_s$, must coincide with the sum of the incoming photon energy, $\hbar\omega_l$, plus (absorption of a phonon) or minus (emission of a phonon) the phonon energy, i.e., $\hbar\omega_s = \hbar\omega_l \pm \hbar\omega_0$. In the case of phonon emission (absorption), the outgoing photon has a lower



Optical Techniques for Nanostructure Characterization, Fig. 6 Feynman diagram of a Raman process in a semiconductor. A photon with energy $\hbar\omega_l$ creates an electron–hole pair. In the first diagram, the electron emits or absorbs an optical phonon of energy $\hbar\omega_0$, and afterward the electron and hole recombine emitting a photon with energy $\hbar\omega_s$. In the second diagram, the hole emits or absorbs the phonon

(larger) energy than the incoming one and the process is labeled as Stokes (anti-Stokes) Raman scattering. At low temperature, the population of optical phonons is very low and the signal corresponding to anti-Stokes Raman scattering is very weak.

Since the electronic states are intermediate states of the process, the electronic band structure plays an important role in Raman scattering. In fact, when either the incoming or the outgoing photon is close to a critical point of the semiconductor, there is a tremendous enhancement of the Raman scattering efficiency, since the virtual states become real. This process is called resonant Raman scattering (RRS).

Since not all the sample volume scatters light, instead of the typical cross section used to describe scattering processes, the intensity of the Raman signal is given in terms of the scattering efficiency (scattering cross section per unit volume). For a Stokes process:

$$\frac{dS}{d\Omega} = \frac{\omega_s^3 n_s^3 n_l}{h^2 c^4 \omega_l} \times \sum_{F,I} \left| \sum_{\mu,v} \frac{\langle I|H_{eR}|\mu\rangle \, \langle \mu|H_{ep}|v\rangle \, (v|H_{eR}|F)}{(E_\mu - \hbar\omega_l)\,(E_v - \hbar\omega_s)} \right|^2 (N_0 + 1),$$

where I and F are the initial and final states of the system; μ and v the virtual intermediate states; H_{eR} and H_{ep} the electron–radiation and electron–phonon interaction Hamiltonians, respectively; E_μ and E_v the energy of the intermediate states; $\hbar\omega_l$ and $\hbar\omega_l$ the incoming and outgoing photon energy, respectively; and N_0 the phonon population, given by the Bose–Einstein distribution function $\left(N_0 = \frac{1}{e^{\hbar\omega_0/kT}-1}\right)$. The anti-Stokes process is proportional to N_0. This is the reason why at very low temperatures, the anti-Stokes scattering signal is negligible. Actually, from the rate of Stokes versus anti-Stokes Raman intensities, the lattice temperature of the sample can be obtained:

$$\frac{\dfrac{dS}{d\Omega(\text{Stokes})}}{\dfrac{dS}{d\Omega(\text{anti}-\text{Stokes})}} = e^{\hbar\omega_0/kT},$$

k being the Boltzmann constant and $\hbar\omega_0$ the energy of the corresponding phonon. When the energy of the incoming photon or the outgoing photon coincides with the gap of the material, the Raman efficiency can increase several orders of magnitude. The denominators are actually non-zero because of the broadening of the electronic states; an imaginary term corresponding to the linewidth must be added to the energy terms in the denominators of the scattering efficiency. By tuning the laser energy around an electronic

transition, RRS can be used to determine the corresponding electronic transition.

Raman Tensor and Selection Rules

The electron–phonon Hamiltonian has two contributions: a short-range and a long-range contribution. The short-range contribution, or deformation potential, is due to the local distortion produced by the optical phonon when it propagates through the lattice. The long-range or Fröhlich interaction is due to the change in the polarization vector $\vec{P}$ (actually on $\vec{\nabla} \cdot \vec{P} = \rho_p$, the polarization charges) carried by the polar optical phonons. The deformation potential Hamiltonian is a range 2 tensor and it depends on the specific phonon and the direction of propagation. The Fröhlich Hamiltonian is a scalar. Fröhlich electron–phonon interaction is inversely proportional to the wave vector of the phonon and it is very important under resonant conditions. It is responsible for the multi-phonon emission processes (multi-phonon RRS) and it competes in intensity with photoluminescence in polar semiconductors.

The electron–radiation interaction Hamiltonian contains, through the vector potential, the polarization of the incoming photon, $\vec{e}_l$, and the outgoing photon, $\vec{e}_s$. The Raman intensity is proportional to [7]:

$$I_{\text{Raman}} \propto \left| \vec{e}_l \cdot \Re \cdot \vec{e}_s \right|,$$

where $\Re$ is the Raman tensor of the corresponding optical phonon and it can be written as a 2×2 matrix. The observation of the phonon in a Raman experiment, or Raman selection rules, depends on the mode symmetry and the polarization of the light. Since most of the technological important semiconductors grow in the zincblende (ZB) or wurtzite (W) structure, the discussion on the selection rules will be particularized to these two special cases.

The zone-center optical phonons in the zincblende structure correspond to a triply degenerate three-dimensional representation, usually denoted by T_2 or F_2. The corresponding Raman tensor of the three modes can be written as:

$$\Re(T_2(x)) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & d \\ 0 & d & 0 \end{pmatrix},$$

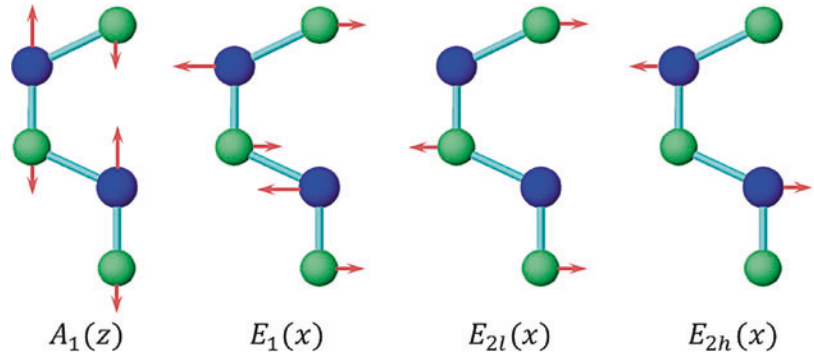
$$\Re(T_2(y)) = \begin{pmatrix} 0 & 0 & d \\ 0 & 0 & 0 \\ d & 0 & 0 \end{pmatrix},$$

$$\Re(T_2(z)) = \begin{pmatrix} 0 & d & 0 \\ d & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

The three axes x, y, and z are along the standard coordinate axes [100], [010], and [001], respectively. The triply degenerated zone-center phonons in the ZB structure split for $q \neq 0$ into a doubly degenerate transverse optical mode (TO) and a longitudinal optical phonon (LO). The reason is the coupling of the transverse electromagnetic field of light with the optical vibrations. In group theory, the zone-center phonons correspond to $q = 0$ and thus are triply degenerated.

The selection rules are written using the notation introduced by S. P. S. Porto [10], indicating the direction of the photon wave vectors and the corresponding polarizations as $\vec{\kappa}_l\left(\vec{e}_l, \vec{e}_s\right)\vec{\kappa}_s$. In the study of semiconductor nanostructures, due to the existence of a nontransparent substrate in most of the cases, the scattering geometries are limited to backscattering and 90° scattering. Forward scattering is only possible if the sample is transparent.

For ZB crystals, in backscattering geometry along the x axis, the Raman tensor $\Re(T_2(x))$ corresponding to the LO phonon can be observed in crossed polarizations, $x(y, z)\bar{x}$. For a TO phonon (the phonon polarization must be perpendicular to the propagation direction), the Raman tensor is a combination of $\Re(T_2(y))$ and $\Re(T_2(z))$. In that case either the laser or scattered polarization must have a component along x. By inspection of the corresponding matrices, it can be checked that the TO phonon mode is forbidden (not observable) on a (100) surface in backscattering configuration, since the light is a transversal wave and cannot have a projection of the polarization vector along the direction of propagation. The



Optical Techniques for Nanostructure Characterization, Fig. 7 Phonons in ZnS (Zn in *green* and S in *blue*) as given by the group theory analysis. Actually, the E_2 modes mix as explained in the text, giving rise to the E_{2l} and E_{2h} Raman modes

selection rules for other surfaces can be found in Ref. [6].

In the case of wurtzite crystals (four atoms in the primitive cell), there are nine optical modes, but two of them are silent modes (only observable with Hyper-Raman scattering). The allowed optical modes are:

$$\Re(E_1(x)) = \begin{pmatrix} 0 & 0 & c \\ 0 & 0 & 0 \\ c & 0 & 0 \end{pmatrix},$$

$$\Re(A_1(z)) = \begin{pmatrix} a & 0 & 0 \\ 0 & a & 0 \\ 0 & 0 & b \end{pmatrix}$$

$$\Re(E_1(y)) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & c \\ 0 & c & 0 \end{pmatrix}$$

$$\Re(E_2) = \begin{pmatrix} d & d & 0 \\ d & -d & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

From these modes (see Fig. 7), the A_1 and E_1 modes are polar modes, since there is a change in the polarizability generated by the atomic movement, while the E_2 mode is nonpolar [11]. In the polar modes, the phonon polarization (direction of the atomic movement) has been indicated. The two E_2 optical modes correspond in principle to the movement of the anion and the cation, respectively, but they mix because they have the same symmetry. In the actual modes, both atoms vibrate; in the E_{2l}, the movement of the heaviest atom (Zn) dominates, while in the E_{2h}, the movement of the lightest atom (S) dominates. These modes can be interpreted as originated from the folder of the zincblende structure in the [111] direction. In the ZB structure, there are actually two modes in the L point where only one of the two atoms vibrates in opposite directions. In the A_1 mode, the atoms vibrate along the c-axis of the wurtzite structure. For a wave propagating along z, the mode is longitudinal ($A_1(LO)$), while for a mode propagating along a direction perpendicular to z, it is transversal ($A_1(TO)$), giving a discontinuity at the Γ point. The same occurs for the $E_1(x)$ vibrating along x; it is longitudinal if the phonon propagates along x, but it is transversal if the phonon propagates along y or z (in other words, the mode is degenerated for a phonon propagating along z, but splits into TO and LO if the phonon wave vector is along the x and y directions). The same applies for the $E_1(y)$ mode. The selection rules for the wurtzite structure can be summarized in Table 1 [11].

In Table 1, all unequal backscattering and 90° scattering geometries and the corresponding Raman allowed modes have been considered.

Confocal Raman Microscopy

In order to analyze thin films or nanostructures, it is essential to use a confocal microscope in order to improve the lateral and depth resolutions; otherwise, the Raman signal of the substrate can completely mask the signal from the thin film. In a micro-Raman setup, the laser beam enters in the system through a spatial filter in order to select the TEM mode. This spatial filter is equivalent to the entrance pinhole of a normal confocal microscope. The collected signal is filtered at the

image plane with a confocal pinhole. The setup of a confocal Raman microscope is basically that shown in Fig. 1. The light at the exit pinhole is focused at the entrance slit of the Raman spectrometer. As in confocal microscopy, by using a motorized *xy* plate and a focusing system, it is possible to obtain three-dimensional Raman maps of the sample.

Direct Raman imaging of a nanostructure or a living cell is also possible if instead of a spectrometer, a band-pass filter is used to select the phonon emission. In order to suppress the stray light from the laser, we need a notch or edge filter.

Coherent Raman Spectroscopies

The Raman scattering technique described in the preceding sections refers to spontaneous Raman scattering, which corresponds to the term $N_0 + 1$ in the expression of the Raman efficiency. Actually, the Raman signal is proportional to this factor. We know that the anti-Stokes signal is proportional to N_0, but the population of phonons is very low even at room temperature. As a consequence, Raman signals are very low. It is possible to surpass this circumstance by stimulated emission. This alternative technique is named "coherent anti-Stokes Raman scattering" (CARS). With two laser beams, of energy $\hbar\omega_1$ and $\hbar\omega_2$, and wave numbers $\vec{\kappa}_1$ and $\vec{\kappa}_2$, we can tune $\hbar\omega_2$ in such a way that $\hbar\omega_1 - \hbar\omega_2 = \hbar\omega_0$. The anti-Stokes scattered signal is produced at (see Fig. 8):

$$\hbar\omega_s = \hbar\omega_1 + \hbar\omega_0 = 2\hbar\omega_1 - \hbar\omega_2$$

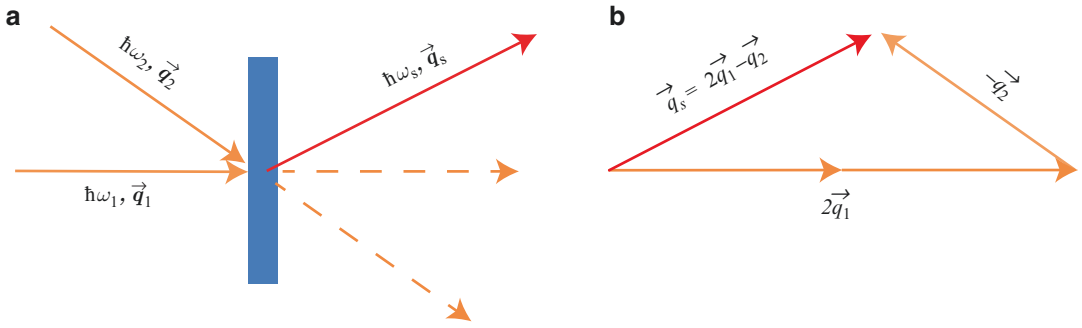
Similarly, we can show that:

$$\vec{\kappa}_s = 2\vec{\kappa}_1 - \vec{\kappa}_2$$

The wave vector of the scattered photon is twice the wave number of one of the incoming photons minus the wave vector of the photon corresponding to the second laser. The signal of a CARS experiment is usually nine or ten orders of magnitude stronger than that of a Raman experiment. The inconvenience of the technique is the existence of suitable tunable lasers in the range of interest. In the visible range, we can use a laser system consisting of a Q-switched Nd:YAG master oscillator, Nd:YAG amplifiers, frequency-doubling crystals, a dye laser, as well as a set of totally reflecting and dichroic mirrors and lenses adapted for purposes of

Optical Techniques for Nanostructure Characterization, Table 1 Possible backscattering and 90° scattering configurations where the Raman allowed optical modes of the wurtzite structure can be observed

Geometry configuration	Phonon mode
$z(y, y)\bar{z}$	$A_1(LO)$, E_2
$z(y, x)\bar{z}$	E_2
$x(y, y)\bar{x}$	$A_1(TO)$, E_2
$x(y, z)\bar{x}$	$E_1(TO)$
$x(y, x)y$	E_2
$x(y, z)y$	$E_1(TO)$, $E_1(LO)$
$x(z, z)y$	$A_1(TO)$
$x(z, x)z$	$E_1(TO)$, $E_1(LO)$
$x(z, y)z$	$E_1(TO)$
$x(y, y)z$	E_2



Optical Techniques for Nanostructure Characterization, Fig. 8 Energy (**a**) and wave number (**b**) conservation in a CARS experiment

O

CARS experiments. The technique has acquired recent interest since the combination with a metallic tip, exciting the plasmons in the material, opens the field of CARS as a local probe [12].

Surface- and Tip-Enhanced Raman Spectroscopy

Surface-enhanced Raman scattering – or spectroscopy – (SERS) and surface-enhanced resonant Raman scattering (SERRS) are extremely sensitive and selective techniques that can be applied to study the Raman response of individual nanostructures or small molecules. These techniques have been used in biochemistry, materials science, or forensic. Although the exact mechanism of enhancement is not completely clear, it is known that the surface plasmon excitations of metallic nanoparticles are responsible for the enhancement. These excitations are produced in the UV or visible range in the case of small metallic nanoparticles. Substrates with ordered nanoparticles can be used, but it is also possible to produce the nanoparticles by chemical methods in a solution, put our sample on that solution, and evaporate the solvent. The diameter of the nanoparticles must be chosen in accordance to the wavelength of the laser used for the Raman experiment. An entry in this book that focuses on SERS is referred to the readers interested in the topic.

A related technique, tip-enhanced Raman spectroscopy (TERS), combines the use of an AFM and a confocal microscope. The Raman signal is enhanced by two mechanisms: (a) the surface plasmon generated at the metallic tip and (b) the lighting rod effect. In the first case, as in the case of SERS, the surface plasmons are excited by the laser, giving rise to an enhancement of the electric field at the tip covered with a noble metal film. In TERS it is essential to illuminate the contact of the tip with the sample and collect the scattered light efficiently. The lighting rod effect consists of the enhancement of the electric field at the apex of the tip when the polarization of the light is parallel to the tip shank. The easiest geometry is to illuminate and collect the light from the bottom using an inverted microscope, but this can only be done if the substrate is transparent or partially transparent, or the nanostructures can be dispersed on a cover slip of a microscope. In the most typical configuration, the illumination of the sample at the tip position is done from the side. TERS is applied nowadays to the study of IV and III-V quantum dots and nanowires, or in biology. A recent review on the topic and perspectives of the techniques can be found in Ref. [13].

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Confocal Laser Scanning Microscopy
- ▶ Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures
- ▶ Optical and Electronic Properties
- ▶ Scanning Near-Field Optical Microscopy

References

1. Born, M., Wolf, E.: Principles of Optics. Cambridge University Press, Cambridge (1999)
2. Garini, Y., Vermolen, B.J., Young, I.T.: From micro to nano: recent advances in high-resolution microscopy. Curr. Opin. Biotechnol. **16**, 1–10 (2005)
3. Richards, D.: Near-field microscopy: throwing light on the nanoworld. Philos. Trans. R. Soc. Lond. A **361**, 2843–2857 (2003)
4. Rockstuhl, C., Märki, I., Scharf, T., Salt, M., Herzig, H.P., Dändliker, R.: High resolution interference microscopy: a tool for probing optical waves in the far-field on a nanometric length scale. Curr. Nanosci. **2**, 337–350 (2006)
5. Losurdo, M., et al.: Spectroscopic ellipsometry and polarimetry for materials and systems analysis at the nanometer scale: state-of-the-art, potential, and perspectives. J. Nanopart. Res. **11**, 1521–1554 (2009)
6. Azzam, R.M.A., Bashara, N.M.: Ellipsometry and Polarized Light. North-Holland, Amsterdam (1987)
7. Yu, P.Y., Cardona, M.: Fundamentals of Semiconductors, 4th edn. Springer, Berlin (2010)
8. Dierre, B., Yuan, X., Sekiguchi, T.: Low-energy cathodoluminescence microscopy for the characterization of nanostructures. Sci. Technol. Adv. Mater. **11**, 043001 (2010)
9. Vasa, P., Roper, C., Pomraenke, R., Lineau, C.: Ultrafast nano-optics. Laser. Photonics Rev. **3**, 483–507 (2009)

10. Rousseau, D.L., Bauman, R.P., Porto, S.P.S.: Normal mode determination in crystals. J. Raman. Spectrosc. **10**, 253–290 (1981)
11. Cantarero, A.: Review on Raman scattering in semiconductor nanowires: I. Theory. J. Nanophotonics **7**, 071598 (2013)
12. Silberberg, Y.: Quantum coherent control for nonlinear spectroscopy and microscopy. Annu. Rev. Phys. Chem. **60**, 277–292 (2009)
13. Yeo, B.S., Stadler, J., Schmid, T., Zenobi, R., Zhang, W.: Tip-enhanced Raman spectroscopy – its status, challenges and future directions. Chem. Phys. Lett. **472**, 1–13 (2009)

Optical Trap

► Optical Tweezers

Optical Tweezers

Martin Hegner, Dorothea Brüggemann and Dunja Skoko
The Naughton Institute, School of Physics, Trinity College Dublin, CRANN, Dublin, Ireland

Synonyms

Laser tweezers; Optical trap

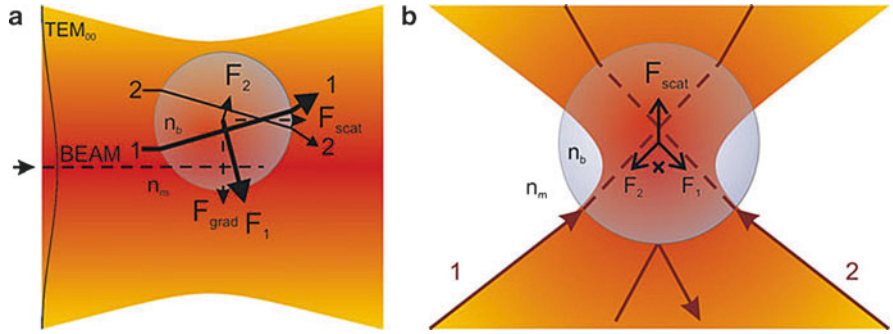
Definition

The technique of optical tweezers and manipulation of small neutral particles by lasers is based on the forces of radiation pressure. These forces arise from the momentum of the light itself.

Overview

It has been first documented by Arthur Ashkin in 1970 that optical forces, or radiation pressure, could be used to trap and accelerate dielectric polarizable micron-sized particles [1]. For these experiments, a stable optical potential well was formed using two, slightly divergent, counter-propagating laser beams. This pioneering study established the groundwork for the well-known optical tweezers (OT) technique, where a single TEM_{00} laser beam is focused by a high numerical aperture (NA) objective lens to a diffraction limited spot.

In the ray optics regime, in which the dimensions of the trapped particle are larger than the wavelength of the laser light, known as the Mie regime, the origin of optical forces can be understood easily. Figure 1 shows the principle of optical trapping in the Mie regime. In Fig. 1a, a dielectric semitransparent sphere with a high index of refraction (n_b; e.g., polystyrene $n_b = 1.59$) is shown. This sphere is immersed in a fluid with index of refraction n_m (e.g., water n_m ~1.33). When exposed to a slightly focused laser beam with a Gaussian intensity profile, the sphere is drawn into the high-intensity region of the laser and then accelerated along the beam axis. Consider two rays "1" and "2" impinging on the surface of the sphere in a symmetrical manner about its center. At this moment, surface reflections are neglected. Most of the rays get refracted through the bead and give rise to forces F_1 and F_2 in the direction of momentum change. For theoretical consideration on the forces arising, the reader is referred to Grange et al. [2]. Due to the fact that the intensity of ray "1" is higher than that of ray "2," the force F_1 is higher than F_2. Adding all such symmetrical pairs of rays striking the sphere, it is imaginable that the force can be separated into two components, a gradient force, F_{grad}, and a scattering force, F_{scat}. The gradient component arises from the electric field gradient pointing toward the highest intensity region of the beam. The scattering force component, F_{scat}, which is pointing in the direction of the laser beam, is caused by the photons scattering on the surface of the sphere. In Fig. 1b, stable optical trapping in three dimensions is illustrated. If the laser beam is highly focused, a backward light force is generated. It can be shown that stable trapping occurs along the optical axis when the gradient force, which is composed of the sum of the two forces F_1 and F_2, overcomes the scattering force F_{scat}. This explains why (1) high NA



Optical Tweezers, Fig. 1 *Optical forces*. (**a**) Origin of scattering force, F_{scat}, and gradient force, F_{grad}, for a dielectric high index of refraction (n_b) sphere in a medium (n_m) in a mildly focused laser displaced from the TEM_{00} beam axis. The gradient component is oriented orthogonally to the axis of the laser beam, and the scattering is parallel. (**b**) The axial trapping requires highly convergent rays. A two-dimensional ray optics model illustrates the scattering and the gradient forces in the Mie regime. Since the axial trapping is weakest in the single beam configuration, counter-propagating laser traps have been developed [2, 16]

objective lenses have to be used and (2) the back aperture of the objective lens has to be overfilled to create the steepest possible gradient. When a particle is stably trapped, the sum of all forces acting on it is zero. If the particle is slightly displaced from the equilibrium point by applying an external force, the gradient force acts in the direction of the field gradient toward the center of the beam. This force depends linearly on the displacement similarly to a spring that follows Hooke's law. Optical forces are however very small, since 100 mW of laser power at the focus only produce forces of tens of piconewtons. At the focal point, not only dielectric spheres can be trapped, but also biological organisms such as cells, viruses, or bacteria as shown in 1987 by Ashkin and coworkers. The trapping of objects much smaller than the wavelength of the laser light, such as cooled atoms, has been demonstrated. This work resulted in the Nobel Prize in physics in 1997. Trapping cooled atoms cannot be understood by simple ray optics. For a more detailed insight into the principle of trapping atoms, the reader is referred to the entries of this encyclopedia where atomic clocks or Bose-Einstein condensates are discussed. For a good overview on the evolution of optical trapping, a review from Ashkin is recommended [3].

Choice of Light/Laser Source

The high-intensity radiation of the laser beam used in optical trapping experiments has the potential of effecting materials exposed to the radiation. In some experiments, laser power of more than a hundred milliwatts (mW) is used as a source. When focused to a micron-sized region, the light power in the trapping spot is equivalent to 10 MW/cm^2, which is over 600 million times greater than the flux of solar radiation on the Earth's surface on a sunny day. For chemical or other physical experiments, where the material manipulated is being modified, lasers with a wavelength in the visible regime are used. For biological research, laser light is in the near-infrared regime (NIR: $\lambda \sim 800$–1,100 nm) and the laser power is chosen depending on the system under investigation. In the high-power NIR, single biomolecules are less susceptible to optical damage than whole cells. The local heating of the surrounding liquid is in the range of only a few degrees Celsius. However, Steven Block's lab demonstrated in 1999 that live microorganism cells were affected by the radiation and showed physiological photodamage of live *Escherichia coli*.

Another technique, laser-tracking microrheology (LTM) was developed in the

mid-1990s to probe the mechanical properties of living cells by examining cellular viscoelastic properties [4]. Not only cells can be probed but also rheological properties of liquids in general. Micron-sized trapped objects such as beads or individual cells are moved with OT and can be used to generate fluid flow in two and three dimensions around the item and this can be combined with rheological measurements. Thereby, the bodies' displacements from the trap center are analyzed during the motion or when the optical trap is temporarily switched off. Such LTM measurements are highly reproducible, minimally invasive, and can be extended to in-depth studies of the rheological properties of liquids, cells, or internal cell compartments and microstructures in lab-on-a-chip devices [5].

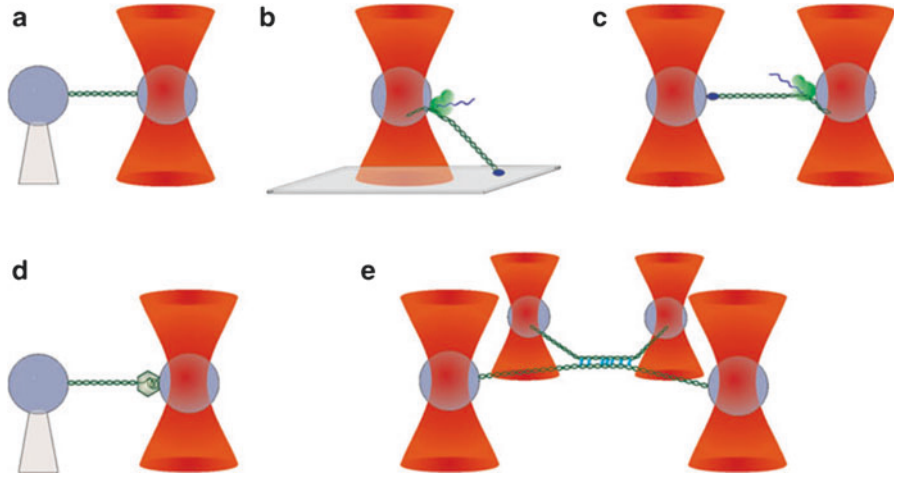
Meanwhile, OTs are a focal point for interdisciplinary research ranging from condensed matter physics and physical chemistry to biology [6]. The fast emerging applications for OT have led to various technical innovations, thus improving the versatility of optical traps. The instrumental advances involve rapid position detection, sample positioning, trap stiffness determination, applying of controlled and calibrated forces. Multiple laser beams and computerized automation of laser beams have been used to study quasicrystals as well as polymer, metallic particles [7] and also aerosol droplets in air [8].

Most commonly TEM_{00} lasers with a Gaussian intensity beam profile are used to form a trap (see Fig. 1a). But physicists also started to use other beam profiles, such as Bessel beams, to enable the manipulation of micrometer-sized particles in multiple planes. Asymmetric beams for instance are used as tools to study the transfer of angular momentum from light to particles for their rotation. An overview on such developments is given in Ref. [9], where a series of experts in the field describe their optical experiments.

Setup of an Experiment

The interest for biologists in the OT technique comes from the fact that minute forces can also be measured with sub-pN accuracy and nanometer spatial resolution on the trapped object. Since such small forces are not accessible by conventional techniques, such as scanning-force-microscopy-based techniques (in aqueous/liquid environment, the natural "habitat" for many both in vitro and in vivo biological systems), OT has become a major investigating tool in biology (for recent reviews on instrumentation in optical trapping see [10–13]).

Biological molecules are often in the nanometer regime and are too small to be directly trapped by OT. The key to such experiments is the immobilization of the biological molecule of interest, such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and protein or molecular motor, on the surface of a micron-sized sphere. The sphere serves as a handle for optical manipulation. In the commonly used OT setups, interacting partner molecules or the other end of the manipulated molecule is attached either to a cover slide or to a second bead (which can also be trapped in three dimensions) see Fig. 2. The external control parameters, which can be manipulated in such experiments, are the end-to-end distance of the investigated biological system linked to the spheres/surfaces or the externally acting force in the range of 0.1–150 pN. In experiments one can vary the distance between the spheres, and then the optical force acting on the bead is recorded. The experimental plots of distance-versus-force give direct access to the mechanical properties of elongated molecules. Force-extension graphs are typically compared to worm-like chain or freely jointed chain models for polymers, and parameters that describe the polymer are extracted – most notably the persistence length, which describes the behavior of the polymer under thermal fluctuations. In the other experiments, where the position of one sphere is fixed and the force is kept constant, the molecular extension is the variable under observation. Such experiments allow direct examination of the kinetics of the involved molecules, may it be a biomolecular motor traveling along a defined path of another biomolecule (e.g., RNA polymerase along a DNA molecule [14]) or a change in molecular mechanics as observed in protein polymerization on DNA [15].



Optical Tweezers, Fig. 2 *Geometry of optical tweezers systems used in single molecule experiments*. (**a**) One end of the DNA molecule is attached to a microsphere held in the optical trap and the other end to a bead held by a glass micropipette by suction. Pulling force is generated by moving the micropipette, which is mounted on a piezoelectric stage [2, 15, 16, 24]. (**b**) RNA polymerase is attached to the trapped microsphere and to one end of the transcribed DNA. The other end of the DNA is attached to the glass slide of the flow cell mounted on the piezoelectric stage. Force is generated by moving the piezoelectric stage [12, 19]. (**c**) Dual optical traps are used to eliminate micromechanical noise coming from the glass slide or micropipette. Individual 0.34-nm base-pair steps of RNA polymerase translocation are observed in this setup (see Fig. 3c) [14]. (**d**) Geometry of the OT used for the investigation of bacteriophage φ29 DNA packing motor [20]. (**e**) Sequences of nonspecific protein-DNA interactions are investigated using four optical traps. The DNA-H-NS-DNA complex is formed by trapping two DNA molecules simultaneously in the presence of nucleotide-associated protein H-NS. Schematics are not drawn to scale

The single molecule approach to biology offers distinct advantages over the conventional approach of taking ensemble measurements. Parameters such as individual kinetics or motion can be investigated (e.g., observing a single molecular motor at work). One should keep in mind that the setup of these intricate experiments must be done in series, which is very time consuming. The required purity of the solutions and biological molecules is very important and different from "normal" molecular biological experiments where ensembles are investigated. The arrangement of an experiment in the OT starts with the production of modified molecules at a nano-molar concentration (~10^{11}/mL). The spheres thereby serve as handles for manipulation in the OT device. The strongest linkage of biomolecules to the interface of the spheres requires covalent binding and chemical activation steps. Each additional step reduces the amount of successfully anchored molecules on the interface. Therefore, a bigger quantity of modified biomolecules has to be produced at the start to end up with one molecular complex available for the OT experiment. Ligand-receptor pairs, such as biotin-streptavidin, histidine tags nitrilotriacetic acid (NTA), or antibody-antigen pairs (e.g., digoxigenin and anti-digoxigenin), are commonly used to provide specific, oriented, and tight binding to microspheres. Biomolecular interactions do not have to be activated like chemical reactions. This allows longer incubation periods to complete the interactions and often provides a successful coupling of the molecule to the surface of the spheres. Constant force experiments with forces exceeding 40 pN for long duration require either biotin-streptavidin or covalent linkage. The biomolecular "bonds" of single antibody-antigen pairs unbind after short periods while being pulled. The externally applied forces lower the activation barrier; therefore, the ligand unbinds faster from its receptor. To strengthen such interactions, some

researchers apply more than one of these bonds in series. This principle has already been used in DNA pulling experiments where multiple digoxigenins were introduced as tags on the DNA to anchor the molecule to the antibody-activated spheres [16].

Biological Applications of Optical Tweezers

In the following sections, experiments with OT, which are used for holding and manipulating a variety of nanometer- to micron-sized biological objects such as DNA, single molecules, membrane bilayers, intracellular structures, vesicles, and plant cells, are reported [12]. Trapping these distinct objects with laser tweezers enables measurements of their mechanical properties such as elasticity, stiffness, rigidity, torque, and dynamic behavior.

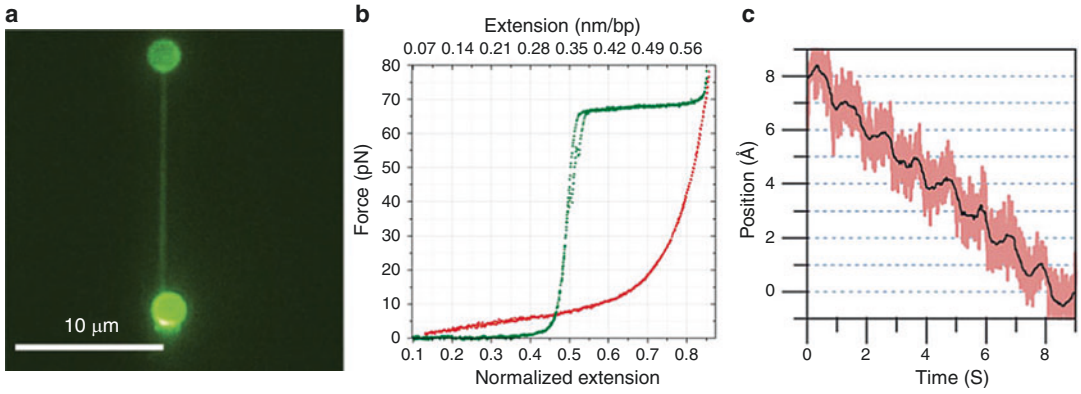
Single Molecular Systems in Optical Tweezers

To date several different homebuilt OT systems have been designed to manipulate single molecules of DNA and RNA, to investigate interactions between DNA and proteins and to measure the forces generated by single molecule motors. Thereby, the position of the trapped microspheres, which act as handles, can be tracked with sub-nanometer spatial resolution and sub-millisecond temporal resolution [17, 18].

Bustamante and coworkers performed one of the first single molecule experiments with OT in order to study DNA elasticity. In their assay, one end of the manipulated DNA molecule was attached to a bead held in the optical trap and the other end to a bead, which was held by a glass micropipette by suction (see Figs. 2a and 3a) [2, 16]. This assay gave detailed insight into the mechanical properties of double-stranded DNA (dsDNA) molecules for forces between 1 and 100 pN [16] (Fig. 3b).

Protein-DNA interactions are the most fundamental biomolecular interactions in the living cell. DNA elasticity plays an important role in the interactions with proteins and protein complexes. It was demonstrated that by cooperative protein binding to DNA, proteins act as a powerful molecular machine [15]. The seminal experiments, which focused on protein-DNA interactions involved in transcription, copying, and packaging of DNA, were performed around the turn of the century by the Block or Bustamante labs [19]. Most of these processes are performed by molecular motors, i.e., proteins with dynamic structures capable of converting energy stored in nucleoside triphosphates (NTPs) into mechanical work. For example, RNA polymerase is an enzyme that constructs RNA chains from DNA as templates in a process called transcription. Single molecule studies performed with OT gave comprehensive insight into the stall force, transcriptional pausing, and backtracking of RNA polymerase. Please refer to the review papers [18, 20] for a good overview. One of the OT geometries used for studying DNA transcription is shown in Fig. 2b. The transcribing RNA polymerase was attached to a microsphere held by the optical trap and the DNA is attached to the surface of a flow cell. The position of the flow cell could be precisely controlled by a piezoelectric stage. During transcription, the polymerase moved along the DNA and therefore shortened the distance from the sphere to the end of the DNA. To maintain a constant force on the RNA polymerase, respectively on the trapped bead, the piezoelectric stage moved along the direction of the polymerase to keep the bead in the laser trap at a constant position (i.e., at constant force). In both experiments, the DNA elasticity experiment (Fig. 2a) and the RNA polymerase experiment (Fig. 2b), the biomolecules were linked to a surface, which was "connected" to the external mechanical environment of the OT (micropipette or flow-cell wall). This connection introduces a positional noise of a few nanometers to the experiment due to the inherent thermal motion of the mechanical objects to which the biomolecules are attached.

A much more stable OT design was obtained by decoupling the spheres from mechanical objects introducing a second optical trap into the system. Instead of using a micropipette or a glass slide, a high-power laser beam was split optically to receive another attachment point with a second sphere, as illustrated in Fig. 2c [17, 18]. One of the beams was



Optical Tweezers, Fig. 3 *Single molecule experiments*. (**a**) Typical image of a combined fluorescence – OT experiment of DNA being stretched by externally applied forces. The DNA molecule in this experiment has a length of approximately 16 μm. The spheres are functionalized with streptavidin or are chemically activated to bind the ends of the modified DNA ends covalently. After successful tethering, the biomolecule is exposed to fluorescent dyes (i.e., *SYBR® green*) and extended. The individual laser power of each of the NIR ($\lambda = 830$ nm) counter-propagating trapping lasers is ~85 mW at the trap. The molecule can be stretched and relaxed numerous times without any change in its mechanics. To visualize the fluorophore-stained DNA, the molecule is illuminated for a couple of seconds by the built-in Argon laser ($\lambda = 488$ nm) and can be visualized with a confocal electron multiplying CCD. If the illumination time by the Argon laser (~4 mW) at the trap is increased to few tens of seconds, the molecule is optically damaged and ruptures due to breakage of covalent bonds. (**b**) Typical force-versus-extension curves of unlabelled double-stranded DNA (contour length = 0.5), and single-stranded DNA (contour length = 1) in buffer (10 mM HEPES pH 7.2, 1 mM EDTA, 150 mM NaCl). Curves are normalized to the contour length of single-stranded DNA (assuming a base-to-base distance of 0.7 nm). The DNA base-to-base distance (obtained by multiplying the normalized extension by 0.7 nm) projected onto the direction of the applied force is shown at the *top* [2, 15, 16]. (**c**) RNA polymerase moves in discrete steps along DNA molecule: Steps resolved for a stiffly trapped bead moved in 1-Å increments at 1 Hz. Data were median filtered with a 5-ms (*grey*) and 500-ms (*black*) window [14]

then moved by steering the beam with one precise optical mirror. This bead-DNA-bead dumbbell configuration allowed Block and coworkers to observe amazingly individual 0.34-nm base-pair steps of RNA polymerase translocation [14] (see Fig. 3c).

When viruses are reproduced by infected cells, they are reassembled from proteins which form a shell (i.e., a capsid) and from their genetic material, which they need to infect subsequent cells. The packaging of viral DNA by a rotary protein motor into the viral capsids was investigated with OT in 2001 by the Bustamante lab [20]. One of these studies (Fig. 2d) revealed that a bacteriophage φ29 motor can work against forces of up to 57 pN, making it one of the most powerful molecular motors known.

Optical traps have been used to probe the motion and mechanisms of various nucleic acid enzymes such as exonuclease, proteins that cut nucleic acids; helicases, proteins that unwind double-stranded DNA; and polymerases, proteins that synthesize RNA and DNA [18]. Today not only polymerase activity on DNA has been investigated on a single molecule level but also the translational activity of ribosomes acting on RNA or polypeptide chains [21, 22].

The genetic material of living cells is packed into condensed structures called chromosomes, the precise architecture of which is yet to be fully elucidated. The great complexity of chromosomal assembly is based on interactions between DNA and numerous proteins that bind DNA in sequence-independent manner. Optical tweezers proved to be an excellent tool for studying these interactions. In one study, researchers combined two optical traps to capture a pair of double-stranded DNA molecules and characterized protein-DNA complexes created in the presence of the DNA-bridging protein H-NS (Fig. 2e) [12].

Apart from protein-nucleic acid interactions, many biological interactions were investigated at

a single molecule level. Kinesin and dynein, molecular motors that transport cellular cargo by "walking" along microtubules, and myosin, a molecular motor that moves on actin filaments and generates the forces in muscles, have been investigated in depth for many years. The interaction forces of these motor proteins were among the first ones to be elucidated by OT [23].

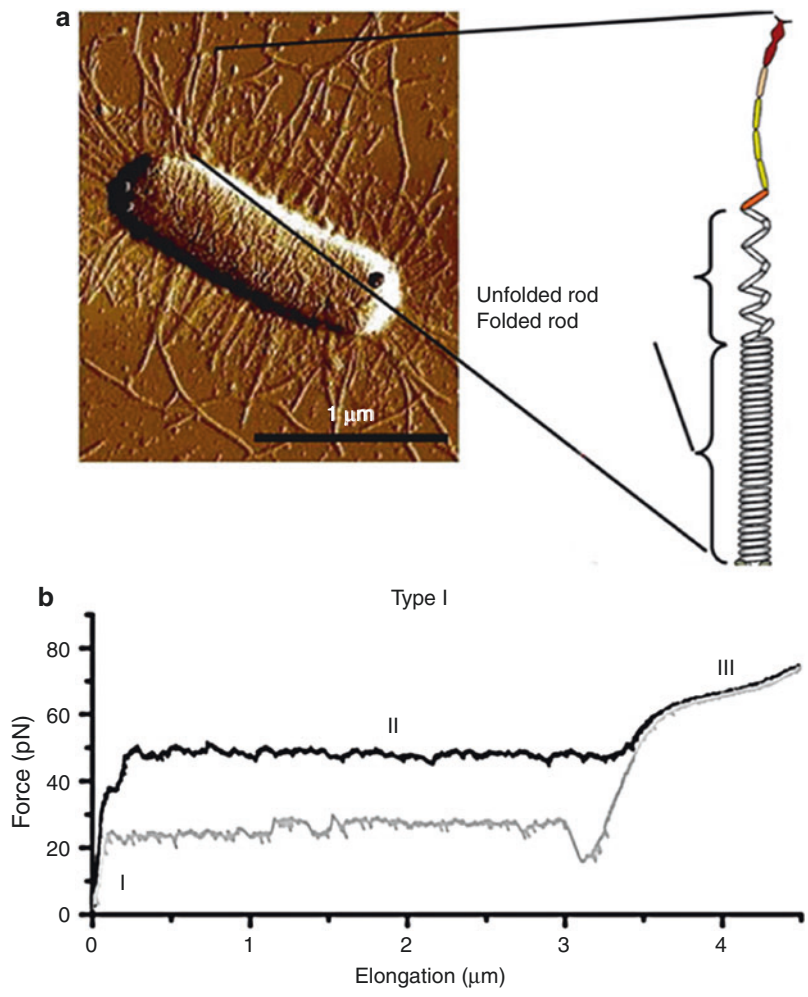
Since the size of biomolecules is in the order of nanometers, they are too small to be observed by classical diffraction limited optical microscopy. To overcome this limit, biomolecules can be fluorescently labeled and visualized using fluorescence microscopy. Fluorescently tagged molecules can be detected one by one. However, measuring single molecule forces and fluorescence at the same time is difficult due to the high level of background light from the optical traps, which is ten orders of magnitude greater than the light emitted by a single fluorophore. Over the last 5 years, instruments capable of simultaneous spatially coincident optical trapping and single molecule fluorescence were introduced to gain a more detailed insight into the nature of biomolecules. This combination was enabled by using improved optical designs, high-performance spectral filters, and automated rapid data acquisition [24, 25].

When applying a mechanical load on single molecules with OT, their response could be monitored using fluorescent probes. In Fig. 3a, an experiment on a single dsDNA molecule is shown [2]. A DNA molecule of 16-μm length was labeled with SYBR® green fluorophores and subsequently stretched with OT. The polystyrene bead in the upper part of the image (diameter 2.9 μm, coated with streptavidin) was trapped by two counter-propagating laser beams and acted as an anchoring point for the biotinylated end of the DNA. For such typical stretching experiments, the bead on the micropipette (with aperture <1 μm), to which the other end of the DNA was covalently coupled, was moved with nanometer precision by a piezoelectric stage until the DNA was extended to ~75 % of its contour length. Thereby, the force was raised above 1.5 pN and the DNA was then visualized in the confocal image plane of the microscope arrangement [2]. When extended further, the force on the molecule, which was measured by the change in light momentum flux, raised till the molecule reached a value of approximately 68 pN. At this force, the molecule undergoes a structural change and could be extended without much force to almost 1.7 times of its contour length [15, 16]. Further examples of combining optical traps with other analytical techniques will be presented later in this report.

Cells and Optical Tweezers

As mentioned before, one has to be careful with the choice of wavelength and the power of the laser used in optical trapping experiments. This is even more pronounced when whole cells are investigated. First experiments with whole cells, where cell membrane mechanics were investigated, were performed by the Sheetz group in the mid-1990s. A valuable application of OT could for instance be found in the study of bacterial motility and their adhesion to surfaces. Many types of bacteria express micrometer-long attachment organelles, the pili, which adhere to the host tissue and have an ability to uncoil and recoil. The pili facilitate bacterial conjugation, which is the exchange of genetic material from cell to cell. The folding mechanisms of piliated bacteria could be studied with force-measuring OT at single molecule level at a resolution in the low pN range [26]. For the first time, this method allowed the operator to assess forces mediated by individual pili on single living bacteria in real time. Recently, this technique was used to study how uropathogenic *E. coli* can withstand shear forces in a flow configuration [27]. Thus, force-measuring OT could help to understand these mechanisms enabling to fight infections caused by antibiotic-resistant bacteria. In Fig. 4, an image of a bacterium, which was manipulated with an OT, is shown.

Besides *E. coli*, also other mammalian cells such as red blood cells, nerve cells, and gametes were studied extensively by optical trapping [28]. Furthermore, OT could be used to spatially fix subcellular structures, which normally move inside a cell. Being optically trapped, these structures could be investigated with confocal microscopy. A recent example of such laser-based



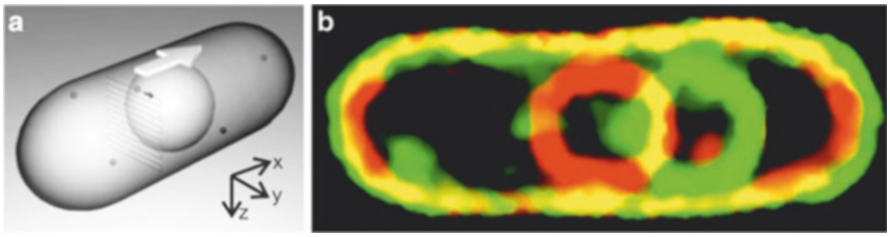
Optical Tweezers, Fig. 4 (**a**) AFM image of a HB101/pPAP5 bacterium expressing pili. The schematic drawing shows the folded and unfolded parts of the pili. (**b**) Typical force-versus-elongation response of a single pili during unfolding (*upper curve*) and refolding (*lower curve*) at an elongation speed of 0.1 µm/s [27]. In region I, the pili are elongated up to a fraction of its relaxed length, and a linear force-elongation response is observed. In the following steady-state region II, a constant force is applied to the pili so that the unfolding process takes place. Region III originates from an overstretching of the pili

nano-surgery was the displacement of the cell nucleus to study the relevance of its intracellular position for the cell division and the resulting cell shape and size [29]. How internal cell structures like a nucleus can be manipulated by OT is shown in Fig. 5 [30]. Here cell internal particles, which diffract light (such as granules or organelles), can be trapped and then be displaced within the cell body. This mechanical motion induces effects onto surrounding structures such as the cell nucleus, which can be monitored.

Trapping outer cell components can also be realized using OT. These studies gave insight into essential cell functions such as cell motility, vesicle-mediated release or uptake of soluble and membrane components, or nanotube-mediated cell-cell communications [31]. Analyzing cell membranes with OT could even help to promote aging research, for instance by exposing endothelial cells to vertical forces using optical trapping. This experiment simulated the effects of high blood pressure to the blood vessels, and cell stiffening was observed indicating that the cells undergo complex morphological changes under the effect of external pressure [32].

Optical Stretching of Cells

Normally, forces acting on a small scale like a single molecule or an organelle are investigated by optical tweezers, as reported above.



Optical Tweezers, Fig. 5 *Scheme of the cell nucleus being displaced due to optical tweezers*. (**a**) A lipid granule (*small sphere*) is trapped close to the nucleus (*large sphere*). The small granule is moved along the shown raster-scanning trajectory in the y-z plane (*small arrow*). At the same time, it is shifted in x direction (*big arrow*), thus pushing the adjacent nucleus forward in x direction [30]. (**b**) Superposition of two images of a single cell: before (nucleus in the *centre*) and after (nucleus moved to the *right*) the manipulation of the nucleus. The nuclear envelope and cell membrane were marked with GFP [29]

Nevertheless the forces acting on whole cells are also important, for instance during cell differentiation. To enable whole cell investigations, the optical stretcher, where two opposed laser beams are used to trap a cell, has been developed. In this experiment, no point force acts on the object. The photons generate surface forces at the opposite membranes and act on the whole structure, thus leading to a stretching of the cell along the beam axis [33]. When stretched, the elasticity of a whole cell can be measured and yield its Young's modulus. Thereby, the cell damage is reduced significantly since the optical stretcher only uses unfocused laser beams. This allows the application of higher laser powers to biological materials. Hence, the forces exhibited by an optical stretcher bridge the gap between OT and atomic force microscopy.

The optical stretcher has already been employed to stretch human erythrocytes, mouse fibroblasts, and PC12 cells [33]. In recent studies, the viscoelastic properties of cells have been correlated with their metastatic potential so that optical stretching has become a valuable technique for gaining new insights into the processes of cell migration and metastasis.

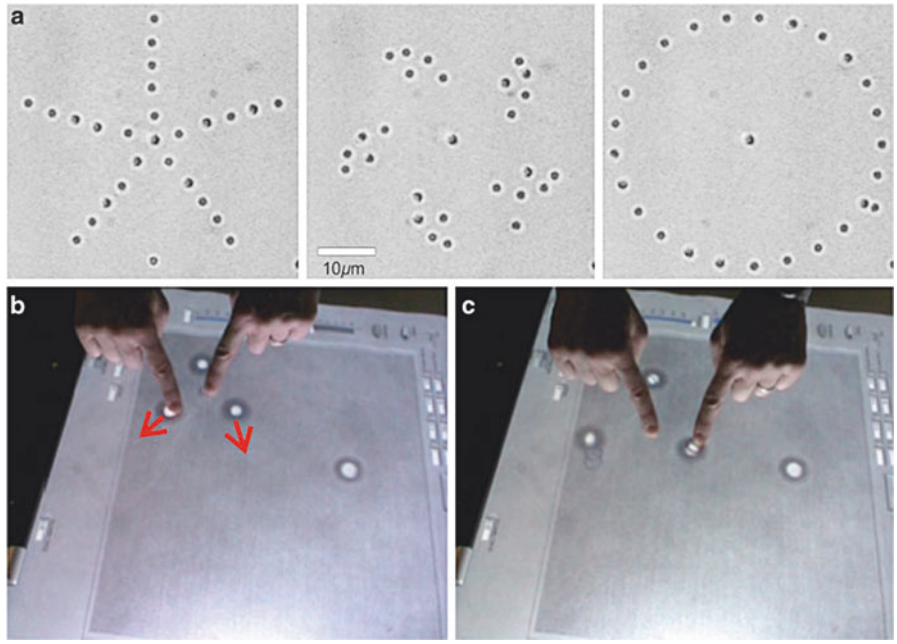
Combining Optical Trapping of Cells with Other Optical Techniques

As already mentioned before, nowadays, OT is often combined with other analytical methods to form powerful tools in single molecule studies. Besides fluorescence microscopy, Raman spectroscopy could be used simultaneously with OT. In a recent study, red blood cells were placed under stress with OT to simulate the stretching and compression the cells experience when passing through small capillaries. Simultaneously, the oxygenation state of the blood cells was monitored by Raman spectroscopy revealing a transition between the oxygenation and deoxygenation states of the cells when being stretched [34]. Raman spectroscopy combined with optical trapping could also be used to identify micrometer-sized particles, e.g., bacterial cells, in an aqueous environment. A single cell was immobilized in the laser trap and then the collection of its Raman signals could be maximized without any unwanted background signals from other particles or substrate surfaces. Thus, the OT analysis of the relative concentration of each type of organism in the aqueous solution was possible [35].

Holographic Applications of Optical Tweezers

With multiple laser beams, an array of several particles can be trapped at the same time. Such spatial arrangements of traps, also known as holographic tweezers (examples are shown in Fig. 6), can be achieved with two different methods:

(a) Scanning between trap positions: A single beam scans rapidly between several discrete



Optical Tweezers, Fig. 6 *Dynamic holographic tweezers in action*. (**a**) Polystyrene spheres with 990 nm in diameter are trapped simultaneously in dynamic tweezers and moved in the same plane from a star pattern into a circle [6, 37]: (*left*) the original configuration; (*middle*) after 16 steps; (*right*) the final configuration after 38 steps. (**b**) An arrangement of SiO_2 spheres is optically trapped, and the spheres can be seen in a multi-touch interface for holographic tweezers. (**c**) The arrangement was moved simultaneously in the direction of the *arrows* shown in **b** [39]

trap positions, which can be computer controlled. Thus, the single beam follows separate light paths, that are recombined before the beam enters the microscope. If the laser beam switches fast enough between the distinct positions, multiple objects can be trapped simultaneously because the viscous drag on them is high enough to keep them in the defined positions while the laser switches to the surrounding positions [36]. However, the number of objects which can be trapped with this technique is limited by the diffusion of the objects and the switching speed of the laser.

(b) Separate laser beams: A computer-addressed spatial light modulator (SLM) is employed to impose a predefined phase shift on the incoming laser beam. Thus, the coherent beam is split into many separate beams, each of which is focused into an optical trap by a strongly converging lens [6, 37, 38]. Computer-generated holograms thereby configure the resulting pattern of traps, and each trap can have individually specified characteristics, arranged in arbitrary three-dimensional configurations. Recently, a novel multi-touch interface for the interactive real-time control of holographic OT was presented to control numerous optical traps independently but simultaneously [39].

With permanent three-dimensional arrangements of OT, several new applications arose. Tweezer-organized structures can for instance be fixed in place by sintering or gelling, depending on the material surrounding the trapped objects. In this nanofabrication process, the intense illumination of the laser tweezers can even be used to drive spatially resolved photochemistry for fabricating small complex three-dimensional structures [6]. Moreover, three-dimensional configurations

of isolated cells such as *E. coli* have been obtained by trapping the cells within a gelatin matrix. When the trap was switched off and the gel was set, a permanent and vital cell architecture was maintained for several days, thus opening new perspectives in cell and tissue engineering using OT [40]. With such results, the development of holographic OT promises exciting new opportunities for research engineering and manufacturing at mesoscopic length scales.

Key Research Findings

The molecular actions of individual molecules or cells can be investigated in optical tweezer experiments in buffers that are very similar to their native environments. Experiments reach an accuracy of sub-nanometer resolution applying sub-piconewton forces. By scaling up to multiple traps using holographic techniques, new manufacturing of three-dimensional structures at mesoscopic length scales is possible.

Cross-References

► Atomic Force Microscopy

References

1. Ashkin, A.: Acceleration and trapping of particles by radiation pressure. Phys. Rev. Lett. **24**(4), 156–159 (1970)
2. Grange, W., et al.: Optical tweezers system measuring the change in light momentum flux. Rev. Sci. Instrum. **73**(6), 2308–2316 (2002)
3. Ashkin, A.: Optical trapping and manipulation of neutral particles using lasers. Proc. Natl. Acad. Sci. U. S. A. **94**(10), 4853–4860 (1997)
4. Wirtz, D.: Particle-tracking microrheology of living cells: principles and applications. Annu. Rev. Biophys. **38**(1), 301–326 (2009)
5. Yao, A., et al.: Microrheology with optical tweezers. Lab Chip **9**(17), 2568–2575 (2009)
6. Grier, D.G.: A revolution in optical manipulation. Nature **424**(6950), 810–816 (2003)
7. Lang, M.J., Block, S.M.: Resource letter: LBOT-1: laser-based optical tweezers. Am. J. Phys. **71**(3), 201–215 (2003)
8. Burnham, D.R., McGloin, D.: Holographic optical trapping of aerosol droplets. Opt. Express **14**(9), 4175–4181 (2006)
9. Molloy, J.E., Dholakia, K., Padgett, M.J.: Preface: optical tweezers in a new light. J. Mod. Opt. **50**(10), 1501–1507 (2003)
10. Kuyper, C.L., Chiu, D.T.: Optical trapping: a versatile technique for biomanipulation. Appl. Spectrosc. **56**(11), 300A–312A (2002)
11. Gutierrez-Medina, B., et al.: An optical apparatus for rotation and trapping. In: Walter, N.G. (ed.) Single Molecule Tools, Pt B. Methods in Enzymology, vol. 475, pp. 377–404. Elsevier/Academic, San Diego (2010)
12. Moffitt, J.R., et al.: Recent advances in optical tweezers. Annu. Rev. Biochem. **77**, 205–228 (2008)
13. Lee, W.M., et al.: Construction and calibration of an optical trap on a fluorescence optical microscope. Nat. Protoc. **2**(12), 1754–2189 (2007)
14. Abbondanzieri, E.A., et al.: Direct observation of base-pair stepping by RNA polymerase. Nature **438**(7067), 460–465 (2005)
15. Grange, W., et al.: VirE2: a unique ssDNA-compacting molecular machine. PLoS Biol. **6**(2), 343–351 (2008)
16. Bustamante, C., Bryant, Z., Smith, S.B.: Ten years of tension: single-molecule DNA mechanics. Nature **421**(6921), 423–427 (2003)
17. Bustamante, C., Chemla, Y.R., Moffitt, J.R.: High-resolution dual-trap optical tweezers with differential detection. In: Selvin, P.R., Ha, T. (eds.) Single-Molecule Techniques: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor (2008)
18. Greenleaf, W.J., Woodside, M.T., Block, S.M.: High-resolution, single-molecule measurements of biomolecular motion. Annu. Rev. Biophys. Biomol. Struct. **36**, 171–190 (2007)
19. Michaelis, J., et al.: DNA based molecular motors. Phys. Life Rev. **6**(4), 250–266 (2009)
20. Bustamante, B., Cheng, W., Mejia, Y.X.: Revisiting the central dogma. Cell **144**, 480–497 (2010)
21. Katranidis, A., et al.: Force measurements of the disruption of the nascent polypeptide chain from the ribosome by optical tweezers. FEBS Lett. **585**, 1859–1863 (2011)
22. Qu, X., et al.: The ribosome uses two active mechanisms to unwind messenger RNA during translation. Nature **475**, 118–121 (2011)
23. Block, S.M.: Kinesin motor mechanics: binding, stepping, tracking, gating, and limping. Biophys. J. **92**(9), 2986–2995 (2007)
24. Gross, P., et al.: Combining optical tweezers, single-molecule fluorescence microscopy, and microfluidics for studies of DNA-protein interactions. In: Walter, N.G. (ed.) Single Molecule Tools, Pt B. Methods in Enzymology, vol. 475, pp. 427–453. Elsevier/Academic, San Diego (2010)

25. Lang, M.J., et al.: Simultaneous, coincident optical trapping and single-molecule fluorescence. Nat. Methods **1**(2), 133–139 (2004)
26. Merz, A.J., So, M., Sheetz, M.P.: Pilus retraction powers bacterial twitching motility. Nature **407**(6800), 98–102 (2000)
27. Andersson, M., et al.: Physical properties of biopolymers assessed by optical tweezers: analysis of folding and refolding of bacterial pili. Chemphyschem **9**(2), 221–235 (2008)
28. Zhang, H., Liu, K.K.: Optical tweezers for single cells. J. R. Soc. Interface **5**(24), 671–690 (2008)
29. Tolic-Norrelykke, I.M., et al.: Nuclear and division-plane positioning revealed by optical micromanipulation. Curr. Biol. **15**(13), 1212–1216 (2005)
30. Sacconi, L., et al.: Optical micromanipulations inside yeast cells. Appl. Opt. **44**(11), 2001–2007 (2005)
31. Pascoal, P., et al.: Membrane nanotubes drawn by optical tweezers transmit electrical signals between mammalian cells over long distances. Lab Chip **10**(17), 2235–2241 (2010)
32. Grigaravicius, P., Greulich, K.O., Monajembashi, S.: Laser microbeams and optical tweezers in ageing research. Chemphyschem **10**(1), 79–85 (2009)
33. Guck, J., et al.: Stretching biological cells with light. J. Phys. Condens. Matter **14**(19), 4843–4856 (2002)
34. Rao, S., et al.: Raman study of mechanically induced oxygenation state transition of red blood cells using optical tweezers. Biophys. J. **96**(1), 209–216 (2009)
35. Chan, J.W., et al.: Reagentless identification of single bacterial spores in aqueous solution by confocal laser tweezers Raman spectroscopy. Anal. Chem. **76**(3), 599–603 (2004)
36. Molloy, J.E., Padgett, M.J.: Lights, action: optical tweezers. Contemp. Phys. **43**(4), 241–258 (2002)
37. Curtis, J.E., Koss, B.A., Grier, D.G.: Dynamic holographic optical tweezers. Opt. Commun. **207**(1–6), 169–175 (2002)
38. Liesener, J., et al.: Multi-functional optical tweezers using computer-generated holograms. Opt. Commun. **185**(1–3), 77–82 (2000)
39. Grieve, J.A., et al.: Hands-on with optical tweezers: a multitouch interface for holographic optical trapping. Opt. Express **17**(5), 3595–3602 (2009)
40. Jordan, P., et al.: Creating permanent 3D arrangements of isolated cells using holographic optical tweezers. Lab Chip **5**(11), 1224–1228 (2005)

Optimization of Nanoparticles

► Integrated Approach for the Rational Design of Nanoparticles

Optimization of Plasmonic Substrates for SERS Application

Marc Lamy de la Chapelle
Sorbonne Paris Cité, Laboratoire CSPBAT, CNRS, (UMR 7244), Université Paris 13, Bobigny, France

Introduction

Surface Enhanced Raman Scattering (SERS) is an analytical technique that provides a huge enhancement of the Raman signal. This technique is essentially based on the exploitation of the optical properties of metallic nanostructures. One of the main contributions of these properties is the excitation of the localized surface plasmon that can induce a huge enhancement of the electromagnetic near-field. By tuning the resonance of the localized surface plasmon, it is possible to optimize and control the SERS signal.

The Raman Spectroscopy

Raman spectroscopy is a vibrational spectroscopy that probes the vibrational modes of an excited molecule. These modes are directly related to the structure of the molecule and their frequencies depend on the mass of the molecule atoms, on the chemical bond nature and on the geometry of the molecules. Each vibration is specific of the molecular architecture as well as the molecule composition and can be used to identify and characterize the molecule, giving information on its chemical group and their spatial distribution. Thus, the Raman spectrum provides specific bands for each chemical or biological compound and is considered as an actual spectral signature for one individual compound.

The Raman spectroscopy is in fact based on an inelastic scattering process of a monochromatic light (a laser excitation for example) induced by an indirect transition between two different vibrational levels. Such transition can occur from the fundamental to the excited vibrational states or

from the excited to the fundamental vibrational states. In the first case, a decrease of the energy and of the frequency of the scattered light is observed (Stokes scattering process). In the other case, an increase of the energy and of the frequency is measured (anti-Stokes scattering process). In both cases, the measured spectral shift corresponds directly to the energy of the vibrational modes.

The intensity of these processes is directly related to the excitation electric field amplitude, E, and to the polarizability of the vibration mode, α, since the amplitude of the induced dipole moment of the vibrational mode, P, is defined as:

$$P = \alpha.E \quad (1)$$

The intensity of the Raman signal is then proportional to the square of the amplitude of P. However, the Raman cross section is very low (in the range of 10^{-30} to 10^{-26} cm^2) depending on the molecules and on its electronic structure. To reach high signal in Raman spectroscopy it is possible to either increase the number of molecules (increase of the concentration for example in solution) or increase the laser power. In the former case, it can be hard to get high amount of material especially in the case of biological molecules that are difficult to produce in large quantity. In the latter case, a high laser power can induce a degradation of the material by heating.

To avoid these two drawbacks, the increase of the Raman signal can also be reached by exploiting the local electromagnetic field enhancement obtained with metallic nanostructures. This method is called Surface Enhanced Raman Scattering (SERS).

The Surface Enhanced Raman Scattering (SERS)

The SERS is a very powerful method to increase the Raman signal of a specific molecule by several orders of magnitude, thanks to the use of metallic nanostructures of various shapes and sizes. Thus, any metallic surface that exhibits a roughness from the nanometric to mesoscopic scales can be considered as a SERS substrate and has the ability to enhance the Raman signal of any molecules in the close vicinity of the metallic surface. SERS can also be observed with metallic nanoparticles of different shapes (spheroids, ellipsoids, oblate, prolate, triangles, cubes…) and sizes (from few nanometers up to several hundreds of nanometers).

The enhancement of the Raman signal observed in SERS can be assigned to two different processes in the SERS: an electromagnetic process and the chemical process.

The chemical one is not completely understood but it is assigned to electronic interaction between the metallic surface and the electron cloud of the molecule. It can be modeled by a charge transfer or by the formation of a chemical bond between the surface and the molecule. In both cases, such interaction induces a modification of the electron distribution around the molecule and thus in favorable case, an increase of the molecule polarizability. This polarizability modification has an influence on the amplitude of the induced dipole moment, P, and thus induces an increase of the Raman signal. However, this enhancement is limited since it has been estimated to be around two orders of magnitude.

In fact the main effect that largely explains the SERS process is the electromagnetic one which contributes to the Raman signal enhancement up to eight orders of magnitude. The electromagnetic process is based on the enhancement of the electric field amplitude at the vicinity of the nanoparticle surface. This field enhancement is directly related to the optical properties of the nanostructure and more especially to the ability to excite the localized surface plasmon (LSP). The LSP is a collective oscillation of the free electron cloud inside the nanostructure. Hence, the LSP can be seen as an oscillation mode and thus its frequency depends strongly on the size, the shape and the metal nature of the nanostructure, and eventually on its coupling with other LSP from other closely packed nanostructures. As any oscillating system with a proper mode, the LSP is defined by a resonance frequency corresponding to the maximum of interaction with an excitation electromagnetic field. The LSP resonance (LSPR) position is

often determined as the maximum of the extinction (absorption and scattering) spectrum measured with the nanoparticles. The LSPR can be tuned in a wide spectral range just by changing the geometrical parameters or the metallic nature of the nanostructures. When excited with a monochromatic laser light close the LSPR, a huge enhancement of the electromagnetic field is created at the vicinity of the surface. This enhancement notably depends on the position of the excitation wavelength compared to the position of the resonance. In enhanced spectroscopies as SERS, it is actually important to optimize the position of the LSPR to reach the highest enhancement factor.

The enhancement factor can be determined taking into account the different light interaction processes between the excitation light, the nanostructure and the molecule. In fact, the nanostructure with the molecule at its vicinity has to be considered as a two bodies system and light interacts with both of them [1, 2]. First, the excitation light (excitation wavelength λ_0) interacts with the LSP and creates a first electric field enhancement. This enhanced electric field, E_{loc}, excites the molecule that scatters an enhanced Raman signal at a redshifted wavelength, λ_R, (Stokes process, decrease of the light frequency and increase of the light wavelength). The enhanced Raman signal can then interact with the nanostructure and the LSP. This process, called the reradiation process, induces a second enhancement of the Raman signal at the Raman wavelength, λ_R. The enhanced Raman dipole moment, P_{SERS}, and the enhanced intensity of the Raman signal, I_{SERS}, can be expressed as follows:

$$P_{SERS} = \alpha . f(\lambda_0) . f(\lambda_R) . E \tag{2}$$

$$\begin{aligned} I_{SERS} &= \alpha^2 . |f(\lambda_0) . f(\lambda_R)|^2 . E^2 \\ &= |f(\lambda_0) . f(\lambda_R)|^2 . I_{\mathrm{Raman}} \approx |f|^4 . I_{\mathrm{Raman}} \end{aligned} \tag{3}$$

with $f(\lambda_0)$, the enhancement factor at the excitation wavelength, λ_0, and $f(\lambda_R)$, the enhancement factor at the Raman wavelength, λ_R. If we assume that the excitation and Raman wavelengths are close (low Raman shift), the two enhancement factors will also be close ($f(\lambda_0) \approx f(\lambda_R) \approx f$) and the SERS signal will be enhanced at the power fourth of the field enhancement factor, f (as shown in Eq. 3). This explains why one can reach giant enhancement since even a low enhancement of the amplitude of the electric near-field can induce a large enhancement of the Raman signal. Both enhancement factors depend on the position of the LSPR comparing to the excitation and the Raman wavelengths.

Thus, the Raman signal of any molecules deposited at the nanostructure surface could be enhanced by several orders of magnitude depending on the LSPR and more generally on the optical properties of the nanostructures. Moreover, this enhancement occurs in the near field (approximately in the distance from the surface lower than the half of the wavelength) of the nanostructures. Indeed, the field created by the LSP is confined at the surface vicinity due to the evanescent nature of the LSP. Thus the amplitude of the enhanced field decreases exponentially from the nanostructure surface with a very short decay length close to a few tens of nanometer in the most favorable cases. Thus, only the first layers of molecules are in fact inside the enhanced field and only they contribute to the SERS signal. SERS is an extremely local spectroscopy to probe a few quantities of molecules and has enabled the single molecule sensitivity.

Optimization of the SERS Signal

To be able to control the enhancement factor obtained in SERS, it is of first importance to control the optical properties of the metallic nanostructures used as SERS substrate. Such approach is effectively necessary in order to reach reproducible SERS signal as well as to reach the best SERS signal. In this section, the discussion will be focused on the different parameters that have an influence on the SERS signal and how they have to be adjusted in order to optimize it. The main rules of optimization will be presented and clarified.

Control the Optical Properties to Tune the SERS Electromagnetic Contribution

First, as explained above, one has to control precisely the optical properties of the nanostructure and more specifically the position and the width of the plasmon resonance. In fact, the LSPR can be finely tuned by controlling the size and the shape of the nanostructures at the nanometer scale as well as their close environment including potential coupling. Thus, to reach such control of the optical properties and to follow precise requirements, the use of lithography methods as production techniques is required. The electron beam lithography (EBL) is the method that offers the highest flexibility since it allows the production of any kind of nanostructure shapes, sizes, and arrays with a precision below 10 nm. The principle of the EBL is the creation of a polymeric mask with the electron beam directly on the substrate. The mask precision is directly related to the beam size. As a consequence, this technique is a powerful tool to provide highly reproducible SERS substrate. However, it has several drawbacks since it is slow and the produced surface is small (width of only few hundreds of micrometer). This technique is expensive and is not suitable for the mass production of SERS substrates. To overcome such limitations, two other methods have been proposed: the nanosphere lithography (NSL) and the nanoimprint lithography (NIL). The NSL is based on the deposition of a monolayer of latex beads at the substrate surface that forms a mask to produce nanotriangles in hexagonal array. The principle of the NIL is the imprint of a nanostructured stamp that reproduces the nanostructures on a polymer to form the mask. Both techniques are cheaper and faster than EBL but they are limited in terms of shape for NSL (only nanotriangles can be produced) and of size for NIL (nanostructures should have size higher than 100 nm to be produced correctly) [2, 3].

SERS Versus LSPR

The main effect that is involved in the enhancement of the electromagnetic field around the nanostructures is the LSP. The close relation between the LSP and the SERS signal can be demonstrated by polarization experiments. Indeed, on elongated nanoparticles, several LSP modes can be observed along each axis of the nanoparticle. For example, with a nanorod with an aspect ratio around two (length around 100 nm and width around 50 nm), one can observe a LSPR along the length close to 700 nm and a LSPR along the width close to 520 nm. Thus, depending on the excitation polarization, it is possible to excite selectively each LSP. Since there is a large spectral shift (around 200 nm) between both LSPR, one can chose an excitation wavelength (633 nm for example) that is close to the LSPR (700 nm along the nanorod length) and out of resonance for the other one (520 nm along the width). When the excitation polarization is then rotated comparatively to the length of the nanorod, the efficiency of the LSP depends on the angle, θ, between the excitation polarization and the length of the nanorod ($\theta = 0°$, full excitation of the LSP; $\theta = 90°$, no excitation of the LSP). When the SERS signal is measured depending on θ, a clear correlation is observed between the LSPR and the SERS intensities [4]. It means that both intensities are maximum for the same angle ($\theta = 0°$ and 180°) whereas the SERS intensity is null when the LSPR is not excited ($\theta = 90°$ and 270°). Moreover, whereas the LSPR intensity follows a Malus law with an angular dependence in $(\cos(\theta))^2$, one can notice that the dependence of the SERS signal with the angle is closer to a $(\cos(\theta))^4$ law, which is a clear evidence of the dependence of the SERS signal is close to the 4th power of the near-field enhancement factor (Eq. 3).

As the LSP is assigned to the oscillation of the electron cloud, it is a resonant effect that induces a high interaction with the light for one specific wavelength (LSPR position) and on a specific wavelength range (width of the LSPR). Out of resonance, it is assumed that there is few or no interaction with light. It means that the field enhancement is directly related to the position of the excitation wavelength compared to position of the LSPR and as a consequence the SERS signal strongly depends on these relative positions. Moreover, as the enhancement factor (EF) in SERS includes one EF at the excitation wavelength as well as one EF at the Raman wavelength

due to the reradiation process, the SERS signal also depends on the position of the Raman wavelength compared to the position of the LSPR. Intuitively, one can easily assume that the best SERS signal is reached for a position of the LSPR between the excitation and the Raman wavelength in order to optimize both EF. Indeed, if the excitation (Raman) wavelength is close to the LSPR, the Raman (excitation) wavelength will be out of resonance that will induce a low global EF [2].

The position of the LSPR compared to the excitation and the Raman wavelengths is then the first parameter to be optimized to reach the highest SERS signal.

Several experiments have been done to determine the best LSPR position and it has been demonstrated that the above rule is respected for gold nanocylinders or nanotriangles produced by EBL or NSL respectively using an excitation wavelength between 514 and 633 nm [2, 4-6]. These experiments show a strong variation of the SERS signal depending on the position of the LSPR. A maximum is observed at the average between the excitation and the Raman wavelength whereas the SERS signal decreases rapidly outside this range. It is then possible to tune the SERS signal by changing the position of the LSPR.

But this rule is no more applicable when the excitation wavelength is close to the near infrared. Indeed, for an excitation wavelength of 785 nm and for nanocylinders produced by EBL, the maximum of the SERS signal is reached for a LSPR close to 750 nm and thus largely outside the excitation and Raman wavelengths range [7]. Such effect has been assigned to a discrepancy between the near-field around the nanostructure and the far field. The SERS process is effectively a near-field effect since the enhanced Raman signal is produced by the enhanced field at the close vicinity of the nanostructure surface whereas the LSPR is measured by extinction spectroscopy in the far-field. It means that the highest field enhancement could not be reached at the wavelength corresponding to the LSPR. Such discrepancy has notably been explained by a spectral shift between the near-field enhancement and the LSPR [8, 9] due to the oscillation damping inside the nanostructures. This spectral shift should then increase with the LSPR width. It is important to take into account not only the position of the LSPR but also the excitation wavelength in order to optimize the SERS signal.

Another parameter that influences the SERS optimization is the shape of the nanostructure. Some experiments have been done on gold nanorods and nanowires produced by EBL at several excitation wavelengths (633 and 676 nm) [2, 10]. Such nanorods have some aspects ratio from 1 to 20 (length from 60 nm up to 1 μm, width 60 nm). For an excitation electric field parallel to the long axis of the nanorods, the SERS signal exhibits large variations depending on the length of the nanorods with several maxima for specific lengths (around 100, 350, 650 and 900 nm at 633 nm and 140, 400 and 700 nm at 676 nm). It has also been observed that for some specific lengths the SERS signal is null (length around 300 nm). The first maximum can be assigned to the first order of LSPR whereas the other ones are assigned to higher orders (3rd order for the second maximum, 5th order for the third one and 7th order for the last one). Indeed, in long nanorods, some higher orders of resonance can be excited since the length of the nanorods is no more negligible compared to the excitation wavelength. To respect the symmetry of the LSP mode along the nanorods, only odd modes can be excited. First, this observation demonstrates that the higher order of resonance can be SERS active in nanorods. This is not the case for nanocylinders since it has been shown that for nanocylinders larger than 200 nm exhibiting 3rd order of LSPR, no specific enhancement is observed. Second, the higher orders of LSPR can exhibit higher SERS signal than the dipolar one. This is notably observed at 633 nm for which the highest SERS signal is obtained for the 5th and 7th orders. Third, for any LSPR orders and any excitation wavelengths, the SERS signal is optimized for a LSPR position close to the Raman wavelength. The SERS behavior of the nanorods is then different from the one of the nanocylinders: different rule of optimization and no dependence on the

excitation wavelength. This specific behavior has been assigned to the elongated shape of the nanorods [4]. The use of electromagnetic model has shown that such shape favors the reradiation effect in SERS and thus the enhancement of the signal at the Raman wavelength. From such results, it is possible to determine that the nanorods acts as nanoantenna at the nanometric scale through the enhanced re-emission of the Raman scattering.

All these results demonstrate that the optimization of the SERS signal depends on the LSPR position and that some optimization rules can be extracted. However, these rules are strongly dependent on the shape and the excitation wavelength essentially due to the near-field interaction between the molecule and the nanostructure. In the aim of maximizing the SERS signal, it is of real importance to take into account such near-field effect and the geometry of the nanostructure is a key parameter for the SERS optimization.

Effect of the Environment

Another way to increase the SERS signal is to control the direct environment of the nanostructure.

As any resonator, the LSP can be damped by the local media that surrounds the nanostructure. The effect of this damping is directly related to the dielectric constant of the surrounding media and more specifically to its imaginary part that induces absorption of the excitation energy or of the LSP energy. As an example, one can consider the adhesion layer used to attach the gold nanostructures at the glass substrate surface. In most of the case, this adhesion layer is made of chromium, which has a large imaginary part of the dielectric constant in the visible range that is known to degrade the LSPR. It has been proposed to change this layer by a dielectric layer as the MPTMS that has no absorption in the visible range [11]. In this latter case, the LSPR has been improved since the use of the MPTMS layer induces an increase of the LSPR intensity and a decrease of the LSPR width and as a consequence an increase of the quality factor of the resonance. Such improvement of the LSPR has also a strong influence on the near-field enhancement. First, it has been demonstrated that the SERS signal is increased by one order of magnitude when the MPTMS is used compared to the Cr adhesion layer. A sharper LSPR induces a higher SERS signal. Second, a narrower LSPR width induces also a narrower SERS signal variation when the LSPR is tuned comparatively to the excitation and the Raman wavelengths. A higher quality factor of the LSPR induces a higher quality factor of the SERS efficiency.

One of the main effect used to increase the SERS signal is to approach several nanostructures together at distance lower than the half of the wavelength. The LSPs of these close nanostructures couple together to create a highly intense near-field in the gap between the nanostructures. The field enhancement is then largely higher than the one obtained with individual nanostructures due to the coupling. Since the amplitude of the LSP decreases exponentially with the distance from the nanostructure surface and since the decay length of the LSP amplitude is between 10 and 20 nm, the gap to get a coupling has to be lower than few tens of nanometers (typically lower than 20 nm). Moreover the field enhancement depends strongly on the gap and it increases exponentially when the gap decreases. Such geometry creates what is called in SERS "hot spots," thus providing huge enhancement factors. As a consequence, the SERS signal follows the same behavior than the one observed for the LSPR coupling, meaning highly intense Raman signal that increases exponentially when the gap decreases [12]. The ideal configuration is then to have the lowest gap between the nanostructures to reach the highest enhancement factor. However, keeping in mind that it is necessary to control the optical properties and the geometrical parameters of the nanostructures, it is difficult with the current lithography techniques to produce some coupled nanostructures with a gap lower than 10 nm with high reproducibility on a large surface. For controllable SERS substrate, such kind of signal optimization by the coupling approach is still not completely achievable.

Conclusion

The objective of this article was to give a large overview of the different parameters that influence the SERS signal and how the latter one can be optimized. First, we emphasized the relation between the LSPR position and the SERS intensity, demonstrating that some optimization rules can be extracted. Such rules depend essentially on the shape of the nanostructures as well as on the excitation wavelengths. It was also demonstrated that the higher orders of LSPR are SERS active and can provide higher SERS signal than the dipolar LSPR. Moreover, the relation between the LSPR position (measured in far field) and the SERS intensity (near-field interaction) is not totally obvious and has to be treated carefully. Second, it is clear that the direct environment of the nanostructures, either the surrounding media or the coupling between several nanostructures, has also a strong influence on the SERS signal. In all the cases, in order to control the optical properties of the nanostructures and as a consequence the SERS signal, it is of real importance to control precisely the geometrical parameters of the nanostructures and then to use lithography techniques that allow the production of nanostructures with high reproducibility. Thus, the control of the nanostructure geometry and architecture is a promising approach to produce some highly reproducible SERS substrates, which is the ideal achievement for any SERS applications.

Cross-References

- ▶ E-Beam Lithography (EBL)
- ▶ Local Surface Plasmon Resonance (LSPR)
- ▶ LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design
- ▶ Nano-optics, Nanoplasmonics
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Plasmonics
- ▶ Surface Plasmon Nanophotonics
- ▶ Surface-Enhanced Raman Scattering for Imaging Biological Cells

References

1. Wokaun, A.W.: Surface enhanced electromagnetic processes. Solid State Phys. **38**, 223–293 (1984)
2. Guillot, N., Lamy de la Chapelle, M.: The electromagnetic effect in surface enhanced Raman scattering: enhancement optimization using precisely controlled nanostructures. J. Quant. Spectrosc. Radiat. Trans. **113**, 2321–2333 (2012)
3. Guillot, N., Lamy de la Chapelle, M.: Lithographied nanostructures as nanosensors. J. Nanophotonics **6**(1), 64506/1–64506/28 (2012)
4. Grand, J., Lamy de la Chapelle, M., Bijeon, J.L., Adam, P.-M., Vial, A., Royer, P.: Role of localized Surface plasmons in surface enhanced Raman scattering of shape- controlled metallic particles in regular arrays. Phys. Rev. B Condens. Matter Mater. Phys. **72**, 033407/1–033407/4 (2005)
5. Haynes, C.L., Van Duyne, R.P.: Plasmon-sampled surface-enhanced Raman excitation spectroscopy. J. Phys. Chem. B **107**, 7426–7433 (2003)
6. Félidj, N., Aubard, J., Lévi, G., Krenn, J.R., Hohenau, A., Schider, G., Leitner, A., Aussenegg, F. R.: Optimized surface-enhanced Raman scattering on gold nanoparticle arrays. Appl. Phys. Lett. **82**, 3095–3097 (2003)
7. Guillot, N., Shen, H., Fremaux, B., Péron, O., Rinnert, E., Toury, T., Lamy de la Chapelle, M.: Surface enhanced Raman scattering optimization of gold nanocylinder arrays: influence of the localized surface plasmon resonance and excitation wavelength. Appl. Phys. Lett. **97**, 023113/1–023113/3 (2010)
8. Grimault, A.S., Vial, A., Lamy de la Chapelle, M.: Modeling of regular gold nanostructures arrays for SERS applications using a 3D FDTD method. Appl. Phys. B-Lasers Opt. **84**(1–2), 111–115 (2006)
9. Zuloaga, J., Nordlander, P.: On the energy shift between near-field and far-field peak intensities in localized plasmon systems. Nanoletters **11**(3), 1280–1283 (2011)
10. Billot, L., Lamy de la Chapelle, M., Grimault, A.S., Vial, A., Barchiesi, D., Bijeon, J.L., Adam, P.M., Royer, P.: Surface enhanced Raman scattering on gold nanowire arrays: evidence of strong multipolar surface plasmon resonance enhancement. Chem. Phys. Lett. **422**, 303–307 (2006)
11. Shen, H., Guillot, N., Rouxel, J., Lamy de la Chapelle, M., Toury, T.: Optimized plasmonic nanostructures for improved sensing activities. Opt. Express **20**(19), 21278–21290 (2012)
12. Kessentini, S., Barchiesi, D., D'Andrea, C., Toma, A., Guillot, N., Di Fabrizio, E., Fazio, B., Marago, O., Gucciardi, P., Lamy de la Chapelle, M.: Gold dimer nanoantenna with slanted gap for tunable LSPR and improved SERS. J. Phys. Chem. C **118**(6), 3209–3219 (2014)

Optoelectrically Enabled Multiscale Manipulation

Han-Sheng Chuang[1,2], Aloke Kumar[3], Stuart Williams[4] and Steven T. Wereley[5]
[1]Department of Biomedical Engineering, National Cheng Kung University, Taiwan
[2]Medical Device Innovation Center, National Cheng Kung University, Taiwan
[3]Department of Mechanical Engineering, University of Alberta, AB, Canada
[4]Department of Mechanical Engineering, University of Louisville, Louisville, KY, USA
[5]School of Mechanical Engineering, Purdue University, Birck Nanotechnology Center, West Lafayette, IN, USA

Synonyms

Hybrid optoelectric manipulation; Optoelectrofluidics

Definition

Multi-scale manipulation refers to manipulation of objects with operational length scales ranging from several millimeters to nanometers. An optoelectric platform, which integrates multiple physical mechanisms, uses optically triggered electrokinetics to achieve multi-scale manipulation. These platforms are characterized by dynamic programmability and multitasking.

Overview

The growing demand for point-of-care analysis has motivated the rapid growth of the lab-on-a-chip (LoC) market. Considering the complexity, diversity, and multitasking nature of the technology, considerable efforts have been made to create a versatile platform capable of handling as many targets as possible. Multi-scale manipulation, or the ability to manipulate media and particles at multiple length scales, has been a significant and sought-out milestone in the LoC field for years. Despite a significant progress over the past decade in micro-/nanofluidics, overcoming the operational gaps of macroscale to microscale and microscale to nanoscale regimes remains a big challenge in the scientific community. Existing prior techniques, such as micropumps, microvalves, micropipettes, microgrippers, and so forth, heavily rely on continuous fluid flow and mechanical structures, thus increasing the difficulties in connections between different length domains. As a result, the multi-scale manipulation was not fully realized until digital microfluidics became practical. Recently, noninvasive manipulation techniques are rapidly emerging in many biomedical applications. Noninvasive manipulation typically utilizes optical, electrical, magnetic, or thermal measures to maneuver targets; therefore, no physical contact is involved. This free-of-contact capability is beneficial to biomedical applications due to the vulnerability and susceptibility of most biological materials to cross-contamination and mechanical stress. The potential of noninvasive techniques in multidimensionality and extensibility makes them the best candidates for carrying out multi-scale manipulation.

Two categories of length scale can be used to distinguish current noninvasive manipulation schemes: millimeters to sub-millimeters and micrometers to nanometers. For the first category, aqueous droplets are targeted due to ease of control, least cross-contamination, free from physical microfluidic components, and definitive geometry. Droplets act as carrying vessels for mediating various movements, such as mixing, splitting, translation, metering, and incubation. Without conventional structural connections, the gap between different length scales is no longer a hassle. Optoelectrowetting (OEW) [1], a technique based on electrowetting and photovoltaics, and its derivative, open optoelectrowetting (O-OEW) [2], are optoelectric techniques used in this regime. Both techniques enable dynamic

manipulation and simplify microfabrication. Unlike the first category, the second one usually deals with more diverse subjects, which include micro-/nanoparticles, quantum dots (QDs), cells, bacteria, nanowires, and nanorods. In this regime, the targets will go through a final processing step to achieve a designated task. Common manipulation scenarios may include separation, concentration, and sorting. Optoelectronic tweezers (OETs) [3] and rapid electrokinetic patterning (REP) [4] have shown more promise when compared to traditional techniques, such as dielectrophoresis (DEP) and electrophoresis (EP) in terms of programmability and maneuverability. The subtle combination of electric and optical domains broadens their applications. Since these techniques are compatible with either electric field, optical field, or both, an integration with other optoelectronic devices for a more powerful tool is viable.

Some attempts have been made to successfully achieve noninvasive multi-scale manipulation. Cho et al. [5] first demonstrated concentration and binary separation of latex particles in liquid vessels based on electrowetting on dielectric (EWOD) and electrophoresis. Later, Fan et al. [6] achieved live and dead cell sorting in droplets based on a platform composed of DEP and electrowetting. The length scales involved in both studies were several micrometers for particles/cells and millimeters for droplets. However, the operational flexibility remained very limited due to the permanent nature of the electrodes. After OEW, OETs, and REP were introduced, attention then shifted toward optoelectric tools to seek for more flexible manipulation techniques. Shah et al. [7] came up with a EWOD-driven device integrated with OETs for cellular isolation and analysis. Although the platform was not fully operated with the optoelectric means, OET performed its programmable potential in such kind of integrated system. Later, Kumar et al. [8] exhibited a fully optoelectric platform driven by O-OEW and REP. All electrode patterns were thus virtually and dynamically generated by incident light images. O-OEW, a modified OEW, was utilized for basic droplet handling. The hybrid system was built with added flexibility by replacing the stereotypical sandwiched structure with an open configuration. Similarly, Valley et al. [9] achieved similar results with OETs, although the underlying physics differ from REP mechanisms.

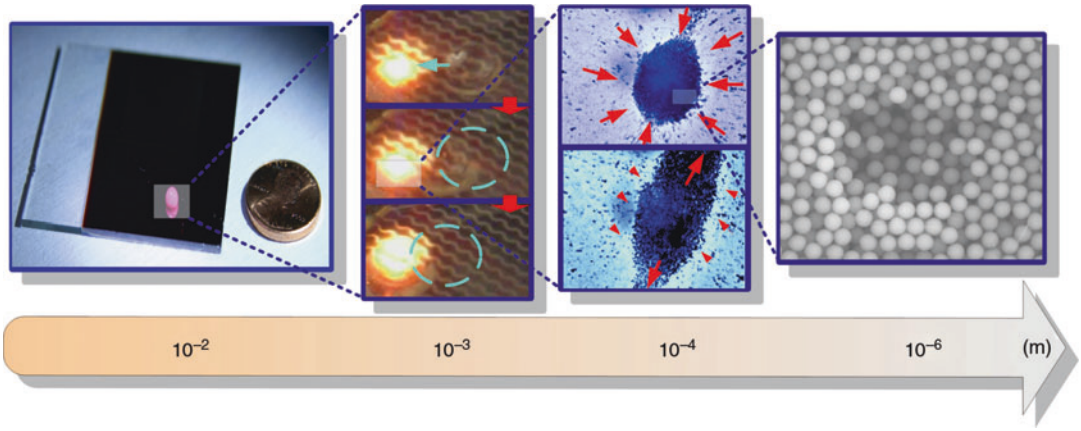
Figure 1 shows a chart of the hybrid optoelectrically enabled multi-scale manipulation and the corresponding length scales involved in different operational regimes. Due to the inherent compatibility between optoelectric techniques, the transitional gaps from one length domain to another become nearly seamless. Such hybrid techniques can engender new applications in lab-on-a-chip systems and may enable novel approaches to the investigation and application of fundamental electrokinetic phenomena.

Methodology

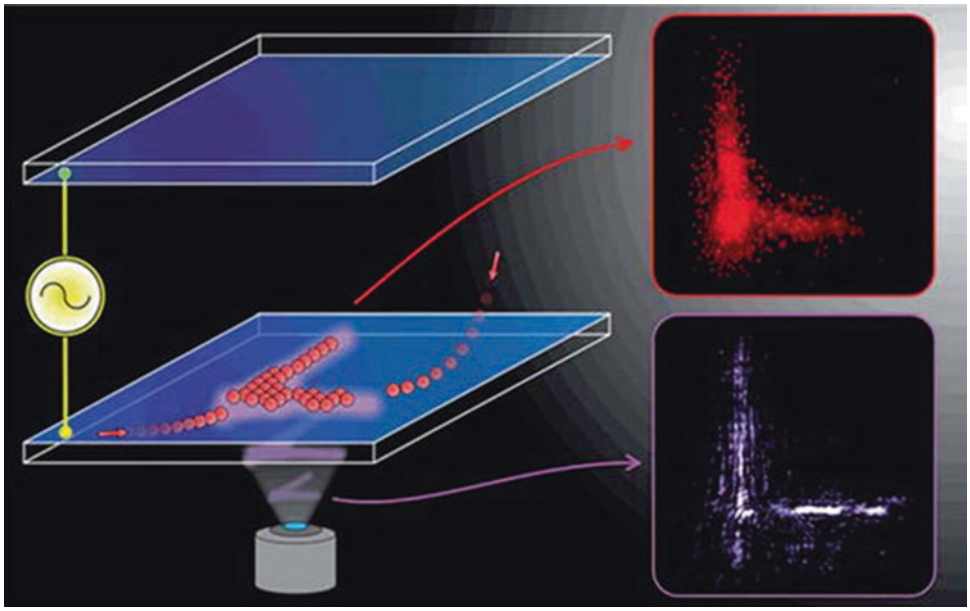
In this section the mechanisms that have been employed as components of different multi-scale tools will be discussed. First REP and OETs – tools which target the micrometer to nanometer length scales – will be discussed. OEW will be discussed last that targets liquid droplets, objects typically within the millimeter length scale.

Rapid Electrokinetic Patterning (REP)

REP is an innovative new micro-/nanomanipulation tool that has been developed at Purdue University, USA. This technique is a noninvasive, real-time, dynamic, optoelectric manipulation tool that was first introduced in 2008 by Williams et al. [4]. The first reported chip configuration consisted of ITO-coated parallel electrodes biased with an AC signal (Fig. 2). It was demonstrated that in a low AC frequency regime (<200 kHz), micrometer-sized particles could be trapped and dynamically configured into various shapes on the electrode surface by using various optical landscapes produced from a laser operating at a wavelength of 1064 nm. Since this report, the technique has been extended to the use of nonuniform electric fields [10] and to particles as diverse as gold particles (200, 250 nm),



Optoelectrically Enabled Multiscale Manipulation, Fig. 1 Multi-scale capabilities of a hybrid optoelectric device. An aqueous droplet is manipulated by O-OEW. The illumination is provided through a microscope objective lens with focused broadband light. The applied AC biasing conditions were 50 V_{rms} at 300 Hz. Inside the droplet, 1 μm polystyrene particles are massively aggregated around an IR-illuminated site at the water-electrode interface. The applied AC biasing conditions were 28 V_{rms} at 25 kHz. The particle clump disperses when the IR illumination is off. Zooming the particle aggregation shows thousands of loosely packed particles stacked into several layers (Reprinted with permission from Kumar et al. [8]. Copyright 2010 American Chemical Society)



Optoelectrically Enabled Multiscale Manipulation, Fig. 2 In REP, laser illumination causes sharp temperature gradients in the medium through liquid and electrode heating, thus increasing mass transfer to the illuminated locations on the electrode surface. On the electrode surface, electrokinetic forces lead to the capture of the particles. In the figure, a laser hologram captures thousands of fluorescent (700 nm) beads into an "L" shape (Williams et al. [4]. Reproduced with permission from The Royal Society of Chemistry)

silica particles, and polystyrene particles (50 nm to 3 μm) [10, 11]. Lately, biological applications, such as bacterial manipulation [12] and fluorescent signal enhancement for bead-based bioassays [13], have also been demonstrated with the technique. The attempts provide an insight to implementing microsurgeries on microorganisms in the near future.

REP, in its classical rendition, relies on the creation of extremely localized temperature gradients to drive electrothermal microvortices in the chip. The temperature increase due to laser heating does not necessarily need to be significant (<5 K); however, tightly focused laser beams lead to an extremely high gradient of temperature ($>10^6$ K/m). Electrothermally driven microvortices increase mass transport to the laser-illuminated spot on an electrode, where low-frequency electrokinetics results in the capture of particles. A full discussion of the underlying mechanisms and chip configurations is not in the scope of current discussion, but interested readers can refer to a review article by Kumar et al. [11].

Optoelectronic Tweezers (OETs)

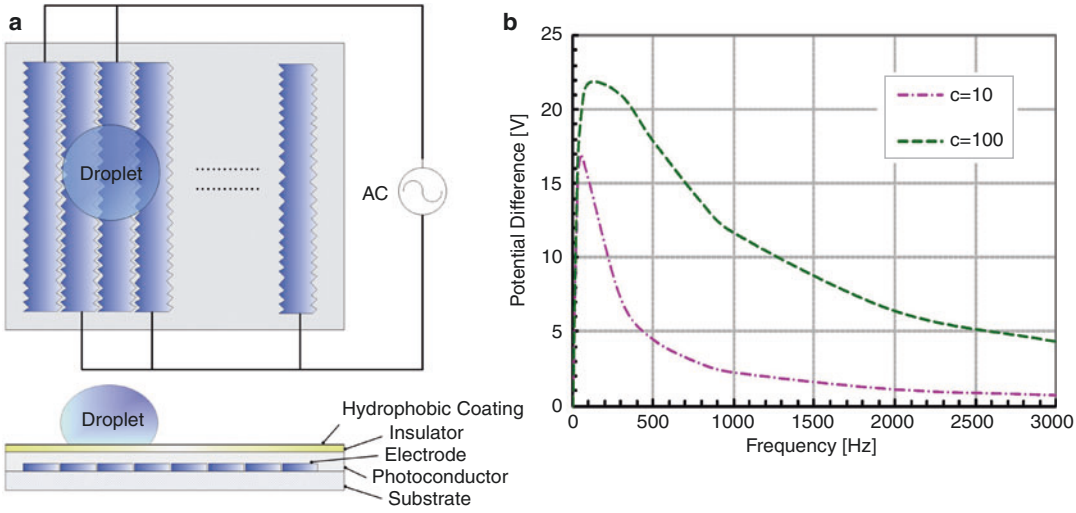
Distinct from REP, OET is an electrokinetic technique, which uses a strongly photoconductive substrate. In OETs, an illumination source locally increases the material's conductivity, creating reconfigurable "virtual" electrodes. Chiou et al. [14] first reported this device which consisted of two parallel ITO electrodes with a photoconductive material, hydrogenated amorphous silicon (a-Si:H), coating one of them. When illuminated a-Si:H increases its electrical conductivity several orders of magnitude, creating a localized increase in field strength, enabling dynamic virtual electrode configurations. Photoconductivity of a-Si:H is most efficient from red to near-infrared wavelengths, though a white light source can also be used.

The reconfigurable illumination sources vary from a digital micromirror device (DMD), liquid crystal display (LCD), or a beam projector. Like REP, OET devices can also use a number of different electrokinetic mechanisms for its operation. They can use DEP, ACEO, or electrothermal flows for particle aggregation. Recent reviewed articles [11, 15] discuss in detail the underlying principles and operational regimes of OETs. OETs have trapped, sorted, and manipulated a variety of nonbiological and biological particles ranging from quantum dots to erythrocytes and leukocytes [16].

Optoelectrowetting (OEW)

Electrowetting (EW) alters the surface energy of the substrate under an aqueous droplet by modulating an external electric field. Charges induced by the electric potential are accumulated on both sides of the insulator. The droplet hence changes its geometric profile in response to the new energy state. Based on the core mechanism of EW, OEW [1] and O-OEW [2] delicately harvest the optical energy and convert it to the electrical energy through a photoconductor. In O-OEW, a photoconductor is arranged between the insulating and conductive layers serving as a binary gate in the presence of AC electric fields (Fig. 3a).

Light can lower the resistance in a photoconductor by activating a photovoltaic effect. It follows that the gate can therefore be opened by light but closed when the light is removed. The physics behind the on/off switching phenomenon is simply attributed to impedance shifting. In the absence of light, the photoconductor has high impedance. The majority of the potential drop hence is within the photoconductive layer in the dark. In the presence of light, however, the high impedance shifts to the insulator due to lowered resistance in the photoconductor. The potential drop across the insulator hence results in the conventional electrowetting effect. The overall switching phenomenon can be modeled with an equivalent circuit. The bright-to-dark potential difference across the insulator escalates as the frequency increases and then gradually declines when the frequency exceeds several hundred Hertz. The empirically operational range for an O-OEW device is estimated between 100 and 800 Hz [2] (Fig. 3b), but may vary with device geometry.



Optoelectrically Enabled Multiscale Manipulation, Fig. 3 (**a**) Schematic diagram of the chip layout. The interdigitated electrodes are connected to ground and electric potential alternately. The jigsaw edge is used to decrease the discontinuity between the gaps. The materials from the *bottom* to the *top* are glass substrate, titanium (Ti) (electrodes), *a*-Si (photoconductor), SiO_2 (insulator), and Teflon (hydrophobic coating). (**b**) Potential difference versus driving frequency. The potential difference is the difference between the dark and illuminated potentials in the insulator. Notation *c* denotes the photoconductivity ratio (light-to-dark conductivity ratio). The driving potential was 50 V_{rms}. The empirically operational frequency range was between 100 and 800 Hz. The frequency out of the range could result in limited or none O-OEW effect

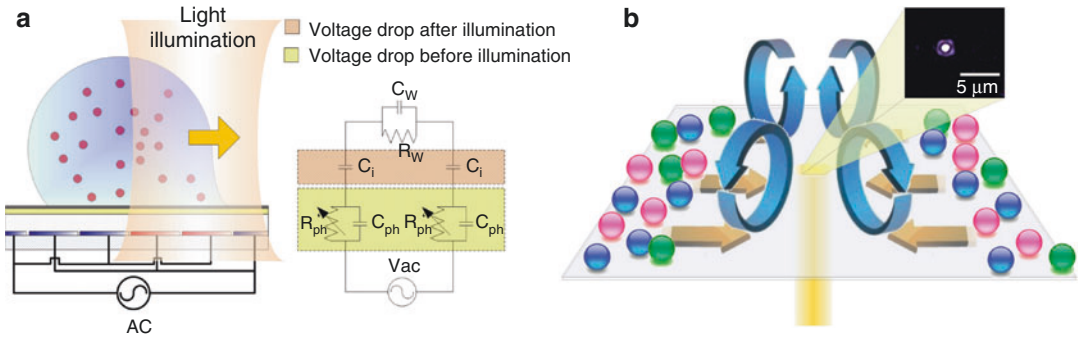
Since the gate can be locally opened by a focused light spot, electrodes are virtually created. This feature enables programmability as well as dynamic manipulation without pixilated electrodes. Unlike EW, virtual electrodes not only simplify microfabrication but also impart O-OEW more maneuverability and flexibility. As compared to conventional OEW's sandwiched configuration, O-OEW had added freedom of sample introduction with coplanar arrangement of interdigitated electrodes. Detailed theories as well as applications of OEW are available [1, 2, 17].

Key Research Findings

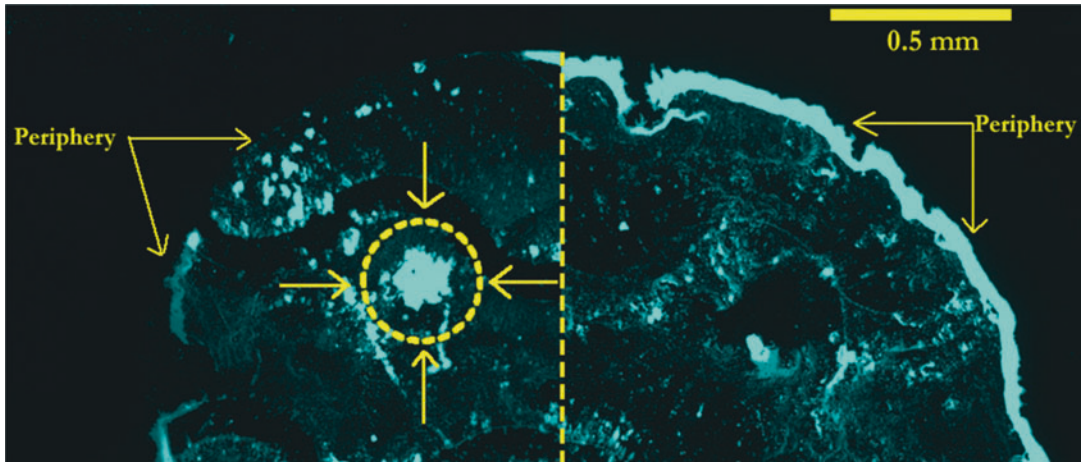
The first all optoelectric multi-scale manipulation was pioneered by Kumar et al. [8], where REP and O-OEW were combined to obtain multi-scale manipulation. This report was followed by Valley et al.'s work [9], where OETs and OEW were used simultaneously. In 2009, Shah et al. [7] had also demonstrated multi-scale manipulation in a device which integrated OETs and EW. Since only one of the techniques involved in Shah et al.'s device is optoelectric, this work will be discussed last.

REP+O-OEW

Kumar et al. [8] first demonstrated an all optoelectric device for multi-scale manipulation (Fig. 4). The multi-scale manipulation requirement was that a microliter aqueous droplet had to be moved to an appropriate work zone and subsequently concentration of suspended 1.0 μm fluorescent polystyrene beads close to the substrate surface had to be achieved. Thus, two distinct length scales, micron length scale for suspended particulates and millimeter length scale for droplet, were involved. To achieve this end, the researchers created an O-OEW device using ITO electrodes. Steering and mobilization of the liquid droplet was achieved using a red laser pointer in conjunction with the O-OEW device (Fig. 4a). Once, the liquid droplet was brought to the appropriate work zone, a focused infrared laser was used to create the REP effect, thus



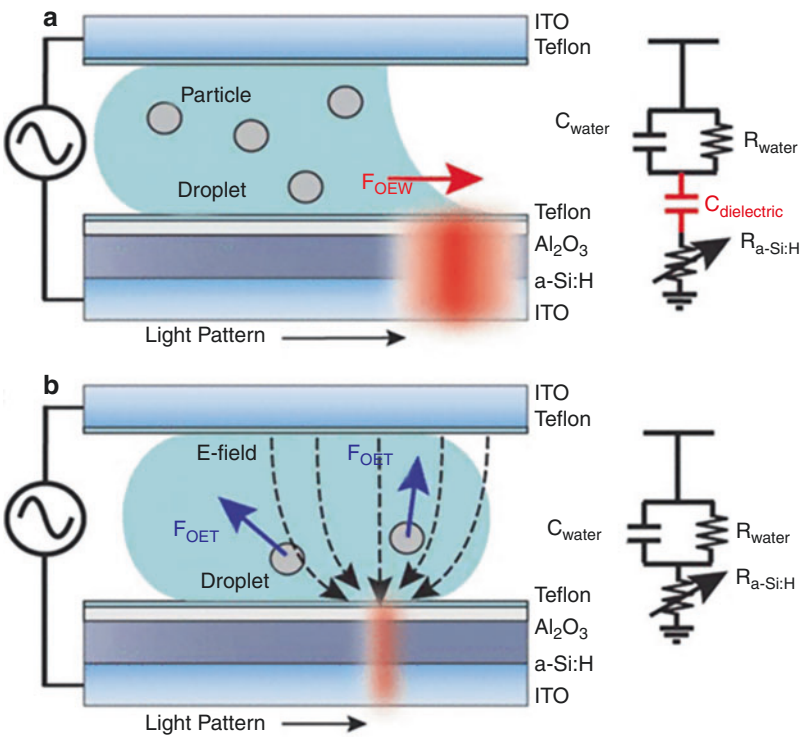
Optoelectrically Enabled Multiscale Manipulation, Fig. 4 REP and O-OEW have been combined for achieving manipulation across length scales. (**a**) A droplet containing micron-sized particles is steered to an appropriate location using O-OEW. (**b**) REP is applied to aggregate micron-sized particles at a user-defined location. The particles aggregate under the influence of lateral ACEO and toroidal electrothermal flows (Reprinted with permission from Kumar et al. [8]. Copyright 2010 American Chemical Society)



Optoelectrically Enabled Multiscale Manipulation, Fig. 5 The combined technique was used to aggregate particles at a known location, while the droplet dried (*left*). Typically, such a drying process leaves behind a ring of particles at the periphery (*right*) (Reprinted with permission from Kumar et al. [8]. Copyright 2010 American Chemical Society)

leading to rapid and massive particle accumulation at the illuminated site on the substrate surface. The REP mechanism in this device is slightly more involved than the one portrayed by Williams et al. [4] The REP effect here is caused by the combined effects of ACEO flows and electrothermal flows (Fig. 4b).

While displaying the ability to manipulate across length scales, the researchers also demonstrated a unique microfabrication technique. They used REP to concentrate particles in a region over many minutes, while the droplet dried. Evaporation of the droplet left behind an assembly of several thousands of particles (Fig. 5). Typically, a drying droplet leaves behind an assembly of particles in the shape of a ring – the effect being popularly known as the "coffee ring effect." The coffee ring effect, other than its popular appeal, has been vigorously studied for its varied applications such as in printing and sensing. The scientific community has long pursued techniques of altering the coffee ring effect by passive



Optoelectrically Enabled Multiscale Manipulation, Fig. 6 Illustration of operational principle. (**a**) Device operated by OEW (valid for frequencies, f < 100 kHz). This causes aqueous to move toward the light spot. Particles within the droplet are transported along with the droplet. (**b**) Device operated by OETs (f > 100 kHz). Particles within the droplet experience a DEP force when in the vicinity of incident optical pattern. In this regime, the device electrically looks identical to standard OETs (Valley et al. [9]. Reproduced with permission from The Royal Society of Chemistry)

techniques. Kumar et al.'s work [8] showed a novel way of actively altering the coffee ring effect.

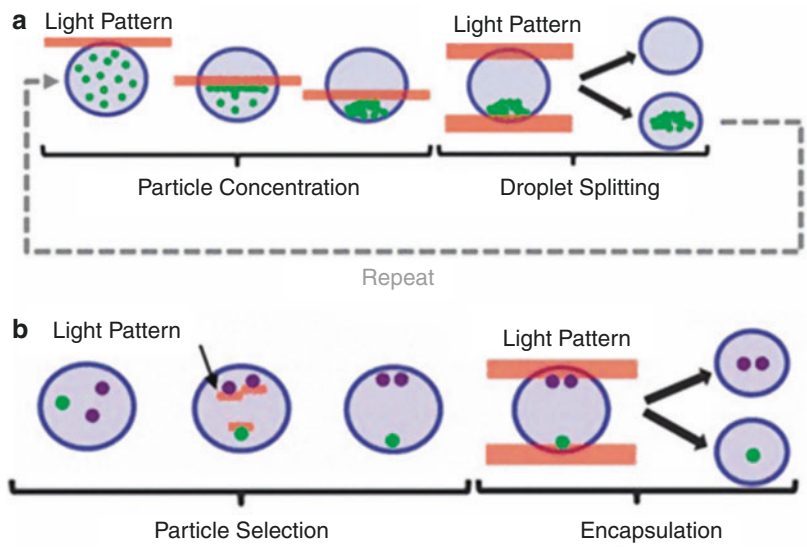
OET+OEW

A unique optoelectric platform integrating OEW and OET was first demonstrated by Valley et al. [9]. The platform is designed to manipulate droplets by OEW and suspended particles by OE-T. The maximum translational speeds of droplet and particle were reported to be 8 mm/s and 60 μm/s, respectively. The illustrative principles of the platform are depicted in Fig. 6. The platform is composed of indium tin oxide (ITO), a-Si:H, aluminum oxide (Al_2O_3), and Teflon for the bottom plate and Teflon and ITO for the top plate. An aqueous droplet containing particles is sandwiched between two plates. The distinct peak for the OEW operation is at 10–20 kHz, while the peak for OET is at 100–200 kHz.

Using this technique, polystyrene particles of 10 μm were concentrated by slowly sweeping a light bar across the droplet from one end to the other with a potential of 16 V_{pp} at 200 kHz (Fig. 7a) [9]. Subsequently, the platform was switched to the OEW mode with a potential of 32 V_{pp} at 10 kHz for droplet splitting. Compared to droplet splitting without OETs, the efficiency increased from 50 % to 93 %. Serially repeating the procedure of concentration and splitting increased the particle density (i.e., particles per volume) because the volume was divided in half each time. A second demonstration showed how a single cell can be selected and encapsulated using the hybrid technique (Fig. 7b). By changing the light image from one single bar to multiple bars, a specific cell was isolated from a group of HeLa cells. After splitting the droplet, the selected cell was solely encapsulated in an aqueous droplet.

OETs+EW

EWOD has been an effective means of manipulating droplets since it was first reported [18, 19]. OET is a recent tool that can handle



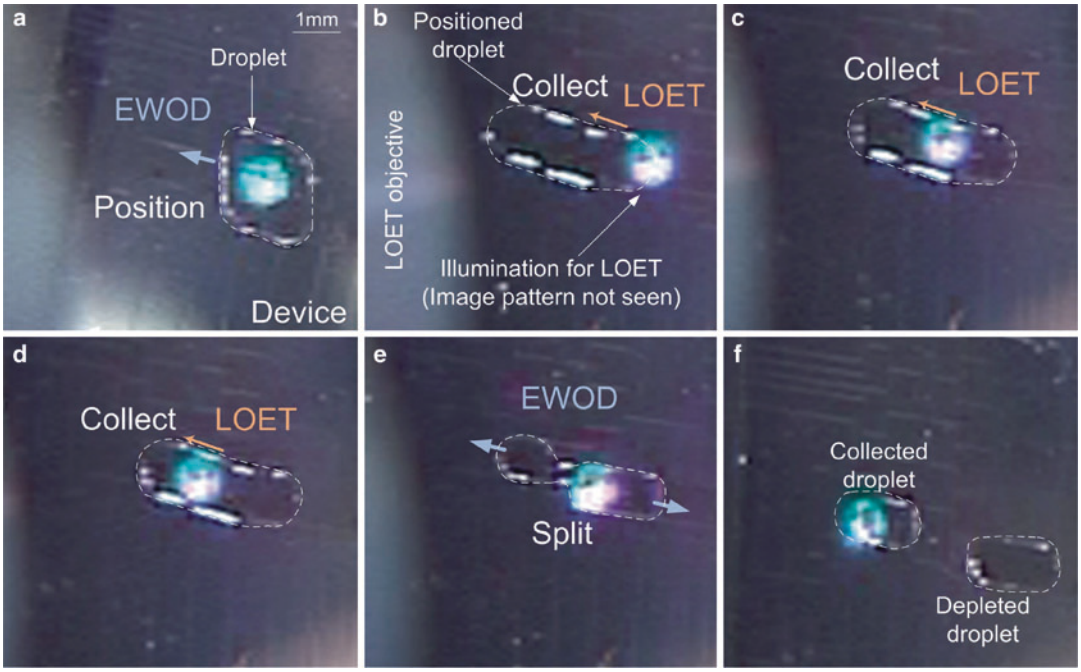
Optoelectrically Enabled Multiscale Manipulation, Fig. 7 (**a**) Procedure of concentration of a clump of particles. (**b**) Procedure of single cell selection and encapsulation (Valley et al. [9]. Reproduced with permission from The Royal Society of Chemistry)

microparticles in a dynamic way. The integration of OETs and EW leads to a brand new platform suitable for multi-scale biological applications. In this case, large discrete electrodes were physically fabricated on a substrate to address droplets. Interdigitated electrodes were fabricated on another substrate to steer particles. It was found that the current arrangement of EWOD and OETs on separate plates yields a better performance than the previous generation which has EWOD and OET devices interweaving together [7]. OETs use a photoconductor to gate the particle actuation, but EWOD uses an insulator, silicon nitride (Si_3N_4), to generate a major potential drop for electrowetting effect. Both plates are covered with a hydrophobic layer, Cytop, to provide high contact angle as well as insulating effect [7].

A driving potential of 3–5 V_{rms} at 200 kHz for OETs and a driving potential of about 120 V_{rms} at 20 kHz for EWOD were employed. The working medium of droplet was DI water, such that the live HeLa cells showed positive OETs at the frequency of the electric field used. The operational procedures are as follows (Fig. 8): (1) stretch the droplet in an elongated shape using EWOD, (2) sweep a light bar over the cells in one direction multiple times, (3) split the droplet, and (4) collect the live cells and dump the dead cells. Although only mammalian cells were demonstrated on this platform, the mechanism can be extended to all other biological species, such as bacteria, yeasts, etc. It should be noted that any targets smaller than sub-micrometers may not experience much OET forces due to the r^3 scaling law of DEP force.

Future Directions

The ability to manipulate fluids and process various biological species for sequential reactions constantly poses challenges to the current lab-on-a-chip applications. The typical challenges while dealing with cross-scale tasks include connections between different scale regimes, compatibility between techniques, sample collecting, cross-contamination, and disturbances from fluidic components or manipulation tools. As the technology advances, researchers are not just satisfied with the knowledge on the microscale but keen to explore the frontier on the nanoscale. However, except for some existing mechanical measures, such as AFM and STM, noninvasive techniques seem still far-fetched. Although, at present, scientists have been developing some tools to handle multi-scale targets, there remains a long journey before fully performing operations crossing more than five orders of length scale. Currently, the optoelectric devices coupled with



Optoelectrically Enabled Multiscale Manipulation, Fig. 8 Images showing the sequential operations of EWOD and OETs on the integrated device. The thick (*blue*) *arrows* represent EWOD actuation and the thin (*orange*) *arrows* represent OET actuation. (**a**) Droplet (350 nL) containing HeLa cells is positioned in the work site. (**b**–**c**) Droplet is stretched by EWOD, while the live cells are swept by OETs from one end to the other. (**e**) When all live cells are concentrated in the left side, EWOD forces then act on both ends of the stretched droplet to split it. (**f**) After the droplet breaks up, the left droplet contains the collected live HeLa cells, but the right droplet contains only dead cells (Shah et al. [7]. Reproduced with permission from The Royal Society of Chemistry)

O

digital microfluidics (i.e., droplet manipulation) provide a niche in multi-scale manipulation. However, most operations are limited to the range from millimeters to mesoscale (sub-micrometers). Although both REP and OETs have manipulated micro- and nanoparticles under special conditions, considerable effort must be taken, and single-molecule manipulation remains difficult. It is highly anticipated that new breakthroughs in optoelectric devices can be made to resolve the aforementioned difficulties, such that the genuine multi-scale manipulation will be achieved to benefit future scientific research.

Acknowledgments A. Kumar performed the work as Eugene P. Wigner's fellow and staff member at the Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy under Contract No. DE-AC05-00OR22725.

Cross-References

- ▶ Digital Microfluidics
- ▶ Electrowetting
- ▶ Lab-on-a-Chip
- ▶ Rapid Electrokinetic Patterning and Its Applications

References

1. Chiou, P.Y., Moon, H., Toshiyoshic, H., Kim, C.J., Wu, M.C.: Light actuation of liquid by optoelectrowetting. Sens. Actuators A **104**, 222–228 (2003)
2. Chuang, H.S., Kumar, A., Wereley, S.T.: Appl. Phys. Lett. **93**, 064104 (2008)
3. Chiou, P.Y., Ohta, A.T., Wu, M.C.: Nature **436**, 370–372 (2005)
4. Williams, S.J., Kumar, A., Wereley, S.T.: Lab Chip **8**, 1879–1882 (2008)
5. Cho, S.K., Zhao, Y.J., Kim, C.J.: Lab Chip **7**, 490–498 (2007)

6. Fan, S.K., Huang, P.W., Wang, T.T., Peng, Y.H.: Lab Chip **8**, 1325–1331 (2008)
7. Shah, G.J., Phta, A.T., Chiou, E.P.Y., Wu, M.C., Kim, C.J.: Lab Chip **9**, 1732–1739 (2009)
8. Kumar, A., Chuang, H.S., Wereley, S.T.: Langmuir **26**, 7656–7660 (2010)
9. Valley, J.K., NingPei, S., Jamshidi, A., Hsu, H.-Y., Wu, M.C.: Lab Chip **11**, 1292–1297 (2011)
10. Kumar, A., Williams, S.J., Wereley, S.T.: Microfluid. Nanofluid. **6**, 637–646 (2009)
11. Kumar, A., Williams, S.J., Chuang, H.S., Green, N.G., Wereley, S.T.: Lab Chip **11**, 2167–2174 (2011)
12. Kwan, J.S., Ravindranath, S., Kumar, A., Irudayaraj, J., Wereley, S.T.: Lab Chip **12**, 4955–4959 (2012)
13. Wang, K.C., Kumar, A., Williams, S.J., Green, N.G., Kim, K.C., Chuang, H.S.: Lab Chip **14**, 3958–3967 (2014)
14. Chiou, P.Y., Wong, W.H., Liao, J.C, Wu, M.C.: Proceeding of the 17th IEEE ICMEMS, Maastricht, Netherlands, pp. 21–24 (2004)
15. Hwang, H., Park, J.-K.: Lab Chip **11**, 33–47 (2011)
16. Ohta, A.T., Chiou, P.Y., Han, T.H., Liao, J.C., Bhardwaj, U., McCabe, E.R.B., Yu, F.Q., Sun, R., Wu, M.C.: J. Microelectromech. Syst. **169**, 491–499 (2007)
17. Chiou, P.Y., Chang, Z., Wu, M.C.: J. Microelectromech. Syst. **17**, 133–138 (2008)
18. Cho, S.K., Moon, H., Kim, C.J.: J. Microelectromech. Syst. **12**, 70–80 (2003)
19. Pollack, M.G., Shenderov, A.D., Fair, R.B.: Lab Chip **2**, 96–101 (2002)

Optoelectrofluidics

► Optoelectrically Enabled Multiscale Manipulation

Optoelectrokinetic Technique

► Rapid Electrokinetic Patterning and Its Applications

Optoelectronic Properties

► Optical and Electronic Properties

Optofluidic Cavities

► Optofluidic Resonators and Sensors

Optofluidic Resonators and Sensors

Zhenyu Li
Department of Biomedical Engineering, The George Washington University, Washington, DC, USA

Synonyms

Optofluidic Cavities

Definition

Optofluidic resonators are optical microcavities that are partially or completely made of liquids. Sensors based on optofluidic resonators use either passive cavity resonance shift or active lasing to detect small changes in refractive index, absorption, gain, or scattering in the vicinity of the resonator.

Overview

Optofluidics refers to the synergistic fusion of microfluidics and photonics [1, 2]. By integrating photonic and fluidic functions on a single chip, optofluidics not only gives rise to highly compact and integrated microsystems but also enables novel adaptive and reconfigurable photonic devices via microfluidic liquid manipulation. Although both photonics and fluidics have been independently miniaturized within the past three decades, it was not until very recently that the monolithic integration of the two became possible. Today, the channel dimensions in microfluidic devices have become comparable to or even smaller than the wavelength of light (i.e., <200 nm for visible light), and the materials

involved are often optically transparent. These two facts led to the realization that it is now feasible to build liquid-core or liquid cladding that are inherently compatible with microfluidic adaptation and integration. Optofluidic resonators are one of such structures.

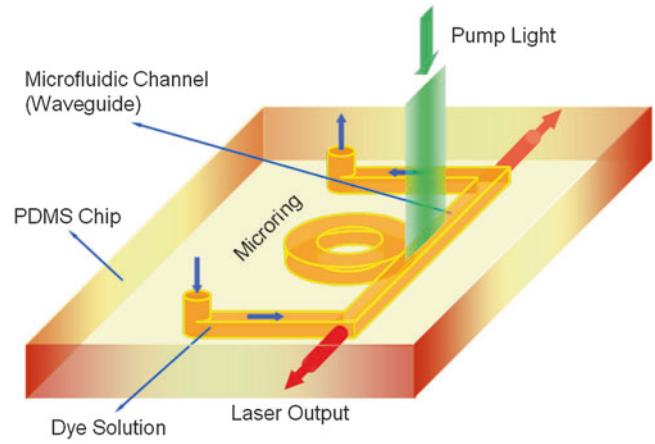
In optofluidic resonators, part of the optical cavity is made of liquids. For example, a liquid droplet suspended in another immiscible liquid with lower refractive index forms a so-called whispering gallery mode (WGM) resonator, in which light is confined by continuous total internal reflections at the liquid/liquid interface [3]. Compared with their solid-state equivalents, optofluidic resonators offer some unique advantages, including

1. Liquid-core or liquid-cladding cavities enable tunable optical resonances with extremely large tuning range (refractive index change Δn can be as high as ~ 0.5 compared with <0.01 in most solid-state materials) by liquid filling, mixing, or replacement.
2. Flexibility of optofluidic resonator structures allows one to easily introduce functional optical materials such as dye molecules (optical gain or absorption), nonlinear optical liquids, liquid crystals, quantum dots, nanoparticles, and biomolecules into/onto the optical resonator. This can give rise to many functional photonic and sensing devices that are inherently tunable and reconfigurable.
3. Molecularly smooth optical surfaces induced by interfacial tensions enable ultrahigh quality factor (Q) resonators. For example, microdroplets, with their molecularly smooth surfaces and small sizes, represent an attractive optical resonator for making low-threshold lasers and ultrasensitive sensors [3]. The Q factors of microdroplet cavities can be as high as 108, and their mode volumes can be as small as a few picoliters [3], which makes them extremely sensitive to small changes in their environments.
4. Inherently compatibility with microfluidics allows for the seamless integration of liquid sample manipulation steps with optical sensing and characterization capabilities. Moreover, lithographically defined accurate optical alignment and no mechanically moving parts give rise to more reproducible, robust, and easy-to-use systems.
5. Simultaneous liquid and optical confinements at micron or submicron scale allow for highly efficient optical interrogation of target molecules or particles in subpicoliter volume liquids, which is ideally suited for single-molecule detection, single-cell analysis, and low-abundance biosensing applications.

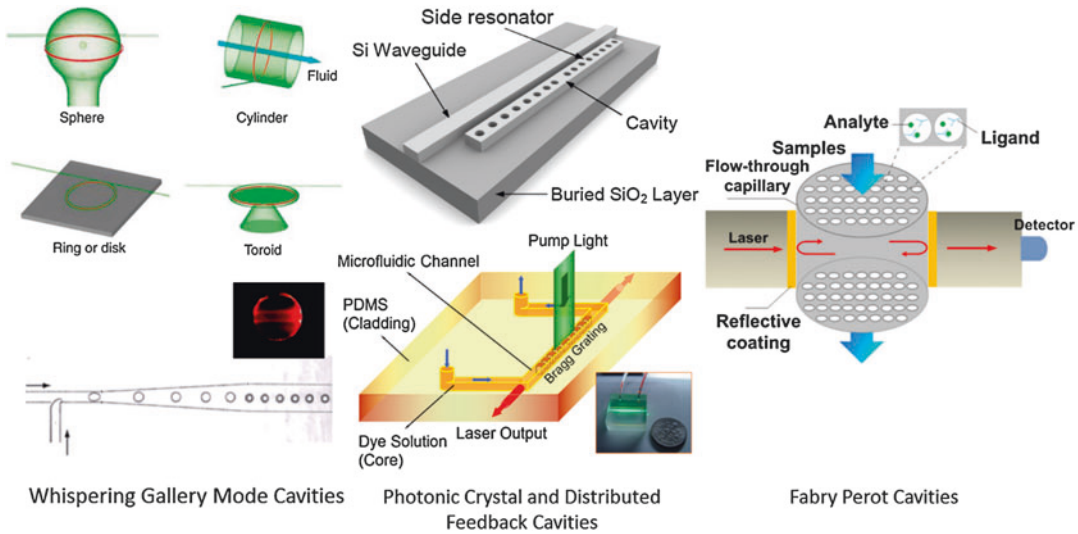
An example of liquid-core microring resonator is described below to illustrate the general principle of operation. Then optofluidic resonators as sensors, in particular biosensors, are discussed.

Principle of Operation

Optical resonators (or cavities) confine light both spatially and temporally in some structurally defined resonance modes [3]. Traditional optical resonators are made of solid-state materials, whereas optofluidic resonators are partially or completely made of liquids. Figure 1 shows an example of a liquid-core guided-wave optofluidic microring resonator for dye lasing application [4]. In this device, the optofluidic resonator is made of a ring-shaped microfluidic channel (2 μm high and 2 μm wide) filled with a higher refractive index liquid (e.g., ethylene glycol, a common organic solvent, $n = 1.43$) than that of the cladding. The cladding of the waveguide and the microring resonator is made of poly(dimethylsiloxane) (PDMS) (RTV 615, $n = 1.406$ at 589 nm), a widely used polymer for microfluidic devices [5]. In such liquid-core optofluidic resonators, both the spatial modes and the resonance wavelengths can be controlled, to some degree, by the choice of the liquid. For example, by mixing two different solvents such as ethylene glycol and ethanol, the core refractive index can be tuned from 1.36 to 1.43. When the core/cladding index contrast is less than 0.003 (i.e., core refractive index is less than 1.409 when using a PDMS cladding), the waveguide supports only the two degenerate fundamental transverse modes. The



Optofluidic Resonators and Sensors, Fig. 1 *Optofluidic ring resonator.* Schematic of a liquid-core optofluidic microring resonator for dye laser application. The liquid core in this device provides the laser cavity, light confinement, and optical gain, which are tunable by liquid manipulation



Optofluidic Resonators and Sensors, Fig. 2 *Classification of optofluidic resonators:* (*left*) whispering gallery mode (*WGM*) cavities [6–8]; (*middle*) photonic crystal and 1D distributed feedback (*DFB*) type cavities [9, 10]; (*right*) Fabry-Perot type cavity [11]. Either the resonator itself (core) or its surrounding (cladding) is liquid

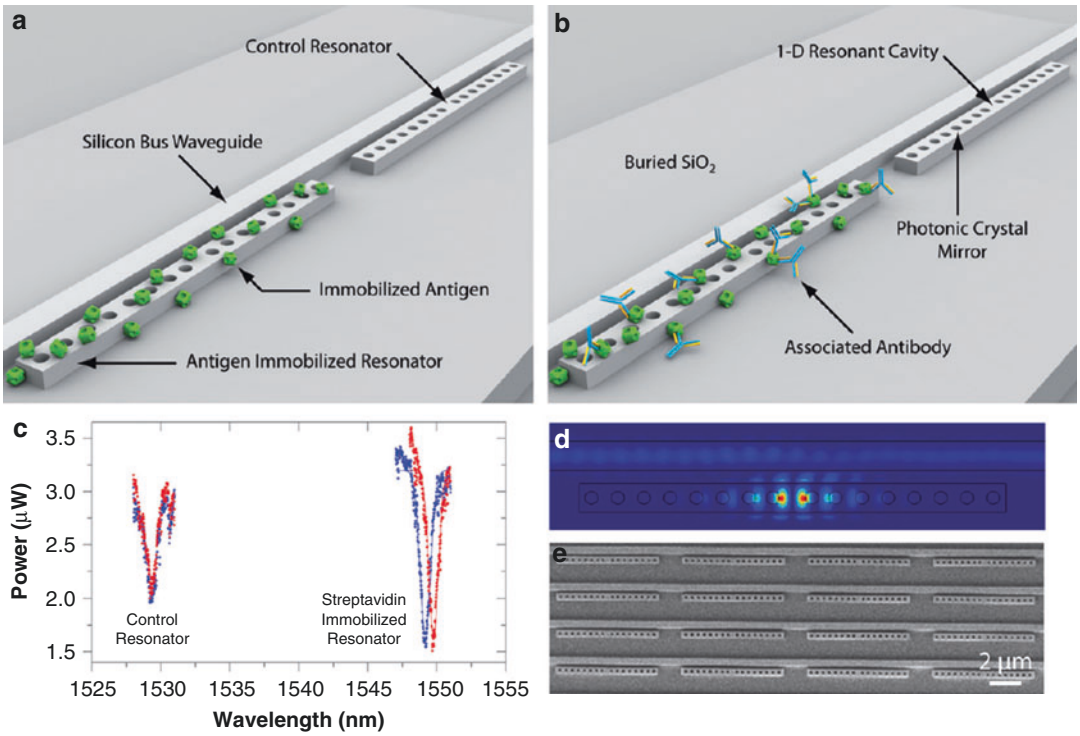
resonant wavelengths of a microring resonator are determined by the condition

$$m\lambda_m = 2\pi n_{eff} R \qquad (1)$$

where λ_m is the mth resonant wavelength, n_{eff} is the effective index of the guided mode, and R is the ring radius. It's clear from this expression that the optofluidic structure enables tunable resonances by changing n_{eff} or m, which provides a convenient tuning mechanism. Changing the refractive index of the liquid core changes the effective index, and switching dye molecules covering different spectral regions can change the resonance order m. This intrinsic optofluidic tunability can also be used for sensing applications because any change in refractive index in the liquid core will lead to a resonance shift.

Types of Optofluidic Resonators

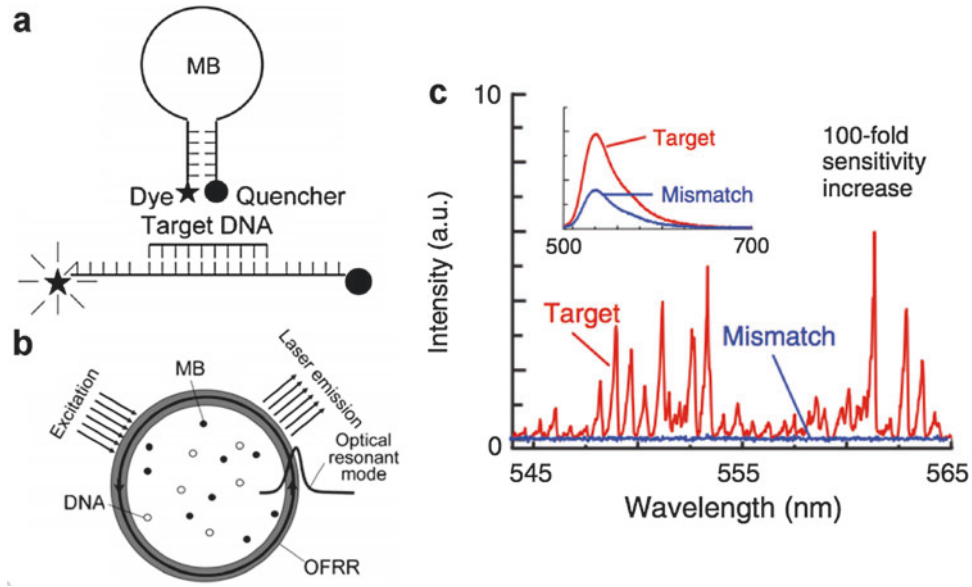
Optofluidic resonators can be classified into liquid-core, liquid-cladding, and all-liquid resonators based on the liquid components. In addition, the type of the cavity can be used to classify optofluidic resonators. Figure 2 shows the three most common cavity types: whispering gallery mode (WGM) cavities [6, 7], photonic crystal or DFB type cavities [8, 9], and Fabry-Perot type cavities [10]. Each type of optofluidic resonator has its own unique advantages and suitable applications. For example, WGM cavities can have ultrahigh quality (Q) factors if the cavity surface is formed by some interfacial tension effect such as a melted glass microsphere or a liquid droplet. High Q factors can give rise to very narrow linewidth cavity resonances favorable for low-threshold lasing and ultrasensitive sensing applications. Photonic crystal or DFB cavities are more amenable to planar lithographic fabrication process, thus more suited for on-chip integration with microfluidics. Microfluidic channels can be cofabricated with the optofluidic cavity structure to allow for on-chip microfluidic tuning and sample introduction. Fabry-Perot cavity is well suited for intracavity sensing and spectroscopy applications because the ease of access to its introcavity space. Thus far, all types of cavities have been demonstrated in optofluidic formats for various applications. For comprehensive review on their implementations and applications, references [12] and [6] are recommended.



Optofluidic Resonators and Sensors, Fig. 3 *Liquid-cladding Optofluidic Photonic Crystal (PC) Resonators for Biosensing.* Core is silicon, and cladding is 1× PBS buffer containing 0.1 % Tween-20 and 0.2 mg/ml bovine albumin (PBST). (**a**) 3D rendering of two optofluidic PC resonators evanescently coupled to a bus waveguide. One resonator is immobilized with an antigen, and the other acts as a control. (**b**) 3D rendering of the binding of the target antibody to the antigen immobilized resonator. (**c**) Experimental data showing the detection of 45 mg/ml antistreptavidin antibody. The blue curve shows the baseline spectrum corresponding to Fig. 3a. The red trace shows the test spectra after the binding of antistreptavidin as illustrated in Fig. 3b. (**d**) Finite difference time domain (*FDTD*) simulation of the steady-state electric field distribution within the PC resonator at the resonant wavelength. (**e**) SEM image showing the multiplexing capability of the optofluidic PC resonator array [9]

Sensing Applications

The liquid-core or liquid-cladding structure combined with compatibility with microfluidic sample manipulation makes optofluidic resonators attractive for biosensing applications. In fact, most optofluidic resonator structures can function as sensors by simply replacing the liquid core or cladding with another liquid carrying the target analytes such as a buffer solution with some biomolecues of interest. Figure 3 shows a passive liquid cladding 1D photonic crystal (PC) resonator for protein detection [9]. In this sensor, a nanofabricated on-chip silicon 1D PC resonator is covered by a PDMS microfluidic chamber. The optofluidic PC resonators were evanescently coupled to act as a control. The liquid (or cladding) is 1× PBS buffer containing 0.1 % Tween-20 and 0.2 g/ml bovine albumin (PBST). When antistreptavidin containing PBST was introduced into the chamber, the resonance wavelength shifted (Fig. 3c). The highly localized optical mode greatly enhanced light-matter interaction and enabled target detection on the order of 63 ag bound mass. The planar structure of the device also enabled multiplexed detection of interleukins 4, 6, and 8 using a sandwich assay on an array of PC resonators. Figure 4 shows another unique biosensing application enabled by active optofluidic resonators, i.e., lasers [13]. Traditionally, single base-pair DNA mismatch is detected by fluorescence-based detections. However such detection methods have low discrimination ratios because the fluorescence different is caused by a single base mismatch. Fan et al. devised a unique signal amplification scheme by using an optofluidic laser as the sensor. An optofluidic ring resonator (OFRR) was used as the laser cavity. Target DNA and molecular beacon (MB) probes were incorporated as part of the laser gain medium. A perfectly matched (PM) DNA and a single-base mismatched (SM) DNA are used to test the system performance. Although a small binding affinity difference between PM DNA and SM DNA causes only a small change in the laser gain, this small difference is amplified by the laser action into two orders of magnitude differences in the



Optofluidic Resonators and Sensors, Fig. 4 *Active Optofluidic Ring Laser for Single Base Mismatch Detection in DNA*. (**a**) Molecular beacon (*MB*) fluorescence is quenched when it is in the closed state and restored upon hybridization with the target DNA. (**b**) Intracavity DNA detection and differentiation with the optofluidic ring resonator (*OFRR*) laser. (**c**) Detection sensitivity was increased 100-fold by using the OFRR laser when compared to fluorescence-based detection [13]

emission intensity because of the strong optical feedback in the laser cavity. Detection sensitivity was increased 100-fold by using the OFRR laser when compared to fluorescence-based detection. The potential of active optofluidic resonators and lasers has been reviewed in [14]. It is noted that Fan et al. has demonstrated an optofluidic laser with a single molecular layer of gain made of GFP or dye-labeled DNA [15].

Cross-References

► Whispering Gallery Mode Resonator Biosensors

References

1. Psaltis, D., Quake, S.R., Yang, C.: Developing optofluidic technology through the fusion of microfluidics and optics. Nature **442**, 381–386 (2006)
2. Monat, C., Domachuk, P., Eggleton, B.J.: Integrated optofluidics: a new river of light. Nat. Photonics **1**, 106–114 (2007)
3. Vahala, K.: Optical Microcavities. World Scientific, Singapore (2005)
4. Li, Z.Y., Zhang, Z.Y., Scherer, A., Psaltis, D.: Optofluidic microring dye laser. IEEE/LEOS Summer Topical Meetings, Jul 2007
5. McDonald, J.C., Whitesides, G.M.: Poly(dimethylsiloxane) as a material for fabricating microfluidic devices. Acc. Chem. Res. **35**(7), 491–499 (2002)
6. Vollmer, F., Arnold, S.: Whispering-gallery mode biosensing: label-free detection down to single molecules. Nat. Methods **5**(7), 591–596 (2008)
7. Fan, X., White, I.M.: Optofluidic microsystems for chemical and biological analysis. Nat. Photonics **5**, 591–597 (2011)
8. Bailey, R.: Future Sci. Bioanalysis **1**(6), 1043–1047 (2009)
9. Mandal, S., Goddard, J., Erickson, D.: A multiplexed optofluidic biomolecular sensor for low mass detection. Lab Chip **9**, 2924–2932 (2009)
10. Li, Z.Y., Zhang, Z.Y., Emery, T., Scherer, A., Psaltis, D.: Single mode optofluidic distributed feedback dye laser. Opt. Express **14**, 696–701 (2006)
11. Gather, M.C., Yun, S.H.: Single-cell biological lasers. Nat. Photonics **5**, 406–410 (2011)
12. Li, Z., Psaltis, D.: Optofluidic dye lasers. Microfluid. Nanofluid. **4**, 145–158 (2008)
13. Sun, Y., Fan, X.: Distinguishing DNA by analog-to-digital like conversion by using optofluidic lasers. Angew. Chem. Int. Ed. **51**, 1236–1239 (2012)
14. Fan, X., Yun, S.-H.: The potential of optofluidic biolasers. Nat. Methods **11**, 141–147 (2014)
15. Chen, Q., Ritt, M., Sivaramakrishnan, S., Sun, Y., Fan, X.: Optofluidic lasers with a single molecular layer of gain. Lab. Chip **14**, 4590–4595 (2014)

Optofluidics

► Rapid Electrokinetic Patterning and Its Applications

Optomechanical Resonators

► Nanomechanical Resonators Based on III–V Semiconductors
► Nano-optomechanical Systems (NOMS)

Optomechanics

► Nano-optomechanical Systems (NOMS)

Opto-mechano-fluidics

► Microfluidic Optomechanics

O

Organic Actuators

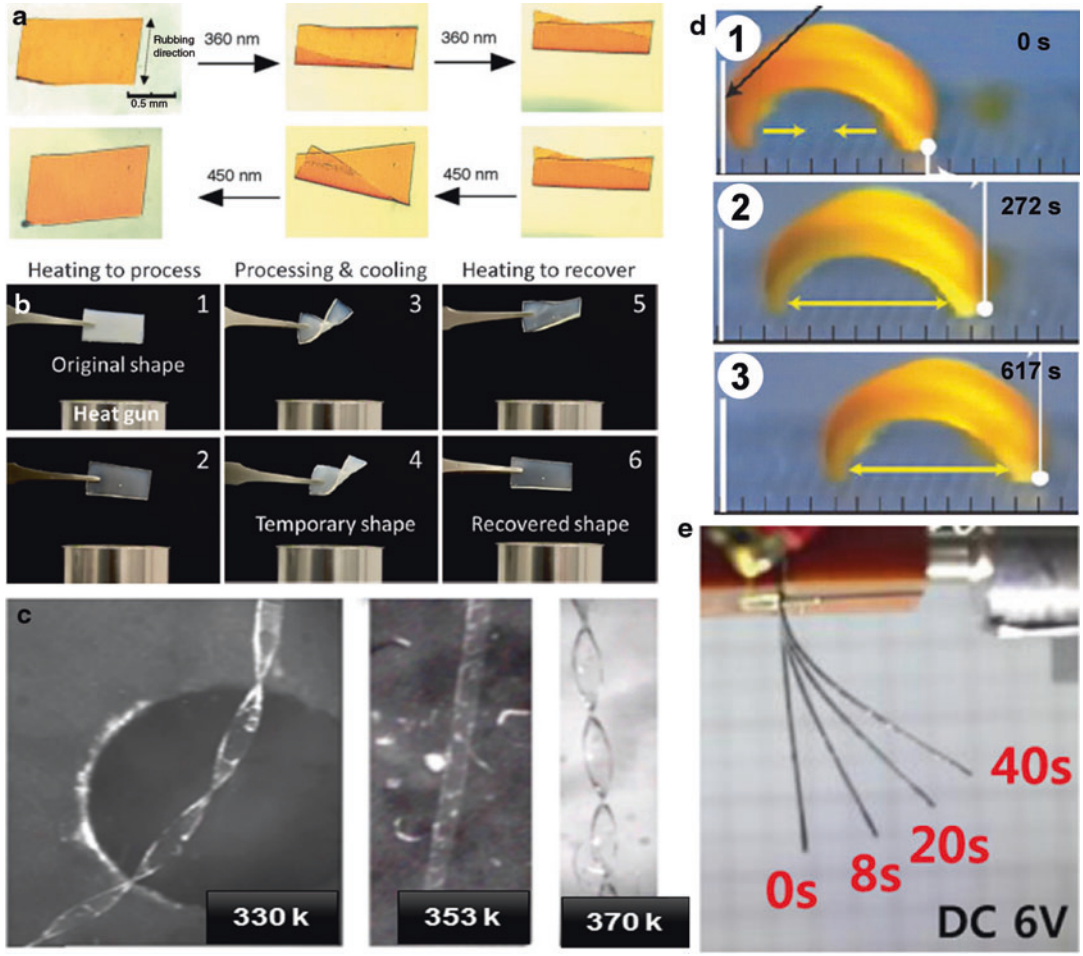
Yuyun Liu and Yanlei Yu
Department of Materials Science and State Key Laboratory of Molecular Engineering of Polymers, Fudan University, Shanghai, China

Synonyms

Soft actuators

Definition

Organic actuators are energy transducers that are made of organic materials such as stimuli-



Organic Actuators, Fig. 1 The various movements of organic actuators: (**a**) Reversible bending of a crosslinked liquid crystal polymer film induced by UV and Vis light irradiation [1]. Copyright 2003 Wiley-VCH. (**b**) Demonstration of the shape-memory behavior of 2-perfluorooctylethyl acrylate induced by heating [2]. Copyright 2014 Wiley-VCH. (**c**) Helicoids formed by the narrow twist-nematic crosslinked liquid crystal polymer films: left-handed at 370 K; almost flat at 353 K; right-handed at 330 K [3]. Copyright 2011 PNAS. (**d**) The self-walking motion of the gel actuator by the Belouzov-Zhabotinsky (BZ) reaction [4]. Copyright 2007 Wiley-VCH. (**e**) The actuators behavior of an ionic polymer driven by electrical stimuli [5]. Copyright 2014 American Chemical Society

responsive gels or polymers and able to act upon their external environment by converting input energy into mechanical work.

Overview

Organic actuators have shown their considerable potential for the applications in medical devices, prostheses, robotics, toys, biomimetic devices, and micro/nanoelectromechanical systems. Compared to conventional metallic and ceramic actuators, the advantages of organic actuators are flexibility, light weight, low cost, and quiet operation compliance. Many actuators that respond to various external stimuli have been developed using polymers as base materials: polymer gels, cross-linked liquid crystal polymers, conducting polymers, and shape memory polymers. As Fig. 1 shows, the promise of developing organic

actuators has inspired scientists for decades and led to the design and operation of polymers that exhibit mechanically relevant movements, such as contraction/expansion, bending, twisting, swimming, and inchworm walking [1–5].

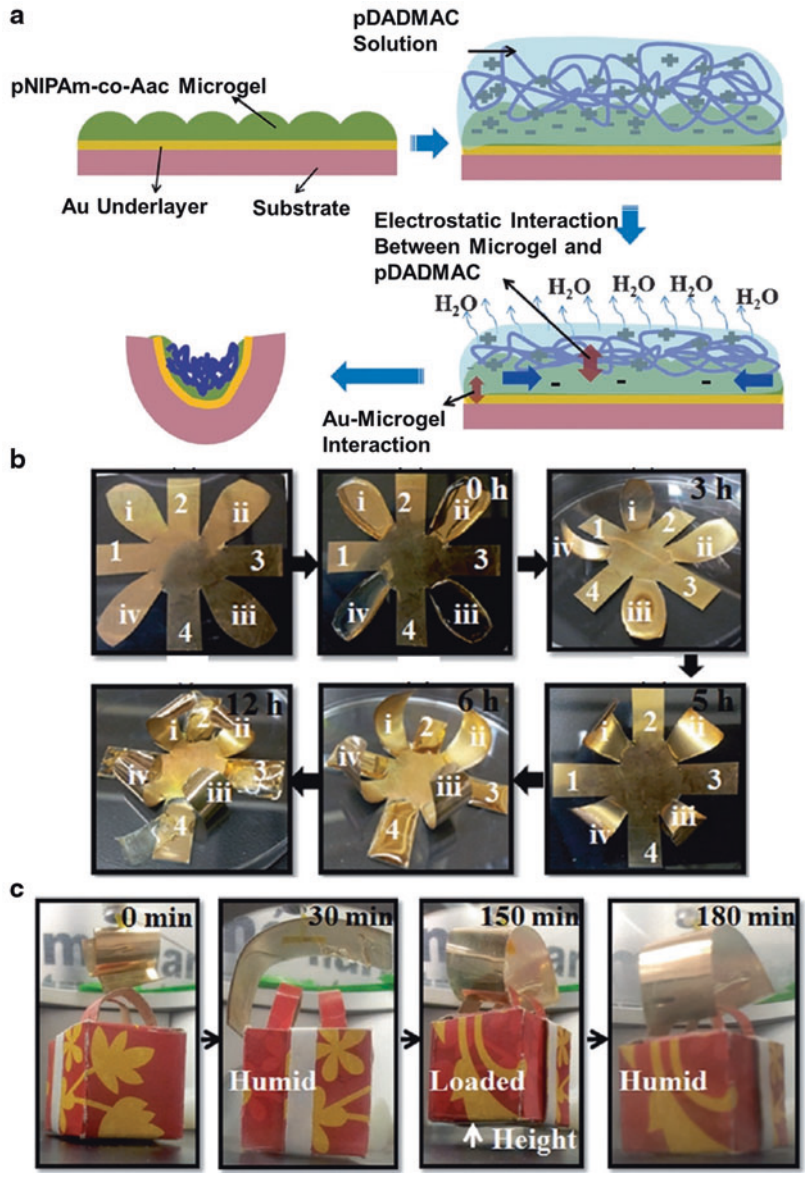
Gel Actuators

A polymer gel contains an elastic cross-linked network and fluid filling in their interstitial space of three-dimensional network, which provides softness as well as high biocompatibility. Gels have the ability to swell or shrink due to the large amount of liquid they can hold and undergo change in shape or volume in response to external stimuli, i.e., thermal, electric energy, light, and pH. Research in stimuli-responsive polymer gels, especially hydrogels, has persisted over several decades, and a great deal of work has been dedicated to develop useful actuators [6].

One of the most studied hydrogels to date is poly(N-isopropylacrylamide) (pNIPAm), which is often used as a matrix of hydrogels. A pNIPAm-microgel-based device is able to do work and lift masses many times the mass of the device, in response to simple changes in the humidity of its environment [7]. Acrylic acid (AAC), a weak acid having a pKa of approximately 4.25, is added to pNIPAm-based microgels. At pH $>$ 4.25, the AAc groups are deprotonated, thus making the microgels negatively charged (polyanionic), while they are neutral at pH $<$ 4.25 owing to AAc protonation. As Fig. 2a shows, the monolayer of pNIPAm-co-AAc microgels is deposited on an Au-coated substrate. Subsequently, a solution containing oppositely charged poly(diallyldimethylammonium chloride) (pDADMAC) is added to the microgel-coated substrate. The pDADMAC solution had a pH of 6.55, which rendered the microgel layer polyanionic; thus, the pNIPAm-co-AAc microgels swell by reason of the charge-charge repulsion in the polymer network of the microgels. Then, the pDADMAC layer contracts after the water evaporation due to the increase of hydrophobic interactions between the pDADMAC chains. Since the electrostatic interactions between pDADMAC and the microgels, and the interaction between microgels and the Au layer, are strong, three of them as an integration bend when the polymer dries. In other words, upon drying, the pDADMAC layer contracts bending the substrate.

As Fig. 2b shows, a flexible substrate with alternate arms was exposed to the same pDADMAC solution used above while keeping the other arms empty. The arms exposed to the pDADMAC solution curled up owing to the strong interactions between the pDADMAC and the microgels after solution drying. However, there was no response when the empty arms were exposed to polyanionic poly(sodium 4-styrenesulfonate) (PSS) solution (20 %) and dried in ambient condition. Figure 2c shows a device manipulated to lift a box by changing the humidity of a chamber. The curled substrate was hung from a string into a humidity chamber (0 min). As the humidity rose, the device uncurled and then loaded the box along with decrease of humidity to 10 %. While maintaining this humidity, the box was lifted off the chamber surface (150 min). A subsequent increase in the humidity allowed the device to reopen and drop the box (180 min). These arms are able to be actuated in response to humidity and act as muscles or arms to resist forces many times its own mass.

The development of remotely manipulating an actuator without any direct contact is of great importance. Yoshinori and coworkers designed a photoresponsive supramolecular actuator by integrating host-guest interactions and photoswitching ability in a hydrogel [8]. As Fig. 3 shows, a photoresponsive supramolecular hydrogel with α-cyclodextrin (αCD) as a host molecule and an azobenzene derivative as a photoresponsive guest molecule exhibit reversible macroscopic deformations in both size and shape when irradiated by UV light at 365 nm or visible light at 430 nm in water. After UV irradiation from the left side of the plate gel, the complex between αCD and cis-azo units decomposes to expand αCD-azo gels and bends to the right. Vis irradiation causes isomerization from the cis to the trans form, and complexation between αCD and the azo units regenerates, restoring the αCD-azo gel.



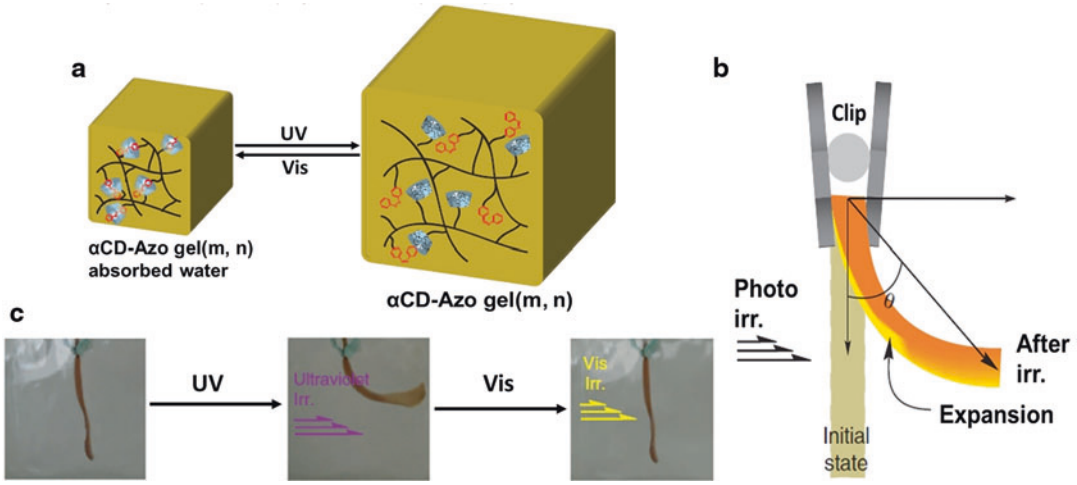
Organic Actuators, Fig. 2 Polymer-based muscle expansion and contraction. (**a**) PNIPAm-co-AAc microgels were deposited on the substrate obtaining a monolithic monolayer of microgels on the Au. PDADMAC solution renders the pNIPAm-co-AAc microgel negatively charged initiating the electrostatic interaction between the microgels and polyelectrolyte. Upon drying, the pDADMAC layer contracts bending the substrate. (**b**) Arms i-v exposed to the pDADMAC solution were bended while Arms 1–4 exposed to the PSS solution were no response after drying for 6 h. (**c**) The device lifted a box by changing the humidity of chamber [7] (Copyright 2013 Wiley-VCH)

Supramolecular hydrogels are hopeful to realize "muscle fibrils" in biological systems.

Shape-Memory Polymer Actuators

Shape-memory effect (SME) is the capability of a material to recover the original shape in a predefined way in response to an appropriate stimulus, such as light, electric fields, alternating magnetic fields, or water. Shape-memory polymers (SMP) are physical- or chemical-nature polymer networks, in which net-points being connected by chain segments determine the permanent shape [9]. Upon applied external stress, these chain segments could make certain an orientation to obtain the required deformability. Until an appropriate stimulus is applied to the shaped body, this temporary shape is induced to recover to the original shape.

So far, the most extensively investigated group of SMPs were thermally induced SMPs. SMPs



Organic Actuators, Fig. 3 Expansion-contraction of photoresponsive artificial muscle regulated by host-guest interactions. (**a**) Schematic illustration of the expansion and contraction of αCD-azo gel irradiated by UV and Vis light. (**b**) Lateral view of αCD-azo plate gel hung with a clip. (**c**) Image of the plate gel bending behavior in response to UV and Vis light [8] (Copyright 2012 Nature Publishing Group)

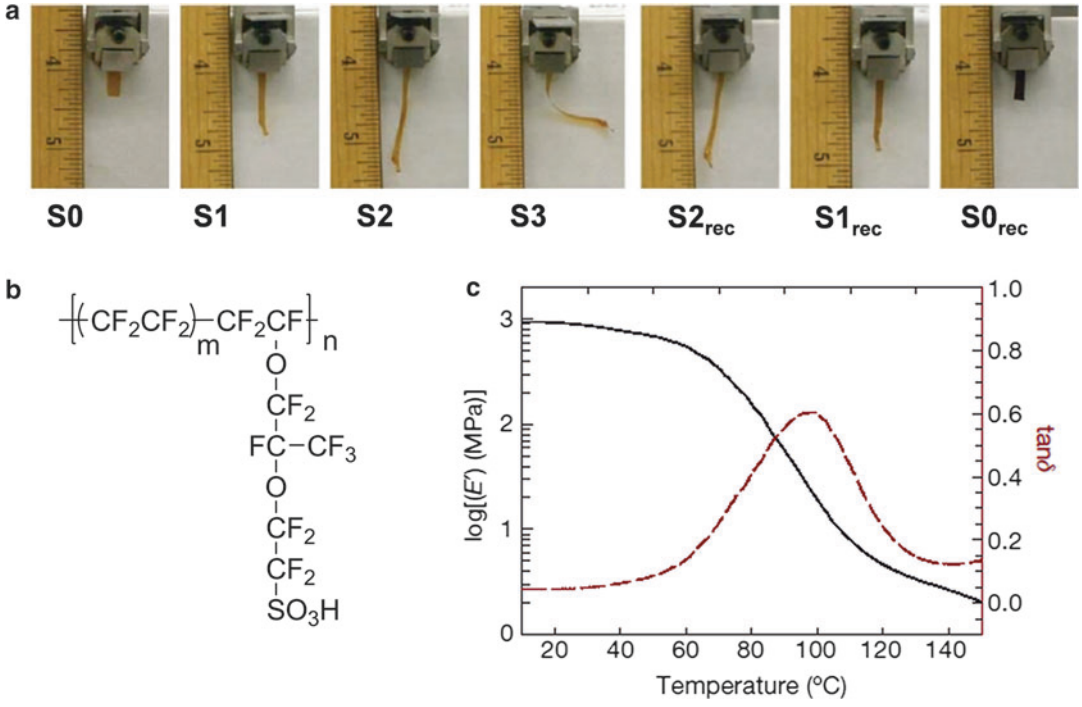
can be deformed by applying external stress upon transition temperature T_{trans} and fixed in a temporary shape. When cooled below T_{trans}, the temporary shape becomes stable. This process is called programming. Then, reheating induces the recovery of the stored permanent shape. Recently, Tao Xie et al. designed a multishape (triple or quadruple-shape) memory polymer and fulfilled sequential recovery to its original (permanent) shape from two, three, or four previously programmed temporary shapes [10, 11]. For example, perfluorosulphonic acid ionomer (PFSA) is a commercial thermoplastic polymer with a polytetrafluoroethylene backbone and perfluoroether sulphonic acid side chains (Fig. 4b). As Fig. 4a shows, starting as an original shape S0, PFSA can memorize three temporary shapes (S1, S2, and S3) in the same shape-memory cycle. Subsequent heating to the relevant temperatures led to the recovered shapes ($S2_{rec}$, $S1_{rec,}$ and $S0_{rec}$). The versatile shape-memory properties are stemmed from its broad glass transition of PFSA (55–130 °C, Fig. 4c). We can think of it as infinitely sharp transition temperatures continuously distributed across the broad transition.

However, SMPs may not be suitable when the material is expected to operate reversibly in response to repeated decreases and increases of temperature.

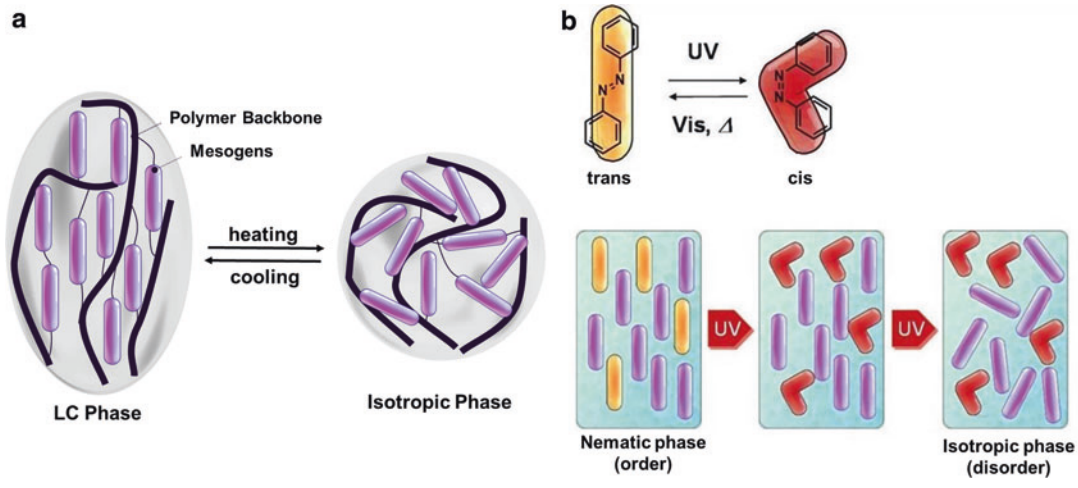
Cross-Linked Liquid Crystal Polymer Actuators

Liquid crystal (LC) is a thermodynamically stable fluid phase characterized by some anisotropy of properties. The LC molecules have a tendency to point along a common axis, called the director. They may be regularly arrayed in either one or two dimensions. Hence, LCs have many unique properties, such as a variety of different LC phases, differing in the positional or orientational order.

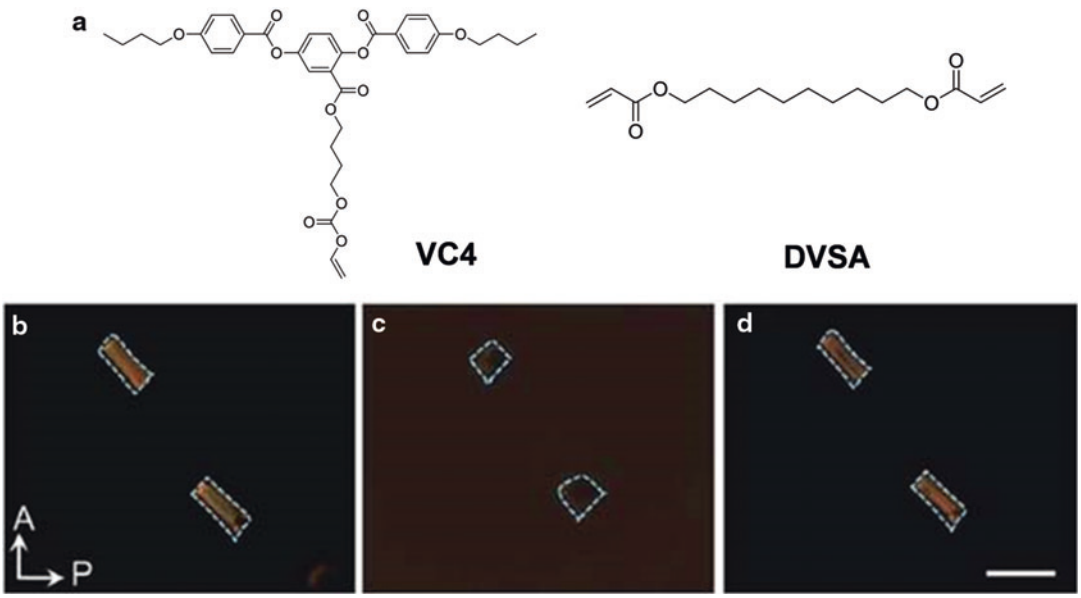
De Gennes proposed the possibility of using cross-linked LC polymers (CLCPs) as artificial muscles, by taking advantage of their substantial contraction in the direction of the director axis and cooperative motion [12]. For example, when the nematic CLCP films are heated above the nematic-isotropic phase transition temperature T_{trans}, the nematic order decreases, and the CLCPs exhibit a general contraction along the alignment direction of the mesogens; CLCP films will revert to their original size by expanding as the temperature drops below T_{trans} (Fig. 5a). Based on the above principle, it was reported that



Organic Actuators, Fig. 4 Tunable polymer multi-shape memory effect. (**a**) Quadruple-shape memory properties of PFSA. S0: permanent shape; S1: first temporary shape (140 °C); S2: second temporary shape (107 °C); S3: third temporary shape (68 °C); $S2_{rec}$: recovered second temporary shape (68 °C); $S1_{rec}$: recovered first temporary shape (107 °C); $S0_{rec}$: recovered permanent shape (140 °C). (**b**) Structure of PFSA; (**c**) Dynamic mechanical analysis curve of PFSA. tanδ is the ratio between the loss modulus and the storage modulus [10] (Copyright 2010 Nature Publishing Group)



Organic Actuators, Fig. 5 (**a**) Schematic illustration of thermal induced anisotropic contraction and expansion of the CLCPs. (**b**) Schematic illustration of photoisomerization of the azobenzene chromophores and photochemical phase transition of the azobenzene-containing CLCPs



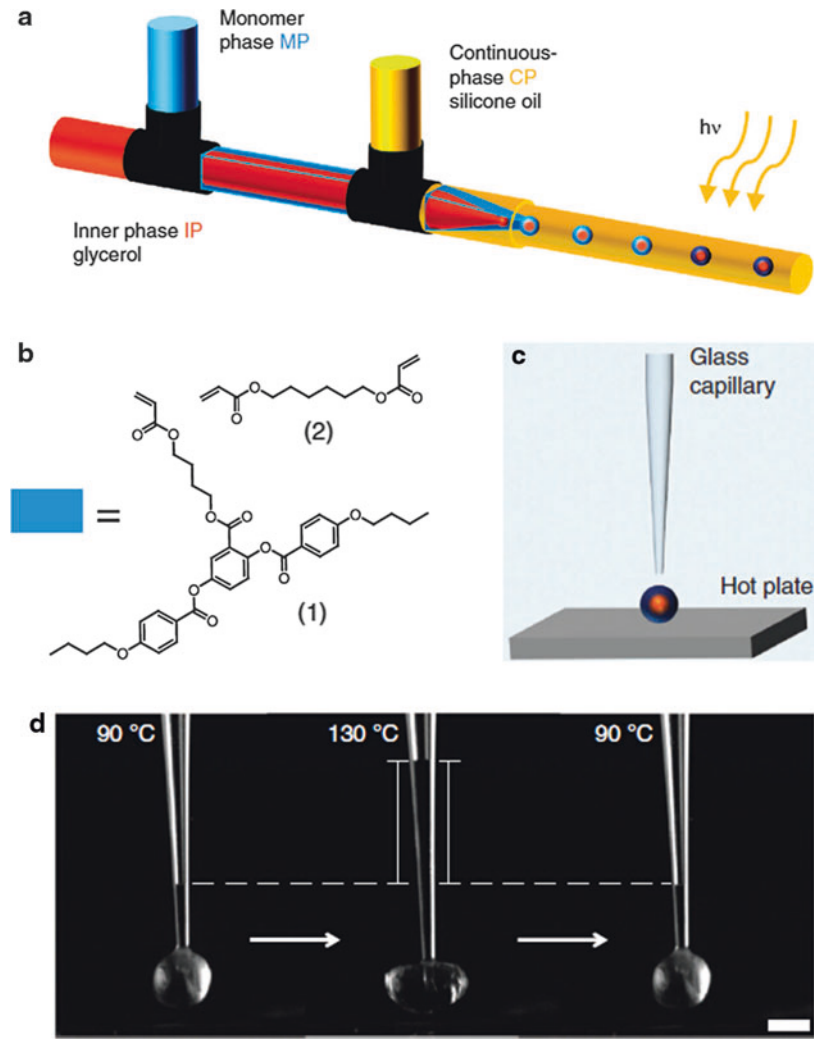
Organic Actuators, Fig. 6 (**a**) Thermal induced contraction of microsized CLCP pillars polymerized by VC4 and DVSA. (**b-d**) The shape of an isolated pillar with the diameter of 20 mm changes with temperature under polarizing optical microscope (from left to right, T =30, 96, and 48 °C) [13] (Copyright 2013 Wiley-VCH)

microsized CLCP pillars prepared by a soft lithography technique underwent a contraction when heated to T_{trans} (Fig. 6) [13]. Besides, Zentel et al. fabricated micropumps from CLCP core-shell particles via a microfluidic double-emulsion process. A liquid core which is wrapped into the LC shell is reversibly pumped into and out of the particle by actuation of the LC shell responsive to the change of temperature (Fig. 7) [14].

Real-world applications of the CLCPs are lacking because the mechanical properties of the CLCPs become unreliable over many actuation cycles and difficult to reprocess. Yan Ji and Eugene Terentjev et al. showed that the processing bottleneck of the CLCPs can be overcome by introducing exchangeable links in place of permanent network cross-links [15]. As Fig. 8a demonstrates, they used a transesterification reaction to generate an exchangeable CLCP. With the added catalyst, the fast breaking and reforming of ester bonds are above the activation barrier at a temperature T_v, that is, the designed xCLCPs can be reshaped or reprocessed at $T > T_v$. At temperatures below T_v, exchange reactions become slow exponentially, and the network acts as a classical covalently bonded thermoset. Reversible transesterification also allows an xCLCP film to bond with a non-LC epoxy elastomer from the reaction of the diglycidyl ether of bisphenol A and diacid above T_v (Fig. 8c). As Fig. 8d shows, as the temperature goes up, the height of the dome-shaped CLCP actuator decreases because of the contraction induced by the thermal phase transiton; on cooling back to the LC phase, the dome reappears in its original shape.

In addition, light-induced large contraction-elongation effects were achieved in the pioneering work by Finkelmann using azobenzene as a sensitive chromophore in a CLCP film [16]. The rod-like trans form of the azobenzenes stabilizes the phase structure of LC phase, whereas the bent cis isomer tends to destabilize the phase structure of mixture (Fig. 5b).

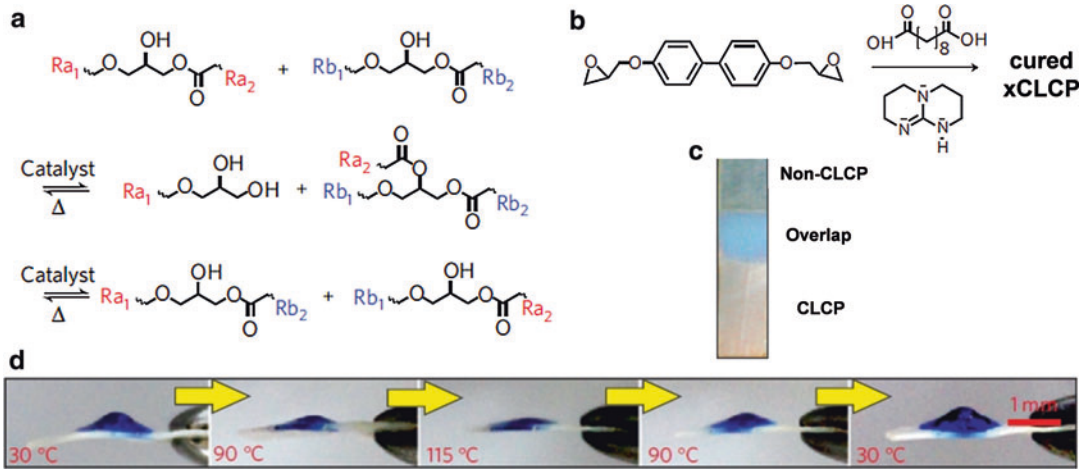
The most interesting property of an azobenzene-based CLCP film is the capability to perform three-dimensional movement, which could be advantageous for a variety of real manipulation applications. By means of the selective absorption of linearly polarized UV light in macroscopically randomly aligned CLCP films, Ikeda



Organic Actuators, Fig. 7 (**a**) The triple co-flowing geometry of the microfluidic set-up allows the fabrication of nematic shell particles filled with a glycerol core. Dispersed in the surrounding continuous-phase silicone oil, the monomer droplets are polymerized by ultraviolet light. (**b**) Chemical structures of the reactive mesogen (1) and crosslinker (2) in the monomer mixture. (**c**) Schematic drawing of the micromanipulator to penetrate the particle shell with a thin glass capillary. (**d**) When heated to the isotropic phase at 130 °C, the phase transition results in a deformation of the LCE shell. The force exerted by the actuating shell pushes the liquid glycerol core into the capillary. Once the particle resumes its former shape in the nematic phase at 90 °C, the glycerol sinks to its original level (scale bar, 100 mm) [14] (Copyright 2012 Nature Publishing Group)

et al. realized that the deformation direction of a polydomain CLCP film was chosen precisely along any direction, as Fig. 9 shows [17]. Because of the high concentration of the azobenzene moieties in the film and the high extinction coefficient of the azobenzene moieties at around 360 nm, most of the incident photons are absorbed by the film surface. Only the surface region facing the incident light undergoes the decrease in alignment order of LC and phase transition. As a result, the contraction is generated only in the surface layer, causing the films to bend toward the irradiation source.

Some interesting designs from the photoresponsive CLCPs present an opportunity to realize soft actuators in more applications. For example, Palffy-Muhoray and coworkers demonstrated that the CLCP sample doped with azobenzene dyes swam away from the laser beam, with an action resembling that of flatfish [5]; Ikeda et al. fabricated the first light-driven plastic motor with the use of CLCP/polyethylene laminated films [18]; Yu et al. utilized photoresponsive CLCPs as microvalves and micropumps [19, 20] and other relative applications, such as light-controlled quick switch of adhesion on a microarrayed CLCP superhydrophobic film and light-switchable behavior on the reflection spectra of a microarray of the azobenzene CLCP induced by the photodeformation [21, 22].



Organic Actuators, Fig. 8 (**a**) A scheme of reversible transesterification. (**b**) The synthesis of xLCE, showing the mesogen and the spacer used at 1:1 ratio, and the triazabicyclodecene catalyst. (**c**) An xCLCP film bonded with a non-liquid-crystal epoxy elastomer could be joined together at a sufficiently high temperature $T > T_v$. (**d**) Reversible cycle of a dome-shaped moulded CLCP actuator becoming completely flat on heating to the isotropic phase [15] (Copyright 2014 Nature Publishing Group)

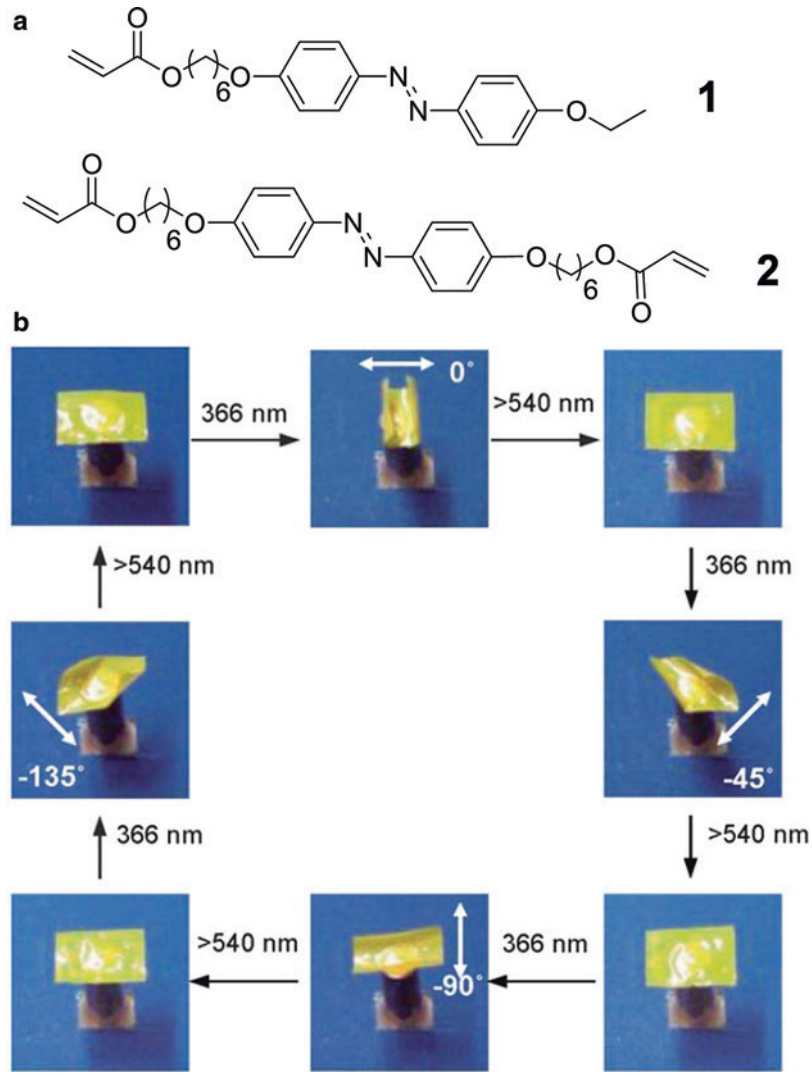
The photoinduced deformations mentioned above are interesting; however, UV light is harmful to our health. Yu et al. fist reported visible-light-induced bending and unbending of azotolane-containing CLCPs, whose deformation even occurred upon irradiation with sunlight [23]; then, they demonstrated a visible-light-driven microrobot, which can pick, lift, move, and place milligram-scale objects [24].

Furthermore, Yu et al. incorporated upconversion nanophosphors (UCNPs) into the azotolane-containing CLCP film and succeeded in generating fast bending of the resulting composite film exposed to continuous-wave (CW) NIR light at 980 nm. Under excitation with a CW 980 nm laser, the as-prepared UCNPs show blue emission, and the main upconversion lamination (UCL) emission peaks at 450 nm and 475 nm overlap the absorption band of the azotolane CLCP film (between 320 and 550 nm) perfectly; thus, the UCL light triggered molecular-scale motion (trans-cis photoisomerization of azotolane), which is converted into large macroscopic deformations (Fig. 10) [25].

Most recently, Yu et al. achieved red-light-controllable CLCP driven by low-power excited triplet-triplet annihilation-based upconversion luminescence (TTA-UCL) [26]. A red-to-blue TTA-based upconversion system was incorporated into rubbery polyurethane film and then assembled with an azotolane-containing CLCP film. The film bent toward the light source irradiated by 635 nm laser at low power of 200 mW cm^{-2} (Fig. 11). This work is a significant step forward to develop a novel red-light-controllable CLCP based on TTA-UCL.

Some other interesting movements of the CLCP films such as cilia-like motion and twisting have also been developed. As Fig. 12 shows, Oosten et al. used an azobenzene-containing film with a splay bent configuration which improved the bending strength of the material and utilized inkjet printing technology to create cilia-shaped structures [27]. The response of A3MA polymer cilia in water to light is shown in Fig. 12b. Two parts of the CLCPs incorporated with different azobenzene chromophores allow for a wavelength-selective control in the bending of the different segments. Thereby, the motion of natural cilia is mimicked.

Based on a combination of a nematic LC and a small amount of chiral dopant, Fletcher

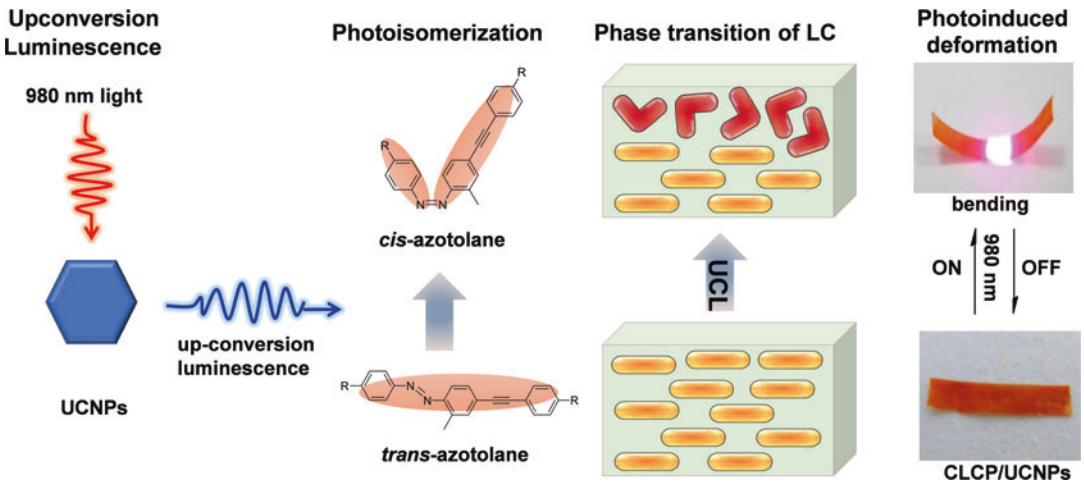


Organic Actuators, Fig. 9 (**a**) Chemical structure of the LC monomer (1) and crosslinker (2) used for preparation of the film. (**b**) Images of the film bending in different directions in response to linearly polarized UV light with different polarization directions at $\lambda = 366$ nm. The film is subsequently flattened again by visible light at $\lambda = 540$ nm. (Reprinted with permission from ref. [17] Copyright 2003 Nature Publishing Group)

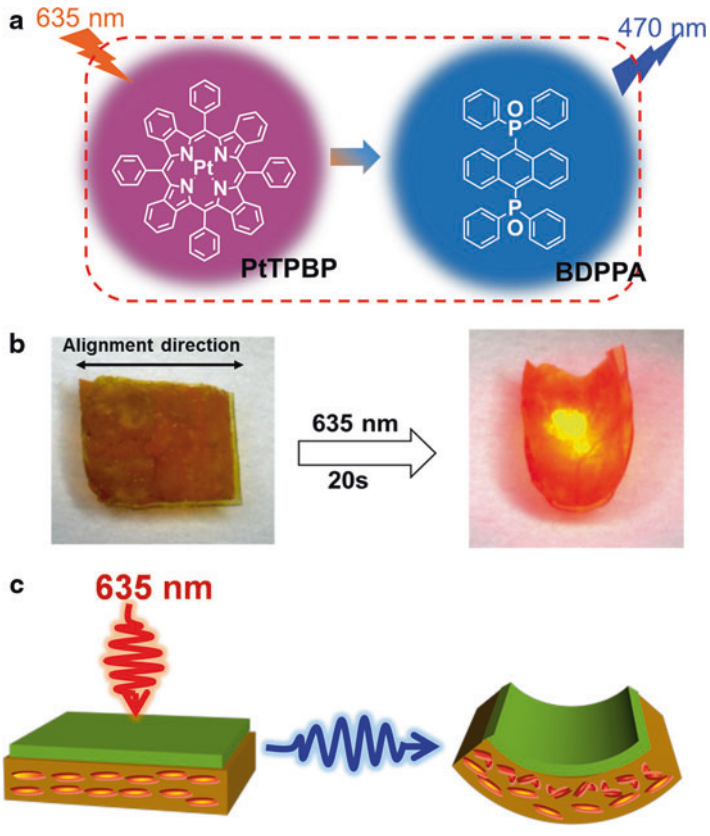
et al. succeeded in spring-like motion of the CLCP toward mimicking the mechanical behavior of plant tendrils [28]. As Fig. 13a shows, the orientation of the LC molecule changes smoothly by 90° from the bottom surface to the top surface. The helical shapes depend on the angle at which the ribbon is cut (Fig. 13b). With UV light irradiation, left-handed spiral ribbons B decreased in their macroscopic pitch, and the corresponding right-handed ribbon D is just the opposite (Fig. 13c). It is remarkable to observe inversion of the helical sense from right-handed to left-handed (ribbons E).

Electroactive Polymer Actuators

Electroactive polymers (EAPs) are polymers that exhibit a change in size or shape when stimulated with an electric field. Compared to their inorganic counterparts (e.g., shape-memory alloys, carbon nanotubes, and piezoelectric ceramics), the polymer-based EAP materials allow many attractive properties and unique characteristics including relatively inexpensive, lightweight, mechanically robust, easily processable, and commercially scalable. Further, they can be configured into almost any shape, and their properties

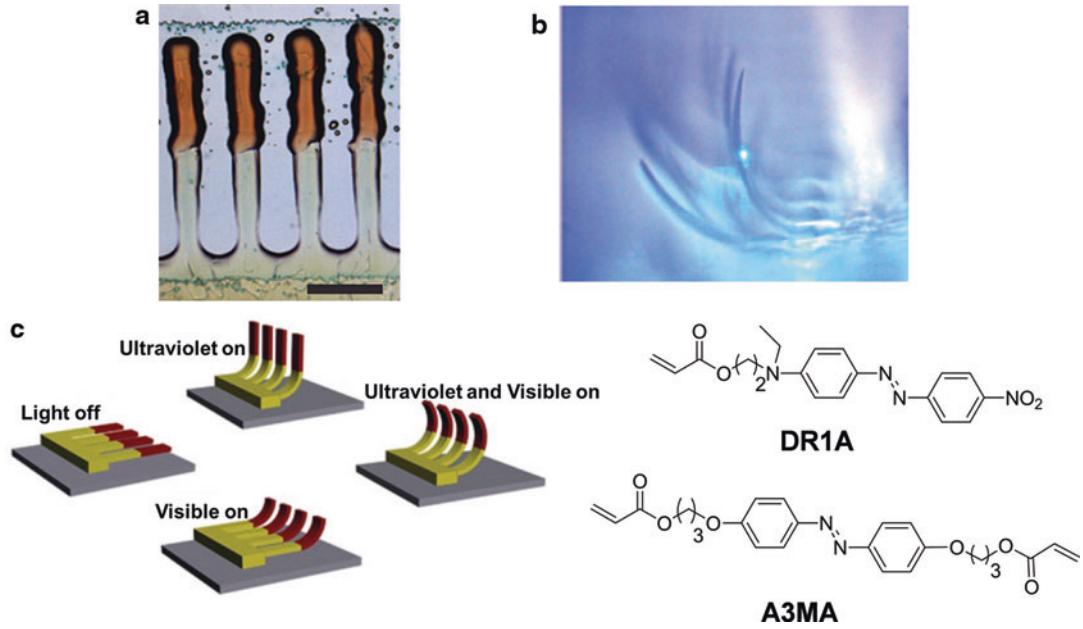


Organic Actuators, Fig. 10 Schematic illustration of the mechanism of CW NIR-lightinduced deformation of the azotolane CLCP/UCNP composite film, and photographic frames of the composite film bending in response to the NIR light at CW 980 nm and recover after removing the light source (Reprinted with permission from ref. [25]. Copyright 2012 Wiley-VCH)



Organic Actuators, Fig. 11 (**a**) Schematic illustration of TTA-UCL emission of PtTPBP (sensitizer) and BDPPA (annihilator) under excitation with 635 nm laser. (**b**) Red-light-induced deformation of the two layer assembly film. (**c**) Mechanism for the photoinduced deformation of the as-prepared assembly films (Reprinted with permission from ref. [26]. Copyright 2013 American Chemical Society)

O



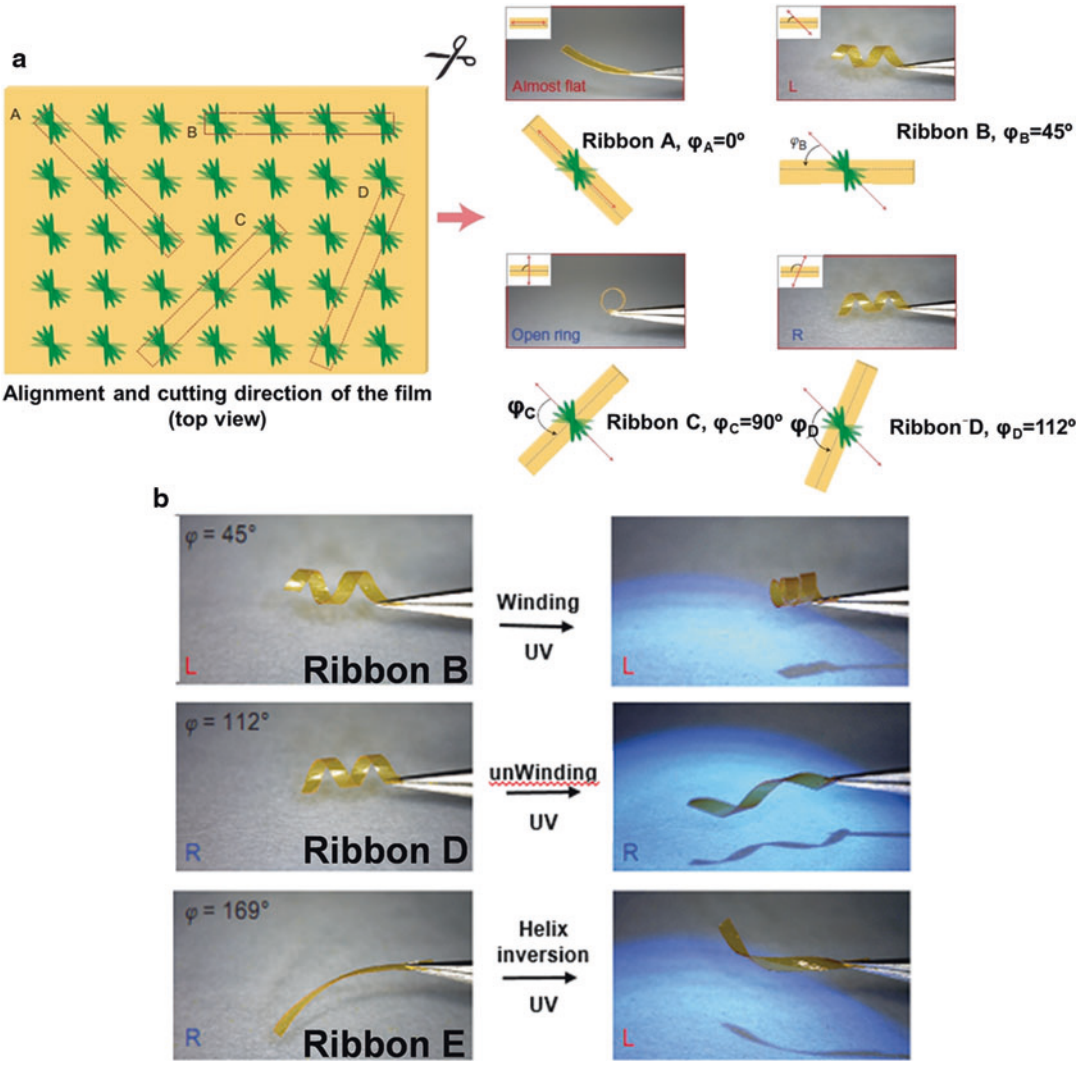
Organic Actuators, Fig. 12 (**a**) Arrays of cilia are made with two separate parts (scale bar 0.5 mm). (**b**) Side view of the actuation of A3MA polymer cilia with ultraviolet light in water. (**c**) Schematic illustration of an asymmetric motion of artificial cilia produced by controlling the wavelength of light. The sample is built with polymer containing azobenzene DR1A at the top (absorbing at < 390 nm) and polymer containing azobenzene A3MA at the bottom (absorbing in the 455–550 nm range) [27] (Copyright 2009 Nature Publishing Group)

can be tailored to suit a broad range of requirements. Impressive advances in improving the actuation strain capability of EAPs are attracting the attention of engineers and scientists from many different disciplines. EAP types are divided into two major groups based on their primary activation mechanisms: electronic electroactive polymers (driven by electric fields and coulomb forces) and ionic electroactive polymers (driven by the drifting or diffusion of ions).

The leading electronic EAP types include ferroelectric polymers, dielectric EAPs, and electrostrictive graft elastomers, which respond quite fast because of no diffusion of charge species. This type of EAPs can be made to hold the induced displacement while activated under a DC voltage without consuming electrical energy, making these EAPs highly efficient for robotic applications. These materials have a high mechanical energy density. However, generating a large deformation requires a high activation field of 100 V/μm or more, which may be close to the electric breakdown level. It is a large challenge to make the field-activated EAP materials applicable for commercial devices, because they are needed to produce large strain using dramatically lower voltages than currently needed [29].

Examples of ionic EAP materials include ionic polymer-metal composites (IPMCs), conductive polymers (CPs), gels, and carbon nanotubes. They undergo volume changes due to a reversible counterion insertion and expulsion during redox cycling. Depending on whether anion or cation exchange predominates, the active layer will either (1) expand upon oxidation due to anion insertion and contract upon reduction due to anion expulsion or (2) contract on oxidation due to cation expulsion and expand on reduction due to cation insertion. Actuators based on CPs can be electrically controlled at low operating voltages (typically 1–3 V), continuously switched between expanded/contracted states.

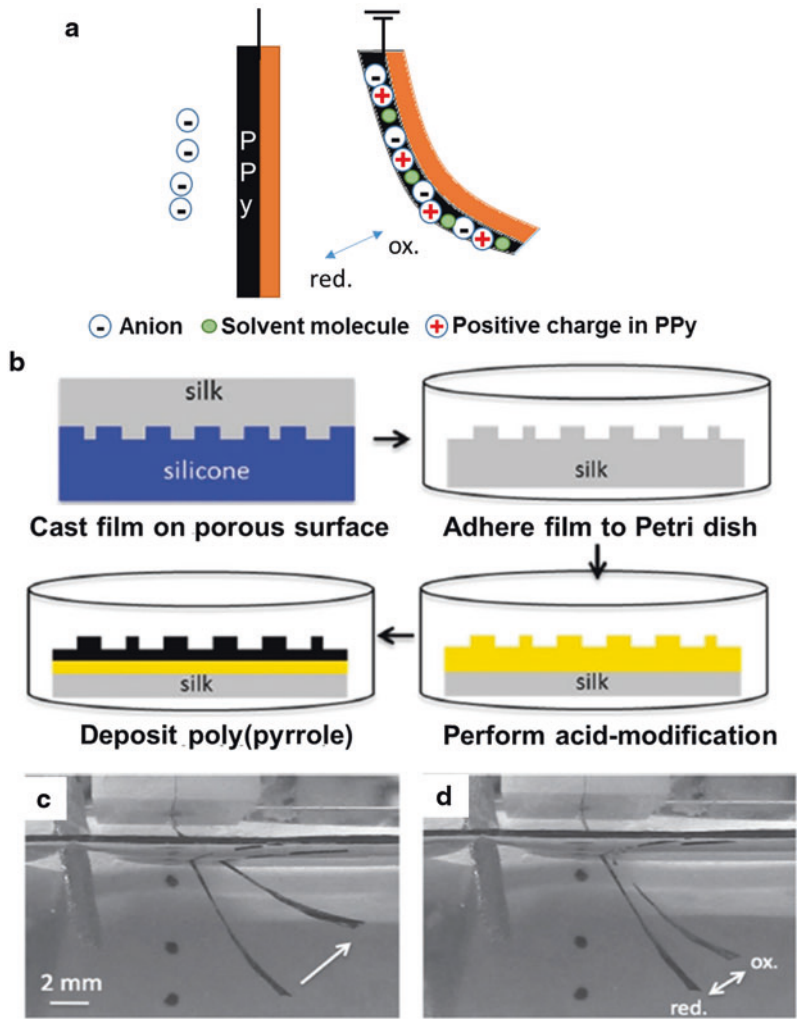
Murphy et al. fabricated a single-component, metal-free, biocompatible, and electromechanical actuator device using a conductive interpenetrating network of silk fibroin and poly



Organic Actuators, Fig. 13 (**a**) An illustration shows the orientation of the LC molecule changing smoothly by 90° from the bottom surface to the top surface. The ribbons display a variety of shapes that depend on the direction in which they are cut, from left-handed or right-handed ribbons to ribbons for which chirality is not expressed macroscopically. (**b**) Spiral ribbons irradiated for two minutes with ultraviolet light ($\lambda = 365$ nm) display isochoric winding, unwinding and helix inversion as dictated by their initial shape and geometry [28] (Copyright 2014 Nature Publishing Group)

(pyrrole) (PPy) [30]. As Fig. 14 shows, aqueous silk solutions were cast onto a silicone baking mat resulting in only one porous face. After acid modification, PPy is covalently attached to the porous silk surface on which it is deposited. PPy is completely integrated with the silk film making no delamination. For maximal actuation, a ± 3.0 V square wave was applied for 50 cycles at 1 Hz firstly aiming to a more consistent macroscale actuation behavior, then the fast-sequence ± 3.0 V square wave was applied to the films at two-minute intervals. During the initial 50-cycle curing sequence, films typically exhibited a net expansion in addition to the observed oscillation. During the second actuator cycle, the silk-PPy composites contracted under reduction as well as expanded under oxidation in PBS. Obviously, the network operates primarily by anion exchange.

O



Organic Actuators, Fig. 14 (**a**) Schematic diagram of PPy bilayer actuator expansion and contraction. (**b**) Illustration of the bilayer actuator film fabrication process. The free-end actuation behavior of a 3 × 10 mm silk-PPy film. (**c**) A ±3.0 V square wave was applied for 50 cycles at 1 Hz. (**d**) by 16 cycles at 1/120 Hz in PBS [30] (Copyright 2014 Wiley-VCH)

In this mechanism, counteranions and accompanying solvent molecules are expelled when the cationic PPy is reduced to its neutral form, resulting in a contracting motion. Upon oxidation, the anions are reinserted, resulting in expansion or stiffening of the film. This is the first demonstrations of a fully biocompatible, metal-free actuator device operated in a biologically relevant environment.

Future Direction for Research

The market potential for organic actuators is quite remarkable, as they have such advantages as flexibility, light weight, low cost, and quiet operation compliance. Many types of organic actuators have already been developed as artificial actuators in response to external stimuli such as change in temperature, light, and electric field. Not only two-dimensional but also three-dimensional motions in any direction have been achieved which are useful for the application as actuators. In the rapidly developing world of advanced materials, the application is determined by the material's performance and production costs. Therefore, to make the organic actuators applicable for commercial devices, further effort is required in tailoring the different material systems to specific demands of selected applications.

Many challenges and opportunities remain for using these smart materials for "real-life" applications, such as realization of low-cycle fatigue, biocompatibility, resistance in harshness of the operating conditions (extreme temperatures, radiation effects, and very high reliability and durability requirements) and design of complex motions.

Cross-References

- Biosensors
- Conduction Mechanisms in Organic Semiconductors
- Flexible Electronics
- Microfluidic Whole-Cell Biosensor
- NanoRobotics
- Organic Bioelectronics
- Organic Sensors
- Smart Hydrogels
- Whispering Gallery Mode Resonator Biosensors

References

1. Ikeda, T., Nakano, M., Yu, Y.L., Kanazawa, A.: Anisotropic bending and unbending behavior of azobenzene liquid-crystalline gels by light exposure. Adv. Mater. **13**, 201–205 (2003)
2. Yao, X., Dunn, S., Kim, P.: Fluorogel elastomers with tunable transparency, elasticity, shape-memory, and antifouling properties. Angew. Chem. Int. Ed. **53**, 4418–4422 (2014)
3. Sawa, Y., Ye, F.F., Urayama, K.: Shape selection of twist-nematic-elastomer ribbons. PNAS. **108**, 6364–6368 (2011)
4. Maeda, S., Hara, Y., Sakai, T.: Self-walking gel. Adv. Mater. **19**, 3480–3484 (2007)
5. Jaehwan, K., Jin-Han, J., Hyun-Jun, K.: Durable and water-floatable ionic polymer actuator with hydrophobic and asymmetrically laser-scribed reduced graphene oxide paper electrodes. ACS Nano. **8**, 2986–2997 (2014)
6. Wei, J., Yu, Y.L.: Photodeformable polymer gels and crosslinked liquid-crystalline polymers. Soft Matter. **8**, 8050–8059 (2012)
7. Islam, M.R., Li, X., Smyth, K., Serpe, M.J.: Polymer-based muscle expansion and contraction. Angew. Chem. Int. Ed. **52**, 10330–10333 (2013)
8. Takashima, Y., Hatanaka, S., Otsubo, M.: Expansion-contraction of photoresponsive artificial muscle regulated by host-guest interactions. Nat. Commun. **3**, 1270–1278 (2012)
9. Behl, B.M., Razzaq, M.Y., Lendlein, A.: Multifunctional shape-memory polymers. Adv. Mater. **22**, 3388–3410 (2010)
10. Xie, T.: Tunable polymer multi-shape memory effect. Nature **464**, 267–270 (2010)
11. Luo, Y.W., Gao, X., Xie, T.: A general approach towards thermoplastic multishape-memory polymers via sequence structure design. Adv. Mater. **25**, 743–748 (2013)
12. De Gennes, P.G.: Physique moleculaire. C. R. Acad. Sci. B. **281**, 101–103 (1975)
13. Wu, Z.L., Buguin, A., Keller, P.: Microstructured nematic liquid crystalline elastomer surfaces with switchable wetting properties. Adv. Funct. Mater. **23**, 3070–3076 (2013)
14. Fleischmann, E., Zentel, R.: One-piece micropumps from liquid crystalline core-shell particles. Nat. Commun. **3**, 1178–1186 (2012)
15. Pei, Z.Q., Terentjev, E.M., Ji, Y.: Mouldable liquid-crystalline elastomer actuators with exchangeable covalent bonds. Nat. Mater. **13**, 36–41 (2014)
16. Finkelmann, H., Nishikawa, E.: A new opto-mechanical effect in solids. Phys. Rev. Lett. **87**, 15501–15504 (2001)
17. Yu, Y.L., Nakano, M., Ikeda, T.: Directed bending of a polymer film by light. Nature **425**, 145–145 (2003)
18. Yamada, M., Yu, Y.L., Ikeda, T.: Photomobile polymer materials: towards light-driven plastic motors. Angew. Chem. Int. Ed. **47**, 4986–4988 (2008)
19. Chen, M.L., Zhu, Y.T., Yu, Y.L.: Photodeformable polymer material: towards light-driven micropump applications. Appl. Phys. A. **100**, 39–43 (2010)
20. Chen, M.L., Zhu, Y.T., Yu, Y.L.: Photodeformable CLCP material: study on photo-activated microvalve applications. Appl. Phys. A. **102**, 667–672 (2011)
21. Li, C., Cheng, F.T., Yu, Y.L.: Light-controlled quick switch of adhesion on a micro-arrayed liquid crystal polymer superhydrophobic film. Soft Matter. **8**, 3730–3733 (2012)
22. Yan, Z., Ji, X.M., Yu, Y.L.: Light-switchable behavior of a microarray of azobenzene liquid crystal polymer induced by photodeformation. Macromol. Rapid Commun. **33**, 1362–1367 (2012)
23. Yin, R.Y., Yu, Y.L., Ikeda, T.: Can sunlight driven the photoinduced bending of polymer films? J. Mater. Chem. **19**, 3141–3143 (2009)
24. Cheng, F.T., Yin, R.Y., Yu, Y.L.: Fully plastic microrobots which manipulate objects using only visible light. Soft Matter. **6**, 3447–3449 (2010)
25. Wu, W., Li, F.Y., Yu, Y.L.: NIR-light-induced deformation of cross-linked liquid-crystal polymers using upconversion nanophosphors. J. Am. Chem. Soc. **133**, 15810–15813 (2011)
26. Jiang, Z., Xu, M., Yu, Y.L.: Red-light controllable liquid crystal soft actuators via low-power excited upconversion based on triplet-triplet annihilation. J. Am. Chem. Soc. **135**, 16446–16453 (2013)

27. Van Oosten, C.L., Bastiaansen, C.W.M., Broer, D.J.: Printed artificial cilia from liquid-crystal network actuators modularly driven by light. Nat. Mater. **8**, 667–682 (2009)
28. Iamsaard, S., Fletcher, S.P., Katsonis, N.: Conversion of light into macroscopic helical motion. Nat. Chem. **6**, 229–235 (2014)
29. Bar-Cohen, Y., Zhang, Q.M.: Electroactive polymer actuators and sensors. MRS Bulletin. **33**, 173–180 (2008)
30. Romero, I.S., Bradshaw, N.P., Murphy, A.R.: Biocompatible electromechanical actuators composed of silk-conducting polymer composites. Adv. Fun. Mater. **24**, 3866–3873 (2014)

Organic Bioelectronics

Agneta Richter-Dahlfors[1] and Magnus Berggren[2]
[1]Department of Neuroscience, Swedish Medical Nanoscience Center, Karolinska Institutet, Stockholm, Sweden
[2]Department of Science and Technology, Linköpings University, Norrköping, Sweden

Synonyms

Organic bioelectronics (OBOE)

Definition

Active materials used for electrochemical devices are either organic or more commonly inorganic metal (−oxides). This entry only covers organic polymer materials and their application in life science research. Organic Bioelectronics [1] (OBOE) are carbon-based (therefore defined as organic) devices engineered from conductive polymers. The active parts of organic bioelectronic devices are usually composed of conjugated polymers, alone or in combination with other materials to form a primary or secondary interface with biological specimens (therefore defined as bio-). The conductivity of the polymer materials makes it possible to design devices that have the same functionality and capability as common electronics (therefore defined as -electronics). The devices can be manufactured to a multitude of different geometrical designs depending on applications targeted.

Overview and Background

Synthetic polymers revolutionized the technology of materials in the beginning of the twentieth century, when the first commercial plastic, bakelite found its use as a lightweight alternative to outmatch iron and steel in many applications. In 1977, Heeger, MacDiarmid, and Shirikawa achieved conductivity in polymers, a major breakthrough, for which they were awarded the Nobel Prize year 2000. This finding opened for the use of polymers in completely new areas such as in electronic industry and today a wide array of conducting polymers are synthesized, whose different levels of conductivity vary from almost insulators to good conductors. Besides the electronic properties of conducting polymers, the materials have also attracted much attention because of their unique electrochemical properties which are utilized in organic electrochemical transistors (OECT), polymer batteries, electrochromic displays, sensors, and actuators. The novelty of organic electronics results in part from the fact that these materials offer new or improved electroactive and opto-chromic features with respect to its inorganic counterparts as they are flexible, can be manufactured using printing tools, and for life science applications generally show a high level of biocompatibility and other desired features such as flexibility, optically transparent at visible wavelengths, and more.

Basic Methodology

The key to electrical conductivity in polymers are the alternating double bonds that build up the conjugated backbone. These bonds offer free electrons or holes in non-hybridized orbitals of carbon of the polymer and are thus able to participate in transporting electrical current. The electrical conductivity can be modulated by doping the polymer, that is, addition of another compound that

forms an ionic complex with the polymer chain that either donates or accepts one valence electron to further enhance the conductivity. Positive doping (p-doping) is achieved by the removal of one electron from the valence band while donating one electron to the valence band is termed negative doping (n-doping).

Since many conductive polymers can be doped and undoped by electrochemical reactions and hence change their intrinsic properties like conductivity, color/transparency, and surface properties, they are very well suited for active electrochemical circuits for biointerfaces.

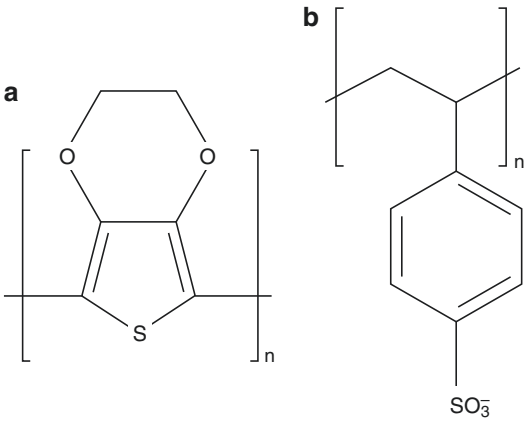
A useful property of many conductive polymers is the ability to conduct not only electrons but also ions, which is accompanied by a mass transport of matter in and out of the material. The one-dimensional characteristics of conjugated polymers provide coupled transport of both electronic and ionic charged species. Such co-existing charge transport can be exemplified by the conducting polymer poly (3,4-ethylenedioxythiophene) (PEDOT) doped with the charge-balanced counter-ion poly(styrene sulphonate) (PSS) which forms the polymer-polyectrolyte PEDOT:PSS (Fig. 1). The entire bulk of a PEDOT:PSS film can undergo electrochemical switching. This requires electronic transport along the conjugated backbone of PEDOT, while ionic transport occurs primarily in the PSS polyelectrolyte phase. PEDOT undergoes electrochemical switching according to the half reaction

$$\mathrm{PEDOT^{+}PSS^{-}} + \mathrm{M^{+}} + \mathrm{e^{-}} \Leftrightarrow \mathrm{PEDOT^{0}} + \mathrm{M^{+}} : \mathrm{PSS^{-}} \quad (1)$$

M^+ represents a monovalent cation and e^- an electron. As the reaction goes to the right, reduction of PEDOT to its neutral state occurs, while in the left direction oxidation of PEDOT takes place. Thus, switching of PEDOT between its neutral and oxidized state requires both cations and electrons to the reaction site, while the polymer chains themselves are effectively immobilized.

The possibility to electronically control the redox state of the conducting polymer material is a key feature in biointerfacing since it serves as a translator of biological signals (ions) and electronic signals. Examples of these applications are shown in the next section.

Mass transport of counter-ions in and out of the material during an electrochemical switch is often associated with a change in volume of the bulk material in part due to solvents (e.g., water) that follow the ions into the material. This can be utilized in actuators, and an example of a biomedical application is presented in the next section.

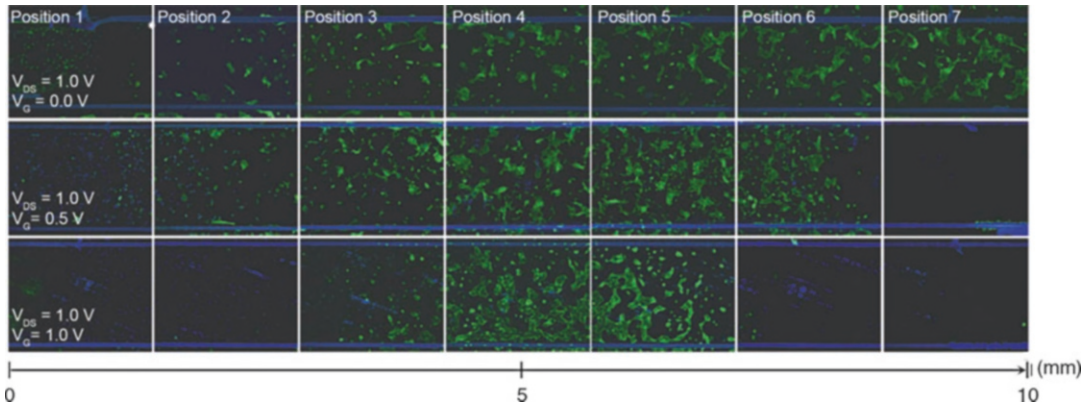


Organic Bioelectronics, Fig. 1 Conducting polymer poly(3,4-ethylenedioxythiophene) (*PEDOT*) (**a**) doped with the charge-balanced counter-ion poly(styrene sulphonate) (*PSS*) (**b**) which forms the polymer-polyelectrolyte PEDOT:PSS

Key Research Findings and Biomedical Applications

Organic Bioelectronics for Modulation of Cell: Surface Interactions

Previous research in the conducting polymer field has shown that these materials offer a possibility to reversibly alter the above-mentioned surface features through electronic control. Submerged in an electrolyte, conducting polymers undergo reduction or oxidation reactions, depending on the direction of the current applied. Using polypyrrole (PPy) as an electroactive cell culture substrate also revealed that electric addressing of the PPy surface promoted neurite outgrowth from cultured neuronal cells, and this effect was



Organic Bioelectronics, Fig. 2 Complex patterning of the extracellular matrix, cells, and tissues under in situ electronic control is the aim of the technique shown above. The distribution of MDCK epithelial cells along the channel of an organic electrochemical transistor is shown to be actively controlled by the gate and drain voltages, as electrochemical gradients are formed along the transistor channel when the device is biased; *top* = 0.0 V; *middle* = 0.5 V; *bottom* = 1.0 V; cells are visualized by F-actin staining (Reproduced with permission from Bolin [5])

correlated to an increase of fibronectin adsorbed to activated PPy surfaces [2].

The conducting polymer PEDOT has become increasingly popular in various bio-applications. In its doped oxidized state, it shows high conductivity and possesses great stability. The material shows high biocompatibility, and it performs very well in aqueous solutions over extended periods of time [3]. The electrochromism as well as other optical properties of the material, that is, optical transparency, lend itself very well to perform microscopy-based analysis, which is a popular tool in cell biology studies.

PEDOT doped with Tosylate (PEDOT:To) is one material that has been investigated to see whether its electronically regulated surface properties could be used to modulate adhesion and proliferation of the renal epithelial cells (Madin Darby canine kidney cell line, (MDCK)). Functionalized cell culture dishes, with conducting polymers coated onto the bottom, is one application, in which cells were subjected to two adjacent electrode areas. By applying low voltage to the electrodes, typically in the range of 0.5–1.5 V, the dish becomes an electrochemical cell with cell culture medium serving as the electrolyte. Electronic addressing of the dish causes redox reactions in the two electrodes: the negatively biased electrode becomes oxidized and the positively addressed electrode turns reduced. Svennersten et al. analyzed whether different redox states of PEDOT:To affected cell adhesion and subsequently cell proliferation. MDCK epithelial cells were shown to readily adhere to, spread, and proliferate on the reduced electrode, while the oxidized surface presented only few, unspecifically bound, apoptotic cells. The oxidized state of PEDOT:To induced a conformational change in surface-bound fibronectin, making the RGD domain, critical for integrin binding, inaccessible. As cells cannot form focal adhesion complexes, cell adhesion is abrogated. This study demonstrates how one physical entity, that is, the redox state of the conducting polymer, acts via several layers of complexity to indirectly regulate something as sophisticated as cell adhesion [4].

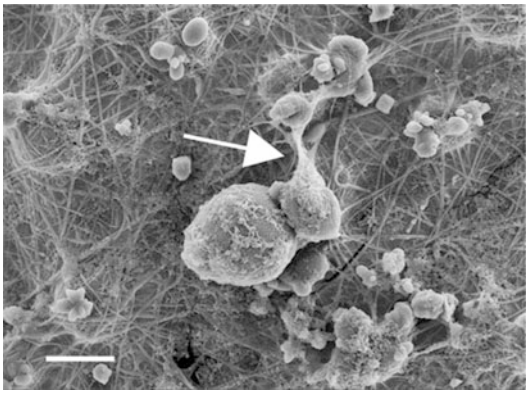
Cell density gradients and chemical gradients act in concert to direct a number of biological functions such as cell migration and morphogenesis. Modulation of the number of adhesion sites on biotic and abiotic surfaces is commonly used to produce cell gradients of different densities. Due to the inherent reversible surface characteristics of conducting polymers, this material has been explored as the basis of the establishment of dynamic gradients [5] (Fig. 2). Organic electrochemical transistors (OECT), a variant of the

organic thin film transistor (OTFT), comprise a channel which has an electrochemical potential gradient. Based on PEDOT:To, a lateral design of the OECT was generated, which had an exposed channel surface [6]. The electrochromic properties of PEDOT can be utilized to image the potential along the channel. The device properties and the functionality of the OECT can be even further optimized, by altering the ratio between the gate versus channel area, as demonstrated by Malliaras and coworkers [7].

While this transistor design establishes a close to exponential gradient of the redox potential and accordingly the cell density, linear redox gradients can also be produced. When PEDOT:To is coated on top of an indium tin oxide (ITO) film, the redox state of the polymer followed the linear potential gradient of the ITO film, and this was also reflected in the cell density. As it appears that linear as well as exponential gradients can be created using the same conducting polymer, it may suggest that these devices will become useful tools when deciphering the role of the vast array of linear and exponential gradients found in biological systems [8].

Organic Bioelectronics for Active 3D Matrix Environments for Cell Interaction

3D scaffold electrodes are considered to offer relatively lower impedance characteristics for electrical interfacing for cultured cells by mimicking the structure of natural extracellular matrix. The structural features of these scaffolds are in the nanometer scale, thus allowing nutrients and small molecules to freely diffuse through the matrix, while cells cannot move without an effort. Much effort is accordingly invested in engineering-suitable 3D matrices for cell cultures. Dependent on the nature of the material, scaffolds are produced by different methods. Nano-scaled fibrous scaffolds are commonly fabricated by self-assembly, phase separation, and electro-spinning. Several studies, mostly concerning neuronal cells and brain tissues, demonstrate the suitability of integrating conducting polymers in synthetic scaffolds. As the polymer's conducting property adds an active function to the nano-scaffolds, a potential can be delivered to depolarize excitable cells.



Organic Bioelectronics, Fig. 3 Scanning electron micrograph of a SH-SY5Y neuroblastoma cell on PEDOT:To coated PET nanofibers. Neurite indicated by *arrow*. Scale bar = 10 mm (Figure reproduced from Bolin [10])

The efficient charge transfer is due to the high surface-to-volume ratio of conducting polymer-coated nano-scaffolds [9].

By depositing the conducting polymer PPy onto nano-fibers made from PLGA (poly (lactic-co-glycolic acid)), these electrically active nano-fibers were found to promote the outgrowth of neurite extensions. Similarly, depolarization of neuroblastoma cells, cultivated on PEDOT:To deposited onto PET (poly-ethylene-traphtalate) nanofibers (Fig. 3), resulted in the activation of voltage-operated Ca^{2+} channels, leading to an increase of the intracellular $[Ca^{2+}]$ [10].

Biological material can also be used as a cast for conducting polymers. This opens for a possibility to build a conducting matrix in situ, within the tissue. Martin and co-workers demonstrated this concept by electropolymerizing PEDOT around neuronal cells to maximize the interaction with the cell [11].

This shows that electroactive nano-fibers as well as direct casting of conductive polymers can work as tight electronic interfaces and may thus be useful as a novel tool for cell biology studies in vitro as well as in vivo, that is, as an interfacing component in neural prosthetics.

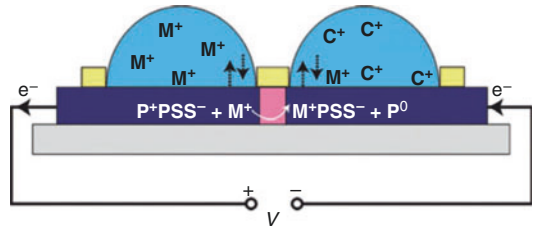
Organic bioelectronic actuators take advantage of the volume change and swelling of the polymer bulk caused by the electrochemical switch. By fixing the polymer to an electrochemically inert

material, the device will stretch, bend, or extend depending on its geometrical design. This feature can be used in biomedical applications and has been shown for micromanipulation of single cells, for example [12].

Organic Bioelectronics for Transport and Delivery of Bio-substances In Vitro and In Vivo

The use of conductive polymers in drug delivery devices has also been demonstrated. By applying a potential, substances such as neurotrophins can be released from delivery electrodes fabricated from conducting polymers [13]. Similarly, PEDOT and PPy-coated nanostructures fabricated on microelectrodes have been used to obtain spatially defined release of the anti-inflammatory drug dexamethasone [14]. Major challenges of electrochemically active drug release polymer electrodes is, however, the difficulty to achieve a proper concentration of the released compound, which preferentially should be in the same range as the physiologically relevant condition, as well as to achieve a device in which the delivery can be switched on and off.

Utilizing the ionic conductivity of conjugated polymers (explained in the previous section) true electronic control over the transport and release of charged molecules can be achieved. Recently, the "organic electronic ion pump" (OEIP) was developed to provide a device for lateral transport and electronically controlled delivery of ions and other positively charged ions and biomolecules [15]. The OEIP is produced from a single film of the conductive polymer PEDOT:PSS patterned into two or more electrodes joined by a "transport channel." Over-oxidation of the PEDOT:PSS renders the channel electronically insulating but ionically conducting. A hydrophobic resist covers the channel and provides openings to position the electrolytes on top of the electrode areas. By applying a voltage between the electrodes an electrochemical circuit is established. Oxidation of PEDOT occurs at the anode/source whereas reduction occurs at the cathode/target (Fig. 4). Migration of ions in the channel resembles the transport of charged molecules, such as DNA and proteins, in conventional gel electrophoresis

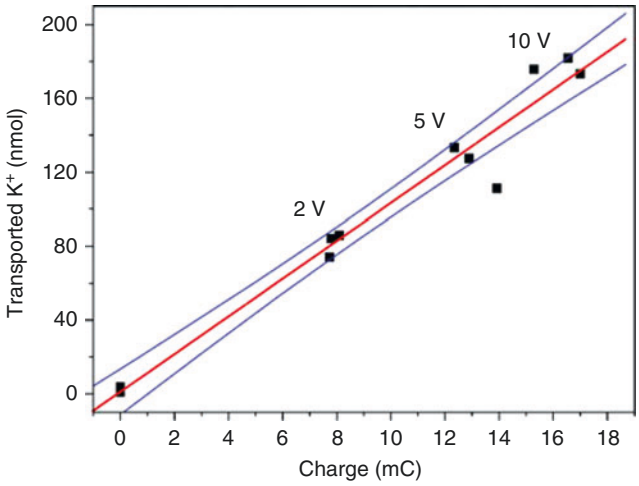


Organic Bioelectronics, Fig. 4 Schematic cross-section of the OEIP. The cation M^+ is present in a relatively high concentration in the source electrolyte (*left*) and binds to PSS in the polymer film. When *V* is applied, PEDOT (denoted P) in the source electrode (*left*) is oxidized. The positive charge on the polymer chain is neutralized with PSS, thereby releasing one M^+. Simultaneously, the target electrode (*right*) receives an electron, and PEDOT:PSS on this side is consequently reduced. There is a net flow of M^+ from the left to the right and every transported (monovalent) ion ideally corresponds to one charge measured between source and target. As the polymer film is very thin (200 nm), ion exchange will immediately occur between the target electrode and the target electrolyte, which also contains other cations (C^+). Hence, the concentration of M^+ increases in the target electrolyte as soon as voltage is applied and as long as source and target electrodes can be oxidized and reduced, respectively (Figure reproduced from Isaksson [15])

systems. In gel electrophoresis, the motion of charged species through an electrically charged field in the gel matrix separates molecules according to their charge and size. In analogy to this, the OEIP offers electrophoretic transport and delivery of positively charged ions and bio-molecules in the absence of a liquid flow and pressure.

The powerful features of the OEIP can be utilized in many biological applications; the most natural is in cell signaling research, where researchers are challenged by the complexity of cellular signaling patterns, both to study them but also to try to mimic them.

The basic functions of the OEIP were first verified by transport of the metal ion K^+. When the device is switched "ON," using solutions of K^+ as source and Na^+ as target electrolyte, K^+ is transported through the channel and delivered into the target electrolyte. The OEIP is connected to a source meter monitoring the current in the electronic branch of the circuit throughout the experiment. By calculating the total charge transported



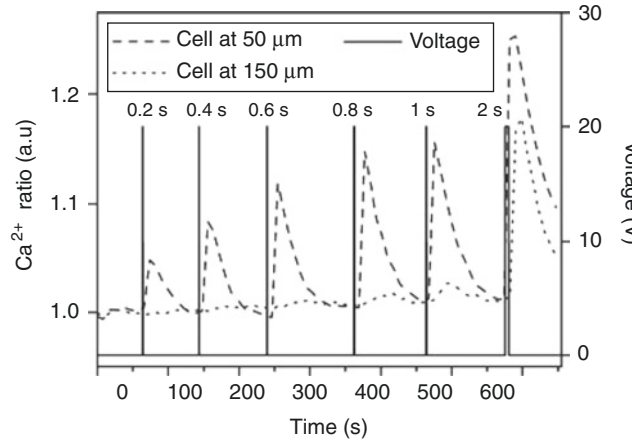
Organic Bioelectronics, Fig. 5 Relationship between K^+ delivered to the target system after 10 min and the total charge measured in the electronic circuit (i.e., integrated current) for driving voltages set to 2 V, 5 V, and 10 V. The middle line shows the linear fit and the two outer lines represent the 95 % confidence interval (Figure reproduced from Isaksson [15])

from the source to the target electrode, that is, the integrated current and comparing it to the amount of K^+ delivered to the target a linear relationship can be established between the amount of material delivered and the total charge. For K^+ the electron-to-ion ratio equals 1, that is, delivery efficiency is 100 %, indicating that for each electronic charge measured in the circuit one K^+ is delivered to the target (Fig. 5). Capacity and efficiency of the OEIP is in part dictated by the current bias but also its dimensions. These criteria allow fine-tuning of substance delivery. The OEIP polymer channel was miniaturized from 4 mm down to 10 μm for fine-tuning purposes but also to fit the dimensions of a single cell. The miniaturization of the channel to sizes compatible to that of individual cells has also the advantage that it allows high spatial and temporal precision of single cell stimulation. This precise delivery mechanism was shown by triggering single neuroblastoma cells with the quaternary ammonium cation acetylcholine (ACh), a small molecular neurotransmitter found in both the peripheral and central nervous system [16].

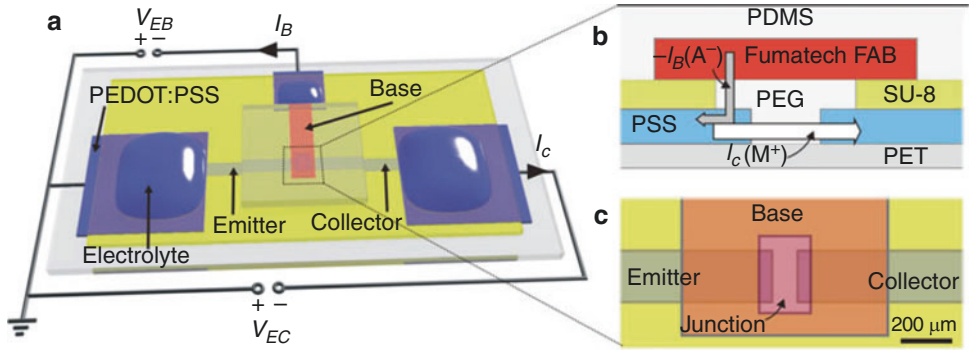
The Ca^{2+} ion is one of the most versatile signaling entities in the cell. Research over the last decades has revealed that cells use repetitive signals, known as Ca^{2+} oscillations, when information must be relayed over longer time periods. The switchable nature of ion delivery in the OEIP provides a novel tool to decipher the role of the frequency and amplitude components of the Ca^{2+} oscillations. Due to the electronic control of the OEIP, switching the device ON/OFF will result in a similar ON/OFF behavior of the electrophoretic delivery. This strategy can accordingly be used to establish temporally device-induced, controlled oscillations. Figure 6 shows a typical example of OEIP-induced Ca^{2+} oscillation in cells. By pulsing the delivery of ACh using the 10 μm-channel OEIP, a repetitive Ca^{2+} response was induced in single neuroblastoma cells (Fig. 6). The strength of the Ca^{2+} response can also be modulated by changing the pulse-length of the electronic induction of the transport. Control of dynamic parameters, such as frequency and amplitude, is not limited to ACh, but has also been demonstrated using H^+.

To further increase the fine-tuning and also ease the integration of multiple circuits for molecular transport an ion bipolar junction transistor (IBJT) has been developed [17] (Fig. 7). This is the ionic analog to conventional semiconductor bipolar junction transistors and shows that amplification, addressing, and processing also can be achieved in more complex bioelectronic bio-molecular circuits.

Abnormal protein aggregates, that is, amyloid fibrils is a common marker for many diseases. Utilizing the H^+ transport of the OEIP, the device has been shown to induce spatially controlled formation of amyloid fibrils, a unique way of producing these in vitro [18].



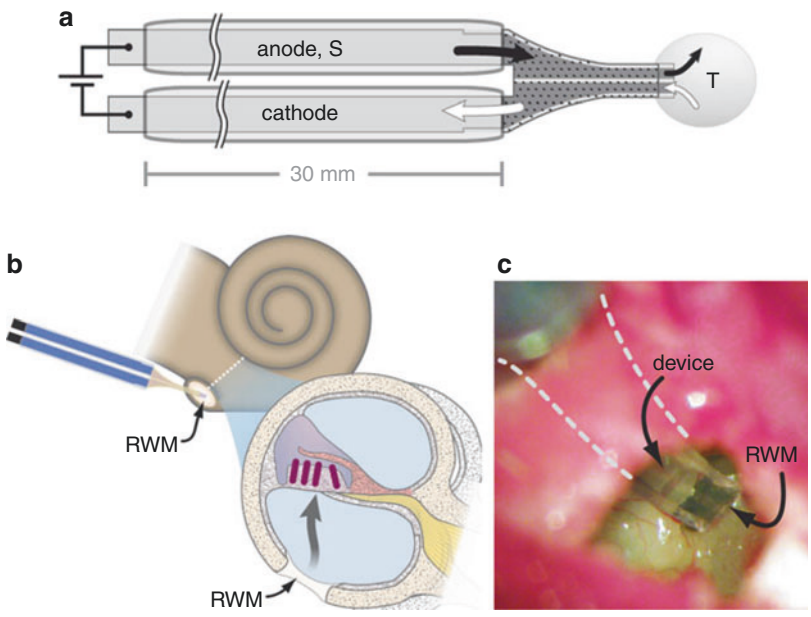
Organic Bioelectronics, Fig. 6 Ca^{2+} responses in SH-SY5Y cells after delivery of ACh from the outlet of the 10 µm-OEIP. Indicated times represent the duration of the pulse (Reproduced with permission from Tybrandt [16])



Organic Bioelectronics, Fig. 7 (**a**) The IBJT biased according to the common-emitter configuration. PEDOT:PSS electrodes with aqueous electrolytes generate the ionic currents within the device. Emitter and collector (cation selective) meet the base (anion selective) defining the junction. (**b**) Vertical cross-section along the emitter-collector configuration (not to scale). The emitter and collector (~250 nm thick PSS) are patterned on *top* of a PET substrate. A cross-linked PEG gel (referred to as the junction) is employed into the opening of the 10 µm thick SU-8 insulating layer. The base (~20 µm thick) contacts the gel and is covered by a PDMS seal layer. *IC* is the flow of cations (M^+) from the emitter to the collector, while *IB*, in steady state, primarily consists of anions (A−) migrating from the base to the emitter. (**c**) Top-down view of the IBJT. Emitter and collector are horizontally separated by 100 µm and the base covers the whole junction area (Figure reproduced from Tybrandt [17])

Stimulating neuronal activity in vivo is generally performed using neural probes that generate electrical signals. Compared to chemical messengers, electrical signals cannot discriminate among the target cells, as all excitable cells in the stimulated area will be affected. A technique allowing local stimulation using a specific chemical compound is highly desirable, and this need has sparked great activities in the area of drug delivery systems. Delivery systems are often classified as "active" or "passive."

A typical example of an active system is microfluidics. These systems are liquid phase–based, generally small, programmable, and self-contained devices for sustained, local release. Hydrogels, on the other hand, represents a passive system, made of cross-linked water-soluble polymers. These systems are commonly used in clinical practice for wound-healing and tissue regeneration. The above-mentioned techniques have been applied in inner ear diseases and disorders, where local delivery has been a



Organic Bioelectronics, Fig. 8 (**a**) Top view of the syringe-like OEIP. *S* source electrolyte, *T* target electrolyte. (**b**) Cartoon depicting the experimental setup in vivo. *Dotted line* indicates a slice through the cochlea, expanded below. *Gray arrow* = diffusion of ions through the RWM. (**c**) Device mounted on the RWM in a guinea pig. The two ion channels are visible as dark stripes on the transparent substrate (Figures reproduced from Simon [19])

long-standing challenge. The cochlea is a complex structure of the inner ear that converts sound into neural activity. The coiled liquid-filled tube includes a membrane stretched across the middle of the tube and covered with sensory hair cells detecting sound using their mechanotransduction apparatus. Because these elements are exceptionally force- and flow-sensitive, and since the cochlea contains only a limited volume of liquid (so-called perilymph), it makes local delivery to this site exceptionally challenging. As substances are transported in the OEIP via electrophoretic transport rather than liquid flow, this device is very attractive to use for local delivery in vivo in general and the cochlea in particular. The inner hair cells utilize glutamate as the primary neurotransmitter and were thus used in a proof-of-concept study where in vivo OEIP–transported Glu modulated the hearing in a guinea pig model [19] (Fig. 8). This study shows that OBOE can be used as an interface for translating electrical signals into chemical messengers that specifically target one type of neuron, thereby modulating mammalian brainstem responses.

The previous section covered the possibility of organic bioelectronic devices to create gradients of surface interfaces. To form controlled gradients of bio-substances within a liquid, that is cell culture medium or other biological liquid interfaces, there is yet another challenge. The non-convective molecular delivery of the OEIP is a unique advantage in this instance, as a bio-substance is transported to the outlet of the OEIP, it spreads by diffusion into the target electrolyte. The migration through the polymer channel occurs at a constant rate, and the applied voltage can then be used to balance the delivery rate to the diffusion characteristics for the specific bio-substance. Using computational analysis, a model of the local concentrations of ions evolving from the channel outlet can be predicted. This has been shown in vitro by creating neurotransmitter gradients of ACh that at different spatial positions of the neuronal cell results in a response or no response [16].

Future Directions for Research

Organic bioelectronics are shown to be powerful tools for both in vitro as well as in vivo applications within life science. At the same time, further development and several areas of improvements are being investigated to increase versatility and performance. On the technology side, development of more advanced components that makes

the electrochemical circuits of OBOEs as complex as their traditional electronic counterparts and allows for more integrated devices with multifunctionality can be foreseen. The recently developed IBJT is a distinct example that in the future will allow for matrix addressing of biological specimens, for example, cells. Adding complexity to the devices also include integration of already existing organic electronic innovations, for example, polymer batteries to make the organic bioelectronic circuit self-supporting or adding in sensors to allow for automation and feedback regulation. Improvements in materials and manufacturing will also yield further miniaturization of the OBOE devices and circuits to even better match the sizes and dimensions encountered in biology. This will allow for the expansion of both in vitro and in vivo applications, where OBOE devices will be able to target and modulate endogenous signal-transduction pathways rather than rely on chemical drugs in drug-delivery applications.

Cross-References

- ▶ Cell Adhesion
- ▶ Doping in Organic Semiconductors
- ▶ Nanomedicine
- ▶ Organic Actuators
- ▶ Organic Sensors

References

1. Berggren, M., Richter-Dahlfors, A.: Organic bioelectronics. Adv. Mater. **19**, 3201–3213 (2007)
2. Schmidt, C.E., Shastri, V.R., Vacanti, J.P., Langer, R.: Stimulation of neurite outgrowth using an electrically conducting polymer. Proc. Natl. Acad. Sci. U. S. A. **94**, 8948–8953 (1997)
3. Asplund, M., et al.: Toxicity evaluation of PEDOT/biomolecular composites intended for neural communication electrodes. Biomed. Mater. **4**, 045009 (2009)
4. Svennersten, K., Bolin, M.H., Jager, E.W., Berggren, M., Richter-Dahlfors, A.: Electrochemical modulation of epithelia formation using conducting polymers. Biomaterials **30**, 6257–6264 (2009)
5. Bolin, M.H., et al.: Active control of epithelial cell-density gradients grown along the channel of an organic electrochemical transistor. Adv. Mater. **21**, 4379–4382 (2009)
6. Nilsson, D., et al.: Bi-stable and dynamic current modulation in electrochemical organic transistors. Adv. Mater. **14**, 51–54 (2002)
7. Cicoira, F., et al.: Influence of device geometry on sensor characteristics of planar organic electrochemical transistors. Adv. Mater. **22**, 1012–1016 (2009)
8. Kholodenko, B.N., Hancock, J.F., Kolch, W.: Signalling ballet in space and time. Nat. Rev. Mol. Cell Biol. **11**, 414–426 (2010)
9. Green, R.A., Lovell, N.H., Wallace, G.G., Poole-Warren, L.A.: Conducting polymers for neural interfaces: challenges in developing an effective long-term implant. Biomaterials **29**, 3393–3399 (2008)
10. Bolin, M.H., et al.: Nano-fiber scaffold electrodes based on PEDOT for cell stimulation. Sensor. Actuator. B Chem. **142**, 451–456 (2009)
11. Richardson-Burns, S.M., et al.: Polymerization of the conducting polymer poly (3,4-ethylenedioxythiophene) (PEDOT) around living neural cells. Biomaterials **28**, 1539–1552 (2007)
12. Smela, E.: Conjugated polymer actuators for biomedical applications. Adv. Mater. **15**, 481–494 (2003)
13. Richardson, R.T., et al.: Polypyrrole-coated electrodes for the delivery of charge and neurotrophins to cochlear neurons. Biomaterials **30**, 2614–2624 (2009). doi:10.1016/j.biomaterials.2009.01.015. S0142-9612(09)00022-2 [pii]
14. Abidjan, M.R., Kim, D.H., Martin, D.C.: Conducting-polymer nanotubes for controlled drug release. Adv. Mater. **18**, 405–409 (2006). doi:10.1002/adma.200501726
15. Isaksson, J., et al.: Electronic control of Ca^{2+} signalling in neuronal cells using an organic electronic ion pump. Nat. Mater. **6**, 673–679 (2007). doi:10.1038/nmatl963. nmatl963 [pii]
16. Tybrandt, K., et al.: Translating electronic currents to precise acetylcholine -induced neuronal signaling using an organic electrophoretic delivery device. Adv. Mater. **21**, 4442–4446 (2009). doi:10.1002/adma.200900187
17. Tybrandt, K., Larsson, K.C., Richter-Dahlfors, A., Berggren, M.: Ion bipolar junction transistors. Proc. Natl. Acad. Sci. U. S. A. **107**(22), 9929–9932 (2010). doi:10.1073/pnas.0913911107. 0913911107 [pii]
18. Gabrielsson, E.O., Tybrandt, K., Hammarstrom, P., Berggren, M., Nilsson, K.P.: Spatially controlled amyloid reactions using organic electronics. Small **6**, 2153–2161 (2010). doi:10.1002/smll.201001157
19. Simon, D.T., et al.: Organic electronics for precise delivery of neurotransmitters to modulate mammalian sensory function. Nat. Mater. **8**, 742–746 (2009). doi:10.1038/nmat2494. nmat2494 [pii]

Organic Bioelectronics (OBOE)

▶ Organic Bioelectronics

Organic Films

► Self-Assembled Monolayers for Nanotribology

Organic Photovoltaics: Basic Concepts and Device Physics

Tingting Xu[1] and Qiquan Qiao[2]
[1]Department of Applied Chemistry, School of Science, Northwestern Polytechnical University, Xi'an, Shaanxi, People's Republic of China
[2]Department of Electrical Engineering and Computer Science, South Dakota State University, Brookings, SD, USA

Synonyms

Organic solar cells

Definition

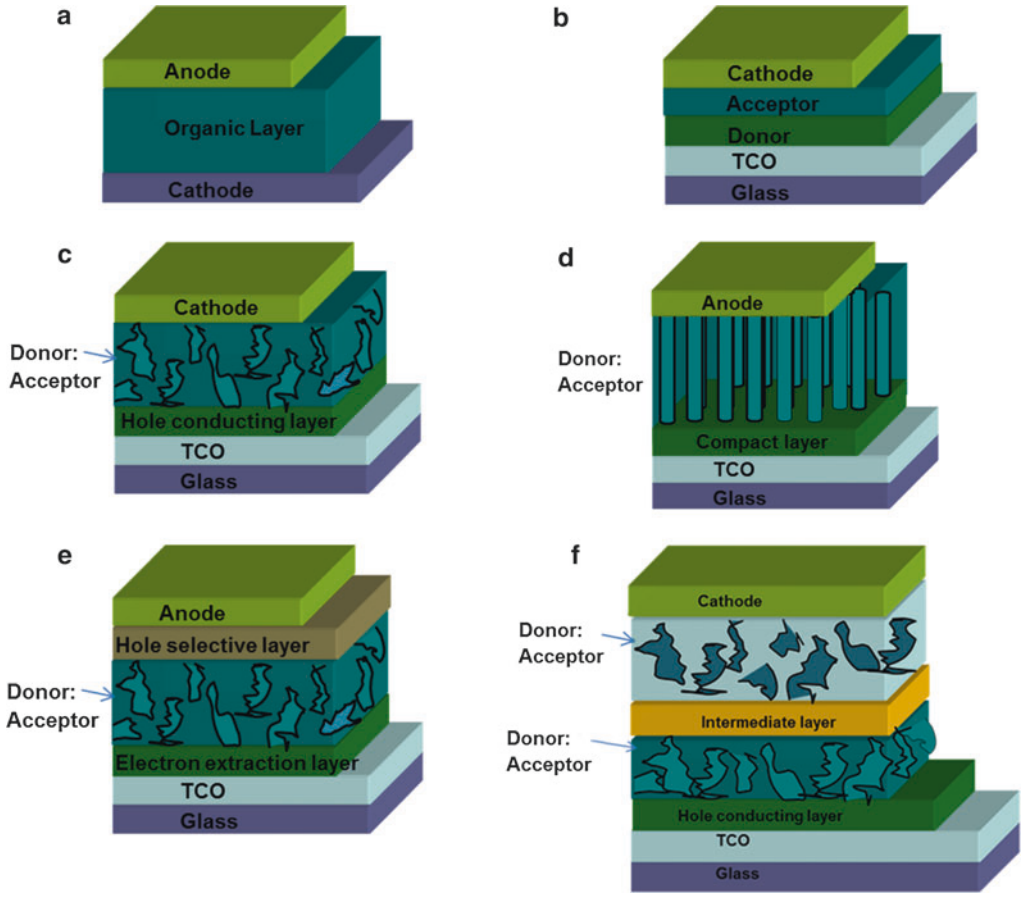
Organic solar cell (OSC) is a type of device made up of thin films of an organic semiconductor material such as conjugated polymer or small molecule as a donor blended with an acceptor material. The donor harvests sunlight. Upon light absorption, exciton (an electron–hole pair) is generated and dissociated at the donor–acceptor interface. The excited electron is transferred to the acceptor and then collected by the cathode, while the hole is transported to the anode. The acceptor materials can be either carbon based as fullerene derivatives (PCBM) or inorganic nanocrystals (e.g., CdSe, TiO_2, ZnO). An efficient organic solar cell requires not only a large interfacial area between the donor and acceptor but also a bicontinuous interpenetrating network that provides one continuous channel for electron transport and another consecutive pathway for hole transfer.

Overview

Organic solar cell (OSC) has gained extensive research interests attributing to its simple fabrication process, abundant raw materials as well as mechanical flexibility and being lightweight. Organic solar cell is a promising cost-effective alternative for the inorganic semiconductor counterparts. Herein, organic solar cell configurations, organic donor and acceptor materials, basic concepts in photovoltaic parameters, and device operation physics are introduced. Organic solar cell can be classified as single-layer, bilayer, bulk-heterojunction (BHJ), ordered-heterojunction, inverted, and tandem configurations. Conjugated polymers such as poly(3-hexylthiophene) (P3HT) and poly(p-phenylene vinylene) (PPV) derivatives are typically used as donor materials for harvesting sunlight; fullerene derivatives PCBM and inorganic semiconductor materials such as CdSe, TiO_2, and ZnO with various nanostructured morphologies are applied as acceptor materials. Short-circuit current density (J_{sc}), open-circuit voltage (V_{oc}), and fill factor (FF) are the key parameters to determine solar cell power conversion efficiency (PCE, η). Six main steps including (i) photon absorption (η_a), (ii) exciton generation (η_{ex}), (iii) exciton diffusion (η_{diff}), (iv) exciton and polaron dissociation (η_{ed}), (v) charge transport (η_{tr}), and (vi) charge collection (η_{cc}) govern solar cell performance. The maximized performance of each step would lead to the highest available efficiency. Further improvement in solar cell efficiency needs to explore novel donor and acceptor materials and new device architectures, as well as the fundamental understanding of the operation mechanism of organic solar cells.

Background

Limited supply and adverse environmental effects from the combustion of fossil fuels have driven people to search for clean and sustainable energy. Solar is an abundant and renewable energy source. Sunlight harvesting and converting it directly into electricity are known as photovoltaics, which has been regarded as one of the



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 1 Configuration of polymer solar cells: (**a**) single-layer structure, (**b**) bilayer heterojunctions, (**c**) bulk heterojunctions (BHJ), (**d**) ordered heterojunctions, (**e**) inverted structure, and (**f**) tandem structure

most promising technologies that can be an alternative to fossil fuels. The commercial solar cell marketing currently is dominated by the inorganic semiconductor-based solar cells such as silicon solar cells, which typically require sophisticated high-temperature processing and complex engineering, rendering them cost prohibitive as an energy source for wide applications. These cells are also hampered by limited mechanical flexibility. Polymer solar cells show great potential as a low-cost alternative to Si solar cells because they can be fabricated using inexpensive painting or simple printing. In addition to their low cost, polymer solar cells offer additional advantages of being lightweight and having mechanical flexibility. Here the basic concepts and device physics of organic photovoltaics will be presented.

Organic Solar Device Configuration

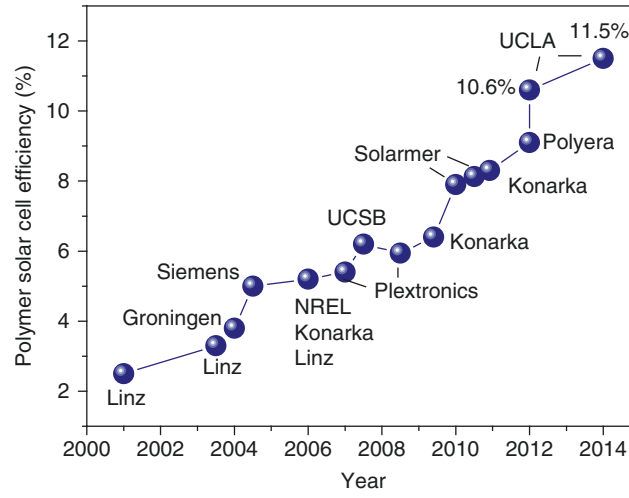
The early stage of organic solar cells was made in a single-layer configuration as shown in Fig. 1a. The devices were fabricated by preparing an organic layer sandwiched between a low-workfunction metal layer (Al) and a high-workfunction metal layer (Au). The solar cells showed poor photovoltaic performance due to inefficient exciton dissociation [1]. Because of low dielectric constant in organic semiconductors,

exciton binding energy (E_b) is relatively large with a typical value of ~0.1–0.4 eV [1]. This is different from Si solar cells, in which the photoexcited electrons and holes have a quite weak binding energy that can be separated by the built-in electric field of a PN junction [2]. However, in organic solar cells, the built-in electric field is not high enough to dissociate the photogenerated excitons.

Then in 1986, Tang [3] reported a bilayer solar cell consisting two distinctive layers (one is donor layer and the other is acceptor layer). The cell was fabricated by thermally evaporated copper phthalocyanine (p-type) and perylene diimide (n-type) derivatives, respectively, between indium tin oxide and silver electrodes. This cell is also called as "Tang cell" and its configuration is shown in Fig. 1b. The cell power-conversion efficiency has reached ~1 % under simulated solar illumination. Cell performance improvement was attributed to the creation of planar interface between the donor and acceptor layer, which provided an energetically favorable interface for exciton dissociation. The acceptor is usually more electronegative and thus electrons have a tendency to transfer to the acceptor, and holes stay with the donor. In order to enhance exciton dissociation and prevent exciton recombination, a large interfacial area is preferred. However, the drawback of bilayer solar cells is that the interfacial area is limited by the fixed planar structure between the donor and acceptor layer. In addition, exciton diffusion lengths in organic semiconductors such as conjugated polymers are only about 5–15 nm, while the light-absorption depth in organic semiconductors is ~100 nm for sufficient light absorption [4]. Since the exciton dissociation area is fixed at the interface between two distinctive layers, only the excitons generated within exciton diffusion lengths can reach the interface for dissociation. Thus, a majority of excitons decay to the ground state rather than separate as free-charge carriers in bilayer devices.

To solve the limitations of bilayer solar cells and substantially increase both the light absorption and the interfacial area for exciton dissociation, the bulk-heterojunction concept (Fig. 1c) was introduced by Yu et al. [5]. Fullerene C_{60} and its derivatives PCBM have been applied as the electron acceptor since the ultrafast photoinduced electron transfer from a conjugated polymer (MEH-PPV) to fullerene (C_{60}) at a time scale of 50 fs was found by Heeger [6], and fullerene and its derivatives have been proved as the most efficient type of acceptors so far. Bulk-heterojunction solar cells are typically assembled by blending the donor and acceptor materials together and then sandwiching the mixture layer between two electrodes. The nanoscale morphology of the donor and acceptor is a crucial factor limiting the cell performance, and solvent selection affects the solubility of fullerene which thus leads to the different roughness of blend films. Besides solvent, thermal and solvent annealing methods are demonstrated as effective approaches to improve cell efficiency since the polymer chain ordering are reorganized through the treatment. The bulk-heterojunction solar cells have been extensively studied, and great progresses have been achieved in increasing organic solar cell efficiencies.

Ordered heterojunction solar cells (Fig. 1d) were also investigated, and they are considered as an ideal configuration for solar cell application, because they have direct charge transport pathway. Inorganic acceptor semiconductors can be made as nanotube-, nanorod-, and nanowire-shaped structures, and they can be vertically aligned onto conducting substrates. The polymer solution can be infiltrated into the nanoporous and nanospaces by spin coating or drop casting. The devices can also be assembled by in situ synthesis of polymer via UV-light irradiation, chemical oxidative method, or electrochemical polymerization methods. However, so far, cell efficiencies have been lower than those of randomly mixed bulk-heterojunction polymer solar cells. A possible reason is that the in situ synthesized polymers might not have the same high quality as the ex situ synthesized polymers because the latter can be purified and optimized before used for solar cell fabrication. Another possibility is that in situ synthesized polymers are generally combined with inorganic acceptors that usually have less photovoltaic performance than fullerene derivative-based solar cells.



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 2 Efficiency evolution of polymer solar cells from 2001 to 2014 [9–11]

Inverted solar cell (Fig. 1e) is made of n-type buffer layer on top of a transparent cathode (conducting glass), then the active layer, p-type buffer layer, and a metal anode layer were stacked subsequently. The inverted structure was first reported by Yang and Li et al. in 2006; they used a low-workfunction interfacial layer Cs_2CO_3 to lower the workfunction of ITO [7]. Transition metal oxides such as V_2O_5, MoO_3, and WO_3 were usually used as a hole-buffer layer placed on the anode side to form the inverted structure. ZnO has been demonstrated to be an efficient p-type buffer layer in the inverted device. The major advantage of inverted design compared with the conventional one is the better cell stability.

Tandem solar cell (Fig. 1f) was firstly invented by Boer and Janssen in 2006 [8]. Tandem solar cell consists of more than one sub-cell with the complementary light-absorption properties from donor materials, and an interconnecting layer such as MoO_3, TiO_2(n type)/PEDOT:PSS (p type), and ZnO(n type)/PEDOT:PSS(p type) was employed to connect the top and the bottom cell. The overall device shows higher performance than that of individual sub-cell.

Figure 2 shows the efficiency evolution of polymer solar cells in the last decade. So far, the highest cell efficiency of 11.5 % with a triple-junction polymer solar cell has been in reach by Yang's group at UCLA.

Organic Photovoltaic Materials

Donor Materials (p-Type Materials). The conductivity in organic semiconductors is attributed to the alternating single and double bonds or alternating single and triple bonds. In the system with alternating single and double bonds, the s orbital and two p orbitals form three hybridized sp^2 orbitals, which form σ bonds; while the fourth orbital (p_z) is perpendicular to the plane of the hybridized sp^2 orbitals, forming a π bond. The single bonds are σ bonds; the double bonds include a σ-bond and a π-bond. In the system of alternating single and triple bonds, the s orbital and one p orbital form two hybridized sp orbitals, while the other two form π orbitals. The single bonds are only σ bonds, while the triple bonds have one σ-bond and two π-bonds. The overlap of π orbitals causes delocalization that is the origin of carrier conductivity. Because of Peierls instability, the system is distorted and asymmetric; the single bonds become a little longer, while the double or triple bonds become a little shorter. Thus, the π electrons produce two delocalization energy bands: one is from lower-energy bonding π orbitals (π band) and the other is from higher-energy anti-bonding π orbitals (π^* band) [4]. The former is generally called the highest occupied molecular orbitals (HOMO) and the latter is the lowest unoccupied molecular orbitals (LUMO). The HOMO–LUMO gap usually determines the

optical properties such as photon absorption and emission. Most organic semiconductors have a bandgap within 1.5–4 eV and a high absorption coefficient up to $\sim 10^5$ cm^{-1}. Therefore, an incident visible-light photon has sufficient energy to excite an electron from the HOMO to the LUMO of conjugated polymers. This makes them well suited to absorb visible light for photovoltaic application.

A variety of conjugated polymers have been used as donor materials in organic solar cells. Polythiophene (PT) and its derivatives poly(3-hexylthiophene) (P3HT), poly(p-phenylene vinylene) (PPV) and its derivatives are two major types of polymers. These polymers typically only absorb light up to 650 nm, showing a limited light-harvesting capability. Further substantial increases in device efficiency will partly depend on significant broadening of absorption spectra to harness low energy photons beyond those of the current polymers. Low-bandgap polymers have been used for broader light harvesting of the solar spectrum. In addition, the polymers need to have a relatively high degree of planarity and high carrier mobility. The polymers should be able to self-assembled into a more organized structure with enhanced backbone packing and charge transport using treatments such as thermal and solvent annealing. The donor polymers should also properly phase separate and create a hole-transport network with efficient hole transport, and the donor–acceptor interfaces need to be finely managed within the exciton reach to increase exciton dissociation. The energy levels (LUMO and HOMO) also need to be optimized to achieve efficient charge transfer as well as a high effective bandgap ($E_{DA} = LUMO_A - HOMO_D$) for high open-circuit voltages (V_{oc}). Generally, the key optical and electronic properties of donor polymers are their bandgaps, LUMOs, and carrier mobilities. Their bandgaps should be close to the optimal $\sim$1.5 eV [1, 10, 12]. Figure 3 lists some conjugated polymers used in organic solar cells [1].

Acceptor materials (n-type materials). In organic solar cells, the acceptor materials need to have a higher electron affinity than the donor polymers. In other words, the LUMO of acceptor materials need to lie well below the related LUMO of various polymers, making them energetically favorable for exciton dissociation and charge transfer at the interfaces. In addition, a high-electron-accepting ability is also required for acceptors used in solar cells. For example, fullerene and its derivatives $PC_{61}BM$ and $PC_{71}BM$ (Fig. 4) are capable of taking multiple electrons. Another property is high electron mobility, which works as a "speed limit" for electron transport. In addition, inorganic nanoscale semiconductors such as CdSe, TiO_2, and ZnO with different kinds of nanostructures can be used as acceptors or electron-transport materials in polymer/inorganic hybrid solar cells. These inorganic nanostructures can be prepared into various morphologies through wet chemistry methods. Different shapes such as nanoparticles, nanorods, nanosheets, tetrapods, and hyperbranched nanostructures of various inorganic were investigated. Figure 5 shows the scanning electron microscopy (SEM) images of ZnO nanofilms with nanoridge (a), nanosheet (b), and nanorod (c) shape. The morphologies mainly affect the interface area between inorganic and polymers, thus influencing the exciton dissociation. See entry on "► Hybrid Solar Cells".

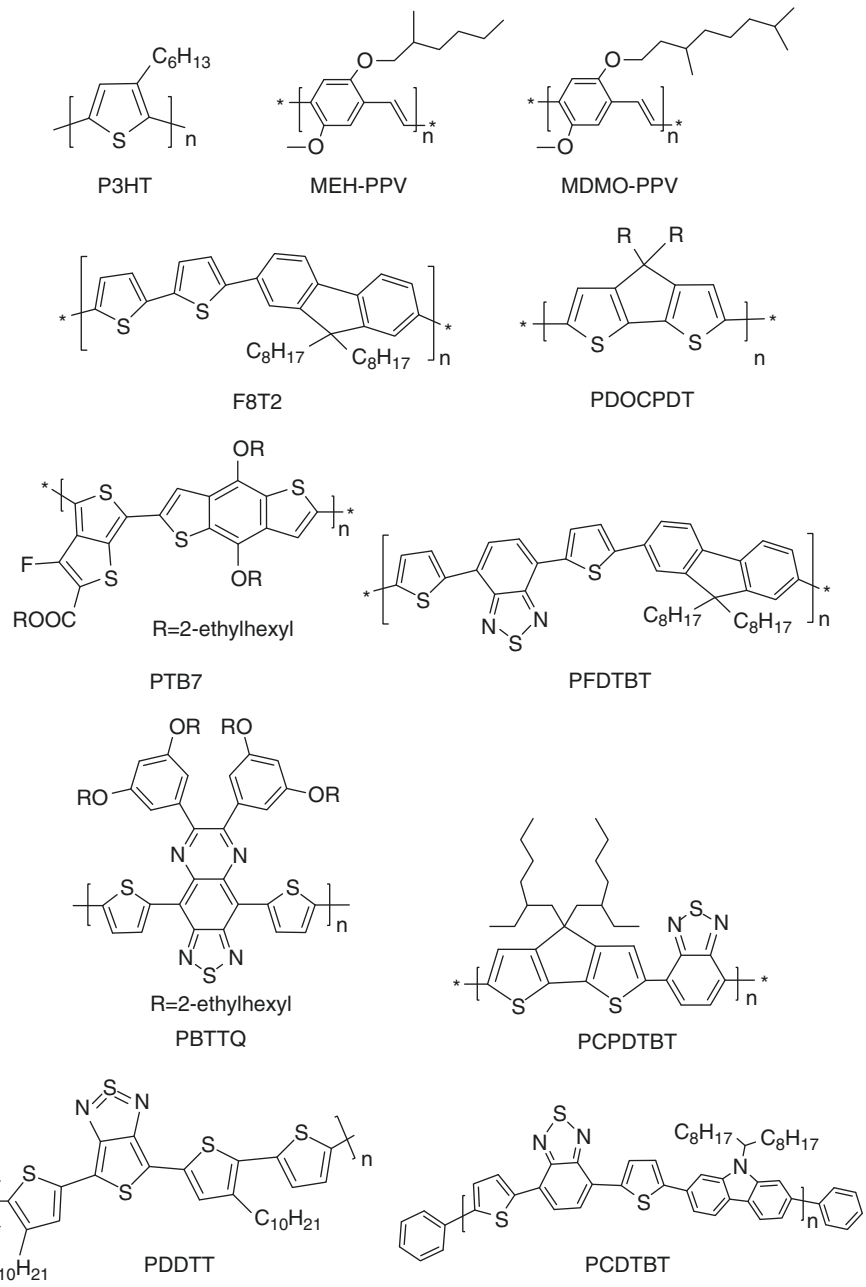
O

Basic Concepts in Organic Photovoltaics

Characterization of a solar cell. Solar cell performance can be illustrated by current density–voltage (J–V) curves, which are shown in Fig. 6a under dark (black curve) and illumination (red curve). The power density ($P = JV$) was plotted (blue curve). From the J–V curve, open-circuit voltage (V_{oc}), short-circuit current (J_{sc}), and energy-conversion efficiency (η) can be obtained.

$$\eta = \frac{J_m V_m}{P_s} = \frac{J_{sc} V_{oc} FF}{P_s} \tag{1}$$

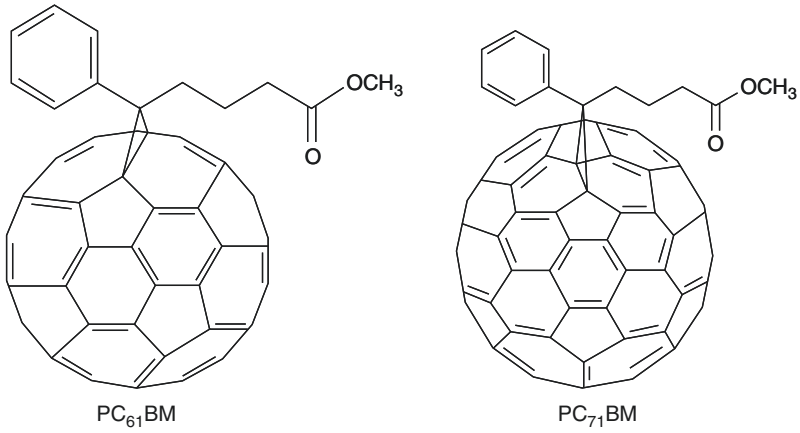
where J_m and V_m are the current and voltage at the maximum power point, respectively, FF is the fill factor, and P_s is the incident light power density.



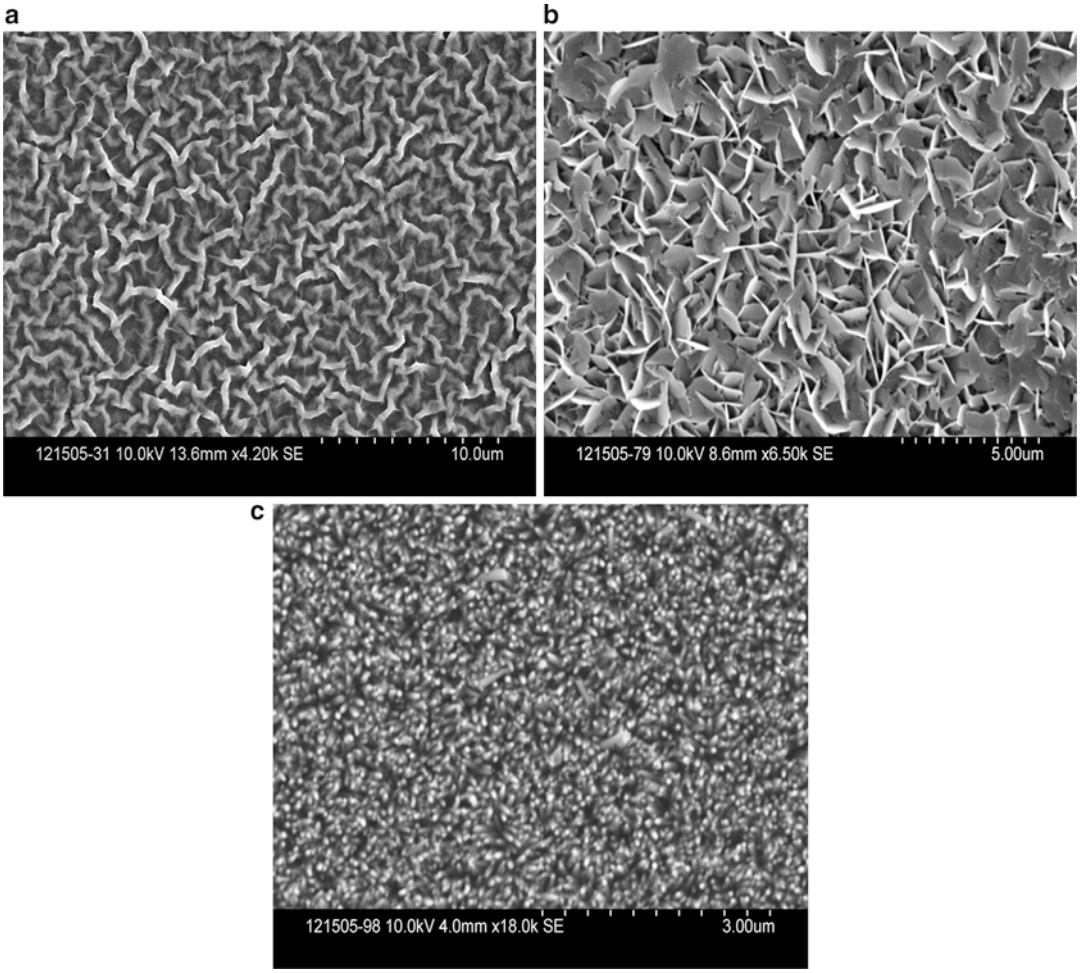
Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 3 Chemical structures of some conjugated polymers used in solar cells

The equivalent circuit of a solar cell is shown in Fig. 6b. A series resistance (R_s) may originate from contact and bulk semiconductor, and a shunt resistance (R_{sh}) may come from poor diode contact [15]. The J–V characteristics can be described as

$$J = J_0\left\{\exp\left[\frac{q(V - JR_sA)}{nk_BT}\right] - 1\right\} + \frac{V - JR_sA}{R_{sh}A} - J_{ph} \quad (2)$$

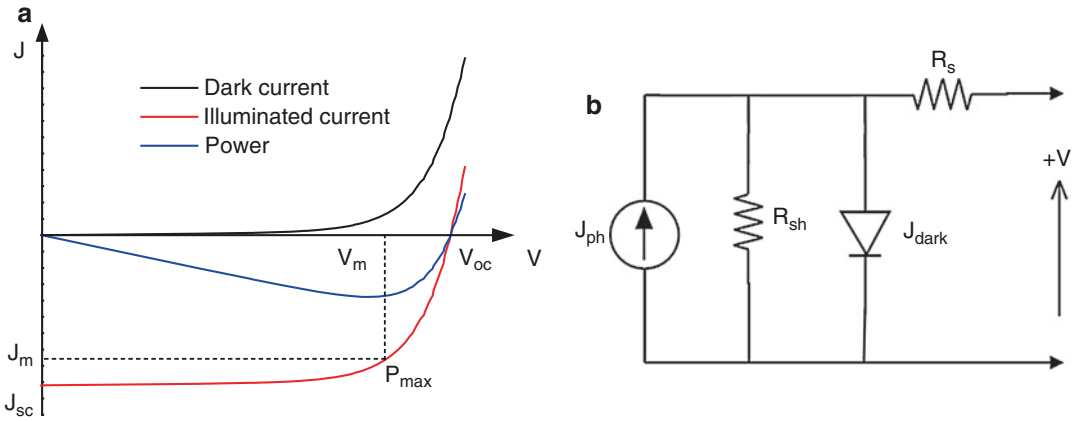


Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 4 Chemical structure of fullerene derivatives $PC_{61}BM$ and $PC_{71}BM$



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 5 ZnO nanostructured films with (**a**) nanoridge [13]; (**b**) nanosheet (unpublished result); (**c**) nanorod [14] morphologies as electron acceptors

O



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 6 (**a**) The characteristic current density- and power-voltage curves under dark and illumination; (**b**) equivalent circuit for polymer photovoltaic cells

where k_B is Boltzmann constant, T is temperature, q is elementary charge, A is device area, n is ideality factor of the diode, J_0 is reverse saturation current density, J_{ph} is photocurrent, R_s is the series resistance, and R_{sh} is the shunt resistance. The J–V curves and photovoltaic parameters including J_{sc}, V_{oc}, and FF strongly depend on the n, J_0, R_s, and R_{sh}.

Short-Circuit Current Density. Short-circuit current density (J_{sc}) is the photocurrent generated by a photovoltaic device under illumination at short-circuit condition. It depends on the incident light intensity and absorption spectra of active materials, which can be described as [10]

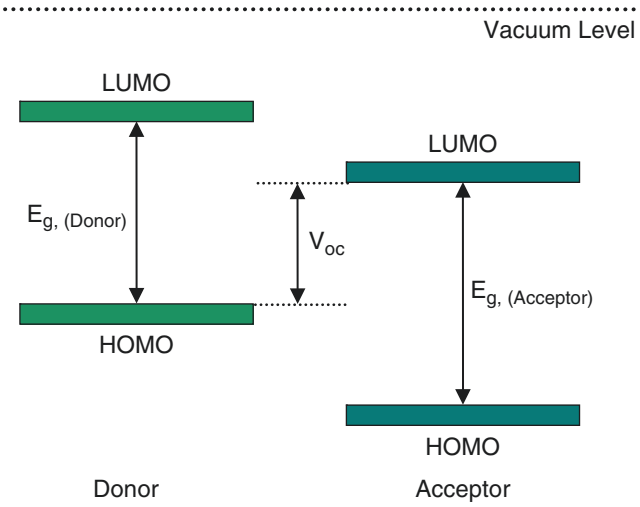
$$J_{sc} = \int_{E_g}^{\infty} e \cdot N_{ph}(E) \cdot EQE(E) dE \qquad (3)$$

where N_{ph} is the spectral photon flux of incident light, E_g is the bandgap of active materials, E is the photon energy, and EQE is the external quantum efficiency.

Open-Circuit Voltage. Open-circuit voltage (V_{oc}) is defined as the voltage across the cell under illumination with a zero current at which the dark current and photocurrent are exactly canceled out [2]. V_{oc} is the largest voltage output from a solar cell [16]. By solving the current density verse voltage equation (Eq. 2) at $J = 0$ and $V = V_{oc}$, V_{oc} can be derived as [17]

$$V_{oc} = \frac{nk_BT}{q} \ln\left(\frac{J_{ph}}{J_0} + 1 - \frac{V_{oc}}{J_0 R_{sh}}\right) \qquad (4)$$

The V_{oc} of polymer–fullerene solar cells have been widely studied. In early days, the V_{oc} was considered to be determined by workfunction difference of the two electrodes originally based on metal–insulator–metal (MIM) devices. Further studies show that the V_{oc} linearly depends on the donor–acceptor effective bandgap ($E_{g,DA} = HOMO_D - LUMO_A$) [18]. Scharber et al. reported $V_{oc} = E_{g,DA}/e - 0.3\text{V}$ for various polymers with different HOMO energy levels and they found that 0.3 V is an empirical value that describes the difference between the V_{oc} and maximum built-in potential (V_{bi}) [12]. More recently, Vandewal et al. [19] studied the relationship between the V_{oc} and $E_{g,DA}$ and found that the $E_{g,DA}$ (a theoretical maximum of V_{oc}) is proportional to the experimental V_{oc} but several 100 meVs larger (e.g., $V_{oc} = E_{g,DA}/e - 0.43\text{V}$). The donor–acceptor properties, mass fraction of crystalline nanofiber polymer in the entire film, and film morphologies can affect $E_{g,DA}$ [4, 20]. The difference between the V_{oc} and $E_{g,DA}$ is possibly caused by energetic disorders, radiative and nonradiative recombination, and band bending due to carrier diffusion [4, 19]. To maximize V_{oc}, this difference between the V_{oc} and $E_{g,DA}$ needs to be minimized.



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 7 Energy-level diagram of a donor and acceptor system with the theoretical maximum ($E_{g,DA} = HOMO_D - LUMO_A$)

Low-bandgap polymers can be used to increase light absorption for a higher photocurrent by broadening the absorption spectrum; however, the decrease of conjugated polymer bandgaps will also affect the V_{oc}. In a bulk-heterojunction solar cell, the theoretical maximum V_{oc} can be described as

$$V_{oc,\max} = E_{g,DA} = E_g - (LUMO_D - LUMO_A) \quad (5)$$

If the E_b is comparable in both the low- and high-bandgap polymers, then the energy offset ($LUMO_D - LUMO_A$) needed to dissociate excitons is also comparable. Thus, according to Eq. 5, a lower bandgap (E_g) will lead to a reduced theoretical maximum of V_{oc}. In order to achieve significantly higher energy conversion efficiency, a balance between donor bandgap (mainly determining J_{sc}) and donor LUMO (mainly determining V_{oc}) needs to be achieved if a specific acceptor material is used with its LUMO or the conduction band is fixed. Figure 7 shows the energy-level diagram of a donor and acceptor system with the theoretical maximum ($E_{g,DA} = HOMO_D - LUMO_A$).

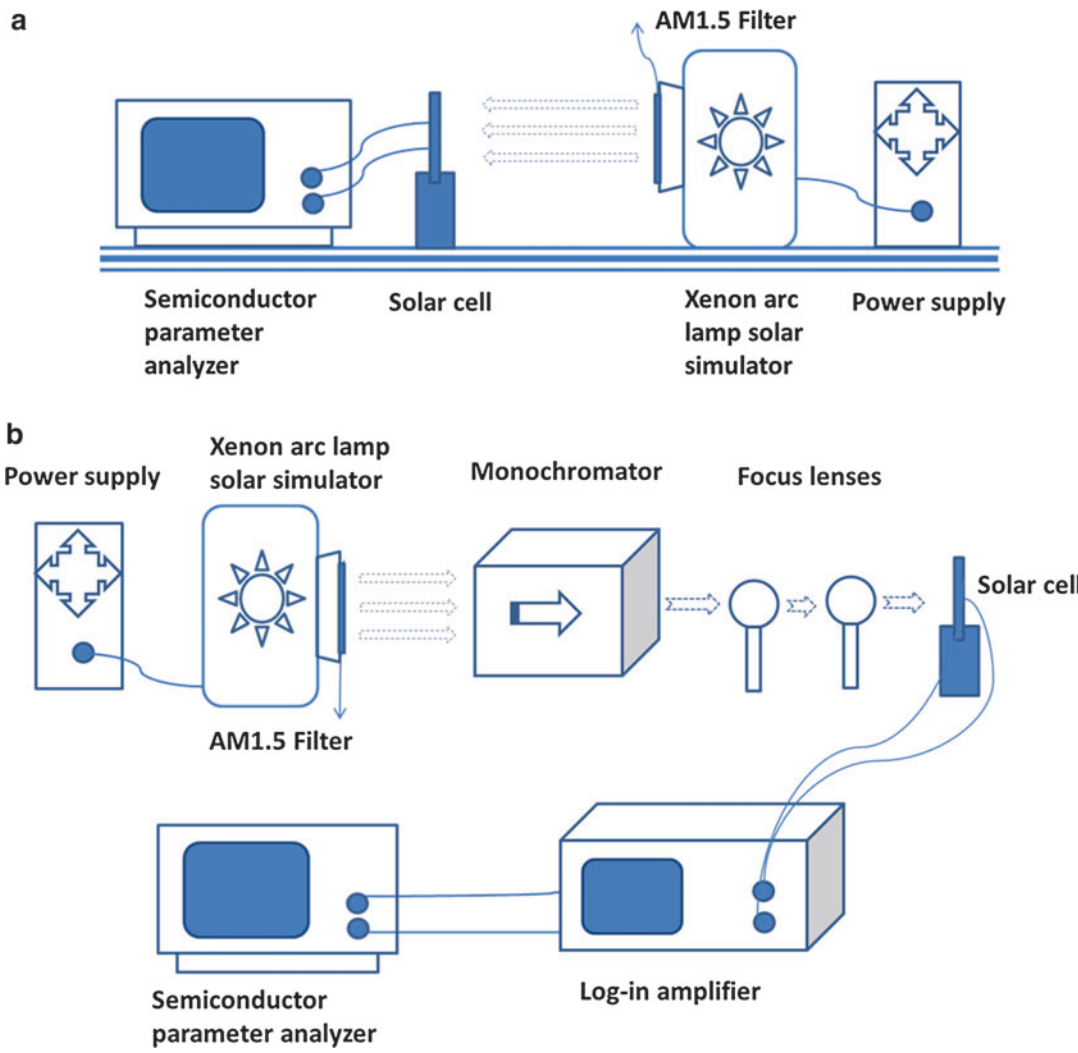
Fill Factor. Fill factor (*FF*) is a parameter, together with J_{sc} and V_{oc}, used to calculate the maximum power output from a solar cell. *FF* is defined as the ratio of the maximum power ($J_m V_m$) to the product of J_{sc} and V_{oc} [16].

$$FF = \frac{P_m}{J_{sc}V_{oc}} = \frac{J_m V_m}{J_{sc}V_{oc}} \quad (6)$$

The *FF* is affected by the series resistance (R_s) and shunt resistance (R_{sh}). The relationship can be derived herein [17]. High *FF* can be achieved with a low R_s and a high R_{sh} (ideally $R_s = 0$, $R_{sh} = \infty$). Therefore, the R_s needs to be minimized and the R_{sh} should be maximized to ensure a high *FF*. The dependence of *FF* on R_s and R_{sh} can be approximated as [17]

$$FF(R_s, R_{sh}) \approx FF(0,\infty)\left[\left(1 - \frac{J_{sc}R_s}{V_{oc}}\right) - \left(\frac{V_{oc}}{J_{sc}R_{sh}}\right)\right] \quad (7)$$

External and Internal Quantum Efficiency. External quantum efficiency (*EQE*), called incident photon-to-electron conversion efficiency (*IPCE*), is defined as a percentage of the number of charge carriers collected at the electrode under short-circuit condition to the number of photons incident on the device. *EQE* can be calculated from Eq. 8. Internal quantum efficiency (*IQE*) is defined as the percentage of absorbed photons at a given wavelength which are converted to the



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 8 (**a**) Schematic of current density versus voltage (J–V) measurement setup; (**b**) schematic of incident photon-to-electron conversion efficiency (IPCE) measurement setup [13]

current. The relationship of *EQE* and *IQE* can be found as Eq. 9.

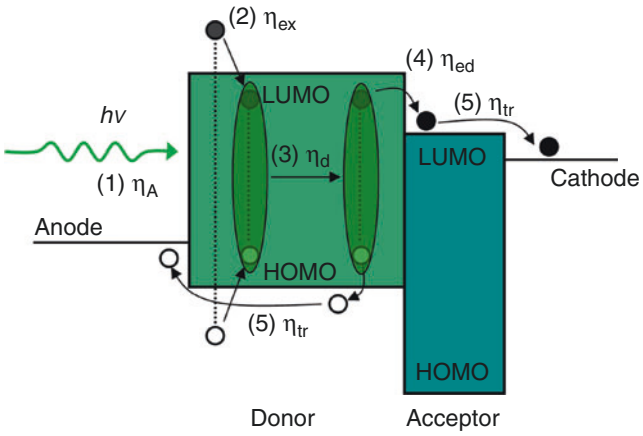
$$EQE = \frac{J_{sc}hc}{P_s\lambda e} = \frac{1240J_{sc}}{P_s\lambda} \tag{8}$$

$$IQE = \frac{EQE}{1-R-T} \tag{9}$$

where h is Planck's constant, c is velocity of light, R is reflectance, and T is transmittance.

Measurements of Solar Cells

Current Density versus Voltage (J–V) Measurements. Dark and illumination current density versus voltage (J–V) measurements can be tested using semiconductor parameter analyzer combined with a xenon lamp AM 1.5 solar simulator as shown in Fig. 8a [21]. The first step is to use a standard silicon photodetector to calibrate the light intensity at 100 mW/cm^2. Then a sample solar cell is located at the fixed position with a



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 9 Energy level diagram of a donor–acceptor organic solar cell with key processes. Note step 6 (charge collection) is integrated into step 5 (charge transport) and thus not shown in this figure (hγ: photon energy) [1]

sweeping voltage from −1 to 1 V under dark and light illumination. The current was then measured by a semiconductor parameter analyzer. Solar cell parameters such as open-circuit voltage (V_{oc}), short-circuit current (J_{sc}), and fill factor (*FF*) can be obtained from the measurement. Solar cell efficiency () is calculated using Eq. 1.

External Quantum Efficiency (EQE). External quantum efficiency (*EQE*) can be tested using the setup as shown in Fig. 8b. Since *EQE* is wavelength dependent, a monochromator is used to sweep the wavelength in a certain range. The voltage from the parameter analyzer through a log-in amplifier is measured. A standard silicon photodetector is first illuminated by the monochromator light via a set of focus lenses and the voltage is recorded. Then the sample solar cells are tested at the same position when light illuminated. Since the *IPCE* value of the standard device is known, the *IPCE* of the sample cell can be calculated comparing to the reference of the standard device cell using the equation below:

$$IPCE_{\text{cell}} = \frac{V_{\text{cell}}}{V_{\text{ref}}} IPCE_{\text{ref}} \qquad (10)$$

Device Physics

Solar cell is a device that produces electricity from sunlight. Upon illumination, photons are absorbed by an active layer, and then electron–hole pairs are generated as excitons. The excitons need to diffuse to a donor–acceptor interface to dissociate into free charges. After that, electrons and holes need to transport to electrodes through their corresponding percolation pathway. During these processes, six main steps affect device performance (Fig. 9): (i) photon absorption (η_a), (ii) exciton generation (η_{ex}), (iii) exciton diffusion (η_{diff}), (iv) exciton and polaron dissociation (η_{ed}), (v) charge transport (η_{tr}), and (vi) charge collection (η_{cc}). The η is the yield of each process.

These six processes determining solar cell performance can be better understood in a connection with *EQE*, which can be expressed as the product of the η in the above steps.

$$EQE = \eta_a \times \eta_{ex} \times \eta_{diff} \times \eta_{ed} \times \eta_{tr} \times \eta_{cc} \qquad (11)$$

Photon Absorption. The photon absorption yield η_a is determined by the absorption spectral band, optical absorption coefficient, and thickness of a photoactive layer, as well as internal reflection [22]. Most semiconducting polymers (e.g., P3HT, MDMO-PPV, etc.) have a bandgap larger than 2 eV, which limits light absorption range less than 650 nm. Therefore, only a small portion of sunlight can be absorbed in polymer solar cells. Usually, the thickness of an active layer is on the order of 100 nm to avoid exciton and charge transport loss. To well balance sufficient light absorption and efficient charge transport, both a wide absorption spectral band and a high absorption coefficient are important for absorbing enough sunlight. The absorption coefficients of conjugated polymers were reported to

be much higher than that of silicon so that a thin layer (e.g., 100 nm) of polymers is thick enough to absorb sufficient light.

Exciton Generation. Upon light absorption, the exciton is generated and a so-called charge transfer (CT) state is formed. Since the dielectric constant in the organic semiconductor is smaller compared with the inorganic semiconductor, the photoinduced electron–hole pairs cannot be dissociated into free electrons and holes by the thermal excitation; thus the tightly bound excitons with a binding energy of 0.3 ∼ 1.4 eV are formed.

Exciton Diffusion. The exciton moves in the organic semiconductor matrix within the limited diffusion length. The efficiency (η_{diff}) of exciton diffusion to a donor–acceptor interface is related to its exciton diffusion length (L_D) and distance (L_i) between the photoexcitation location and the donor–acceptor interface which serves as a dissociation center. The L_D is equal to $(D\tau)^{1/2}$, where D is diffusion coefficient and τ is exciton lifetimes [23]. Excitons can diffuse to a donor–acceptor interface with $L_i < L_D$; otherwise, they recombine with a reduced η_{diff} [24]. Typically, exciton diffusion length is in the range of 5 ∼ 15 nm for conjugated polymers [1].

Exciton Dissociation. The efficiency (η_{ed}) of exciton dissociation into free holes and electrons relies on the donor and acceptor LUMO energy offsets and internal electric field at a donor–acceptor heterojunction [24]. The minimum energy required to dissociate an exciton in a conjugated polymer is the same amount of energy that is needed to overcome the exciton binding energy. This energy can be provided by the offset between the LUMO energy levels of the donor and acceptor [24]. Once the exciton is dissociated, the hole and electron stay on the donor and acceptor material, respectively, but they are still bound as a polaron pair by Coulomb attraction [4]. To get free electrons and holes, such polaron pairs need to be separated. The polaron pair dissociation can be described by the commonly used Braun–Onsager model, in which the polaron pair dissociation probability is [4]

$$P(F) = \frac{k_d(F)}{k_d(F) + k_f} \tag{12}$$

where F is the electric field, k_f is the polaron pair decay rate to the ground state, and k_d is the polaron pair dissociation rate, which can be expressed as

$$k_d(F) = \frac{3\gamma}{4\pi r_{pp}^3} \exp\left(-\frac{E_b}{k_B T}\right) \frac{J_1\left(2\sqrt{-2b}\right)}{\sqrt{-2b}} \tag{13}$$

$$\gamma = \frac{q}{\varepsilon_0 \varepsilon_r} (\mu_e + \mu_h) \tag{14}$$

$$E_b = \frac{q^2}{4\pi\varepsilon_0 \varepsilon_r r_{pp}} \tag{15}$$

$$b = \frac{q^3 F}{8\pi\varepsilon_0 \varepsilon_r (k_B T)^2} \tag{16}$$

where γ is the Langevin recombination factor, r_{pp} is the radius of the initial polaron pair, E_b is the polaron pair Coulomb binding energy, k_B is the Boltzmann constant, T is the temperature, J_1 is the first-order Bessel function, μ_e is the electron mobility, μ_h is the hole mobility, ε_0 is the permittivity in vacuum, and ε_r is the relative permittivity.

Carrier Transport. The η_{tr} is carrier transport efficiency. The holes are transported in a donor material, while the electrons are transported in an acceptor material. Both the donor and acceptor materials are required to form highly efficient percolation networks spanning the entire active layer to provide efficient charge transport. Drift and diffusion current are the major driven forces for the transport of electrons and holes to their corresponding electrodes. The drift current relates to the potential gradient in the solar cell, which is decided by the workfunction difference of the electrodes, while the diffusion current is formed by the diffusion of the carriers through the carrier concentration gradient. In order to achieve efficient charge transport, a balanced hole and electron is needed to avoid the accumulation of carriers in the interface of the active layer and the electrode. The low mobility of the organic materials is a major factor that limited the carrier

transport; thus, the active layer has to be thin enough to allow the carrier to arrive at the electrode in their lifetime. The polymers also need to have a higher degree of planarity for efficient backbone stacking to achieve a high hole mobility. Through the treatments such as thermal and solvent annealing, the polymers should also be able to self-assemble into a more organized structure. In a typical organic solar cell made by the donor–acceptor blends, the overlap of wave functions between adjacent sites in both the donor and acceptor phase is not strong; the carrier transport states are also localized. Thus, the separated charges can only transport by hopping from one localized site to the next with a low carrier mobility. The hopping transport rate can be described by a Miller–Abrahams hopping rate [4, 25]

$$v_{ij} = v_0 \exp(-\gamma r_{ij}) \begin{cases} \exp\left(-\frac{E_j - E_i}{k_B T}\right) & E_j > E_i \\ 1 & E_j \le E_i \end{cases} \tag{17}$$

where v_0 is the maximum hopping rate or attempt-to-escape frequency, γ is the inverse of localization radius, r_{ij} is the distance between sites i and j, and $E_j - E_i$ is the height of energetic barrier (energy difference between sites i and j). The term $\exp(-\gamma r_{ij})$ is from tunneling, which is related to the wave-function overlap between sites i and j. If the hopping transport is upward, the term $\exp\left(-\frac{E_j - E_i}{k_B T}\right)$ is a thermally activated process. Using Miller–Abrahams transition rate and Monte Carlo simulations within Gaussian density of states (DOS) without considering site correlation, the carrier mobility in a donor polymer and its dependence on temperature and electric field can be predicted as [26, 27]

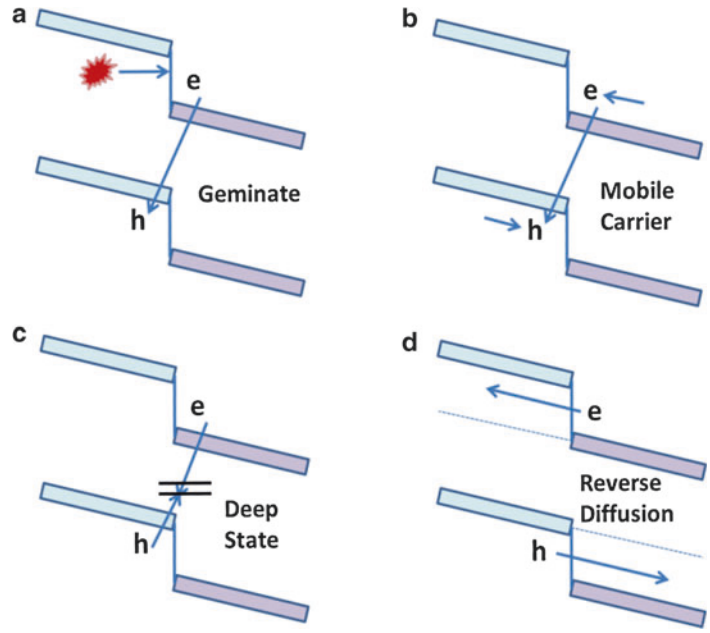
$$\mu = \mu_0 \exp\left\{-\left(\frac{2\sigma}{3k_B T}\right)^2 + C\left[\left(\frac{\sigma}{k_B T}\right)^2 - \sum^2\right]\sqrt{F}\right\} \tag{18}$$

where μ_0 is the mobility when $T \to \infty$, σ is the width of the Gaussian DOS, C is a constant related to site spacing, $\sum$ is the degree of spatial disorders, and F is the electrical field.

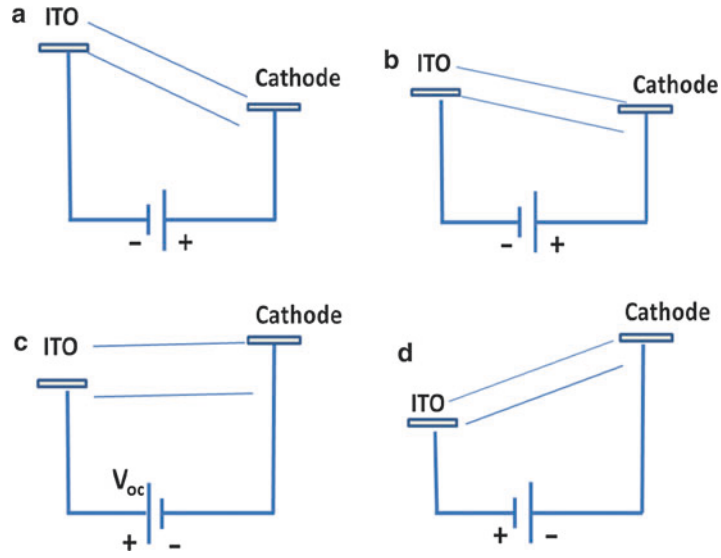
Charge Collection/Extraction. The η_{cc} is charge collection/extraction efficiency. It is the fraction of the charges transported from the active layer to the electrodes with respect to the total free charges that are supposed to transport to the electrodes. Once separated, the polarons (electrons and holes) must travel to their corresponding electrodes to be collected. When the carrier transports to the active layer/electrode interface, there needs to have an ohmic contact to ensure the efficient charge collection, i.e., the workfunction of anode matches with the HOMO of donor, and the workfunction of cathode matches with the LUMO of acceptor. The η_{cc} depends on the device architecture, energy levels of the active layer and the electrodes, and surface recombination at the organic–metal interface [4, 24].

Recombination Loss. Recombination loss is one of the major reasons for low solar cell performance. There are four main recombination mechanisms existing in organic solar cells, namely, geminate, mobile carrier, deep-state, and reverse diffusion recombinations [28] as shown in Fig. 10. Germinate recombination occurs when the electrons and holes recombine before they dissociate. Mobile free-carrier recombination is also called Langevin recombination, and it is the recombination of mobile electrons and holes. Localized deep-state recombination involves the recombination of trapped carriers and mobile carriers. It correlates to the deep-state density. Reverse diffusion to the contacts happens when an exciton dissociates close to one of the electrodes, and the carrier diffuses to the reverse electrode and recombines at the contact.

Operation of Solar Cells. Solar cells can be operated under various biases. Figure 11a shows that under large reverse bias, exciton dissociation and charge transport are enhanced due to the strength of built-in electric field by a large bias. Drift current is dominant at this case. In Fig. 11b, when the applied bias is close to zero, carriers are driven by the built-in field and then collected by respective electrodes. Figure 11c shows that the applied bias reaches to a point that is equal to the built-in field. In this case, diffusion current dominates since the electric field is quite small inside



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 10 Four main recombination mechanisms occurred in the bulk-heterojunction (BHJ) solar cells: (**a**) geminate recombination, (**b**) mobile carrier recombination, (**c**) deep-state recombination, and (**d**) reverse diffusion



Organic Photovoltaics: Basic Concepts and Device Physics, Fig. 11 Operation of the solar cell at different biases: (**a**) large reverse bias; (**b**) small reverse bias; (**c**) positive bias when the internal field is equal to zero, which is called as open-circuit condition; and (**d**) large positive bias

the device. When the applied bias switches to the positive condition, as in Figure 11d, the applied field is larger than the built-in field, and the potential gradient in the device is reversed, carrier injection occurs via tunneling mechanism resulting in the positive current.

Summary

In summary, organic solar cell device structures, materials, basic concepts, and device physics are introduced. To make organic solar cells available for commercial applications, additional basic

research in material synthesis, fundamental understanding of device physics, and device engineering are still needed. Organic photovoltaic technology becomes a potential cost-effective energy source, and further research also needs to improve device efficiency and stability.

Acknowledgments This work was partially supported by the "Fundamental Research Funds for the Central Universities" under grant no. 3102014JCQ01098 as well as NSF CAREER (ECCS-0950731), NASA EPSCoR (NNX13AD31A) and NSF MRI (CBET-1428992).

Cross-References

- ▶ Doping in Organic Semiconductors
- ▶ Electrode–Organic Interface Physics
- ▶ Flexible Electronics
- ▶ Hybrid Solar Cells
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanostructures for Energy
- ▶ Toward Bioreplicated Texturing of Solar-Cell Surfaces

References

1. Xu, T., Qiao, Q.: Conjugated polymer-inorganic semiconductor hybrid solar cells. Energy Environ. Sci. **4**, 2700–2720 (2011)
2. Nelson, J.: The Physics of Solar Cells. Imperial College Press, London (2003)
3. Tang, C.W.: Two-layer organic photovoltaic cell. Appl. Phys. Lett. **48**, 183–185 (1986)
4. Deibel, C., Dyakonov, V.: Polymer–fullerene bulk heterojunction solar cells. Rep. Prog. Phys. **73**, 096401 (2010)
5. Yu, G., Gao, J., Hummelen, J.C., Wudl, F., Heeger, A. J.: Polymer photovoltaic cells – enhanced efficiencies via a network of internal donor-acceptor heterojunctions. Science **270**, 1789–1791 (1995)
6. Sariciftci, N.S., Smilowitz, L., Heeger, A.J., Wudl, F.: Photoinduced electron transfer from a conducting polymer to buckminsterfullerene. Science **258**, 1474–1476 (1992)
7. Li, G., Chu, C.W., Shrotriya, V., Huang, J., Yang, Y.: Efficient inverted polymer solar cells. Appl. Phys. Lett. **88**, 253503 (2006)
8. Hadipour, A., de Boer, B., Wildeman, J., Kooistra, F. B., Hummelen, J.C., Turbiez, M.G.R., Wienk, M.M., Janssen, R.A.J., Blom, P.W.M.: Solution-processed organic tandem solar cells. Adv. Funct. Mater. **16**, 1897–1903 (2006)
9. Kazmerski, L.L.: Solar photovoltaics R&D at the tipping point: a 2005 technology overview. J. Electron. Spectrosc. Relat. Phenom. **150**, 105–135 (2006)
10. Siddiki, M., Li, J., Galipeau, D., Qiao, Q.: A review on polymer multijunction solar cells. Energy Environ. Sci. **3**, 867–883 (2010)
11. Chen, C.-C., Chang, W.-H., Yoshimura, K., Ohya, K., You, J., Gao, J., Hong, Z., Yang, Y.: An efficient triple-junction polymer solar cell having a power conversion efficiency exceeding 11 %. Adv. Mater. **26**, 5670–5677 (2014)
12. Scharber, M.C., Wuhlbacher, D., Koppe, M., Denk, P., Waldauf, C., Heeger, A.J., Brabec, C.L.: Design rules for donors in bulk-heterojunction solar cells – towards 10 % energy-conversion efficiency. Adv. Mater. **18**, 789–794 (2006)
13. Xu, T.: Novel strategies to prepare polymer-inorganic hybrid nanostructure systems for cost-effective energy conversion. Ph.D. dissertation, South Dakota State University (2012)
14. Xu, T., Chen, Q., Lin, D.-H., Wu, H.-Y., Lin, C.-F., Qiao, Q.: Self-assembled thienylsilane molecule as interfacial layer for ZnO nanowire/polymer hybrid system. J. Photonics Energy **1**, 011107 (2011)
15. BERNÈDE, J.C.: Organic photovoltaic cells: history, principle and techniques. J. Chil. Chem. Soc. **53**, 1549–1564 (2008)
16. Honsberg, C., Bowden, S.: PVCDROM, the Beta2 edition of the PVCDROM. http://pvcdrom.pveducation.org/
17. Rand, B.P., Genoe, J., Heremans, P., Poortmans, J.: Solar cells utilizing small molecular weight organic semiconductors. Prog. Photovolt. Res. Appl. **15**, 659–676 (2007)
18. Brabec, C.J., Cravino, A., Meissner, D., Sariciftci, N. S., Fromherz, T., Rispens, M.T., Sanchez, L., Hummelen, J.C.: Origin of the open circuit voltage of plastic solar cells. Adv. Funct. Mater. **11**, 374–380 (2001)
19. Vandewal, K., Tvingstedt, K., Gadisa, A., Inganas, O., Manca, J.V.: On the origin of the open-circuit voltage of polymer-fullerene solar cells. Nat. Mater. **8**, 904–909 (2009)
20. Vandewal, K., Oosterbaan, W.D., Bertho, S., Vrindts, V., Gadisa, A., Lutsen, L., Vanderzande, D., Manca, J. V.: Varying polymer crystallinity in nanofiber poly (3-alkylthiophene):PCBM solar cells: influence on charge-transfer state energy and open-circuit voltage. Appl. Phys. Lett. **95**, 123303 (2009)
21. Shrotriya, V., Li, G., Yao, Y., Moriarty, T., Emery, K., Yang, Y.: Accurate measurement and characterization of organic solar cells. Adv. Funct. Mater. **16**, 2016–2023 (2006)
22. Hoppe, H., Sariciftci, N.S.: Organic solar cells: an overview. J. Mater. Res. **19**, 1924–1945 (2004)
23. Brédas, J.-L., Norton, J.E., Cornil, J., Coropceanu, V.: Molecular understanding of organic solar cells: the challenges. Acc. Chem. Res. **42**, 1691–1699 (2009)

24. Saunders, B.R., Turner, M.L.: Nanoparticle-polymer photovoltaic cells. Adv. Colloid Interface Sci. **138**, 1–23 (2008)
25. Miller, A., Abrahams, E.: Impurity conduction at low concentrations. Phys. Rev. **120**, 745 (1960)
26. Bässler, H.: Charge transport in disordered organic photoconductors a Monte Carlo simulation study. Phys. Status Solidi B **175**, 15–56 (1993)
27. Laquai, F., Wegner, G., Bässler, H.: What determines the mobility of charge carriers in conjugated polymers? Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **365**, 1473–1487 (2007)
28. Dou, L., You, J., Hong, Z., Xu, Z., Li, G., Street, R.A., Yang, Y.: 25th anniversary article: a decade of organic/polymeric photovoltaic research. Adv. Mater. **25**, 6642–6671 (2013)

Organic Sensors

Soyoun Jung[1] and Taeksoo Ji[1,2]
[1]Samsung Mobile Display Co., Ltd, Young-in, South Korea
[2]School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea

Synonyms

Flexible sensors; Polymer sensors

Definition

The demand for small, low-cost sensors that can be built on flexible plastics has driven the rapid maturation of synthetic methodology in the field of organic material-based electronics. This is because organic materials can be processed at low temperatures even close to room temperature offering the potential for very-low-cost manufacturing. The flexibility and light weight that organic materials possess allows a variety of simple low-cost deposition techniques including solution casting, ink-jet printing, stamping, dip-coating, and spray-on and spin-on coating, which enable roll-to-roll manufacturing. Even though organic semiconductors are not desirable for use in applications needing very high switching speeds because of their poor carrier mobility, about three orders of magnitude lower than that of single crystalline silicon, they provide ease of fabrication, flexibility, and conformability able to accommodate the requirements to develop biochemical and physical microsensors that possess good portability, low-power dissipation, and compact device integration. Sensors based on organic electronics can be implemented for diverse applications in medical diagnostics and care, environmental monitoring, food monitoring, and military defense.

Charge Transport in Organic Semiconductors

Organic semiconductors have been studied since the late 1940s [1], especially concerning conjugated small molecules and polymers. However, their organic semiconducting characteristics were too poor for utilization in any real device development. The interest in organic semiconductors faded until the late 1970s because of their severe limitations, such as very high density of defects and traps and very low carrier mobility [2]. Discovery of the conducting properties of polymer polyacetylene in 1977 by Shirakawa, MacDiarmid, and Heeger (Noble Prize in Chemistry 2000) led to a large amount of research into the new field of organic electronics. Since this discovery, organic materials have received considerable attention as potential replacements for their inorganic counterparts in flat panel display driver circuitry, light-emitting elements, solar cells, as well as in thin-film transistor (TFT) and sensor applications since they exhibit the advantage of simple, cheap, low-temperature, thin-film processing.

Organic semiconductors can be classified into two groups: conjugated polymers and short conjugated oligomers (small molecules). The former includes, for instance, poly-3-hexylthiophene, poly-p-phenylene vinylene, polyacetylene, and its derivatives, while the latter contains pentacene, anthracene, and rubrene. Organic molecular crystals are known to be formed by rather weak Van der Waals attraction, resulting in pronounced

polaron formation due to the strong electron–lattice interaction, so the charge transport in organic semiconductors must be profoundly different from those of covalently/ionically bonded inorganic semiconductors. However, the charge transport in organic TFTs is still not well understood even though there have recently been several models reported to explain the mechanism, such as multiple trapping and release [3] and variable range hopping [4] models.

Organic molecules are commonly described as compounds containing carbon atoms combined with other atoms, which typically possess a wide energy bandgap (1–4 eV) and a small bandwidth (0.1–0.5 eV). Among them, such conjugated small molecules and polymers as pentacene, α-sexithiophene (α-6 T), and poly-3-hexylthiophene (P3HT) have been especially preferred as the active materials in organic devices owing to their relatively high mobility. For inorganic semiconductors, the structures are generally sustained by strong covalent or ionic bonds that keep atoms in the lattice, and their charge carrier transport in a material, typically explained by the single-electron approach, is described as the movement of highly delocalized plane waves in wide bands. However, this band transport model is not simply applicable to organic semiconductors with complicated structural organizations, requiring still further studies to understand the real nature of transport mechanisms in organic semiconductors.

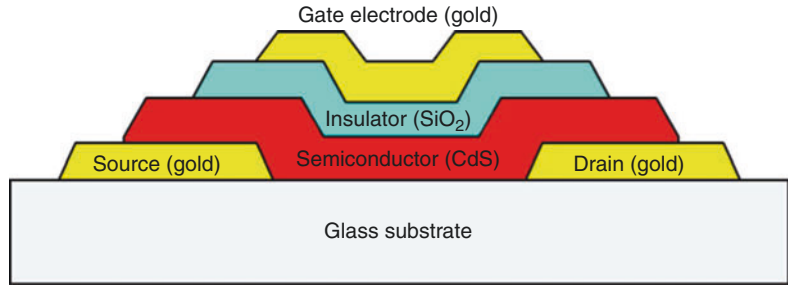
The molecular structure of organic semiconductors is composed of regularly alternated single (σ) and double (π) chemical bonds. Their molecular structure can be mainly categorized into two different types: intramolecular structure of individual molecules composed of covalently bonded atoms, and intermolecular structure of molecules in the lattice consisting of discrete molecules held together by weak Van der Waals forces. In the intramolecular structure, the covalent bonds between atoms are formed by overlapping sp and p_z orbitals in which the former results in σ bonds with localized σ-electrons and the latter leads to π bonds with delocalized π-electrons spread out over the molecule, and across the conjugated backbone. The π-electrons are responsible for physical and chemical properties as well as intermolecular interaction related to the formation of crystal structures in organic semiconductors. In the intermolecular structure, the orientation of molecules is determined by the summation of interaction energies, the Van der Waals interaction, and the quadrupolar interaction. The van der Waals interaction stemming from a fluctuating dipole moment on a molecule is weak and prefers face-to-face orientation of the molecules. On the other hand, the quadrupolar interaction arising from the permanent quadrupole moments by π orbitals favors edge-to-face orientation of molecules. Hence, the crystal structure is determined by the interplay of the two interactions, and, as a consequence, the most common crystals, such as naphthalene, tetracene, and pentacene, exhibit a herringbone arrangement of the molecules [5].

Charge transport is significantly dependent on the intermolecular interaction which determines π overlapping between neighboring molecules. The π orbital overlap is affected by the orientation of molecules, which results from the interplay of van der Waals and the quadrupole interaction. Therefore, it is very important to understand the molecular structure in order to explicitly explain the charge transport in organic molecules.

Organic Thin-Film Transistors (OTFTs)

After Lilienfeld's new idea on field effect devices was proposed in the 1930s [6], Weimer [7] first developed thin-film transistors based on a top-gate staggered structure using thin films of polycrystalline cadmium sulfide as the active semiconductor layer in 1962. Figure 1 shows the early TFT device structure that he employed. The device was fabricated onto a glass substrate by evaporating all the components of the TFTs through shadow masks. The current–voltage characteristics of this device were similar to those of a metal-oxide-semiconductor field effect transistor (MOSFET).

Since Weimer reported the first TFT based on cadmium sulfide, TFTs have been developed with different semiconductor materials as the active layer such as cadmium sulfide, cadmium selenide,



Organic Sensors, Fig. 1 Cross section of Weimer's top-gate staggered TFT structure [7]

tellurium, and lead sulfide. In the late 1970s, Spear and LeComber [8] reported that hydrogenated amorphous silicon (a-Si:H) could be doped, and shortly after the report, LeComber et al. [9] first proposed to use amorphous silicon in active matrix flat panel display applications. Since then, a-Si:H thin-film transistors have been intensively researched and developed in spite of their poor electronic properties compared to crystalline silicon films. The technology of TFTs based on a-Si:H is attractive in large-area applications such as portable computer displays, digital scanners, solar cells, medical imaging sensors, and active matrix liquid crystal displays (AMLCD).

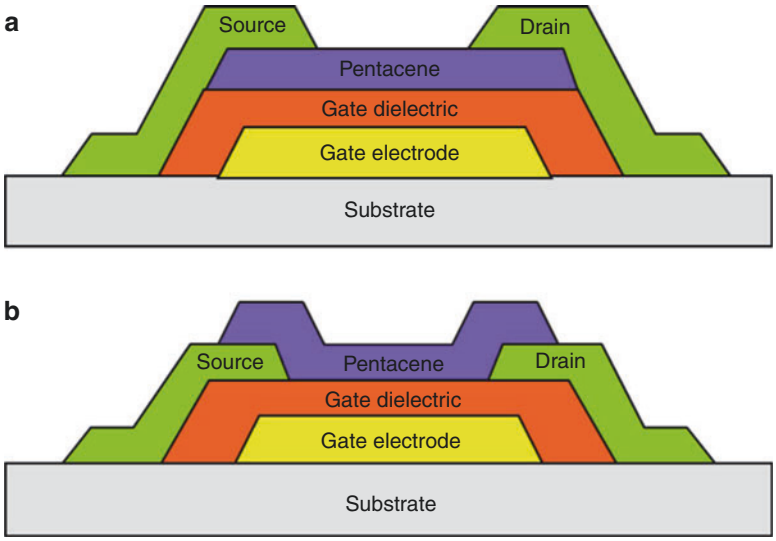
One of the main advantages of using amorphous silicon is that a uniform thin film can be easily achieved over large areas, for instance, using plasma-enhanced chemical vapor deposition (PECVD). Furthermore, amorphous silicon can be deposited at relatively lower temperatures ($\sim$250 °C) than other materials, such as polysilicon ($\geq$600 °C) allowing low weight, and rugged polymeric substrate. Typically, a-Si:H TFTs exhibit about 1 $cm^2/V \cdot s$ field effect mobility, around 1 V threshold voltage, and high on/off ratio of $\sim 10^8$.

However, the processing temperature of a-Si:H TFTs is still not low enough to enable low-cost, flexible, and large-area applications. For instance, commercially available flexible polymeric films typically limit process temperature to less than 200 °C. However, electronics based on organic semiconductors capable of being processed at lower temperatures than a-Si can be aimed at such flexible applications where ease of fabrication, mechanical flexibility, low-cost, and large-area applicability are of particular importance.

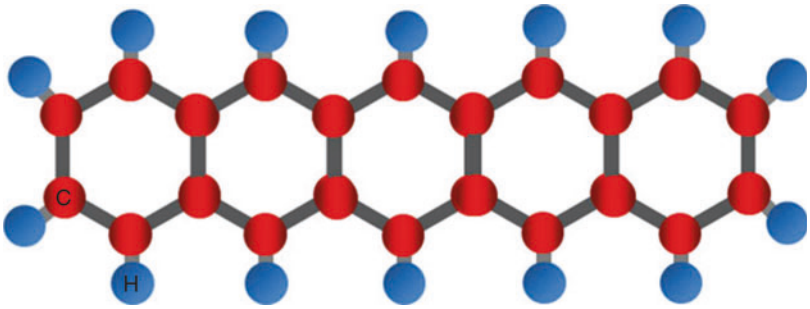
Since the first demonstration of OTFTs in 1983 [10], their performance has been improved dramatically, greatly increasing the carrier mobility from $10-^4$ to 1 $cm^2/V \cdot s$ [11]. OTFT technology involves the use of organic semiconductor compounds in electronic components, flat panel displays, organic light-emitting diodes (OLED), and so forth. The interest in the study of OTFT, in particular, was derived from the fact that the device is easy to fabricate compared to traditional silicon-based transistors. More specifically, low-process temperatures may allow organic TFTs to be realized on inexpensive plastic substrates, rather than on rigid materials as traditionally necessary.

There are two common device configurations used in OTFTs, as shown in Fig. 2. The first one is a top-contact device with source and drain electrodes evaporated onto the organic semiconducting layer (Fig. 2a). The other is a bottom-contact device with the organic semiconductor deposited onto prefabricated source and drain electrodes (Fig. 2b). The current–voltage characteristics of OTFTs show behavior similar to a MOSFET, thus enabling the employment of the MOSFET model to analyze OTFT behavior. The major difference between OTFTs and MOSFETs is the mode in which the conducing channel is formed. Thus, OTFTs operate in the accumulation mode, while MOSFETs form the conducting channel in the inversion mode.

The current and voltage behavior in an OTFT, based on the gradual channel approximation, is as follows:



Organic Sensors, Fig. 2 OTFT device configuration. (**a**) Top-contact structure (**b**) bottom-contact structure



Organic Sensors, Fig. 3 A molecular structure of pentacene

For the non-saturation region: $V_{DS} < V_{GS} - V_T$

$$I_{DS} = \mu_{FE} C_o \frac{W}{L}\left[(V_{GS} - V_T)V_{DS} - \frac{V_{DS}^2}{2}\right] \quad (1)$$

For the saturation region: $V_{DS} > V_{GS} - V_T$

$$I_{DS} = \frac{\mu_{FE} C_o W}{2L}(V_{GS} - V_T)^2 \quad (2)$$

where W and L are the channel width and length, Co is the gate dielectric capacitance, VT is the threshold voltage, VDS is the drain-source voltage, VGS is the gate-source voltage, and μEF is the field effect mobility that can be extracted from:

$$\mu_{EF} = \frac{2L}{WC_o}\left(\frac{\sqrt{\partial I_{DS}}}{\sqrt{\partial V_{GS}}}\right)^2 \quad (3)$$

Figure 3 shows the molecular structure of pentacene, a well-known p-type organic material, which is currently under extensive investigation for use in TFTs as the active layer because of its highest field effect mobility among all organic semiconductors reported to date. Since the carrier mobility strongly depends on the π-orbital overlap of neighboring molecules, the ordering of the organic materials controlled by relatively weak Van der Waals interactions is the key to improvement in the mobility. It has been well known that the formation of highly ordered pentacene films resulting in large grains is significantly affected by both growth conditions and substrate surfaces such as deposition rate, substrate temperature, substrate morphology, and chemical properties of the substrate. Since the nucleation of the first monolayer is extremely sensitive to defects and/or impurities on the substrate, careful treatment of the substrate surface is strictly required to obtain

large as well as highly ordered grains. Grain boundaries associated with trap sites act as carrier barriers that dominate charge transport by trapping at the boundaries. This intergrain transport is mostly responsible for the low mobility of the devices. Thus, the investigation of close correlation between the morphology and the structural properties of organic semiconductors should be first required to obtain the most advantageous structures and morphological characteristics for improved performance of organic devices.

Organic Sensors

Over the last decade, micro-fabrication technologies have been extensively applied to yield a miniaturized biochemical and physical sensor in order to monitor various parameters such as ion diffusion, temperature, strain, and electrolyte conductivity leading to the successful development of multi-parameter measuring systems including electronic nose/tongue, micro total analysis system (μTAS), and lab-on-chip. It is obvious that microsensors based on the principle of the field effect and solid state transport in semiconductor structures offer a number of advantages over traditional physiobiological sensors with regard to real-time electronic readout, small size, and rapid response.

In the last several years, much research effort has been devoted to the use of OTFTs as multi-parameter sensors, pursuing good levels of selectivity, reliability, and reproducibility at low cost. This is due to the fact that OTFT sensors can provide a variety of information from changes in multiple properties on exposure to biochemical as well as physical quantities, such as the bulk conductivity of an organic thin film, the saturation current, the threshold voltage, and the field effect mobility. It is obvious that the multi-functionality in such a multi-parameter sensor operating in different modes enables both envisioning and realization of a simple "higher order" sensor network, which leads to low-cost manufacturing and easy commercialization. In addition, the inherent characteristics of field effect devices such as signal amplification are expected to provide transistor-based sensors with enhanced detection limits and sensitivity compared to usual chemiresistor, amperometric, or potentiometric sensors. Recently, the possibility of using OTFTs for sensing devices has been demonstrated by many research groups. In this case, the conductivity of the organic semiconductor is modulated due to the direct interaction with gas, pressure, humidity, and strain.

As one of the promising applications of organic electronics, large-area strain sensors integrated with transistor matrix capable of monitoring shape, strain, force, and/or pressure in a subject material have been envisioned for a broad spectrum ranging from microsurgery, point-of-care (POC) diagnostics, and monitoring of respiration rate to arrays such as flexible artificial skins of robots and structural health monitoring systems for aging aircrafts. Especially, the key to POC diagnostics is the development of small, light, portable, and implantable sensor devices that are capable of monitoring and controlling diseases in real time so that health-care professionals can make rapid and accurate diagnosis of diseases. This will ensure the effectiveness of therapy along with early detection. Thus, there has been high demand at the POC for development of biochemical and physical microsensors that possess good portability, high sensitivity, low-power dissipation, and compact device integration.

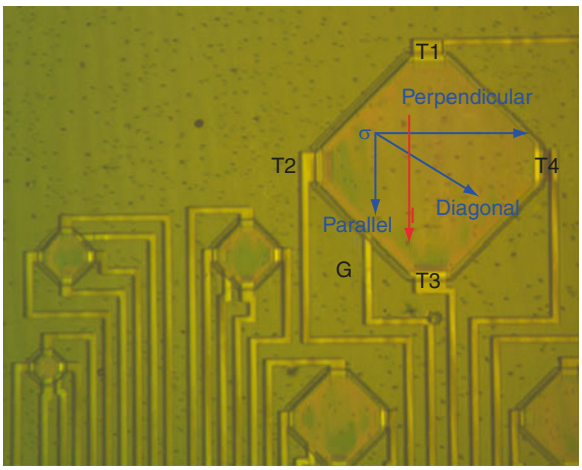
It is obvious that such flexible strain sensors will help construct a POC system for continuous health monitoring when integrated into clothing to provide patients with a safe and comfortable environment for home health care, illness prevention, and citizen medicine. The system monitors various physiological vital parameters measured from the skin including thermal, electrical, geometrical, and mechanical parameters to enable early detection, which help suppress the occurrence of acute events and complications leading to hospitalization and extended hospital treatment. From geometrical and mechanical parameters, it is expected that strain sensors constituting a wearable monitoring system are capable of collecting health data such as respiration rate and amplitude, heart rate, blood pressures, and detection of falls.

Strains sensors, usually bonded to the structures to be tested, measure the change in the resistance of conductors as directly proportional to mechanical motion (strain) experienced by the sensor. Inorganic semiconductors such as amorphous and microcrystalline silicon have recently been employed as the strain-sensing element and fabricated on a flexible film to overcome the limitation of conventional strain sensors based on metallic foils and rigid piezoresistive semiconductors [12, 13]. However, the high Young's modulus (∼200 GPa) of the inorganic materials limits the use for applications with large flexible polymeric substrates due to the large stiffness mismatch that is generated in the interface of the inorganic semiconductor elements and the flexible films. This may lead to irreversible plastic substrate deformations, degrading the sensor performance in terms of reliability and repeatability. Thus, the use of organic materials with low Young's modulus (∼5 GPa) as the sensing element is expected to minimize the induced stress concentration.

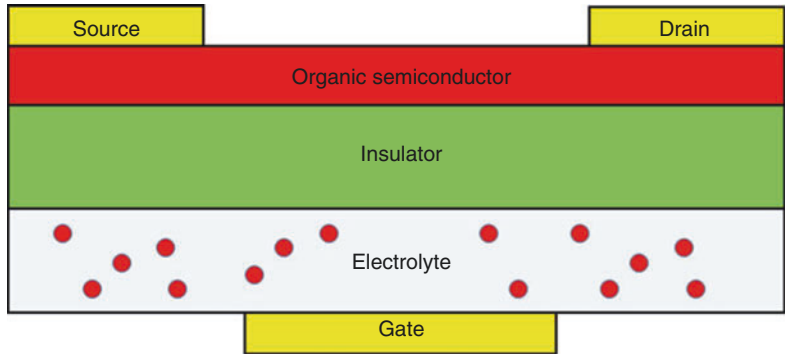
Ji et al. first developed organic flexible strain sensors that employed a thin-film transistor-like structure in conjunction with a Wheatstone bridge configuration, as shown in Fig. 4 [14]. The conductivity of the sensor is field controllable by the bias applied to the gate electrode, which tailors the sensor's sensitivity. This strain sensor consisted of five electrodes, one of which was the gate (G) electrode capable of controlling the field-induced conducting layer of pentacene to turn the sensor on and off, thereby tailoring the sensor sensitivity. This pseudo-TFT sensor operated in a depletion mode (normally on) in which it exhibited typical field effect transistor characteristics in a gate voltage range of 40 to −60 V, revealing a good saturation behavior at high drain current.

Pressure sensors, consisting of OTFTs and pressure-sensitive rubber developed by Someya et al. [15], provide the possibility of realizing a practical artificial skin thanks to the inherent flexibility and low-cost processing of organic circuits. A variation of drain current is measured under the application of different pressures from 0 to 30 kPa. Subramanian et al. recently developed printed RFID electronics including passive components (inductors and capacitors) and organic transistors for item-level tracking applications [16].

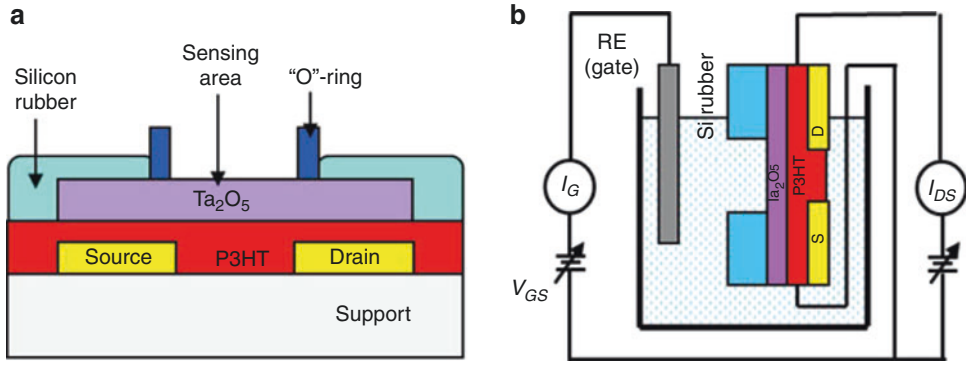
Zhu et al. [17] proposed a humidity sensor based on OTFTs that can be envisioned to monitor relative humidity in moisture-sensitive environments such as use in semiconductor manufacturing and packaging, medical, pharmaceutical, and food manufacturing. Humidity can affect the charge transport properties of organic active layers in many ways. When H_2O molecules diffuse into the crevices of polycrystalline organic materials and react with the trapped carriers in there while changing their morphology, they will alternate the electric field at the grain boundaries, which



Organic Sensors, Fig. 4 Field controllable flexible strain sensors employing a transistor-like structure in conjunction with a Wheatstone bridge configuration. *T1* and *T3* input bias, *T2* and *T4* bridge output, *G* gate, and *J* input current flow direction. Strains (σ) were applied either parallel (0°), diagonal (45°), or perpendicular (90°) to the direction of the input current flow (J)



Organic Sensors, Fig. 5 Schematic cross section of an ion-sensitive organic field effect transistor where the insulator is exposed to an electrolyte containing the analyte, and the gate electrode is separated from the insulator by the electrolyte. The electric field across the insulating gate dielectric is controlled by ions at the electrolyte–insulator surface



Organic Sensors, Fig. 6 (**a**) Schematic of the device structure used for pH sensor, and (**b**) schematic of the measurement setup [18]. This pH sensor detects Ph variations occurring in the bulk of the solution through field-enhanced conductivity of the organic semiconductor using Ta_2O_5 material that possesses a number of proton-sensitive surface sites

decrease the rate of charge transport by increasing the amount of energetic disorder through charge–dipole interactions. All these factors will affect the carrier mobility, saturation current, on/off ratio, threshold voltage, and so on. They measured the sensor output with respect to time dependence of the saturation current (I_{DS}) when the OTFT was alternatively exposed to vacuum and N_2 gas with different levels of relative humidity (RH). It turned out that the saturation current of the OTFT drops dramatically by nearly one order of magnitude when exposed to 70 % RH.

Ion-sensitive field effect transistors (ISFET) using OTFTs, as shown in Fig. 5, have been used to detect ions where the gate insulator is exposed to an electrolyte containing the analyte. The source-drain current is modulated by the electric field across the insulating gate dielectric which is controlled by ions at the electrolyte–insulator interface. Bartic et al. presented OTFTs fabricated with poly (3 hexylthiphene) regioregular (P3HT) as the semiconductor layer to be used as a pH sensor that can detect both charged and uncharged chemical species in aqueous media, as shown in Fig. 6 [18]. This pH sensor

detects pH variations occurring in the bulk of the solution through field-enhanced conductivity of the organic semiconductor using Ta_2O_5 material that possesses a number of proton-sensitive surface sites. They also realized a glucose biosensor with the same structure by anchoring an enzymatic layer of glucose oxidase (GOx) on the dielectric of the organic transistor where the biocatalytic reaction between glucose and GOx modifies the local concentration of proton on the dielectric surface leading to the variation in drain current.

Crone et al. proved that a variety of organic semiconducting materials being used in OTFTs can offer advantages to develop sensor arrays for selective gas detections [19]. Eleven different organic sensor materials based TFTs were exposed to 16 various gaseous analytes including alcohol, ketones, thions, and ring compounds while measuring the source-drain current of the TFTs which gave rise to selective patterns associated with each analyte due to different semiconductor-analyte pair response mechanisms. Crone et al. have also pointed out the fact that the OTFT sensor can be readily integrated into a circuit system with superior performance to discrete chemiresistor sensors. For instance, the ring oscillator sensor they developed exhibited a higher sensitivity to octal vapor with a faster recovery time than discrete transistor sensor [20].

Future Directions for Research

The tremendous progress in the performance of organic electronics, including OTFTs, opens wide horizons for the use of organic devices as compact sensing systems. However, developments in sensors using organic semiconductors still remain at the laboratory stage. It can be predicted that the impact of organic electronics on sensing applications will be dependent on the stability of the sensors in a real environment, the ability to design a new structure of sensors that is suitable for the integration of sensor arrays, and the cost of the sensors themselves, which determines whether this technology will move from the research state to potentially commercial phases.

Cross-References

- ▶ CMOS MEMS Biosensors
- ▶ Flexible Electronics
- ▶ Microfluidic Whole-Cell Biosensor
- ▶ Nanostructure Field Effect Transistor Biosensors
- ▶ Nanostructured Materials for Sensing
- ▶ NEMS Resonant Chemical Sensors
- ▶ Organic Actuators
- ▶ Organic Bioelectronics
- ▶ Organic Photovoltaics: Basic Concepts and Device Physics

References

1. Pope, M., Swenberg, C.E.: Electronic Processes in Organic Crystals and Polymers, 2nd edn. Oxford University Press, New York (1999)
2. Chamberlain, G.A.: Organic solar cells: a review. Sol. Cell **8**, 47–83 (1983)
3. Horowitz, G., Hajlaoui, R., Delannoy, P.: Temperature dependence of the field effect mobility of sexithiophene. Determination of the density of traps. J. Phys. III **5**, 355–371 (1995)
4. Vissenberg, M.C.J.M., Matters, M.: Theory of the field effect mobility in amorphous organic transistors. Phys. Rev. B **57**, 12964–12967 (1998)
5. Mattheus, C.C., Dros, A.B.., Baas, J., Oostergetel, G. T., Meetsma, A., de Boer, J.L., Palstra, T.T.M.: Identification of polymorphs of pentacene. Synth. Met. **138**, 475–481 (2003)
6. Lilienfeld, J.E.: Device for controlling electric current. US Patent 1,900,018, 7 Mar 1933
7. Weimer, P.K.: The TFT-new thin-film transistor. Proc. IRE **50**, 1462–1465 (1962)
8. Spear, W.E., LeComber, P.: Substitutional doping of amorphous silicon. Solid State Commun. **17**, 1193–1196 (1975)
9. LeComber, P.G., Spear, W.E., Gaith, A.: Amorphous-silicon field effect device and possible application. Electron. Lett. **15**, 179–181 (1979)
10. Ebisawa, F., Kurokawa, T., Nara, S.: Electrical properties of polyacetylene/polysiloxane interface. J. Appl. Phys. **54**, 3255–3259 (1983)
11. Mitzi, D.B., Chondroudis, K., Kagan, C.R.: Organic–inorganic electronics. IBM J. Res. Dev. **45**, 29–46 (2001)

12. Caputo, D., Cesare, G., Gavesi, M., Palma, F.: a-Si:H alloy for stress sensor application. J. Non Cryst. Solid **338–340**, 725–728 (2004)
13. Zhou, L., Jung, S., Brandon, E., Jackson, T.N.: Flexible substrate micro-crystalline silicon and gated amorphous silicon strain sensors. IEEE Trans. Electron. Dev. **53**, 380–385 (2006)
14. Ji, T., Jung, S., Varadan, V.K.: Field-controllable flexible strain sensors using pentacene semiconductors. IEEE Electron Dev. Lett. **28**, 1105–1107 (2007)
15. Takao, S., Tsuyoshi, S., Shingo, I., Yusaku, K., Hiroshi, K., Takayasu, S.: A large-area, flexible pressure sensor matrix with organic field-effect transistors for artificial skin applications. Proc. Natl. Acad. Sci. U. S. A. **101**, 9966–9970 (2004)
16. Subramanian, V., Chang, P.C., Lee, J.B., Molesa, S.E., Volkman, S.K.: Printed organic transistors for ultra-low-cost RFID application. IEEE Trans. Compon. Packag. Technol. **28**, 742–747 (2005)
17. Zhu, Z.T., Mason, J.T., Diekmann, R., Malliaras, G. G.: Humidity sensors based on pentacene thin-film transistors. Appl. Phys. Lett. **24**, 4643–4645 (2002)
18. Bartic, C., Campitelli, A., Borghs, S.: Field-effect detection of chemical species with hybrid organic/inorganic transistors. Appl. Phys. Lett. **82**, 475–477 (2003)
19. Crone, B., Dodabalapur, A., Gelperin, A., Torsi, L., Katz, H.E., Lovinger, A.J., Bao, Z.: Electronic sensing of vapors with organic transistors. Appl. Phys. Lett. **78**, 2229–2231 (2001)
20. Crone, B., Dodabalapur, A., Sarpeshkar, R., Gelperin, A., Katz, H.E., Bao, Z.: Organic oscillator and adaptive amplifier circuits for chemical vapor sensing. J. Appl. Phys. **91**, 10140–10146 (2002)

Organic Solar cells

► Nanomaterials for Excitonic Solar Cells
► Organic Photovoltaics: Basic Concepts and Device Physics

Organic/Inorganic Interactions

► Simulations of Oxide/Polymer Hybrids

Organic–Inorganic Heterojunctions

► Hybrid Solar Cells

Organic–Inorganic Photovoltaics

► Hybrid Solar Cells

Organosilanes

► Nanostructures for Surface Functionalization and Surface Properties

Orientation Sensors

► Gyroscopes

Oscillator

► Nonlinear and Parametric NEMS Resonators

Osteogenic Materials

► Synthesis of Functional Materials for Bone Regeneration

Osteoinduction Materials

► Synthesis of Functional Materials for Bone Regeneration

P

Parameter-Free Study of Electrochemical Interfaces

► Electrochemical Interfaces for Energy Storage and Conversion

Parametric

► Nonlinear and Parametric NEMS Resonators

Parametrically Excited Nonlinear MEMS Resonators

► Nonlinear Parametric MEMS

Passive Pumping

► Capillary Flow

Patterned Hydrogels

► Nanoengineered Hydrogels for Cell Engineering

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Patterning

► Nanostructures for Surface Functionalization and Surface Properties

PBPK Modeling

► In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles

Perfluorocarbon Nanoparticles

Patrick M. Winter[1], Gregory M. Lanza[2] and Samuel A. Wickline[2]
[1]Department of Radiology, Imaging Research Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[2]C-TRAIN Labs, Washington University, St. Louis, MO, USA

Synonyms

PFC nanoparticles; PFOB nanoparticles

Definition

Perfluorocarbon (PFC) nanoparticles represent a multifunctional platform consisting of a PFC core

encapsulated within a phospholipid monolayer. The particles have been developed for magnetic resonance (MR) molecular imaging to noninvasively map the expression of important biomarkers of disease.

Introduction

Perfluorocarbon (PFC) nanoparticles consist of a liquid PFC core encapsulated by a phospholipid monolayer. PFCs represent a family of modified hydrocarbon molecules in which the hydrogens are replaced with fluorine atoms [1–3]. The carbon–fluorine bond is very stable and biologically inert due to the dense electron cloud, which prevents chemical reactions with other compounds [2]. PFCs with molecular weights between 460 and 520 are biocompatible, showing no toxicity, carcinogenicity, mutagenicity, or teratogenic effects. PFCs have tissue half-life residencies ranging from 4 to 65 days. They are not metabolized. Instead, they are reintroduced into the blood circulation from clearance organs, such as the liver, by lipid carriers and are exhaled via the lungs [2]. PFC nanoparticles have been formulated for a wide range of molecular imaging and targeted drug delivery applications. The particles are typically 250 nm in diameter, which is large enough to constrain them to the intravascular space. As a result, PFC particles are often targeted to intraluminal biomarkers expressed by endothelial cells, such as integrins, selectins, or adhesion molecules [2, 3].

Magnetic resonance (MR) molecular imaging aims to identify and map the expression of important biomarkers on a cellular scale. Typical MR imaging techniques lack sufficient resolution and sensitivity to directly detect these molecular signatures due to their very low concentrations in vivo. To overcome this limitation, nanoparticle agents are often employed to specifically bind to the biomarker of interest, accumulate at the site of pathology, and generate sufficient image contrast. Nanoparticles offer a large surface area for the incorporation of multiple binding ligands (to improve targeting efficacy) and multiple paramagnetic chelates (to amplify the signal enhancement). In some cases, the nanoparticle may also serve as a drug delivery agent, providing diagnostic and therapeutic information via noninvasive MR imaging.

A number of other medical imaging modalities are capable of molecular imaging, including positron emission tomography (PET), single photon emission computed tomography (SPECT), and optical imaging [4]. Like MR imaging, these modalities typically rely on specifically designed contrast agents that bind to the molecular biomarker of interest. For PET and SPECT imaging, the contrast agents consist of a radioactive element for detection. In optical imaging, the contrast agent will utilize fluorescence or bioluminescence for mapping expression of cellular receptors. MR imaging, however, has certain important advantages over these other modalities for molecular imaging applications. MR imaging offers higher spatial and temporal resolution than PET or SPECT and provides greater tissue penetration than optical imaging. This allows MR imaging to combine anatomical and/or functional information with molecular imaging, much like the multimodality benefits of PET/CT. In addition, MR imaging may be better suited for serially tracking disease progression or therapeutic response compared to nuclear imaging due to concerns of cumulative radiation dose.

Molecular imaging with PFC nanoparticles has been applied to a wide array of disease biomarkers, therapeutic drugs, and imaging modalities [2, 3]. PFC nanoparticle formulations have been developed for ultrasound imaging [5], computed tomography [6], nuclear imaging [7], optical imaging [8–10], MR [3], as well as multimodality applications. For example, combined ^{99m}Tc SPECT and MR imaging of $\alpha_v\beta_3$-integrin-targeted PFC nanoparticles has been used to map tumor angiogenesis with high sensitivity and high resolution [7]. The $\alpha_v\beta_3$-integrin is a heterodimeric transmembrane glycoprotein expressed by numerous cell types, including endothelial cells, macrophages, platelets, lymphocytes, smooth muscle cells, and tumor cells [2]. Endothelial expression of $\alpha_v\beta_3$-integrin is required for the migration and invasion of vascular endothelial cells, and therefore, serves as a

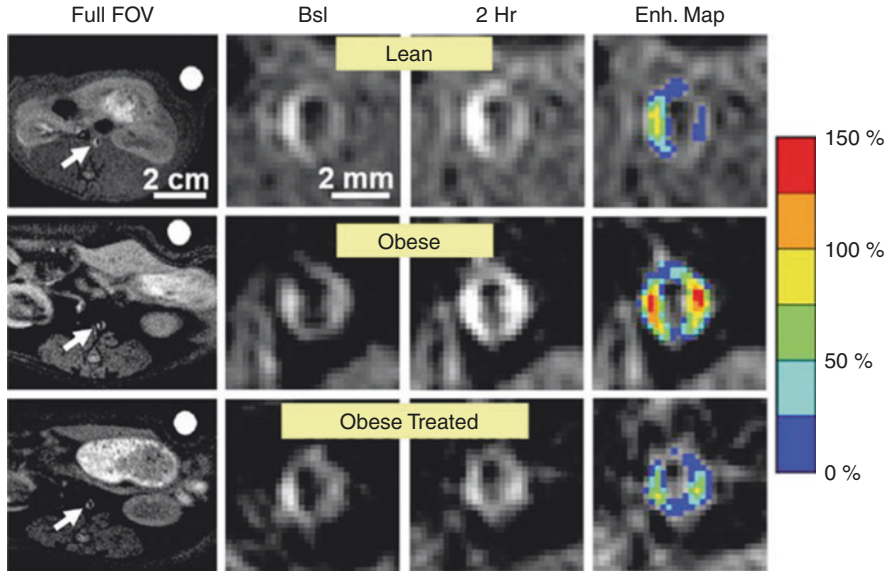
biomarker of angiogenesis [11]. Since PFC nanoparticles are constrained to the vasculature, they are not exposed to nonendothelial integrin-expressing cells, which improves binding specificity to the neovasculature [2]. Nuclear imaging offers very high sensitivity, which is preferred for the detection of occult tumors or metastases within large tissue regions. However, the limited spatial resolution of nuclear imaging precludes detailed characterization of the pathology. Incorporating a ^{99m}Tc nuclear imaging agent onto $\alpha_v\beta_3$-integrin-targeted gadolinium nanoparticles yielded a high-sensitivity nuclear signal that was used to guide further high-resolution MR imaging and characterization of tumor angiogenesis [7]. MR molecular imaging demonstrated a coherent, asymmetric, patchy pattern of angiogenesis along the outer aspects of the tumor mass, consistent with an angiogenic phenotype.

In the case of ultrasound imaging, PFC nanoparticles represent a significant departure from conventional contrast agent design. Most ultrasound contrast agents consist of gas bubbles to maximize the backscatter of ultrasound waves and thereby increase the image contrast. However, these bubbles are quite large (more than 1 μm in diameter) and are quickly entrapped within the lungs. This severely limits the practical application of bubbles in molecular imaging because they do not circulate long enough to bind to biomarkers of disease and accumulate at the sites of pathology. PFC nanoparticles, however, have a long circulatory half-life and are inherently echogenic due the differences in the speed of sound between water and the PFC core. This ultrasound backscatter is typically not sufficient to visualize particles in the blood stream. Instead, the particles must accumulate at the target site, forming an "acoustic mirror" to generate significant image contrast. Targeted PFC nanoparticles have been used for molecular imaging of fibrin in acute arterial thrombus, high-frequency intravascular imaging, and imaging with novel ultrasound receivers, called information theoretical detectors [5].

Along with MR molecular imaging, PFC nanoparticles have been employed to label cells and noninvasively track implantation and migration of cell therapies [1, 3]. While MR molecular imaging with PFC nanoparticles typically utilizes paramagnetic chelates for detection with ^{1}H MR, tracking of cells labeled with PFC nanoparticles often relies on ^{19}F MR. ^{19}F has a relatively high sensitivity, 83 % compared to ^{1}H, and there is virtually no native background signal. Thus, the PFC core can provide a definitive and quantitative MR signature [1]. PFC nanoparticles can be internalized by stem cells without the need of a transfection agent and ^{19}F MR imaging can quantitatively track the cells in vivo [1]. The high sensitivity of ^{19}F MR allows the in vivo detection of as few as 2,000 labeled cells. For some applications, modification of the surfactant surface of the particles improves cellular uptake. For example, PFC nanoparticles formulated with cationic phospholipids on the particle surface can be effectively internalized by neural stem cells [1].

Molecular Imaging of Therapeutic Response

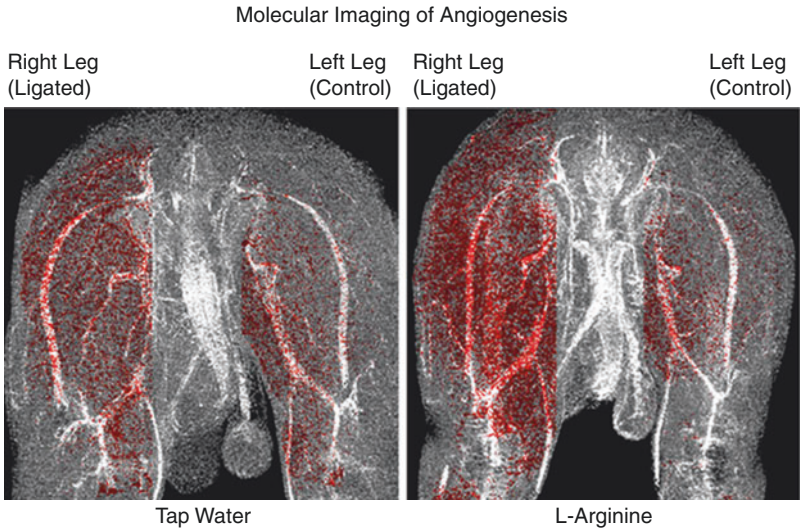
One potential clinical application of MR molecular imaging with PFC nanoparticles is to serially map biomarkers of disease progression and/or therapeutic response. Noninvasive imaging could be used to monitor the biochemical signatures of disease, providing earlier and more specific detection of pathology compared to tracking symptomology or the occurrence of clinical events. For example, molecular imaging of angiogenesis can serve as a biomarker of atherosclerosis and can be used to monitor the physiological effects of a range of therapeutic treatments [3]. Paramagnetic $\alpha_v\beta_3$-integrin-targeted PFC nanoparticles have been utilized to characterize angiogenesis in the aortic wall of an atherosclerotic rat model and to monitor the effects of benfluorex (an appetite suppressant) on aortic angiogenesis [12]. The JCR:LA-cp rat is a model of metabolic syndrome characterized by obesity, insulin resistance, hyperlipidemia, and vasculopathy. The obese rats displayed higher food consumption and body weight, as well as increased insulin, leptin, cholesterol, and triglyceride levels compared to the lean controls. In conjunction with this typical metabolic syndrome



Perfluorocarbon Nanoparticles, Fig. 1 Full field-of-view ("Full FOV") images of JCR rats showing large abdominal fat deposits in the obese (*middle*) and obese-treated (*bottom*) animals that were absent in the lean (*top*) rat (*arrows* denote aortas). Magnified views of the abdominal aorta in lean, obese, and obese-treated rats at baseline ("Bsl") and 2-h ("2 Hr") post injection of targeted nanoparticles displaying signal enhancement in the aorta wall. Percent signal enhancement maps ("Enh. Map," false-colored from *blue* to *red*) demonstrate higher MR signal enhancement in the untreated obese animal, indicating active angiogenesis supporting development of atherosclerotic lesions (Reprinted with permission from Cai et al. [12])

phenotype, MR molecular imaging revealed increased aortic angiogenesis in the obese animals, indicative of actively developing atherosclerotic plaques in the arterial wall (Fig. 1). Benfluorex reduced food consumption, body weight, insulin, and leptin in obese-treated rats, but had no effect on serum cholesterol or triglyceride levels. In a similar fashion, benfluorex reduced aortic angiogenesis in the obese-treated animals to the same level as the control lean rats. These molecular imaging results were corroborated by histological measurements of microvascular density. Therefore, MR molecular imaging with $\alpha_v\beta_3$-integrin-targeted PFC nanoparticles can noninvasively monitor atherosclerosis in a rat model of obesity and can monitor the neovascular effects of an appetite suppression therapy.

However, atherosclerosis does not always lead to an increase in angiogenesis. In a rabbit model of peripheral vascular disease, hypercholesterolemia causes a reduction in the angiogenic response to chronic muscle ischemia, which can be normalized with a pro-angiogenic therapy, such as L-arginine supplementation [13]. MR molecular imaging was performed 10 days after femoral artery ligation with paramagnetic $\alpha_v\beta_3$-integrin-targeted PFC nanoparticles. Molecular imaging revealed diffuse angiogenesis throughout the ischemic limb for both L-arginine-treated and untreated groups (Fig. 2). However, the angiogenic response was significantly increased in L-arginine-treated animals compared to the untreated controls. Furthermore, the integrin-targeted formulation produced two times higher MR enhancement compared to nontargeted particles, demonstrating improved identification of angiogenic vasculature with biomarker targeting. X-ray angiography was performed 40 days postligation to measure the development of collateral vessels feeding the ischemic muscle tissue. More collateral vessels were observed in L-arginine-treated rabbits compared to the untreated controls. Histologic staining of muscle capillaries revealed a denser pattern of microvasculature in L-arginine-treated animals, confirming



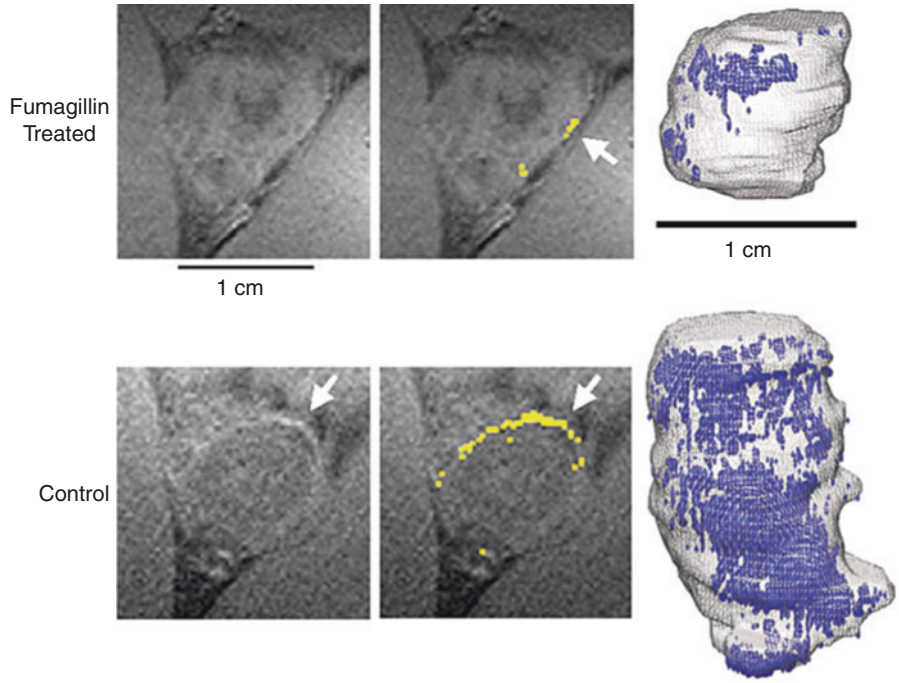
Perfluorocarbon Nanoparticles, Fig. 2 Molecular imaging of angiogenesis shows highly diffuse enhancement throughout the ligated (*right*) leg of cholesterol-fed rabbits with only slight enhancement of the control (*left*) leg. The L-arginine-treated rabbit (*right panel*) shows a more dense distribution of angiogenesis in the ligated limb compared with the control animal (*left panel*) (Reprinted with permission from Winter et al. [13])

the MR and X-ray imaging results. Noninvasive molecular imaging with PFC nanoparticles could be used for early detection of response to angiogenic therapies directed at ischemic diseases, such as peripheral vascular disease or myocardial infarction, to enable personalized optimization of treatment strategies.

Targeted Antiangiogenic Therapy

In the realm of targeted drug delivery, PFC nanoparticles have been utilized to carry a number of antiangiogenic drugs. In particular, the potent antiangiogenic drug, fumagillin suppresses angiogenesis by inhibition of methionine aminopeptidase 2 (MetAP2) [9]. Targeting the particles to a biomarker of angiogenesis, such as the $\alpha_v\beta_3$-integrin, provides a means to specifically direct potent drugs to the sites of pathology, improving the therapeutic efficacy and decreasing the occurrence of unwanted side effects. Combining a specifically targeted molecular imaging agent with a therapeutic drug could benefit patients in a wide range of clinical scenarios. The imaging agent would allow confirmation and quantification of local drug delivery and enable personalization of treatment protocols based on the pharmacokinetics in the target tissue. In addition, specific targeting of the agent could improve uptake and retention of the drug at the sites of disease while lowering the exposure to other susceptible organs and reducing the occurrence of side effects.

Paramagnetic PFC nanoparticles targeted to the $\alpha_v\beta_3$-integrin have been used for sensitive detection of angiogenesis in Vx2 tumors implanted in New Zealand white rabbits [2]. In vivo competition studies demonstrated that the targeted particles specifically bind to the $\alpha_v\beta_3$-integrin. Furthermore, MR molecular imaging with the targeted nanoparticles was able to differentiate between growing tumors and inflammatory lesions resulting from rejected cancer cells. These tumor cell remnants appeared identical to the viable growing tumors on standard T2-weighted MR images, but molecular imaging of angiogenesis showed negligible neovasculature in these masses. Antiangiogenic therapy has also been demonstrated in this animal model using $\alpha_v\beta_3$-integrin-targeted fumagillin nanoparticles [14]. Treatment with $\alpha_v\beta_3$-integrin-targeted fumagillin nanoparticles reduced the tumor volume significantly compared with animals given nontargeted fumagillin nanoparticles, $\alpha_v\beta_3$-integrin-targeted nanoparticles without drug or saline (Fig. 3). The MR imaging data was used to generate a 3D map of tumor angiogenesis, which showed a coherent asymmetric peripheral distribution characterized by dense neovessel regions interspersed with a finer, reticular pattern.



Perfluorocarbon Nanoparticles, Fig. 3 *Left*: Molecular imaging of targeted antiangiogenic therapy (fumagillin) in rabbit Vx-2 tumors showing diminished MR contrast enhancement in the treated animal (*top*) versus the control animal (*bottom*). *Middle*: MR enhancement, color coded in *yellow* (*arrows*), demonstrating reduced tumor angiogenesis in the fumagillin-treated animal. *Right*: 3D maps of tumor angiogenesis showing an asymmetric distribution of neovasculature (*blue*) that is dramatically reduced in the fumagillin-treated rabbit (Reprinted with permission from Winter et al. [14])

Another possible biomarker of angiogenesis is Robo4, an endothelial specific guidance receptor that is expressed on tumor neovasculature but not mature vessels [15]. Paramagnetic PFC nanoparticles were coupled to a Robo4 antibody and evaluated in a mouse melanoma tumor model [15]. Both Robo4-targeted and $\alpha_v\beta_3$-integrin-targeted nanoparticle formulations generated greater MR image contrast than particles lacking a targeting ligand or coupled to scrambled targeting ligands. Microscopic analysis of particle binding was performed with Robo4-targeted rhodamine nanoparticles and $\alpha_v\beta_3$-integrin-targeted AlexaFluor 488 nanoparticles. Binding of $\alpha_v\beta_3$-integrin-targeted particles was generally more prevalent than the binding of Robo4-targeted nanoparticles, particularly in the tumor periphery. Thus, PFC nanoparticles can be used for both macroscopic MR molecular imaging as well as light microscopy to define cell binding of the particles and the coexpression of different cellular receptors.

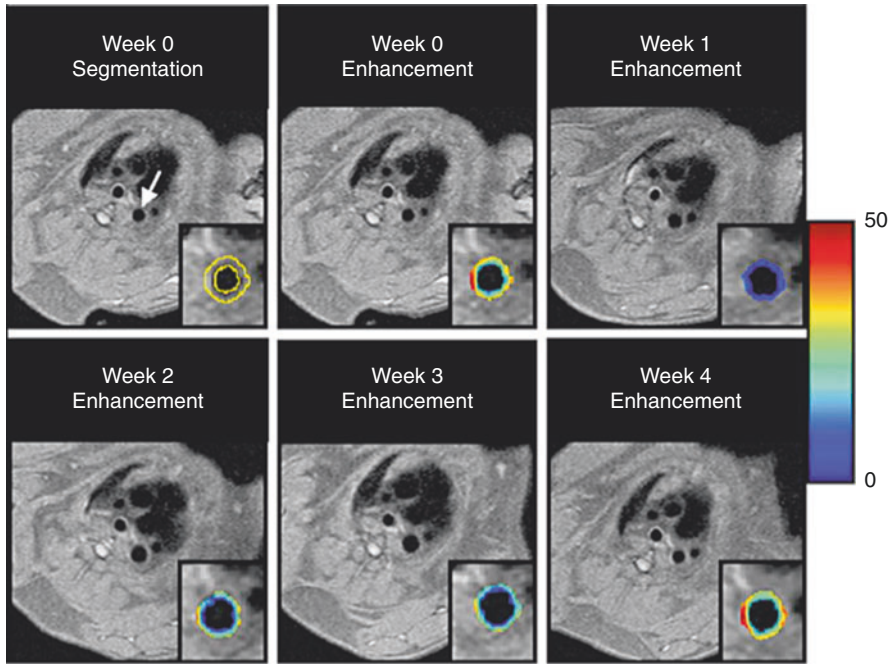
MR molecular imaging of angiogenesis has also been demonstrated in human melanoma tumors (40 mm^3) implanted in immunocompromised mice using $\alpha_v\beta_3$-integrin-targeted PFC nanoparticles [2]. Again, in vivo competition studies demonstrated the high specificity of ligand-directed targeting, which was further corroborated microscopically using $\alpha_v\beta_3$-integrin-targeted fluorescent PFC nanoparticles. As an alternative to targeting the $\alpha_v\beta_3$-integrin for mapping angiogenesis, the $\alpha_5\beta_1$-integrin has been explored as a biomarker of tumor neovasculature [8]. The $\alpha_5\beta_1$-integrin is an important adhesion molecule, which regulates endothelial cell migration and survival during angiogenesis. The $\alpha_5\beta_1$-integrin is not expressed on normal quiescent blood vessels, but its expression is induced on tumor blood vessels and in response to angiogenic

factors, including basic fibroblast growth factor, interleukin-8, tumor necrosis factor-alpha, and the angiomatrix protein Del-1. The utility of a dual $\alpha_5\beta_1(\alpha_v\beta_3)$-integrin targeting ligand for molecular imaging and targeted delivery of fumagillin was compared with the use of a targeting ligand specific for only the $\alpha_v\beta_3$-integrin. MR signal enhancement maps revealed a sparse network of neovessels predominantly located in the tumor periphery. Treatment with $\alpha_5\beta_1(\alpha_v\beta_3)$-integrin-targeted nanoparticles containing fumagillin further decreased the already small amount of neovessel enhancement to an almost negligible level; however, the tumor volume was not decreased. Since this tumor model displayed a very sparse neovasculature and was relatively slow growing, it is not surprising that an antiangiogenic therapy did not significantly decrease the tumor size. Clinically, MR molecular imaging of angiogenesis could be used to characterize small tumors, such as breast, lung, or colon, and stratify patients into antiangiogenic therapies based on the abundance of neovascular biomarkers. Such a procedure for screening potential patients may yield improved therapeutic benefits with decreased risk of side effects.

Similar to the growth of tumors, the development of atherosclerotic plaques requires the recruitment of new blood vessels to meet the high metabolic demands of the abnormal tissue. In atherosclerosis, angiogenesis occurs in the vasa vasorum of large caliber arteries, such as the aorta, carotids, and coronaries. Since antiangiogenic therapies can slow the progression of cancer, it is reasonable to investigate if these treatments could be employed to normalize atherosclerotic vessels. The acute antiangiogenic response to $\alpha_v\beta_3$-integrin-targeted paramagnetic nanoparticles containing fumagillin has been studied in atherosclerotic rabbits [16]. MR imaging was used to estimate drug deposition in the aortic wall. One week later, the level of neovascular $\alpha_v\beta_3$-integrin expression was assessed using $\alpha_v\beta_3$-targeted paramagnetic nanoparticles without fumagillin. Aortic areas displaying high MR enhancement at the time of treatment had the largest subsequent reduction in $\alpha_v\beta_3$-integrin-targeted MR signal 1 week later, suggesting that combining imaging with therapy may not only confirm and quantify the local delivery of chemotherapeutics but may also provide early predictions of the subsequent treatment effects.

In addition to the acute response to therapy, MR molecular imaging could provide serial monitoring of the end organ effects with the ultimate goal of optimizing drug treatment regiments and customizing patient protocols. The antiangiogenic effect of atorvastatin with and without targeted delivery of fumagillin was serially monitored for 8 weeks by MR imaging in atherosclerotic rabbits [9]. Atorvastatin was dosed continuously via incorporation into the feed, while fumagillin treatment was provided once every 4 weeks. During 8 weeks of study, atorvastatin treatment did not reduce $\alpha_v\beta_3$-integrin expression in the aortic wall. Fumagillin treatment, on the other hand, resulted in a transient (2–3 weeks) reduction in image enhancement, indicating successful antiangiogenic treatment (Fig. 4). Combining the fumagillin and atorvastatin treatments yielded a persistent decrease in the MR signal, suggesting that chronic statin treatment could be used to prolong the effects of discrete doses of an antiangiogenic agent. Histology revealed that the MR image enhancement was strongly correlated to the microvascular density in the aortic wall in a logarithmic fashion. An effective and sustained antiangiogenic treatment may halt plaque progression and stabilize atheromas by decreasing intramural hemorrhage and inflammation.

One of the benefits of using a molecular imaging contrast agent for targeted drug delivery is that imaging can provide a noninvasive measurement of drug deposition at the site of pathology and elimination from the body. Paramagnetic $\alpha_v\beta_3$-integrin-targeted PFC nanoparticles have been utilized to quantify the pharmacokinetics of the particles in both the blood pool as well as the target tissue in an atherosclerotic rabbit model [3]. Noninvasive monitoring of the particle biodistribution and elimination could be used as a surrogate marker for the tissue levels of lipophilic drugs incorporated in the nanoparticle formulation. Typically, pharmacodynamic measurements rely on blood sampling without consideration of the drug accumulation in the target



Perfluorocarbon Nanoparticles, Fig. 4 Serial imaging of aortic wall angiogenesis for monitoring targeted antiangiogenic therapy with fumagillin nanoparticles. *Top*: Black blood image of the thoracic aorta (*arrow*) prior to antiangiogenic treatment (Week 0) demonstrating segmentation of the vessel wall (*outlined in yellow*) and patchy areas of high angiogenesis within the aortic wall (color-coded as percent enhancement of MR signal). At week 1, the MR enhancement is drastically reduced due to the antiangiogenic effect of fumagillin. *Bottom*: The level of signal enhancement gradually increases at Weeks 2 and 3 after fumagillin treatment. At Week 4, the level of enhancement is practically identical to the Week 0 image (Reprinted with permission from Winter et al. [9])

tissue. Noninvasive mapping of drug concentrations in the end organ could provide significant clinical advantages in monitoring targeted therapeutic agents and determining the local drug concentrations beyond the blood pool. Atherosclerotic rabbits were injected with either $\alpha_v\beta_3$-targeted or nontargeted paramagnetic nanoparticles at a dose of 1 mL/kg (0.0046 mmol Gd^{3+}/kg), which is 20-fold lower than the standard dose of conventional gadolinium agents (0.1 mmol Gd^{3+}/kg). Blood sampling was performed to determine the bulk pharmacokinetic behavior of the nanoparticles, and T1-weighted MR images of the descending aorta were collected over 24 h to ascertain binding to the angiogenic microvasculature in the vessel wall.

A three-compartment pharmacokinetic model was developed to describe the in vivo behavior of the nanoparticles in both the blood pool as well as the aortic wall. Compartments 1 and 2 represent the bulk distribution of nanoparticles throughout the blood stream. The third compartment consists of the vasa vasorum of the aortic wall where the nanoparticles can specifically bind to the $\alpha_v\beta_3$-integrin. The constants that describe the transfer into and out of this compartment, k_{13} and k_{31}, are lumped parameters that are used to describe both passive transfer and active binding.

Blood sampling after IV injection of nanoparticles demonstrated the half-life for the distribution phase was 20.2 min, while the elimination half-life was 11.9 h. MR imaging of the aortic wall revealed that the targeted nanoparticle agent generated at least twofold higher signal enhancement at 0.5 and 1 h post injection than the nontargeted nanoparticles. Fitting both the blood clearance and tissue uptake data with a three-compartment model showed no significant differences in k_{12}, k_{21}, v_1, or k_{31} between the

targeted and nontargeted groups. However, k_{13} (representing transfer from the blood into the aortic wall) and k_e (elimination from the blood pool) were significantly higher for the targeted formulation. While k_{13} represents the passive transfer from the blood into the third compartment in a standard three-compartment model, the nanoparticles are not just passively transferred but also actively binding within the third compartment. These two effects cannot be differentiated with this model and combine together, resulting in an apparent increase in k_{13} for the targeted nanoparticles compared to the nontargeted formulation. The pharmacokinetic parameters were used to calculate the tissue exposure to the nanoparticles. The local tissue concentration of targeted nanoparticles was doubled compared to nontargeted particles (1.64 vs. 0.84 nmol Gd^{3+}/g tissue/h, respectively). The amount of gadolinium that reached the aortic vasa vasorum was 0.18 % of the total dose for nontargeted particles and 0.38 % for targeted nanoparticles. These results demonstrate that MR can be used to noninvasively track binding of a molecularly targeted contrast agent, which has previously only been accomplished with nuclear and optical imaging. The gadolinium chelate on the particle surface could provide a surrogate marker for tracking the biodistribution of a lipophilic drug payload [9, 16, 17]. Rather than relying on blood pool measurements of drug concentrations, noninvasive imaging could serially monitor the actual concentration of the drug in the end organ and lead to more accurate modeling of therapeutic response.

Targeted Antiproliferative Therapy

In addition to antiangiogenic drugs, PFC nanoparticles have been formulated to deliver drugs that inhibit the pathological proliferation of various cell types. The progression of both cancer and atherosclerosis typically involve uncontrolled cell division and a wide range of antiproliferative drugs have been developed to fight these diseases. For example, smooth muscle cells line the luminal aspect of arteries and are involved in the progression of atherosclerosis. Tissue factor, a transmembrane glycoprotein with a prominent role in angiogenesis, thrombosis, cell signaling, hemostasis, and mitogenesis, is expressed on smooth muscle cells in response to vascular injury. MR molecular imaging of tissue factor targeted PFC nanoparticles showed that the targeted particles bind to cultured smooth muscle cells at a concentration of 468 pM, while nontargeted particles only reach a concentration of 88 pM [3]. The MR measurements were validated by gas chromatography of the samples, revealing a particle concentration of 530 pM for the targeted nanoparticles and 111 pM for the nontargeted nanoparticles. Smooth muscle cell proliferation portends restenosis following coronary angioplasty. Therefore, the utility of PFC nanoparticles for delivery of antiproliferative drugs, including doxorubicin and paclitaxel, to smooth cells has been studied [2]. Exposure of the cells to targeted doxorubicin or paclitaxel nanoparticles significantly inhibited proliferation. Including the targeting ligand on the particle surface significantly increased the antiproliferative effect, particularly for paclitaxel. In vitro dissolution studies revealed that nanoparticle drug release persisted over 1 week. Targeted antiproliferative results were dependent on the hydrophobic nature of the drug and noncovalent interactions with other surfactant components. MR molecular imaging of the bound nanoparticles was demonstrated with both ^{1}H and ^{19}F MR imaging. A targeted paramagnetic nanoparticle carrying an antiproliferative payload to specific cell types could provide a novel, MR-visualizable, and quantifiable drug delivery system for the prevention of restenosis after angioplasty.

Clinically, drug-eluting stents are often used to locally release antiproliferative drugs for inhibiting restenosis following angioplasty. However, in-stent thrombosis can occur with these devices due to endothelial damage from the antiproliferative drugs. An alternative approach has been pursued with PFC nanoparticles loaded with antiproliferative agents and targeted to cellular biomarkers or the extracellular matrix [18]. Furthermore, these nanoparticle agents can be used to visualize ruptures in the artery wall

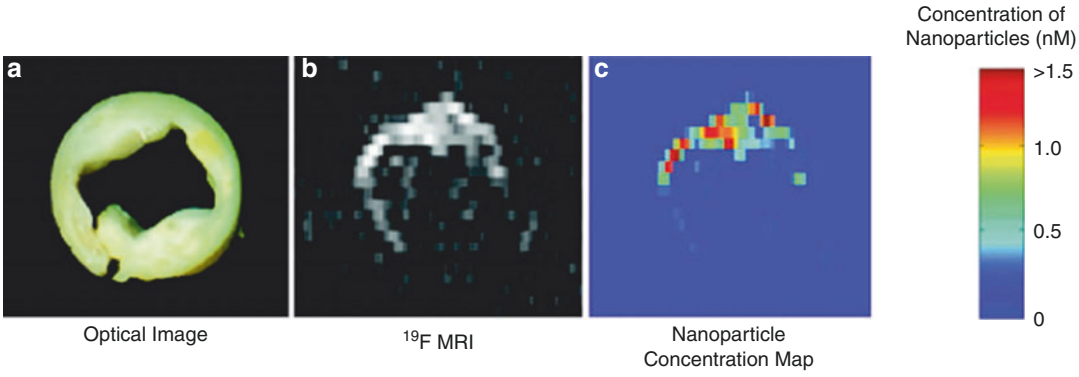
resulting from the angioplasty balloon. Paramagnetic $\alpha_v\beta_3$-integrin-targeted PFC nanoparticles containing rapamycin were used to locally deliver a drug payload through microfractures created by balloon overstretch injury. The local delivery of rapamycin inhibited arterial stenosis without impeding endothelial healing [18]. MR imaging demonstrated that the integrin-targeted nanoparticles remained bound to the injured arterial wall 30–40 min after balloon overstretch and reestablishment of blood flow. Two weeks after injury, MR imaging showed significant luminal stenosis in vessels treated with control formulations: $\alpha_v\beta_3$-integrin-targeted nanoparticles without drug, nontargeted nanoparticles with rapamycin or saline. In vessels treated with $\alpha_v\beta_3$-integrin-targeted rapamycin nanoparticles, no lumen irregularities were observed on MR angiograms. Histology showed $\alpha_v\beta_3$-integrin-targeted rapamycin nanoparticles decreased neointimal proliferation by 40 % compared to controls. Endothelial healing was evaluated 1, 7, 14, and 28 days after balloon injury and showed no differences between $\alpha_v\beta_3$-integrin-targeted rapamycin nanoparticles and the controls.

Another therapeutic approach to inhibit cellular proliferation uses host defense peptides that embed in the phospholipid cell membrane to create pores and disrupt normal cellular physiology. Melittin is a cytolytic peptide derived from honeybee venom that ruptures lipid membranes. PFC nanoparticles conjugated with melittin have been used as a therapeutic delivery vehicle in cancer models [19]. Melittin PFC nanoparticles were injected intravenously in mice bearing MDA-435 tumors [19]. Ultrasound imaging showed that treated tumors grew 25 % slower than controls. In a mouse melanoma model, four intravenous injections of melittin nanoparticles reduced the tumor size by 80 % compared to control injections of saline or nanoparticles lacking melittin. In a genetically engineered model of squamous carcinoma, treatment with integrin-targeted melittin nanoparticles resulted in an 80 % reduction in dermal papillae, the site of dysplasia in this model. In cultured cells, free melittin causes necrosis resulting in the release of lactate dehydrogenase (LDH) due to the disruption of cell membranes and cytochrome c indicating mitochondria damage. Cultured cells exposed to targeted melittin nanoparticles showed cytochrome c release, but only minimal LDH, suggesting a shift to apoptotic cell death rather than necrotic cell death. Thus, drug delivery with PFC nanoparticles could be utilized to modulate the therapeutic mechanism of action.

^{19}F MR Molecular Imaging

MR detection of paramagnetic PFC nanoparticles typically relies on the effect of gadolinium on the ^{1}H MR signal. The paramagnetic nature of gadolinium increases the water relaxation rate and generates signal enhancement on T1-weighted MR images. An alternative approach achieves direct detection of the PFC core of the nanoparticles via ^{19}F MR imaging [1, 3]. ^{19}F has a relatively high sensitivity, 83 % compared to ^{1}H, and there is virtually no native background signal. Thus, the PFC core can provide a definitive and quantitative MR signature [1]. For example, ^{19}F MR of fibrin-targeted PFC nanoparticles could be used to map the formation of thrombi on ruptured atherosclerotic plaques and quantify the extent of ruptures in the fibrous cap. Human carotid endarterectomy samples were treated with fibrin-targeted paramagnetic nanoparticles and imaged at 4.7 T [20]. ^{1}H MR imaging showed high levels of signal enhancement along the luminal surface due to nanoparticle binding to fibrin deposits (Fig. 5). A ^{19}F projection image of the artery showed an asymmetric distribution of nanoparticles around the vessel wall corroborating the ^{1}H signal enhancement. Co-registration of the quantitative nanoparticle map with the ^{1}H image permitted visualization of both anatomical and pathological information in a single image. Combining information from ^{1}H and ^{19}F MR imaging could allow prediction of subsequent occlusion or distal embolization from unstable or disrupted plaques, and aid clinical decision-making for acute invasive intervention versus pharmaceutical therapies.

PFC nanoparticles can be formulated with different perfluorochemicals to yield particles that



Perfluorocarbon Nanoparticles, Fig. 5 Direct detection of fibrin-targeted PFC nanoparticles utilizing ^{19}F MR imaging. (**a**) Optical image of a cross section of a human carotid endarterectomy sample showing moderate luminal narrowing as well as several atherosclerotic lesions. (**b**) A ^{19}F projection image of the carotid artery sample showing high signal along the lumen due to nanoparticles bound to fibrin. (**c**) Concentration map of bound nanoparticles on the carotid artery (Reprinted with permission from Morawski et al. [20])

can be differentiated by ^{19}F MR. Fibrin-targeted nanoparticles consisting of perfluorooctylbromide (PFOB) or perfluoro-15-crown-5-ether (CE) have been distinguished based on the unique chemical shifts of each compound [1]. Both ^{19}F imaging and spectroscopy could identify PFOB and CE nanoparticles. The signal to noise for PFOB was lower than CE (10 vs. 25, respectively), presumably due to the single CE peak (20 equivalent fluorine atoms) compared to the multiple PFOB peaks (17 fluorine atoms distributed over 5 peaks). A clear linear relationship between the ^{19}F signal intensity and nanoparticle concentration was demonstrated for both PFOB and CE using both imaging and spectroscopy. Multiple PFC nanoparticle agents could be used to target different epitopes and achieve a noninvasive analogy to immunohistochemistry. For example, simultaneous quantification of angiogenesis in the vessel wall and fibrin deposition on the plaque cap could be used to evaluate the pathophysiological stage of an obstructive atherosclerotic lesion.

Tissue inflammation is another biological marker of unstable atherosclerotic plaques, inducing the release of cytokines and upregulating the production of proteins like vascular cell adhesion molecule-1 (VCAM-1). PFC nanoparticles targeted to VCAM-1 were injected into genetically engineered ApoE$^{-/-}$ mice, which mimic the clinical progression of atherosclerosis, to map inflammation [3]. These mice display focal inflammation and macrophage infiltration in the kidneys. To definitively identify nanoparticle binding in the kidneys, ^{19}F MR imaging was performed on a 11.7 T research scanner 2 h after injection of VCAM-1 targeted PFC nanoparticles. The ^{19}F signal arising from the PFC core provided an unambiguous marker of particle accumulation, without the innate signal variations that can confound typical ^{1}H signal enhancement with traditional paramagnetic or superparamagnetic MR contrast agents. VCAM-1-targeted nanoparticles accumulated in ApoE$^{-/-}$ kidneys to a greater extent than nontargeted nanoparticles (3.7 billion particles/g of tissue vs. 0.9 billion particles/g). The uptake of targeted nanoparticles was also higher in the kidneys of ApoE$^{-/-}$ mice compared to non-ApoE$^{-/-}$ controls (3.7 vs. 1.6 billion particles/g). Control animals displayed no significant difference in the uptake of targeted versus nontargeted nanoparticles (1.6 vs. 1.5 billion particles/g).

Typically, targeted MR contrast agents generate image enhancement as a result of both nonspecific blood pool signal as well as the specific binding of the agent to the biomarker of interest. Separating out these two contributions can be very difficult to achieve in vivo. One method to suppress the nonspecific signal utilizes diffusion-weighted ^{19}F spectroscopy to null the

signal arising from moving particles, representing the unbound fraction in the blood pool, while retaining the signal from stationary particles that are specifically bound to the target epitope [1]. A genetically engineered mouse model of squamous cell cancer, derived by incorporating human papilloma virus into the mouse genome, was studied with a 11.7 T research scanner. Transgenic and age-matched control mice were scanned 90 min after injection of $\alpha_v\beta_3$-integrin-targeted PFC nanoparticles to assess angiogenesis in these precancerous lesions. Both the transgenic and control mice displayed decreased ^{19}F signal with increasing b-values. However, 60–100 % of the ^{19}F signal remained at b-values near 60,000 s/mm^2 in the transgenic animals, while no detectable ^{19}F signal was observed in control mice at b-values of 1,500 s/mm^2. The calculated ADC of PFC nanoparticles was 33.1 m^2/s in the transgenic mice, significantly lower than the 19,563 m^2/s ADC in the controls.

Angiogenesis is a prominent feature in the progression of aortic valve stenosis. Cholesterol-fed rabbits develop aortic valve sclerosis, characterized by gross thickening, macrophage infiltration, calcification, and eventual bone formation that mimics the clinical presentation of the disease. Cholesterol-fed rabbits underwent ^{19}F MR imaging after injection of $\alpha_v\beta_3$-integrin-targeted nanoparticles to quantify angiogenesis in the aortic valve leaflets [1]. The cholesterol feeding caused gross thickening of the aortic valves accompanied by extensive foam cell infiltration, non-calcified bone deposition, activation of myofibroblasts, abnormal microvascular proliferation, and upregulation of $\alpha_v\beta_3$-integrin expression. None of these abnormalities were observed in the normal valve tissue from control animals.

The crown ether (CE) signal arising from the nanoparticle contrast agent was readily detected and distinguished from a perfluorooctylbromide (PFOB) quantification reference based on the chemical shifts of these perfluorocarbon species, allowing quantification of the total volume of bound nanoparticles. The volume of targeted nanoparticles bound to the valves was 19.5 nL, which was more than three times higher than the amount of nontargeted nanoparticles (5.6 nL). Competitive inhibition of $\alpha_v\beta_3$-integrin binding reduced the amount of nanoparticles in the valves by about half (10.3 nL). Valves from healthy rabbits treated with targeted nanoparticles contained almost nine times fewer nanoparticles (2.3 nL) than the valves from cholesterol-fed rabbits. These techniques may be useful for assessing atherosclerotic components of preclinical aortic valve disease in patients and could assist in defining efficacy of medical therapies. The sensitivity of this approach for molecular detection of sparse quantities of inflammatory epitopes in very thin structures at high field strengths establishes a basis for future efforts to develop localized spectroscopic methods at clinical field strengths that could be useful for detecting disease and monitoring therapies.

Conclusions

PFC nanoparticles have been utilized for a wide range of molecular imaging, cell tracking, and targeted drug delivery applications. The relatively large size of the particles constrains them to the intravascular space, allowing specific targeting to endothelial expression of integrins, selectins, and adhesion molecules. PFC nanoparticles have a large surface area, which provides a means to incorporate multiple binding ligands and multiple paramagnetic chelates to the particle surface. The particle surface can also serve as a drug reservoir that selectively delivers a therapeutic payload to the targeted cell population. The dual role of PFC nanoparticles as both a molecular imaging contrast agent and a drug delivery vehicle can be exploited to provide diagnostic and therapeutic information via noninvasive MR imaging, or most of the other common clinical imaging modalities, including ultrasound, nuclear imaging, and computed tomography. Although PFC nanoparticles not currently approved for clinical use, the future applications of these agents in the management of patients with chronic diseases could aid in diagnostics, therapeutics, and/or follow-up of treatment efficacy or disease

recurrence, by noninvasively mapping biomarkers associated with cancer or cardiovascular disease and specifically directing highly potent drugs directly to the sites of pathology.

Cross-References

- Angiogenesis
- Imaging the Human Body Down to the Molecular Level
- Intravital Microscopy Analysis
- Liposomes
- Microfabricated Probe Technology
- Nanomedicine
- Nanotechnology in Cardiovascular Diseases

References

1. Chen, J., Lanza, G.M., Wickline, S.A.: Quantitative magnetic resonance fluorine imaging: today and tomorrow. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. **2**, 431–440 (2010)
2. Pan, D., Caruthers, S.D., Chen, J., Winter, P.M., Senpan, A., Schmieder, A.H., Wickline, S.A., Lanza, G.M.: Nanomedicine strategies for molecular targets with MRI and optical imaging. Future Med. Chem. **2**, 471–490 (2010)
3. Winter, P.M., Caruthers, S.D., Lanza, G.M., Wickline, S.A.: Quantitative cardiovascular magnetic resonance for molecular imaging. J. Cardiovasc. Magn. Reson. **12**, 62 (2010)
4. Culver, J., Akers, W., Achilefu, S.: Multimodality molecular imaging with combined optical and SPECT/PET modalities. J. Nucl. Med. **49**, 169–172 (2008)
5. Lanza, G.M., Winter, P.M., Caruthers, S.D., Hughes, M.S., Hu, G., Schmieder, A.H., Wickline, S.A.: Theragnostics for tumor and plaque angiogenesis with perfluorocarbon nanoemulsions. Angiogenesis **13**, 189–202 (2010)
6. Winter, P.M., Shukla, H.P., Caruthers, S.D., Scott, M. J., Fuhrhop, R.W., Robertson, J.D., Gaffney, P.J., Wickline, S.A., Lanza, G.M.: Molecular imaging of human thrombus with computed tomography. Acad. Radiol. **12**(Suppl 1), S9–S13 (2005)
7. Lijowski, M., Caruthers, S., Hu, G., Zhang, H., Scott, M. J., Williams, T., Erpelding, T., Schmieder, A.H., Kiefer, G., Gulyas, G., et al.: High sensitivity: high-resolution SPECT-CT/MR molecular imaging of angiogenesis in the Vx2 model. Invest. Radiol. **44**, 15–22 (2009)
8. Schmieder, A.H., Caruthers, S.D., Zhang, H., Williams, T.A., Robertson, J.D., Wickline, S.A., Lanza, G.M.: Three-dimensional MR mapping of angiogenesis with alpha5beta1(alpha nu beta3)-targeted theranostic nanoparticles in the MDA-MB-435 xenograft mouse model. FASEB J. **22**, 4179–4189 (2008)
9. Winter, P.M., Caruthers, S.D., Zhang, H., Williams, T. A., Wickline, S.A., Lanza, G.M.: Antiangiogenic synergism of integrin-targeted fumagillin nanoparticles and atorvastatin in atherosclerosis. JACC Cardiovasc. Imaging **1**, 624–634 (2008)
10. Zhou, H.F., Chan, H.W., Wickline, S.A., Lanza, G.M., Pham, C.T.: Alphavbeta3-targeted nanotherapy suppresses inflammatory arthritis in mice. FASEB J. **23**, 2978–2985 (2009)
11. Horton, M.A.: The alpha v beta 3 integrin "vitronectin receptor". Int. J. Biochem. Cell Biol. **29**, 721–725 (1997)
12. Cai, K., Caruthers, S.D., Huang, W., Williams, T.A., Zhang, H., Wickline, S.A., Lanza, G.M., Winter, P.M.: MR molecular imaging of aortic angiogenesis. JACC Cardiovasc. Imaging **3**, 824–832 (2010)
13. Winter, P.M., Caruthers, S.D., Allen, J.S., Cai, K., Williams, T.A., Lanza, G.M., Wickline, S.A.: Molecular imaging of angiogenic therapy in peripheral vascular disease with alphanubeta3-integrin-targeted nanoparticles. Magn. Reson. Med. **64**, 369–376 (2010)
14. Winter, P.M., Schmieder, A.H., Caruthers, S.D., Keene, J.L., Zhang, H., Wickline, S.A., Lanza, G.M.: Minute dosages of alpha(nu)beta3-targeted fumagillin nanoparticles impair Vx-2 tumor angiogenesis and development in rabbits. FASEB J. **22**, 2758–2767 (2008)
15. Boles, K.S., Schmieder, A.H., Koch, A.W., Carano, R. A., Wu, Y., Caruthers, S.D., Tong, R.K., Stawicki, S., Hu, G., Scott, M.J., et al.: MR angiogenesis imaging with Robo4- vs. alphaVbeta3-targeted nanoparticles in a B16/F10 mouse melanoma model. FASEB J. **24**, 4262–4270 (2010)
16. Winter, P.M., Neubauer, A.M., Caruthers, S.D., Harris, T.D., Robertson, J.D., Williams, T.A., Schmieder, A. H., Hu, G., Allen, J.S., Lacy, E.K., et al.: Endothelial alpha(v)beta3 integrin-targeted fumagillin nanoparticles inhibit angiogenesis in atherosclerosis. Arterioscler. Thromb. Vasc. Biol. **26**, 2103–2109 (2006)
17. Lanza, G.M., Yu, X., Winter, P.M., Abendschein, D. R., Karukstis, K.K., Scott, M.J., Chinen, L.K., Fuhrhop, R.W., Scherrer, D.E., Wickline, S.A.: Targeted antiproliferative drug delivery to vascular smooth muscle cells with a magnetic resonance imaging nanoparticle contrast agent: implications for rational therapy of restenosis. Circulation **106**, 2842–2847 (2002)
18. Cyrus, T., Zhang, H., Allen, J.S., Williams, T.A., Hu, G., Caruthers, S.D., Wickline, S.A., Lanza, G.M.: Intramural delivery of rapamycin with alphavbeta3-targeted paramagnetic nanoparticles inhibits stenosis after balloon injury. Arterioscler. Thromb. Vasc. Biol. **28**, 820–826 (2008)
19. Soman, N.R., Baldwin, S.L., Hu, G., Marsh, J.N., Lanza, G.M., Heuser, J.E., Arbeit, J.M., Wickline, S.A.,

Schlesinger, P.H.: Molecularly targeted nanocarriers deliver the cytolytic peptide melittin specifically to tumor cells in mice, reducing tumor growth. J. Clin. Invest. **119**, 2830–2842 (2009)
20. Morawski, A.M., Winter, P.M., Yu, X., Fuhrhop, R. W., Scott, M.J., Hockett, F., Robertson, J.D., Gaffney, P.J., Lanza, G.M., Wickline, S.A.: Quantitative "magnetic resonance immunohistochemistry" with ligand-targeted (19)F nanoparticles. Magn. Reson. Med. **52**, 1255–1262 (2004)

Perfluoropolyethers

▶ Boundary Lubrication

Petal Effect

▶ Rose Petal Effect

Petroleum Lubricants

▶ Boundary Lubrication

PFC Nanoparticles

▶ Perfluorocarbon Nanoparticles

PFOB Nanoparticles

▶ Perfluorocarbon Nanoparticles

Pharmacokinetics

▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles

Phase Transformation of Carbon to Diamond

▶ Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering

Phonon Transport

▶ Heat Transfer in Semiconductor Nanostructures

Phononic Crystals

▶ Heat Transfer in Semiconductor Nanostructures

Photocatalysis

▶ Influence of Defects in Photocatalysis

Photocatalytic

▶ Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Photoetching

▶ Chemical Milling and Photochemical Milling

Photofabrication

▶ Chemical Milling and Photochemical Milling

Photolithography

Marc Madou[1] and Chunlei (Peggy) Wang[2]
[1]Department of Mechanical and Aerospace Engineering and Biomedical Engineering, University of California at Irvine, Irvine, CA, USA
[2]Department of Mechanical and Materials Engineering, Florida International University, Miami, FL, USA

Synonyms

Optical lithography; UV lithography

Definition

Photolithography is a process used in microfabrication to selectively pattern a thin film substrate using light to transfer a geometric pattern from a photomask to a light-sensitive photoresist [1].

Introduction

Lithography is a technique used to transfer copies of a master pattern onto the surface of a solid material such as a silicon wafer. The word *lithography* (Greek for the words *stone* [*lithos*] and to *write* [*gráphein*]) refers to the process invented in 1796 by Aloys Senefelder, who inked Bavarian limestone and transferred a carved image from stone onto paper. The most widely used form of lithography is photolithography. In the IC industry, pattern transfer from masks onto thin films is accomplished almost exclusively via photolithography. Photomasking, followed by chemical processing, led to the photolithography now used in fabricating ICs (integrated circuits) and in miniaturization science. The first applications of the printed circuit board (PCB) (also printed wiring board or PWB) were invented in 1943 by the Austrian Paul Eisler. By 1961, Fairchild Semiconductor had introduced the first commercial integrated circuit in which photoetching processes produce large numbers of transistors on a thin slice of silicon (Si). Now, integrated circuits can be found in almost every electronic device, microprocessors, audio and video equipment, dishwashers, garage openers, security systems, automobiles, etc. With the first ICs, patterns had a resolution not better than 5 μm. Today, photolithography, x-ray lithography, and charged-particle lithography all achieve submicron printing accuracy. Intel introduced the Prescott Pentium IV chip in the fourth quarter of 2003, an IC chip incorporating 90 nm sized features, using UV photolithography. Chips based on a 65 nm technology became available in 2006, and the 32 nm node was achieved in 2009. Furthermore, first consumer-level CPU deliveries started in April 2012 using a 22 nm technology. In 2014, the 14 nm technology was also reached by semiconductor companies.

The combination of accurate registration and exposing a series of successive patterns leads to complex multilayered ICs. This essentially two-dimensional (2D) process has a limited tolerance for nonplanar topography, creating a major constraint for building non-IC miniaturized systems, which often exhibit extreme topographies. Photolithography has matured rapidly, and its ability to resolve ever-smaller features is constantly improving. For the IC industry, this continued improvement in photolithography resolution has impeded the adaptation of alternative, higher-resolution lithography techniques, such as x-ray lithography. Research in high-aspect-ratio resists features to satisfy the needs of both IC and non-IC miniaturization is also improving dramatically photolithography's capacity to cover wide ranges of topography. Performance of a photolithographic process is determined by its resolution, the minimum feature size that can be transferred with high fidelity, the registration, how accurately patterns on successive masks can be aligned, and throughput, the number of wafers that can be transferred per hour. Photolithography and pattern transfer involve a set of processes including photoresist coating, UV exposure, baking, developing, etc.

Wafer Cleaning and Clean Room

An important step, even before lithography proper, is wafer cleaning. Contaminants include solvent stains (methyl alcohol, acetone, trichloroethylene, isopropyl alcohol, xylene, etc.), dust from operators and equipment, smoke particles, etc. Stains or films may lead to adverse effects during oxidation and evaporation processes. Particulates, chunks of granular matter, may cause undesirable masking effects and scratches on the photomask during contact printing. All lithography processes take place inside a semiconductor clean room, which is a specially constructed enclosed area environmentally controlled with respect to airborne particulates, temperature (±0.1 °F), air pressure, humidity (from 0.5 % to 5 % RH), vibration, and lighting. In a Class 1 clean room, the particle count does not exceed 1 particle per cubic foot with particles of a size of 0.5 μm and larger, and in a Class 100 clean room, the particle count does not exceed 100 particles per cubic foot with particles of a size of 0.5 μm and larger. The allowable contamination particle size in IC manufacture has been decreasing hand in hand with the ever-decreasing minimum feature size.

Many different dry and wet methods for wafer cleaning currently in use, such as RCA1 and RCA2, developed by W. Kern, use mixtures of hydrogen peroxide and various acids or bases followed by deionized (DI) water rinses. Others include vapor cleaning; thermal treatment, for example, baking at 1,000 °C in vacuum or in oxygen; and plasma or glow discharge techniques, for example, in Freons with or without oxygen. Mechanical methods include ultrasonic agitation, polishing with abrasive compounds, and supercritical cleaning. Ultrasonic cleaning, which is excellent for removing particulate matter from the substrate, is unfortunately prone to contamination and mechanical failure of deposited films.

Masks

The stencil used to repeatedly generate a desired pattern on resist-coated wafers is called a *mask*. In typical use, a photomask – a nearly optically flat glass (transparent to near ultraviolet [UV]) or quartz plate (transparent to deep UV) with an absorber pattern metal (e.g., an 800 Å thick chromium layer) – is placed above the photoresist-coated surface, and the mask/wafer system is exposed to UV radiation. The absorber pattern on the mask can be generated by e-beam lithography which yields higher resolution than photolithography. A *positive* or *dark-field* mask is a mask on which the pattern is clear with the background dark. A *negative* or *clear-field* mask is a mask on which the pattern is dark with the background clear. A light-field or dark-field image, known as *mask polarity*, is then transferred to the semiconductor surface. This procedure results in a 1:1 image of the entire mask onto the silicon wafer.

Masks, making direct physical contact with the substrate, are called *contact masks*. Unfortunately, these masks degrade faster due to wear than noncontact, proximity masks (also referred to as *soft contact masks*), which are slightly raised, say 10–20 μm, above the wafer. Contact mask and proximity mask printing is collectively known as *shadow printing*. A more reliable method of masking is *projection printing* where, rather than placing a mask in direct contact with (or in proximity of) a wafer, the photomask is imaged by a high-resolution lens system onto the resist-coated wafer. The imaging lens can reduce the mask pattern by 1:5 or 1:10, making mask fabrication less challenging. Lithography is still mostly carried out using masks, but due to problems caused by masks, such as expense and time in fabricating them, contamination introduced by them, their disposal, and the difficulties in their alignment, research into maskless optical projection lithography (MOPL) is growing rapidly and broadly. One MOPL, an approach already on the market (http://www.intelligentmp.com), is based on the digital micromirror device (DMD) chip from Texas Instruments Inc. (TI) and relies on the same spatial light modulation technology used in DLP (digital light processing) projectors and HDTV (high-definition television) arrays in a DMD chip to project images on the photoresist. The resolution of DMD-based maskless photolithography (about 5 μm) is significantly less than with e-beam

lithography (0.25 μm) or laser writers (<1.0 μm), but it is a parallel technique, and for many applications, that is, in microfluidics, the lower resolution might not be an obstacle. Maskless zone plate array lithography (ZPAL) is another optical maskless technology which operates on the principle of diffraction rather than refraction.

Process

Photoresist Deposition

Lithography is the most expensive step in microelectronics technology, representing up to 35 % of the wafer manufacturing cost. Within lithography, photoresist coating is one of the more expensive steps. For silicon integrated circuits, the resist thickness after a prebake (see below for what a prebake does) typically ranges between 0.5 and 2 μm. For miniaturized three-dimensional (3D) structures, much greater resist thicknesses are often required, and complex topography might also call for a conformal resist coat over very high-aspect-ratio features. For thick resist coats, techniques such as casting and the use of thick sheets of dry photoresists replace the ineffective resist spinners.

As the first step in the lithography process itself, a thin layer of an organic polymer, a photoresist sensitive to ultraviolet radiation, is deposited on the oxide surface. The photoresist is dispensed onto the wafer which is held by a vacuum chuck. A speed of about 500 rpm is commonly used during the dispensing step enabling the spread of the fluid over the substrate. After the dispense step, it is common to accelerate to a relatively high speed to thin the fluid to near its final desired thickness. Typical spin speeds for this step range from 1,500 to 6,000 rpm, depending on the properties of the fluid (mostly its viscosity) as well as the substrate. This step can take from 10 s to several minutes. The combination of spin speed and time selected for this step will generally define the final film thickness. At these speeds, centrifugal force causes the solution to flow to the edges, where it builds up until expelled when surface tension is exceeded. The resulting polymer thickness, T, is a function of spin speed, solution concentration, and molecular weight (measured by intrinsic viscosity).

As alternatives to spin coating, several other methods can be used for photoresist application, including spray coating, electrostatic spraying or electrostatic deposition (ED), meniscus coating, silk screening, plasma deposition, electrophoretic photoresist deposition, dip coating, roller coating, curtain coating, extrusion coating, etc.

Soft Baking or Prebaking

After resist coating, the resist still contains up to 15 % solvent and may contain built-in stresses. The wafers are therefore soft baked (also preexposure baked or prebaked) at 90–100 °C for about 20 min in a convection oven, or at 75–85 °C for 1–3 min with a vacuum hot plate, to remove solvents and stress and to promote adhesion of the resist layer to the substrate. This is a critical step in that failure to sufficiently remove the solvent will affect the resist profile. Excessive baking destroys the photoactive compound and reduces sensitivity. Thick resists may benefit from a longer bake time. The resist thickness, for both negative and positive resists, is typically reduced by 10–25 % during soft baking. Hot plating the resist is faster, is more controllable, and does not trap solvent like convection oven baking. In convection ovens, the solvent at the surface of the resist is evaporated first, and this can cause an impermeable resist skin, trapping the remaining solvent inside. So heating in a convection oven must proceed slowly to avoid solvent bursts. Commercially, microwave heating or IR lamps are also used in production lines. The optimization of the prebaking step may substantially increase device yield.

Exposure and Postexposure Treatment

After soft baking, the resist-coated wafers are transferred to an illumination or exposure system where they are aligned with the features on the mask. For any lithographic technique to be of value, it must provide an alignment technique capable of a superposition precision of mask and wafer that is a small fraction of the minimum feature size of the devices under construction. In the simplest case, an exposure system consists of a

UV lamp illuminating the resist-coated wafer through a mask without any lenses between the two. The purpose of the illumination is to deliver light with the proper intensity, directionality, spectral characteristics, and uniformity across the wafer, allowing a nearly perfect transfer or printing of the mask image onto the resist in the form of a latent image.

In photolithography, wavelengths of the light source used for the exposure of the resist-coated wafer range from the very short wavelengths of extreme ultraviolet (EUV) (10–14 nm) to deep ultraviolet (DUV) (150–300 nm) to near ultraviolet (UV) (350–500 nm). In near UV, one typically uses the g-line (435 nm) or i-line (365 nm) of a mercury lamp. The brightness of shorter-wavelength sources is severely reduced compared to that of longer-wavelength sources, and the addition of lenses further reduces the efficiency of the exposure system. In general, the smallest feature that can be printed using projection lithography is roughly equal to the wavelength of the exposure source, in this example, 248 nm would be expected. The same laser, in combination with sophisticated resolution-enhancing techniques (RETs), may be used to produce more advanced circuits with transistor gate features of 160 nm and below. RET methods (see below) enable one to go quite a bit beyond the conventional Rayleigh diffraction limit. Other exposure systems now available include two DUV excimer lasers, the ArF at 193 nm and the F_2 at 157 nm, and extreme UV lithography (EUV) at 13.4 nm. The current generation of lithography is using 193 nm light from ArF lasers. In the case of EUV, a plasma or synchrotron source and all reflective reduction optics (4×) (at this wavelength, all materials absorb!) are used.

Postexposure Treatment

A postexposure treatment of the exposed photoresist is often desired, because the reactions initiated during exposure might not have run to completion. To halt the reactions or to induce new ones, several postexposure treatments are in use: postexposure baking (PEB), flood exposure with other types of radiation, treatment with a reactive gas, and vacuum treatment. Postexposure baking (sometimes in a vacuum) and treatment with reactive gas are used in image reversal and dry resist development. In the case of a chemically amplified resist, the postexposure bake is most critical. Although reactions induced by the catalyst that forms during exposure take place at room temperature, their rate is greatly increased by baking at 100 °C. The precise control of this type of postexposure bake critically determines the subsequent development itself.

Development

Development is the dissolution of unpolymerized resist that transforms the latent resist image, formed during exposure, into a relief image that will serve as a mask for further subtractive and additive steps. During the development of an exposed resist, selective dissolving takes place. Two main technologies are available for development: wet development, which is widely used in circuit and miniaturization manufacture in general, and dry development, which is starting to replace wet development for some of the ultimate line-width resolution applications.

During batch spray development, fan-type sprayers direct fresh developing solution across wafer surfaces. Positive resists are typically developed in aqueous alkaline solutions and negative resists in organic ones. For alkaline developers, simple solutions of NaOH (Shipley 351) or KOH (AZ/Clariant 400 K) may be used, but because of the possibility of mobile ion contamination in Metal-oxide-semiconductor (MOS) devices, developers free of metal ions are preferred. The latter are usually tetramethylammonium hydroxide (TMAH) based (Shipley CD-26, MF-321, AZ/Clariant 300-MIF). Each developer has a different dilution, and some require longer development times than others. Developers are generally matched to a type of photoresist. Though they may be interchangeable to some extent, changing the type of developer used in a process will usually change the exposure time necessary to resolve the pattern. It should be noted that all of the abovementioned alkaline developers etch aluminum. If the micromachine or microchip, under construction, contains aluminum features, Shipley Microposit Developer Concentrate (MDC),

a mixture of proprietary alkaline salts (mostly phosphates) with the slowest aluminum etch rate, is a better developer choice.

The use of organic solvents leads to some swelling of the resist (especially for negative resists; see below) and loss of adhesion of the resist to the substrate. Dry development overcomes these problems, as it is based either on a vapor phase process or a plasma. In the latter, oxygen-reactive ion etching (O_2-RIE) is used to develop the latent image. The image formed during exposure exhibits a differential etch rate to O_2-RIE rather than differential solubility to a solvent.

De-scumming and Post-baking

A mild oxygen plasma treatment, the so-called de-scumming, removes unwanted resists left behind after development. Negative, and to a lesser degree positive, resists leave a thin polymer film at the resist/substrate interface. The problem is most severe in small (<1 μm) high-aspect-ratio structures where the mass transfer of a wet developer is poor. Patterned resist areas are also thinned in the de-scumming process, but this is usually of little consequence. Before etching the substrate or adding a material, the wafer must be post-baked. Post-baking or hard baking removes residual coating solvent and developer and anneals the film to promote interfacial adhesion of the resist that has been weakened either by developer penetration along the resist/substrate interface or by swelling of the resist (mainly for negative resists). Hard baking also improves the hardness of the film and avoids solvent bursts during vacuum processing. Improved hardness increases the resistance of the resist to subsequent etching steps. Post-baking frequently occurs at higher temperatures (120 °C) and for longer times (say 20 min) than soft or prebaking. The major limitation for heat application is excessive flow or melt, which degrades wall profile angles and makes it more difficult to remove the resist. Post-bake does induce some stress and resist shrinkage. Special care needs to be taken when the baking temperature is above the glass transition temperature, T_g, when impurities are easily incorporated into the resist due to the plastic flow of the resist.

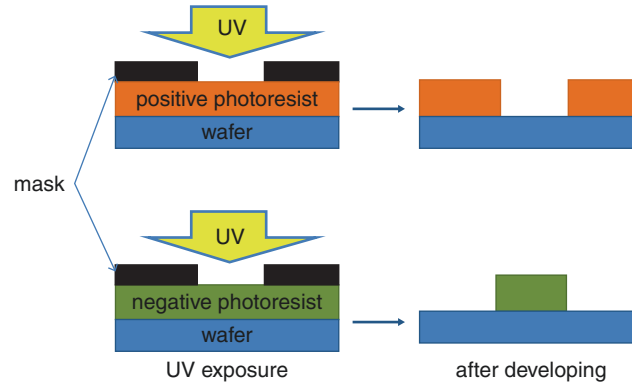
Resists

The principal components of photoresists are a polymer (base resin), a sensitizer, and a casting solvent. The polymer changes structure when exposed to radiation; the solvent allows spin application and formation of thin layers on the wafer surface; sensitizers control the chemical reactions in the polymeric phase. Resists without sensitizers are single-component or one-component systems, whereas sensitizer-based resists are two-component systems. Solvent and other potential additives do not directly relate to the photoactivity of the resist.

If the photoresist is of the type called *positive* (also *positive tone*), the photochemical reaction during exposure of a resist weakens the polymer by rupture or scission of the main and side polymer chains, and the exposed resist becomes more soluble in developing solutions (say ten times more soluble). In other words, the development rate, R, for the exposed resist is about ten times faster than the development rate, R_0, for the unexposed resist. Two well-known families of positive photoresists are the single-component poly(methyl methacrylate) (PMMA) resists and the two-component DNQ resists comprised of a photoactive component (PAC) such as diazonaphthoquinone ester (DNQ) (20–50 wt%) and a phenolic novolac resin.

If the photoresist is of the type called *negative* (also *negative tone*), the reaction strengthens the polymer by random cross-linkage of main chains or pendant side chains, becoming less soluble (slower dissolving). Exposure, development, and pattern-transfer sequences for negative and positive resists are shown in Fig. 1.

The first negative photoresists were based on free radical-initiated photo-cross-linking processes of main or pendant polymer side chains, rendering the exposed parts insoluble. Commonly used negative-acting, two-component resists are bis(aryl)azide rubber resists, whose matrix resin is cyclized poly(cis-isoprene), a synthetic rubber. An example of a commercial, two-component negative photoresist is the Kodak KTFR [an azide-sensitized poly(isoprene) rubber] with a lithographic sensitivity (also photospeed) of



Photolithography, Fig. 1 Positive and negative resist: exposure, development, and pattern transfer

75–125 mJ/cm^2. Negative photoresists, in general, adhere very well to the substrate, and a vast amount of compositions are available (stemming from R&D work in paints, UV-curing inks, and adhesives all based on polymerization hardening). Negative resists are highly resistant to acid and alkaline aqueous solutions as well as to oxidizing agents. As a consequence, a given thickness of negative resist is more chemically resistant than a corresponding thickness of positive resist. A comparison of negative and positive photoresist features is presented in Table 1. This table is not exhaustive and is meant only as a practical guide for selection of a resist tone.

Resists typically are removed (stripped) once they have served their function as temporary stencils. Some negative resists, hardened through UV exposure, are used as permanent components of miniature devices. Two prominent examples in this category are polyimide and SU-8-based resists. Polyimides are a class of polymers synthesized from two monomers: a dianhydride and a diamine. Commercial products are supplied as soluble polyamic acid (PAA) intermediates, which undergo a thermal imidization with the evolution of water to form an insoluble polyimide. Polyimide is used as an interlayer dielectric because of its desirable characteristics such as a high thermal stability with a glass transition temperature, Tg $>$ 300 °C, a low dielectric constant, a good chemical stability, and an ease of processing.

Most resists in IC and Micro-Electro-Mechanical Systems (MEMS) fabrication are deposited as liquids, whereas resists used in printed wiring board (PWB) manufacture are usually dry film resists that come in rolls (ranging from 2 to 60 in. wide and 125 to 1,000 ft. long) and are laminated onto the substrate instead of being spin coated on it. Dry film resist formulations are sandwiched between a polyolefin release sheet and a polyester base, rolled up on a support core. Dry film resists offer advantages such as excellent adhesion on most substrates, no liquid handling since there is no solvent, high process speed, excellent thickness uniformity over a whole substrate, simple handling, no formation of edge beads, low exposure energy, low cost, short processing time, and near vertical sidewalls. There are a variety of dry film photoresists widely used and commercially available, examples include Riston®, Ordyl BF 410, Etertec® 5600, DF 4615, and DFR-15. All are used in the manufacture of circuit boards and can be made quite thick, and all are candidates for broad use in MEMS as well.

Photoresist stripping, in slightly oversimplified terms, is organic polymer etching. The primary consideration is the complete removal of the photoresist without damaging the device under construction. There are many commercial strippers available; some are specific for positive resists (e.g., ACT-690 C from Ashland), others for negative resists, and still others are universal strippers (e.g., ACT-140 from Ashland). Other popular commercial strippers are Piranha and RCA clean. Wet stripping solutions lose potency in use, causing stripping rates to change with time.

Photolithography, Table 1 Comparison of traditional negative and positive photoresists

	Resist type	
Characteristic	Positive resist	Negative resist
Adhesion to Si	Fair (priming required)	Excellent (priming not required)
Available compositions	Many	Vast
Baking	In air (+)	In nitrogen (−)
Contrast γ	Higher, e.g., 2.2	Lower, e.g., 1.5
Cost	More expensive	Less expensive
Developer	Temperature sensitive (−) and aqueous based (ecologically sound)	Temperature insensitive (+) and organic solvent (−)
Developer process window	Small	Very wide, insensitive to overdeveloping
Influence of oxygen	No (+)	Yes (−)
Lift-off	Yes	Yes
Mask type	Dark field: lower defects	Clear field: higher defects
Opaque dirt on clear portion of mask	Not very sensitive to it	Causes printing of pinholes
Photospeed	Slower	Faster
Pinhole count	Higher	Lower
Pinholes in mask	Prints mask pinholes	Not so sensitive to mask pinholes
Plasma etch resistance	Not very good	Very good
Proximity effect	Prints isolated holes or trenches better	Prints isolated lines better
Residue after development	Mostly at <1 μm and with high aspect ratio	Often a problem
Resolution	High	Low (>1 μm)
Sensitizer quantum yield Φ	0.2–0.3	0.5–1
Step coverage	Better	Lower
Strippers of resist over		
Oxide steps	Acid	Acid
Metal steps	Simple solvents	Chlorinated solvent compounds
Swelling in developer	No	Yes
Thermal stability	Good	Fair
Wet chemical resistance	Fair	Excellent

Accumulated contamination in solutions can be a source of particles, and liquid-phase surface tension and mass transport tend to make photoresist removal difficult and uneven. Dry stripping or oxygen plasma stripping, also known as ashing, has become more popular as it poses fewer disposal problems with toxic, flammable, and dangerous chemicals. Dry stripping is more controllable than liquid stripping, less corrosive with respect to metal features on the wafer, and, more importantly, it leaves a cleaner surface under the right conditions. Finally, it does not cause the undercutting and broadening of photoresist features that can be caused by wet strippers.

Critical Dimension and Overall Resolution

The absolute size of a minimum feature in an IC or a miniature device, whether it involves a line-width, spacing, or contact dimension, is called the *critical dimension* (CD). The overall resolution of a process describes the consistent ability to print a minimum size image, a critical dimension, under conditions of reasonable manufacturing variation. Many aspects of the process, including hardware, materials, and processing considerations, can limit the resolution of lithography. Hardware limitations include diffraction of light or scattering of charged particles (in the case of charged-particle lithography or hard x-rays), lens aberrations, mechanical stability of the system, etc. The resist material properties that impact resolution are contrast, swelling behavior, thermal flow, chemical etch resistance, etc. The most important process-related resist variables include swelling (during development) and stability (during etching and baking steps). Resolution frequently is measured by line-width measurements using either transmitted or reflected light or other metrology techniques. Optical techniques perform satisfactorily for features of 1 μm and larger, providing a precision of ± 0.1 μm. The successful performance of devices depends upon the control of the size of critical structures across the entire wafer and from one wafer to another, referred to as line-width control. A rule of thumb is that the

dimensions must be controlled to tolerances of at least ±1/5 of the minimum feature size.

Correct feature size must be maintained within a wafer and from wafer to wafer, as device performance depends on the absolute size of the patterned structures. In the shadow printing mode, including contact and proximity arrangements of mask and wafer, optical lithography has a resolution with limits set by a variety of factors. These include diffraction of light at the edge of an opaque feature in the mask as the light passes through an adjacent clear area, alignment of wafer to mask, nonuniformities in wafer flatness, and debris between mask and wafer. Diffraction causes the image of a perfectly delineated edge to become blurred or diffused. The theoretical resolution, R, that is, the minimum resolved dimension in a grating mask ($b_{\min}$ for a line or a space), is given by

$$\mathrm{R} = b_{\min} = k\sqrt{\lambda\left(\mathrm{s} + \frac{\mathrm{z}}{2}\right)} \qquad (1)$$

where

- $b_{\min}$ = half the grating period and the minimum feature size transferable
- s = the gap between the mask and the photoresist surface
- λ = wavelength of the exposing radiation
- z = the photoresist thickness
- k = a constant which theoretically is ~1.5

Depending on the relative magnitude of s and z, one distinguishes between contact and proximity printing. The square root relation in contact and proximity printing is a consequence of the near-field or Fresnel diffraction theory valid in the near-field region just below the mask openings; this contrasts with the far-field behavior (Fraunhofer diffraction) of projection lithography (see below). Self-aligned printing can be regarded as an extreme form of contact printing where there is no gap between the mask and the substrate. In contact printing, a photomask is pressed against the resist-covered wafer with pressures in the range of 0.05–0.3 atm., and the s, in Eq. 1, is zero. The theoretical maximum resolution is seldom achieved, however, as only diffraction effects were taken into account to derive Eq. 1. The other factors mentioned above (wafer flatness, mask alignment, etc.) and shadowing effect (penumbral blur) usually conspire to make the resolution worse.

In proximity printing, spacing of the mask removed from the substrate (by at least 10 μm) minimizes defects that result from contact. On the other hand, diffraction of the transmitted light reduces the resolution. The degree of reduction in resolution and image distortion depends on the wafer-to-substrate distance, which may vary across the wafer.

In projection printing, wafer contact is completely avoided; a high-resolution lens projects an image of the photomask onto the photoresist-covered wafer. Imaging in this case is far-field or Fraunhofer diffraction dominated, and the Rayleigh criterion for far-field resolution can be considered. A dimensionless constant k_1 can be used to define the practical limiting resolution R in projection printing as

$$\mathrm{R} = k_1 \frac{\lambda}{\mathrm{NA}} \qquad (2)$$

where k_1 = experimentally determined dimensionless parameter that depends on resist parameters, process conditions, mask aligner optics, etc.

To achieve higher resolutions, instrumentation designers want to work at shorter wavelengths λ, higher NA, and smaller k_1.

Since the mid-1980s, the demise of optical lithography has been predicted as being only a few years away, but each time a new resolution limit approaches, some new method extends the useful life of the technology. The three main technologies involved in printing ICs and other miniaturized devices are resist technology, mask technology, and the exposure tool, all of which need to be addressed to optimize lithography resolution. Strategies for improved resolution can be achieved through using chemically amplified resists, antireflection coatings, improved mask technology (such as optical proximity correction), improved exposure equipment (such as off-axis

illumination), etc. DUV photolithography is expected to continue until the 45 nm node through these resolution enhancing techniques. The 32 nm node is viewed as beyond the scope of DUV lithography.

Next-Generation Lithographies (NGLs) and Lithography Research

In the IC industry, continuous improvements to DUV photolithography have postponed the industrial adoption of alternative next-generation lithographies (NGLs), such as extreme ultraviolet lithography (EUVL), x-ray lithography, charged-particle beam lithography based on electrons and ions (such as electron and ion projection techniques), and imprint lithography (nanoimprint lithography [NIL] and step-and-flash imprint lithography [SFIL]), because of the huge financial investment in existing photolithography equipment. These NGLs are regarded today as important technologies for the beyond-the-DUV-lithography era, either for mask making or for actual IC production. In addition, many lithography approaches in the R&D stage, including lithography based on very thin resist layers and block copolymers, zone plate array lithography (ZPAL), quantum lithography (two-photon lithography), and proximal probe-based techniques such as atomic force microscopy (AFM), scanning tunneling microscopy (STM), dip-pen lithography (DPL), near-field scanning optical microscopy (NSOM), and apertureless near-field scanning optical microscopy (ANSOM), are attracting more and more research interest. It is quite possible that some of the alternative lithography tools in the R&D category will emerge as serious next-generation lithographies (NGLs) in the coming years.

Extreme Ultraviolet Lithography (EUVL)

Extreme ultraviolet lithography (EUVL), using wavelengths in the 10–14 nm range to carry out projection imaging, is perhaps the most natural extension of optical projection lithography as, in principle, it only differs in terms of the wavelength. This type of radiation is also referred to as *soft x-ray radiation* and *vacuum UV*. Sources for this type of radiation are laser-produced plasmas and synchrotrons. EUV is strongly absorbed in virtually all materials, and, consequently, imaging must be carried out in vacuum; also, all camera optics as well as masks used must be reflective rather than refractive. New resists and processing techniques must be developed as well.

Photolithography, Table 2 Light sources for various types of lithography

Wavelength [nm]	Source	Range
436	Hg arc lamp	G-line
405	Hg arc lamp	H-line
365	Hg arc lamp	I-line
248	Hg/Xe arc lamp, KrF excimer laser	Deep UV (DUV)
193	ArF excimer laser	DUV
157	F2 laser	Vacuum UV (VUV)
~10	Laser-produced plasma sources	Extreme UV (EUV)
~1	X-ray tube, synchrotron	X-ray

X-ray Lithography

X-ray lithography employs a shadow printing method similar to optical proximity printing. The x-ray wavelength (4–50 Å) is much shorter than that of UV light (Table 2). Hence, diffraction effects are reduced and higher resolution can be attained. For instance, for an x-ray wavelength of 5 Å and a gap of 40 μm, resolution can be around 0.2 μm. In contrast, with electron lithography and ion beam lithography, no charged particles are directly involved in x-ray exposures, which makes the need for vacuum less stringent.

X-ray lithography is superior to optical lithography due to its use of shorter wavelengths and its very large Depth of field (DOF) and because exposure time and development conditions are not as stringent. Reproducibility is high since results are independent of substrate type, surface reflections, and wafer topography. Another important benefit is that x-ray lithography is immune to low-atomic-number (Z) particle contamination

(dust). With an x-ray wavelength on the order of 10 Å or less, diffraction effects generally are negligible, and proximity masking can be used, increasing the lifetime of the mask. In x-ray lithography, there are essentially no optics involved, and, although this sounds like an advantage, it also presents one major disadvantage since one can only work with 1:1 shadow printing. No image reduction is possible, so the mask fabrication process is very complicated. In the United States, IBM remains the only major champion of x-ray lithography for next-generation lithography. At the end of 1999, IBM fabricated several PowerPC 604e microprocessor batches to demonstrate the viability of the method. In Japan, there are still many players involved such as NTT, Toshiba, Mitsubishi, and NEC.

LIGA

The LIGA technique (a German acronym for Lithographie, Galvanoformung, Abformung) was invented about 20 years ago. LIGA exploits all of the advantages of x-ray lithography listed above. The LIGA process involves a thick layer of resist (from micrometers to centimeters), high-energy x-ray radiation, and resist development to make a resist mold. By applying galvanizing techniques, the mold is filled with a metal. The resist structure is removed, and metal products result. Alternatively, the metal part can serve as a mold itself for precision plastic injection molding. Several types of plastic molding processes have been tested, including reaction injection molding, thermoplastic injection molding, and hot embossing. LIGA enables new building materials and a wider dynamic range of dimensions and possible shapes.

Cross-References

- ▶ BioPatterning
- ▶ Biosensors
- ▶ CMOS MEMS Biosensors
- ▶ CMOS-CNT Integration
- ▶ Dip-Pen Nanolithography
- ▶ Dry Etching Processes
- ▶ DUV Photolithography and Materials
- ▶ Electron Beam Lithography (EBL)
- ▶ Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)
- ▶ MEMS
- ▶ Microcontact Printing
- ▶ Nanoimprinting
- ▶ NEMS Piezoelectric Switches
- ▶ NEMS Resonant Chemical Sensors
- ▶ SU-8 Photoresist

References

1. Madou, M.: Fundamentals of Microfabrication, 2nd edn. CRC Press, Boca Raton (2002)

Photomilling

- ▶ Chemical Milling and Photochemical Milling

Photonic Crystal Nanobeam Cavities

Parag B. Deotare and Marko Loncar
Electrical Engineering, Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA

Definition

Photonic crystal nanobeam cavities are high quality factor, small mode volume optical resonators formed by using photonic crystal Bragg mirrors in the longitudinal direction and total internal reflection in the others.

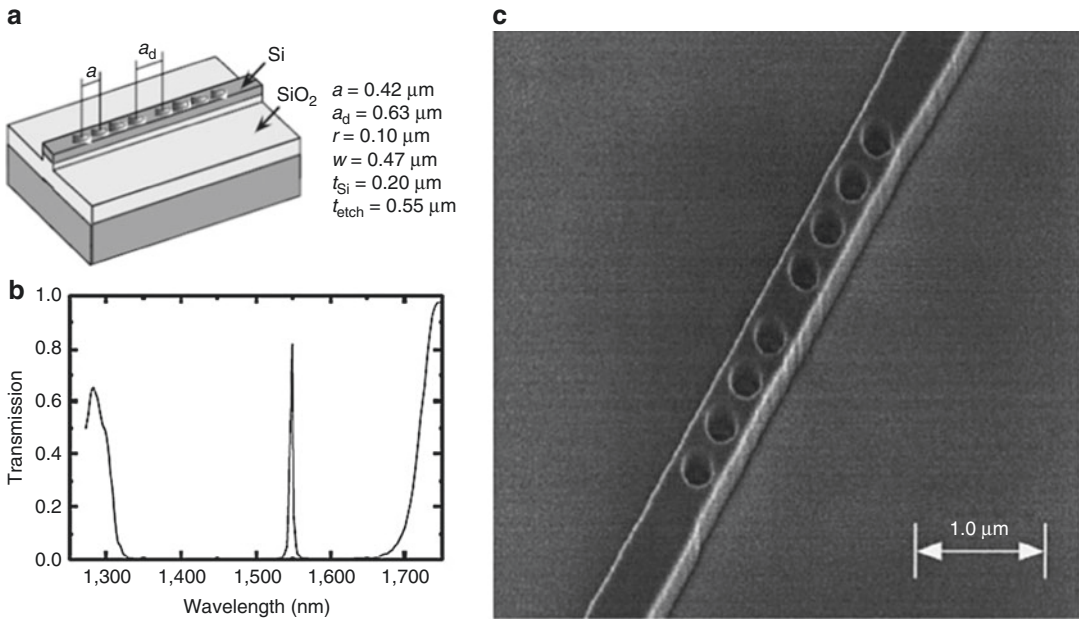
Introduction

On-chip manipulation of light is of great importance for applications ranging from optoelectronics and nonlinear optics to biochemical sensing and cavity quantum electrodynamics. Ultra-high quality factor (Q) optical resonators are

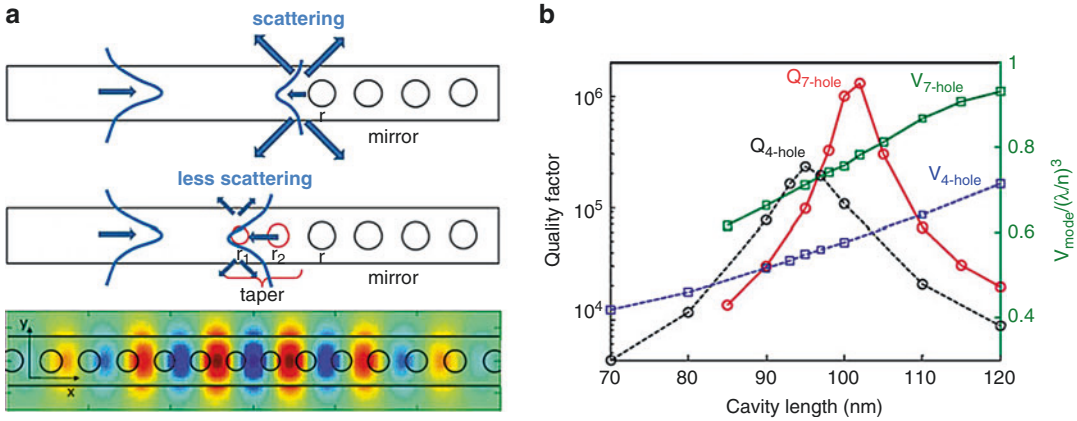
indispensable for these applications since they enhance the probability of light-matter interaction due to their ability to store photons for many optical cycles [1]. Traveling-wave-based resonators, and micro-toroid resonators in particular, have been used to achieve some of the largest Q factors to date [1]. For many applications, however, the figure of merit is proportional to the ratio between resonator's Q and mode volume (V_{mode}), which can be quite large in traveling-wave geometries. Therefore, standing-wave optical resonators, capable of having very small mode volume, have been explored as an attractive alternative [1]. These structures, based on a Fabry-Perot geometry, typically consist of the region in which light can propagate (e.g., an optical waveguide), terminated at its both ends with optical mirrors. The mirrors are often in the form of a 1-D Bragg mirror, as in the case of vertical cavity surface emitting laser (VCSEL), for example. In 1987, the Bragg mirror idea was extended to 3-D by Yablonovitch [2] and John [3] who proposed 3-D photonic crystals as a platform suitable for realization of ultimate cavities with wavelength-scale mode volumes and ultra-high Qs. Unfortunately, fabrication of these 3-D periodic optical structures has proven to be challenging and in spite of great effort no ultra-high Q small mode volume cavity based on 3-D photonic crystals has been realized to date. This prompted researchers to consider alternative approaches, based on lower-dimensionality photonic crystals. In particular, planar photonics crystals, based on 2-D periodic lattices realized in thin semiconductor membranes [4] received considerable interest due to the ease of fabrication and compatibility with standard planar microfabrication technology. State-of-the-art fabrication techniques, combined with creative cavity designs, have resulted in wavelength-scale optical cavities with $Q > 10^6$ (experimental) and Q/V_{mode} values larger than traveling wave resonators. Recently, however, it has been demonstrated that photonic crystals with even lower dimensionality, resembling some of the early distributed feedback (DFB) cavity designs, can be used to realize cavities with performance rivaling and even outperforming those based on 2-D planar photonic crystals. These photonic crystal nanobeam cavities (PCNCs), based on chirped 1-D optical lattices, also have a physical footprint which is exactly the same as that of an optical waveguide and therefore, represent the smallest cavity geometry that can be made using dielectric materials only. Furthermore, nanobeams enable realization of high-Q small-V_{mode} cavities even in low-index materials (e.g., SiO_2, polymers) [5, 6], thus vastly increasing the library of photonic materials suitable for realization of ultra-high Q/V_{mode} cavities. The near-field of PCNCs is highly accessible, which allows for easy light-material interaction and which is useful for sensing [6], optical trapping [7], and optomechanics [8, 9]. PCNCs can be easily and naturally integrated with optical waveguides for efficient delivery of light [10, 11]. Cavities that support ultra-high Q modes with both TE and TM polarization [12] can be easily realized. All of these features, many of which are hard to achieve using 2-D photonic crystals, make photonic crystal nanobeam cavities an ideal platform for practical applications.

Design Evolution

The idea of waveguide-based micro-cavities has been around for more than 15 years. It was first proposed at MIT [13] in 1995 and then experimentally realized 2 years later [14]. Figure 1a shows the schematic of the device which consisted of a waveguide segment surrounded with two lattices of holes that play the role of Bragg mirrors. Light trapping is achieved in this system via a combination of Bragg scattering (in longitudinal direction) and total internal reflection (in two transverse directions), resulting in an optical resonance as shown in Fig. 1b. The theoretical value of Q was calculated to be $Q = 280$ whereas the experimentally measured value was 265 (Fig. 1c). Such a low value of Q factor was due to the suboptimal design. In particular, light scattering at the interface between the Bragg mirror and the central cavity region was responsible for the low Q factor (Fig. 2a) due to the mismatch in the effective mode indices of the two regions.



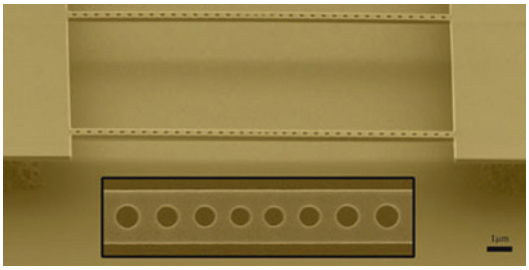
Photonic Crystal Nanobeam Cavities, Fig. 1 (**a**) Schematic of photonic crystal nanobeam cavity (PCNC) (**b**) simulated transmission of the device showing the bandgap and the cavity resonance [13] (**c**) SEM image of the fabricated device. The measure Q of the device was 265 [14]



Photonic Crystal Nanobeam Cavities, Fig. 2 (**a**) Schematic showing the effect of taper on scattering loss due to mismatch of the waveguide mode and the Bloch mode. Cavity mode electric field (Ey) for a 7-hole taper (**b**) simulated results from [16] showing the variation in Q and mode volume for different taper and cavity lengths. Cavity length was defined as the distance between two innermost holes (indicated in Part **a**)

In order to increase the Q of the cavity it is necessary to *taper* propagating mode of the cavity region into the exponentially decaying mode of the Bragg mirror [15, 16]. This mode matching, analogous to impedance matching problem in electronics, can be achieved using a multi-hole taper toward the cavity [15]. By doing so, the waveguide mode index can slowly be reduced to match the Bragg index which significantly reduces the scattering and provides exponential increase in cavity Q, at an expense of slight, linear increase in the cavity mode volume. By increasing

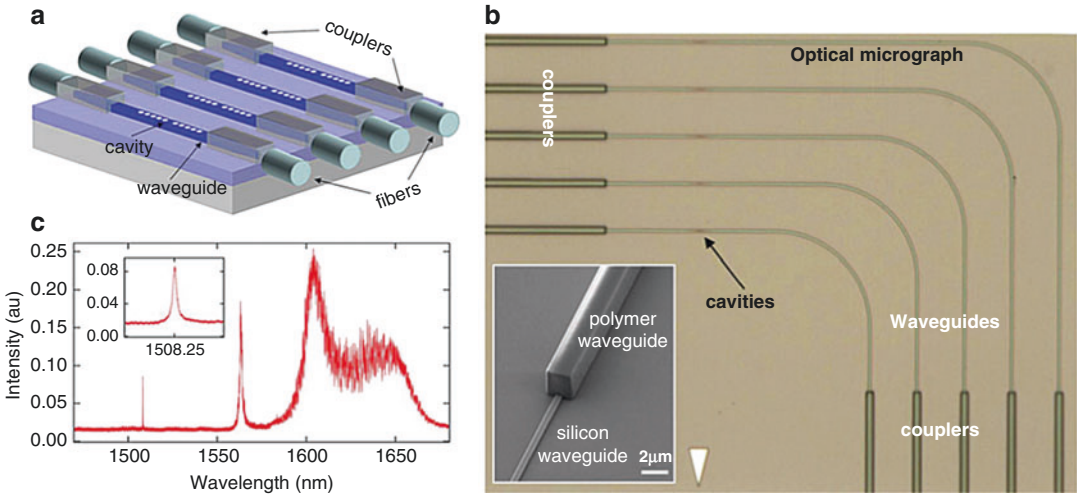
the number of holes used in the taper, near adiabatic conversion between the two modes could be achieved. By putting two of such engineered tapered (chirped) mirrors back to back, ultra high-Q cavity could be realized. The electric field of a cavity mode formed by using a 7-hole taper is shown in Fig. 2a. The effect of tapered holes on the Q and mode volume was first studied in silicon nitride material system for operation in visible [16] and the results are shown in Fig. 2b. The figure shows Q as a function of cavity length, defined as distance between two central holes in the structure, for two different numbers of tapered holes used in the mirror. It can be seen that the drastic increase in Q can be achieved for an optimal value of the cavity length. Also, it is apparent that the Q values depend strongly on the precise cavity length which puts stringent requirement on the fabrication quality needed for the realization of ultra-high Q cavities.



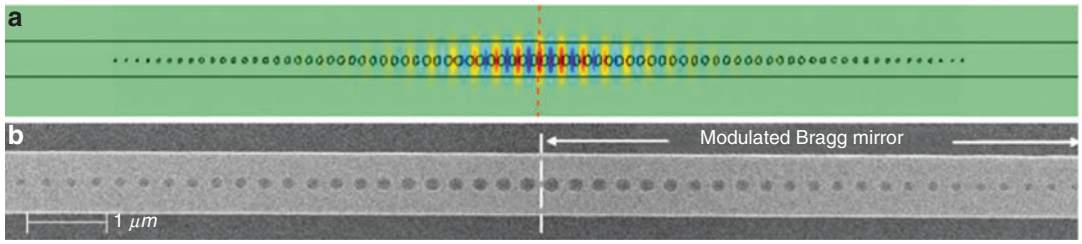
Photonic Crystal Nanobeam Cavities, Fig. 3 SEM image of the fabricated structure with a measured Q of 7.5×10^5 [17]. The inset shows the zoomed cavity region

Ultra-High Q Nanobeam Cavities

In order to validate the designs, devices were fabricated in silicon-on-insulator (SOI) material platform (Fig. 3) using state-of-the art nanofabrication technology. The highest Q measured in such cavities that operated in telecom wavelength range was 750,000 [17]. The cavities were probed using a free-space cross-polarization technique popularly known as resonance scattering. This technique is relatively straightforward to implement but suffers from poor in- and out-coupling of light, in particular in the case of high-Q cavity modes. Other characterization techniques that have been used to characterize nanobeam cavities include evanescent coupling using a tapered fiber [8] and butt coupling [10] (Fig. 4a). The latter technique, particularly suitable for realization of integrated on-chip



Photonic Crystal Nanobeam Cavities, Fig. 4 (**a**) Schematic showing butt coupling technique using spot-size converter couplers based on polymer waveguides and inverse-tapered silicon waveguides. (**b**) Optical micrograph showing the fabricated device with the in- and out-coupling pads, silicon waveguide, and PCNC. Inset shows an SEM image of the polymer coupling pads. (**c**) Transmission from one of the devices shown in (**b**)



Photonic Crystal Nanobeam Cavities, Fig. 5 Ultra-high Q cavity designed using deterministic approach [10] (**a**) Electric field profile (**b**) SEM image of structure fabricated in SOI

networks, takes advantage of spot-size converters to couple light from tapered optical fiber into the on-chip optical waveguide. In this case, the converter is in the form of polymer (SU-8 photoresist) waveguides (ref. Fig. 4b) whose cross-section and resulting mode profile is comparable to the spot emerging from the tapered optical fiber. Light is then adiabatically coupled from polymer waveguide into the silicon waveguide using the inverse taper geometry, and then from the silicon waveguide into the nanobeam cavity using adiabatically tapered photonic crystal [10, 11].

Efficient coupling of light into the cavity is achieved by using hole taper similar to the one described earlier [11]. Light transmitted through the system is collected by a second spot-size converter and tapered fiber. Figure 4b shows an optical micrograph of such a fabricated structure with the input coupling pads, silicon waveguides, cavities, and output coupling pads. Inset shows the zoomed-in image of the polymer coupling pad. Coupling efficiencies as high as 80 % over more than 150 nm could be achieved using this technique. The transmission spectrum shown in Fig. 4c features the excellent quality of signal to noise achieved using the above technique. The inset shows the zoomed region around the cavity resonance.

There have been quite a few variations in the design of the periodic structure of the ultra-high Q nanobeam cavity, including ladder cavities [18] with rectangular holes, bookshelf/domino structures [18], and parabolic width-taper cavity [19]. However, the working principle of all of these approaches is the same and is based on efficient effective index matching between cavity and mirror regions. Another common feature of all the different design strategies is that they are all intuition-based and involve extensive parameter search and trial-and-error-based optimization in order to find the optimal taper profile (the radius and spacing of the holes in the taper section). To overcome these limitations and streamline the design process, thus significantly reducing the time needed for the design of ultra-high Q structures, a deterministic method to design ultra-high Q cavities has recently been proposed and experimentally demonstrated [10]. This approach not only eliminates trial-and-error-based shifting and resizing of holes, but also results in cavities that feature high transmission at the cavity resonance: $Q > 10^7$ with transmission of 97 % (Fig. 5).

As mentioned in the introduction, nanobeam cavities enable realization of ultra-high Q cavities, with small mode volumes, even in materials with low refractive index. This is typically very hard to achieve using 2-D planar photonic crystals since the bandgap is typically closed due to low-index contrast. Nanobeam cavities do not suffer from this problem since 1-D bandgap is always open even for the smallest index modulation. Recently, high-Q all-polymer cavities with index contrast between core (cavity material) and cladding as low as 1.15 have been experimentally demonstrated [6].

Finally, it has been demonstrated that photonic crystal nanobeam cavities also play a role of *phononic crystals* and are capable of localizing phonons [9]. Furthermore, simultaneous confinement of photons and phonons allows for strong interaction between mechanical and optical excitations, thus enabling both exploration of exciting physics at nanoscale and realization of novel

devices. For example, electromagnetically induced transparency (EIT) in photonic/phononic crystal nanobeam cavity was demonstrated and successfully used to alter the group velocity of light thus realizing optical delay lines of interest for optical buffering, filtering, and signal processing [20].

Applications

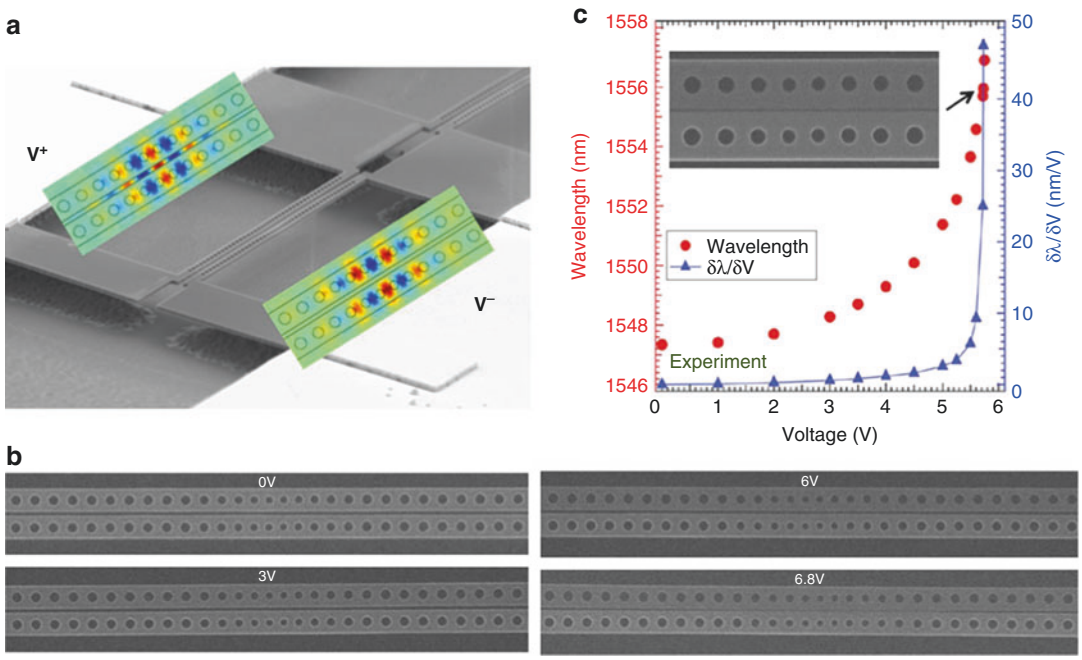
One promising application of nanobeam cavities is realization of compact optical filters and spectrometers. However, due to fabrication limitations, some degree of post-fabrication wavelength trimming is needed in order to bring the resonance of the filter (cavity) to the desired wavelength. Moreover, optical devices that could be dynamically and reversibly configured to operate at vastly different wavelengths (optical analog to FPGAs) would be of great interest for practical applications. Conventional techniques, based on thermal effects and free carrier injection, suffer from high steady-state power consumption and small tuning range, respectively. These limitations can be overcome by taking advantage of mechanical degrees of freedom of suspended photonic crystal nanobeam resonators.

Using coupled photonic crystal nanobeam cavities [21], a reconfigurable photonic crystal filter [22] that can be dynamically, continuously, and reversibly tuned over a wide wavelength range was proposed and demonstrated. These devices, which combine the fields of NEMS and nanophotonics, consist of two doubly clamped photonic crystal nanobeam cavities separated by a gap smaller than 100 nm. Coupling between cavity modes results in two super modes with even and odd symmetry [8, 21] – the situation analogous to coupled quantum wells.

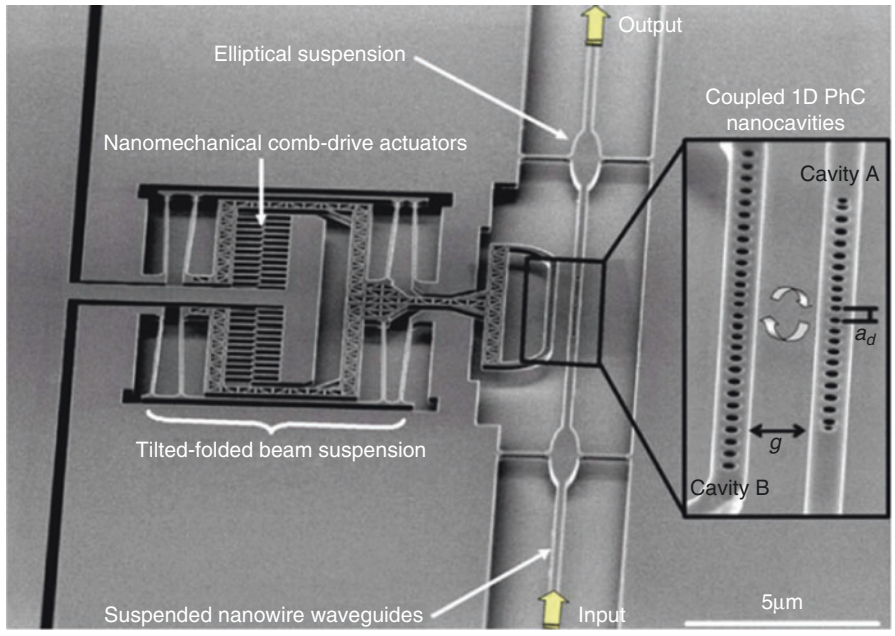
The symmetric mode of such coupled system is extremely dispersive to the separation between the nanobeams which allows for the tuning of this mode by controlling the separation. On the other hand, the asymmetric mode of the system is virtually insensitive to the changes in the separation, which can be attributed to higher-order effects such as a coupling-induced frequency shift [21]. The separation between nanobeams can be controlled using capacitive force, for example: An external bias voltage can be applied to two gold electrodes (Fig. 6a) that are in contact with Si with which the cavities were made of. In this way, the two cavities form the plates of a parallel plate capacitor. The capacitive force is strong enough to bend the beams toward each other as shown in Fig. 6b, which in turn has a strong effect on the resonant wavelength of the structure. Using this approach, tunable filters with a tuning range of $\sim$10 nm for less than 6 V bias and negligible steady-state power consumption were demonstrated. The latter is a very important feature, because once reconfigured, *they do not dissipate any power*. In addition, due to their low mass and flexibility, the energy required to deflect nanobeams is extremely low, and is estimated to be well below the femtojoule level. Therefore, nanoscale optomechanical devices can be used to realize programmable optical systems that can be pre-configured (and reconfigured) using electrical signals, and will likely find applications in low-power optical circuit switching and programmable photonics. It should also be noted that when properly biased, only a few mV of additional electrical signal is enough to move the nanobeams and induce the shift in cavity resonance larger than one linewidth, thus completely detuning the cavity from the incoming optical signal (originally in resonance with the structure) [22].

Significant improvements in the tuning range can be obtained by using more sophisticated actuation methods. For example, Fig. 7 shows a tuning technique that uses a mechanical comb drive to move one of the cavities thus changing the gap between the cavities and resonance of the system [23]. This technique can achieve extremely small gaps as it does not suffer from the pull-in effects which can be detrimental for capacitive actuation approach discussed above.

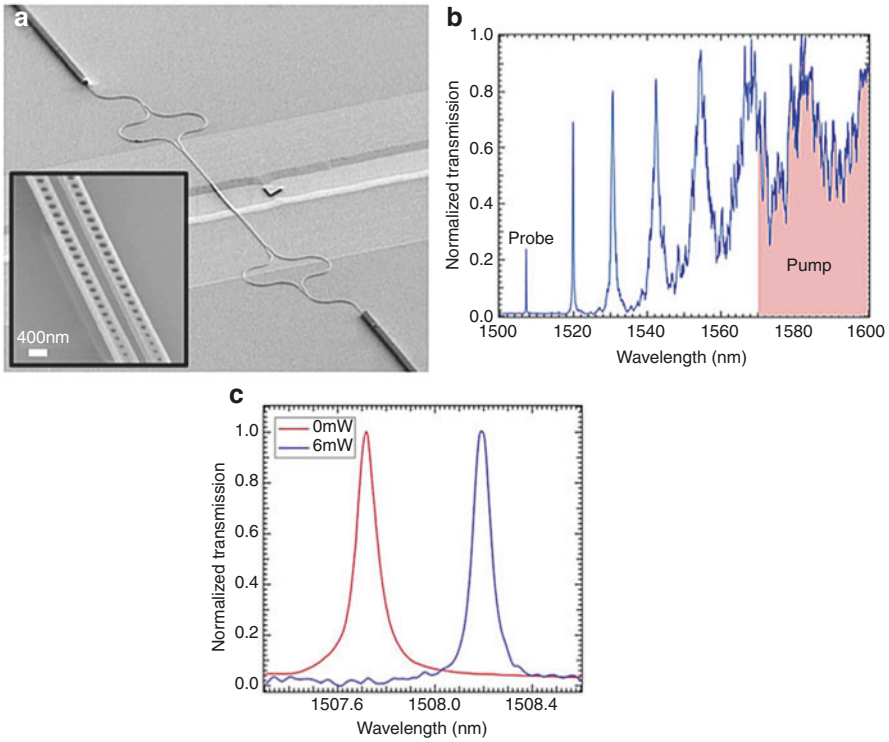
In addition to electrical actuation, nanoscale optical devices can be actuated using light [8, 24]. Figure 8a shows a structure consisting of freestanding coupled nanobeam cavities interfaced with optical waveguides and fiber couplers (for efficient in- and out-coupling of light). Optical transmission through the system



Photonic Crystal Nanobeam Cavities, Fig. 6 (**a**) SEM image of the fabricated structure with gold contact pads. The *insets* show the field profiles for even and odd super modes. (**b**) SEM images showing the reduction of gaps for different applied potentials between the two nanobeams; (**c**) *red graph* shows the tuning of the even mode for various applied voltages while the *blue graph* plots the sensitivity at the operating points [22]



Photonic Crystal Nanobeam Cavities, Fig. 7 SEM image showing the PCNCs and comb drive used to achieve tuning of the resonant wavelength [23]

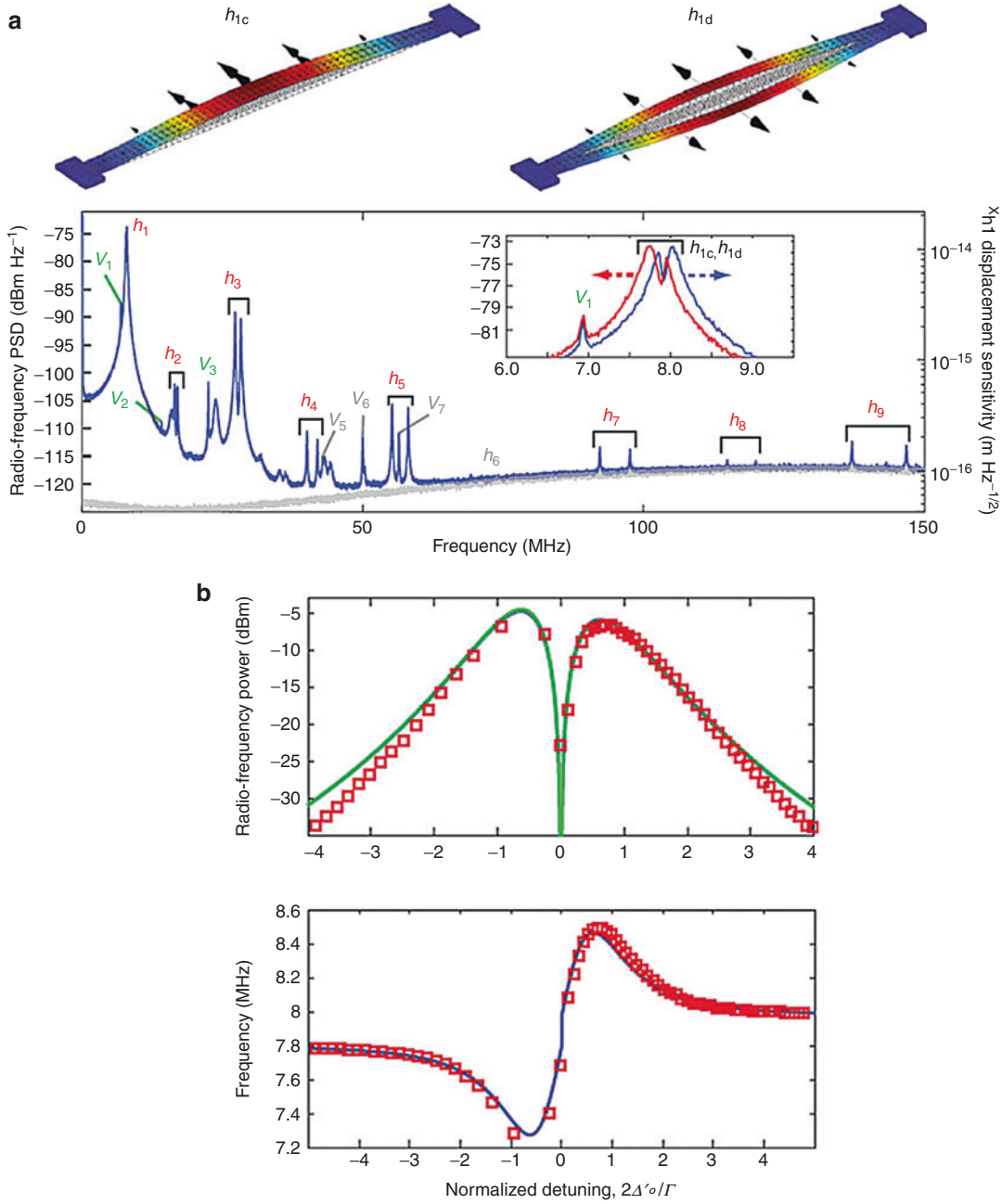


Photonic Crystal Nanobeam Cavities, Fig. 8 Integrated optomechanical system based on coupled nanobeam cavities (**a**) SEM image showing the complete device with the fiber-coupling pads, interferometer, waveguides, and the suspended nanobeam cavity region (inset). (**b**) A transmission spectrum of the fabricated cavity. The even cavity mode – *probe* – is at ~1,507 nm (Q ~15,000) and the *shaded* region can be used to pump outside the bandgap. (**c**) Cavity resonance was tuned from 1,507.72 to 1,508.19 nm when 6 mW of *pump* power at 1,583 nm was used. Cavity tuning was due to contribution from thermo-optical, nonlinear, optical, and optomechanical effects

(experimental) is shown in Fig. 8b, whereas the tuning of cavity resonance for 6 mW of pump power at 1,583 nm is shown in Fig. 8c. The cavity tuning was due to contributions from thermo-optical, nonlinear optical, and optomechanical effects [24].

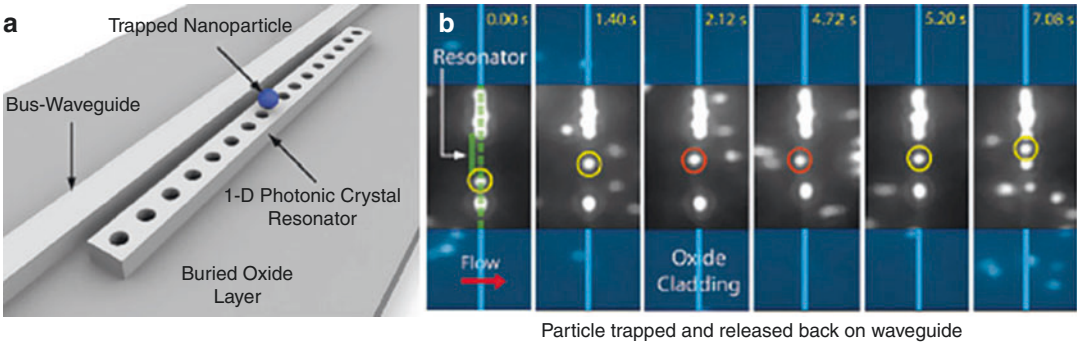
Alternatively, optomechanical systems can be used to read out mechanical motion optically. Figure 9 shows the mechanical spectrum of the structure in air (mechanical modes excited by Brownian motion), detected by monitoring optical transmission through the device. However, it is important to note that this measurement also affects the mechanical properties of the devices, as seen in Fig. 9: Optical signal can stiffen or soften mechanical resonances (induce an additional effective spring constant) via optical gradient force [8].

Owing to their small mode volume and footprint, nanobeam cavities are also excellent candidates for the realization of low-threshold nanolasers with a large spontaneous emission (β) factor [19, 25]. Furthermore, it is possible to realize tunable/programmable nanolasers by taking advantage of mechanical degrees of freedom of coupled nanobeam cavities, as discussed above [26]. PCNCs can also be designed to support TE and TM modes simultaneously, which is of interest for nonlinear frequency conversions [12]. Scaling properties of photonic crystals can be used in PCNCs to design cavities in microwave regime allowing for integrating of RF with photonics. Finally, PCNCs are excellent candidates for applications in sensing [6, 7] and for manipulation of particles suspended in fluid (optical trapping). Figure 10a shows the schematic of the device



Photonic Crystal Nanobeam Cavities, Fig. 9 Zipper cavities (**a**) FEM simulation showing the lowest-order common and differential mechanical modes. Detected radio frequency spectrum showing various in-plane and out-of-plane mechanical modes of zipper structure. (**b**) Power and resonance frequency of fundamental mechanical mode (common) for various optical probe detunings

while Fig. 10b reports snapshots from a movie showing particle manipulation. Polystyrene particles were propelled along the waveguides by the guided modes and trapped by the intense field in the cavity region when the laser was tuned to the cavity resonance.



Photonic Crystal Nanobeam Cavities, Fig. 10 (**a**) Schematic of side-coupled PCNC used for trapping polystyrene particles; (**b**) snapshots showing particle being trapped by the cavity field and propelled by the guided modes of the waveguide [26]

Outlook

Nanobeam photonic crystal cavities have developed significantly since they were first proposed more than a decade ago. Greater understanding and improvements in the design methods along with the advancement in nanofabrication processes has significantly helped research outgrow the challenges in making ultra-high Q wavelength-scale optical resonators. The last few years of research has seen nanobeam photonic crystal cavities mature as a photonic device and allowed researchers to not only solve current technological problems but also explore new research areas involving enhanced light-matter interaction. Their use in tunable filters, switches, delay lines, sensing, particle manipulation, etc., has already been demonstrated. Their wavelength-scale dimensions, flexible structure, and small footprint area along with the scalable nature make them ideal to be used for applications in the entire electromagnetic spectrum. Emerging areas of research include RF, Terahertz, and Mid-IR photonics.

Cross-References

- ▶ Capacitive MEMS Switches
- ▶ Light Localization for Nano-optical Devices
- ▶ Nanophotonic Structures for Biosensing
- ▶ Nanostructures for Photonics
- ▶ Whispering Gallery Mode Resonator Biosensors

References

1. Vahala, K.J.: Optical microcavities. Nature **424**(6950), 839–846 (2003)
2. Yablonovitch, E.: Inhibited spontaneous emission in 3 dimensionally modulated periodic dielectric structures. J. Phys. **48**(C-5), 615–616 (1987)
3. John, S.: Strong localization of photons in certain disordered dielectric superlattices. Phys. Rev. Lett. **58**(23), 2486–2489 (1987)
4. Joannopoulos, J.D., et al.: Photonic Crystals: Molding the Flow of Light. Princeton University Press, Princeton (2008)
5. Gong, Y., Vuckovic, J.: Photonic crystal cavities in silicon dioxide. Appl. Phys. Lett. **96**(3), 031107 (2010)
6. Quan, Q.M., et al.: High-Q, low index-contrast polymeric photonic crystal nanobeam cavities. Opt. Express **19**(22), 22191–22197 (2011)
7. Mandal, S., Serey, X., Erickson, D.: Nanomanipulation using silicon photonic crystal resonators. Nano Lett. **10**(1), 99–104 (2010)
8. Eichenfield, M., et al.: A picogram- and nanometre-scale photonic-crystal optomechanical cavity. Nature **459**(7246), 550–U79 (2009)
9. Eichenfield, M., et al.: Optomechanical crystals. Nature **462**(7269), 78–82 (2009)
10. Quan, Q.M., Deotare, P.B., Loncar, M.: Photonic crystal nanobeam cavity strongly coupled to the feeding waveguide. Appl. Phys. Lett. **96**(20), 203102 (2010)
11. Zain, A.R.M., et al.: Ultra high quality factor one dimensional photonic crystal/photonic wire microcavities in silicon-on-insulator (SOI). Opt. Express **16**(16), 12084–12089 (2008)
12. McCutcheon, M.W., et al.: High-Q transverse-electric/transverse-magnetic photonic crystal nanobeam cavities. Appl. Phys. Lett. **98**, 111117 (2011)
13. Villeneuve, P.R., et al.: Air-bridge microcavities. Appl. Phys. Lett. **67**(2), 167–169 (1995)

14. Foresi, J.S., et al.: Photonic-bandgap microcavities in optical waveguides. Nature **390**(6656), 143–145 (1997)
15. Lalanne, P., Hugonin, J.P.: Bloch-wave engineering for high-Q, small-V microcavities. IEEE J. Quantum. Electron. **39**(11), 1430–1438 (2003)
16. McCutcheon, M.W., Loncar, M.: Design of a silicon nitride photonic crystal nanocavity with a quality factor of one million for coupling to a diamond nanocrystal. Opt. Express **16**(23), 19136–19145 (2008)
17. Deotare, P.B., et al.: High quality factor photonic crystal nanobeam cavities. Appl. Phys. Lett. **94**(12), 121106 (2009)
18. Notomi, M., Kuramochi, E., Taniyama, H.: Ultrahigh-Q nanocavity with 1D photonic gap. Opt. Express **16**(15), 11095–11102 (2008)
19. Ahn, B.H., et al.: One-dimensional parabolic-beam photonic crystal laser. Opt. Express **18**(6), 5654–5660 (2010)
20. Safavi-Naeini, A.H., et al.: Electromagnetically induced transparency and slow light with optomechanics. Nature **472**(7341), 69–73 (2011)
21. Deotare, P.B., et al.: Coupled photonic crystal nanobeam cavities. Appl. Phys. Lett. **95**(3), 31102 (2009)
22. Frank, I.W., et al.: Programmable photonic crystal nanobeam cavities. Opt. Express **18**(8), 8705–8712 (2010)
23. Chew, X.Y., et al.: Dynamic tuning of an optical resonator through MEMS-driven coupled photonic crystal nanocavities. Opt. Lett. **35**(15), 2517–2519 (2010)
24. Deotare, P.B., et al.: All optical reconfiguration of optomechanical filters. Nat. Commun. (2012). doi:10.1038/ncomms1830. accepted
25. Zhang, Y., et al.: Photonic crystal nanobeam lasers. Appl. Phys. Lett. **97**(5), 051104 (2010)
26. Perahia, R., et al.: Electrostatically tunable optomechanical "zipper" cavity laser. Appl. Phys. Lett. **97**(19), 191112 (2010)

Photonics

► Biomimetics of Optical Nanostructures

Photopolymerization Induced by Localized Surface Plasmon Resonance

► Nanoplasmonics Involved in Localized Photopolymerization

Photovoltaic

► Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Photovoltaic Devices

► Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components

Physical Colors

► Nanostructures for Coloration (Organisms Other Than Animals)

Physical Dry Etching

► Dry Etching Processes

Physical Modification

► Nanostructures for Surface Functionalization and Surface Properties

Physical Vapor Deposition

Yoke Khin Yap and Dongyan Zhang
Department of Physics, Michigan Technological University, Houghton, MI, USA

Synonyms

Anodic arc deposition; Arc discharge; Boron nitride nanotubes (BNNTs); Carbon nanotubes

(CNTs); Catalyst; Cathodic arc deposition; Chemical beam epitaxial (CBE); Electron beam evaporation; Electron beam physical vapor deposition (EBPVD); Gallium arsenide (GaAs); Gas-phase molecular beam epitaxy (gas-phase MBE); Graphene; Ion beam; Magnetron sputtering; Metal-organic molecular beam epitaxy (MOMBE); Molecular beam epitaxy (MBE); Nanorods; Nanotubes; Nanowires; Plasma; Pulsed laser deposition (PLD); Si nanotubes; Sputtering; Thermal evaporation

Definition

Physical vapor deposition (PVD) is referred to deposition processes of thin films and nanostructures through the evaporation of solid precursors into their vapor phase by physical approaches followed by the condensation of the vapor phase on substrates. The whole process consists of three stages: (1) evaporation of the solid source, (2) vapor phase transport from the source to the substrates, and (3) vapor condensation on the substrates. Since PVD technique can convert solid materials into vapor phase without chemical processes, they are very convenient and possible to deposit many types of materials into thin films and nanostructures.

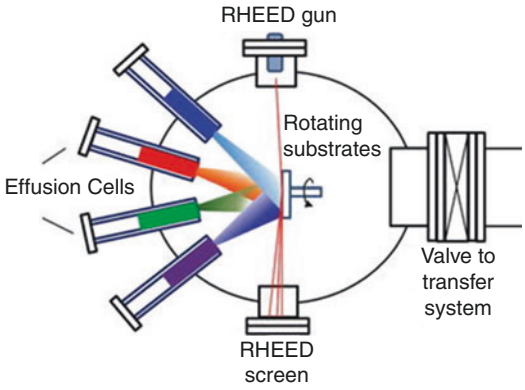
Classification

PVD techniques can be classified based on the techniques used to evaporate the solid source materials into their vapor phase [1, 2]. These techniques include (1) thermal evaporation, (2) ion sputtering, and (3) arc discharge. In the following subsections, examples of PVD based on these evaporation techniques will be described.

PVD by Thermal Evaporation

Molecular Beam Epitaxy

Molecular beam epitaxy (*MBE*) is a PVD technique conducted under ultrahigh vacuum (UHV, better than 10^{-9} mbar) where epitaxial deposition of thin film crystal on crystalline substrates can be obtained by using one or more vapor sources (*effusion cells* or sometimes called *Knudsen cells*). An effusion cell consists of a crucible (made of graphite, pyrolytic boron nitride, quartz, or tungsten) where the solid source can be heated and evaporated by hot filaments (Fig. 1). For example, powders of gallium (Ga) and arsenic (As) can be evaporated in separated cells until they are sublimate into vapors. When these separated beams of element vapors are condensed on a heated Si substrate, crystallized thin films of gallium arsenide (GaAs) will be formed. A typical MBE system has a reflection high-energy electron diffraction (RHEED) system to provide in situ monitoring of the crystal growth processes.



Physical Vapor Deposition, Fig. 1 Schematic drawing of a typical molecular beam epitaxy system

The major advantage of MBE comes from its controllable deposition rate at the atomic or molecular scale. This can be achieved due to the use of effusion cells that enable the generation of "molecular beam," where the generated vapors have low interparticle collisions before they reach to the substrate surface. This means the evaporated atoms escape from the orifice of the cells through effusion and have long mean free path. They neither interact with each other nor with the gases in the vacuum chamber. Due to this capability, MBE was recently used for the growth of graphene, but the outcome is debatable [3, 4]. Finally, gases were also used as the precursors instead of solid sources, and such configuration is called *gas-phase MBE* or more appropriately *chemical beam epitaxial* (*CBE*) [5].

In this case, the pyrolysis of either trimethylindium (TMIn) or triethylindium (TEIn) and trimethylgallium (TMGa) or triethylgallium (TEGa) at the heated substrates was used as the In and Ga source, while As_2 and P_2 were obtained by thermal decomposition of the trimethylarsine (TMAs) and triethylphosphine (TEP). Such approach is more like a combined MBE and *metal-organic chemical vapor deposition* (*MOCVD*) and is sometimes called *metal-organic molecular beam epitaxy* (*MOMBE*).

In fact, *MBE* is an advanced form of thermal evaporation technique with highly controllable deposition thickness (and thus with the drawback of slow deposition rate) due to the use of effusion cells and high vacuum. For the deposition of amorphous thin films, much simple setup can be employed with less demanding vacuum level ($\sim 10^{-2}$ mbar in a glass bell jar chamber). This approach is simply called *direct resistive evaporation* and can be achieved by passing a large current through a resistive wire or foil containing the solid material to be deposited. The heating element is often referred to as an "evaporation source." The evaporation source can appear in many forms including tungsten wires in the form of filaments or baskets and thin tungsten, tantalum, or molybdenum foils in the form of boats.

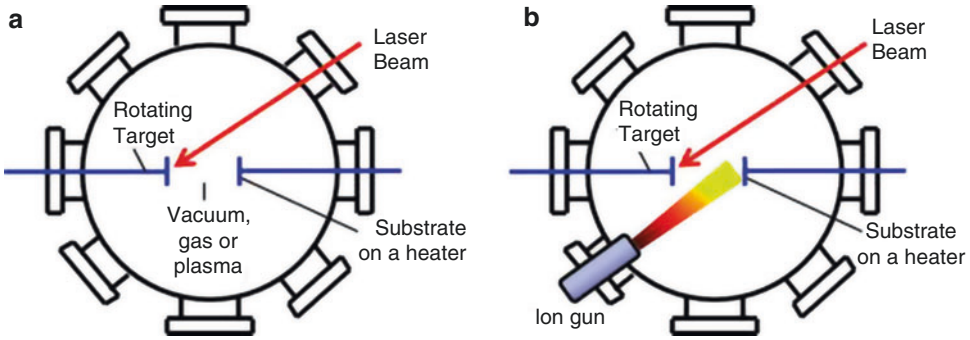
PVD by Electron Beam (E-Beam) Evaporation

In contrast to *MBE*, e*lectron beam physical vapor deposition* (*EBPVD*) provides very high deposition rate on large deposition areas. *EBPVD* is referring to thin film deposition initiated by electron beam evaporation. The deposition rate is typically ranging from tens of nanometer/minute to hundreds of micrometer/minute, although a rate as high as 50 μm/s was reported [6]. *EBPVD* is based on electron beam evaporation of anode materials (usually in ingot form) in high vacuum (10^{-4} mbar). A typical *EBPVD* system can have multiple electron guns and sources. The electron beams can be generated by either thermionic emission or field emission of filaments. These beams are then accelerated to a high kinetic energy and focused on the ingot. The energy of these electrons is converted into thermal energy as the beam bombards the surface of the ingot. The high deposition rate of *EBPVD* without corrosive products meets the needs of industrial applications. A potential drawback of *EBPVD* is that filament degradation in the electron gun results in a nonuniform evaporation rate. In addition, *EBPVD* may cost contamination of the deposition chamber as the vapors generated from the source materials are coated everywhere in the vacuum chamber, the so-called line-of-sight process.

Pulsed Laser Deposition

Pulsed laser deposition (*PLD*) is a *PVD* technique where a high-power pulsed laser is used to generate vapors (i.e., laser ablation) from a solid target followed by the condensation of the vapors on substrates. The experimental setup for *PLD* is relatively simpler than many PVD techniques, but the laser ablation mechanism is quite complex [7]. As shown in Fig. 2a, a *PLD* system consists of a pulsed laser, a vacuum chamber with a rotating target holder, and a substrate holder with a heater. *PLD* can be performed in vacuum at various substrate temperatures. During film deposition, laser pulses are focused on the surface of a solid target through a viewport of the chamber by using a thin lens. This will generate a laser spot with very high *irradiance* (*I*, power per unit area) and local electric field, where $I = \frac{1}{2}\varepsilon_o c E_o^2$ [8]. Here, E_o is the amplitude of electric field of the laser, c is the speed of light in vacuum, and ϵ_0 is the vacuum permittivity. The local electric field is sufficient to cause atomic bond breaking and dielectric breakdown of the target materials to initiate vaporization of the target materials. This vapor will absorb the remaining portion of the laser pulse, and the electric field of the pulse will accelerate electrons and ions inside the vapor to create rapid heating and to form laser plasma. Continued absorption of the laser pulse will cause a highly directional expansion of the laser plasma perpendicular to the target surface. This expanding vapor is called laser plume that propagates toward the substrate placed several centimeters opposite to the target.

PLD can be conducted in vacuum or in a low-pressured gas ambient. When *PLD* is conducted in reactive gas ambient (say oxygen or nitrogen), the technique is called *reactive PLD*. In this case, the expanding laser plume



Physical Vapor Deposition, Fig. 2 Schematic drawings of a typical pulsed laser deposition (*PLD*) system operated (**a**) in vacuum or with auxiliary plasma or (**b**) with auxiliary ion beam

will form shock waves and cause chemical reaction between the target vapor and the gas molecules. *PLD* can also be modified into various configurations. For example, plasma can be applied to enhance the interaction of the vapors with the gas ambient, and the technique is called *plasma-enhanced or plasma-assisted PLD*. For example, such an approach was shown to control the phases of nitride thin films [9, 10]. Sometimes, an auxiliary ion beam is irradiated on the surface of the substrates during thin film deposition as shown in Fig. 2b, and the technique is called *ion beam-assisted PLD*.

In fact, the mechanism of laser ablation and the quality of samples deposited by PLD depend strongly on the wavelength and energy of the pulsed laser. This can be understood by the absorption of the laser light by the laser plasma when the plasma index of refraction, $n(\omega)$, becomes complex. According to the relation $n^2(\omega) = 1 - [\omega_p/\omega]^2$, a complex $n(\omega)$ will occur when $\omega_p > \omega$ [8]. This means, shorter laser wavelength (larger ω) will reduce the chances of such plasma heating effect. However, one should also consider that the laser plasma frequency, $\omega_p = [N\ e^2\ m_e\ \varepsilon_0]^{1/2}$, where m_e is the mass of electron and N is the electron density. In this case, to avoid plasma heating, N must be minimized, which can be obtained by keeping the laser energy or irradiance low. Plasma heating will lead to the so-called photothermal ablation. For example, photothermal ablation of h-BN target will transfer h-BN target to the substrates and prevent the formation of cubic phase (c-BN) [8]. Photothermal ablation of graphite target at longer laser wavelength will also prevent the formation of diamond-like carbon film [11] as the sp^2 species from the graphite are transferred to the substrates.

PVD by Sputtering

Sputtering is a process where atoms are released from a solid target material due to bombardment of the target by energetic particles such as ions and atoms. The atoms released from the target can then be deposited as thin film on substrates and is commonly referred as PVD by sputtering. The physics behind the process is momentum transfer from the incident ions/atoms to the target materials through collisions. The average number of atoms ejected from the target per incident particle is called the sputter yield. The yield depends on the ion incident angle, the energy of the ion, the masses of the ion and target atoms, the surface binding energy of atoms in the target, and the crystalline orientation of the target when crystals are used as the target. There are many approaches to generate sputter including the use of DC, AC, and RF plasmas or external ion beams.

Sputtering offers several advantages over other PVD techniques including (1) large area deposition as compared to PLD, (2) convenience for thin film deposition of alloy/composite and materials with high melting points, and (3) avoidance of device damaging from X-rays generated by electron beam evaporation. However, sputtering also has the following disadvantages: (1) more

complicated setup than PLD, (2) contamination more likely to occur due to the use of plasma as well as relatively lesser vacuum level, and (3) the film morphology that may be rougher and even damaged due to bombardment of energetic growth species and clustering of the growth species in the relatively high deposition pressures.

PVD by DC, AC, RF, and Magnetron Sputtering

PVD can be obtained by applying plasma in various configurations. The simplest one is to apply a DC potential across the substrate (anode) and target (cathode). In this case, plasma can be generated when gases are introduced (10–100s of mTorr) at sufficient electric field strength ($\sim$ a few kV/cm). For example, argon gas is often used, as argon ion can lead to high sputter yield and uniform glow discharge. These Ar ions will bombard the target surface to generate vapors that will be deposited to the substrates. Despite the simplicity, the DC configuration will have the following issues: (1) positive charges will build up on the target surface and cause sparking within the plasma and (2) not suitable for insulating target. To avoid these issues, AC and RF configurations are often used. In these cases, the potential between the target and substrate is alternating so that the target will be negative in potential more often (to generate sputtering) than positive potential. These configurations will also allow lower operation pressure ($\sim$10–30 mTorr).

The use of magnetron can further enhance the deposition rate at even lower pressures by the use of DC, AC, or RF electric fields. RF potential is of advantage, since it can be used for insulating targets and is applicable in an antenna configuration, that is, without positive bias voltages on the substrates (less electron bombardments and less substrate heating). In this case, the target is placed on the magnetron surrounded by strong electric and magnetic fields. Electrons in the RF plasma will follow helical paths around the magnetic field lines and generate more ionizing collisions near the target surface with the surrounding gas. Thus, more ions will be created and will lead to a higher sputtering and deposition rates. Because of higher ionization rates, plasma can be sustained at a lower pressure.

PVD by Ion Beam Sputtering

Ion beam sputtering (*IBS*) is a PVD technique in which the target is evaporated by an external ion source (ion beam). In a typical ion source, ions are generated by collisions of neutral gas atoms/molecules with electrons. They are then accelerated by the electric field by a grid electrode toward the target. Before these ions leave the source, they are neutralized by electrons from a second external filament. The major advantage of PVD by *IBS* is that the energy and flux of ions can be controlled independently. In addition, the *IBS* approach is applicable to insulating and conducting targets since the flux of "ions" (which are neutralized as they are emitted from the ion beam) that strikes the target is composed of neutral atoms/molecules.

Reactive Sputtering

Like the PLD process, reactive sputtering can be achieved by mixing reactive gases (N_2, O_2, etc.) with the inert gas (Ar, Kr, etc.). While the inert gas is responsible for the sputtering of the target materials, the reactive gases are used to initiate chemical reactions with the target vapors to form oxide or nitride films. Like *PLD*, reactive sputtering can also be achieved by various configurations. For example, plasma can be applied to enhance the interaction between the target vapors and the reactive gas ambient, and the technique is called *plasma-enhanced or plasma-assisted deposition*. Sometimes, an auxiliary ion beam is irradiated on the surface of the substrates during thin film deposition, and the technique is called *ion beam-assisted deposition* (*IBAD*).

PVD by Arc Discharge

In addition to PVD by thermal evaporation and sputtering, the desired vapors for thin film deposition can be obtained by arc discharge of the target. This is sometimes called *arc vapor deposition* [1], where a high-current (10^4–10^6 A/cm^2), low-voltage (tens of voltages) discharge was applied on the target. For example, *cathodic arc deposition* refers to arc vapor deposition using the target as the cathode. In this case, a high-voltage "trigger arc" will be ignited between the cathode and a sharp auxiliary anode. This will form "seed"

electrons and ions to initiate a low-voltage, high-current discharge between the target cathode and the adjacent anode.

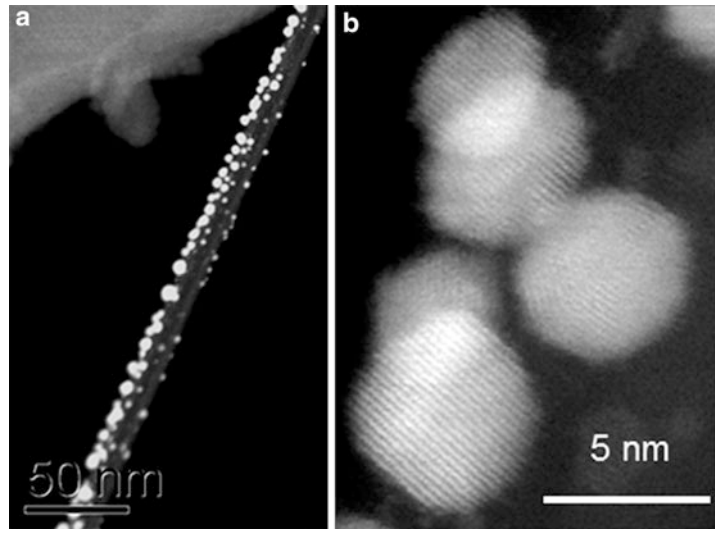
On the other hand, target materials can be evaporated by using the target as the anode of the arc discharge. This PVD process is referred as *anodic arc deposition*. In fact, this process is similar to e*lectron beam physical vapor deposition* (*EBPVD*) discussed earlier, where arc discharge is generated by irradiating an unfocused electron beam on the target anode. The electrons can be made to spiral in a magnetic field to further ionize the evaporated materials from the target prior to deposition. Like most other PVD processes, *reactive arc vapor deposition* can be performed by both *anodic arc deposition* and *cathodic arc deposition* using reactive gases like nitrogen and oxygen. This will lead to the formation of nitride and oxide films.

The deposition rates of *arc vapor deposition* are usually higher than that of PVD by sputtering. This is one of the advantages of *anodic arc deposition* and *cathodic arc deposition* in addition to the low operation voltage. On the other hand, contamination is the potential drawback of most plasma-based PVD processes. In addition, particulates or the so-called macros are undesired for thin films deposited by *anodic arc deposition* and *cathodic arc deposition*. Macros are formed by ablation of molten or solid particles by thermal shock during arc discharges.

Examples of PVD Approaches for Nanotechnology

Physical evaporation techniques like arc discharge [12] and laser ablation [13] are important on the discovery of carbon nanotubes (CNTs). However, PVD techniques discussed here are not popular for the growth of nanomaterials in comparison to chemical vapor deposition (CVD). The major reason is the tendency of forming thin films and amorphous by-products on the surfaces of catalysts which cause "poisoning" effect and prohibit the formation of nanotubes or nanowires [14]. However, there have been some pioneering works reported on the growth of nanotubes and nanowires by PVD techniques such as sputtering, PLD, and MBE. For example, vertically aligned carbon nanotubes (CNTs) were first successfully deposited by RF plasma-assisted pulsed laser deposition [15]. In this approach, two RF plasmas were employed. The first one was applied on the substrates to form a negative bias voltage during deposition. The second one was applied on a ring electrode located between the graphite target and the substrates. Hydrogen gas was used for the plasma generation, and Fe powders (20 nm in diameter) were employed as the catalysts. Although the density of these CNTs is low, these nanotubes have high structural order with internal nanotube diameter as small as 0.4 nm. The formation of these smallest possible CNTs was confirmed by transmission electron microscopy (TEM). Later, CNTs were deposited by magnetron sputtering with a dc bias voltage on substrates [16]. Prior to the growth of CNTs, Ni films (10 nm thick) were deposited by e-beam evaporation to form nanoparticles that later function as the catalyst for the vapor–liquid–solid (VLS) process. However, these CNTs are defective with bamboo-like structures. Later, RF plasma-assisted pulsed laser deposition was used to achieve low-temperature growth of boron nitride nanotubes (BNNTs) [14, 17]. In this case, pyrolytic boron nitride pallets were used as the targets, and a RF plasma was applied on the substrates to create bias voltages, which are important to eliminate the formation of boron nitride thin films. Again, these BNNTs have internal nanotubes of 1 nm in diameter with high structural order.

Various PVD approaches were also employed for the synthesis of nanowires/nanorods. For example, ZnO and Zn_{1-x} Mg_x O nanorods were grown by MBE on oxidized Si substrate coated with Ag catalyst [18]. In both cases, an ozone/oxygen mixture was used as the oxidizing source. The Zn and Mg cation fluxes were provided by a Knudsen effusion cell using high purity Zn and Mg metals as the sources. The growth of vertically aligned ZnO [19] and MgO [20] nanowires can also be obtained by PLD using ZnO or MgO targets, respectively. Gold particles were used as the catalysts for both cases.



Physical Vapor Deposition, Fig. 3 (**a**) Low and (**b**) high magnification of gold nanoparticles deposited on one side of a boron nitride nanotubes (Adapted with permission from Lee, C.H. et al., Adv. Mater. **25**, 4544–4548 (2013). Copyright (2013) WILEY-VCH Verlag GmbH & Co)

More recently, PLD was shown to produce gold nanoparticles on the wide band gap (~6 eV) boron nitride nanotubes (BNNTs) [21]. These nanoparticles have functionalized the electrically insulating BNNTs to become the tunneling channel of tunneling field effect transistors (TFETs) without involving any semiconducting properties. Such transistors without the use of semiconductors are interesting as these TFETs by-pass many fundamental issues of semiconducting FETs such as leakage current, short channel effect, and high contact resistance when the devices are scaled down to the quantum limits. Because of the related tunneling effect, these gold nanoparticles' functionalized BNNTs are called quantum dots' (QDs) functionalized BNNTs (QDs-BNNTs). As shown in Fig. 3, gold QDs smaller than 10 nm are preferentially coated on one side of the BNNT.

Acknowledgments Yoke Khin Yap acknowledges the support from the National Science Foundation (Award number DMR-1261910).

Cross-References

- Atomic Layer Deposition
- Carbon Nanotube-Metal Contact
- Carbon Nanotubes for Chip Interconnections
- Carbon-Nanotubes
- Chemical Vapor Deposition (CVD)
- Electron-Beam-Induced Deposition
- Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)
- Graphene
- Growth of Silica Nanowires
- Nanomaterials for Electrical Energy Storage Devices
- Nanomaterials for Excitonic Solar Cells
- Nanotechnology
- Quantum Dot
- Synthesis of Carbon Nanotubes

References

1. Mattox, D.M.: *Handbook of Physical Vapor Deposition (PVD) Processing: Film Formation, Adhesion, Surface Preparation and Contamination Control (Hardcover)*. Noyes, Westwood, N.J (1998)
2. Mahan, J.E.: Physical Vapor Deposition of Thin Films. Wiley-Interscience, New York (2000)
3. Park, J., et al.: Epitaxial graphene growth by carbon molecular beam epitaxial (CMBE). Adv. Mater. **22**, 4140–4145 (2010)
4. Hackley, J., et al.: Graphitic carbon growth on Si (111) using solid source molecular beam epitaxy. Appl. Phys. Lett. **95**, 133114 (2009)
5. Tsang, W.T., et al.: Chemical beam epitaxy of InP and GaAs. Appl. Phys. Lett. **45**, 1234–1236 (1984)
6. Schiller, S., Jäsch, G.: Deposition by electron beam evaporation with rates of up to 50 μm s^{-1}. Thin Solid Films **54**, 9–21 (1978)
7. Chrisey, D.B., Hubler, G.H. (eds.): Pulsed Laser Deposition of Thin Films. Wiley-Interscience, New York (1994)

8. Wang, J., Yap, Y.K.: Growth of adhesive cubic phase boron nitride films without argon ion bombardment. Diamond Relat. Mater. **15**, 444–447 (2006)
9. Yap, Y.K., Kida, S., Aoyama, T., Mori, Y., Sasaki, T.: Influence of negative dc bias voltage on structural transformation of carbon nitride at 600 °C. Appl. Phys. Lett. **73**, 915 (1998)
10. Yap, Y.K., Aoyama, T., Kida, S., Mori, Y., Sasaki, T.: Synthesis of adhesive c-BN films in pure nitrogen radio-frequency plasma. Diamond Relat. Mater. **8**, 382–385 (1999)
11. Yamamoto, K., Koga, Y., Fujiwara, S., Kokai, F., Heimann, B.: Dependence of the sp^3 bond fraction on the laser wavelength in thin carbon films prepared by pulsed laser deposition. Appl. Phys. A **66**, 115–117 (1998)
12. Iijima, S.: Helical microtubules of graphitic carbon. Nature **354**, 56–58 (1991)
13. Iijima, S., Ichihashi, T.: Single-shell carbon nanotubes of 1-nm diameter. Nature **363**, 603–605 (1993)
14. Wang, J., et al.: Low temperature growth of boron nitride nanotubes on substrates. Nano Lett. **5**, 2528–2532 (2005)
15. Yap, Y.K., Yoshimura, M., Mori, Y., Sasaki, T., Hanada, T.: Formation of aligned-carbon nanotubes by RF-plasma-assisted pulsed-laser deposition. Special issue for Tsukuba symposium on carbon nanotubes in commemoration of the 10th anniversary of its discovery; S. Ijima et al. eds. Phys. B. **323**, 341–343 (2002)
16. Lee, K.-F., et al.: Synthesis of aligned bamboo-like carbon nanotubes using radio frequency magnetron sputtering. J. Vac. Sci. Technol. B **21**, 1437–1441 (2003)
17. Xie, M., Wang, J., Yap, Y.K.: Mechanism for low temperature growth of boron nitride nanotubes. J. Phys. Chem. C **114**, 16236–16241 (2010)
18. Heo, Y.W., Kaufman, M., Pruessner, K., Norton, D.P., Ren, F., Chisholm, M.F., Fleming, P.H.: Optical properties of $Zn_{1-x}Mg_xO$ nanorods using catalysis-driven molecular beam epitaxy. Solid-State Electron. **47**, 2269–2273 (2003); Heo, Y.W., Varadarajan, V., Kaufman, M., Kim, K., Norton, D.P., Ren, F., Fleming, P. H.: Site-specific growth of Zno nanorods using catalysis-driven molecular beam epitaxy. Appl. Phys. Lett. **81**, 3046–3048 (2002)
19. Rahm, A., et al.: Pulsed-laser deposition and characterization of ZnO nanowires with regular lateral arrangement. Appl. Phys. A **88**, 31–34 (2007)
20. Nagashima, K., Yanagida, T., Tanaka, H., Kawai, T.: Epitaxial growth of MgO nanowires by pulsed laser deposition. J. Appl. Phys. **101**, 124304 (2007)
21. Lee, C.H., et al.: Room-temperature tunneling behavior of boron nitride nanotubes functionalized with gold quantum dots. Adv. Mater. **25**, 4544–4548 (2013)

Physical-Chemical Etching

► Dry Etching Processes

Physicochemical Properties of Nanoparticles in Relation with Toxicity

Jérôme Rose[1,2], Mélanie Auffan[3,4], Olivier Proux[2,5], Vincent Niviere[2,6] and Jean-Yves Bottero[7]
[1]CEREGE UMR 6635–CNRS–Université Paul Cézanne Aix–Marseille III, Aix–Marseille Université, Aix–en–Provence, France
[2]GDRI ICEINT: International Center for the Environmental Implications of Nanotechnology, CNRS–CEA, Aix–en–Provence, France
[3]CEREGE, UMR 6635 CNRS/Aix–Marseille Université, Aix–en–Provence, France
[4]iCEINT, International Consortium for the Environmental Implications of Nanotechnology, Center for the Environmental Implications of NanoTechnology, Aix–en–Provence, France
[5]OSUG, Grenoble, France
[6]Laboratoire de Chimie et Biologie des Métaux, UMR 5249, iRTSV–CEA Bat. K', Grenoble, France
[7]Aix–Marseille Université, CNRS, IRD, CEREGE UM34, Aix–en–Provence, France

Synonyms

Engineered nanoparticles; Nano-effects; Nanomaterials and nanoproducts; Nanotoxicology

Definition

Enhancement or modifications of properties in the nanometric range is the core of nanotechnologies. Nanoparticles exhibit size-related properties that differ significantly from those observed in fine particles or bulk materials. Properties can be intrinsic (e.g., electrical, magnetic, optical, etc.) or extrinsic (e.g., aggregation state, surface area, etc.). Then what makes nanoparticles so attractive also raises questions of "new" toxic "nano-effects." It is then worth trying to identify which specific properties can generate biological harmful effects.

More than these "new" nano-properties aspects, the second issue is to manufacture materials that incorporate those nanoparticles. Then exposure to nanoparticles will depend not only on intrinsic properties of nanoparticles, but also on extrinsic properties related to the stability and durability of their incorporation.

Overview

Introduction

The success story of nanoparticles in many scientific and technological fields is based on their unique physicochemical properties and reactivity. Indeed engineered nanoparticles are designed to exhibit specific optical, electronic, magnetic, or chemical properties that are strongly related to the exponential increase in the number of atoms localized at the surface as the size decreases [1]. However, the uniqueness of nanoparticles that makes them so attractive may also raise questions about the environmental and health risk they could generate. Indeed is there any relation between their new physicochemical properties and potential new biological hazards?

Moreover the tremendous development of nanotechnologies leads to a huge variety of manufactured nanoparticles with varying chemical composition, size, shape, crystal structure, surface area, surface chemistry and charge, solubility... etc., e.g., [1].

It is therefore challenging to determine which of those unique intrinsic or extrinsic properties can be at the origin of their biological danger. In the following, we will focus on the different properties, how to determine their evolution in biological media and how they can affect living organisms.

Physicochemical Properties of Nanoparticles

Even if there is no consensus concerning the definition of engineered nanoparticles, most of the definitions indicate that besides the size, "*Novel properties differentiate nanoparticles from the bulk material typically develop at a critical length scale of under 100 nm*" or "*phenomena not observed in larger structures start to become apparent . . .*". But what are these novel properties or phenomena not observed for larger structures? What is the size at which the latter properties occur? One of the most emblematic example concerns gold. Indeed gold has been a topic of fascination since its discovery in almost all ancient and modern civilizations. Gold preserve its brightness for years which made his high value for art. Everybody has in mind that the magnificent death mask of Egyptian pharaoh Tutankhamun is made of gold. Moreover, gold chemical inertness is related to its high monetary value [2]. But its chemical inertness is a matter of size since at the nanoscale it can be exceptionally active as a catalyst for instance but only for particles below 2–3 nm [3].

Among the vast diversity of engineered nanoparticles, many other properties can be tailored in different scientific and technological fields. Indeed optical properties can be modified to use nanoparticles as biological dyes for instance. QDots are of good examples to trace, label, or image biological organisms or cells. In the case of QDots, the band gap energy (related to the fluorescence wavelength that is to say the color) is constant for large particles whatever their size but only if larger than 6–8 nm. The bandgap energy strongly increases when size of particles decreases below 6–8 nm. Photocatalytic properties may also be tailored not only as function of size, but also as function as the crystal structure. TiO_2 anatase phase exhibits optimum photocatalytic degradation of organic contaminants at various sizes that are much below 100 nm (e.g., ~7 nm for trichloroethylene or around 11 nm for chloroform). But in any case, anatase is much more efficient than rutile, indicating that besides size, crystal structure also strongly affects physicochemical properties.

Thermal properties may also vary in relation with size decrease. For instance, tin and indium melting point can be strongly reduced by 80 °C and 120 °C, respectively (e.g., in [1]) when their diameters decrease from 100 to 10 nm. More surprisingly, their normalized heat of fusion that is supposed to remain constant in classical

thermodynamics exponentially decreases when size decreases below 20 nm in diameter. Solubility is also affected by size decrease. It is intuitive to predict that the dissolution kinetics of particles is function of the specific area and aggregation state, with of course faster dissolution kinetic for the smallest particles, but what about the thermodynamic crystal solubility K_b? K_b is supposed to be constant and generally similar to the solubility product K_{sp}.

$\ln K_b = \ln K_{sp} + f(\gamma/r)$ (with γ is the surface tension and r is the characteristic length of the crystal).

For particles larger than 25 nm, $f(\gamma/r)$ can be approximated as a constant function. But this simplification fails for smaller nanoparticles as the size dependence of the morphology and γ cannot be ignored.

Other examples of physicochemical properties modification with size could be given (magnetic, electrical, etc.). Besides these intrinsic properties, it is worth detailing nanoparticles when introduced in aquatic media. Indeed some confusion may exist between dissolution and dispersion. Due to their small size, when nanoparticles are homogeneously distributed in aquatic media, particles cannot be detected by eyes and appear as if particles may be totally dissolved to ionic species. However, (colloidal) dispersion is the term referred to suspended material in the 1–250 nm range. Colloidal dispersions are thermodynamically unstable even if the aggregation rate can be very slow.

Latter examples are good illustrations of the high variety of properties that can change or be enhanced as size of particles decreases in the nanometric size range. It is therefore challenging to prioritize which properties are of main concern in relation with toxicity.

Determination of Nanoparticle's Properties in Biological Media

Generalities

Even if challenging, determining the evolution of physicochemical properties of nanoparticles in relation to human and environmental exposure requires to answer to, at least, two questions, in order to define the appropriate useful techniques.

First of all, which structural and characterization information are considered to be important? As detailed before their aggregation in the media, their (mean) size and shape, dissolution or solubility, their surface area, redox stability, surface charge, and surface chemistry are of prime interest.

Secondly, in which conditions the nanoparticles have to be characterized? A three-step procedure have been formalized in which the characterizations need ideally to be performed (1) *on as-synthesized nanoparticles* in their native state (dry or hydrated state, depending on the synthesis), (2) in situ *on nanoparticles dispersed in the media* used for the environment and toxicity tests (ultrapure water, phosphate-buffered saline solution cell culture media. . .), and (3) in vivo or in vitro *on particles interacting with the living systems* tested (bacteria, cells, animals, plants, etc.).

Numerous characterization and detection techniques exist to determine these physicochemical properties of nanoparticles in a toxicological or biological context. There is no best technique, each of them has their specificities, advantages, and disadvantages. However, in order to limit at the maximum the discrepancies between the results, the same technique has to be employed for each step of the characterization. Several recent review papers summarize different and complementary analytical techniques (e.g., [4]. . .). We will here focus on recent techniques using X-ray. Their large penetration depth, especially in the hard X-rays regime, is interesting in a large range of applications because of its compatibility with in situ or in vivo conditions, in most of the cases without any special sample preparation, directly on the nanoparticles in suspension, for example. Moreover, specific interactions of X-rays with matter provide elemental and chemical sensitivity that have made X-ray techniques a very attractive tool.

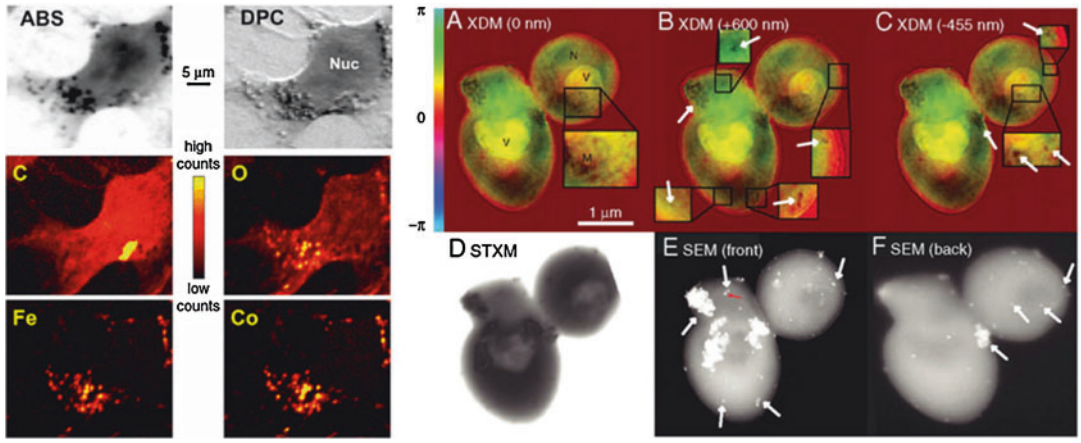
X-Ray Imaging

Advances in X-ray spectromicroscopy techniques based on new third-generation synchrotron

sources have opened unprecedented opportunities in terms of spatial resolution to analyze nano-sized compounds in biological systems [5]. Different possibilities exist to improve the contrast using X-ray microscopes with respect to optical or electron microscopes. They can be schematically described looking at the different interaction processes occurring between the incident photons (or the associated wave) and the sample. (1) The *photon absorption* which strongly depends on the chemical nature of the element (the higher the atomic number of an element, the stronger will be the absorption). (2) The *fluorescence emission* resulting from the decay process. (3) The *phase shift* of the incoming wave induced by the matter/sample interaction. (4) The *photon diffraction* by the sample. These processes give rise to different kinds of X-ray microscopy, among them Transmission X-ray Microscopy (TXM), Scanning Transmission X-ray Microscopy (STXM), X-ray fluorescence with microbeam (μXRF), and X-ray Diffraction Microscopy (XDM).

Examples of imaging performed with these techniques are shown on Fig. 1, on fibroblast cells exposed to nanoparticles using STXM in absorption (ABS) and differential phase contrast (DPC) modes combined with μXRF (Fig. 1, *left*) [5], and on yeast using XDM compared with Scanning Electron Microscopy (SEM) and STXM (Fig. 1, *right*) [6].

First example deals with the toxicity of $CoFe_2O_4$ nanoparticles, nano-sized magnets that can be potentially used in combined cancer therapy if their biocompatibility is demonstrated. μXRF maps indicate that, for the probed concentrations, these nanoparticles do not penetrate the nucleus. Precise localization of the elements with respect to the different constituting parts of the cell is possible by acquiring at the same time an image, the best contrast being obtained in DPC mode (Fig. 1, *left*, upper part). Spatial resolution in this soft-X-ray microscope is sub-50 nm. An improvement of the resolution can be achieved by shifting to DCM microscopy as shown by Nelson et al. [6] who obtained a "resolution between 11–13 nm, the highest resolution in X-ray imaging of a biological specimen." *Saccharomyces cerevisiae* yeasts were imaged after labeling using colloidal gold particles. These labels were positioned to correlate the images obtained by the three different ways, to have a probe to estimate the resolution and the probing depth. The 3D XDM image obtained after reconstruction allows to image the internal part of the cell (which can be compared to the image obtained in transmission using STXM) or the two opposite walls of the cell



Physicochemical Properties of Nanoparticles in Relation with Toxicity, Fig. 1 *Left*. Uptake and interaction of $CoFe_2O_4$ nanoparticles with fibroblast cells imaged using absorption (*ABS*) and differential phase contrast (*DPC*) combined with μXRF analysis of the C, O, Fe, and Co (From [5]). *Right. Saccharomyces cerevisiae* yeast imaged using XDM (three images corresponding to three different analyzing depth), STXM, and SEM (Ref. [6])

(which can be compared in this case to the two images obtained in SEM). This technique is really promising even if it still remains quite challenging, to image thick cells, too thick to be studied by transmission electron microscopy.

From Long-Range to Short-Range Orders in Nanomaterials

X-ray scattering and X-ray Absorption Spectroscopy techniques (XAS) techniques provide valuable information on long- and short-range orders existing inside (X-Ray Diffraction, XRD) and between (Small-Angle X-ray Scattering, SAXS) nanoparticles, with a great sensitivity to their shape.

SAXS experiments have been performed to follow the kinetics of aggregation or dissolution of ZnS nanoparticles as a function of the mean size of the particles and of the pH of the solution [7]. Levard et al. [8] have combined SAXS and XAS in situ experiments to study the growth mechanisms of imogolite-like aluminogermanate nanotubes at various stages of their formation, from the precursor (rooftile-shaped particles) to the nanotubular structure. 20 % of Ge vacancies have been also determined in the Ge layer from XAS analysis. Finally, structure of iron oxide nanoparticles in aqueous suspension was probed by high-energy powder XRD on a large angular range coupled to atomic pair distribution function (PDF) analysis, well suited for determining the internal atomic ordering and the geometry of nanoparticles [9]. This PDF technique provided detailed structural information for γ-Fe_2O_3 spherical and tetrapod-shaped nanoparticles.

These techniques are performed using synchrotron radiation in order to perform kinetics experiments and XAS spectroscopy, but SAXS and XRD experiments can also be performed using laboratory sources, preferably at high energy (Mo or Ag Kα anode).

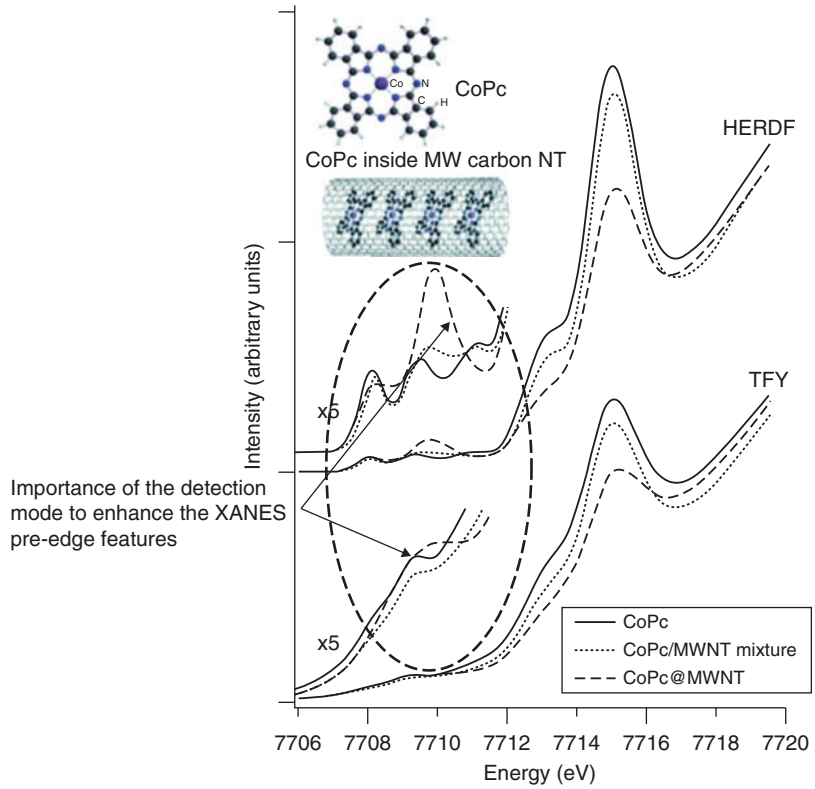
Electronic Structure and Short-Range Order in Nanomaterials

XAS is now a well-established structural technique that allows to determine the electronic structure (valence) and the local order (geometry of the site, chemical nature of the neighbors, etc.) of a selected element in a sample. Being only sensitive to the local order, XAS is well suited to study either amorphous, ill-ordered, or well-crystallized samples, in the solid or liquid states, and has been then quite naturally used for the nanoparticles studies [10]. Numerous examples of the application of this technique for nanotoxicology studies can be found (e.g., [11, 12]). In this particular field, two conditions are required. Acquisitions need to be performed in the fluorescence mode, to be sensitive to a small amount of nanoparticles, using, for example, of conventional solid state detector, with a medium energy resolution (ca. 200–250 eV). Moreover, a high resolution is necessary to detect accurately the subtle variations upon interaction with a solution or a biological media. This second requirement is especially true in the pre-edge and edge part of the signal (X-ray Absorption Near Edge Structure, XANES).

High-Energy Resolution Fluorescence Detection (HERFD) allows to further enhance XANES sensitivity. Even if not extensively used at present, this technique has gained a lot of interests and under development on many synchrotron beamlines. Using a detector with an energy resolution better than the core hole lifetime broadening gives XANES spectra with sharper spectral features compared to conventional total fluorescence (TFY) spectra. Example of such sharpening is illustrated in Fig. 2. The structure of Cobalt phthalocyanine (CoPc), free or encapsulated inside multiwalled carbon nanotubes has been probed using either TFY or HERFD [13]. The differences in the pre-edge using conventional TFY spectra are really small, until those obtained by HERFD clearly put in evidence a modification of the Co conformation when CoP is free or encapsulated. Complete analysis shows that Co atoms move out of the molecular plane by a fraction of Angstrom.

Physicochemical Properties: Biological Effects Relationships

Some recent reviews inventoried NPs that are toxic to living organisms, e.g., [14–16]. There are now obvious examples of the toxic effects of



Physicochemical Properties of Nanoparticles in Relation with Toxicity, Fig. 2 TFY XANES pre-edge and edge region compared to HERFD spectra for three CoPc samples. The pre-edge is magnified by 5 for clarity (From Ref. [13])

C_{60} fullerenes to bacteria, to daphnia, etc., Nano-TiO_2, nano-CeO_2, nano-Ag, etc., can also be toxic to various organisms ([14, 15] etc.), or even colloidal silica to rat and mice [16]. But despite the growing body of literature, many conclusions remain contradictory. For instance, dose–response relationship are not always observed even for equivalent systems. Lovern et al. [17] clearly observed such relationship in the case of daphnia mortality exposed to 25 nm TiO_2, while Hund-Rinke et al. [18] were unable to determine a LC50 value for 30 nm TiO_2 in contact with Daphnia due to chaotic results. In many cases, the degree of physicochemical characterization of nanoparticles is not equivalent and one cannot identify why results are inconsistent. As previously discussed, nanoparticles can differ by many aspects due to variation in size, shape, redox properties, surface charge, crystal structure, chemical composition, adhesion to surfaces, etc., and beyond a list of results, it is worth addressing the effects of various physicochemical parameters of NPs on their toxic effects.

Effects of Aggregation State and Size

Most of pioneer published results dealing with NPs toxicity mainly focused on size effects. Can toxic effects only be due to the small size of particles? Size was mainly measured by TEM that implied dry state characterization. However, the size was not characterized in nutritive or aquatic test media. Indeed the dispersion state of NPs (aggregation vs. dispersion) in contact with living organisms can change from one study to another. At this stage, it is worth detailing that here aggregation is referred to an assembly of particles which are loosely attached to each other. It means that aggregation is a reversible phenomenon and that NPs can aggregate in aquatic media but can disaggregate in presence of stabilizing agents like proteins. Then one has to keep in mind that aggregation can be reversible. There is a generally accepted idea assuming that aggregation of NPs impedes the determination of specific properties of NPs. It is obvious that the transport and mobility are strongly affected. Nevertheless, intrinsic or specific "nano"

physicochemical properties of NPs are not necessarily (or even significantly) affected by aggregation processes. In the particular case of surface chemical reactivity, Auffan et al. [1, 19] have shown that surface normalized quantity of probe molecules sorbed to iron oxide particles is strongly enhanced for particles smaller than 12 nm. Using arsenite as surface reactivity molecular probe, the quantity of arsenite adsorbed to the maghemite (γFe_2O_3) surface per surface unit reached 3.6 As/nm^2 for 300 nm and 20 nm particle size while it increased up to 11 As As/nm^2 for 12 nm particles. This surface reactivity was not affected by aggregation [19]. The example of Nano-Fe_2O_3 in contact with E. coli can illustrate that the aggregation state as well as the size may not be the most important parameters in terms of toxicity [20]. Indeed, [20] have shown that 8 nm maghemite (Fe_2O_3) had no toxic effect toward E. coli regardless of the aggregation state of the NPs.

Effects of the Redox State Evolution and Instability

As function of their chemical composition redox states of NPs can change in biological redox/pH conditions. Indeed most elements exhibit various possible redox states that are or are not stable as function of Eh and pH values. It is obvious for all metals that they can exist under metallic or oxidized forms. But even when oxidized, some elements can exist under various redox states. Metals like Fe, Mn, Cr, or rare earth like Ce illustrate such variability (Fe^{2+}, Fe^{3+}; Mn^{2+}, Mn^{3+}, Mn^{4+}, Mn^{7+}; Cr^{3+}, Cr^{6+}; Ce^{3+}, Ce^{4+}, etc.). Living organisms and especially eukaryotic cells metabolism cannot support large pH variation. For most eukaryotic cells, pH is buffered in the 7–8 pH range. In the case of redox potential (noted Eh), oxidative phosphorylation in mitochondria is based on a series of redox reactions at near circumneutral pH for which potentials are from −0.32 (NAD^+/NADH) to 0.29 V (cytochromes). Besides oxidative phosphorylation (respiration), extracellular Eh is generally controlled by thiol/disulfide redox systems (mainly GSH/GSSH and Cys/CySS) for which Eh vary in the −0.140/−0.08 V range. In such conditions, many elements can be redox unstable leading to electron exchange between NP surface and surrounding media. This could then be the starting point of disequilibrium of the redox balance and then to oxidative stress (See Box 1). In the specific case of iron, Auffan et al. [21] have shown that nano-zero valent iron as well as nano-magnetite (Fe_3O_4) can be toxic to E. coli. The study reported that the oxidation of iron in contact to bacteria can be at the origin of the toxicity. The potential of the redox couple characteristic of Fe^0 nanoparticles at pH = 7 is in the $Eh_{Fe^{2+}/Fe(0)} = -0.40/-0.7$ V range as function of the Fe concentration [22] which is lower than the typical Eh in biological media. This allowed the oxidation of Fe^0(s) in contact with E. coli. This oxidation is directly responsible of toxic effects through the generation of an oxidative stress, as demonstrated by an enhanced toxicity of these Fe NPs toward a mutant strain of E. Coli deprived of defense mechanism against oxidative stress. Other examples in the literature confirm the importance of the redox "instability" of nanoparticles. Using XANES at Ce L3-edge, Thill et al. [12] and Auffan et al. [11] have shown that the cytotoxicity/genotoxicity of CeO_2 was due to the reduction of surface Ce (IV) ions to Ce(III).

Box 1

Oxidative stress is a major problem that concerns all the cells living in the presence of dioxygen. It originates from the production of highly toxic reactive oxygen species (ROS), including superoxide radical $O_2^{\bullet -}$, hydrogen peroxide H_2O_2 and hydroxyl radical $HO^{\bullet}$. Although cells have developed efficient antioxidant systems to detoxify ROS, any deficiency of these systems or overload by overproduction of ROS can result in severe cellular damages such as proteins, lipids, and DNA oxidations, which can ultimately lead to cell death. It is now well documented that oxidative stress is involved in the pathogenesis of a variety of diseases such as cancers, asthma,

(continued)

pulmonary diseases, Alzheimer, Parkinson, and in ageing [1].

$$\underset{oxygen}{O_2} \xrightarrow{e^-} \underset{\substack{speroxide\\ anion}}{O_2^{\bullet -}} \xrightarrow[2H^+]{e^-} \underset{\substack{Hydrogen\\ peroxide}}{H_2O_2} \xrightarrow[OH^-]{e^-} \underset{\substack{Hydroxyl\\ radical}}{OH^{\bullet}} \xrightarrow[H^+]{e^-} H_2O$$

: formation and elimination of ROS

Several studies have related the toxicity of some NPs to an oxidative stress toward biological models [2]. In order to better understand the cellular oxidative stress pathways induced by NPs, identification and quantification of the specific ROS produced directly or indirectly by NPs represent an important issue.

In vivo detection and quantification of ROS are far from being trivial and require the use of specific techniques [3, 4]. This is mainly due to the fact that ROS are unstable species and are present at very low steady-state concentration in the cells. Various methods using synthetic probes have been developed to detect ROS in vivo. Some of these probes lead to a fluorescence signal upon oxidation by ROS, which is very well adapted for in cellulo detection by fluorescence microscopy. However, some of them lack of specificity and may lead to false interpretations for the presence of ROS [3, 4]. For example, the use of dichlorofluorescein (DCF) as a fluorescent peroxide probe is complex and crucially lacks of selectivity in vivo [5]. Hydroethidine (HE, MitoSOX™), a cell permeable compound which is specifically oxidized by $O_2^{\bullet -}$ into the fluorescent 2-hydroxyethidium (2-OH-E+) compound, has been described as an interesting $O_2^{\bullet -}$ probe [6]. However, in the cell, HE also generates other fluorescent compounds that are not related to the reaction with $O_2^{\bullet -}$. Consequently, direct quantification of 2-OH-E+ production in whole cells by fluorescence microscopy is not possible, and chromatography analysis

(continued)

of extracts obtained after cell lysis must be carried out in order to quantify 2-OH-E+ production [7].

In vitro, ROS detection is much less challenging since control experiments can be usually carried out in order to make corrections for nonspecific reactions. Probes like cytochrome c and tetrazolium salts XTT have been widely used to detect $O_2^{\bullet -}$ in vitro [3, 4]. Spin trap nitrones, like DMPO or DEPMPO, associated with electron paramagnetic resonance spectroscopy (EPR) are well adapted to detect the unstable radical species $O_2^{\bullet -}$ and $HO^{\bullet}$ [3]. However, in vivo, their use is more difficult since the nitrone adducts resulting from the reaction of the spin trap with $O_2^{\bullet -}$ or $HO^{\bullet}$ are unstable and can be easily reduced by cellular reductants [3].

1. Winterbourn, C.: Reconciling the chemistry and biology of reactive oxygen species. Nat. Chem. Biol. **5**, 278–286 (2008)
2. Nel, A., Xia, T., Madler, L., Li, N.: Toxic Potential of Materials at the Nanolevel. Science **311**, 622–627 (2006)
3. Bartosz, G.: Use of spectroscopic probes for detection of reactive oxygen species. Clin. Chim. Acta **368**, 53–76 (2006)
4. Wardman, P.: Fluorescent and luminescent probes for measurement of oxidative and nitrosative species in cells and tissues: Progress, pitfalls, and prospects. Free Rad. Biol. Med. **43**, 995–1022 (2007)
5. Zielonka, J., Kalyanaraman, B.: ROS-generating mitochondrial DNA mutations can regulate tumor cell metastasis" – a critical commentary. Free Rad. Biol. Med. **45**, 1217–1219 (2008)
6. Robinson, K.M., Janes, M.S., Pehar, M., Monette, J.S., Ross, M.F., Hagen, T.M., Murphy, M.P., Beckman, J.S.: Selective fluorescent imaging of superoxide

(continued)

in vivo using ethidium-based probes. Proc. Natl. Acad. Sci. **103**, 15038–15043 (2006)
7. Zielonka, J., Vasquez-Vivar, J., Kalyanaraman, B.: Detection of 2-hydroxyethidium in cellular systems: a unique marker product of superoxide and hydroethidine. Nat. Protocol. **3**, 8–21 (2008)

Effects of NPs Shape and Aspect Ratio

The toxicity of particles as function of their shape does not concern only nanoparticles. Indeed, many studies concerning asbestos toxicity have demonstrated that the toxicity can be partly related to the aspect ratio. Then comparison between asbestos and carbon nanotubes legitimated the worry concerning probable toxic effects of NPs. Poland and coworkers [23] have shown that "*long multiwalled carbon nanotubes result in asbestos-like, length dependent, pathogenic behavior*." However, the mechanisms at the origin of the toxicity of long (>15–20 μm) nanotubes still remain unclear. In the case of nano-silver, it is possible to produce spherical, truncated triangular, rods nanoparticles. Pal et al. [24] have shown that that the biocidal effects depend on silver NPs shape with the higher effect for truncated triangular NPs. It has been emphasized that the {111} lattice plane exhibit the highest reactivity to sulfur- or phosphorus-containing soft bases, such as R-S-R, R-SH, RS, or PR3. Then sulfur-containing proteins in the membrane interacting with Ag NPs (with high affinity {111} lattice plane) can lead to membrane disruption or strong damage.

Effects of NPs Solubility

As discussed in the "Physicochemical properties of Nanoparticles" part, NPs solubility may substantially vary from thermodynamic prediction deduced from larger particles. The small size and consequently large surface area is of high concern to address possible NPs dissolution. It has been mentioned that in the case of toxic metals the dissolution will increase the concentration of free metal ions in contact with living organism. This is particularly true for silver nanoparticles for which the dissolution is associated with a change of the redox state (oxidative dissolution). It has been reported that upon contact with a bacterial cell, biocide Ag^{+} is released from the surface of NPs, which then interacts with bacterial cell membranes and inhibits respiratory enzymes (e.g., [25]). In the case of ZnO NPs, dissolution and release of Zn^{2+} play an important role in the biological toxicity (e.g., [26]). Dissolution is also discussed in the case such as Quantum Dots, even if recent developments prevent or strongly decrease QDots solubility via specific surface coating.

Effects of NP Adhesion to Cells

NPs can be chemically unstable (dissolution, oxidation, reduction, etc.), inducing biological effects. However, the toxicity level may be enhanced or alleviated by additional parameters. For instance, it has been shown that nano-CeO_2 can be reduced when both E. coli and Synechocystis are exposed [27]. But the reduction of Ce^{4+} to Ce^{3+} induced a much more intense effect on E. coli. Indeed, at the highest nano-CeO_2 concentration, the percentage of survival cells was still 35 % for Synechocystis while no cell was alive in the case of E. coli. This difference may be explained by the production of extracellular polymeric substances (EPS) preventing direct contact between cell membranes and nano-CeO_2 due to strong adsorption and adhesion of NPs. Thus, in the event of redox phenomena, the added distance between NPs and the membranes and the EPS shield itself may provide at least partial protection against the toxicity.

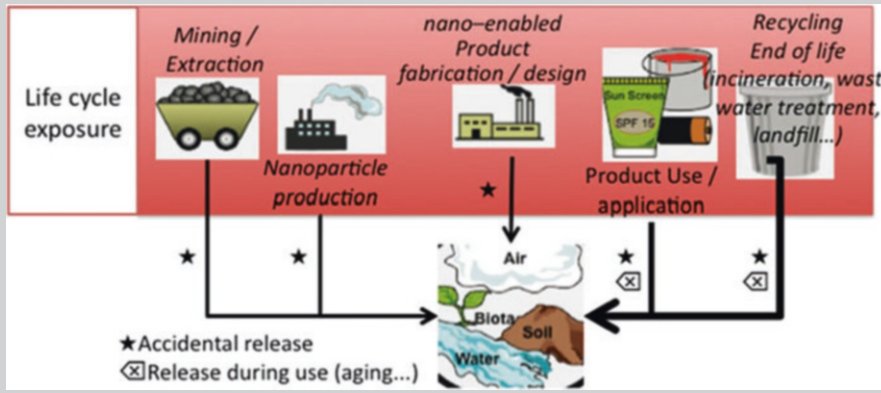
Effects of Chemical Composition, Crystal Structure. . .

Besides NPs properties discussed above, other parameters can induce biological effects. The chemical composition of NPs is of utmost importance. Many metals and metalloids are known to

be toxic like for instance Cd or Se. Their use by nanotechnology to elaborate nanoproducts or nanomaterials must be done with care. Any dissolution of NPs incorporating Cd (for instance) will lead to toxicity, and it is worth developing any specific design to enhance their chemical stability and prevent at best their direct contact with external media. Most importantly, one has to define whether or not other less toxic element can be used to reach the same level of performance.

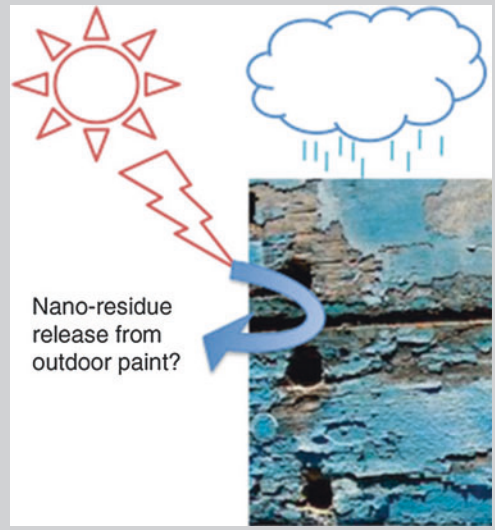
Box 2: Exposure During Life Cycle of Nanoparticles and Nanomaterials

Most of the time, nanoparticles are surface modified prior to their incorporation in final products. Therefore, it is very likely that nanoparticles will not be directly released in the environment.



The degradation of products due to external conditions (sun, rain, etc.) may generate the release of by-products containing nanoparticles.

For instance:



From Nanoparticles to Nano-Residues of the Nanomaterials Life Cycle

Nanoparticles and nanomaterials may be released from point sources like factories or landfills that may be more related to accidental events or technical defects during the processes. But most of the release may be due to indirect exposure through the life cycle of the nanomaterials. Indeed, for most applications, nanoparticles are surface modified and generally are embedded in the final product. Therefore, they do not come into direct contact with consumers or the environment. Their surface modification can strongly affect their behavior and especially their toxicity. NPs in commercial sunscreens can illustrate the difference between bare nanoparticles and surface modified nanoproducts. Indeed, the nanoparticles of TiO_2 are coated with AlOOH and polydimethylsiloxane for their use as sunscreens. The release of nano-residues from commercial NPs-based sunscreens is no longer a hypothesis, but has become a fact [28–30]. The major effect of the remaining Al-based coating is a protection against the generation of superoxide ions from the photoactive/phototoxic TiO_2 core. This surface modification of the nano-TiO_2 results in a substantially lower toxic potential of the residue as long as the protective layer remains stable (See Box 2).

Conclusion

This section discussed the relationships between the physicochemical properties of nanoparticles and their biological effects. As described above, nanoparticles differ from each other by many aspects and not only size and shape. Aggregation state, chemical stability, solubility, etc. can change from one to another NP leading to difficult comparison when toxicity occurs. Then, even if many issues remain unsolved, it clearly appears that the toxicity strongly depends on an accurate physicochemical characterization of nanoparticles and nanomaterials. However, while characterizing several properties like chemical composition, size, shape, crystal structure, surface area, surface chemistry, charge, and solubility of the initial NPs is now well accepted (but not always performed!), information is still lacking as, more than initial product characterization, it is worth determining the evolution of important parameters during the interaction between NPs and biological organisms. However, such accurate determination of nanoparticles in biological media remains highly challenging as discussed in this section.

Acknowledgments The authors would like to thank the Centre National de la Recherche Scientifique (CNRS) and the Commissariat à l'Energie atomique (CEA) for funding the International Consortium for the Environmental Implications of NanoTechnology (GDRi iCEINT).

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles
- ▶ In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
- ▶ Nanoparticle Cytotoxicity
- ▶ Selected Synchrotron Radiation Techniques

References

1. Auffan, M., Rose, J., Bottero, J.Y., Lowry, G.V., Jolivet, J.P., Wiesner, M.R.: Towards a definition of inorganic nanoparticles from an environmental, health and safety perspective. Nat. Nanotechnol. **4**, 634–641 (2009)
2. Hutchings, G.J., Brust, M., Schmidbaur, H.: Gold – an introductory perspective. Chem. Soc. Rev. **37**, 1759–1765 (2008)
3. Turner, M., Golovko, V.B., Vaughan, O.P.H., Abdulkin, P., Berenguer-Murcia, A., Tikhov, M.S., Johnson, B.F.G., Lambert, R.M.: Selective oxidation with dioxygen by gold nanoparticle catalysts derived from 55-atom clusters. Nature **454**, 981–983 (2008)
4. Hassellöv, M., Readman, J.W., Ranville, J.F., Tiede, K.: Nanoparticle analysis and characterization methodologies in environmental risk assessment of engineered nanoparticles. Ecotoxicology **17**, 344–361 (2008)
5. Kaulich, B., Thibault, P., Gianoncelli, A., Kiskinova, M.: Transmission and emission x-ray microscopy: operation modes, contrast mechanisms and

applications (topical review). J. Phys. Condens. Matter **23**(8), 083002 (2011)
6. Nelson, J., Huang, X., Steinbrener, J., Shapiro, D., Kirz, J., Marchesini, S., Neiman, A.M., Turner, J.J., Jacobsen, C.: High-resolution x-ray diffraction microscopy of specifically labeled yeast cells. Proc. Natl. Acad. Sci. **107**, 7235–7239 (2010)
7. Zhang, H., Chen, B., Banfield, J.F.: Particle size and pH effects on nanoparticle dissolution. J. Phys. Chem. C **114**, 14876–14884 (2010)
8. Levard, C., Rose, J., Thill, A., Masion, A., Doelsch, E., Maillet, P., Spalla, O., Olivi, L., Cognigni, A., Ziarelli, F., Bottero, J.Y.: Formation and growth mechanisms of imogolite-like aluminogermanate nanotubes. Chem. Mater. **22**, 2466–2473 (2010)
9. Petkov, V., Cozzoli, D., Buonsanti, R., Cingolani, R., Ren, Y.: Size, shape, and internal atomic ordering of nanocrystals by atomic pair distribution functions: a comparative study of γ-Fe_2O_3 nanosized spheres and tetrapods. J. Am. Chem. Soc. **131**, 14264–14266 (2009)
10. Modrow, H.: Tuning nanoparticle properties – the X-ray absorption spectroscopic point of view. Appl. Spectrosc. Rev. **39**, 183–290 (2004)
11. Auffan, M., Rose, J., Orsiere, T., De Meo, M., Thill, A., Zeyons, O., Proux, O., Masion, A., Chaurand, P., Spalla, O., Botta, A., Wiesner, M.R., Bottero, J.-Y.: CeO_2 nanoparticles induce DNA damage towards human dermal fibroblasts in vitro. Nanotoxicology **3**, 161–171 (2009)
12. Thill, A., Zeyons, O., Spalla, O., Chauvat, F., Rose, J., Auffan, M., Flank, A.M.: Cytotoxicity of CeO2 nanoparticles for *Escherichia coli*. Physico-chemical insight of the cytotoxicity mechanism. Environ. Sci. Technol. **40**, 6151–6156 (2006)
13. Swarbrick, J.C., Weng, T.-C., Schulte, K., Khlobystov, A.N., Glatzel, P.: Electronic structure changes in cobalt phthalocyanine due to nanotube encapsulation probed using resonant inelastic X-ray scattering. Phys. Chem. Chem. Phys. **12**, 9693–9699 (2010)
14. Klaine, S.J., Alvarez, P.J.J., Batley, G.E., Fernandes, T.F., Handy, R.H., Lyon, D.Y., Mahendra, S., Mclaughlin, M.J., Lead, J.R.: Nanomaterials in the environment: behavior fate, bioavailability, and effects. Environ. Toxicol. Chem. **27**, 1825–1851 (2008)
15. Handy, R.H., von der Kammer, F., Lead, J.R., Hassellöv, M., Owen, R., Crane, M.: The ecotoxicology and chemistry of manufactured nanoparticles. Ecotoxicology **17**, 287–314 (2008)
16. Napierska, D., Thomassen, L.C.J., Lison, D., Martens, J.A., Hoet, P.H.: The nanosilica hazard: another variable entity. Part. Fibre Toxicol. **7**(1), 39 (2010)
17. Lovern, S.B., Klaper, R.: Daphnia magna mortality when exposed to titanium dioxide and fullerene [C60] nanoparticles. Environ. Toxicol. Chem. **25**, 1132–1137 (2006)
18. Hund-Rinke, K., Simon, M.: Ecotoxic effect of photocatalytic active nanoparticles (TiO_2) on algae and daphnids. Environ. Sci. Pollut. Res. **13**, 225–232 (2006)
19. Auffan, M., Rose, J., Proux, O., Borschneck, D., Masion, A., Chaurand, P., Hazemann, J.L., Chaneac, C., Jolivet, J.P., Wiesner, M.R., Van Geen, A., Bottero, J.Y.: Enhanced adsorption of arsenic onto maghemites nanoparticles: As(III) as a probe of the surface structure and heterogeneity. Langmuir **24**, 3215–3222 (2008)
20. Auffan, M., Decome, L., Rose, J., Orsiere, T., De Meo, M., Briois, V., Chaneac, C., Olivi, L., Berge-Lefranc, J.L., Botta, A., Wiesner, M.R., Bottero, J.Y.: In vitro interactions between DMSA-coated maghemite nanoparticles and human fibroblasts: a physicochemical and cyto-genotoxical study. Environ. Sci. Technol. **40**, 4367–4373 (2006)
21. Auffan, M., Achouak, W., Rose, J., Roncato, M.A., Chaneac, C., Waite, D.T., Masion, A., Woicik, J.C., Wiesner, M.R., Bottero, J.Y.: Relation between the redox state of iron-based nanoparticles and their cytotoxicity toward *Escherichia coli*. Environ. Sci. Technol. **42**, 6730–6735 (2008)
22. Pourbaix, M.: Atlas of Electrochemical Equilibria in Aqueous Solutions, 2nd edn. National Association of Corrosion Engineers, Houston (1974)
23. Poland, C.A., Duffin, R., Kinloch, I., Maynard, A., Wallace, W.A.H., Seaton, A., Stone, V., Brown, S., MacNee, W., Donaldson, K.: Carbon nanotubes introduced into the abdominal cavity of mice show asbestos-like pathogenicity in a pilot study. Nat. Nanotechnol. **3**, 423–428 (2008)
24. Pal, S., Tak, Y.K., Song, J.M.: Does the antibacterial activity of silver nanoparticles depend on the shape of the nanoparticle? A study of the gram-negative bacterium *Escherichia coli*. Appl. Environ. Microbiol. **73**, 1712–1720 (2007)
25. Choi, O., Hu, Z.: Size dependent and reactive oxygen species related nanosilver toxicity to nitrifying bacteria. Environ. Sci. Technol. **42**, 4583–4588 (2008)
26. Xia, T., Kovochich, M., Liong, M., Madler, L., Gilbert, B., Shi, H.B., Yeh, J.I., Zink, J.I., Nel, A.E.: Comparison of the mechanism of toxicity of zinc oxide and cerium oxide nanoparticles based on dissolution and oxidative stress properties. ACS Nano **2**, 2121–2134 (2008)
27. Zeyons, O., Thill, A., Chauvat, F., Menguy, N., Cassier-Chauvat, C., Orear, C., Daraspe, J., Auffan, M., Rose, J., Spalla, O.: Direct and indirect CeO_2 nanoparticles toxicity for *E. coli* and *Synechocystis*. Nanotoxicology **3**, 284–295 (2009)
28. Auffan, M., Pedeutour, M., Rose, J., Masion, A., Ziarelli, F., Borschneck, D., Chaneac, C., Botta, C., Chaurand, P., Labille, J., Bottero, J.Y.: Structural degradation at the surface of a TiO_2-based nanomaterial used in cosmetics. Environ. Sci. Technol. **44**, 2689–2694 (2010)
29. Labille, J., Feng, J.H., Botta, C., Borschneck, D., Sammut, M., Cabie, M., Auffan, M., Rose, J., Bottero, J.Y.: Aging of TiO2 nanocomposites used in

sunscreen. Dispersion and fate of the degradation products in aqueous environment. Environ. Pollut. **158**, 3482–3489 (2010)
30. Botta, C., Labille, J., Auffan, M., Borschneck, D., Miche, H., Cabié, M., Masion, A., Rose, J., Bottero, J.-Y.: TiO_2-based nanoparticles released in water from commercialized sunscreens in a life-cycle perspective: structures and quantities. Environ. Pollut. **159**(6), 1543–1550 (2011)

Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing

Shen Xu[1], Lijun Zhang[2], Yanan Yue[3] and Xinwei Wang[1]
[1]Department of Mechanical Engineering, Iowa State University, Ames, IA, USA
[2]Department of Mechanical Engineering, Shanghai Ocean University, Shanghai, People's Republic of China
[3]School of Power and Mechanical Engineering, Wuhan University, Wuhan, Hubei, People's Republic of China

Synonyms

Laser-assisted tip-based heating and fabrication; Laser-assisted tip–substrate interaction; Near-field focusing and heating

Definition of the Subject

Laser near-field nanomanufacturing is manufacturing nanoscaled materials and complex structures at the nanoscale by means of phase change/explosion with the near-field focused laser beam, whose spot size is usually smaller than its wavelength.

Introduction

Nanofabrication and nanomanufacturing involve manufacturing of materials, structures, devices, and systems down to the nanoscale. There are two basic categories of fabrication, top-down and bottom-up or self-assembly processes, containing a growing number of chemical and physical techniques, such as chemical methods, chemical vapor deposition, molecular beam epitaxy, atomic layer epitaxy, lithography, self-assembly, and countless newly developed techniques. Among those methods, laser-assisted nanomanufacturing plays an important role due to its highly coherent feature. To achieve a resolution below the diffraction limit, various near-field mechanisms of laser-assisted technologies have been developed in past decades, such as the laser-assisted nanopatterning and nanoimprinting lithography, near-field scanning optical microscopy (NSOM), superlens, nanoscale aperture, and scanning probe microscopes (SPM).

The laser-assisted tip-based fabrication is a large branch of laser-assisted manufacturing techniques by using SPM tips. The SPM tip is an ideal tool for apertureless NSOM due to its tiny but sharp geometry. When a laser irradiates the SPM tip, it can bend the propagation direction of the laser and focus and enhance the optical field in an extremely small area in proximity to the tip apex. The enhancement factor could be more than 100 by using a scanning tunneling microscope (STM) tip. This strongly enhanced optical field is then utilized to develop a new fabrication method by using laser-irradiated STM tip [1]. The absorption of this enhanced optical field could generate heat in the sample surface and induce phase change at the nanoscale, finally modifying the surface. Similar to STM, atomic force microscopes (AFM) can focus the laser beam to a nanoscale spot as well to achieve surface manufacturing at the nanoscale.

This entry focuses on experimental and computational works about laser-assisted SPM nanostructuring. Section Near-Field Focusing and Heating by the Micro/Nanoscale AFM Tip: Effect on the Tip introduces experiment and simulation of the temperature and optical field distribution in the AFM tip when it is heated by near-field focusing laser. Section Near-Field Focusing and Heating by the Micro/Nanoscale AFM tip: Effect on the Substrate discusses the study about

the temperature measurement and optical field distribution simulation in a silicon substrate next to the apex of the AFM tip. Sections Near-Field Focusing by Micron/Submicron Particles: Experimental Study and Near-Field Thermal, Optical, and Mechanical Phenomena: Modeling are for experiments and simulation study, respectively, for the near-field thermal, optical, and mechanical phenomena of a silicon plate under the micro/submicron particles. Section Material Structure Evolution Under Near-Field Laser Heating mainly focuses on the simulation work about material's structure change under near-field laser heating, including structure evolution and the effect of shock waves.

Near-Field Focusing and Heating by the Micro-/Nanoscale AFM Tip: Effect on the Tip

The enhanced optical distribution in and around the apex of a tip has been widely studied. The thermal evolution and thermal distribution, however, was studied little. Chen et al. [2] has conducted high fidelity and systematic simulation of near-field thermal transport in an AFM tungsten tip under laser irradiation. The electromagnetic field and temperature distribution is simulated in a domain housing a tungsten SPM tip and silicon substrate system.

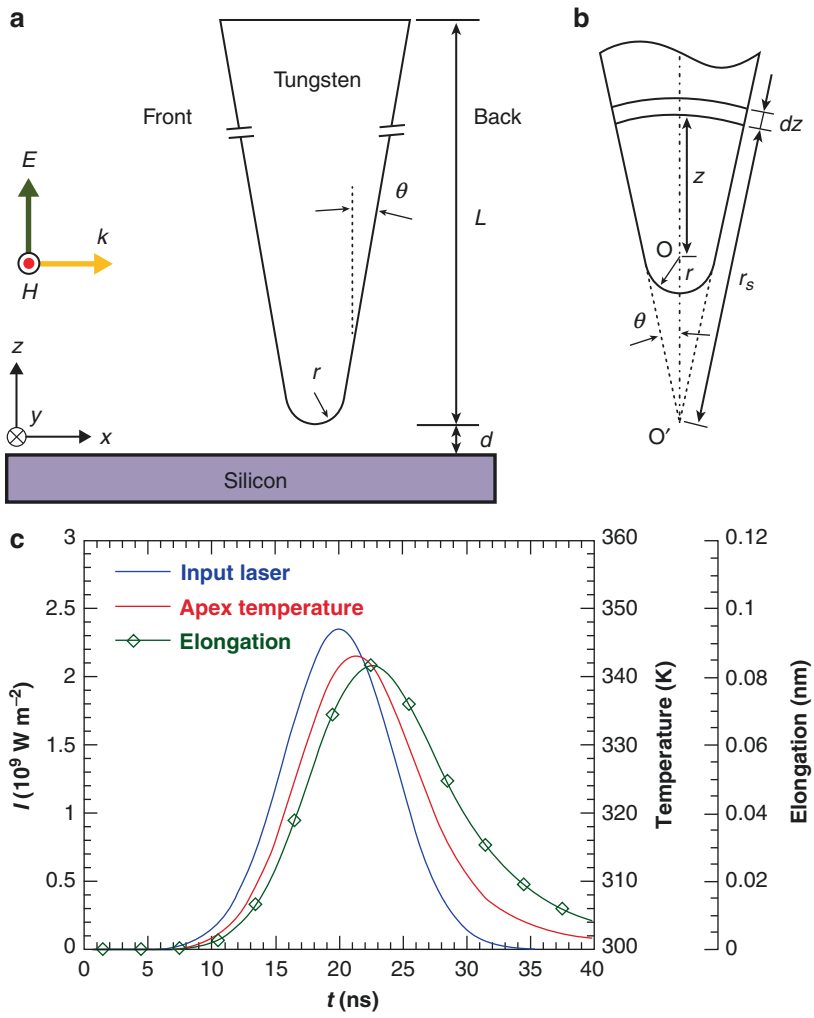
FEM Modeling for Both Optical and Thermal Physics in AFM Tip Under Irradiation

The model (Fig. 1a) employs a conical tungsten tip ended with a hemisphere on the silicon substrate. The modeling is performed by using a high-frequency structure simulator (HFSS V12.1 ANSYS, Inc.), a full-wave high-frequency 3-D finite element modeler of Maxwell's equations. Maxwell's equations are solved across the vacuum domain. Tetrahedral elements with a length shorter than one-fourth of the incident wavelength are set in the whole domain. They are further refined in areas where the high field gradient may occur. Absorbing boundaries are applied for the domain. The incident plane wave is along the x direction of the tip.

The geometry of the tip is shown in Fig. 1b. The length L of the tip model is carefully chosen as 600 nm which is a good approximation of 15 μm long commercial tips. The tip apex radius $r = 30$ nm and the half taper angle $\theta = 10°$ are used in the simulation. The distance d between the tip and substrate is set as 5 nm. The Poynting vector simulation illustrates the incident laser is bent around the tip and concentrated under the apex, and the extremely intensive electric field appears there. The largest enhancement factor around 15 under the apex in the model is given by the electric field simulation. In contrast, the electric field is much less enhanced inside the tip, since only a small fraction of laser penetrates into the tip. The electric field is not symmetrical inside the tip. Its gradient decreases along the laser's incident direction.

Three factors significantly affect the electric field enhancement: the laser polarization direction, the geometry of the tip, and the distance between the tip apex and the substrate. The intensity enhancement of the electric field depends on $\cos^2\varphi$, where φ is the polarization angle to the tip axis. Only the component along the tip axis of the incident laser makes a contribution to the electric field enhancement under the tip apex. For the geometry factor, the optimum r for electric field enhancement is about 9 nm when the half taper angle θ is 10°. For a tip with a fixed radius of 30 nm, the field enhancement first increases along with θ increasing when θ is less than 25°, and the field enhancement keeps pretty much constant in the range of 25–30° and then decreases as θ goes beyond 35°. Moreover, the peak field enhancement decreases exponentially as the tip–substrate distance increases, and a shift of field peak position is observed being away from the tip axis. This shift vanishes when the tip touches the substrate.

A single pulse is applied on the tip to study the temperature evolution and distribution in the tip at room temperature (300 K). The simulation of the temperature distribution in the tip is conducted with commercial computational software FLUENT (V12.0.1 ANSYS, Inc.). The intensity of the electric field is calculated according to the Poynting vector. The maximum temperature in the tip apex is delayed (Fig. 1c) due to the thermal conduction in the tip when compared with the

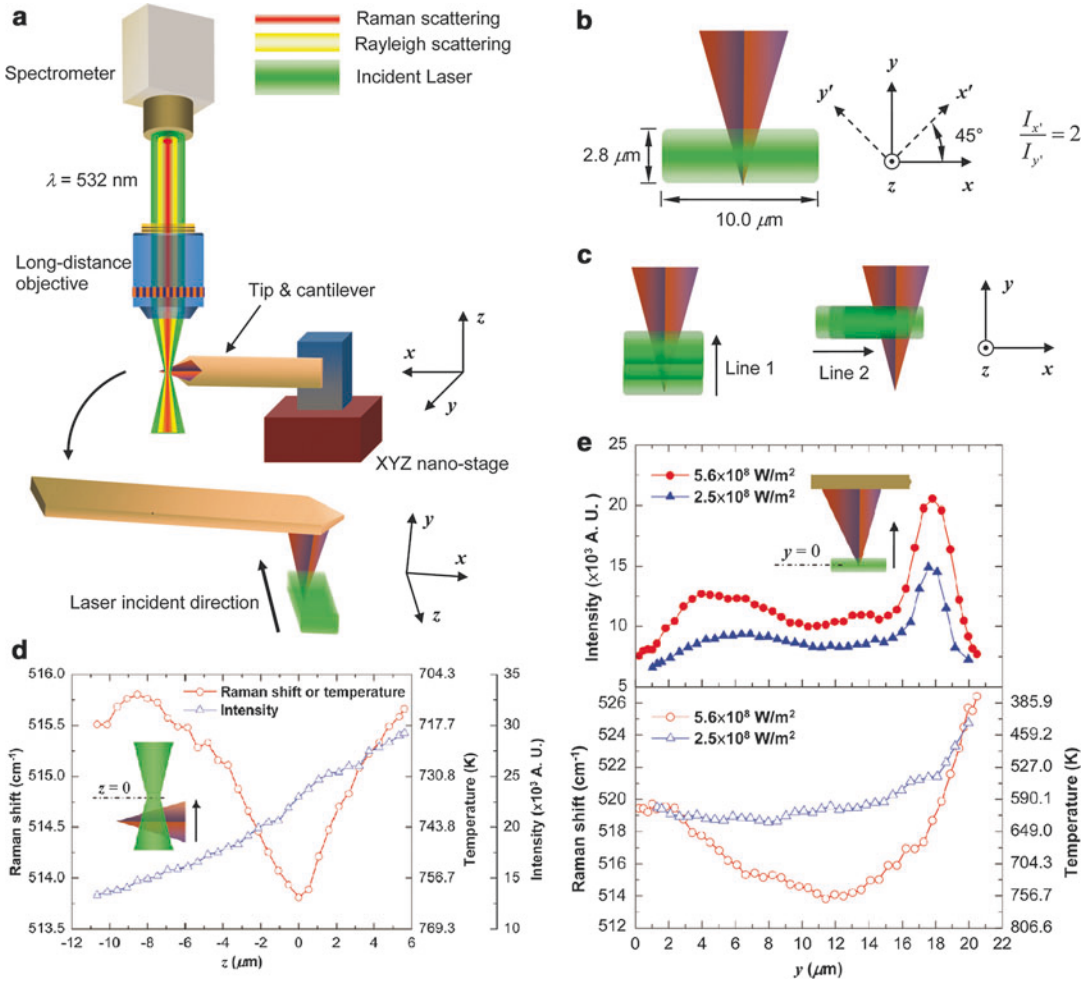


Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 1 (**a**) Schematic of the tip–substrate system, (**b**) geometric structure of the tip model, and (**c**) the temperature evolution against the time (Reprinted from ref. [2], © IOP Publishing. Reproduced by permission of IOP Publishing. All rights reserved)

maximum energy in the laser pulse. The resulting temperature distribution is affected by two factors: the heat source due to laser absorption and the geometric shape of the tip. As the polarization angle or apex radius increases, the peak apex temperature decreases. The half taper angle strongly affects the thermal transport in the tip. The peak apex temperature declines with an increased taper angle, even though the laser intensity inside the tip is increasing in the meanwhile.

Experimental Details of Temperature Measurement in the AFM Tip

The corresponding experimental work is conducted with a freestanding AFM silicon tip [3] under the laser irradiation with a micrometer spatial resolution. A single crystal silicon (antimony doped) tip (CSG10, K-Tek nanotechnology) is mounted on a three-dimensional piezo-actuated nanostage with a positioning resolution of 20 nm. The stage with the sample is placed under the laser spot at the focal level of a confocal Raman microscope system (Voyage, B&W Tek, Inc.). The tip axis is horizontal and along the y direction of the nanostage. The cantilever is along the x direction, and the laser is incident vertically along the z direction. The whole system is placed in air. The laser beam is focused through a 20× microscope objective (NA = 0.40, WD = 12.0 mm) with a focal spot size of $2.8 \times 10.0\ \mu m^2$. The laser beam is partially polarized along the x' and y' directions as shown in



Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 2 (**a**) Schematic of experimental setup (no to scale). (**b**) Configuration of the laser beam at the focal spot when the laser illuminates the side surface of tip. (**c**) Laser spot moving directions with respect to the tip during experiment. (**d**) Raman shift, temperature, and Raman intensity when the tip moves upward along the z direction in the vicinity of focal spot. (e) Raman intensity and Raman shift along line 1 [as shown in Fig. 2c, in the y direction] under energy fluxes of 5.6×10^8 and 2.5×10^8 W/m^2, respectively (Reprinted from ref. [3], © IOP Publishing. Reproduced by permission of IOP Publishing. All rights reserved)

Fig. 2b. The nanostage is aligned to make sure that the laser is coming in along the z direction. The tip axis is perpendicular to the longer side of the laser spot. In this case, the tip axis has an angle of 45° with respect to the direction in which the polarized laser has the strongest intensity. The laser with a wavelength of 532 nm is used as both Raman probe and heat source. To investigate the heating effect caused by laser, the tip is illuminated with different laser powers of 4, 7, 12.7, and 15.9 mW.

The Raman shift method of Raman thermometry is used to determine the temperature rise of the Si AFM tip under laser irradiation. Raman shift will become smaller when the temperature of the probing area goes up. The exact temperature rise is then calculated based on the theoretical calculation introduced by McCarthy et al. [4]. To precisely extract the Raman shift, the silicon Raman peak in all measured spectra is fitted with the Gaussian function. In order to exclude the

focusing effect on the experimental result, the diffraction image of the tip under laser illumination is examined to precisely monitor the focusing level.

The temperature rise in the tip is first studied as the tip moves along the z direction, and the whole travel covers the focal spot (Fig. 2d). When the tip moves upward closer to the focal spot, the Raman shift decreases and the temperature increases; analogously, when the tip passes the focal spot and keeps moving upward, the Raman shift increases and the temperature goes down. Nevertheless, the Raman intensity increases during the whole process. When the laser spot is moved along the tip axis at the focal level from tip apex to cantilever, the temperature in the tip increases first and then decreases (Fig. 2e). The corresponding temperature increases from 603 to 754 K and then declines to 728 K. This is a trade-off between the increasing of laser absorbing area and the increasing heat conduction to the base of the tip even the cantilever. The incident laser intensity also strongly affects the heating condition.

To investigate the mechanism of the near-field heating in the experiment, the electric field distribution around the tip and inside the tip is calculated using the similar method as described in section FEM Modeling for Both Optical and Thermal Physics in AFM tip Under Irradiation. The highest field enhancement is 5.19 and 2.86 for under the tip apex and inside the tip, respectively. A one-dimensional heat transfer model is used to calculate the temperature distribution along the tip axis. Details of the calculation can be found in the reference [3]. The highest temperature of 448 K is obtained at the hemisphere apex from the calculation, while the measured temperature is approximately 600 K. The difference between them could be induced by further thermal conductivity reduction of the Si tip by its dopants and the underestimated laser scattering/absorption in the tip.

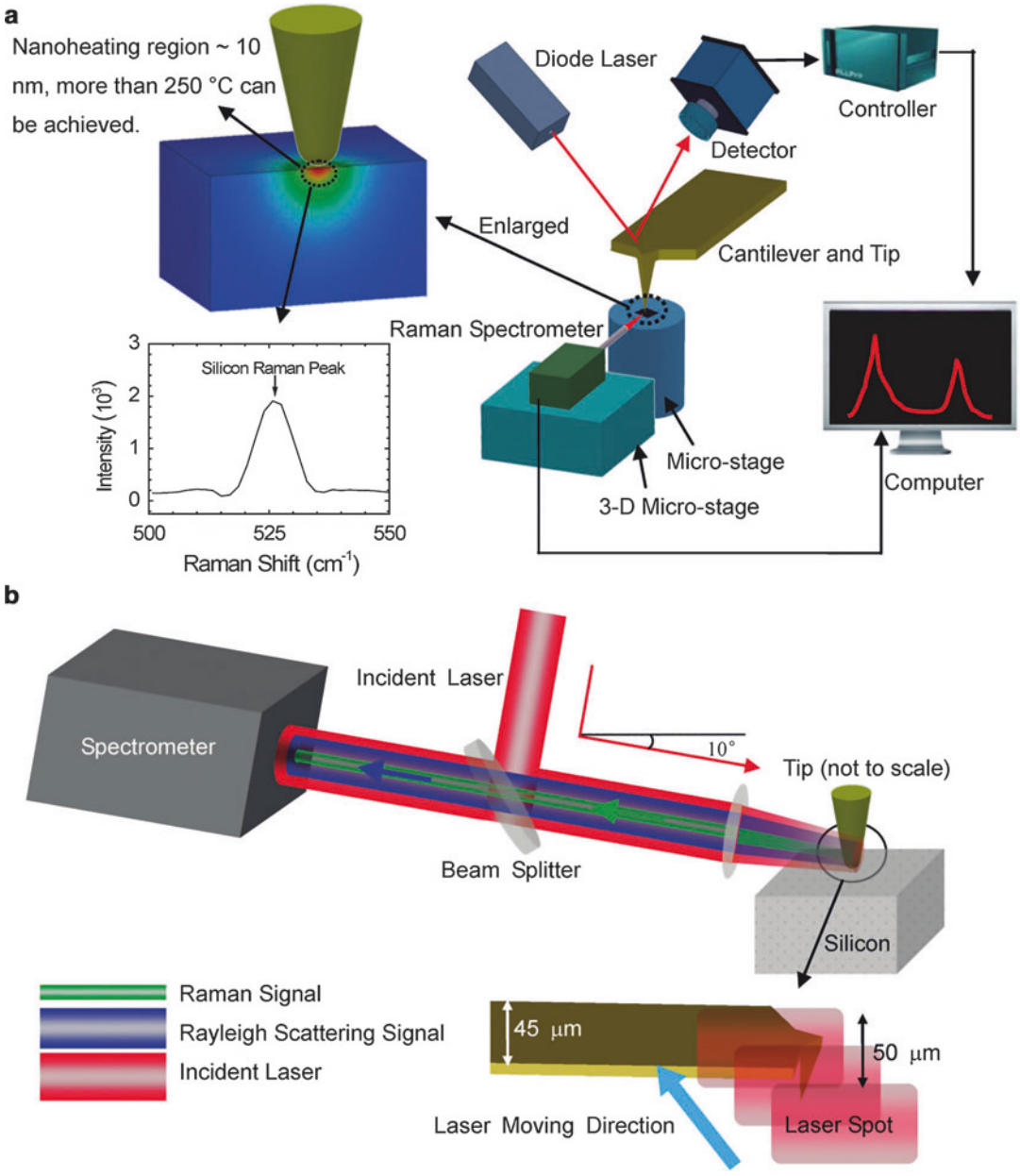
Near-Field Focusing and Heating by the Micro-/Nanoscale AFM Tip: Effect on the Substrate

According to Chen's study, an AFM tip can induce the near-field enhancement on electromagnetic field next to the tip apex when an external laser irradiates the tip. Besides, the local heating in the substrate beneath the tip can also induce a considerable temperature rise. The first experimental work for the direct temperature measurement was done by Yue et al. [5]. The Raman shift method is applied to measure the temperature variation in a sub-10 nm area of the silicon substrate beneath the apex of the AFM tip.

Experimental Details

Before the temperature measurement, thermal response calibration is conducted between Raman shift of Si peak and temperature from room temperature to 230 °C. A silicon sample is heated using a power-adjustable electrical heater and temperature is sensed by a T-type thermocouple. The Raman signals are excited by using the 785 nm probing laser from the Raman spectrometer (B&W Tek). The Raman spectra are then fitted with the Gaussian function to obtain precise Raman shift of the Si peak. The obtained linear relationship between Raman shift and temperature is used to determine the temperature in the substrate in the experiment.

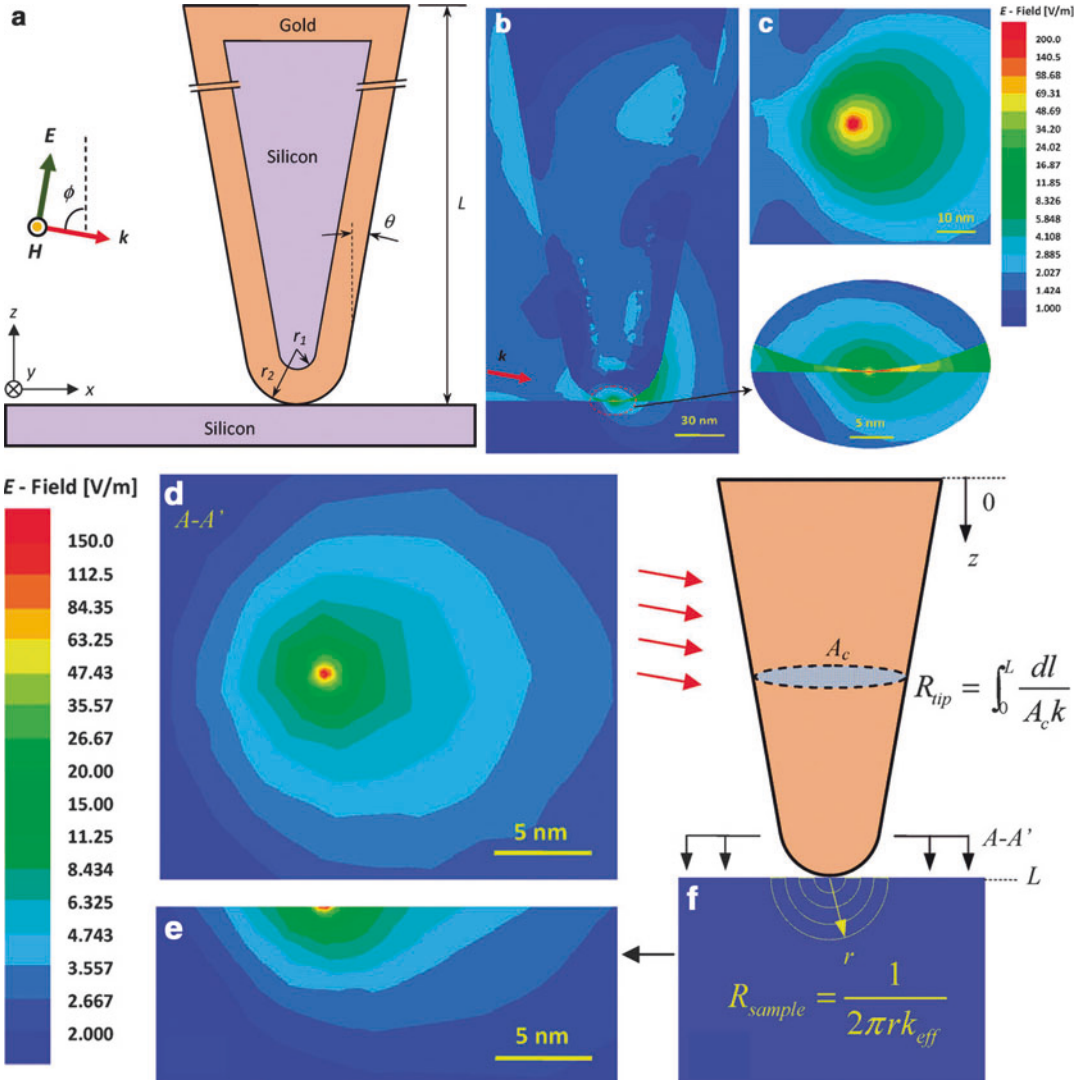
The experimental setup is shown in Fig. 3. A silicon substrate is mounted on a 3-D nanostage in the AFM system, and the 20 nm gold-coated AFM silicon tip is fixed above the substrate. A Raman spectrometer (B&W Tek) is mounted on another 3-D microstage and placed in front of the fixed AFM tip–substrate system. A beam splitter and a lens are placed on the optical path between Raman probe and the tip. The incident Raman laser is focused on the tip by using the lens, and the reflected incident laser is then used to monitor the irradiating position. The incident angle with respect to the tip axis is kept to be around 80° in order to reduce the Raman signal directly from the unheated silicon. Three cases with three different polarized angles, 15°, 30°, and 75° (with respect to the tip axis), are studied. In 15° and 30° cases, the temperature rise is more than 200 °C in the sub-10 nm area in the silicon substrate due to the near-field focus effect next to the AFM tip apex.



Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 3 (**a**) Schematic of experimental setup for the apertureless NSOM (not to scale) for thermal sensing. (**b**) Schematic of the optical system. The incident laser passes through a beam splitter and is focused on the tip by a convex lens. The excited Raman signal and Rayleigh scattering signal go along the same optical trace back to the spectrometer (Reprinted with permission from ref. [5], Copyright (2011) American Chemical Society)

Simulation Study of the Near-Field Optical Effect and Thermal Effect

The simulation of corresponding electric field enhancement is also conducted to well explain the mechanism behind the above experiment. The HFSS (V13 ANSYS, Inc.) is used for calculating the electric field distribution in a tip–substrate system. The simulated model is



Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 4 (**a**) Schematic of tip–substrate model in the simulation. Electric field distribution around the tip for (**b**) front view in the $y = 0$ plane and (**c**) top view of the cross section under the tip apex under vacuum. (**d**) The electric field distribution in the A-A' plane, (**e**) cross-sectional view of the distribution, and (**f**) schematic of the thermal resistance calculation of the tip and Si substrate (Reprinted with permission from reference [5], Copyright (2011) American Chemical Society)

shown in Fig. 4a: a well-defined rectangular domain contains a 20 nm gold-coated silicon tip and a 200 nm thick silicon substrate. The length L of the tip in the model is 600 nm, the half taper angle θ is 10°, and the apex radius r_1 is 10 nm. The polarized laser illuminates the tip along the x direction with an incident angle ϕ of 80° to the z direction. The simulation result (Fig. 4b–f) shows a large enhancement of the electric field only in the vicinity of the contact point between the tip and substrate. This extremely strong electric field focuses into a hemisphere with a diameter of sub-10 nm in the silicon substrate.

P

Nonlinear Optical Absorption and Ballistic Heat Conduction in the Near-Field Focused Region in the Substrate

To calculate the temperature rise in the near-field heating region, a thermal model is established as $\int_{\Delta V} I\beta\, dv = (T - T_\infty)/R_{\text{th}}$ which means an equilibrium between heat generation from the enhanced electric field and heat transfer from the heating region to the outside. ΔV is the near-field heating domain ($\Delta V \approx 2\pi r_e^3/3$). r_e is the effective radius of this heating region. It can be derived from nonlinear photon absorption induced by the high intensity of the optical field in the silicon. The electric field under the tip is first enhanced due to the near-field effect on the incident laser. This enhanced electric field then excites the tip again which contributes to an additional radiated electric field enhancement. Both enhancements of the electric field by the incident laser and the excited tip are second order; meanwhile, they are symmetrical in the same wavelength. Therefore, the overall magnitude of the intensity would be $\sim M^4$ [6]. M is the magnitude of enhanced electric field. The optical field intensity under low polarization angles (15° and 30°) is as high as 10^{14} W/m^2 within a 1 nm radius hemispherical region in silicon. The avalanche ionization process then occurs due to the short absorption depth (75 nm) in this region [6], where the electrical field intensity exceeds the threshold. Thus, the effective radius r_e of the hemisphere region is calculated as 0.8 nm in the case of 15° polarization angle.

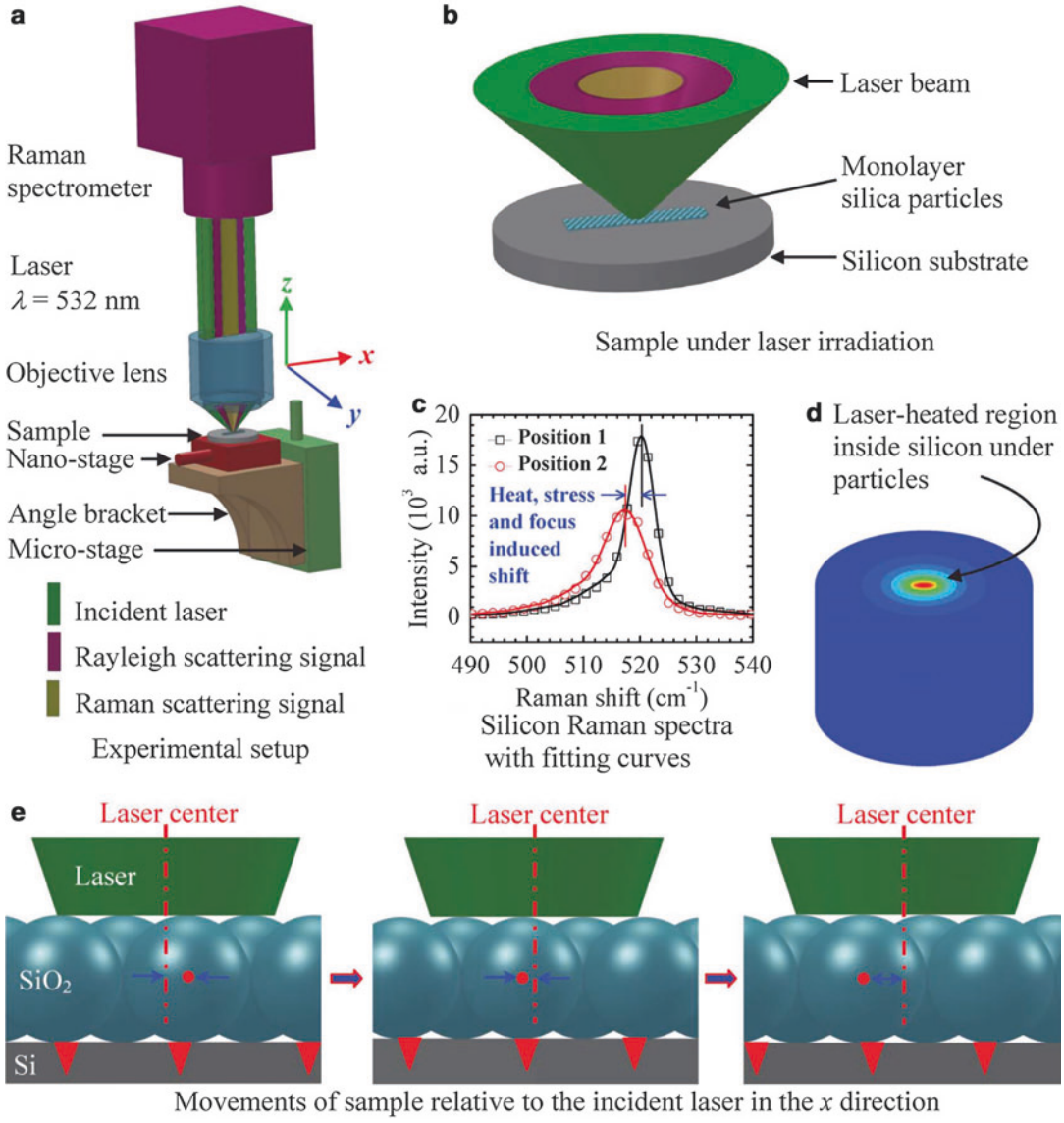
To calculate the theoretical temperature rise, the thermal resistance from the heating region to the outside is also needed. R_{th} is expressed as $1/(2\pi r_e k_{\text{eff}})$, where k_{eff} is the local effective thermal conductivity. k_{eff} is determined by considering the strong ballistic heat transfer effect due to the extremely small heating region. This heating domain (within 10 nm scale) is even smaller than the mean free path of silicon as 43 nm at room temperature. Thermal equilibrium cannot be achieved because of not enough scattering and collisions of phonons [7]. The thermal conductivity (k_{eff}) is then used for the heating region in silicon, which can be expressed as $3k/(3 + 4\tau)$, where k is the thermal conductivity of bulk silicon, and $\tau = r/\Lambda$, where Λ is the mean free path of the phonon in bulk silicon. The effective thermal conductivity is evaluated as 2.5 W/(m·K) in the calculation region. Thus, the temperature rise in this sub-10 nm domain is calculated as 235 °C and 105 °C for the 15° and 30° cases, respectively, which agree well with the experimental result. For the 75° case, a very weak optical field intensity enhancement occurs, and the temperature rise is negligible.

Near-Field Focusing by Micron/Submicron Particles: Experimental Study

Micro-/nanoparticle is another kind of useful geometry for inducing ultra-focusing and heating into sub-wavelength. This technique has been widely used in laser-assisted nanopatterning and nanolithography on a large-area substrate. Knowledge of the temperature and stress in the nanoscale beneath the particles is critical for process control and optimization. At present, the extremely small size of this heating area (~100 nm or less) remains a big challenge for nanoscale temperature, stress, and optical field probing because the heating area is very small and not immediately accessible for sensing. Tang et al. [8–10] did the pioneering and systematic experimental study of the effects of temperature, stress, and near-field focusing in the silicon plate under a single silica particle, a glass fiber, and a layer of particles, with Raman spectroscopy.

Experimental Details

The sample is prepared by spreading a single layer of glass nanoparticle mixture on a clean silicon wafer using a tilting technique. Different sizes of particles, 200, 400, 800, and 1,210 nm (diameters), are used. Details can be found in reference [8]. The experimental setup for temperature, stress, and optical field mapping is shown in Fig. 5. The sample is mounted on a combined stage under the confocal Raman system (VoyageTM, B&W Tek, and Olympus BX51): a



Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 5 Schematic of experimental setup for far-field nanoscale imaging (not to scale). (**a**) A sample is located under an objective-focused laser beam from a Raman spectrometer. The movement of sample in the x direction is controlled by a piezo-actuated nanostage. The focal level of the laser on the sample in the z direction is controlled by a motorized microstage. (**b**) The sample consists of a silicon substrate and a monolayer of silica particles. The spot size of the incident laser is about 0.5 μm in the x–y plane on a silicon substrate. (**c**) The Raman spectrum shifts to left due to the near-field laser heating, the stress, and the out-of-focus effect. (**d**) The silicon substrate is heated in a sub-wavelength region ($r \sim 200$ nm) right beneath the particles. (**e**) During the experiment, the position of the laser beam is fixed, and the sample moves along the x direction controlled by the nanostage electrically without any touch of the sample and other equipment. The step of movement is 27 nm in a range of about 4 μm (Reprinted with permission from reference [8])

P

piezo-actuated nanostage with a resolution of 20 nm (ThorLabs NFL5DP20) in the x direction (imaging direction) and a motorized translation stage (ThorLabs MT1-Z8) in the z direction. The mapping process is conducted along the x direction. The motorized stage is fixed tightly to exclude the movement in the z direction. The Raman laser is the only laser source used for both near-field excitation and probing, which is focused on the top of glass nanoparticles by a 100× objective lens (NA = 0.8). When passing through the SiO_2 particle, the laser is further focused to an even smaller spot size. Temperature and stress mapping are carried out by measuring Raman spectra along the x direction. Four energy levels of incident laser are applied in the experiment.

De-conjugation of Thermal, Stress, and Optical Effects

The obtained Raman spectra are first fitted using the Gaussian function to get the precise Raman intensity, Raman shift, and linewidth. Raman intensity, Raman shift, and linewidth of the Si peak will vary against temperature. The Raman intensity and shift decrease, while the linewidth broadens as the temperature rises. Among these three parameters of the Raman peak, the linewidth is closely relevant to the phonon lifetime. It is strongly affected by the temperature in collection area, but the stress has little effect on it. Therefore, the linewidth method is available for direct temperature analysis. Raman shift depends on both the local stress and temperature, and Raman intensity is mainly affected by the intensity of the incident laser and the temperature of the silicon substrate. The optical field intensity variation and the stress distribution can be obtained through analyzing Raman intensity and shift, respectively. Therefore, the periodic variation of the combined effect of the near-field optical heating, local stress, and the optical field in space can be separately detected by analyzing the variation of the Raman intensity, Raman shift, and linewidth along the mapping path.

The optical alignment in the system and laser energy will affect the Raman intensity, but they are independent of the sample. To quantitatively obtain the nanoscale local stress and temperature in the silicon substrate based on the intensity in this particle–substrate model, the Raman spectrum for bare silicon is referred at the same measurement condition. The effect induced by the incident laser can be eliminated when the Raman intensity ratio from silicon under particles to that from bare silicon, I_{SiO2}/I_{Si}, is applied. The difference between I_{SiO2} and I_{Si} is caused by the temperature rise under the irradiation. The silica particle-induced near-field focusing effect largely raises the temperature of the silicon substrate in the probed area. The intensity ratio I_{SiO2}/I_{Si} is then normalized by dividing the extrapolation value at 0 laser energy $[I_{SiO2}/I_{Si}]_{E=0}$. The normalized intensity ratio $[I_{SiO2}/I_{Si}]/[I_{SiO2}/I_{Si}]_{E=0}$ reduces when energy flux increases. With the calibration of I_{Si} against temperature, the temperature rise ΔT induced by 1,210 nm particles are obtained from 10.0 to 56.1 K with an uncertainty of ±7.0 K, when the laser energy percentage goes from 25 % to 100 %.

The thermal stress σ inside the silicon under particles is then determined by using both Raman shift ω and linewidth Γ. Due to the fact that Γ is less sensitive to stress than ω, the experimental linewidth and Raman shift can be expressed as $\Delta\Gamma_{exp} = \Delta\Gamma_{\Delta T} + \Delta\Gamma_{focus}$ and $\Delta\omega_{exp} = \Delta\omega_{\Delta T} + \Delta\omega_{focus} + \Delta\omega_{\sigma}$, respectively, when the temperature rise and out-of-focus effect induced by the particles are considered. With the slopes for linewidth and Raman shift against temperature increase in the calibration and the calculated ΔT from intensity, $\Delta\Gamma_{focus}$ can be easily obtained. Another calibration is conducted to obtain the relationship between $\Delta\Gamma_{focus}$ and $\Delta\omega_{focus}$ induced by the out-of-focus effect: $\Delta\omega_{focus} = -0.21–0.77\Delta\Gamma_{focus}$. Thus, $\Delta\omega_{\sigma}$ and even σ can be both obtained easily. A proportionality constant of $-3.6\ cm^{-1}$/GPa [11] is adopted for calculating the stress σ inside silicon based on the variation of Raman shift $\Delta\omega_{\sigma}$. According to the calculation, σ increases from 140 to 370 MPa with an uncertainty of 640 MPa as the energy percentage of laser goes up from 25 % to 100 %.

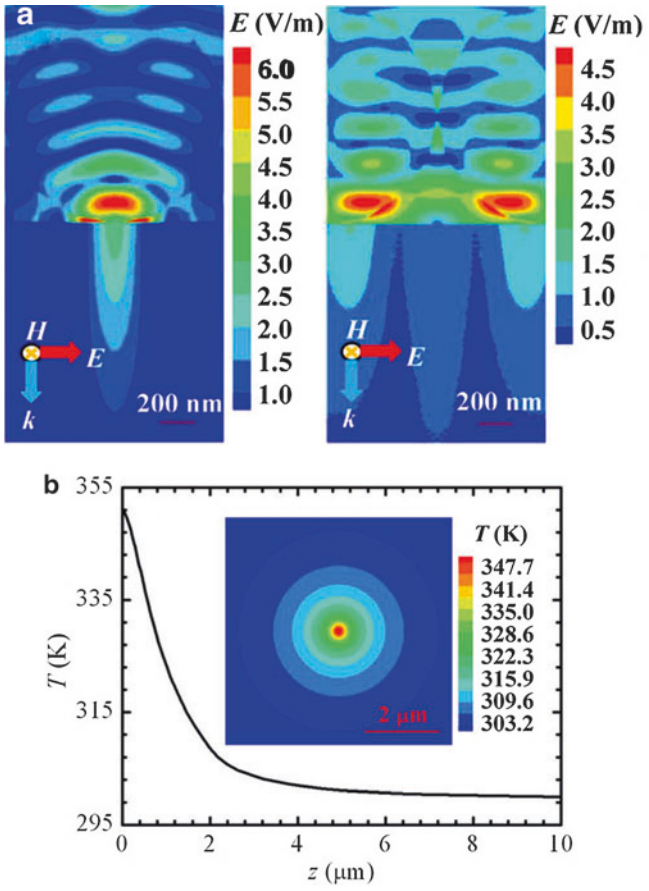
The cases of silicon under particles of smaller sizes, 800, 400, and 200 nm, are also studied. It is found that the maximum Raman intensity ratio (I_{max}/I_{min}) decreases as the particles decreases from 1,210 to 160 nm. 140 nm diameter particles

would be the minimum size which can be detected by using this confocal Raman system. The lateral resolution improves with the increase of the particle size. The best resolution of about 20 nm is achieved when the particle diameter is 1,210 nm.

Near-Field Thermal, Optical, and Mechanical Phenomena: Modeling

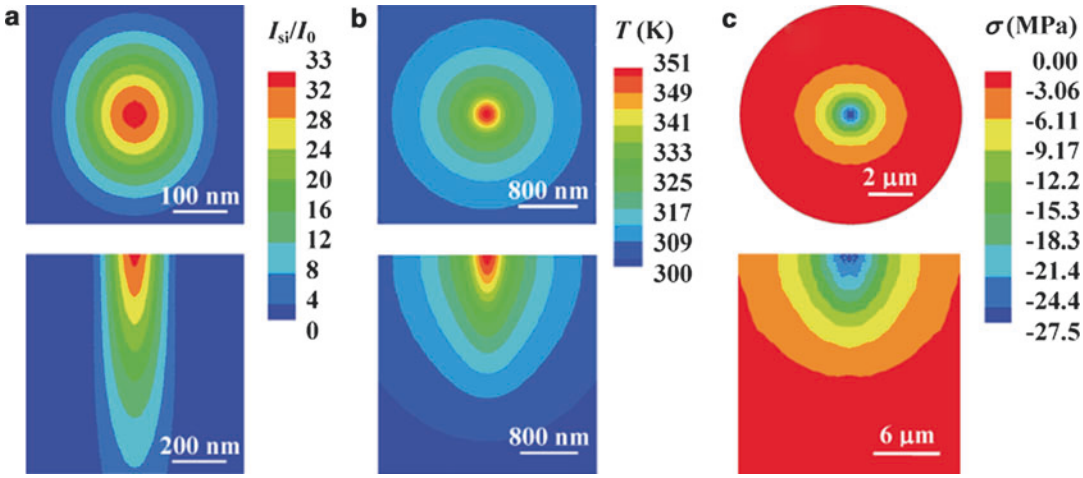
The 1,210 nm particle–substrate case is also studied with the finite element method by using HFSS (V14 ANSYS, Inc) [8–10]. A quarter of the silica particle and the substrate are used due to the consideration of the amount of mesh density and calculation in HFSS. A plane wave ($\lambda = 532$ nm) irradiates normally from the top of the particle. In the experiment, the laser spot size is smaller than the diameter of the 1,210 nm particle. To meet the experimental condition, only the parts of particles inside the laser spot area remain in the model. The rest parts are cut off to avoid receiving the plane wave. Perfect H and Perfect E symmetry boundaries are applied for symmetrical planes. Absorbing (radiation) boundaries are applied for other boundary planes in the domain. Two typical cases regarding different laser-particle positions in an imaging period are simulated, and the resulting electric field distribution is shown in Fig. 6a.

The temperature distribution inside the silicon substrate is calculated by using FLUENT (V12.0.1 ANSYS, Inc) based on the electric field distribution. Because of the symmetry of the temperature distribution inside the silicon, a quarter cylinder computational domain with a radius of 5 mm and a height of 10 mm is adopted in the simulation. Both vertical cross sections use symmetrical boundary conditions and the top surface is set as adiabatic. The temperature of peripheral and bottom surfaces of the domain and the initial



Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 6 (**a**) HFSS modeling of a plane wave passing through a 1.21 μm silica sphere (e = 2.13 + 0i) in air above a silicon substrate (e = 17.22 + 0.428i). The amplitude of electric field is equal to the enhancement factor. In the left figure, the particle center is under the laser spot center. In the right figure, the particle center is at the fringe of the laser spot area. (**b**) Temperature profile inside a silicon substrate beneath a 1210 nm silica particle under laser irradiation. The *inset* shows the temperature distribution on the top of the substrate (Reprinted with permission from ref. [8])

P



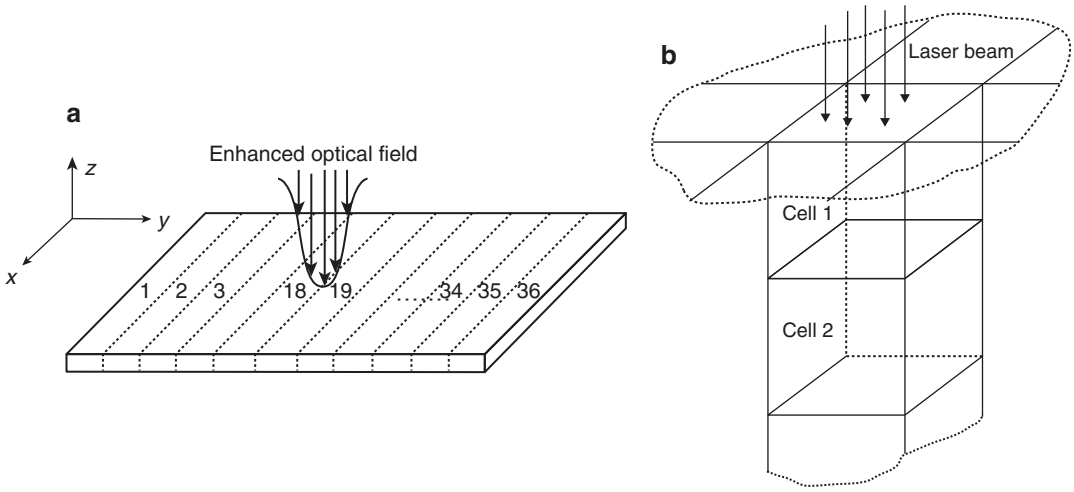
Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 7 (**a**) Optical intensity distribution inside the silicon substrate. The amplitude indicates the optical intensity enhancement relative to the incident wave. (**b**) Temperature distribution inside the silicon substrate. The initial temperature in silicon is 300 K. (**c**) The thermal stress distribution induced by temperature rise inside the silicon substrate. The upper figures show the top of the substrate, and the lower figures show the vertical planes (side view) in silicon under the particle center (Reprinted with permission from ref. [9])

temperature of the substrate are set at 300 K. The calculated temperature profile in the silicon substrate beneath the silica particle is shown in Fig. 6b.

The stress distribution in the silicon substrate beneath a single silica particle is also studied [9]. The thermal stress distribution inside the silicon substrate (Fig. 7) is simulated with ANSYS V14. The model as well as the temperature distribution is imported from the FLUENT calculation. The peripheral and foot end surfaces of the model are fixed. The top boundary is not constrained. Both vertical cross sections are symmetrical. The calculated maximum thermal stress in the single particle case is around 30 MPa, which is much lower than the experimental value (160 Mpa) determined from Raman shift and linewidth. The difference between the experiment and the simulation is mainly induced by the setting of the incident laser and the out-of-focus effect in experiment.

Material Structure Evolution Under Near-Field Laser Heating

In laser-assisted STM nanostructuring, as a direct consequence of nanoscale heating, the nanodomain of a sample just below the STM tip experiences intense heating, phase change, phase explosion, stress development and propagation, and rapid structural evolution. The enhanced optical field, light confinement, and only the thermal evolution of the sample surface under the STM tip without phase change have been widely studied. Furthermore, in surface nanostructuring, the final surface nanostructure is strongly affected by solidification of the molten material. The solidification process is driven by the heat conduction in the material and is much slower than the heating and melting processes that are largely driven by the laser energy input. Previous works mainly focused on the simulation in a small scale and short time duration. In this section, studies for a longtime behavior of a large system consisting of more than one hundred million atoms are introduced for the material structure evolution: heating, melting, phase change, solidification, and defect formation in the laser-assisted STM nanostructuring. The model used in the simulation is designed with true dimensions though the simulated result is based on several artificial assumptions.



Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 8 (**a**) Schematic of the MD configuration. (**b**) Schematic of the laser beam absorption in the material (Reprinted from ref. [14], © IOP Publishing. Reproduced by permission of IOP Publishing. All rights reserved)

Methodologies of Large-Scale Parallel Molecular-Dynamics (MD) Simulation

At nanoscale, the thermal movements of molecules/atoms show strong statistical variations in space since thermal equilibrium cannot be well established. MD simulation, which directly tracks the movement of molecules/atoms, is capable of exploring the physical phenomena down to molecular/atomic levels. A designed thin free-standing argon film at 50 K is irradiated with a near-field focused laser beam. The distribution of the optical field in space and time for this laser beam is constructed as

$$I = I_0 \exp\left(\frac{-\left(\vec{r} - \vec{r}_0\right)^2}{r_g^2}\right) exp\left(\frac{-(r - r_0)^2}{r_g^2}\right), \tag{1}$$

where I_0 is a laser beam intensity constant, r is the location of the laser beam, r_0 is the center of the sample surface, r_g is a length constant determining the size of the laser beam, t is the time, t_0 is a time constant determining the peak location of the laser intensity in the time domain, and t_g is a time constant determining the pulse width of the laser beam.

To simulate the laser–argon film interaction, the movement of atoms is governed by the Newtonian equation,

$$m_i \frac{d^2 r_i}{dt^2} = \sum_{j \neq i} F_{ij}, \tag{2}$$

where m_i and r_i are the mass and position of atom i, respectively. F_{ij} is the interaction force between atoms i and j and is computed from the Lennard-Jones (LJ) potential as $F_{ij} = -\partial \phi_{ij} / \partial r_{ij}$. The Lennard-Jones potential ϕ_{ij} follows the form

$$\phi_{ij} = 4\varepsilon \left[\left(\frac{\sigma_e}{r_{ij}}\right)^{12} - \left(\frac{\sigma_e}{r_{ij}}\right)^{6} \right], \tag{3}$$

The half-step leapfrog scheme is used in the simulation. For argon, the cutoff distance r_c adopts the widely accepted value $2.5\sigma_e$. Computation of the force between an atom and its neighbors is arranged by the cell structure and the linked-list method.

When studying the longtime behavior of a large system, the physical domain is divided into N subdomains of the same size as shown in Fig. 8a, and these subdomains are distributed over a parallel computer consisting of

N computing nodes. Each node is responsible for the MD simulation of one subdomain. The parallel program is developed using MPICH. Periodical boundary conditions are applied to the boundaries in the x and y directions and free boundary conditions to the boundaries in the z direction. Details of the parallel treatment are discussed in reference [12]. Nowadays, various MD simulation packages are available for parallel computation. Still this subdomain scheme saves significant time for data exchange among computing nodes.

The numerical treatment of the laser absorption in the sample is shown in Fig. 8b. The MD domain is divided into cubic cells with a little larger size than cutoff distance. Take cells 1 and 2 as examples; cell 1 is the first cell to absorb the laser energy when the laser propagates from one layer to another along the z direction. The laser energy absorbed by cell 1 within each time step is

$$\delta E_1 = E_1\left(1 - exp\left(\frac{-\delta z}{\tau_1}\right)\right), \qquad (4)$$

where E_1 is the laser energy reaching cell 1 within one time step δt, $E_1 = \int_A I\delta t \cdot dA$. δz is the z dimension of cell 1. τ_1 is the modified artificial absorption depth due to heating, explosion, and other physical processes. The artificial absorption depth is designed to reflect the volumetric absorption of the laser beam in the sample rather than to represent the practical experimental condition. When laser arrives at cell 2, the energy is $E_2 = E_1$ - δE_1. The laser energy absorbed by cell 2 can be calculated with eqn. 4. The laser keeps propagating and the energy absorption process is continued in the following cells until the energy is totally absorbed. The kinetic energy of atoms in the material is excited by the energy transferred from the laser beam, while the momentum of atoms is conserved during laser energy absorption.

The Material Structure Evolution: Heating, Melting, Phase Change, Solidification, and Defect Formation

The solidification and crystallization of an argon film of 100,776,960 atoms is simulated over 2 ns [13] after the irradiation of a near-field enhanced pulse laser. The film is comprised of 648 fcc unit cells (350.83 nm) in the x and y directions and 60 fcc unit cells (32.48 nm) in the z direction. The length of the film is much larger than the beam spot of 25 nm, and the thickness is much thicker than the optical absorption depth 10 nm. The pulse energy is taken as 8×10^{-15} J, which corresponds to an energy density level of 50 MW/cm^2. Before applying laser heating, the simulation is run for 100 ps, during which the velocity of atoms is scaled to make the sample to reach 50 K gradually. After this velocity scaling, the sample is simulated for another 100 ps to eliminate the disturbance introduced by the velocity scaling. Free spaces are added below and above the sample to capture the movement of atoms escaping from the sample. The entire computational domain measures 350.83 nm in the x and y directions and 759.10 nm in the z direction with the back side of the sample located at 213.71 nm.

Epitaxial regrowth process is observed in this longtime simulation. The atomic dislocation is induced by thermal strain in the epitaxial regrowth. The normal compressive stress and negligible shear stress decline when the melting occurs. The distance between atoms is studied to reflect the crystallinity and distinguish solid and liquid. The solid–liquid interface moves at an average speed of 3–5 m/s which increases with time, and the spatial distribution of the moving speed at the solid–liquid interface in nonuniform. Two edges of the liquid region in the nanohole solidify first, and their solidification speed is faster than that at the bottom because the heat conduction area at the two edges is much stronger, which leads to the appearance of nonuniform epitaxial regrowth.

Irregular perturbations at the edge of the nanohole are observed in simulating the surface nanostructuring with a pulse-laser-assisted STM [14]. To explore the thermal and mechanical mechanisms, another system consisting of more than 200 million atoms is modeled by employing large-scale parallel MD simulations. The sample is designed as 648 fcc unit cells (350.83 nm) in the x and y directions and 60 fcc unit cells (32.48 nm) in the z direction. The pulse energy is taken as 1.6×10^{-14} J, which corresponds to an energy

Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Table 1 Detailed information about the three domains studied

	Sample 1	Sample 2	Sample 3
FCC cubes in the domain ($x \times y \times z$)	$5 \times 324 \times 60$	$5 \times 648 \times 60$	$5 \times 324 \times 120$
Domain size (nm^3)	$2.707 \times 175.4 \times 32.48$	$2.707 \times 350.8 \times 32.48$	$2.707 \times 175.4 \times 64.97$
Atom number	388,800	777,600	777,600

Reprinted from reference [15], Copyright (2008), with permission from Elsevier

density level of 100 MW/cm^2. Strong compressive stress develops beneath the sample surface due to the absorption of the strong laser power. These compressive stresses then drive atoms' movement to induce phase explosion. In addition, the anisotropic propagation of the stress occurs in the material because of the direction-dependent stress wave speeds in the crystal. Visible subsurface structural damages are observed in the direction of 45° with respect to the z axis. The strong tensile stress resulting from the reflection of the compressive stress at the bottom plays the most significant role in inducing the structural damage.

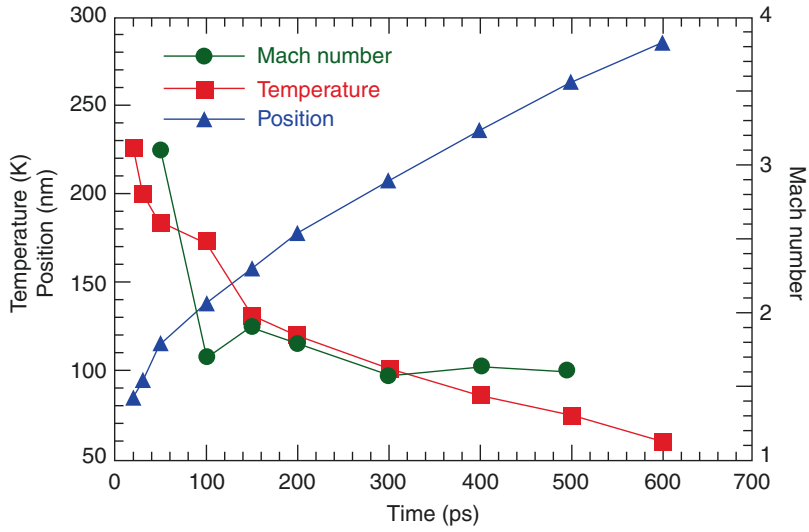
The solidification process of the nanohole after laser heating is also studied by varying the power of the pulse laser [15]. The MD simulation is performed on a very small surface area (tens of nanometers in diameter) in argon crystals, with eight different levels of laser beam energy input: 0.05, 0.1, 0.2, 0.35, 0.5, 0.75, 1, and 1.1 fJ. The corresponding maximal energy density levels are about 12, 24, 48, 84, 120, 180, 240, and 260 MW/cm^2, respectively. The detailed sizes of simulated domains are listed in Table 1. The effect of stress reflection from the bottom of the sample on the final nanostructure is reduced by using suitable large domain size. The diameter, depth, and edge protrusion of the nanohole increase when the laser fluence slows down. The largest melting thickness is reached after the laser heating stops because the melting procedure still continues due to the fact that heat conduction is continuing and its rate is much lower than the heating speed.

Shock Wave

During the process of laser-assisted SPM nanostructuring, intense laser heating will make the solid-state material transfer directly into gas phase to result in phase explosion. The ejected atoms have high kinetic energy and movement speed. When an ambient gas is present, this high-speed movement could compress ambient gas atoms in a very short time, leading to a strong shock wave. It is a very important phenomenon though it is little useful in nanomanufacturing. The first atomistic modeling work on shock wave was done by Zhang et al. [16] in picosecond laser–material interaction. The formation of the shock wave in the nanoscale domain could significantly affect the laser–material interacting zone and the nanostructuring process.

In the shock wave phenomena, the kinetic and physical properties of the disturbed gas compression are explored in a nanodomain in near-field laser–material interaction [17]. The gas compression is driven by the high-speed movement of the molten particulates ejected from the solid target in a nanodomain. The quick interaction between solid and gas atoms compresses the gas and forms a steep shock wave front, which moves at a supersonic speed. The fast compression of gas also induces a steep interface of density, temperature, and pressure distribution, which are viewed as typical characteristics of nanoscale shock waves. Evolutions of shock wave front position, velocity, and Mach number are shown in Fig. 9, illustrating quick decay during wave propagation. Because of energy dissipation, the temperature, pressure, and speed of the shock wave experience a fast decay, but the Mach number always remained higher than 1, meaning the shock wave always persisted in the simulation.

A one-dimensional model is established to simply describe the shock wave front in nanosecond laser–material interaction. The longtime



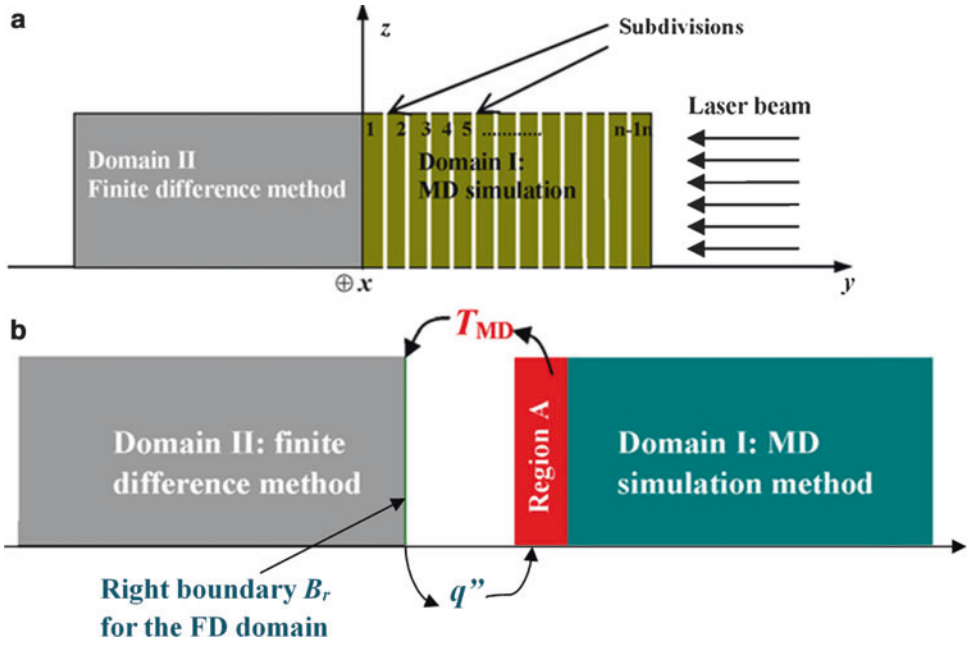
Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 9 Evolution of the shock wave front temperature, position, and Mach number against time (Reprinted from ref. [17], Copyright (2007), with permission from Elsevier)

phase change, including solid–liquid interface speed, temperature, maximum melting depth, and ablation rate, is systematically studied [18]. The whole modeled domain is divided into two subdomains as shown in Fig. 10. Subdomain I is directly subjected to the nanosecond laser irradiation, and its length along the y direction is much longer than the laser melting depth. Subdomain II only experiences heat conduction, and its length is longer than the thermal diffusion length in the material. Subdomain I and II are simulated with finite difference method and MD method, respectively. The temperature and heat flux are continuous at the interface. Periodical boundary conditions are used along the x and z directions as shown in Fig. 10. Strong fluctuation is observed at the solid–liquid interface and surface of the molten pool. No visible superheating is observed at the solid–liquid interface. The melting stops later. A higher peak melting speed appears with higher laser fluence. An almost linear relationship is observed between the ablation yield and the laser fluence, indicating weak phase explosion.

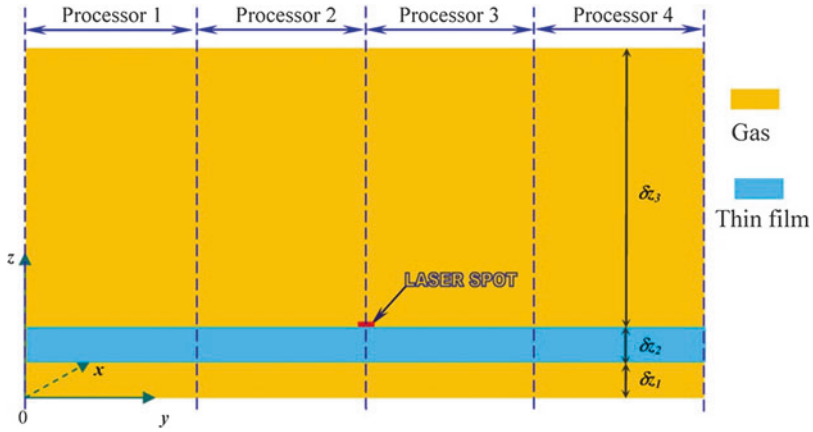
The properties of background gas also have effects on the shock wave [19, 20]. The molecular weight and density of background gas are considered on the laser-induced shock wave. A quasi-three-dimensional model (Fig. 11) is constructed to study a free argon thin film in a gas environment. This simulated system consists of over 2 million atoms so the parallel computation is employed. Eight cases of different background gases are studied in two groups. Details are listed in Table 2. The lower ambient pressure or lighter background gas causes the larger area being affected and a weaker shock wave front. The molecular weight of background gas has little effect on the final kinetic energy of the ablated material and the energy of the background gas regardless of the ambient pressure and laser fluence. When the ambient pressure is reduced, more kinetic energy is carried out by the ablated material, while less energy is transferred to the ambient gas. By studying the ablation in the temporal domain, a heavier ambient gas is observed to bounce back the ablated material.

Concluding Remarks

The experimental and computational works mainly focus on the laser-assisted tip–substrate interaction. The near-field focus effect largely enhances the optical field in the substrate under the tip other than inside the tip. The large parallel MD simulations are carried out to study longtime surface evolution after an intensive pulse applied on it including material structure evolution: heating, melting, phase change, solidification, and defect formation. A shock wave is observed



Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 10 (**a**) Schematic of the computational domain; (**b**) schematic for treatment of the interface between the MD and FD domains (Reprinted from ref. [18], Copyright (2008), with permission from Elsevier)



Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Fig. 11 Schematic of the MD configuration for shock wave study and parallel computation (Reprinted from ref. [19], © IOP Publishing. Reproduced by permission of IOP Publishing. All rights reserved)

and systematically studied. The effect of gas background is also considered.

There are some problems still faced in nanomanufacturing. The nanostructuring is a dynamic process, and the processing area is very small (low to the nanoscale). The phenomena occurring in this area are difficult to observe and measure. In the SPM tip-assisted nanomanufacturing, the near-field effect will be easily changed due to vulnerable tip. Either the bump damage or the contamination of the apex of the tip will broaden the tip and finally changes the near-field optical field distribution in and around the tip. Furthermore, in this tip–substrate system, the tip and the substrate are so close to be considered in contact with each other. Therefore, any thermal expansion from the tip and the substrate induced by the local laser heating will result in

Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing, Table 2 Conditions, domain sizes, and total number of atoms for different cases

Group	Conditions	δz_1 (nm)	δz_1 (nm)	δz_1 (nm)	l_x (nm)	l_y (nm)	N
1	$\beta = 2, \gamma = 1$	21.656	32.484	541.4	2.707	920.38	2480960
	$\beta = 1, \gamma = 1$	21.656	32.484	541.4	2.707	920.38	2480960
	$\beta = 0.5, \gamma = 1$	21.656	32.484	541.4	2.707	920.38	2480960
	$\beta = 0.25, \gamma = 1$	21.656	32.484	541.4	2.707	920.38	2480960
	$\beta = 0.125, \gamma = 1$	21.656	32.484	541.4	2.707	900.89	2429440
2	$\beta = 1, \gamma = 0.25$	34.650	34.650	1732.479	2.707	900.89	2469376
	$\beta = 1, \gamma = 0.0626$	34.650	34.650	1732.479	2.707	900.89	2214784
	$\beta = 1, \gamma = 0.0157$	34.650	34.650	1732.479	2.707	900.89	2151136

Reprinted from reference [19], © IOP Publishing. Reproduced by permission of IOP Publishing. All rights reserved

measurement instability. Right now the measurement for laser-assisted tip-based nanomanufacturing takes a longer time compared to the time of the dynamic process in the manufacturing. The manufacturing is confined in the nanoscale as well. Future improvement of this technique is required to develop the quick response of the measurement and scale up the manufacturing and improve the scalability in industries.

Cross-References

- Heat
- Heat Capacity
- Heat Conduction
- Heat Conductivity
- Molecular Dynamics Method
- Nanofabrication
- Nanoscale Heat Transport
- Nano-Scale Structuring
- Nanoscale Thermal Analysis
- Nanotechnology
- Near-Field Optics
- Near-Field Scanning Optical Microscopy (NSOM)
- Thermal Conductance
- Thermal Resistance
- Thermal Transport
- Thermometry

References

1. Lu, Y.F., Hu, B., Mai, Z.H., Wang, W.J., Chim, W.K., Chong, T.C.: Laser-scanning probe microscope based nanoprocessing of electronics materials. Jpn. J. Appl. Phys. **1**(40), 4395–4398 (2001)
2. Chen, X.W., Wang, X.W.: Near-field thermal transport in a nanotip under laser irradiation. Nanotechnology **22**, 075204 (2011)
3. Chen, X.W., Wang, X.W.: Microscale spatially resolved thermal response of Si nanotip to laser irradiation. J. Phys. Chem. C **115**, 22207–22216 (2011)
4. McCarthy, B., Zhao, Y.M., Grover, R., Sarid, D.: Enhanced Raman scattering for temperature measurement of a laser-heated atomic force microscope tip. Appl. Phys. Lett. **86**, 111914 (2005)
5. Yue, Y.A., Chen, X.W., Wang, X.W.: Noncontact Sub-10 nm temperature measurement in near-field laser heating. ACS Nano **5**, 4466–4475 (2011)
6. Downes, A., Salter, D., Elfick, A.: Finite element simulations of tip-enhanced Raman and fluorescence spectroscopy. J. Phys. Chem. B **110**, 6692–6698 (2006)
7. Chen, G.: Nonlocal and nonequilibrium heat conduction in the vicinity of nanoparticles. J. Heat Trans-T. ASME **118**, 539–545 (1996)
8. Tang, X.D., Xu, S., Wang, X.W.: Nanoscale probing of thermal, stress, and optical fields under near-field laser heating. Plos One **8**, e58030 (2013)
9. Tang, X.D., Xu, S., Wang, X.W.: Thermal probing in single microparticle and microfiber induced near-field laser focusing. Opt. Express **21**, 14303–14315 (2013)
10. Tang, X.D., Yue, Y.N., Chen, X.W., Wang, X.W.: Sub-wavelength temperature probing in near-field laser heating by particles. Opt. Express **20**, 14152–14167 (2012)
11. Beechem, T., Graham, S., Kearney, S.P., Phinney, L. M., Serrano, J.R.: Invited article: simultaneous mapping of temperature and stress in microdevices using micro-Raman spectroscopy. Rev. Sci. Instrum. **78**, 061301–9 (2007)
12. Wang, X., Lawrence, C.: In: Proceedings of the ASME heat transfer/fluids engineering summer conference 2004. Charlotte, NC. American Society of Mechanical Engineers, New York, 2004
13. Wang, X.W., Lu, Y.F.: Solidification and epitaxial regrowth in surface nanostructuring with laser-assisted scanning tunneling microscope. J. Appl. Phys. **98**, 114304 (2005)

14. Wang, X.W.: Large-scale molecular dynamics simulation of surface nanostructuring with a laser-assisted scanning tunnelling microscope. J. Phys. D Appl. Phys. **38**, 1805–1823 (2005)
15. Feng, X.H., Wang, X.W.: Effects of laser fluence on near-field surface nanostructuring. Appl. Surf. Sci. **254**, 4201–4210 (2008)
16. Zhang, L.J., Wang, X.W.: Dynamic structure and mass penetration of shock wave in picosecond laser-material interaction. Jpn. J. Appl. Phys. **47**, 964–968 (2008)
17. Feng, X.H., Wang, X.W.: Nanodomain shock wave in near-field laser-material interaction. Phys. Lett. A **369**, 323–327 (2007)
18. Zhang, L.J., Wang, X.W.: Hybrid atomistic-macroscale modeling of long-time phase change in nanosecond laser-material interaction. Appl. Surf. Sci. **255**, 3097–3103 (2008)
19. Guo, L.Y., Wang, X.W.: Effect of molecular weight and density of ambient gas on shock wave in laser-induced surface nanostructuring. J. Phys. D Appl. Phys. **42**, 015307 (2009)
20. Li, C., Zhang, J.C., Wang, X.W.: Phase change and stress wave in picosecond laser-material interaction with shock wave formation. Appl. Phys. A-Mater. **112**, 677–687 (2013)

Phytotoxicity

► Toxicology: Plants and Nanoparticles

Piezoelectric Effect at Nanoscale

Zhong Lin Wang and Ying Liu
School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Synonyms

Nano-piezoelectricity

Definition

Piezoelectricity is a phenomenon of strain-induced electric polarization in certain crystals, which can be used to create a mechanical action by applying an external voltage for sensors and actuators or a mechanical straining that can produce a voltage for energy conversion.

Piezoelectricity

Piezoelectricity was discovered by Jacques and Pierre Curie during their systematic studies of the effect of pressure on the generation of electrical charge by crystals in 1880. It is a reversible physical phenomenon, as there is the direct piezoelectric effect and the reverse piezoelectric effect.

Direct piezoelectric effect is defined as: *electric polarization produced by mechanical strain in crystals belonging to certain classes, the polarization being proportional to the strain and changing sign with it* [1].

Reverse (or inverse) piezoelectric effect is defined as: *a piezoelectric crystal becomes strained, when electrically polarized by an amount proportional to the polarizing field*.

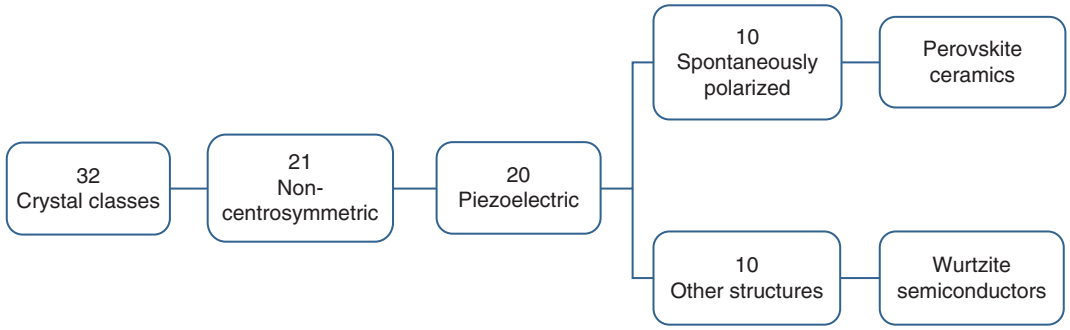
Structures of Piezoelectric Crystals

The nature of piezoelectricity comes from the lack of centrosymmetry/inversion symmetry in the crystal (Fig. 1). Twenty one out of all 32 crystal classes lack inversion symmetry, and 20 of them exhibit piezoelectric effect [2]. The two most important classes include perovskite structure, which occurs in ceramics such as lead zirconate titanate (PZT) and barium titanate ($BaTiO_3$), and wurtzite structure, which occurs in semiconducting materials such as zinc oxide (ZnO), gallium nitride (GaN), indium nitride (InN), and zinc sulfide (ZnS).

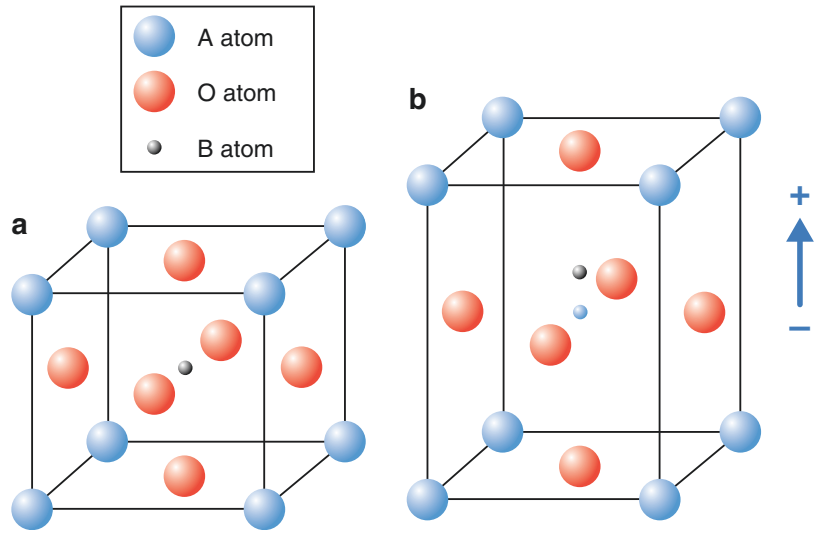
Perovskite Structure

The perovskite structure gets its name from the mineral $CaTiO_3$ and is adopted by many oxides that have the chemical formula of ABO_3 [3]. The most widely used piezoelectric ceramics, including barium titanate ($BaTiO_3$) and lead zirconate titanate (PZT), are formed in this structure.

As shown in Fig. 2a, in the unit cell of perovskite compound, type "A" atom is in the cube corner



Piezoelectric Effect at Nanoscale, Fig. 1 Scheme of piezoelectric crystal classes

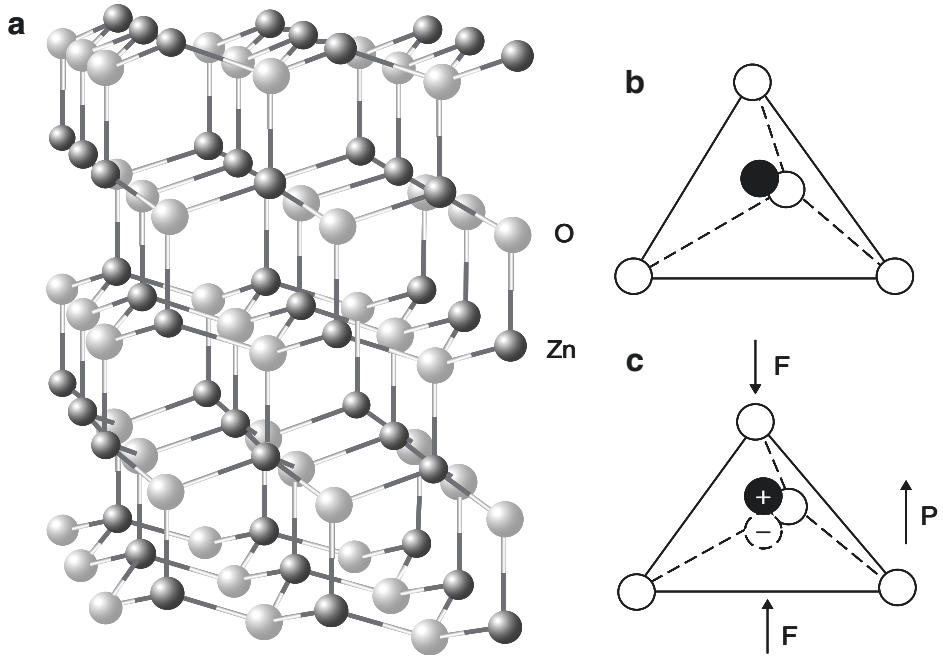


Piezoelectric Effect at Nanoscale, Fig. 2 (**a**) Ideal perovskite structure, (**b**) perovskite structure with distortion and polarization along *c*-axis

positions (0, 0, 0), type "B" atom is in the body center position (1/2, 1/2, 1/2), and oxygen atoms are in the face-centered positions (1/2, 1/2, 0). However, in real cases, since ion sizes hardly meet the stringent requirement for stability, most perovskite structures are distorted and lose their cubic symmetry and commonly shift into orthorhombic or tetragonal phases [4]. When "B" atoms are displaced in such distortion, as shown in Fig. 2b, the center for negative charge and positive charge in the crystal no longer coincides with each other and results in an electric dipole. This is where the piezoelectricity comes from.

Since the piezoelectricity originates from rather randomly oriented distortion, it is much more complicated and is correlated with the concept of ferroelectricity. In most cases, perovskite ceramics forms polycrystalline structures with multiple electric "domains." Each domain has a size in the order of 10–100 nm, and even though each individual domain is piezoelectric itself, the whole ceramic does not show macroscopic dipole without poling, as they have differently oriented electric dipoles. Electrical alignment of the dipoles, or electric poling as stated in most cases, is required to activate the piezoelectric effect in the ceramics. Similar to ferromagnetic materials, a characteristic temperature, the Curie temperature (T_C), exists in perovskite ceramics, above which they lose their electric polarization induced in the poling process.

There are multiple methods for fabricating perovskite ceramic nanostructures [5]. The top-down methods include electron-beam-



Piezoelectric Effect at Nanoscale, Fig. 3 ZnO with wurtzite structure and illustration for the formation of dipoles ([8], Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission)

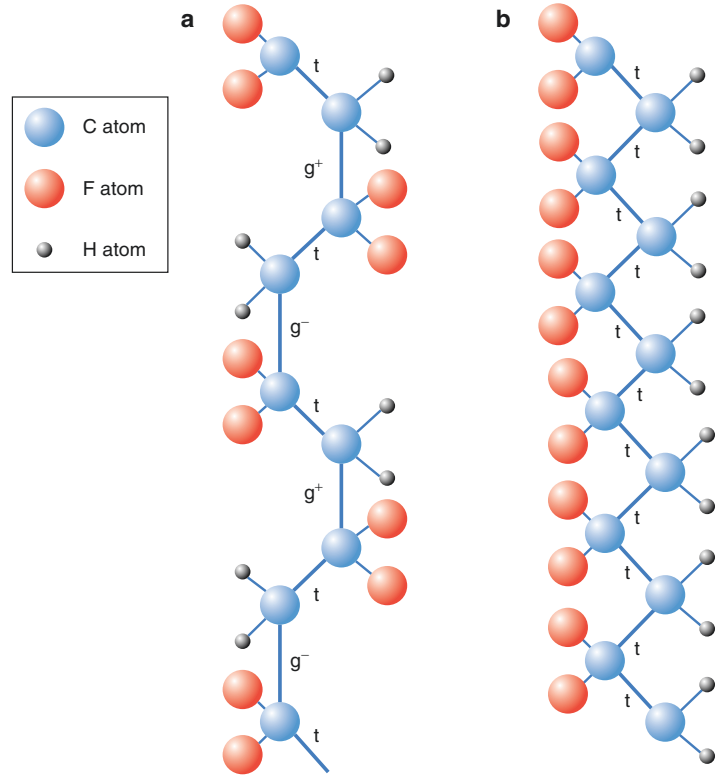
assisted fabrication, nanoprinting, and self-assembling and they provide high-precision positioning and size control. However, top-down methods are limited in resolution and prone to processing damage. In comparison, the bottom-up approaches allow much smaller feature sizes but suffer from poor registration and thus require substrate patterning that in turn involves top-down processing. For one-dimensional nanostructure, solution-phase decomposition and hydrothermal chemical synthesis [6] can result in single crystalline nanorods and nanowires, sol–-gel method associated with nanochannel templates or electrospinning can produce polycrystalline nanowires, and wetting of the porous templates can produce nanotubes.

In all of the conventional piezoelectric materials, PZT has the best performance and the widest application. Its piezoelectric coefficient d_{33} can be as high as 400 pC/N and can be higher than 600 pC/N when doped with lanthanum [2, 7]. However, since it is a lead-containing chemical, it is not environmentally and biologically friendly without proper packaging, and its applications in the nanoscale region are thus restricted, as such applications are often related to biomedical or environmental issues.

Wurtzite Structure (ZnO, GaN)

The wurtzite crystal structure is named after the mineral wurtzite and is the crystal structure for various binary compounds. Wurtzite structure occurs in most piezoelectric semiconductors, including zinc oxide (ZnO), gallium nitride (GaN), cadmium sulfide (CdS), silicon carbide (SiC), and so on, and among them, ZnO is the most widely used in one-dimensional nanostructure and the key material for nanogenerators, piezotronics, and piezophototronics, as will be discussed later.

As shown in Fig. 3, in a Wurtzite structure, each of the two individual atom types forms a sublattice which is hexagonal close pack (HCP) type, and they form tetrahedral coordination with each other [8]. Distinct from Perovskite structure, Wurtzite structure crystals do not need distortion



Piezoelectric Effect at Nanoscale,
Fig. 4 Illustration of the two main conformations of PVDF. (**a**) α conformation, (**b**) β conformation

or external poling to break the centrosymmetry. In single crystal Wurtzite materials, there are strong and uniform dipoles along the *c*-axis, and the dipole direction is determined by the stacking order of the anions and cations and, more specifically, *+c* for cation to anion and *-c* for anion to cation. The orientation and polarity determine the direction of the piezoelectricity in the crystal. In ZnO, the interaction of the polar charges at the surface results in the growth of a wide range of unique nanostructures, such as nanobelts, nanosprings, nanorings, and nanohelices.

The synthesis of ZnO nanostructures is easy and convenient. To fabricate large-scale nanowire arrays, chemical hydrothermal at 80 °C can be carried out on any substrate, while physical vapor liquid solid (VLD) or physical or chemical vapor solid (PVD or CVD) growth can give results with higher aspect ratio and higher density on substrate with proper lattice match.

As will be discussed in latter sections, one-dimensional wurtzite nanostructures, especially ZnO nanowires and nanobelts, promote the new concept of "piezotronics," which is the coupling between piezoelectricity and semiconductor transport property and "piezophototronics," which is the result of three-way coupling of piezoelectricity, photonic excitation, and semiconductor transport property.

Piezoelectric Polymer

Piezoelectricity also occurs in semicrystalline polymer materials. The most famous and widely used of them are polyvinylidene fluoride (PVDF) and its copolymers [9].

In PVDF, hydrogen atoms are positively charged, and fluoride atoms are negatively charged with respect to the carbon atoms so that each monomer has an inherent dipole moment. As shown in Fig. 4, PVDF has several conformations: α (tg^+tg^-), β (all trans), and γ ($tttg^+ttg^-$). In α conformation, the dipoles along the chain cancel each other so that the overall piezoelectric effect is zero, and this conformation is the non-piezoelectric conformation. In the β conformation, the alignment of all its dipoles is in the

same direction normal to the chain axis. In this way, once strain is applied on the β-phased polymer and the chain is either stretched or compressed, the dipole strength will change, and extra net charge will be induced so that this conformation is piezoelectric. Since the nature of PVDF is ferroelectric, electric poling is also needed for the dipoles on the polymer chains to align so that the polymer will have piezoelectric response.

The crystallinity of pure PVDF is usually 50 ∼ 60 %. In its copolymer with TrFE and TFE, the crystallinity can be increased to nearly 100 % after annealing, and the portion of β conformation is also larger, and the piezoelectric response is thus greater. The piezoelectric coefficient in PVDF can be as high as a negative value of −33 pm/V, and that of PVDF-TrFE can be as high as −38 pm/V.

The synthesis of piezoelectric polymer nanofibers can be easily carried out through the process of electrospinning.

Theory of Piezoelectricity in One-Dimensional Nanostructures

In theoretical concern, piezoelectricity refers to the linear coupling of electricity and mechanics [10]. Considering electrical behavior, we have Maxwell's equation, $D = \varepsilon E$, where D is electric displacement, ε is permittivity, and E is electric field. Considering mechanical behavior, we have Hooke's law, $S = sT$, where S is strain, s is compliance, and T is stress. To describe piezoelectricity, the coupled equations are given as

$$\{S\} = [s^E]\{T\} + [d^t]\{E\}$$
$$\{D\} = [d]\{T\} + [\epsilon^T]\{E\}$$

where $[d]$ is the direct piezoelectric effect matrix and its transpose $[d^t]$ is the inverse piezoelectric effect matrix, and their specific form is related to the point group to which the crystal belongs. The most frequently used parameter is the piezoelectric coefficient d_{33}, with expression $d_{33} = (\partial D_3/\partial T_3)^E = (\partial S_3/\partial E_3)^T$, so the unit of d_{33} is either C/N or m/V, and generally in materials d_{33} is often in the order of 10^{-10}–10^{-12} C/N or m/V. When normal stress is applied along the polarization axis and on the surface charge is collected, d_{33} can be simplified as the volume change when subject to an electric field or the electric polarization when subject to stress. A similar parameter d_{31} is applied when stress is applied at right angle with the polarization axis.

Theoretical approaches have been applied on calculating piezoelectric potential distribution in piezoelectric nanostructures. A continuum model was built for calculating the electrostatic potential in a laterally bent nanowire (NW) [11].

To derive the relationship between the potential distribution in a NW and the dimensionality of the NW and magnitude of the force applied, four sets of equations for static piezoelectric material are first listed:

1. Mechanical equilibrium equation, $\nabla \cdot \sigma = \vec{f}_e^{(b)} = 0$, where σ is the stress tensor.
2. Constitutive equations, $\begin{cases} \sigma_p = c_{pq}\epsilon_q - e_{kp}E_k \\ D_i = e_{iq}\epsilon_q + \kappa_{ik}E_k \end{cases}$, where c_{pq} is the linear elastic constant, e_{kp} is the linear piezoelectric coefficient, and k_{ik} is the dielectric constant.
3. Geometrical compatibility equation, $e_{ilm}e_{jpq}\left(\partial^2\epsilon_{mp}/\partial x_l \partial x_q\right) = 0$ where e_{ilm} and e_{jpq} are Levi-Civita antisymmetric tensors. Small deformation of the NW is also assumed for simplicity of the derivation.
4. Gaussian equation of electric field, $\nabla \cdot \vec{D} = \rho_e^{(b)} = 0$, under the assumption that the piezoelectric NW has no conductivity.

For analytical analysis, a perturbation theory has been developed based on these equations. Under different orders of approximation, the four equations correspond to different extends of decoupling and coupling between the electric field and mechanical deformation: the zeroth-order solution stands for purely mechanical deformation without piezoelectricity; the first order is the result of direct piezoelectric effect, where stress/strain generates an electric field in the

NW; and the second order shows the first coupling of the piezoelectric field to the strain in the material. For the calculation of piezoelectric potential in an insulating NW, the first-order approximation is sufficient, and it gives analytical results less than 5 % away from finite element method (FEM) numerical calculation for potential distribution in a typical NW with diameter $d = 50$ nm and length $l = 600$ nm at a lateral bending force of $f = 80$ nN.

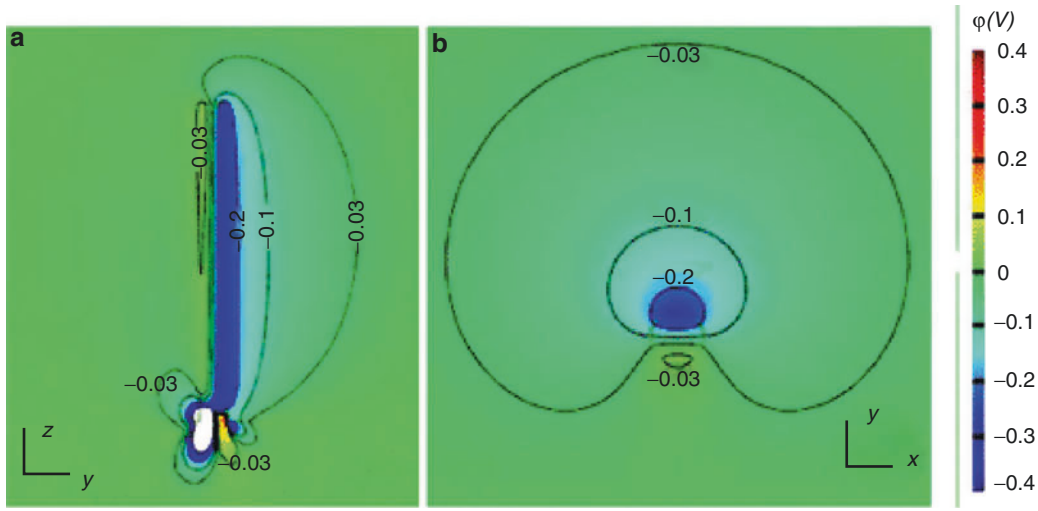
For further investigation on the interaction of piezoelectric potential with semiconductor transportation properties, charge carrier concentration can be substituted into the Gaussian equation of electric field, and the equation is modified as $\nabla \cdot \vec{D} = (\partial/\partial x_i)\left(e_{iq}\epsilon_q + \kappa_{ik}E_k\right) = \rho_e^{(b)} = ep - en + eN_D^+ - eN_A^-$, where p is the hole concentration in the valance band, n is the electron concentration in the conduction band, N_D^+ is the ionized donor concentration, and N_A^- is the ionized acceptor concentration. For n-type nanowires, it can be assumed that $p = N_A^- = 0$ and vice versa for p-type. Considering band structure and charge carrier distribution equations, there can be numerical solution of the modified equations. As shown in Fig. 5, FEM solutions based on these equations are given for a typical ZnO NW with moderate doping (Ref. [36] of [12]).

Further investigation is also carried out considering the contact of semiconducting NW with metals [13]. When a ZnO NW is under nonuniform deformation, the local piezoelectric potential distribution at different positions shows significant effect on and distinct trend of variation in the charge carrier transport characteristic. This can help understanding the characteristics of piezotronic devices based on NW of wurtzite materials by controlling local contact position and contact size.

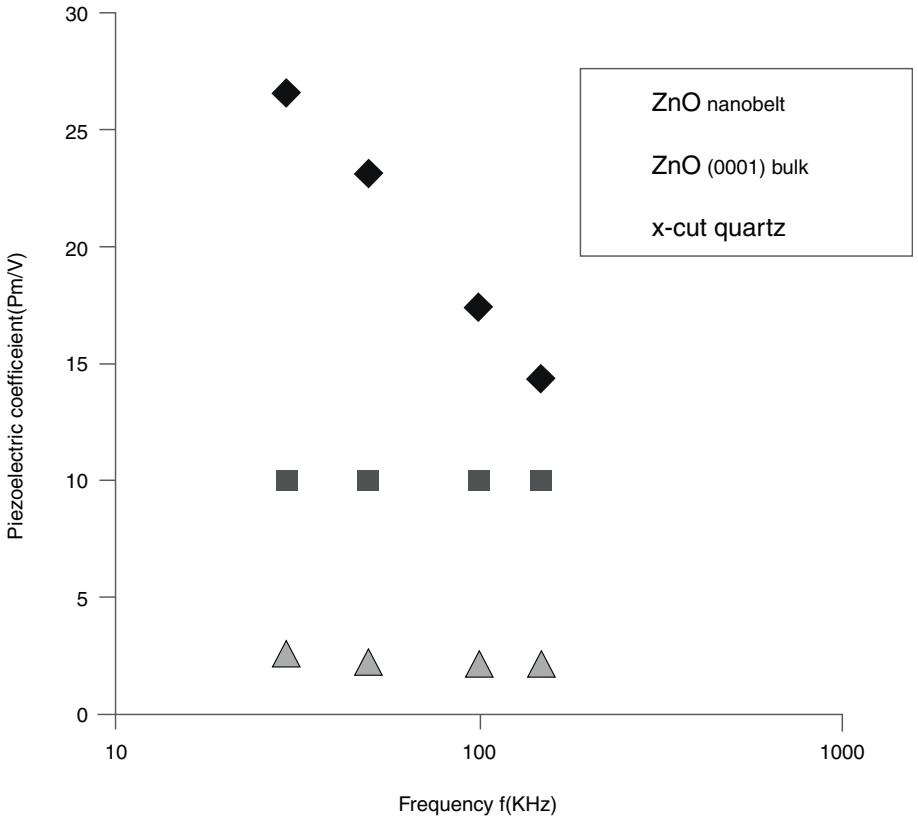
Effect of Size on Piezoelectricity

Size Effect in Wurtzite Nanocrystals

At nanoscale, many interesting effects occur, and this can affect physical properties of materials, including piezoelectric properties. In nanowires (NW) and nanobelts (NB), crystal defects such as dislocation and vacancies will concentrate and eliminate on the surface, and perfect single crystallinity is more easily achieved at this scale of dimension. As a result, during piezoelectric



Piezoelectric Effect at Nanoscale, Fig. 5 Plot of calculated piezoelectric potential φ. Regarding a ZnO NW with diameter $d = 50$ nm and length $l = 600$ nm at a lateral bending force of $f = 80$ nN for the NW is the same as in Fig. 4 except that it is n-type semiconducting with $N_D = 1 \times 10^{17}$ cm^{-3} (Adapted with permission from Ref. [36] of [12] "Gao, Y.F., Wang, Z.L.: Equilibrium potential of free charge carriers in a bent piezoelectric semiconductive nanowire. Nano Lett. **9**, 1103–1110 (2009)." Copyright 2009 American Chemical Society)



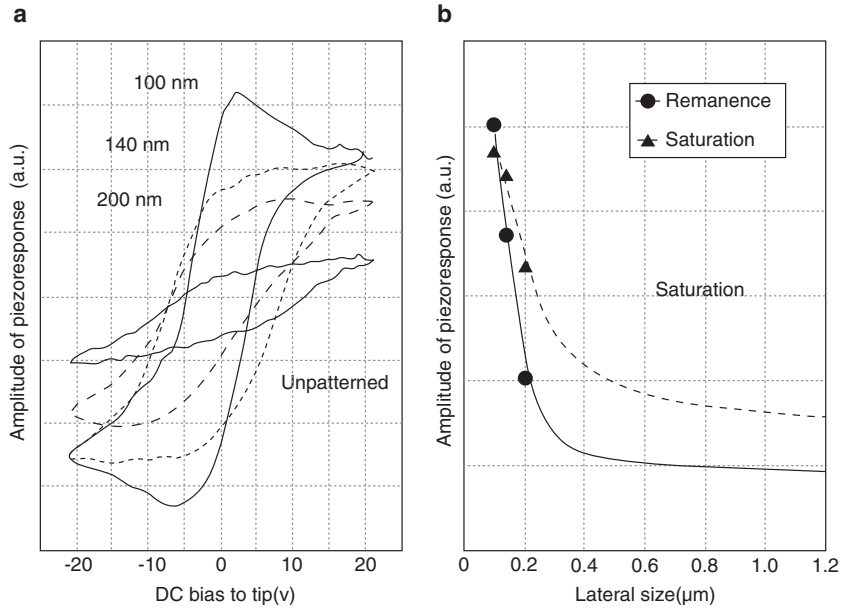
Piezoelectric Effect at Nanoscale, Fig. 6 Frequency dependence of piezoelectric coefficient of ZnO nanobelts, ZnO bulk material, and x-cut quartz (Adapted with permission from Ref. [41] of [12] "Zhao, M.H., Wang, Z.L., Mao, S. X.: Piezoelectric characterization of individual zinc oxide nanobelt probed by piezoresponse force microscope. Nano Lett. **4**, 587–590 (2004)." Copyright 2004 American Chemical Society)

measurement of individual zinc oxide NB with dimension of tens of nanometers in thickness, hundreds of nanometers in width and tens of micrometers in length, the effective piezoelectric coefficient d_{33} is revealed to be frequency dependent and varies from 14.3 to 26.7 pm/V, which is about onetime larger than that in bulk ZnO, around 10 pm/V, as shown in Fig. 6.

Size Effect in Perovskite Nanostructures

Since in perovskite materials, piezoelectric performance is tightly related to the concept of domain, the case of size effect is much more complicated, and there are multiple factors to take into consideration. Intrinsically, at nanoscale, competition exists between thermal vibrations and the correlation energy which aligns the dipoles, but greater surface area also helps eliminate defects. Extrinsically, grain boundaries and other microstructural factors can also have various influences.

Between 10 and 1,000 nm, which is bigger than or equal to the scale of a typical single domain and similar to the state of art piezoelectric nanodevices, controversial size effects can occur in differently processed materials regarding the sequence of patterning and crystallization of the nanostructure. This controversy is caused by extrinsic factors related with post annealing. In structures that were patterned with high-energy ion etching after crystallization, which leads to higher concentration of oxygen and lead vacancies, several reports reveal the inherent reduction in the piezoelectric coefficient in ultrathin PZT



Piezoelectric Effect at Nanoscale, Fig. 7 Ferroelectric loop showing increased piezoelectric response with shrinking size. (**a**) Loops on patterns with different lateral sizes and on the unpatterned film, and (**b**) remnant polarization and saturation as a function of lateral size [14]

films. With post annealing, nonswitching of pinned domains helps stabilize and enhance piezoelectric effect. In thin film studies, the observed increase of piezoelectric response amplitude can be as high as 300 % when decreasing the feature size of patterns on epitaxial PZT thin film from 200 to 100 nm, as shown in Fig. 7. The crystalline nature of patterns might also have contributed to this difference: single crystalline tends to have better piezo performance with smaller scale than polycrystalline. In template method, synthesized PZT NW enhanced piezoelectric response than the value in thin film was also discovered [5].

Under 10 nm, intrinsic factors have more influence on the piezoresponse. People have used phenomenological models and first principal/ab initial calculations to predict the behavior of such nanostructures, and their prediction is based on the competition and cancelation of thermal drift. In general, the results are shown as (1) depolarizing fields exist near the surfaces to reduce polarization and (2) piezoelectric state becomes unstable below certain sizes, which is estimated to be 7–10 nm along each of the three directions of well-known materials such as $BaTiO_3$. Microscopy study in this scale has been done and verified these predictions to some extent. This gives us important information that fabricating piezoelectric devices under the scale of 10 nm using perovskite structure materials will hardly result in better performance.

Applications of Nanoscale Piezoelectric Properties

As mentioned in previous sections, for materials with perovskite structure, the enhancement of performance at the nanoscale is not always significant, and related application is very limited. For materials with wurtzite structures however, especially for ZnO NW and NB, the interaction of piezoelectricity with semiconducting transportation process brings about new phenomena and draws great research interests, into the new concept of "piezotronics" and even "piezophototronics." These applications include piezoelectric field-effect transistor, piezoelectric diode, piezoelectricity-based nanosensors, piezoelectricity-enhanced photodetector, and,

most importantly, the piezoelectric nanogenerators.

Piezoelectric Nanogenerator

The future of nanotechnology research is likely to focus on the areas of integrating individual nanodevices into a nanosystem that acts like living specie with sensing, communicating, controlling, and responding. A nanosystem requires a nanopower source to make the entire package extremely small yet with high performance. The goal is to make self-powered nanosystem that can operate wirelessly, independently, and sustainably. Harvesting energy from the environment is a choice for powering nanosystems especially for biomedical applications. Since mechanical energy is a conventional form of energy that exists in our living environment, it is essential to explore innovative nanotechnologies for converting mechanical energy (such as body movement and muscle stretching), vibration energy (such as acoustic/ultrasonic wave), and hydraulic energy (such as body fluid and blood flow) into electric energy that will be used to power nanodevices without using battery. This is a key step toward *self-powered nanosystems* [15].

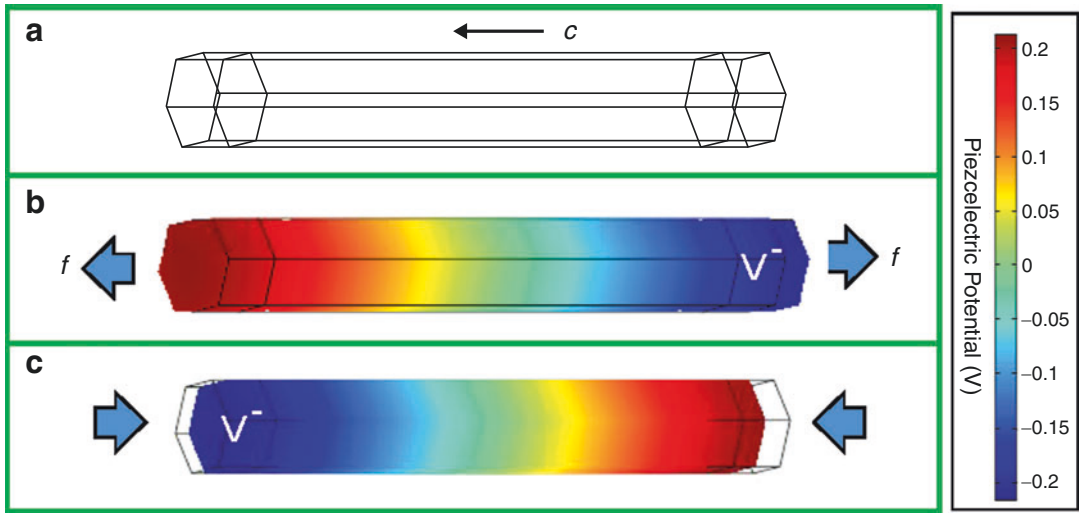
Dr. Z.L. Wang's group in Georgia Institute of Technology is the first to invent an innovative approach for converting mechanical energy into electric energy by piezoelectric zinc oxide nanowire arrays [16]. The operation mechanism of the nanogenerator (NG) relies on the piezoelectric potential created by an external strain; a dynamic straining of the nanowire results in a transient flow of the electrons in the external load due to the driving force of the piezopotential. The nanogenerator technology has been developed from fundamental science, to engineering integration and to technological scale-up. As today, a gentle straining can output 1–3 V at an instant output power of $\sim$2 μW from an integrated nanogenerator, using which a self-powered nanosensor has been demonstrated [6]. This technology has the potential applications for powering MEMS/NEMS that requires a power of the microwatt to milliwatt range.

The fundamental of nanogenerator relies on the presence of the piezoelectric potential (piezopotential) generated in a nanowire/nanobelt via dynamic straining. As demonstrated in previous sections, ZnO has the wurtzite structure, in which the tetrahedral coordinated O^{2-} and Zn^{2+} are stacked layer by layer along the *c*-axis. The lack of central symmetry results in the piezoelectric property of ZnO, which is vital for the mechanic-electric energy conversion with the ZnO-based nanogenerator.

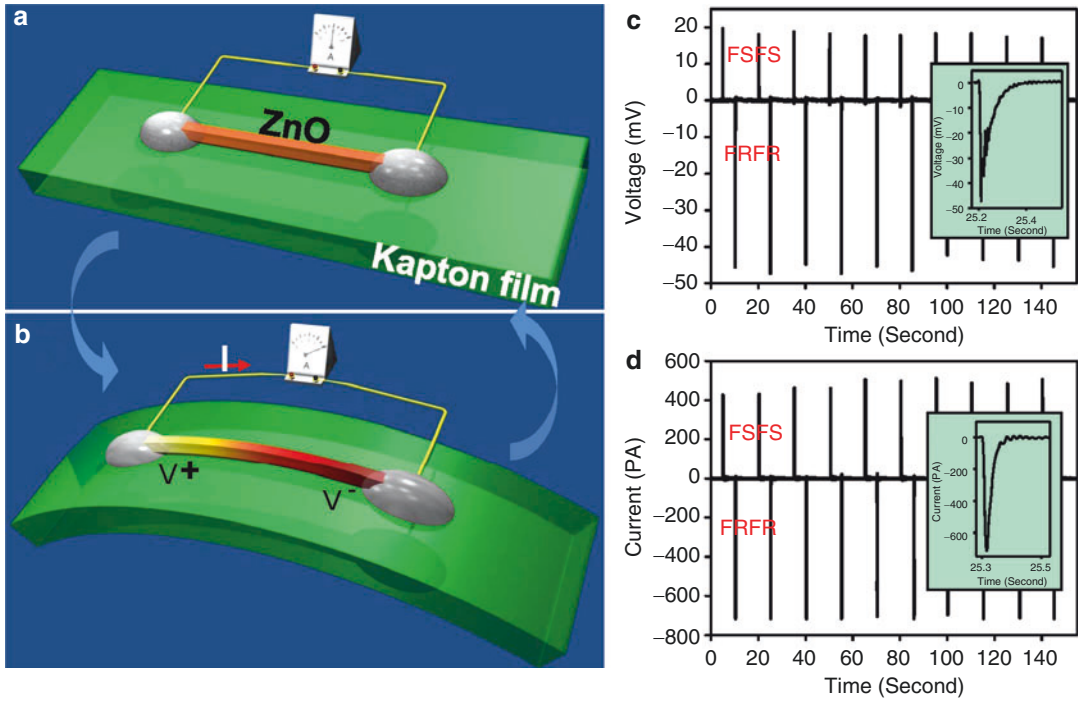
The cations and anions are tetrahedral coordinated in the wurtzite-structured ZnO. At the strain-free status, the charge center of the cations and that of anions coincides with each other. When an external strain is applied, the structure is deformed so that the charge centers for cations and anions separate and result in an electric dipole. Because the ionic charges are not free to move and the intrinsic free-charge carriers can only partially screen them if the doping level is low, the piezoelectric field is preserved as long as the NW is strained. The potential created by the polar ions is called *piezoelectric potential or piezopotential*.

The presence of piezopotential is the fundamental physical basis of the nanogenerators and piezotronics (Fig. 8). When a strained crystal is connected to an external load, the electrons in the circuit are driven to flow in to partially screen the piezopotential, which is the energy conversion process. Therefore, *the principle of the nanogenerator is the transient flow of electrons in external load as driven by the piezopotential created by dynamic straining*.

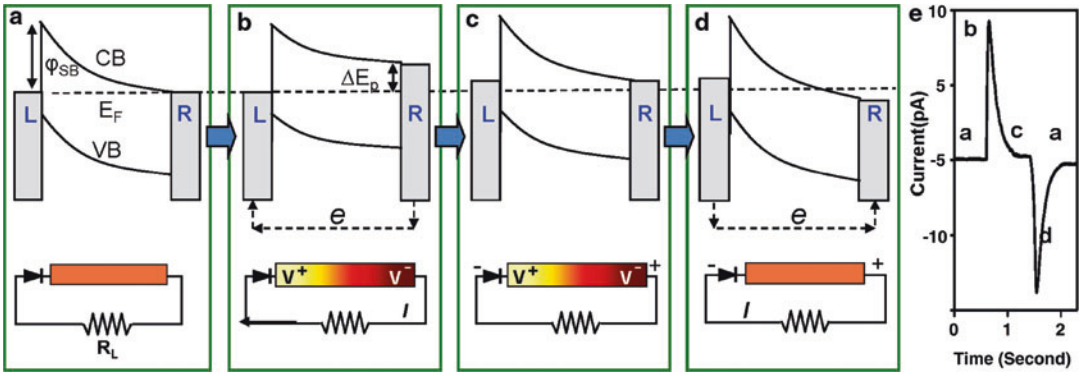
Piezopotential-driven transient flow of electrons in an external load is the principle of the nanogenerator. Figure 9 presents a single wire generator (SWG) using a laterally packaged ZnO wire. In brief, a single ZnO wire was placed laterally on the flexible polyimide film. Silver paste can be used to fix two ends of the ZnO wire to the substrate and link the wire to the external measuring instrument through metallic wires. Electric measurement shows that a functioning SWG usually has rectifying *I-V* characteristic, indicating the existence of the Schottky contact at least at one end. Before the SWG is deformed, there is no measurable potential drop from the ZnO wire. When the substrate is bent, both the substrate



Piezoelectric Effect at Nanoscale, Fig. 8 Numerical simulation of the piezopotential distribution in an insulating ZnO nanowire (**a**), the piezopotential distribution and the deformed shape of a ZnO nanowire grown along *c*-axis under a stretching force of 85 nN (**b**), or a compressing force of 85 nN (**c**) (Adapted with permission from Ref. [13] of [12] "Gao, Z.Y., Zhou, J., Gu, Y.D., Fei, P., Hao, Y., Bao, G., Wang, Z.L.: Effects of piezoelectric potential on the transport characteristics of metal-ZnO nanowire-metal field effect transistor. J. Appl. Phys. **105**, 113707 (2009)." Copyright 2009, American Institute of Physics)



Piezoelectric Effect at Nanoscale, Fig. 9 Design and power output from a single wire generator. (**a**) and (**b**) schematic diagram of single wire generator before and after stretching. Open-circuit voltage (**c**) and short-circuit current (**d**) of the single wire generator, which was cyclically fast stretched and fast released (Adapted from Ref. [16] of [12] "Yang, R.S., Qin, Y., Dai, L.M., Wang, Z.L.: Power generation with laterally packaged piezoelectric fine wires. Nat. Nanotechnol. **4**, 34–39 (2009)")



Piezoelectric Effect at Nanoscale, Fig. 10 Current generation mechanism explained with energy band diagram of the cyclically stretched NW. Energy band diagram of the ZnO NW in equilibrium and free condition (**a**), nonequilibrium and tensile-strained condition (**b**), rereached equilibrium and tensile-strained condition (**c**), and nonequilibrium and free condition (**d**). After (**d**), the NW will reach equilibrium as in (**a**) after charges flow. E_F, CB, VB, Φ_{SB}, and ΔE_p indicate Fermi level of the electrode, conduction band, and valence band of ZnO, Schottky barrier height, and piezopotential difference, respectively. (**e**) Short-circuit current measurement in which the labels of *a*, *b*, *c*, and *d* indicate the corresponding process shown in subparts *a*–*d* (Adapted from Ref. [16] of [12] "Yang, R.S., Qin, Y., Dai, L.M., Wang, Z.L.: Power generation with laterally packaged piezoelectric fine wires. Nat. Nanotechnol. **4**, 34–39 (2009)")

and the ZnO wire are under strain. Because the thickness of the substrate is much larger than the dimension of the ZnO wire, the ZnO wire is solely experiencing tensile strain when the bend is concave downward, as shown in Fig. 9b. As discussed earlier, tensile strain will result in piezopotential, such that the +*c*-axis side gains positive potential (light yellow) and −*c*-axis side gains negative potential (dark red). The potential difference between the two ends can then be measured either as open-circuit voltage (Fig. 9c) or short-circuit current (Fig. 9d).

The working mechanism of the SWG can be understood with the energy band diagram in Fig. 10. The Schottky contact at least at one end is necessary for the energy generation. It is assumed that the Schottky contact is at the +*c*-axis side (left side in Fig. 10a) and the energy generation follows in a similar way when the Schottky contact is on the other side or both sides. The entire wire is in equilibrium state without any strain and power output (Fig. 10a, e). When the wire is stretched, the tensile strain induces polarization of atoms, as well as piezoelectric field, in the crystal. Consequently, the +*c*-axis side holds positive potential and the −*c*-axis side holds negative potential. The potential difference is ΔE_p. The conduction band edge, as well as the Fermi level of the metal of the right-side electrode, rises up for the same amount of ΔE_p. The electrons in the external circuit should flow from the right-hand side to the left-hand side to compensate the energy difference, which will generate the first output signal if the flow rate is sufficiently large, as indicated in Fig. 10b, e. The electrons cannot flow across the interface due to the presence of a Schottky barrier ϕ_{SB}. The accumulation of the electrons at the interface will partially screen the piezopotential built in the wire. When the electrons and the piezoelectric field reach equilibrium, the Fermi energy on both sides is on the same level, and there is no more current flow (Fig. 10c, e). When the SWG is fast released, there is no strain inside the NW. As a result, the polarization and piezoelectric field vanish as well, and the equilibrium with the accumulated electrons is broken. The accumulated charge carriers flow back from the left-hand side to the right-hand side through the external circuit, producing the second output signal in the opposite direction, as shown in Fig. 10d, e.

Integration of SWGs is a major step toward the practical applications. Taking advantage of the small size of the NW, multiple SWGs can be built on a single substrate such that the hamster

can drive all SWGs simultaneously. Integration of up to four SWGs in serial has been demonstrated with the open-circuit voltage of ~0.1–0.15 V. The output current can also be improved with multiple SWGs in parallel. The work on the human finger and living hamster clearly demonstrated the capability of the SWG to scavenge mechanical energy from the biosystem. It also confirms the feasibility of using the SWG for harvesting the energy created by regular and irregular motions.

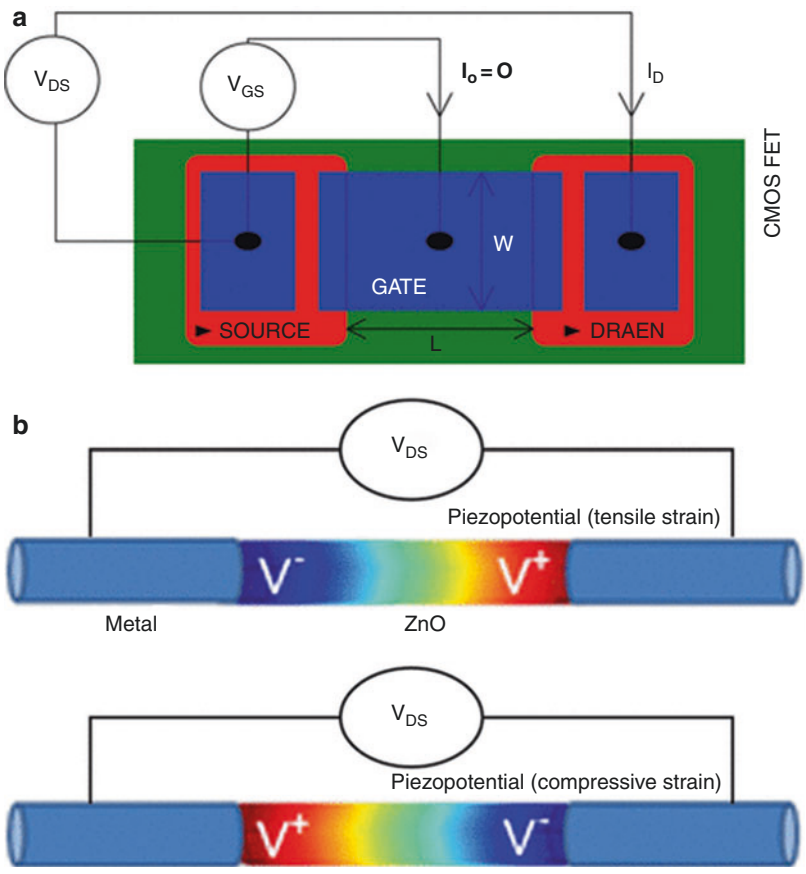
Piezotronics

A most simple FET is a two-end bonded semiconductor wire, in which the two electric contacts at the ends are the source and drain, and the gate voltage can be applied either at the top of the wire through a gate electrode or at its bottom on the substrate [8, 13]. When a ZnO NW is strained axially along its length, the piezoelectric potential continuously drops from one side of the NW to the other, which means that the electron energy continuously increases from the one side to the other. Meanwhile, the Fermi level will be flat all over the NW when equilibrium is achieved, since there is no external electrical field. As a result, the effective barrier height and/or width of the electron energy barrier between ZnO and metal electrode will be raised at one side and lowered at the other side; thus, it has a nonsymmetric effect on the source and drain. This is the *piezotronic effect* [11].

Once a strain is created in the semiconductor that also has piezoelectric property, a negative piezopotential at the semiconductor side effectively increases the local SB height, while a positive piezopotential reduces the barrier height. The polarity of the piezopotential is dictated by the direction of the *c*-axis for ZnO. The role played by the piezopotential is to effectively change the local contact characteristics through an internal field; thus, the charge carrier transport process is tuned/gated at the metal-semiconductor (M-S) contact. Figure 11 compares the piezopotential-gated FET and the traditional CMOS FET, which are different only by replacing the gate voltage by the inner-crystal piezopotential. With considering the change in piezopotential polarity by switching the strain from tensile to compressive, the local contact characteristics can be tuned and controlled by the magnitude of the strain and the sign of strain. This is the core of piezotronics.

The fundamental working principle of the p-n junction and the Schottky contact is that there is an effective barrier that prevents the charge carriers of both sides from crossing. The height and width of the barrier are the characteristic of the device. In piezotronics, the role played by the piezopotential is to effectively change the width of p-n junction or height of Schottky barrier (SB) by piezoelectricity, which has applications in maximizing the performance of the solar cell, photon detector, and LED by tuning the height of the local SB for achieving the optimum charge separation.

A simple piezotronic device is a polarity switchable diode that is made of a ZnO NW contacted with metal contacts at the two ends on an insulating polymer substrate. From the initial I-V curve measured from the device before applying a strain as shown in Fig. 12, the symmetric shape of the curve indicates that the SBs present at the two contacts are about equal heights. The equivalent circuit model of the device is a pair of back-to-back Schottky diodes, as illustrated in the inset in Fig. 12. Under tensile strain, the piezoelectric potential at the right-hand side of this NW was lower (denoted by blue color in the inset in Fig. 12), which raised the local barrier height (denoted by a large diode symbol in the inset). Since the positive piezoelectric potential was partially screened by free electrons, the SB height at the left-hand side remained almost unchanged. As a result, under positive bias voltage with the left-hand side positive, the current transport was determined by the reverse-biased SB at the right-hand side. While under the reverse-biased voltage with the right-hand side positive, the current transport depended on the reverse-biased SB at the left-hand side, which had a much lower barrier height than the right-hand one. Experimentally, the device thus exhibited a rectifying behavior in the positive voltage region, and the *I-V* curve in the negative voltage region overlapped with that of the original curve without straining. By the same token, under compressive strain, the device



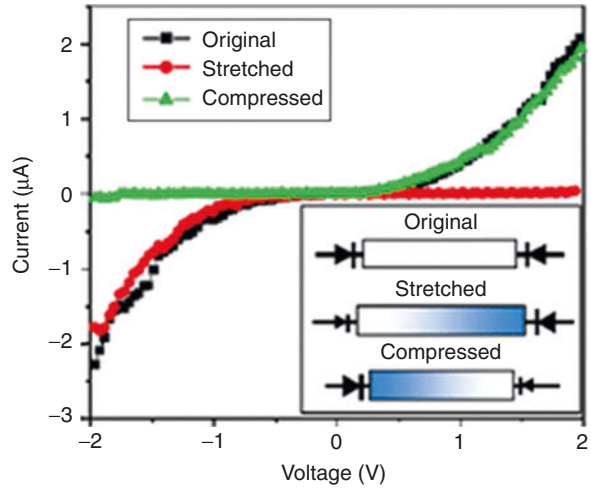
Piezoelectric Effect at Nanoscale, Fig. 11 A comparison of a CMOS FET and a piezotronic FET. The key difference is that the externally applied gate voltage V_{GS} is replaced by the stain-generated piezopotential, which tunes the charge transport insets are the enlarged view of the boxed area for one cycle of deformation. This is the core of piezotronics

exhibited a rectifying behavior in the negative voltage region, and the *I-V* curve in the positive voltage region overlapped with that of the original curve without straining, as shown by the green line in Fig. 12.

Piezotronic Logic Operations

The existing semiconductor NW logic devices are based on electrically gated field-effect transistors (FETs), which function as both the drivers and the active loads of the logic units by adjusting the conducting channel width [17]. Moreover, the currently existing logic units are "static" and are almost completely triggered or agitated by electric signals, while the "dynamic" movable mechanical actuation is carried out by another unit possibly made of different materials. It is highly desirable to integrate the static logic operations with the dynamic mechanical action into one unit utilizing a single material. The first piezoelectric-triggered mechanical-electronic logic operation has been demonstrated using the piezotronic effect, through which the integrated mechanical actuation and electronic logic computation are achieved using only ZnO NWs.

A strain-gated transistor (SGT) is made of a single ZnO NW with its two ends, which are the source and drain electrodes, being fixed by silver paste on a polymer substrate (Fig. 13a). Once the substrate is bent, a tensile/compressive strain is created in the NW since the mechanical behavior of the entire structure is determined by the substrate. Utilizing the piezopotential created inside the NW, the gate input for a NW SGT is an externally applied strain rather than an electrical signal. I_{DS}-V_{DS} characteristic for each single ZnO NW SGT is obtained as a function of the strain created in the SGT (Fig. 13a) before further assembly into logic devices. A NW SGT is defined as forward biased, if the applied bias is

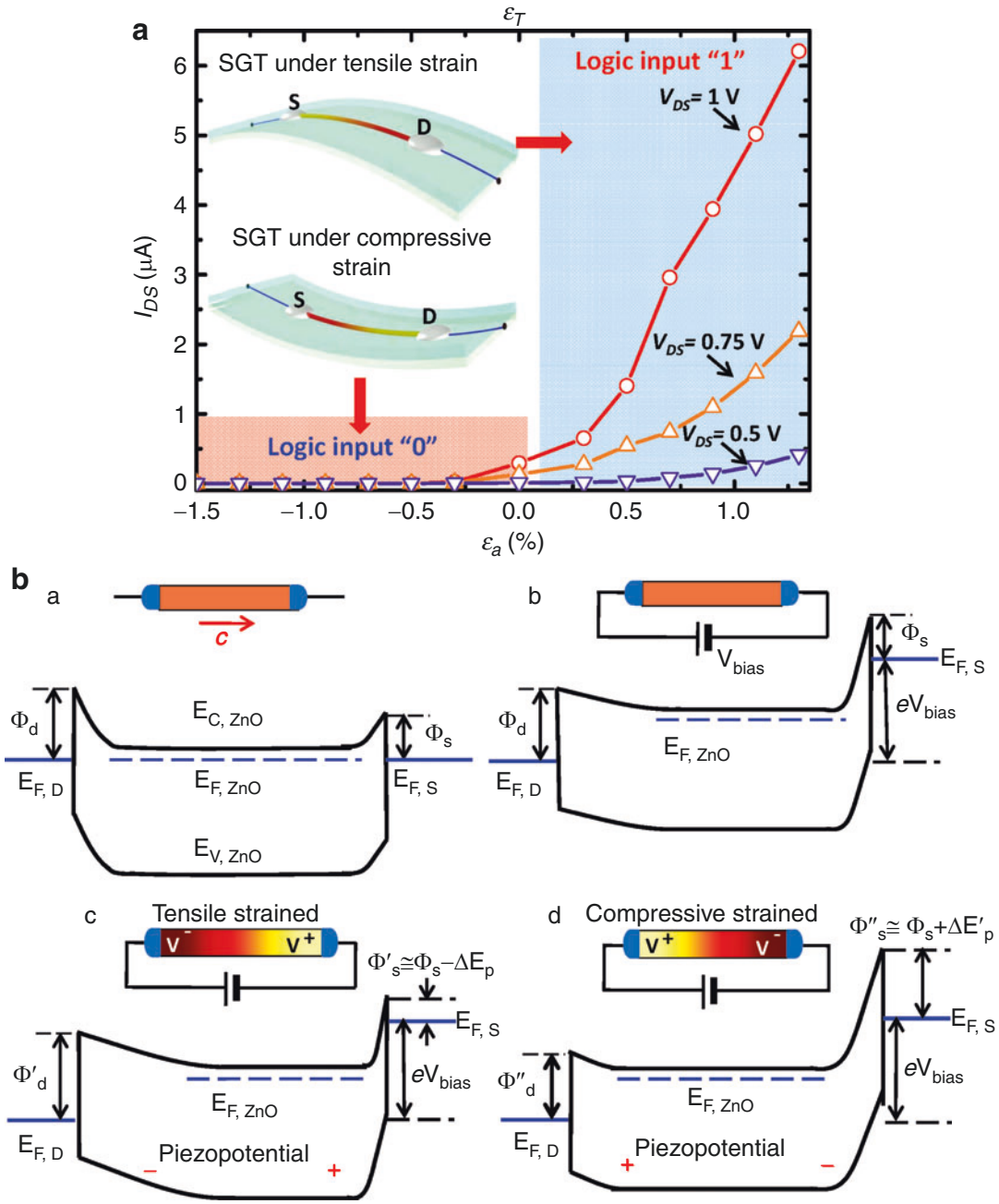


Piezoelectric Effect at Nanoscale, Fig. 12 Piezotronic strain sensor/switch. Changes of transport characteristics of an Ag/ZnO-nanowire/Ag device from symmetric *I-V* characteristic (*black*) to asymmetric rectifying behavior when stretching (*red*) and compressing (*green*) the wire. The inset is the equivalent circuit models of the device in corresponding to the observed *I-V* curves; different sizes of diode symbol are used to illustrate the asymmetric Schottky contacts at the two ends of the nanowire. The *blue* side is the negative potential side, and the other side is the positive side (Ref. [22] of [12], Adapted with permission from "Zhou, J., Gu, Y.D., Fei, P., Mai, W.J., Gao, Y.F., Yang, R.S., Bao, G., Wang, Z.L.: Flexible piezotronic strain sensor. Nano Lett. **8**, 3035–3040 (2008)." Copyright 2008 American Chemical Society)

connected to the drain electrode (Fig. 13a). For a SGT, the external mechanical perturbation-induced strain (ε_g) acts as the gate input for controlling the "on"/"off" state of the NW SGT. The positive/negative strain is created when the NW is stretched/compressed. The SGT behaves in a similar way to an *n*-channel enhancement-mode MOSFET, apparently indicating the working principle of the SGT.

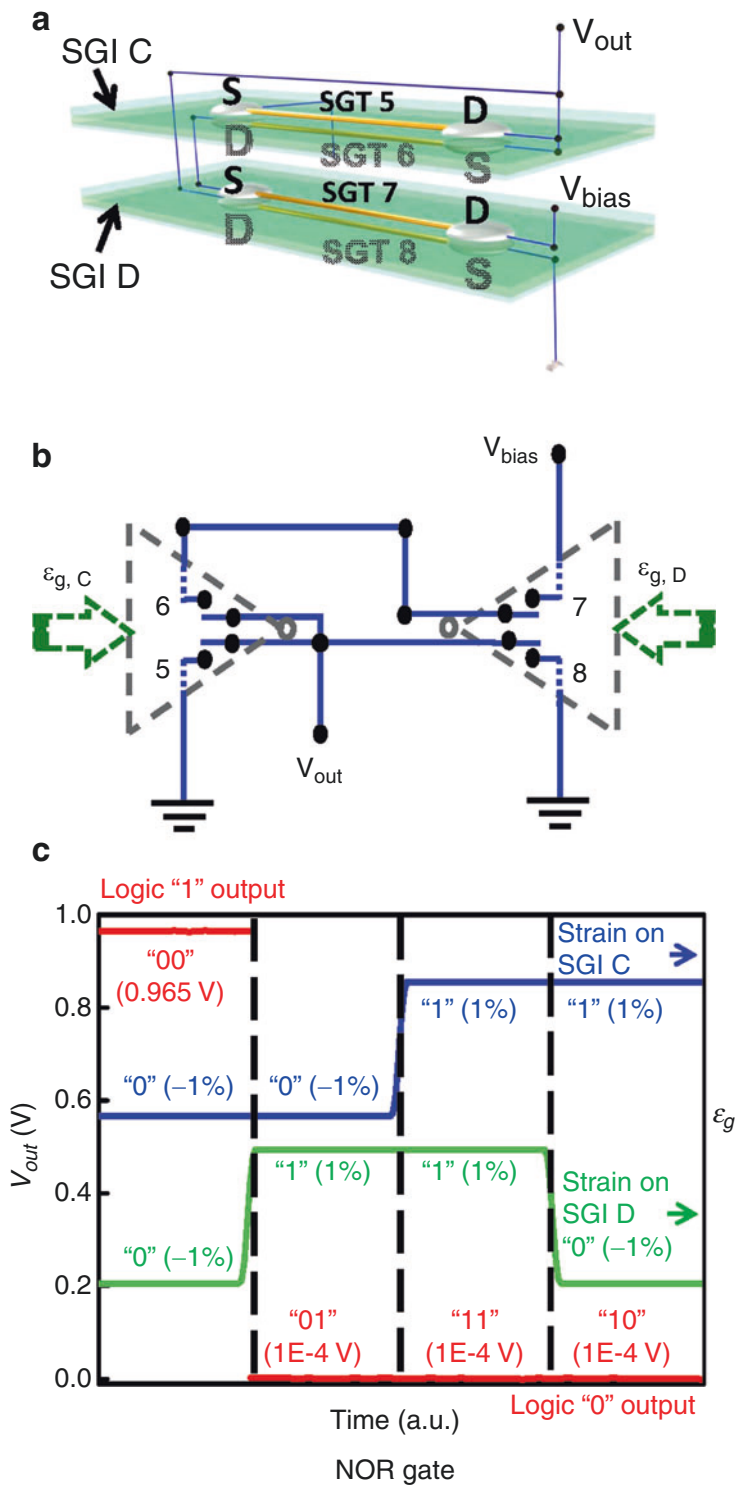
The working principle of a SGT is illustrated by the band structure of the device. A strain-free ZnO NW may have Schottky contacts at the two ends with the source and drain electrodes but with different barrier heights of Φ_S and Φ_D, respectively (Fig. 13b-a). When the drain is forward biased, the quasi-Fermi levels at the source ($E_{F,S}$) and drain ($E_{F,D}$) are different by the value of eV_{bias}, where V_{bias} is the applied bias (Fig. 13b-b). An externally applied mechanical strain (ε_g) results in both band structure change and piezoelectric potential field in a ZnO NW. The former leads to the piezoresistance effect, which is a nonpolar and symmetric effect at the both source and drain contacts. Since ZnO is a polar structure along *c*-axis, straining in axial direction (*c*-axis) creates a polarization of cations and anions in the NW growth direction, resulting in a piezopotential drop from V^+ to V^- along the NW, which produces an asymmetric effect on the changes in the SB heights (SBHs) at the drain and source electrodes. Under tensile strain, the SBH at the source side reduces from Φ_S to $\Phi'_S \cong \Phi_S - \Delta E_P$ (Fig. 13b-c), where ΔE_P denotes the effect from the locally created piezopotential and it is a function of the strain, resulting in increased I_{DS}. For the compressively strained SGT, the sign of the piezopotential is reversed; thus, the SBH at the source side is raised from Φ_S to $\Phi''_S \cong \Phi_S + \Delta E'_P$ (Fig. 13b-d), where $\Delta E'_P$ denotes the piezopotential effect on the SBH at source side, resulting in a large decrease in I_{DS}. Therefore, as the strain ε_g is swept from compressive to tensile regions, the I_{DS} current can be effectively turned from "on" to "off," while V_{DS} remains constant. This is the fundamental principle of the SGT.

Logic operations of NW strain-gated NOR gate have been realized by integrating SGTs, which are gated individually by the applied strains, according to corresponding layout rules



Piezoelectric Effect at Nanoscale, Fig. 13 Strain-gated transistor (SGT). (**a**) Current – strain (I_{DS}–ε_g) transfer characteristic for a ZnO SGT device with strain sweeping from $\varepsilon_g = -0.53$ to 1.31 % at a step of 0.2 %, where the V_{DS} bias values were 1, 0.75, and 0.5 V, respectively; (**b**) the band structures of the ZnO NW SGT under different conditions for illustrating the mechanism of SGT. The crystallographic *c*-axis of the nanowire directs from drain to source. (*a*) The band structure of a strain-free ZnO NW SGT at equilibrium with different barrier heights of Φ_S and Φ_D at the source and drain electrodes, respectively; (*b*) the quasi-Fermi levels at the source ($E_{F,S}$) and drain ($E_{F,D}$) of the ZnO SGT are split by the applied bias voltage V_{bias}; (*c*) with tensile strain applied, the SBH at the source side is reduced from Φ_S to $\Phi'_S \cong \Phi_S - \Delta E_P$; (*d*) with compressive strain applied, the SBH at the source side is raised from Φ_S to $\Phi''_S S \cong \Phi_S + \Delta E'_P$ (Adapted with permission from Ref. [29] of [12] "Wu, W.Z., Wei, Y.G., Wang, Z.L.: Strain-gated piezotronic logic nanodevices. Adv. Mater. **22**, 4711–4715 (2010)." Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission)

P



Piezoelectric Effect at Nanoscale,

Fig. 14 Strain-gated piezotronic logic NOR gate. (**a**) Schematic of the ZnO NW strain-gated NOR logic gate, which is composed of two SGIs, SGI C, and SGI D. The strain input C for SGI C is defined in reference to the strain applied to SGT 5, and the strain input D for SGI D is defined in reference to the strain applied to SGT 8, (**b**) layout for ZnO NW strain-gated NOR logic gate connecting two ZnO NW SGIs, (**c**) logic operations and experimental truth table of the ZnO NW strain-gated NOR logic gate. *Red line* is the electrical output of the NOR gate. *Blue* and *green lines* represent the strain inputs applied on SGI C and SGI D, respectively (Adapted with permission from Ref. [29] of [12] "Wu, W.Z., Wei, Y.G., Wang, Z.L.: Strain-gated piezotronic logic nanodevices. Adv. Mater. **22**, 4711–4715 (2010)." Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission)

(Figs. 14a, b). The output voltages of NOR gate versus the input gate strains are shown in Fig. 14c. It can also be seen that NW strain-gated NOR gate with active loads exhibit better overall performance, such as larger logic swing, compared to passive-load NOR gate. Using the SGTs as building blocks, other universal logic components such as inverters and NAND and XOR gates have been demonstrated for performing piezotronic logic calculations, which have the potential to be integrated with the current NEMS technology for achieving advanced and complex functional actions in nanorobotics, microfluidics, and micro/nanosystems.

Conclusion

With reduction in material size, its mechanical robustness, elastic modulus, and toughness improve dramatically. It can easily go way beyond the nonlinear but still elastic mechanical deformation without causing plastic damage. The piezoelectric effect at nanoscale is expected to be largely enhanced so that they can be effectively used for ultrasensitive sensors/actuators and energy converters.

The two most recent important advances in using the nanoscale piezoelectric effect have been introduced here. One is the nanogenerator that converts mechanical energy into electricity utilizing tiny physical motion. Besides its potential for large-scale applications such as harvesting ocean wave and wind energy, it finds application in driving wireless nanodevices and nanosystems for medical science, environmental/infrastructure monitoring, defense technology, and even personal electronics. It is a key step for building self-powered nanotechnology.

The second important concept in science is the piezotronics. Electronics fabricated by using inner-crystal piezopotential as a "gate" voltage to tune/control the charge transport behavior is named *piezotronics*. *Piezo-phototronic effect* is a result of three-way coupling among piezoelectricity, photonic excitation, and semiconductor transport, which allows tuning and controlling of electro-optical processes by strain-induced piezopotential. These designs are different from the CMOS field-effect transistor by replacing the externally applied gate voltage using a crystal generated inner potential (piezopotential) for controlling the charge transport. It is a transistor controlled by a force/strain. It will find applications in human-computer interfacing, mechanical-triggered electronics, nanomachine, electronic signature, medical diagnostics, and sensor.

Cross-References

- Ab Initio DFT Simulations of Nanostructures
- AFM Probes
- Biosensors
- Carbon Nanotube-Metal Contact
- CMOS MEMS biosensors
- Electrospinning
- Finite Element Methods for Computational Nano-Optics
- Flexible Electronics
- Hybrid Solar Cells
- Nanorobotics
- Nanomaterials for Excitonic Solar Cells
- Nanostructures for Energy
- Nanotechnology
- NEMS Piezoelectric Switches
- Piezoelectric MEMS Switch
- Sol–gel Method

References

1. Cady, W.G.: Piezoelectricity: An Introduction to the Theory and Applications of Electromechanical Phenomena in Crystals. Dover, New York (1964)
2. Haertling, G.H.: Ferroelectric ceramics: history and technology. J. Am. Ceram. Soc. **82**(4), 797–818 (1999)
3. Bhalla, A.S., Guo, R., Roy, R.: The perovskite structure – a review of its role in ceramic science and technology. Mater. Res. Innov. **4**, 3–26 (2000)
4. Johnsson, M., Lemmens, P.: Crystallography and chemistry of perovskites. In: Kronmüller, H. (ed.) Handbook of Magnetism and Advanced Magnetic Media. Wiley, New York (2006)
5. Gruverman, A., Kholkin, A., Ramesh, R.: Nanoscale ferroelectrics: processing, characterization and future trends. Rep. Prog. Phys. **69**, 2443–2474 (2006)

6. Xu, S., Hansen, B.J., Wang, Z.L.: Piezoelectric-nanowire-enabled power source for driving wireless microelectronics. Nat. Commun. **1**, 93 (2010)
7. Turner, R.C., Fuierer, P.A., Newnham, R.E., Shrout, T. R.: Materials for high temperature acoustic and vibration sensors: a review. Appl. Acoust. **41**, 299–324 (1994)
8. Wang, Z.L.: Nanopiezotronics. Adv. Mater. **19**, 889–892 (2007)
9. Murayama, N., Nakamura, K., Obara, H., Segawa, M.: The strong piezoelectricity in polyvinylidene fluoride (PVDF). Ultrasonics **14**, 15–24 (1976)
10. Yang, J.S.: Introduction to the Theory of Piezoelectricity. Springer, New York (2005)
11. Wang, Z.L.: Towards self-powered nanosystems: from nanogenerators to nanopiezotronics. Adv. Funct. Mater. **18**, 3553–3567 (2008)
12. Wang, Z.L.: Piezopotential gated nanowire devices: piezotronics and piezo-phototronics. Nano Today **5**(6), 540–552 (2010)
13. Yan, Z., Hu, Y.F., Xiang, S., Wang, Z.L.: Effects of piezopotential spatial distribution on local contact dictated transport property of ZnO micro/nanowires. Appl. Phys. Lett. **97**, 033509 (2010)
14. Bühlmann, S., Dwir, B., Baborowski, J., Muralt, P.: Size effect in mesoscopic epitaxial ferroelectric structures: increase of piezoelectric response with decreasing feature size. Appl. Phys. Lett. **80**, 3195–3197 (2002)
15. Wang, Z.L.: Self-powering nanotech. Sci. Am. **298**, 82–87 (2008)
16. Wang, Z.L., Song, J.H.: Piezoelectric nanogenerators based on zinc oxide nanowire arrays. Science **312**, 242–246 (2006)
17. Thorsen, T., Maerkl, S.J., Quake, S.R.: Microfluidic large-scale integration. Science **298**, 580–584 (2002)

Piezoelectric MEMS Switch

R. G. Polcawich, J. S. Pulskamp and R. M. Proie
US Army Research Laboratory, Adelphi, MD, USA

Synonyms

Piezoelectric Switches

Definition

A piezoelectric switch is an electromechanical relay or switch relying on piezoelectric actuation to provide the force necessary to close or produce a gap between electrical contacts. In practical use, these switches can be designed for systems operating with carrier frequencies from DC to tens or hundreds of GHz.

Overview

The routing of electrical signals is controlled by switches of various types. For mechanical switches, a number of transduction techniques have been utilized including electrostatic, electromagnetic, thermomechanical, and piezoelectric. Fundamental to most radio frequency (RF) circuits, a switch is used not only to control the path of electrical circuits but also the phase and timing of the circuits. The continuous miniaturization of military radar and communication systems requires development of smaller, more energy-efficient switches for continuous control of a wide variety of electronic signals [1]. Additionally, switches are required for ever-increasing frequencies, making existing solid-state technologies less attractive to the system designer because of larger power losses and decreases in linearity as the frequency increases [2]. In short, microelectromechanical systems (MEMS) switches provide a very low-loss, low-power, and highly linear, with respect to input power, alternative to existing solid-state switch technologies. Moreover, they are significantly smaller than macro-switch technologies. Unlike electronic switches, PIN diodes and field-effect transistors (FET), a well-designed MEMS switch is extremely linear with respect to input power and does not create distortion, such as harmonics or intermodulation products, in the electrical signal [2]. Next-generation missile seeker radar, smart cellular phone technologies, and communication-on-the-move systems require highly linear, low-power, broadband RF switches as a fundamental building block [1, 3].

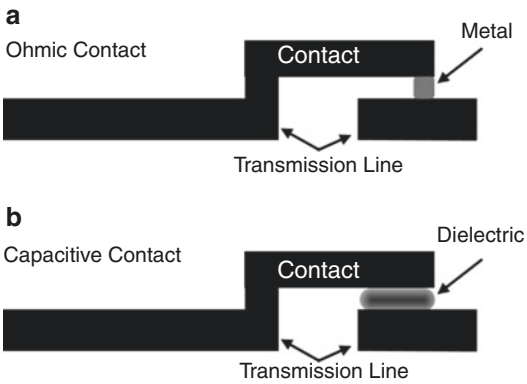
History and Brief Background of MEMS Switch Development

The development of MEMS switches began with the effort of Petersen at IMB in 1979 [4] and then continued with Hughes Research Laboratories on

micromachined microwave actuators (MIMACs) that produced the first switch capable of operating up to 50 GHz [5]. Following this research, a concentrated effort began in developing switches in both series and shunt configurations for a wide variety of applications, including lower-frequency commercial cell phones and high-frequency military radar and communication systems [1–3].

All MEMS switches have a few basic components for their operation. This includes the signal line, switch contact(s), and actuator(s). The signal line can be designed for use with signals ranging from DC to sub-THz. For low-frequency operation, a simple electrical trace commonly comprised of gold or other conductive metal can be utilized. For higher-frequency operation, the signal line requires greater design consideration. For MEMS switches, the signal line is commonly configured using coplanar waveguide or microstrip transmission line geometries.

MEMS switches are defined by two different types of contacts: ohmic and capacitive (see Fig. 1). Capacitive contact switches utilize a thin dielectric film between two metals with adjustable gap. At elevated frequencies, the RF energy is capacitively coupled through the dielectric to the switch contact. As a result, capacitive switches are generally limited to higher-frequency operation (>6 GHz) where capacitive coupling is feasible with reasonably sized capacitive contacts. In contrast, ohmic contacts use two metals to create the switch contact. Ohmic switches provide the opportunity for very broadband operation using carrier frequencies from DC on upward. It should be noted that the type of contact material is crucial for high-performance switches with long lifetime. Although beyond the scope of this entry, dielectric materials with low charge-trapping densities are preferred for capacitive contacts (e.g., silicon nitride and alumina), and metals with high hardness and low probability of creating frictional polymers are preferred for ohmic contacts (e.g., ruthenium and ruthenium oxide).

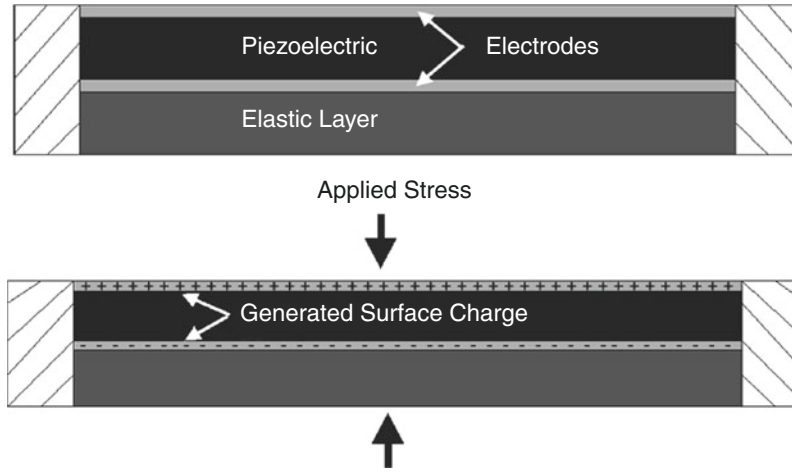


Piezoelectric MEMS Switch, Fig. 1 Schematic illustration of (**a**) ohmic and (**b**) capacitive switch contacts

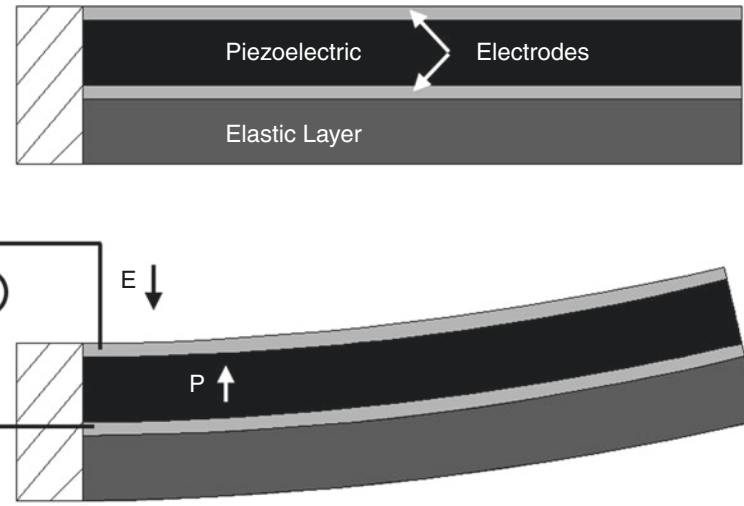
The final critical component is the actuator. Historically, most reported MEMS switches have utilized electrostatic actuators in which a metal beam is suspended over both a bias electrode and one of the switch contacts [2]. The remainder of this entry, however, will focus on an alternative actuation mechanism, piezoelectricity, and the piezoelectric actuators' role in MEMS switch technologies. When compared with switches actuated by electrostatic force, one critical advantage of piezoelectric switches is the full control of their displacement when the switch is making contact. In contrast to piezoelectric switches, electrostatic switches experience a "pull-down voltage" wherein there is an instability in the gap between the actuator and its bias electrode. Once the pull-down voltage is reached and the gap of the electrostatic capacitor is reduced to 2/3 of the gap at 0 V, a positive feedback is created in the device. A smaller gap creates larger closing force, which in turn moves the switch in the direction of even smaller gaps. This process of constant acceleration ends when the contact is closed against a typically immovable transmission line. In contrast, piezoelectric MEMS switches move only in response to the magnitude of the DC actuation voltage. They do not exhibit positive feedback and their dynamic behavior is well controlled. As a result, switches with piezoelectric actuation can be operated where the contacts close with only a modest acceleration resulting in minimal contact bounce, thus improving switch cycle reliability [6].

Piezoelectricity

Piezoelectricity is comprised of two electromechanical responses, a direct effect and a converse



Piezoelectric MEMS Switch, Fig. 2 Illustration of the direct piezoelectric effect applied to a piezoelectric unimorph membrane



Piezoelectric MEMS Switch, Fig. 3 Illustration of the converse piezoelectric effect applied to a unimorph cantilever with the vertical scale exaggerated

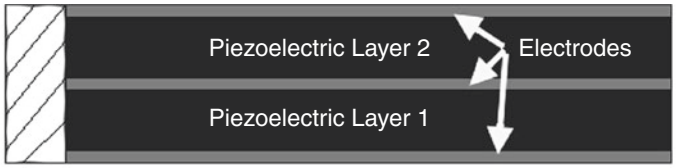
effect. The direct effect refers to the generation of charge under the application of an applied stress and is governed by Eq. 1 where the induced electric displacement (D_i) is proportional to the piezoelectric coefficient (d_{ijk}) and the applied stress (T_{jk}). The converse effect refers to the generation of strain under an applied electric field and is governed by Eq. 2, where the induced strain (x_{jk}) is proportional to the piezoelectric coefficient and electric field (E_i).

$$D_i = d_{ijk} T_{jk} \tag{1}$$

$$x_{jk} = d_{ijk} E_i \tag{2}$$

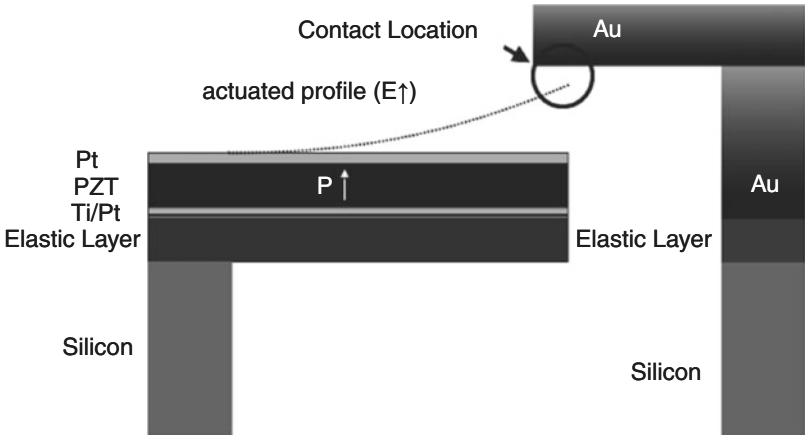
In application, the direct effect can be used for sensing and energy harvesting in which the applied stresses are used to generate surface charges on the piezoelectric (see Fig. 2). For actuators, an applied electric field generates a strain in the piezoelectric layer producing flexure in unimorph actuators (see Fig. 3) and bimorph actuators (see Fig. 4). In switch applications, the piezoelectric unimorph can be combined with overhanging contact cantilevers to create a functioning switch (see Fig. 5).

The tensor notation for the aforementioned relationships can be simplified using matrix notation where the subscripts i, j, and k having





Piezoelectric MEMS Switch, Fig. 4 Illustration of the converse piezoelectric effect applied to a bimorph cantilever with the vertical scale exaggerated



Piezoelectric MEMS Switch, Fig. 5 Cross-section schematic of a PZT unimorph and electrical contacts configured for use as switch

values 1–3 are replaced with the subscripts p and q with values of 1–6 [7]. As a result, d_{ijk} is transformed to d_{ip}. The d_{ip} (strain) and e_{iq} (stress) piezoelectric coefficients are the most useful for thin-film sensor and actuator applications and are related by the elastic constant, c_{pq}^{E}, in Eq. 4.

$$e_{iq} = d_{ip} c_{pq} \tag{3}$$

For thin-film composites where the piezoelectric layer is rigidly attached to a supporting elastic layer (i.e., thin film or substrate), the out-of-plane stress, X_3, is generally set to zero while the longitudinal strain, x_3, varies from a combination of the piezoelectric effect and elastic coupling from the Poisson effect. As a result, an effective thin-film piezoelectric coefficient, $e_{31,\,f}$, can be related to the common piezoelectric coefficients, d_{31} and e_{31} (see Eq. 5), by compensating for the clamping effect experienced by piezoelectric thin films in composite structures [8, 9].

$$e_{31,f} = e_{31} - \frac{c_{31}^{E} e_{33}}{c_{33}^{E}} = \frac{d_{31}}{\left(s_{11}^{E} + s_{12}^{E}\right)} \tag{4}$$

Piezoelectric Materials: Ferroelectrics and Non-ferroelectrics

The important piezoelectric and dielectric material properties for the three most common thin-film piezoelectrics, ZnO, AlN, and lead zirconate titanate (PZT), are listed in Table 1. A few

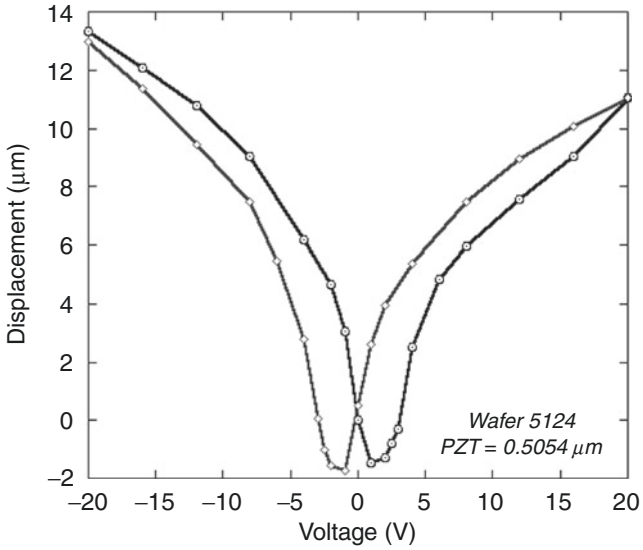
Piezoelectric MEMS Switch, Table 1 Comparison of thin-film piezoelectric materials

	ZnO	AlN	$Pb(Zr_xTi_1\text{-}x)O_3$
Crystal structure	Wurtzite	Wurtzite	Perovskite
$e_{31,f}$ (C/m^2)	$-0.4 \rightarrow -0.8$	$-0.9 \rightarrow -1.1$	$-8 \rightarrow -18$
$d_{33,f}$ (pC/N)	$10 \rightarrow 17$	$3.4 \rightarrow 6.5$	$90 \rightarrow 150$
ε_{33}	$8 \rightarrow 12$	$10.1 \rightarrow 10.7$	$800 \rightarrow 1200$
Tan δ			$0.02 \rightarrow 0.05$
Density (g/cm^3)	5.68	3.26	$7.5 \rightarrow 7.6$
Young's modulus (GPa)	$110 \rightarrow 150$	$260 \rightarrow 380$	$60 \rightarrow 80$
Acoustic velocity (m/s)	6.07×10^3	11.4×10^3	2.7×10^3
References	[11–14]	[11, 12, 15, 16]	[11, 12]

distinctive features to recognize are the differences in crystal structure and the effective thin-film piezoelectric coefficient, $e_{31,f}$, of these three compounds. The difference in crystal structures can be summarized as a distinction between purely polar materials and ferroelectric materials. Pure polar materials such as AlN and ZnO do not permit the polarization vector to change orientations relative to the crystal lattice. The polar direction is crystallographically defined along the c-axis of the wurtzite crystal structure. In contrast, ferroelectric materials, such as PZT, have a spontaneous electric dipole moment that can be reoriented between crystallographically defined stable states by a realizable electric field (i.e., before dielectric breakdown occurs) [10]. As a result, extremely large polarization values and subsequently large dielectric constants and piezoelectric coefficients can be achieved in ferroelectric materials.

Another important characteristic distinction between purely polar materials and ferroelectric materials is the use of the constitutive equations of piezoelectricity, Eqs. 1 and 2, which describe the linear response of piezoelectric materials. These, generally, can be applied successfully to polar piezoelectric thin films like AlN and ZnO in most circumstances. However, ferroelectrics, like PZT, can display pronounced stress and electric field induced nonlinear material responses. This is particularly true for thin films and actuators. Thin films permit the application of extremely large electric fields with modest operating voltages due to the film thicknesses involved. Piezoelectric MEMS actuator applications typically drive the material well beyond the coercive field. For a half-micron thin film of PZT (52/48), for example, the coercive field corresponds to an applied voltage of approximately 2.5 V. Ferroelectrics even show significant nonlinearity below the coercive field [17]. Despite this fact it is surprising that, for both bulk material and thin-film applications, the linear piezoelectric equations are often used to describe the device response in ferroelectric materials. To accurately model ferroelectric MEMS device response, under high operating field (actuators) and/or bias field (sensors) conditions, these significant nonlinearities must be taken into account.

At the high fields encountered in MEMS actuators, the total strain response is due to the combination of the linear response and nonlinear piezoelectricity including saturation effects, domain wall motion, and electrostriction. All dielectrics display the property of electrostriction, whereby externally applied electric fields induce a strain response that is proportional to the square of the field strength. Equation 5 describes the total strain response in ferroelectric materials where Q is the appropriate electrostrictive coefficient, P_s is the spontaneous polarization, E is the applied electric field, and κ is the appropriate dielectric constant. The first term defines the remnant strain. The second term is equal to the linear piezoelectric coefficient (d_{ij}) multiplied by the electric field. The last term is the strain due to electrostriction. As the applied field strength increases, the electrostrictive term contributes more to the overall strain response in the material.



Piezoelectric MEMS Switch,
Fig. 6 Displacement versus voltage plot for a PZT thin-film cantilever comprised of a 0.5 μm elastic layer ($SiO_2/Si_3N_4/SiO_2$)/0.1 μm Ti/Pt bilayer, 0.505 μm PZT (52/48), and a 0.1 μm Pt top electrode. Key aspects of this diagram include the hysteretic displacement characteristics upon bipolar voltage swings. Additionally, all but a small fraction of the displacement occurs in one direction

$$x = QP_s^2 + 2\varepsilon_o \kappa QP_s E + Q(\varepsilon_o \kappa E)^2 \qquad (5)$$

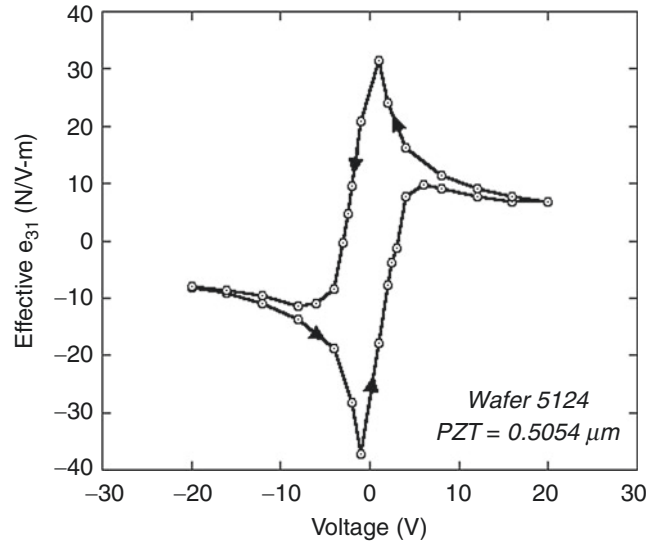
At large applied stresses and electric fields, significant contributions to the total strain response (Eq. 5) are due to nonlinear piezoelectricity and the extrinsic response. These nonlinear material properties used for design purposes are best measured for specific materials and processing conditions. Due to these complex nonlinear effects, the ratio of total strain to applied electric field is referred to as the effective electroactive coefficients. These coefficients can be used with the models described above using field-dependant functions.

Perhaps the most important consequence of these nonlinear effects is the absence of a large bipolar strain response when operated above the coercive field. Figure 6 shows voltage-displacement data from a simple unimorph PZT MEMS-actuated cantilever, illustrating the so-called butterfly loop. The plot was obtained by sweeping the actuation voltage up to a positive voltage, then reducing to zero, then sweeping down to a negative voltage, and finally returning to zero field. The extracted effective electroactive piezoelectric stress constant, e_{31eff}, corresponds to the displacement data in Fig. 6 and is illustrated in Fig. 7. This effective piezoelectric constant includes the small signal piezoelectric effect, electrostriction, and strain induced by large polarization charges associated with overcoming a pinched ferroelectric hysteresis loop.

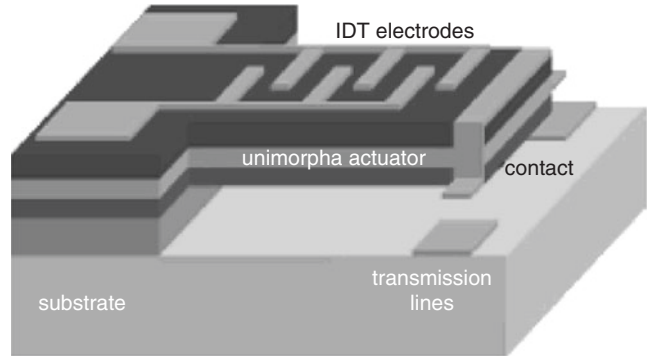
The response of the device is seen to be bipolar only in the regime between the positive and negative coercive voltages where the negative displacement reverts to a positive displacement (the "V" regions of the lower curves). Thus for large electric fields (i.e., in excess of the coercive field) applied to the piezoelectric material, the sense of the piezoelectric strain is independent on the applied field polarity, therefore only a single sense of the piezoelectric strain is possible. In terms of the device, the in-plane contraction of the piezoelectric material at large fields gives a negative sense to the piezoelectric actuation force. The standard configuration of the d_{31} mode MEMS unimorph, with the neutral axis below the midplane of the piezoelectric layer, gives a positive sense of the moment arm. Piezoelectric MEMS devices based on ferroelectric materials are most often operated under unipolar conditions and hence avoid much of the hysteresis observed in Fig. 6.

Piezoelectric Switch Actuator History

One of the key limitations of electrostatically actuated MEMS switches is the inability to



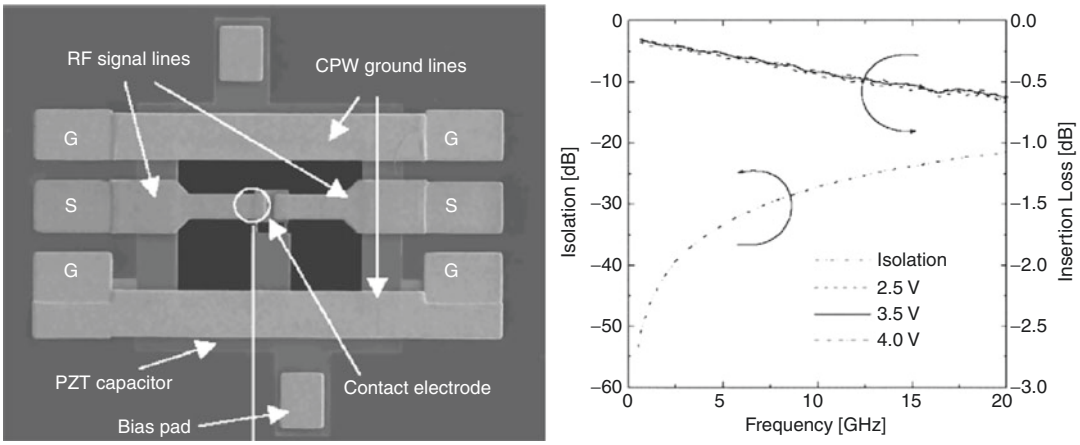
Piezoelectric MEMS Switch, Fig. 7 Effective piezoelectric stress constant, $e_{31,eff}$, versus voltage extracted from the displacement-voltage measurements for the PZT thin-film unimorph described in Fig. 6



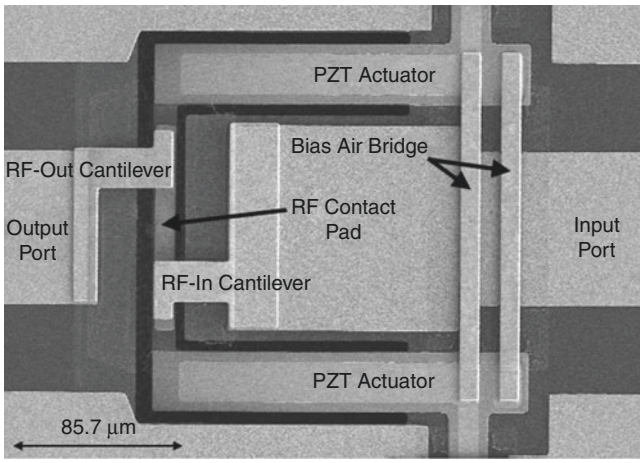
Piezoelectric MEMS Switch, Fig. 8 Illustration of a thin-film PZT unimorph switch developed at Penn State University [20] (Reprinted with permission. Copyright 2003 American Institute of Physics)

combine a low actuation voltage with excellent DC and RF performance (i.e., contact force and insertion loss, respectively) and reliability. Attempts to lower the actuation voltage with electrostatic switches generally rely on reducing either the mechanical stiffness of the released structures or the gap between the mechanical bridge/cantilever and the corresponding biasing pad [18]. Reductions in the stiffness can limit the restoring force of the switch and can lead to stiction failures [19]. Decreasing the electrode gap limits the high-frequency RF performance (in particular, the isolation in the open state for series switches) [2].

The first functioning piezoelectric MEMS DC relay was demonstrated by Gross et al. using in-plane poled PZT film actuators [20]. These devices used a cantilever beam with a supporting elastic layer, a PZT thin film as the actuator, and patterned gold structures to act as the electrodes (see Fig. 8). This device operated as low as 20 V with a switching on time as low as 2 μs. The first reported working RF MEMS switch using PZT thin films used bulk micromachining to release a unimorph actuator and a suspended transmission line [21]. This design utilized a cantilever that is perpendicular to the wave propagation direction along the RF conductor of the coplanar waveguide (CPW). The switch operated as low as 2.5 V with an insertion loss better than 0.7 dB and isolation better than −21 dB up to 20 GHz (see Fig. 9). The first surface micromachined approach for creating a PZT-actuated RF MEMS switch was developed by Polcawich



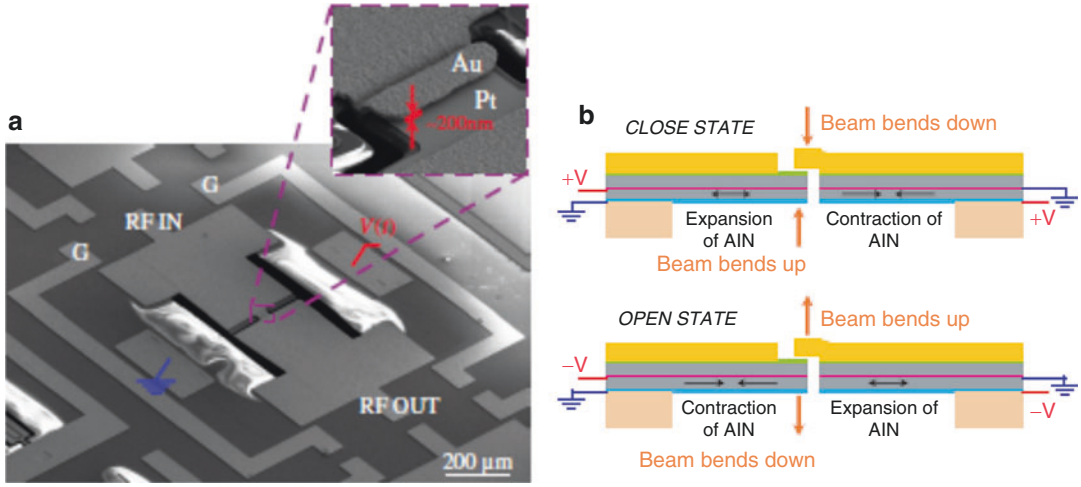
Piezoelectric MEMS Switch, Fig. 9 SEM image of a bulk micromachined RF MEMS series switch using a thin-film PZT actuator to complete the signal path [21] (Reprinted with permission. Copyright 2005 IEEE)



Piezoelectric MEMS Switch, Fig. 10 SEM image of a PZT-actuated RF MEMS switch [22] (Reprinted with permission. Copyright 2007 IEEE)

et al. [22]. These devices operated at less than 10 V and demonstrated temperature-stable performance from −25 °C to 100 °C for normally open series switches with isolation less than −20 dB and insertion loss of better than −1.0 dB up to 40 GHz [23] (Fig. 10).

The interest in radio and cell-phone frequency-compatible AlN-based filters has motivated the development of integrated piezoelectric switches and contour-mode filters both operated with AlN thin films. Mahameed et al. created the first AlN-based MEMS switches demonstrating good RF performance up to the low GHz regime [24]. These devices feature an innovative approach to address residual stress deformation, temperature sensitivity, and the modest piezoelectric coefficients of AlN. A symmetric "dual-beam" design features pairs of actuators that experience similar residual stress- and temperature-induced deformations that minimize variations in the open-state contact gap despite the mechanical fluctuations (see Fig. 11). The device shows immunity to residual stresses, fast switching speeds (~2 µs), moderate actuation voltage (~20 V), and good low-frequency isolation (>26 dB) and insertion loss (<0.7 dB) at 2 GHz. These devices have also been successfully co-fabricated on the same wafer as contour-mode AlN RF MEMS resonators. The device fabrication leverages the process used for AlN



Piezoelectric MEMS Switch, Fig. 11 (**a**) SEM image of AlN RF MEMS switch with inset indicating the nano-gap contact location and (**b**) illustration of principle of operation of "dual-beam" architecture [24] (Reprinted with permission. Copyright 2008 Institute of Physics)

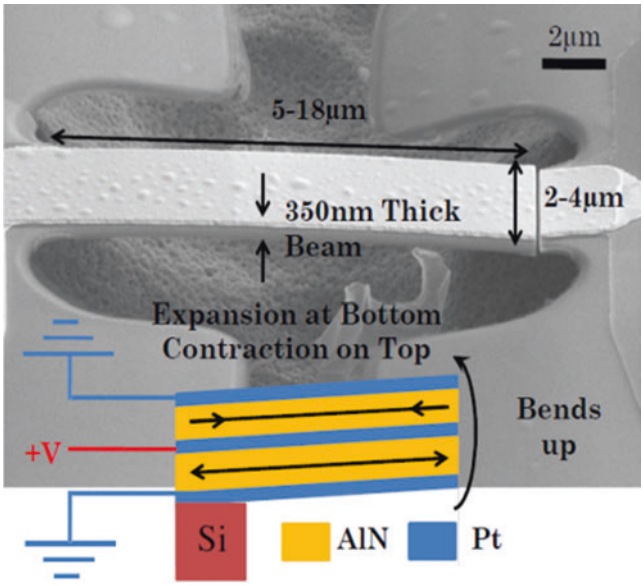
contour-mode resonators. Once these features are defined, including multiple AlN and Pt electrode deposition and patterning steps, an amorphous silicon sacrificial layer is deposited and patterned with liftoff. This is followed by the deposition and patterning of a thick-electroplated Au layer to provide RF transmission lines and unimorph actuator features. The switch is released using a xenon difluoride dry release to complete the process.

Switch Scaling

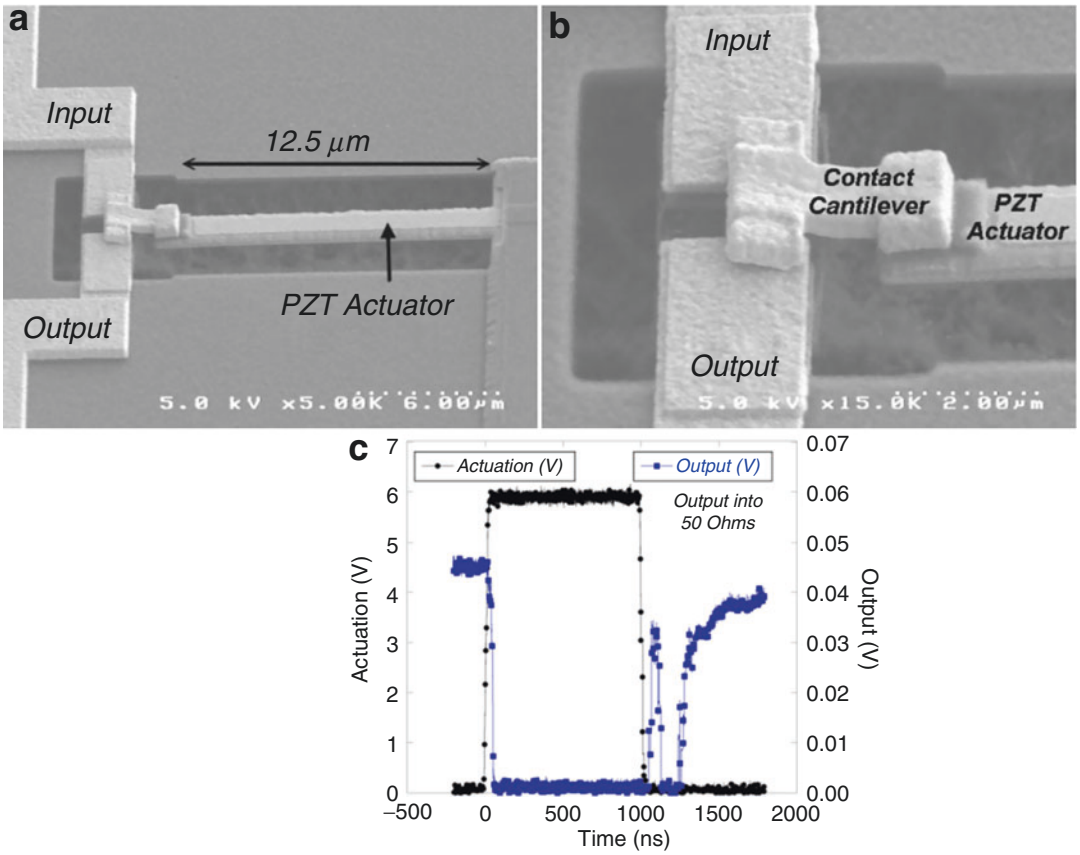
A number of groups are investigating mechanical logic technologies to address the limitations imposed in complementary metal-oxide-semiconductor (CMOS) transistors by gate oxide leakage. Static power consumption now rivals dynamic power consumption in CMOS and presents a significant challenge to the continued scaling of CMOS logic devices. Motivated by low-power nanomechanical logic applications, electromechanical relays based on either AlN or PZT have been investigated to provide practical solutions for low-power embedded microcontrollers. Switches have recently been demonstrated on both technologies using combinations of electron-beam lithography and focused ion beam milling [25–27].

The AlN-based devices consisted of a 100 nm AlN layer and were implemented in nanoscale bimorph cantilevered actuators similar to those described previously. The actuators were fabricated using a five-mask post-CMOS compatible process similar to the process outlined for the fabrication of the AlN RF MEMS switch discussed earlier. Displacements of 40 nm were achieved at 2 V in 18 µm long cantilever devices. A key difference in the process however is the use of a focused ion beam (FIB) tool. The nearly complete suspended clamped-clamped AlN beam is severed near one of the anchors with the FIB to create the cantilevered device seen in Fig. 12.

Similar to the scaled version of the AlN switch, a 3-D scaled version of the aforementioned PZT RF MEMS switch has been demonstrated. The switch shown in Fig. 13 was fabricated using an 11-step aligned electron-beam lithography process following the same process flow as its larger relative. The key differences were reductions in film thickness and gap along with reductions in lateral dimensions for the switch actuators and contact cantilever. In the normally closed configuration, the switch operated at 6 V demonstrated an opening response ranging from 17 to 33 ns with a corresponding closing response ranging from 33 to 43 ns.

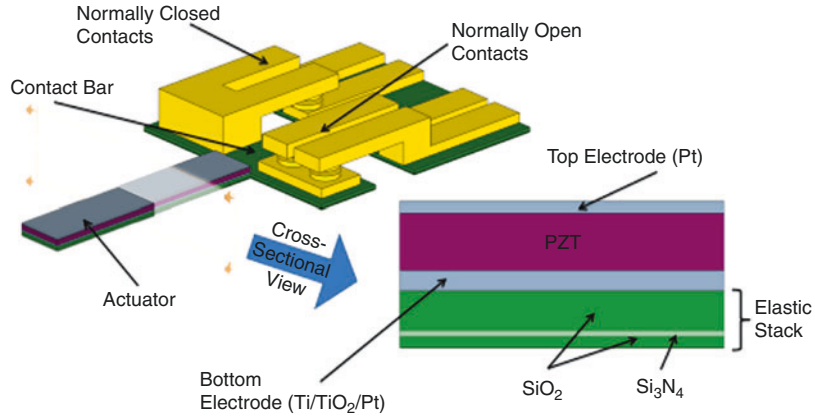


Piezoelectric MEMS Switch, Fig. 12 SEM image of AlN-cantilevered bimorph nano-actuator [25] (Reprinted with permission. Copyright 2008 IEEE)

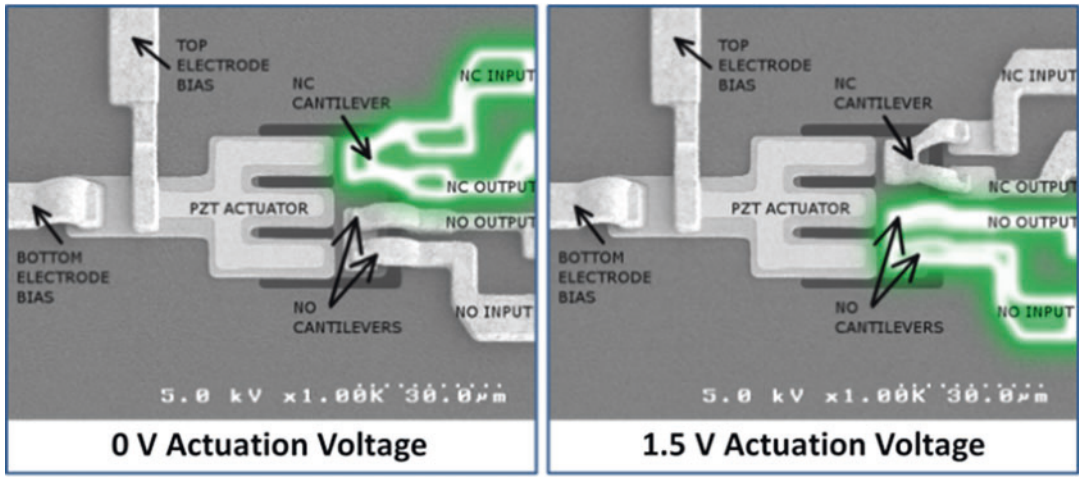


Piezoelectric MEMS Switch, Fig. 13 Images (**a** and **b**) of a PZT-nanoscaled relay fabricated from an 11-step electron-beam lithography process and (**c**) electrical output response during one cycle of operation [27] (Reprinted with permission. Copyright 2008 IEEE)

P



Piezoelectric MEMS Switch, Fig. 14 Single-pole double-throw relay schematic



Piezoelectric MEMS Switch, Fig. 15 Electron micrograph of a fabricated relay highlighting the two paths of current flow

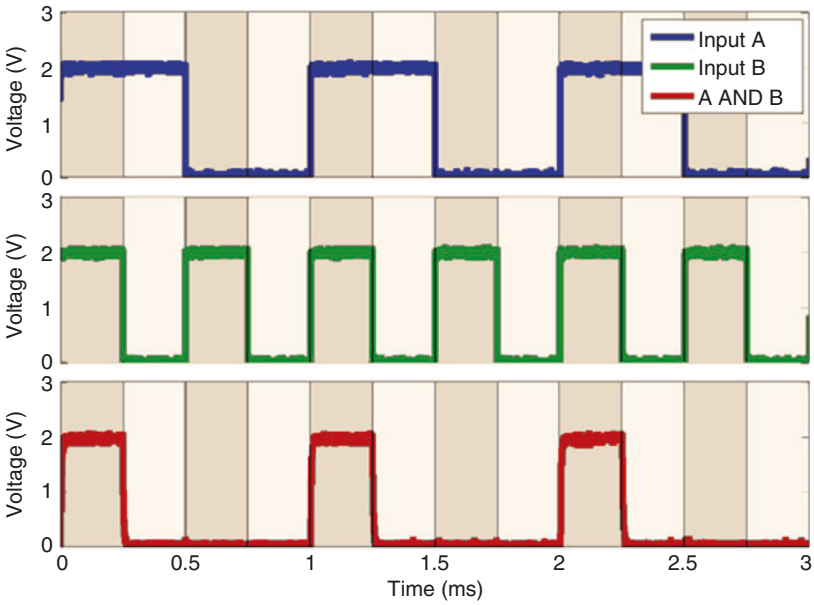
Piezoelectric-Enabled Mechanical Logic

Piezoelectric RF switches have also become enabling in very low-power digital logic circuits. Leveraging the results highlighted in the switch-scaling section, one can extend the technology to enable a fully functional electromechanical-based logic. Both AlN and PZT thin films have successfully demonstrated mechanical logic functionality. To date, the AlN logic relays have been assembled to yield both NAND and inverter gates using a body bias on the actuator [28, 29]. The body bias is used to reduce the contact gap and enable dynamic operation at less than 1 V.

The low leakage current observed on a PZT-based actuator has led to the development of a digital, single-pole, double-throw (SPDT) mechanical relay with zero leakage current in one state and under 30 f. of leakage current in the other [30]. The device is shown schematically in Fig. 14 and with the two independent current paths highlighted in green via electron micrograph in Fig. 15. In addition to low leakage currents, the structure's material composition provides innate radiation hardening [31] and the SPDT configuration of this device, highlighted in Fig. 15, makes it optimally suited for complimentary logic circuits.

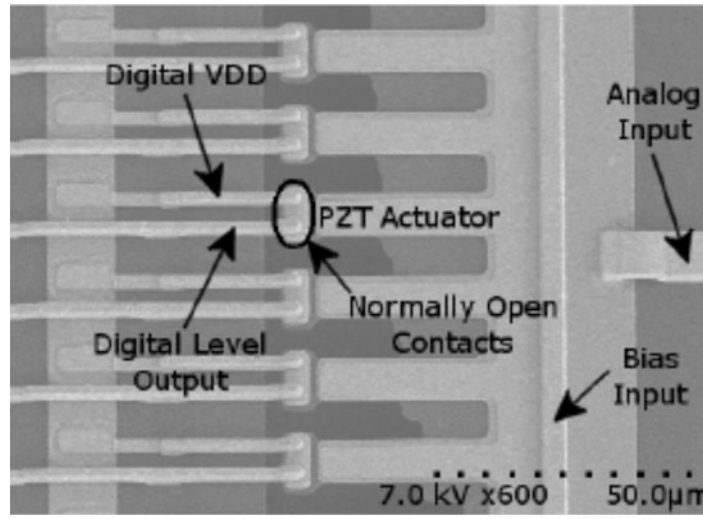
Piezoelectric MEMS Switch, Table 2 Boolean logic functions possible with PiezoMEMS SPDT relay

Boolean function	Top Pt	Bottom Pt	NC input	NC output	NO output	NO input
Always 0	–	–	GROUND	OUTPUT		INPUT B
A NOR B	VDD	INPUT A	GROUND	OUTPUT		NOT INPUT B
(NOT A) AND B	GROUND	INPUT A	INPUT B	OUTPUT		GROUND
NOT A	GROUND	INPUT A	VDD	OUTPUT		GROUND
A AND (NOT B)	GROUND	INPUT B	INPUT A	OUTPUT		GROUND
NOT B	GROUND	INPUT B	VDD	OUTPUT		GROUND
A XOR B	INPUT A	INPUT B	GROUND	OUTPUT		VDD
A NAND B	GROUND	INPUT A	VDD	OUTPUT		NOT INPUT B
A AND B	GROUND	INPUT A	GROUND	OUTPUT		INPUT B
A XNOR B	INPUT A	INPUT B	VDD	OUTPUT		GROUND
B	–	–	INPUT B	OUTPUT		INPUT B
(NOT A) OR B	GROUND	INPUT A	VDD	OUTPUT		INPUT B
A	–	–	INPUT A	OUTPUT		INPUT A
A OR (NOT B)	GROUND	INPUT B	VDD	OUTPUT		INPUT A
A OR B	GROUND	INPUT A	INPUT B	OUTPUT		VDD
Always 1	–	–	VDD	OUTPUT		VDD

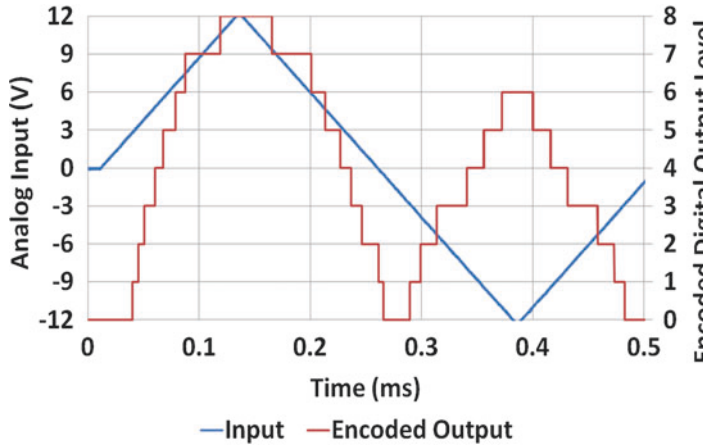


Piezoelectric MEMS Switch, Fig. 16 Measured results of Boolean AND function being performed via a PiezoMEMS relay

Due to PZT's large $e_{31,f}$ coefficient, large deflections, several hundred nanometers for nanomechanical logic, can be achieved at modest voltages. This enables the relay to have an open-state contact gap sufficient to support up to 40 V hot switched across its contacts while still allowing for fully functional Boolean logic operations at fewer than 2 V (see Table 2). Additionally, due to the ambipolar nature of metal-to-metal contacts, the device is well suited to functioning in a pass-gate logic configuration, allowing for a reduction in gate count without harming the signal integrity. Figure 16 demonstrates this behavior with a Boolean AND operation. Here, the top electrode of Fig. 15 is grounded, the bottom electrode is connected to "Input A," the normally closed (NC) input is grounded, the normally open (NO) input is connected to "Input B," and



Piezoelectric MEMS Switch, Fig. 17 SEM image of a piezoelectric thin-film mechanical relay-based analog-to-digital converter (ADC) [33]



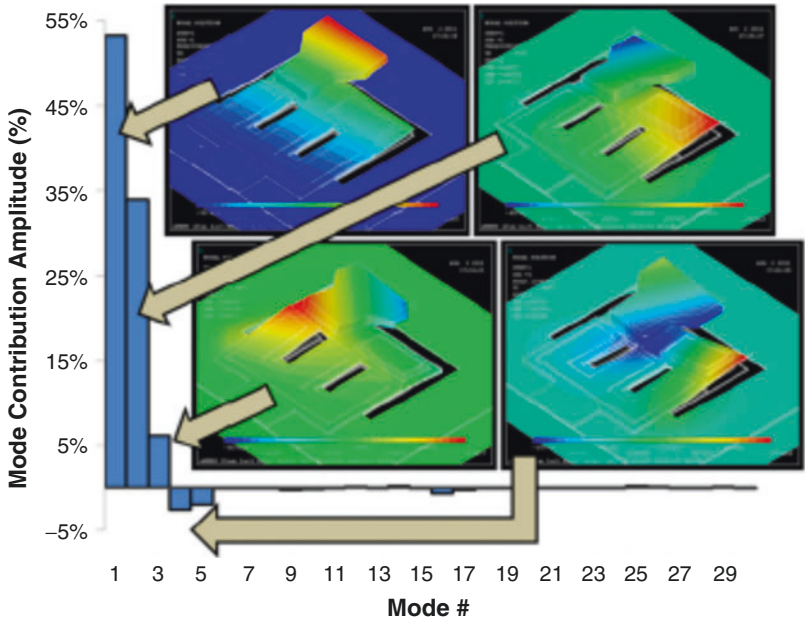
Piezoelectric MEMS Switch, Fig. 18 Comparison between the analog input voltage (*blue*) to the encoded digital output level (*red*) for the 3-bit ADC illustrated in Fig. 17

both the NC and NO outputs are connected together. In this configuration, when "Input A" is false, ground is shorted to the output, and when "Input A" is true, "Input B" is passed to the output. The additional Boolean functions and the related device configurations are listed in Table 2. The device can also operate as either of the two classes of digital memory: dynamic and static, and for additional information, see Refs. [32, 33].

In addition to basic Boolean operators, a slight modification to the relay can enable a multi-bit analog-to-digital converter. As shown in Fig. 17, an array of variable length piezoelectric actuators can be designed to realize a 3-bit ADC [33]. In this example, the length was varied from 11 to 29 µm with actuation voltages expected to range from 2 to 12 V. The resultant ADC encoded output along with the analog actuation voltage is shown in Fig. 18.

Switching Dynamics

A unique attribute of the design illustrated in Fig. 15 is the switching dynamics based on the electromechanical design. A reported dynamic model describes well the measured actuation voltage dependence on the propagation delay of PiezoMEMS relays [34]. The model explains the observed improvement in switching speed of



Piezoelectric MEMS Switch, Fig. 19 Plot of the peak modal displacements (in% of total) for the first 30 vibration modes. The modeled mode shapes for the first, second, third, and fourth modes are also included and the corresponding displacements are indicated. The peak displacements are the maximum values for each mode over a 4 μs period [34]

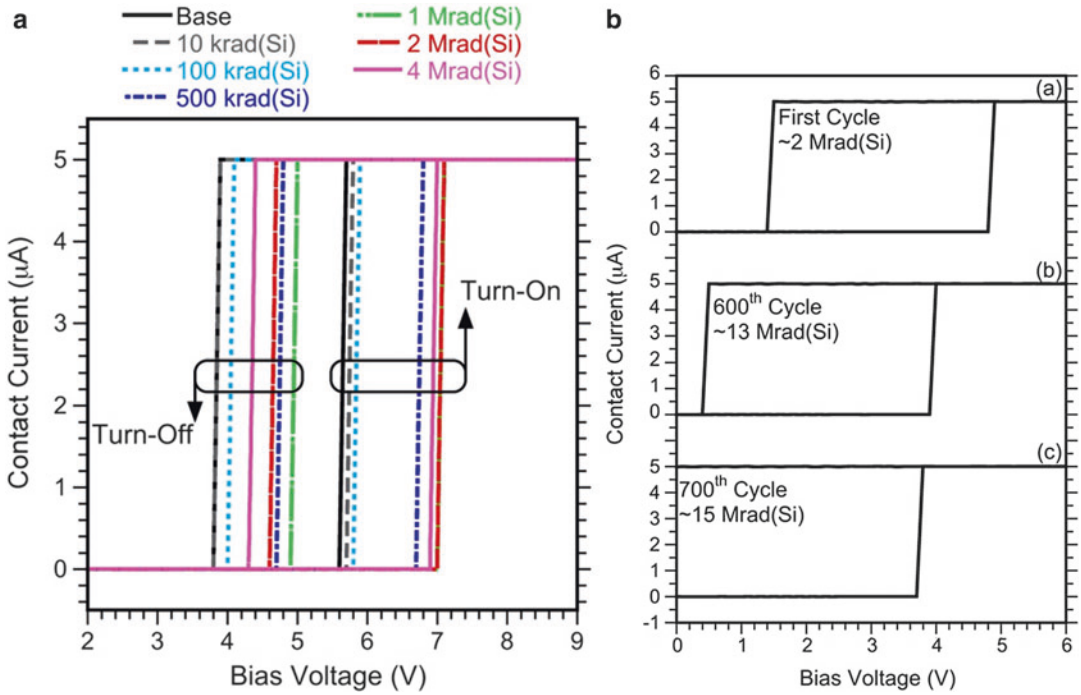
these devices relative to single degree of freedom switching speed models and suggests the strong potential for improved performance in RF MEMS and mechanical logic devices through the exploitation of higher-order vibrational modes (see Fig. 19). The model enables speed optimization of such devices via structural and electrode design to enhance the contribution of favorable vibrational modes and to suppress modes that reduce switching speed. For a detailed review of the model and the resulting improvements in switching speed, see [34].

Radiation Effects

To explore the radiation tolerance of the electromechanical relays, ^{60}Co exposure was used to examine the effects of total ionizing dose (TID) radiation on the piezoelectric properties and viability of PZT-based actuators to serve as devices within radiation-intensive environments [35]. The devices showed continued operation after 4 Mrad (Si) exposures in both the 0 V and 8 V fixed bias cases (see Fig. 20a). Moreover, under typical biasing conditions, the devices were able to operate correctly during exposure (855.6 rad(Si)/s) to a TID in excess of 13 Mrad(Si); see Fig. 20b. Data suggests that the changes detected are due to creation of defects in the PZT and subsequent pinning of domains and domain walls by the defects within the PZT material.

Conclusion

This entry provides a general overview of piezoelectric MEMS-based switch technologies and their application for RF circuits and mechanical digital relays. Both AlN and PZT thin films have demonstrated success operation of switches for both RF and digital applications. The fundamental material differences between the AlN and PZT result in the use of bimorph actuators for AlN and unimorph actuators for PZT. The key advantages of a piezoelectric switch can be summarized as low operating or drive voltage (typically sub-10 V) reducing the need for large area and power-consuming charge pumps for integrated drive applications and full control of actuator displacement when the switch is making contact to improve contact dynamics and improve reliability. Additionally, the piezoelectric N/MEMS switch has been shown to be capable of operation



Piezoelectric MEMS Switch, Fig. 20 Experimental operation (turn on/turn off) of a mechanical relay (similar to Fig. 15) exposed to ^{60}Co (855.6 rad(Si)/s): (**a**) radiation exposure under 0 V bias on actuator and (**b**) cycle operation during exposure (Reprinted with permission. Copyright 2008 IEEE)

under exposure to gamma radiation and able to survive to in excess of 13 Mrad(Si).

Acknowledgments The authors would like to thank Dr. Tony Ivanov, Daniel Judy, Dr. Roger Kaul, Joel Martin, Brian Power, and Dr. Luz Sanchez of the US Army Research Laboratory and Dr. Cory Cress of the US Naval Research Laboratory for their useful comments and their roles in device fabrication, characterization, and testing.

Cross-References

► Piezoelectric Effect at Nanoscale

References

1. Brown, E.: RF-MEMS switches for reconfigurable integrated circuits. IEEE Microw. Theory Tech. **46**(11), 1868–1880 (1998)
2. Rebeiz, G.: RF MEMS Theory, Design, and Technology. Wiley, Hoboken (2003)
3. Rebeiz, G., Tan, G., Hayden, J.: RF MEMS phase shifters: device and applications. IEEE Microw. Mag. **3**, 72–81 (2002)
4. Petersen, K.E.: Micromechanical membrane switches on silicon. IBM J. Res. Dev. **23**, 376–385 (1979)
5. Larson, L.E., Hackett, R.H., Melendes, M.A., Lohr, R. F.: Micromachined microwave actuator (MIMAC) technology – a new tuning approach for microwave integrated circuits. Millimeter-Wave Monolithic Circuits Symposium Digest, pp. 27–30. Boston (1991)
6. McCarthy, B., Adams, G.G., McGruer, N.E.: A dynamic model, including contact bounce, of an electrostatically actuated microswitch. J. Microelectromech. Syst. **11**, 276–283 (2002)
7. Setter, N.: ABC of piezoelectricity and piezoelectric materials. In: Piezoelectric Materials in Devices, pp. 1–28. Ceramics Laboratory EPFL, Lausanne (2002)
8. Muralt, P.: Piezoelectric thin films for MEMS. Integr. Ferroelectr. **17**, 297–307 (1997)
9. Dubois, M., Muralt, P.: Measurement of the effective transverse piezoelectric coefficient e31, f of AlN and Pb(ZrxTi1-xO3) thin films. Sensors Actuators **77**, 106–112 (1999)
10. IEEE Standard on Piezoelectricity, ANSI/IEEE Std 176 (1987)
11. Muralt, P.: PZT thin films for microsensors and actuators: where do we stand? IEEE Trans. Ultrason. Ferroelectr. Freq. Control **47**, 903–914 (2000)
12. Polcawich, R.G., Pulskamp, J.S.: Additive processes for strain-dependent materials: piezoelectric

MEMS. In: MEMS Processing Handbook, pp. 273–353. Springer, New York (2011)
13. Fang, T.-H., Chang, W.-J., Lin, C.-M.: Nanoindentation characterization of ZnO thin films. Mater. Sci. Eng. A **452–453**, 715–720 (2007)
14. Yang, P.-F., Wen, H.-C., Jian, S.-R., Lai, Y.-S., Wu, S., Chen, R.-S.: Characteristics of ZnO thin films prepared by radio frequency magnetron sputtering. Microelectron. Reliab. **48**, 389–394 (2008)
15. Tonisch, K., et al.: Piezoelectric properties of polycrystalline AlN thin films for MEMS application. Sensors Actuators A **132**, 658–663 (2006)
16. Andrei, A., Krupa, K., Jozwik, M., Delobelle, P., Hirsinger, L., Gorecki, C., Nieradko, L., Meunier, C.: AlN as an actuation material for MEMS applications. The case of AlN driven multilayered cantilevers. Sensors Actuators A **141**, 565–576 (2008)
17. Damjanovic, D.: Ferroelectric, dielectric and piezoelectric properties of ferroelectric thin films and ceramics. Rep. Prog. Phys. **61**, 1267–1324 (1998)
18. Peroulis, D., Pacheco, S.P., Sarabandi, K., Katehi, L.: Electromechanical considerations in developing low voltage RF MEMS switches. IEEE Trans. Microw. Theory Tech. **51**, 259–270 (2003)
19. Majumber, S. et al.: Measurement and modeling of surface micromachined, electrostatically actuated microswitches. In: International Conference on Solid-State Sensors and Actuators, pp. 1145–1148. Chicago (1997)
20. Gross, S.J., Tadigadapa, S., Jackson, T.N., Trolier-McKinstry, S., Zhang, Q.Q.: Lead-zirconate-titanate-based piezoelectric micromachined switch. Appl. Phys. Lett. **83**, 174–176 (2003)
21. Lee, H.C., Park, J.Y., Bu, J.U.: Piezoelectrically actuated RF MEMS DC contact switches with low voltage operation. IEEE Microw. Wireless Compon. Lett. **15**(4), 202–204 (2005)
22. Polcawich, R.G., Pulskamp, J.S., Judy, D., Ranade, P., Trolier-McKinstry, S., Dubey, M.: Surface micromachined microelectromechanical ohmic series switch using thin-film piezoelectric actuators. IEEE Trans. Microw. Theory Tech. **55**, 2642–2654 (2007)
23. Polcawich, R.G., Judy, D., Pulskamp, J.S., Trolier-McKinstry, S., Dubey, M.: Advances in piezoelectrically actuated RF MEMS switches and phase shifters. In: Proceedings of IEEE International Microwave Symposium, pp. 2083–2086. Honolulu (2007)
24. Mahameed, R., et al.: Dual-beam actuation of piezoelectric AlN RF MEMS switches monolithically integrated with AlN contour-mode resonators. J. Micromech. Microeng. **18**, 105011 (2008)
25. Sinha, N. et al.: Ultra thin AlN piezoelectric nanoactuators. In: Proceedings of the Transducers 2009, pp 469–472. Denver (2009)
26. Judy, D., Polcawich, R.G., Pulskamp, J.: Low Voltage Complementary MEMS Logic using Piezoelectric Actuators, Solid State Sensor, Actuator and Microsystems Workshop, pp. 328–331. Hilton Head (2008)
27. Judy, D.C., Pulskamp, J.S., Polcawich, R.G., Currano, L.: Piezoelectric nanoswitch. In: Proceedings of the 22nd IEEE MEMS Conference, pp. 591–594. Sorento (2009)
28. Sinha, N., Jones, T.S., Guo, Z., Piazza, G.: Demonstration of low voltage and functionally complete logic operations using body-biased complementary and ultra-thin AlN piezoelectric mechanical switches. In: Proceedings of 2010 MicroElectroMEchanical Systems Conference (MEMS 2010), Hong Kong. (2010)
29. Sinha, N., Jones, T.S., Guo, Z., Piazza, G.: Body-biased complementary logic implemented using AlN piezoelectric MEMS switches. In; Proceedings of 2009 IEEE. International Electron Devices Meeting (IEDM 2009), Baltimore, MD (2009)
30. Proie Jr., R.M., Polcawich, R.G., Pulskamp, J.S., Ivanov, T., Zaghloul, M.: Development of a PZT MEMS switch architecture for low-power digital applications. J. Microelectromech. Syst. **20**(4), 1032–1042 (2011)
31. Benedetto, J.M., Moore, R.A., McLean, F.B., Brody, P.S., Dey, S.K.: The effect of ionizing radiation on sol-gel ferroelectric PZT capacitors. IEEE Trans. Nucl. Sci. **37**(6), 1713–1717 (1990)
32. Proie, R., Polcawich, R.G., Pulskamp, J.S., Ivanov, T., Zaghloul, M.: Nano -electromechanical storage element for a low power complimentary logic architecture using PZT relays. In: Proceedings of Transducers, Beijing, China (2011)
33. Proie, R., Pulskamp, J.S., Polcawich, R.G., Ivanov, T., Zaghloul, M.: Low power 3-bit piezoelectric MEMS analog to digital converter. In: MEMS 2011, pp. 1241–1244. Cancun (2011)
34. Pulskamp, J.S., Proie, R.M., Polcawich, R.G.: Nano- and micro-electromechanical switch dynamics. J. Micromech. Microeng. **24**, 015011 (2014) (11pp)
35. Proie, R.M., Polcawich, R.G., Cress, C.D., Sanchez, L.M., Grobicki, A.D., Pulskamp, J.S., Roche, J.-H.: Total ionizing dose effects in piezoelectric MEMS relays. IEEE Trans. Nucl. Sci. **60**, 4505–4511 (2013)

Piezoelectric Switches

► NEMS Piezoelectric Switches
► Piezoelectric MEMS Switch

Piezoresistance

► Piezoresistivity

Piezoresistive Effect

▶ Piezoresistivity

Piezoresistivity

Woo-Tae Park
Seoul National University of Science and Technology, Seoul, South Korea

Synonyms

Piezoresistance; Piezoresistive effect; Semiconductor piezoresistance

Definition

Piezoresistivity is defined as change of resistance of semiconductors from the applied mechanical stress.

History

The discovery of piezoresistance dates back to the late nineteenth century (1856), when Lord Kelvin discovered the change of resistance in copper and iron wires. He compared the difference in the change of resistance of the two materials, while the elongation was the same [1]. The term "piezoresistance" was coined by Cookson in 1935, when he defined the term as the change in conductivity with stress [2]. The term "piezo" is rooted from the term "piezen" from the Greek meaning *to press*. "Piezoelectricity" sounds similar, but with a very different effect, where charge is generated from applied stress.

When Nobel laureates, Bardeen and Shockley invented the transistor, they also predicted relatively large conductivity change with deformation in single crystal semiconductors [3]. Later C.S. Smith, who was a visiting scientist at Bell labs, reported his pioneering results in the large piezoresistance in semiconductors [4]. Smith's work marked the start of semiconductor piezoresistive sensors (Table 1).

Piezoresistivity, Table 1 Values of the derived piezoresistance coefficients in units of 10^{-12} cm^2 dyn^{-1}. Resistivities are in ohm-cm

Type, material, and resistivity	Π_{44}	Π_{11}–Π_{12}	$-\frac{1}{P}\frac{\delta\rho}{\rho} = \Pi_{11} + 2\Pi_{12}$	
			Adiabatic	Isothermal
n Ge 1.5	−138.1	+1.0	(−8.7)[a]	−7.0
n Ge 4.2[b]				−8.3
n Ge 5.7	−136.8	+1.2	−10.5	−8.8
n Ge 8.4[b]				−9.8
n Ge 9.9	−137.9	+0.3	−14.7	−13.0
n Ge 16.6	−138.7	+0.3	−16.2	−17.9
n Ge 18.5[b]				−27.1
p Ge 0.9–3.3[c]				+2.5
p Ge 1.1	+96.7	−6.9	+2.7	+5.2
p Ge 15.0	+98.6	−15.6	−0.6	+1.9
n Si 11.7	−13.6	−115.6	+4.6	+5.7[d]
p Si 7.8	+138.1	+7.7	+4.4	+6.0

After Smith [4]. © 1954 American Physical Society, http://www.prola.aps.org/abstract/PR/v94/i1/p42_1
[a]This value extrapolated from *n* Ge 5.7 and *n* Ge 9.9
[b]P. W. Bridgman (see Ref. [13]) and by latter. Values are from the initial pressure effect
[c]P. W. Bridgman (see Ref. [13]). There were six specimens in this resistivity range, five of which were consistent at the value shown and one of which showed zero pressure effect
[d]Measured directly

In the 1960s, at least a couple of companies started manufacturing commercial silicon strain gauges, and in 1962, an "integrated" device using diffused piezoresistors was used to make pressure sensors at Honeywell [5].

From the early piezoresistive strain gauges to the present implanted piezoresistive pressure sensors and accelerometers, this technology required the three-dimensional micromachining of silicon. This requirement led to the development of bulk etching technology, and the development of MEMS technology.

Piezoresistance Essentials

Theory

Electrical resistance can be defined by a function of the resistivity of the material and the dimensions, as given in the following equation:

$$R = \rho \frac{l}{a}$$

where R = resistance, ρ = resistivity, l = length, and a = cross-sectional area.

So, the change of resistance is a function of the dimensional change and the resistivity change.

$$\frac{\Delta R}{R} = (1 + 2\upsilon)\,\varepsilon + \frac{\Delta \rho}{\rho}$$

where υ = Poisson's ratio and ε = strain.

The former term comes from the dimensional change, and the latter term is the resistivity change. The change of resistance to the applied stress is defined as the gauge factor.

$$GF = \frac{\Delta R/R}{\varepsilon}$$

where GF = gauge factor.

For metals, geometric effect alone approximately accounts for 1.4–2.0 of gauge factor, and the change in resistivity is almost zero. However, for single crystal semiconductors, such as silicon and germanium, the change in resistivity can be 50–100 times larger than the geometric term. Because of this exceptional large resistivity change in semiconductors, a new term called "piezoresistive coefficient" denoted by "π" is defined as the following:

$$\frac{\Delta \rho}{\rho} = \pi_t \sigma_t + \pi_l \sigma_l$$

where π = piezoresistive coefficient, σ = stress, and the subscripts t = transverse, and l = longitudinal. Transverse stress and piezoresistive coefficient are defined as mechanical stress and electrical field being orthogonal, and longitudinal stress and piezoresistive coefficient are defined as mechanical stress and electrical field being parallel.

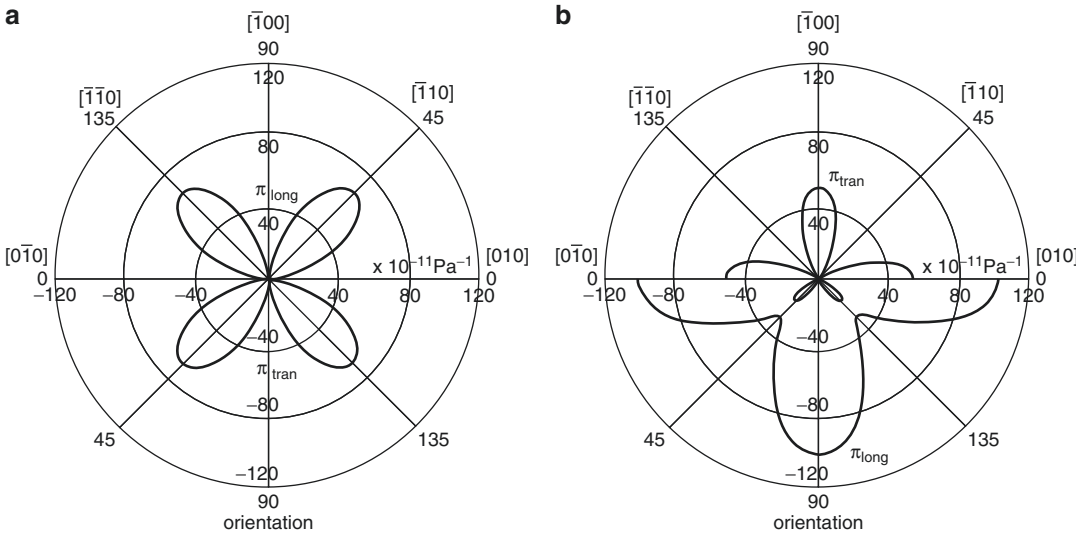
The theories to explain this large piezoresistive property of semiconductors are based on one-dimensional descriptions of electron and hole transport in crystalline structures under strain. The models are based on bandgap energy models, wave mechanics, and quantum effects; interested reader should refer to [6] for detailed explanation of the models.

N-type and p-type piezoresistors exhibit opposite trends in resistance change and different direction dependence under strain. The magnitude and sign of the piezoresistive coefficient depend on a number of factors including doping concentration, temperature, crystallographic direction, and relation between the current and stress direction. For n-type silicon, Herring explained piezoresistance in silicon or germanium using the Many-Valley model [7].

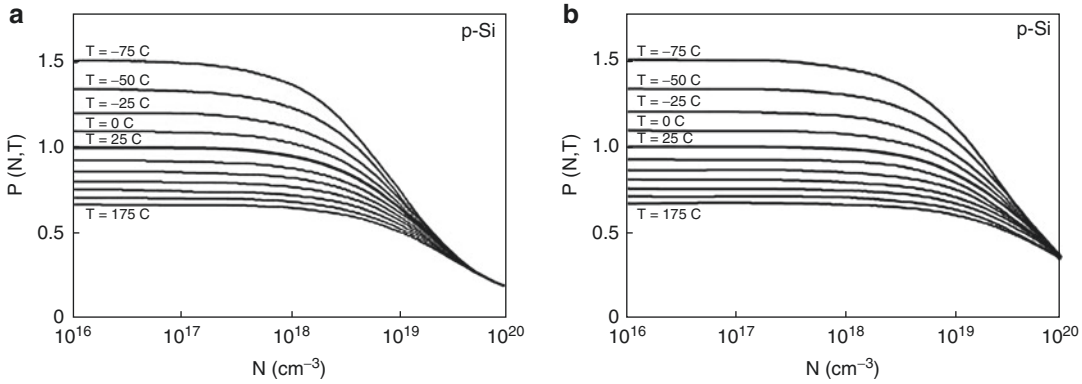
Fabrication

The most commonly used fabrication processes for doping piezoresistors are diffusion and ion implantation. Although both methods have its advantage for doping silicon, ion implantation is the dominant method used for doping piezoresistors.

1. Diffusion: Diffusion is the migration of dopant atoms from a high-concentration region to a low-concentration region. The fabrication steps for diffusion, most commonly involve pre-deposition and drive-in step. During pre-deposition, the wafers are placed in high temperature (900–1300 °C) with a gas-phase or a solid-phase dopant source. The dopant source (e.g., diborane, phosphine, or arsine) is carried in an inert gas (N_2 Ar). The resulting dopant profile is approximated by a complementary error function. After the pre-deposition step, the dopants are driven in deeper with high-temperature annealing. Diffusion method is still useful for high-dose doping (near solid solubility).
2. Ion implantation: Ion implantation provides better control of the dopant dose, which is why it is the preferred method to dope piezoresistors today. The process starts by accelerating the dopant ions at high energy (KeV to MeV) into the substrate. The implanted ions leave a cascade of damage in the crystal structure. Dopant distribution is approximated by a symmetric Gaussian distribution. During implantation, 7° tilt angle used



Piezoresistivity, Fig. 1 Piezoresistive coefficients in the popular (100) plane for (**a**) p-type silicon (**b**) n-type silicon (After Kanda [6], © 1982 IEEE)



Piezoresistivity, Fig. 2 The piezoresistance factor P(N, T) as a function of doping concentration and temperature for (**a**) p-type silicon and (**b**) n-type silicon (After Kanda [6], © 1982 IEEE)

to avoid ion channeling effect. After implant, a short high-temperature annealing is used to restore the crystal lattice. Typical masking material is photoresist, silicon dioxide, or silicon nitride.

Design and Process Effects on Performance

Design and process conditions strongly affect the sensitivity and noise of the piezoresistor. Doping concentration and piezoresistor crystallographic orientation strongly affect the sensitivity. Choosing the device geometry, doping, and anneal conditions, the piezoresistive device designer must also consider the temperature coefficients of sensitivity and resistance, nonlinearity with strain and temperature, and noise limits.

Kanda established the graphical representations of the piezoresistive coefficients in arbitrary orientations in the commonly used crystal substrates (Fig. 1) [8]. Kanda also showed theoretical calculations of the piezoresistive coefficients and dopant concentration (Fig. 2).

Resistors are generally sensitive to temperature. Silicon-doped resistors are also sensitive to

pressure. Temperature affects the absolute resistor values, but also affects the sensitivity. General trend is that the higher the sensitivity, the higher the temperature coefficient of the sensitivity. However, the rise in temperature sensitivity is faster than that of piezoresistive sensitivity. Sensor manufacturers use the Wheatstone configuration of sensors and conditioning circuits to compensate this temperature sensitivity.

Piezoresistive sensitivity also shows nonlinearity at high strain. So designers should avoid exerting strain higher than 0.1 %.

Noise

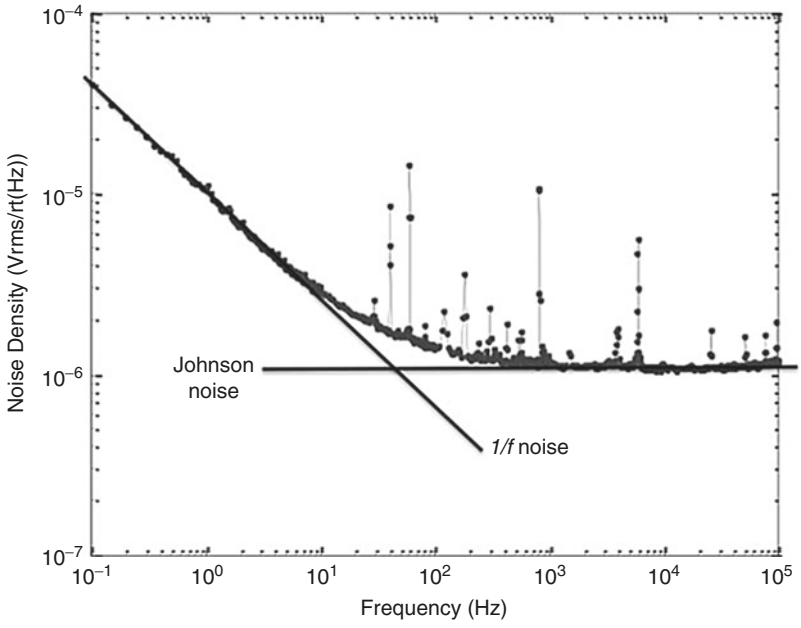
Noise in piezoresistors manifests as random fluctuations of the potential. Apart from other noise sources in the system, noise in the piezoresistor itself sets the fundamental limit in the resolution of the piezoresistor. Noise contributions come from two main components: Johnson noise (thermal noise) and 1/f noise (flicker noise). Figure 3 shows a representative example of piezoresistor noise.

The Johnson noise is determined by resistor value. So higher resistors have higher Johnson noise [9].

1/f noise is, as the name implies, inversely proportional to the frequency. So this noise component is the main culprit for low-frequency measurements. This is important since piezoresistors are used in mostly low-frequency static measurements. Although there are still debates on the source of this noise, there are two main popular theories. Hooge model views the fluctuation as carriers' mobility fluctuation by scattering centers [10]. Since the scattering centers depend on the quality of the crystal, annealing to repair internal damage reduces the noise. McWhorter model views the fluctuation as a surface effect scattering [11].

Alternative Piezoresistive Materials and Structures

Today, most commercially available piezoresistive sensors use silicon as the material. With the advances in material science, researchers studied new materials or scaling of the silicon to enhance the properties in sensitivity, and higher melting temperature.



Piezoresistivity, Fig. 3 Typical noise curve of a piezoresistor. The sloped *solid line* is the noise dominated by 1/f noise component plus the underlying Johnson noise. The *flat line* represents the inherent Johnson noise

Silicon Carbide (SiC)

Silicon carbide was attractive because of its superior mechanical properties over silicon such as higher modulus (424 GPa), higher sublimation temperature (1800 °C), and inertness to corrosive environments. Although the fabrication is not as straightforward as silicon, and growing oxide is a challenge due to the carbon in the film, SiC-based piezoresistive devices were able to achieve gauge factors as large as −35 [12] and withstand harsh operating environments, such as high temperatures (200–500 °C) [13] and high impact (40,000 g) [14].

Diamond

Diamond has higher Young's modulus, yield strength, and thermal conductivity than silicon. The reported piezoresistive gauge factors of single crystal diamond range up to 3836. However, it is extremely difficult to grow single crystal diamond, and poly diamond has similar sensitivities as silicon carbide [15].

Carbon Nanotube (CNT)

Carbon nanotubes are the strongest and most resilient materials known in nature. Electronically, based on the manner in which the graphene films are rolled, it can be metallic, semiconducting, or insulator. Gauge factors range from 400 to 850 [16] for CNT response adhered to pressurized membranes. However, the difficulty is in the ohmic connection to the CNT. A more manufacturable process was used mixing CNT to SU8, but the attained sensitivities were lower than what is possible with bare CNTs [17].

Silicon Nanowires

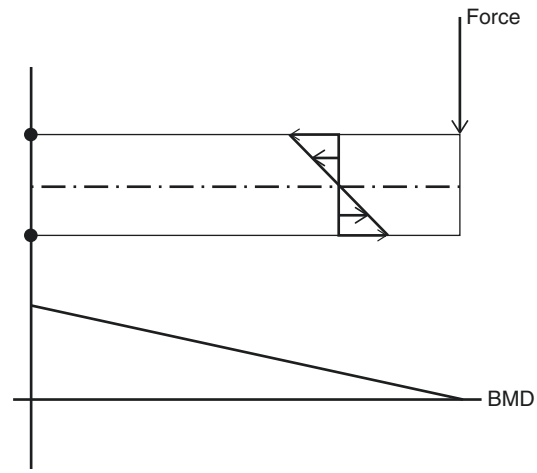
Silicon nanowires are simply reducing the dimension of conventional single crystal silicon components. Very high gauge factor, higher than 6000, was reported for silicon nanowires in dimensions under 100 nm [18]. Reducing the dimensions down to 100 nm is not a big challenge for modern lithography and further shrinking by self-limiting oxidation is possible. Neuzil reported that top-down manufactured (by standard clean room process) nanowires can reach up to 5000 of gauge factor [19]. However, recently Milne reported that his new measurement techniques revealed that the high change of signal is most likely due to long-term drift rather than high gauge factor [20].

Applications

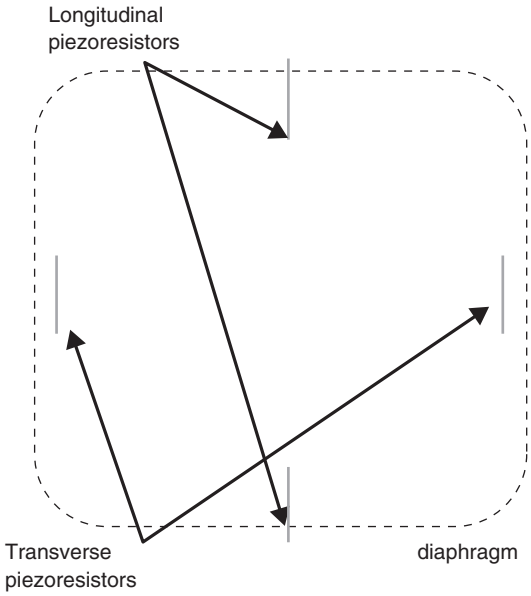
Piezoresistors have the longest history of usage as MEMS devices since the 1950s [4]. Since they measure strain, piezoresistors have wide applications such as cantilevers, pressure sensors, and accelerometers.

Cantilevers

Cantilevers are used as force sensors, biosensors, or imaging (atomic force microscopy (AFM)). Figure 4 is showing where the piezoresistor should be placed, which is the maximum stress location of the cantilever. For force sensing applications, scaling is important to measure smaller cantilevers to measure smaller force. Piezoresistors do have advantage in this area, since they can more easily scale if compared to capacitive or optical detection methods. Tortonese et al. first used piezoresistive sensing for cantilevers and achieved 0.1 Å resolution [21] (Fig. 4).



Piezoresistivity, Fig. 4 A cantilever sensor representation. Upon applied normal force, the bending moment amplitude is shown by the bending moment diagram (*BMD*). The internal stress/strain due to the bending moment is shown by the *arrows*. Therefore, the maximum stress/strain points are at the root of the cantilever and at the *top* and *bottom* surfaces of the beam



Piezoresistivity, Fig. 5 Typical piezoresistive pressure sensor layout. The *dotted line* is representing the micromachined diaphragm. Upon applied pressure, the two longitudinal piezoresistors change resistance in opposite direction as the two transverse piezoresistors. Forming a complete Wheatstone bridge with these four resistors, this configuration allows the maximum output

Pressure Sensors

Pressure sensing is the most reported application for piezoresistive MEMS sensors. With four piezoresistors at the four edges forming a Wheatstone bridge configuration, the output is maximized in a differential sensing configuration. Upon applied pressure, the two longitudinal piezoresistors change in opposite direction as the two transverse piezoresistors (Fig. 5).

Wheatstone bridge also improves the temperature sensitivity by canceling the temperature coefficient in each neighboring resistor. Piezoresistive sensing is still the most dominant sensing type for all MEMS pressure sensors. The only downside of piezoresistive sensing for pressure sensors is the inability to adapt to polysilicon surface micromachining. However, recent work from Bosch using porous 0silicon showed a very low-profile manufacturing process without the need of backside etching [22].

Accelerometers

One of the very first MEMS accelerometers reported used piezoresistive transduction [23]. Numerous research articles reported the use of piezoresistive sensing in accelerometers. Recent products from Hitachi metals and Hokuriku use a novel three-axis sensing structure for piezoresistive accelerometers. However, the most dominant sensing method used for commercial accelerometers is capacitive. This is due to the fact that capacitive sensing can be more sensitive by employing very flexible springs without the need to have large strain concentration on the springs themselves. Capacitive sensing has proven to be more scalable for accelerometers.

Conclusions

Short after the discovery of transistors in 1947, piezoresistive property has been discovered in 1954 [4]. Through more than 50 years of steady development, piezoresistive sensors have transformed the way we live. It is used in numerous medical, automotive, aerospace, and consumer markets. While it is one of the easiest sensing techniques to construct, the fundamental physics is still not well understood. With the advances in material science, new materials are being proposed and studied. Yet, still the simplest implanted piezoresistors are the most common forms in use today. Piezoresistive sensing will always be in the MEMS designers' toolbox and employed in many devices in the years to come.

Cross-References

▶ Atomic Force Microscopy
▶ Piezoresponse Force Microscopy and Spectroscopy

References

1. Thomson, W.: On the electro-dynamic qualities of metals: effects of magnetization on the electric conductivity of nickel and of iron. Proc. R. Soc. Lond. **8**, 546–550 (1856)

2. Cookson, J.W.: Theory of the piezo-resistive effect. Phys. Rev. **47**, 194–195 (1935)
3. Bardeen, J., Shockley, W.: Deformation potentials and mobilities in non-polar crystals. Phys. Rev. **80**, 72–80 (1950)
4. Smith, C.S.: Piezoresistance effect in germanium and silicon. Phys. Rev. **94**, 42–49 (1954)
5. Tufte, O.N., Chapman, P.W., Long, D.: Silicon diffused-element piezoresistive diaphragms. J. Appl. Phys. **33**, 3322–3327 (1962)
6. Sun, Y., Thompson, S.E., Nishida, T.: Physics of strain effects in semiconductors and metal-oxide-semiconductor field-effect transistors. J. Appl. Phys. **101**, 104503 1–104503 22 (2007)
7. Herring, C.: Transport and deformation potential theory for many-valley semiconductors with anisotropic scattering. Phys. Rev. **101**, 944 (1956)
8. Kanda, Y.: A graphical representation of the piezoresistance coefficients in silicon. IEEE Trans. Electron. Dev. **29**, 64–70 (1982)
9. Nyquist, H.: Thermal agitation of electric charge in conductors. Phys. Rev. **32**, 110–113 (1928)
10. Hooge, F.N.: 1/f noise is no surface effect. Phys. Lett. A. **29**, 139–140 (1969)
11. McWhorter, A.L.: Semiconductor Surface Physics. University of Pennsylvania Press, Philadelphia (1957)
12. Okojie, R.S., Ned, A.A., Kurtz, A.D., Carr, W.N.: Characterization of highly doped n- and p-type 6H-SiC piezoresistors. Electron. Dev. IEEE Trans. **45**, 785–790 (1998)
13. Eickhoff, M., Moller, H., Kroetz, G., Berg, J.V., Ziermann, R.: A high temperature pressure sensor prepared by selective deposition of cubic silicon carbide on SOI substrates. Sens. Actuat. A Phys. **74**, 56–59 (1999)
14. Atwell, A.R., Okojie, R.S., Kornegay, K.T., Roberson, S.L., Beliveau, A.: Simulation, fabrication, and testing of bulk micromachined 6H-SiC high-g piezoresistive accelerometers. Sens. Actuat. A. Phys **104**, 11–18 (2003)
15. Taher, I., Aslam, M., Tamor, M., Potter, T.J., Elder, R. C.: Piezoresistive microsensors using p-type CVD diamond films. Sens. Actuat. A. Phys. **45**, 35–43 (1994)
16. Grow, R.J., Wang, Q., Cao, J., Wang, D., Dai, H.: Piezoresistance of carbon nanotubes on deformable thin-film membranes. Appl. Phys. Lett. **86**, 93104–93107 (2005)
17. Dai, H.: Carbon nanotubes: opportunities and challenges. Surf. Sci. **500**, 218–241 (2002)
18. He, R., Yang, P.: Giant piezoresistance effect in silicon nanowires. Nat. Nanotechnol. **1**, 42–46 (2006)
19. Milne, J.S., Rowe, A.C.H., Arscott, S., Renner, C.: Giant piezoresistance effects in silicon nanowires and microwires. Phys. Rev. Lett. **105**, 226802 (2010)
20. Neuzil, P., Wong, C.C., Reboud, J.: Electrically controlled giant piezoresistance in silicon nanowires. Nano Lett. **10**, 1248 (2010)
21. Tortonese, M., Barrett, R.C., Quate, C.F.: Atomic resolution with an atomic force microscope using piezoresistive detection. Appl. Phys. Lett. **62**, 834–836 (1993)
22. Armbruster, S., Schafer, F., Lammel, G., Artmann, H., Schelling, C., Benzel, H., Finkbeiner, S., Larmer, F., Ruther, P., Paul, O.: A novel micromachining process for the fabrication of monocrystalline Si-membranes using porous silicon. In: The 12th International Conference on Solid-State Sensors, Actuators, and Microsystems, Boston, pp. 246–249 (2003)
23. Roylance, L.M.: A miniature integrated circuit accelerometer for biomedical applications. Ph.D., Electrical Engineering Department, Stanford University (1978)

Piezoresponse Force Microscopy and Spectroscopy

Brian J. Rodriguez[1], Sergei V. Kalinin[2], Dmitry Pelegov[3] and Andrei L. Kholkin[3,4]
[1]University College Dublin, Dublin, Ireland
[2]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[3]Institute of Natural Sciences, Ural Federal University, Ekaterinburg, Russia
[4]Department of Materials and Ceramic Engineering and CICECO, University of Aveiro, Aveiro, Portugal

Definition

Piezoresponse force microscopy and spectroscopy (PFM and PFS, respectively) refer to a family of scanning probe microscopy (SPM) techniques based on the coupling between electrical bias and mechanical response (electromechanical detection). PFM and PFS are broadly used for the characterization of piezoelectric and ferroelectric properties (domain structure, switching dynamics, etc.) at the nanometer scale and are also applicable to polar biological and macromolecular systems. PFM is closely related to electrochemical strain microscopy (ESM), an SPM method used for the characterization of ionic transport and electrochemical reactivity on the nanoscale in ionic conductors.

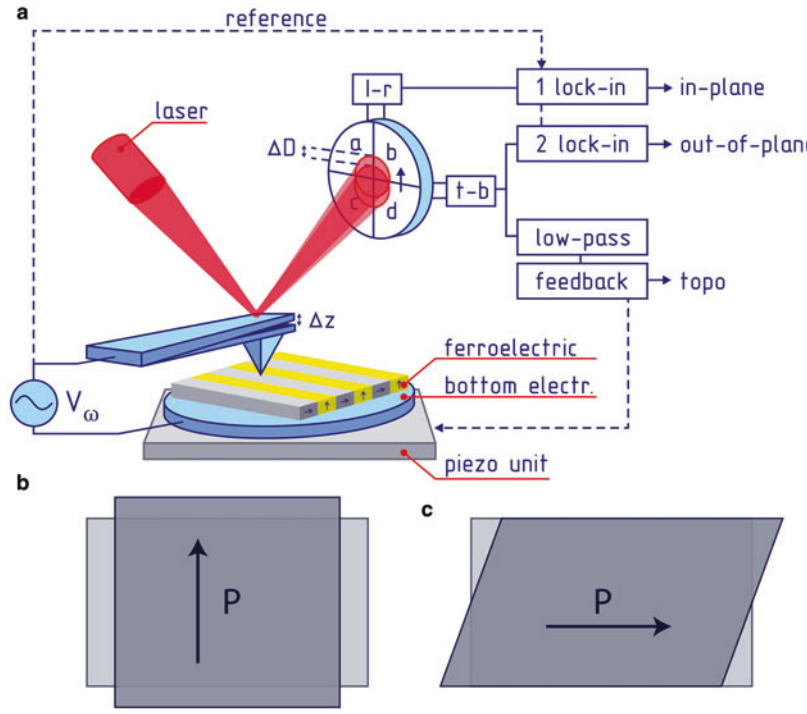
Introduction

PFM and PFS were originally developed for the characterization of ferroelectric materials. The history of PFM can be traced to the first studies of local bias-induced sample deformations in piezoelectric materials using scanning acoustic microscopy [1] and scanning tunneling microscopy [2]. The seminal work by Kolosov and Gruverman [3] gave rise to the application of this method for imaging and manipulating domain structures in ferroelectric materials. Recent advances in PFM and PFS have allowed a wide variety of materials to be studied, including piezoelectric semiconductors, ferroelectric polymers, biological polymers, and energy materials. Many historical aspects of PFM development are discussed in recent books and reviews [4–10]. The unique aspect of PFM compared to other SPM techniques is the detection principle, which utilizes the coupling between bias and strain, often allowing for quantitative measurements.

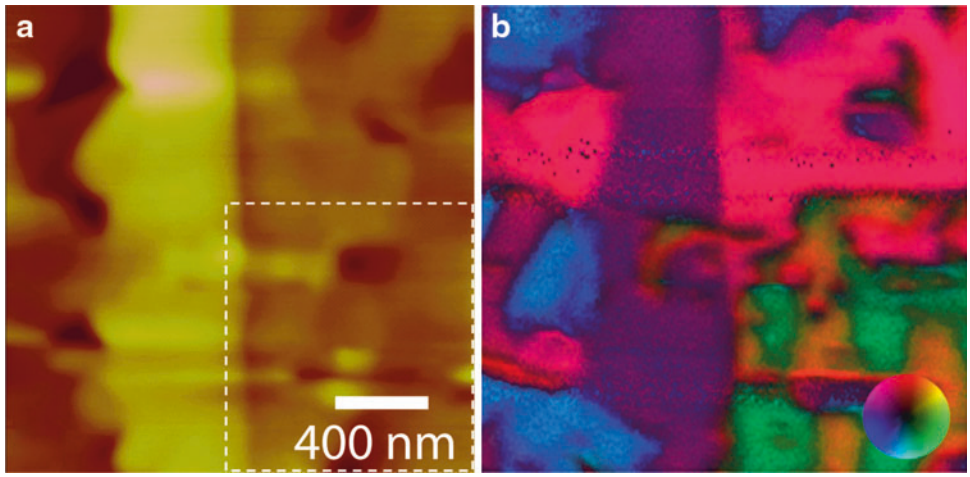
Principles of PFM

PFM is implemented using an atomic force microscope (AFM) operated in contact mode. A nanometer-sized probe tip attached at the end of a microfabricated cantilever is brought into contact with a sample. The interaction force between the sample and the tip causes the cantilever to deflect, which can be detected using an optical beam method, i.e., through the displacement of a laser beam reflected off the backside of the cantilever and into a photodetector (Fig. 1). Typically, a feedback loop is employed which adjusts the tip-sample separation to keep the cantilever deflection and hence the (static) loading force constant. In PFM, a conducting probe is used, and an alternating current bias is applied to the conducting cantilever to create a highly localized time-dependent electric field at the probe tip apex. When the sample under investigation is piezoelectric, the material will deform, i.e., expand and contract, in response to the applied field. The tip can simultaneously be used as the sensor to detect this bias-induced deformation. The frequency of the ac-bias (typically 3 kHz–3 MHz) is chosen to be much higher than the scan rate of the AFM (~1 Hz), and thus, using a lock-in amplifier technique, the sample topography and piezoelectric response can be recorded simultaneously. The piezoelectric response is generally measured as an amplitude and phase signal. The PFM amplitude signal corresponds to the distance the tip is deflected and is thus directly related to the deformation of the sample. The latter is determined by piezoelectric tensor coefficients (in the laboratory coordinate system), which in turn depend on the crystal structure of the material under investigation and the crystal orientation. The PFM phase signal is related to the phase-lag of the piezoelectric response of the material compared to the sinusoidal ac-bias used to induce the deformation. When the direction of the applied electric field is parallel to the polarization direction of the ferroelectric domain under investigation, the domain will expand. Conversely, when the electric field and polarization are antiparallel, the domain will contract. Thus, the PFM phase provides information about the orientation of the ferroelectric polarization. The nanoelectromechanics (tip bias- and strain-induced phenomena) of PFM is reported in detail elsewhere [10]. A simple schematic of PFM is shown in Fig. 1. Note that similar to contact mode AFM, PFM can be operated in air, liquid, and vacuum environments.

For materials with general crystallographic orientation, the surface deformation is not necessarily purely out-of-plane (Fig. 1b), and the surface can shift laterally (shear deformation) depending on the polarization direction and the piezoelectric coefficients (Fig. 1c). This shear motion results in cantilever torsion for a tip in contact with the sample surface. Thus, in order to develop a full picture of the polarization orientation, it is important to measure both the tip deflection (vertical PFM) and torsion (lateral PFM) [11] (Fig. 1). In such a way, it is possible to reconstruct the three-dimensional real-space polarization direction in the material. An example of such vector PFM [12] imaging of an epitaxial bismuth ferrite film



Piezoresponse Force Microscopy and Spectroscopy, Fig. 1 (**a**) Schematics of the PFM experimental setup and (**b**) out-of-plane and (**c**) in-plane sample deformations which result in cantilever deflection (t-b signal output) and torsion (l-r output), respectively



Piezoresponse Force Microscopy and Spectroscopy, Fig. 2 (**a**) Topography and (**b**) vector PFM images of a bismuth ferrite sample. The *dashed line* in (**a**) indicates a region subjected to a dc-bias prior to imaging. The *color wheel* in (**b**) indicates the ferroelectric polarization orientation and magnitude (Reproduced with permission from Ref. [13], copyright 2010, Wiley-VCH Verlag GmbH & Co. KGaA)

is shown in Fig. 2. Note that torsion can occur on domain walls, and, e.g., even 180° domain walls in tetragonal (001)-oriented material can be visible in LPFM amplitude images.

Domain Writing by PFM

In PFM imaging, an ac-bias is applied to the tip. This allows for local probing of the piezoelectric

properties and for ferroelectric domain orientation imaging. The application of a dc-bias via the tip to the sample can cause local polarization reversal if the highly localized electric field at the tip exceeds the local coercive field of the material. Thus, PFM can be used for local domain writing and for subsequent imaging of the polarization reversal. Dense arrays of written domains have been envisioned for polarization-based information storage. Information bits can be written as small domains with opposite polarization, and the read cycle can be achieved using the same tip by measuring the amplitude and the phase of the piezoelectric response. These memory devices are expected to achieve high densities, limited only by the size of the written ferroelectric domains and the presence of possible cross talk [14].

Polar surfaces of ferroelectrics are characterized by the presence of surface charges and electric potential and can be used in a broad set of applications, including local molecular self-assembly, physical adsorption/desorption, chemical reactions, and charge transfer reactions, and for the manipulation of individual organic molecules. Exploring the effect of ferroelectric polarization on chemical reactivity, an approach for assembling complex nanostructures consisting of oxide substrates, metal nanoparticles, and organic/biological molecules [15, 16] has been demonstrated (ferroelectric lithography). Controllable and selective deposition of molecular and metal species onto nanoscale domain-patterned ferroelectric templates may provide an alternative route to lithographic fabrication methods.

PFM Versus Other SPMs

The unique aspect of PFM is that for piezoelectric materials in the classical continuous mechanics regime, the electromechanical surface displacement is independent on the contact area. In other words, the displacement produced by the application of 1 V at the center of a 1 m, 1 mm, and 10 nm electrode will be the same. This behavior is fundamentally different compared to, e.g., contact mode mechanical AFMs, in which adhesion force, contact stiffness, and other parameters describing the properties of the tip-surface junction scale linearly with the tip-sample contact size. Correspondingly, on non-flat surfaces, the effects of material properties and topography variation cannot be unambiguously decoupled (direct topographic cross talk). In PFM, the response is intrinsically quantitative, albeit it can still be affected by the position-dependent transfer function of the cantilever (indirect cross talk), the tip sliding along the sample surface, etc.

Similarly to other force-based SPM techniques, the signal-to-noise ratio in PFM can be greatly improved if imaging is performed close to the contact resonance frequency of the cantilever. The quality factor of the resonances can be as high as ~100, allowing for a two-order increase in signal strength compared to the off-resonance case. Quantitative PFM measurements can be performed if the frequency of the electric signal driving the cantilever is adjusted to follow the contact resonance at each point on the surface. However, classical strategies based on phase-locked loop detection are inapplicable to PFM, since the phase of the local response depends on the local polarization direction. In the past several years, techniques have been developed to enhance the signal-to-noise ratio in PFM and to quantify the piezoresponse signal, including dual-amplitude resonance tracking (DART) PFM [17] and band excitation (BE) PFM [18]. In DART-PFM, two ac-biases are sent to the tip, with frequencies chosen such that one is on each side of the contact resonance. A feedback loop is employed to keep the difference in the amplitude measured at each frequency equal to zero, allowing the contact resonance to be tracked, i.e., remain in between the two frequencies during scanning. In BE-PFM, the cantilever is electrically excited across a band of frequencies, typically chosen to contain the contact resonance peak at all surface locations, allowing the resonant frequency and the quality factor of the system to be measured directly.

PFM Spectroscopies

PFM can be extended to spectroscopic modes, in which the response is measured as a function of (i) time during or after the application of a

bias pulse (time spectroscopy), (ii) bias during a triangular sweep (voltage spectroscopy), and (iii) ac-bias amplitude. More complex measurement sequences (e.g., first-order reversal curve (FORC) measurements) have also been demonstrated. These spectroscopic modes can in turn be implemented as spectroscopic imaging modes (similar to, e.g., force volume mode imaging in AFM), in which the spectroscopic sweeps are performed at every point in an image.

Switching Spectroscopy Piezoresponse Force Microscopy

In switching spectroscopy PFM (SS-PFM), the ac-bias is superimposed onto a dc-bias waveform, which cycles between a maximum and minimum dc voltage value (i.e., a triangular voltage sweep). SS-PFM allows local piezoelectric hysteresis loops to be measured at each pixel in an image, providing a nanoscale analog to macroscopic ferroelectric polarization-voltage loops. Each loop can be analyzed to determine the width of the loop, the nucleation biases, the remanent polarization, etc. These extracted switching parameters can then be plotted as two-dimensional maps, allowing correlations between switching properties and local topography to be identified. Note that SS-PFM can be operated in single-frequency, DART, and BE-PFM modes.

As an example of the SS-PFM imaging mode, maps of piezoelectric response and switching properties for $BiFeO_3$-$CoFe_2O_4$ (BFO-CFO) heterostructures are shown in Fig. 3. The BFO film is ferroelectric, whereas the CFO nanopillars are non-ferroelectric (and ferromagnetic). The ferroelectric behavior is clearly seen in the PFM amplitude image in Fig. 3b, in which the BFO regions are bright (i.e., have high piezoresponse) and show clearly visible domain walls, whereas the response from the CFO pillars is negligible. In the SS-PFM mode, the hysteresis loops are acquired over the same region on a rectangular grid. Shown in Fig. 3c is the work of switching map (i.e., the area under the hysteresis loop). Note that the loops are open (Fig. 3d) when measured on the BFO matrix and are closed when measured on the CFO nanopillars. The SS-PFM image illustrates that the polarization switching within BFO is very uniform, indicative of a highly uniform material, which is not affected by proximity to the second-phase inclusions.

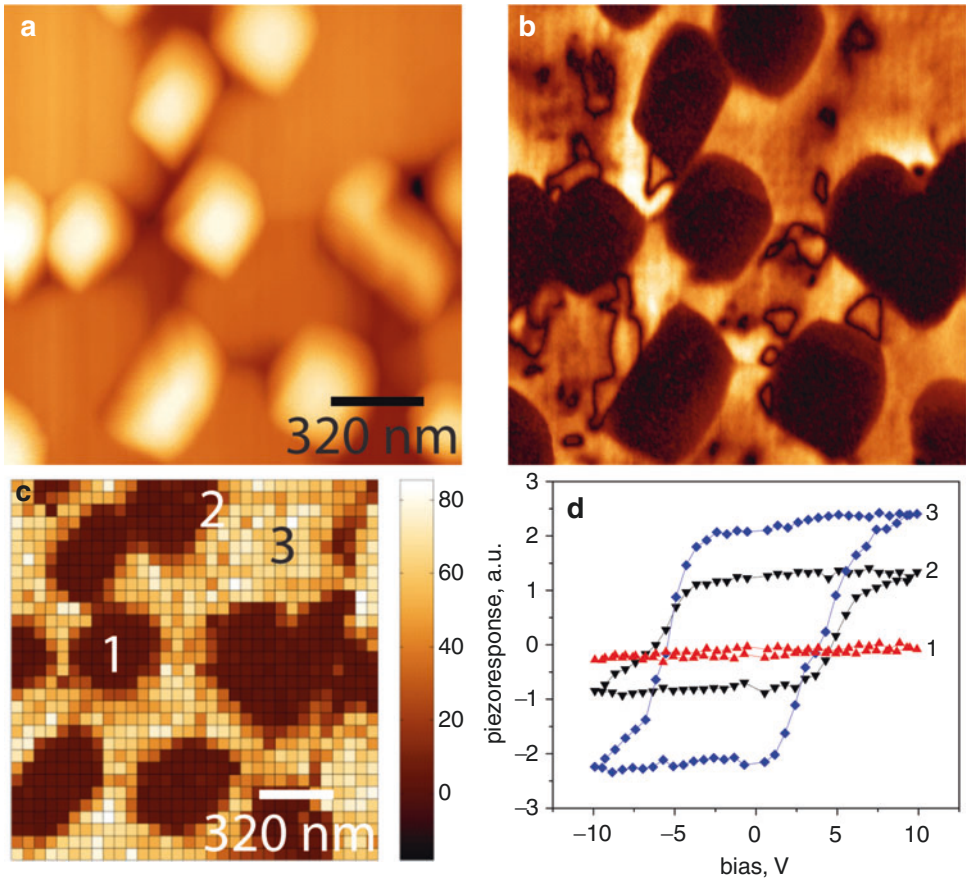
Time Spectroscopy in PFM

PFM can be extended to study time-dependent dynamic properties of materials. In time-resolved (TR) PFM, the local piezoelectric response is measured during the application of a dc-bias pulse and upon the completion of the pulse. As with SS-PFM, the response curves can be modeled to reveal information about local relaxation behavior of polymers and ferroelectric relaxors. Note that TR-PFM can also be operated in single-frequency, DART, and BE-PFM modes.

An example of TR-PFM is shown in Fig. 4 for the ferroelectric PMN-PT (solid solution of lead magnesium niobate (PMN) and lead titanate (PT)) relaxor. Here, polarization response at each point following the application of voltage pulse is collected. The response curves are fitted by the logarithmic function, $PR(t) = B_0 + B_1 \ln t$, and the spatial maps of offset, $B_0(x, y)$, and slope, $B_1(x, y)$, are shown in Fig. 4e, f, respectively. Note that both maps indicate the presence of mesoscopic structures and contain a number of uncorrelated features, indicative of the validity of the analysis and demonstrating the presence of intrinsic *dynamic* inhomogeneities within the material.

More Complex Spectroscopies

The characteristic aspect of hysteretic systems is the fact that the response is generally history dependent. For example, for a linear or nonlinear resistor, each value of current corresponds to only one value of applied external voltage. In a hysteretic system, the instantaneous value of current depends on the voltage history. However, the knowledge of the hysteresis loop for one type of field sweep does not allow for the prediction of a different one – e.g., knowledge of the hysteresis behavior in the (−10 V,10 V) bias window is insufficient to describe the (0,10) or (−5,10) bias windows. For certain classes of systems, this limitation can be bypassed using FORC measurements that allow one to reconstruct the response for an arbitrary field history. Another aspect of



Piezoresponse Force Microscopy and Spectroscopy, Fig. 3 (**a**) Topography of BFO-CFO and (**b**) corresponding PFM amplitude. (**c**) SS-PFM map of the work of switching and (**d**) representative loops from CFO (*1*), the CFO-BFO heterostructure interface (*2*), and BFO (*3*) (Reproduced with permission from Ref. [19], copyright 2007, Institute of Physics)

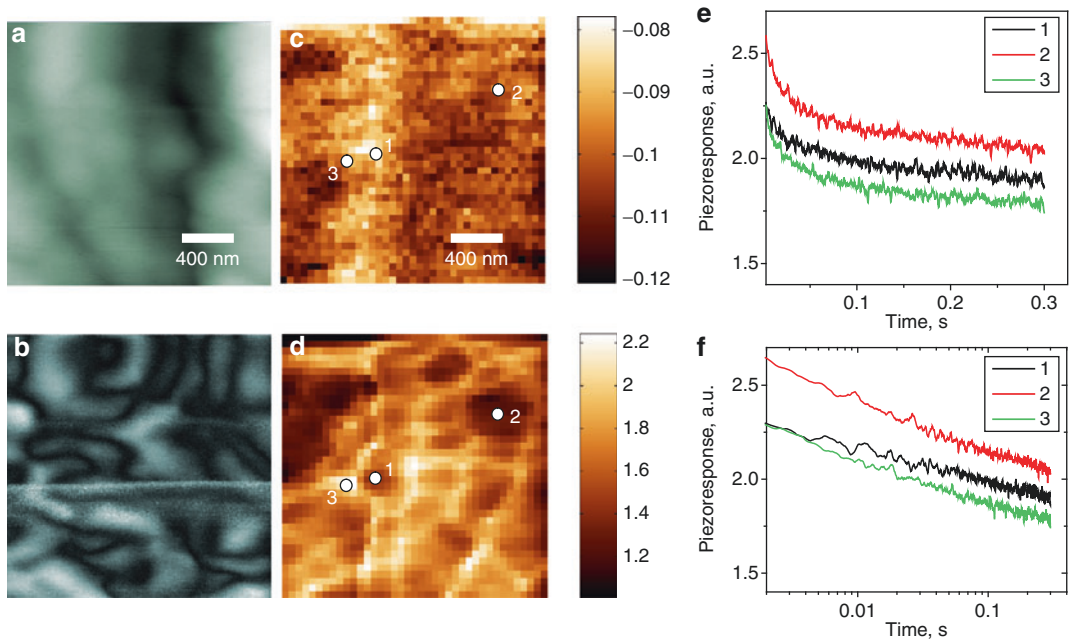
these materials is that the response can be time dependent, necessitating the study of the frequency or time dependence of the signal. To address these phenomena locally, more complicated time- and voltage-dependent spectroscopies in PFM are being developed.

PFM of Functional Materials

The materials that can be imaged by PFM are by definition functional materials in that they exhibit electromechanical coupling – they mechanically deform in an applied field. The classes of materials that have been investigated by PFM include ferroelectrics, biopolymers and polysaccharides, and piezoelectric polymers. Many recent advances in investigating functional materials using PFM have been published in Special Topics issues of the *Journal of Applied Physics* [21].

Ferroelectrics

Ferroelectric materials are a subset of pyroelectric materials, which in turn are a subset of piezoelectric materials. Piezoelectric materials exhibit linear electromechanical coupling, while pyroelectric materials have a permanent spontaneous polarization. In ferroelectric materials, the direction of this spontaneous polarization can be changed, i.e., switched between stable directions corresponding to different ferroelectric domains. These domains are readily imaged by PFM in

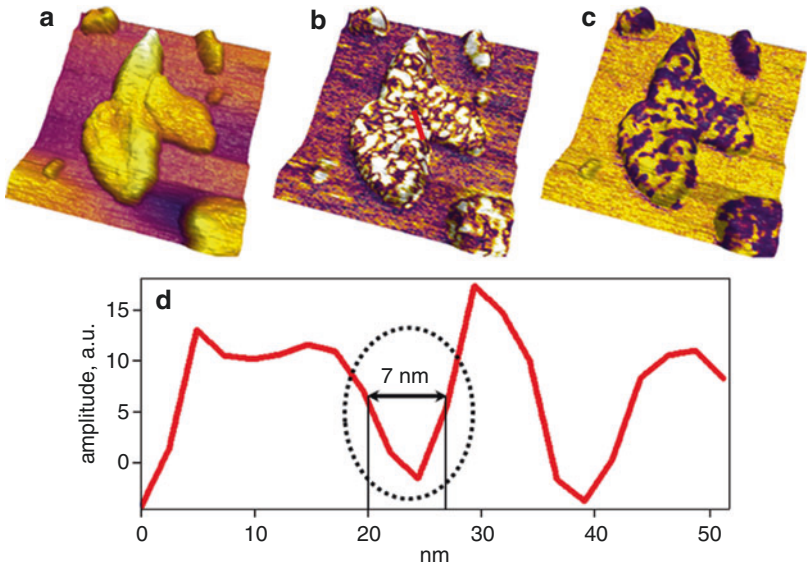


Piezoresponse Force Microscopy and Spectroscopy, Fig. 4 (**a**) Topography and (**b**) PFM amplitude images of a PMN-PT ferroelectric relaxor crystal. Spatially resolved maps of (**c**) slope, B_1, and (**d**) offset, B_0, of the logarithmic relaxation law ($2 \times 2\ \mu m^2$). Relaxation curves in selected locations in (**e**) linear and (**f**) logarithmic scale (Reproduced with permission from Ref. [20], copyright 2009, American Institute of Physics)

various ferroelectric families such as perovskites, Aurivillius layered structures, tungsten bronze ferroelectrics, hexagonal manganites, and organic materials and, recently, in a number of multiferroics (i.e., materials having both ferroelectricity and ferromagnetism simultaneously) and various composites [8, 22]. These measurements are indispensable for understanding the functionality of ferroelectrics at the nanoscale (e.g., within a single grain in polycrystalline ferroelectrics) and various size effects in as-grown and structured materials [23]. A ferroelectric relaxor is a special case of a ferroelectric material where long-range polarization is disrupted by defects and very small mesoscopic structures are observed on the surface on a length scale much smaller than the "normal" micron-sized ferroelectric domains in these materials. These properties have recently been accessed by time-, temperature-, and depth-dependent PFM measurements, demonstrating the power and capabilities of PFM to uncover the polarization mechanism in complex materials with coexisting order parameters and disorder [24].

Biosystems

The piezoelectric properties of biopolymers and polysaccharides have been studied on the macroscopic scale for some time, and several researchers have speculated on a relation between piezoelectricity in these materials and biological functionality such as bone remodeling. PFM is ideally suited for such studies and has been applied to a wide variety of biomaterials and in amino acid microcrystals [25]. These measurements are found to be especially important in macromolecules such as collagen fibers and, recently, in bioinspired peptide nanotubes [26]. The exceptional piezoelectric properties of the latter observed by PFM have renewed the interest in artificial biomaterials for biological sensor- and actuator-based applications. It has been also shown that PFM can be used as a high-resolution nanoscale tool for uncovering



Piezoresponse Force Microscopy and Spectroscopy, Fig. 5 Ferroelectric domains in a PVDF nanomesa probed by PFM. (**a**) Topography, (**b**) PFM amplitude, and (**c**) PFM phase images. (**d**) The resolution in (**b**) is found to be 7 nm (Reproduced with permission from Ref. [28], copyright 2011, American Chemical Society)

new features in complex biological materials, including cells and tissues, such as aortic walls [27].

Polymers

Many synthetic polymers, including polypropylene, polystyrene and poly(methyl methacrylate), semicrystalline polyamides, and amorphous polymers such as vinyl acetate, are piezoelectric. However, piezoelectric effects in these materials are relatively weak and often unstable and, therefore, are of limited practical significance. Strong and stable piezoelectric properties have primarily been found in the synthetic polymer polyvinylidene fluoride (PVDF) and PVDF copolymers. PFM has been found to be extremely useful for studying local piezoelectric properties in PVDF because it allows one to relate the surface topography to the local polarization response in these semicrystalline materials where a mixture of material phases is often observed. It has also been demonstrated that PFM can effectively probe nanostructured PVDF nanomesas (Fig. 5). Studying polarization switching in these soft ferroelectric materials is one of the challenges for PFM because of the possible coupling present between electrostatics, piezoelectricity, and elasticity.

PFM of Device Structures

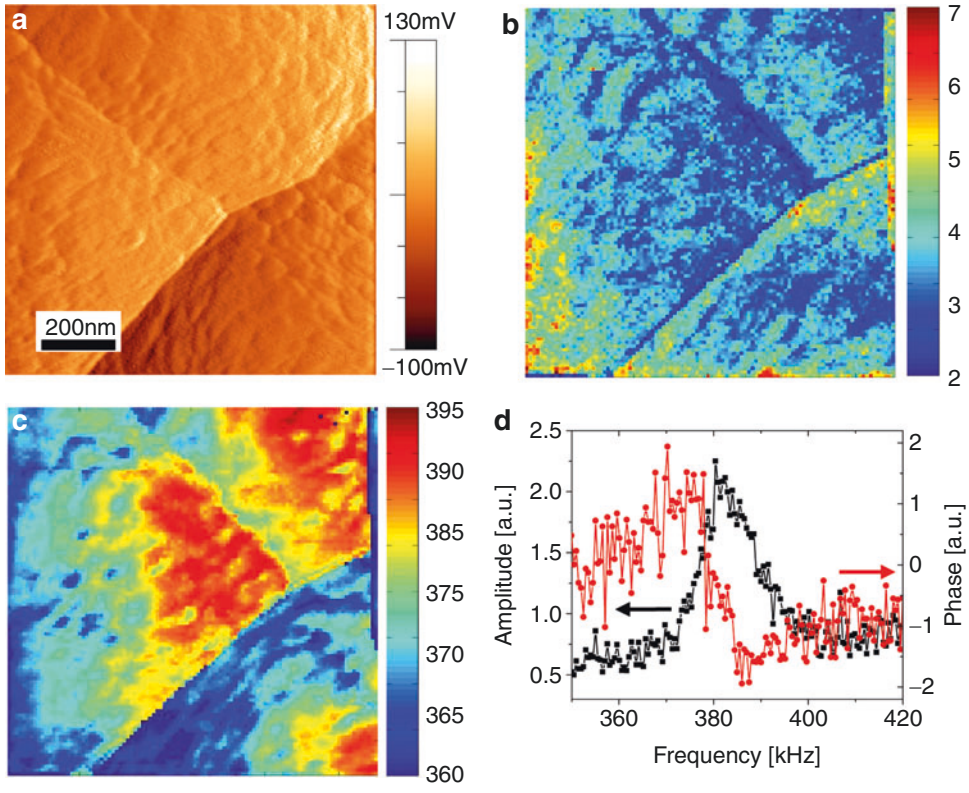
PFM is essentially a nanoscale polarization probe and, as such, can be used to study ferroelectric devices based on polarization dynamics and switching, including nonvolatile ferroelectric random access memories (NV-FRAM), thin-film capacitors, piezoelectric elements for microelectromechanical systems (MEMS), and novel multiferroic devices with multiple memory states [29]. As for NV-FRAM, the bit value is represented by one of the two distinguishable remanent polarization states of a ferroelectric material in a capacitor configuration. The spatial variation of the ferroelectric properties, e.g., the static and dynamic domain configurations, becomes critical as the NV-FRAM capacitor size decreases. If the ferroelectric properties of a material vary on a micron scale, then micron-sized capacitors would exhibit a significant capacitor-to-capacitor variation in their electrical properties, often resulting in a failure of the device. These properties have been extensively studied by a number of groups [30–33], and it has been shown that the polarization dynamics differ significantly as the lateral size of the ferroelectric capacitors is scaled down. Analogously, ferroelectric-based MEMS have been found to

be sensitive to scaling in both lateral and vertical dimensions. This makes PFM a useful industrial tool, e.g., for quality control in the production of memory chips and micromechanical devices based on the piezoelectric effect.

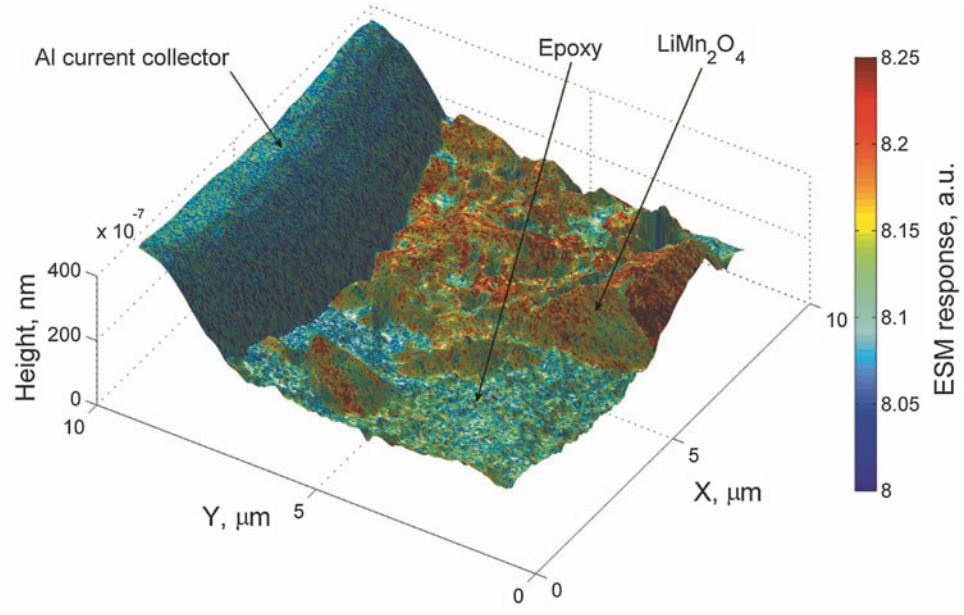
Electrochemical Strain Microscopy

Bias-strain coupling can be driven by electrochemical mechanisms as a consequence of ionic motion in applied electric fields. PFM can be used to probe ac-bias-induced ionic motion in energy materials, giving rise to electrochemical strain microscopy (ESM) [34–37]. In ESM, a biased SPM tip concentrates an electric field in a nanometer-scale volume of the material, inducing interfacial electrochemical processes at the tip-surface junction and ionic currents through the solid. The intrinsic link between the concentration of ionic species and/or the oxidation states of the host cation and the molar volume of the material results in electrochemical strain and surface displacement.

The ESM imaging mode is based on detecting the strain response of a material to an applied electric field through a blocking or electrochemically active SPM tip. The high-resolution (<10 nm) data shown in Fig. 6 demonstrates ESM imaging on a Li-intercalated Si-anode surface [38]. A full suite of spectroscopic imaging modes can also be implemented in ESM. To date, ESM has been demonstrated for a variety of lithium-ion materials, including layered transition metal oxide cathodes, silicon anodes, and electrolytes such as LISICON; oxygen electrolytes,



Piezoresponse Force Microscopy and Spectroscopy, Fig. 6 Mapping of the electrochemical strain response in a Si anode. (**a**) Deflection signal of a 1 × 1 μm area showing a triple boundary point. (**b**) Contact resonance amplitude and (**c**) resonance frequency map showing heterogeneous strain response and a strong correlation between resonance frequency and topography. (**d**) A single-point contact resonance peak from the boundary region. The data was acquired in BE mode (Adapted from Ref. [38])



Piezoresponse Force Microscopy and Spectroscopy, Fig. 7 Electrochemical activity of the cross section of a commercial Li battery. The ESM signal is inhomogeneous in Li_2MnO_4 particles embedded in epoxy. The data was acquired in single-frequency mode (Copyright © 2014, Cambridge University Press, all rights reserved. Permission to reprint Fig. 6 from Ref. [39] has been granted)

including yttria-stabilized zirconia and samarium-doped ceria; mixed electronic-ionic conductors for fuel cell cathodes; and some proton conductors. Even cross sections of commercial Li batteries have been recently studied [39]. An example of ESM of a Li battery composite cathode is shown in Fig. 7, where active Li_2MnO_4 particles demonstrate much higher ESM signal as compared to the other parts of the composite. Inhomogeneous ESM signal distribution is directly related to the Li activity within the grains providing nanometer resolution. Such studies should shed light on the mechanisms of degradation related to the loss of battery charging capability.

The ability to probe electrochemical processes and ionic transport in solids is invaluable for a broad range of applications for energy generation and storage (e.g., batteries and fuel cells). The viability of electric vehicles and grid storage as key components of a renewable energy technology hinges on advances in battery energy densities and lifetimes. An example of ESM has the potential to aid in these advances with two major improvements over existing technologies: (a) the resolution to probe nanometer-scale volumes and (b) imaging capability extended to a broad range of spectroscopy techniques. In addition, because electrochemical strains are ubiquitous in virtually all solid-state ionics, ESM is applicable to all battery and fuel cell materials in energy technologies and electroresistive/memristive materials in information technologies.

Summary

Electromechanical coupling is ubiquitous in systems ranging from piezoelectrics and ferroelectrics to polar macromolecules and solid-state ionics. Piezoresponse force microscopy and spectroscopy are the techniques that have opened these phenomena to scientific inquiry, providing advances ranging from sub-10 nm domain imaging and lithography to probing intricate polarization structures in ferroelectric relaxors, polar polymers, and biosystems and to revealing details of electrochemical processes in battery materials. The future of this field clearly contains many more serendipitous discoveries.

References

1. Güthner, P., Glatz-Reichenbach, J., Dransfeld, K.: Investigation of local piezoelectric properties of thin copolymer films. J. Appl. Phys. **69**, 7895 (1991)
2. Birk, H., Glatz-Reichenbach, J., Jie, L., Schreck, E., Dransfeld, K.: The local piezoelectric activity of thin polymer films observed by scanning tunneling microscopy. J. Vac. Sci. Technol. B **9**, 1162 (1991)
3. Kolosov, O., Gruverman, A., Hatano, J., Takahashi, K., Tokumoto, H.: Nanoscale visualization and control of ferroelectric domains by atomic force microscopy. Phys. Rev. Lett. **74**, 4309 (1995)
4. Gruverman, A., Auciello, O., Tokumoto, H.: Imaging and control of domain structures in ferroelectric thin films via scanning force microscopy. Annu. Rev. Mater. Sci. **28**, 101 (1998)
5. Alexe, M., Gruverman, A. (eds.): Nanoscale Characterization of Ferroelectric Materials. Springer, Heidelberg (2004)
6. Hong, S. (ed.): Nanoscale Phenomena in Ferroelectric Thin Films. Kluwer, Dordrecht (2004)
7. Kalinin, S.V., Gruverman, A. (eds.): Scanning Probe Microscopy of Electrical and Electromechanical Phenomena at the Nanoscale. Springer, Berlin (2007)
8. Gruverman, A., Kholkin, A.: Nanoscale ferroelectrics: processing, characterization and future trends. Rep. Prog. Phys. **69**, 2443 (2006)
9. Kalinin, S.V., Rodriguez, B.J., Jesse, S., Karapetian, E., Eliseev, E.A., Morozovska, A.N.: Nanoscale electromechanics of ferroelectric and biological systems: a new dimension in scanning probe microscopy. Annu. Rev. Mater. Res. **37**, 189 (2007)
10. Kalinin, S.V., Morozovska, A.N., Chen, L.-Q., Rodriguez, B.J.: Local polarization dynamics in ferroelectric materials. Rep. Prog. Phys. **73**, 056502 (2010)
11. Eng, L.M., Abplanalp, M., Günter, P.: Mapping the domain distribution at ferroelectric surfaces by scanning force microscopy. Appl. Phys. A **66**, S679 (1998)
12. Kalinin, S.V., Rodriguez, B.J., Jesse, S., Shin, J., Baddorf, A.P., Gupta, P., Jain, H., Williams, D.B., Gruverman, A.: Vector piezoresponse force microscopy. Microsc. Microanal. **12**, 206 (2006)
13. Rodriguez, B.J., Choudhury, S., Chu, Y.H., Bhattacharyya, A., Jesse, S., Seal, K., Baddorf, A.P., Ramesh, R., Chen, L.Q., Kalinin, S.V.: Unraveling deterministic mesoscopic polarization switching mechanisms: spatially resolved studies of a tilt grain boundary in bismuth ferrite. Adv. Funct. Mater. **19**, 2053–2063 (2009)
14. Tybell, T., Paruch, P., Giamarchi, T., Triscone, J.-M.: Domain wall creep in epitaxial ferroelectric $Pb(Zr_{0.2}\text{-}Ti_{0.08})O_3$ thin films. Phys. Rev. Lett. **89**, 097601 (2002)
15. Kalinin, S.V., Bonnell, D.A., Alvarez, T., Lei, X., Hu, Z., Ferris, J.H., Zhang, Q., Dunn, S.: Atomic polarization and local reactivity on ferroelectric surfaces: a new route toward complex nanostructures. Nano Lett. **2**, 589 (2002)
16. Heredia, A., Bdikin, I.K., Machado, M., Yudin, S., Fridkin, V.M., Delgadillo, I., Kholkin, A.L.: Preferred deposition of phospholipids onto ferroelectric P (VDF-TrFE) films via polarization patterning. J. Phys. D **43**, 335301 (2010)
17. Rodriguez, B.J., Callahan, C., Kalinin, S.V., Proksch, R.: Dual-frequency resonance-tracking atomic force microscopy. Nanotechnology **18**, 475504 (2007)
18. Jesse, S., Kalinin, S.V., Proksch, R., Baddorf, A.P., Rodriguez, B.J.: The band excitation method in scanning probe microscopy for rapid mapping of energy dissipation on the nanoscale. Nanotechnology **18**, 435503 (2007)
19. Rodriguez, B.J., Jesse, S., Baddorf, A.P., Zhao, T., Chu, Y.H., Ramesh, R., Eliseev, E.A., Morozovska, A.N., Kalinin, S.V.: Spatially resolved mapping of ferroelectric switching behavior in self-assembled multiferroic nanostructures: strain, size, and interface effects. Nanotechnology **18**, 405701 (2007)
20. Kalinin, S.V., Rodriguez, B.J., Jesse, S., Morozovska, A.N., Bokov, A.A., Ye, Z.-G.: Spatial distribution of relaxation behavior on the surface of a ferroelectric relaxor in the ergodic phase. Appl. Phys. Lett. **95**, 142902 (2009)
21. Prefaces to "Special topic: piezoresponse force microscopy. J. Appl. Phys. **118**, 07901 (2015) *ibid.* **116**, 066701 (2014); *ibid.* **113**, 187101 (2013); *ibid.* **112**, 051901 (2012); *ibid.* **110**, 051901 (2011); *ibid.* **108**, 041901 (2010)
22. Balke, N., Bdikin, I.K., Kalinin, S.V., Kholkin, A.L.: Electromechanical imaging and spectroscopy of ferroelectric and piezoelectric materials: state-of-the-art and prospects for the future. J. Am. Ceram. Soc. **92**, 1629 (2009)
23. Kholkin, A.L., Bdikin, I.K., Kiselev, D.A., Shvartsman, V.V., Kim, S.-H.: Nanoscale piezoelectric characterization of polycrystalline ferroelectrics. J. Electroceramics **19**, 81 (2007)
24. Kholkin, A.L., Kiselev, D.A., Bdikin, I.K., Rodriguez, B.J., Wu, P., Bokov, A.A., Ye, Z.-G., Dkhil, B., Chen, L.-Q., Kosec, M., Kalinin, S.V.: Surface domain structures and mesoscopic phase transition in relaxor ferroelectrics. Adv. Funct. Mater. **21**, 1977 (2011)
25. Heredia, A., Meunier, V., Bdikin, I.K., Gracio, J., Balke, N., Jesse, S., Tselev, A., Agarwal, P., Sumpter, B.G., Kalinin, S.V., Kholkin, A.L.: Nanoscale ferroelectricity in crystalline γ-glycine. Adv. Funct. Mater. **22**, 2996 (2012)
26. Kholkin, A.L., Amdursky, N., Bdikin, I., Rosenman, G., Gazit, E.: Strong piezoelectric activity in peptide nanotubes. ACS Nano **4**, 610 (2010)
27. Liu, Y., Zhang, Y., Chow, M.-J., Chen, Q.N., Li, J.: Biological ferroelectricity uncovered in aortic walls by piezoresponse force microscopy. Phys. Rev. Lett. **108**, 078103 (2012)
28. Sharma, P., Reece, T.J., Ducharme, S., Gruverman, A.: High-resolution studies of domain switching behavior in nanostructured ferroelectric polymers. Nano Lett. **11**, 1970 (2011)

29. Kalinin, S.V., Setter, N., Kholkin, A.L. (eds.): Electromechanics on the nanometer scale: emerging phenomena, devices, and applications. MRS Bull. **34**, 634 (2009)
30. Gruverman, A., Rodriguez, B.J., Kingon, A.I., Nemanch, R.J., Cross, J.S., Tsukada, M.: Spatial inhomogeneity of imprint and switching behavior in ferroelectric capacitors. Appl. Phys. Lett. **82**, 3071 (2003)
31. Stolichnov, I.A., Tagantsev, A.K., Setter, N., Cross, J. S., Tsukada, M.: Top-interface-controlled switching and fatigue endurance of (Pb, La)(Zr, Ti)O_3 ferroelectric capacitors. Appl. Phys. Lett. **74**, 3552 (1999)
32. Kim, D.J., Jo, J.Y., Kim, T.H., Yang, S.M., Chen, B., Kim, Y.S., Noh, T.W.: Observation of inhomogeneous domain nucleation in epitaxial Pb(Zr, Ti)O_3 capacitors. Appl. Phys. Lett. **91**, 132903 (2007)
33. Kim, Y., Lu, X., Jesse, S., Hesse, D., Marin, A., Kalinin, S.V.: Universality of polarization switching dynamics in ferroelectric capacitors revealed by 5D piezoresponse force microscopy. Adv. Func. Mater. **23**, 3971 (2013)
34. Bonnell, D.A., Kalinin, S.V. (eds.): Scanning Probe Microscopy for Energy Research. World Scientific, New Jersey (2013)
35. Balke, N., Jesse, S., Morozovska, A.N., Eliseev, E., Chung, D.W., Kim, Y., Adamczyk, L., García, R.E., Dudney, N., Kalinin, S.V.: Nanoscale mapping of ion diffusion in a lithium-ion battery cathode. Nat. Nanotechnol. **5**, 749 (2010)
36. Kumar, A., Ciucci, F., Morozovska, A.N., Kalinin, S. V., Jesse, S.: Measuring oxygen reduction/evolution reactions on the nanoscale. Nat. Chem. **3**, 707 (2011)
37. Jesse, S., Vasudevan, R.K., Collins, L., Strelcov, E., Okatan, M.B., Belianinov, A., Baddorf, A.P., Proksch, R., Kalinin, S.V.: Band excitation in scanning probe microscopy: recognition and functional imaging. Annu. Rev. Phys. Chem. **65**, 519 (2014)
38. Balke, N., Jesse, S., Kim, Y., Adamczyk, L., Tselev, A., Ivanov, I.N., Dudney, N.J., Kalinin, S.V.: Real space mapping of Li-ion transport in amorphous Si anodes with nanometer resolution. Nano Lett. **10**, 3420 (2010)
39. Luchkin, S., Romanyuk, K., Ivanov, M., Kholkin, A.L.: Single- and multi-frequency detection of surface displacements via scanning probe microscopy. Microsc. Microanal. **21**, 154 (2015)

Placenta

► Use of Nanotechnology in Pregnancy

Plasma

► Physical Vapor Deposition

Plasma Etching

► Dry Etching Processes

Plasma Grafting

► Nanostructures for Surface Functionalization and Surface Properties

Plasma Jet Machining (PJM)

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Plasma Micro-contact Patterning

► Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Plasma Patterning

► Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Plasma Polymerization

► Nanostructures for Surface Functionalization and Surface Properties

Plasma-Assisted Chemical Etching (PACE)

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Plasma-Enhanced Chemical Vapor Deposition (PECVD)

▶ Chemical Vapor Deposition (CVD)

Plasmon Resonance Energy Transfer from Metallic Nanoparticles to Biomolecules

▶ Plasmon Resonance Energy Transfer Nanospectroscopy

Plasmon Resonance Energy Transfer Nanospectroscopy

Gang Logan Liu
Micro and Nanotechnology Laboratory, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Synonyms

Plasmon resonance energy transfer from metallic nanoparticles to biomolecules

Definition

Plasmon resonance energy transfer is the energy stored in the collective movement of free electrons in metallic nanoparticles being transferred to the adsorbed chemical and biomolecules with match electronic transition energy.

Overview

Nanoparticle Plasmon Resonance

A plasmon is defined as the collective and periodic free electron movement in metallic structures under electromagnetic excitation. Free electrons in the metal conduction band are not bounded to the nucleus, so they can move in the applied alternating electromagnetic field within the physical boundary of metal materials producing an alternating electron current. Like the resonance modes in mechanical movements, there exist plasmon resonance modes with minimal energy damping. If the plasmon resonance modes fall into the optical frequency range, especially the visible frequency range, the resonating free electrons in metallic nanostructures form an optical dipole to reemit light.

The plasmon resonance modes are dependent on the dielectric constants of the metallic nanoparticle and the surrounding material. Within the optical frequency range, the complex dielectric constants or the permittivity of the metallic materials such as gold and silver changes rapidly with the light wavelength while the dielectric constants of dielectric materials remain nearly unchanged. The plasmonic polarizability has to be modified when the metallic structure approaches sub-100 nm scale. At this scale, the dimension of the metallic nanostructure becomes comparable to the mean free path length of the conduction electrons in metal. The physical boundary of the metallic nanostructures poses a boundary condition for the plasmon resonance mode. Nanoplasmonic structures are thus referred to the metallic nanostructures such as nanoparticles with discrete plasmon resonance modes which are dependent on the nanostructure geometries and surrounding medium.

Plasmon Resonance Energy Transfer

Nanoparticle plasmon resonance [1, 2] is distinctive from the propagating surface plasmon resonance on metallic thin film [3]. It is spatially confined within the physical boundary of the nanoparticle; however it can be continuously transferred to adjacent metallic nanoparticles [4, 5] through plasmon coupling. It has also been conjectured for a long time that the plasmon resonance energy can be possibly transferred to chemical or biological molecules adsorbed on metallic nanostructures [6, 7]. This entry will review the highest-sensitivity-ever nanoscopic biomolecular absorption spectroscopy enabled

by plasmon resonance energy transfer (PRET) from a single metallic nanoparticle to a conjugated biomolecule and chemical molecule. Due to PRET, quantized plasmon quenching dips are observed in single nanoparticle scattering spectra which correspond to the absorption spectral peaks of conjugated biomolecules. The hybrid nanoparticle–biomolecule PRET system allows near single molecular sensitivity absorption spectroscopy with nanoscale spatial resolution for in vivo functional molecular imaging.

Although not directly observed before, PRET was accounted for one of possible explanations for surface-enhanced Raman scattering [8, 9] and fluorescence [10] on single nanoparticles. Experimentally the plasmon resonance of gold and silver nanoparticles conjugated with various biomolecules such as DNA [11], peptide [12], and biotin-streptavidin [13] has been studied by single particle Rayleigh scattering spectroscopy [14, 15]. All these previous studies demonstrated the shift of plasmon resonant wavelength by changing dielectric medium due to structural changes of biomolecule conjugated on the surface of single metallic nanoparticles. Since most of the previous cases have conjugated biomolecules with optical absorption or electronic resonance peaks in ultraviolet (UV) or far-infrared range on gold and silver nanoparticles with visible plasmon resonance peaks, only the shift of plasmon resonance peak was observed. However in this chapter, metalloprotein Cytochrome c (Cyt c) was conjugated on a single 30 nm gold nanoparticle in order to observe the direct quantized plasmon resonance energy transfer from the nanoplasmonic particle to Cyt c, as seen in Fig. 1.

Theory

The strict modeling of plasmon resonance energy transfer between metallic nanoparticle and adsorbed resonance chemical and biological molecules requires quantum mechanical calculation of electron interaction Hamiltonian between nanoparticle and molecules.

$$H_I = \frac{1}{4\pi\varepsilon_0} \times \int dr_{\text{molecule}} \int d^3 r_{\text{metal}} \frac{\rho(r_{\text{molecule}})\rho(r_{\text{metal}})}{|R - r_{\text{metal}} + r_{\text{molecule}}|} \tag{1}$$

The above equation describes the interaction Hamiltonian where ρ is the electron density distribution in molecule or metallic nanoparticle and r is the radial distance of the electrons from the center of the molecule or nanoparticle and R is the distance between the molecule center and nanoparticle center. Apparently it is extremely difficult to precisely describe the three-dimensional spatial distribution of electron density; therefore some assumption and approximation have to be made to carry out the analytical calculation. Fortunately near the matched nanoparticle plasmon resonance and molecule electronic resonance the electron distribution may be approximated as dipolar distributions. Then the electron interaction becomes dipole–dipole interactions as described in Eq. 2.

$$H_I^{\text{dip}} = \frac{1}{4\pi\varepsilon_0} \times \left[\frac{\mu_{\text{molecule}} \cdot \mu_{\text{metal}} - 3(\mu_{\text{molecule}} \cdot \hat{d}(\mu_{\text{metal}} \cdot \hat{d})}{d^3} \right] \tag{2}$$

Here μ is the dipole moment operator and defined as

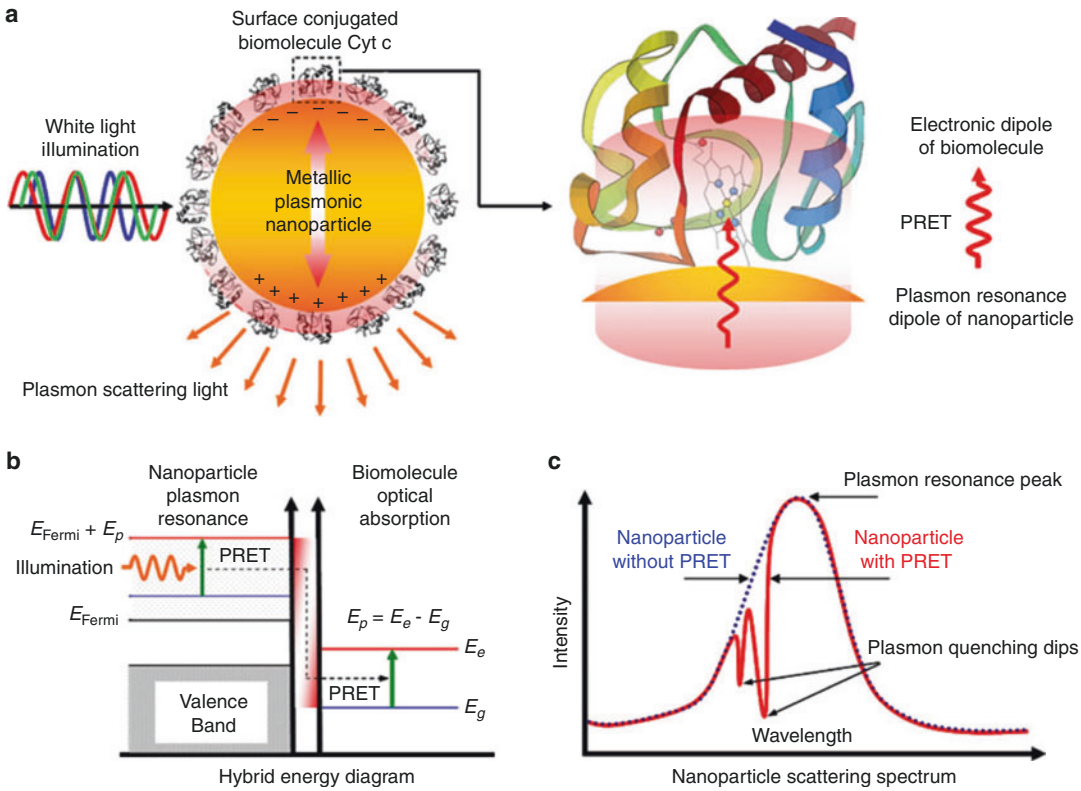
$$\mu_{\text{metal}} = \int r_{\text{metal}} \rho_{1,0}(r_{\text{metal}}) d^3 r_{\text{metal}}$$
$$\mu_{\text{molecule}} = \int r_{\text{molecule}} \rho(r_{\text{molecule}}) dr_{\text{molecule}} \tag{3}$$

Where ρ is simplified as the radial distribution of the electron density and especially for metallic nanoparticle $\rho_{1,0}$ becomes the dipolar plasmon charge density which is defined as

$$\nabla^2 \rho_{1,0}(r_{\text{metal}}) + \alpha^2 \rho_{1,0}(r_{\text{metal}}) = 0 \tag{4}$$

The constant α is the function of fundamental plasmon resonance mode frequency $\omega_{1,0}$ and bulk plasmon resonance frequency ωp as well as the electron velocity at Fermi energy vF. The expression is as the following:

P



Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 1 Schematic diagram of PRET-enabled nanoscopic biomolecular absorption spectroscopy. (**a**) PRET from a single metallic nanoparticle to surface-conjugated biomolecules. The wavelength-specific plasmon resonance (collective free electron oscillation) in metallic nanoparticle is excited by white light illumination. The plasmon resonance dipole can interact with the biomolecular dipole and transfer energy to biomolecules. (**b**) Hybrid energy diagram showing quantized energy transfer process. With optical excitation, the free electrons in the conduction band of metallic nanoparticle are elevated from Fermi to higher energy level forming resonating plasmon. The plasmon resonance energy is transferred to the metalloprotein biomolecules (i.e., Cyt c) conjugated on the nanoparticle surface when matched with the electronic transition energy in biomolecule optical absorption. (**c**) Typical Rayleigh scattering spectrum of the single PRET probe. The energy transition in PRET is represented as quenching dips in nanoparticle scattering spectrum, and the dip positions correspond to the biomolecule optical absorption peaks

$$\alpha^2 = \frac{{\omega_{1,0}}^2 - {\omega_p}^2}{3{v_F}^2/5} \tag{5}$$

The dipole moment operator of the metallic nanoparticle at the fundamental plasmon resonance mode is then

$$\mu_{\text{metal}} \approx \left(\frac{4\pi^3 \varepsilon_0 \hbar \omega_p}{9\sqrt{3}}\right)^{1/2} a^{3/2} \tag{6}$$

Basic Methodology

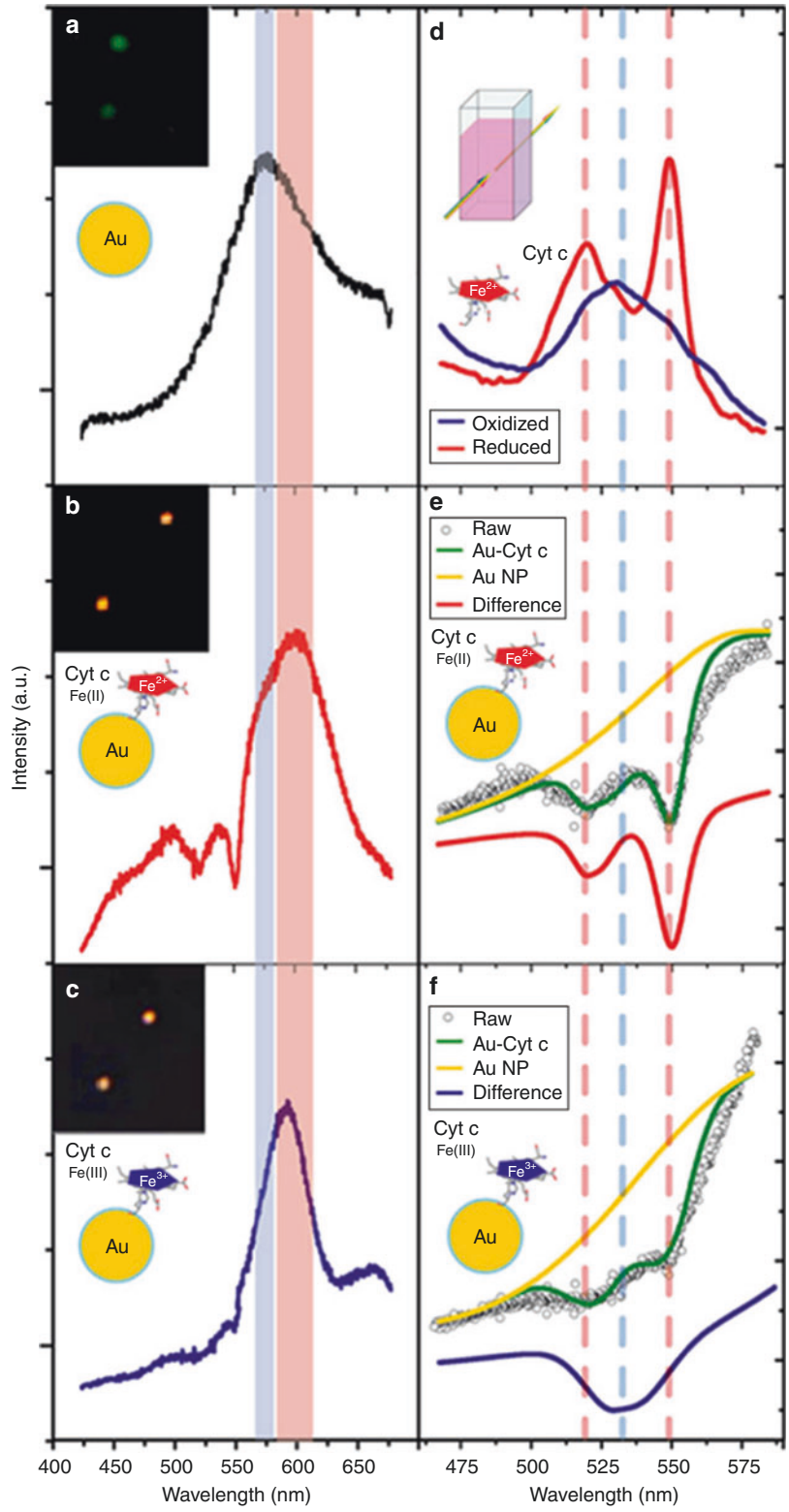
PRET-Enabled Biomolecular Absorption Spectroscopy

The intentional overlap of the absorption peak positions of desired biomolecules with the plasmon resonance peak of the metallic nanoparticle generates distinguishable spectral quenching dips on the Rayleigh scattering spectrum of single nanoparticle, which also allows near single

molecular level nanoscopic absorption spectroscopy (Figs. 1 and 2). Cyt c, a metalloprotein in cellular mitochondria membrane, acts as the charge transfer mediator [16] and plays a crucial role in bioenergy generation, metabolism, and cell apoptosis [17]. Unlike many other proteins, Cyt c has several optical absorption peaks in visible range around 550 nm coinciding with the 30 nm gold nanoparticle plasmon resonance, and more importantly it is a natural energy acceptor with electron tunneling channels [18]. Similar to the donor–acceptor energy matching in fluorescent (or Förster) resonance energy transfer (FRET) between two fluorophores, the critical matching of the localized resonating plasmon kinetic energy *Ep* in gold nanoparticles with the electron transition energy from ground to excited state *Ee*−*Eg* in Cyt c molecules permits the PRET process (Fig. 1b). The quantized energy is likely transferred through the dipole–dipole interaction between the artificial alternating dipole – resonating plasmon in nanoparticle and the biomolecular dipole. Previous work on surface plasmon–mediated FRET process [19, 20], superlens imaging [21], surface plasmon resonance shift of redox molecules [22], and very recent work on bulk optical extinction spectroscopy of nanoplasmonic particle clusters with conjugated resonant molecules [23] also indicate the possibility of such dipole–dipole interaction. Distinctively in this experiment, the plasmon energy quenching of nanoparticle due to PRET is represented as the "spectral dips" in the single nanoparticle scattering spectrum (Figs. 1c and 2) and the positions of dips match with the molecular absorption peak positions (Figs. 1c and 2). PRET is a direct energy transfer process and thus much more efficient and faster than optical energy absorption [10], so the absorption spectral peaks of conjugated Cyt c molecules on single gold nanoparticles can be detected with a simple optical system, which is otherwise impossible using conventional visible absorption spectroscopic methods.

The Cyt c conjugated 30 nm gold nanoparticles are dispersedly tethered on the surface of a transparent glass slide. The glass slide is mounted on a white light darkfield microscopy system with a true-color camera and a spectrometer to characterize the scattering image and spectrum of individual gold nanoparticles as well as the hybrid PRET probes (i.e., specific metallic nanoparticles with conjugated Cyt c molecules).

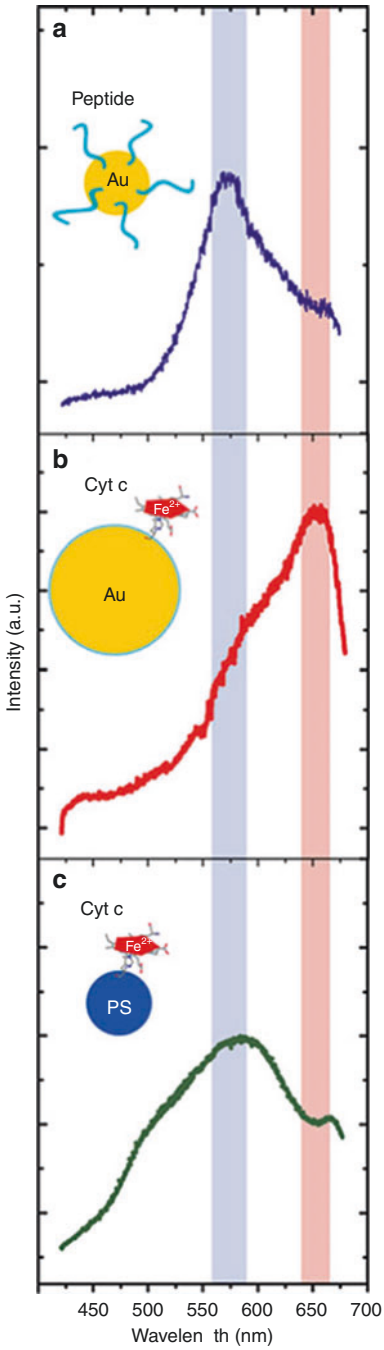
In comparison with the visible scattering spectrum of gold nanoparticles coated with only Cysteamine cross-linker molecules (Fig. 2a), the raw scattering spectra of gold nanoparticles conjugated with reduced (Fig. 2b) and oxidized Cyt c (Fig. 2c) show not only a scattering peak (plasmon resonance peak) but distinctive dips next to it. The spectral dips modulated on the nanoparticle scattering spectrum can be decoupled and converted to the visible absorption peaks of the reduced and oxidized Cyt c molecules (Fig. 2e and f). In accordance with the conventional visible absorption spectrum of bulk Cyt c solutions (Fig. 2d), the processed spectra have matched absorption peaks of reduced Cyt c around 525 and 550 nm, and oxidized Cyt c at 530 nm [24]. Although the near-field optical excitation efficiency from the nanoparticle scattering light is much higher than far field optical excitation (i.e., the photon scattered from the nanoparticle is more likely to transmit through and be absorbed by the surface conjugated biomolecules than those far away from the nanoparticle), the optical absorption at 550 nm by the ferrocytochrome c molecule monolayer only accounts for 0.03 % of the nanoparticle scattering light even for 100 % excitation efficiency (The absorption coefficient ε of horse heart Cyt c at 550 nm is 20.4 mM^{-1} cm^{-1}. For a completely packed monolayer of ~3 nm-in-diameter Cyt c molecules on 30 nm gold nanoparticle surface, the local concentration *C* of Cyt c on single nanoparticle is 47 mM. The optical path length *L* of a Cyt c monolayer is ~3 nm, so the optical absorbance is $A = \varepsilon CL \sim = 3 \times 10^{-4}$ assuming 100 % of scattering photon passing through the biomolecule monolayer.); therefore, the dramatic spectral dips are not a result of the direct optical absorption of Cyt c molecules.



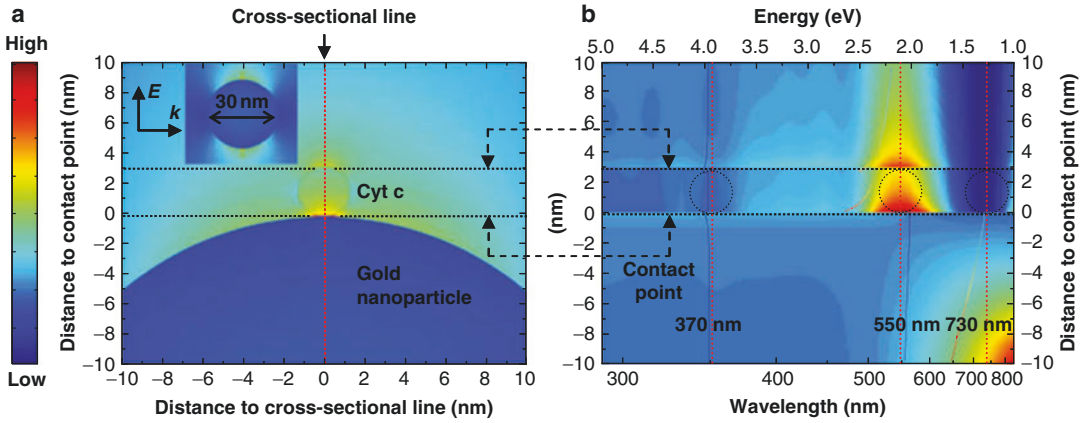
Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 2 Experimental results of PRET from single gold nanoparticle to conjugated Cyt c molecules. The Rayleigh scattering spectrum of a single gold nanoparticle coated with (**a**) only Cysteamine coating, (**b**) Cysteamine cross-linker and reduced Cyt c and (**c**) Cysteamine and oxidized Cyt c. The Rayleigh scattering spectrum was obtained using 1 s integration time. (**d**) The visible absorption spectra of Cyt c bulk solution in reduction form (*blue solid line*), and in oxidation form (*red solid line*) measured in conventional UV–vis absorption spectroscopy. (**e**) The fitting curve for the spectrum in (**b**). *Black open circle*: raw data, *Green solid line*: fitting curve of the raw data, *Yellow solid line*: Lorenzian scattering curve of bare gold nanoparticle, *Red solid line*: processed absorption spectra for the reduced conjugated Cyt c by subtracting yellow curve from the green curve. (**f**) The fitting curve for the spectrum in (**b**). *Black open circle*: raw data, *Green solid line*: fitting curve of the raw data, *Yellow solid line*: Lorenzian scattering curve of bare gold nanoparticle, *Blue solid line*: processed absorption spectra for the oxidized conjugated Cyt c by subtracting *yellow curve* from the *green curve*

The energy matching condition in PRET is further confirmed by three negative control experiments. For the first control experiment, synthesized peptides which have absorption peaks out of wavelength range of the plasmon resonance of a 30 nm gold nanoparticle are intentionally conjugated. As expected, the scattering spectrum of this hybrid system shows only the scattering peak because the absorption peaks of peptide do not coincide with the plasmon resonance spectrum of the nanoparticle (Fig. 3a). For the second control experiment, the importance of matching resonant frequency and molecular absorption peaks was tested by using a large gold nanoparticle cluster which has a plasmon resonance wavelength beyond 650 nm. As anticipated, the conjugated Cyt c absorption peaks can be hardly observed, including the 525 and 550 nm peaks for reduced Cyt c (Fig. 3b). For the third control experiment, dielectric polystyrene nanoparticles were conjugated with Cyt c and characterized. As expected, the plasmon quenching spectral dips cannot be found on the scattering spectrum of single dielectric polystyrene nanoparticle without plasmon resonance even though the dielectric nanoparticle also scatters light, which indicates the presence of excited free electrons is necessary for the PRET process (Fig. 3c).

The average surface density of Cyt c molecules on an individual gold nanoparticle is controlled by the molar concentration ratio used in the conjugation process. Considering the effective cross-sectional area of single Cyt c molecules and the surface area of single 30 nm gold nanoparticle, maximally around 400 Cyt c molecules can be tethered on a 30 nm gold nanoparticle. The scattering spectrum of many individual 30 nm nanoparticles was measured and extracted the reduced Cyt c visible absorption peaks. Due to the nonuniformity of the surface molecule numbers on each particle and nanoparticle plasmon resonance wavelength, the plasmon quenching or Cyt c absorption peak intensity shows variations from particle to particle (Fig. 4a), whereas the spectral measurement on each individual nanoparticle is repeatable and stable (Fig. 4b) and no photochemical changes are observed.



Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 3 Negative control results showing the importance of critical energy matching for PRET. (**a**) The scattering spectrum spectra of a 30 nm gold nanoparticle coated with Cys-(Gly-Hyp-Pro) [6] peptides. (**b**) The scattering spectrum of a single large gold nanoparticle cluster conjugated with Cysteamine and Cyt c. (**c**) The scattering spectrum of a 40 nm amine-modified polystyrene bead conjugated with Cyt c



Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 4 Simulation of nanoparticle plasmon resonance coupling to a single Cyt c molecule. (**a**) Time-averaged total electromagnetic (EM) energy at 550 nm vertically polarized light excitation around the interface of single 30 nm gold nanoparticle and single 3 nm spherical molecules. The dielectric nanosphere is used to simulate single reduced Cyt c molecule with a wavelength-dependent complex refractive index (16). The EM energy is transferred to the single molecule and forms the dipolar energy distribution across the molecule. The inset image of the whole nanoparticle shows the energy coupling only occurs in the light polarization direction. (**b**) Time-averaged total energy profile at the cross-section line in (**a**) as the function of the excitation wavelength or energy. The energy distribution at each wavelength is normalized to the energy at the inner surface of the nanoparticle. The representative line plots of the energy profile at 370, 550, and 730 nm are superposed on the 2D color-coded energy distribution at corresponding wavelength positions. The dipolar energy of nanoparticle plasmon resonance is electromagnetically coupled to the single molecule around 550 nm, while at other wavelengths much less energy transfer is observed

Unpolarized white light source is used in all the above experiments.

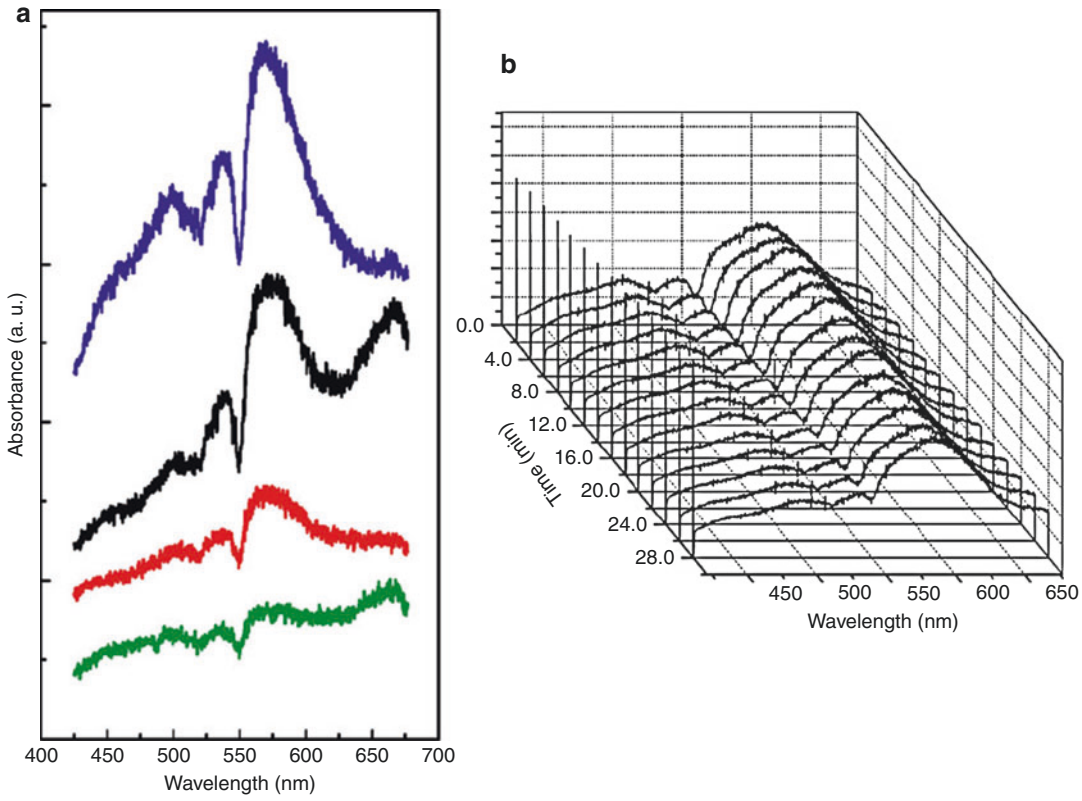
Similar to the energy transfer process in FRET, the PRET efficiency is possibly dependent on the distance from the spectrally active moiety of biomolecules, e.g., Heme group for Cyt c, to the plasmonic nanoparticle surface as well as the relative orientations between the polarized plasmon resonance dipole and molecular dipoles (Fig. 5a). On the other hand, the simulated single nanoparticle PRET spectra (Fig. 5b) shows that the strongest plasmon resonance mode for a single Cyt c conjugated 30 nm gold nanoparticle occurs around 550 nm. This resonant frequency (i.e., energy) matching condition potentially explains why the plasmon quenching peak amplitude at 550 nm is relatively higher than at 525 nm (Fig. 2e) compared to the peak intensity ratio in Cyt c bulk solution absorption measurement (Fig. 2e). The scattering peak wavelength of the gold nanoparticle in experiments is higher than the simulated results due to larger numbers of Cyt c and Cysteamine molecules conjugated on surface.

The PRET effect can be also observed for hemoglobin molecules conjugated with single silver nanoparticles. Hemoglobin molecule has a distinctive absorption peaks around 407 nm which is the sorbet band of the heme group. The absorption spectra of hemoglobin with and without binding to oxygen are shown in Fig. 6. Using PRET nanosensor, plasmon quenching dips were observed and corresponding to the Soret band (407 nm) of hemoglobin molecules on the surface of a single nanoparticle, as shown in Fig. 7.

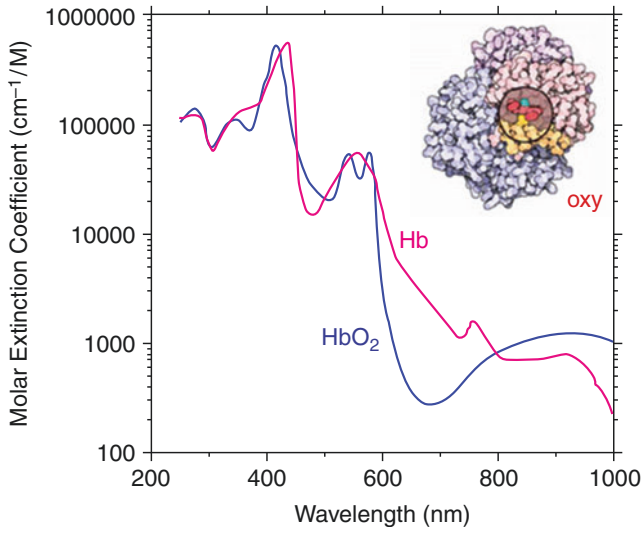
Key Research Findings

PRET Imaging of Nanoplasmonic Particles in Living Cells

The nanoplasmonic particles can be internalized into living cells via multiple ways, either by natural endocytosis of cells or assisted endocytosis such as lipo-transfection technique. The nanoplasmonic particles can be effectively internalized in living cancer cells using Fugene

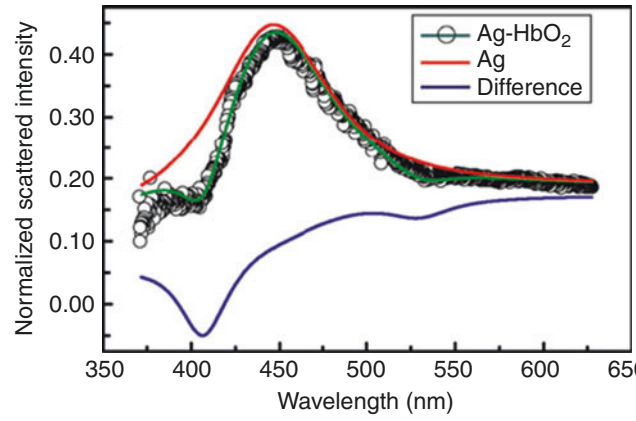


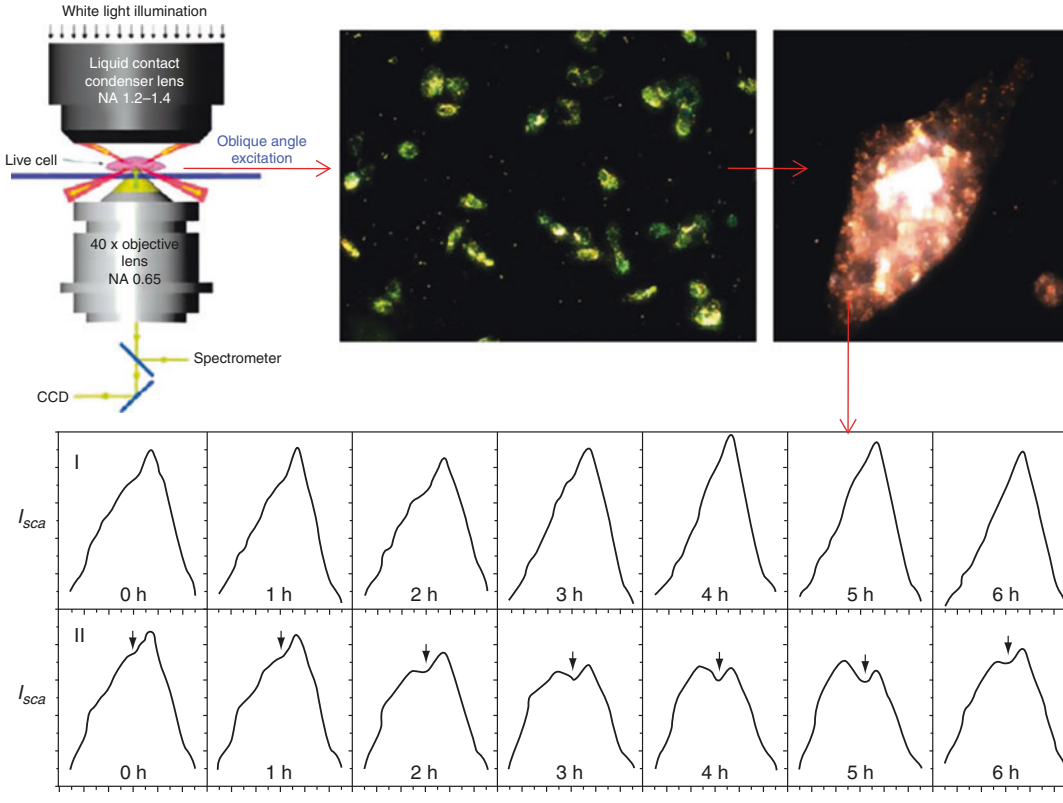
Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 5 (**a**) Raw scattering spectra of four representative gold nanoparticles conjugated with reduced Cyt c molecules. The nanoparticle plasmon resonance peaks and PRET-induced plasmon quenching dips have variable intensities from particle to particle due to the nonuniformity of the conjugated molecule number and particle geometry; however the plasmon quenching peak positions are consistent. More than 50 individual nanoparticles were tested and PRET can be consistently observed. (**b**) Time-lapse measurement of scattering spectra of a single gold nanoparticle conjugated with reduced Cyt c molecules. The plasmon quenching spectral dips remain nearly constant during the whole time period of measurement. No photobleaching effect was observed



Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 6 Absorption spectra of oxyhemoglobin and deoxyhemoglobin bulk high-concentration solutions

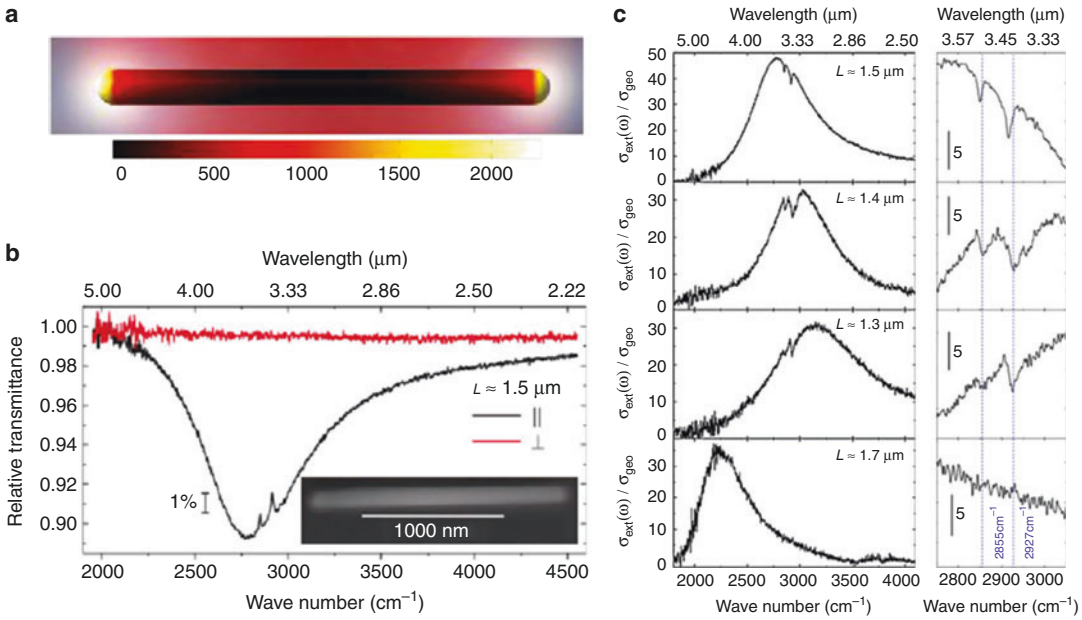
Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 7 Absorption spectrum of hemoglobin molecules on single nanoparticle surface and the spectrum is acquired and can be only acquired using our PRET nanomolecular sensor





Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 8 Darkfield imaging of PRET nanoparticles in living cancer cells and the time-lapse observation of intracellular Cyt c molecules using PRET nanospectroscopy [26]

6 (Roche Diagnostics) transfection reagent [25]. The nanoplasmonic particles can be clearly imaged using our nanospectroscopy imaging system. Figure shows the dark field scattering image of individual liver cancer cells with many internalized nanoplasmonic particles. Recent study also shows that intracellular Cyt c molecules can be imaged using the PRET nanospectroscopy in a long-term continuous measurement [26] as shown Fig. 8.



Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 9 Infrared PRET on gold nanowire. (**a**) Fundamental plasmon resonance mode along the long axis of the 1.34 μm long gold nanowire. (**b**) Relative transmittance spectrum of gold nanowire with (*black curve*) and without (*red curve*) molecule adsorption. (**c**) Spectral dips due to PRET in the infrared absorption spectra measured from gold nanowires in various lengths [27]

Near-Infrared PRET

The aforementioned examples primarily occur in visible light regime due to the visible plasmon resonance wavelength of sub-100 nm gold or silver nanoparticles. As the matter of fact, the PRET effect is not limited to visible wavelength range. When using long gold nanowires with the microscale lengths, the PRET effect can be observed even in mid-infrared region [27]. As shown in Fig. 9a, the gold nanowire with 100 nm in diameter and 1.31 μm in length has the fundamental plasmon resonance mode along the long axis at about 3.41 μm. The gold nanowire will show very different absorption spectra when excited by perpendicularly and parallel polarized light (Fig. 9b). Only the excited light polarized along the long axis of the gold nanowire will be absorbed by the nanowire due to the fundamental plasmon resonance mode at 3.41 μm.

When certain chemical molecules are adsorbed on the surface of the gold nanowire, a few spectral dips can be clearly observed and they overlap on top of the nanowire absorption peak. The positions of the spectral dips exactly match with the infrared absorption peak wavelengths of the chemical molecule at 2855 and 2927 cm^{-1}. Even if the plasmon resonance wavelengths shift due to the nanowire length variation, the PRET spectral dip positions at infrared region remain identical and show high specificity to molecule species. However, the amplitude of the infrared PRET spectral dips is dependent on the extent how the infrared plasmon resonance peak of gold nanowire matches with the infrared absorption peaks of the adsorbed molecules. Since most of biological and chemical molecules have distinctive infrared absorption peaks, the PRET nanospectroscopy on gold and silver nanowires at infrared wavelength has even wider applications to detect small amount of adsorbed molecules without labeling process.

Applications

Whole-Field PRET Imaging

All previous PRET imaging works have been limited to the spectroscopy of several particles at

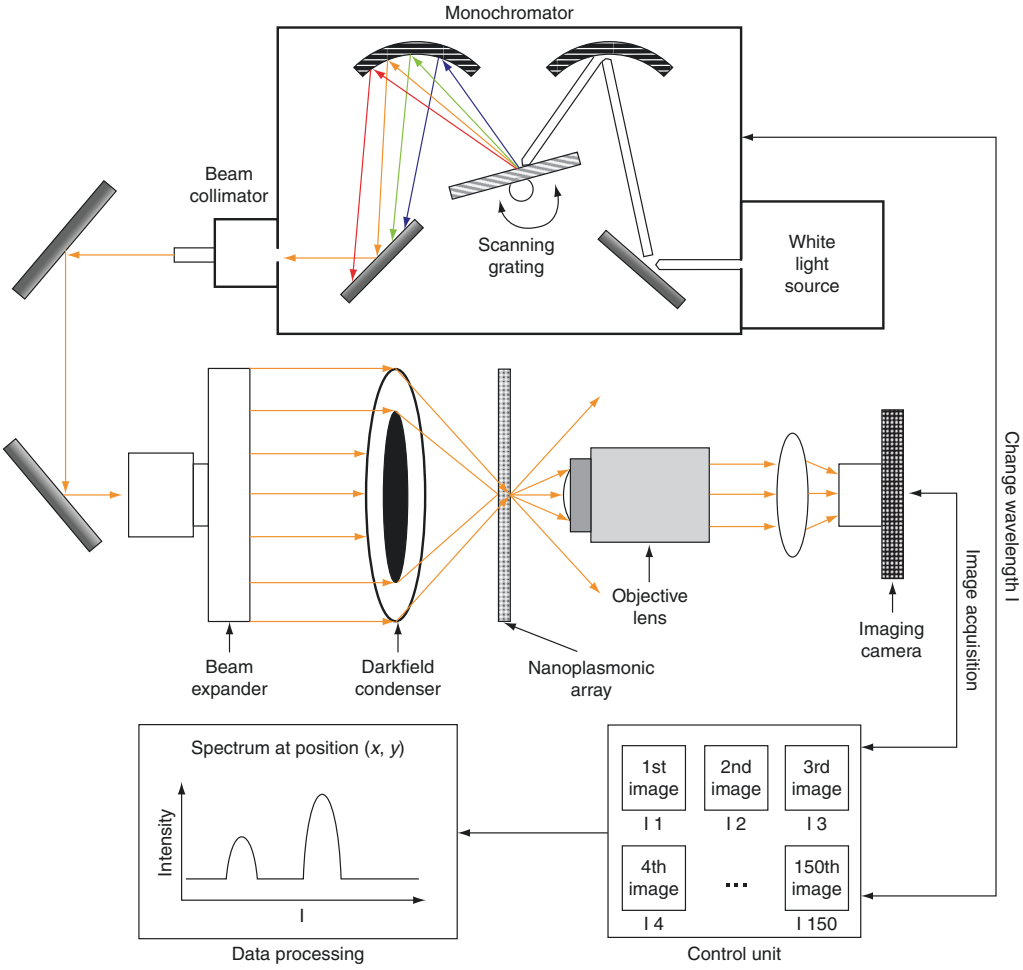
most, and whole-field intracellular PRET imaging has not been implemented to date due to the limitations imposed by current spectral imaging system. All existing imaging and spectroscopy systems for metallic nanostructures are based upon a microscopy system with a white light illumination source, a darkfield condenser, and a polychromator spectrograph, as shown before. In such a system, the field of view is limited to several microns by the width of the entrance slit to the polychromator even though the field of view of the objective lens is on the order of several hundred microns; therefore only a few nanoparticles or nanostructures that are aligned in a column can be imaged spectrally at once.

The polychromator entrance slit is required to limit data pollution for the high-resolution spectroscopy of each individual nanoparticle. Hence, in the case of spectrally imaging hundreds or even thousands of sparsely distributed nanostructures, the sample stage of the microscopy system must be translated either manually or automatically, which has two disadvantages. First, the spectra of separated nanostructures cannot be simultaneously captured and the acquired data will be from different instants in time, e.g., 20∼30 min for a 100 × 100 array, which voids their applicability in the real-time monitoring of multiple biomolecular interactions. Second, the focal plane will tend to drift due to the stage translation which may change the condition of the immersion liquid between the sample and darkfield condenser lens, and thus manual refocusing is often required.

A new multispectral imaging system was constructed to simultaneously monitor the individual scattering spectra and plasmon resonance wavelength of large numbers of nanoplasmonic particles distributed within the field of view of a microscopy objective lens without mechanically scanning the sample [28]. The multispectral imaging system currently supports frame rates as high as 2 s per frame (wavelength) that could potentially be increased by using a light source with a higher power or an image detector with greater sensitivity, by which the image signal-to-noise ratio can also be increased. In contrast with the configurations of previous imaging systems, our whole field plasmonic imaging system consists of a multispectral illumination source synchronized with an intensity imaging camera rather than a polychromatic spectrograph. As shown in Fig. 10, the white light from a halogen lamp is coupled into a scanning monochromator which is controlled by a computer program to output a monochromatic light beam with varying wavelength and specified spectral step size. The monochromatic light beam is then focused and illuminated on the sample by a darkfield condenser (N. A. = 1.2∼1.4). A 40× microscopy objective lens (N. A. = 0.8) is used to collect the scattered light, which is captured by a CCD camera. The monochromator and image acquisition control software are integrated and they are synchronized to capture a single image at each wavelength of interest. The image acquisition can be finished in a few minutes and the acquisition time is dependent on the chosen spectral range and resolution.

All images are stored as grayscale data files and are analyzed by an image processing program. The bright spot regions (typically 1∼10 pixels) in each image are recognized by the analysis program as individual nanoplasmonic structures of interest. The mean intensity value of these small regions is extracted from the image at each wavelength as the raw scattering spectra data. The mean intensity value in a large, empty (black) region is also measured at each wavelength as the background spectrum, which is subsequently subtracted from the raw scattering spectra. The difference spectra are then scaled according to the previously stored spectrum of the light output from the monochromator to yield the final scattering spectra. The process of the image analysis and the spectral data reconstruction is completely automated by the computer program.

As a demonstration, randomly dispersed Au colloidal nanoparticles were used on a glass slide as the imaging sample. The diameters of the Au nanoparticles vary from 20 to 80 nm, so their plasmon resonance wavelengths, and thus their scattering colors are different. Figure 11a shows the true-color scattering image of the nanoparticles within ∼1/10 of whole view field of the objective lens. The true-color image is taken

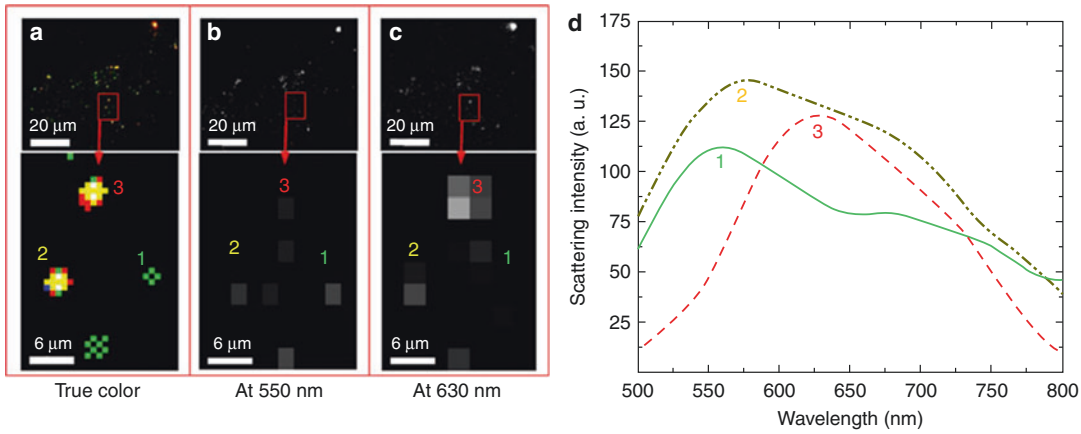


Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 10 Configuration of the multispectral imaging system for the scattering spectra measurement of nanoplasmonic particles in the whole field of view [28]

in the same darkfield microscopy system but with a white-light illumination source and color camera. Figures 11b and 2c show the scattering intensity images of the same nanoparticles within the same field of view taken by our system at 550 and 630 nm, respectively. Figure 11c shows the scattering spectra of three representative nanoparticles marked in Fig. 11a–c. The plasmon resonance wavelengths (spectral peaks) of these three nanoparticles are respectively 560, 580, and 630 nm, which agree well with their colors (green, yellow, and red) in Fig. 11a and their relative intensities in Figs. 11b and 10c. The Au "particle" with red scattering color could be a cluster of a few Au nanoparticles, because the plasmon resonance wavelength of individual Au nanoparticle is shorter than 600 nm according to Mie scattering theoretical predictions. Although only the scattering spectra of three typical particles are shown here, the spectral information for all the other nanoparticles in the field of view are also stored at once and can be reconstructed in the same fashion.

Future Directions for Research

Although only the PRET in the visible wavelength range is observed here due to the optical properties of gold nanoparticles and Cyt c molecules, the PRET process at UV and near infrared range could be envisioned by using different properties (i.e.,



Plasmon Resonance Energy Transfer Nanospectroscopy, Fig. 11 (**a**) True-color scattering image of thousands of dispersed Au nanoparticles. For the clarity of the image, a field of 100 × 100 μm is cropped from the whole view field of ~300 × 300 μm. The lower picture is the zoom-in image from the marked area (*square*) of the upper image. (**b**) and (**c**) show the scattering intensity image of the same Au nanoparticles as in (**a**) with 550 and 630 nm monochromatic illumination, respectively. (**d**) Scattering spectra of three representative particles marked as *1*, *2*, and *3*, respectively, in the images [28]

size, shape, free electron density, etc.) of metallic nanoparticles with UV or near infrared plasmon resonance wavelength. The PRET-based ultrasensitive biomolecular absorption spectroscopy on single metallic nanoparticle could be used for molecular imaging such as genetic analysis of small copies of latent nucleotides, activity measurements of small numbers of functional cancer biomarker proteins, and rapid detection of little biological toxin, pathogen, and virus molecules. Additionally PRET could be applied in intracellular biomolecule-conjugated nanoparticle sensors to detect localized in vivo electron transfer, oxygen concentration, and pH value changes in living cells with nanoscale spatial resolutions. Furthermore the optical energy in advanced plasmonic devices [29] could be potentially tuned by functional biomolecules taking advantage of the PRET process.

References

1. Bohren, C.F., Huffman, D.R.: Absorption and Scattering of Light by Small Particles, pp. 335–336. Wiley, New York (1998)
2. Mock, J.J., Smith, D.R., Shultz, S.: Local refractive index dependence of plasmon resonance spectra from individual nanoparticles. Nano Lett. **3**, 485–491 (2003)
3. Otto, A.: Excitation of nonradiative surface plasma waves in silver by the method of frustrated total reflection. Z. Phys. **216**, 398–410 (1968)
4. Maier, S.A., Kik, P.G., et al.: Local detection of electromagnetic energy transport below the diffraction limit in metal nanoparticle plasmon waveguides. Nat. Mater. **2**, 229–232 (2003)
5. Maier, S.A., Atwater, H.A.: Plasmonics: localization and guiding of electromagnetic energy in metal/dielectric structures. J. Appl. Phys. **98**, 011101 (2005)
6. Moskovits, M.: Surface-enhanced spectroscopy. Rev. Mod. Phys. **57**, 783–826 (1985)
7. Lombardi, J.R., Birke, R.L., Lu, T.H., Xu, J.: Charge-transfer theory of surface enhanced Raman spectroscopy – Herzberg-Teller contributions. J. Chem. Phys. **84**, 4174–4180 (1986)
8. Nie, S., Emory, S.R.: Probing single molecules and single nanoparticles by surface-enhanced Raman scattering. Science **275**, 1102–1106 (1997)
9. Kneipp, K., et al.: Single molecule detection using surface-enhanced Raman scattering (SERS). Phys. Rev. Lett. **78**, 1667–1670 (1997)
10. Das, P., Metiu, H.: Enhancement of molecular fluorescence and photochemistry by small metal particles. J. Phys. Chem. **89**, 4680–4687 (1985)
11. Liu, G.L., et al.: A nanoplasmonic molecular ruler for measuring nuclease activity and DNA footprinting. Nat. Nanotechnol. **1**, 47–52 (2006)
12. Endo, T., Kerman, K., Nagatani, N., Takamura, Y., Tamiya, E.: Label-free detection of peptide nucleic acid-DNA hybridization using localized surface plasmon resonance based optical biosensor. Anal. Chem. **77**, 6976–6984 (2005)

13. Rascheke, G., et al.: Biomolecular recognition based on single gold nanoparticle light scattering. Nano Lett. **3**, 935–938 (2003)
14. Sönnichsen, C.: Plasmons in metal nanostructures. Ph.D. dissertation, University of Munich (2001)
15. van Dijk, M.A., et al.: Absorption and scattering microscopy of single metal nanoparticles. Phys. Chem. Chem. Phys. **8**, 3486–3495 (2006)
16. Wuttke, D.S., Bjerrum, M.J., Winkler, J.R., Gray, H. B.: Electron-tunneling pathways in cytochrome-c. Science **256**, 1007–1009 (1992)
17. Yang, J., Liu, X.S., et al.: Prevention of apoptosis by Bcl-2: release of cytochrome c from mitochondria blocked. Science **275**, 1129–1132 (1997)
18. Lange, C., Hunte, C.: Crystal structure of the yeast cytochrome *bc1* complex with its bound substrate cytochrome *c*. Proc. Natl. Acad. Sci. U. S. A. **99**, 2800–2805 (2002)
19. Andrew, P., Barnes, W.L.: Energy transfer across a metal film mediated by surface plasmon polaritons. Science **306**, 1002–1005 (2004)
20. Van Duyne, R.P.: Molecular plasmonics. Science **306**, 985–986 (2004)
21. Fang, N., Lee, H., Sun, C., Zhang, X.: Sub-diffraction-limited optical imaging with a silver superlens. Science **308**, 534–537 (2005)
22. Bard, A.J., Faulkner, L.R.: Electrochemical Methods: Fundamentals and Applications, 2nd edn, p. 685. Wiley, New York (2001)
23. Ozbay, E.: Plasmonics: merging photonics and electronics at nanoscale dimensions. Science **311**, 189–193 (2006)
24. Boussaad, S., Pean, J., Tao, N.J.: High-resolution multiwavelength surface plasmon resonance spectroscopy for probing conformational and electronic changes in redox proteins. Anal. Chem. **72**, 222–226 (2000)
25. Lee, E.S., Liu, G.L., Kim, F., Lee, L.P.: Remote optical switch for localized control of gene interference. Nano Lett. **9**, 562–570 (2009)
26. Choi, Y., Kang, T., Lee, L.P.: Plasmon resonance energy transfer (PRET)-based molecular imaging of cytochrome c in living cells. Nano Lett. **9**, 85–90 (2009)
27. Neubrech, F., Pucci, A., Cornelius, T.W., Karim, S., Garcia-Etxarri, A., Aizpurua, J.: Resonant plasmonic and vibrational coupling in a tailored nanoantenna for infrared detection. Phys. Rev. Lett. **101**, 157403 (2008)
28. Liu, G.L., Doll, J.C., Lee, L.P.: High-speed multispectral imaging of nanoplasmonic array. Opt. Express **13**, 8520–8525 (2005)
29. Haes, A.J., Zou, S., Zhao, J., Schatz, G.C., Van Duyne, R.P.: Localized surface plasmon resonance spectroscopy near molecular resonances. J. Am. Chem. Soc. **128**, 10905–10914 (2006)

Plasmon-Assisted Two-Photon Polymerization

► Nanoplasmonics Involved in Localized Photopolymerization

Plasmon-Enhanced Fluorescence

► Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons

Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons

Jakub Dostalek
BioSensor Technologies, AIT-Austrian Institute of Technology GmbH, Vienna, Austria

Synonyms

Plasmon-enhanced fluorescence; Plasmonic amplification of fluorescence; Plasmonic fluorescence biosensor; Surface plasmon-coupled emission

Definition

Plasmonics offers efficient means for enhancing sensitivity of fluorescence bioassays for detection of chemical and biological analytes. In this method, specific capture of target analyte from analyzed liquid sample at metallic sensor surface with attached recognition elements is observed by using fluorophore labels (e.g., organic dyes or quantum dots). These labels are probed by the confined field of surface plasmons that originate from collective oscillations of charge density at a surface of metallic films or metallic nanoparticles. The excitation of surface plasmons is accompanied with strongly increased intensity of electromagnetic field which couples with fluorophores. The combination of surface plasmon-enhanced excitation rate fluorophore absorption wavelength, improved quantum yield, and directional emission at the emission wavelength allows enhancing the detected fluorescence intensity per captured molecule by several orders of magnitude. When implemented to heterogeneous

assays, limit of detection can be improved by similar factor.

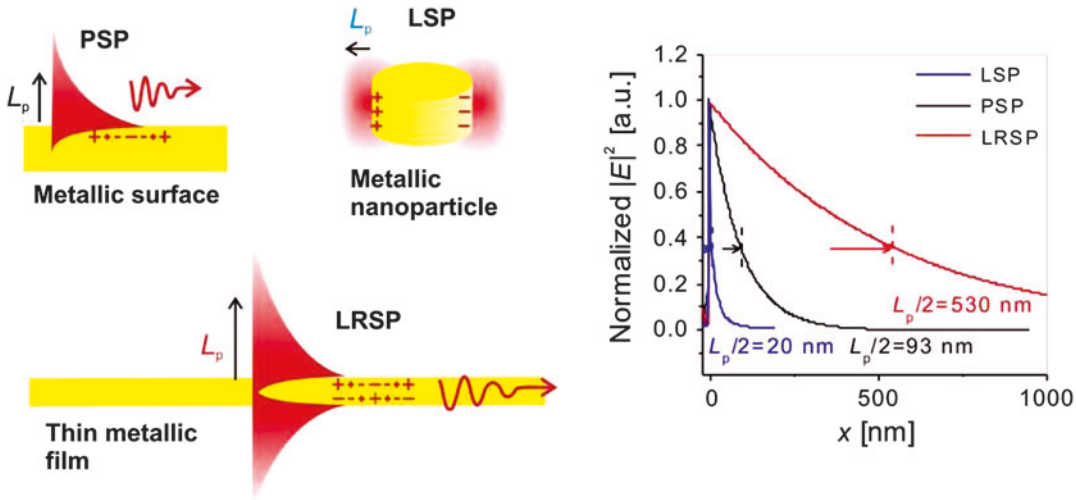
Overview

The emission of fluorophores placed above a metallic surface was investigated since the 1970s of the last century [1, 2]. These studies revealed that the lifetime of an excited fluorophore is dramatically reduced due to the coupling with so-called surface lossy waves (leading to quenching at distances below Förster radius of ~10 nm) and propagating surface plasmons (that is observed at longer distances up to ~100 nm). The emitted photons that are trapped by propagating surface plasmon (PSP) modes can be partially recovered by reverse Kretschmann configuration [3] or by periodic corrugation of the metal surface [4, 5]. These techniques for out-coupling of PSPs allow extracting the emitted fluorescence photons from the surface to a highly directional beam and deliver them to a detector.

The coupling of fluorophores with PSPs provides efficient means for increasing sensitivity of heterogeneous fluorescence bioassays. In this type of assays, fluorophore emitters are used as labels to report affinity binding of target molecules from analyzed liquid sample at the surface with attached recognition elements such as antibodies or oligonucleotides [6–9]. The interaction of fluorophore labels with PSPs at their absorption λ_{ab} and emission λ_{em} wavelengths can increase their "brightness" and thus offers means to detect smaller number of captured molecules on the surface. In analytical applications, this translates to the possibility of detecting smaller concentrations of analyte in a sample and improved limit of detection (LOD).

Surface Plasmons

As Fig. 1a illustrates, PSP modes travel along a continuous metallic film and probe the adjacent dielectric medium above the surface with enhanced electric field intensity $|E|^2$. This field exponentially decays as $\exp(-2x/L_p)$, where x states for a distance from the metal surface. The probing depth is typically quantified by a half of penetration depth $L_p/2$ that is equal to about 90 nm for PSPs on a gold surface in contact with aqueous medium at a red wavelength of $\lambda = 633$ nm. The probing depth can be changed



Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons, Fig. 1 Examples of surface plasmon modes on metallic films and metallic nanoparticles (*left*). Comparison of probing depth of regular propagating surface plasmons (PSPs) and long-range surface plasmons (LRSPs) on a gold surface in contact with water and of localized surface plasmons supported by a gold disk with diameter of 110 nm and height of 50 nm (*right*). Wavelength of $\lambda = 633$ nm was assumed

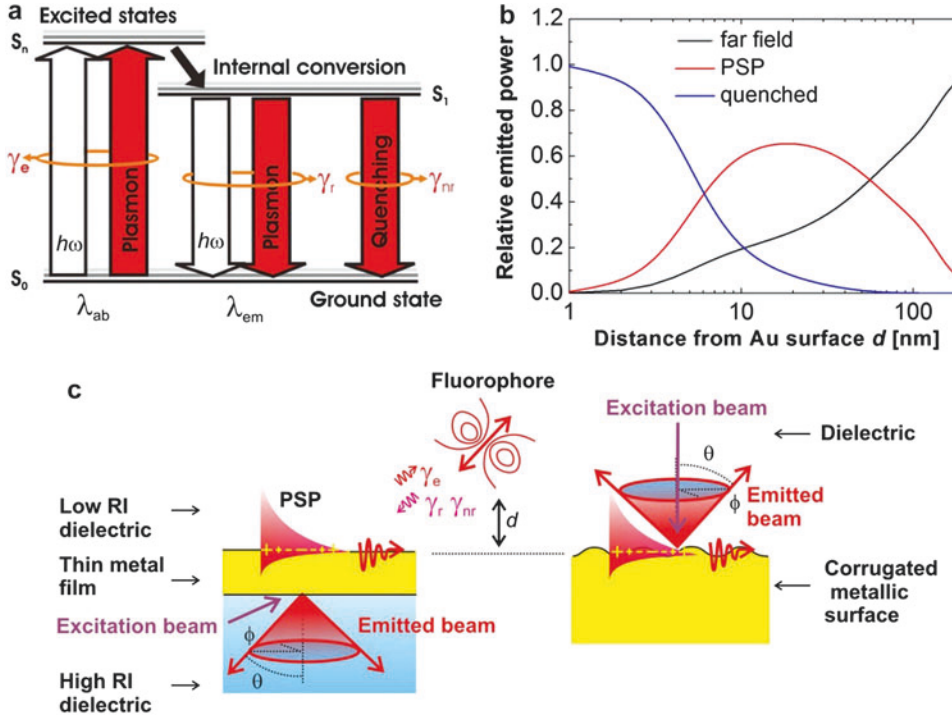
by a design of the surface plasmon-supporting metallic structure. For example, the coupling of PSP modes at opposite interfaces of a thin metallic film gives rise to new plasmonic modes with tunable characteristics. This coupling can occur when a thin metallic film is sandwiched between dielectrics with similar refractive index [10]. For probing of aqueous medium that exhibits refractive index close to $n = 1.33$, such geometry can be realized by depositing a thin metal film (with thickness of few tens of nanometers) on the top of a low refractive index fluoropolymer such as Cytop from Asahi Inc. ($n = 1.34$) or Teflon AF from DuPont Inc. ($n = 1.31$) [11, 12]. The near-field coupling of PSPs at opposite interfaces of a metal film gives rise to the occurrence of long-range surface plasmons (LRSP) that exhibit weaker confined field and thus probe deeper away from the metal. An example in Fig. 1 shows that for 20 nm thick gold film on the Teflon AF surface, the aqueous medium is probed by the extended LRSP field with the probing depth of $L_p/2 = 540$ nm at a wavelength of $\lambda = 633$ nm. Let us note that the probing depth can be tuned between that of regular PSPs and (virtually) infinity by changing the thickness metal and by introducing a small perturbation to the refractive index symmetry [13].

Plasmonic Coupling with Fluorescence Emitters

The resonant coupling of light with PSP modes confines its energy to a small volume in vicinity to a metal surface which leads to the strong enhancement of electromagnetic field intensity. The field intensity at metallic surface $|E|^2$ can be normalized with that of the excitation beam $|E_0|^2$ in order to quantify field intensity enhancement. Depending on the confinement and losses of excited PSP modes, the field enhancement $|E/E_0|^2$ can reach values up to $\sim 10^2$ (it should be noted that much stronger enhancement was observed upon the excitation of localized surface plasmons on sharp metallic nanoparticles that feature more confined field [14]). As Jablonski diagram in Fig. 2a illustrates, PSPs couple with a fluorophore at both emission and excitation wavelengths λ_{ab} and λ_{em}, respectively. This interaction strongly depends on the distance d between the fluorophore and metallic surface. Figure 2b shows the relative power emitted by a fluorophore at a red wavelength $\lambda_{em} = 670$ nm above a gold surface. At distances d below $\sim$10 nm, the fluorescence intensity is mostly quenched due to the Förster energy transfer. At the distance of about $\sim$20 nm, the majority of energy is emitted via PSPs. When moving farther away from the surface, the fluorophore emits preferably to the far field.

Figure 2c shows two common schemes that allow converting the fluorescence light that is trapped by PSPs into waves propagating into far field. PSPs that are generated by fluorescence at λ_{em} can leak through a thin metallic film when the reverse Kretschmann geometry of attenuated total reflection (ATR) method is used. For randomly oriented emitters (fluorophores), the fluorescence light is reemitted to a cone beam that propagates below the thin metal at a polar angle θ at which surface plasmon resonance occurs. Alternatively, diffraction on periodically corrugated metallic surface allows for out-coupling of PSPs into the far field.

For the plasmonic amplification of fluorescence, the emitter (which is typically a fluorophore label anchored to a biomolecule) is placed on a metallic surface and exposed to the intense surface plasmon field. The fluorescence intensity that is delivered to a detector can be increased by three effects. Firstly, the coupling of an excitation beam with PSPs increases the excitation rate of emitter that is (far from saturation) proportional to the field enhancement $|E/E_0|^2$ at the absorption wavelength λ_{ab}. Secondly, the emitter radiates fluorescence light at the emission wavelength λ_{em} via PSPs that can be afterward reemitted to a specific direction toward a detector. This feature improves the fraction of emitted photons that are delivered to the detector from above the plasmonic structure f_d compared to that without it f_{d0}. Thirdly, near-field coupling of the emitter with PSPs can improve its quantum yield η (particularly for emitters with low intrinsic quantum yield η_0). Let us note that these three effects can be combined and the total enhancement factor EF (defined as fluorescence intensity emitted by a



Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons, Fig. 2 (**a**) Jablonski diagram showing surface plasmon-mediated transitions between the fluorophore ground state and higher excited states. (**b**) Comparison of the emitted power from randomly oriented dipole that is quenched, emitted via PSPs, and radiated directly to the far field depending on the distance from a flat gold surface *d*. (**c**) Schematic out-coupling of PSP modes via reverse Kretschmann geometry of ATR method (*left*) and by using diffraction on periodically corrugated metallic surface (*right*)

fluorophore at a metallic structure F normalized with that without the structure F_0) can be expressed as

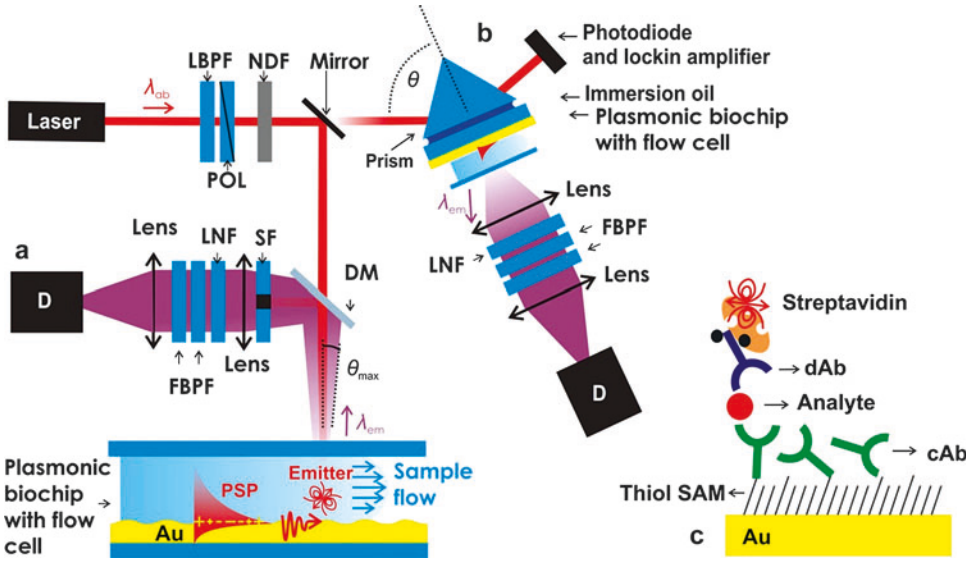
$$EF = F/F_0 \propto |E/E_0|^2 \times f_d/f_{d0} \times \eta/\eta_0. \quad (1)$$

It should be noted that the EF strongly depends on the emitter orientation which is in most experiments not controlled. For ensembles of fluorophores, Eq. 1 typically needs to be averaged over all possible orientations and locations of emitter on plasmonic structure (mode details on such models can be found in literature [15, 16]).

Implementations of Plasmon-Enhanced Fluorescence Biosensors

The employing of PSPs for the amplification of fluorescence signal mostly utilizes epifluorescence or total internal reflection fluorescence (TIRF) geometries that are depicted in Fig. 3a, b, respectively. In the epifluorescence-based sensors, an excitation beam at a wavelength that coincides with the absorption band of used fluorophore label at λ_{ab} is made incident on a sensor chip a with plasmonic structure. The surface plasmon-enhanced field intensity at this wavelength probes the assay on the surface (see Fig. 3c) and excites fluorophore labels that report the target analyte binding events. The fluorophore emits light at the emission wavelength λ_{em} directly to the far field or via PSPs that are subsequently out-coupled toward a detector.

The TIRF-like geometry allows for combined surface plasmon resonance (SPR) and fluorescence measurements based on ATR method with Kretschmann geometry (see Fig. 3b). In this implementation, the resonantly excited surface plasmon waves on a thin metal film can probe



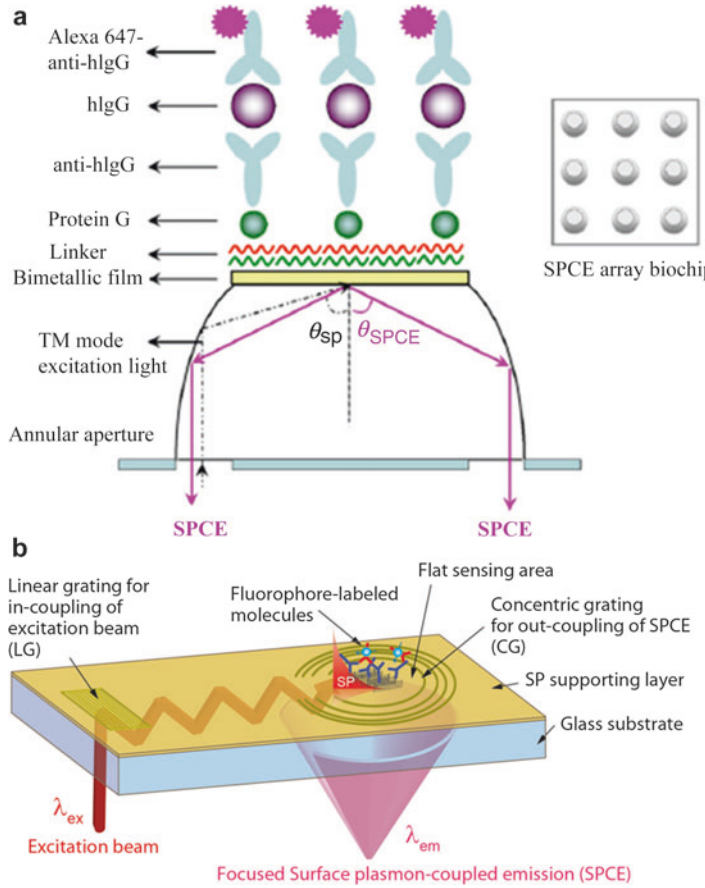
Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons, Fig. 3 Schematics of (**a**) total internal reflection fluorescence (TIRF) and (**b**) epifluorescence geometry with the surface plasmon-based amplification of fluorescence light intensity. (**c**) An example of sandwich immunoassay with the fluorescence detection that utilizes capture antibody (cAb) attached to a surface and biotinylated detection antibodies (dAb) that are decorated with fluorophore-labeled streptavidin

the binding events by (i) changes in refractive index due to the capture of target analyte leading to a shift of the SPR angle and (ii) fluorescence light that is excited by the enhanced surface plasmon field intensity when fluorophore labels are used [7]. The fluorescence light that is directly emitted perpendicular to the surface through the sample can be detected. Optical filters are used in order to decrease background signal (see Fig. 3). The laser band-pass filter (LBPF) blocks light propagation at wavelengths different from the excitation laser wavelength that is close to the absorption wavelength of used fluorophore λ_{ab}. Laser notch filter (LNF) is employed to block the laser light intensity that is typically much stronger than that of emitted fluorescence light at λ_{em}. In addition, fluorescence band-pass filters (FBPF) are used to select only the light at wavelengths around λ_{em}. In general, the fluorescence intensity is weak, and thus low light intensity detectors (D) such as photomultiplier, avalanche photodiode, or cooled CCD are used.

In the TIRF-like geometry depicted in Fig. 3b, the majority of fluorescence light intensity is leaking through the thin metallic film and propagates in the form of a characteristic surface plasmon-coupled emission (SPCE) cone. In order to retrieve this fluorescence light and deliver it to a detector, glass hemispherical prisms were used [17]. Later on, this approach was implemented by using miniaturized parabolic elements that can be embossed to polymer chips [18]. As Fig. 4a shows, such element allows directly imaging SPCE intensity to a detector and offers means for measurements with array detection format. Alternatively, concentric diffraction grating was proposed for out-coupling of SPCE signal and its imaging to a detector [19]; see Fig. 4b. This approach can be implemented by mass production-compatible technologies such as nanoimprint lithography and holds potential for simplification of the more traditionally used optical schemes.

Examples of Applications

Since the end of the last century, numerous fluorescence biosensors that utilize plasmonic amplification of fluorescence signal were developed for detection of nucleic acids [20], proteins [6], as well as whole cells [21]. These efforts were

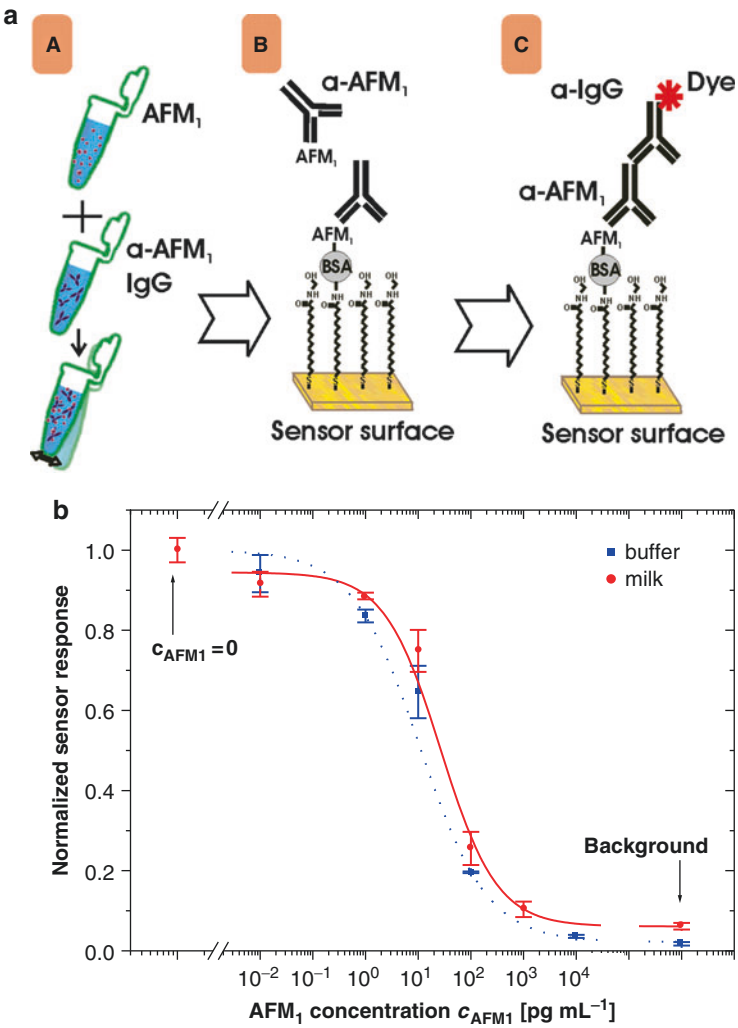


Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons, Fig. 4 Optical elements for the extraction of SPCE utilizing reverse Kretschmann configuration and (**a**) paraboloid elements and (**b**) concentric diffractive lens (Reproduced with permission from Yuk et al. 18 and Toma et al. 9)

mainly motivated by the increasing needs for assays with enhanced sensitivity in important areas of medical diagnostics (analysis of biomarkers), food control, and security (detection of harmful compounds such as toxins or pathogens). This section illustrates the plethora of sensor architectures and achieved performance characteristics by several examples. It should be noted that it does not cover the whole area and does not present a critical review in this field.

Plasmon-enhanced fluorescence biosensors were utilized for detection of low molecular weight analytes by using an inhibition competitive immunoassay. For instance, Wang et al. detected aflatoxin M_1 (AFM_1) with the molecular weight of 328 Da by using an optical arrangement similar to that depicted in Fig. 3b [22]. This harmful analyte is a metabolite of mycotoxin aflatoxin B_1 produced mainly by *Aspergillus flavus* and *Aspergillus parasiticus* pathogens. The gold sensor surface was functionalized by using a thiol self-assembled monolayer with a conjugate of AFM_1 and bovine serum albumin for the inhibition competitive immunoassay; see Fig. 5a. Monoclonal rat antibody against AFM_1 was incubated with a sample containing AFM_1, and the unreacted antibody was flowed over the sensor surface and detected by using a secondary anti-rat antibody that was labeled with a Cy5 dye. As the calibration curve in Fig. 5b shows, the limit of detection of AFM_1 present in milk was determined to be 0.6 pg/ml (1.8 pM) and the whole analysis required 53 min.

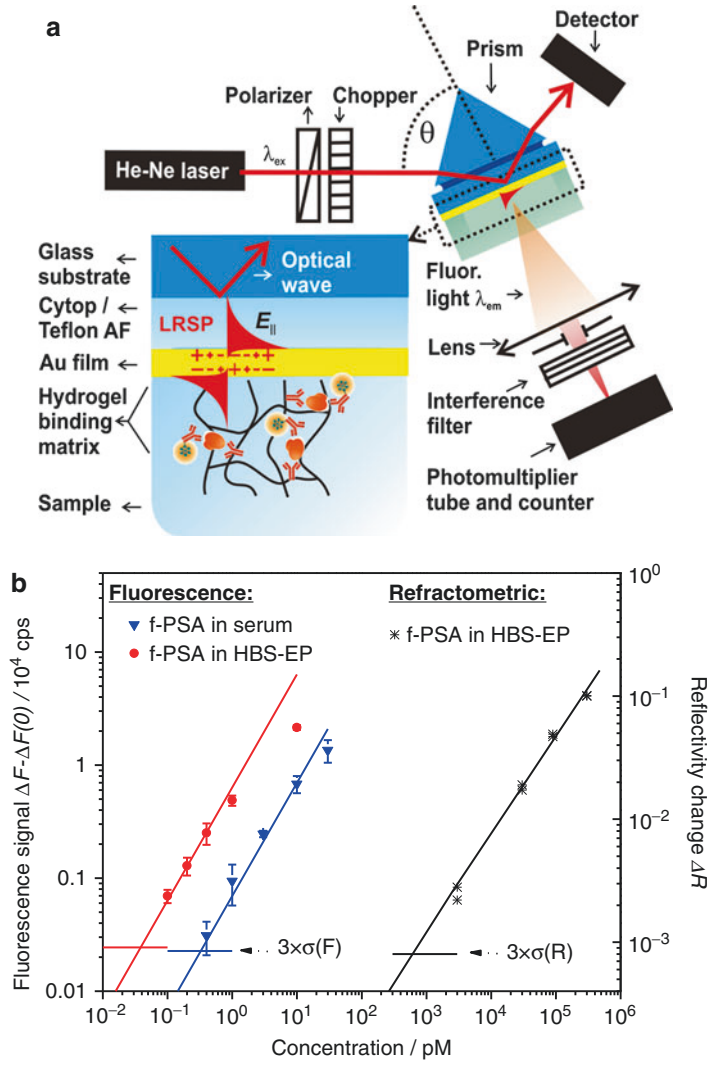
Prostate-specific antigen (PSA) is a biomarker with a molecular weight of 33–34 kDa that is present in the blood at around nM concentrations and used for a diagnosis of prostate cancer. Improving the sensitivity of PSA assays for the analysis at lower concentrations has been pursued for point-of-care diagnosis (POC) of female breast



Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons, Fig. 5 (**a**) Schematic of inhibition competitive immunoassay for detection of aflatoxin M1 and (**b**) a calibration curve for the aflatoxin M1 surface plasmon-enhanced fluorescence biosensor (Reproduced with permission from Wang et al. 22)

cancer [108], for early identification of prostate cancer relapse [109], and in forensic applications [110]. A biosensor for detection of free prostate-specific antigen (f-PSA) using long-range surface plasmon-enhanced fluorescence spectroscopy (LRSP-FS) and photo-cross-linked carboxymethylated dextran hydrogel matrix (depicted in Fig. 6a) was reported by Wang et al. [23]. In this approach, there was used a three-dimensional hydrogel binding matrix with a thickness of around 1 μm that offers the advantage of large binding capacity which can translate to higher efficiency in capturing of target analyte from a sample. In order to probe the whole volume of the binding matrix, long-range surface plasmons with increased probing depth were used. By using the sandwich immunoassays, the sensor allowed for detection of f-PSA in buffer and human serum with the limit of detection (LOD) of 34 f. and 0.33 pM, respectively, and the analysis time was of 35 min. This LOD was about four orders of magnitude better than that for SPR-based detection as can be seen from calibration curves presented in Fig. 6b.

Similar optics that allows probing to about micrometer distance from the metallic sensor surface was implemented for detection of large bacterial pathogens with around micrometer size by

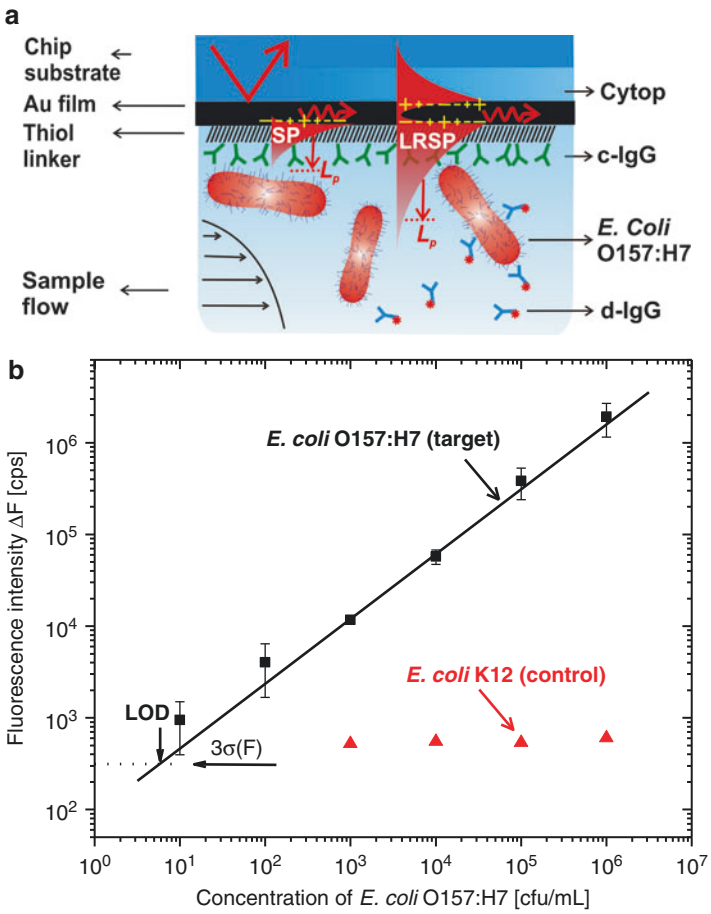


Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons, Fig. 6 (**a**) Surface plasmon-enhanced fluorescence biosensor with a biointerface architecture based on three-dimensional hydrogel matrix that is probed by LRSPs. (**b**) Comparison of calibration curves of free prostate-specific antigen (f-PSA) assay with direct SPR readout (*black curve*) and with plasmonically amplified fluorescence readout for serum (*blue curve*) and buffer (*red curve*) samples (Reproduced with permission from Wang et al. 23)

Huang et al. [21]. As depicted in Fig. 7a, sandwich immunoassay was used, and labeled detection antibodies were bound to the analyte that was affinity captured on the sensor surface. By using the PSP modes with increased probing depth, fluorophore labels on the whole area of the analyte were excited and contributed to the sensor signal. This fluorescence readout principle was carried out for *E. coli* O157:H7 and provided the limit of detection of six colony-forming units in one milliliter (cfu/mL). As can be seen in the calibration curve presented in Fig. 7b, the assay was highly specific and required 20 min.

Conclusion and Perspectives

Plasmonics provides highly efficient amplification in important analytical tools based on surface-enhanced Raman spectroscopy (SERS), surface-enhanced infrared absorption spectroscopy (SEIRA), and plasmon-enhanced fluorescence spectroscopy (PEF). In the fluorescence spectroscopy, plasmonic structures supporting propagating surface plasmons can provide enhancement of fluorescence signal by a factor $EF > 10^2$, and in conjunction with immunoassays, it offers the limit of detection at low fM



Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons,

Fig. 7 Plasmonically amplified fluorescence biosensor for detection of bacterial pathogens by using sandwich immunoassay. (**a**) Schematic of the interface architecture and (**b**) calibration curve for the *E. coli* O157:H7 (Reproduced with permission from Huang et al. [21, 24] respectively)

concentrations and the analysis time of several ten minutes. By tuning the probing depth of intense field of surface plasmons, a range of analytes can be detected and various biointerface architectures can be used. In order to translate this method to viable commercial technology, plasmonic sensor chips that are compatible with established fluorescence scanners and microscopes are pursued [25, 26]. For these applications, plasmonic structures that can be prepared by using mass production – compatible methods such as nanoimprint lithography – are of high interest [27]. In parallel, dedicated devices for research applications where fluorescence detection is combined with surface plasmon resonance were described [7]. In addition, the plasmonic amplification of fluorescence signal is subject of implementation to miniaturized devices for sensitive analysis in field that does not rely on using specialized laboratory environment [28].

Cross-References

- LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design
- Photopolymerization Induced by Localized Surface Plasmon Resonance
- Plasmonics
- Quantum Dot
- Surface Plasmon Resonance

References

1. Drexhage, K.H.: Interaction of light with monomolecular dye layers. In: Wolf, E. (ed.) Progress in Optics XII. North-Holland, Amsterdam (1974)

P

2. Chance, R.R., Prock, A., Silbey, R.: Lifetime of an emitting molecule near a partially reflecting surface. J. Chem. Phys. **60**, 2744 (1974)
3. Weber, W.H., Eagen, C.F.: Energy-transfer from an excited dye molecule to the surface-plasmons of an adjacent metal. Opt. Lett. **4**, 236–238 (1979)
4. Knoll, W., Philpott, M.R., Swalen, J.D.: Emission of light from Ag metal gratings coated with dye monolayer assemblies. J. Chem. Phys. **75**, 4795–4799 (1981)
5. Kitson, S.C., Barnes, W.L., Sambles, J.R.: Photoluminescence from dye molecules on silver gratings. Opt. Commun. **122**, 147–154 (1996)
6. Attridge, J.W., Daniels, P.B., Deacon, J.K., Robinson, G.A., Davidson, G.P.: Sensitivity enhancement of optical immunosensors by the use of a surface-plasmon resonance fluoroimmunoassay. Biosens. Bioelectron. **6**, 201–214 (1991)
7. Liebermann, T., Knoll, W.: Surface-plasmon field-enhanced fluorescence spectroscopy. Colloids. Surf. A **171**, 115–130 (2000)
8. Yu, F., Persson, B., Lofas, S., Knoll, W.: Attomolar sensitivity in bioassays based on surface plasmon fluorescence spectroscopy. J. Am. Chem. Soc. **126**, 8902–8903 (2004)
9. Matveeva, E.G., Gryczynski, Z., Lakowicz, J.R.: Myoglobin immunoassay based on metal particle-enhanced fluorescence. J. Immunol. Methods **302**, 26–35 (2005)
10. Sarid, D.: Long-range surface-plasma waves on very thin metal films. Phys. Rev. Lett. **47**, 1927–1930 (1981)
11. Nenninger, G.G., Tobiska, P., Homola, J., Yee, S.S.: Long-range surface plasmons for high-resolution surface plasmon resonance sensors. Sens. Actuators B **74**, 145–151 (2001)
12. Kasry, A., Knoll, W.: Long range surface plasmon fluorescence spectroscopy. Appl. Phys. Lett. **89**, 101106 (2006)
13. Dostalek, J., Kasry, A., Knoll, W.: Long range surface plasmons for observation of biomolecular binding events at metallic surfaces. Plasmonics **2**, 97–106 (2007)
14. Ward, D.R., Hueser, F., Pauly, F., Carlos Cuevas, J., Natelson, D.: Optical rectification and field enhancement in a plasmonic nanogap. Nat. Nanotechnol. **5**, 732–736 (2010)
15. Bharadwaj, P., Novotny, L.: Spectral dependence of single molecule fluorescence enhancement. Opt. Express **15**, 14266–14274 (2007)
16. Bauch, M., Dostalek, J.: Collective localized surface plasmons for high performance fluorescence biosensing. Opt. Express **21**, 20470–20483 (2013)
17. Matveeva, E.G., Gryczynski, Z., Malicka, J., Lukomska, J., Makowiec, S., Berndt, K.W., Lakowicz, J.R., Gryczynski, I.: Directional surface plasmon-coupled emission: application for an immunoassay in whole blood. Anal. Biochem. **344**, 161–167 (2005)
18. Yuk, J.S., MacCraith, B.D., McDonagh, C.: Signal enhancement of surface plasmon-coupled emission (SPCE) with the evanescent field of surface plasmons on a bimetallic paraboloid biochip. Biosens. Bioelectron. **26**, 3213–3218 (2011)
19. Toma, K., Vala, M., Adam, P., Homola, J., Knoll, W., Dostalek, J.: Compact surface plasmon-enhanced fluorescence biochip. Opt. Express **21**, 10121–10132 (2013)
20. Liebermann, T., Knoll, W., Sluka, P., Herrmann, R.: Complement hybridization from solution to surface-attached probe-oligonucleotides observed by surface-plasmon-field-enhanced fluorescence spectroscopy. Colloids Surf. A **169**, 337–350 (2000)
21. Huang, C.J., Sessitsch, A., Dostalek, J., Knoll, W.: Long range surface plasmon-enhanced fluorescence spectroscopy biosensor for ultrasensitive detection of *E. coli* O157:H7. Anal. Chem. **83**, 674–677 (2011)
22. Wang, Y., Dostalek, J., Knoll, W.: Biosensor for detection of aflatoxin M_1 in milk based on long range surface plasmon enhanced fluorescence spectroscopy. Biosens. Bioelectron. **24**, 2264–2267 (2009)
23. Wang, Y., Brunsen, A., Jonas, U., Dostalek, J., Knoll, W.: Prostate specific antigen biosensor based on long range surface plasmon-enhanced fluorescence spectroscopy and dextran hydrogel binding matrix. Anal. Chem. **81**, 9625–9632 (2009)
24. Huang, C.J., Sessitsch, A., Knoll, W., Dostalek, J.: SPR bacterial pathogen biosensors: the importance of fluidic conditions and probing depth. Talanta **122** (2014)
25. Tawa, K., Hori, H., Kintaka, K., Kiyosue, K., Tatsu, Y., Nishii, J.: Optical microscopic observation of fluorescence enhanced by grating-coupled surface plasmon resonance. Opt. Express **16**, 9781–9790 (2008)
26. Bauch, M., Hageneder, S., Dostalek, J.: Plasmonic amplification for bioassays with epi-fluorescence readout. Opt. Express 22(26), 32026–32038 (2014)
27. Boltasseva, A.: Plasmonic components fabrication via nanoimprint. J. Opt. Pure Appl. Opt. **11**, 11 (2009)
28. Zhang, B., Kumar, R.B., Dai, H.J., Feldman, B.J.: A plasmonic chip for biomarker discovery and diagnosis of type 1 diabetes. Nat. Med. **20**, 948–953 (2014)

Plasmonic Amplification of Fluorescence

► Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons

Plasmonic Fluorescence Biosensor

► Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons

Plasmonic Photothermal Therapy (PPTT)

▶ Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites

Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites

Kostiantyn Turcheniuk[1], Charles-Henri Hage[3], Jolanda Spadavecchia[2], Laurent Heliot[3], Rabah Boukherroub[1] and Sabine Szunerits[1]
[1]Institut d'Electronique, de Microélectronique et de Nanotechnologie (IEMN), UMR CNRS 8520, Université Lille 1, Villeneuve d'Ascq, France
[2]Laboratoire de Réactivité de Surfaces, UMR CNRS 7197, Université Pierre & Marie Curie – Paris VI, Ivry-sur-Seine, France
[3]Laboratoire de Physique des Lasers, Atomes et Molécules (PhLAM, UMR 8523), Université Lille 1, Villeneuve d'Ascq cedex, France

Synonyms

Cancer; Gold nanorods; Plasmonic photothermal therapy (PPTT); Reduced graphene oxide

Definition

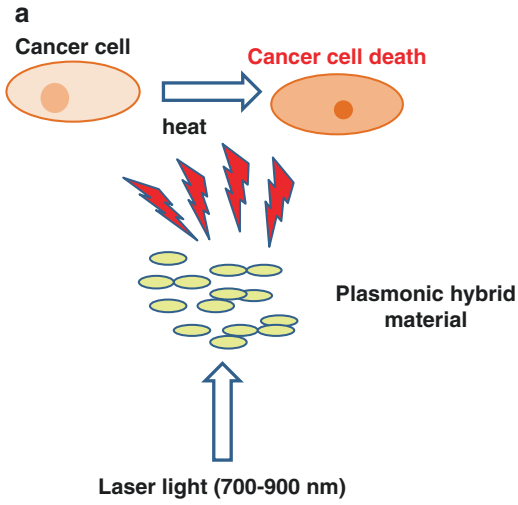
Plasmonic photothermal therapy is based on the use of metallic nanostructures such as gold nanorods as efficient photoabsorbing agent for achieving photothermal ablation of tumors.

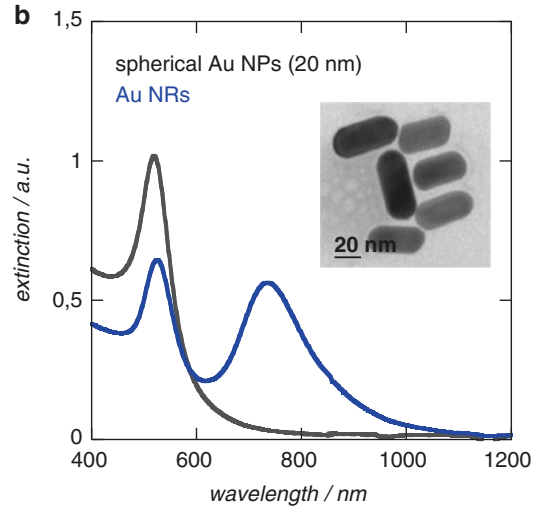
Background

According to the International Agency for Research on Cancer (IARC), the specialized cancer agency of the World Health Organization, an estimated 14.1 million new cancer cases and 8.2 million cancer-related deaths occurred in 2010, compared with 12.7 million and 7.6 million, respectively, in 2008 [1]. The major therapeutic approach for the treatment of a wide range of cancers remains chemotherapy. However, nonspecific delivery of chemotherapeutic agents leads to systematic side effects of normal tissues and limited efficacy in killing the tumor. Indeed, a disadvantage of many drug-based therapies particularly when administered orally or intravenously is that the compound does not localize to the target site. Recently, the use of nanoparticle-based therapy has been proposed for achieving a higher level of precision. The objective of nanoparticle-based therapy is to specifically accumulate the nanostructures in the tumor rather than the normal tissue by engineering the particle's shape, size, and/or chemical property (Fig. 1a). Larger nanoparticles have been shown to be retained in the liver and spleen while passing relatively freely through other organs. This type of passive targeting has been used in radiotherapeutic treatments of liver cancers [2]. Similarly, the fact that tumor vasculature is more permeable than that of healthy tissue can be exploited to concentrate nanoparticles in tumorous tissue due to enhanced permeability effect [3]. A more sophisticated approach is to modify the surface of the nanoparticles with ligands showing high affinity to receptors on tumor or tumor vasculature cell surfaces [4, 5].

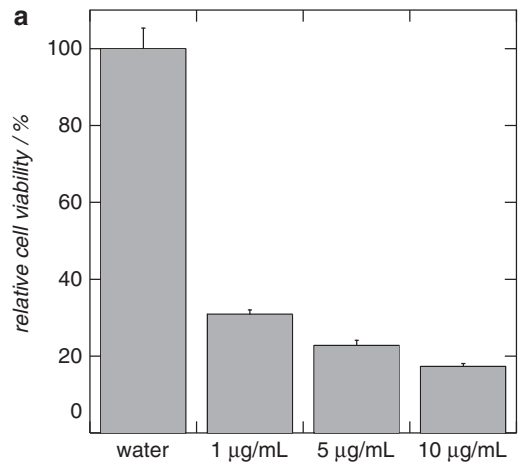
Among the nanoparticles that are currently considered for therapy, gold nanostructures have shown to be particularly well suited for tumor treatments. Spherical gold nanostructures absorb light strongly in the visible region due to the coherent oscillations of the metal conducting electrons in strong resonance with visible light frequencies of light at ≈520 nm (Fig. 1b). Their synthesis is based on the reduction of $HAuCl_4$ in the presence of citric acid, which acts as both reducing and stabilizing agent (capping agent).

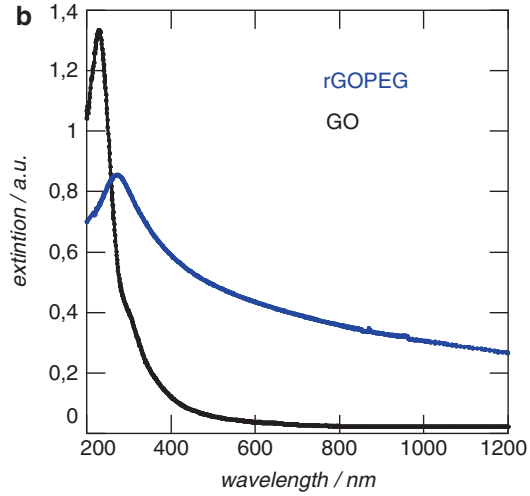
Unlike spherical nanoparticles, more complex shapes such as nanorod or nanoshell structures, the excitation wavelength can be shifted to the near infrared (NIR) with absorption maxima between 700 and 900 nm (Fig. 1b). We have





Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites, Fig. 1 (**a**) Cancer therapy by plasmon photothermal therapy and (**b**) optical extinction spectra of citrate capped spherical gold nanoparticles (*black*) and gold nanorods (*blue*) (inset: TEM image of Au NRs) [6]





Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites, Fig. 2 (**a**) Relative cell viability of MDA-MB-231 cells with different concentrations of CTAB-coated Au NRs. (**b**) UV–vis absorption spectra of GO (*black*) and rGOPEG (*blue*)

shown recently that cetyltrimethylammonium bromide (CTAB)-modified gold nanorods (Au NRs) can be obtained by a simple seed-mediated procedure [6]. The Au NRs are well dispersed in size and shape with an average length of 33 ± 1 nm and width of 13 ± 1.8 nm (TEM image in the inset of Fig. 2b). The UV–vis spectrum displays two extinction bands at 548 and 706 nm. The resonance band at 706 nm corresponds to the longitudinal plasmon oscillations, while the band at 548 nm is due to the transverse plasmon oscillation band, confirming the presence of elongated Au NRs. The NIR spectrum is particularly interesting as water, melanin, and hemoglobin have absorption minima, and light between 700 and 900 nm is most likely to pass directly

through tissues without significant absorption and heat generation [7]. Exposure of Au NR-marked tumors to infrared light will thus result in a strong local light absorption and generation of localized heat (Fig. 1a). The highly efficient production of heat from the absorbed light energy by Au NRs makes these structures of great promise in photothermal therapy (PTT) of cancers and other diseases [8]. Indeed, from the point of cancer therapy, the strong light absorption of Au NRs ensures affective laser therapy at relatively low energies, rendering the therapy method in addition minimally invasive.

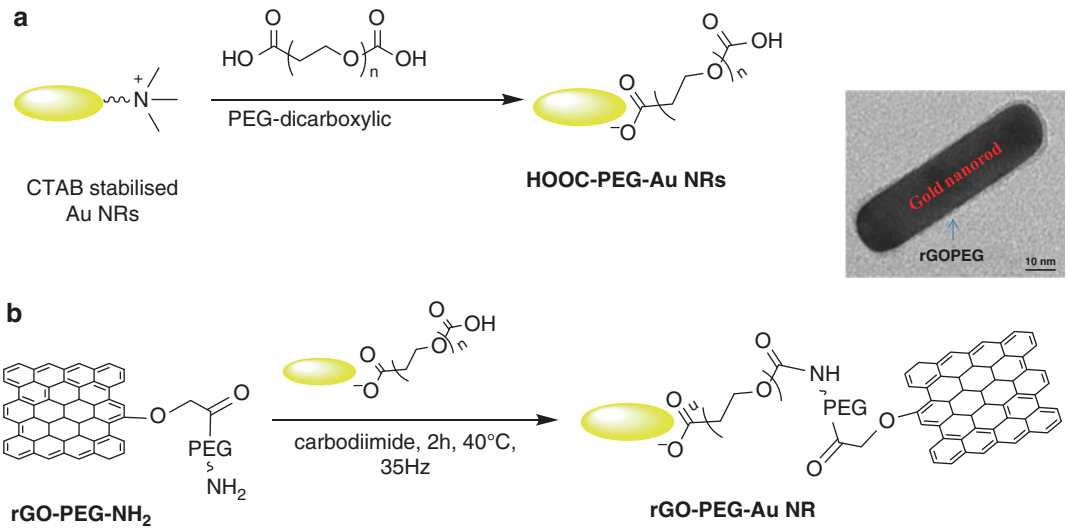
The first report on the use of gold particles as photothermal agent in photothermal therapy was by Halas and coworkers in 2003 [9]. NIR irradiation of gold-on-silica nanoshells led to rise in the temperature of the target region to 40–50 °C, which selectively destroyed the target breast carcinoma cells [10]. Au NRs have generated a significant interest for photothermal applications due to their relatively small size and high theoretical absorption coefficients [11]. However, one of the major concerns using Au NRs was the cytotoxicity of the surfactant, cetyltrimethylammonium bromide (CTAB), used in the manufacturing of the nanorods (Fig. 2a). CTAB molecules can induce cytotoxicity by disrupting cellular membranes [12]. CTAB is not only toxic to cells but also very difficult to remove completely [13, 14]. In addition, CTAB-coated Au NRs lack cancer-targeting capabilities, something of critical importance for active targeting cancer cells.

Graphene oxide (GO) and in particular reduced graphene oxide (rGO) have recently emerged as most promising nanomaterials for photothermal therapy [15–21]. The strong optical absorption across the visible for GO (Fig. 2b) and in the NIR spectrum for water-soluble polyethylene glycol modified rGO (rGOPEG) (Fig. 2b) coupled with high chemical and thermal stability allowed rapid temperature rise and efficient way of heating. GO possesses unique features such as facile synthesis, abundant functional groups, and good biocompatibility making it highly suitable as surface coating. Yang et al. reported one of the first NIR photothermal therapies of cancer cells by using graphene nanosheets functionalized with polyethylene glycol [21]. The effect of the chemistry and lateral size of GO on the photothermal efficiency was investigated by the same group [22].

However, there are some important limitations of graphene-based materials for photothermal therapy. Graphene exhibits some concentration-dependent cytotoxicity effects usually originating from (i) generation of reactive oxygen species [23], (ii) interaction with extremely sharp edges of graphene [24], or (iii) trapping within aggregating rGO sheets [25]. For clinical cancer treatments, graphene concentrations of $\approx$10 μg mL^{-1} or less are recommended [26]. In line with this recommendation, Robinson et al. reported an in vitro photothermal treatment of cancer cells by using 6.6 μg mL^{-1} of rGOPEG using a NIR laser power of 15.3 W cm^{-2} for 8 min [18]. Such high optical power causes efficient ablation of tumor located deeply inside the body but can substantially damage the surface of normal tissues. This limitation was recently overcome by using rGO nanomesh for photothermal treatment of cancer cells with 10 μg mL^{-1} rGO nanomesh at 0.1 W cm^{-1} for 7 min [27].

Coupling Graphene to Gold Nanorods

A different, but highly promising, design of an efficient photothermal agent would be the encapsulation of plasmonic nanostructures with rGO (Fig. 3). Indeed, some groups reported the synergetic effect that hybrid matrices composed of gold nanostructures and GO or rGO have on photothermal energy conversion efficiency [15, 16, 28, 29]. The hybrid materials were prepared by embedding gold particles within a graphene matrix or through electrostatic interactions between positively charged gold nanostructures and negatively charged GO. We have shown lately the interest of rGO covalently modified with gold nanostructures as a photothermal agent. Such a hybrid was synthesized via amide bond formation between amine-PEG-terminated rGO (rGO-PEG-NH_2) and HOOC-PEG-Au NRs (Fig. 3). The synthesis of



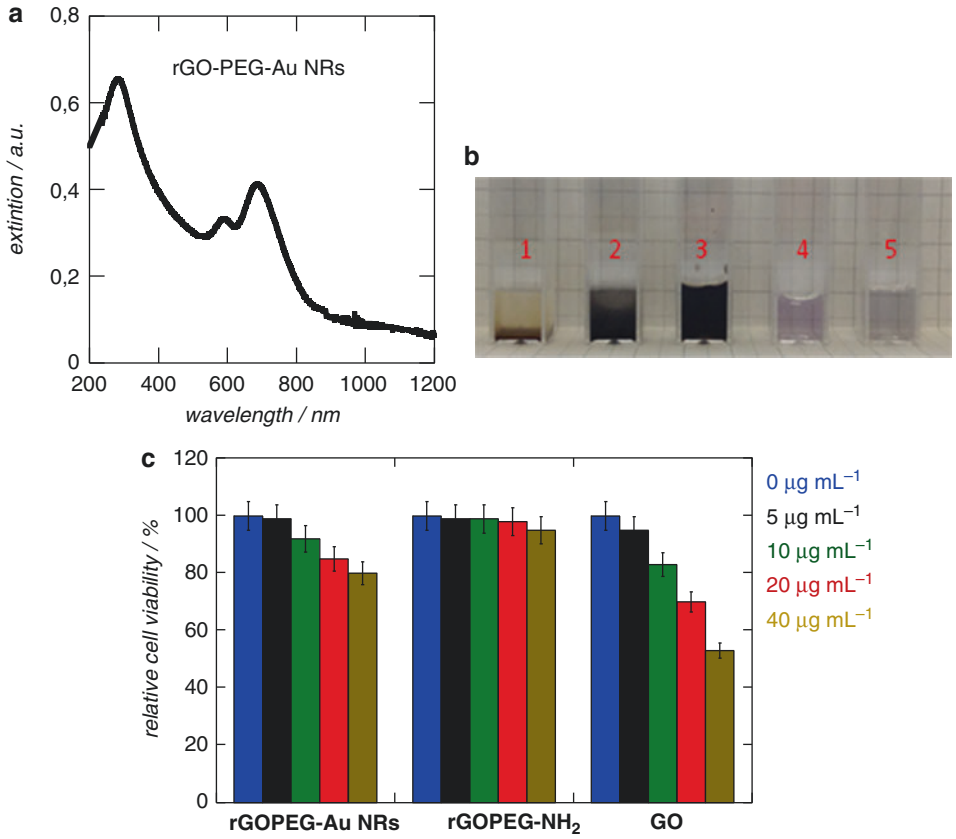
Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites, Fig. 3 (**a**) Synthesis of HOOC-PEG-Au NRs. (**b**) Covalent linkage of rGO-PEG-NH_2 with HOOC-PEG-Au NRs; TEM image of a Au NR coated with rGO-PEG-NH_2

the water dispersible HOOC-PEG-Au NRs was performed following the well-established seed-mediated procedure in the presence of CTAB and dicarboxylic polyethylene glycol in the growth solution [30, 31]. Reaction of carboxylated GO sheets [32] with diamine-PEG (1.5 kDa) through amide bond formation yields highly soluble and stable amine-terminated PEG-rGO of 50–70 nm in size. Reaction of rGO-PEG-NH_2 with HOOC-PEG-Au NRs resulted in Au NRs wrapped with rGOPEG (Fig. 3). A representative TEM image of the rGO-PEG-Au NR hybrid shows the presence of a rGOPEG layer of ≈1.5 nm around the Au NRs.

The UV–vis spectrum of the rGO-PEG-Au NR hybrid displays two bands at 589 and 680 nm corresponding to the transverse and longitudinal plasmon oscillations of the Au NRs, respectively (Fig. 4a). The presence of rGO is furthermore evidenced by the absorption maximum at 270 nm and an overall increase in the absorption from the visible to the NIR. The rGO-PEG-Au NR hybrid (average size of 60 ± 15 nm) is in addition highly dispersible in aqueous media and stable for several months (Fig. 4b). In contrast to the HOOC-PEG-Au NRs, the in vitro toxicity of the hybrid to MDA-MB-231 cells was drastically improved (Fig. 4c). No obvious decrease of cell viability was noticed for MDA-MB-23 cells when incubated with rGO-PEG-Au NRs (10 μg mL^{-1}) for 78 h in strong contrast to Au NRs (Fig. 2a), where at the same concentration, cell viability was around 20 %. The hybrid also proved to be less toxic than GO, but somewhat more cytotoxic than rGO-PEG-NH_2.

Photothermal Efficiency

The photothermal efficiency in solution was studied under NIR illumination at 700 nm (close to plasmon band of Au NRs at 680 nm) in 96-well plates using a pulsed titanium/sapphire laser (chameleon model, coherent). Solutions (1 mL) of the different hybrids were illuminated for up to 10 min at a power of 10 W cm^{-2} and then imaged with an IR camera (ThermoVision A40, ThermaCAM Researcher Pro 2.9 software), allowing measurements of temperature changes (Fig. 5a). Irradiation of equal mass concentrations (10 μg mL^{-1}) of Au NRs and rGO-PEG-NH_2 for 10 min resulted in temperature increase for distilled water from 20 ± 1.5 °C to 32 ± 1.5 °C for rGO-PEG-NH_2 ($\Delta T = 12 \pm 1.5$ °C with $\Delta T =$

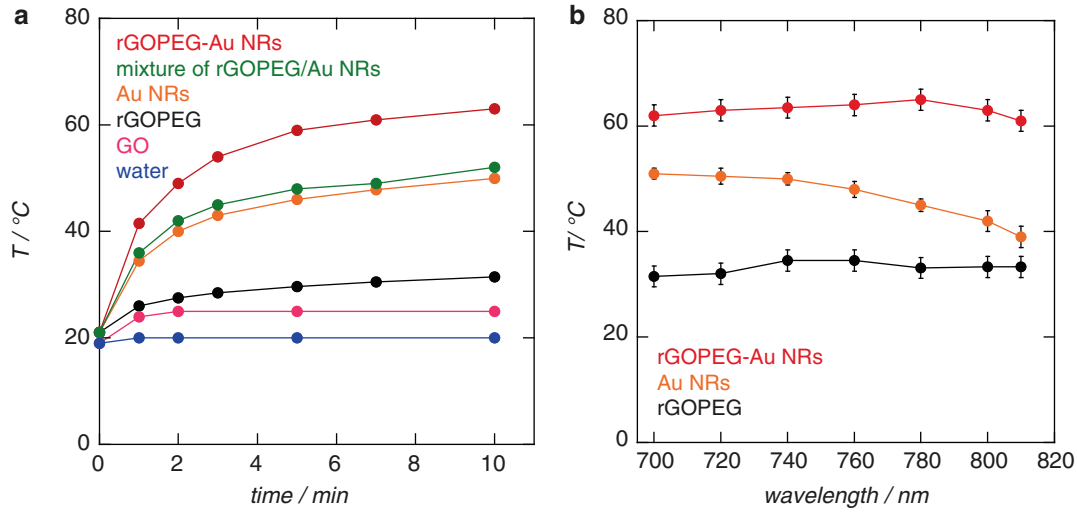


Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites, Fig. 4 (**a**) UV–vis of rGO-PEG-Au NRs. (**b**) Stability of different matrices in DMEM buffer (100 μg mL^{-1}): (1) GO, (2) rGOCOOH, (3) rGO-PEG-NH_2, (4) Au NRs, and (5) rGO-PEG-Au NRs. (**c**) Relative cell viability of MDA-MB-231 cells to incubation with different concentrations (0–40 μg mL^{-1}) of rGO-PEG-Au NRs, rGO-PEG-NH_2, and GO for 78 h

T_{final}−20 °C) and 50 ± 2 °C for Au NRs (ΔT = 30 ± 2 °C). At equal mass concentration, Au NRs are much more efficient photothermal probes. The covalent coating of Au NRs with rGO-PEG-NH_2 induced significant temperature changes, reaching in the case of rGO-PEG-Au NRs 63 ± 2 °C (ΔT = 43 ± 2 °C) after 10 min irradiation. These independent measurements demonstrate the greater photothermal effect of Au NRs coated with rGO-PEG-NH_2 as recently highlighted by Kim et al. [28]. The enhancement of the photothermal effect of the hybrid is attributed to additional light absorption by rGO-PEG-NH_2 to that of Au NRs. Moreover, the covalent linkage of rGO-PEG-NH_2 onto Au NRs (Fig. 3) is crucial for the synergetic effect. Indeed, simple mixing of Au NRs with rGOPEG (same mass ratio 1:1) did not show any photothermal enhancement when compared to Au NRs alone (Fig. 5a, green line). We can, therefore, conclude that the close interaction between the gold plasmons and rGO-PEG-NH_2 is critical for enhanced photothermal effect seen with rGO-PEG-Au NRs.

Interestingly, while as expected the Au NRs show a clear dependence of the photothermal efficiency on the irradiation wavelength, the photothermal profile of the rGO-PEG-Au NR hybrid remains unchanged (Fig. 5b).

The effect of the irradiation intensity on the photothermal efficiency of the rGO-PEG-Au NR hybrid can be evaluated from Fig. 6a. While a maximum temperature of 63 ± 2 °C was achieved

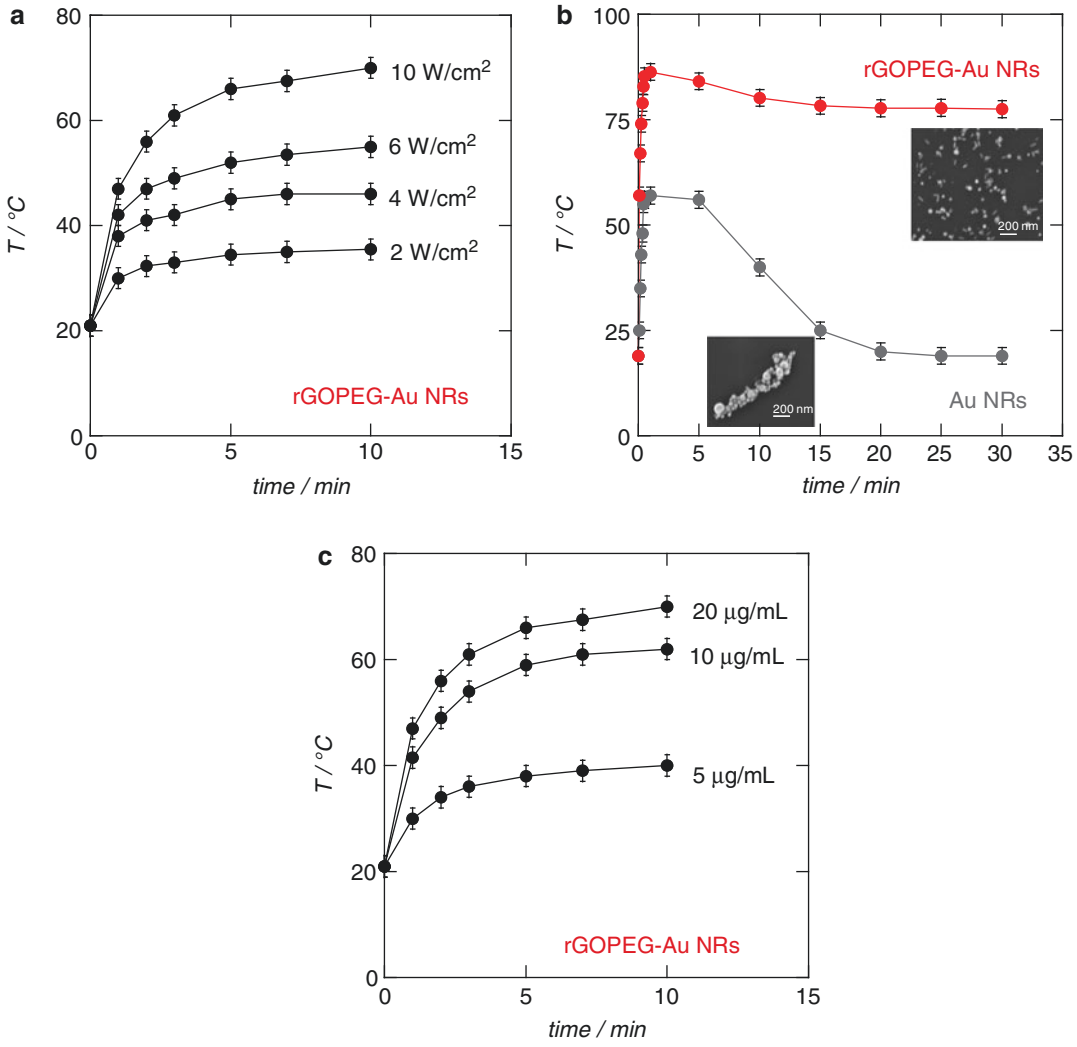


Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites, Fig. 5 (**a**) Photothermal heating curves of the different matrices dissolved in distilled water (10 μg mL^{-1}) under NIR illumination (700 nm) at 10 W cm^{-2} for 10 min using a pulsed titanium/sapphire laser for distilled water alone (*blue*), GO (*pink*), rGO-PEG-NH_2 (*black*), Au NRs (*orange*), mixed Au NRs and rGO-PEG-NH_2 (1/1) (*green*), and rGO-PEG-Au NRs (*red*). (**b**) Wavelength dependence of the photothermal response between 700 and 810 nm

using 10 W cm^{-2} for water, decreasing the laser power to 2 W cm^{-2} resulted in merely 35 ± 1 °C. An additional interesting observation with the hybrid is its high stability under laser irradiation. Indeed, increasing the power density up to 30 W cm^{-2} showed no effect on the hybrid morphology over time, while Au NRs are not stable under such irradiation conditions (Fig. 6b). As seen in the SEM images in the inset of Fig. 6b, gold nanoclusters are formed and the plasmonic heating effect is lost. These results suggest that the rGO-PEG-NH_2 coating not only enhanced the photothermal properties of Au NRs but also protected them from degradation over time. Figure 6c depicts the photothermal response as a function of rGO-PEG-Au NR hybrid concentration of an aqueous solution irradiated at 700 nm for 10 min at 10 W cm^{-2}. The result clearly indicates a concentration dependence response of the hybrid material. According to the data in Fig. 6, the optimal parameters for a photothermal therapy on cultured cells are a laser power of ≈5 W cm^{-2} and a concentration as low as 10–20 μg mL^{-1} of the hybrid.

Conclusion and Perspectives

The use of reduced graphene oxide coatings on metallic nanostructures has shown to have favorable effects on the photostability and photothermal properties. In addition, the presence of PEG units in the core/shell structure makes this multifunctional material well dispersed in aqueous media. The biocompatibility of the material is in addition largely increased and neglectable in vitro toxicity has been achieved. Such nanostructures could thus serve as powerful photothermal agents for in vitro and in vivo plasmonic photothermal therapeutic treatment of cancer cell killing but also for effective killing and deactivation of different pathogens. The emergence of antibiotic-resistance bacterial strains makes the development of alternative strategies for combating bacterial infection of high priority. The use of gold/graphene-based photothermal agents for in-line photothermal sterilization may become possible. Furthermore, the rGO-PEG-NH_2 shell provides anchor points for the conjugation of



Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites, Fig. 6 (**a**) Photothermal heating curves of rGO-PEG-Au NRs (10 µg mL^{-1}) in water at 700 nm for 10 min as a function of the laser power. (**b**) Heating profile of rGO-PEG-Au NRs and Au NRs at a high power density of 30 W cm^{-2} (*inset*: SEM images of the nanostructures after irradiation for 30 min at 30 W cm^{-2}). (**c**) Photothermal heating curves of rGO-PEG-Au NRs at 700 nm for 10 min as a function of rGO-PEG-Au NR concentration (power density of 10 W cm^{-2})

cancer and pathogen receptors (e.g., hyaluronic acid, glycans, etc.) as well as for the loading of drugs and other therapeutic agents. This allows in the long run the formulation of structures where combined therapeutic treatments as well as diagnostic and treatments are feasible.

Cross-References

- ▶ Gold Nanorods
- ▶ Hyperthermia Treatment
- ▶ Photopolymerization Induced by Localized Surface Plasmon Resonance
- ▶ Reduced Graphene Oxide

References

1. Bray, F., Ren, J.S., Masuyer, E., Ferlay: Int. J. Cancer **132**, 1133 (2013)
2. Moghimi, S.M., et al.: FASEB J. **19**, 311 (2005)
3. O'Neal, D.P., et al.: Cancer Lett. **209**, 171 (2004)
4. Kam, N.W.S., O'Connel, M., Wisdom, J.A., Dai, H.: Proc. Natl. Acad. Sci. U. S. A. **102**, 11600 (2005)
5. Oyelere, A.K., Chen, P.C., Huang, X.H., El-Sayed, I.H., El-Sayed, M.A.: Bioconjugate Chem. **18**, 1490 (2007)
6. Spadavecchia, J., Barras, A., Lyskawa, J., Woisel, P., Laure, W., Pradier, C.-M., Boukherroub, R., Szunerits, S.: Anal. Chem. **85**, 3288 (2013)
7. Weissleder, R.: Nat. Biotechnol. **19**, 316 (2001)
8. Huang, X., Jain, P.K., El-Sayed, I.H., El-Sayed, M.A.: Lasers Med. Sci. **23**, 217 (2008)
9. Hirsch, L.R., et al.: Proc. Natl. Acad. Sci. U. S. A. **100**, 13549 (2003)
10. Loo, C., et al.: Technol. Cancer Res. Treat. **3**, 33 (2004)
11. Jain, P.K., Lee, K.S., El-Sayed, I.H., El-Sayed, M.A.: J. Phys. Chem. B **110**, 7238 (2006)
12. Goodman, C.M., McCusker, C.D., Ylmaz, T., Rotello, V.M.: Bioconjugate Chem. **15**, 897 (2004)
13. Niidome, T., Yamagata, M., Okamoto, Y., Akiyama, Y., Takahashi, H., Kawano, T., Katayama, Y., Niidome, Y.: J. Control. Release **114**, 343 (2006)
14. Hauck, T.S., Ghazani, A.A., Chan, W.C.W.: Small **4**, 153 (2008)
15. Xu, C., Yang, D., Mei, L., Li, Q., Zhu, H., Wang, T.: ACS Appl. Mater. Interfaces **5**, 12911 (2013)
16. Zedan, A.F., Moussa, S., Terner, J., Atkinson, G., El-Shall, M.S.: ACS Nano **7**, 627 (2013)
17. Yang, K., Zhang, S., Zhang, G., Sun, X., Lee, S.T., Liu, Z.: Nano Lett. **10**, 3318 (2010)
18. Robinson, J.T., Tabakman, S.M., Liang, Y., Wang, H., Casalongue, H.S., Vinh, D., et al.: J. Am. Chem. Soc. **133**, 6825 (2011)
19. Jung, H.S., Kong, W.H., Sung, D.K., Lee, M.-Y., Beack, S.E., Keum, D.H., kim, K.S., Yun, S.H., Hahn, S.K.: ACS Nano. **28**, 8(1):260–8 (2014)
20. Chen, J., Liu, H., Zhao, C., Qin, G., Xi, G., Li, T., Wang, X., Chen, T.: Biomaterials **35**, 4986 (2014)
21. Yang, K., Feng, L., Shi, X., Liu, Z.: Chem. Soc. Rev. **42**, 530 (2013)
22. Yang, K., Wan, J., Zhang, S., Tian, B., Zhang, Y., Liu, Z.: Biomaterials **33**, 2206 (2011)
23. Zhang, Y., Ali, S.F., Dervishi, E., Xu, Y., Li, Z., Casciano, D., Biris, A.S.: ACS Nano **4**, 3181 (2010)
24. Akhavan, E., Ghaderi, E.: ACS Nano **4**, 5731 (2010)
25. Akhavan, E., Ghaderi, E.: J. Phys. Chem. B **115**, 6279 (2011)
26. Akhavan, E., Ghaderi, E.: Biomaterials **33**, 8017 (2012)
27. Akhavan, E., Ghaderi, E.: Small **9**, 3593 (2013)
28. Lim, D.-K., Barhoumi, A., Wylie, R.G., Reznor, G., Langer, R.S., Kohane, D.S.: Nano Lett. **13**, 4075 (2013)
29. Shi, X., Gong, H., Li, Y., Wang, C., Cheng, L., Liu, Z.: Biomaterials **34**, 4786 (2013)
30. Jana, N.R., Gearheart, L., Murphy, C.: J. Adv. Mater. **13**, 1389 (2001)
31. Jana, N.R., Gearheart, L., Murphy, C.J.: J. Phys. Chem. B **105**, 4065 (2001)
32. Liu, Z., Robinson, J.T., Sun, X., Dai, H.: J. Am. Chem. Soc. **130**, 10876 (2008)

Plasmonic Structures for Solar Energy Harvesting

Jao van de Lagemaat
National Renewable Energy Laboratory, Golden, CO, USA
Renewable and Sustainable Energy Institute, Boulder, CO, USA

Synonyms

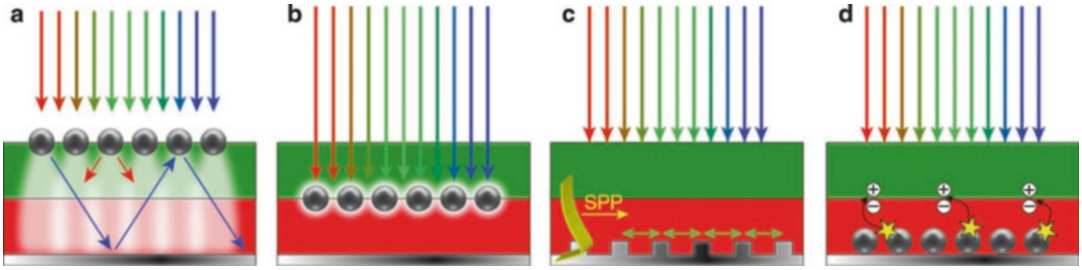
Plasmonics for photovoltaics; Plasmonics for solar cells; Surface-plasmon-enhanced solar energy conversion

Definitions

Plasmonic structures for solar energy harvesting describe the use of surface plasmons to increase the energy conversion efficiency of solar cells.

Introduction

The field of plasmonics describes the interaction of light with matter in or near nanostructured metals and owes its existence to so-called surface plasmons at the interface of metal nanostructures and dielectrics [1]. Plasmonics have found widespread use in analytical chemistry and are being researched for a variety of applications in photonics such as optical computers, nanoscale lasers, and more. Starting in the early 1990s, publications in the field of surface-plasmon (SP) research have seen geometric growth, with almost 4,000 papers



Plasmonic Structures for Solar Energy Harvesting, Fig. 1 Four different strategies proposed for plasmonically enhancing photovoltaics: (**a**) forward scattering and confinement illustrated by the *lighter shaded* areas, (**b**) direct field enhancement, (**c**) waveguiding, and (**d**) creation of surface plasmons inducing altered excited-state dynamics leading to more efficient exciton creation. The photovoltaic device is shown as a generic p-n junction or an acceptor–donor bilayer that is represented by *red* and *green* layers, respectively. Graphic inspired by a figure in Ref. [4]

published in 2010. In contrast, publications on the use of surface plasmons in photovoltaic (PV) applications have only recently taken off, with about 50 papers published per year currently.

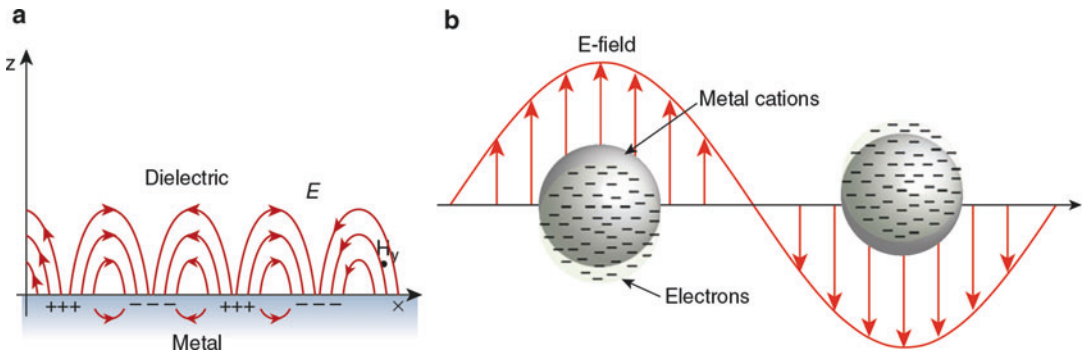
Surface plasmons are collective oscillations of conduction electrons sustained on nanoscale metals. They are well known to enhance optical electromagnetic fields in their vicinity and show extraordinary light-scattering properties. This realization leads to the intriguing idea of using such metal nanostructures in PV applications. The energy of surface plasmons also depends strongly on the shape and size of the nanostructured metals. As a consequence, a solar cell designer can tune the spectral response of SP active systems in such a manner that the solar spectrum is harvested equally well at all wavelengths [2].

Numerous approaches are being pursued to use surface plasmons in PV applications. Figure 1 illustrates these approaches. Firstly, SPs are used to more efficiently couple light into solar cells by efficient forward scattering, as well as efficient internal reflection of light (Fig. 1a). Secondly, the well-known property of SPs where they strongly enhance the optical electrical field is used to enhance optical absorption in the absorber (Fig. 1b). Thirdly, a grating pattern is used that couples light waves into SP polariton waveguide modes in the absorber region of the solar cell (Fig. 1c). Approaches a to c differ mostly by the placement of the plasmonic structures. Lastly, the creation of hybrid exciton plasmon states leads to new light-absorbing states or to altered excited-state pathways that can lead to increased overall efficiency (Fig. 1d). A fifth method to employ surface plasmons in solar energy conversion will not be extensively discussed here. In this method, the surface plasmon absorbs the solar energy directly and a "rectenna" structure is used to tap the electrical energy. Although intriguing, it is currently of little practical relevance but could become so in the future.

Surface plasmons are caused by the interaction of surface charges with the electromagnetic field incident on the surface. They comprise collective oscillations of the conduction electrons in the surface. When one solves Maxwell's equation for the interface between a metal and a dielectric, the dispersion relation for surface plasmons is given by [1]:

$$k_{SP} = \frac{\omega}{c}\sqrt{\frac{\varepsilon_d \varepsilon_m}{\varepsilon_d + \varepsilon_m}}. \quad (1)$$

For SPs to be possible, the complex dielectric constant of the metal ε_m must be opposite in sign of that of the dielectric ε_d on top. With metals, this is possible because ε_m is negative over a large frequency range. Equation 1 also shows that the momentum of the SP mode $\hbar k_{SP}$ is always greater than the momentum of the free-space photon of the same frequency $\hbar k_{fs} = \frac{\omega}{c}$. This leads to the realization that if one wants to generate SPs using



Plasmonic Structures for Solar Energy Harvesting, Fig. 2 (**a**) Schematic representation of a surface-plasmon polariton (Reprinted by permission from Macmillan Publishers Ltd: Nature, Ref. [1], copyright 2003). (**b**) Interaction of light with a spherical nanoparticle. Figure 2b was inspired by Ref. [3]

light, extra momentum must be supplied. Such extra momentum can be supplied by three different methods. The first is by using prism coupling and is of little practical relevance to solar energy conversion. Secondly, extra momentum can be supplied by scattering by a nanoscale feature such as an isolated metal nanoparticle, or a topological defect such as a nanohole. Lastly, periodic structuring of a metal surface (e.g., a grating) at wavelength scales of visible light allows the light to be momentum matched to the SPs on the metal surface. The last two methods discussed above are of great significance to solar energy conversion and are the basis of all strategies for including SP formation in solar cells.

The surface plasmon that arises on a periodically structured metal interface is called a surface-plasmon polariton (SPP). A SPP can be thought of as a photon caught at the interface between the metal and dielectric. These collective oscillations have a combined electromagnetic and surface charge character as shown in Fig. 2a [1]. The SPP waves are transverse magnetic (TM) in character and their electromagnetic field intensity is highest at the surface and decays exponentially away from the interface. The increased electrical field above the surface can be used to enhance optical absorption. They can also be used to guide light along the surface horizontally into the active layer of a device, causing increases of optical absorption because of the improved length of interaction of the light wave with the active layer of the solar cell.

The second type of surface plasmon is observed in isolated nanostructures and is visualized in Fig. 2b. It is easiest explained by visualizing a small metal nanoparticle in the electromagnetic field of an incoming light wave. The electromagnetic field moves the sea of conduction electrons up and down in unison at the frequency of the incoming light. The combination with the stationary positive ions of the metal causes a fluctuating dipole. The resonant oscillation frequency of such a nanoparticle where the movement of the conduction electrons is maximized is determined by several factors: the density of conduction electrons, their effective mass, and the shape and size of the distribution of charges (i.e., the particle size and shape). Metal nanoparticles can show multiple resonances such as the aforementioned dipole resonance, as well as higher modes such as a quadrupole mode. Not surprisingly, the oscillating dipole and multipole modes cause strong electromagnetic fields in the neighborhoods of such particles that can enhance optical absorption and strong light scattering (i.e., the particle reradiates the energy) that can direct light into light-absorbing structures. See Ref. [3] for a comprehensive discussion of nanoparticle resonances and their behavior.

Forward Light Scattering Using Metal Nanostructures

The strategy of using metal nanostructures on the front surface of solar cells is well established and is shown schematically in Fig. 1a. Metal particles

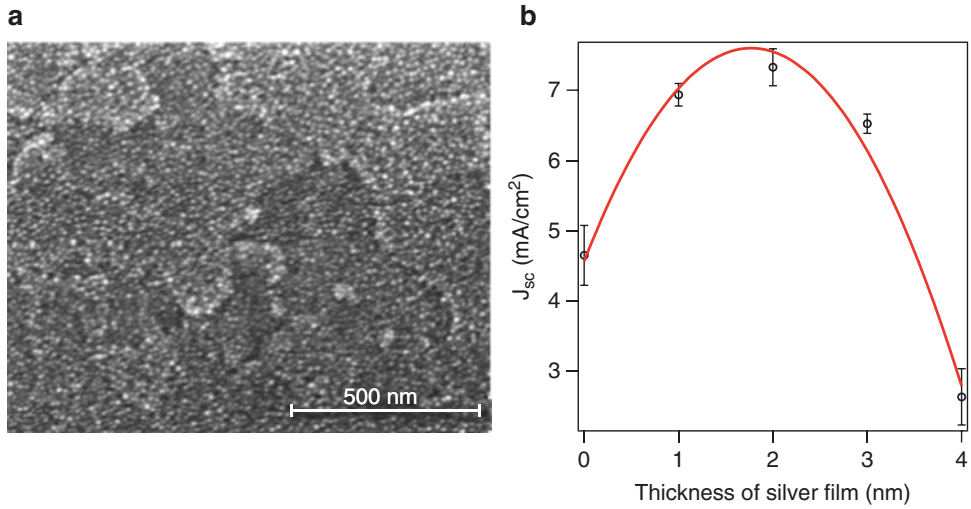
embedded in a homogenous medium scatter light evenly in all directions. However, when such a particle sits at or in the interface between two dielectrics, it will scatter preferentially into the medium with the highest dielectric constant [4]. Analogous to structuring the front surface of solar cells, which is common especially for silicon-based cells, these nanostructures will scatter the light sideways into the active region of the solar cell, allowing a substantial increase of optical path length. At the same time, because a major fraction of the light is now traveling at an angle inside the active layer, it is less likely to escape from the active layer (it is trapped) because it is incident on the top surface again at an angle more shallow than the critical angle. For these reasons, the inclusion of particles at the top surface can cause strong increases in optical absorbance probability by the active layer.

The front-scattering strategy has been extensively studied, especially in silicon photodiodes and solar cells and has been shown to be effective [5, 6]. One question that remains is whether metal particles in such a geometry are more effective than using simple light-scattering structures such as oxide nanoparticles, pyramids, and others. When designing an effective light-coupling strategy, one must weigh the benefit of improved forward scattering with the losses that occur inside the plasmonic nanoparticles due to simple ohmic heating upon absorption of light – something that does not occur in simple dielectric scattering structures. It is known that smaller particles, in principle, are more effective at scattering light into the active layer, but also that ohmic losses are considerable in the smallest particles. Since ohmic (absorption) losses scale more or less with the volume V of the particles, whereas light-scattering scales with V^2, larger particles improve this trade-off [4]. Therefore, one would expect an optimum particle size for scattering efficiency. Catchpole and Polman recently carried out a study of different particle geometries and found that cylindrical and hemispherical geometries are more effective than simple spherical particles, and that parallel dipole orientation of the surface plasmon to the surface can lead to more effective light scattering than the Lambertian value [4]. On the other hand, a direct comparison of the enhancement in the case of metal particles versus that when using silica nanosphere scatterers showed that silica nanospheres provide a similar enhancement to the photocurrent as plasmonic gold nanoparticles. Such silica particles are also good candidates for increasing the light in-coupling in silicon solar cells [7]. This result indicates that possibly there is no convincing case from the perspective of cost for using plasmonic structures for forward-scattering geometries because scattering oxide particles are more economically produced than metal nanoparticles and are also more stable to air exposure.

Direct Field Enhancement Using Embedded Nanostructures

Metal nanostructures can be used to enhance the optical field in their direct vicinity and can therefore be used to enhance optical absorption. The strategy to use this effect is shown schematically in Fig. 1b. The optical enhancement effect has been used widely to enhance the Raman cross section of analyte molecules by orders of magnitude, even enabling the detection of individual molecules. The enhancement of the optical fields around metal nanostructures is usually expressed as an enhancement of the square magnitude of this field because optical absorption processes to a first approximation scale linearly in this quantity. Enhancements of 4 orders of magnitude are attainable, but factors of more than 10 are more common [8].

The direct enhancement approach has been shown to work well in organic solar cell structures where silver particles or other silver nanostructures have commonly been sandwiched between layers or embedded in the active layer of the device [9–12], but also have been used in this manner in inorganic semiconductor-based devices [8, 13]. Figure 3 shows the enhancement of the photocurrent in a bulk heterojunction device fabricated by vapor depositing silver on top of an indium tin oxide conducting front electrode and subsequent deposition of PEDOT:PSS (a transparent conducting polymer), a P3HT/PCBM layer, and a

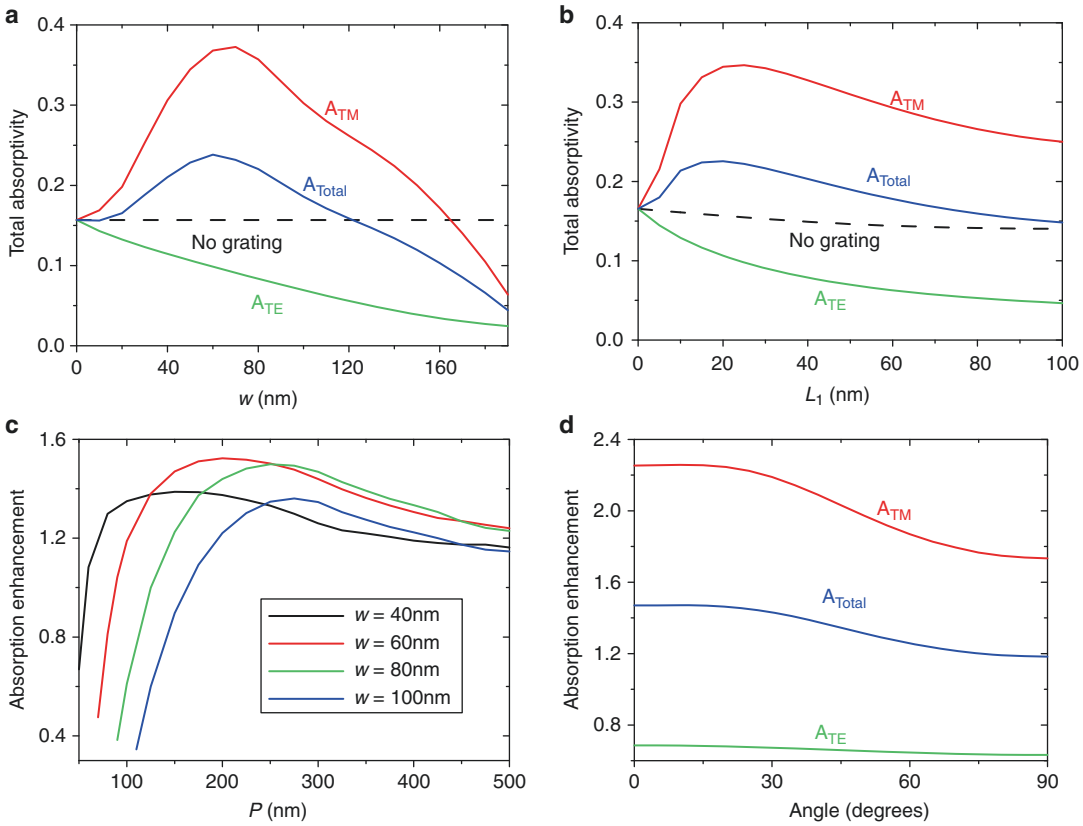


Plasmonic Structures for Solar Energy Harvesting, Fig. 3 (**a**) Silver nanoparticles created by vapor deposition on an ITO electrode. (**b**) Observed enhancement of the photocurrent of an organic photovoltaic device deposited on *top* of the structure shown in Fig. 3a. Data previously published by the author in Ref. [11]

Ca/Al top contact. In such devices, the conversion efficiency can be almost doubled by including silver nanoparticles [11]. In the structures discussed here, especially those where light is incident through the nanoparticle arrays, it is sometimes difficult to separate the effect of field enhancement and enhanced in-coupling of light due to enhanced forward scattering or decreased reflectance. It can be controversial whether any observed enhancement is actually due to field enhancement.

When there is direct contact between the metal nanostructures and the active layer, strong enhancements of the field are expected, but a major problem that can be encountered is that metal nanostructures can act like recombination centers or can quench excited states by energy transfer [10]. This can happen either by capturing both electrons and holes from the semiconductor or by energy transfer from an exciton in the semiconductor to the metal particle, followed by thermalization. Also, in certain PV technologies (such as organic solar cells, which rely on the semiconductor having a low dielectric constant to maintain a constant electrical field over the device to aid in charge separation), the presence of metal nanoparticles can collapse this field because of their very large negative dielectric constant. Direct contact is essential for the greatest absorption enhancements because field enhancement around a metal structure decays approximately exponentially with distance. Therefore, a maximum in the enhancement is expected somewhere in the neighborhood of 5–10 nm, where energy transfer is not efficient but where the evanescent field from the metal nanostructure has not yet fully decayed. This indicates that insulating spacers could be effective when the enhancing particles are in direct contact. Such spacers have typically been provided by buffer layers such as oxides and transparent polymers [11, 13, 14]. Because the spacer layer lowers the expected field enhancement, the benefit of the plasmonically active nanoparticles as direct field enhancers has been limited to small increases in efficiency.

An interesting use of silver nanoparticles that exploits both the optical field enhancement and the strong tendency to cause recombination is to use such particles in the interconnect between two subcells in a tandem device. In this region, charge carriers are supposed to recombine to allow the voltages of the subcell to add up and to mitigate the recombination problem. Such a use of silver nanoparticles has been demonstrated on small molecule organic solar cells [15].



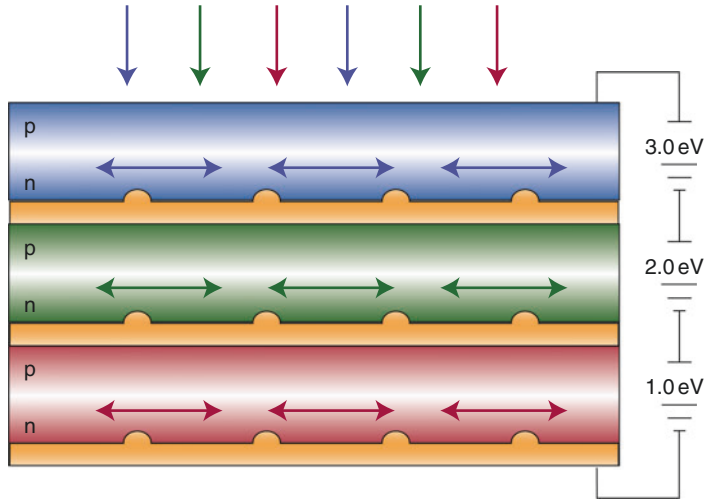
Plasmonic Structures for Solar Energy Harvesting, Fig. 4 Absorption enhancement in a thin-film organic solar cell using CuPc and PTCBI as an active layer by including a metallic grating as a function of the width of the metal lines (w), their thickness (L_1), and period (P) for normally incident light-polarized transverse magnetic (*TM*) and transverse electrical (*TE*) with respect to the grating repeat direction (Reprinted with permission from Ref. [17], Copyright 2010, American Institute of Physics)

P

The aforementioned applications are only scratching the surface of the possible enhancement in solar conversion applications by using direct field enhancement by metal nanoparticles. The enhancement that has been demonstrated to date has mostly been in non-optimized designs, and there is much room for further engineering of this enhancement strategy in organic PV, as well as many other types of solar cells.

SPP Waveguide-Enhanced Photovoltaics

Metallic, plasmonically active grating structures can be used to direct light into waveguide modes in solar cell structures by generating SPPs (see Fig. 1c). This strategy basically rotates the travel direction of incoming light waves 90°, allowing it to travel as SPPs at the interface between a metal contact and the active layer of the device. This allows for extraordinary improvements in the optical path length of light inside the active area of a solar cell. When light is resonant with the plasmon resonance frequency, significant losses are expected from Joule heating of the metal. However for wavelengths to the red side of this frequency, there can be significant gains in absorption efficiency. An analysis of this trade-off can be found in [16]. To illustrate the potentially very large gains that can be made using the SPP waveguide approach, Fig. 4 shows the expected absorption enhancement in an organic solar cell structure that includes a grating structure, which enhances optical absorption of the active layer significantly [17] even if the effect of loss in



Plasmonic Structures for Solar Energy Harvesting, Fig. 5 Plasmonic tandem solar cell concept with layers tuned to different wavelengths of solar light. Semiconductors with different bandgaps are stacked on *top* of each other, separated by a metal contact layer with a plasmonic nanostructure that couples different spectral bands of the solar spectrum into the corresponding semiconductor layer (Reprinted by permission from Macmillan Publishers Ltd: Nature Materials, Ref. [4], Copyright 2010)

the metal structures is taken into account. Such waveguide approaches are expected to be especially useful in solar cell designs where thin layers have to be used either because of charge transport and recombination limitations in the active layer, or because of economical or material supply reasons.

An intriguing idea for SPP waveguide-like modes in a tandem solar cell concept was recently proposed by Atwater and Polman [4]. Figure 5 shows this concept schematically. In each layer, a metallic grating is employed that is tuned to enhance the absorption of only the wavelength of light that is desired in that layer. Such a structure could reduce the required thickness of the active layers of the device, which currently are often deposited using expensive vapor-phase methods and non-earth-abundant elements.

Contrary to the large amount of theoretical investigations into these structures, few have been demonstrated experimentally, and the potential gain in efficiency remains largely theoretical at the moment. Furthermore, a remaining question is whether such grating structures can be fabricated economically over large areas. SPP-active gratings are currently made using relatively expensive methods such as focused-ion-beam etching or e-beam lithography. More promising methods include conformal imprint lithography using rubber stamps that can generate large-area deposition mask patterns, as well as nanosphere lithography and laser interference lithography.

Enhancement of Third-Generation Approaches

A novel approach to employing plasmonics in solar cells is shown schematically in Fig. 1d. In this approach, the presence of surface plasmons fundamentally alters the excited-state dynamics of the photoactive system, driving a photosystem to a normally unlikely outcome. This approach can be especially significant in systems that employ third-generation photoconversion concepts. Such concepts potentially transcend the limits of single-bandgap photoconverters by converting energy that would normally be wasted as heat. Limiting theoretical efficiencies for such solar cells have been estimated to be around 50 % under solar illumination, which is much higher than the limit of about 30 % for a traditional, first-generation solar cell. Examples of such approaches where plasmonic nanostructures can be helpful are multiple-exciton generation (MEG) and singlet

fission (SF). Both of these processes are so-called quantum cutting processes, where a single incoming photon is converted into multiple excited states. In the case of MEG, a semiconductor nanoparticle (i.e., a quantum dot) is used that is so small that carriers in it behave like particles in a box, and the bandgap of the material widens. At the same time, the energy levels go from a continuum to a set of discrete levels. Under the right conditions, it becomes possible for a single highly excited state (i.e., an exciton) to break into multiple excitons at the bandgap energy. If one could design solar cell structures that separate these excitons faster than the recombination of the excitons, a commensurate increase in efficiency could be obtained. Although some promising structures have been demonstrated, it seems likely that new structures and concepts are needed. It has been theorized that surface plasmons can aid in the efficiency of long-range energy transfer between quantum dots and inorganic PV devices [18]. This could potentially lead to efficient energy funnel structures, aided by SP active nanostructures that lead the multiple excitations created by the MEG process away from the originating quantum dot before recombination occurs. At the moment, however, such structures are still purely theoretical.

Singlet fission is another process that can lead to much higher efficiency solar energy conversion. The process is observed in molecular systems and molecular crystals. In short, an excited molecule transfers about half of its energy to another similar molecule or another moiety on the same molecule. The molecule itself will go from the singlet excited state to a triplet state that is about halfway down in energy from the singlet excited state. At the same time and in a coherent fashion, the partner molecule undergoes a transition into the triplet state. The overall double-triplet state still has singlet spin character, making the transition spin allowed. Again, as in MEG, a single excitation is converted into multiple excited states, leading to a decrease in the loss of energy to thermalization of charge carriers. Singlet fission is observed in a large array of different molecular systems, as well as in molecular crystals such as rubrene. It has been postulated that modification of excited-state dynamics can be responsible for more efficient SF processes. Recently, it was observed that pentacene deposited on a silver nanohole array shows more efficient production of triplets and slower triplet–triplet recombination [19]. Possibly, this is due to the surface plasmons creating an efficient pathway to decohere the spin states of the two triplets that start out in coherence.

As mentioned before, SPs can also be used to enhance long-range energy transfer. This effect was recently demonstrated in an organic solar cell structure where light absorbed in a rubrene antenna that had no electrical connection to the cell was effectively transferred to the active device using SPPs resonant in a purposely designed cavity [20]. Such a configuration could be used to design solar cells where light absorption is uncoupled from charge generation and transport.

It is likely that future third-generation photoconversion systems will rely on surface plasmons to increase their likelihood of leading to charge carriers not just by increasing the absorbance probability of light, but also, by fundamentally altering the excited-state dynamics. Using surface plasmons as partner states in such processes is a promising avenue to strongly increased photoconversion efficiencies in the future.

Acknowledgments The author wishes to acknowledge funding from the Division of Chemical Sciences, Geosciences, and Biosciences, Office of Basic Energy Sciences of the US Department of Energy. The submitted manuscript has been offered by an employee of the Alliance for Sustainable Energy, LLC (Alliance), a contractor of the US Government under Contract No. DE-AC36-08GO28308.

Cross-References

- ▶ Active Plasmonic Devices
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanostructures for Energy
- ▶ Nanostructures for Photonics
- ▶ Organic Photovoltaics: Basic Concepts and Device Physics
- ▶ Plasmon Resonance Energy Transfer Nanospectroscopy
- ▶ Surface Plasmon-Polariton-Based Detectors

References

1. Barnes, W.L., Dereux, A., Ebbesen, T.W.: Surface plasmon subwavelength optics. Nature **424**(6950), 824–830 (2003)
2. Cole, J.R., Halas, N.J.: Optimized plasmonic nanoparticle distributions for solar spectrum harvesting. Appl. Phys. Lett. **89**(15), 153120 (2006)
3. Kelly, K.L., Coronado, E., Zhao, L.L., Schatz, G.C.: The optical properties of metal nanoparticles–The influence of size, shape, and dielectric environment. J. Phys. Chem. B **107**, 668–677 (2003)
4. Atwater, H.A., Polman, A.: Plasmonics for improved photovoltaic devices. Nat. Mater. **9**(3), 205–213 (2010)
5. Lim, S.H., Mar, W., Matheu, P., Derkacs, D., Yu, E.T.: Photocurrent spectroscopy of optical absorption enhancement in silicon photodiodes via scattering from surface plasmon polaritons in gold nanoparticles. J. Appl. Phys. **101**(10), 104309 (2007)
6. Pillai, S., Catchpole, K.R., Trupke, T., Green, M.A.: Surface plasmon enhanced silicon solar cells. J. Appl. Phys. **101**(9), 093105 (2007)
7. Sundararaian, S.P., Grady, N.K., Mirin, N.A., Halas, N.J.: Nanoparticle-induced enhancement and suppression of photocurrent in a silicon photodiode. Nano Lett. **8**(2), 624–630 (2008)
8. Mendes, M.J., Luque, A., Tobías, I., Martí, A.: Plasmonic light enhancement in the near-field of metallic nanospheroids for application in intermediate band solar cells. Appl. Phys. Lett. **95**(7), 071105 (2009)
9. Kim, K., Carroll, D.L.: Roles of Au and Ag nanoparticles in efficiency enhancement of poly (3-octylthiophene)/C-60 bulk heterojunction photovoltaic devices. Appl. Phys. Lett. **87**(20), 203113 (2005)
10. Yoon, W.-J., Jung, K.-Y., Liu, J., Duraisamy, T., Revur, R., Teixeira, F.L., Sengupta, S., Berger, P.R.: Plasmon-enhanced optical absorption and photocurrent in organic bulk heterojunction photovoltaic devices using self-assembled layer of silver nanoparticles. Sol. Energy Mater. Sol. Cells **94**(2), 128–132 (2010)
11. Morfa, A.J., Rowlen, K.L., Reilly, T.H., Romero, M.J., van de Lagemaat, J.: Plasmon-enhanced solar energy conversion in organic bulk heterojunction photovoltaics. Appl. Phys. Lett. **92**(1), 013504 (2008)
12. Kim, S.-S., Na, S.-I., Jo, J., Kim, D.-Y., Nah, Y.-C.: Plasmon enhanced performance of organic solar cells using electrodeposited Ag nanoparticles. Appl. Phys. Lett. **93**(7), 073307 (2008)
13. Rockstuhl, C., Lederer, F.: Photon management by metallic nanodiscs in thin film solar cells. Appl. Phys. Lett. **94**(21), 213102 (2009)
14. Standridge, S.D., Schatz, G.C., Hupp, J.T.: Distance dependence of plasmon-enhanced photocurrent in dye-sensitized solar cells. J. Am. Chem. Soc. **131**(24), 8407 (2009)
15. Rand, B.P., Peumans, P., Forrest, S.R.: Long-range absorption enhancement in organic tandem thin-film solar cells containing silver nanoclusters. J. Appl. Phys. **96**(12), 7519–7526 (2004)
16. Dionne, J., Sweatlock, L., Atwater, H.A., Polman, A.: Planar metal plasmon waveguides: frequency-dependent dispersion, propagation, localization, and loss beyond the free electron model. Phys. Rev. B **72**(7), 075405 (2005)
17. Min, C., Li, J., Veronis, G., Lee, J.-Y., Fan, S., Peumans, P.: Enhancement of optical absorption in thin-film organic solar cells through the excitation of plasmonic modes in metallic gratings. Appl. Phys. Lett. **96**(13), 133302 (2010)
18. Govorov, A.O., Lee, J., Kotov, N.A.: Theory of plasmon-enhanced Förster energy transfer in optically excited semiconductor and metal nanoparticles. Phys. Rev. B **76**(12), 1–16 (2007)
19. Johnson, J.C., Reilly, T.H., Kanarr, A.C., van de Lagemaat, J.: The ultrafast photophysics of pentacene coupled to surface plasmon active nanohole films. J. Phys. Chem. C **113**(16), 6871–6877 (2009)
20. Heidel, T., Mapel, J., Singh, M., Celebi, K., Baldo, M.: Surface plasmon polariton mediated energy transfer in organic photovoltaic devices. Appl. Phys. Lett. **91**(9), 093506 (2007)

Plasmonic Waveguides

▶ Light Localization for Nano-optical Devices

Plasmonics

▶ Integration of Nanostructures Within Microfluidic Devices

Plasmonics for Photovoltaics

▶ Plasmonic Structures for Solar Energy Harvesting

Plasmonics for Solar Cells

▶ Plasmonic Structures for Solar Energy Harvesting

Plasmonics in Self-Organized Media

Francesco Bisio[1], Christian Martella[2], Luca Anghinolfi[2,3], Maria Caterina Giordano[2], Michael Caminale[2,4], Maurizio Canepa[2] and Francesco Buatier de Mongeot[2]
[1]CNR-SPIN, Genova, Italy
[2]Dipartimento di Fisica, Università di genova, Genova, Italy
[3]Laboratory for Micro- and Nanotechnology, Paul Scherrer Institut (PSI), Villigen, Switzerland
[4]CEA, INAC–SPINTEC, Grenoble, France

Synonym

Bottom-up fabrication of plasmonic materials

Definition

Plasmonics in self-organized media addresses the characteristics of localized surface plasmon resonances in nanostructured/nanopatterned materials realized by bottom-up fabrication methods.

Fundamentals of Plasmonics

The localized surface plasmon resonance (LSPR) is a resonant oscillation of the free-electron gas confined within a metallic nanostructure driven by an incident electromagnetic (EM) wave. Physically, the charges displaced by the EM field experience a restoring force due to their mutual attraction; when the characteristic oscillation frequency ω_0 associated to the restoring force matches that of the EM field, the resonance condition is met for the plasma oscillation [1]. At the LSPR, the extinction of EM radiation is greatly enhanced as a result of an increase of both scattering and absorption cross sections.

The enhanced scattering arises from the reradiation of the EM field that is induced in the metallic medium, whereas the absorption is related to the non-radiative decay of the plasma oscillations that typically leads to the dissipation of the power associated with the EM wave within the metallic nanostructures. The accumulation of a large net charge in proximity of the nanostructures' surfaces at the LSPR frequency leads to large enhancement of the local EM field (near-field regime) in the immediate proximity of the structures' surface with respect to its free-space value. Such an enhancement is exploited for nonlinear optical spectroscopies, like surface-enhanced Raman scattering (SERS), whose yield scales nonlinearly with the local field, thereby promoting plasmonic systems as platforms for high-sensitivity molecular detection [2].

Since the LSPR frequency, amplitude, and linewidth depend upon a multitude of electronic and geometric effects, controlled modifications of the nanostructures' constituent materials (electron density, material's losses), dielectric environment (supporting substrates, host media, adsorbed molecules), and geometrical parameters (nanostructures' size and shape) allow to flexibly tune the plasmon's characteristics. Such a tunability probably ranks as the single most relevant aspect that contributed to establish plasmonics as an invaluable tool in fundamental and applied photonics research. In order to make the most of plasmonics' numerous applications, it is crucial to tailor the LSPR characteristics according to the desired requirements. For plasmon-enhanced spectroscopies or optical-signal waveguiding, matching the LSPR frequency against narrow-linewidth excitations is required; for plasmon-enhanced photovoltaics, broadband plasmon features are often more advantageous. In the following, we will briefly introduce the main concepts of plasmonics, the key factors that contribute to the LSPR tunability, and the configurations that are typically fabricated for maximizing the EM-field enhancement ratio for nonlinear spectroscopy applications.

Let us consider an ellipsoidal metallic particle with semiaxes a_γ ($\gamma = x, y, z$) oriented along the Cartesian axes, immersed in a transparent dielectric host and far from any other polarizable entity. Let the dielectric constants of the metal and the host be ϵ_m and ϵ_h, respectively. When the particle is much smaller than the wavelengths $a_\gamma \ll \lambda$ (quasi-static approximation, QSA), we can

neglect spatial retardation effects. If the host is homogeneous and isotropic, then $\boldsymbol{E}_{loc} = \epsilon_h \boldsymbol{E}_0$, and we can write:

$$\boldsymbol{p} = \epsilon_0 \epsilon_h \boldsymbol{\alpha} \otimes \boldsymbol{E}_0 \tag{1}$$

where $\boldsymbol{\alpha}$ is the polarizability tensor of the particle. For ellipsoidal particles, α is diagonal, with principal values α_γ given by [1]

$$\alpha_\gamma = v \frac{\epsilon_m - \epsilon_h}{\epsilon_h + L_\gamma(\epsilon_m - \epsilon_h)} \quad \gamma = x, y, z \tag{2}$$

where $v = (4/3)\pi a_x a_y a_z$ is the particle's volume and L_γ are the *depolarization factors* that can be written as

$$L_\gamma = \frac{a_x a_y a_z}{2} \int_0^\infty \frac{dq}{\left(q + a_\gamma^2\right) \sqrt{\prod_{\eta=x,y,z}\left(q + a_\eta^2\right)}} \tag{3}$$

and satisfy the sum rule $\sum_\gamma \mathrm{L}_\gamma = 1$.

For spherical particles, we have $a_x = a_y = a_z \equiv a$, and the geometrical factors reduce to $L_\gamma = 1/3$ in all directions: the polarizability is isotropic, and from Eq. 2 we find

$$\alpha_{\mathrm{sph}} = 3v \frac{\epsilon_m - \epsilon_h}{\epsilon_m + 2\epsilon_h} \tag{4}$$

If $\mathcal{I}m[\epsilon_m(\omega)]$ is small or slowly varying, the resonance condition for spherical particles simplifies to

$$\mathcal{R}e[\epsilon_m(\omega)] = -2\epsilon_h \tag{5}$$

which is called the *Fröhlich condition* for the LSP resonances.

Another interesting case is given by the cylindrical limit, i.e., when $a_z \gg a_x, a_y$. Now, the depolarization factor along the main axis (L_z) vanishes, implying that the cylinder can be polarized only by electric fields lying in the *xy* plane. For the case of a prolate ellipsoid, the LSPR in the longitudinal geometry (electric field polarized along the major axis) redshifts as the aspect ratio (defined as the ratio of the major to the minor axis) increases, while the LSPR in transverse excitation (electric field polarized along the minor axis) correspondingly blueshifts.

Beyond the quasi-static regime, retardation effects become relevant, and multipolar charge distributions set in due to the inhomogeneous EM-field distribution over the nanoparticle (NP). Under these conditions, the optical response of a metallic NP is determined by the superposition of LSPR corresponding to the different orders of the multipolar charge distribution, giving rise to broader resonances; for increasing NP size, the LSPR of all orders are monotonically redshifted.

The LSPR are heavily influenced by the dielectric characteristics of the surrounding medium, which can be simply composed of a homogeneous dielectric host, or include other particles within the near-field EM-coupling range (i.e., with a mutual spacing in the order of 1/10 of the LSPR wavelength or below). As apparent from Eq. 5, an increase in the NP-host dielectric constant translates into a LSPR redshift. More in detail, the LSPR redshift exhibits an exponential decay upon the thickness of the medium surrounding the particle, meaning that the LSPR is sensitive to the variation of dielectric function within few tens of nanometers from the NP surface [3]. This local sensitivity forms the basis of the exploitation of LSPR for sensing of adsorbed molecules, provided selective binding sites at the surface of the metal nanoparticles are created by functionalization [4].

Whenever NPs are brought closer and closer, the mutual influence of the NP local fields affects the collective LSPR characteristics and the EM-field distribution. The most striking and interesting example of this many-particle effects is provided by the so-called dimer configuration, where two identical plasmonic NPs are placed close to each other, few nanometers apart. In such a configuration, an enhanced field *hot spot* is generated within the interparticle gap at the collective LSPR frequency. Series of NPs aligned in linear chains can be built as infinite series of dimers, thus generating a multitude of hot spots, and analogously so for NP arrays, that thus become prototypical substrates for plasmon-enhanced nonlinear spectroscopy [5].

Fabrication Issues

The fabrication schemes suited for realizing plasmonic nanostructures can be grouped for convenience into two broad categories, the so-called "top-down" and "bottom-up" approaches. Top-down methods include all the diverse forms of lithography that allow the fabrication of nanostructures with a high degree of morphological precision, albeit sometimes at the expense of a higher price and lower throughput. These methods are called for in those specific applications for which the precision and control in the NP shape and/or in the architecture of many-particle systems is of paramount importance. This includes cutting-edge research applications like ultrahigh-sensitivity plasmonic detectors and sub-wavelength optical waveguiding.

Bottom-up methods instead exploit spontaneous processes to produce nanostructured materials. They target the achievement of metastable configurations characterized by the tendency to develop regular patterns on specific, system-dependent length scales, harnessing the process' kinetics. Bottom-up strategies are typically employed in all those circumstances in which large-area plasmonic systems must be grown at low fabrication costs and in which the physical effect under investigation is not vulnerable to the unavoidable morphological disorder. Clearly, while bottom-up approaches cannot challenge lithography in terms of the variety and degree of precision of the achievable nano-architectures, specific niche applications exist in which spontaneously fabricated structures retain a significant advantage over lithography, even aside from fabrication-cost issues. Some significant examples are related to applications in which a broad size distribution and the consequent broadband plasmonic response provide an inherent advantage. In this respect we can, e.g., mention disposable plasmonic sensors, photon outcoupling in light-emitting devices, or broadband photon harvesting in thin-film solar cells induced by a self-organized arrays of plasmonic scatterers whose resonances match the solar spectral range [6]. In addition, bottom-up methods are typically more flexible material wise and in general lend themselves to be more easily applied under ultra-controlled conditions (e.g., ultrahigh vacuum), thereby allowing to target reactive metals as well.

Self-organized materials naturally exhibit a certain degree of morphological disorder (NP size, shape, geometrical arrangement), as a result of the spontaneous amplification of fluctuations in the systems' morphology which are the basis of the self-organization process itself. Such a dispersion in the morphological parameters of the constituent building block can be fully probed in ensemble measurements. As far as the plasmonic response is concerned, disorder impacts the system properties in two manners: ensembles of plasmonic nanostructures which are sufficiently sparse, i.e., located outside one another's near-field coupling range, exhibit a plasmonic response determined by the incoherent summation of the individual response (inhomogeneous broadening). If the nanostructures' size distribution is large enough that size-dependent redshift and broadening of the plasmonic response are present, the collective response will fully bear trace of these effects. Conversely, in ensembles of near-field-coupled nanostructures, superradiant dipolar modes become effectively delocalized over several NPs, smoothening the effects of morphological disorder and achieving an ensemble response that is spectrally narrower than the incoherent superposition of all the constituents' individual response [7]. Depending upon the application envisaged (broadband response over spectrally narrow features), one of the two approaches might be followed.

Plasmonics systems are composed of a metallic and an insulating counterpart. The insulator, supporting or even embedding the nanostructured metallic material, guarantees the required electron confinement. Different bottom-up fabrication approaches may involve the direct nanopatterning of the metal, regardless of the insulator morphology, or the preliminary patterning of the insulator that later acts as a template for realizing metallic nanostructures. We will in the following discuss examples from both classes.

Plasmonic Arrays by Direct Ion-Beam Sputtering

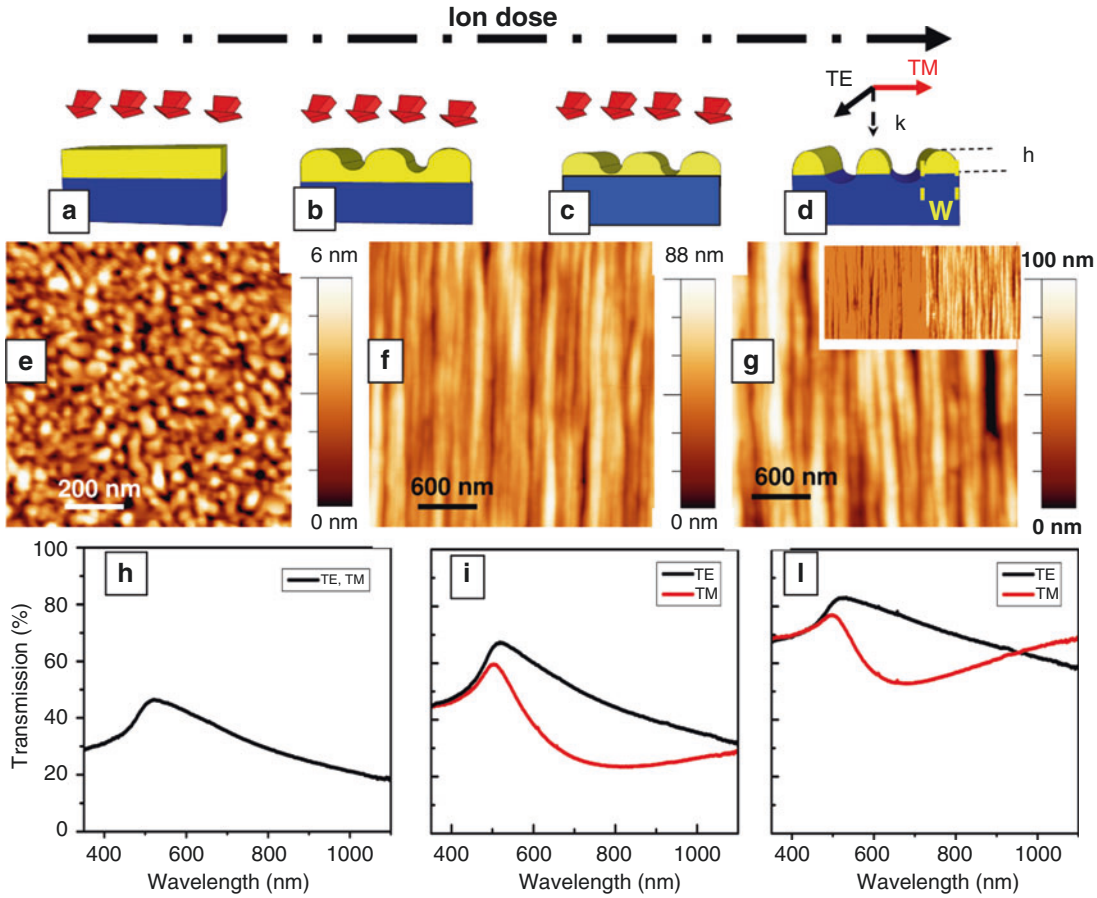
Among the approaches which have been exploited for the bottom-up fabrication of nanoscale patterns, the use of defocused ion-beam

sputtering (IBS) has gained a growing attention due to the ease of implementation, the flexibility, and the scalability over macroscopic substrate areas employing standard industrial equipment. Planar substrates of a wide class of materials, after irradiation with noble gas ions at energies in the range from few hundreds of eV to several tens of keV, develop periodic patterns when the ion fluency $\mathcal{F}$ exceeds 10^{16}–10^{17} ions/cm^2; depending on the sputtering conditions, and in particular on the incidence angle of the ions, a rich variety of patterns can be observed [8]. In an oversimplified classification, near normal incidence, a smoothing regime is predominant for single component systems, while, for some compound semiconductor substrates, close-packed arrays of dots or mounds can be formed [9]. At higher incidence angles, below a critical value θ_c (in the range of 60–70°), amorphous materials like glasses or materials which become amorphized following ion impact (semiconductors) develop a ripple pattern with ridges running perpendicular to the ion projection. For grazing incidence angles above θ_c, the ripple pattern is instead rotated by 90° with ridges running parallel to the ion projection. A classical model describing the early stages of pattern formation (small amplitude approximation) of an initially flat substrate under IBS conditions has been proposed by Bradley and Harper in the continuum limit [10]. The basic idea of the model is that a periodic ripple undulation arises as the result of the interplay between the destabilizing action of the ion beam, which tends to roughen the surface, and the stabilizing surface relaxation mechanisms which reduce the interface width. The erosion velocity at a given surface point turns out to depend on the local curvature of the topographic profile and is faster for the valleys (regions with positive curvature) than for the peaks (regions with negative curvature). The smoothing terms which counteract ion-induced roughening instead arise from thermally and/or ion-induced diffusion of the mobile defects which are determining surface relaxation. At higher ion fluencies, the simplified small amplitude approximation does not hold anymore, and nonlinear terms must be taken into account in order to describe the experimental observations within the continuum approximation [11]. The general lines of this description hold for a broad class of transparent dielectric materials employed as substrates in low-cost photonic and plasmonic applications such as soda lime and borosilicate glasses, silicon, silica, and alumina.

The case of metal surfaces and films, of interest in plasmonic applications, is more complex because the crystalline properties of the substrate are preserved even after prolonged sputtering at low ion fluxes and at sufficiently high temperature (*diffusive regime*). Thermally activated surface diffusion of mobile defects is in fact no more isotropic, but oriented along the high symmetry directions of the substrate unit cell. Additionally, the presence of an extra energy barrier for defect crossing at atomic step edges (the Ehrlich-Schwoebel barrier) biases mass transport in the uphill direction so that adatoms are preferentially attracted by the *ascending* step edges while vacancies by the *descending* step edges; as a consequence of the *kinetic* competition between the destabilizing uphill currents and the diffusive relaxation currents (Villain instability), a periodic mounded pattern develops in the direction of the dominant uphill current [12]. At high ion fluxes and/or low temperature, the anisotropic diffusive relaxation mechanisms are hindered, and, in analogy to amorphous substrates, the pattern formation process is dominated by the erosive action of the ion beam. Thus, under *erosive* IBS conditions (high ion fluxes and/or low substrate temperature at grazing incidence), both single-crystal [8] and polycrystalline metal surfaces can develop an undulated ripple pattern-oriented parallel to the ion-beam projection [13].

Nanowires by Direct IBS

The steps leading to the transformation of a rippled polycrystalline metal film into a self-ordered one-dimensional array of disconnected nanowires (NWs) by *direct* IBS patterning are outlined with the help of the sketches of Fig. 1a–d and of the atomic force microscopy images in Fig. 1e–g. The initial morphology of the pristine metal film is characterized by the size of the polycrystalline grains [14]; in the example shown in Fig. 1e, thermal evaporation of a 150 nm Au film leads



Plasmonics in Self-Organized Media, Fig. 1 (**a**–**d**) Sketches of the direct IBS at $\theta = 82°$ of a polycrystalline Au film as a function of the ion dose. Low doses: thick film (**a**) and rippled pattern (**b**). Higher doses: disconnection threshold (**c**) and arrays of laterally disconnected Au NWs (**d**). (**e**–**g**) AFM topographies corresponding to the sketches (**a**, **c**, **d**). (**h**–**l**) Optical transmission spectra corresponding to the topographies (**e**–**g**), for light polarization parallel (TE) and perpendicular (TM) to the ripple ridges

to the formation of grains with an average size in the 80 nm range.

Increasing $\mathcal{F}$ beyond 10^{16}–10^{17} ions/cm^2, elongated ridges are formed which can exceed several micrometers in length, with a lateral periodicity in the range of 100 nm and with an amplitude which can reach several tens of nm (Fig. 1b). Extinction (absorption) spectroscopy of polarized light in the VIS-NIR spectral range allows to follow the evolution of the optical properties of the patterned Au film and to correlate them in a straightforward way to the morphological modifications induced by the IBS process. The optical spectra of samples irradiated at increasing ion doses are shown in Fig. 1h–l. Two characteristic linear polarizations of the incident light, i.e., transverse magnetic (TM – electric field perpendicular to the ripple's ridges, red traces) and transverse electric (TE – electric field parallel to the ripple's ridges, black traces), are used. In Fig. 1h, one can see the optical spectra of a still connected rippled film corresponding to the sketch in Fig. 1b. The spectral signatures are similar to those of a flat Au metal film with no evidence of polarization dependence. For $\mathcal{F}$ in the range of 10^{18} ions/cm^2, film erosion increases up to the point that the ripple valleys reach the dielectric substrate, and the metal film decomposes into an array of laterally disconnected NWs separated by linear nanogaps as shown in Fig. 1c and f. The NW array is

characterized by a lateral periodicity in the range 100–120 nm, while the length of individual NWs can exceed several micrometers. The individual nanowires are composed by polycrystalline grains with predominant (111) in-plane texture and with densely packed domain boundaries [14]. The corresponding optical spectra shown in Fig. 1i reveal that at the onset of Au NW disconnection, a polarization dependent extinction minimum becomes evident in the TM transmittance spectra, a typical fingerprint of the LSPR [15]. The TE spectra still resemble the behavior of a connected Au film, since metal NWs are elongated for several micrometers in the long-axis direction. It is worth to notice that the plasmonic extinction band centered around 800 nm spans a broad range from about 550 to 1000 nm. This is a consequence of the intrinsic dispersion of the lateral size (w) of the Au NWs in the early stages of the disconnection process (inhomogeneous broadening) and of a redshift of the dipolar LSP resonances due to near-field coupling between adjacent NWs separated by small gaps. Further increasing $\mathcal{F}$ in the 10^{19} ions/cm^2 range, the average NW width w is reduced, and the interwire gap amplitude is increased from few nm up to several tens of nm as evidenced in Fig. 1d. The inset in Fig. 1g evidences on the left side a region in which the NW array has been removed by means of a micrometric tip thus exposing the underling glass substrate with deep trenches where local erosion has occurred in correspondence to the gaps. Looking at Fig. 1l, one observes a progressive increase of the transmittance due to the thinning of the metal film and a narrowing of the LSP resonance band essentially related to a reduction of the NW size distribution and to an increase of the gap width. Moreover, as the average w/h out-of-plane aspect ratio of the NWs reduces, a blueshift of the plasmonic band down to 650 nm occurs. This example highlights the potential of *direct* ion-beam sputtering as a self-organized nanopatterning approach capable of transforming a polycrystalline metal film supported on a low-cost dielectric substrate into an array of nanowires with broadband plasmonic response. Tailoring of the LSP resonance is possible by changing the ion dose and/or the metal (e.g., Au, Ag, Cu, or Al).

Near-Field Enhancement and SERS

IBS-fabricated NW arrays naturally lend themselves to be applied as substrates for plasmon-enhanced SERS, thanks to the high density of *hot spots* generated in the nanometric gaps between the NWs. Recurring to commercial software for EM solving, one can calculate the near-field intensity and distribution in the proximity of arbitrary shaped and spaced metal nanoparticles. As an example, one can consider two identical gold prolate nanoellipsoids placed in close proximity side by side; when the nanocavity width reads 2 nm, at the LSP resonant condition, the EM intensity in the gap region is enhanced almost 50-fold and rapidly drops to almost zero as the nanocavity width is increased to 60 nm. The tailored near-field coupling of the plasmonic nanostructures thus represents one of the most challenging issues for the development of high-sensitivity spectroscopy applications exploiting the fluorescence and/or Raman signal.

In this context, the local field enhancement in the vicinity of the NW arrays of Fig. 1 has been probed detecting the SERS signature of a monolayer of methylene blue molecules chemisorbed inside the nanocavities [16] and directly mapping the near-field distribution in the vicinity of the metal nanostructures acquired via scanning near-field optical microscopy (SNOM). An enhancement factor of the order of 10^4 is found for the Raman signature when the pump laser wavelength matches the plasmonic resonance of the nanowires and when light is polarized along the nanocavity axis (i.e., along the NW's short axis). The enhancement factor rapidly drops by almost 2 orders of magnitude when the polarization of the pump laser is rotated by 90°, i.e., parallel to the NW long axis, so that the LSP resonance is no more excited. The possibility to switch off the LSP resonance by simply rotating the laser pump polarization allows to separate the chemical and electromagnetic contribution to the Raman enhancement factor since the same sample location and molecular ensemble are probed. Furthermore, the SERS signal turns out to be polarized along the nanocavity axis independently from the polarization of the pump. This proves that only the component of the molecule's Raman dipole

parallel to the nanocavity axis is selectively excited by the plasmonic resonance and that the measured SERS signal arises from those molecules which are located within the narrowest nanocavities, i.e., those hot spot regions in which the NWs are near-field coupled.

It is also worth to highlight that the broad and tunable plasmonic band of the gold NWs is interesting for promoting photon harvesting in photovoltaic devices through light trapping and scattering [6]. In fact, the LSP resonances of the self-organized arrays of plasmonic nanoparticles overlap a large part of the solar spectrum due to the intrinsic size dispersion of the NW array. Even more interestingly, the NW arrays form a percolated network which combines anisotropic electrical conduction with sheet resistance as low as few Ohm/square and a high optical transparency in the 70–80 % range [17]. NW arrays can thus be considered as an alternative to ITO-based transparent conductive oxides with the additional benefit of supporting LSP resonances tunable over the VIS-NIR solar spectral range.

Template-Driven Fabrication of Plasmonic Nanostructures

The template-driven fabrication of plasmonic nanostructures exploits an existing morphological nanopattern on the insulator's surface to steer the formation of metallic nanostructures. Thus, when metal atoms are deposited on the insulator nanopattern, the template morphology acts as a partitioner for the deposited material, forcing the fractioning of an otherwise continuous metal film into disconnected plasmonic units, with size and periodicity dictated by the regular spatial patterns of the template. The shape of the metal nanostructures, now all roughly similar in volume, might then be dictated by the kinetics and geometry of the deposition process or further modified by bringing the system to a thermodynamically stable state.

Methods to fabricate the insulating template include, among others, the patterning by ion-beam sputtering of an initially flat insulator (crystalline or amorphous), the miscut or faceting of insulating single-crystalline materials that exploits in various ways the strong dependence of the surface energy upon the crystallographic orientation in single-crystal materials, or the self-assembly of arrays of close-packed nanospheres. Fabricating the metallic (plasmonic) nanostructures on top of a nanopatterned insulator typically relies on the deposition of a nominally uniform metal layer, sometimes in peculiar geometric configuration (e.g., glancing angle deposition) to boost the conformality to the template. The deposition might be sometimes followed by post-deposition treatments such as annealing.

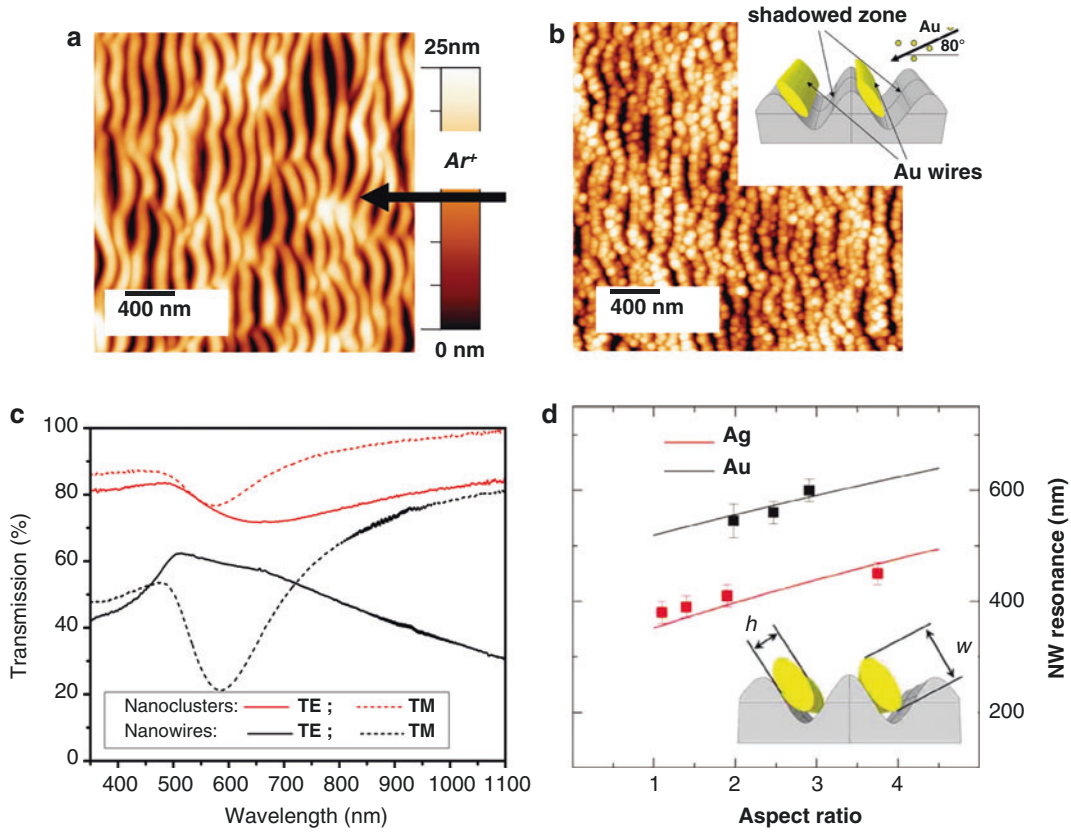
The use of prestructured dielectric templates allows to circumvent some of the inconvenients that arise when performing the *direct* patterning of a metal. For example, the synthesis of ordered metal NW arrays by *direct* IBS nanopatterning depends crucially from the self-organization properties of the metal film, while the flat supporting substrate plays a negligible role. This means that adopting the same direct IBS patterning conditions (ion energy, ion flux, ion incidence angle), the morphological parameters of the NW distribution will be different when different materials like Ag, Cu, or Al are employed [14, 13].

In the following, we describe an alternative approach for the synthesis of arrays of plasmonic nanoparticles which does not depend on the specific properties of the metal employed. The morphological parameters of the array can in fact be constrained geometrically if nucleation of the metal nanoclusters is achieved by glancing angle deposition on a faceted template, according to the sketch in the inset of Fig. 2b or Fig. 3c.

In particular, in view of plasmonic applications, the method allows a great flexibility in the choice of noble metal clusters such as Au, Ag, and Cu which are characterized by a plasmon resonance in the visible-NIR spectral and Al in the UV spectral range.

Arrays of One-Dimensional Structures on Templates Prepared by Ion-Beam Sputtering

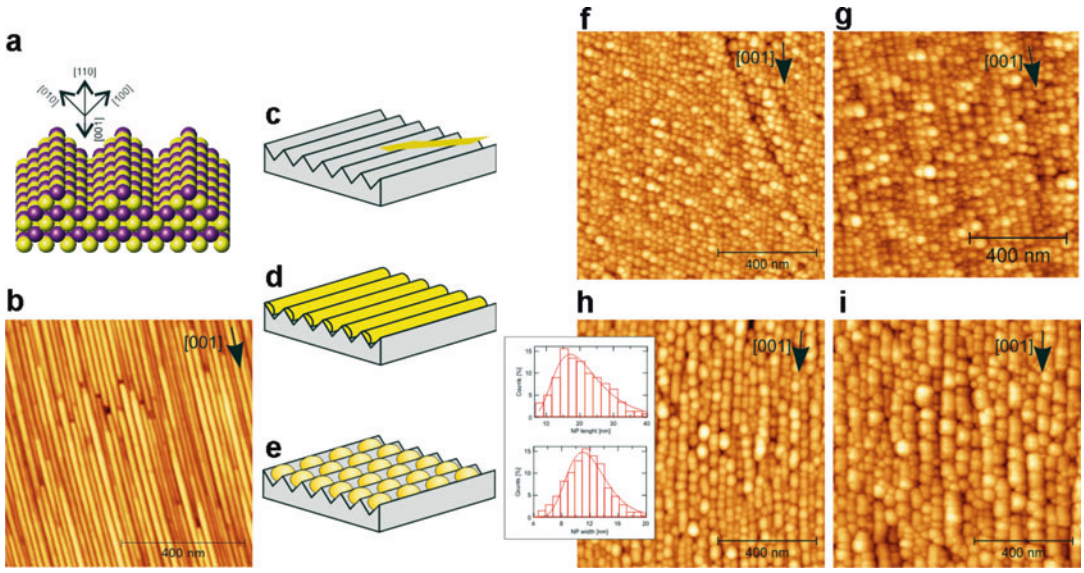
Ion-beam sputtering at non-grazing incidence angles $\theta < \theta_c$ has been reported to produce periodic patterns on a broad class of amorphous



Plasmonics in Self-Organized Media, Fig. 2 (**a**) AFM topography of a typical ripple pattern obtained by IBS on a soda lime glass substrate at $\theta = 35°$. (**b**) AFM topography acquired after deposition of 16 nm of Au at grazing angle (80°) on the rippled templates showing 1D chains of connected nanoclusters localized on the illuminated facets. (**c**) Optical extinction spectra of 1D arrays of disconnected Au nanoclusters (*red traces*) after deposition of 8 nm at 70° and connected NWs (*black traces*) after deposition of 16 nm at 80°. (**d**) Graph of the spectral position of LSP for Au and Ag NWs as a function of the out-of-plane aspect ratio *w/h*

substrates. In view of low-cost plasmonic applications, the possibility to employ as templates transparent dielectric substrates like soda lime glass is particularly attractive. IBS irradiation of glass substrates at $\theta = 35°$ leads to the formation of regular faceted ripples with tunable periodicity in the range 100÷350 nm, peak-to-valley height in the range 5÷60 nm, and length of the ridges in the range 1÷3 μm [18].

As an example, the topography of Fig. 2a reveals the typical rippled morphology of a glass template nanostructured by IBS after $\mathcal{F} = 6 \cdot 10^{18}$ ions/cm^2, with a periodicity around 150 nm and a height of about 20 nm. The typical slope of the tilted ripple facets is in the range 10°–20°. If Au evaporation is performed at a perpendicular azimuthal angle with respect to the ripple ridges, and at a polar glancing angle around 80°, the spatial distribution of the metal atom flux becomes modulated by the local facet angle which determines an alternating sequence of shadowed regions with no Au flux and illuminated regions with high flux. Preferential homogeneous nucleation of the clusters is thus forced near the top of the illuminated ridges, where the Au flux is the highest (Fig. 2b) [19]. For the lowest metal equivalent thickness (defined as the thickness of the homogeneous film that the same quantity of deposited material forms on a flat surface), nucleation of 1D chains of disconnected nanoclusters takes place on the illuminated facet, in registry with the underlying



Plasmonics in Self-Organized Media, Fig. 3 (**a**) Schematic atomic arrangement of nano-faceted rock salt (110) surfaces. (**b**) Atomic force microscopy image of a nano-faceted LiF(110) surface. (**c**–**e**) Elementary steps involved in fabricating metallic nanoparticles on a nano-faceted LiF(110) surface: grazing angle deposition, formation of nanowires, and temperature-induced dewetting. (**f**–**i**) AFM images recorded for Au/LiF(110) systems obtained depositing an equivalent thickness of Au t_{Au} on nanopatterns with periodicity Λ: (**f**.) $t_{Au} = 4.5$ nm, $\Lambda = 30$ nm; (**g**) $t_{Au} = 4.8$ nm, $\Lambda = 40$ nm; (**h**) $t_{Au} = 5.4$ nm, $\Lambda = 45$ nm; and (**i**) $t_{Au} = 7.2$ nm, $\Lambda = 55$ nm. The deposition was followed by annealing at 400 °C. Inset of panel (**h**): size distribution of the NP semiaxis in the direction parallel and perpendicular to the LiF ridges

glass pattern, while for higher Au doses, the 1D chains coalesce from a percolated nanowire as shown in Fig. 2b.

The morphological anisotropy of the Au clusters is reflected in their dichroic response as shown by the optical transmission spectra of Fig. 2c. The red traces refer to 1D chains of disconnected Au nanoclusters grown at a reduced Au equivalent thickness of 8 nm. For light polarized perpendicularly to the ripple ridges, the TM spectrum (red dashed line) is characterized by a localized extinction minimum around 570 nm due to the excitation of the localized surface plasmon resonances along the short axis of the clusters. For longitudinal polarization (TE spectrum, red continuous line), a broader extinction minimum is observed which is redshifted around 670 nm. The anisotropic excitation of the broadband LSP resonances confirms that for low metal doses, the growth of disconnected and anisotropic clusters with a broad size distribution is favored parallel to the ripple ridges. By further increasing the metal coverage to 16 nm, laterally ordered arrays of connected polycrystalline NWs are formed, due to agglomeration, coalescence, and coarsening of the primitive 1D chains of clusters, as shown in the AFM image of Fig. 2b. The change in the optical response is evidenced in Fig. 2c: for longitudinal polarization (TE, black continuous line), the extinction spectrum resembles those of a continuous connected Au film, while for TM polarization, a deep extinction minimum around 580 nm can be attributed to the excitation of LSP dipolar resonances polarized transversely to the NW axis.

In this nanopatterning approach, for a given template pattern, the spectral position of the LSP resonance can be simply tailored by changing the evaporation conditions, i.e., metal dose and incidence angle (which affect the nanocluster out-of-plane aspect ratio w/h) or alternatively the metal itself since the LSP resonance depends on the free-electron density n_e through the bulk plasmon frequency:

$$\omega_p = \sqrt{\frac{n_e e^2}{m^* \epsilon_0}}.$$

As an example in Fig. 2d, the characteristic spectral range of TM-polarized LSP resonances of Au and Ag nanoclusters is plotted as a function of the out-of-plane aspect ratio w/h of the NWs. A change of this aspect ratio from 1 to 4, obtained by varying the thickness h of the metallic wire, allows to spectrally shift the plasmon band by about 100 nm both for Au and Ag NWs. A stronger spectral shift of about 200 nm is instead observed when comparing Au and Ag NWs with the same out-of-plane aspect ratio, due to the modified dielectric constant of the material. The tunability of the LSP resonance and the strong field localization effects of the nanoparticle arrays supported on the templates can also be exploited in order to enhance weak nonlinear optical effects such as in second harmonic generation (SHG). As an example, when the tilted arrays of Au nanowires such as those of Fig. 2b are illuminated by a pump beam resonant with the LSP, a strong amplification of the SHG signal is observed in comparison to a reference flat Au film [20].

2D Plasmonic Arrays

The plasmonic nanoparticle arrays prepared by glancing angle deposition on self-organized rippled templates show a good degree of morphological control in the direction perpendicular to the ripple ridges, since nucleation takes place in registry with the undulated glass facets, but do not allow to control size selection and lateral ordering of the nanoclusters in the direction parallel to the elongated facets. The self-organized nanopatterning approaches based on ion-beam sputtering so far have succeeded in preparing 2D arrays of nanodots only at surfaces of some selected compound semiconductor [9] but not on the low-cost transparent dielectric substrates of interest in photonic applications. This can be obviated in several ways, thus serving the double purpose of selecting the lateral size of the nanoparticles and their periodicity along both symmetry directions. The simplest strategy is the adoption of self-assembled templates endowed with a two-dimensional periodicity, while a further scheme exploits the thermal dewetting of regular NW arrays to effectively split them along their translationally invariant axis. As the dewetting occurs in a constrained manner, the NWs tend to split in a regular fashion, creating a quasi-periodic arrangement of clusters. In the following, we will discuss the fabrication of 2D arrays by constrained dewetting of nanowires and employing self-assembled monolayers of polymeric microspheres as a deposition template [21].

Ionic Crystal Faceting

As long as the constrained dewetting is concerned, we will address the plasmonic media fabricated exploiting the self-organized faceting of ionic crystals. Ionic crystals like NaCl, LiF, and MgO exhibit the characteristic rock salt crystallographic structure, where (100) surfaces are the most energetically favorable. Starting from a suitably oriented ionic crystal surface, typically with (110) or (111) orientation, either thermodynamic processes (annealing) or out-of-equilibrium procedures (molecular beam homoepitaxy) lead to a surface rearrangement that leaves the lowest-energy crystal face exposed, as schematically shown in Fig. 3a [22, 23]. In the first stages of the process, surface instabilities, in the form of (100)-oriented facets, are generated in random positions on the surface. Going on, these facets typically grow larger in size evolving toward well-defined nanostructures in the form of triangular waves or three-sided pyramids for (110)-cut surfaces and for (111) surfaces, respectively. The nanostructures typically exhibit a remarkably narrow size distribution, thanks to the strong size selection occurring during their development.

The mean nanostructure periodicity Λ in these systems can be tuned in a certain window (typically 20 nm to 70 nm) as a function of the process temperature T, due to the temperature dependence of parameters involved in the faceting process, such as the adatom mobility, providing a simple method to tune the typical nanostructures' size. Characteristic periodicities down to $\Lambda \approx 20$ nm are thus achievable over areas as large as the available crystals' size, typically larger than the cm^2.

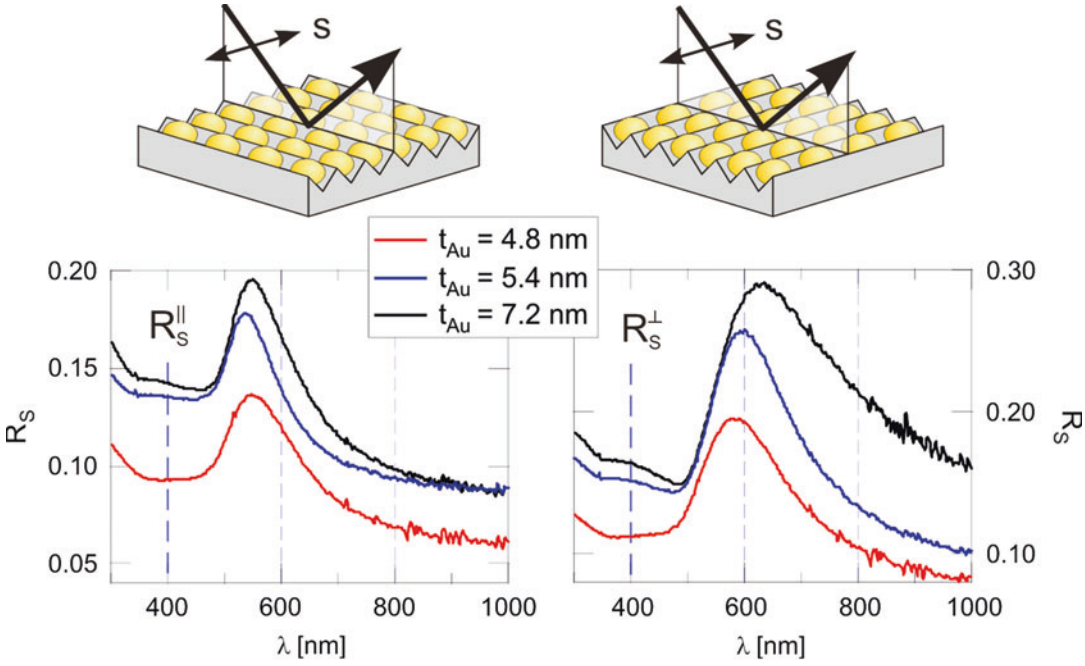
We notice that this periodicity nicely complements, at the lowest size limit, the ones achievable by IBS, so that a broad periodicity range becomes accessible by a suitable combination of both techniques.

An example of coherently oriented triangular wave nanostructures obtained on a LiF (110) surface is shown in the atomic force microscopy image of Fig. 3b. The insulating crystals do not undergo amorphization during the faceting process, thus maintaining their pristine bandgap, a factor particularly appealing for plasmonic application in the UV and DUV, where the formation of defective states in the insulator can negatively affect the LSPR detection. We will discuss in more depth two examples of metal/insulator plasmonic structures based on LiF self-organized substrates, in which the plasmonic counterpart is respectively represented by Au and Al. The general scheme for the realization of the plasmonic systems is represented in Fig. 3c–e. Starting from the bare nanopatterned LiF (110) surface (c), a metal film is deposited at grazing incidence to form NWs (d) that dewet to form regular NP arrays upon annealing (e). We will discuss two cases. The first, featuring Au as the plasmonic metal, will be representative of the plasmonic response tuning achieved by acting upon macroscopic variables. The second, featuring Al, is an example of the way in which the smallest-NP-size limit in self-organized materials can be exploited for pushing the high-energy limit of plasmonics.

In Fig. 3f–i, we report a few atomic force microscopy images representative of Au NPs formed onto a nanopatterned LiF (110) surface. The images (f) to (i) were obtained depositing an equivalent thickness t_{Au} of Au on nanopatterns with periodicity of Λ, respectively: (f) $t_{Au} = 4.5$ nm, $\Lambda = 30$ nm; (g) $t_{Au} = 4.8$ nm, $\Lambda = 40$ nm; (h) $t_{Au} = 5.4$ nm, $\Lambda = 45$ nm; and (i) $t_{Au} = 7.2$ nm, $\Lambda = 55$ nm. The deposition was followed by annealing at 400 °C. The images clearly show the formation of regular arrays of Au agglomerates, aligned along the LiF nanoridges. While the array pitch in the direction transverse to the LiF ridges is dictated by the template periodicity, the regular NP periodicity along the ridges stems from the constrained dewetting of the Au deposited within each individual nanogroove. The morphological bounds given by the LiF ridges restrict the mass transport, resulting in structures that are much more regular with respect to those obtained by unconstrained dewetting. Increasing the Au equivalent thickness from 4.5 to 7.2 nm allows indeed to change the in-plane aspect ratio of the NPs from circular to ellipsoidal. This demonstrates the tunability of a nanoscale property by means of a macroscopic parameter, one of the major advantages of bottom-up methods [24]. The tunable substrate can be then fruitfully employed as support for near-field-coupled optically active materials whose optical response can be tuned via the influence of the LSPR [25].

In the inset of Fig. 3h, we report the size distribution of the NP semiaxis extracted from the corresponding image. The mean semiaxis of the particles is 17 ± 7 nm in the direction along the ripples and 11 ± 3 nm in the transverse direction, well within the quasi-static approximation limit for Au. It must be noted that all the images clearly show a limited degree of disorder, in the NP size, shape, and array arrangement. The mean NP-array pitch is around 40 nm, implying that each NP finds itself within the EM near-field interaction range of the neighboring particle.

The plasmonic response of the arrays of Fig. 3g–i is reported in Fig. 4. There, reflectivity data measured with s-polarized radiation with the incidence plane oriented either parallel (right panel, TE geometry) or perpendicular (left panel, TM geometry) to the LiF ridge orientation are reported, as schematically represented in the top panels. The LSPR gives rise to the prominent peak in the spectra. The arrays of Fig. 3g, having the in-plane aspect ratio closer to 1, exhibit isotropic optical response (red line), meaning the LSPR wavelength does not change between the TE and TM excitation geometries. A macroscopic birefringence of the samples is induced in correspondence of an increased in-plane aspect ratio of the NPs (blue and black lines, corresponding to Fig. 3h and 3i). There, the TE configuration excites the longitudinal plasmon in the particles that redshifts due to the NP elongation.

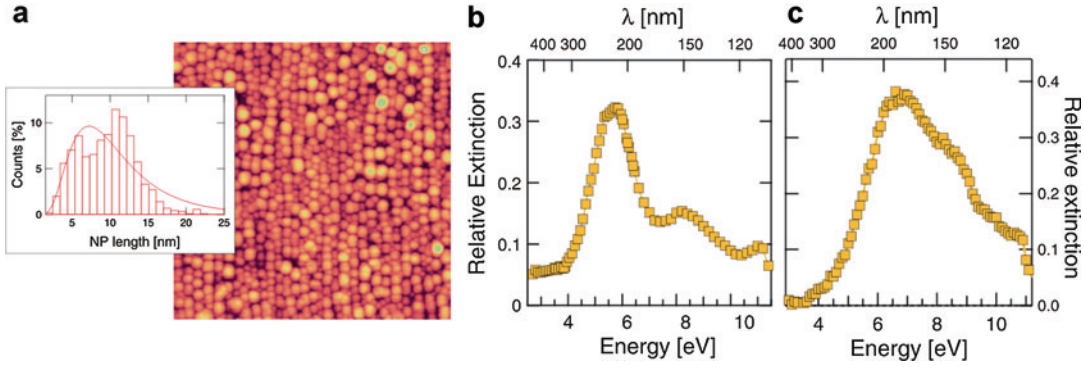


Plasmonics in Self-Organized Media, Fig. 4 *Top left* (*right*), optical geometry for the reflectivity measurements with the light polarization oriented across (along) the LiF nanoridges. *Bottom*, s-polarized reflectivity spectra for the Au/LiF NP arrays reported in the AFM images of Fig. 3 (**g–i**) (*black*, *blue*, and *red line*, respectively). On the *left* (*right*) are reported spectra obtained with light polarization across (parallel) the LiF ridges

The close proximity of the NPs also implies that the dipolar plasmon mode, the only one clearly observable in this case, is effectively delocalized over neighboring NPs, due to the near-field coupling. Thus, despite the large size dispersion and the remarkable positional disorder of the NPs, the plasmon peaks stay narrow and well defined, and the LSPR becomes largely insensitive to the *single particle* size/shape dispersion. Incidentally, the extremely high achievable areal density of NPs (of the order of 10^{11} NP/cm^2) also makes these substrates ideal for SERS, due to the large number of plasmonic *hot spots*. At variance with the NW case, SERS enhancement is achieved for both TE and TM excitation geometries, being particularly intense in TE configuration, where hot spots are created in between the NPs lying within the same LiF groove.

The capability of bottom-up techniques to fabricate NPs with remarkably small size finds a particularly interesting application in the plasmonics of aluminum. Historically, indeed, the most widely employed plasmonic material is the noble metal Au, due to its low dielectric losses in its LSPR spectral window and, above all, its extremely high resistance to corrosion and to plasma-etching procedures. Ag outperforms all metals in terms of plasmonic performances (field enhancement, LSPR intensity), but its reactivity limits its applicability. For both metals, the achievable LSPR spectral range is however upward limited in energy by the presence of strong interband absorption at 2.5 eV (Au) and just below 4 eV (Ag). Other metals, like the free-electron metal aluminum, do not suffer from such drawbacks, allowing to stretch the achievable LSPR range to the deep ultraviolet (DUV) spectral range.

For Al nanospheres in vacuum, the theoretical LSPR in the quasi-static approximation is found at a wavelength $\lambda \approx 140$ nm. The NP-size constraints for the quasi-static approximation are therefore more stringent than for Au or Ag, implying that LSPR from aluminum particles can quickly redshift and broaden even for relatively small NP size. For Al, the dipolar limit, for which



Plasmonics in Self-Organized Media, Fig. 5 (**a**) AFM image of an Al NP array obtained depositing 2.5 nm equivalent thickness of Al on a nano-faceted LiF (110) surface. (**b**) Extinction spectra of arrays of core/shell Al NPs. (**c**) Extinction spectra of arrays of purely metallic Al NPs

the LSPR has the highest energy and spectral purity, requires a particle diameter below 10 nm. Bottom-up methods can here fruitfully compare with lithography, whose lower size limit for Al is around 40 nm of diameter.

A further issue, as far as Al is concerned, is the quick formation of a few-nm-thick oxide layer on the surface of the NPs upon exposure to the atmosphere. From the plasmonic point of view, the formation of the oxide shell has the effect of locally increasing the environmental dielectric function, hence, redshifting the LSPR. This effect can be fruitfully exploited to achieve Al LSPR in the visible range, yet it is a drawback for DUV plasmonics, and as such might be circumvented only by fabrication and storage under controlled environments, such as UHV.

At its lowest limit, the self-organized faceting of LiF produces nanostructures with characteristic periodicity down to 20 nm. On this template, Al NPs with typical diameter within the dipolar-limit constraints can be thus fabricated. In Fig. 5a, we report an AFM image of an array of Al NPs fabricated on a nanopatterned LiF (110) surface, in a vacuum vessel with a base pressure $p < 1 \times 10^{-8}$ mbar. The Al NPs were realized depositing a 2.5 nm equivalent thickness film on a LiF (110) template with $\Lambda \approx 25$ nm, followed by annealing at 670 K. The exposure of the chip to the atmosphere leads to the formation of a 2 nm thick oxide around their metallic core. In full analogy with Au, the AFM image shows that Al forms ordered arrangements of clusters aligned along the LiF nanoridges. In the case under scrutiny, the mean size of the metallic Al core of the NPs is around 13 ± 5 nm, with an oxide shell thickness of 2 nm, well within the QSA size limit.

In Fig. 5b, polarized light absorption measurements of the Al NP ensemble are reported. Due to the high photon energies involved and in order to avoid spurious absorption by water vapor, the transmission measurements were performed making use of a synchrotron radiation light source and under high-vacuum conditions. The transmission spectra were measured at normal incidence and room temperature, with the electric field polarization oriented transverse to the direction of the LiF ridges (TM configuration). The absorption spectra show a marked extinction peak at energies as high as 5.8 eV. This peak represents the dipolar plasmonic mode of the array, and its high energy clearly demonstrates the successful achievement of a plasmonic response in the DUV spectral range for Al NPs [26].

A further blueshift of the LSPR was achieved by fabricating the NP arrays within the same high-vacuum environment under which their optical characterization is performed, as shown in Fig. 5c. At variance with the previous case, the arrays were not exposed to the atmosphere between the fabrication and the measurements, thus preventing the formation of an oxide shell. Under these conditions, the dipolar LSPR is pushed to 6.8 eV of photon energy ($\approx$ 180 nm),

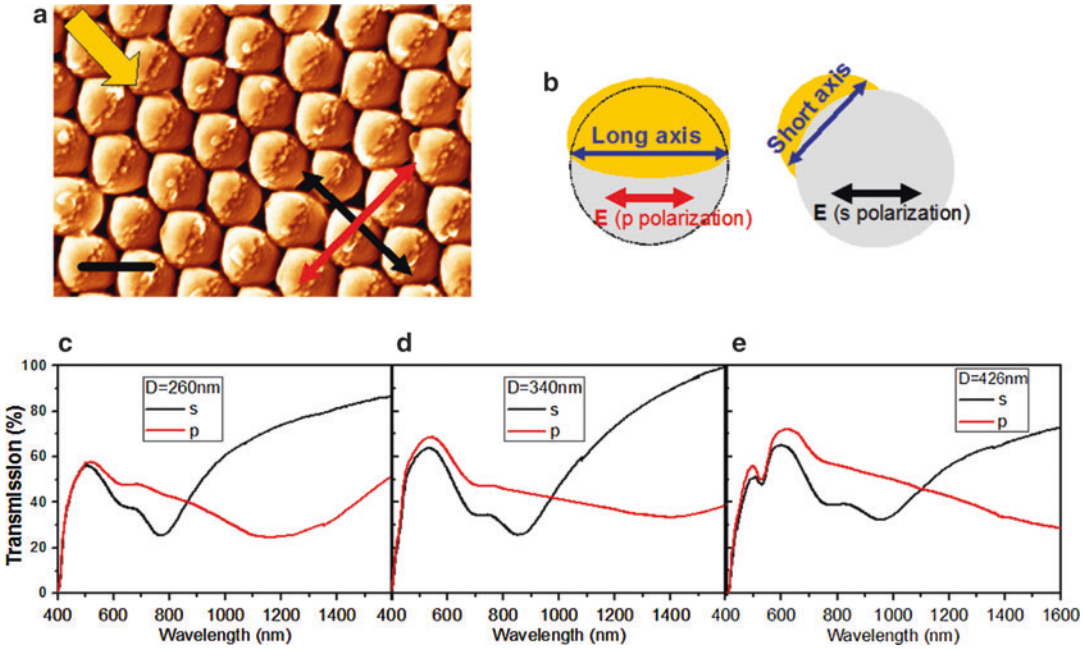
so far the highest value reported for optical excitation of metallic nanoparticles [27].

Interestingly, the dipolar peak of Al NP arrays stays remarkably sharp, despite the much stronger influence of the NP size on the LSPR energy. This is again a positive consequence of the above-described strong near-field coupling of the NPs, in absence of which significantly broader peaks are to be expected. Unlike the previous case of Au, the dipolar LSPR peak is accompanied by sizable sub-peaks on its high-energy side that originate, as shown by calculations, from higher-order plasmon resonances. Comparing Au and Al, no significant differences in mean NP size or in the degree of disorder of the arrays can be noticed. The high-order resonances are much more pronounced in the Al case simply due to the different dielectric characteristics of the two materials. In Au, the QSA plasmon resonance occurs around 550 nm in the arrays, and the NPs fully satisfy the condition $d \gg \lambda_P$ This leads to negligible high-order plasmonic resonances that are further damped by excitation of interband transitions. In Al, much stronger multipolar modes are expected due to the more stringent QSA limit, and those experience a very weak damping due to the absence of dissipation channels in Al at those energies.

Nanosphere Lithography

A simple alternative for the formation of ordered templates which cover large areas of a glass substrate is to employ self-assembled monolayers of polymeric microspheres [21]. In most approaches, the nanosphere monolayers have been employed as a sacrificial stencil mask for depositing ordered arrays of metal clusters on the underlying substrate by evaporating metal at normal incidence through the gaps of the close-packed hexagonal layer [28]. Here instead we describe an approach in which the nanosphere monolayer is directly employed as the supporting template for the array of plasmonic nanoparticles formed by glancing angle evaporation. In Fig. 6a, a scanning electron microscopy (SEM) micrograph shows a close-packed monolayer of 260 nm polystyrene microspheres self-assembled from a colloidal solution on a glass substrate. Au evaporation at glancing angles above 70° forces nucleation of the metal atoms only on the illuminated portions of the spheres which are shaped as half-moon crescents (brighter regions), while the shadowed portions of the sphere remain uncovered by gold (darker regions). In this way, by changing the diameter of the nanospheres, it is possible to tune both the periodicity of the array and the lateral size of the individual Au nanocrescents. As highlighted in the sketch of Fig. 6b, the nanocrescents are strongly anisotropic with a characteristic short axis, parallel to the evaporation direction, and a long axis running perpendicular to it.

The morphological anisotropy of the nanocrescents is reflected in their optical extinction spectra as highlighted in Fig. 6c. Transmission of polarized light through the Au nanocrescents acquired for light polarization parallel (s) and perpendicular (p) to their short axis (sketch in Fig. 6b) evidences a strongly dichroic behavior. For s polarization (black traces), a sharp extinction minimum is observed around 800 nm due to the excitation of the dipolar LSP resonance along the short crescent axis. On the other hand, for p polarization (red traces), the spectrum is characterized by an extinction minimum which is redshifted toward the IR spectral range, due to the dipolar LSP excitation parallel to the long axis of the structures. Furthermore, an additional feature, ascribed to a higher-energy quadrupolar mode, is observed for both light polarizations around 650 nm. The dipolar LSP resonances of the nanocrescents can be tailored in a simple way by changing the geometrical parameters of the template (nanosphere diameter) or the aspect ratio of the nanocrescents (evaporated metal thickness and deposition angle). In Fig. 6c–e, we compare the optical extinction spectra of the Au nanocrescent arrays when increasing the diameter of the polystyrene spheres from 260 to 340 and 426 nm, at a fixed grazing incidence angle $\theta = 70°$ and for the same Au thickness 10 nm. The general trend, analogously to the data observed for grazing incidence evaporation on the rippled glass template, is that by increasing the width to thickness aspect ratio of the nanocrescents, the dipolar plasmon resonance is



Plasmonics in Self-Organized Media, Fig. 6 (**a**) SEM topograph of a monolayer of 260 nm polystyrene microspheres decorated with Au nanocrescents. Metal deposition of 10 nm is performed at $\theta = 70°$ with the Au beam parallel to the *yellow arrow*. The black scale bar corresponds to 300 nm. (**b**) Sketch of the morphology of the crescents, characterized by a short axis (parallel to the evaporation direction) and by a long axis (perpendicular). (**c**–**e**) The optical transmission spectra are measured for light polarization parallel (p, *red trace*) and perpendicular (s, *black trace*) to the long axis of the crescents for increasing the sphere diameter from 260 to 340 and 426 nm.

redshifted. By increasing the sphere diameter, the short-axis plasmon mode observed in the s-polarized spectra thus shifts from 780 to 880 and 950 nm, respectively; the broad resonance corresponding to the long-axis dipolar plasmon mode observed in the p-polarized spectra instead shifts deep into the NIR spectral range, starting from 1150 to 1400 nm and finally extending beyond 1600 nm. We can thus conclude that the combination of low-cost self-ordered transparent templates with grazing incidence deposition of noble metals is a viable alternative for the controlled production of two-dimensional metal nanocluster arrays with tunable plasmonic response in the VIS-NIR spectral range [29]. The strong field localization effects induced in the presence of the dipolar LSP resonances supported by the nanocrescents are also at the basis of the nonlinear enhancement of the SHG signal which has been recently observed also for such kind of nanostructures [30].

Cross-References

- Bottom-Up Nanofabrication
- Local Surface Plasmon Resonance (LSPR)
- Nano-Optics, Nanoplasmonics
- Nano-patterning
- Nanostructured Materials
- Plasmonics
- Self-Organized Nanostructures
- Surface Plasmon Nanophotonics

References

1. Bohren, C.F., Huffman, D.R.: Absorption and Scattering of Light by Small Particles. Wiley, New York (1998)
2. Willets, K.A., Van Duyne, R.P.: Localized surface plasmon resonance spectroscopy and sensing. Annu. Rev. Phys. Chem. **58**, 267–297 (2007)
3. Malinsky, M.D., Kelly, K.L., Schatz, G.C., Van Duyne, R.P.: Chain length dependence and sensing capabilities of the localized surface plasmon resonance

of silver nanoparticles chemically modified with alkanethiol self-assembled monolayers. J. Am. Chem. Soc. **123**, 1471–1482 (2001)
4. Vigderman, L., Khanal, B.P., Zubarev, E.R.: Functional gold nanorods: synthesis, self-assembly, and sensing applications. Adv. Mater. **24**, 4811–4841 (2012)
5. Halas, N.J., Lal, S., Chang, W.-S., Link, S., Nordlander, P.: Plasmons in strongly coupled metallic nanostructures. Chem. Rev. **111**, 3913–3961 (2011)
6. Atwater, H.A., Polman, A.: Plasmonics for improved photovoltaic devices. Nat. Mater. **9**, 205 (2010)
7. Slaughter, L.S., Willingham, B.A., Chang, W.S., Chester, M.H., Ogden, N.: Toward plasmonic polymers. Nano Lett. **12**, 3967–3972 (2012)
8. Valbusa, U., Boragno, C., Buatier de Mongeot, F.: Nanostructuring surfaces by ion sputtering. J. Phys. Condens. Matter **14**, 8153 (2002)
9. Facsko, S., Dekorsy, T., Koerdt, C., Trappe, C., Kurz, H., Vogt, A., Hartnagel, H.L.: Formation of ordered nanoscale semiconductor dots by ion sputtering. Science **285**, 1551 (1999)
10. Bradley, R.M., Harper, J.M.E.: Theory of ripple topography induced by ion bombardment. J. Vac. Sci. Technol. A **6**, 2390 (1988)
11. Chan, W.L., Chason, E.: Making waves: kinetic processes controlling surface evolution during low energy ion sputtering. J. Appl. Phys. **101**, 121301 (2007)
12. Villain, J.: Continuum models of crystal growth from atomic beams with and without desorption. J. Phys. I **1**, 9 (1991)
13. Toma, A., Chiappe, D., Šetina Batič, B., Godec, M., Jenko, M., Buatier de Mongeot, F.: Erosive versus shadowing instabilities in the self-organized ion patterning of polycrystalline metal films. Phys. Rev. B. **78**, 153406 (2008)
14. Toma, A., Šetina Batič, B., Chiappe, D., Boragno, C., Valbusa, U., Godec, M., Jenko, M., Buatier de Mongeot, F.: Patterning polycrystalline thin films by defocused ion beam: the influence of initial morphology on the evolution of self-organized nanostructures. J. App. Phys. **104**, 104313 (2008)
15. Toma, A., Chiappe, D., Boragno, C., Buatier de Mongeot, F.: Self-organized ion-beam synthesis of nanowires with broadband plasmonic functionality. Phys. Rev. B. **81**, 165436 (2010)
16. Fazio, B., D'Andrea, C., Bonaccorso, F., Irrera, A., Calogero, G., Vasi, C., Gucciardi, P.G., Allegrini, M., Toma, A., Chiappe, D., Martella, C., Buatier de Mongeot, F.: Re-radiation enhancement in polarized surface-enhanced resonant raman scattering of randomly oriented, molecules on self-organized gold nanowires. ACS Nano **5**(7), 5945–5956 (2011)
17. Chiappe, D., Toma, A., Buatier de Mongeot, F.: Transparent plasmonic nanowire electrodes via self-organised ion beam nanopatterning. Small **9**, 913–919 (2013)
18. Buatier De Mongeot, F., Valbusa, U.: Applications of metal surfaces nanostructured by ion beam sputtering. J. Phys. Condens. Matter **21**, 224022 (2009)
19. Toma, A., Chiappe, D., Massabò, D., Boragno, C., Buatier de Mongeot, F.: Self-organized metal nanowire arrays with tunable optical anisotropy. Appl. Phys. Lett. **93**, 163104 (2008)
20. Belardini, A., Larciprete, M.C., Centini, M., Fazio, E., Sibilia, C., Chiappe, D., Martella, C., Toma, A., Giordano, M., Buatier de Mongeot, F.: Circular dichroism in the optical second-harmonic emission of curved gold metal nanowires. Phys. Rev. Lett. **107**, 257401 (2011)
21. Hulteen, J.C., Van Duyne, R.P.: Nanosphere lithography: a materials general fabrication process for periodic particle array surfaces. J. Vac. Sci. Technol. A **13**, 1553 (1995)
22. Sugawara, A., Mae, K.: Surface morphology of epitaxial LiF(110) and CaF_2(110) layers. J. Vac. Sci. Technol. B **23**(2), 443 (2005)
23. Kitahara, T., Sugawara, A., Sano, H., Mizutani, G.: Optical second-harmonic spectroscopy of au nanowires. J. Appl. Phys. **95**, 5002 (2004)
24. Anghinolfi, L., Moroni, R., Mattera, L., Canepa, M., Bisio, F.: Flexible tuning of shape and arrangement of au nanoparticles in 2-dimensional self-organized arrays: morphology and plasmonic response. J. Phys. Chem. C **115**, 14036 (2011)
25. Caminale, M., Anghinolfi, L., Magnano, E., Bondino, F., Canepa, M., Mattera, L., Bisio, F.: Tuning the magneto-optical response of iron oxide nanocrystals in Au- and Ag-based plasmonic media. ACS Appl. Mater. Interfaces **5**, 1995 (2013)
26. Maidecchi, G., Gonella, G., Zaccaria, R.P., Moroni, R., Anghinolfi, L., Giglia, A., Nannarone, S., Mattera, L., Dai, H.-L., Canepa, M., Bisio, F.: Deep ultraviolet plasmon resonance in aluminum nanoparticle arrays. ACS Nano **7**, 5834–5841 (2013)
27. Bisio, F., Zaccaria, R.P., Moroni, R., Maidecchi, G., Alabastri, A., Gonella, G., Giglia, A., Andolfi, L., Nannarone, S., Mattera, L., Canepa, M.: Pushing the high-energy limit of plasmonics. ACS Nano **8**, 9239–9247 (2013)
28. Haynes, C.L., Van Duyne, R.P.: Nanosphere lithography: a versatile nanofabrication tool for studies of size-dependent nanoparticle optics. J. Phys. Chem. B **105**, 5599 (2001)
29. Robbiano, V., Giordano, M.C., Martella, C., Di Stasio, F., Chiappe, D., Buatier de Mongeot, F., Comoretto, D.: Hybrid plasmonic photonic nanostructures: gold nanocrescents over opals. Adv. Opt. Mater. **1**(389) (2013)
30. Belardini, A., Benedetti, A., Centini, M., Leahu, G., Mura, F., Sennato, S., Sibilia, C., Robbiano, V., Giordano, M.C., Martella, C., Comoretto, D., Buatier de Mongeot, F.: Second harmonic generation circular dichroism from self-ordered hybrid plasmonic-photonic nanosurfaces. Adv. Opt. Mater. **2**, 207 (2014)

Plasticity of Nanostructured Solids

► Simulations of Bulk Nanostructured Solids

Plasticity Theory at Small Scales

Lorenzo Valdevit[1] and John W. Hutchinson[2]
[1]Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA
[2]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Synonyms

Higher-order plasticity theory; Strain gradient plasticity theory

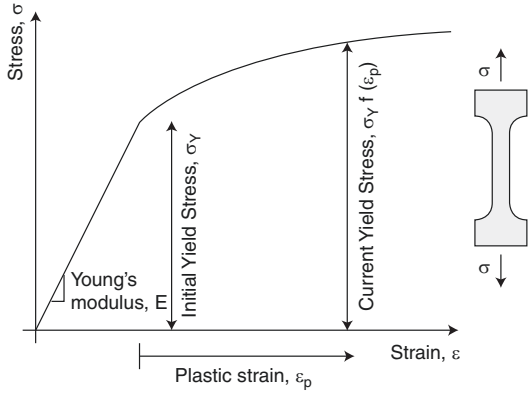
Definition

Plasticity theory is the mathematical formalism that describes the constitutive model of a material undergoing permanent deformation upon loading. For polycrystalline metals at low temperature and strain rate, the J_2 theory is the simplest adequate model. Classic plasticity theory does not include any explicit length scale, and as a result, the constitutive behavior is independent of the sample dimensions. As the characteristic length of a sample is reduced to the micro (and nano) scale, careful experimental observations clearly reveal the presence of a size effect that is not accounted for by the classical theory. Strain gradient plasticity is a formalism devised to extend plasticity theory to these smaller scales. For most metals, strain gradient plasticity is intended to apply to objects in the range from roughly 100 nm to 100 μm. Above 100 μm, the theory converges with the classical theory and below 100 nm surface and grain boundary effects not accounted for in the theory begin to dominate the behavior. By assuming that the plastic work (or in some theories, the yield strength) depends not only on strain but also on strain gradients (a hypothesis physically grounded in dislocation theory and, in particular, in the notion of geometrically necessary dislocations (GND) associated with incompatibility due to strain gradients), an intrinsic length scale is naturally introduced, allowing the theory to capture size effects. According to most theories, the intrinsic length scale is of the order of the distance between dislocation-clipping obstacles or cellular dislocation structures (typically, submicron to tens of microns). This continuum theory is appropriate for length scales that remain large relative to the distance between dislocations. As the sample length scale is dropped below this level, dislocations must be modeled individually, and discrete dislocations simulations (DSS) are the preferred approach. At even smaller scales, molecular dynamics (MD) becomes the applicable tool. This article presents a brief overview of one of the simplest continuum strain gradient plasticity theories that reduces to the classical J_2 theory when the scale of the deformation becomes large compared to the material length scale. This simple theory captures the essence of the experimental trends observed to date regarding size effects in submicron to micron scale plasticity.

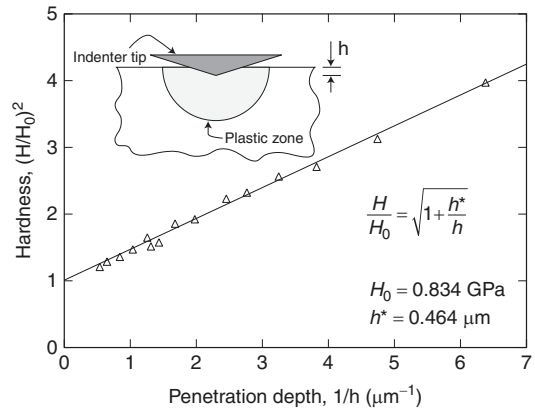
Plastic Flow of Materials and the Emergence of Size Effects at Small Scales

When ductile materials are loaded beyond their elastic limit, permanent deformation will result, involving long-range displacement of atoms (i.e., displacements much larger than the atomic spacing). In metals, this deformation occurs through the motion of dislocations. A typical stress–strain curve for uniaxial tensile loading is depicted in Fig. 1. In polycrystalline samples at low temperatures ($T < 0.5\ T_m$, with T_m the melting point of the material on the absolute scale) and relatively low strain rates ($\dot{\varepsilon} < 10/s$), a simple constitutive equation relating uniaxial stress to uniaxial strain is typically adequate. The flow stress can be defined as:

$$\sigma = \sigma_Y f\left(\varepsilon_p\right) \tag{1}$$



Plasticity Theory at Small Scales, Fig. 1 Typical stress–strain curve for an elastoplastic material in uniaxial tension. Young modulus, initial yield strength, and hardening behavior are illustrated



Plasticity Theory at Small Scales, Fig. 2 Dependence of hardness of cold worked polycrystalline copper on the indentation depth. The data fit a linear model almost perfectly (Modified from Ref. [2])

with σ_Y the initial yield strength, ε_p as the uniaxial plastic strain, and $f(\varepsilon_p)$ the hardening function. A power-law expression for f is often adequate. The plastic work per unit volume can be expressed as:

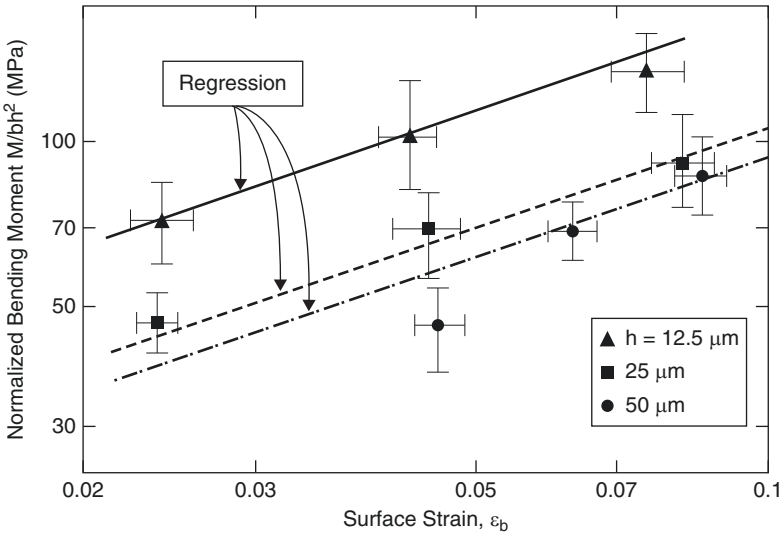
$$U_P(\varepsilon_p) = \int_0^{\varepsilon_p} \sigma d\varepsilon_p = \sigma_Y \int_0^{\varepsilon_p} f(\varepsilon_p) d\varepsilon_p \qquad (2)$$

Generalization of this behavior to general three-dimensional states of stress and strain is the objective of plasticity theories and is reviewed in the next section.

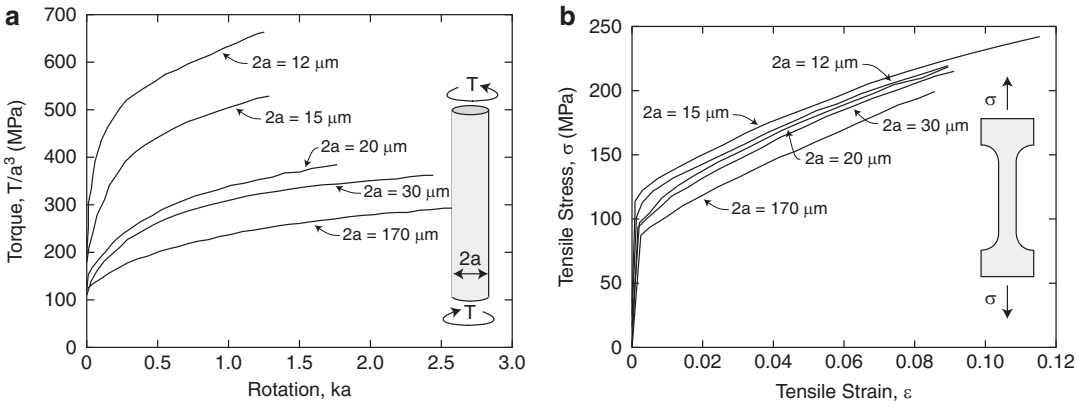
For samples with characteristic dimensions larger than 10–100 μm (depending on the material), yield strength and hardening are independent of sample dimension, rendering (Eq. 1) scale independent. Traditional plasticity theories embody this critical concept. Over the past two decades, though, as technological advances allowed measurements of mechanical properties in smaller and smaller systems, unexpected size effects became evident, typically suggesting that smaller samples are stronger. A large body of experimental literature supports this finding. Here, for the sake of conciseness, three key observations are reviewed. For a more exhaustive list, see [1].

(a) *Materials are harder when indented at smaller depths* [2]. A number of careful nanoindentation experiments were conducted and compiled by Nix and Gao [2], revealing a linear dependence of hardness on indentation depth: $(H/H_0)^2 = 1 + (h^*/h)$, where H is the hardness for a given depth of indentation, H_0 the hardness in the limit of infinite depth, and h^* a characteristic length of order 1 μ that depends on the shape of the indenter, the shear modulus of the material, and H_0 (Fig. 2). The effect becomes negligible as the indentation depth is increased to a depth well beyond h^*.

(b) *Thinner foils are stronger in bending than thicker ones* [3]. Experiments conducted by bending thin nickel foils around a mandrel revealed that thinner samples are stronger and strain harden more than thicker samples. Scaling arguments based on classic plasticity theory predict that the normalized moment, M/bh^2 (with b and h the foil width and thickness, respectively), should depend uniquely on the plastic strain at the foil surface, $\varepsilon_b = h/2R_0$, with R_0 the mandrel radius. For foils of different thickness but identical material, the results should superimpose in the absence of an intrinsic length scale, a conclusion clearly negated by experimental evidence for foil thicknesses between 12.5 and 50 μm (Fig. 3). Importantly, an opposite (albeit less significant) trend is observed in tension, i.e., thinner samples are weaker. Although this conclusion has not been fully rationalized,



Plasticity Theory at Small Scales, Fig. 3 Size effect in bending on thin nickel foils. Substantial strengthening is observed for thinner films, consistently with predictions from SGP theories. Importantly, the same effect is not observed in tension (Reprinted from Ref. [3])



Plasticity Theory at Small Scales, Fig. 4 Experimentally measured size effect in torsion of thin copper wires. (**a**) A significant size effect is reported for pure torsion. (**b**) Nearly no size effect is observed in uniaxial tension, confirming that the strengthening entails the presence of strain gradients (Modified from Ref. [4])

it clearly demonstrates that strengthening at small length scales entails the presence of strain gradients.

(c) *Thinner wires are stronger in torsion than thicker ones* [4]. In this classic experiment by Fleck et al. spearheading intense activity in the development of strain gradient plasticity theories, thin copper wires in diameters ranging from 15 to 170 μ were tested in tension and torsion. Dimensional analysis based on conventional plasticity theory predicts a dependence of the normalized torque, T/a^3 (with a the wire radius), on the plastic shear strain at the surface of the wire, $\gamma_s = \kappa a$, with κ the angle of twist per unit length. Again, data from samples of different radius should superimpose, when plotted in these coordinates. Experimental results clearly contrast with this prediction (Fig. 4a). Once again, tensile tests do not show the same trend (Fig. 4b), confirming that the strengthening is associated with the presence of strain gradients.

A constitutive theory incorporating an explicit dependence of the stress on strain gradients

P

(in addition to strain) could in principle explain all the size effects described above. An elementary theory encompassing this principle is discussed below. As explained above, observations (b) and (c) clearly indicate that the strengthening effect disappears as the gradients are removed. It is important to mention that this conclusion changes as the sample size is reduced to sub-micrometric dimensions: in these regimes, substantial size-dependent strengthening is observed under uniform stresses in single crystals (see entry on "▶ Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures" in this volume).

Conventional Plasticity Theory (J_2 Deformation Theory)

The goal of plasticity theories is the generalization of the one-dimensional model to general, three-dimensional states of stress and strain. For the sake of simplicity, in this article the attention is limited to plastically isotropic materials (i.e., materials for which the flow stress and hardening behavior are independent of the orientation of the coordinate system). It is also assumed that the loading is monotonic and proportional ($\sigma_{ij} = \lambda\sigma^0_{ij}$, with σ^0_{ij} a constant stress state and λ a loading parameter; throughout this article, indicial notation is adopted: when two indices appear in one term of an equation, summation over those indices is implicitly assumed; see [5] for details) at least approximately. This condition justifies the use of deformation theory, a constitutive model relating plastic strain to stress independently of the loading path. For more general loading scenarios (including all cyclic loadings), a more elaborate theory relating increments of stress and plastic strain (flow theory) is required [5, 6].

The yield point in a uniaxial tensile test (Fig. 1) generalizes to a yield surface, defined in a six-dimensional stress space as the envelope of all stress states resulting in elastic deformation: $f(\sigma_{ij}) = 0$. If $f < 0$, the behavior is completely elastic, whereas if $f = 0$ plastic deformation occurs.

Under the common assumption that a hydrostatic state of stress ($\sigma_{ij} = -p\delta_{ij}$) can be neglected in characterizing plastic behavior, f is taken as a function of the deviatoric stress tensor ($s_{ij} = -\sigma_{ij} - (1/3)\sigma_{kk}\delta_{ij}$). For isotropic materials, the most general dependence can be expressed as $f(J_2, J_3) = 0$, where $J_2 = (1/2)s_{ij}s_{ij}$ and $J_3 = (1/3)s_{ij}s_{jk}s_{ki}$ are the second and third invariants of the deviatoric stress tensor. In the J_2 theory of plasticity, the dependence of f on J_3 is ignored, and thus the yield surface can be simply represented as $f(J_2) = 0$. From Eq. 1, we have $J_2 = \sigma_Y^2/3$, with σ_Y the yield strength in uniaxial tension. This is the well-known Von Mises criterion [7]. Similarly, the hardening function f can be chosen as an isotropic function of the plastic strain tensor, ε^p_{ij} (isotropic hardening). In the simplest theories, it is also common to assume that hardening depends only on the second invariant of ε^p_{ij}, with $\varepsilon^p_{kk} = 0$ (plastic incompressibility). Thus, the yielding condition for an isotropic material in a general state of stress can be still written in scalar form as:

$$\sigma_e = \sigma_Y f(\varepsilon_p) \tag{3}$$

with:

$$\sigma_e = \sqrt{\frac{3}{2}s_{ij}s_{ij}} \text{ and } \varepsilon_p = \sqrt{\frac{2}{3}\varepsilon^p_{ij}\varepsilon^p_{ij}} \tag{4}$$

The numerical factors 2/3 and 3/2 are introduced so that the equivalent stress, σ_e, and the plastic strain magnitude, ε_p, coincide with stress and plastic strain for uniaxial tensile loading.

The plastic work per unit volume can then be expressed as:

$$U_p(\varepsilon_p) = \int_0^{\varepsilon_p} \sigma_e d\varepsilon_p = \sigma_Y \int_0^{\varepsilon_p} f(\varepsilon_p) d\varepsilon_p \tag{5}$$

which is identical to Eq. 2, as long as stress and plastic strain are defined as in Eq. 4.

As the Cauchy stress tensor σ_{ij} and the strain tensor ε_{ij} are work-conjugate, the plastic strain tensor can be expressed as:

$$\varepsilon_{ij}^{p} = \frac{\partial U_p}{\partial \sigma_{ij}} = \frac{\partial U_p}{\partial J_2}\frac{\partial J_2}{\partial \sigma_{ij}} = \frac{\partial U_p}{\partial J_2} s_{ij} = \frac{3}{2}\frac{\varepsilon_p}{\sigma_e} s_{ij} \quad (6)$$

where the relation $J_2 = \sigma_e^2/3$ was adopted. The implication is that according to J_2 deformation theory, the plastic strain tensor is codirectional with the stress deviator.

Strain Gradient Plasticity Theories

As mentioned earlier, the theory presented in the previous section does not contain any intrinsic length scale, and is thus incapable of predicting the size effects depicted in Figs. 2, 3, and 4. The most natural way of introducing a length scale in the constitutive model is to assume that the plastic work is a function of an enhanced strain measure that incorporates both the plastic strain, ε_{ij}^{p}, and the plastic strain gradients, $\varepsilon_{ij,k}^{p}$. The deformation version intended for application to proportional or near-proportional loading was introduced by Fleck and Hutchinson [4] and represents the application to plasticity theory of higher-order elasticity theories first introduced by Toupin [8] and Mindlin [9]. The simplest implementation defines the plastic work per unit volume as:

$$U_p(E_p) = \int_0^{E_p} \sigma_e d\varepsilon_p \quad (7)$$

where:

$$E_p = \left((\varepsilon_p)^{\mu} + \left(\ell \varepsilon_p^{*}\right)^{\mu} \right)^{1/\mu} \quad (8)$$

The measure of the plastic strain gradient, ε_p^{*}, can be defined in many different ways [10, 11]. Here the representation $\varepsilon_p^{*} = \sqrt{\varepsilon_{p,i}\,\varepsilon_{p,i}}$, based on the choice made by Zibb and Aifantis [12], is adopted for mathematical convenience. Based on dimensional considerations, the dependence of E_p on ε_p^{*} requires the introduction of an intrinsic length scale ℓ. This definition generates a family of theories, as a function of the parameter μ. The choice $\mu = 1$ appears to give the best agreement with experimental observations (e.g., it predicts the linear dependence of hardness on indentation depth, Fig. 2 [1]), but most formulations have used $\mu = 2$ for mathematical reasons. This section surveys the basic predictions and general trends of such theories. A physical interpretation of the intrinsic length scale ℓ is provided in the following section.

The general trends predicted by this theory are easily elucidated. Imagine a perfectly plastic object ($f(\varepsilon_p) = 1$) of characteristic dimension h, subject to an average plastic strain, ε_p. Dimensionally, the plastic strain gradient is given by $\varepsilon_p^{*} = c\,\varepsilon_p/h$, with $c \approx 1$. According to Eq. 7 and Eq. 8, the plastic work per unit volume can be expressed as:

$$U_p = \sigma_Y(1 + c\ell/h)\varepsilon_p \quad (9)$$

Comparing Eq. 9 to the classic J_2 theory prediction, $U_p = \sigma_Y \varepsilon_p$, one concludes that this strain gradient plasticity theory predicts a yield strength elevation of $1 + c\ell/h$. As the length scale of the object, h, approaches the characteristic materials length scale, ℓ, the yield strength elevation becomes significant.

In general, the potential energy functional can be written as:

$$\Phi = \int_V \left(\frac{1}{2} C_{ijkl} \varepsilon_{ij}^{e} \varepsilon_{kl}^{e} + U_p(E_p) \right) dV - \int_{S_T} \left(T_i u_i + t \varepsilon_p \right) dS \quad (10)$$

with $\varepsilon_{ij}^{e} = \varepsilon_{ij} - \varepsilon_{ij}^{p} = \varepsilon_{ij} - \frac{3}{2}\frac{\varepsilon_p}{\sigma_e} S_{ij}$. Here, C_{ijkl} are the elastic moduli (isotropic here) with $\sigma_{ij} = C_{ijkl}\varepsilon_{kl}^{e}$ and where T_i and $t = \tau_i n_i$ are prescribed on S_T. This class of theories, referred to as higher-order theories, inevitably introduces a new stress measure, τ_i, that is work-conjugate to the plastic strain gradient, i.e., $\tau_i = \partial U_p/\partial \varepsilon_{p,i}$. Specification of an additional boundary condition is required, involving t, ε_p or a combination of both, thus allowing solution to a wider variety of problems, intractable with conventional plasticity theory. For example, passivation layers can be

applied on the surface of a solid, forcing the plastic strain to vanish at the surface. (For more details, see [11].)

The solution to Eq. 10 represents a minimum with respect to u_i and ε_p, assuming a monotonic uniaxial stress–strain curve (Eq. 1). The material length scale, ℓ, must be obtained by fitting one set of experimental data that brings in strain gradient effects.

Application to the torsion problem presented in Fig. 4 illustrates the predictive capability of the theory. Full details of the analysis are presented in Ref. [1]. Elastic strains constitute a very small contribution to the tensile and torsion data in Fig. 4, justifying the adoption of a power-law tensile stress–strain curve of the form $\sigma_e = \sigma_0 \varepsilon_p^N$. The strain measure ε_p must be replaced by E_p for all loading cases involving strain gradients, as explained above. The ensuing plastic work can then be expressed as:

$$U_p\left(E_p = \frac{1}{N+1}\right)\sigma_0 {E_p}^{N+1} \tag{11}$$

In torsion, with κ as the twist/length, the shear strain varies with distance, r, from the center of the wire according to $\gamma = \kappa r$, such that $\varepsilon_p = \kappa r/\sqrt{3}$ (with $\kappa > 0$), $\varepsilon_p^* = \kappa/\sqrt{3}$, and $E_p = \frac{\kappa}{\sqrt{3}}\left[r^\mu + \ell^\mu\right]^{1/\mu}$. The torque, T, under monotonically increasing κ is:

$$\frac{T}{\sigma_0 a^3} = \left(\frac{\kappa a}{\sqrt{3}}\right)^N \frac{2\pi}{\sqrt{3}} \int_0^{\mathrm{I}} \left[\zeta^\mu + (\ell/a)^\mu\right]^{\frac{N+1}{\mu}} \zeta d\zeta \tag{12}$$

where a is the radius of the wire.

The tensile stress–strain curve in Fig. 5a is plotted with $\sigma_0 = 275MPa$ and $N = 0.2$, giving an approximate fit to the experimental curves in Fig. 4a. With these same values and with the choice $\ell = 7\mu m$, the trends of the experimental torsion data in Fig. 4a are approximately captured by the theoretical curves from Eq. 12 in Fig. 5b. $\mu = 2$ was chosen for all calculations.

Although in this brief survey the attention was limited to the deformation theory version of strain gradient plasticity (both for mathematical convenience and transparency of scaling effects), more general flow theory versions (relating plastic strain increments to the increments of stress in a manner which has intrinsic strain-path dependence) have been developed [13–16]. The constitutive behavior based on flow theory is applicable to any loading history (including cyclic loading, in principle, although none of the existing versions have been vetted for such loadings) and almost always requires a finite element formulation to analyze problems of interest.

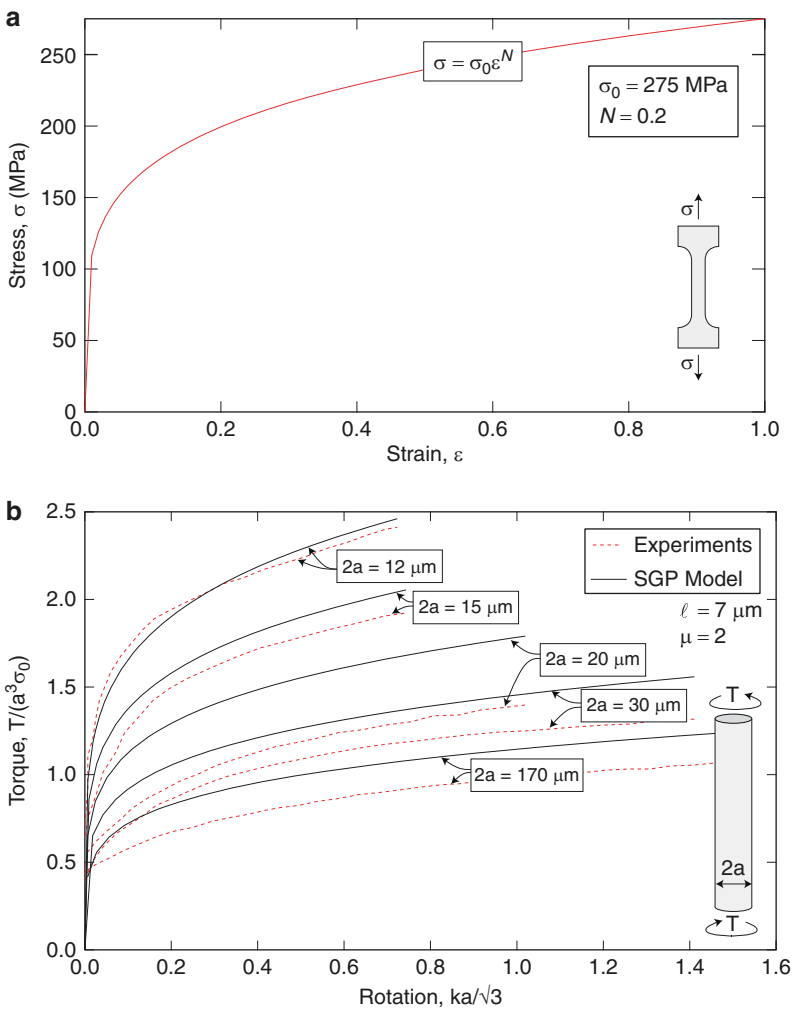
In the next section, the intrinsic length scale ℓ will be qualitatively related to basic notions from dislocation theory.

Physical Interpretation of the Intrinsic Material Length Scale

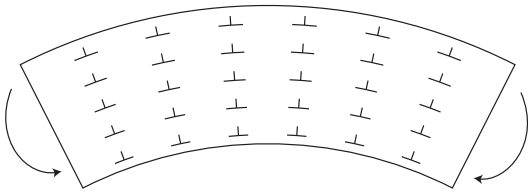
Conventional strain hardening and yield stress elevation are attributed to the interaction of mobile dislocations with relative immobile *statistically stored dislocations* (SSD). The density of SSD, ρ_{SSD}, increases proportionally to the accumulated plastic strain, according to $\varepsilon_p \approx \rho_{\mathrm{SSD}} bd$, with b the burger's vector and d the average distance traveled by a dislocation (defined by the spacing of obstacles). As emphasized by Ashby [17], accumulation of plastic strain gradients requires the introduction of additional dislocations to ensure geometric compatibility (Fig. 6): unlike SSDs, these *geometrically necessary dislocations* (GND) must be arranged in space to accommodate the incompatibility associated with the plastic strain gradient and scale as $\varepsilon_p^* \approx \rho_{\mathrm{GND}} b$. The motion of both families of dislocations contributes to the plastic work. If one assumes that SSD and GND do not strongly interact, a simple additive contribution is appropriate. According to this model, the plastic work can be expressed as:

$$\begin{aligned} U_p &\approx \sigma_Y(\rho_{\mathrm{SSD}} bd + \rho_{\mathrm{GND}} bd) \\ &\approx \sigma_Y\left(\varepsilon_p + d\varepsilon_p^*\right) \end{aligned} \tag{13}$$

Comparison with Eq. 9 immediately reveals $\ell \approx d$. As the average distance traveled by a dislocation is



Plasticity Theory at Small Scales, Fig. 5 (**a**) Power-law stress–strain curve in pure tension (extrapolated at large strains), giving an approximate fit to the data in Fig. 4a. (**b**) Strain gradient plasticity prediction (Eq. 12, with $\ell = 7\mu$m, $\mu = 2$) for the torsion data of Fig. 4b, based on the stress–strain law in (**a**)



Plasticity Theory at Small Scales, Fig. 6 Cartoon illustrating the formation of geometrically necessary dislocations (GND) in pure bending

comparable to the average spacing between obstacles, this simple analysis suggests that usually $\ell \approx 1 - 20\mu$m, consistent with experimental results and modeling prediction (Fig. 5).

Modeling Plasticity of Very Small Samples

The strain gradient plasticity theory described in this entry is still a continuum theory (i.e., it does not model individual dislocations). As the size of the sample is reduced to the point that only a handful of dislocations exist across the characteristic length scale, this fundamental assumption breaks down. In this case, numerical approaches which explicitly model dislocation interactions are required. Molecular dynamics (MD) is a natural modeling strategy, but its application is limited to relatively small samples and extremely high strain rates. Discrete dislocations simulations (DDS) present an approach that bridges the two

theories [18–20]. These models generally utilize a FE continuum framework to solve for stresses, strains, and displacements, and treat dislocations as singularities, which both affect and are affected by the global strain and stress fields. Dislocations motion is governed by the Peach–Koehler equation and standard dislocation–dislocation interaction laws. Dislocation sources like Frank–Read and single-arm sources are introduced at statistically random locations. These models (both in two-dimensional and three-dimensional) are computationally more efficient than full-scale MD simulations, but have not yet fully succeeded in duplicating key aspects of the experimentally measured size effects in plasticity [21].

Open Issues, Future Developments, and Final Considerations

One reason why strain gradient plasticity theories have not been widely embraced is the ambiguity surrounding the definition of the intrinsic length scale. Although the scaling arguments presented in this article ascribe its physical basis and define its order of magnitude, a quantitative physical model allowing expression of ℓ as a function of measurable materials parameters is still lacking. Attempts to define such a model based on dislocation theory are still ongoing. At the same time, efforts are still underway in developing a fully consistent flow theory version of strain gradient plasticity and in implementing such theories in finite element codes.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Finite Element Methods for Computational Nano-Optics
- ▶ Nanoindentation
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures

References

1. Evans, A.G., Hutchinson, J.: A critical assessment of theories of strain gradient plasticity. Acta Mater. **57**, 1675–1688 (2009)
2. Nix, W.D., Gao, H.: Indentation size effects in crystalline materials: a law for strain gradient plasticity. J. Mech. Phys. Solids **46**(3), 411–425 (1998)
3. Stolken, J., Evans, A.: A microbend test method for measuring the plasticity length scale. Acta Mater. **46**(14), 5109–5115 (1998)
4. Fleck, N., et al.: Strain gradient plasticity – theory and experiments. Acta. Metall. Mater. **42**(2), 475–487 (1994)
5. Lubliner, J.: Plasticity Theory. Dover, New York (2008)
6. Hill, R.: The Mathematical Theory of Plasticity. Oxford University Press, Oxford (1998)
7. Mises, R.V.: Mechanik der festen Körper im plastisch deformablen Zustand. Göttin. Nachr. Math. Phys. **1**, 582–592 (1913)
8. Toupin, R.A.: Elastic materials with couple stresses. Arch. Ration. Mech. Anal. **11**(5), 385–414 (1963)
9. Mindlin, R.D.: Micro-structure in linear elasticity. Arch. Ration. Mech. Anal. **16**(1), 51–78 (1964)
10. Fleck, N., Hutchinson, J.: Strain gradient plasticity. Adv. appl. Mech. **33**, 295–361 (1997)
11. Niordson, C.F., Hutchinson, J.W.: Basic strain gradient plasticity theories with application to constrained film deformation. J. Mech. Mater. Struct. **6**(1–4), 395–416 (2011)
12. Zibb, H., Aifantis, E.: On the gradient-dependent theory of plasticity and shear banding. Acta Mech. **92** (1–4), 209–225 (1992)
13. Fleck, N.A., Willis, J.R.: A mathematical basis for strain-gradient plasticity theory. Part II: tensorial plastic multiplier. J. Mech. Phys. Solids **57**(7), 1045–1057 (2009)
14. Fleck, N.A., Willis, J.R.: A mathematical basis for strain-gradient plasticity theory-Part I: scalar plastic multiplier. J. Mech. Phys. Solids **57**(1), 161–177 (2009)
15. Gudmundson, P.: A unified treatment of strain gradient plasticity. J. Mech. Phys. Solids **52**(6), 1379–1406 (2004)
16. Gurtin, M., Anand, L.: A theory of strain-gradient plasticity for isotropic, plastically irrotational materials. Part I: small deformations. J. Mech. Phys. Solids **53**(7), 1624–1649 (2005)
17. Ashby, M.F.: Deformation of plastically non-homogeneous materials. Philos. Mag. **21**(170), 399 (1970)
18. Deshpande, V.S., Needleman, A., Van der Giessen, E.: Plasticity size effects in tension and compression of single crystals. J. Mech. Phys. Solids **53**(12), 2661–2691 (2005)
19. Tang, H., Schwarz, K.W., Espinosa, H.D.: Dislocation-source shutdown and the plastic behavior

of single-crystal micropillars. Phys. Rev. Lett. **100**(18), 185503 (2008)
20. Van der Giessen, E., Needleman, A.: Discrete dislocation plasticity – a simple planar model. Model. Simul. Mater. Sci. Eng. **3**(5), 689–735 (1995)
21. Uchic, M.D., Shade, P.A., Dimiduk, D.M.: Plasticity of micrometer-scale single crystals in compression. Annu. Rev. Mater. Res. **39**(1), 1–23 (2009)

Plating

Seajin Oh and Marc Madou
Department of Mechanical and Aerospace Engineering and Biomedical Engineering, University of California at Irvine, Irvine, CA, USA

Synonyms

Metal plating

Definition

Plating is a film coating process on a substrate through either chemical or electrochemical reactions in the solution. In chemical plating, several chemical reactions occur in the plating solution to deposit metal on a substrate. In electrochemical plating, known as electroplating, metal ionic species in the plating solution react with electrons supplied by an external electrical power supply to deposit metal onto a substrate.

Overview

Plating is used to change the surface properties of a substrate, for example, hardness, wear resistance, friction, adhesion, corrosion, and electrical or thermal conductivity, and also to decorate objects. Films are deposited by either chemical or electrochemical reactions which are distinguished by whether an outside source of electric current is involved in the reaction. Chemical plating is a non-galvanic plating method where crystallization occurs on the substrate, providing films with no upper limit in thickness. The electroless plating deposits a metal film on nonconductor such as synthetic plastics, which already makes this process important in the printed circuit industry. The list of synthetic plastics to be plated is growing because of the importance of the process for a variety of industrial applications. Electrochemical plating or electroplating is referred to as electrodeposition and is a fairly simple process. A negative charge is placed on the object that is electroplated, which is called cathode and a positive charge is placed on the anode. Both the cathode and the anode are immersed in an electrolyte solution containing the charged particles of dissolved metal that deposit metal on the substrate by passing current through an external circuit between the cathode and the anode. The surface of substrate must be cleaned thoroughly. Recessed areas plate less than other external surface due to the difference of proximity to plating source, often resulting in making a scratch more prominent, rather than being smoothened by the plating process. The thickness of the electroplated films is generally controlled by adjusting time and current in the electroplating process, enabling to control film thickness with nanometer scale accuracy. Three-dimensional nanostructures are produced by the electroforming process in the nanoscale templates such as a metal nanowire array grown through nanopores formed in a polymer matrix. The electroplating process is called anodizing when the object is anodic and an oxide film is formed on the surface.

Basic Methodology

Chemical Plating

Chemical plating is classified into (1) deposition by homogeneous reaction of a metal ion with a reducing agent in solution and (2) deposition by heterogeneous reaction on the substrate specified for the autocatalytic reaction [1].

1. Homogeneous reduction reaction takes place in a solution containing a metal ion to be

deposited and a reducing agent that reacts with the metal ion and crystallizes to grow a metal film on the substrate. One old art is silver mirrors that are deposited on glass or synthetic plastics. This process involves a solution of silver salt (e.g., $AgNO_3$) as one solution and a reducing agent (e.g., sucrose) dissolved in a second solution. The two solutions are mixed before using, or sprayed from a two-nozzle gun on the substrate to be silvered. This method is also used to deposit Cu, Ni, Au, and PbS.

2. Autocatalytic chemical reaction takes place only on catalytic surfaces immersed in the solution where a chemical reducing agent provides electrons for the reaction $M^{n+} + ne = M_0$ with no use of external electric power source. This inhomogeneous process requires that the deposited metal layer itself must be catalytic to continue plating on the growing layer and enables continuous deposition of a metal or alloy simply by immersion in the plating solution. In that regard, at least theoretically there is no upper limit to the film thickness to grow. Compared to electroplating, the deposition process has advantages of simplicity and essentially less porous films than electroplating. There are two critical steps before plating. In the first step a reducing agent (e.g., $SnCl_2$) is adsorbed on the plastic surface (sensitization step), which in the following step reduces $PdCl_2$ in a dilute solution to Pd deposited on the surface at discrete sites [nucleation step]. This method deposits Ni, Co, Pd, Pt, Cu, Au, Ag, and some alloys containing P or B in the metals. Among them, Ni is the most important film that is deposited from hypophosphite baths, formulated as

$$2H_2PO_2^- + 2H_2O + Ni^{2+} = Ni + H_2 + 4H^+ + 2HPO_3^{-2} \quad (1)$$

Electroplating

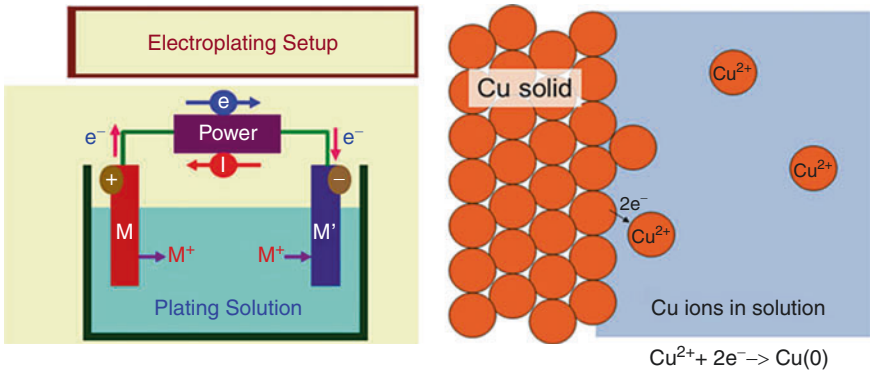
Electroplating involves passing an electric current between two electrodes immersed in an electrolyte. When current enters or leaves an electrode, a chemical change takes place at the interface between electrode and electrolyte. For example, a metal cation in the electrolyte receives electron at the interface to deposit as a metal layer. An electrolyte is an ionic solution – solution of acids, bases, salts, fused salts, or some gases – where ions are moved by the electrical potential gradient applied between two electrodes. Ions in the solution are usually hydrated or complexed. Metals in the solution are either in the cationic or in the anionic form. In the latter, the metal is complexed with a ligand or coordinating agent [e.g., $Zn(CN)_2{}^{-4}$]. The conductivity of electrolyte is far lower than metal electrode, for example, the conductivity of 1 % solution of NaCl is about 5×10^7 times lower than that of copper. The electrode where reduction takes place is the cathode and the electrode where oxidation takes place is the anode. In the electrolyte cell where electroplating takes place, the cathode is the negative electrode and the anode is the positive electrode (Fig. 1).

In electroplating, the quantity of electricity that passes through the electrolyte cell measures the amount of chemical change. Theoretically if the chemical change is solely the deposition of metal and no current leakage occurs, the mass of metal deposit can be exactly calculated as follows:

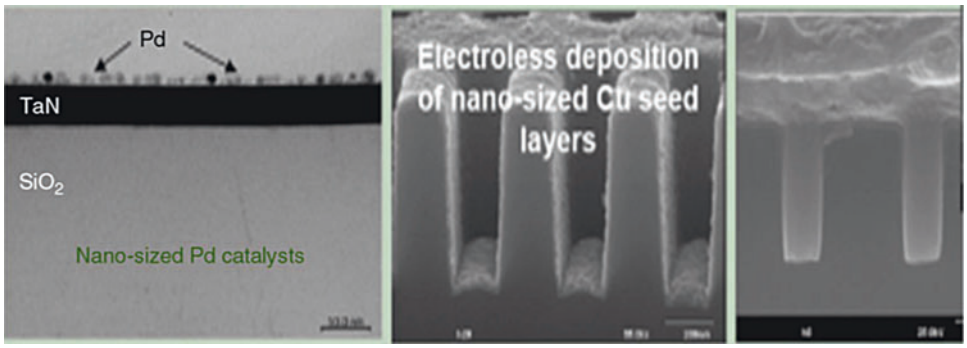
$$G = Iet = 96,500 \quad (2)$$

where G is grams of metal, I the current (A), e is the electrochemical equivalent weight (gram) which is atomic weight divided by the valance charge, and t is the time (second). In most cases, the electroplated metal weight is less than predicted by Eq. 2 because of additional reactions involved, such as the co-deposition of hydrogen with metal at the cathode and the evolution of oxygen at the anode which is dissolving in electrolyte. Current efficiency in electroplating varies widely and depends on such factors as electrolyte bath composition, temperature, agitation, pH, and current density.

Anodization is the electroplating process that takes place at anode where the surface metal is converted to a form of its oxide while hydrogen evolves at the cathode. The principal application is to Al for corrosion resistance, coloring, and electrical insulation. Other metals to be anodized



Plating, Fig. 1 Electroplating. Metal deposition involves reduction of a metal cation in solution by means of electrical current flow



Plating, Fig. 2 Copper electroless plating and electroplating. Trenches (200 nm width) in the dielectric layers are filled by copper electroplating Ref. [2]

P

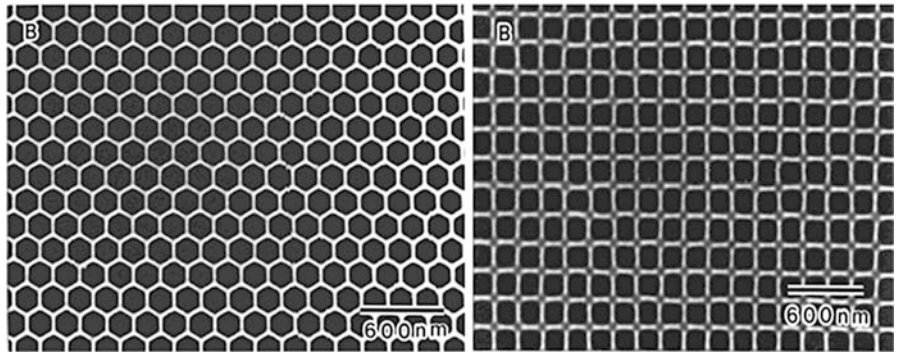
are Mg, Cu, Ag, Cd, Ti, and steel mostly for decorative effects.

Key Research Findings in Nanotechnology

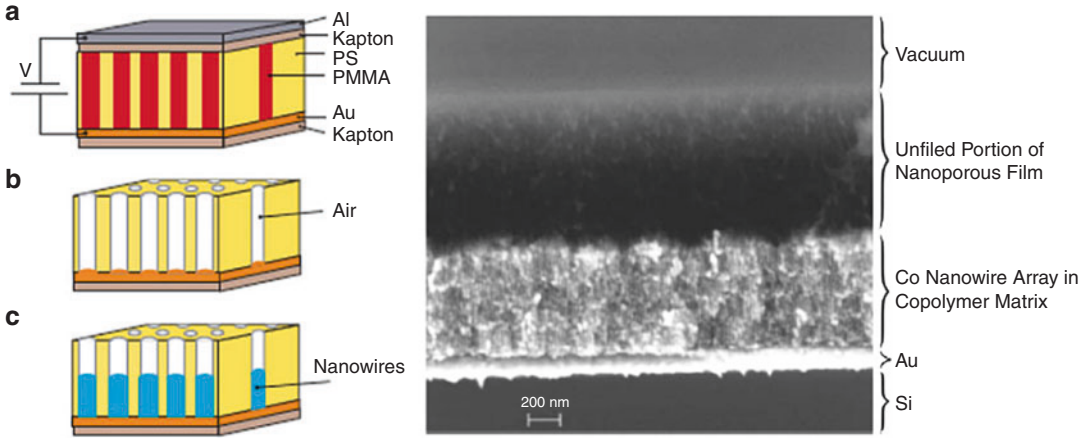
Copper metallization in damascene process. In the modern microprocessors, the wiring material changes from aluminum to copper to accommodate a large number of transistors interconnected in the restricted space. Copper electroplating takes place to fill the trenches and holes in the dielectric layers with copper in the damascene process for wiring (Fig. 2).

Porous anodic alumina [3]. Porous alumina layers are produced by the anodization of aluminum in an acidic electrolyte. The pores are circular or arranged in a close-packed array of square, triangle, and hexagonal opening. This process starts with an artificial laying-out of the initiating indentation sites in the arrays patterned by photolithography and from which corresponding shapes are produced as anodization proceeds in an acidic electrolyte. The pore size, pore spacing, and film thickness are all controllable (Fig. 3).

Electroforming through lithographic masks. Combining electroplating with novel lithography techniques becomes a powerful electroforming technique in the fabrication of microstructures or nanostructures. A representative example is the LIGA process where X-ray lithography creates high aspect ratio structures of photoresist mask, which allows for the fabrication of three-dimensional MEMS devices of a few hundred microns in height [4]. Using nanolithography



Plating, Fig. 3 The pore opening in the anodic alumina formed by indentation of triangular array (*left*) and square array (*right*) Ref. [3]



Plating, Fig. 4 Nanowire grown in diblock copolymer template [5]. (*Left*) A schematic representation of high-density nanowire fabrication in a polymer matrix. (**a**) An asymmetric diblock copolymer annealed above the glass transition temperature of the copolymer between two electrodes under an applied electric field, forming a hexagonal array of cylinders oriented normal to the film surface. (**b**) After removal of the minor component, a porous film is formed. (**c**) By electrodeposition, nanowires are grown in the porous template, forming an array of nanowires in a polymer matrix. (*Right*) SEM image of a fracture surface of an array of Co nanowires grown within an array of nanopores formed from block copolymers. The growth of the nanowires is terminated before the template is completely filled

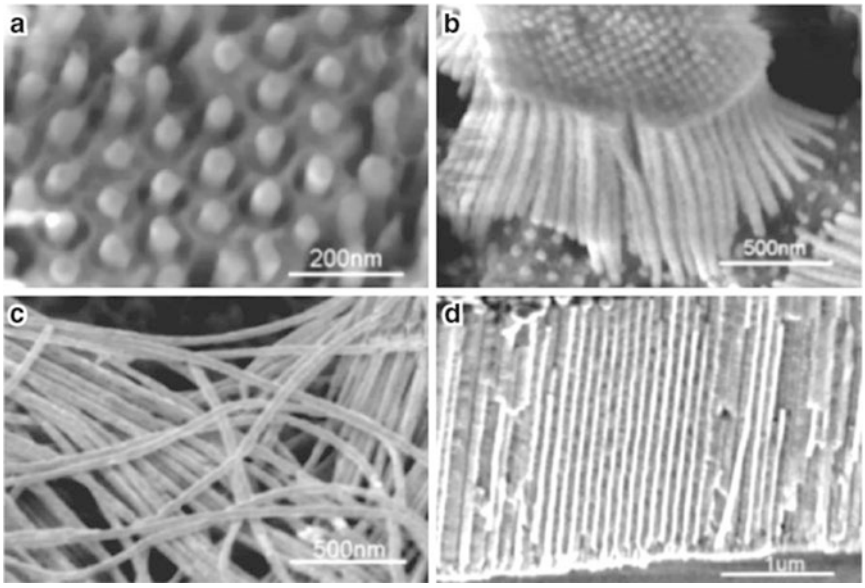
techniques produces templates including nanoscale features that can produce high-fidelity replicas.

Electrodeposited nanowires. Nanowire arrays are produced by electroplating in diblock copolymer template [5] or porous anodic alumina template [6] (Figs. 4, 5, and 6).

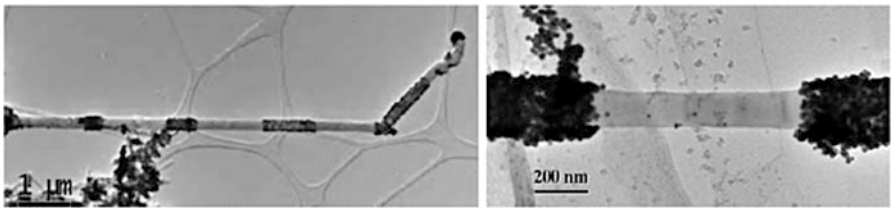
Selective plating on silicon nanowire. P-n junctions are grown in situ in semiconductor nanowires for various electronic or optical applications. Au is selectively coated onto the n-type segments of silicon nanowire p-n junction which allows to measure the position of p-n junction and segment length [7].

Future Directions for Research

Plating is less expensive than any other thin and thick film deposition methods. Because of the economical advantage continuous effort will be made to produce plated films with desirable



Plating, Fig. 5 Nanowires grown in porous anodic alumina template [6]. (**a**) Top view of Ni nanowire arrays (**b**) bundle of Ni nanowires in a 2 mm anodized aluminum oxide film after aluminum oxide was partially dissolved (**c**) part of the Bi nanowires in a 25 mm anodized aluminum oxide template after aluminum oxide was partially dissolved (**d**) cross section of the Bi nanowire array, in which the bright lines are Bi nanowires exposed by cracking the template



Plating, Fig. 6 TEM image showing nanoparticle morphology of gold plated on n + segments Ref. [7]

mechanical, physical, chemical, and optical properties that meet particular application requirements. Electroplating is simple to control the deposition rate by controlling the current. In that regard, electroplating is a superior technique for the production of nanoscaled structures made of crystalline pure metals, alloys as well as nanocomposites. Electroplating of metal in templates enables the fabrication of miniaturized structures from submillimeter to nanometer-scaled dimensions with extremely high precision replicas. Research will be continued to provide various three-dimensional nanostructures (e.g., nanowires and nanofilm) in conjunction with the progress of innovative nanotemplate fabrication techniques.

Cross-References

- ▶ DUV Photolithography and Materials
- ▶ Nanoindentation
- ▶ Nanotechnology
- ▶ SU-8 Photoresist

References

1. Lowenheim, F.A.: Deposition of inorganic films from solution. In: Vossen, J.L., Kern, W. (eds.) Thin Metal Process, pp. 209–257. Academic, New York (1978)
2. OM group Inc, power point presentation. www.omgi.com
3. Masuda, H., Asoh, H., Watanabe, M., Nishio, K., Nakao, M., Tamamura, T.: Square and triangular nanohole array architectures in anodic alumina. Adv. Mater. **13**, 189–192 (2001)
4. Romankiw, L.T.: A path: from electroplating through lithographic masks in electronics to LIGA in MEMS. Electrochim. Acta **42**(20–22), 2985–3005 (1997)
5. Thurn-Albrecht, T., Schotter, J., Kastle, G.A., Emley, N., Shibauchi, T., Krusin-Elbaum, L., Guarini, K., Black, C. T., Tuominen, M.T., Russell, T.P.: Ultrahigh-density nanowire arrays grown in self-assembled diblock copolymer templates. Science **290**, 2126–2129 (2000)
6. Yin, A.J., Li, J., Jian, W., Bennett, A.J., Xu, J.M.: Fabrication of highly ordered metallic nanowire arrays by electrodeposition. Appl. Phys. Lett. **79**(7), 1039–1041 (2001)
7. Eichfeld, C.M., Wood, C., Liu, B., Eichfeld, S.M., Redwing, J.M., Mohney, S.E.: Selective plating for junction delineation in silicon nanowires. Nano Lett. **7**(9), 2642–2644 (2007)

PmCP

► Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Polarization-Induced Effects in Heterostructures

Katja Tonisch
Institut für Mikro- und Nanotechnologien, Technische Universität Ilmenau Fachgebiet Nanotechnologie, Ilmenau, Germany

Synonyms

Pyroelectricity; Spontaneous polarization

Definition of Crystal Polarization

According to the number of rotational axes and reflection planes, crystals can be divided into 32 classes (point groups). Twenty of these classes exhibit a non-centrosymmetric structure leading to a direct piezoelectric behavior, i.e., an electric field is generated when the crystal is submitted to mechanical force and vice versa. Of these 20 piezoelectric classes, 10 are additionally polar, which implies that they possess a spontaneous polarization, having a dipole in their unit cell. These crystals are called pyroelectric. If this dipole can be reversed by an external electric field, the material is also called ferroelectric. Any dielectric material develops a dielectric polarization (electrostatics) when an electric field is applied, but only substances exhibiting a natural charge separation in the absence of external electric fields are polar. This material property is determined solely by its crystal structure.

- Piezoelectric crystal classes: 1, 2, m, 222, mm^2, 4, −4, 422, 4 mm, −42 m, 3, 32, 3 m, 6, −6, 622, 6 mm, −62 m, 23, −43 m
- Pyroelectric crystal classes: 1, 2, m, mm^2, 3, 3 m, 4, 4 mm, 6, 6 mm

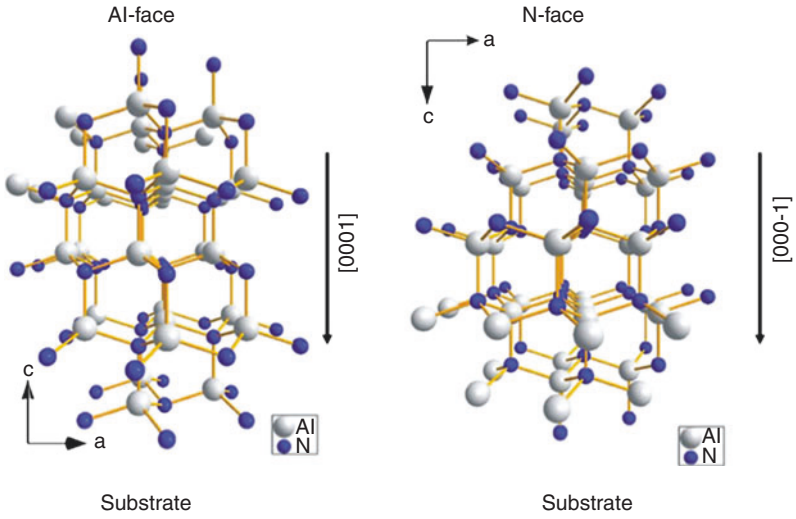
Examples for materials exhibiting both, a spontaneous and a piezoelectric polarization, are given in Table 1. In most cases, two or more polytypes exist, meaning that the same composition can occur in different crystal structures. Here, only the polar crystal structure (i.e., wurtzite or hexagonal) is given, the corresponding polytype is indicated in case of SiC. The polar properties of ternary and quaternary alloys such as AlGaN, AlInN, and AlInGaN can be interpolated from the involved binary composites. However, this interpolation is not strictly linear but exhibits a parabolic component.

Origin of the Spontaneous Polarization

Crystallographic polarity along certain crystal axes can occur, when two different sequences of the atomic layering in the two opposing directions can be observed. For binary A–B compounds with wurtzite structure, the sequence of the atomic layers of the constituents A and B is reversed along the (0001) and (000–1) directions.

Polarization-Induced Effects in Heterostructures, Table 1 Overview of III-V nitride, silicon carbide, and II-IV oxide materials with wurtzite structure exhibiting both spontaneous and piezoelectric polarization. Values are given in (C/m^2) for the main polar axis

Material	Crystal class	Spontaneous polarization P_{SP} (C/m^2)	Piezoelectric polarization P_{PE} (C/m^2)		
			e_{15}	e_{31}	e_{33}
AlN [1, 2]	$P6_3mc$	−0.081	−0.48	−0.60	1.46
GaN [1, 2]	$P6_3mc$	−0.029	−0.3	−0.49	0.73
InN [1, 2]	$P6_3mc$	−0.032		−0.57	0.97
2H-SiC [3]	$P6_3mc$	−0.007	0.08	−0.10	0.20
ZnO [4]	$P6_3mc$	−0.0503	−0.489	−0.664	1.243
BeO [4]	$P6_3mc$	−0.0363	−0.275	−0.436	0.804
MgO [4]	$P6_3mc$	−0.134	−0.428	−0.763	2.004
CdO [4]	$P6_3mc$	−0.107	−0.831	−0.790	1.584



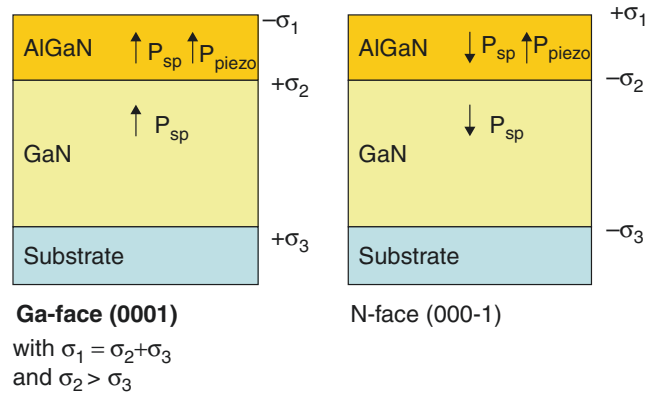
Polarization-Induced Effects in Heterostructures, Fig. 1 Crystal structure of AlN illustrating the polarity along the main crystal axis in wurtzite binary A-B compounds [5]

The corresponding <0001> and <000–1> faces are the A-face and B-face, respectively (Fig. 1). Please note that they also differ in their chemical behavior.

Each A-B (cation-anion) bond has a dipole character; its strength depends on bond length and charge distribution of the binding electron due to the different electron negativities of cation and anion. In case of wurtzite crystals, the tetrahedral binding structure consists of one vertical and three in-plane cation-anion bonds between an atom and its nearest neighbors. In case of an ideal crystal, the dipoles which correspond to these four bonds compensate each other. However, in an ideal wurtzite crystal, the distance to the second nearest neighbor is by 13 % shorter along the polar axis than the distance to the second nearest neighbors within the (basal) plane. This is not the case in cubic crystals. This in fact finally leads to an electric field (i.e., electric polarization) in the direction of the polar axis of the crystal [6]. Real crystals can still deviate from the ideal structure in bond lengths and angles, leading to an additional lattice distortion which in return changes the values of the spontaneous polarization.

Charge Carrier Accumulation at Heterostructure Interfaces

Typical applications of polar semiconductors require the epitaxial growth of heterostructures,



Polarization-Induced Effects in Heterostructures, Fig. 2 Scheme of a simple AlGaN/GaN heterostructure with its spontaneous and piezoelectric polarization; the GaN layer is assumed to be unstrained

i.e., a layer stacking with multiple layers of polar materials exhibiting a similar crystal structure is grown on a (hetero-) substrate. A multitude of possibilities exist to stack various layers of binary, ternary, or even quaternary compounds within one material class (e.g., combinations of AlN, InN, GaN, AlInN, InGaN, AlGaN, and AlGaInN layers) or even in combining different classes (ZnO/GaN, GaAs/GaN, AlN/SiC). Each single layer then possesses an individual internal electrical field which is the result of its polarization. Since pseudomorphically grown layers are strained in most cases, the total polarization of each layer is the sum of its spontaneous polarization due to the pure material property and of its piezoelectric-induced polarization due to strain. Such layer stackings can be achieved in both polarization directions (i.e., A- or B-face), which defines the direction of the internal electric field. In all cases, the polarization exhibits an abrupt change in its value or even a change in its sign (positive or negative) at the layer interfaces, resulting in a net sheet charge. This effect is commonly used intentionally in devices based on polar materials. To simplify matters, a detailed explanation will be given for one material combination only.

A commonly used structure (compare section Application) is based on thin $Al_xGa_{1-x}N$ layers grown on thick GaN buffer layers (Fig. 2). Without external electric fields, the total polarization of a GaN or AlGaN layer is the sum of the spontaneous polarization $\mathbf{P}_{SP}$ of the ideal crystal lattice, and of the strain-induced or piezoelectric polarization $\mathbf{P}_{PE}$. The piezoelectric polarization can be calculated from the deviation of the real crystal from its ideal structure. In case of epitaxial grown layers, a biaxial strain occurs due to the alignment of the lateral lattice constant of one layer to the layer beneath. Then, the lateral strain $\varepsilon_x, \varepsilon_y$ is calculated from the difference between the as-grown lattice constant a_{real} and a_0 (depending on the composition ratio x):

$$\varepsilon_x = \varepsilon_y = \frac{a_{real} - a_0(x)}{a_0(x)} \tag{1}$$

The corresponding vertical strain ε_z represents the deviation of the vertical lattice constant c and can be calculated using the elastic constants C_{11} and C_{33}:

$$\varepsilon_z = \frac{c_{real} - c_0(x)}{c_0(x)} = -2\frac{C_{13}}{C_{33}}\left(\frac{a_{real} - a_0(x)}{a_0(x)}\right) \tag{2}$$

The piezoelectric polarization is then given by:

$$P_{PE} = e_{33}(x) \cdot \varepsilon_z + e_{31}(x) \cdot (\varepsilon_x + \varepsilon_y) \tag{3}$$

$$P_{PE} = 2 \cdot \frac{a_{real} - a_0}{a_0} \cdot \left(e_{31} - e_{33} \cdot \frac{C_{13}}{C_{33}}\right) \tag{4}$$

The piezoelectric polarization is negative for tensile and positive for compressive strained AlGaN barriers over the whole range of

compositions. Since the spontaneous polarization for group III-nitrides is negative, it always points toward the substrate in case of metal-face heterostructures. As a consequence, the alignment of the piezoelectric and spontaneous polarization is parallel in the case of tensile strain, and antiparallel in the case of compressively strained top layers. The piezoelectric, as well as the spontaneous polarization, changes its sign, when N-face structures are considered instead. The polarization field abruptly changes at all heterostructure interfaces resulting in a net sheet charge density σ, which is defined by the polarization difference ΔP of the adjacent layers:

$$\begin{aligned}\sigma(P_{SP}+P_{PE}) &= \Delta P \\ &= P_{SP}(A1GaN) - P_{SP}(GaN) \\ &\quad + P_{PE}(A1GaN)\end{aligned} \quad (5)$$

If this sheet charge density is positive as it is the case of the above-considered heterostructure, it will be compensated by free electrons. This results in a two-dimensional accumulation of electrons at the interface, which is also called two-dimensional electron gas (2DEG). Its sheet charge density can be calculated according to [5]:

$$\begin{aligned}n_s(x) = &\frac{\sigma(x)}{e} - \frac{\varepsilon_0 E_F}{e^2}\left(\frac{\varepsilon_{AlGaN}(x)+\varepsilon_{GaN}}{d_{AlGaN}}\frac{\varepsilon_{GaN}}{d_{GaN}}\right) \\ &- \frac{\varepsilon_0\varepsilon_{AlGaN}(x)}{e^2\cdot d_{AlGaN}}\left(e\phi_{AlGaN}(x) + E_F(x) - \Delta E^C_{AlGaN}(x)\right)\end{aligned} \quad (6)$$

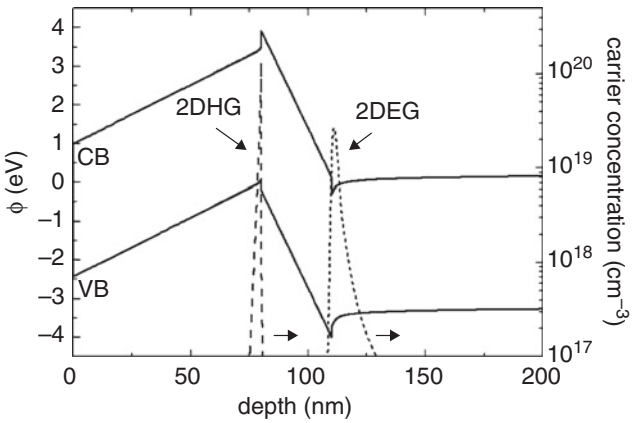
with the Fermi energy E_F of GaN, the relative dielectric constants ε_{GaN} and ε_{AlGaN}, the Schottky barrier height eϕ at the surface and the conduction band offset ΔE_c at the interface. Vice versa, a negative sheet charge density will cause an accumulation of holes at the interface. As the Schottky barrier height at the surface directly influences the sheet carrier density, the choice of the contact material as well as applied external voltages at the surface leads to an increase or depletion of the sheet carrier density. This phenomenon is the typically used effect in applications based on the field transistor configuration (see section "Application").

In case of heterostructures containing more than one interface, electron and hole accumulations can occur at the respective interface, resulting in a coexistence of 2D electron and hole gases within the same structure. In Fig. 3, the simulation results of a triple heterostructure GaN/AlGaN/GaN is displayed, showing the course of the conduction band and a slightly asymmetric charge distribution at the interfaces.

This coexistence recently has been proven by electro-reflectance measurements [7]. Shifting the net internal electric field by applying an external voltage leads to a corresponding decrease or increase of the electron and hole density (Fig. 4).

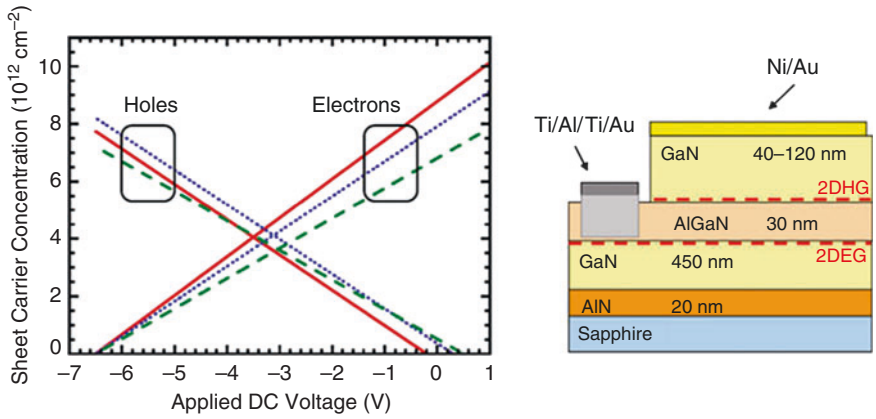
Application

The electron transport properties at heterostructure interfaces in polar semiconductors provide a new field of pyroelectronics and pyrosensors. Thus, the pyroelectric character of



Polarization-Induced Effects in Heterostructures, Fig. 3 Simulation results for a GaN/AlGaN/GaN heterostructure. Displayed are valence band (*VB*), conduction band (*CB*), and the resulting 2D hole gas (2DHG) and 2D electron gas (2DEG) at the interfaces

P



Polarization-Induced Effects in Heterostructures, Fig. 4 Electron and hole concentration at the GaN/AlGaN and AlGaN/GaN interface in GaN/AlGaN/GaN heterostructures depending on the vertically applied external voltage (*left*) [7]. Scheme of the analyzed structure (*right*)

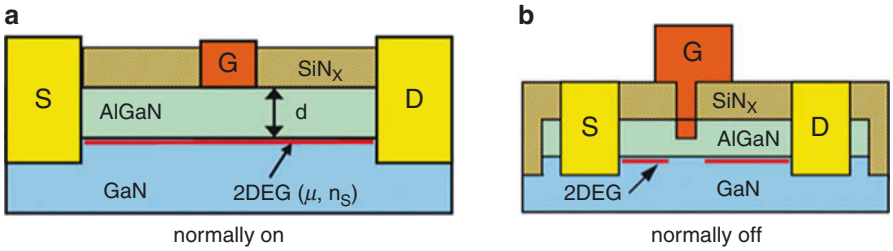
such semiconductors yields a novel degree of freedom in designing and tailoring devices for modern micro- and nanoelectronic as well as for micro- and nanomechanical applications. The interface charges, which are induced by the above-described spontaneous and piezoelectric polarization, respond very sensitively to changes of the overall electric field caused by mechanical strain, the adsorption of ions, wetting by polar liquids, or exposure to gases.

High Electron Mobility Transistors (HEMTs)

Generally, devices for high currents and high frequencies require a material with high charge carrier mobilities and high saturation velocities in the electron channel. The v-E characteristic defines the electron mobility μ_n, as the slope of the v-E characteristic at low-electric field, and the saturated velocity v_{sat} as the constant, field independent value of the electron velocity at high electric field. Compared to Ga(In)As as a classic semiconductor for rf applications, electron and hole low field mobilities in GaN are three times smaller, which results in lower maximum operation frequencies. GaN, however, has a larger saturation electron velocity. The fairly good electron mobility and the outstanding high saturation velocity in connection with the high thermal stability and the large band gap of GaN makes AlGaN/GaN-based transistors promising candidates for microwave power amplification (frequencies $>$ 1 GHz, power $>$ 1 W/mm). For this application, high electron mobility transistors (HEMTs) are the most attractive kind of transistors. Originally, HEMTs were developed based on AlGaInAs/(In)GaAs heterostructures on GaAs or InP substrates [8]. These devices are also known as MODFETs (Modulation-doped FETs), HFETs (Heterojunction FETs), TEGFETs (Two-dimensional electron gas FETs), or SDHT (Selectively doped heterostructure transistors). While in GaAs-based HEMTs, the 2DEG is formed at the heterostructure interface as a consequence of strong asymmetrically doping, in GaN-based HEMTs, the interface accumulation is due to polarization charges as described before (Figs. 2 and 3) without intentionally doping. Electron confined in those channels might have a substantially higher mobility due to their separation from their heavily doped source region, and their accumulation in a low-doped region, where impurity scattering has no dominant influence. The mobility in heterostructures increases with Al content [9]. Due to the increased polarization gradient (caused by higher spontaneous polarization as well as higher strain-induced piezoelectric polarization), the 2DEG sheet charge density increases. The mobility peaks for an Al-mole fraction, which is dependent on the barrier thickness and decreases for both lower and higher mole fractions.

Polarization-Induced Effects in Heterostructures, Table 2 Comparison of electron mobilities and velocities in AlGaAs/GaAs and AlGaN/GaN heterostructures and bulk material

	μ_{bulk} (cm^2/Vs)	μ_{bulk} (cm^2/Vs)	v_{peak} (cm/s)	v_{sat} (cm/s)	Ref.
AlGaAs/GaAs	10,000 (10 K)	10,00,000 (10 K)	5×10^7 (77 K)		[11]
AlGaN/GaN	2070 (115 K)	75,000 (77 K)	3×10^7 (300 K)	2.5×10^7 (300 K)	[12, 14]

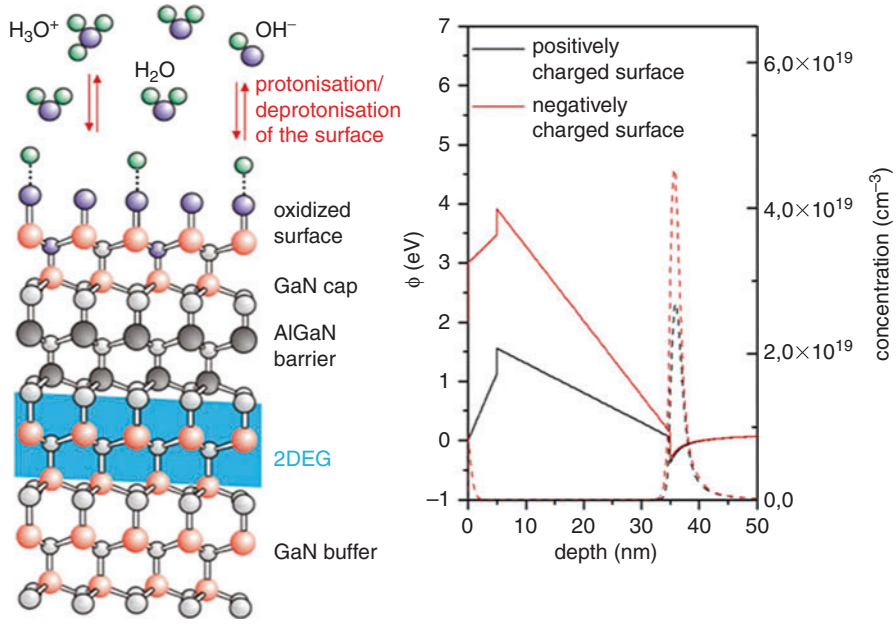


Polarization-Induced Effects in Heterostructures, Fig. 5 Basic structure of an AlGaN/GaN HEMT

Typically, room temperature mobility in the range of $\mu_n \sim 1000–1500$ cm^2/Vs is obtained for device applications. This mobility is affected by several scattering mechanisms. Limiting mechanisms in the AlGaN/GaN heterostructures are scattering at (1) threading dislocations, (2) due to the interface roughness between AlGaN and GaN, and (3) due to the random distribution of Al and Ga atoms within the alloy. The electron enhancement in AlGaN/GaN heterostructures was first reported in 1991 [10]. It was attributed to the two-dimensional character of the electrons, which has been demonstrated by mobility enhancement at lower temperatures as well as Shubnikov-de-Haas oscillations. Now, sheet concentrations of 10^{13} cm^{-2} or higher are achievable close to the interface, well in excess of those achievable in other III–V material systems [5]. In addition, the piezoelectric polarization of the strained top layer is more than five times larger as compared to AlGaAs/GaAs structures, leading to a significant increase of the sheet carrier concentration at the interface (Table 2).

The application of AlGaN/GaN HEMTs for microwave applications was proposed in 1994 [15]. Such HEMTs can be used in power as well as small-signal applications. In the first case, the device has to withstand high drain source voltages and drain currents. Here, the large breakdown voltage, the achievable high channel conductivity, and the good thermal properties of GaN are advantageous for the devices. A schematic cross section is given in Fig. 5. Typically, AlGaN/GaN HEMT devices are established on sapphire or SiC substrates, seldom also on silicon substrates [16]. First, thick GaN films are grown for reduction of growth defects and device isolation. The 2DEG channel is formed between this GaN buffer and an AlGaN top layer with a thickness of 15–30 nm. Occasionally, thin GaN (~3 nm) are grown on top of the heterostructure in order to improve the stability. Drain and Source contacts are usually Ti/Al based and are alloyed in a high temperature annealing step. The gate contact is crucial for operation at high frequency and has typically T- or Y-shaped cross sections. Additional field plates are employed to reduce the local field enhancement at the gate contact edges, thus enhancing the breakdown voltage. Finally, the device has to be passivated for stable operation, which is commonly achieved by deposition of SiN_x, SiO_2, Sc_2O_3, or related materials.

The achievable frequency roughly scales with the transconductance of the device. Thus, thin barriers (thickness d) as well as small gate lengths L_g are preferable for high-frequency operation. Small-signal amplifier transistors usually have small gate lengths (<200 nm); the breakdown voltage, however, is low (~ < 10 V). Power HEMTs, on the other hand, have rather uncritical



Polarization-Induced Effects in Heterostructures, Fig. 6 The hydroxyl groups on the GaN cap layer of an AlGaN/GaN heterostructure reacting with (**a**) water, (**b**) acidic, and (**c**) alkaline solutions and the resulting impact on the 2DEG sheet carrier density [17]

gate lengths. Especially for blocking voltages exceeding 100 V, gate lengths >1 μm are sufficient.

The strong polarization in AlGaN/GaN heterostructures is considered to be one source of some of the experienced reliability problems. To minimize such effects, alternate barrier materials such as InAlN have been employed, which are less polar. In addition, InGaN layers have been used instead of a pure GaN buffer as well as AlN interlayers at the AlGaN/GaN interface to improve the confinement of the 2DEG electrons.

Electrolyte-Gated Field Effect Transistors

The development of biosensors based on group III-nitrides received an increasing interest in the last years. In analogy to the development of biologically applied FETs (BioFETs) based on silicon enzymes, DNA, antibodies, and whole cells have already been successfully been used as transducing element. Especially the strong bond strength and the resulting high band gap are responsible for the excellent chemical stability and biocompatibility of the nitride materials. Therefore, no functionalization is required for the adhesion and proliferation of cells on the surface. In order to attach different biomolecules, the modification methods, which have been developed for Si-based BioFETs, can easily be adopted. Sensitivity and limit of detection for the first reported AlGaN/GaN-BioFETs were shown to be at least similar and often exceeding silicon-based devices and their background noise is significantly lower.

The strong responses of AlGaN/GaN field effect structures have been explained by the site-binding model, which has previously been applied to Si-ISFETs. According to this model, the metal oxide forms amphoteric hydroxyl groups (M-OH) at the surface when the surface is exposed to aqueous solutions. Depending on the pH value of the solution, these sites can either be protonated ($M\text{-}OH_{2+}$), neutral (M-OH), or deprotonated (M-O-) (Fig. 6). At a low pH value, i.e., a high concentration of H_3O_+, they act as a proton acceptor, and at high pH values, i.e., a high concentration of OH^-, they act as proton donators. This results in a change of the surface charge which in turn enhances/depletes the sheet charge density n_s in the 2DEG at low and high pH, respectively.

The pH response of the AlGaN/GaN heterostructure can also be used to record a pH change induced by a biochemical reaction in the solution, e.g., of an enzyme assay.

For the functionalization of group III-nitrides with biomolecules, different methods of covalent immobilization were applied which can be categorized as follows:

- Direct chemisorption of biomolecules on GaN via amine chemistry
- Deposition on organosilanes, e.g., aminopropyltriethoxysilane (APTES) followed by immobilization of biomolecules via amine chemistry
- Using the high affinity of thiols to a Au-coated gate directly via thiol-modified DNA or indirectly binding biomolecules through ester groups

One of the main advantages of group III-nitrides based ISFETs is their chemical stability under physiological conditions. Silicon is known to be attacked by many biological important agents; thus, an appropriate surface modification by native oxidization or by deposition of other metal oxides such as TaO_5 is required. Also, other semiconductors require a thorough passivation (e.g., GaAs in order to achieve a biocompatible surface). In contrast to that, (Al)GaN-based surfaces are naturally chemically inert, which greatly simplifies the device technology.

Microelectromechanical Systems

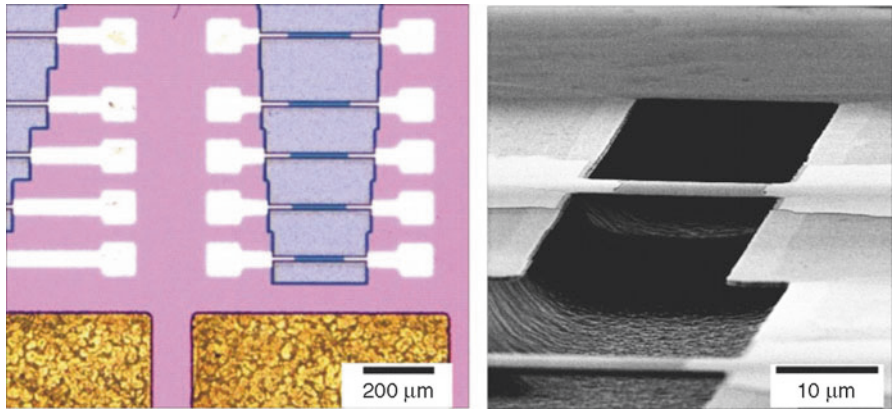
With the increasing requirements for microelectromechanical systems (MEMS) regarding stability, miniaturization, and integration, novel materials such as wide band gap semiconductors receive more and more attention. However, only few applications in MEMS have been reported yet, since the commonly used sapphire and silicon carbide heterosubstrates exhibit an extraordinary chemical stability and hinder the realization of freestanding functional layers by selective undercutting.

Nevertheless, wide band gap semiconductor materials, namely SiC, AlN, GaN, and AlGaN/GaN heterostructures, offer superior properties concerning mechanical, thermal, chemical, and biochemical stability. They show no or very low reaction with molecules from the air. For MEMS and NEMS, this is particularly advantageous as surface influences increasingly compromise their performance with decreasing device dimensions. Wide band gap materials exhibit outstanding ratios of the Young's modulus to the mass density, which, for a given resonator geometry and size, effects in significant higher mechanical operation frequencies than other materials, e.g., silicon, would allow [18].

The piezo- and pyroelectric properties of the group III-nitrides provide an elegant way for an integrated electromechanical coupling mechanism (Fig. 7). Thereby, the 2DEG serves as back electrode for the piezoelectric coupling of both, actuation and read-out [19]. Although other materials, e.g., functional ceramics such as Pb(Zr,Ti)O_3 (PZT), offer higher piezoelectric constants, the group III-nitrides convince by their structural quality and the dimensional control down to the monolayer level. The above-described 2DEG at the interface of AlGaN/GaN heterostructures can also be used for designing sophisticated MEMS-based sensor concepts. Due to its dependence on the spontaneous polarization of the interfacing materials, the 2DEG is sensitive to mechanical stress and other environmental influences such as temperature or chemical surface modifications. Additionally, this material allows the combination of MEMS and HEMT structures, thus opening the way to sensing concepts with integrated signal amplification.

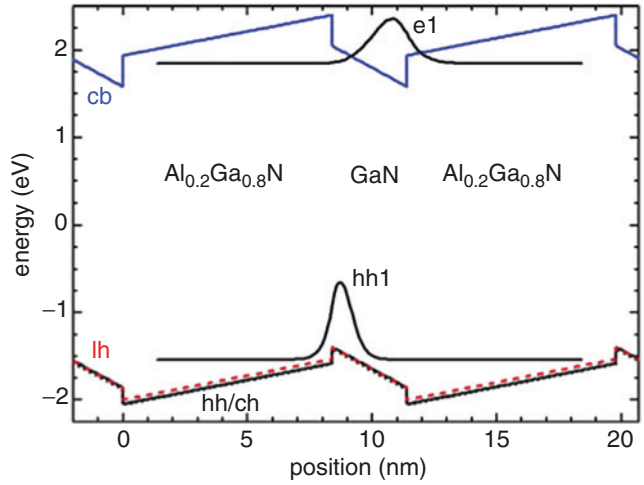
Unintentional Polarization Effects

The quantum-confined Stark effect (QCSE) describes the effect of an external electric field upon the light absorption spectrum of a quantum well (QW). Without electric fields, electrons and holes within the quantum well only occupy states within a discrete set of energy subbands. Consequently, only a discrete set of frequencies of light may be absorbed by the system. Under the influence of an external electric field, the electron states shift to lower energies, while the hole states shift to higher energies, thus reducing the permitted light absorption frequencies. Additionally, the



Polarization-Induced Effects in Heterostructures, Fig. 7 Double clamped beam structures of AlGaN/GaN on silicon. *Left*: optical micrograph, the size varies from 275 to 350 µm, *right*: scanning electron micrograph of a 20-µm beam [20]

Polarization-Induced Effects in Heterostructures, Fig. 8 Quantum-confined Stark Effect illustrated at an AlGaN/GaN/AlGaN quantum well structure, as simulated by nextnano^3 [20]



external electric field shifts electrons and holes to opposite sides of the well, decreasing the overlap integral, which in turn reduces the recombination efficiency (i.e., fluorescence quantum yield) of the system due to a spatial separation. In case of polar materials, the above-described shifts in the energy subband levels and in the localization of charge carriers are not caused by external fields, but inherently caused by the polarization-induced internal electric field. Figure 8 shows the conduction (cb) and valence (heavy hole (hh), light hole (lh), and crystal-field split-off hole (ch)) band edges of an AlGaN/GaN heterostructure, including the effects of strain, piezo- and pyroelectricity. The ground state electron (e1) and the ground state heavy hole wave functions (hh1) are shown. Due to the built-in piezo- and pyroelectric fields, the electron wave functions are shifted to the right and the hole wave function to the left.

Cross-References

- ▶ Basic MEMS Actuators
- ▶ Biosensors
- ▶ NEMS Piezoelectric Switches
- ▶ Piezoelectric Effect at Nanoscale

References

1. Ambacher, O.: Growth and applications of group III-nitrides. J. Phys. D Appl. Phys. **31**, 2653–2710 (1998)
2. Bernardini, F., Fiorentini, V., Vanderbilt, D.: Spontaneous polarization and piezoelectric constants of III-V nitrides. Phys. Rev. B **56**, R10024–R10027 (1997)
3. Yu, G., Levinshtein, M.E., Rumyantsev, S.L.: Silicon carbide (SiC). In: Levinshtein, M.E., Rumyantsev, S. L., Shur, M.S. (eds.) Properties of Advanced Semiconductor Materials GaN, AlN, SiC, BN, SiC, SiGe, pp. 93–148. Wiley, New York (2001)
4. Gopal, P., Spaldin, N.A.: Polarization, piezoelectric constants and elastic constants of ZnO, MgO and CdO. J. Electron. Mater. **35**, 538–542 (2006)
5. Ambacher, O., et al.: Two-dimensional electron gases induced by spontaneous and piezoelectric polarization charges in N- and Ga-face AlGaN/GaN-heterostructures. J. Appl. Phys. **85**, 3222–3233 (1999)
6. Ambacher, O., et al.: Pyroelectric properties of Al(In)GaN/GaN hetero- and quantum well structures. J. Phys. Condens. Matter **14**, 3399–3434 (2002)
7. Buchheim, C., Goldhahn, R., Gobsch, G., Tonisch, K., Cimalla, V., Niebelschuetz, F., Ambacher, O.: Electric field distribution in GaN/AlGaN/GaN heterostructures with 2DEG and 2DHG. Appl. Phys. Lett. **92**, 013510 (2008)
8. Yu, P.Y., Cardona, M.: Fundamentals of Semiconductors: Physics and Materials Properties. Springer, Heidelberg (2010)
9. Arulkumaran, S., Egawa, T., Ishikawa, H., Jimbo, T.: Characterization of different-Al-content AlGaN/GaN heterostructures and high-electron-mobility transistors on sapphire. J. Vac. Sci. Technol. **B21**, 888–894 (2003)
10. Khan, M.A., Van Hove, J.M., Kuznia, J.N., Olson, D.T.: High electron mobility GaN/AlxGa$_{1-x}$N heterostructures grown by low-pressure metalorganic chemical vapor deposition. Appl. Phys. Lett. **58**, 2408–2410 (1991)
11. Yngvesson, S.: Microwave Semiconductor Devices. Kluwer, Boston/Dordrecht/London (1991)
12. Manasreh, O. (ed.): III-Nitride Semiconductors: Electrical Structural and Defects Properties. Elsevier, Amsterdam (2000)
13. Manfra, M.J., Pfeiffer, L.N., West, K.W., Stormer, H. L., Baldwin, K.W., Hsu, D.V., Lang, J.W.P., Molnar, R.J.: High-mobility AlGaN/GaN heterostructures grown by molecular-beam epitaxy on GaN templates prepared by hydride vapor phase epitaxy. Appl. Phys. Lett. **77**, 2888–2890 (2000)
14. Khan, M.A., Kuznia, J.N., Olson, D.T., Schaff, W.J., Burm, J.W., Shur, M.S.: Microwave performance of a 0.25 μm gate AlGaN/GaN heterostructure field effect transistor. Appl. Phys. Lett. **65**, 1121–1123 (1994)
15. Tonisch, K., Jatal, W., Niebelschuetz, F., Romanus, H., Baumann, U., Schwierz, F., Pezoldt, J.: AlGaN/GaN-heterostructures on (111) 3C-SiC/Si pseudo substrates for high frequency applications. Thin Solid Films **520**, 491–496 (2011)
16. Cimalla, I., Gebinoga, M., Lübbers, B., Cimalla, V., Ambacher, O., Schober, A.: Group III nitride based sensors – a promising bio-sensing generation. In: Ho-Young, C. (ed.) Advanced Semiconductor Materials and Device Research: III-nitrides and SiC. Research Signpost, Kerala, (2011). ISBN: 978-81-7895-371-7
17. Brueckner, K., Niebelschuetz, F., Tonisch, K., Foerster, Ch., Cimalla V., Stephan, R. Pezoldt, J., Stauden, T., Ambacher, O., Hein, M.A.: Micro- and nano-electromechanical resonators based on SiC and group III-nitrides for sensor applications, Phys. Status Solidi A, 1–20 (2010)
18. Tonisch, K., Buchheim, C., Niebelschütz, F., Cimalla, B., Schober, A., Gobsch, G., Ambacher, O., Goldhahn, R.: Piezoelectric actuation of (GaN/)AlGaN/GaN heterostructures. J. Appl. Phys. **104**, 084516 (2008)
19. Brückner, K., Niebelschuetz, F., Tonisch, K., Michael, S., Dadgar, A., Krost, A., Cimalla, V., Ambacher, O., Stephan, R., Hein, M.A.: Two-dimensional electron gas based actuation of piezoelectric AlGaN/GaN microelectromechanical resonators. Appl. Phys. Lett. **93**, 173504 (2008)
20. Birner, S.: Nextnano^3 Tutorial. www.nextnano.de (2012)

Polyethylene Carbonate (PEC)

► SU-8 Photoresist

Polymer Coatings

Enrica De Rosa[1], Joseph Fernandez-Moure[1] and Ennio Tasciotti[2]
[1]Nanomedicine Department, The Methodist Hospital Research Institute, Houston, TX, USA
[2]Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Synonyms

Polymeric surface modifications

Definition

A polymer coating is defined as the chemical or physical binding of polymeric molecules on the surface of a substrate material.

Introduction

In the last two decades, the physical, chemical, and functional properties of polymers have been exploited to enhance the properties of nanomaterials. Polymer coatings are capable of shielding nanoparticles from adverse physiological environments, retarding their degradation or erosion, or protecting the body from the direct exposure to their toxic components. Recent advancements in polymer technology have also produced polymers able to change properties over time or according to environmental conditions (temperature, pH, etc.). Polymer-coated nanoparticles incorporated in nanocomposites have improved filler dispersion and developed materials with novel and sensational properties. Both natural and synthetic polymers can be used to modify nanomaterials through wet and dry surface-surface coupling techniques. Polymer coatings are currently considered as a valuable tool for the improvement of nanomaterials due to the fine control and tuning of their properties and due to the ability to add multiple functionalities to the engrafted surfaces. They have been employed in different settings, from medicine to oil industries, opening new promising application possibilities.

Natural and synthetic polymers have been used in medical applications for decades, thanks to their biocompatibility, biodegradability, and tunable properties. They have been injected, implanted, and used to host cells, to release drugs, to replace tissues, or to drive tissue regeneration. Historically, polymer coatings have been used in the surgical setting to improve prosthesis biocompatibility, facilitate bone implant fixation, and reduce fibrosis and contracture of mammary implants. Since the introduction of polymer coating in the field of medicine and surgery, their use has applied to the treatment, diagnosis, and monitoring of more complex pathologies such as cancer, cardiovascular disorders, and degenerative or metabolic diseases. The ability to obtain thin polymeric coatings and to control their self-assembly makes their use advantageous and versatile, especially in combination with nanomaterials and, in particular, with nanoparticles. This integration has served to not only protect nanoparticles from sequestration or degradation but also to lend them multiple functionalities. In essence, polymers can be considered the true enabler of the use of nanoparticles in the biomedical field, and today, multifunctional nanostructures combine the biomimetic and targeting properties enabled by the presence of complex coatings with their unique properties like remote control and photothermal activation.

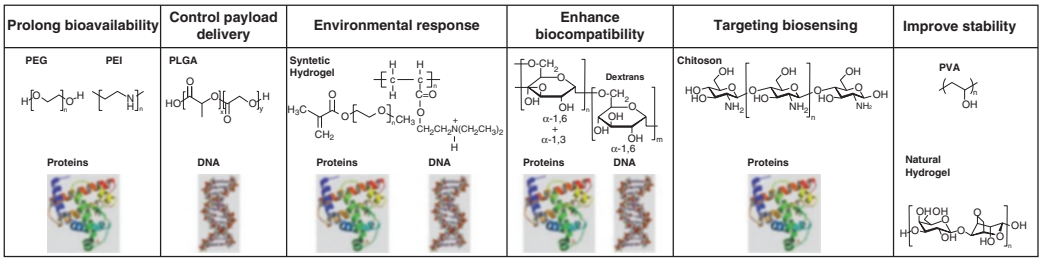
For these reasons, also in the oil industry, the use of polymer-coated nanoparticles has been investigated for enhanced oil recovery (EOR) processes. EOR aims to recover trapped oil left in reservoirs after primary and secondary recovery. Nanoparticles, being interfacially active and able to modify surface properties, have been explored for use in a remarkable range of applications [1]. Even if they have already been able to change the wettability of rock or stabilize emulsions and foams, in order to be used for EOR processes, they must be able to migrate through porous media and be dispersible in water/brine, inexpensive, and injectable into a reservoir [2]. Nanoparticles with polymer chains grafted to the surface, showing improved stability and solubility, greater stabilization of foams and emulsions, and more facile transport through porous media, may represent an emerging, superior class of materials for EOR.

For similar reasons, the use of nanoparticles as filler for polymeric matrices has generated materials with exceptional properties. Crucial for the proper integration of these particles is the interface between the nanoparticles and the polymeric matrix; recently, grafting to the nanoparticles, the homopolymer brush with the same chemistry of the matrix has shown less agglomeration and better mechanical properties [3].

Polymer Coating of Nanoparticles

The crucial aspect of any therapeutic treatment is the identification of the cellular and molecular targets and the localization of the therapeutic molecules in the right place at the right time and dose. Nanotechnology offers solutions to address the biodistribution and bioavailability of the drug of

Polymer Coatings, Table 1 Scheme of polymeric coating applications and relative coating materials



choice and to achieve higher therapeutic efficacy and reduce systemic toxicity [4]. Some nano-based delivery systems lack the ability to control their trafficking inside the cell where biological barriers such as the lysosomal sequestration limit the advantage of using nanoparticles. To overcome these limitations, physical or chemical coatings of nanomaterials with biocompatible and biofunctional polymers have been the focus of many laboratories in the past two decades [5] (Table 1). By chemically conjugating the appropriate ligands, antibodies, or targeting moieties to the reactive functional groups on the polymer coating, nanoparticles have been targeted to specific body districts as needed [6]. Coating nanoparticles with hydrophilic polymers such as poly(ethylene glycol) (PEG) reduces the interfacial energy in an aqueous environment, preventing unwanted aggregation. Nanoparticles coated with hydrophilic polymers are less recognized by macrophages and opsonins in the body. This allows the nanoparticles to stealthily circulate in the blood for a longer period of time, thus increasing their possibility to reach their target site. Some of the substrate materials used to synthesize nanoparticles (iron oxide nanocrystals, gold nanoshells, quantum dots, and nanofibers) and nanodevices (nanosensors, micro and nanofluidic devices) [7] include metal and metal oxide particles (gold, silver, iron, zinc, and silicon) [8]. These compounds are ideally suited for many biomedical applications given their ability to be tracked in vivo, to be triggered by external fields or forces, and to react with the tissue microenvironment. The drawback of some of these nanoparticles (such as quantum dots made of an inorganic mixture of cadmium selenide and zinc sulfide) is their inherent toxicity [9]. Natural and synthetic polymers can be used to increase the biocompatibility of this class of nanoparticles and made possible their use in the clinical and preclinical settings.

Natural Biomaterials and Polymers

A variety of natural biomaterials and polymers can be adapted for biomedical applications. Because of their intrinsic biocompatibility, they represent an excellent solution for in vitro and in vivo applications. They include: polysaccharides, proteins, peptides, and DNA.

Polysaccharides such as starch, chitosan, dextran, agarose, alginate, and cellulose are a class of long-chain sugars and carbohydrates consisting of repeated monosaccharide units. Coating using these materials not only improves biocompatibility and the colloidal stability of nanoparticles in a physiological environment but also provides binding sites for the further functionalization of nanoparticles with various biological molecules such as peptides, antibodies, enzymes, DNA, and RNA. Dextran coating was initially introduced to reduce particles' toxicity. Glycochitosans, hydrophobically modified polysaccharides able to self-assemble in a core-shell structure, were used to reduce tumor growth through the delivery of RGB peptides and paclitaxel [10]. Heparin nanoparticles have prolonged the half-life and ID50 of the encapsulated drug and resulted in increased therapeutic efficiency [11]. DNA layer-by-layer capsules, able to shrink

in response to environmental stimuli, have been proposed for the controlled release of drugs [12].

Proteins, peptides, and poly(amino acids) may provide hydrophobic, hydrophilic, or amphiphilic coatings to nanoparticles due to the secondary and tertiary arrangement of their moieties. Bovine serum albumin (BSA), collagen, gelatin, and silk are widely used as protein-based coatings for tissue-engineering scaffolds and biomimetic approaches [13]. Proteins such as β-galactosidase were used to enhance immune response [14], while proteins that undergo solution pH-dependent association and dissociation could be used to induce reversible association-dissociation transitions by varying pH [15]. These coatings can be used to target the nanoparticles to a specific tissue or organ in the body and to respond to the local microenvironment to activate the release of the drug. The main drawback of these applications is the long-term retention of the biological activity of the protein coating in circulation which still remains a major challenge of the field.

DNA, or deoxyribonucleic acid, is a complex polymer based upon nucleic acid units bound by ionic and hydrogen bonds. Multilayered DNA coatings have been proven to be cyto- and histocompatible and to increase fibroblast proliferation without altering cell morphology [16]. It was reported that DNA coatings on magnetic Fe_2O_3 nanoparticles could decrease the magnetic interparticle interaction while retaining individual magnetism of Fe_2O_3 nanoparticles. DNA coatings affected the osteoblast-like cell differentiation through an increase in osteocalcin deposition indicative of enhanced differentiation.

Synthetic Polymers

Numerous synthetic polymers have been developed to replace or mimic naturally occurring molecules for various bioapplications, most of which derive from hydrocarbon-based monomers. They have been used in the coating of nanoparticles for therapeutic delivery, evasion of cellular sequestration, and environmental interaction. Although varied in chemical formulation, the principles of biocompatibility and biodegradability inherent to these compounds are paramount to their success and to the success of other novel formulations.

Poly(ethylene glycol) (PEG) is a hydrophilic polymer that can be chemically/physically cross-linked via acrylate modification to form a hydrogel. Based on the method of cross-linking, layer thickness and physical and chemical properties of PEG-based hydrogels may be precisely tuned. Its chemical structure allows for limited protein adhesion or immune response. PEG coatings can therefore improve the biocompatibility of nanoparticles while simultaneously preventing their aggregation, thereby enhancing their colloidal stability in body fluid [17]. PEG modification also provides protection for nanoparticles from being taken up by macrophages and monocytes, allowing them to temporarily avoid the mononuclear phagocyte system and have a prolonged circulation time in vivo. Such polymers prevent the particle surface from adsorbing plasma components due to the steric barrier and therefore, accumulation of blood proteins and recognition by macrophages is greatly reduced. In vivo studies observed that PEG coating allowed formation of stealth particles that globally distributed much better through the animal, while the PEG molecular weight did not affect probe biodistribution [18]. A successful example of PEG-lipid conjugation now advanced into clinical testing phases is DOXIL, a liposomal formulation of the chemotherapeutic agent, Doxorubicin (DOX). The in vivo data showed 90 % suppression of 4T1 murine mammary carcinomas and reduced toxicity compared to DOX [19]. DOXIL is now FDA approved for ovarian and breast cancer treatment. However, recent studies also assessed that repeated intravenous injections of PEGylated liposomes are characterized by reduced circulation time. This finding suggests that prolonged treatment with PEGylated liposomes can generate antibodies to PEG and lead to unexpected pharmacokinetic behavior.

Another synthetic polymer is *polyvinyl alcohol (PVA)* prepared by hydrolysis of polyvinyl acetate. It is biocompatible, biodegradable, and readily functionalized. PVA is water soluble, but

Polymer Coatings, Table 2 Summary of nanomaterial substrates, coating materials, methods, and applications

Nanomaterial	Coating material	Coating method	Function/application
Nanoparticles	Synthetic	Emulsion/solution	Controlled drug delivery
Nanoporous microparticles	PEG	CVD	Prolonged bioavailability
Nanowires	PVA	PVD	Improved stability/stabilized suspension
Implant	PEI	Chemical, physical binding	Enhanced loading efficiency
Microfluidic device	PLGA/PLLA	Electro-nanospray	Modified biointeraction
	PCL	Plasma	
	Natural	LBL	
	Polysaccharide	RAFT	
	Protein/peptide		
	DNA/RNA		

PEG poly(ethylene glycol), *PVA* polyvinyl alcohol, *PEI* Polyethylenimine, *PLGA* poly(lactic-co-glycolic acid), *PLLA* poly-L-lactide, *PCL* polycaprolactone, *DNA* deoxyribonucleic acid, *RNA* ribonucleic acid, *CVD* chemical vapor deposition, *PVD* physical vapor deposition, *LBL* layer-by-layer assembly, *RAFT* reversible addition-fragmentation chain transfer

resistant to oils and organic solvents, thus it can protect nanoparticles from agglomeration and allow them to form a uniform colloid in an aqueous biological environment [20].

Polyethylenimine (PEI), another synthetic polymer, has been widely used for gene delivery. PEI can form complexes with DNA and condense them into compact nanoparticles [21]. It is also capable of destabilizing the endosomal membrane by triggering osmotic swelling in the endosomes to increase DNA release [22]. PEI has become one of the most efficient nonviral transfection agents. The PEI-coated nanoparticles have been used as a successful vector for gene delivery. In the last decades, different PEI modifications, including thiol cross-linking, polymer size, covalent attachment, and branching degrees, achieved high transfection efficiency of plasmid DNA and siRNA into cells and live animals. It is important to consider that high molecular weight PEI can cause toxicity. Recent studies have shown that it is possible to reduce toxicity-maintaining effective nucleic acid delivery by using intermediate-length polymers, and coating with low molecular weight PEI polymers, while inefficient for gene delivery, is nonetheless efficient to increase delivery of antitumor agents into cancer cells [23].

Methods of Surface Coating

The surface properties of nano- and microparticles can be tuned by coating a thin film of polymeric materials to achieve the desired physical, chemical, optical, electrical, mechanical, and biomedical properties. There are two conventional approaches for nano- and microparticle coating, categorized as dry or wet methods depending on the environment of polymerization reaction [24] (Table 2). Dry methods include physical vapor deposition, chemical vapor deposition, plasma treatment, and pyrolysis of polymeric or nonpolymeric polyethylene glycol organic materials for in situ precipitation of nanoparticles. Wet methods include sol-gel processes, emulsification and solvent evaporation techniques, and layer-by-layer assembly methods. Combination and repetition of one or more coating methods recently have also been investigated to better control molecular variables of the attached chain and improve properties such as filler dispersion.

Dry Surface Coating

The two principal methods for dry coating of nanomaterials are *physical vapor deposition (PVD)* and *chemical vapor deposition (CVD)*. PVD is a technique that condenses vaporized

materials in vacuum and at high temperature to form a thin-film coating onto the substrate surface. The vaporized materials are usually carried by H_2, N_2, or inert gasses, and there is no chemical reaction involved in this process [25]. CVD is a chemical method to deposit gaseous reactants to the surface of substrate materials to form a thin solid film through chemical reaction or decomposition of the precursor in an activated (heat, light, plasma) environment. CVD method has been used as a coating method for various materials including nano/microparticles, nanocrystals, nanowires, etc. [26].

Another option for dry coating is *plasma treatment* which can be used to directly attach polymers to a desired surface, generating highly branched or cross-linked polymers. The principle of this technique is that the ionized and excited molecules and radicals activated by the electrical field collide and react with the substrate surface. The plasma technique is environmentally friendly and usually performed at room temperature, which provides this technique with flexibility for surface modification or coating of almost all substrates. However, there are challenges to the deposition of thin films on nanoparticles due to their severe aggregation and high surface-to-mass ratio. In order to achieve uniform coating by this technique, it is necessary to homogeneously expose the surface to the plasma as any unexposed area will be minimally coated. To solve this problem, a fluidized bed reactor has been used to serve as an ideal tool for gas-particle reactions. The combination of these two technologies provides an innovative way for nanoparticle surface modification [27].

Wet Surface Coating

Wet surface coating refers to the coating of nanoparticles in a liquid (usually aqueous) solution. It can be achieved by *sol-gel processes*, emulsification/evaporation, or layered assembly. Sol-gel is a chemical technique where a chemical or colloid solution (sol) acts as a precursor for integrated network nano/microparticles (gel). This method is characterized by a gel-like diphasic system containing both a liquid and a solid phase. It can be applied to the synthesis of solid materials (such as glass or ceramics) at lower temperatures than the conventional approaches or in nano/microparticle coating. A thin film is deposited onto the surface of the nano/microparticles and the composition of the film can be tuned by controlling the reaction parameters (temperature, composition, etc.) [28]. *Emulsification* involves two or more immiscible liquids: a liquid phase (dispersed phase), composed of the desired coating materials dispersed into small droplets in a second phase (continuous phase) by constant stirring. Surfactants can be added to the second phase to stabilize the newly formed droplets of the dispersed phase against the coalescence. By resuspending metal, inorganic, or organic nanoparticles into the dispersed phase, their surface is coated with thin polymer layers [29]. *Layer-by-layer (LbL) assembly*, a versatile and easy-to-use technique, has become a nanofabrication method for a broad range of applications. Basically, when LbL assembly is applied to nanoparticle coating, it is used to sequentially absorb positively or negatively charged species onto the nanoparticles by alternatively submerging the different materials into solutions. The multilayered coatings can alter the intrinsic surface characteristics of nanoparticles for desired physical, chemical, mechanical, thermal, and optical properties pertaining to diverse biomedical applications [30] as in the following examples. The electrostatic and covalent LbL assembly strategy of Dhuram et al. has been used to engineer nanoparticles with multilayer shells, small size distribution, high stability in physiological media, low toxicity, protection against macrophage uptake, triggered release and activation only after endocytosis, and enzymatic cleavage. Another successful modification of the LbL method has been developed by Caruso and coworkers using colloidal particles to coat the surfaces of nanomaterials. LbL assembly of a lipid bilayer and polyelectrolyte structures can be used in drug and gene delivery, thanks to the lipid bilayer's ability of accommodating biomaterials. Katsuhiko Ariga's group has developed mesoporous nanocompartment films through LbL assembly of mesoporous silica capsules.

These films were able to modulate the release of the payload by means of a nonequilibrium evaporation process. Mohwald's emulsion encapsulation combines the advantages of direct encapsulation of a liquid colloidal core with the accuracy and multifunctionality of LbL polyelectrolyte deposition. Their capsules show two different destruction mechanisms upon interaction with solid surfaces and hence may find interesting application in the pharmacy and medicine industry. Synthesis through LbL is a very flexible and easy coating method that may provide multiple functions and enhance particle functionality.

Reversible addition-fragmentation chain transfer (RAFT) polymerization is currently used to synthesize polymer brushes on nanoparticles. Recently, also bidisperse/bimodal polymer brush-anchored nanoparticles have been synthesized through a double RAFT polymerization that enables to independently control molecular weight, composition, and graft density. Indeed, bimodal brush-grafted colloidal nanoparticles improve interaction and dispersion within the matrix [31].

Future Perspectives

Polymer coating is a versatile powerful tool in nanomedicine and other fields, such as the oil industry and material sector, and all evidence suggests that its use will continue to grow. Tuning nanoparticle surface properties and stability, polymer coatings are a promising instrument for nanomaterial composites synthesis and in EOR processes. The different polymer coatings developed so far are able to overcome several limitations particular to drug delivery and nanotechnology, such as biocompatibility, bioavailability, and gene delivery efficiency. Furthermore, advances in polymer chemistry and nanotechnology will continue to provide additional compounds to develop more effective nanostructures able to target the desired tissue and to release the entire payload in a successful fashion. Through the exploration of polymer combinations and through the advent of new coating techniques, it has been possible to pave a new avenue in the biomedical and commercial application of polymers. RAFT polymerization, for example, was used to synthesize binary polymer brush-anchored nanoparticles and produce composite materials with improved mechanical and thermal properties. By improving stability and tuning surface interactions such as adsorption, aggregation, viscosity, and emulsion generation, under different environmental conditions, polymer-coated nanoparticles may more easily propagate through a reservoir and lead to more efficient EOR processes. As polymer coating continues to add more functionalities to nanodevices and nanoparticles, the innovation in their synthesis and modification will continue to bring a fresh breath of life into the field and contribute to the use of nanomaterials to the growing biomedical field. In particular, environmental responsiveness offers a unique opportunity to add functionality to the already versatile nanosystems. By being able to interact with the surrounding environment, polymer coatings can generate a new paradigm of smart therapeutics that only act in the presence of a specific stimulus. Finally, the versatility and variety of polymer coatings could offer a solution to the issue of patient-to-patient variability encountered in any approach of personalized medicine. By providing an optimal therapeutic regimen for each individual patient in response to their specific condition and background, they might facilitate the introduction of the concept of individualized therapy in the clinical arena.

Cross-References

- ▶ Chemical Vapor Deposition (CVD)
- ▶ Nanomedicine
- ▶ Nanoparticles
- ▶ Nanostructured Functionalized Surfaces
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Nanotechnology
- ▶ Physical Vapor Deposition
- ▶ Self-Assembly
- ▶ Sol–Gel Method

References

1. Matteo, C., Candido, P., Vera, R., Francesca, V.: Current and future nanotech applications in the oil industry. Am. J. Appl. Sci. **9**, 784–793 (2012)
2. ShamsiJazeyi, H., Miller, C.A., Wong, M.S., Tour, J. M., Verduzco, R.: Polymer-coated nanoparticles for enhanced oil recovery. J. Appl. Polym. Sci. **131**, n/a 40576 (2014)
3. Akcora, P., Kumar, S.K., Moll, J., Lewis, S., Schadler, L.S., Li, Y., Benicewicz, B.C., Sandy, A., Narayanan, S., Ilavsky, J., Thiyagarajan, P., Colby, R.H., Douglas, J.F.: "Gel-like" Mechanical Reinforcement in Polymer Nanocomposite Melts, Macromolecules **43**, 1003 (2010)
4. Sahoo, S.K., Parveen, S., Panda, J.J.: The present and future of nanotechnology in human health care. Nanomed. Nanotechnol. Biol. Med. **3**, 20–31 (2007)
5. Mikhaylova, M., Kim, D.K., Bobrysheva, N., Osmolowsky, M., Semenov, V., Tsakalakos, T., Muhammed, M.: Superparamagnetism of magnetite nanoparticles: dependence on surface modification. Langmuir **20**, 2472–2477 (2004)
6. Wang, L., Zhao, W., Tan, W.: Bioconjugated silica nanoparticles: development and applications. Nano. Res. **1**, 99–115 (2008)
7. Moghimi, S.M., Hunter, A.C., Murray, J.C.: Nanomedicine: current status and future prospects. FASEB J. **19**, 311–330 (2005)
8. Ju-Nam, Y., Lead, J.R.: Manufactured nanoparticles: an overview of their chemistry, interactions and potential environmental implications. Sci. Total Environ. **400**, 396–414 (2008)
9. Clapp, A.R., Goldman, E.R., Mattoussi, H.: Capping of CdSe-ZnS quantum dots with DHLA and subsequent conjugation with proteins. Nat. Protoc. **1**, 1258–1266 (2006)
10. Kim, J.H., Kim, Y.S., Kim, S., Park, J.H., Kim, K., Choi, K., Chung, H., Jeong, S.Y., Park, R.W., Kim, I. S., Kwon, I.C.: Hydrophobically modified glycol chitosan nanoparticles as carriers for paclitaxel. J. Control. Release **111**, 228–234 (2006)
11. Park, K., Lee, G.Y., Kim, Y.S., Yu, M., Park, R.W., Kim, I.S., Kim, S.Y., Byun, Y.: Heparin-deoxycholic acid chemical conjugate as an anticancer drug carrier and its antitumor activity. J. Control. Release **114**, 300–306 (2006)
12. Johnston, A.P.R., Caruso, F.: Stabilization of DNA multilayer films through oligonucleotide crosslinking. Small **4**, 612–618 (2008)
13. Quaroni, L., Chumanov, G.: Preparation of polymer-coated functionalized silver nanoparticles. J. Am. Chem. Soc. **121**, 10642–10643 (1999)
14. Cui, Z., Mumper, R.J.: Coating of cationized protein on engineered nanoparticles results in enhanced immune responses. Int. J. Pharm. **238**, 229–239 (2002)
15. Meziani, M.J., Sun, Y.-P.: Protein-conjugated nanoparticles from rapid expansion of supercritical fluid solution into aqueous solution. J. Am. Chem. Soc. **125**, 8015–8018 (2003)
16. van den Beucken, J.J.J.R., Walboomers, X.F., Vos, M. R.J., Sommerdijk, N.A.J.M., Nolte, R.J.M., Jansen, J. A.: Biological responses to multilayered DNA-coatings. J. Biomed. Mater. Res. B **81B**, 231–238 (2007)
17. Jiang, J.-S., Gan, Z.-F., Yang, Y., Du, B., Qian, M., Zhang, P.: A novel magnetic fluid based on starch-coated magnetite nanoparticles functionalized with homing peptide. J. Nanopart. Res. **11**, 1321–1330 (2009)
18. Maldiney, T., Richard, C., Seguin, J., Wattier, N., Bessodes, M., Scherman, D.: Effect of core diameter, surface coating, and PEG chain length on the biodistribution of persistent luminescence nanoparticles in mice. ACS Nano **5**, 854–862 (2011)
19. Charrois, G.J., Allen, T.M.: Multiple injections of pegylated liposomal doxorubicin: pharmacokinetics and therapeutic activity. J. Pharmacol. Exp. Ther. **306**, 1058–1067 (2003)
20. Lin, C., Tsai, H., Tung, C.: Preparation of Silver Nanoparticles/Pseudo-Thermoplastic Polyvinyl Alcohol (PT-PVA) Films by the Synchronous Chemical Reduction Method, Polym. Plast. Technol. Eng. **48**, 1171–1175 (2009)
21. Dunlap, D.D., Maggi, A., Soria, M.R., Monaco, L.: Nanoscopic structure of DNA condensed for gene delivery. Nucleic Acids Res. **25**, 3095–3101 (1997)
22. Remy-Kristensen, A., Clamme, J.P., Vuilleumier, C., Kuhry, J.G., Mely, Y.: Role of endocytosis in the transfection of L929 fibroblasts by polyethylenimine/DNA complexes. Biochim. Biophys. Acta **1514**, 21–32 (2001)
23. Xia, T., Kovochich, M., Liong, M., Meng, H., Kabehie, S., George, S., Zink, J.I., Nel, A.E.: Polyethyleneimine coating enhances the cellular uptake of mesoporous silica nanoparticles and allows safe delivery of siRNA and DNA constructs. ACS Nano **3**, 3273–3286 (2009)
24. Wang, X., Zhou, L., Ma, Y., Li, X., Gu, H.: Control of aggregate size of polyethyleneimine-coated magnetic nanoparticles for magnetofection, Nano. Res. **2**, 365–372 (2009)
25. Choy, K.L.: Chemical vapour deposition of coatings. Prog. Mater. Sci. **48**, 57–170 (2003)
26. Zhang, B., Liao, Y.C., Girshick, S.L., Roberts, J.T.: Growth of coatings on nanoparticles by photoinduced chemical vapor deposition. J. Nanopart. Res. **10**, 173–178 (2008)
27. Tatoulian, M., Brétagnol, F., Arefi-Khonsari, F., Amouroux, J., Bouloussa, O., Rondelez, F., Paul, A. J., Mitchell, R.: Plasma deposition of allylamine on polymer powders in a fluidized bed reactor. Plasma Processes Polym. **2**, 38–44 (2005)
28. Klein, L.C.: *Sol-Gel Optics: Processing and Applications* (1994)

29. Hielscher, T.: Ultrasonic production of nano-size dispersions and emulsions. Proceedings of European Nanosystems Conference ENS'05 Paris (2005)
30. Varahramyan, K., Lvov, Y.: Proceedings of the institution of mechanical engineers. Part N J. Nanoeng Nanosyst **220**, 29–37 (2006)
31. Rungta, A., Natarajan, B., Neely, T., Dukes, D., Schadler, L.S., Benicewicz, B.C.: Grafting bimodal polymer brushes on nanoparticles using controlled radical polymerization. Macromolecules **45**, 9303–9311 (2012)

Polymer Nanocapsules

► Acoustic Nanoparticle Synthesis for Applications in Nanomedicine

Polymer Pen Lithography

► Dip-Pen Nanolithography

Polymer Sensors

► Organic Sensors

Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing

Tommaso Baldacchini[1,2], Lijia Jiang[2], Wei Xiong[2] and Yongfeng Lu[2]
[1]Technology and Applications Center, Newport Corporation, Irvine, CA, USA
[2]Department of Electrical Engineering, University of Nebraska–Lincoln, Lincoln, NE, USA

Definition

In laser direct writing, two- and/or three-dimensional patterns are fabricated by the interaction of a focused laser beam with a target. Complex geometries are easily attainable by free-form writing with feature sizes as small as hundreds of nanometers. Two-photon polymerization is a specific case of direct laser writing where a photosensitive material is polymerized using a laser by means of multiphoton absorption.

Overview of Laser Direct Writing

Lasers exhibit a number of features that make them unique sources of energy for several material processing methods. For example, they are indispensable tools for drilling, welding, and cutting in industrial manufacturing where high average and peak powers are required. Furthermore, by virtue of their high spatial coherence, lasers can produce patterns with resolutions that extend over four orders of magnitude from millimeters to hundreds of nanometers. And lastly, materials can be removed, added, or modified without physical contact between a tool and the target substrate allowing rapid and inexpensive turnover.

Many manufacturing techniques that use lasers to create patterns on the surface or in the volume of a material can be described as laser direct-write (LDW) processes [1]. In LDW, the interaction between the laser and the material can create different effects: voids by means of ablation, material state modifications by means of thermal cycles, and addition of diverse materials by means of chemical reactions or induced forward transfer procedures.

LDW creates patterns in a "spot-by-spot" fashion where the target substrate is moved around a fixed laser beam and/or the laser beam is scanned following predetermined outlines. Although this serial approach to create parts is slow compared to more traditional parallel approaches, two characteristics of LDW methods render it attractive and competitive in many applications. First, there are almost no topological constraints for creating structures. Second, by choosing the appropriate laser and experimental conditions virtually, any material can be processed. Examples of materials processed by LDW can be found among

conductors, semiconductors, dielectrics, and ceramics [2].

Both continuous wave and pulsed lasers have been successfully employed in LDW. In particular, the ones that have found the most uses thus far are infrared CO_2 gas lasers, ultraviolet excimer gas lasers, and visible solid-state lasers. When interacting with these kinds of lasers, materials absorb light in a linear fashion, and heating effects play an important role on the quality of the created patterns. If ultrashort (less than a picosecond) pulsed lasers are used for LDW, then energy is deposited in the material in a time range shorter than the electron–phonon coupling time (a few picoseconds) rendering thermal effects negligible [3]. Moreover, the likelihood to induce multiphoton absorption when using ultrashort pulsed laser is such that the matter/light interaction can be spatially confined to femtoliter volumes enabling high-precision and high-resolution patterning. Consequently, LDW processes that use ultrashort pulsed lasers have found a large number of applications in diverse fields as microfluidics, photonics, MEMS, cell development, tissue engineering, and medicine [4].

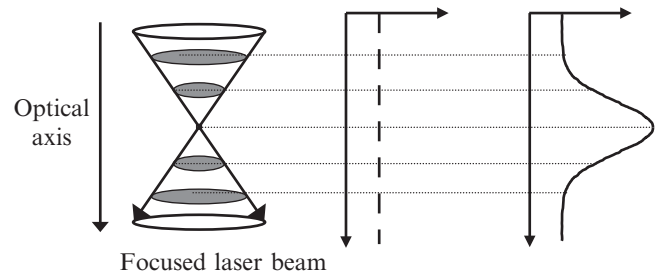
For many technologies, smaller is synonymous with better. Shrinking the size of devices brings many advantages: more components per unit area, lower power consumption, lower cost, faster response, and higher performance. An area of research that has affected almost every aspect of society because of device downsizing is microelectronics. Staggering examples of these changes can be found across diverse fields including automotive, computers, communications, and medicine. The successful story of microelectronics is based on the development of a set of tools that goes under the name of integrated chip (IC) technology and comprises processes such as photolithography and electron beam lithography [5]. By means of IC technology, microprocessors with millions of transistors can be routinely fabricated and launched in enormous numbers of calculations per second.

With the increasing complexity of micro-/nano-devices needed in most modern applications though, IC technology presents some limitations [6]. In particular these techniques are incompatible with many chemical and biological environments, creating features smaller than 100 nm is difficult, and they are not capable of fabricating structures with complex three-dimensional shapes. Furthermore, they cannot fabricate features on a curved or uneven substrate, and generally require expensive facilities. For these reasons, there has been a tremendous effort from the research community to develop unconventional micro-/nanofabrication techniques. Nano-imprint lithography, proximity probe lithography, and soft lithography were born from the desire to break free from the constraints of conventional lithography [7].

Two-photon polymerization (TPP) is a laser direct writing process for the fabrication of polymeric three-dimensional microstructures [8]. Because of its ability to produce geometries with no topological constraints with features smaller than 100 nm, TPP is a unique tool in the arsenal of unconventional microfabrication methods now available to scientists. Although the first report in the scientific literature of photopolymerization induced by a multiphoton absorption process appeared more than 40 years ago, it was a paper of Kawata and coworkers published in 1997 that established TPP as a technique for the fabrication of three-dimensional microstructures [9, 10]. Since then the number of works based on TPP has grown rapidly, allowing researchers in diverse fields such as nanophotonics, microelectronics, and bioengineering to create and use devices previously impossible to manufacture with conventional microfabrication procedures.

Three-Dimensional Micro-/Nanofabrication by Two-Photon Polymerization

The excited state of a molecule can be reached by the absorption of one photon of an appropriate wavelength or by the simultaneous absorption of two photons, each having half the energy required for the transition. In this case, two-photon



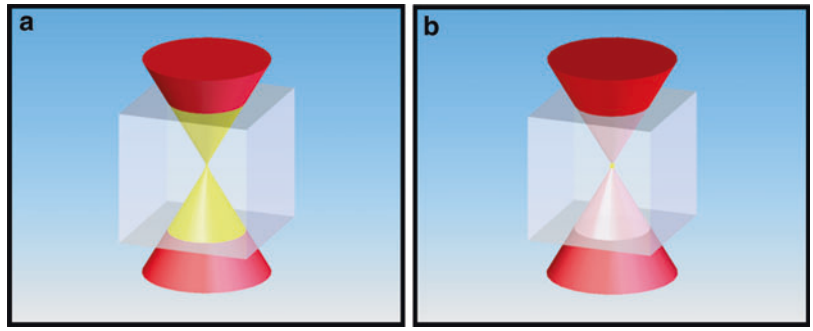
Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 1 Integration of the intensity (—) and the intensity square (—) in each transverse section of the focused laser beam along the optical axis. While the probability of one-photon absorption is always the same, the probability of two-photon absorption is highest at the focal point

absorption is said to be degenerate. When the photons promoting the transition are of different energies (wavelengths), the process is said to be nondegenerate. Because of the simplicity of the experimental setup for degenerate two-photon absorption, most of the applications have used this technique. Two-photon cross sections of most molecules are very small, usually on the order of 1 GM or less, where 1 GM corresponds to 10^{-50} cm^4 s $photon^{-1}$ (GM stands for Göppert-Mayer in honor to the scientist who theoretically predicted multiphoton absorption). As a result, high local photon fluxes are needed to promote two-photon absorption. Commercially available Ti:sapphire lasers emitting at wavelengths in the near-infrared region of the spectrum and producing pulses shorter than 100 fs can meet this requirement. Each pulse has a peak power on the order of a kilowatt, while at a typical repetition rate of 80 MHz, the average power is on the order of milliwatts. When focused into tight spots, these lasers allow reaching the photon densities needed for two-photon absorption to occur.

Since the probability of a two-photon absorption event is proportional to the second power of the light intensity, excitation can be spatially localized. Figure 1 illustrates the total one-photon (dashed line) and two-photon (continuous line) absorption per transverse section of a focused laser beam. While the absorption probability is the same along the optical axis for a one-photon process, a maximum coinciding with the focal point is present for the two-photon process. Researchers have used the ability to confine excitation in a volume as small as a femtoliter by two-photon absorption in several applications, including photolysis, optical data storage, photochromism, and photodynamic therapy. Among these areas of research, the one that has experienced the most explosive growth is two-photon fluorescence microscopy [11].

In polymerization induced by a one-photon process, the excitation beam is usually greatly attenuated by linear absorption before reaching the focal point. Consequently, only the first layer of the photoresist (photosensitive material) is polymerized. If the intensity of the laser is increased, all the material along the optical axis may polymerize (Fig. 2a), or worse the energy dumped into the system is high enough to boil the material away. In either case the lack of polymerization control forbids the use of one-photon absorption for the fabrication of three-dimensional microstructures. Conversely, polymerization is confined to the focal point when excitation occurs via two-photon absorption (Fig. 2b).

In TPP the selective excitation of molecules within the focal volume of a tightly focused beam allows for a spatial confinement of matter transformation. If the photoresist can be solidified by two-photon absorption, then three-dimensional microstructures with submicron resolution can be fabricated by means of accurate positioning of the laser focal point. As long as the solubility properties of the solidified and unsolidified photoresist



Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 2 One-photon (**a**) and two-photon (**b**) polymerization. Only when nonlinear excitation occurs (2b), spatial confinement of the chemical reaction is observed

are different, the unpolymerized material can be washed away to leave the freestanding structures. A diagram representing the three main steps of TPP are shown in Fig. 3 where the proportions of the objects depicted have been exaggerated for viewing clarity.

Since the index of refraction of the photoresist changes upon polymerization, TPP can be monitored in real time by the aid of white light microscopy. Although in this way large failures of the procedure can be detected and corrected in time, the fine details of the microstructures cannot be observed. Furthermore, a three-dimensional view is impossible. Hence, prior to being employed in experiments, microstructures fabricated by TPP are carefully investigated by scanning electron microscopy (SEM) in order to check if the desired geometry and surface roughness were obtained. Figure 4 is an SEM image of a microstructure fabricated by TPP. The sample was positioned at a 45° tilt in order to reveal the complex topology.

This fabrication process offers a unique combination of advantages. First, no topological constraints are present in the fabrication of three-dimensional structures. Second, sub-diffraction-limited feature sizes can be attained by employing laser intensities just above the intensity threshold at which polymerization will occur. Third, the optical penetration depth is greatly increased with the use of near-IR light. Fourth, movable components can easily be created without the use of sacrificial layers. Finally, the carbon-based nature of the materials can be used as a chemical handle to fabricate structures with tunable physical and chemical properties such as hardness, shrinkage, index of refraction, and chemical specificity, depending upon their application.

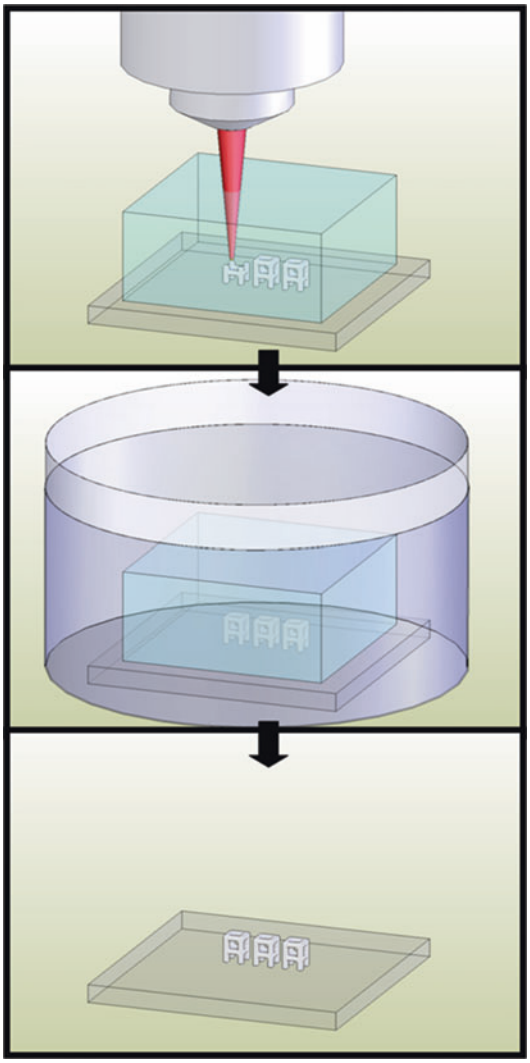
Size Characterization of Features Fabricated by Two-Photon Polymerization

The ultimate resolution and complexity of three-dimensional microstructures fabricated by TPP depends upon factors influencing the shape and size of the smallest polymerized volume element or voxel. It is the accurate overlapping of these voxels that produces the desired surface topology of the final structure.

The modification of the photoresist state from liquid to solid occurs in the region with the highest concentration of radicals. In TPP this corresponds to the central part of the square light intensity distribution at the focal plane. As a consequence, the shape of the polymerized voxels resemble an ellipsoid with the two minor axes equal to each other and lying in the geometric focal plane.

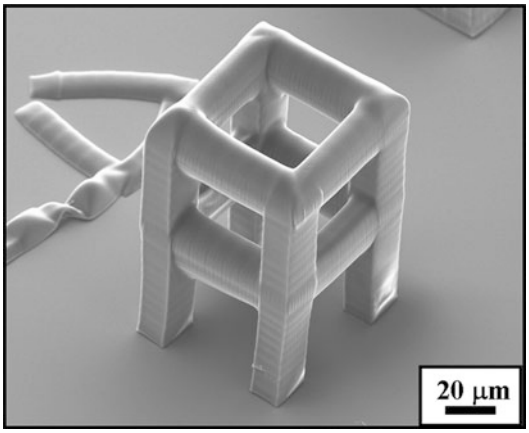
The volume of the polymerized voxel is proportional to λ^3, where λ is the wavelength of the laser beam used to induce two-photon absorption. Although shorter wavelengths will form smaller voxels, the choice of the excitation energy is dictated mainly by the photoinitiator efficiency. In general, the working wavelength of choice optimizes the product of the photoinitiator two-photon cross-section area and quantum yield for radical generation.

By assuming a Gaussian distribution of light intensity at the focal volume and a constant photon flux during a laser pulse, an equation predicting the voxel dimensions as a function of experimental parameters can be found. As expected, both voxel diameter and length increase with increasing values of exposure times and average powers.



Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 3 Free-standing three-dimensional microstructures can be fabricated by TPP. First the desired pattern is "written" by precise positioning of the laser focal point in the photoresist. Then, the unsolidified remainder of the photoresist is washed away by means of an organic solvent, leaving only the newly fabricated microstructures on the substrate

When the conditions for TPP are optimized to achieve the best resolution, the ratio between the axial and lateral dimensions of the polymerized voxel (μ) varies from three to five. By increasing either exposure time or laser average power, this ratio becomes larger implying that the length of the voxel grows more rapidly than its diameter.

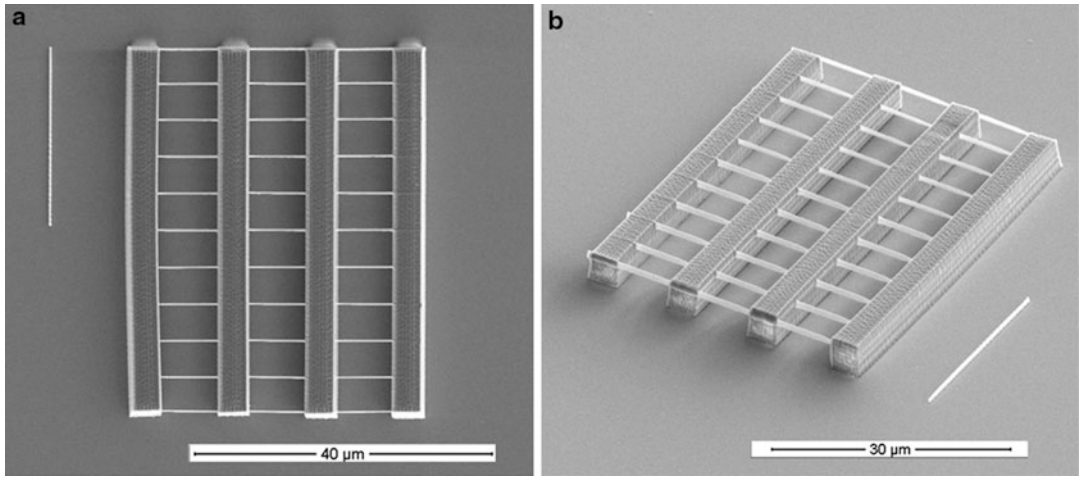


Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 4 SEM image of a three-dimensional microstructure fabricated by TPP

Furthermore, μ is affected in larger degree by the laser average power than by the exposure time, and it saturates at long exposure times.

To obtain the highest fidelity of fabrication when writing a complex three-dimensional microstructure by two-photon polymerization, it is of uttermost importance the characterization of the voxel lateral and axial dimensions. One experimental technique employed to retrieve this information is the so-called ascending method. Separate voxels are fabricated on the substrate's surface keeping all writing conditions the same, but the laser focal volume positions in reference to the substrate/resin interface. Because a polymerized voxel will have a defined length, some of the voxels will be fully anchored to the substrate, while other will be lying sideway due their weak adhesion to the substrate. By imaging these latter voxels by SEM, it is possible to measure both their length and diameter.

Although this method has the advantage of reveling in one image both dimensions of a voxel, it has the drawback of not giving a direct measure of the voxel size when performing TPP. In almost all cases, TPP occurs by moving either the sample around a fixed laser beam or the laser beam within a fixed resin. In both cases, the microstructure is fabricated with a continuous overlap of voxels. Hence, a more direct way to get information on the feature size used in TPP is



Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 5 SEM images of a test microstructure used to measure the width and the length of suspended lines fabricated by TPP. When the specimen is not tilted (**a**), a clear visualization of the line width is possible. When the specimen is tilted 45° (**b**), the axial dimension of the line is revealed

to fabricate lines suspended from the substrate by means of tall supports. These lines, generated by a single laser passes, can then reveal both the width and length of the polymerized feature size when imaged by SEM at different angles. An example of such a study is shown in Fig. 5.

Linear Versus Nonlinear Optical Absorption in Laser-Induced Photopolymerization

In two-photon polymerization, an intensity laser threshold exists below which no polymerization occurs. In the specific case of resins that polymerize through a radical mechanism, the presence of this threshold is a direct consequence of the chemical nature of the resins' components. The growth of macromolecular chains within the resin depends greatly on delicate ratios of the rate constants among several processes that occur at the same time. These are initiation, propagation, recombination, and termination. If radicals are not continuously generated during the initial stages of the polymerization reaction, no chains will be formed. In polymerization induced by light absorption, this mechanism results in the presence of an intensity threshold.

In first approximation, the accumulated exposure dose D of a single voxel induced for a N-photon absorption process can be described as:

$$D \propto (P - P_{th})^N t, \tag{1}$$

where P is the laser average power, P_{th} is the laser threshold power, and t is the exposure time [12].

According to this model, in one-photon polymerization ($N = 1$), the exposure dose is linearly proportional to the laser average power. In two-photon polymerization ($N = 2$), the exposure dose becomes proportional to the squared of the laser average power. To take advantage of this model from an experimental point of view, a series of voxels are fabricated using different values of both laser average powers and exposure times. Among these voxels, there will be some with identical width that are created with different combinations of laser average powers and exposure times. Assuming that identical voxels are fabricated when the same exposure dose is employed, a log–log plot of exposure times and laser average powers used to create identical feature sizes can reveal the value of N in Eq. 1. Hence, careful dosimetry experiments, such as the ones described in the previous section, not only produce useful information regarding

writing conditions to use when performing high-resolution TPP, but they can also offer insightful evidence on the nature of the light absorption process that starts polymerization.

Characterization of Microstructures Fabricated Using Two-Photon Polymerization Using Raman Microspectroscopy

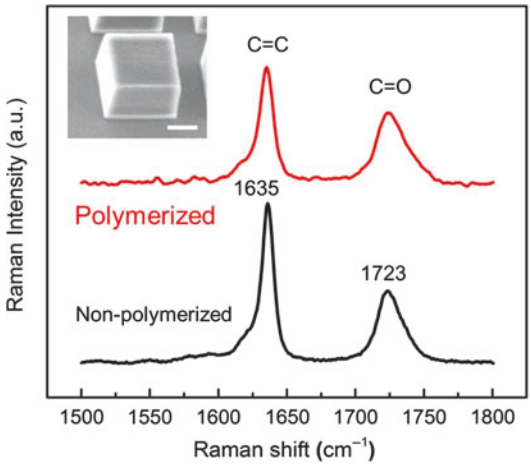
Although TPP has become an enabling technology for several applications, there are still some fundamental aspects that have not yet been fully characterized. One of these aspects is the influence of the scanning pattern on the mechanical properties of the microstructures. Even if the same object can be realized utilizing different writing strategies, some approaches are better than others in producing accurate and dimensionally robust microstructures. Another aspect of TPP that requires a better understanding is the influence of the developing process on polymer swelling and solvent inclusion which might have a large impact on the structural integrity of the microstructures. To answer these and other questions, it is required to use an analytical technique capable of investigating the materials used in TPP at the molecular lever before and after laser processing. Raman microspectroscopy is a formidable tool for the characterization of microstructures fabricated by TPP. It allows for in situ and nondestructive monitoring of specific molecular bond vibrations with high spectral resolution and minimal sample preparation [13].

Raman provides comprehensive information about the vibrational modes of molecules by using light scattering. The Raman-scattered light from the molecules in the probed sample is shifted in the frequency domain relative from the excitation light center frequency, and it contains information about the vibrational frequencies and anisotropy of molecular vibrations. Vibrational modes are like fingerprints of molecules. Therefore, Raman scattering is endowed with the sensitivity to discern different molecular species. Raman spectroscopy has broad applications, ranging from fundamental physical chemistry, inorganic chemistry, biochemistry, art, and archaeology to highly practical forensic examinations.

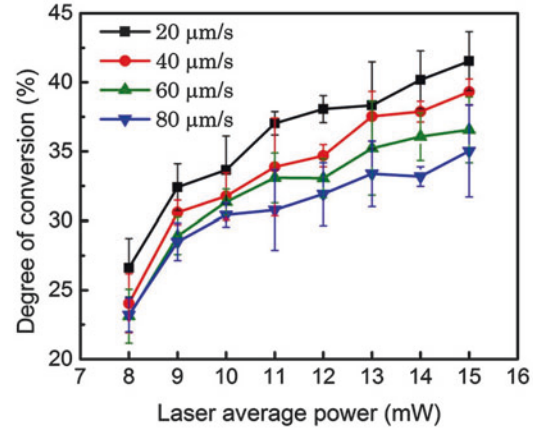
Most of the work in LDW by TPP has been performed using acrylic-based resins. To illustrate the utility of Raman microspectroscopy in characterizing TPP is thus useful to briefly describe the polymerization process involved. Acrylic-based resins are composed mainly of two molecular components, a photoinitiator and a mixture of monomers/oligomers [14]. The first one is the molecule capable of undergoing excitation by two-photon absorption and of generating active species (radicals) that start the polymerization process. In the presence of excited photoinitiators, the monomer/oligomer mixture forms a complex linked network, a polymer. Monomers and oligomers are present in the resin in much larger numbers than photoinitiators.

Radicals generated during laser irradiation react with the monomer/oligomer mixture to form high molecular weight materials through a chain growth mechanism. Several reactions with different rates for initiation, propagation, chain transfer, and termination occur at the same time. In chain polymerization, monomers and oligomers react only with the propagating reactive center, not with other monomers and oligomers, and chain addition ceases when the active species are depleted by a number of termination reactions. The molecular weight of chain polymers increases rapidly during polymerization, and monomer/oligomer to polymer conversion (degree of conversion) can widely range between 20 % and 90 % depending on the chemical nature of the monomers and oligomers. In acrylic monomers and oligomers, chain addition takes place in the carbon–carbon double bonds (C=C) present in the molecules ester moiety. Thus, during TPP the number of C=C bonds diminishes in favor to the formation of new intermolecular carbon–carbon single bonds (C–C).

The Raman spectra of a typical resin employed in TPP before and after polymerization are shown in Fig. 6. Two distinctive peaks are observed at 1,635 cm^{-1} and 1,723 cm^{-1} 1,635 which are ascribed to the vibrational modes of C=C and the carbonyl (C=O) groups of the monomer,



Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 6 Raman spectra of a typical acrylic-based resin used in TPP before and after polymerization. The inset shows a scanning electron micrograph of a solid cube created by TPP and used to perform Raman microspectroscopy (5 μm scale bar)



Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 7 Polymeric degree of conversion of microstructures fabricated by TPP using Raman microspectrometry

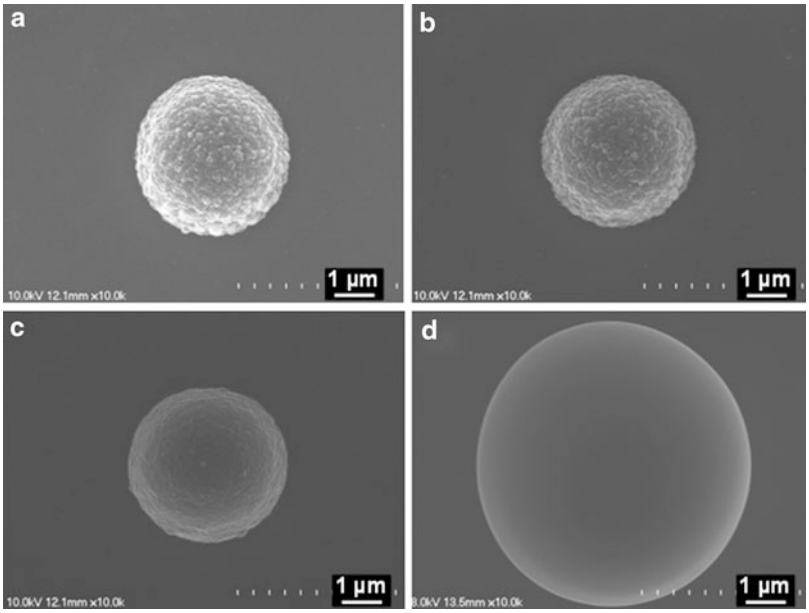
respectively. As a consequence of the diminished concentration of the C=C bonds following polymerization, the intensity of the peak at 1,635 cm^{-1} in the polymerized resin is smaller than in the unpolymerized resin. On the other hand, the intensity of peak at 1,723 cm^{-1} remains the same before and after polymerization because the carbonyl group does not participate in the chain reaction.

The area of a peak in a Raman spectrum is proportional to the concentration of the oscillators responsible for that particular Raman active mode. Thus, spectra such as the ones in Fig. 3 can be used to measure the degree of conversion (DC) in resins following TPP. DC represents the number of C=C bonds consumed during polymerization and is a major factor influencing the mechanical properties of the ultimate microstructure; the higher the DC, the stronger the polymer. Since the number of carbonyl groups in the resin remains the same before and after polymerization, the peak at 1,723 cm^{-1} can be used as internal reference for DC estimation. The percentage of DC is obtained using the following equation:

$$\mathrm{DC} = \left[1 - \frac{A_{C=C}/A_{C=O}}{A'_{C=C}/A'_{C=O}}\right], \qquad (2)$$

where $A_{C=C}$ and $A_{C=O}$ are the integrated intensities of the 1,635 cm^{-1} and 1,720 cm^{-1} peaks in the polymerized resin, respectively, while $A'_{C=C}$ and $A'_{C=O}$ are the integrated intensities of the same peaks in the unpolymerized resin.

To evaluate the usefulness of Raman microspectroscopy in TPP, a series of microcubes were fabricated first and then characterized using a confocal Raman micro-spectrometer. An SEM image of a microstructure used in the study is shown in the inset of Fig. 6. The micro-cubes were fabricated using several laser average powers and writing speeds. For each microstructure, Raman spectra like the ones in Fig. 6 were recorded and then used to measure DCs with Eq. 2. The results are shown in Fig. 7. At constant writing speeds, DC values monotonically increase with increasing laser average power. The highest DC value is around 43 %; larger DCs were not attainable because the resin will start to boil if higher laser average powers were used. During TPP, highly cross-linked macromolecules are formed that can severely limit molecular mobility. The probability that an unreacted C=C bond will encounter a radical is dramatically reduced with



Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 8 Surface quality of 3D microstructures fabricated by TPP using different laser scanning methodologies: (**a**) raster scanning mode, 200-nm step; (**b**) raster scanning mode, 100-nm step; (**c**) raster scanning mode, 20-nm step, 1-h processing time; (**d**) annular scanning mode, 10-nm step, 20-min processing time

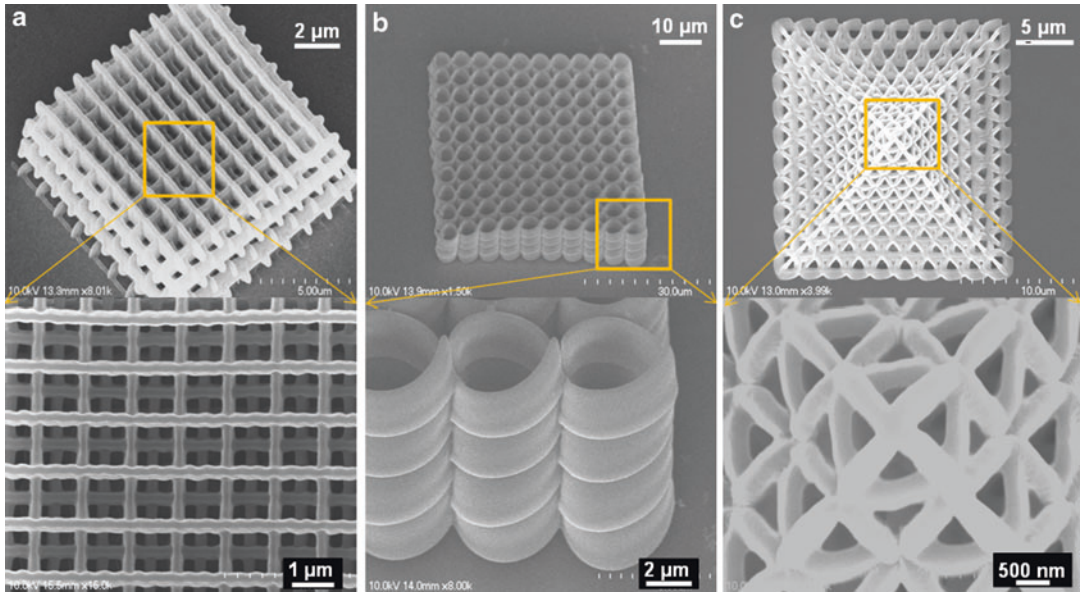
the evolving polymerization, and a complete conversion (DC = 100 %) is never achieved.

Array of Complex 3D Microstructures Fabricated by Two-Photon Polymerization

TPP is an additive femtosecond laser direct writing technique that uses photosensitive resins in 3D micro-/nanofabrication. Compared with other micro-/nanofabrication approaches, TPP offers both high resolution and relatively high throughput. TPP is intrinsically a three-dimensional microfabrication technique. Because of the nonlinear optical nature of light absorption, polymerization is confined in the focused laser beam. By moving the sample relative to the fixed focal position, arbitrary paths can be written into the resin. It can be simply described as using a laser beam to inscribe micro-/nanostructures in three dimensions. Compared to conventional stereolithography, TPP provides much smaller feature sizes allowing the reproduction of three-dimensional objects at the micron scale with unprecedented fidelity.

Microstructures targeted for optical application, such as micro-lens arrays, require high-quality surface profiles. When using TPP to fabricate these microstructures, the writing process needs to be optimized accordingly in order to obtain the desired surface smoothness. For example, it was found that raster scanning mode can hardly get high-quality surface as shown in Fig. 8a–c. In order to obtain a smoother micro-lens, an annular scanning mode cooperated with continuous variable scanning steps was adopted. By using this new scanning method, spherical micro-lens with high surface quality was readily produced (Fig. 8d).

Besides manufacturing micro-lens with excellent surface quality, TPP can easily produce another type of passive optical elements that otherwise will be difficult or impossible to make, and this is a three-dimensional photonics crystal. Photonic crystals are structures that present periodic alternation of materials with different indices of refractions. The most common photonic crystal is the one with a log–pile geometry that consists of layers of parallel rods in which the rod orientation in each layer is perpendicular to that of the adjoining layers. Photonic crystals have the interesting property of forbidding the propagation of certain modes of the electromagnetic spectrum. It turns out that this photonic band gap phenomenon occurs at wavelengths corresponding



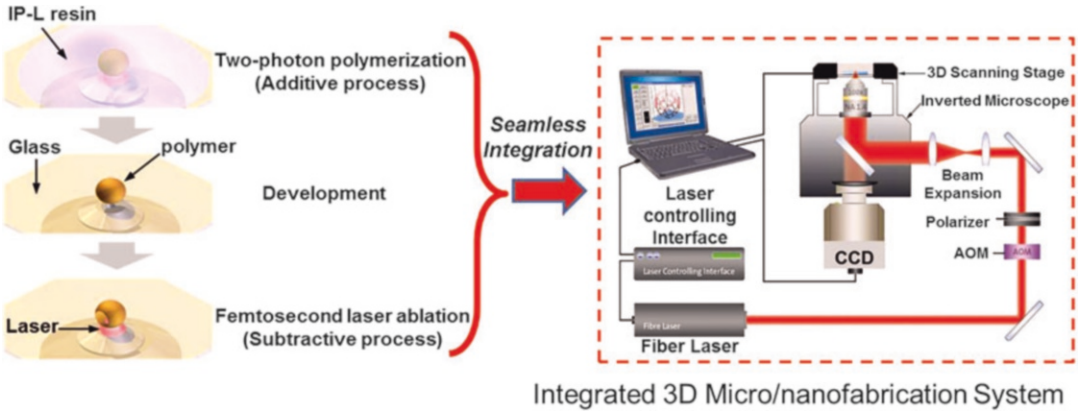
Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 9 SEM images of three-dimensional photonic crystals with different geometries. Woodpile (**a**), spiral (**b**), and pyramid (**c**) arrangements can be directly written by TPP with high spatial control

approximately to the distance between the rods. Thus, TPP is the preferred technique for the fabrication of photonic crystals that operate in the visible and near-IR regions of the spectrum. Furthermore, TPP permits to investigate the effect of different geometries than the log–pile stacks have on the performance of photonic crystals. A variety of three-dimensional photonic crystal structures fabricated using by TPP are shown in Fig. 9.

Comprehensive Micro-/Nanofabrication Combining Two-Photon Polymerization and Multiphoton Ablation

Fabricating advanced devices with complex 3D geometries calls for the use of both additive and subtractive processes. For example, a seamless spherical shell is difficult to be fabricated by either additive or subtractive process. In additive processes, the solid shell of the sphere prevents the dissolving of unexposed photoresist inside the sphere to produce a void. While in subtractive processes, a solid sphere has to be created before subtracting the inner part to make a hollow shell. Therefore, a comprehensive 3D micro-/nanofabrication method possessing both additive and subtractive functionalities is not only desirable but also in demand for advanced device fabrication. By combining TPP with multiphoton-induced laser ablation (MPA) in one experimental setup, a comprehensive 3D micro-/nanofabrication method integrating additive and subtractive processes can be achieved. The integrated fabrication method not only inherits the merits of both TPP and MPA such as writing feature sizes smaller than the diffraction limit and objects with sharp and clean edges, but it also offers the possibility to produce novel device structures which are difficult to be fabricated by either TPP or MPA alone [15].

A schematic design of the integrated 3D micro-/nanofabrication system combining both additive TPP and subtractive MPA processes is shown in Fig. 10. The whole fabrication process consists of three steps. First, TPP is employed to fabricate 3D solid microstructures inside a negative resin. Second, unsolidified resin is washed away by rinsing the sample in a developing solution. Third, a subtractive process based on MPA is



Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 10 Schematic diagram of the comprehensive 3D micro-/nanofabrication combining additive TPP and subtractive MPA processes

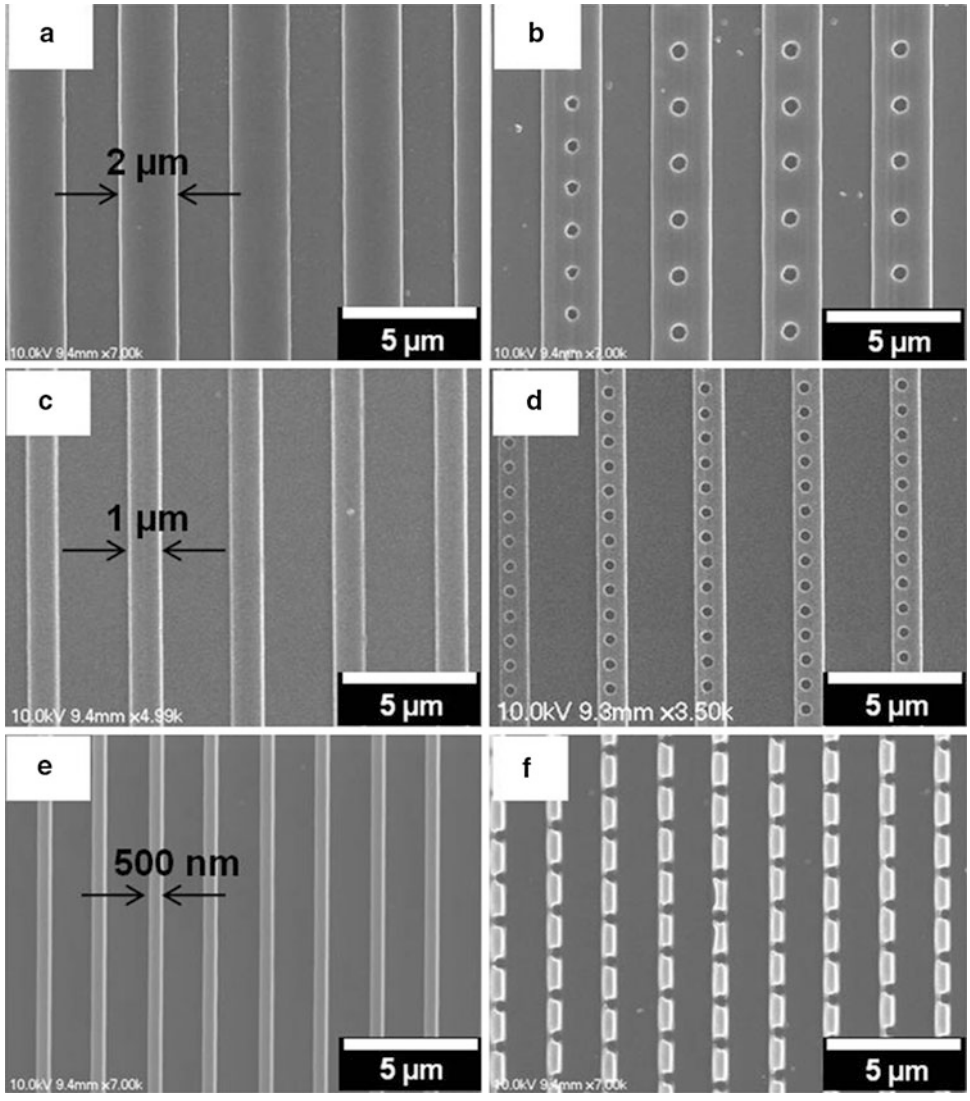
carried out in air to tailor the cured resin to a desired 3D geometry. Depending on the size of the ultimate microstructure, TPP is performed by using laser average powers that vary between 3.5 mW and 8 mW. MPA, instead, requires laser average powers which are two to seven times higher than that used for TPP. To maintain a high degree of spatial confinement when performing both TPP and MPA, a high numerical aperture objective (1.4) is used.

An example of the capabilities of the "TPP + MPA" method in micro-/nanofabrication is shown in Fig. 11. It consists of arrays of polymer fibers with different diameters created by TPP. These dimensions were achieved by controlling the number of laser passes used for the realization of the fibers. The refractive index of polymerized resin is around 1.5. Therefore, these microstructures could be employed as light waveguides for integrated optics. Following TPP, a subtractive MPA process was performed. By removing selectively part of the polymerized resin, periodic patterns of holes with diameters of 500 nm were fabricated along the polymer fibers. The diameter and the periodicity of the holes can be adjusted by using different laser average powers and/or exposure times. The final result of the TPP/MPA process is a complex optical element that can function as a fiber Bragg gratings and that would be difficult to be fabricated by either TPP or MPA alone.

Future Directions for Research

All of the grand goals for nanoscience are dependent upon reliable ways to fabricate nanostructures. Methods that use laser direct writing process to achieve these goals show promising results in the fabrication of various nanomaterials and micro-/nanostructures of various dimensions which are challenging to be fabricated via conventional fabrication methods. An example of laser direct writing process that has achieved remarkable results in the production of complex three-dimensional microstructure is TPP. This technique is truly unique in that it can create three-dimensional microstructures with arbitrary geometries and feature sizes smaller than 100 nm.

Although TPP is almost ready to be implemented in a complimentary fashion to other more conventional microfabrication methods for industrial production, there are still two obstacles to its widespread use that need to be addressed. The first obstacle is speed. Being a serial process, TPP can produce only one object at a time slowing manufacturing progress and increasing its cost. The second obstacle is the limited array of materials available. It is then imperative to focus future research activities to solve these two shortcomings of TPP in order for this novel microfabrication method to be used in the development of advanced devices such as integrated optical circuits and lab-on-a-chip systems.



Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing, Fig. 11 Polymeric microstructures fabricated by combining TPP with MPA. Waveguide of different widths were first created by TPP (**a**, **c**, **e**), and then periodic hole patterns were generated in the waveguides (**b**, **d**, **f**) by MPA

References

1. Arnold, C.B., Piqué, A.: Laser direct-write processing. MRS Bull. **32**, 9–15 (2007)
2. Chrisey, D.B.: The power of direct writing. Science **289**, 879–881 (2000)
3. Cahill, D.G., Yalisove, S.M.: Ultrafast lasers in materials research. MRS Bull. **31**, 594–600 (2006)
4. Gattass, R.R., Mazur, E.: Femtosecond laser micromachining in transparent materials. Nat. Photon. **2**, 219–225 (2008)
5. Moreau, W.M.: Semiconductor Lithography. Principles, Practice, and Materials. Springer, New York (1987)
6. Xia, Y., Rogers, J.A., Paul, K.E., Whitesides, G.M.: Unconventional methods for fabricating and patterning nanostructures. Chem. Rev. **99**, 1823–1848 (1999)
7. del Campo, A., Arzt, E.: Generating Micro- and Nanopatterns on Polymeric Materials. Wiley-VCH, Weinheim (2011)
8. LaFratta, C.N., Fourkas, J.T., Baldacchini, T., Farrer, R.A.: Multiphoton fabrication. Angew. Chem. Int. Ed. **46**, 6238–6258 (2007)

9. Pao, Y.H., Rentzepis, P.M.: Laser-induced production of free radicals in organic compounds. Appl. Phys. Lett. **6**, 93–95 (1965)
10. Maruo, S., Nakamura, O., Kawata, S.: Three-dimensional microfabrication with two-photon-absorbed photopolymerization. Opt. Lett. **22**, 132–134 (1997)
11. Diaspro, A.: Confocal and Two-Photon Microscopy. Wiley-Liss, New York (2002)
12. Thiel, M., Fisher, J., von Freymann, G., Wegener, M.: Direct laser writing of three-dimensional submicron structures using a continuous-wave laser at 532 nm. Appl. Phys. Lett. **97**, 221102–1 (2010)
13. Jiang, L.J., Zhou, Y.S., Xiong, W., Gao, Y., Huang, X., Jiang, L., Baldacchini, T., Silvain, J.F., Lu, Y.F.: Two-photon polymerization: investigation of chemical and mechanical properties of resins using raman microspectrometry. Opt. Lett. **39**, 3034–3037 (2014)
14. Baldacchini, T., LaFratta, C.N., Farrer, R.A., Teich, M. C., Saleh, B.E.A., Naughton, M.J., Fourkas, J.T.: Acrylic-based resin with favorable properties for three-dimensional two-photon polymerization. J. Appl. Phys. **95**, 6072–6076 (2004)
15. Xiong, W., Zhou, Y.S., He, X.N., Gao, Y., Mahjouri-Samani, M., Jiang, L., Baldacchini, T., Lu, Y.F.: Simultaneous additive and subtractive three-dimensional nanofabrication using integrated two-photon polymerization and multiphoton ablation. Light Sci. Appl. **1**, e6 (2012)

Polymeric Surface Modifications

▶ Polymer Coatings

Polymer–Metal Oxide Solar Cells

▶ Hybrid Solar Cells

Polymethyl Methacrylate (PMMA)

▶ SU-8 Photoresist

Polypropylene Carbonate (PPC)

▶ SU-8 Photoresist

Postexposure Bake (PEB)

▶ SU-8 Photoresist

Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Remigio Picone[1], Buzz Baum[2] and Rachel McKendry[3]
[1]Harvard Medical School and Dana-Farber Cancer Institute, Boston, MA, USA
[2]Medical Research Council – Laboratory for Molecular Cell Biology, University College London, London, UK
[3]London Centre for Nanotechnology and Department of Medicine, University College London, London, UK

Synonyms

Bio-adhesion; Biopatterning; Cell adhesion; Cell micro-patterning; Cell patterning; Molecular patterning; Nano-patterning; Plasma micro-contact patterning; Plasma patterning; PmCP; PuCP; PμCP; Surface micro-patterning

Definition

Plasma micro-contact patterning (PμCP) is a simple, efficient, and cost-effective method for the precise patterning of molecules on surfaces. It combines the use of low-pressure plasma with an elastomeric 3D mask to spatially control the removal of molecules, such as proteins, from a surface. The entire PμCP process is subdivided into three main steps: surface pre-coating, plasma micro-patterning, and surface post-coating. Surfaces are first pre-coated with a molecular species and then placed in close contact with the 3D mask. This allows the formation of two distinct regions: an unmasked open region which is accessible to

the plasma, from which the surface layer is removed, and a contact region which is physically protected from exposure to the plasma. In the final step, a second molecule is added to backfill the pattern generated through plasma treatment.

The PμCP technique allows the patterning of virtually any organic molecules on different surface materials and geometries (e.g., flat and curved surfaces and 3D microstructures). Moreover, it is a simple and robust procedure. The main advantages of this approach over traditional micro-contact printing are twofold: the stability of molecule binding to plasma-treated surfaces and the separation of the surface functionalization step from the actual micro-patterning step, which enables the precise control of concentration and uniformity of patterned molecules. In conclusion, PμCP is a simple way to generate surface patterns that are highly reproducible, stable, and uniform, making it a useful method for many applications.

Introduction

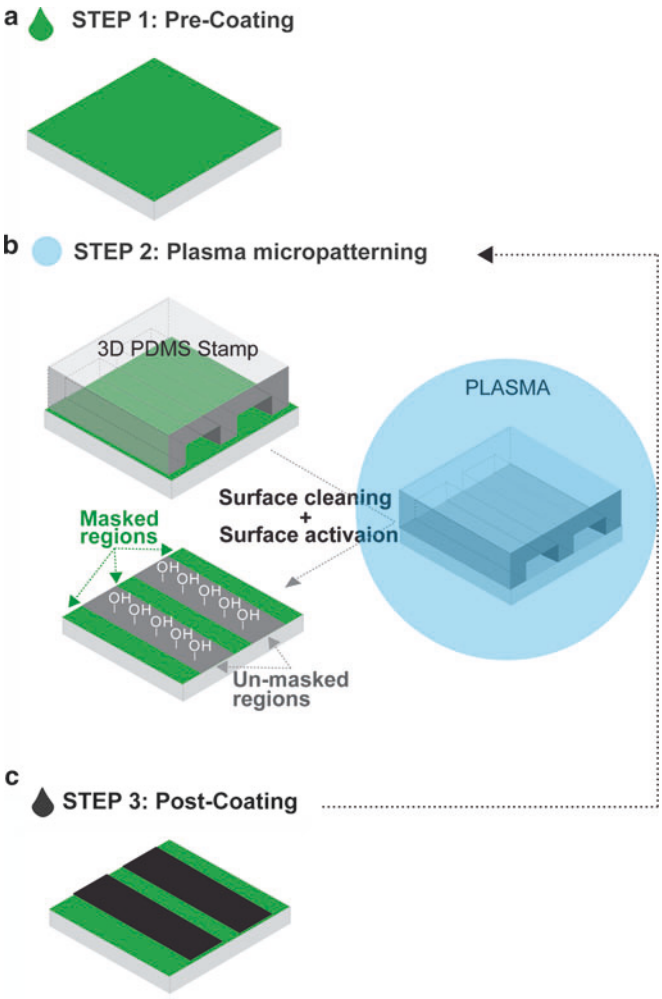
Micro-patterning techniques allow surfaces to be patterned with molecules at small-length scales. The need to fabricate surface patterns emerged with the miniaturization of electrical components in the first integrated circuits. Since then, micropatterning has been applied using a variety of methods [1, 2]. In general, a surface pattern is formed from a foreground pattern that is coated with a specific molecule of interest and a background pattern, generated by backfilling the areas in between foreground features (Fig. 1). The molecules used vary according to the specific patterning application, and the technique can be extended to generate more complex patterns made using >2 types of molecule.

Many biological studies have demonstrated the importance of surface-patterning techniques for the study of fundamental biological processes [3–10]. An established application is to use micro-patterns to confine living cells to specific regions to study the effects of cell shape and spread area on cell behavior, e.g., as tracks for cell movement, and/or to impose a polar form on cells or to generate populations of cells with identical forms that can be used as an aid to quantification by enabling the simple averaging of behavior over large numbers of identically shaped cells [6, 10, 7, 3–5].

To confine the cells to regions with defined geometries, ligands known to bind native cell adhesion molecules are patterned to form foreground regions, while a cell-repellent molecule is used to backfill the background (Fig. 1). When plated onto this type of surface, cells specifically adhere to and spread onto the patterned regions, while retracting extensions that reach onto cell-repellent regions. As a result, this technique can be used to ensure that cells assume and maintain defined geometries imposed by the surface pattern.

The PμCP technique is based on plasma surface cleaning and is a simple, efficient, and cost-effective method for the precise patterning of molecules on surfaces. The PμCP combines the use of a 3D mask with plasma to produce highly reproducible, stable, and uniform molecular patterns. In particular, here is described a simple protocol for biopatterning using this technique, which can be easily adapted to other applications.

The PμCP is a simple procedure. In the first step, surfaces are coated with the molecule of interest (Fig. 1a). This can be done using virtually any surface materials or functionalization technique. Below, a simple protocol based on adsorption of fibronectin (a cell adhesive extracellular matrix protein) onto glass is described. A 3D mask patterned with geometrical features in relief is placed into conformal contact with the pre-coated surface to mask inked surfaces while leaving a 3D free-volume region accessible to plasma. The 3D mask with the glass attached is then plasma treated to remove the molecules from exposed surface regions (Fig. 1b), leaving the patterned region intact underneath the mask. Importantly, the plasma treatment leaves the cleaned surface with a hydrophilic and oxidized monolayer, which facilitates the binding affinity of subsequent coated molecules. In the final step, the mask is removed from the surface and a second molecule added to backfill the pattern generated through plasma cleaning (Fig. 1c). Steps 2 and 3 can then be repeated to form complex



Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique, Fig. 1 PμCP process. (**a**) Step 1: the surface is coated with the molecule of interest. (**b**) Step 2: the 3D mask is put in contact with pre-coated surface to form a free plasma accessible volume. A plasma treatment is then applied to remove the molecules bound to exposed surface regions. Pattern surface regions in contact with 3D mask remain intact. (**c**) Step 3: the second molecule of interest is added to backfill the surface regions previously cleaned by the plasma treatment. Importantly, Steps 2 and 3 can be repeated to make multi-patterns of different shapes and molecules (*black arrow*)

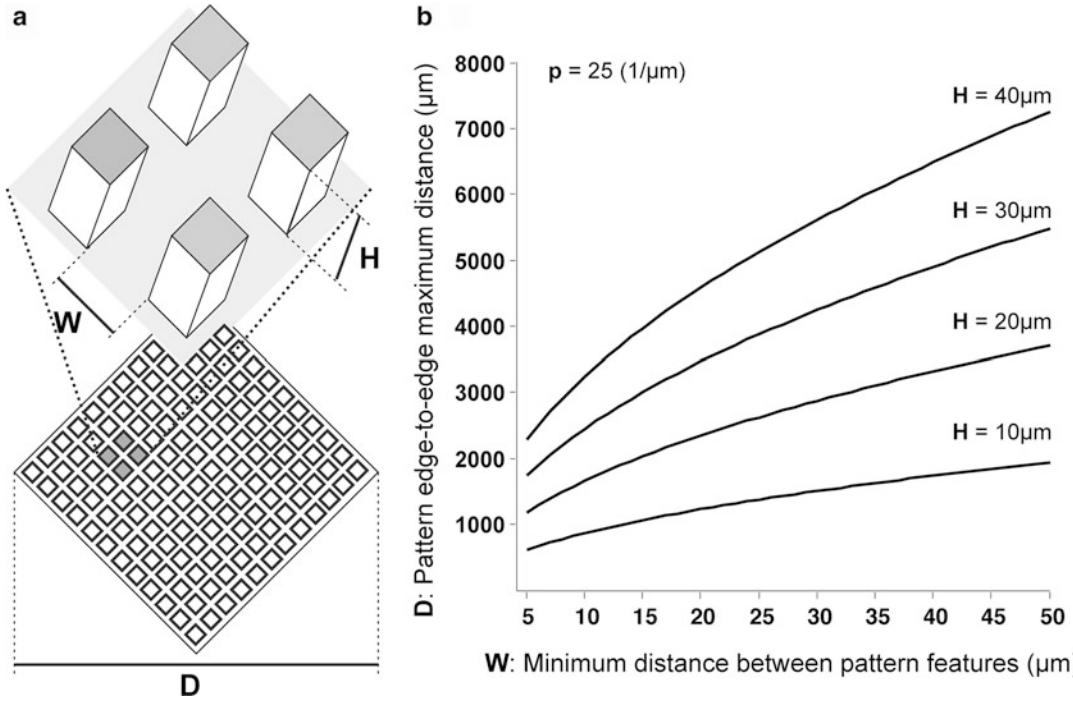
patterns using different molecules on the same surface (Fig. 1b, c).

In subsequent paragraphs the single PμCP steps will be explained in detail, and a simple PμCP protocol for cell patterning will be described.

Methods

3D Mask Design and Fabrication

To produce the 3D mask, polydimethylsiloxane (PDMS, Sylgard 184, Dow Corning) was used. PDMS molds with very high fidelity to the master template (refer to the next paragraph for master fabrication) and its elastic properties allow the mask to make intimate contact with the surface. To fabricate the 3D PDMS mask with geometrical relief features, the prepolymer is mixed with a curing agent, poured onto the master, and cured at 70 °C for 3 h to cross-link the polymer. The standard PDMS to curing agent ratio is of 10:1. A 7:1 ratio for the PμCP technique was preferred. This produces stiffer masks improving feature stability when placed in contact with a surface. This is particularly important for masks with low-aspect-ratio features. Using a higher PDMS to curing agent ratio will make the 3D mask more elastic, a property that can be useful in producing patterns on curved surfaces or 3D microstructures (e.g., plastic tubes for implants, microfluidic channels).



Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique, Fig. 2 (**a**) Diagram showing the critical dimensions for the fabrication of the 3D mask. W: minimum pattern intra-features distance; H: pattern sidewall height; D: maximum pattern width distance. (**b**) A plot showing the relationship between *D*, *H*, and *W* according to the PμCP plasma penetration model described above given a specified range of values

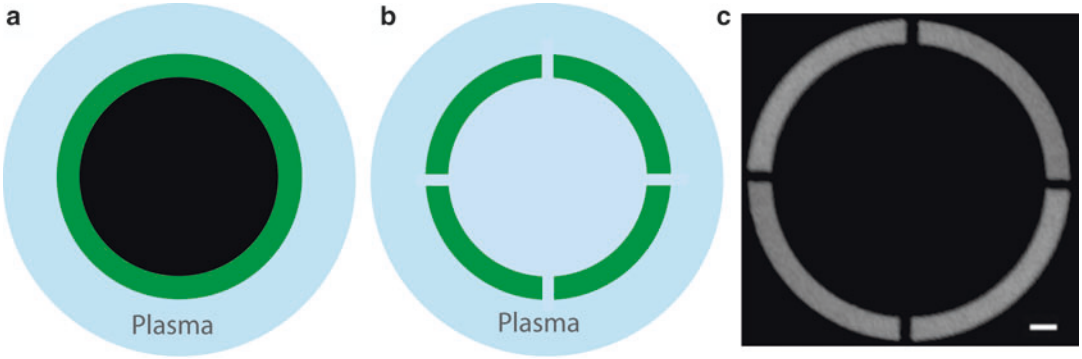
To avoid problems associated with removing the 3D mask from the coated surface, the mask is first passivated using a silanizing agent. To do so, the wafer is put into a vacuum desiccator with 100ul of trimethylchlorosilane (silanizing agent) for 15 min. This work should be performed under a fume hood (refer to the material safety data sheet).

The 3D mask and photomask design for the PμCP should follow two main requirements. First, the free volume formed by the 3D mask in contact with the surface has to be externally accessible from the plasma. Therefore, when the pattern and photomask are designed, it has to be ensured that the geometrical features are not isolated inside a closed perimeter. Second, the plasma penetration is directly proportional to the 3D mask sidewall height (*H* in Fig. 2a) and to the minimum pattern intra-feature distance (*W* in Fig. 2a) and inversely proportional to the maximum mask width (*D* in Fig. 2a) (Fig. 2a). Numerical relationships between *W* and *H* and *D* were calculated by the following simple formula inferred empirically:

$$\frac{D}{H * \sqrt{W}} \leq p$$

where *p* is the "PμCP plasma penetration constant," calculated empirically to give optimal penetration value between 20 and 30 (1/μm).

The plot in Figure 2b shows the relationship between *D*, *H* and *W* according to the PμCP plasma penetration formula shown above for p=25(1/m) and a given range of values for *H*. Low *p* values give high plasma penetration efficiency into the 3D mask. Notably, high *H* value can give rise to feature buckling, and small *H* values and/or high *W* values can make the 3D mask roof collapse onto the surface [11]. Note that this does not take into account the flow of plasma through the device, which may be affected by the geometry of features and their connectedness.



Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique, Fig. 3 (**a**) An example of circular-shaped pattern with an internal hollow circle (*black region*) which cannot be accessed by the plasma. (**b**) Design of a 3D mask with small channels connecting the internal hollow shape with the 3D mask free volume, which can be accessed by the plasma. (**c**) Fibronectin pattern produced by using the PμCP with the 3D mask design shown in B. 3D mask dimensions: $W = 50$ μm, $H = 20$ μm, and $D = 3$ mm. Scale bar 5 μm

Mask stability can be improved by increasing the stiffness of the 3D mask and/or by gluing the top of a thin mask to a flat rigid surface (e.g., glass coverslip). To pattern large surface areas with a single PμCP application, multiple 3D masks on the same surface were placed.

A limitation of the PμCP technique is the difficulty of patterning hollow shapes (Fig. 3a). A solution to this problem can be obtained by designing the 3D mask in such a way that the internal hollow shapes are connected, by a small channel of a few microns, to the external free volume of the 3D mask (Fig. 3b).

Master Fabrication

A master, or template, is used to produce the 3D PDMS mask. A common way to produce a master is by exposing, through a photolithographic mask (photomask), a silicon wafer previously spin coated by a photoresist.

The PμCP technique does not require a special type of photolithographic mask (i.e., photomask). The minimum size of the pattern features determines the least expensive photomask that can be used without problems. The photomask can be designed by using commercial software (e.g., L-Edit, CleWin, AutoCAD, LayoutEditor) that produces GDSII, CIF, or DXF file formats. Once the photomask has been designed, it can be ordered from one of several photomask producers (e.g., JD Photo-Tools, Advance Reproductions). For the design requirements of the pattern shapes, sizes, and arrangement, refer to the above 3D mask fabrication paragraph. To make the process dust-free, all procedural steps are performed in a clean room.

Photoresist producers supply (e.g., MicroChem) clear processing guidelines for master fabrication. To enhance the release of the PDMS mold from the fabricated master, the wafer was passivated by a silanizing agent by putting the wafer into a vacuum desiccator with 100ul of trimethylchlorosilane (silanizing agent) for 15 min (work under a fume hood and refer to the material safety data sheet).

Surface Pre-coating

One of the main advantages of the plasma micropatterning over the standard micro-contact printing techniques is the possibility of separating the surface characterization from the actual surface-patterning step. In particular, by using PμCP, the molecules of interest can be applied to the surface before or after plasma treatment. This allows the concentration and uniformity of the molecules used to form the pattern to be precisely controlled (refer to the section "Precise Control of the Surface Patterning" for a detailed discussion on the precise control of the surface micro-patterning by using the PμCP). Here, a simple surface-coating

protocol based on adsorption of fibronectin onto a glass surface is described. Fibronectin adsorbs strongly and uniformly onto glass surfaces without compromising its ability to bind cells through integrin-mediated adhesion.

Materials
- Glass coverslips (Fisher Scientific) or glass bottom dishes/plates (MatTek)
- Ethanol 96 % (Sigma)
- Deionized H_2O
- Fibronectin (Sigma) or Fibronectin Alexa Fluor 488 nm (Invitrogen)

Equipment
- Laminar flow hood (NuAire)
- Nitrogen gas cylinder
- Sonicator (facultative, Branson, 2510)
- Vacuum storage system (facultative, custom produced)

Procedure
1. Glass surfaces are cleaned by bathing in 75 % ethanol, rinsed in deionized H_2O, and blow-dried with nitrogen.
   Optionally, glass cleaning can be improved by sonicating in 75 % EtOH for 15 min and rinsing with deionized H_2O. To keep the surfaces sterile, work under a laminar flow hood. If necessary, surfaces can also be cleaned using plasma or acid/alkali treatments.
2. Dried glass surfaces are immersed completely in 10–20 μg/ml fibronectin in deionized H_20 for 20 min. To generate a homogeneous surface coating, the entire glass surface must be immersed in the fibronectin solution. The fibronectin concentration used depends on the specific application.
3. Coated glass surfaces are then washed by deionized H2O and gently blow-dried with nitrogen.

Plasma Micro-patterning (Step 2)

Once surfaces are coated with the molecule of interest (in this protocol fibronectin), they are then treated with plasma to create the surface pattern. To do so, a 3D mask with the pattern features in relief is placed onto the pre-coated surface to create a free volume within the channels of the mask in which the plasma can penetrate to remove surface molecules. The 3D mask contact regions protected from the plasma will be left intact to form surface patterns. For this purpose, polydimethylsiloxane (PDMS) is an ideal material for the 3D mask because its elastic properties allow good contact and simple removal from the surface (see section "3D Mask Design and Fabrication" for details about 3D mask design requirements and fabrication).

For the plasma micro-patterning, oxygen plasma generated in a commercial plasma cleaner machine by electrically exciting an oxygen gas at low pressure (0.1–2 Torr) in a vacuum chamber was used. The plasma so generated has low temperature, positively and negatively charged molecules along with photons in the shortwave ultraviolet (i.e., vacuum UV or VUV) range and other reactive molecules. The surface removal is carried out by two main actions: (i) the breaking of organic bonds (i.e., C–H, C–C, C = C, C–O, and C–N) by the UV and (ii) the reaction of the plasma oxygen species (O_2^+, O_2^-, O_3, O, O^+, O^-, ionized ozone, metastable excited oxygen, and free electrons) with organic molecules on the surface to form H_2O, CO, CO_2, and lower molecular weight hydrocarbons, which are subsequently evacuated from the chamber. Thus, the plasma surface treatment removes the exposed surface molecules and leaves the bare surface with a hydrophilic oxidized monolayer, which facilitates the binding of subsequent molecules (Fig. 1b).

Materials
- Coated glass surfaces obtained from "surface pre-coating" step (see section "3D Mask Design and Fabrication")
- Deionized H_2O
- 75 % ethanol (in deionized H_2O)
- 3D PDMS mask (see section "3D Mask Design and Fabrication" for the 3D mask fabrication)

Equipment
- Plasma cleaner. It used the Femto plasma cleaner from Diener Electronic, Germany. This model of plasma cleaner allows the

manual control of the gas pressure, plasma time, power, and number of cycles.
– Sonicator (optional, Branson, 2510).

Procedure

1. The 3D mask is washed by sonicating in 75 % ethanol for 15 min and then rinsed with deionized water and blow-dried by nitrogen. It is important to ensure that the 3D mask and glass surfaces are dry for the next step. No washing is needed when using a newly fabricated 3D mask.
2. The clean, dry 3D mask is gently placed on top of the coated glass surface. A gentle pressure is applied with tweezers to allow even adhesion of the 3D mask to the surface. Adhesion can be monitored by close inspection.
3. The 3D mask with the glass surface attached is then placed into the plasma cleaner. The vacuum is switched on and three cycles of plasma treatments of 10 s each at 50 W power are performed. The parameters for plasma surface removal can vary between devices. Before PμCP routine use, a series of tests on the plasma cleaner are needed to find the optimal parameters (i.e., plasma time, power, and cycles). As a rule of thumb, the plasma time, power, and number of cycles need to be minimized. High plasma exposure could microblast the surface and modify the bulk material of the surface and the 3D mask.
4. The 3D mask is then gently removed from the glass surface and made ready for the post-coating step (see section "Surface Post-coating (Step 3)"). The 3D mask can be used straightaway for a new PμCP cycle (Fig. 1b, c). If needed, another cycle of PμCP can be carried out on the same patterned surfaces to make complex patterns with additional surface molecules. After lifting off the mask, inspect the patterned surface under a microscope to ensure that no PDMS residue remains on the surface. If PDMS residue is observed, then reduce plasma processing time, power, and/or number of cycles. If the problem persists, then use a new 3D mask.

Surface Post-coating (Step 3)

After the plasma micro-patterning step, a second molecule is added to coat the surface region removed by the plasma (pattern background). PLL-g-PEG was used as pattern background molecule as it is a cell-repellent molecule, which has been successfully used for cell micropatterning [12].

Materials

– Plasma micro-patterned glass surfaces obtained with the above protocols.
– PLL-g-PEG (0.1 mg/ml in HEPES). For the stock solution preparation, refer to the instructions from the producer (SuSoS, Switzerland, PLL(20)-g[3.5]-PEG(2)).
– Nitrogen cylinder with filtered spray gun (Innotech).

Protocol

1. Glass plasma micro-patterned surfaces are incubated in 0.1 mg/ml PLL-g-PEG solution at room temperature for 1 h.
2. Glass surfaces are rinsed in deionized H_2O and gently dried using nitrogen gas.

The patterned surfaces are now ready to be used. For best results, the patterned surfaces should be used the same day of the preparation. Optionally, patterned surfaces can be stored for months under vacuum and nitrogen or argon gasses at 4C. A low-cost and very simple design for a storage nitrogen/argon vacuum system can be requested from the authors.

Cell Patterning

Here a protocol for patterning adherent cells on PμCP-patterned surfaces is described. The PμCP protocol has been successfully used to pattern both adherent (e.g., RPE1, HeLa, MFC7A MFC10A, MDCK, HTC-116, A431, S2R+ cells) and nonadherent cells (e.g., *Drosophila* S2 cells and human T cells). In the case of nonadherent cells, patterning is achieved using 20 μg/ml concanavalin A (Sigma, C2272) on glass by the same protocol described above. Concanavalin A is a tetravalent lectin protein that binds a variety of glycosylated cell surface proteins and lipids.

Materials

- Cell dish/flask
- Trypsin (Gibco, the concentration depends on the cell type)
- 1× PBS (Gibco)
- Cell culture medium (the type and recipe of the culture medium depend on the cell type being used)

Equipment

- Laminar flow hood (NuAire, TYPF A/B)
- Cell incubator (Fisher Scientific)
- Microscope (Nikon Diaphot inverted)
- Cell counter (Beckman) or hemacytometer
- Centrifuge (Thermo Scientific)

Protocol

1. Cell are washed with PBS, then detached from their culture surface using trypsin and resuspended in fresh culture medium.
2. Cells in suspension are centrifuged for 4 min at 1,200 rpm and resuspended in fresh culture medium. To detach cell clumps, gently pipette up and down the resuspended cells.
3. Cell numbers are determined and fresh medium is added to reach the desired density. Cell density for patterning needs to be determined empirically as it depends on the number and dimension of pattern features, overall surface area, and specific application.
4. Cells are plated onto the patterned surfaces and incubated to allow cell attachment and spreading.
5. After about 10 min, cells are examined under the microscope to confirm adhesion and confinement to the pattern. Attachment time can vary between cell types and cellular adhesive patterned molecule. In general cells attach and confine onto the pattern in 30–50 min after plating.
6. Nonattached cells are removed by gently exchanging the culture medium. Medium exchanges can be repeated to obtain the desired number of cells on the pattern. Ensure that the surface is left always wet to avoid cells dying during the culture medium change
7. Cells are incubated for at least 1 h to allow for complete cell spreading. Significantly, cells stably bind to plasma-patterned substrates for several days, without their destroying the pattern. This is not the case for micro-patterns generated by simple PDMS stamping (Fig. 7).

Results

Precise Control of the Surface Patterning

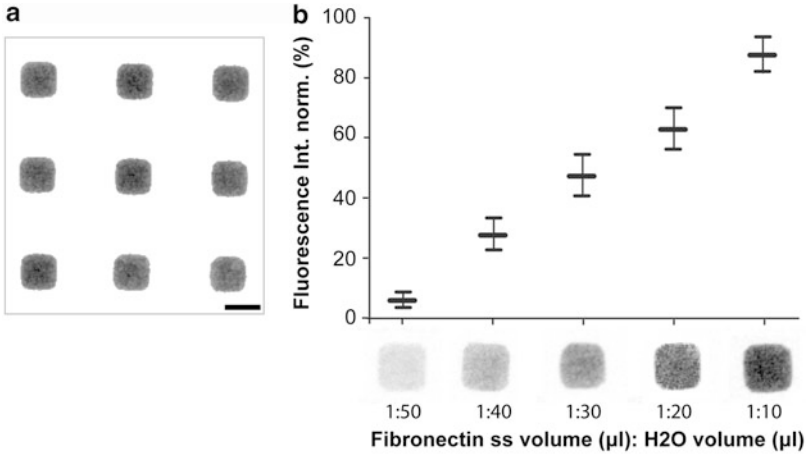
An important advantage of PμCP is the ease of controlling the concentration and uniformity of the patterned surfaces. To show that this is the case, the method described above was used to coat glass with five different concentrations of fluorescently conjugated fibronectin protein (Fig. 4) prior to PμCP. The results of this test showed that the patterns generated using PμCP are reproducible, having a maximum average variation of about 10 % across a small area of patterned substrate, and that concentration titration of the pre-coated fibronectin on glass results in a reproducible linear increase in surface fibronectin density as measured by fluorescence intensity (Fig. 4).

PμCP-Inverted Techniques: Comparison with the Standard PμCP

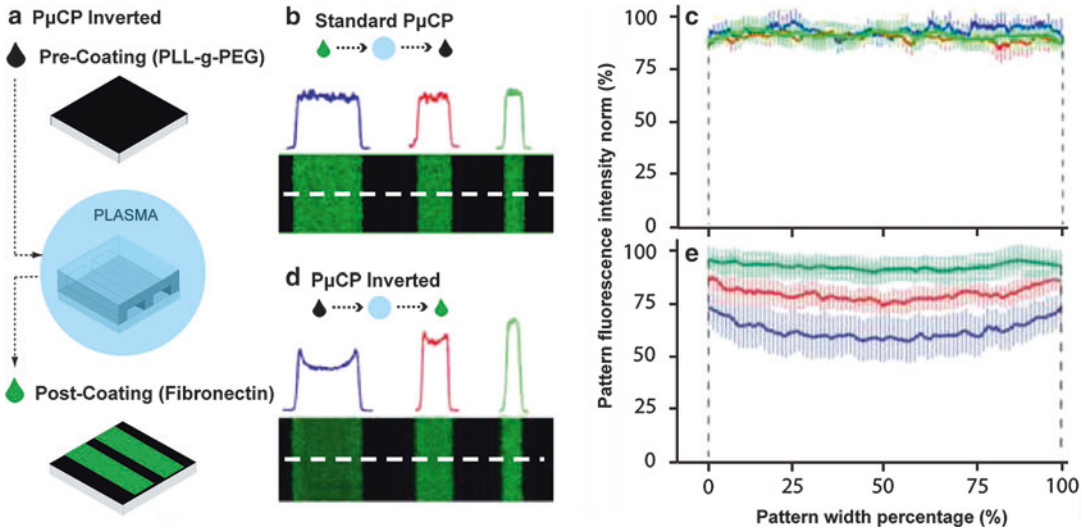
A variation to the PμCP procedure can be implemented by inverting the standard PμCP protocol above described to first coat the background layer (e.g., with PLL-g-PEG) before plasma treatment (micro-patterning step) and then application of the foreground (e.g., fibronectin) (Fig. 5a). This inverted PμCP technique (PμCP-inv) can only be applied to pattern features that form 3D mask with a 3D open space accessible to the plasma.

Interestingly, when fibronectin lines micropatterned by the standard PμCP were compared with the ones obtained by using the PμCP-inv technique, an increased concentration of molecules at the edges of the pattern produced by using the PμCP-inv technique (Fig. 5d, e) was observed. Furthermore, this pattern "edge effect" was dependent on the pattern size (i.e., line width) (Fig. 5d, e).

This edge effect was reproducible and seems likely to be due to differences in the diffusion of fibronectin over the micro-patterned PLL-g-PEG



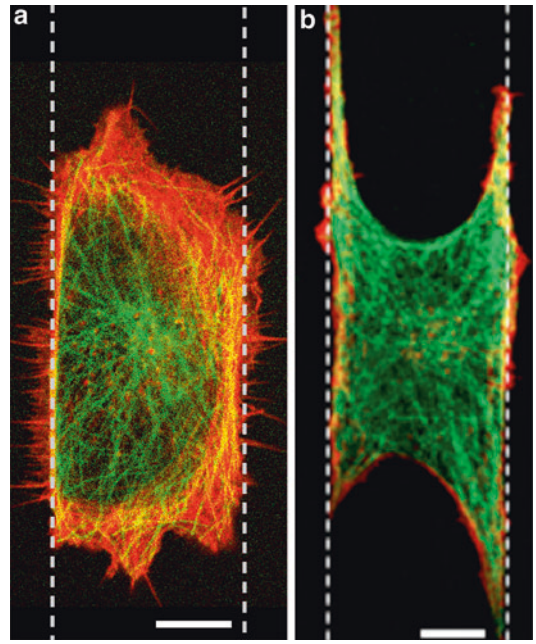
Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique, Fig. 4 Precise control of pattern concentration and uniformity using PμCP. (**a**) 10 × 10 μm fluorescently labeled fibronectin square-shaped patterns on glass produced by the PμCP technique. (**b**) A plot to show the fluorescence intensities of square-shaped fibronectin patterns generated using five different concentrations of fibronectin (ss = stock solution of 1 mg/ml) via the protocol described above. To compare images across acquisitions, fibronectin fluorescence images were corrected and the fluorescence intensity values expressed as percentage of the minimum and maximum intensity values. 3D mask dimensions: $W =$ 17.5 μm, $H = 20$ μm, and $D = 2$ mm. Scale bar 10 μm



Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique, Fig. 5 (**a**) PμCP-inverted (PμCP-inv) protocol. Images show fibronectin lines (*green*) of different widths (*thick*, *middle, and thin*) micro-patterned using the standard PμCP (**b**) or PμCP-inv (**d**), together with the corresponding normalized fluorescence intensity diagrams (direction indicated by the *dotted white line*). The images were acquired by confocal microscopy. PLL-g-PEG was used as background molecule (*black* areas in **b** and **d**). Graphs (**b** and **e**) show comparison between fluorescence intensities of the three fibronectin line width classes: thick (23–32 μm, in *blue*, $N = 13$), middle (13–22 μm, in *red*, $N = 11$), and thin (3–12 μm, in *green*, $N = 12$) micro-patterned using PμCP (**b**) or PμCP-inv (**c**) techniques. To compare fluorescence intensities of the different pattern line widths, the intensities were plotted against normalized line width. Graph bars indicate standard deviations. 3D mask dimensions: $W =$ 50 μm, $H = 20$ μm, and $D = 3$ mm

surface. Notably, this edge effect is not visible in the patterns generated by using the standard PμCP technique because the fibronectin coating is applied uniformly prior to plasma treatment (Fig. 5b, c). Nevertheless, the standard PμCP technique could in principle produce increased PLL-g-PEG concentration at the pattern edges. This does not represent a physical barrier as cells growing on thin fibronectin lines visibly extend and retract lamellipodium outside of the pattern onto the PEG. However, a more concentrated PLL-g-PEG rim all around the patterns could enhance the micro-patterning cell confinement.

Notably, when cells on the patterns were plated with increased concentration at the edges, it was found that cells show a nonuniform spreading. Cells preferentially spread along pattern edges forming a characteristic concave border (Fig. 6b). Importantly, this suggests that the correct control of the molecule pattern concentration and uniformity is crucial to avoid misinterpretation of the phenomena under observation.



Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique, Fig. 6 HeLa cells on fibronectin lines micro-patterned by using PμCP (**a**) or PμCP-inv (**b**) technique (f-actin in *red* and microtubules in *green*). In A the fibronectin pattern was produced by the standard PμCP protocol described above. In B to produce the edge effect, the fibronectins were coated at low concentrations (5 μg/ml) and short seeding time (15 min). 3D mask dimensions: $W = 50$ μm, $H = 20$ μm, and $D = 3$ mm. Scale bar 10 μm

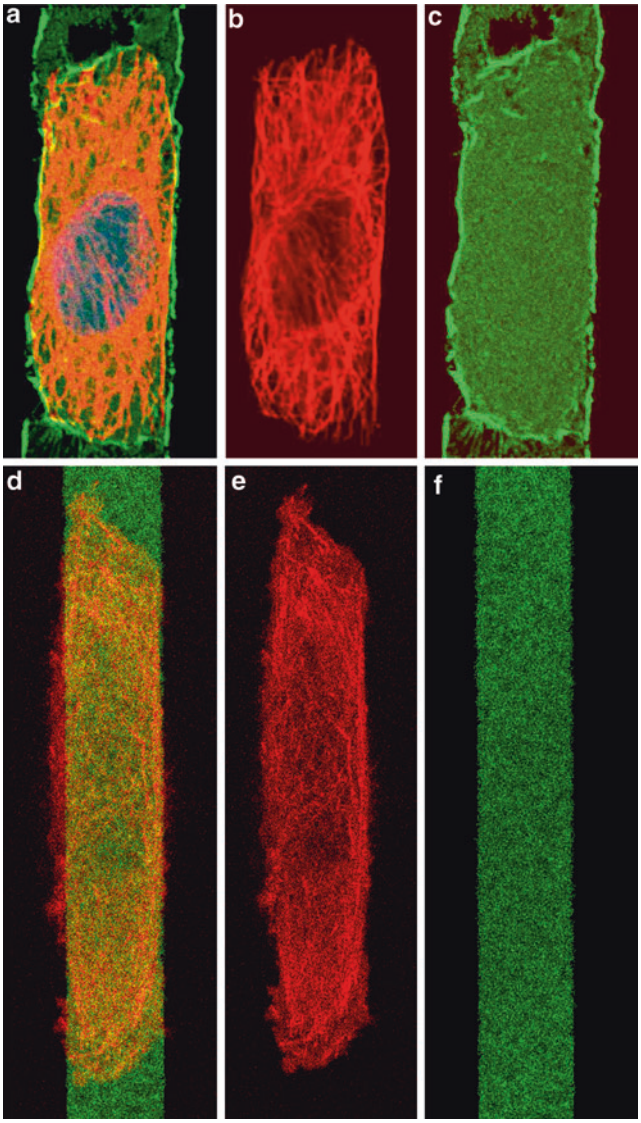
Discussion

The PμCP technique is simple to implement and allows the precise control of surface characterization and micro-patterning steps. Furthermore, it allows the patterning of organic molecules on different surface materials and at different conformations (e.g., flat and curved surfaces and 3D microstructures). This makes the PμCP particularly suitable for biomedical applications but can be also easily adapted to other types of applications and different fields.

In the last 10 years, there has been a proliferation of new methods for micro-patterning, but two methods are mostly used in biology laboratories: the soft-lithography-based [1], in particular μCP, and UV-based techniques [13].

In comparison to the μCP, the PμCP has the advantage of having a better control over individual patterning steps. It aids the generation of reproducible and uniform patterns, which can be difficult to achieve using simple μCP [14]. Moreover, in our hands, μCP-patterned surfaces are not as resistant to contractile forces exerted by cells over long periods as those generated using PμCP (Fig. 7) [14]. The use of gold self-assembly on classical substrate for cell biology (e.g., plastic or glass) can be used to overcome in part to this problem [15], but with the drawback of adding complexity to the overall micro-patterning process. Nevertheless, the difficulty of controlling the loading and successive unloading and transfer of the molecules from the stamp to the surface still remains [16].

Using PμCP, it is possible to overcome these limitations by using the elastomeric stamp as a mask (3D mask) for the plasma micro-patterning and by separating the surface functionalization steps from the micro-patterning step. Furthermore, the use of the elastomeric stamp as 3D mask allows the micro-patterning of molecules on flat and non-flat substrates (e.g., 3D



Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique, Fig. 7 After 12h, HeLa cells tend to remove the line-shaped fibronectin pattern (A and C, in green) generated using the standard CP technique. This was not the case for the line-shaped fibronectin pattern generated with the PCP (D and F, in green), which remained stable. Cell microtubules are shown (B and E, in red)

microstructures, microfluidic channels, or tubing for biomedical applications). In addition, the elastomeric 3D mask can be cut and resized to allow the micro-patterning of surfaces inside chambers, dishes, and multi-well plates.

Potential problems associated with PμCP are the need of photolithography tools and skills for the master fabrication and plasma cleaner. Although these are not traditionally skills and tools available in a biological laboratory, the procedure for the master fabrication is simple, well established, and requires only about 2 h to be accomplished; and only a single master needs to be fabricated as many can be replicated from it by a simple epoxy replica molding process [17]. Alternatively, the fabrication of the 3D mask can be inexpensively purchased (e.g., BioTray). A plasma cleaner machine can be also purchased at a reasonable cost from different producers (e.g., Harrick Plasma, Plasma Etch, Glow Research, Diener Electronic). Alternatively, low-cost oxygen plasma cleaner machines can be built or purchased unassembled (e.g., Epsilon Micro Devices).

Another micro-patterning technique which is procedurally similar to the PμCP (in particular to

the PμCP-inv) which also offers good control on the micro-patterning process is the UV-based techniques [13]. In this case, deep UV light and photomask with transparent micrometer pattern features is used to locally irradiate a pre-coated PEG-based cell-repellent molecule monolayer. The pattern is then obtained by backfilling with a molecule of interest (e.g., fibronectin) the UV-irradiated areas. Unlike PμCP and μCP, the UV-based micro-patterning technique has the advantage of not requiring the master or stamp/3D mask fabrication. However, it requires the use of specialized equipment (i.e., UV ozone oven) and of an expensive photomask allowing the penetration of the deep UV light.

The PμCP- and UV-based techniques have in common the separation of the surface functionalization from the micro-patterning step [13]. However, the UV-based technique employs the inverse PμCP process (PμCP-inv is described in section "PμCP-Inverted Techniques: Comparison with the Standard PμCP"). The cell-repellent molecule is coated before the micro-patterning step (UV illumination), and the cell adhesive protein (e.g., fibronectin) is coated after it.

UV-based techniques give good control on the surface passivation but have the potential drawback of generating "edge effects" described above when low concentration or/and short incubation times of the post-coating molecule are applied. Consequently, good molecule uniformity can be only obtained by saturating patterned regions with a specific molecule at high concentration and/or long incubation times, compromising the control of pattern molecule concentration and uniformity. The PμCP it is not subject to this problem as the coating of the patterning molecule is applied to the non-confined surface before the plasma treatment.

A secondary advantage of the plasma treatment is the formation of an oxidized surface layer aiding subsequent surface chemistries and the binding of the plasma-cleaned micro-patterned surface to microfluidic devices. The PμCP also induces very efficient removal of the unmasked surface molecules by using a low-temperature plasma oxygen, which forms a "cloud" of reactive oxygen species that breaks organic bonds, cleans the surface, and evacuates residues, something that UV alone does not mimic. Note that short plasma exposure times and low power must be applied to prevent the surface bulk material being damaged. Moreover, a specific 3D mask design must be implemented to allow the efficient plasma penetration into the mask channels (see section "3D Mask Design and Fabrication").

One of the major limitations of micro-patterning technique is the difficulty of patterning at the nanometer scale. This limitation is mainly due to the use of light for the micro-patterning process. The plasma micro-patterning removal is mainly dependent on the penetration of atomic reactive species into the confined 3D mask open volume. Thus, PμCP has the potential to be used for nano-patterning, but further studies are needed to investigate the plasma penetration and efficiency in nano-compartments.

In conclusion, the PμCP technique is highly reproducible, is easy and fast to apply, allows the precise control of the patterned molecule concentration and uniformity, and has the potential to be applied to patterning molecules at nanometer scale.

Cross-References

- ▶ Adhesion
- ▶ Bioadhesion
- ▶ Bioadhesives
- ▶ Molecular Patterning
- ▶ Molecular Printing
- ▶ Nanopatterned Substrata
- ▶ Nano-patterning
- ▶ Patterned Hydrogels
- ▶ Patterning
- ▶ Plasma
- ▶ Plasma Etching
- ▶ Plasma Polymerization
- ▶ Single Cell Analysis

References

1. Whitesides, G.M., Ostuni, E., Takayama, S.: Soft lithography in biology and biochemistry, Annu. Rev. Biomed. Eng. 3, 335–373 (2001)

2. Théry, M.: Micropatterning as a tool to decipher cell morphogenesis and functions. J. Cell Sci. **123**(24), 4201–4213 (2010)
3. Godinho, S.A., Picone, R, Burute, M., Dagher, R., Su, Y., Leung, C. T., Polyak, K., Brugge, J. S., Théry, M., and Pellman, D.: Oncogene-like induction of cellular invasion from centrosome amplification. Nature, doi:10.1038/nature13277 (2014)
4. Lancaster, O. M., Le Berre, M., Dimitracopoulos, A., Bonazzi, D., Zlotek-Zlotkiewicz, E., Picone, R., Duke, T., Piel, M., and Baum, B.: Mitotic Rounding Alters Cell Geometry to Ensure Efficient Bipolar Spindle Formation. Dev. Cell, **25**, 1–14 (2013)
5. Elliott, P.R., Irvine, A.F., Jung, H.S., Tozawa, K., Pastok, M.W., Picone, R., Badyal, S.K., Basran, J., Rudland, P.S., Barraclough, R., Lian, L.-Y., Bagshaw, C.R., Kriajevska, M., Barsukov, I.L.: Asymmetric mode of Ca^{2+}-S100A4 interaction with nonmuscle myosin IIA generates nanomolar affinity required for filament remodeling. Structure **20**(4), 654–666 (2012)
6. Théry, M., Racine, V., Pépin, A., Piel, M., Chen, Y., Sibarita, J.-B., Bornens, M.: The extracellular matrix guides the orientation of the cell division axis. Nat. Cell Biol. **7**(10), 947–953 (2005)
7. Picone, R., Ren, X., Ivanovitch, K.D., Clarke, J.D.W., McKendry, R.A., Baum, B.: A polarised population of dynamic microtubules mediates homeostatic length control in animal cells. PLoS Biol. **8**(11), e1000542 (2010)
8. Lafaurie-Janvore, J., Maiuri, P., Wang, I., Pinot, M., Manneville, J.-B., Betz, T., Balland, M., Piel, M.: ESCRT-III assembly and cytokinetic abscission are induced by tension release in the intercellular bridge. Science **339**(6127), 1625–1629 (2013)
9. Godinho, S.A., Kwon, M., Pellman, D.: Centrosomes and cancer: how cancer cells divide with too many centrosomes. Cancer Metastasis Rev. **28**(1), 85–98 (2009)
10. Chen, C.S., Mrksich, M., Huang, S., Whitesides, G. M., Ingber, D.E.: Geometric control of cell life and death. Science **276**(5317), 1425–1428 (1997)
11. Delamarche, E., Schmid, H., Michel, B., Biebuyck, H.: Stability of molded polydimethylsiloxane microstructures. Adv. Mater. Weinheim **9**(9), 741–746 (1997)
12. Csucs, G., Michel, R., Lussi, J.W., Textor, M., Danuser, G.: Microcontact printing of novel co-polymers in combination with proteins for cell-biological applications. Biomaterials **24**(10), 1713–1720 (2003)
13. Azioune, A., Storch, M., Bornens, M., Théry, M., Piel, M.: Simple and rapid process for single cell micropatterning. Lab Chip **9**(11), 1640–1642 (2009)
14. Fink, J., Théry, M., Azioune, A., Dupont, R., Chatelain, F., Bornens, M., Piel, M.: Comparative study and improvement of current cell micro-patterning techniques. Lab Chip **7**(6), 672–680 (2007)
15. Ostuni, E., Whitesides, G.M., Ingber, D.E., Chen, C. S.: Using self-assembled monolayers to pattern ECM proteins and cells on substrates. Methods Mol. Biol. **522**, 183–194 (2009)
16. Tan, J.L., Liu, W., Nelson, C.M., Raghavan, S., Chen, C.S.: Simple approach to micropattern cells on common culture substrates by tuning substrate wettability. Tissue Eng. **10**(5), 865–872 (2004)
17. Koerner, T., Brown, L., Xie, R., and Oleschuk, R. D.: Chemical - Epoxy resins as stamps for hot embossing of microstructures and microfluidic channels. Sensors and Actuators B **107**, 632–639 (2005)

Pregnancy

► Use of Nanotechnology in Pregnancy

Prenucleation Clusters

Denis Gebauer
Department of Chemistry, Physical Chemistry, University of Konstanz, Konstanz, Germany

Synonyms

DOLLOP (Dynamically Ordered Liquid-Like Oxyanion Polymer); Stable clusters; Stable pre-critical clusters

Definition

Prenucleation clusters are thermodynamically stable associates of atoms, ions, or molecules forming in solution, approximately 1–3 nm in size on average. They are solutes themselves and form based upon dynamic chemical equilibrium in undersaturated as well as supersaturated (metastable) solution states, playing a key role during nucleation of a second phase from the latter.

Overview

Relevance

How crystallization may be controlled is an important question, since many processes and applications depend on the control of

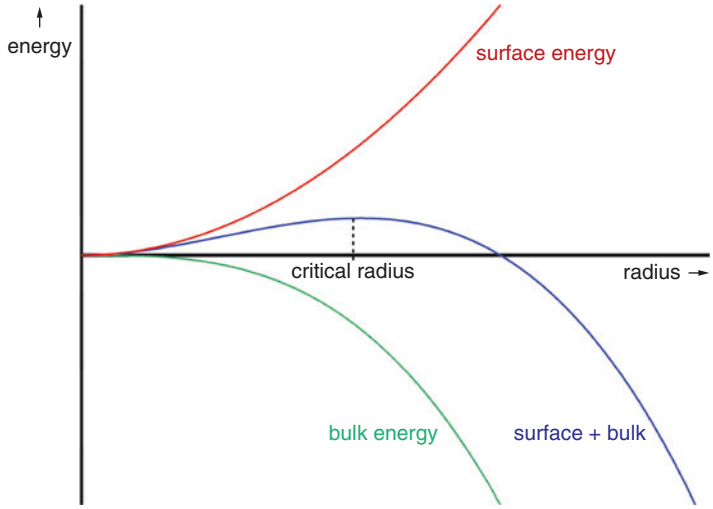
crystal polymorphism, morphology, size, and other crystal characteristics. However, today, the design of bottom-up approaches to tailor-made crystalline materials is still widely settled in the realm of trial and error. In order to improve these empirical approaches, (nano)technology depends on a molecular understanding of the underlying processes [1]. A look at living nature shows that organisms are able to control crystallization in sophisticated ways, and form mineralized structures – biominerals like bone, teeth, or nacre – that show outstanding material properties that are tailor made with respect to function through advanced micro- and nanostructures [2]. Many studies have indicated that biomineralization processes can proceed via so-called nonclassical, that is, particle-mediated pathways, which may also involve liquid and/or amorphous intermediates [3]. In contrast, classical theories presented in most crystallization textbooks only consider ion-by-ion, atom-by-atom, or molecule-by-molecule reaction channels for nucleation and growth [4].

The notion of stable prenucleation clusters can be regarded as an important expansion of nonclassical concepts toward the most fundamental step in crystallization: nucleation [5]. The term "nucleation" refers to the onset of phase separation in a system that has become supersaturated. The existence of stable clusters is not taken into account within classical nucleation theory, which is the framework that is used to access nucleation phenomena in the majority of cases. Hence, an alternative reaction channel referred to as the "prenucleation cluster pathway" has been proposed [5], which includes a nanoscopic liquid-liquid separation as central event. In the proposed mechanism, stable prenucleation clusters play a key role as direct molecular solute precursors of phase-separated nanodroplets. There are strong indications that this pathway is relevant for many materials forming from aqueous solution, including the most important biominerals calcium carbonate and phosphate, as well as silica, iron oxides, calcium oxalate, but also small organic molecules like amino acids, urea, or citrate [5, 6].

Background: Classical Nucleation Theory

Nucleation means the onset of a phase transition, that is, the irreversible formation of the first nuclei of a nascent phase in a system that has become supersaturated. The work of Becker and Döring is commonly known as classical nucleation theory (CNT), which was originally derived for the formation of nuclei from supersaturated water vapors [7, 8]. The fundamental assumption of CNT is that the bulk energy of a forming nucleus drives nucleation, while the generation of a phase interface, and with it of interfacial surface, is the antagonist and impedes nucleus growth. The bulk energy (favorable, negative sign) scales with the cube of the radius of the nuclei, and their surface energy (unfavorable, positive sign) increases with the square of the radius. Hence, the bulk energy begins to balance the energetic costs due to the generation of a phase interface at a certain point (Fig. 1). A nucleus of this size is the critical nucleus. The critical size corresponds to a metastable state of equilibrium, and even infinitesimal changes toward smaller or larger sizes will lead to nucleus dissolution or growth without limit, respectively. The change in free enthalpy for the formation of (pre) critical nuclei is positive, hence formation of these species is thermodynamically improbable and can only occur through fluctuations on microscopic scales.

CNT makes several assumptions, among which the so-called capillary assumption is the most debatable. The oversimplification was realized already by Gibbs when he first formulated it: Nanoscopic nuclei of a nascent phase are assumed to behave as if they were macroscopic. There have been several empirical and semiempirical treatments of CNT that aim at the improvement of quantitative predictions of nucleation rates, which can deviate from measurements by many orders of magnitude [9]. Under purely qualitative considerations, though, CNT is successful in describing nucleation phenomena in principle. The hallmark of nucleation is the presence of a barrier, which scales with supersaturation (and temperature) and prevents nucleation from proceeding spontaneously, at least close to the saturation limit. CNT assumes that a nucleus of critical size is fundamental to the existence of this thermodynamic barrier.



Prenucleation Clusters, Fig. 1 Classical representation of the energy of a nucleus (*blue line*) in dependence of its radius in a supersaturated system. The surface energy gives a positive (unfavorable, *red line*) and the bulk energy a negative energetic contribution (*green line*), which are proportional to the square and the cube of the radius, respectively. Hence, at the critical size, the bulk term begins to balance the surface term

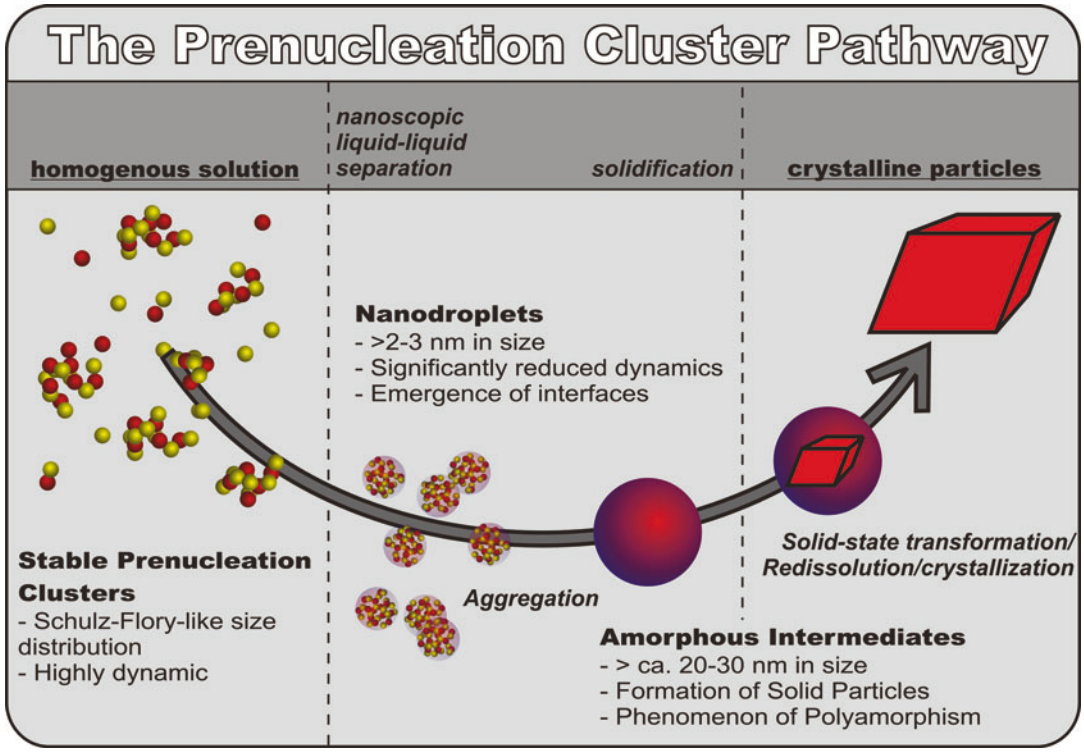
Prenucleation Clusters: An Alternative Concept

Prenucleation clusters, as opposed to classical nuclei, are considered to be solutes. Their formation is thus not linked to the problem of phase separation, at least during the initial stages. Mechanistically, prenucleation cluster formation can be conceived of being analogous with polycondensation reactions, yielding mostly chain-like assemblies of monomers. Hence, solute association leads to Schulz-Flory-like size distributions of a population of clusters [10]. Since a broad distribution of many clusters is entropically favorable, thermodynamically stable clusters will not grow without limit to macroscopic sizes. Also, the solute clusters will not aggregate, because there is no driving force for this due to the absence of interfacial surfaces.

Accordingly, five key characteristics have been suggested to define chemical species as prenucleation clusters [5]:

1. They are composed of chemical entities that also constitute the final solid but may include additional molecules like water.
2. Prenucleation clusters are thermodynamically stable solutes, around 1–3 nm in size on average, and there is formally no phase boundary with the surrounding solution.
3. They are direct molecular precursors to phase-separated species.
4. Stable solute clusters are highly dynamic and change configuration on timescales of hundreds of picoseconds, typical for rearrangements in solution.
5. They can exhibit distinct structural motifs and thereby give rise to the phenomenon of polyamorphism as observed in transient solid intermediates [11].

Computer simulations evidence that upon reaching a certain size, prenucleation clusters can internally develop higher coordination numbers than in the initial chain-like forms. This slows down their dynamics distinctly; thus, the clusters become different from the solution and should then be regarded as nanodroplets of a second phase. The resulting phase interface is subsequently reduced via aggregation, yielding larger, eventually solid amorphous nanoparticles most likely through solidification. However, the direct formation of (nano)crystalline species from prenucleation clusters should not be categorically excluded [5]. Returning to the hallmark of nucleation, the barrier of phase separation close to the solubility limit in the prenucleation cluster pathway is considered to arise from considerable energetic costs for condensing small dynamic solute clusters toward bulk-like states with distinctly higher coordination. Only fluctuations toward larger cluster sizes can lead to the development of somewhat higher coordination numbers,



Prenucleation Clusters, Fig. 2 Schematic illustration of the prenucleation cluster pathway. Stable prenucleation clusters form in solution based on chemical equilibrium. They are direct molecular precursors to phase-separated nanodroplets, forming upon a nanoscopic liquid-liquid separation event. The key difference between prenucleation clusters and nanodroplets is the distinctly reduced dynamics of the latter. This corresponds to the emergence of interfacial surfaces, which are reduced via aggregation, yielding amorphous intermediates. These may solidify and eventually crystallize. For explanation see text

slowing down the dynamics, and phase separation yielding at first liquid nanodroplets can be initiated as described above. Physically, this event can be conceived of as entering a nanoscopic liquid-liquid miscibility gap [5]. A schematic overview of the prenucleation cluster pathway is illustrated in Fig. 2.

Key Research Findings

In case of electrolyte solutions, the prenucleation cluster concept essentially represents an expansion of classical ion-pairing models [5]. In older literature, it is postulated that ion association must be limited to ion pairs when corresponding experimental ion-binding profiles are linear. The basic argument in this case is that the formation of defined larger species would lead to curved ion-binding behavior, as can be shown easily utilizing the law of mass action [5]. The most prominent example for this observation is arguably calcium carbonate. The first central key research finding here is the realization that linear ion-binding profiles are in fact ambiguous and that the formation of higher associated states cannot be unequivocally ruled out based upon ion-binding profiles alone. This ambiguity arises when successive binding events that lead to the generation of larger associates (e.g., in case of calcium carbonate, binding of a calcium ion to a carbonate ion, binding of a calcium ion to a calcium carbonate ion pair, binding of an ion pair to an ion pair, and so forth) are assumed to be all equal and independent, that is, bound by equivalent energetics. Then, all association events are

macroscopically indistinguishable in the first place and may appear like simple ion pairing from the point of view of the law of mass action, despite the presence of large clusters. While this basic assumption may seem like an oversimplification, extensive computer simulations do show that it is indeed sustainable [5]. Thus, in any case, the unambiguous identification, or exclusion, of higher associated states crucially depends upon additional techniques like analytical ultracentrifugation, which can also show that the bound ions quantitatively reside in entities that are much larger than ion pairs on average [5].

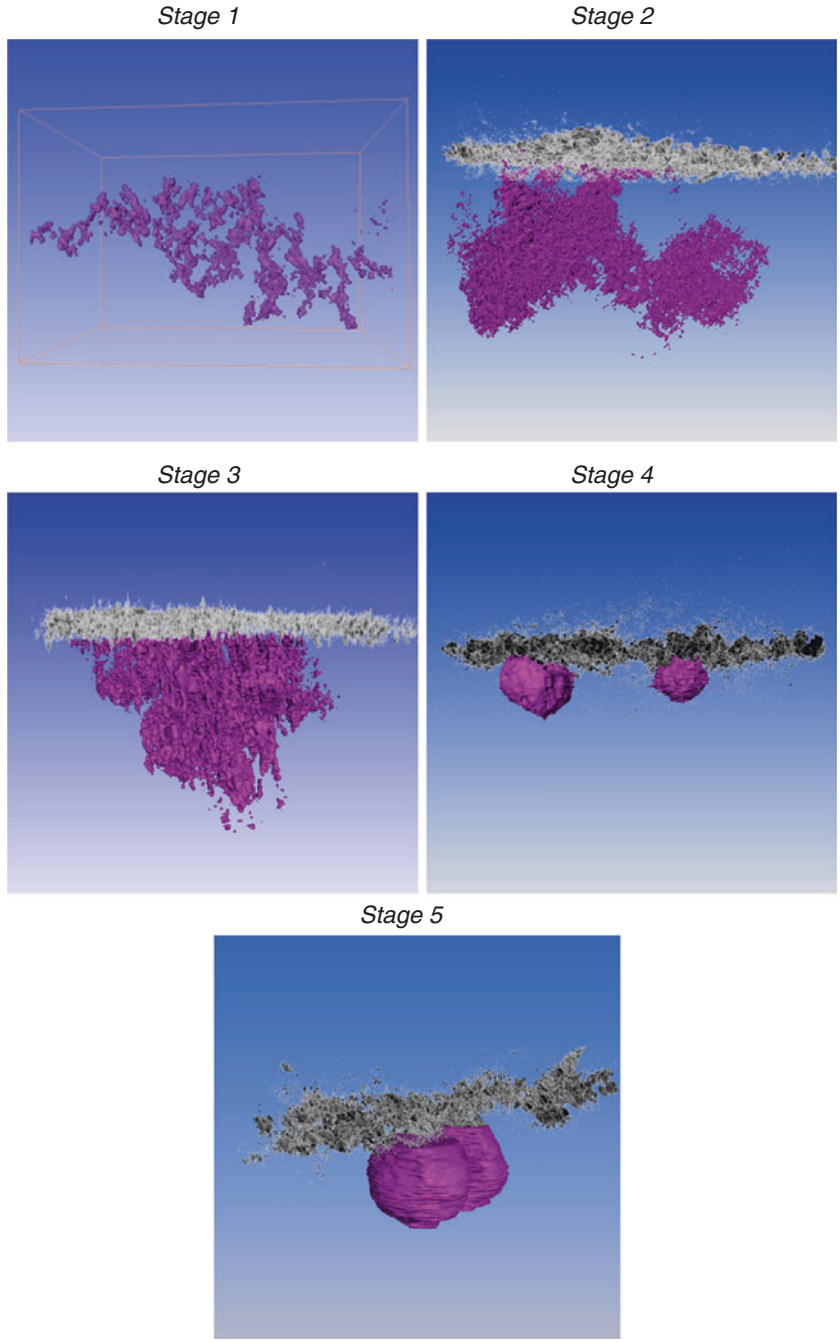
Second, the thermodynamic stability of prenucleation clusters (equilibrium constant of formation $K >> 1$ and a corresponding change in free enthalpy of $\Delta G < 0$) is essential to explain the occurrence of a significant population of larger clusters, which can be analytically detected in the first place. In contrast, metastable species are in fact very rare (owing to an equilibrium constant of formation $0 < K < 1$, and thus $\Delta G > 0$) and usually cannot be detected analytically. Because a metastable state is regarded to be central to the nucleation event within the notions of CNT, stable species have to be regarded as inactive during phase separation processes according to this theoretical framework, because they cannot participate in metastable fluctuations that would lead to the random formation of a critical nucleus. Stable clusters would sit in a free energy trap and hamper formation of a critical nucleus, at least from a classical point of view [5, 12].

However, in the prenucleation cluster pathway, the stable clusters do not yield a solid metastable nucleus of critical size but form amorphous nanodroplets via a structurally more direct transformation, and these nanophases subsequently aggregate to form a new macroscopic phase. At this point, it is important to note that the existence of a nucleation barrier shows that phase separation is a generic consequence of metastable fluctuations, which do not necessarily require the existence of a classical critical nucleus. Thus, the third major key research finding is that metastable fluctuations can also occur in a stable cluster population. They lead to the random occurrence of large clusters through microscopic fluctuations, which correspond to metastable deviations from the equilibrium cluster size distribution. In turn, the random occurrence of a significant population of rather large clusters, which can condense internally as opposed to smaller ones, would be at first reversible and physically correspond to a binodal nanoscopic liquid-liquid separation event [5].

Future Directions for Research

Despite significant progress when it comes to the understanding of the prenucleation cluster pathway in the recent years, many open questions remain that need to be addressed. The following list does not demand to be complete and is supposed to give a rather generic overview:

- *Nucleation theory*: It is most important to include the notion of stable prenucleation clusters into a revised quantitative theory of nucleation. In this context, modeling and simulation approaches appear to be crucial. The implementation concerns homogeneous nucleation theory, but cluster-based pathways do play a role in heterogeneous phase separation processes as well (Fig. 3). The latter are of immediate importance for various industrial and technological problems but also in biomineralization and medical issues.
- *Biomineralization and additive-controlled crystallization*: It is now evident that prenucleation clusters play important roles during surface-induced crystallization in biomimetic approaches (Fig. 3), but their role in biological systems has not been shown directly. Importantly, the prenucleation cluster pathway can mechanistically explain the occurrence of liquid intermediate forms of minerals, which are considered to play a central role in biomineralization [13]. The latter phenomenon provides a direct link to additive-controlled mineralization, since such liquid intermediates can be stabilized by certain polymers. This immediately raises the question whether this phenomenon can be exploited more generally than previously assumed and



Prenucleation Clusters, Fig. 3 Computer-aided 3D visualization of tomograms from cryo-TEM of different stages of the surface-induced mineralization of calcium phosphate from simulated body fluid. Prenucleation clusters (*Stage 1*) that are approximately 0.9 nm in size aggregate upon tranformation into nanodroplets on an extrinsic surface (*Stages 2 & 3*) to form initially amorphous nanoparticles (*Stage 4*), which subsequently crystallize (*Stage 5*; data kindly provided by A. Dey and N. A. J. M. Sommerdijk)

how the prenucleation cluster pathway may improve the understanding of additive-controlled mineralization, as suggested for biomineralization [14].

- *Generality*: Prenucleation clusters have been shown to exist for a series of compounds, including most biominerals but also small organic molecules. The prenucleation cluster pathway may be a more general phenomenon that may apply to the formation mechanism of many solids, perhaps not only in aqueous solution. The generality of the prenucleation cluster concept has to be tested, including the identification of conditions under which cluster-based mechanisms of phase separation may or may not dominate.

Cross-References

- ▶ Biomimetics
- ▶ Nanoparticles
- ▶ Nanoscale Properties of Solid–liquid Interfaces
- ▶ Nanotechnology in Cardiovascular Diseases
- ▶ Self-assembly of Nanostructures
- ▶ Transmission Electron Microscopy

References

1. Rieger, J., Kellermeier, M., Nicoleau, L.: Formation of nanoparticles and nanostructures – an industrial perspective on $CaCO_3$, cement and polymers. Angew. Chem. Int. Ed. **53**, 12380–12396 (2014)
2. Lowenstam, H., Weiner, S.: On Biomineralization. Oxford University Press, New York (1989)
3. Cölfen, H., Antonietti, M.: Mesocrystals and Nonclassical Crystallization. Wiley, Chichester (2008)
4. Mullin, J.: Crystallization. Butterworth-Heinemann, Oxford (2001)
5. Gebauer, D., Kellermeier, M., Gale, J.D., Bergström, L., Cölfen, H.: Pre-nucleation clusters as solute precursors in crystallisation. Chem. Soc. Rev. **43**, 2348–2371 (2014)
6. Davey, R.J., Schroeder, S.L.M., ter Horst, J.H.: Nucleation of organic crystals – a molecular perspective. Angew. Chem. Int. Ed. **52**, 2166–2179 (2013)
7. Nielsen, A.E.: Kinetics of Precipitation. Pergamon, New York (1964)
8. Kashchiev, D.: Nucleation: Basic Theory with Applications. Butterworth-Heinemann, Oxford (2000)
9. Vekilov, P.G.: The two-step mechanism of nucleation of crystals in solution. Nanoscale **2**, 2346–2357 (2010)
10. Flory, P.J.: Fundamental principles of condensation polymerization. Chem. Rev. **39**, 137–197 (1946)
11. Cartwright, J.H., Checa, A.G., Gale, J.D., Gebauer, D., Sainz-Díaz, C.I.: Calcium carbonate polyamorphism and its role in biomineralization: how many amorphous calcium carbonates are there? Angew. Chem. Int. Ed. **51**, 11960–11970 (2012)
12. De Yoreo, J.J.: Crystal nucleation: more than one pathway. Nat. Mater. **12**, 284–285 (2013)
13. Gower, L.B.: Biomimetic model systems for investigating the amorphous precursor pathway and its role in biomineralization. Chem. Rev. **108**, 4551–4627 (2008)
14. Evans, J.S.: "Liquid-like" biomineralization protein assemblies: a key to the regulation of non-classical nucleation. CrystEngComm **15**, 8388–8394 (2013)

Preparation of Complex Oxide Unidimensional Nanostructures

▶ Growth of 1-D Oxide Nanostructures

PRINT®

▶ Nanoscale Printing

Probe Technology of Scanning Probe Microscopy (SPM) and Its Fabrication Technology

▶ Microfabricated Probe Technology

Product of Random Matrices

▶ Harnessing Disorder at the Nanoscale

Production of Gold Nanoparticles

▶ Synthesis of Gold Nanoparticles

Production of Graphene or Graphene Production

► Synthesis of Graphene

Properties Related to Applied Stress or Strain

► Mechanical Properties of Nanostructures

Propylene Glycol Methyl Ether Acetate (PGMEA)

► SU-8 Photoresist

Prostheses

► MEMS Neural Probes

Protein Nano-crystallization

► Macromolecular Crystallization Using Nano-volumes

Pseudoelasticity

► Superelasticity and the Shape Memory Effect

PuCP

► Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Pulsed Focused Ultrasound

► Pulsed High-Intensity Focused Ultrasound

Pulsed HIFU

► Pulsed High-Intensity Focused Ultrasound

Pulsed High-Intensity Focused Ultrasound

Brian E. O'Neill[1] and King C. Li[2]
[1]Department of Radiology Research, The Methodist Hospital Research Institute, Houston, TX, USA
[2]Department of Bioengineering, Rice University, Houston, TX, USA

Synonyms

Pulsed focused ultrasound; Pulsed HIFU

Definition

Pulsed high-intensity ultrasound (pHIFU) is a technique used to physically modify the permeability and hydraulic conductivity characteristics of living tissue in order to increase penetration and accumulation of large molecular drugs or nanoparticles in the interstitial space. Pulsed HIFU is a form of *ultrasound-enhanced drug delivery*; however, while the latter may be assisted by the use of exogenous agents such as *ultrasound contrast agents* or *microbubbles*, those who use pHIFU generally rely on the physical interaction of the ultrasound with the tissue alone. The use of pulsed energy deposition is specifically designed to enhance the mechanical over the thermal effects of the focused ultrasound by allowing for deposited heat to dissipate even at high powers.

Microbubble contrast agents are avoided to prevent *cavitation* in the tissue.

Ultrasound and the Nanoparticle Delivery Problem

The transport of agents to tissues can be divided into steps based on the different physiological barriers presented, from administration to final interaction with individual cells. The relative importance of each step is dependent on the type of agent (size, surface chemistry), the type of tissue, and the type of administration (intravenous, direct injection). Assuming systemic injection, the amount of agent that ultimately reaches an individual cell depends, in some part, on eight factors: (1) agent-specific blood pool pharmacokinetics, (2) tissue perfusion, (3) local vascular topography, (4) agent-specific local vascular permeability, (5) agent-specific interstitial conductivity, (6) interstitial topography, (7) cellular affinity, and, finally, (8) vascular/lymphatic drainage efficiency. In the case of a direct (intratissue) injection, the first four factors are much less important but probably still play a role in the distribution of agents whose transport is convection dominated.

Nanoparticles pose a specific challenge to the natural transport scheme because of their sizes and surface chemistry. Certain issues, such as blood pool pharmacokinetics, have been largely addressed with various surface coatings and proper particle size; however, the issue of getting nanoparticles across vascular, endothelial, and cell membrane barriers and into the tumor parenchyma and cell is still a challenge. First, within the vasculature of tumors there are pockets of stasis and poor perfusion, due in large part to the improperly constituted and chaotic network of blood vessels in these environments [1]. This situation might be improved by increases in perfusion due to ultrasound hyperthermia or other vascular effects of ultrasound. Huge increases in perfusion are known to accompany hyperthermia in temperature-sensitive tissue such as muscle. Increases in tumor are much more modest and limited [2] but still important. Second, most of the currently successful nanoparticles targeted to solid tumors rely heavily on the atypically high fenestration of the endothelial barrier in tumors compared to most other tissues. This means that particles in the size range of ~50–200 nm in diameter have a tendency to accumulate preferentially in tumor tissue even without further considerations (known as the enhanced permeability and retention or EPR effect). Unfortunately, the exact size cutoff is highly dependent on tumor type and can be highly variable across tumors and even within the same tumor [3]. In tissues with poor fenestration, there is evidence that ultrasound treatment can enhance the movement of nanoparticles across the endothelium. Third, nanoparticles generally move most efficiently by convection; however, in the tumor microenvironment, convective flow beyond the vasculature is virtually nonexistent. The interstitial conductivity and topography that determine how an agent penetrates and distributes in the interstitial space by either convection or diffusion is not favorable to the movement of nanoparticles through tumors. Cellular crowding due to rapid growth and the production of a collagenous extracellular matrix can both result in barriers to large particle movement. As well, the lack of lymphatics in tumors contributes to poor drainage, causing high interstitial fluid pressures and sluggish flow. A number of preliminary experiments involving pHIFU and direct intratumoral injection of agents have suggested improved distribution in the interstitial space. A final issue is the difficulty of getting nanoparticles into the cell through the cell membrane. Although many potential techniques exist to address this problem, many of them leave the nanoparticle sequestered inside the cell only to be ejected later, or are not practical for clinical work. The effects of ultrasound cavitation can transiently increase membrane permeability (sonoporation), and may be able to provide a physical means of achieving intracellular accumulation of nanoparticles. All these considerations suggest strongly that there is a need to remodel or re-engineer the transportation dynamics to be more favorable to nanoparticle traffic, and that ultrasound treatment provides some opportunity to do so.

Basic pHIFU Physics

Ultrasound is a coherent traveling pressure disturbance that easily transports mechanical and thermal energy through biological soft tissues [4]. As ultrasound travels, the coherent wave is attenuated by absorption and scattering processes. A traveling ultrasonic pressure plane wave might generally be written as

$$p(x,t) = p_0 \sin\left(\omega\left(\frac{x}{c} - t\right)\right)\exp(-\mu_a x),$$

where p_0 is the pressure amplitude, ω is the frequency, c is the speed of sound, and μ_0 is the attenuation of the wave. The intensity of the wave is given by $I = \frac{p_0^2}{2\rho c}$ with ρ the density of the medium. At frequencies around 1 MHz, traveling at a speed of 1,500–1,600 m/s in water or tissue, the ultrasonic wavelength is on the order of 1.5–1.6 mm, much too long to directly couple to objects on the cellular scale. The low compressibility of water means that ultrasonic strains remain quite low even when pressures reach tens of atmospheres. Thus, the pressure disturbance itself, even at powers as high as 10^4 W/cm^2, is generally regarded as having no direct biological effects, unless cavitation is involved (see below). In the absence of cavitation, part of the wave energy is still continuously being reflected or absorbed as heat, a fact modeled with an attenuation coefficient. The attenuation coefficient, μ_a, determines how the wave transmits energy and momentum to the media through which it passes. Attenuation in general incorporates both scattering and absorption phenomena; however, often absorption is thought to be dominant. The absorption of the energy creates heat, and this heat per unit volume is given by

$$\dot{Q} = \mu_a I$$

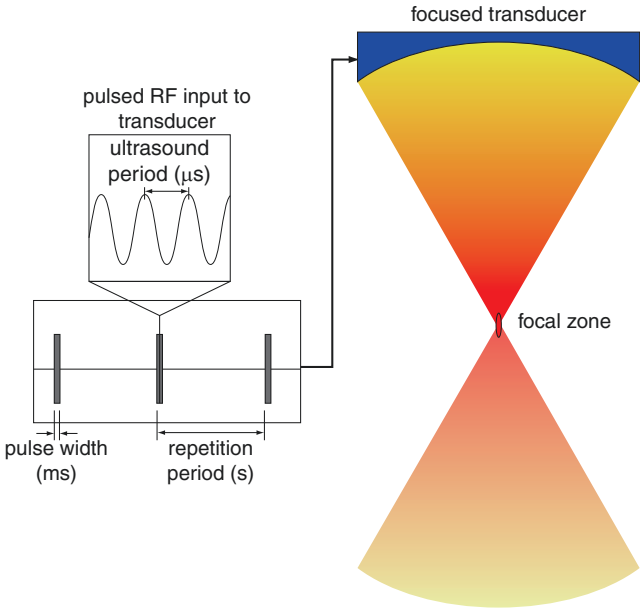
As well, both absorption and scattering result in a transfer of momentum from the wave to the medium. The resulting body force is known as the radiation pressure and is often written (ignoring scattering) as

$$F = \frac{\mu_a I}{c}$$

Generally, in the low megahertz frequency range, attenuation is not very large; however, at very high intensity the effect is still biologically significant. Such high intensity may be achieved by concentrating the ultrasound from a large source into a small volume. Focused ultrasound may be achieved by means of a curved transducer or an acoustic lens. Either way, the result is a concentration of much of the acoustic energy produced by the transducer into a diffraction-and-aperture-limited ellipsoidal focal spot (Fig. 1). In typical pulsed HIFU experiments, using a 1.0 MHz transducer with equal aperture and focal depth ($f = 1$), the spot size full-width half-maximum is about 1.5 mm in diameter and 7 mm in length. With this design, local peak intensities can range up to 10^4 W/cm^2. Pulsed HIFU powers are typically maintained in the range of 10^2–10^3 W/cm^2.

Ultrasonic heating is the basis of HIFU ablation and ultrasound hyperthermia treatment. The principal goal of using pulsed rather than continuous HIFU is to avoid overheating the tissue by allowing the heat deposited in the focal zone to diffuse, while still maintaining high peak powers to induce the mechanical effects. In fact, there will always be some unavoidable heating, but because the thermal response is relatively slow (order of seconds), it is governed by the average power. In contrast, the mechanical response is rapid (order of tens of milliseconds), and thus is related more to the peak power.

If sufficient levels of gas exist, either dissolved or in suspension, then acoustic cavitation becomes a possibility. In that case, rapid and large expansion of the gaseous phase can happen during the negative-pressure portion of the cycle, a phenomenon known as acoustic cavitation [5]. Because of the huge acoustic impedance mismatch between gas and liquid and because of the resonant expansion and contraction of gas microbubbles in an acoustic wave, cavitation produces an enormous concentration of energy in and around the microbubbles that can have a number of biological effects. Acoustic cavitation phenomena are divided into two types: the first, known as



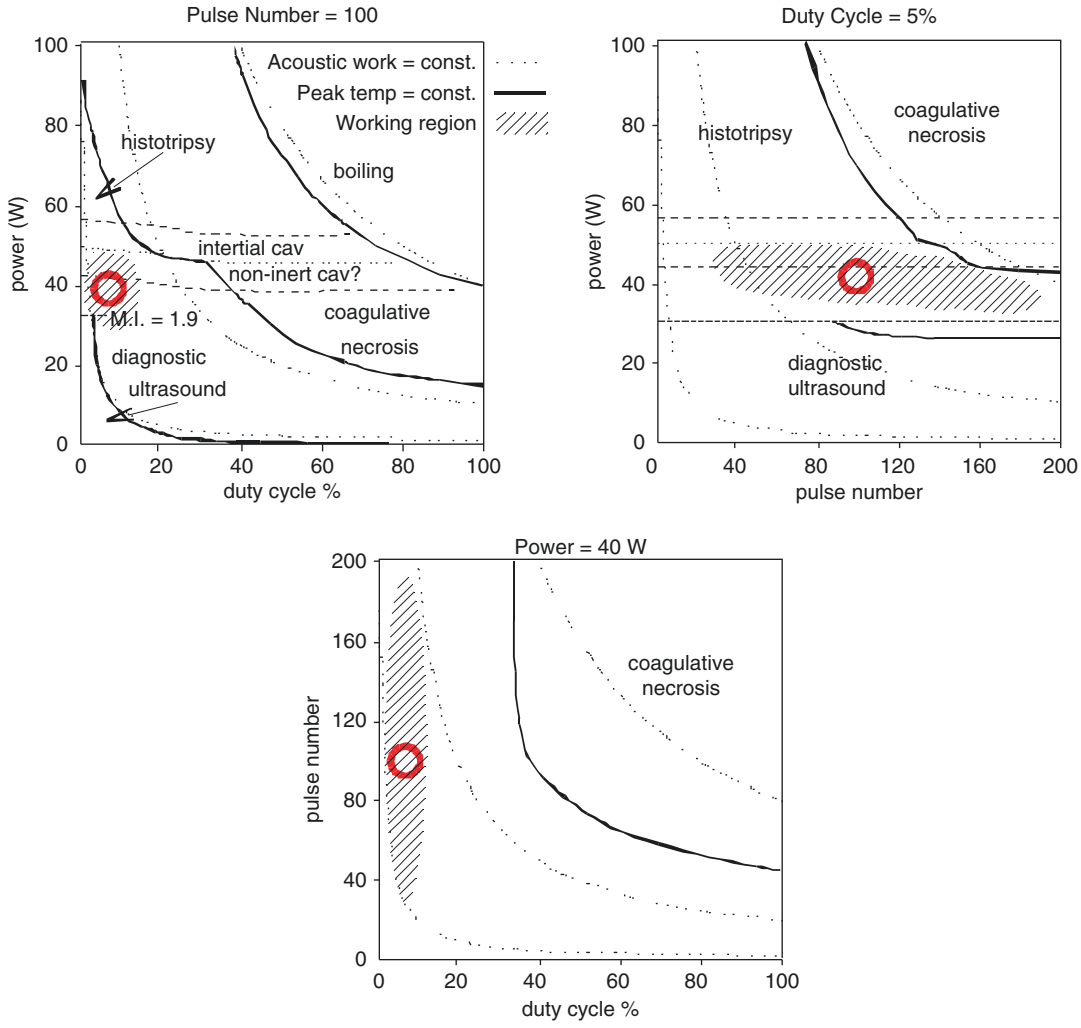
Pulsed High-Intensity Focused Ultrasound, Fig. 1 Schematic of pulsed high-intensity focused ultrasound. Pulsed input to focused ultrasonic transducer results in high peak intensity, low average intensity beams, with highest energy density in and around the focal zone. The focal zone is scanned across the tissue to enable treatment of large volumes

noninertial or stable cavitation, involves gas bubbles resonating around an equilibrium size, while the second, known as inertial or unstable, involves the rapid expansion and subsequent violent collapse of gas bubbles. Generally, the existence of cavitation in tissue is inferred from its effect on the spectral content of the wave; in particular, the noninertial cavitation will create an unambiguous increase in half harmonic content, while inertial cavitation also causes an increase in broadband noise level. Both processes involve the concentration of acoustic, mechanical, and thermal energy in the neighborhood of the bubble, except that in the second instance, this energy is focused even more into a tiny volume during collapse, with resulting extremes in temperature and pressure. These in turn can have profound effects on tissue, including enhanced heating, membrane poration, and cell lysis.

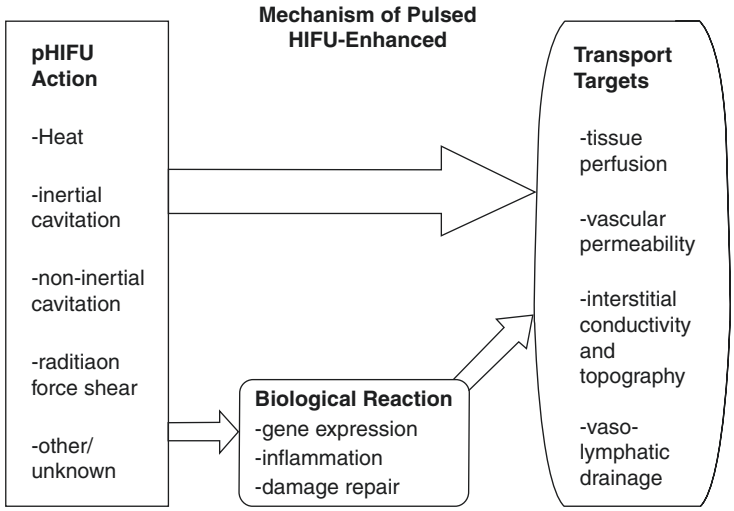
Biological Impact of pHIFU

The set of acoustic parameters most often used for pHIFU applications are 100 pulses at 1 MHz center frequency, 5 % duty cycle, and 1 Hz repetition rate (50 ms on, 950 ms off), 40 W peak acoustic power, using an f = 1 transducer (approximately 2,200 W/cm^2 in water). Another common set of parameters employed roughly half the power and double the duty cycle. (The earliest studies described below used much higher total energy and peak powers; however, because the efficiency of the transducer was not measured, it is not clear precisely what these were.) Under these study conditions, mostly carried out in mouse models, researchers have demonstrated changes in normal and tumor tissues that effectively increase the rates of nanoparticle uptake by two to three folds over untreated tissues. The effects appear to be accompanied by edema and often inflammation. While the precise mechanism remains uncertain, some possibilities exist. It is widely accepted that inertial and possibly noninertial cavitation could cause such effects. Indeed, introducing cavitating microbubbles in the form of ultrasound contrast agents is a technique that has been employed for blood–brain barrier disruption [6]. However, under the conditions given, there is no evidence that the cavitation threshold is being reached during pHIFU in the absence of exogenous agents. Cavitation is very difficult to achieve in tissues, first, because the levels of available gas are small, and second, because the threshold for cavitation is increased

with increasing viscosity. Heating is a more likely factor, although preliminary studies have generally concluded that the heating and thermal dose produced is rather low when compared to standard hyperthermia treatments. However, many of the effects seen do appear to mimic those of hyperthermia. Finally, tissue shear is created by the radiation forces and also when the ultrasonic wave reflects from a boundary at some angle and undergoes mode conversion. The resulting highly attenuating shear waves propagate at a much lower velocity, approximately 30–100 m/s, therefore having a more biologically relevant wavelength of 30–100 μm. It is generally accepted that shear forces can mediate biological effects, particularly in the vasculature [7], where normal blood flow shear is known to be a factor in the proper maintenance of the endothelial layer. Such processes might be disrupted or mimicked by radiation force-induced streaming in the blood vessels. It is



Pulsed High-Intensity Focused Ultrasound, Fig. 2 Pulsed HIFU neighborhood in tissue "phase space," in relation to cavitation, thermal ablation, and diagnostic ultrasound. The effect on tissue of ultrasound in this region is incompletely understood. Axes are the principal treatment parameters in common units (reflecting the choice of f = 1, 1 MHz focused transducer, and 1 Hz pulse repetition rate): acoustic power (50 W acoustic power implies ~2,800 W/cm^2 in the focal zone), duty cycle (5 % implies pulse length of 50 ms), and total number of applied pulses



Pulsed High-Intensity Focused Ultrasound, Fig. 3 Pulsed HIFU acts to alter tissue transport either directly, by altering the physical properties of transport targets, or indirectly, by inducing biological changes that then influence transport properties

less clear how the relatively small shear strains of radiation forces, or the high frequency, rapidly attenuating shear waves, might affect the cells of bulk tissue.

As well as the direct physical effects of pHIFU on tissues, there is always the possibility of triggering secondary biological processes, such as inflammation. Indeed, pHIFU often appears to trigger a clear inflammatory response, at least in muscle. It is not so clear whether this is also true in tumor or other tissue, nor whether the observed vascular and parenchymal permeability is the cause or the effect of inflammation.

If tissue changes are considered as comparable to material phase changes, then pHIFU effects lie in volatile and unexplored regions of this phase space. This paradigm is illustrated in Fig. 2, which shows where treatment might be relative to a tissue phase map using the three fundamental treatment parameters as axes. Figure 3 lists the candidates for pulsed HIFU action and possible tissue targets.

Examples of Application

Pulsed HIFU has been explored in combination with a diverse set of agents, including small-molecule chemotherapeutics, thrombolytics, large-molecule therapeutics, antibodies, plasmid DNA, viral vectors, and nanoparticle drug carriers. The main goal of the technology since it was first conceived in the mid-1990s was to allow the uniform penetration of nanoparticle drug carriers throughout the tumor parenchyma. In that early study [8], the nanoparticle was a polymerized liposome containing gadolinium contrast agent of mean diameter around 100 nm. MRI was used to monitor the delivery of the particle to rabbit muscle. The delivery of plasmid DNA (fluorescent reporter gene) was successfully demonstrated in a mouse squamous cell carcinoma more than 5 years later [9]. Later, the same technique was used to enhance delivery of TNF alpha plasmid and the liposomal nanoparticle Doxil and for tumor therapy [10, 11]. In 2006–2007, pHIFU was used to deliver radioactive antibodies [12] as well as genes via viral vector to mouse tumor models [13]. It was also combined with small molecular weight chemotherapeutics [14] and thrombolytic agents [15]. During various mechanistic studies, the delivery of nanoparticle-sized fluorescent dextrans and microspheres were used as tracking agents. Most of the mechanistic studies were done in mouse tumor models or mouse muscle. Comprehensive reviews of these and other related studies are available [16, 17].

Future Directions

Pulsed HIFU produced by ultrasound and magnetic resonance image-guided systems will be

used for assisted delivery of a variety of tissue targeted and untargeted nanoparticles. Virtually any large molecule or nanoscale therapeutic agent can take advantage of this technology for improved, targeted delivery. In fact, to be truly efficacious, it is anticipated that many nanoparticle systems will require the assistance of this or similar technology. Finally, it is possible to combine image-guided pulsed HIFU with various ultrasound and energy-sensitive nanoconstructs for further control local and systemic therapeutic concentrations.

Cross-References

- ▶ Liposomes
- ▶ Nanoencapsulation
- ▶ Nanoparticles
- ▶ Thermal Cancer Ablation Therapies Using Nanoparticles

References

1. Jain, R.K.: Transport of molecules in the tumor interstitium: a review. Cancer Res. **47**(12), 3039–3051 (1987)
2. Song, C.W.: Effect of local hyperthermia on blood flow and microenvironment: a review. Cancer Res. **44**, 4721s–4730s (1984)
3. Netti, P.A., Hamberg, L.M., Babich, J.W., Kierstead, D., Graham, W., Hunter, G.J., Wolf, G.L., Fischman, A., Boucher, Y., Jain, R.K.: Enhancement of fluid filtration across tumor vessels: implication for delivery of macromolecules. Proc. Natl. Acad. Sci. U. S. A. **96**(6), 3137–3142 (1999)
4. Hill, C.R., Bamber, J.C., ter Haar, G.R. (eds.): Physical Principles of Medical Ultrasonics, 2nd edn. Wiley, Chichester (2004)
5. Kimmel, E.: Cavitation bioeffects. Crit. Rev. Biomed. Eng. **34**(2), 105–161 (2006)
6. McDannold, N., Vykhodtseva, N., Hynynen, K.: Targeted disruption of the blood-brain barrier with focused ultrasound: association with cavitation activity. Phys. Med. Biol. **51**(4), 793–807 (2006)
7. Hillsley, M.V., Tarbell, J.M.: Oscillatory shear alters endothelial hydraulic conductivity and nitric oxide levels. Biochem. Biophys. Res. Commun. **293**(5), 1466–1471 (2002)
8. Bednarski, M.D., Lee, J.W., Callstrom, M.R., Li, K.C.: In vivo target-specific delivery of macromolecular agents with MR-guided focused ultrasound. Radiology **204**(1), 263–268 (1997)
9. Dittmar, K.M., Xie, J., Hunter, F., Trimble, C., Bur, M., Frenkel, V., Li, K.C.: Pulsed high-intensity focused ultrasound enhances systemic administration of naked DNA in squamous cell carcinoma model: initial experience. Radiology **235**(2), 541–546 (2005)
10. Yuh, E.L., Shulman, S.G., Mehta, S.A., Xie, J., Chen, L., Frenkel, V., Bednarski, M.D., Li, K.C.: Delivery of systemic chemotherapeutic agent to tumors by using focused ultrasound: study in a murine model. Radiology **234**(2), 431–437 (2005)
11. Frenkel, V., Etherington, A., Greene, M., Quijano, J., Xie, J., Hunter, F., Dromi, S., Li, K.C.: Delivery of liposomal doxorubicin (Doxil) in a breast cancer tumor model: investigation of potential enhancement by pulsed-high intensity focused ultrasound exposure. Acad. Radiol. **13**(4), 469–479 (2006)
12. Khaibullina, A., Jang, B.-S., Sun, H., Le, N., Yu, S., Frenkel, V., Carrasquillo, J.A., Pastan, I., Li, K.C.P., Paik, C.H.: Pulsed high-intensity focused ultrasound enhances uptake of radiolabeled monoclonal antibody to human epidermoid tumor in nude mice. J. Nucl. Med. **49**, 295–302 (2008)
13. Sun, H., Khaibulina, A., Li, K.: Bioluminescence studies: pulsed high-intensity focused ultrasound enhances the viral gene delivery in tumor models. Mol. Imaging **5**, 336 (2006)
14. Poff, J.A., Allen, C.T., Traughber, B., Colunga, A., Xie, J., Chen, Z., Wood, B.J., Waes, C.V., Li, K.C.P., Frenkel, V.: Pulsed high-intensity focused ultrasound enhances apoptosis and growth inhibition of squamous cell carcinoma xenografts with proteasome inhibitor bortezomib. Radiology **248**, 485–491 (2008)
15. Dromi, S., Frenkel, V., Luk, A., Traughber, B., Angstadt, M., Bur, M., Poff, J., Xie, J., Libutti, S.K., Li, K.C.P., Wood, B.J.: Pulsed-high intensity focused ultrasound and low temperature-sensitive liposomes for enhanced targeted drug delivery and antitumor effect. Clin. Cancer Res. **13**, 2722–2727 (2007)
16. Frenkel, V., Li, K.C.P.: Potential role of pulsed-high intensity focused ultrasound in gene therapy. Future Oncol. **2**, 111–119 (2006)
17. O'Neill, B., Li, K.C.P.: Augmentation of targeted delivery with pulsed high intensity focused ultrasound. Int. J. Hyperthermia **24**, 506–520 (2008)

Pulsed Laser Deposition (PLD)

► Physical Vapor Deposition

Pyroelectricity

► Polarization-Induced Effects in Heterostructures

PμCP

► Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Q

Quantum Cluster

► Synthesis of Subnanometric Metal Nanoparticles

Quantum Dot

► Synthesis of Subnanometric Metal Nanoparticles

Quantum Dot Nanophotonic Integrated Circuits

Lih Y. Lin
Department of Electrical Engineering, University of Washington, Seattle, WA, USA

Definition

Quantum dot nanophotonic integrated circuits are photonic integrated circuits composed of nanoscale optoelectronic components made with semiconductor quantum dots.

The modern electronics industry was revolutionized by the invention of the transistor. Since then, the transistor has been the fundamental building block of all modern electronic devices widely deployed in various places including communication and computer systems. The number of transistors on an integrated circuit has roughly doubled every 2 years, following Moore's Law. Moore himself, however, has stated that this law cannot be sustained indefinitely due to the fundamental limits to miniaturization.

Although the demand for higher processor chip speeds in a computer system continues to drive the bandwidth and density requirements and performance for on-chip and chip-to-chip interconnects, using electrons as the information carrier is now facing the challenges of speed, power consumption, and crosstalk due to electromagnetic interference between channels. The limits of Moore's law call for fundamentally new ways to transmit, process, and compute information, not electronically, but perhaps optically. In general, smaller and faster photonic integrated circuit systems with versatile functionalities are desirable for applications not only in computation, but also in other areas such as optical communication and sensing. The need has fueled the development of new photonic devices, new fabrication methods like self-assembly, and the synthesis of new optoelectronic materials like nanocrystals and nanowires. With these advances, it is now possible to realize lasing, waveguiding, optical modulation, and detection in submicron scale devices. Furthermore, it is possible to pick and place nanoparticles on silicon substrates by merely exposing the substrate surface to a liquid droplet of suspended particles. With these new capabilities and tools, building nanoscale photonic

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

integrated circuits has begun to be widely investigated, in an effort to achieve an optical analogy of VLSI with higher speed, lower power consumption, and higher integration density.

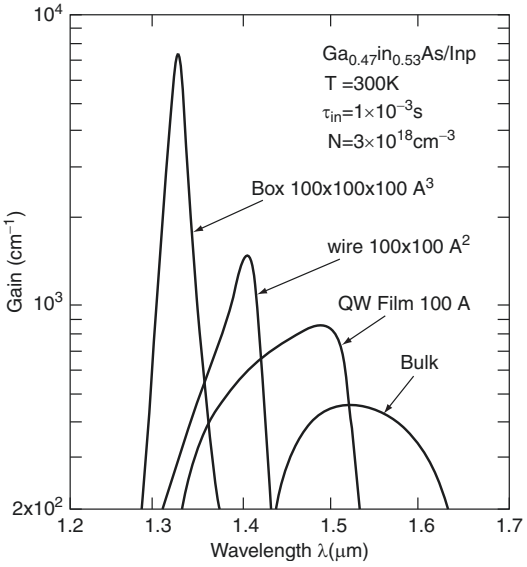
The development of photonic integrated circuits has nonetheless been hindered by the diffraction limit, which restricts the extent of optical wave confinement to $\sim\lambda/2$ for guiding and transmitting optical energy in conventional optical waveguides, and constrains the integration density one can achieve using photons as information carriers. While electronic integrated circuit technologies are advancing fast in the sub-100 nm lithographic resolution regime, the diffraction limit must be overcome for photonics to successfully complement and perhaps replace some of the electrical components. The advances in nanofabrication technologies have facilitated the miniaturization and integration density of device technologies, with semiconductor quantum dots (QD), also referred to as nanocrystals (NC), as materials with high promise.

Since the early 1980s, the quantum dot has gained increasing notice and visibility. Combined with modern material science techniques, the focus has shifted from modeling behavior to implementing and taking advantage of the QD's unique properties for various applications. Originally, the quantum dot was conceived as the next logical step in lasing devices after quantum well and quantum film structures. Progressing from one- to two- and then to three-dimensional confinement of electron and hole wave functions narrows the gain spectrum and increases maximum gain, which reduces the threshold current to begin laser operation. Furthermore, the quantization of the electron and hole energy states and the material and size-specific absorption characteristics are responsible for enabling the high extinction ratios of QD optical labels in biomedical research, and provide a basis for several photonic devices, including transistors, switches, and, naturally, lasers.

Characteristics of QDs

In general terms, the quantum dot is a form of nanocrystals, or a monodisperse crystalline metal or semiconductor particle whose diameter is adjustable on the nanoscale. QDs are of particular interest from a device standpoint as the electron and photon behavior in the materials differ greatly from that of bulk materials. When deposited in an ordered manner and formed into a solid, collective traits may arise that are also distinctive. As an individual particle, the performance of a QD may be tuned through variations in size and shape, leading to changes in electron–hole generation and recombination dynamics, tunneling behavior, and energy transfer. As in bulk semiconductors, population inversion can be achieved with a sufficient excitation rate and energy, which results in an emission coefficient greater than the absorption coefficient leading to net optical gain. Furthermore, the three-dimensional confinement of electron–hole pairs (exciton wave function) leads to discretized energy levels and delta-function-like distribution of the density of states. Under excitation, the emission of QDs is spectrally much sharper than their bulk material counterparts, which results in a sharp gain spectrum and therefore high gain, low loss, and high quantum efficiency (Fig. 1) [1]. The spectral response of QDs can be easily tuned by particle size and composition. These unique properties make QDs ideal materials for optoelectronic applications.

Conventional QD photonic devices utilize self-organization at the interface of strained epitaxially grown III–V or II–VI semiconductor layers to create pyramidal QD nanostructures due to lattice mismatch. The fabrication is often done through metal-organic chemical vapor deposition (MOCVD). The Stranski–Krastanov process is the basis for manufacture of the first QD lasers, which have been theoretically and experimentally demonstrated with improved gain characteristics and reduced lasing current thresholds compared to quantum well and quantum film structures. Although high-quality optoelectronic devices have been obtained using the self-organization method, the process itself is often not compatible with silicon CMOS fabrication. Another process to fabricate QDs is through chemical synthesis, or solution-based processes. The resulting QDs are in colloidal form and have been widely used as fluorescent tags in biomedicine. The colloidal



Quantum Dot Nanophotonic Integrated Circuits, Fig. 1 Gain spectra for $Ga_{0.47}In_{0.53}As/InP$ 100 Å × 100 Å × 100 Å cubic QD, 100 Å × 100 Å quantum wire, 100 Å-thick quantum well film, and bulk crystal. The calculation was done assuming $T = 300$ K, intraband relaxation time $\tau_{in} = 1 \times 10^{-13}$ s, and electron–hole density $N = 3 \times 10^{18}$ cm^{-3} [1]

QDs can then be subsequently deposited on a variety of substrates including silicon and mixed with different materials using various methods to be described later, resulting in rapid expansion in applications such as whispering gallery lasers, sensors, organic LEDs, and solar cells.

In the colloidal case, where individual nanoparticles are fabricated in a solution suspension, the composition materials often fall in the semiconductor category such as II–VI and III–V. The process often involves controlled injection of reagents into the solvent to nucleate and grow the compound of interest. Recently, synthesis of group IV semiconductor QDs has also been investigated. In the case of CdSe nanocrystals, dimethylcadmium, and organophosphine selenide combine to form a core QD structure. A shell layer, such as ZnS, may be subsequently added by solution growth to suppress surface oxidation and improve the fluorescence behavior. The diameters of the core and shell thicknesses determine the energy levels in the QD, which affect the absorption spectrum and correspond to the emission wavelength of the photon generated after the excitation and recombination of an electron–hole pair [2]. The shell layer also allows further surface chemical treatment to enable specific reactions with surrounding media. By design, a shell material with a larger bandgap is commonly selected to enable better electron–hole pair confinement, whereas the surface is treated with stabilizers and can be conjugated with molecules for specialized binding activity, which is dependent on the application.

Colloidal QDs offer high fabrication and integration flexibility. This article focuses on the fabrication, deposition, and photonic devices made of colloidal QDs.

QD Fabrication

Solution-based fabrication of QDs involves techniques to achieve colloids in a controlled manner [3]. To begin, a solvent containing stabilizing and reacting chemicals is poured into a vessel and typically heated to a temperature ideal for stimulating NC formation. Then, a rapid injection of metal-organic material supersaturates the solution and breaks above the nucleation threshold. Subsequently, small particles form, or nucleate, to reduce the energy within the system. Over time, if the rate of injection is below the consumption rate of reagents, then no new nuclei are created and the existing NCs incorporate more metal-organic precursors to become larger.

With regards to specific compounds, long alkane chains attached to phosphorous or phosphorous oxide provide the active ingredients in the solvent mixture. For II–VI semiconductor NC fabrication, the precursors include metal alkyls consisting of the group II metal element and organophosphine chalcogenides or bistrimethylsilylchalcogenides incorporating the group VI chalcogenide element. Taking the CdSe QD as an example, dimethylcadmium and organophosphine selenide are commonly used.

As for ensuring monodisperse nanocrystals, the size distribution will tend toward homogeneity

given that the initial nucleation event produced particles whose eventual dimensions eclipse the original volume. Moreover, a secondary growth phase known as Ostwald ripening may aid the process by dissolving smaller NCs with high surface energy in favor of material deposition on the larger ones. Overall, the procedure may be calibrated and timed to synthesize particles of a desired size. Additionally, reactants and NCs are removed from the solution as appropriate.

For creating core/shell QDs, the materials must have compatible deposition conditions and surface energies at the interface. Another requirement is that the nucleation threshold of the shell material be lower than the core, and interdiffusion between the core and shell elements must not occur. The production method of bilayer QDs generally appends to the core synthesis procedure by selecting out NCs of the desired size and adding them to a new reaction solution. Upon reaching the optimal temperature, precursors with the shell material are added at a rate which allows for gradual deposition onto the original particles but avoids a secondary nucleation [3]. To add the final capping layer, an exchange process is used whereby the stabilizing molecules attached to the NCs are interchanged through competition with an excess of another molecule, molecule group, or ion. Furthermore, clumping of particles is avoided if the repulsion strength of outermost layer is greater than the van der Waals attraction force.

On the whole, production of core and core/shell structures results in a size distribution of 5–10 % standard deviation from the mean. Further filtration and separation steps may be carried out to identify and collect nanocrystals of better uniformity. Some standard structural characterization methods for NCs include high resolution transmission electron microscopy (TEM) and both small- and wide-angle X-ray scattering (SAXS, WAXS) while analysis of the chemical constitution may be accomplished through nuclear magnetic resonance (NMR), X-ray fluorescence, and photoelectron spectroscopy.

Due to its nontoxicity and lower material cost, silicon (Si) QD is an actively pursued research topic and much effort in developing-related technologies that lead to future commercialization is being invested. Development in colloidal Si QD synthesis thus far lags behind the progress of II–VI QDs; nevertheless, processes that lead to Si QDs with good material characteristics have been demonstrated. Hydrogen-terminated colloidal Si QDs can be synthesized using a low temperature solution-based process [4]. The Si QDs can be prepared by a solution reaction using sodium silicide and ammonium bromide in dimethoxyethane (DME, 80 °C) and dioctyl ether (DOE, 260 °C). The chemical reaction of the process follows $NaSi + NH_4Br \rightarrow NaBr + NH_3 + Si/H + H_2$, where the desired product, the hydrogen-terminated Si NCs, is noted as Si/H. Once the surface of the Si QDs is hydrogen passivated, further functionalization via photoinitiated hydrosilation can be performed.

Si QDs can also be obtained using electrochemical etching [5]. The process starts with placing p-type boron-doped Si wafers with (100) orientation and 5–20 Ω-cm resistivity in a mixture of hydrofluidic acid (HF), methanol, hydrogen peroxide (H_2O_2), and polyoxometalates (POMs), with the latter two functions as catalysts. At high current density (>10 mA/cm^2) and an etching time of several hours, typical etched surface structures are micropores of diameter around 2 μm, while at low current density (<10 mA/cm^2) and short etching time, nanoscale porous surface is formed. Ultrasonication is used to disintegrate the etched wafer into colloidal Si QDs. Prior to this step, hydrosilylation reaction with chloroplatinic acid as catalyst can result in an octane-modified Si surface, which prevents aggregation of the QDs.

QD Deposition Process

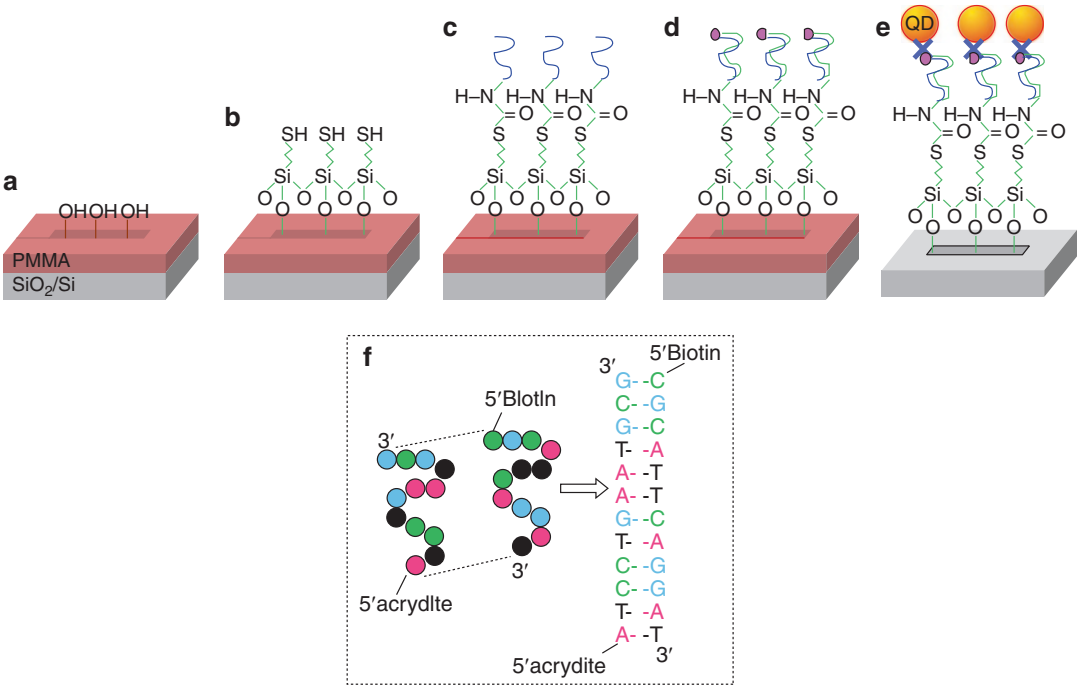
Once the QDs are synthesized, they need to be positioned onto a substrate for further processing into nanophotonic devices. This can be achieved by various methods. In this section, a few examples are listed: Deoxyribonucleic acid (DNA)-directed self-assembly, APTES-mediated molecular self-assembly, solution-based drop-casting, and electrostatic layer-by-layer deposition.

DNA-Directed Self-assembly

The process flow to assemble QDs through the DNA-mediated method is illustrated in Fig. 2 [6]. The area for QD deposition is defined in a layer of polymethylmethacrylate (PMMA) on an oxidized Si substrate using electron-beam lithography (EBL). The sample is then treated with oxygen plasma to generate hydroxyl (−OH) groups, which covalently bind to the dangling ends of silicon oxide and act as hooks for the next chemical reaction. Next, the first self-assembled monolayer (SAM) is formed by gas phase deposition of an organo-silane molecule, 3′mercaptotri-methoxysilane (MPTMS). The silane terminal anchors to the oxygen element of the hydroxyl groups, thus freeing the hydrogen as an energy minimization step while the mercapto, or thiol (SH), group becomes oriented to the outer surface.

The next step is to deposit the second monolayer, composed of a 12 base pair (5′acrydite-ATCCTGAATGCG-3′) chain of DNA terminated by acrylamide on the sample in solution form. This often involves diluting the base DNA with phosphate buffer solution (PBS) and $MgCl_2$. Immersing the sample in buffered acrylic acid followed by PBS rinse leads to passivation of inactivated MPTMS molecules and readies the sample for addition of the complementary DNA (cDNA). Biotin-linked cDNA is used to establish the third SAM. Now, the sample is ready for QD deposition. Streptavidin-bound QDs are combined in PBS and dropped on the surface. The QD assembly step utilizes the well-known biological recognition avidin-biotin binding pair where streptavidin, which is a protein, has four sites that preferentially attract and bind to biotin molecules. The last step is to remove the PMMA layer using toluene submersion.

One chief advantage of DNA-directed self-assembly approach is simultaneous deposition of different types of QDs to form a diverse range of



Quantum Dot Nanophotonic Integrated Circuits, Fig. 2 DNA-directed QD deposition process: (**a**) EBL pattern PMMA-coated substrate; (**b**) deposit MPTMS monolayer; (**c**) covalently bind with 5′acrydite-DNA; (**d**) hybridize with biotin-modified cDNA; (**e**) bind streptavidin-QDs to biotin-cDNA sites, remove PMMA; (**f**) detail of DNA chains as hybridized

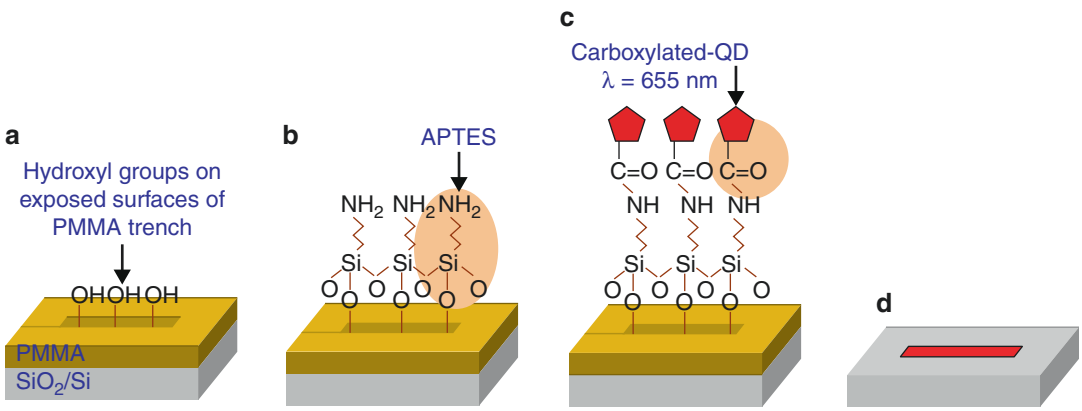
QD structures through specific DNA-cDNA binding recognition. As streptavidin-conjugated structures will preferentially bind to biotin, the critical step is to prepare the QDs with the corresponding biotin-cDNA molecules before reacting them en masse to the base DNA on the substrate. The details of the procedure are much like before, but a few additional steps are required. In particular, use of EBL is needed to open up each of the sections upon which different QDs will be bound. Consequently, after the initial EBL round, the substrate would be treated with MPTMS followed by the passivation buffer and the first acrydite-terminated DNA. With adequate concentration and exposure time, the packing of the base layer with quench molecules will be dense and prevent assembly of subsequent DNA strands. The next EBL writing creates new patterns that are developed and the second round of MPTMS and acrydite-DNA sequence SAMs are added. The process is repeated for as many QD types, and hence DNA-cDNA pairings, that are desired. After the final set of base DNA is added and prior to assembly with the complementary strands, the biotin-cDNA chains are mixed with the matching QD solutions specific to the device design. Subsequently, the QD-cDNA solutions are deposited on the sample as a mass self-assembly step.

APTES-Mediated Molecular Self-assembly

Similar to the DNA-directed technique, the APTES-mediated molecular self-assembly method (Fig. 3) starts with oxygen plasma treatment of an oxidized Si substrate following EBL. Then, the process diverges with sample immersion into 3′aminopropyltriethoxysilane (APTES) thoroughly mixed in an isopropyl alcohol (IPA) and deionized water solvent. In contrast to MPTMS, the chemisorption of APTES leads to an amine ($-NH_3$) rather than a mercapto ($-SH$) group presented at the terminal. Following this, a droplet of carboxylated QDs mixed in deionized water with 1-ethyl-3-(3′dimethylaminopropyl)-carbodiimide (EDC), a coupling reagent that aids in binding amine and carboxyl groups, is deposited directly on the patterned region.

Solution-Based Drop-Casting

Drop-casting is a method previously used in large-area thin-film devices as a convenient and cost-effective means of making a QD thin film [7]. The method involves dispersing a concentration of QDs in a solvent and then micropipetting a liquid droplet onto a substrate. As the solvent evaporates, the QDs pack together to form a thin membrane on the surface of the substrate. The resultant thin film forms a nanocrystal solid while retaining some of the unique optoelectronic properties of



Quantum Dot Nanophotonic Integrated Circuits, Fig. 3 APTES-mediated molecular self-assembly of QDs: (**a**) EBL pattern a SiO_2/Si sample coated with PMMA; develop PMMA; treat with oxygen plasma to create hydroxyl groups on surface; (**b**) solution phase deposition of APTES; (**c**) carboxylated QDs covalently bind to the amine groups of APTES; (**d**) strip PMMA to leave the patterned QDs region

the individual QDs. The quality and thickness of the QD thin film is determined by several factors, including the type of solvent used, the concentration of QDs, the liquid volume of the droplet, the hydrophobicity of the substrate, the type of atmosphere, and the drying temperature. To achieve high-quality nanoscale thickness, using a mixture of hexane/octane as the solvent often achieves a better result. For larger scale thin-film photodetectors, the most common solvent is toluene. Typically, commercial QDs are also suspended in toluene. The hexane/octane mixture prevents ring formations on the resultant film, in part due to the higher rate of evaporation. A high concentration of surface ligands, which prevent the QDs from clumping together, can also result in an opaque or "dirty" thin film. To prevent this, the QDs can be washed by precipitating out of the host solvent with a reagent-like ethanol or methanol. The ethanol acts to strip the organic capping ligands from the shell of the QDs. Once the solution is centrifuged, the supernatant liquid is poured off and the QDs are allowed to dry in a desiccator. The dried QDs can then be redispersed in a solvent of choice. For example, commercially available QDs originally suspended in toluene can be washed and then resuspended in the hexane/octane mixture.

Electrostatic Layer-by-Layer Deposition

The deposition methods described previously achieves either a single layer of QDs or a QD film with unspecified thickness. To achieve controllable film thickness, electrostatic layer-by-layer deposition of QDs has been developed [8]. The patterned areas are pre-covered with a monolayer of charged molecules which serve as traps to capture the QDs by electrostatic bonding. Utilizing a similar electrostatic process, the QDs carrying opposite charges are deposited alternately, forming a multilayer film. By virtue of the layer-by-layer self-assembly, the film thickness can be controlled by the number of layers. Such fabrication process can be performed on a variety of substrates once their surfaces are modified with dangling hydroxyl groups, including silicon oxide and flexible substrates.

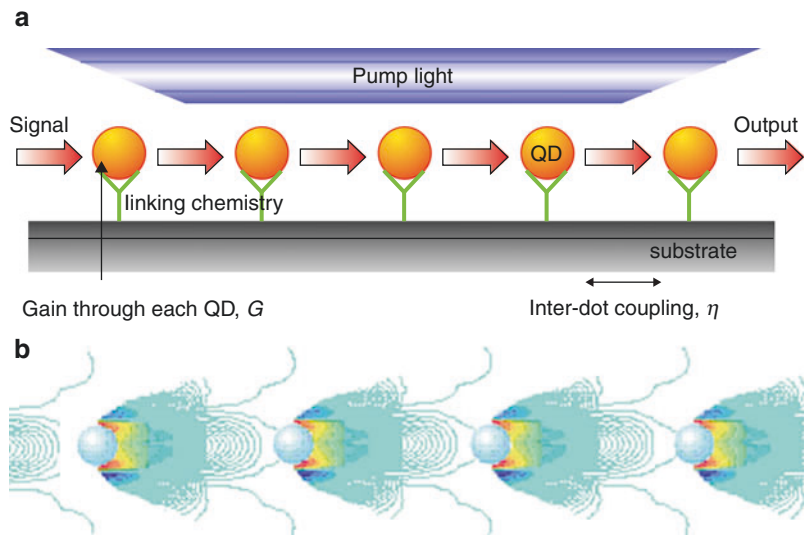
The length of ligands on the QD surface is critical to achieving short tunneling barrier and therefore high carrier mobility in the QD thin film. Extremely short ligands can be achieved using 2-mercaptoethylamine (MA) for positive capping and thioglycolic acid (TGA) for negative capping. The lengths of MA (∼0.4 nm) and TGA (∼0.3 nm) are short enough such that adjacent QDs' surfaces almost contact each other. TGA and MA also work as negative and positive thiol-stabilizers during CdTe QD synthesis. The precursors are formed in the ratio of Cd^{2+}: Te^{2-}: thiol stabilizer = 1: 0.5: 2.4 under appropriate pH conditions (11.5 for TGA and 5.75 for MA). The nanocrystals are grown by refluxing the reaction mixture at 100 °C and the particle size is controlled by refluxing time.

To prepare the substrate for electrostatic layer-by-layer deposition, an oxide-covered substrate is treated with oxygen plasma and dipped into NaOH to increase the density of hydroxide on the surface. Subsequently, the sample is immersed in (3-Aminopropyl)triethoxysilane (APTES) solution, followed by protonating the amines in diluted HCl solution for a few seconds. Upon this step, a high area density of positively charged amines on the surface of the substrate can be achieved. The layer-by-layer deposition of QDs is performed by immersing the chip first into TGA (negative)-CdTe QDs solution, then MA (positive)-CdTe QDs solutions. After washing away excess QDs with deionized water, each time of self-assembly will result in deposition of a monolayer of CdTe QDs. Repeating this process leads to a thin film of QDs with a controlled number of layers.

QD Nanophotonic Waveguide

The quantum dot waveguide consists of an array of densely packed QDs anchored to a substrate (Fig. 4a) [6]. A pump light illuminates the sample from above to create excited electron–hole pairs within the QD, which are then triggered by a signal light placed at the waveguide edge to recombine and emit a photon. The stimulated emission of photons occurs in a cascade fashion

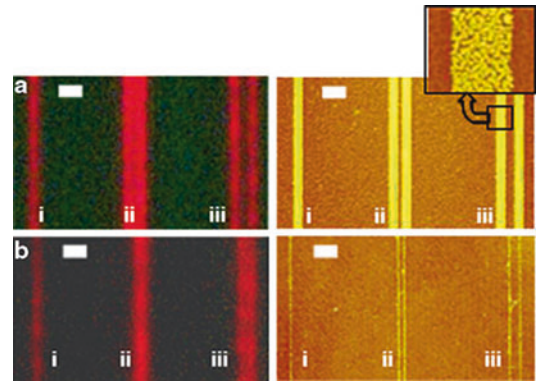
Q



Quantum Dot Nanophotonic Integrated Circuits, Fig. 4 (**a**) Schematic drawing of the QD nanophotonic waveguide. (**b**) Poynting vector distribution within the QD waveguide

downstream to generate a response at the output edge. The stimulated emitted photons are identical to the input signal photons in terms of wavelength, polarization, phase, and propagation direction. The overall transmission efficiency at the end of the waveguide depends on the inter-QD coupling efficiency and the gain available in each QD. To avoid wavelength mixing, the pump light energy is set at a higher level than the signal light energy. The finite-difference time-domain (FDTD) simulation of the Poynting vector distribution, representing the power flow and indicating photon transmission behavior in a QD waveguide, is shown in Fig. 4b. Upon absorption of the pump light and triggering of the signal light, the accepting QD produces energy as photons via stimulated emission, with the majority of energy aimed in the propagation direction to interact with the neighboring QD, causing a cascade of energy transmission.

Fabrication of the QD nanophotonic waveguides has been accomplished through either the DNA-directed or the APTES-mediated molecular self-assembly fabrication described earlier. Fluorescence micrographs and atomic force microscope (AFM) images, in Fig. 5a, b, show attachment of 655 nm emission QDs aligned in the 100 and 500 nm width lines. An additional AFM image is shown as an inset in Fig. 5a to provide more details on the QD shape and



Quantum Dot Nanophotonic Integrated Circuits, Fig. 5 Fluorescence and corresponding AFM images of (**a**) QD waveguide with 500 nm width; the *inset* shows granular QD shape; and (**b**) 100 nm-wide QD waveguides in (*i*) single and pair formations spaced, (*ii*) 200 nm, and (*iii*) 500 nm apart. Scale bar is 1 μm in length

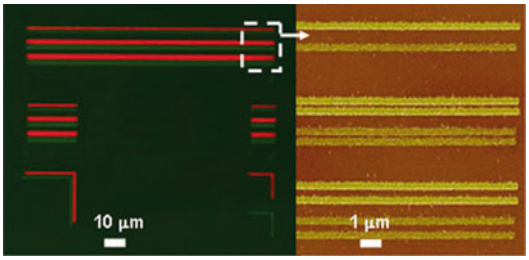
distribution within the waveguide. Measurement results on straight waveguides and waveguides with 90° bend demonstrated low-loss light transmission in sub-diffraction limit scale [6].

By repeating the fabrication process of the APTES-mediated self-assembly process, it is possible to create waveguides using QDs of different emission wavelengths to operate at targeted transmission wavelengths. Figure 6 shows the fluorescence microscope and AFM images of multiple QD waveguides with red (655 nm) and green (565 nm) emission wavelengths. The AFM

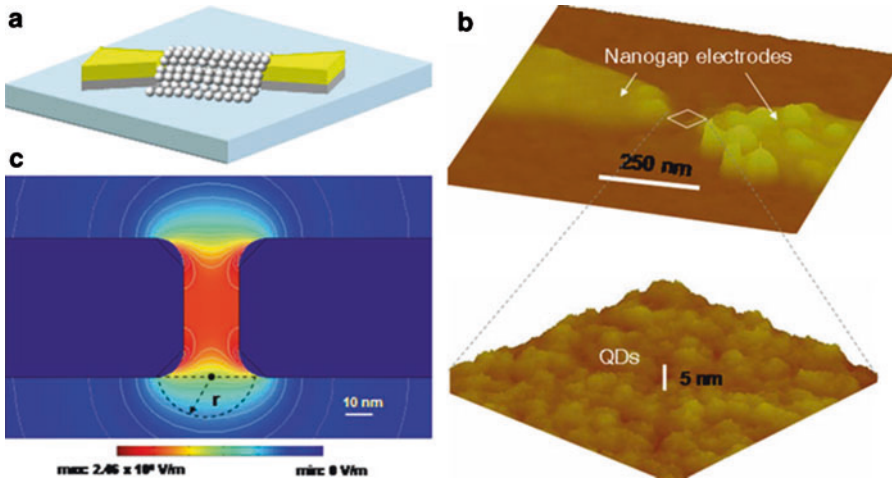
images represent the topological profile of the sample, showing well-packed monolayer coverage of QDs. The height is taller for the red patterns, corresponding to larger nanoparticles, as demonstrated by the brighter coloring consistent within the confines of the line.

Nanoscale QD Photodetector

As the integration density of photonic integrated circuits progresses toward the submicron regime, it becomes imperative to detect the optical signal with nanometer resolution as well in order to preserve the on-chip integration density. The requisite photodetection devices are critical for integration of nanophotonics with electronics. The devices should also have high fabrication and integration flexibility with other nanophotonic devices. This can be achieved by a nanoscale QD photodetector with a simple configuration schematically depicted in Fig. 7a [9]. Colloidal QDs are placed between two electrodes spaced with a nanogap, so that electrons can tunnel between the electrodes through the QDs. Electron transport in colloidal QDs embedded in a metal-QD-metal structure typically involves significant potential barriers at the QD–QD interface as well as the QD–electrode interface. Transport between QDs can, therefore, be considered as a problem of hopping between localized states. Carriers with higher energies have a higher probability of tunneling through the QDs and contributing to the tunneling current. The effect of optical excitation on the electrical transport of such structures is to reduce the effective potential barrier from the QD to the electrode by exciting photo-generated carriers into higher energy states. The tunneling current increases when the QDs are photo-excited, thus achieving photodetection.



Quantum Dot Nanophotonic Integrated Circuits, Fig. 6 Fluorescence micrograph (*left*) and AFM image of zoomed-in region (*right*) of multiple QD-type waveguides. The emission wavelengths are 655 and 565 nm, corresponding to larger and smaller QDs. The different colors in the AFM image demonstrate the height difference between the two types of QDs



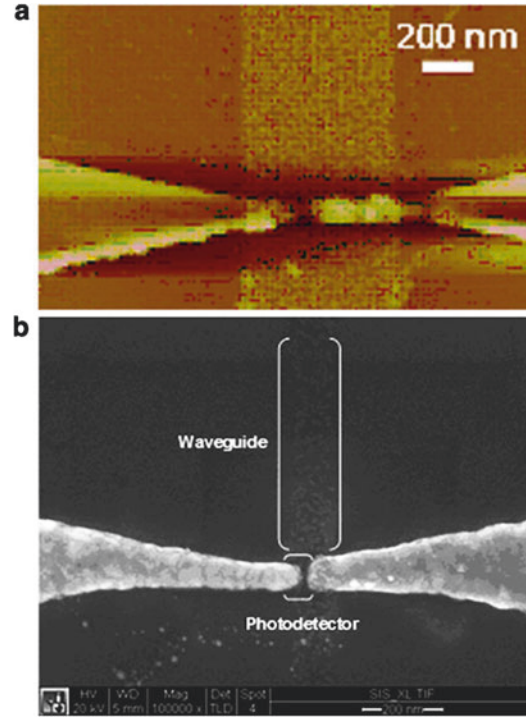
Quantum Dot Nanophotonic Integrated Circuits, Fig. 7 (**a**) Illustration of the nanoscale QD photodetector device structure. (**b**) AFM image with three-dimensional rendering of the device. (**c**) Finite element simulation of the electric field intensity in and around a 20 nm nanogap

The metal electrodes, spaced at nanometers apart, can be fabricated using electron-beam lithography (EBL). The QDs can be attached to the substrate through the various deposition methods described earlier. The fabrication process of the QD photodetectors consists of two simple steps: patterning of the metal electrodes and attachment of QDs. Both steps, either in sequence or separated, can be readily integrated with the fabrication of a wide range of nanophotonic components and devices. Figure 7b shows an AFM image in three-dimensional view of a device, and zoom-in of the QD film in the nanogap region.

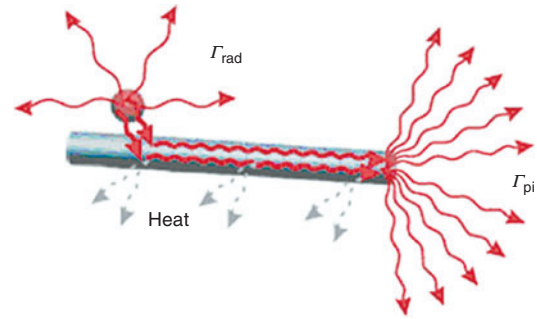
Electron transport through the device is limited by the field-induced ionization rate of the photogenerated excitons, which in this case is proportional to the electron-tunneling rate between neighboring nanocrystals. Thus, for high responsivity, it is necessary to have high field intensities across the junction, which can be obtained most easily in a nanoscale gap. Figure 7c shows finite element simulation of the DC electric field distribution in the gap that is highly concentrated in and around the nanogap. The short distance between electrodes limits the number of tunneling steps required for electrons to traverse the gap, which also increases responsivity and enhances the speed of the device.

QD photodetectors are expected to achieve high sensitivity due to the high quantum efficiency of the nanocrystals. To achieve high-speed operation of the devices remains challenging, due to the tunneling barriers between QDs. Less than 500 f. light detection sensitivity and >125 kHz modulation bandwidth have been reported for the nanoscale QD photodetectors [9].

Integrated fabrication of the nanoscale QD photodetector with the QD waveguides is straightforward due to the nature of self-assembly deposition of QDs. The fabrication result in Fig. 8a shows the AFM topology of a 500 nm-wide QD waveguide aligned to the center of metal electrodes created by a break-junction process spaced approximately 22 nm apart. Figure 8b shows a SEM image of a 50 nm-wide QD waveguide aligned to the center of metal electrodes created using EBL spaced about 12 nm apart.



Quantum Dot Nanophotonic Integrated Circuits, Fig. 8 (**a**) AFM image of a 500 nm-wide QD waveguide integrated with a 22 nm break-junction nanogap electrode. (**b**) SEM image of a 50 nm-wide QD waveguide integrated with a nanogap electrode spaced 12 nm apart



Quantum Dot Nanophotonic Integrated Circuits, Fig. 9 The emission from a single QD can be coupled to a plasmonic nanowire, creating a single plasmon source [10]

Quantum Dot – Plasmonic Waveguide Integration

Plasmonic structures are nanoscale metallic structures that can concentrate light into deep-subwavelength volumes. Therefore, they provide

a promising approach to interface diffraction-limited optical components with nanophotonic structures. Integration of a single QD with a plasmonic waveguide has been explored [10]. A CdSe QD is placed in close proximity to a silver nanowire (Fig. 9). When optically excited, the emission from the QD couples directly to guided surface plasmons in the nanowire, causing the end of the wire to light up and acting as a source of single, quantized plasmons. This approach can realize strong nonlinear interactions at the single-photon level, making a single-photon transistor possible [11].

Cross-References

- ▶ Active Plasmonic Devices
- ▶ AFM
- ▶ Atomic Force Microscopy
- ▶ Electron Beam Lithography (EBL)
- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Finite-Difference Time-Domain Technique
- ▶ Nanoparticles
- ▶ Nanostructured Functionalized Surfaces
- ▶ Nanostructures for Photonics
- ▶ Nanotechnology
- ▶ Scanning Electron Microscopy
- ▶ Self-Assembled Monolayers
- ▶ Self-Assembly

References

1. Asada, M., Miyamoto, Y., Suematsu, Y.: Gain and the threshold of three-dimensional quantum-box lasers. IEEE J. Quantum Electron. **22**, 1915–1921 (1986)
2. Bruchez Jr., M., Moronne, M., Gin, P., Weiss, S., Alivisatos, A.P.: Semiconductor nanocrystals as fluorescent biological labels. Science **281**, 2013–2016 (1998)
3. Murray, C.B., Kagan, C.R., Bawendi, M.G.: Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. Annu. Rev. Mater. Sci. **30**, 545–610 (2000)
4. Neiner, D., Chiu, H.W., Kauzlarich, S.M.: Low-temperature solution route to macroscopic amounts of hydrogen terminated silicon nanoparticles. J. Am. Chem. Soc. **128**, 11016–11017 (2006)
5. Tu, C.-C., Tang, L., Huang, J., Voutsas, A., Lin, L.Y.: Solution-processed photodetectors from colloidal silicon nano/micro particle composite. Opt. Express **18**, 21622–21627 (2010)
6. Lin, L.Y., Wang, C.-J.: Quantum dot nanophotonic waveguides. In: Ohtsu, M. (ed.) Nanophotonics and nanofabrication, pp. 215–240. Wiley, Weinheim (2009)
7. Konstantatos, G., Sargent, E.H.: Solution-processed quantum dot photodetectors. Proc. IEEE **97**, 1666–1683 (2009)
8. Tu, C.-C., Lin, L.Y.: High efficiency photodetectors fabricated by electrostatic layer-by-layer self-assembly of CdTe quantum dots. Appl. Phys. Lett. **93**, 163107 (2008)
9. Hegg, M.C., Horning, M.P., Baehr-Jones, T., Hochberg, M., Lin, L.Y.: Nanogap quantum dot photodetectors with high sensitivity and bandwidth. Appl. Phys. Lett. **96**, 101118 (2010)
10. Akimov, A.V., Mukherjee, A., Yu, C.L., Chang, D.E., Zibrov, A.S., Hemmer, P.R., Park, H., Lukin, M.D.: Generation of single optical plasmons in metallic nanowires coupled to quantum dots. Nature **450**, 402–406 (2007)
11. Chang, D.E., Sørensen, A.S., Demler, E.A., Lukin, M.D.: A single-photon transistor using nanoscale surface plasmons. Nat. Phys. **3**, 807–812 (2007)

Quantum Dot Solar Cells

- ▶ Nanomaterials for Excitonic Solar Cells

Quantum Dot Toxicity

Wafa' T. Al-Jama1 and Kostas Kostarelos
Nanomedicine Laboratory, Centre for Drug Delivery Research, The School of Pharmacy, University of London, London, UK

Synonyms

Semiconductor nanocrystals

Definition

Semiconductor quantum dots (QD) are 1–10 nm nanocrystals consisting mainly of a semiconductor core with or without an inorganic passivation shell

usually made of zinc sulfide (ZnS) [1]. The shell is generally used to protect the QD core from oxidation and photolysis and to minimize the associated toxicity related to the release of Cd^{2+} ions.

Currently QD are not only composed of CdS or CdSe but of many different semiconducting materials. These are derived from the II–VI elemental groups (e.g., zinc sulfide [ZnS], zinc selenide [ZnSe], and cadmium telluride [CdTe]), or the III–V elemental groups (e.g., indium phosphide [InP], indium arsenate [InAs]), or the IV–VI elemental groups (e.g., lead selenide [PbSe], lead sulfide [PbS]). In addition to the huge advances in semiconductor synthesis, novel QD such as CdTe/CdSe (core/shell), CdSe/ZnTe (core/shell), cadmium-free QD, and Mn-doped ZnSe have also been developed [2].

In biological sciences QD exhibit great potential as fluorescent probes for optical imaging and diagnostic applications in vitro and in vivo. Due to their bright fluorescence, pronounced photostability and broad excitation spectra, simultaneous detection of QD in different colors has enabled multiplexed imaging for tracking cancer cell metastasis and identification of tumor nodules in vivo [3–6].

QD Solubility and Ligation

QD can be prepared in either water-based or organic solvent environments. The latter approach produces uniformly dispersed QD with a range of emission colors ranging from ultraviolet to infrared. However, QD synthesized in apolar solvents contain organic shells that compromise their water solubility and consequently their compatibility with the biological milieu. Many strategies have been developed to overcome this limitation based on the surface ligation of QD with hydrophilic groups. Several hydrophilic ligands have been utilized to exchange the hydrophobic trioctylphosphine oxide (TOPO) coat on the QD surface with hydrophilic moieties, including: (1) thiol-containing molecules, such as mercaptoacetic acid (MAA), dihydrolipoic acid (DHLA) and mercaptopropyltris (methoxy) silane (MPS); (2) peptides; (3) dendrons; and (4) oligomeric phosphine. In spite of maintaining small mean diameters, surface modification methods generally tend to cause QD aggregation and decrease the efficiency of their optical properties (e.g., fluorescence intensity). Moreover, when labile ligands detach from the QD surface, QD-induced toxicity will increase correspondingly due to exposure to the QD core.

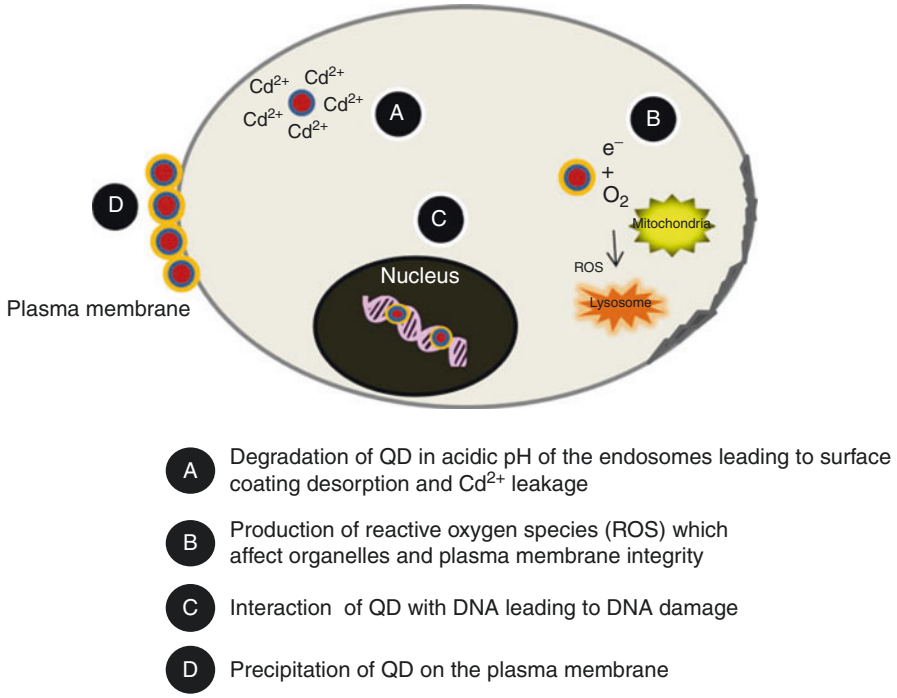
An alternative approach to make QD water dispersible and therefore more biocompatible is based on the hydrophobic interactions between the organic QD surface and the self-assembly of various amphiphilic molecules. Embedding or encapsulating organic QD into phospholipid micelles, lipid bilayers, and amphiphilic diblock or triblock copolymers was found to significantly increase QD diameter, but more importantly to preserve QD photostability, colloidal stability and enable QD application to physiological conditions without the release of the toxic Cd^{2+} ions.

QD Toxicity In Vitro

The toxicity of QD is mainly derived from their intrinsic core composition such as CdSe and CdTe [7, 8]. Cd^{2+} ions have been shown to be toxic upon their release from the QD core due to photolysis and/or oxidation (Fig. 1A). Other mechanisms contributing to QD-induced cytotoxicity have also been identified including: the formation of reactive oxygen species (ROS) that induce cell damage [9] (Fig. 1B) and the interaction of QD nanoparticles with the individual cell components (e.g., DNA or proteins) (Fig. 1C) or with the cell membrane (Fig. 1D).

QD-induced cell dysfunction is accompanied by apoptotic and necrotic biochemical changes including: morphological alteration in the plasma membrane; mitochondria and nucleus damage; lysosome enlargement; reduction in cytochrome C concentration; loss of mitochondrial membrane potential; and upregulation of peroxidized lipids [10].

The correlation between cytotoxicity and free Cd^{2+} ions has been established with the occurrence of significant cell death in the range of 100–400 μM Cd^{2+} ions. The blue shift in QD



Quantum Dot Toxicity, Fig. 1 Mechanism of QD toxicity

fluorescence spectrum has been proposed as an indication of QD size reduction and Cd^{2+} release. In addition, QD-induced cytotoxicity dramatically increased in the case of QD exposure to oxygen or ultraviolet (UV) light. In fact the principle of QD phototoxicity has been exploited for the development of photodynamic therapeutics [6].

Several attempts to reduce the inherent QD toxicity have been described. For instance, ZnS coating protects the QD core from oxidation, which minimizes Cd^{2+} leakage and reduces the QD-induced cytotoxicity. In addition, the use of antioxidants, such as *N*-acetylcysteine (NAC) has been shown to be effective in reducing QD cytotoxicity. To date, most cytotoxicity studies have used QD solubilized by direct exchange of the organic coat (TOPO) with hydrophilic ligands, such as mercaptopropionic acid (MPA-QD), mercaptoacetic acid (MAA-QD), mercaptoundecanoic acid (MUA-QD), cysteamine (QD-NH_2), and thioglycerol (QD-OH). In such cases, ligand detachment from the QD surface may also occur due to the weak interaction between the QD surface and the ligands, especially under unfavorable conditions like those in the endosomal compartment. Several studies have indicated that cell incubation with QD solubilized by ligand surface exchange has mostly been associated with severe cell death with increasing QD concentrations, similar to that induced by Cd^{2+} ions, which indicates the need for more stable QD. In contrast, lower cytotoxicity has been observed with polymer-coated QD.

Overall, the in vitro toxicity studies carried out so far have shown that the key determinants of QD toxicity are the nanocrystal composition and the type of surface functionalization. However, other factors including: cell type, QD size, concentration, and QD exposure to oxygen and UV light have also been found to influence QD cytotoxicity.

QD Toxicity In Vivo

The concern about the potential toxicity of QD in vivo is mainly due to the well-established toxicological profile of cadmium, the physicochemical

characteristics of certain QD surface coatings, and the size of QD. In general, the small diameter of nanoparticles provides a large available surface for interaction with different biological molecules. Moreover, nanocrystals of a few nanometers in size, such as QD, can enter vital organs such as the heart, lung, liver, and brain after intravenous administration. Therefore, studies that will help determine the toxicological risk and tolerated exposure levels of QD in vivo will be needed for any type of QD developed for biomedical use.

It has been previously described that QD nanocrystals and other nanoparticles could be inhaled and deposited in the respiratory tract, mainly depending on their size [11, 12]. It has also been shown that QD are capable of penetrating intact skin [13]. These preliminary studies suggest that extra care should be taken during QD handling especially during their manufacturing, since systemic toxicity can occur via QD inhalation and direct skin contact.

Preclinical research has shown that QD can be used to visualize lymph nodes and solid tumors in vivo following local or systemic administration. Many studies have also shown that intravenously administered QD mainly reside in the reticuloendothelial system tissues (liver, spleen, and lymph nodes) with no degradation or excretion for up to 2 years; however, no clinical signs of toxicity have been reported. More recent studies have shown that only QD of 5–6 nm in diameter, which is below the renal filtration threshold, were found to be excreted via the urine 4 h post-injection, while QD of larger diameters undesirably remained in the liver up to a few months, which may increase the potential toxicological risks from these nanoparticles in the long term. Interestingly, a recent study has reported that the QD dose in combination with the surface coating and charge play an important role in QD toxicity in vivo. For instance, carboxylated QD injected at a high dose in mice can activate blood coagulation and induce thrombus formation in the lung that can be lethal. Such studies further emphasize the need to investigate the effect of surface coatings on QD toxicity.

In conclusion, QD offer tremendous potential based on their optical properties. Even though a lot of interest and improvements have been achieved in terms of aqueous dispersion and compatibility in physiological environments, there is an inherent risk associated with exposure to QD nanocrystals, in particular those consisting of cadmium. Since the most meaningful clinical translation of QD technology is as probes for biomedical imaging, the acceptable toxicological risks should be minimal or nonexistent. This seems the most formidable challenge that QD technology has to overcome, because as can be seen from the available pharmacological studies, QD can reside in different organs when injected in living animals depending on various QD characteristics (size, dose, surface charge, and coating) and the route of administration. Furthermore, such studies have indicated that QD can accumulate in some tissues for extensive periods of time which requires further investigation to identify the long-term toxicological risks of QD before embarking on any clinical use of QD. In addition, even if QD are excreted from the body the toxicity profile of these nanocrystals on the environment (water and soil) should also be investigated.

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles
- ▶ Genotoxicity of Nanoparticles
- ▶ Nanoparticle Cytotoxicity

References

1. Rogach, A.L.: Semiconductor Nanocrystal Quantum Dots: Synthesis, Assembly, Spectroscopy, 1st edn. Springer, New York (2008)
2. Nagarajan, R., Hatton, T.A.: Nanoparticles Synthesis, Stabilization, Passivation and Functionalization. Oxford University Press, New York (2011)
3. Hotz, C.Z., Bruchez, M.: Quantum Dots: Applications in Biology. Humana Press, New York (2007)
4. Osinski, M., Yamamoto, K., Jovin, T.M.: Colloidal Quantum Dots for Biomedical Applications II. SPIE, San Jose (2007)
5. Pinaud, F.: Peptide-Coated Quantum Dots: Applications to Biological Imaging of Single Molecules in

Live Cells and Organisms. VDM Verlag Dr. Müller, Saarbrücken (2009)
6. Kuma, C.S.S.R.: Semiconductor Nanomaterials. Nanomaterials for Life Sciences. Wiley (VCH), Weinheim (2010)
7. Jin, Y., Zhao, X.: Cytotoxicity of photoactive nanoparticles. In: Webster, T.J. (ed.) Safety of Nanoparticles – From Manufacturing to Medical Applications, pp. 19–29. Springer, New York (2009)
8. Sahu, S.C., Casciano, D.A.: Nanotoxicity: From In Vivo and In Vitro Models to Health Risks. Wiley, Chichester (2009)
9. Nadeau, J.: Quantum dot reactive oxygen species generation and toxicity in bacteria: mechanisms and experimental pitfalls. In: Callan, J., Raymo, F. (eds.) Quantum Dot Sensors: Technology and Commercial Applications. Pan Stanford Publishing, Singapore (2011)
10. Dave, S.R., White, C.C., Kavanagh, T.J., Gao, X.: Luminescent quantum dots for molecular toxicology. In: Balls, M., Combes, R., Bhogal, N. (eds.) New Technologies for Toxicity Testing. Springer, New York (2011)
11. Gehr, P., Mühlfeld, C., Rothen-Rutishauser, B., Blank, F.: Particle-Lung Interactions, (Lung Biology in Health and Disease). Informa Healthcare, London (2009)
12. Marijnissen, J.C., Gradon, L.: Nanoparticles in Medicine and Environment: Inhalation and Health Effects. Springer, Berlin (2009)
13. Ryman-Rasmussen, J.P., Riviere, J.E., Monteiro-Riviere, N.A.: Penetration of intact skin by quantum dots with diverse physicochemical properties. Toxicol. Sci. **91**(1), 159–165 (2006)

Quantum Dynamics

► Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Quantum Kinetics

► Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism

Quantum Nanoplasmonics

► Quantum System Near Metallic Particle

Quantum System Near Metallic Particle

C. H. Raymond Ooi
Department of Physics, University of Malaya, Kuala Lumpur, Malaysia

Synonyms

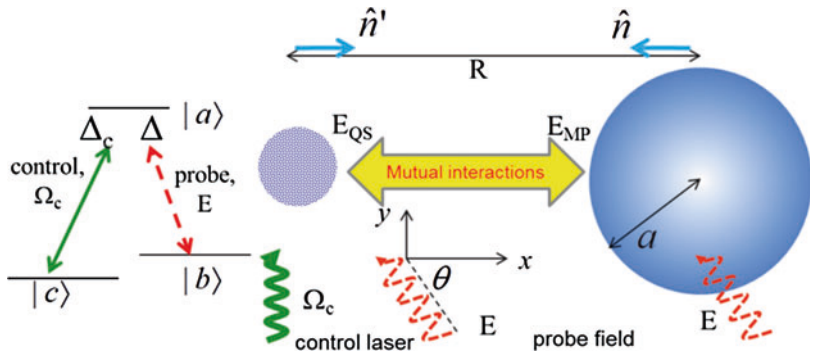
Quantum nanoplasmonics; Surface plasmon resonance

Definition

The optical properties of quantum system including atoms, molecules, and quantum dots can be drastically modified by a metallic particle through surface plasmon resonance effect.

The presence of plasmonic metallic particle (MP) can affect the optical properties of a quantum system (QS) in a remarkable way. The mechanism here is the mutual effect of polarizations similar to the induced dipole moment through dipole-dipole interactions. This mechanism should be contrasted from the quantum electrodynamic effect due to the presence of boundaries such as cavity mirrors or microstructures that alter the electromagnetic density of states of the quantum vacuum. We develop a general quantum nonlinear formalism with exact vectorial description for the scattered photons by the QS. The formalism enables us to study the variations of the dielectric function and photon spectrum of the QS with the particle distance between QS and MP, exciting laser direction, polarization, and phase in the presence of surface plasmon resonance (SPR) in the MP. The quantum formalism also serves as a powerful tool for studying the effects of these parameters on the nonclassical properties of the scattered photons. The plasmonic effect of nanoparticles has promising possibilities as it provides a new way for manipulating quantum optical properties of light in nanophotonic systems.

Recent research interest is in the study of quantum dot-metallic nanoparticle system. Metallic



Quantum System Near Metallic Particle, Fig. 1 (Color online) Configuration of a quantum system (QS) and a metallic particle (MP) probed by a laser with field **E**. The internal states of the QS are described by a three-level system with *a*–*c* transition driven by a strong external control laser

nanoparticle can act as a nanoamplifier and pulse controller for quantum dot [1], as well as enhancement and suppression of emission between nanoparticles in the semiconductor-metal nanoparticle assemblies [2]. The presence of metallic surface can enhance the fluorescence emission rate of nanocrystal [3] and field localization through nanofocusing [4] and plasmonic nanoantenna in laser diode [5].

Interesting current research focuses on quantum plasmonics where surface plasmon resonance (SPR) can affect the strength of light-matter interactions. In return quantum coherence can affect the optical properties of the QS-MP hybrid system not only at near field but also at larger interparticle separation R or far field. To illustrate this remarkable effect, we consider the interaction between a quantum system (QS) and plasmonic metallic nanoparticle (MP) coupled through their own polarizations as shown in Fig. 1. QS is modeled by three-level atoms which exhibit electromagnetically induced transparency (EIT) of probe field in the presence of control laser. We analyze how the optical properties of the QS are affected by factors such as the distance R, laser parameters (frequency, direction, and polarization), and quantum coherence effects.

Plasmonic effect or surface plasmon resonance can be found in historical artifacts such as the famous Lycurgus cup and buildings such as colorful glass ceilings. Gold nanoparticles with various shapes and sizes give rise to different colors observed. Potential applications as chemical sensors and optical chip have attracted many researchers into this topic.

It is well known that metallic nanoparticles or microparticles generate quantized plasmon in the presence of electromagnetic radiation leading to nanoplasmonic effects with large intensity outside the metallic surface and surface mode with large and discrete wavevector. However, there are also important quantum effects in plasmonic system composed of atoms, molecules, or quantum dots, such as the Fano dip that has been explored [6]. Thus, the study will involve quantum theoretical formalism, particularly the density matrix $\hat{\rho}$ and the Liouville-von Neumann equation $\frac{d}{dt}\hat{\rho} = \left[\hat{H},\hat{\rho}\right]/i\hbar$.

The Hamiltonian of the QS with three-level system is

$$\hat{H} = \hat{H}_0 + \hbar\left(\Omega_b e^{-ivt}|a\rangle\langle b| + \Omega_c e^{-iv_c t}|a\rangle\langle c| + \text{adj}\right) \quad (1)$$

leading to a set of 3^2 coupled density matrix equations. The Rabi frequencies of the *effective* probe field $\mathbf{E}_{QS}$ and control field $\mathbf{E}_c$ are $\Omega_b = \frac{1}{\hbar}\wp_{ab}\cdot\mathbf{E}_{QS}$ and $\Omega_c = \frac{1}{\hbar}\wp_{ac}\cdot\mathbf{E}_c$, respectively. The *incident* probe laser field which depends on polarization angle ϕ and direction θ is given by

$$\mathbf{E}_p = E\left(\alpha\sin\theta, \alpha\cos\theta, \beta e^{i\phi}\right) \quad (2)$$

where θ gives the laser direction and $\alpha, \beta = \sqrt{1-\alpha^2}, \phi$ determines the ellipticity of polarization.

The three-level system is coupled by two fields: probe field and control field. The probability amplitude for the system to excite from level b to level a can interfere destructively in the presence of the control field, creating optical transparency window and preventing absorption. This is electromagnetically induced transparency which leads to transparency at resonant frequency ω_{ab} with two absorption peaks around $\omega_{ab} \pm \Omega_c$.

Exact expression of the electric field $\mathbf{E}_{dip}$ created by a dipole emitter with dipole moment $\mathbf{p}$, at any distance R, can be obtained from

$$\mathbf{E}_{dip} = \frac{K}{\varepsilon_{eff}} \left[A \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} + B \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} \mathbf{p} \cdot \hat{n} \right] \tag{3}$$

$$A(k,R) = \frac{k^2}{R} - \frac{1}{R^3} + \frac{ik}{R^2} \tag{4}$$

$$B(k,R) = \frac{3}{R^3} - \frac{3ik}{R^2} - \frac{k^2}{R} \tag{5}$$

$$\mathbf{p} \cdot \hat{n} = p_x n_x + p_y n_y + p_z n_z \tag{6}$$

where $\hat{n}$ is along the line from the dipole to the observation point, $K = \frac{1}{4\pi\varepsilon_0}$, k is the wavevector, and $\varepsilon_{eff} = \frac{2\varepsilon_b + \varepsilon_{MP}}{3\varepsilon_b}$ is the effective dielectric function that depends on that of the MP, ε_{MP}. The coefficients A and B are exact such that the theory is valid for arbitrary R.

The fields of both the QS and MP $\begin{pmatrix} E_{MPx} \\ E_{MPy} \\ E_{MPz} \end{pmatrix}$ affect each other leading to two coupled equations. The electric field at MP due to dipole in QS is

$$\begin{pmatrix} E_{MPx} \\ E_{MPy} \\ E_{MPz} \end{pmatrix} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} + \frac{K}{\varepsilon'_{eff}} \widetilde{\rho}_{ab} \left\{ A \begin{pmatrix} \wp_x^* \\ \wp_y^* \\ \wp_z^* \end{pmatrix} + B \begin{pmatrix} n'_x \\ n'_y \\ n'_z \end{pmatrix} \wp * \cdot \hat{n}' \right\} \tag{7}$$

$$= \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} + \frac{KM'}{\varepsilon'_{eff}} \widetilde{\rho}_{ab} \begin{pmatrix} \wp_x^* \\ \wp_y^* \\ \wp_z^* \end{pmatrix} \tag{8}$$

$$M' = \begin{pmatrix} A + Bn'_x 2 & Bn'_x n'_y & Bn'_x n'_z \\ Bn'_x n'_y & A + Bn'_y 2 & Bn'_y n'_z \\ Bn'_x n'_z & Bn'_y n'_z & A + Bn'_z 2 \end{pmatrix} \tag{9}$$

where $\varepsilon'_{eff} = \frac{2\varepsilon_b + \varepsilon_{QS}}{3\varepsilon_b}$. The effects of MP on a QS where the vectorial nature of the fields is included provide additional possibilities for manipulating the optical properties and mechanism of the QS by changing the direction, polarization, and phase of the probe laser. On the other hand, the electric field at QS due to polarizability of the MP is

$$\begin{pmatrix} E_{QSx} \\ E_{QSy} \\ E_{QSz} \end{pmatrix} = \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} + \frac{K}{\varepsilon_{\text{eff}}} \left\{ A \begin{pmatrix} p_{MPx} \\ p_{MPy} \\ p_{MPz} \end{pmatrix} + B \begin{pmatrix} n_x \\ n_y \\ n_z \end{pmatrix} \mathrm{p}_{MP} \cdot \hat{n} \right\}$$

$$= \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} + \frac{KM\alpha_{MP}}{\varepsilon_{\text{eff}}} \begin{pmatrix} E_{MPx} \\ E_{MPy} \\ E_{MPz} \end{pmatrix} \tag{10}$$

$$M = \begin{pmatrix} A + Bn_x^2 & Bn_x n_z & Bn_x n_z \\ Bn_x n_y & A + Bn_y^2 & Bn_y n_z \\ Bn_x n_z & Bn_y n_z & A + Bn_z^2 \end{pmatrix} \tag{11}$$

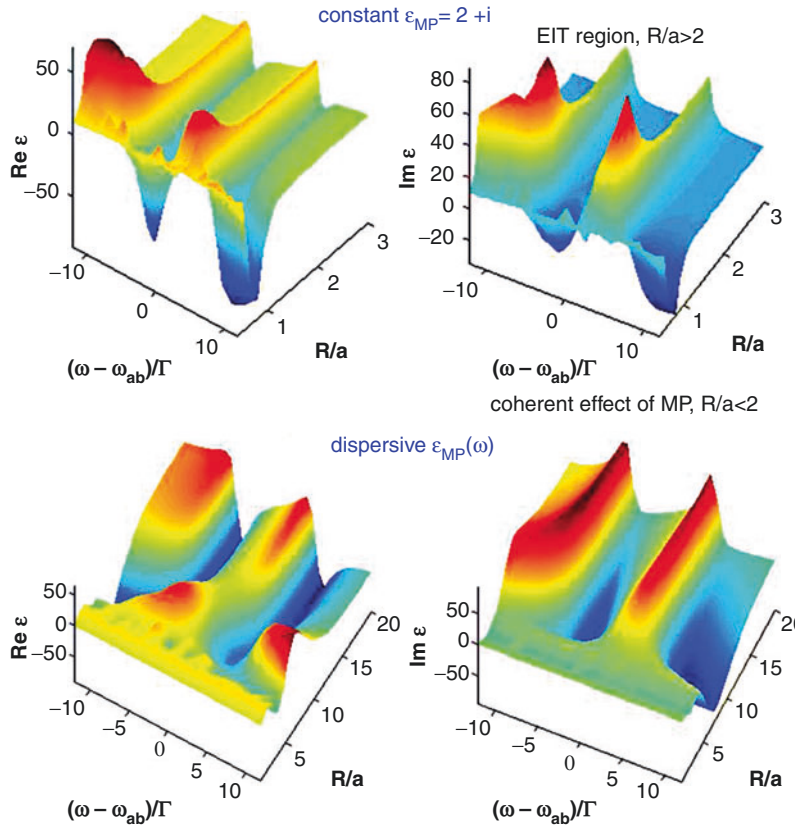
where we use $p_{MPj} = \alpha_{MP} E_{MPj}$ with polarizability α_{MP} and $j = x, y$. Solving the coupled equations gives the exact and general and vectorial local field description of the QS:

$$\begin{pmatrix} E_{QSx} \\ E_{QSy} \\ E_{QSz} \end{pmatrix} = \Lambda \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} + G\widetilde{\rho}_{ab} \begin{pmatrix} \wp_x^* \\ \wp_y^* \\ \wp_z^* \end{pmatrix} \tag{12}$$

$$\Lambda = I + \frac{\alpha_{MP} KM}{\varepsilon_{eff}}, G = \alpha_{MP} \frac{KM}{\varepsilon_{eff}} \frac{KM'}{\varepsilon'_{eff}} \tag{13}$$

The vector field $\begin{pmatrix} E_{QSx} \\ E_{QSy} \\ E_{QSz} \end{pmatrix}$ is coupled to the

Q



Quantum System Near Metallic Particle, Fig. 2 (Color online) The spectra for the real and imaginary parts of the dielectric function versus R, the spacing between the QS and MP for different dielectric function of the metallic particle. (**a**) Constant $\varepsilon_{MP} = 2 + i$ and (**b**) dispersive $\varepsilon_{MP}(\omega) = \varepsilon_\infty - \frac{\omega_p^2}{\omega(\omega+i\Gamma_m)}$ with SPR

density matrix equations through the Rabi frequency $\Omega_b = \frac{1}{\hbar}\wp_{ab} \cdot \mathbf{E}_{QS}$.

Solving for the density matrix element ρ_{ab} numerically gives the dielectric function $\varepsilon(\omega) = 1 + \chi$ with $\frac{\chi = \frac{N|\wp_{ab}|^2}{\varepsilon_0}\rho_{ab}}{\wp_{ab}\cdot\mathbf{E}_p}$ for the linear optical properties of the QS. It is important to note that Im $\varepsilon(\omega) > 0$ (< 0) corresponds to absorption (gain). Results for the near-field $kR \sim 1$ can be found in [7].

We can learn about the coherent quantum and plasmonic effects by considering the following cases shown in Fig. 2. In the case of constant or nondispersive MP, $\varepsilon_{MP} = 2 + i$ (with no SPR), we only have a small enhancement and the EIT effect diminishes at $R \tilde{} 2a$. When dispersion is introduced (such as in the case of Ag) and with the surface plasmon resonance (SPR) condition $\varepsilon_{MP}(\omega_{SP}) + 2\varepsilon_b \simeq 0$, the effective dielectric function $\varepsilon_{eff} = \frac{2\varepsilon_b + \varepsilon_{MP}}{3\varepsilon_b}$ is close to zero. Here, we have a large plasmonic enhancement to the field $\mathbf{E}_{QS}$, where $\varepsilon_{MP}(\omega)$ is the dielectric function of metal at SPR frequency ω_{SP}. Typically, for gold or silver nanoparticles, we have negative value in the real part of ε_{MP} at the frequency ω_{SP}. To see the plasmonic effects on the EIT profile, we set $\omega_{ab} = \omega_{SP}$. With the coexistence of the SPR and the EIT effect, we can see how the SPR affects the double-resonant peaks of the EIT. Figure 2 shows that the EIT peaks are affected at a much larger distance $R \sim 20a$, which is about one order of magnitude larger than the case without SPR. This is due to the plasmonic effect which gives rise to long-range Foerster interaction. At small R, the EIT effect is quenched and the optical properties of QS are drastically changed.

The presence of the MP gives rise to the coherent interactions between the QS and MP, producing gain when the imaginary part of the dielectric function takes an opposite sign.

This effect provides laser switching of quantum optical properties between the regimes of EIT, absorptive and gain. With SPR effect of the MP, the local field of the QS is significantly enhanced by many orders of magnitude. This effect has profound implications on the atom-field coupling which can reach nonlinear regime even with a few photons. It opens up the practicality of quantum nonlinear optics, whereby quantum effects can coexist with nonlinear effects at microscopic level. The model shown can be extended to many particle system or mesoscopic particles composed of the quantum particles. This enables the quantum plasmonics to be extended to mesoscopic objects such as nanoparticles and microparticles.

Cross-References

- ▶ Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures
- ▶ Quantum Dot

References

1. Sadeghi, S.M.: The inhibition of optical excitations and enhancement of Rabi flopping in hybrid quantum dot–metallic nanoparticle systems. Nanotechnology **20**, 225401 (2009)
2. Govorov, A.O., Bryant, G.W., Zhang, W., Skeini, T., Lee, J., Kotov, N.A., Slocik, J.M., Naik, R.R.: Exciton–Plasmon Interaction and Hybrid Excitons in Semiconductor–Metal Nanoparticle Assemblies. Nano Lett. **6**, 984 (2006)
3. Shimizu, K.T., Woo, W.K., Fisher, B.R., Eisler, H.J., Bawendi, M.G.: Surface-Enhanced Emission from Single Semiconductor Nanocrystals. Phys. Rev. Lett. **89**, 117401 (2002)
4. Kato, K., Ono, A., Inami, W., Kawata, Y.: Plasmonic nanofocusing using a metal-coated axicon prism. Opt. Express **18**(13), 13580 (2010)
5. Cubukcu, E., Kort, E.A., Crozier, K.B., Capasso, F.: Plasmonic laser antenna. Appl. Phys. Lett. **89**, 093120 (2006)
6. Zhang, W., Govorov, A.O.: Quantum theory of the nonlinear Fano effect in hybrid metal-semiconductor nanostructures: The case of strong nonlinearity. Phys. Rev. B **84**, 081405(R) (2011)
7. Raymond Ooi, C.H., Tan, K.S.: Controlling Double Quantum Coherence and Electromagnetic Induced Transparency with Plasmonic Metallic Nanoparticle. Plasmonics **8**(2), 891–898 (2013)

R

Radiofrequency

► Thermal Cancer Ablation Therapies Using Nanoparticles

Random Matrix Theory

► Harnessing Disorder at the Nanoscale

Rapid Atomic Processing (RAP)

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Rapid Electrokinetic Patterning and Its Applications

Katherine Clayton[1], Jian-Wei Khor[1] and Steven T. Wereley[2]
[1]Birck Nanotechnology Center, School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA
[2]School of Mechanical Engineering, Purdue University, Birck Nanotechnology Center, West Lafayette, IN, USA

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Synonyms

Electrokinetics; Hybrid opto-electric technique; Optoelectrokinetic Technique; Optofluidics

Definition

Rapid electrokinetic patterning (REP) is a hybrid optoelectric technique that concentrates and manipulates particles via the combination of a uniform alternating current (AC) electric field and laser light. Alone, the AC electric field induces electric double layer (EDL) polarization of the dielectric particles. Coupling the AC electric field with the laser light produces a toroidal electrothermal flow. This flow sweeps colloids suspended in the solution and brings them close to the electrode surface where particles are trapped by particle-electrode interactions.

Overview

The ability to concentrate and manipulate colloids on chip provides potential solutions in fields of micromanufacturing, disease diagnostics, biology, and plasmonics [1–3]. The ability to dynamically configure clusters of particles decreases time-to-assemble and manufacture structures, which is a continuing issue in many lab-on-a-chip (LOC)-based systems [4]. Many optical and

electrokinetic techniques have attempted to alleviate these issues however certain limitations still exist. Purely optical techniques, such as optical tweezers, allow for only small numbers of particles to be captured within the beam waist of a laser beam [5]. On the other hand, purely electric techniques, such as dielectrophoresis, require complex fabrication of microelectrode that cannot be changed dynamically [6].

Rapid electrokinetic patterning is a dynamic, high-throughput technique that allows for micro- and nano-based patterning of colloids. REP requires the application of a uniform AC electric field and optical illumination in order to concentrate particles within solution. Optical illumination induces an axisymmetric temperature gradient within the chip, with the highest temperature at the beam focus. This temperature variation causes physical property variation, most importantly the electrical conductivity and permittivity. The applied electric field interacts with these material property variations to create a body force within the fluid and consequently an electrothermal vortex. The vortex transports colloids to the center of the torus where particles are trapped by particle-electrode interactions. This flow is produced whether the laser is focused at the top electrode or bottom electrode. Figure 1a provides an example of a laser focused on the bottom electrode. When focusing the laser at the top electrode, it would look like Fig. 1a but vertically reflected in the horizontal midplane.

REP enables particle manipulation, translation, sorting, and patterning. This wide variety of processes that the technique enables make it useful in fields like micromanufacturing, plasmonics, biological detection and cell sorting, as well as chemical analysis. Using light patterning, particles can be concentrated into different patterns and structures on the electrode surface [8]. For example, detection of pathogenic bacteria requires cell sorting and concentration. The ability to sort such entities with ease could be potentially resolved with REP [9, 10]. These tasks are highly desirable in the biomedical community, potentially allowing full-scale laboratory analysis with small sample volumes and on platforms the size of a credit card [4].

REP can simultaneously collect particles of a specific size and manipulate the colloidal cluster. This technique requires a heat source, such as a laser or microheaters, to provide a temperature gradient along with an AC electric field. REP provides several benefits. First, it is a noncontact high-speed manipulation method, accumulating a cluster of particles in a matter of seconds. Second, REP is highly repeatable, as under the same settings, the same particle clusters will form each time. Third, REP chips require simple fabrication and manufacturing.

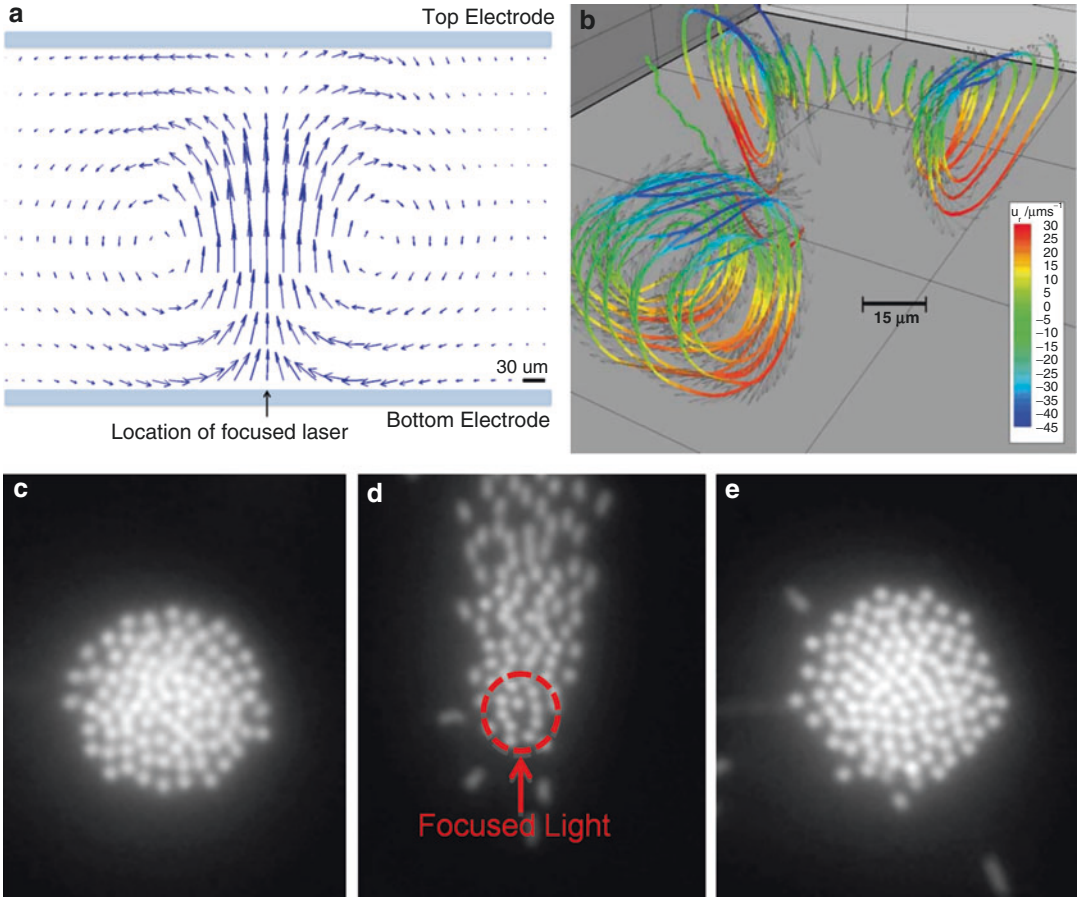
REP is a promising technique in its potential to contribute to a wide variety of fields in nano- and micro-technologies.

Physics

REP combines optics (using laser light) with electrokinetics (using an AC electric field). When coupled with one another, an electrothermal flow is produced in a toroidal fashion. This flow pattern "sweeps" particles toward the center of the torus' minor radius, where the temperature gradient generated by the nonuniform light source is maximum. The cluster of particles can be trapped onto the top or bottom electrode surface (depending on where the laser light is focused). An example of this is provided in Fig. 2b.

When the light source is removed from the apparatus, particles are no longer collected into the trap. Smaller-sized particles (i.e., <3 μm), post-light deactivation, are repelled from one another due to dipole-dipole repulsion (Fig. 2c). However, this repulsive effect is largely due to the AC electric field being maintained in chip. If the alternating current is removed, but the laser light is maintained, natural convection is seen within the apparatus [11] but usually at a speed an order of magnitude slower than that created by the electrical field. The coupling of the light source and AC electric field is vital for maintaining the concentration of particles.

In more detail, dielectric particles are polarized due to the introduction of the electric field. There is a dipole force that acts between two neighboring particles. Due to their repulsion and self-organization on the electrode surface when

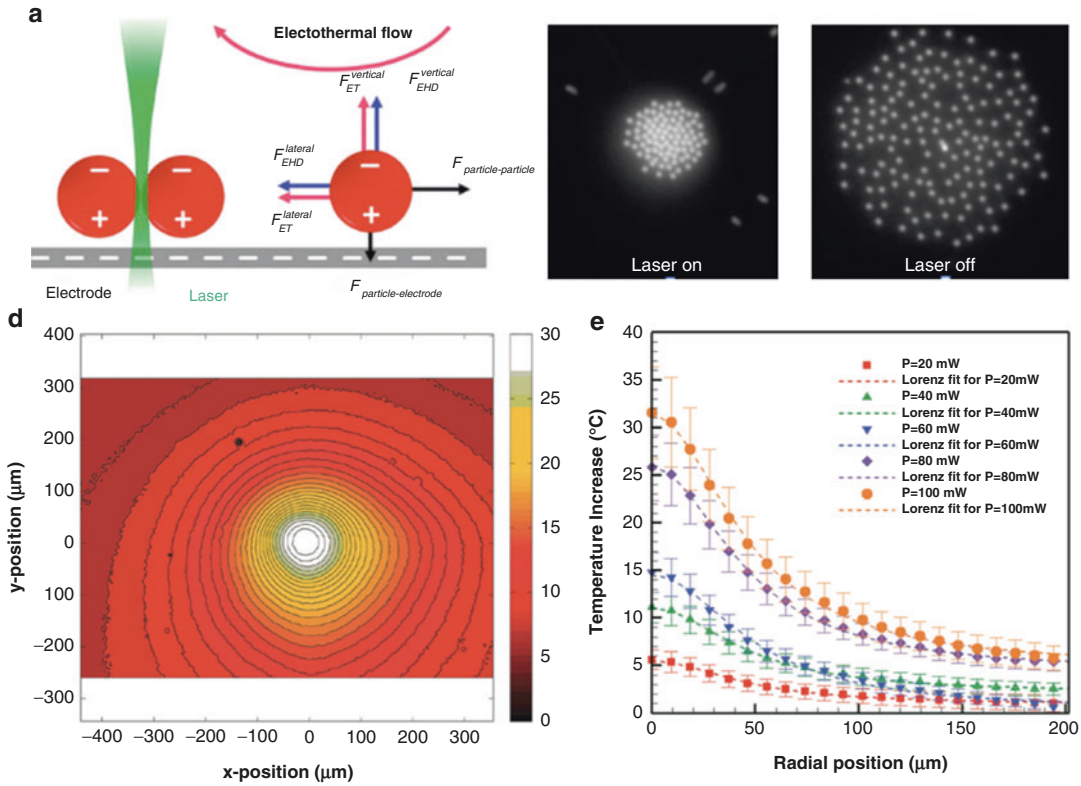


Rapid Electrokinetic Patterning and Its Applications, Fig. 1 **a** A 2D projection of electrothermal flow in REP. In this circumstance, a laser is focused on the bottom electrode of an REP chip, producing an upward flow on the inner radius of the toroidal fluid motion. **b** A 3D trajectory of the electrothermal flow at a frequency of 200 kHz. In this case the laser is focused at the top electrode surface (**b** Printed with permission from Kumar et al. Copyright 2010 Springer) [7]. **c** 1 µm particles are trapped in a monolayer using REP. These particles can be manipulated with ease, such as with **d** where the laser light is suddenly moved to an area below its original spot, and the particles are swept to the new location, until they are again in equilibrium in the trap, as seen in (**e**)

solely affected by the electric field, their lines of centers remain perpendicular to the electric field lines. As the laser light is combined with these electric field effects, the temperature gradient provided by the light induces a gradient for the conductivity and permittivity of the solution. This combination produces a toroidal electrothermal vortex, which is characterized by the time-averaged electrothermal body force equation:

$$\langle f_e \rangle = \frac{1}{2}\mathrm{Re}\left[\left(\frac{(\sigma\nabla\varepsilon - \varepsilon\nabla\sigma)\bullet E}{\sigma + i\omega\varepsilon}\right)E^* - \frac{1}{2}|E|^2\nabla\varepsilon\right] \quad (1)$$

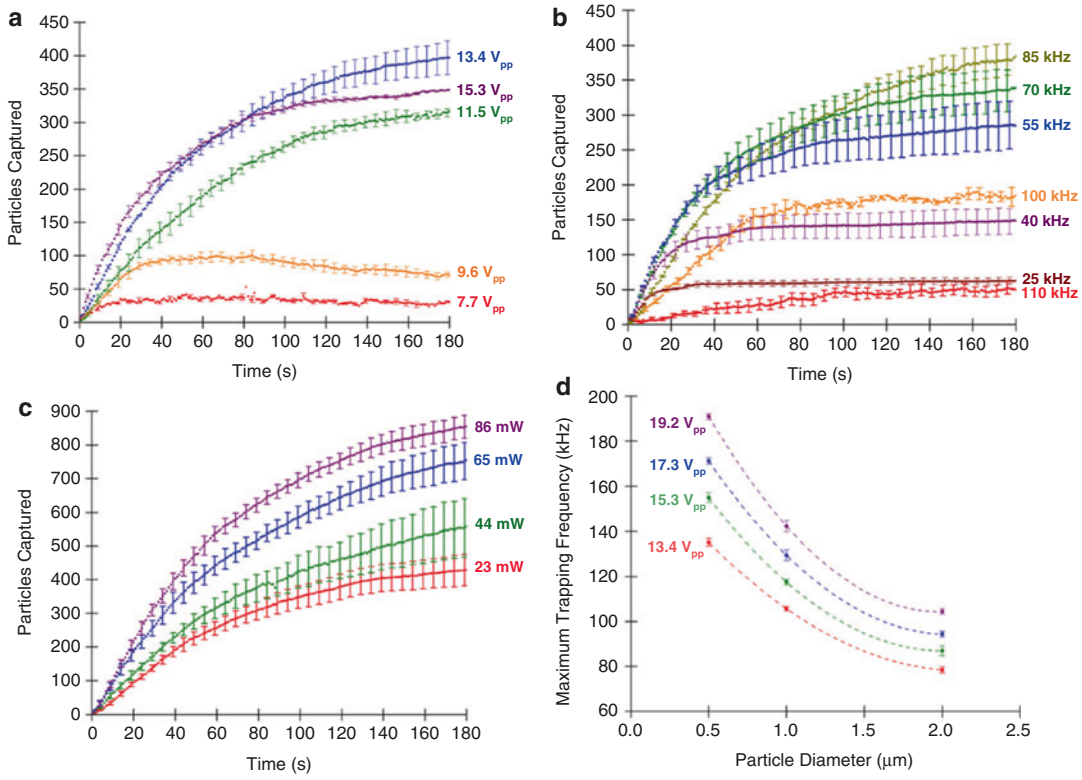
In the time-averaged fluid body force equation, Re represents taking the real valued portion of the expression. The electric field is represented by E and its complex conjugate is E^*. The equation encompasses the gradient of conductivity, σ, and permittivity, ε, of the solution, which is created due to the temperature gradient produced by the optical source, and then finally the ω term represents the angular frequency. On the right-hand side of the equation, the left-hand term represents the Coulomb force, while the right-hand term is the dielectric force component [12].



Rapid Electrokinetic Patterning and Its Applications, Fig. 2 **a** The force relationships on a particle undergoing REP – the bottom electrode is negatively charged, where the particle with an induced dipole interacts with its surface. Both lateral and vertical induced drag force, and electrohydrodynamic forces act on the particle as well. Additionally, the particle-particle interaction force is shown (Reprinted with permission from Kwon et al. Copyright 2012 Springer) [13]. **b** shows the congregation of particles during REP trap activation and **c** dipole-dipole repulsion of the particles after REP is "turned off." **d** The temperature increase on an ITO coverslip introduced to a 150 mW powered laser. The laser is focused at the origin. **e** The effect of different laser powers (in mW) and its corresponding temperature increase, starting at the origin of the laser spot and propagating out radially (**d** and **e** Reprinted with permission from Kumar et al. Copyright 2010 American Chemical Society) [15]

From current understanding, four main electrokinetic relationships can be analyzed. The first relationship is the force between particles. Denoted $F_{particle}$, this force is what particles experience between one another (i.e., $F_{particle}$- $F_{particle}$). Under the AC electric field, the dipole-induced particles experience repulsion from one another. The second is the force between the electrode and particle ($F_{electrode}$- $F_{particle}$). As the electrothermal flow sweeps the particle into a monolayer formation, the particles undergo a physical interaction with the electrode. This electrothermal flow aids in a drag force that can be broken into vertical and lateral component ($F^{lateral}_{ET}$, $F^{vertical}_{ET}$). The fourth is an attractive particle-particle force ($F^{lateral}_{EHD}$, $F^{vertical}_{EHD}$). Polarized particles produce a local electrohydrodynamic flow due to the interaction between the electric field and an ionic double layer on the particle and electrode [13]. This force balance is demonstrated in Fig. 2a.

At a specific frequency, (1) the drag force due to electrohydrodynamic flow, (2) the force between the electrode and particle, (3) the force between particles, and (4) the drag force due to the electrothermal vortex balance one another at the trapping point. If the frequency is increased, the flow becomes so strong that particles are



Rapid Electrokinetic Patterning and Its Applications, Fig. 3 **a** Accumulation of 1.0 um particles at the ITO electrode over a period of 3 min for various V_{pp}. Laser power is 23 mW and AC frequency is 70 kHz. **b** Accumulation of 1.0 um particles at the ITO electrode over a period of 3 min at various frequencies. Laser power is 23 mW and AC voltage is 15.3 V_{pp}. **c** Accumulation of 1.0 um particles at the ITO electrode over a period of 3 min for various frequencies. AC peak-to-peak voltage is 15.3 V_{pp}. and AC frequency is 70 kHz. **d** Maximum trapping frequency for particles of various diameters to cluster at different applied AC peak-to-peak voltage. Laser power is 23 mW (Reprinted with permission from Williams et al. Copyright 2009 IOP) [12]

rapidly swept and unable to be maintained within their respective cluster. However, this relationship is still under investigation.

Although these relationships do remain under investigation, Fig. 3 represents trends seen experimentally in REP. In Fig. 3a the number of particles captured increases linearly and then plateaus as the particle cluster reaches steady state. Furthermore, it is observed that higher applied voltages cause larger numbers of particles trapped (except for the 15.3 V_{pp} case). This 15.3 V_{pp} outlier case may be due to an increased electrothermal body force. This increase may overcome particle-particle and particle-electrode forces, ripping particles out of their original trap space. In Fig. 3b it is observed that the particles captured for various frequencies increase linearly and then plateau as the cluster reaches steady state. Figure 3c shows that the number of particles within the cluster increases linearly and then begins to reach steady state. The increase in laser power increases the number of particles captured. This may be due to the fact that as the laser power increases, the electrothermal flow velocity increases. This is due to a steeper temperature gradient, leading to faster particle circulation. This circulation brings particles closer to the colloid cluster, increasing the probability of particles to be pulled into the REP trap. Figure 3d shows the maximum trapping frequency for particles of different diameters before it fails to form a cluster. A particle of smaller diameter will remain trapped at a higher frequency when compared to a larger particle. Furthermore, the maximum trapping

frequency of the same particle increases as the applied AC peak-to-peak voltage increases [14].

Methodology

Fabrication of REP chips is relatively simple. The chip is made of a thin-film electrode coated onto a glass substrate. Indium tin oxide (ITO)-coated glass substrates are typically used to conduct REP. This coating is highly transparent allowing for observation within the REP chip under a microscope.

Two of these substrates are placed with the electrodes facing one another with a spacer in between. This spacer is designed to define microchannels and wells. The spacer material is adhered to both electrodes and is nonreactive with the sample. In some cases, a commercial double-sided tape can be used to conduct short-term REP studies. The typical setup for conducting REP requires a laser source and an AC function generator. The laser is applied to a desired spatial location on chip. A representation of this setup is presented in Fig. 4.

The channel is often illuminated using a near-infrared (NIR) laser of 1064 nm wavelength. However, alternative wavelengths may be used. Traditionally, the laser power used in REP is less than 200 mW. Furthermore, the laser frequency has to be considered so that the heat absorption of the laser onto the sample is minimal. Objective lenses are typically chosen to be at a magnification of 20× or larger. This ensures that the laser spot size is focused enough to create a steep temperature gradient. A schematic of the REP setup can be found in Fig. 5.

For electrical effects, an AC electric field with a frequency less than 200 kHz must be used. This induces polarization of the dielectric material. In higher frequency regimes, the particles relax, hindering the collection of the particles under REP. Very low-frequency regimes often result in self-organization of particles on the electrode surface, but in turn induce strong particle-electrode interactions.

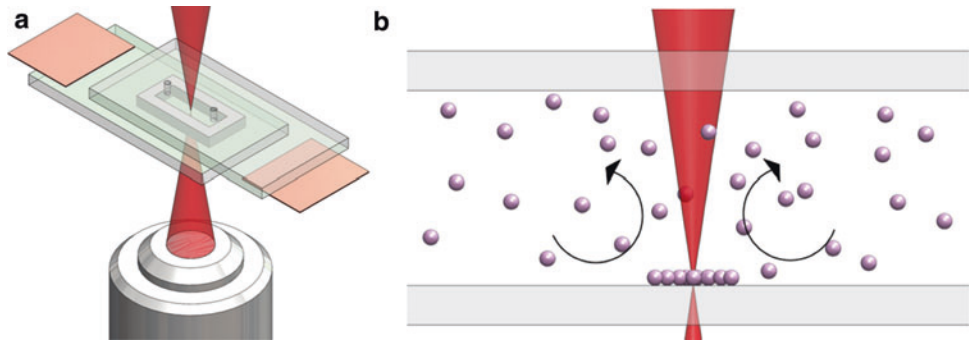
Particles must be suspended in low conductivity solutions (i.e., $\sigma \leq 10$ mS/m) [14, 15]. When particles are submerged in higher conductivity solutions, the electric double layer becomes thinner, increasing the particle-electrode adhesion.

Applications

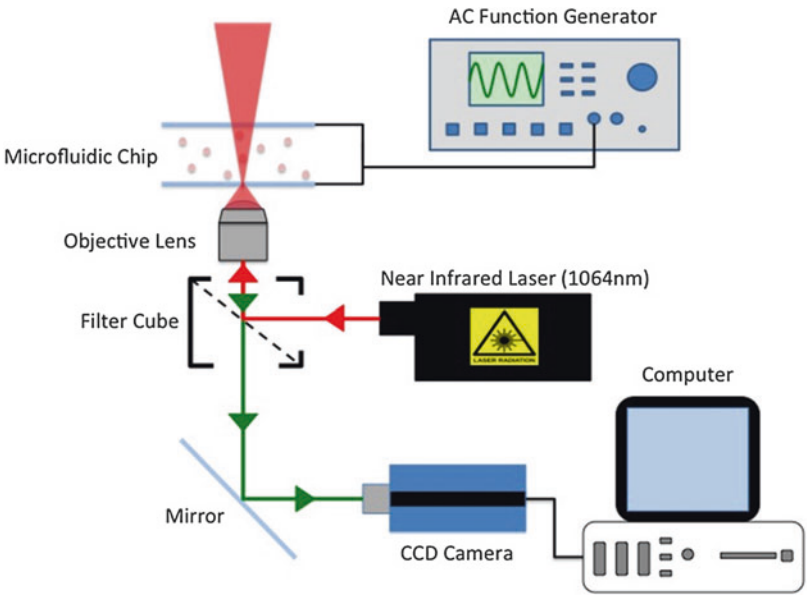
In recent years, researchers have investigated potential applications of rapid electrokinetic patterning. This spawned a wide spectrum of uses for this opto-electrokinetic technique, from biology to plasmonics.

In Biology

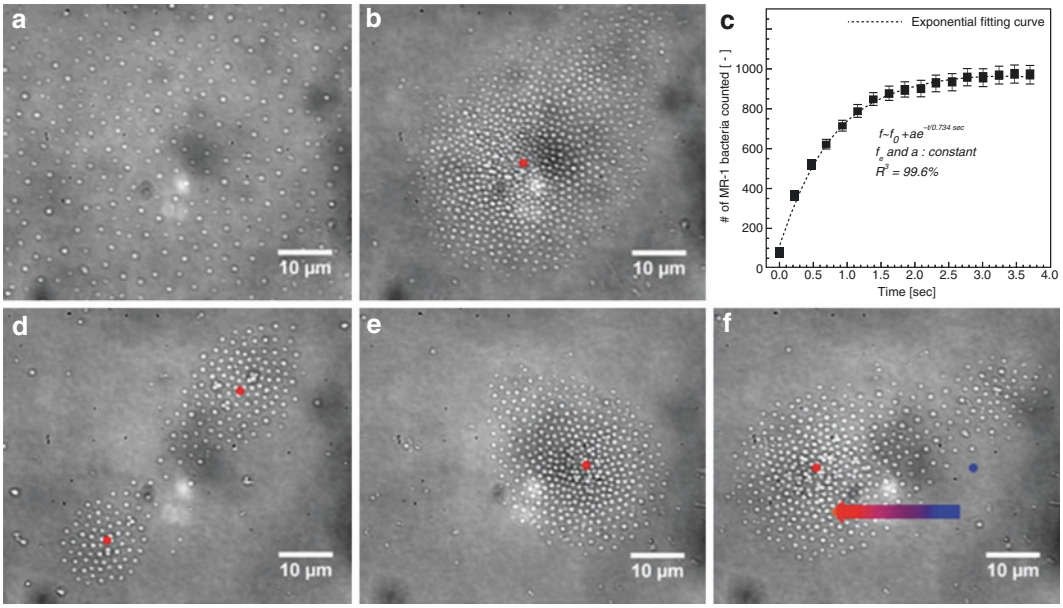
Biology wet labs are looking for alternative systems that allow for appropriate sampling of



Rapid Electrokinetic Patterning and Its Applications, Fig. 4 In image **a**, a demonstration of an REP chip with laser light focused on a particular plane, where **b** shows how at the beam waist of the laser particles are trapped on the electrode surface; note the *black arrows* correspond with the flow direction of the electrothermal vortex

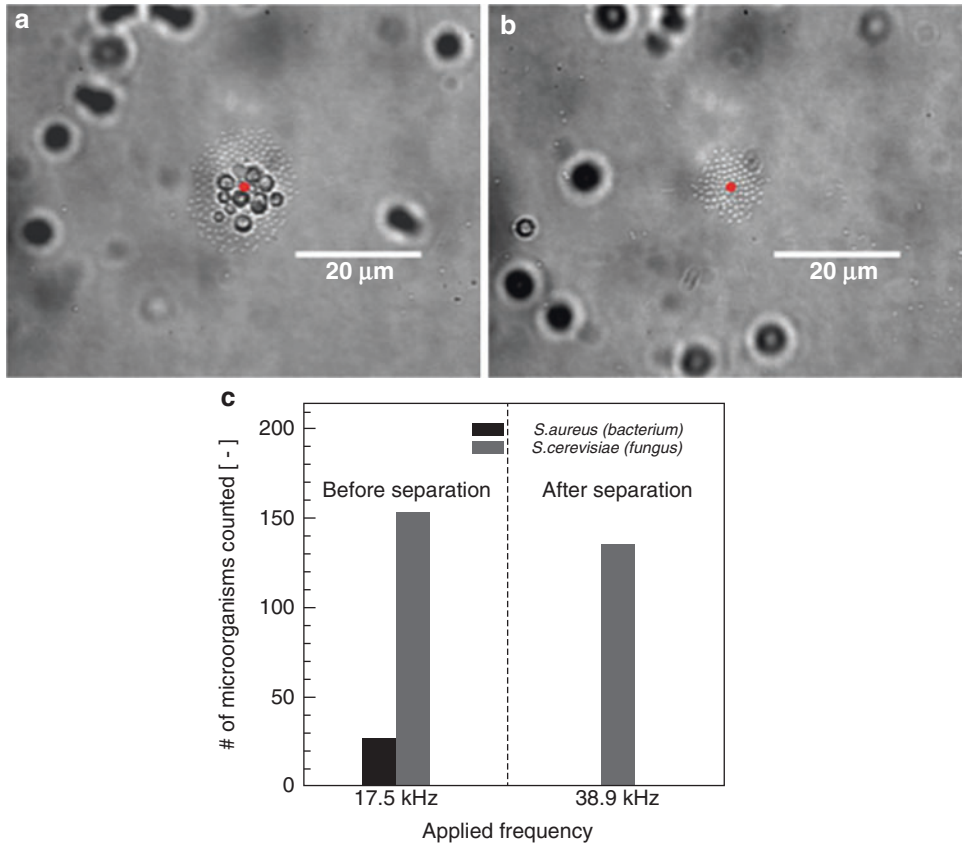


Rapid Electrokinetic Patterning and Its Applications, Fig. 5 A schematic of a standard REP setup. A fabricated REP chip undergoes an AC electric field from a function generator as well as heating from laser light. The 1064 nm laser is often directed to the substrate via an objective lens from the microscope. The CCD camera records images from the objective



Rapid Electrokinetic Patterning and Its Applications, Fig. 6 From Kwon et al. image **a** shows *S. oneidensis* prior to REP trapping and **b** during trapping. In **c** it is shown how over time more bacteria are collected in the REP trap, eventually reaching a plateau, image **d** shows two laser spots and the particles trapped about the focused light, and **e** and **f** demonstrate how moving the light about the chip allows for dynamic manipulation of the particles on chip (Reprinted with permission from Kwon et al. Copyright 2012 Royal Society of Chemistry) [3]

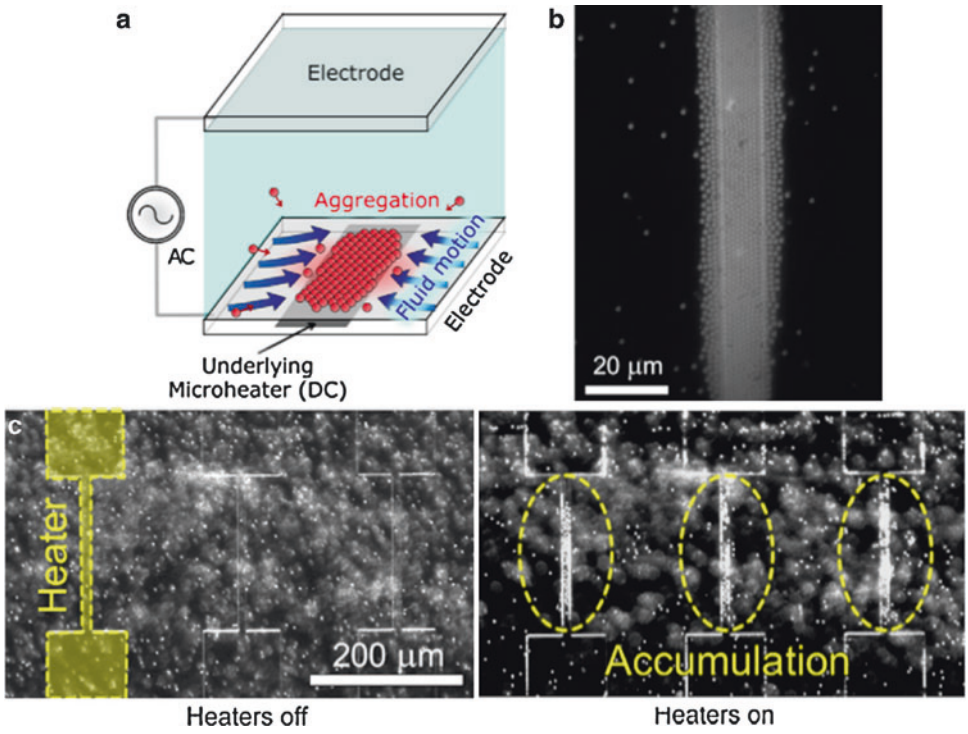
solutions. Current techniques often require large amounts of consumables (both from reagents and materials), which are both costly and environmentally unfriendly. Many optical and electrokinetic-based methods wish to alleviate these issues. However, biological applications are difficult to implement in optoelectric systems due to issues with their charge properties, uniformity in sample,



Rapid Electrokinetic Patterning and Its Applications, Fig. 7 **a** Showing how multiple species – *S. aureus* and *S. cerevisiae* – are collected with REP. **b** After changing frequency, the larger *S. cerevisiae* are swept out of the trap, and only the *S. aureus* remains. **c** Providing how in image (**a**) there are 150 bacteria and 8 fungi trapped and how after in (**b**) there are 125 bacteria and no fungi within the trap (Reprinted with permission from Kwon et al. Copyright 2012 Royal Society of Chemistry) [3]

and viability/compatibility with the system. Therefore, like with many optical and electrokinetic techniques, biological use with REP is currently being investigated. Kwon et al. published the first work, using REP as a bioassay platform. *Shewanella oneidensis*, a 1 µm rod-shaped bacteria, was examined with rapid electrokinetic patterning. This gram-positive MR-1 mutant was placed under an AC electric field, where the bacterium's major axis aligned with the electric field. After introducing the laser, the bacteria collected onto the electrode surface in their vertical orientation. Within 0.73 s, 700 organisms were maintained in the trap, saturating out at ~1,000. Figure 6a, b shows the *S. oneidensis* before and during trapping. This demonstrates the ability to quickly and dynamically collect the species under REP. Likewise, multiple laser spots allowed for the collection of multiple *S. oneidensis* clusters, and through movement of the laser, these traps could translate about the chip, as seen in Fig. 6d [3].

Kwon et al. also demonstrated collecting multiple organisms within an REP trap, depending on its critical frequency effects (Fig. 7). *Saccharomyces cerevisiae* (a 5 µm spherical-shaped fungus) and *Staphylococcus aureus* (a ~1 µm spherical-shaped bacteria) were investigated. These two geometrically similar species were chosen due to their shape and similar zeta potentials. Under a 17.5 kHz frequency, 10.07 V_{pp}, and 20 mW laser power, both species clustered together within the



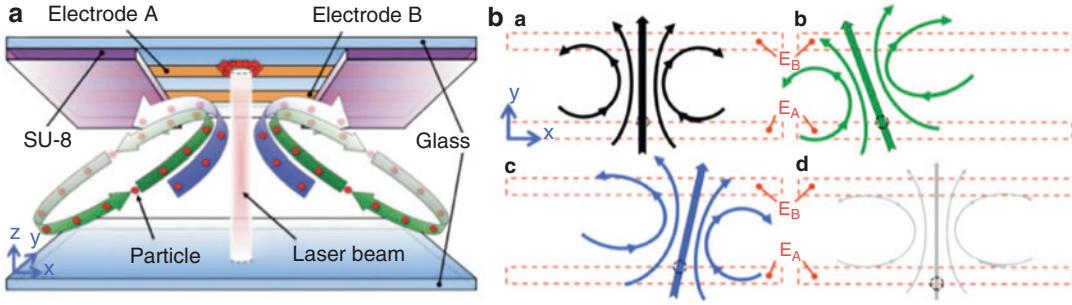
Rapid Electrokinetic Patterning and Its Applications, Fig. 8 **a** A schematic of a fabricated microheater to induce a temperature gradient on the bottom electrode surface, allowing for REP to take place without laser heating. **b** Accumulation of polystyrene particles along the microheater. **c** A display of an array of three microheaters. Initially when the heaters are off, particles are freely dispersed in solution. After the heaters are turned on with the electric field, particles begin to accumulate on the surface of the electrode, in the location of the thin-film heaters (Reprinted with permission from Velasco et al. Copyright 2012 Elsevier) [17]

REP trap. However, by increasing the frequency to almost twice as much, the *S. cerevisiae* were swept away from the trap, whereas the *S. aureus* sample was maintained. This proved the capability of using REP as a method for separating small microbial samples on the basis of size [3].

Alternative Platforms

Producing a cost-effective, easily accessible, and minimally power-consuming REP system is a key requirement for developing REP-based portable devices. Currently, a 1,064 nm laser is used to illuminate the REP chip. This is easily controllable within the laboratory space but provides little-to-no mobility of the system. Williams et al. investigated thin-film heaters as an alternative for laser-induced electrothermal flow [16]. Velasco et al. took this theory further by experimentally fabricated resistors onto a glass surface and tested 1–2 µm polystyrene spheres in the system [17]. A 10 µm wide wire trace was used which produced a 0.5 K temperature gradient. The system allowed for trapping, patterning, and sorting of colloidal particles. The electrothermal vortex was slower as compared to traditional REP [17]. Examples of this work are provided in Fig. 8a–c. This successful work showed the potential capabilities of microfabricating REP systems that can be used in the field. This opens up a potential opportunity for future use as a portable electrokinetic system.

Park et al. investigated twin opposing microvortices (TOMV). Fabricating two coplanar ITO electrodes onto glass surface (partly covered with SU-8), an AC electric field was applied between them while introducing laser light through the chip. Depending on where the



Rapid Electrokinetic Patterning and Its Applications, Fig. 9 In **a** is a schematic of microfabricated ITO electrodes, partially blocked by a layer of SU-8 on a glass substrate. These coplanar electrodes are alternative to the traditional parallel ITO coverslips used. This design creates twin opposing microvortices, collecting particles at the electrode surface. In **b**, where the laser light is focused (demonstrated by the point drawn on the *bold colored line*), changes the orientation of the opposing microvortices (Reprinted with permission from Park et al. Copyright 2012 Royal Society of Chemistry) [1]

light was focused, the TOMV changed their orientation within the chip (Fig. 9). These tunable microvortices are of interest due to the specific control of location and mixing of particles. Asymmetric flows and interesting mixing effects are possible with this technique [1].

Ndukaife et al. fabricated plasmonic nanostructures to increase heating effects in a traditional REP chip [2]. This nanostructure array absorbed light optimally at 1,064 nm. By exciting the structure at resonance, a rapid heat gradient was produced. Due to the increased heating effect, the velocity of the electrothermal vortex increased. This showed the potential of plasmonic structures as an REP platform. This could be used further in biosensors as well for surface-enhanced Raman spectroscopy (SERS) [2]. However, it is important to note that the fabrication of plasmonic structures is cost and labor intensive due to the use of e-beam lithography. A carefully selected ITO electrode can produce a sufficiently strong flow without the need of nanostructures.

Conclusions

Rapid electrokinetic patterning poses an array of current and future applications in a variety of fields. Now that the technique has been developed, work is being done to apply it. As this space is being pursued, REP will be used for novel applications within a large spectrum of nano- and microscale phenomena.

Cross-References

- AC Electrokinetics
- Dielectrophoresis
- Lab-on-a-Chip

References

1. Park, C., Wereley, S.T.: Rapid generation and manipulation of microfluidic vortex flows induced by AC electrokinetics with optical illumination. Lab Chip **13**, 1289–1294 (2013)
2. Ndukaife, J.C., Mishra, A., Guler, U., Nnanna, A.G. A., Wereley, S.T., Boltasseva, A.: Photothermal heating enabled by plasmonic nanostructures for electrokinetic manipulation and sorting of particles. ACS Nano **8**, 9035–9043 (2014)
3. Kwon, J.-S., Ravindranath, S.P., Kumar, A., Irudayaraj, J., Wereley, S.T.: Opto-electrokinetic manipulation for high-performance on-chip bioassays. Lab Chip **12**, 4955–4959 (2012)
4. Franke, T.A., Wixforth, A.: Microfluidics for miniaturized laboratories on a chip. Chem. Phys. Chem. **9**, 2140–2156 (2008)
5. Arai, F., Ng, C., Maruyama, H., Ichikawa, A., El-Shimy, H., Fukuda, T.: On chip single-cell separation and immobilization using optical tweezers and thermosensitive hydrogel. Lab Chip **5**, 1399–1403 (2005)

6. Pethig, R.: Review article-dielectrophoresis: Status of the theory, technology, and applications. Biomicrofluidics **4**, 1–35 (2010)
7. Kumar, A., Cierpka, C., Williams, S.J., Kähler, C.J., Wereley, S.T.: 3D3C velocimetry measurements of an electrothermal microvortex using wavefront deformation PTV and a single camera. Microfluid. Nanofluid. **10**, 355–365 (2010)
8. Williams, S.J., Kumar, A., Wereley, S.T.: Electrokinetic patterning of colloidal particles with optical landscapes. Lab Chip **8**, 1879–1882 (2008)
9. Bridle, H., Miller, B., Desmulliez, M.P.Y.: Application of microfluidics in waterborne pathogen monitoring: A review. Water Res. **55**, 256–271 (2014)
10. Yoon, J.-Y., Kim, B.: Lab-on-a-chip pathogen sensors for food safety. Sensors (Basel) **12**, 10713–10741 (2012)
11. Ramos, A., Morgan, H., Green, N.G., Castellanos, A.: Ac electrokinetics: A review of forces in microelectrode structures. J. Phys. D. Appl. Phys. **31**, 2338–2353 (1998)
12. Mishra, A., Kwon, J.-S., Thakur, R., Wereley, S.: Optoelectrical microfluidics as a promising tool in biology. Trends Biotechnol. **32**, 414–421 (2014)
13. Kwon, J., Thakur, R., Wereley, S.T.: Encyclopedia of Nanotechnology. Springer Netherlands, Dordrecht (2012)
14. Williams, S.J., Kumar, A., Green, N.G., Wereley, S.T.: Optically induced electrokinetic concentration and sorting of colloids. J. Micromech. Microeng. **20**, 015022 (2010)
15. Kumar, A., Kwon, J.-S., Williams, S.J., Green, N.G., Yip, N.K., Wereley, S.T.: Optically modulated electrokinetic manipulation and concentration of colloidal particles near an electrode surface. Langmuir **26**, 5262–5272 (2010)
16. Williams, S.J.: Enhanced electrothermal pumping with thin film resistive heaters. Electrophoresis **34**, 1400–1408 (2013)
17. Velasco, V., Williams, S.J.: Electrokinetic concentration, patterning, and sorting of colloids with thin film heaters. J. Colloid Interface Sci. **394**, 598–603 (2013)

Rate Sensors

► Gyroscopes

Reactive Current Clamp

► Dynamic Clamp

Reactive Empirical Bond-Order Potentials

J. David Schall[1], Paul T. Mikulski[2], Kathleen E. Ryan[3], Pamela L. Keating[3], M. Todd Knippenberg[4] and Judith A. Harrison[3]
[1]Department of Mechanical Engineering, Oakland University, Rochester, MI, USA
[2]Department of Physics, United States Naval Academy, Annapolis, MD, USA
[3]Department of Chemistry, United States Naval Academy, Annapolis, MD, USA
[4]Department of Chemistry, High Point University, High Point, NC, USA

Synonyms

Bond-order potential; REBO

Definition

Reactive empirical bond-order potentials are interatomic energy functions used in molecular dynamics simulation and modeling of nanosystems where an accurate but efficient description of chemical reactivity is required.

Genesis of Reactive Bond-Order Potentials

Atomistic simulation of a large number of atoms using molecular dynamics (MD) is a powerful tool for understanding the fundamental mechanisms present in nanomaterials systems. The ability to accurately model chemistry including reactivity, bonding, charge transfer, polarizability, mixing, etc., is central to this understanding. Underpinning these calculations is the atomic interaction potential. Ideally, the atomic interactions would be obtained directly from first-principles, i.e., through solution of Schrodinger's equation. However, such calculations are orders of magnitude too slow for the large number of energy evaluations required to study nanosystems

of reasonable size and practical interest. To obtain useful information in a reasonable amount of time, empirical, and semiempirical approximations to the atomic potentials have been developed. In this entry, a survey of reactive empirical bond-order and related interatomic potentials is presented.

An effective analytic potential should have a relatively simple functional form, which captures the essential essence of the underlying quantum mechanical bonding, and will be able to mimic experimental quantities by utilizing empirically derived functions and parameters. Empirical potentials are likely to include a wide range of fitting parameters, which are selected to reproduce cohesive energies, elastic constants, lattice constants, and surface energies. Ideally, the potential will also have some degree of transferability, having the ability to describe structures not included in the fitting database, at least in a qualitative sense. Finally, the resulting function should be relatively efficient computationally. The process of developing a potential, sometimes referred to as an art as well as a science, requires a combination of chemical insight, trial and error, and tenacity on the part of its developer.

Covalent materials, such as silicon and carbon, form strong directional bonds. This poses a challenge for potential development for this important class of materials. Many standard potential functions, such as the Lennard-Jones (LJ) potential or the embedded atom method (EAM), do not include any bond directionality. The Stillinger–Weber (SW) potential for solid and liquid phases of silicon was one of the first attempts to use a classic potential to overcome this challenge [1]. Stillinger and Weber based their potential model on a many-body approach. In this approach, the total energy is given as a linear superposition of terms representing different types of interatomic interactions, e.g., stretching, bending, rotation, and torsion. For the SW potential, the total energy includes only two of these terms, a pair- and triplet-term based on the geometric factors of bond length and bond angle, respectively. The total potential energy, *Etot*, is given by

$$E_{tot} = \frac{1}{2}\sum_{\substack{ij \\ (i\neq j)}} \phi_{ij}(R_{ij}) + \sum_{ijk} g(R_{ij})\, g(R_{ik}) \left(\cos\theta_{ijk} + \tfrac{1}{3}\right)^2. \quad (1)$$

Here ϕ_{ij} is the pair-term (bond stretching) representing electrostatic interactions between atoms i and j. The second summation is a three-body term, which represents bond bending between atoms i, j, and k. The term $g(r_{ij})$ is a decaying function with a cutoff between the first- and second-neighbor shell, and θ_{ijk} is the bond angle described by two neighbors j and k of atom i. The inclusion of the three-body term allows the potential to achieve an acceptable description of short-range order and of atom-exchanging diffusive motion in the liquid phase. While this potential is reasonably accurate when used for modeling solid silicon in the diamond cubic phase, it is biased toward the tetrahedral bond angle through the explicit inclusion of the factor $\left(\cos\theta_{ijk} + 1/3\right)$ in the three-body term. The inclusion of this trigonometric factor discriminates in favor of pairs of bonds with the tetrahedral geometry, i.e., $\left(\cos\theta_{ijk} = -1/3\right)$, and limits the transferability of the potential. For instance, it cannot accurately predict the energies for various non-tetrahedrally bonded, high-pressure phases of silicon, it does not correctly predict relative energies of surface structures, and the coordination of liquid-phase silicon is too low.

The family of potentials, CHARMM, AMBER, OPLS, etc., derived from the many-body approach described above for the Stillinger–Weber potential is often used in simulation of organic systems. These so-called force-field models are able to model structural and dynamical properties of very large molecules with a high degree of accuracy. However, it is important to note that the force-field methods generally do not allow for bond-forming or breaking to occur during simulation. Connectivity must be determined a priori. For a review of these potentials see for example Mackerell [2].

Abell presented a very general description of bonding based on the observation of a universal relationship between binding energy and bond

length in 1985 [3]. This relationship placed the bonding in crystalline solids and molecules on the same footing. The only criterion for bonding preference was the optimization of the binding energy with respect to the local coordination. The local coordination is, in turn, the dominant topological variable in the determination of binding energy. Soon after the publication of Abell, Tersoff developed a potential constructed to guarantee that this universal behavior was obtained [4]. The Tersoff potential was the first to attempt to incorporate the structural chemistry of covalently bonded systems into empirical potential energy function. The general form of the potential is

$$\begin{aligned} E &= \sum_i E_i = \frac{1}{2}\sum_{i\neq j} V_{ij}, \\ V_{ij} &= f_c\left[V_R\left(R_{ij}\right) - B_{ij}V_A\left(R_{ij}\right)\right], \end{aligned} \tag{2}$$

where E is the total energy of the system, E_i is the site energy for site i, V_{ij} is the interaction energy between atoms i and j, and R_{ij} is the distance between these atoms. The sum is over the j nearest neighbors of i; $V_R(R)$ and $V_A(R)$ are pair-additive repulsive and attractive interactions, respectively; f_c is a cutoff function to limit the range of the potential. For the Tersoff potential, the repulsive and attractive terms are represented by the Morse-type functions

$$\begin{aligned} V_R &= A\exp\left(-\lambda_1 R_{ij}\right) \\ V_A &= B\exp\left(-\lambda_2 R_{ij}\right), \end{aligned} \tag{3}$$

where A, B, λ_1, and λ_2 are all positive constants with $\lambda_1 > \lambda_2$. All deviations from the simple-pair potential are ascribed to the dependence of the bond-order B_{ij} on the local bonding environment or coordination. More specifically, the bonding strength for a bonded pair should be a monotonically decreasing function of the number of competing bonds, the strength of the competing bonds, and the cosines of the angles of the competing bonds. The innovative feature of the Tersoff potential is that it does not assume different forms for the angular functions for different hybridizations. In his work, B_{ij} has the form

$$\begin{aligned} B_{ij} &= \left(1+\beta^n \zeta_{ij}^n\right)^{-1/2n}, \\ \zeta_{ij}^n &= \sum_{k\neq i,j} f_c\left(r_{ijk}\right) g\left(\theta_{jik}\right) \exp\left[\lambda_3^3\left(r_{ij}-r_{ik}\right)^3\right], \\ g(\theta) &= 1 + c^2/d^2 - c^2/\left[d^2 + (h - \cos\theta)^2\right], \end{aligned} \tag{4}$$

where θ_{jik} is the bond angle between bonds ij and ik. The angular function $g(\theta)$ is determined globally by a fit to solid structures with different coordination. This gives the function a high degree of transferability. Tersoff extended his original silicon potential to include germanium, carbon, and combinations thereof. This potential form has also been extended to a variety of III-V semiconductors.

An empirical bond-order expression that described both hydrocarbon molecules and solid-state carbon was developed by Brenner [5] to model chemical vapor deposition of diamond films. Taking inspiration from the Tersoff potential, the form of the reactive empirical bond-order potential (REBO) potential allows for bonds to form and break with changes in hybridization. This first-generation Brenner potential is very similar to the Tersoff potential. Pair repulsive and attractive terms are again represented by Morse-type potentials. The primary difference is the way in which the bond order is handled for hydrocarbon molecules. The expression for the total bond order is

$$\overline{B}_{ij} = \frac{B_{ij} + B_{ji}}{2} + F_{ij}\left(N_i^t, N_j^t, N_{ij}^{\text{conj}}\right) \tag{5}$$

where

$$\begin{aligned} B_{ij} = &\left[1 + \sum G_i\left(\theta_{jik}\right) f_c\left(r_{ik}\right) e^{\alpha\left[\left(r_{ij}-R_{ij}^E\right)-\left(r_{ik}-R_{ik}^E\right)\right]}\right. \\ &\left. + H_{ij}\left(N_i^H, N_i^C\right)\right]^{-\delta}, \end{aligned} \tag{6}$$

The quantities Ni^C and Nj^H are the number of carbon and hydrogen atoms bonded to atom i. The total number of neighbors, Ni^{t} of atom i is $(Ni^{\text{C}} + Nj^{\text{H}})$, Nij^{conj} depends on whether a bond between carbon atoms i and j is part of a conjugated system, $G(\theta)$ is a function of the angle between bonds i-j and i-k and has the same form

as the Tersoff potential. The two- and three-dimensional cubic splines, H_{ij} and F_{ij}, have discrete nodal values that are modified so that the potential reproduces the energies of various hydrocarbon molecules. The resulting potential successfully describes the different bonding characteristics of hydrogen and carbon radicals and nonconjugated double and triple carbon-carbon bonds.

Extensions of the First-Generation REBO Potential

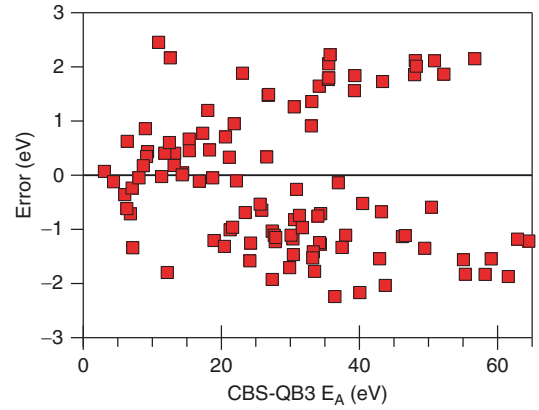
Because the Tersoff-Brenner formalism has been used successfully to simulate a wide range of processes such as chemical vapor deposition, diamond growth, carbon nanotubes, graphene, diamond-like carbon, etc., it was extended to other systems by various groups. In what follows, some examples of ways in which the first-generation REBO potential has been extended are provided. This analysis is not meant to include all such examples, however.

Parameterization of Additional Atom Types

Because bond-order potentials were designed to model chemical reactivity in covalent systems, it was logical to adopt the first-generation REBO formalism to model Si-H systems [6]. Parameterizations also exist for Si-F and Si-Cl. The first attempts to construct bond-order potentials capable of modeling three different atom types were two independently developed C-Si-H potentials. The C-Si-H potential of Beardmore and Smith is a hybrid potential which makes use of the C-H, Si-H, and Si-C parameters developed by Brenner [5], Murty and Atwater [6], and Tersoff, respectively. Because it is a hybrid potential, it inherits the strengths and weaknesses of the potentials on which it is based. The C-Si-H potential of Dyson and Smith, known as the extended Brenner (XB) potential, adopts the C-H parameters of Brenner and develops three new sets of parameters for Si-Si, Si-C, and Si-H [7] while maintaining the same functional form.

When chemical reactions occur during a simulation, the bond order of atoms must change in a continuous manner from reactant to product values. This is accomplished in the first-generation REBO hydrocarbon potential through the use of a bicubic spline $H_{ij}(N_i^H, N_i^C)$ (Eq. 6), which is a function of the number of neighbors surrounding atom i. To extend a bond-order potential to three atom types, such as Si-C-H, the spline function can be extended into ways. First, the cubic-spline function can be extended to three dimensions, such that it has the form $H_{ij}(N_i^H, N_i^C, N_i^{Si})$. This method provides the most flexibility in adjusting the bond order but can be difficult to code. Alternatively, because C and Si have similar electronegativities and possess similar bonding characteristics, a two-dimensional cubic spline that considers the number of neighbors of C and Si together, i.e., $H_{ij}\left(N_i^H, \left(N_i^C + N_i^{\mathrm{Si}}\right)\right)$, could also be used. This approach was adopted in both the Si-C-H potentials discussed above. This simplification has the advantage of simplifying the MD code. However, it is at the expense of a considerable loss in flexibility in fitting and, therefore, accuracy in reproducing a wide range of properties. For example, the errors in the energies of organosilane molecules calculated with the XB potential systematically increase with increasing molecule size (Fig. 1).

The performance of the XB potential was evaluated by Sbraccia et al. in 2002 [8]. Several



Reactive Empirical Bond-Order Potentials, Fig. 1 Error in atomization energies of $Si_xC_yH_z$ molecules predicted using the mXB potential relative to the atomization energy calculated using the CBS-QB3 method in the ab initio code Gaussian03 illustrating increasing error with atomization energy E_A (increasing molecule size)

problems, such as spurious minima in some potential energy curves, the incorrect energetic ordering for chemisorption of hydrocarbon molecules on Si surfaces, and over-coordinated carbon and silicon atoms were identified. These problems were corrected by altering a minimal set of parameters and the C–C interaction cutoff values. This parameterization is known as the modified extended Brenner (mXB) potential. It should be noted, however, that these changes do not improve the elastic properties of the solid phases of carbon or silicon [9] or the energies of organosilane molecules shown in Fig. 1.

Extension to Metals

Recently, a classical many-body potential for transition metal (Fe, Co, Ni) carbide clusters based on the bond-order type potential was developed [10]. The covalent interactions are based on the first-generation REBO potential. Deviations from the REBO potential formalism occur in the bond-order terms for carbon–metal interactions. The bond order for carbon–metal interactions has the form where b and δ are fitting parameters. For metal–metal interactions, instead of $B^* = \left(1 + b\left(N^C - 1\right)\right)^{\delta}$, multiplying V^A by B^*, the binding energy D_e and equilibrium bond length R_e are expressed as functions of the metal coordination number N^M and are given by

$$D_e = D_{e1} + D_{e2} exp\left\{-C_D\left(N^M - 1\right)\right\} \quad (7)$$

$$R_e = R_{e1} + R_{e2} exp\left\{-C_R\left(N^M - 1\right)\right\}. \quad (8)$$

The reader is referred to Shibuta and Maruyama [10] for all parameter values appearing in Eqs. 7 and 8.

Limitations

Before undertaking MD simulations with empirical potentials, careful thought should be given to the selected potential's inherent strengths and weaknesses and how these may impact the phenomena to be simulated. For example, multiple potential energy functions exist for modeling Si, including three versions of Tersoff's bond-order potential [4,11], SW [1], and the environment-dependent interatomic potential (EDIP) [12]. The majority of empirical potentials for Si do not accurately describe the three elastic constants for diamond cubic silicon [13]. In contrast, the EDIP and the 2B-Si [9] potentials do a good job describing elastic constants. However, the bond lengths for many of the crystalline silicon structures are overestimated when using the EDIP potential. For a more complete comparison of the strengths and weaknesses of Si potentials the reader is referred to Balamane et al. [13] and Schall et al. [9].

In general, four limitations have been identified with the first-generation REBO potential. First, the functional form of the pair potentials was not flexible enough to allow a parameter set that could simultaneously reproduce structural energies and force constants. In his original work, Brenner provided two parameter sets, one that provided an excellent fit to bond energies and a second that provided a to fit force constants [5]. Second, because the zero-Kelvin elastic constants C_{11}, C_{12}, and C_{44} for diamond were not included in the fitting database, they are not accurately reproduced by the potential [14]. Third, both the attractive and repulsive pair terms have finite values at zero separation. Thus, in highly energetic atomic collisions, it may be possible for an atom pass through another atom without being repulsed. Finally, the derivatives of the cutoff function $f_{ij}(r)$ are nonrealistic and may lead to spurious minima in energies for certain structures, particularly amorphous carbon [5].

Second-Generation REBO Potentials

To correct the shortcomings of the original potential, Brenner and coworkers developed a second generation of the REBO potential [15]. The repulsive and attractive pair terms have the forms

$$V^R(r) = f^c(r)(1 + Q/r)Ae^{-\alpha r} \quad (9)$$

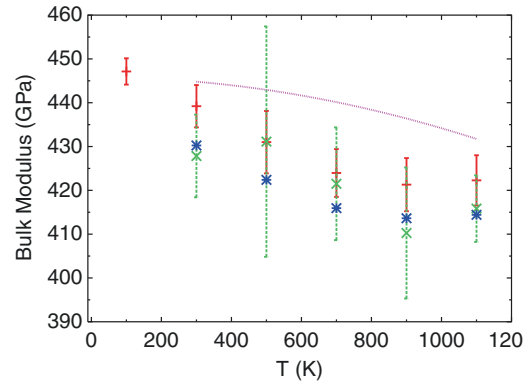
and

$$V^A(r) = f^c(r)\sum_{n=1,3} B_n e^{-\beta_n r} \quad (10)$$

where A, Q, α, B_n and β_n are fitting parameters. The screened Coulomb function used for the repulsive pair interaction goes to infinity as interatomic distances approach zero, and the attractive term has sufficient flexibility to simultaneously fit the bond properties that could not be fit with Morse-type terms. The bond-order B_{ij} term is also significantly different from either the Tersoff or first-generation REBO potential. Separate terms are included that depend on local coordination and bond angles, radical character and conjugation, and dihedral angle for C–C double bonds.

By altering the functional form and expanding the fitting database, the second-generation REBO potential is able to provide significantly better descriptions of bond energies, lengths, and force constants for hydrocarbon molecules relative to the first-generation REBO potential. In addition, the zero-Kelvin elastic properties, interstitial defect energies, and surface energies for diamond are well reproduced. This potential also qualitatively reproduces the correct behavior of the elastic constants C_{11} and C_{44} of diamond with temperature. Coincidentally, the relatively poor agreement of the calculated C_{12} with the experimentally determined data combined with a fast softening of C_{11} with temperature results in bulk moduli that agree fairly well with the experimental values (Fig. 2) [14].

As a first step toward developing a Si–C–H bond-order potential, Schall et al. recently published parameters for silicon potential (2B-Si) based on the second-generation REBO formalism [9]. This potential does a reasonable job producing a broad range of properties of diamond cubic and amorphous silicon. In principle, the second-generation REBO formalism could be extended to a wide range of interatomic interactions to model chemical reactivity in nanoscale systems that contain multiple atom types. For example, Sinnott and coworkers used density functional theory to develop parameters within the second-generation REBO formalism for C–O, O–O, and O–H [16]. Because this potential is based on the second-energy REBO potential, it possesses all of its inherent strengths and weaknesses, such as a lack of intermolecular forces.



Reactive Empirical Bond-Order Potentials, Fig. 2 Bulk modulus of diamond as a function of temperature. Bulk values were calculated from the elastic constants using the relationship $B = \frac{1}{3}(C_{11} + 2C_{12})$. (+ signs), (*crosses*), and (*asterisks*) symbols represent values calculated with the direct, strain fluctuation, and stress fluctuation methods, respectively

The addition of a third atom type complicates the functional form of the bond order as discussed above. This potential suffers from the limitation discussed above by making use of a two-dimensional cubic spline instead of a three-dimensional cubic spline in the bond-order function. In addition, O is treated in a similar manner as C, despite the inherent differences in electronegativities of the two atoms. Additional parameter sets have also been developed recently by this group for Mo–S and C–F.

Covalent + Intermolecular Forces

To model chemical reactions in condensed phases, such as liquids, graphite, and self-assembled monolayers (SAMs), intermolecular forces must be added to the REBO potential. Goddard and coworkers added intermolecular forces to the first-generation REBO potential for hydrocarbons. Distance-based switching functions were used to turn the intermolecular forces "on" and "off." Intermolecular forces and torsional interactions have also been added to the second-generation REBO potential [17]. This potential, known as the adaptive intermolecular reactive empirical bond-order (AIREBO) potential, uses a unique algorithm to determine the conditions under which the intermolecular forces should be "on" and "off."

The AIREBO potential includes intermolecular interactions between nonbonded atoms and torsional interactions associated with a connected sequence of three bonds. This extension makes the AIREBO potential a particularly attractive tool for studying interfacial/tribological systems and liquids. The discussion below focuses on how intermolecular interactions are introduced without compromising the reactivity of the potential. For a discussion of how torsional interactions are modeled, the reader is referred to the original publication [17].

Intermolecular interactions are modeled through a Lennard-Jones (LJ) potential,

$$V^{LJ} = 4\varepsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right). \tag{11}$$

Only four parameters are taken as independent: σ_{CC}, σ_{HH}, ε_{CC} and ε_{HH}. The heterogeneous parameters are fixed by Lorentz-Berthelot combining rules,

$$\sigma_{CH} = \frac{1}{2}(\sigma_{CC} + \sigma_{HH}) \tag{12}$$

and

$$\varepsilon_{\mathrm{CH}} = \sqrt{\varepsilon_{CC}\varepsilon_{HH}}. \tag{13}$$

To merge this with a reactive potential, the LJ potential may be either completely or partially "turned off" in response to the chemical environment of the interacting pair. The key feature of the AIREBO potential is that it provides a means of smoothly interpolating between purely bonded and nonbonded interactions through a set of three switching functions S_{distance}, S_{bond}, and $S_{\mathrm{connectivity}}$. Each one of these switching functions may turn off the LJ interaction partially or entirely,

$$E^{\mathrm{LJ}} = (1 - S_{\mathrm{distance}}S_{\mathrm{bond}})\left(1 - S_{\mathrm{connectivity}}\right)V^{\mathrm{LJ}}. \tag{14}$$

For each switching function, a value of 1 is associated with turning off the LJ interaction completely, 0 is associated with a full LJ interaction, and values in between are associated with partial LJ interactions.

The distance-based switch, S_{distance}, is 1 for distances below r^{LJmin} and 0 for distances above r^{LJmax}. A cubic-spline function is used to interpolate at distances falling between the setpoints r^{LJmin} and r^{LJmax} to give a value for the switch between 0 and 1. The setpoints are fixed by requiring that the LJ potential minimum remain unchanged and that no artificial repulsive barrier is present as the LJ potential is switched off. These constraints require that $r^{\mathrm{LJmin}} = \sigma_{ij}$ and $r^{\mathrm{LJmax}} = 2^{1/6}\sigma_{ij}$.

The hypothetical bond-order-based switch S_{bond} interpolates to give values between 0 and 1 in the region $b^{\mathrm{min}} < b^{*} < b^{\mathrm{max}}$ using a cubic spline, 0 for a bond-order below this range and 1 for a bond-order above this range. The bond-order b^{*} is evaluated at r^{LJmin} for pairs separated by intermolecular distances (distances larger than r^{LJmin}), and thus at these distances b^{*} is a *hypothetical* bond-order that assesses the potential for these atoms to bond they approach one another. In the calculation of b^{*}, the distances of each atom in the pair to its neighbors remain unchanged.

The distance-based switch $S_{\mathrm{connectivity}}$ takes into account the set of all one-, two-, and three-bond sequences that connect the pair of atoms under consideration. For each atom pair in each sequence of bonds, a weight is calculated that is 1 for distances below r^{LJmin} and 0 for distances above r^{LJmax}. These setpoints are the same as in S_{distance}; however, for compatibility with the original REBO potential, a shifted half-period cosine function is used to interpolate between r^{LJmin} and r^{LJmax}. A weight for a sequence of bonds is calculated as the product of the sequence's individual bond weights. The $S_{\mathrm{connectivity}}$ switch is taken to be the largest sequence-weight found in the total set of sequences. In the adopted form for E^{LJ}, the switches S_{distance} and S_{bond} appear together as a product while $S_{\mathrm{connectivity}}$ appears on its own. Consequently, a full LJ interaction will be included for atom pairs that are not (1,2), (1,3), or (1,4) neighbors and are *either* beyond the cutoff distance r^{LJmax} *or* have a bond-order below b^{min}.

The treatment of the van der Waals interactions can be improved in the AIREBO potential by

treating the long-range forces so that the chemical environment is taken into account [18]. In the spirit of the REBO potential, the LJ parameters for carbon are functions of the number of N^C and N, or $\sigma_C(N^C, N^H)$ and $\varepsilon_C(N^C, N^H)$. Two-dimensional cubic-spline functions are used to interpolate between values with integer numbers of neighbors.

Modeling Reactivity in Systems Containing Charge

The electronic properties and nature of bonding in molecules and solids stems from the distribution of charges within the system of interest. The simplest way to add charges to MD simulations is to assign a fixed charge to each atom in the simulation. Fixed-charge schemes have several limitations; however, the most important of which is their inability to redistribute charges in a physically realistic manner. Charge clouds can be distorted when molecules are brought into close proximity, without reacting. When chemical reactions occur, charge needs to be transferred. For instance, the charge states of hydrogen and chlorine change dramatically when they approach each other to form an HCl molecule.

The first atomic-level, charge transfer model was developed by Rappé and Goddard [19] and is known as the electronegativity equalization (QEq) method. In this method, the charges in the system distribute themselves so that the electronegativity at each atomic site is equalized for a given nuclear configuration. This is equivalent to saying that the driving force to shift charge from one atom to another is zero, or the derivatives of the potential with respect to charge are zero. The self-energy $E_i^s(qi)$ on particle i is a function of the charge of the ion and is expressed in terms of the charge, electronegativity χ, and atomic hardness (a term related to the change in energy as a function of charge). Obtaining the charge on each atom for each nuclear configuration of N atoms requires taking the inverse of an N by N matrix. This method was utilized by Streitz and Mintmire in their examination of the aluminum/alumina interface using their ES + EAM potential [20]. Inversion of an N by N matrix at each MD time step usually scales as $O(N^2)$ and becomes computationally intractable as the size of the system increases. Several alternative numerical methods have been developed to solve the N linear equations in charge associated with the N by N matrix. To date, these methods have not found widespread application.

To avoid the matrix inversion, Rick et al. developed an approximate method to equalize the electronegativity based on an extended Lagrangian approach. In this method, charges are treated as dynamic variables that evolve explicitly with time [21]. Equations of motion for the positions and charges of the atoms can be shown to be:

$$m_i \ddot{r}_i = \frac{\partial}{\partial r_i} E_T(\{r_i\}\{q_i\}) \tag{15}$$

and

$$s_i \ddot{q}_i = \frac{\partial}{\partial q_{r_i}} E_T(\{r_i\}\{q_i\}) \tag{16}$$

where m_i, r_i, and q_{r_i} are the mass, position, and charge of the ith atom. The term s_i is a fictitious mass of the charge, which is chosen to provide numerical stability. The individual charges respond to deviations of the electronegativity equilibration by moving to a new charge which more closely satisfies the equilibrium condition. One advantage of this approach is that both equations of motion Eqs. 15 and 16 can be integrated using standard numerical integration algorithms. In addition, because the matrix-based solution for determining charge is not used, the code remains $O(N)$, a crucial requirement for large-scale simulations.

The Lagrangian-based dynamic-charge method, which amounts to an approximate implementation of the electronegativity equalization condition, suffers from one significant drawback, however. The dynamic-charge method is closely related to a single iteration, steepest-descent search for the electrostatic energy minimum. The magnitude of the charge transfer is controlled by a

constant factor for all atoms and simulation steps. Thus, the charge "lags" behind the atomic positions used in minimizing the energy and, as a result, the electronegativity equalization condition is not satisfied. To remedy this situation, Ma and Garofalini have introduced the iterative fluctuation charge (IFC) model [22]. The IFC model is able to satisfy the electronegativity equalization condition by performing multiple iterations for electrons (charge) while keeping the atomic positions fixed. The exact numerical criterion, or convergence factor Δ, for electronegativity equalization is $\Delta\left(\frac{eV}{atom}\right)=\frac{\sqrt{\left(\chi_i-\chi_{eq}\right)^2}}{N}=0$, where χ_i and χ_{eq} are the electronegativities on the ith atom and the equilibrium electronegativity. In practice, a small convergence factor, such as $\Delta=1.0\times10^{-8}$ is used in MD simulations. In addition, charges will oscillate about their fully converged, equilibrium values unless a damping factor is used. The selection of the damping factor and the fictitious charge-mass values can be optimized to achieve optimal damping while not preventing thermal coupling between atomic and electronic degrees of freedom.

Recently, the extended Lagrangian approach with the damping factor discussed by Ma and Garofalini [22] was used to examine the interface between Si and SiO_2 surfaces using a charge optimized many-body (COMB) potential [23]. The potential has the form:

$$E_T=\sum_i\left[E_i^S+\frac{1}{2}\sum_{j\neq1}V_{ij}\left(r_{ij},q_i,q_j\right)+E_i^{BB}\right], \tag{17}$$

where E_T is the total potential energy of the system, E_i^S is the self-energy term of atom i, V_{ij} is the interatomic potential between the ith and jth atoms, r_{ij} is the distance of the atoms i and j, and q_i and q_j are charges of the atoms, and E_i^{BB} is the bond-bending term of atom i. The interatomic potential energy V_{ij} consists of four components: short-range repulsion, U_{ij}^R, short-range attraction, U_{ij}^A, long-range Coulombic interaction, U_{ij}^I, and long-range van der Waals energy, U_{ij}^V, which are defined as

$$V_{ij}\left(r_{ij},q_i,q_j\right)=U_{ij}^R\left(r_{ij}\right)+U_{ij}^A\left(r_{ij},q_i,q_j\right)+U_{ij}^I\left(r_{ij},q_i,q_j\right)+U_{ij}^V\left(r_{ij}\right), \tag{18}$$

The repulsive and attractive terms in Eq. 18 are of the Tersoff form:

$$U_{ij}^R=f(r)Ae^{-\lambda r}\text{ and }U_{ij}^A=f(r)b_{ij}Be^{-\alpha r} \tag{19}$$

In Tersoff's original work, A, B, λ, and α are constants and $f(r)$ is a distance-based cutoff function, which limits the range of the potential. In shorthand notation, the energy contribution due the repulsive and attractive terms (i.e., the short-range pair potential) can be written as

$$E_{pair}=\sum_{j\neq i}f(r)Ae^{-\lambda r}+\sum_{j\neq i}f(r)b_{ij}Be^{-\alpha r}=V^R+bV^A \tag{20}$$

where b_{ij} is the bond-order function.

Modeling Charge Transfer During Chemical Reactions

Model fluctuating charges using the extended Lagrangian approach seems ideally suited to examine chemical reactions using MD. However, the use of this approach requires the definition of a charge-neutral entity, such as a molecule or the simulation cell. For example, Yu et al. recognized the problem with defining a charge-neutral simulation cell when simulating Si/SiO_2 interfaces. To circumvent this problem, the self-energy terms of the COMB potential were modified [23]. Rick et al. defined each water molecule as a charge-neutral entity in their examination of the structural properties of water [21]. Charge equilibration (QE) methods have been applied successfully to relatively homogeneous small-molecule systems near equilibrium. However, application of these methods to large molecules yields atomic charges and molecular dipole moments that are too large. This problem has been traditionally addressed by constraining sub-entities of large molecules to be charge neutral. This is problematic in a reactive environment where it would not be possible to

meaningfully adhere to a fixed set of molecule-based charge-neutral entities.

Recently, Mikulski et al. developed a method to integrate charge equilibration with bond-order potentials that does not require the assignment of charge neutrality to molecules [24]. The bond-order potential/charge equilibration (BOP/SQE) method is a bond-centered approach, where the charge of atom i is the sum of all charges transferred to it across each of its bonds. This has been referred to as a split-charge equilibration (SQE). The problem of the overestimation of dipole moments of large molecules is not solved by merely adopting the bond-centered (SQE) approach. A novel method is needed whereby each bond partly decouples from its bond network and settles into its own equilibrium. The BOP/SQE method connects the bond order of each bond to an amount of shared charge in each bond and interprets the split charge as an imbalance in where the shared charge is concentrated. Split charges cannot grow in size beyond this shared-charge limit, and if the nature of the bond is covalent, the size of the split charges should not get close to this limit. Thus, a fractional split charge is defined as the ratio of the split charge $\overline{f}_{ij}$ on atom i transferred from atom j to its shared charge $\overline{q_{ij}^{\max}}$, charge on atom i is expressed as

$$Q_i = \sum\nolimits_j \overline{f}_{ij}\, \overline{q_{ij}^{\max}}. \tag{21}$$

The shared-charge limit $\overline{q_{ij}^{\max}}$ is calculated directly from the bond order bij assigned to the bond from the BOP, $\overline{q_{ij}^{\max}} = \overline{q_{ij}^{\max}}\left(b_{ij}\right)$. Focus is now shifted from equilibrating split charges to equilibrating fractional split charges.

The equilibration process for bond ij is controlled by three factors: (1) The constant pull to increase charge separation associated with the traditional constant electronegativity difference $\left|\chi_i - \chi_j\right|$; (2) Coulomb interactions within the bond and with the surrounding environment; (3) A restoring force that approaches and infinite wall as the shared-charge limit is approached. The first two factors are analogous to the SQE condition but are applied to $\overline{f}_{ij}$ rather than $\overline{q}_{ij}$. The third, unique, element is implemented through inverse hyperbolic tangent function added to the charge-dependent electronegativity of each fractional split charge,

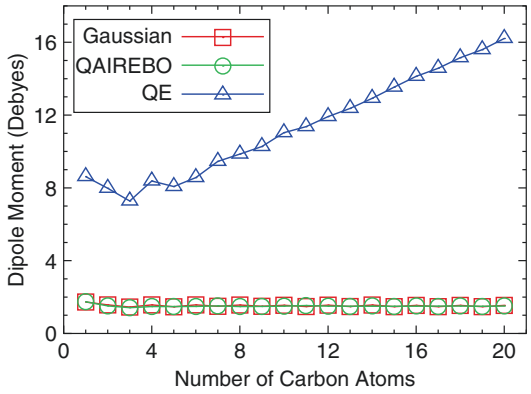
$$\overline{\chi}_{ij} = \left[\chi_i + J_{it}^{o} Q_i + \sum\nolimits_{k \neq i} J_{ij} Q_k\right] + \overline{\epsilon}_{ij} \tan h^{-1}\left(\overline{f}_{ij}\right). \tag{22}$$

The scale parameter $\overline{\epsilon}_{ij}$ can be tailored to each pair of atom types. Large values of $\overline{\epsilon}_{ij}$ cause the wall to gradually rise as soon as the split charges grow. The last term in this equation, or the charge-asymmetry penalty, serves to physically constrain the growth of the split charges and partially decouples each bond from its bond network. Therefore, the BOP/SQE formalism becomes a locally driven process that does not map back to the global equilibration of an entire molecule. With the selection of the scale parameter $\overline{\epsilon}_{ij}$ to reproduce the dipole moment of water, this method was shown recently to reproduce the correct trends in molecular dipole moments of long-chain alcohols (Fig. 3).

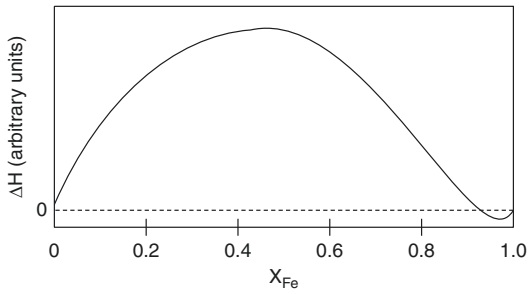
It should be noted, that other methods for handling chemical reactions where charge is included are being developed [25]. Because these methods are not bond-order potentials, they are not included in this overview.

Improved Description of Alloys

In a recent study of the thermodynamics an empirical potential description of Fe–Cu alloys, Caro et al. concluded that to study alloys in an empirical framework, heteronuclear interactions require additional modifications to the empirical formalism to incorporate concentration dependent interactions that reproduce the magnitude of excess enthalpy of mixing and asymmetry around the equi-atomic composition [26]. To illustrate this point, data for the heat of mixing in the Fe-Cr system as determined in an ab initio study conducted by Olsson has been replotted in Fig. 4. This figure clearly illustrates the deviation from ideality with both positive and negative heat of mixing in evidence.



Reactive Empirical Bond-Order Potentials, Fig. 3 Molecular dipole moments of straight chain alcohols calculated using Gaussian, charge equilibration (QE), and a fluctuating-charge MD code that uses the BOP/QE method of charge equilibration and qAIREBO potential



Reactive Empirical Bond-Order Potentials, Fig. 4 The enthalpy of mixing ΔH for the Fe–Cr system as a function of Fe fraction. (replotted from the work of Olsson) The solid line was calculated via first-principles. The dashed line indicates a linear interpolation corresponding to an ideal solution

Traditionally, empirical models for alloys use the Lorentz-Berthelot mixing rules (Eqs. 12 and 13). These rules reproduce the ideal heat of mixing, which is, strictly speaking, only accurate for dilute solutions of solute atoms in the host matrix. Caro et al. has proposed a method for reproducing the thermodynamic properties of concentrated alloys through a modification of the embedded atom method (EAM) formalism. The basic EAM description gives the total energy as a the sum of a pair potential, V, and an embedding function, F, which in turn depends on the electron density, ρ

$$E_{EAM} = \frac{1}{2}\sum V(r) + \sum F\left(\sum \rho(r)\right) \quad (23)$$

Consideration of first-principles calculations gives important information about the behavior of these functions. First, the embedding energy relative to the free atom energy goes to zero for zero electron density and should have a negative slope and positive curvature for the background electron density in metals. Second, the pair interaction term $V(r)$ is purely repulsive.

Caro et al. eschewed the commonly used Lorentz-Berthelot mixing rules and developed a new functional form for the cross potential for random alloy mixtures:

$$E_{RAND} = x_A^2 V_{AA} + x_B^2 V_{BB} + 2x_A x_B V_{AB} + x_A F_A(\widetilde{\rho}) + x_B F_B(\widetilde{\rho}). \quad (24)$$

Here x_A and x_B are the atomic fractions of alloy element A and B, respectively. The pair potential for each element denoted as V_{AA} or V_{BB}. The model adopts the assumption that atoms are embedded in the same average environment, through $F(\widetilde{\rho})$ where $\widetilde{\rho}$ is the average electron density seen by either atom type A or B in the host material. Caro et al. found that the contributions to the formation energy from the embedding term were negligible, leaving only the pair potential as the sole contributor to the formation energy for alloys. By assuming V_{AB} is a function of both composition x and distance r and can be separated into a product $h(x)$ $u_{AB}(r)$, an appropriate choice of the cross term is

$$V_{AB}(x,r) = h(x)\frac{1}{2}[V_A(r) + V_B(r)] \quad (25)$$

Where $h(x)$ is a fourth-order polynomial. This selection of the cross term allows for the description of any type of formation energy curves, with an ideal solution of $h(x) = 1$, for a regular solution. Through appropriate fitting of the polynomial coefficients, a positive or negative heat of mixing can be achieved. Stukowski et al. recently published a version of this potential function.

The methodology described above can also be extended to Tersoff-type potentials. The foundation of this approach was laid out by Brenner who

recognized that the EAM potential and Tersoff many-body potential are formally identical in the appropriate limits. The essential result is that it is reasonable to assume that embedding function $F(\rho(r))$ in the EAM formalism is equal to the attractive part of the Tersoff pair potential, bV^A. Similarly, it is a logical to assume that the repulsive terms in both EAM and Tersoff are equivalent. With these points in mind, a combination of the Tersoff potential with Caro's compositional dependence method can be expressed in shorthand notation as

$$E_{RAND} = x_A^2 V_{AA}^R + x_B^2 V_{BB}^R + 2x_A x_B V_{AB}^R + x_A \tilde{b} V_{AA}^A + x_B \tilde{b} V_{BB}^A. \quad (26)$$

where V_{AA} and V_{BB} are the repulsive and attractive pair terms, denoted by superscripts ^{R} or ^{A}, respectively, for the pure elements A or B as given in Eq. 3. The repulsive cross term V_{AB} is given in Eq. 26. Schall et al. have fit a multicomponent bond-order potential for Si–C–H systems, and it was found that the average bond-order $\tilde{b}$ must be normalized using a scaling factor to ensure compatibility between bond orders for derived for different pure element systems. The average bond order is then given as

$$\tilde{b} = \chi_{ij} \frac{b_{ij} + b_{ji}}{2}. \quad (27)$$

The normalizing factor χ_{ij} is equal to one for homonuclear atom pairs. For heteronuclear atom pairs, χ_{ij} is determined by fixing the average bond order to a value of one for a specified crystal structure. For example, in the Si–C system, χ_{ij} was determined by setting the average bond order for the β-Zn phase of SiC solving Eq. 27 for χ_{ij}. A similar scheme was utilized by Tersoff for multicomponent systems.

Summary and Future Directions

In the foregoing sections, a discussion of reactive empirical bond-order potentials was given. Such potentials were developed to describe the universal relationship between binding energy and bond length and have been extended from simple descriptions of single elements to multicomponent systems and then to a combination of solids and molecules. The motivation for the development of these potentials was to provide a computationally efficient means to model chemical reactivity, including an accurate description of bond-breaking and forming processes, while still using sound quantum mechanical principles. The original formalism has since been expanded to include descriptions of long-range interactions, torsion, charge transfer, and descriptions of alloys. This family of potentials has since been used to describe a wide variety of processes in nanomaterials systems. One serious limitation of the family of reactive empirical bond-order potentials is that only a relative few materials systems have been parameterized. As mentioned previously, parameterization of a potential involves quite a bit of art as well as science. In the future, semi-automated methods such as force matching that combine ab initio force calculations systematically with parameter fitting could help overcome this hurtle. In addition, the incorporation of more than three elements presents a significant challenge due to the complexity of the spline functions used to interpolate the bond order correction terms. To overcome this barrier, either compromises in the accuracy of the potential or new developments in multidimensional interpolation functions must be made.

Cross-References

▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling

References

1. Stillinger, F.H., Weber, T.A.: Computer simulation of local order in condensed phases of silicon. Phys. Rev. B **31**, 5262 (1985)

2. Mackerell, A.D.: Empirical force fields for biological macromolecules: overview and issues. J. Comput. Chem. **25**, 1584 (2004)
3. Abell, G.C.: Empirical chemical pseudopotential theory of molecular and metallic bonding. Phys. Rev. B **31**, 6184 (1985)
4. Tersoff, J.: New empirical model for the structural properties of silicon. Phys. Rev. Lett. **56**, 632 (1986)
5. Brenner, D.W.: Empirical potential for hydrocarbons for use in simulating chemical vapor deposition of diamond films. Phys. Rev. B **42**, 9458 (1990)
6. Murty, M.V.R., Atwater, H.A.: Empirical interatomic potential for Si-H interactions. Phys. Rev. B **51**, 4889 (1995)
7. Dyson, A.J., Smith, P.V.: Extension of the brenner empirical interatomic potential to C-Si-H systems. Surf. Sci. **355**, 140 (1996)
8. Sbraccia, C., Silvestrelli, P.L., Ancilotto, F.: Modified XB potential for simulating interactions of organic molecules with Si surfaces. Surf. Sci. **516**, 147 (2002)
9. Schall, J.D., Gao, G., Harrison, J.A.: Elastic constants of silicon materials calculated as a function of temperature using a parametrization of the second-generation reactive empirical bond-order potential. Phys. Rev. B **77**, 115209 (2008)
10. Shibuta, Y., Maruyama, S.: Bond-order potential for transition metal carbide cluster for the growth simulation of a single-walled carbon nanotube. Comput. Mater. Sci. **39**, 842 (2007)
11. Tersoff, J.: Empirical interatomic potential for silicon with improved elastic properties. Phys. Rev. B **38**, 9902 (1988)
12. Justo, J.F., Bazant, M.Z., Kaxiras, E., Bulatov, V.V., Yip, S.: Interatomic potential for silicon defects and disordered phases. Phys. Rev. B **58**, 2539 (1998)
13. Balamane, H., Halicioglu, T., Tiller, W.A.: Comparative study of silicon empirical potentials. Phys. Rev. B **46**, 2250 (1992)
14. Gao, G.T., Van Workum, K., Schall, J.D., Harrison, J. A.: Elastic constants of diamond from molecular dynamics simulations. J. Phys. Condens. Matter **18**, S1737 (2006)
15. Brenner, D.W., Shenderova, O.A., Harrison, J.A., Stuart, S.J., Ni, B., Sinnott, S.B.: Second generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. J. Phys. Condens. Matter. **14**, 783 (2002)
16. Ni, B., Lee, K.H., Sinnott, S.B.: A reactive empirical bond order (REBO) potential for hydrocarbon-oxygen interactions. J. Phys. Condens. Matter **16**, 7261 (2004)
17. Stuart, S.J., Tutein, A.B.., Harrison, J.A.: A reactive potential for hydrocarbons with intermolecular interactions. J. Chem. Phys. **112**, 6472 (2000)
18. Liu, A., Stuart, S.J.: Empirical bond-order potential for hydrocarbons: adaptive treatment of van der Waals interactions. J. Comput. Chem. **29**, 601 (2008)
19. Rappe, A.K., Goddard, W.A.: Charge equilibration for molecular-dynamics simulations. J. Phys. Chem. **95**, 3358 (1991)
20. Streitz, F.H., Mintmire, J.W.: Electrostatic potentials for metal-oxide surfaces and interfaces. Phys. Rev. B **50**, 11996 (1994)
21. Rick, S.W., Stuart, S.J., Berne, B.J.: Dynamical fluctuating charge force-fields: applicatoin to liquid water. J. Chem. Phys. **1001**, 6141 (1994)
22. Ma, Y., Garofalini, S.H.: Iterative fluctuation charge model: a new variable charge molecular dynamics method. J. Chem. Phys. **124**, 234102 (2007)
23. Yu, J.G., Sinnott, S.B., Phillpot, S.R.: Charge optimized many-body potential for the Si/SiO2 system. Phys. Rev. B **75**, 085311 (2007)
24. Mikulski, P.T., Knippenberg, M.T., Harrison, J.A.: Merging bond-order potentials with charge equilibration. J. Chem. Phys. **131**, 241105 (2009)
25. van Duin, A.C.T., Dasgupta, S., Lorant, F., Goddard, W.A.: ReaxFF: a reactive force field for hydrocarbons. J. Phys. Chem. A **105**, 9396 (2001)
26. Caro, A., Turchi, P.E.A., Caro, M., Lopasso, E.M.: Thermodynamics of an empirical potential description of Fe-Cu alloys. J. Nucl. Mater. **336**, 233 (2005)

Reactive Ion Beam Etching (RIBE)

▶ Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Realtime Green's Function Approach

▶ Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism

REBO

▶ Reactive Empirical Bond-Order Potentials

Reduced Graphene Oxide

▶ Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites

Refractometric Sensing Using Plasmonic Nanoparticles

Erik Martinsson and Daniel Aili
Division of Molecular Physics, Department of Physics, Chemistry and Biology, Linköping University, Linköping, Sweden

Synonyms

LSPR-based sensing; Nanoplasmonic sensing

Definition

Refractometric sensing using plasmonic nanoparticles exploits the excitation of localized surface plasmons in metal nanostructures to monitor changes in the dielectric properties in the close vicinity of the nanoparticle surface, caused by, for instance, the binding of a biomolecule to immobilized receptors. The sensitivity to small changes in refractive index enables real-time and label-free detection of molecular interactions at the metal nanoparticle surface.

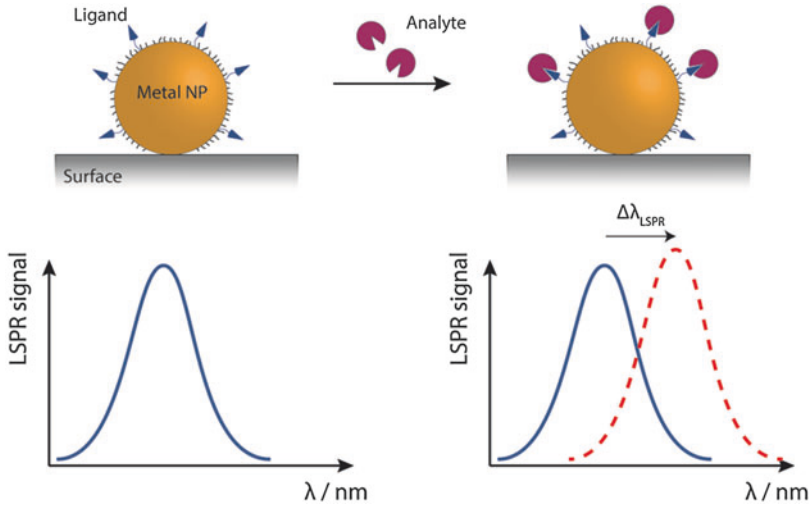
Overview

Surface plasmons are collective oscillations of the free electron gas at the surface of metals that can be induced by an external electric field, including electromagnetic radiation. When the size of the metal structures is reduced to nanoscale dimensions, the electron oscillations are confined in the nanostructures creating a phenomenon known as localized surface plasmon resonance (LSPR). The excitation of LSPR gives rise to a strong enhancement of electromagnetic fields in the near-field region, and the ability to concentrate electromagnetic radiation into particles of nanoscale dimensions has made plasmonic materials attractive for a large variety of applications, including molecular sensing. For most types of silver and gold nanostructures, the resonance frequency is in the visible region of the electromagnetic spectrum, which gives rise to distinct colors for nanoparticle ensembles. The colors emanate from the extensive absorption and scattering (extinction) of light by the nanoparticles at the LSPR frequency. The resonance conditions are highly dependent on a number of parameters, including (i) the nature of the nanostructures, i.e., their size, shape, and composition; (ii) their interaction with other, closely spaced metal nanostructures; and (iii) the dielectric properties of the surrounding medium. The latter enables metal nanostructures to be used as sensitive probes for detecting changes in the refractive index (RI) in their close proximity caused by for instance the adsorption of biomolecules to the metal surface. The RI dependency of the LSPR is the basis of nanoplasmonic refractometric sensing (Fig. 1). The highly localized sensing volume around metal nanostructures makes these materials attractive signal transducers in miniaturized sensors with high sensitivity to local RI changes.

Metal nanoparticles of different sizes and geometries, such as spherical, rod shaped, and triangular, can be prepared by wet-chemical synthesis and display plasmon resonance frequencies over the entire visible and near-infrared region of the electromagnetic spectrum. Moreover, metal nanoparticles can be immobilized on solid substrates by self-assembly for cost-efficient production of large nanoplasmonic arrays (>1 cm^2), which facilitates development of inexpensive LSPR-based biosensing devices.

Historical Background

The vibrant colors of metal nanoparticles have fascinated mankind for centuries and have been extensively used to decorate glass and ceramics. However, an explanation of the origin of these colors was not presented until the mid-nineteenth century by the pioneering work of Michael Faraday. Faraday synthesized gold nanoparticles by reducing gold salt with phosphorus, and after conducting a series of experiments, he concluded that the colors were caused by the presence of small particles of metallic gold that were suspended in the aqueous media [1]. Some decades



Refractometric Sensing Using Plasmonic Nanoparticles, Fig. 1 Schematic illustration of the basic concept behind refractometric sensing using plasmonic nanoparticles. Molecular interactions that take place at the surface of a metal nanoparticle change the conditions for the LSPR excitation that result in measurable changes in the extinction spectra, which enables real-time and label-free detection of the binding event

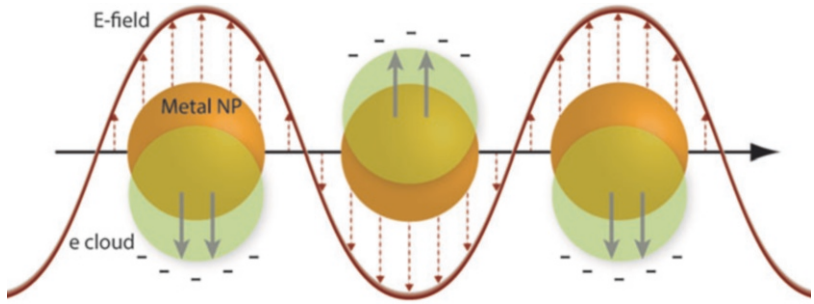
later, Richard Zsigmondy could verify the nanoscale dimensions of these particles by investigating them using an ultramicroscope. The techniques for fabrication and characterization of nanomaterials have advanced considerably since then, and it is now possible to produce well-defined nanostructures with respect to physical and chemical properties and to investigate their properties in great detail.

Refractometric sensing using surface plasmons was first described by Liedberg et al. in 1983 [2]. In this seminal work, propagating surface plasmons at the interface between a planar silver film and a dielectric were exploited to monitor antigen-antibody interactions. It took an additional 15 years before Englebienne showed that the same sensing principle also could be applied to localized surface plasmons in metal nanoparticles [3]. Both sensing schemes offer label-free and real-time detection of molecular interactions taking place at (or close to) the metal surface. Biosensing based on surface plasmon resonance (SPR) is today a commercially established technique that is frequently used to characterize a wide variety of biomolecular interactions. LSPR-based techniques are just about to find their way into the commercial market. In contrast to SPR, LSPR enables cheap, miniaturized, and multiplexed sensing using very simple optical setups.

Theory

Metals have partially filled conduction bands, which give rise to "free" electrons that can move within the material. When metal nanostructures are exposed to an external electric field, these electrons are displaced relative to the positive core of the atoms. The resulting induced dipole will oscillate with the same frequency as the external electric field (e.g., electromagnetic radiation). For small nanostructures the electrons will oscillate coherently at a specific frequency (Fig. 2). This creates a distinct peak in the extinction spectrum at the resonance frequency, known as the LSPR peak.

In 1908, Gustav Mie described the interaction between light and small metal nanoparticles in a homogenous medium by analytically solving Maxwell's equations [4]. He derived an expression for the optical extinction (sum of absorption and scattering) cross section (σ_{ext}) for metal nanoparticles much smaller than the wavelength of light ($2r << \lambda$):



Refractometric Sensing Using Plasmonic Nanoparticles, Fig. 2 Schematic illustration of localized surface plasmon resonance in spherical gold nanoparticles induced by an external electric field

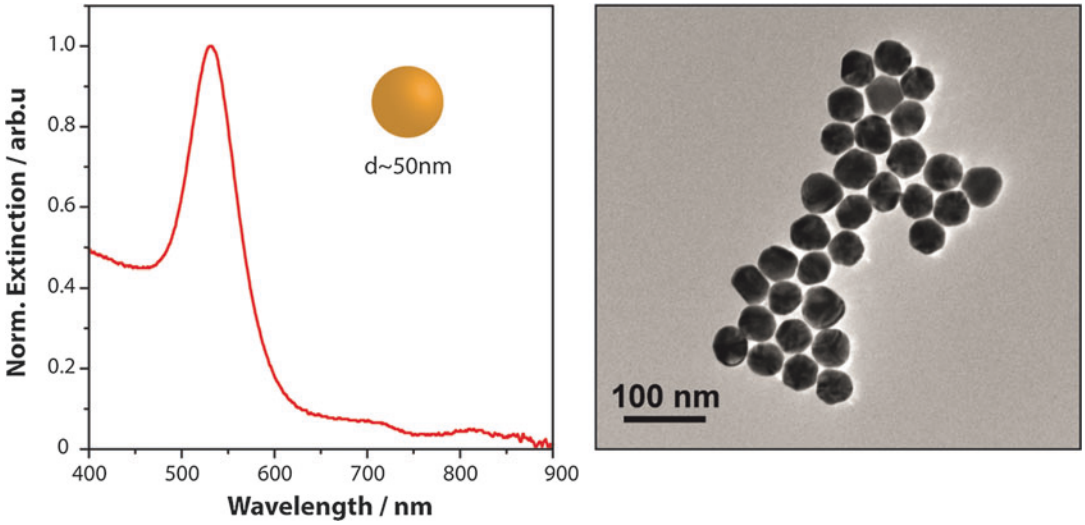
$$\sigma_{ext}(\lambda) = \frac{24\pi^2 r^3 \varepsilon_d^{3/2} N}{\lambda \ln(10)} \frac{\varepsilon_i(\lambda)}{(\varepsilon_r(\lambda) + 2\varepsilon_d)^2 + \varepsilon_i(\lambda)^2} \quad (1)$$

where r is the radius of the nanoparticle, ε_d is the dielectric constant of the surrounding medium, N is the electron density, and ε_r and ε_i are the real and imaginary part of the complex dielectric function of the bulk metal, respectively. The plasmon resonance condition is fulfilled when $\varepsilon_r(\lambda) = -2\varepsilon_d(\lambda)$ if the imaginary part of the dielectric function is small (low losses). For small gold and silver nanoparticles, this resonance frequency falls in the visible region of the electromagnetic spectrum (gold ~ 520 nm and silver ~380 nm) which explains their distinct colors. However, the resonance frequency can be changed by varying the size and shape of the nanoparticles; thus, the plasmon resonance peak can be tuned over the entire visible region. For larger nanoparticles, the dipolar approximation is no longer valid since the electrons will not experience a homogeneous electric field (phase differences) which means that the expression in Eq. 1 cannot be used to accurately determine their optical properties. Instead, complex numerical methods like discrete dipolar approximation (DDA) and finite-difference time-domain (FDTD) simulations are frequently employed to describe the optical response for larger, spherical, or anisotropic nanoparticles.

Equation 1 shows that the extinction depends on the size and the optical properties of the metal nanoparticles as well as the dielectric properties of the surrounding medium. Consequently, the LSPR excitation will change when the refractive index of the surrounding medium changes. Biomolecules have a higher refractive index (~1.46) than water (1.33). A binding event of a molecular analyte to a ligand immobilized on the nanoparticle surface displaces water molecules and replaces them with the biomolecule, resulting in an effective increase in the refractive index of the medium surrounding the nanoparticle and a concomitant change in the extinction cross section.

Synthesis of Metal Nanoparticles

Procedures for metal nanoparticle synthesis have improved significantly during the last decades, and it is now possible to produce metal nanoparticles with a range of different structures and morphologies and that are composed of either a single or several different materials, resulting in nanostructures with unique optical properties. One of the most commonly used wet chemistry techniques to produce metal nanoparticles is based on metal ion nucleation using citrate reduction. This method was developed by Turkevich et al. in 1951, where gold chloride ($HAuCl_4$) was reduced by sodium citrate to generate fairly monodisperse gold nanoparticles in the size range of approximately 10–50 nm in diameter (Fig. 3) [5]. The sodium citrate acts both as a reducing agent and electrostatic stabilizer of the resulting nanoparticles. The size of the nanoparticles can be tuned by varying the ratio between the concentrations of gold salt and sodium citrate in water. A common synthesis method to obtain anisotropic gold nanoparticles is the seed-mediated growth method, which



Refractometric Sensing Using Plasmonic Nanoparticles, Fig. 3 Extinction spectra (*left*) and transmission electron microscopy (*TEM*) image (*right*) of spherical gold particles with an average diameter of 50 nm synthesized by citrate reduction of gold salt

utilizes reagents, typically surfactants in order to restrict the growth at certain crystal facets. This method has been used to realize a wide variety of nanostructures including nanorods, nanocubes, stare-shaped nanoparticles, and nanoprisms, which all show different LSPR frequencies.

Refractive Index Sensitivity

Plasmonic metal nanoparticles can be utilized as transducer elements in refractometric sensors because of the sensitivity of LSPR to small changes in refractive index in the immediate vicinity of their surface. The sensitivity (η_B) is defined as the magnitude of the spectral response caused by a change in refractive index (Δn):

$$R = \eta_B \Delta n \quad (2)$$

where the spectral response (R) can be measured as a change in the plasmon peak wavelength position (nm) or in the signal intensity (extinction). This linear relationship refers to RI changes in the entire sensing volume around the nanoparticles, and η_B is therefore normally referred to as the bulk RI sensitivity, usually expressed in nm/RIU or ext/RIU. In order to achieve a large response for a specific RI change, the bulk RI sensitivity should obviously be as high as possible. Consequently, there are numerous reports on the synthesis of metal nanoparticles exhibiting different RI sensitivities, ranging from ~50 nm/RIU for small, spherical gold nanoparticles to >1000 nm/RIU for more complex, anisotropic nanoparticles [6]. However, Miller et al. have shown that the bulk RI sensitivity for metal nanoparticles is strongly related to their plasmon resonance frequency, and a linear relationship between the two parameters was obtained for particles with a plasmon resonance below 800 nm [7]. Accordingly, nanoparticles that exhibit a plasmon resonance at longer wavelength will display higher bulk RI sensitivity, which means that the aim of creating nanoparticles with high bulk RI sensitivity has merely been quests for synthesizing nanoparticles with longer plasmon resonance wavelength, i.e., lower energy. Silver nanoparticles generally exhibit higher sensitivity as compared to gold nanoparticles because of the different dielectric properties of the two metals. Silver nanoparticles can, however, easily undergo oxidation, which makes them less suitable for most sensing applications.

Metal nanoparticles that exhibit a high bulk RI will not necessarily exhibit a high sensing

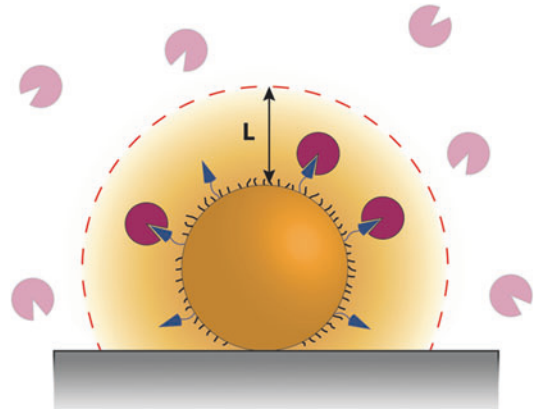
performance since also the resolution must be taken into account. Large nanoparticles that display a plasmon resonance at longer wavelengths (in the near-infrared region) usually have a broad plasmon band due to dephasing and increased radiative damping. This makes it harder to accurately determine the plasmon peak maxima, which has a negative effect on the resolution. The figure of merit (FOM) is a measure that takes both the sensitivity and the resolution into account and is expressed as:

$$\mathrm{FOM} = \frac{\eta_B}{\mathrm{fwhm}} \tag{3}$$

where η_B is the bulk RI sensitivity and fwhm is the full width at half maximum, i.e., the plasmon peak width (in nm) at half its maximum value. Otte et al. have shown that the FOM value for gold nanorods is maximized for a plasmon resonance close to 700 nm, i.e., where an optimum relationship between sensitivity and resolution can be observed [8]. Also, it is advantageous to stay in this region since cheaper optical components can be used for plasmon resonance excitation and detection.

Sensing Volume

The sensing volume around a metal nanoparticle depends on the spatial distribution of the electromagnetic near field created as a result of the excitation of localized surface plasmons. The intensity of the near field decays exponentially from the metal surface, and the decay length is defined as the distance where the magnitude has decreased by 1/e (~37 %). Both the decay length (~10–20 nm) and the intensity of the near field depend on the physical properties of the metal nanoparticle and the dielectric properties of the surrounding medium. The sensing distance, i.e., the average distance from the metal surface where a change in RI can be detected, is strongly correlated to the decay length of the electric field. Numerous studies have been carried out in order to determine the sensing distance for various metal nanoparticle structures using different layer-by-layer techniques. In general, the reported



Refractometric Sensing Using Plasmonic Nanoparticles, Fig. 4 Schematic illustration of the highly localized sensing volume around a gold nanoparticle. The sensing volume depends strongly on the decay length (L) of the electromagnetic near field

sensing distances range between 20 and 30 nm. Hence, the ability of metal nanoparticles to detect changes in their surroundings is strictly limited to a small volume around the nanoparticles, which makes them suitable to probe highly localized RI changes, caused by for instance a molecular binding event to an immobilized receptor (Fig. 4). As the sensing response obtained as a result of adsorption of molecules on the nanoparticle surface is more relevant than the bulk RI sensitivity, an expression for the plasmon shift that can be expected when a molecular layer is formed around a metal nanostructure has been derived [9]:

$$\Delta\lambda_{\max} = \eta_B\left(n_{\mathrm{layer}} - n_{\mathrm{medium}}\right)\left(1 - e^{-\frac{2d}{L}}\right) \tag{4}$$

where n_{layer} and n_{medium} are the refractive indices of the adsorbed layer and the surrounding medium, respectively, d is the thickness of the adsorbed molecular layer, and L is the decay length of the electric field. Although, this expression is based on empirical findings, it gives a rather good prediction of the sensing response that can be obtained when a thin film of molecules are adsorbed at the surface of a metal nanoparticle.

Conventional SPR, based on propagating surface plasmons, exhibits a bulk RI sensitivity that is roughly one order of magnitude higher than LSPR. Despite this fact, the two sensing schemes

show a similar sensing performance when considering molecular adsorption close to the metal interface [10]. As the sensing volume of metal nanoparticles is small, the background signal, also known as the bulk signal, can be significantly lower than in conventional SPR-based sensors, which is advantageous in the analysis of low molecular compounds that will only induce small RI changes.

However, the strictly limited sensing volume associated with LSPR-based sensing limits the usage of bulkier surface chemistries and large recognition molecules (e.g., antibodies), since the molecular interactions of interest then will be located outside the actual sensing volume. Thin surface coatings and small recognition elements are thus appropriate to employ in LSPR-based sensing, ideally keeping the molecular interactions in the direct vicinity of the nanoparticle surface. However, it is a great challenge to realize a thin, functional surface coating that exhibits a high resistance against unspecific molecular adsorption.

Single Nanoparticle Detection

The highly localized sensing volume around metal nanoparticles makes them suitable transducer elements for miniaturized sensing. Single nanoparticle detection has been an area of intense research since this technique enables a significant reduction in the absolute detection limit, i.e., fewer analytes per nanoparticle, which is important both for miniaturization and multiplexed sensing. Single nanoparticles also show narrower plasmon bands as compared to an ensemble of particles, which improve the spectral resolution since smaller LSPR shifts can be detected. Dark-field microscopy techniques enable simultaneous single particle spectral plasmon mapping of multiple nanoparticles for single nanoparticle analysis. Although metal nanoparticles have a large extinction cross section, the scattered light intensity from a single nanoparticle is still rather weak, and the signal-to-noise ratio in the obtained extinction spectra is generally relatively low. The overall sensing performance is thus usually better when an ensemble of nanoparticles is used for detection since this gives a higher signal intensity and an improved signal-to-noise ratio.

Substrate Effect

Nanoplasmonic sensing devices where the metal nanoparticles are immobilized on a solid support have several practical advantages as compared to using dispersed nanoparticles since immobilization facilitates handling, washing, storage, and regeneration of the sensor surface. Immobilization of nanoparticles can be achieved by self-assembly, using, for instance, thiol chemistry, or by physisorption using, e.g., electrostatic interactions between charged species adsorbed on the nanoparticle surface and the substrate. The self-assembly process can be controlled by an appropriate and defined surface chemistry at both the substrate and the metal nanoparticles and tuned by changing, e.g., the pH, ionic strength, nanoparticle concentration, or incubation time. However, tethering nanoparticles to a solid support will influence their RI sensitivity since the substrate will occupy a certain fraction of the sensing volume around the nanoparticles. This is known as the substrate effect and results in a reduction in the RI sensitivity. The magnitude of this reduction depends strongly on the fractional surface area of the nanoparticle that is in direct contact with the underlying substrate [11]. Thus, a flat structure like a nanodisk or nanoprism will be affected to a larger extent by the substrate effect than, for example, a nanosphere. Consequently, in order to reduce the influence of a substrate immobilization on the RI sensitivity, the contact area between the nanoparticle and the substrate should be minimized.

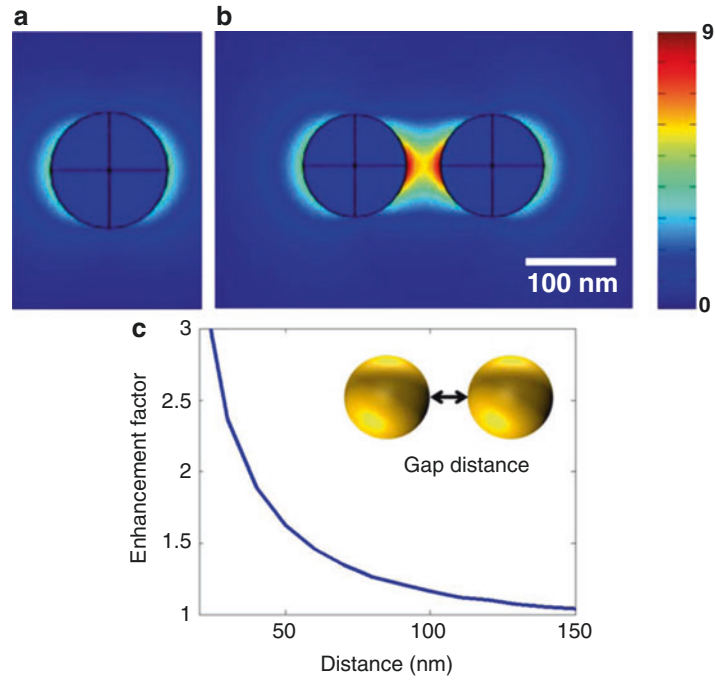
In addition, if the substrate has a higher RI value than the surrounding medium, the electric field created around the nanoparticles upon excitation can get redistributed, shifting a large part of the field toward the substrate. This will increase the substrate effect additionally since the RI detection is associated with the electric field. The redistribution of the electric field can, however, be reduced when using substrates with a low RI value, like Teflon ($n \approx 1.32$).

Plasmonic Coupling

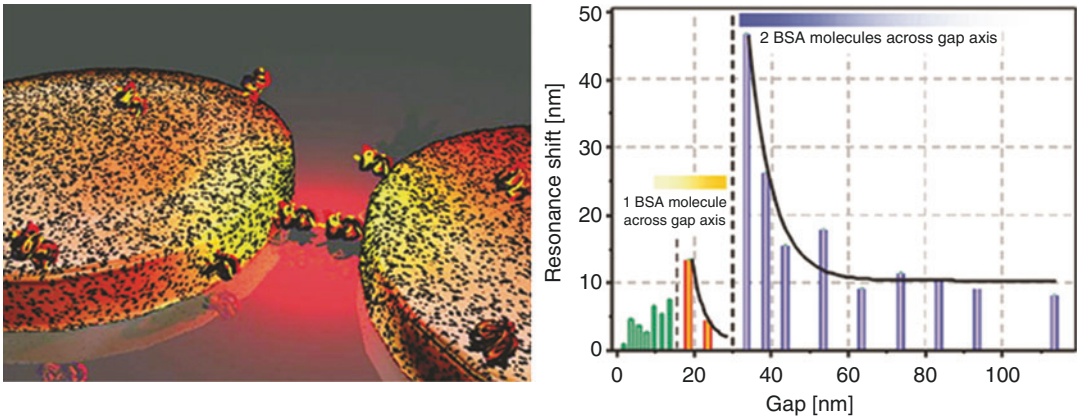
Another way to tune the plasmonic response of metal nanoparticles is to utilize dipolar interactions between closely spaced nanoparticles. This is known as plasmonic coupling and results in appearance of hybridized plasmon modes where the plasmon frequency is highly dependent on the separation between the nanoparticles. Plasmonic coupling between closely spaced nanoparticles gives rise to regions where the electromagnetic field is strongly enhanced as compared to the incident electric field that is used for the excitation (Fig. 5) [12, 13]. These so-called hot spots are of large interest for sensing applications and have been extensively utilized in surface-enhanced Raman spectroscopy (SERS) where the signals from Raman-active molecules can be significantly enhanced. In addition, these regions have also been shown to exhibit a higher RI sensitivity, and an improved sensing performance can be achieved if the molecular interactions can be guided into these areas [14]. Plasmonic coupling is also utilized in colorimetric sensing schemes based on analyte recognition-induced changes in colloidal stability. This method was pioneered by Mirkin et al. who used DNA to facilitate an aggregation of gold nanoparticles [15].

Nanoplasmonic Biosensing

Numerous examples of LSPR-based biosensing schemes have been proposed for detection of a wide variety of analytes including DNA, proteins, and small organic molecules using different recognition elements. Due to the limited sensing volume around metal nanoparticles, LSPR-based biosensing is a suitable method for detection of small analytes. The smaller the analytes are, the smaller the RI change will be upon binding to the ligand, which consequently requires a high sensitivity in order to obtain an acceptable limit of detection. For example, Kabashin et al. showed that small molecular weight compounds can be detected by localized surface plasmons created in gold nanopillars, and they were able to detect biotin (244 Da) instead of streptavidin (52.8 kDa), which is normally detected [16]. In addition, Hall et al. showed that LSPR can be used to monitor



Refractometric Sensing Using Plasmonic Nanoparticles, Fig. 5 Electric field distribution plots for (**a**) a single gold nanoparticle and (**b**) a nanoparticle dimer. (**c**) Field enhancement factor for a nanoparticle dimer at different gap distances with respect to a single nanoparticle (Reproduced with permission from Ref. [12])



Refractometric Sensing Using Plasmonic Nanoparticles, Fig. 6 Binding of BSA in the gap of a gold nanodimer. By varying the gap distance, the binding of one or two BSA molecules could be identified and discriminated (Reprinted with permission from Ref. [18]. Copyright 2009 American Chemical Society)

protein conformational changes using an array of silver nanoprisms fabricated by nanosphere lithography [17]. Aćimović et al. were able to discriminate between the bindings of one or two BSA (bovine serum albumin) molecules by varying the gap distance between gold nanodisks (Fig. 6) [18]. Recently, a LSPR biochip based on gold nanorods was realized that enabled a multiplexed detection of cancer biomarkers (human alpha-fetoprotein and prostate-specific antigen) down to concentrations of 500 pg/mL [19]. These examples demonstrate the great potential of LSPR in the development of high-performance sensing platforms that can be produced at a low cost.

Future Perspective

Refractometric sensing using plasmonic nanoparticles is a promising strategy for detecting small, local changes in the surrounding dielectric environment. This real-time, label-free sensing technique enables miniaturization and multiplexed analysis, which is difficult to achieve using conventional SPR-based sensing. However, many challenges remain before cost-effective refractometric sensing platforms based on LSPR that exhibit high sensitivity, specificity, and robustness can be practically implemented and can compete with existing sensing technologies. Some of these challenges are discussed below:

Biofunctionalization: The highly localized sensing volume around metal nanoparticles is advantageous in many aspects but also put constraints on the surface chemistry that can be used since the distance between the recognition moiety and the nanoparticle surface must be kept to a minimum. A reliable and biofunctional surface chemistry, which both exhibit selectivity and low unspecific binding, is critical in the development of high-performing sensing devices. In order to fully exploit the advantages of LSPR-based sensing devices, the surface coating must hence be very thin while still displaying low nonspecific adsorption and selective recognition of target analytes.

Reproducibility: Although the techniques for synthesis of metal nanoparticles has advanced significantly during the last decade, there is still room for improvements with respect to batch-to-batch reproducibility and minimizing size and shape polydispersity that give rise to broad plasmon bands and lower the overall sensing performance. Lithographic techniques can be used to make nanostructures and patterns with high reproducibility, but generally at a higher production cost and facing problems with up-scaling as compared to colloidal self-assembly. Self-assembly of metal nanoparticles thus offers a simple and cost-effective way to produce nanoplasmonic substrates, but the assembly conditions needs to

be tightly controlled in order to produce nanopatterns with high reproducibility and to facilitate multiplexed sensing.

Localization: Plasmonic coupling between adjacent nanoparticles gives rise to areas with enhanced electromagnetic fields. If the molecular sensing can be guided into these regions, a significantly higher sensitivity can be obtained. Thus, it would likely be more beneficial to be able to control the orientation and interparticle coupling between closely spaced nanoparticles rather than optimizing the structure of individual nanoparticles. However, it is a great challenge to both obtain well-defined plasmon coupling and to guide the immobilization of recognition elements into these hot-spot regions to exploit the enhanced sensitivity.

Cross-References

- Active Plasmonic Devices
- Biosensors
- Gold Nanorods
- Light Localization for Nano-optical Devices
- Nanoparticles
- Optical Properties of Metal Nanoparticles
- Plasmon Resonance Energy Transfer Nanospectroscopy
- Plasmonic Structures for Solar Energy Harvesting
- Surface Plasmon-Polariton-Based Detectors
- Synthesis of Gold Nanoparticles

References

1. Faraday, M.: The bakerian lecture: experimental relations of gold (and other metals) to light. Philos. T. Roy. Soc. **147**, 145–181 (1857)
2. Liedberg, B., Nylander, C., Lundstrom, I.: Surface-plasmon resonance for gas-detection and biosensing. Sensor Actuator. **4**, 299–304 (1983)
3. Englebienne, P.: Use of colloidal gold surface plasmon resonance peak shift to infer affinity constants from the interactions between protein antigens and antibodies specific for single or multiple epitopes. Analyst **123**, 1599–1603 (1998)
4. Mie, G.: Beiträge zur Optik trüber Medien, speziell kolloidaler Metallösungen. Ann. Phys. **3**, 377–445 (1908)
5. Turkevich, J., Stevenson, P.C., Hillier, J.: A study of the nucleation and growth processes in the synthesis of colloidal gold. Discuss. Faraday Soc. **11**, 55–75 (1951)
6. Charles, D.E., Aherne, D., Gara, M., Ledwith, D.M., Gun'ko, Y.K., Kelly, J.M., Blau, W.J., Brennan-Fournet, M.E.: Versatile solution phase triangular silver nanoplates for highly sensitive plasmon resonance sensing. ACS Nano **4**, 55–64 (2010)
7. Miller, M.M., Lazarides, A.A.: Sensitivity of metal nanoparticle surface plasmon resonance to the dielectric environment. J. Phys. Chem. B **109**, 21556–21565 (2005)
8. Otte, M.A., Sepulveda, B., Ni, W.H., Juste, J.P., Liz-Marzan, L.M., Lechuga, L.M.: Identification of the optimal spectral region for plasmonic and nanoplasmonic sensing. ACS Nano **4**, 349–357 (2010)
9. Willets, K.A., Van Duyne, R.P.: Localized surface plasmon resonance spectroscopy and sensing. Annu. Rev. Phys. Chem. **58**, 267–297 (2007)
10. Svedendahl, M., Chen, S., Dmitriev, A., Kall, M.: Refractometric sensing using propagating versus localized surface plasmons: a direct comparison. Nano Lett. **9**, 4428–4433 (2009)
11. Martinsson, E., Otte, M.A., Shahjamali, M.M., Sepulveda, B., Aili, D.: Substrate effect on the refractive index sensitivity of silver nanoparticles. J. Phys. Chem. C **118**, 24680–24687 (2014)
12. Chung, T., Lee, S.Y., Song, E.Y., Chun, H., Lee, B.: Plasmonic nanostructures for nano-scale bio-sensing. Sensors **11**, 10907–10929 (2011)
13. Jain, P.K., El-Sayed, M.A.: Plasmonic coupling in noble metal nanostructures. Chem Phys Lett **487**, 153–164 (2010)
14. Feuz, L., Jonsson, P., Jonsson, M.P., Hook, F.: Improving the limit of detection of nanoscale sensors by directed binding to high-sensitivity areas. ACS Nano **4**, 2167–2177 (2010)
15. Mirkin, C.A., Letsinger, R.L., Mucic, R.C., Storhoff, J.J.: A DNA-based method for rationally assembling nanoparticles into macroscopic materials. Nature **382**, 607–609 (1996)
16. Kabashin, A.V., Evans, P., Pastkovsky, S., Hendren, W., Wurtz, G.A., Atkinson, R., Pollard, R., Podolskiy, V.A., Zayats, A.V.: Plasmonic nanorod metamaterials for biosensing. Nat. Mater. **8**, 867–871 (2009)
17. Hall, W.P., Anker, J.N., Lin, Y., Modica, J., Mrksich, M., Van Duyne, R.P.: A calcium-modulated plasmonic switch. J. Am. Chem. Soc. **130**, 5836–5837 (2008)
18. Acimovic, S.S., Kreuzer, M.P., Gonzalez, M.U., Quidant, R.: Plasmon near-field coupling in metal dimers as a step toward single-molecule sensing. ACS Nano **3**, 1231–1237 (2009)
19. Acimovic, S.S., Ortega, M.A., Sanz, V., Berthelot, J., Garcia-Cordero, J.L., Renger, J., Maerkl, S.J., Kreuzer, M.P., Quidant, R.: LSPR chip for parallel, rapid, and sensitive detection of cancer markers in serum. Nano Lett. **14**, 2636–2641 (2014)

Relaxation

► Nanomechanical Properties of Nanostructures

Relaxation Calorimetry

► Nanocalorimetry

Reliability of Nanostructures

Tuncay Alan
Mechanical and Aerospace Engineering Department, Monash University, Victoria, Australia

Synonyms

Mechanical characterization of nanostructures; Nanomechanics

Definition

The continuous reduction in the size of structural device components promises improved efficiency and, often, new functionalities, which would otherwise be impossible with conventional macroscale devices. As a result, in the past decades numerous microsystem devices with nanoscale mechanical parts have been demonstrated and used in a variety of engineering applications, ranging from electrical circuit components to sensors, actuators, and biomedical applications [1]. This ongoing drive to create smaller new devices to perform critical functions requires significant attention to the mechanical reliability of materials used at nanoscales. To make use of the full mechanical potential of materials at decreasing sizes, fundamental properties, such as the elastic modulus, yield, fracture and fatigue properties, adhesion and intrinsic stresses, should be determined by testing relevant sized samples that have been fabricated or processed using the same procedures as the structure of interest and subjected to the same environmental conditions.

Overview

Even though bulk elastic properties are generally expected to remain constant down to a length scale of several nanometers, the mechanical reliability of materials depends significantly on the specimen size [2]. Despite their high theoretical strengths, most materials used in practical applications cannot sustain very high stresses due to intrinsic defects. An estimate for the minimum defect size that will cause failure can be obtained by an order of magnitude calculation considering Griffith fracture criterion (which suggests that an infinitesimal crack will propagate when the available strain energy is larger than that needed to create new surfaces) [3]. Accordingly, the plane-stress fracture strength of the material is given by

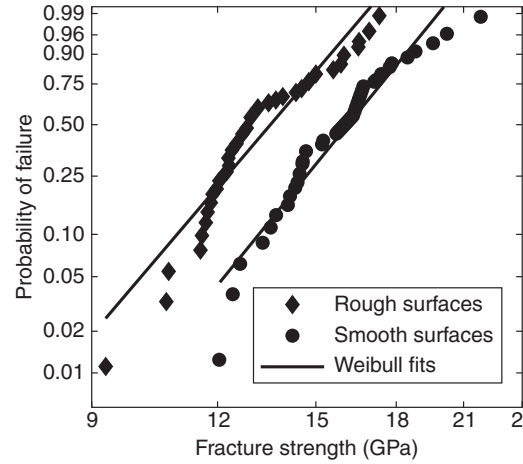
$$\sigma_f = \frac{1}{1.12}\sqrt{\frac{2E\gamma}{\pi l_c}}, \qquad (1)$$

where lc is the characteristic size of an ideal defect, γ is the surface energy per unit area, and E is the elastic modulus. Considering the theoretical strength of Si (calculated by atomistic simulations), $\sigma f = 23$ GPa [4], $\gamma = 1{:}25$ J/m^2 [5] and E = 130 GPa, Eq. 1 suggests that *a single 1-nm-long edge crack will reduce the strength of an otherwise defect-free Si component by 60 % to 9 GPa!*

Key Research Findings

As the dimensions of a material decrease down to nanometers, approaching the typical size of critical defects, the strength is expected to increase significantly with a potential to reach the theoretical limit. Recent experiments demonstrated that near-theoretical strength values that are more than 25 times higher than typical macroscale strengths could indeed be achieved for 190-nm-thick single crystal silicon structures [6, 7] and 15-nm-thick

low-pressure chemical vapor deposited low-stress SiN [8]. These ultrahigh strengths are very promising as they enable novel MEMS devices, which can reliably operate under very high stresses [8]. Yet, at decreasing size scales, the number of intrinsic volume defects decreases and surface-to-volume ratio increases. Hence, surfaces and process parameters gain importance. The above experiments on silicon [6] as well as stochastic simulations [7] show that atomic scale surface features, which would otherwise be negligible in larger structures start acting as defects (Eq. 1) and that they may significantly compromise the reliability of the nano-material. For instance, a slight increase in process induced root-mean-square (rms) surface roughness (from 0.4 to 1.5 nm) decreases the strength of Si nanobeams by approximately 20 % (Fig. 1). A similar effect is observed when the samples are exposed to air for an extended period. When surfaces are not effectively protected against oxidation, 1-nm-thick native oxide islands that form randomly on the surface have been shown to reduce the strength of structures by 30 % [9].



Reliability of Nanostructures, Fig. 1 Fracture strength statistics of nanobeams with different surface roughnesses. The estimated strength of the beams increases from 14.2 to 16.8 GPa as the surface roughness decreases from 1.5 to 0.4 nm

The above examples emphasize that at decreasing size scales, mechanical performance of structures is often dominated by a competition between size, grain structure, surface defects, and environmental conditions. How would failure statistics change as the average defect size approaches the size of the device? At what size scales would the environmental factors start dominating failure? Could the mechanical service life of structures be improved by modifying process parameters? Answering these three questions is crucial to effectively characterize mechanical reliability of any material used at the nanoscale.

Methodology

So far, various nanomechanical test procedures have been proposed, and a few commercial devices have appeared in the market as ultimate nanoscale test platforms. Initial studies have focused on commonly used MEMS materials, and more recently, nanomechanical experiments were also adapted to characterize biological nanostructures such as cells and proteins. Due to differences in experimental techniques, geometric properties of the samples used and fabrication procedures, there is a significant variation in the failure statistics reported by different groups.

Because of the number of parameters affecting mechanical response, there are still no standard techniques and it is difficult to come up with a universal method. The available experiments generally involve variations to well-established macroscale tensile, bulge, and bending tests. The choice of procedure depends on several factors among which, the most important are: size (a micron and nano-sized sample cannot always be tested by the same instrument), required statistical accuracy (due to their complexity some methods do not allow easy repetition), tested material (not all materials are compatible with the fabrication procedures necessary for the preparation of the samples).

Tensile Tests

The main advantage of a tensile test is its analytical simplicity. Since the gauge length of the specimen is under uniform stress and strain fields, load–displacement curves can easily be used to determine elastic modulus and failure properties

directly. Required displacement and load resolutions can generally be accommodated by commercially available piezoelectric or capacitive actuators and high-precision load cells [2]. Yet, handling tensile test specimens remains a major challenge: It is very difficult to grip an individual micron-sized sample from both ends to apply a uniform load. Sharpe and collaborators got around this problem by electrostatic gripping. They used an external probe that was temporarily attached to hundreds of micrometer long test samples by electrostatic attraction forces. The probe, which was connected to a load cell was then manipulated to load the samples to failure, while displacements were monitored via image processing. There have been many variations of this procedure as reviewed in detail elsewhere [2]. Using a different approach, Gaspar and coworkers [10] developed a homebuilt setup which uses a commercial load sensor and an automated stage to test micron-scale tensile samples that were supported by MEMS springs. The biggest advantage of the method, despite the relatively large length scales, is that it enables a very large number of tests to be performed at a short time.

The experiments that focused on smaller devices have generally employed specifically designed MEMS-based test setups, where the sample is integrated with the test device. For instance, Haque and coworkers [11] developed a tensile tester that allows in situ analysis of deformation inside an electron microscope while the sample is being loaded by MEMS beams. Similarly, some recent commercial devices by Hysitron [12] and others have promised similar functionalities. A comprehensive review of the test methods can be found in [2, 13, 14].

Bulge Tests

As another approach, bulge tests can effectively be used to test membranes with thicknesses approaching several nanometers while keeping the lateral dimensions in the order of hundreds of micrometers. This well-established method consists of measuring the out-of-plane deflection of a thin membrane under the influence of increasing pressure, applied from its backside [14]. The central deflection of the membrane is monitored continuously and the maximum stress at failure can be inferred from the corresponding pressure–deflection curves. Shrinking only the thickness (while keeping the other dimensions at μm/mm scales relaxes the experimental constraints on detection considerably. Recently, a homebuilt bulge test setup was successfully used to statistically characterize membranes that are as thin as 10 nm [8], while displacements were measured optically. Bulge tests also provide a very good alternative to study 2D materials such as graphene. However, the major disadvantage of the technique is the complicated analysis that is required to interpret the experimental data accurately.

Nanoscale Bending Tests with AFM

The third approach, which is reviewed in more detail here, is the nanoscale bending test. Bending tests offer many benefits over the above methods. Most importantly, gripping of the test samples is no longer a problem and the experimental procedures are generally simpler than in tensile tests. Since the test samples and test setup are independent from each other, there are fewer process restrictions and the effects of different parameters on reliability can be more easily studied. Finally, in a bending test, the area of fracture initiation can be effectively controlled without needing external notches. For an un-notched tensile test sample, fracture load can be reached anywhere along the beam: Sidewall surfaces and beam top or undersides have equal stresses and hence the defect distribution on all sides should be carefully considered. On the other hand, bending samples can be designed so that maximum stresses occur on the critical sites, hence, simplifying the corresponding statistical analysis. This point proves to be very useful when studying the effects of specific defects on mechanical performance [7]. Yet, the major drawback of the method is that the stress–strain distributions within the sample and its failure characteristics must be acquired indirectly, by processing the experimental load–displacement measurements.

Overview of the Method

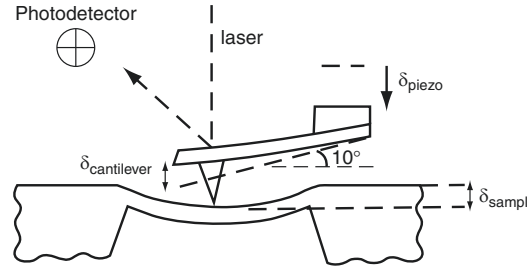
Among all of the available methods, atomic force microscopy (AFM) is arguably the most popular

one for nanoscale bending tests, being extensively employed to characterize both mechanical device components and biological samples. AFM is a powerful analytical tool used mainly to determine the surface topography of structures with the help of a (usually) compliant cantilever, which is attached to a piezo scanner. For instance, in contact mode operation, the cantilever scans the surface of the material while it is in contact with the sample. A feedback control loop moves the piezo scanner vertically at each scan point to maintain a fixed force between the tip and the sample. This vertical piezo displacement can then be transformed into a topographic image of the surface. A nanomechanical AFM-based test consists of using an AFM cantilever, with known stiffness, to deflect a test sample until it fails. The applied force, F, and the sample deflection, δ_{sample}, are simultaneously measured during loading and the fracture strength of the tested structure is obtained from either a simple theoretical calculation or finite element analysis (FEA).

The first step in the experiment is to determine the position of the AFM cantilever tip relative to the test sample accurately. For this purpose, the sample of interest is imaged by a contact mode scan and the load application point is picked from the AFM image. Using the piezoelectric controls of the instruments, the cantilever tip is offset to the loading point with an accuracy of a few nanometers. Next, the sample is loaded. The schematic in Fig. 2 shows a cantilever loading a micron-sized test beam. When the piezo scanner is extended by δ_{piezo}, the AFM cantilever pushes against the sample, which is deflected by δ_{sample}, while the cantilever undergoes an upward displacement of $\delta_{\text{cantilever}}$. Since the total piezo extension, δ_{piezo} equals the sum of the cantilever and sample displacements, deflection of the test structure can be expressed as

$$\delta_{\text{sample}} = \delta_{\text{piezo}} - \delta_{\text{cantilever}} \tag{2}$$

Deflection Measurement To obtain load–deflection curves for individual samples, $\delta_{\text{cantilever}}$ is continuously recorded throughout the experiment by monitoring the position of the laser light, which is reflected off the back of the AFM cantilever and collected by a position-sensitive photodetector. The upward cantilever deflection results in a change in the photodetector output signal (S_{p}). Hence, every time a cantilever is attached to the piezo scanner, the relation between $\delta_{\text{cantilever}}$ and the S_{p} (in Volts) must be carefully calibrated. To relate S_{p} (V) to $\delta_{\text{cantilever}}$ (m), the cantilever is brought into contact with a noncompliant surface, such as a rigid Si piece. The piezo is extended by a known distance, δ_{piezo}, pushing the cantilever against the surface and the corresponding change in S_{p} is monitored. Throughout the calibration, since the surface remains rigid (neglecting elastic contact deformations), $\delta_{\text{cantilever}} = \delta_{\text{piezo}}$. Considering the nonlinearity often observed at higher deflections, this relation is best characterized by a third-order polynomial ($\delta_{\text{cantilever}} = c_1 S_{\text{p}}^{\,3} + c_2 S_{\text{p}}^{\,2} + c_3 S_{\text{p}}$), the constants of which are determined by a least-squares fit.



Reliability of Nanostructures, Fig. 2 Schematic of a nanoscale bending test with an AFM cantilever loading a double clamped test beam

Force Measurement The force, F, applied by the cantilever on the test structure is equal in magnitude and opposite in direction to the load that causes the cantilever to deflect by $\delta_{\text{cantilever}}$. As seen in Fig. 2, the AFM cantilever makes a 10° angle with the horizontal. Hence, any load, F, applied on the test structure will have a vertical component, $Fv = F\cos(10^\circ)$ and a much smaller, negligible, horizontal component, $Fh = F\sin(10^\circ)$. The vertical force applied on the beam is given by

$$F_v = k_{\text{cantilever}}\delta_{\text{cantilever}}, \tag{3}$$

where $k_{\text{cantilever}}$ is the stiffness of the AFM cantilever.

Accuracy of the force measurements strongly depends on $k_{\text{cantilever}}$. Using commercially available cantilevers may save lengthy fabrication steps; however, since the specifications are not precisely known, a careful calibration process is needed to accurately determine $k_{\text{cantilever}}$.

AFM cantilever stiffness calibration. When uncoated single-crystal Si cantilevers are used, the stiffness can be obtained from the cantilever dimensions and known elastic properties of Si. An SEM can be used to measure the lateral dimensions and the offset of the tip, which defines the effective length of the cantilever. Similarly, the tip height can be obtained using an optical surface profilometer. Measuring the cantilever thickness accurately is often the most important, yet, most difficult step, due to the inclined sidewalls resulting from the fabrication process.

An iterative method based on resonant measurements was recently proposed to accurately determine h and calculate $k_{\text{cantilever}}$ [6]. The resonant frequency of the cantilever in air, f_{air}, is measured with the AFM and the vacuum frequency, f_{vac}, is estimated by correcting for the effects of the air damping.

$$f_{\text{vac}} = \left[1 + \frac{\pi b \rho_{\text{air}}}{4h\rho_{\text{cantilever}}}\right]^{0.5} f_{\text{air}}, \qquad (4)$$

where ρ_{air} and $\rho_{\text{cantilever}}$ are the densities of air and the cantilever material respectively, b is the average width, and h is the thickness of the cantilever. (For typical cantilevers used in the experiments in [6], $f_{\text{vac}} \sim 1.002\, f_{\text{air}}$.) The vibration of the cantilever is then simulated by FEA while the thickness, h, is iterated until simulated and measured frequencies match. The stiffness, $k_{\text{cantilever}}$, is then calculated from a static FEA, which incorporates known elastic properties and the exact cantilever geometry.

Determining the Fracture Strength Finally, the experimental load–deflection curves are processed to obtain the Young's modulus, E, of the samples and the deformation of each sample is modeled with a nonlinear FEA simulation. The FEA model takes into account the maximum load at fracture, load application point, geometric properties, and calculated E values to determine the fracture stress, *maximum tensile stress along the length of the sample*, for each sample. Results from repeat tests can then be analyzed statistically.

Accuracy of the Measurements

The precision of the strength measurements strongly depends on the accuracy of the force measurement, since the fracture load is used to extract fracture strength from FEA. An approximate expression for the error in strength measurements is given by:

$$\frac{\Delta\sigma}{\sigma}; \frac{\Delta F_v}{F_v}; \frac{\Delta k_{\text{cantilever}}}{k_{\text{cantilever}}} + \frac{\Delta_{\text{cantilever}}}{\delta_{\text{cantilever}}} \qquad (5)$$

For this analysis, the cantilever stiffness can be expressed as $k \approx 3EI/L^3$, where $I = bh^3/12$, with a cantilever width, b, thickness, h, and length, L. Similarly, the thickness $h \approx cf_{\text{vac}}L^2$, where c is a constant and f_{vac} is the resonant frequency in vacuum. Hence, Eq. 5 becomes:

$$\frac{\Delta\sigma}{\sigma} = \frac{\Delta b}{b} + \frac{3\Delta f_{\text{vac}}}{f_{\text{vac}}} + \frac{6\Delta L}{L} - \frac{3\Delta L}{L} + \frac{\Delta\delta_{\text{cantilever}}}{\delta_{\text{cantilever}}} \qquad (6)$$

The relative errors in the width and length measurements are related to the resolution of the SEM images, while $\frac{\Delta f_{\text{vac}}}{f_{\text{vac}}}$ depends on the sensitivity of the frequency measurement. All of these are much smaller than 1 %. On the other hand, the precision of the AFM cantilever displacement is directly related to the precision of the AFM piezo scanner and the photodetector readout and will dominate the measurement error, depending on the setup being used.

Future Directions in Research

AFM has great potential to become the standard testing method for a large class of materials. No doubt, the biggest advantage of the method is the use of commonly available standard instruments,

hence allowing similar structures to be tested at different locations without needing specialized test apparatus. Being highly flexible, it can effectively characterize the effects of process defects and environmental degradation on mechanical reliability, hence addressing the questions postulated in the "Key Research Findings" section. Moreover, there are no material restrictions, and, upon correct choice of cantilevers, samples ranging from hundreds of nanometers to a few micrometers can repetitively be tested with very good-to-excellent accuracy. However, further studies are still needed to better understand the interaction between the tip and the sample during loading, determine the optimum tip and cantilever geometry to test different samples, and decide upon standard calibration procedures. One of the current challenges is also to develop new generation of AFM cantilevers in order to adapt the existing methodology to perform tensile tests at the same resolutions.

Cross-References

- Ab Initio DFT Simulations of Nanostructures
- AFM
- AFM Probes
- Atomic Force Microscopy
- Finite Element Methods for Computational Nano-optics
- Mechanical Properties of Hierarchical Protein Materials
- Nanomechanical Properties of Nanostructures

References

1. Senturia, S.D.: Microsystem Design. Springer, New York (2004)
2. Hemker, K.J., Sharpe Jr., W.N.: Microscale characterization of mechanical properties. Annu. Rev. Mater. Res. **37**, 93–126 (2007)
3. Anderson, T.J.: Fracture Mechanics: Fundamentals and Applications, 3rd edn. CRC Press, Boca Raton (2004)
4. Roundy, M., Cohen, M.L.: Ideal strength of diamond, Si, and Ge. Phys. Rev. B **64**, 212103 (2001)
5. McCarty, A., Chasiotis, I.: Description of brittle failure of non-uniform MEMS geometries. Thin Solid Films **515**(6), 3267–3276 (2006)
6. Alan, T., Hines, M.A., Zehnder, A.T.: Effect of surface morphology on the fracture strength of silicon nanobeams. Appl. Phys. Lett. **89**, 091901 (2006)
7. Alan, T., Zehnder, A.T.: A Monte-Carlo simulation of the effect of surface morphology on the fracture of nanobeams. Int. J. Fract. **148**, 2 (2007)
8. Alan, T., Yokosawa, T., Gaspar, J., Pandraud, G., Paul, O., Creemer, F., Sarro, P.M., Zandbergen, H. W.: Microfabricated channel with ultra-thin yet ultra-strong windows enables electron microscopy under 4-bar pressure. Appl. Phys. Lett. **100**, 081903 (2012)
9. Alan, T., Zehnder, A.T., Sengupta, D., Hines, M.A.: Methyl monolayers improve the fracture strength and durability of silicon nanobeams. Appl. Phys. Lett. **89**, 231905 (2006)
10. Gaspar, J., Schmidt, M.E., Held, J., Paul, O.: Wafer-scale microtensile testing of thin films. J. Microelectromech. Syst. **18**(5), 1062–1076 (2009)
11. Haque, M.A., Espinosa, H.D., Lee, H.J.: MEMS for in situ testing-handling, actuation, loading, and displacement measurements. MRS Bull. **35**(5), 375–381 (2010)
12. Kiener, D.A., Minor, A.M.: Source truncation and exhaustion: insights from quantitative insitu TEM tensile testing. Nano Lett. **11**(9), 3816–3820 (2011)
13. Haque, M.A., Saif, T.: Mechanical testing at the micro/nanoscale. In: Sharpe, W.N. (ed.) Springer Handbook of Experimental Solid Mechanics. Springer, New York (2008)
14. Tabata, O., Tsuchiya, T.: Reliability of MEMS: Testing of Materials and Devices. Wiley-VCH, Weinheim (2007)
15. Vlassak, J.J., Nix, W.D.: A new bulge test technique for the determination of Young's Modulus and Poisson's ratio of thin films. J. Mater. Res. **7**(12), 3242–3249 (1992)

Remotely Powered Propulsion of Helical Nanobelts

Gilgueng Hwang[1] and Stéphane Régnier[2]
[1]Laboratoire de Photonique et de Nanostructures (LPN-CNRS), Site Alcatel de Marcoussis, Paris, France
[2]Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie, Paris, France

Synonyms

Wireless propulsion of helical nanobelt swimmers

Definition

Remotely powered propulsion of helical nanobelts is the swimming propulsion of nanoswimmers created by helical nanobelts.

Overview

Since swimming mobile micro-agents at low Reynolds number experience severe viscous drag, their swimming performances are much limited. Considering the most of in vivo environments are dynamic system, highly dynamic microswimmers would have a big technological impact in many fields. Engineering artificial flagella is inspired from nature's bacteria having their excellent motility. They are created from self-scrolled helical nanobelts. These swimmers are designed with a head and a tail, similar to nature's micro-organisms such as bacteria and their flagella. In this entry, the highly dynamic propulsion of biologically inspired artificial flagella which are remotely powered by electroosmotic force is introduced. They can swim as fast as 24 times of their body lengths per second and apply pressure higher than 300 Pa from their high energy conversion efficiency. Their swimming performance outran the other inorganic microscale swimmers and even nature's bacteria like *Escherichia coli*. These electroosmotic propelled helical nanobelt swimmers could be the potential biomedical carriers, wireless manipulators, and as local probes for rheological measurements.

Introduction

Rapid advancement in semiconductor electronics and microelectromechanical systems (MEMS) has accelerated the size reduction of functional systems like sensors or actuators. Considering much increased needs in less invasive biomedical or clinical tools to prolong human life, micro- or nanoscale systems have many promising features. One of the most important features is to reduce their size and disconnect them from wires. In this context, recent interest toward developing untethered biomedical micro- and nanorobots could be an alternative way to achieve less invasive and targeted medical therapy. Although they seem to increase the complexity of conventional medications, they are still promising due to their controllability, local therapy, and repeatability. Recent developments of in vivo biomedical nanorobots draw much interest in discovering new and efficient wireless power transfer methods and in controlled locomotion mechanisms. This kind of novel tools would have many potential applications, such as targeting, diagnosing, and treating blood clots or cancer cells, for neural cell probing and so on.

Although they are promising to such applications, they are still remained in laboratory experiments due to several reasons. The micro- or nanoscale mobile agents experience much more viscous drag due to the decreased Reynolds number which results in severely limited swimming performance. Considering the capillary of human's circulatory system has the average flow speed which can go up to 1 mm/s, mobile microagents should swim faster than this flow speed for their survival and task continuity. It should be noted that nature's bacteria have well adapted in such a harsh environment.

Many approaches aim to mimic natural bacteria for the locomotion of artificial nanostructures because of their excellent motility, around 5–10 times of body sizes per second [1]. The Nature has adopted several different propulsion techniques. The swimming of real bacteria is mainly divided into corkscrew-type rotating propulsion and the oscillation of flagella tail [2]. Several studies exploring similar propulsion techniques exist in the literature, with different efficiencies. Purcell uses a definition of swimming efficiency which compares the power used to propel a body at a given velocity to the power required to simply pull the body through the fluid at the same velocity (ε = power required to simply pull body/power consumed during propulsion) [2]. Under this definition, direct pulling with field gradient is 100 % efficient (i.e., $\varepsilon = 1$), and other methods would always perform less. Recently, direct pulling of nanowires, nanotubes and microtubes with excellent swimming performances (2 mm/s^{-1} which is

approximately 50–100 times per body size) was demonstrated by catalytic decomposition of hydrogen peroxide (H_2O_2). However, this technique only works in a solution like H_2O_2, allowing a catalytic reaction with Pt or Ag [3].

Among different propulsion methods, the use of magnetic effect is the most widely encountered. A rotating magnetic field was used to drive the macroscale flagella-like structure [4]. Microbeads were attached to natural bacteria and moved by flagella propulsion [5], 30-μm-scale bacterial flagella–like propulsion was demonstrated by an attached ferromagnetic metal pad under an external rotating magnetic field [6]. A linear chain of colloidal magnetic particles linked by DNA and attached to a red blood cell can act as a flexible artificial flagellum [7]. Further size reduction and applying same principle to drive artificial bacteria was achieved [8]. Lately, helical propellers are shown as an efficient solution at low Reynolds numbers [9].

Concerning the driving power source, there are several limitations of using an external field. The magnetic field gradient rapidly decreases with the distance from the source. Therefore, although the helical morphology is advantageous at low Reynolds numbers by reducing the viscous drag, the swimming performance of artificial structures is still much lower than the natural bacteria. To improve the swimming performance, increasing field intensity would be necessary.

However, the major challenge toward potential in vivo biomedical applications is the lack of closed loop motion control either by teleoperation or autonomous navigation. This lack is related to position-tracking issues. Widely used noninvasive biomedical sensors or imaging devices include radiography, computed tomography (CT), ultrasound echography, and magnetic resonance imaging (MRI). Considering the real-time navigation and high spatial resolution requirements, MRI is one of the most promising solutions. Ferromagnetic objects, or even magnetotactic bacteria, were tracked and navigated under real-time MRI [10]. In case of smaller objects at low Reynolds numbers, an additional propulsion-dedicated field gradient generator would probably be required. However, magnetic field gradient–based propulsion would be largely limited when it has to be used with an observation device such as MRI because of the conflict between the imaging and propelling magnetic fields. This problem can be avoided by using nonmagnetic propulsion.

As nonmagnetic approaches, the flapping or undulating motion of flagella was achieved in mesoscale using a commercial piezoelectric bimorph [11]. Microrobots that harness natural bacteria have also been demonstrated [5]. There are very few works on electric field–based wireless micro/nanorobots. For example, diodes were actuated using electroosmotic pumping [12]. It showed that electroosmotic field can be a good candidate for actuating micro/nano robots. However, the work was only demonstrated in macro scale and not scaled down to micro or nano.

In this entry, remote propulsion of artificial micro swimmers with high dynamic mobility (5–30 body lengths per second) at low Reynolds numbers by electroosmotic effect is described. These swimmers are based on helical nanobelt (HNB) structures. As different types of HNBs showed different swimming performances, choices of surface material and geometry is discussed along with the fabrication processes. The electroosmotic propulsion principle and experimental setups are presented. Finally, the experimental results on electroosmotic propulsion of HNBs in comparison with different locomotion techniques are described.

Self-Scrolled Helical Nanobelts as Artificial Flagella

Advantages of HNBs as Artificial Flagella

There are several different types of bacterial locomotion, such as rotating and undulating motions. Their major features, especially in terms of surface condition, geometry, and motility have evolved to swim efficiently at low Reynolds numbers. For example, *E. coli* bacteria consist of a highly elastic body formed of protein. Therefore, they can self-transform their morphology to increase their swimming velocity in a viscous liquid environment. It is considered that such

ability of natural bacteria is mainly attributed to the nonlinear mechanics of their flagella. To mimic similar performances of natural bacteria, the ultraflexible nanostructures with large nonconstant mechanics are considered to be necessary. The mechanical durability is also an important issue especially in harsh environments.

HNBs are one of the most flexible nanostructures that can be fabricated using standard microfabrication techniques with controlled geometry. Helical morphology appears as an adapted solution for electroosmotic propulsion because of the large surface-to-volume ratio to maximize the external energy reception. Another advantage is their elasticity to passively self-adapt their morphology in thick liquid environments by reducing viscous drag. Different types of nanobelts were synthesized and mechanically characterized [13]. Recently, HNBs revealed very large nonlinear mechanical properties along with giant piezoelectricity [14]. In case of electroosmotic propulsion, different surface coatings with thin films can also be achieved by standard microfabrication techniques to improve electrokinetic energy. The high surface-to-volume ratio of HNBs can disadvantageously increase friction at the liquid/solid interface. HNBs with hydrophobic surface coatings can reduce this friction to improve the swimming performance.

Additionally, harvesting energy from the environment would be an important advantage for mobile micro/nanoagents. Piezoelectric nanogenerators based on zinc oxide nanowire arrays were firstly proposed [15]. Nanowire-based, self-powered nanosensor has also been demonstrated [16]. Passive motions of HNBs during their swimming at low Reynolds number can self-harvest electrokinetic energy due to their giant piezoelectricity [14].

Therefore, considering the limitations of conventional propulsion mechanisms, remotely powered gradient pulling of HNBs should be efficient to swim in low Reynolds number liquid [9].

Fabrication Process

Four different types of HNB-based nanoswimmers, varying the materials and geometries, summarized in Table 1 are produced for testing purposes. Their propulsion efficiency is then compared in electroosmotic propulsion experiments.

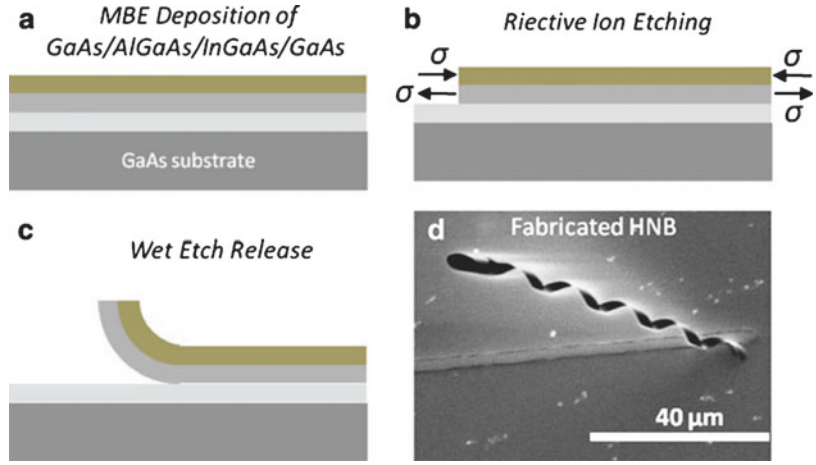
Figure 1 illustrates the fabrication process. The initial layers were grown on semiinsulating GaAs (AXT wafers) using a molecular beam epitaxy system (VEECO, Gen II MBE) equipped with a valved cracker for As and solid sources for Ga and In (Fig. 1a). For n-type doping, a Si source was used. Substrate temperature was measured by a pyrometer. After a thin GaAs buffer layer, the 500-nm thick sacrificial AlGaAs layer is deposited. The layer contains 20 % Ga in order to prevent oxidation. The sacrificial layer in previous designs was made of AlAs. The oxidation of these layers after exposure damaged the InGaAs/GaAs bilayer. On top of the sacrificial layer the InGaAs/GaAs layer is deposited which later self-forms into a nanospring. An In content of 15 % in the InGaAs layer was determined by X-ray diffraction (XRD) measurements. The thickness of this layer must be smaller than the critical thickness to maintain elastic strain. The layer properties along with other specifications of the structures are summarized in Table 1.

During the deposition of the InGaAs/GaAs bilayer on this wafer for HNB 1, it was attempted to get a slightly lower doping concentration than on the wafer that was used for HNB 2, 3, and 4. From the results, it seems that the doping concentration was too high which resulted in a partial self-compensation and, therefore, in a decrease in the effective doping concentration. The doping concentrations of the structures are 4.4×10^{18} and 8.5×10^{18} cm^{-3}. The initial pattern can be created through photolithography. Reversible photoresist AZ5214 was used as a resist. After the development of the resist, reactive ion etching (RIE) with a $SiCl_4$ gas was used to transfer the pattern to the InGaAs/GaAs bilayers. For HNB4, the thickness of photoresist was reduced around 50 nm with O_2 plasmas. For metallic HNB3, the Cr/Ni layers on the surface and on the heads to which the structures are fixed at the end were created before RIE through a lift-off process with negative photoresist AZ5214.

The Cr layer is 10 nm thick and serves as an adhesion layer. The 10-nm thick Ni layer is used for metal HNB3. Finally, a 2 % HF aqueous

Remotely Powered Propulsion of Helical Nanobelts, Table 1 Specifications of different types of artificial bacteria (HNB)

	HNB 1 low	HNB 2 high	HNB 3 Cr/Ni	HNB 4 photoresist
Thickness $In_{0.15}$ $Ga_{0.85}As$/GaAs (nm)	11.6/15.6	8~10/15.6	11.6/15.6	
Thickness Cr/Ni or photoresist (nm)	NA		Cr/Ni 10/10	Photoresist ~50
Diameter (tail = head), pitch and width (tail) (μm)	2.1, 14 and 2.5			
Number of turns (tail)	4.5 turns			
Length (head/total) (μm)	12/74			
InGaAs/GaAs doping (*ND*)	4.4×10^{18} cm^{-3}	8.5×10^{18} cm^{-3}		



Remotely Powered Propulsion of Helical Nanobelts, Fig. 1 Basic fabrication process sequence of an HNB. (**a**–**c**) Initial planar bilayer, patterned through conventional microfabrication techniques, folds itself into a 3-D nanostructure during wet etch release. (**d**) Produced HNB with a tail and a head (SEM photo)

solution at 4 °C was used to selectively etch the AlGaAs sacrificial layer under the InGaAs/GaAs heterostructures for the self-forming of the nanostructures (Fig. 1a). During this wet etch, the patterned bilayer curled up along a <100> direction, releasing the internal strain and forming 3-D structures. The direction of the scrolling is determined by the anisotropy in stiffness of the InGaAs/GaAs bilayer. After the wet etch release, the chips were rinsed in deionized water and subsequently in isopropyl alcohol [14]. Samples are also conserved in isopropyl alcohol in order to prevent oxidation by water of heads to which the structures are fixed.

Electromechanical Property

To understand the electrokinetic function of the applied strain by liquid environment, the piezoresistivity of HNBs is characterized. For the intrinsic property characterizations, the HNBs with metal pads attached on both sides were fabricated using microfabrication techniques [14]. With the metal connectors, good electrical contact can be achieved. Besides the electromechanical characterization, such connectors also allow for the integration of these structures into more complex assemblies. Nanomanipulation inside an SEM was used for their electromechanical property characterization. The experimental results showed that the structures showed unusually high piezoresistive response. Moreover, electrostatic actuation was used to excite the structures at their resonance frequency and investigate their resistance to fatigue. With their low stiffness, high strain capability, and good fatigue resistance, the HNBs can be used as high-resolution and large-range force sensors. By variation of design parameters, such as the number of turns, thickness, diameter, or pitch, an HNB with the required stiffness can be designed through simulation.

For the quantitative comparison with other piezoresistors, the piezoresistance coefficient can be obtained as,

$$\pi_l^{\sigma} = \frac{1}{X}\frac{\Delta\sigma}{\sigma_0} \quad (1)$$

where σ_0 is the conductivity under zero stress and X is the stress. The piezoresistance coefficient was defined with resistivity as $\pi_l^{\sigma}=\Delta\rho/\rho_0$. The conversion is $\pi_l^{\sigma} = -\pi_l^{\rho}$ for small σ. Uniaxial stresses were applied on HNB along their lengths. Since the linearity of each cycle from the measured data is almost constant, the widest measured region of the cycles was chosen to calculate the piezoresistance coefficient. Geometry information from Table 1 was used as the parameters.

These obtained piezoresistance coefficients of HNBs based on Eq. 1 are compared with other elements of the piezoresistors. Each reported piezoresistance coefficients of Bulk Si, boron-doped Si, and Si-NW, and CNT were summarized in Fig. 2 in comparison with HNB [14]. Especially, Si-NW was recently reported as showing giant piezoresistive effect. The revealed piezoresistance coefficients of the HNBs are even much higher than the one of Si-NW. These results are straightforward from the benefit of HNB's flexibility considering the Eq. 1 since the tiny input stress (X) can cause the resistance change ($\Delta\sigma/\sigma_0$). At least the torsional effect seems to affect much to the high piezoresistive response of the HNBs. As was discussed on the difficulty in exact modeling of the piezoresistive HNBs, mechanism behind such a high piezoresistive response was not understood yet. However, there is the fact that the HNBs are the promising nanostructures to be used as the elements of ultraflexible electromechanical devices such as force sensors and bioinspired systems.

Mechanical Property

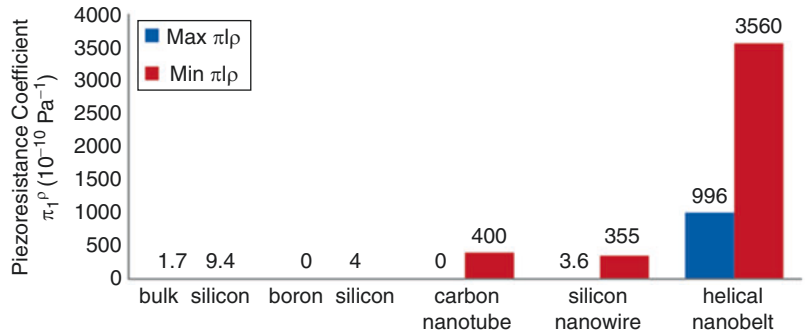
In addition to piezoresistivity of HNBs, they can perform like a spring to save energy from large-range mechanical force based on their nonconstant mechanics. Tuning fork force calibration tool was utilized to characterize the mechanics of HNBs in full-range tensile elongations. The tuning fork as force interaction tools have been proposed mainly to replace the AFM optical cantilevers for imaging [17]. The mechanical property (spring constant) of HNBs was measured by attaching them to the tuning fork tip. The force is obtained from the stiffness of HNBs measured by tuning fork gradient force sensor and by the displacement measurement of HNBs with SEM imaging. Figure 3 summarizes the measured mechanical property while HNBs are under tensile elongation. The nonconstant stiffness behavior of HNBs during their controlled tensile elongation was clearly revealed in their full-range elongation. The obtained stiffness ranges from 0.009 to 0.297 N/m with a resolution of 0.0031 N/m during full elongation and 0.011 to 0.378 N/m with a resolution of 0.0006 N/m for the nondestructive method. It was transformed into full elongation tensile forces as high as 2.95 μN for the first experiment and 1.56 μN for the second.

Wireless Propulsion of Helical Nanobelts

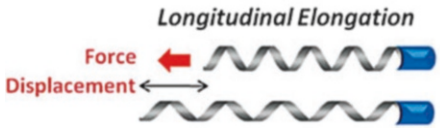
Principle of HNB Electroosmotic Propulsion

The propulsion of HNBs involves two conversions: transforming energy from an external source to a force and transforming this force into a motion. The proposed method to supply power remotely to HNBs is an electrokinetic process. Its main advantages for in vivo applications are the good tolerance of living organisms to electric fields and its compatibility with medical MRI imaging.

In general, the surface charge that develops at the solid–liquid interfaces plays an important role in a number of electrokinetic processes. Due to the presence of a thin interfacial layer named the Stern layer, it results in a nonzero electric potential at the liquid interface, also known as the zeta potential ζ. This potential is screened by mobile counterions diffusing in the liquid in a layer generally named Electrical Diffuse Layer (EDL), which has a typical thickness of a few tens of nanometers depending on the solution ionic strength. Under an electric potential difference, the resulting electric field puts in motion the excess counterions in the EDL. This moving layer drags the whole



Remotely Powered Propulsion of Helical Nanobelts, Fig. 2 Comparison of longitudinal piezoresistance coefficients in different nanostructures (bulk silicon, boron silicon, carbon nanotube, silicon nanowire, and helical nanobelt) [14]



Stiffness K (N/m)	Displacement (μm)	Force (μN)
~ 0.009	0	0
0.297 ~ 0.378	4.13 ~ 9.95	1.56 ~ 2.95

Remotely Powered Propulsion of Helical Nanobelts, Fig. 3 Mechanical property of helical nanobelt with nonconstant stiffness for large-range force sensing

liquid inside the fluidic medium, resulting in "pluglike" flows also referred to as electroosmotic flows (EOF). In addition, one should also consider the surface charge developing at the HNB surface. In the case of a HNB floating in liquid, its own Stern layer generates a flow which applies a hydrodynamic pressure on the surface of the robot, propelling it in the opposite direction.

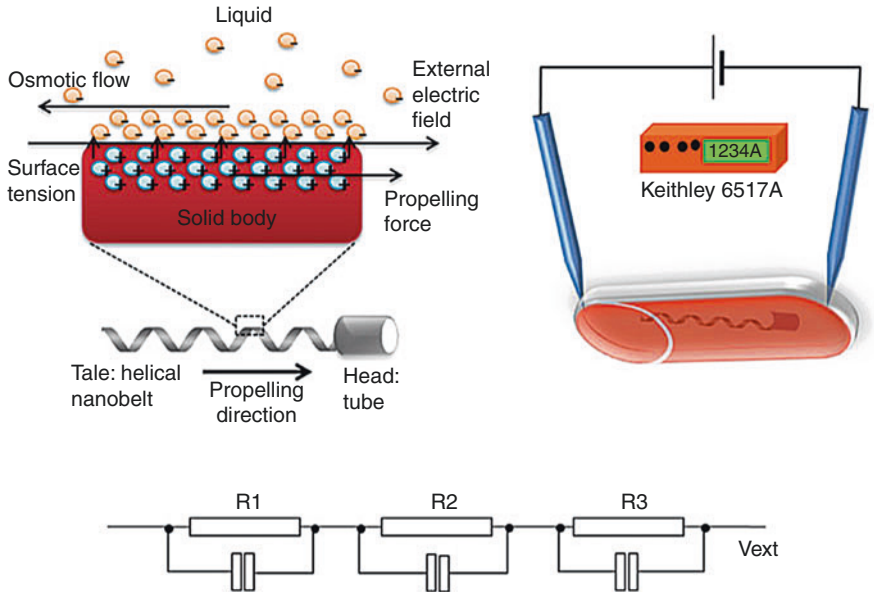
Figure 4 depicts the electroosmotic propulsion mechanism through the interface between HNB's thin membrane surface and liquid medium solution. Since the propulsion force occurs through the whole surface, the large surface-to-volume ratio of HNBs is very advantageous.

HNBs' motility results from a local electroosmotic flux powered by an external field. The specific direction of HNB's propulsion along the cathode and anode probes indicates that a DC field gradient along the HNB is responsible for this propulsion. The equivalent circuit can predict that the electroosmotic propulsion force is affected by the difference of electrical resistivity between medium solution and the HNB (Fig. 4). The electric voltage applied to the HNB by the external field can be estimated from a serially connected resistors model describing the ionic conductance through the liquid medium and the capacitors for the ionic layers. In a DC electric field, the resistance is the leading contribution. The resulting DC voltage of magnitude V_d induced in the HNB is modeled as

$$V_d = \frac{R_2}{R_1 + R_2 + R_3} V_{ext} \tag{2}$$

where V_{ext} is the DC input voltage applied to the probes in the petri dish, and R_1 (resistance between left electrode and HNB), R_2 (resistance of HNB), and R_3 (resistance between HNB and right electrode) are depicted in Fig. 4. Assuming that the resistance of the liquid is linearly proportional to the distance between the probes, Eq. 2 can be simplified as $V_d = E_{ext}\, l_d$, where l_d is the length of the HNB body and E_{ext} is the external DC field. The external DC field between the electrodes makes electroosmotic fluid flow along the HNB body.

Considering a liquid electrolyte, consisting of positive and negative particles in liquid solution, the interface between the electrolyte and the container wall generally forms a double layer in equilibrium, where a nonzero surface charge is screened by a very thin diffuse layer of excess ionic charge of width λ the Debye screening



Remotely Powered Propulsion of Helical Nanobelts, Fig. 4 Schematic diagrams of the experiments for HNB swimming. A DC bias was applied between two electrodes located through axial direction of HNB body. Thus, the propulsion direction can be controlled and predicted by the configuration of electrodes. For controlling the direction of electric field, micromanipulators with tungsten probes were utilized to form electrodes

length (typically 1–100 nm). The double layer is effectively a capacitor skin at the interface, which has a small voltage ζ across called the zeta potential. The effect on the HNB and also the environment under electric field is further detailed in the following model which decomposes the part of intrinsic electrophoretic force directly generated to HNB and the electroosmotic flow from the surface interaction. Considering a tangential electric field E_{ii} applied in parallel to a flat surface, the electric field acts on ions in the diffuse part of the double layer, which drags the fluid to produce an effective slip velocity outside the double layer by the Helmholtz-Smoluchowski formula (Eq. 3) [18],

$$u_{11} = \left(\frac{2\varepsilon_0 \varepsilon_r \zeta_{HNB} H}{3\eta} - \frac{\varepsilon_0 \varepsilon_r \zeta}{\eta} \right) \vec{E} \qquad (3)$$

where $\varepsilon_0 \varepsilon_r$ and η are the electrolyte permittivity and fluid viscosity, which are constants (environment parameters). Increasing the speed of the HNB requires increasing ζ_{HNB} and H which are the zeta potential around the surface of the HNB and the Henri function, which depends on the geometry and hydrodynamic properties and the external field (E). On the negative term of the equation, the Zeta potential ζ on the borders of the surface is inversely proportional to the resulting speed, thus it should be minimized and controlled by applying experiments under closed microfluidic channel.

An important application of electroosmotic flow is capillary electroosmosis, where an electric field is applied down a capillary tube to generate a uniform plug flow (Eq. 3), driven by the slip at the surface. Typical flow speeds of 100 μm/s are produced by a field of 10 V/mm. This value will be used to compare the swimming performance of HNBs. It should be noted here that the electroosmotic propulsion does not depend on the size of the microagents. Thus, microscopic HNB could move theoretically about as fast as a macrosized one.

Electroosmotic Propulsion Experimental Setup

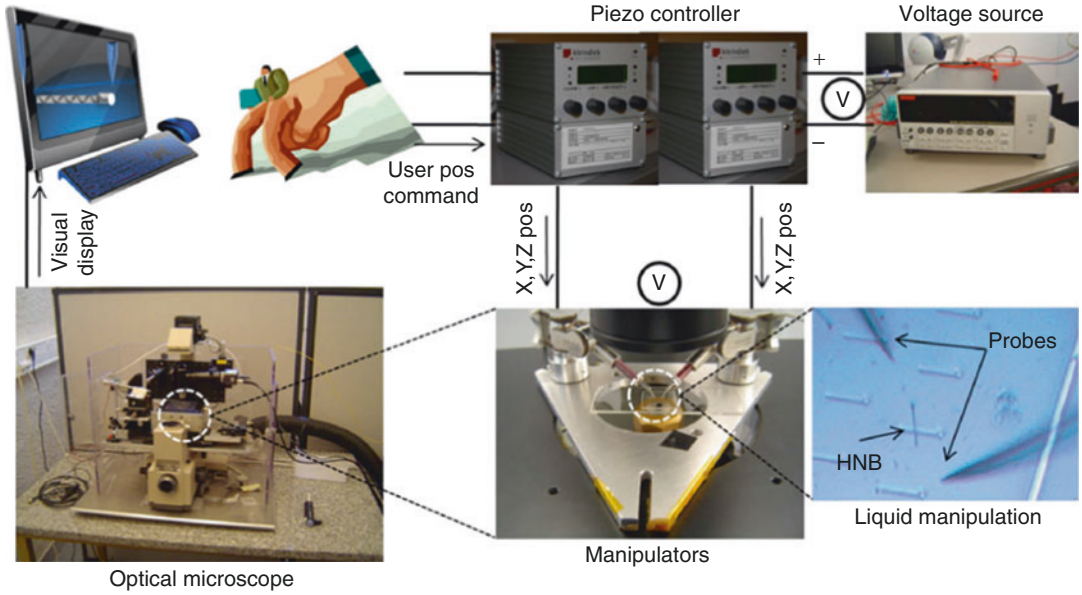
HNBs are detached from the substrate by micromanipulation. At the start of the experiments, the fixed end of produced HNBs are cut and released

from chips using a setup with two probes with high positioning accuracy under an optical microscope. Same probes are also used to generate the electric field through the liquid medium and HNBs. Total experimental setup is shown in Fig. 5. Two nanorobotic manipulators (Kleindiek, MM3A) are installed under optical microscope; each has three degrees of freedom and, respectively, 5, 3.5, and 0.25 nm resolutions in X, Y, and Z directions. For the application of the electrical field and high resistance measurement, an electrometer with a DC power supply (Keithley 6517A) is used. Two probes attached to each side of manipulators are positioned in less than 1 mm distance. To avoid the optical reflection from the meniscus between probes and liquid medium interface, longer probes (~5 mm) with sharp tips (diameter around 100 nm) are used.

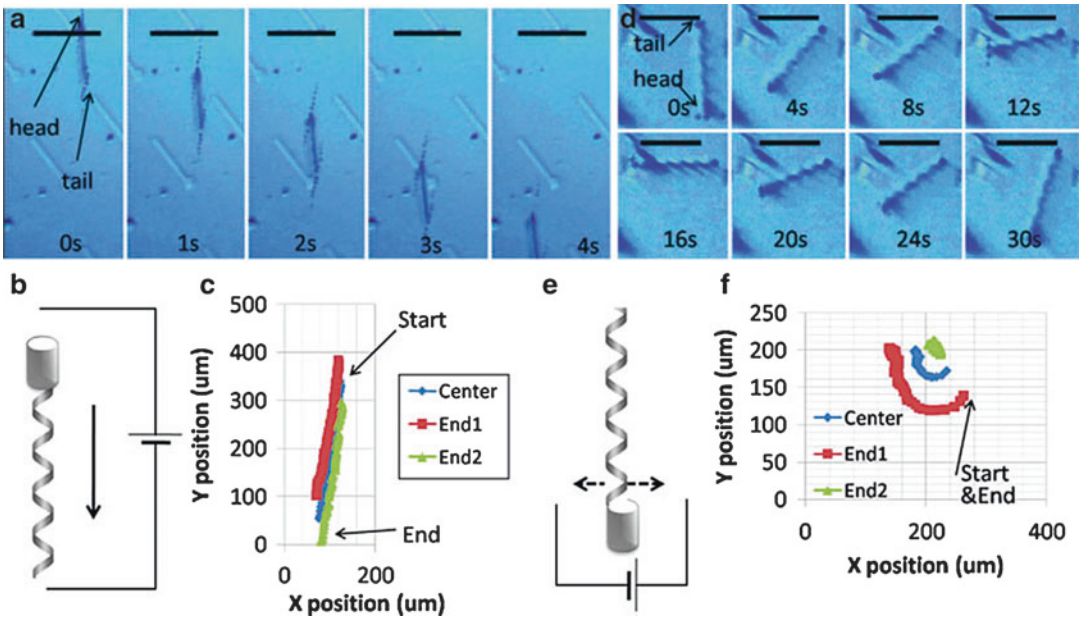
Controlled Swimming Experiments

Figure 6 demonstrates the forward and rotating propulsion of HNBs. Experiments are depicted in Fig. 6b, e. The swimming propulsions of HNBs are achieved by modifying the electrodes' configuration. A series of snap shots during forward (Fig. 6a) and rotating (Fig. 6d) propulsions were taken during the propulsion experiments. The trajectories of each swimming experiment are displayed in Fig. 6c, f. For trajectories, video analysis software and a custom-made C++ program are utilized. Rotation of HNBs was achieved by their aligning to the applied field direction. The probes attached to manipulators can change the configuration to generate the desired electric field gradient. Note that in this case the rotation was achieved using an HNB fixed in one extremity, while the other is free. The electrode configuration was fixed to make the HNB rotate. Furthermore, the forward and backward swimming direction can be converted without the complicated tumbling motion of *E. coli* bacteria.

It is confirmed that the swimming directions of HNBs could be easily controlled by differentiating the field gradient. For the steering control, the field gradient needs to be aligned to the rotational axes of the HNBs. In case of an untethered HNB, the rotation is achieved by controlling the field gradient. The high surface-to-volume ratio of



Remotely Powered Propulsion of Helical Nanobelts, Fig. 5 System setup for electroosmotic propulsion experiments: optical microscope with a camera is used to give user visual feedback and thus proper manipulation in liquid. User operates the piezo controller to manipulate tungsten probes, which are used to manipulate HNBs and to apply the electric field generated by an external voltage source



Remotely Powered Propulsion of Helical Nanobelts, Fig. 6 Different swimming propulsions of HNBs are achieved by modifying electrode configuration. Series of snap shots during backward (**a**) and rotating, (**d**) propulsions are taken during the experiments (**b** and **e**). The trajectories of each swimming experiment are displayed in (**c**) and (**f**). The scale bars in (**a**) and (**d**) are, respectively, 80 and 60 μm

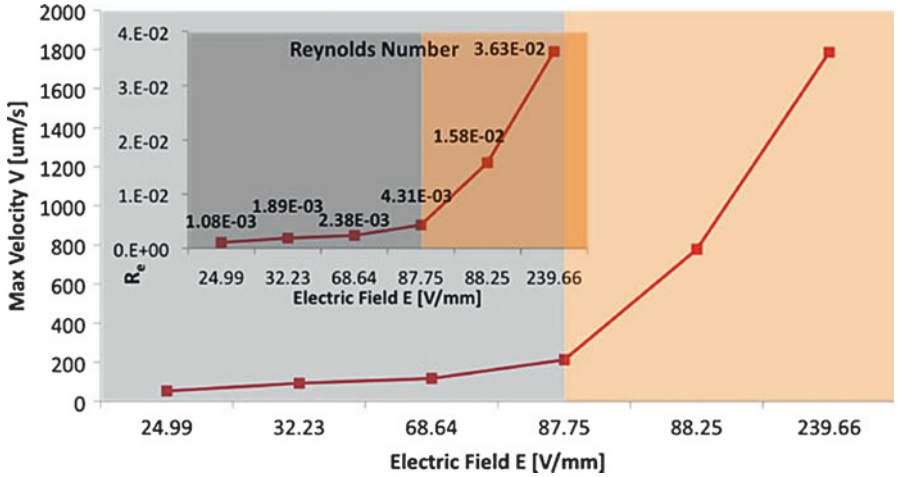
HNBs increases the swimming efficiency by increasing the charged area in a given volume and mass.

The maneuverability of the currently demonstrated electroosmotic HNBs is still limited to the basic motions compared to the magnetic field based approaches. However, this issue can be addressed using an array of embedded electrodes.

Comparison of Swimming Performance

Electrokinetic Effect

To further understand the repeatable swimming behaviors due to an electroosmotic force and to find the optimized design parameters, especially the surface profile, the swimming performances of four different types of HNB were characterized. High- and low-doping HNB, a metallic (Cr/Ni) one and a dielectric counterpart were analyzed. Figure 7 shows the average and maximum propulsion velocity and acceleration normalized by the applied electric field. HNB performance was investigated in several steps (6–11 steps) ranging from 24 to 401 V/m. It was difficult to compare four HNBs in identical single field intensity because the experimental configuration of probes and HNB is difficult to maintain between each experiment. Therefore, a statistical average velocity was chosen to describe the general behavior of different HNB. It should also be noted that experiments on different types of HNBs were conducted in different days which might slightly change electrical resistivity of the liquid medium. Even though the exact resistance values of the utilized HNBs were not verified because of the difficulty in collecting them after experiments, the intrinsic resistance of HNB can be predicted as $R_{HNB4} > R_{HNB1} > R_{HNB2} > R_{HNB3}$. Hence, the average and maximum velocities and accelerations depend on the resistance of HNB as was expected from the Eq. 3. Highly doped HNB2 showed better swimming performance than low-doped HNB1. This can be explained by the fact that the InGaAs layer thickness (estimated as 8–10 nm) of HNB2 is slightly thinner than the one of HNB1 (11 nm). It results in better performance in HNB2 than HNB1 from the reduced gravity



Remotely Powered Propulsion of Helical Nanobelts, Fig. 7 Dependence of HNB velocity on applied electric field intensity. The plot was obtained during the swimming experiments of a 74-μm-long HNB in isopropyl alcohol medium. The nonlinearity of the swimming performance is demonstrated at the region with higher electric field intensity than 87.75 V/mm. The *inset* figure shows the Reynolds numbers in different velocities versus applied electric field

effect and hydrodynamic friction, regardless of the higher conductivity. It reveals that the doping concentration does not play much role in the swimming performance compared to the geometry or volume effect. It should also be noted that the HNB3 and HNB4 are slightly thicker than HNB1 and HNB2 because of additional metallic or dielectric layers (Cr/Ni 10/10 nm, AZ5214 photoresist around 50 nm Table 1).

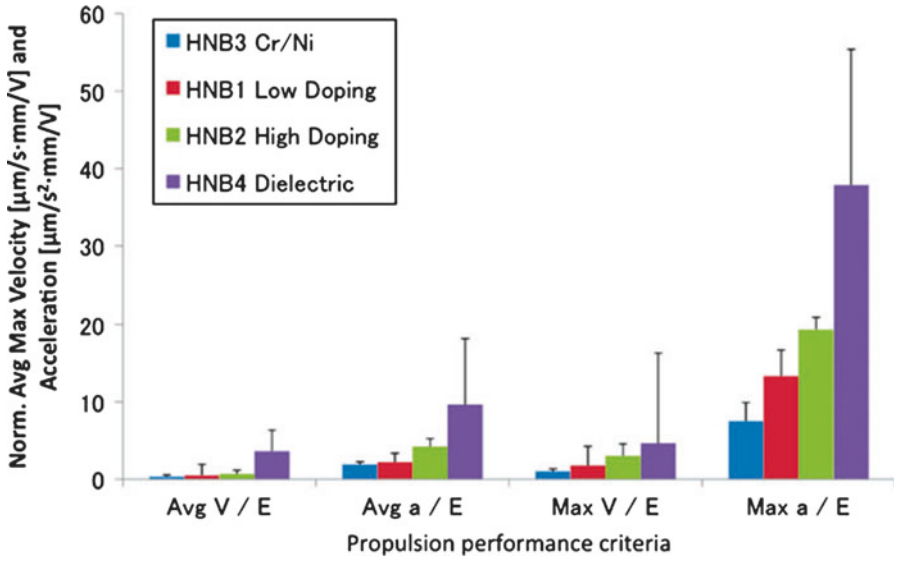
However, the HNB3 had lower performance, while HNB4 was better than the others. In the case of HNB3, this can be explained by the increase in hydrodynamic friction during swimming which slow down HNB. Furthermore, in case of metallic HNB3, the gravitational force can also increase the surface interaction force by dragging it down to the substrate. The detailed calculation and comparison of buoyancy force and gravitational force are not described here; however, the resulting nonfluidic forces (as the difference between the gravitational force and the buoyancy force) on HNB3 is 0.94 pN, which is two times higher than on HNB1 (0.43 pN). These negative effects on HNB3 reduce its swimming performance compared to the others (Fig. 8). The result of HNB4 is contrary to the HNB3 while additional layer was deposited onto the surface. It can be explained by the fact that the nonfluidic force of HNB4 was decreased by the increased buoyancy force from the photoresist, which has much lower density than both gold and nickel. The result clearly shows the high dependency on the surface charge and thus zeta potential in dielectric surface coating to HNB compared to conductive or semiconductor materials (Fig. 8).

Swimming Performance Comparison with Other Swimmers

The force generated by the HNB's swimming is estimated from the drag it overcomes, considering its shape approximately as an ellipsoid [8].

$$F = \frac{4\eta\pi a v}{\left[\ln\left(\frac{2a}{b}\right) - \frac{1}{2}\right]} \tag{4}$$

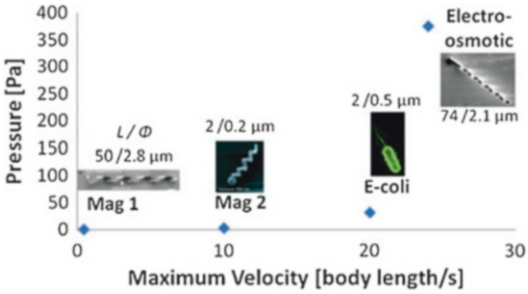
where a and b are the dimensions along the longitudinal and lateral axes respectively, η is the viscosity of the liquid medium; and v is speed of the HNB. For a 74-μm-long HNB with a diameter of 2.1 μm moving at 1.8 mm/s in isopropyl alcohol medium, the generated force of 1.3 nN (Eq. 4) is an order of magnitude higher than the conventional optical trap force which is less than 200 pN. Furthermore, HNBs can apply a large pressure



Remotely Powered Propulsion of Helical Nanobelts, Fig. 8 Dependence of normalized velocity and acceleration of propulsion on four types of HNBs. High- and low-doped semiconductor HNB showed higher swimming performance than metal (Cr/Ni)-coated HNB, and best performances are achieved with type 4 (dielectric)

up to 375.5 Pa, which can be advantageous for local manipulation, such as on biological membranes.

Compared to two rotating magnetic propulsions of helical nanostructures, the demonstrated electroosmotic propulsion of HNBs showed much higher swimming performance in terms of maximum swimming velocity and manipulation force. The demonstrated velocity (1,785 μm/s) can reach 90 times faster and the manipulation force and pressure reaching at around 1.3 nN and 375.5 Pa which are much higher than others respectively (Fig. 9).



Remotely Powered Propulsion of Helical Nanobelts, Fig. 9 Comparison of swimming performances of swimmers: Magnetic 1 [6], Magnetic 2 [8], *E. coli* [1, 19] and Electroosmotic HNB

Conclusion

This entry introduced biologically inspired microswimmers based on helical nanobelts. Helical nanobelts microswimmers have excellent electromechanical and mechanical properties which can mimic their original model in nature. The demonstrated direct pulling of HNBs by electro-osmotic pumping was proven to be an efficient energy conversion mechanism for microscopic artificial swimming objects and overcomes easily the viscous drag and gravitational forces in an aqueous environment at low Reynolds numbers. Swimming velocities as fast as 1.8 mm/s (24 times the body length per second) and pressure as high as 375.5 Pa are achieved by HNBs propelled by an electro-osmotic force under an external electric field. These performances are higher than other state-of-the-art microscale artificial swimmers with physical energy conversion. Note that the velocity of 24 body lengths per second is even faster than natural bacterial propulsion which is

their original model. Although helical morphology with a thin film is beneficial to such a fast swimming at low Reynolds numbers, the swimming performance appears to depend highly on the surface electrokinetic effects. Among the tested four different types, dielectric polymer coated HNB showed the best swimming performance compared to others with metallic or semiconductor surfaces. These HNBs can be used as wireless liquid manipulators, assemblers and nanorobots for biomedical or MEMS/NEMS applications or remote physical or chemical detection by functionalizing with proper readouts.

Cross-References

- ▶ Biomimetics
- ▶ Dielectrophoresis
- ▶ Microfabricated Probe Technology

References

1. Berg, H.C.: *E. coli* in Motion. Springer, New York (2004)
2. Purcell, E.M.: Lift at low Reynolds number. Am. J. Phys. **45**, 3–11 (1977)
3. Sundararajan, S., Lammert, P.E., Zudans, A.W., Crespi, V.H., Sen, A.: Catalytic motors for transport of colloidal cargo. Nano Lett. **8**, 1271–1276 (2008)
4. Honda, T., Arai, K.I., Ishiyama, K.: Micro swimming mechanisms propelled by external magnetic fields. IEEE Trans. Magn. **32**, 5085–5087 (1996)
5. Behkam, B., Sitti, M.: Bacterial flagella-based propulsion and on/off motion control of microscale objects. Appl. Phys. Lett. **90**, 023902(1–3) (2006)
6. Zhang, L., Abbott, J.J., Dong, L.X., Peyer, K.E., Kratochvil, B.E., Zhang, H., Bergeles, C., Nelson, B. J.: Characterizing the swimming properties of artificial bacterial flagella. Nano Lett. **9**, 3663–3667 (2009)
7. Dreyfus, R., Baudry, J., Roper, M.L., Fermigier, M., Stone, H.A., Bibette, J.: Microscopic artificial swimmers. Nature **437**, 862–865 (2005)
8. Ghosh, A., Fischer, P.: Controlled propulsion of artificial magnetic nanostructured propellers. Nano Lett. **9**, 2243–2245 (2009)
9. Abbott, J.J., Peyer, K.E., Lagomarsino, M.C., Zhang, L., Dong, L.X., Kaliakatsos, I.K., Nelson, B.J.: How should microrobots swim? Intl. J. Rob. Res. **28**, 1434–1447 (2009)
10. Martel, S., Tremblay, C., Ngakeng, S., Langlois, G.: Flagellated magnetotactic bacteria as controlled MRI-trackable propulsion and steering systems for medical nanorobots operating in the human microvasculature. Intl. J. Rob. Res. **28**, 571–582 (2009)
11. Kosa, G., Jakab, P., Hata, N., Jolesz, F., Neubach, Z., Shoham, M., Zaaroor, M., Szekely, G.: Flagella swimming for medical micro robots: theory, experiments and application. Proceeding of the 2nd IEEE RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics, pp. 258–263. Scottsdale (2008)
12. Chang, S.T., Paunov, V.N., Petsev, D.N., Velev, O.D.: Remotely powered self-propelling particles and micropumps based on miniature diodes. Nat. Mater. **6**, 235–240 (2007)
13. Gao, P.X., Mai, W., Wang, Z.L.: Superelasticity and nanofracture mechanics of ZnO nanohelices. Nano Lett. **6**, 2536–2543 (2006)
14. Hwang, G., Hashimoto, H., Bell, D.J., Dong, L.X., Nelson, B.J., Schon, S.: Piezoresistive InGaAs/GaAs nanosprings with metal connectors. Nano Lett. **2**, 554–561 (2009)
15. Wang, Z.L., Song, J.H.: Piezoelectric nanogenerators based on zinc oxide nanowire arrays. Science **312**, 242–246 (2006)
16. Xu, S., Qin, Y., Xu, C., Wei, Y., Yang, R., Wang, Z.L.: Self-powered nanowire devices. Nat. Nanotechnol. **5**, 366–373 (2010)
17. Karrai, K., Grober, R.D.: Piezoelectric tip-sample distance control for near field optical microscopes. Appl. Phys. Lett. **66**, 1842 (1995)
18. Hunter, R.J.: Foundations of Colloid Science. Oxford University Press, New York (2010)
19. Berg, H.C., Brown, D.A.: Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking. Nature **239**, 500–504 (1972)

Replica Molding

- ▶ Microcontact Printing
- ▶ Nanoscale Printing

Resistive Switching

- ▶ Graphene Nanostructures for Memristive Devices

Resonant Evanescent Wave Biosensor

- ▶ Whispering Gallery Mode Resonator Biosensors

Retina Implant

▶ Artificial Retina: Focus on Clinical and Fabrication Considerations

Retina Silicon Chip

▶ Artificial Retina: Focus on Clinical and Fabrication Considerations

Retinal Prosthesis

▶ Epiretinal Prosthesis

Reversible Adhesion

▶ Gecko Effect

RF Filters

▶ Tunable RF MEMS Filters: A Review

RF MEMS Reconfigurable Filter

▶ Tunable RF MEMS Filters: A Review

RF MEMS Switches

▶ Capacitive MEMS Switches

RF MEMS/NEMS

▶ Graphene NEMS

RF-MEMS/NEMS

▶ Nanotechnology

Rheology

▶ Dissipative Particle Dynamics, Overview

Riblets

▶ Shark Skin Drag Reduction

Rigorous Maxwell Solver

▶ Finite Element Methods for Computational Nano-optics

RNA Interference (RNAi)

▶ RNAi in Biomedicine and Drug Delivery

RNAi in Biomedicine and Drug Delivery

Michèle Riesen and Nuria Vergara-Irigaray
Department of Genetics Evolution and Environment, Institute of Healthy Ageing, University College London, London, UK

Synonyms

RNA interference (RNAi); Short-interfering RNA (siRNA); siRNA delivery

Definitions

RNA interference (RNAi) is the targeted down regulation of gene expression, most commonly

on the translational level by introducing synthetic double-stranded or short interference RNA (dsRNA, siRNA) into the cell. Gene expression is then reduced by employing an existing cellular mechanism for processing the dsRNA into the active, single-stranded form incorporated in a complex called RISC, which targets and leads to cleavage of the specific messenger RNA (mRNA) substrate.

Biomedicine is the field of medicine that develops concepts or mechanisms discovered or used in basic research into clinical applications. This process is called *From Bench to Bedside*. In the context of RNAi or drugs that are delivered to the organism or target tissue by using nanoparticles, biomedicine is sometimes referred to as nanomedicine. The size of the particles ranges between 1 and 300 nm.

Overview

The recent advances in the understanding of the molecular mechanisms of RNA interference (RNAi) have led to the concept that dsRNA can be delivered into cells to modulate pathogenic functions of genes. This mechanism opens new possibilities to cure diseases since the underlying cause of nearly every pathological process is the dysregulation of one or several genes at different steps in a physiological process. This therapeutic application of RNAi as a new type of drug has been quickly embraced by many researchers as a promising strategy for diseases which were hitherto considered to be difficult or nigh impossible to treat. However, the delivery of dsRNA into cells is not a trivial task due to their large size of approximately 14 kDa and negative surface charge. Additionally, RNA has a rather short lifespan under physiological conditions due to the efficient removal by nucleases.

In order to successfully employ RNAi to target pathological processes, the stability of siRNA has to be addressed. This is generally achieved by introducing chemical modifications to the nucleotides without inducing toxicity or through the selected drug vesicle. Since RNAi is highly sequence-specific, an appropriate sequence in the gene of interest has to be found and preferentially tested using different siRNAs against the same target. Further important factors which should be controlled for are off-target effects (OTEs), for example, cleavage of other mRNA than the sequence-defined target, which can arise despite the precision of the mechanism, and activation of the immune response which can be induced both sequence dependently and independently. The major challenge is still the delivery of the siRNA to the target organ for the treatment. This then expands the list of considerations to the choice of delivery vehicles and routes of administrations. For an excellent and comprehensive introduction into the field please see [11, 17].

Some of these issues, in particular the delivery of the siRNA, are solved by combining the expertise of material, chemical, and physical science to devise efficient vectors for delivery of the dsRNA into the target tissue. The development of carrier molecules in the nano-range, such as polymers, liposomes, peptides, and aptamers, goes hand in hand with the design of new therapeutics for various pathologies, such as cancer, genetic, and viral diseases. Notably, in less than 10 years since the discovery of RNAi, several drugs have been successfully developed and are currently tested in Phase I–III clinical trials in the United States (www.clinicaltrials.gov).

It is worth mentioning that RNAi can originate from either synthetic siRNA, which has to reach the cytosol for final processing and activation, or as plasmid-borne short hairpin RNA (shRNA). These plasmids must be delivered to the nucleus where the RNA is transcribed and undergoes a first step of processing before it is transported to the cytosol. This type of RNAi is generally carried by viral vectors and has been subject of extensive investigation. However, the focus of this review is nonviral delivery of siRNA; hence, the reader is referred to elsewhere for a further discussion of this area. Finally, bacteria have recently been employed as delivery vehicles for shRNA or siRNA in a process termed transkingdom RNAi (tkRNAi) or bacteria-mediated RNAi (bmRNAi).

In this entry, the basic cellular mechanism of RNAi will be introduced. Important criteria for the design of siRNA such as sequence design and

chemical modifications will be discussed together with observed off-target effects (OTEs). This will be followed by an overview over physiological barriers and approaches chosen to deal with these challenges. Further, different types of delivery strategies will be presented which are currently employed in clinical applications according. Finally, how nanotechnology is integrated with RNAi will be discussed with an overview and discussion of a selection of particles presently utilized for in vitro, in vivo, and clinical trials.

Basic Mechanisms

RNA Interference: A Shared to Defend the Genome

In 2006, Andrew Fire and Craig Mello were awarded the Nobel Prize in Medicine or Physiology for their discovery of gene silencing through double-stranded RNA (dsRNA) in *Caenorhabditis elegans (C. elegans)*. Notably, their work describing this phenomenon had only been published for 8 years before they received the honors [5]. The observation that gene silencing could be achieved by the introduction of plasmid DNA into plants was first described some 20 years ago, but the importance of this curious finding was not recognized until it was reported by Andrew Fire and colleagues in one of the most widely used model organisms, the nematode worm *C. elegans*.

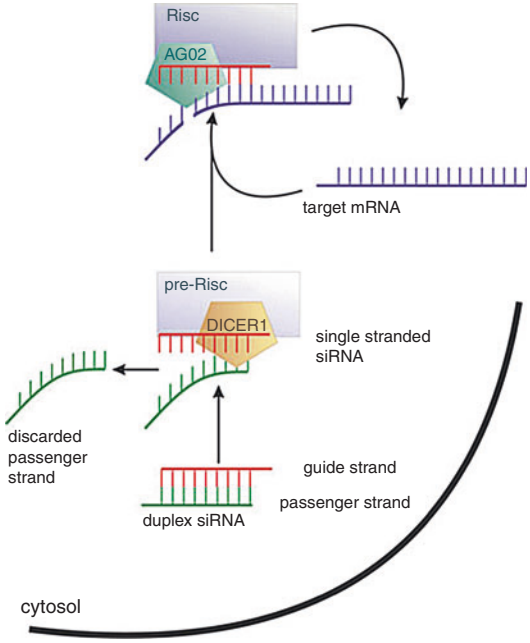
Although the exact mechanisms of RNA interference (RNAi) were unknown when it was first reported, not much time passed before it became clear that the strategies and functions of RNAi are conserved in plants, invertebrates and vertebrates, as well as many of the components of the pathway. Once long dsRNA has entered the cell, it is trimmed into small duplexes of approximately 20–25 nucleotides in the cytosol by an RNase III enzyme called DICER1. This step is followed by integration into a complex called pre-RNA induced silencing complex (pre-RISC), also thought to be mediated by DICER1, where further processing into single-stranded RNA (ssRNA) takes place. RISC contains DICER1 and several proteins from the Argonaute family, AGO1–4. The passenger (sense) strand is then degraded and the mature RISC is formed with the guide (antisense) strand integrated into RISC. The guide strand is complementary to mRNA and determines the specificity of RISC. When RISC encounters its target mRNA and perfect Watson–Crick base-pairing occurs, the mRNA is cleaved by the RISC-member AGO2. If the base-pairing is imperfect, the protein translation is disabled, but the mRNA is not cleaved by AGO2. This complex can be assembled stably over an extended period of time and catalytically regulate gene expression on the translational level (Fig. 1).

Since the discovery of RNAi, the knowledge of the field has progressed tremendously. Species possess a vast arsenal of their own small RNAs, for example, microRNAs (miRNA) or PIWI-interacting RNAs (piRNAs). The different classes of RNAs fulfill different functions such as regulating gene expression at the translational level, or they are expressed in specific tissues, for example, the germ line or neurons. Furthermore, RNA was found in the nucleus where it was found to regulate gene silencing at the transcriptional level by modulating histone modifications or DNA methylation levels. Further evidence proposes RNAi as some kind of nuclear pest control to defend the genome against insertion of mobile elements such as L1 transposons and viral DNA. Despite the rapid succession of discoveries of RNAi-governed processes, the detailed mechanisms by which these regulations are executed remain yet to be elucidated. Nevertheless, the impact the study of RNAi has had on the understanding of gene regulatory processes over the past decade must not be diminished. For comprehensive and recent reviews over the different pathways, processes, and the components of RNAi please refer to [12, 19].

Design of Synthetic Short RNA for Clinical Applications

Sequence Design

Intensive research into sequence design resulted in a set of criteria that greatly facilitates the development of suitable RNA duplexes for any given gene. There are a number of web tools available from either commercial or academic sources in order to identify appropriate RNAi target



RNAi in Biomedicine and Drug Delivery, Fig. 1 siRNA-induced silencing. Double-stranded siRNA is delivered to the cytosol, where it is recognized by DICER1, and incorporated into a pre-RISC complex upon which the duplex is separated and the passenger strand discarded. This processing step is also mediated by DICER1. The mature RISC complex is then formed and guided to target mRNA. Perfect match interactions between the guide (antisense) strand and mRNA lead to cleavage of the substrate mRNA by AGO2. A fully assembled RISC complex is stable and can catalytically inactivate a large number of mRNA. The model is a simplified depiction of the siRNA pathway only and omits transport of siRNA into the cell as well as other small RNA processing pathways. The nomenclature refers to human proteins

sequences [15]. The basic structure of the siRNA generally contains a central domain of 19 basepairs with an overhang of two bases at the 3′-end and a 5′-phosphate group. The reason for this type of synthetic siRNA to be common is because it is identical with the end product of the cellular RNAi pathway and DICER1. However, other siRNA varying in length between 19 and 27 basepairs and with either blunt or asymmetric ends have been shown to be efficient in vitro and to be processed by DICER1. Secondary structures of the siRNA as well as the target sequence should be considered together with the overall thermodynamic properties (T_m), which influence loading into RISC. As a rule, siRNA is more efficient when the 5′-end of the guide strand has a lower melting temperature than the 5′-end of the passenger strand by adjusting the A/U and G/C content, respectively. The positioning of individual bases has also been found to have an effect on efficiency of siRNA [2, 15].

Chemical Modifications to the Sugar or Phosphodiester Backbone

Chemical modifications of the dsRNA in order to increase the stability (T_m) under physiological conditions or reduce the susceptibility to RNase attacks include the introduction or replacement of various reactive groups on the base, sugar or phosphodiester backbone of the RNA. In general, the sugar backbone is modified at the 2′*O*- position, for example, to 2′-Fluoro. This modification does not interfere with loading into RISC or cleavage by DICER1. Another commonly used modification is 2′-*O*-Methyl. This nucleotide is occurring naturally in tRNA or ribosomal components and insertion of these modified nucleotides has been shown to abolish the induction of the innate immune system in mice. Locked nucleic acids (LNA), where the sugar ring is locked in the 3′-*endo* conformation due to the presence of a methylene bridge between the 2′-*O* and the 4′-C of the ribose, have been reported to increase stability successfully.

Modifications to the phosphodiester backbone include, for example, phosphorothioate, which has been shown to increase T_m and resistance to nuclease-mediated degradation. This modification can be easily integrated at defined positions, but it has been related to toxicity, loss of function, and nonspecific protein binding if large numbers of nucleotides in the siRNA are replaced with this modification. The replacement of the phosphodiester bond with a boranophosphate greatly increases stability of the dsRNA, but synthesis of this type of backbone requires in vitro transcription and the position of the insertion of boranophosphate can therefore not be defined. For excellent and detailed reviews of the design of siRNA and modifications, please refer to [1, 2, 8].

Off-Target Effects

A serious concern when developing siRNA as therapeutics are OTEs. They arise most commonly for the following reasons: activation of the innate immune system, sequence homology with other mRNAs additionally to the specific target, and when siRNA enters the cellular miRNA pathway. Initially, it was assumed that the short RNA duplexes could not induce an immune response. However, with the increasing number of work employing siRNA in vitro and in vivo, siRNAs were reported to stimulate the innate immune system and induce a strong interferon response. This activation has been shown to be both sequence dependent and independent, as well as cell-type specific. Intracellular receptors from the Toll-like family, in particular TLR3, 7, and 8 recognize RNA, either dsRNA (TLR3) or ssRNA (TLR7 and 8), and mediate the induction of IFN-α or IFN-β. These receptors are found either on intracellular organelles such as the endosome or lysosomes, as well as cells that constitute the innate immune system such as B cells or monocytes. Under given circumstances, siRNA can therefore induce the innate immune system [22].

In order to avoid OTEs based on sequence homology, it is recommended to carefully screen the transcriptome for similarities with the required siRNA. It is important to note that although single mismatches have been reported to abolish the efficiency of siRNA, short similarities of five to eight bases can direct RISC to unintentional targets. If siRNA engage in the miRNA pathway, the specific effect of siRNA on one gene is replaced by effects on many genes due to the nature of miRNA-mediated gene regulation. Generally, OTEs are dose-dependent and can therefore be circumvented by establishing dose–response curves and choosing the lowest efficient dose. Since it has been shown that downregulations of the targeted gene transcripts were in fact due to OTEs and not to the direct mediation of the administered siRNA, it is strongly recommended to test at least two or three different siRNAs against the same gene. This should result in reproducible effects and confirm that the expected and observed downregulation of the transcripts is due to the direct effect of siRNA [8, 15].

Delivery of Synthetic Short Interference RNA for Clinical Applications

Physiological Barriers

After the successful design, synthesis, and in vitro tests of a suitable siRNA compound, the next and likely more challenging obstacle to be overcome is the delivery of siRNA to its site of action. There is a consensus in the field that the vast potential of RNAi cannot be fully exploited yet because of the lack of efficient delivery vehicles to the target tissue. The body has various highly efficient strategies for clearance of foreign particles from the system, which are briefly described below in order to understand the difficulties in delivery and how they are addressed by the choice of appropriate vesicles or siRNA modification or both combined. Tremendous efforts are directed toward the development of a range of particles with different physicochemical properties which are equipped to reach the target organ and pass the physiological barriers [7].

If siRNAs are not immediately degraded by free nucleases then the next barrier is renal clearance, which is thought to determine the half-life of modified or naked siRNA in vivo. The small size of around 14 kDa, which is far below the cutoff of 40 kDa for glomerular filtration, and the hydrophilic nature of siRNA target these molecules for removal by the kidneys. The next trial is the clearance via the reticuloendothelial system (RES). The RES consists of mobile, circulating phagocytic cells, such as monocytes, or resident tissue macrophages that remove pathogens and cellular debris which remains from apoptosis, wound healing, or tissue remodeling. Large numbers of RES-associated cells reside in the liver, the Kupffer cells, and in the spleen or generally in other tissues with high blood flow and fenestrated vasculature.

Should carriers have successfully dodged clearance by any of the above-described pathways, they will then encounter the endothelial barrier lining the vascular lumen. Endothelial cells are tightly connected with the underlying extracellular matrix through integrins and they are joined together via a number of strictly regulated junctions. Endothelial cells are also

perforated with a large number of small pores of about 45 Å and rare large pores of 250 Å. Most of the transport of vesicles or siRNA is thought to be mediated via leaks in junctions rather than the pores themselves; however, there is some controversy regarding this matter. In many tissues, particles larger than 5 nm will not be able to extravasate through the endothelial lining, whereas organs such as the liver or spleen are easily accessible because of the extensive vasculature of these organs.

One specific type of endothelial barrier is the blood brain barrier (BBB) at the boundary to the astrocytes and pericytes where the endothelium forms an exceptionally tight lining to protect the brain. As progress in research into neurodegenerative diseases such as Alzheimer's reveals an ever-increasing number of genes which are potentially "drugable," the brain becomes an attractive target for siRNA. Under normal physiological conditions, this barrier is intact. However, the integrity of the BBB can be compromised during disease or through administration of certain drugs. Also, specific endothelial reporters have been described which could be employed to aid passage of particles across the BBB to the brain. Therefore, development of carriers that can deliver drugs across the BBB will become increasingly important to treat this kind of diseases.

The final step for successful delivery is the cellular uptake and, importantly, endosomal release into the cytosol. In general, particles enter a cell via receptor-mediated endocytosis. Recognition and binding to the receptor induces the formation of a vesicle which leaves the cell membrane on the cytosolic site. From there, the vesicles undergo a series of intracellular trafficking steps and maturation until, finally, late endosomes or lysosomes have been formed. The last challenge the particles face is the release from the endosome into the cell. The *proton sponge effect* has been described as a mechanism by which the trapped drug can exit the endosome. Briefly, acidification of the endosome leads to protonation of amine groups on the vesicle. This leads to an influx of protons and chloride ions, followed by water to equilibrate the changes in osmolarity, eventually causing rupture of the vesicle and release of its contents into the cytosol. For further reading, please see [7, 20].

Types of Delivery

In vivo application of siRNA can be classified, for example, into active or passive systemic and local/cellular delivery. Passive systemic delivery does not require any specific targeting sequences or molecules. It is based on the property of nanoparticles to accumulate in certain tissues under certain conditions. For example, nanoparticles or naked siRNA have been shown to accumulate in liver or spleen because of the specific structure of the endothelium, with large fenestrations of the microvasculature of these organs. Delivery of vesicles to inflamed tissue is thought to be mediated by the RES. Accumulation of particles in tumors has been attributed to the *enhanced permeability and retention effect* (EPR) in tumors. This is thought to be caused by reduced lymphatic drainage, leaky microvasculature, and enhanced angiogenesis of cancerous tissue.

Active or targeted systemic delivery is more complex because it requires the carrier to be equipped with components such as antibodies, ligands against cell-surface receptors, or peptides from viral proteins targeting the nanoparticle to the desired cell type or tissue. Antibodies are used as whole entities, or fragments thereof, and they are generally directed against receptors which are increasingly expressed on tumor cells. Receptors that have been targeted include, for example, the epidermal growth factor receptors (EGFR) or human epidermal growth factor receptor-2 (HER-2), the iron-binding protein transferrin (Tf), which targets the transferrin receptor (TfR) responsible for iron uptake, as well as Toll-like receptors (e.g., TLR3). Successfully employed peptides are the rabies virus glycoprotein (RVG) or arginine-glycine-asparagine (RGD). Sugars, i.e., glucose or mannose, are recognized by a number of cell-type specific receptors and have also been successfully employed in delivery to specific cell types or across the BBB.

Naked siRNA has been injected using the hydrodynamic delivery method into mouse tail veins and was subsequently found to accumulate

in the liver and to a lesser extent in other organs such as the lungs. This method requires large volumes to be rapidly injected, and although successfully employed in mice and pigs, it is doubtful whether this way of administering systemic siRNA would be of therapeutic use in humans. The volume for injection needed is around 10 % of the body weight, and this would most likely lead to heart failure and death. Direct injection of siRNA is feasible if the target is easily accessible, such as is used, for example, for treatment of wet age-related macular disease or other diseases of the eyes where siRNA is injected directly into the eye. Several Phase I–III clinical trials have been approved by the US FDA. Other means of delivery that have been tested include topical administration which is useful for exposed sites, such as the skin. Inhalation is the means of choice for treatment of respiratory tract diseases. Electroporation has been used for delivery into muscle tissue. For more detailed information, please see [14, 18, 21, 22].

Delivery Vesicles and Nanoparticles

Considerations for the Design of an siRNA Carrier Investigation into suitable carriers over the past decade has resulted in a vast amount of materials and combinations of several components which all have advantages and disadvantages for delivery of specific cargo, the route of administration, and the target tissues. Criteria to be considered are the size and size distribution of particles, which is, for example, inversely related to cellular uptake and delivery to tumor tissue. Since the smallest microvessels have a diameter of 5–6 μm, particles must be considerably below this diameter in order to avoid accumulation and clogging which could lead to embolism. The surface properties of nanoparticles dictate whether they will be recognized and removed by the host immune system; non-modified hydrophobic particles are generally swiftly removed, whereas vesicles decorated with biodegradable materials which are hydrophilic such as polyethylene glycol (PEG) or polysorbate 80 (Tween 80) are better tolerated.

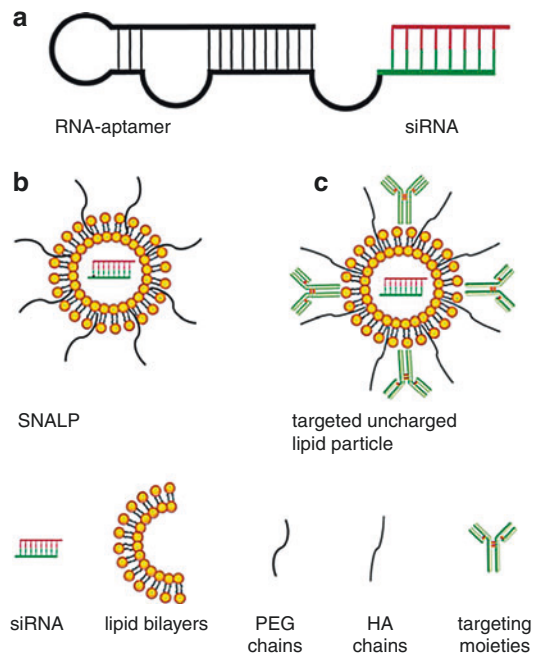
Drug loading capacities and the release characteristics are further important points. Loading can be achieved by two methods. The drug can be incorporated at the time of nanoparticle assembly or it can be absorbed after the formation of the carrier by incubation in a concentrated drug-solution. Highly efficient loading has been shown for siRNA due to the negative charge of the oligonucleotide and positively charged vesicles, such as cationic lipids or polymers. Drug release depends, for example, on the degradation of the carrier or how fast the drug can diffuse across the polymer membrane, as well as the solubility of the drug. A concise selection of the most frequently and successfully employed carriers is briefly described below. For a detailed description of these design considerations, please see [4, 18]. The reader is also referred to cross-referenced entries which elaborate in more detail the properties and design of nanoparticles specifically for drug delivery.

Aptamers Among biological macromolecules RNA has a unique place because of the ease of manipulation and synthesis similar to DNA, which is combined with a structural versatility found in proteins. These useful properties predestine RNA to be used itself as a building block for the assembly of nanoparticles called RNA-aptamers (Fig. 2a). Generally, aptamers are highly structured oligonucleotides that are obtained by SELEX (systematic evolution of ligands by exponential enrichment) for the required purpose. This is a protocol whereby a library of RNA-aptamers is subjected to iterative rounds of counter-selection and selection to yield aptamers which recognize, for example, specific receptors on target cells, proteins, as well as organic compounds. The prime features which make aptamers attractive carriers are their small size of a few nanometers compared to cationic polymers or liposomal carriers (upto 300 nm), that modified nucleotides can easily be integrated to increase stability in physiological conditions, they have low toxicity and are biodegradable. Instances where aptamers have been employed as carriers for siRNA include murine xenografts of prostate cancer where tumor growth and regression was mediated by aptamer-mediated siRNA delivery, or treatment against viral infection by

either targeting HIV gp120, a glycoprotein which recognizes and utilizes the CD4 receptors on T lymphocytes for entry, or the CD4 receptor itself [6, 10].

Cationic Polymers and Dendrimers The cationic polymer cyclodextrin (CDP) has been widely and successfully used as a carrier directed in particular against solid tumors. Advantages of this material are the low toxicity; it is not metabolized in humans and does not induce an immune response. CDPs are commonly of a tri-composite structure, where CDP is employed as the carrier for siRNA due to strong ionic interactions between the cationic polymer and the anionic nucleic acids. The particle is stabilized with the hydrophilic cycloalkane adamantane conjugated to PEG (AD-PEG). Targeting of the vesicle to tumors is achieved by adding an AD-PEG-Transferrin (AD-PEG-Tf) moiety because Tf is expressed abundantly in a number of tumors. This formulation carrying siRNA against the ribonucleotide reductase subunit 2 (RRM2), called CALAA-01 by Calanda Pharmaceuticals, was approved for Phase I clinical trials by the FDA in 2008. The reader is referred to [3, 9] for a comprehensive review on the development of CALAA-01.

Another cationic species that is frequently used for similar properties is polyethylene-imine (PEI), which can be synthesized to be of linear or branched structure and is then classified as a dendrimer. PEI efficiently forms stable nanoparticles with oligonucleotides and promotes endosomal release via the *proton sponge effect*. Because of its branched structure, it can carry a large payload. Despite the supreme quality of PEI as a siRNA delivery vesicle, its identified toxicity toward most cells is not a desirable side effect. Three approaches have been taken to address the toxicity issue of PEI. First, the dendrimer is equipped with targeting moieties such as RGD to reduce the toxicity to nontarget tissue. Second, maintaining the particle size below 10 kDa substantially reduced toxicity compared to larger particles of 25 kDa. Finally, ketalized carriers have been synthesized with the added advantage that these linkages are sensitive to acidic pH, and hence readily dissolve in endosome to release the cargo [13].



RNAi in Biomedicine and Drug Delivery, Fig. 2 Delivery vehicles for siRNA administration. Three types of vesicles are shown for delivery of siRNA. (**a**) RNA-aptamer with siRNA ligated. (**b**) Stable nucleic acid particle decorated with PEG (SNALP). (**c**) Uncharged lipid particle decorated with targeting moieties, i.e., single chain antibody fragment directed against Tf and HA chains

Liposomal Carriers Lipids as gene delivery vehicles have been used since the early 1980s. Cationic lipids are often used as vesicle due to the facile handling and preparation since they readily associate with the anionic nucleic acids. Moreover, routinely used reagents for in vitro transfection of cell culture are based on cationic liposomes, such as *Lipofectamine 2000* by Invitrogen or other equivalent products. The most commonly used cationic lipid is 1,2-dioleoyl-3-trimethylammonium-propane (DOTAP) because it efficiently associates with nucleic acids. However, DOTAP was again shown to be cytotoxic; hence, this carrier was equipped with modifications which reduce this side effect, such as lactosylation. Another strategy to inhibit the growth of pancreatic xenografts on a murine model involved DOTAP decorated with histidine-lysine peptides. This facilitated endosomal escape and single-chain antibody

fragment directed against Tf, and this vesicle did not induce an IFN-response and could be administered at low doses due to the target specificity.

Another promising type of vesicle based on cationic lipids is the stable nucleic acid-lipid particle (SNALP) (Fig. 2b). The siRNA is embedded within the lipid envelope and the composition of SNALPs contains reduced cationic lipid content and is PEG-ylated. SNALPs have been subjected to extensive study as siRNA carrier against in vivo viral models with considerable success of full protection against infection of guinea pigs with the deadly Zaire strain of the Ebola virus. Another trial employed SNALPs with siRNA against apolipoprotein B levels in nonhuman primates, which was well tolerated and showed reduction of Apolipoprotein B (ApoB) and serum cholesterol among other physiological improvements. Following on from these promising results, several therapeutic siRNA-SNALP formulations against tumors or metabolic diseases are now in preclinical or have been admitted to Phase I clinical trials, and targets include VEGF and ApoB among others [8]. Further information on these trials can be found on www.clinicaltrials.gov or the web sites listed.

Since cationic liposomes have been associated with significant toxicities, development of vesicles composed of uncharged lipids has been resumed with considerable success. One lipid in particular has emerged as a promising carrier: 1,2-dioleoyl-sn-glycero-3-phosphatidylcholine (DOPC). These vesicles successfully reduced tumor xenografts of different origins. Targeted vesicles against the thrombin receptor (TR) expressed on some tumors greatly reduced not only TR but also angiogenic and invasive tumor markers, such as vascular enodothelial growth factor (VEGF) and matrix metalloproteinase 2 (MMP2). Techniques developed in order to maximize the payload of these nanoparticles utilize lyophilization followed by rehydration with siRNA. The half-life in circulation under physiological conditions of these particles was increased by surface coating with hyaluronic acid (HA), a naturally occurring glucosaminoglycan with properties equivalent to PEG (Fig. 2c). Further information can be found in [17].

Other Types of Carriers: Atellocollagen, Cationic Cell Penetrating Peptides (CPPs), Polylysine There are a number of other options for building siRNA-nanocarriers. Atellocollagen is a preparation of digested cationic calf dermal collagen successfully employed and well tolerated in animal models of inflammatory disease or tumor xenografts. Despite the fact that the studied atellogen-siRNA nanoparticles were not decorated with a specific targeting moiety, they were found to accumulate in tumors and inflamed areas presumably due to the EPR effect in these tissues. CPPs are short peptides with a high content of positively charged amino acids, in particular arginine, which can be synthesized or are derived from naturally occurring peptides, e.g., TAT from HIV-1 to facilitate cellular uptake. Initially thought to be fully soluble, it appears that their efficiency is due to aggregation into nanoparticles. Successful trials in vivo used polyarginine fused to cholesterol and siRNA against VEGF in a tumor xenograft. Another polyarginine-based CPP coupled with RVG was found to cross the BBB and reduce green fluorescent protein expression in a mouse model. Another type of carrier is based on polylysine (K) stretches. Initial problems with toxicity such as complement activation were overcome by PEG-modifications and the addition of histidine (H), which facilitates endosomal escape. Further, it was found that the efficiency as siRNA carriers was improved by increasing the histidine ratio compared to lysine [9].

Further Reading

Currently, 20 studies are listed on www.clinicaltrials.gov when searched for the keyword siRNA. This clearly shows that the challenges of delivery posed by the physiological barriers or physicochemical properties of the particles can be overcome. The marriage between advances in material sciences and the developing understanding of the physiological and pathophysiological processes culminates in the development of novel RNAi-based therapeutics for the treatment of diseases ranging from viral to inherited and age-related pathologies.

Due to space restraints, a number of additional excellent reviews could not be referenced in this entry; however, they can be found when searching Pubmed or Web of Science with keywords such as "siRNA AND nanotechnology." Further information on research and development of siRNA-based therapeutics can be found on the web sites of the leading companies.

http://www.alnylam.com/
http://www.opko.com/
http://www.quarkpharma.com/
http://www.silence-therapeutics.com/
http://www.sirna.com/
http://www.tekmirapharm.com/Home.asp

An insightful and curated web site on the recent development of RNAi therapeutics, career opportunities, consulting, and overview over the market and enterprise portfolios can be found here:

http://rnaitherapeutics.blogspot.com/

Cross-References

- ▶ Genotoxicity of Nanoparticles
- ▶ Lab-on-a-Chip for Studies in *C. elegans*
- ▶ Liposomes
- ▶ Nanomedicine
- ▶ Nanoparticle Cytotoxicity
- ▶ Nanoparticles
- ▶ Nanotechnology
- ▶ Nanotechnology in Cardiovascular Diseases

References

1. Behlke, M.A.: Progress towards in vivo use of siRNAs. Mol. Ther. **13**, 644–670 (2006)
2. Behlke, M.A.: Chemical modification of siRNAs for in vivo use. Oligonucleotides **18**, 305–319 (2008)
3. Davis, M.E.: The first targeted delivery of siRNA in humans via a self-assembling, cyclodextrin polymer-based nanoparticle: from concept to clinic. Mol. Pharm. **6**, 659–668 (2009)
4. Fattal, E., Barratt, G.: Nanotechnologies and controlled release systems for the delivery of antisense oligonucleotides and small interfering RNA. Br. J. Pharmacol. **157**, 179–194 (2009)
5. Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., Mello, C.C.: Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature **391**, 806–811 (1998)
6. Guo, P.: RNA nanotechnology: engineering, assembly and applications in detection, gene delivery and therapy. J. Nanosci. Nanotechnol. **5**, 1964–1982 (2005)
7. Juliano, R., Bauman, J., Kang, H., Ming, X.: Biological barriers to therapy with antisense and siRNA oligonucleotides. Mol. Pharm. **6**, 686–695 (2009)
8. Lares, M.R., Rossi, J.J., Ouellet, D.L.: RNAi and small interfering RNAs in human disease therapeutic applications. Trends Biotechnol. **28**, 570–579 (2010)
9. Leng, Q., Woodle, M.C., Lu, P.Y., Mixson, A.J.: Advances in systemic siRNA delivery. Drugs Future **34**, 721 (2009)
10. Levy-Nissenbaum, E., Radovic-Moreno, A.F., Wang, A.Z., Langer, R., Farokhzad, O.C.: Nanotechnology and aptamers: applications in drug delivery. Trends Biotechnol. **26**, 442–449 (2008)
11. Martin, S.E., Caplen, N.J.: Applications of RNA interference in mammalian systems. Annu. Rev. Genomics Hum. Genet. **8**, 81–108 (2007)
12. Moazed, D.: Small RNAs in transcriptional gene silencing and genome defence. Nature **457**, 413–420 (2009)
13. Mok, H., Park, T.G.: Functional polymers for targeted delivery of nucleic acid drugs. Macromol. Biosci. **9**, 731–743 (2009)
14. Oliveira, S., Storm, G., Schiffelers, R.M.: Targeted delivery of siRNA. J. Biomed. Biotechnol. **2006**, 63675 (2006)
15. Peek, A.S., Behlke, M.A.: Design of active small interfering RNAs. Curr. Opin. Mol. Ther. **9**, 110–118 (2007)
16. Raemdonck, K., Vandenbroucke, R.E., Demeester, J., Sanders, N.N., De Smedt, S.C.: Maintaining the silence: reflections on long-term RNAi. Drug Discov. Today **13**, 917–931 (2008)
17. Schroeder, A., Levins, C.G., Cortez, C., Langer, R., Anderson, D.G.: Lipid-based nanotherapeutics for siRNA delivery. J. Intern. Med. **267**, 9–21 (2010)
18. Singh, R., Lillard Jr., J.W.: Nanoparticle-based targeted drug delivery. Exp. Mol. Pathol. **86**, 215–223 (2009)
19. Siomi, H., Siomi, M.C.: On the road to reading the RNA-interference code. Nature **457**, 396–404 (2009)
20. Tiemann, K., Rossi, J.J.: RNAi-based therapeutics-current status, challenges and prospects. EMBO Mol. Med. **1**, 142–151 (2009)
21. Tokatlian, T., Segura, T.: siRNA applications in nanomedicine. Wiley Interdiscip. Rev. Nanomed. Nanobiotechnol. **2**, 305–315 (2010)
22. Weinstein, S., Peer, D.: RNAi nanomedicines: challenges and opportunities within the immune system. Nanotechnology **21**, 232001 (2010)

Robot-Based Automation on the Nanoscale

Sergej Fatikow, Daniel Jasper, Christian Dahmen, Florian Krohs, Volkmar Eichhorn and Michael Weigel-Jech
Division Microrobotics and Control Engineering (AMiR), Department of Computing Science, University of Oldenburg, Oldenburg, Germany

Synonyms

AFM; Atomic force microscopy; Carbon-nanotubes; DNA manipulation; Manipulating; Nanorobotic assembly; Nanorobotic manipulation of biological cells; Nanorobotics; Nanorobotics for bioengineering; Nanorobotics for MEMS and NEMS; Scanning electron microscopy; Visual servoing for SEM

Definition

The handling of micro- and nanoscale objects is an important application field of robotic technology. Nanohandling includes the finding, grasping, tracking, cutting, etc., of objects as well as different characterization methods such as indenting or scratching on the nanoscale. Measurement of different features of the object, probe positioning with nanometer accuracy, structuring or shaping of nanostructures, and generally all kinds of changes to matter at the nanolevel could also be defined as nanohandling in the broadest sense. This entry addresses several approaches that can be automated with the help of nanohandling robots. The development of nanohandling robot systems working in constricted work spaces (in a vacuum chamber of an SEM or under an optical microscope) and with a nanometer resolution and accuracy is a big technological challenge for the robotics research community. Advanced actuator and sensor technologies suitable for nanohandling have to be investigated and implemented as well as vision-based feedback methods for tracking and data acquisition. Another crucial issue is the development of real-time robot control methods that meet the demands of automated nanomanipulation. This entry presents several promising solutions and applications of robot-based nanohandling.

Robot-Based Automation on the Nanoscale

Introduction to the Topic

The handling of micro- and nanoscale objects is an important application field of robotic technology. It is often referred to as nanohandling, having in mind the range of aspired positioning accuracy for the manipulation of micro- and nanoscale objects of different nature. The nanohandling of objects may include their finding, grasping, moving, tracking, releasing, positioning, pushing, pulling, cutting, bending, twisting, etc. Additionally, different characterization methods like indenting or scratching on the nanoscale, measurement of different features of the object, probe positioning with nanometer accuracy, structuring or shaping of nanostructures, and generally all kinds of changes to matter at the nanolevel could also be defined as nanohandling in the broadest sense. As in the field of industrial robotics, where humans leave hard, unacceptable work to robots, robots with nanohandling capabilities can help humans to handle extremely small objects with very high accuracy. The size of these robots also plays an important role in many applications [1, 2]. Highly miniaturized robots, often referred to as microrobots, are able to operate in constricted work spaces, for example, under a light microscope or in the vacuum chamber of a scanning electron microscope (SEM). Microsystem technology (MST) and nanotechnology require this kind of robot, since humans lack sensing capabilities, precision, and direct manipulation at those scales. Automated nanohandling by microrobots will have a great impact on both these technologies [3]. This entry will address several aspects of nanohandling automation.

The development of nanohandling robot systems is a big technological challenge for the

robotics research community. Advanced actuator and sensor technologies suitable for nanohandling have to be investigated and implemented. Another crucial task is the development of real-time robot control methods that meet the demands of automated nanomanipulation. The latter is challenging, especially due to the difficulty of getting real-time visual feedback and the lack of advanced control strategies able to deal with changing and uncertain physical parameters and disturbances.

Trends in Nanohandling

The following three approaches are being pursued in the nanohandling research community:

- Top-down approach utilizing serial nanohandling by microrobotic systems. The main goal is the miniaturization of robots, manipulators, and their tools as well as the adaptation of robotic technology (sensing, actuation, control, automation) to the demands of MST and nanotechnology. This approach is the major topic of this entry.
- Bottom-up approach or self-assembly, utilizing parallel nanohandling by autonomous organization of micro- and nanoobjects into patterns or structures.
- The use of a scanning probe microscope (SPM) as a nanohandling robot. In this approach, the (functionalized) tip of an atomic force microscope (AFM) probe or of a scanning tunneling microscope (STM) probe acts as a robot end-effector that is applied to a nanoscale part. This approach has been actively investigated in recent years [1, 2] and is discussed in section "AFM as a Nanohandling Robot."

Several other approaches like the use of optical tweezers or electrophoresis might also be adapted for automated nanohandling.

Automated Microrobot-Based Nanohandling

Microrobotics for handling micro- and nanoscale parts has been established as a self-contained research field for nearly 15 years and, in recent years, a trend toward the microrobot-based automation of nanohandling processes has emerged [4, 5]. Process feedback is the most crucial aspect of nanohandling automation. It is difficult to obtain reliable information when handling microscale and especially nanoscale parts. Vision feedback often is the only way to control a nanohandling process. For this reason, the vacuum chamber of a scanning electron microscope is, for many applications, the best place for a nanohandling robot. It provides an ample work space, very high resolution up to 1 nm, and a large depth of field. However, real-time visual feedback from changing work scenes in the SEM containing moving nanorobots is a challenging problem.

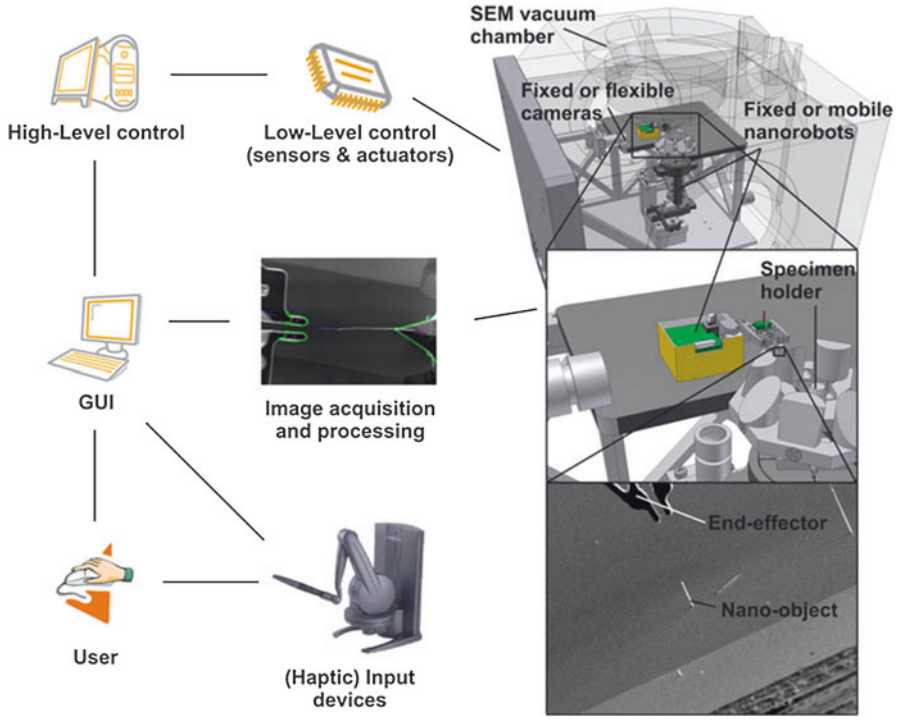
Automated Microrobot-Based Nanohandling Station (AMNS)

Figure 1 presents a generic concept of an automated microrobot-based nanohandling station (AMNS).

The microrobots are usually driven by piezoactuators that enable positioning resolutions down to sub-nanometer ranges. The microrobots of the station have a nanomanipulator integrated in their mobile platform, which makes them capable both of moving over longer distances and of manipulating with nanometer accuracy. This leads to more flexibility as the robots can be deployed anywhere inside the SEM vacuum chamber. Stationary robots can also be used in an AMNS. They are easier to control compared to mobile robots, which makes them more suitable for high-throughput automation. Depending on the application, different combinations of both robot types can be implemented.

The sensor system of the AMNS includes SEM, video cameras, force sensors, as well as position sensors integrated into the robots axes. The sensor data is sent to the station's control system for real-time signal processing. The control systems task is to calculate the positions of the robots and their tools as well as the positions of the parts to be handled or other objects of interest. The calculated positions serve as input data for the closed-loop robot control.

Teleoperation often is the first step on the way to automation. For this reason, a user interface is an important component of the AMNS.



Robot-Based Automation on the Nanoscale, Fig. 1 Generic concept of the AMNS [2]

The user can influence the handling process by a haptic interface and/or by a graphical user interface (GUI). A good overview of state of the art teleoperation techniques and applications is given in [6].

The positioning accuracy of the microrobots during automated nanohandling is affected by several factors, so that a powerful robot control system is required. On the low level of the control system, driving voltages for the robot actuators are calculated in real time, which keep the robot and its end-effector on the desired path. The high-level control system is responsible among others for path planning, error handling, and the time-saving parallel execution of tasks. GUI and haptic interface are supported by the high-level control system as well.

The AMNS concept has been implemented in different application fields for (semi-) automated nanomanipulation [2, 3]. Relevant aspects of this work will be introduced in the following sections.

Structure of the Essay

The following is a brief summary of the topics covered in the following sections.

Section "Real-Time Vision Feedback" covers the vision feedback in nanohandling robot stations. The focus is on the use of SEM in combination with real-time image processing algorithms, which is proposed as the near-field sensor for the automation of nanohandling tasks.

Section "Position Control Inside SEM" introduces fundamental considerations for automated positioning inside the SEM. The section encompasses actuation types, motion principles, closed-loop control mechanisms for nanorobots, SEM-based visual servoing, as well as the difference between macro- and microscale control engineering.

Section "AFM as a Nanohandling Robot" introduces the atomic force microscope as a robotic system for manipulation at the nanoscale. The AFM tip acts as a nanohandling tool to

change the position or orientation of a nanoobject or to modify the surface of a substrate with nanometric precision. Different application fields are introduced and current challenges in this field of research are discussed.

Section "Application Example: Handling of CNT" describes nanorobotic strategies for the microgripper-based pick-and-place handling of individual carbon nanotubes. For a reliable alignment in z-direction, a shadow-based depth detection technique is presented. Furthermore, special handling strategies are introduced for placing and releasing the CNT on the target structure. As an example, the automated assembly of so-called CNT-enhanced AFM probes is demonstrated.

Section "Application Example: Handling of Biological Materials" deals with the characterization and manipulation of biological objects by an atomic force microscope. An overview of relevant parameters for AFM measurements of samples in liquids and of soft samples is given. After a short introduction of the biological background, the current state of the art in automated AFM-based handling of biomaterials is given, ranging from the design of DNA-based nanoelectric circuits to the structuring of biosensor chips for medical and forensic use.

Real-Time Vision Feedback

Manipulation and handling of objects on the nanoscale is subject to strong uncertainty on position and status of objects. This is due to the low mass of the objects of interest and the relatively high parasitic and interacting forces. Additionally, the choice of imaging modalities available at this scale is limited. Optical microscopes cannot deliver resolutions sufficient for imaging nanosized objects. Scanning probe microscopes like the AFM, though having the advantageous possibility to be used as a manipulator, do not exhibit the necessary speed for closed-loop feedback. The most flexible solution for imaging in this context is the SEM. It allows for observing processes at the nanometer range and still is fast enough to enable closed-loop control feedback. Drawbacks of the SEM are the sensitivity to noise and movement artifacts during fast scans, the possibility of distortions due to magnetic or electric influences generated by the AMNS, and the lack of three-dimensional information. Various approaches have been developed to deal with these conditions. The important properties of the SEM imaging and the tasks and approaches to solve the problems associated with it will be described in the following sections.

Imaging Source Properties

Certain properties of the SEM and its imaging process are important for the applied computer vision approaches. The most important negative property is the high level of noise. Noise can be reduced by frame averaging or pixel averaging which, however, also reduces the frame rate. Figure 2 shows the SEM images with low speed and high speed and the corresponding noise level.

The noise in the SEM can be modeled as additive noise:

$$g_{\mathrm{SEM}} = \mathrm{g} + \mathrm{g}_{\mathrm{noise}}$$

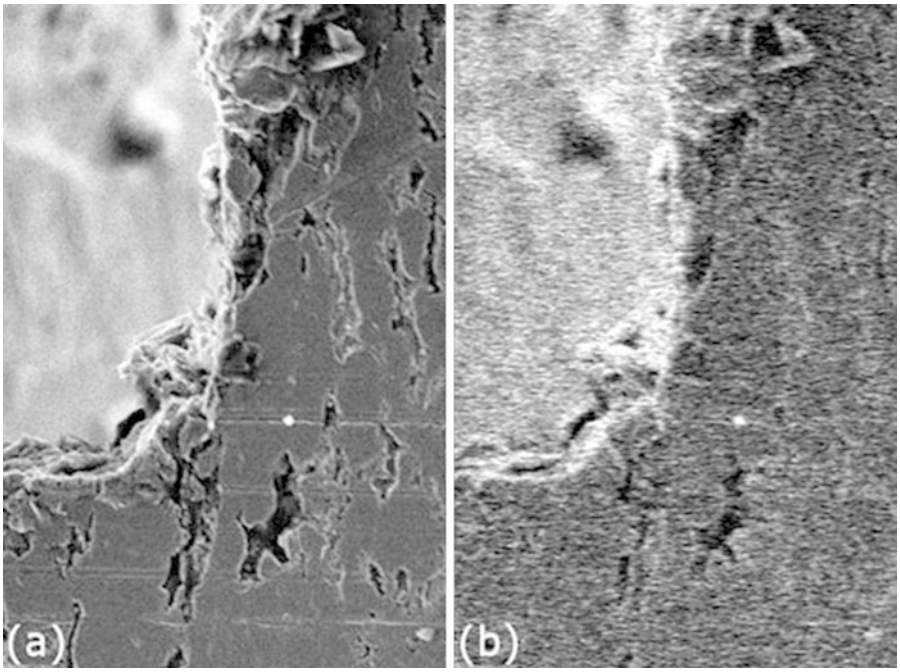
With g the true gray level value, $\mathrm{g}_{\mathrm{noise}}$ the noise component, and g_{SEM} the acquired gray level value.

Because the generation of secondary electrons is a statistic process and the number of electrons generated and detected can be assumed as Poisson distributed, the gray level fluctuation also has a Poisson characteristic.

Due to the scanning principle of the SEM, fast moving objects appear distorted in the images. This is due to the fact that the object has moved during the time needed to finish each line of the scan. This can significantly reduce the performance of computer vision algorithms. Therefore, the speed of robots has to be restricted if computer vision algorithms are required.

Typical Visual Feedback Tasks

For a successful nanohandling automation, visual feedback is needed for specific tasks. These can be described as object recognition, classification, and position determination. The latter can be executed in two dimensions and in three dimensions, respectively. Important especially for the position determination is speed in order to allow for closed-loop control. Still, even with fast imaging,



Robot-Based Automation on the Nanoscale, Fig. 2 SEM images at different scan speeds, acquired with a LEO 1,450 tungsten emitter SEM, (**a**) slow scanning, (**b**) fast scanning

control is often restricted to a look-then-move strategy, because the actuation is orders of magnitude faster than the imaging and subsequent visual feedback. When using SEM image acquisition with computer vision approaches as an image-based sensor, this is a limitation which cannot be overcome easily.

Object Recognition

Object recognition has the task of determining if an object of interest is in the view field or not. Many different approaches are known in traditional macro- or microscale imaging and computer vision. The approaches used at the nanoscale include template matching, statistical pattern recognition, and recognition based on local features. Object recognition algorithms used for nanohandling automation in many cases do not have special and elevated needs for noise robustness, because the recognition task is not time critical and low frame rates with high averaging can be used. Approaches used in nanohandling applications use support vector machines and various feature extraction algorithms. Applications include CNT-detection and detection of TEM lamellae [7].

Two-Dimensional Position Determination and Tracking

The positions of objects and tools are not known precisely through any sensor except for the SEM. Object position determination and tracking allows image-based visual servoing approaches to work. Example methods which have been developed to work on SEM images rely on template matching [8], active contours [3], and rigid body models [9]. What is common in all these methods is that some measures have been included to diminish the effect of noise on the algorithm performance, either by use of inherent noise robustness like in the case of correlation-based template matching, the use of regions to cancel noise influence, or the usage of multiple or even manifold features and the selection and weighting of these features according to their expected significance and robustness.

Different algorithms may be suitable for different needs, depending on the degrees of freedom

and the boundary conditions imposed by the system setup and the object properties. If the objects of interest are rigid, their appearance constant and known, their expected movement restricted to translations, and no magnification changes are required, template matching using cross correlation has turned out to be a robust and powerful yet simple approach. The idea is to calculate the cross correlation of a template image calculate the cross correlation of a template image and an input image I:

$$C(x, y) = T(x, y) * I(x, y)$$

In the resulting matrix, each value corresponds to the similarity of the image part to the template. Assuming that the object of interest is in the image, the position is with very high probability at the position of the maximal value in this matrix:

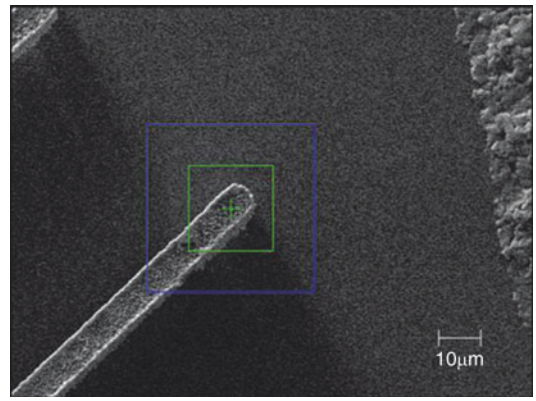
$$C(x_0, y_0) = \max(C(x, y))$$

The algorithm delivers a single position. It may well determine the position with subpixel accuracy, using weighting or interpolating techniques. The speed can be improved by using the frequency domain for calculation. In Fig. 3, the result of the algorithm can be seen. The template is marked at the place the best match is found.

If the objects may rotate or change in size, active contours or rigid body models are more suitable, depending on whether an a priori model of the object is available or not. Also, the occurrence of occlusions is less problematic with these algorithms. The principle behind active contours is that a contour represented by a polygon or spline is combined with an energy function which will determine its shape and position. The energy function consists of two parts:

$$J = J_{\text{int}} + J_{\text{ext}}$$

J_{int} is the internal energy which purely depends on contour properties like length, compactness, or smoothness. For compactness, the internal energy J_{int} can be formulated as:



Robot-Based Automation on the Nanoscale, Fig. 3 Tip of a microgripper tracked with correlation tracking

$$J_{\text{int}} = \frac{L^2}{A}$$

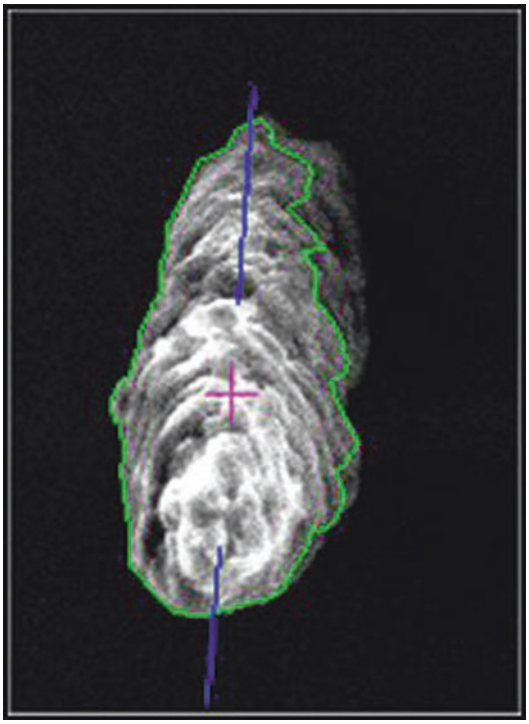
The second part is the external energy, which depends on the image content or image features. Possible implementations of this evaluate the distance to edges or the region statistics. The energy function is minimized for each frame by applying transformations to the contour. These transformations can be either single point translations or Euclidean transformations of the whole contour.

Figure 4 shows an active contour tracking an object in SEM images.

Using single point translations, deformable and nonrigid objects can be tracked, but occlusion is problematic. Using Euclidean transformations of the entire contour, the algorithm is robust against occlusion at the expense of trackability of deformable objects. Using these kinds of algorithms, full automation of nanohandling applications has been performed.

Three-Dimensional Position Determination

A difficult issue with image-based position determination using SEM images is the extraction of the third dimension. Any setup used for nanohandling and nanomanipulation is essentially a three-dimensional one, which means that in many cases, a two-dimensional position extracted from an image is insufficient for reliable and flexible automation. Several approaches exist to solve this problem:



Robot-Based Automation on the Nanoscale, Fig. 4 Object tracked with active contours in SEM images

- Restrict the setup or degrees of freedom to a plane, which is an unpractical solution.
- Use different image sources to enable depth estimation from multiple views. This includes the usage of additional microscope cameras to roughly estimate the information missing in the SEM image, as well as the use of beam deflection mechanisms to acquire SEM images from different angles [3].
- Use models of the SEM image generation and the lens distortions to extract the third dimension. This has been successfully done with smaller magnifications by [9].
- Use additional information from the images, such as defocus [20] or shadow to estimate the missing information or at least determine if the object or tool is at the desired location in space.

All these approaches have different advantages and disadvantages. Restricting the setup or degrees of freedom as mentioned has the advantage that no additional measures are needed to deal with the three-dimensional problem. This restricts the possibilities for handling and manipulation to an unacceptable extent. The use of different sources may deliver everything from rough estimates in case of microscope cameras to very precise (nm resolution) measurements in case of well-designed stereo imaging SEMs or SEM/FIB combinations. Drawbacks here are the reduction in speed when using stereo SEM and the complex correspondence problem in very dissimilar images when using a camera or the SEM/FIB combination. The use of models for the SEM imaging has produced quite good results, but depends strongly on the accuracy of the imaging model and on the accuracy of the rigid body model of the object which is tracked. This restricts the application to some extent to objects for which a model can be assumed, and to a certain parameter range of the SEM imaging. The use of additional information from the image mainly concentrates on defocus and shadow. Using defocus of objects can deliver an accuracy up to 5 μm using time-consuming focus sweeps, and it can also be used to estimate the out-of-focus displacement and therefore the position of any object for a given frame, but with reduced accuracy and reliability. This is partly due to the complex nature of the electron beam optics and the high number of factors influencing the image, especially image and object sharpness.

Position Control Inside SEM

A common challenge for the automation of handling and assembly tasks in the SEM is the precise alignment of two or more objects carried by different robots. Such objects can be wafers with micro- or nanostructured components as well as different tools including grippers, probes, and cantilevers. Dependent on the size of the involved objects, the accuracy of this alignment has to be in the micrometer or even nanometer range. Accuracy in this sense stands for the precision of the relative alignment, that is, a nanoobject needs to be brought precisely into the opening of a gripper. For automated positioning, a closed-loop controller is necessary.

Actuation Types and Motion Principles

To facilitate closed-loop motion control inside SEMs, actuators capable of generating the required resolution need to be designed. Although there are efforts to scale conventional actuators such as electromotors to enable such a resolution, for example, employing backlash-free gears, novel actuation principles are better suited. In general, an actuator capitalizes on a *physical effect* in order to create motion. Most of the currently employed actuators for nanohandling are based on the inverse piezoelectric effect, that is, the controlled deformation of a piezoelectric element using an electric field. Piezoelectric actuators are popular owing to their large forces, high precision, and fast response time. Such actuators can create forces in the kN-range and respond to an input signal within microseconds. Major downsides are the limited deformation of about 0.1 % as well as a strongly nonlinear behavior affected by hysteresis and creep. Other physical effects that are used include electrostatic, thermal, and magnetostrictive actuators as well as shape memory alloys [4].

Piezoelectric actuators can be further categorized into scanning and step-wise actuators. *Scanning actuators* use a backlash-free cinematic structure based on flexure hinges to amplify the relatively small displacements of the piezoelectric elements. Although commercial actuators can reach working ranges of up to 500 μm, the working range is a severe limitation and usually a second actuator is required for coarse positioning. Furthermore, the electric field used for actuation cannot be generated with arbitrary precision and without noise. Thus, there is a trade-off between working range and resolution. *Step-wise actuators* are a solution to overcome these downsides. Instead of amplifying the piezoelectric element's motion using a mechanical structure, the stroke is used to perform small steps. Step-wise actuators have a virtually unlimited working range. Most step-wise actuators are hybrid actuators, so that the piezoelectric element can also be used in a scanning mode to create movements smaller than single steps.

Scanning actuators have the advantage of generating a smooth motion over the entire working range. Furthermore, due to the high force that can be generated by a piezoelectric actuator, very high accelerations and speeds are possible. Another advantage is that a rough knowledge of the actuator's position can be directly derived from the applied voltage. Homing is not required for scanning actuators and even open-loop movements are reliable to a certain extent. A downside is that a scanning actuator always returns to its initial position on power loss. Thus, microrobotic systems have to be designed in a way that this movement cannot result in a severe damage to the employed robots, tools, or objects. Furthermore, the nonlinear characteristics including hysteresis and creep are scaled with respect to the full working range. Thus, these effects have a significant influence and make sensor feedback necessary.

In contrast, step-wise actuators have opposed properties. Their working range is only restricted by the mechanical guides, usually in the centimeter range. They require a sensor because the executed steps are not sufficiently repeatable. Furthermore, dependent on the sensor, homing is required to find the initial location of the actuators. In addition to the large working ranges, major advantages of step-wise actuators are that they do not perform a significant movement if the power is lost and that hysteresis and creep are scaled with the step length.

The most commonly applied step-wise motion principle based on piezoelectric actuators is the stick–slip actuation principle [4]. This principle makes use of the high movement speed and acceleration of a piezoelectric actuator. A mobile part is connected to a piezoelectric actuator mechanically using a contact with a specific friction. If the piezoelectric actuator is deformed slowly, the friction forces the mobile part to perform the same motion. If the actuator is deformed rapidly, however, the inertia of the mobile part keeps it in place and the friction force is overcome generating a net displacement between actuator and mobile part. Alternating slow and rapid deformations into opposite directions generate a step-wise motion over long travel ranges.

Vision-Based Closed-Loop Control

On the macroscale, the location of the manipulated objects in a world coordinate system is either

known from the process design or can be measured by a variety of available sensors. To perform closed-loop positioning, a robot can move to this location with sufficient accuracy using its internal sensors. On the nanoscale, the position of objects has to be derived by processing images of the employed SEM (see section "Real-Time Vision Feedback"). SEMs, however, are designed to create highly magnified images and not to create highly accurate measurements. Thus, the resulting position is often time variant and nonlinear, so that the object's location in a world coordinate system cannot be accurately derived. In addition, the positioning itself has limited accuracy due to the size ratio between the robotic system and the handled objects. Commonly, a robot with a size of a few centimeters is used to manipulate an object with a precision of a few nanometers, that is, the robot is about six orders of magnitude larger than the required accuracy. Effects such as thermally- induced drift and limited mechanical stiffness lead to a nonlinear and time-variant relation between the position measured by a robot's internal sensors and the real position of the robot's tool. Thus, internal sensors are insufficient for reliable nanopositioning.

As known from the macroworld, the limitations of positioning based on internal sensors can be overcome by using the information generated by a visual sensor. Using vision feedback as a sensor for robotic manipulation is promising as it is a noncontact measurement and mimics human behavior. The initial technique, which is still used for most SEM-based positioning, extracts a position deviation from an image and then moves the robot either open loop or based on internal sensors to decrease this deviation. When the robot stops, another image is taken, again extracting and compensating for the deviation. This is called the *look-then-move* approach [10]. Two advantages of the approach are that low throughput rates of image processing are acceptable and that the entire setup can be handled in a quasi-static fashion neglecting dynamic aspects. The effectiveness, however, entirely depends of the accuracy of the visual sensor and the robot's open-loop control.

An alternative is to use the visual feedback as a low-level element in the motion control loop of a robot. This is called *visual servoing* [10]. Two challenges need to be solved to achieve effective visual servoing. First, the image processing needs to be fast. As the available processing power has significantly increased over the recent years, several image processing algorithms can be used for visual servoing. More complex algorithms, however, still require computation times on the order of seconds and do not allow for effective servoing. Second, the dynamic aspects of visual servoing need to be considered. The robot's movement as well as resulting vibrations leads to movement during image acquisition. This movement creates image distortions rendering most of the current image processing algorithms useless. Enabling an effective visual servoing, however, is a necessary step on the way of high-speed micro- and nanohandling viable for industrial application.

Closed-Loop Control

Based either on an internal sensor or on vision feedback, a closed-loop controller can be implemented. Closed-loop control on the micro- and nanoscale is significantly different from macroscale control [1]. Macroscale control mostly solves challenges arising from the dynamic behavior of systems, for example, inertia. On the micro- and nanoscale, these effects become negligible. As an example, a microrobot might carry a mass m of 30 g and can create a force F of approximately 0.02 N. Thus, its maximum acceleration a is:

$$a = F/m = 2/3\left[\mathrm{m/s^2}\right]$$

If it needs to accelerate to a velocity v of 500 μm/s, which is high for a micro- or nanooperation, it requires:

$$t = v/a \ = 0.375\ [\mathrm{ms}]$$

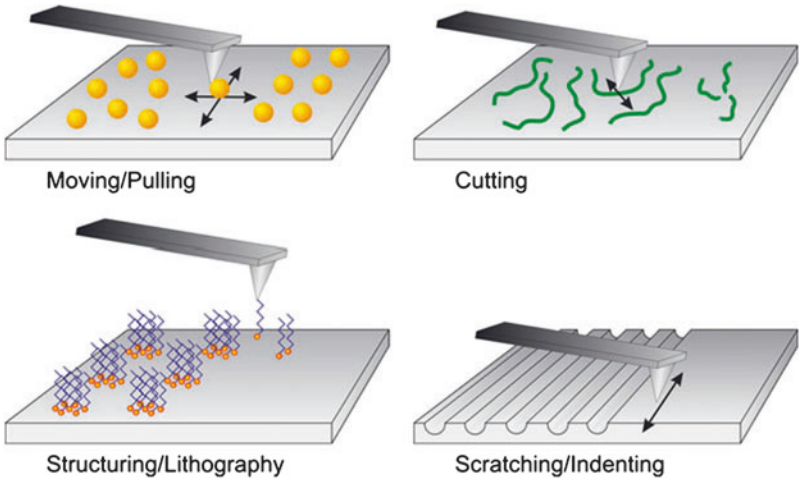
This is significantly faster than the update rate of any visual sensor. Thus, acceleration effects can be neglected. Therefore, for many applications and actuator types, a simple proportional controller is sufficient to perform an effective alignment.

AFM as a Nanohandling Robot

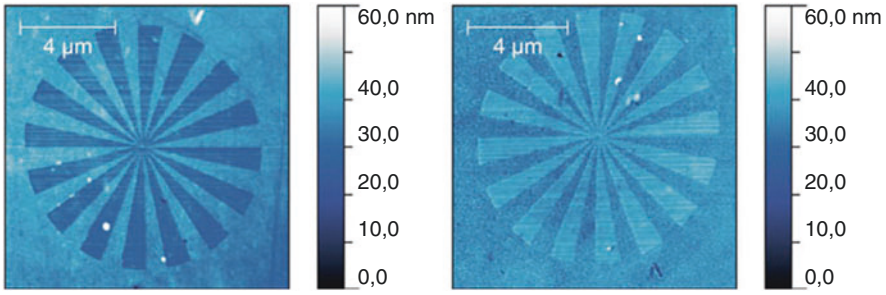
Since its development by Binnig, Quate, and Gerber [11], AFM has been increasingly recognized not only as a tool for imaging purposes, but also for manipulating at the micro-, nano-, and atomic scale. The main advantages of the AFM are its high resolution in the sub-nanometer range as well as its flexibility for operating under physiological conditions and in vacuum or liquid environments, depending on the specific application. In addition to its visualization capabilities, the AFM can be used for force spectroscopy and, with functionalized tips, for measuring electric, magnetic, or chemical properties. Moreover, the AFM tip can be used to manipulate objects or materials with nanometric precision and thus the AFM can also be seen as a nanohandling robot. AFM-based manipulation benefits from the fact that the AFM directly measures forces that are exerted during the manipulation and thus gives important feedback on the progress of the manipulation. Potential applications of AFM-based manipulations are manifold, ranging from the fabrication of nanoscale devices, reparation or modification of nanostructures to the characterization and handling of biological samples. Different types of manipulation can be conducted with the AFM including pushing, pulling, cutting, indenting, and structuring of a wide range of materials (Fig. 5).

Starting with unintended manipulations of Au-Clusters, reported 1987 by Baró et al. in [12], the AFM as a tool for manipulating single nanoentities became subject of increasing interest in the research community. A controlled manipulation of nanoscale objects by an AFM has been actively investigated over the last decade. Many publications exist describing the utilization of the AFM as an instrument for manipulating spherical nanoobjects such as gold nanoparticles as well as longish nanoobjects such as carbon nanotubes (CNTs), nanorods, or also biological objects such as DNA [2]. Figure 6 illustrates a successful manipulation of multiwalled CNTs that was conducted by the AFM tip. A commercially available dynamic-mode cantilever was used for both – the image acquisition as well as the manipulation. For imaging, conventional intermittent mode was used, whereas for manipulation the tip was brought into proximity with the surface and then moved laterally to mobilize the CNTs.

To perform a controlled movement of a nanoscopic object, it has to be localized and identified on the substrate first. A major drawback of AFM-based manipulation is the lack of an additional, visual sensor. Therefore, before any manipulation is conducted, the AFM usually has to be utilized to obtain a detailed image of the scenery in order to plan the execution of the manipulation. This imaging scan is most often performed in dynamic mode (intermittent or noncontact) in order to avoid any damage or unintended manipulation of the sample. For the manipulation, the AFM tip is brought into proximity with the object and moved – either in dynamic or contact mode, but without any feedback to control the z-direction – toward a predefined position. As a result, the object is then pushed forward by the repulsive forces. After a manipulation step, reimaging of the area of interest often becomes necessary to reveal the result of the manipulation. The disadvantage of this "look and manipulate" scheme is that the image acquisition process usually needs several minutes depending on the required quality. Moreover, the manipulation itself cannot be monitored by means other than the measured forces exerted by the cantilever and thus have to be performed in a "blind" way. Unfortunately, due to a variety of uncertainties, the results of the manipulation are often not satisfying and require frequent trial and error experiments. Real-time imaging would greatly help to make the manipulation process faster, more robust, and accurate. However, optical microscopy cannot be used as an additional real-time capable imaging method because its resolution does not fit the requirements for the manipulation of objects with nanometric dimensions. An alternative is given by the integration of the AFM into the vacuum chamber of an SEM to allow for real-time observation of the manipulated object and the AFM probe. However, this entails the disadvantage of restriction to vacuum and to some extent conductive sample materials, which is not feasible for many applications.



Robot-Based Automation on the Nanoscale, Fig. 5 Schematic view of manipulation tasks that can be performed by AFM-based robotic systems [1, 2]



Robot-Based Automation on the Nanoscale, Fig. 6 AFM topography star with diameter of 10 µm, fabricated on a gold layer (*left*) using mechanical AFM-lithography and the replica of this structure fabricated on polypropylene (*PP*) using the microinjection molding technique (*right*)

Besides its ability to manipulate nanoscale objects, the AFM tip can also be used to modify surfaces with nanometric precision or to change an object's shape, for example, by scratching, indenting, cutting, and dissecting. For nanomachining purposes, the AFM tip can be exploited as a nanohandling tool in order to act as, for example, a milling cutter, nanoscalpel, or nanoindenter. However, the type of interaction between AFM tip and substrate is not only mechanical, but may also involve chemical, thermal, or electrical impact. Over the last two decades, a large number of promising nanolithographic technologies have evolved and are still emerging that are based on such AFM-based nanomanipulation of surfaces. These techniques can be used among others for mask-free lithography on the nanoscale and provide a promising alternative to conventional technologies. Using AFM-based surface modification, material can either be removed (e.g., through direct mechanical scratching), chemically changed (e.g., resist exposure techniques or oxidation), or deposited (e.g., dip-pen lithography). A detailed review of such SPM-based lithography can be found [13]. An example for a novel process chain for small series production of nanostructures that is based on AFM-based mechanical lithography, which is the most basic type of AFM lithography, was introduced [14]. Figure 7 depicts how the mechanical AFM-lithography is utilized to fabricate a nanostructure that was subsequently replicated several times by means of the microinjection molding technique.

Up to now, the AFM as a robot for the nanomanipulation can only be found in



Robot-Based Automation on the Nanoscale, Fig. 7 Multiwalled carbon nanotubes on HOPG substrate manipulated by the AFM. When moving the AFM tip laterally, the height control loop was turned off in order to avoid slipping of the AFM tip over the CNT. It could be observed that, depending on the parameters chosen for manipulation (tip velocity, damping amplitude), cutting as well as pushing and pulling of the CNTs can be accomplished

laboratories for prototypical issues. For industry, the serial AFM-based approach is still not reliable and fast enough to be successfully applied for productive use. To increase throughput and reliability of this sequential process, automation can be considered as a desirable goal. To reach this goal, however, a large number of challenging prerequisites have to be fulfilled.

Besides the lack of real-time imaging of the manipulation process discussed above, spatial uncertainties constitute an import error source. Although positioning inaccuracies caused by creep and hysteresis of piezoactuated scanning stages can nowadays most often be compensated by closed-loop operation, thermal drift of the AFM components lead to spatial positioning errors that are challenging to counteract. Solutions to overcome this problem have been addressed [1, 2, 15]. Moreover, sticking effects also play a crucial role in AFM-based nanohandling and therefore, substrate, cantilever, and tip material have to be carefully chosen in order to minimize these effects. A promising strategy that is pursued by several research groups aims at the precise modeling of the nanohandling task. To compensate for the uncertainty introduced by the inevitable "blind" manipulation, a model of the manipulation process is utilized enabling a manipulation in open-loop mode. Having a valid model of the nanomanipulation, including all relevant interactions between nanoobject, tip, and substrate, it is possible to calculate the expected position of the nanoobject during manipulation in real time. Using such a model for an augmented reality system, where a virtual reality is augmented by the calculated nanoworld situation, enables an additional visual feedback. However, this approach requires exact knowledge of the nanomanipulation phenomena, which is not yet available in the current state of nanosciences.

Application Example: Handling of CNT

Nanomaterials such as different kinds of nanowires and especially carbon nanotubes (CNTs) [16] have unique physical properties and can thus improve and enhance existing microsystems in a multitude of applications. Two main application areas have been identified where in particular the integration of CNTs into known microsystems is of high impact: CNT-enhanced AFM probes [17] for the analysis of high-aspect ratio structures and CNT-interconnects in microchip technology [18]. Even though parallel fabrication techniques will be required for a future bulk production of such devices, enabling technologies for the reliable manipulation of individual nanostructures and their prototypic integration into microsystems have to be developed. Automated robot-based assembly techniques are very promising to implement a fast

and systematic prototyping of CNT-based devices. In the following section, a microgripper-based strategy for the automated assembly of CNT-enhanced AFM probes is presented.

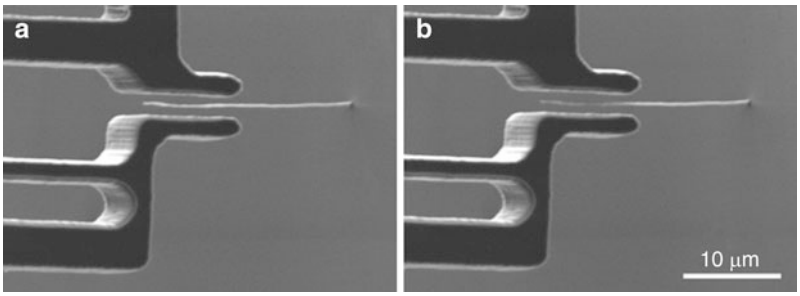
Nanorobotic Strategies for Microgripper-Based Handling of CNTs

In prior experiments, electrothermally actuated microgrippers have already been used for teleoperated pick-and-place handling of individual CNTs [19]. The CNTs used for the experiments have a typical diameter of 200 nm and are grown vertically aligned on a substrate with a spacing of 50 μm between nanotubes. In order to automate the microgripper-based handling of CNTs, special nanorobotic strategies have been developed. Most important are methods for the z-approach of gripper and CNT and for the reliable gripping and placing of CNTs that lead to a well-defined system state.

Methods for z-Approach of Microgripper and CNT An AMNS has been developed that can be integrated into the SEM [2], so that visual feedback from the microscope is available during the nanohandling tasks in addition to the integrated sensors of the positioning systems. Image processing algorithms for object recognition and tracking have been developed (compare section "Real-Time Vision Feedback") providing the x, y-position of gripper and nanotube. This two-dimensional position information can be used to automate the approach in x,y-direction. However, to receive information on the gripper's and CNT's relative z-position, additional contact sensors or three-dimensional systems are required. Such systems cause additional costs and reduce the available space within the SEM's vacuum chamber [20].

A novel so-called shadow-based depth detection method has been developed that, together with the focus information of the SEM, facilitates the automated z-approach of microgripper and CNT in any available SEM. First of all, the "depth-from-focus" technique [20] is used to coarsely align both objects in z-direction down to 50 μm. Then, for the fine approach of gripper and CNT, a shadow can be detected as soon as the CNT is located exactly between the gripper jaws. The gripper jaws block the secondary electrons from the detector. Figure 8 shows SEM images illustrating the shadow-based depth detection method. The clearly detectable shadow on the CNT in Fig. 8b is used as a stop criterion for the z-approach.

Gripping and Removing CNTs from the Growth Substrate Different gripping techniques for the reliable picking and removing of individual CNTs that are grown on a silicon substrate by chemical vapor deposition (CVD) have been discussed in [21]. The so-called shear gripping technique has been identified as the most reliable approach, requiring smaller gripping forces compared to tensile gripping. The microgripper is closed around the nanotube and



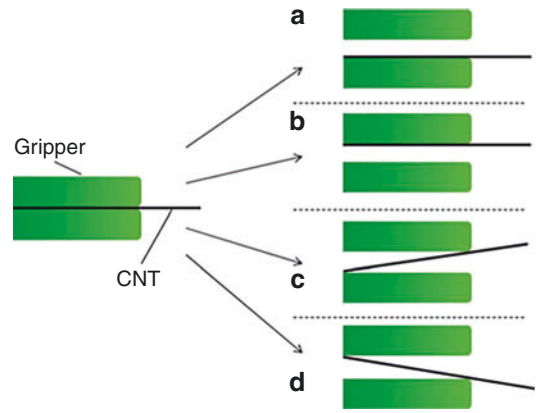
Robot-Based Automation on the Nanoscale, Fig. 8 Shadow-based depth detection for the alignment of microgripper and CNT in z-direction. (**a**) CNT is located below the gripper jaws, no shadow is detectable. (**b**) CNT is between the gripper jaws, a shadow is clearly visible

moved sideways to break the CNT off the substrate. After that, the gripped CNT can be transferred to the target structure.

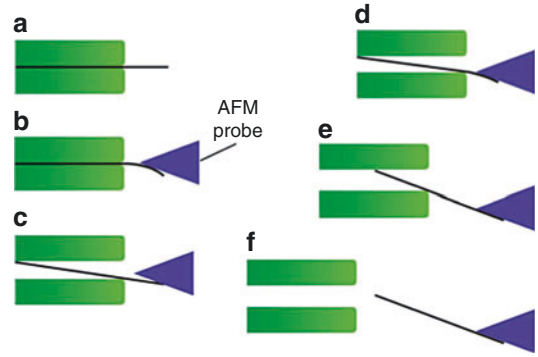
Placing and Releasing CNTs on the Target Structure A novel handling strategy for placing and releasing CNTs onto the target structure has been developed. This strategy relies on the adhesion forces between CNT and microgripper jaws and thus avoids the need of additional joining techniques. After reopening the microgripper, the nanotube can be oriented in four different configurations. Either the CNT is connected by a line contact to one gripper jaw (see Fig. 9a, b) or the CNT is connected by two-point contacts to the gripper jaws (see Fig. 9c, d).

In case of the two-point contacts, the adhesion forces are smaller compared to the line contacts. For this reason, the two-point contact is preferred in order to easily place and release the CNT on the target structure. In addition, obtaining a well-defined system state after reopening the gripper is required to allow full automation of the placing and releasing process. For the automated microgripper-based mounting of an individual CNT onto the tip of an AFM probe, the strategy shown in Fig. 10 has been developed.

The closed microgripper holding the CNT (Fig. 10a) is approached to the tip of an AFM probe until a bending of the nanotube can be detected (Fig. 10b). This bending is detected automatically and used as the stop criterion of the z-approach. At the same time, the exerted force assures that CNT and gripper jaws are forming a two-point contact with defined orientation after reopening the microgripper (Fig. 10c). After this, the CNT is moved toward the AFM tip until a line contact is achieved (Fig. 10d). Because the adhesion forces of the line contact between CNT and AFM tip are much higher than the forces of the two-point contacts between CNT and gripper jaws, the nanotube can be easily released from the gripper jaws by retracting the microgripper (Fig. 10e–f).



Robot-Based Automation on the Nanoscale, Fig. 9 The four possible configurations after reopening the microgripper. (**a**, **b**) Line contact between CNT and gripper jaw. (**c**, **d**) Two-point contact between CNT and gripper jaws



Robot-Based Automation on the Nanoscale, Fig. 10 Nanorobotic handling strategy for automated placing a CNT onto the tip of an AFM probe and releasing it from the microgripper

Automated Assembly of CNT-Enhanced AFM Probes

The position control techniques described in section "Position Control Inside SEM" have been used to fully automate the nanohandling task of mounting a CNT onto the tip of an AFM probe. For this purpose, the whole automation sequence has been divided into three subsequences: The initialization process, the gripping and removing process, and the process of placing and releasing the CNT. In the following, these three subsequences are described in detail with respect to automation purposes.

Initialization Sequence The first step of an automated robotic micro-nano-integration is the

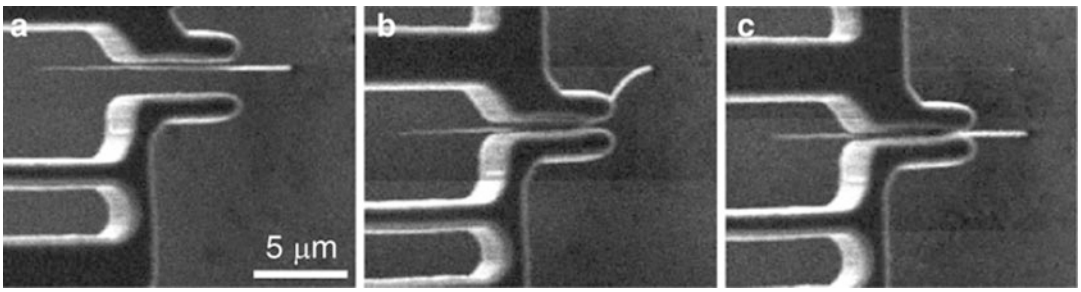
initialization. For this purpose, all actuators are moved to a well-known starting position and the coordinate system of the internal position sensors are mapped to the coordinates of the SEM image. Then, the electrothermal microgripper mounted to the mobile microrobots is calibrated. A pattern of the microgripper is recorded that allows for tracking the position of the gripper by applying cross-correlation algorithms to the SEM images (compare section "Real-Time Vision Feedback"). After this, the CNT substrate is scanned automatically to index all CNTs. For each CNT, length and orientation are checked as a quality criterion and only the positions of suitable CNTs are stored to the database. Finally, the AFM probes are indexed by saving the exact position of the probes' tips to the database.

Picking Sequence As described in section "Gripping and Removing CNTs from the Growth Substrate", the shear gripping technique is used to grip and remove an individual CNT from the sample substrate. Figure 11 shows SEM images of the gripping sequence illustrating the process of breaking the CNT off the substrate. First, so-called coarse and fine positioning steps are performed placing the CNT between the gripper jaws (see Fig. 11a). Second, the microgripper is actuated and the CNT is detached by moving the gripper sideways (see Fig. 11b). Third, the gripped CNT can be transferred to the target structure.

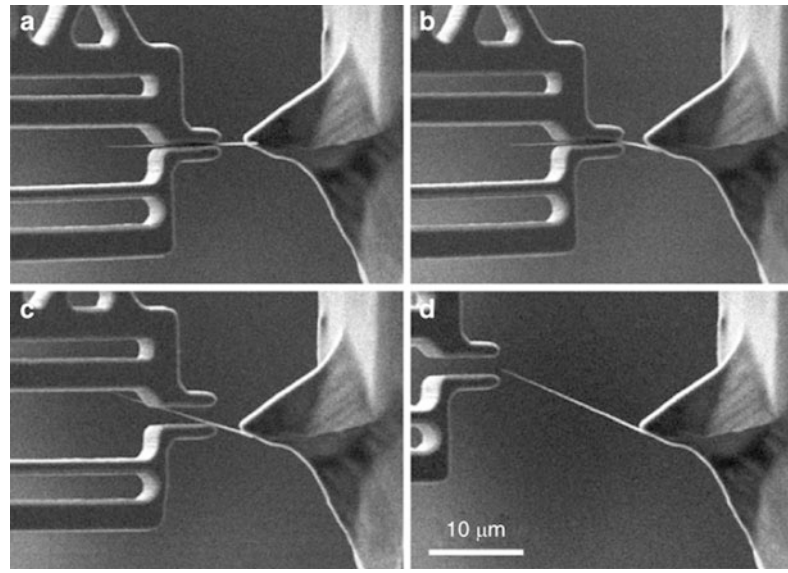
Placing Sequence For the automated placing of the CNT onto the tip of an AFM probe, the nanorobotic strategy described in section "Placing and Releasing CNTs on the Target Structure" is used. Figure 12 shows SEM images of the automated placing sequence. After the coarse and fine approach between CNT and AFM probe has been realized in x,y-direction (see Fig. 12a), the final z-approach is performed until a bending of the CNT is detected (see Fig. 12b). Then, the gripper is reopened and a two-point contact between CNT and gripper jaws as well as a line contact between CNT and AFM probe tip is established (see Fig. 12c). Finally, the CNT can be easily released from the gripper jaws and placed onto the tip of the AFM probe by retracting the microgripper (see Fig. 12d).

Application Example: Handling of Biological Materials

Current research areas in molecular and cell biology, medicine, and process sensor technology (sensor systems monitoring various processes in industry and science) often require advances in the nanoengineering technologies to look for molecular phenomena with the highest possible resolution in order to enlighten the "black-box" which still shades most metabolic reactions. The aim of present work, starting from current medical problems, is to propose novel studies especially in molecular and cellular microbiology as well as virology with direct applications in medicine and industry, closing the bottom-up cycle. Thanks to advances in micro- and nanofabrication and micro- and nanorobotics, robotic systems which offer the possibility of characterizing and



Robot-Based Automation on the Nanoscale, Fig. 11 SEM images of the automated gripping sequence. (**a**) Coarse and fine alignment of gripper of CNT. (**b**) Removing the CNT from the substrate by shear gripping. (**c**) Closed gripper holding the detached CNT



Robot-Based Automation on the Nanoscale, Fig. 12 SEM images of the automated assembly. (**a**) Coarse and fine alignment of gripped CNT and AFM tip. (**b**) Final z-approach until a bending of the CNT is detected. (**c**) Creating a two-point contact between CNT and gripper jaws and a line contact between CNT and AFM probe. (**d**) Releasing the CNT by retracting the gripper

manipulating biological objects can be developed today. The use of such systems enable new studies of single cell phenomena with nanometer resolution (e.g., studying the local mechanical and electrical properties of single bacteria for characterization and evaluation of the resistance to antibiotics or to the spread of infections) up to complete cell compounds (e.g., studying the mechanical properties of bacterial biofilms) and deepen the appreciation of the processes at the nanoscale. The following section will give a brief overview of the handling of biological objects using automated AFM-based approaches.

Manipulation and Handling of Biomaterials Using Atomic Force Microscopy

Section "AFM as a Nanohandling Robot" has already shown the basics of the AFM for the automated handling of nanoscale materials. In this subsection, a brief overview of the differences for biomaterials is given.

For characterization or manipulation, samples have to be immobilized on an adequate substrate. A common method is the immobilization on freshly cleaved mica. Mica has a hydrophilic, charged surface that binds proteins and other biomolecules easily. Occasionally, mica is treated with buffer systems, which change the charging of the surface. Gold substrates can also be used if coated with protein reactive monolayers.

The characteristics of the used cantilevers are determined by material, tip shape, and mechanical parameters such as spring constant, resonant frequency, and Q factor. For biological applications, cantilevers are usually made of silicon or silicon nitride. These materials are chemically inert and can be doped to induce electrical charges. Soft cantilevers ($C < 1$ N/m) are used for the contact mode and hard cantilevers ($C \approx 42$ N/m) are used for the tapping mode. If operated in liquids, the pH value and the presence of electrolytes influence the forces between tip and sample. Although a smaller tip radius results in more accurate images, it may be advantageous for soft biological samples, to use a duller tip reducing the risk of accidentally damaging the sample.

Compared to other methods (e.g., optical tweezers, dielectrophoresis) the AFM has some major advantages for the handling of biomaterials. The most important points are [1, 2, 22]:

- The AFM can be used in liquids, in vacuum, and at ambient conditions meaning that biological samples can be characterized and manipulated at their necessary physiological conditions. Furthermore, living samples and

their reaction to mechanical stimulation can be studied.
- The AFM can be used for imaging as well as for manipulation nearly at the same time.
- The possible resolution lies in the sub-nanometer range.
- Three-dimensional topological information can be obtained.
- No special preparation methods are necessary.

However, compared to the other methods AFM-based manipulation entails major drawbacks and limitations. Knowing these limitations is important to choose the right method or tool to handle or characterize a specific sample. The most important limitations are:

- A three-dimensional force measurement is not yet possible.
- The scanning speed is generally too low for real-time mapping.
- The scanning range for high-speed AFMs is still very low (about 3×3 µm).

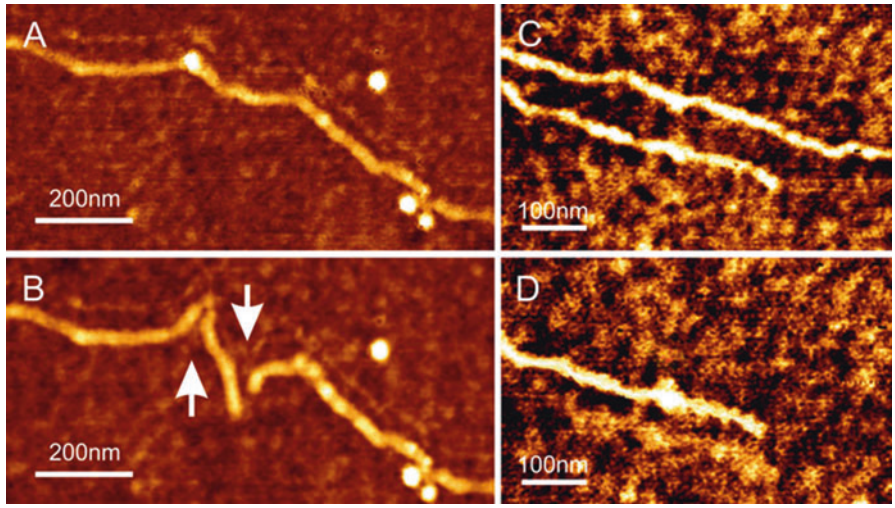
Examples of Automated Handling of Biomaterials with an AFM

In the following sections, current work on the field of handling, characterization, and manipulation of biomaterials is described. Current research mainly deals with the development of fully automated solutions for handling and manipulation of DNA, for solving of packaging problems in the future of nanoelectronics, and for structuring of biosensors for medical and forensic use.

Handling and Manipulation of DNA In accordance with the progressing miniaturization of electric circuits, new ways to realize smaller channel widths are needed. To overcome the limitations of silicon-based processing, the use of objects like CNTs or biomolecules like DNA is a promising solution. Nanowires produced via the metallization of DNA are promising candidates for nanoelectronic devices of future generations. DNA can be considered a basic building block for nanostructure fabrication because it provides unique self-recognition properties. DNA constitutes an ideal template for the organization of metallic and semiconductor particles into wire-like assemblies. Nanoscopic (thickness below 10 nm) and regular wire-like metallic structures can be efficiently produced because DNA molecules have a diameter of only two nanometers and a length in the micrometer range. Moreover, both the ends of DNA molecules and the surface of solid substrates can be functionalized through a variety of methods to create specific links between DNA and substrate, allowing for the integration of DNA molecules into specific, predefined sites of microstructured electronic circuits. By exploiting the specific recognition of complementary nucleotide sequences, complex structures made of DNA can be fabricated and metalized at a later stage. Packaging problems of next-generation nanoelectronics technologies, (e.g., single-electron transistors, quantum automata, molecular electronics, etc.) might also be addressed by using DNA strands.

Recently, many researchers have proven the feasibility of DNA manipulation by an AFM tip, primarily in liquid conditions. The next step toward high-throughput fabrication is automation of DNA handling. Figure 13 show two alternative immobilization procedures of DNA, which offer the potential of automated DNA manipulation in dry ambient conditions. Figure 13a shows a topographic scan of a DNA strand. The height of the DNA strand is about 1.5 nm. Figure 13b shows the same DNA strand after several manipulations by the AFM cantilever tip. The DNA was manipulated and successfully moved about 25–100 nm. Figure 13c, d shows an experiment to demonstrate the transport of DNA. Two separate DNA strands are shown. The upper DNA strand (Fig. 13c) was moved by several AFM strokes up to 300 nm. As a result, this DNA strand was removed from its location (Fig. 13d), without affecting the position of the lower DNA.

The presented methods for the handling of biomaterials show the potential for different usage scenarios in industry and scientific research. In case of the handling of DNA, the experiments demonstrate that a DNA manipulation and handling can be performed in dry ambient conditions.



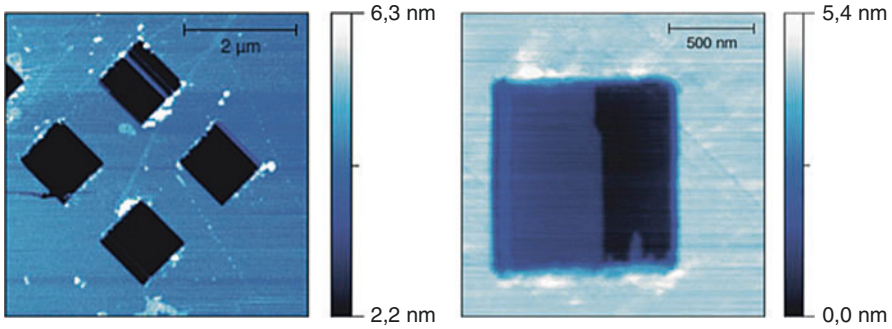
Robot-Based Automation on the Nanoscale, Fig. 13 (**a**, **b**) DNA immobilized on silicon without surface modifications (**a**) and the same DNA after the manipulation process (**b**). On the *left* side of the strand, two short distance movements from *bottom* to *top* were implemented. As a result, the corresponding DNA parts have been pushed about 25 nm. On the *right* side, two long distance movements were implemented, resulting in DNA displacement by 100 nm. (**c**, **d**) DNAs immobilized on silicon without surface modifications (**a**) and the same substrate with *upper* DNA strand pushed away on the surface by the AFM tip, without affecting the *lower* strand (**b**)

Automated Structuring of Biological Materials for Biosensors Biosensors consist of a sensitive biological component, a transducer, or detector element, and associated electronics or signal processors. An accurate and well-defined preparation of the sensitive components made, for example, of microorganisms, cell receptors, proteins, enzymes, anti-bodies, or nucleic acids is crucial part of the sensor design. Commonly used methods are microstamping and microprinting, both of which are unable to decrease the structures of the biological component down to the nanometer range. However, higher density is necessary to increase the measurement accuracy and the number of different, simultaneously useable biocomponents. In this section, current work on automated design of high-density sensitive components using AFM-based lithography is introduced.

The compensation of spatial uncertainties plays an important role in automation of AFM-based handling. Although the effects of hysteresis and creep of the PZT-based scanner can be reduced by closed-loop control, spatial uncertainties caused by thermal drift are crucial and less straightforward to deal with. Especially in life science applications, where the AFM is often operated in humid environments, the effect becomes strong and often prevents successful nanohandling, let alone automation of the handling task. To achieve reliable results, the recently developed method for real-time drift compensation described in section "AFM as a Nanohandling Robot" and in [1, 15] was applied, both for the presented manipulation of DNA as well as for the AFM-based nanotooling described below.

As a first step toward the automated fabrication of biosensor components, rectangular areas were successfully structured into an APTES monolayer on mica (Fig. 14). To further investigate the correlation between the forces applied during manipulation and the resulting machining depths, a series of experiments was performed to scratch multiple areas.

By using the AFM, an automated structuring of biological materials for biosensors can be done.



Robot-Based Automation on the Nanoscale, Fig. 14 Height image of 1 × 1 µm area scratched into APTES monolayer by AFM machining. The two visible machining depths arise from different force set points used during the processing (*left*) and AFM height image of a series of 1 × 1 µm sized scratched areas (*right*)

The automation of the necessary processes will also lead to a feasible industrial use, as well as for clinical and research laboratories. The size of the resulting structures correlates with the resolution of the commonly used read-out techniques such as fluorescence microscopy or surface plasmon resonance spectroscopy.

Future Trends

In the future, closed-loop motion control in SEMs needs to be changed from the quasi-static look-then-move-approach to an effective visual servoing in order to implement high-speed micro- and nanohandling viable for industrial application. Furthermore, to allow for a more effective avoidance of both collisions and vibrations, the commonly implemented position control needs to be extended to trajectory control so that the robot moves with a predetermined velocity, acceleration and jerk at all times. Considering visual feedback, an important future task is to enable more reliable and faster feedback especially in case of fast moving objects, because of the artifacts and distortions generated by the object movement. Also speed, accuracy, and resolution of the z-position estimation still is a topic where significant improvements can be made.

The AFM has emerged as an important tool for the fabrication of nanoscale structures, as well as the handling and manipulation of biological objects and for the manipulation of nanoscale objects. To broaden the applicability of those techniques for productive use, it will become necessary to increase throughput significantly. For the lithographic techniques used to structure surfaces, the usage of multiple AFM tips in parallel will greatly enhance throughput. To speed up the assembly of individual nanoentities, achieving a high level of automation is an important research goal for the future. Further miniaturization of gripping tools can facilitate the nanorobotic pick-and-place handling of CNTs with smaller diameters. The optimization of the presented automated assembly sequence of CNT-enhanced AFM probes will lead to a high-throughput prototyping of CNT-based devices, which is an important pre-stage toward a direct integration of CNTs into microstructures by CMOS-compatible CVD-based fabrication techniques.

Cross-References

- ▶ AFM
- ▶ AFM in Liquids
- ▶ AFM Probes
- ▶ Atomic Force Microscopy
- ▶ Basic MEMS Actuators
- ▶ BioPatterning
- ▶ Biosensors

- Carbon Nanotubes for Chip Interconnections
- Carbon-Nanotubes
- Chemical Vapor Deposition (CVD)
- Dielectrophoresis
- Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes
- Dip-Pen Nanolithography
- DNA Manipulation Based on Nanotweezers
- Electric-Field-Assisted Deterministic Nanowire Assembly
- Electron-Beam-Induced Deposition
- Manipulating
- Microfabricated Probe Technology
- Nanogrippers
- Nanoparticles
- Nanorobotic Assembly
- Nanorobotic Manipulation of Biological Cells
- Nanorobotic Spot Welding
- Nanorobotics
- Nanorobotics for Bioengineering
- Nanorobotics for NEMS Using Helical Nanostructures
- Nanostructured Functionalized Surfaces
- Nanostructured Materials for Sensing
- Optical Tweezers
- Organic Bioelectronics
- Piezoelectric Effect at Nanoscale
- Piezoelectric MEMS Switch
- Scanning Electron Microscopy
- Scanning Tunneling Microscopy
- Self-Assembled Monolayers
- Self-Assembly
- Self-Assembly for Heterogeneous Integration of Microsystems
- Self-Assembly of Nanostructures
- Thermal Actuators
- Transmission Electron Microscopy

References

1. Sattler, K.D.: Handbook of Nanophysics – Nanomedicine and Nanorobotics. Taylor & Francis/CRC Press, Boca Raton (2010)
2. Fatikow, S., Wich, T., Dahmen, C., Jasper, D., Stolle, C., Eichhorn, V., Hagemann, S., Weigel-Jech, M.: Nanohandling robot cells. In: Sattler, K.D. (ed.) Handbook of Nanophysics – Nanomedicine and Nanorobotics, pp. 47.1–47.31. Taylor & Francis/CRC Press, Boca Raton (2010)
3. Fatikow, S. (ed.): Automated Nanohandling by Microrobots. Springer, London (2008)
4. Fatikow, S., Rembold, U.: Microsystem Technology and Microrobotics. Springer, Berlin/Heidelberg/New York (1997)
5. Fatikow, S.: Mikroroboter und Mikromontage (in german). B.G. Teubner, Stuttgart/Leipzig (2000)
6. Ferreira, A., Mavroidis, C.: Virtual reality and haptics for nanorobotics. IEEE Robot. Autom. Mag. **13**, 78–92 (2006)
7. Fatikow, S., Dahmen, C., Wortmann, T., Tunnell, R.: Visual feedback methods for nanohandling automation. Int. J. Inf. Acquis. **6**, 159–169 (2009)
8. Sievers, T., Fatikow, S.: Real-time object tracking for the robot-based nanohandling in a scanning electron microscope. J. Micromechatron. **3**(3–4), 267–284 (2006). Special Issue on Micro/Nanohandling
9. Kratochvil, B.E., Dong, L.X., Nelson, B.J.: Real-time rigid-body visual tracking in a scanning electron microscope. Int. J. Robot. Res. **28**, 498–511 (2009)
10. Hutchinson, S., Hager, G.D., Corke, P.I.: A tutorial on visual servo control. IEEE Trans. Robot. Autom. **12**, 651–670 (1996)
11. Binnig, G., Quate, C.F., Gerber, C.: Atomic force microscope. Phys. Rev. Lett. **56**, 930–933 (1986)
12. Baró, A.M., Bartolome, A., Vazquez, L., García, N., Reifenberger, R., Choi, E., Andres, R.P.: Direct imaging of 13-Å-diam Au clusters using scanning tunneling microscopy. Appl. Phys. Lett. **51**, 1594–1596 (1987)
13. Tseng, A.A., Notargiacomo, A., Chen, T.P.: Nanofabrication by scanning probe microscope lithography: a review. J. Vac. Sci. Technol. B **23**, 877–894 (2005)
14. Brousseau, E., Krohs, F., Caillaud, E., Dimov, S., Gibaru, O., Fatikow, S.: Development of a novel process chain based on atomic force microscopy scratching for small and medium series production of polymer nanostructured components. ASME J. Manuf. Sci. Eng. **132**, 030901 (2010)
15. Krohs, F., Onal, C., Sitti, M., Fatikow, S.: Towards automated nanoassembly with the atomic force microscope: a versatile drift compensation procedure. ASME J. Dyn. Syst. Meas. Control **131**, 061106 (2009)
16. Iijima, S.: Helical microtubules of graphitic carbon. Nature **354**, 56–58 (1991)
17. Dai, H., Hafner, J.H., Rinzler, A.G., Colbert, D.T., Smalley, R.E.: Nanotubes as nanoprobes in scanning probe microscopy. Nature **384**, 147–150 (1996)
18. Srivastava, N., Li, H., Kreupl, F., Banerjee, K.: On the applicability of single-walled carbon nanotubes as VLSI interconnects. IEEE Trans. Nanotechnol. **8**, 542–559 (2009)
19. Sardan, O., Eichhorn, V., Petersen, D.H., Fatikow, S., Sigmund, O., Bøggild, P.: Rapid prototyping of nanotube-based devices using topology-optimized microgrippers. Nanotechnology **19**, 495503 (2008)

20. Eichhorn, V., Fatikow, S., Wich, T., Dahmen, C., Sievers, T., Andersen, K.N., Carlson, K., Bøggild, P.: Depth-detection methods for microgripper based CNT manipulation in a scanning electron microscope. J. Micro-Nano Mechatron. **4**, 27–36 (2008)
21. Andersen, K.N., Petersen, D.H., Carlson, K., Mølhave, K., Sardan, O., Horsewell, A., Eichhorn, V., Fatikow, S., Bøggild, P.: Multimodal electrothermal silicon microgrippers for nanotube manipulation. IEEE Trans. Nanotechnol. **8**, 76–85 (2009)
22. Castillo, J., Dimaki, M., Svendsen, W.E.: Manipulation of biological samples using micro and nano techniques. Integr. Biol. **1**, 30–42 (2009)

Robotic Insects

► Insect Flight and Micro Air Vehicles (MAVs)

Robotic Platforms for Nanomanipulations

► Nanorobotics for Bioengineering and Medical Interventions

Role of Cavitation Bubble Dynamics on Nanoparticle Generation

► Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids

Role of Plasma Confinement on Nanoparticle Generation

► Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids

Rolled-Up Nanostructure

► Nanorobotics for NEMS Using Helical Nanostructures

Rolled-Up Nanotechnology

► Electrical Impedance Tomography for Single-Cell Imaging

Rose Petal Effect

Bharat Bhushan[1] and Michael Nosonovsky[2]
[1]Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA
[2]Department of Mechanical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Synonyms

Petal effect

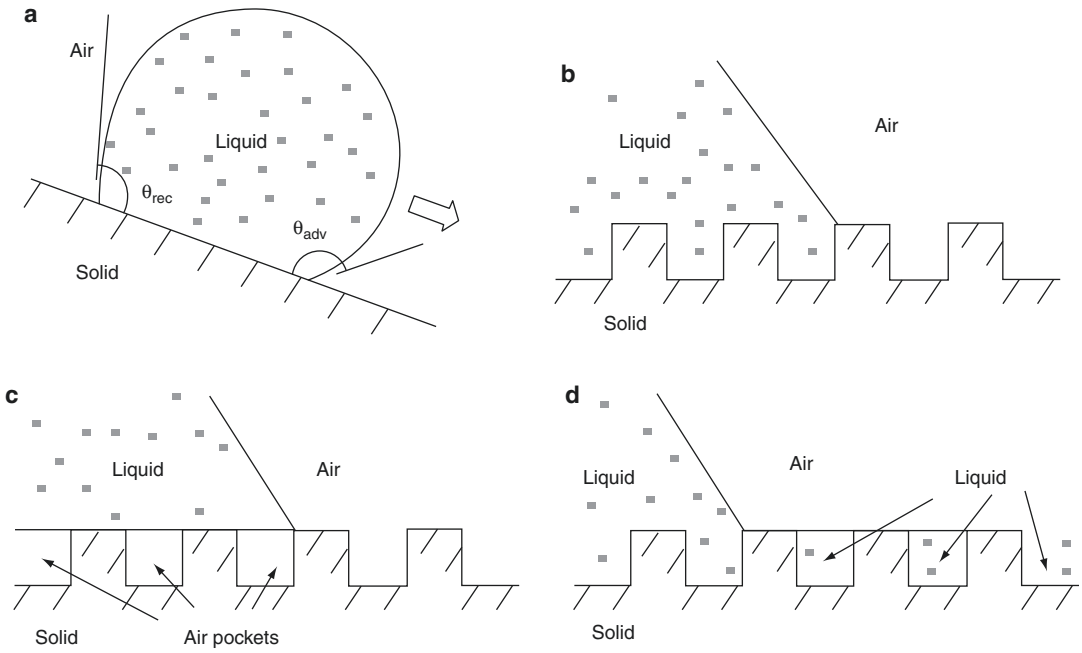
Definition

Rose petal effect is the ability of certain rough surfaces to have a high contact angle with water simultaneously with high adhesion (large contact angle hysteresis) with water.

Theoretical

Recent experimental findings and theoretical analyses made it clear that the early Wenzel [1] and Cassie and Baxter [2] models do not explain the complexity of interactions during wetting of a rough surface which can follow several different scenarios. As a result, there are several modes of wetting of a rough surface, and therefore, wetting cannot be characterized by a single number such as the contact angle (CA). Contact angle hysteresis (the difference between the advancing and receding CA) is another important parameter which characterizes wetting (Fig. 1).

Wetting of a solid is dependent upon the adhesion of water molecules to the solid. On the one



Rose Petal Effect, Fig. 1 (**a**) Schematics of a droplet on a tilted substrate showing advancing (θ_{adv}) and receding (θ_{rec}) contact angles. The difference between these angles constitutes the contact angle hysteresis. Configurations described by (**b**) the Wenzel equation for the homogeneous interface, (**c**) Cassie–Baxter equation for the composite interface with air pockets, and (**d**) the Cassie equation for the homogeneous interface

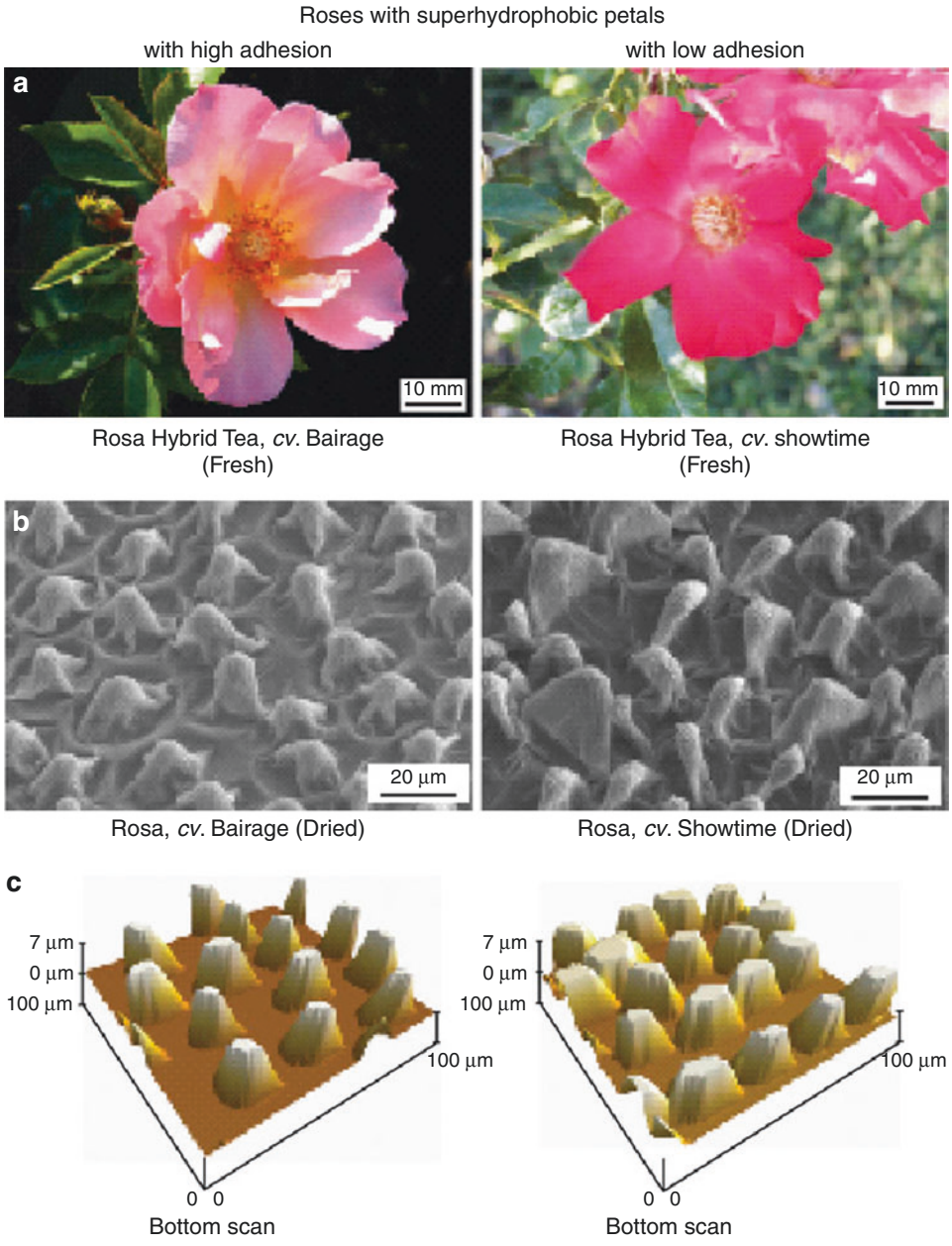
hand, a high CA is a sign of low liquid–solid adhesion. On the other hand, low CA hysteresis is a sign of low liquid–solid adhesion as well. It is now widely believed that a surface can be superhydrophobic and at the same time strongly adhesive to water [3–9]. The so-called petal effect is exhibited by a surface that has a high CA, but also a large CA hysteresis and strong adhesion to water.

Main Findings

Plant leaves and petals provide an example of surfaces with high CA, and high and low CA hysteresis. Bhushan and Her [3] studied two kinds of superhydrophobic rose petals: (1) Rosa Hybrid Tea, cv. Bairage and (2) Rosa Hybrid Tea, cv. Showtime, referred to as Rosa, cv. Bairage and Rosa, cv. Showtime, respectively. Figure 2 shows optical micrographs, scanning electron microscopy (SEM) images, and atomic force microscope (AFM) surface height maps of two rose petals. Figure 3 shows a sessile and a suspending water droplet on Rosa, cv. Bairage, demonstrating that it can simultaneously have high CA, high adhesion, and high CA hysteresis.

The surface roughness of the two rose petals was measured with the AFM, and the results for the peak-to-base height of bumps, the mid-width, peak radius, and bump density are summarized in Table 1. The data indicates that the low adhesion specimen (Rosa, cv. Showtime) has higher density and height of the bumps, indicating that the penetration of water between the microbumps is less likely. Wetting of a rough surface with a single level of hierarchy of roughness details can follow several scenarios (Table 2).

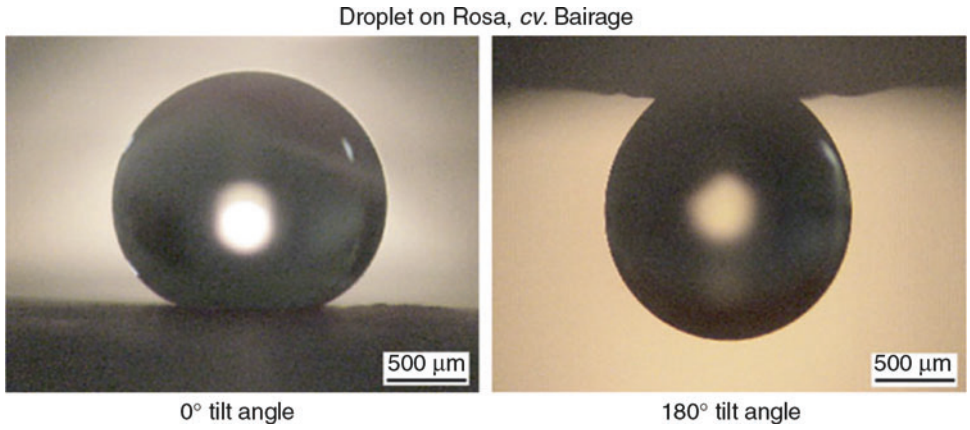
For a hierarchical structure with small bumps on top of larger bumps, a larger number of scenarios is available, and they are summarized in Table 3 and Fig. 4. Water can penetrate either into the micro- or nanostructure, or into both. In addition, the micro- or nanostructure can be



Rose Petal Effect, Fig. 2 (**a**) Optical images, (**b**) scanning microscope micrographs, and (**c**) atomic force microscope roughness maps of petals of two roses (Rosa Hybrid Tea, cv. Bairage) (Rosa, cv. Bairage), and Rosa Hybrid Tea, cv. Showtime (Rosa, cv. Showtime) (Adapted from Bhushan and Her [3])

impregnated by water or air. The regimes with water penetrating into the microstructure can have high solid–water adhesion and therefore, high CA hysteresis.

Bhushan and Her [3] conducted a series of carefully designed experiments to decouple the effects of the micro- and nanostructures. They synthesized microstructured surfaces with pillars out of epoxy resin. The epoxy surfaces were reproduced from model Si templates and were created by a two-step molding process producing a dual replica (first a negative replica and then a



Rose Petal Effect, Fig. 3 Optical micrographs of water droplets on Rosa, cv. Bairage at 0° and 180° tilt angles. Droplet is still suspended when the petal is turned upside down [3]

Rose Petal Effect, Table 1 Surface roughness statistics for the two rose petals [3]

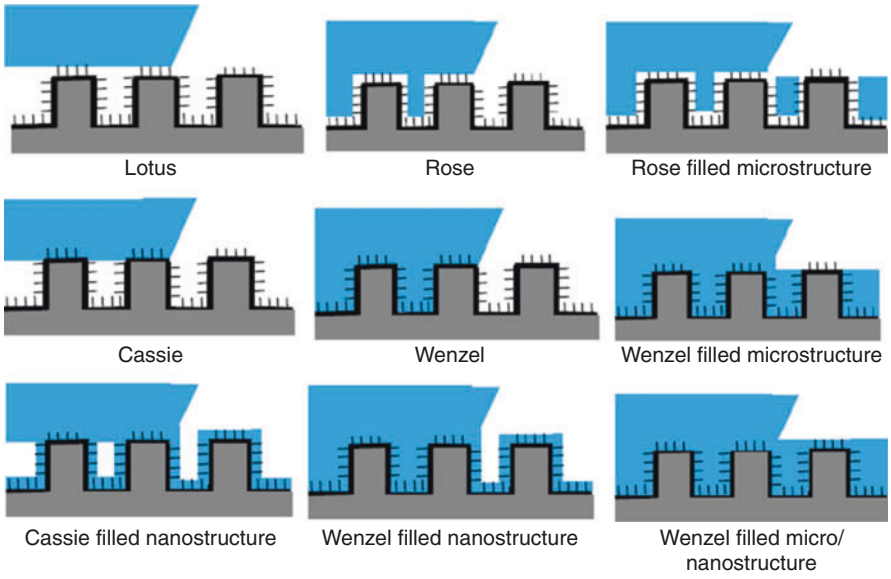
	Peak-to-base height (µm)	Mid-width (µm)	Peak radius (µm)	Bump density (1/10,000 µm^2)
Rosa, cv. Bairage (high adhesion)	6.8	16.7	5.8	23
Rosa, cv. Showtime (low adhesion)	8.4	15.3	4.8	34

Rose Petal Effect, Table 2 Wetting regimes of a surface with a single level of hierarchy of roughness

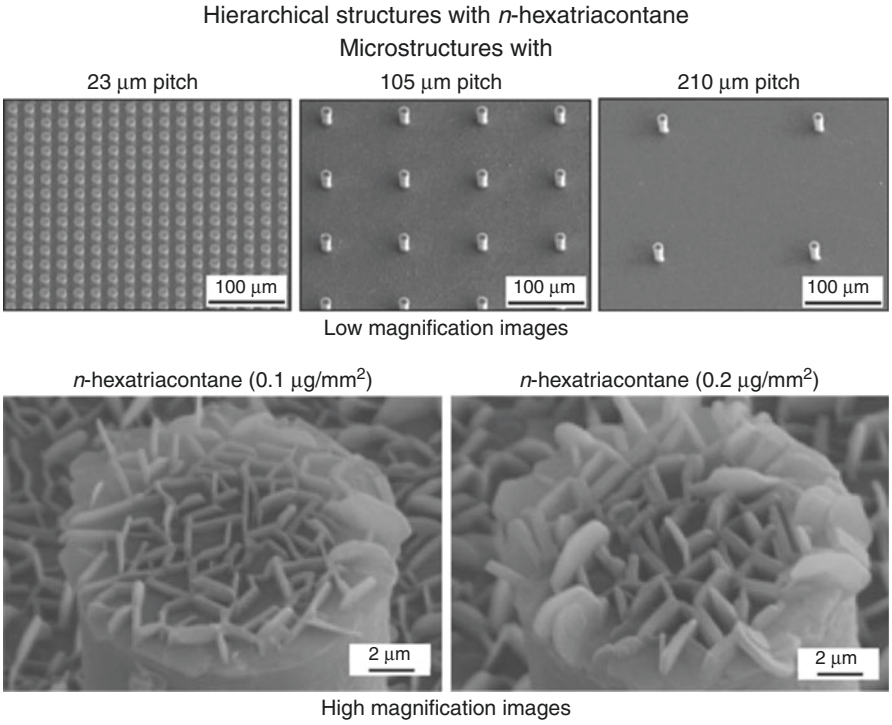
State	Cassie–Baxter	Wenzel	Impregnating Cassie
Cavities	Air	Water under droplet	Water everywhere
CA	High	High	High
CA hysteresis	Low	Can be high	Low

Rose Petal Effect, Table 3 Different regimes of wetting of a surface with dual roughness

	Air in microstructure	Water under droplet in microstructure	Water impregnating microstructure
Air in nanostructure	Lotus, high CA, low CA hysteresis	Rose, high CA, high CA hysteresis	Rose-filled microstructure
Water under droplet in nanostructure	Cassie (air-filled microstructure, water in nanostructure), high CA, low CA hysteresis	Wenzel (water in micro- and nanostructure), high CA, high or low CA hysteresis	Wenzel-filled microstructure
Water impregnating nanostructure	Cassie-filled nanostructure	Wenzel-filled nanostructure	Wenzel-filled micro and nanostructure



Rose Petal Effect, Fig. 4 Schematics of nine wetting scenarios for a surface with hierarchical roughness



Rose Petal Effect, Fig. 5 SEM micrographs of the microstructures and nanostructures fabricated with two different masses of n-hexatriacontane for hierarchical structure. All images were taken at 45° tilt angle. All samples are positive replicas, obtained from negative replica with dental wax and Si micropatterned master template (14 μm diameter and 30 μm height) fabricated with epoxy resin coated with n-hexatriacontane [3]

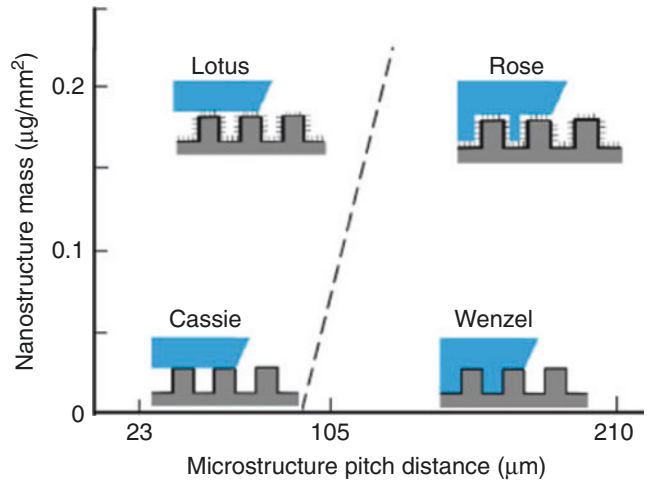
R

positive replica of the original Si template). Surfaces with a pitch (the periodicity of the structure of the pillars) of 23, 105, and 210 μm and with the same diameter (14 μm) and height (30 μm) of the pillars were produced. After that, nanostructures were created on the microstructured sample by self-assembly of the alkane n-hexatriacontane ($CH_3(CH_2)_{34}CH_3$) deposited by a thermal evaporation method. Alkanes of varying chain lengths are common hydrophobic compounds of plant waxes. On smooth surfaces, alkanes can cause a large contact angle and a small contact angle hysteresis for water droplets. To fabricate the nanostructure, various masses of n-hexatriacontane were coated on a microstructure. The nanostructure is formed by three-dimensional platelets of n-hexatriacontane. Platelets are flat crystals, grown perpendicular to the surface. They are randomly distributed on the surface, and their shapes and sizes show some variation. Figure 5 shows selected images. When different masses of wax are applied, the density of the nanostructure is changed.

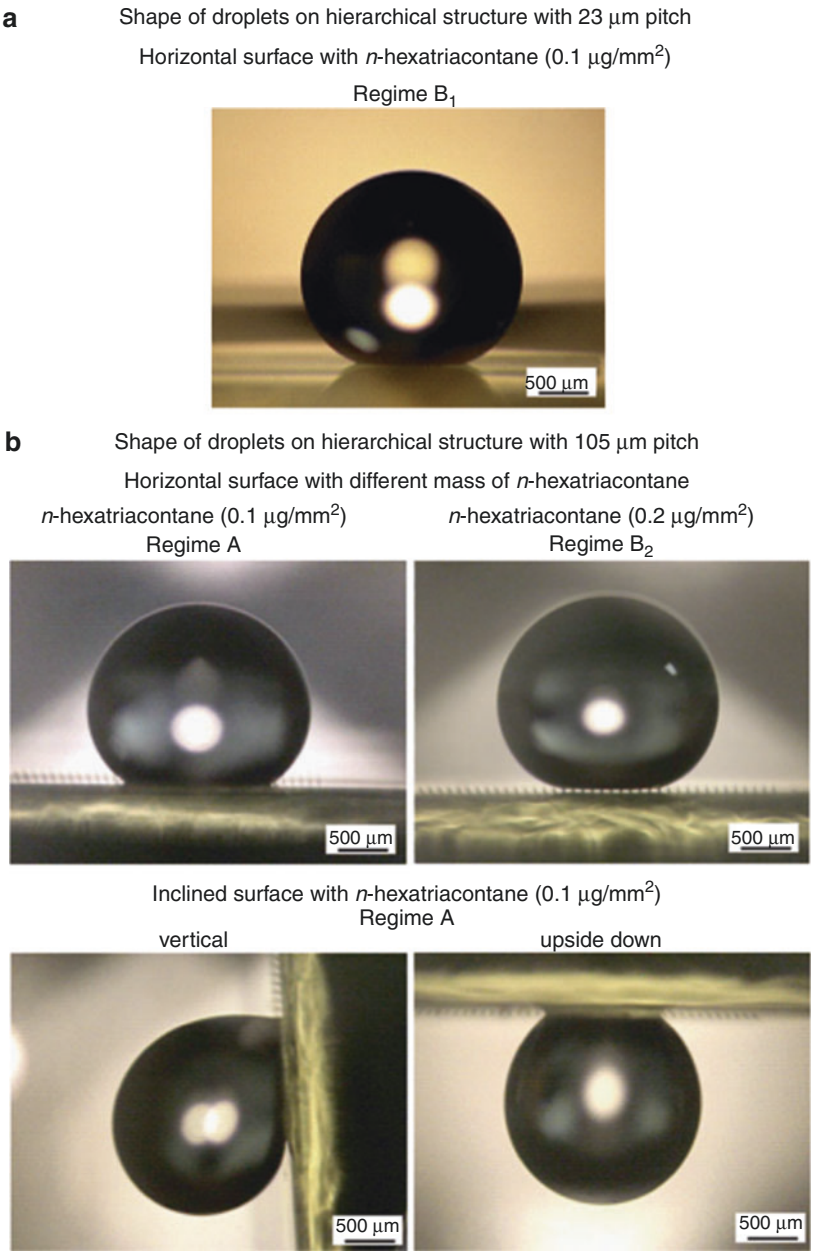
For surfaces with a small pitch of 23 μm, while the mass of n-hexatriacontane is changed, there are only small changes in the static contact angle and contact angle hysteresis values, which means that they are always in the "Lotus" wetting regime. On the surface with a 210 μm pitch value, as the mass of n-hexatriacontane is increased, the static contact angle is increased, and the reverse trend was found for the contact angle hysteresis. This was interpreted as evidence that the nanostructure is responsible for the CA hysteresis and low adhesion between water and the solid surface. The results are summarized in Table 4. The wetting regimes are shown schematically in Fig. 6 as a function of the pitch of the microstructure and the mass of n-hexatriacontane. Small mass of the nanostructure material

Rose Petal Effect, Table 4 CA and CA hysteresis for surfaces with various micro- and nanoroughness (Based on Bhushan and Her [3])

Mass of n-hexatriacontane (μg/mm^2)	Pitch					
	23 μm		105 μm		210 μm	
	CA	CA hysteresis	CA	CA hysteresis	CA	CA hysteresis
0.1	164°	3°	152	87	135	45
0.12	165°	3°	153	20	135	42
0.16	166°	3°	160	5	150	12
0.2	167°	3°	168	4	166	3



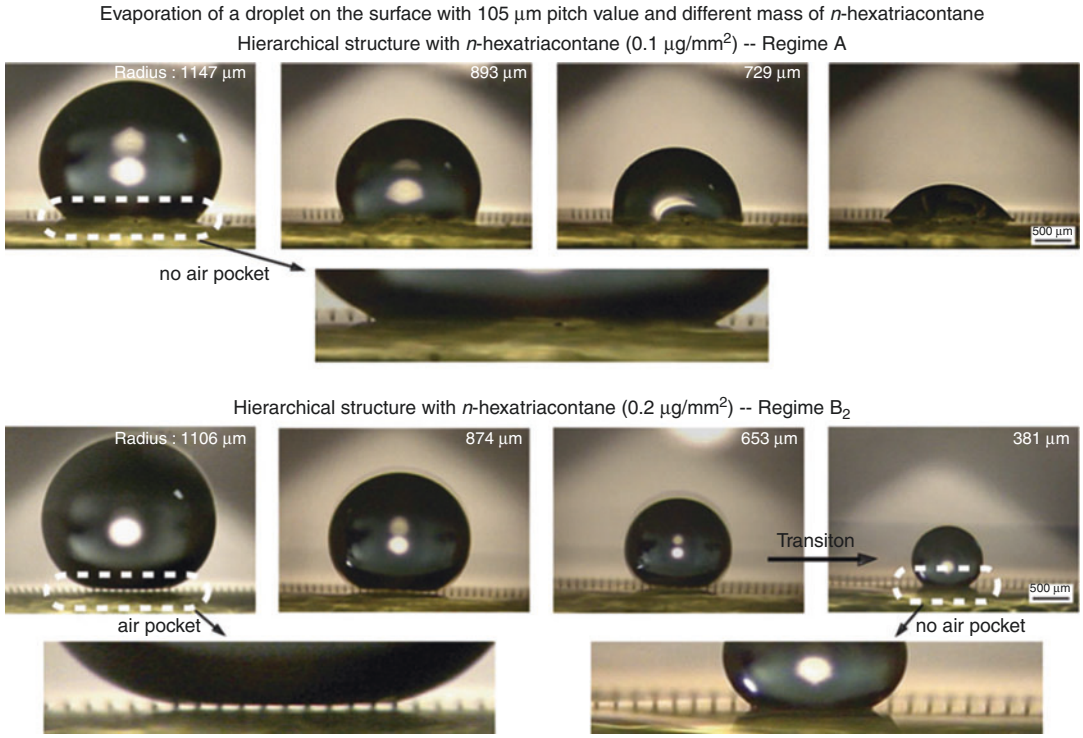
Rose Petal Effect, Fig. 6 Schematic of a wetting regime map as a function of microstructure pitch and the mass of nanostructure material. The mass of nanostructure material equal to zero corresponds to microstructure only (with the Wenzel and Cassie regimes). Higher mass of the nanostructure material corresponds to higher values of pitch, at which the transition occurs



Rose Petal Effect, Fig. 7 (**a**) Droplet on a horizontal surface of hierarchical structure with 23 μm pitch and n-hexatriacontane (0.1 μg/mm^2) showing air pocket formation and (**b**) droplet on a hierarchical structure with 105 μm pitch and n-hexatriacontane (0.1 μg/mm^2) and 0.2 μg/mm^2 showing no air pocket and air pocket formation, respectively. Also shown is the image taken on the inclined surface with hierarchical structure with 0.1 μg/mm^2 showing that droplet is still suspended [3]

correspond to the Cassie and Wenzel regimes, whereas high mass of the nanostructure corresponds to the Lotus and Rose regimes. The Lotus regime is more likely for larger masses of the nanostructure material. Figure 7 shows a droplet on a horizontal surface of a hierarchical structure with 23 and 105 μm pitches and n-hexatriacontane (0.1 μg/mm^2). Air pockets are observed in the first case and not observed in the second case, indicating the difference between the two regimes [3].

To further verify the effect of wetting states on the surfaces, evaporation experiments with a droplet on a hierarchical structure coated with two different amounts of n-hexatriacontane were performed. Figure 8 shows the optical



Rose Petal Effect, Fig. 8 Optical micrographs of droplet evaporation on the hierarchical structured surfaces with 105 μm pitch value. n-Hexatriacontane (0.1 μg/mm^2)-coated sample has no air pocket formed between the pillars in the entire contact area until evaporation was completed. Hierarchical structure with n-hexatriacontane (0.2 μg/mm^2) has air pocket, and then the transition from the "Lotus" regime to the "Rose petal" regime occurred [3]

micrographs of a droplet evaporating on two different hierarchical structured surfaces. On the n-hexatriacontane (0.1 μg/mm^2)-coated surface, an air pocket was not visible at the bottom area of the droplet. However, the droplet on the surface has a high static contact angle (152°) since the droplet still cannot completely impregnate the nanostructure. The footprint size of the droplet on the surface has only small changes from 1820 to 1791 μm. During evaporation, the initial contact area between the droplet and hierarchical structured surface does not decrease until the droplet evaporates completely, which means complete wetting between droplet and microstructures. For the n-hexatriacontane (0.2 μg/mm^2)-coated surface, the light passes below the droplet, and air pockets can be seen, so to start with the droplet is in the Cassie-Baxter regime. When the radius of the droplet decreased to 381 μm, the air pockets are not visible anymore. The footprint size of the droplet on the surface is changed from 1177 to 641 μm, since droplet remained on only a few pillars until the end of the evaporation process [9].

The experimental observations of the two types of rose petals show that hierarchically structured plant surfaces can have both adhesive and non-adhesive properties at the same time with high CA. This is due to the existence of various modes of wetting of a hierarchical surface, so that water can penetrate either into macro- or nanoroughness, or into both. Water penetration into the microroughness tends to result in high adhesion with the solid surface, whereas the presence of the nanoroughness still provides high CA. As a result, two distinct modes of wetting are observed. One can be called the "Lotus" mode (with low adhesion) and another is the "rose" mode with high adhesion.

Cross-References

▶ Biomimetics
▶ Lotus Effect

References

1. Wenzel, R.N.: Resistance of solid surfaces to wetting by water. Ind. Eng. Chem. **28**, 988–994 (1936)
2. Cassie, A., Baxter, S.: Wettability of porous surfaces. Trans. Faraday Soc. **40**, 546–551 (1944)
3. Bhushan, B., Her, E.K.: Fabrication of superhydrophobic surfaces with high and low adhesion inspired from rose petal. Langmuir **26**, 8207–8217 (2010)
4. Bormashenko, E., Stein, T., Pogreb, R., Aurbach, D.: "Petal Effect" on surfaces based on lycopodium: high-stick surfaces demonstrating high apparent contact angles. J. Phys. Chem. C **113**, 5568–5572 (2009)
5. Chang, F.M., Hong, S.J., Sheng, Y.J., Tsao, H.K.: High contact angle hysteresis of superhydrophobic surfaces: hydrophobic defects. Appl. Phys. Lett. **95**, 064102 (2009)
6. Feng, L., Zhang, Y., Xi, J., Zhu, Y., Wang, N., Xia, F., Jiang, L.: Petal effect: a superhydrophobic state with high adhesive force. Langmuir **24**, 4114–4119 (2008)
7. Gao, L., McCarthy, T.J.: Teflon is hydrophilic. Comments on definitions of hydrophobic, shear versus tensile hydrophobicity, and wettability characterization. Langmuir **24**, 9184–9188 (2008)
8. Jin, M.H., Feng, X.L., Feng, L., Sun, T.L., Zhai, J., Li, T.J., Jiang, L.: Superhydrophobic aligned polystyrene nanotube films with high adhesive force. Adv. Mater. **17**, 1977–1981 (2005)
9. Bhushan, B., Nosonovsky, M.: The rose petal effect and the modes of superhydrophobicity. Philos. Trans. Royal. Soc. A. **368**, 4713–4728 (2010)

Rotary Nanoelectromechanical Systems (NEMS)

▶ Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices

Rotary Nanomachines

▶ Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices

Rotary Nanomotors

▶ Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices

Rotation Sensors

▶ Gyroscopes

S

Sand Skink

► Friction-Reducing Sandfish Skin

SANS

► Small-Angle Scattering of Nanostructures and Nanomaterials

SAXS

► Small-Angle Scattering of Nanostructures and Nanomaterials

Scanning Electron Microscopy

Yimei Zhu[1] and Hiromi Inada[2]
[1]Center for Functional Nanomaterials, Brookhaven National Lab, Upton, NY, USA
[2]Hitachi High-Technologies Corporation, Hitachinaka, Japan, Japan

Synonyms

Scanning electron microscopy and secondary-electron imaging microscopy

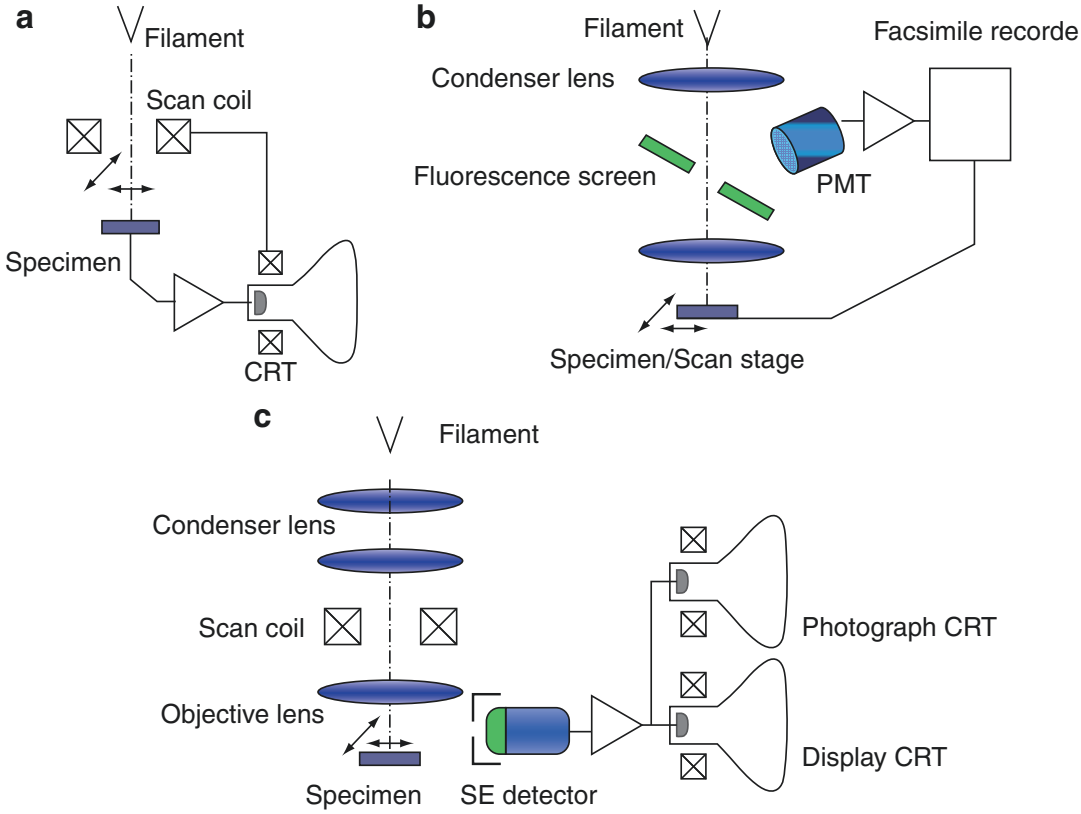
© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Definition

A scanning electron microscope (SEM) is an instrument that uses high-energy electrons in a raster scan pattern to form images, or collect other signals, from the three-dimensional surface of a sample.

Introduction

The scanning electron microscope (SEM) is one of the most popular and user-friendly imaging tools that reveal the surface topography of a sample. It is also widely used for structural characterization of materials and devices, especially in the field of nanotechnology. Today there are in excess of 50,000 SEMs worldwide, and it is often seen as a "must-have" apparatus for research institutes and industry laboratories. In the SEM, incident electrons interact with the atoms that make up the sample producing signals that contain information about the sample's surface morphology, composition, and other physical and chemical properties. The most common imaging mode in SEM lies in using secondary electrons. Since secondary electrons have very low energies, they are generated in and escape from regions near the sample surface. Combined with various detection systems, a SEM also can be used to determine the sample's chemical composition through energy-dispersive x-ray spectroscopy and Auger electron spectroscopy and identify its phases through



Scanning Electron Microscopy, Fig. 1 Schematics of configurations of the scanning electron microscopes proposed and developed by some notable pioneers in the SEM history. (**a**) Knoll in 1935, (**b**) Zworykin in 1945, and (**c**) Cambridge Instruments in 1965

analyzing electron diffraction patterns, mostly via high-energy backscattered electrons. Besides backscattered electrons, Auger electrons, characteristic x-rays, and other signals are generated from the interactions of the incident beam, and the sample under a SEM includes plasmons, bremsstrahlung radiation (noncharacteristic x-rays), cathodoluminescence, and electron-beam-induced current. This entry is focused on secondary-electron imaging related instrumentation, signal generation processes, and state-of-the-art imaging capabilities in SEM.

A Brief History

SEM was invented by Max Knoll in 1935 in Germany [1, 2] to study the targets of television tubes. The instrument consisted of electron-beam deflection coils that scan the beam on a plate as the sample in a cathode ray tube (CRT) and an amplifier that boosts the plate current to display the signal on another CRT (Fig. 1a). Two years later, Manfred von Ardenne built an electron microscope with a highly demagnified probe using two condenser lenses for scanning transmission electron microscopy and also tried it as a SEM. Zworykin and his coworkers of the RCA Laboratories in the USA designed and built a dedicated SEM in 1942. Its electron optics includes three electrostatic lenses with scan coils placed between the second and third ones. A photomultiplier was first used to detect secondary electrons (Fig. 1b). The essential components of this apparatus are similar to those used in modern SEMs. The probe size of the incident, or primary, electron beam had a diameter of about 10 nm. However, compared with transmission

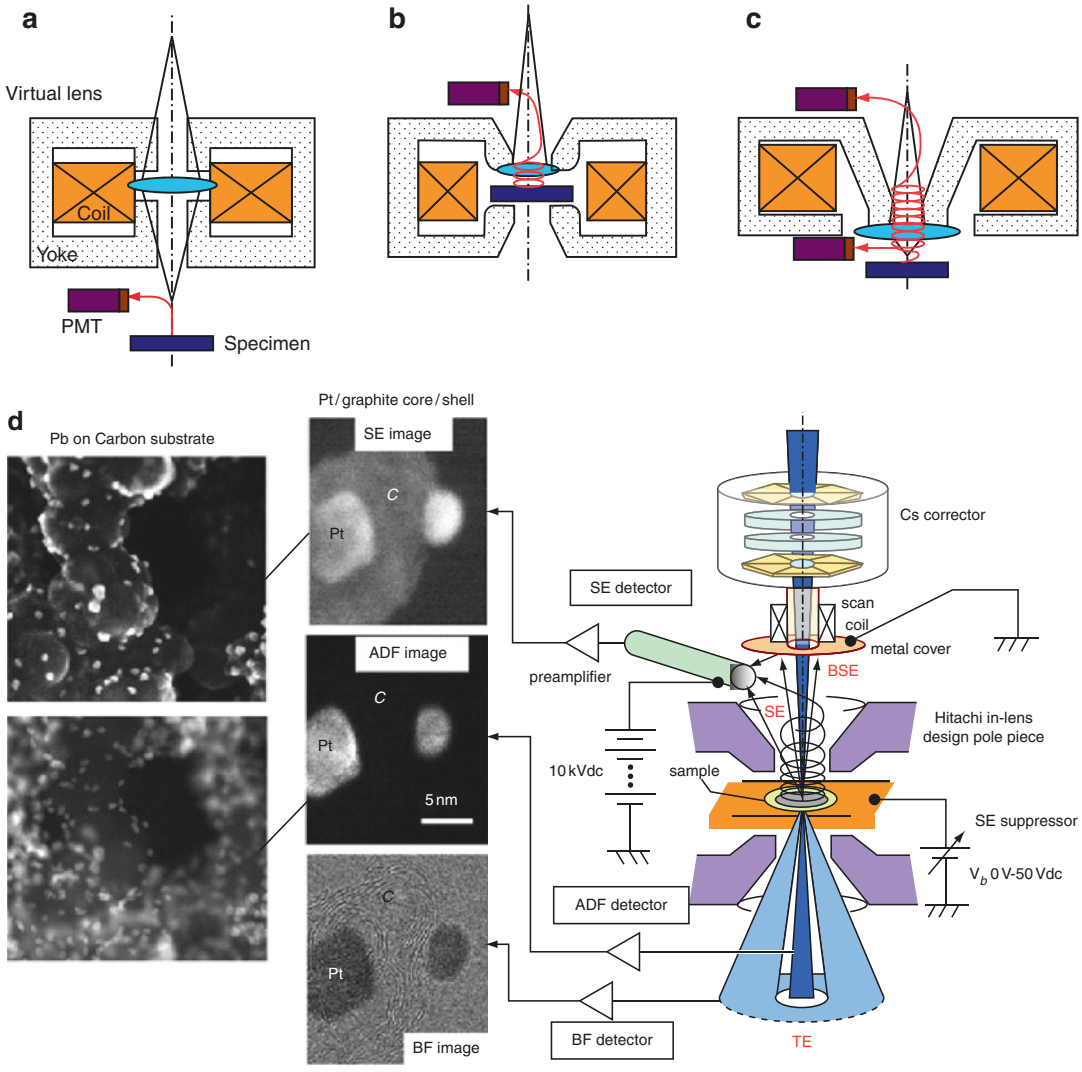
electron microscopes (TEMs), at that period, it could not image secondary electrons satisfactorily due to the poor signal-to-noise ratios of the images [3]. Sir Charles Oatley and Dennis McMullan built their first SEM at Cambridge University in 1948. The SEM technology was further pioneered by many postgraduate students at Cambridge including Gary Stewart. The first commercial instrument, named as "Stereoscan," was launched in 1965 by the Cambridge Scientific Instrument Company for DuPont [4]. The instrument consists of electron multiplier detector with beryllium-copper dynodes to detect scattered electrons from the specimen surface. Images were displayed on a CRT, while another synchronized CRT recorded them on camera film. An Everhart-Thornley-type secondary-electron detector [5] significantly improved the detection efficiency of the low-energy secondary-electron signals (Fig. 1c). JEOL produced first commercial Japanese SEM, JSM-1, in 1966, while Hitachi commercialized its SEM, HSM-2, in 1969.

Instrumentation

A typical SEM consists of an electron gun, an electron lens system, various electron-beam deflection coils, electron detectors, and display and recording devices [6]. In a SEM, an electron beam is emitted from an electron gun sitting on a thermionic filament cathode, or from a field-emission needle tip. Tungsten is often used for thermionic electron guns due to its low cost, high melting point, and low vapor pressure so that it tolerates heating to about 2,800 K for electron emission. Other types of electron emitters include lanthanum hexaboride (LaB6) cathodes, which offer higher brightness but require a better vacuum to avoid oxidizing the gun. Field-emission guns (FEGs) often used in modern SEMs can be the thermally assisted Schottky type, using emitters of zirconium oxide (ZrO), or the cold cathode type using tungsten <310> single crystal emitters and operated at room temperature. A cold field-emission gun has a much smaller source size (5–10 nm) than a tungsten filament (1–10 μm) with three to four orders of magnitude larger current density and brightness [7].

The typical energy range of the electron beam used in SEM is from 0.5 to 40 keV. The electron-condenser lens system usually demagnifies the electron source more than hundreds of times to form a small probe on the sample. The beam passes through pairs of deflection coils, or scanning coils, in the electron column, typically in the final lens, which deflect the beam in the x and y axes by applying an incremental current into the scan coils, so that it scans in a raster fashion over a rectangular area of the sample surface.

Historically, SEM incorporates an objective lens. Unlike the objective lens in optical microscopes or transmission electron microscopes (TEMs), its purpose in SEM is not to image the sample but to focus the small probe on the sample. There are three types of objective lenses: out-lens (Fig. 2a), in-lens (Fig. 2b), and semi-in-lens (Fig. 2c). Most early SEMs had the simplest out-lens design, in which the sample sits beneath the lens leaving a large area available in the sample chamber. However, the yoke gap across the optical axis acts as a lens with leakage field, thus yielding significant imaging aberration. The in-lens pole piece, originally designed for TEMs, was adopted for SEM to reduce spherical aberration. Hitachi developed the first commercial SEM with such a design in 1985 [8]. Their microscope operating at 30 kV reaches a probe size of 0.5 nm with spherical aberration coefficient C_s of 1.6 mm. The drawback of the design is that a conventional thin sample, similar to that used in TEM, is required because the sample sits inside the pole-piece gap. A design resulting from a compromise between the out-lens and in-lens is the semi-in-lens pole piece (Fig. 2c), offering reasonable spatial resolution but with larger open space for the sample chamber so that a thick sample can be used. Such a design allows the sample to be placed a few mm from the pole piece and a large magnetic field to be applied to produce a smaller focal length and less spherical aberration. SEMs with the semi-in-lens design played a significant role in characterizing devices for the semiconductor industry in early 1990s.



Scanning Electron Microscopy, Fig. 2 (**a**–**c**) Three different objective lens designs. (**a**) Out-lens, (**b**) in-lens, and (**c**) semi-in-lens. (**d**) Schematic of the lens detector configuration of the aberration-corrected scanning electron microscope, Hitachi HD2700C, that routinely achieves an atomic resolution in imaging using secondary electrons and transmitted electrons. Two examples are shown on the *left*; one is Pb particles on a carbon support, and the other is Pt particles with carbon-graphite shells. *BF* bright-field, *ADF* annular dark-field, *SE* secondary electrons, *TE* transmitted electrons, and *BES* backscattered electrons. The SE images clearly give a topographic view of the area and higher brightness of the light element C, compared with the corresponding ADF images

It is noteworthy that the in-lens design shown in Fig. 2b is very similar to that used in conventional TEM and/or STEM (scanning transmission electron microscope). Thus, with an efficient secondary-electron detector, a TEM or a STEM also can image a sample surface [9, 10]. Figure 2d shows a schematic of the objective lens (in-lens design), the sample, and the arrangement of detectors in the Hitachi HD2700C aberration-corrected SEM/STEM [11]. The instrument operates at 80–200 kV, allowing simultaneous acquisition of bright-field (BF), annular dark-field (ADF), and secondary-electron (SE) images. In the BF and ADF modes, the transmitted electrons are used

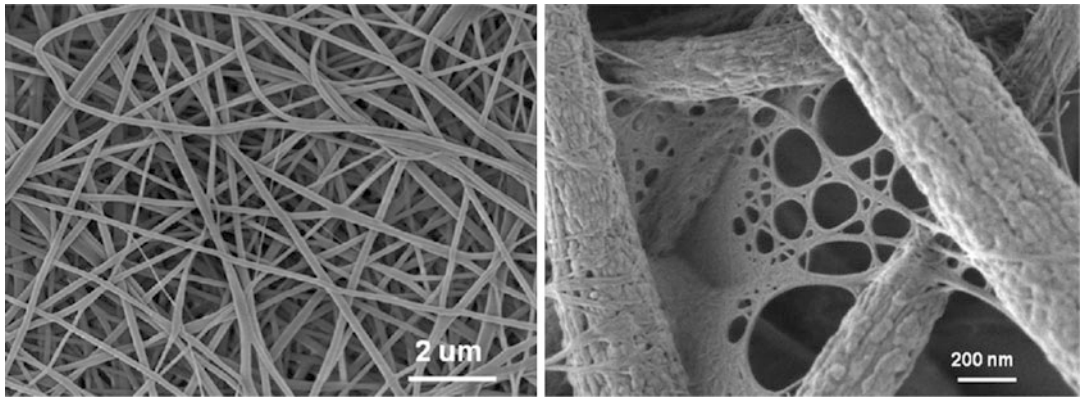
to form the images, thus providing structural information from the sample's interior. In contrast, in the SE mode, electrons emerging from the surface with low energies, or short escape length, are used, and thus, the signals are surface sensitive. Different imaging modes have their own advantages and limitations. Since the BF signals are close to the phase contrast seen in TEM, they offer high spatial resolution, but are difficult to interpret. ADF images, on the other hand, are based on Rutherford scattering, and thus, their image intensity is directly related to the sample's atomic number Z (the so-called Z-contrast imaging). Since BF and ADF images are projected images, they give little structural information in the direction of the beam's trajectory. In contrast, SE imaging offers depth information on the surface topography and is more sensitive to the light elements than is ADF imaging (see Fig. 2d). Combining these different imaging modes in an electron microscope with an in-lens design, it has been demonstrated that simultaneously imaging both surface (SE) and bulk (ADF) at atomic resolution is possible in thin samples for a wide range of elements, from uranium and gold to silicon and carbon [11, 12].

The electron detector is another important part of the SEM instrumentation. Although detector itself does not determine the image resolution in SEMs, it is essential to improve the SEM resolving power in terms of the signal-to-noise ratio. The commonest detector used in SEMs today still is that developed by Everhart and Thornley in 1957 [5], the so-called E-T detector. The detector consists of a photomultiplier, a light guide, and a positively biased scintillator. To attract low-energy electrons effectively, the scintillator is applied in a 10 kV dc bias to accelerate the electrons. The energized electrons cause the scintillator to emit flashes of light (cathodoluminescence) that then are transmitted to the photomultiplier. The amplified output of the electrical signals by the photomultiplier is displayed as a two-dimensional intensity distribution that can be viewed and photographed on an analogue video display. The Everhart-Thornley detector, which normally is positioned to one side above the specimen, exhibits low efficiency in detecting backscattered electrons because few such electrons are emitted in the solid angle subtended by the detector. Furthermore, its positive bias cannot readily attract the high-energy backscattered electrons (close to the energy of the incident electrons). Backscattered electrons are usually collected above the sample in a "doughnut-type" arrangement, concentric with the electron beam, to maximize the solid angle of collection.

Secondary-Electron Signal Generation

Secondary-electron (SE) imaging is the most frequently used mode of imaging in SEM. Secondary electrons, defined as the electrons with energy below 50 eV, are generated along the primary electrons' trajectories within the sample, but are subject to elastic and inelastic scattering during their passage through the sample. These electrons can be valence electrons or are ejected from the orbits of the inner shells (most likely the k-shells) of the sample. The consequence of their low kinetic energy is their shallow escape depth, which is about 1 nm for metals and up to 10 nm for insulators. The probability of escape decreases continuously with the increase of the depth below the surface. However, there is a nonzero probability of secondary-electron emission arising from inelastic scattering below the escape depth, as, for example, when a primary electron creates a fast secondary electron (energy > 50 eV) that travels toward the surface and generates a lower-energy secondary electron within the escape depth. The production of the secondary electron signals involves the generation, propagation, escape from the surface, and arrival at the detector. These four processes are detailed in the atomic imaging section.

The secondary electrons discussed here are often referred to as SE_I, i.e., the secondary electrons generated by the incident beam upon entering the sample. Secondary electrons generated by backscattered electrons when leaving the sample are termed SE_{II}. Secondary electrons generated when the backscattered electrons strike a lens pole piece or the sample chamber's wall and by primary electrons hitting the aperture are, respectively, called SE_{III} and SE_{IV}. Although SE_{III}



Scanning Electron Microscopy, Fig. 3 SEM micrographs of a polymer membrane that consists of electrospun fibrous scaffold for water filtration. The fibrous scaffold has an average diameter of 200 nm. The high-resolution SEM image on the *right* shows that cellulose nanofibers (5–10 nm in diameter) infused into the scaffold form a cellulose network to enhance the membrane's ability to remove bacteria and viruses. The images were taken with JEOL7600 SEM at operation voltage of 0.5 eV. A thin layer of carbon was coated on the sample to avoid charging. Note the bright contrast at the edge of the fibers generates a topological view of the fibrous structure

contain the information on the sample, SE_{IV} do not. Furthermore, there are fast secondary electrons, with energy higher than 50 eV. Bias experiments, wherein a positive dc voltage is applied to the sample to suppress the emission of the secondary electrons, were mainly designed to separate the SE_I from backscattered electrons; they cannot distinguish SE_I from SE_{II}, which for a very thin specimen should be negligible. Measuring other types of secondary electrons, including those with high energy, would require a different bias experiment. Heavy elements (high atomic number) that backscatter electrons more strongly than do light elements (low atomic number), and thus appear brighter in the image, can be used to yield contrast that contains information of a sample's chemical composition.

Figure 3 shows SEM images of a nanofiber containing polymer membrane developed for water filtration with enhanced ability to remove bacteria and viruses. The image on the right shows the structural network formed by the 5–10 nm diameter cellulose nanofibers. It reveals astonishing topographic details on how the cellulose nanofibers are interwoven with scaffold. In SEM images, flat surfaces give even contrast, while curved surfaces and sharp edges often appear brighter (high image intensity). This is due to the increased escape of secondary electrons from the top and side surfaces when the interaction volume intercepts them. Since the secondary-electron detector is biased with a high acceleration voltage, a surface facing away from the detector still can be imaged. Surfaces tilted away from the normal to the beam allow more secondary electrons to escape [13].

Atomic Imaging Using Secondary Electrons

In the last decade or so, high-resolution SEM has proven an indispensable critical dimension metrology tool for the semiconductor industry. The road map for semiconductor nanotechnology identifies the need for ultrahigh-resolution SEM in the quest for ever-decreasing device sizes. In a SEM, the size of the imaging probe often determines the instrument's resolution power. The probe size d (measured in full width and half maximum) is a function of the beam convergence; half-angle α is an incoherent sum of contributions from source size, diffraction limit, spherical aberration, and chromatic aberration and is given by

$$d^2 = \left(\frac{4i_p}{\beta\pi^2\alpha^2}\right)^2 + \left(\frac{0.6\lambda}{\alpha}\right)^2 + (0.5C_s\alpha^3)^2 + \left(C_c\alpha\frac{\Delta E}{E}\right)^2$$

where i_p is the probe current, α is the convergence half-angle, β is the source brightness, λ is the electron wavelength at beam energy E, ΔE is the energy spread, and C_s and C_c are, respectively, the spherical and chromatic aberration coefficients of the probe-forming lens. The first term on the right side of the equation is the probe size defined by the source size, the second term is due to the diffraction limit, and the third and the fourth terms are due, respectively, to the spherical and chromatic aberrations. Recent advancement on correcting spherical aberration in SEM and STEM diminishes C_s to zero, thus eliminating the third term and produces a small probe. It is important to note that the probe size also depends on the energy spread in the fourth term that includes the energy distribution of the primary electron beam and the fluctuation of the instruments' acceleration voltage.

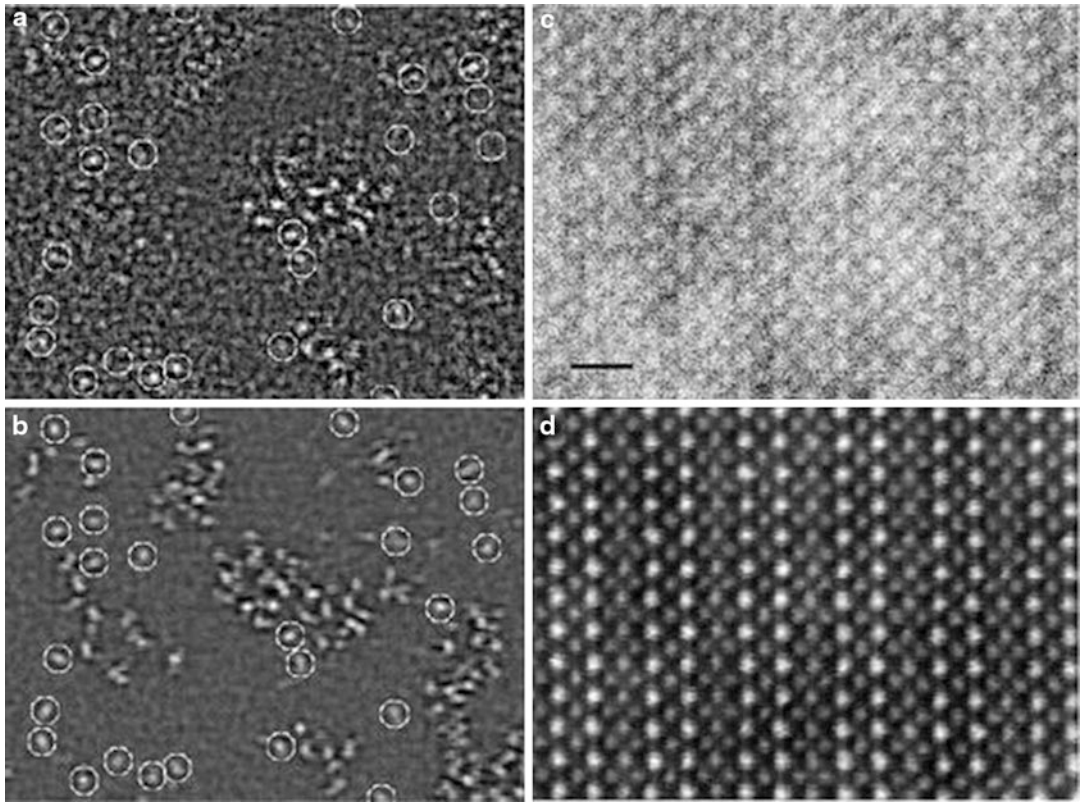
In SEM, the image resolution depends not only on the instrument but also on the sample, or, more accurately, on the sampling volume of the sample from which the signal is generated. When the incident electrons impinge on a point of the sample's surface, they interact with atoms in the sample via elastic (change trajectory) and inelastic (lose energy) interactions. The size of the interaction volume depends on the electron's landing energy, the atomic number of the sample, and its density. The simultaneous energy loss and change in trajectory spread beam into the bulk of the sample and produce an interaction volume therein that, for a thick sample, can extend from less than 100 nm to around 5 μm into the surface, i.e., more than an order of magnitude larger than the original probe size. For a very thin sample (a few nm thick), the interaction volume, as the first approximation, might be defined as the probe size.

Figure 4 illustrates the atomic resolution images using secondary electrons on single uranium atoms (a) and the (010) surface of a $YBa_2Cu_3O_7$ crystal (c) recorded on the Hitachi HD2700C SEM/STEM (Fig. 2d). For comparison, the corresponding annular dark-field images using transmitted electrons, (b) and (d), respectively, are included. The equally sharp images in SE and ADF suggest negligible imaging delocalization. Such an attainable resolution was attributed to the combination of several factors: better design of the electro-optics of the instrument (including ultrahigh electric and mechanical stabilities) and the detector, aberration correction that reduces the probe size and increases the probe current, and the higher operation voltage that beneficially assures a very small volume of beam interaction for a thin sample.

The physical mechanism of producing the low-energy secondary electrons traditionally is ascribed to inelastic scattering and decay of collective electron excitation with the incident electron giving up, say, 20 eV, to produce a secondary electron with energy 20 eV energy minus the work function of the surface. Since the momentum transfer (scattering angle) of the scattering with 20 eV energy loss is small, the transfer should be delocalized to an area >1 nm; thus, atomic imaging using secondary electrons was not considered possible. Recent studies suggest that this is not the primary mechanism for secondary-electron imaging at least on a thin sample [11]. The secondary electrons responsible for atomic-scale resolution are generated by inelastic scattering events with large momentum transfer, including those from inner shell orbitals, which give rise to a sharp central peak in the point spread function for signal generation.

In general, four steps are involved in producing the signal that is used to form a secondary-electron (SE_I) image [12]: (1) the generation of secondary electrons through the inelastic scattering of primary electrons in the sample, at a generation rate, G; (2) random motion of these secondary electrons, which are scattered by atoms of the specimen both elastically and inelastically (potentially creating other secondary electrons of lower energies), such that, on average, T electrons reach the sample surface for each secondary electron generated; (3) the escape of secondary electrons over the potential barrier at the sample surface of the specimen, with an average probability P; and (4) the acceleration of the emitted electrons in vacuum, such that a fraction D reaches the electron detector.

The secondary-electron signal S is a product of these four factors: $S = G{\cdot}T{\cdot}P{\cdot}D$. To generate contrast in a scanned probe image, one or more of the



Scanning Electron Microscopy, Fig. 4 (**a**–**b**) Simultaneous atomic imaging using secondary electrons (secondary-electron mode, **a**) and transmitted electrons (annual dark-field mode, **b**) of uranium individual atoms on a carbon support (raw data). The *circles mark* the single uranium atoms. The atoms shown in (**b**) but not in (**a**) are presumably those on the back side of the support. (**c**–**d**) Simultaneous atomic imaging using secondary electrons (**c**) and transmitted electrons (**d**) of $YBa_2Cu_3O_7$ superconductor viewing along the [010] direction (raw data). The scale bar in (**c**) is 0.7 nm

above steps must depend on the x-coordinate of the electron probe in the scan direction, i.e.,

$$\frac{dS}{dx} = I_0 TPD\frac{dG}{dx} + I_0 GPD\frac{dT}{dx} + I_0 GTD\frac{dP}{dx} + I_0 GTP\frac{dD}{dx}$$

For most non-atomically resolved SE images obtained in a SEM, $\frac{dT}{dx}$ provides the main contrast mechanism: Secondary electrons created at an inclined surface or close to a surface step have an increased probability of escape, resulting in surface-topography contrast [14]. Less commonly, variations in surface work function contribute additional contrast by providing a nonzero $\frac{dP}{dx}$. In voltage-contrast applications, changes in surface voltage provide a nonzero $\frac{dD}{dx}$. Atomic number contrast is possible if the specimen is chemically inhomogeneous and G varies with atomic number, yielding a nonzero $\frac{dG}{dx}$. However, for atomic imaging in a thin sample (Fig. 4), the dominant mechanism can be quite different. $\frac{dG}{dx}$ is likely to play an important role in the atomic-scale contrast as a consequence of Z-dependence inelastic scattering cross section and channeling effect for crystals (for thickness in the order of extinction distance, it should be minor). Because secondary electrons are generated through inelastic scattering of the incident electrons, $\frac{dG}{dx}$ is limited by the delocalization of the scattering process, which may be described by the point

spread function for inelastic scattering. The term $\frac{dT}{dx}$ should be small, for adatoms or surface atoms that lie on the detector side of the sample. $\frac{dP}{dx}$ would not become important unless the effective work function varies on an atomic scale, and $\frac{dD}{dx}$ must be also negligible at atomic scale. These assumptions are reasonable because the scattering process disperses secondary electrons over a range of x that is comparable to the escape depth, typically 1–2 nm. Consequently, T, P, and D are x-averages that vary little with x on an atomic scale. For uranium atoms on a carbon substrate, the argument is even simpler; these atoms lie outside the solid, so the terms T and P are not applicable.

Future Remarks

Secondary-electron imaging is the most popular mode of operation of the scanning electron microscope (SEM) and traditionally is used to reveal surface topography. Nevertheless, this imaging method never was regarded as being on the cutting edge of performance, due to its perceived limited spatial resolution in comparison with its TEM or STEM counterparts using transmitted electrons. Recent work using aberration-corrected electron microscopes demonstrated that secondary-electron signals in the SEM can resolve both crystal lattices and individual atoms, showing SEM's unprecedented and previously unrealized imaging capabilities. Furthermore, the work demonstrates the incompleteness of present understanding of the formation of secondary-electron images. Secondary-electron imaging using high acceleration voltage with thin samples can now compete with TEM on spatial resolution and provide new capabilities, such as depth-resolved profiles, at the atomic level. There seems to be no fundamental reason why atomic resolution in secondary-electron imaging could not be achieved at the accelerating voltages of 0.5–40 keV that are currently used in conventional SEMs. It remains to be seen whether the integrated spherical and chromatic aberration of a probe-forming objective lens can be corrected to a sufficient degree at low operation voltages. One clear outcome thus far is the importance of preparing samples with clean surfaces in order to obtain interpretable and reproducible results.

Cross-References

- ▶ Robot-Based Automation on the Nanoscale
- ▶ Scanning Tunneling Microscopy
- ▶ Transmission Electron Microscopy

References

1. Knoll, M.: Aufladepotentiel und sekundäremission elektronenbestrahlter körper. Z. Tech. Phys. **16**, 467–475 (1935)
2. von Ardenne, M.: Das Elektronen-Rastermikroskop. Praktische Ausführung. Z. Tech. Phys. **19**, 407–416 (1938) (in German)
3. Wells, O.C., Joy, D.C.: The early history and future of the SEM. Surf. Interface Anal. **38**, 1738–1742 (2006)
4. Oatley, C.W.: The early history of the scanning electron microscope. J. Appl. Phys. **53**, R1–R13 (1982)
5. Everhart, T.E., Thornley, R.F.M.: Wide-band detector for micro-microampere low-energy electron currents. J. Sci. Instrum. **37**, 246–248 (1960)
6. Goldstein, G.I., Newbury, D.E., Echlin, P., Joy, D.C., Fiori, C., Lifshin, E.: Scanning Electron Microscopy and X-ray Microanalysis. Plenum, New York (1981)
7. Pawley, J.: The development of field-emission scanning electron microscopy for imaging biological surfaces. Scanning **19**, 324–336 (1997)
8. Tanaka, K., Mitsushima, A., Kashima, Y., Osatake, H.: A new high resolution scanning electron microscope and its application to biological materials. In: Proceedings of the 11th International Congress on Electron Microscopy, Kyoto, pp. 2097–2100 (1986)
9. Liu, J., Cowley, J.M.: High resolution SEM in a STEM instrument. Scanning Microsc. **2**, 65–81 (1988)
10. Howie, A.: Recent developments in secondary electron imaging. J. Microsc. **180**, 192–203 (1995)
11. Zhu, Y., Inada, H., Nakamura, K., Wall, J.: Imaging single atoms using secondary electrons with an aberration-corrected electron microscope. Nat. Mater. **8**, 808–812 (2009)
12. Inada, H., Su, D., Egerton, R.F., Konno, M., Wu, L., Ciston, J., Wall, J., Zhu, Y.: Atomic imaging using secondary electrons in a scanning transmission electron microscope: experimental observations and possible mechanisms. Ultramicroscopy **111**(7), 865–876 (2011). doi:10.1016/j.ultramic.2010.10.002. Invited articles for the special issue in honor of John Spence
13. Joy, D.C.: Beam interactions, contrast, and resolution in the SEM. J. Microsc. **136**, 241–258 (1984)
14. Reimer, L.: Scanning Electron Microscopy, 2nd edn. Springer, New York (1998)

Scanning Electron Microscopy and Secondary-Electron Imaging Microscopy

▶ Scanning Electron Microscopy

Scanning Force Microscopy in Liquids

▶ AFM in Liquids

Scanning Kelvin Probe Force Microscopy

▶ Kelvin Probe Force Microscopy

Scanning Near-Field Optical Microscopy

Achim Hartschuh
Department Chemie and CeNS,
Ludwig-Maximilians-Universität München,
Munich, Germany

Synonyms

Near-field scanning optical microscopy (NSOM)

Definition

Scanning near-field optical microscopy (SNOM) is a microscopic technique for nanostructure investigation that achieves sub-wavelength spatial resolution by exploiting short-ranged interactions between a sharply pointed probe and the sample mediated by evanescent waves. In general, the resolution of SNOM is determined by the lateral probe dimensions and the probe-sample distance. Images are obtained by raster-scanning the probe with respect to the sample surface corresponding to other scanning-probe techniques. As in conventional optical microscopy, the contrast mechanism can be combined with a broad range of spectroscopic techniques to study different sample properties, such as chemical structure and composition, local stress, electromagnetic field distributions, and the dynamics of excited states.

Introduction

Optical microscopy forms the basis of most of the natural sciences. In particular, life sciences have benefited from the fascinating possibility to study smallest structures and processes in living cells and tissue. Optical techniques feature extremely high detection sensitivity reaching single molecule sensitivity in fluorescence, Raman scattering and absorption spectroscopy. Besides the direct visualization, chemically specific information is obtained through Raman spectroscopy.

The resolution of conventional optical microscopes, however, is limited by diffraction, a consequence of the wave nature of light, to about half the wavelength. Concepts extending optical microscopy down to nanometer length scales below the diffraction-limit are distinguished into far-field and near-field techniques. Far-field techniques rely on the detection of propagating waves at distances from the source larger than the wavelength, while near-field techniques exploit short ranged evanescent waves.

Scanning near-field optical microscopy, initiated by pioneering work of Pohl, Lewis and others in the late 1980s and the beginning of 1990s gave access to nanoscale resolution for the first time. The history of near-field optics is reviewed in [1]. In addition numerous review articles and books exist describing fundamentals and applications (see e.g., [2–4]).

This entry introduces first the key physical principles beginning with the role of evanescent and propagating waves, and the loss of spatial information upon light propagation. Concepts of near-field detection are shown using different pointed probes. The next sections describe the experimental realization and present several

applications of SNOM. The outlook addresses future prospects of SNOM and remaining challenges.

Key Principles and Concepts

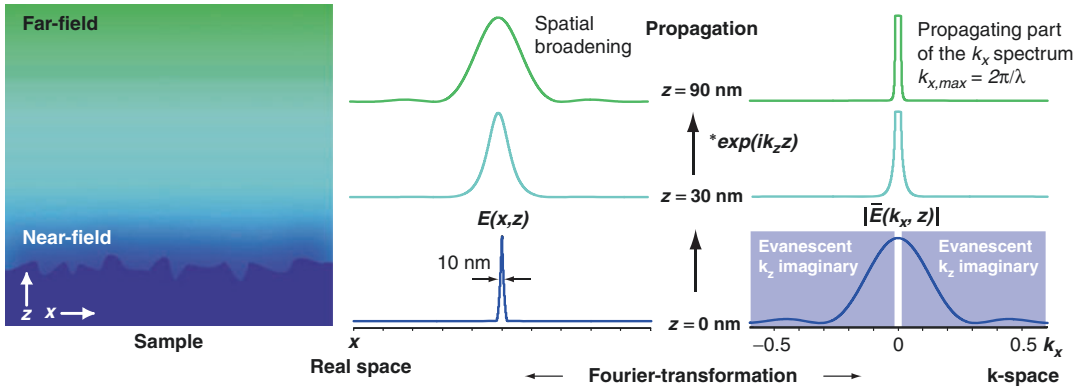
Near-field optics has its origin in the effort of overcoming the diffraction limit of optical imaging. The physical origin of this limit is sketched in the following starting with the distinction between propagating waves that form the optical far-field of a radiation source and their evanescent counterpart that dominate the optical near-field. A powerful tool to describe wave propagation is the so-called angular spectrum representation of fields expressing the electromagnetic field E in the detector plane at z as the superposition of harmonic plane waves of the form $exp\left(i\vec{k}\,\vec{r} - i\omega t\right)$ with amplitudes $\overline{E}\left(k_x, k_y, z = 0\right)$ emanating from the source plane at $z = 0$ [2].

$$E(x,y,z) = \int_{-\infty}^{+\infty}\int \overline{E}\left(k_x, k_y, z = 0\right) \times e^{i\left(k_x x + k_y y\right)} e^{\pm i k_z z} dk_x dk_y \quad (1)$$

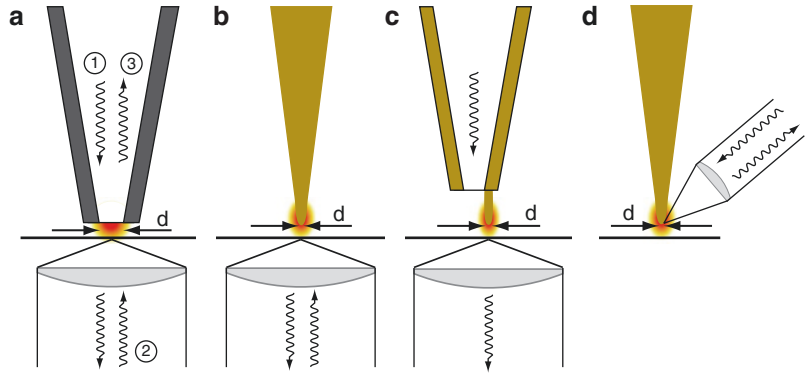
The wave vector $\vec{k}$ describing the propagation direction of the wave is represented by its components $\vec{k} = \left(k_x, k_y, k_z\right)$ while its length is fixed by the wavelength of light λ and the refractive index of the medium n through $\left|\vec{k}\right| = \sqrt{k_x^2 + k_y^2 + k_z^2} = 2\pi n/\lambda$. In Eq. 1, the time dependence of the fields has been omitted for clarity. For simplicity the following discussion is limited to the x-z-plane and $n = 1$ such that $k_z = \sqrt{4\pi^2/\lambda^2 - k_x^2}$. In Eq. 1 the term $e^{\pm i k_z z}$ controls the propagation of the associated wave: For $k_x \leq 2\pi/\lambda$ the component k_z is real and the corresponding wave with amplitude $\overline{E}(k_x, z = 0)$ propagates along the z-axis oscillating with $e^{-ik_z z}$. If $k_x > 2\pi/\lambda$ the component kz becomes complex and $e^{-|k_z z|}$ describes an exponential decay of the associated wave that is therefore evanescent. As a result, only waves with $k_x \leq 2\pi/\lambda$ can propagate and contribute to the field far from the source forming the far-field. Figure 1 schematically illustrates this behavior: In the center, the electric field $E(x, z)$ emanating from a narrow sub-wavelength source at $z = 0$ is shown together with its angular spectrum $\overline{E}(k_x, z = 0)$ calculated by the inverse of Eq. 1. The wave amplitudes $\overline{E}$ result from the Fourier-transformation of E with respect to the spatial coordinate x. As for the correlation between time and frequency domain, where a short optical pulse requires a broad frequency spectrum, a sharp field distribution requires a broad spectrum of spatial frequencies k_x.

Since only waves with limited spatial frequencies can propagate, the spectral width rapidly decreases with increasing distance from the source z leading to fast broadening of the electric field distribution in real space. In other words, propagation corresponds to low-pass filtering with frequency limit $k_{x,\max} = 2\pi/\lambda$. The far-field thus contains limited spatial frequencies equivalent to limited spatial information. To overcome this limitation different near-field concepts have been developed that are outlined in the following.

The key concept of SNOM is the probing of the sample near-field that contains the evanescent waves using a sharply pointed probe. Since evanescent waves decay rapidly for increasing distance to the source, the probe needs to be in close proximity to the sample. Waves with large kx components that carry high spatial information decay most rapidly following $e^{-|k_z|z}$ as can be seen from Fig. 1. Hence the spatial resolution obtained in an SNOM experiment drops fast with increasing tip-sample distance z. As for other scanning probe techniques that exploit short-ranged interactions, such as atomic force and scanning tunneling microscopy, AFM and STM, respectively, the lateral resolution is also determined by the lateral dimension of the probe. Two conceptually different types of probes can be distinguished: The first confines and samples electromagnetic fields using an aperture with sub-wavelength diameter (Fig. 2a). The second exploits the antenna concept that couples locally enhanced near-fields to propagating waves and vice versa (Fig. 2b–d). The two types, termed aperture and antenna probe, respectively, are illustrated in the following.



Scanning Near-Field Optical Microscopy, Fig. 1 Scheme illustrating the propagation of waves and the loss of spatial information. Initial field distribution $E(x, z = 0)$ at a 10-nm wide source in the x-z-plane (*center*) and corresponding angular spectrum $\tilde{E}(k_x, z = 0)$ (*right*). Near the source the spectrum contains both evanescent and propagating waves. Upper panels illustrate the evolution of the fields at $z = 30$ nm and $z = 90$ nm distance for a source wavelength of $\lambda = 500$ nm in vacuum. Only waves with $k_x \leq 2\pi/\lambda$ propagate. Evanescent waves decay exponentially following $e^{-|k_z|z}$. The decay of high spatial frequencies leads to spatial broadening and loss of spatial information in the far-field

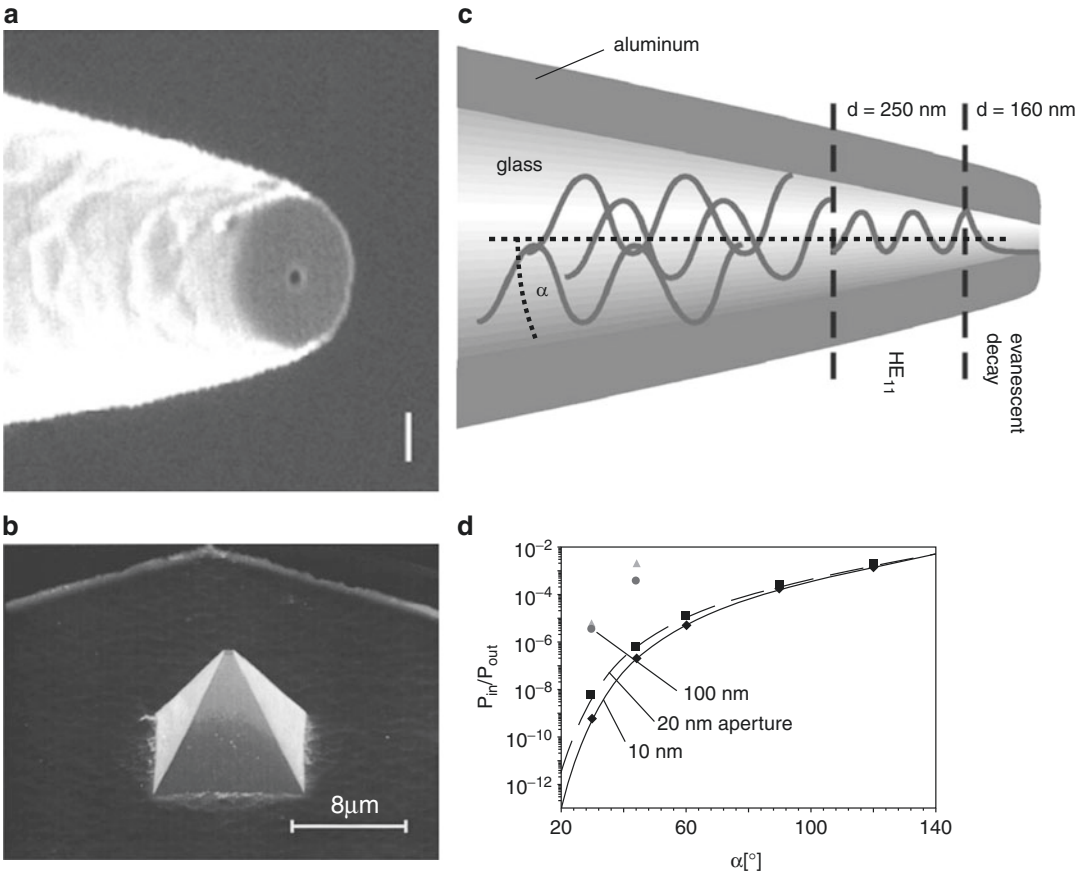


Scanning Near-Field Optical Microscopy, Fig. 2 Schematics of the most common SNOM probes and configurations: (**a**) Aperture probes are utilized in different modes: Near-field excitation – far-field collection ①, far-field excitation – near-field collection ② and near-field excitation – collection mode ③. (**b**) Tip-enhanced near-field optical microscopy (*TENOM*) using local field enhancement at a sharp metal antenna probe upon far-field excitation. (**c**) Tip-on-aperture (*TOA*) probe using local field enhancement at an antenna probe that is excited in the near-field of an aperture probe. (**d**) Scattering-SNOM (*s-SNOM*) based on far-field illumination and detection of local scattering at a sharp antenna probe. The label *d* indicates the structural parameter that determines the achievable spatial resolution

Aperture Probes

Aperture probes confine light by squeezing it through a sub-wavelength hole (Fig. 2a). This approach, termed aperture-SNOM, provides an enormous flexibility regarding signal formation. Different operation modes can be used that are capable of local sample excitation and/or local light collection. Depending on which step of the experiment exploits near-field interactions to obtain sub-wavelength resolution, aperture-SNOM can be implemented in excitation ①, collection ② and excitation-collection ③ mode (Fig. 2a).

The original scheme was proposed by Synge in 1928. He suggested to use a strong light source behind a thin, opaque metal film with a 100-nm

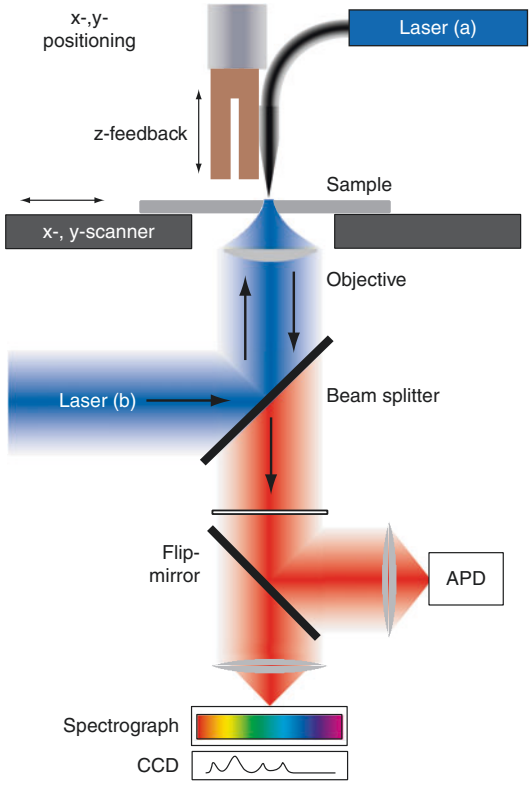


Scanning Near-Field Optical Microscopy, Fig. 3 (**a**) Scanning-electron microscopy (*SEM*) image of an aperture probe formed by a metal-coated tapered fiber with an aperture diameter of 70 nm (scale bar 200 nm) (Reprinted with permission from Veerman et al. [5]. Copyright 1999, John Wiley and Sons) (**b**) SEM image of a metallic hollow aperture probe microfabricated on a Si cantilever. The aperture diameter is about 130 nm (Reprinted with permission from Mihalcea et al. [6]. Copyright 1996, American Institute of Physics) (**c**) Schematic of the mode propagation in a tapered aperture probe. For probe diameters below the cut-off diameter, here $d = 160$ nm, the intensity decays exponentially toward the aperture. (**d**) Transmission of tapered probes determined as the ratio of input versus output power P_{in}/P_{out} as a function of the cone angle α defined in (**c**). For smaller cone angles, the distance between cut-off and aperture increases leading to extremely low transmission (**c** and **d**) (Reprinted with permission from Hecht et al. [7]. Copyright 2000, American Institute of Physics)

S

diameter hole in it as a very small light source. In 1984, two groups adopted this scheme and presented the first experimental realizations in the optical regime [1]. The aperture was formed at the apex of a sharply pointed transparent probe tip coated with metal. Raster-scanning the probe was made possible by the scanning technology developed in the context of scanning tunneling microscopy (STM). In Fig. 3a, a scanning electron microscopy (SEM) image of an aperture probe consisting of a metal-coated tapered glass fiber is shown. At the front surface a well-defined aperture with diameter of 70 nm is seen.

Analytical expressions for the electric field distribution in a sub-wavelength aperture in a metallic screen were already presented in 1944 and 1950 by Bethe and Bouwkamp. Numerical simulations for metal-coated tapered fiber probes show that strong fields pointing in axial direction occur at the rim due to local field-enhancement by the metal coating (see Fig. 4). The center of the aperture is dominated by a weaker horizontal



Scanning Near-Field Optical Microscopy, Fig. 4 Schematic of an experimental setup applicable to aperture-probe SNOM (laser (**a**), excitation mode) and TENOM (laser (**b**)) in case of transparent samples. In the case of TENOM the fiber probe would be replaced by an optical antenna, for example, an etched metal wire. The probe tip is positioned in the focus of the microscope objective by piezo-electric actuators. The sample is raster-scanned using a closed-loop *x*-*y*-scanner while both laser and probe position remain fixed. The tip-sample distance is controlled using a tuning-fork shear-force feedback scheme. The optical signal is collected by the objective and detected either by a highly sensitive avalanche photodiode (*APD*) or energetically resolved using a spectrograph and a CCD

component [2]. The optical field distribution can be determined experimentally by raster-scanning single fluorescent molecules that act as point-like dipoles across the aperture while recording the fluorescence intensity [5] (see section "Fluorescence Microscopy" and Fig. 3).

Tapered aperture probes suffer from low light transmission due to the cut-off of propagating wave-guide modes. For probe diameters below the cut-off diameter only evanescent waves remain and the intensity decays exponentially toward the aperture (Fig. 2c, d). Probe designs, therefore, aim at maximizing the cone angle that determines the distance between aperture and cut-off diameter. Hollow-cantilever probes feature relatively large cone-angles as compared to fiber-based probes (Fig. 2b). On the other hand, the input power needs to be limited because of the damage threshold of the metal coating in case of fiber-based probes. Due to the limited transmission and the skin-depth of the optical fields on the order of several tens of nanometers, most aperture-SNOM measurements are carried out with apertures of 50–100 nm.

Antenna Probes

Antenna probes act as transmitter and receiver coupling locally enhanced near-fields to propagating waves and vice versa (Fig. 2b–d) [8]. To distinguish this approach from the earlier implementations based on apertures, it is also termed apertureless-SNOM or a-SNOM. Antenna probes can be used in two different techniques: (1) Scattering type microscopy [9, 10], also termed scattering-SNOM or s-SNOM, in which the tip locally perturbs the fields near a sample surface. The response to this perturbation is detected in the far-field at the frequency of the incident light corresponding to elastic scattering (Fig. 2d). (2) Tip-enhanced near-field optical microscopy (TENOM) in which locally enhanced fields at laser-illuminated metal structures are used to increase the spectroscopic response of the system at frequencies different from that of the incident light [8, 11] (Fig. 2b, c). The flexibility of this technique allows the study of a variety of spectroscopic signals including Raman scattering (tip-enhanced Raman spectroscopy (TERS)), and fluorescence as well as time-resolved measurements. In the following, the signal formation in the case of optical antennas is sketched.

Elastic scattering signal. The near-field interaction can be treated within a simplified model in which the tip is replaced by a polarizable sphere. Due to the antenna properties of the tip, laser excitation with incident fiel E_i creates a dominating dipole oriented along the tip axis in z-direction normal to the sample surface. This dipole induces

a mirror dipole in the sample depending on its dielectric properties. The mirror dipole's field, decreasing with the third power of distance, interacts with the tip dipole. Solving the system of electrostatic equations that describes the multiple interaction between tip and mirror dipoles neglecting retardation yields an effective polarizability of the coupled tip-sample system which fully expresses the influence of the sample.

$$\alpha_{\text{eff}} = \left(\frac{(\alpha(1+\beta)/1-(\alpha\beta))}{\left(16\pi(\alpha+z)^3\right)} \right) \qquad (2)$$

The scattered field can then be calculated from $E_s \propto \alpha_{\text{eff}} E_i$ reflecting the short-ranged interaction required for sub-diffraction resolution. Since the laser illuminates a greater part of the tip and the sample, elimination of background-scattering contributions is crucial. Efficient background suppression can be achieved by demodulating the detected intensity signal at higher harmonics of the tapping mode frequency (see section "Instrumentation"). Besides the amplitude of the scattered field, its phase can be retrieved using interferometric heterodyne detection [9].

Raman scattering signal. In the case of Raman scattering, the total signal depends on the product of the excitation and emission rates $k^{\text{ex}}(\lambda^{\text{ex}})\, k^{\text{rad}}(\lambda^{\text{rad}})$. As a consequence, the total signal enhancement scales with the fourth power of the field enhancement for small differences between the excitation λ^{ex} and emission wavelength λ^{rad} and assuming that the field enhancement at the tip does not depend sensitively on the wavelength.

$$M_{\text{Raman}} = \frac{k^{\text{ex}}_{\text{tip}}}{k^{\text{ex}}_{0}} \cdot \frac{k^{\text{rad}}_{\text{tip}}}{k^{\text{rad}}_{0}} \approx f^4 \qquad (3)$$

The factor f measures the ratio between tip-enhanced E_{tip} and non-enhanced electric field E_0 in the absence of the tip. For the general case of surface-enhanced Raman scattering (SERS), Raman enhancement factors are reported reaching up to 12 orders of magnitude for particular multiple particle configurations involving interstitial sites between particles or outside sharp surface protrusions. Since the signal scales with the fourth power, already moderate field enhancement, predicted for a single spherical particle to be in the range of $f = 10$–100 is sufficient for substantial signal enhancement.

Fluorescence signal. The fluorescence intensity depends on the excitation rate k^{ex} and the quantum yield η denoting the fraction of transitions from excited state to ground state that give rise to an emitted photon. The quantum yield is expressed in terms of the radiative rate k^{rad} and the non-radiative rate k^{nonrad} through $\eta = k^{\text{rad}}/(k^{\text{rad}} + k^{\text{nonrad}})$. Accordingly, the fluorescence enhancement due to the presence of the metal tip can be written as

$$M_{\text{Flu}} = \left(\frac{E_{\text{tip}}}{E_0}\right)^2 \left(\frac{\eta_{\text{tip}}}{\eta_0}\right) = f^2 \left(\frac{\eta_{\text{tip}}}{\eta_0}\right). \qquad (4)$$

Here, it is assumed that the system is excited far from saturation. From Eq. 4 it is clear that TENOM works most efficiently for samples with small fluorescence quantum yield η_0 such as semiconducting single-walled carbon nanotubes [11]. For highly fluorescent samples such as dye molecules, the quantum yield η_0 is already close to unity and cannot be enhanced further.

Because of the small separation between emitter and metal tip required for high spatial resolution, non-radiative transfer of energy from the electronically excited state to the metal followed by non-radiative dissipation in the metal has to be taken into account. This process represents an additional competing non-radiative relaxation channel and reduces the number of detected fluorescence photons. Metal-induced fluorescence quenching can also be exploited for image contrast formation (see e.g., [12]). In this case M_{Flu} in Eq. 4 becomes smaller than unity.

While the theory of energy transfer between molecules and flat metal interfaces is well understood in the framework of phenomenological classical theory, nanometer-sized objects are more difficult to quantify [2]. Tip- and particle-induced radiative rate enhancement and quenching has been studied in literature both experimentally and theoretically. Experiments on model systems formed by single dipole emitters such as molecules and semiconductor nanocrystals and

spherical metal particles revealed a distance-dependent interplay between competing enhancement and quenching processes. While semiconducting tips cause less efficient quenching, they also provided weaker enhancement because of their lower conductivity at optical frequencies.

Polarization and angular-resolved detection of the fluorescence signal of single emitters demonstrated that the fluorescence rate enhancement provided by the optical antenna also results in a spatial redistribution of the emission [13]. The same redistribution can be expected to occur for tip-enhanced Raman scattering. The spatial distribution of the enhanced electric field follows approximately the outer dimensions of the tip apex. Since the signal enhancement scales with higher orders of field enhancement, the optical resolution can surpass the size of the tip [11]. Stronger fields and field confinement are observed for so-called gap modes formed by metal tips on top of metal substrates.

The scheme depicted in Fig. 2b shows that in addition to the signal resulting from the near-field tip–sample interaction the confocal far-field signal contribution is detected, representing a background. This background originates from a diffraction-limited sample volume that is far larger than the volume probed in the near-field. The near-field signal has to compete with this background, and strong enhancement is required to obtain clear image contrast. This requirement is relaxed in case of low-dimensional sample structures such as spatially isolated molecules or one-dimensional nanostructures [11]. The near-field signal to background ratio can be improved by exploiting the non-linear optical response of sample and tip. Examples include two-photon excitation of fluorescence using a metal tip antenna and the application of the four-wave mixing signal of a metal particle dimer as local excitation source.

Instrumentation

Near-field optical microscopy exploits short-ranged near-field interactions between sample and probe. SNOM instruments thus require a mechanism for tip-sample distance control working on the scale of nanometers. Typical implementations utilize other non-optical short-ranged probe–sample interactions such as force or tunneling current used for topography measurements in AFM and STM, respectively. During image acquisition by raster-scanning the sample with respect to the tip, optical and topographic data are thus obtained simultaneously. Due to the strong tip-sample distance dependence of the near-field signal, cross-talk from topographic variations is possible that can lead to artifacts in the optical contrast. Tip-sample distance curves need to be measured to prove unequivocally the near-field origin of the observed image contrast. Since the optical signal results from a single sample spot only, SNOM instruments are often based on a confocal laser scanning optical microscope equipped with sensitive photodetectors.

Scattering-SNOM is often implemented with intermittent-contact-mode AFM in which the tip-sample distance is modulated sinusoidally at frequencies typically in the range of 10–500 kHz. The elastically scattered laser light intensity is demodulated by the tapping-mode frequency using lock-in detection for background suppression.

Since Raman and fluorescence signals are typically weak requiring longer acquisition times, signal demodulation is more challenging. In the case of single-photon counting time-tagging can be used to retrieve the signal phase with respect to the tapping oscillation. While this can be applied to single-color experiments, spectrum acquisition using CCD cameras in combination with spectrometers is not feasible at typical tapping frequencies.

In most aperture-SNOM and TENOM experiments, the tip-sample distance is kept constant by either using STM, contact/non-contact AFM, or shear-force feedback. Sensitive piezoelectric tuning-fork detection schemes operating at small interaction forces have been developed and are used for fragile fiber and antenna probes [2].

Probe Fabrication

The near-field probe forms the crucial part of an SNOM setup since both optical and topographic

signals are determined by its short-ranged interactions with the sample. Instrument development, therefore, focuses mainly on the design of optimized tip concepts and tip geometries as well as on fabrication procedures for sharp and well-defined probes with high reproducibility. These continuing efforts benefit substantially from improving capabilities regarding nanostructuring and nanocharacterization, using, for example, focused ion beam milling (FIB) or electron beam lithography (EBL).

Aperture Probes

Optical fiber probes. Sharply pointed optical fiber probes were the first to provide sub-diffraction spatial resolution and are widely used as nanoscale light sources, light collectors or scatterers (Figs. 2a and 3a). Probe fabrication requires a number of steps starting with the formation of a tapered optical fiber. Typically two different methods are used. Chemical etching of a bare glass fiber dipped into hydro-fluoric (HF) acid yields sharp tips [4, 7]. The surface tension of the liquid forms a meniscus at the interface between air, glass, and acid. A taper is formed due to the variation of the contact angle at the meniscus, while the fiber is etched and its diameter decreased. Chemical etching allows reproducible production of larger quantities of probes in a single step. A specific advantage is that the taper angle can be tuned and optical probes with correspondingly large transmission coefficient can be produced.

The second method combines local heating using a CO_2 laser or a filament and subsequent pulling until the fiber is split apart. The resulting tip shapes depend heavily on the temperature and the timing of the heating and pulling, as well as on the dimensions of the heated area. The pulling method has the advantage of producing tapers with very smooth surfaces, which positively influences the quality of the evaporated metal layer. Etched probes, on the other hand, typically feature rough surfaces. Pulled fibers, however, have small cone angles and thus reduced optical transmission as well as flat end-faces limiting the minimum aperture size.

The aperture is formed during the evaporation of aluminum. Since the evaporation takes place under an angle slightly from behind, the deposition rate of metal at the apex is much smaller than on the sides. This geometrical shadowing effect leads to the self-aligned formation of an aperture at the apex.

The ideal aperture probe should have a perfectly flat end face to position a sample as close as possible into the near-field of the aperture. Conventional probes generally have a roughness determined by the grain size of the aluminum coating, which is around 20 nm at best. Due to the corrugated end face the distance between aperture and sample increases, which lowers the optical resolution and decreases the light intensity on the sample. Furthermore, these grains often obscure the aperture, which makes the probe ill-defined and not suited for quantitative measurements [4, 5, 7]. Subsequent focused ion beam (FIB) milling can be used to form high definition SNOM probes with well-defined end face (Fig. 3).

Microfabricated cantilevered probes. Hollow-pyramid cantilevered probes can be batch-fabricated with large taper angles. Lithographic patterning of an oxidized silicon wafer first defines the position of the aperture and the dimensions of the cantilever beam by structuring the oxide layer [6]. Anisotropic etching of the exposed silicon with buffered HF forms a pyramidal groove for the tip and trenches for the cantilever beam. After removing the oxide layer on the opposite side anisotropic etching is used to open a small aperture in the pyramidal groove. A 120-nm chromium layer is deposited on the back side forming the hollow pyramidal tip that is finally freed by isotropic reactive ion etching.

Si_3N_4 tips can be fabricated by dry etching in a CF_4 plasma and covering with thin aluminum films. Microfabricated probes based on quartz tips attached to silicon cantilevers have also been reported [2, 7]. First sharp quartz tips were produced in hydrofluoric acid followed by coating with thin films of aluminum and silicon nitride. Reactive ion etching was then used to selectively remove the silicon nitride from the tip apex, while the remaining film served as a mask for

wet-etching of the protruding aluminum in a standard Al-etching solution, leaving a small aperture on the apex of the tip. For light coupling, windows are etched into the backside of the levers at the position of the tip.

Microfabricated tips provide several advantages over fiber-based probes. The mechanical stability is typically increased and often sufficient to measure also in contact AFM mode without destroying the tip. A reproducible fabrication of the tips leads to a well-defined aperture shape. Large taper angles of around 70° shift the position of the mode cut-off closer to the aperture, resulting in higher transmission.

Antenna Probes

Sharp metal tip are fabricated in a single step by simple electrochemical etching. In the case of gold, pulsed or continuous etching in hydrochloric acid (HCL) or a mixture of HCL and ethanol routinely yields tips with diameters in the range of 30–50 nm. These tips can either be used in shear-force mode after gluing to the prong of a tuning-fork or in STM-mode. Silicon or silicon nitride cantilevered probes can be coated by a thin metal film through evaporation. Subsequent nanostructuring by FIB-milling can be used to optimize tip parameters.

Tip-on-aperture probes need to be fabricated in a series of sequential steps [14] (Fig. 2c). First, a fiber-based aperture probe is produced on which a well-defined end face is formed by FIB-milling. In the second step, a nanoscale tip is grown by electron beam–induced deposition of carbon. Next, the carbon tip is coated by a thin layer of chromium to improve adhesion of an aluminum layer evaporated during the final preparation step. Since the length of the tip can be controlled during electron beam deposition, TOA probes can be tailored to provide optimum antenna enhancement for a chosen wavelength [8].

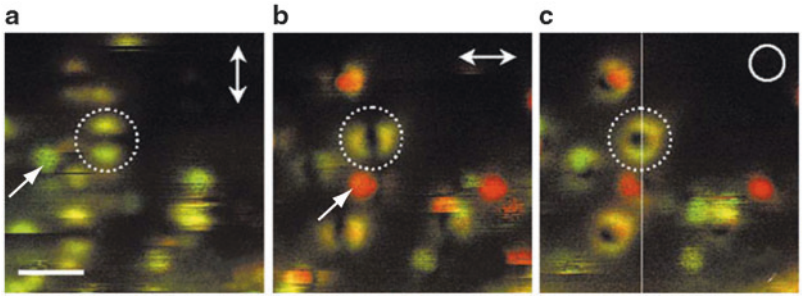
Applications in Nanoscience

The energy of light quanta – photons – is in the range of electronic and vibrational excitations of materials. These excitations are directly determined by the chemical and structural composition of matter. Optical spectroscopy, the energy-resolved probing of the material response to light exposure, thus provides a wealth of information on the static and dynamical properties of materials. Combining spectroscopy with near-field microscopy is particularly interesting since spectral information is obtained spatially resolved at the nanoscale. In the following several representative examples covering different material responses including fluorescence, Raman scattering and elastic scattering are briefly illustrated to highlight the capabilities of SNOM techniques.

Fluorescence Microscopy

Fluorescence measurements were among the first applications of SNOM. In these experiments, aperture-based probes were used to probe highly fluorescent dye-molecules on substrates. Aluminum-coated fiber tips were used for excitation while the fluorescence was collected in the far-field by a high-numerical aperture objective. A molecule is excited only if the optical electric field is polarized parallel to its transition dipole moment. The resulting fluorescence patterns rendered by a single molecule with known orientation can thus be used to visualize the local electric field distribution at the probe. Conversely, the molecular orientation can be determined for known field distributions. These experiments showed that the strongest electrical fields do not occur in the center of the aperture, but at the rims of the metal coating. This is the result of local field enhancement at the thin metal rim (see section "Antenna Probes"). Two lobes with strong fields oriented in axial direction occur located on opposite sides of the aperture in the direction of the polarization of the incident linearly polarized light. Molecules with a transition dipole moment oriented parallel to the tip axis are excited efficiently by these field components as can be seen in Fig. 5 [5]. Rotating the polarization of the incident linearly polarized light is seen to rotate the resulting double lobe pattern that indicates the area with strongest fields.

Fluorescence measurements typically do not require high excitation densities and in many cases, the small light transmission of aperture



Scanning Near-Field Optical Microscopy, Fig. 5 Series of three successive aperture-SNOM fluorescence images of the same area (1.2 × 1.2 μm) of a sample of dye molecules embedded in a thin transparent polymer film. The excitation polarization, measured in the far-field, was rotated from linear vertical (**a**) to linear horizontal (**b**) and then changed to circular polarization (**c**). Circular features marked by arrows result from molecules with transition dipole moments oriented parallel to the sample plane. The double lobe structure marked by the dashed circle results from a molecule with perpendicularly oriented transition dipole moment. This molecule senses the strong electrical fields that occur at the rim of the metal aperture at positions determined by the far-field polarization. Scale bar 300 nm (Reprinted with permission from Veerman et al. [5]. Copyright 1999, John Wiley and Sons)

probes represents no major drawback. In fact, fluorescence imaging of single dye molecules with 32 nm spatial resolution has been demonstrated using a microfabricated cantilevered glass tip covered with a 60-nm-thick aluminum film. Due to the thickness of the virtually opaque film, possible contributions from surface plasmons propagating on the outside of the film have been discussed [4]. Examples of fluorescence microscopy measurements based on aperture probes also include, for example, studies of single nuclear pore complexes and the kinetics of protein transport under physiological conditions.

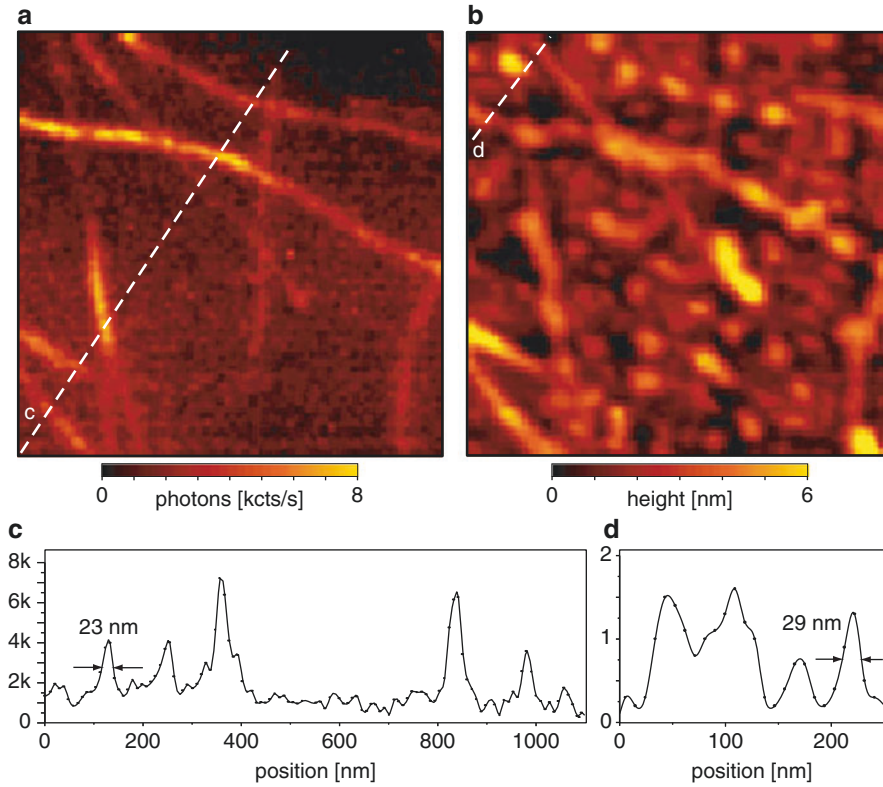
Optical antennas have been used for a variety of samples and materials including photosynthetic proteins, polymers, semiconductor quantum dots, and carbon nano-tubes [11]. The image contrast was based on local field enhancement provided by the tip. The spatial resolution achieved in these experiments was essentially determined by the diameter of the tip and ranged between 10 and 20 nm. Imaging can be combined with local spectroscopy to visualize emission energies on the nanoscale. Single-molecule experiments revealed the field distribution at the tip-antenna in analogy to the discussion made above for aperture probes [14].

While most of the reported studies were exploiting local signal enhancement, distance-dependent metal-induced quenching of fluorescence can also be used for high-resolution imaging. This approach provides sub 10 nm spatial resolution and has been applied to single fluorescent organic molecules and inorganic semiconductor nanorods (see e.g., [12]). In these experiments, the spectrally integrated fluorescence signal was demodulated by the tapping-mode frequency of the AFM cantilever after recording photon-arrival times.

Raman Microscopy

Raman scattering probes the unique vibrational spectrum of a sample and directly reflects its chemical composition and molecular structure. A main drawback of Raman scattering is the extremely low scattering cross-section which is typically 10–14 orders of magnitude smaller than the cross-section of fluorescence in the case of organic molecules. Raman measurements thus require higher laser intensities and in many cases the low transmission of aperture probes prohibits their application. The signal enhancement provided by the antenna tip in TENOM is substantial for the detection of nanoscale sample volumina. In the following, a review of selected examples is given to illustrate the possibilities of tip-enhanced Raman scattering (TERS) (see e.g., [11, 15]).

In Fig. 6, simultaneous near-field Raman and topographic imaging of individual single-walled



Scanning Near-Field Optical Microscopy, Fig. 6 Tip-enhanced Raman spectroscopy (*TERS*) of single-walled carbon nanotubes on glass. Simultaneous near-field Raman image (**a**) and topographic image (**b**). Scan area $1 \times 1\ \mu m^2$. The Raman image is acquired by detecting the intensity of the G′ band upon laser excitation at 633 nm. No Raman scattering signal is detected from humidity-related circular features present in the topographic image. (**c**) Cross-section taken along the *dashed line* in the Raman image indicating a spatial resolution around 25 nm. (**d**) Cross-section taken along the indicated dashed line in the topographic image. Vertical units are photon counts per second for c and nanometer for d (Reprinted with permission from Hartschuh et al. [16]. Copyright 2003, American Physical Society)

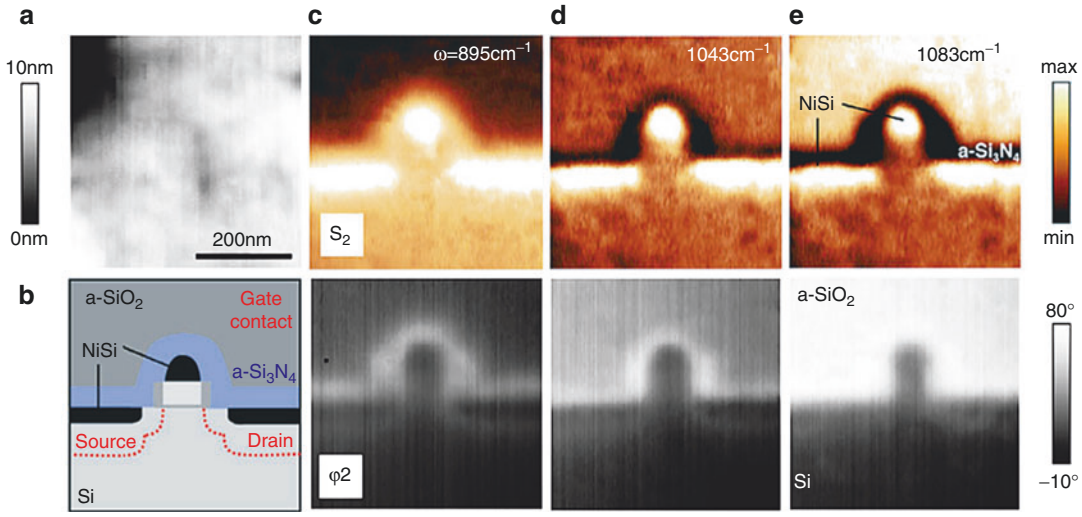
carbon nanotubes is shown. The optical image in (a) reflects the intensity of the G′ band, a particular Raman-active vibrational mode of carbon nanotubes. The optical resolution obtained in this experiment was about 25 nm as can be seen from the width of the peaks in the cross-section in Fig. 6c.

The strong fields required for sufficient enhancement of the Raman scattering signal can cause laser-induced decomposition and photochemical reactions in the presence of oxygen. TERS of single electronically resonant molecules has been demonstrated for ultra-high vacuum conditions. A review focusing on single-molecule surface- and tip-enhanced Raman scattering can be found in [15].

Elastic Scattering Microscopy

Elastic scattering SNOM probes the dielectric properties of the sample and has been used from the visible to the microwave regime of the electromagnetic spectrum. Reviews of the fundamentals of the technique and representative applications can be found in [9, 10]. The majority of s-SNOM experiments have been reported for the IR to THz spectral range. Applications include detection of the Mott-transition in nanodomains, mapping of the doping concentration in semiconductors, surface characterization with a sensitivity of a single monolayer, strain-field mapping, and infrared spectroscopy of a single virus.

As an example nanoscale infrared spectroscopic near-field mapping of single nano-



Scanning Near-Field Optical Microscopy, Fig. 7 Material-specific mapping of transistor components using s-SNOM: Cross-sectional images of a single transistor fabricated at the 65 nm technology node. (**a**) Topography. (**b**) Sketch of the transistor with materials indicated. (**c**–**e**) Near-field amplitude and phase images recorded at three different laser frequencies. Amorphous SiO_2 and Si_3N_4 render reversed optical contrast and are clearly distinguished. A spatial resolution better than 20 nm has been achieved (Reprinted with permission from Huber et al. [17]. Copyright 2010, IOP)

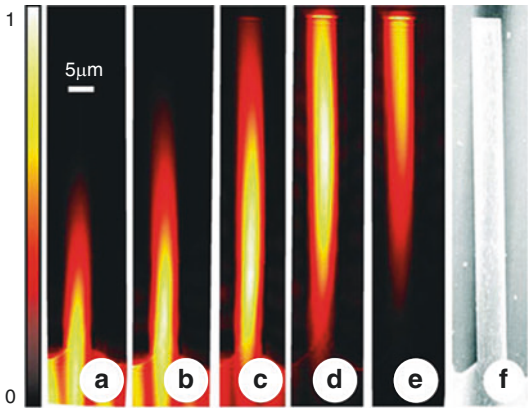
transistors is shown in Fig. 7. A cantilevered metallized Si-tip operating in tapping-mode with an oscillation frequency of 300 kHz and an amplitude of about 60 nm was used [17]. The data clearly demonstrates the potential of s-SNOM for infrared spectroscopic recognition of materials within individual semiconductor nanodevices.

Based on the antenna approach, s-SNOM typically provides 10–20 nm spatial resolution determined by the diameter of the tip-apex. In most of the s-SNOM experiments to date, monochromatic laser sources were used. Since only the optical response at this frequency is determined, the acquisition of scattering spectra or spectrally resolved images can only be obtained sequentially with a series of image scans at different laser frequencies. New developments exploiting broadband NIR laser sources aim at overcoming this limitation, and recently obtained a spectral bandwidth exceeding 400 cm^{-1}.

Plasmonics and Photonic Nanostructures

SNOM plays a vital role in the field of plasmonics which deals with the study of optical phenomena related to the electromagnetic response of metals [18]. Near-field optical probes are particularly important for two reasons. First, they provide a means to locally excite propagating surface plasmon polaritons (SPPs) in metal films, a process that is not possible in case of propagating light waves because of momentum (k-vector) mismatch. The broad k-spectrum associated with the near-field of the probe contains sufficient bandwidth for efficient SPP-excitation (see Fig. 1). Second, near-field probes can be used simultaneously to convert SPPs back into propagating waves, thereby probing the local distribution of electromagnetic fields in the vicinity of metallic nanostructures. As an example, the near-field associated with SPPs has been visualized along gold nanowires using an aperture probe in collection mode (see Fig. 2a, [4]). This approach is also termed photon scanning tunneling microscopy (PSTM) to illustrate the analogy between evanescent electromagnetic waves and the corresponding exponentially decaying electron wavefunctions within the tunnel barrier of an STM. PSTM has been widely used to spatially resolve light wave propagation also in dielectric photonic nanostructures [4].

Besides the visualization of static field distributions, optical spectroscopy also allows for the



Scanning Near-Field Optical Microscopy, Fig. 8 Phase-sensitive and ultrafast near-field microscopy of a surface plasmon polariton (*SPP*) waveguide. The local electric field is collected by an aperture-probe and detected interferometrically in a Mach-Zehnder–type configuration. (**a**–**e**) Normalized amplitude information of the SPP wavepacket E-field. Succeeding frames are new scans of the probe. In between the frames the delay line is lengthened to 14.4 μm. Therefore, the time between two frames is 48 fs. The scan frame is 15 × 110 μm^2, scan lines run from *top* to *bottom*. (**f**) Topography of the SPP waveguide obtained by shear-force feedback (Reprinted with permission from Sandtke et al. [19]. Copyright 2008, American Institute of Physics)

study of their temporal evolution and the propagation of pulses. Figure 8 illustrates ultrafast and phase-sensitive imaging of the plasmon propagation in a metallic waveguide by PSTM [19]. In this case the near-field microscope uses an aperture-probe in collection mode and incorporates a Mach-Zehnder–type interferometer enabling heterodyne time-resolved detection.

Scattering-SNOM with antenna tips has been used extensively to study localized surface plasmon polaritons (LSPP) in different metal nanostructures. By varying the laser excitation frequency near-field optical imaging allowed for distinguishing higher order plasmonic resonances [20].

Perspectives

During the last 25 years SNOM has demonstrated its capabilities for sub-wavelength optical imaging and spectroscopy of surfaces and sub-surface features. The strength of SNOM results from its enormous flexibility with respect to sample types as well as measurement configurations and in particular, from its combination with a broad range of spectroscopic techniques. Ongoing developments aim at increasing antenna efficiencies and new aperture-type schemes [18]. In addition, the combination of nano-optical approaches and ultrafast laser technique is explored to achieve enhanced light localization and the control of optical near-fields on the time scale of few optical cycles.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Confocal Laser Scanning Microscopy
- ▶ Light Localization for Nano-optical Devices
- ▶ Nanostructures for Photonics
- ▶ Scanning Tunneling Microscopy

References

1. Novotny, L.: The history of near-field optics. In: Wolf, E. (ed.) Progress in Optics, vol. 50, pp. 137–184. Elsevier, Amsterdam (2007)
2. Novotny, L., Hecht, B.: Principles of Nano-optics. Cambridge University Press, Cambridge (2006)
3. Kawata, S., Shalaev, V.M. (eds.): Advances in Nano-optics and Nano-photonics Tip Enhancement. Elsevier, Amsterdam (2007)
4. Kawata, S., Shalaev, V.M. (eds.): Handbook of Microscopy for Nanotechnology. Kluwer, Dordrecht (2006)
5. Veerman, J.A., Garcia-Parajo, M.F., Kuipers, L., van Hulst, N.F.: Single molecule mapping of the optical field distribution of probes for near-field microscopy. J. Microsc. **194**, 477–482 (1999)
6. Mihalcea, C., Scholz, W., Werner, S., Münster, S., Oesterschulze, E., Kassing, R.: Multipurpose sensor tips for scanning near-field microscopy. Appl. Phys. Lett. **68**, 3531–3533 (1996)
7. Hecht, B., Sick, B., Wild, U.P., Deckert, V., Zenobi, R., Martin, O.J.F., Pohl, D.E.: Scanning near-field optical microscopy with aperture probes: fundamentals and applications. J. Chem. Phys. **112**, 7761–7774 (2000)
8. Bharadwaj, P., Deutsch, B., Novotny, L.: Optical antennas. Adv. Opt. Photon. **1**, 438–483 (2009)

9. Keilmann, F., Hillenbrand, R.: Near-field microscopy by elastic light scattering from a tip. Philos. Trans. R. Soc. Lond. A **362**, 787–805 (2004)
10. Bründermann, E., Havenith, M.: SNIM: scanning near-field infrared microscopy. Annu. Rep. Prog. Chem., Sect. C: Phys. Chem. **104**, 235–255 (2008)
11. Hartschuh, A.: Tip-enhanced near-field optical microscopy. Angew. Chem. Int. Ed. **47**, 8178–8198 (2008)
12. Ma, Z., Gerton, J.M., Wade, L.A., Quake, S.R.: Fluorescence near-field microscopy of DNA at sub-10 nm resolution. Phys. Rev. Lett. **97**, 260801–260804 (2006)
13. Taminniau, T.H., Stefani, F.D., Segerink, F.B., van Hulst, N.F.: Optical antennas direct single-molecule emission. Nat. Photonics **2**, 234–237 (2008)
14. Frey, H.G., Witt, S., Felderer, K., Guckenberger, R.: High resolution imaging of single fluorescent molecules with the optical near field of a metal tip. Phys. Rev. Lett. **93**, 200801–200804 (2004)
15. Pettinger, B.: Single-molecule surface-and tip-enhanced Raman spectroscopy. Mol. Phys. **108**, 2039–2059 (2010)
16. Hartschuh, A., Sánchez, E.J., Xie, X.S., Novotny, L.: High-resolution nearfield Raman microscopy of single-walled carbon nanotubes. Phys. Rev. Lett. **90**, 095503–4 (2003)
17. Huber, A.J., Wittenborn, J., Hillenbrand, R.: Infrared spectroscopic near-field mapping of single nanotransistors. Nanotechnology **21**, 235702–6 (2010)
18. Schuller, J.A., Barnard, E.S., Cai, W., Jun, Y.C., White, S.W., Brongersma, M.L.: Plasmonics for extreme light concentration and manipulation. Nat. Mater. **9**, 193–204 (2010)
19. Sandtke, M., Engelen, R.J.P., Schoenmaker, H., Attema, I., Dekker, H., Cerjak, I., Korterik, J.P., Segerink, F.B., Kuipers, L.: Novel instrument for surface plasmon polariton tracking in space and time. Rev. Sci. Instrum. **79**, 013704–10 (2008)
20. Dorfmüller, J., Vogelgesang, R., Khunsin, W., Rockstuhl, C., Etrich, C., Kern, K.: Plasmonic nanowire antennas: experiment, simulation, and theory. Nano Lett. **10**(9), 3596–3603 (2010)

Scanning Probe Microscopy

► Atomic Force Microscopy

Scanning Surface Potential Microscopy

► Kelvin Probe Force Microscopy

Scanning Thermal Microscopy

Li Shi
Department of Mechanical Engineering,
The University of Texas at Austin, Austin,
TX, USA

Synonyms

Microthermal analysis; Scanning thermal profiler

Definition

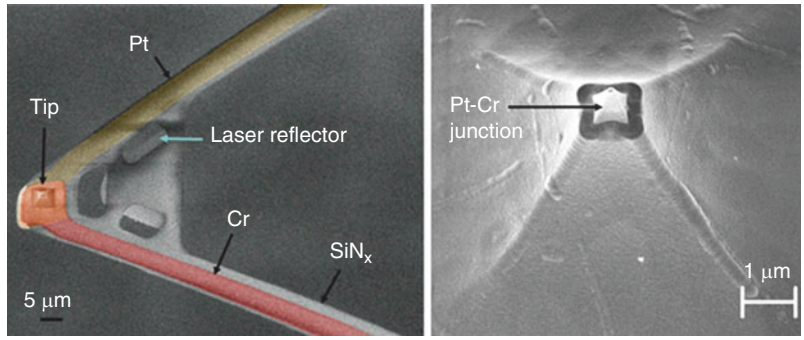
Scanning Thermal Microscopy (SThM) is a class of experimental methods for high spatial resolution mapping of the surface temperature distribution of an operating device or the thermal property variation of a structure with the use of a sensor fabricated on a scanning probe.

Thermal Probes

In 1986, Williams and Wickramasinghe [1] pioneered a so-called scanning thermal profiler technique based on a thermocouple sensor fabricated at the end of a probe tip for Scanning Tunneling Microscopy (STM). The thermocouple sensor consisted of two dissimilar conductors that made a junction at the end of the STM tip. An insulator separated the two conductors in all areas remote from the tip. The size of the thermocouple junction could be made as small as 100 nm. When a temperature difference (ΔT) exists between the thermocouple junction and the ends of the two lead wires, thermal diffusion of electrons through the two dissimilar wires results in a thermoelectric voltage between the two lead wires. The magnitude of the thermovoltage measured with the use of a voltmeter is

$$V = (S_1 - S_2)\Delta T \quad (1)$$

where S_1 and S_2 are the Seebeck coefficient of the two thermocouple wires.



Scanning Thermal Microscopy,
Fig. 1 Scanning electron micrographs of a SiN_x AFM cantilever probe (**a**) with a sub-micron Pt-Cr thermocouple junction (**b**) formed at the apex of the SiO_2 tip

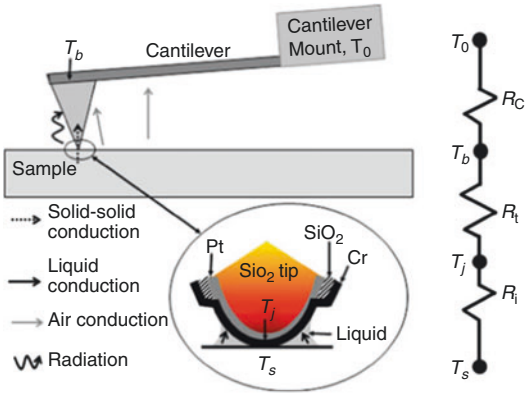
The main purpose of this first scanning thermal profiler was not for mapping temperature distribution of a surface but to regulate the tip-sample distance using the relationship between the thermocouple tip temperature and the distance between the tip and the sample. The work of Williams and Wickramasinghe stimulated intense efforts to develop a scanning probe for high spatial resolution mapping of the temperature distribution or thermal properties on a surface. In 1993, Majumdar and co-workers [2] introduced a wire thermocouple probe that could be used in an atomic force microscope (AFM) for simultaneous mapping of topography and temperature. Since then, different designs of scanning thermal probes have been fabricated. A common feature of these thermal probes is a thermal sensor fabricated at the end of an AFM or STM tip. The thermal sensor can be a thermocouple, a resistance thermometer, a Schottky diode, or a fluorescence particle [3–5].

As discussed above, a thermocouple measures the temperature difference between the junction of its two constituent metals and the other ends of the two metal wires. The current–voltage (I-V) characteristics of a Schottky junction depends on the temperature, and can thus serve as a local temperature sensor. A resistance thermometer is usually made of a metal or degenerately doped semiconductor with a relatively large and constant temperature coefficient of resistance. Because the resistance of a nanoscale resistor can be too small for sensitive electronic detection, resistance thermometer cannot be miniaturized as readily as a thermocouple sensor. Another approach is to use the emission band of some fluorescence particles, which shifts with increasing temperature. This feature has been used for making a SThM probe with a fluorescence particle located at the end of a scanning probe tip [5]. Besides these temperature-sensing techniques, the thermally induced bending of a biomaterial AFM cantilever has been used for temperature measurements [4]. Moreover, the thermal expansion of a sample has been measured with a regular AFM tip, and used to infer the sample temperature [4].

Different methods have been reported for fabricating different SThM probes. Some of the methods use sequential (or one at a time) fabrication methods [4], whereas batch fabrication processes have been developed for wafer scale fabrication of the SThM probes [6]. While the probe batch fabrication processes are somewhat similar to those that have been developed for the manufacturing of ultra large scale integrated (USLI) devices, they often involve additional etching steps to make free standing cantilever probes, as well as novel processes to form a thermal sensor such as a thermocouple at the end of the cantilever probe tip. Figure 1 shows a batch-fabricated thermocouple SThM cantilever probe that consists of a 0.5 μm thick low stress silicon nitride (SiN_x) cantilever, a 8 μm tall silicon dioxide (SiO_2) pyramid tip with a $\sim$ 20 nm tip radius, and a sub-micron Pt-Cr thermocouple junction formed at the apex of the SiO_2 tip [6]. Several other unique fabrication processes have been developed for wafer scale fabrication of a thermocouple junction, Schottky diode, Pt-C or doped silicon resistance thermometer [4].

Special probe holders with necessary electrical contacts or optical access can allow for the use of the SThM sensor probes in a commercial AFM or STM. STM requires the sample surface to be

conducting because the tip-sample gap is controlled based on the tunneling current. This issue has limited the use of SThM probes in STM. In comparison, AFM with a cantilever SThM probe can allow simultaneous topography and thermal mapping of both conducting and non-conducting samples. In a typical contact-mode AFM, the tip-sample spacing is regulated by the force acting on the probe tip. The tip-sample interaction force can be obtained from the AFM cantilever bending that is measured with the use of several methods. One of these methods is to measure the laser beam reflected by the cantilever with the use of a position-sensitive detector. Another method is to employ a built-in piezoresistive sensor in the cantilever to measure the cantilever bending. Although the optical detection method provides superior sensitivity, the laser heating of the thermal sensor at the end of the cantilever needs to be minimized.



Scanning Thermal Microscopy, Fig. 2 Schematic diagram showing the heat transfer mechanisms between the thermal probe and the sample as well as a thermal resistance circuit where the heat transfer through the air gap is ignored (Reproduced from [6] with permission by ASME)

Theoretical Analysis of Heat Transfer Mechanisms

For many of the various SThM sensor probes, the sensitivity and spatial resolution of the thermal imaging technique depend on the mechanisms of heat transfer between the thermal probe and the sample. For SThM measurements conducted in air, heat is transferred between the probe and the sample via conduction through the solid–solid contact, a liquid meniscus at the tip-sample contact, and the air gap between the probe, and via radiation, as illustrated in Fig. 2 heat transfer through the air gap and via radiation is not localized at the tip-sample contact, and can deteriorate the spatial resolution of the SThM probes.

For a circular contact area with radius b, the spreading thermal resistance of the sample is obtained as [7]

$$R_s = \frac{1}{4\kappa_s b} \tag{2}$$

where κ_s is the thermal conductivity of the sample. For $\kappa_s \approx 5$ W/m-K, $b \approx 30$ nm, $R_s \approx 1.7 \times 10^6$ K/W.

The thermal interface resistance through the solid–solid contact is given as [8]

$$R_{i,s} \approx \frac{4}{\alpha C_s v_s \pi b^2} = \frac{4K}{3\alpha\kappa_s \pi b} \tag{3}$$

where C_s and v_s are the specific heat and phonon group velocity of the sample, α is the phonon transmission coefficient from the sample into the tip, $K \equiv l_s/b$ is the Knudsen number, and l_s is the phonon mean free path in the sample. The second equation is obtained with the use of the kinetic theory, $\kappa_s = C_s v_s l/3$ [8]. For $\alpha = 0.3$, $\kappa_s = 5$ W/m-K, $b = 30$ nm, and $ls = 5$ nm, $R_{i,s} = 1.5 \times 10^6$ K/W.

A liquid meniscus often forms at the tip-sample junction when the tip is scanned on the sample in air. The interface thermal resistance through the liquid meniscus ($R_{i,\,l}$) at the tip-sample junction depends on the humidity and surface properties. The value of $R_{i,\,l}$ has been estimated by Majumdar to be on the order of 10^5 K/W [4], which is lower than the solid-solid thermal interface resistance. The total interface tip-sample thermal resistance is

$$R_i = \left(R_{i,s}^{-1} + R_{i,l}^{-1}\right)^{-1} \tag{4}$$

The spreading thermal resistance of the conical tip can be estimated as

$$R_t \approx \frac{1}{\pi \kappa_t b \tan\theta} \tag{5}$$

where κ_t is the thermal conductivity of the tip and θ is the half angle of the conical tip. For $\kappa_t \approx 5$ W/m-K, $b = 30\,\text{nm}, \theta = \pi/8, R_t = 5 \times 10^6$ K/W. Hence, if the thermocouple junction can be made as small as the contact size, the high spreading thermal resistance of the tip can be utilized to thermally isolate the junction from the cantilever.

An approximation expression for the tip-sample air thermal conductance per unit area is given here as

$$g_a \approx \frac{\kappa_a}{z + l_a} \tag{6}$$

where κ_a and l_a are the thermal conductivity and mean free path of air molecules, and z is the air gap. In the ballistic limit of $z \ll l_a$, this expression is reduced to $g_a = \kappa_a/l_a$. Based on the kinetic theory, $\kappa_a = C_a v_a l_a/3$, where C_a and v_a are the specific heat and velocity of air molecules, so that $g_a = C_a v_a/3$ for the ballistic limit. This result is close to the ballistic thermal conductance per unit area [8], $C_a V_a/4$, for heat transfer of gas molecules between two close parallel plates when the thermal accommodation coefficient is unity, corresponding to the case that the scattered molecules take the temperature of the surface. In the diffusive limit of $z \gg l_a, g_a = \kappa_a/z$ is the diffusive thermal conductance per unit area of the air gap. The total tip-sample thermal conductance through the air gap can be obtained as

$$G_{g,t-s} = \int_b^{b+H\tan\theta} \frac{\kappa_a}{l_a + (r-b)/\tan\theta} 2\pi r dr = 2\pi\kappa_a H \tan^2\theta \left[1 + \frac{b - l_a\tan\theta}{H\tan\theta} \ln\frac{l_a + H}{H}\right] \tag{7}$$

where H is the tip height. For $H = 8\,\mu\text{m}$ and $\theta = \pi/8$, $Gg,\ t\text{-}s \approx 2.2 \times 10^{-7}$ W/K, corresponding to a resistance $R_{g,\,t\text{-}s} \equiv 1/G_{g,\,t\text{-}s} = 4.4 \times 10^6$ K/W [6].

Far-field radiation conductance between the tip and the sample is approximated as that of a two-surface enclosure as

$$G_{\text{rad},t-s} \approx \frac{4A_t T^3}{(1-\varepsilon_t)/\varepsilon_t + 1/F_{t-s} + A_t(1-\varepsilon_s)/A_s\varepsilon_s} \tag{8}$$

where σ is the Stefan–Boltzmann constant, A_t and A_s are the surface areas of the tip and sample, ε_t and εs are the surface emissivity of the tip and sample, and F_{t-s} and T are the view factor and the average temperature between the tip and sample. When the temperature is close to 300 K, G_{rad}, $t - s$ is on the order of 1×10^{-10} W/K. Both the far-field and near-field radiation heat transfer between the tip and the sample can be ignored compared to those via the air gap, solid–solid interface, and liquid meniscus [4], unless for a very high tip or sample temperature.

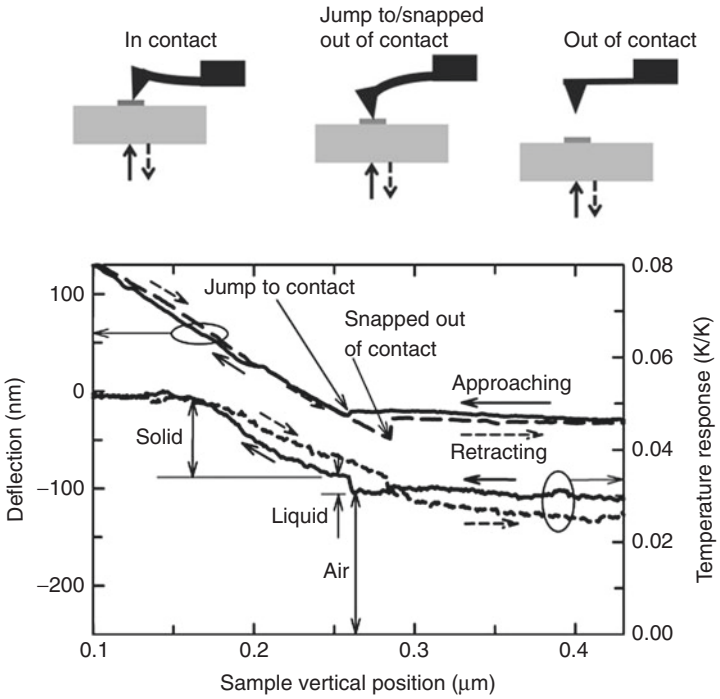
Because of heating by the sample surface, the temperature of the air molecules surrounding the cantilever, T_g, can be different from the temperature at the cantilever mount, T_0, which is usually close to the room temperature. In this case, heat transfer from the tip into the cantilever can be obtained as the fin heat transfer rate expressed as [7]

$$q_{t-c} = M\frac{\cosh mL - \theta_0/\theta_b}{\sinh mL} \tag{9}$$

where $M = (hP\kappa_c A_c)^{1/2}\theta_b$, $m = (hP/\kappa_c A_c)^{1/2}$, h is the heat transfer coefficient between the cantilever and the surrounding gas, κ_c and A_c are the thermal conductivity and cross-section of the cantilever, P is the perimeter of the cantilever cross-section, $\theta_0 = T_0 - Tg$, $\theta_b = T_b - T_g$, and T_b is the temperature at the joint between the tip and the cantilever. When only a small sample area is at a temperature different from the room temperature so that $T_g \approx T_0$, the thermal resistance of the cantilever can be defined as

$$R_c = \frac{T_b - T_0}{q_{t-c}} = \frac{\tanh mL}{\sqrt{hp\kappa_c A_c}} \tag{10}$$

For $\kappa_c \approx 5$ W/m-K, and 0.5 μm by 10 μm cantilever cross-section, $L = 150$ μm, $h \approx k_a(\pi/A_c)^{1/2}$, $R_c = 3 \times 10^5$ K/W.



Scanning Thermal Microscopy, Fig. 3 Cantilever deflection and temperature response of a thermocouple probe as a function of the sample vertical position when a 350-nm wide heater line sample was raised toward and then retracted from the tip (Reproduced from [6] with permission by ASME)

On the other hand, when a large area of the sample is at a temperature higher than T_0, the air molecules surrounding the cantilever can be heated by the sample surface to a temperature T_g that is considerably higher than T_0. In this case, the cantilever is heated by the surrounding air, resulting in an additional heat transfer path from the sample to the probe.

Experimental Characterization of Heat Transfer Mechanisms

Experimentally, the heat transfer path through the air gap between the sample and the cantilever was found to be significant when the size of the heated zone on the sample surface was not small [6]. In the experiment [6], a 350-nm wide metal line was joule heated to 5.3 K above room temperature. The cantilever deflection and temperature rise of the thermocouple junction were recorded simultaneously when the sample was approached and then retracted from a thermocouple SThM probe tip. When the sample approached the tip, the cantilever deflection signal remained approximately constant before the sample contacted the tip, as shown in the deflection curve in Fig. 3. In this region, the junction temperature rise was caused by air conduction between the probe and the sample. As the tip-sample distance was reduced, the junction temperature rise due to air conduction increased slowly. Before the sample made solid–solid contact to the tip, the adsorbed liquid layers on the tip and the sample bridged each other. Initially, this liquid bridge pulled the tip down by a van der Waals force, as being seen in the dip labeled as "jump to contact" in the deflection curve. Coincidentally, there was a small jump in the junction temperature due to heat conduction through the liquid bridge. As the sample was raised further, both the solid–solid contact force and the junction temperature increased gradually, until the cantilever was deflected for more than 100 nm. After this point, the junction temperature remained almost constant as the contact force increased.

As the sample was retracted from the tip, the junction temperature remained almost constant until at a cantilever deflection of 100 nm, after which the junction temperature rise decreased

roughly linearly but at a smaller slope than that found in the approaching cycle. As the sample was lowered further, the tip was pulled down together with the sample by surface tension of the liquid bridge until after a certain point, the restoring spring force of the cantilever exceeded the surface tension, and the tip "snapped out of contact" with the sample. Associated with the breaking of the liquid bridge, there was a small drop in the junction temperature.

This experiment shows several mechanisms. First, before the tip contacted the sample, air conduction contributed to a junction temperature rise up to 0.03 K per K sample temperature rise, which was about 60 % of the maximum junction temperature rise at the maximum contact force of the experiment. Second, conduction through a liquid meniscus was responsible for the sudden jump and drop in junction temperature when the tip "jumped to contact" to and "snapped out of contact" from the sample, respectively. Third, solid–solid conduction resulted in the almost linear relationship between the junction temperature rise and the contact force. This is a well understood feature for macroscopic solid–solid contacts. When the contact force was increased further, the junction temperature approached a constant likely because the contact size approached the maximum possible value limited by the diameter of an asperity located at the tip-sample interface.

The point contact experiment was repeated for a 5.8 μm wide and 2,000 μm long heater line on an oxidized silicon wafer. While the increase of the normalized junction temperature due to solid and liquid conduction was similar in magnitude to those observed in the narrower line, the normalized junction temperature rise due to air conduction was one order of magnitude higher than the corresponding one in the narrower line. In addition, the temperature rise of the thermocouple junction was measured when the tip was in contact with three heater lines of different line widths. It was found that the temperature rise at the thermocouple junction of the tip was about 53 %, 46 %, and 5 % of that of a 50, 3, and 0.3 μm wide line, respectively, showing a trend of increasing temperature rise in the SThM probe with increasing heater line width. These results suggest that the cantilever was heated more by air conduction between the tip and the larger hot area of the wider heater lines. This trend indicates that air conduction plays an important role in probe-sample heat transfer, especially when the hot area on the sample surface is large in comparison to the tip size.

The unwanted heat transfer through the air gap can be eliminated by conducting SThM measurements in vacuum. In this case, the tip-sample air conductance $G_{g,\,t-s} = 0$, and the thermal resistance of the cantilever is the conduction thermal resistance given as

$$R_c = \frac{L}{\kappa_c A_c} \tag{11}$$

For a cantilever length $L = 150$ μm, thermal conductivity $\kappa_c = 5$ W/m-K, and cross-section $A_c = 0.5\ \mu\text{m} \times 10\ \mu\text{m}$, $R_c = 6 \times 10^6$ W/K based on the above equation. In addition, the heat transfer from the sample to the tip in vacuum can be described with the use of the thermal resistance circuit shown in Fig. 2.

However, vacuum operation of the SThM probe is considerably much more challenging than operation in open air environment. Moreover, the liquid meniscus at the tip-sample junction is also eliminated by the vacuum environment. Although this can further enhance the spatial resolution of the thermal imaging technique, the tip-sample thermal interface conductance could be reduced considerably. Consequently, the thermocouple junction temperature at the tip end could become rather different from the sample temperature, resulting in low sensitivity and large uncertainty in the measured temperature.

An alternative method to eliminate the unwanted air conductance while keeping the heat transfer via the liquid meniscus is based on a dual scan measurement approach [9]. In this approach, after a line scan of the SThM probe on the sample surface in the contact mode, the probe is scanned on the same line in the lift mode with a small fixed tip-sample distance between the tip and the sample. The difference between the two thermal sensor signals obtained in the contact mode and the lift

mode is attributed to the local heat transfer through the solid–solid contact and the liquid meniscus at the tip-sample junction, and can be used to obtain the local sample surface temperature at the contact point, as discussed below.

Surface Temperature Mapping

When the heat transfer through the probe-sample gap is eliminated by the dual scan technique, the difference in the measured thermocouple junction temperature values in the contact mode and the lift mode is still different from that for vacuum measurement given by the thermal resistance circuit of Fig. 2 as

$$T_j = (T_s - T_0)\left[1 + \frac{R_i}{R_c + R_t}\right]^{-1} + T_0 \qquad (12)$$

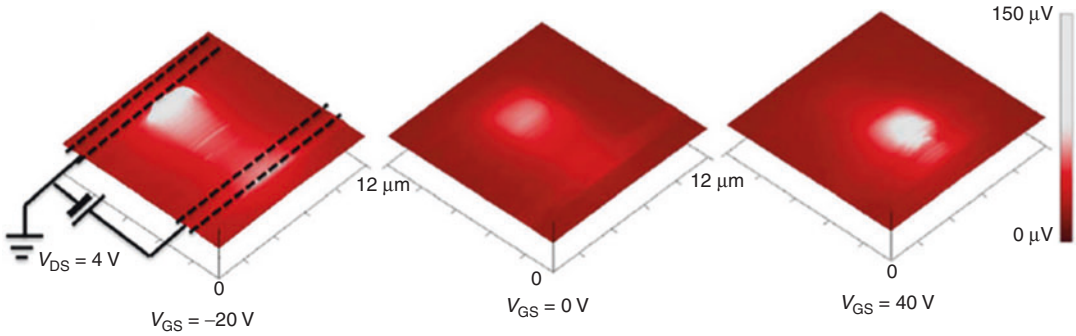
where T_s is the local sample temperature at the contact point, T_0 is the ambient temperature.

It is important to increase R_t and R_c relative to R_i and R_s, so as to maximize ΔT_j and minimize the sample temperature change due to the contact by the tip. For this reason, it is desirable to use low-thermal conductivity dielectrics such as SiO_2 and SiN_x to fabricate the pyramid tip and the cantilever, and low-thermal conductivity conductors such as Cr and Pt to fabricate the thermocouple sensor. The geometry of the sensor, tip, and cantilever also needs to be optimized to minimize heat loss.

Even for state-of-the-art thermocouple SThM probes, R_t and R_c are still not much larger than R_i. Hence, the ΔT_j obtained from the differential dual scan measurement needs to be calibrated in a separate experiment, where the sample is heated with an external heater and the sample surface temperature is determined using a different method, such as a resistance thermometer fabricated on the sample surface [10].

SThM methods operated in the contact mode or the dual scan mode have been employed to measure the surface temperature distribution of silicon, carbon nanotube, and graphene nanoelectronic devices, with a spatial resolution better than 100 nm [10, 11]. The very high electric field in electrically biased nanoelectronic devices can result in non-equilibrium transport of electrons, optical phonons, and acoustic phonons [12]. Energetic hot electrons accelerated in the high field lose their energy to lattice vibration via emission of high frequency optical phonons. The optical phonons are further scattered with acoustic phonons that make a larger contribution to the lattice thermal conductivity because of a much higher group velocity than the optical phonons. The scattering time between optical phonons and acoustic phonons can be much longer than that between hot electrons and optical phonons in nanoelectronic devices made of silicon and carbon nanotubes. Consequently, energy transfer from optical phonons and acoustic phonons becomes a bottle neck in the heat dissipation process. This bottle neck can result in different temperatures for acoustic phonons and optical phonons. The local temperatures of different energy carriers in operating nanoelectronic devices can be measured with different methods. For example, infrared spectroscopy has been used to measure the electronic temperature of electrically biased graphene devices by fitting the infrared emission from the hot electrons in the device with the Planck distribution [13]. The intensity ratio between the anti-Stokes peak to the Stokes peak in the Raman spectrum can be used to probe the local temperature of the zone center or zone boundary optical phonons that are active in the Raman scattering processes [14]. Complementary to these optical thermometry methods, the thermocouple SThM probe can be used to map the low frequency phonon temperature distribution because the phonon transmission coefficient at the tip-sample interface increases with decreasing phonon frequency. This capability has been employed to observe bias-dependant and asymmetric acoustic phonon temperature distribution in electrically biased graphene with a superior spatial resolution of about 100 nm [10], as illustrated in Fig. 4.

Besides the dual scan method, an active SThM method has been reported by Nakabeppu and Suzuki [15] to obtain quantitative temperature map of a sample surface in vacuum. Their probe for active SThM consists of two thin film



Scanning Thermal Microscopy, Fig. 4 Measured SThM thermovoltage maps of an electrically biased graphene device for a constant drain-source bias ($V_{DS} = 4$ V) and different gate-source voltage of −20 V, 0 V, and 40 V, respectively. The scan size is 12 μm × 12 μm

thermocouples (TCs) and a micro-heater fabricated on a cantilever. When the cantilever was scanned on a heated sample, heat flow from the sample into the cantilever resulted in a temperature drop (ΔT) along the cantilever, which was measured using the differential thermocouple. The micro-heater power was adjusted by using a feedback loop until the differential thermocouple reads $\Delta T = 0$. Under this condition, the heat flow from the sample to the cantilever was zero so that the measured tip temperature is the same as the sample surface temperature. Alternatively, the active SThM method can also be implemented using the dual scan operation in air [9]. In this implementation, the SThM probe is heated to different temperatures and is scanned on a hot sample at each heating rate for the probe. When the measured sensor temperature of the heated SThM probe is the same during the contact-mode and lift-mode scanning of the probe, the sensor temperature is taken as the same as the local sample temperature, and the heat flow between the tip and the sample is nullified.

It is worth emphasizing that the key issue in SThM measurements is that the thermal sensor temperature can be quite different from the sample temperature. This issue was addressed in the aforementioned measurements via either a detailed calibration or actively heating the sensor until the sensor temperature was the same as the sample temperature. While four-point measurement methods with minimum current leakage into the voltage measurement devices are commonly used to address problems caused by contact electrical resistance in electrical measurements, a thermal analogue of this approach has not been effective for addressing the problem caused by thermal contact resistance in SThM measurements, because it is difficult to eliminate heat loss into a temperature measurement device, or the thermal probe in SThM.

Thermal Property Mapping

Scanning thermal probes consisting of a resistance heater and thermometer at the tip end have been used to measure the local thermal property near the surface in a solid sample. For this measurement, the temperature of the heated probe tip decreases as the tip touches the sample. A larger drop in the tip temperature is expected when the tip touches a higher thermal conductivity region of the sample because of a smaller spreading thermal resistance in the sample. Alternatively, the probe heating rate can be feedback-controlled to maintain a constant tip temperature during tip scanning. In this case, a higher heating power is needed to achieve a constant tip temperature when the tip touches a higher thermal conductivity region. Although thermal conductivity contrast has been observed under dc operations, accurate thermal conductivity measurement remains a challenge especially for low-thermal conductivity samples [4]. This is in part because a significant

fraction of the power may be dissipated through the cantilever instead of into the sample. The heat loss through the cantilever can be reduced and the spatial resolution can be improved by using microfabricated probes with a low-thermal conductivity cantilever and a sharp pyramid Si resistance heater and thermometer tip. However, with a decreased contact radius b, the tip-sample thermal interface resistance increases according to $1/b^2$, and can dominate the spreading thermal resistance of the sample that scales as $1/b$. The large interface thermal resistance can make it difficult to quantify the sample thermal property when the tip size is small.

The resistance thermometer probe can also be operated in the ac mode, as shown by Pollock, Hammiche, and coworkers [3]. The sensor probe used in their measurements was 5-μm-diameter Pt resistance wire bent to form a tip at the bent. An ac electrical heating current was used to modulate the probe temperature by 5 °C at 10 kHz around a dc temperature of 40 °C, the measured amplitude and the phase shift in the ac voltage drop across their Pt wire thermometer probe depended on the thermal properties of the sample. One advantage of measuring the phase lag is that it is independent of the temperature dependence of the electrical resistivity of Pt wire and the power input to the probe, because the phase lag is given as [4]

$$\tan\phi = -\frac{G_{\mathrm{im}}}{G_{\mathrm{r}}} \tag{13}$$

where G_{r} and G_{im} are the real and imaginary parts of the complex AC thermal conductance, which can be expressed as

$$\begin{aligned} G_{\mathrm{r}} &\approx 2\pi\kappa_s b\left(1 + b\sqrt{\frac{\omega}{2\alpha_s}}\right) + A_c\kappa_c\sqrt{\frac{\omega}{2\alpha_c}} \\ G_{\mathrm{im}} &\approx \omega mC + 2\pi\kappa_s b^2\sqrt{\frac{\omega}{2\alpha_c}} \end{aligned} \tag{14}$$

where b is the contact radius, κs and αs are the thermal conductivity and thermal diffusivity of the sample, κ_c and α_c are the thermal conductivity and thermal diffusivity of the cantilever probe, and the mC product is the thermal mass of the temperature sensor. For simplicity, the thermal interface resistance is ignored in this analysis [4]. For the phase lag to be sensitive to the thermal properties of the sample, the thermal conductance of the cantilever probe cannot be much larger than that of the sample. In addition, it should be noted that both the amplitude and the phase of the signal are influenced by the penetration depth of $\sqrt{2\alpha_s/\omega} = \delta$. Hence, varying the frequency can allow for depth profiling and subsurface imaging.

Pollock, Hammiche, and coworkers [3] have also used their Pt wire resistance heater and thermometer probe to perform localized ac calorimetry [16]. Here they ramped the temperature of both the sample and the probe at about 15 °C/min while adding an ac temperature modulation of 1 °C at 10 kHz by the probe. They found that any phase transition in the sample gave rise to a change in the phase signal in ac mode. This is observed more clearly in the first derivative of the phase as a function of temperature. By scanning the sample under this mode, calorimetric analysis can be performed locally.

Summary and Future Directions

A number of SThM probes have been designed and fabricated. Different operating methods including the dual scan and the zero heat flux techniques have been employed for quantitative mapping of the surface temperature distribution of operating devices. Both dc and ac operations of a resistance heater and thermometer probe have allowed the mapping of local thermal property variation of nanostructured materials. Further enhancement of the spatial resolution and thermal measurement accuracy of the SThM methods can be made by continuous miniaturization of the sensor size at the probe tip as well as improved thermal isolation of the sensor. Detecting the phase lag in the ac SThM signal can potentially allow for locating sub-surface defects that generate localized heating in operating electronic devices, as suggested in [17].

The SThM probes based on a thermal-to-electrical signal transduction method can be complemented with non-contact optical methods for mapping the local temperature distribution of different phonon populations or different thermal properties. Although the spatial resolution of conventional far-field optical imaging methods is limited by diffraction to be on the order of the wavelength, the diffraction barrier has been broken by near-field scanning optical microscopy (NSOM) methods [18] as well as several far-field optical nanoscopy methods [19]. Near-field infrared evanescence wave emitted from a sample surface has been collected using a solid immersion lens [11] and/or scattered into far-field signal by an apertureless metal tip [20]. These techniques have the potential for profiling the surface temperature distribution of devices with a spatial resolution of a fraction of the IR wavelength. Similarly, the spatial resolution of Raman thermograph or thermal reflectance techniques can potentially be improved with either a near-field or far-field nanoscopy technique.

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Carbon-Nanotubes
- ▶ Graphene
- ▶ Scanning Tunneling Microscopy
- ▶ Thermal Conductivity and Phonon Transport

References

1. Williams, C.C., Wickramasinghe, H.K.: Scanning thermal profiler. Appl. Phys. Lett. **49**, 1587–1589 (1986)
2. Majumdar, A., Carrejo, J.P., Lai, J.: Thermal imaging using the atomic force microscope. Appl. Phys. Lett. **62**, 2501–2503 (1993)
3. Pollock, H.M., Hammiche, A.: Micro-thermal analysis: techniques and applications. J. Phys. D Appl. Phys. **34**, R23–R53 (2001)
4. Majumdar, A.: Scanning thermal microscopy. Annu. Rev. Mater. Sci. **29**, 505–585 (1999)
5. Aigouy, L., Tessier, G., Mortier, M., Charlot, B.: Scanning thermal imaging of microelectronic circuits with a fluorescent nanoprobe. Appl. Phys. Lett. **87**, 3 (2005)
6. Shi, L., Majumdar, A.: Thermal transport mechanisms at nanoscale point contacts. J. Heat Trans-T. ASME. **124**, 329–337 (2002)
7. Incropera, F.P., Dewitt, D.P., Bergman, T.L., Lavine, A.S.: Fundamentals of Heat and Mass Transfer. Wiley, Hoboken (2007)
8. Chen, G.: Nanoscale Energy Transport and Conversion: A Parallel Treatment of Electrons, Molecules, Phonons, and Photons. Oxford University Press, New York (2005)
9. Chung, J., Kim, K., Hwang, G., Kwon, O., Jung, S., Lee, J., Lee, J.W., Kim, G.T.: Quantitative temperature measurement of an electrically heated carbon nanotube using the null-point method. Rev. Sci. Instrum. **81**, 5 (2010)
10. Jo, I., Hsu, I.-K., Lee, Y.J., Sadeghi, M.M., Kim, S., Cronin, S., Tutuc, E., Banerjee, S.K., Yao, Z., Shi, L.: Low-frequency acoustic phonon temperature distribution in electrically biased graphene. Nano Lett. (2010). doi:10.1021/nl102858c
11. Cahill, D.G., Goodson, K., Majumdar, A.: Thermometry and thermal transport in micro/nanoscale solid-state devices and structures. J. Heat Trans-T. ASME. **124**, 223–241 (2002)
12. Tien, C.L., Majumdar, A., Gerner, F.M.: Microscale Energy Transport. Taylor & Francis, Washington, DC (1998)
13. Berciaud, S., Han, M.Y., Mak, K.F., Brus, L.E., Kim, P., Heinz, T.F.: Electron and optical phonon temperatures in electrically biased graphene. Phys. Rev. Lett. **104**, 227401 (2010)
14. Chae, D.H., Krauss, B., von Klitzing, K., Smet, J.H.: Hot phonons in an electrically biased graphene constriction. Nano Lett. **10**, 466–471 (2010)
15. Nakabeppu, O., Suzuki, T.: Microscale temperature measurement by scanning thermal microscopy. J. Therm. Anal. Calorim. **69**, 727–737 (2002)
16. Price, D.M., Reading, M., Hammiche, A., Pollock, H. M.: Micro-thermal analysis: scanning thermal microscopy and localised thermal analysis. Int. J. Pharm. **192**, 85–96 (1999)
17. Kwon, O., Shi, L., Majumdar, A.: Scanning thermal wave microscopy (STWM). J. Heat Trans-T. ASME. **125**, 156–163 (2003)
18. Dunn, R.C.: Near field scanning optical microscopy. Chem. Rev. **99**, 2891–2928 (1999)
19. Hell, S.W.: Far-field optical nanoscopy. Science **316**, 1153–1158 (2007)
20. De Wilde, Y., Formanek, F., Carminati, R., Gralak, B., Lemoine, P.A., Joulain, K., Mulet, J.P., Chen, Y., Greffet, J.J.: Thermal radiation scanning tunneling microscopy. Nature **444**, 740–743 (2006)

Scanning Thermal Profiler

- ▶ Scanning Thermal Microscopy

Scanning Tunneling Microscopy

Ada Della Pia and Giovanni Costantini
Department of Chemistry, The University of Warwick, Coventry, UK

Definition

A scanning tunneling microscope (STM) is a device for imaging surfaces with atomic resolution. In STM, a sharp metallic tip is scanned over a conductive sample at distances of a few Å while applying a voltage between them. The resulting tunneling current depends exponentially on the tip-sample separation and can be used for generating two-dimensional maps of the surface topography. The tunneling current also depends on the sample electronic density of states, thereby allowing to analyze the electronic properties of surfaces with sub-nm lateral resolution.

Overview and Definitions

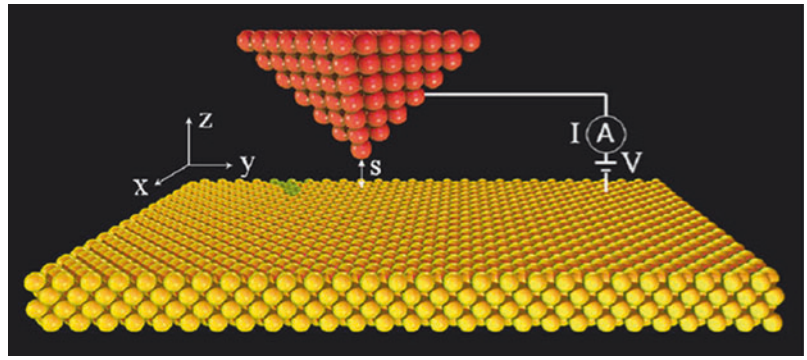
If two electrodes are held a few Å apart and a bias voltage is applied between them, a current flows even though they are not in contact, due to the quantum mechanical process of electron tunneling. This current depends exponentially on the electrode separation, and even minute, subatomic variations produce measurable current changes. In 1981, Gerd Binnig and Heinrich Rohrer at IBM in Zürich realized that this phenomenon can be used to build a microscope with ultrahigh spatial resolution [1], if one of the electrodes is shaped as a sharp tip and is scanned across the surface of the other (Fig. 1). Moreover, since the tunneling current depends also on the electronic properties of the electrodes, this microscope has the ability to probe the electronic density of states of surfaces at the atomic scale. A few years later, Don Eigler at IBM in Almaden, showed that, due to the extremely localized interaction between tip and sample, it is also possible to use this instrument to manipulate individual atoms, to position them at arbitrary locations and therefore to build artificial structures atom-by-atom [2]. This remarkable achievement brought to reality the visionary predictions made by Richard Feynman in his famous 1959 lecture "*There's plenty of room at the bottom*" [3].

The construction of this instrument, dubbed the scanning tunneling microscope (STM), was awarded the 1986 Nobel Prize in Physics and has since then revolutionized contemporary science and technology. The STM has enabled individual atoms and molecules to be imaged, probed, and handled with an unprecedented precision, thereby essentially contributing to our current understanding of the world at the nanoscale. Together with its offspring, the atomic force microscope (AFM) [▶ AFM], the STM is considered as the main innovation behind the birth of nanotechnology.

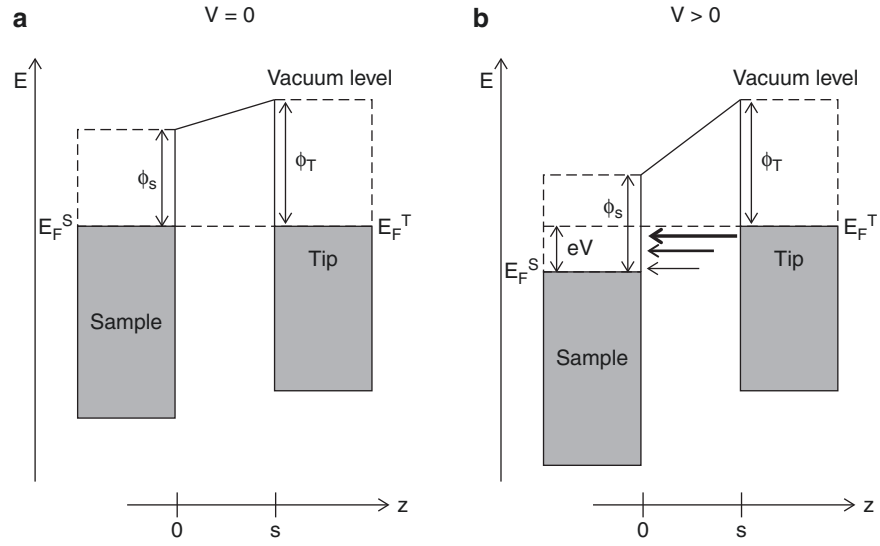
This entry will start with a discussion of the physical principles and processes at the heart of STM in section "Theory of Tunneling." This will be followed by a description of the experimental setup and the technical requirements needed for actually operating such a microscope in section "Experimental Setup." Section "STM Imaging" is dedicated to the most frequent use of STM, namely imaging of surfaces, while section "Scanning Tunneling Spectroscopy" gives a brief account of the spectroscopic capabilities of this instrument. Finally, section "Applications" discusses several applications and possible uses of STM.

Theory of Tunneling

Figure 2 is a schematic representation of the energy landscape experienced by an electron when moving along the z axis of a metallic-substrate/insulator/metallic-tip tunneling junction. The following treatment can easily be extended to include also semiconducting tips or samples. Usually, the tip and the sample are not made of the same material and therefore have different work functions, ϕ_T and ϕ_S, respectively. At equilibrium, the two metals have a common Fermi level, resulting in an electric field being established across the gap region and in different local vacuum levels,



Scanning Tunneling Microscopy, Fig. 1 Schematic representation of a STM. Tip and sample are held at a distance s of a few Å and a bias voltage V is applied between them. The resulting tunneling current I is recorded while the tip is moved across the surface. The coordinate system is also shown



Scanning Tunneling Microscopy, Fig. 2 Energy potential perpendicular to the surface plane for an electron in a tip-vacuum-sample junction. z is the surface normal direction, and s is the tip-sample distance. The *gray boxes* represent the Fermi–Dirac distribution at 0 K. $\phi_{T,S}$ and $E_F^{T,S}$ are the work functions and the Fermi levels of tip and sample, respectively. (**a**) Tip and sample in electrical equilibrium: a trapezoidal potential barrier is created. (**b**) Positive sample voltage V: the electrons tunnel from occupied states of the tip into unoccupied states of the sample. The thickness and the length of the *arrows* indicate the exponentially decreasing probability that an electron with the corresponding energy tunnels through the barrier

depending on the difference $\phi_T - \phi_S$ (Fig. 2a). Since the work functions in metals are of the order of several eV, the potential in the gap region is typically much higher than the thermal energy kT and thus acts as a barrier for sample and tip electrons. A classical particle cannot penetrate into any region where the potential energy is greater than its total energy because this requires a negative kinetic energy. However, this is possible for electrons which, being quantum mechanical objects, are described by delocalized wave functions. This phenomenon goes under the name of *quantum tunneling*. In an unpolarized tip-sample junction the electrons can tunnel from the tip to the sample and vice versa, but there is no net tunneling current. On the contrary, if a voltage V is applied between sample and tip, the Fermi level of the former is shifted by $-eV$ and a net tunneling current occurs, whose direction depends on the sign of V (Fig. 2b). Here the convention is adopted to take the tip as a reference since experimentally the voltage is often applied to the sample while the tip is grounded. If

V is the bias voltage, the energy for an electron in the sample will change by $-eV$, that is, it will decrease for positive values of V.

The tunneling current can be evaluated by following the time-dependent perturbation approach developed by Bardeen [4, 5]. The basic idea is to consider the isolated sample and tip as the unperturbed system described by the stationary Schrödinger equations:

$$(\mathcal{T} + \mathcal{U}_s)\psi_\mu = E_\mu \psi_\mu \tag{1}$$

and

$$(\mathcal{T} + \mathcal{U}_\mathcal{T})\chi_\nu = E_\nu \chi_\nu \tag{2}$$

where $\mathcal{T}$ is the electron kinetic energy. The electron potentials $\mathcal{U}_s$ and $\mathcal{U}_\mathcal{T}$ and the unperturbed wavefunctions ψ_μ and χ_ν are nonzero only in the sample and in the tip, respectively. Based on this, it can be shown [5] that the transition probability per unit time $w_{\mu\nu}$ of an electron from the sample state ψ_μ to the tip state χ_ν is given by Fermi's golden rule:

$$w_{\mu\nu} = \frac{2\pi}{\hbar} |M_{\mu\nu}|^2 \delta(E_\nu - E_\mu) \tag{3}$$

where the matrix element is:

$$M_{\mu\nu} = \int \chi_\nu^* \left(\vec{x}\right) \mathcal{U}_\mathcal{T} \left(\vec{x}\right) \psi_\mu \left(\vec{x}\right) d^3 \vec{x} \,. \tag{4}$$

The δ function in Eq. 3 implies that the electrons can tunnel only between levels with equal energy, that is, (Eq. 3) accounts only for an *elastic tunneling* process. The case of an inelastic tunneling process will be considered in section "Scanning Tunneling Spectroscopy." The total current is obtained by summing $w_{\mu\nu}$ over all the possible tip and sample states and by multiplying this by the electron charge e. The sum over the states can be changed into an energy integral by considering the density of states (DOS) $\rho(E) : \sum \to 2\int f(\varepsilon)\rho(\varepsilon)d\varepsilon$, where the factor 2 accounts for the spin degeneracy while f, the Fermi–Dirac distribution function, takes into consideration Pauli's exclusion principle and the electronic state population at finite temperatures.

As a consequence, the total current can be written as:

$$I = \frac{4\pi e}{\hbar} \int_{-\infty}^{\infty} \left[f_T\left(E_F^T - eV + \varepsilon\right) - f_S\left(E_F^S + \varepsilon\right)\right] \times \rho_T\left(E_F^T - eV + \varepsilon\right)\rho_S\left(E_F^S + \varepsilon\right)|M|^2 d\varepsilon \tag{5}$$

where EF is the Fermi energy and the indexes T and S refer to the tip and the sample, respectively. Equation 5 already accounts for the movement of electrons from the sample to the tip and vice versa.

Several approximations can be made to simplify Eq. 5 and to obtain a manageable analytical expression for I. If the thermal energy $k_B\mathrm{T} \ll eV$, the Fermi–Dirac distributions can be approximated by step functions and the total current reduces to:

$$I = \frac{4\pi e}{\hbar} \int_0^{eV} \rho_T\left(E_F^T - eV + \varepsilon\right)\rho_S(E_F^S + \varepsilon)|M|^2 d\varepsilon \tag{6}$$

(Note that Eq. 6 is valid only for $V > 0$. For $V < 0$ the integrand remains identical but the integration limits become $-e|V|$ and 0). In this case, only electrons with an energy differing from E_F by less than eV can participate to the tunneling current. This can be directly seen in Fig. 2b for the case of positive sample bias: tip electrons whose energy is lower than $E_F^T - eV$ cannot move because of Pauli's exclusion principle, while there are no electrons at energies higher than E_F^T. The main problem in determining expression (Eq. 5) is, however, the calculation of the tunneling matrix elements M since this requires a knowledge of the sample and the tip wave functions, which can be very complicated. On the other hand, for relatively small bias voltages (in the ± 2 V range), Lang [6] showed that a satisfactory approximation of $|M|^2$ is given by a simple one-dimensional WKB tunneling probability. In the WKB approximation [7], the probability $D(\varepsilon)$

that an electron with energy ε tunnels through a potential barrier $U(z)$ of arbitrary shape is expressed as:

$$D(\varepsilon) = \exp\left\{-\frac{2}{\hbar}\int_0^s [2m(U(z)-\varepsilon)]^{\frac{1}{2}}dz\right\}. \quad (7)$$

This semiclassical approximation is applicable if ($\varepsilon \ll U$) which is generally satisfied in the case of metal samples where the work function is of the order of several eV. In order to obtain a simple analytical expression for D, the trapezoidal potential barrier of a biased tip-sample junction (see Fig. 2b) is further approximated with a square barrier of average height $\phi_{\text{eff}}(V) = (\phi_T + \phi_S + eV)/2$. By using this, the integral in Eq. 7 becomes:

$$D(\varepsilon, V, s) = \exp(-2ks) \quad (8)$$

where

$$k = \sqrt{\frac{2m}{\hbar^2}(\phi_{\text{eff}} - \varepsilon)}. \quad (9)$$

In order to evaluate k, it must be noted that electrons closest to the Fermi level experience the lowest potential barrier and are therefore characterized by an exponentially larger tunneling probability (see Fig. 2b). Thus, in a first approximation, it can be assumed that only these electrons contribute to the tunneling current which, for positive bias, is equivalent to set $\varepsilon \approx eV$ in Eq. 9. Moreover, if the bias is much smaller than the work functions, eV can be neglected, resulting in

$$k \cong \frac{\sqrt{m(\phi_T + \phi_S)}}{\hbar} = 5.1\sqrt{\frac{\phi_T + \phi_S}{2}}\,\text{nm}^{-1} \quad (10)$$

where the work functions are expressed in eV. Using typical numbers for metallic work functions, the numerical value of the inverse decay length 2 k in Eq. 8 becomes of the order of 20 nm^{-1}. Therefore, variations in s of 1 Å correspond to one order of magnitude changes in the tunneling probability and, as a consequence, in the measured current. This very high sensitivity provides the STM with a vertical resolution in the picometer regime. The lateral resolution of STM depends on how different points of the tip contribute to the total tunneling current. By considering a spherical tip shape with radius R, most of the current originates from the central position since this is closest to the surface. A point laterally displaced by Δx from the tip center is $\Delta z \approx \frac{\Delta x^2}{2R}$ further away from the substrate (higher order Δx terms are neglected in this evaluation). As a consequence, with respect to the tip center, the corresponding tunneling probability is reduced by a factor:

$$\exp\left(-2k\frac{\Delta x^2}{2R}\right). \quad (11)$$

By considering a tip radius $R \approx 1$ nm, the current changes by one order of magnitude for variations $\Delta x = 3$ Å. The actual lateral resolution is typically smaller than this upper limit and can reach down to fractions of an Å. Its specific value however depends on the precise shape of the tip which is unknown a priori. These values, together with the vertical resolution discussed above, lie at the basis of the STM atomic imaging capabilities.

Finally, if the tunneling probability (Eq. 8) is substituted for the tunneling matrix $|M|^2$ in Eq. 6, the total tunneling current can be expressed as:

$$I = \frac{4\pi e}{\hbar}\int_0^{eV} \rho_T\left(E_F^T - eV + \varepsilon\right)\rho_S\left(E_F^S + \varepsilon\right)e^{-2ks}d\varepsilon. \quad (12)$$

Therefore, for a fixed lateral position of the tip above the sample, the tunneling current I depends on the tip-sample distance s, the applied voltage V and the tip and sample density of states ρ_T and ρ_S, respectively.

Experimental Setup

As seen in the previous section, variations of 1 Å in s induce changes in the tunneling probability of one order of magnitude. The exponential dependence in Eq. 8 is responsible for the ultimate spatial resolution of STM but places stringent

constraints on the precision by which s must be controlled, as well as on the suppression of vibrational noise and thermal drift. Moreover, typical tunneling currents are in the 0.01–10 nA range, requiring high gain and low noise electronic components. The following subsections are dedicated to a general overview of technologies and methods used to meet these specifications.

Scanner and Coarse Positioner

The extremely fine movements of the tip relative to the sample required for operating an STM are realized by using piezoelectric (▶ Piezoresistivity) ceramic actuators (*scanners*) which expand or retract depending on the voltage difference applied to their ends. In a first approximation, the voltage-expansion relation can be considered as linear with a proportionality factor (piezo constant) usually of few nanometer/Volt. The main requirements for a good scanner are: high mechanical resonance frequencies, so as to minimize noise vibrations in the frequency region where the feedback electronics operates (see section "Electronics and Control System"); high scan speeds; high spatial resolution; decoupling between x, y, and z motions; minimal hysteresis and creep; and low thermal drift. Although several types of STM scanner have been developed, including the bar or tube tripod, the unimorph disk and the bimorph [8], the most frequently used is a single piezoelectric tube whose outer surface is divided into four electrode sections of equal area. By applying opposite voltages between the inner electrode and opposite sections of the outer electrode, the tube bends and a lateral displacement is obtained. The z motion is realized by polarizing with the same voltage the inner electrode in respect to all four outer electrodes. By applying several hundred Volts to the scanner, lateral scan widths up to 10 μm and vertical ones up to 1 μm can be obtained, while retaining typical lateral and vertical resolutions of 0.1 and 0.01 nm, respectively.

While scanning is typically done by one individual piezoelectric element, larger displacements up to several millimeters are needed to bring the tip in close proximity to the sample, to move it to different regions of the surface or to exchange samples or tips. These are achieved by mounting the scanner onto a coarse position device. Several designs have been developed to this aim including micrometric screws driven either manually or by a stepper motor, piezoelectric walkers like the louse used in the first STM [9] or the inch-worm [10], magnetic walkers where the movement is obtained by applying voltage pulses to a coil with a permanent magnet inside and piezoelectric driven stick–slip motors, as the Besocke-beetle [11] or the Pan motor [12].

Electronics and Control System

The voltages driving the piezoelectric actuators and their temporal succession and duration are generated by an electronic control system. The electronics are also used to bias the tunneling junction, to record the tunneling current and to generate the STM images. In most of the modern instruments, these tasks are digitally implemented by a computer interfaced with digital to analog (DAC) and analog to digital (ADC) converters. The tunneling current is amplified by a high gain I–V converter (10^8–10^{10} V/A) usually positioned in close proximity of the tip, so as to reduce possible sources of electronic interference. This signal is then acquired by an ADC and processed by the control system. DACs are used to apply the bias voltage (from a few mV to a few V) between tip and sample and, in conjunction with high voltage amplifiers, to polarize the piezo elements. A feedback loop is integrated into the control system and is activated during the frequently used *constant current* imaging mode (see section "STM Imaging"). By acting on the z motion of the scanner, the feedback varies s to keep the tunneling current constant. This is controlled by a proportional-integral and derivative (PID) filter whose parameters can be set by the operator. Finally, a lock-in amplifier is often used to improve the signal-to-noise ratio in scanning tunneling spectroscopy (STS) measurements (see section "Scanning Tunneling Spectroscopy").

Tip

Sharp metal tips with a low aspect ratio are essential to optimize the resolution of the STM images and to minimize flexural vibrations of the tip,

respectively. Ideally, in order to obtain atomically resolved topographies and accurate spectroscopic measurements, the tip should be terminated by a single atom. In this case, because of the strong dependence on the tip-sample separation (see section "Theory of Tunneling"), most of the tunneling current would originate from this last atom, whose position and local DOS would precisely determine the tunneling conditions. In practice, however, it is almost impossible to determine the exact atomic configuration of the tip and the actual current is often due to a number of different atoms. This is still compatible with good tunneling conditions as long as these atoms are sufficiently localized (in order to avoid "multiple tip effects") and their structural and chemical state remains constant during scanning.

The most commonly used methods to produce STM tips are to manually cut or to electrochemically etch thin wires of platinum-iridium and tungsten, respectively. These materials are chosen because of their hardness, in order to prevent tips becoming irreversible damaged after an accidental crash. Other metallic elements and even semiconductor materials have been used as tips for specific STM applications. Due to their chemical inertness, Pt-Ir tips are often used to scan in air on atomically flat surfaces without the need of any further processing. However, they typically have inconsistent radii, while etched W tips are characterized by a more reproducible shape. These latter have the drawback that a surface oxide up to 20 nm thick is formed during etching or exposure to air. For this reason, W tips are mostly used in ultrahigh vacuum (UHV) where the oxide layer can be removed through ion sputtering and annealing cycles. Prior to use, tips are often checked by optical microscopy, scanning electron microscopy (▶ SEM), and field ion microscopy or transmission electron microscopy (▶ TEM). The quality of a tip can be further improved during scanning by using "*tip forming*" procedures, including pulsing and controlled crashing into metal surfaces. These processes work because the desorption of adsorbed molecules or the coating with atoms of the metallic substrate can produce a more stable tip apex. If STM is performed in polar liquids (▶ EC-STM), electrochemical processes might generate Faradaic or non-Faradaic currents which can be of the same order of magnitude or even larger than the tunneling current. In order to minimize these effects, the tip, except for its very apex, must be coated with an insulating material.

Vibration Isolation

A low level of mechanical noise is an essential requirement for any type of scanning probe microscopy. For this reason, the core of a STM, where the tip-sample junction is located, is always equipped with one or several types of vibration damping systems. These can be stacks of metal plates separated by elastic spacers, suspension springs, or eddy current dampers composed of copper elements and permanent magnets. The low-frequency components of mechanical noise (<10 Hz), which are the most difficult to eliminate, are minimized by building a small and rigid STM with a high resonance frequency. Depending on the overall size and weight of the microscope, further noise damping strategies can be adopted. Smaller, typically ambient conditions STMs can be placed on metal or granite slabs suspended by springs or bungee cords or floating on pneumatic isolators. Sometimes, piezo-driven, feedback-controlled active vibration suppressors are also combined with passive systems. Larger versions of pneumatic isolators and active damping are often used to float the frames and the chambers of big UHV STMs. The laboratory where a STM instrument is located also plays an essential role for its performance. Ground floor rooms are always preferred since they minimize low-frequency natural building oscillations, which can be very difficult to counteract. High-resolution instruments are sometimes placed on large concrete blocks which are separated from the rest of the laboratory floor and rest either on a sand bed, an elastomer barrier or on second-stage pneumatic isolators. Moreover, they are also often surrounded by an acoustically insulating box. All these systems essentially act as low-pass mechanical filters whose effectiveness improves with decreasing cutoff frequencies, that is, with increasing mass and decreasing rigidity. For this reason, the body of a STM is typically a relatively heavy block of metal and the frames, slabs, and

vacuum chambers supporting or containing the microscope often have a considerable weight.

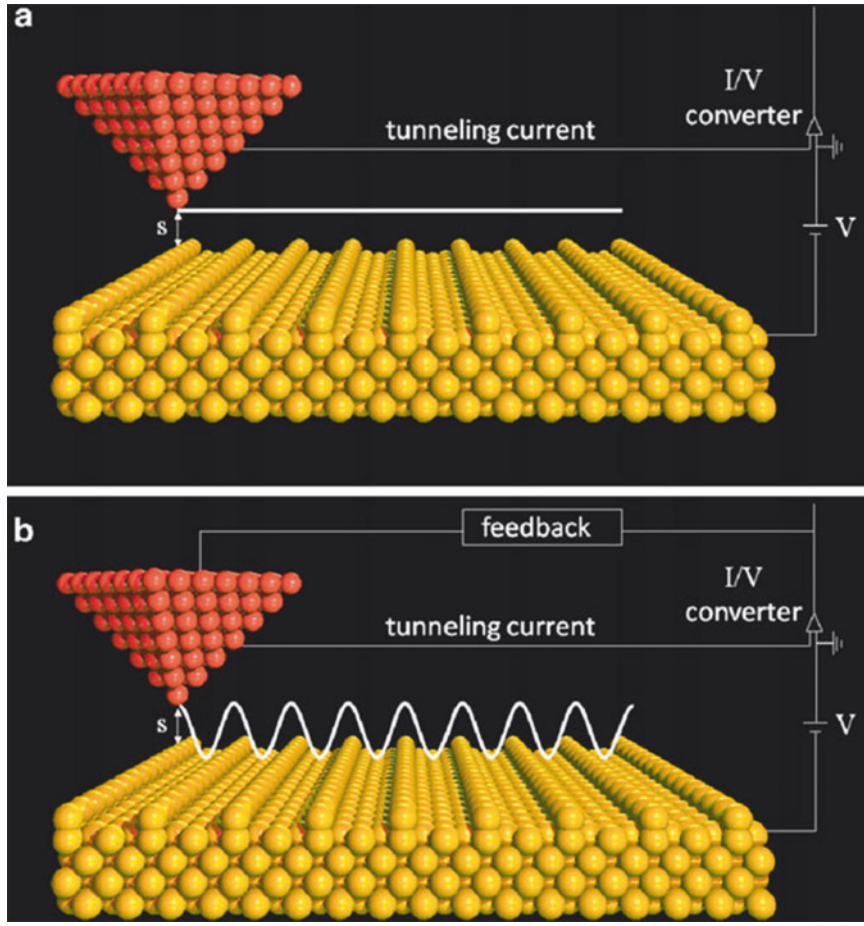
Setups for Different Environments and Temperatures

Different types of STMs have been developed that can operate in various environments such as air, inert atmosphere (N_2, Ar), vacuum, high pressure, liquid, or in an electrochemical cell. The core of the different instruments is essentially the same, although the experimental chambers and setups in which they are located can vary substantially. Ambient condition STMs are typically quite compact and rigid and do not need elaborated anti-vibrational mechanisms. On the other hand, since sound waves represent a major problem, atmospheric pressure STMs are usually contained in an acoustic enclosure. A STM operating in vacuum must be hosted in a chamber with vibration-free pumps (typically ionic pumps for UHV) and must be equipped with sophisticated sample and tip manipulation mechanisms. Such systems often also have an in situ surface preparation stage allowing the handling of samples without air exposure.

STM can be performed at high pressures (1–30 bar) by installing the microscope head into gas manifolds under conditions similar to those used in industrial catalytic processes. Also in this case, sample and tip manipulation and preparation stages are mandatory parts of the system. Since these types of studies are typically performed at elevated temperatures (up to 600 K) and in the presence of highly reactive gases, the metallic parts of the STM scanner and of the chamber are often gold plated, the volume of the STM chamber is kept as small as possible and the tip material is chosen to be inert toward the gases [13]. Moreover, low voltages are used for polarizing the piezos in order to avoid gas discharges at intermediate pressures (10^{-3}–10 mbar) and shields are added to protect the STM from the deposition of conductive materials which could create electrical shorts.

STM at the liquid/solid interface and electrochemical STM (EC-STM) (▶ EC-STM) need the tip and sample to be inside a liquid cell which, in turn, may be placed in a humidity-controlled atmosphere. In the case of low vapor pressure liquids, the STM can be simply operated under ambient conditions by dipping the tip into a liquid droplet deposited on the sample. A special coating must be applied to the tip when working with polar liquids (see section "Tip").

STM can also be performed at different temperatures (in vacuum or controlled atmosphere chambers): variable temperature STM (VT-STM) able to cover the 5–700 K range, low temperature STM (LT-STM) operating at 77 K or 5 K and even milli-Kelvin STM instruments are currently available. A VT-STM is typically used to study thermally activated processes such as diffusion and growth, phase transitions, etc. These systems have sample heating and cooling stages which can be operated in a combined way so as to achieve a very precise temperature stabilization. Resistive heating is normally employed to increase the temperature, while both flow and bath cryostats with liquid nitrogen or helium as cryogenic fluids are used to reduce it. Continuous flow cryostats offer a high flexibility in temperature but are characterized by lower thermal stability, by inherent mechanical vibrations and do not easily attain temperatures below 20 K. Bath cryostats are more stable, are able to reach lower temperatures but are often also much bulkier (e.g., in order to limit the He consumption rate, a liquid He cryostat is actually a double-stage cryostat with an outer liquid nitrogen mantle). For most of these instruments the variable temperature capabilities refer to the possibility of choosing different (fixed) temperatures at which the microscope is run. However, few systems endowed with specific position tracking and drift compensating capabilities allow a "true" variable temperature operation where the same surface area can be imaged with atomic resolution while its temperature is changed. LT-STMs are operated at a fixed temperature and are typically inserted inside double stage cryostats which significantly complicates the tip and sample access. However, these instruments are extremely stable with a very low thermal drift and are therefore the best choice for STS and manipulation experiments (see sections "Scanning Tunneling Spectroscopy" and "Applications"). Milli-Kelvin STMs enable temperatures to be reached where extremely interesting magnetic, quantum Hall physics and superconductivity phenomena occur. Moreover, the thermal broadening of electronic



Scanning Tunneling Microscopy, Fig. 3 Schematic representation of (**a**) the constant height and (**b**) the constant current imaging modes, respectively. The *thick lines* represent the trajectory followed by the tip

features is strongly reduced, which is required for high-resolution measurements. These systems operate based on the evaporative cooling of liquid ^{3}He to temperatures of about 300 mK or liquid ^{3}He and ^{4}He mixtures below 10 mK. The STM heads can be further placed inside large-bore superconducting magnets (up to 15 T), enabling the low temperature and high magnetic field conditions necessary to access superconductive phase transitions or to detect single spin flip processes.

STM Imaging

STM images are generated by recording the tunneling current as a function of the tip position while the tip is scanned across the sample surface. This can be done in two different ways which define the two main STM imaging modes:

- *Constant height mode*. The z section of the piezo scanner is kept fixed while the tip is moved over the substrate at a constant bias voltage (Fig. 3a). Variations of the tip-sample distance due to the surface topography produce a corresponding variation of the tunneling current which is recorded point-by-point and used to build the STM gray-level image. This mode is employed only in small areas of extremely flat surfaces, where the probability of crashing into protrusions such as steps or defects is relatively

small. Very high scanning speeds can be used because of the absence of a feedback control.
- *Constant current mode*. While the x and y sections of the piezo scanner are used to laterally move the tip across the surface, the z section is driven by the electronic feedback so as to maintain a constant tunneling current (Fig. 3b). The corresponding z-voltage applied to the scanner (feedback signal) is recorded point-by-point and used to build the STM gray-level image. This mode can be employed for any type of surface topography and is therefore the most frequently used.

Since the constant height mode is applied to atomically flat surfaces with sub-Å height variations, the exponential $I-s$ relation derived from Eq. 12 can be approximated by a linear dependence. As a consequence, constant height STM images are a good representation of flat surfaces. On the other hand, for less planar substrates, one must use the constant current mode which directly reproduces the surface height due to the linear voltage-extension relation of piezoelectric materials. However, even constant current STM images are a reliable representation of the "true" surface topography only if the sample local DOS does not vary across the scanned area. If this is not the case, a constant current profile corresponds to a complex convolution of topographical and electronic features (see Eq. 12) which can be particularly relevant for surfaces covered with molecular adsorbates.

Scanning Tunneling Spectroscopy

Besides complicating the interpretation of STM images, the dependence of the tunneling current on the sample DOS also offers the unique opportunity of probing the electronic characteristics of surfaces with sub-nm spacial resolution. Having fixed the tip lateral position, the tunneling current I is a function of the applied bias voltage V and the tip-sample separation s only, the precise relation being established by Eq. 12. In a STS experiment, the relation between two of these three parameters is measured while the remaining one is kept constant (STS). $I(V)$ spectroscopy, where the tunneling current is measured as a function of the bias voltage for a constant tip-sample separation, is the most widely used technique because it provides indications about the DOS of the sample.

Due to the spatial localization of the tunneling current (see section "Theory of Tunneling"), STS enables the characterization of the electronic properties of individual atoms and molecules in relation to their structure, bonding and local environment. Moreover, STS can also be used to create 2D maps of the sample DOS with sub-nm resolution. Such measurements are particularly interesting for quantum confined electronic systems (e.g., quantum dots or quantum corrals) or for determining the shape of molecular orbitals [14] (*wavefunction mapping*). By changing the polarity of the bias voltage, STS gives access to both the occupied and the unoccupied states of the sample. In this sense, it is often considered as complementary to ultraviolet photoemission spectroscopy (UPS), inverse photoemission spectroscopy (IPS) and electron energy loss spectroscopy (EELS), where the signal is averaged over a large area of the surface (between 0.1 and 2 mm in diameter). On the other hand, STS does not provide direct chemical information and tip artifacts can strongly influence the spectroscopic data.

So far we have assumed that electrons conserve their energy during the tunneling process (see Eq. 3). However, electrons can also tunnel inelastically between the tip and the sample by exchanging energy and inducing the excitation of vibrational modes, spin-flips, magnons, plasmons, excitons, etc. These extra tunneling channels become available only above specific voltage thresholds since only beyond these values a part of the electron energy can be converted into the excitation. The additional inelastic pathways increase the overall tunneling probability and therefore show up as discrete step-like features in the tunneling conductivity or as slope changes in $I(V)$ curves. This technique is called inelastic electron tunneling spectroscopy (IETS) and benefits from the same spatial resolution as STM and STS. IETS has been used to measure vibrational modes of individual molecules, spin excitations of single magnetic atoms, collective

plasmon excitations in 2D materials and magnons in ferromagnets.

A different way of detecting tunneling-induced molecular vibrations by means of a STM is to rely on their coupling with dynamical processes such as molecular motions. In particular, by measuring the frequency of molecular hopping events as a function of the applied bias voltage, it is possible to create so-called *action spectra* which reflect the vibrational spectrum of an individual molecule in a quantitative manner [15]. Optical excitations can also be revealed in an alternative way by coupling the STM with a photon detection system able to collect and analyze the luminescence stimulated by inelastically tunneling electrons [16]. Such a setup has been used to characterize plasmon emission from metallic surfaces and luminescence from semiconductor quantum structures and adsorbed molecules.

Applications

Since the first STM images of the surfaces of $CaIrSn_4$ and Au [1] were published back in 1982, STM has been used to analyze a wide range of materials: clean and adsorbate covered metal surfaces, semiconductors, superconductors, thin insulating layers, small and large organic molecules, individual atoms, liquid–solid interfaces, magnetic layers and surfaces, quasicrystals, polymers, biomolecules, nanoclusters, and carbon nanotubes. Imaging is the most frequent application of STM used to determine the structural properties of substrates and their reconstructions, the presence of defects, sites of adsorption for adatoms and molecules and the symmetry and periodicity of adsorbate superstructures. Nevertheless, right from the beginning, it became clear that the ultimate spatial resolution of STM, in combination with its dependence on the electronic properties of tip and sample, could allow a much wider range of applications of this instrument. These include the characterization of surface electronic, vibrational, optical, and magnetic properties, the measurement of single molecule conductivities, and the study of dynamic processes. In the following, we will only touch upon some of the most frequent applications, without any presumption of being exhaustive.

Equation 12 shows that the tunneling current depends on the sample DOS close to the Fermi energy E_F. As a consequence, at a typical bias of a few Volts, it should be possible to image conductors, superconductors and small-gap or doped semiconductors but not molecules and insulating materials due to the vanishing DOS in the probed energy range (for most molecules the highest occupied and the lowest unoccupied orbitals are separated by an energy gap of several eV). However, the great majority of molecules adsorbed on metallic substrates can be easily imaged at moderate bias voltages. This is due to the formation of a metal-organic interface which can modify the molecular electronic properties leading to a broadening of the initial discrete energy levels, to a reduction of the gas-phase energy gap and even to the development of new states if covalent molecule-substrate bonds are established. All these effects contribute to the DOS at E_F and allow the imaging process. Regarding insulating materials, STM can only be done on films deposited onto conductive substrates if they are thin enough to allow the tunneling of electrons. These films are often used to electronically decouple organic adsorbates from metallic substrates.

The mechanism which allows the imaging of biomolecules such as DNA and proteins is currently still under debate [17]. As these molecules have a very large energy gap (5–7 eV) they can be considered as insulating materials and the current measured in STM experiments might be mediated by the thin water layer surrounding the molecules in air. Metalloproteins have also been imaged in their "natural" environment by using EC-STM (► EC-STM). Several reports have shown that when these redox active molecules are imaged under potentiostatic control, the tunneling current can be mediated by their metal redox-center, with enhanced conductivities measured for bias voltages close to the redox potential.

Although STM is a surface sensitive method, it can be also used to analyze buried interfaces and structures in cross-sectional STM (XSTM) [18]. The specific sample preparation in this technique requires brittle materials such as oxide samples or

semiconductor wafers. A cross section of the structure to be analyzed is prepared by cleaving the sample and positioning the STM tip onto the exposed edge. In this way, various physical properties can be probed, including the morphology and abruptness of buried nanostructures and interfaces, the alloying in epitaxial layers, the spatial distribution of dopants and their electronic configuration and the band offsets in semiconductor heterojunctions. XSTM has also been used to study, in real time, the changes occurring in semiconductor quantum well laser devices under operating conditions.

STM can further be employed for tracking dynamic surface processes, provided that the corresponding characteristic times are longer than the acquisition time. By choosing optimized designs for the piezo scanners and the electronic feedback, video-rate instruments have been developed able to record several tens of images per second and thereby to follow mobility and assembly processes in real time [19, 20].

When a tip with spin polarized electrons is used in STM, besides the parameters already indicated in Eq. 12, the local sample magnetization also influences the tunneling current. In fact, due to the different density of states at E_F of "spin-up" and "spin-down" electrons in magnetic materials, a spin polarized tip causes a tunnel magnetoresistance effect which results in a further contrast mechanism. This technique, called spin-polarized STM (SP-STM), has been used both in the presence and absence of an external magnetic field for detecting magnetic domain structures and boundaries in ferro- and antiferromagnetic materials, visualizing atomic-scale spin structures and determining spatially resolved spin-dependent DOS. An essential aspect of SP-STM is the ability to control the magnetization direction of the tip which can be achieved by evaporating different types of ferromagnetic or antiferromagnetic thin films on nonmagnetic tips. This technique is preferred to the use of bulk magnetic tips since it reduces the magnetic stray fields which can significantly modify the sample magnetization [21].

The ability of STM to identify and address individual nano-objects has been used to measure the conductivity of single molecules absorbed on metal surfaces. While the tip is approached to the molecule of interest at constant bias voltage, the current flowing in the junction can be measured, thereby generating an $I(s)$ curve. Alternatively, $I(V)$ curves can be recorded at different s values. Since tip and substrate act as electrodes, both methods enable information to be obtained about the conductance of the individual molecule embedded in the junction. These measurements are often complemented by IETS experiments in the same configuration. IETS might in fact help to determine the arrangement and the coupling of the junction, which has a significant influence on the electronic and structural properties of the molecule. Single molecule STM conductance experiments represent an important source of information for understanding mechanisms of electron transport in organic molecules with applications in organic electronics and photovoltaics. They complement narrow gap electrode and break junction techniques, having the significant advantage of a highly localized electrode which allows to address and characterize individual molecules.

A similar type of application, although typically not aimed at individual molecules, is at the basis of the four point probe STM, where four STM tips, in addition to imaging, are used for local four point electric conduction measurements. A scanning electron microscope is installed above the STM enabling the positioning of the tips on the contact. The purpose of such very complex instruments is to measure the charge transport through individual nanoelectronic components (in particular self-assembled ones) and to correlate this information with a local high-resolution structural characterization.

S

Tip-Induced Modification

Besides being an extraordinary instrument for the characterization of structural, electronic, vibrational, optical, and magnetic properties of surfaces with subnanometer resolution, STM has also developed as a tool to modify and nanoengineer matter at the single molecule and atom scale.

By decreasing the distance between the tip and the sample in a controlled way, indentations can

be produced in the substrate with lateral sizes down to a few nm. Nanolithography can also be performed by tunneling electrons into a layer of e-beam photoresist (▶ SU-8 Photoresist), thereby reaching a better resolution compared to standard electron beam lithography (EBL). Many other STM-based nanopatterning and nanofabrication techniques have been developed based on a number of physical and chemical principles including anodic oxidation, field evaporation, selective chemical vapor deposition, selective molecular desorption, electron-beam induced effects, and mechanical contact. All these methods exploit the extreme lateral localization of the tunneling current and can be applied in air, liquids and vacuum.

However, the nanotechnological application that gained most attention is the ability to manipulate individual atoms and molecules on a substrate. This is possible due to a controlled use of tip-particle forces and is typically done in UHV and at low temperatures. The first atomic manipulation experiment was performed by Eigler and Schweizer in 1989 [2]. This phenomenal result fulfilled Richard Feynman's prophecy that "ultimately-in the great future-we can arrange the atoms the way we want; the very atoms, all the way down!" [3].

During a lateral manipulation experiment, the tip is first placed above the particle to be moved (for example an atom) and the tunneling current is increased while keeping a constant voltage. This results in a movement of the tip toward the atom, see Eq. 12. If their separation is reduced below 0.5 nm, Van der Waals forces start to come into play together with attractive and repulsive chemical interactions. When these forces equal the diffusion energy barrier, a lateral displacement of the tip can induce a movement of the atom parallel to the surface. After the desired final position is reached, the tip is retracted by reducing the tunneling current to the initial value, leaving the atom in the selected place. Depending on the tip-particle distance and therefore on the strength and nature of the interaction, different manipulation modes including pulling, pushing, and sliding [22] were identified and used to move different types of atoms and molecules.

Thanks to this technique, it was possible to fabricate artificial nanostructures such as the *quantum corral* [23] and to probe quantum mechanical effects like the quantum confinement of surface state electrons or the *quantum mirage*. Lateral STM manipulation has also been used to switch between different adsorption configurations and conformations of molecules on surfaces and to modify their electronic properties in a controlled way [24].

A further application of STM manipulation is the synthesis of new molecular species based on the ability of STM to form and break chemical bonds with atomic precision. Reactants are brought close together on the surface and the actual reaction is realized by applying a voltage pulse or by exciting vibrational modes through inelastically tunneling electrons. Examples of this technique include the dissociation of diatomic molecules, the Ullmann reaction, the isomerization of dichlorobenzene and the creation of metal-ligand complexes.

The STM tip has also been used to perform vertical manipulations of nanoparticles where an atom (or molecule) is deliberately transfered from the surface to the tip and vice versa by using the electric field generated by the bias voltage. In contrast to the lateral manipulation, here the bonds between the surface and the atom are broken and re-created [25]. By approaching the tip at distances of a few Å from the chosen particle chemical interactions are established that reduce the atom-surface binding energy. If a voltage pulse is applied under these conditions, the resulting electric field (of the order of 10^8 V/cm) can be enough to induce the particle desorption. The vertical manipulation technique has also been used as a means to increase the lateral resolution of STM. In fact, the controlled adsorption of a specific molecule onto the tip often makes it "sharper" and can add a chemical resolution capability if the DOS of the extra molecule acts as an "energy filter."

A related effect is exploited in the recently proposed scanning tunneling hydrogen microscopy (STHM) technique. In STHM, the experimental chamber is flooded with molecular hydrogen while the tip is scanned in constant

height mode at very close distances over the surface. H_2 can get trapped in the tip-sample junction and its rearrangement during scanning of the surface generates a new contrast mechanism based on the short-range Pauli repulsion. This is extremely sensitive to the total electron density, thereby endowing the STM with similar imaging capabilities to non-contact AFM (▶ AFM, Noncontact Mode) and making it able to resolve the inner structure of complex organic molecules [26].

Acknowledgments This work was supported by EPSRC (EP/D000165/1); A. Della Pia was funded through a WPRS scholarship of the University of Warwick. J. V. Macpherson, T. White, and B. Moreton are gratefully thanked for their critical reading of the manuscript.

Cross-References

- ▶ AFM, Noncontact Mode
- ▶ Atomic Force Microscopy
- ▶ Electrochemical Scanning Tunneling Microscopy
- ▶ Electron Beam Lithography (EBL)
- ▶ Piezoresistivity
- ▶ Scanning Electron Microscopy
- ▶ SU-8 Photoresist
- ▶ Transmission Electron Microscopy

References

1. Binnig, G., Rohrer, H., Gerber, C., Weibel, E.: Surface studies by scanning tunneling microscopy. Phys. Rev. Lett. **49**, 57 (1982)
2. Eigler, D.M., Schweizer, E.K.: Positioning single atoms with a scanning tunneling microscope. Nature **344**, 524 (1990)
3. Feynman, R.P.: There's plenty of room at the bottom: an invitation to enter a new field of physics. Eng. Sci. **23**, 22 (1960)
4. Bardeen, J.: Tunneling from a many-particle point of view. Phys. Rev. Lett. **6**, 57 (1961)
5. Gottlieb, A.D., Wesoloski, L.: Bardeen's tunnelling theory as applied to scanning tunnelling microscopy: a technical guide to the traditional interpretation. Nanotechnology **17**, R57 (2006)
6. Lang, N.D.: Spectroscopy of single atoms in the scanning tunneling microscope. Phys. Rev. B **34**, 5947 (1986)
7. Landau, L.D., Lifshitz, E.M.: Quantum Mechanics: Non-relativistic Theory. Pergamon Press, Oxford (1977)
8. Chen, C.J.: Introduction to Scanning Tunneling Microscopy. Oxford University Press, Oxford (2008)
9. Binnig, G., Rohrer, H.: Scanning tunneling microscope. Helv. Phys. Acta **55**, 726 (1982)
10. Okumura, A., Miyamura, K., Gohshi, Y.: The STM system constructed for analytical application. J. Microsc. **152**, 631 (1988)
11. Besocke, K.: An easily operable scanning tunneling microscope. Surf. Sci. **181**, 145 (1987)
12. Pan, S.H., Hudson, E.W., Davis, J.C.: ^{3}He refrigerator based very low temperature scanning tunneling microscope. Rev. Sci. Instrum. **70**, 1459 (1999)
13. Laegsgaard, E., et al.: A high-pressure scanning tunneling microscope. Rev. Sci. Instrum. **72**, 3537 (2001)
14. Repp, J., Meyer, G., Stojkovic, S.M., Gourdon, A., Joachim, C.: Molecules on insulating films: scanning-tunneling microscopy imaging of individual molecular orbitals. Phys. Rev. Lett. **94**, 026803 (2005)
15. Sainoo, Y., et al.: Excitation of molecular vibrational modes with inelastic scanning tunneling microscopy processes: examination through action spectra of cis-2-butene on Pd(110). Phys. Rev. Lett. **95**, 246102 (2005)
16. Gimzewski, J.K., Reihl, B., Coombs, J.H., Schlittler, R.R.: Photon emission with the scanning tunneling microscope. Z. Phys. B: Condens. Matter **72**, 497 (1988)
17. Davis, J.J.: Molecular bioelectronics. Philos. Trans. R. Soc. A **361**, 2807 (2003)
18. Feenstra, R.M.: Cross-sectional scanning-tunneling-microscopy of III-V semiconductor structures. Semicond. Sci. Technol. **9**, 2157 (1994)
19. Rost, M.J., et al.: Scanning probe microscopes go video rate and beyond. Rev. Sci. Instrum. **76**, 053710 (2005)
20. Petersen, L., et al.: A fast-scanning, low- and variable-temperature scanning tunneling microscope. Rev. Sci. Instrum. **72**, 1438 (2001)
21. Wiesendanger, R.: Spin mapping at the nanoscale and atomic scale. Rev. Mod. Phys. **81**, 1495 (2009)
22. Bartels, L., et al.: Dynamics of electron-induced manipulation of individual CO molecules on Cu (111). Phys. Rev. Lett. **80**, 2004 (1998)
23. Crommie, M.F., Lutz, C.P., Eigler, D.M.: Confinement of electrons to quantum corrals on a metal-surface. Science **262**, 218 (1993)
24. Moresco, F., et al.: Conformational changes of single molecules induced by scanning tunneling microscopy manipulation: a route to molecular switching. Phys. Rev. Lett. **86**, 672 (2001)
25. Avouris, P.: Manipulation of matter at the atomic and molecular-levels. Acc. Chem. Res. **28**, 95 (1995)
26. Weiss, C., et al.: Imaging Pauli repulsion in scanning tunneling microscopy. Phys. Rev. Lett. **105**, 086103 (2010)

Scanning Tunneling Spectroscopy

Amadeo L. Vázquez de Parga and
Rodolfo Miranda
Department of Física de la Materia Condensada, Universidad Autónoma de Madrid and Instituto Madrileño de Estudios Avanzados en Nanociencia (IMDEA-Nanociencia), Madrid, Spain

Definition

Scanning tunneling spectroscopy (STS) is a technique that allows the study of the electronic structure of surfaces with atomic resolution.

Overview

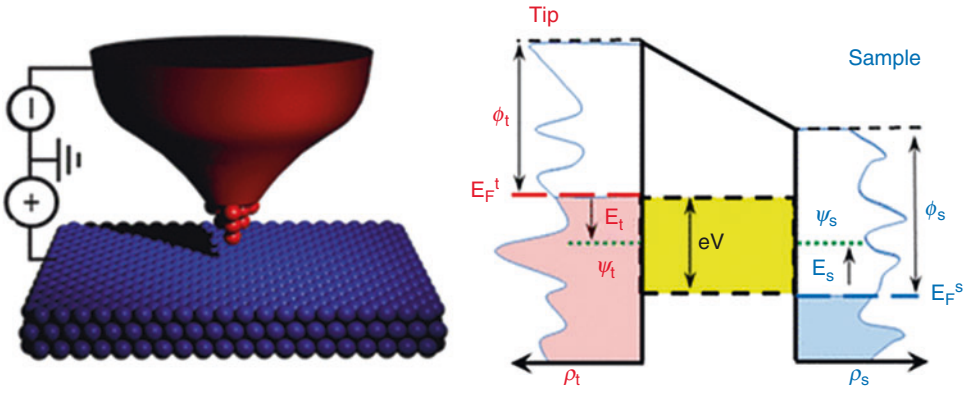
Scanning tunneling microscopy (STM) was historically the second technique that could image individual atoms one by one. It was invented in 1981–1982 by Gerd Binnig and Heinrich Rohrer [1], long after the technique of field ion microscopy (FIM) developed in 1951 by Erwin Müller [2].

In STM a sharp tip probes the surface of interest by allowing electrons to tunnel quantum mechanically between the tip and the surface. Because such tunneling is extremely sensitive to the distance between tip and surface, one gets high resolution perpendicular to the surface. Assuming a constant density of states on the surface, when the STM tip is scanned over the sample surface while keeping the tunneling current constant, the tip movement depicts the surface topography, because the separation between the tip apex and the sample surface is always constant. It is worth noting that STM not only converts the spatial change in the tunneling current into a highly detailed topographic image of surfaces with constant density of states but also the tunneling current changes with the available surface electronic states. This dependence of the tunneling current on the surface electronic structure together with the high spatial resolution of STM allows us to study the electronic structure of the surfaces with atomic resolution. The technique is known as scanning tunneling spectroscopy (STS).

Scanning Tunneling Spectroscopy Theory

Most theoretical treatments applied today to describe the tunneling process in an STM start from the formalism of the transfer Hamiltonian developed by Bardeen in 1961 [3] for the study of superconducting tunnel junctions. In this approach, the electronic structure and electron wave functions of both electrodes are calculated assuming no interaction between them and afterward the tunneling current is calculated [3]. Figure 1a shows the scheme of the tunnel junction in an STM where one of the electrodes is a tip. Figure 1b shows the energy diagram of the tunnel junction. In this diagram, the vertical axis represents energy. E_t and Ψ_t are the energy and wave function of the states of the electrode "tip" in the absence of electrode "sample." E_s and Ψ_s are the energy and wave function of the states of the electrode "sample" in the absence of electrode "tip." ϕ_t, ϕ_s, E_F^t, E_F^s, ρ_t, and ρ_s are the work functions, Fermi energies, and densities of states (DOS) of electrode "tip" and "sample," respectively, and V is the voltage applied to electrode "sample." When the distance between the electrodes is small enough, the overlap between their wave functions is significant, and the probability of electron transfer between the two electrodes by tunneling starts to be noticeable. In the absence of applied voltage, the Fermi levels of the two electrodes are aligned and no net tunneling current flows. However, by applying a voltage V, the Fermi levels move with respect to each other opening an energy window, eV, where electrons from one electrode can tunnel to the empty states of the other and, thus, the tunneling current starts to flow.

In 1983, Tersoff and Hamann applied the Bardeen's formalism to the STM, replacing one of the electrodes by a point [4–6]. The tip was



Scanning Tunneling Spectroscopy, Fig. 1 *Left panel*: Configuration of a positively biased tunnel junction in an STM. The bias voltage is applied to the sample and the tunneling current is measured on the tip. *Right panel*: Energy diagram of an STM tunnel junction. In this diagram, the *vertical axis* represents energy and the *horizontal axes* distance between tip and sample and density of states. E_t, Ψ_t, E_s, and Ψ_s are the energy and wave function of the states of the electrode tip and sample, respectively. ϕ_t, ϕ_s, E_F^t, E_F^s, ρ_t, and ρ_s are the work functions, Fermi energies, and densities of states (DOS) of electrode tip and sample, respectively, and V is the voltage applied to electrode sample

shaped like an s orbital centered at the tip position and the calculated tunneling matrix elements proved to be proportional to the amplitude of the wave functions of the sample at the position of the tip. If the distance between tip and sample is not very large (few angstroms), the bias voltage small, and the temperature low, the tunneling current can be written as follows:

$$I \propto \int_o^{eV} \rho_s\left(\vec{r}_s, E\right)\rho_t\left(\vec{r}_s, E - eV\right)T(E, eV, d, \phi)dE \tag{1}$$

where $\vec{r}_s$ is the tip position over the sample surface, d is the distance between tip and sample, and T is the transmission probability that depends on the energy of the states involved, the bias voltage applied between tip and sample, the distance between tip and sample, and the tunneling barrier height, which is related with the tip and surface work functions. This equation indicates that the tunneling process depends, for a given energy, on three interconnected parameters, i.e., the tunneling current I, the bias voltage V, and the tip sample separation d. Almost all attempts to explore these complex dependences of the tunneling current (and simultaneously extend the performance of STM) have been demonstrated by the late 1980s [7]. Scanning tunneling spectroscopy measures the relation between any two of while keeping fixed the third one. This gives three modes of spectroscopy measurements: (1) I-V curves, where the variation of the tunneling current with the bias voltage is measured for a fixed distance between tip and sample; (2) I-z curves, where the variation of the tunneling current with the distance between tip and sample is measured for a fixed bias voltage V; and (3) V-z curves, where the variations in the tip sample distance are measured as function of the bias voltage for a fixed tunneling current.

In these three modes, energy conservation for the tunneling electrons is assumed. If electrons change their energy during the tunneling process by an inelastic process, the inelastic electron tunneling spectroscopy (IETS) mode is possible. Finally, if the STM tip and the sample are magnetic, the tunneling current depends on the relative orientation of the magnetization of both tip and sample. This mode is called spin-polarized scanning tunneling microscope (SP-STM) and allows the study of magnetic properties with atomic resolution. In the following, the five modes will be discussed in detail.

Scanning Tunneling Spectroscopy Modes

I-V Curves

I-V measurements are the most widely used spectroscopic technique in STM experiments. If the tunneling current (Eq. 1) is differentiated with respect to the bias voltage, the following expression is obtained:

$$\frac{dI}{dV} \propto \rho_t(0)\rho_s(eV)T(eV, eV, d, \phi) + \int_0^{eV} \rho_t(E - eV)\rho_s(E)\frac{dT(E, eV, d, \phi)}{dV}dE + \int_0^{eV} \rho_s(E)\frac{d\rho_t(E - eV)}{dV}T(E, eV, d, \phi)dE$$

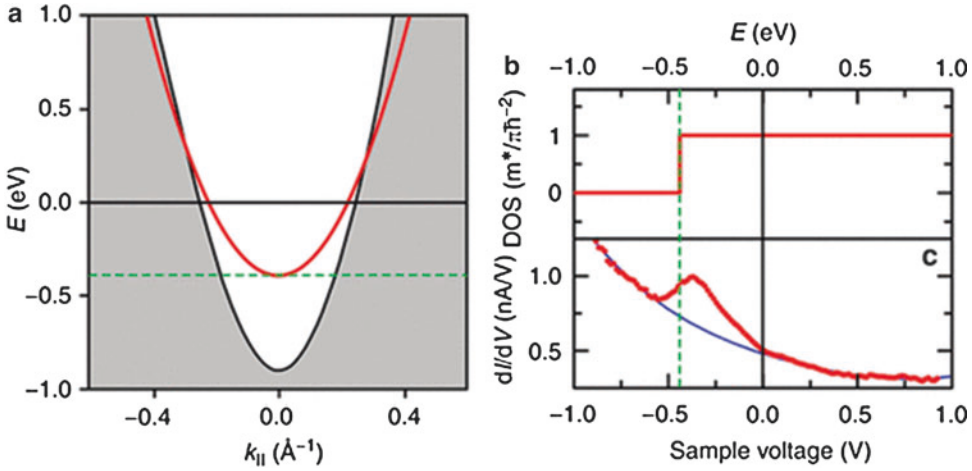
Assuming a constant density of states for the tip, the third term is zero, but, it should be mentioned that, often, the tip electronic states have a strong influence in the STS spectra [8]. Another common simplification is to assume that the transmission coefficients are constant in the voltage range explored in the measurement. Then the second term also vanishes and the expression becomes

$$\frac{dI}{dV} \propto \rho_t(0)\rho_s(eV)T(eV, eV, d, \phi)$$

In general, the experimentally determined differential tunneling conductance is widely accepted as a good approximation to the DOS of the surface (modulated by the specific transmission of the barrier) at an energy value of eV, with V being the bias voltage applied between tip and sample.

Figure 2 shows an STS experiment performed on the Cu(111) surface in which the transmission probability (see Eq. 1) can be seen to depend on the energy parallel to the surface. Figure 2a shows the bulk band projection of Cu(111) along the $\overline{\Gamma M}$ direction of the surface Brillouin zone. Bulk bands are represented in gray and the projectional bandgap in white. The gray line corresponds to the dispersion relation of Shockley surface state of Cu (111). The surface state can be seen as a two-dimensional (2D) electron gas with the bottom of the band at -0.44 eV and an effective mass $m^* = 0.40\ m_e$. Therefore, disregarding any contribution from the bulk electronic structure, the LDOS expected around the Fermi level of Cu (111) is a step function centered at the bottom of the band. This step function is shown in Fig. 2b. Figure 2c shows an STS spectrum taken on the Cu (111) surface. The experimental data correspond to the dots. The spectrum has a peak at -0.38 eV superimposed on a background which decays with increasing energy. The bottom of the surface state band corresponds to the point halfway up the peak (dashed line in Fig. 2c). The line in the graphic is a fit to the background that reflects the contribution of the bulk states. As the bias voltage approaches the Fermi level from below, the $k_{\parallel}$ of the accessible bulk states increases, so that they present a smaller effective perpendicular energy which means a smaller transmission probability and, thus, they contribute less to the spectrum. On the other hand, the peak at -0.38 eV corresponds to the sharp increase in the LDOS associated to the bottom of the surface state, superimposed on the background due to the bulk bands. Although the LDOS of a 2D electron gas is a step function to a constant value, the peak in the spectrum instead of staying constant for energies higher than the bottom of the band (-0.44 eV) decreases with increasing energy. This reduction in the signal with energy reflects the dispersion of the surface state, i.e., the fact that the $k_{\parallel}$ of the surface state increases also when the energy increases. Accordingly, the transmission probability (and the signal in the spectrum) gets smaller.

Experimentally in order to record I-V curves, the distance between tip and sample has to be kept constant during the measurement time. This can be done in different ways. In one of them, the tip is placed at a desired position on the surface and at the desired distance from the surface. The distance between tip and sample is dictated by the values of the tunneling current and the bias voltage used in the topographic image. The feedback circuit, which keeps the tunneling current constant adjusting the tip sample distance, is disconnected and then the voltage V is ramped and the tunneling current is recorded over the desired bias voltage range. The dI/dV values are obtained by numerical differentiation of the I-V curves. If the measurement is repeated in every pixel of a topographic image, the method is called *current imaging*



Scanning Tunneling Spectroscopy, Fig. 2 (**a**) The bulk band projection of Cu(111) along the $\overline{\Gamma M}$ direction of the surface Brillouin zone. Bulk bands are represented in *gray*, and in *white* the projectional bandgap. The *gray line* corresponds to the dispersion relation of surface state of Cu (111). (**b**) Expected density of states for a 2D electron gas. (**c**) An STS spectrum taken on the surface Cu(111). The experimental data correspond to the *dots*

tunneling spectroscopy (CITS) and provides with a map of the spatial distribution of the LDOS on the surface.

Another method is to detect directly the *dI/dV* signal using a lock-in amplifier. In order to do that, a small high-frequency sinusoidal signal, $V_{\text{mod}}\sin(\omega t)$, is superimposed on the bias voltage between tip and sample. The modulation causes a sinusoidal response in the tunneling current, and the amplitude of the modulated current is sensitive to *dI/dV*. For a small applied sinusoidal signal, the modulated current can be Fourier decomposed on the applied modulation frequency ω:

$$I(V_{\text{bias}},t) = I(V_{\text{bias}}) + \frac{dI(V_{\text{bias}})}{dV}V_{\text{mod}}\sin(\omega t) + \frac{d^2I(V_{\text{bias}})}{dV^2}\frac{V_{\text{mod}}^2}{4}\sin(2\omega t) + \cdots \quad (2)$$

The first harmonic, which is proportional to the differential conductance (*dI/dV*), can be extracted by means of lock-in detection, and the spatial variation of the *dI/dV* signal can be mapped in certain area of the surface. During a constant current topographic image, the *dI/dV* signal is simultaneously recorded in each point of the image at a certain bias voltage. The result is a map that reflects the LDOS of a surface area at a defined energy *eV*. Since the feedback loop is connected during the topographic image, the frequency of the modulated signal needs to be higher than the cutoff frequency of the feedback loop response in order to keep the tip distance constant during the acquisition of the data.

I-Z Curves

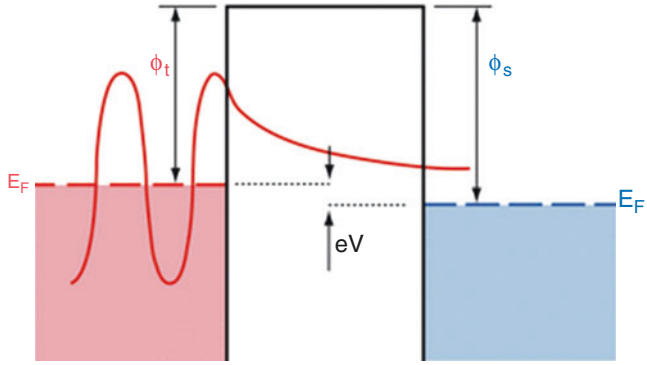
In the simplest theoretical treatment of the tunneling process for a metal-vacuum-metal tunnel junction [7] where the bias voltage is much smaller than the work function (assumed to be the same for both metals), the problem is reduced to a square potential barrier as shown in Fig. 3. The wave function that describes an electron in the tunneling barrier is given by:

$$\psi(z) = \psi(0)e^{-\kappa z}$$

where

$$\kappa = \frac{\sqrt{2m(\phi)}}{\hbar}$$

is the decay constant that describes the probability of finding the electron along the $+z$ direction and depends on the surface work function (ϕ).



Scanning Tunneling Spectroscopy, Fig. 3 One-dimensional metal-vacuum-metal tunnel junction. Both the sample, in *red*, and the tip, in *blue*, are modeled as semi-infinite pieces of free electron metals. The tunneling probability depends on the exponential decay of the electron wave function into the vacuum barrier. The bias voltage is small enough to consider a square barrier

The tunneling probability for a given distance, d, between tip and sample is

$$P \propto |\psi(0)|^2 e^{-2\kappa d}$$

and the tunneling current is proportional to the number of occupied states available in the energy interval defined by the bias voltage applied between tip and sample, and therefore, the tunneling current can be written as follows:

$$I \propto \sum_{E_n=E_F-eV}^{E_F} |\psi_n(0)|^2 e^{-2\kappa d}$$

If the bias voltage is small enough to consider the density of states constant, the equation can be written in terms of the local density of states (LDOS) at the Fermi level. At a position d and energy E, the LDOS of the sample can be expressed as

$$\rho_s(d,E) \equiv \frac{1}{\epsilon} \sum_{E_n=E-\epsilon}^{E} |\psi_n|^2$$

for small enough ε. The tunneling current in terms of the surface LDOS at the Fermi level is

$$I \propto V \rho_s(0,E_F) e^{-2\kappa d}$$

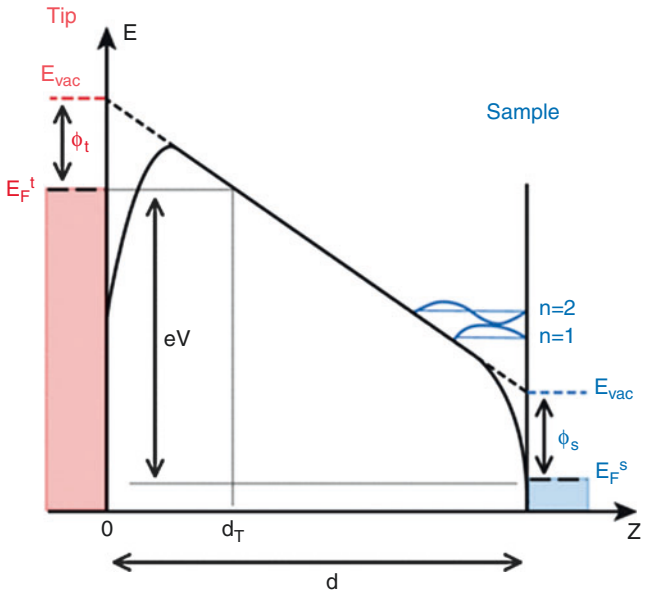
Assuming a typical value of 4 eV for the work function, $\kappa = 1.025$ Å^{-1}. From the expression above, the dependence of the logarithm of the tunneling current with respect to distance is a measure of the work function or, more precisely, of the tunneling barrier height. The corresponding expression is

$$\phi \approx 0.95 \left(\frac{dlnI}{dz} \right)^2$$

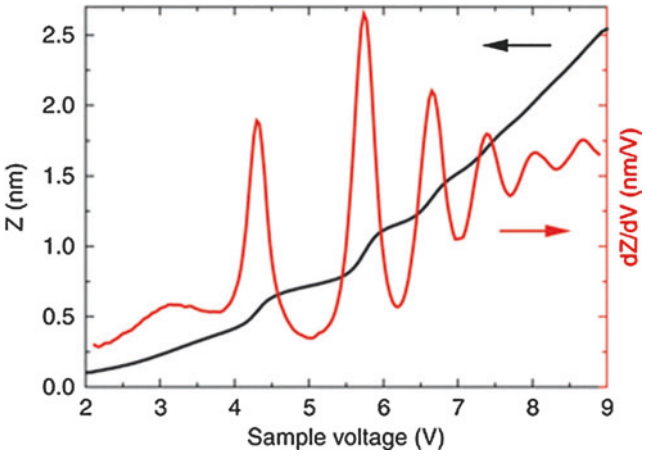
The measurement of the apparent barrier height can be carried out by approaching or retracting the tip from the sample and recording the tunneling current. In order to measure the spatial change in apparent barrier height, a small modulation in the separation between the tip and the sample is introduced at high frequency, and the modulated tunneling current is measured using a lock-in amplifier. This type of measurement gives the apparent barrier height at certain bias voltage and at certain distance from the surface. Measurements performed with different bias voltages or different tip sample distances may give different values for the apparent barrier height [7]. It is important to realize that the apparent barrier height is different from the work function in traditional surface science but is closely related; the apparent barrier height measures the spatial correlation of the overlap between the wave functions of the tip and the sample.

Z-V Curves

In Z-V measurements, the bias voltage is ramped at a fixed tunneling current, and the tip sample



Scanning Tunneling Spectroscopy,
Fig. 4 Diagram of the tunnel junction when the positive bias voltage (eV) is larger than the work function of the sample. The electrons from the tip tunnel through a narrow (d_T) triangular potential barrier to be afterward trapped in a triangular potential well. The first two quantum well states are shown



Scanning Tunneling Spectroscopy,
Fig. 5 *Black curve*, tip displacement as function of the tunneling bias voltage. The field emission resonances appear as wiggles on the curve. *Gray curve*, dZ/dV, the field emission resonances appear as well-defined peaks that allow a precise determination of the energy

separation is constantly adjusted. When the applied voltage exceeds the sample or tip work function (depending on the sign of the bias voltage), there is a transition from the vacuum tunneling regime to the field emission regime. In the field emission regime, a triangular potential well is formed between tip and sample due to the bias voltage applied in the tunneling junction. In this triangular potential well, the existence of quantum well states leads to resonances in the electron transmission at certain energies, as illustrated in Fig. 4.

These transmission resonances show up as wiggles in the Z-V data (black curve in Fig. 5) and are closely related to the image states. Image states are unoccupied states bound by the classical image charge response of metallic surfaces and have a free electron-like dispersion parallel to the surface. The inverse dependence on distance from the surface of the image potential leads to a Rydberg-like series of states that converges to the continuum at the vacuum level (E_{vac}). Inverse photoemission studies of the image potential states have shown experimentally that the energy

position of the Rydberg series is tied to the local surface potential of the material. In STM, the electric field across the tunnel junction causes a Stark shift of these states, expanding the image state spectrum into a resonance spectrum associated with the triangular potential well (Fig. 5). Following the analysis performed by Gundlach in the 1960s [9], the resulting energy spectrum can be written as follows:

$$E_n = \phi + \alpha(n - 0.25)^{2/3} F^{2/3}$$

where ϕ is the surface work function, α is a constant, F is the electric field between tip and sample, and n is the quantum number of the states. These field emission resonances (FERs) were experimentally observed in field ion microscopy (FIM) by Jason [10] and with an STM by Binnig et al. [11] and since then have been used to chemically identify different transition metals on surfaces, to obtain atomic resolution on insulating surfaces (e.g., diamond), or to study local changes in the surface work function [12].

The experiments are typically performed with the feedback loop on to keep the current constant. The tip movement is recorded as a function of the bias voltage, and afterward, the curves are numerically differentiated to obtain the energy position of the field emission resonances, as shown in Fig. 5 (red curve).

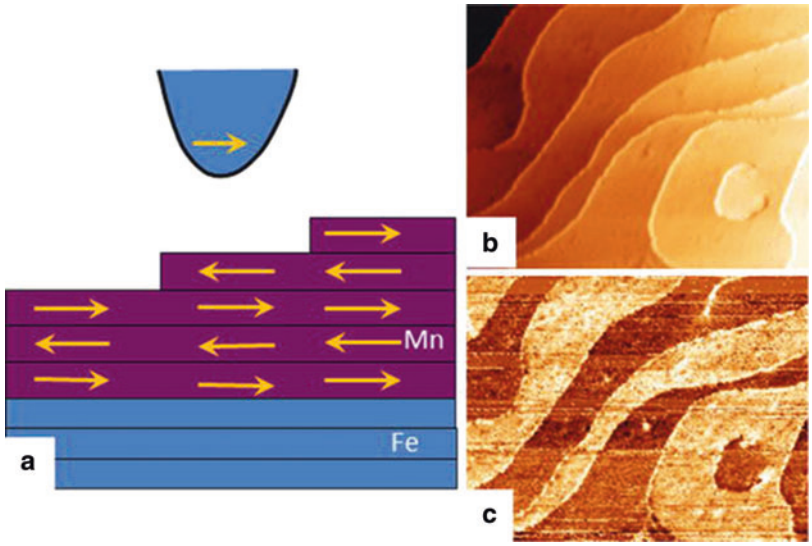
Spin-Polarized Tunneling Spectroscopy

In 1975 Julliére [13] discovered spin-dependent tunneling between two planar ferromagnetic electrodes separated by an insulating tunnel barrier, which has become the basis for the development of magnetic random access memories and the spin-polarized version of the STM. In fact, the tunneling current between a magnetic sample and an STM tip (covered with a magnetic thin film) shows an asymmetry in the spin population. The magnitude of the tunneling conductance between two magnetic electrodes with directions of the respective magnetization differing by a certain angle depends on the cosine of this angle. Spin-polarized tunneling with an STM was observed in 1990s by Wiesendanger [14]. The spectroscopic mode of spin-polarized STM is based on using the different intensity of certain features in differential conductance spectra as source of contrast to image magnetic domains and domain walls with atomic resolution [14].

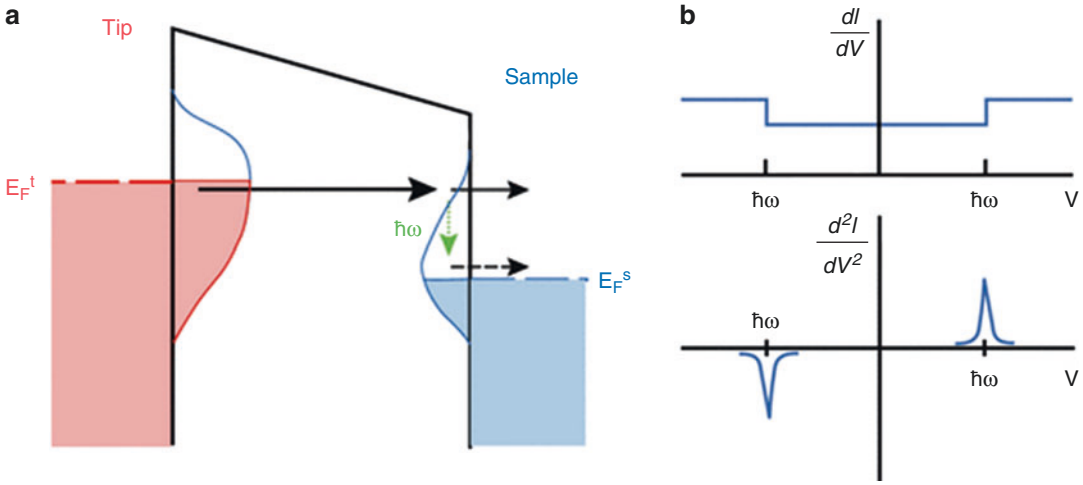
Magnetic domain observations in antiferromagnetic materials have been difficult in the past due to the limited number of experimental techniques that are sensitive to domain states in antiferromagnetic crystals. Another aspect that hampered the studies of antiferromagnetic materials was the limited spatial resolution of the available techniques. Mn and Cr crystals or thin films of these materials exhibit quite complex spin structures. The crystallographic structure of Mn thin films grown on Fe(001) is body-centered tetragonal. In this crystallographic structure, the Mn magnetic structure is layered antiferromagnetic. This means that the magnetization orientation rotates by 180° in every layer, as can be seen in Fig. 6a. The tip magnetization direction is constant in the experiment, and the magnetization direction of the sample rotates by 180° every time the tip crosses a step on the surface. The change in the magnitude of the tunneling conductance can be measured on every pixel of the topographic image, and an image of the magnetic domains is obtained, as can be seen in Fig. 6b, c. In those domains where the magnetization of tip and sample is aligned, the tunneling conductance is higher (brighter color), and on those areas where antiparallel, the tunneling conductance is lower (darker colors).

Vibrational Spectroscopy (Inelastic Electron Tunneling Spectroscopy)

In the previous sections, elastic tunnel processes, where the tunneling electrons do not change their energy, have been discussed. However, in certain cases, i.e., for molecules adsorbed on surfaces or samples with easy excitation of phonons, there is a small fraction of electrons that lose energy in the tunneling process [15]. For bias voltages larger than corresponding the quantum of vibration, $\hbar\omega$, a new tunneling channel, i.e., inelastic channel, opens up (as illustrated in Fig. 7a). The inelastic channel acts in addition to the elastic channel and increases slightly the differential conductance (dI/dV) of the junction (Fig. 7b upper panel).



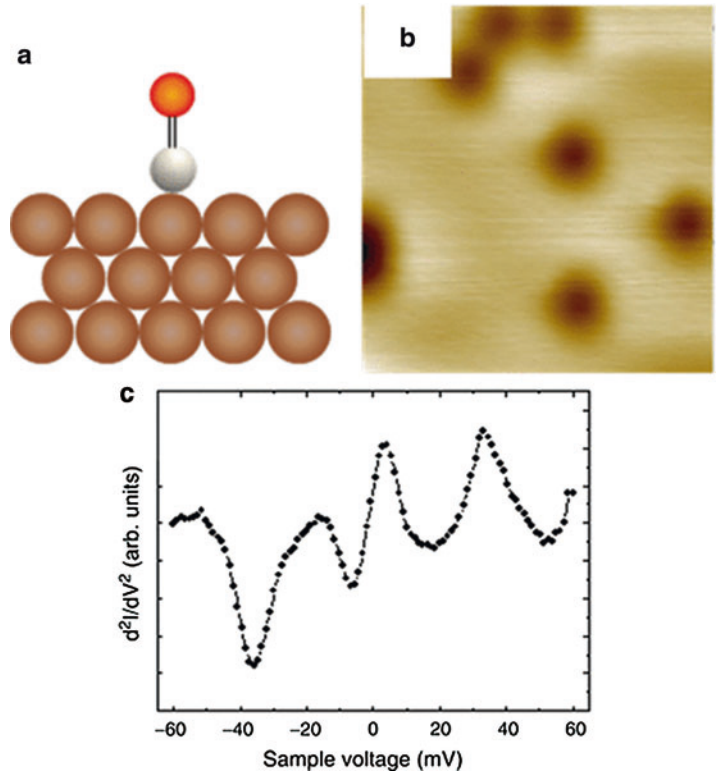
Scanning Tunneling Spectroscopy, Fig. 6 (**a**) Model showing the layered antiferromagnetic structure of Mn films grown on Fe (001). (**b**) STM topographic (100 × 78 nm) image of 6.5 mL of Mn grown on Fe (001). (**c**) Spatially resolved spectroscopic image measured simultaneously with the topography shown in (**b**) where the dI/dV signal shows low and high levels depending on the relative directions of magnetization of tip and surface terrace revealing the topological antiferromagnetic order of the Mn(001) surface



Scanning Tunneling Spectroscopy, Fig. 7 (**a**) Energy distance diagram of the tunneling processes with an applied bias voltage V. When eV is larger than the energy of the molecular vibration (ħω), empty final states at the Fermi level of the sample become accessible, and the inelastic channel opens up. (**b**) The opening of the inelastic channel causes a sharp increase in the tunneling conductance (*upper panel*) or peaks in the second derivative (*lower panel*). The activation channel is symmetric with respect to the Fermi level

Although vibrations of molecules were detected in the 1960s by tunneling in extended tunneling junctions with insulating layer spray coated with molecules, the use of STM facilitates the acquisition of vibrational spectra in single molecules in well-characterized environments and was pioneered by Wilson Ho [15].

In practice the change in conductance is smaller than 10 % and can be detected only under very severe conditions of stability of the tunnel junction and energy resolution (i.e., with the STM at low temperatures). The vibrational modes are detected as peaks in the second derivative of the tunneling current (Fig. 7b lower panel) measured by means of lock-in techniques. Equation 2 shows that the magnitude of the second harmonic of the tunneling current is proportional to d^2I/dV^2. In practice, a small (≈3 mV) ac



Scanning Tunneling Spectroscopy, Fig. 8 (**a**) Model of the adsorption of CO molecules on Cu(111). The molecule adsorbs perpendicular to the surface with the carbon (in *gray*) chemically bonded to the surface copper atoms. (**b**) STM topographic image measured using a tunneling current of 1 nA and a sample bias voltage of 0.25 V. (**c**) Vibrational spectra of CO on Cu(111): the peaks at ±5 mV and ±35 mV are due to excitation of the CO frustrated rotation and translation with respect to the Cu(111) surface

component is added to the bias voltage, the dc component of the bias voltage is scanned across the selected energy range, and the variations in d^2I/dV^2 are recorded. The width of the peaks is given by the Fermi energy distribution, i.e., the FWHM is $3.5k_BT$ (1.2 mV at 4 K).

A common observation in IETS spectra is that peaks at certain values of positive voltages appear as *dips* at opposite polarity. The symmetry position with respect to the zero bias of the features observed is a fingerprint of their inelastic origin. Differences in the density of states of the electrodes, however, may change the peak intensity. The observed symmetry implies that the inelastic processes are accessible for electron tunneling on both directions. Selection rules for which modes are detectable, unlike Raman or infrared spectroscopy, seem to depend on the symmetry of the molecular state involved in the tunneling process.

A well-studied system are CO molecules adsorbed on Cu(111). It is known that the CO molecules adsorb on the surface on top of the copper atoms with the oxygen pointing toward the vacuum and the carbon chemically bonded to the copper surface, as shown in Fig. 8a. Figure 8b shows a topographic STM image of several CO molecules adsorbed on Cu(111) at 4.6 K measured with a tunneling current of 1 nA and a bias voltage of −0.25 V. With these values, the CO molecules are imaged as a round depression because the chemical bond between the molecule and the Cu (111) surface reduces the electron density around the Fermi level. In order to measure the vibrational spectra, the tip is positioned over the center of the CO molecule. With the feedback off, the sample bias voltage is ramped over the range of the vibrational peaks while a sinusoidal bias modulation is superimposed. The derivative of the conductance exhibits peaks at the molecular vibration energy. For CO on Cu(111), the vibration spectra are characterized by two features at about 5 and 35 meV, as can be seen in Fig. 8c. These peaks are assigned to the two degenerated transverse vibration modes: the frustrated translation and the frustrated rotation, respectively.

Cross-References

- ▶ Field Electron Emission from Nanomaterials
- ▶ Scanning Tunneling Microscopy
- ▶ Surface Electronic Structure

References

1. Binnig, G., Rohrer, H.: Scanning tunneling microscopy. Helv. Phys. Acta **55**, 726 (1982)
2. Müller, E.W.: Das Feldionenmikroskop. Z. Phys **31**, 136 (1951)
3. Bardeen, J.: Tunneling from many particle point of view. Phys. Rev. Lett. **6**, 57 (1961)
4. Tersoff, J., Hamann, D.R.: Theory and application for the scanning tunneling microscope. Phys. Rev. Lett. **50**, 1998 (1983)
5. Tersoff, J., Hamann, D.R.: Theory of the scanning tunneling microscope. Phys. Rev. B **31**, 805 (1985)
6. Lang, N.D.: Spectroscopy of single atoms in the scanning tunneling microscope. Phys. Rev. B **34**, 5947 (1986)
7. Chen, C.J.: Introduction to Scanning Tunneling Microscopy. Oxford University Press, New York (1993)
8. Vázquez de Parga, A.L., Hernán, O.S., Miranda, R., LeviYeyati, A., Mingo, N., Martín-Rodero, A., Flores, F.: Electron resonances in sharp tips and their role in tunneling spectroscopy. Phys. Rev. Lett. **80**, 357 (1998)
9. Gundlach, K.H.: Zu berechnung des tunnelstroms durch eine trapezförmige potentialstufe. Solid State Electron. **9**, 949 (1966)
10. Jason, A.J.: Field induced resonance states at a surface. Phys. Rev. **156**, 266 (1966)
11. Binnig, G., Frank, K.H., Fuchs, H., Garcia, N., Reihl, B., Rohrer, H., Salvan, F., Williams, A.R.: Tunneling spectroscopy and inverse photoemission: image and field states. Phys. Rev. Lett. **55**, 991 (1985)
12. Borca, B., Barja, S., Garnica, M., Sánchez-Portal, D., Silkin, V.M., Chulkov, E.V., Hermanns, C.F., Hinarejos, J.J., Vázquez de Parga, A.L., Arnau, A., Echenique, P.M., Miranda, R.: Potential energy landscape for hot electrons in periodically nanostructured graphene. Phys. Rev. Lett. **105**, 036804 (2010)
13. Julliere, M.: Tunneling between ferromagnetic films. Phys. Lett. **54A**, 225 (1971)
14. Wiesendanger, R.: Soin mapping at the nanoscale and atomic scale. Rev. Mod. Phys. **81**, 1495 (2009)
15. Person, B.N.J., Baratoff, A.: Inelastic electron tunneling from a metal tip: the contribution from resonant processes. Phys. Rev. Lett. **59**, 339 (1987)
16. Stipe, B.C., Rezaei, M.A., Ho, W.: Single molecule vibrational spectroscopy and microscopy. Science **280**, 1732 (1998)

Scanning X-Ray Diffraction Microscopy (SXDM)

- ▶ Selected Synchrotron Radiation Techniques

Scanning-Probe Lithography

- ▶ Dip-Pen Nanolithography

Scincus officinalis

- ▶ Friction-Reducing Sandfish Skin

Scincus scincus

- ▶ Friction-Reducing Sandfish Skin

Scrolled Nanostructure

- ▶ Nanorobotics for NEMS Using Helical Nanostructures

Selected Synchrotron Radiation Techniques

Antoine Barbier[1], Cristian Mocuta[2] and Rachid Belkhou[2]
[1]CEA-Saclay, DSM/IRAMIS/SPCSI, Gif-sur-Yvette, France
[2]Synchrotron SOLEIL, L'Orme des Merisiers, Gif-sur-Yvette, France

Synonyms

Small angle X-ray scattering in grazing incidence geometry; Scanning X-ray diffraction microscopy (SXDM); Soft X-ray microscopy; Spectromicroscopy; X-ray diffraction with micron sized X-ray beams

Definition

Synchrotron radiation is produced when highly energetic charged particles are deviated in a magnetic field. The generated spectrum consists of a large energy range from infrared to gamma X-rays. Harder X-ray ranges require larger synchrotron storage rings. Recent rings offer simultaneously energy tunability for photons with high brightness, high photon fluxes, and low divergence [1]. Photon energy tunability and variable polarization provide chemical, electronic structure and/or magnetic sensitivity. The tunable penetration depth using glancing incident and/or exit scattering angles enables the study of buried interfaces. Importantly, X-ray techniques, especially at high photon energies, can be used with various sample environments including (and not limited to) high pressure, high temperatures, ultrahigh vacuum, liquids, and low temperatures. This unique combination of characteristics allowed developing genuine experimental methods and techniques particularly suitable to investigate micro- and nanostructures that are intended to complement laboratory experiments in order to obtain in-depth information on specific issues.

The use of synchrotron radiation has also some drawbacks. The available beam time is limited. Organic and biological samples are often unable to bear the highly intense X-ray beams. X-ray diffraction techniques measure scattered intensities, and the phase information is lost; unless the coherence of the synchrotron radiation or statistical methods are used the data interpretation requires fit procedures and the input of some models.

The goal of the present essay is here to exemplify some approaches proven to be relevant to investigate nanostructures of current interest. The interested reader may consider more general textbooks or reviews dedicated to synchrotron radiation techniques [2, 3]. Within the field of nanostructure investigations three techniques that are particularly illustrative were chosen: Grazing Incidence Small Angle X-ray Scattering (GISAXS), micro X-Ray Diffraction (μ-XRD), and X-ray Photo-Emission Electron Microscopy (X-PEEM). These techniques are essentially nondestructive, noninvasive, and allow investigating surfaces and buried interfaces. GISAXS and μ-XRD allow studying samples in various environments (e.g., temperature, gas or liquid cells for "real" environments, external magnetic or electric fields, etc.). X-PEEM is more limited within this respect and requires good vacuum conditions in order to detect electrons but provides full field real space imaging capabilities.

Overview

X-ray scattering probes materials properties and, depending on the technique used, can be sensitive to chemical species, magnetic and/or crystallographic ordering, and even density fluctuations near surfaces and buried interfaces. The process can be elastic (i.e., photon energy conservative) or inelastic, specular or diffuse [4], and can be used in spectroscopies that provide chemical information about the sample composition. Photon energy tunability can be used to greatly enhance the chemical contrast between several chemical species (a specific contrast is obtained when the photon energy is tuned to a given electron level edge), i.e., anomalous x-ray scattering conditions [2, 3].

For the analysis of materials at the micro- and the nanoscale, well-established microscopy-based techniques can probe different properties at short – down to atomic – length scale. Techniques like the transmission electron microscopy (▶ TEM), scanning probe methods (scanning electron microscopy ▶ SEM, atomic force microscopy ▶ AFM, microphotoluminescence μPL, etc.) have an undeniable contribution in understanding what is happening at the small scale for nanostructured materials. From some of their drawbacks, one could mention here their surface sensitivity, invasive sample preparation methods, presence of vacuum environment, or sensitivity to external fields (magnetic, electric). Alternative X-ray-based techniques can offer complementary insights and overcome some of these drawbacks. Some examples are detailed hereafter.

GISAXS is a powerful technique to study nanostructured surfaces, thin films, and assemblies of nano-objects, combining small-angle

X-ray scattering (SAXS) and the surface sensitivity of grazing incidence diffraction (GID) [5]. The method is fully nondestructive for materials science samples and gives access to morphological (e.g., size/shape, dispersion) statistical information of an assembly of nano-objects (separated by an average distance in the several 10–100 nm range), averaged over a large area (several square millimeters); it characterizes the "mean" object in reciprocal space. During a GISAXS experiment the incident photon energy is mainly chosen with respect to the reciprocal space resolution that is expected. For chemical contrast studies the photon energy may be tuned to the absorption edge of a given element. The pattern is conveniently observed on an area detector (2D camera) because at high photon energy the Ewald sphere region corresponding to these small scattering angles can reasonably be approximated by the plane tangent to the sphere.

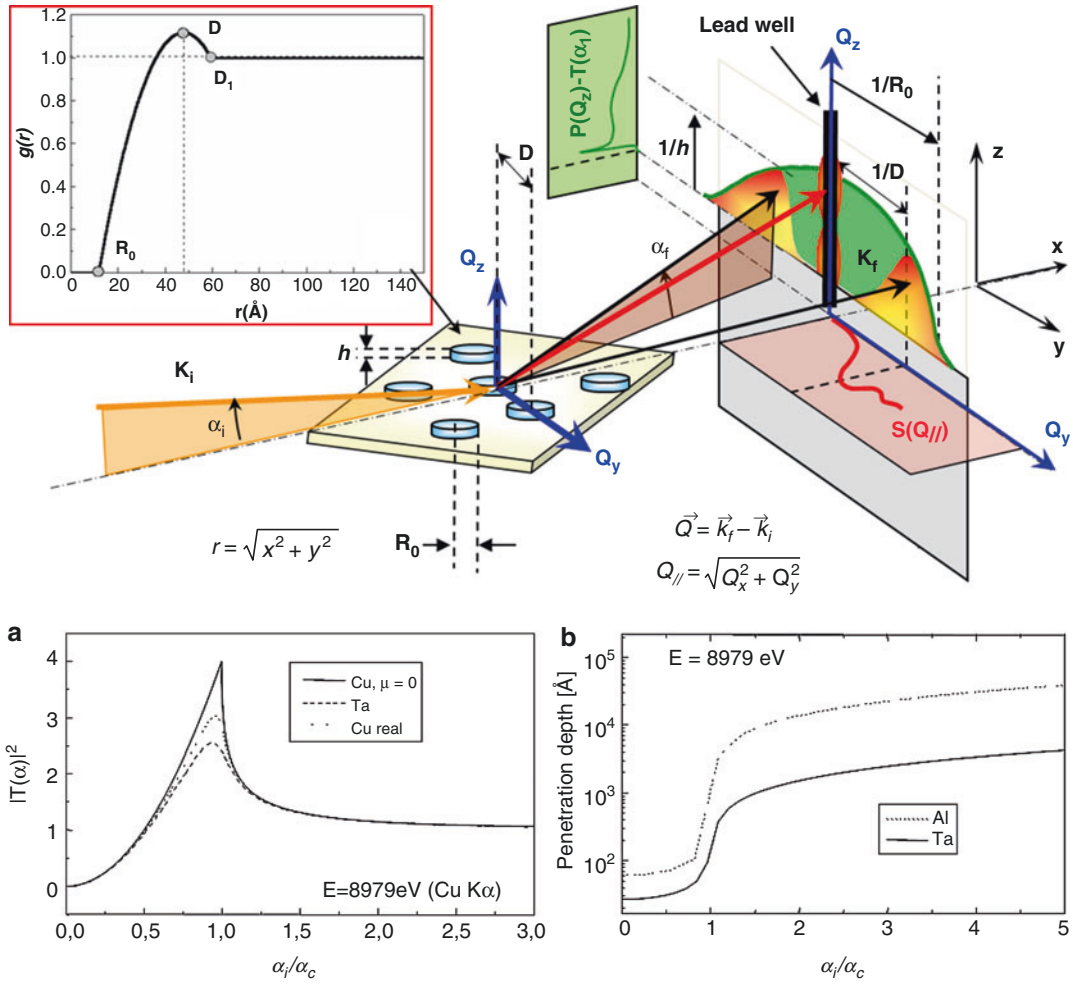
Microbeam X-ray diffraction (**μ-XRD**) experiments use highly intense and very small (focused to micron or submicron sizes) X-ray beams in order to address locally the properties of matter by diffraction [4, 6]. The lateral resolution is essentially given by the size of the X-ray spot. XRD uses typically hard X-rays (∼10 keV energy range). It is a scanning microscopy-like approach and consists of investigating a single structure at a time. The signal originating from individual micro/nanostructures is recorded by using local probe techniques, and a two-dimmensional raster image of the sample is obtained. It can give access to crystalline structure, strain, defects, composition (by exploiting the variation of the lattice unit cell size with the composition), etc.; the strain can be measured in a simple experiment with an accuracy (10^{-4}) hard to achieve with other techniques. The resulting data (images or maps in the so-called reciprocal space, which is the Fourier transform space) are however usually not straightforward to visualize and understand (as compared to real space imaging techniques).

XPEEM (X-ray PhotoEmission Electron Microscopy) spectromicroscopy is a derivative of the classical PEEM/LEEM [7, 8] employing parallel imaging techniques making use of special electron optics with resolution in the 10 nm ranges. If a photon energy just above the photothreshold is used, the photoelectron yield is mainly determined by the differences in the work function ϕ of the sample. The local variations of ϕ result in images with high contrast. This ultraviolet (UV)-PEEM mode of operation is ideally suited to study surface chemical reactions in real time. It allows accessing well-established techniques like Ultraviolet Photoemission Spectroscopy (UPS), X-ray Photoemission Spectroscopy (XPS), and X-ray Absorption Spectroscopy (XAS) at the nanoscopic level, thus leading to element selective imaging. Note that photoemission techniques require an additional analyzer enabling energy filtering of the photoemitted electrons. Information on the spatial distribution of the electronic structure, chemical composition and nature, or the local magnetization at the surface can be obtained. This technique is well suited to investigate modern nanostructures, fabricated using various methods (bottom-up or top-down, lithography, self-organized growth methods, etc.), at the nanoscopic level combining good spatial (few nm) and time (subnanosecond) resolutions. The increasing sophistication level of the nanostructures prompted considerable progress in high-resolution microscope probes that are sensitive to the compositional, electronic, chemical, and magnetic microstructure.

Grazing Incidence X-ray Scattering (GISAXS)

Geometry

In a typical GISAXS experiment the incident beam, of wave vector $\mathbf{k}_i$, makes a small incident angle α_i with respect to the sample surface. The incident beam is supposed to be monochromatic, parallel, and free from background contributions; the beam is cleaned using scattering slits whenever necessary. The direction of the scattered wave vector $\mathbf{k}_f$ spans the (Q_y, Q_z) reciprocal plane (Fig. 1, top) that includes information about island size, height, and inter-island distance (or correlation). The pattern is observed on a 2D detector as far as possible from the sample in order to increase the reciprocal space resolution. Beam



Selected Synchrotron Radiation Techniques, Fig. 1 (*Top*) Typical geometry of a GISAXS experiment. $\mathbf{k}_i$ and $\mathbf{k}_f$ are the incident and scattered wave vectors, respectively, yielding the momentum transfer (i.e., the reciprocal space vector) $\mathbf{Q} = \mathbf{k}_f - \mathbf{k}_i$. The angles α_i, α_f and $2\theta_f$ are related to the components of the momentum transfer, either parallel (Q_x and Q_y) or perpendicular (Qz) to the sample surface, by the equations $\mathbf{k}_i = \mathbf{k}_0[\cos(\alpha_i), 0, -\sin(\alpha_i)]$, corresponding to the wavelength $\lambda = 2\pi/k_0$ and $k_f = k_0[\cos(\alpha_f) \cdot \cos(2\theta_f), \cos(\alpha_f) \cdot \sin(2\theta_f), \sin(\alpha_f)]$ of equal modulus. $\mathbf{k}_0$ is known to be the incident wave vector modulus. (*Bottom*) (**a**) Fresnel transmission coefficients calculated at Cu Kα photon energy for non-absorbing Cu, Ta and Cu using real δ and β values. (**b**) Penetration depth of X-rays calculated at Cu Kα photon energy for Al and Ta. The incidence angle scale is normalized by the critical angle of total external reflection

stops hiding the direct and specular beam and rod are necessary because of the limited dynamical range of detectors (charge-coupled device, CCD camera in this particular case). The scattered intensity distribution depends on α_i, Q_y, and Q_z.

The experiment can be run in transmission for thinned samples or in grazing incidence conditions. Using focused X-ray beams and transmission geometries can provide a scanning microscopy approach valuable to investigate inhomogeneities in the sample [9]. The grazing incidence geometry is better adapted for the study of nano-objects and their correlations without sample preparation.

Grazing Incidence

To understand the usefulness of grazing incidence some elements of photon-matter interaction have

to be recalled. Consider a sharp interface between vacuum and a surface of a material of wavelength-dependent index of refraction n.

The incident beam is supposed to be a linearly polarized plane wave that impinges on this surface at a shallow angle α_i with wave vector $\mathbf{k}_i$; the reflected and transmitted beams leave at angles α_f and α_t with wave vectors $\mathbf{k}_f$ and $\mathbf{k}_t$ respectively. Snell's law writes as

$$\cos(\alpha_t) \cdot n = \cos(\alpha_i) \text{ and } \alpha_f = \alpha_i \quad (1)$$

As long as $n > 1$ total reflection cannot occur when light travels from the vacuum to the material, even if $\alpha_i = 0$. Fortunately, and unlike visible light, hard X-rays have an index of refraction less than unity and can be written as

$$n = 1 - \delta - i\beta \quad (2)$$

with $\delta = \frac{\lambda^2}{2\pi} r_e \rho_e$

$$\text{and} \quad \beta = \frac{\lambda}{4\pi}\mu \quad (3)$$

in which λ is the X-ray wavelength, r_e the Bohr atomic radius, ρ_e the electronic density of the material, and μ the material linear absorption coefficient at the considered wavelength.

Consequently, for hard X-rays, the transmitted beam will be deflected always toward the sample internal surface, and total external reflection occurs on the vacuum side with a critical angle for total external reflection $\alpha_c \approx \sqrt{2\delta}$ in the 0.1–0.6° range because δ and β are, respectively, in the 10^{-5} and 10^{-6} ranges.

When $\alpha_i < \alpha_c$, the component of the transmitted wave vector normal to the surface becomes imaginary. The refracted wave is exponentially damped as a function of the distance below the surface and is an evanescent wave traveling parallel to the surface.

Considering the Fresnel transmission coefficient that writes as

$$T = |t|^2 = \left| \frac{2\sin(\alpha_i)}{\sin(\alpha_i) - \sqrt{n^2 - \cos(\alpha_i)}} \right|^2$$

and the penetration depth given by

$$\Lambda(\alpha_i) = \frac{1}{-2k_0 \,\mathrm{Im}\left(\sqrt{\alpha_i^2 - \alpha_c^2 - 2i\beta}\right)} \quad (4)$$

one can see that the surface sensitivity is strongly enhanced when α_i become close to α_c because the incident and reflected waves are nearly in phase. The penetration depth of the x-rays becomes nanometric, whatever the considered material, and the Fresnel transmission coefficient is maximal (see Fig. 1, bottom). One may note that higher absorption coefficients reduce slightly the Fresnel coefficient maximum and conversely decrease the penetration depth at any angle. Importantly, for $\alpha_i > 3 \times \alpha_c$ surface effects become negligible.

Grazing incidence scattering geometries provide simultaneously enhanced surface sensitivity and limit the unwanted bulk scattering (diffuse or not) that may otherwise overcome the surface or nanostructure feeble signal. It opens additional experimental possibilities as compared to simple specular reflectivity for which incidence and exit angles are kept equal. Grazing incidence is thus a configuration of choice in order to nondestructively investigate nanostructures deposited on – or buried under – a substrate.

Pattern Analysis

The interpretation and simulation of the measured scattering patterns is challenging and requires accurate models and, if size/distance distributions are considered, is demanding tedious computing. However, a complete modeling is not always necessary, and the data evaluation can go from quite simple for a crude first approximation to tedious when the full pattern reproduction is sought [10]. A practical illustration of all important terms is given in Fig. 1. In an extremely crude and semiqualitative approach, and neglecting particle shape, correlation, and anisotropy, one can deduce from the position and shape (width) of the diffusion lobes directly an approximation of the interisland distance (D), height (h), and size (R_0) as illustrated in Fig. 1.

The scattering signal contains a coherent contribution and an incoherent scattering term as soon

as the size distribution of the nanostructures is not monodisperse. The coherent term is the product of the Fourier transform of the nanostructure shape with the interference function $S(Q_{//})$. For the incoherent scattering the two limit cases are (i) the Decoupling Approximation (DA), assuming no nanostructure correlations, and (ii) the Local Monodisperse Approximation (LMA), assuming full correlation between nanostructure sizes at a scale corresponding to the photon coherence length.

Hard X-ray photons interact weakly with matter, and except for perfect crystals or at glancing angles, the kinematical Born approximation (BA), which neglects multiple scattering effects, is valid. At angles below $3 \times \alpha_c$ dynamical effects of reflection and refraction at interfaces become important and have to be taken into account using the popular distorted wave Born approximation (DWBA) [11].

Within surface physics and nanoscience two approaches have to be distinguished depending on the level of interpretation that is sought:

(i) The more intuitive effective layer Born approximation (ELBA), which applies for isotropic buried islands with simple shapes. The index of refraction is that of the effective layer in which the islands are supposed to be buried. In ELBA the resulting scattering pattern is the weighted sum of the scattering arising from each subelement since an assembly of objects can be considered as the sum of the individual scatterings as far as the relative intensities remain linked to the total number of electrons in each object within the x-ray coherence length.

(ii) The DWBA, which applies in the vicinity of critical scattering geometries for islands deposited on a substrate. In the complete treatment, an anisotropic signal originating from island facets can also be considered. The effective form factor corresponds to the coherent interference of four waves corresponding to the four possible scattering events (weighted by the corresponding reflection coefficients) experienced by the incoming and exiting beams on a given island.

Within the ELBA, to first order and a rapid interpretation strategy assuming weak reflectivity, the scattered intensity I(Q) is proportional to the product of three terms: the form factor P(Q) which is the Fourier transform of the island volume, the interference function S(Q), and the transmission factor $T(\alpha_f)$ that is supposed to simulate the dependence of the scattered intensity as function of α_f.

The interference function S(Q) is related to the correlation length, which characterizes a distribution of islands on the surface. It is also related to the real space correlation function g(r) via the formula

$$S(Q) = 1 + 2\pi \cdot \rho_s \cdot \int_0^r (g(r) - 1) \cdot J_0 \cdot (Q_c r).r.dr \tag{5}$$

where ρ_s is the particle density per surface unit and J_0 the Bessel function of zero order. Many analytical different pair correlation functions can be used within models. One may consider, for example, a Gaussian pair correlation function (illustrated in inset of Fig. 1, top) given by

$$g(r) = \begin{cases} 0, & \text{for } 0 \le r \le R_0 \\ \dfrac{e^{-\frac{(r-D)^2}{\omega^2}} - e^{-\frac{(R_0-D)^2}{\omega^2}}}{e^{-\frac{(D_1-D)^2}{\omega^2}} - e^{-\frac{(R_0-D)^2}{\omega^2}}}, & \text{for } R_0 \le r \le D_1 \\ 1, & \text{for } D_1 \le r \le \infty \end{cases} \tag{6}$$

The most reliable input remains the direct evaluation of $g(r)$ from real space measurements made, for example, by complementary microscopy techniques.

Within the DWBA the effective form factor corresponds to the coherent interference of four waves corresponding to the four possible scattering events (weighted by the corresponding reflection coefficients) experienced by the incoming and exiting beams on a given island. The DWBA scattering intensity I_{DW} from islands deposited onto a substrate may be summarized as

$$I_{DW} = \left\{ \frac{|1 - n_\Delta^2|}{2 \cdot \mathrm{Re}(1 - n_\Delta)} \cdot \left[\widetilde{\chi}_\Delta\left(q_{//}, q_z\right) + R^f \widetilde{\chi}_\Delta\left(q_{//}, -p_z\right) + R^i \widetilde{\chi}_\Delta\left(q_{//}, p_z\right) + R^i R^f \widetilde{\chi}_\Delta\left(q_{//}, -q_z\right) \right] \right\}^2 \tag{7}$$

where n_Δ is the index of refraction of the islands, $\widetilde{\chi}_\Delta\left(q_{//}, q_\perp\right)$ is the form factor of the islands, $q_{//} = (q_x, q_y)$ is the component of the momentum transfer q parallel to the surface, q_z corresponds to the net momentum transfer perpendicular to the substrate surface, $p_z = k_z^f + k_z^i$ where k_z^f and k_z^i are the normal to the substrate plane wave vectors of the incoming and exiting waves. The reflectivity coefficients for the incoming and outgoing waves are denoted R^i and R^f, respectively. The Born approximation is restored with $R^i = R^f = 0$.

The specular reflectance of discontinued multilayer systems with interfacial roughness can be calculated by use of a recursive classical exact layer-by-layer method, based on Fresnel coefficients within the Parratt optical formalism and the dispersion model.

Interestingly, variations of electronic densities are the only requirement mandatory to enable the observation of GISAXS patterns. Objects (of lower electronic densities) in a larger electron density matrix will yield such signals. The extreme case will then be the presence of structured voids or pores. For reactive interfaces like NiO/Cu(111) [12], self-organized holes in the surface or the presence of Ni decorated corrals lead to GISAXS patterns similar to growing nanoparticles.

Size distributions can be included in the pattern evaluation in any framework. The obvious effect of the size distributions is to smooth the scattering pattern. Height distribution and cross-correlation between lateral size and height distributions are very difficult to extract from the GISAXS pattern.

Examples of Application

In 1989, GISAXS was introduced to study the dewetting of gold deposited on a glass surface and later, using synchrotron light metal agglomerates on surfaces and in buried interfaces were considered.

Determining the morphology of islands on a substrate, embedded clusters in matrices as well as the fabrication control of nanometer-sized objects can be successfully addressed using GISAXS. It has become a popular technique because it can be applied to a broad range of samples in various samples environments: investigation of samples in standard conditions [5, 10], porous layers, precipitates and quantum dots [13], growing particles in ultra-high vacuum conditions [4, 10], complex nanostructures like carbon nanotubes [14] and shape modifications upon various treatments. In recent years GISAXS studies have shifted towards the study of sample processing conditions and in-situ treatments [15–17]. The method in itself does not require any particular sample preparation – it is the thin film growth method requiring ultra-high vacuum environment. A major step has been realized in the last decade because of an in-depth understanding of the underlying phenomena and the availability of algorithms enabling the simulation of the experimental patterns. In the most advanced studies coherent GISAXS experiments are considered enabling full particle shape reconstruction [13, 18]. In general, GISAXS can be applied to characterize self-assembly and self-organization at the nanoscale in thin films.

The technique becomes highly interesting in situations where charge build-up is a problem like for investigating metal/insulating oxide systems. Electron based techniques are generally hampered by the insulating character of the sample and the charge build-up. Many such systems are characterized by Volmer-Weber or Stranski-Krastanov 3D growth of particles following a nucleation, growth and coalescence scheme. In GISAXS such a growth sequence leads to diffusion lobes moving towards the reflected beam in reciprocal space with respect to film thickness as illustrated in Fig. 2 for the Co/NiO(111) system. Even without pattern simulation the position and

shape changes of the diffusion lobes indicates directly the growth mode.

From sequences of images taken at various film thicknesses and sample conditions (like substrate temperatures or quality) one can extract the overall behavior of a metal/oxide interface and evaluate for example the effect of defects in the cluster and/or on the substrate. The study of Co deposits, performed in ultra-high vacuum conditions using substrates of different qualities, i.e., with variable amounts of nucleation centers is reported in Fig. 2. The experiment was performed on BM32 beamline (ESRF, Grenoble, France) using a beam energy of 10 keV. GISAXS was also combined with grazing incidence diffraction measurements (not shown here) to access the crystallinity of the sample. When the growth is performed at low temperature (T = 450 K) the Co clusters have numerous defects and poor crystalline quality. For Co thicknesses below 0.8 nm the islands diameters are similar whatever the surface quality or Co island crystalline quality. Thus in the nucleation regime, for cluster sizes below the onset of defects in the clusters, the number of nucleation centers on the surface does not lead to noticeable effects. Above 0.8 nm the diameters fairly diverge indicating that when coalescence starts to play an important role, the island mobility increases with temperature. This effect is consistent with the inter-island distance behavior of the islands. From the island height behavior it appears that the deposition at high temperature (good Co crystalline quality) on a substrate with numerous nucleation centers is equivalent to low temperature deposition (poor Co crystalline quality) on a high quality substrate. Such investigations can hardly be made using other techniques.

GISAXS is not only highly efficient to investigate in situ grown particles, it can also tackle very complex ex situ samples like carbon nanotubes (CNT) [14] organized in aligned long tubes perpendicular to the sample surface. Each tube includes on its top a Co nucleation particles used to promote the growth of the CNT. Figure 3 reports measurements made on beamline ID01 at the ESRF (Grenoble, France). The experimental pattern includes GISAXS scattering and reflectivity contribution. The calculated patterns required the use of a core-shell C-Co structure inspired from transmission electron microscopy additional measurements. The height/size distribution where fairly large and the DWBA adequate parameters where retrieved from a first ELBA approximation. Such an analysis enables the evaluation of the effects of preparation conditions changes like the ammonia content in the reactive gas mixture [14].

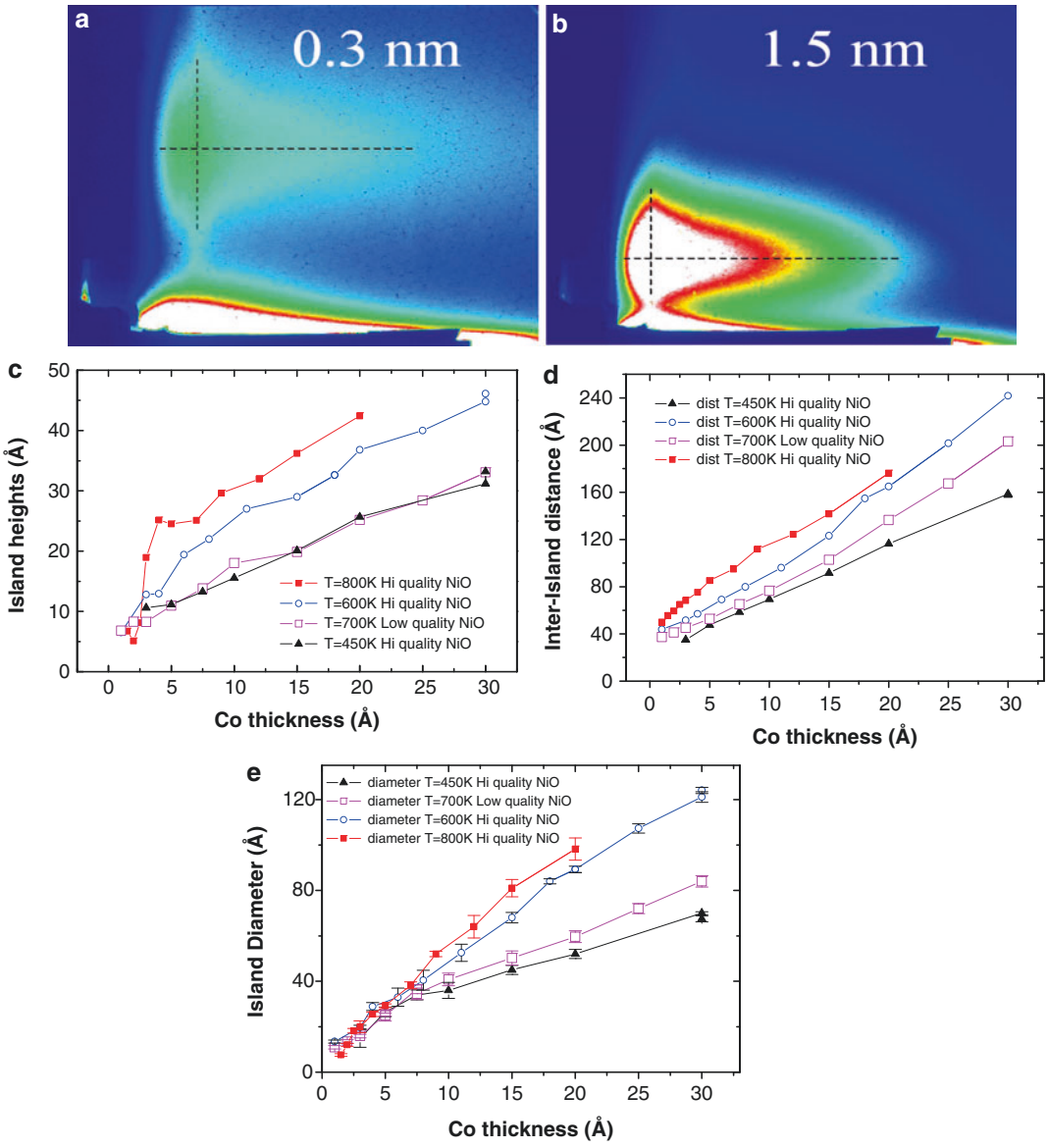
Recent GISAXS studies tackle a number of important issues concerning the driving forces of self-organization processes in block copolymers [19], nanoparticles [20] and organic electronics [17, 21]. The technique provides new and original contribution to these ongoing fields.

High Resolution Micro-Diffraction (μ-XRD)

In the local probe diffraction approach (μXRD), it is necessary to illuminate with the X-ray beam only a single object. The probe size yields the lateral resolution. As a consequence, the volume of nanostructures probed by the X-rays will be highly reduced (from several 10^4–10^6 objects in a 'standard' XRD experiment to a single one), thus the detected scattered signal is reduced. Indeed, this one is proportional to the incident X-ray photon flux and the probed volume. In order to compensate for the very small probed volume, the X-ray beam has to be focused to small sizes. Several focusing alternatives are nowadays available.

A Short Introduction to Hard X-ray Focusing Optics

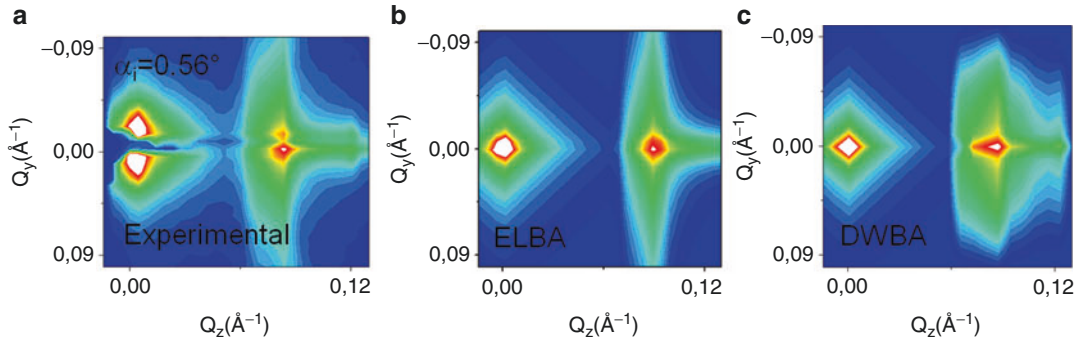
The following discussion is centered mostly around hard x-ray (several keV energy) optics used at synchrotron sources, although laboratory sources with X-ray beams of several 10 μm are available and can be used for certain XRD experiments with local (lateral) resolution. But a stable, highly brilliant and low divergent X-ray source is a "must have" for obtaining very small and intense X-ray spots. The synchrotron x-ray source is generally situated far from the experimental station (several 10 m) which allows for large demagnification factors for the focusing optics.



Selected Synchrotron Radiation Techniques, Fig. 2 (**a**, **b**) Part of the GISAXS signal recorded on a CCD camera for (**a**) 0.3 nm and (**b**) 1.5 nm nominal thickness of Co deposited on NiO(111). The beam energy was 10 keV; the vertical of the CCD is parallel to the sample surface and the horizontal direction is perpendicular to it. The lead well at the bottom of the pictures hides the direct and reflected beams. The dashed lines are guide for the eye to evidence the changes in shape and position of the GISAXS signal. (**c**, **d**, **e**) Morphology of Co islands grown on NiO(111): Island height (**c**), inter-island distance (**d**) and island diameter (**e**) with respect to the deposited Co thickness, as extracted from GISAXS measurements. Hi- and low- quality NiO substrates had mosaic spreads of 0.05° and 0.5° respectively

Consequently, major progress was done at synchrotron facilities, where these focusing devices can be used up to their limits: beam sizes of several 10 nm have been demonstrated and highly intense beams in the sub-μm size range are available at a number of synchrotrons [1].

A detailed description of the various focusing/collimation schemes for X-rays proposed in the

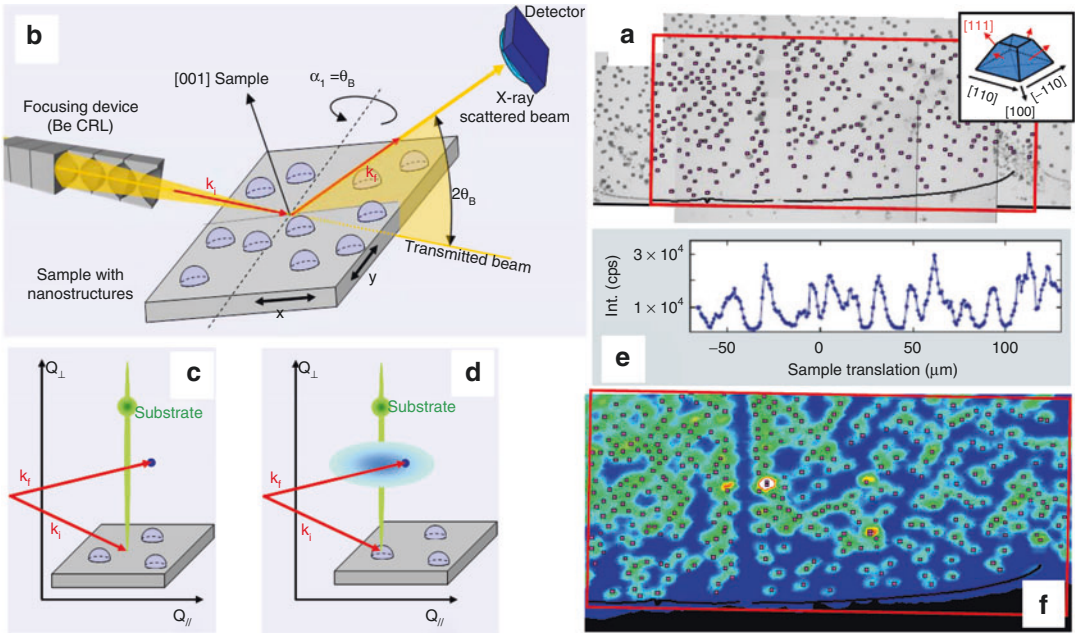


Selected Synchrotron Radiation Techniques, Fig. 3 (**a**) Experimental GISAXS pattern of CNTs prepared with 1 % ammonia concentration in the reactive gas mixture, collected at $a_i = 0.56°$. The sample surface is vertical and located on the *left side* of the pattern. A horizontal beam-stop stops the intense direct and specular beams. (**b**) ELBA pattern including Co cylinders, C tubes, Co-C intermixing (Co-C core-shell 14.8 nm average object and intermixed C-Co intermixed tubes); Co and C radii 2.5 and 3.3 nm respectively, inter-tube distance 55.6 nm, (**c**) DWBA pattern using the parameters retrieved from the ELBA model. The colour scale is logarithmic and identical for all patterns

last years can be found in Ref. [6]. Their respective advantages and drawbacks have to be carefully considered when designing a micro/nano X-ray beam experiment.

μ-XRD Setup and Experimental Approach

Using X-ray focusing optics, it is possible to build various X-ray microscopes working with different contrast mechanisms, or combinations of them: fluorescence, luminescence, transmission, diffraction. The particular case in which a diffraction contrast is used will be briefly described here. The setup is a combination of a X-ray focusing element and an accurate diffractometer. It consists of a scanner stage used for laterally positioning the sample in the X-ray beam, coupled to precise rotation stages both for the sample angles (incidence, azimuth, tilts) and the X-ray detector (at or close to Bragg diffraction angle) for probing, in the reciprocal space, any desired lattice parameter. Optionally a high resolution optical microscope pointing at the precise position of the diffractometer's centre (and X-ray beam) is used for sample pre-aligning and finding of the region of interest. The instrument can function in two ways: one in which the sample surface is *imaged point by point like in a scanning probe technique*, and a second one in which, for particular regions (microstructures) of the sample, detailed *high resolution μ-XRD* data are recorded. Both approaches will be illustrated hereafter. Figure 4 shows the principle of the method. The sample considered here consists of SiGe islands (square based truncated pyramids, inset Fig. 4, panel a) [22]. Panel (a) shows an optical image of the sample surface, image on which the individual SiGe islands can easily be identified (dark squares). The x-ray beam is focused onto a small spot on the sample. The elastically scattered x-ray photons are collected by the detector. By rotating the sample and detector angles, the intensity distribution around different reciprocal lattice point positions can be probed [4]. This is done by appropriately setting the Bragg angles (panel (b), beam impinging incidence angle given by the $\mathbf{k_i}$ vector and the position of the detector given by $\mathbf{k_f}$). If the sample is laterally translated in the X-ray beam without changing any of the angles, one expects to see differences in the scattered intensity depending if an island or the bare substrate is illuminated. In the case of the substrate, the intensity distribution shows a sharp Bragg peak (panel (c)), and a streak perpendicular to the sample surface, the so called crystal truncation rod (CTR) [4, 5]. If the focused x-ray beam hits an island, additional and broader features will appear in reciprocal space at different position (panel (d)), since a different lattice parameter, the one



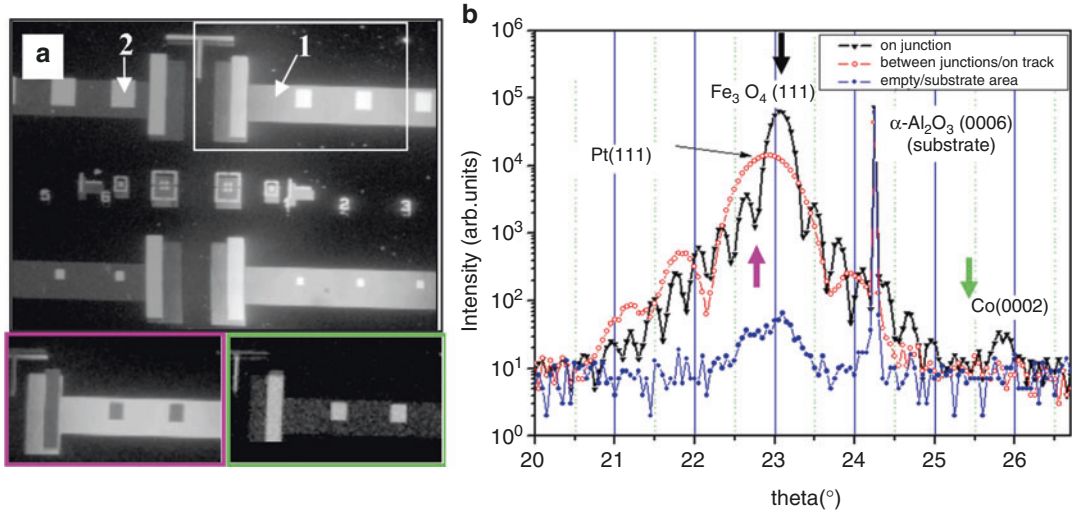
Selected Synchrotron Radiation Techniques, Fig. 4 Principle of the SXDM approach (see also Refs. [6, 23, 24]). (**a**) Optical image of the SiGe/Si pyramids sample. The inset shows a sketch of the shape of the SiGe square-based truncated pyramids. The red rectangle shows the area imaged using the μXRD approach (panel **f**). The optical image, obtained in back-reflected light, makes that the side facets are deflecting the light out of the optical axis of the objective, thus appearing dark. Each dark square represents a single SiGe island. (**b**) Illustration of the setup, ki and kf being the scattering vectors. (**c**, **d**) The x-ray spot (represented by the tip of the ki *arrow*) illuminates the substrate only or a single island (respectively). Only in the latter case a broad x-ray diffraction signal (different lattice spacing distribution inside the island) is observed, at a different position with respect to the substrate signal. Tuning the diffraction angle such to be sensitive in the reciprocal space to a position characteristic to the islands (*dot* pointed by the tip of the *arrow* kf) and scanning the sample laterally, higher intensity is observed only when the X-ray spot illuminates an island (panel **e**). If this is done in both x and y directions, a map of the surface of the sample (SXDM) is obtained (panel **f**), with enhanced intensity (*green* and *yellow* color) when an island is probed. Overlaid to this image, the position of the SiGe islands extracted from the optical image (shown in panel **a**) is superposed as red squares

of the SiGe island, is probed. The origin of this different lattice parameter can be multiple, some of the possible causes can be: a various composition, strain, shape, relaxation, etc.

Examples of Applications

Scanning X-ray Diffraction Microscopy (SXDM): Imaging with Diffraction Contrast

It is now straightforward to perform a "microscopy"-like experiment, the result being an image of the surface of the sample. The diffraction angles are fixed at the expected values to record diffraction from the crystalline planes inside the nanostructures and the position of the sample is scanned laterally while recording, at each point, the scattered intensity (Fig. 4 panel (e)). The resulting contrast is diffraction based and is related to the islands presence. If the mapping is performed in both directions (x and y), the resulting map is the result of the Scanning X-ray Diffraction Microscopy (SXDM) approach. Panel (f) shows a SXDM image (color scale) superimposed to the optical image of the same region of the sample (from panel (a)). It is possible to find precisely a particular object: in this case, the region close to the 'defect' exhibited as the vertical zone without any SiGe islands. Markers could also be envisaged. Indeed, it is of utmost



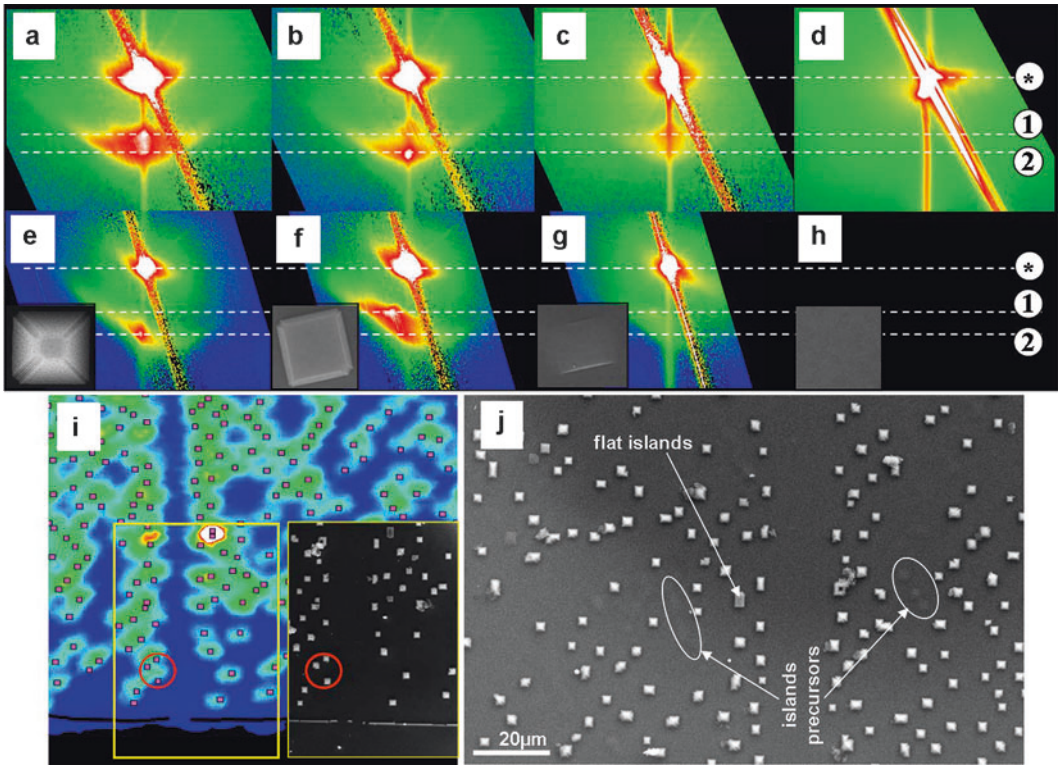
Selected Synchrotron Radiation Techniques, Fig. 5 (**a**) Scanning μXRD image (SXDM) of the MTJ surface showing defects due to mask align in the optical lithography process. Areas where only the Pt layer (1) remained after the lithography process or the full MJT (2) is present are shown. (**b**) μXRD data (θ-2θ geometry) at different locations on the sample: ▼on a MTJ; ○ on contact track (Pt buffer); ● on the empty area (substrate). The *colored arrows* point to reciprocal space positions characteristic of the different crystalline structures (Pt, Fe3O4 and Co), positions for which SXDM images of the sample surface (shown in panel **a**) were performed (see also the colored frame insets in panel *a*)

importance not only to be able to measure an isolated object out of an ensemble, but to precisely know which particular object was measured: the properties of these SiGe islands can change dramatically from object to object and can be correlated with their shape obtained by ▶ SEM.

The second example below is showing the power of the method and its sensitivity not only to (local) crystalline changes, but to thickness as well. The sample (metal oxide Magnetic Tunnel Junction, MTJ) consists of a stacking of several metallic and oxide layers on which structures of lateral size in the 10–100 μm range were made by optical lithography [25]. This particular sample suffered a mask misalign during one of the lithography steps. Although this could be easily detected by simple optical microscopy examination of the surface, μXRD can also bring some information about how this happened and which layers suffered; this sample is a good example to show how the different structures can be imaged and differentiated. The corresponding XRD recorded at these different locations are shown in Fig. 5 panel (b). Signal originating from the various crystalline structures of the layers (Pt, Fe_3O_4 and Co) can easily be differentiated and used as probe in imaging the surface of the sample, after tuning the diffraction angles to these positions (as highlighted by the colored arrows in panel b). The contrast on the surface images changes accordingly. Moreover, sensitivity to thickness is achieved: the arrow pointing to the maximum of the Pt signal also corresponds to a minimum contrast (interference) for the Fe_3O_4 layer. This explains the reversed contrast obtained in the left bottom panel (a) of Fig. 5. For more details, the reader can see Ref. [26] and references therein.

High Resolution X-ray Micro-Diffraction (HR – μXRD) and Combination with SEM

With the approach depicted above it is possible to measure not only an *individual* object but a very *particular* one [6, 23, 24]. The case of the SiGe islands will be used as example. On the SXDM image of the surface of the sample (Fig. 4), individual objects are chosen and measured in HR-μXRD. Once the lateral positions of the



Selected Synchrotron Radiation Techniques, Fig. 6 (**a**–**h**) Reciprocal space maps (logarithmic intensity scale) in the vicinity of the (004) and (115) Bragg peaks, recorded using a X-ray focused beam. The representative SEM image for each type of object is shown in the inset. The Si substrate Bragg peak is indicated by (*). (**i**) Image of the sample surface, performed with diffraction contrast (SXDM), compared to a SEM image of the same area, with particular objects identified (*red circle*). (*j*) SEM image (details) of the sample, around the region with the 'vertical path' free of SiGe islands, in the *middle* of the image

sample are fixed such that the X-ray spot illuminates the object of interest, the angles (sample and detector) are scanned in order to describe the reciprocal space in the vicinity of the chosen Bragg position. The recorded data show characteristics allowing to be grouped in three categories corresponding to three types of SiGe objects: the scattered intensity might vary slightly for the same type of object (depends on the particular imaged object), but, for one family the same characteristics are found. The differences are from one type to another. Figure 6a–h shows such measurements. The differences from the three mentioned types are visible both close to the (004) and (115) Bragg peaks. Since these μXRD measurements were performed at known positions of the sample (cf. image of the surface, as described in Fig. 4), it is possible, at the end of the diffraction experiment, to image in detail precisely the objects in question. The corresponding SEM images of the objects are reported as insets in Fig. 6a–h. Comparing the scattered intensities measured for SiGe pyramids and flat islands, some major differences can be highlighted: the intensity distribution close to the SiGe peak is changing. If the truncated pyramid exhibits scattered intensity concentrated around two positions (two maxima, labeled (1) and (2) in the figure), the flat islands show distributions mostly around the position (2). In the case of the (004) map, the position (1), closer to the Si Bragg location, shows the presence of a lattice parameter (out of the surface plane) closer to the Si one, i.e., indication of a lower Ge concentration. The signal at position (2) indicates a higher Ge concentration and a larger lattice parameter.

The availability of two types of maps, around symmetric (004) and asymmetric (115) Bragg peaks, bring complementary information: the peak's position on the (115) maps has also an in-the surface plane component of the SiGe lattice parameter [24].

This example shows the need and the power of combining several methods on the very same micro/nano-object, for detailed characterization and modeling: the shape and the sizes are deduced from SEM images and then injected into a model used to simulate the diffraction data (finite elements methods for strain distribution, then semi-kinematical scattering theory), more details can be found in Refs. [6, 24].

In-situ Combination of μXRD with Atomic Force Microscopy (AFM)

The previous example showed the combination of μXRD with (ex-situ) ▶ SEM images, yielding to an understanding of the structure of microscopic semiconductor samples. The next example shows the in-situ combination of a μXRD experiment with an Atomic Force Microscope (AFM) [27, 28]. This approach allows measuring in the same time and for the same individual small object the mechanical behavior, the internal strain and the response to external stress. Such measurements, especially when approaching the nanoscale, are of the highest importance for materials characterization, both in the elastic and the plastic deformation regimes.

The SiGe island sample described before was used as well as model sample to prove the principle of this experiment. The experimental setup is schematically shown in Fig. 7a: an ▶ AFM is mounted on the diffractometer and the focused X-ray beam is aligned with the apex of the tip. During point by point mapping of the sample surface with the X-ray spot (SXDM), the electron current induced in the AFM tip by the X-ray beam is also recorded. The resulting two contrast images (acquired simultaneously) are then combined (green and red colors respectively) to yield the image shown in panel (b). Panel (c) shows precisely the same sample area imaged in SEM (ex-situ) previous to the experiment – some of the missing islands in panel (b) were 'destroyed' by plastic deformation attempts during the first part of the AFM experiment.

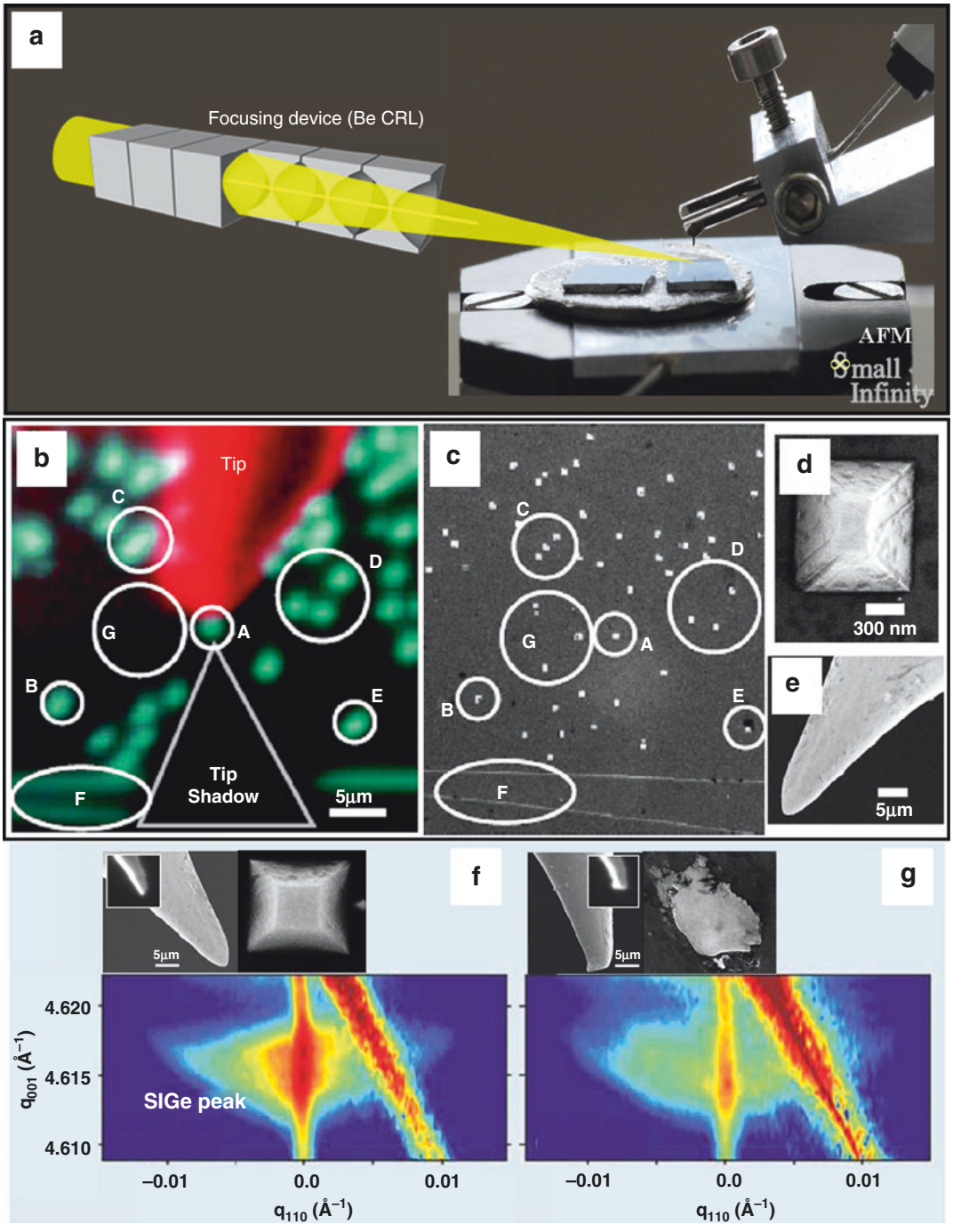
The applied force (and thus the pressure) can be extracted from the changes in the resonance frequency of the AFM tip and the lattice parameter (average value over the whole island) is extracted from the Bragg position in μXRD experiments, at each applied external pressure. It was also possible to access the plastic regime. Figure 7 (panels f and g) shows μXRD measurements (reciprocal space maps) close to the (004) Bragg position for a SiGe pyramid before and after pushing. The insets show the corresponding SEM images of the island and AFM tip before and after the experiment. Although, the AFM tip bends, the applied pressure can be large enough to completely destroy the SiGe island, which result in a major change of the μXRD signal.

This example validates the use of the coupled μXRD and AFM not only for studying the elastic properties of small structures, but also for in-situ indentation of materials (▶ Nanoindentation) (a stiffer tip, and with an adapted shape for indentation, should then be used). In such an indentation experiment, by applying laterally resolved μXRD, the resulting strain field around the indented area could be mapped and modeled.

X-Ray PhotoEmission Electron Microscopy

XPEEM Instrumentation and Operating Principle

XPEEM microscopy is a parallel imaging method that combines X-ray electron spectroscopy and electron microscopy. It is based on soft X-ray-in/Electrons-out principle and was originally pioneered by B. Tonner. The sample is illuminated by a monochromatic soft X-ray beam (<1500 eV). The method does not require extreme focusing of the X-rays. Moderate focusing (few μm^2) using dedicated X-ray optics (Kirkpatrick – Baez, Volter mirrors, etc.) is used so that the beam spot matches the useful microscope field of view. After the absorption of the X-ray photons, the emitted photoelectrons are collected, magnified and projected onto a detector



Selected Synchrotron Radiation Techniques, Fig. 7 (**a**) Cartoon of the combination of μXRD with AFM (simplified setup). Once the X-ray focused beam aligned with the tip of the AFM, it is used to 'image' (high resolution XRD) the object placed beneath the tip. b) superposed maps (images) of the sample, with the X-ray beam aligned on the AFM tip and a single SiGe island placed just beneath the tip: in green the detector signal (sensitive to SiGe) is recorded, while in red is the electron current induced in the tip when this one is hit by the X-ray beam. (**c**) SEM image of the sample showing precisely the region measured in panel (**b**). The different areas on the sample are labeled. Note the absence of the islands labeled 'G' – during the

using a dedicated (electrostatic or magnetic) low electron energy microscopy column.

There is a certain number of XPEEM microscopes available commercially. A version that combines PEEM and LEEM microscopy is nowadays the most popular one used with synchrotron radiations sources. This apparatus is very powerful because it allows to combines in a single instrument structural (LEEM) and spectroscopic (PEEM) methods, enabling a true multi technique approach to study surfaces, buried interfaces and nanostructures.

The electron column used in the XPEEM microscopy is similar to those used in Electron microscopy, the main difference arising from the use of low kinetic energy photoelectrons. The XPEEM instrument in its simpler version consists of an objective lens, a projective lenses, an image detector and a contrast aperture (see Fig. 8). A first image, collected and magnified by the objective lens, is formed in the back-focal plane of the objective lens. An aperture placed within this plane (contrast aperture) allows to reduce the energy spread of the photoelectrons and limits their angular dispersion. The image is further magnified by a set of projective lenses. The final image is collected via a detector consisting of a Micro-Channel plate, a florescence screen and a CCD camera. In the most sophisticated instruments, a certain number of lenses are additionally implemented: corrector lenses for astigmatism, retarding grid as high-pass energy filter, field lenses, etc.

Although the incorporation of an energy filter in XPEEM microscope is not mandatory for imaging, its installation in the most recent and advanced microscopes opens up the way for core level and valance band photoemission microscopy. The energy filter is not only helpful for imaging with primary photoelectrons but also useful for secondary electron imaging. It allows to select a narrow energy window around the maximum of the secondary electron energy distribution and, thus, to improve the spatial resolution without unacceptable loss of transmission.

The main requirements for X-ray microscopy, in particular for PEEM/LEEM, are the spatial resolution and the contrast mechanism.

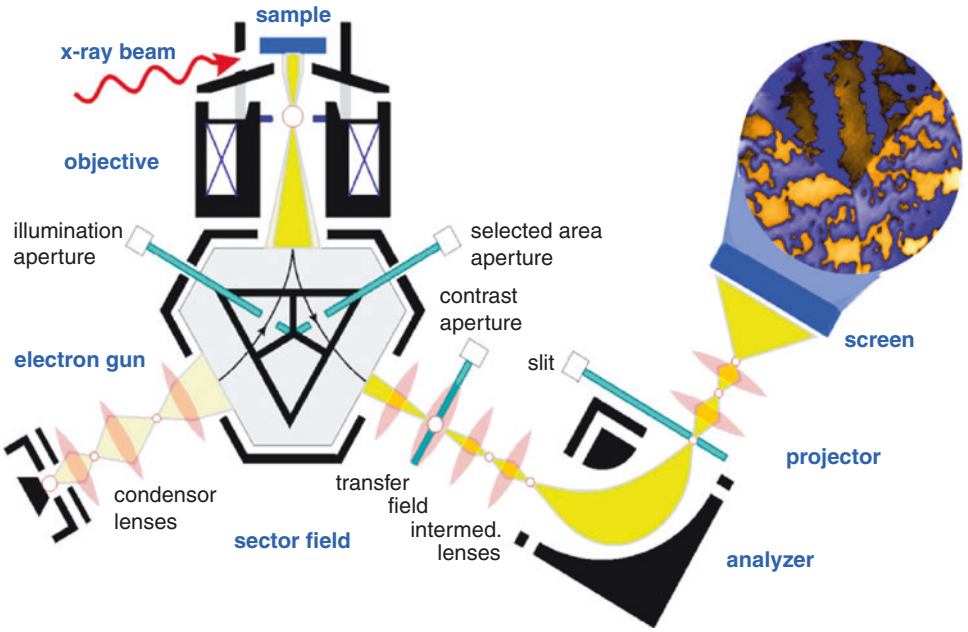
Spatial Resolution

A spatial resolution of 22 nm using synchrotron radiation has already been achieved and a resolution of 6 nm using LEEM. A key limitation to the performance of electron-optical X-ray microscopes is the severe chromatic and spherical aberration of the immersion lens used. This problem constrains the X-PEEM microscope to a fairly small transmission in an attempt to offset these problems. Rempfer's group has demonstrate a solution to this problem, by designing and testing an electron mirror, which has spherical and chromatic aberrations of similar magnitude to the objective lens, but opposite in sign. When used in combination with a properly designed objective lens and a highly symmetric beam separator, the aberration can be highly reduced. Therefore both transmission and resolution can be increased. Several projects using correcting mirror are under development [29].

Contrast Mechanism

The contrast mechanism used in XPEEM microscopy arises from the spectroscopic capabilities and the characteristic binding energies of the

Shark Skin Effect, Fig. 7 (continued) experiment and previously taking this image, these islands were 'smashed' (destroyed) or laterally displaced using the AFM tip. (**e**) SEM image of a SiGe single pyramid of ~1 μm lateral size. (**e**) Image of the blunt tip used for pushing experiments. Note the relatively large apex radius of curvature (with respect to standard AFM tips) – this feature makes easier pushing on the small objects. (**f**) Reciprocal space map (close to the (004) Bragg peak) for a single SiGe island. The Si substrate Bragg position is not shown on the map ($q_{si(004)} = 0.4628\ nm^{-1}$). (**g**) Same map after pushing with the AFM and reaching the plastic regime (destroying the SiGe island). The intensity distribution changed completely. The insets show the SiGe island and the AFM tip (SEM images) before and after plastic deformation. The electron yield images of the tip are also shown for each case



Selected Synchrotron Radiation Techniques, Fig. 8 Schematic sketch of the LEEM-PEEM microscope

atomic core electrons. The illumination of the specimen with X-rays excites a broad electron spectrum consisting of primary photoelectrons and inelastically scattered secondary photoelectrons. Two operation modes can be therefore performed:

- XPS (X-ray Photoelectron Spectroscopy). The XPEEM detects and selects, using an appropriate energy filter, the primary photoemitted electrons from atomic core levels. The photon energy is fixed and the kinetic electron energy is given by: $E_K = h\nu - E_b - \phi$, where $h\nu$ is the photon energy, E_b is the core level binding energy and ϕ the work function i.e., the energy barrier that the photoelectrons have to overcome to escape from the sample surface.
- XAS (X-ray Absorption Spectroscopy). In this case the XPEEM detects the secondary photoelectron as function of the impinging photon energy. When the photon energy matches the absorption threshold of the element, the photoelectron spectrum shows strong resonances due to electron transition arising from the core level to the unfilled valence band states.

The choice of XAS or XPS modes is mainly made on the basis of sample depth probe requirements (1/e) which are directly correlated to the electron inelastic mean free path (e) and the photoelectron kinetic energy. XPS allows to achieve a high surface sensitivity with a probing depth below 1 nm, meanwhile the XAS is more bulk sensitive (<10 nm).

In both cases, the XPS or XAS intensities are proportional to the number of emitter atoms within the probing depth, and thus provides direct and quantitative mapping of the chemical composition. Additional information can be also obtained from the lineshape analysis as they are a fingerprint of the emitter chemical state (valence state, site location, etc.). Finally, local spectroscopy can be performed selecting a small region of the specimen: local spectroscopy (μ-XAS or μ-XPS), photoelectron diffraction (μ-XPD), angle resolved photoelectron spectroscopy (μ-ARPES), etc.

The high spatial resolution of XPEEM microscopy, it chemical sensitivity and spectroscopic capability open up a broad field of application which extend from thin film studies,

nanostructures, magnetism to more exotic applications for surface science such as in tribology and geology.

Examples of Application: Magnetic Imaging

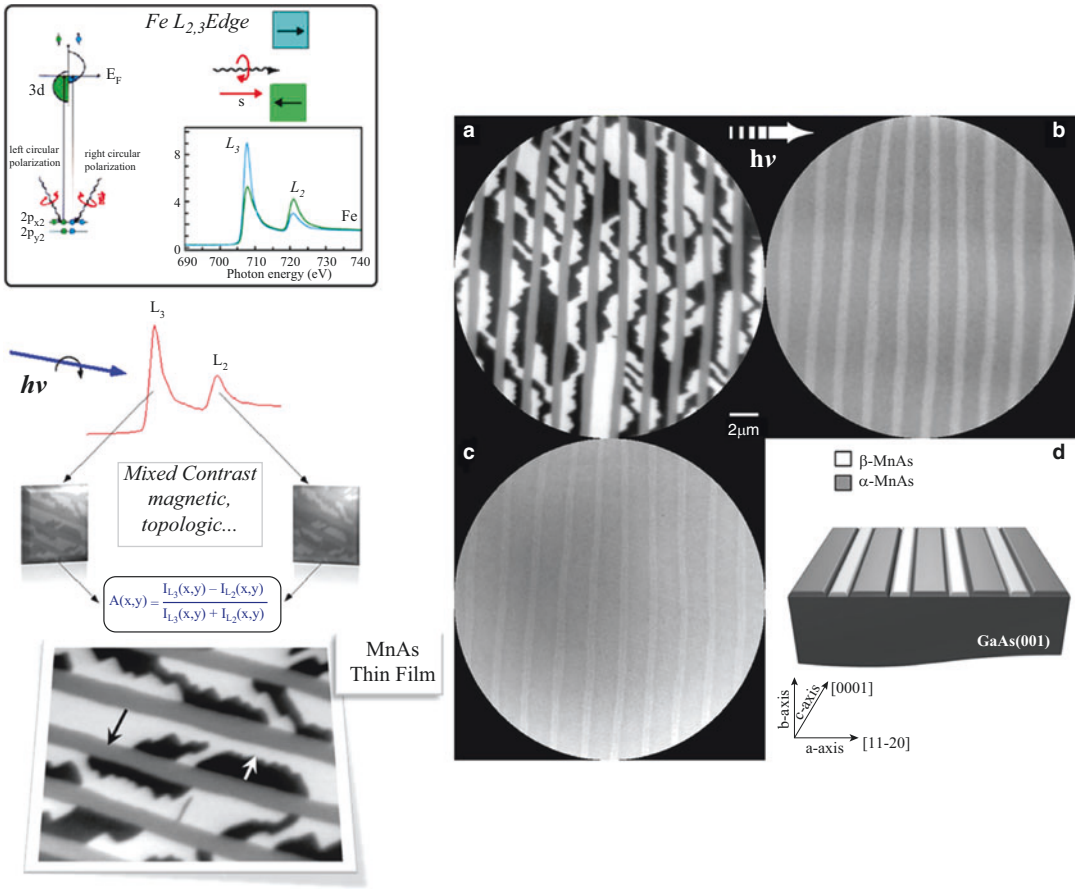
The interest in magnetic domain imaging in the nanometer range has been rapidly increasing during the last decade. A considerable impetus is coming from the development of high-density magnetic storage devices and from the forthcoming achievement of spin electronics. In order to tailor the magnetic behaviour of these systems to specific needs (for instance a certain response to magnetization reversal), a detailed understanding of the structure and of the dynamics of magnetic domains is mandatory. In addition, the thin film nature of such devices emphasizes the surface aspect of magnetism. This situation requires magnetic domain-imaging techniques which combine surface sensitivity and high spatial resolution. Moreover, for many applications element specificity is even more important than high lateral resolution. Magnetic storage media or building elements of spin-electronic devices are often composed of several chemical elements or intermetallic compounds, each of which distinctly contributes to the magnetic behavior. All these requirements pose a considerable challenge to conventional magnetic domain imaging techniques such as magneto-optical Kerr microscopy, Lorentz microscopy, scanning electron microscopy (SEMPA), Magnetic Force Microscopy (MFM), etc.

XPEEM magnetic microscopy is today a good candidate for an ideal surface magnetic imaging technique, as it combines the magnetic and element selectivity with a spatial resolution below the size of the magnetic domains. One may identify three important length scales for magnetic imaging which span over two orders of magnitude. The first one is about 1 μm, set by the size of lithographically manufactured magnetic cells such as in spin valve read heads or magnetic memory cells. The second one is about 10 nm, corresponding to the crystallographic grain size of typical magnetic materials and to domain walls. The last one is 0.1 nm, i.e., the atomic size. The actual spatial resolution of the XPEEM microscope allow to image straightforwardly the magnetic domain, the further improvements of the resolution may allow access to the second characteristic length scale of 10 nm.

The elemental specificity in XPEEM magnetic microscopy arises from the characteristic binding energies of the atomic core electron, as mentioned above. Both X-ray photoelectron spectroscopy (XPS) and X-ray absorption (XAS) can be used for magnetic imaging, although the latter is the most used because it is less demanding in terms of instrumentation and photon flux.

The use of polarized synchrotron radiation enables studies of the electronic and magnetic anisotropies, and thus allows magnetic contrast for the XPEEM [30]. A simple description of the photon polarization by a biaxial vector for linear polarization and a vector for left/right handed circular polarization is the physical basis for probing various anisotropies of the sample. In general, linearly polarized light can only detect anisotropy of electronic charge. In contrast, helicity resolved circularly polarized light can measure a dipolar or vector quantity, in the present case the modulus and direction of the electron angular moment and spin.

For magnetic XPEEM spectromicroscopy, X-ray Magnetic Circular Dichroism (XMCD) is exploited. This effect is widely used to determine the size, the direction and the anisotropy of the atomic magnetic moments in magnetic material at a macroscopic scale. In the simple figure of a 3d transition metal, the magnetisation or more precisely the spin moment, is given by the imbalance between the spin-up and spin-down electrons (or holes) within the d shell and bellow the Fermi level. The use of circularly polarised light will make the absorption process spin-dependant and therefore enable access to the spin moment. In other word, the spin-split valence band will act as a detector for the spin of the excited photoelectron: the XAS intensity is simply proportional to the number of empty d states of a given spin. The measurements of the XAS signal anisotropy with respect to the direction of the light for a given



Selected Synchrotron Radiation Techniques, Fig. 9 (*Left* panel) *Top*: Principles of X-ray magnetic circular dichroism (*XMCD*) in the case of L Fe edge. The XMCD spectrum reflects the anisotropy of L absorption edge with respect to the relative direction between the incident light and the magnetization direction at a given circular polarized light. Bottom: Principles of XMCD-PEEM in the case of a MnAs thin film. The measurements are performed at the Mn L edge (see text for explanation). (*Right*) XMCD and XMLD-XPEEM images. MnAs sample thickness 330 nm: (**a**) XMCD image at the L_3 Mn edge highlighting the FM Mn domains in the α-phase. (**b**) XMLD image taken using the L_3 multiplets asymmetry at vertical linear polarization evidencing the AF domains. The gray level corresponds to the isotropic α-phase and the light region to the orthorhombic β-phase. (**c**) LEEM image taken at the same sample area showing the α-β phase coexistence. (**d**) Schematic sketch of the FM/AF configuration in the MnAs thin films

S

circular polarisation enable directly the access to the projection of the magnetisation as illustrated in the Fig. 9. In XMCD-PEEM, a set of two images are acquired at respectively the L_3 and L_2 edges for a given light helicity and direction. These two images will contain a mixture of different contrast: magnetic, structural, chemical.... In order to cancel all the contrast mechanism except the magnetic one, an asymmetric image is calculated by subtracting the above two mentioned data. This image reflects directly the local variation of the $L_{2,3}$ asymmetries and therefore gives access to the projection of the magnetisation moment with respect to the light direction: the black and white regions in the image reflect the domains where the magnetic axis is aligned parallel or anti parallel to the direction of the light at fixed polarization. The grey area corresponds to domains where the magnetic axis is perpendicular to the direction of the light or that have no magnetic moment. It is

worth noting that the same XMCD-PEEM image can be obtained by varying the circular light helicity and keeping the photon energy fixed at the same absorption edge (L_3 or L_2).

The study of antiferromagnetic (AF) surfaces and interfaces has posed an even larger challenge because conventional techniques are mainly bulk sensitive and antiferromagnet do not carry any net external magnetic dipole moment. These limitations were overcome by the use of X-ray Magnetic Linear Dichroism (XMLD) spectroscopy. In contrast to XMCD which directly measures the magnetic moment, XMLD measures the projected value of the square of the magnetic moment. XMLD can therefore be applied for all uniaxial magnetic system, i.e., antiferromagnets as well. Recently it was shown that XMLD spectroscopy in conjunction with X-PEEM microscopy is capable of imaging the detailed antiferromagnetic domain structure of a surface and interface [31]. This has been shown on a NiO(001) thin film and cleaved sample. The XMLD-PEEM images reveal antiferromagnetic contrast corresponding of the different in-plane projection of the antiferromagnetic axis.

The ability to combine in the same instrument XMCD and XMLD XPEEM microscopy with high spatial resolution and chemical selectivity has open a new route to investigate magnetic materials, as illustrated in the following examples.

Surface Magnetism in MnAs Thin Films

Ferromagnetic (FM) MnAs is a promising candidate for electrical spin injection into GaAs and Si based semiconductors, since it exhibits a large carrier spin polarization, small coercive field and relatively high saturation magnetization and Curie temperature. Bulk MnAs is FM at Room Temperature (RT) (α phase) and shows close to 40 °C a first order transition to the paramagnetic β phase. On the contrary, epitaxial MnAs films on GaAs substrate, which are more appropriate for the injection applications, show at RT the coexistence of both phases. The phase coexistence results in the formation of self-organized stripes of alternating α and β phases. The XMCD-PEEM image of Fig. 9 (right) illustrates this phase coexistence. The black/white regions correspond to ferromagnetic domains in the α-phase, the vertical grey stripes correspond to the non-magnetic α-phase. The easy magnetisation direction is perpendicular to the stripes direction, pointing in opposite directions in the black and white regions. The period of the stripes is closely correlated to the film thickness (4.8 × thickness). The phase coexistence is due to anisotropic strain applied by the GaAs substrate. As a consequence the magnetization is pointing not along the stripe direction as expected from shape anisotropy considerations but is perpendicular to it, predominantly in-plane. This leads to a complex, thickness-dependent magnetic domain structure in the interior of the α-MnAs which is reflected in the complexity of magnetic images of the surface of the film.

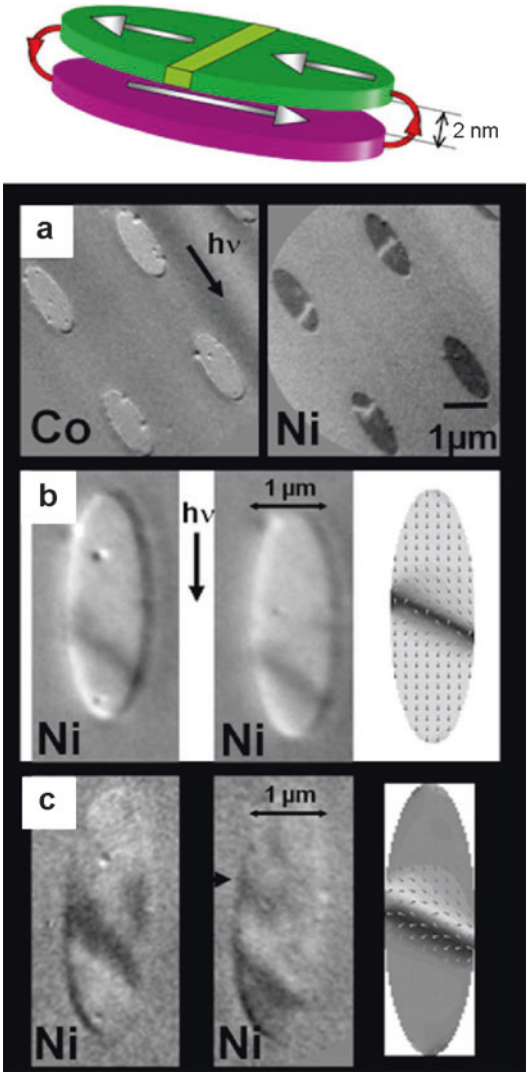
The nature of the non-magnetic β-phase is still controversial. In the coexistence range exchange bias and giant magnetoresistance effects of possible practical importance have been reported. Until recently β-MnAs has generally been considered to be paramagnetic but the effects just mentioned and first principles calculations suggest antiferromagnetism. A more thorough first principles calculation comes to the conclusion that β-MnAs is paramagnetic. The XMLD-PEEM microscopy enables to elucidate this controversy. Figure 9 evidences a weak antiferromagnetism in the β-phase which explain the reported exchange bias and giant magnetoresistance [32].

Micromagnetism of Patterned Nanostructures and Magnetic Tunnel Junction

Magnetic tunnel junctions (MTJ) are probably one of the most studied devices in the so called spintronics research field. They are usually composed by two active ferromagnetic electrodes separated by a thin insulating layer (<3 nm) that act as a tunnel barrier for the spin polarised electrons i.e., a spin dependant tunnel current exists and can be induced between the two ferromagnetic electrodes. The most attractive property of MTJs concerns their strong electrical resistance variation with the magnetic configuration of the electrodes, the so called magnetoresistance. The resistance can vary from more than 100 %. The Magnetic Random Access Memories (MRAM) and the read

heads of hard drive disks take benefits from this property. In order to increase storage capacity, a strong effort is made toward a size reduction of the MTJs. However, in the course of miniaturizing a unit cell for magnetic device a certain number of questions related to the finite size effect arise. Size effect starts then to play an important role and can affect drastically the electrodes magnetic properties. The results obtained on thin films cannot be straightforwardly extrapolated to patterned objects and nanostructures. Combining the high spatial resolution of XPEEM magnetic imaging, it chemical sensitivity and micromagnetic simulations, allowed to demonstrate the strong influence of a dipolar magnetic coupling on the magnetization reversal of MTJs [33]. Such studies are mandatory because understanding the domain wall formation process is crucial since it affects the magnetic – electrical properties of MTJs.

When the MTJs are structured in micrometer sized elements, an antiferromagnetic coupling tends to align the magnetization of each electrode in an antiparallel configuration contrary to what is usually observed in the thin film case. This coupling originates from the magnetic stray field at the nanostructures edges (see red arrows in Fig. 10). The MTJs is formed by two ferromagnetic layers: (permalloy and cobalt) separated by a thin aluminium oxide layer: Co(4 nm)/Al_2O_3(2 nm)/$Fe_{20}Ni_{80}$(4 nm). The MTJs have been patterned in ellipse-shaped structures in order to re-enforce the stray field effect. The high spatial resolution of the X-ray Photoemission Electron Microscopy (X-PEEM) combined to X-ray Magnetic Circular Dichroism (XMCD) enable to directly image the magnetic configuration of each electrode of the MTJs. This powerful technique allows to image independently the magnetic configuration in each electrodes if composed of different elements, thanks to it elemental selectivity. Figure 10a shows that, when no magnetic field is applied, both layer magnetizations are mainly in an antiparallel configuration. Moreover, if the Co layer magnetization is uniform, surprisingly this is not always the case in the NiFe layer. The (b) and (c) parts of Fig. 10 present images obtained in geometries where the technique is sensitive to the magnetization component either



Selected Synchrotron Radiation Techniques, Fig. 10 *Top*: Schematic representation of a MTJs. The two magnetic electrodes, composed of a nickel-iron alloy and of cobalt, are respectively represented in *green* and *purple*. They are separated by 2 nm thick layer of oxidized aluminium (not represented for clarity reason). The magnetic stray field sketched in *red* tends to align the magnetization each electrode in an antiparallel configuration. *Bottom*: XMCD-PEEM images of four ellipses recorded at the Co and at the Ni edges. The photons incidence direction is aligned along the ellipses long axis from the *top left corner*. The *white* and *black* contrasts correspond to magnetization components aligned along the long ellipses axis. XMCD-PEEM images of two ellipses (1 × 3 μm) measured at the Ni edge. The *grey* level distribution corresponds to the scalar projection of the local magnetization with respect to the light incidence: (**b**) parallel and (**c**) perpendicular to the ellipses long axis

along the ellipses long axis (b) or along the ellipses short axis (c). The combination of this imaging technique with *in silico* modelling (right part of the Fig. 2b, c) allows to understand the nature and the formation process of the non-uniform magnetization distribution presents in the NiFe layer. In this region, the magnetization rotates continuously by 360°, forming an object separating two regions of uniform magnetization. This magnetic object is called a 360° domain wall. By simulations, only one chirality is obtained for the wall, although three different cases can be observed experimentally: no wall and two chiralities (Fig. 10). This difference has been attributed to local magnetic anisotropy fluctuations at the ellipses extremities which drive the magnetization curling direction during the reversal process.

Future Directions for Research (Outlook)

Future challenges of GISAXS may lie in the use of the x-ray coherence [13, 18] in order to reconstruct single objects without any model assumptions. As far as the corresponding nano-objects can bear the X-ray beam, soft matter systems like polymers, block copolymers or even biological applications, such as proteins, peptides and viruses attached to surfaces or in lipid layers, have become a growing field unfolding the full potential of the GISAXS technique.

For μXRD experiments aiming at a characterization of individual micro and nano-structures [6, 23, 24], the actual demand is that of availability of brighter and smaller x-ray spots, with increased stability. Last generation synchrotron sources can fulfill in a certain measure these demands, and dedicated experimental stations are built [1]. Coupling μXRD with other techniques is achieved [6] and allows obtaining complementary results from various techniques, which completes the understanding about these systems and their properties. Another experimental path using focused X-ray beams is the one using a white (continuous energy spectrum) x-ray microbeam – the Laue microdiffraction experiments. It can give access to grain distributions (orientation, strain) in polycrystalline samples [34]. This particular topic needs achromatic focusing elements (e.g., Kirkpatrick-Baez mirrors, capillaries) and was not detailed here. Alternative techniques like the lensless x-ray microscopy (Coherent Diffraction Imaging CDI [35], X-ray holography [36], ptychography [37],. . .) using coherent x-ray beams are also a path opened for accessing non-destructively the properties of small structures: 3D imaging and strain determination with spatial resolutions of 10 nm and below were proved. Tremendous gain in photon flux and coherence can be achieved by the use of free electron laser sources. These approaches will not be detailed here; for the reader looking for more details, one can refer to the reviews [6, 38]. The non-invasive character of the XRD method and the relatively large penetration depth of the hard X-rays make possible experiments in which the diffraction approach is combined (in- or ex-situ) with other analysis methods. The particular cases of SEM and AFM have been examined. Ultra High Vacuum (UHV) environment with in-situ preparation facilities of thin metallic films was also addressed, but other environments (high temperature, clean environment, high pressure, reaction cells, electric or magnetic fields, . . .) can easily be imagined. Experiments performed on model electronic devices while they were functioning [39] were also realized.

One of the main challenges for XPEEM microscope in the future is to combine the high spatial resolution with time-resolved experiments. In the case of magnetic materials for example, the dynamics of the magnetization reversal in thin magnetic films has become a matter of high interest for the future of magnetic recording and non-volatile magnetic memories. Parallel to the evolution towards smaller magnetic bits and memory cells, writing and reading times approaching the ns range will be required in a few years from now. A complete understanding of the magnetization dynamics in these structures requires the ability to probe the magnetization of the individual layers as well as their mutual interaction. Time-resolved X-PEEM measurements are very challenging, since the secondary electrons that are used for the image are strongly perturbed by the magnetic field necessary to switch the magnetization direction. Therefore,

the time-resolved XPEEM experiment can be only performed in stroboscopic mode. The magnetic pulses can be synchronized with the X-ray pulses coming from the storage ring to perform dynamic measurements in a pump-probe scheme. X-PEEM images cannot be acquired during the field pulses, but time resolution of 50 ps has been already demonstrated [40]. The advent of fourth generation synchrotron and more specifically X-ray Free Electron Laser [38] will certainly open up the way for high resolution imaging at a time scale of femtosecond.

Finally, many efforts are done to improve the spatial resolution. Although the ultimate resolution will be always limited by the diffraction limit (1 nm), the development of new aberration corrected setups will certainly open up the way to image nanostructure with resolution bellow 5 nm. Other aberrations correction methods using numerical algorithms or phase retrieval have been also proposed.

Cross-References

- ▶ AFM
- ▶ Nanoindentation
- ▶ SEM

References

1. See for example http://www.lightsources.org/ for links to existing facilities
2. Wiedemann, H.: Synchrotron Radiation. Springer, Berlin (2002). ISBN 978–3540433927
3. Reimers, W., Pyzalla, A.R., Schreyer, A., Clements, H.: Neutrons and Synchrotron Radiation in Engineering Materials Science: From Fundamentals to Material and Component Characterization. Wiley-WCH/GmbH & Co. KGaA, Waintheim (2008). ISBN 978-3-527-31533-8
4. Als-Nielsen, J., McMorrow, D.: Elements of Modern X-Ray Physics, 2nd edn. Wiley, New York (2010)
5. Barbier, A., Mocuta, C., Renaud, G.: Characterization and spectroscopy of thin films, Chapter 11. In: Nalwa, H.S. (ed.) Handbook of Thin Film Materials, vol. 2. Academic, San Diego (2002). ISBN 0-12-512910-6
6. Stangl, J., Mocuta, C., Chamard, V., Carbone, D.: Nanobeam X-Ray Scattering: Probing Matter at the Nanoscale. Wiley-VCH, Weinheim (2013). ISBN 978-3-527-41077-4
7. Locatelli, A., Bauer, E.: Phys. Condens. Matter **20,** 093002 (2008); Kuch, W.: Imaging Magnetic Microspectroscopy. Spinger, Berlin (2003)
8. Bauer, E.: Low energy electron microscopy. Rep. Prog. Phys. **57**, 895–938 (1994). doi:10.1088/0034-4885/57/9/002
9. Riekel, C.: New avenues in x-ray microbeam experiments. Rep. Prog. Phys. **63**, 233 (2000); Davies, R.J., Burghammer, M., Riekel, C.: A combined microRaman and microdiffraction set-up at the European Synchrotron Radiation Facility ID13 beamline. J. Synchr. Rad. **16**, 22 (2009)
10. Renaud, G., Lazzari, R., Leroy, F.: Probing surface and interface morphology with grazing incidence small angle X-Ray scattering. Sur. Sci. Rep. **64**, 255–380 (2009)
11. Rauscher, M., et al.: Grazing incidence small angle x-ray scattering from free-standing nanostructures". J. Appl. Phys. **86**, 6763 (1999). doi:10.1063/1.371724
12. Barbier, A., Stanescu, S., Boeglin, C., Deville, J.-P.: Local morphology and correlation lengths of reactive NiO/Cu(111) interfaces. Phys. Rev. B **68**, 245418-1-7 (2003)
13. Zozulya, A.V., Yefanov, O.M., Vartanyants, I.A., Mundboth, K., Mocuta, C., Metzger, T.H., Stangl, J., Bauer, G., Boeck, T., Schmidbauer, M.: Imaging of nanoislands in coherent grazing-incidence small-angle x-ray scattering experiments. Phys. Rev. B **78**, 121304 (2008)
14. Mane Mane, J., Cojocaru, C.S., Barbier, A., Deville, J. P., Jean, B., Metzger, T.H., Thiodjio Sendja, B., Le Normand, F.: GISAXS study of carbon nanotubes grown on SiO2/Si(100) by CVD. Phys. Status Solidi (RRL) **1**, 122–124 (2007)
15. Gilles, R., Rémi, L., Christine, R., Antoine, B., Marion, N., Olivier, U., Frédéric, L., Jacques, J., Yves, B., Henry, C.R., Jean-Paul, D., Fabrice, S., Jeannot, M.-M., Olivier, F.: Real-time monitoring of growing nanoparticles. Science **300**, 1416 (2003)
16. Papadakis, C.M., Di, Z., Posselt, D., Smilgies, D.-M.: Structural instabilities in Lamellar Diblock Copolymer thin films during solvent vapor uptake. Langmuir **24**, 13815–13818 (2008)
17. Smilgies, D.-M., Li, R., Giri, G., Chou, K.W., Diao, Y., Bao, Z., Amassian, A.: Look fast: crystallization of conjugated molecules during solution shearing probed in-situ and in real time by X-ray scattering. Phys. Status Solidi (RRL) **7**, 177–179 (2013)
18. Sun, T., Jiang, Z., Strzalka, J., Ocola, L., Wang, J.: Three-dimensional coherent X-ray surface scattering imaging near total external reflection. Nat. Photonics **6**, 586–590 (2012)
19. Zhang, J., Posselt, D., Sepe, A., Shen, X., Perlich, J., Smilgies, D.-M., Papadakis, C.M.: Structural evolution of perpendicular Lamellae in Diblock Copolymer thin films during solvent vapor treatment investigated by grazing-incidence small-angle X-Ray scattering. Macromol. Rapid Commun. **34**, 1289–1295 (2013)

20. Senesi, A.J., Eichelsdoerfer, D.J., Macfarlane, R.J., Jones, M.R., Auyeung, E., Lee, B., Mirkin, C.A.: Stepwise evolution of DNA-programmable nanoparticle superlattices. Angew. Chem. Int. Ed. **52**, 6624–6628 (2013)
21. Perlich, J., Schwartzkopf, M., Körstgens, V., Erb, D., Risch, J.F.H., Müller-Buschbaum, P., Röhlsberger, R., Roth, S.V., Gehrke, R.: Pattern formation of colloidal suspensions by dip-coating: an in situ grazing incidence X-ray scattering study. Phys. Status Solidi (RRL) **6**, 253–255 (2012)
22. Schmidbauer, M.: X-Ray Diffuse Scattering from Self Organized Mesoscopic Semiconductor Structures. Springer Tracts in Modern Physics, vol. 199. Springer, Berlin (2004)
23. Dubslaff, M., Hanke, M., Schoder, S., et al.: X-ray nanodiffraction at individual SiGe/Si(001) dot molecules and its numerical description based on kinematical scattering theory. Appl. Phys. Lett. **96**(13), 133107 (2010)
24. Mocuta, C., Stangl, J., Mundboth, K., et al.: Beyond the ensemble average: x-ray microdiffraction analysis of single SiGe island. Phys. Rev. B **77**, 245425 (2008)
25. Mocuta, C., Barbier, A., Ramos, A.V., et al.: Effect of optical lithography patterning on the crystalline structure of tunnel junctions. Appl. Phys. Lett. **91**(24), 241917 (2007)
26. Mocuta, C., Barbier, A., Stanescu, S., et al.: X-ray diffraction imaging of metal-oxide epitaxial tunnel junctions made by optical lithography: use of focused and unfocused X-ray beams. J. Synchrotron Radiat. **20**, 355–365 (2013)
27. Ren, Z., Mastropietro, F., Davydok, A., et al.: Scanning force microscope for in situ nanofocused X-ray diffraction studies. J. Synchrotron Radiat. **21**, 1128–1133 (2014)
28. Cornelius, T.W., Mastropietro, F., Thomas, O., Schulli, T.U.: In situ nanofocussed X-ray diffraction combined with scanning probe microscopy, Chapter 10. In: Shih, K, (ed) X-Ray Diffraction: Structure, Principle and Applications, New York, USA, pp. 233–259. Nova Science (2013)
29. Fink, R., et al.: J. Electron Spectrosc.Relat. Phenom. **84**, 1249 (1997); Feng, J., et al.: J. Phys. Condens. Matter **17,** S1339 (2005)
30. Stöhr, J., et al.: Element-specific Magnetic Microscopy with Circularly Polarized X-rays. Science **259**, 658 (1993)
31. Ohldag, H., et al.: Spin Reorientation at the Antiferromagnetic NiO(001) Surface in Response to an Adjacent Ferromagnet. Phys. Rev. Lett. **86**, 2878 (2001)
32. Bauer, E., Belkhou, R., Cherifi, S., Locatelli, A., Pavlovska, A., Rougemaille, N.: Magnetostructure of MnAs on GaAs revisited. J. Vac. Sci. Technol. B **25**, 1470 (2007)
33. Hehn, M., et al.: 360° domain wall generation in the soft layer of magnetic tunnel junctions. Appl. Phys. Lett. **92**, 072501 (2008)
34. Budai, J.D., Yang, W., Tamura, N., et al.: X-ray microdiffraction study of growth modes and crystallographic tilts in oxide films on metal substrates. Nat. Mater. **2**, 487 (2003)
35. Beutier, G., Verdier, M., Parry, G., et al.: Strain inhomogeneity in copper islands probed by coherent X-ray diffraction. Thin Solid Films **530**, 120–124 (2013)
36. Chamard, V., Stangl, J., Carbone, D., et al.: Three-dimensional X-ray Fourier transform holography: the Bragg case. Phys. Rev. Lett. **104**, 165501 (2010)
37. Godard, P., Carbone, D., Allain, M., et al.: Three dimensional X-ray Bragg ptychography: high resolution imaging of extended crystalline nanostructures. Nat. Commun. **2**, 568 (2011)
38. Vartanyants, I.A., Robinson, I.K., McNulty, I., et al.: Coherent scattering and lensless imaging at the European XFEL Facility. J. Synchrotron Radiat. **14**, 453 (2007)
39. Hrauda, N., Zhang, J., Wintersberger, E., et al.: X-ray nanodiffraction on a single SiGe quantum dot inside a functional field-effect transistor. Nano Lett. **11**(7), 2875–2880 (2011)
40. Choe, S.-B., et al.: Vortex Core–Driven Magnetization Dynamics. Science **304**, 420 (2004)

Self-Assembled Monolayers

▶ Nanostructures for Surface Functionalization and Surface Properties

Self-Assembled Monolayers for Nanotribology

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

Synonyms

Molecularly thick layers; Monolayer lubrication; Organic films; Self-organized layers

Definition

Organized and dense molecular-scale layers of long-chain, organic molecules are referred to as

self-assembled monolayers (SAMs). SAMs are molecularly thick, well-organized, and chemically bonded to the substrate. Ordered molecular assemblies with various properties can be engineered using chemical grafting of various polymer molecules with suitable functional head groups, spacer chains, and surface terminal groups.

Overview

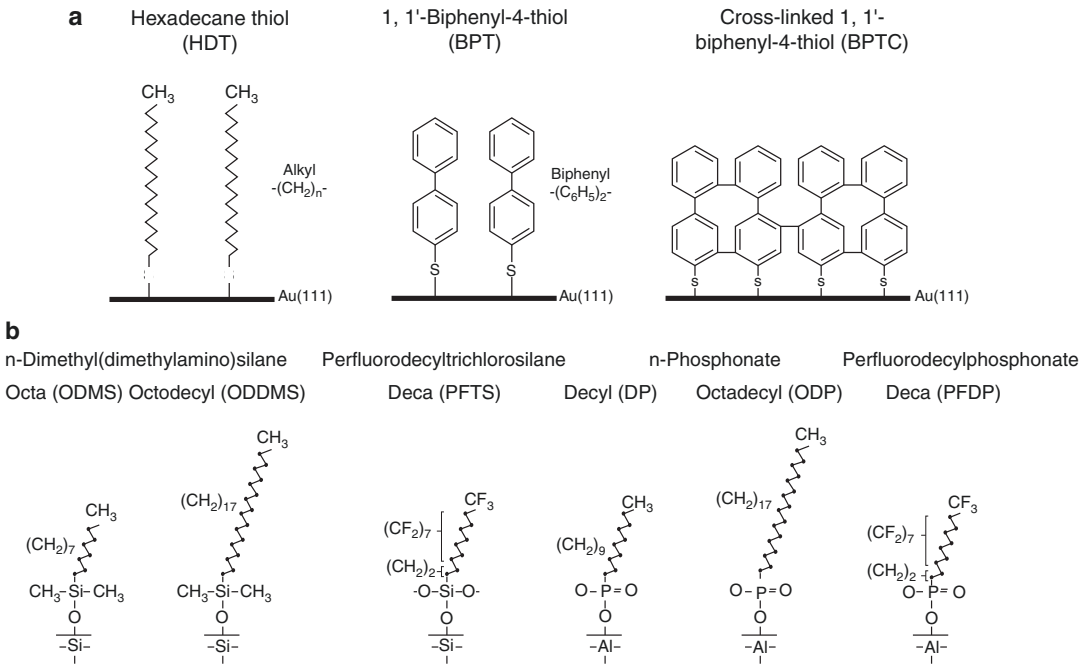
Reliability of various micro- and nanodevices, also commonly referred to as micro/nanoelectromechanical systems (MEMS/NEMS) and BioMEMS/BioNEMS, requiring relative motion, as well as magnetic storage devices (which include magnetic rigid disk and tape drives) requires the use of hydrophobic and lubricating films to minimize adhesion, stiction, friction, and wear [2–4, 7, 9–12, 15, 36, 39]. In various applications, surfaces need to be protected from exposure to the operating environment. For example, in various biomedical applications, such as biosensors and implantable biomedical devices, undesirable protein adsorption, biofouling, and biocompatibility are some of the major issues [9–11]. In micro- and nanofluidic based sensors, the fluid drag in micro- and nanochannels can be reduced by using hydrophobic coatings [42]. Selected hydrophobic films are needed for these applications.

For lubrication, an effective approach involves the deposition of organized and dense molecular layers of long-chain molecules. Two common methods to produce monolayers and thin films are the Langmuir-Blodgett (L-B) deposition and self-assembled monolayers (SAMs) by chemical grafting of molecules. LB films are physically bonded to the substrate by weak van der Waals attraction, while SAMs are chemically bonded via covalent bonds to the substrate. Because of the choice of chain length and terminal linking group that SAMs offer, they hold great promise for boundary lubrication of MEMS/NEMS. A number of studies have been conducted to study tribological properties of various SAMs deposited on Si, Al, and Cu substrates [14, 16–18, 21, 24–31, 33–35, 38, 40].

Bhushan and Liu [14], Liu et al. [35], and Liu and Bhushan [8] studied adhesion, friction and wear properties of alkylthiol and biphenylthiol SAMs on Au(111) films. They explained the friction mechanisms using a molecular spring model in which local stiffness and intermolecular forces govern the friction properties. They studied the influence of relative humidity, temperature, and velocity on adhesion and friction. They also investigated the wear mechanisms of SAMs by a continuous microscratch AFM technique.

Fluorinated carbon (fluorocarbon) molecules are known to have low surface energy and are commonly used for lubrication [5, 7, 8, 12]. Bhushan and Cichomski [13] deposited fluorosilane SAMs on polydimethylsiloxane (PDMS). To make a hydrophobic PDMS surface chemically active, PDMS surface was oxygenated using an oxygen plasma, which introduces silanol groups (SiOH). They reported that SAM coated PDMS was more hydrophobic with lower adhesion, friction, and wear. Bhushan et al. [17, 19], Kasai et al. [30], Lee et al. [31], Tambe and Bhushan [38], and Tao and Bhushan [41] studied the adhesion, friction, and wear of methyl- and/or perfluoro- terminated alkylsilanes on silicon. They reported that perfluoroalkylsilane SAMs exhibited lower surface energy, higher contact angle, lower adhesive force, and lower wear as compared to that of alkylsilanes. Kasai et al. [30] also reported the influence of relative humidity, temperature, and velocity on adhesion and friction. Tao and Bhushan [40] studied degradation mechanisms of alkylsilanes and perfluoroalkylsilane SAMs on Si. They reported that oxygen in the air causes thermal oxidation of SAMs.

Tambe and Bhushan [38], Bhushan et al. [18], Hoque et al. [24, 25] and DeRose et al. [21] studied the nanotribological properties of methyl- and perfluoro-terminated alkylphosphonate, perfluorodecyldimethylchlorosilane, and perfluorodecanoic acid on aluminum, of industrial interest. Hoque et al. [26] and DeRose et al. [21] studied the nanotribological properties of alkylsilanes and perfluroalkylsilanes on aluminum. Hoque et al. [27–29] studied the nanotribological properties of alkylphosphonate and perfluoroalkylsilane SAMs on copper. The



Self-Assembled Monolayers for Nanotribology, Fig. 1 Schematics of the structures of (**a**) hexadecane and biphenyl thiol SAMs on Au(111) substrates, and (**b**) perfluoroalkylsilane and alkylsilane SAMs on Si with native oxide substrates, and perfluoroalkylphosphonate and alkylphosphonate SAMs on Al with native oxide

authors found that these SAMs on aluminum and copper perform well irrespective of the substrate used. They confirmed the presence of respective films using X-ray photoelectron spectroscopy (XPS).

Hoque et al. [27–29] studied the chemical stability of various SAMs deposited on Cu substrates via exposure to various corrosive conditions. DeRose et al. [21] studied the chemical stability of various SAMs deposited on Al substrates via exposure to corrosive conditions (aqueous nitric acid solutions of a low pH of 1.8 at temperatures ranging from 60 °C to 80 °C for times ranging from 30 to 70 min). The exposed samples were characterized by XPS and contact angle measurements. They reported that perfluorodecanoic acid/Al is less stable than perfluorodecylphosphonate/Al and octadecylphosphonate/Al, but more stable than perfluorodecyldimethylchlorosilane/Al, which has implications in digital micromirror devices (DMD) applications. In general, chemical stability data of various SAMs deposited on Cu and Al surfaces to corrosive environments has been reported by these authors. Based on these studies, it was concluded that chemisorption occurs at the interface and is responsible for strong interfacial bonds.

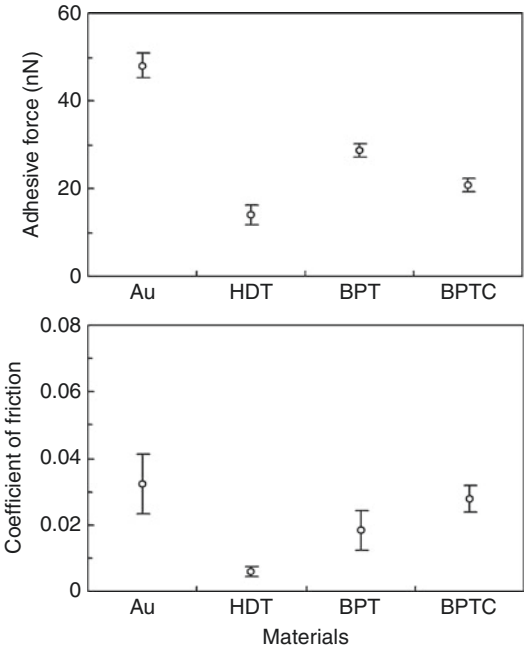
To date, the contact angle and nanotribological properties of alkanethiol, biphenylthiol, alkylsilane, perfluoroalkylsilane, alkylphosphonate and perfluoroalkylphosphane SAMs have been widely studied. In the following, the nanotribological properties of various SAMs are reviewed having alkyl and biphenyl spacer chains with different surface terminal groups (–CH_3, –CF_3) and head groups (–S–H, –Si–O–, –OH, and P–O–) which have been investigated by AFM at various operating conditions [14, 17, 18, 30, 33–35, 38, 40]. Hexadecane thiol (HDT), 1, 1′-biphenyl-4-thiol (BPT), and crosslinked BPT (BPTC) were deposited on Au(111) films on Si(111) substrates by immersing the substrate in a solution containing the precursor (ligand) that is reactive to the substrate surface (Fig. 1a). Crosslinked BPTC was produced by irradiation of BPT monolayers with low energy electrons. Perfluoroalkylsilane and alkylsilane

SAMs were deposited on Si(100) by exposing the substrate to the vapor of the reactive chemical precursors (Fig. 1b). Perfluoroalkylphosphonate and alkylphosphonate SAMs were deposited on sputtered Al film on Si substrate as well as bulk Al substrates (Fig. 1b).

Hexadecane Thiol and Biphenyl Thiol SAMs on Au(111)

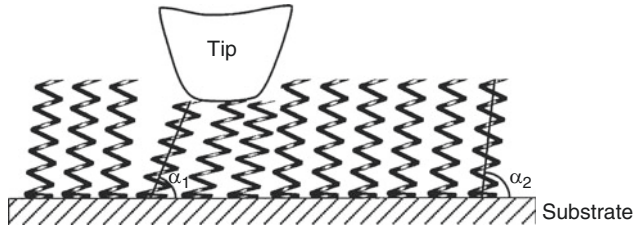
Bhushan and Liu [14] studied the effect of film compliance on adhesion and friction. They used hexadecane thiol (HDT), 1,1,biphenyl-4-thiol (BPT), and crosslinked BPT (BPTC) solvent deposited on Au(111) substrate, Fig. 1a. The average values and standard deviation of the adhesive force and coefficient of friction are presented in Fig. 2. Based on the data, the adhesive force and coefficient of friction of SAMs are less than the corresponding substrates. Among various films, HDT exhibits the lowest values. Based on stiffness measurements of various SAMs, HDT was most compliant, followed by BPT and BPTC. Based on friction and stiffness measurements, SAMs with high-compliance long carbon chains exhibit low friction; chain compliance is desirable for low friction. Friction mechanism of SAMs is explained by a so-called "molecular spring" model (Fig. 3). According to this model, the chemically adsorbed self-assembled molecules on a substrate are just like assembled molecular springs anchored to the substrate. An asperity sliding on the surface of SAMs is like a tip sliding on the top of "molecular springs or brush." The molecular spring assembly has compliant features and can experience orientation and compression under load. The orientation of the molecular springs or brush under normal load reduces the shearing force at the interface, which in turn reduces the friction force. The orientation is determined by the spring constant of a single molecule as well as the interaction between the neighboring molecules, which can be reflected by packing density or packing energy. It should be noted that the orientation can lead to conformational defects along the molecular chains, which lead to energy dissipation.

An elegant way to demonstrate the influence of molecular stiffness on friction is to investigate



Self-Assembled Monolayers for Nanotribology, Fig. 2 Adhesive forces and coefficients of friction of Au (111) and various SAMs

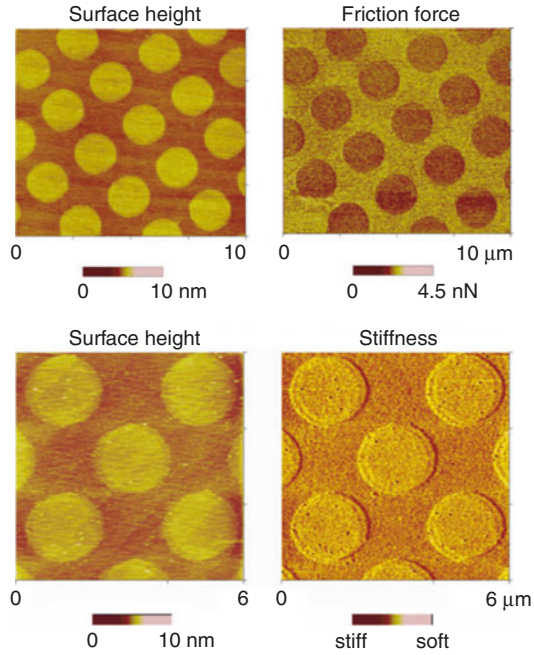
SAMs with different structures on the same wafer. For this purpose, a micropatterned SAM was prepared. First the biphenyldimethylchlorosilane (BDCS) was deposited on silicon by a typical self-assembly method [33]. Then the film was partially crosslinked using mask technique by low energy electron irradiation. Finally the micropatterned BDCS films were realized, which had the as-deposited and crosslinked coating regions on the same wafer. The local stiffness properties of this micropatterned sample were investigated by force modulation AFM technique [22]. The variation in the deflection amplitude provides a measure of the relative local stiffness of the surface. Surface height, stiffness, and friction images of the micropatterned biphenyldimethylchlorosilane (BDCS) specimen are obtained and presented in Fig. 4 [33]. The circular areas correspond to the as-deposited film, and the remaining area to the crosslinked film. Figure 4a indicates that crosslinking caused by the low energy electron irradiation leads to about 0.5 nm decrease of the surface height of BDCS films. The corresponding stiffness images indicate that the



Self-Assembled Monolayers for Nanotribology, Fig. 3 Molecular spring model of SAMs. In this figure, $\alpha_1 < \alpha_2$, which is caused by the orientation under the normal load applied by AFM tip. The orientation of the molecular springs reduces the shearing force at the interface, which in turn reduces the friction force. The molecular spring constant, as well as the inter-molecular forces can determine the magnitude of the coefficients of friction of SAMs. In this figure, the size of the tip and molecular springs do not in the exactly scale [14]

crosslinked area has higher stiffness than the as-deposited area. Figure 4b indicates that the as-deposited area (higher surface height area) has a lower friction force. Obviously, these data of the micropatterned sample prove that the local stiffness of SAMs has an influence to their friction performance. Higher stiffness leads to larger friction force. These results provide a strong proof of the suggested molecular spring model.

The SAMs with high-compliance long carbon chains also exhibit the best wear resistance [14, 33]. In wear experiments, the wear depth as a function of normal load curves show a critical normal load, at which film wears rapidly. A representative curve is shown in Fig. 5. Below the critical normal load, SAMs undergo orientation; at the critical load SAMs wear away from the substrate due to relatively weak interface bond strengths, while above the critical normal load severe wear takes place on the substrate.



Self-Assembled Monolayers for Nanotribology, Fig. 4 (**a**) AFM Grayscale surface height and stiffness images, and (**b**) AFM grayscale surface height and friction force images of micropatterned BDCS [33]

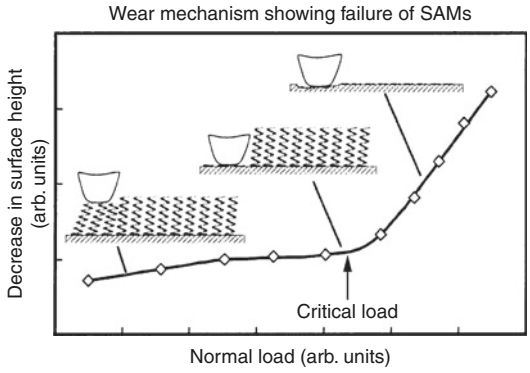
Perfluoroalkylsilane and Alkylsilane SAMs on Si(100), and Perfluoroalkylphosphonate and Alkylphosphonate SAMS on Al

Perfluorodecyltricholorosilane (PFTS), CF_3–$(CF_2)_7$–$(CH_2)_2$–$SiCl_3$, n-octyldimethyl (dimethyl amino)silane (ODMS), CH_3–$(CH_2)_n$–Si $(CH_3)_2$–$N(CH_3)_2$ (n = 7), and n-octadecyldimethyl(dimethylamino)silane (n = 17) (ODDMS) vapor deposited on Si(100) substrate and perfluorodecylphosphonate (PFDP)

$$CF_3-(CF_2)_7-(CH_2)_2-\underset{\underset{O}{\|}}{\overset{\overset{O}{|}}{P}}-OH,\ \text{decylphosphonate}$$

$$(DP),\ CH_3-(CH_2)_n-\underset{\underset{O}{\|}}{\overset{\overset{O}{|}}{P}}-OH \quad (n=9)\ \text{and}$$

octadecylphosphonate (ODP) (n = 17) by liquid deposition on sputtered Al film on Si substrate were selected (Fig. 1b). Perfluoro SAMs were selected because fluorinated films are known to have low surface energy. Two chain lengths of



Self-Assembled Monolayers for Nanotribology, Fig. 5 Illustration of the wear mechanisms of SAMs with increasing normal load [33]

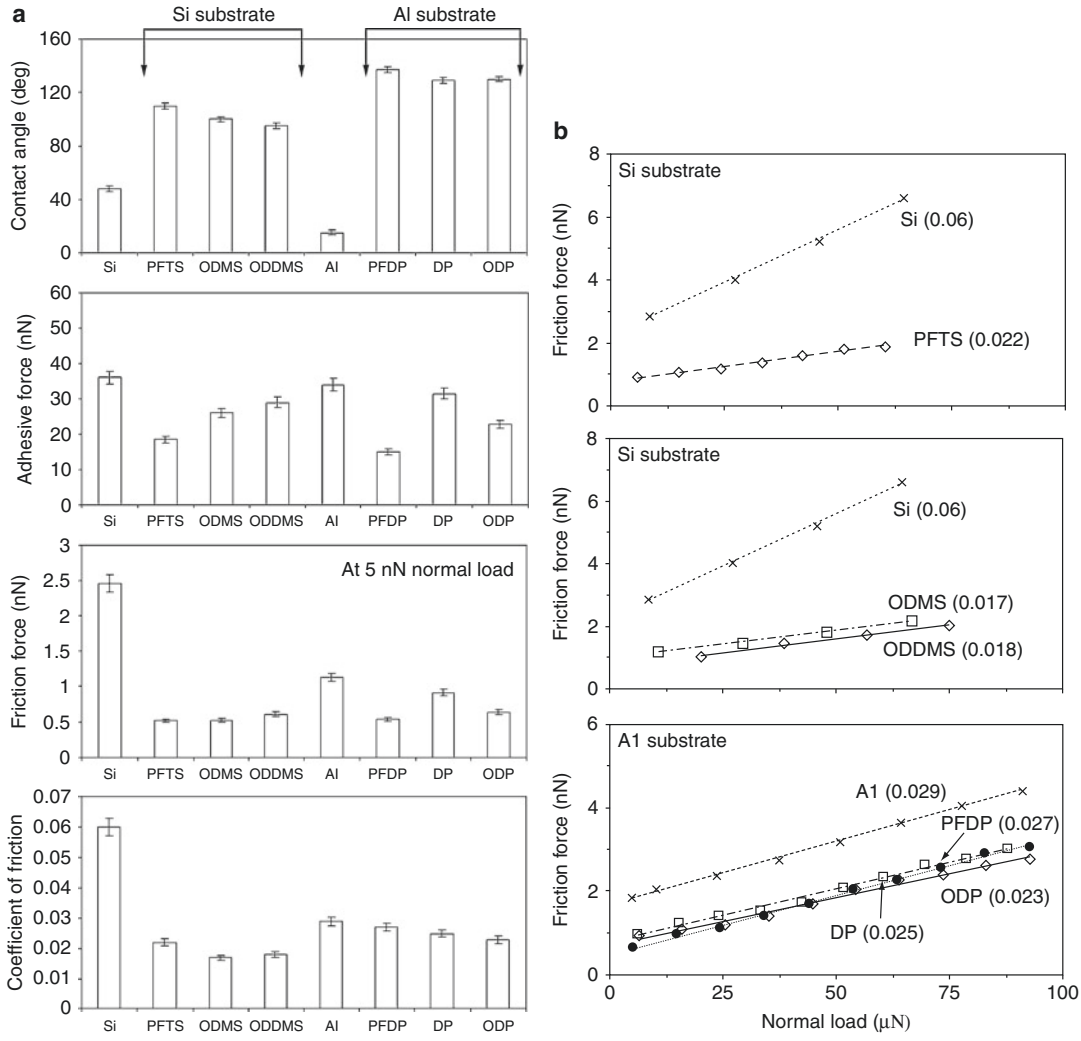
alkylsilanes (with 8 and 18 carbon atoms) were selected to compare their nanotribological performance with that of the former as well as to study the effect of chain length. Al substrate was selected because of the application of Al micromirrors in digital projection displays. Perfluoroalkylphosphane (with ten carbon atoms) and alkylphosphonate SAMs (with 10 and 18 carbon atoms) on Al were selected.

The measured values are compared among the samples in Fig. 6a [17, 18]. Significant improvement in the water repellent property was observed for perfluorinated SAMs as compared to bare Si and Al substrates. Static contact angles of alkylsilanes and alkylphosphonates were also higher than corresponding substrates, but lower than corresponding perfluorinated films. The contact angle generally increases with a decrease in surface energy [23], which is consistent with the data obtained. The contact angles can be influenced by the packing density as well as the sample roughness [37]. The higher contact angles for the SAMs deposited on Al substrates than those on Si substrate are probably due to this effect. The $-CH_3$ groups in ODMS, ODDMS, DP, and ODP are non-polar and are known to contribute to the water repellent property. Perfluorinated SAMs exhibited the highest contact angle among the SAMs tested in this study.

Figure 6a shows the adhesive force, friction force, and the coefficient of friction measured under ambient conditions using an AFM, and Fig. 6b shows the friction force vs. normal load plots for various SAMs deposited onto the Si and Al substrates [17, 18]. The bare substrates showed higher adhesive force than the SAMs coatings. ODMS and ODDMS shows an adhesive force comparable to DP and ODP despite their lower water contact angles. These SAMs have the same tail groups, and during AFM measurements the AFM tip interacts only with the tail groups, whereas the contact angles can also be influenced by the head groups in these SAMs. This is probably the reason as to why the adhesive forces for these SAMs are comparable. PFTS and PFDP, which have the highest contact angles, showed the lowest adhesion.

The effect of relative humidity for various SAMs on adhesion and friction was studied. Adhesive force, friction force at 5 nN of normal load, coefficient of friction, and microwear data are presented in Fig. 7 [18, 30]. The result of adhesive force for silicon showed an increase with relative humidity, Fig. 7. This is expected since the surface of silicon is hydrophilic, as shown in Fig. 6a. More condensation of water at the tip-sample interface at higher humidity increases the adhesive force due to capillary effect. On the other hand, the adhesive force for the SAMs showed a very weak dependency on the change in humidity. This occurs since the surface of the SAMs is hydrophobic. The adhesive force of ODMS/Si and ODDMS/Si showed a slight increase from 75 % to 90 % RH. Such an increase was absent for PFTS/Si, possibly because of the hydrophobicity of PFTS/Si.

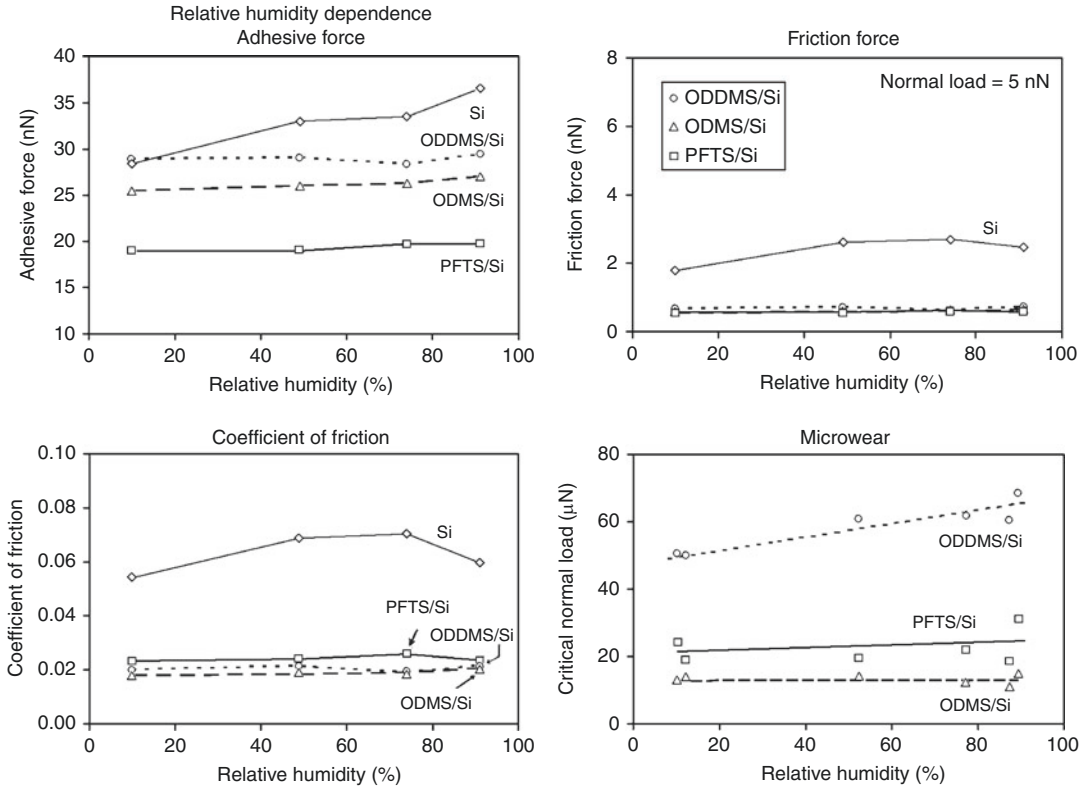
The friction force of silicon showed an increase with relative humidity up to about 75 % RH and a slight decrease beyond this point, see Fig. 7. The initial increase can result from the increase in adhesive force. The decrease in friction force at higher humidity could be attributed to the lubricating effect of the water layer. This effect is more pronounced in the coefficient of friction. Since the adhesive force increased and coefficient of friction decreased in this range, those effects cancel each other out and the resulting friction force showed slight changes. On the other hand, the friction force and coefficient of friction of SAMs showed very small changes with relative humidity



Self-Assembled Monolayers for Nanotribology, Fig. 6 (**a**) The static contact angle, adhesive force, friction force and coefficient of friction measured using an AFM for various SAMs on Si and Al substrates, and (**b**) friction force vs. normal load plots for various SAMs on Si and Al substrates [17, 18]

like that found for adhesive force. This suggests that the adsorbed water layer on the surface maintained a similar thickness throughout the relative humidity range tested. The differences among the SAM types were small within the measurement error, however a closer look at the coefficient of friction for ODMS/Si showed a slight increase from 75 % to 90 % RH as compared to PFTS/Si, possibly due to the same reason for the adhesive force increment. The inherent hydrophobicity of SAMs means that they did not show much relative humidity dependence.

Figure 8 shows the effect of temperature on adhesive force, friction force at 5 nN of normal load, and coefficient of friction for various SAMs on Si substrate [30]. The adhesive force showed an increase with the temperature, from room temperature (RT) to about 55 °C, followed by a decrease from 55 °C to 75 °C, and eventually leveled off from 75 °C to 100 °C. The initial increase of adhesive force at lower temperatures is not well understood. The observed decrease could be attributed to the desorption of water molecules on the surface. After almost full depletion of the water

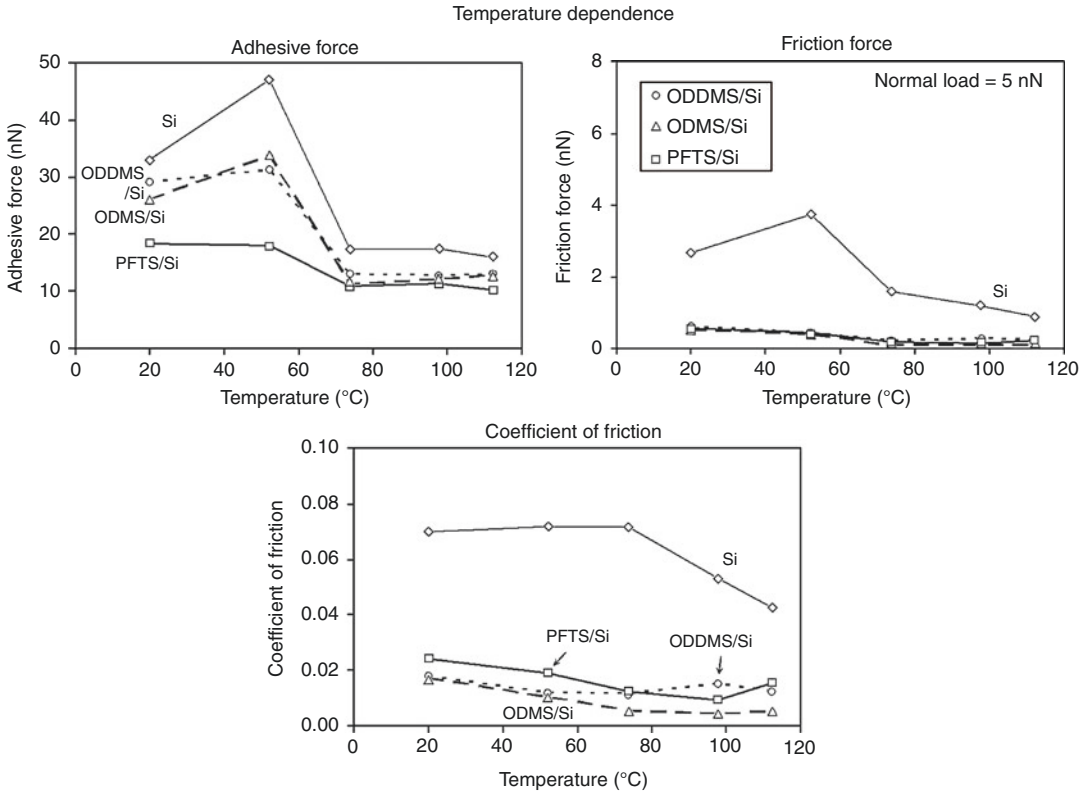


Self-Assembled Monolayers for Nanotribology, Fig. 7 Relative humidity effect on adhesive force, friction force, coefficient of friction and microwear for various SAMs on Si substrates [30]

layer, the adhesive force remains constant. The SAMs with hydrocarbon backbones showed similar behavior as that of the Si substrates, but the initial increase in the adhesive force with temperature was smaller. The SAMs with fluorocarbon backbone chains showed almost no temperature dependence. For the SAMs with hydrocarbon backbone chains, the initial increase in adhesive force is believed to be caused by the melting of the SAM film. The melting point for a linear carbon chain molecule such as $CH_3(CH_2)_{14}CH_2OH$ is 50 °C [32]. With an increase in temperature, the SAM film softens, thereby increasing the real area of contact and consequently the adhesive force. Once the temperature is higher than the melting point, the lubrication regime is changed from boundary lubrication in a solid SAM to liquid lubrication in the melted SAM [33].

The friction force for silicon showed an increase with temperature followed by a steady decrease. The friction force is highly affected by the change in adhesion. The decrease in friction can result from the depletion of the water layer. The coefficient of friction for silicon remained constant followed by a decrease starting at about 80 °C. For SAMs, the coefficient of friction exhibited a monotonic decrease with temperature. The decrease in friction and coefficient of friction for SAMs possibly results from the decrease in stiffness. As introduced before, the spring model suggests a smaller friction for more compliant SAMs [33]. The difference among the SAM types was not significant. PFTS could maintain its stiffness more than ODMS and ODDMS when temperature is increased [20], however, it was not pronounced in the results.

Figure 9a shows the relationship between the decrease in surface height as a function of the normal load during wear tests [18, 30]. As shown in the figure, the SAMs exhibit a critical



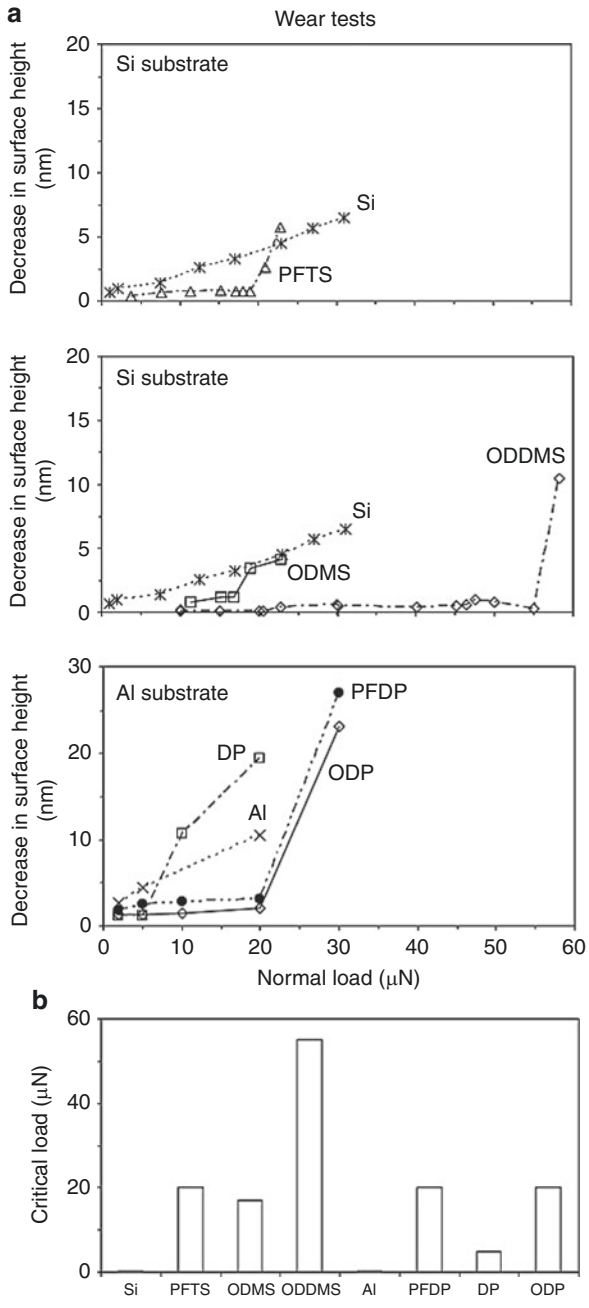
Self-Assembled Monolayers for Nanotribology, Fig. 8 Temperature effect on adhesive force, friction force, and coefficient of friction for various SAMs on Si substrates [30]

normal load, beyond the point of which the surface height drastically decreases. Figure 9a also shows the wear behavior of the Al and Si substrates. Unlike the SAMs, the substrates show a monotonic decrease in surface height with the increasing normal load with wear initiating from the very beginning, i.e., even for low normal loads. Si (Young's modulus of elasticity, E = 130 GPa [1], hardness, H = 11 GPa [6]) is relatively hard in comparison to Al (E = 77 GPa, H = 0.41 GPa), and hence the decrease in surface height for Al is much larger than that for Si for similar normal loads.

The critical loads corresponding to the sudden failure of SAMs are shown in Fig. 9b. Amongst all the SAMs, ODDMS shows the best performance in the wear tests, and this is believed to be because of the longer chain length effect. Fluorinated SAMs – PFTS and PFDP show a higher critical load as compared to ODMS and DP with similar chain lengths. ODP shows a higher critical load as compared to DP because of its longer chain length. The mechanism of failure of compliant SAMs during wear tests has been presented earlier in Fig. 5. It is believed that the SAMs fail mostly due to shearing of the molecule at the head group, that is, by means of shearing of the molecules off the substrate.

Closure

The contact angle, adhesion, friction, and wear properties of SAMs having alkyl, biphenyl, and perfluoroalkyl spacer chains with different surface terminal groups (–CH_3 and –CF_3) and head groups (–SH, –Si–O–, –OH, and P–O–) studied are presented in this article. It is found that the adhesive force varies linearly with the work of adhesion value of SAMs, which indicates that capillary condensation of water plays an important role to the adhesion of SAMs on the



Self-Assembled Monolayers for Nanotribology, Fig. 9 (**a**) Decrease in surface height as a function of normal load after one scan cycle for various SAMs on Si and Al substrates, and (**b**) comparison of critical loads for failure during wear tests for various SAMs [18, 30]

nanoscale at ambient conditions. SAMs with high-compliance long carbon spacer chains exhibit the lowest adhesive force and friction force. The friction data are explained using a molecular spring model, in which the local stiffness and intermolecular force govern frictional performance. The results of the stiffness and friction characterization of the micropatterned sample with different structures support this model. Perfluoroalkylsilane and perfluoroalkylphosphonate SAMs exhibit lower surface energy, higher contact angle, and lower adhesive force as compared to that of alkylsilane and alkylphosphonate SAMs, respectively. The substrate had

S

little effect. The coefficient of friction of various SAMs were comparable.

The influence of relative humidity on adhesion and friction of SAMs is dominated by the thickness of the adsorbed water layer. At higher humidity, water increases friction through increased adhesion by meniscus effect in the contact zone. With an increase in temperature, the desorption of the adsorbed water layer and reduction of the surface tension of water reduces the adhesive force and friction force. A decrease in adhesion and friction with temperature was found for all films.

PFTS/Si showed a better wear resistance than ODMS/Si. ODDMS/Si showed a better wear resistance than ODMS/Si due to the chain length effect. Wear behavior of the SAMs is mostly determined by the molecule-substrate bond strengths. Similar trends were observed for films on Al substrates.

Cross-References

- ▶ Nanotechnology
- ▶ Nanotribology
- ▶ Reliability of Nanostructures

References

1. Anonymous: Properties of silicon, EMIS data reviews series No. 4, INSPEC, Institution of Electrical Engineers, London (1988) (see also Anonymous, MEMS materials database, http://www.memsnet.org/material/ (2002))
2. Bhushan, B.: Tribology and Mechanics of Magnetic Storage Devices, 2nd edn. Springer, New York (1996)
3. Bhushan, B. (ed.): Tribology Issues and Opportunities in MEMS. Kluwer, Dordrecht (1998)
4. Bhushan, B. (ed.): Handbook of Micro/Nanotribology, 2nd edn. Boca Raton, CRC Press (1999)
5. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
6. Bhushan, B.: Chemical, mechanical and tribological characterization of ultra-thin and hard amorphous carbon coatings as thin as 3.5 nm: recent developments. Diam. Relat. Mater. **8**, 1985–2015 (1999)
7. Bhushan, B. (ed.): Modern Tribology Handbook. Principles of Tribology, Vol. 1. Materials, Coatings, and Industrial Applications, vol. 2. CRC Press, Boca Raton (2001)
8. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
9. Bhushan, B.: Nanotribology and nanomechanics of MEMS/NEMS and BioMEMS/BioNEMS materials and devices. Microelectron. Eng. **84**, 387–412 (2007)
10. Bhushan, B.: Nanotribology and nanomechanics in nano/biotechnology. Philos. Tr. R. Soc. A **366**, 1499–1537 (2008)
11. Bhushan, B.: Springer Handbook of Nanotechnology, 3rd edn. Springer, Heidelberg (2010)
12. Bhushan, B.: Nanotribology and Nanomechanics I – Measurement Techniques and Nanomechanics, II – Nanotribology, Biomimetics, and Industrial Applications, 3rd edn. Springer, Heidelberg (2011)
13. Bhushan, B., Cichomski, M.: Nanotribological characterization of vapor phase deposited fluorosilane self-assembled monolayers deposited on polydimethylsiloxane surfaces for biomedical micro-/nanodevices. J. Vac. Sci. Technol. A **25**, 1285–1293 (2007)
14. Bhushan, B., Liu, H.: Nanotribological properties and mechanisms of alkylthiol and biphenyl thiol self-assembled monolayers studied by atomic force microscopy. Phys. Rev. B **63**, 245412-1–245412-11 (2001)
15. Bhushan, B., Israelachvili, J.N., Landman, U.: Nanotribology: friction, wear and lubrication at the atomic scale. Nature **374**, 607–616 (1995)
16. Bhushan, B., Kulkarni, A.V., Koinkar, V.N., Boehm, M., Odoni, L., Martelet, C., Belin, M.: Microtribological characterization of self-assembled and Langmuir-Blodgett monolayers by atomic and friction force microscopy. Langmuir **11**, 3189–3198 (1995)
17. Bhushan, B., Kasai, T., Kulik, G., Barbieri, L., Hoffmann, P.: AFM study of perfluorosilane and alkylsilane self-assembled monolayers for anti-stiction in MEMS/NEMS. Ultramicroscopy **105**, 176–188 (2005)
18. Bhushan, B., Cichomski, M., Hoque, E., DeRose, J. A., Hoffmann, P., Mathieu, H.J.: Nanotribological characterization of perfluoroalkylphosphonate self-assembled monolayers deposited on aluminum-coated silicon substrates. Microsyst. Technol. **12**, 588–596 (2006)
19. Bhushan, B., Hansford, D., Lee, K.K.: Surface modification of silicon surfaces with vapor phase deposited ultrathin fluorosilane films for biomedical devices. J. Vac. Sci. Technol. A **24**, 1197–1202 (2006)
20. Callister, W.D.: Materials Science and Engineering, 4th edn. Wiley, New York (1997)
21. DeRose, J.A., Hoque, E., Bhushan, B., Mathieu, H.J.: Characterization of perfluorodecanote self-assembled monolayers on aluminum and comparison of stability with phosphonate and siloxy self-assembled monolayers. Surf. Sci. **602**, 1360–1367 (2008)
22. DeVecchio, D., Bhushan, B.: Localized surface elasticity measurements using an atomic force microscope. Rev. Sci. Instrum. **68**, 4498–4505 (1997)

23. Eustathopoulos, N., Nicholas, M., Drevet, B.: Wettability at High Temperature. Pergamon, Amsterdam (1999)
24. Hoque, E., DeRose, J.A., Hoffmann, P., Mathieu, H.J., Bhushan, B., Cichomski, M.: Phosphonate self-assembled monolayers on aluminum surfaces. J. Chem. Phys. **124**, 174710 (2006)
25. Hoque, E., DeRose, J.A., Kulik, G., Hoffmann, P., Mathieu, H.J., Bhushan, B.: Alkylphosphonate modified aluminum oxide surfaces. J. Phys. Chem. B **110**, 10855–10861 (2006)
26. Hoque, E., DeRose, J.A., Hoffmann, P., Bhushan, B., Mathieu, H.J.: Alkylperfluorosilane self-assembled monolayers on aluminum: a comparison with alkylphosphonate self-assembled monolayers. J. Phys. Chem. C **111**, 3956–3962 (2007)
27. Hoque, E., DeRose, J.A., Hoffmann, P., Bhushan, B., Mathieu, H.J.: Chemical stability of nonwetting, low adhesion self-assembled monolayer films formed by perfluoroalkylsilazation of copper. J. Chem. Phys. **126**, 114706 (2007)
28. Hoque, E., DeRose, J.A., Bhushan, B., Mathieu, H.J.: Self-assembled monolayers on aluminum and copper oxide surfaces: surface and interface characteristics, nanotribological properties, and chemical stability. In: Bhushan, B., Fuchs, H., Tomitori, M. (eds.) Applied Scanning Probe Methods Vol. IX – Characterization, pp. 235–281. Springer, Heidelberg (2008)
29. Hoque, E., DeRose, J.A., Bhushan, B., Hipps, K.W.: Low adhesion, non-wetting phosphonate self-assembled monolayer films formed on copper oxide surfaces. Ultramicroscopy **109**, 1015–1022 (2009)
30. Kasai, T., Bhushan, B., Kulik, G., Barbieri, L., Hoffmann, P.: Nanotribological study of perfluorosilane SAMs for anti-stiction and low wear. J. Vac. Sci. Technol. B **23**, 995–1003 (2005)
31. Lee, K.K., Bhushan, B., Hansford, D.: Nanotribological characterization of perfluoropolymer thin films for BioMEMS applications. J. Vac. Sci. Technol. A **23**, 804–810 (2005)
32. Lide, D.R.: CRC Handbook of Chemistry and Physics, 85th edn. CRC Press, Boca Raton (2004)
33. Liu, H., Bhushan, B.: Investigation of nanotribological properties of alkylthiol and biphenyl thiol self-assembled monolayers. Ultramicroscopy **91**, 185–202 (2002)
34. Liu, H., Bhushan, B.: Orientation and relocation of biphenyl thiol self-assembled monolayers. Ultramicroscopy **91**, 177–183 (2002)
35. Liu, H., Bhushan, B., Eck, W., Stadler, V.: Investigation of the adhesion, friction, and wear properties of biphenyl thiol self-assembled monolayers by atomic force microscopy. J. Vac. Sci. Technol. A **19**, 1234–1240 (2001)
36. Man, K.F., Stark, B.H., Ramesham, R.: A Resource Handbook for MEMS Reliability. Rev. A. JPL Press, Jet Propulsion Laboratory, California Institute of Technology, Pasadena (1998). See also, Man, K.F.: MEMS reliability for space applications by elimination of potential failure modes through testing and analysis. http://www-rel.jpl.nasa.gov/Org/5053/atpo/products/Prod-map.html (2002)
37. Ren, S., Yang, S., Zhao, Y., Yu, T., Xiao, X.: Preparation and characterization of ultrahydrophobic surface based on a stearic acid self-assembled monolayer over polyethyleneimine thin films. Surf. Sci. **546**, 64–74 (2003)
38. Tambe, N.S., Bhushan, B.: Nanotribological characterization of self assembled monolayers deposited on silicon and aluminum substrates. Nanotechnology **16**, 1549–1558 (2005)
39. Tanner, D.M., Smith, N.F., Irwin, L.W., et al.: MEMS Reliability: Infrastructure, Test Structure, Experiments, and Failure Modes. Sandia National Laboratories, Albuquerque (2000). SAND2000–0091
40. Tao, Z., Bhushan, B.: Degradation mechanisms and environmental effects on perfluoropolyether self assembled monolayers and diamond like carbon films. Langmuir **21**, 2391–2399 (2005)
41. Tao, Z., Bhushan, B.: Surface modification of AFM silicon probes for adhesion and wear reduction. Tribol. Lett. **21**, 1–16 (2006)
42. Wang, Y., Bhushan, B.: Boundary slip and nanobubble study in micro/nanofluidics with atomic force microscope. Soft Matter **6**, 29–66 (2010)

Self-Assembled Nanostructures

▶ Self-Assembly of Nanostructures

Self-Assembled Protein Layers

▶ Interfacial Investigation of Protein Films Using Acoustic Waves

Self-Assembly

▶ Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly
▶ Self-Assembly for Heterogeneous Integration of Microsystems

Self-Assembly for Heterogeneous Integration of Microsystems

J. H. Hoo[1], K. S. Park[1], R. Baskaran[1,2] and K. F. Böhringer[1]
[1]Department of Electrical Engineering, University of Washington, Seattle, WA, USA
[2]Components Research, Intel Corporation, Santa Clara, CA, USA

Synonyms

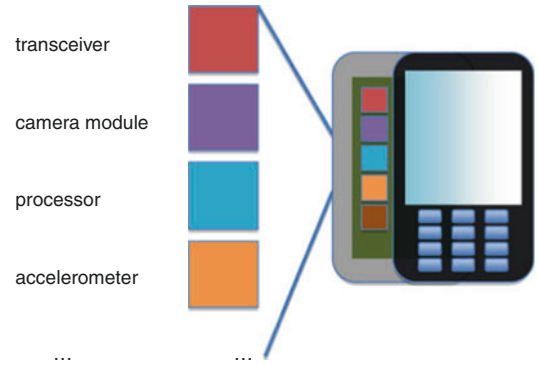
Heterogeneous integration; Self-assembly; Stochastic assembly

Definition

Self-assembly is defined as the autonomous organization of components into ordered patterns or structures without human intervention. In this entry, the scope of self-assembly will be restricted within the microelectronics field – typically, the process of transporting, aligning, and permanently adhering discrete components onto various substrates, achieving heterogeneous integration.

Introduction

The packaging and integration of microscale modules, such as sensors, actuators and transceivers, and data processing, microfluidic, power management, electromechanical and optoelectronic devices into heterogeneous multifunctional systems is one of the big challenges in the field of microelectromechanical systems (MEMS). The fabrication processes and the material prerequisites for the components are predominantly incompatible, thereby eliminating the possibility of fabricating them all on a common substrate, as it would have been done in a monolithic approach. Consequently, in a heterogeneous approach, the components of a system are fabricated separately under respective optimal conditions before being assembled to construct a functional system (Fig. 1).

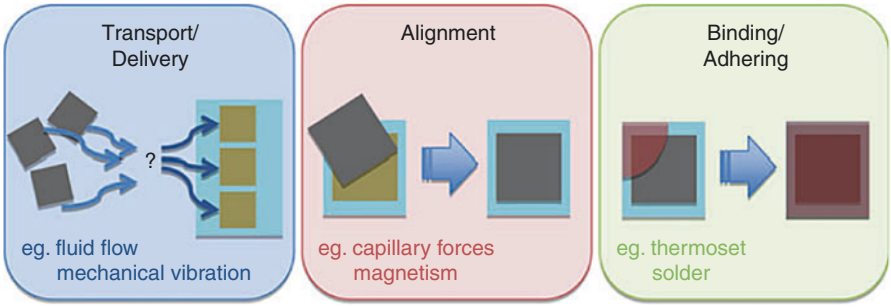


Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 1 A modern cell phone: an example of a multifunctional device put together by discrete components from incompatible fabrication processes

The prevailing method used for assembly and packaging in microelectronic manufacturing is robotic pick-and-place. This technique has proven accurate and reliable for large component scales, and its throughput is satisfying for consumer electronics applications [1]. However, pick-and-place is confronted with a trade-off between throughput and component placement accuracy. Furthermore, the method is serial, it requires closed-loop control, and it becomes more expensive as device dimensions shrink, and registration constraints become more stringent. Furthermore, stiction problems [2] set in for device sizes smaller than 300 μm [1].

Self-assembly, defined as *the autonomous organization of components into ordered patterns or structures without human intervention* [3], provides a promising alternative to pick-and-place machinery. Self-assembly is parallel in nature and can be applied to a wide range of sizes, from the millimeter to nanometer scales. Self-assembly processes can be engineered to avoid direct contact manipulation of components, thereby eliminating the issues with stiction.

Self-assembly is certainly not a suite of techniques exclusive to the microelectronic field. Self-assembling processes (also satisfying the stated definition) from the natural environment and reported in other scientific disciplines range from the noncovalent association of organic molecules to the formation of crystals from precipitates to



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 2 Self-assembly process can be separated into the transport, alignment, and permanent binding/adhering phases. During the transport phase, components are brought to the immediate vicinity of binding/assembly sites; the alignment phase corrects the orientation of the component; the binding phase adheres the component onto the binding site permanently

DNA origami [4]. The scope of this entry will be restricted to the application of self-assembly within the microelectronics field – typically, the process of transporting, aligning, and permanently adhering discrete components onto various substrates (Fig. 2). It is important to note the stochastic nature of the transportation phase; it is impossible, and unnecessary, for a user of self-assembly to determine the specific component that gets delivered to a specific binding location. Table 1 tabulates some notable microscale self-assembly techniques in the field.

Self-assembly techniques can be classified by the underlying mechanisms that drive the transportation and alignment assembly steps (Table 1). Having introduced self-assembly as a possible alternative for prevailing methods of microcomponent assembly and packaging, metrics necessary to the evaluation and comparison of self-assembly techniques are also listed, including the dimensions of the discrete components to be assembled (lateral size, thickness, and aspect ratio), how closely can components be assembled on the substrate (packing density), if there is control over the orientation of the placement of the components (orientation specificity), and the average rate of completeness of the assembly processes (yield).

The successful application of self-assembly requires not only a match between the dimensions of the discrete components reported in Table 1 and that of the components one wishes to assemble. It is also important to consider the assembly mechanisms, which are closely related to the environment in which the self-assembly takes place. For example, one should avoid any self-assembly techniques requiring magnetic force should the discrete components be sensitive to magnetic force. In the following sections of the entry, most of the assembly methodologies will be further examined.

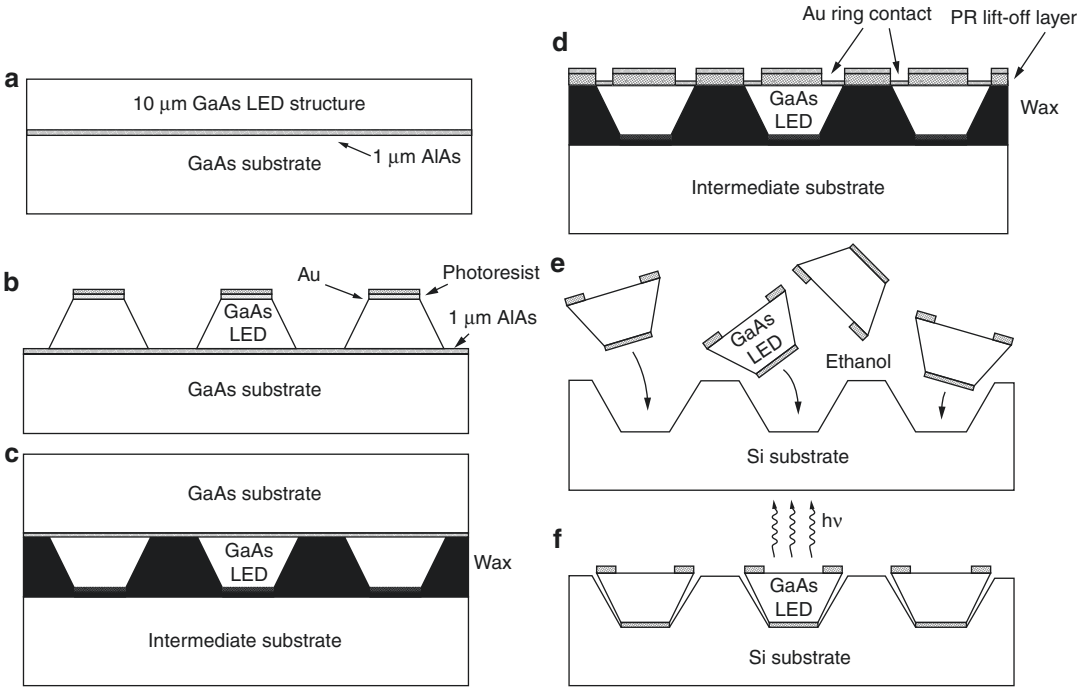
Fluidic Transport–Shape-Matching Alignment

Yeh et al. pioneered the work on self-assembly based on fluidic flow and shape matching [5]. Trapezoidal-shaped light-emitting diode (LED) devices were fabricated on GaAs substrates (Fig. 3) and released into ethanol or methanol, the carrier fluids of choice, which inhibits the oxidation and, therefore, degradation of the GaAs structures.

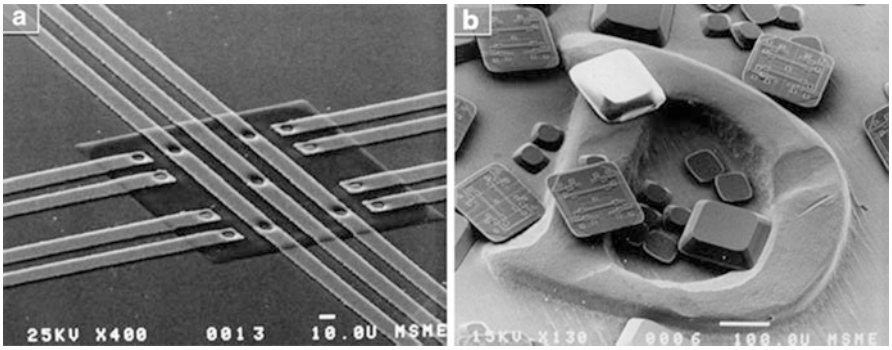
The LEDs were flowed over silicon host wafers fabricated with corresponding trapezoidal holes to capture them (Fig. 3). After the holes were (mostly) filled, the carrier fluid was evaporated, leaving the GaAs LEDs sitting in the holes, attached to the silicon surface only by van der Waals forces. The flowing of LEDs over the silicon substrate is identified as the transportation component of a self-assembly process. The alignment step was realized by shape matching between the trapezoidal LEDs and the corresponding silicon holes. The final procedure

Self-Assembly for Heterogeneous Integration of Microsystems, Table 1 Survey of current self-assembly process. Components are square in shape except where marked by [C], and [R], which are circular and rectangular respectively. Part sizes of components are taken to be the length for square and rectangular parts, and the diameter for circular components

Researcher, institution	Assembly mechanism					Part size (lateral dim) (μm)	Thickness (μm)	Aspect ratio (lateral/ thickness)	Packing density (in plane) (%)	Orientation specificity	Reported yield (%)
	Transport		Aligment								
	Fluidic	Mechanical agitation	Shape matching	Capillary forces	Magnetic						
Yeh, Berkeley	✓		✓			18	9.9	1.8	–	No	>90
Parviz, UW	✓		✓	✓		100[C]	10–20	5–10	20	Yes	97
Srinivasan, Berkeley	✓			✓		150–400	15–50	8–10	16	No	100
Jacobs, UMN & Harvard	✓			✓		20	10	2	25	No	>98
Böhringer, UW	✓			✓		1,000–2,000	100	10–20	25	No	100
Böhringer, UW	✓			✓	✓	2,000	100	20	40	Yes	100
Böhringer, UW		✓	✓			370	150	2.47	80	No	100
Böhringer, UW		✓	✓			400[R]	200	2	35	No	100
Ramadan, IME		✓	✓		✓	1,000[C]	350	2.86	6	No	97
Fischer, KTH		✓	✓		✓	35[C]	350[L]	–	–	No	>95



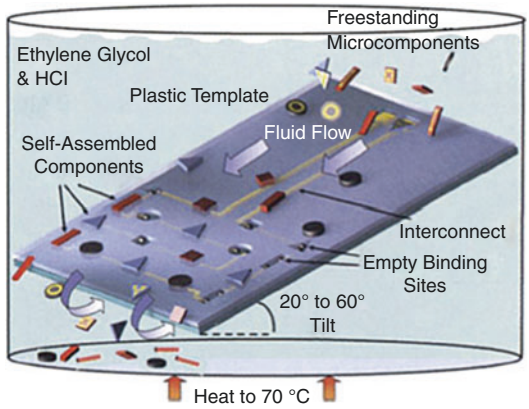
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 3 Schematic diagram of the fluid self-assembly process: (**a**) Molecular beam epitaxy (MBE) grown structure with 1 μm AlAs etch-stop layer, (**b**) trapezoidal GaAs mesa definition, (**c**) bonding to intermediate substrate with wax, (**d**) top-side ring contact metallization, (**e**) solution containing the GaAs blocks dispensed over patterned Si substrate and (**f**) Si substrate with GaAs, light-emitting diodes integrated by fluidic self-assembly (Reprinted with permission from [5], © 1994 IEEE)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 4 Scanning electron microscope (SEM) picture of Yeh's devices: (**a**) a device assembled using fluidic self-assembly techniques – the metallic connections are deposited over the device such that mechanical and electrical connection is established. (**b**) Devices placed on a US dime for scale (From [6], reprinted with permission from Alien Technology)

of binding was performed by the deposition of metallic connections between the devices and the host substrate, establishing electrical and mechanical connection, permanently adhering the chips and the substrate (Fig. 4). Yeh's self-assembly technique was subsequently commercialized, and has been adapted to be a high volume manufacturing compatible process, specializing in the production of radio frequency identification (RFID) tags [6].

By combining Yeh's concept of shape matching under fluidic flow with capillary forces as drivers of self-assembly, Parviz et al. demonstrates the possibility to assemble multiple device types simultaneously onto a variety of substrates [1], including flexible plastics [7]. Figure 5 summarizes the features of Parviz's multi-device assembly process.

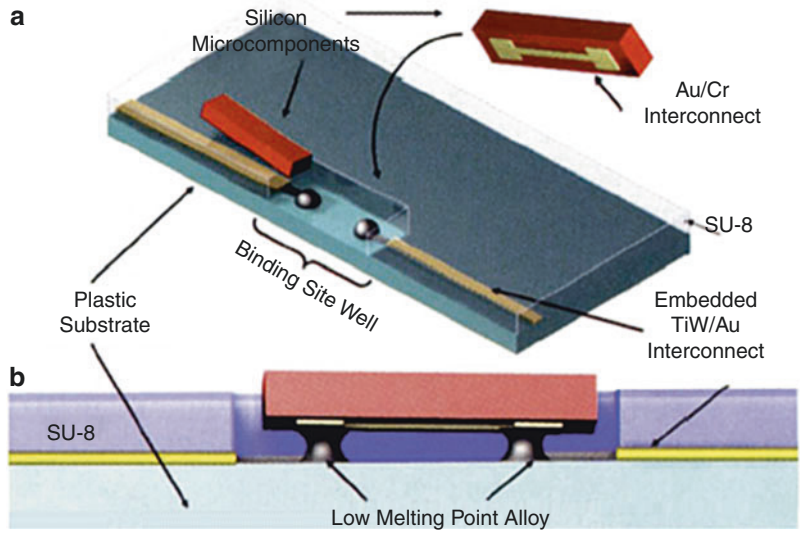


Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 5 The heterogeneous self-assembly process. Microcomponents are introduced over a template submerged in a liquid medium and moved with the fluid flow. Self-assembly occurs as microcomponents first fall into complementarily shaped wells and then become bound by the capillary forces resultant from a molten alloy (Reprinted from [7], with permission from IOP)

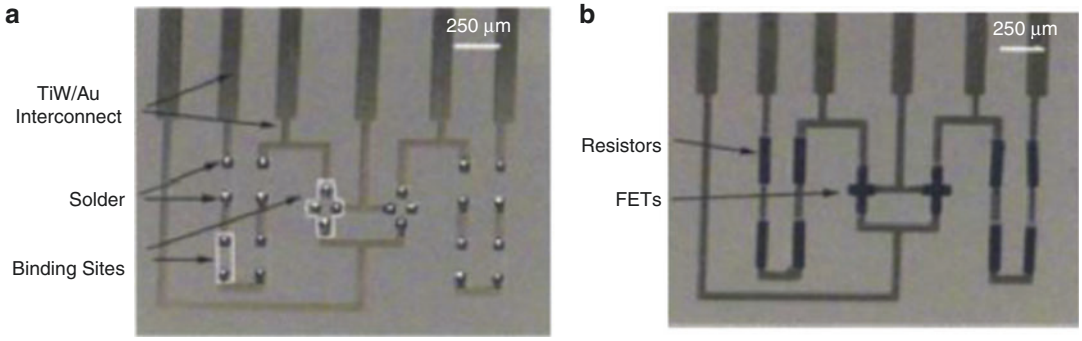
Components of distinct shapes were flowed down the inclined surface of a template. Binding sites patterned on the template correspond to the shapes of components to be assembled. Component–binding site pairs are designed to only be compatible with each other, disallowing unintended assembly. The system is also designed to allow binding sites to only accept components when it happens to approach with the correct surface facing the binding site (Fig. 6).

The assembly process (Figs. 5 and 6) was maintained above the melting temperature of the low-melting-point alloy solder deposited at the contact pads of the template, within the binding sites. When the metal contact pads on a single microcomponent comes into contact with the molten solder, the component will be pulled to the template as the molten solder wets the contact pad on the chip. Conversely, if a component approaches the binding site with the incorrect side, capillary forces will not be exerted on the microcomponent, as molten solder will not wet on silicon surfaces. Excess components collected at the bottom of the container will be reused repeatedly until assembly is complete (Fig. 5).

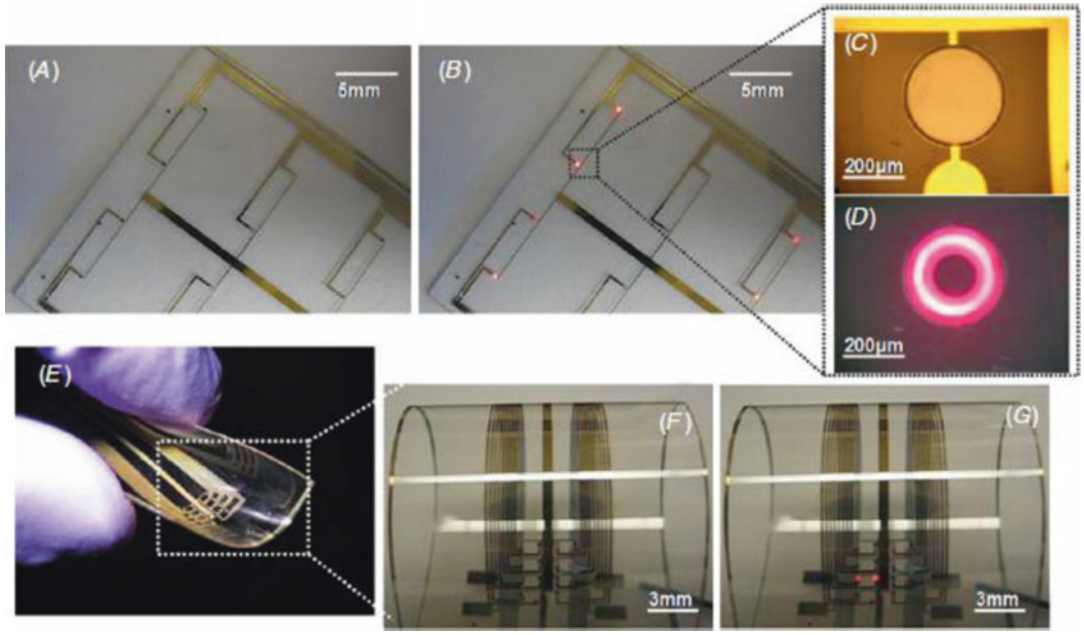
When assembly was observed to be complete, the temperature of the self-assembly setup was lowered to room temperature, thereby solidifying the molten solder, establishing electrical and



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 6 Details of the self-assembly process for a single microcomponent. (**a**) The microcomponent approaches a binding site with complementary shape. (**b**) The microcomponent is held by capillary force resultant from molten-alloy-bridging the metal pads positioned on the microcomponent and on the template (Reprinted from [7], with permission from IOP)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 7 Self-assembly of field effect transistors (FETs). (**a**) Optical microscopic image of a plastic substrate with empty binding sites (two outlined with *white lines* for clarity); (**b**) The template after completion of the self-assembly process showing the position of FETs and diffusion resistors (Reprinted from [7], with permission from IOP)

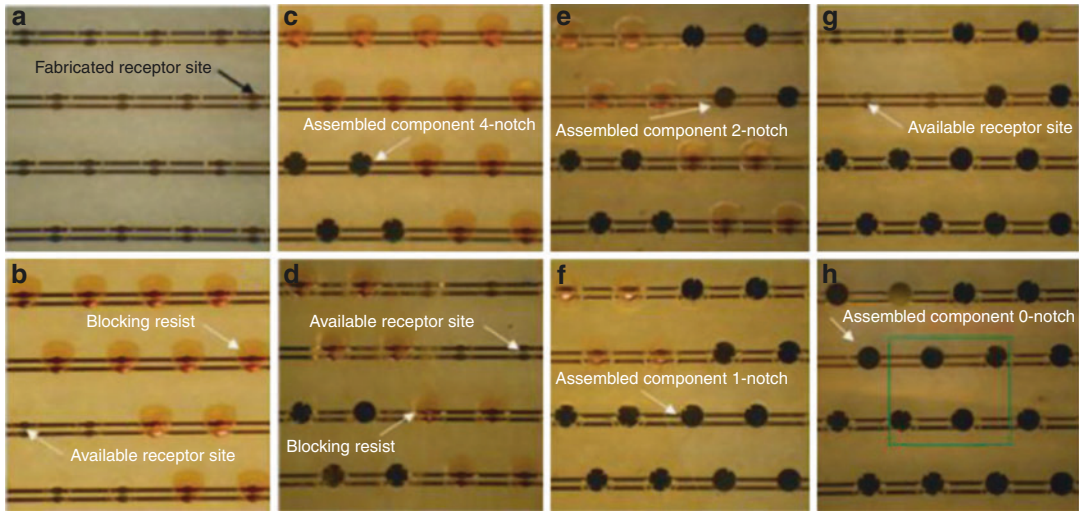


Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 8 (**a**) A glass template after the conclusion of the self-assembly process. The micro (µ)-LEDs are assembled on the template. (**b**) 6 µ-LEDs are turned ON on the same glass template via the application of a 5 V bias. (**c**), (**d**) Close-up images of 1 µ-LED turned OFF and ON, on the template under an optical microscope; (**e**) A flexible plastic template after the completion of the self-assembly process; (**f**, **g**) The same plastic template, bent over itself, with µ-LEDs OFF and then ON upon the application of a 4 V bias (Reprinted from [7], with permission from IOP)

mechanical connection between the template and the discrete components. While the fluidic flow transportation mechanism is similar to Yeh et al., the alignment of microcomponents is achieved through shape-matching and solder capillary action. Finally, binding of the discrete components was achieved through the solidification of the molten solder bumps. Final products are shown in Figs. 7 and 8.

Parviz et al. added programmability to their fluidic shape-matching self-assembly process with a technique that involves the use of

S



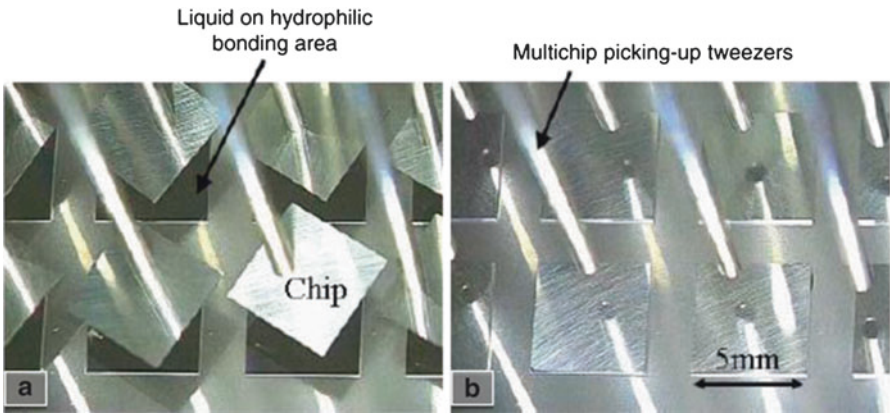
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 9 Optical microscope images of a sequence of events during the self-assembly process of four different microcomponent types. (**a**) Fabricated receptor site wells and C-shaped traps on a plastic template prior to self-assembly; (**b**) Positive photoresist was patterned on the template to block all the receptor sites, except the four receptor sites located on the *bottom left*. (**c**) The first type of components (circular 4-notch) was assembled within the 2×2 parallelogram pattern of available receptor sites. (**d**) A thin layer of photoresist was coated on the whole template to fix the assembled microcomponents. The blocking resist was then removed from the four top right receptor sites designated for the second type of microcomponents. (**e**) The second type of components (circular 2-notch) was assembled in available receptor sites. (**f**) The third type of components (circular 1-notch) was assembled in the *bottom left corner* of the image in a similar fashion. (**g**) The resist was removed from the remaining four receptor sites on the *top left*. (**h**) Final type of components (circular 0-notch) was self-assembled (Reprinted from [8], with permission from MRS)

photosensitive materials to control the access of binding sites to microcomponents being flowed across the template [8]. Programmable binding locations were created by depositing photoresist within receptor sites that uses the principle of shape matching to capture microcomponents that are flowed over them. To activate a group of receptor sites, the region of interest is selectively exposed to ultraviolet radiation to remove the blocking photoresist material, thereby availing said sites to capture incoming components. By successive steps of site activation and component flowing, different components can be assembled onto a single type of receptor sites (Fig. 9).

It is instructive, at this point, to definite "programmability" in the context of micro self-assembly as *the ability to direct different reliably reproducible outcomes with the same set of components in a self-assembly process*. Programmability, which will be revisited in other works, will add to the appeal of self-assembly in industry.

Fluidic Transport–Capillary Forces Alignment

Capillary forces dominate inertial forces and gravity as one approaches submillimeter scales, given that capillarity scales with length, while gravity scales with length cubed. There are many studies that concentrate on the alignment properties of capillary forces instead of an entire self-assembly process as defined in Fig. 2. Notably, work from Koyanagi et al. focuses on the aligning and bonding of silicon chips on micro-fabricated silicon pedestals with corresponding dimensions [9] (Fig. 10).



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 10 Self-alignment process from Koyanagi et al. (**a**) Introduction of chips using a multi-chip vacuum pick system. (**b**) Chips self-aligned after placement on bonding areas (Reprinted with permission from [9], © 2008 IEEE)

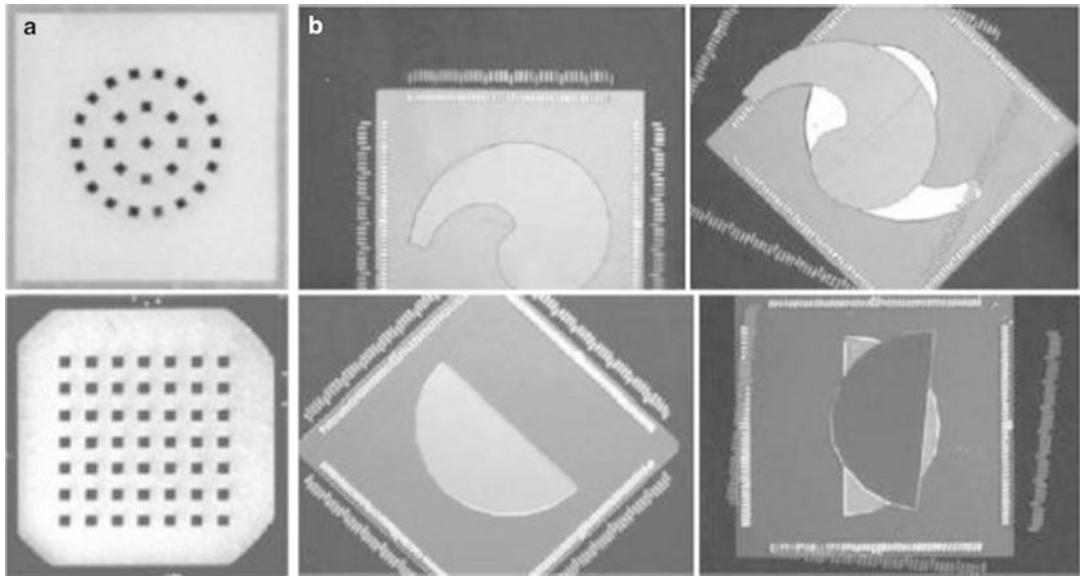
Droplets of an aqueous solution of hydrofluoric acid (HF) were placed on the pedestals before introducing silicon chips with a multichip vacuum pick system. The faces of the chips and the pedestals were treated to be hydrophilic so that the droplet underneath the chips will wet both surfaces and self-align the chips to the pedestals underneath to minimize the interfacial energies. Alignment will improve as the HF solution evaporates, and eventually dissipates entirely, while bonding chips to pedestals permanently with strong SiO_2–SiO_2 bonds.

There are many other works titled "self-assembly" that deal exclusively with alignment rather than the entire self-assembly process which were presented in Fig. 2. "Self-alignment" is probably a more appropriate title for such studies, drawing the distinction between investigations focused on alignment and works that also include a stochastic transportation mechanism that brings discrete components to the vicinity of the intended assemblies.

Srinivasan et al. demonstrated a complete self-assembly technique with fluidic component transportation and capillary force capture and alignment [10]. They patterned a substrate with an array of hydrophobic self-assembled monolayer (SAM)–coated gold-binding sites. By inserting the substrate into water through a film of hydrophobic adhesive floating on the water surface, the hydrophobic binding sites were coated with the adhesive (Fig. 11).

Microparts fabricated from silicon-on-insulator (SOI) wafers were then introduced through a pipette toward the substrate in water. When the hydrophobic pattern on the microparts came into contact with the adhesive, the parts were pulled to the hydrophobic binding sites and self-alignments occurred spontaneously. Finally, when all sites were occupied, the adhesive was polymerized by heat (thermoset) or UV radiation, depending on the type of adhesive used, bonding the chips to the substrate permanently. Binding sites of shapes with in-plane rotational symmetries such as squares gave alignment yields of up to 100 %, with translation and rotational misalignments of less than 0.2 μm and 0.3°, respectively. Component–binding site pairs without in-plane rotational symmetries (aimed at specific in-plane alignment) such as semicircles and commas gave alignments yields of approximately 30–40 %.

Jacobs et al. developed a similar self-assembly process that uses low-temperature melting solder to mount LED arrays on flexible cylindrical templates [11]. The assembly template was patterned with copper squares, and a simple dip-coating process applied molten solder on these copper squares, given that copper has good wetting



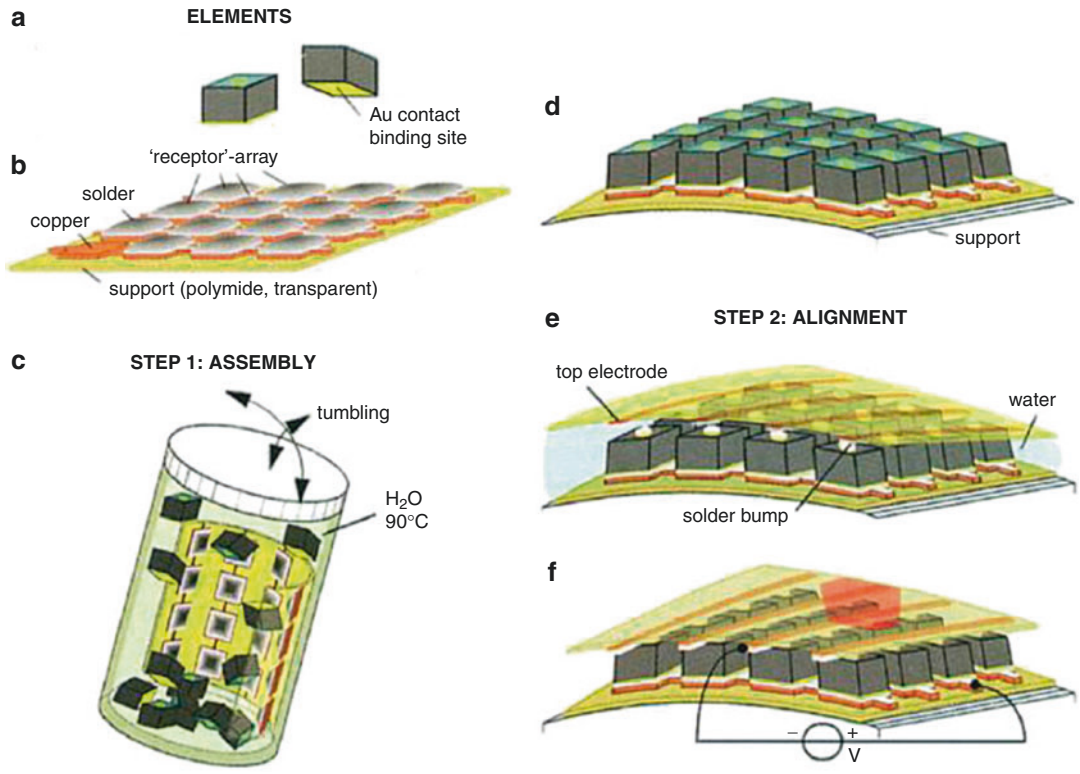
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 11 Optical micrographs of capillary-driven self-assembly of flat microparts: (**a**) *Square* parts on quartz substrates in *ring* and grid configurations and (**b**) correct and wrong alignments for *semicircle*- and *comma-shaped* binding sites (Reprinted with permission from [10], © 2001 IEEE)

characteristics for solders. The dip-coating process was performed in an acidic aqueous environment to prevent oxidation of the solder surface.

Hundreds of LEDs and a flexible assembly template were placed inside a vial; the vial was filled with water and heated to a temperature above the melting point of the solder. The LEDs were tumbled inside the vial, allowing for the LEDs to come into contact with the molten solder, be pulled to a copper square by the solder, and self-aligned (Fig. 12c). The cathode and anode of the LEDs were distinguished by a small round contact and a larger square contact (the size of the copper squares on the template), respectively, on two faces of the discrete LEDs (Fig. 12a). The agitation intensity of the tumbling process was tuned to dissociate assemblies between the solder and the smaller round contact while not severe enough to affect LEDs that were caught with their square contact face. This gives the assembly process side-selectively.

A permanent bond between each LED and the template was established when the template was allowed to cool, solidifying the solder. To make the LED array functional, a transparent film patterned with electrical circuitry was aligned and bonded to the assembled array, establishing the electrical connections to the cathodes of the LEDs. By activating the LEDs (Fig. 13), the defect rate was visually determined to be approximately 2 %.

Jacobs et al. also worked on self-assembly based on the delivery of components on a fluid-fluid interface [12]. Components were fabricated out of silicon and SU-8, with one face coated with gold. The gold surfaces were treated with a mercaptoundecanoic acid (MUA) SAM to render them hydrophilic. The silicon faces were treated to be hydrophobic using 3-glycidoxypropyltrimethoxysilane (GPTMS), creating an orientation preference for components when they are introduced at a silicone oil–water fluid-fluid interface (SU-8 surface are hydrophobic and thus required no further surface modification). By mildly agitating the system, components at the silicone oil–water interface will predominantly orient themselves such that the hydrophilic gold surface will face toward water, and the hydrophobic silicon/SU-8 will face the direction of the silicone oil.

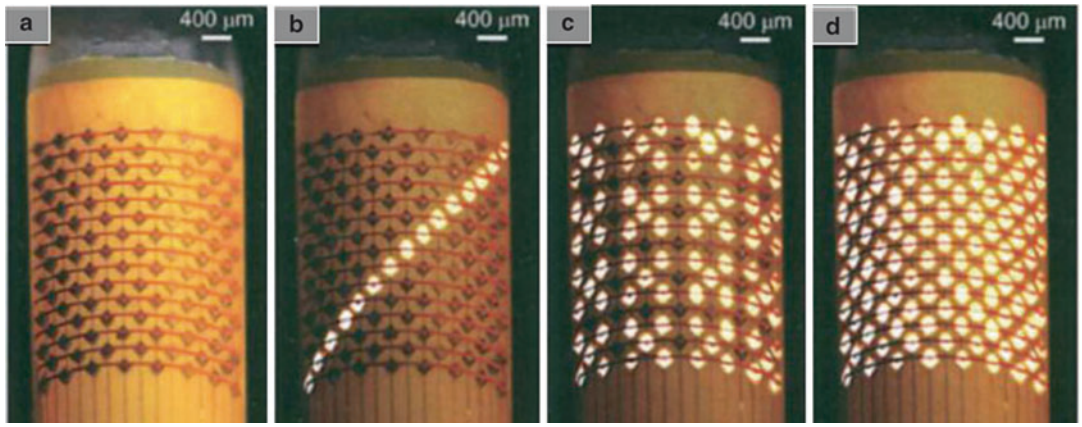


Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 12 Procedure used to assemble a functional cylindrical display. (**a**) *Top* and *bottom* views of an LED segment that has two contacts: a small circular gold contact (cathode) on the *front*, and a large square gold contact (anode) on the *back*. (**b**) Array of solder-coated copper squares supported on a flexible substrate; these squares are of the same size as the anodes of the LEDs and act as receptors for the LEDs during the self-assembly. (**c**) The components are tumbled in a vial at a temperature above the melting point of the solder. (**d**) Two-dimensional array of the assembled LEDs. (**e**) Alignment of a top electrode: The copper wires of the top electrode and the cathodic contacts on the front of the LEDs were first dip coated with solder. The array of wires is prealigned with the array of cathodic solder bumps. At a temperature above the melting point, electrical connections form and the structure self-aligns. (**f**) Test of the self-assembled display-prototype (From [11], reprinted with permission from AAAS)

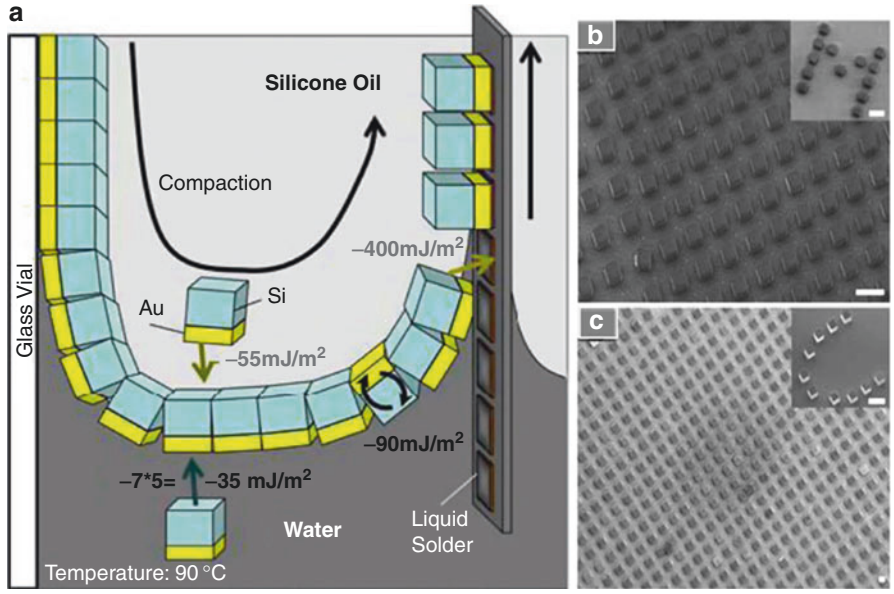
Two types of substrates have been used: silicon and flexible propylene terephthalate (PET). Copper contact pads, corresponding to the sizes of the components, were patterned on the surface. A dip-coating process was used to deposit low-melting-temperature solder onto the copper pads.

The transfer of the components is shown in Fig. 14. A container containing silicone oil, water, and components on the interface between the two fluids was first heated to a temperature beyond the melting point of the solder. Substrates were pulled upward through the liquid-liquid interface, allowing the components to come into contact with the molten solder. Upon contact, the molten solder will pull components in, transferring the components onto the substrates. While it is possible to deliver components to over 90 % of the copper pads on a substrate in a single pass, multiple passes were required to achieve full coverage. Using this method, chips $20 \times 20 \times 10$ and $60 \times 60 \times 20\ \mu m^3$ in size were assembled onto rigid and flexible substrates at reported speeds of up to 62,500 chips in 45 s.

Böhringer et al. studied the assembly of thin microcomponents in a fluidic environment, delivering components at the air-water interface [13]. Components ($1{,}000 \times 1{,}000 \times 100\ \mu m^3$) with one face coated in gold are fabricated in silicon. Gold-coated silicon assembly templates were prepared with square openings, exposing



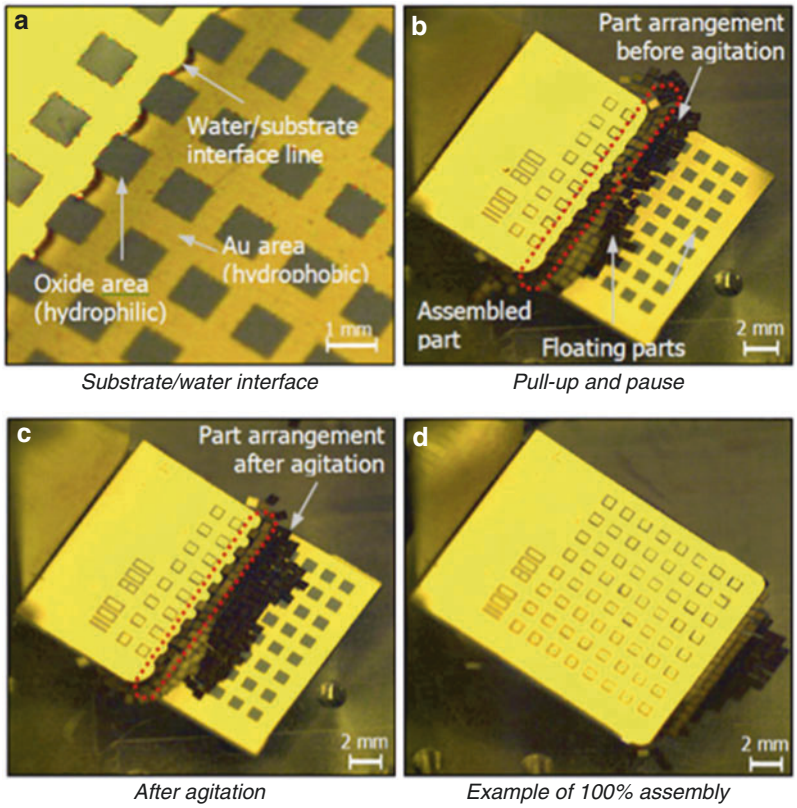
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 13 (**a**) Photograph of the display after self-alignment of the top electrode; (**b**–**d**) Photographs of the operating display after the alignment of the top electrode. The display contains 113 LEDs that are assembled in an interleaved fully addressable array with eight columns of eight receptors interleaved with seven columns of seven on the *bottom* electrode that connect to 15 rows of crossing copper wires on the *top* electrode (From [11], reprinted with permission from AAAS)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 14 (**a**) Procedure of surface tension–directed self-assembly at a liquid-liquid–solid interface. SEM of (**b**) SU-8 (20 μm side length) and (**c**) Si chiplets (20 μm and 60 μm side length) assembling in regular arrays and arbitrary text patterns (*insets*) (Reprinted from [12], © 2010 National Academy of Sciences, USA)

silicon, that were designed to be slightly larger than the lateral dimensions of the components. All gold surfaces were then treated to be hydrophobic using dodecanethiol SAM.

Components were introduced onto a water surface. By applying controlled agitation, the components will preferentially orient themselves with their gold faces upward due to the surface



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 15 Parts rearrangement by agitation. *Darker squares* are hydrophilic silicon surfaces

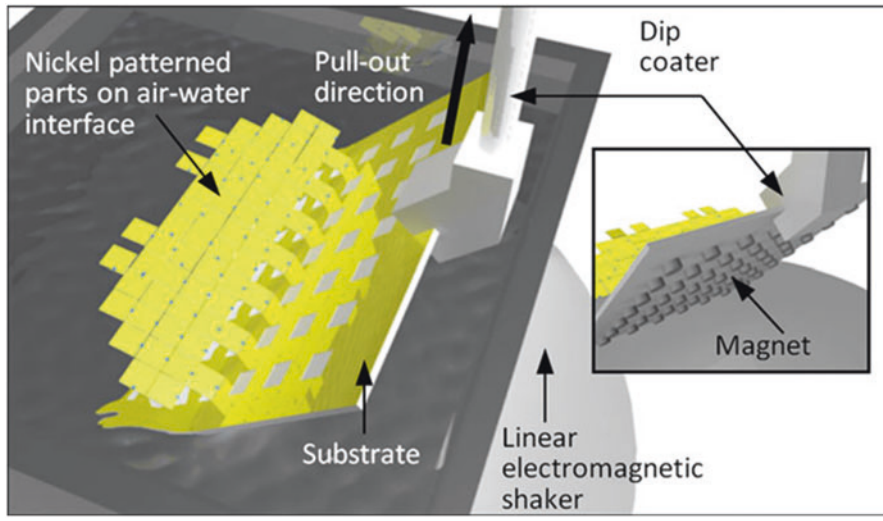
treatment. The assembly template was then pulled up, at an angle, through the air-water interface, in the immediate proximity of the floating components.

When the level of the assembly template reaches the top edges of a row of assembly sites (Fig. 15a), controlled agitation was applied to cause the components to arrange themselves such that the edge of a single part was pinned to the edge of a single site (Fig. 15b, c). At this point, the assembly template was pulled up, allowing an entire row of parts to be transferred and self-aligned to the template. By repeating this pull-up, pause and agitate cycle, entire assembly templates can be perfectly filled. The most significant aspect of this work lies in the aspect ratio of the components being self-assembled; no other self-assembly technique has been demonstrated to work well with very thin components.

The programmability of this process was demonstrated by the ability to assemble only on alternate rows of binding sites. After a single (first) row of parts have been transferred onto the assembly substrate, one can avoid assembling on the subsequent row by agitating the water surface while the substrate is extracted through the air-water interface, bypassing the second row and stopping at the assembly position of the third. Using similar methods, arbitrary row assemblies can be reproduced reliably.

Fluidic Transport–Capillary/Magnetic Forces Alignment

By depositing nickel squares near the edges of chips, and positioning a magnet underneath each binding site underneath the assembly substrate, Böhringer et al. extended the previous fluidic transport–capillary forces alignment self-assembly method to include orientation specificity (Fig. 16).



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 16 The experimental setup consists of a water container, linear electromagnetic actuator, dip coater, assembly substrate, magnets, and test parts floating at the air-water interface

Similar to the previous method, controlled agitation of the water surface ensures one-to-one addressing of parts to binding sites. During this agitation, the interaction of the magnetic forces between the magnets underneath each binding site and the ferromagnetic nickel squares ensures also that chips approach the binding sites with the nickel-patterned edge, allowing for orientation specificity previously unachievable (Fig. 17).

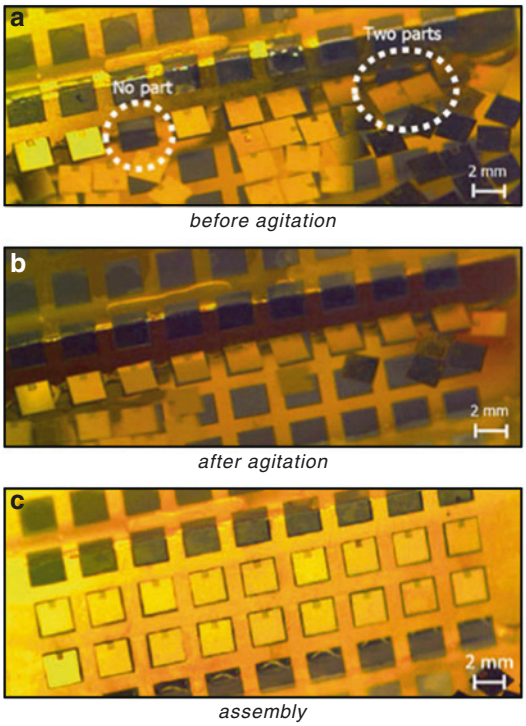
Mechanical Agitation Transport–Capillary Forces Alignment

For discrete components and/or substrates that are adverse to fluidic environments, there are also self-assembly processes that are conducted in dry or "semidry" [14–16] conditions. Dry environment self-assembly processes predominantly require the use of mechanical vibration/agitation to distribute the microcomponents.

Böhringer et al. devised a suite of self-assembly methods to be performed in an air environment that uses a two-step assembly process consisting of a shape-matching step and a capillary force–driven alignment step [14–16] for square silicon parts as well as surface mount technology (SMT) passive devices.

The latest iteration in the technique for the assembly of silicon chips is shown in Fig. 18. A silicon fabricated assembly template was first aligned on top of a transfer substrate. The transfer substrate was fabricated by first depositing gold onto a silicon wafer. Square openings (silicon) corresponding to the size of the microparts were subsequently developed, and the gold surfaces were treated to be hydrophobic with a SAM coating. Each aperture in the assembly template addresses a single silicon opening underneath (Fig. 18h), and can only contain a single discrete component standing on its side.

Square-shaped silicon microcomponents with one gold-coated (hydrophobic SAM-treated) side were delivered to every aperture on the assembly template. When the delivery was completed, the assembly template was removed, and steam was introduced to the system to form water droplets on the hydrophilic silicon surfaces. When the droplets become large enough, they will exert enough



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 17 Assembly process of 2,000 × 2,000 × 100 μm^3 components on the assembly substrate: (**a**) A binding site does not yet have a corresponding component, and another has two test parts in its vicinity. (**b**) After applying specific Faraday waves using agitation, every binding site gets a single component. (**c**) Two rows of components assembled onto the assembly substrate

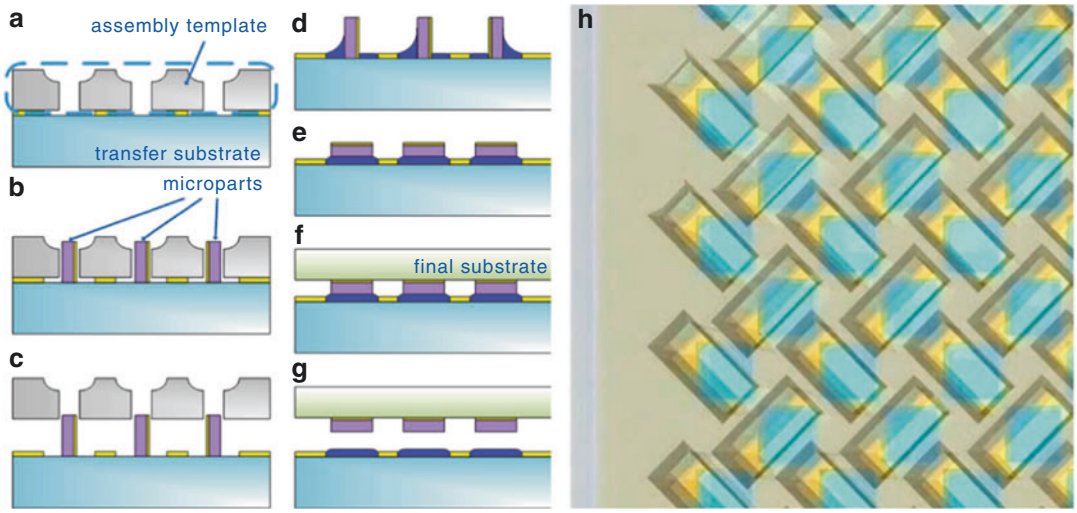
force to pull the silicon face of the microparts flat onto the transfer substrate, and self-align the components to the squares. The aligned microcomponents can then be transferred onto a destination substrate to be permanently mounted. Similar techniques have been tested with parts ranging from 370 × 370 × 150 μm^3 to 790 × 790 × 330 μm^3 in size.

Böhringer et al. also developed a system of control for the precise manipulation of microcomponents on a vibrating surface [15]. Parts can be made to jump or walk (where parts seem to crawl across the assembly template) in predictable directions and speeds. Using these controllable part-motion modes, the transportation phase of the self-assembly process can be programmed to deliver components to arbitrary regions of apertures or even empty out filled apertures. In the case of a programmed "feedback-driven" delivery process [15] (Fig. 19), visual feedback is used to identify empty apertures and direct excess components toward them, guaranteeing the complete delivery of components.

Finally, Böhringer et al. also adapted the template-based methodology (complete with the feedback-driven assembly process) to the assembly of surface mount technology (SMT) thin-film resistors and monolithic ceramic capacitors of the 01005 standard (0.016″ × 0.008″, 0.4 × 0.2 mm) [16]. In this instance, two templates were used, one stacked on top of another. Template 1 (Fig. 20) ensures that only one component will be allowed into template 2, and template 2 guides the in-plane orientation of the component. During the solder reflow process, the molten solder wets the metallized contact-ends of the SMT components and aligns them to the contact pads on the target substrate. At the end of the solder reflow, mechanical and electrical bonds between components and substrate were achieved.

Mechanical Agitation/Magnetic Forces Transport–Shape-Matching Alignment

A dry environment self-assembly technique combining transportation via mechanical vibration and magnetic attractive force with shape-matching alignment was devised by Ramadan et al. [17] (Fig. 21). A host substrate, featuring physical recesses, was superimposed to a master array comprising embedded NdFeB magnets, whose position exactly matched that of the recesses. Target chips, bearing a CoNiP soft magnetic film on their nonfunctional side, were distributed over the substrate, which was mechanically vibrated and tilted. This motion causes chips to be distributed across the entire substrate, until being collected in an enclosure at the lower end of the tilt.



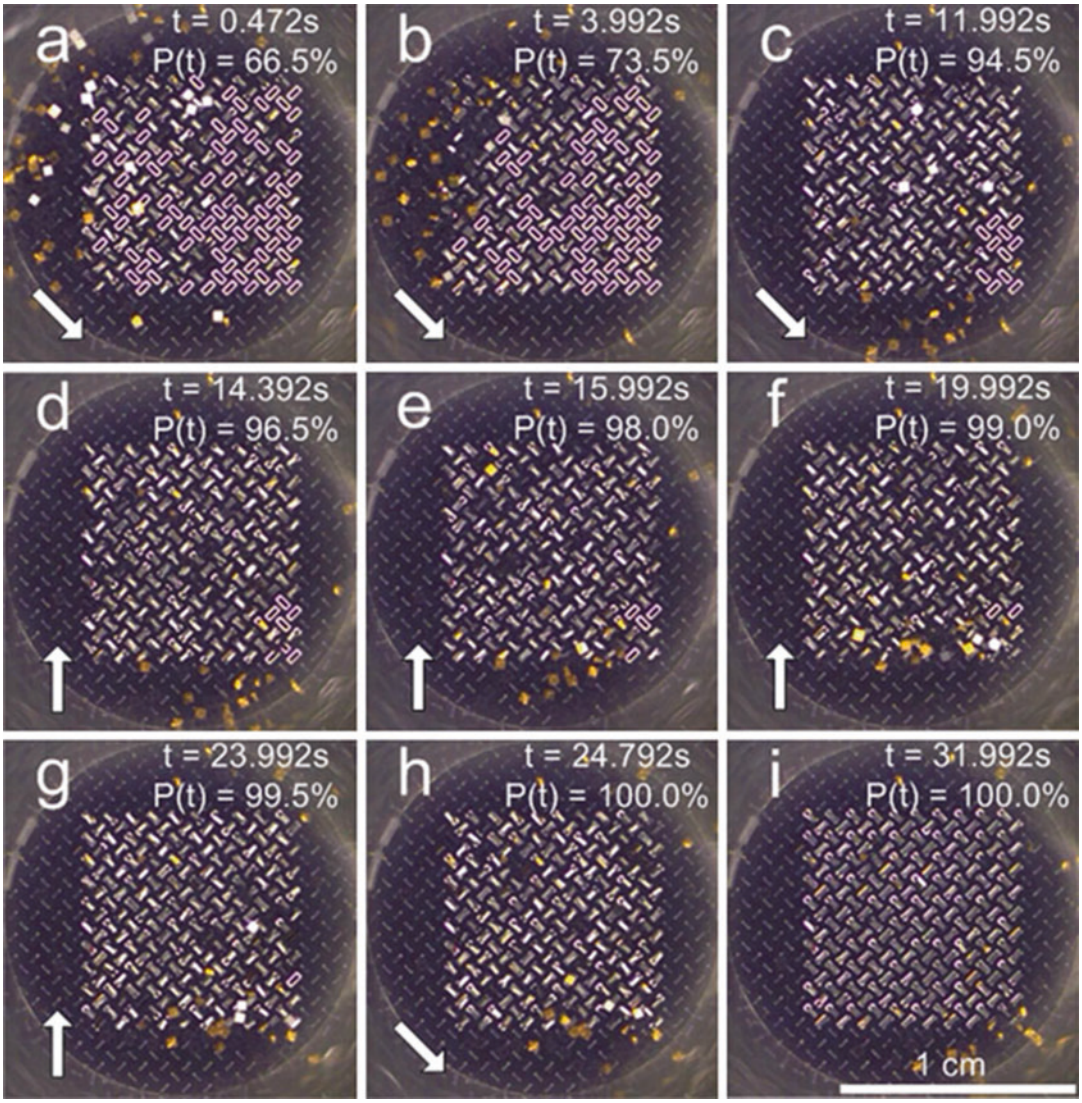
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 18 Full assembly process flow (**a**–**g**): (**a**) Mount assembly template on a gold-patterned transfer substrate. (**b**) Assemble parts. (**c**) Remove template. (**d**) Apply moisture. (**e**) Vibrate the setup gently to have the parts fall on hydrophilic side and self-align. (**f**) Attach parts to destination substrate/device. (**g**) Complete the process by removing the assembled substrate/device from transfer template. (**h**) *Top-down* view of assembly template placed over transfer substrate (Reprinted with permission from [15], © 2009 IEEE)

The amplitude of the substrate vibration and the thickness of chips were tailored so that the only stable configuration possible for the chips to be assembled is to have the functional side facing upward. The in-plane alignment of the chips was controlled by shape matching between the chips and the recesses of the host substrate. After assembly, the master array could be removed and reutilized while the assembled substrate could be further processed. An assembly of 2,500 chips in 5 min with 97 % yield was reported for this technique.

Fischer et al. proposed a scheme to implement through silicon vias (TSVs) by using magnetic force [18]. High-aspect-ratio holes are first etched on a silicon substrate using deep reactive ion etching (DRIE). Commercially available nickel wires (length: 350 μm, diameter: 35 μm) were deposited on the surface of the silicon substrate. A magnet was introduced at the bottom of the substrate, causing the wires to stand upright. By moving the magnet laterally, wires were dragged along the surface (standing), and pulled into the holes (Fig. 22).

After the wires have been assembled into the holes (Fig. 23), the excess space within a wire-occupied hole was filled by a manual application of the thermosetting polymer bisbenzocyclobutene (BCB). The BCB was then cured, permanently adhering the nickel wires within the silicon substrate. Subsequent steps involve steps to expose the ends of the wires on two sides of the silicon, and depositing metal contact lines per specific application. The filling rate of a 30 by 30 array of via holes with a pitch of 120 μm was reported at about 80 % in 20 s, and near complete (>95 %) in about 2 min.

TSVs are vertical electrical connections passing through die layers. They are important for heterogeneous integration solutions involving the stacking of dies. While seemingly out of place with previously mentioned assembly methodologies for chips/dies and passive devices, it is significant to show that there is a viable self-assembly solution for TSV placements, demonstrating that there are self-assembly solutions for every aspect of the device assembly process.

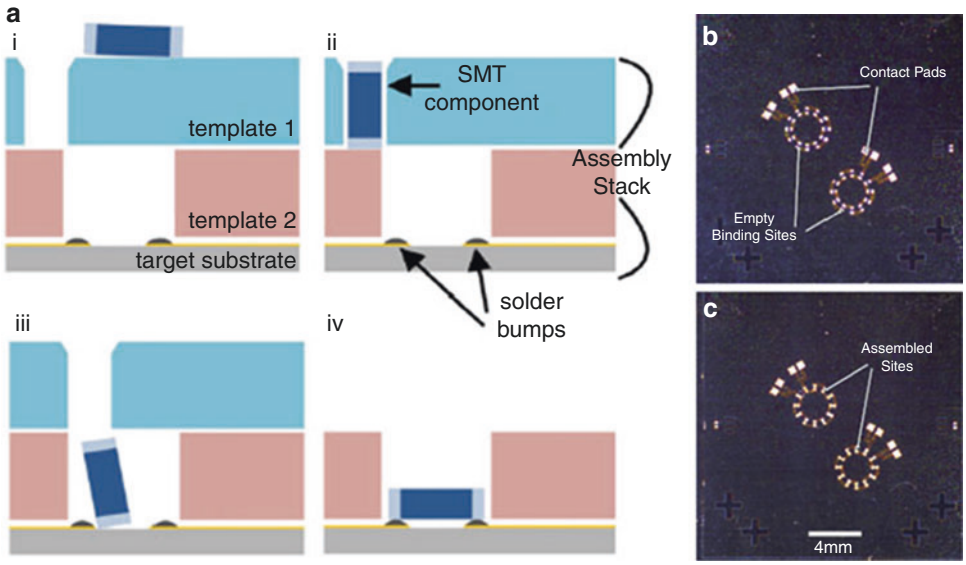


Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 19 Stages of a feedback-driven assembly: (**a**) shows the instance when jumping mode (with a *top left corner* bias) is switched to a walking mode with part-motion direction as indicated by the *white arrow* in the *box*. (**b**) and (**c**) show the progressive filling of the entire assembly area. *Right* after (**c**), when assembly-percentage, P(t), has plateaued, a walking mode in the upward direction as seen in (**d**–**g**) is activated, moving excess parts below the assembly area to the empty sites. After achieving 100 % assembly in (**h**), a walking mode directed toward the lower-right corner is activated, moving excess parts away from the assembly area, finishing the process in (**i**); note: size-scale provided in image (**i**) (Reprinted with permission from [15], © 2009 IEEE)

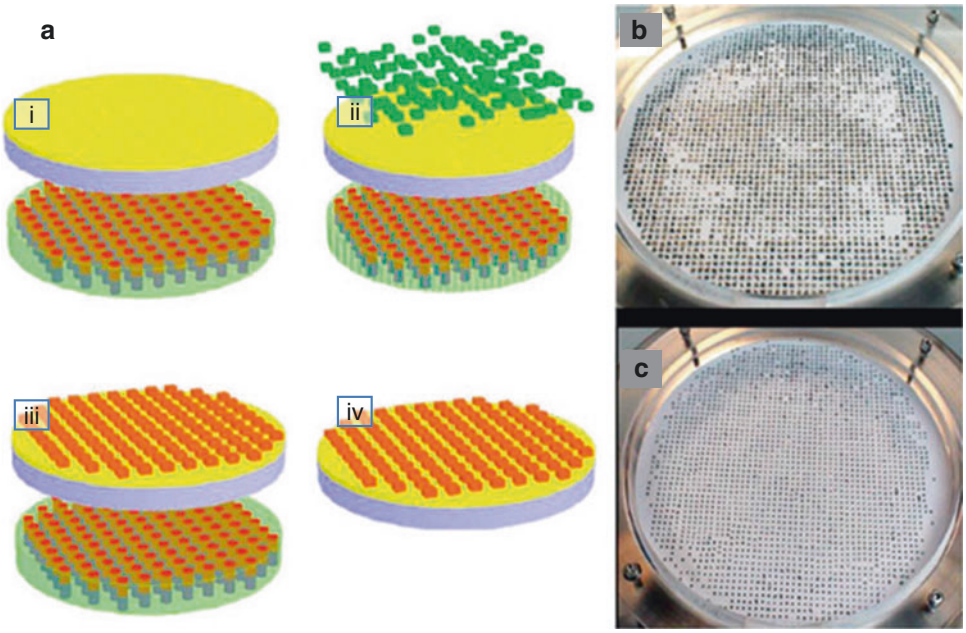
S

Conclusion

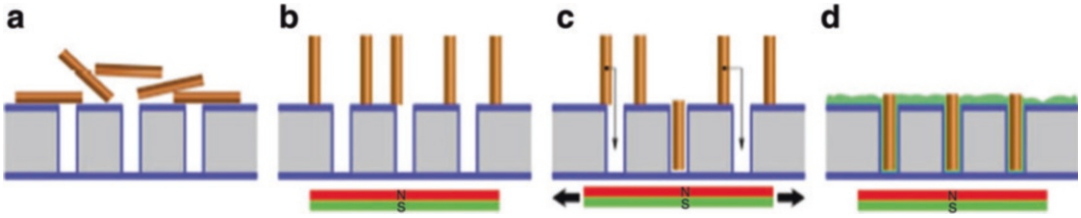
Microscale self-assembly has been presented as a suite of techniques that can enable the advancement of the electronics industry through heterogeneous integration. A selection of self-assembly methodologies that are representative of techniques employing various modes of transportation and alignment, and the associated environments in which they are performed have been



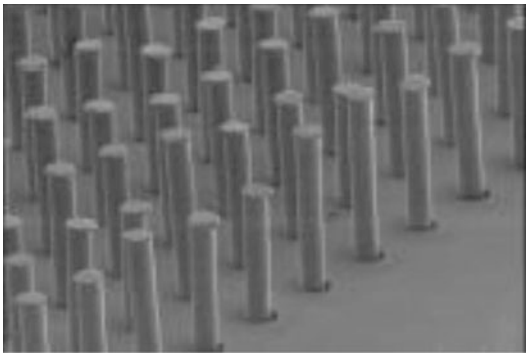
Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 20 (**a**) Assembly process: (*i*) Walking mode [3] component delivery performed on assembly stack in Configuration 1; (*ii*) a single component is captured by template 1 near the binding location; (*iii*) Configuration 2 – component is allowed to drop into template 2; (*iv*) slightly agitating the system aligns the component to the orientation of the aperture on top of the binding location on the target substrate. After step (*iv*), solder reflow is performed to bond the component to the target substrate mechanically and electrically. Assembly of SMT passive components onto a circular test device: (**b**) before assembly; (**c**) after assembly (Reprinted with permission from [16], © 2009 IEEE)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 21 Assembly process (**a**): (*i*) Host substrate aligned to the master magnetic array, (*ii*) chip seeding, (*iii*) applying vibration, and (*iv*) chip final alignment and master magnetic array removal. Photo of 1 mm^2 assembly (**b**) 2 mins and (**c**) 5 mins into the vibration process (Reprinted with permission from [17], © 2007 American Institute of Physics)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 22 Nickel wire assembly process: (**a**) nickel wires are introduced to a substrate etched with corresponding holes, (**b**) a magnet is placed under the substrate, causing the nickel wires to stand upright due to magnetic forces, (**c**) the magnet is moved laterally under the substrate, moving wires to unoccupied holes and trapping them, (**d**) BCB is applied over the assembled wires. Processing steps after (**d**) include curing the BCB, polishing the substrate to expose the ends of the nickel wires, and depositing metal lines for electrical connection (Reprinted with permission from [18], © 2011 IEEE)



Self-Assembly for Heterogeneous Integration of Microsystems, Fig. 23 SEM image of a 30 × 30 array with a pitch of 120 μm of nickel wires placed in via holes prior to the filling of BCB (Reprinted with permission from [18], © 2011 IEEE)

introduced. By identifying self-assembly techniques with suitable discrete component dimensions as well as carefully considering the transport and alignment mechanism and assembly environment, one can harness the parallel and scalable nature of self-assembly, thereby streamlining the existing assembly processes or enabling revolutionary pathways.

References

1. Morris, C.J., Stauth, S.A., Parviz, B.A.: Self-assembly for microscale and nanoscale packaging: steps toward self-packaging. IEEE Trans. Adv. Packag. **28**, 600–611 (2005)
2. Fearing, R.S.: Survey of sticking effects for micro parts handling. Presented at the proceedings of the international conference on intelligent robots and systems, vol. 2, 1995
3. Whitesides, G.M., Grzybowski, B.: Self-assembly at all scales. Science **295**, 2418–2421 (2002)
4. Rothemund, P.W.K.: Folding DNA to create nanoscale shapes and patterns. Nature **440**, 297–302 (2006)
5. Yeh, H.J.J., Smith, J.S.: Fluidic self-assembly for the integration of GaAs light-emitting diodes on Si substrates. IEEE Photonics Technol. Lett. **6**, 706–708 (1994)
6. Alien Technology. Available: http://www.alientechnology.com (2011). Accessed 1 Mar 2011
7. Saeedi, E., Kim, S., Parviz, B.A.: Self-assembled crystalline semiconductor optoelectronics on glass and plastic. J. Micromech. Microeng. **18**, 7 (2008)
8. Saeedi, E., Etzkorn, J.R., Parviz, B.A.: Sequential self-assembly of micron-scale components with light. J. Mater. Res. **26**, 1 (2011)
9. Fukushima, T., Konno, T., Tanaka, T., Koyanagi, M.: Multichip self-assembly technique on flexible polymeric substrate. In: 58th Electronic Components and Technology Conference 2008 (ECTC 2008), pp. 1532–1537 (2008)
10. Srinivasan, U., Liepmann, D., Howe, R.T.: Microstructure to substrate self-assembly using capillary forces. J. Microelectromech. Syst. **10**, 17–24 (2001)
11. Jacobs, H.O., Tao, A.R., Schwartz, A., Gracias, D.H., Whitesides, G.M.: Fabrication of a cylindrical display by patterned assembly. Science **296**, 323–325 (2002)
12. Knuesel, R.J., Jacobs, H.O.: Self-assembly of microscopic chiplets at a liquid-liquid–solid interface forming a flexible segmented monocrystalline solar cell. Proc. Natl. Acad. Sci. U. S. A. **107**, 993–998 (2010)
13. Park, K.S., Xiong, X., Baskaran, R., Böhringer, K.F.: Fluidic self-assembly of millimeter scale thin parts on preprogrammed substrate at air-water interface. In: IEEE International Conference on Micro Electro Mechanical Systems (MEMS), pp. 504–507 (2010)
14. Fang, J., Böhringer, K.F.: Parallel micro component-to-substrate assembly with controlled poses and high surface coverage. J. Micromech. Microeng. **16**, 721–730 (2006)

15. Hoo, J., Baskaran, R., Böhringer, KF.: Programmable batch assembly of microparts with 100 % yield. In: Transducers, Denver, pp. 829–832 (2009)
16. Hoo, J., Lingley, A., Baskaran, R., Xiong, X., Böhringer, K.F.: Parallel assembly of 01005 surface mount technology components with 100 % yield. In: IEEE International Conference on Micro Electro Mechanical Systems (MEMS), pp. 532–535 (2010)
17. Ramadan, Q., Uk, Y.S., Vaidyanathan, K.: Large scale microcomponents assembly using an external magnetic array. Appl. Phys. Lett. **90**, 172502/1–172502/3 (2007)
18. Fischer, A.C., Roxhed, N., Haraldsson, T., Heinig, N., Stemme, G., Niklaus, F.: Fabrication of high aspect ratio through silicon vias (TSVs) by magnetic assembly of nickel wires. In: IEEE 24st International Conference on Micro Electro Mechanical Systems 2011(MEMS 2011), Cancun (2011)

Self-Assembly of Nanostructures

Wei Lu
Department of Mechanical Engineering,
University of Michigan, Ann Arbor, MI, USA

Synonyms

Self-assembled nanostructures; Self-organized nanostructures

Definition

Self-assembly of nanostructures is a process where atoms, molecules or nanoscale building blocks spontaneously organize into ordered structures or patterns with nanometer features without any human intervention. It is the most promising practical low-cost and high-throughput approach for nanofabrication.

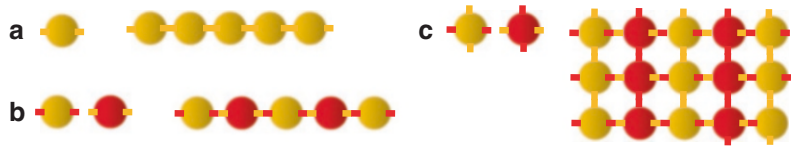
The Concept of Self-Assembly

The term "self-assembly" can be understood from its two components. The first component "self" implies "spontaneous and on its own," which suggests that it is a process that happens without human intervention or operation from outside of the system. The second component "assembly" indicates "forming or putting together," which suggests that the result is a structure built up by lower level building blocks or parts. Self-assembly of nanostructures refers to structures or patterns with nanometer features that form spontaneously from the basic building blocks such as atoms, molecules, or nanoparticles. Cells and living organisms are perfect examples of fairly sophisticated structures self-assembled in nature. These functional assemblies have motivated the study and design of nonliving systems.

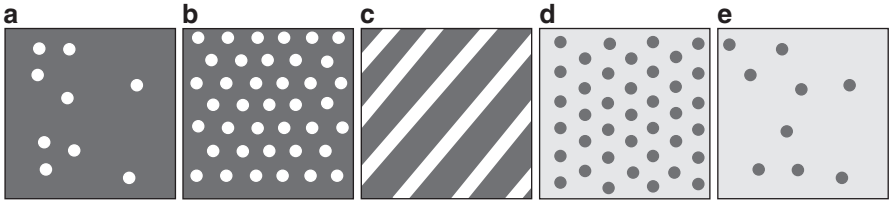
From a scientific point of view, the spontaneous formation of ordered nanostructures from a random disordered state is inherently an intriguing process in any system. The fundamental mechanisms behind these behaviors raise significant research interests. In addition, studying self-assembly in nonliving systems may provide the critical understanding toward decoding the living systems, which is far more complicated. From an engineering point of view, self-assembly as the fundamental principle of a bottom-up approach is possibly the most promising technique for nanofabrication. Thus self-assembly plays a central role in the broad nanotechnology field. The self-assembly of nanostructures may enable a wide range of applications [1], such as nanoelectronic devices, ultrasensitive biosensors, carriers for drug delivery, high capacity lithium batteries, efficient photovoltaic devices, and advanced materials with unique mechanical, electrical, magnetic, or photonic properties.

Discrete System Versus Continuum System

Depending on the type of building blocks, self-assembled systems can be classified into two categories: discrete and continuum. A discrete system uses prefabricated building blocks with fixed sizes and shapes. Figure 1 shows an example, where the building blocks are spherical nanoparticles with engineered bonding sites through surface treatment. When put in a liquid, these nanoparticles collide randomly due to the



Self-Assembly of Nanostructures, Fig. 1 An example shows the self-assembly of a discrete system



Self-Assembly of Nanostructures, Fig. 2 An example shows the self-assembly of a continuum system. The domain patterns evolve from *dots* to *parallel stripes*, and to *inverted dots* with the increasing coverage of the bright phase

Brownian motion. When two particles are close to each other with matching orientations so that their bonding sites meet, a bond between them forms. A chain structure as shown in Fig. 1a will appear when the system has a single type of nanoparticles. Figure 1b shows a system composed of two types of nanoparticles. A yellow particle has two red bonding sites, indicating that it will bond to two red particles. A red particle has two yellow bonding sites, indicating that it will bond to two yellow particles. This selective bonding leads to a self-assembled chain of alternating particles. Figure 1c shows a case where each particle has four bonding sites. They self-assemble into a network of alternating particle chains. Similar approach can be used to construct three-dimensional structures as well. Various structures can self-assemble by engineering the bonding properties.

In addition to local bonding properties, electric and magnetic fields as well as shear forces and spatial constraints have been used to direct the assembly of nanoparticles and nanorods into different configurations [2]. The electric and magnetic field method induce a dipole-type long-range interaction to act as the driving force to bring nanoparticles together, while the shear force method uses hydrodynamic interactions. The assembly of nanowire arrays is more challenging than that of nanoparticles and nanorods due to their highly anisotropic shape. Nanowire self-assembly usually results in partially ordered, small superlattices. This issue has been addressed by several new methods to direct the process, including the use of microfluidic channels and electric fields [3, 4].

In contrast, a continuum system exploits the spontaneous formation of nanoscale domains. Examples include self-assembled domain patterns in binary monolayers [5], block copolymers [6], and organic molecular adsorbates on metal surfaces [7, 8]. A binary monolayer on an elastic substrate may separate into two phases, and self-assemble into ordered patterns, such as triangular lattice of dots, parallel stripes, or serpentine stripes. The feature size is on the order of 1–100 nm, and stable against annealing. Block copolymers are polymers consist of at least two chemically distinct, immiscible polymer fragments that are joined by a covalent bond. These systems have been shown to develop a variety of regular domain patterns via phase separation [9, 10]. The size and the period of the structures are typically on the order of 10–100 nm, depending on the conditions of preparation and the relative chain lengths of the participating polymers.

Figure 2 shows an example of domain patterns formed by two phases. A notable feature is the dependence of the pattern on the average concentration. Take Cu and Pb monolayer on Cu (111) substrate as an example. The Pb and Cu mixture monolayer on a Cu substrate forms

two phases: a Pb overlayer and a disordered Pb–Cu surface solution. The two phases in the monolayer self-assemble into ordered, nanoscale patterns. When the average concentration of Pb atoms increases, the system can experience a series of patterns from (a) to (e), where the bright phase is Pb while the dark phase is the Pb–Cu surface alloy. When the average concentration is very low, isolated dots are obtained, as shown in Fig. 2a. The bright phase is embedded in the continuous matrix of the dark phase. No long-range ordering is observed. When the average concentration increases, the area covered by the bright dots increases. The dots order and form patterns close to a triangular lattice, as shown in Fig. 2b. When the bright phase occupies roughly half of the surface, stripe structure as shown in Fig. 2c forms. "Inverted" dots are observed when the bright phase dominates, as shown in Fig. 2e.

A continuum system offers several unique features. For instance, domains and their patterns self-assemble simultaneously, so that there is no need to pre-synthesize the building blocks; a significant degree of process flexibility and control can be achieved; and the approach may be applied to diverse systems.

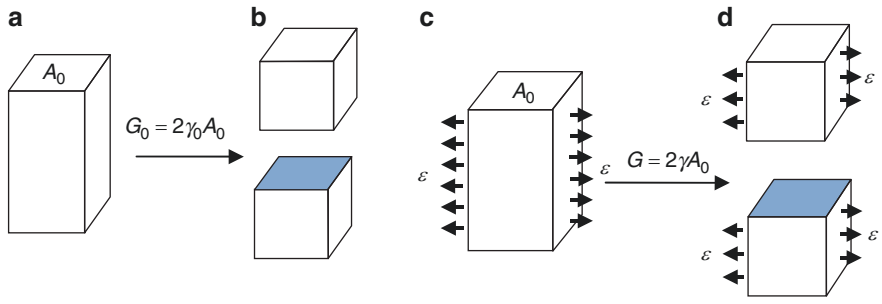
Static Self-Assembly Versus Dynamic Self-Assembly

Depending on the nature of interactions, self-assembly systems can be classified into two categories: static and dynamic. Static self-assembly is a process driven by energy minimization to form static structures which are at global or local equilibrium. Many examples fall into this category, such as colloidal crystals, microphase separation, or spontaneous domain patterns. Dynamic self-assembly involves interactions that only occur when the system is dissipating energy. The structure is determined by an energetic minimum relying on an influx of energy to the system. When the energy flux stops, the minimum configuration does not exist anymore, leading to disintegration of the structure. For instance, rotating nanodisks in a liquid can generate hydrodynamic interactions due to the flow of liquid. When the disks are close to each other, the hydrodynamic interaction causes them to form a nicely ordered hexagonal pattern with a regular inter-disk spacing. When the rotation stops, the interaction disappears and the pattern vanishes. Another example is the oscillating reaction–diffusion reaction, where an evolving pattern develops from a quiescent medium upon the influence of stimuli. A living organism is possibly the most typical example of dynamic self-assembly.

Compared to static self-assembly, dynamic self-assembly is even less understood [11]. This process involves interacting building blocks that can adapt or react to a chemical or physical stimulus from the environment. Unlike the forces that drive static self-assembly through a reduction of the free energy toward equilibrium, the interactions in dynamic self-assembly may drive the formation of structures and patterns away from the thermodynamic equilibrium sustained by a continuous energy input. Such patterns can thus be sensitive to external stimuli and adaptive in response to the surrounding conditions [12].

Forces That Drive Self-Assembly

Forces of different physical origins may contribute to the interactions between the building blocks. The short-range interaction is generally controlled by forces such as ionic, covalent, or metallic bonding; hydrogen bonding; and van der Waals forces. The long-range interaction has an effective range much longer than the atomic distance, and can originate from elasticity, electrostatic or magnetic field, colloidal and capillary forces, or hydrodynamic interaction. The short-range interaction alone typically leads to either a homogenous structure as shown in Fig. 1a, or a structure with only local modulation whose wavelength is on the order of the size of the building blocks, such as those in Fig. 1b, c. The long-range interaction is essential to form characteristic feature sizes larger than the dimension of the building blocks. Thus long-range interaction is especially important for the self-assembly in a continuum system to form nanoscale feature size, which is much larger than the building block of a single atom or molecule.



Self-Assembly of Nanostructures, Fig. 3 Schematic representation illustrates the concept of surface energy and surface stress

In a discrete system of nanoscale building blocks, the feature size is determined by the competition between the attractive and repulsive forces. Thus force balance analysis is a useful approach. While such a concept may still be applied to a continuum system, the connection between the domain size and their interaction is much less direct. Therefore the energy method is usually used to analyze a continuum system. The feature size is determined by two competing actions: a coarsening action which tends to increase the domain size and a refining action which tends to reduce the domain size. The following discusses a few representative forces and their roles in self-assembly.

Surface Stress

Surface stress plays crucial roles in a variety of surface phenomena in solids. To explain the idea, consider a uniform and infinitely large solid shown in Fig. 3a. The body is unstrained and is taken as the reference state. In Fig. 3b, the body is cut into two parts, and the atoms on the surfaces are allowed to find their equilibrium configuration. The energy in state (b) is higher than that in state (a). The energy difference between the two states, G_0, can be written as $G_0 = 2\gamma_0 A_0$, where γ_0 is the excess free energy per unit area owing to the existence of the surface, and A_0 is the surface area. G_0 is called surface energy. γ_0 is usually called surface energy density, which is the reversible work per unit area to create a surface. Figure 3c is the same solid in (a) but under strain ε. In Fig. 3d, the strained body is cut into two parts. The surface energy can be written as $G = 2\gamma A_0$. The two energies, G and G_0, are different. The former depends on the strain, i.e., $G = G(\varepsilon)$. Accordingly, the surface energy density also depends on the strain state, $\gamma = \gamma(\varepsilon)$. Hence the surface energy depends on both the created surface area and the strain state, namely, $d(\gamma A_0) = A_0 f d\varepsilon$, where $f = \partial\gamma/\partial\varepsilon$ is called surface stress. Surface stress becomes a tensor when generalized to biaxial strains.

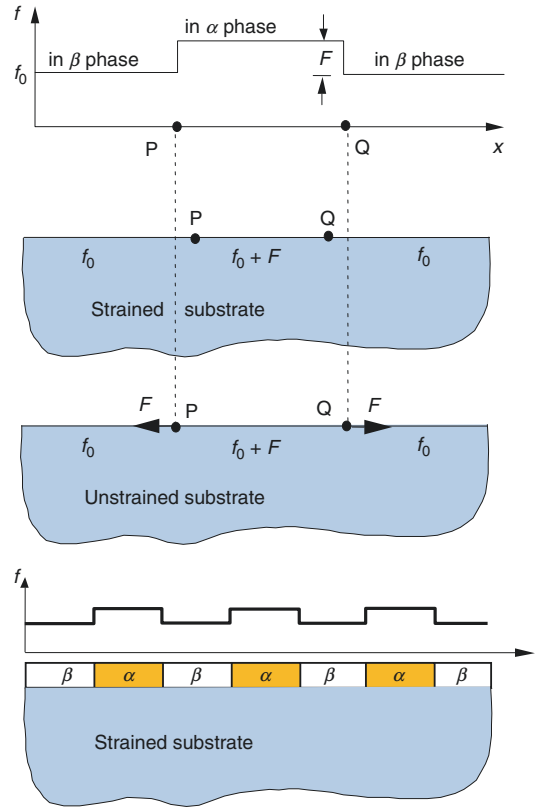
The physical origin of surface stress is related to the difference of bonding between atoms at the surface and the interior atoms. A typical approach to measure surface stress is the cantilever bending method. Upon deposition of atoms on the top surface, surface stress causes the cantilever to bend. By measuring the curvature and some geometric parameters, such as the thickness of the cantilever, one can calculate surface stress. The surface stress of a domain depends on its composition. For a binary epilayer, a composition modulation will cause surface stress nonuniformity.

Consider a thin binary layer that grows epitaxially on a solid substrate with fixed thickness, say a monolayer. The two species can relocate by diffusion within the layer. Several physical ingredients contribute to the spontaneous domain formation behavior. Phase separation is a commonly observed phenomenon in various material systems, which can be explained in terms of the free energy of mixing. Consider a mixture of two atomic species A and B, and define concentration

by the fraction of one species. The free energy density is given by a double well function of the concentration. A tangent line contacts the function at two concentrations, corresponding to the equilibrium A-rich and B-rich phases. A mixture with an average concentration between these two concentrations will separate into the two phases to reduce the energy. The atoms at the phase boundaries have excess free energy. To reduce the free energy, the total area of the phase boundaries must reduce. Consequently, atoms leave small domains, diffuse in the matrix, and join large domains. Over time the small domains disappear, and the large ones become larger. Thus phase boundary energy causes phase coarsening, which cannot form stable patterns. The observed stable periodic patterns suggest that there must be a refining action in the regular nanoscale structures. This action is provided by the surface stress.

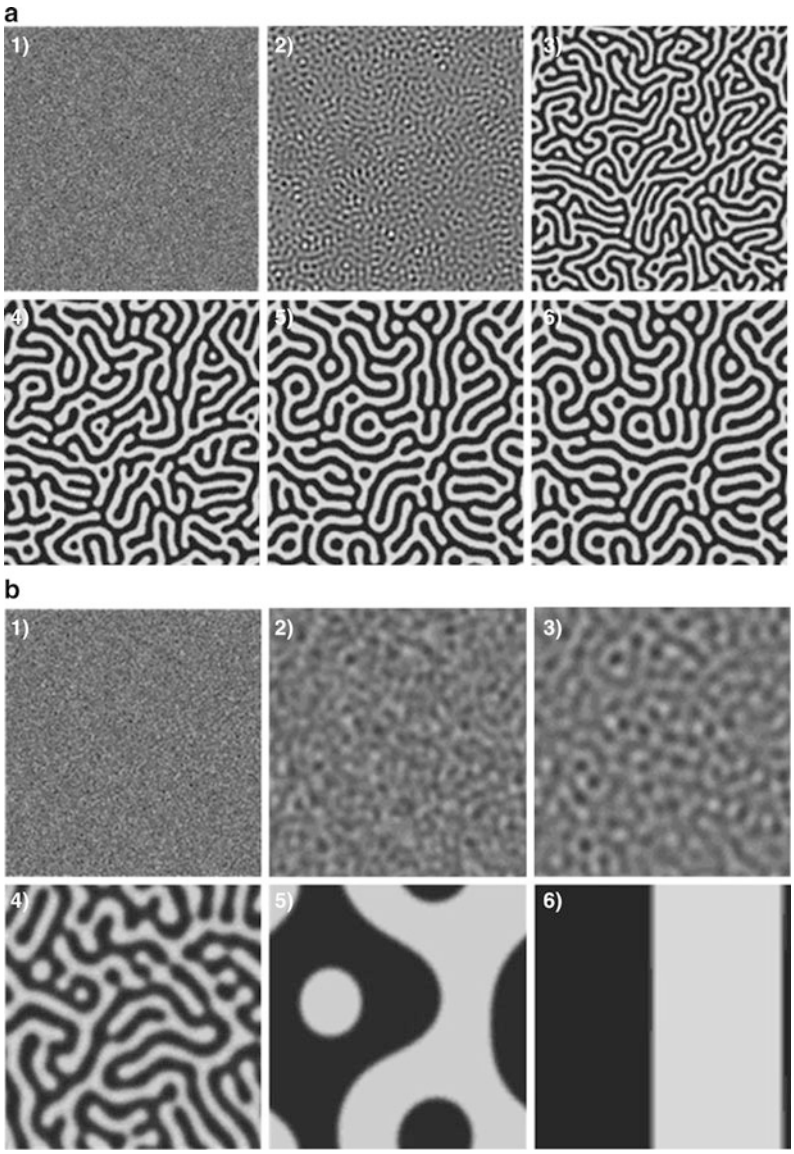
Figure 4 shows an important concept: substrate deformation allows the nonuniform surface stress to reduce the total energy. The surface stress is f_0 everywhere except for the portion of the surface between points P and Q, where the surface stress steps up by F. The nonuniformity causes P and Q to move toward each other, giving rise to a strain field in the substrate. Imagine that a pair of forces are applied to move the points P and Q away from each other. When the applied forces reach the magnitude F, the substrate becomes unstrained. Because the forces do work to the body, the unstrained state has a higher energy than the strained state. The energy change depends only on F, but not on f_0. Whether the overall surface stress is tensile, compressive, or vanishing does not make any difference to the energy reduction. When the domain size is refined, more pairs of F will contribute so that the energy is reduced further. Consequently, the elastic energy in the film-substrate composite tends to refine the domains. The refinement, however, adds more domain boundary, which increases the phase boundary energy. The phase boundary energy tends to coarsen the domains. The two competing actions – refining and coarsening – can select an equilibrium domain size.

These physical ingredients have been incorporated into an energy framework to simulate



Self-Assembly of Nanostructures, Fig. 4 Substrate deformation allows the nonuniform surface stress to reduce the energy, and surface stress provides a refining action

various nanostructures [13]. Figure 5 clearly shows the refining effect of the surface stress. The average concentration is 0.5. Each phase takes half of the surface area. In sequence (a), shortly after phase separation, the two phases form serpentine stripes. The width of the stripes stabilizes very fast. From $t = 1{,}000$ to $t = 10^6$, the widths are almost invariant. In contrast, in sequence (b), the two phases try to increase their sizes as much as possible. The system finally evolves into a state that one phase takes half of the calculation cell and the other phase takes the rest half. This reproduces the classical spinodal decomposition. The pattern depends on the average concentration. Changing the average concentration away from 0.5 can generate dots instead of stripes. Anisotropic surface stress induces interesting patterns such as parallel stripes and herringbone structures [14].

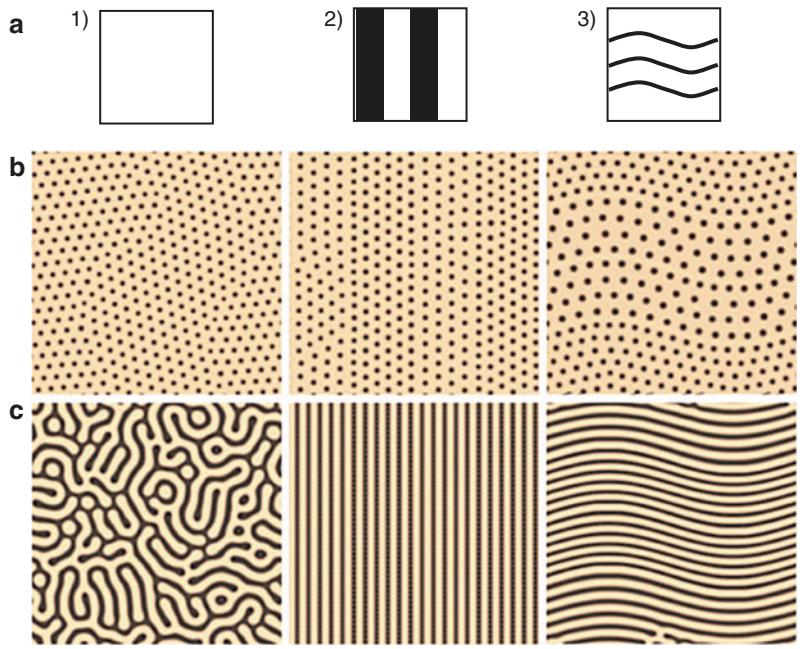


Self-Assembly of Nanostructures, Fig. 5 Simulation starts with a random initial condition with an average concentration of 0.5. (**a**) The phases form nanoscale serpentine stripes with surface stress. (**b**) Spinodal decomposition without surface stress. The phases always coarsen. The times are (*1*) $t = 0$, (*2*) $t = 10$, (*3*) $t = 100$, (*4*) $t = 1{,}000$, (*5*) $t = 10^5$, (*6*) $t = 10^6$ (Adapted from Ref. [14])

Guided or Templated Self-Assembly

Guided or templated self-assembly uses a coarse scale external field to influence the fine scale self-assembly behavior. An essence of the approach is that the wavelength of the guiding field or template can be much larger than the nanoscale structures to be formed. Consequently, preparing the coarse scale field involves low cost and can be easily applied to a very large area. The guiding field can be electrostatic, geometric pre-patterns, surface chemistry, or elastic field [15].

Figure 6 shows an example to guide the surface patterns with strain field on the substrate surface. Surface strain is applied in the black region in Fig. 6a. Three cases are shown: (1) no guidance, (2) guiding strain in wide stripes, and (3) guiding strain in wavy stripes. In practice, there are many ways to induce an elastic field on the substrate surface. In addition to direct mechanical loading, pre-patterning a substrate with different materials by photolithography or applying an electric field to a substrate embedded with piezoelectric particles produce diverse well-defined strain fields.



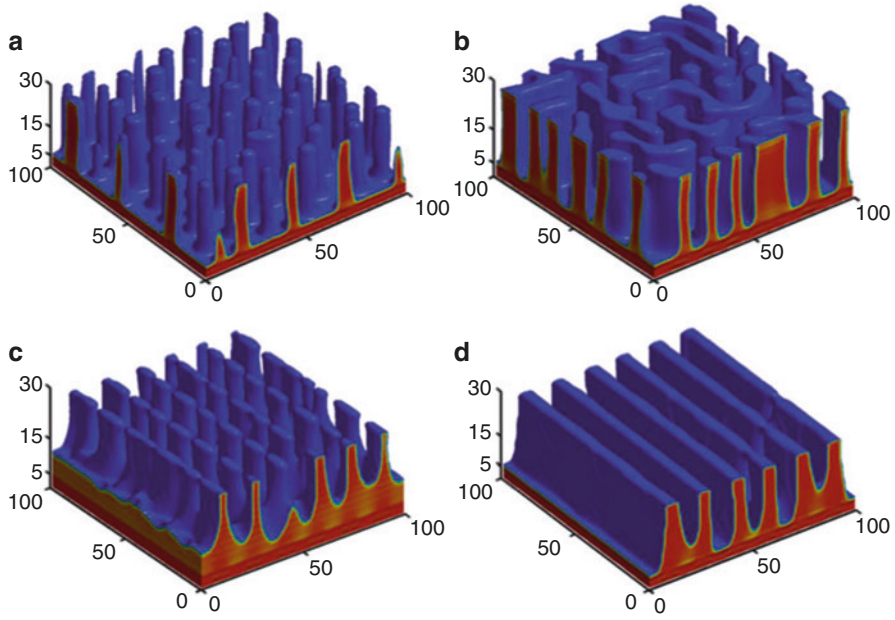
Self-Assembly of Nanostructures, Fig. 6 Guided self-assembly with strain field on the substrate surface. (**a**) A constant strain is applied in the *black region* to guide the self-assembly on the substrate surface. (**b**) Patterns for an average concentration of 0.3. (**c**) Patterns for an average concentration of 0.5 (Adapted from Ref. [15])

Figure 6b shows patterns for an average concentration of 0.3. The monolayer separates into two phases and evolves into a multi-domained triangular lattice of dots when there is no guidance. The middle column shows that in case [2], the monolayer evolves into pattern colonies. The dots form an almost perfect triangular lattice within each colony, and line up with the close packed direction parallel to the edge of the stripes. The lattices in the white and black stripe areas are self-contained, maintaining their own lattice spacings. The right column shows that in case [3], the dots orientate themselves along the wavy edges of the three curves. The narrow curves preclude the formation of any dots inside. Similar guiding effect is also shown for an average concentration of 0.5 in Fig. 6c. These results suggest that external loading can effectively change size, shape, and orientation of self-assembled surface structures. Similar guided self-assembly has also been achieved by surface chemistry [16].

Figure 7 shows another example where electric field is used to guide self-assembly. A three-dimensional model has been developed to allow the simulation of the entire self-assembly process and electric field design [17]. It is shown that a melted thin polymer film subjected to an electrostatic field may lose stability at the polymer-air interface, leading to uniform self-organized pillars emerging out of the film surface. The film thickness can affect the patterns formed. With patterned electrodes, parallel stripes replicating an electrode pattern emerge.

Electric Dipoles

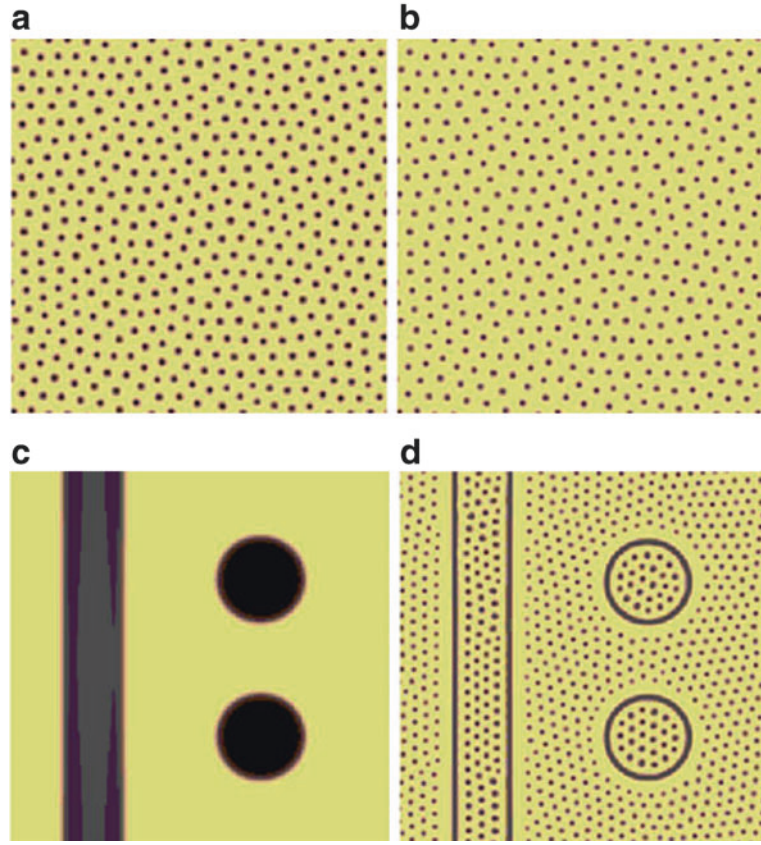
An adsorbate molecule usually carries an electric dipole. Even if the molecules are nonpolar, the act of binding onto a substrate breaks the symmetry and causes the formation of dipoles. A molecule can be engineered to carry a large electric dipole moment by incorporating a polar group. Electric dipole interaction can give rise to domain patterns. Examples include Langmuir films at the air–water interface, ferrofluids in magnetic fields, and organic molecules on metal surface. Despite the difference of these systems, similar phenomenology and mechanism can be identified. The adsorbed molecules are mobile. Domains coarsen to reduce the domain boundary and refine to reduce the dipole interaction energy. The competition leads to equilibrium patterns. This mechanism may be used to make two-dimensional



Self-Assembly of Nanostructures, Fig. 7 Self-assembled morphological patterns of a melted thin polymer film induced by two electrodes parallel to the initial flat film surface, one above the film and the other beneath the film. (**a**) Induced nanopillars in a thin film. (**b**) A relatively thick film. The film continues to evolve after touching the ceiling electrode and then spread. (**c**) Patterned top electrode causes the nanopillars to line up. (**d**) Replication of the top electrode pattern (Adapted from Ref. [17])

nanostructures. Specifically, molecule monolayers composed of electric dipoles can be manipulated with an electric field induced by an AFM tip, a ceiling above the layer, or an electrode array in the substrate.

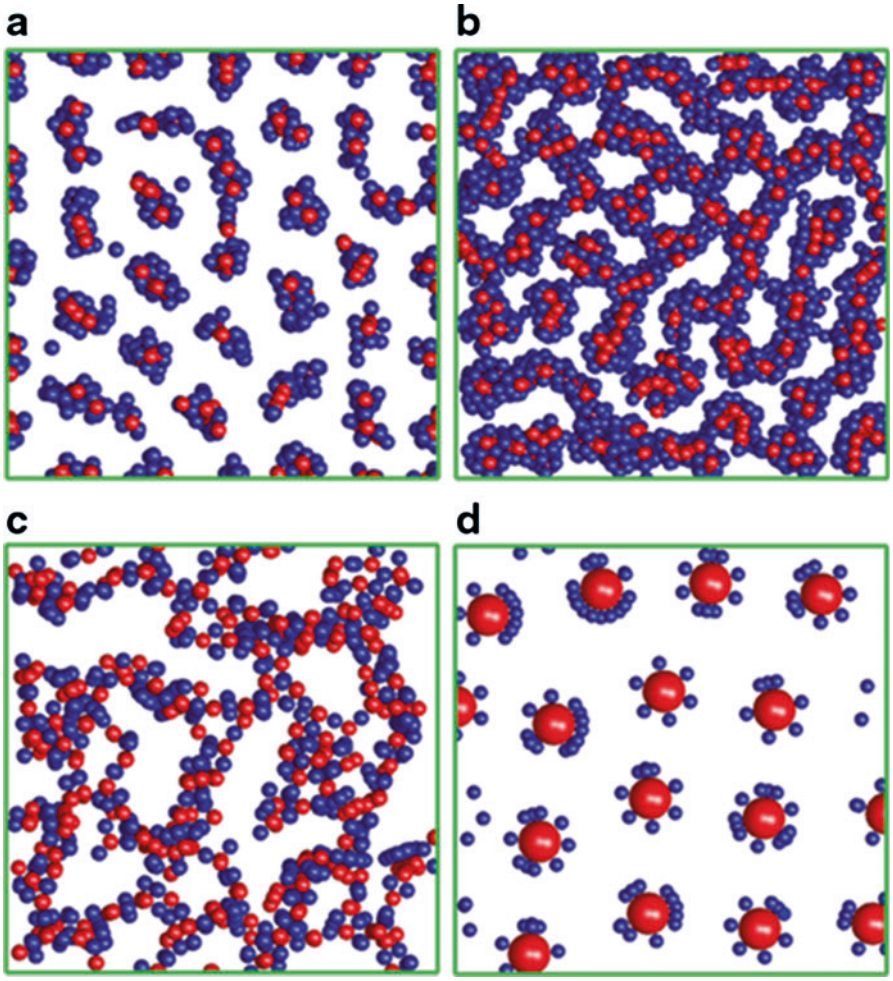
Multilayers of dipole molecules have the capability for making complex structures, especially the formation of nanointerfaces and three-dimensional nanocomposites. Studies have revealed self-alignment between layers, reduction of feature sizes, and guided self-assembly by layer–layer interaction and embedded electrodes [18]. Figure 8 gives an example. Each layer is composed of two kinds of molecules with different dipole moments, which form two domains. Figure 8a shows that the first layer self-assembles into a triangular lattice of dots at an average concentration of 0.3. The dots have uniform size and form multiple grains. Figure 8b shows the second layer pattern at an average concentration of 0.2 and grows on top of the pattern in Fig. 8a. The second layer evolves from a completely different random initial condition. The dots of the second layer stay at exactly the same positions as those in Fig. 8b, suggesting the anchoring effect of the first layer. The dot size of the second layer is smaller due to lower average concentration. The observations suggest that the first layer determines the ordering and lattice spacing, while the second layer determines the feature size. A scaling down of size can be achieved via multilayers. The interesting behavior suggests a potential fabrication method. In additional to self-assembly, the first layer pattern can be defined by embedded electrodes, proximal probe technique, or nanoimprinting. Figure 8c shows the first layer pattern with the application of a high voltage, which sweeps off any self-assembled features so that the monolayer replicates the voltage pattern. The second layer evolves into a pattern shown in Fig. 8d. The dots align themselves along the edges of the first layer pattern, and form triangular lattice. It is interesting to note the formation of pairing dark and white lines following the contour of the first layer pattern. Two nearby regions separated by the lines have different preference for



Self-Assembly of Nanostructures, Fig. 8 Self-assembled patterns by electric dipole interaction. (**a**) First layer, average concentration 0.3. (**b**) Second layer on *top* of (**a**). (**c**) First layer pattern replicating the electrode pattern. (**d**) Second layer on *top* of (**c**) (Adapted from Ref. [18])

the two dipole species. This causes local accumulation and depletion of the two dipoles, resulting in the observed phenomena.

Another example of electric dipole interaction is the self-assembly of nanoparticles [19]. Under an applied electric field, each particle will acquire an effective induced dipole. The dipole interaction causes particles to move in a liquid medium and line up into chains following the electric field direction. The formation of an ordered three-dimensional structure is further affected by the interaction between the chains. The self-assembly of binary nanoparticles can lead to a wide class of nanocomposite materials with properties not attainable by a single particle component. Figure 9a shows a top view of the structure. Behind each visible particle shown in the figure is a particle chain lining up. The two kinds of particles are denoted by red and blue colors. They have the same diameter and are both more polarizable than the medium, while the red particle is more polarizable than the blue one. The volume fraction of the particles is 10.2 %. Under these situations the particles assemble into isolated columns with a core-shell configuration. The core is composed of the more polarizable red particles, while the shell is composed of the blue ones. From an energetic point of view, the attraction between two red particle chains is stronger than the interaction between a red particle chain and a blue particle chain. The latter is strong than the attraction between two blue particle chains. As a result, the red chains aggregate to form columns. The blue chains tend to get close to the red columns as much as possible, leading to the formation of shells. It is exciting to note that this self-organized functionally gradient structure offers a gradual transition of the permittivity from that of the core to that of the medium. This approach could be potentially very useful to construct functional gradient nanocomposites from bottom up.



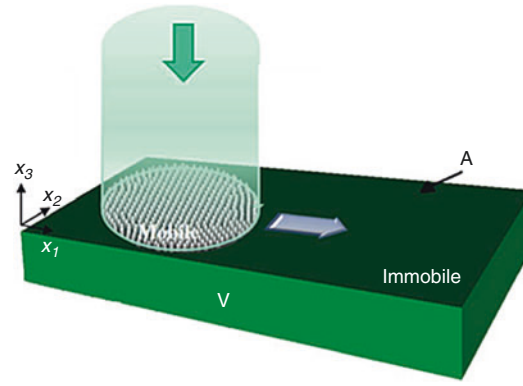
Self-Assembly of Nanostructures, Fig. 9 *Top* view of structures self-assembled in a system of binary nanoparticles. (**a**) Gradient core-shell structure. (**b**) A continuous structure with isolated holes in between. (**c**) Network structure. (**d**) A single layer (Adapted from Ref. [19])

Figure 9b shows a higher volume fraction of particles, 30.6 %. The system evolves into a continuous structure with isolated holes in between. The blue particles enclose the red particles, and form the peripheral regions around the holes. The particle columns demonstrate local BCT structures. Another common feature is that same kind particles tend to aggregate. Figure 9c demonstrates the structure when one kind of particle is more polarizable than the medium while the other is less polarizable. The particles also form pure chains along the field direction with each chain composed of single kind particles. A distinct feature in Fig. 9c is that the red and blue chains are highly dispersed and form a network. This morphology is in contrast to that in Fig. 9a, where same color chains aggregate.

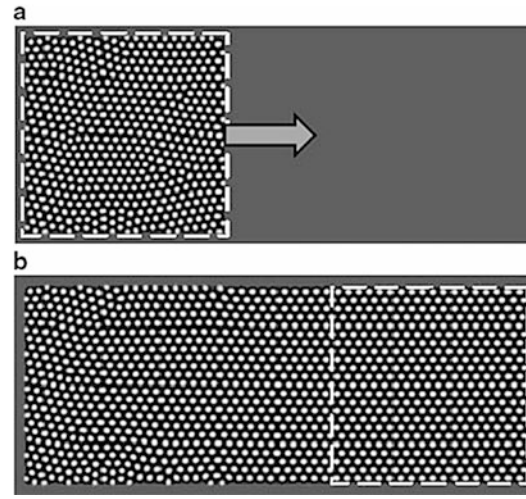
Assembling a single layer of nanoparticles on a substrate has many potential applications. A relevant simulation is shown in Fig. 9d for a layer of two kinds of particles. The repulsion between the red particles causes them to form a nicely ordered triangular lattice. The attraction between the blue and red particles causes the formation of blue rings surrounding the red cores.

Sequential Activation of Self-Assembly

The lack of long-range order has been a major challenge for self-assembled nanostructures since high regularity is crucial to many applications. A general kinetics-based template-free approach has been developed recently to grow nanostructured superlattices over a very large area [20]. The essence is sequential activation of self-assembly. The idea is illustrated in Fig. 10. A binary monolayer is used as an example. Typically multiple grains of dots will form due to simultaneous self-assembly at different locations, producing a pattern lack of long-range order. Here self-assembly is first activated in a finite mobile region, where atoms are allowed to diffuse and form domain patterns. This initial mobile region will serve as a "seed." The seed does not need to have a perfect lattice. Then the mobile region is shifted like scanning. The self-assembly in the newly activated region will be influenced by the pattern already formed in the seed. In experiments this process can be achieved by laser or ion beam scanning to control the local temperature, so that diffusion is activated sequentially at each spot along the scanning path. This sequential activation will lead to a large long-range ordered domain pattern even if started with an imperfect seed. The pattern quickly improves and converges to a perfect superlattice along with the sequential activation.

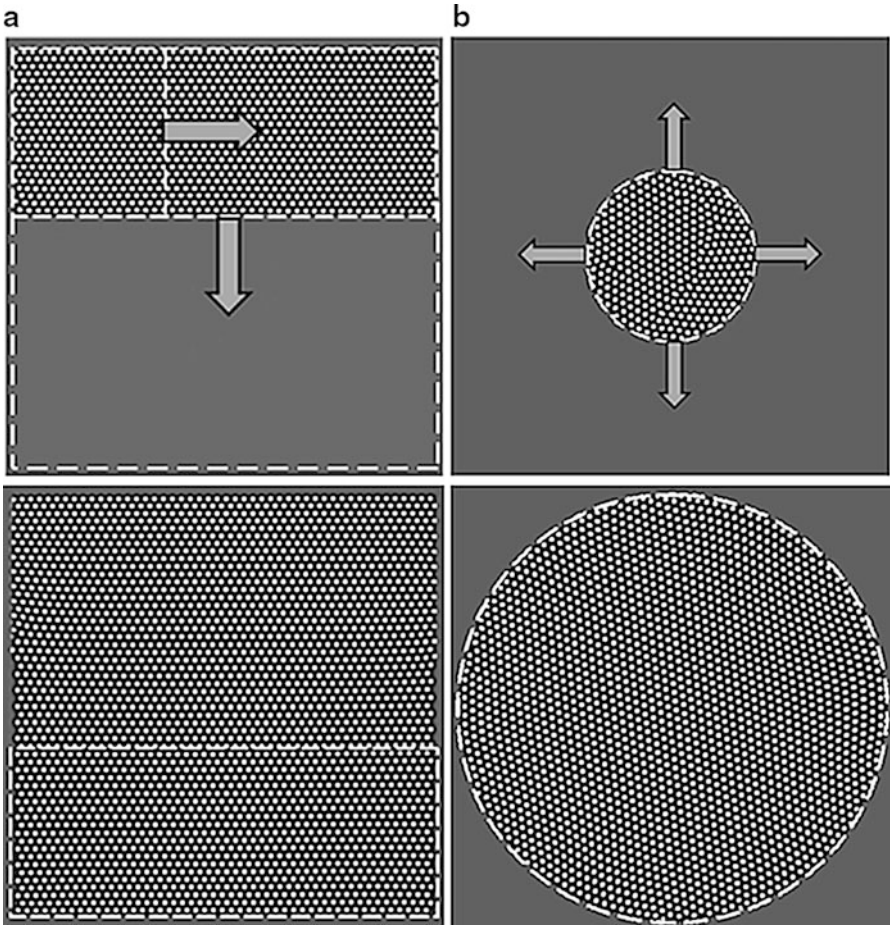


Self-Assembly of Nanostructures, Fig. 10 A schematic of sequential activation of self-assembly. Self-assembly is first activated in a finite mobile region, where atoms are allowed to diffuse and form domain patterns. This initial mobile region serves as a "seed." The mobile region is then shifted like scanning. Self-assembly in the newly activated region will be under influence of patterns already formed in the seed (Adapted from Ref. [20])



Self-Assembly of Nanostructures, Fig. 11 Growth of superlattice from seeds. (**a**) A square seed. (**b**) A band of nicely ordered hexagonal superlattice formed after scanning over the width. The lattice improved to perfect along with the scanning, demonstrating tolerance of defects in the seeds (Adapted from Ref. [20])

Figure 11 demonstrates the growth of long-range ordered superlattices from a seed. Rather than using a perfect lattice directly, the seed is grown on site. Take Fig. 11a as an example. The constraint of kinetics leads to a square seed composed of dots. The seed is larger than the size of a typical single grain. Therefore defects such as misalignment and multiple grains appeared in this seed. Next the mobile region is shifted to the right. This process is called scanning. The shift distance in each step is much smaller compared to the seed size, which creates a continuous scanning effect. The scanning velocity is chosen to ensure ample time for new domains to develop. Figure 11b shows the structure after scanning over the width of the calculation cell, which forms a band of nicely ordered hexagonal superlattice. Noticeably, the lattice improves to perfection along with the scanning, demonstrating the tolerance of defects in the seeds. High output is essential to nanostructure applications.



Self-Assembly of Nanostructures, Fig. 12 Two schemes for scaling-up growth. (**a**) Alternate the scanning directions, using the superlattice created in previous step as a large seed. (**b**) Increase the size of the mobile in two-dimension (Adapted from Ref. [20])

Two schemes have been proposed to facilitate large-scale fabrication. (1) As shown in Fig. 12a, the scanning directions are alternated between left-right scanning and up-down scanning, using the superlattice created in each previous step as a large seed. This scheme allows the growth rate (area of lattice created per unit time) to increase exponentially with time, greatly accelerating large area fabrication. (2) As shown in Fig. 12b, the size of the mobile area is increased in two dimensions, rather than scanning along one direction. This scheme allows the growth rate to be quadratic of the scanning velocity.

Cross-References

- ▶ Charge Transport in Self-Assembled Monolayers
- ▶ Nanomaterials for Electrical Energy Storage Devices
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanoparticles
- ▶ Nanostructured Functionalized Surfaces
- ▶ Nanotechnology
- ▶ Self-Assembled Monolayers
- ▶ Self-Assembly
- ▶ Self-Assembly for Heterogeneous Integration of Microsystems

References

1. Lu, W., Sastry, A.M.: Self-assembly for semiconductor industry. IEEE Trans. Semicond. Manuf. **20**, 421–431 (2007)
2. Pileni, M.P.: Nanocrystal self-assemblies: fabrication and collective properties. J. Phys. Chem. B **105**, 3358–3371 (2001)
3. Huang, Y., Duan, X., Wei, Q., Lieber, C.M.: Directed assembly of one-dimensional nanostructures into functional networks. Science **26**, 630–633 (2001)
4. Smith, P.A., Nordquist, C.D., Jackson, T.N., Mayer, T. S., Martin, B.R., Mbindyo, J., Mallouk, T.E.: Electric-field assisted assembly and alignment of metallic nanowires. Appl. Phys. Lett. **77**, 1399–1401 (2000)
5. Plass, R., Last, J.A., Bartelt, N.C., Kellogg, G.L.: Nanostructures – self-assembled domain patterns. Nature **412**, 875 (2001)
6. Fasolka, M.J., Mayes, A.M.: Block copolymer thin films: physics and applications. Annu. Rev. Mater. Res. **31**, 323–355 (2001)
7. Love, J.C., Estroff, L.A., Kriebel, J.K., Nuzzo, R.G., Whitesides, G.M.: Self-assembled monolayers of thiolates on metals as a form of nanotechnology. Chem. Rev. **105**, 1103–1169 (2005)
8. Schneider, K.S., Lu, W., Owens, T.M., Fosnacht, D.R., Banaszak Holl, M.M., Orr, B.G.: Monolayer pattern evolution via substrate strain-mediated spinodal decomposition. Phys. Rev. Lett. **93**, 166104 (2004)
9. Krausch, G., Magerle, R.: Nanostructured thin films via self-assembly of block copolymers. Adv. Mater. **14**, 1579–1583 (2002)
10. Lopes, W.A., Jaeger, H.M.: Hierarchical self-assembly of metal nanostructures on diblock copolymer scaffolds. Nature **414**, 735–738 (2001)
11. Whitesides, G.M., Boncheva, M.: Beyond molecules: self-assembly of mesoscopic and macroscopic components. Proc. Natl. Acad. Sci. U. S. A. **99**, 4769–4774 (2002)
12. Ozin, G.A., Hou, K., Lotsch, B.V., Cademartiri, L., Puzzo, D.P., Scotognella, F., Ghadimi, A., Thomson, J.: Nanofabrication by self-assembly. Mater Today **12**, 12–23 (2009)
13. Lu, W., Suo, Z.: Dynamics of nanoscale pattern formation of an epitaxial monolayer. J. Mech. Phys. Solids **49**, 1937–1950 (2001)
14. Lu, W., Suo, Z.: Symmetry breaking in self-assembled monolayers on solid surfaces: anisotropic surface stress. Phys. Rev. B **65**, 85401 (2002)
15. Lu, W., Kim, D.: Engineering nanophase self-assembly with elastic field. Acta Mater. **53**, 3689–3694 (2005)
16. Lu, W., Kim, D.: Patterning nanoscale structures by surface chemistry. Nano Lett. **4**, 313–316 (2004)
17. Kim, D., Lu, W.: Three-dimensional model of electrostatically induced pattern formation in thin polymer films. Phys. Rev. B **73**, 035206 (2006)
18. Lu, W., Salac, D.: Patterning multilayers of molecules via self-organization. Phys. Rev. Lett. **94**, 146103 (2005)
19. Park, J., Lu, W.: Self-assembly of functionally gradient nanoparticle structures. Appl. Phys. Lett. **93**, 243109 (2008)
20. Zhao, Z., Lu, W.: Growing large nanostructured superlattices from a continuum medium by sequential activation of self-assembly. Phys. Rev. E **83**, 041610 (2011)

Self-Cleaning

► Lotus Effect

Self-Energy and Excitonic Calculations in Many-Body Systems

► Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches

Self-Healing Materials

► Self-Repairing Materials

Self-Organized Layers

► Self-Assembled Monolayers for Nanotribology

Self-Organized Nanostructures

► Self-Assembly of Nanostructures

Self-Regeneration

► Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components

Self-Repairing Materials

Michael Nosonovsky
Department of Mechanical Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI, USA

Synonyms

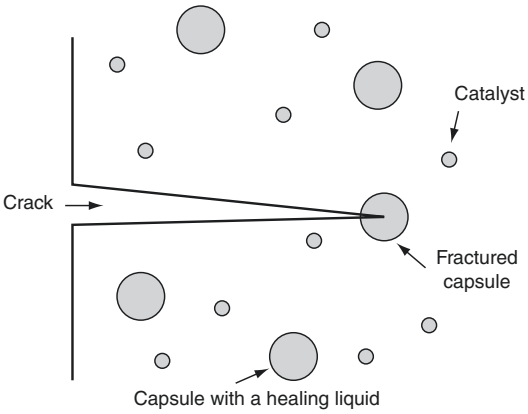
Self-healing materials

Definition

Self-repair or self-healing is the ability of a material (usually, a composite or nanocomposite) to partially or completely repair damage, such as voids and cracks, occurring during its service lifetime [1, 2]. Self-repair is achieved by embedding a nano-, micro-, or multi-scale repair mechanism into the structure of the material (Fig. 1). Autonomous self-healing refers to the ability to repair damage without any external activation. Nonautonomous self-repair refers to healing mechanisms, which require external activation, such as heating the material.

Occurrence and Key Findings

Self-healing is related to the general ability of many systems for self-organization. To facilitate self-organization, usually a special nano-, micro-, or multi-scale structure of a material is needed. Self-healing in biological objects is a source of inspiration for this new research area. Most living tissues or organisms can regenerate and heal themselves, provided the incurred damage is small or moderate. Most engineered materials, however, deteriorate with time irreversibly due to wear, brittle fracture, fatigue, creep, and other modes of degradation, which limits the life of various components and sometimes causes catastrophic damage. Biological mechanisms of healing are very complex and cannot be directly borrowed for artificial materials. Therefore, it is preferable to speak about bio-inspired, rather than biomimetic, self-healing materials.



Self-Repairing Materials, Fig. 1 Self-repair by embedding microcapsules with healing agent

Polymers

Self-healing has been applied most successfully in polymers since they have relatively high rates of diffusion and plasticity due the presence of the cross-molecular bonds. One way to create self-healing polymers is to use thermosetting polymers that have the ability to cure (toughening or hardening by cross-linking of polymer chains), such as the thermosetting epoxy. Epoxy is a polymer formed by a reaction of an epoxide resin with polyamine hardener. Epoxy can serve as a healing agent that is stored within thin-walled inert brittle macrocapsules embedded into the matrix along with a catalyst or hardener. The catalyst or hardener is also embedded in the matrix, but separately from the healing agent. When a crack propagates, the capsules fracture, the healing agent is released and propagates into the crack due to capillarity. Then the healing agent mixes with the catalyst embedded in the matrix, which triggers the cross-linking reaction and hardening of the epoxy that seals the crack [3].

A different approach involves thermoplastic polymers with various ways of incorporating the healing agent into the material. In this approach, heating is often required to initiate healing [3]. A self-repair method involving two concentric cylinders of conducting material filled with a liquid

solution containing electromagnetic particles of polystyrene or silica was suggested. When damage occurs, voltage is applied between the inner pipe and outer pipe, the current density naturally increases at the location of the damage, this increase in current density causes particle coagulation at the damage site, which in a way to heal the damage.

Ceramics

Besides the polymers, ceramic self-healing materials are being developed. For example, a concrete composite was produced with glass fibers containing an air-curing sealant embedded in the concrete matrix. This composite exhibited the self-healing behavior, but it suffered from a significant (10–40 %) loss of stiffness compared with standard concrete due to fibers. This is a typical situation, when a compromise between self-healing and mechanical properties should be sought. In another project involving self-healing ceramics, researchers have studied the crack-healing behavior and mechanical properties of a mullite composite toughened by the inclusion of 15 % (by volume) SiC whiskers. Self-healing ceramic materials often use oxidative reactions because the volume of oxide exceeds the volume of the original material, and therefore, products of these reactions can be used to fill small cracks [3]. Self-healing nanocomposites constitute another area of research.

Metallic Materials

It is much more difficult to heal metallic materials, than polymers, because metallic atoms are strongly bonded and have small volumes and low diffusion rates. Currently, there are three main directions which have been taken in the development of self-healing metallic systems. First is the formation of precipitates at the defect sites that immobilize further growth until failure. This mechanism is called sometimes "damage prevention" because it prevents the formation of voids by the diffusion of the precipitate. The atomic transport of matter to voids and defects is provided by a supersaturated solid solution in alloys having decreasing solid solubility of solute elements with decreasing temperature (e.g., Al–Cu). Such an "under-aged" alloy, when quenched from high temperature, becomes supersaturated or metastable. However, in order to facilitate precipitation of the solute, heterogeneous nucleation sites are needed. Sites with high surface energy, such as voids, defects, grain boundaries, and free surfaces, become nucleation sites. The driving mechanism for the diffusion is the excess surface energy of microscopic voids and cracks that serve as nucleation centers of the precipitate that plays the role of the healing agent. As a result, the newly formed void is sealed before it grows and thus minimizing the creep and fatigue [1, 2].

Second is reinforcement of an alloy matrix with microfibers or wires made of a shape-memory alloy (SMA), such as Nitinol (NiTi). SMA wires have the ability to recover their original shape after some deformation has occurred if they are heated above certain critical temperature. If the composite undergoes crack formation, heating the material will activate the shape recovery feature of the SMA wires which then shrink and close the cracks [1, 2].

The third approach is to use a healing agent (such as an alloy with a low-melting temperature) embedded into a metallic solder matrix, similarly to the way it is done with the polymers. However, encapsulation of a healing agent into a metallic material is much more difficult task than in the case of polymers. The healing agent should be encapsulated in microcapsules which serve as diffusion barriers and which fracture when crack propagates [1, 2].

Self-healing mechanisms in metals are summarized in Table 1, showing typical materials for the matrix and reinforcement, parameters that characterize microstructure, degradation, and healing, characteristic length scales for the degradation and healing mechanisms, the type of phase transition involved into the healing and what property is improved in the self-healing alloy, the nature of the healing force and details of healing mechanisms will be discussed in the consequent section.

Surface Healing

Surface is often the most vulnerable area of a material. Several approaches have been

Self-Repairing Materials, Table 1 Self-healing mechanisms in metals

Mechanism	Precipitation	SMA reinforcement	Healing agent encapsulation
Type	Damage prevention (solid)	Damage management (solid, possibly liquid-assisted)	Damage management (liquid-assisted)
Matrix material	Al–Cu, Fe–B–Ce, Fe–B–N, etc.	Sn–Bi, Mg–Zn	Al
Reinforcement materials	–	NiTi	Sn–Pb
Microstructure parameter (ψ)	Solute fraction	Concentration of microwires	Concentration of microcapsules or low-melting-point alloy
Degradation measure (ξ)	Volume of voids	Volume of voids	Volume of voids
Healing measure (ζ)	Amount of precipitated solute	SMA strain	Amount of released healing agent
Characteristic length of degradation	Void size (microscale)	Void/crack size (macroscale)	Void/crack size (macroscale)
Characteristic length of the healing mechanism	Atomic scale (atomic diffusion)	Microwires diameter (macro or microscale)	Microcapsule size (microscale)
Phase transition involved	Solute precipitation	Martensite/austenite	Solidification of the solder
Healing temperature	Ambient	Martensite/austenite transition	Melting of the low-melting-point alloy
Property improved	Creep resistance	Restored strength	Restored strength and fracture toughness

suggested to for fixing surface damage. One approach is to mimic the healing of skin that has a network of vessels and veins so that cutting the skin triggers the blood flow, its coagulation, and sealing of the cut. The vascular network is a tree-like hierarchical structure that provides a uniform and continuous distribution of fluids throughout the material volume. The vascularization has been successfully used for polymeric composites [4–6].

A biomimetic approach combining self-healing with the Lotus-effect has been suggested, mimicking healing of a leaf surface. In that approach, micro-reservoirs with coating liquid (e.g., wax paraffin) were attached to the surface with the supply of coating to the damage area.

Various types of self-lubrication (e.g., self-replenishing solid and liquid lubrication, reservoirs for liquid lubricant, lubricant supply facilitated by oxidation or a tribochemical reaction, etc.) can also be viewed as surface healing mechanisms.

Theoretical Considerations

From the thermodynamic point of view, self-repair can be viewed as a nonequilibrium self-organization process that leads to increasing orderliness of the material and thus to decreasing entropy. In most schemes of self-repair, the self-healing material is driven away from the thermodynamic equilibrium either by the deterioration process itself or by an external intervention, such as heating. After that, the composite slowly restores thermodynamic equilibrium, and this process of equilibrium restoration drives the healing [5].

In most situations, damage repair at a certain length scale is achieved at the expense of the deterioration at the lower-length scale, making self-repairing materials multi-scale systems. For example, macroscale cracks are healed at the expense of fracturing macrocapsules of the healing agent embedded in the matrix of the material. In the precipitation-induced damage prevention mechanisms, micro-voids are healed by atomic-scale material transport. Multi-scale

organization is typical for biological materials and tissues, it is therefore not surprising that it plays a central role in biomimetic self-repairing materials.

In order to characterize degradation, it is convenient to introduce a so-called "degradation parameter" ξ, to represent, for example, the wear volume, and a corresponding generalized thermodynamic force, Y^{deg}. When a self-healing mechanism is embedded in the system, another generalized coordinate, the healing parameter, ζ, can be introduced, for example, the volume of released healing agent together with the corresponding generalized force, Y^{heal}. The generalized degradation and healing forces are external forces that are applied to the system, and flows are related to the forces by the governing linear Onsager's equations

$$\begin{aligned} J^{\mathrm{deg}} &= LY^{\mathrm{deg}} + MY^{heal} \\ J^{heal} &= NY^{\mathrm{deg}} + HY^{heal} \end{aligned} \qquad (1)$$

where L, M, N, H are corresponding Onsager coefficients [5].

The degradation force Y^{deg} is an externally applied thermodynamic force that results in the degradation. The healing force Y^{heal} is an external thermodynamic force that is applied to the system. In most self-healing mechanisms, the system is placed out of equilibrium and the restoring force emerges, so this restoring force could be identified with Y^{heal}. Since the restoring force is coupled with the degradation parameter ξ by the negative coefficients $N = M$, it also causes degradation decrease or healing.

Measure of Healing

A quantitative measure is required in order to characterize the efficiency of a healing mechanism and to compare different mechanisms. Several parameters have been suggested, such as (1) restored strength divided by the original strength, (2) fracture toughness divided by the original toughness, and (3) creep life of the material (in the case of damage-prevention mechanisms). Most self-repair mechanisms are not capable to provide 100 % restoration of the original properties of the material.

Cross-References

► Lotus Effect

References

1. Ghosh, S.K. (ed.): Self-Healing Materials: Fundamentals, Design Strategies, and Applications. Wiley, Weinheim (2009)
2. van der Zwaag, S. (ed.): Self Healing Materials – An Alternative Approach to 20 Centuries of Materials Science. Springer, Dordrecht (2007)
3. van der Zwaag, S.: Self-healing behaviour in man-made engineering materials: bioinspired but taking into account their intrinsic character. Phil. Trans. R. Soc. A **367**, 1689–1704 (2009)
4. Nosonovsky, M., Amano, R., Lucci, J.M., Rohatgi, P.K.: Physical chemistry of self-organization and self-healing in metals. Phys. Chem. Chem. Phys. **11**, 9530–9536 (2009)
5. Nosonovsky, M.: Self-organization at the frictional interface for green tribology. Phil. Trans R. Soc. A. **368**, 4755–4774 (2010)
6. Nosonovsky, M., Bhushan, B.: Thermodynamics of surface degradation, self-organization, and self-healing for biomimetic surfaces. Phil. Trans. R. Soc. A. **A367**, 1607–1627 (2009)

Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components

Moon-Ho Ham[1], Ardemis A. Boghossian[1], Jong Hyun Choi[2] and Michael S. Strano[1]
[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]School of Mechanical Engineering, Purdue University, West Lafayette, IN, USA

Synonyms

Biomimetic energy conversion devices; Photovoltaic devices; Self-regeneration; Solar cells

Definition

Self-repairing photoelectrochemical cells are synthetic solar energy conversion devices that illustrate principles of self-regeneration and repair through a reversible assembly/disassembly cycle. Motivated by the repair mechanism in living chloroplasts, this cycle can potentially extend the lifetime of a photoelectrochemical cell indefinitely.

Introduction

Advancements in nanotechnology toward the synthesis and manipulation of materials have enabled the direct interfacing of synthetic components with biological complexes to create new functions. Direct coupling of synthetic, nanoscale components with biological complexes has been used to study biological systems [1, 2] as well as develop new biomimetic electronics [3–5]. In most recent developments, protein complexes have been interfaced with nanoparticles and single-walled carbon nanotubes (SWCNTs) in the synthesis of biomimetic solar cells [6–10]. This new generation of nano-bio solar cells often relies on a combination of biologically derived components and biologically inspired mechanisms for device fabrication to minimize costs and enhance device efficiencies. As technology progresses in this field, the mechanisms demonstrated by the new devices increasingly mimic processes found in nature. In this work, the hydrophobicity of SWCNTs for the self-assembly of nanoscale components was employed to mimick the self-repair process used by plants [6]. In this sense, a thorough understanding of the natural processes is essential for developing such biomimetic devices.

Photosynthesis: Chloroplast Structure

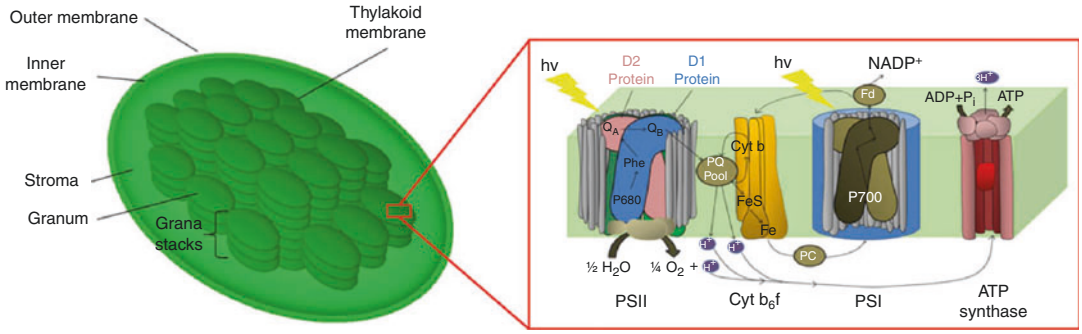
Natural light-harvesting mechanisms in plants are primarily carried out in the chloroplast, an organelle responsible for the conversion of sunlight into energy for the plant (Fig. 1) [11, 12]. Like most organelles, the chloroplast is surrounded by an outer as well as an inner membrane. A thick, aqueous fluid, the stroma, fills the organelle. The stroma contains genetic materials, including deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and several ribosomes. The most pronounced containment within the chloroplast is the stacks of thylakoid membranes, or grana stacks. Each of the thylakoid membranes contains four key protein complexes used in photosynthesis: photosystem II (PSII), cytochrome b_6f (cyt b_6f), photosystem I (PSI), and adenosine-5′-triphosphate (ATP) synthase. These complexes are responsible for carrying out the first stage of photosynthesis, the light-dependent reactions [11].

The light-dependent reactions are initiated by the absorption of light in PSII. In PSII, light is primarily absorbed by the P680 site at approximately 680 nm. The light-harvesting antennas surrounding by the protein complexes contain an array of pigments and chlorophyll molecules that absorb light at various wavelengths throughout the solar spectrum, enhancing the overall absorbance of light by the plant. The photons absorbed by these antennas are then transferred to the P680 site via resonance energy transfer. Absorption of light at the P680 site excites an electron, transferring it to the pheophytin (Phe) site of the PSII reaction center (RC). The hole remaining in the P680 site is used in the oxidation of water for the production of oxygen and hydrogen.

$$2H_2O \xrightarrow{h\nu} O_2 + 4e^- + 4H^+$$

Meanwhile, the electron is transferred to the Q_A and Q_B sites of PSII, where it is ultimately shuttled via a redox carrier to the next major transmembrane protein complex, cyt b_6f.

Cyt b_6f is responsible for electron transfer between the two major RCs of the membrane, PSI and PSII. The electron transfer between the two centers is coupled to the establishment of a transmembrane proton gradient between the outer stroma and inner thylakoid space. The electron undergoes a series of redox reactions demoting electron energy in exchange for proton transfer



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 1 Chloroplast structure and function. The chloroplast (*left*) consists of an outer membrane surrounding stacks of thylakoid membranes, or grana. Each thylakoid membrane is embedded with several protein complexes (*right*) including PSII, cyt b6f, PSI and ATP synthase. Electron–hole separation occurs in PSI upon light absorption. The hole is used in the oxidation of water into hydrogen and oxygen, while the electron is transmitted through cyt b6f and PSI. Synthesis of ATP for chemical storage of energy ultimately occurs in ATP synthase

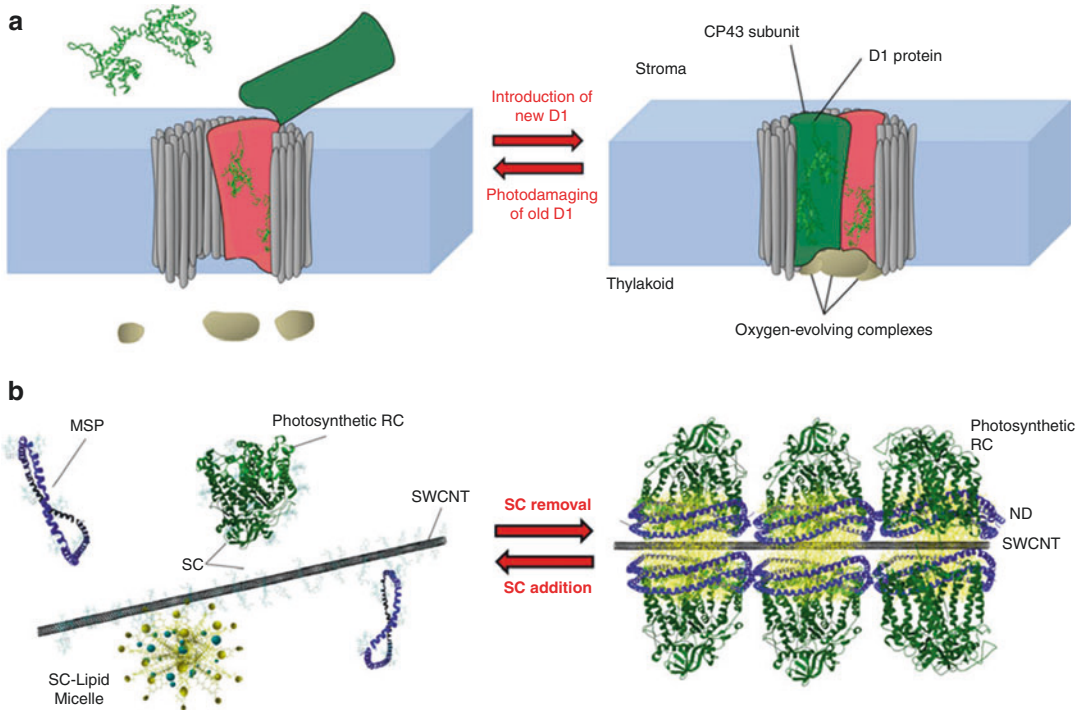
from the outer stroma to the inner thylakoid lumen. At the conclusion of this series of redox reactions, the electron is emitted at a lower energy state.

The electron emitted from cyt b_6f is transferred via a redox carrier to the next major protein complex, PSI. Like PSII, PSI is surrounded by a variety of pigment- and chlorophyll-containing antennas. The P700 site of PSI is responsible for light absorption at approximately 700 nm. As before, the surrounding antennas broaden the absorbance of the chloroplast and transfer this absorbed energy to the P700 site via resonance energy transfer. Upon energy absorption, the photoelectric effect is used to excite the electron at the P700, where it is emitted toward the ferrodoxin site for pickup by the next major complex. Excited electrons emitted by PSI are then subject to a variety of chemical pathways, one of which is nicotinamide adenine dinucleotide phosphate (NADPH) production. In this reduced state, the NADPH's reducing power is used to power the Calvin cycle, or the second stage of photosynthesis in the light-independent series of reactions, which takes place in the stroma.

The fourth, and final, major complex embedded within the thylakoid membrane, ATP synthase, is not directly involved in the preceding series of light-dependent reactions. In ATP synthase, a hydrogen ion is extracted from other sites in the membrane, including the water-oxidizing site of PSII, to produce ATP. ATP, which is also used in the Calvin cycle mentioned above, is the storage of light-absorbed energy in chemical form. Energy storage and release is carried out by phosphate bond formation and scission, respectively.

Natural Self-repair Cycles in Plants

The aforementioned system of light-dependent reactions relies on a series of embedded protein complexes under nearly continuous illumination that become highly susceptible to photodamage. In particular, the D1 subunit, which contains the P680, Phe, and the Q_B sites of PSII, is vulnerable toward protein damage. To address this, plants have evolved sophisticated mechanisms of self-repair wherein the damaged D1 protein is replaced with a newly synthesized protein (Fig. 2a) [13–15]. The initiation of the self-repair cycle is the damage of the D1 protein, which undergoes a change in conformation and phosphorylation levels when denatured. The dysfunctional protein signals the spontaneous, partial disassembly of the PSII complex. With complex disassembly, the damaged protein is released from the system and diffuses toward the unappressed, stromal region of the membrane. Meanwhile, new D1 protein biosynthesized elsewhere in the cell diffuses toward the appressed



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 2 Natural and synthetic mechanisms of self-assembly. (**a**) The self-repair in plants relies on the molecular recognition of components and trapping of meta-stable thermodynamic states. In the disassembled state (*left*), the damage D1 protein is released from the dysfunctional PSII and replaced with a new D1 protein. Introduction of the new protein triggers the self-assembly of PSII into a functional state (*right*). (**b**) The self-assembly of the RC-ND-SWCNT relies on the removal of the surfactant from the system. In the presence of SC, the systems exists in the disassembled state with surfactant dispersion of individual components (*left*). Upon SC removal, the proteins, lipids and nanotubes self-assemble into the complex shown (*right*)

region of the membrane. The selective diffusion of the damaged protein toward unappressed regions and new protein toward appressed regions of the membrane is attributed to the varying phosphorylation levels of the two proteins, which vary the degree of protein hydrophilicity. The more hydrophilic protein is driven toward the appressed region of the membrane whereas the more hydrophobic, damaged protein is driven toward the unappressed region. Introduction of the new protein to the disassembled complex triggers the spontaneous reassembly of a functional PSII incorporating the new D1 protein.

The described self-repair cycle consists of transitions between a series of metastable states in an energy-minimized fashion. Aside from the synthesis of a new protein, this cycle requires the input of no additional energy; the reversible disassembly and reassembly of PSII is completely driven by the formation of thermodynamically and kinetically trapped states. The self-assembly in particular is driven by hydrophobic/hydrophilic interactions and the assortment of diffusivities demonstrated by the proteins on the molecular scale. Synthetic replication of this reversible self-assembly process therefore hinges on the ability to control molecular interactions on the nanoscale.

Use of Biological Light-Harvesting Components in Solar Conversion Devices

Biological components such as PSI and photosynthetic RCs have external quantum efficiencies of

almost 100 % and energy yields of approximately 58 %. Recent works have demonstrated that the biological light-harvesting components can be incorporated into optoelectronic devices including photovoltaic devices [6–10].

Jennings and coworkers [10] reported that biomimetically inspired energy conversion devices were fabricated with spinach-derived PSIs which were bound to an Au electrode surface through covalent imine bonds. To increase the surface area of the electrode, the nanoporous leaf-like structure of an Au electrode was employed, which led to the threefold enhancement of the photocurrent due to the increased density of PSI complexes compared to planar electrodes.

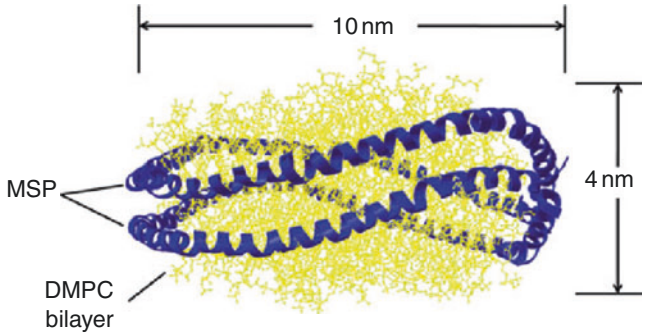
Carmeli and coworkers [8] directly bound PSI as a monolayer onto an Au surface. The PSIs were isolated from the cyanobacteria *Synechocystis* sp. PCC6803 and modified at cysteine sites near the P700 for covalent attachment. Kelvin probe force microscopy (KPFM) images show distinct photocurrent generation domains corresponding to PSIs covalently bound to the Au surface with unidirectional orientation. They extended their work to improve the photo-efficiency and apply it to other substrates. Orientated multilayers of PSI complexes were fabricated on the Au surface. In addition, they covalently attached PSI directly to carbon nanotubes and indirectly to GaAs surfaces via a small linker molecule.

A self-assembled monolayer of RCs, isolated from the purple bacterium *Rhodobacter* (*Rb.*) *sphaeroides*, on an Au electrode with a His tag was incorporated into solid-state photovoltaic cells, as reported by Baldo and coworkers [7]. The photocurrent spectrum of the device matched the absorption spectrum of the RCs, producing internal quantum efficiencies of ~12 %. Lebedev and coworkers [9] used cytochrome *c* as a conductive wire between photosynthetic RCs and the electrode, where the RCs were the same as those in Baldo's group, leading to a remarkable enhancement of the photocurrent. They also used arrayed carbon nanotube electrodes to improve the photo-conversion efficiency, where the RCs were encapsulated inside carbon nanotube arrays. The efficiency was considerably improved by increasing the number of RCs attached to the electrode surface by about fivefold compared to that obtained with the same proteins when immobilized on a planar graphite (HOPG) electrode.

Until now, most of these efforts have been centered on binding photoactive components onto an electrode surface in a unidirectional orientation to enhance photo-efficiencies. However, even with such an enhancement, these arrangements, like those often found in nature, undergo photo-degradation and protein damage. To address issues of photo-degradation, plants have evolved self-repair mechanisms whereby damaged proteins are replaced with photoactive, newly synthesized proteins. In fact, without this self-assembling mechanism, plants would produce less than 5 % of their typical photosynthetic yields with lifetimes on the order of minutes under intense illumination.

Synthetic Regeneration Cycles in Photoelectrochemical Cells

One approach to mimicking this self-assembly process is to use nanoscaled materials with controllable properties to interface with biologically derived photoactive components to create a photoelectrochemical cell capable of plant-like regeneration. In a recent study [6], this approach was used to develop the first photoelectrochemical cell capable of mimicking key steps in the self-repair cycle. An aqueous solution consisting of SWCNTs, RCs isolated from *R. Sphaeroides*, the phospholipid dimyristoylphosphatidylcholine (DMPC), membrane scaffold proteins (MSPs), and the surfactant sodium cholate (SC) (Fig. 2b, left) is placed within a dialysis bag with a 10,000 molecular weight cutoff and dialyzed against buffer to selectively remove SC from the system. The removal of SC spontaneously triggers the self-assembly of the photoactive complex shown in Fig. 2b (right). This complex contains a series of agglomerates sequentially aligned along the length of a SWCNT. These agglomerates consist of a lipid bilayer disk, or nanodisk (ND) (Fig. 3). In the ND, the DMPC



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 3 ND structure. NDs consist of DMPC lipids arranged into a bilayer surrounded by two strands of MSPs. Atomic force microscopy (AFM) measurements reveal that these lipid bilayer disks are approximately 10 nm wide and 4 nm high

molecules arrange themselves such that their hydrophilic heads face outward toward the aqueous solution and their hydrophobic tails are sandwiched within the bilayer. Wrapped around the circumference of this disk are two strands of MSP, the length of which determines the overall disk diameter. The particular strands of MSP used in this study result in NDs that are approximately 10 nm wide and 4 nm high. These NDs, which align along the length of the nanotube in the assembled state, house photoactive RCs. Although the self-assembly process occurs spontaneously, the RCs are specifically orientated such that their hydrophobic area (the region near the P680 site) faces the SWCNT and their hydrophilic area (the area near the Q_B site) faces outward toward its aqueous surroundings. This self-assembly process is completely reversible, such that the re-addition of surfactant to the solution decomposes the complex back into its initial micellar state, which is monitored using the photoluminescence shift of the nanotubes [6].

Understanding and Quantifying the Self-assembly Process

This spontaneous assembly of nano- and bio-based materials into a precise configuration via chemical signaling alone can be attributed to the comparative scalabilities of the synthetic and biological components and the controlled, molecular interactions that occur on the nanoscale. The removal of surfactant from the system and, most importantly, from the hydrophobic SWCNT surface triggers the formation of (meta) stable hydrophobic/hydrophilic agglomerates that minimizes SWCNT exposure to its aqueous surroundings. The precise locations and sizes of the hydrophobic/hydrophilic regions in the RC, the thickness of the ND bilayer, and the unidimensionality and length of the SWCNT are all factors that contribute to the precise molecular arrangement of the RCs along the nanotube length. Even the slightest alteration in sizes, diffusivities, and hydrophobicities of the nano-components are enough to perturb the dynamics of the system and even altogether inhibit the formation of such a photoactive complex. To develop a more thorough, quantitative understanding behind the formation of these complexes, the formation of various agglomerates upon surfactant removal was modeled as a series of reactions [16].

$$A[SC]_{free} \underset{k_{1r}}{\overset{k_{1f}}{\rightleftarrows}} [SC]_{micelle} \tag{1}$$

$$B[DMPC] + C[SC]_{free} \underset{k_{2r}}{\overset{k_{2f}}{\rightleftarrows}} [SC - DMPC]_{micelle} \tag{2}$$

$$[MSP] + D[SC]_{free} \underset{k_{3r}}{\overset{k_{3f}}{\rightleftarrows}} [SC - MSP] \tag{3}$$

$$[SWCNT] + E[SC]_{free} \underset{k_{4r}}{\overset{k_{4f}}{\rightleftarrows}} [SC - SWCNT] \quad (4)$$

$$F[DMPC] + [SWCNT] \underset{k_{5r}}{\overset{k_{5f}}{\rightleftarrows}} [DMPC - SWCNT] \quad (5)$$

$$G[MSP] + [SWCNT] \underset{k_{6r}}{\overset{k_{6f}}{\rightleftarrows}} [MSP - SWCNT] \quad (6)$$

$$2[SWCNT] \underset{kr}{\overset{k_{7f}}{\rightleftarrows}} [SWCNT_2] \quad (7)$$

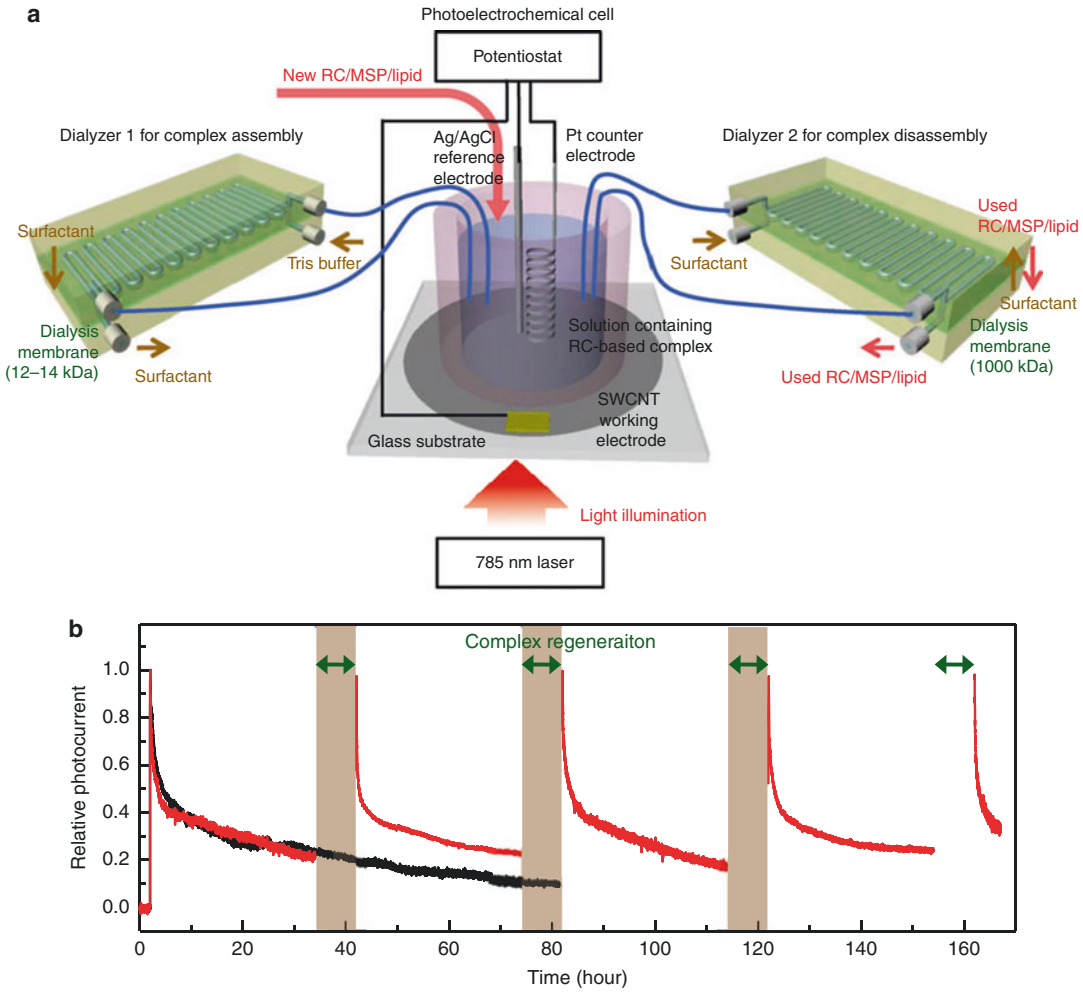
$$[SC] \overset{k_{8f}}{\rightarrow} [SC]_{removed} \quad (8)$$

$$H[DMPC] + I[MSP] \underset{k_{9r}}{\overset{k_{9f}}{\rightleftarrows}} [ND] \quad (9)$$

$$J[ND] + [SWCNT] \underset{k_{10r}}{\overset{k_{10f}}{\rightleftarrows}} [ND - SWCNT] \quad (10)$$

where $[SC]_{free}$ is free sodium cholate concentration in monomeric form, $[SC]_{micelle}$ is sodium cholate concentration in micellar form, [DMPC] is free lipid concentration in monomeric form, $[SC–DMPC]_{micelle}$ is SC-lipid mixed micelle concentration, [MSP] is MSP concentration, [SC–MSP] is SC-suspended MSP concentration, [SWCNT] is free SWCNT concentration, [SC–SWCNT] is SC-suspended SWCNT concentration, [DMPC–SWCNT] is lipid–SWCNT aggregate concentration, [MSP–SWCNT] is protein–lipid aggregate concentration, $[SWCNT_2]$ designates SWCNT bundle concentration, $[SC]_{removed}$ is the concentration of SC that has been removed from the system, [ND] is a ND concentration, and [ND–SWCNT] is ND–SWCNT concentration. In these reactions, k_f is the forward rate constant, k_r is the reverse rate constant, and A-J are stoichiometric coefficients. Fitting this model to experimental data on ND–SWCNT concentration over the course of dialysis, the best-fit rate constants for ND and ND-SWCNT formation were 79 $mM^{-1}s^{-1}$ and 5.4×10^2 mM^{-1} s^{-1}, respectively. In a diffusion-controlled process, one would expect the ND–SWCNT rate constant to be smaller than that of the ND, since the bulkier NDs and SWCNTs are expected to have smaller diffusivities than the individual lipid and MSP molecules. However, the reverse is the case, and the calculated ND–SWCNT rate constant is larger than that of the ND, indicating that the system is not under diffusion-controlled conditions. Because of the strongly hydrophobic nature of the SWCNT, the adsorption of the ND is expected to be nearly instantaneous, and in this case, the interaction is expected to be closer to diffusion-controlled conditions than ND formation [16].

The hydrophobic nature of SWCNTs not only induces the formation of ND–SWCNT complexes, but also the formation of other agglomerates, including SWCNT bundles. As discussed above, removal of surfactant from the SWCNTs forces the nanotubes to seek relatively hydrophobic surfaces to shield them from the aqueous surroundings. These surfaces may be lipid-formed agglomerates, such as NDs, or other nanotubes. In a diffusion-controlled dialysis, the structure with the largest diffusion coefficient is expected to diffuse toward the exposed nanotube surface more quickly. Large, bulky nanotubes with lengths on the order of microns demonstrate smaller diffusion coefficients than the more agile NDs; hence, they are expected to diffuse more slowly toward one another than the ND, favoring the faster ND-SWCNT formation. However, although diffusion occurs more slowly, the nanotube bundling is nearly irreversible: once nanotube bundles are formed, their dissociation would require the use of high-energy perturbations via sonication. Over extended periods of time, SWCNTs are expected to occupy their thermodynamically favored state in bundles, where strong, continuous, hydrophobic interactions along the entire length of the nanotube maintain irreversible bundle formation. Slower dialysis conditions would thus favor the formation of kinetically slower, but thermodynamically favored nanotube bundles. Faster dialysis conditions, on the other hand, favor the formation of kinetically faster, but thermodynamically metastable, ND–SWCNT complexes over smaller time scales [6].



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 4 Regenerating photoelectrochemical cell with self-assembled RC-ND-SWCNT complexes. (**a**) Schematic of the setup of photoelectrochemical systems which consists of a photoelectrochemical cell incorporated to two re-circulating membrane dialyzers. Dialyzers 1 and 2 are used for assembly and disassembly of the complex, respectively. The photodamaged RCs are removed during dialysis for disassembly of the complex and replaced with new ones while SWCNT are retained. (**b**) Temporal photoresponse of the RC-ND-SWCNT complexes with (*red line*) and without (*black line*) regeneration. (Reproduced from [6] © 2010 Nature Publishing group)

Photoelectrochemical Measurements of Self-assembled Complexes

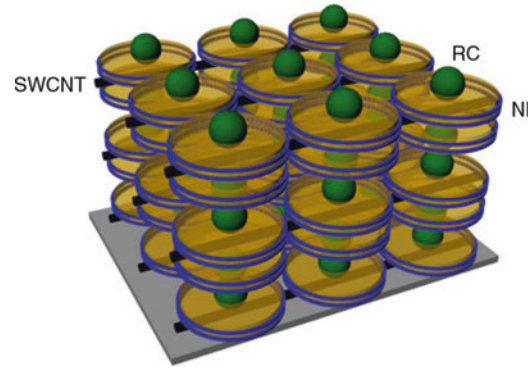
To exemplify the advantages of using the kinetically and thermodynamically favored formation of nanotube-based complexes in dynamic solar cells, the self-assembly was used as a means of regeneration in photoelectrochemical measurements. Photoelectrochemical measurements were carried out using a three-electrode system: a casted SWCNT working electrode, a Pt counter-electrode, and a Ag/AgCl standard reference electrode. A ferricyanide/ubiquinone dual mediator was used to optimize device efficiency. Illumination of a 700 nM solution of RC-ND-SWCNT yielded a per nanotube external quantum efficiency (EQE) of approximately 40 %. Upon introduction of SC to the illuminated solution, the

S

photoactive complex disassembles, and no photocurrent is recovered [6].

Although the exact mechanism behind the demonstrated photoactivity of the assembled complex remains unclear, it is hypothesized that electron–hole generation occurs in the RC in a manner similar to that found in nature, where the hole ultimately occupies the P680 site and the electron is shuttled to the Q_B. In nature, the Q_B site contains ubiquione. Ubiquinone molecularly consists of an electron-withdrawing head and a conjugated tail. In vivo, ubiquinone occupies a thermodynamically favored configuration within the RC wherein the head is docked in the Q_B site and the tail is extended outward toward to the surface of the protein complex. Since photoactivity is only observed in the presence of ubiquinone, it is hypothesized that the ubiquinone docks to the RC in an analogous manner, serving as a means of electron extraction from the Q_B site buried deep within the RC. On the other hand, the P680 site, which is much more easily accessible, faces the SWCNT, which may serves as a hole-conducting wire in this configuration.

Although photoactivity has been demonstrated in the assembled state, it remains to be seen whether reversible self-assembly could be used as a means of solar cell regeneration, as is the case with the plants. To address this, photoelectrochemical measurements were taken over an extended period of continuous illumination, provoking photodamage and diminishing photocurrents (Fig. 4). When the photocurrent approaches approximately 30 % of its initial value, SC is introduced to the system to disassemble the complex, the damaged proteins are replaced with new ones, and surfactant is once more removed from the system to reassemble the complex and recover photocurrent to its initial value. The synthetic regeneration cycle is hence conducted in analogy to the regeneration cycle used by plants, whereby chemical signaling is used to transition between the metastable assembled and disassembled states to replace photodamaged proteins with new ones. Using this regeneration cycle in the synthetic device, photoeletrochemical cell lifetime can be extended indefinitely while increasing efficiency by over 300 % over 168 h.



Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components, Fig. 5 High-density stacks for optimal efficiencies. Arrays of RC-ND-SWCNT can be aligned to maximize concentration of these complexes. Alignment of these complexes relies on recent advancements on SWCNT alignment

Nanotechnology as a Means of Modulating Bio-inspired Devices

The development of the regenerable, self-assembling solar device hinges on the ability to interface SWCNTs with biologically derived components. The precise arrangement of hydrophobic and hydrophilic regimes in biological complexes, such as RCs, is confined to regions that are at most a few nanometers apart. To utilize these variations in non-covalent interactions, one must subject these complexes to materials with properties that can be controlled on the nanoscale. For instance, nanotubes offer unidimensional confinement of hydrophobic surfaces that allows for the one-dimensional alignment of light-harvesting complexes. Confinement of these complexes into high-density, unidirectional arrays allows for device optimization and design. In fact, preliminary findings [6] indicate a linear increase in photoefficiency with increasing RC–ND–SWCNT complexes, suggesting that maximum efficiency (e.g., 40 %) could be achieved by maximizing concentration. Maximum concentration can be realized through high-density RC–ND–SWCNT stacks (Fig. 5). The key in synthesizing these stacks is controlling the alignment of not only RC-NDs along the nanotube, but also the nanotubes relative to one another. Recent advancements in nanotube

alignment [17–20] have expanded researchers' ability to precisely control device fabrication on the nanoscale, making the fabrication of such high-density stacks feasible.

Cross-References

- ▶ Carbon Nanotubes
- ▶ Hybrid Solar Cells
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanostructures for Energy
- ▶ Self-Assembly of Nanostructures

References

1. Rosi, N.L., Mirkin, C.A.: Nanostructures in biodiagnostics. Chem. Rev. **105**, 1547–1562 (2005)
2. Wong, S.S., Joselevich, E., Woolley, A.T., Cheung, C. L., Lieber, C.M.: Covalently functionalized nanotubes as nanometre-sized probes in chemistry and biology. Nature **394**, 52–55 (1998)
3. Oregan, B., Gratzel, M.: Low-cost, high-efficiency solar-cell based on dye-sensitized colloidal TiO2 films. Nature **353**, 737–740 (1991)
4. Colfen, H., Mann, S.: Higher-order organization by mesoscale self-assembly and transformation of hybrid nanostructures. Angew. Chem. Int. Ed. **42**, 2350–2365 (2003)
5. Lowe, C.R.: Nanobiotechnology: the fabrication and applications of chemical and biological nanostructures. Curr. Opin. Struct. Biol. **10**, 428–434 (2000)
6. Ham, M.H., Choi, J.H., Boghossian, A.A., Jeng, E.S., Graff, R.A., Heller, D.A., Chang, A.C., Mattis, A., Bayburt, T.H., Grinkova, Y.V., Zeiger, A.S., Van Vliet, K.J., Hobbie, E.K., Sligar, S.G., Wraight, C.A., Strano, M.S.: Photoelectrochemical complexes for solar energy conversion that chemically and autonomously regenerate. Nat. Chem. **2**, 929–936 (2010)
7. Das, R., Kiley, P.J., Segal, M., Norville, J., Yu, A.A., Wang, L.Y., Trammell, S.A., Reddick, L.E., Kumar, R., Stellacci, F., Lebedev, N., Schnur, J., Bruce, B.D., Zhang, S.G., Baldo, M.: Integration of photosynthetic protein molecular complexes in solid-state electronic devices. Nano Lett. **4**, 1079–1083 (2004)
8. Frolov, L., Wilner, O., Carmeli, C., Carmeli, I.: Fabrication of oriented multilayers of photosystem I proteins on solid surfaces by auto-metallization. Adv. Mater. **20**, 263 (2008)
9. Lebedev, N., Trammell, S.A., Spano, A., Lukashev, E., Griva, I., Schnur, J.: Conductive wiring of immobilized photosynthetic reaction center to electrode by cytochrome c. J. Am. Chem. Soc. **128**, 12044–12045 (2006)
10. Ciesielski, P.N., Scott, A.M., Faulkner, C.J., Berron, B.J., Cliffel, D.E., Jennings, G.K.: Functionalized nanoporous gold leaf electrode films for the immobilization of photosystem I. ACS Nano **2**, 2465–2472 (2008)
11. Asimov, I.: Photosynthesis. Basic Books, New York (1968)
12. Stern, K.R., Bidlack, J.E., Jansky, S.: Introductory Plant Biology, 11th edn. McGraw-Hill Higher Education, Boston (2008)
13. Aro, E.M., Virgin, I., Andersson, B.: Photoinhibition of photosystem-2 – inactivation, protein damage and turnover. Biochim. Biophys. Acta **1143**, 113–134 (1993)
14. Melis, A.: Photosystem-II damage and repair cycle in chloroplasts: what modulates the rate of photodamage in vivo? Trends Plant Sci. **4**, 130–135 (1999)
15. Melis, A.: Dynamics of photosynthetic membrane-composition and function. Biochim. Biophys. Acta **1058**, 87–106 (1991)
16. Boghossian, A.A., Choi, J.H., Ham, M.H., Strano, M. S.: Dynamic and reversible self-assembly of photoelectrochemical complexes based on lipid bilayer disks, photosynthetic reaction centers, and single-walled carbon nanotubes. Langmuir **27**, 1599–1609 (2011)
17. Xie, X.L., Mai, Y.W., Zhou, X.P.: Dispersion and alignment of carbon nanotubes in polymer matrix: a review. Mater. Sci. Eng. R Rep. **49**, 89–112 (2005)
18. Vigolo, B., Penicaud, A., Coulon, C., Sauder, C., Pailler, R., Journet, C., Bernier, P., Poulin, P.: Macroscopic fibers and ribbons of oriented carbon nanotubes. Science **290**, 1331–1334 (2000)
19. Haggenmueller, R., Gommans, H.H., Rinzler, A.G., Fischer, J.E., Winey, K.I.: Aligned single-wall carbon nanotubes in composites by melt processing methods. Chem. Phys. Lett. **330**, 219–225 (2000)
20. Murakami, Y., Chiashi, S., Miyauchi, Y., Hu, M.H., Ogura, M., Okubo, T., Maruyama, S.: Growth of vertically aligned single-walled carbon nanotube films on quartz substrates and their optical anisotropy. Chem. Phys. Lett. **385**, 298–303 (2004)

S

SEM

▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells

Semiconductor Nanocrystals

▶ Quantum Dot Toxicity

Semiconductor Piezoresistance

▶ Piezoresistivity

Semiempirical Methods

▶ Tight-Binding Simulations of Nanowires

Sense Organ

▶ Arthropod Strain Sensors

Sensors

▶ Arthropod Strain Sensors

Shark Denticles

▶ Shark Skin Drag Reduction

Shark Skin Drag Reduction

Amy Lang[1], Maria Laura Habegger[2] and Philip Motta[2]
[1]Department of Aerospace Engineering and Mechanics, University of Alabama, Tuscaloosa, AL, USA
[2]Department of Integrative Biology, University of South Florida, Tampa, FL, USA

Synonyms

Riblets, Shark denticles, Shark skin separation control

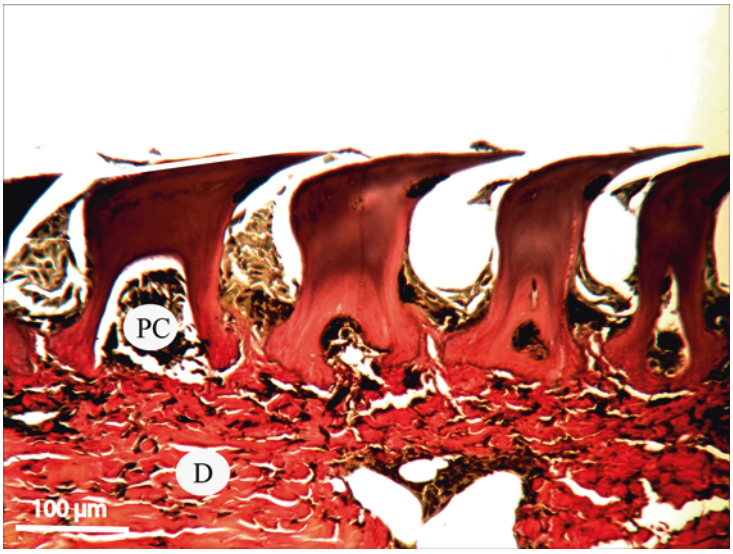
Definition

The scales, or denticles, on fast-swimming sharks have evolved two mechanisms for controlling the boundary layer flow over the skin surface leading to a reduction in drag. The first, and most widely known and studied, consists of the small streamwise keels covering the surface of the scales also known as riblets which reduce turbulent skin friction drag. The second mechanism is attributed to loosely embedded scales that are located on key regions of the body. When actuated to bristle by the flow, these scales potentially act as a means of controlling flow separation, thereby minimizing pressure drag during swimming maneuvers. Shark scales display a wide variation in geometry both across species while also varying with body location, but on faster swimming sharks, they typically range in size from 180 to 500 μm in crown length.

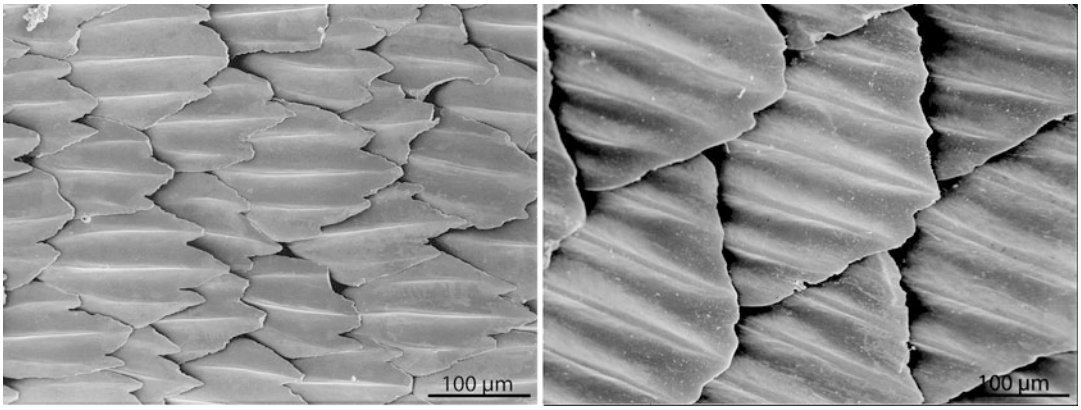
Overview

The Shark Skin

Sharks are covered with minute scales, also known as denticles or placoid scales because of their toothlike nature. The scales have a pulp cavity and a hard enameloid covering and are anchored at the base of the scale to the collagenous layer of the skin known as the stratum laxum (Fig. 1). The interlocking crowns of each scale make up the surface of the shark exposed to the water, and it is on the crown where many species have developed small riblets, or keels, orientated in the streamwise direction of the flow (Fig. 2). A reduction in the length of the base relative to the length of the crown and a change of shape of the base for some species over certain regions of the body appear to be the means by which certain scales have developed the capability to bristle or erect upon flow reversal (Fig. 3; flow would normally pass over the surface from left to right; flow reversal proceeding right to left can induce scale bristling as shown in Fig. 4 with a schematic shown in Fig. 5). The length of the scales is typically fixed for specific regions of the body within a species but differs among regions and



Shark Skin Drag Reduction, Fig. 1 Lateral view of sectioned placoid scales of a shortfin mako shark *Isurus oxyrinchus* in the region midway between the leading and trailing edge of the pectoral fin. Note the relatively long scale base relative to the crown length. The pulp cavity (*PC*) of the scale on the *left* is visible, and the base of the scales is anchored by collagen fibers to the dermis (*D*) of the skin

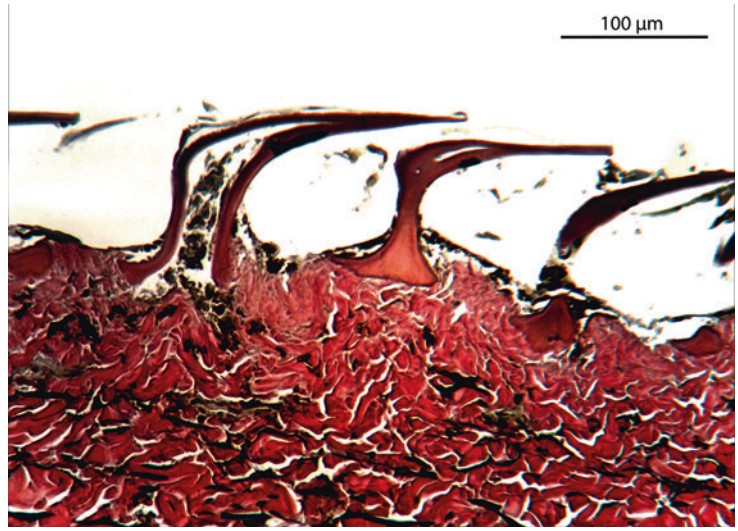


Shark Skin Drag Reduction, Fig. 2 Scanning electron micrograph (200×) of the placoid scales of a shortfin mako shark *Isurus oxyrinchus* (*left*) and blacktip shark *Carcharhinus limbatus* (*right*) from the dorsal body wall anterior to the dorsal fin. The mako shark scale has three keels or riblets, whereas the blacktip shark has five. Anterior is to the *left*

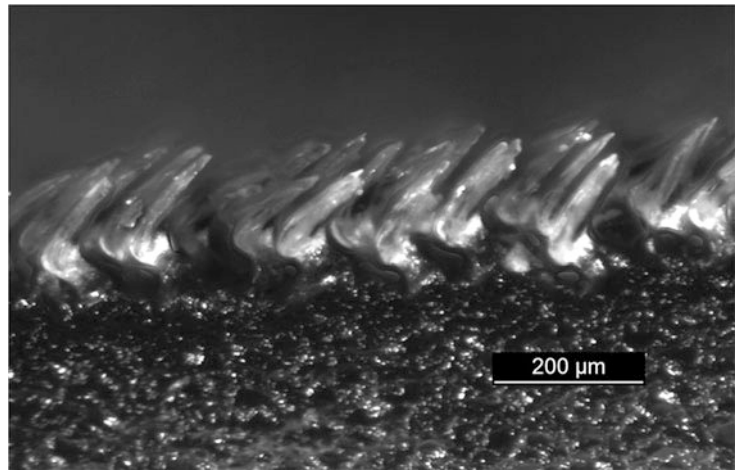
species. Similarly, the number of keels per scale is also consistent per location for a species. For instance, on the fast-swimming shortfin mako (*Isurus oxyrinchus*), the flank scales have a crown length of approximately 0.18 mm; each crown typically has three keels, each having a height of 0.012 mm and a spacing of 0.041 mm. The slower swimming blacktip shark (*Carcharhinus limbatus*) has flank scales typically 0.32 mm in length; each crown typically has five keels with a height of 0.029 mm and a spacing of 0.065 mm (Fig. 2).

Specifications and Fluid Dynamics

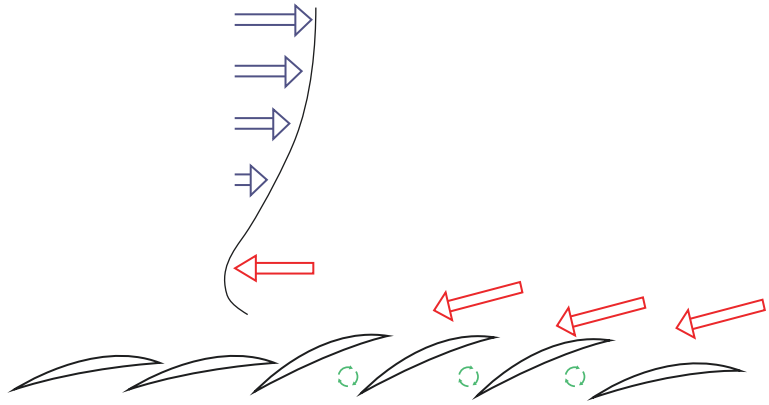
A shark swimming through water experiences two major sources of drag due to the viscous resistance of the fluid flow in the direction of motion. It is generally accepted that faster swimming sharks, such as the shortfin mako (*Isurus oxyrinchus*), have the capability to reduce both types of drag through evolutionary adaptations to their denticles. The first source of drag is skin friction drag; though not necessarily the largest contributor to overall drag, it is the one most associated with friction of the flow moving over the body of



Shark Skin Drag Reduction, Fig. 3 Lateral view of sectioned placoid scales in the flank region of a shortfin mako shark *Isurus oxyrinchus* in the region midway between the dorsal fin and the pectoral fin. Note the relatively long scale crown relative to the shorter base length



Shark Skin Drag Reduction, Fig. 4 Side view through the shortfin mako skin (from the flexible flank area) showing scales that have been manually erected. Because of the individual manual erection, not all scales are erected to the same degree. Flow would normally pass over the skin from *left* to *right*, and reversed flow, as occurs during separation, is hypothesized to cause bristling as shown



Shark Skin Drag Reduction, Fig. 5 Schematic showing forward flow in the boundary layer (*blue arrows*) subject to an adverse pressure gradient causing flow reversal (*red arrows*). The reversed flow close to the scales causes localized bristling to occur and the formation of cavity vortices (*circular green arrows*)

the shark. Skin friction drag is a result of the no-slip boundary condition between the water and the surface of the shark which results in the formation of a boundary layer. Because of the high speeds a shark can achieve (often greater than U = 10 m/s, where U is the swimming speed), this leads to a very high Reynolds number ($Re = Ux/\nu$, where x is length and ν is kinematic viscosity) in the boundary layer forming over most portions of the body and indicates the development of turbulent flow over most of the shark's body. For instance, at this speed, the typical transition location when the local $Re = 5 \times 10^5$ occurs at a location just x = 5 cm from the nose. While a turbulent boundary layer is less prone to separate from the body, it will have a skin friction magnitude 5–10 times larger than if the flow were to remain laminar. This riblet drag reduction mechanism has demonstrated the potential to reduce the turbulent skin friction drag by approximately 8–10 % in man-made applications [1, 2]. The size and spacing of the riblets are indicative of the speed at which a shark may swim with faster species having closer spaced keels.

The second type of drag, and the most important to be controlled, is due to a difference in pressure around the body and is often referred to as form drag. It is highly dependent on whether the flow remains attached or separates during swimming. Flow separation causes regions of low pressure on the downstream portions of the body leading to an imbalance of the pressure fore and aft and thus a dramatic increase in drag. The first means of decreasing pressure drag is streamlining, which consists of smoothing sharp corners and elongating the body with a tapered downstream portion. As a result, sharks and other marine animals have evolved a very streamlined body shape. Body proportions for fast-swimming sharks such as the shortfin mako are in part determined by its thunniform swimming ability. However, when it undergoes turning maneuvers, the fusiform body will undergo relatively greater body curvature than when swimming in a straight line. Turning may thus induce flow separation, but recent experimental evidence suggests that loosely embedded scales on the flank and behind the gills, in the region of maximum girth and further downstream, may act as a means of flow control.

Flow separation first involves a reversal of the fluid particles in a thin section adjacent to the surface; this is induced by a region of adverse pressure gradient occurring aft of the point of maximum girth in the streamwise direction. In other words, a suction pressure upstream induces the flow with the least momentum, that closest to the surface, to reverse. Scale bristling likely inhibits the process leading to reversed flow which thereby can control flow separation. Preventing separation will also favorably affect the flow further downstream over the caudal fin which can lead to higher thrust production. While, for a swimming shark, the production of thrust and drag occur simultaneously over the body and are inexorably linked, evidence signifies that sharks have evolved mechanisms within the structure of their skin to reduce drag leading to increases in thrust production and allowing greater maneuverability at high speeds.

Reif, a German biologist, working in the late 1970s is generally considered to be the first to report in literature the hypothesis as to the drag-reducing properties of shark skin [3]. Engineering research began at about the same time in both America [4] and Germany [5] as to the functional aspect of riblet surfaces for reducing skin friction drag. Seminal work into riblet design was completed by Bechert et al. in 1997, which demonstrated a maximum turbulent skin friction drag reduction for man-made, blade-like riblets of 9.9 % [6]. The bristling of shark scales leading to a mechanism for separation control was always hypothesized to function in a manner similar to a man-made technique known as vortex generators utilized since the 1940s [7, 8]. In more recent years, Lang et al. [9] have posited a new mechanism, whereby flow reversal leads to a region of localized scale bristling, leading to a flow-actuated separation control method which can be derived from the shark skin and for which research is ongoing.

Key Research Findings

Since the 1970s when studies began, riblets have become a well-accepted means of reducing

turbulent skin friction drag with an upper limit of reduction of just under 10 %. As previously stated, the most exhaustive testing of various riblet geometries, with comparisons made to other researchers working in the field, was completed in 1997 by Bechert et al. [6]. Previous work by Walsh [4] had focused on a sawtooth geometry, which resulted in a maximum drag reduction of about 5 %. While flow field measurement and visualization were not carried out to fully understand the mechanism behind the drag reduction, an exhaustive series of drag measurement experiments was carried out in a specially designed oil channel facility [6]; these measurements allowed for a determination of the most effectual geometry for man-made riblet applications.

Bernard and Wallace provide a summarization of the basic research that has been performed to map out the key characteristics found in turbulent flows [10]. To understand the mechanism whereby drag is reduced by surface modifications, some basic aspects of a turbulent boundary layer flow field need to be understood. A wall-bounded turbulent flow consists of a layer of vorticity within which fluid located further away from the surface is moving faster than that nearer to the surface due to the no-slip condition, and this results in an overall rotational characteristic of the flow in the clockwise direction for flow moving from left to right where the surface itself is stationary (this is the common reference frame to be used for studying a boundary layer flow). Within this layer, a complex assortment of horseshoe- or hairpin-shaped vortices of various sizes and stages of growth/decay are formed and interact. Scaling of the flow within the boundary layer can be achieved by considering a viscous length scale defined as $\delta_v = \nu/u^*$, where u^* is the friction velocity. The friction velocity is a function of the shear stress, or skin friction, at the surface (τ) and the fluid density (ρ) such that $u^* = (\tau/\rho)^{1/2}$. The viscous length scale determines the characteristic sizing of the fluid scales found in the boundary layer. For instance, in a given flow with fixed viscosity, ν, and at a particular downstream location, x, within a boundary layer consider a variation in free-stream velocity, U. As U is increased, the thickness of the boundary layer at that location will decrease, and the local average shear stress at that same location will increase. This results in an increase in the friction velocity and thus a decrease in the viscous length scale. The resulting decrease in viscous length scale reduces the characteristic sizing of the vortices forming within the boundary layer.

In the region close to the surface, longitudinal vortices with an axis of rotation in the streamwise direction are found to form and persist. These vortices have a characteristic diameter of approximately $30\delta_v$ and a streamwise length ranging from a few hundred up to 1,000 viscous length scales. When these vortices pair up, a region of low-speed fluid is lifted up from the wall, resulting in the formation of a low-speed streak. It is the instability of these streaks located in a region of high shear that leads to the process known as a turbulent burst and subsequent turbulent sweep. A burst is a sudden ejection of low-speed fluid up into the boundary layer, and a sweep is a sudden injection of high-speed fluid down toward the wall. It is the sweep of this high-momentum fluid onto the surface that results in localized, time-varying patches of high skin friction that are the main causation of increased drag in a turbulent boundary layer. It is the interaction of the shark skin with these components of the boundary layer flow that is essential to controlling the flow.

Results from the work of Bechert et al. [6] plotted the relative change in skin friction ($\Delta\tau/\tau$) (compared to a flat surface) versus spacing (s^+), where riblet spacing(s) is nondimensionalized such that $s^+ = s/\delta_v$. They found that the reduction in skin friction increases and reaches a maximum in the vicinity of $s^+ \sim 16$ and thereupon begins to decrease such that larger spacing can actually result in an increase in drag. It is noteworthy that this value corresponds to half the characteristic diameter of the longitudinal vortices forming close to the wall. Thus the riblets sized correctly restrain these near-wall vortices, which for reduced skin friction and reduced viscous length scale will now form at a slightly larger size, and these in turn induce the formation of low-speed streaks [1]. It was also found that a height of the riblets corresponding to half the

spacing ($h = 0.5$ s) gave optimal results. The other key result to garner from these experiments is the variation in maximum decrease with geometry, where a blade-like shape provided the upper limit for skin friction reduction (9.9 %). However, for durability of the surface in real applications, a trapezoidal configuration was tested and found to provide improved performance over the sawtooth geometry. Finally, it is remarkable that the geometry found on shark skin scales, with keels of a scalloped shape closely resembling the trapezoidal shape but with smoothed corners, was also tested by Bechert et al. [6] and found to perform comparably to that of the trapezoidal shape. Many shark species also appear to have the approximate $h = 0.5$ s ratio found to be optimal in experiments. Both of these are indicators that sharks have indeed evolved a scale and riblet geometry for turbulent skin friction reduction; however, the keels may play a role in separation control as well.

Experiments to discern the benefits of shark scale bristling began with Bechert et al. [8]; they used a shark skin replica whereby an overlapping array of individual shark scales was built and tested in their oil tunnel facility. In this man-made model of the shark skin, they meticulously built 800 small replicas of a hammerhead (*Sphyrna zygaena*) shark denticle scaled up 100 times in size (resulting in a crown length of 19 mm and riblet spacing of 4 mm). Each scale was anchored with a compliant spring with variation in stiffness to discern if denticle bristling could result in any additional mechanism to further decrease turbulent skin friction drag. The model was used for cases where the scales laid flat or were wholly bristled at a collective angle of attack. Bristled cases only resulted in increased drag, with higher spring stiffness resulting in higher drag. Flat, aligned scales gave comparable results to riblet surfaces with the only difference being that the greatest drag reduction achieved around $s^+ = 15$ had a value of about 3 %. This decrease in performance was attributed to the construction of the model with small gaps and other imperfections preventing a smoother surface. Thus, if bristling of the scales is to be advantageous to the shark, it must be something that is only activated upon demand and thus concurs with the postulation that bristling is utilized as a means to control flow separation.

Initial speculation began with the hypothesis that bristled shark scales act as vortex generators [8]. These devices produce streamwise vortices which energize the flow close to the surface. Vortex generators need to be placed at a specific downstream location within a boundary layer for maximum performance and typically upstream of the point of separation [9]. Another method of controlling flow separation, used to date at a more global scale, consists of movable flaps; for airfoil applications, these are placed close to the trailing edge (~10 % of chord length or larger) and have been shown to delay the onset of stall resulting in greater lift [8]. When the flap itself was given a 3D jagged trailing edge, they were also found to act more effectively.

More recent work by the authors [11–15] has investigated the mechanism by which shark skin bristling may lead to the development of a passive, flow-actuated mechanism for separation control. The current working hypothesis is that the passive, flexible scales of the shark work as microflaps to locally control flow separation as needed on crucial regions of the body where it most often occurs during swimming maneuvers. Recent observations on shark skin bristling angles on the shortfin mako (*Isurus oxyrinchus*) indicate that, for this particular species known for its swimming capability, only certain portions of the body have very flexible scales. Bristling capability was measured on dead specimens, and the effect of body pressurization was also considered as a potential bristling mechanism. However, results showed that subcutaneous skin pressurization did not cause scale bristling and had no effect on bristling angles; furthermore, such scales have no muscles attached to them as they lie in the deeper layer of the skin, the dermis. Because the flank scales are very loosely attached to the skin and highly mobile, these observations led us to infer that the scales are most likely bristled by reverse flow actuation. Sixteen body locations [12–14] were considered, six on fins and ten on the body. Scales were manipulated by a fine acupuncture needle and remained bristled once manipulated as shown in Fig. 4.

Scale angles vary with body location, but the most flexible scales are found along the flank of the body extending behind the gills to the tail; here scales are found to be easily flexible with slight manipulation on dead specimens to angles of 50° or greater. Highly flexible scales are also found at the trailing edge of the pectoral fins as compared to the leading edge where there was zero scale flexibility; this indicates the scales may be used to control dynamic stall (unsteady separation) to maintain lift forces on these surfaces during swimming and thereby maintain control. Contragility during swimming, or the ability to change direction quickly and easily, requires low pressure drag as well as high musculature control. These recent findings, herein reported as to variation in scale flexibility, corroborate the hypothesis that the flank region, extending from the location of maximum girth to the tail, is where a shark with a side-to-side swimming motion requires separation control to increase contragility. The scale flexibility appears to be a result of a reduction in size of the scale base anchored in the skin and a change in the shape of the base (as can be seen by the histological data shown in Figs. 1 and 3). The reduction appears to occur in the length of the base relative to its width, where the portion of the base that would pivot up and out of the skin at high bristling angles shows a decrease in length. Thus, the scales with greater bristling angles are less firmly anchored in the anterior to posterior direction and can pivot more freely within the skin.

Cassel et al. [16] describe the process leading to unsteady flow separation, as would occur in a turbulent boundary layer. In the presence of an adverse pressure gradient, the fluid closest to the wall, which also has the lowest momentum, is where flow reversal is first initiated (Fig. 5). This region of fluid moves back upstream and thickens and then subsequently erupts from the surface leading to a patch where the flow is separated from the surface. In a turbulent boundary layer, the fluid with the lowest momentum is that contained in the low-speed streaks, and separation of a streak results in the formation of a horseshoe-shaped structure as observed in computational studies of a separating turbulent boundary layer [17]. Thus for the shark skin, made up of a staggered array of scales, flow reversal in long, thin patches of fluid may locally bristle the scales, thereby interrupting the flow separation process. This feature of shark skin resulting in a surface with a favored flow direction is likely key to its capability to inhibit flow separation.

Furthermore, previous experiments over a bristled shark skin model established the existence of embedded cavity vortices, axis of rotation in spanwise direction, forming between replicas of the scales [11]. Thus, if flow is induced to form between the scales when bristled, there are two supplementary mechanisms that may help to control the flow. The formation of embedded vortices, similar to those created by golf ball dimples, would allow the flow to pass over the surface with an ensuing partial slip condition, thereby leading to higher momentum adjacent to the skin. Secondly, with a turbulent boundary layer flow forming above the cavities, there may be added momentum exchange whereby high-momentum fluid is induced at a greater rate to move toward the skin and into the cavities. This latter mechanism, resulting in turbulence augmentation [1], is another possible means to enhance the momentum overall in the flow closest to the wall. These three mechanisms may be working in combination to control flow separation over the skin of the shark. Finally, experimental verification that shark skin can control flow separation was recently obtained [15]. In two different experiments using real shark skin specimens, a pectoral fin and hydrofoil covered with flank skin, flow separation was controlled for both laminar and turbulent boundary layer conditions.

Past and Future Applications

The application of drag-reducing techniques to vehicles in both air and water has obvious advantages and at the same time limitations. Typical riblet spacing for both water and air applications, at speeds normally encountered, falls in the range of ~0.035 mm. Thin plastic films with adhesive on one side and riblets on the other have been made commercially available; thus a wide array of testing and use on aircraft has already taken place

where drag reduction was documented. However, practical limitations from cost, added weight, and maintenance of the riblet film (particularly with respect to particulates clogging the surface) with only an associated total decrease in drag of 1–3 % have prevented widespread use in most aircraft applications [1].

Separation control cannot only reduce drag but can also lead to increased maneuverability for vehicles. Decreasing drag overall can lead to increased fuel efficiency, payload and range in both military and commercial applications. Flow separation is also an important issue for maintaining use of control surfaces on vehicles (i.e., prevention of stall), which can also have relevance to helicopter rotors and turbine/compressor blades. Passive control mechanisms, including those found on shark skin, have been and will continue to be applied in all these applications.

Cross-References

- ▶ Biomimetics
- ▶ BioPatterning
- ▶ Shark Skin Effect

References

1. Gad-el Hak, M.: Flow Control: Passive, Active and Reactive Flow Management. Cambridge University Press, Cambridge, UK (2000)
2. Bhushan, B.: Shark skin effect. In: Encyclopedia of Nanotechnology. Springer, Berlin (2012)
3. Reif, W.-E.: Protective and hydrodynamic function of the dermal skeleton of elasmobranchs. Neues Jahrb. Geol. Palaontol. Abh. **157**, 133–141 (1978)
4. Walsh, M.: Drag characteristics of V-groove and transverse curvature riblets. In: Hough, G.R. (ed.) Viscous Flow Drag Reduction. Progress in Astronautics and Aeronautics, vol. 72, pp. 168–184. AIAA, New York (1980)
5. Bechert, D., Hoppe, G., Reif, W.: On the drag reduction of the shark skin. American Institute of Aeronautics and Astronautics (AIAA) Paper No. 85-0546 (1985)
6. Bechert, D., Bruse, M., Hage, W., Van der Hoeven, J., Hoppe, G.: Experiments on drag-reducing surfaces and their optimization with an adjustable geometry. J. Fluid Mech. **338**, 59–87 (1997)
7. Bushnell, D., Moore, K.: Drag reduction in nature. Annu. Rev. Fluid Mech. **23**, 65–79 (1991)
8. Bechert, D., Bruse, M., Hage, W., Meyer, R.: Fluid mechanics of biological surfaces and their technological application. Naturwissenschaften **80**, 157–171 (2000)
9. Lin, J.: Review of research on low-profile vortex generators to control boundary-layer separation. Prog. Aerosp. Sci. **38**, 389–420 (2002)
10. Bernard, P., Wallace, J.: Turbulent Flow: Analysis, Measurement, and Prediction. Wiley, Hoboken (2002)
11. Lang, A., Motta, P., Hidalgo, P., Westcott, M.: Bristled shark skin: a microgeometry for boundary layer control? Bioinspir. Biomim. **3**, 046005 (2008)
12. Lang, A., Habegger, M., Motta, P.: Shark skin boundary layer control. In: Proceedings of the IMA Workshop "Natural Locomotion in Fluids and on Surfaces: Swimming, Flying, and Sliding," 1–5 June 2010. IMA Volumes in Mathematics and its Applications (2012)
13. Lang, A., Motta, P., Habegger, M., Hueter, R., Afroz, F.: Shark skin separation control mechanisms. Mar. Technol. Soc. J. **45**(4), 208–215 (2011)
14. Motta, P., Habegger, M., Lang, A., Hueter, R., Davis, J.: Scale morphology and flexibility in the shortfin mako *Isurus oxyrinchus* and the blacktip shark *Carcharhinus limbatus*. J. Morphol. **273**(10), 1096–1110 (2012)
15. Lang, A., Bradshaw, M., Smith, J., Motta, P., Habegger, M., Hueter, R.: Movable shark scales act as a passive dynamic micro-roughness to control flow separation. Bioinspir. Biomim. **9**, 036017 (2014)
16. Cassel, K., Smith, F., Walker, J.: The onset of instability in unsteady boundary-layer separation. J. Fluid Mech. **315**, 223–256 (1996)
17. Na, Y., Moin, P.: Direct numerical simulation of a separated turbulent boundary layer. J. Fluid Mech. **374**, 379–405 (1998)

Shark Skin Effect

Bharat Bhushan
Nanoprobe Laboratory for Bio- and Nanotechnology and Biomimetics, The Ohio State University, Columbus, OH, USA

S

Synonyms

Low fluid drag surface

Definition

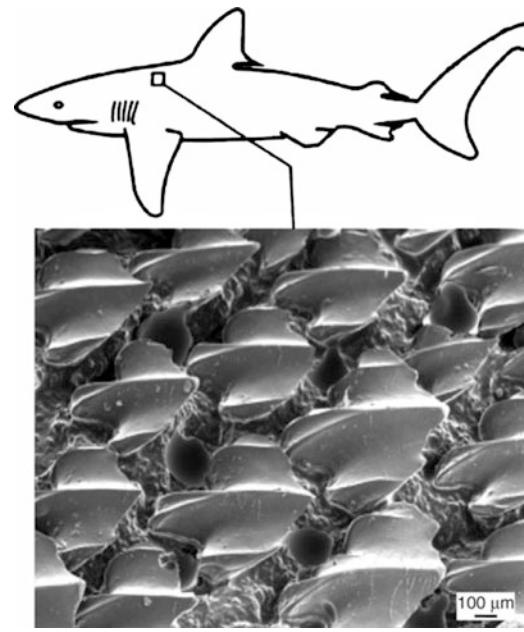
The shark skin effect is the reduction of the fluid drag during swimming at fast speeds and

protection of its surface against biofouling. The presence of surface microstructure on the skin surface is responsible for this effect.

Overview

Many structures, materials, and surfaces found in nature can be exploited for commercial applications. As an example, nature has created ways of reducing drag in fluid flow, evident in the efficient movement of fish, dolphins, and sharks [10, 14]. The mucus secreted by fish causes a reduction in drag as they move through water, and also protects the fish from abrasion by making the fish slide across objects rather than scrape, and it makes the surface of the fish difficult for microscopic organisms to adhere to [28]. It has been known for many years that by adding as little as a few 100 parts per million guar, a naturally occurring polymer, friction in pipe flow can be reduced by up to two thirds. Other synthetic polymers provide an even larger benefit [18]. The compliant skin of the dolphin has also been studied for drag-reducing properties. By responding to the pressure fluctuations across the surface, a compliant material on the surface of an object in a fluid flow has been shown to be beneficial. Studies have reported 7 % drag reduction [12].

Another set of aquatic animals which possess multipurpose skin is fast swimming sharks [14]. The skin of fast swimming sharks reduces the drag experienced by sharks as they swim through water and protects against biofouling. The tiny scales covering the skin of fast swimming sharks, known as dermal denticles (skin teeth), are shaped like small riblets and aligned in the direction of fluid flow (Fig. 1). Shark skin–inspired riblets have been shown to provide a drag reduction benefit up to 9.9 % [5]. The spacing between these dermal denticles is such that the riblets may not be very effective against very small (micro-)organisms – they probably work best against larger organisms such as mussels, algae, and barnacles. Prevention of undesirable accumulation of microorganisms protects the surface against biofouling. Slower sharks are covered in dermal denticles as well, but these are not shaped like riblets and do not provide much drag reduction benefits.



Shark Skin Effect, Fig. 1 Scale patterns on fast-swimming sharks (*Squalus acanthias*) [19]

The effect of riblet structures on the behavior of fluid drag, as well as the optimization of their morphology, is the focus of this entry.

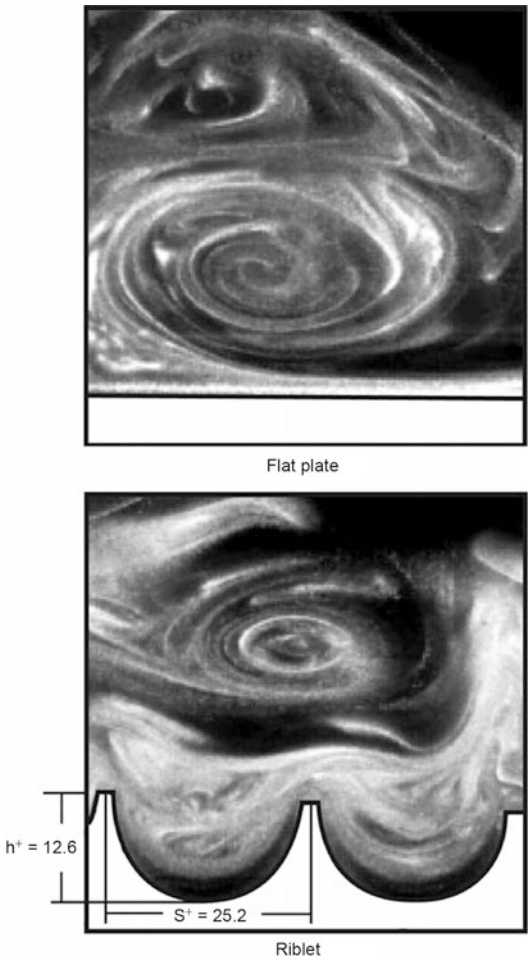
Mechanisms of Fluid Drag and Role of Riblets in Drag Reduction

Fluid drag comes in several forms, the most basic of which are pressure drag and friction drag [14]. Pressure drag is the drag associated with the energy required to move fluid out from in front of an object in the flow, and then back in place behind the object. Much of the drag associated with walking through water is pressure drag, as the water directly in front of a body must be moved out and around the body before the body can move forward. The magnitude of pressure drag can be reduced by creating streamlined shapes. Friction or viscous drag is caused by the interactions between the fluid and a surface parallel to the flow, as well as the attraction between

molecules of the fluid. Friction drag is similar to the motion of a deck of cards sliding across a table. Fluids of higher viscosity – the attraction between molecules – have higher apparent friction between fluid layers, which increases the thickness of the fluid layer distorted by an object in a fluid flow. An increase in drag occurs as fluid velocity increases. The drag on an object is in fact a measure of the energy required to transfer momentum between the fluid and the object to create a velocity gradient in the fluid layer between the object and undisturbed fluid away from the object's surface.

The above discussion of friction drag assumes all neighboring fluid molecules move in the same relative direction and momentum transfer occurs between layers of fluid flowing at different velocities. Fully developed turbulent flow is commonly said to exhibit complete randomness in its velocity distribution, but distinct regions exist within fully developed turbulent flow that exhibit different patterns and flow characteristics [20]. As these vortices rotate and flow along the surface, they naturally translate across the surface in the cross-flow direction. The interaction between the vortices and the surface, as well as between neighboring vortices that collide during translation, initiates bursting motions where vortices are rapidly ejected from the surface and into the outer boundary layers. As vortices are ejected, they tangle with other vortices and twist such that transient velocity vectors in the cross-stream direction can become as large as those in the average flow direction [20]. The translation, bursting of vortices out of the viscous sublayer, and chaotic flow in the outer layers of the turbulent boundary layer flow are all forms of momentum transfer and are large factors in fluid drag. Reducing the bursting behavior of the stream-wise vortices is a critical goal of drag reduction, as the drag reduction possibilities presented by this are sizable.

The vortices have been visualized using flow visualization techniques to capture cross-sectional images, shown in Fig. 2, of the stream-wise vortex formations above both flat-plate and riblet surfaces [26]. The average cross-stream wavelength of these high- and low-speed streaks, the added



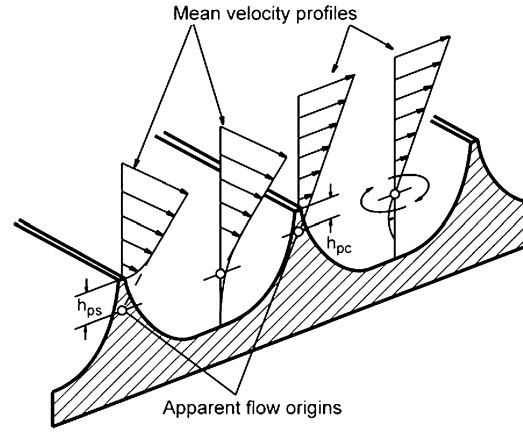
Shark Skin Effect, Fig. 2 Turbulent flow visualization of stream-wise vortices in a vertical cross section over flat-plate and riblet surfaces (Adapted from Ref. [26])

widths of one high-speed streak and one low-speed streak, is equal to the added diameters of two neighboring vortices and has been measured at 70–100 wall units [6, 20, 34]. This corresponds to a vortex diameter of 35–50 wall units. The flow visualizations in Fig. 2 show vortex cross sections and relative length scales, demonstrating vortex diameters smaller than 40 wall units [26]. (As flow properties change, the dimensions of the turbulent flow structures change as well. As such, it is useful to use non-dimensional length values to better compare studies performed in different flow conditions. Dimensionless wall units, marked $^{+}$, are used for all length scales,

which are calculated by multiplying the dimensional length by V_τ/ν. For example, $s^+ = sV_\tau/\nu$, where s^+ is the non-dimensional riblet spacing, s is the dimensional riblet spacing, ν is the kinematic viscosity, and $V_\tau = (\tau_0/\rho)^{0.5}$ is the wall stress velocity, for which ρ is the fluid density and τ_0 is the wall shear stress. Wall shear stress can be estimated for round pipe flow using the equation $\tau_0 = 0.03955\nu^{1/4}\rho V^{7/4}d^{-1/4}$, where V is the average flow velocity and d is the hydraulic diameter. For flow in rectangular pipes, the equation for hydraulic diameter $d = 4A/c$ can be applied, where A is the cross-sectional area and c is the wetted perimeter.)

The small riblets that cover the skin of fast swimming sharks work by impeding the cross-stream translation of the stream-wise vortices in the viscous sublayer. As vortices form above a riblet surface, they remain above the riblets, interacting with the tips only and rarely causing any high-velocity flow in the valleys of the riblets. Since the higher-velocity vortices interact only with a small surface area at the riblet tips, only this localized area experiences high-shear stresses. The low velocity fluid flow in the valleys of the riblets produces very low shear stresses across the majority of the surface of the riblet. By keeping the vortices above the riblet tips, the cross-stream velocity fluctuations inside the riblet valleys are much lower than the cross-stream velocity fluctuations above a flat plate [26]. This difference in cross-stream velocity fluctuations is the evidence of a reduction in shear stress and momentum transfer near the surface, which compensates the potentially drag-increasing effect of a larger surface area. Though the vortices remain above the riblet tips, some secondary vortex formations do occur that enter the riblet valleys transiently. The flow velocities of these transient secondary vortices are such that the increase in shear stress caused by their interaction with the surface of the riblet valleys is small.

Protruding into the flow without greatly increasing fluid drag allows the riblets to interact with the vortices to reduce the cross-stream translation and related effects. As the riblets protrude into the flow field, they raise the effective flow origin by some distance. The amount by which the height of the riblets is greater than the apparent vertical shift of the flow origin is referred to as the effective protrusion height. By calculating the average stream-wise velocity in laminar flow at heights over riblet surfaces and comparing them to the average stream-wise velocities in laminar flow at heights over a flat plate, the effective stream-wise protrusion height, h_{ps}, is found for laminar flow. The effective cross-stream protrusion height, h_{pc}, is similarly found for laminar flow by comparing the cross-stream velocities over a riblet surface to those over a flat plate. A schematic of stream-wise and cross-stream flow velocity profiles and effective protrusion heights is shown in Fig. 3. The difference between the vertical shifts in the stream-wise and cross-stream origin, $\Delta h = h_{ps} - h_{pc}$, for any riblet geometry has been proposed to be the degree to which that riblet geometry will reduce vortex translation for low Reynolds number (Re) flows [5]. As Re increases, the degree to which increased surface area affects the overall fluid drag increases, and the drag reduction correlation to the laminar flow theories deteriorates.



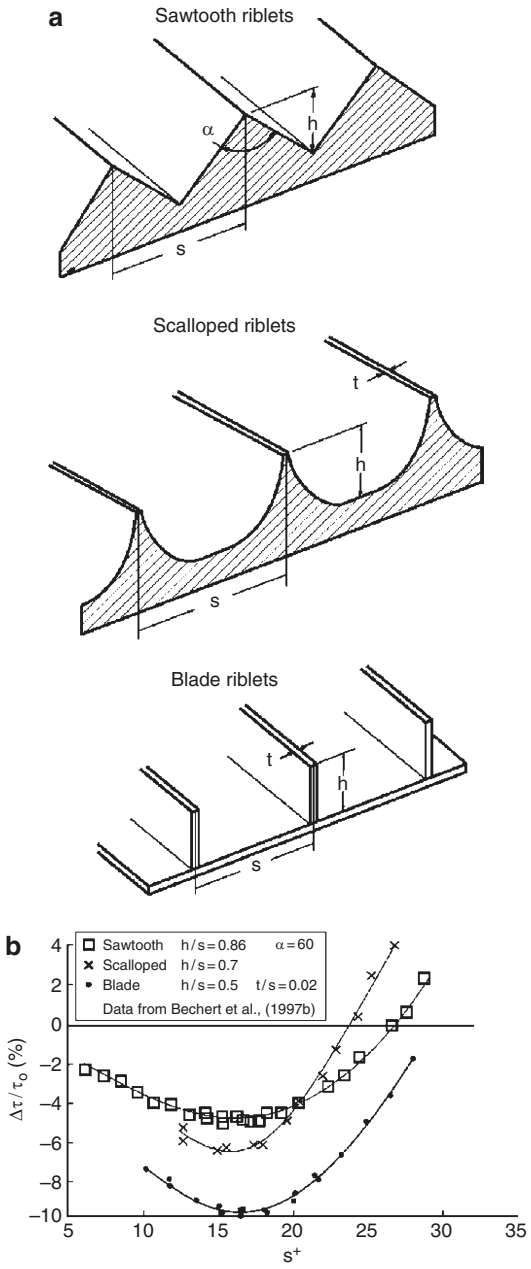
Shark Skin Effect, Fig. 3 Schematic representation of the mean velocity profiles and effective protrusion heights for flow in both the stream-wise direction, h_{ps}, and in the cross-flow direction, h_{pc} (Adapted from Ref. [5])

Optimization of Riblet Geometry

The cross-sectional shape of riblets on fast swimming sharks varies greatly, even at different locations on the same shark. Many types of riblets

have been studied, the shapes of which have been chosen for several reasons [14]. Riblet shapes have been chosen for their similarity to natural riblets, for their ease of fabrication, and for purposes of drag reduction optimization. Two-dimensional (2D) riblets, which have a continuous extrusion of a simple cross section in the stream-wise direction, have been most extensively characterized. The most thorough characterization has been completed for symmetrical 2D riblets with sawtooth, scalloped, and blade cross sections as shown in Fig. 4a [1, 2, 4, 6, 7, 29–32, 35, 36]. Alternative riblet geometries have, in general, shown no increased benefit. These riblets, including asymmetrical riblets, hierarchical riblets, and riblets with rounded or notched peaks have been studied in detail and do not improve upon the benefit of standard riblet geometries [29, 30, 32]. Other 2D riblet shapes which have been studied include alternating brother-sister-type riblets [4] and hierarchical riblets with small riblets on top of larger riblets [35]. Three-dimensional (3D) riblets, which include segmented 2D riblets as well as shark skin moldings and replicas have also been studied. Riblet types characterized include aligned segmented-blade riblets [35], offset segmented-blade riblets [6], offset-3D blade riblets [6], and 3D shark skin replicas [7, 19, 24].

Most studies are done by changing the non-dimensionalized spacing, s^+, by varying only fluid velocity and collecting shear stress data from a shear stress balance in a wind tunnel or fluid flow channel. The use of non-dimensional characteristic dimensions for riblet studies, namely, nondimensional spacing, s^+, is important for comparison between studies performed under different flow conditions. Non-dimensionalization accounts for the change in size of flow structures like vortex diameter, which is the critical value to which riblets must be matched. When comparing the optimal drag reduction geometries for sawtooth, scalloped, and blade riblets shown in Fig. 4b, it is clear that blade riblets provide the highest level of drag reduction, scalloped riblets provide the second most, and sawtooth riblets provide the least benefit [14]. A summary of comparison features for sawtooth, scalloped, and blade riblets is presented in Table 1.



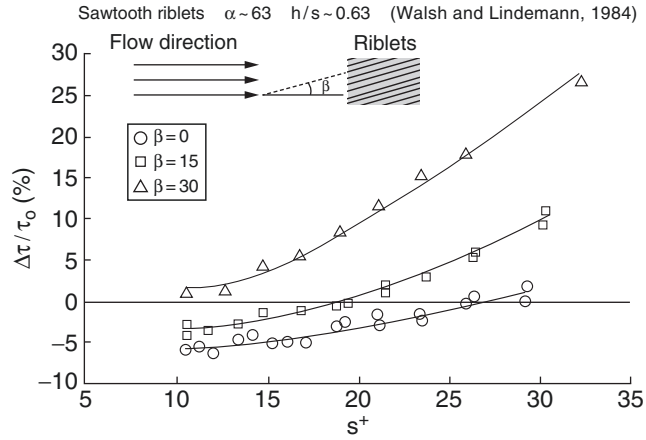
Shark Skin Effect, Fig. 4 (**a**) Schematic representation of riblet dimensions and (**b**) drag reduction comparison for sawtooth, scalloped, and blade riblets (Adapted from Ref. [5])

In general, it can be seen in Fig. 4b that each type of riblet is most beneficial near $s^+ \sim 15$, which is between 1/3 and 1/2, the width of the stream-wise vortices. Larger s^+ will cause vortices to begin

Shark Skin Effect, Table 1 Summary and comparison of optimum riblet geometry for various riblet shapes [14]

Riblet shape	Relative rank[a]	Maximum drag reduction (%)[b]	Optimum geometry[b]	Comments
Sawtooth	3	5	h/s ~ 1, α ~ 60°	Most durable
Scalloped	2	6.5	h/s ~ 0.7	
Blade	1	9.9	h/s ~ 0.5	Drag reduction increases as riblet thickness, t decreases. Durability is an issue.

[a]1 corresponds to greatest drag reduction
[b]Based on published data in Ref. [5]

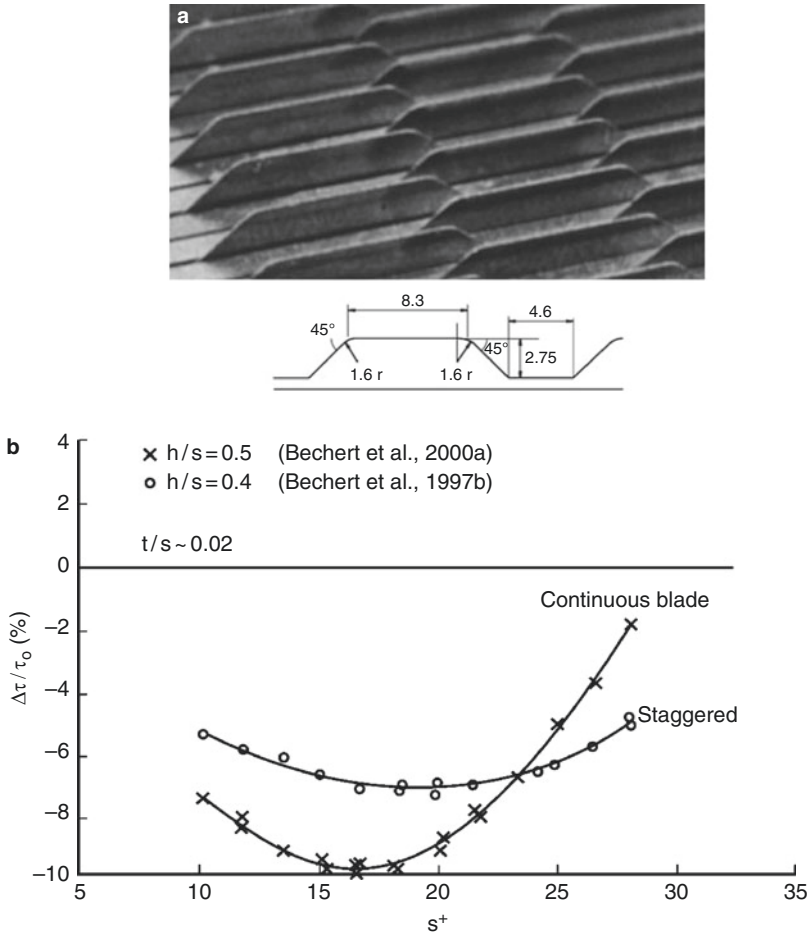


Shark Skin Effect, Fig. 5 Drag reduction dependence on yaw angle, β, of sawtooth riblets in free stream for h/s ≈ 0.62 (Adapted from Ref. [32])

falling into the gap between the riblets, which increases the shear stress at the surface between riblets. As s^+ decreases below optimum, the overall size of the riblets decreases to a point below which they cannot adequately impede vortex translation.

An additional concern to the application of riblets is the sensitivity of drag reduction to yaw angle, the angle between the average flow direction and the riblet orientation. Figure 5 shows the effects of yaw angle on riblet performance for flow over sawtooth riblets. Yaw angle has a deleterious effect on the drag reduction benefits of riblet surfaces. Riblet surfaces become drag inducing above $\beta = 30°$, but small drag reductions can still be seen up to $\beta = 15°$ [32].

Riblets on shark skin exist in short segments and groups, not as continuous structures. Riblets with 3D features have been created to better approximate the performance of actual shark skin and to determine if there are methods of drag reduction not yet understood from 2D riblet studies. Studies have explored the effects of compound riblet structures and 3D riblets comprised of aligned, segmented-blade riblets [36]. No improvement in net drag reduction was realized when compared to corresponding performance of continuous riblet geometries. More recently, experiments with similarly shaped segmented-blade riblets at spacing s with a matching set of segmented-blade riblets staggered between each row of blades at a spacing of s/2 from either side have been performed [5]. A schematic and image of staggered trapezoidal blade riblets is shown in Fig. 6a. Using these and other staggered riblets, Bechert et al. [5] hoped to achieve the same vortex elevation and anti-translation effects of continuous riblets with less effect on the flow origin. Experimental data comparing the largest drag reduction achieved with staggered segmented-blade riblets to optimum continuous blade riblets can be seen in Fig. 6b. Again, no net benefit in drag reduction was achieved, and after comparison of data, the conclusion was made that it is



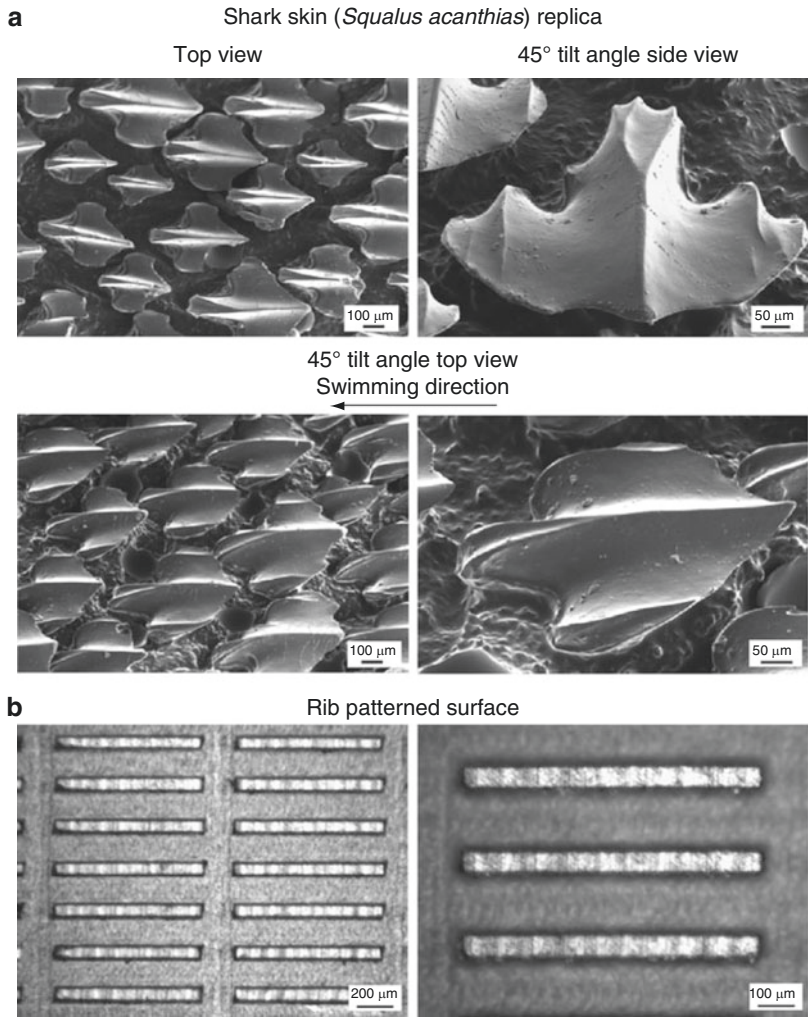
Shark Skin Effect, Fig. 6 Comparison of drag reduction over optimum continuous blade riblets with optimum segmented trapezoidal blade riblets. (**a**) Segmented riblets were staggered as shown. Spacing between offset rows is s/2, while spacing between corresponding rows is s. (**b**) Optimal hour to second ratio for staggered blade riblets is 0.4. Staggered blade riblets provide less drag reduction benefit than continuous blade riblets (Adapted from Ref. [5, 6])

unlikely that 3D riblets comprised of segmented 2D riblets will greatly outperform continuous 2D riblets.

The scales have been molded in epoxy resin from the skin of the Spiny Dogfish (*Squalus acanthias*) by Jung and Bhushan [19], as shown in Fig. 7a. A decrease in pressure drop corresponding to a decrease in fluid drag was achieved as compared to a smooth surface in a rectangular flow cell experiment, as shown in Fig. 8a. Pressure drop from inlet to outlet of a pipe is a measure of drag with large pressure drop occurring as a result of high drag. Aligned riblets were fabricated on acrylic by using micromachining (Fig. 7b), and a minimal decrease in pressure drop was realized as compared to a smooth acrylic test section (Fig. 8b) [19].

Riblet Fabrication and Applications

Riblets have been fabricated for studies and large-scale applications [14]. Typical microscale manufacturing techniques are ill-fitted for large-scale application due to the associated costs. Even for studies, most researchers have opted for traditional milling or molding methods over the microfabrication techniques used in the microtechnology industry. Though non-dimensional units allow for comparison between flow fields of different fluids and at different conditions, the accurate microscale manufacture of riblets for experimentation has been a field of study in its own right. The largest difficulty in optimizing riblet geometries has been the fabrication of riblet series with incremental changes in characteristic dimension. Riblets used in airflow require

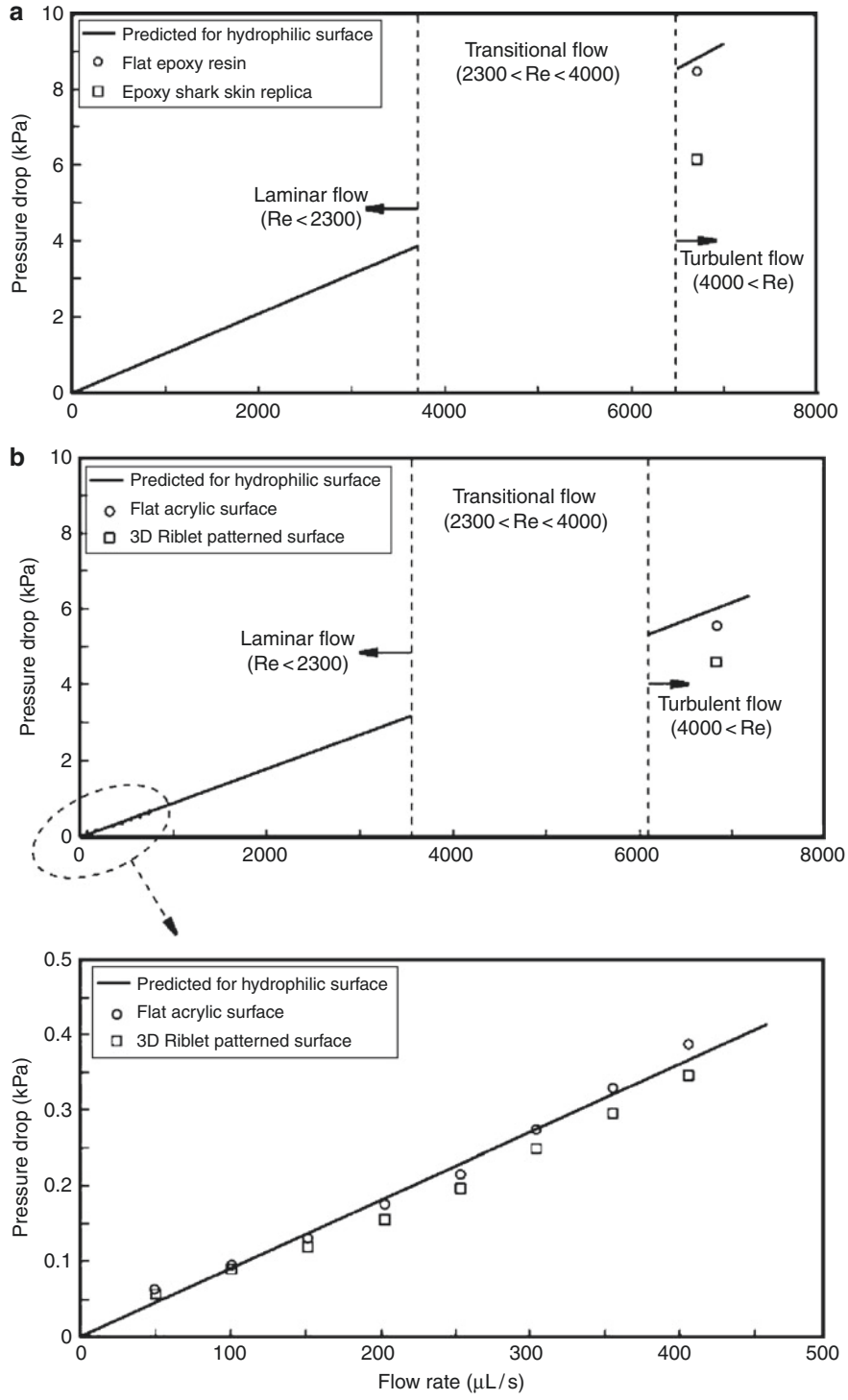


Shark Skin Effect, Fig. 7 (**a**) Scanning electron microscope (SEM) micrographs of shark skin replica patterned in epoxy, and (**b**) segmented bade-style riblets fabricated from acrylic [19]

spacings at or below 1 mm due to the low viscosity of air and the high speed at which wind tunnels must operate to create accurately measurable shear stresses on a test surface. Conversely, studies in an oil channel have been carried out in flow that is both highly viscous and slower moving. This allows for riblets to be made with spacings in the 3–10 mm range [3].

Commercial and experimental application of riblets outside wind tunnels and test stands is also limited by the high costs for less-than-optimal riblet performance. Application of riblets on a large scale has been done for several studies as well as for competition and retail purposes. Sawtooth riblets on vinyl films produced by 3 M riblets have been applied on surfaces ranging from boat hulls to airplanes. Racing swimsuits produced by Speedo and others also employ a riblet pattern on the surface to reduce drag during the streamline portion of each lap of a race [23]. Additionally, a novel surface-scratching technique has been applied to the inside surface of pipelines to create a faux-riblet surface [33].

Beginning in the mid-1980s, vinyl film sawtooth riblets have been applied to boat hulls for racing. Both an Olympic rowing boat and an America's Cup sailing yacht have been covered with riblets during competition. Because skin friction of an airplane accounts for as much as 48 % of total drag, vinyl film riblets have also been applied to test planes of both Boeing and Airbus. These films have not seen use on standard commercial



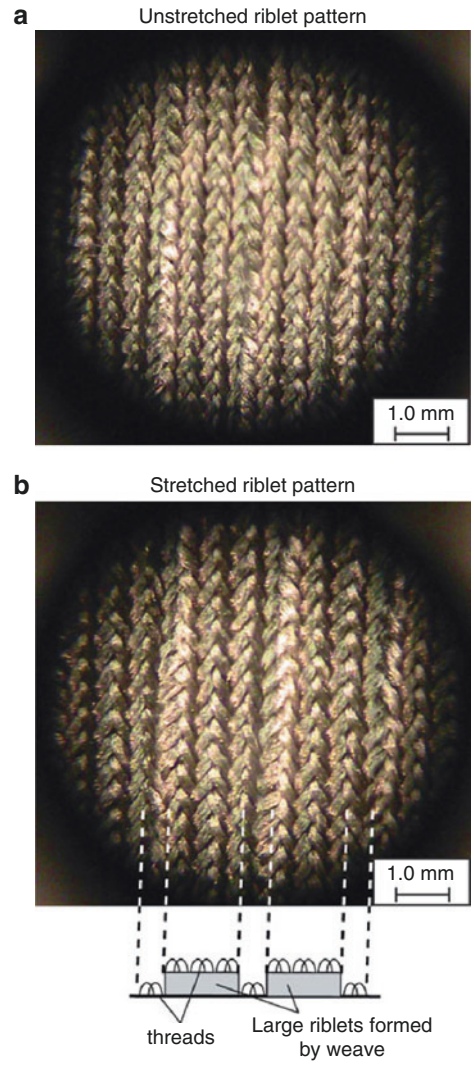
Shark Skin Effect, Fig. 8 (**a**) Comparison of pressure drop in rectangular pipe flow over flat epoxy surface with shark skin replica surface. (**b**) Comparison of pressure drop in rectangular pipe flow over flat acrylic surface and segmented blade riblets. Data are compared with the predicted pressure drop function for a hydrophilic surface (Adapted from Ref. [19])

flights yet, but the benefits seen in testing should not go unmentioned. Application of riblets to an airplane requires that several concessions are made. Several locations that would be covered by riblets must be left uncovered due to environmental factors; windows are not covered for the sake of visibility; several locations where dust and debris contact the airplane during flight are left bare because the riblets would be eroded during flight; and locations where deicing, fuel, or hydraulic fluid would come in contact with the riblets are left bare. After these concessions, the riblets covering the remaining 70 % of the aircraft have provided 3 % total drag reduction. This 3 % drag reduction correlates to a similar 3 % savings in fuel costs [4].

Another large commercial application for riblet technologies is drag reduction in pipe flow. Machining the surface or applying vinyl film riblets proves difficult in the confines of most pipes, and an alternate solution must be used. Experimental application of a scratching technique to the inside surface of pipes has created a riblet-like roughness that has provided more than 5 % drag reduction benefit [33]. Stemming from an old sailors' belief that ships sail faster when their hulls are sanded in the longitudinal direction, Weiss fabricated these riblets by using a steel brush moved through the pipeline to create a ridged surface. Studies have shown as much as a 10 % reduction in fluid flow with the combined effect of cleaning the pipe and ridging the surface. Tests on a 10-mile gas pipeline section have confirmed this benefit during commercial operation [4].

The dominant and perhaps only commercial market where riblet technology for drag reduction is commercially sold is competitive swimwear [14]. The general population became aware of shark skin's drag reduction benefits with the introduction of the FastSkin® suits by Speedo in 2004. Speedo claimed a drag reduction of several percent in a static test compared to other race suits. However, given the compromises of riblet geometry made during manufacturing, it is hard to believe the full extent of the drag reduction.

It is clear that creating surface structures by weaving threads is difficult. As a result, riblet geometries woven from thread have limited options of feasible riblet shapes. By the pattern woven into the FastSkin® swimsuits, riblets are formed which resemble wide blade riblets with small grooves on top. The larger riblets are formed by the macro-weaving pattern, and the smaller riblets are created by the individual weaves of thread aligned with the macro-riblets. Both of these riblet-like shapes are distinguishable in Fig. 9. As shown in Fig. 9a, unstretched



Shark Skin Effect, Fig. 9 Images of riblet geometries on (**a**) unstretched and (**b**) stretched Speedo FastSkin® swimsuit. (**c**) Schematic showing apparent hierarchical riblet structure formed by threads [14]

riblets are tightly packed. As the fabric stretches, the riblet width and spacing increase (Fig. 9b). The associated decrease in h/s ratio depends on the dimensions of each swimmer's body, which is another compromising factor in the design. Riblet thickness is also a factor considered in the design. Aside from the limitations imposed by the weaving patterns available, flexibility in the riblet tips will hinder the fabric's ability to impede the cross-stream translation of streamwise vortices. Thicker riblets are probably needed, used for strength, and cause a decrease in the peak drag reduction capability compared to thinner riblets.

3 M vinyl film riblets [27] have been applied to many test surfaces, including the inside of various pipes for pipe flow studies [22], flat plates in flow channels and wind tunnels, boat hulls in towing tanks [11], airplane wings [17], and airplane fuselages. Similar riblet films have been fabricated using bulk micromachining of silicon to create a master for molding of Polydimethylsiloxane (PDMS) to create a thin, flexible riblet film. This film has been used in flow visualization tests [25].

Grinding and rolling methods of riblet fabrication have been studied for application in both research and large-scale application. A profiled grinding wheel has been used to fabricate several riblet geometries based on sawtooth riblets with $h = 20$ μm and $s = 50$ μm [15]. Dressing of the grinding wheel was done with diamond-profile roller used in a two-step process in which the profile roller dresses every second tooth on the first pass, shifts axially the distance of one riblet spacing, and dresses the remaining teeth on a second pass. One downside of the grinding process is the lack of hardening on the final riblet surface. Alternatively, rolling methods can be used to strain harden the riblets during fabrication. Using a roller with the profile of two riblets on its outer face, a linearly patterned rolling process has been used to fabricate scalloped riblets in a titanium alloy with $h = 162$ μm and $s = 340$ μm [21]. The strain hardening, favorable grain patterns, and residual compressive stresses in the riblet surface after fabrication provide advantages in riblet strength for production applications.

Effect of Fish Mucus and Polymers on Fluid Drag

Fish are known to secrete mucus during swimming [14]. Though it is not known whether the mucus is present at all times, it is known that certain environmental factors cause, alter, or enhance the production of mucus. These environmental stressors may present a need for increased swimming speed to catch or avoid becoming prey, for protection against non-predatory threats such as microorganisms, or to resist abrasion while swimming near rocky surfaces. Regardless of which events cause fish to secrete mucus, the drag reduction benefits during mucus-assisted swimming are known. Numerous experiments have demonstrated the drag reduction possible with fish-covered mucus compared to non-mucus-covered shapes [18]. In an experiment comparing the drag on wax models to a mucus-covered fish, a reduction in skin friction drag of 50 % was seen [13].

Similar to these fish mucus experiments, polymer additives in pipe flows have been known for many years to reduce the drag in fluid flows by extreme amounts. Polymers are known to have low shear strength which results in low friction [8, 9]. In a pipe flow study comparing various injection techniques of polymer solutions into water, drag reductions of up to 80 % were achieved [16]. Additionally, the drag reduction benefit increases with increased Reynolds number. While this works well for pipe flows, in which the polymer remains mixed and active throughout the length of the pipe, its application to external flows is much more difficult. Mucus on fish does not mix well with water in static contact, but does mix and provide drag reduction during dynamic contact. By this feature, the mucus use of fish is minimized. Unfortunately for any long range application of polymer drag reduction on an external flow, the polymer solution must be continuously injected. This would cause large quantities of the solution to be used and likely render the strategy inefficient in terms of overall energy use.

Though sharks do not secrete enough mucus to use this mechanism for drag reduction, small amounts of mucus are present on the skin of

sharks [2]. It is possible that shark skin mucus secretion is similar to fish, where only environmental stressors or swimming causes an increase in output, but the total quantity of mucus on the surface at any given time is likely quite low. One possible mechanism by which this trace quantity of mucus could be useful is in changing the flow characteristics in the riblet valleys or at the riblet peaks, where shear stresses are highest. In the riblet valleys, a trace secretion of mucus could increase flow velocity and decrease the overall momentum transfer from the shark to the surrounding water by condensing the overall structure of the velocity gradient. Alternatively, injection at the riblet peaks may cause a reduction in shear stresses where they are at their maximum and again cause a reduction in drag. These small effects near the surface may propagate into a larger benefit as the lines of constant velocity in the flow shift and condense [14].

Closure

Fluid drag in the turbulent boundary layer is in large part due to the effects of the stream-wise vortices formed in the fluid closest to the surface. Turbulence and associated momentum transfer in the outer boundary layers is in large part due to the translation, ejection, and twisting of these vortices. Additionally, the vortices also cause high velocities at the surface which create large shear stresses on the surface. Riblets impede the translation of the stream-wise vortices, which causes a reduction in vortex ejection and outer layer turbulence. In addition, riblets lift the vortices off the surface and reduce the amount of surface area exposed to the high-velocity flow. By modifying the velocity distribution, riblets facilitate a net reduction in shear stress at the surface.

Various riblet shapes have been studied for their drag-reducing capabilities, but sawtooth, scalloped, and blade riblets are most common. Drag reduction by riblet surfaces has been shown to be as high as nearly 10 % given an optimal geometry of h/s ~0.5 for blade riblets with a no-slip condition. The maximum reliable drag reduction provided by scalloped riblets and sawtooth riblets is about 6 % at h/s ~0.7 and 5 % at ridge angle $\alpha \sim 60^\circ$, respectively. Experimentation of other shapes has provided similar benefits of around 5 % drag reduction. Though the optimum shape for drag reduction performance is blade riblets, the fragile nature of these blades makes their commercial application of little use. Scalloped and sawtooth riblets, which provide considerably less drag reduction benefit, are much stronger shapes mechanically speaking and should be used for application in environments where contact may occur with non-fluid materials.

Commercial applications of riblets include competition swimsuits, which use a thread-based riblet geometry, as well as experimental applications to airplanes. Drag reductions in riblet application have been accomplished, and flight applications have seen fuel savings of as much as 3 %. Manufacturing techniques for riblets must also be chosen specific to their application. Vinyl film riblets are the easiest method, as application of a film to a surface requires less for work small-scale application than other methods. Rolling or grinding methods of riblet application should be investigated for turbine blades or high volume commercially sold pieces.

Cross-References

- Biognosis
- Biomimetics
- Biomimicry

References

1. Becher, D.W., Hoppe, G.: On the drag reduction of the shark skin. Paper # AIAA-85-0564, presented at AIAA Shear Flow Control Conference, Boulder, 12–14 Mar (1985)
2. Bechert, D.W., Bartenwerfe, R.M., Hoppe, G., Reif, W.-E.: Drag reduction mechanisms derived from shark skin. Paper # ICAS-86-1.8.3, Proceedings of the 15th ICAS congress, vol. 2 (A86-48-97624-01), pp. 1044–1068. AIAA, New York (1986)
3. Bechert, D.W., Hoppe, G., van der Hoven, J.G.T., Makris, R.: The Berlin oil channel for drag reduction research. Exp. Fluids **12**, 251–260 (1992)

4. Bechert, D.W., Bruse, M., Hage, W., Meyer R.: Biological surfaces and their technological application – laboratory and flight experiments on drag reduction and separation control. Paper # AIAA-1997-1960, presented at AIAA 28th Fluid dynamics conference, Snowmass Village, 29 June–2 Aug (1997)
5. Bechert, D.W., Bruse, M., Hage, W., van der Hoeven, J.G.T., Hoppe, G.: Experiments on drag reducing surfaces and their optimization with an adjustable geometry. J. Fluid Mech. **338**, 59–87 (1997)
6. Bechert, D.W., Bruse, M., Hage, W.: Experiments with three-dimensional riblets as an idealized model of shark skin. Exp. Fluids **28**, 403–412 (2000)
7. Bechert, D.W., Bruse, M., Hage, W., Meyer, R.: Fluid mechanics of biological surfaces and their technological application. Naturwissenschaften **87**, 157–171 (2000)
8. Bhushan, B.: Principles and Applications of Tribology. Wiley, New York (1999)
9. Bhushan, B.: Introduction to Tribology. Wiley, New York (2002)
10. Bhushan, B.: Biomimetics: lessons from nature – an overview. Phil. Trans. R. Soc. A **367**, 1445–1486 (2009)
11. Choi, K.S., Gadd, G.E., Pearcey, H.H., Savill, A.M., Svensson, S.: Tests of drag-reducing polymer coated on a riblet surface. Appl. Sci. Res. **46**, 209–216 (1989)
12. Choi, K.S., Yang, X., Clayton, B.R., Glover, E.J., Altar, M., Semenov, B.N., Kulik, V.M.: Turbulent drag reduction using compliant surfaces. Proc. R. Soc A **453**, 2229–2240 (1997)
13. Daniel, T.L.: Fish mucus: *in situ* measurements of polymer drag reduction. Biol. Bull. **160**, 376–382 (1981)
14. Dean, B., Bhushan, B.: Shark-skin surfaces for fluid-drag reduction in turbulent flow: a review. Phil. Trans. R. Soc. A **368**, 4775–4806 (2010); 5737
15. Denkena, B., de Leon, L., Wang, B.: Grinding of microstructures functional surfaces: a novel strategy for dressing of microprofiles. Prod. Eng. **3**, 41–48 (2009)
16. Frings, B.: Heterogeneous drag reduction in turbulent pipe flows using various injection techniques. Rheol. Acta **27**, 92–110 (1988)
17. Han, M., Huh, J.K., Lee, S.S., Lee, S.: Micro-riblet film for drag reduction. In: Proceedings of the Pacific Rim Workshop on transducers and micro/nano technologies, Xiamen (2002)
18. Hoyt, J.W.: Hydrodynamic drag reduction due to fish slimes. Swim. Fly. Nat. **2**, 653–672 (1975)
19. Jung, Y.C., Bhushan, B.: Biomimetic structures for fluid drag reduction in laminar and turbulent flows. J. Phys.: Condens. Matter **22**, #035104 (2010)
20. Kline, S.J., Reynolds, W.C., Schraub, F.A., Runstadler, P.W.: The structure of turbulent boundary layers. J. Fluid Mech. **30**, 741–773 (1967)
21. Klocke, F., Feldhaus, B., Mader, S.: Development of an incremental rolling process for the production of defined riblet surface structures. Prod. Eng. **1**, 233–237 (2007)
22. Koury, E., Virk, P.S.: Drag reduction by polymer solutions in a riblet-lined pipe. Appl. Sci. Res. **54**, 323–347 (1995)
23. Krieger, K.: Do pool sharks really swim faster? Science **305**, 636–637 (2004)
24. Lang, A.W., Motta, P., Hidalgo, P., Westcott, M.: Bristled shark skin: a microgeometry for boundary layer control? Bioinspir. Biomim. **3**, 1–9 (2008)
25. Lee, S.J., Choi, Y.S.: Decrement of spanwise vortices by a drag-reducing riblet surface. J. Turbul. **9**, 1–15 (2008)
26. Lee, S.-J., Lee, S.-H.: Flow field analysis of a turbulent boundary layer over a riblet surface. Exp. Fluids **30**, 153–166 (2001)
27. Marentic, F.J., Morris, T.L.: Drag reduction article. US Patent 5,133,516, 1992
28. Shephard, K.L.: Functions for fish mucus. Rev. Fish Biol. Fish. **4**, 401–429 (1994)
29. Walsh, M.J.: Drag characteristics of v-groove and transverse curvature riblets. Viscous Flow Drag. Red. **72**, 169–184 (1980)
30. Walsh, M.J.: Turbulent boundary layer drag reduction using riblets. Paper # AIAA-82-0169, presented at AIAA 20th Aerospace sciences meeting, Orlando, 11–14 Jan 1982
31. Walsh, M.J., Anders, J.B.: Riblet/LEBU research at NASA Langley. Appl. Sci. Res. **46**, 255–262 (1989)
32. Walsh, M.J., Lindemann, A.M.: Optimization and application of riblets for turbulent drag reduction. Paper # AIAA-84-0347, presented at AIAA 22nd aerospace sciences meeting, Reno, 9–12 Jan 1984
33. Weiss, M.H.: Implementation of drag reduction techniques in natural gas pipelines. Presented at 10th European drag reduction working meeting, Berlin, 19–21 Mar 1997
34. Wilkinson, S.P.: Influence of wall permeability on turbulent boundary-layer properties, Paper # AIAA 83–0294, presented at 21st aerospace sciences meeting of the american institute of aeronautics and astronautics, Reno, 10–13 Jan 1983
35. Wilkinson, S.P., Lazos, B.S.: Direct drag and hot-wire measurements on thin-element riblet arrays. Presented at the IUTAM symposium on turbulence management and relaminarization, Bangalore, 19–23 Jan 1987
36. Wilkinson, S.P., Anders, J.B., Lazos, B.S., Bushnell, D.M.: Turbulent drag reduction research at NASA Langley: progress and plans. Inter. J. Heat Fluid Flow **9**, 266–277 (1988)

Shark Skin Separation Control

► Shark Skin Drag Reduction

Short (Low Aspect Ratio) Gold Nanowires

▶ Gold Nanorods

Short-Interfering RNA (siRNA)

▶ RNAi in Biomedicine and Drug Delivery

Shrinkable and Stretchable Nanomanufacturing

Clifford J. Engel and Teri W. Odom
Department of Chemistry, Northwestern University, Evanston, IL, USA

Definition

Shrinkable and stretchable nanomanufacturing extends the capabilities of soft lithography by allowing dynamic control of nanopatterns without the need to remake a master.

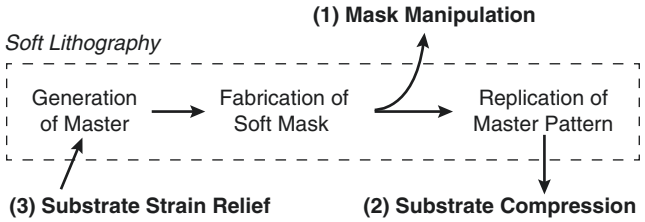
Overview

Early advances in nanomanufacturing were driven by innovations in the electronics industry. The invention of the integrated circuit in 1958 by Jack Kilby at Texas Instruments created a need for manufacturing tools that could pattern multiple components (e.g., transistors, interconnects, and other functional elements) on a single substrate. While the first commercially available chips did not have nano-sized components, future progress required the ability to reduce the size of components. For example, the deep-UV excimer laser lithography system was introduced in 1982, which overcame the 500 nm resolution limits of mercury-xenon lamps while increasing throughput by two orders of magnitude [1]. Advances in shorter-wavelength light sources were critical in improving the resolution of photolithography from 500 nm in 1990 to 45 nm and below in 2000 [2].

For emerging industries and consumer products besides chip-based electronics, like DNA microarrays and wearable sensors, a primary goal of nanomanufacturing has not been simply miniaturization but the ability to pattern nonplanar and soft material substrates at reduced costs. Thus, alternative strategies, such as soft lithography, to conventional microfabrication methods were developed [3]. Soft lithography offers a conceptually different approach to nanofabrication and consists of two major steps: (1) the fabrication of a master and (2) the replication of the master pattern. Typically, a master is generated using tools from the microelectronics industry, and the replication process involves a soft, elastomeric mask created by molding a polymer against a master pattern. The most common mask material, poly(dimethylsiloxane) (PDMS), is commercially available as Sylgard 184 (Dow Corning) and can replicate feature sizes that are 500 nm and larger with high fidelity. Composite stamps consisting of a thin, hard PDMS layer supported on a flexible PDMS backing can extend the pattern transfer and replication capabilities down to 50 nm in lateral dimensions [3] and < 2 nm in the vertical dimension [4]. To create replicas of the master on a desired surface, the soft mask is brought into contact with a patterning material. As a result, the PDMS mask determines the structural and geometrical characteristics of the molded and printed material.

The patterning material can vary depending on the desired chemical or physical properties of the pattern for an application. Printing techniques such as microcontact printing transfer the 2D structure of the PDMS mask by bringing molecular "ink" into contact with a metal thin film. Monolayers of alkanethiols can be printed on noble and coinage metals (Au, Ag, Pd, and Pt), while alkylsiloxanes can be printed on Si or glass. The end functional groups of the molecular inks control chemical properties of the patterned surface, such as hydrophobicity or reactivity [3].

In contrast to printing, molding retains the 3D structure of the mask, where the precursors are



Shrinkable and Stretchable Nanomanufacturing, Fig. 1 *Overview of shrinkable and stretchable nanomanufacturing:* Soft lithography processes involve generation of a master pattern through photolithography, fabrication of a soft mask molded against the master pattern, and replication of the master pattern into the patterning substrate. Modifying the nanopattern can be accomplished by (**1**) manipulating the soft mask, (**2**) compressing the substrate after patterning, or (**3**) relieving the strain in a thin film to form spontaneous features

usually liquid prepolymers such as UV-curable polyurethane or thermally curable epoxies [3]. The precursor material undergoes a transition from liquid to solid while in direct physical contact with the PDMS mask. The prepolymer is selected for optimal physical properties such as optical transparency or elasticity. Despite the versatility of soft lithography, the end goal of this suite of techniques is to replicate or duplicate the pattern on the master. Applications that require a range of structures, therefore, need to make a new master for each desired pattern, which can be time-consuming, tedious, and costly.

A new class of techniques—shrinkable and stretchable nanomanufacturing—can expand the capabilities of soft lithography to generate patterns without needing to remake the master or, in some cases, eliminate the need for a master entirely (Fig. 1). This article will describe how such patterns can be achieved by (1) manipulating the mask during the replication process, (2) post-processing the patterned substrate by compressing or stretching, and (3) avoiding a mask and relying on the spontaneous generation of patterns from unpatterned films.

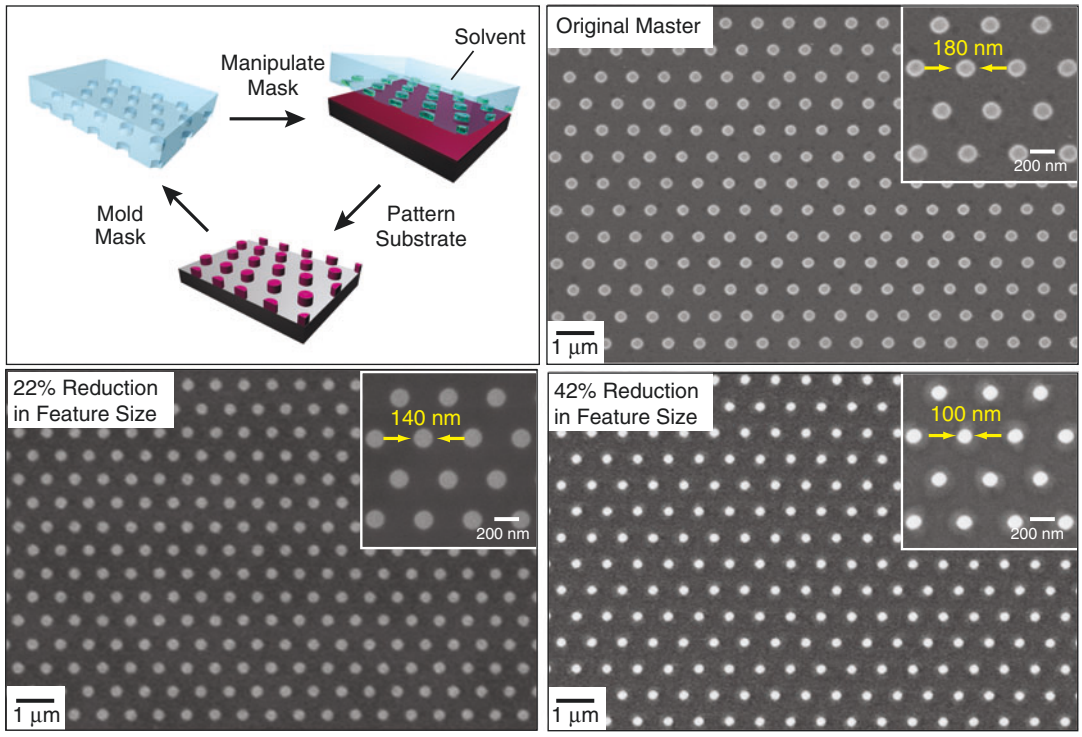
Manipulating the Soft Mask

Although PDMS is a ubiquitous elastomer for preparing masks in soft lithography, a major disadvantage is that it swells when in contact with nonpolar solvents [5]. Swelling of PDMS distorts the mask and is a major limitation for techniques that directly replicate the mask pattern. The swelling ratio S (the ratio of the length of PDMS in solvent over the length of dry PDMS) can be as large as 2.13 in diisopropylamine, 1.30 in toluene, and 1.22 in chlorobenzene [5]. As a result, most procedures use only non-swelling solvents such as water, dimethylformamide, or ethanol ($S = 1.00–1.02$) [5].

Solvent-assisted nanoscale embossing (SANE) is an all-moldable nanofabrication platform that has exploited the swelling properties of PDMS to produce, from a single master, nanoscale patterns with programmable feature sizes. Using solvents with specific S, the degree of swelling can shrink the size of the features accordingly in the mask during the embossing process *without* affecting the periodicity of the patterns (Fig. 2). For example, changing the solvent from a low-swelling solvent like dimethylformamide to a high-swelling solvent like dichloromethane can reduce feature sizes from 140 nm to 100 nm (≈30 %). Since swelling is a reversible process, a single PDMS mask can be dried and soaked in different solvents during iterative embossing processes to generate patterned surfaces with varying feature sizes without different master patterns [6].

Post-processing of Patterned Substrates

If patterns are created on hard or stiff substrates such as Si, the separation between features cannot be modified post-patterning. In contrast, patterning on soft substrates enables possibilities to tune the spacing and symmetry of features by compressing or stretching the substrate. Although

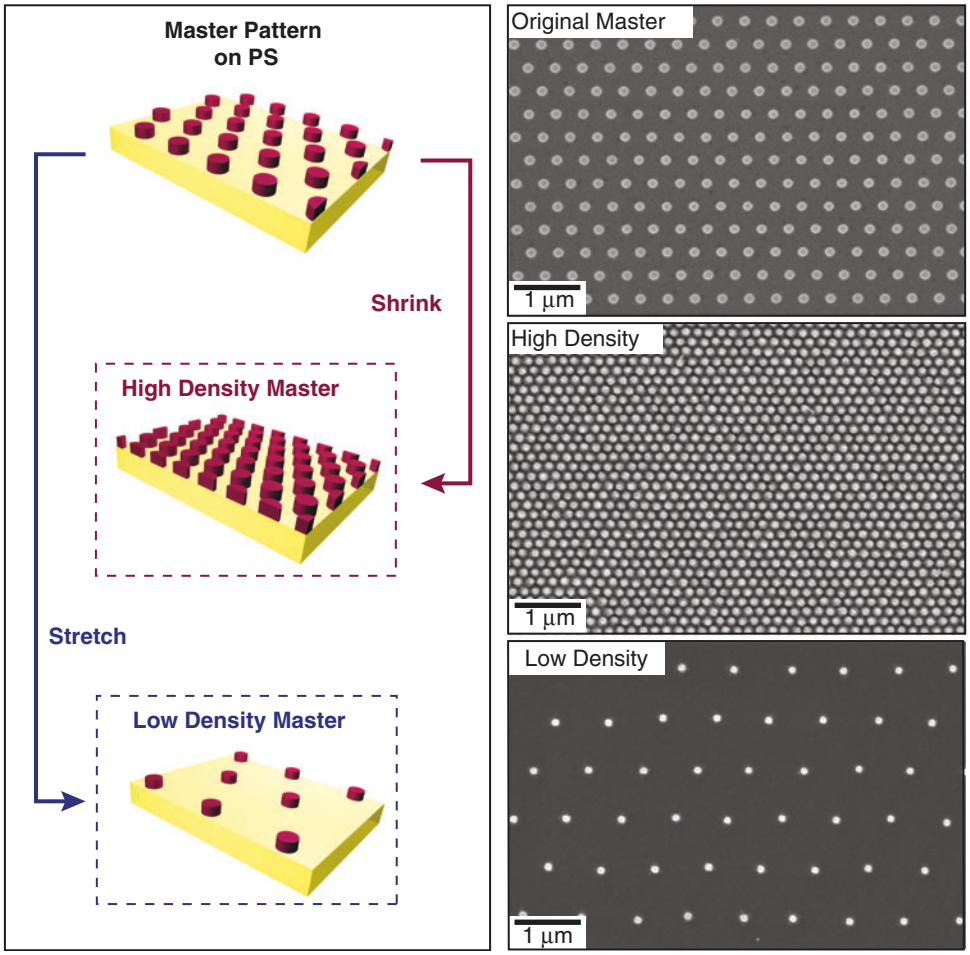


Shrinkable and Stretchable Nanomanufacturing, Fig. 2 *Manipulating the soft mask:* Solvent-assisted nanoscale embossing (SANE) involves the swelling of a PDMS mask with varying solvents during the replication process. The swelling of PDMS reduces the feature sizes of the patterned substrate without changing the periodicity of the master pattern. The patterned substrate with reduced feature size can act as a template for molding a new mask and repeating the process. Feature sizes were reduced through the replacement of the non-swelling solvent ethanol (*original master*) with high-swelling dimethylformamide (22 % reduction) and dichloromethane (42 % reduction) solvents

PDMS is one option for such substrates [7], the elastomer PDMS returns back to its original dimensions when the strain is removed. In contrast, thermoplastics such as polystyrene (PS) require no external force to maintain the stability of stretched or compressed substrates once heated and cooled above their glass transition temperature (T_g), representing an advantage over PDMS.

A stretchable and shrinkable manufacturing technique—inverse solvent-assisted nanoscale embossing (inSANE)—effectively SANE on thermoplastics, can maintain the patterned feature sizes dictated by the PDMS mask yet with tunable spacing and symmetry [6]. Features patterned with inSANE on a PS surface can undergo controlled shrinking to reduce feature periodicity by up to 50 % (Fig. 3). Importantly, this pattern with reduced spacing can now function as a new master for subsequent use in soft lithography. The major breakthrough of SANE/inSANE is that a single master pattern can now be used to create a near-infinite range of master molds with varying feature size, spacing, and orientation. No other existing nanopatterning methods can both prototype arbitrary patterns with small separation and scale for less than $100.

In addition, shrinkable and stretchable nanomanufacturing can also be incorporated into other nanofabrication techniques, such as patterning metals via deposition through a shadow mask. The size of the patterned metal is determined by the size of the gaps in the mask (limited to ca. 100 nm by photolithography). Replacing the mask material with a thermoplastic enables reduction of the gap size through controlled shrinking.



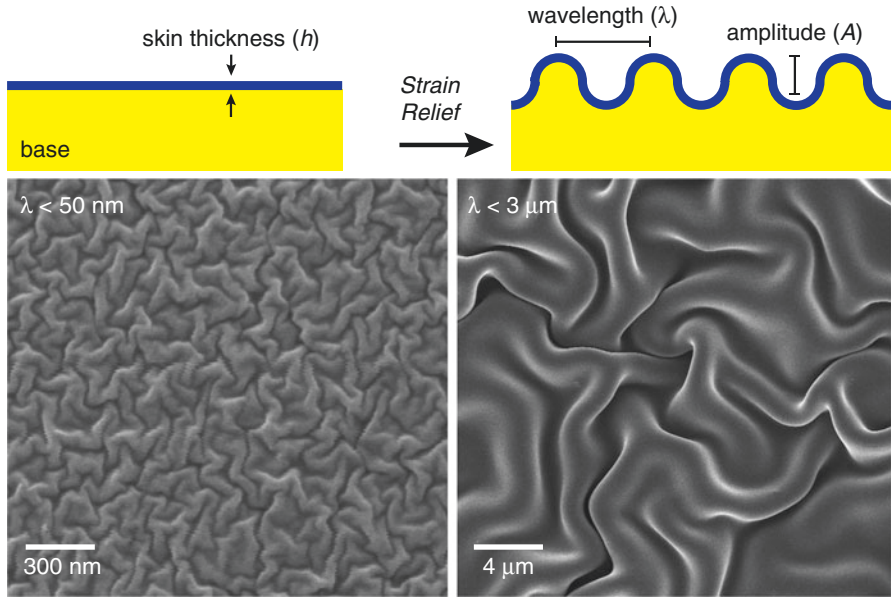
Shrinkable and Stretchable Nanomanufacturing, Fig. 3 *Post-processing of patterned substrates:* The master pattern is first replicated on a thermoplastic PS. After heating above T_g, the thermoplastic can be either shrunk to increase the pattern density or stretched to decrease the pattern density. In a single step, the master pattern periodicity (400 nm) is reduced to 200 nm (*high density*) by shrinking or increased to 800 nm (*low density*) by stretching

S

For example, the final gap size depends on the time and temperature at which the thermoplastic was heated above T_g and not the feature size of the initial slit patterning controlled by photolithography. As a result, controlling the shrinking of the mask enables the tuning of patterned metal features from 2 µm down to 21 nm [8].

Nanopatterns Without Needing a Mask

When a thin film (skin) bound to a soft layer (base) is compressed, out-of-plane buckles can form. The formation of such features is known as wrinkling, a spontaneous and mask-less process whose patterned area is scalable to hundreds of square meters. For comparison, as of 2015, the largest photomask available for purchase for fabricating PDMS masters is 32" by 32" with feature resolution down to 1 µm [9]. Useful as a mask-less method, the out-of-plane buckles are periodic, where the distance between features is characterized by an average wrinkle wavelength (λ). The mechanical properties of the skin and base prior compression dictate the final wrinkle λ. Changing the thickness of the skin and



Shrinkable and Stretchable Nanomanufacturing, Fig. 4 *Nanopatterns without needing a mask:* After applying strain relief to a stiff skin (*h*) on a soft base, periodic out-of-plane buckling will occur. The wrinkles are characterized by their wavelength (λ) and amplitude (*A*). The wrinkle wavelength of a polymer stiff skin grown on a thermoplastic substrate can be continuously tuned from 30 nm to several μm by changing the skin thickness prior to strain relief

skin/base materials can tune the wavelength from tens of nanometers to hundreds of meters. Wrinkles have been incorporated into applications like organic photovoltaics [9], flexible electronics [10], and fluorescence sensing of biomolecules [11], which benefit from large-area patterns.

The selection of materials for both the skin and base controls the geometric characteristics of the wrinkle patterns. For example, the deposition of thin metal (Au) films on a thermally expanded PDMS base can result in wrinkles with λ from 100 μm down to 4 μm upon the PDMS relaxing back to its normal size at room temperature. The thickness of the Au controls the final wrinkle λ. Unfortunately, during the physical deposition of Au onto the PDMS base, the top region stiffens, which ultimately limits the lower limit of skin thickness and resultant wavelengths of this system [12]. To overcome this limitation, the metal skin can be replaced with a silica-like film formed on PDMS by plasma-mediated growth. By using an oxygen plasma to control silica skin thickness, submicron wrinkles (λ ≈ 500 nm) were formed [12]. Ultraviolet/ozone (UVO) radiation (skin ≈ 5 nm) of the PDMS slab has resulted in a modified PDMS wrinkle systems with λ as low as 80 nm [12].

Alternative wrinkle systems have replaced the PDMS base layer with thermoplastics such as PS in order to expand the capabilities of wrinkle fabrication. For example, a Au film deposited on PS, when heated above T_g, can generate wrinkles with λ from several microns down to 250 nm, four times smaller than for Au of a similar thickness on PDMS [13]. These lower wavelengths, which approach the nanoscale regime, are possible as the Young's moduli ratio of skin and base (E_S/E_B) decreases. When the Au film is replaced with a polymer skin grown on a thermoplastic substrate, for example, wrinkle λ can be continuously tuned down to a lower limit of λ ≈ 30 nm (Fig. 4). The use of a polymer thin film for the stiff skin resulted in a 6 times reduction in E_S/E_B compared

to silica on PDMS [14]. Pushing the lower limits of nanowrinkle fabrication can be achieved by optimization of the mechanical properties of the skin and base.

Besides λ, the amplitude *A* plays an important role in determining the physical properties of a nanowrinkled substrate. Unlike other nanopatterns, wrinkles have a sinusoidal nature, where *A* is reduced when the vertical feature size is reduced. *A* can be maximized for a given wavelength by increasing the applied strain until a critical strain threshold is reached, after which nonlinear effects cause structural breakdowns and folds, tears, or self-similar wrinkles form [12]. The critical strain threshold limits *A* in most wrinkle systems, including PDMS, to ca. 0.1 λ. As a result, nanowrinkles ($\lambda \leq 100$ nm) are limited to *A* less than 10 nm. Alternative wrinkle material systems can overcome this limitation. For example, the polymer skin grown on thermoplastic PS has advantages since both the polymer skin and base can mitigate nonlinear effects at high strains, increasing the critical strain threshold to over 50 %. As a result, wrinkles with $A > \lambda$ can be formed, which is a tenfold increase in amplitude over PDMS wrinkle systems with similar λ [15].

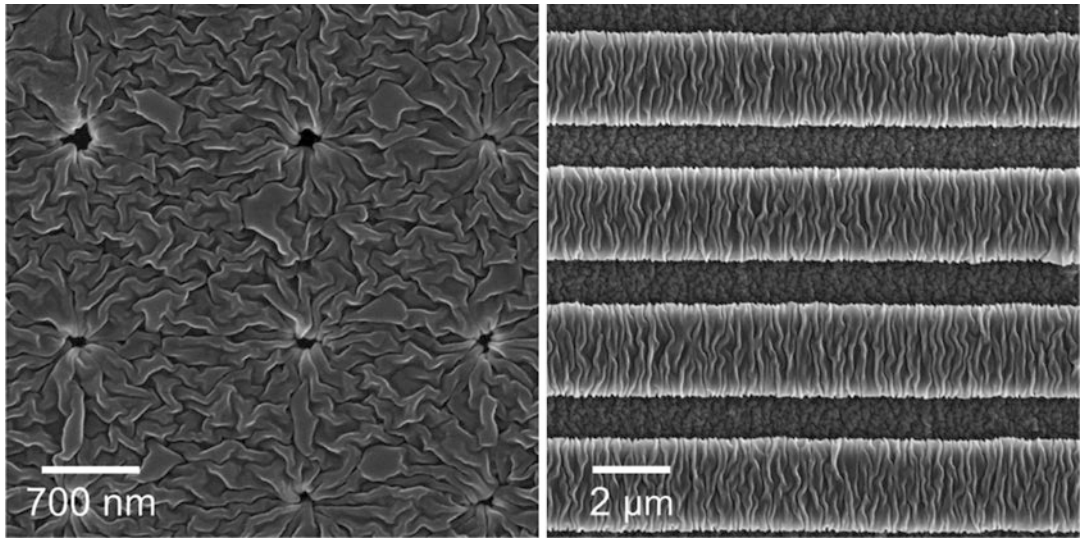
Hybrid Mask and Mask-Free Techniques

Hybrid nanofabrication approaches combine two or more techniques to produce nanopatterns not feasible with a single method and, importantly, can also overcome limitations of the individual techniques. For example, a drawback of wrinkles is that they are disordered within a characteristic wavelength; however, patterning the stiff skin prior to shrinking will orient the wrinkles perpendicular to a patterned feature [12]. This controlled order is a result of nonuniform local strain at the patterned/unpatterned interface; wrinkles will remain ordered for distances 20λ–100λ away from the patterned feature [12, 15]. Disrupting the local strain to introduce order will have a negative effect on the wrinkle *A* and structure. As a result, the ratio between patterned feature widths (*w*) and wrinkle λ must be controlled to optimize the final wrinkle structure. At the micron scale, patterning lines with 200 μm to 500 μm *w* into a Au film prior to applying compressive force produced ordered wrinkles with wavelengths of 20–50 μm [12]. To order sub-100 nm wrinkles, inSANE was combined with wrinkling on PS. The inSANE pattern acted as a mask for the growth of the patterned polymer skin while the thickness of the polymer skin determined the final wrinkle λ, resulting in ordered 50 nm wrinkles [15]. The *A* and wrinkle order was measured with varying line patterns and skin thicknesses to optimize the relationship between λ and *w*. It was found that as *w* approached 20λ, wrinkles would become ordered without any observable decreases in *A*. As *w* was further decreased to 2λ, the strain manipulation became large enough to collapse the wrinkles into folds (Fig. 5). Since nanowrinkles and nanofolds spontaneously form across the entire surface, a hybrid scheme provides a parallel method to create complex ordered and disordered patterns with nanoscale spatial resolution.

Hybrid nanofabrication techniques can also be used to form hierarchical patterns. Multi-scale structures are abundant in nature; macroscale features impart mechanical stability, and nanoscale features control functionality such as wettability, optical properties, and adhesion [13]. Control over features at all length scales is key to fabricating engineered systems that can mimic or eventually outperform natural systems. Physical imprinting followed by swelling of a soft polymer produced a hierarchical pattern with nanoscale lines (500 nm) and macroscale wrinkle features ($\lambda = 15$–35 μm) [16]. In addition to forming a hierarchical structure, controlling the embossing depth and direction of the nanoscale lines tuned the previously uncontrollable periodicity and orientation of the macroscale wrinkles.

Future of Shrinkable and Stretchable Nanomanufacturing

Shrinkable and stretchable strategies enable unique applications such as the manipulation of



Shrinkable and Stretchable Nanomanufacturing, Fig. 5 *Hybrid mask and mask-free techniques:* Combination of inSANE and nanowrinkle fabrication results in ordered nanowrinkles. The shape and periodicity of the nanoembossed pattern determine the degree of wrinkle order. With a single patterned feature, the wrinkles orient radially around the hole. When the hole is replaced with a line with a width less than 20λ, wrinkles become completely ordered

cell adhesion with nanopatterns [17], the formation of large-area antibacterial surfaces [18], and the fabrication of bioinspired hydrophobic surfaces [13]. Designs for these emerging fields require patterning techniques that can control feature shapes and sizes from 10^{-9} to 10^{-3} m while simultaneously reducing costs associated with nanopatterning [13], a major strength of shrinkable and stretchable nanomanufacturing. The development of simple yet transformative approaches that can achieve independent control over feature size, periodicity, and orientation of patterns is the future of nanomanufacturing.

Cross References

- ▶ Flexible Electronics
- ▶ Mechanical Properties of Nanocrystalline Metals
- ▶ Micro- and Nanomanipulation for Nanomanufacturing
- ▶ Microcontact Printing
- ▶ Nanoimprinting
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanotechnology
- ▶ Reliability of Nanostructures

References

1. Xia, Y., Rogers, J.A., Paul, K.E., Whitesides, G.M.: Unconventional methods for fabricating and patterning nanostructures. Chem. Rev. **99**, 1823–1848 (1999)
2. Basting, D., Marowsky, G.: Excimer laser technology. Springer Science & Business Media, Berlin (2005)
3. Qin, D., Xia, Y., Whitesides, G.M.: Soft lithography for micro- and nanoscale patterning. Nat. Protoc. **5**, 491–502 (2010)
4. Gates, B.D., Whitesides, G.M.: Replication of vertical features smaller than 2 nm by soft lithography. J. Am. Chem. Soc. **125**, 14986–14987 (2003)
5. Lee, J.N., Park, C., Whitesides, G.M.: Solvent compatibility of poly(dimethylsiloxane)-based microfluidic devices. Anal. Chem. **75**, 6544–6554 (2003)
6. Lee, M.H., Huntington, M.D., Zhou, W., Yang, J.-C., Odom, T.W.: Programmable soft lithography: solvent-assisted nanoscale embossing. Nano Lett. **11**, 311–315 (2011)
7. Xia, Y., Whitesides, G.M.: Reduction in the size of features of patterned SAMs generated by microcontact printing with mechanical compression of the stamp. Adv. Mater. **7**, 471–473 (1995)

8. Zhang, B., Zhang, M., Cui, T.: Low-cost shrink lithography with sub-22 nm resolution. Appl. Phys. Lett. **100**, 133113 (2012)
9. Kim, J.B., Kim, P., Pégard, N.C., Oh, S.J., Kagan, C. R., Fleischer, J.W., Stone, H.A., Loo, Y.-L.: Wrinkles and deep folds as photonic structures in photovoltaics. Nat. Photon **6**, 327–332 (2012)
10. Rogers, J.A., Someya, T., Huang, Y.: Materials and mechanics for stretchable electronics. Science **327**, 1603–1607 (2010)
11. Sharma, H.; Digman, M. A.; Felsinger, N.; Gratton, E.; Khine, M.: Enhanced emission of fluorophores on shrink-induced wrinkled composite structures. Opt Mater Express **4**, 753–763 (2014)
12. Genzer, J., Groenewold, J.: Soft matter with hard skin: from skin wrinkles to templating and material characterization. In Soft Matter **2**, 310 (2006)
13. Bae, W.-G., Kim, H.N., Kim, D., Park, S.-H., Jeong, H.E., Suh, K.-Y.: 25th Anniversary article: scalable multiscale patterned structures inspired by nature: the role of hierarchy. Adv. Mater. **26**, 675–700 (2013)
14. Huntington, M.D., Engel, C.J., Hryn, A.J., Odom, T. W.: Polymer nanowrinkles with continuously tunable wavelengths. ACS Appl. Mater. Interfaces **5**, 6438–6442 (2013)
15. Huntington, M.D., Engel, C.J., Odom, T.W.: Controlling the orientation of nanowrinkles and nanofolds by patterning strain in a thin skin layer on a polymer substrate. Angew. Chem. Int. Ed. **53**, 8117–8121 (2014)
16. Li, Y.; Dai, S.; John, J.; Carter, K. R.: Superhydrophobic surfaces from hierarchically structured wrinkled polymers. ACS Appl. Mater. Interfaces **5**, 11066–11073 (2013)
17. Yang, P., Baker, R.M., Henderson, J.H., Mather, P.T.: In vitro wrinkle formation via shape memory dynamically aligns adherent cells. Soft Matter **9**, 4705–4714 (2013)
18. Freschauf, L.R., McLane, J., Sharma, H., Khine, M.: Shrink-induced superhydrophobic and antibacterial surfaces in consumer plastics. PLoS One **7**, e40987 (2012)

Si Nanotubes

► Physical Vapor Deposition

Silent Flight of Owls

► Silent Owl Wings

Silent Owl Wings

Thomas Bachmann[1] and Hermann Wagner[2]
[1]Institute for Fluid Mechanics and Aerodynamics, Technische Universität Darmstadt, Darmstadt, Germany
[2]Institute for Biology II, RWTH Aachen University, Aachen, Germany

Synonyms

Silent flight of owls

Definition

Wings of owls are equipped with macro- and microstructured devices that reduce flight noises significantly. These birds of prey hunt in both twilight and darkness. Since visual information is limited at that time of day, owls use acoustic information to detect and track potential prey. The silent flight affords the detection of prey in flight and makes the owl inaudible for its prey. Different wing and feather specializations, such as serrations at the leading edge of the wing, a velvet-like upper surface and fringes along the inner vanes, influence the airflow and thus are important for the reduced flight noise in owls.

Overview

Fundamentals of Bird Flight

In the course of evolution, birds have conquered the air for locomotion. A great variety of wing geometries have evolved for different needs of birds. Wings can be long or short, narrow or broad, thick or thin. Furthermore, wings can be pointed or have a rounded tip, which may be smooth or fingered. Considering this variety, each wing appears to be adapted to the distinct flight conditions of the particular species and the ecological niche it uses. However, all wings have at least one aerodynamic feature in common: They are cambered to the dorsal side. Only curved

wings are able to produce enough lift at moderate angles of attack and low flight speeds to allow a bird to become airborne. Bird wings are mostly cambered in proximal regions. Thickness as well as camber decrease towards the wing tip. The resulting reduction of mass at the wing tip causes a decrease of inertia to allow high wing-beat frequencies (2–12 Hz) [1] at low energy consumption [2].

A bird has to defy gravity to become airborne. Hence, flight consumes a lot of energy, especially during takeoff and landing [2]. High lift is achieved either by an intense beating of the wings or by forming wings with high-lift properties [3]. Such wings are characterized by a high camber and a large wing area. Apart from gravity, birds have to deal with an additional force, the drag. Drag is the resistance of the body and wings while moving through the air. It is comprised of three components [3]: first, the friction drag between the airflow and the surface, second, the form drag of the bird's body and the leading edges of the wings, and third, the induced drag. The induced drag is the component of drag force on the bird's wing that is caused by an induced downward velocity. The drag is also influenced by the microstructure of the plumage. Several anatomical specializations of feathers have been reported that influence the airflow around the bird wing, most of them in owls [4–8].

Function of Silent Flight in Owls

Owls have an almost global distribution, which is reflected in the variety of species and subspecies. These birds spend a high proportion of their time searching for prey, either by sitting on a perch, or flying slowly over fields and vegetation [9, 10]. The prey of most owls is mainly active at night. Since visual information is limited in low-light conditions, owls use acoustical cues to detect their prey. The prey emits noise by rustling through the vegetation or through intra- and interspecific communication sounds which can be detected by the owl even in absolute darkness [11]. Several anatomical and behavioral adaptations of the hunting strategy have been discovered in owls. One conspicuous specialization in many owl species is the facial ruff with its specialized feathers that direct sound toward the ears like a dish antenna. The facial ruff is also responsible for the low hearing threshold [12]. Asymmetrically arranged ear flaps in front of the outer ear canals help to localize the elevation of sound sources. Furthermore, nuclei in the brain that process acoustic information are enlarged compared with other similar-sized birds [13]. With this set of adaptations, owls are able to detect and localize their prey precisely. Nevertheless, to be able to utilize this ability during flight, owls should fly silently. If that would not be the case, noise emitted by prey would be masked by the owl's own flight noise and the prey might be warned by the approaching owl. Consequently, owls rely on a silent flight, which is achieved by adaptations of the wing, the plumage, and the flight behavior [4–10].

Specializations of the Owls' Plumage

Feathers are the main aerodynamic components of a bird's wing. Interaction with the airflow during flight and protection against cold, heat, and wetness are essential functions. Furthermore, feathers are used for display or camouflage. A feather consists of a central shaft (rachis) and two laterally attached vanes. The feather vane is made up of parallel barbs that branch off from either side of the rachis. Barbs are interlinked via hook and bow radiates to form a closed surface. While each feather quill is inserted into the integument, all flight feathers (remiges) are additionally associated with the underlying skeleton. Secondary remiges are connected to the posterior edge of the ulna; primary remiges are supported by bones of the hand and fingers. Coverts arise from the anterior integument membrane, forming smooth and closed upper and lower wing surfaces. All wing feathers together form a wing that is light and flexible, but strong enough to meet the bird's requirements to fly.

In owls, several microstructured plumage specializations are known that influence the flow over the wing [4–8]: The dorsal surface of the feathers has an increased roughness caused by elongations of hook radiates, termed pennula. Those feathers that form the leading edge of the distal wing are equipped with serrations at their outer vanes.

Every feather of the wing is surrounded by a fringed structure which is due to unconnected barb endings. Furthermore, owl feathers are very flexible and pliant due to a reduced number of radiates and thus a lower number of interlinks between the barbs compared to other birds [8]. Finally, cross-sectional profiles of the rachises are less structured compared to other birds of similar size. These plumage specifications have been implicated in noise suppression during flight by damping noise above 2 kHz [5]. Thus, flight noise is reduced within the typical hearing spectrum of the owls' prey (>3 kHz) [5] and within the owl's own best hearing range (5–9 kHz) [14]. The reduction of noise is mainly achieved by airflow control and friction reduction of single feathers.

Basic Methodology

Morphometrics

Bird wings, together with many other parameters, have been morphometrically characterized in order to classify bird species and to create a systematic order within birds (*Aves*). However, aerodynamic parameters are not included in this approach. Aerodynamic properties such as wing span, wing area, and aspect ratio can be measured in living birds or well-preserved museum specimens. Camber and thickness distribution are more prone to errors due to drying processes or unnatural fixation. Therefore, anesthetized or recently deceased birds yield better results. Best results are gained from experiments with living birds. Hereby, the wing geometry is obtained during free-flight experiments. Computer-aided analyses of three-dimensional reconstructions of the wings provide high-resolution profile data.

Free-flight experiments are conducted to measure the noise emission of flying owls in different flight modes while the birds fly over an extreme sensible microphone array at a defined flight speed and height. The measured noise spectrum can then be correlated with the flight parameters. Finally, the results are compared to other non-silently flying birds [15].

Substructures of the feathers are investigated with different digital imaging techniques having high spatial resolutions. Light microscopy and scanning electron microscopy (SEM) enable to some extent two-dimensional measurements and morphometric characterizations. Confocal laser scanning microscopy (CLSM) and micro-computed-tomography (μ-CT) allow the precise digitization of the three-dimensional shape of microstructures. In all cases, the data are then processed and analyzed with adequate visualization and measuring software to determine relevant morphometric and aerodynamic parameters. After specifying the shape and geometry of the owl-specific structures, these can be artificially manufactured and applied to wing models for further analysis, e.g., in wind tunnel experiments. By this means, their aerodynamic function may be clarified.

Behavioral Studies

Birds use their wings to maneuver through the air. The form and geometry but also the movements of the wings are important for an efficient locomotion. Hence, studying the flight behavior of birds is one fundamental piece of the large puzzle of understanding how bird wings work. The study of wing motion in different flight modes, ideally in combination with flow visualization, helps to understand the functions of each wing component. Flight speed, wing-beat frequency, and wing-beat amplitude are some parameters that influence the flight of birds. Different flight modes like flapping flight, gliding, or soaring exist. Furthermore, some birds adapt their flight behavior during hunting, escaping, courtship, and other modes. Nowadays, high-speed cameras outperform traditional bird watching with binoculars and cameras in many aspects. Images with high temporal and spatial resolution of wing and feather movements during beating of the wings allow a much better investigation of each wing element and its aerodynamic function. Such techniques are used in field observations or wind tunnel experiments with living birds.

Wind Tunnel Experiments

Fixed wings of birds and artificial models of bird wings are investigated in wind tunnel

experiments. Performing experiments in wind tunnels guarantee reliable and repeatable results. The use of wing models instead of living animals allows the manipulation of certain parameters of the wing. The impact of several wing geometries can then be compared and general conclusions can be drawn based on these results. Flow visualization techniques such as particle-image velocimetry (PIV), oil-flow pattern or pressure measurements are carried out to understand the influence of the wing profiling and each wing element on the airflow. In case of owl wings, the specific feather adaptations are added separately on a wing model or are removed step by step from fixed owl wings. Their potential function is revealed by comparing wings either with or without the owl's specializations in fluid-mechanical experiments.

Key Research Findings

Owl Wings

Owl wings differ from those of other birds. Figure 1a shows an example wing of a barn owl (*Tyto alba*). The large wing area, in combination with a relatively low body mass, results in a low wing loading allowing the owl to fly slowly and to carry large prey. Specific camber and thickness distributions (Fig. 1b, c) are also characteristic for owl wings [6]. While the proximal wing is highly cambered and anteriorly thickened, the distal wing is thin and less cambered. Anatomical studies revealed the construction plan of the wing (Fig. 1d). Skeletal elements, the muscle mass, and the coverts are responsible for the thick anterior part of the wing. By contrast, the distal wing and the posterior arm part of the wing are formed by remiges. These areas are extremely thin and lightweight. Hand and arm part of the wing have different functions (lift production and thrust) expressed by different geometries and profiling. At both wing parts, however, the airflow tends to separate in wind tunnel experiments when tested with a smooth surface without any surface or edge modifications [5, 16]. Owls, however, evolved microstructures on their feathers to influence the airflow around the wing in such a way that the airflow stays attached especially during critical flight maneuvers [5, 16] such as takeoff, landing, or striking. Hence, the wings produce a huge amount of lift even at low flight speeds.

Owl Feathers

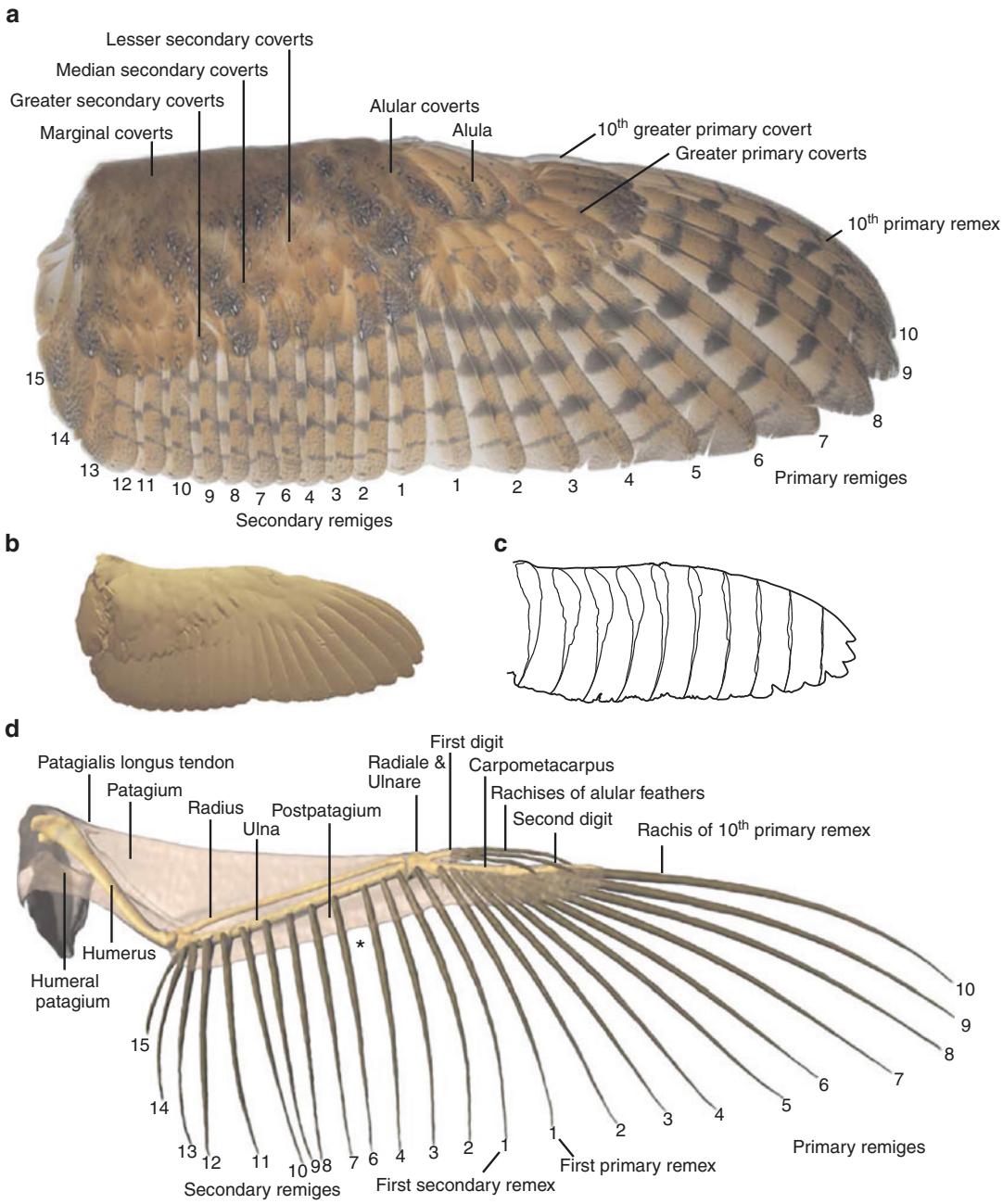
Owl feathers share the anatomy of typical contour feathers [8]. However, they differ in some detail. As feathers are the main aerodynamic components of the bird wing, their geometry and dimensions affect the overall wing geometry. In owls, the feather vanes are large and at the same time extremely light and flexible. Fewer interlinks of the barbs cause these effects [8]. Hereby, owl feathers are very air-transmissive and flexible, whereby they can react rapidly to varying airflow conditions. Their bending behavior is passively driven. It depends on the material properties of feather keratin and the geometry of the rachis and the barbs, respectively [17].

Compared to other birds, owls stand out due to a large count of contour feathers [9]. Specifically, head, feet, and body are densely covered. This upholstery of the body serves the suppression of noise to some extent, but also thermal insulation since owls have only small fat reserves. The coloration of the plumage is mostly adapted to the environment of the relevant owl species with respect to an effective camouflage, but also to sexual display to some extent. Snowy owls (*Bubo scandiacus*), for instance, have a white plumage, while species that live in wooded environment have a brownish patterned plumage (see Fig. 2). Different color intensities between males and females are found, for instance, in barn owls with the females being slightly darker.

The most conspicuous attribute of owl remiges are several anatomical microstructures that are responsible for the silent flight [4, 8]. Interestingly, fishing owls like the Malay fish owl (*Bubo ketupu*) lack such feather specialization [4]. Fish cannot hear the approaching owl. Hence, these owls do not need to fly silently.

Microstructures of Owl Feathers

Owls evolved several surface and edge modifications of their wings that are in turn specializations



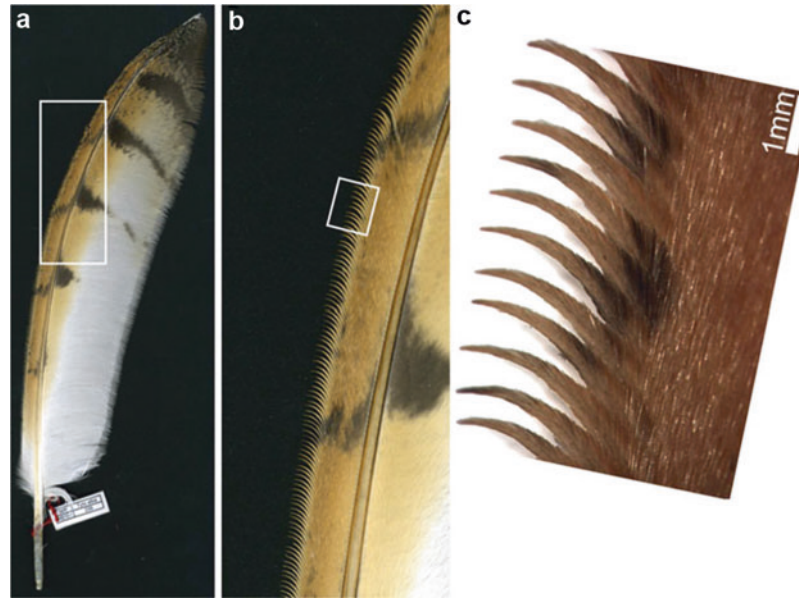
Silent Owl Wings, Fig. 1 Anatomy of a barn owl (*Tyto alba*) wing: (**a**) Topography of the dorsal wing (photograph). (**b**) Surface image of the digitized wing (surface scan). (**c**) Profiles of the owl wing in steps of 10 % of the half wing span. (**d**) 3D reconstruction of skeletal elements, skin, and rachises of the remiges (processed CT scan). *The fifth secondary remex is missing in all owl species

of the plumage. Three structures are mainly worth mentioning: the leading-edge comb-like serrations, the velvet-like upper surface, and the inner vane fringes [4, 7, 8].

Each feather that functions as a leading edge of the wing is equipped with comb-like serrations (Fig. 3). The number of remiges forming the leading edge varies among owl species, but is in the

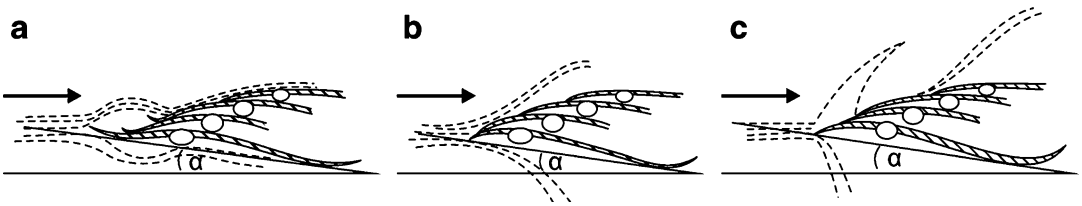


Silent Owl Wings, Fig. 2 *F*eathers of different owl species: (**a**) Tenth primary remex of a barn owl (*Tyto alba*) wing. The outer vane is equipped with tiny serrations, the inner vane is lined with fringes. (**b**) Contour feather from the body of a snowy owl (*Bubo scandiacus*). Note the symmetry of the vanes in comparison to the flight feathers (**a**) and (**c**) and the elongated plumulaceous barbs (fringes at the end of the vanes). (**c**) Primary remex of an eagle owl (*Bubo bubo*). As the eagle owl is the largest owl, so are its feathers. Scale bars represent 2.5 cm



Silent Owl Wings, Fig. 3 Serrations at the leading edge of a tenth primary of a barn owl (*Tyto alba*) in different magnifications: Barb endings separate and bow toward the *upper side* (**a**, **b**, **c**). Note also the fringes at the inner vane (**a**) and the dorsal surface texture

range of 1–3 remiges. Barb endings separate and bend upward, changing their form and function. Differences in the anatomical phenotype are small within one feather, though small variations occur (Fig. 3) [8]. While long serrations with small spacing are found in basal and central regions of the feathers, shorter serrations with a larger distance are located at the tip of the feather.



Silent Owl Wings, Fig. 4 Scheme of the airflow at the leading edge of a distal bird wing: (**a**) Intact wing of a tawny owl (*Strix aluco*) in the region of the serrations. Note the laminar flow conditions around the wing. (**b**) After manual removing of the serrations at the wing of the tawny owl, the flow separates early. (**c**) Flow conditions around a wing of a mallard duck (*Anas platyrhynchos*). The *arrow* indicates the flow direction; alpha is the angle of attack (Drawings after Neuhaus et al. [5])

Functionally, serrations affect the airflow and noise production only marginally during cruise flight [5]. The stagnation pressure at the leading edge of the wing prevents a functional airflow around the serrations. However, in critical flight maneuvers, such as landing or striking, the angle of attack is much higher and the serrations cause the airflow to stay attached to the wing. This is achieved by dividing the airflow into many small micro-vortices that extent chordwise over the upper wing surface. As a consequence, the boundary layer is more energetic and flow separation is prevented (Fig. 4a). By this means, lift production is maintained even at high angles of attack and low flight speed. This effect gets lost after manual removal of the serrations at the leading edge (Fig. 4b). Here, the flow field around the wing resembles that of a mallard duck (Fig. 4c) which produces much more noise [5].

The second specialization is found on the upper wing surface. Each feather of the owl's plumage is covered with a velvet-like dorsal texture (Fig. 5). This structure is formed by elongations of the hook radiates. Each hook radiate terminates in a filament structure, termed pennula. Due to their large number, the surface becomes very fluffy and porous. Functionally, two aspects are at least associated with this structure.

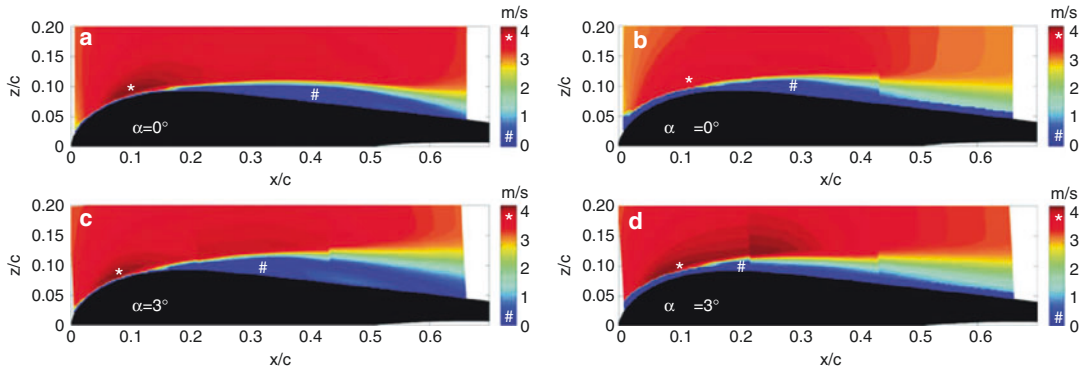
On the one hand, the velvet-like dorsal surface of the inner vane, which is mostly covered by the adjacent feather, is very thick and well developed. The pennula are long and interconnect at large angles with the feather's surface. This in turn causes a friction-reductive device on the upper surface and enables a smooth and silent gliding of the feathers. In all other birds investigated so far, the parallel-oriented barbs of adjacent feathers rub against each other producing a well-noticeable high-frequency sound, especially during flapping flight. In owls, however, such noises are prevented by the velvety surface. Those noises that still occur are dampened by the porous feather texture acting like acoustic absorbers.

On the other hand, the velvet-like structure is also found at the outer vane. Here, the texture affects the aerodynamics of the wing. As mentioned above, the airflow tends to separate at wings with the specific camber and thickness distribution of owls and a smooth wing surface [16]. Figure 6 shows the PIV results of the velocity distribution of the flow around an owl-based airfoil at Reynolds number 40,000, two different angles of attack (0° and 3°), and two different surfaces. The velocity of the flow is color coded. At the wing with a clean and smooth surface (Fig. 6a, c), a separation bubble occurs on the suction side that increases with the higher angle of attack and leads to a complete flow separation beyond 3° angle of attack (not shown) [16].

Increasing the surface roughness as it is found in owls is one means in the direction of flow control. An artificial velvety surface texture applied to the owl-based wing geometry has a dramatic influence on the flow field (Fig. 6b, d). The velvety surface shifts the separation toward higher angles of attack indicated by a smaller separation bubble in comparison to the clean wing model. The boundary layer of the wing is controlled in such a way that occurring separation bubbles are reduced in size and shifted toward the



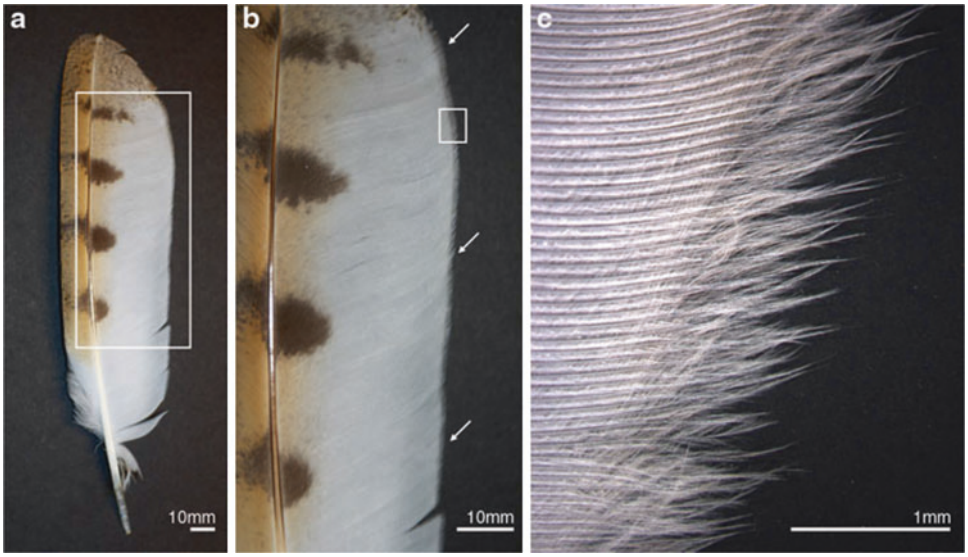
Silent Owl Wings, Fig. 5 Velvet-like dorsal surface of a barn owl's fourth secondary in different magnification: Elongations of the hook radiates create this filigree texture which is supposed to influence the airflow and to reduce friction noise. (**a**, **b**) Photographs, (**c**) scanning electron microscope image



Silent Owl Wings, Fig. 6 Averaged velocity fields at Re = 40,000 and an angle of attack of 0° and 3° for a clean wing (**a**, **c**) and the same wing geometry with a velvety surface (**b**, **d**). The velocity of the airflow is *color* coded (online version)

leading edge of the wing. Consequently, the airflow is stabilized even at low flight speeds [16].

A further specialization can be found at the inner vanes of the remiges. Here, barb endings separate by a reduction of the hooklets of the hook radiates. Barbs are no longer interconnected and fringes are formed (Fig. 7). Since the tips of the inner vanes of remiges form



Silent Owl Wings, Fig. 7 Fringes at the inner vane of a barn owl's fifth primary in different magnification: Hooklets of the hook radiates are absent causing the barbs to separate. Hook and bow radiates change their length and thereby support the formation of fringes. (**a**) and (**b**) dorsal, (**c**) ventral

the trailing edge of the wing, fringes are also found here. The appearance of fringes is very fluffy and unstable in shape. They are made up of three components: the central barb shaft and the laterally attached hook and bow radiates. The radiates support the formation of fringes by a parallel orientation to the barb shafts [8]. Since the barb shaft does not change its shape, as is the case for the serrations, the structure is pliant and the fringes can float freely resulting in a thin, almost two-dimensional structure. Functionally, fringes are assumed to decrease the sound intensity by a reduction of turbulence at the trailing edge of the wing [7, 18]. While sharp edges of wings and flat plates are known to produce a well-noticeable noise even at low Reynolds numbers [19], fringed trailing edges of owl remiges prevent such phenomena. Each wing feather has its own trailing edge that produces noise during flight independently from being part of the wing or separated, for example, during flapping flight. In general, fringes act as a pressure release mechanism that interrupt scattering phenomena which can be found when turbulent eddies pass over the posterior wing region that gradually transitions from the loaded wing to the freestream conditions [18, 19]. Without such fringes, turbulent boundary layers being convected past the trailing edge into the wake would generate an intense, broadband scattering noise. In owls, fringes prevent sharp edges at the trailing edge such that the turbulent boundary layer is merged smoothly into the airflow.

Behavioral Adaptations

Nocturnal owls are unable to benefit from thermal upwind like other birds of prey. Hence, owls combine active flapping flight and gliding phases [9, 10]. The large wing area and the specific profiling of the wing enable the owl to glide and maneuver with little movements of the wing. The flight speed is relatively low and depends on the illumination of the environment. Barn owls (*Tyto alba*), for instance, can reduce their flight speed down to 2.5 m/s [9, 10] which per se leads to a silent flight, since only little turbulences occur at that velocity. Apart from the slow gliding flight, some adaptations in the motion of the owl already prevent rising noises. During hunting, owls reduce their wing beat frequency and amplitude [9]. Consequently, less friction between feathers occurs during flight which would cause

rubbing noises. Those noises that still rise are efficiently dampened by the porous upper wing surface [4, 7].

Future Directions for Research

Urbanization and the increasing demand for air freight and renewable energy have led to a dramatic increase of noise pollution in the last decades. At ground level, airplanes produce an intense noise by their engines and high-lift configurations of the wings. Furthermore, wind farms are constantly upgraded. Their noise emission does not only annoy humans, but also disturbs migratory animals such as birds, or in case of offshore wind farms, whales, and fishes. How flight noise can efficiently be reduced is demonstrated by owls. These birds evolved noise-reduction and noise-suppression devices in the course of evolution. The implementation of the underlying mechanisms in modern aircraft and rotors of wind-power plants would help to reduce noise pollution. Until then more studies and experiments have to be carried out. In joint biological and fluid-mechanical efforts, the basic aerodynamic characteristics and the fluid-mechanical details that define the drag and noise reduction mechanisms of owl wings have to be investigated in more detail. This will be done using preserved material from zoological collections, experiments with living owls, and wing models in wind tunnel experiments. The knowledge gained will be used to create devices that operate at higher Reynolds numbers, since copying the natural structure alone would not deliver satisfying results. In the long run, designing silent wings of modern aircraft will bring us closer to the goal of a quieter environment.

Cross-References

- ▶ Biomimetic Flow Sensors
- ▶ Confocal Laser Scanning Microscopy
- ▶ Friction-Reducing Sandfish Skin
- ▶ Insect Flight and Micro Air Vehicles (MAVs)
- ▶ Scanning Electron Microscopy
- ▶ Shark Skin Drag Reduction

References

1. Pennycuik, C.J.: Wing beat frequency of birds in steady cruising flight: new data and improved predictions. J. Exp. Biol. **199**, 1613–1618 (1996)
2. Rayner, J.M.V.: On aerodynamics and the energetics of vertebrate flapping flight. Contemp. Math. **141**, 351–400 (1993)
3. Rüppell, G.: Vogelflug. Rowohlt Taschenbuchverlag GmbH, Hamburg (1980)
4. Graham, T.: The silent flight of owls. J. Roy. Aero. Soc. **38**, 837–843 (1934)
5. Neuhaus, W., Bretting, H., Schweizer, B.: Morphologische und funktionelle Untersuchungen über den lautlosen Flug der Eule (*Strix aluco*) im Vergleich zum Flug der Ente (*Anas platyrhynchos*). Biol. Zentr. Bl. **92**, 495–512 (1973)
6. Nachtigall, W., Klimbingat, A.: Messungen der Flügelgeometrie mit der Profilkamm-Methode und geometrische Flügelkennzeichnung einheimischer Eulen. In: Nachtigall, W. (ed.) Biona Report 3, pp. 45–86. Gustav Fischer Verlag, Stuttgart/New York (1985)
7. Lilley, G.: A study of the silent flight of the owl. AIAA Paper, pp. 98–2340 (1998)
8. Bachmann, T., Klän, S., Baumgartner, W., Klaas, M., Schröder, W., Wagner, H.: Morphometric characterisation of wing feathers of the barn owl (*Tyto alba*) and the pigeon (*Columba livia*). Front. Zool. **4**, 23 (2007)
9. Mebs, T., Scherzinger, W.: Die Eulen Europas. Franckh-Kosmos, Stuttgart (2000)
10. Taylor, I.: Barn Owls: Predator – Prey Relationship and Conservation. Cambridge University Press, Cambridge (1994)
11. Knudsen, E.: The hearing of the barn owl. Sci. Am. **245**(69), 113–125 (1981)
12. Hausmann, L., Campenhausen, V.M., Endler, F., Singheiser, M., Wagner, H.: Improvements of sound localization abilities by the facial ruff of the barn owl (*Tyto alba*) as demonstrated by virtual ruff removal. PLoS One **4**(11), e7721 (2009)
13. Konishi, M., Takahashi, T., Wagner, H., Sullivan, W., Carr, C.: Neurophysiological and anatomical substrates of sound localization in the owl. In: Auditory Function – Neurobiological Bases of Hearing, pp. 721–745. Wiley, New York (1988)
14. Konishi, M.: How the owl tracks its prey. Sci. Am. **61**, 414–424 (1973)
15. Sarradj, E., Fritzsche, C., Geyer, T.: Silent owl flight: bird flyover noise measurements. AIAA Paper 2010–3991 (2010)

16. Klän, S., Bachmann, T., Klaas, M., Wagner, H., Schröder, W.: Experimental analysis of the flow field over a novel owl based airfoil. Exp. Fluids **46**, 975–989 (2008)
17. Bonser, R.H., Purslow, P.: The young's modulus of feather keratin. J. Exp. Biol. **198**, 1029–1033 (1995)
18. Herr, M.: Experimental study on noise reduction through trailing edge brushes. Note. N. Fl. Mech. Mul. D **92**, 365–372 (2006)
19. Moreau, D.J., Brooks, L.A., Doolan, C.J.: On the aeroacoustic tonal noise generation mechanism of a sharp-edged plate. JASA E. L (2011). doi:10.1121/1.3565472

Silica Gel Processing

▶ Sol–Gel Method

Silicon Microphone

▶ Electrostatic MEMS Microphones

Silicon Thin Films

▶ Heat Transfer in Semiconductor Nanostructures

Silks of Spiders as Model Bio-polymers

▶ Spider Silk

Silver (Ag)

▶ In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles

Simulating Nanoscale Heat Transport

Giuseppe Romano[1], Jean-Philippe M. Peraud[2] and Jeffrey C. Grossman[1]
[1]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Synonyms

Modeling nanoscale heat transport

Introduction

Heat conduction has been modeled for almost two centuries by the well-known Fourier's law. Jean-Baptiste Joseph Fourier stated that "the quantity of heat which flows uniformly, during unit of time, across unit of surface taken on any section whatever parallel to the sides, all other things being equal, is directly proportional to the difference of the extreme temperatures, and inversely proportional to the distance which separates these sides" [1]. Fourier's law can be conveniently written in its local form:

$$\mathbf{J} = -\kappa \nabla T, \tag{1}$$

where κ is the thermal conductivity (for silicon, it is about 150 $Wm^{-1}\ K^{-1}$), T is the lattice temperature, and $\mathbf{J}$ is the heat flux. Equation 1 together with the continuity equation for thermal flux

$$\rho C_p \frac{\partial T}{\partial t} + \nabla \cdot \mathbf{J} = H, \tag{2}$$

gives the well-known diffusive heat conduction equation. In Eq. 2, C_p is the specific heat capacity, ρ the material density, and H the sum of all the heat sources in the system.

While Fourier's law is valid at the macroscale, it has been experimentally proved that in some nanostructured materials, such as nanowires [2], thin films [3], or nanoporous materials [4], the description of thermal transport as diffusive heat conduction breaks down. Furthermore, as we will see later, Fourier's law is violated in transient heat conduction experiments [5, 6].

The departure of heat transport from the diffusive regime can be qualitatively explained by considering heat as being carried by material waves, with their own frequency, group velocity, and dispersion curve [7]. Wave effects can be taken into account by the so-called molecular dynamics (MD) simulations, which essentially apply Newton's law to a system of atoms, whose internal forces are modeled by either empirical or first-principles calculations. The thermal conductivity computed by a MD-based formalism inherently includes both harmonic and anharmonic effects and can be computed by the well-known Green-Kubo formula:

$$\kappa_{\alpha\beta} = \frac{1}{Vk_BT^2}\int_0^\infty < J_\alpha(0)J_\beta(t) > dt \qquad (3)$$

where k_B is the Boltzmann constant, V the volume of the system, and $\kappa_{\alpha\beta}$ the thermal conductivity tensor. In Eq. 3, $< A >$ is the average ensemble of the observable A and can be replaced by a time average if the simulation is long enough to satisfy ergodicity [8]. Although MD is a powerful approach to model thermal conductivity at very small scales, it can become computationally prohibitive when dealing with realistic structures of mesoscale dimensions, such as thin films and porous materials. To overcome this limit, one may model the quantized lattice vibrations – or phonons – as particles and compute the temperature as well as the thermal fluxes by applying numerical schemes solving for particle transport problems.

This simplification is justified by noting that at room temperature, the dominant phonons in certain materials, such as Si, have wavelengths below 10 nm. Within this assumption, phonon transport can be safely modeled by the Boltzmann transport equation (BTE) [9]. The BTE has been gaining much attention recently in modeling thermal transport in nanostructured materials, especially for thermoelectric applications. In the following, we will focus on the BTE and the two main approaches used to solve it in complex structures. Readers interested in modeling wave effects at the mesoscale can refer to [10]. Also, while here we discuss the potential of the BTE in nanoscale heat transport simulations, there are many equally important methods left aside. A comprehensive review on computational tools for heat transfer can be found in [11].

The Boltzmann Transport Equation

In order to introduce the BTE for phonons, we first define $f(\mathbf{r}, \mathbf{k}, t)$ as the phonon distribution function at a given time t, at position $\mathbf{r}$ with wave vector $\mathbf{k}$. With no loss of generality, we assume the Brillouin zone to be isotropic. We further define the quantity $I(\mathbf{r}, \mathbf{s}, \omega, p) = \frac{1}{4\pi}|\mathbf{v}(\omega, \mathrm{p})|fD(\omega, p)\hbar\omega$, where $D(\omega, p)$ is the density of states, $\mathbf{v}(\omega, p)$ the group velocity, $\hbar\omega$ the phonon energy, and p the phonon branch. The group velocity can be obtained from the dispersion relations $\omega(\mathbf{k})$ as $\mathbf{v} = \frac{\partial\omega(\mathbf{k})}{\partial\mathbf{k}}$. The quantity $I(\mathbf{r}, \mathbf{s}, \omega, p)$ represents the intensity of phonons for a given frequency within a unit solid angle [12], traveling along the direction $\mathbf{s} = \frac{\mathbf{v}}{|\mathbf{v}|}$. In the absence of any external driving forces, such as an applied temperature gradient, the intensity $I(\mathbf{r}, \mathbf{s}, \omega, p)$ equals the equilibrium intensity $I_0(T_0)$, defined as $I_0(T_0) = \frac{1}{4\pi}|\mathbf{v}(\omega, \mathrm{p})|f_0(\omega, T_0)D(\omega, p)\hbar\omega$. The term $f_0(T_0)$ is the Bose-Einstein distribution $f_0(T_0) = \left[\exp\left(\frac{\hbar\omega}{k_BT_0}\right) - 1\right]^{-1}$. Assuming that all phonon scattering events are uncorrelated, the nonequilibrium phonon intensity can be modeled by

$$\frac{1}{|\mathbf{v}|}\frac{\partial I}{\partial t} + \mathbf{s}\cdot\nabla I = \frac{I_0(T) - I}{\tau|\mathbf{v}|}, \qquad (4)$$

which is the BTE under the so-called relaxation time approximation [12]. In Eq. 4, τ is the intrinsic scattering time [13]. Energy conservation among

modes can be obtained by applying the continuity equation for the energy flux, which results in

$$\sum_p \int_0^{\omega_M^p} \frac{I_0(T)}{\tau} d\omega = \sum_p \int_0^{\omega_M^p} \frac{< I >}{\tau} d\omega, \quad (5)$$

where $I_0(T)$ is the Bose-Einstein distribution at the temperature T, ω is the phonon angular frequency, p is the phonon branch, and ω_M^p is the maximum angular frequency for the branch p. In practical thermal conductivity calculations, a difference of temperature ΔT is applied to a simulation domain, and once Eqs. 4 and 5 are solved consistently, the thermal flux is computed by

$$\mathbf{J}(\mathbf{r}) = 4\pi \sum_p \int_0^{\omega_M^p} |\mathbf{v}| < I\mathbf{s} > d\omega, \quad (6)$$

and the *effective* thermal conductivity is deduced by $J = -\kappa_{eff} \nabla T$. The BTE described in Eq. 4 has to be solved for the whole phonon spectrum and is called the *frequency-dependent* BTE. Generally speaking, the BTE can be solved either deterministically or stochastically. In the following, we devote a section to each approach.

Deterministic Solution of the BTE

Here we show how the BTE can be useful for calculating the steady-state thermal conductivity values in nanostructured materials. Following the approach described in [14], we consider only steady-state transport. Furthermore, we assume that the bulk thermal conductivity is isotropic and that a very small temperature difference is applied across the sample. Under these simplifications, the BTE becomes

$$\Lambda \mathbf{s} \cdot \nabla \tilde{T} + \tilde{T} = \gamma \int_0^{\infty} \frac{K}{\Lambda'^2} < \tilde{T} > d\Lambda', \quad (7)$$

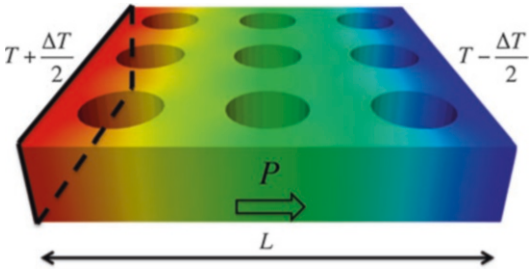
where $\tilde{T}$ is the departure of a temperature associated with a given phonon mode from the equilibrium, normalized by the applied temperature difference. In Eq. 7, the term $K(\Lambda)$ is the bulk phonon mean free path distribution, a quantity that can be obtained either theoretically [8] or experimentally [15, 16], and $\gamma = \left[\sum_p \int_0^{\infty} \frac{K}{\Lambda^2} d\Lambda\right]^{-1}$ is a material. The notation $< x >$ stands for the angular average $< x >= \frac{1}{4\pi} \int_{4\pi} x d\Omega$. The right-hand side of Eq. 7 is related to the normalized *effective* lattice temperature.

Practically, Eq. 7 requires the discretization of $K(\Lambda)$, as opposed to the discretization of the phonon frequencies typically used in a frequency-dependent approach. For each MFP, the BTE is solved by means of the finite-volume method, whereas the solid angle is discretized by the discrete ordinate method [17]. Equation 7 is named phonon MFP-BTE and has the following advantages:

- It retains the accuracy of the frequency-dependent approach.
- It requires the knowledge of only the bulk MFP distribution, which is a quantity that can be obtained from experiments [15].
- The requirements in the discretization of the bulk MFP distribution are less demanding than in a typical frequency-dependent approach, leading to a significant improvement in the computational efficiency [14].
- It is relatively easy to parallelize. In fact, each phonon mode is only coupled through the integral appearing on the right-hand side of Eq. 7. In a typical iterative solver, each BTE for a given MFP can be run independently, using the effective lattice temperature from the previous step.

In the following, we show an application of the MFP-BTE to nanoporous silicon, a promising material for thermoelectric applications, thanks to its capability to suppress thermal transport with little degradation in the electrical conductivity [4].

As shown in Fig. 1, a difference of temperature ΔT is applied to the simulation domain. Then, after the MFP-BTE converges, the thermal power, P, is computed on either the cold or hot



Simulating Nanoscale Heat Transport, Fig. 1 Simulation of a porous material. A difference of temperature ΔT is applied across the two ends of the domain

side. The effective thermal conductivity κ_{eff} is computed by using Fourier's law $\kappa_{eff} = \frac{PL}{A\Delta T}$, where A is the contact area. In Fig. 2a, the discretization of the solid angle is shown. As the actual simulated system is two-dimensional, we only consider the upper hemisphere and then apply symmetry. Figure 2b shows the discretization of the spatial domain. The MFP distribution, shown in Fig. 2c, is obtained by means of a frequency-dependent model, as described in [14].

The phonon thermal conductivity (PTC) is computed by varying the length of the unit cell from the nanoscale to the macroscale. As shown in Fig. 3, for very small unit cells, the thermal conductivity is reduced by roughly seven times with respect to the diffusive value, which is given by the approximated formula [18]

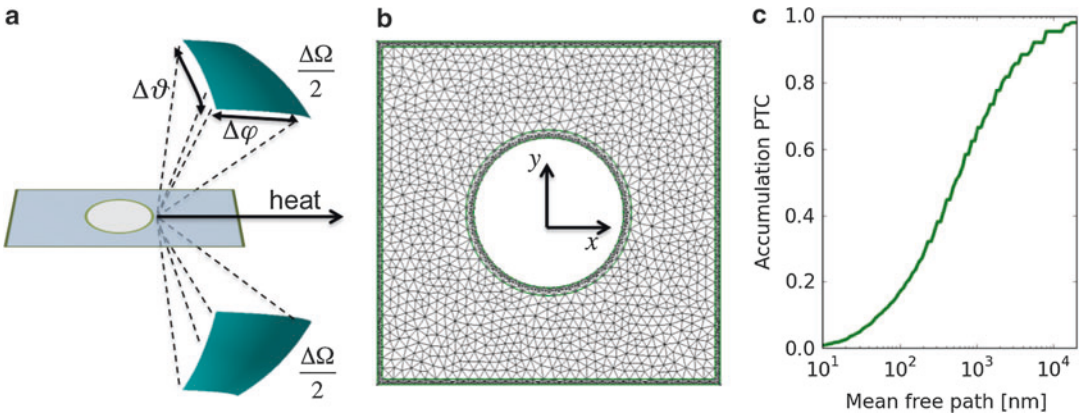
$$\kappa_{eff} = \kappa_{bulk} \frac{1-\phi}{1+\phi}, \tag{8}$$

where ϕ is the material porosity. In our case, we choose $\phi = 0.25$, which leads to the PTC $\kappa_{eff} \approx 77 W/mK$. For very large unit cells, the PTC correctly recovers the diffusive limit. In our calculations, the pore walls diffusively scatter phonons. In general, depending on the surface roughness, phonon wavelength, and temperature, there might be a fraction of phonons specularly reflected [9]. The BTE-MFP has also been applied to the realistic case described in [4], finding good agreement with experiments. These calculations represent a validation of the developed code and a starting point for PTC minimization in porous materials with different pore configurations. For example, in [19], it has been shown that triangular pores arranged in misaligned columns bring a reduction in the PTC of about 60 % with respect to the aligned case with the same porosity.

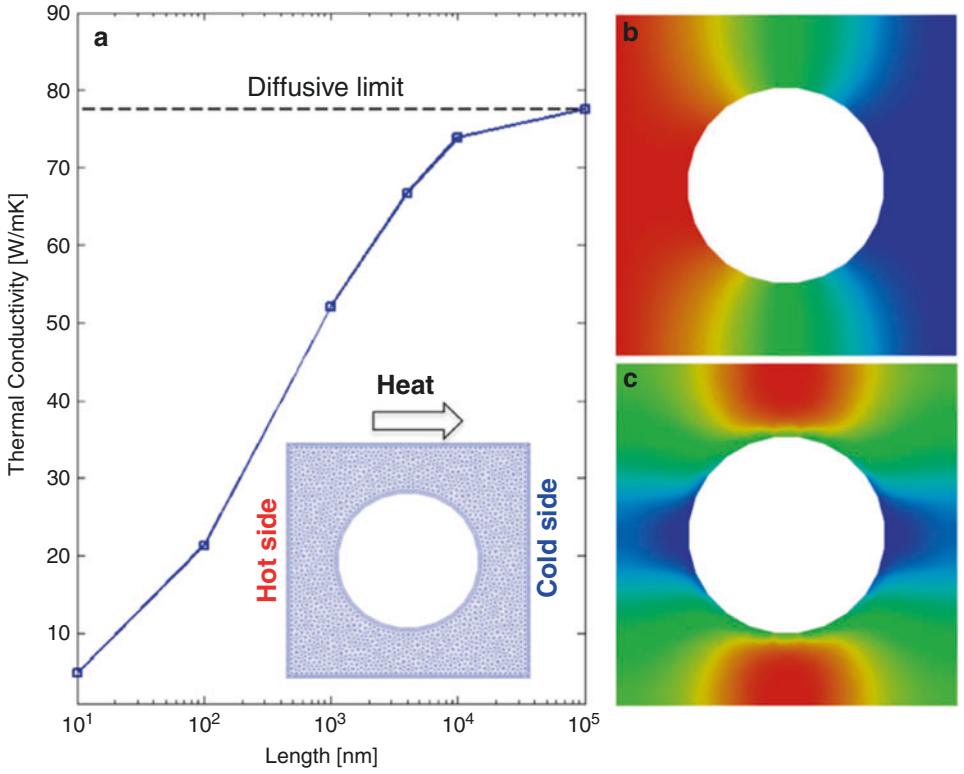
Monte Carlo Methods

One of the major limitations of traditional deterministic solution methods lies in the discretization of the six dimensions – seven dimensions for transient problems – of the phase space. Particle Monte Carlo methods have gained popularity thanks, in part, to their ability to avoid the full discretization of the phase space. They are commonly used for solving the Boltzmann equation for a range of physics problems such as neutron transport, radiation, rarefied gases, or electron transport. The idea of using such an approach for solving phonon transport problems was first introduced by Klitsner [20] and further improved by Peterson [21] and Mazumder et al. [22]. Various improvements were added between the years 2000 and 2014 and made it a powerful method.

Monte Carlo methods for particle transport solve the Boltzmann equation by *directly simulating* the physical processes involved. Phonons are characterized by their position, their frequency, their polarization, and their traveling direction, although alternative descriptions, using, for instance, the wave vectors instead of the frequencies, are also adequate. The group velocity vector is deduced from these properties, using the dispersion relation. Simulated phonon properties are randomly drawn from specified distributions in the phase space, and the rules for their evolution in time are deduced from the physical formulation. Quantities of interest, such as temperature or heat flux, are calculated by ensemble averaging. The Monte Carlo algorithm that solves for the transient nonlinear BTE with the relaxation time approximation starts with an initialization step where the initial properties of a population of N computational phonons are drawn from a known distribution defined by the initial



Simulating Nanoscale Heat Transport, Fig. 2 (**a**) Uniform scheme for solid-angle discretization. (**b**) Discretization of the spatial domain. (**c**) Cumulative thermal conductivity for Si obtained by a frequency-dependent model. We acknowledge ASME for granting us the permission for using Fig. 1 of the article G. Romano and J. C. Grossman. *Journal of Heat Transfer* 137.7 (2015): 071302



Simulating Nanoscale Heat Transport, Fig. 3 (**a**) Thermal conductivity versus the length of the unit cell. For very large unit cells, thermal transport reaches the diffusive limit. In the inset, the discretization of the simulation domain is shown. Periodic boundary conditions are applied along the direction of the heat flux. (**b**) Normalized temperature distribution. (**c**) Normalized thermal flux distribution. Due to phonon-boundary scattering, heat travels in areas that are far away from pore boundaries

condition. The evolution in time of this population is then calculated through a split algorithm where advection and scattering of phonons are treated separately. The algorithm is described in detail in [22]; further details are provided in [23, 24]. A timestep Δt is defined, and starting from the positions at time t, positions at time $t + \Delta t$ are deduced assuming that particle trajectories are collision-free. This is the advection step. The scattering step consists of simulating the scattering events that occur between times t and $t + \Delta t$. Given the relaxation time τ of a computational particle, the probability of occurrence of a scattering event may be calculated by $1 - \exp(-\Delta t/T)$. Thus, this probability law is used to randomly select particles undergoing a scattering event. Selected particles are replaced by new particles whose properties are drawn from the local "post-scattering" distribution $I^0(T)/(v\tau)$.

The scattering process must conserve energy. In early work [22–24], this was traditionally done approximately by adding and deleting particles until the post-scattering energy matched the pre-scattering one. Recently, a method for enabling exact energy conservation was proposed [25]. It relies on the simulation of computational particles representing a fixed amount of energy instead of a fixed number of phonons. As a result, energy is rigorously conserved by simply conserving the number of computational particles. Finally, the scattering step requires the knowledge of the local temperature at every timestep. This is usually done by defining a spatial grid of computational cells in order to sample the energy density of the particle population and relating it to the temperature of the corresponding equilibrium (Bose-Einstein) distribution. This algorithm is very similar to the direct simulation Monte Carlo (DSMC) method used for rarefied gases [26].

The main source of error from Monte Carlo methods is the statistical uncertainty, or noise, inversely proportional to the square root of the number of independent samples (the particles) used. Noise is an important issue for problems featuring low deviations from the Bose-Einstein equilibrium since it tends to obscure the signal. Such problems are ubiquitous in nano-engineering and yet cannot be treated by the particle Monte Carlo method presented above without resorting to massively parallel computing. Recently, a variance reduction scheme, also called "deviational" algorithm, was proposed for rarefied gases [27, 28] and for phonon transport [25] and was shown to capture arbitrarily low deviations from equilibrium at constant computational cost. The method relies on the concept of control variates. Instead of simulating random particles representing the absolute phonon distributions, the variance-reduced algorithm simulates particles representing the *deviation* from a known equilibrium. In other words, if the temperature of the system is expected to feature low deviations from a given temperature T_{eq}, we can introduce the following algebraic decomposition:

$$I = I^0\left(T_{\mathrm{eq}}\right) + \left(I - I^0\left(T_{\mathrm{eq}}\right)\right). \tag{9}$$

to obtain the deviational BTE

$$\frac{1}{v}\frac{\partial I^{\mathrm{d}}}{\partial t} + \mathbf{s}\cdot\nabla_{\mathrm{x}} I^{\mathrm{d}} = \frac{I^0(T_{\mathrm{loc}}) - I^0\left(T_{\mathrm{eq}}\right) - I^{\mathrm{d}}}{v\tau}. \tag{10}$$

The deviational algorithm, which solves for $I^{\mathrm{d}} = I - I^0(T_{\mathrm{eq}})$, retains the main features of the non-variance-reduced algorithm and adapts them for solving the deviational BTE. Notably, I^{d} may be negative and simulated particles must carry a sign. The particles representing the initial condition are drawn from the deviational initial distribution and the advection step of the split algorithm is unchanged. Although the selection rule for scattered particles is also unmodified, the post-scattering properties should now be drawn from the local deviational distribution $(I^0(T_{\mathrm{loc}}) - I^0(T_{\mathrm{eq}}))/\tau$. The total energy density is found by adding the stochastic deviational energy density to the deterministically calculated energy density corresponding the to equilibrium distribution.

The variance-reduced algorithm, discussed in detail in [25], offers the three following advantages:

– It does not introduce any approximation with respect to the traditional algorithm.

– The reduction of the statistical uncertainty essentially results from the fact that the amplitude of the simulated deviational distribution is small with respect to $I^0(T_{eq})$. In particular, arbitrarily low deviations from equilibrium can be simulated with no additional cost.
– The algebraic decomposition is inherently multiscale. It focuses the calculation effort to the regions where deviation from equilibrium is nonzero. By extending this idea to simulations of deviation from a Fourier solution, one can achieve algorithms which use particles *only* when the deviation from the Fourier solution is nonzero, that is, where size effects are important – the definition of a multiscale method. The simulation of a thermoreflectance experiment presented in [25] is a compelling example of this effect.

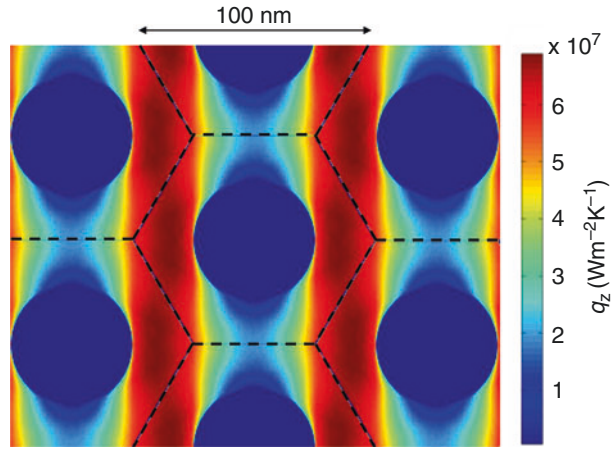
Additional computational benefits can be obtained by combining the deviational formulation with the linearization of the BTE, which amounts to linearizing the collision operator as follows:

$$\frac{\left(I^0(T_{loc}) - I^0(T_{eq})\right)}{\tau(\omega, p, T)} \approx \frac{\left(T_{loc} - T_{eq}\right)}{\tau\left(\omega, p, T_{eq}\right)} \frac{\partial I^0(T_{eq})}{\partial T} \tag{11}$$

In other words, the distribution from which the post-scattering properties are drawn does not depend on the local temperature since, after normalization, the temperature term $T_{loc} - T_{eq}$ cancels. We recall that the algorithms presented above require the use of a timestep Δt because knowledge of the temperature field (which features in $I^0(T_{loc})$) is required for simulating the scattering of the particles. Within the linearized approximation, such information is not needed. The scattering rates can be taken at temperature T_{eq}, and the post-scattering properties are all taken from a fixed distribution. Finally, we already explained that energy conservation is ensured by conserving the number of computational particles in the energy-based formulation. These observations yield the following consequences. Particles may be simulated one by one, independently from one another (a timestep is not needed). For a given particle, the time between each scattering event may be simply calculated from an exponential distribution with survival parameter $\tau(\omega, p, T_{eq})$. In other words, the traveling time between each scattering event is calculated by the formula $\Delta t = -\tau(\omega, p, T_{eq}) \ln(R)$, where R is a random number uniformly drawn in the range (0, 1). The resulting "kinetic-type" algorithm is similar to existing algorithms used for neutron or photon transport, where particles are assumed to interact with the underlying medium only. Details on the implementation may be found in Refs. [29, 30]. Reference [31] shows that, at an equilibrium temperature of 300 K, the linearized approximation is reasonably accurate up to a deviation of 30 K.

Although particle methods are inherently explicit in time, it is shown in [29] that most steady problems can be efficiently (and rigorously) treated using this approach. Traditional time-based Monte Carlo algorithms for phonon transport typically solve steady-state problems by letting a time-dependent system evolve toward steady state. Since quantities of interest are then only sampled when the steady state is reached, the number of timesteps needed to compute the transition from the initial condition to the steady state wastes significant computational resources. On the contrary, by treating each particle independently, the "kinetic-type" approach can deliberately ignore the transitory regime and only simulate particles at the steady state. Such particles are emitted from the steady sources only, and particle trajectories are terminated only when they leave the spatial domain, independently of how long the particle has been staying in the system. The resulting method rigorously solves the steady Boltzmann equation. Figure 4 shows an example of steady-state calculation in a nanoporous material.

The kinetic Monte Carlo method yields two immediate advantages. It features substantial memory savings since particles are simulated one by one, and eliminating the need of a timestep significantly reduces the cost of calculating each trajectory. Speedups of a factor 100–1,000 with

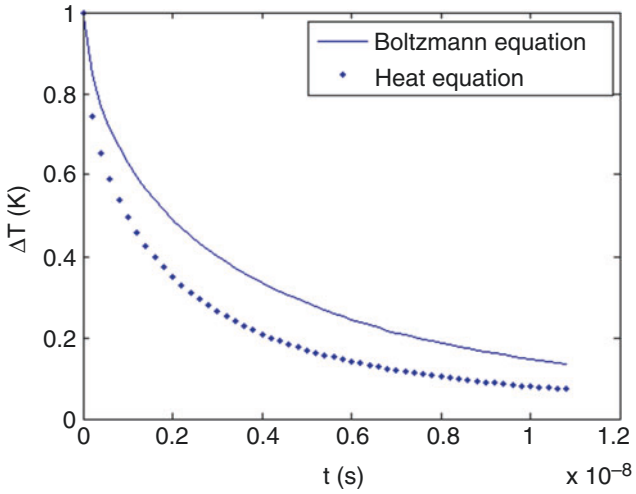


Simulating Nanoscale Heat Transport, Fig. 4 A component of the heat flux in a periodic nanostructure with circular pores arranged into a honeycomb structure. The material parameters used for this simulation are the silicon parameters used in [25]. The diameter of the circular pores is 50 nm. The heat flux results from an externally applied temperature gradient of 1×10^6 Km^{-1} (see Ref. [29] for detailed information on the formulation that enables the simulation of an externally applied temperature gradient with periodic boundary conditions). The component of the heat flux shown here is parallel to the applied gradient. In this calculation, the boundaries of the nanopores diffusely reflect the phonons

respect to the timestep-based deviational algorithm have been reported [29]. Finally, the method is inherently multiscale in time. Relaxation times in typical materials such as silicon are known to span several orders of magnitude. Since the time between each scattering process is computed independently of other phonons, the process automatically adapts itself to the time scale of the relaxation time of each step. It should also be noticed that this algorithm completely removes the need for spatial and time discretization. Not only does it remove the associated errors, it also allows to simulate systems of infinite sizes. Figure 5, which shows the time-dependent Boltzmann solution in an infinite system, highlights these features.

An important aspect of particle Monte Carlo methods lies in the fact that they primarily return *moments* of the underlying solution. For instance, the temperature at a given point in space is calculated in an average sense, within a volume surrounding this point. The accuracy of the estimate thus depends on the number of particle trajectories intersecting the volume and contributing to the estimate. Consequently, estimates calculated within small volumes tend to feature high statistical uncertainties. Recently, the adjoint Boltzmann transport equation for phonons was introduced as a means of addressing this limitation [31]. The adjoint BTE describes a particle problem where particles travel backward in time and where sources and detectors are switched. Thanks to this formulation, estimates can be produced in arbitrarily small volumes in the phase space, including surfaces and points. This is useful, for instance, for producing the contributions of individual phonon modes to the heat flux [32, 33].

The techniques presented above focused exclusively on the relaxation time approximation which, while relatively convenient and reasonably accurate for a number of materials, fails to capture the specific features of three-phonon scattering. The three-phonon scattering operator, which incorporates both energy and momentum conservation, has been widely acknowledged to be a more accurate description [34]. In addition, while the relaxation time approximation is semiempirical and requires the fitting of parameters, the three-phonon scattering may be entirely derived from ab initio calculations.



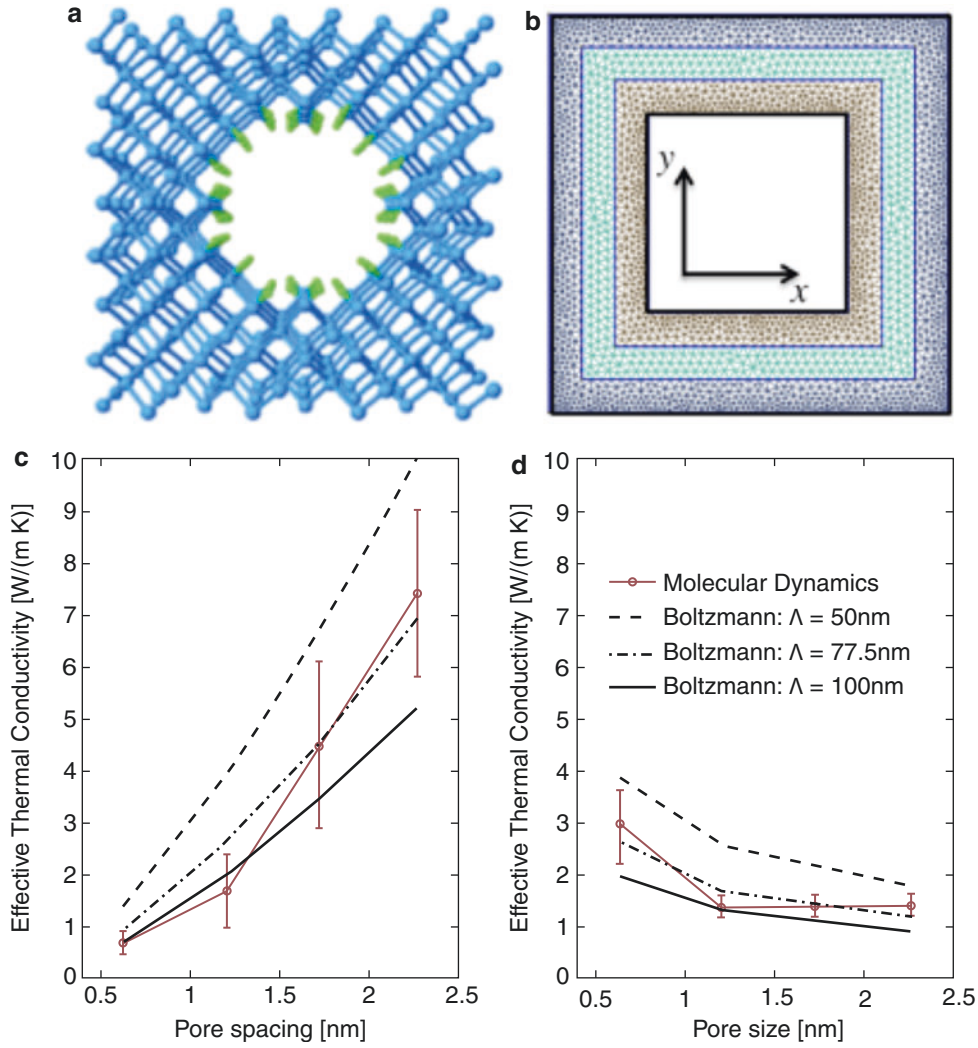
Simulating Nanoscale Heat Transport, Fig. 5 In this figure, a homogeneous infinite material is considered (the material properties of silicon were used for the calculation). The computational domain is $\mathbf{R}^3$. The initial condition is as follows: $T(x, y, z, t = 0) = 301$ K within the cube defined by $0 < x < L$, $0 < y < L$, and $0 < z < L$, with $L = 2$ μm, and $T(x, y, z, t = 0) = 300$ K outside of the cubic region. Following this initial heating, the linearized Monte Carlo technique can be used to calculate the average temperature within the 2 μm cube against time. The deviational distribution is defined with respect to the equilibrium at temperature $T_{eq} = 300\ K$. The quantity ΔT therefore corresponds to $\Delta T = T - T_{eq}$. In this configuration, the solution of the heat equation can be calculated analytically. The analytical solution requires the knowledge of the diffusivity, which can be computed from the material parameters

Unfortunately, the resulting BTE is so complicated that very few general methods incorporating three-phonon scattering have so far been developed. Notably, efficient solution techniques for solving the *homogeneous* equation were recently developed [35]. Monte Carlo methods are well known for their ability to solve highly complex partial differential equations and, as such, hold significant potential. Landon and Hadjiconstantinou recently developed a promising solution method based on the deviational approach for treating 2D materials such as graphene [36].

Need for Multiscale Modeling

In the previous sections, we have described two solution techniques for BTE for nano- and mesoscale thermal transport. Also, in the very beginning, we have briefly described the molecular dynamics method and its importance in certain nanoscale systems such as those dominated by interfaces. In Fig. 6, we report a comparison between the MD and BTE calculations, varying pore distances, and sizes in porous silicon.

Specifically for this case, we used the so-called gray approximation, where all phonons are assumed to the have the same MFP, which is used as a parameter to fit MD data. Although MD and BTE results agree qualitatively with each other, the BTE is not as accurate as MD, especially because the phonon dispersion differs from the bulk on very small scales. It is, therefore, paramount for future research to identify the range of validity of both models and ultimately bridge them in accurate yet computationally affordable simulations. In reference to the porous system, when pores are filled with a guest material, interfacial effects may become important and a detailed treatment of phonon transmission



Simulating Nanoscale Heat Transport, Fig. 6 (**a**) Porous silicon modeled atomistically. The pore walls have been hydrogenated. (**b**) Porous silicon modeled by means of a continuum domain. (**c**) Effective thermal conductivity versus pore spacing. (**d**) Effective thermal conductivity versus pore size

across such an interface is necessary. On the other hand, performing MD over the whole domain is computationally prohibitive. As a consequence, a good approach could be to split the simulation domain in two parts, one where MD is performed and one governed by the BTE. A good multiscale method should then ensure flux conservation between the two regions and define appropriate rules for the atomistic-to-continuum coupling [37].

Cross-References

- ▶ Heat Transfer in Semiconductor Nanostructures
- ▶ Nanostructured Thermoelectric Materials
- ▶ Thermal Conductivity and Phonon Transport

References

1. Fourier, J.B.J.: Théorie Analytique de la Chaleur. Cambridge University Press, New York (1822)

2. Li, D., Wu, Y., Kim, P., Shi, L., Yang, P., Majumdar, A.: Thermal conductivity of individual silicon nanowires. Appl. Phys. Lett. **83**(14), 2934 (2003)
3. Ju, Y.S., Goodson, K.E.: Phonon scattering in silicon films with thickness of order 100 nm. Appl. Phys. Lett. **74**(20), 3005 (1999)
4. Song, D., Chen, G.: Thermal conductivity of periodic microporous silicon films. Appl. Phys. Lett. **84**(5), 687 (2004)
5. Maasilta, I., Minnich, A.J.: Heat under the microscope. Phys. Today **67**, 27–32 (2014)
6. Wilson, R.B., Cahill, D.G.: Anisotropic failure of Fourier theory in time-domain ther-moreflectance experiments. Nat. Commun. **5**, 5075 (2014)
7. Chen, G.: Nanoscale Energy Transport and Conversion: A Parallel Treatment of Electrons, Molecules, Phonons, and Photons. Oxford University Press, New York (2005)
8. Esfarjani, K., Chen, G., Stokes, H.T.: Heat transport in silicon from first-principles calculations. Phys. Rev. B. **84**, (2011). doi:10.1103/phys-revb.84.085204
9. Ziman, J.M.: ELECTRON-ELECTRON INTERACTION, in Electrons and Phonons, pp. 159–174. Oxford University Press, Oxford (2001)
10. Davis, B.L., Hussein, M.I.: Nanophononic metamaterial: thermal conductivity reduction by local resonance. Phys. Rev. Lett. **112**, (2014). doi:10.1103/physrevlett.112.055505
11. Chen, G.: Multiscale smulation of phonon and electron thermal transport. Annu. Rev. Heat Transf. (2014). doi:10.1615/annualrevheattransfer.2014011051
12. Majumdar, A.: Microscale heat conduction in dielectric thin films. J. Heat Transf. **115**(1), 7 (1993)
13. Broido, D., Malorny, M., Birner, G., Mingo, N., Stewart, D.: Appl. Phys. Lett. **91**, 231922 (2007)
14. Romano, G., Grossman, J.C.: Multiscale phonon conduction in nanostructured materials predicted by bulk thermal conductivity accumulation function. arXiv preprint arXiv:1312.7849 (2013)
15. Minnich, A.J., Johnson, J., Schmidt, A., Esfarjani, K., Dresselhaus, M., Nelson, K.A., Chen, G.: Thermal conductivity spectroscopy technique to measure phonon mean free paths. Phys. Rev. Lett. **107**(9), 095901 (2011)
16. Regner, K.T., Sellan, D.P., Su, Z., Amon, C.H., McGaughey, A.J., Malen, J.A.: Broad-band phonon mean free path contributions to thermal conductivity measured using frequency domain thermoreflectance. Nat. Commun. **4**, 1640 (2013)
17. Chandrasekhar, S.: Radiative Transfer. Courier Dover Publications, New York (1960)
18. Nan, C.-W., Birringer, R., Clarke, D.R., Gleiter, H.: Effective thermal conductivity of particulate composites with interfacial thermal resistance. J. Appl. Phys. **81**(10), 6692 (1997)
19. Romano, G., Grossman, J.C.: Toward phonon-boundary engineering in nanoporous materials. Appl. Phys. Lett. **105**, 033116 (2014)
20. Klitsner, T., VanCleve, J., Fischer, H., Pohl, R.: Phonon radiative heat transfer and surface scattering. Phys. Rev. B **38**, 7576–7594 (1988)
21. Peterson, R.B.: Direct simulation of phonon-mediated heat transfer in a debye crystal. J. Heat Transf. **116**(4), 815 (1994)
22. Mazumder, S., Majumdar, A.: Monte Carlo study of phonon transport in solid thin films including dispersion and polarization. J. Heat Transf. **123**(4), 749 (2001)
23. Lacroix, D., Joulain, K., Lemonnier, D.: Monte Carlo transient phonon transport in silicon and germanium at nanoscales. Phys. Rev. B. **72**, (2005). doi:10.1103/phys-revb.72.064305
24. Hao, Q., Chen, G., Jeng, M.-S.: Frequency-dependent Monte Carlo simulations of phonon transport in two-dimensional porous silicon with aligned pores. J. Appl. Phys. **106**(11), 114321 (2009)
25. Peraud, J.-P. M., Hadjiconstantinou, N.G.: Efficient simulation of multidimensional phonon transport using energy-based variance-reduced Monte Carlo formulations. Phys. Rev. B. **84**, (2011). doi:10.1103/phys-revb.84.205331
26. Bird, G.: Molecular Gas Dynamics and the Direct Simulation of Gas Flows, vol. 1. Clarendon Press, Oxford (1994)
27. Homolle, T.M., Hadjiconstantinou, N.G.: A low-variance deviational simulation Monte Carlo for the Boltzmann equation. J. Comput. Phys. **226**, 2341–2358 (2007)
28. Radtke, G.A., Peraud, J.M., Hadjiconstantinou, N.G.: On efficient simulations of multiscale kinetic transport. Philos. Trans. R Soc. A Math. Phys. Eng. Sci. **371**, 20120182–20120182 (2012)
29. Peraud, J.-P.M., Hadjiconstantinou, N.G.: An alternative approach to efficient simulation of micro/nanoscale phonon transport. Appl. Phys. Lett. **101**(15), 153114 (2012)
30. Peraud, J.-P.M., Hadjiconstantinou, N.G.: Deviational phonons and thermal transport at the nanoscale. In: volume 7: Fluids and Heat Transfer Parts A, B, C, and D. ASME (2012)
31. Peraud, J.-P.M., Landon, C.D., Hadjiconstantinou, N. G.: Monte Carlo methods. Annu. Rev. Heat Transf. **17**, 205–265 (2014)
32. Peraud, J.-P.M., Hadjiconstantinou, N.G. (in press)
33. Hua, C., Minnich, A.J.: Semicond. Sci. Technol. **29**, 124004 (2014)
34. Ziman, J.M.: Electrons and Phonons. Clarendon, Oxford (1960)
35. Mingo, N., Stewart, D.A., Broido, D.A., Lindsay, L., Li, W.: Length-Scale Dependent Phonon Interactions, pp. 137–173. Springer, New York (2014)
36. Landon, C., Hadjiconstantinou, N.G.: J. Appl. Phys. **116**, 163502 (2014)
37. Wagner, G.J., Jones, R., Templeton, J., Parks, M.: Comput. Methods. Appl. Mech. Eng. **197**, 3351 (2008)

Simulation of Supported Metal Clusters

Giovanni Barcaro[1], Luca Sementa[2] and Alessandro Fortunelli[3]
[1]CNR-IPCF, Consiglio Nazionale delle Ricerche, Pisa, Tuscany, Italy
[2]CNR-ICCOM, Consiglio Nazionale delle Ricerche, Pisa, Tuscany, Italy
[3]Istituto per la Chimica dei Composti Organometallici (Institute for the Chemistry of Organometallic Compounds, ICCOM), Consiglio Nazionale delle Ricerche (National Research Council, CNR), Pisa, PI, Italy

Synonyms

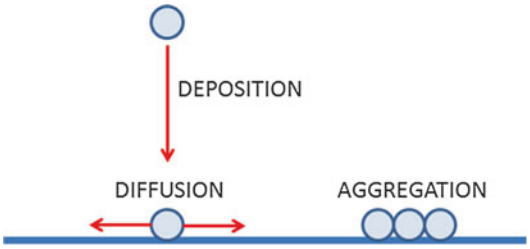
Deposited metal nanostructures; Heterogeneous catalysts; Nanoparticles; Nanoscale metals

Definition

A metal cluster is an aggregate composed of metallic elements, ranging in size between a few and few thousand atoms, while larger entities are usually referred to as "particles." A metal cluster can exist in different forms: free, coated by a protective shell of ligands, or deposited on a support, in which the last case the expression "supported metal cluster" is employed. In these structures, nanoscale confinement effects appreciably modify the response properties of the metallic nanosystem [1, 2]. For example, the high surface/volume ratio assures that a great part of metal atoms lie at the surface in low-coordinated sites, which are therefore available for chemical sensing and catalysis. As another example, the collective oscillation of conduction band electrons which gives rise to plasmonic phenomena in bulk metals is modulated by the presence of metal/environment interfaces, thus producing the so-called surface plasmon resonances. Magnetic properties are also exalted by the reduction of the electronic density of states in such confined systems, hence producing enhanced magnetic moments and anisotropy, etc. All these properties are very sensitive not only to size, shape, and external fields but also – in the case of supported metal clusters – to the interaction of the cluster with the support. From the possibility of modifying the cluster properties by tuning cluster/substrate interactions, a great freedom in the physical and chemical behavior of these materials results and therefore their interest in science and technology. This interest has been translated into an intense theoretical effort aimed at the prediction and understanding of the structure and properties of metal clusters via simulations, which is the topic of the present entry.

Introduction

A general problem in the application of metal clusters and nanoparticles is their intrinsic instability: due to their high surface/volume ratio, nanoparticles tend to decrease their energy by coalescing into larger particles (Ostwald ripening and sintering processes) [3]. Being intrinsically unstable, nanoclusters can only survive in the presence of kinetic barriers which avoid mass transfer and agglomeration processes, as due, for example, to specific interactions of nanoclusters with a support, with metal oxides as very common substrates [2]. Metal-on-oxide systems (as they are thus often called) can be created by depositing atoms or preformed clusters onto oxide surfaces in various heterogeneous environments (liquid/solid, gas/solid), followed by the processes of adsorption, diffusion, and self-organization of the aggregates (see Fig. 1). Supported metal clusters can be grown by chemical vapor deposition (CVD) or physical vapor deposition (PVD) techniques on top of a surface. In CVD the precursor metal atoms or clusters are protected by ligands which are eliminated via pyrolysis (the vaporized precursors are introduced into a reactor and adsorb onto a substrate held at elevated temperatures where they decompose producing metal species), whereas PVD involves condensation from the vapor phase of the bare precursor in the form of atoms or clusters. To control these processes



Simulation of Supported Metal Clusters, Fig. 1 Surface processes in physical vapor deposition (PVD) of metal species onto a substrate and growth of nanoparticles

and the resulting structural, catalytic, optical, etc., properties, detailed information on metal/support interactions and elementary and collective mechanisms of growth and diffusion is needed.

The properties of metal clusters in fact often depend sensitively on the microscopic details of their structure. Among other tools, microscopy techniques such as high-resolution transmission electron microscopy (TEM) can be used to derive information on the geometric structure of a cluster with real-space resolution currently reaching the atomic scale. Especially interesting is atomic-resolution environmental TEM (ETEM) [4] which is able to operate under controlled and close-to-realistic conditions of pressure and temperature. An application of this technique is reported in Fig. 2, which shows the evolution of the shape of Cu clusters (supported on ZnO), under the action of a mixture of H_2, H_2O, and CO at different values of temperature and total pressure [5].

Supported metal clusters find important applications, most notably as heterogeneous catalysts [2]. In this respect metal clusters benefit not only from their high surface/volume ratio but also from the cluster greater structural freedom with respect to extended surfaces which can better accommodate incoming ligand species and promote their diffusion, with the presence of the underlying support also playing a role in tuning the cluster propensitivity, as pictorially exemplified in Fig. 3.

Such detailed geometric and mechanistic information as depicted in Figs. 1, 2, and 3 is difficult to obtain at the experimental level. Therefore, theory and simulation play a crucial role in this field as complementary tools of characterization, analysis, and understanding. In this entry, a synthetic description of the approaches developed to meet this demand of theoretical information will be presented, especially focusing on the methods that are employed to predict the structural and electronic properties of supported metal clusters.

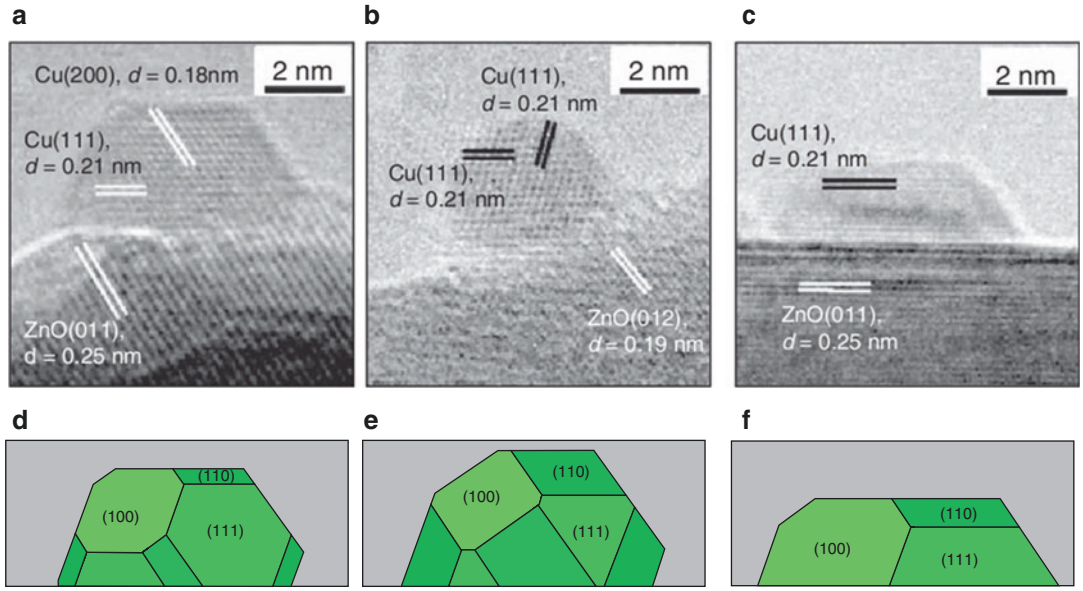
Simulation Methods

Calculations of energy and forces: to investigate the structural properties of supported metal clusters, what is needed is an expression for the system energy as a function of the structural degrees of freedom. In supported metal systems, especially for small clusters, a full atomistic description is usually employed, i.e., the degrees of freedom correspond to explicitly specifying the positions (the Cartesian coordinates) of all the atoms of the cluster and the support. One then needs to specify a function: E(**X**), where E is the energy and **X** are the atomic coordinates. Such an energy expression can be derived using different methods.

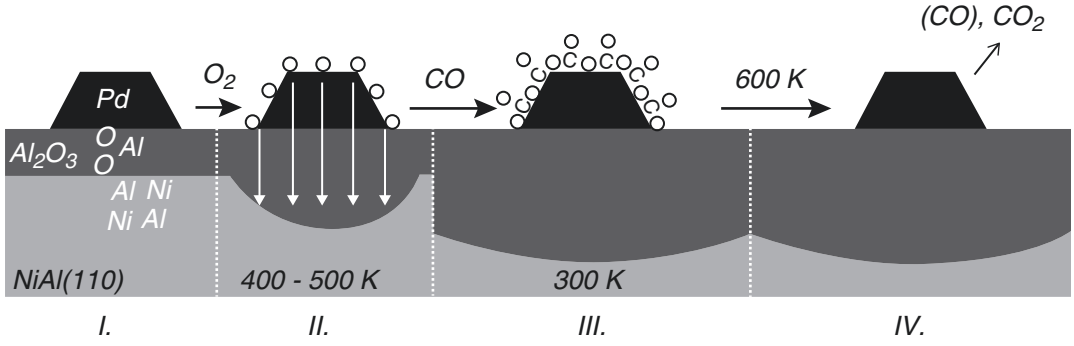
At the most fundamental level, one can use first-principles methods (i.e., methods not introducing parameters other than basic physical constants). Among these, the most widely employed in the field of metal clusters is the density functional theory (DFT), as it realizes the best compromise between chemical accuracy and computational effort. DFT in its most common form implies solutions of single-particle Kohn–Sham equations [7]:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V(\boldsymbol{r}, \boldsymbol{X}) + \int \frac{(r')}{|r-r'|}d\boldsymbol{r}' + V_{xc}[(\boldsymbol{r})](\boldsymbol{r})\right)\varphi_i(\boldsymbol{r}) = \lambda_i\varphi_i(\boldsymbol{r}) \tag{1}$$

where the Hamiltonian operator (in parenthesis) is expressed as a functional of the electronic density $\rho(\boldsymbol{r})$ of the system and $\varphi_i(\boldsymbol{r})$ and λ_i are single-particle wave functions and energies, respectively. The Hamiltonian operator contains the contributions of kinetic energy, external potential



Simulation of Supported Metal Clusters, Fig. 2 In situ TEM images (**a**, **b**, and **c**) of a Cu/ZnO catalyst in various gas environments together with the corresponding Wulff constructions of the Cu nanocrystals (**d**, **e**, and **f**). (**a**) The image was recorded at a pressure of 1.5 mbar of H_2 at 220 °C. The electron beam is parallel to the [011] zone axis of copper. (**b**) Obtained in a gas mixture of H_2 and H_2O, H_2:H_2O = 3:1 at a total pressure of 1.5 mbar at 220 °C. (**c**) Obtained in a gas mixture of H_2 (95 %) and CO (5 %) at a total pressure of 5 mbar at 220 °C (Reproduced with permission from Hansen et al. [5]. Copyright 2002 HighWire Press)



Simulation of Supported Metal Clusters, Fig. 3 The proposed mechanism for the interaction of Pd particles deposited on Al_2O_3/NiAl(110) with O_2 at elevated temperatures. It involves dissociation of oxygen on Pd, its diffusion through the oxide film, and reaction with the metallic substrate underneath (**II**). At temperatures above 400 K, the oxide film becomes thicker by about 3–5 oxygen sub-layers. Surface oxygen reacts with CO during the reduction step (**III**), followed by CO and CO_2 desorption upon heating to 600 K (**IV**). The resulting surface exhibits adsorption–desorption properties very similar to the pristine samples (Reproduced with permission from Shaikhutdinov et al. [6]. Copyright 2002 Elsevier)

(depending on the coordinates $\boldsymbol{X}$ of the nuclei), coulombic repulsion, and exchange–correlation potential. DFT is a sophisticated and computationally demanding first-principles approach, with an explicit description of the electronic wave function of the system, and can only be used in practice when the size of the system is relatively small – say, clusters between subnanometer and 2 nm as typical dimensions, even though advances in software and hardware

are continuously extending the scope of systems amenable to accurate DFT modeling.

When considering a supported cluster at the DFT level, one usually considers also an atomistic and first-principles description of the substrate. This gives rise to two separate issues.

First, it necessitates either employing a code with periodic boundary conditions (such that a unit cell encompassing the metal cluster and a piece of its underlying support is periodically replicated in two or three dimensions) or developing a finite-size model (usually named a "cluster" model, not to be confused with the metal cluster itself) which is embedded into the neighboring environment via properly chosen boundary conditions (such as an external potential or pseudo-atoms mimicking the environment). The use of DFT periodic computational codes is currently the most common choice. The explicit description of the substrate entails a remarkable increase of the computational cost, thus resulting in a limitation in the size of the systems which can be treated at this level of theory.

Second, the substrate itself may not be properly described by a standard DFT approach. Considering, for example, transition metal or rare earth oxide supports, one can face the problem that standard DFT tends to excessively delocalize the *d*-band or *f*-band electrons and simultaneously decrease the associated magnetism. A convenient way to circumvent the problem is to add an energy term to the Hamiltonian operator: the Hubbard *U* term, so named after its proposer [8], which corresponds to an energy penalty for doubly occupying a given orbital. The *U* term can be self-consistently inserted into the Kohn–Sham operator of Eq. 1, and by choosing a proper value for the *U* term, one can achieve a much better description of the oxide support. The ensuing method is called DFT + *U*.

For larger metal aggregates, say over 2 nm in size, DFT becomes excessively demanding, and one has to switch to a description of the system in terms of analytic potentials. In these approaches the energy function E(**X**) is expressed in terms of analytic expressions of the atomic coordinates. As the number of such coordinates is huge, further simplifications must be sought. Let us focus first on the metal–metal interactions. A common practice is to first define E(**X**) as a sum of site atomic energies $\mathrm{E}^{i,\mathrm{atomic}}(X)$:

$$\mathrm{E} = \sum\nolimits_i \mathrm{E}^{i,\mathrm{atomic}}(X) \tag{2}$$

where the sum runs over all *i*-atoms. One then expresses the dependence of the $\mathrm{E}^{i,\mathrm{atomic}}(X)$ functions on the coordinates of atoms other than the *i*th through generalized collective variables (GCV). The GCV, i.e., are a limited number of functions which describe the distribution of neighbors around the given atom and depend on an analytically compact form upon their number, distance, and orientation. What distinguishes the various analytic potentials is the choice of the GCV and the analytic form by which $\mathrm{E}^{i,\mathrm{atomic}}$ depend on the GCV. One of the simplest analytic potentials to describe the metal–metal interactions is the second moment approximation (SMA) tight-binding potential [9], in which each atomic energy is written as a sum of a pair repulsive ($E^{i,\mathrm{rep}}$) and a many-body attractive ($E^{i,\mathrm{att}}$) component:

$$E = \sum_{i=1}^{N} \left\{ E^{i,\mathrm{rep}}(i) + E^{i,\,\mathrm{att}}(i) \right\} \tag{3}$$

where

$$E^{i,\mathrm{rep}}(i) = \sum_{j \neq i}^{N} A \exp\left\{ -p\left(\frac{r_{ij}}{r_0} - 1 \right) \right\} \tag{4}$$

$$E^{i,\mathrm{att}}(i) = \left[\sum_{j \neq i}^{N} \xi^2 \exp\left\{ -2q\left(\frac{r_{ij}}{r_0} \right) - 1 \right\} \right]^{1/2} \tag{5}$$

and the sums in Eqs. 3, 4, and, 5 run over all *j*-atoms different from the *i*th. The parameters *A*, r_0, ξ, *p*, and *q* are characteristic of the given metal element and are usually fitted to the experimental quantities such as the cohesive energy, the equilibrium lattice parameter, and independent elastic constants for the reference crystal structure. It should be stressed that in the SMA analytic potential, the repulsive term in $E^{i,\mathrm{atomic}}(X)$ is linear, i.e.,

S

simply additive, whereas the attractive term has a many-body character and is highly nonlinear since it is proportional to the square root of a sum of additive terms. This *many-body* or *nonlinear behavior* as a function of coordination is an essential feature of the metallic bond. This feature allows metallic elements to sustain a great number of possible coordination numbers whose strength however tends to *saturate* as the coordination number increases.

While many different variants of analytical potentials are available for describing metal–metal binding, the choice is much more restricted for the metal/oxide interaction. One of the most accurate so far proposed refers to the square-symmetry MgO(100) surface and is one in which the substrate is assumed to be rigid, i.e., the oxide atoms are assumed to be frozen in their equilibrium configurations [10]. The metal/oxide interaction is then expressed as a sum of the contribution coming from each metal atom:

$$E^{Pd-MgO} = \sum_{i=1}^{N} E_i(x, y, z, C) \tag{6}$$

where N is the number of metal atoms in the cluster. The atomic energies E_i depend upon the spatial position (x,y,z) and the metal coordination (C) of each metal atom. The dependence upon the height of the metal atom with respect to the substrate plane (x,y) is described via a simple Morse-type function:

$$E_i(x, y, z, C) = a_1 + e^{-2a_2(z-a_3)} - 2e^{-a_2(z-a_3)} \tag{7}$$

where the parameters a_i, i = 1–3, depend on (x, y) and the coordination number C. It should be noted that the dependence of the a_i parameters on metal coordination is *nonlinear* or *many body*. In fact, it is intuitive that the metal/oxide interaction is strongest in the case of adsorption of isolated adatoms and decreases progressively in a nonlinear way as a function of an increasing number of metallic neighbors. A three-point interpolation formula describing an exponentially decreasing dependence of the interaction energy on the coordination of the metal atom C is used in the parametrization which hence reads

$$a_i(x, y, C) = b_{i,1} + b_{i,2}e^{-C/b_{i,3}} \tag{8}$$

The parameters $b_{i,j}$, i,j = 1–3, finally depend on (x,y) via a linear combination of trigonometric functions preserving the square symmetry of the MgO(100) unit cell:

$$\begin{aligned} b_{i,j}(x, y) = {} & c_{i,j,1} + c_{i,j,2}[\cos(x) + \cos(y)] \\ & + c_{i,j,3}[\cos(x + y) + \cos(x - y)] \end{aligned} \tag{9}$$

In this way, a set of 27 c_{ijk}, i,j,k = 1–3, parameters is sufficient to describe the oxide/metal interaction. This procedure has been extended and successfully applied to other metals interacting with the same oxide surface. Unfortunately, analytic potentials describing the metal/oxide interaction in the case of nonrigid oxide substrates are not yet available.

Algorithms for structural search: supported metal clusters can exhibit an impressive variety of different structures and morphologies (see Ref. [11] and the discussion below). In the case of small clusters (below 100–200 metal atoms), the majority of the atoms are on the surface, whence the importance of reducing the surface energy. This reduction can be achieved by adopting noncrystalline structures, which manage to decrease the surface energy at the expense of increasing the cluster internal energy due to bond strain. The Mackay icosahedron is probably the most common noncrystalline motif: this structure is limited by (111) close-packed facets only, which are usually the surfaces with the lowest energy, thus minimizing the cluster surface energy [11]. However, Mackay icosahedra are highly stressed structures (radial inter-shell bonds are highly compressed, whereas intra-shell bonds are highly expanded) and are hence favorable only at small sizes. Moreover, their interaction with a substrate is often disfavored so that they are not common in supported metal clusters. Another noncrystalline motif is represented by the decahedron: this structure is formed by two pentagonal

pyramids sharing a base; its surface has only (111) close-packed facets. A truncation of the five edges limiting the common basis of the pyramids to expose (100)-like facets is advantageous and produces the so-called Ino decahedron. Other cuts with reentrances exposing further (111) close-packed facets separating neighboring (100)-like facets can also be advantageous (Marks decahedra). Decahedra are also strained structures, but the strain is much smaller than for icosahedra, making them favorable in an intermediate range of sizes [11]. When the cluster grows in size, crystalline motifs become preferred as the ratio between surface and bulk atoms quickly decreases: volume contribution is hence optimized, at the expense of a higher surface energy.

All these motifs can be adopted by both isolated and supported metal clusters. Moreover, such motifs are also often close in energy, so that the metal clusters can transition among them, i.e., exhibit what is usually called a "fluxional" behavior. The fact that the PES (potential energy surface) of a metal cluster system is characterized by a large number of local minima with relatively small energy differences among them represents a challenge at the theoretical/computational level if accurate quantitative predictions are sought for. To be predictive, it is necessary to utilize theoretical methods able to perform a systematic sampling of the system PES. Appropriate methods to this effect are based on global optimization techniques, i.e., computational techniques aimed at finding the absolute minimum (or global minimum) of a mathematical function which can be very complex and exhibit very many minima differing by small energy amounts.

One of the simplest and most efficient global optimization techniques is the so-called Monte Carlo with minimization or basin-hopping (BH) algorithm, which has been proposed more than 25 years ago [12] but is still the most commonly used. The BH algorithm consists of the following steps:

1. An initial random configuration of the metal cluster is chosen, a local geometry optimization is performed, and the final energy (the so-called "fitness" parameter) is registered as E_1.
2. Starting from the relaxed configuration, the atoms of the metal cluster are subjected to a random move, a new local geometry optimization is performed, and the final energy is registered as E_2.
3. A random number (rndm) between 0 and 1 is generated, and the movement of step 2 is accepted only if $\exp[(E_1–E_2)/k_B\ T] > \text{rndm}$ (Metropolis criterion).
4. Steps 2 and 3 – the Monte Carlo steps – are repeated a given number of times.

Depending on the k_BT parameter, which plays the role of a fictitious temperature, some high-energy configurations are accepted, and the search is able to explore different structural motifs (belonging to different funnels of the PES) of the metal cluster. The thoroughness of the BH search is much increased when the BH protocol is coupled with *structural recognition* algorithms, i.e., algorithms which classify geometrical configurations as belonging to a given structural motif (or structural family) [11].

The BH algorithm can be in principle combined with any method for calculating energy and forces, but the feasibility of such a systematic sampling approach depends on the size of the system, so that a multi-scale hierarchy of theoretical methods can be envisaged [11].

For small supported metal clusters (with a number of atoms, N, less than a few tens, say $N \leq 40$), it is computationally feasible to conduct a global optimization (in the form of the BH algorithm) using energy and forces derived from DFT in a DF-BH method.

For medium-size supported metal clusters (say, $40 \leq N \leq 200$), the DF-BH approach becomes progressively less feasible, and more approximate methods such as analytic potentials (AP) must be employed. An efficient possibility is to replace the DF-BH approach with a combination of DFT and AP simulations, in the DF-AP method. In this approach an AP for the given system is first selected (or developed). Second, a thorough global optimization search using, e.g., the BH technique with structural recognition algorithms

is conducted using this AP. Third, the low-energy configurations produced by the BH search are grouped into structural families or basins. Fourth, a few of the low-energy configurations belonging to each structural basin are subjected to local geometry relaxation using a DFT method, and the changes in the relative energy ordering of the structural families are analyzed. Finally, if these changes highlight inconsistencies or inaccuracies in the AP, the DFT results are used to refine the AP in a self-consistent process.

For yet larger supported metal clusters (say $200 \leq N \leq$ few $\cdot$ 1,000), even performing DFT calculations becomes progressively unfeasible, at least conducting them in large numbers, and one has to rely entirely on global optimization searches based on AP approaches. The size of the system justifies the use of an AP and thus an average description of the metallic bond.

At some point (the current limit is around 1,000–2,000 supported metal atoms) even the BH algorithm using AP becomes computationally unaffordable. An efficient alternative solution is to use extrapolations based on structural motifs, which is discussed in the following subsection. The idea is that for these large particles the competing motifs have been singled out and are known, so that large clusters belonging to these motifs can be constructed and locally optimized and their energy can be compared to predict energetic crossover and phase transformations among structural families.

Large Particles: Generalized Wulff Construction

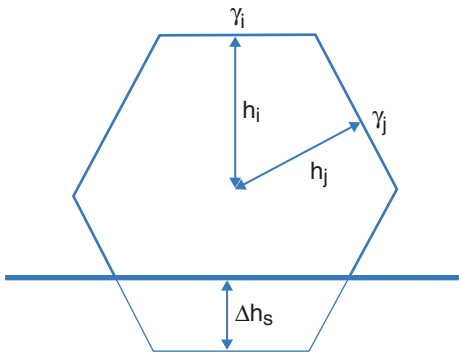
Let us assume that the structural family of the metal cluster is known, e.g., it corresponds to a crystalline motif. In this family, the atoms of the clusters are positioned in the crystalline sites characterizing the metal in its bulk form (e.g., face centered cubic, fcc). Still, the cluster shape needs to be determined. The "best" way to cut a gas-phase (or isolated) fcc particle was investigated by Wulff [13], and the corresponding building criterion (originally proposed for macroscopic aggregates but which can be translated into atomistic systems) goes under the name of Wulff construction. In this construction, taking for simplicity a crystal only limited by (111) and (100) faces, the best shape fulfills the following condition:

$$\frac{\gamma(100)}{\gamma(111)} = \frac{d(100)}{d(111)} \tag{10}$$

where $\gamma(100)$ and $\gamma(111)$ are the surface energies of the (100) and (111) faces, respectively, whereas d(100) and d(111) are the distances of the facets from the center of the cluster. The Wulff criterion has a general validity and can also be applied to higher-Miller-index surfaces – as the (110), etc. – and also noncrystalline structures, once low-energy surfaces are determined in these motifs.

The Wulff construction is valid for isolated clusters or particles. When considering the equilibrium structure of a cluster which is in contact with an environment, the Wulff construction can be straightforwardly generalized by considering the role played by the environment. Suppose, for example, that the cluster is surrounded by an atmosphere of interacting molecules. The adsorption of such species on the facets of the cluster will modify the corresponding surface energies. A generalized Wulff construction in which the *bare* surface energies are replaced by surface energies corresponding to facets "vested" by the ligand species can be used. This often gives rise to a phenomenon known as *particle restructuring* or *reshaping* and leads to substantial changes in the particle morphology, as observed experimentally in fair agreement with theoretical predictions (see, e.g., Ref. [14]). For example, in reactive oxidative conditions one typically expects that more open facets such as (100) are more reactive with respect to (111) and therefore are more stabilized by the interaction with oxygen species, so that they will grow at the expense of the more inert ones with respect to more reducing conditions.

A similar, properly modulated reasoning holds for supported clusters, and the corresponding generalized Wulff construction goes under the name of Wulff–Kaishev construction [15]. Suppose for the moment to ignore the influence of adsorbed



Simulation of Supported Metal Clusters, Fig. 4 Schematic representation of the equilibrium shape of a supported crystal. The Wulff shape of the free crystal is truncated at the interface by Δ_{hs}, which is proportional to the adhesion energy

species. A supported cluster can then be considered as isolated except for one face which is in contact with the substrate. It suffices to change the surface energy of such facet by taking into account the adhesion energy to the substrate to obtain the Wulff–Kaishew construction, illustrated in Fig. 4.

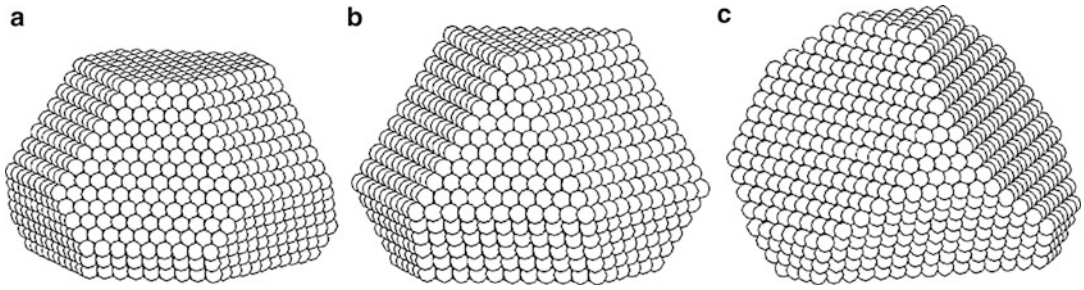
As it can be observed in Fig. 4, starting from the equilibrium shape dictated by the Wulff construction, a portion of the cluster can be cut off more or less profoundly depending on the adhesion energy, thus resulting in an asymmetric shape with respect to the central plane of the particle. In other words, the adhesion energy of the particle to the substrate reduces the surface energy of the particle face in contact with the substrate, also reducing the distance between this surface and the center of the nanoparticle. A simple example taking into account both the influence of ligand species and a support is given in Fig. 5, where reshaping of Ag clusters deposited on an amorphous alumina support under an O_2 atmosphere is depicted [14].

From the discussion of the Wulff–Kaishev construction, it results that it is crucial to be able to predict accurately the adhesion energy of a cluster to its support. In general, the adhesion of a metal cluster to a substrate depends on the detailed atomistic relationships between the metal particle and the substrate, also called epitaxy or epitaxial relationships. As an example, let us examine the results of global optimization searches carried out on four pure metals (Ag, Au, Pd, and Pt) adsorbed on MgO(100) [16]. The simulations on Ag, Au, Pd, and Pt particles adsorbed on MgO(100) singled out three dominant motifs (see Fig. 6):

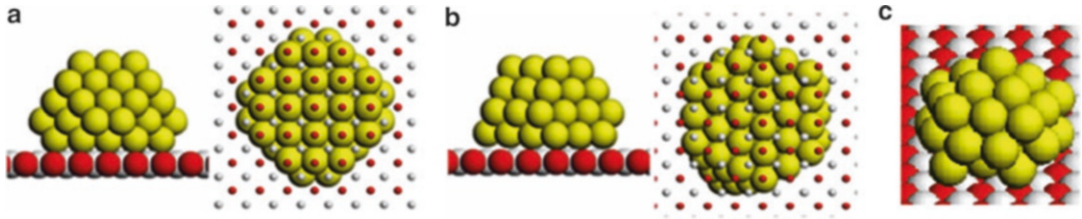
1. Fcc clusters in (001) epitaxy
2. Fcc clusters in (111) epitaxy
3. Decahedral cluster which adhere to the substrate with a pseudo-(001) facet

Icosahedral clusters do not achieve a good matching with the substrate and are therefore not favorable for these metal/support combinations. Decahedra can lie on a pseudo-(100) facet and achieve a better matching. Concerning fcc clusters, the competition between the (001) and the (111) epitaxy is due to the following reason. The (001) epitaxy has a stronger adhesion energy per unit contact area, due to its better matching with the oxygen atoms of the square-symmetry substrate, especially when the lattice mismatch between metal and oxide is small. However, in the (111) epitaxy this can be compensated by a larger contact area, so that the total adhesion energy may become even larger than in the (001) epitaxy.

The previous example shows how subtle can be the competition among different epitaxies in supported clusters. This becomes even more complicated when the substrate does not simply play a weak "perturbative" effect on the metallic particle equilibrium shape, but a strong interaction between the metallic cluster and the substrate completely alters the equilibrium shape of the former with respect to the gas phase. The simplest way in which this reconstruction can be realized is via the creation of a network of dislocations. This solution is often found in larger aggregates and thin films, for which a substantial adhesion energy drives the system to better match the substrate but is not sufficient to compensate the loss of metallic energy within the particle due to a transition to a completely different structural motif. An alternative possibility is that the interaction with the support can drive the metal cluster toward exotic



Simulation of Supported Metal Clusters, Fig. 5 Morphology of amorphous-alumina-supported silver particles in case of (**a**) nonoxidized surfaces and oxygen-covered surfaces at (**b**) $p(O_2) = 5 \times 10^{-3}$ atm and (**c**) $p(O_2) = 1$ atm. The lateral dimension of the clusters is about 6 nm (Reproduced with permission from Molina et al. [14]. Copyright 2011 Elsevier)



Simulation of Supported Metal Clusters, Fig. 6 Metal clusters deposited on a MgO(100) surface. (**a**) Side and bottom views of a fcc (001) cluster. (**b**) Side and bottom views of a fcc (111) cluster. (**c**) Side/top view of a decahedral cluster. Oxygen atoms are in *red* and magnesium atoms are in *light gray*. In the bottom views (*right part* of **a** and **b**), substrate atoms are represented by small spheres so that the contact epitaxy of the cluster is visible (Reproduced with permission from Ferrando et al. [16]. Copyright 2009 American Institute of Physics)

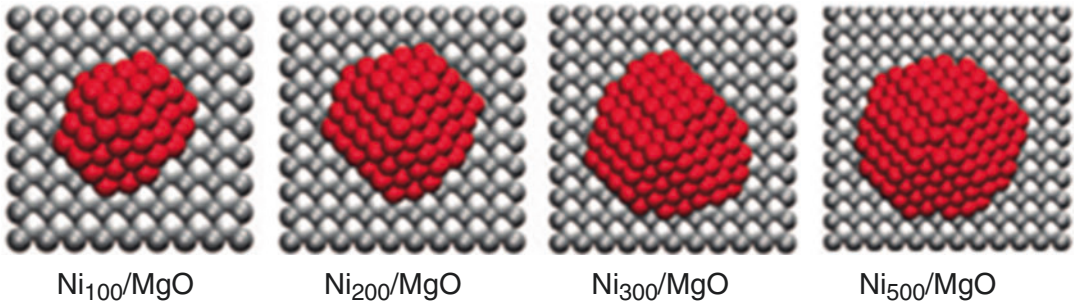
interface-stabilized structures which may even have no counterpart in the gas phase. A well-understood example is given by Ni clusters grown on MgO(100) [17]. In this system Ni atoms would like to be positioned on top of oxygen atoms of the support but – due to the large size mismatch between the nearest-neighbor distance in bulk Ni and the O–O distance in the MgO substrate – the preferred structural motif turns out to be hcp-like (note that Ni is fcc in the bulk), with close-packed planes alternating in . . .ABAB. . . stacking and oriented perpendicular to the oxide substrate along its [100] direction (see Fig. 7). Theoretical simulations show that, as the size of the clusters increases above $N = 40$, the lowest-energy structures are hcp for almost all sizes until for clusters larger than about 2,500 atoms, the interfacial energy is not sufficient anymore to compensate the transformation from fcc to hcp structure, and fcc-like clusters with a (111)/(100) epitaxial relationship to the support become energetically favorable. These predictions are in excellent agreement with experiment [17].

Electronic Effects

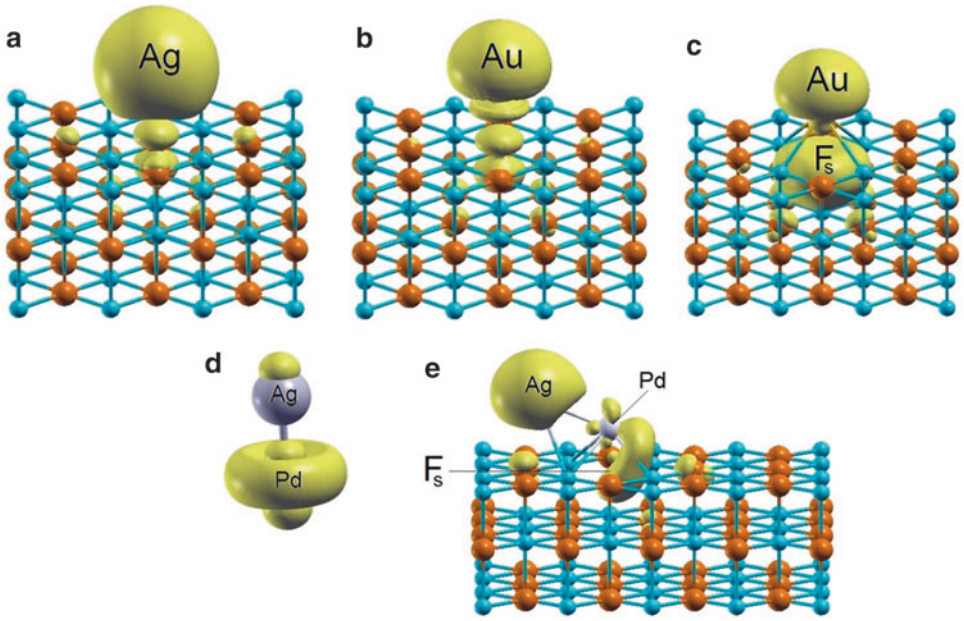
To conclude this entry, electronic effects, of great importance in catalysis, are briefly discussed.

The cluster electronic structure can be significantly affected by the interaction with the support: phenomena such as the quenching or at least lowering of the spin state of the metal cluster upon adsorption or charge transfer effects at the support/cluster interface are well documented [2, 11].

In general, when describing the metal/oxide interaction, it is customary to single out four main components: charge transfer, covalent binding, polarization, and dispersion interactions.



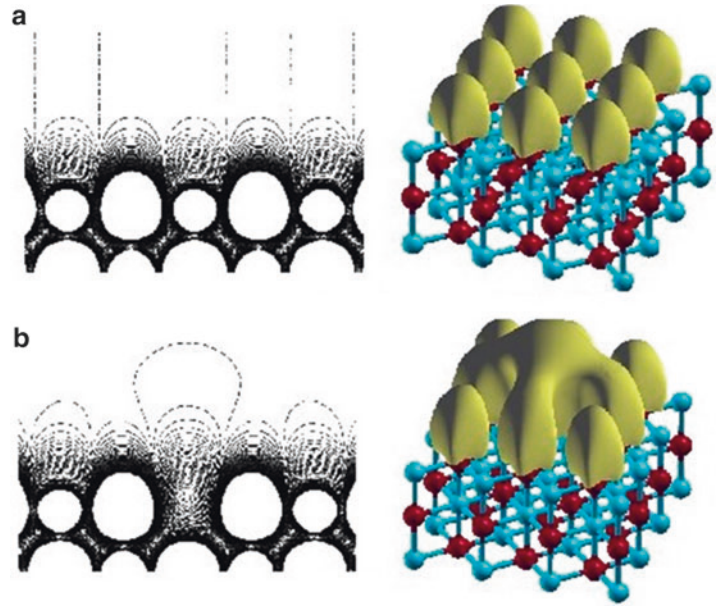
Simulation of Supported Metal Clusters, Fig. 7 Global minima of Ni/MgO nanodots (of hcp structure) shown for $N = 100$, 200, 300, and 500 atoms (Reproduced with permission and adapted from Ferrando et al. [17]. Copyright 2008 American Chemical Society)



Simulation of Supported Metal Clusters, Fig. 8 Spin density plots of (**a**) an Ag and (**b**) an Au adatom interacting with the regular (100) MgO surface, (**c**) an Au adatom interacting with the F_s-defected (100) MgO surface, (**d**) the gas-phase PdAg dimer, (**e**) the PdAg dimer interacting with the F_s-defected (100) MgO surface. Isosurfaces at a value of 0.001 AU are plotted (Reproduced with permission from Negreiros et al. [19]. Copyright 2014 Elsevier)

On simple oxides, it is often found that charge transfer between the metal and the oxide atoms is not large on regular surfaces. It can however be increased when the work function of the oxide is reduced by creating electronic states in the band gap via defects such as an oxygen vacancy. Charge transfer from/to the metal cluster can also be appreciable on oxides which act as a medium for electron transfer from an underlying electron reservoir as it occurs for oxide ultrathin films grown on low-work-function metal surfaces [18].

As an illustration, Fig. 8 shows a comparison of plots of the spin density for different systems. From this figure the changes due to a change in metal and the present of defects are apparent: Ag on the regular MgO(100) surface exhibits weak charge transfer, Au on the regular surface exhibits



Simulation of Supported Metal Clusters, Fig. 9 Electrostatic potential generated by (**a**) the regular and (**b**) an F_s-defected (100) MgO surface is shown in a 2D map (*left side*) and 3D map (*right side*). In the case of the 3D map, an isosurface of 0.0005 AU is plotted (Reproduced with permission from Barcaro et al. [20]. Copyright 2009 Springer)

a less weak charge transfer (and correspondingly a bit stronger bonding), and Au on an F_s center exhibits strong bonding and a very strong charge transfer. In Fig. 8 the case of a binary dimer (AgPd) is also shown: one can see that AgPd in the gas phase contains a Pd atom whose electronic configuration has been promoted from d^{10} (which is the configuration of the free atom) to d^9s^1 (which corresponds to the valence state of the fully formed metallic bond in bulk systems), whereas when AgPd is deposited on an F_s center of MgO(100), the electronic configuration of Pd is switched back from d^9s^1 to d^{10}; this shows that the interaction with the substrate can decrease magnetism (of Pd in this case) as mentioned above. Note also that, as the chosen metal systems contain an unpaired electron, the spin density gives good indications on chemical bond effects.

In addition to charge transfer, one less recognized effect of an oxide support is the strong polarization due to the electric field generated by such a charge-separated substrate. Even on the regular MgO(100) surface (see Fig. 9 for a plot of the electrostatic potential generated by the regular oxide surface, in which the electrostatic potential generated by an F_s-defected surface is also shown for comparison) and thus for metal/oxide interactions without or with little charge transfer, this electric field can completely alter the PES of small supported clusters. For example, breaking of O_2 on small Au_3 and Ag_3 clusters becomes thermodynamically favorable, opposite to the situation in the gas phase [11]. The effect of the support electrostatic field is connected with the polarization component of the metal/oxide interaction and can be important for both small and larger clusters. This effect is also the origin of the so-called "metal-on-top" effect, i.e., the enhancement of metal adhesion to the substrate due to the presence of other metal atoms on top of those which are directly in contact with the substrate [11]. The metal-on-top effect can play an important role in determining the structure of supported metal clusters, e.g., favoring upright with respect to epitaxial configurations of small metal clusters.

Conclusions

In conclusion, simulations represent an effective tool to characterize and predict the structure and properties of supported metal clusters, i.e., small- or intermediate-size metal aggregates deposited

on a support. Several theoretical approaches are used in this field, ranging from first-principles methods to analytic potentials to simplified constructions based on surface and adhesion energetics. In conjunction with these methods, global optimization explorations of the potential energy surfaces of such systems allow one to accurately predict their structural features, which serve as a basis for a deeper understanding and knowledge of their striking structural, catalytic, optical, magnetic, etc. response properties. Electronic interactions between the cluster and its support are particularly interesting, as they affect profoundly the activity and efficiency of supported clusters in heterogeneous catalysis, one of their main technological applications. The substrate can play a role simply as a stabilizing or anchoring support or as a component actively interacting with and strongly modifying the properties of the metal cluster, and this two-fold role has been underlined here. Attention has been mostly focused on general phenomena, while more detailed information on properties other than structural (and the theoretical approaches used to predict them) can be found in the cross-reference essays.

Cross-References

- ▶ Ab initio DFT Simulations of Nanostructures
- ▶ Bond-Order Potential
- ▶ Cluster
- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Engineered Nanoparticles
- ▶ Integrated Approach for the Rational Design of Nanoparticles
- ▶ Magnetic Nanostructures and Spintronics
- ▶ Metal Nanoparticles from First Principles
- ▶ Multiscale Modeling
- ▶ Nanoalloy Simulation
- ▶ Nanocluster
- ▶ Nanoscale Particle
- ▶ Optimization of Nanoparticles
- ▶ Mechanical Properties of Nanocrystalline Metals

References

1. McWeeny, R.: Coulson's Valence, 3rd edn. Oxford University Press, Oxford (1980)
2. Ertl, G., Knotziger, H., Schuth, F., Weitkamp, J. (eds.): Handbook of Heterogeneous Catalysis, 2nd edn. Wiley, New York (2008)
3. Parker, S.C., Campbell, C.T.: Reactivity and sintering kinetics of Au/TiO(2)(110) model catalysts: particle size effects. Top. Catal. **44**, 3–13 (2007)
4. Giorgio, S., Joao, S.S., Nitsche, S., Chaudanson, D., Sitja, G., Henry, C.R.: Environmental electron microscopy (ETEM) for catalysts with a closed E-cell with carbon windows. Ultramicroscopy **106**, 503–507 (2006)
5. Hansen, P.L., Wagner, J.B., Helveg, S., Rostrup-Nielsen, J.R., Clausen, B.S., Topsoe, H.: Atom-resolved imaging of dynamic shape changes in supported copper nanocrystals. Science **295**, 2053–2055 (2002)
6. Shaikhutdinov, S., Heemeier, M., Hoffmann, J., Meusel, I., Richter, B., Bäumer, M., Kuhlenbeck, H., Libuda, J., Freund, H.J., Oldman, R., Jackson, S.D., Konvicka, C., Schmid, M., Varga, P.: Interaction of oxygen with palladium deposited on a thin alumina film. Surf. Sci. **501**, 270–281 (2002)
7. Kohn, W., Sham, L.J.: Self-consistent equations including exchange and correlation effects. Phys. Rev. **140**, 1133A (1965)
8. Hubbard, P.S.: Rev. Mod. Phys. **33**, 249–264 (1961)
9. Cleri, F., Rosato, V.: Tight-binding potentials for transition metals and alloys. Phys. Rev. B **48**, 22–33 (1993)
10. Vervisch, W., Mottet, C., Goniakowski, J.: Theoretical study of the atomic structure of Pd nanoclusters deposited on a MgO(100) surface. Phys. Rev. B **65**, 245411 (2002)
11. Fortunelli, A., Barcaro, G.: Density-functional theory of free and supported metal nanoclusters and nanoalloys. In: Mariscal, M.M., Oviedo, O.A., Leiva, E.P.M. (eds.) Metal Clusters and Nanoalloys: From Modeling to Applications. Springer, Berlin (2013). doi:10.1007/978-1-4614-3643-0_2
12. Li, Z., Scheraga, H.A.: Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc. Natl. Acad. Sci. U. S. A. **84**, 6611–6615 (1987)
13. Wulff, G.: On the question of speed of growth and dissolution of crystal surfaces. Z. Krystallogr. **34**, 449–530 (1901)
14. Molina, L.M., Lee, S., Sell, K., Barcaro, G., Fortunelli, A., Lee, B., Seifert, S., Winans, R.E., Elam, J.W., Pellin, M.J., Barke, I., von Oeynhausen, V., Lei, Y., Meyer, R.J., Alonso, J.A., Rodriguez, A.F., Kleibert, A., Giorgio, S., Henry, C.R., Meiwes-Broer, K.H., Vajda, S.: Size-dependent selectivity and activity of silver nanoclusters in the partial oxidation of propylene to propylene oxide and acrolein: a joint experimental and theoretical study. Catal. Today **160**, 116–130 (2011)

15. Kaischew, R.: Arbeitstagung Festkorper Physik, Dresden, p. 81 (1952)
16. Ferrando, R., Rossi, G., Levi, A.C., Kuntová, Z., Nita, F., Barcaro, G., Fortunelli, A., Jelea, A., Mottet, C., Goniakowski, J.: Structures of metal nanoparticles adsorbed on MgO(001). I. Ag and Au. J. Chem. Phys. **130**, 174702 (2009)
17. Ferrando, R., Rossi, G., Nita, F., Barcaro, G., Fortunelli, A.: Interface-stabilized phases of metal-on-oxide nanodots. ACS Nano **2**, 1849–1856 (2008)
18. Repp, J., Meyer, G., Olsson, F.E., Persson, M.: Controlling the charge state of individual gold atoms. Science **305**, 493–495 (2004)
19. Negreiros, F., Barcaro, G., Sementa, L., Fortunelli, A.: Concepts in theoretical heterogeneous ultrananocatalysis. C. R. Chim. **17**, 625–633 (2014)
20. Barcaro, G., Causà, M., Fortunelli, A.: A comparison between the adsorption properties of the regular and F_s-defected MgO(100) surface. Theor. Chem. Acc. **118**, 807–812 (2007)

Simulations of Bulk Nanostructured Solids

Donald W. Brenner
Department of Materials Science and Engineering, North Carolina State University, Raleigh, NC, USA

Synonyms

Plasticity of nanostructured solids

Definition

A bulk nanostructured solid is a material available in macroscopic quantities that has nanometer-scale structural features in at least one dimension.

Computer modeling has emerged to become a third pillar of research alongside experiment and theory. There are many reasons for this emergence, including an exponential increase in processing speed, relatively inexpensive platforms for parallel computing and data storage, new visualization capabilities, and the development of powerful algorithms that take full advantage of these advances in hardware. Advances in simulation methodologies have also made the results of atomic-level computer modeling reliable to the extent that in many cases simulations can replace difficult and expensive experiments. Methods that enable atomistic dynamics using full electronic energies now allow processes involving up to several thousand atoms to be accurately modeled, with significantly larger systems on the horizon. Similarly, advances in materials theory have produced relatively simple analytic potential energy functions that can capture essentials of quantum mechanical bonding for simulations involving well over a billion atoms [1].

Computer modeling has played an especially important role in developing the current understanding of nanometer-scale structures and processes [2]. Indeed, among the many scientific and technological advances coming from nanotechnology is the ability for computer modeling and experiment to characterize phenomena on a common scale. Atomic-level computer modeling is now commonly used to explore and even predict new properties and processes that can be probed experimentally, to suggest new materials and structures with unique and desirable properties, to provide insight into the results of experiments, to generate data for larger-scale analysis, and to test scaling laws and analytic theories. In the case of nanostructured solids, computer modeling is allowing researchers to understand at the atomic level the structure, deformation mechanisms, and thermal-mechanical properties of these new materials with unprecedented detail [3, 4].

The two standard methods for modeling nanometer-scale systems are molecular dynamics and Monte Carlo simulation. In the former, coupled classical equations of motion for the atoms are integrated stepwise in time. Time steps generally range from one-half to tens of femtoseconds depending on the highest frequency vibration of interest, and simulations are typically carried out for picoseconds to tens of nanoseconds depending on system size, the phenomena to be studied, and the method in which interatomic forces are calculated. In a typical equilibrium Monte Carlo simulation, atomic configurations are generated with a probability that is proportional to their Boltzmann factor. Thermodynamic quantities in a Canonical ensemble are then obtained by averaging over the

properties of each configuration. Alternatively, in kinetic Monte Carlo simulation, time-ordered configurations are generated, typically using some rate expression.

Equilibrium Monte Carlo modeling requires specifying a potential energy as a function of atomic positions to calculate Boltzmann factors, while molecular dynamics simulations require interatomic forces, which are typically obtained as partial derivatives of the potential energy. In general, two approaches are used to calculate interatomic energies and forces. In the least computationally intensive approach, the interactions between atoms due to the electrons are replaced with effective interactions that are described with an analytic potential energy function. At present, there is no definitive mathematical form for the potential energy function, and forms that range from relatively simple pair-additive expressions to complicated many-body forms are used depending on the system, type of bonding, and phenomena to be studied. In the second approach to calculating potential energies, explicit electronic degrees of freedom are retained, the energy for which is calculated either using first principles methods or through a simplified semiempirical Hamiltonian. The calculation of total energies from first principles is well defined in terms of basis sets, electron correlation for ab initio methods, or choice of density functional expression (and pseudopotential) for density functional theory calculations. This is in contrast to analytic potentials, where a set of parameters (and often entirely new functional forms) must be developed for each system.

Nanostructured materials can be defined as materials that have at least one dimension in the "nanoscale" (typically 1–100 nm). Depending on which dimension this is, they can be classified into nanoparticles, layered or lamellar structures, filamentary structures, and bulk nanostructured materials. This entry is focused on the latter, which can be thought of as a traditional material with grain sizes at the nanometer scale. The nanometer-scale grains introduce unusual properties compared to more traditional micron-scale grain sizes, including unique combinations of strength and ductility as discussed in more detail below. The origin of these differences, and how they are being revealed by computer modeling, makes up the bulk of the material discussed below.

Atomic positions for bulk nanostructured solids have been generated by a number of different methods. Some of the methods are based on geometrical constraints imposed by the simulation conditions, for example, ensuring that active slip systems are properly oriented with respect to periodic boundaries [5, 6]. Crystallization dynamics have also been used to generate nanostructures. These methods can be based on a Johnson-Mehl or a Potts model construction, both of which produces a log normal grain size distribution, or by using a molecular dynamics simulation to model crystallization from a melt [3, 7, 8]. Other researchers have generated nanostructures by simulating compaction and sintering of nanoclusters [9–11]. Other methods use random grain centers, with grain boundaries chosen based on a Voronoi construction. Variations on this method include picking grain orientations to produce a particular range of tilt angle (e.g., low angle grains), or picking grain centers to produce a log normal distribution of grain sizes [4, 12].

Many of the atomic simulations of sintering and grain growth dynamics – and many of the simulation studies in general on nanostructured materials – have focused on understanding how processes and rates at the nanometer scale differ from their counterparts in macroscopic-scale systems. At the macroscopic scale, six distinct mechanisms contribute to the sintering dynamics of crystalline particles: surface diffusion; lattice diffusion from the surface, from grain boundaries and through dislocations; vapor transport; and grain boundary diffusion. There is a much larger degree of surface curvature at the nanometer scale and a much higher ratio of interface to bulk atoms, both of which may lead to very different sintering mechanisms. The details regarding these differences, and some new and unexpected effects at the nanometer scale, have been revealed from computer simulations. Molecular dynamics simulations, for example, have been used to study surface energies, grain boundary mobility, and sintering of metal nanoparticle arrays at different temperatures [13, 14]. The results suggest that of the macroscopic-scale mechanisms associated with sintering, only surface and

grain boundary diffusion contribute significantly to nanometer-scale sintering dynamics, and that these two processes are accelerated due to the large interfacial forces. Simulations have also suggested three unconventional mechanisms that contribute to the early stages of nanometer-scale sintering: mechanical rotation, plastic deformation via dislocation generation and transmission, and amorphization of subcritical grains. This grain rotation mechanism has also been observed in structures with columnar grains, where simulations suggest that necessary changes in the grain shape during grain rotation in the nanostructure can be accommodated by diffusion either through the grain boundaries or through the grain interior [15].

In conventional metals with micron-scale grains, plastic deformation occurs by motion of dislocations. This dislocation motion can be inhibited by grain boundaries, which leads to the well-established Hall–Petch relation that the yield strength is proportional to the inverse square root of the grain size. However, there appears to be a threshold grain size below which materials become softer with decreasing grain size. This so-called inverse Hall–Petch behavior has been attributed to a transition from dislocation-mediated plastic deformation to grain boundary sliding for some critical grain size. This transition has been observed in several sets of molecular dynamics simulations that predict a transition grain size of about 10–15 nm in good agreement with experimental measurements. At the same time, the simulations have also revealed a rich and unanticipated set of dynamics near the threshold region that can be related to the fundamental properties of the bulk materials. These unique dynamics include an enhanced role of grain rotation (analogous to sintering dynamics), cooperative inter-grain motion, and formation of stacking faults via motion of partial dislocations across grains.

The deformation of strained nanocrystalline copper with grain sizes that average about 5 nm has been studied with molecular dynamics simulations. These simulations showed a material softening for small grain sizes, in agreement with experimental measurements. These simulations suggest that plastic deformation in the inverse Hall–Petch region occurs chiefly by grain boundary sliding with dislocation motion having a minimal influence on deformation mechanisms. In related studies, molecular dynamics simulations have been used to understand the deformation of nanostructured nickel and copper with grain sizes ranging from 3.5 to 12 nm [5]. For grain sizes less than about 10 nm, deformation occurred mainly by grain boundary sliding, while for the larger grain sizes, deformation occurred by a combination of dislocation motion and grain boundary sliding. Detailed mechanisms of strain accommodation have been characterized; these include both single atom motion and correlated motion of several atoms, as well as stress-assisted free volume migration [16].

Molecular simulations of the deformation of columnar structures of aluminum have shown emission of partial dislocations during deformation that form at grain boundaries and triple junctions [17]. Atomic simulations have also suggested that these structures can be reabsorbed upon removal of the applied stress, which may contribute to an inability to observe dislocations experimentally in systems of this type after external stresses are released. In addition, near the grain size where plastic deformation transitions from dislocation-mediated plasticity to grain boundary sliding, the motion of single partial dislocations across nanograins during tensile loading has been observed in simulations. Without emission of a trailing partial dislocation, an intrinsic stacking fault is formed along the width of the nanograin. From the simulations, it appears that the formation of a low-energy ordered grain boundary drives emission of a full dislocation and the resulting absence of a stacking fault.

It has been argued that nucleation of the initial partial dislocation and the atomic rearrangement at the grain boundary associated with its emission sufficiently lowers the grain boundary energy such that emission of the trailing partial dislocation is not always needed to further relax the system [18]. Based on simulations of aluminum with a columnar nanostructure, it has been further suggested that the stacking fault width, and hence the intrinsic stacking fault energy, as determined by the distance between two partial dislocations, is the critical quantity that defines the transition from full to partial dislocation emission as grain

sizes approach the critical size for the onset of inverse Hall–Petch behavior [19]. On the other hand, it has been argued that the relation between the emission of partial dislocations does not correlate well with calculated stacking fault energies for nickel, copper, and aluminum. Instead, it has been suggested that full dynamics associated with the nucleation of a partial dislocation from a grain boundary must be considered, and therefore that the full planar fault energy, which includes the stable and unstable stacking fault energy as well as twin fault energies, must be used to understand and ultimately predict the relation between mechanical deformation and grain size.

Extensive simulations of crack propagation in nanostructured metals have also been carried out to better understand how fracture, fatigue, and toughness depend on grain scale [20–22]. These simulations have revealed crack propagation mechanisms that are similar to the plastic response of fully dense samples as well as key differences resulting from the presence of the crack tip. For example, in simulations of nanocrystalline nickel with grain sizes in the inverse Hall–Petch region, mode I crack propagation occurred by inter-grain decohesion via a mechanism involving coalescence of nanovoids that form in front of the crack tip. Plastic deformation leading to both full and partial dislocations was also observed in the neighboring grains.

Compared to their mechanical properties under tensile loading, much less is understood about the influence of grain size on the shock loading properties of nanostructured solids. Atomic simulations that have been carried out, however, suggest a strong coupling between nanostructure and shock loading dynamics, as well as unique and very important properties of shocked-loaded nanostructured metals. It has been noted, for example, that the mechanisms associated with the mechanical deformation of nanostructured metals depend strongly on pressure, temperature, and strain rate, and therefore, these materials may show ultrahigh strength under shock loading depending on the shock loading conditions and system [23]. The fast temperature and pressure rises associated with shock fronts freeze out deformation mechanisms that require diffusion. Similarly, production of dislocations that requires nucleating events is inhibited. In the case of grain boundary accommodation, increasing the pressure results in an increase in the threshold stress for sliding plasticity. Similarly, the threshold for dislocation plasticity increases with increasing pressure because of an increase in the shear modulus with increasing pressure. Taking these effects into account, and assuming that the maximum in hardness as a function of grain size occurs where stress for sliding and dislocation plasticity are equal, it has been shown that ultrahigh hardness can be achieved by shock loading of nanostructured solids. These arguments have been validated by using molecular dynamics simulations to model the shock loading of nanostructured copper with different grain sizes. At relatively low shock velocities (i.e., low stresses), grain boundary sliding is observed, which results in a relatively low hardness value that increases with increasing grain size. At intermediate stresses, the hardness of the copper increases with increasing shock strength for all grain sizes, with a shift in the maximum hardness toward lower grain sizes compared to deformation at lower strain rates. This leads to a net increase in the maximum hardness of the material. At even higher stresses, simulations predict a drop in strength due to an increase in temperature and an associated increase in dislocation nucleation and motion.

It is clear from the discussion in the preceding sections that atomic simulations have provided new and exciting insights into the unique properties of nanosystems in general and nanostructured solids in particular. This remains a very active area of research within which molecular simulation will continue to provide new insights into relations between structural mechanical and thermodynamic properties.

Cross-References

- ▶ Computational Study of Nanomaterials: From Large-scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Molecular Dynamics Simulations of Nanobiomaterials
- ▶ Nanomechanical Properties of Nanostructures

References

1. Brenner, D.W., Shenderova, O.A., Areshkin, D.A.: Quantum-based analytic interatomic forces and materials simulation. In: Lipkowitz, K.B., Boyd, D.B. (eds.) Reviews in Computational Chemistry, pp. 213–245. VCH, New York (1998)
2. Brenner, D.W., Shenderova, O.A., Schall, J.D., Areshkin, D.A., Adiga, S., Harrison, J.A., Stuart, S. J.: Contributions of molecular modeling to nanometer-scale science and technology, Chapter 24. In: Goddard, W., Brenner, D., Lyshevski, S., Iafrate, G. (eds.) Nanoscience, Engineering and Technology Handbook. CRC Press, Boca Raton (2002)
3. Wolf, D., Yamakov, V., Phillpot, S.R., Mukherjee, A., Gleiter, H.: Deformation of nanocrystalline materials by molecular dynamics simulations: relation to experiment? Acta Mater. **53**, 1 (2005)
4. Kumar, K.S., Van Swygenhoven, H., Suresh, S.: Mechanical behavior of nanocrystalline metals and alloys. Acta Mater. **51**, 5743 (2003)
5. Van Swygenhoven, H., Spaczer, M., Caro, A., Farkas, D.: Competing plastic deformation mechanisms in nanophase metals. Phys. Rev. B **60**, 22 (1999)
6. Yanakov, V., Wolf, D., Salazar, M., Phillpot, S.R., Gleiter, H.: Length-scale effects in the nucleation of extended dislocations in nanocrystalline Al by molecular-dynamics simulation. Acta Mater. **49**, 2713 (2001)
7. Phillpot, S.R., Wolf, D., Gleiter, H.: Molecular-dynamics study of the synthesis and characterization of a fully dense, three-dimensional nanocrystalline material. J. Appl. Phys. **78**, 847 (1995)
8. Phillpot, S.R., Wolf, D., Gleiter, H.: A structural model for grain boundaries in nanocrystalline materials. Scripta Metall. Mater. **33**, 1245 (1995)
9. Zhang, Y.W., Liu, P., Lu, C.: Molecular dynamics simulations of the preparation and deformation of nanocrystalline copper. Acta Mater. **52**, 5105 (2004)
10. Kodiyalam, S., Kalia, R., Nakano, A., Vashashta, P.: Multiple grains in nanocrystals: effect of initial shape and size on transformed structures under pressure. Phys. Rev. Lett. **93**, 203401 (2004)
11. Chatterjee, A., Kalia, R.K., Nakano, A., Omeltchenko, A., Tsuruta, K., Vashishta, P., Loong, C.-K., Winterer, M., Klein, S.: Sintering, structure, and mechanical properties of nanophase SiC: a molecular-dynamics and neutron scattering study. Appl. Phys. Lett. **77**, 1132 (2000)
12. Schiotz, J., Di Tolla, F.D., Jacobsen, K.W.: Softening of nanocrystalline metals at very small grain sizes. Nature **391**, 1223 (1998)
13. Zeng, P., Zajac, S., Clapp, P.C., Rifkin, J.A.: Nanoparticle sintering simulations. Mater. Sci. Eng. A **252**, 301 (1998)
14. Xiao, S., Hu, W.: Molecular dynamics simulations of grain growth in nanocrystalline Ag. J. Cryst. Growth **286**, 512 (2006)
15. Haslam, A.F., Phillpot, S.R., Wolf, D., Moldovan, D., Gleiter, H.: Mechanisms of grain growth in nanocrystalline fcc metals by molecular-dynamics simulation. Mater. Sci. Eng. A **318**, 293 (2001)
16. Van Swygenhoven, H., Derlet, P.M.: Grain boundary sliding in nanocrystalline fcc metals. Phys. Rev. B **64**, 224105 (2001)
17. Yamakov, V., Wolf, D., Salazar, M., Phillpot, S.R., Gleiter, H.: Length scale effects in the nucleation of extended dislocations in nanocrysalline Al by molecular dynamics simulation. Acta Mater. **49**, 2713 (2001)
18. Van Swygenhoven, H., Derlet, P.M., Froseth, A.G.: Stacking fault energies and slip in nanocrystalline metals. Nat. Mater. **3**, 399 (2004)
19. Yamakov, V., Wolf, D., Phillpot, S.R., Mukherjee, A. K., Gleiter, H.: Deformation mechanism map for nanocrystalline metals by molecular dynamics simulation. Nat. Mater. **3**, 43 (2004)
20. Latapie, A., Farkas, D.: Molecular dynamics investigation of the fracture behavior of nanocrystalline α-Fe. Phys. Rev. B **69**, 134110 (2004)
21. Farkas, D., Van Petegem, S., Derlet, P.M., Van Swygenhoven, H.: Dislocation activity in nano-void formation near crack tips in nanocrystalline Ni. Acta Mater. **53**, 3115 (2005)
22. Farkas, D., Sillemann, M., Hyde, B.: Atomistic mechanisms of fatigue in nanocrystalline metals. Phys. Rev. Lett. **94**, 165502 (2005)
23. Bringa, E.M., Caro, A., Wang, Y., Victoria, M., McNaney, J.M., Remington, B.A., Smith, R.F., Torralva, B.R., Van Swygenhoven, H.: Ultrahigh strength in nanocrystalline materials under shock loading. Science **309**, 1838 (2005)

Simulations of Oxide/Polymer Hybrids

Maria Ilenia Saba and Alessandro Mattoni
Istituto Officina dei Materiali, Consiglio Nazionale delle Ricerche, CNR-IOM Cagliari (SLACS), Monserrato, CA, Italy

Synonyms

Atomistic simulations; Molecular dynamics; Organic/inorganic interactions

Definitions

Polymer/oxide hybrids are systems composed mainly by polymers and oxides.

Oxides are intended here as a class of inorganic solid compounds having at least one oxygen and one other element (e.g., metal or semimetal) in their chemical formula. They typically contain the oxygen anion in the -2 oxidation state. Oxides are abundant in nature; most of the Earth's crust consists of solid oxides, due to the oxidation of the elements in air or in water.

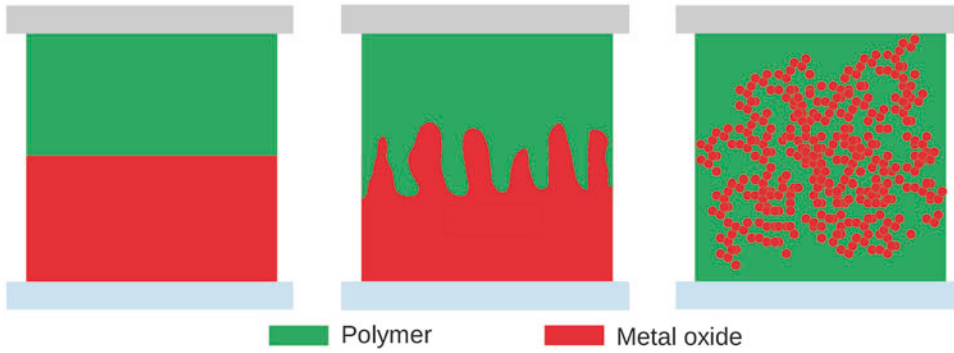
Polymers are a class of natural or synthetic chemical compounds formed by the repetition of small units (monomers). Polymers formed by organic units are named organic polymers. An important class of polymers are the conjugated ones, having a backbone of alternating double and single bonds that provides conductive properties. Polymer/oxide systems have applications in bioengineering, textiles, magnetism, optoelectronics, and photovoltaics.

Molecular dynamics is a computational technique that consists in calculating the classical trajectories of a set of interacting atoms representing the material of interest. The trajectories obey the Newton's equation of motions $\left(\vec{F} = m\ \vec{a}\right)$, where the interatomic forces $\vec{F}$ are calculated as the derivative of a suitable potential. In model potential molecular dynamics, the potential does not depend explicitly on the electronic degrees of freedom, but it is an empirical function of the nuclei positions with parameters that are calibrated to reproduce a set of physical properties of the material. Model potential molecular dynamics is a valuable tool to study hybrid polymer/oxide systems since it makes affordable the study of large-scale atomistic models with complex morphologies while retaining an atomistic (i.e., fundamental) description of the material.

Technological Relevance of Polymer/Oxide Hybrids

Polymer/oxide materials have high technological impact in textiles, bioengineering, molecular electronics, catalysis, magnetism, optoelectronics, and photovoltaics [1]. Their hybrid nature makes possible to join, at a low production cost, the tailorable properties and the flexibility of the organic polymers with the thermal and mechanical stability or the good transport properties of the inorganic components. A large number of synthetic or natural polymers can be used in hybrids, including organic, inorganic, and biopolymers. In technical textiles, for example, organic polymers such as polyolefins, nylons, polyesters, and polyurethanes are widespreadly used because of easy fabrication, processability, durability, and relatively low cost [2]. The inorganic particles improve the tensile strength or stiffness via reinforcement mechanisms described by theories for nanocomposites. The addition of inorganic particles (such as metal oxides and silica) not only provides mechanical and thermal stability but also new functionalities that depend on the chemical nature, the structure, the size, and the crystallinity of the inorganic nanoparticles. For example, titania (TiO_2) can be mixed with poly(amide-imide) forming composite membranes for gas separation or with poly(3-hexylthiophene) for photovoltaics; vanadium oxide (V_2O_5) can be used with poly(3,4-ethylenedioxythiophene) for cathode materials for rechargeable lithium batteries; zinc oxide (ZnO) particles with polystyrene resin are useful for electrical applications and with poly(3-hexylthiophene) for optoelectronics and photovoltaics; alumina (Al_2O_3) mixed with polycarbonate can be applied as optical materials; and Fe_3O_4 with poly(vinylidene difluoride) is useful for magnetic nanocomposite and γ-Fe_2O_3 with polyimide for superparamagnetic films. Finally, silica (SiO_2) can be combined with polycaprolactone, polyimide, poly(methyl methacrylate), polyethyl acrylate, poly(p-phenylene vinylene), and polycarbonate giving rise to hybrids applicable for bone-bioerodible and skeletal tissue repair, microelectronics, dental application and optical devices, catalysis support and chromatography, nonlinear optical material for optical waveguides, and abrasion-resistant coating, respectively [1–3].

Among polymers, conductive ones are the most relevant when considering hybrids for optoelectronics and photovoltaics. They are characterized by a backbone formed by conjugated chemical groups with alternating double and single bonds, along which the charge carriers



Simulations of Oxide/Polymer Hybrids, Fig. 1 Microstructure of polymer/metal oxide hybrids: bilayers (*left*), polymer/porous metal oxide bulk heterojunction (*center*), and polymer/nanocrystal bulk heterojunction (*right*)

(in most cases holes) can diffuse easily. One of the most commonly used is poly(3-hexylthiophene) (P3HT) that has good hole mobility ($0.1\ cm^2\ V^{-1}\ s^{-1}$) and thermal stability. The backbone of P3HT consists of π-conjugated thiophenes, responsible for the good intrachain hole mobility, and alkyl side chains, providing polymer solubility and processability in several solvents.

Concerning oxides, the most used for hybrid polymer/oxide composites are ZnO, TiO_2, and SiO_2. Zinc oxide (ZnO) is a wide band gap semiconductor (3.37 eV) that crystallizes in two main forms: hexagonal wurtzite (the most common at ambient conditions) and cubic zincblende. ZnO has very good electron mobility ($205\ cm^2\ V^{-1}\ s^{-1}$); it is nontoxic, and it can be grown in a variety of highly crystalline nanostructures which are commonly used as electron acceptors. Nanorods with $(10\bar{1}0)$ nonpolar surfaces are commonly used in hybrid bulk heterojunctions for photovoltaics [4].

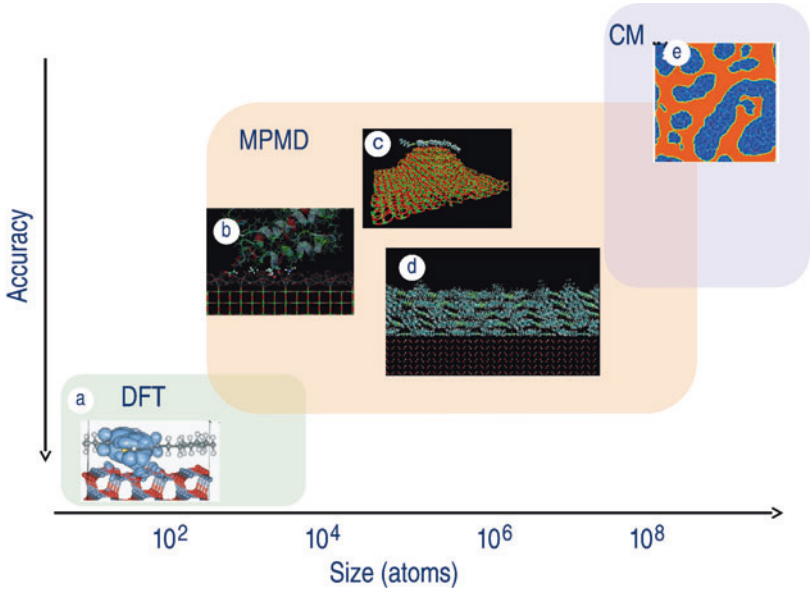
Titanium dioxide (TiO_2) is a semiconductor used as sensor, pigment, and photocatalyst and in solar cell for the production of hydrogen and energy. TiO_2 crystallizes in three major structures: tetragonal rutile (the most common and stable), anatase, and rhombohedrical brookite. The most stable crystalline surfaces are the (110) for the rutile phase and the (101) for the anatase one.

Silica (SiO_2) is one of the most abundant oxides in nature. In addition to its amorphous disordered phases, it has a number of crystalline structures (polymorphs). The only stable crystalline structure under normal thermodynamic conditions is the α-quartz that is formed by SiO_4 tetrahedra linked by shared corner oxygen atoms. Due to its stability and biocompatibility, SiO_2 is largely used in biomedical applications.

Atomistic Modeling of Hybrid Systems

The microstructure of polymer/metal oxide hybrids can be very different depending on the synthesis methods, on the technological applications, and on the compatibility between the organic and the inorganic components. Simple planar interfaces (also called bilayers; see Fig. 1 left) are sometimes used for sensors, optoelectronic devices, or coatings. More complex microstructure can be found in hybrids for structural or optoelectronic applications. For example, in the bulk heterojunction architectures used in photovoltaics, the polymer is infiltrated within a mesoporous oxide scaffold (Fig. 1 center) or mixed with oxide nanoparticles (right) so forming two interpenetrated networks of polymer and oxide. Such kind of microstructures provides a large interface-to-volume ratio necessary to optimize the light absorption and the collection of carriers through percolation paths to the contacts.

Due to the complex polymer/metal oxide microstructure, the overall macroscopic properties of the hybrid depend on several physical and chemical processes occurring at different length



Simulations of Oxide/Polymer Hybrids, Fig. 2 Accuracy versus size in atomistic simulation of polymer/metal oxide hybrids: (*a*) ab initio P3HT/ZnO system (Adapted with permission from Dag et al., Nano Letters 2008 8 (12), 4185–4190. Copyright (2008) American Chemical Society), (*b*) protein adsorbed on TiO_2 surface studied by molecular dynamics (Adapted with permission from Kang et al. J. Phys. Chem. C 2010 114 (34), 14496–14502. Copyright (2010) American Chemical Society), (*c*) snapshot of MPMD simulation of a P3HT polymer chain on TiO_2 surface (Adapted with permission from Melis et al. J. Phys. Chem. C 2010 114 (8), 3401–3406, Copyright (2010) American Chemical Society), (*d*) P3HT/ZnO interface studied by MPMD (Adapted with permission from Saba et al. J. Phys. Chem. C 2014 118 (9), 4687–4694. Copyright (2014) American Chemical Society), (*e*) snapshot of the time evolution of Janus nanorod-filled binary polymer mixture investigated by using a cell dynamics system (CDS) method (Adapted with permission from Li et al. Macromolecules, 2013, 46 (18), pp 7465–7476. Copyright (2013) American Chemical Society)

and timescales: at the mesoscopic scale, where the material can be described as a continuum blend, and down to the atomic scale, where chemical bonds, electrostatic, and dispersive forces between atomic constituents control, for example, the interface properties.

Atomistic methods are necessary to model predictively the physical properties of hybrid materials and are complementary to the experimental characterization, as well as the continuum modeling approaches. Among the other atomistic methods, the model potential molecular dynamics (MPMD) is a powerful tool to study the complex microstructure of hybrid materials since, while retaining an atomistic (i.e., fundamental) description, it makes affordable the study of relatively large systems (typically composed by 10^4–10^6 atoms). At variance with ab initio molecular dynamics, that is typically limited to systems as small as 10^2–10^3 atoms because of the high computational cost associated to the explicit treatment of electrons, in MPMD the electronic degrees of freedom are integrated into an effective interatomic potential that only depends on the nuclei positions. The accuracy and the transferability of the method must be validated in each case, but the computational workload is highly reduced. It typically scales linearly with the number of atoms (order-N method), making possible to study the thermodynamics and the microstructure evolution of systems as large as 10^1–10^3 nm (up to several millions of atoms) over the 10^1–10^3 ns timescale (see Fig. 2). MPMD is a suitable method to study hybrid systems where a large number of atoms are necessary to represent the structural disorder induced by macromolecules and interfaces. Furthermore, long-range electrostatic and dispersive interactions that are

important in hybrid materials are easily incorporated into MPMD by using two-body Coulomb and Lennard-Jones-type potentials, respectively. The drawback of the method is that the accurate description of interatomic forces is not guaranteed and careful calibrations against experimental data or ab initio calculations are necessary.

The Force Fields for Oxide/Polymer Systems

The interatomic potential for hybrid polymer (P)/oxide (X) systems to be used in model potential molecular dynamics can be written in most cases as the sum of organic–organic, organic–inorganic, and inorganic–inorganic interactions:

$$U = U_{PP} + U_{PX} + U_{XX}$$

The basic requirement of any interatomic potential for hybrid systems is to provide a good description of the separate organic and inorganic phases.

Several potentials exist for the U_{PP} term that describes the chemical bonding between atoms of the organic polymers, as well as intermolecular interactions. A possible classification of the U_{PP} force fields is based on the number of atomic degrees of freedom simulated in the calculation. All-atom force fields treat explicitly all the atoms of the system, including hydrogen; this is the case of AMBER (Assisted Model Building with Energy Refinement) [5], CHARMM (Chemistry at HARvard Molecular Mechanics) [6], and OPLS (Optimized Potentials for Liquid Simulations) [7] force fields. United-atom force fields (such as Gromos (GROningen MOlecular Simulation package) [8]) do not include explicit representation of nonpolar hydrogen atoms; only polar hydrogens are included in the force field definition. Finally, in the coarse-grained force fields, specific chemical groups are replaced by a single or a few pseudo-atoms representing the collective behavior of the group. In this way, the number of degrees of freedom of the simulation is smaller and the computational workload is strongly reduced making affordable larger systems and timescale. An example of coarse-grained force field developed for biomolecular systems is the Martini force field [9]. It has been parameterized to reproduce the partitioning free energies between polar and apolar phases of a large number of chemical compounds.

The AMBER force field is the typical all-atom classical force field to study organic materials. It takes into account bonded and nonbonded interactions [5]. The former are necessary to model the chemical bonds between atoms. The list of interacting atoms must be specified according to the topology of the molecules and it is fixed during the simulation (molecular mechanics). Bonding interactions consist of three contributions [5]:

$$U_{\text{bonded}} = U_{\text{bonds}} + U_{\text{angles}} + U_{\text{dihedrals}} \qquad (1)$$

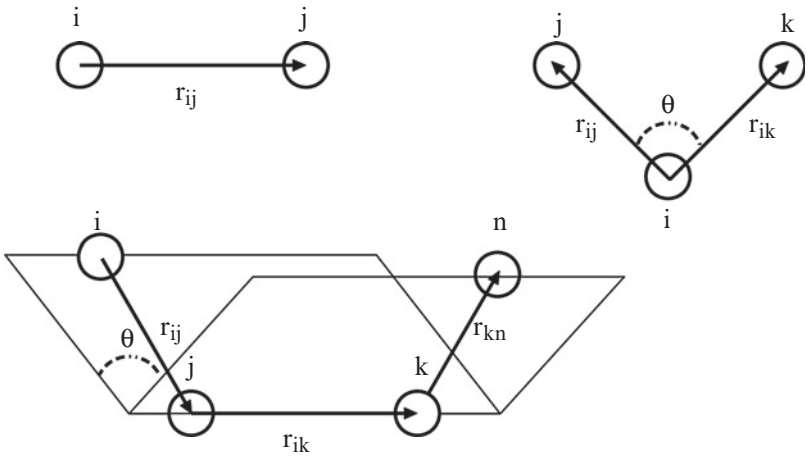
$U_{\text{bonds}} = \sum_{ij} \frac{1}{2} K_{ij}^{b} \left(r_{ij} - r_{ij}^{0} \right)^2$ describes the two-body elastic part of the bonding. K_{ij}^{b} is the constant of the force, r_{ij}^{0} is the equilibrium distance between two atoms ij, and r_{ij} is the corresponding instantaneous distance of the bonded atoms (Fig. 3 top left).

$U_{\text{angle}} = \sum_{ijk} \frac{1}{2} K_{ijk}^{a} \left(\theta_{ijk} - \theta_{ijk}^{0} \right)^2$ describes the energy cost of the deformation of the angle formed by three adjacent atoms ijk. K_{ijk}^{a} and θ_{ijk}^{0} are the stiffness and the equilibrium angle, and θ_{ijk} is the instantaneous bond that they form (Fig. 3 top right).

$U_{\text{dihedral}} = \sum_{ijkn} \frac{1}{2} V_{ijkn}^{\phi} \left(1 + \cos \left(n_{ijkn} \phi_{ijkn} - \phi_{ijkn}^{0} \right) \right)$ describes the interaction arising from torsional forces in molecules. They are sometimes referred to as torsion potentials and require the specification of four atomic positions representing the instantaneous angle between the plane defined by the atoms ijk and that defined by the atoms jkn. V_{ijkn}^{ϕ}, ϕ_{ijkn}^{0}, and n_{ijkn} are the stiffness, the equilibrium dihedral angle, and the multiplicity, respectively, associated with the torsion involving $ijkn$ types of atoms (Fig. 3 bottom).

Improper dihedrals, involving nonconsecutive atoms, can also be considered and are generally used to enforce planarity around sp^2 central atoms [5].

The bonded energy U_{bonded} prevents bond breaking, and it is suitable for describing phenomena in which the chemical groups do not dissociate during the simulation.



Simulations of Oxide/Polymer Hybrids, Fig. 3 Bonding (*top left*), angular (*top right*), and dihedral (*bottom*) interaction between two, three, and four atoms

The nonbonded terms make it possible to establish interactions between atoms that are not chemically bonded or that are separated by three o more bonds. It is modeled as a sum between the Coulombic interactions between atomic partial charges q_i and the van der Waals forces described by the Lennard-Jones 12–6 potential:

$$U_{\text{non-bonded}} = U_{vdW} + U_{\text{Coul}} = \sum 4\epsilon_{ij}\left(\frac{\sigma_{ij}^{12}}{r_{ij}} - \frac{\sigma_{ij}^{6}}{r_{ij}}\right) + \sum \frac{q_i q_j}{4\pi\varepsilon r_{ij}} \tag{2}$$

where ϵ is the depth of the van der Waals potential well, σ is the finite distance at which the interparticle potential is zero, and r_{ij} is the distance between the particles and ε is the dielectric constant. The r_{ij}^{-12} term describes the Pauli repulsion at short distances due to electron overlapping orbitals, while the attractive long-range r_{ij}^{-6} term corresponds to the interactions between fluctuating dipoles (dispersion forces). Other U_{PP} potentials such as CHARMM and OPLS force fields have a functional form very similar to that of AMBER.

There exists a second generation of interatomic potentials with higher-order and nondiagonal mixing terms for the bonded interactions [10]. Among them, there are CFF [11], UFF [12], and the MM series force fields of Allinger [13]. Derived from CFF, COMPASS [14] force field combines ab initio results and the empirical fitting of condensed-phase properties to obtain parameters for valence, charge, and van der Waals terms.

Another classification for the force fields is related to the way in which the atomic charge is treated. In first-generation force fields, the charge is kept fixed, usually centered on atoms, without taking into account the electrostatic environment and the electronic polarization. In the last years, a new generation of force field incorporating models for polarizability has been proposed. The most studied approach is the point dipole method [10], included, for example, in the AMOEBA (Atomic Multipole Optimized Energetics for Biomolecular Applications) [15] force field, where polarization energy is achieved by the interactions between static charges and their induced dipole moments [10]. In the fluctuating charge model (e.g., included in the UFF and CHARMM force fields), a charge flows between atoms equalizing their electronegativities. In this way, a molecule's charge-state distribution is coupled to its environment and the effects of polarization are included [10]. In Drude oscillator methods (or shell methods), the electronic polarizability is included adding an extra charged site to each ion connected via a spring. The magnitude of both charges is fixed, and the electronic polarization is mimicked by relative displacement of both charges due to an external electrostatic field [10].

Concerning the inorganic–inorganic term U_{XX}, it is usually of nonbonded type. For example, for ZnO and TiO_2, U_{XX} can be modeled by adding to the long-range electrostatic interactions, a two-body Buckingham potential [16, 17]

$$U_{\text{buck}} = \sum_{ij} A\exp\left(-\frac{r_{ij}}{B}\right) - \frac{C}{r_{ij}^6} \qquad (3)$$

where A, B, and C are fitting parameters. The atomic partial charges used in this approach are fixed and polarizability effects can be included, for example, by using shell methods.

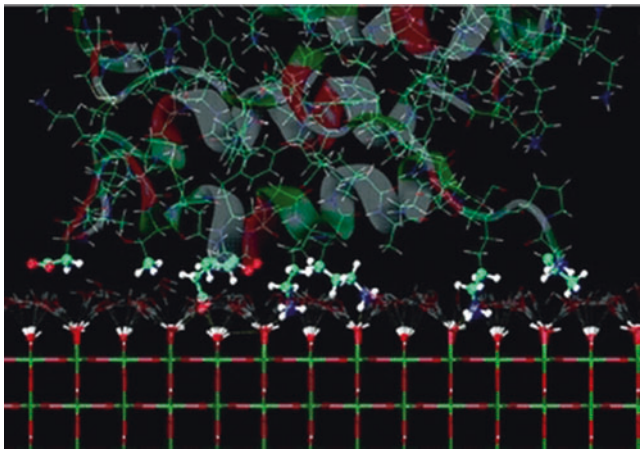
Many-body effects can be included by environment-dependent potentials, such as the reactive force field (ReaxFF) [18]. It is a bond-order interaction model with two-body, three-body, and four-body short-range interaction terms. ReaxFF have been applied to ZnO [18] and TiO_2 [19]. It allows the redistribution of charges, can simulate the breaking and reforming of bonds, and can reproduce the structures and mechanical properties of condensed phases but requires a high computational cost and a very large number of fitting parameters.

As for the modeling of SiO_2, several force field have been applied, in order to take into account its significant ionic contribution and the strong, directional covalent bond. Woodcock et al. were the first to perform simulations on vitreous silica using an ionic pair potential [20], while Feuston and Garofalini introduced the representation of the covalent bonding by a three-body interatomic potentials [21]. Vessal et al. used full ionic charges and developed two different models to take into account the directionality of the Si–O bond [22]. In both cases, they used a two-body (a Buckingham potential) plus a three-body interaction (called screened and truncated) [22]. The Beest, Kramer, and van Santen (BKS) model uses partial charges to simulate partial covalent bonds and short-range interactions of the Buckingham form [23]. Morse potentials [24, 25] as well as ReaxFF [26] have been developed for silica.

The last potential contribution of hybrid polymer/oxides is the U_{PX} that contains the cross atomic interactions between polymer and the oxide. In most cases, there are not covalent bonds between the polymer and the oxide (unless specific ligands with anchoring groups are present). Under these conditions, the interaction can be treated by nonbonding forces of dispersive or electrostatic nature, provided that a proper fit of the charges is used. U_{PX} models of nonbonded type can be successfully applied to reproduce the interaction of aromatic rings (e.g., thiophene and benzenes within conjugated polymers) with an oxide surface (e.g., ZnO, TiO_2, SiO_2). Also in the presence of chemical ionic binding between the polymer and the oxide, the same approach can be effective. Most organic ligands such as fatty acids (e.g., oleic acid) have a carboxylic COOH terminal group able to anchor to the oxide surface. MPMD simulations are able to reproduce monodentate and bidentate COO−X anchoring modes of protonated and deprotonated molecules for several ligands. The same strategy can be successfully employed for POOH−X interactions and heterocycles containing oxygen (e.g., THF [27]) or nitrogen (such as pyridines on titania [28], for which an effective N–Ti binding is reproduced), as well as for organometallic compounds such as zinc phthalocyanine on ZnO [29].

As for the computational cost of the above force fields, the nonbonded electrostatic interactions are in most cases the most expensive, with the direct computation time proportional to the square of the number of atoms N^2. An efficient method to spare computational time is represented by the Ewald summation. In this method from each point charge spreads out radially, up to a cutoff distance, a Gaussian charge distribution of the same magnitude and opposite sign. The interaction is separated into short-range interaction, representing the screening interaction between neighboring charges and calculated in the real space, and long-range one, representing the canceling charge distribution of the same sign as the original charge calculated in the reciprocal space [30]. In this way, the time calculation is proportional to $N^{3/2}$. By calculating the Ewald summation using the fast Fourier transform (FFT) technique (particle-mesh Ewald method), the computational workload can scale as $N\log(N)$.

Concerning the cost of the U_{PP} part, in most cases, its terms (even in the case of environment-



Simulations of Oxide/Polymer Hybrids, Fig. 4 Protein adsorption on a hydroxylated TiO_2 surface (Adapted with permission from Kang et al. J. Phys. Chem. C 2010 114 (34), 14496–14502. Copyright (2010) American Chemical Society)

dependent contributions) decrease fast with interatomic distances and they are set to zero beyond a cutoff length. Accordingly, the overall workload of the force fields $U = U_{PP} + U_{PX} + U_{XX}$ scales linearly with the number of atoms N in the hybrid system.

Applications of Atomistic Methods to Hybrid Systems

Atomistic simulations can be applied to study a large number of physical properties of interest for hybrids such as adhesion and interface energies, molecular order and crystallinity, diffusivity and assembling phenomena, vibrational properties, solvent effects, and so on. In most atomistic studies, the central problem is related to the modeling of the polymer/oxide interface from planar interfaces to bulk heterojunctions. Below we illustrate a few paradigmatic examples for biology, photovoltaics, and optoelectronics.

Atomistic Studies for Biological Applications

An important problem related to hybrid systems for biology is the interaction of biopolymers (such as proteins) with oxides and, in particular, with TiO_2. Titanium has been extensively employed for medical implants such as dental ones, artificial joints, and blood-contacting devices. It is regarded as a bioactive material with good biocompatibility, which mainly derives from the presence of the surface oxide film that protects the metal from further oxidation [31]. Biocompatibility requires to selectively control the hybrid interactions and the adsorption between the biopolymers and the implant surfaces. The interactions among the proteins, the TiO_2, surface, and the solvent, as well as the dynamic mechanism of the adsorption/desorption process, are the key points to design implants with optimal bioactivity and antifouling properties (i.e., inhibiting the accumulation of undesired materials on wetted surfaces through nonspecific interactions).

Many proteins such as albumin, fibronectin, and laminin have been found to be able to adsorb onto titanium dioxide surfaces. For example, Kang et al. [31] have studied by model potential molecular dynamics the role of hydroxylation in the absorption of human serum albumin onto TiO_2 in water environment (Fig. 4). The model potential adopted was the AMBER force field for the protein, the Buckingham potentials for the titania, and the Lennard-Jones and Coulombic interactions for the hybrid interactions. By molecular dynamics simulations, it was shown that the hydroxylated surface has a stronger interaction with the protein and provides a larger number of adsorbed residues with respect to the non-hydroxylated case. The hydroxyl modification makes it easier for the protein to move closer to surface by reducing the hydrogen bonds between the surface and the water molecules.

The same model potential for TiO_2 has been used in the study of the titania interaction with the biopolymers surrounding the cell membranes.

In particular, Ham et al. by using the OPLS force field investigated the adhesion on titania of a glycopeptoid that mimics the composition and abilities of glycocalyx [32]. This is a glycoprotein–polysaccharide with antifouling properties that surrounds the cell membranes of some bacteria, epithelia, and other cells. The computational results revealed an aqueous interface enriched in highly hydrated saccharide residues with the formation of a high number of hydrogen bonds with water molecules possibly contributing to fouling resistance.

Biopolymeric films on metal oxides have been also exploited in preventing blood coagulation and thrombosis. A class of copolymers based on poly(l-lysine)-g-poly(ethylene glycol) (PLL-g-PEG) was found to spontaneously adsorb from aqueous solutions onto several metal oxide surfaces, such as TiO_2, $Si_{0.4}$ $Ti_{0.6}O_2$, and Nb_2O_5 [33]. The resulting adsorbed layers are highly effective in reducing the adsorption both of blood serum and of individual proteins such as fibrinogen, which is known to play a major role in the cascade of events that lead to biomaterial-surface-induced blood coagulation.

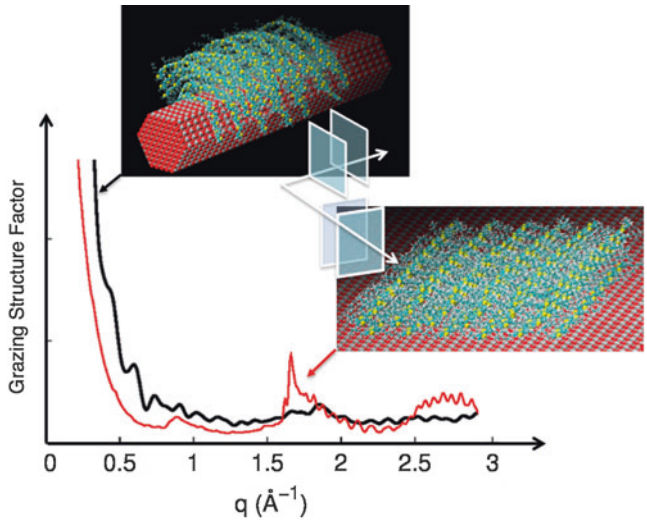
Though classical molecular dynamics is the method of choice for most atomistic investigations in biophysics, the applications to polymer/oxide systems are relatively few. A rapid growth of the field is expected in the next years.

Atomistic Studies for Photovoltaic and Optoelectronic

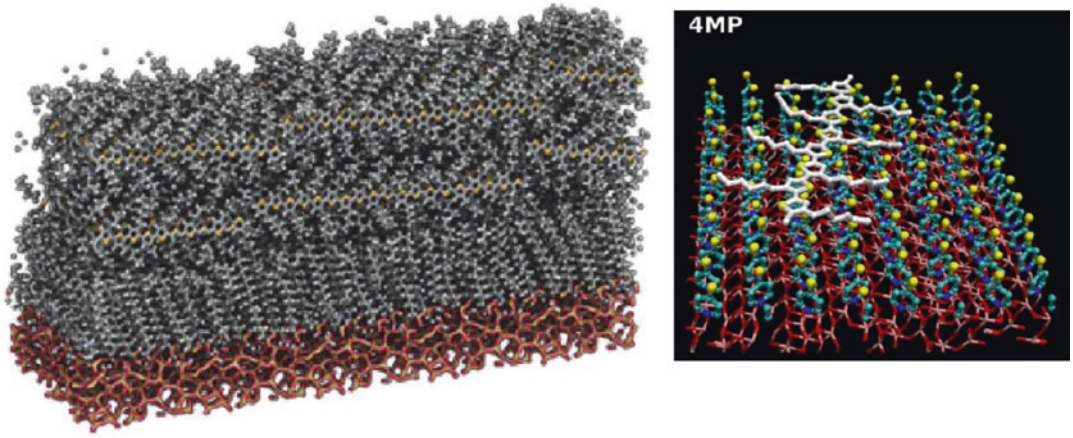
A class of polymer/oxide hybrid films formed by organic conductive polymers (such as P3HT) and metal oxides (typically TiO_2 or ZnO) have attracted a large interest as candidates for environmentally friendly and solution-processable films for low-cost photovoltaics. In these hybrids, the polymer and the metal oxide form a donor–acceptor pair with photovoltaic properties in which the polymer absorbs the sunlight, donates the electrons to the oxide, and transports the holes to the contact. The electrical and optical properties of the system are very susceptible to the microstructure and to the crystalline assembling of the molecular constituents. There is in fact a strong dependence of the intermolecular electronic properties (e.g., transfer integral) on the ordered stacking of the polymer chains. When the polymer order is lost, as is the case of amorphous polymers, an increase of resistivity and a blue shift of the light absorption are observed. Accordingly, the polymer order and crystallinity at the polymer/oxide interface has been the subject of intense investigation by atomistic methods.

In this context, one example is that of P3HT/ZnO hybrid interface that has been studied to understand the spontaneous formation of disordered polymer close to the interface with the oxide [34]. By atomistic simulations (based on AMBER, Buckingham, and Coulombic plus dispersive model potentials), it has been shown that the polymer/ZnO adhesion is driven by the electrostatic interactions between the polymer backbones and the oxide. On planar ZnO substrates (Fig. 5 right inset), though some disorder is present, the polymer films can be crystalline with predominance of the face-on molecular orientation. However, the local curvature of the oxide substrate can induce the bending of the polymer backbones [35], perturbing the molecular film with the possible formation of amorphous layers (Fig. 5 top inset). Specific methods to analyze the degree of order/disorder in the atomistic models are necessary, such as the structure factor analysis. It has been applied to the P3HT/ZnO surface showing a decrease of the diffraction peak of the polymer backbones with the substrate curvature (Fig. 5), in agreement with experimental evidences.

A further important problem of interest for atomistic simulation is related to the use of interfacial layers between the oxide and the polymers. The use of the interlayer represents a versatile approach to engineer and design hybrid interfaces with controlled order and improved properties. For example, the case of P3HT polymer on SiO_2 that is relevant in the field of hybrid and organic transistors has been studied by Meredig et al. (see left panel of Fig. 6). By using the COMPASS force field, they have shown that the presence of a disorder self-assembled monolayer (SAM) of octadecyltrichlorosilane (OTS) between the polymer and the silica substrate [36] induces a more stable edge-on orientation for P3HT.



Simulations of Oxide/Polymer Hybrids, Fig. 5 Planar and curved P3HT/ZnO interfaces and comparison between their structure factors (Adapted with permission from Saba et al. J. Phys. Chem. C 2014 118 (9), 4687–4694. Copyright (2014) American Chemical Society)



Simulations of Oxide/Polymer Hybrids, Fig. 6 *Left*: edge-on P3HT chains on amorphous SiO_2 surface with a SAM interlayer of OTS molecules in between (Adapted with permission from Meredig et al. ACS Nano 2009 3 (10), 2881–2886. Copyright (2009) American Chemical Society) *Right*: example of ternary P3HT/4MP/TiO_2 system (Adapted with permission from Malloci et al. J. Phys. Chem. C 2013, 117 (27), 13894–13901. Copyright (2013) American Chemical Society)

Interfacial layers are also useful to improve the mobility of polymer molecules on the oxide surface. Such a mobility enhancement is expected to improve the quality and the amount of polymer/oxide interface area. For example, atomistic simulations at the P3HT/TiO_2 interface showed a highly increased polymer mobility in the presence of self-assembled interlayers of 4-mercaptopyridines (4MP) [28] (see Fig. 6 right). These molecules (similar to pyridines that are typically used as exchange ligand for titania nanoparticles) are able to efficiently bind to the titania forming a crystalline monolayer that exposes a flat surface of thiol groups to the polymer. This improves the ability of the polymer to infiltrate within mesoporous titania.

Interlayers are also important to enhance optical absorption, charge separation, and injection at the polymer/oxide interface. For these purposes, the molecular interlayer can be formed by optically active molecules such as dyes. There exist several examples in the field of solid-state

dye-sensitized solar cells in which a polymer (P3HT) is mixed with an oxide (titania or zinc oxide) sensitized by organic absorbing molecules. An example of atomistic simulation in this context studies the role of phthalocyanines in P3HT/ZnO systems [37]. A multi-scale combination of atomistic methods, including model potential molecular dynamics, has been applied to understand the thermodynamic stability of the system and its optoelectronics properties. The molecules are found to bind in parallel configuration to the substrate through Zn-O interactions, and they self-assemble into oriented stripes that cover the oxide surface. The formation of the ZnPc film has been shown to be beneficial for absorption, charge injection, and energy level alignment [37].

Model potential molecular dynamics methods are emerging as powerful tools for hybrid polymer/oxide systems. It is expected that its role in the design of novel hybrid materials will grow together with the increase of computer powers and the development of more advanced interatomic potentials. The combination of model potentials to first-principle methods, on the one hand, and to coarse-grained or continuum methods, on the other, will further extend the relevance of the atomistic simulations of polymer/oxide systems.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Hybrid Solar Cells
- ▶ Molecular Dynamics Simulations of Nanobiomaterials
- ▶ Nanomaterials for Excitonic Solar Cells

References

1. Sanchez, C., Beatriz, J., Belleville, P., Popall, M.: Applications of hybrid organic–inorganic nanocomposites. J. Mater. Chem. **15**, 3559–3592 (2005)
2. Jeon, I.-Y., Baek, J.-B.: Nanocomposites derived from polymers and inorganic nanoparticles. Materials **3**, 3654–3674 (2010)
3. Camargo, P.H.C., Satyanarayana, K.G., Wypych, F.: Nanocomposites: synthesis, structure, properties and new application opportunities. Mater. Res. **12**(1), 1–39 (2009)
4. Bouclé, J., Ackermann, J.: Solid-state dye-sensitized and bulk hetero-junction solar cells using TiO_2 and ZnO nanostructures: recent progress and new concepts at the borderline. Polym. Int. **61**(3), 355–373 (2012)
5. Ponder, J.W., Case, D.A.: Force fields for protein simulations. Adv. Protein. Chem. **66**, 27–85 (2003)
6. Brooks, B.R., Brooks III, C.L., Mackerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caflisch, A., Caves, L., Cui, Q., Dinner, A.R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R.W., Post, C.B., Pu, J.Z., Schaefer, M., Tidor, B., Venable, R.M., Woodcock, H.L., Wu, X., Yang, W., York, D.M., Karplus, M.: CHARMM: the biomolecular simulation program. J. Comput. Chem. **30**(10), 1545–1615 (2009)
7. Jorgensen, W.L., Tirado-Rives, J.: The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J. Am. Chem. Soc. **110**(6), 1657–1666 (1988)
8. Schmid, N., Eichenberger, A.P., Choutko, A., Riniker, S., Winger, M., Mark, A.E., van Gunsteren, W.F.: Definition and testing of the GROMOS force-field versions 54a7 and 54b7. Eur. Biophys. J. **40**, 843–856 (2011)
9. Marrink, S.J., Risselada, H.J., Yefimov, S., Tieleman, D.P., de Vries, A.H.: The MARTINI force field: coarse grained model for biomolecular simulations. J. Phys. Chem. B **111**(27), 7812–7824 (2007)
10. Cieplak, P., Dupradeau, F.-Y., Duan, Y., Wang, J.: Polarization effects in molecular mechanical force fields. J. Phys. Condens. Matter **21**(33), 333102 (2009)
11. Niketić, S.R., Rasmussen, K.: The Consistent Force Field: A Documentation. Lecture Notes in Chemistry, vol. 3. Springer, Berlin\Heidelberg (1977)
12. Rappe, A.K., Casewit, C.J., Colwell, K.S., Goddard III, W.A., Skiff, W.M.: UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J. Am. Chem. Soc. **114**(25), 10024–10035 (1992)
13. Allinger, N.L., Chen, K.-H., Lii, J.-H., Durkin, K.A.: Alcohols, ethers, carbohydrates, and related compounds. I. the MM4 force field for simple compounds. J. Comput. Chem. **24**(12), 1447–1472 (2003)
14. Sun, H., Ren, P., Fried, J.R.: The COMPASS force field: parameterization and validation for phosphazenes. Comput. Theor. Polym. Sci. **8**(1–2), 229–246 (1998)
15. Ponder, J.W., Wu, C., Ren, P., Pande, V.S., Chodera, J.D., Schnieders, M.J., Haque, I., Mobley, D.L.,

Lambrecht, D.S., DiStasio Jr., R.A., Head-Gordon, M., Clark, G.N.I., Johnson, M.E., Head-Gordon, T.: Current status of the AMOEBA polarizable force field. J. Phys. Chem. B **114**(8), 2549–2564 (2010)
16. Matsui, M., Akaogi, M.: Molecular dynamics simulation of the structural and physical properties of the four polymorphs of TiO_2. Mol. Simul. **6**(4–6), 239–244 (1991)
17. Kulkarni, A.J., Zhou, M., Ke, F.J.: Orientation and size dependence of the elastic properties of zinc oxide nanobelts. Nanotechnology **16**(12), 2749–2756 (2005)
18. Raymand, D., van Duin, A.C.T., Baudin, M., Hermansson, K.: A reactive force field (ReaxFF) for zinc oxide. Surf. Sci. **602**(5), 1020–1031 (2008)
19. Raju, M., Kim, S.-Y., van Duin, A.C.T., Fichthorn, K. A.: ReaxFF reactive force field study of the dissociation of water on titania surfaces. J. Phys. Chem. A C **117**(20), 10558–10572 (2013)
20. Woodcock, L.V., Angell, C.A., Cheeseman, P.: Molecular dynamics studies of the vitreous state: simple ionic systems and silica. J. Chem. Phys. **65**(4), 1565–1577 (1976)
21. Feuston, B.P., Garofalini, S.H.: Empirical three-body potential for vitreous silica. J. Chem. Phys. **89**(9), 5818–5824 (1988)
22. Vessal, B., Amini, M., Fincham, D., Catlow, C.R.A.: Water-like melting behavior of SiO_2 investigated by the molecular dynamics simulation technique. Philos. Mag. B **60**(6), 753–775 (1989)
23. van Beest, B.W.H., Kramer, G.J., van Santen, R.A.: Force fields for silicas and aluminophosphates based on ab initio calculations. Phys. Rev. Lett. **64**(16), 1955–1958 (1990)
24. Huff, N.T., Demiralp, E., Çagin, T., Goddard III, W.A.: Factors affecting molecular dynamics simulated vitreous silica structures. J. Non Cryst. Solids **253**(1–3), 133–142 (1999)
25. Takada, A., Richet, P., Catlow, C.R.A., Price, G.D.: Molecular dynamics simulations of vitreous silica structures. J. Non Cryst. Solids **345–346**, 224–229 (2004)
26. van Duin, A.C.T., Strachan, A., Stewman, S., Zhang, Q., Xu, X., Goddard III, W.A.: ReaxFF $_{SiO}$ reactive force field for silicon and silicon oxide systems. J. Phys. Chem. A **107**(19), 3803–3811 (2003)
27. Saba, M.I., Calzia, V., Melis, C., Colombo, L., Mattoni, A.: Atomistic investigation of the solid-liquid interface between the crystalline zinc oxide surface and the liquid tetrahydrofuran solvent. J. Phys. Chem. C **116**(23), 12644–12648 (2012)
28. Malloci, G., Binda, M., Petrozza, A., Mattoni, A.: Role of molecular thermodynamical processes at functionalized polymer/metaloxide interfaces for photovoltaics. J. Phys. Chem. C **117**(27), 13894–13901 (2013)
29. Melis, C., Raiteri, P., Colombo, L., Mattoni, A.: Self-assembling of Zinc Phthalocyanines on ZnO $(10\bar{1}0)$ surface through multiple time scales. ACS Nano **5**(12), 9639–9647 (2011)
30. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Clarendon, Oxford (1989)
31. Kang, Y., Li, X., Tu, Y., Wang, Q., Ågren, H.: On the mechanism of protein adsorption onto hydroxylated and nonhydroxylated TiO_2 surfaces. J. Phys. Chem. C **114**(34), 14496–14502 (2010)
32. Ham, H.O., Park, S.H., Kurutz, J.W., Szleifer, I.G., Messersmith, P.B.: Antifouling glycocalyx-mimetic peptoids. J. Am. Chem. Soc. **135**(35), 13015–13022 (2013)
33. Kenausis, G.L., Vörös, J., Elbert, D.L., Huang, N., Hofer, R., Ruiz-Taylor, L., Textor, M., Hubbell, J.A., Spencer, N.D.: Poly(l-lysine)-g-poly(ethylene glycol) layers on metal oxide surfaces: attachment mechanism and effects of polymer architecture on resistance to protein adsorption. J. Phys. Chem. B **104**(14), 3298–3309 (2000)
34. Saba, M.I., Mattoni, A.: Effect of thermodynamics and curvature on the crystallinity of P3HT thin films on ZnO: insights from atomistic simulations. J. Phys. Chem. C **118**(9), 4687–4694 (2014)
35. Caddeo, C., Dessí, R., Melis, C., Colombo, L., Mattoni, A.: Poly(3-hexylthiophene) adhesion on zinc oxide nanoneedles. J. Phys. Chem. C **115**(34), 16833–16837 (2011)
36. Meredig, B., Salleo, A., Gee, R.: Ordering of poly (3-hexylthiophene) nanocrystallites on the basis of substrate surface energy. ACS Nano **3**(10), 2881–2886 (2009)
37. Mattioli, G., Dkhil, S.B., Saba, M.I., Malloci, G., Melis, C., Alippi, P., Filippone, F., Giannozzi, P., Thakur, A.K., Gaceur, M., Margeat, O., Diallo, A.K., Videlot-Ackermann, C., Ackermann, J., Bonapasta, A. A., Mattoni, A.: Interfacial engineering of P3HT/ZnO hybrid solar cells using phthalocyanines: a joint theoretical and experimental investigation. Adv. Energy Mater., **4**,1301694 (2014)

Simulations of Solid Interfaces

► Surface Modeling of Ceramic Biomaterials

Single Cell Analysis

► Electrical Impedance Tomography for Single-Cell Imaging

Single Cell Impedance Spectroscopy

► Electrical Impedance Cytometry

Single-Cell Electrical Impedance Spectroscopy

▶ Single-Cell Impedance Spectroscopy

Single-Cell Impedance Spectroscopy

Jian Chen[1], Nika Shakiba[2], Qingyuan Tan[2] and Yu Sun[3]
[1]State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing, People's Republic of China
[2]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada
[3]Department of Mechanical and Industrial Engineering and Institute of Biomaterials and Biomedical Engineering and Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada

Synonyms

Impedance measurement of single cells; Impedance spectroscopy for single-cell analysis; Single-cell electrical impedance spectroscopy

Definition

Single-cell impedance spectroscopy is a technique that operates by applying a frequency-dependent excitation signal on a single cell positioned in between two measurement microelectrodes. The current response is measured as a function of frequency. By interpreting the impedance profile, dielectric properties of a single cell such as cell membrane capacitance and cytoplasmic resistance are obtained.

Overview

Historical Development

In conventional cell electrical impedance spectroscopy, the impedance of multiple cells as a whole is measured using an AC excitation signal. The cell suspension is held in a container with two electrodes. Current is measured as a function of frequency to determine the electrical properties of the cells in the suspension.

Recent advances in microfabrication and lab-on-a-chip technologies enable the development of electrical impedance spectroscopic devices to measure impedance profiles of cells at the single-cell level, providing useful biophysical characteristics of single cells and promising potentially noninvasive, label-free approaches for analyzing cells.

Working Principle

Single-cell impedance spectroscopy measures impedance profiles of a cell positioned in between two microelectrodes [1]. A low-magnitude AC voltage, $\tilde{U}(j\omega)$, over a range of frequencies is used as the excitation signal. The current response, $\tilde{I}(j\omega)$, is measured. The impedance $\tilde{Z}(j\omega)$ is

$$\tilde{Z}(j\omega) = \frac{\tilde{U}(j\omega)}{\tilde{I}(j\omega)} = \tilde{Z}_{RE} + j\tilde{Z}_{IM} \tag{1}$$

where $\tilde{Z}_{RE}$ and $\tilde{Z}_{IM}$ are the real and imaginary parts of impedance. The real part is termed resistance while the imaginary part is termed reactance. The magnitude and phase angle are

$$\left|\tilde{Z}\right| = \sqrt{\tilde{Z}_{RE}^2 + \tilde{Z}_{IM}^2} \tag{2}$$

and

$$\angle\tilde{Z} = \arctan\left(\frac{\tilde{Z}_{IM}}{\tilde{Z}_{RE}}\right) \tag{3}$$

Advantages and Applications

Compared to impedance measurement on a cell suspension, single-cell impedance spectroscopy interrogates the property of a single cell (vs. a cell population). Theoretical analysis can also become simpler because it is not necessary to take into account electrical interactions among cells [2].

The technique of single-cell impedance spectroscopy has been used as a noninvasive method to quantify the physiological state of single cells. As a biophysical marker, single-cell impedance profiles have also been utilized to distinguish normal cells from abnormal cells (e.g., human cancer cells with different metastatic potential and malaria-infected red blood cells) [3, 4].

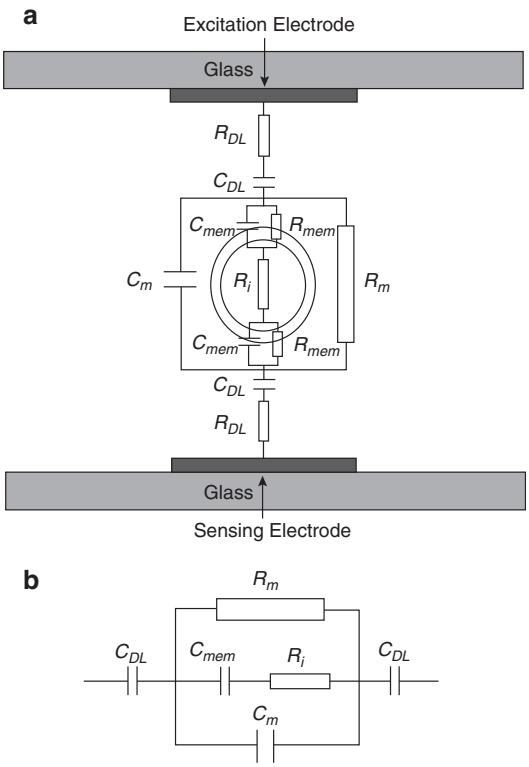
Methodology

Equivalent Circuit Model

Figure 1a [1] shows an equivalent circuit for interpreting single-cell impedance measurement data [1]. In this circuit model, the cell is represented by a capacitor in series with a resistor, with elements connected in parallel with the capacitor and resistor (i.e., the cell) representing the suspending medium. At low frequencies, the thin cell membrane gives rise to a large capacitance. As the frequency increases, the reactive component of this element tends to zero out, and the cell internal properties are represented by the resistor. In this circuit model, the membrane is assumed to have a high resistance and the cytoplasmic capacitance is ignored. All the circuit elements are functions of the volume fraction or cell size.

The electrical double layer has an influence on measurements at low frequencies (below 1 MHz for high-conductivity solutions). This is generally modeled as a constant phase angle, represented by a resistor (R_{DL}) and capacitor (C_{DL}) in series. As shown in Fig. 1a, the double layer is in series with the network model. In the simplified model shown in Fig. 1b, the double layer is assumed to be purely capacitive (C_{DL}), with a value given by the product of the inverse Debye length and the permittivity of the medium.

The simplified circuit in Fig. 1b [1] shows that at very low frequencies current flow is blocked by the double layer capacitor, and only the impedance of the double layer is measured. As the measurement frequency increases, this capacitor is gradually short-circuited and the excitation voltage charges the cell in suspension. It takes a finite amount of time



Single-Cell Impedance Spectroscopy, Fig. 1 (**a**) The equivalent circuit model for single-cell impedance spectroscopy. R_{DL} and C_{DL} represent the resistance and capacitance of the electrical double layer, R_m and C_m the resistance and capacitance of the medium, R_{mem} and C_{mem} the resistance and capacitance of the cell membrane and R_i the resistance of the cytoplasm. (**b**) The simplified equivalent circuit model neglecting the electrical double layer resistance and the membrane resistance (Reproduced with permission from Ref. [1])

to charge the membrane through the extracellular and intracellular fluid, resulting in two dielectric dispersions in the frequency range of 1–100 MHz. The lower frequency dispersion is governed by the polarization of the cell membrane. Measurement of this parameter provides information about the dielectric properties of the membrane. The higher-frequency dispersion is governed by the polarization of the cytoplasm and the suspending medium as the membrane is short-circuited at these frequencies. This second dispersion is generally small and difficult to measure using impedance spectroscopy techniques.

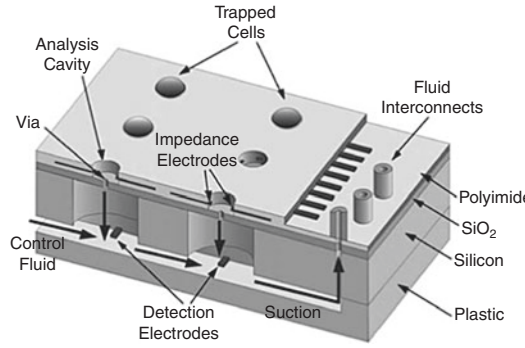
With the effect of the double layer taken into account, the total impedance of the circuit is

$$\tilde{z}(j\omega) = \frac{1}{j\omega C_{DL}} + \frac{R_m(1 + j\omega R_i C_{\text{mem}})}{j\omega R_m C_{\text{mem}} + (1 + j\omega R_i C_{\text{mem}})(1 + j\omega R_m C_m)} \quad (4)$$

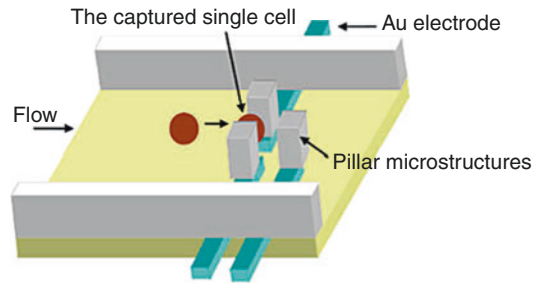
Single-Cell Impedance Spectroscopy on Stationary or Moving Cells

Single-cell impedance spectroscopy in microfluidic devices can be divided into two main categories: measurements on either stationary or moving cells. For measuring a stationary cell, a cell can be positioned/trapped in between microelectrodes using several approaches. Cells can also be controlled to pass the measurement electrodes at the speed of microfluidic flow.

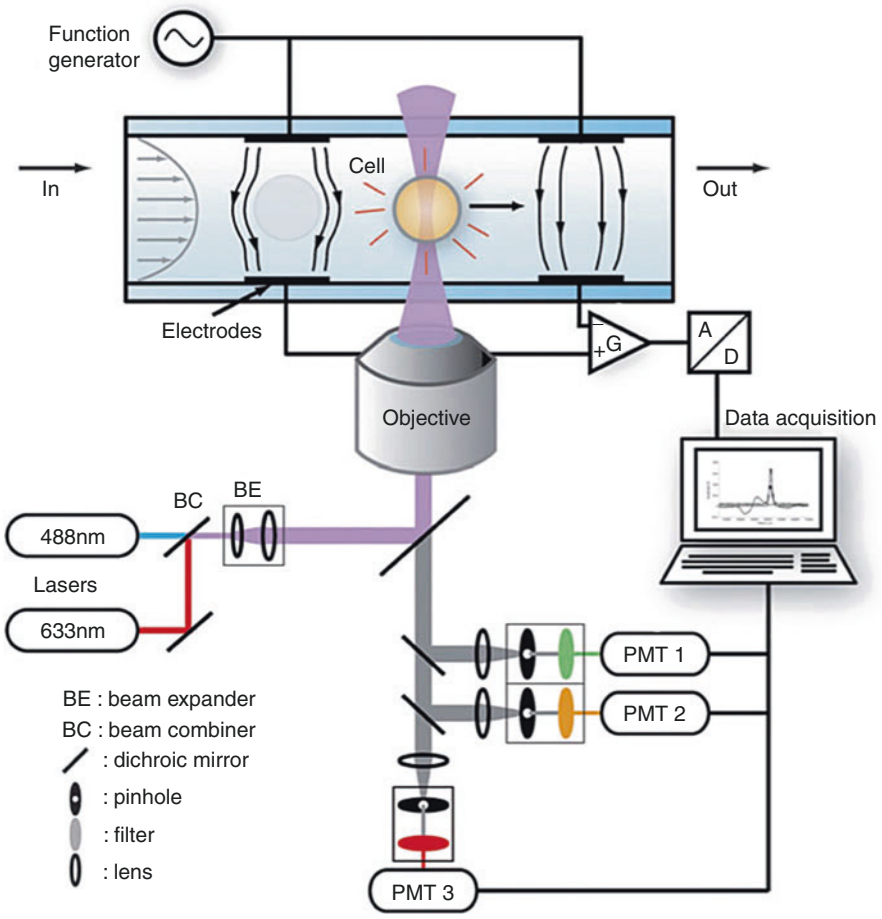
1. Cell aspiration [5]
   (a) Working mechanism: In the aspiration approach, a negative pressure is applied to trap a single cell in a channel opening. Measurement electrodes can be built into such channel openings. A tapped cell is forced by the applied negative pressure to contact the electrodes (Fig. 2) [5]. The impedance profile of the cell is then obtained.
   (b) Advantages: The main advantage of the aspiration approach for capturing a cell is the seal formed between the measurement electrodes and the cell. The suction force is sufficient to hold the cell in place throughout the measurement process, especially when the cell-capturing channel has a more or less circular cross section.
   (c) Potential limitations: It has been suggested that mechanical deformations of a cell due to the suction force may lead to changes in the electrical properties of the cell, thus, affecting the impedance profile measured.
2. Hydrodynamic trapping [6]
   (a) Working mechanism: In hydrodynamic trapping, cells are trapped by microfabricated weirs, used as filters on top of electrodes. By adjusting flow rates, cells can be captured on measurement electrodes for impedance data recording (Fig. 3) [6].
   (b) Advantages: The main advantage of this approach is the feasibility to realize large-array single-cell trapping and potentially high-throughput impedance measurement.
   (c) Potential limitations: Contact between the trapped cell and the measurement



Single-Cell Impedance Spectroscopy, Fig. 2 Schematic view of an aspiration-based single-cell impedance spectroscopy system. The system is composed of an array of analysis cavities, each capable of analyzing a single cell at a time, each containing a mechanical fluid via with flow control to hold the cell in position during the impedance analysis operation, and multiple electrodes surrounding the circumference of the cell to measure the electrical characteristics of the captured cell (Reproduced with permission from Ref. [5])



Single-Cell Impedance Spectroscopy, Fig. 3 Schematic view of a microfluidic device for single-cell impedance spectroscopy using the concept of hydrodynamic trapping for single-cell immobilization (Reproduced with permission from Ref. [6])



Single-Cell Impedance Spectroscopy, Fig. 4 Schematic view of differential impedance micro flow cytometry. Each cell's impedance signal is recorded by a differential pair of microelectrodes using media without a single cell passing as a reference. In this setup, the effect of the electrical double layer on impedance profiles is canceled (Reproduced with permission from Ref. [7])

microelectrodes underneath can be problematic since there is no force to hold the cell in close contact with the microelectrodes.

3. Flow cytometry [7]
   (a) Working mechanism: Flow cytometry is a technique that allows for the analysis of cells in suspension with single-cell precision. A laminar flow carries suspended cells through the measurement location. Each cell's impedance signal is recorded by a differential pair of microelectrodes, using the surrounding media without a cell as reference. Microfabricated devices and electronic circuits allow simultaneous impedance measurements at multiple frequencies (Fig. 4) [7].
   (b) Advantages: Flow cytometry is advantageous in that it allows for rapid analysis of cells in suspension with a high throughput. Furthermore, flow cytometry can also allow for the sorting of cells into subpopulations based on their impedance measurement profiles.
   (c) Potential limitations: There is no contact between the cell and measurement electrodes and thus, the impedimentary contribution of the current leakage in the medium around the cell is significant.

Key Research Findings

Effect of Electrical Double Layer on Impedance Profile

When a polarizable electrode (i.e., an electrode operated at a regime where no charge transfer reaction occurs at the surface) comes into contact with an electrolyte, charges from the electrolyte opposite in sign to the charges present on the surface of the electrode move the electrode/electrolyte interface and provide a localized condition of electroneutrality as well as a layer of charge, termed the electrical double layer.

The electrical double layer has an influence on cell impedance measurements at low frequencies, which is generally represented as a capacitor. The most effective way to minimize the electrical double layer is to maximize the electrolyte-electrode interface area. This interface area enhancement can be achieved either by mechanically roughing the electrode surface to an electrode-electrolyte interface area that is effectively larger than the actual electrode surface, or to use chemical treatments that lead to a high electrode-electrolyte interface area.

Comparison of Impedance Measurement Mechanisms

The essential feature of measurements on a stationary cell trapped/positioned between electrodes is a tight seal between the cell and electrodes. In this configuration, the impedimentary contribution of the cell shunt (i.e., current leakage in the medium around the cell) is insignificant.

The advantage of impedance cytometry is differential impedance measurement that minimizes the effect of the electrical double layer. In an alternating fashion, each measurement electrode-ground electrode pair without a cell serves as a reference to the other pair, over which a cell passes.

Future Work

A single-cell impedance spectroscopy measurement system ideally should have a high testing throughput and the capability of sorting single cells based on impedance measurement results. The potential combination of single-cell impedance spectroscopy with other detection/measurement methods, such as fluorescent detection and the use of biochemical markers, may prove powerful for a range of applications, such as disease diagnostics and rare cell isolation.

References

1. Morgan, H., Sun, T., Holmes, D., Gawad, S., Green, N. G.: Single cell dielectric spectroscopy. J. Phys. D Appl. Phys. **40**, 61–70 (2007)
2. Valero, A., Braschler, T., Renaud, P.: A unified approach to dielectric single cell analysis: impedance and dielectrophoretic force spectroscopy. Lab Chip **10**, 2216–2225 (2010)
3. Sun, T., Morgan, H.: Single-cell microfluidic impedance cytometry: a review. Microfluid. Nanofluid. **8**, 423–443 (2010)
4. Cheung, K.C., Di Berardino, M., Schade-Kampmann, G., Hebeisen, M., Pierzchalski, A., Bocsi, J., Mittag, A., Tarnok, A.: Microfluidic impedance-based flow cytometry. Cytometry A **77A**, 648–666 (2010)
5. Han, K.H., Han, A., Frazer, A.B..: Microsystems for isolation and electrophysiological analysis of breast cancer cells from blood. Biosens. Bioelectron. **21**, 1907–1914 (2006)
6. Jang, L.S., Wang, M.H.: Microfluidic device for cell capture and impedance measurement. Biomed. Microdevices **9**, 737–743 (2007)
7. Holmes, D., Pettigrew, D., Reccius, C.H., Gwyer, J.D., Berkel, C.V., Holloway, J., Daviesb, D.E., Morgan, H.: Leukocyte analysis and differentiation using high speed microfluidic single cell impedance cytometry. Lab Chip **9**, 2881–2889 (2009)

Single-Walled Carbon Nanotubes (SWCNTs)

▶ Chemical Vapor Deposition (CVD)

Sintered

▶ Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm

siRNA Delivery

▶ RNAi in Biomedicine and Drug Delivery

Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures

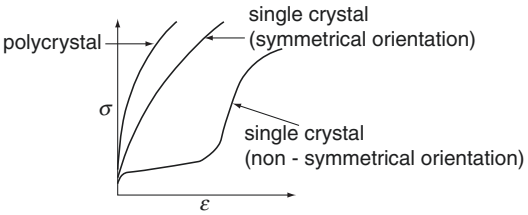
Julia R. Greer
Division of Engineering and Applied Sciences, California Institute of Technology, Pasadena, CA, USA

Definition

"Size-dependent plasticity" here refers to the strength of metallic samples being a strong function of their size when their dimensions are reduced to the micron and below scales. The notion of reduced sample size applies to all three dimensions, i.e., stand-alone, or one-dimensional (1D) nano and microstructures rather than thin films (2D), where only their thicknesses are reduced to nano- and micro-scales, or to the small-scale deformation volumes within a bulk matrix (3D) as would be the case during nanoindentation experiments, for example.

Introduction and Overview of Stress Vs. Strain for Bulk Metals

Pure bulk single crystalline metals like Cu, Ni, Mo, Ti, etc. generally exhibit a convex, continuous stress–strain relationship upon uniaxial deformation – tension or compression – as can be found in any classic "Mechanical Behavior of Materials" book and as is also schematically shown in Fig. 1. For the low-symmetry orientations, i.e., where only a single slip system is activated, the deformation immediately following yield is called the "easy glide" region, which is characterized by a low hardening rate. Such low hardening rate stems from the dislocations traveling large distances unimpeded in their glide



Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures, Fig. 1 Schematic of typical stress–strain curves for bulk metals: comparing polycrystalline and single crystalline (in high- and low-symmetry orientations) microstructures of the same metal

planes as only one set of crystallographic slip planes is active, and the dislocations are not forced to overcome closely spaced obstacles such as impurities or other dislocations in the course of their motion. A simple way to think of it is the following: Each dislocation travels a distance L before encountering an obstacle, which pins it. The shear strain associated with this motion is: $d\gamma = bLd\rho$, where b is the Burgers vector, L is the dislocation mean free path, and ρ is the mobile dislocation density [1]. The distance traveled by each dislocation then scales with $1/\sqrt{\rho}$, i.e., $L = \beta/\sqrt{\rho}$, where β may be on the order of 1,000 since the dislocations traveling in parallel crystallographic planes have very limited interactions. The hardening corresponding to this "easy glide" region is a result of dislocation storage through the well-known Taylor relation: $\tau = \alpha\mu b\sqrt{\rho}$, where $\alpha \sim 0.2$, which results in a very low hardening rate $d\tau/d\gamma_{low-symmetry} = 10^{-4}\mu$. In high-symmetry orientations, multiple symmetric slip systems are activated, and dislocations come into close encounters with one another as they are traveling toward one another in the symmetric slip planes, resulting in a pronounced dislocation density increase, which in turn, drives the high rate of hardening: $d\tau/d\gamma_{high-symmetry} = 0.01\mu$. This hardening in bulk single crystals stems from dislocation multiplication processes arising from the well-established dislocation interactions, which are followed by the dislocation networks formation processes. Therefore, in bulk single crystals – regardless of the sample size – the

dislocations multiply in the course of deformation, thereby creating dense networks and dislocation sub-structures, which require the application of higher stresses to move the additional gliding dislocations through these obstacles, thereby carrying out plastic deformation. This holds true for crystals of any size with greater-than-several-microns dimensions. As a result, bulk samples of the same single crystalline material also exhibit identical yield strengths, flow stresses, and hardening rates.

Emergence of Size Effects in Single Crystals at the Nanoscale

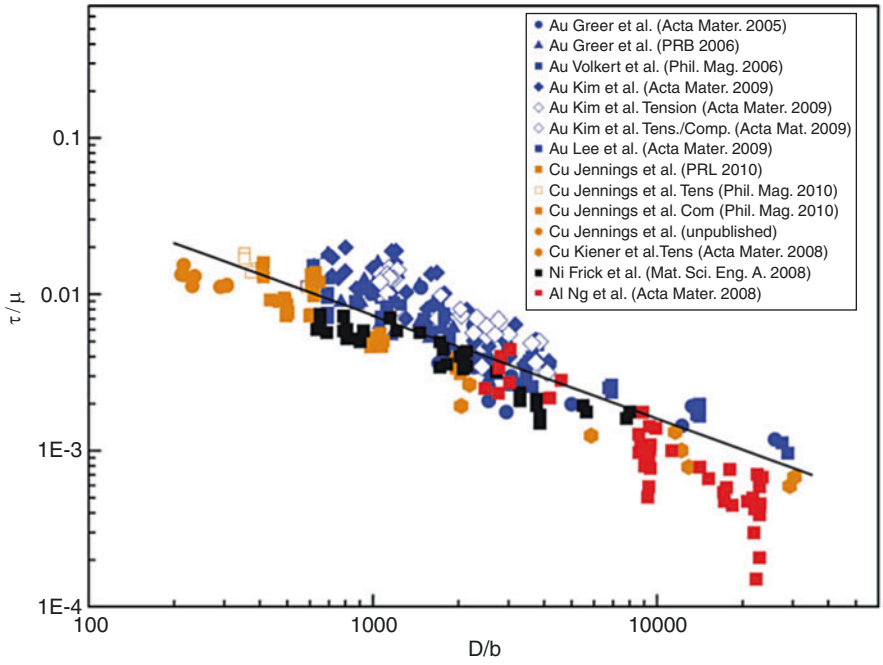
In a striking deviation from this classical depiction, in the last 5–6 years, it was ubiquitously demonstrated that at the micron- and sub-micron scales, the sample size dramatically affects crystalline strength – even in the absence of any constraining effects, strain gradients, and grain boundaries – as revealed by room-temperature uniaxial compression experiments on a wide range of single crystalline metallic nano-pillars (for reviews, see [2–4]). In these studies, cylindrical nano-pillars were fabricated mainly by the use of the focused ion beam (FIB), as well as with some FIB-less methodologies, and remarkably, the results of these reports for all metallic single crystals show power law dependence between the flow stress and sample size, regardless of fabrication technique, of the form $\sigma = C \times D^{-k}$ with the exponent k strongly dependent on the initial crystal structure (i.e., face-centered cubic (fcc), body-centered cubic (bcc), hexagonal close-packed (hcp), etc.), experimental aspects (i.e., sample aspect ratio, lateral stiffness of the instrument, etc.), and dislocation density. A particularly auspicious example is the unique scaling of strength with pillar diameter with the exponent of ~ -0.6 exhibited by nearly all non-pristine (i.e., containing initial dislocations) fcc nano-pillars subjected to uniaxial compression or tension, as illustrated in Fig. 2.

The notion of "size effects" applied specifically to the strength of single crystalline metals dates back to the works of Brenner [5, 6], who demonstrated that dislocation-free Cu and Ag whiskers attained very high strengths upon uniaxial tension. Although significant efforts have since been dedicated to studying single crystalline plasticity in confined dimensions, until ~5 years ago, most of these research thrusts were focused on thin films, whose yield strengths increased with decreasing film thickness (for example, [7]). The renaissance of small-scale mechanical testing on free-standing vertical pillars is largely due to the original work of Uchic et al. who reported higher compressive strengths attained by focused ion beam (FIB)-machined cylindrical Ni micro-pillars [8]. Greer and Nix then extended this robust and elegant methodology into the nanoscale regime, where <001> −oriented Au nano-pillars with diameters below 1 μm were reported to be 50 times higher than bulk [9], and today numerous groups are pursuing this type of uniaxial nano-mechanical testing of materials ranging from single crystalline metals to lithiated battery anodic materials, ceramics, irradiated materials, shape memory alloys, nano-twinned and nanocrystalline metals, metallic glasses, superalloys, and nanolaminates. The results of many of these studies are overviewed in detail in three existing reviews: [2–4]. Figure 3 shows representative images of before- and after-testing single crystalline Nb nano-pillars, as well as the representative stress–strain curves with clear discrete characteristics and size effect [10]. Further adapting this methodology, the exploration of size effects in plasticity for a large variety of materials ensued, albeit mainly focusing on fcc structures. More recently, investigating plasticity in small volumes has been further advanced through uniaxial tensile experiments, usually conducted inside of in-situ mechanical deformation instruments custom built by some research groups [9–12].

Key Research Findings

Experimental Findings

To date, uniaxial compression and tension tests have been performed on Ni and Ni-based superalloys, Au, Cu, and Al (as-fabricated and intentionally passivated) [2–4]. Beyond single

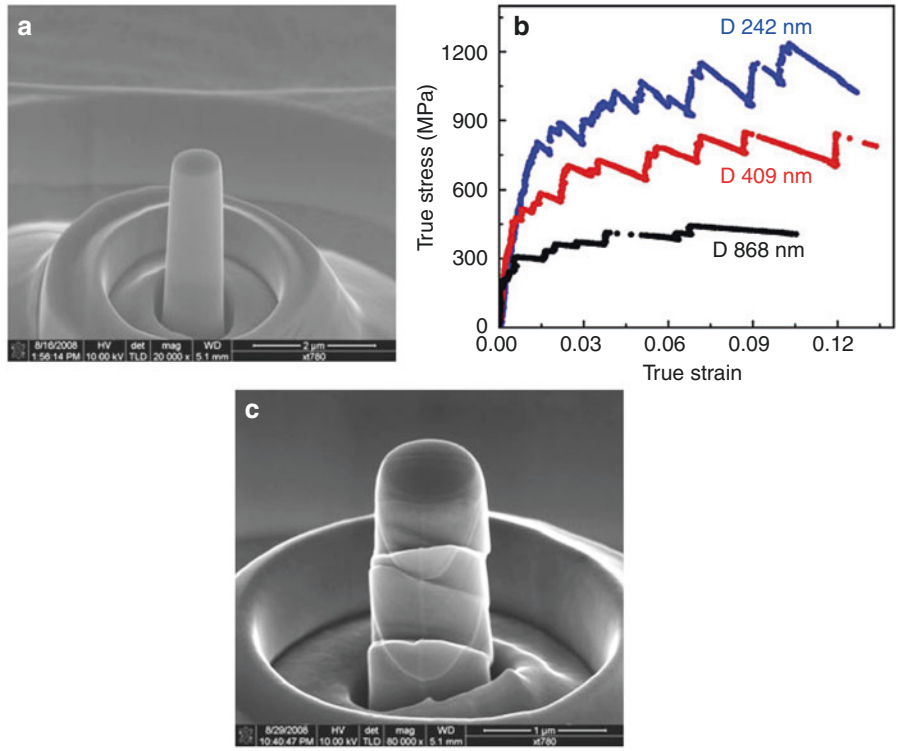


Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures, Fig. 2 Resolved shear strength normalized by the shear modulus as a function of diameter (normalized by the Burgers vector) of experimentally determined compressive strengths of most face-centered cubic metallic nano-pillars published to date (Reprinted with permission from Ref. [2])

crystalline fcc metals, mechanical behavior of other single crystals has been published to date: bcc metals (W, Nb, Ta, Mo, and V), hcp metals (Mg, Ti-Al alloy, and Ti), tetragonal metals (In) [2–4]. The reports on nearly all single crystalline metals with non-zero initial dislocation densities (i.e., not whiskers or nano-wires) unanimously demonstrate that their strengths significantly rise with reduced size, with ~100 nm-diameter samples sometimes attaining a flow stress ~10× higher than bulk. Intriguingly, unlike in bulk, where the dislocation multiplication processes result in Taylor hardening (introduced in section "Introduction and Overview of Stress Vs. Strain for Bulk Metals)", the flow stresses in small structures do not appear to scale with the dislocation density. Rather, the global dislocation density appears to decrease upon mechanical loading, while the applied stress required to deform the structure increases (see, for example [11]). Consistent with this "upside down" behavior in the nano- and micron-sized crystals, whereby samples containing fewer mobile dislocations are stronger than their bulk counterparts, it has been shown that introduction of additional dislocations into the structure (for example, via pre-straining) actually *weakens* these crystals [12, 13]. These findings are antipodal to classical plasticity, where dislocation interactions lead to multiplication, causing higher dislocation densities and requiring higher applied stresses for deformation to continue, as described in the section "Introduction and Overview of Stress Vs. Strain for Bulk Metals". The initial dislocation density indeed plays a critical role in the onset of the size effect, as has now been shown by several research groups [2–4].

Computational Findings

Several models attempting to explain the origins of size-dependent flow stress in the absence of strong strain gradients, as well as the stochastic nature of deformation, have been put forth. There are generally three classes of these models: (1) phenomenological theories, which attempt to

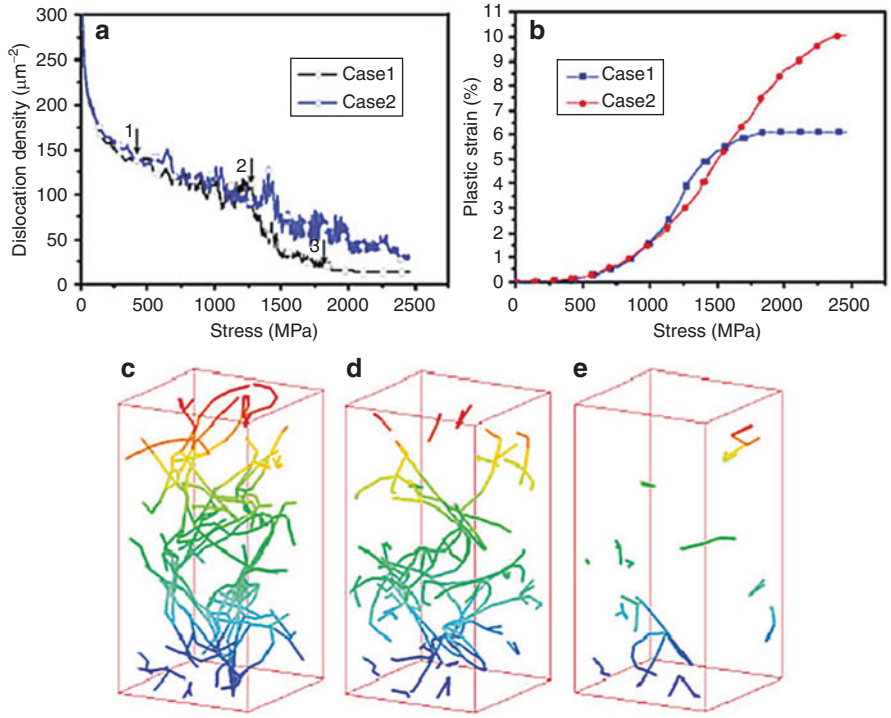


Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures, Fig. 3 SEM images of (**a**) single crystalline Nb nano-pillar as fabricated by the focused ion beam, (**c**) same pillar after compression showing pronounced crystallographic slip lines. (**b**) Stress versus strain curves clearly revealing the size effect (smaller is stronger) and the stochastic signature of small-scale deformation (Reprinted with permission from Ref. [10])

describe the physical processes occurring in a nano-single-crystal upon deformation, (2) discrete dislocation dynamics (two- and three-dimensional) simulations (DD), which rely on the entered initial dislocation density and distribution, as well as on the dislocation mobility laws, and (3) molecular dynamics (MD) simulations, which are generally limited to small (50 nm and below) scales and unrealistically high strain rates ($\sim 10^{-8}$ s^{-1}). An example of (1) is the "hardening by dislocation starvation" theory which hypothesizes that the mobile dislocations inside a small nano-pillar have a greater probability of annihilating at a free surface than of interacting with one another, thereby shifting plasticity into nucleation-controlled regime [14–16]. Other models include source exhaustion hardening, source truncation, and weakest link theory, all of whose general premise involves representing dislocation source operations in a discrete fashion, and then evaluating the effect of sample size on the source lengths, and therefore on the strengths of their operation [17]. These models capture the ubiquitously observed stochastic signature of the experiments results and show either marginal dislocation storage or no storage at all. An example of an evolved microstructure as revealed by dislocation dynamics simulations is shown in Fig. 4 and clearly shows that upon deformation, the dislocation density decreases, rendering multiplication processes unlikely and shifting plasticity into dislocation nucleation-controlled regime.

Discussion: What Causes the Size effect?

Despite the definition of "ultimate tensile strength" according to Wikipedia stating that



Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures, Fig. 4 Variations in (**a**) dislocation density and (**b**) plastic strain with increasing stress. A dislocation avalanche occurs during the second stage. (**c**–**e**) are the microstructures corresponding to points *1*, *2*, and *3* in (**a**), respectively (Reprinted with permission from Ref. [20])

UTS is an intensive property and "therefore its value does not depend on the size of the test specimen," there is now a large body of work that convincingly demonstrates that this does not hold true in the micron and sub-micron size regime. The ubiquitously reported size-dependent strengths in small-scale single crystalline micro- and nano-pillars show the "smaller is stronger" phenomenon as revealed by the existence of a power law dependence between the attained flow stress (σ_{flow}) and pillar diameter (D): $\sigma_{flow} = C \times D^{-n}$, where n varies with the specific crystal structure, initial dislocation density, and deformation type. Intriguingly, it appears that the power law slope for all fcc metals is unified, on the order of −0.6, as shown in Fig. 2. This is not the case for body-centered cubic metals, whose power law slopes vary within this crystal structure family, possibly due to the unique potential energy landscape of each bcc metal that a gliding screw dislocation has to overcome. The power law slopes of bcc metals appear to correlate with the intrinsic lattice resistance of each metal, a.k.a. the Peierls barrier, and can be separated into groups according to the low, mid- and high Peierls barrier: low-barrier group containing Nb (−0.93, −0.48, −1.07) and V (−0.79); mid-barrier group containing Mo (−0.44, −0.38) and Ta (−0.43, −0.41); and high-barrier group containing W (−0.21, −0.44) [2–4]. Beyond these two cubic crystalline structures, other types of crystals were also found to exhibit size-dependent strengths, with smaller generally being stronger. For example, hexagonal close-packed (hcp) metals, Mg and Ti, oriented for slip (as opposed to twinning) were also characterized by power law size effects: [11–21] Ti oriented for prismatic slip has the critical resolved shear stress (CRSS) scale inversely with the diameter, resulting in the slope of −1, while the [3–964] orientation of Mg (basal slip) has the slope of −0.64 [2].

There is currently no physics-based theory that captures these size effects as a function of metal, size, and initial microstructure. Several semi-empirical theories have been proposed, and nearly all identify the initial dislocation density as a key factor in defining the natural microstructural length scale, which in turn, drives the size effect. In an attempt to explain the emergence of higher strengths in small-scale crystals, multiple deformation mechanisms have been proposed, with two distinct ones: (1) single-arm source (SAS) theory first developed by Parthasarathy and Rao et al. [17] applicable to micron-sized pillars, and (2) hardening by dislocation starvation proposed by Greer and Nix [9, 14] for the much smaller, nano-sized pillars. In the SAS theory, the creation of dislocations occurs by the operation of truncated Frank-Read spiral sources, a.k.a. single-arm sources (SAS), whose strength, τ_s, is inversely proportional to their average length, λ, through $\tau_s = k_s \mu \frac{\ln(\bar{\lambda}/b)}{\bar{\lambda}/b}$, where, k_s is a source-hardening constant [17]. The smaller micron-sized pillars are not capable of accommodating large single-arm sources, and therefore require the application of higher stresses to activate the stronger, "shorter" sources, giving rise to higher strengths in smaller pillars.

In the dislocation starvation theory, applicable to the much smaller pillars, with diameters deeply in the sub-micron regime, new dislocations are created via surface nucleation rather than through a multiplicative process of SAS operation as is the case in the micron-sized samples. The surface nucleation gets activated only after a significant fraction of the pre-existing gliding dislocations within the pillar has exited through the free pillar surface, i.e., "starving" the crystal of the necessary mobile dislocations to carry plastic strain. A necessary condition for dislocation starvation is that the largest distance that any gliding dislocation has to travel is smaller than the so-called breeding length, or the distance before it replicates itself, as defined by Gilman [18]. This is likely to be the case in the samples with diameters deeply in the submicrometer range since Ag, for example, has the breeding distance on the order of 0.7 μm. The nucleation stress for a surface source may be represented as $\sigma = \frac{Q^*}{\hat{Q}} - \frac{k_B T}{\hat{Q}} \ln \frac{k_B T N \nu_0}{E \dot{\varepsilon} \hat{\Omega}}$, where the first term corresponds to the athermal contribution and the last term represents thermally activated nature of such events [19]. The activation of surface sources has a significant thermal activation component, as is evident from the $T \ln T$ and strain rate ($\dot{\varepsilon}$) dependence of the nucleation stress. Recent experimental studies performed on very small (below 500 nm) Cu nano-pillars are consistent with this surface nucleation model and lay out the single-arm source versus surface source operation regimes as a function of both the pillar size and strain rate [2]. While these theories may appear to be competing, it is likely that they both take place at different pillar sizes: with SAS strengthening occurring in the micron-sized pillars (perhaps, down to ~500 nm) and with dislocation starvation followed by surface nucleation prevailing at the smaller sizes, deep in the sub-micron regime.

Summary

- Recent experimental and computational results convincingly demonstrate that the strength of nano- and micron-sized single crystalline metals is indeed size dependent and appears to be well represented by a power law of the form $\sigma = C \times D^{-k}$ with the exponent strongly dependent on the initial microstructure (i.e., dislocation density, lattice resistance, and the presence of impurities).
- With the advent of highly capable nano-fabrication and analysis instruments, as well as of the sophisticated computational tools, the society is ever closer to developing a more complete understanding of size-dependent mechanical behavior of small-scale structures. To date, it is generally agreed that non-pristine (i.e., containing dislocations) micron-sized fcc structures attain higher strengths at smaller sizes through dislocation multiplication processes produced by the operation of single-arm dislocation sources, whose lengths scale with the sample size. The operation of

these truncated sources, which become exhausted in the course of deformation, drives the corresponding strength increase in these small-scale structures. On the contrary, when the sample dimensions are further reduced to the nano-sized regime, plasticity likely occurs via dislocation nucleation from surface sources (rather than through a multiplicative process), and the higher strengths arise due to the lower probability of finding weaker dislocation sources in smaller structures.

Acknowledgments JRG gratefully acknowledges the financial support of the National Science Foundation (NSF) CAREER grant (DMR-0748267) and the Office of Naval Research (ONR) Grant No. N00014-09-1-0883. The author is particularly grateful to W.D. Nix, A.T. Jennings, D. Jang, J.-Y. Kim, Q. Sun, A. Ngan, C. Weinberger, J. Li, and D. Gianola for useful discussions.

References

1. Nix, W.D.: MSE 208: Mechanical behavior of materials class notes. (1980)
2. Greer, J.R., de Hosson, J.T.M.: Critical review: plasticity in small-sized metallic systems: intrinsic versus extrinsic size effect. Prog. Mater. Sci. **56**, 654–724 (2011)
3. Kraft, O., Gruber, P.A., Monig, R., Weygand, D.: Plasticity in confined dimensions. Annu. Rev. Mater. Res. **40**, 293 (2010)
4. Uchic, M.D., Shade, P.A., Dimiduk, D.M.: Plasticity of micrometer-scale single crystals in compression. Annu. Rev. Mater. Res. **39**, 361–386 (2009)
5. Brenner, S.S.: Tensile strength of whiskers. J. Appl. Phys. **27**, 1484–1490 (1956)
6. Brenner, S.S.: Growth and properties of "whiskers". Science **128**, 568 (1958)
7. Thompson, C.V.: The yield stress of polycrystalline thin films. J. Mater. Res. **8**, 237 (1993)
8. Uchic, M.D., Dimiduk, D.M., Florando, J.N., Nix, W. D.: Sample dimensions influence strength and crystal plasticity. Science **305**, 986 (2004)
9. Greer, J.R., Oliver, W.C., Nix, W.D.: Size dependence of mechanical properties of gold at the micron scale in the absence of strain gradients. Acta Mater. **53**, 1821 (2005)
10. Kim, J.-Y., Jang, D., Greer, J.R.: Insights into deformation behavior and microstructure evolution in Nb single crystalline nano-pillars under uniaxial tension and compression. Scr. Mater. **61**, 300 (2009)
11. Tang, H., Schwartz, K.W., Espinosa, H.D.: Dislocation escape-related size effects in single-crystal micropillars under uniaxial compression. Acta Mater. **55**, 1607 (2007)
12. Bei, H., Shim, S., Pharr, G.M., George, E.P.: Effects of pre-strain on the compressive stress–strain response of Mo-alloy single-crystal micropillars. Acta Mater. **56**, 4762 (2008)
13. Lee, S., Han, S., Nix, W.D.: Uniaxial compression of fcc Au nanopillars on an MgO substrate: the effects of prestraining and annealing. Acta Mater. **57**, 4404 (2009)
14. Greer, J.R., Nix, W.D.: Nanoscale gold pillars strengthened through dislocation starvation. Phys. Rev. B **73**, 245410 (2006)
15. Shan, Z.W., Mishra, R., Syed, S.A., Warren, O.L., Minor, A.M.: Mechanical annealing and source-limited deformation in submicron-diameter Ni crystals. Nat. Mater. **7**, 115–119 (2008)
16. Weinberger, C., Cai, W.: Surface controlled dislocation multiplication in metal micro-pillars. Proc. Natl. Acad. Sci. U. S. A. **105**, 14304 (2008)
17. Rao, S.I., Dimiduk, D.M., Parthasarathy, T.A., Uchic, M.D., Tang, M., Woodward, C.: Athermal mechanisms of size-dependent crystal flow gleaned from three-dimensional discrete dislocation simulations. Acta Mater. **56**, 3245 (2008)
18. Gilman, J.J.: Micromechanics of Flow in Solids. McGraw-Hill, New York (1969)
19. Zhu, T., Li, J., Samanta, A., Leach, A., Gall, K.: Temperature and strain rate dependence of surface dislocation nucleation. Phys. Rev. Lett. **100**, 025502 (2008)
20. Liu, Z.L., Liu, X.M., Zhuang, Z., You, X.C.: Atypical three-stage-hardening mechanical behavior of Cu single-crystal micropillars. Scr. Mater. **60**, 594 (2009)

Small Angle X-Ray Scattering in Grazing Incidence Geometry

► Selected Synchrotron Radiation Techniques

S

Small Unilamellar Vesicle (SUV)

► Liposomes

Small-Angle Neutron Scattering

► Small-Angle Scattering of Nanostructures and Nanomaterials

Small-Angle Scattering

▶ Small-Angle Scattering of Nanostructures and Nanomaterials

Small-Angle Scattering of Nanostructures and Nanomaterials

M. Laver
Laboratory for Neutron Scattering, Paul Scherrer Institut, Villigen, Switzerland
Materials Research Division, Risø DTU, Technical University of Denmark, Roskilde, Denmark
Nano–Science Center, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
Department of Materials Science and Engineering, University of Maryland, College Park, MD, USA

Synonyms

SANS; SAXS; Small-angle neutron scattering; Small-angle scattering; Small-angle X-ray scattering

Definition

Small-angle scattering of nanostructures and nanomaterials encompasses the measurements of scattering from structures with length scales ranging between the near-atomic (nanometer) to the near-optical (micrometer), using beams of nanometer wavelengths or less.

Overview

Small-angle scattering reveals structural features on length scales between the near-atomic (nanometer) to the near-optical (micrometer). With such versatility, the technique has made an astounding impact in many fields of research including polymer systems [1, 2], complex fluids [3], biology [4], condensed matter physics, and materials science [1, 5]. The methodology for performing small-angle scattering studies extends back to the 1930s [6], following the first small-angle X-ray scattering (SAXS) studies. Subsequently, synchrotron sources have enabled explorations with ever smaller sample amounts, in complex sample environments. Laboratory SAXS instruments are now routinely used for sample characterization in many research institutions. The creation of large-scale facilities producing neutrons for research has proved to be instrumental for small-angle neutron scattering (SANS) studies. As X-rays and neutrons interact differently with matter, SAXS and SANS are quintessentially complementary, and a particular technique may frequently become an indispensable probe for a particular application.

This entry is organized as follows: in section "Key Principles", the general concepts of small-angle scattering are explained. In section "X-Rays or Neutrons", the advantages and practical aspects of X-ray and neutron techniques are compared. In section "Magnetic Neutron Scattering", the discussion is focused on small-angle neutron scattering from magnetic structures, and neutron polarization analysis (section "Neutron Polarization Analysis") is provided as an example of a developing field for small-angle studies. The concluding section (section "Summary and Outlook") incorporates an outlook toward novel directions for the small-angle scattering technique.

An overview of symbols used repeatedly throughout this entry is provided in Table 1.

Key Principles

Bragg Diffraction

Small-angle scattering patterns are frequently continuous in nature, rather than consisting

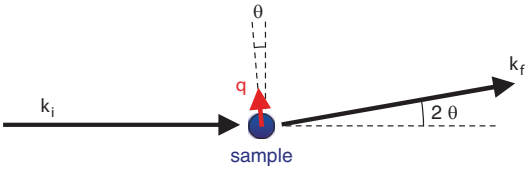
Small-Angle Scattering of Nanostructures and Nanomaterials, Table 1 List of symbols used throughout this entry

θ	Bragg angle of diffraction
$\vec{k}_i$, $\vec{k}_f$	Wavevector of incoming, final (scattered) beams
λ	Wavelength of beam
$\vec{q}$, q	Scattering vector, its magnitude
ζ	Angle of q in detector plane
d	Spacing between Bragg planes
D	Characteristic length scale within sample
$\vec{r}$, r	(Real space) position vector in sample or within nano-object, its magnitude
b, b_x, b_n, b_m	Scattering length, that of X-rays, that of neutrons from nuclei, that of neutrons from atomic moments
ρ, ρ_0	Scattering length density, that of solvent or matrix
F	Form factor of single nano-object
I	Scattered intensity
M, N	Magnetic, nuclear parts of (neutron) form factor of single nano-object
P	Form factor $P = \|F\|^2$
S	Structure factor
C	Component of magnetic moment perpendicular to $\vec{k}_i$
ϕ	Angle of magnetic moment in detector plane
Z	Component of magnetic moment parallel to $\vec{k}_i$
$\vec{A}$	Halpern–Johnson vector

of crystalline diffraction peaks. Nevertheless, Bragg's law encapsulates a fundamental relationship

$$\lambda = 2d \sin\theta = \frac{4\pi}{q} \sin\theta \qquad (1)$$

showing that the angle of diffraction θ varies inversely with the separation d of the diffracting lattice planes. It can be seen that the scattering from objects on nano- to micrometer scales falls predominantly in the small-angle regime $<1°$. A beam of $\lambda \simeq 7$ Å wavelength, for example, would scatter from objects with a d-spacing of $\simeq$1000 Å into an angle $2\theta \simeq 0.4°$ separated from the unscattered beam. Practical constraints on the wavelengths used are imposed by the spectrum of the source available and by the absorption



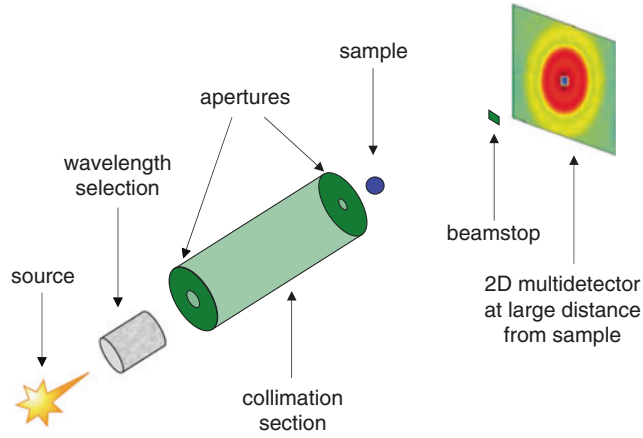
Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 1 Schematic depicting the relation for the scattering vector $\vec{q} = \vec{k}_f - \vec{k}_i$.$2\theta$ is the angle between the incident beam $\vec{k}_i$ and the scattered beam $\vec{k}_f$

of the beam by the sample, which increases with wavelength for both X-rays and neutrons. Typically, "soft" X-rays ($\approx$1 Å, or equivalently, 12 keV) or neutrons ($\approx$5 Å) from a cold ($\lesssim$40 K) moderator are used in small-angle studies. All scattering measurements are made in reciprocal (Fourier transform) space. The scattering vector $\vec{q}$, whose magnitude q satisfies (Eq. 1), parameterizes the scattering pattern in reciprocal space and is helpful when comparing patterns between different instruments. The vectorial version of (Eq. 1), $\vec{q} = \vec{k}_f - \vec{k}_i$, is illustrated in Fig. 1, where $\vec{k}_i$, $\vec{k}_f$ are, respectively, the wavevectors of the incident and final (scattered) beams.

Experimental Method

Small-Angle Scattering

The fundamental task of a small-angle instrument is to separate the intense, unscattered direct beam from the weaker scattering at small angles 2θ. A common experimental setup, depicted in Fig. 2, resembles the classic pinhole camera. Following the source, the wavelength λ of the incoming beam is selected. Due to the low scattering angles, beam intensity can be maintained using a large wavelength spread $\Delta\lambda/\lambda$ without significant detriment to the q-resolution. On small-angle X-ray scattering (SAXS) instruments, a monochromator crystal such as Si (111) is used if a wavelength spread narrower than the source distribution is required. On small-angle neutron scattering (SANS) instruments at pulsed neutron sources, a neutron's wavelength may be calculated by timing its journey to the detector



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 2 Schematic of a typical small-angle scattering instrument

(known as the *time of flight* technique). At continuous (e.g., reactor) neutron sources, velocity selectors are mainly used. Velocity selectors comprise of channels or slotted disks forming helical pathways about an axle; only neutrons of a particular velocity are able to pass through the rotating helical pathway. The desired neutron wavelength is selected by varying the rotational speed. The opening angle of the helical pathway is typically set to give wavelength spreads of $\Delta\lambda/\lambda \approx 0.06$.

After the wavelength selection, a collimation section that may consist of a pair of apertures or a slit system ensures that the beam has a tight angular spread $<0.1^\circ$ when it is incident on the sample. The precise collimation and resulting angular spread are often tuned to match the desired measurement range in 2θ (or, equivalently, in q) and to optimize beam intensity. A two-dimensional multidetector, comprising an array of pixels in a plane normal to the unscattered beam, is placed at a sufficient distance behind the sample so as to resolve the small-angle scattering in the desired q range. For example, the D11 SANS instrument at the Institut Laue-Langevin in Grenoble has a multidetector of 128×128 pixels, each pixel 7.5×7.5 mm^2, and features maximum collimation and sample-to-detector distances of 40.5 m and 39 m, respectively, a configuration that permits scattering vectors as low as $q \approx 0.0003$ Å^{-1} to be measured. The 2D area detectors used in SANS or SAXS are designed to have high detection efficiencies (>80 %), and may need to be shielded from intense fluxes, namely, the unscattered beam, that may saturate and can even damage the detector. A beamstop may be placed in front of the detector for this purpose.

Ultra-Small-Angle Scattering

For some systems one would like to extend the range of observable structures to length scales larger than those which may be probed using the habitual pinhole setup (Fig. 2) and measure scattering at "ultra" small angles. To achieve this, an experimental geometry developed by Bonse and Hart [7], which is now commonplace, utilizes a pre-sample monochromator comprising of a single crystal out of which a channel has been precisely cut. The beam passes in through one end of the channel at an incident angle θ, and appears at the other end (also at an angle θ) after having bounced multiple times from Bragg reflections off both sides of the channel. The resulting beam has an extremely small angular divergence $\lesssim 10^{-4}$ degrees. A similar channel-cut crystal placed behind the sample analyzes the angles of the scattered beam. With this setup, measurements are possible very close to the unscattered beam, down to scattering vectors $q \lesssim 10^{-5}$ Å^{-1}, in other words, up to length scales in the tens of micrometer range. Unlike the pinhole SANS setup where a single measurement with the 2D detector captures an entire surface of scattering vectors $\vec{q}$, the Bonse-Hart geometry can only probe one point in $\vec{q}$ at a time and the measurement times required with this technique are accordingly much longer.

General Theory

In this and in the following subsection (section "Spherically Symmetric Forms") the concepts of *scattering length density, contrast, structure factor, form factor, polydispersity*, and *Guinier* and *Porod regimes* are introduced. These are also explained in several introductory textbooks [1, 3, 6]. The q-dependence of the measured scattered intensity is described by the "differential cross section," denoted simply as $I(\vec{q})$. For now the scattering is presumed to be elastic ($\left|\vec{k}_i\right| = \left|\vec{k}_f\right|$ i.e., there is no energy transfer between beam and sample) and coherent (produces interference effects). Detailed treatments of inelastic and incoherent scattering may be found in the textbooks on diffraction methods [8].

Born Approximation and Scattering Length

One presumption of particular pertinence to small-angle scattering experiments is the *Born approximation*, in which the scattering is regarded to be weak (such that the scattering potential may be treated analytically as a perturbation). Within the Born approximation, the photon or neutron scatters no more than once during its passage through the sample. Then the measured scattered intensity is

$$I\left(\vec{q}\right) \sim \left|\sum_j b_j e^{i\vec{q}.\vec{R}_j}\right|^2 \qquad (2)$$

where $\vec{R}_j$ denotes the position of the jth scatterer in the sample, and b_j is the *scattering length*, epitomizing the capability of the jth object to scatter the beam. Neutrons scatter from nuclei and from magnetic moments [8], while X-rays scatter from electrons (Thomson scattering) [1, 3]: the scattering length of one electron is 2.8×10^{-5} Å. The differences between neutron and X-ray scattering will be discussed in section "X-rays or Neutrons". Here and throughout this work, leading prefactors are omitted for simplicity. These prefactors would, for example, ensure that the scattered intensity scales with the illuminated sample volume. In practice $I(\vec{q})$ is scaled onto absolute units through the measurement of either a standard sample or the incident beam flux, together with corrections for the attenuation of the beam by the sample and instrument-dependent factors such as electronic noise and detector efficiency [9]. With $I(\vec{q})$ converted to absolute units, intensities obtained on different small-angle instruments may be directly compared, and sample properties, for example, the stoichiometry of that part of the sample giving rise to scattering, may be estimated quantitatively.

Most model functions for small-angle scattering data presuppose that the Born approximation holds. In practice, for strong scatterers, care should be taken to verify that this applies before fitting models to $I(\vec{q})$. If multiple scattering is a problem, the probability of scattering may be diminished by making the sample thinner or, where possible, diluting it or decreasing the *contrast*. For grazing incidence experiment geometries such as the grazing-incidence SAXS (GISAXS) technique that are used to resolve in-plane structures on films [1], the Born approximation is sufficient for very small or thin scatterers, otherwise quantitative analysis of the GISAXS profile requires what is known as the "distorted-wave" Born approximation, where the scattering potential is divided into two parts such that an exact solution to the scattering problem may be obtained [10]. The grazing incidence geometry is not expounded here; further details may be found in Refs. [1] and [10].

Scattering Length Density and Contrast

Theoretical small-angle scattering profiles can be calculated directly from models comprised of individual atomic scatterers, but for fitting purposes the computational effort required is too great at present. As the objects leading to small-angle scattering are generally on larger scales, a more continuous, "coarse-grained" description of the scattering potential landscape would seem appropriate. A continuous function $\rho(\vec{r})$, known as the *scattering length density*, is accordingly defined [6]:

$$\rho\left(\vec{r}\right) = \sum_j b_j \delta\left(\vec{r} - \vec{R}_j\right)$$

S

and the measured scattered intensity becomes

$$I(\vec{q}) \sim \left| \int e^{i\vec{q}\cdot\vec{r}} \rho\left(\vec{r}\right) d\ \vec{r} \right|^2 \tag{3}$$

$$\sim \left| \int e^{i\vec{q}\cdot\vec{r}} \rho_0 d\ \vec{r} + \int e^{i\vec{q}\cdot\vec{r}} \left(\rho\left(\vec{r}\right) - \rho_0\right) d\ \vec{r} \right|^2 \tag{4}$$

The last equation serves to illustrate the concept of *contrast*. Here it is useful to consider the scattering from nanoparticles in solution, or equally, nanoscale heterogeneities in a solid matrix. ρ_0 would be the scattering length density of the matrix or solvent, while $\rho(\vec{r})$ would be that of the nanoparticles or heterogeneities. The first term in (Eq. 4) involves just the matrix or solvent and is nonzero only at $q = 0$ [the Fourier transform of a constant is $\delta(0)$], so no small-angle scattering originates from the matrix or solvent. In the second term in (Eq. 4), the particles or heterogeneities give forth to small-angle intensity proportional to the squared difference $(\rho(\vec{r}) - \rho_0{}^2)$. This is referred to as the "contrast factor." Systematic alterations of the solvent ρ_0 may thus be used to determine the scattering length density of a solute ρ_1, since the intensity will go to zero at the match point where $\rho_1 = \rho_0$. This is called "contrast matching." In multicomponent systems, separate contrast-variation analyses for different q-regimes or for different features in the scattering profile can also be used with great effect to reveal the sizes and shapes of the underlying components. Contrast methods were recently demonstrated, for example, in a prominent study of the nanoscale phases within cement [11].

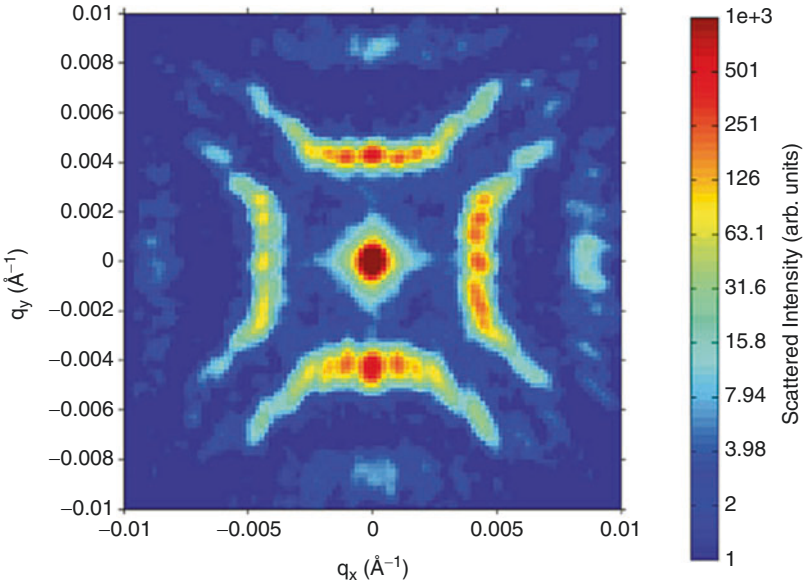
Structure Factor and Form Factor

The notions of *structure factor* and *form factor* are familiar concepts in crystallographic work [8]. Considering N identical scattering nano-objects, each centered at a position $\vec{s}_j$, the scattered intensity becomes

$$I\left(\vec{q}\right) \sim \left| \sum_j^N e^{i\vec{q}\cdot\vec{s}_j} \right|^2 \times \left| \int_{\text{nano-object}} e^{i\vec{q}\cdot\vec{r}} \rho\left(\vec{r}\right) d\vec{r} \right|^2 = S\left(\vec{q}\right) \times \left| F\left(\vec{q}\right) \right|^2$$

The left side of the multiplication sign defines the *structure factor* $S(\vec{q})$. The integral on the right side ranges over an isolated nano-object, and this term defines the single particle *form factor* $F(\vec{q}) = \int e^{i\vec{q}\cdot\vec{r}} \rho(\vec{r}) d\vec{r}$. It is also commonplace to refer to the function $P(\vec{q}) = \left|F(\vec{q})\right|^2$ as the form factor too. In the limit where the nano-objects arrange into a crystal with positions $\vec{s}_j$ forming a perfect lattice, the structure factor becomes a series of δ-functions $S(\vec{q}) = \Sigma\imath\, \delta(\vec{q} - \vec{G}\,\imath)$, where $\vec{G}\,\imath$ are the reciprocal lattice vectors of the crystal [8]. Crystalline Bragg-diffraction peaks appear in the scattering profile, as illustrated in Fig. 3, with intensities modulated by the form factor. To measure each $\vec{G}\imath$ peak, the sample must be rotated with respect to the neutron beam in order to satisfy the condition $\vec{q} = \vec{G}\imath$. For situations in the opposite limit where nano-objects are well separated, with no interactions between nano-objects, $S(\vec{q}) = 1$ for all $\vec{q}$ and the small-angle intensity depends only upon $P(\vec{q})$, that is to say, upon the form of an isolated particle. Such a situation is called the "dilute" limit because, with small-angle scattering from solutions, it may be attained by sufficient dilution of the solute [4, 9]. The near-dilute situation is also often realized, where the interparticle interactions are finite but small, such that $S(\vec{q})$ is essentially uniform in the q-range of interest. In situations where the structure factor and the form factor are comparable, for example where the volume fraction of nano-objects is high but no long-range order emerges, the $I(\vec{q})$ must be fitted with models for both $S(\vec{q})$ and $P(\vec{q})$. Details of the many model functions available may be found in the literature [1–6, 13–15].

There are also several numerical methods for modeling $P(\vec{q})$ or $S(\vec{q})$. Numerical methods often start from a radial pair distribution function, which can be obtained by an indirect Fourier transform of $I(q)$ [3]. One approach, commonly used to reconstruct the shape of biological macromolecules, imitates the scattering length density profile $\rho(\vec{r})$ of the macromolecule by several 100 close-packed "beads," each of constant ρ. Genetic algorithms and simulated annealing



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 3 Example of a crystal diffraction pattern on a small-angle neutron detector, here from a flux line lattice [12] in a single crystal of superconducting niobium. To obtain this picture, several detector images were summed as the sample was rotated over a range $\pm 1^\circ$ of rocking angles about the vertical and horizontal axes. Here 28 distinct first order Bragg spots are discernible, due to four scalene triangular and two square flux line lattices coexisting in different regions of the sample. The peaks on the left side of the picture appear weaker due to detector efficiency. The unscattered beam, not covered up by a beamstop in this experiment, appears in the center at $q = 0$

techniques are able to refine the virtual bead assemble to fit the experimental data [4, 16]. Reverse Monte Carlo methods have also been used to recreate the structure factor $S(\vec{q})$ in situations where quasi-long-range order is observed [17].

Spherically Symmetric Forms

Most texts on small-angle scattering begin by considering a spherically centrosymmetric nano-object [3, 6, 13], in other words, a nano-object whose scattering length density $\rho(\vec{r})$ depends only on $r = |\vec{r}|$. Here this scenario is also pursued in order to illustrate concepts such as contrast and polydispersity and furthermore, to highlight the different *regimes* in the scattering profile. To obtain the form factor for a spherical object, it is useful to select a spherical polar coordinate system such that the pole lies along $\vec{q}$; the angular components of the integral within $F(\vec{q})$ may then be evaluated, yielding

$$F(q) = 4\pi \int \rho(r) r^2 \frac{\sin(qr)}{qr} dr$$

The form factor is here a function only of the magnitude $q = |\vec{q}|$ of the scattering vector. Even when the nano-object is not spherical, but a large number of randomly oriented nano-objects are contained in the sample, these orientations may be averaged over and the result becomes again a function of $q = |\vec{q}|$. It is usual to average the scattering measured on the 2D detector (cf. Fig. 2) over the detector's azimuthal direction, so reducing the scattering profile to one dimension, i.e., an I versus q dataset. In section "Magnetic Neutron Scattering" an example is given of where the scattering shows an azimuthal dependence arising for a reason other than crystalline order (cf. Fig. 3). The azimuthal information can impart vital clues as to the nature of the underlying nanostructures.

Core-Shell Form Factor

Core-shell nano-objects are frequently encountered in small-angle scattering studies, e.g., aerosol droplets, polymer micelles [2], inorganic nanoparticles [18]. In addition, core-shell models may provide a useful quantitative starting point even when nano-objects depart from being spherically symmetric. The spherical core-shell model form factor is

$$F(q) = 4\pi\int_0^{r_c} \rho_c r^2 \frac{\sin(qr)}{qr} dr + 4\pi\int_{r_c}^{r_s} \rho_s r^2 \frac{\sin(qr)}{qr} dr$$
$$= 4\pi(\rho_c - \rho_s) r_c^3 \frac{j_1(qr_c)}{qr_c} + 4\pi\rho_s r_s^3 \frac{j_1(qr_s)}{qr_s} \quad (5)$$

where the sphere has a uniform core of radius r_c and scattering length density ρ_c, encapsulated by a uniform shell of scattering length density ρs extending to radius $r_s > r_c$, and $j_1(x) = \frac{\sin x}{x^2} - \frac{\cos x}{x}$ is a spherical Bessel function. For brevity, ρ_c and ρ_s are measured relative to the scattering length density ρ_0 of the surrounding solution or matrix; equivalently, ρ_0 may be considered to be zero. Equation 5 may be simplified by introducing volume terms such as $V_c = \frac{4}{3}\pi r_c^3$. Figure 4b exemplary $P(q) = |F(q)|^2$ curves for this core-shell model are plotted.

Polydispersity

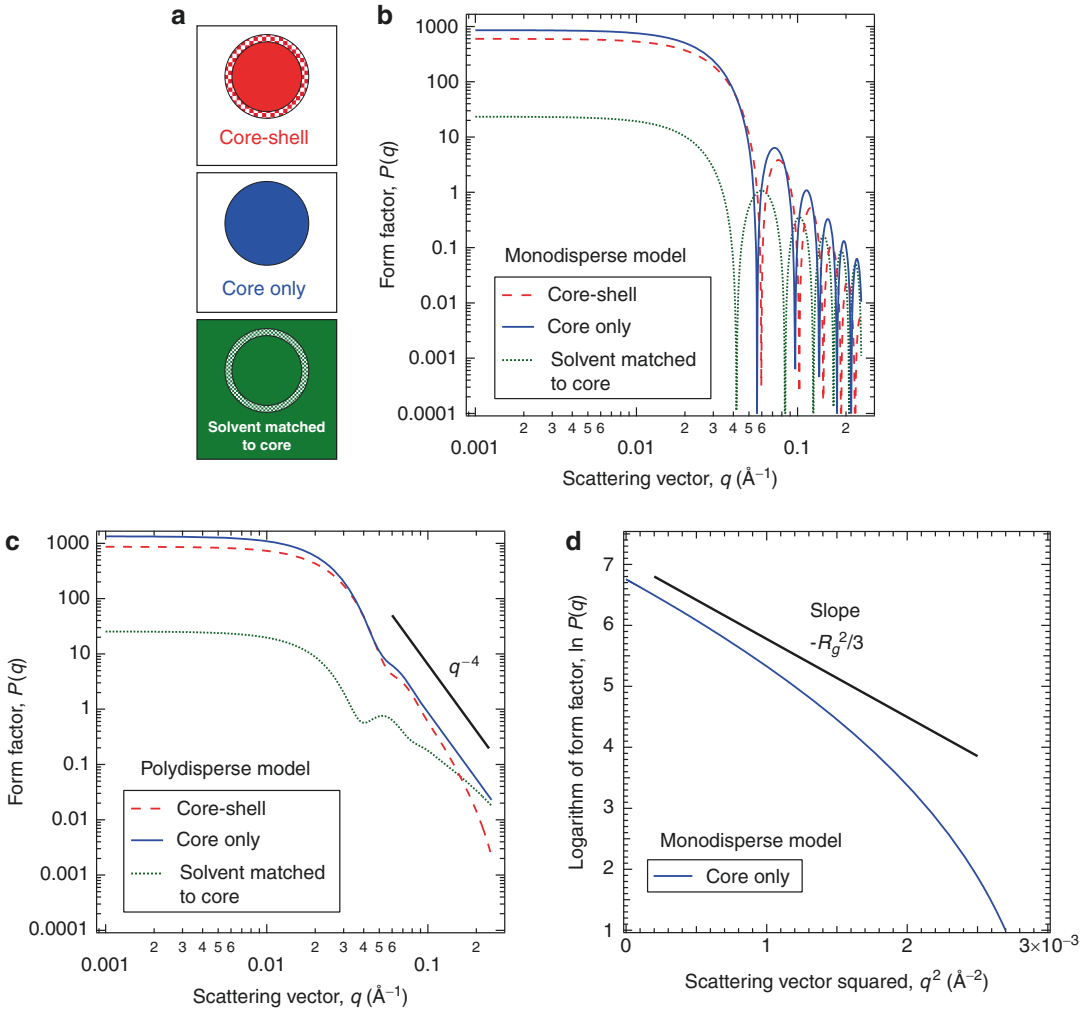
In reality, most scattering systems show a distribution in their sizes and shapes; this is known as *polydispersity*. Form factors can be straightforwardly modified to average over a distribution in size or shape, and several functions defining distributions for the polydispersity are available [19]. The effects of polydispersity on the small-angle scattering profile are now examined briefly. In Fig. 4c the core-shell model is modified to include a distribution in core radii with standard deviation 20 % of the mean core radius. By comparing this figure with Fig. 4b, the adverse effects of polydispersity – a smearing in q of the scattering profile – can readily be seen. A smearing of the scattering profile also results from a finite instrument resolution; nonetheless it is clear that sizeable benefits are gained if samples for small-angle scattering studies are as monodisperse as possible. Figure 4 furthermore illustrates how contrast variation may be exploited, where feasible: by matching the solvent ρ_0 to the core ρ_c (dotted green line in Fig. 4), the interference bumps in the profile are more discernible, notably in polydisperse situations.

Regimes in the Scattering Profile

The appearance of distinctive q regimes in the scattering profile may also be seen in Fig. 4. The solid sphere model (continuous blue lines in Fig. 4) features a characteristic length scale D, namely, the sphere radius 80 Å. The scattering features a high-q regime for $qD \gg 1$, where the scattering profile appears to decay algebraically with q, and a regime at low q where the profile falls off more slowly in q. These regimes emerge independently of the absolute intensity and of any model, and invariably provide useful clues as to the size and homogeneity of the nano-object. In the high-q regime $qD \gg 1$, known as the *Porod regime*, the exponent α of the power-law decay $I(q) \sim q^{-\alpha}$ reveals the "fractal dimension" of the scattering objects [14]. The $\alpha = 4$ observed in Fig. 4c, for example, indicates smooth surfaces, as expected for the model of uniform spheres with boundaries sharp in $\rho(r)$.

In the *Guinier regime* as $q \to 0$, the scattering follows a Gaussian dependence $\sim \exp\left(-\frac{1}{3}q^2R_g^2\right)$ where Rg is the "radius of gyration" of the nano-object [6], and is analogous to the radius of gyration in mechanics. For uniform solid spheres of radius r_c, $R_g^2 = \frac{3}{5}r_c^2$. To reveal this behavior at low q, a "Guinier plot" is useful, where the logarithm of the scattering is plotted versus q^2; the plot will be linear with slope $-\frac{1}{3}R_g^2$ at sufficiently low q. A Guinier plot for the sphere model is shown in Fig. 4d, together with the expected low-q slope (black line in Fig. 4d). Comparing the two, it may be seen that the Guinier approximation holds reasonably well when $q^2 \lesssim 0.0008$ Å^{-2}, that is, for $qRg \lesssim 1.3$. The upper limit for the Guinier regime is found to depend somewhat on the scattering model [13].

Almost every small-angle scattering article incorporates an analysis of the low- or high-q



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 4 The spherical core-shell form factor $P(q)$. (**a**) Schematic showing the various parameters for the model curves plotted in (**b**–**d**): in (**b**) the particles are monodisperse; in (**c**) the particles have polydisperse core sizes with the standard deviation of the polydispersity distribution set at 20 % of the mean core radius r_c. The *dashed red* and *dotted green lines* indicate $P(q)$ with $r_c = 70$ Å and a shell 10 Å thick; the *dashed red line* shows $P(q)$ when the core scattering length density ρ_c is twice that of the shell ρ_s; the *dotted green line* shows $P(q)$ when the solvent ρ_0 matches ρ_c. The continuous *blue line* shows $P(q)$ for a core only model with $r_c = 80$ Å; in (d), this line is redrawn on ln $P(q)$ versus q^2 axes. *Black lines* are guides to the eye

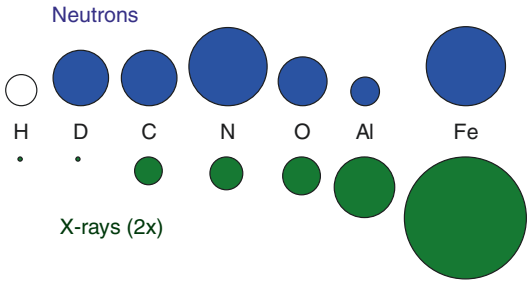
regimes. Extensive examples of Guinier plots and associated methods (e.g., the Zimm and Kratky plots) that quantify scattering regimes without resorting to detailed model fitting may also be found in the literature [3]. There have also been efforts to encapsulate both low-q and high-q regimes in unified models. A common example is the Beaucage model, applicable to situations where a nanostructure exhibits a hierarchy of scales, each discernible scale contributing a pair of Guinier and Porod regions to the scattering profile [15].

X-rays or Neutrons?

Scattering Length

The basic principles of small-angle scattering, as overviewed in previous sections, have the same

S



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 5 Schematic indicating the propensity for scattering of X-rays or neutrons [20] by selected atoms. The radii of the circles are drawn proportional to the scattering length. For X-rays, the scale has been 2× magnified for clarity. The scattering length of hydrogen for neutrons is negative, in contrast to the other atoms depicted

Small-Angle Scattering of Nanostructures and Nanomaterials, Table 2 The scattering properties, for neutrons, of selected substances

Substance	Neutron attenuation length (cm)	Neutron scattering length density ($\times 10^{-6}$ Å^{-2})
H_2O	0.18	−0.56
D_2O	1.55	6.33
C	1.59	7.53
B	0.0036	6.91
^{10}B	0.00067	−0.14
^{11}B	1.35	8.51
Al	7.74	2.08
Fe	0.629	8.02
Cd	0.0031	2.27

For the calculation of these values, room temperature densities are used, and the elements with no isotope specified are deemed to comprise of their isotopes in naturally abundant ratios

form for both X-rays and neutrons. However, there are fundamental distinctions between the techniques arising from how the X-rays or neutrons interact with matter. One such distinction is patent in the atomic scattering length b [cf. (Eq. 2)]. For neutrons, b_n depends on the structure of the nucleus and is isotope dependent. For X-rays, b_x depends on the electronic structure and scales with the total number of electrons Z in the atom. At the length scales probed by small-angle scattering, nuclei look like point sources of scattering and b_n is q-independent, whereas for X-rays (atomic absorption edges neglected) $b_x(q)$ falls off slowly with q – this is essentially the Guinier regime of the form factor for the atomic electron clouds, with $b_x(0) = Z \times 2.8 \times 10^{-5}$ Å. Low-resolution modeling of SAXS profiles can be achieved with bx considered constant. The scattering lengths for selected atoms are pictured in Fig. 5 for both neutrons [20] and X-rays.

Generation of Contrast

The scattering length density ρ of an elementary volume V (e.g., a molecular volume) within the sample is calculated by summing the scattering lengths of atoms within V, that is, $\rho = \Sigma_j b_j/V$. Here, a uniform electron density needs to be assumed for X-rays. The scattering length densities for neutrons of selected substances is listed in Table 2. Comparing with Fig. 5, it is clear that there is an enormous distinction between hydrogen and deuterium. The former has $b_n = -3.74 \times 10^{-5}$ Å, the latter 6.67×10^{-5} Å [20], and this is reflected in ρ as shown for H_2O and for D_2O in Table 2. Great contrast may be achieved by using hydrogenated solutes in deuterated solvents. Polystyrene in toluene, for example, presents a contrast of 0.48×10^{-6} Å^{-2} when both are hydrogenated, whereas the contrast of hydrogenated polystyrene in deuterated toluene is 4.2×10^{-6} Å^{-2}. In contrast-variation studies, tuning the solvent scattering length density ρ_0 may be accomplished by using a binary mixture whose components have widely different scattering length densities; these are mixed in a linear ratio to yield the desired ρ_0. For X-rays, the two components would have different electron densities; for neutrons, they may be deuterated and hydrogenated versions of the same solvent. Isotope substitution may also be used to create SANS contrast *inside* a particle by specific isotope labeling of a particular part of interest, which can be used, for example, to determine how this part fits in with the rest of the particle. In SAXS, a novel way to induce contrast is to measure scattering profiles at different photon energies in the vicinity of an atomic absorption edge of one of the elements within the sample: for this element, the form factor bx varies across the absorption edge, while the rest of the scattering potential landscape is essentially

invariant in the small range of photon energies measured. This contrast-variation technique is known as anomalous SAXS (ASAXS) and may be exploited for the large number of elements having an absorption edge in the range 5–25 keV [1, 9, 21].

Neutron Absorption

Another aspect of neutron scattering that is heavily dependent on the isotope is absorption. A few elements, for example boron and cadmium (c.f. Table 2), are outstanding neutron absorbers in their naturally abundant forms. Not surprisingly, boron-containing substances and cadmium are common shielding materials. For scattering explorations of these and other such materials, samples need to be prepared using elements devoid of the absorbing isotopes; boron-11, for example, is seen from Table 2 to have an acceptable attenuation length. Separating elements into their different isotopes, however, becomes increasingly difficult and expensive for heavier atoms. It is also apparent from Table 2 that some materials, such as aluminum, are rather poor neutron absorbers. This is providential for the construction of bulky or demanding sample environments such as cryostats or pressure cells.

General Considerations

The majority of elements have neutron attenuation lengths on the order of a centimeter, reflecting the weak interaction of neutrons with matter. In general, neutron scattering explores the bulk of samples and, for similar reasons, effectively no radiation damage to the sample occurs with the neutron fluxes currently available (up to $\approx 10^8$ neutrons cm^{-2} s^{-1}) on SANS instruments. For SAXS experiments at synchrotron sources, the high brilliance of the X-ray beam ($\approx 10^{16}$ photons cm^{-2} s^{-1}) can damage soft matter and biological samples within milliseconds [9]. This may affect the validity of the scattering profile measurement. For SAXS the sample thickness may also need to be chosen to obtain the optimum scattered intensity, since at the photon energies (≈10 keV) used in SAXS, the X-ray attenuation length is on the order of millimeters for the lighter elements [9], which make up most soft matter systems. The many orders of magnitude more intense fluxes available on SAXS beamlines are, of course, invaluable for high-resolution, low uncertainty measurements of scattering profiles. The associated fast data acquisition times also facilitate the characterization of time-dependent effects. Likewise, relatively small sample quantities are typically required (<50 μL for SAXS) compared to SANS (≈1 mL).

Neutron Incoherent and Spin Scattering

In previous sections, it has been shown that variations in scattering length density on the nanoscale give rise to coherent small-angle scattering. However, if the scattering length b varies in a disordered fashion for different atoms of the same element, then for neutrons, since the nuclei are effectively point sources, this gives rise to incoherent (q-independent) scattering [8]. This adds to the background of the coherent scattering profile and in a few cases may be significant enough to limit the measurement even when longer count times are used. A variation in b for the same element can arise due to different isotopes, present in sufficient abundance, having widely different scattering lengths. Another reason for a variation in b occurs when the nuclei have nonzero spin and the scattering length depends on the spin state. Neutrons are sensitive to such spin disorder because the neutron itself has spin $\frac{1}{2}$. A notorious example is hydrogen: this has an incoherent scattering length of 25.3×10^{-5} Å, compared to 4.1×10^{-5} Å for deuterium. Deuterated solvents should therefore be used where possible.

The spin of the neutron is rather advantageous for the exploration of magnetic structures. Small-angle scattering can arise from nano- to micrometer-scale variations in the magnetic field, as well as correlations over similar length scales between atomic moments. This is discussed in the following section.

Magnetic Neutron Scattering

The spin-$\frac{1}{2}$ nature of the neutron and its associated magnetic moment make neutron scattering techniques a natural probe of magnetic systems. Indeed, magnetic structures are considerably better

S

understood following the dawn of high-flux neutron sources. The origin of an observed scattering signal – whether it is nonmagnetic (i.e., from nuclei) or magnetic in origin – can be determined by measurements at either side of the appropriate thermodynamic phase transition. In other cases, for example if magnetic and structural order appear simultaneously, the origin of the signal may be determined by analyzing the neutron polarization and/or the azimuthal dependence of the scattering on the 2D detector. Furthermore, neutron polarization analysis can convey additional information as to the orientation of the magnetic moments behind the scattering signal. The small-angle applications of neutron polarization analysis are introduced in the second part of this section (section "Neutron Polarization Analysis").

Unpolarized Neutrons

The interaction potential of the neutron with a magnetic field $\vec{B}$ is $-\vec{\mu}.\vec{B}$, where $\vec{\mu} = \gamma\mu_N \vec{\sigma}$ is the magnetic moment operator for the neutron, γ is the neutron gyromagnetic ratio, and μ_N is the nuclear magneton [8]. $\vec{\sigma}$ is the Pauli spin operator describing the neutron spin state; its average value characterizes the polarization of the neutron beam. All the general principles explained in the previous sections can be used to describe the small-angle scattering from magnetic structures, whether the neutrons are polarized or not. Noting, for example, that the scattering interaction potential $-\vec{\mu}.\vec{B}$ is already continuous, (Eq. 3) for the unpolarized elastic coherent scattering becomes, with a collinear field B in the sample

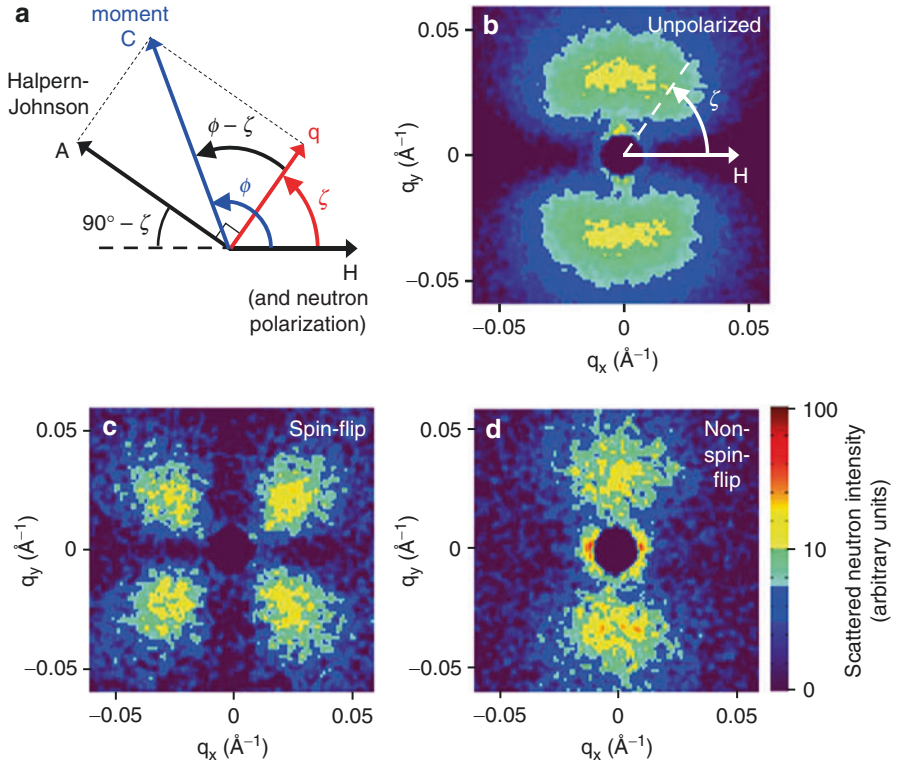
$$I\left(\vec{q}\right) \sim \int e^{i\vec{q}.\vec{r}} B\left(\vec{r}\right) d\vec{r}$$

This prescription is very convenient for flux line lattices that form in Type-II superconductors [12], where crystalline diffraction peaks are frequently observed, as illustrated in Fig. 3. The long-range order and nanometer d-spacings of magnetic skyrmion lattices also result in SANS diffraction peaks [22].

The magnetic scattered intensity may also be estimated in terms of magnetic scattering lengths of individual atoms. Like the atomic form factors in SAXS, the magnetic scattering length $b_m(q)$ is close to constant at low-q [8], with $b_m = m \times 2.7 \times 10^{-5}$ Å/μ_B where m is the atomic moment in Bohr magnetons μ_B [5, 23]. Under the proviso that the model chosen to fit the scattered intensity is representative, this relation can be utilized to estimate the underlying moment behind a scattering feature. With SANS, studies may be directed at the magnetism of individual atoms, since in ferri- and ferromagnetic systems in particular, the structure factor at low-q affords a measure of the interactions between individual atomic moments separated by large distances. An example is the collection of "random anisotropy" magnets, where SANS arises from random variations in the magnetic anisotropy fields. SANS studies on these materials are the subject of a recent review [5], in which various structure factor models pertinent to magnetic scattering are also detailed.

Magnetic domain walls in ferri- or ferromagnets also give rise to small-angle scattering. Such walls are normally observed to be spatially uncorrelated so that the scattering is dominated by the form factor. Because the domain sizes D are typically micrometers or above, a pinhole SANS instrument is able only to measure the high-q (Porod) regime of this scattering. The full profile may be accessible with an ultra-small-angle instrument [7, 24]. An example of where structure-factor-dominated and form-factor-dominated magnetic scattering features occur in the same sample is demonstrated in the phase-separated perovskite oxides, which exhibit colossal magnetoresistance [25]. In these systems, micrometer-sized ferromagnetic regions materialize in the nonmagnetic matrix below a critical temperature T_c. The Porod regime of their form factor appears as SANS intensity $I(q) \sim q^{-4}$. At the same time, at temperatures in the vicinity of T_c, quasielastic critical fluctuations give (Ornstein–Zernike) scattering, which has a Lorentzian profile $I(q) \sim 1/\left[q^2 + (1/\xi)^2\right]$ where ξ is the correlation length.

The dipole nature of the interaction between the neutron and atomic magnetic moments means that only moments $\vec{m}$ which are perpendicular to the scattering vector $\vec{q}$ are effective in scattering [8, 23]. The relevant part of $\vec{m}$ is expressed by the Halpern–Johnson vector



Small-Angle Scattering of Nanostructures and Nanomaterials, Fig. 6 2D SANS detector images from magnetically distinct nanoscale heterogeneities in a ferromagnet. (**a**) Schematic drawn in the detector plane depicting how detector anisotropy arises from the component of the moment C lying at an angle ϕ in the detector plane. (**b**–**d**) Measurements from an iron–gallium single crystal under saturating field $\vec{H}$ applied along $\zeta = 0$ in the detector plane such that $\phi = 0$: (**b**) unpolarized neutrons, (**c**) polarized neutrons that scatter with spin flip or (**d**) without spin flip

$$\vec{A} = \hat{\mathbf{q}} \wedge \left(\vec{m} \wedge \hat{\mathbf{q}}\right) = \vec{m} - \left(\vec{m}.\hat{\mathbf{q}}\right)\hat{\mathbf{q}}$$

where $\hat{\mathbf{q}}$ is the unit vector $\vec{q} \,/\left|\vec{q}\right|$. In terms of $\vec{A}$, the unpolarized neutron scattering (Eq. 2) becomes

$$I\left(\vec{q}\right) \sim \sum_{j,l} \left(b_{nj} b^*_{nl} + b_{mj} b^*_{ml} \vec{A}_j \vec{A}_l\right) e^{i\vec{q}.\left(\vec{R}_j - \vec{R}_l\right)}$$

Spin-dependent (incoherent) scattering from nuclei has been excluded here; this would enter in the form of terms similar to those in $\vec{A}$ [23]. The Halpern–Johnson relationship can lead to an anisotropy on the SANS detector. To describe this anisotropy, it is convenient to use cylindrical polar coordinates with the longitudinal axis along the beam direction $\vec{k}_i$. The polar plane of this coordinate system is drawn schematically in Fig. 6a. The scattering vector $\vec{q} \simeq (q, \zeta, 0)$ at small angles. Magnetically distinct heterogeneities within a ferromagnet provide a useful demonstration: the moment of the jth nanoscale heterogeneity, or rather, the contrast of this nanoscale moment with the ferromagnetic matrix, is described by (C_j, ϕ_j, Z_j) in the cylindrical polar coordinates. The unpolarized small-angle scattered intensity is

$$I\left(\vec{q}\right) \sim \sum_{j,l} \left(N_j\left(\vec{q}\right) N^*_l\left(\vec{q}\right) + M_j\left(\vec{q}\right) M^*_l\left(\vec{q}\right)\right)\left(Z_j Z_l + C_i C_l \sin\left(\zeta - \phi_j\right) \times \sin\left(\zeta - \phi_l\right)\right) e^{i\vec{q}.\left(\vec{R}_j - \vec{R}_l\right)} \tag{6}$$

where $N(q)$ and $M(q)$ are, respectively, the nuclear and magnetic components of the form factor of a

S

single nanoscale heterogeneity. $(\zeta - \phi)$ is the angle between the moment and $\vec{q}$; the component perpendicular to the scattering vector $\vec{q}$ is $\sin(\zeta - \phi)$, as shown in Fig. 6a. Magnetic heterogeneities with moments co-aligned along a direction ϕ in the detector plane will give rise to a $\sin^2(\zeta - \phi)$ anisotropy on the detector, as illustrated in Fig. 6b. Correspondingly, any observed sine-squared anisotropy on the detector could indicate that the small-angle scattering is magnetic in origin. Nuclear scattering can also be anisotropic, for example, if elongated particles are co-aligned in a direction perpendicular to the beam. For further confirmation that the observed scattering is magnetic, analysis of the neutron polarization can be performed.

Neutron Polarization Analysis

A multilayer or "supermirror" (e.g., of Fe and Si) is commonly used to polarize the incident cold neutron beam. A similar device can in principle be employed to analyze the spin of the scattered neutrons. For SANS, however, the angular acceptance of these devices is not sufficient to cover the 2θ range of scattering angles and these devices may also confer an undesirable small-angle background. The recent development of ^{3}He cells that resolve these issues has made it possible to routinely perform polarization analysis on small-angle scattering profiles. Helium-3 has a particularly spin-dependent neutron absorption: with 100 % of the nuclei in the ^{3}He cell polarized, one spin state of the neutrons (that parallel to the ^{3}He spin) is negligibly absorbed, while the other is strongly absorbed. Thus one is able to determine whether, during scattering, a neutron flips its spin ("spin-flip" scattering) or not ("non-spin-flip" scattering). In other words, the polarization is analyzed fully in one direction; this is called "longitudinal" polarization analysis. The scattered intensity breaks down into four components, denoted ++, +−, −+, and −−; the first symbol designates the incident neutron spin, the second the spin of the scattered neutron.

Longitudinal polarization analysis allows an unambiguous distinction to be made between coherent nuclear scattering and spin (incoherent) nuclear scattering, or between magnetic scattering and nuclear scattering [21, 23]. The latter is aptly demonstrated with the example of magnetically distinct heterogeneities in a ferromagnetic matrix. The contribution from nuclear spins is excluded, as before. The non-spin-flip channel is found to be sensitive to nuclear scattering and to magnetic scattering that has a component of the Halpern–Johnson vector $\vec{A}$ parallel to the neutron polarization. Meanwhile, all the scattering observed in the spin-flip channel is magnetic in origin, arising from the components of the Halpern–Johnson vector $\vec{A}$ perpendicular $(\vec{A}\ \perp)$ to the neutron polarization. The neutron polarization at the sample follows the applied magnetic field $\vec{H}$, that is, along $\zeta = 0$ in this example. Then, the polarized scattered intensity

$$I^{\pm\mp}\left(\vec{q}\right) \sim \sum_{j,l} M_j M_l^* \left(Z_j Z_l + C_j C_l \sin\left(\zeta - \phi_j\right) \sin\left(\zeta - \phi_l\right) \cos^2\zeta\right) e^{i\vec{q}\cdot\left(\vec{R}_j - \vec{R}_l\right)} \tag{7}$$

$$I^{\pm\mp}\left(\vec{q}\right) \sim \sum_{j,l} \left(N_j N_l^* \pm \left(M_j N_l^* C_j \sin\left(\zeta - \phi_j\right) + N_j M_l^* C_l \sin\left(\zeta - \phi_l\right)\right) \sin\zeta + M_j M_l^* C_j C_l \sin\left(\zeta - \phi_j\right) \sin\left(\zeta - \phi_l\right) \times \sin^2\zeta\right) e^{i\vec{q}\cdot\left(\vec{R}_j - \vec{R}_l\right)} \tag{8}$$

where the form factors are $\vec{q}$-dependent functions $N(\vec{q})$ and $M(\vec{q})$. Comparing (Eq. 7) and (Eq. 8) to the unpolarized scattering (Eq. 6), additional $\sin(90° - \zeta)$ and $\cos(90° - \zeta)$ terms appear in the spin-flip and non-spin-flip channels, respectively. The origin of these terms, as illustrated in Fig. 6a, lies in the sensitivity of polarized neutrons to the components of $\vec{A}$. In the spin-flip channel (Eq. 7), a $\vec{A}_\perp\ \vec{A} \wedge \vec{A}_\perp$ term has been omitted that arises from chiral magnetic structures, such as spin helices or skyrmions [26]. The small-angle scattering from magnetic nanoparticles provides an example of where a significant nuclear component $N(\vec{q})$ of the form factor arises in conjunction with a magnetic component $M(\vec{q})$ from the same nano-object [18]. For this particular example of magnetically distinct heterogeneities illustrated in Fig. 6, the nuclear form factor $N(\vec{q})$ is found to be

negligible. With moments co-aligned at magnetic saturation along $\phi = 0$, the anisotropy on the detector follows a $\sin^4 \zeta$ dependence for the non-spin-flip channel and a $\sin^2 \zeta \cos^2 \zeta$ dependence for the spin-flip channel (Fig. 6c, d). It is clear that these anisotropies have provided unambiguous confirmation that the origin of the small-angle scattering is magnetic.

Summary and Outlook

Small-angle scattering techniques have made an enormous impact on condensed matter research, spanning across several disciplines. In this short entry, a comprehensive review of the wealth of publications in this realm has been avoided. Instead, the basic principles that form the quoin of topical small-angle scattering articles have been overviewed. The advantages and practical aspects of small-angle neutron and X-ray techniques have also been compared. For further reading, in-depth reviews focusing on selected disciplines are available, in particular for studies of soft matter [1, 2, 4, 9, 10]. These include discussions of the applications of small-angle scattering for biology [4], polymers [1, 2], disordered solids [1], and SAXS on solutions in general [3].

The examples presented in this entry tend toward "hard" condensed matter research, as despite significant topical results, there are presently few reviews in this area. Exceptions may be found in Ref. [5], where the focus is on magnetic SANS, in particular for nanocrystalline materials and magnetic fluids. In particular, magnetic neutron scattering and neutron polarization analysis have been introduced. Neutron polarization analysis has only just become routinely available on small-angle neutron instruments due to the recent development of ^{3}He apparatus to analyze the spin of the scattered neutron. This enables the separation of spin-incoherent from coherent scattering [1, 21]. Furthermore, through the particular character of the interaction between neutron spin and atomic moments [23], small-angle polarized neutron studies can provide tremendous insights into the magnetic nature of nanostructures [5, 18].

In conjunction with advances in beamline apparatus, significant progress is also resulting from the availability of increasingly powerful sources of X-rays and neutrons, which allow the imaging of nanostructures in unprecedented ways. The advent of the free electron laser – a source providing super-intense, femtosecond pulses of X-rays – allows unrivaled scattering studies in real time on ultrashort timescales, from the functional dynamics of biomolecules to magnetic spin-flip processes [27]. Concurrently, advancement steadily continues with the enhancement of existing techniques [28] and the introduction of new scattering models and numerical methods [4, 5, 15–17].

Acknowledgments DanScatt is acknowledged for financial support and Annabel J. Lingham is thanked for a careful reading of the manuscript.

Cross-References

Note: With the exception of 'Synchrotron Radiation Techniques', these cross-references are essentially a list of applications where it is apparent that small-angle scattering techniques have been employed and made a nontrivial contribution to the field. The length of the list reflects that small-angle scattering provides a staple technique across the realm of nanotechnology, and is routinely used for sample characterization.

- ▶ Carbon Nanotubes
- ▶ Chitosan Nanoparticles
- ▶ Dermal and Transdermal Delivery
- ▶ Electric Field–Directed Assembly of Bioderivatized Nanoparticles
- ▶ Fullerenes for Drug Delivery
- ▶ Gas-Phase Nanoparticle Formation
- ▶ Gold Nanorods
- ▶ Hollow Gold Nanospheres
- ▶ Light-Element Nanotubes and Related Structures
- ▶ Liposomes
- ▶ Macromolecular Crystallization Using Nanovolumes

- Magnetic-Field-Based Self-assembly
- Nano-Concrete
- Nanomaterials for Electrical Energy Storage Devices
- Nanomaterials for Excitonic Solar Cells
- Nanoparticles
- Nanoscale Water Phase Diagram
- Nanostructured Thermoelectric Materials
- Nanostructures Based on Porous Silicon
- Nanostructures for Coloration (Organisms other than Animals)
- Nanostructures for Photonics
- Optical Properties of Metal Nanoparticles
- Prenucleation Clusters
- Selected Synchrotron Radiation Techniques
- Self-assembly
- Self-assembly of Nanostructures
- Spider Silk
- Structural Color in Animals
- Wetting Transitions

References

1. Naudon, A., Schmidt, P.W., Stuhrmann, H.B.: In: Brumberger, H. (ed.) Modern Aspects of Small-angle Scattering, pp. 1–56, 181–220, and 221–254. Kluwer, Dordrecht (1995)
2. Hammouda, B.: SANS from polymers – review of the recent literature. J. Macromol. Sci. Part C: Polym. Rev. **50**, 14–39 (2010); Melnichenko, Y.B., Wignall, G.D.: Small-angle neutron scattering in materials science: recent practical applications. J. Appl. Phys. **102**, 021101 (2007)
3. Glatter, O., Kratky, O. (eds.): Small Angle X-ray Scattering. Academic, London (1982); Feigin, L.A., Svergun, D.I.: Structure Analysis by Small-Angle X-ray and Neutron Scattering. Plenum, New York (1987)
4. Putnam, C.D., Hammel, M., Hura, G.L., Tainer, J.A.: X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Q. Rev. Biophys. **40**, 191–285 (2007); Kasai, N., Kakudo, M.: X-ray Diffraction by Macromolecules. Springer, Berlin (2005)
5. Michels, A., Weissmüller, J.: Magnetic-field-dependent small-angle neutron scattering on random anisotropy ferromagnets. Rep. Prog. Phys. **71**, 066501 (2008); Avdeev, M.V., Aksenov, V.L.: Small-angle neutron scattering in structure research of magnetic fluids. Phys. – Uspekhi **53**, 971–993 (2010)
6. Guinier, A., Fournet, G.: Small-Angle Scattering of X-rays. Wiley, New York (1955)
7. Bonse, U., Hart, M.: Tailess X-ray single-crystal reflection curves obtained by multiple reflection. Appl. Phys. Lett. **7**, 238–240 (1965)
8. Squires, G.L.: Introduction to the Theory of Thermal Neutron Scattering. Cambridge University Press, Cambridge (1978); Lovesey, S.W.: Theory of Thermal Neutron Scattering from Condensed Matter, vol. 1. Oxford University Press, Oxford (1984); Balcar, E., Lovesey, S.W.: Theory of Magnetic Neutron and Photon Scattering. Oxford University Press, Oxford (1989)
9. Narayanan, T., Grillo, I.: In: Borsali, R., Pecora, R. (eds.) Soft Matter Characterization, pp. 725–782 and 900–952. Springer, Berlin (2008)
10. Tolan, M.: X-ray Scattering from Soft-matter Thin Films. Springer, Berlin (1999); Okuda, H., Kato, M., Kuno, K. Ochiai, S., Usami, N., Nakajima, K., Sakata, O.: A grazing incidence small-angle x-ray scattering analysis on capped Ge nanodots in layer structures. J. Phys.: Condens. Matt. **22**, 474003 (2010)
11. Allen, A.J., Thomas, J.J., Jennings, H.M.: Composition and density of nanoscale calcium-silicate-hydrate in cement. Nat. Mater. **6**, 311–316 (2007)
12. Huxley, A.: In: Huebener, P., Schopohl, N., Volovik, G.E. (eds.) Vortices in Unconventional Superconductors and Superfluids, pp. 301–339. Springer, Berlin (2002)
13. Glatter, O., May, R.: Small-angle techniques. In: Prince, E. (ed.) International Tables of Crystallography, vol. C, pp. 89–112. Kluwer, Dordrecht (2004)
14. Teixeira, J.: Small-angle scattering by fractal systems. J. Appl. Crystallogr. **21**, 781–785 (1988); Bale, H.D., Schmidt, P.W.: Small-angle X-ray-scattering investigation of submicroscopic porosity with fractal properties. Phys. Rev. Lett. **53**, 596–599 (1984); Ruland, W.: Small-angle scattering of two-phase systems: determination and significance of systematic deviations from Porod's law. J. Appl. Crystallogr. **4**, 70–73 (1971)
15. Beaucage, G.: Approximations leading to a unified exponential/power-law approach to small-angle scattering. J. Appl. Crystallogr. **28**, 717–728 (1995); Hammouda, B.: Analysis of the Beaucage model. J. Appl. Crystallogr. **43**, 1474–1478 (2010)
16. Chacón, P., Fernando Díaz, J., Morán, F., Andreu, J. M.: Reconstruction of protein form with X-ray solution scattering and a genetic algorithm. J. Mol. Biol. **299**, 1289–1302 (2000); Svergun, D.I.: Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. Biophys. J. **76**, 2879–2886 (1999); Svergun, D.I., Petoukhov, M.V., Koch, M.H.J.: Determination of domain structure of proteins from X-ray solution scattering. Biophys. J. **80**, 2946–2953 (2001)
17. Laver, M., Forgan, E.M., Abrahamsen, A.B.., Bowell, C., Geue, T., Cubitt, R.: Uncovering flux line correlations in superconductors by reverse Monte Carlo refinement of neutron scattering data. Phys. Rev. Lett. **100**, 107001 (2008)

18. Krycka, K.L., Booth, R.A., Hogg, C.R., Ijiri, Y., Borchers, J.A., Chen, W.C., Watson, S.M., Laver, M., Gentile, T.R., Dedon, L.R., Harris, S., Rhyne, J. J., Majetich, S.A.: Core-shell magnetic morphology of structurally uniform magnetite nanoparticles. Phys. Rev. Lett. **104**, 207203 (2010)
19. Walter, G., Kranold, R., Gerber, T., Baldrian, J., Steinhart, M.: Particle size distribution from small-angle X-ray scattering data. J. Appl. Crystallogr. **18**, 205–213 (1985)
20. Sears, V.F.: Neutron scattering lengths and cross-sections. Neutron News **3**, 26–37 (1992)
21. Stuhrmann, H.B.: Contrast variation in X-ray and neutron scattering. J. Appl. Crystallogr. **40**, s23–s27 (2007)
22. Mühlbauer, S., Binz, B., Jonietz, F., Pfleiderer, C., Rosch, A., Neubauer, A., Georgii, R., Böni, P.: Skyrmion lattice in a chiral magnet. Science **323**, 915–919 (2009)
23. Moon, R.M., Riste, T., Koehler, W.C.: Polarization analysis of thermal-neutron scattering. Phys. Rev. B **181**, 920–931 (1969); Halpern, O., Johnson, M.H.: On the magnetic scattering of neutrons. Phys. Rev. **55**, 898–923 (1939)
24. Wagner, W., Bellmann, D.: Bulk domain sizes determined by complementary scattering methods in polycrystalline Fe. Physica B **397**, 27–29 (2007)
25. Wu, J., Lynn, J., Glinka, C.J., Burley, J., Zheng, H., Mithcell, J.F., Leighton, C.: Intergranular giant magnetoresistance in a spontaneously phase separated perovskite oxide. Phys. Rev. Lett. **94**, 037201 (2005)
26. Grigoriev, S.V., Chernyshov, D., Dyadkin, V.A., Dmitriev, V., Maleyev, S.V., Moskvin, E.V., Menzel, D., Schoenes, J., Eckerlebe, H.: Crystal handedness and spin helix chirality in $Fe_1 - xCoxSi$. Phys. Rev. Lett. **102**, 037204 (2009); Pappas, C. Lelièvre-Berna, E., Falus, P., Bentley, P.M., Moskvin, E., Grigoriev, S., Fouquet, P., Farago, B.: Chiral paramagnetic skyrmion-like phase in MnSi. Phys. Rev. Lett. **102**, 197202 (2009)
27. Treusch, R., Feldhaus, J.: FLASH: new opportunities for (time-resolved) coherent imaging of nanostructures. New J. Phys. **12**, 035015 (2010)
28. Hura, G., Menon, A.L., Hammel, M., Rambo, R.P., Poole, F.L., Tsutakawa, S.E., Jenney, F.E., Classen, S., Frankel, K.A., Hopkins, R.C., Yang, S., Scott, J.W., Dillard, B.D., Adams, M.W.W., Tainer, J.A.: Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nat. Methods **6**, 606–614 (2009)

Small-Angle X-Ray Scattering

▶ Small-Angle Scattering of Nanostructures and Nanomaterials

Smart Carbon Nanotube-Polymer Composites

▶ Active Carbon Nanotube-Polymer Composites

Smart Drug Delivery Microchips

▶ Stimuli-Responsive Drug Delivery Microchips

Smart Hydrogels

Alessandro Parodi, S. M. Khaled, Iman K. Yazdi, Michael Evangelopoulos, Naama E. Toledano Furman, Xin Wang, Federico Urzi, Sarah Hmaidan, Kelly A. Hartman and Ennio Tasciotti
Department of Nanomedicine, Houston Methodist Research Institute, Houston, TX, USA

Synonyms

Stimulus-responsive polymeric hydrogels

Definition

Smart hydrogel is defined as the polymer network able to respond to external stimuli through abrupt changes in the physical nature of the network.

Polymer Science in Medicine

The first application of polymers in the medical field can date back to the 1940s when polymethylmethacrylate (PMMA) was used for the replacement of damaged corneas. Since then, the mechanical, physical, and chemical properties of polymers have been extensively investigated and utilized for numerous medical applications. In particular, regenerative medicine has benefited greatly from polymer research and development.

Polymers can now replace metal devices used in orthopedic settings, and the investigation into the development of new biomaterials to repair and substitute body tissues is proceeding with great momentum. Today, the use of polymers has an extensive array of applications in medicine. They have a major role in replacing damaged bones, increasing efficacy of wound repair, and fabricating external devices for dialysis, heart valves, vascular grafts, prosthesis, implantable lenses and dental materials [1]. Nevertheless, polymer science still remains far from perfectly mimicking nature's ability to engineer biomacromolecule conjugates in terms of their structure, versatility, adaptability, and synthetic processes.

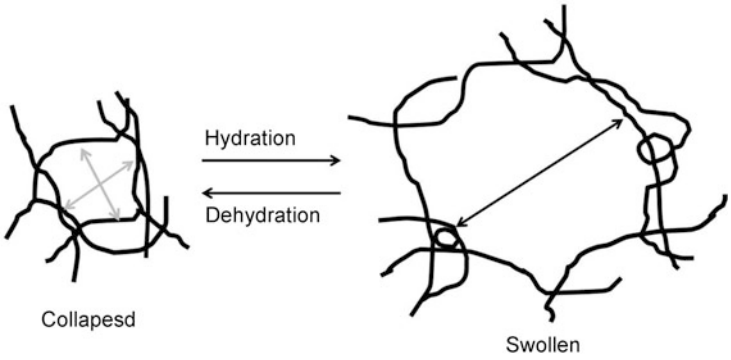
In 1975, Ringsdorf introduced the concept of a "pharmacologically active polymer," specifically referring to the chemical conjugation between a polymer and a drug, which represented a revolutionary approach in drug delivery [2]. Abuchowski et al. were the first to show the conjugation of PEG (polyethylene glycol) to a protein (bovine serum albumin) in 1977 [3]. The resulting conjugate was found to be less immunogenic as the PEG exerted its shielding effects on the incorporated protein. Since then, a variety of polymer-biomacromolecule conjugates have been developed and applied in a wide range of areas including bioseparation, drug/siRNA delivery, enzymatic catalytic processes, diagnostics and biosensing, and cell culture processes. Effective synthesis techniques are based on free radical polymerization. A more advanced method is the living radical polymerization (LRP), recently termed as reversible-deactivation radical polymerization [4]. This radical polymerization is aimed to reduce termination reactions (in which two active chains react with each other to become inactive), thus providing the ability to control the final molecular weight and molecular weight distribution. It also permits polymer chains to be extended as needed with the addition of monomers once the dedicated feed is done. In this respect, LRP is highly useful in producing functional "smart" polymers, as will be discussed later [5].

Construction of responsive polymer bioconjugates is usually achieved by one of three methods: "grafting-to," "grafting-from," and "grafting-through." The most common method is "grafting-to" which involves the covalent or noncovalent attachment of a preformed polymer to a biomacromolecule by reactive coupling or affinity [6]. "Grafting- from" is a more recent method in which the biomacromolecule is functionalized with a moiety capable of initiating polymerization of monomers in solution. Lastly, "grafting-through" involves the attachment of the biomacromolecule to the monomer before polymerization, the generated conjugates potentially having multiple biological species attached along the polymer backbones [7].

Another aspect to be considered is the bioproperties of the polymeric chain itself. Biodegradable polymers were the first type of polymers intended for drug delivery because of the wide range of payloads that can be easily loaded and released during the degradation of the polymers. This class of polymers has been extensively studied over the last decades with successful examples such as calcium alginate and PEG-based polymer networks [8]. In addition, the advancement of nanomedicine and of its biomedical applications sparked a new breakthrough in the use of polymers for drug delivery. This application of polymers to drug delivery has been classified into five main categories: polymeric drugs, polymer-drug conjugates, polymer-protein conjugates, polymeric micelles, and polynucleic acids [9]. The improvements in synthesis, surface modifications, and characterization techniques have transformed polymers into ideal partners for many of the nanoparticles currently used in drug delivery. The fusion of these two sciences enables nanovectors to exhibit enhanced bioavailability and specificity with respect to the targeted biological sites.

Introduction of Hydrogel in Medical Field

In order to be used in the medical field, polymers must meet strict biocompatibility criteria. They have to exhibit low to no toxic effects, be hydrosoluble, dissolve into safe by-products in



Smart Hydrogels, Fig. 1 Swelling of hydrogel. *Left*: Hydrogel in the unswollen or dry state. The polymer chains are in close proximity and may interact with each other. However, when fluid enters the hydrogel, the polymer chains undergo hydration that leads to swelling due to this hydrophobic or electrostatic interaction. *Light gray arrows* indicate pressure being exerted on the system that leads the swelling. *Right*: The hydrogel in the swollen or hydrated state where the polymer chains are fully extended and only the crosslinks prevent the material from dissolution

aqueous environments, and be nonimmunogenic to avoid activating the inflammatory response [10]. Hydrogels, also defined as hydrophilic three-dimensional polymeric networks, represent a promising option in this regard. Featured with its biocompatibility to the surrounding environment and its biodegradability, hydrogels allow for the introduction of new aspects and possibilities in the medical field. The surface of hydrogels can be reconstructed and engineered into scaffolds for the growth of synthetic tissues (pancreas, cornea, skin, bones) [11]. These materials are advanced in regenerative medicine because of their improved tolerance in the recipients for implantation and their suitability for soluble injection.

Hydrogels also provide a venue for controlled and localized drug loading, transportation, and release [12]. Recently, the applications of hydrogels have been expanded. For example, a hydrogel with embedded target biomolecules can be used as biosensor for small-scale analyte detection in diagnosis. They can be reprogrammed into tunable or self-healing coatings to interact with different microenvironments. They can also provide the opportunity for dynamic control of permeation of chemicals and drugs or serve as carriers for catalytic nanoparticles and enzymes in chemical and biochemical catalysis. Finally, hydrogels can be actuated to mimic the action of muscle in the field of biomimetic medicine.

Composed of hydrophilic cross-linked polymers, hydrogels do not dissolve in water and can hold a large amount of water or biological fluid while maintaining their structure. The ability of hydrogels to absorb water arises from the hydrophilic functional groups attached to their polymeric backbone. Their resistance to dissolution arises from cross-linking between the network chains. Figure 1 represents the polymeric network structure of a typical hydrogel interacting with the aqueous phase. While in a dehydrated or deswollen state, its polymer chains are in close proximity to each other, leaving little room for the diffusion of molecules. As the material swells, the polymer chains separate to an extent determined by the properties of the solvent. During the swelling process, the polymer chains of the hydrogel extend, decreasing the interaction that takes place between them. In this state, the swelling of the hydrogel is counteracted by the cross-linkers present within the hydrogel matrix. At its maximum hydration, the diffusion of small molecules (e.g., drugs) approaches the diffusion coefficient in pure fluid.

Smart Hydrogels

Hydrogels are characterized by a dynamic balance with water or other biological fluids that drives the absorption and release of drugs in the desired

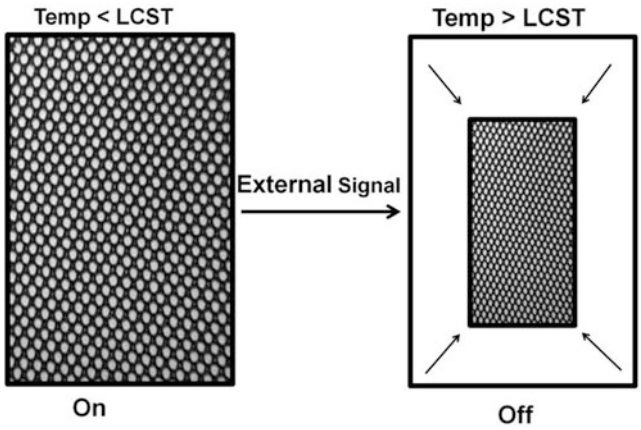
environment. The presence of water in the hydrogel matrix determines the physical and chemical properties of these polymeric structures in the drug delivery field and increases their biocompatibility. Hydrogels represent a very interesting tool in regenerative medicine because they reduce the probability of inflammatory responses due to mechanical frictions with the surrounding tissues in implant surgeries. In addition, these matrices can be formulated in injectable solutions, which represent an appreciable advantage over the conventional surgical procedures.

During the last decades, hydrogels gained increasing interest in medical field because they can be designed or tailored to undergo discrete or continuous volume transformation in response to infinitesimal changes of environmental stimuli such as pH, temperature, electric field, solvent composition, salt concentrations/ionic strength, light/photon, pressure, and coupled magnetic or electric fields. Better known as smart or stimuli-responsive hydrogels, these highly maneuverable and adaptive polymers can sense changes in the environment and respond by inducing structural changes (increasing or decreasing their degree of swelling) without requiring an external driving force. This "volume-changing" phenomenon is particularly useful in drug delivery applications as drug release can be triggered upon these environmental changes.

Smart hydrogels are ideal candidates for the development of self-regulated drug delivery systems with enhanced therapeutic efficacy. Temperature and pH are the most commonly used stimuli to trigger the hydrogel's curative action because they have biological and physiological relevance [13]. The ability for hydrogels to swell as a result of external environmental factors relies largely on three forces: (1) polymer and solvent interaction; (2) polymer elasticity; and (3) ion osmotic force, commonly referred to as electrostatic forces [14]. In the first case, the interaction with the solvent results in swelling due predominantly to the increase in translational entropy, a result of the solvent's molecules invading the polymer network. The second force, polymer elasticity, is determined by the elastically active chains intertwining the crosslinks within the polymer network. Specifically, this is determined by the mean number of chains per crosslink and the crosslink density increasing with temperature. While early models developed by Flory and Rehner were capable of predicting polymer swelling based on these two forces, a third force requires additional tuning [15]. In the presence of ionized groups, the third force (i.e., ion osmotic force or electrostatic force) must be considered. Similar to the Gibbs-Donnan effect, in order for electroneutrality to be enforced everywhere except at the polymer interface, counterions must be drawn into the hydrogel while coions are simultaneously released.

Temperature-Sensitive Hydrogels

Temperature-sensitive hydrogels are probably the most universally studied class of environmentally sensitive polymer systems in drug delivery research. These hydrogels are able to swell or collapse as a result of temperature fluctuation in surrounding fluid and are classified into negatively and positively thermosensitive hydrogels. Thermoresponsive hydrogels are composed of polymer chains that possess either moderately hydrophobic groups, such as methyl, ethyl, and propyl, or a mixture of hydrophilic and hydrophobic segments. Negative temperature-sensitive hydrogels are characterized by their lower critical solution temperature (LCST). Below the LCST, the hydrogel swells in solution, and above this temperature the polymer contracts. On the other hand, positive temperature-sensitive hydrogels are characterized by their upper critical solution temperature (UCST). Below UCST, the polymer shrinks, and above UCST it swells. As a result of adjusting the critical solution temperature within a physiological range, these hydrogels acquire intriguing biological activities that make them reliable candidates for drug delivery in environmental conditions in which the temperature varies. Polymeric materials such as poly(*N*-isopropyl acrylamide (PNIPAM) or poly (γ-2-(2-(2-methoxyethoxy)-ethoxy)ethoxy-ε-caprolactone)-b-poly(γ-octyloxy-ε-caprolactone) have been preferred as block copolymers for



Smart Hydrogels, Fig. 2 Schematic illustration of the "on-off release" from a squeezing hydrogel device for drug delivery (Adapted from Ref. [17])

thermosensitive nanocarriers demonstrating marked transition temperatures, allowing improved drug release at low hyperthermia. It can also occur during a brief temperature decrease (also called cold shock or cryotherapy). In this case, a thermally reversible swelling or deswelling of the block copolymers such as Pluronic F127–polyethyleneimine (PEI) leads to diffusion of the encapsulated small interfering RNA (siRNA) delivery into the cytosol [16].

Under the critical point of negative temperature-sensitive hydrogels, hydrogen bonding between hydrophilic segments of the polymer chain and water molecules dominates, leading to enhanced dissolution in water. As the temperature increases, however, interactions among hydrophobic segments are strengthened, while hydrogen bonding weakens. The net result is deswelling of the hydrogels due to interpolymer chain association through hydrophobic interactions [13]. The LCST can be changed by adjusting the ratio of hydrophilic and hydrophobic segments in the hydrogel system. It can also be done by developing copolymers of hydrophobic (e.g., *N*-isopropylacrylamide (NIPAAm)) and hydrophilic (e.g., acrylic acid (AA)) monomers. Hydrogels with negative temperature sensitivity have the potential to show an "on-off" drug release with "on" at low temperature and "off" at high temperature allowing pulsatile drug release. Figure 2 shows a schematic illustration of the on-off release [17].

Polymers having LCST below human body temperature have the potential for injectable applications and are often combined with polysaccharides such as chitosan, alginate, cellulose, and dextran [18]. This can improve the material's properties in terms of biocompatibility, swelling ratio, pH-sensitivity, swelling dynamics (swelling and deswelling rate), modulation of LCST, and thermal stability. Many injectable hydrogels are based on this particular feature as they remain in the Newtonian fluid form at room temperature and shrink once injected into the patient's body. This allows for a prolonged delivery of loaded drugs, antineoplastic or analgesic drugs being just a few examples of the many possibilities. For example, Poly(*N*-isopropylacrylamide) (PNIPAAm)-based hydrogels have a great potential for biomedical applications, especially for in vivo *drug delivery.* PNIPAAm's LCST (32 °C) is in the physiological range and above the temperature of an operating room (e.g., body temperature). These hydrogels can be obtained with a wide range of crosslinking methods, and the desired features for the polymer can be easily tuned by modifying the chemical formulation of the hydrogel. For example, the LCST of PNIPAAm can be changed by adding different comonomers which will increase LCST and vice versa.

On the other hand, this polymer is not fully biocompatible as it could activate a platelet coagulation cascade. In order to overcome this limitation, the grafting of PNIPAAm with PEG (triblock copolymer) allows for the formulation of a biocompatible and biodegradable hydrogel to deliver the drug in situ for several hours. In addition, the

features of this copolymer can be tuned in function of the molecular weight of the PEG (up to a limit of MW 10,000). This parameter can be changed in order to modify the porosity of the hydrogel, where a higher MW causes the hydrogel to have a limited number of bigger pores because the high MW of the PEG limits the mobility of PNIPAAm [19].

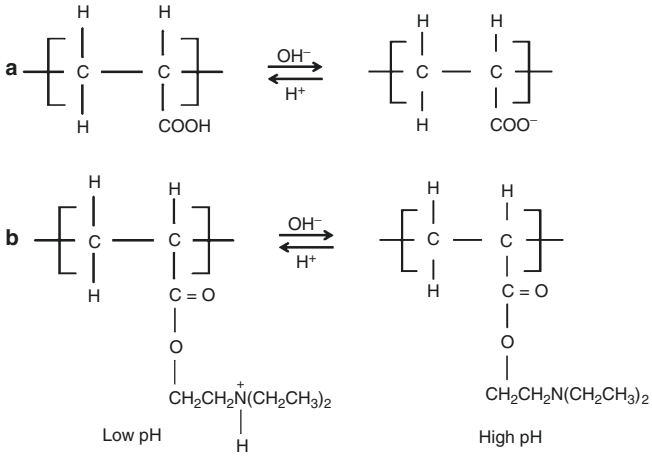
Another copolymer worth mentioning is the triblock PLGA-PEG-PLGA, which showed thermosensitive biodegradability and biocompatible properties. This copolymer has been used to deliver a wide range of drugs such as insulin, calcitonin, porcine growth hormone, and testosterone, and its use can be applicable to the delivery of hydrophilic and hydrophobic molecules. The release of the drug in the body is due to both drug diffusion and the degradation of the PLGA releasing the incorporated molecules, where hydrophobic molecules are mainly released by the degradation of the copolymer and are "delivered" over a longer temporal window. The features of this class of polymers can be modified by changing some variables such as the molecular weight of PEG and PLGA, as well as the LA/GA ratio. The applications of these kinds of hydrogels range from drug delivery (cancerous, ocular, arteriovenous) to tissue engineering. Another issue of the thermoresponsive polymers is represented by temperature-sensitive peptides often associated with the widely investigated Elastin-like Polypeptide (ELP) polymers. This polymer is composed of pentapeptide monomers, derived from human tropoelastin [20], which tend to aggregate in a reversible way under heating, entrapping the payload. This biopolymer showed extraordinary biological properties as it is fully biocompatible, completely biodegradable without any immunogenic reaction, and it appeared to be ideal for the delivery of protein-based treatments. The degradation of the polymer is constant and slow when compared to the free form of the monomer.

The temperature-sensitive peptides represented a breakthrough in polymer science in the biomedical field. The exploitation of standard recombinant DNA techniques via genetic engineering allowed them to be easily synthesized and modified in the primary sequence. This changed the chemical, physical, and biological properties of the polymer while permitting a controlled design of the peptide, according to the nature of the cargo. Furthermore, they could easily be attached to a protein-based payload during the synthesis process.

The two main features of ELP polymers are directionality and reversibility. Directionality describes the conformational state of the polymer under a specific stimulus. The ability of the peptide to associate or disassociate under heating is essential if the treatment requires rapid or prolonged drug release. Reversibility indicates the capacity of the polymer to return to its initial state after the thermal stimulus [13]. A reversible polymer could be very useful for concentrating the payload in a region of the body and obtaining a targeted release of the drug after heating through an external stimulus. Conversely, an irreversible polymer could allow drug delivery in a controlled fashion through slow degradation. This category of thermosensitive polymers is emerging as a new opportunity in cancer cell therapy [13], and future research aims to increase and optimize the biocompatibility and the tenability of these delivery systems. The challenge in the design of these thermoresponsive polymeric delivery systems lies in the use of materials that are both safe and sensitive enough to respond to slight temperature changes around the physiological temperature.

pH-Sensitive Hydrogels

Another class of smart hydrogels undergoes a volume transition with the variation of the environmental pH. pH variations have been exploited to control the delivery of molecules to intracellular compartments (such as endosomes or lysosomes) or specific organs (such as the gastrointestinal tract or the vagina) to trigger the release when subtle environmental changes are associated with pathological conditions such as cancer or inflammation. Polyacid or polybase hydrogels with ionizable groups and the hydrogels with acid-sensitive bonds whose cleavage enables the release of molecules anchored at



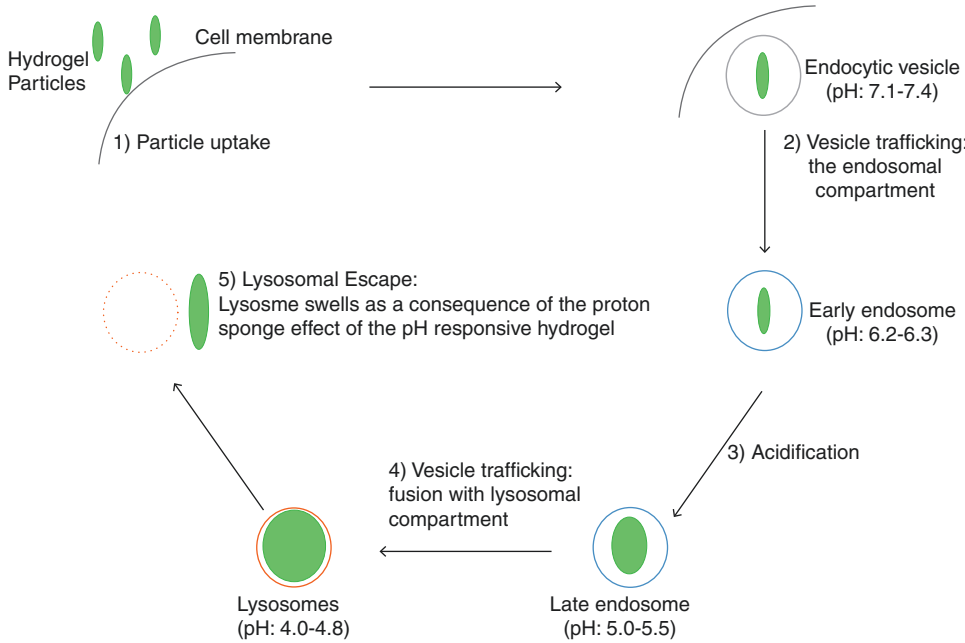
Smart Hydrogels, Fig. 3 pH–dependent ionization of hydrogels. (**a**) Poly (acrylic acid) and (**b**) poly (*N*, *N*/-diethylamionethyl methacrylate)

the backbone are the two strategies used to utilized response to environmental pH variation [21].

These hydrogels are characterized by the presence of ionic moieties on the polymeric backbone. These ions either accept or release protons in response to pH changes. Weak acidic groups such as acrylic and cationic acids or weak basic groups such as amines provide pH sensitivity to the polymer chain. Swelling of hydrogels sharply changes in the vicinity of the pKA or pKB values of the acidic or basic functional groups. Carboxylic pendant groups accept hydrogen at low pH but exchange it for other cations above the pKA value, and become ionized at higher pH. The hydrodynamic volume and swelling capability of these polymer chains increases sharply when their carboxylic groups become ionized and reach a plateau around pH 7. Amines, on the other hand, accept protons at low pH and become positively ionized at and below their pKB values. Hence, their swelling capability increases sharply in acidic solutions. Figure 3 shows structures of anionic and cationic polyelectrolytes and their pH-dependent ionization. Poly(acrylic acid) (PAA) becomes ionized at high pH, while poly(*N*,*N*-diethylaminoethyl methacrylate) (PDEAEM) becomes ionized at low pH.

The swelling and pH-responsiveness of polyelectrolyte hydrogels can be adjusted by using neutral comonomers, such as 2-hydroxyethyl methacrylate, methyl methacrylate, and maleic anhydride [13]. Different comonomers provide different hydrophobicity to the polymer chain, leading to different pH-sensitive behavior. pH-responsive polymers were first applied as oral drug delivery systems due to the variation of the pH that characterizes the different areas of the gastrointestinal tract. Several anionic polymers such as methyl acrylic acid, methyl methacrylate, and hydroxypropyl methylcellulose phthalate have been commercially used as enteric coatings for oral delivery of protein-based drugs. In the acidic environment of the stomach, the carboxylic acid groups of the aforementioned polymers are unionized, and the particle retains its therapeutic cargo. They were shown to release the encapsulated drug molecules into the small intestine in response to the alkaline pH, because the carboxylic acid group becomes ionized and the particle swells. PNIPAAm can be classified as a thermal and pH-responsive polymer. It is used for oral drug delivery applications, exploiting its degradation characteristics in the alkaline environment of the intestine. It was shown that the polymer synthesized in combination with buthylmethacrylate (BMA) and acrylic acid (AAc) is very stable at low pH, while resulting in full degradation and release of payload within the high pH environment [22].

Vesicles derived from endocytosis fuse their membranes with vesicles of the early endosomal compartment. Lysosome formation follows after several other membrane fusion steps (with late endosomal and lysosomal vesicles). During



Smart Hydrogels, Fig. 4 Schematic of the lysosomal escape. Upon cell internalization by endocytosis, hydrogel particles are trapped in the endolysosomal compartment. During their transition from early to late endosomes and finally to lysosomes, particles undergo increased swelling due to protonation. The proton sponge effect they exert eventually leads to the breakage of the lysosomal vesicle

these membrane fusions inside the cytoplasm, the vesicles are proton rich, decreasing their internal pH through an active import of H^+. This proton enrichment is permitted by an ATPase that works using ATP to pump in hydrogen ions and a negative counterion (Cl^-) to satisfy electroneutrality. Due to the presence of many ionizable groups in the backbone, a pH-responsive polymer is usually able to work as a "proton sponge" with a buffering effect on the surrounding environment. Therefore, the ATPase continues to work with the result of increasing the concentration of the counterion and affecting the osmotic balance of the vesicles. The final result of this process is the breaking of the vesicle's membrane, caused by an increase in osmotic pressure.

Many polymers increase their volume when subjected to an acidic environment. This produces a mechanical stress on the lysosomal membranes, affecting the integrity of these organelles. Polymers that exhibit this behavior show a proton-sponge effect, and lysosomal escape is very useful for nonviral gene delivery and could potentially be more efficient and less toxic than standard virus-based carriers. These molecules are usually cationic and therefore likely to be efficient in forming macrocomplexes with DNA (polyplexes) or creating new therapies based on RNA (e.g., siRNA, miRNA). Figure 4 represents a schematic of lysosomal escape performed by a pH-responsive cationic hydrogel. The polymer in this case plays three roles: (1) to physically carry the plasmid, (2) to shield the DNA payload from degradation, and (3) to perform lysosomal escape and facilitate the nuclear delivery of the cargo. Poly(ethylene imine) (PEI), poly(amidoamine) (PAMAM), Poly (*N*,*N*-dimethylaminoethylmethacrylate) (PDEAEM), poly (L-Lysine), and modified chitosan represent a few examples of all the polymers studied for this application.

Thermo- and pH-responsive hydrogels allow (with or without the use of copolymers) the loading and the triggered release of many drugs for the treatment of different pathological states. Table 1 reports some examples of these polymer-drug formulations and their potential clinical applications.

Smart Hydrogels, Table 1 Some applications of temperature and pH-responsive polymer

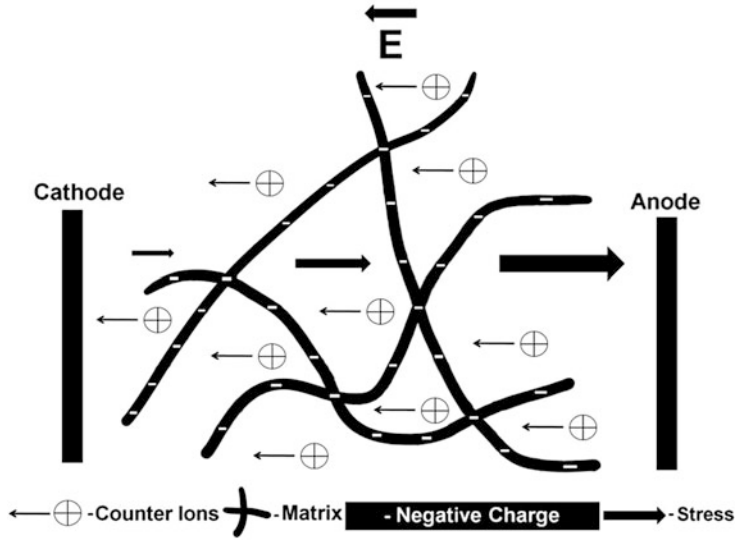
Polymer	Enviromental stimulus	Copolymer	Loaded drug	Application
PNIPAAm	Temperature	Chitosan	Diclofenac	Anti-inflammatory
Glycidyl methacrylate	Temperature	Chitosan	Doxorubicin	Chemotherapeutic
β-glycerophosphate	Temperature	Chitosan	Chitosan	Antibiotic
Poly(propylene)-g-AA-g	Temperature	NIPAAm/ Chitosan	Chitosan	Antibiotic
Triblock polymer	Temperature	Poly(lactide-co-glycolide, polylactic acid, poly(ethylene glycol)	Adriamycin	Chemotherapeutic
PNIPAAm	Temperature	Alginate	Indomethacin	Anti-inflammatory
PNIPAAm	pH	Acrylic acid	Growth factors	Wound repair
Poly(ethylene imine)	pH		Nucleic acid	Gene delivery
poly(amidoamine)	pH		Nucleic acid	Gene delivery
Poly (*N*,*N*-dimethylaminoethylmethacrylate)	pH		Nucleic acid	Gene delivery
poly (L-Lysine)	pH		Nucleic acid	Gene delivery

Other Responsive Compounds

Recently, electroresponsive hydrogels stimulated the curiosity of the scientific community because they can unswell or bend, depending on the shape and orientation of the gel along an electric field. The gel bends when it is parallel to the electrodes, whereas unswelling occurs when the hydrogel lies perpendicular to them. This kind of polymer is usually based on a polyelectrolyte matrix (polymers that contain a relatively high concentration of ionizable groups along the backbone chain) so these polymers are also pH responsive. Weak electric fields can be used to achieve pulsed or sustained release of drug molecules through various mechanisms. For example, polypyrrole hydrogel exhibited tailored drug release as a result of a synergistic process of electrochemical reduction–oxidation and electric-field-driven movement of drug molecules [23].

The electric field generates an interaction between the mobile counterions and the immobile charged groups of the gel's polymeric network in order to induce the phase transition. For example, in partially hydrolyzed polyacrylamide gels, the mobile H+ ions migrate toward the cathode while the negatively charged immobile acrylate groups in the polymer networks are attracted toward the anode (Fig. 5). Thus, the anionic polymeric gel network is pulled toward the anode. This pull creates a uniaxial stress along the gel axis, being maximal at the anode and minimal at the cathode. This stress gradient contributes to the gel deformation, and when triggered by an electric filed, these hydrogels undergo the swelling and contraction necessary to release the payload. Electroresponsive polymers represent one of the most complex challenges in biomedical polymer synthesis. Even if the electric stimulus can be finely tuned in terms of strength, duration, and impulse frequency, the use of these techniques is limited by objective safety restrictions. This kind of hydrogel was conceived to obtain a pulsatile delivery of hormones (e.g., insulin, hydrocortisone), proteins and peptides, and glucose. A few clinical applications for electroresponsive polymers are currently being investigated for dermal and transdermal drug delivery [24].

Ultrasound-triggered hydrogel-based delivery system represents a noninvasive effective approach to achieve spatiotemporal control of drug release at the desired site while preventing



Smart Hydrogels, Fig. 5 The effect of an electric field on a polyelectrolyte gel. Positively charged counterions migrate toward the cathode, while the immobile polymeric anionic groups are attracted toward the anodes (Adapted from Ref. [34])

any possible side effects to surrounding tissues with flexibility to regulate the tissue penetration depth by tuning frequency, duty cycles, and time of exposure. Ultrasound waves trigger the release of the drug molecule through the thermal and/or mechanical effects generated by cavitation phenomena or radiation forces [25]. Combination of thermosensitive hydrogels with high-intensity focused ultrasound waves enables triggered drug release to obtain high delivery efficiency with only a mild temperature increase.

Stimulus-Sensitive Connection

This approach allows not only for the development of drug or gene delivery tools but also adaptive structures, self-healing materials, and sensing detectors that are responsive to the composition of the biological microenvironment [26]. These materials are often designed to achieve additive conformational changes of the individual subunits producing a coherent mechanical response to an external stimulus.

Stimuli can include heat, stress, pressure, electrical current/voltage, magnetic field, pH change in solvent, water, or moisture, light (photoresponsive materials), and ultrasound-responsive materials [27]. In the case of biomacromolecule conjugates, the stimuli may confer the responsive nature of the following polymer by the ability to perform fine tuning of the solubility, stability, and/or bioactivity of the resulting conjugate. This new generation of polymer bioconjugates exhibits properties and activities that can be controlled and tuned rather than simply enhanced [28].

In this case, the polymer's function is limited to being a carrier for the cargo and targeting the area with the correct environmental conditions for drug release. Polymers such as *N*-(2-hydroxypropyl) methacrylate (HPMA), already exploited for steric stabilization of bioactive proteins, were used to coat biological or synthetic particles to provide protection against antibodies and complement factors. HPMA was also used to bind doxorubicin via a pH-sensitive linker. This polymer-bound drug showed significant improvement in the ability to kill MCF-7 breast cancer cells when compared with the drug alone. Another example of drug delivery system based on a peptide linker (Pro-Val-Gly-Leu-Ile-Gly) was composed of a chemotherapeutic agent and a dextran-based polymer. This linker is sensitive to the action of the overexpressed cancer metalloproteinases 2 and 9, while remaining stable in blood circulation. Therefore, the drug release is more concentrated at the site of the tumor where these enzymes are usually overexpressed [29].

A new frontier of material science is aiming to develop self-healing materials with an elongated lifespan through the formation of supramolecular hydrogels designed in a host-guest interaction pattern. The selective complementary interactions between host and guest molecules are versatile and can be used for the preparation of self-healing/self-repairing polymers. They can be used as both architectural materials and external coatings. The sol–gel phase transition can occur through different stimuli such as heat, pH, light, or redox environment. Multipoint crosslinks play an important role in forming the supramolecular hydrogel: usually their structure is formed by a main chain containing host molecules with a sufficient length and an appropriate number of guest molecules. For example, Cyclodextrin (CD) is used as a host molecule because it has a diverse set of applications while ferrocene (Fc) is used as a guest molecule. Polyacrylic acid (pAA) possessing CD and pAA possessing Fc are mixed together to form a supramolecular hydrogel. Fc derivatives have redox-responsive properties and variations and could induce a reversible sol–gel phase transition in a supramolecular hydrogel, resulting in a hydrogel with self-healing properties such as readhesion between cut surfaces. There are several research groups attempting to use them for a variety of different applications such as artificial molecular muscles, drug delivery systems, and vascular embolization.

Recently, stimuli-responsive polymers exhibited an ability to change their shape in reaction to a particular right stimulus. When the shape change is accompanied by the stimulus, this is the shape change material (SCM) [28]. Elastic rubber, electroactive polymer, piezoelectric material, and liquid crystal are few examples of such SCMs. Other types of stimuli-responsive materials demonstrating the ability to actively changing shape can be differentiated into shape memory materials (SMM). Although massive progress has been made in the past several years, SCMs and SMMs are starting to show their potential for different bio-based applications, although the research in this area is far from maturity.

Sensing and Signal Transduction by Swelling and Deswelling

As previously discussed, hydrogels have the potential to provide many beneficial properties in the field of drug delivery. For example, by sensing changes in hydrogel volume or weight, it is possible to measure enzymatic conversions or binding of solutes among other factors. For instance, this can be particularly beneficial in monitoring glucose levels in the tear film of the eye. This is accomplished by placing polymerized crystalline colloid array hydrogels in the front of the eye.

Unfortunately, measuring the changes in volume of such a small hydrogel can pose some challenges. One solution to this issue is to place a small hydrogel onto the tip of a fiber optic that is connected to a light source. This alternative method is advantageous because of its retention and ability to accurately measure blood metabolites. Other techniques devised by Peppas et al. and Siegel et al. involve the use of a micromachined cantilever beam with pH-sensitive hydrogel polymerized either at the top or base of the beam, measuring the deformation of the beam that occurs upon swelling of the hydrogel [8, 30].

Detection of small molecules through implantable sensors requires the consideration of additional factors such as exclusion of large molecules and isolation from the host tissue. Previously, this was accomplished by confining the hydrogel between a rigid, semipermeable membrane and a transducer. By measuring the pressure buildup from the resulting confinement, it was possible to distinguish analyte-induced changes within the hydrogel. Other techniques have been employed to measure changes in pH, glucose levels, and ionic strength [31]. Nevertheless, the confinement of a hydrogel into a tight space is not always easy, and complications can occur. For example, if the hydrogel is too small, the empty space could make it difficult to measure the pressure changes that occur upon swelling. Conversely, stress imposed when the hydrogel is too large can also make it difficult to measure pressure fluctuations. To overcome this issue, one group embedded superparamagnetic iron oxide nanoparticles into the hydrogel network [32]. After lamination onto a planar coil,

the swelling of the hydrogel can be measured by quantifying the magnetic permeability following stimulation with radio frequency.

Although many beneficial measurements can be obtained by evaluating fluctuations in weight and volume, strategically using the swelling and deswelling effects for drug release can also prove useful. Specifically, the expansion in mesh size resulting from swelling has previously been shown to provide favorable drug release properties [33]. Similarly, strong contractions in the hydrogel network can also provide increased release properties. Together, these features have resulted in a multitude of delivery vectors and continue to transform the field of nanotechnology.

Summary

Smart hydrogels, a special class of polymer networks, have offered new breakthroughs in the medical field due to their exceptional properties, engineering flexibility, natural abundance, and ease of manufacturing. This entry represents a basic overview of classifications and material properties of smart hydrogels for guided drug delivery and applications in regenerative medicine. While providing a brief representation on the current trends and advancement of polymer science in medicine, we have also revealed some cutting-edge progress in the field originating from the blending of nanotechnology with polymer science.

Cross-References

- ▶ Drug Delivery
- ▶ Nanocomposites
- ▶ Nanomedicine
- ▶ Nanoparticles

References

1. Navarro, M., Michiardi, A., Castano, O., Planell, J.: Biomaterials in orthopaedics. J. R. Soc. Interface **5**, 1137–1158 (2008)
2. Ringsdorf, H.: Structure and properties of pharmacologically active polymers. J. Polym. Sci. Polym. Symp. **51**, 135–153 (1975)
3. Abuchowski, A., McCoy, J.R., Palczuk, N.C., van Es, T., Davis, F.F.: Effect of covalent attachment of polyethylene glycol on immunogenicity and circulating life of bovine liver catalase. J. Biol. Chem. **252**, 3582–3586 (1977)
4. Jenkins, A.D., Jones, R.G., Moad, G.: Terminology for reversible-deactivation radical polymerization previously called“ controlled” radical or“ living” radical polymerization (IUPAC recommendations 2010). Pure Appl. Chem. **82**, 483–491 (2009)
5. Cobo, I., Li, M., Sumerlin, B.S., Perrier, S.: Smart hybrid materials by conjugation of responsive polymers to biomacromolecules. Nat. Mater. **14**, 143–159 (2015)
6. Lutz, J.-F., Börner, H.G.: Modern trends in polymer bioconjugates design. Prog. Polym. Sci. **33**, 1–39 (2008)
7. Ivanov, A.E., Edink, E., Kumar, A., Galaev, I.Y., Arendsen, A.F., Bruggink, A., et al.: Conjugation of penicillin acylase with the reactive copolymer of *N*-Isopropylacrylamide: a step toward a thermosensitive industrial biocatalyst. Biotechnol. Prog. **19**, 1167–1175 (2003)
8. Keys, K.B., Andreopoulos, F.M., Peppas, N.A.: Poly (ethylene glycol) star polymer hydrogels. Macromolecules **31**, 8149–8156 (1998)
9. Schmaljohann, D.: Thermo- and pH-responsive polymers in drug delivery. Adv. Drug Deliv. Rev. **58**, 1655–1670 (2006)
10. Anderson, J.: The future of biomedical materials. J. Mater. Sci. Mater. Med. **17**, 1025–1028 (2006)
11. Alijotas-Reig, J., Garcia-Gimenez, V.: Delayed immune-mediated adverse effects related to hyaluronic acid and acrylic hydrogel dermal fillers: clinical findings, long-term follow-up and review of the literature. J. Eur. Acad. Dermatol. Venereol. **22**, 150–161 (2008)
12. Peppas, N.A.: Hydrogels and drug delivery. Curr. Opin. Colloid Interface Sci. **2**, 531–537 (1997)
13. Chilkoti, A., Dreher, M.R., Meyer, D.E., Raucher, D.: Targeted drug delivery by thermally responsive polymers. Adv. Drug Deliv. Rev. **54**, 613–630 (2002)
14. Erman, B., Flory, P.J.: Critical phenomena and transitions in swollen polymer networks and in linear macromolecules. Macromolecules **19**, 2342–2353 (1986)
15. Flory, P.J., Rehner, J.: Statistical theory of chain configuration and physical properties of high polymers. Ann. N. Y. Acad. Sci. **44**, 419–429 (1943)
16. Lee, S.H., Choi, S.H., Kim, S.H., Park, T.G.: Thermally sensitive cationic polymer nanocapsules for specific cytosolic delivery and efficient gene silencing of siRNA: swelling induced physical disruption of endosome by cold shock. J. Control. Release **125**, 25–32 (2008)
17. Gutowska, A., Seok Bark, J., Chan Kwon, I., Han Bae, Y., Cha, Y., Wan, K.S.: Squeezing hydrogels for controlled oral drug delivery. J. Control. Release **48**, 141–148 (1997)

18. Webber, R.E., Shull, K.R.: Strain dependence of the viscoelastic properties of alginate hydrogels. Macromolecules **37**, 6153–6160 (2004)
19. Alexander, A., Ajazuddin, Khan, J., Saraf, S., Saraf, S.: Polyethylene glycol (PEG)–poly(*N*-isopropylacrylamide) (PNIPAAm) based thermosensitive injectable hydrogels for biomedical applications. Eur. J. Pharm. Biopharm. **88**, 575–585 (2014)
20. Urry, D.W.: Physical chemistry of biological free energy transduction as demonstrated by elastic protein-based polymers. J. Phys. Chem. B **101**, 11007–11028 (1997)
21. Mura, S., Nicolas, J., Couvreur, P.: Stimuli-responsive nanocarriers for drug delivery. Nat. Mater. **12**, 991–1003 (2013)
22. Serres, A., Baudyš, M., Kim, S.: Temperature and pH-sensitive polymers for human calcitonin delivery. Pharm. Res. **13**, 196–201 (1996)
23. Ge, J., Neofytou, E., Cahill, T.J., Beygui, R.E., Zare, R.N.: Drug release from electric-field-responsive nanoparticles. ACS Nano **6**, 227–233 (2011)
24. Peppas, N.A., Bures, P., Leobandung, W., Ichikawa, H.: Hydrogels in pharmaceutical formulations. Eur. J. Pharm. Biopharm. **50**, 27–46 (2000)
25. Epstein-Barash, H., Orbey, G., Polat, B.E., Ewoldt, R. H., Feshitan, J., Langer, R., et al.: A microcomposite hydrogel for repeated on-demand ultrasound-triggered drug delivery. Biomaterials **31**, 5208–5217 (2010)
26. Yang, Y., Urban, M.W.: Self-healing polymeric materials. Chem. Soc. Rev. **42**, 7446–7467 (2013)
27. Sun, L., Huang, W.M., Ding, Z., Zhao, Y., Wang, C.C., Purnawali, H., et al.: Stimulus-responsive shape memory materials: a review. Mater. Des. **33**, 577–640 (2012)
28. Lu, H.B., Huang, W.M., Yao, Y.T.: Review of chemo-responsive shape change/memory polymers. Pigm. Resin Technol. **42**, 237–246 (2013)
29. Liu, W., MacKay, J.A., Dreher, M.R., Chen, M., McDaniel, J.R., Simnick, A.J., et al.: Injectable intratumoral depot of thermally responsive polypeptide–radionuclide conjugates delays tumor progression in a mouse model. J. Control. Release **144**, 2–9 (2010)
30. Lei, M., Ziaie, B., Nuxoll, E., Iván, K., Noszticzius, Z., Siegel, R.A.: Integration of hydrogels with hard and soft microstructures. J. Nanosci. Nanotechnol. **7**, 780–789 (2007)
31. Lin, G., Chang, S., Hao, H., Tathireddy, P., Orthner, M., Magda, J., et al.: Osmotic swelling pressure response of smart hydrogels suitable for chronically implantable glucose sensors. Sensors Actuators B Chem. **144**, 332–336 (2010)
32. Song, S.H., Park, J.H., Chitnis, G., Siegel, R.A., Ziaie, B.: A wireless chemical sensor featuring iron oxide nanoparticle-embedded hydrogels. Sensors Actuators B Chem. **193**, 925–930 (2014)
33. Lin, C.C., Metters, A.T.: Hydrogels in controlled release formulations: network design and mathematical modeling. Adv. Drug Deliv. Rev. **58**, 1379–1408 (2006)
34. Murdan, S.: Electro-responsive drug delivery from hydrogels. J. Control. Release **92**, 1–17 (2003)

Soft Actuators

▶ Organic Actuators

Soft Lithography

▶ Microcontact Printing
▶ Nanoscale Printing

Soft Matter

▶ Dissipative Particle Dynamics, Overview

Soft X-Ray Lithography

▶ EUV Lithography

Soft X-Ray Microscopy

▶ Selected Synchrotron Radiation Techniques

Soil/Terrestrial Ecosystem/Terrestrial Compartment

▶ Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms

Solar Cells

▶ Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components

Sol–Gel Method

Bakul C. Dave and Sarah B. Lockwood
Department of Chemistry and Biochemistry, Southern Illinois University Carbondale, Carbondale, IL, USA

Synonyms

Chemical solution deposition; Gel chemical synthesis; Silica gel processing; Wet chemical processing

Definition

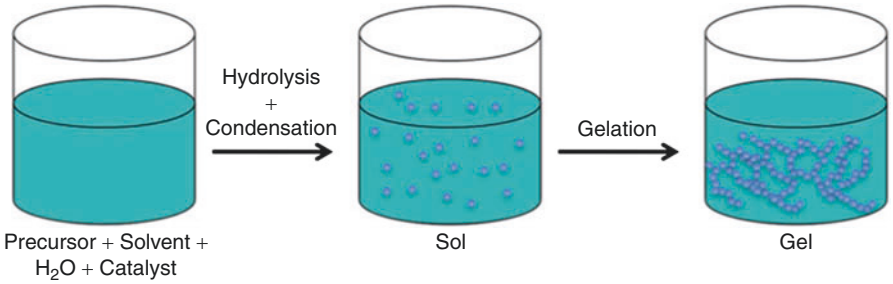
The sol–gel method is a wet chemical process of making oxide-based materials starting from hydrolyzable precursors via hydrolysis and condensation. The precursors usually contain weaker ligands as compared to water such as halides, nitrates, sulfates, alkoxides, or carboxylates. The hydrolyzed precursors then condense together to form small colloidal nanoparticles suspended in a liquid called a sol. Further polycondensation of the sol particles leads to an extended oxo-bridged network of polymeric oxide-based materials. The as-formed gels obtained by the sol–gel method are biphasic materials that contain gel network along with a significant amount of liquid phase. Drying of these gels, either under ambient conditions or at elevated temperature, can lead to expulsion of solvent phase to form dense materials. Because the method utilizes molecular precursors, it offers appealing prospects for the design of nanomaterials with a degree of control over their structure, composition, dimension, morphology, organization, geometry, and bulk architecture. While the method can be used to make bulk monoliths, it is particularly suited for fabricating nanomaterials of varied compositions in different shapes, geometries, and dimensionalities including nanoparticles, fibers, thin films, and coatings.

Overview

The sol–gel method is a solution-based method of making ceramic particles, powders, coatings, films, and monolithic objects [1]. The process is commonly used for making oxides, however, silica-based materials constitute a primary archetype system. With highly electropositive metals, the precursor molecules can react rapidly with water to undergo facile hydrolysis under ambient conditions. However, precursors based on alkoxysilanes usually require the use of catalysts, refluxing, or use of ultrasound for the reaction to proceed at an appreciable rate. The liquid sol containing nanoscale particles can be used to form powders, fibers, coatings, or monolithic forms by allowing the reaction to proceed under specific processing conditions. In addition to the formation of pure oxides, the method can also be used for making organic–inorganic hybrids and organically modified silicas as well as bioceramic hybrids [2].

In the sol–gel method, a suitable molecular precursor is hydrolyzed to generate a solid-state polymeric oxide network (Fig. 1). Initial hydrolysis of the precursor generates a liquid sol, which ultimately turns to a solid, porous gel. The gels formed this way are porous and contain a substantial amount of solvent phase. Slow drying of the gels under ambient conditions (or at elevated temperature) leads to evaporation of the solvent phase. The dried gels – termed xerogels – are substantially less porous than wet gels. Since the method begins from a molecular species, it offers a degree of control over structure and properties that can be tailored at the molecular level. The method can be used to prepare a wide range of oxides. However, the most well-known and extensively studied examples are those based on silica chemistry where there is a good understanding of how the chemical parameters can be used to control the properties of the final product.

In general, the applications of the sol–gel method can be loosely divided into three interdependent aspects of synthesis, processing, and properties, which are a direct function of the structural composition, morphology, and microstructure. The sol–gel



Sol–Gel Method, Fig. 1 Schematic depiction of the sol–gel method

method for synthesis of materials is particularly appealing because with a chemical modification of the precursor at molecular level, the functional properties of the final material can be changed. Therefore, by selectively integrating specific organic functional groups into the precursor at the molecular level, it is possible to introduce desired properties into the product sol–gel inorganic-organic hybrid material.

The sol–gel chemistry is feasible with any metal precursor that is hydrolytically unstable. Typical precursors used are alkoxides, halides, or carboxylates. With silica chemistry, alkoxides [$Si(OR_4)$] are predominantly used. In addition, it is possible to use organically modified alkoxides [$(OR)_3SiR'$] to incorporate organic functionalities in glasses. The R′ group can be any organic group, including polymers or oligomers. When a precursor containing an organic group is hydrolyzed, it leads to formation of an ORganically MOdified SILicate (ORMOSIL) glass. The organic R′ groups can be simple hydrocarbons or can include other functional groups. In addition, it is possible to make glasses with precursors of the type [$(OR)_3Si\text{-}R\text{-}Si[(OR)_3]$], where the organic functionality is present as a spacer group. The precursors can be used alone, or alternatively, a combination of precursors can be used to make hybrid or composite materials. Finally, prior to gelation, other species such as organic-, inorganic-, and biomolecules and/or polymers can be added to the liquid sol to make composite materials. When large biomolecules are added to the glass, they usually become physically entrapped in the glass but nonetheless retain their structure and properties and the resultant materials exhibit biological function and activities [3].

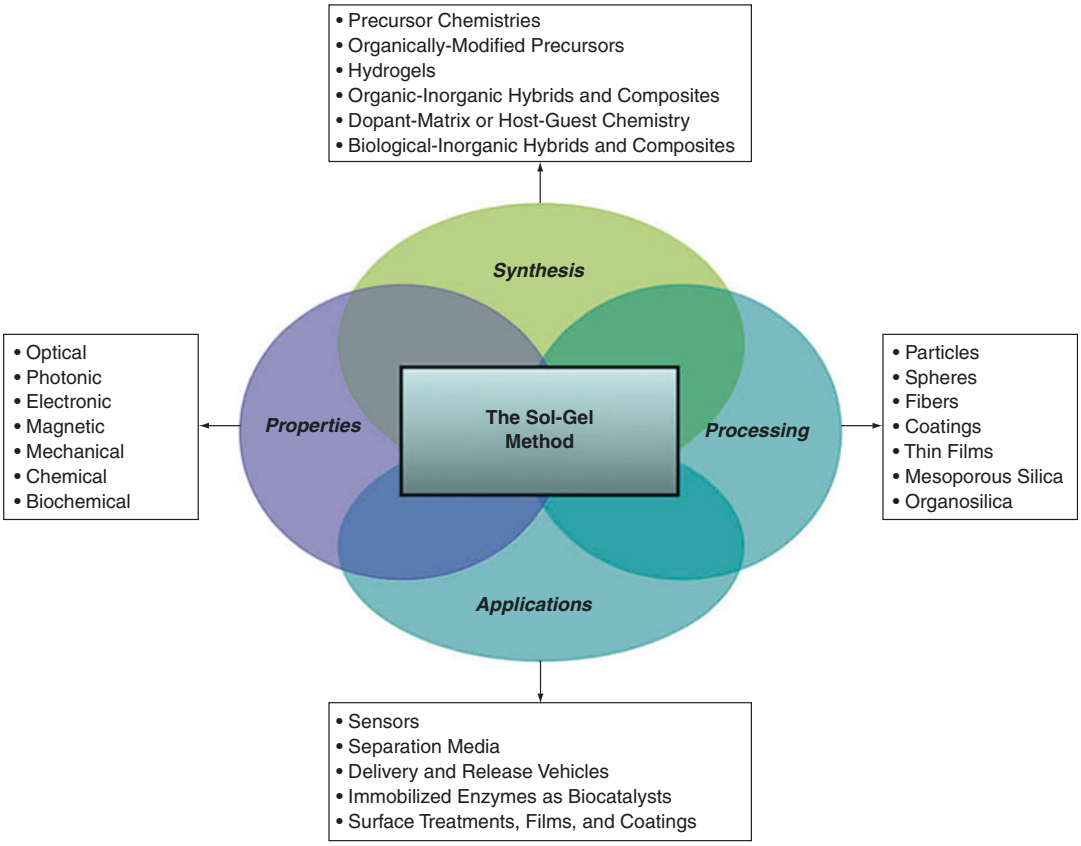
The sol–gel method is primarily a method at the interface of synthesis and processing. Synthesis refers to the making of new materials by means of specialized chemical precursors through arrangement of precise functionalities at the molecular level. On the other hand, processing refers to the specific organization of matter at the nano-, micro-, and macroscopic level. Processing of gels is greatly facilitated by the stability of the liquid sol, which enables construction of materials with good control over their nanostructure, shape, and overall geometry. The final gels can be fabricated in a variety of shapes, geometries, and configurations, including monoliths, coatings, thin films, fibers, and powders. The liquid sol can be used to make monolithic geometries whose final shape is determined by the container used. From the perspective of nanotechnology, the method provides a rather facile pathway to construction of different nanomaterials. Formation of oxide-based nanoparticles is a key area in use of the sol–gel method due to the unique opportunities provided by the solution-based method that proceeds via the sol state characterized by dispersed nanoscale particulate matter. The liquid sol can be used for dip-coating, spin-coating, and spray-coating to prepare films and coatings of varied thicknesses on diverse surfaces including glass, metals, composites, or plastics. The versatility of the method has resulted in a burgeoning utility of the sol–gel method in diverse fields.

Applications are the key drivers in the utilization of the sol–gel method in technology. Because of its versatility, the method has found widespread applications in all walks of science and engineering including, optics, photonics, microelectronics, and biotechnology to name a few. The intrinsic properties of sol–gel materials such as optical transparency, porosity, surface area along with chemical and physical stability confer a unique vantage point for their applications in different areas. Different functions and properties can be imparted to the parent glass by incorporation of additional organic, inorganic, or biological entities. Particularly important in this direction are the organosilica sol-gels and hybrid materials, which integrate the properties of both organic and inorganic materials and constitute a novel platform for development of new materials and devices from a molecular perspective. Similarly, the access to formation of bio-hybrid via integration of biological species makes the method particularly suited for the fabrication of biological-inorganic hybrids and composites. For the design of advanced materials, the advantage of using sol–gel derived glasses is that the parent silica material is structurally inert, functionally inactive, and operationally nonresponsive. Therefore, by selectively integrating specific (bio)chemical entities into the glass (either through precursor modification or through encapsulation), it is possible to introduce desired structural, functional, and operational properties in a modular fashion. The introduction of properties can be singular, sequential, or even multiple. The strategy offers a powerful approach for designing a diverse range of sol–gel-derived materials whose properties can be tailored with a degree of control over the compositional, morphological, and functional characteristics of the product materials. This flexibility enables exploring new opportunities, and materials made using the sol–gel method have found a variety of applications in optics, photonics, magnetics, biocatalysis, detection, sensing, controlled release, and separation in addition to providing access to a range of structures and morphologies in the form of particles, fibers, coatings, and other nanoscale, nanostructured, and nanoporous materials.

The sol–gel method has brought about a fundamental change in synthesis and processing of inorganic materials. The fact that final materials in different geometries and morphologies can be prepared from a solution circumvents the need to go via conventional solid-state high temperature processing pathways. The solution-based method provides considerable flexibility in making pure, homogenous materials as well as composites and hybrids. The method can be used to make materials by using a given precursor, or alternatively, by mixing a plurality of different precursors, it is possible to form multicomponent mixed-oxides. Similarly, the method can be used to make hybrids and composites via integration of polymers or other organic/biological components to obtain materials with varied properties and applications (Fig. 2). The low temperature at which the materials can be prepared makes the method economical and environmentally friendly for practical applications.

Basic Methodology

The sol–gel method of making glasses begins with the reaction of hydrolyzable precursors with water. The most commonly used precursors are alkoxides. The process is illustrated herein with the alkoxysilanes as an example; however, the reaction sequences are analogous for other metals. These precursors are dissolved in a compatible solvent such as an alcohol followed by addition of water to initiate hydrolysis. Often a catalyst is used to accelerate the reaction. The nonpolar nature of alkoxysilanes makes them immiscible with water, and therefore, the reactions proceed at very slow rates even in the presence of a catalyst. The mixing can be enhanced by use of ultrasound, or alternatively, the reaction mixture is refluxed at elevated temperature to facilitate the reaction. Under these conditions, first the alkoxides react with water and form hydroxylated species which then condense to form colloidal nanoparticles containing oxo-bridges. As the reactions proceed, the sol particles combine to form an extended network and the reaction mixture becomes viscous, ultimately leading to the formation of a



Sol–Gel Method, Fig. 2 Representative summary of significant aspects of the sol–gel method

solid gel. The as-prepared gels are biphasic porous materials with interstitial solvent phase. Drying of the gels under ambient conditions results in evaporation of a majority of the solvent phase (water and alcohol) and the gels shrink by up to 80 % in volume. These dried glasses – termed xerogels – are mechanically and dimensionally stable materials that contain some solvent phase, which can be further dried at elevated temperatures to form dense glasses. The overall method of making oxides from alkoxides is represented as

$$\text{Precursor} \xrightarrow[\text{Condensation}]{\text{Hydrolysis}} \text{Sol} \xrightarrow{\textit{Gelation}} \text{Gel} \xrightarrow{\textit{Drying}}$$

$$\text{Xerogel} \xrightarrow{\textit{Densification}} \text{Glass}$$

The sequence is characterized by a series of reactions along with physical and chemical changes that take place along the precursor to sol to gel to xerogel to glass transformation steps. The properties of materials depend upon the conditions employed and by controlling the parameters under which these steps take place one can modulate, regulate, and fine-tune the properties of the materials at each stage.

The Precursor: The nature, composition, and structure of precursor molecules is central to the properties of final product. Controlling the number of hydrolyzable versus nonhydrolyzable groups on the precursor provides an easy pathway to influence the chemical makeup of the materials obtained by the sol–gel method. Therefore, in order to obtain pure oxide-based materials, one can use precursors of the type $Si(OR)_4$ wherein all the alkoxides are replaced with oxygen atoms in the product due to hydrolysis and condensation. Use of nonhydrolyzable ligands on the precursors

ensures that these groups will remain in the final material thereby altering the structural, morphological, and functional characteristics of the product. Precursors of the type $(OR)_3SiR'$ where R′ is organic group provide an easy access to introduce organic functional groups that can impart specific properties. Additionally, use of multiple precursors with different functional groups provides further access to tailoring the final properties.

Alkoxides are the preferred precursors of choice due to the fact that the by-product of their hydrolysis is the corresponding alcohol, which can easily be evaporated to obtain relatively pure materials [4]. The rates of hydrolysis and condensation depend upon the nature of the alkoxide groups, which can lead to dramatic variations in the rates of the reactions as well as product morphologies. The rates of hydrolysis of alkoxides depend upon the polarity of M-OR bond as well as the steric environment around the central metal atom which can influence the approach of water molecule for hydrolysis of the bond. In general, more polar alkoxides are more susceptible to hydrolysis in a polar solvent environment. The inductive and steric effects of the alkoxide ligands can also alter the hydrolysis rates by altering the polarity of the M-OR bonds, and by modulating the access of water molecules to effect the hydrolysis step. Usually, the hydrolysis step is much slower with longer chain and/or branched alkoxides which inhibit the access of water molecules to the metal site. A catalyst is typically used, especially with silicon alkoxides, to accelerate the hydrolysis step. On the other hand, alkoxides of transition metals and other highly electropositive elements usually exhibit much faster hydrolysis due to the intrinsic polarity of the M-OR bond along with the ability of these elements to accommodate the incoming water molecules via coordination to the metal site.

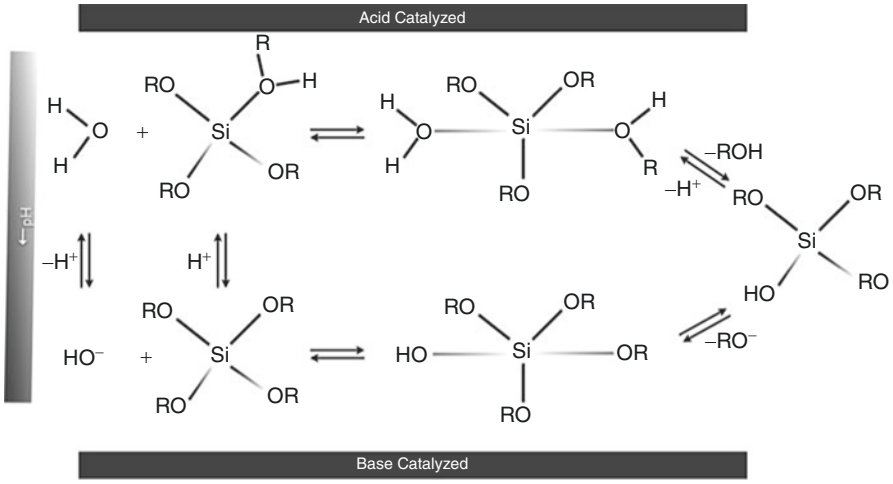
Hydrolysis: The first step of the sol–gel reaction is the replacement of alkoxide groups in the precursor molecule with water. The alkoxide groups leave as the corresponding alcohols along with formation of hydroxylated silanol species. Since alcohol is a by-product of the hydrolysis reaction, carrying out the reaction in the same alcohol as the corresponding alkoxide can alter the reaction equilibrium. If a different alcohol is used as a solvent then, due to its predominance, there is usually a replacement of the alkoxide ligand via transesterification. The ease with which the reaction occurs depends upon the miscibility of the precursors with water, its hydration characteristics, and its ability to accommodate water molecules in the primary coordination sphere. The products of the hydrolysis reaction are the hydroxylated precursor and the corresponding alcohol. The general reactions is given as

$$(OR)_4\,Si\ +\ H_2O \rightleftarrows (OR)_3\,Si\ -\ ROH$$

The hydrolysis reaction proceeds via associative (or S_N2) pathways such that the transition state is characterized by an increase in coordination number. The reaction is usually acid or base catalyzed to make it occur at an appreciable rate. The general reaction under both conditions proceeds via associative pathways through the aid of the predominant catalytic species, namely, protons in acidic conditions and the hydroxide ions under basic conditions.

Under acidic conditions, the reaction occurs in three steps (Fig. 3). First, the alkoxide group gets protonated, which makes the metal-alkoxide bond weaker and the silicon atom electron-deficient. Second, the water molecule binds to the silicon atom to form a positively charged pentacoordinate transition state. Finally, the release of alcohol molecule completes the reaction. Under basic conditions, the reaction also occurs in three steps (Fig. 3). First, the water molecule loses its proton to form the hydroxide ion. Second, the hydroxide ion binds to the silicon atom to form a negatively charged transition state. Finally, the alkoxide ion dissociates to form the hydroxylated species.

The replacement of alkoxides to form hydroxylated species proceeds at different rates under acidic and basic conditions. Under acidic conditions, subsequent hydrolysis steps are retarded after the initial substitution of alkoxide, while in basic medium, these follow-on steps are accelerated. In other words, the hydrolysis of $[(OR)_4 - {}_nSi(OH)_n]$ (where n varies from 0 to 4) species in acidic medium is faster for smaller values of n



Sol–Gel Method, Fig. 3 Hydrolysis mechanisms in acid or base catalyzed reactions

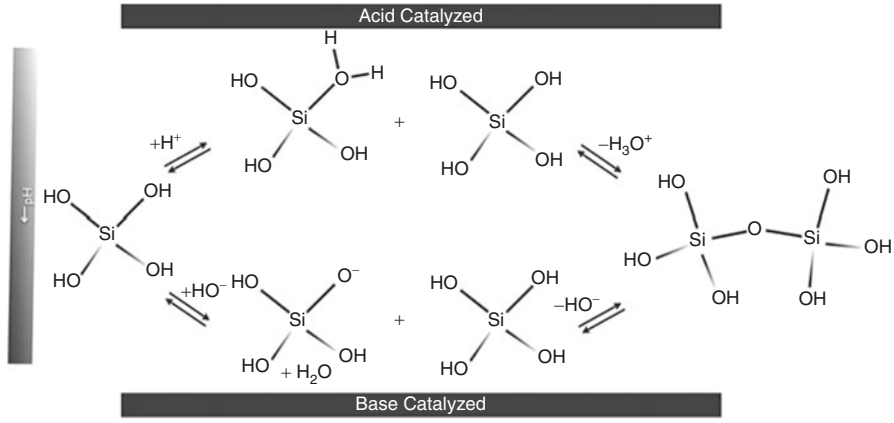
while in a basic medium the hydrolysis is faster for larger values of n. As such, the formation of multihydroxylated species predominates in the basic medium. The rate and extent of hydrolysis can be altered by use of suitable solvents that can change the polarity of the medium, and consequently, the stability of polar versus nonpolar species to shift the equilibrium. Solvents that interact with water by hydrogen bonding interactions also alter the activity of water in the solution and affect the rate and extent of hydrolysis. Solvents that can homogenize the reactants to make them more miscible augment the rates and shift the equilibrium toward products. Finally, use of ultrasound can also facilitate mixing of the precursor with water and can accelerate the reaction.

Condensation: Once the reactive hydroxylated species are formed, they begin to undergo condensation reaction to form Si-O-Si linkages to generate oligomers and polymers. The reaction can be represented as

$$(\mathrm{OR})_3\,\mathrm{Si} - \mathrm{OH} + \mathrm{OH} - \mathrm{Si}(\mathrm{OR})_3 \rightleftarrows (\mathrm{OR})_3\mathrm{Si} - \mathrm{O} - \mathrm{Si}(\mathrm{OR})_3 + \mathrm{H_2O}$$

As with hydrolysis, the condensation reactions are dependent upon pH of the medium and can be acid- or base-catalyzed (Fig. 4). Under conditions of acid catalysis, the reaction proceeds in three steps. First, there is a rapid protonation of the silanol group to form a positively charged species. Second, the protonated positively charged species is attacked by a neutral hydroxylated species. Finally, an elimination of the hydronium ion completes the reaction. Base catalysis follows a slightly different sequence. First, there is rapid deprotonation of the silanol group to form an anionic silicate species. Second, the negatively charged species coordinates with the neutral hydroxylated species via formation of a pentacoordinate transition state. Finally, the expulsion of hydroxide ions completes the reaction sequence.

Analogous to hydrolysis reactions, the rates of the condensation reaction are dependent upon pH of the medium. Condensation reaction for $[(\mathrm{OR})_4 - {}_n\mathrm{Si}(\mathrm{OH})_n]$ (where n varies from 1 to 4) species in acidic medium is faster for smaller values of n while in a basic medium it is faster for larger values of n. As such, base-catalyzed systems, with a predominance of multihydroxylated species, are characterized by faster gelation times as compared to acid-catalyzed reactions. Silanol (Si-OH) groups become more acidic as the silicon atom gets increasingly surrounded by the electron withdrawing oxo groups. This high acidity of silanols with the increasing extent of condensation results in maximum condensation rates in the intermediate pH range of 3–5. In very high pH region



Sol–Gel Method, Fig. 4 Condensation mechanisms in acid or base catalyzed reactions

(pH $\geq$ 10), condensation reactions occur but gelation is thwarted by repulsion due to surface charges leading to formation of discrete particles. As water is a by-product of condensation reactions, the amount of water in the reaction mixture influences the equilibrium. Reaction media high in water content favor hydrolysis but hinder condensation. Similarly, more polar and hydrogen bonding solvents can stabilize the hydroxylated species to retard the condensation process. Finally, the steric effects of the remaining alkoxo groups can also influence the rate with bulkier groups hindering the condensation reactions.

Gelation: As the hydrolysis and condensation reactions continue, the reaction medium becomes more viscous due to increased degree of polymerization. The viscous sol continues to evolve until the viscosity reaches a critical point where the liquid can no longer maintain its fluidity. Gelation is characterized as the point along the hydrolysis-condensation reaction coordinate when the liquid sol turns to a solid gel. Gelation is typically defined by gelation time (t_g) as the time taken from the start of the reaction till the formation of a solid gel. Gelation times are dependent upon a variety of factors, as discussed above, that control the rate and extent hydrolysis and condensation reactions. At gelation time, the growing polymer structure essentially gets "frozen" into an immobile state. The initial gels formed have high viscosity but they lack the mechanical stability due to lack of an extensive interconnected network. Even though the physical state of the reaction mixture changes from liquid to solid, the hydrolysis and condensation reactions still continue to evolve leading to formation of additional Si-O-Si linkages. The increasing degree of condensation results in enhanced mechanical strength over time even after gelation has occurred.

Aging: The as-formed gel is a porous structure with pendent silanols and even some remaining unhydrolyzed alkoxides on the polymeric network. Aging of the gels under closed conditions, usually extending over a period of hours to days, is employed to enable completion of these reactions. During aging, continuation of hydrolysis and condensation reactions leads to an increase in cross-linking and overall stiffening of the gel network. During aging, the gels shrink and expel the solvent phase due to a decrease in internal pore volume caused by extensive condensation of silanols to siloxane linkages. Aging is a necessary step in processing of bulk gels to ensure mechanical stability of these materials during subsequent stages.

Drying: Drying of the gels either under ambient conditions or elevated temperature is used to expel solvent from the gels. The elimination of solvent from the pores of the gels compacts the network and facilitates further condensation reactions between silanols on the surface of pores. The ambiently dried gels can lose up to 70–80 % of

the solvent and shrink considerably due to compaction of the pore network. In order to completely dry the gels, use of elevated temperatures is necessary. Drying enhances the mechanical stability of the gels, and the dried gels are considerably stiffer and stronger materials as compared to aged gels. The dried gels have a greater density and are dimensionally invariant under ambient conditions.

Sintering/Densification: Heat treatment of the gels further eliminates the porous structure to obtain ceramics or dense glasses. During heat treatment, adsorbed and hydrogen-bonded solvent is removed and the collapsing pore structure causes further condensation of silanols to siloxane linkages. Heat treatment makes the gels nonporous and harder with densities and structure approaching close to fused silica.

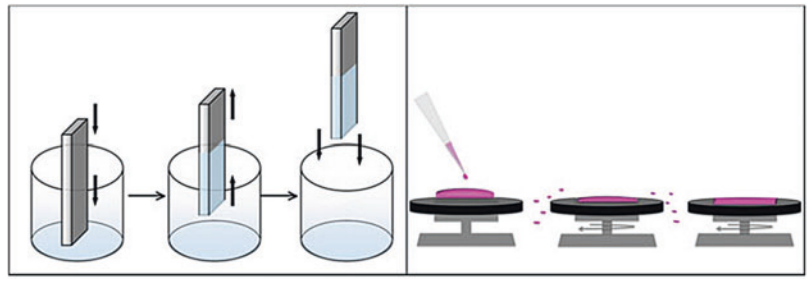
Nonhydrolytic sol–gel method: The sol–gel method depends upon water to initiate the hydrolysis of precursors. This presents a challenge for system that are either unstable in water or too reactive. In order to circumvent these issues, the nonhydrolytic sol–gel method [5] utilizes nonaqueous solvent medium and does not require the hydrolysis step. While there exist many different variants, the general nonhydrolytic sol–gel method can be represented as

$$\mathrm{M(OR)_n + X_nM \rightleftarrows \left[(OR)_{n-1}M - O - M(OR)_{n-1}\right] + RX}$$

where X = halide, -OR′, or -O_2CR′. The reaction bypasses the hydrolysis process altogether and leads to direct formation of oxo-bridged linkages. The reaction is essentially a condensation pathway that is usually carried out at elevated temperatures. The nonhydrolytic process is advantageous in a system wherein a greater degree of control over the relative rates of hydrolysis and condensation to fine tune product microstructure, porosity, and morphology may not be possible with the hydrolytic pathway.

Formation of nanoparticles and powders: The molecule-nanomaterial interface along the precursor to sol reaction coordinate provides a convenient access to a variety of nanoscale systems. Powders such as precipitated or fumed silica have been long used in commercial applications. A key challenge has been to obtain uniform monodisperse nanoscale particles with a control over their nuclearity, shape, and dimensions. The issue of uniformity and homogeneity, to a certain extent, is addressed by utilizing specific reaction conditions to prevent interparticle interactions. Two methods commonly used to stabilize isolated discrete nanoparticles are based on manipulating electrostatic and steric factors. As discussed above, highly basic conditions favor rapid hydrolysis and condensation while producing negative surface charges which favor the formation of particulate structures. The Stober method of forming particles specifically takes advantage of this approach. Uniform size spherical particles can be obtained under basic conditions by using dilute solutions to prevent interparticle aggregation and to allow dissolution-growth phenomena to occur without precipitation. Usually the concentration of the precursor in the reaction mixture is optimized to prevent interparticle polycondensation and prevent formation of larger particles due to particle growth and coalescence. Sometimes, a synthesis strategy of combined acid and base catalysis is used for the initial formation of a stable sol under acidic conditions followed by addition of base to facilitate particle growth. Optimized reaction conditions that favor dynamic surface reorganization and reconstruction to provide uniform spherical particles are commonly utilized. Use of microemulsions has been another method that is employed to form spherical nanoparticles wherein the emulsion provides a sterically enclosing environment for isolated particles to form without aggregation.

Formation of thin films and coatings: The sol–gel method has found wide applicability in the construction and formation of two-dimensional nanostructures such as thin films and coatings. The films made via the sol–gel method provide a better dimensional control with the ability to coat complex shapes, and can even be used to form multilayer coatings. In order to form a coating, a stable sol with controlled viscosity is desired. Usually stable sols can be obtained under acidic conditions and by adjusting the amount of the precursors, solvent, and pH it is possible to control the viscosity of the sol. The films can be formed



Sol–Gel Method, Fig. 5 Schematic depiction of dip-coating (*left*) and spin-coating (*right*) processes

by dip-coating, spin-coating, or spray-coating methods. The dip-coating method of thin film deposition is perhaps more important technologically since a uniform coating can be deposited onto substrates of large dimensions and complex geometries. In the dip-coating process, a suitable substrate immersed in a low viscosity sol is withdrawn slowly at a constant speed (Fig. 5). The adhesion of the sol layer on the substrate aided by gravitational draining leads to formation of a thin film. In order to deposit uniform homogeneous films, a steady vibration-free withdrawal of the substrate from the sol is necessary. Film thickness is controlled by adjusting the viscosity of the sol and the rate of pulling the substrate. Another method of forming thin films is via spin-coating wherein an excess of sol is placed onto a substrate that is rotated at high speeds to disperse the liquid uniformly (Fig. 5). The film thickness in this case depends upon viscosity of the sol and the speed of rotation along with the rate of evaporation of the solvent. The conversion of deposited sol to a porous gel is caused mainly by solvent evaporation. Unlike the bulk gel system where gelation, aging, and drying occur sequentially over a period of several days, all of these processes typically occur within 30 s to 1 min in the thin-film such that the drying stage overlaps the gelation and aging stages. Loss of solvent molecules accelerates hydrolysis and condensation steps such that the two steps occur concurrently leading to the eventual formation of a porous xerogel structure. The important microstructural characteristics of coatings and films such as porosity, surface area, and pore size are determined by pH, concentration, and nature of precursors, viscosity of the initial sol, the solvent and its rate of evaporation, the rate of sol dispersion on the substrate, the ambient temperature, the rate of drying, and the relative rates of evaporation and condensation reactions during drying.

Organically modified materials: Perhaps the most useful aspect of the sol–gel method is the ability to make materials with covalently integrated organic functionalities via hydrolysis of precursors containing hydrolytically stable organic groups. This enables a wide range of functional groups to be incorporated within the oxide-based network to from a diverse array of new materials with unique, novel, and interesting properties. More than one type of precursors with different functional groups can also be used to include a multiplicity of functional groups within a given material. An important consideration in use of multiple precursors is their differential reactivities toward hydrolysis and condensation which can lead to phase separation and/or spatially isolated nanodomains. Strategies to overcome these are based on prehydrolysis of the more reactive precursor followed by addition of the other precursors as well as use of different alkoxides groups on the precursors to modulate their reactivities. The organically modified materials contain less than four siloxane linkages per silicon atom and are more elastic as compared to pure oxides. The ability to tune functionality in the organically modified materials provides appealing prospects for applications based on the nature of organic group incorporated.

Hybrid Materials: Different types of hybrids can be made via the sol–gel method. These include organic–inorganic hybrids and nanocomposites. The organic–inorganic hybrids combine the properties of organic materials with the mechanical strength of the oxides. The main issue in formation of hybrids relates to their miscibility which can be enhanced by means of suitable solvents or by use of

functional groups that can noncovalently interact with silica pore structure via hydrogen bonding or electrostatic interactions. The simplest way to make a hybrid material is by encapsulation of organic molecules in the pores of the gels. The organic molecules in the pores of gels can be used to alter the properties and structure of the inorganic network. Typical examples of these include glasses containing encapsulated catalysts, dyes, and other optically active organic molecules. Another approach is based on forming interpenetrating networks of organic polymer with the silica network. These nanocomposites combine the properties of both inorganic and organic polymers, and usually influence the structure, morphology, and properties of each other. These synergistic interactions make the hybrid materials structurally and functionally superior as compared to their component parts.

An important development in formation of hybrid materials by the sol–gel method centers on formation of bio-hybrids. The solution based sol–gel pathway can be used for direct integration of biological species such as proteins, enzymes, DNA, RNA, and even whole cells. Optimized pH regions are necessary during preparation to stabilize the biological component. Typical procedure of enzyme immobilization involves initial acid hydrolysis followed by raising the pH to neutral region. The biological macromolecules are usually added to a buffered sol and get encapsulated in the growing gel network as the sol turns to a gel. The water-filled porous structure of the gel along with electrostatically charged and hydrogen bonding pore walls provide a noncovalently interacting environment for the biological molecules. The encapsulated biological entity retains its structural and functional characteristics and is able to interact with exogenous molecules that can diffuse in and out of the porous structure. The last two decades have seen tremendous advances in the design and development of bio-silica hybrids made by the sol–gel method [3].

Key Research Findings

Structure and Synthesis: The sol–gel method has been used for several decades to make silica-based materials [6]. Oxides of different metals including mixed oxides have been prepared [7]. Initial research in this area focused on mainly pure and mixed oxides but in recent times the focus has shifted to organically modified, hybrid, and composite system. A classification of organic–inorganic materials synthesized using the sol–gel method has been developed to categorize them [8]. Nanotechnology has also spurred an active interest fabrication of sol–gel-derived particles [9]. These particles are usually synthesized in a microemulsion and can be solid, hollow, or contain core-shell structures.

Synthesis of organically modified glasses has been studied in much detail and a wide range of precursor chemistries have been identified. The organically modified precursors provide access to a diverse range of materials with tunable properties including chemical reactivity, optical properties, electronic properties, enhanced mechanical stability, and elasticity. A major area of research using these precursors has been in the construction of mesoporous materials by means of using suitable templates wherein the assembly is aided by electrostatic and hydrogen bonding interactions between the template and the growing silica network [10]. Gels made from organosilicate precursors made with spacered linkers containing both hydrophobic and hydrophilic groups exhibit hydrogel-like properties and have been shown to undergo structural and volume change with respect to different environmental stimuli. These materials are characterized by enlarged pores due the presence of spacer units along with elasticity which confers the necessary flexibility to undergo shape and volume changes [11]. These materials also provide a suitable matrix for immobilization of biological molecules.

An important aspect of protein and enzyme immobilization has been elucidation of local environment of the immobilized biomolecule and its interactions with the sol–gel matrix [12]. The immobilizing environment of the pore restricts translational diffusion, rotational mobility, and conformational flexibility. Proteins and enzymes have been shown to exhibit enhanced stability against denaturation when encapsulated in the sol–gel-derived glasses. The enzymes in the gel

matrix retain their catalytic activities and the materials have been used as solid-state catalysts for various transformations. Studies have shown that when large proteins get encapsulated in the growing gel network, they act as self-specific templates for the formation of their own pore environment. Recent research also shows that cells can be used for directing assembly of sol–gel materials. Furthermore, proteolytic enzymes have been shown to catalyze the sol–gel reactions leading to formation of gels starting from alkoxide precursors.

Function and Properties: The transparency of the sol–gel-derived glasses has been widely utilized for the development of functions related to modulation of photon flux through the materials [13]. Incorporation of dyes and luminescent fluorophores has been used as a means to develop materials for lasers, luminescent materials, and solar concentrators. Incorporation of photochromic molecules in the pores of the gels has been employed as a means to impart photochromic properties to the resulting glasses. Electronic properties have been developed in the sol–gel glasses via incorporation of conducting polymers, metal particles, or formation of mixed oxides. The nanopores of the gels have been used to form nanoscale superparamagnetic particles of transition metal oxides. The use of organically modified precursors also provides access to tailoring of mechanical properties such as strength, elasticity, and hardness depending upon the functional groups used. In addition to development of physical properties, use of precursors with specific functional groups has also been used to tune their chemical properties.

Applications: The applications of the sol–gel-derived systems have been continuously increasing [14]. The transparency of silica-based systems makes them especially suitable for development of optical applications and sol–gel-derived materials have been used as optical sensors based on changes in color or luminescence upon reaction with an analyte [15]. The analytical applications of sol–gel materials as electrochemical sensors have also been developed [16]. Additionally, the presence of specific functional groups on organosilica materials confers selectivity in their interactions and these materials have been used as separation media in chromatography columns [17]. The use of porous particles for controlled release applications has opened up new opportunities in market and commercial products have been designed for acne treatment and sunscreen release based on particles made using the sol–gel method. A new area in this direction has been applications in fabric-care systems and environmentally sensitive organosilica hydrogels have been demonstrated as controlled release vehicles for water-triggered release of enzymes.

Coatings made from sol–gel pathways have been utilized for an array of applications including corrosion protection and surface passivation, tailoring water affinity of surfaces, enhancing optical gloss, thermal insulation, scratch resistance, and fingerprint resistance among a host of other applications [18]. The ability to modify hydrophobicity and surface microstructure of these coatings has enabled construction of surface features that mimic the lotus leaf effects. An offshoot of this research has also resulted in textile treatments to make them stain- and water-repellent.

In recent times, the use of bio-sol–gel hybrids has been an area of much activity. Initial applications related to immobilization of protein and enzymes inside the pores of silica sol-gels to make solid-state biocatalysts for various chemical transformations. The enzyme sol–gel hybrids have also been used as optical and electrochemical sensors, and as media for bio-affinity chromatography. The use of sol–gel-derived materials has also been established for immobilization of whole cells and microorganisms [19]. Organosilica gels as matrices for bio-immobilization have provided access to a wide range of biocatalysts whose activities can be exploited in practical applications. The applications domain of materials incorporating biological species continues to grow and the biological-inorganic composite area remains a topic of continued development.

Future Directions for Research

The sol–gel method of making materials provides a versatile pathway to not only making new materials, but also making known materials more useful by obtaining them in different morphologies and length scales. The last decade has witnessed significant advances in construction of nanoscale particles, assemblies, and organized structures made using the sol–gel method. Fabrication of different oxides, nanohybrids, and nanocomposites characterized by distinct nanoscale structures, surfaces, and interfaces that can be tailored to obtain the desired functional responses would continue to remain an area of vital investigation in the near future. Use of nanoscale building blocks obtained by the sol–gel method and their organization into higher-order structures along with control, manipulation, and regulation of the emergent properties at different length scales to generate specifically tailored responses would provide an appealing paradigm for future studies. An additional key issue to address in the future development of nanoscale systems made via the sol–gel method would be resolving, manipulating, and controlling the functional dichotomy of properties originating from the nanoscale as well as collective properties due to the ensemble nature of the materials.

Being able to tailor properties of materials made from molecular precursors with specific chemistries offers enticing prospects for controlling and manipulating their organization as they undergo their structural evolution from molecules to clusters to particles and other higher order structures. While there have been some advances in supramolecular assemblies of sol–gel-derived materials, the complexity and sophistication to achieve self-assembling structures with distinct interactions, interfaces, and organization over varied length scales remain as yet elusive. An area of ambitious exploration, given the formidable challenges, would be the development of new precursors and chemistries to fine-tune covalent and noncovalent interactions to generate supramolecular systems with programmed assembly and organization. Tuning this dynamic interplay of structure and organization would require new methods and techniques of sol–gel chemistry that would provide access to far from equilibrium structures. In the near term, the structural-organizational paradigms from nature would continue to provide the necessary inspiration toward achieving these objectives.

The ability to fabricate bio-inorganic hybrids and composites has opened another exciting avenue in the design of novel materials with useful properties and applications. The integration of biological entities into the nanopores of materials, to a certain extent, restores the spatial and supramolecular organization that exists in the native cellular environment which is lacking in freely mobile biomolecules dispersed in an isotropic solution phase. The next extension of this method would generate organized assemblies that would take advantage of the specialized interactions at the biological-inorganic interfaces. Consolidated nanostructures integrating biological materials would become increasingly significant in creating a new generation of bio-hybrids that would combine the structural integrity of the sol–gel-derived materials with the functional and operational diversity of biomaterials.

During the last decade there have been significant advances in mimicking nature with sol–gel-derived materials including biocatalysis, immunodetection, surfaces that mimic lotus leaf, and biomineralized supramolecular architectures. The opportunities provided by the facile access to nanomaterials, internal porous architecture, surfaces, and interfaces would continue to evolve as chemists, materials scientists, biochemists, and physicists attempt to develop correlations between structure, morphology, form, and function. The next evolution of these materials lies in the design of nanomaterials with adaptive capabilities that can adapt to their environment and exhibit active functional responses to generate dynamically active systems capable of self-diagnosis and self-correction. The interplay of

structure, form, function, and operation that exists at different length scale would provide the guidelines for the development of next generation of materials capable of intelligent responses. While recent developments have seen the emergence of materials with smart functions, many of the systems, for the most part, remain monofunctional and passive. In the long range, design and development of new materials with hierarchical functions that span a range of functional length scales and tunable active responses would generate the next wave in this line of research.

The last decade has also seen utilization of sol–gel methods to fabricate devices for detection, diagnosis, separation, sensing, controlled release, and delivery to name a few. The real potential of these materials as nanoscale devices remains to be explored. The ability to tailor size, porosity, and the surfaces of sol–gel nanomaterials would find new avenues of research activity in the development of nanodevice technology based on a combination of inorganic, organic, and biological components. Multifunctional systems that can detect, sense, separate, sort, release, and deliver molecules to targeted sites and in a predetermined fashion would constitute the long range evolution of nanodevices made using sol–gel-derived systems. In the near term, effective utilization of strategies to adjust and modulate physical and chemical interactions of exogenous molecules with sol–gel systems would provide the necessary means of control and regulation to design nanoscale sensors, filters, separators, sorters, and delivery agents. Development of nanodevice systems for biomedical applications would also constitute an important area of research and development.

Finally, another area that would become increasingly important is the development of new "green" technologies utilizing the sol–gel method. The opportunities offered by the sol–gel method would provide the ideal framework for development of novel (bio)catalytic systems, molecular filters, sorters, or separators to reduce, remove, and/or remediate toxic and harmful substances. The sol–gel method is a greener alternative to conventional methods, however, use of volatile solvents and formation of by-products remain as issues to be addressed. Development of new processing techniques that minimize waste and improve atom economy by eliminating by-products would be essential to the development of eco-friendly manufacturing processes based on sol–gel methods. In the long range, sustained development and refinement of the sol–gel method would be critical to displace the current manufacturing processes in large-scale production technologies.

Cross-References

- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Biosensors
- ▶ Self-Assembly of Nanostructures
- ▶ Smart Hydrogels

References

1. Wright, J.D., Sommerdijk, A.J.M.: Sol-Gel Materials – Chemistry and Applications. CRC Press, Boca Raton (2001)
2. Sanchez, C., Belleville, P., Popall, M., Nicole, L.: Applications of advanced hybrid organic–inorganic nanomaterials: from laboratory to market. Chem. Soc. Rev. **40**, 696–753 (2011)
3. Avnir, A., Coradin, T., Lev, O., Livage, J.: Recent bio-applications of sol–gel materials. J. Mater. Chem. **16**, 1013–1030 (2006)
4. Hubert-Pfalzgraf, L.G.: Alkoxides as molecular precursors for oxide-based inorganic materials – opportunities for new materials. New J. Chem. **11**, 663–675 (1987)
5. Mutin, P.H., Vioux, A.: Nonhydrolytic processing of oxide-based materials – simple routes to control homogeneity, morphology, and nanostructure. Chem. Mater. **21**, 582–596 (2009)
6. Hench, L.L., West, J.K.: The sol-gel process. Chem. Rev. **90**, 33–72 (1990)
7. Livage, J., Henry, M., Sanchez, C.: Sol-gel chemistry of transition metal oxides. Prog. Solid State Chem. **18**, 259–341 (1988)
8. Sanchez, C., Ribot, F.: Design of hybrid organic-inorganic materials synthesized via sol-gel chemistry. New. J. Chem. **18**, 1007–1047 (1994)
9. Ciriminna, R., Sciortino, M., Alonzo, G., de Schrijver, A., Pagliaro, M.: From molecules to systems – sol-gel microencapsulation in silica based materials. Chem. Rev. **111**, 765–789 (2011)
10. Rivera-Munoz, E.M., Huirache-Acuna, R.: Sol-gel-dervied SBA-16 mesoporous material. Int. J. Mol. Sci. **11**, 3069–3086 (2010)
11. Rao, M.S., Grey, J., Dave, B.C.: Smart glasses – molecular programming of dynamic responses in

organosilica sol-gels. J. Sol-Gel. Sci. Technol. **26**, 553–560 (2003)
12. Menaa, B., Menaa, F., Aiolfi-Guimaraes, C., Sharts, O.: Silica-based nanoporous glasses – from bioencapsulation to protein folding studies. Int. J. Nanotechnol. **7**, 1–45 (2010)
13. Penard, A.-P., Thierry, G., Boilot, J.-P.: Functionalized sol-gel coatings for optical applications. Acc. Chem. Res. **40**, 895–902 (2007)
14. Aegerter, M.A., Menning, M.: Sol-gel Technologies for Glass Producers and Users. Kluwer, Boston (2010)
15. Tran-Thi, T.-H., Dagnelie, R., Crunaire, S., Nicole, L.: Optical chemical sensors based on hybrid organic–inorganic sol–gel. Chem. Soc. Rev. **40**, 62–639 (2011)
16. Walcarius, A., Collinson, M.M.: Analytical chemistry with silica sol-gels – traditional routes to new materials for chemical analysis. Annu. Rev. Anal. Chem. **2**, 121–143 (2009)
17. Kloskowski, A., Pilarczyk, M., Chrzanowski, W., Namiesnik, J.: Sol-gel technique – a versatile tool for adsorbent preparation. Crit. Rev. Anal. Chem. **40**, 172–186 (2010)
18. Aegerter, M.A., Almeida, R., Soutar, A., Tadanaga, K., Yang, H., Watanabe, T.: Coatings made by sol–gel and chemical nanotechnology. J. Sol-Gel. Sci. Technol. **47**, 203–236 (2008)
19. Livage, J., Coradin, T.: Living cells in oxide glasses. Med. Mineral. Geochem. **64**, 315–332 (2006)

Solid Lipid Nanocarriers

▶ Solid Lipid Nanoparticles (SLN)

Solid Lipid Nanoparticles (SLN)

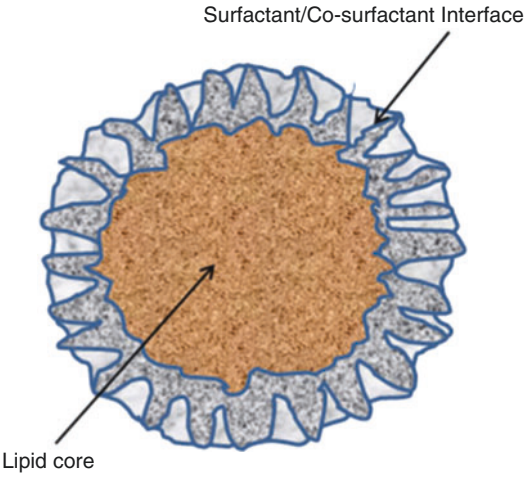
Claudia Musicanti and Paolo Gasco
Nanovector srl, Torino, Italy

Synonyms

Lipospheres; Solid lipid nanocarriers; Solid lipid nanospheres; Solid lipid-based nanoparticles

Definition

Solid lipid nanoparticles (SLN) are drug carriers in submicron size range (50–500 nm) made of biocompatible and biodegradable lipids solid at room and body temperature.



Solid Lipid Nanoparticles (SLN), Fig. 1 Model representation of general structure of SLN

Introduction

Solid lipid nanoparticles (SLN) were developed at the beginning of the 1990s as alternative colloidal carriers to emulsions, liposomes, and polymeric nanoparticles. SLN have attracted increasing attention as delivery system for hydrophobic drugs: prepared with lipids, they can be administered by different routes of administration such as oral, parenteral, dermal, transdermal, ocular and pulmonary.

A typical model of the structure of SLN consists of a solid lipid core surrounded by an emulsifier interface which stabilizes the particle (Fig. 1).

SLN can offer several advantages for drug delivery

1. Possibility of controlled release of drug from lipid matrix
2. Possibility of drug targeting (active and passive)
3. Protection of incorporated drug from degradation
4. Increasing drug bioavailability

5. Feasible incorporation of hydrophobic and also hydrophilic drugs (lipid drug conjugates or other techniques)
6. Very low when absent toxicity
7. Affinity for biological barriers
8. Modification of pharmacokinetic parameters and drug distribution in organs
9. Possibility of large-scale production

On the other hand SLN show certain possible limitations:

1. Poor drug loading capacity
2. Possible drug release during storage
3. Possible stability problems during long-term storage (aggregation, component degradation, tendency to form gel)

The principal components used for the preparation of SLN are:

1. Lipids, such as triglycerides (trilaurin, trimyristin, tripalmitin, tristearin), mono/di/triglycerides mixtures (glyceryl stearate, glyceryl palmitostearate, glyceryl behenate), fatty acids (stearic acid, palmitic acid, behenic acid), waxes (cetyl palmitate), cholesterol esters
2. Emulsifiers (surfactants), such as phospholipids (e.g., lecithins), polysorbates (Tween) and sorbitan esters (Span), polymers (e.g., poloxamers), generally used to stabilize interface between water and lipid
3. Co-emulsifiers (cosurfactants), such as bile salts, short-chain fatty acids and alcohols

Emulsifiers are amphiphilic molecules which possess both hydrophobic and hydrophilic portions; they are able to stabilize emulsion system by placing at the oil/water (o/w) interface (orienting the hydrophobic portion to the oil phase and the hydrophilic portion to the aqueous phase) and lowering the interfacial tension between the two phases. Co-emulsifiers intercalate between the emulsifier molecules and contribute to reduce the interfacial tension between oil and water.

Additional ingredients are used to specifically modify surface of SLN:

1. Stealth agents, such as polyethylene glycol (PEG) for improving circulation time
2. Charge modifiers to modify surface charge
3. Targeting molecules, such as peptide or antibody fragment, for linking to specific site receptor (active targeting)

Antioxidants, antimicrobial agents, and thickening agents can be usually added to SLN (dried or in dispersion) as for other pharmaceutical products.

Preparation Methods of SLN

General issue in the preparation of SLN is to obtain liquid phase where lipid and water can be emulsified before SLN formation: lipids, which are solid at room temperature, are melted or they are dissolved into solvents. Main methods to prepare SLN are reported below.

High-Pressure Homogenization (HPH)

High-pressure homogenization is a well-established technique on the large scale and it is already available in the pharmaceutical industry. High-pressure homogenizers have been used extensively in the production of nanoemulsions for parenteral nutrition.

High-pressure homogenizers push a liquid with high pressure (100–2,000 bar) through a micron size gap: by applying high pressure, the liquid accelerates to high velocity (over 1,000 km/h), and the resulting shear stress and cavitation forces break down the accelerated particles to submicron size.

HPH technique can be applied by two main different approaches: hot high-pressure homogenization (HHPH) technique and cold high-pressure homogenization (CHPH) [1].

For both techniques, the drug is dissolved or dispersed in the melted lipid at approximately 5–10 °C above its melting point.

For the HHPH, the drug-loaded lipid is dispersed under high-speed stirring in a hot aqueous surfactant solution maintained at same temperature, to form a pre-emulsion. The obtained pre-emulsion is then subjected to high-pressure homogenization, still at temperature above melting point of lipid, to produce a hot o/w nanoemulsion. The homogenization process can be repeated several times: in most cases three to five homogenization cycles at 500–1,500 bar are performed. Increasing the homogenization pressure or the number of cycles often results in an increase of the particle size due to particle coalescence occurring for the high kinetic energy of the particles.

Obtained o/w nanoemulsion is then cooled down to room temperature where SLN solidify.

In CHPH, the drug-loaded lipid is rapidly cooled using liquid nitrogen or dry ice to obtain a solid solution which is then ground (milled) to obtain microparticles, approximately in the range 50–100 μm, then the solid microparticles are dispersed in a cold surfactant solution to obtain a pre-suspension. The pre-suspension is subjected to high-pressure homogenization at room temperature or below to obtain solid lipid nanoparticles: usually larger particle size and broader size distributions are obtained compared to hot homogenization.

Cold homogenization minimizes the thermal exposure of the drug, although it does not avoid it, due to dissolution of drug in melted lipid in the initial step.

Main reported disadvantages of HPH are: high energy-intensive operating conditions and potential thermal degradation of drug and excipients during production process.

Warm Microemulsion Technique

Microemulsions are transparent (clear), optically isotropic and thermodynamically stable systems. They are dispersions of two immiscible liquids (oil and water) stabilized by surfactants and optionally by cosurfactants.

Depending on the composition of the system different structures can be formed:

1. Oil dispersion in water medium – o/w microemulsion
2. Water dispersion in oil medium – w/o microemulsion
3. Bicontinuous structure

In o/w and w/o microemulsions, the dispersed phases are in shape of very small drops (10–100 nm) stabilized by an interfacial film of surfactant and cosurfactant molecules.

In bicontinuous structures, regions of water and oil are interdispersed with no spherical geometry: it occurs when the amount of water and oil are comparable.

The thermodynamic stability of microemulsion is due to the very low interfacial tension which allows the spontaneous formation of microemulsions without high energy input.

Warm microemulsions obtained with melted lipids have been used as precursor for the preparation of SLN: after their preparation at temperature ranging from 55 °C to 85 °C depending on melted lipid used, they are then dispersed in cold aqueous medium where melted lipid drops solidify into SLN [2].

The hydrophobic drug is dissolved in mixture composed of melted lipid and surfactant maintained at warm temperature, the cosurfactant is dissolved in water and the obtained solution is heated at the same temperature of the lipid mixture; cosurfactant aqueous solution is then added to the mixture composed of lipid, drug and surfactant and by mild mixing an o/w microemulsion is obtained and successively dispersed in cold water at 2–3 °C under mechanical stirring.

Also hydrophilic drugs have been loaded into SLN by water in oil in water (w/o/w) double microemulsion technique: water, where hydrophilic drug is dissolved, constitutes the internal phase of w/o microemulsion, and the melted lipid (55–85 °C) the external phase; the w/o warm microemulsion is then added to an external aqueous solution of surfactant and cosurfactant maintained at the same temperature and after mixing w/o/w microemulsion is obtained and dispersed in cold water to obtain SLN.

The SLN dispersion, obtained by warm microemulsion technique, is usually purified by tangential ultrafiltration in order to remove surfactant and cosurfactants not incorporated into SLN, and to concentrate dispersion when needed.

Advantage of the production of SLN by warm microemulsion include: low mechanical energy input, flexibility of interphase composition that allow surface functionalization, regular spherical shape. Potential drawbacks are thermal exposure, which can cause degradation of drug and of excipients, use of higher amount of surfactant and cosurfactant, which sometimes need to be removed.

High Shear Homogenization and Sonication

Following this process, drug dispersed in melted lipid phase is homogenized by rotor-stator homogenizer with a hot surfactant water solution to obtain nanoemulsion that is then cooled to form SLN.

Process parameters that can affect particle size are: emulsification time, stirring rate, and cooling conditions.

Homogenization can be used in association with sonication: melted lipid phase loaded with drug is homogenized with a hot surfactant solution to obtain a coarse emulsion that is then ultrasonicated to obtain final nanoemulsion before SLN are formed upon its cooling to room temperature [3].

Main possible disadvantages of this method include high energy input, broad particle size distribution, and potential damage of sensitive biomolecules.

When sonication is used, potential metal contamination and temperature increase have to be considered.

Solvent Emulsification-Evaporation Method

In this method, drug, lipid matrix, and emulsifier are dissolved in a water-immiscible organic solvent which is then emulsified in an aqueous phase containing a water-soluble cosurfactant and homogenized. Upon evaporation of the solvent under reduced pressure, SLN are formed by precipitation of the lipid in the aqueous medium [4].

Parameters that affect mean particle size of nanoparticles obtained by this method are: surfactant/cosurfactant blend and lipid concentration in the organic phase.

Advantages of this technique are: avoidance of heat application, production of small size nanoparticles that can be sterilized by filtration through 0.2 μm pores.

The major possible disadvantage of this method is the presence of organic solvent residues which can cause toxicity.

Solvent Emulsification-Diffusion Method

The process consists in dissolving the lipid in a partially water-soluble solvent (previously saturated with water) at room or controlled temperature, depending on solubility of lipid in the solvent. This organic phase is emulsified with an aqueous solution (saturated with solvent) containing the stabilizing agent and maintained at the same temperature, by a rotor-stator homogenizer. This o/w emulsion is then diluted with water maintaining a constant stirring and controlled temperature for promoting the diffusion of solvent of the internal phase toward the external phase, causing lipid aggregation and precipitation in form of SLN.

Depending on its boiling point, solvent can be eliminated from SLN dispersion under reduced pressure or by washing (ultrafiltration).

The selection of the water-miscible solvent and the stabilizers are critical parameters to obtain lipid particles in the nanometric range: usually solvents with high water miscibility and stabilizers able to form stable emulsions are preferred.

It is possible to reduce the particle size by increasing the process temperature, the stirring rate, the amount of stabilizer, and by lowering the amount of lipids [5].

This technique is efficient and easy to be implemented, so feasible for industrial upscaling: no high energy is required and low physical (thermal and mechanical) stress is applied.

Possible drawbacks of this method are solvent residues which need to be cleaned up and large dilution produced to obtain diffusion of solvent which needs further concentration steps.

Solvent Injection Method

In this procedure, solid lipid is dissolved in a water-miscible solvent or solvents mixture, and then rapidly injected through an injection needle into a stirred aqueous phase with or without surfactants.

The SLN production by this technique is based on the rapid diffusion of the solvent across the solvent-lipid and solvent-water interface, causing precipitation of nanoparticles.

Two simultaneous effects contribute to the effective formation of SLN:

1. Gradual solvent diffusion out of lipid-solvent droplets into water causes reduction of droplet size and simultaneously increases lipid concentration
2. Diffusion of pure solvent from the lipid-solvent droplet causes local variations in the interfacial tension at droplet surface, inducing reduction of size of droplets

In this process particle size of SLN can be influenced and controlled by variation of process parameters such as injected solvent, lipid concentration, injected volume of solvent, lipid concentration in the solvent phase and viscosity of the aqueous phase [6].

This technique is simple and fast, without the need of sophisticated equipment; possible disadvantage of this method is the use of organic solvent which has to be considered for pharmaceutical applications and for this reason a complete removal of the solvents by ultrafiltration, evaporation, or freeze-drying is required.

Coacervation Method

This process allows to obtain SLN made of fatty acids by acidification of micellar solution of their alkaline salts: fatty acids sodium salts are dispersed in a polymer water solution (use of steric stabilizer polymer is essential to avoid particle aggregation) and heated under stirring just above the Krafft point of the fatty acid sodium salt to obtain a clear solution (Krafft point is the temperature at which the solubility of sodium soap increases dramatically and the solution becomes isotropic and transparent): an acidifying solution is then added drop-wise until pH about 4 is reached and the obtained suspension is then cooled down to 15 °C in a water bath under stirring.

Parameters that can influence size of nanoparticles and polydispersity index are: type and molecular weight of polymer and lipid concentration in the dispersion [7].

Although average dimensions are reported in a bigger range of 260–500 nm, this process presents several advantages like absence of solvents and need for common apparatus: the method is feasible and suitable for laboratory production and it is easy to be scaled up; studies are ongoing for loading of different drugs.

Membrane Contactor Method

This method employs a cylindrical membrane module: aqueous phase containing a surfactant is circulated in internal channel of membrane and melted lipid is pressed through pores of membrane into internal water flow allowing the formation of small droplets which are swept away by the aqueous phase; water is maintained at lipid melting temperature. SLN are then formed by cooling the preparation to room temperature.

The membrane contactor allows the preparation of SLN with a lipid phase flux between 0.15 and 0.35 $m^3/h/m^2$ [8].

The advantages of this new process are its simple use, the control of the SLN size by an appropriate choice of process and membrane parameters, and its scaling-up abilities. Drug loading and surface modification, for targeting needs, have to be fully assessed and developed.

Characterization of SLN

Particle Size, Size Distribution and Shape

Particle size, shape and stability over time may influence several important pharmaceutical features of SLN such as drug release, suitable administration route (e.g., particle size is a critical issue for intravenous administration, where particle size must not aggregate or enlarge by protein binding to exceed diameter of blood vessels) and in vivo biodistribution.

Particle size and morphology of SLN are mainly affected by method of production and composition (lipid matrix, type and amount of emulsifiers, viscosity of aqueous phase, etc.).

Main techniques commonly used for characterization of SLN are summarized below.

Photon correlation spectroscopy (PCS) also known as dynamic light scattering (DLS), is widely used method for the measurement of particle size in range from nanometer to few micrometers.

This technique determines hydrodynamic diameter of particles by measuring fluctuation of the intensity of scattered light which is due to Brownian motion of particles.

Instruments usually report hydrodynamic diameter (Zaverage) and polydispersity index (PI), which give estimation of width of particle size distribution.

Electron microscopy methods are very useful techniques for determining shape, morphology and size of lipid nanoparticles.

The most commonly reported method for morphological characterization is *transmission electron microscopy* (TEM): different techniques can be applied for sample preparation (negative staining, freeze-fracture, sample vitrification (cryo-TEM)) and can provide different information about colloidal lipid structures; sample preparation is a crucial point as it can lead to structural alterations of the sample that need to be taken into consideration. As an example, TEM micrographs of SLN (obtained by warm microemulsion method) are reported in Fig. 2 [9]: images show SLN in water dispersion before administration and same SLN in biological fluids coming from in vivo sampling after administration.

Other methods to determine SLN particle size and morphology are based on *scanning electron microscopy* (SEM).

Nonconductive samples, such as lipid-based nanoparticles, may be visualized without metal coating using specialized SEM instrumentation such as *environmental SEM* (*ESEM*) in which the sample is placed in a internal chamber at higher pressure than vacuum: positively charged ions generated by beam interactions with the gas, present in the chamber, help to neutralize the negative charge on the surface of sample.

In *field emission SEM* (*FESEM*), the electron beam is produced with a cold cathode field emitter instead of a thermoionic emitter (tungsten filament heated with an electric current) as in conventional SEM: this allows for much higher resolution showing particle shape and size. As an example, images obtained from two different formulations of SLN, produced by warm microemulsion technique, are reported in Fig. 3.

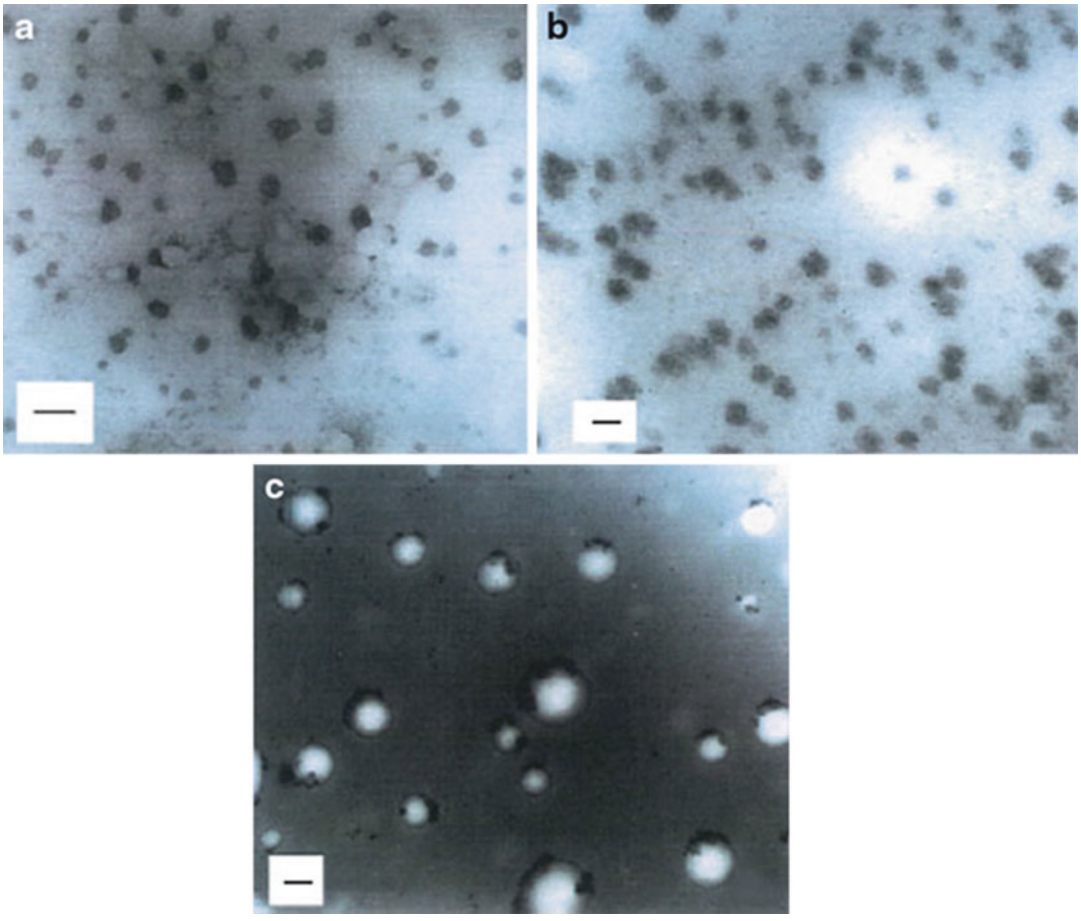
Atomic force microscopy (AFM) is another technique used to determine particle size and shape. AFM can be operated in different modalities depending on the specific application requirements: sample does not require any special treatments (such as coating) and technique also allows to work in ambient conditions, even in liquid environment. As an example, AFM characterization of SLN produced by solvent-diffusion method is reported in Fig. 4 [10]: SLN produced in miniemulsion used as nanoreactor system is compared with SLN obtained by conventional method.

Surface Charge

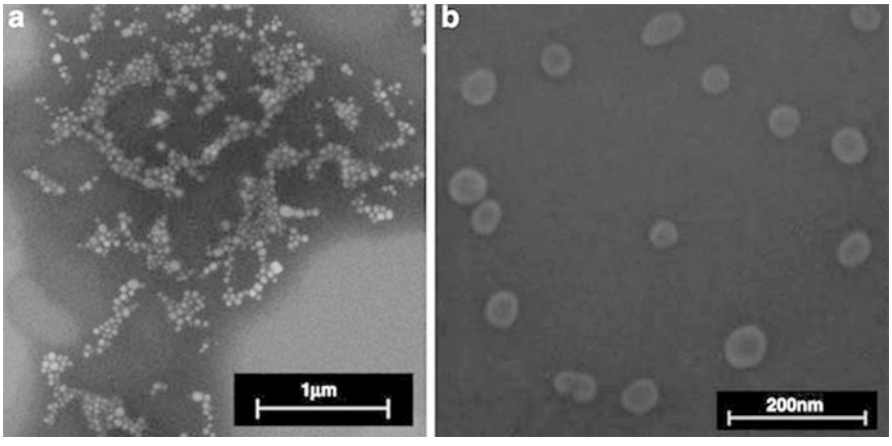
Nanoparticles dispersed in a liquid medium are surrounded by an electrical double layer composed by an inner layer (stern layer) where the ions are strongly bound to particle surface and an outer layer (diffuse layer) where ions are less firmly associated with the particle.

Within the diffuse layer there is a boundary (shear plane) inside of the ions and the particle form a stable entity: when the particle moves, only ions within the boundary move with the particle, while ions beyond the boundary remain in the bulk dispersant.

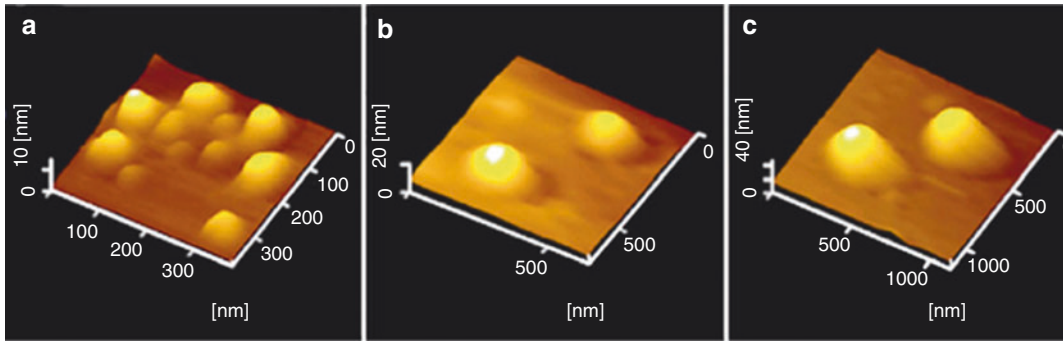
Zeta potential (Z potential) is the electric potential at this boundary and it gives measureof surface charge. It is used to assess stability of colloidal systems as it can express tendency of particles to repulse or to attract: higher values of zeta potential (absolute value > 30 mV) usually imply more stable dispersions due to electrical



Solid Lipid Nanoparticles (SLN), Fig. 2 TEM micrographs of SLN obtained by warm microemulsion method and containing 2.5 % of tobramycin in different physiological media: (**a**) aqueous dispersion before administration to rats, (**b**) lymph, and (**c**) plasma after duodenal administration to rats (bar = 100 nm) [9]



Solid Lipid Nanoparticles (SLN), Fig. 3 FESEM image of SLN obtained by warm microemulsion method: (**a**) SLN with lipid matrix of Stearic acid, (**b**) SLN with lipid matrix of cholesteryl butyrate



Solid Lipid Nanoparticles (SLN), Fig. 4 Atomic force microscopy images of SLN prepared by solvent diffusion method in a nanoreactor system. (**a**) Blank SLN, (**b**) SLN with 10 % of clobetasol propionate charged (**c**) SLN prepared by conventional method with 10 % of clobetasol propionate charged [10]

repulsion, while low values can indicate possible colloid instability which could lead to aggregation.

Z potential affects not only dispersion stability but also interaction with the biological surroundings (interaction with proteins and cells) and the in vivo fate of nanoparticles (biodistribution); it depends on the composition of the particle and on its surrounding medium: pH, electrolyte concentration, and concentration of components of the formulation.

Z potential is determined by measuring the velocity of the particles dispersed in a liquid medium under the influence of an applied electric field (electrophoresis measurements): charged particles move toward the electrode under applied electric field with velocity dependent on the strength of electric field or voltage gradient, the dielectric constant and the viscosity of the medium, and the Z potential of the particle.

Electrophoretic mobility, defined as the velocity of the particle in a unit electric field, is related to Z potential by Henry equation and is usually measured by light scattering technique.

Crystallinity, Polymorphism, Structure, and Stability

Crystallinity and polymorphic transitions of lipids in the dispersed state may differ from bulk material due to the small particle size of colloidal system, the presence of emulsifiers used for the stabilization of SLN and the preparation method. In thermal analysis SLN dispersions usually show lipid melting point slightly shifted to lower temperatures and broader melting peak than the bulk lipid, polymorphic transitions are generally faster in SLN than in the bulk material [1, 11].

Crystallization and polymorphic behavior of SLN are correlated with drug incorporation and drug release: highly ordered crystal lattices (monoacid triglycerides) in SLN matrix can lead to drug expulsion, while less ordered crystal lattice (mixture of mono-, di-, and triglycerides), because imperfections which provide space to accommodate the drug, can allow better drug incorporation.

Polymorphic transitions may cause alterations of lipid packing and thus of the internal structure of the nanoparticles, which may have negative consequences for drug loading. The course of polymorphic transitions depends on the type of lipid matrix and can be modified by other components of the dispersions [11].

As an example, triglycerides crystallize in three main polymorphic forms: α, β', and β; the β form is highly ordered with high thermodynamic stability, the α form is a less ordered form with low thermodynamic stability, the β' form has

intermediate characteristics between α and β forms. SLN based on triglyceride matrix tend to crystallize in the metastable α form: generally thermodynamically unstable configurations allow lipid molecules to have higher mobility causing lower density and a higher capability to incorporate drug molecules, but during storage the α form transform via β′ form in the more thermodynamically stable configuration β, characterized by higher packing density which can cause expulsion of the drug form structure.

SLN may change their shape during polymorphic transitions: as an example, tripalmitin SLN have a spherical shape in α form, but they have a platelet shape in the β form. During storage the polymorphic transition from the α to the β configuration is connected with an increase of the particle surface due to preferred formation of platelets; surfactant molecules could not provide any more complete coverage of the lipid surface and particle aggregation can occur, leading to gelation (irreversible transformation of SLN dispersion into a viscous gel) [1].

SLN can display a lower crystallization tendency than the bulk material and may not crystallize properly after preparation and storage, at a temperature below melting point of lipid, forming super-cooled melts, which are liquid lipid nanoemulsions and not solid lipid dispersion. This phenomenon is particularly pronounced in SLN made from short-chain monoacid triglycerides such as trimyristin and trilaurin, and it is mainly due to size dependency of crystallization process that requires a critical number of nuclei to start, which cannot be reached in small droplets: tendency to formation of super-cooled melts increases with decreasing of droplet size [1, 11].

Hence, the physical state is a very important parameter which affects performance of SLN as drug delivery system: the techniques usually used to investigate physical state of SLN are *differential scanning calorimetry* (DSC) and *X-ray diffraction* (XRD) techniques.

DSC is a thermal analysis technique which measures the difference in heat flow between the sample and a reference when they are both subjected to the same controlled temperature program: DSC quantifies the enthalpy changes during endothermic and exothermic process and it is a useful technique to determine melting point, melting enthalpy, degree of crystallinity, glass transitions of amorphous material, and to identify different polymorphic forms.

In XRD techniques, an X-ray beam is focused on the sample, the scattered X-rays interfere with each other giving particular diffraction patterns: crystalline materials display many diffractions bands which depend on the disposition of the atoms within the unit cell of the crystalline lattice, while amorphous compounds present more or less a regular baseline.

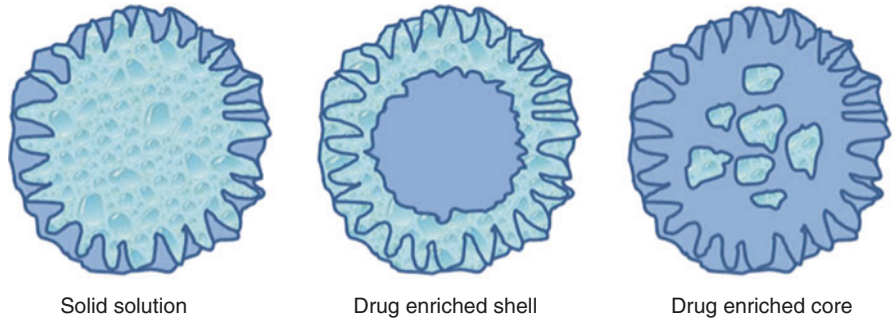
XRD techniques allow to differentiate between crystalline and amorphous material and to identify different polymorphic forms in SLN dispersions: the most commonly used are *powder X-ray diffraction* (PXRD), *small angle X-ray scattering* (SAXS), and *wide angle X-ray scattering* (WAXS).

Nuclear magnetic resonance (NMR) is another useful technique to investigate structure of SLN for surfactant distribution on particle surface, drug distribution in the particle and potential formation of super-cooled melts.

Due to different chemical shifts it is possible to attribute the NMR signals to particular molecules or their segments (functional groups); NMR active nuclei of interest are ^{1}H, ^{13}C, ^{19}F, and ^{31}P. Width and amplitude of NMR signals can be related to liquid and semisolid/solid state of nuclei: nuclei in the liquid state give sharp signals with high amplitude, while nuclei in the semisolid/solid state give weak and broad signals [1].

Drug Incorporation

A broad range of drugs, mainly with hydrophobic properties, has been already incorporated into SLN, although the crystalline nature of the lipid matrix can offer limited space for drug incorporation inside the particle core [11].



Solid Lipid Nanoparticles (SLN), Fig. 5 Schematic representation of drug incorporation models for SLN

Drug loading capacity of SLN is defined as the percentage of drug incorporated into nanoparticles referred to the total weight of the lipid phase or to the content of dispersed material; value of this parameter is usually in range 1–20 %.

Drug entrapment efficiency is defined as the percentage of drug incorporated into nanoparticles referred to the total drug added for the SLN preparation; values of this parameter is usually relatively high (80–99 %).

The distribution (localization) of the drug within SLN structure (Fig. 5) can vary considerably and three *drug incorporation* models have been proposed [12]:

1. *Solid solution model*. The drug is molecularly dispersed in the solid lipid matrix (SLN matrix as a solid solution).
2. *Drug-enriched shell model*. The drug is concentrated in the outer shell of the SLN. This situation can be explained by a higher solubility of the drug in the aqueous-surfactant outer phase at increased temperature in the production process; during cooling, lipid in the core starts to solidify and becomes less/not accessible for the repartition of the drug into the lipid inner phase leading to enrichment in the particle shell.

   High temperatures employed in the preparation process can increase solubility of the drug in the aqueous-surfactant phase promoting drug localization at the surface region.
3. *Drug-enriched core model*. The drug is concentrated in the core of the lipid particle and it is surrounded by a lipid shell. This situation can be explained by a precipitation of the drug before the lipid crystallizes during cooling: this takes place preferentially when drug concentration in the lipid at process temperature is at its saturation solubility, so during cooling a super saturation and subsequent drug precipitation are achieved

Drug-loading capacity of SLN is affected by different factors:

1. Solubility of the drug in the melted lipid. High solubility is a prerequisite to obtain sufficient drug loading; the amount of drug that can be dissolved in the lipid formulation may exceed the pure value of its solubility in lipid alone and a greater solubility in the whole formulation suggests that the drug can localize both in inner lipid phase and in external part on SLN (interfacial region) where surfactants are present. Solubility of drugs in the lipid can be improved by preparing lipid prodrugs such as stearic or palmitic acid derivatives (lipid drug conjugates).
2. Physical structure of solid lipid matrix. Complex lipids which form less ordered crystals with many imperfections can favor drug incorporation.

3. Polymorphic state of lipid matrix. Lipid transformation in more stable form reduces the number of imperfections in crystal lattice and can determine drug expulsion.

Determination of drug-loading capacity can be usually performed by separation of free drug from the dispersion medium by employing different techniques such as ultrafiltration (tangential or centrifugal), ultracentrifugation, gel filtration chromatography, and dialysis.

Drug Release

Drugs incorporated into SLN are released by degradation and surface erosion of the lipid matrix and by diffusion of drug molecules through the lipid matrix. SLN are composed of physiological lipids for which metabolic pathways exist in the body: most important enzymes involved in in vivo SLN degradation are lipases, which are present in various organs and tissues [1].

Drug release from SLN is mainly affected by localization of the drug [12]:

1. Localization of the drug within the core of solid lipid matrix offers the possibility to obtain a prolonged drug release.
2. Localization of drug molecules on particle surface often leads to burst effect (fast initial drug release). SLN can show a biphasic drug release profile: an initial burst release, due to the drug localized at the surface, is followed by a more gradual release due to the drug localized in the lipid matrix.

Possible presence of alternative colloidal species always has to be taken into account for drug release characterization: stabilizing agent cannot be localized exclusively on the lipid surface as expected, but also in the aqueous phase forming micelles, mixed micelles or liposomes that can solubilize the drugs and constitute alternative drug incorporation sites [1].

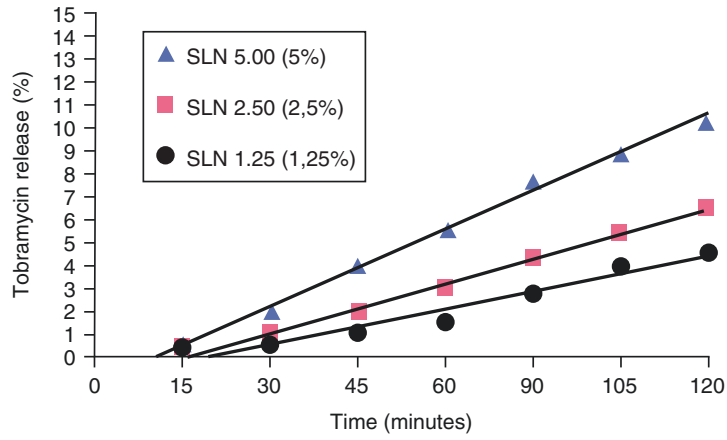
Composition of SLN is other important parameter that can affect drug release: in vitro experiments indicate that SLN show different degradation rates by lipases as a function of their composition (kind of lipid matrix and emulsifier). Degradation of triglycerides SLN show dependence to the length of the fatty acid chain: longer fatty acid chain shows slower degradation, while degradation of SLN made of waxes (cetylpalmitate) is slower compared to glyceride matrices [1].

Emulsifiers containing polyethylene glycol (PEG) chains can reduce SLN degradation by a hindering effect which reduces the anchoring of the enzyme on the SLN surface [1].

Since SLN are degraded by surface erosion, in SLN with small size and higher surface area, drug release is expected to be more rapid; nonspherical particles, such as thin platelet shape, are characterized by larger surface area and require shorter time for degradation and drug diffusion to particle surface [11].

Drug release can be studied by different techniques such as dialysis membranes (flat or bag) and Franz diffusion cells. Release kinetics is affected by in vitro release conditions: sink or non-sink conditions, release medium, dilution of nanoparticles dispersion that has to be considered in particular if the formulation is intended for oral and intravenous administration; release experiments have to be performed mimicking in vivo conditions in order to avoid distorted release profile [11].

As an example of in vitro study [9], drug release was investigated, working in sink conditions, for three types of SLN incorporating different percentage of tobramycin (1.25 %, 2.50 %, and 5.00 %) by using a multicells rotating tool, where donor and acceptor compartments were separated by double hydrophilic/hydrophobic membranes: results showed that release kinetic was pseudo-zero order for all three types of SLN and the amount of drug released was higher for Tobra-SLN 5.00 % than for Tobra-SLN 2.50 % which was higher than for Tobra-SLN 1.25 %, showing that in vitro release can be modified by changing amount of drug loaded into SLN and consequently their physical characteristics (Fig. 6) [9].



Solid Lipid Nanoparticles (SLN), Fig. 6 In vitro release – percentage of tobramycin diffused through a double membrane versus time for SLN loaded at three different percentages of tobramycin [9]

Routes of Administration/Applications

SLN have been tested for drug delivery applications by different routes of administration including intravenous, oral, ocular, topical dermal, transdermal, and pulmonary.

Parenteral Administration: Intravenous Route (IV)

Overview

The evaluation of SLN interaction with blood and other tissues is very important for IV administration: many studies have been performed to assess toxicity of SLN both in in vitro and in vivo tests.

Cellular toxicity of SLN has been investigated using several cell lines: in vitro experiments demonstrate that toxicity is dependent on composition of SLN (nature of lipid matrix and type of surfactant used) and on concentration of SLN in the culture medium [13].

In vivo toxicity studies confirmed that SLN are well tolerated in living system after their iv administration [13].

In order to obtain SLN suitable for parenteral administration, pharmaceutically acceptable excipients must be employed, and sterility must be assessed. Studies showed some SLN formulations can be sterilized by sterilizing filtration (such filtration is applicable for particles with size <0.2 μm), some other can be sterilized by autoclaving generally at 121 °C for at least 15 min (SLN melt during autoclaving and recrystallize during cooling); physicochemical properties of SLN formulation should not change during the sterilization process. Autoclaving temperature can promote chemical degradation and can affect physical stability of SLN because their structure can be lost when particles melt and recrystallize in noncontrolled way; performed studies show that critical parameters for SLN sterilization by autoclaving are temperature and timing of sterilizing process and composition of formulation [1].

Generally, all in vivo studies performed showed drug-loaded SLN change pharmacokinetic parameter of carried drug when compared to reference drug solution: much higher mean residence time (MRT) and area under the curve (AUC) induced by SLN formulation mean higher drug availability and consequent possible higher efficacy.

Performed studies confirmed as well that important limiting factor for IV administration of SLN is the uptake of particles by macrophages of reticuloendothelial system (RES). SLN are recognized as foreign substances and quickly removed from blood circulation. Colloidal particles interact with blood plasma proteins called opsonins which are adsorbed on particle surface (opsonization): opsonins mediate RES recognition by interacting with specific membrane receptors of macrophages. The capacity of nanoparticles to avoid opsonization and

consequent macrophages uptake depends on size of particles, on their surface charge, and hydrophobic characteristics.

In order to avoid RES recognition, SLN surface can be functionalized with different molecules having hydrophilic and flexible chains which form a hydrophilic steric barrier that cover and protects nanoparticles from interaction with plasma proteins and prolongs their blood circulation time. Such sterically stabilized nanoparticles (now reported as *stealth*) are mainly obtained by using lipid derivatives of polyethylene glycol (PEG): amount of stealth agents and length of their chain can affect degree of surface coverage and rate of uptake by macrophages.

Macromolecular drugs and colloidal drug carriers such as SLN, liposomes, and polymeric nanoparticles, can accumulate preferentially in tumor tissue, a phenomenon called enhanced permeation and retention effect (EPR effect), due to the presence of a discontinuous endothelium and to the lack of efficient lymphatic drainage in the tumor tissue: by reducing RES uptake, nanoparticles can further passively accumulate inside such tumor tissues (*passive targeting*).

In order to further increase drug accumulation in target tissue or to produce higher and more selective therapeutic activity, specific functionalization of colloidal drug carriers has been developed (*active targeting*) by decorating SLN surface with ligand molecules able to recognize specific site on target cells inducing internalization of carrier after binding. Targeting molecules include: monoclonal antibodies or fragments, peptides, glycoproteins, and receptor ligands.

Main Applications of SLN by Parenteral/IV Administration

Cancer Therapy

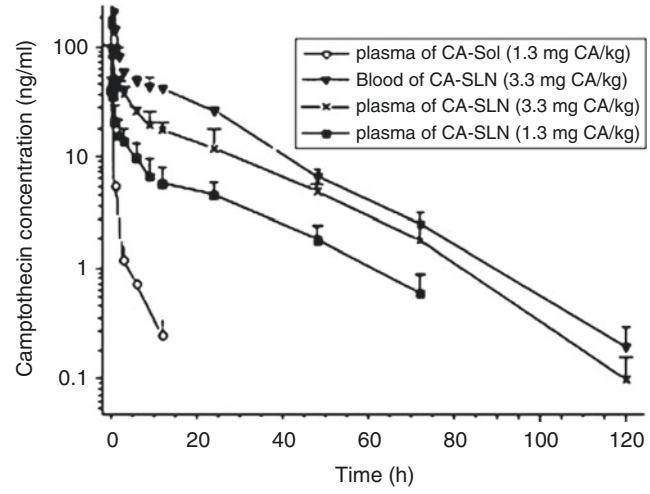
SLN have been investigated as drug delivery system by IV administration mainly for cancer treatment: several antitumor drugs have been incorporated into SLN such as doxorubicin, camptothecin, paclitaxel, etoposide, idarubicin, vinorelbine, mitoxantrone, 5-fluorouracil, and their in vitro and in vivo distribution and efficacy have been evaluated [13].

SLN are considered suitable potential carriers in oncology for ability to incorporate antitumor drugs with different chemical properties for improvement of drug stability (drug protection), enhancement of drug efficacy, reduction of drug toxicity (lowering of side effects), improvement of pharmacokinetics parameters, passive targeting, and the possibility to reach difficult districts like the brain [2, 13].

One of the first in vivo studies on these carriers in oncology has been performed on camptothecin (CA) loaded into SLN prepared by high pressure homogenization method and administered intravenously into mice: the concentration of camptothecin was determined in various organs and compared to a camptothecin control solution [13]. The results showed that in tested organs, the area under the curve (AUC) versus dose and the mean residence times (MRT), pharmacokinetic parameters relevant to drug bioavailability, of CA-SLN were much higher than those of CA-Solution (Fig. 7) [13].

As another example, in vivo tests have investigated pharmacokinetics and tissue distribution of doxorubicin incorporated in both stealth and not stealth SLN prepared by warm microemulsion technique and injected intravenously into rabbits [2]. AUC increased as a function of the amount of stealth agent present in SLN and doxorubicin was still present in the blood 6 h after the injection of both stealth and non-stealth SLN, while none was any more detectable after IV injection of doxorubicin solution.

The pharmacokinetics and tissue distribution of doxorubicin SLN were also studied in rat animal model after intravenous administration and compared with commercial solution of doxorubicin [2]: doxorubicin was still present in the blood 24 h after the injection of stealth and non-stealth SLN, while it was not detectable after the injection of the commercial solution. The results confirmed the prolonged circulation time of SLN, in particular SLN decreased the concentration of doxorubicin in the heart indicating lower cardiotoxicity compared to doxorubicin solution.



Solid Lipid Nanoparticles (SLN),
Fig. 7 Concentration-time curves of camptothecin after IV administration of CA-SLN with doses of 1.3 (■) and 3.3 (x) mg CA/kg in plasma and 3.3 (▼) mg CA/kg in blood, and CA-SOL with a dose of 1.3 (○) mg CA/kg in plasma. Results represent means ± SD of four animals [13]

Diseases of the Central Nervous System (CNS)

Pharmacological treatment of diseases of the CNS, such as brain tumor, neurological, and neurodegenerative diseases, is limited by the presence of the *blood–brain barrier* (BBB) that restricts enormously the transport of many important drugs from the blood into the brain. The BBB is formed by the endothelial cells of the cerebral capillaries which differ from endothelial cells in the rest of the body for the absence of fenestrations, for the presence of particular and more extensive tight junctions, and for minimal pinocytic vescicular transport.

In order to evaluate transport of SLN across the BBB, both stealth and non-stealth radiolabeled SLN, prepared by warm microemulsion method, were injected intravenously into rats [2], and tissue distribution was monitored for 60 min: radioactivity in liver and in lung was lower for stealth SLN than for non-stealth SLN confirming difference in their uptake. Both types of SLN were detected in brain and cerebrospinal fluid, although at low percentage.

SLN incorporating baclofen have been investigated as new pharmaceutical preparation of this drug which is used in treatment of spasticity [2]: SLN were prepared by warm microemulsion method and were injected intraperitoneally to rats at increasing dosage. As an important finding, drug effects were detectable with lower doses of baclofen when loaded into SLN, in comparison with the needed amount of baclofen when in solution.

In vivo results showed a good correlation with plasma and tissue concentration of baclofen: after 2 and 4 h, only baclofen-SLN produced detectable baclofen plasma concentrations, while 2 h after the administration of baclofen solution, the amount of baclofen in plasma was undetectable. Moreover, baclofen concentration in the brain 2 h after SLN administration was almost double than after baclofen solution, suggesting that baclofen may pass the BBB in much higher amount when formulated in SLN.

In order to improve brain uptake, SLN modified with specific ligands on surface have also been tested (active targeting): the presence in the BBB of receptor-mediated transport systems for endogenous molecules can be exploited to gain access to the brain.

Transferrin-conjugated SLN have been investigated for their ability to target quinine hydrochloride to brain [13]: biodistribution studies showed that quinine hydrochloride concentration in brain was significantly higher in case of transferrin-conjugated SLN as compared to that of quinine-SLN not functionalized and of quinine hydrochloride solution.

Imaging

Imaging technologies are important tools for diagnosis of diseases and for monitoring of therapies.

SLN have been proposed as carriers for hydrophobic imaging agents in order to increase concentration of the agent in the target tissue.

SLN have been investigated as carrier of contrast agents for magnetic resonance imaging (MRI) such as iron oxides nanoparticles (NP) [2]. SLN loaded with iron oxides NP were prepared by warm microemulsion method and were studied in in vitro and in vivo tests and compared to Endorem®, commercial preparation of iron oxides NP. Iron oxide-SLN showed in vitro relaxometric properties similar to those of Endorem® and after intravenous administration to rats showed to have slower blood clearance than Endorem®.

Gene Therapy

A potential approach for the treatment of human genetic disorders is gene therapy. This is a technique whereby the absent or faulty gene is replaced by a working gene, so that the body can make the correct enzyme or protein and consequently eliminate the root cause of the disease. Process of introducing nucleic acids into cells is generally called transfection.

Polynucleotides molecules (DNA or RNA) are large, hydrophilic, and negatively charged molecules. They are very labile in biological environment and their spontaneous entry inside cells is a very inefficient process.

Nucleic acids can be delivered by two main classes of vectors: viral vectors and nonviral vectors [14]. Although viral vectors are very efficient in nucleic acid delivery, their immunological risk is high (insertional mutagenesis, adverse immunogenic responses, inflammatory reactions) and there is limitation on dimension of nucleic acid to be incorporated: for these reasons nonviral carriers can offer preferable alternative. Most important tested nonviral vectors are liposomes, complexes made by DNA or RNA and cationic polymers or cationic lipids, and SLN [14].

The first obstacle for systemic delivery of DNA or RNA molecules is the extracellular environment, extreme pH, proteases and endonucleases, and immune defense.

The second barrier is cellular membrane; nucleic acids alone are unable to cross it because of their high negative charge.

One of the main routes of internalization of vectors carrying RNA or DNA is endocytosis: the internalized particle exists in endosomes, compartments of the endocytic membrane transport pathway, that either fuse with lysosomes, the main hydrolytic compartment of the cell where RNA or DNA are degraded losing their activity. Therefore, escape from endosomes is crucial for efficient transfection. Certain lipids have the ability to destabilize endosomal membranes favoring the escape of nucleic acids from this compartment: those lipids are called fusogenic lipids (e.g., dioleoyl phosphatidylethanolamine DOPE) and they are usually included in SLN formulation for this reason.

For their interference with lysosomes, lysosomotropic agents, such as cloroquine, are reported as well to enhance gene expression [14].

Another barrier against DNA transfection is nuclear envelope, containing pore regulating passive transport for molecules with maximum mass of 70 kDa or mean diameter size of 10 nm.

Cationic SLN, with positive surface charge, carry the nucleic acids by means of electrostatic interaction: different in vitro studies showed their efficacious transport of nucleic acid through cellular barriers into the cytoplasm or nucleus [12, 14]

Cationic SLN can be obtained by using different cationic lipids such as: dioleoyl trimethylammonium-propane (DOTAP), cetylpyridinium chloride, dimethyldioctadecylammonium bromide (DDAB), DC-cholesterol.

The cationic SLN are incubated with negatively charged nucleic acids; electrostatic interactions occur resulting in the formation of SLN-DNA or SLN-RNA complexes.

Positive charge of SLN-DNA/SLA-RNA surface is usually preferable for in vitro transfection, as it helps to promote the cellular uptake of the complex because of negative charge of cell membranes.

On the other side, use of some cationic lipids for in vivo delivery is limited by their toxicity [15].

The capacity of SLN-DNA vectors to in vivo transfect after intravenous administration in mice has been evaluated: as an example, cationic SLN were prepared by solvent emulsification evaporation method and by using DOTAP as cationic lipid, SLN-DNA vectors were obtained by mixing SLN with the plasmid pCMS-EGFP, which encodes the enhanced green fluorescent protein (EGFP). The intravenous administration in mice led to transfection in hepatic tissue and spleen, protein expression was detected from the third day after administration and it was maintained for at least 7 days; the results showed the capacity of SLN-DNA vectors to induce expression of foreign proteins in the spleen and in the liver assessing the potential of SLN for gene therapy [15].

SLN have been formulated also to incorporate nucleic acids inside their solid lipid matrix: this kind of nanoparticles are mainly prepared by (w/o/w) double warm microemulsion technique.

Efficacy of SLN incorporating vascular endothelial growth factor antisense oligonucleotide (VEGF-AS-ODN) to downregulate VEGF expression has been evaluated in rat glioma cells (in vitro) and in experimental murine model of glioma (in vivo) [2]: both experiments demonstrated that cellular VEGF expression was significantly reduced in tumor cells with SLN carrying VEGF-AS-ODN.

Oral Administration

SLN have been studied for oral administration of several drugs such as cyclosporin A, clozapine, tobramycin, idarubicin, and apomorphine.

SLN can improve oral bioavailability of drugs by different mechanisms [12]:

1. Protection of incorporated drugs from gastrointestinal fluids and enzymatic degradation
2. Transportation into lymphatic system avoiding first-pass metabolism
3. Absorption enhancing effect of lipids: lipids can promote the absorption of poorly water soluble active compounds
4. Bioadhesion of nanoparticles (due to their small size) to the intestinal wall increasing the residence time in intestinal tract
5. Effect of surfactants which may contribute to increase permeability of the intestinal membrane or to improve affinity between lipid particles and intestinal membrane

Gastrointestinal uptake of radiolabeled SLN administered to duodenum of rats by canula was studied [2]. SLN were observed in the lymph by using electron microscopy; the radioactivity data confirmed targeting of particles to lymph and blood.

In other in vivo study, in order to evaluate gastrointestinal absorption of drugs incorporated into SLN, tobramycin was selected as model drug because it is not absorbed by the gastrointestinal tract. SLN containing different percentages of tobramycin were administered in rats by canula to duodenum: the time-concentration curves showed different profiles (Fig. 8) [9] and pharmacokinetics parameters varied considerably among the three types of tobramycin-SLN. Possible reasons for the different behavior are number of SLN administered, particle size, total surface area, and drug concentration in each nanoparticle [9].

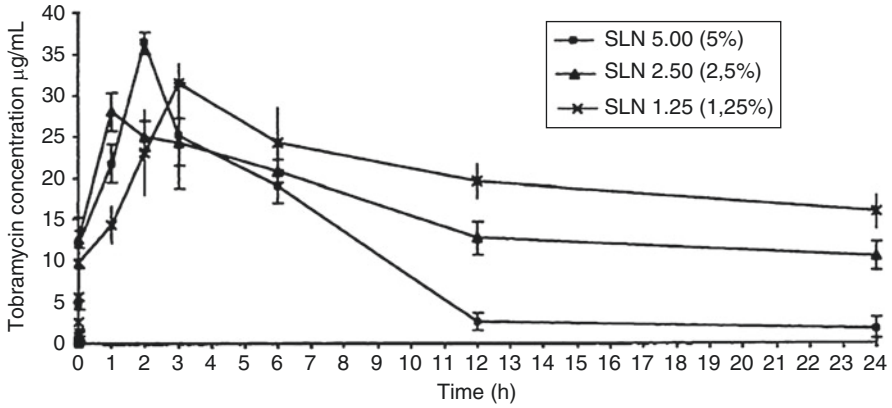
The results of this study confirmed transmucosal transport of drug when loaded in SLN to lymphatic system and the lymphatic uptake of tobramycin-SLN was detected also by TEM micrographs (Fig. 2) [9].

A critical parameter that has to be considered in oral administration is physicochemical stability of SLN into gastrointestinal fluids. SLN aggregation can occur in the stomach due to the acidity and high ionic strength of the gastric environment [1].

Drying of SLN dispersion into a dry powder (by freeze-drying and spray-drying process) can be necessary for the administration of solid dosage forms such as capsules and tablets.

Skin Application

SLN show many advantageous features for skin application. They can provide controlled release profiles, they are composed of physiological and biodegradable lipids with good tolerability, their small size allows a close contact with stratum corneum (horny layer) and causes formation of an adhesive film on the skin; film formation leads to an occlusive effect which increases skin



Solid Lipid Nanoparticles (SLN), Fig. 8 Tobramycin plasma concentrations versus time ± SD after duodenal administration of Tobra-SLN, loaded at different percentage of tobramycin [9]

hydration, improving drug penetration into the skin. Furthermore, SLN can enhance chemical stability of compounds sensitive to light, oxidation, and hydrolysis [16].

The horny layer is an efficient barrier that protects humans from excessive water loss, toxic agents, and microorganisms: it is formed by corneocytes embedded into epidermal lipids forming highly structured layers. Below the horny layer there is the epidermis followed by the dermis.

In general lipid nanoparticles do not penetrate the horny layer, but a follicular uptake has been reported for particulate systems.

Incorporation of drugs into SLN can be exploited for both topical administration and for transdermal administration [2, 16].

SLN have been proposed for improvement of treatment of skin diseases (such as atopic eczema, psoriasis, acne, skin mycosis, and inflammations): drugs incorporated into SLN and investigated for dermal application have been glucocorticoids, retinoids, and antimycotics.

Experimental data showed that SLN can enhance drug penetration into the skin increasing treatment efficiency: improved bioavailability at the site of action reduces the required dose and reduces dose-dependent side effects of the drug. SLN can act as a drug reservoir and this can be an important tool when it is necessary to supply the drug over a prolonged period of time and when drug produces irritation in high concentration [16].

SLN have been investigated also as carriers in cosmetics, especially for UV blockers. SLN by themselves have a sun-protective effect, due to their solid particulate character, and are capable of reflecting UV radiation leading to photoprotection: molecular sunscreens showed to be much more effective after incorporation into SLN for synergistic effects [1, 16].

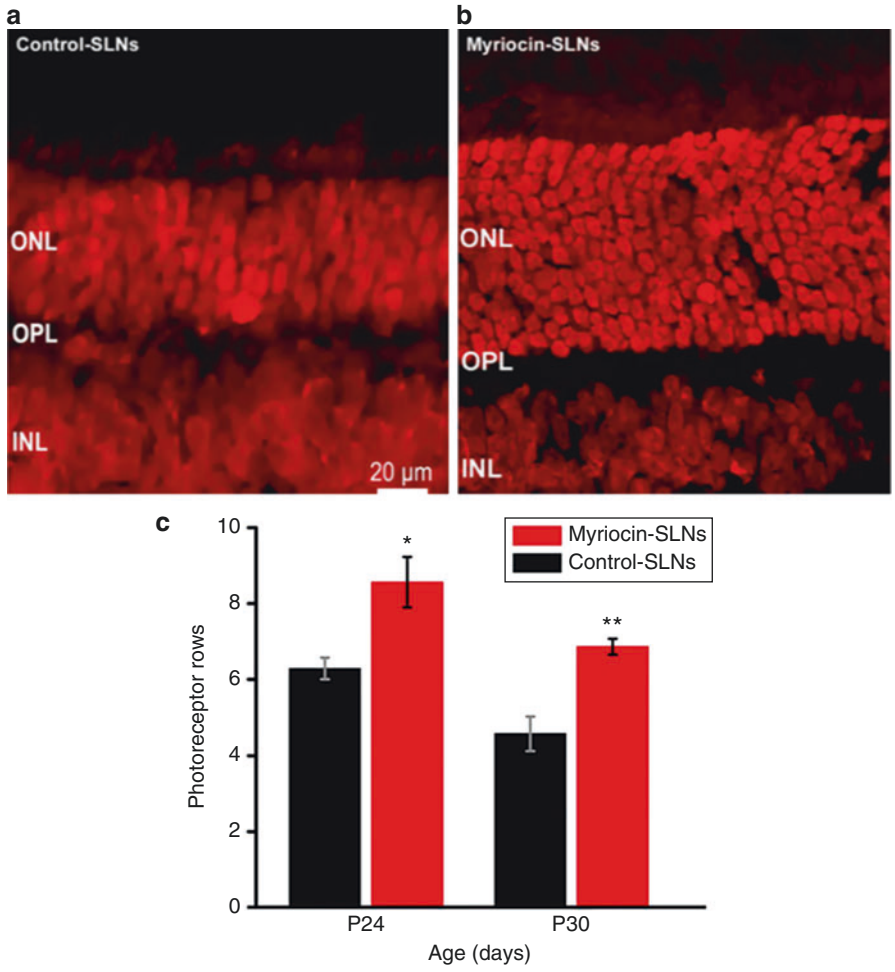
In order to allow skin application, SLN dispersion need to be incorporated into a cream or gel base which must not induce dissolution nor aggregation of SLN.

Ocular Topical Administration

Ocular bioavailability of drugs is limited by the complex structure of the eye: for ophthalmic application, drugs are usually formulated as eyedrops and administered topically but limited permeability of the cornea and other different mechanisms (such as lachrymal secretions, nasolacrimal drainage, drug adsorption into systemic circulation, drug spillage due to the limited capacity of human cul-de-sac, and drug metabolism caused by enzymes) strongly contribute to limit efficacy of the applied drug. Due also to this removal, several drug applications in a day are required to achieve therapeutic effect [17].

SLN can be useful system to enhance ocular bioavailability of both hydrophilic and hydrophobic drugs.

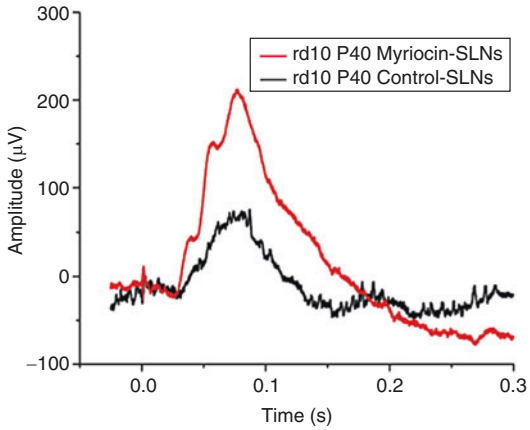
Possible advantages of SLN for improvement in ocular drug delivery include:



Solid Lipid Nanoparticles (SLN), Fig. 9 Effects of myriocin-SLN on retinal morphology: Vertical retinal sections from rd10 mice treated with control SLN (**a**) and myriocin-SLN (**b**) for 10 days (from P14 to P24). The outer nuclear layer (*ONL*) of the myriocin-treated retina is thicker because it contains more photoreceptor rows than the control retina. *INL* inner nuclear layer, *OPL* outer plexiform layer. (**c**) Quantification of photoreceptor rows at P24 and P30 in rd10 mice treated with control SLN or myriocin-SLN [18]

1. Bioadhesive properties, which can prolong ocular surface residence time, with better drug penetration and reduced drainage, increasing drug ocular bioavailability and decreasing dosage and relevant side effects
2. Controlled release of drugs avoiding frequent administrations
3. Protection of incorporated drug from action of metabolic enzymes present on the ocular surface
4. Good tolerability as SLN can provide low irritation and good compatibility with ocular tissues
5. Eyedrops formulation, which is self-administered by patients, improving compliance

SLN have been investigated in animal model for ocular delivery of different drugs such as tobramycin, diclofenac, timolol, and cyclosporin A: specific SLN formulation allowed to assess



Solid Lipid Nanoparticles (SLN), Fig. 10 Cone-driven ERG (Electroretinography) responses from rd10 mice (*red* trace) treated with myriocin-SLN or control SLN (*black* trace). Animals were age P40 and treatment had initiated at P14 [18]

efficacy of myriocin (MYR) in retinitis pigmentosa animal model [18].

MYR is a hydrophobic drug, it can inhibit intracellular synthesis of ceramide, a physiological lipid involved in apoptotic cascade. MYR has been proposed as drug for treatment of retinitis pigmentosa (RP), a disease where degeneration of photoreceptor causes blindness: there is not efficacious treatment approved for RP at this time. Efficacy of MYR loaded SLN has been showed in RP animal model (RD10 mouse model): three times a day administration by eyedrops of MYR-SLN, lasting a period of 35 days, did not show toxicity in mice and physiological registration of activity of photoreceptors by electroretinography (ERG) showed reduced degeneration occurred in animals treated with MYR-SLN (Figs. 9 and 10) [18].

Pulmonary Administration

The large inner surface of the lung and the thin alveolar epithelium allow rapid drug absorption avoiding first-pass metabolism; pulmonary administration using solid lipid nanoparticles represents an alternative for both local and systemic drug delivery.

SLN show different advantages for pulmonary drug delivery, such as the possibility to have controlled release profile, high tolerability, and possibility to reach lymphatic system through the action of alveolar macrophages. For pulmonary administration, SLN dispersions can be nebulized; inhalation device (design of nebulizer, flow rate) and physicochemical properties of formulations can affect aerosol nebulization efficiency and site of aerosol deposition.

Toxicological studies have been performed by using in vitro and in vivo tests [19]: A549 cells and murine precision-cut lung slices (PCLS) were exposed to increasing concentrations of SLN in order to estimate the toxic dose of SLN. The in vitro experiments showed toxic effects begin at concentrations of about 500 μg/mL, while for in vivo experiments toxicological potential of SLN was determined in a 16-days repeated dose inhalation study using mice which were daily exposed to different concentration of SLN (1–200 μg deposit dose): results showed that repeated inhalation exposure to SLN is safe in murine inhalation model.

Biodistribution of inhaled radiolabeled SLN in rats was studied [20]. Results show an important and significant SLN uptake into the lymph, few minutes after inhalation SLN translocate and accumulate into regional lymph nodes suggesting translocation mechanism of SLN may involve phagocytosis by bronchoalveolar macrophages followed by migration to lymphatic system.

Cross-References

- ▶ Liposomes
- ▶ Nanoencapsulation
- ▶ Nanomedicine
- ▶ Nanoparticle Cytotoxicity
- ▶ Nanoparticles
- ▶ RNAi in Biomedicine and Drug Delivery

References

1. Mehnert, W., Mäder, K.: Solid lipid nanoparticles: production, characterization and applications. Adv. Drug Deliv. Rev. **47**, 165–196 (2001)
2. Gasco, M.R., Mauro, A., Zara, G.P.: In vivo evaluations of solid lipid nanoparticles and microemulsions. In: Drug Delivery Nanoparticles Formulation and Characterization, pp. 219–238. Informa Healthcare, New York (2009)
3. Manjunath, K., Venkateswarlu, V.: Pharmacokinetics, tissue distribution and bioavailability of clozapine solid lipid nanoparticles after intravenous and intraduodenal administration. J. Control. Release **107**, 215–228 (2005)
4. Siekmann, B., Westesen, K.: Investigations on solid lipid nanoparticles prepared by precipitation in o/w emulsions. Eur. J. Pharm. Biopharm. **43**, 104–109 (1996)
5. Quintanar-Guerrero, D., Tamayo-Esquivel, D., Ganem-Quintanar, A., Allemann, E., Doelker, E.: Adaptation and optimization of the emulsification-diffusion technique to prepare lipidic nanospheres. Eur. J. Pharm. Sci. **26**, 211–218 (2005)
6. Schubert, M.A., Müller-Goymann, C.C.: Solvent injection as a new approach for manufacturing lipid nanoparticles – evaluation of the method and process parameters. Eur. J. Pharm. Biopharm. **55**, 125–131 (2003)
7. Battaglia, L., Gallarate, M., Cavalli, R., Trotta, M.: Solid lipid nanoparticles produced through a coacervation method. J. Microencapsul. **27**, 78–85 (2010)
8. Charcosset, C., El-Harati, A., Fessi, H.: Preparation of solid lipid nanoparticles using a membrane contactor. J. Control. Release **108**, 112–120 (2005)
9. Cavalli, R., Bargoni, A., Podio, V., Muntoni, E., Zara, G.P., Gasco, M.R.: Duodenal administration of solid lipid nanoparticles with different percentages of tobramycin. J. Pharm. Sci. **92**, 1085–1094 (2003)
10. Yuan, H., Huang, L.-F., Du, Y.Z., Ying, X.Y., You, J., Hu, F.Q., Zeng, S.: Solid lipid nanoparticles prepared by solvent diffusion method in a nanoreactor system. Colloids Surf. B Biointerfaces **61**, 132–137 (2008)
11. Bunjes, H.: Lipid nanoparticles for the delivery of poorly water-soluble drugs. J. Pharm. Pharmacol. **62**, 1637–1645 (2010)
12. Muchow, M., Maincent, P., Muller, R.H.: Lipid nanoparticles with a solid matrix (SLN, NLC, LDC) for oral drug delivery. Drug Dev. Ind. Pharm. **34**, 1394–1405 (2008)
13. Joshi, M.D., Muller, R.H.: Lipid nanoparticles for parenteral delivery of actives. Eur. J. Pharm. Biopharm. **71**, 161–172 (2009)
14. Bondì, M.L., Craparo, E.F.: Solid lipid nanoparticles for applications in gene therapy: a review of the state of the art. Expert Opin. Drug Deliv. **7**, 7–18 (2010)
15. Del Pozo-Rodriguez, A., Delgado, D., Solinis, M.A., Pedraz, J.L., Echevarria, E., Rodriguez, J.M., Gascon, A.R.: Solid lipid nanoparticles as potential tools for gene therapy: in vivo protein expression after intravenous administration. Int. J. Pharm. **385**, 157–162 (2010)
16. Pardeike, J., Hommoss, A., Müller, R.H.: Lipid nanoparticles (SLN, NLC) in cosmetic and pharmaceutical dermal products. Int. J. Pharm. **366**, 170–184 (2009)
17. Seyfoddin, A., Shaw, J., Al-Kassas, R.: Solid lipid nanoparticles for ocular drug delivery. Drug Deliv. **17**, 467–489 (2010)
18. Strettoi, E., Gargini, C., Novelli, E., Sala, G., Ilaria, P., Gasco, P., Ghidoni, R.: Inhibition of ceramide biosynthesis preserves photoreceptor structure and function in a mouse model of retinitis pigmentosa. Proc. Natl. Acad. Sci. U. S. A. **107**, 18706–18711 (2010)
19. Nassimi, M., Schleh, C., Lauenstein, H.D., Hussein, R., Hoymann, H.G., Koch, W., Pohlmann, G., Krug, N., Sewald, K., Rittinghausen, S., Braun, A., Muller-Goymann, C.: A toxicological evaluation of inhaled solid lipid nanoparticles used as potential drug delivery system for the lung. Eur. J. Pharm. Biopharm. **75**, 107–116 (2010)
20. Videira, M.A., Botelho, M.F., Santos, A.C., Gouveia, L.F., De Lima, J.J., Almeida, A.J.: Lymphatic uptake of pulmonary delivered radiolabelled solid lipid nanoparticles. J. Drug Target. **10**, 607–613 (2002)

Solid Lipid Nanospheres

► Solid Lipid Nanoparticles (SLN)

Solid Lipid-Based Nanoparticles

► Solid Lipid Nanoparticles (SLN)

Solid–Liquid Interfaces

► Nanoscale Properties of Solid–Liquid Interfaces

Solid-State Heat Convertors

Joseph P. Heremans
Department of Mechanical and Aerospace Engineering, Department of Physics and Department of Materials Science and Engineering, The Ohio State University, Columbus, OH, USA

Definition

Solid-state Heat Converters are solid-state devices that convert heat into electrical work without any moving parts, for example, by using thermoelectric effects or heat-induced changes in magnetic effects.

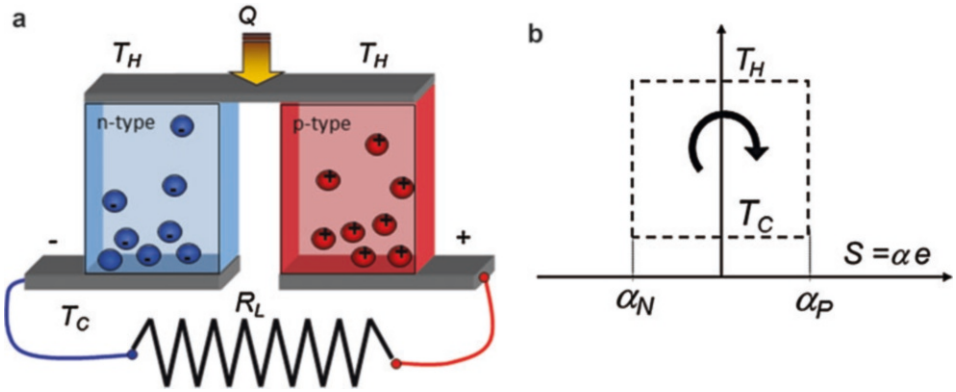
Motivation

Thermoelectric energy conversion emerged in the middle of the twentieth century. Early pioneers were Maria Telkes [1, 2], Abram Ioffe [3], and Julian Goldsmid [4], who developed thermoelectric semiconductors and showed that they could be used for electrical power generation from heat and for solid-state cooling. Our understanding of thermoelectricity itself dates back almost two centuries, with the discovery of the thermoelectric power of solids (the Seebeck coefficient) by Thomas Seebeck [5] in 1821–1822. Lord Kelvin [6] discovers the Thomson heat and understood thermoelectric effects to be part of classical thermodynamics. He derived his reciprocity relations by treating electrical currents like fluids in conventional mechanical heat engines. Lord Rayleigh [7] realized that thermoelectric effects could theoretically be used for solid-state heat-to-electricity conversion. He pointed out not only that the efficiency of thermopiles is proportional to the square of the thermopower, but also that "...the steam-engine and dynamo are not likely to be superseded by German-silver (a Cu-Ni alloy related to constantan) and iron thermopiles." Altenkirch [8] established the concept of the thermoelectric figure of merit ZT, which will be described further in these pages. In practice, before 1945 the only use of thermoelectric effects was in metallic thermocouples used for thermometry.

When a solid is subjected to a temperature gradient (∇T), a heat flux (j_Q) flows through it. To the first order, the heat flux is linearly related to ∇T via Fourier's law, $j_Q = \kappa \nabla T$, where κ is the solid's thermal conductivity. Both ∇T and j_Q are vectors; in the absence of external magnetic fields, they are collinear. When the solid is electrically conducting (metals or semiconductors), ∇T also generates an electric field E, again aligned with ∇T in the absence of an external magnetic field. In linear transport theory, both are again proportional and the thermoelectric power α of the solid, also known as thermopower or Seebeck coefficient, is the proportionality constant, defined as $\alpha \equiv {}^{E}/_{\nabla T}$. It has been shown by Callen [9] that, for a single electron and in strictly reversible thermodynamics, the thermopower equals the entropy S of that electron divided by its charge e, $\alpha = {}^{S}/_{e}$. Consequently:

1. The thermopower is a state equation: it does not matter how the electron goes from point A to B in a sample, just what the potential and temperature are at points A and B.
2. The thermopower is an absolute property of a material, not a difference. It does not require a thermocouple made from two materials to measure a thermopower. The absolute value of thermopower is measured by calorimetry by integrating the Thomson heat [10] over temperature.
3. The Nernst principle (third law of thermodynamics) holds to the thermopower: in the limit for zero absolute temperature, the entropy and therefore the thermopower are zero.
4. One can build heat engines with thermopower, either to generate power from heat or to operate as cooling/refrigeration cycles.

Figure 1a shows a thermoelectric power generator, which consists of two thermoelectric



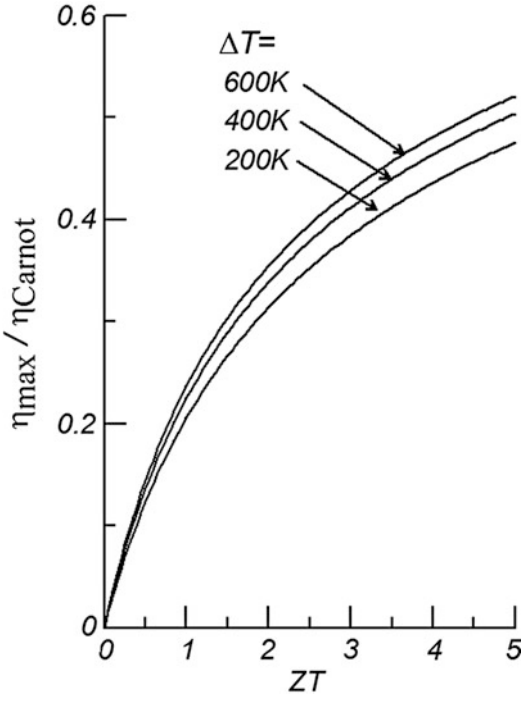
Solid-State Heat Convertors, Fig. 1 (**a**) Schematic diagram of a thermoelectric power generator consisting of an n-type semiconductor and a p-type semiconductor connected as shown. (**b**) The thermodynamic cycle an electron goes through as it travels through the power generator, represented in a (temperature, entropy) diagram which is the equivalent of a (temperature, thermopower) diagram

semiconductors: one is doped n-type in which the majority carriers are electrons, and the thermopower is negative ($\alpha_N < 0$); and other one is doped p-type in which the majority carriers are holes, and the thermopower $\alpha_P > 0$. These semiconductor elements are connected in a thermopile, with a hot contact at temperature T_H and a cold contact at temperature T_C; the difference in temperature between the contacts is maintained by adding heat Q to the hot contact. The electrical contacts are connects in series, as shown in Fig. 1a: an electrical load R_L is attached on the cold side between the n- and p-type semiconductors. The temperature difference T_H–T_C creates a voltage between the two bottom plates, which, in open circuit conditions (i.e., if $R_L \to \infty$), would be equal to $(\alpha_P–\alpha_N)(T_H–T_C)$ with the polarity shown in Fig. 1a. Assume further that α_P and α_N are not temperature-dependent between T_H and T_C. Looking at Fig. 1a from the point of view of one electron, starting from the bottom left (cold side, n-type), as that electron rises in the structure, its temperature increases at constant value of α_N, i.e., isentropically at entropy $S_N = \alpha_N e$, under Callen's condition that the process were reversible. That progress can be followed in the (T,S) diagram (Fig. 1b). When the electron has reached the top hot plate, it moves over from left to right isothermally at T_H, as can be followed in Fig. 1b. It then cools back down to T_C in the p-type semiconductor, another isentropic transformation, at a constant entropy $S_P = \alpha_P e$, again assuming the process is considered reversible. Finally, it flows through the load R_L isothermally again at the cold temperature T_C. Therefore, in reversible thermodynamics, a thermoelectric power generator as shown in Fig. 1a would have a thermodynamic cycle as shown in Fig. 1b, which is a Carnot cycle, and the efficiency would be the Carnot efficiency

$$\eta = \eta_C = \frac{T_H - T_C}{T_H}. \tag{1}$$

Peltier coolers, under the same assumption, would also have the Carnot Coefficient of Performance. Clearly, none of these assertions are true, because the losses in thermoelectric generator and Peltier cooler are dominated by two main irreversibilities:

1. Heat is lost by thermal conduction in the semiconductors. This loss is quantified by the thermal conductivity of the materials (κ_N and κ_P).
2. Joule heating is also an irreversible process. It is due to the internal resistance of the thermopile though which the electrical current must pass in order to do electrical work in the load. Joule heating is minimized by maximizing the electrical conductivity σ of each of the semiconductors.



Solid-State Heat Convertors, Fig. 2 Thermal efficiency of the thermoelectric generator of Fig. 1, as a function of the average *ZT* of the n- and p-type semiconductors, for the values of temperature difference (ΔT) indicated. The efficiency is represented as a function of Carnot efficiency and includes the irreversible thermodynamic losses of the device

It is Altenkirch [7] who first quantified the role of these irreversibilities and related the efficiency of the thermoelectric generator in Fig. 1 to the thermoelectric figure of merit *ZT* of each of the two semiconductors. This is defined by:

$$ZT \equiv \frac{\alpha^2 \sigma}{\kappa} T. \tag{2}$$

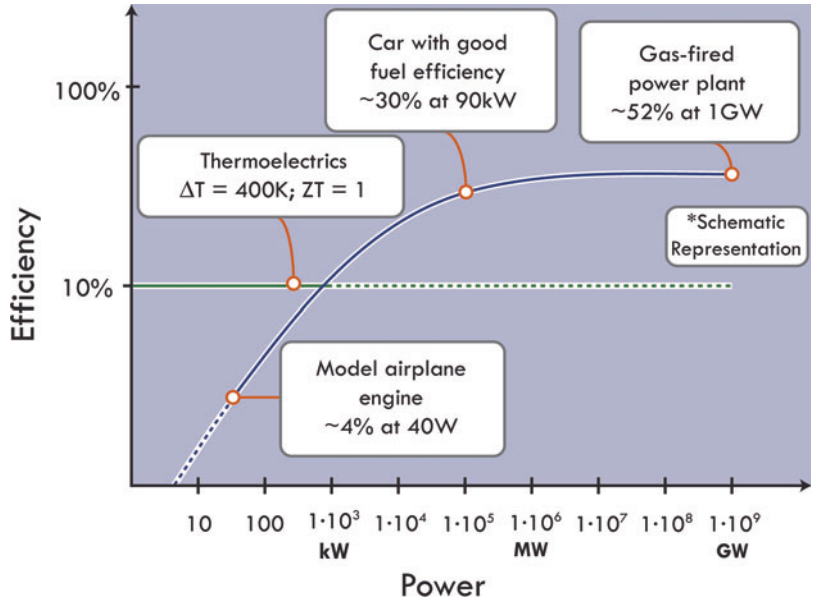
The *ZT* must be characterized separately for the n- and p-type materials, but for simplicity we will assume that they are equal and omit the indices (or, to a first approximation, assume that for the device *ZT* is the average between ZT_N and ZT_P). If so, the efficiency is shown as a function of *ZT* in Fig. 2. We also define the thermoelectric power factor *PF*:

$$PF \equiv \alpha^2 \sigma. \tag{3}$$

The *PF*, which appears in the numerator of *ZT*, is a measure of how much electrical power per unit volume of material a generator can produce under a constant temperature difference. Indeed, under a given (T_H–T_C), α determines the voltage, $\alpha\sigma$ the current density, and the product of the two the power density.

The TE generators such as shown in Fig. 1 have many practical advantages over conventional internal combustion engines or steam-power cycles. They have no moving parts, do not wear out, require no maintenance, are extremely reliable, and have almost infinite lifetimes. These generators have been used to produce power in interplanetary space probes (where the source of heat are radioactive isotopes) where they have operated continuously for more than half a century. They also have extremely high power densities (the number of Watts per unit volume of material), which implies that they scale well to low power densities (1 W or less). The main drawback of TE generators is that at large power levels they are far less efficient than conventional heat engines [11], because the *ZT* of today's materials is too low. This relative disadvantage turns into an advantage in low-power applications, because conventional fluid-based machines lose their efficiency with power level. Figure 3 shows that there is a power range (<1 kW, perhaps <100 W) where TE generators compete well [12] with mechanical fluid-based engines. Nevertheless, in the last 20 years, a significant effort is devoted to finding ways to improve the *ZT* other TE semiconductors, mainly based on the introduction of nanotechnologies [12].

The main difficulty in thermoelectric research is that α, σ, and κ are interrelated and almost mutually counter-indicated in the way they enter (Eq. 2). Yet, they have to be optimized on the same material. To illustrate this, one can re-write the *ZT* recalling that the electrical conductivity is a function of the density of charge carriers n and their mobility μ, and by recalling that the thermal conductivity is due to contributions of κ_E from the electrons and κ_L from the lattice vibrations or phonons:



Solid-State Heat Convertors, Fig. 3 Schematic comparison of the efficiency of conventional heat engines and thermoelectric converters as a function of power level. Thermoelectric solid-state heat engines are competitive in only low-power applications

$$ZT = \left(\alpha^2 n\right)\left(\frac{\mu}{\kappa_E + \kappa_L}\right)eT. \tag{4}$$

The factor $(\alpha^2 n)$ in Eq. 4 contains two mutually counter-indicated properties α and n. Indeed, the thermoelectric power decreases as the charge carrier concentration increases in most solids, a relation attributed to Mr. Pisarenko [13] and shown Fig. 4. As a result, the power factor has a maximum at the optimal charge carrier concentration n_{OPT}, typically between 1×10^{18} cm^{-3} for solids with an electron effective mass m^* below 0.1 m_e (the free electron mass) to 5×10^{20} cm^{-3} for solids with $m^* > m_e$. In the more common thermoelectric semiconductors with $0.25\ m_e \leq m^* \leq 0.7\ m_e$, we have 2×10^{19} cm$^{-3} \leq n_{OPT} \leq 7 \times 10^{19}$ cm^{-3}. Further, the higher the value of the density of states effective mass m^*, the higher the value of the optimum power factor. This illustrates the best way maximize $(\alpha^2 n)$: select solids with high effective masses.

The factor $\mu/(\kappa_E + \kappa_L)$ in Eq. 4 contains another two other mutually counter-indicated properties μ and κ_L. Indeed, adding defects to the system to decrease the thermal conductivities of either electrons or phonons usually also decreases the electron mobility. To illustrate how to maximize $\mu/(\kappa_E + \kappa_L)$, we re-writing Eq. 4 making use of the Wiedemann-Franz law that relates κ_E and σ.

$$\kappa_E = LT\sigma. \tag{5}$$

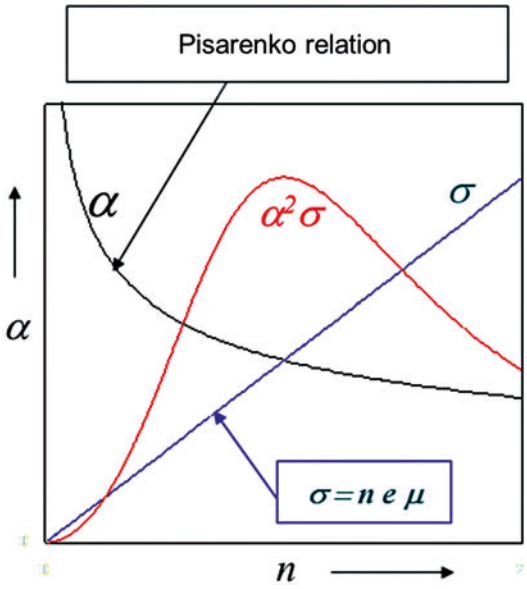
Where L is the Lorentz ratio. For free electrons, $L = L_0$, a smattering of universal constants giving $L_0 \equiv 2.4 \times 10^{-8}$ V^{2}K^{-2}. In practice, L varies little from L_0, and is typically $0.6\ L_0 < L < 1.4\ L_0$. Substituting this in Eq. 2 gives:

$$ZT = \frac{\alpha^2}{L}\frac{T}{1 + \frac{\kappa_L}{\kappa_E}}. \tag{6}$$

This illustrates that thermoelectrics research must aim at reducing the lattice thermal conductivity, not necessarily in absolute numbers, but with regards to the electronic thermal conductivity. It also illustrates the importance of maximizing the thermopower α as long as the electrical conductivity is high enough. In the extreme case of very highly conducting metals, where the fraction of the heat carried by the phonons is zero, $ZT = {}^{\alpha^2}\!/_{L}$, and the primary goal of research has now shifted from optimizing ZT to optimizing a single transport property, the thermoelectric power α.

In the light of the above discussion, the following strategies are used to optimize ZT.

1. Phonon engineering to reduce lattice thermal conductivity
   1a. Engineering nanostructures that scatter phonons more than electrons
   1b. Engineering localized phonon modes (rattlers) to scatter acoustic phonons
   1c. Engineering and maximizing the anharmonicity of the chemical bonds to promote phonon-phonon interactions
2. Band structure engineering to enhance thermopower $\alpha(n)$
   2a. Engineering nanostructures to size-quantize electrons, using two-dimensional structures, quantum wells, quantum wires, and quantum dots.
   2b. Engineering resonant impurities that produce similar effects in the band structure, but now in bulk solids.
   2c. Engineering spin polarization in Kondo physics or correlated systems
3. Add an additional optimization parameter to mitigate the problem of having to optimize counter-indicated properties of a single solid: magnetism and spin physics

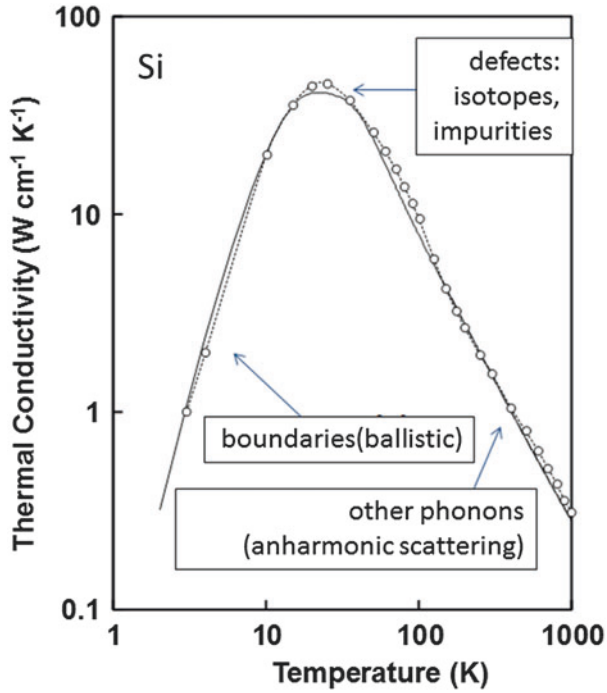


Solid-State Heat Convertors, Fig. 4 Schematic relation between thermopower (α), electrical conductivity (σ), and thermoelectric power factor ($\alpha^2\sigma$) and the concentration n of charge carriers (electrons or holes)

From Eq. 6 and the discussion above, it is also clear that reducing the relative role of the lattice thermal conductivity can be achieved by either reducing κ_L itself or increasing σ and κ_E as long as α is not affected: it is thus more effective to increase the power factor than to decrease the thermal conductivity. These strategies are discussed below.

Phonons

Lattice thermal conductivity κ_L is the only heat-carrying mechanism in nonmagnetic dielectric solids, and, in crystals, it is mainly due to the contribution of the acoustic phonons to the thermal conductivity [14]. These phonons are scattered by either other phonons (phonon-phonon interactions) or various defects. The temperature dependence of the thermal conductivity of high-purity single-crystal silicon [15], shown in Fig. 5, illustrates the various regimes. The defects are generally classified by their size in relation to the wavelength of the phonons they interact with. The phonons that carry heat are dominantly acoustic phonons, which come in three modes, one longitudinal (LA) and two transverse (TA) acoustic modes. Each mode has a dispersion at temperatures much below the Debye temperature Θ_D of the solid given by $E = \hbar\omega = \hbar k v_G$, where v_G is the phonon mode's group velocity (the sound velocity); k is the phonon wave vector, the inverse of its wavelength $\Lambda = {}^1/_k$. Because the average phonon energy E is on order of $k_B T$, the above equation immediately shows the inverse relation $k_B T = \hbar v_G/\Lambda$ between temperature and wavelength: the abscissa axis of Fig. 4 therefore can be seen as proportional to the inverse wavelength of the phonons of average energy at that given temperature, called "thermal phonons." There is an intuitive direct relation between wavelength and mean free path of sound waves: longer wavelength phonons are scattered by larger-scale obstacles. Therefore, Fig. 4 is explained in terms of the length scales of the obstacles that scatter acoustic phonons.



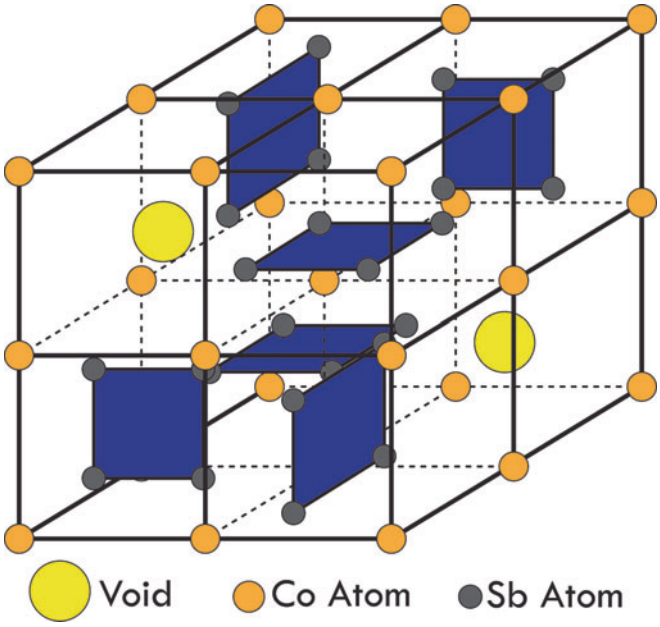
Solid-State Heat Convertors, Fig. 5 Thermal conductivity of single-crystal silicon as a function of temperature. The thermal conductivity is entirely due to phonons. Below 10 K it is limited by interactions between phonons and the sample boundaries, i.e., transport is ballistic. Above 80 K it is limited by interactions between phonons, which are governed by the anharmonic nature of the bonds of the crystal. In the intermediate temperature range, the thermal conductivity has a maximum, the value of which is limited by interactions with phonons with various defects, such as various isotopes of Si or impurities or dislocations

Ballistic Phonons

The amount of heat carried by each mode is given by the kinetic formula, where phonons are considered quasi-particles that behave like an ideal gas: $\kappa_L = {}^1\!/_3 C v_G \ell$. Here the mean free path $\ell = v_G \tau$ is expressed as a function of the group velocity and the scattering time τ, which is the inverse of the scattering rate τ^{-1}. The scattering rate depends on the nature of the obstacle that scatters the phonon. At low temperature (left, in Fig. 5), where phonon wavelengths are very long, the phonon mean free path can reach millimeters at cryogenic temperatures and ℓ is limited by the crystal boundaries. Phonon transport becomes ballistic. It is characterized by a scattering time τ_B, calculated for a crystal whose smallest size is defined as t to be $\ell = t, \tau_B = {}^t\!/_{v_G}$. To reduce κ_L in this temperature range, good thermoelectrics are prepared as fine-grained crystals, typically $t < 1$ μm and on the order of l00 nm, by ball-milling and sintering [16]. This has led to important reductions in κ_L and increases in the ZT of the most widely used thermoelectrics for Peltier cooling at room temperature, p-type alloys of approximate formulas $(Bi_{0.25}Sb_{0.75})_2Te_3$, and n-type alloys of a composition near $Bi_2(Se_{0.1}Te_{0.9})_3$ [16].

Phonon-Phonon Interactions

At a temperature on the order of $T \leq \Theta_D/3$, the phonon wavelengths become of the order of one or a few interatomic distances or crystal unit cells, and phonon-phonon scattering (with a scattering probability $\tau_\phi{}^{-1}$) dominates. The scattering frequency, $\tau_\phi{}^{-1}$, is the product of three factors: (1) the number of phonons available for scatting (i.e., the phonon concentration), (2) the number of empty states available for phonons to scatter into, and (3) the probability of the interaction of each phonon with



Solid-State Heat Convertors, Fig. 6 The structure of the unfilled skutterudite $CoSb_3$

another phonon. All three of these properties can be rigorously evaluated, but they involve nested integrals over the entire Brillouin zone of all phonon modes [17]. Solids can be selected that maximize any of these three properties and have minimum lattice thermal conductivity; some solids can even be engineered to induce these phonon properties.

Selecting solids that have a low density of states in acoustic modes is typically done by using large unit cells. Indeed, this implies that the reciprocal cell and the Brillouin zone are small and have a low density of states. Many good thermoelectrics have long and complex chemical formulae, which achieves this goal. A corollary is that what would otherwise be acoustic modes in solids with simple unit cells become optical modes in more complex solids built from a multiplicity of such simple unit cells. Optical modes not only have low group velocities, and therefore do not carry heat efficiently, but they are also very helpful with the next mechanism.

Selecting solids with a large number of available phonon modes for acoustic phonons to scatter into generally is done by selecting solids with high densities of optical modes, as described above, with energies coinciding with the energies of the acoustic phonons. This approach has been pushed into phonon-mode engineering, typically applied to a family of high-temperature thermoelectrics of general formula $CoSb_3$ that belongs to the crystallographic family called skutterudites after a natural mineral $(Co,Ni,Fe)As_3$ ore. While n-type $CoSb_3$ has a good power factor, its lattice thermal conductivity is too high to have a good *ZT*. The unit cell of the crystal consists of eight elementary cubes with Co atoms on their corners, as shown in Fig. 6. Two of these cubes are empty, and the other six are filled with planar rings of Sb, arranged so as to form octahedra with a Co atom in the middle. The empty cubes act as voids that can accept an electronegative atom in the middle, creating "filled" skutterudites. When this atom is small, it can "rattle" in the void, creating an optical phonon mode that can accept energy from the heat-carrying acoustic modes, dramatically lowering the lattice thermal conductivity [18–20]. This effect has resulted in alloy compositions for high-temperature thermoelectrics that have very high *ZT* values reaching $ZT \approx 1.7$ at 800 K [21].

Selecting solids in which the phonon-phonon interactions are maximized has also been achieved

recently. Returning to the temperature dependence of κ_L at $T > 100$ K in Fig. 4, one observes that it scales roughly as a T^{-1} law, which is characteristic for a conductivity limited by phonon-phonon scattering. A very simplified expression is given by Berman [14], valid for temperatures of about $T > \Theta_D/2$:

$$\kappa_L = A \frac{\overline{M} \Theta_D{}^3 V_{\text{atom}}^{1/3}}{\gamma^2 n^{2/3} T} \tag{7}$$

Here, $\overline{M}$ is the average mass of atoms, V_{atom} the volume per atom, n the number of atoms in the unit cell, A a collection of universal constants (3×10^{-5} if SI units are used consistently), and γ is the average Grüneisen constant. The Grüneisen constant is a mode- and frequency-dependent property that characterizes the change of the frequency ω of a given phonon with the volume V, $\gamma(\omega) \equiv d\ln(\omega)/d\ln(V)$ and is mapped out across the Brillouin zone. The Grüneisen constant arises from the anharmonicity of the chemical bond: if the interatomic bonds behaved like perfect harmonic oscillators, or like balls and springs with position-independent spring constants, $\gamma = 0$. The fact that the spring constants do depend the atomic displacements (via V), i.e., the anharmonicity, is reflected on the acoustic properties of the medium and phonon eigenfrequencies ω. The effect of γ on the probability of phonon-phonon interactions arises as follows. In an anharmonic solid, when a first phonon passes, the atomic displacements involved change the local values of ω, thereby modifying the acoustic property of the medium for a second phonon, should one pass by. The second phonon is thereby more likely to be reflected by the first phonon in a solid with a higher γ; the probability of interaction $\tau_\phi{}^{-1}$ therefore scales with γ^2, which is reflected in Eq. 7.

How does one select solids with large anharmonicities? Solids where the atoms have high coordination numbers (octahedral bonds or more) are more anharmonic than those with lower numbers (tetrahedral bonds), favoring rocksalt (PbTe) or tetradymite (Bi_2Te_3) structures over diamond (Si) or wurtzite (ZnS) structures. Recent progress has started from identifying the role of the lone pair electrons [22] in compounds of the class of the I-V-VI_2 compounds. Here, I represents a group I element: an alkali or a noble metal atom; V represents a pnictogen atom: As, Sb, or Bo; and VI represents a chalcogen atom: S, Se, or Te. These I-V-VI_2 compounds crystallize in a cubic inverted $CaCl_2$ structure. The paradigm for such materials, $AgSbTe_2$, has a ZT exceeding 1 [23]. While one reason is its very favorable band structure and power factor [24], the main effect is due to a lattice thermal conductivity that is pinned at the amorphous limit by phonon-phonon interactions [25] due to the extremely high γ. It was later shown that this is a general property of all rocksalt I-V-VI_2 compounds [22], due to the extreme sensitivity of the valence electrons whose orbital nature corresponds to the s-electrons of the group V element (the "lone pair" electrons in chemical terms) to heat-carrying acoustic phonons, resulting in an extremely high value of γ for those specific phonons.

There are several other structures where similar anharmonicities are at work, although not necessarily on the same phonons or due to the same parts of the valence band structure. One example involves the copper atoms in Cu-Sb-Se tetrahedrites [26], a naturally occurring mineral that, with some chemical modifications, can be made into an excellent thermoelectric. The role of lone pair electrons on the pnictogen atom in the thermal conductivity of tetradymites (Bi_2Te_3, Bi_2Se_3, Sb_2Te_3) has not yet been studied. In contrast, the role of Pb and Sn on that in PbTe is extensively investigated. It is proven that PbTe hosts some extraordinarily strong interactions between acoustic and optical phonons [27], an effect that contributes much to the very high ZT of PbTe. The interactions in the lead and tin salts involve optical phonons. Thus, they are subtly different from those that affect the acoustic phonons of the I-V-VI_2 compounds. However, such considerations in no way detract from the fact that anharmonicity is the key to the high ZT of the lead salts and of the record-holding $ZT = 2.6$ obtained in SnSe [28]. This result is made more remarkable by the fact that it was obtained on single crystals of a simple binary compound, without any alloying or

nanostructuring [29], and must therefore be intrinsic, such as only phonon-phonon interactions are known to be.

Nanostructures

At an intermediate temperature scale on the order of $\Theta_D/10$ to $\Theta_D/2$, the lattice thermal conductivity of very pure large single crystals shows a maximum (see Fig. 5) where boundary scattering and phonon-phonon interactions are both low. Since ℓ of the thermal phonons is very long at the temperature of this maximum, any extrinsic defects remaining in the sample will scatter the phonons. The phonon wavelength here is on the order of $\Lambda \sim 100$ nm to 10 μm, and defects that can limit ℓ at these temperatures are those at those at a length scale about ten times smaller, namely a few to a few hundred lattice unit cells. Some defects are unavoidable, such as naturally occurring isotopes. Dislocations and foreign atoms are other impurities that scatter phonons most prominently here; they have characteristic frequency and temperature dependences [14], but often these impurities tend to scatter electrons as much as phonons and do not increase *ZT* efficiently. It is here that nanotechnologies have had their strongest influence on reducing κ_L in thermoelectric semiconductors.

There are two ways to create nanostructures in thermoelectric semiconductors, metallurgical heat treatments and powder metallurgy. Metallurgical heat treatments can be used to create nanoprecipitates of a second phase in compound semiconductors of the parent phase, mimicking the formation of perlitic steels or Guinier-Preston zones in duralumin. Early on, excess Pb nanoparticles were precipitated in PbTe [30]. In order to improve *ZT* values, the precipitates must let electrons through but scatter phonons, thereby improving the $\mu/(\kappa_E + \kappa_L)$ ratio in Eq. 4. Kanatzidis and his group discovered that the best way to achieve this is to find secondary phases in the thermoelectric semiconductors that are also semiconductors but have electronic bands that align with those of the host material. These secondary phases should have very similar crystal structures and grow in crystallographic registry ("endotaxially"). Examples are SrTe nanoparticles in PbTe and also PbS in PbTe. PbS within the PbTe matrix further beneficially modifies the density of electron states (see below) to allow for enhancement of the power factor to give a ZT of 1.8 at 800 K [31]. Other groups have used this technique successfully on various other materials.

Powder metallurgy can be used to create nanoinclusions, or at least crystals of nanometer-sized grains, that will scatter intermediate-energy phonons when made at the 10–100 nm scale, as mentioned in the section on ballistic phonons. This method was used successfully to prepare tetradymite materials that hold the record for *ZT* in their temperature ranges [16].

All-Scale Hierarchical Phonon Scattering

The record value of *ZT*, prior to the value reached on SnSe, is obtained on PbTe that includes nanoscale SrTe. It is prepared by ball-milling and spark plasma sintering and reaches $ZT = 2.2$ at 900 K [32]. This material, in essence, combines all of the techniques described above to reduce κ_L without affecting the mobility too much. The SrTe nanostructures scatter phonons with short and medium mean free paths (3–100 nm), but leave phonons with longer mean free paths unaffected. The phonons with longer mean free paths (0.1–1 mm) are affected by the grain-boundary phonon scattering. At the highest temperature, the strong anharmonicity promotes the phonon-phonon interactions that scatter the shortest-wavelength phonons (<3 nm), an intrinsic mechanism that is present in all PbTe samples. The authors label this the all-scale hierarchical architecture for designing thermoelectric materials.

Electrons

Hicks and Dresselhaus pointed out in two seminal papers [33, 34] in 1993 how quantum size effects could open a new way to design thermoelectric materials and had the potential to increase the *ZT* of narrow-gap semiconductors and semimetals to values of up to 5. The basic idea, which was subsequently fleshed out by Mahan and Sofo [35], stems from the Mott relation [36] between

the thermopower and the energy-dependent "Mott" conductivity $\sigma(E)$, a quantity which, when integrated over energy, gives the electrical conductivity $\sigma = -\int \sigma(E)\left({}^{\partial}f/_{\partial E}\right)dE$. The Mott relation is:

$$\alpha = \frac{k_B}{e\sigma}\int \sigma(E)\frac{E - E_F}{k_B T}\left(\frac{\partial f}{\partial E}\right)dE. \qquad (8)$$

Here, f is the Fermi-Dirac distribution function and E_F the Fermi energy. For non-degenerately doped semiconductors, the Mott relation yields the Pisarenko relation mentioned in section "Motivation." For metals and degenerately doped semiconductors, the Bethe-Sommerfeld expansion (a second-order Taylor expansion around the Fermi energy E_F) can be applied to yield:

$$\alpha = \frac{\pi^2}{3}\frac{k_B}{e}k_B T\left\{\frac{d[\ln(\sigma(E))]}{dE}\right\}_{E=E_F}. \qquad (9)$$

When $\sigma(E)$ is attributed to band conduction (as opposed to hopping or strong localization) as is the case for thermoelectric materials, its relation to the density of states (DOS) $g(E)$, charge carrier concentration $n = \int g(E)f(E, E_F, T)dE$, and mobility μ is such that (Eq. 9) becomes:

$$\alpha = \frac{\pi^2}{3}\frac{k_B}{e}k_B T\left\{\frac{g(E)}{n(E)} + \frac{1}{\mu(E)}\frac{d\mu(E)}{dE}\right\}_{E=E_F}. \qquad (10)$$

Therefore, when $g(E)$ has a sharp maximum at $E = E_F$, the first factor in the thermopower is maximized, since $n(E)$ is essentially the area under the $g(E)$ curve up to $E = E_F$. When this formula is further worked into the ZT, classical thermoelectric theory [4] shows that the ZT is maximized in solids that have a high density of states effective mass. The traditional way to achieve this is to make use of solids where there the Fermi surface consists of several pieces, combining either a single heavy band or several separate ones.

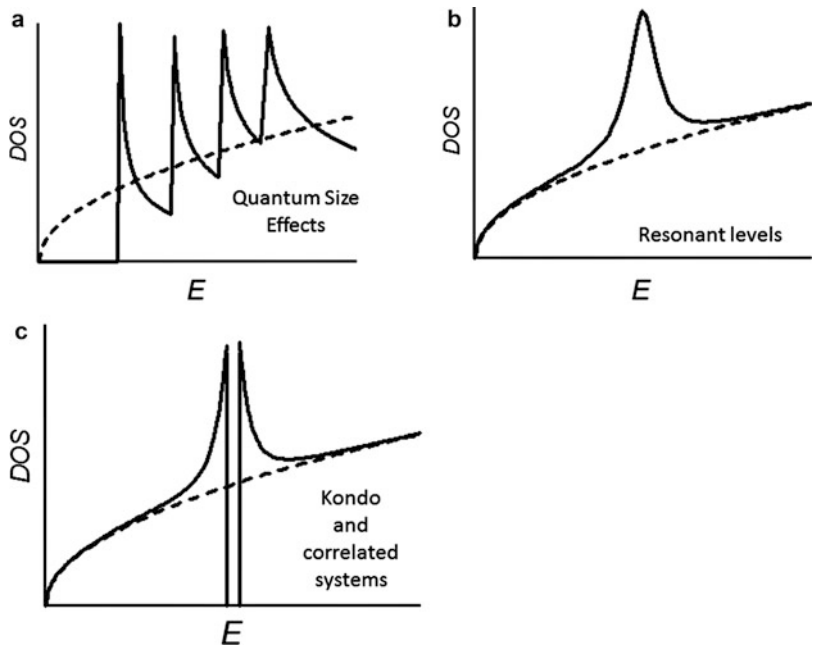
More recent developments involved engineering the DOS energy dependence $g(E)$ in the shapes shown in Fig. 7. In that figure, the DOS is shown for a conventional energy band as a dashed line. The techniques used to obtain sharp maxima in the DOS are illustrated frames (a)–(d) and are described in the following paragraph. Mahan and Sofo [35] point out one very important limitation: the effect of these approaches on ZT is strongly degraded when the background DOS (the dashed line) constitutes an important fraction of the DOS at E_F.

Size-Quantization

Semiconductor quantum wires are wires where the physical size of the structure is narrower than wavelength of the electrons (Fig. 7a). In such wires, electron motion is possible only in the long direction of the wire and quantized energy levels form in the transverse direction [37]. This was originally conceptualized by Hicks and Dresselhaus [34]. It has been implemented experimentally mainly in Bi [38–40] and Sb [41] nanowires. Strong enhancements of the thermopower were observed [42], but there are limitations on how narrow one can make these nanowires, as localization effects set in at around 10 nm diameters. In fact, in metals like Zn, where size-quantization effects require much narrower structures than in Bi, localization dominates [43]. Arrays of quantum wires in templates are difficult to use in thermoelectric materials, because the templates constitute a thermal short. Precipitating nanoparticles small enough that size-quantization effects inside them produce high thermopowers has also been tried using techniques similar to those described above. As a result, the thermopower of PbTe was shown to be enhanced [44].

Resonant Levels

Conventional doping of semiconductors consists of adding donor impurities or acceptor impurities to a host lattice (Fig. 7b). In a donor impurity, the impurity has one more electron on its outer shell than the atoms that constitute the host solid, such as P in Si. In an acceptor impurity, the impurity has one less electron on its outer shell than the atoms that constitute the host solid, such as Ga in Si. Consider only n-type doping of the conduction



Solid-State Heat Convertors, Fig. 7 Energy dependence of the electronic density of states (*DOS*) for several engineered band structures, designed to create favorable thermoelectric properties. The DOS of a conventional bulk semiconductor is shown as the dashed line. Frame (**a**) shows the DOS in quantum wires, where size-quantization squeezes the electron wave function into a one-dimensional shape. Frame (**b**) corresponds to the DOS in a system where a resonant impurity interacts with the band structure of the conventional semiconductor. Frame (**c**) is similar to frame (**b**), but now the interacting impurities add magnetic interactions to the system, and the peak splits in two peaks in which the electrons have opposite spin polarization

band for simplicity since the case of p-type doping is exactly symmetric. At very low temperature, the excess electron of the donor impurity stays localized on the donor, but it can be thermally excited and then become a mobile conduction electron. Translating this in the band model, one states that the excess electron has an energy state in the band gap of the semiconductor at an energy E_D below the conduction band edge (CB). The excess electron is localized on the band gap at low temperature, but it can be thermally activated from that band gap at temperatures T such that $k_BT > E_D$. The CB itself is determined by the properties of the host semiconductor, and the presence of the donor or acceptor does not modify it (the rigid band model, the dashed line in Fig. 3b). Some particular impurities in specific host semiconductors have electronic energy levels for which $E_D < 0$. The impurity level falls inside the conduction band. Its energy coincides with the energy of an extended state of the host. The two resonate to build up two extended states of slightly different energies; these in turn will have the same energies as other extended states with whom they will resonate in turn, and so on. Consequently, the resonant impurity [45] catalyzes the formation of an excess DOS (Fig. 7b). This excess DOS has, in fact, only very little of the character of the electronic levels of the impurity and dominantly consists of energy levels of the host solid, such that those states conduct electricity only slightly less than the original semiconductor. Nevertheless, because of their shape (Fig. 7b), and because of Eq. 10, they increase the thermopower, and thus the *ZT*, of the semiconductor (Tl in PbTe [46], Sn in Bi_2Te_3 [47]) in which they are observed.

Kondo and Correlated Systems

Figure 7c gives a very schematic representation of a concentrated Kondo/hybridized system [45].

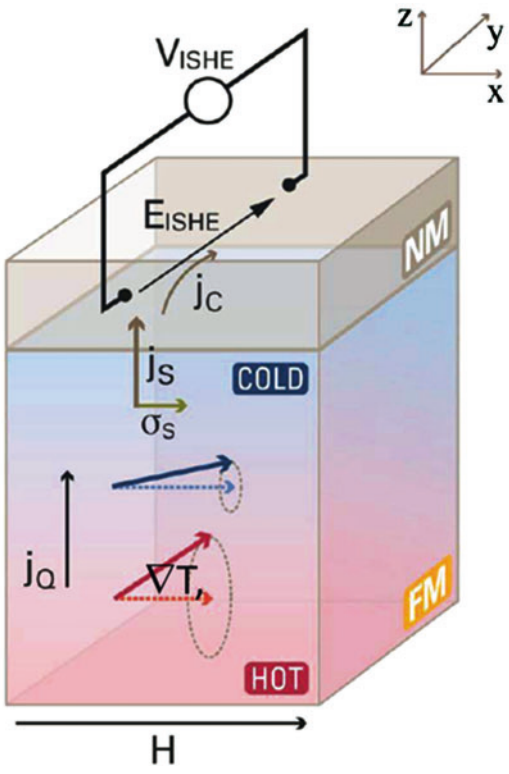
The boost in thermopower obtained from Kondo physics is an effect of the scattering of electrons (the second term in Eq. 10), which becomes indistinguishable from the DOS effect, and it is related to the spin polarization of that scattering – a link into the next chapter. The dilute Kondo effect (DK) was first observed in dilute alloys of metals where the dominant character of the bands correspond to s-orbitals (Cu, Ag, or Au) or p-orbitals to which small amounts transition metal impurities (Fe, Cr, Mn...) are added. The thermopower, which is proportional to T in metals, is greatly enhanced following a Gaussian peak in $\log(T)$ centered on a "Kondo temperature" T_K. T_K can reach 10^3 K [48], but the amplitude of the peak in S rarely exceeds a few tens of $\mu V\ K^{-1}$ in practice [49]. The resistivity, normally a monotonically decreasing function with decreasing T, has a minimum at T_K. Kondo explained [50] the effect on the basis of the fact that conduction electrons undergo a change of spin when they scatter on magnetic impurities, with the details rooted in the calculation of the scattering probabilities. Blatt et al. [51] represent the DOS in DK alloys as similar to that of the resonant levels (Fig. 7b), but with a magnetic splitting of the excesses in DOS into a spin-up and a spin-down bump.

When there are sufficient quantities of them, the magnetic atoms become a major constituent in the solid, such that their electronic levels contribute majorly to the band structure. When this occurs, the nearly dispersionless energy band that originates from the *d*- or *f*-levels of a rare-earth or transition metal constituent intersects a band with *sp* character in a semiconductor or metal. As a result, those two energy bands can hybridize [52]. A hybridization gap then opens in the DOS as depicted in Fig. 7c, and the local DOS near the gap is strongly distorted. Classical Kondo insulators are CeNiSn, $CePd_3$, and $Ce_3Bi_4Pt_4$ [53]. Many decades of research in this field have shown that hybridization gaps are not easy to use in the design of thermoelectric materials, particularly when they involve *f*-levels. Yet the $CePd_2Pt$ [54] and $YbAl_2$ [55] systems hold some record ZT values in cryogenic cooling.

Spin

Very recently, spin has been added as a parameter to circumvent the difficulty inherent in thermoelectrics research, that is, the counter-indicated nature of the design parameters σ, κ, and α. Kondo systems rely on spin-dependent scattering, but the potential of thermal spin transport is still being investigated and the present understanding and future avenues of research are reviewed in this section. For the last 6 years, the exploration of spin caloritronic [56] effects has gained larger attention as a way of developing fundamentally new means for engineering thermal effects in materials, as well as providing methods to thermally generate pure spin fluxes. The first discovery in that field was the spin-Seebeck effect, found to exist in permalloy at Tohoku University in 2008 [57]. In 2010, the group at the Ohio State University discovered [58] the effect on a magnetic semiconductor, GaMnAs, and the Tohoku group found it on a ferromagnetic insulator and yttrium-iron garnet (YIG) at the same time [59]. Up to that time, the effect was measured in a geometry now called the transverse spin-Seebeck effect (TSSE). The Tohoku group discovered a simpler geometry on which the effect could be measured, the longitudinal Spin-Seebeck effect (LSSE). Because it can be confused with a conventional Nernst effect, the LSSE can be isolated only in electrically insulating ferromagnets [60]. Much larger TSSE effects were then found in InSb [61] using spin polarization of electrons by external magnetic fields rather than via ferromagnetism. Here, the giant spin-Seebeck effect can reach values of up to 8 mV/K, rivaling and even surpassing charge-based thermoelectrics, but only at cryogenic temperatures and high magnetic fields. The operating mechanism of the LSSE is now well established [62] and outlined next.

Figure 8 shows a diagram of an LSSE measurement using a YIG/Pt structure [56]. A 7 nm-thick Pt film is evaporated on the top surface of a crystal of YIG. A temperature gradient applied to that YIG from top to bottom generates a heat flux $j_Q = j_{QP} + j_{QM}$ that consists of heat carried by phonons (j_{QP}) and by spin waves or magnons (j_{QM}). The exact distribution between the phonon



Solid-State Heat Convertors, Fig. 8 Illustration of the longitudinal spin Seebeck effect (*LSSE*) geometry, a direct thermal spin injection from a magnon system in a ferromagnet (*FM*) into a normal metal (*NM*). The spin current polarizes the conduction electrons in the NM where, in the presence of spin-orbit interactions, it generates an electric field by the inverse spin-Hall effect (*ISHE*). *Energy Environ. Sci.*, 7 885 (2014) – Reproduced by permission of The Royal Society of Chemistry

and magnon conductivity was determined recently [63]. The magnon heat flux j_{QM} is directly related to the spin flux j_S carried by the magnons [56] $j_S = \left(\hbar j_{QM}/k_B T\right)$. This spin flux crosses the YIG/Pt interface at the top and spin-polarizes the conduction electrons in the Pt over a very small distance near the interface ($\sim$7 nm, the spin-diffusion length in Pt). Indeed, spin-flip transitions in the Pt quickly re-equilibrate the distribution of the electrons in the Pt so as to restore the same density of spin-up and spin-down electrons after one spin-diffusion length. The Pt film is thin enough that the electrons are spin-polarized through most of its thickness. Spin-polarized electrons in Pt are subject to the inverse spin-Hall effect (ISHE) [56]: due to skew scattering and other intrinsic effects, in solids made from heavy atoms, electrons scatter at different angles depending on their spin polarization. If there are more spin-up than spin-down electrons, the excess electrons accumulate on one side of the sample, creating an electric field proportional to the flux of spin-polarized electrons, which is equal to the spin flux. This is the inverse spin-Hall field $\mathbf{E_{ISHE}} = D(\mathbf{j_S} \times \boldsymbol{\sigma})$, where, in vector notation, $\mathbf{j}_S$ is the vector of spin flux (which, as we have seen above is thermally driven) and $\boldsymbol{\sigma}$ is the spin polarization vector (parallel to the applied magnetic field $\mathbf{H}$). Figure 8 shows the orientations of all these vectors. The end result is that a voltage appears across the Pt when a temperature gradient is applied to the ferromagnet, just like in a conventional thermoelectric or, more precisely, like in a Nernst effect measurement (the electric field induced by the Nernst effect, like $\mathbf{E_{ISHE}}$, is perpendicular to the temperature gradient). While it has been shown [61] in InSb that the field to temperature gradient ratio (the spin-Seebeck coefficient) can be as large as for classical thermopower effects, calculations [64] of the *ZT* of the YIG/Pt system is very poor (10^{-3}). The causes are reviewed here, as well as avenues of ongoing research.

The two main steps for the energy conversion is the transformation of the heat flux into a spin flux, and then the spin flux into a charge flux and voltage. The first step is similar to the case of classical thermoelectrics: the heat carried by electrons or magnons is equivalent to a charge or spin flux, both of which are useful, but the heat carried by phonons (the lattice) is a pure loss. Next, there are three loss mechanisms specific to the LSSE, and further research could prove that all three may be circumvented. First, the spin flux must cross the ferromagnet/metal interface, which is characterized by a parameter called the spin-mixing conductance ($g_{\uparrow\downarrow}$). This parameter is very well characterized [65]. Its physics is based on the interactions between the conduction electrons in the Pt and the core d-electrons on the ferromagnetic atoms in the YIG. It is quite akin to the mechanisms that limit the electrical resistivity of transition metals like Fe, where conduction

electrons interact with core d-electrons. The fact that this mechanism leads to a resistivity that is 10–100 times higher in transition metals than in copper proves that this is very efficient. While $g_{\uparrow\downarrow}$ depends on the quality of the interface, generally it is not a major cause of loss of spin polarization. The second loss is the conversion of spin flux into a voltage, i.e., the efficiency of the inverse spin-Hall effect. A review of this effect is given by Hoffmann [66]. From an engineering perspective, the conversion efficiency – included in the parameter D in the formula above – is at most 1–3 %, except in some narrow-gap semiconductors (InSb) which are not well studied yet. This loss is specific to LSSE. The third, and by far the worst, loss mechanism in the LSSE is the fact that only a very small fraction of the volume of the ferromagnet contributes to the LSSE. This was proven by Kehlberger et al. [67] who measured the LSSE as in Fig. 8, but using films of YIG grown on a nonmagnetic substrate (a gadolinium gallium garnet, GGG). They plotted the spin-Seebeck coefficient as a function of YIG film thickness and showed that the intensity of the effect increased with film thickness up to about 150 nm and saturated thereafter. This proves that only the spin flux generated in the 150 nm layer under the Pt gets converted into electrical energy. Because the temperature gradient is applied across the entire film plus substrate thickness (0.5 mm), only 0.03 % of that gradient does any useful work. This is the worst energy loss mechanism, but also the easiest to address: one could make multilayers of 100 nm YIG/7 nm Pt [68] – leaving the losses in the spin-Hall conversion as dominant loss mechanism. Such work is ongoing.

Finally, there is a connection between spin caloritronics and conventional thermoelectrics as pertains to metals: magnon drag. Usual metals have a thermopower limited to $|S| < 10\ \mu V\ K^{-1}$. In reality, metals used in thermocouples have values as high as $\pm 50\ \mu V\ K^{-1}$ or even higher; but even those values are not sufficient to give a good ZT. One thermocouple material is iron, with $\alpha > 15\ V\ K^{-1}$ at 300 K. As far back as the 1960s, this unusual behavior was speculatively interpreted as arising from a spin-based mechanisms called magnon drag [69]. We now understand [70] that magnon drag is in fact a form of self-spin-Seebeck effect, where the electrons that give rise to the voltage are not in a layer adjacent to the ferromagnet, but in the ferromagnet itself. Such an effect was observed on metallic glasses [71].

There is an extensive theoretical literature on magnon drag in metals, but the basic hydrodynamic theory [72] is presented here and similar to the McDonald theory [73] for phonon drag. Consider both magnons in ferromagnets and electrons as quasi-particles with properties similar to ideal gasses, including the fact that they have parabolic dispersion relations between energy and momentum. Note that the case of phonons and of magnons in antiferromagnets, which have linear dispersions, can be treated very similarly, but differ by a numerical factor of two. In all drag situations, the "dragging" quasi-particle, here the magnon, is assumed to undergo collisions that are predominantly with the "dragged" particles, here the electrons. This means that we consider the situation where magnon-electron interactions, at frequency ${\tau_{ME}}^{-1}$, constitute the dominant magnon scattering mechanism, i.e., ${\tau_{ME}}^{-1} \approx {\tau_M}^{-1}$ (the total magnon collision frequency being ${\tau_M}^{-1}$). Consider the magnons as an isotropic ideal gas, with an internal energy U: the magnons will exert a pressure on the electrons with which they collide given by $p = 2/3U$. Now, apply a temperature gradient dT/dx along a specific direction x: this will give rise to a pressure gradient in the magnon population. The pressure gradient will give rise to a force F_x per unit volume, selectively applied to the electron population, that, since the problem is one-dimensional, is given by:

$$F_x = -\frac{dp}{dx} = -\frac{2}{3}\frac{dU}{dx} = -\frac{2}{3}\frac{dU}{dT}\frac{dT}{dx}. \qquad (11)$$

This is the force that impels a unidirectional momentum onto the electron population, so that their momentum along the direction x, i.e., $\hbar k_x$, is modified. This will give rise to an instantaneous magnon-drag thermoelectrical current in the sample proportional to $-dT/dx$. When no current is permitted to flow, it will result in an accumulation of electrons and the creation of an electric field E_x

that will act on the electrons to generate a force such that the force neE_x. The balance of forces means that:

$$F_x + neE_x = 0$$
$$neE_x = \frac{2}{3}\frac{dU}{dT}\frac{dT}{dx} = \frac{2}{3}C_M\frac{dT}{dx}. \quad (12)$$

where C_M is the magnon-specific heat. Given the definition of the thermopower α as $E_x/(dT/dx)$, it results that its magnon-drag component α_{MED} is then simply:

$$\alpha_{MED} = \frac{2C_M}{3ne}. \quad (13)$$

Reconsider now the situation where there are other mechanisms that scatter magnons besides interactions with electrons, i.e., where ${\tau_{ME}}^{-1} < {\tau_M}^{-1}$. The other scattering mechanisms will absorb a fraction of the magnon dragging force that would have been impelling the electrons in the above discourse, reducing it by a fraction ${\tau_{ME}}^{-1}/{\tau_M}^{-1}$. The same fraction then is reflected in the magnon-drag thermopower, which becomes:

$$\alpha_{MED} = \frac{2C_M}{3ne}\frac{\tau_M}{\tau_{ME}}. \quad (14)$$

In reality, one has to consider the above argument for each specific mode and energy of the entire magnon and electron populations, so that the exact solutions for α_{MED} involve integrals over the whole Brillouin zone of both electron and magnon dispersions, and this is what is done in the specialized literature. Nevertheless, one can extract important physical conclusions about α_{MED} from just Eq. 14. First, drag becomes more important as the carrier concentration is reduced; for the case of phonon drag, values exceeding 10 mV/K are easily observed in Si and Ge. Second, at low temperature and in large, perfect crystals where the condition ${\tau_{ME}}^{-1} \approx {\tau_M}^{-1}$ holds, magnon-drag is expected to mimic the temperature dependence of $C_M \propto T^{3/2}$, just like phonon drag follows the Debye lattice-specific heat, a T^3 law. Third, at a sufficiently high temperature where ${\tau_{ME}}^{-1} < {\tau_M}^{-1}$, α_{MED} will follow the temperature dependence of $C_M \times \tau_M$, because there is not much temperature dependence to magnon-electron interactions, which are generally due to s-d interactions. It is on this point that magnon drag differs strongly from phonon drag: in the case of phonons, the Debye-specific heat saturates at the Debye temperature Θ_D. Further, in the very pure crystals where phonon drag is observed, anharmonic scattering is the dominant mechanism limiting the lattice thermal conductivity for all temperatures $T > \Theta_D/3$ to $\Theta_D/10$. Phonon-drag therefore is a low-temperature phenomenon in most solids with $\Theta_D < 1000$ K. Not so for magnon drag, because the magnon density keeps increasing up to the Curie temperature, which can be quite high in ferromagnetic metals, and also because s-d scattering remains strong at all temperatures. Magnon-drag or other spin-mediated thermopowers in ferromagnetic metals easily extend to and exceed room temperature.

Conclusions

In summary, the Hicks and Dresselhaus papers [33, 34] gave a renewed impetus to research on thermoelectric solid-state energy conversion, which has great potential to recover heat otherwise wasted into the environment and convert it to useful energy or to provide thermodynamic topping cycles that increase the thermal efficiencies. The last 20 years of research have resulted in an increase of the thermoelectric figure of merit from that value it was at in the 1970s $ZT \approx 1$, to the value obtained on SnSe this summer, $ZT \approx 2.6$ [28]. This ZT reflects the increase in thermal efficiency of future devices that could be built out of the new materials, although the development of thermoelectric devices has not kept pace with the scientific discoveries yet [12]. The role of spin in thermoelectrics and spin caloritronics is a new avenue, largely unexplored, that promises more progress. Speculatively, perhaps future emphasis could be placed on materials that are more earth abundant than the elements currently used, and also easier to produce, shape, and contact than semiconductors. A good place to start would be to develop high-ZT metals.

Acknowledgments The author acknowledges help with the manuscript editing and figures from Ms. Renee Ripley and support from the NSF MRSEC program, Grant No. DMR 1420451.

Cross-References

- Heat Capacity
- Heat Conduction
- Heat Conductivity
- Heat Transfer in Semiconductor Nanostructures
- Nanostructures for Energy
- Spintronic Devices
- Thermal Conductance
- Thermal Conductivity and Phonon Transport
- Thermal Resistance
- Thermal Resistivity
- Thermal Transport

References

1. Telkes, M.: The efficiency of thermoelectric generators. I. J. Appl. Phys. **18**(12), 1116–1127 (1947)
2. Telkes, M.: Power output of thermoelectric generators. J. Appl. Phys. **25**(8), 1058–1059 (1954)
3. Ioffe, A.F.: Semiconductor Thermoelements and Thermoelectric Cooling. Inforsearch, London (1957)
4. Goldsmid, H.J.: Thermoelectric Refrigeration. Plenum Press, New York (1964)
5. Seebeck, T.J.: Magnetische Polarisation der Metalle und Erze durch Temperatur-Differenz. In: Abhandlungen der Preussischen Akad, Wissenschaften, pp. 265–373 (1822–1823). Reprinted W. Engelmann, Leipzig (1895)
6. Thomson, W.: On the dynamical theory of heat. Part V. Thermo-electric currents. Trans. R. Soc. Edinb. **21**, 123–171 (1857). https://archive.org/stream/transactionsofro21royal#page/n3/mode/2up
7. Lord Rayleigh, F.R.S.: On the thermodynamic efficiency of the thermopile. Philos. Mag. **20**, 361–363 (1885)
8. Altenkirch, E.: Über den Nutzeffekt der Thermosäule. Phys. Ztg. **10**(16), 560–568 (1909)
9. Callen, H.B.: Thermodynamics and an Introduction to Thermostatistics. Wiley, New York (1960)
10. Roberts, R.B.: Absolute scales for thermoelectricity. Measurement **4**(3), 101–103 (1986). doi:10.1016/0263-2241(86)90016-3; Roberts, R.B.: Absolute scale of thermoelectricity. Philos. Mag. **36**, 91 (1977); Roberts, R.B.: Absolute scale of thermoelectricity II. Philos. Mag. B **43**, 1123 (1981); Roberts, R.B., Righini, F., Compton, R.C.: Absolute scale of thermoelectricity III. Philos. Mag. B **52**, 1147 (1985)
11. Vining, C.B.: An inconvenient truth about thermoelectrics. Nat. Mater. **8**, 83–85 (2009)
12. Heremans, J.P., Dresselhaus, M.S., Bell, L., Morelli, D.T.: When thermoelectrics reached the nanoscale. Nat. Nanotechnol. **8**, 471–473 (2013)
13. Ioffe, A.F.: Physics of Semiconductors. Academic, New York (1960)
14. Berman, R.: Thermal Conduction in Solids. Clarendon, Oxford (1976)
15. Morelli, D.T., Heremans, J.P., Slack, G.A.: Estimation of the isotope effect on the lattice thermal conductivity of group IV and group III-V semiconductors. Phys. Rev. B **66**, 195304 (2002)
16. Poudel, B., Hao, Q., Ma, Y., Lan, Y., Minnich, A., Yu, B., Yan, X., Wang, D., Muto, A., Vashaee, D., Chen, X., Liu, J., Dresselhaus, M.S., Chen, G., Ren, Z.: High-thermoelectric performance of nanostructured bismuth antimony telluride bulk alloys. Science **320**, 634–638 (2008)
17. Lindsay, L., Broido, D.A., Reinecke, T.L.: Ab initio thermal transport in compound semiconductors. Phys. Rev. B **87**, 165201 (2013)
18. Morelli, D.T., Meisner, G.P.: Low temperature properties of the filled skutterudite CeFe4Sb12. J. Appl. Phys. **77**, 3777 (1995)
19. Meisner, G.P., Morelli, D.T., Hu, S., Yang, J., Uher, C.: Structure and lattice thermal conductivity of fractionally filled skutterudites: solid solutions of fully filled and unfilled end members. Phys. Rev. Lett. **80**, 3551 (1998)
20. Sales, B.C., Mandrus, D., Williams, R.K.: Filled skutterudite antimonides: a new class of thermoelectric materials. Science **272**, 1325–1328 (1996)
21. Shi, X., Yang, J., Salvador, J.R., Chi, M.F., Cho, J.Y., Wang, H., Bai, S.Q., Yang, J.H., Zhang, W.Q., Chen, L.D.: Multiple-filled skutterudites: high thermoelectric figure of merit through spearately optimizing electrical and thermal transports. J. Am. Chem. Soc. **133**, 7837 (2011)
22. Nielsen, M.D., Ozolins, V., Heremans, J.P.: Lone pair electrons minimize lattice thermal conductivity. Energy Environ. Sci. **6**(2), 570–578 (2013)
23. Jovovic, V., Heremans, J.P.: Doping effects on the thermoelectric properties of $AgSbTe_2$. J. Electron. Mater. **38**, 1504–1509 (2009)
24. Jovovic, V., Heremans, J.P.: Energy band gap and valence band structure of $AgSbTe_2$. Phys. Rev. B **77**, 245204 (2008)
25. Morelli, D.T., Jovovic, V., Heremans, J.P.: Intrinsically minimal thermal conductivity in cubic I-V-VI_2 semiconductors. Phys. Rev. Lett. **101**, 035901 (2008)
26. Lu, X., Morelli, D.T., Xia, Y., Zhou, F., Ozolins, V., Chi, H., Uher, C.: High performance thermoelectricity in earth-abundant compounds based on natural mineral tetrahedrites. Adv. Energy Mater. **3**, 342–348 (2012)
27. Delaire, O., Ma, J., Marty, K., May, A.F., McGuire, M.A., Du, M.-H., Singh, D.J., Podlesnyak, A., Ehlers, G., Lumsden, M.D., Sales, B.C.: Giant anharmonic

phonon scattering in PbTe. Nat. Mater. **10**, 614–619 (2011)
28. Zhao, L.-D., Lo, S.-H., Zhang, Y., Sun, H., Tan, G., Uher, C., Wolverton, C., Dravid, V.P., Kanatzidis, M. G.: Ultralow thermal conductivity and high thermoelectric figure of merit in SnSe crystals. Nature **508**, 373–377 (2014)
29. Heremans, J.P.: Thermoelectricity: the ugly duckling. Nature **508**, 327–328 (2014)
30. Heremans, J.P., Thrush, C.M., Morelli, D.T.: Thermopower enhancement in PbTe with Pb precipitates. J. Appl. Phys. **98**, 063703 (2005)
31. Girard, S.N., He, J., Zhou, X.Y., Shoemaker, D., Jaworski, C.M., Uher, C., Dravid, V.P., Heremans, J. P., Kanatzidis, M.G.: High performance Na-doped PbTe-PbS thermoelectric materials: electronic density of states modification and shape-controlled nanostructures. J. Am. Chem. Soc. **133**, 16588–16597 (2011)
32. Biswas, K., He, J.Q., Blum, I.D., Chun, I.W., Hogan, T.P., Seidman, D.N., Dravid, V.P., Kanatzidis, M.G.: High-performance bulk thermoelectrics with all-scale hierarchical architectures. Nature **490**, 414–418 (2012)
33. Hicks, L.D., Dresselhaus, M.S.: Effect of quantum-well structures on the thermoelectric figure of merit. Phys. Rev. B **47**, 12727–12731 (1993)
34. Hicks, L.D., Dresselhaus, M.S.: Thermoelectric figure of merit of a one-dimensional conductor. Phys. Rev. B **47**, 16631–16634 (1993)
35. Mahan, G.D., Sofo, J.O.: The best thermoelectric. Proc. Natl. Acad. Sci. U. S. A. **93**, 7436–7439 (1996)
36. Cutler, M., Mott, N.F.: Observation of Anderson localization in an electron gas. Phys. Rev. **181**, 1336 (1969)
37. Heremans, J.P.: Low-dimensional thermoelectricity. Acta Phys. Polon. **108**, 609–634 (2005)
38. Murata, M., Nakamura, D., Hasegawa, Y., Komine, T., Taguchi, T., Nakamura, S., Jovovic, V., Heremans, J. P.: Thermoelectric properties of bismuth nanowires in a quartz template. Appl. Phys. Lett. **94**, 192104 (2009)
39. Heremans, J., Thrush, C.M.: Thermoelectric power of bismuth nanowires. Phys. Rev. B **59**, 12579 (1999)
40. Murata, M., Nakamura, D., Hasegawa, Y., Komine, T., Taguchi, T., Nakamura, S., Jaworski, C.M., Jovovic, V., Heremans, J.P.: Mean free path limitation of thermoelectric properties of bismuth nanowire. J. Appl. Phys. **105**, 113706 (2009)
41. Heremans, J., Thrush, C.M., Lin, Y.-M., Cronin, S.B., Dresselhaus, M.S.: Transport properties of antimony nanowires. Phys. Rev. B **63**, 085406 (2001)
42. Heremans, J.P., Thrush, C.M., Morelli, D.T., Wu, M.-C.: Thermoelectric power of bismuth nanocomposites. Phys. Rev. Lett. **88**, 216801 (2002)
43. Heremans, J.P., Thrush, C.M., Morelli, D.T., Wu, M.-C.: Resistance, magnetoresistance and thermopower of zinc nanowire composites. Phys. Rev. Lett. **91**, 076804 (2003)
44. Heremans, J.P., Thrush, C.M., Morelli, D.T.: Thermopower enhancement in lead telluride nanostructures. Phys. Rev. B **70**, 115334 (2004)
45. Heremans, J.P., Wiendlocha, B., Chamoire, A.M.: Resonant levels in bulk thermoelectric semiconductors. Energy Environ. Sci. **5**, 5510–5530 (2012)
46. Heremans, J.P., Jovovic, V., Toberer, E.S., Saramat, A., Kurosaki, K., Charoenphakdee, A., Yamanaka, S., Snyder, G.J.: Enhancement of thermoelectric efficiency in PbTe by distortion of the electronic density of states. Science **321**, 554–558 (2008)
47. Jaworski, C.M., Kulbachinskii, V.A., Heremans, J.P.: Tin forms a resonant level in Bi_2Te_3 that enhances the room temperature thermoelectric power. Phys. Rev. B **80**, 233201 (2009)
48. Daybell, M.D., Steyert, W.A.: Localized magnetic impurity states in metals: some experimental relationships. Rev. Mod. Phys. **40**, 380 (1968)
49. Heeger, A.J.: Localized moments and nonmoments in metals. In: Seitz, F., Turnbull, D., Ehrenreich, H. (eds.) Solid State Physics, vol. 23, pp. 284–407. Academic, New York (1969)
50. Kondo, J.: Resistance minimum in dilute magnetic alloys. Prog. Theor. Phys. **34**, 372 (1965)
51. Blatt, F.J., Schroeder, P.A., Foiles, C.L., Greig, D.: Thermoelectric Power of Metals. Plenum Press, New York (1976)
52. Mahan, G.D.: Good thermoelectrics. In: Ehrenreich, H., Spaepen, F. (eds.) Solid State Physics, vol. 51, pp. 81–152. Academic, New York (1997)
53. Fisk, Z., Sarrao, J.L., Thompson, J.D.: Heavy fermions. Curr. Opin. Solid State Mater. Sci. **1**, 42 (1996), Ce3Bi4Pt3, and CePd3 42 (1996)
54. Boona, S.R., Morelli, D.T.: Enhanced thermoelectric properties of $CePd_{3-x}Pt_x$. Appl. Phys. Lett. **101**, 101909 (2012)
55. Lehr, G.J., Morelli, D.T., Jin, H., Heremans, J.P.: Enhanced thermoelectric power factor in Yb_{1-x} Sc_x Al_2 alloys using chemical pressure tuning of the Yb valence. J. Appl. Phys. **114**, 223712 (2013)
56. Boona, S.R., Myers, R.C., Heremans, J.P.: Spin caloritronics. Energy Environ. Sci. **7**, 885–910 (2014). doi:10.1039/C3EE43299H
57. Uchida, K., Takahashi, S., Harii, K., Ieda, J., Koshibae, W., Ando, K., Maekawa, S., Saitoh, E.: Observation of the spin Seebeck effect. Nature **455**, 778–781 (2008)
58. Jaworski, C.M., Yang, J., Mack, S., Awschalom, D.D., Heremans, J.P., Myers, R.C.: Observation of the spin-Seebeck effect in a ferromagnetic semiconductor. Nat. Mater. **9**, 898–903 (2010)
59. Uchida, K., Xiao, J., Adachi, H., Ohe, J., Takahashi, S., Ieda, J., Ota, T., Kajiwara, Y., Umezawa, H., Kawai, H., Bauer, G.E.W., Maekawa, S., Saitoh, E.: Spin Seebeck Insulator. Nat. Mater. **9**, 894–897 (2010). doi:10.1038/NMAT2856
60. Uchida, K.-i., Adachi, H., Ota, T., Nakayama, H., Maekawa, S., Saitoh, E.: Observation of longitudinal

spin-Seebeck effect in magnetic insulators. Appl. Phys. Lett. **97**, 172505 (2010)
61. Jaworski, C.M., Myers, R.C., Johnston-Halperin, E., Heremans, J.P.: Giant spin Seebeck effect in a non-magnetic material. Nature **487**, 210–213 (2012)
62. Hoffman, S., Upadhyaya, P., Tserkovnyak, Y.: Landau-Lifshitz theory of the longitudinal spin Seebeck effect. Phys. Rev. B **88**, 064408 (2013)
63. Boona, S.R., Heremans, J.P.: Magnon thermal mean free path in yttrium iron garnets. Phys. Rev. B **90**, 064421 (2014). doi:10.1103/PhysRevB.90.064421
64. Kovalev, A.A., Tserkovnyak, Y.: Magnetocaloritronic nanomachines. Solid State Commun. **150**, 500–504 (2010)
65. Weiler, M., Althammer, M., Schreier, M., Lotze, J., Pernpeintner, M., Meyer, S., Huebl, H., Gross, R., Kamra, A., Xiao, J., Chen, Y.-T., Jiao, H.J., Bauer, G.E.W., Goennenwein, S.T.B.: Experimental test of the spin mixing interface conductivity concept. Phys. Rev. Lett. **111**, 176601 (2013)
66. Hoffmann, A.: Spin Hall effects in metals. IEEE Trans. Magn. **49**, 5172–5193 (2013)
67. Kehlberger, A., Ritzmann, U., Hinzke, D., Guo, E.-J., Cramer, J., Jakob, G., Onbasali, M. C., Kim, D. H., Ross, C. A., Jungfleisch, M. B., Hillebrands, B., Nowak, U., and Kläui, M.: Length scale of the spin Seebeck effect, Phys. Rev. Lett. **115**, 096602 (2015)
68. Heremans, J., Jaworski, C.: Spin thermoelectric generator with multiple magnetic layers. Ohio State University Technology Commercialization Office, Invention disclosure T2012-251 (2012)
69. Blatt, F.J., Flood, D.J., Rowe, V., Schroeder, P.A., Cox, J.E.: Magnon-drag thermopower in iron. Phys. Rev. Lett. **18**, 395 (1967)
70. Lucassen, M.E., Wong, C.H., Duine, R.A., Tserkovnyak, Y.: Spin-transfer mechanism for magnon-drag thermopower. Appl. Phys. Lett. **99**, 262506 (2011)
71. Jin, H., Yang, Z., Myers, R.C., Heremans, J.P.: Spin-Seebeck like signal in ferromagnetic bulk metallic glass without platinum contacts. Solid State Commun. **198**(Special Issue on Spin Mechanics), 40–44 (2014)
72. Watzman, S.J., Duine, R.A., Tserkovnyak, Y., Jin, H., Prakash, A., Zheng, Y., and Heremans, J.P.: Magnon-drag thermopower and Nernst coefficient in Fe and Co, arXiv 1603.03736 (2016)
73. MacDonald, D.C.K.: Thermoelectricity: An Introduction to the Principles. Wiley, New York (1962)

Sound Propagation in Fluids

▶ Acoustic Nanoparticle Synthesis for Applications in Nanomedicine

Spectromicroscopy

▶ Selected Synchrotron Radiation Techniques

Spectroscopic Techniques

▶ Optical Techniques for Nanostructure Characterization

Spectroscopy of Ancient Documents

M. Missori[1], O. Pulci[1] and A. Mosca Conte[2]
[1]Institute for Complex Systems, National Research Council, Rome, Italy
[2]ETSF, and Department of Physics, University of Rome Tor Vergata, Rome, Italy

Synonyms

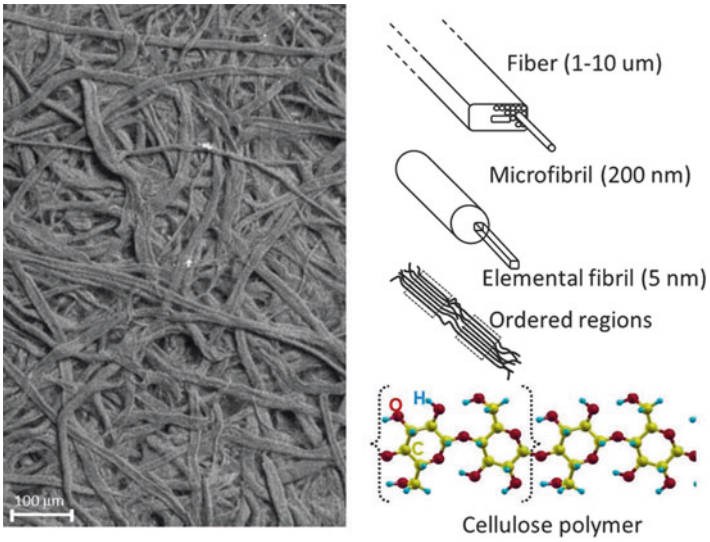
Optical reflectance spectroscopy of ancient paper; Optical spectroscopy for ancient paper diagnostic

Definition

Optical spectroscopy of ancient paper is the study of the interaction between paper and ultraviolet, visible and near infrared radiation. This approach is aimed to nondestructive evaluation of the degradation of ancient artifacts.

Cellulose and Ancient Paper

Cellulose has had a unique role in civilization because it has been widely used for the acquisition, storage, and dissemination of human culture. In fact, one of the main cellulose derivatives is paper, which was invented in China in the second century A.D. and, independently, in Mesoamerica in the seventh century A.D. The technological



Spectroscopy of Ancient Documents, Fig. 1 Image of Whatman no. 1 paper sample, without prior preparation, obtained by using a Zeiss EVO40 variable pressure electron microscope. A random network of many thick fibers tangled together to form a complex structure can be observed (*left*). Schematic view of the supermolecular architecture of cellulose polymers within a fiber (*right*)

knowledge of making paper spread, in the Middle Ages, from the Far East to Arab World and Europe. As a result, for centuries, a growing number of books, documents, and artistic drawings have been accumulating in archives, libraries, and museums, all over the world [1].

The preservation of this cultural heritage is based on limiting the degradation of the paper materials. To this aim, optical spectroscopy allows an advanced knowledge of the nanoscale characteristics of paper materials and their degradation processes.

Paper is a complex multicomponent material consisting mainly of cellulose fibers. A paper sheet is obtained from a dilute suspension of cellulose fibers in water that are then drained through a sieve, pressed and dried, to obtain a network of randomly interwoven fibers. The paper composition varies depending on the production period and technology employed. In the Middle Ages in Europe, paper was made up of pure cellulose fibers (>90 % in weight) from cotton, linen, or hemp, usually obtained from rags with the addition of animal glue as a sizing agent [1]. Besides cellulose, most of modern paper also contains a relevant quantity of lignin and fillers.

Cellulose is the most abundant organic compound on Earth representing 40 % of the annual production of biomass [2]. It is a linear homopolymer composed of β-D-glucopyranose units $(C_6H_{10}O_5)_n$, which are linked together by β-(1,4)-glycosidic bonds up to form chains with n from about 100 to about 10,000 elements (see Fig. 1). The cellulose chains have a strong tendency to aggregate into highly ordered structural entities through an extended network of both intra- and intermolecular hydrogen bonds. As a consequence, a hierarchical arrangement of polymers is formed, from elemental fibrils through microfibrils up to fibers, whose diameter ranges approximately from 1 to 10 µm [3, 4].

Elemental fibrils, basic components of the cellulose supramolecular structure, include assembly of highly ordered (crystalline) domains and of disordered (amorphous-like) regions [2–4] (Fig. 1). Crystalline domains constitute from around 60 % to 70 %, of total cellulose material depending on its origin and history.

Aging of Cellulose and Paper

On macroscopic scale, paper degradation occurs by the weakening of mechanical properties of the sheets and by yellowing. At the nanoscale, cellulose degradation can be seen as the combination of two most important processes: acid hydrolysis of β-(1,4)-glycosidic bonds, and oxidation of the β-D-glucopyranose units with the subsequent development of various chemical products [3, 5].

The degradation extent in paper after a period of time depends on the environmental conditions to which it has been subjected [5]. While the mechanism of acidic hydrolysis is rather known and well settled in the literature, cellulose oxidation running through the radical mechanism initiated by active oxygen species is a complex process with many possible reactions still to be clarified [2–5].

Pristine cellulose does not absorb light up to about 200 nm (below 6 eV of photon energy) [5]. Paper yellowing is due to some oxidation products which are called chromophores. The yellow color seen in aged papers is mainly due to the fact that chromophores in paper absorb the higher energy band of visible light (corresponding to violet and blue) and largely scatter the yellow and red portion, thereby producing the characteristic yellow-brown hue [6].

Optical Spectroscopy of Inhomogeneous Materials

A medium is optically inhomogeneous when it is characterized by variations of the dielectric function ε over distances comparable to those of the wavelength of light impinging on the sample. This induces the scattering of light, namely, the deflection of photons from a straight path [7]. As a consequence, the usual laws related to optical properties of homogeneous materials are no more valid [8]. In general, scattering of photons is caused by, for example, irregularities in the propagation medium, presence of particles, or roughness in the interface between two media. In the presence of scattering, the reflectivity of materials is commonly referred to as diffuse reflectance [8]. Paper sheets being composed by rough-surfaced fibers and voids represent a striking example of materials whose optical properties are strongly governed by light-scattering effects.

In order to separate the scattering from the absorption, a number of theories have been developed [8]. Among those, some phenomenological models have been applied to describe paper optical properties [9]. However, the most successful approach is the Kubelka-Munk (KM) model [8], an approach using two radiative fluxes to the general theory of radiation transfer in optically inhomogeneous substances (such as a sheet of paper). Using this model, it is possible to correlate reflectance measurements to the absorption coefficient of the elements that make up the inhomogeneous material.

KM model assumes that light propagates in just two opposing directions in an inhomogeneous material, with the flux variation at any point in the medium being linearly proportional to the two local opposing fluxes. The proportionality constants called scattering, S, and absorption, K, coefficients are assumed dependent on the intrinsic scattering and absorption properties of the medium. The radiative flux I which propagates downward ($-z$ direction) in a material represents the averaged intensity of all rays directed toward the lower hemisphere. Similarly, the flow J which propagates upward ($+z$ direction) is the average intensity of all rays directed toward the upper hemisphere. The behavior of J and I is described by the system of differential equations:

$$\begin{aligned} -\frac{dI}{dz} &= -(K+S)I + SJ \\ \frac{dJ}{dz} &= -(K+S)J + SI \end{aligned} \tag{1}$$

where z is the upward coordinate of the thickness of the material and S and K coefficients have both physical dimensions of the reciprocal of length. The system of Eq. 1 can be solved analytically and provides the so-called *Kubelka-Munk function*:

$$A_{KM} = \frac{K}{M} = \frac{(1-R_\infty)^2}{2R_\infty} \tag{2}$$

where R_∞ is the measured reflectance, i.e., the ratio of the reflected flux to the incident flux at the sample surface, when the sample is sufficiently thick that no incident radiation passes through it.

K and S coefficients can also be calculated individually for thin samples from KM model. To this aim, two separate reflectance measurements obtained by laying the sheets on a white (R_w) backing, and then a black (R_b) backing must

be carried out. The reflectance values for the white (R_{wb}) and black (R_{bb}) backings must be measured separately.

Hence, the reflectance R_∞ that an infinitely thick layer of the same sample would have, can be calculated by using the results of the original KM model [8, 9], from R_w, R_b, R_{wb}, and R_{bb}:

$$R_\infty = a - \sqrt{a^2 - 1} \tag{3}$$

and

$$a = \frac{1}{2}\frac{(R_{wb} - R_{bb})(1 + R_w R_b) - (R_w - R_b)(1 + R_{wb}R_{bb})}{R_b R_{wb} - R_w R_{bb}} \tag{4}$$

When R_∞ is known, it is possible to derive the phenomenological scattering S and absorption K coefficients:

$$S = \frac{1}{t\left(\frac{1}{R_\infty} - R_\infty\right)} \ln \frac{(1 - R_b R_\infty)(R_\infty - R_{bb})}{(1 - R_\infty R_{bb})(R_\infty - R_b)} \tag{5}$$

and

$$K = \frac{S(1 - R_\infty)^2}{2R_\infty} \tag{6}$$

where t is the thickness of the inhomogeneous sample [8].

The KM theory has been extended by Yang and Miklavcic (YM) to take into account samples with stronger optical absorption [10]. Following this approach, the intrinsic absorption α and scattering s coefficients of paper, in the case of uniform distribution of the incident and reflected light (obtained, e.g., by means of an integrating sphere), can be recovered as

$$\alpha = \frac{K}{2\mu} \qquad s = \frac{S}{\mu} \tag{7}$$

The quantity $\mu = \mu(s,\alpha)$ is the scattering-induced path variation (SIPV) factor [10]. It describes the influence of light scattering on the total path length and is nonlinearly dependent on both the absorption and scattering properties of the medium.

Recently, an extension of the YM model to optically thick samples has been developed [11].

A new expression for the scattering coefficient of light s_∞ was proposed in this regime:

$$s_\infty \propto \sqrt{\frac{e^{-\frac{1}{\sqrt{1+\frac{2}{A_{KM}}}}}}{\left[4\left(A_{KM}^2 + 2A_{KM}\right)\right]^{\frac{1}{4}}}}, \tag{8}$$

and its validity for strongly absorbing paper samples was demonstrated in Ref. [11].

In this way, the cellulose absorption coefficient can be recovered up to the highly absorbing ultraviolet region from nondestructive reflectance measurements.

In order to measure the absolute diffuse reflectance of paper samples, an experimental setup using an integrating sphere to illuminate the sample with diffuse radiation and to collect radiation reflected over all angles is necessary in order to apply the KM and YM models [9, 11].

Recovering the optical absorption of cellulose fibers is an important step in order to make comparisons with theoretical simulations and obtain information on the chemical degradation of paper induced by aging and causing its yellowing.

S

Theoretical Optical Spectroscopy

Because of the complexity of paper structure, it is not possible to associate optical absorption peaks to specific oxidized groups by a simple comparison with spectra of reference compounds. Due to the important role of the local chemical environment around the oxidized groups, which can result in geometrical distortions and in energy level shifts, their optical properties can be strongly modified making the above comparison unreliable.

A possible pathway to tackle this problem relies on the comparison of the experimental optical spectra with those obtained by computational simulations based on ab initio methods. These allow to simulate physical properties of materials (such as the geometry, the electronic and optical properties, and so on) starting from the microscopic knowledge of their chemical structure. The optical spectra of paper can therefore be calculated using these techniques and compared with the experiments.

Density functional theory [12], for which Walter Kohn was awarded with the Nobel Prize for Chemistry in 1998, represents by now the state of the art of the first principle methods for the calculations of the geometry and all ground-state properties of any system of interacting electrons and ions.

Within DFT the many-body problem:

$$H_e\Psi = \left[-\sum\frac{1}{2}\nabla_i^2 - \sum\frac{Z_A}{r_{i,A}} + \sum\frac{1}{r_{i,j}} + \sum\frac{Z_A Z_B}{R_{A,B}}\right]$$
$$\Psi(r_1, r_2, \ldots r_N) = E\Psi(r_1, r_2, \ldots r_N) \tag{9}$$

is mapped into a system of noninteracting particles that can be solved by finding the solution of a single particle equation (known as Kohn-Sham equation):

$$\left(-\frac{1}{2}\nabla^2 + V_{\text{ext}}(r)\right)\psi_{n,k}^{KS}(r) = \varepsilon_{n,k}^{KS}\psi_{n,k}^{KS}(r) \tag{10}$$

with, by construction, ground-state electron density identical to the one of the interacting system. In Eq. 9, the ions are considered as fixed point charges with electric charges Z_A and Z_B and respective distances $R_{A,B}$, while electrons are instead considered by quantum mechanics governed by the Schrodinger equation. The elements of the Hamiltonian reported in Eq. 9 are in order: the electronic kinetic energy, the electron-ion interaction, the electron-electron interaction, and the ion-ion interaction.

In Eq. 10, V^{ext} is the external potential and includes the electron-ion interaction, the Hartree potential, and the so-called exchange and correlation potential, which is defined as the functional derivative of the exchange and correlation energy. The external potential is a functional of the electronic charge density and is constructed in such a way that the noninteracting electron system ground-state charge density is the same as the one of the interacting electron system. The total energy of the interacting system is univocally determined by its ground-state electron density, as demonstrated by the Hohenberg and Kohn theorem within the density functional theory [12]; hence the knowledge of its ground-state density (through the solution of Eq. 10) should in principle give us access to the ground-state energy. The problem arises from the fact that the expression of the total energy as a functional of the electron density is unknown. In particular, the analytical form of one important ingredient, the so-called exchange and correlation energy, is unknown, and as a consequence, also the exact exchange and correlation potential is unknown. Approximations have to be used, such as LDA, GGA, and so on. For this reason, even if DFT is a formally exact theory, the total energy of the interacting system can only be evaluated within approximations. The structure and the main theorems on which the DFT is based can be found in several references (see, e.g., Ref. [12]), as well as the several approximation used to evaluate the exchange and correlation potential.

If ground-state properties can be successfully calculated within DFT, in order to compute optical spectra, which are an excited state property, a different approach has to be used. Time-dependent density functional theory (TDDFT [13]) can be considered as an extension of DFT to time-dependent external potentials and a (in principle) rigorous approach for optical spectra calculations.

Methods based on TDDFT allow to calculate the response function that relates the time-dependent external electric field to the response of the electronic charge density inducing a polarization in the material:

$$\chi(r, r'; t - t')_{\alpha,\beta} = \frac{\partial P_\alpha(r,t)}{\partial E_\beta(r', t')} \tag{11}$$

Here, P is the polarization of the material and E the external electric field. The response function χ is related to the optical properties of the material (dielectric function, optical absorption, reflection, and so on). TDDFT allows to obtain the response function of an interacting electron system from the one calculated for a noninteracting electron system (χ_0) by the following recursive Dyson-like equation:

$$\chi = \chi_0 + \chi_0(v_c + f_{xc})\chi \qquad (12)$$

where v_c is the Coulomb potential and f_{xc} is the exchange-correlation kernel. The kernel is a functional of the electronic charge density. Its exact analytical expression is not known, and several approximations are used to estimate it (as, e.g., ALDA, APBE, ABLYP, and other usually taken from DFT with adiabatic approximation) [13].

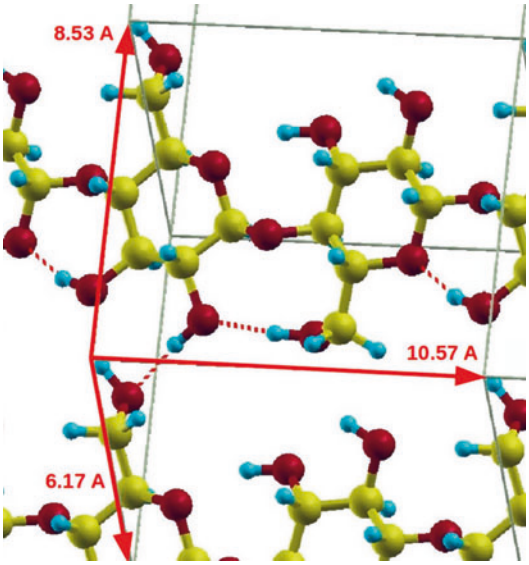
TDDFT calculations of optical spectra in complex systems represent by now the state of the art especially in molecules, nanoclusters, and in biological systems.

Theoretical Optical Spectra for Oxidized Cellulose

In order to interpret the experimental optical absorption spectra and recover chemical information at the nanoscale, the knowledge of the optical absorption spectra of the various chromophores, which are supposed to be responsible for aged paper yellowing, is needed.

Density functional theory and time-dependent density functional theory methods can be applied to the study of cellulose, as reported in Refs. [6, 14]. In these works, as a structural model for cellulose, an infinitely extended crystal of cellulose has been chosen as an approximation of cellulose crystalline domains. Cell parameters of cellulose crystal were obtained by X-ray diffraction data [2–4] and refer to the monoclinic crystallographic phase called cellulose-I_β, which is the most common polymorph for higher-plant cellulose [3, 4] (see Fig. 2).

Cellulose with several forms of oxidized groups has been considered in the simulations.



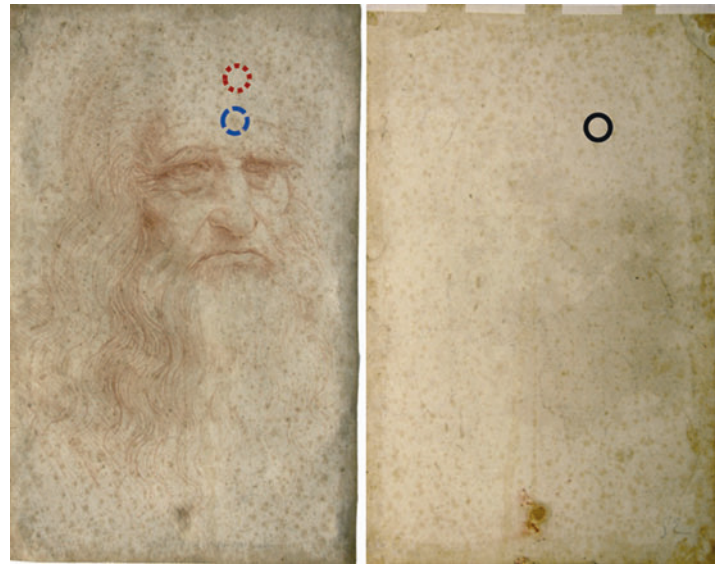
Spectroscopy of Ancient Documents, Fig. 2 Monoclinic crystallographic phase of cellulose-I_β, with cell parameters taken from X-ray diffractions [3, 4] and atomic coordinates inside the periodic cell calculated by DFT. Light-gray (*yellow*), dark-gray (*red*), and small (*blue*) spheres represent, respectively, carbon, oxygen, and hydrogen atoms. *Dashed red lines* indicate hydrogen bonds

Simulated theoretical optical spectra can be compared to experimental ones by a fitting procedure. This comparison allows the characterization and estimation of the quantity of oxidized groups acting as chromophores inside aged paper starting from nondestructive optical experimental measurements.

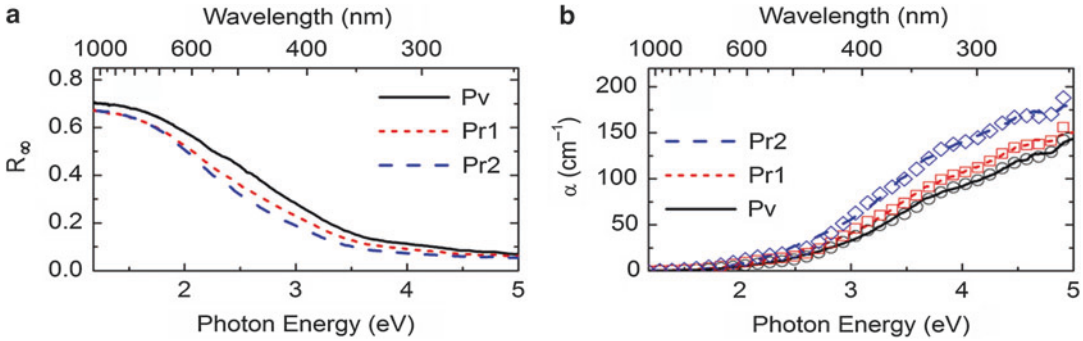
This approach has been applied to several ancient paper documents, some of which are of great importance for cultural heritage, and has also been applied to the famous Leonardo Da Vinci's self-portrait [14].

Applications

The uniqueness and fragility of ancient pieces of art on paper require a noninvasive and nondestructive diagnostic method such as optical spectroscopy. For this reason, diffuse reflectance measurements have been used to evaluate the degradation of several paper artifacts.



Spectroscopy of Ancient Documents, Fig. 3 Front and back of the Leonardo Da Vinci's self-portrait. *Red short dashed*, *circle blue long dashed*, and *circle black-solid circle* indicate the spots where reflectance measurements were carried out

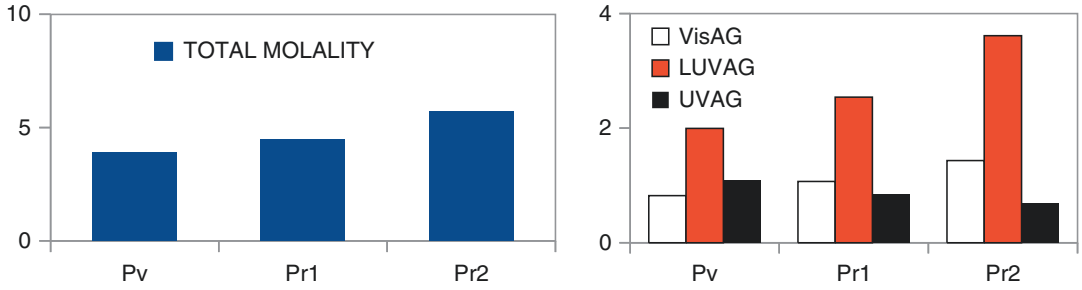


Spectroscopy of Ancient Documents, Fig. 4 Optical reflectance (panel **a**) of the Leonardo Da Vinci's self-portrait measured in the three spots shown in Fig. 3 and respective absorption spectra derived by using the Kubelka-Munk approach described above (panel **b**, lines). Theoretical absorption spectra of spots are also reported in symbols in panel **b**

For example, the diffuse absolute reflectance spectra of the Leonardo Da Vinci's self-portrait were measured in two spots on the front of the sheet (blue and red dashed circles in Fig. 3) and in one spot on its back (black continuous circle in Fig. 3).

The spectra of the artwork were acquired over a wavelength range from 250 to 1050 nm (4.96–1.18 eV of photon energy) (Fig. 4). From the reflectance spectra, the experimental intrinsic absorption coefficient of the constituent cellulose fibers of the paper of the Leonardo self-portrait was recovered by using the extension for strongly absorbing samples (see Eq. 8) of the YM and KM models. For each oxidized form of cellulose (including ketones, aldehydes, and diketones within the β-D-glucopyranose units), the optical spectra were obtained by TDDFT calculations [6, 14].

To analyze our results, chromophores were divided into three main categories according to their optical response: visible (VisAG), low-ultraviolet (LUVAG), and ultraviolet (UVAG) absorbing groups. The VisAG category contains the diketones. The LUVAG category contains aldehydes and single ketones. The latter are characterized by strong H-bonds with the hydroxyl groups. Although the chromophores of the LUVAG category mainly absorb in the low



Spectroscopy of Ancient Documents, Fig. 5 Total (*left* panel) and partial (*right* panel) concentrations of chromophores measured by spectroscopic approach in the three spots of Fig. 3

UV range, they also display a smooth tail extending into the short end of the Vis region (blue violet). Chromophores classified as VISAG and LUVAG are the ones that are the chief ones responsible for the yellowing of paper. Finally, the UVAG category contains single ketones free of hydrogen bond-type interactions which do not directly give rise to visible damage.

Comparison between experimental and theoretical spectra allowed the calculation of the total concentration of chromophores in the measured spots. Values were found to range between 3.3 and 6.1 mmol/100 g of cellulose, figures similar to those found in ancient samples of a comparable age.

The chromophore concentrations found in the front spots resulted to be about 20–80 % greater than that of the back spot (Fig. 5). This points to the front part of the work having been heavily exposed to external degradation agents. This observation implies that any further exposure of the self-portrait to such agents will increase the difference in optical degradation between the front and back of the sheet.

Comparing the concentration of UVAG, LUVAG, and VisAG chromophores to those found in modern papers which were artificially aged in different environmental conditions [14] highlights the important role played by humidity in the degradation of the Leonardo's self-portrait. In addition to the presence of moisture, degradation products within a sealed housing may have catalyzed further oxidative processes in addition to those which were already underway.

A periodic repetition of the same analysis would provide an ongoing quantitative assessment of its rate of degradation and would increase our understanding of the inevitable degradation processes that are underway.

Another example of application of optical spectroscopy of paper is the analysis of degradation phenomena in paper used as insulator of higher voltage winding in power transformers [11]. This kind of paper is made of pure cellulose (without lignin and fillers). The environmental conditions within power transformers are very harsh, with temperatures around 150 °C.

Therefore, the level of oxidation is one order of magnitude larger than that found in ancient paper samples aged during several centuries. The total concentration of chromophores is about 18 mmoles/100 g of cellulose.

The possibility to quantify the level of visual degradation of these paper materials by a nondestructive and noninvasive approach and observe its evolution in time represents invaluable information for conservators and restorers. Moreover, paper has several industrial applications; therefore, the field of application of this approach can be extended beyond cultural heritage.

Optical spectroscopy has a wide range of applications for cellulose-based materials, like paper, textiles, and other manufactured products of great industrial and cultural interest. It can also be extended to other strongly absorbing inhomogeneous materials.

References

1. Hunter, D.: Papermaking: The History and Technique of an Ancient Craft. Dover, New York (1978)

2. Krassig, H.A.: Cellulose: Structure, Accessibility, and Reactivity. Gordon and Breach Science, Singapore (1993)
3. Klemm, D., Philipp, B., Heinze, T., Heinze, U., Wagenknecht, W.: Comprehensive Cellulose Chemistry; Volume I: Fundamentals and Analytical Methods. Wiley VCH, Weinheim (1998)
4. Klemm, D., Heublein, B., Fink, H.-P., Bohn, A.: Cellulose: fascinating biopolymer and sustainable raw material. Angew. Chem. Int. Ed. **44**, 3358–3393 (2005). doi:10.1002/anie.200460587
5. Zervos, S.: Natural and Accelerated Ageing of Cellulose and Paper: A Literature Review in Cellulose: Structure and Properties, Derivatives and Industrial Uses edited by A. Lejeune, and T. Deprez, p. 155. Nova, New York (2010)
6. Mosca Conte, A., Pulci, O., Knapik, A., Del Sole, R., Lojewska, L., Missori, M.: Role of Cellulose Oxidation in the Yellowing of Ancient Paper. Phys. Rev. Lett. **108**, 158301 (2012)
7. Kerker, M.: The Scattering of Light and Other Electromagnetic Radiation. Academic, New York (1969)
8. Kortum, G.: Reflectance Spectroscopy (Principles, Methods, Applications). Springer, Berlin (1969)
9. Hubbe, M.A., Pawlak, J.J., Koukoulas, A.A.: Paper's appearance: a review. Bioresources **3**(2), 627–665 (2008)
10. Yang, L., Miklavcic, S.J.: Revised Kubelka–Munk theory. III. A general theory of light propagation in scattering and absorptive media. J. Opt. Soc. Am. A **22**, 1866 (2005)
11. Missori, M., Pulci, O., Teodonio, L., Violante, C., Kupchak, I., Bagniuk, J., Lojewska, J., Mosca Conte, A.: Optical response of strongly absorbing inhomogeneous materials: Application to paper degradation. Phys. Rev. B **89**, 054201 (2014)
12. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. Oxford University Press, New York (1989). ISBN 0-19-504279-4. ISBN 0-19-509276-7
13. Runge, E., Gross, E.K.U.: Density-functional theory for time-dependent systems. Phys. Rev. Lett. **52**(12), 997–1000 (1984). doi:10.1103/PhysRevLett.52.997. Bibcode:1984PhRvL..52..997R
14. Mosca Conte, A., Pulci, O., Misiti, M.C., Lojewska, J., Teodonio, L., Violante, C., Missori, M.: Visual degradation in Leonardo da Vinci's iconic self-portrait: A nanoscale study. Appl. Phys. Lett. **104**, 224101 (2014)

Spider Silk

Fritz Vollrath
Department of Zoology, University of Oxford, Oxford, UK

Synonyms

Silks of Spiders as model Bio-polymers

Definitions

Silks are animal fibers (or more rarely ribbons or sheets) of proteinageous biomaterials that are, by definition, extrusion spun [1]. While the evolutionary origins and taxonomic placement of silk feedstocks can differ widely across the arthropods, filaments can be surprisingly similar [2]. *Capture silks* are sticky materials depolying either nanoscale filaments or aqueous glycoprotein glues that have evolved from dry silks [3].

Outline

Silks are fascinating biological products and have evolved several times independently in the arthropods. Spiders and moths are the best-known and best-studied of silk spinners, but there are others ranging from mites to bees [2]. In each taxon the diversity of silks has evolved in only one ancestor but then radiated quickly (over millions of years) into many different types fit for the various purposes required by the animal – be it integration into a cocoon composite or use as a single safety line. Spiders are unique among silk spinners in that an individual has a veritable armory of silk glands with each silk being tailored to a specific use. Accordingly the silks of spiders show a surprising diversity ranging from tough dry threads to soft wet fibers and from sticky aqueous glue to adhesive nanoscale filaments [4]. This very diversity provides a window into the interaction of a silk's molecular structure with the animal's ecology and thus delivers deep insights into the material's function-structure relationship on the macroscopic, organismic scale. In addition, and of relevance to the subject of nanotechnology, such diversity also offers examples for function-structure relationships on the micro- and nanoscale, and these are the dimensional domains explored here.

However, in order to better understand the scope of the paradigm, and to better appreciate the range of possible solutions, it may be helpful to first briefly outline the biology of the system, and in this way examine the constraints as well as demands imposed on the material "silk". This

more generic introduction will be followed by specific lessons learned so far from the study of spider silks and from comparisons to the commercially much more important mulberry and wild insect silks. The essay concludes with the briefest of outlooks into the future of silk as a model material for academic investigations into nanoscale processes.

Key Principles

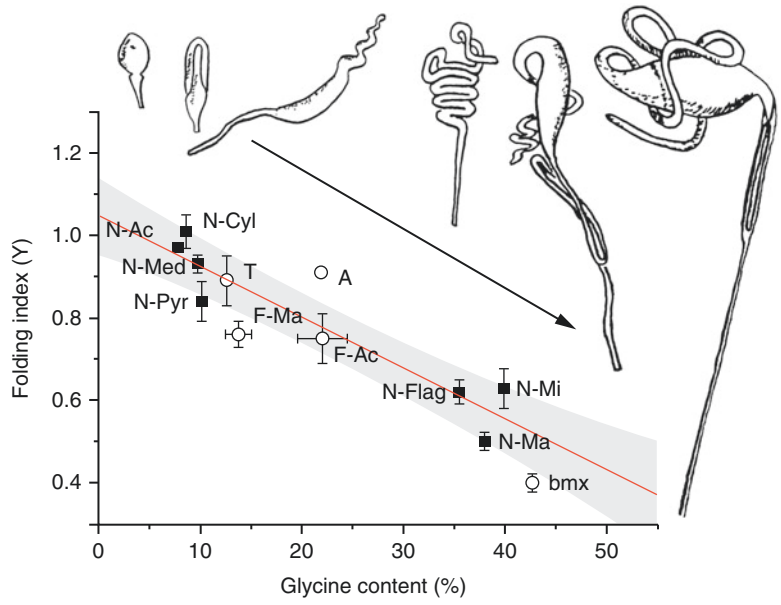
Importantly, silks have evolved to function in the dead state, i.e., detached from the animal and typically after having been dehydrated and denatured [1]. This means that silks provide a model in which we can study – under realistic functional conditions and in great detail – the thermomechanical properties of a biological material. Indeed, there are very few (if any) other biological materials that offer comparable access to in-depth investigations under fully natural conditions. This feature of "natural function even in the fully detached state" provides detailed as well as biologically realistic data about a given silk's material properties. By applying general biological knowledge one can speculate about property–-function relationship (▶ Plasticity Theory at Small Scales). But, in order to understand a silk's structure and its potential relationships with the material's properties, one requires more than biological insight, i.e., structural data are needed. Such data would come from fine-grained structural analysis such as NMR, X-ray, and neutron-scattering as well as Raman, IR, and CD/LD spectroscopy as well as data on gene and molecular sequences and insights into biological constraints such as the overbearing role of water in biological systems and the evolutionary history of the organism and the silks under investigation. Not surprisingly, while a good understanding of a specific silk's material properties is emerging, the links to function and structure are still rather weak [5].

Accordingly it might be more realistic, at this stage, to view the study of silks as research aiming to link *properties* to *function* and to *structure* – rather than taking the more common approach of exploring a "straight" structure-function relationship. This view notwithstanding, with modern molecular techniques (ranging from sequencing to scattering and modeling) the protein and bond interactions that underlie the functional properties are rapidly being uncovered, and a composite picture is beginning to emerge with clear-cut hypotheses that are ready for rigorous testing [6].

Why Study Spider Silks?

The threads of the commercial mulberry silkworm *Bombyx mori* are the fibers spun by the larva of a lepidopteran moth as part of its cocoon composite constructed to shelter the pupa during its metamorphosis. These fibers have been the mainstay of a key textile industry for over 6000 years and as such have seen thousands of years of ever more sophisticated R&D. However, as the worm spins from converted salivary glands through the mouth, it can cut the thread, which (as we will see) poses certain problems for studying the extrusion process itself. Moreover, lepidopteran silks have evolved for integration into multilayered cocoon complexes and thus have never been optimized for single-fiber strength or toughness but for the specific qualities required of threads in a composite.

Spider silks, on the other hand, have evolved for a much wider range of applications (Fig. 1). The best-studied of these silks, the dragline fibers of the *Major Ampulate Glands*, are optimized for effective and efficient deployment as single-thread safety lines and for integration as key structural members in a web's open-mesh network. In both applications, these threads' primary function is always the mitigation of kinetic energy, be it for body support or prey impact and restraint [8]. Many spider silks, therefore, are excellent examples for naturally evolved protein (bio)materials optimized for toughness. Moreover, spiders spin from modified, abdominal leg glands through specialist spinnerets, far away from any cutting mouthpart. Hence spiders can be "silked" easily and under controlled experimental conditions [9].



Spider Silk, Fig. 1 Relationship between the Glycine content of a silk and the "folding index" γ of its proteins in the prespun liquid stage. γ is taken to be indicative of the protein intrinsic disorder. The correlation and model are taken to quantitatively explain the structure-function relationship by describing the molecular conformation i.e., β-sheet propensity. The data suggest that, in order to achieve specialization and performance, silks require higher structural flexibility at the expense of reduced stability and increased conversion energy. A γ value near 1 would denote helix-type folding, while γ values <0.5 would signify mostly unfolded chains having been calculated from the ratio of the circular dichroism spectrum bands at 208 and 220 nm (at 20 °C). The *arrow* shows the direction of gland evolution and the insets depict schematically the overall gland shape (not to scale). *Nephila edulis* (golden orb spider) major ampullate (N-Ma), minor ampullate (N-Mi), flagelliform (NFlag), cylindriform (N-Cyl), aciniform (N-Ac), pyriform (N-Pyr), median (N-Med); *Kukulkania hibernalis* (Filistatidae) major ampullate (F-Ma) and acinous (F-Ac). *Antrodiaetus unicolor* (Antrodiaetidae) single type glands (A), and *Aphonopelma chalcoldes* (Theraphosidae) acinous (T), as well as the commercial mulberry silk from *Bombyx mori* (Insecta: Bombicidae) (bmx – silkworm silk) (For details see Ref. [7])

Such experimental access allows us to explore the contribution of environmental key parameters (such as temperature, pH, and flow characteristics during extrusion spinning) on the conformation and interactions of supramolecular structures using (even online as the animal spins) high-resolution analytical tools such as X-ray scattering or Raman spectroscopy [9]. Hence much of what is known to date about the molecular structure-function relations in the biological elastomer "silk" originated from studies of spider silks rather than insect silks. Not surprisingly, all relevant studies showed that the key determinant for the silk's outstanding mechanical properties seems to be its notable structure, which, in turn, relays on the scaling of its semicrystalline morphology [10].

The Importance of Extrusion Spinning

Importantly, the morphology of a silk is the outcome of extrusion processing, not of growth processes. Extrusion "spinning" relies on the refolding of molecules under physical forces such as flow elongation and shear in a tubular duct typically combined with shifts in the chemical environment such as salts and pH changes [11]. Chaperoning molecules might provide additional guidance. Whatever the details of the individual extrusion processing of specific silks, as the fiber is drawn, the solvent water is spontaneously ejected in a self-denaturing solution-gel-solid conversion with the molecules refolding and rearranging themselves from one conformation into another. While the molecular

conformation in the feedstock is adapted to safe storage (sometimes for weeks) in an aqueous solution, the final conformation in the thread is adapted to stability (sometimes for years) against chemical, physical, and biological agents of the environment. The transition between the two conformations typically happens in milliseconds and appears to require very little energy [11]. At present it is known *where* it happens (in the duct), it is known roughly *what* happens (molecular alignment in the flow accompanied by unfolding and hydrogen bond "cross-linking" within and between molecules), but it is not known exactly *how* it happens [12, 13].

One may assume that a silk, be it in the liquid or solid state, adopts shapes and conformations that are dictated by the interactions between polar and nonpolar moieties [13]. A fiber's macroscale properties are determined by bulk orientations with optimal axial stiffness on the nanoscale. To achieve this, the molecules must be allowed (during the transition from solution to solid) to self-organize into their extended configurations and be accompanied by an efficient intermolecular lock-in. However, in order to maintain flow viscosity, the number of cross-links must be limited. α-helical structures are stable as isolated secondary structures (while β-structures are not) and tend to engage and stabilize interactions with neighboring strands thus forming an intermolecular gelation network. Control of the silk molecules seems to be achieved by a number of reactions. For example, calcium ions stabilize silk proteins at high concentrations but destabilize (and apparently promote β-sheet structures) at low concentrations. Sodium ions also stabilize the silk proteins (important for storage) and are exchanged during the extrusion process by β-sheet promoting potassium ions. Interestingly, similar conditions will produce the formation of amyloids in globular and other fibrous proteins [13].

These (and other) processes during extrusion "spinning" emphasize the importance of a molecular structure and conformation that, while still in the prespun phase, has its future conformations as solid fiber embedded in the amino acid sequence and their accompanying hydrogen bonds [13]. Importantly for the commercialization of bio-inspired ideas, extrusion spinning is a process that is well established in industrial applications and commercial polymer production, while highly controlled biological growth processes are still outside industrial exploitation (other than via microbial intermediaries). Polymer theory provides a possible tool to explore the various nanoscale structures that self-assemble in the silks during spinning [1]. Ideally, the lessons thus learned can then be used by practitioners to design and manufacture new polymers using novel (and hopefully more sustainable) production methods (▶ Bioinspired Synthesis of Nanomaterials; ▶ Self-Assembly; ▶ Sol–Gel Based Nanostructures). After all, silks (and spider silks especially) have many properties that are not only highly eco-friendly but also economically interesting, ranging from full biocompatibility to exceptional mechanical properties.

Mechanical Properties

Spider silks show an enormous diversity of morphologies and mechanical properties, which is not surprising given the great diversity and age of the material [8]. Moreover, spider silks have adapted to a wide range of purposes ranging from "wall papering" burrows to aerial nets, and range in fiber diameters (which is a key parameter of actual strength) from about 10 μm down to a few tens of nanometers [4]. As a rough guide one could take spider silk fibers to have engineering properties (i.e., mechanical properties compensated for fiber dimension) with moduli in the range 1 kPa for a highly hydrated "gel" to about 20 GPa for the stiffest dragline silk, and strengths from almost zero values of yield stress to about 1.6 GPa respectively. In comparison, a decent commercial silk fiber (as used in textiles) would have a modulus and strength of about 10 GPa and 400 MPa respectively [1]. The combination of excellent strength and high strain tolerance gives spider silks their exceptional toughness.

Importantly for the polymer scientist, silks offer a wide range of combinations of stiffness and toughness that might provide, if the

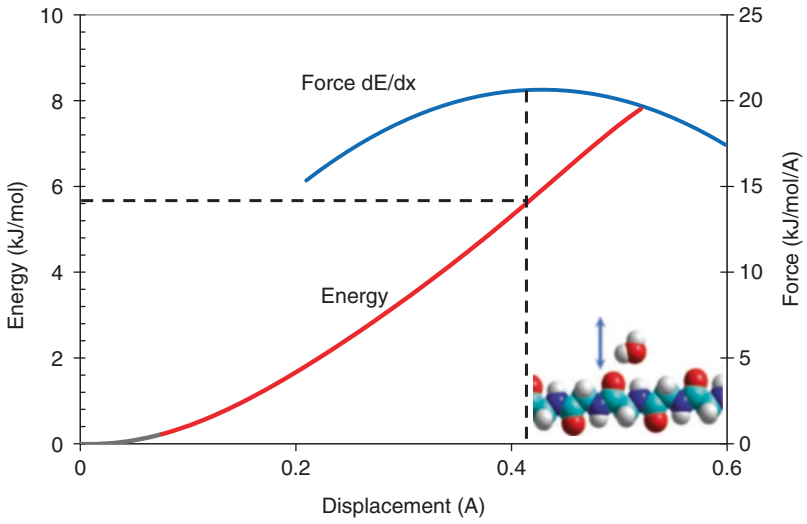
underlying principles are understood, attractive opportunities for the bio-inspired design of novel materials. For example, a dragline of *Nephila clavipes* with a fiber diameter of about 6 μm has a breaking strength of about 1.6 GPa at strains of around 30 %. The estimated yield stress (ca 0.2–0.3 GPa for a 10 GPa modulus) suggests that the spider deploys postyield plastic flow and strain hardening to dampen its movement in a sophisticated combination of viscoelastic properties including the interesting feature of shape-memory [1]. Models and simulations allow the exploration of the molecular processes underlying the stress–strain curves that, for spider silks, can easily be obtained under a wide range of experimental conditions, as outlined earlier.

Modeling Approaches

Surprising fits between observation and interpretation can be achieved by suitable models and simulations that set out to predict ab initio the energy interactions between properties and structure in silk feedstocks and fibers (▶ Ab Initio DFT Simulations of Nanostructures). For example, quantum mechanics can be used to predict elastic instability energy conditions for hydrogen-bonded interactions between water and the silk proteins. Such energy conditions can then be transformed into parameters such as temperature and stress by using clearly defined relations for zero-point and thermal energies that contribute to the denaturation and glass transition temperatures. A realistic model (e.g., 14) for an energy instability criterion for protein denaturation would have four key stages: (i) calculate the potential energy of water-amide hydrogen bonding as a function of bond length, (ii) calculate the energy at which water becomes free to move away from the amide segments, (iii) calculate the kinetic/entropic energy countering the bond potential energy, and (iv) calculate the probability of water-amide bond dissociation for denaturation as a function of time, temperature, and applied stress.

A specific simulation of stress–strain processes in a spider's dragline silk might serve to demonstrate here how the general mechanical properties of a dragline silk can be related to internal structure [13]. Molecular dynamics can be used to develop a model that is able to match observed stress–strain profiles with hypothetical "snapshots" of the deforming structures [1]. One insightful model leads to the conclusion that the extraordinary toughness of spider draglines derives from an energy equilibration, as does the postyield strain hardening, which in turn is critical for the strength-toughness balance [14]. This model takes a poly(glycine) "string-of-beads" structure with periodic boundary conditions in an energy-minimized structure to have a density of 1.3 g/cc at 300 K. The purely elastic reference modulus B_e would be given by the cohesive binding energy density with the molecular interactions derived from the Lennard-Jones potential function providing a proportionality constant of 18 for the molecular interactions. Consequently, a cohesive energy of about 40 kJ/mol and a volume of 50 cc/mol for a generic silk peptide segment would give $B_e \approx 14$ GPa. The energy dissipated through the broad secondary relaxation due to hydrocarbon segments would result in an isotropic tensile modulus of about 9 GPa at a Poisson's ratio of about 0.4. The evaluation of such molecular dynamics simulations (see Fig. 2) suggest that the gradual postyield regeneration of hydrogen bonding under strain can absorb large amounts of energy of deformation, which would increase the elastic modulus to its low-strain initial value [16]. There are, of course, also other approaches to molecular modeling, and the field is advancing rapidly.

Of paramount importance for any modeling approach aiming to relate predictions of silk properties to structural features is the reduction of complex protein/polypeptide structures into "ordered" and "disordered" fractions [17]. The ordered domains are more rigid and deform significantly less than the disordered domains, which (like in any amorphous, viscoelastic polymer) would be responsible mainly for energy dissipation, i.e., toughness. Experiments have shown that silk properties are fine-tuned by water acting (as plasticizer) on the disordered fraction where



Spider Silk, Fig. 2 Energy and force as a function of separation distance for a silk water-amide bond, with the elastic instability condition marked by the horizontal *dashed line*, predicted for the structure shown as an insert using dft simulations (For details see Refs. [13–15])

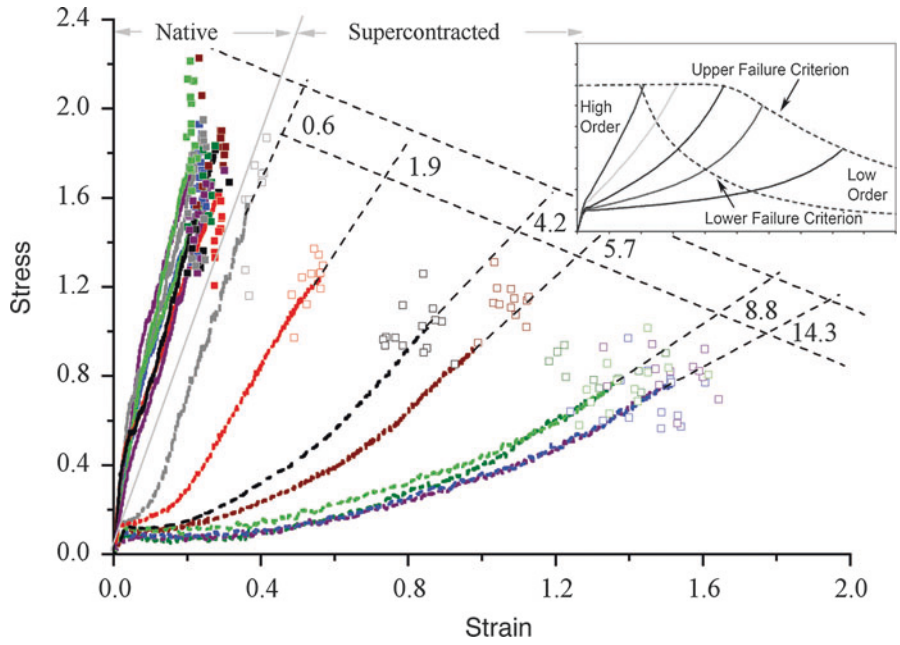
it binds specifically to polar groups and thus reduces the glass transition temperature. As polar amide-amide bonding determines the glass transition temperature, T_g, (of ca 200 °C in dry silk), increasing hydration of a silk will reduce not only T_g but also modulus and yield stress [10].

Clearly, as in other organic and inorganic materials, the nanoscale order–disorder organization coupled with degree of hydration are the two key features that define property-structure relationships. An experimental window into these relationships is provided by the comparison of natural silks and silk-like materials prepared from silks that have been dissolved in chaotropic agents to be respun using dehydrating agents. These materials (often called Reconstituted Silk Fibroin or RSF) share molecular sequence but not, apparently, molecular structure/conformation with native silks. Nor do they share mechanical properties, which opens the window for studying the details of property-structure relationships experimentally with the aim of testing specific hypotheses [17].

Functional Properties

Energy is the currency of living organisms. And water is the primary commodity for life and thus the key to understanding biological interactions. Silk molecules are formed in a fully hydrated state and stored in an aqueous solution. Extraction of water obliges them to go through a sol–gel transition, become immobile, and form the fiber. The rate and ratio of dehydration tunes the mechanical properties of the material via the nature of the hydrogen bonds and their intra- as well as intermolecular interactions.

There is no question that spider silks and their properties can be defined by their sensitivity to water hydration and, indeed, even more powerful chaotropic agents. Some silks swell and shrink when exposed to high humidity or submersion in water while others show no response at all. It seems that the proportion of the amino acid proline is a major contributor to such high water sensitivity – also called supercontraction – reflecting the silk's observable behavior [18]. Proline, given its shape, would be a side group that brings "disorder" to the folding of the molecular chain. Its presence would thus affect a silk's material properties by skewing the key ratio of order and disorder (Fig. 3). Interestingly, high-proline content silks not only take up water readily but also become rather "stiff" when redried. Moreover, if such a wetted and redried silk is stretched and restretched repeatedly (i.e., cyclically loaded) it reverts to its original stress–strain behavior. This suggests that there is a reversible change in the order–disorder



Spider Silk, Fig. 3 Representative stress–strain curves of silks with different percentages of proline relating to different nanoscale structural arrangements. The silks were all major ampulate gland MAA *drag-line* silks forcibly reeled from a range of species. The reeling condition had been tuned to produce silk samples with a native breaking strain of 25 % and the silks were tested either in the native (*solid lines*/*squares*) or the supercontracted state (*dashed lines*/*open squares*). The proline content and ordered fraction of supercontracted MAA silks are marked. The *straight black* line defines the upper limit of native MAA silk, where the ordered fraction is 1.0; the *straight gray* line roughly separates the stress–strain curves of native and supercontracted silks (for details see Ref. [1]). The inset shows the calculated mechanical properties of threads with different relationships of order/disorder (For details see Ref. [18])

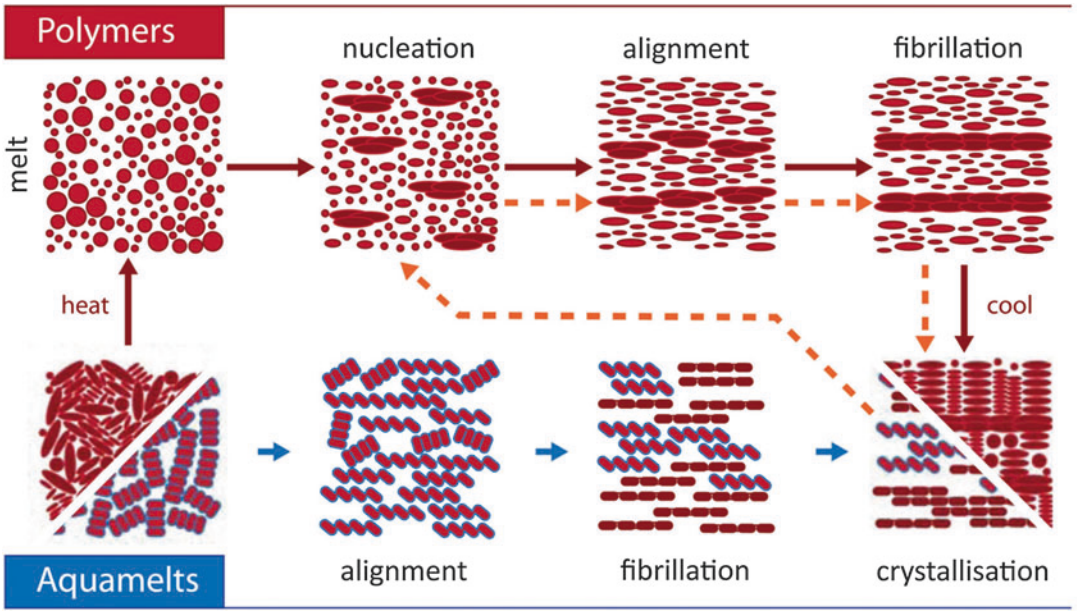
ratio even in the finished fiber, which reflects the level of hydration. It could be argued that the hydrogen bonds of the molecular chain segments around the proline segment are being forced apart under hydration only to be fixed into new positions during dehydration. Rehydration would allow water molecules to reaccess the protein opened up at the proline site and assist in the formation of new hydrogen bonds. This would relax the mechanical strain "frozen" into the material and open the fiber for ingress of water and the exhibition of supercontraction with its accompanying changes in mechanical properties.

Clearly, water in its interaction with the different motifs in the molecular chain of the silk protein can control a silk's material properties – presumably via the hydrogen bonds at occupied and potential bond sites.

The Importance of Spinning

Generically speaking, the "typical" spider silk consists of two proteins, one rather large (in hundreds of kD) and one significantly smaller. There are notable exceptions, such as the silks of bees, which have three or four smallish proteins (in tens of kD) or the silks of some midges, which seem to consist of hundreds of even smaller proteins [2]. Together, these and some of the unusual spider and lepidopteran silks should provide important test grounds for hypotheses (derived from the "standard" spider silks outlined earlier in this essay) aiming to probe the relationships between structure, function, and properties of silks in general.

The principal hypothesis of nanoscale order–disorder as key to silk properties centers around the importance of a number of processes

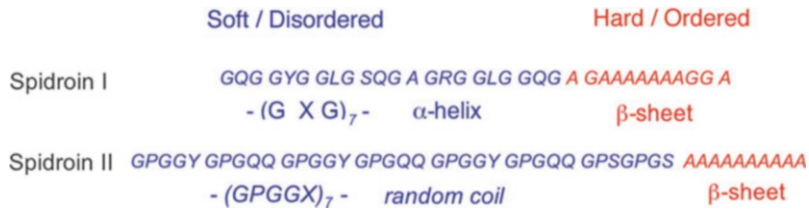


Spider Silk, Fig. 4 Fibre formation in polymers (*red arrows*) and their proposed subclass aquamelts (*blue arrows*). Horizontal arrows depict shearing steps with length corresponding to the relative amount of work required for fibrillation. Colored objects represent components responsible for fibrillation: long chain molecules (their radius of gyration or shape) in polymers and proteins (secondary and tertiary structures) in aquamelts with differences in shade corresponding to phase differences. The *blue* outline of aquamelts represents the outer shell of hydration of the protein. *Orange* arrows represents the path reconstituted silk follows in order to be reprocessed into a fiber (For details see Ref. [19])

during the formation of the fiber. Indeed, the observable nanoscale structures and interactions appear highly evolved and, by conjecture and experiment, seem to confer on silks the impressive properties that, from the human standpoint, are highly desirable (Fig. 4). Importantly, the molecular interactions have a fixed component (i.e., genetically transferred) as well as a flexible component (i.e., environmentally responsive). Thus, to use a human analogy, a spun silk fiber embodies both nature (the genetic blueprint) and nurture (the milieu during maturation).

Structure (having both a fixed and a flexible component) provides the opportunity to experimentally probe the interactions between structure, properties, and function on the molecular and supramolecular, i.e., nanoscale, level. Firstly, sequencing the silk genes exposes the genetic blueprint and gives the exact positioning of the amino acids that make up the side chains of all silk-polymer macromolecules. Secondly, extracted liquid silk precursor (feedstock or dope) from the gland of the animal allows examination of molecular conformations in that prespinning state. Thirdly, molecular conformation in the spun fiber state can be studied in the fiber. Importantly, it is also possible to examine the molecular transformations from dope to fiber, although these kinds of study are much more difficult than the others given the small dimensions and quick timescales involved [20, 21].

But it is during extrusion where the protein molecules self-assemble into the complex containing the correct intra- and intermolecular assemblies [11]. Clearly, parameters associated with flow through very fine ducts are important, and not surprisingly given this constraint, the block copolymer molecules of the silk proteins display characteristics of liquid crystals. However, it is still debated whether they really are true liquid crystals or semicrystalline complexes behaving as such [1]. However this may be,



Spider Silk, Fig. 5 Schematic of the two principal proteins identified in a benchmark spider dragline silk outlining the link between sequence and properties. Amino acid codes are *A* alanine, *G* glycine, *P* proline, *Y* thyrosine, *L* leucine, *Q* glutamine

already during the synthesis of the silk protein the molecular complex takes on the distinct morphology of a tightly folded nanoscale rod. Minute changes in pH imposed during the travel down the duct, in combination with mechanical forces, are exerted by the accelerating flow and elongate the fiber as it moves through the hyperbolic section of the extruder [11]. Water is pressed out as the proteins align and interact. Further down in the duct, which is now tubular, additional mechanisms (about which little is known) impose further interactions that give the thread its final shape, inside and out, as well as determining the mechanical properties, which are of course tightly correlated with internal and external structure.

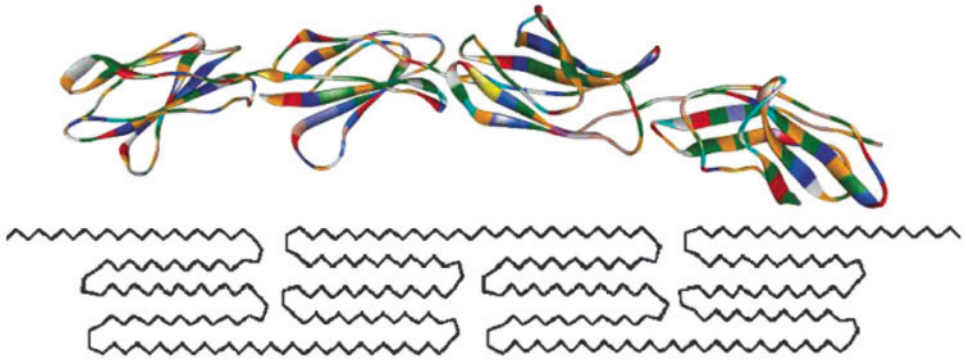
Molecular Sequences

Silks are mostly made up of proteins and contain a range of amino acid combinations. The peptide segments/motifs have a core configuration of -NH-CO-CHR- where R represents 1 of 20 different side groups, which may exhibit highly divergent chemical functionalities (Fig. 5). Spider silk proteins typically have very high fractions of the amino acid glycine, which has the simplest (smallest) of all side groups (G, R = H) and which, because of lack of steric hindrance, allows a wide range of folding conformations. The next important, and still rather simple, peptide groups are alanine (A, R = CH_3) and serine (S, R = CH_2OH) introducing hydrophobicity and polarity, respectively. Finally, there is proline (P, R = CH_2 CH_2 CH_2), which with its large cyclic and rather rigid side group intrinsically disorders the chains by twisting them out of their usual highly regular conformer torsional angles.

Importantly, it appears that in spiders the more "advanced" silks seem to have higher ratios of glycine coupled with rather complex spinning glands (Fig. 1). This suggests that molecular flexibility in a silk feedstock requires more control during extrusion; however, the inverse is also possible, i.e., that increased complexity during processing has led to increased molecular flexibility [13].

Proline is another key component of silk nanostructure by forcing the molecular chain that contains it in regular spacings into the string-of-beads morphology mentioned earlier (Fig. 6). This conformation (as yet hypothetical but based on strong evidence) consists of "beads" of beta folds, typically 10 units long and with a chain axis length of about 4 nm each [10]. Such a "string-of-beads" would facilitate processing in the flow field of the narrow extrusion duct [1] by allowing reorientation but not entanglement, thus conferring liquid-crystalline non-Newtonian flow behavior [11].

Clearly, analysis of the primary structure already allows for the formulation of testable hypotheses and related interpretations of the secondary structures. And these are the structures that form most silks' semicrystalline, nanometer-scale morphology of "hard", highly ordered domains in a "soft" matrix of more disordered polymer chains (Fig. 3). Degree of hydration is the key as the secondary conformations are held in place by hydrogen bonds connecting the amide groups of the main-chain backbone either directly (hard) or via a water molecule (soft).



Spider Silk, Fig. 6 Silk's nanoscale "string of beads" hairpin folding morphology induced by the proline side chains (For details see Refs. [7, 22])

It must be assumed, although the final verdict is still out, that strong spider dragline fibers have a hybrid structure of about 50/50 hard/ordered to soft/disorderd domains whereby the hard beta sheet folded domains are embedded in a matrix of chains in various degrees of helical conformations thus acting like an amorphous matrix. Indeed, perhaps it is wise to hedge about the minute details of structural function-property relationships until better data has come in from fine-grained analytical spectroscopic techniques such X-ray and Neutron Scattering, Nuclear Magnetic Resonance Fourier Transform InfraRed and Ultra-Violet Spectroscopy, as well as Dynamic Mechanical Thermal Analysis, Differential Scanning Calorimetry, and Thermogravimetric Analysis.

Conclusion

More than 400 million years of evolution will have seen to it that the hierarchical morphology in the patterning and mesophase assembly of spider silk molecular motifs is spatially optimized for each of the many specific functions required for survival of the spider. It has been argued that the evolutionary optimization of energy resources provides a perfect framework for multiscale models probing the intrinsic property-structure relationships of silks. In order to investigate the distribution and exchange of energy at the nanoscale, such a model would deploy the control of energy storage (strength) and dissipation (toughness) at the molecular level. If correct, then the model should predict the full stress–strain profile of silks to failure, covering the full range right from the strongest dragline threads to the most compliant capture threads in the web. Importantly, it appears that this can be done [10].

The model explaining key aspects of the behavior of finished fibers can be extended to integrate key aspects of the formation of the fiber from the dope [9]. Experimental data demonstrate that "live" (i.e., native) silk dope differs significantly from "dead" (i.e., spun denatured then reconstituted) silk [17]. Studies that theoretically examine this conversion from unspun to spun *native* silk seem to indicate that a silk, once spun, cannot be "unspun" and spun again [20]. This insight, if correct, will have important implications for the goal of producing spinnable silk dopes either by reconstitution from fibers [6] or by protein extraction from microbial expression systems [21]. According to present insights [13], both ways of producing semi-native dope would fatally deconfigure the natural molecular structures; yet these (if the models are correct) would be key requirements for the natural spinning process to function. And that, after all, is one of the most amazing traits of silks and their energetically so efficient and clean production: it uses only water as solvent at ambient pressure and temperature yet produces a biopolymer fiber that can

hold its own proudly in comparison with mechanical properties of man-made polymer fibers.

As outlined in this essay, spiders provide us with a generic class of natural materials, silks, which lend themselves as models for a wide range of other biological elastomers. Unlike these, which naturally perform in the hydrated state, silks have evolved to operate in a wide range of states of hydration, from dry webs and cocoons through to air sacs for underwater spiders, and with the subtle control of tightening sagging webs by supercontraction with dew condensation. Understanding the interaction of biological elastomers with water is a key requirement if we are to produce synthetic biomimetic analogues. Biological functionality, after all, relies on wet engineering coupled with nanoscale dimensions often integrated into structural "hierarchies" of differing degrees of order.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Plasticity Theory at Small Scales
- ▶ Self-Assembly
- ▶ Sol–Gel Method

References

1. Vollrath, F., Porter, D.: Silks as ancient models for modern polymers. Polymer **50**, 5623–5632 (2009)
2. Sutherland, T.D., Young, J., Weisman, S., Hayashi, C. Y., Merrit, D.: Insect silk: one name, many materials. Annu. Rev. Entomol. **55**, 171–188 (2010)
3. Brunetta, L., Craig, C.: Spider Silk: Evolution and 400 Million Years of Spinning, Waiting, Snagging, and Mating, pp. 1–229. Yale University Press, New Haven/London (2010)
4. Vollrath, F.: Spider webs and silks. Sci. Am. **266**(3), 46–52 (1992)
5. Fu, C., Shao, Z., Vollrath, F.: Animal silks: their structures, properties and artificial production. Chem. Commun. **43**, 6515–6529 (2009)
6. Omenetto, F., Kaplan, D.L.: New opportunities for an ancient material. Science **329**, 528–531 (2010)
7. Porter, D., Vollrath, F.: Silk as a biomimetic ideal for structural polymers. Adv. Mater. **21**, 487–492 (2009)
8. Harmer, A.M.T., Blackledge, T.A., Madin, J.S., Herberstein, M.E.: High-performance spider webs: integrating biomechanics, ecology and behaviour. J. R. Soc. Interface **8**, 457–471 (2011)
9. Vollrath, F., Porter, D., Dicko, C.: The structure of silk. In: Eichhorn, S.J., Hearlem, J.W.S., Jaffe, M., Kikutani, T. (eds.) Handbook of Textile Fibre Structure, vol. 2, pp. 146–198. Woodhead Publishing, Oxford/Cambridge, MA/New Delhi (2009)
10. Vollrath, F., Knight, D.P.: Liquid crystal silk spinning in nature. Nature **410**, 541–548 (2001)
11. Aldo Leal-Egaña, A., Scheibel, T.: Silk-based materials for biomedical applications. Biotechnol. Appl. Biochem. **55**, 155–167 (2010)
12. Dicko, C., Porter, D., Vollrath, F.: Silk: relevance to amyloids. In: Riggaci, S., Bucciantini, M. (eds.) Functional Amyloid Aggregation, pp. 51–70. Research SignPost, Trivandrum (2010)
13. Porter, D., Vollrath, F.: The role of kinetics of water and amide bonding in protein stability. Soft. Matter. **4**, 328–336 (2008)
14. Holland, C., Vollrath, F.: Biomimetic principles of spider silk for high-performance fibres. Chapter 7. In: Ellison, M.S., Abbott, A.G. (eds.) Biologically Inspired Textiles. Woodhead Publishing, Cambridge, MA (2008)
15. Liu, Y., Sponner, A., Porter, D., Vollrath, F.: Proline and processing of spider silks. Biomacromolecules **9**, 116–121 (2008)
16. Porter, D.: Group Interaction Modelling of Polymers. Marcel Dekker, New York (1995)
17. Holland, C., Terry, E.A., Porter, D., Vollrath, F.: Rheological characterisation of native spider and silkworm dope. Nat. Mater. **5**, 870–874 (2006)
18. Spiess, K., Lammel, A., Scheibel, T.: Recombinant spider silk proteins for applications in biomaterials. Macromol. Biosci. **10**, 998–1007 (2010)
19. Holland, C., Vollrath, F., Ryan, A.J., Mykhaylyk, O.O.: Silk and synthetic polymers; reconciling 100 degrees of separation. Adv. Mater. **24**, 105–109 (2012)
20. Porter, D., Vollrath, F.: Water mediated proton hopping empowers proteins. Soft. Matter. **9**, 643–646 (2013)
21. Porter, D., Vollrath, F.: Water mobility, denaturation and the glass transition in proteins. Biochim. Biophys. Acta (BBA) Proteins Proteomics **1824**, 785–791 (2012)
22. Vollrath, F., Porter, D.: Spider silk as archetypal protein elastomer. Soft. Matter. **2**(5), 377–385 (2006)

Spiders

- ▶ Arthropod Strain Sensors

Spintronic Devices

- ▶ Magnetic Nanostructures and Spintronics

Spontaneous Polarization

▶ Polarization-Induced Effects in Heterostructures

Spray Technologies Inspired by Bombardier Beetle

Alexander Booth[1], Andy C. McIntosh[1], Novid Beheshti[2], Richard Walker[3], Lars Uno Larsson[4] and Andrew Copestake[5]
[1]Energy and Resources Research Institute, University of Leeds, Leeds, West Yorkshire, UK
[2]Swedish Biomimetics 3000® Ltd, iBIC – Birmingham Science Park Aston, Birmingham, UK
[3]ICON plc, Marlow, Buckinghamshire, UK
[4]Swedish Biomimetics 3000® AB, Stockholm, Sweden
[5]Swedish Biomimetics 3000® Ltd, University of Southampton Science Park, Southampton, UK

Synonyms

Flash evaporation liquid atomization; Liquid atomization through vapor explosion; µMist®

Definition

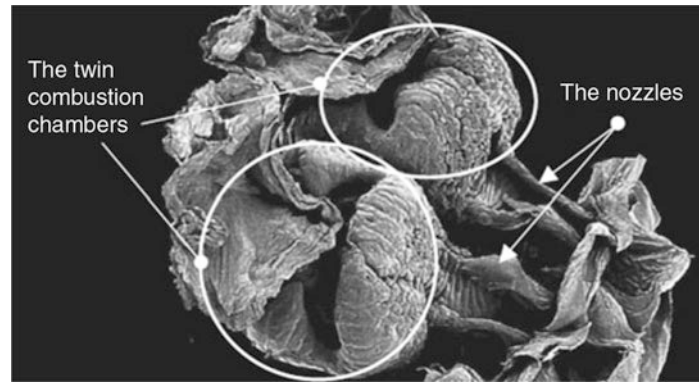
µMist® is a liquid spray and atomization system inspired by a method that is used as a defense mechanism by several types of bombardier beetle. The liquid atomization, spray formation, and its propulsion/emanation are achieved by causing a liquid in a sealed chamber to flash evaporate on the release of an exhaust valve. This flash evaporation in the confined space of the chamber causes a vapor explosion, the force of which overcomes the surface tension of the liquid, causing it to break down into small droplets and exit the valve with a considerable momentum as a spray.



Spray Technologies Inspired by Bombardier Beetle, Fig. 1 African bombardier beetle

The µMist® System Development

Initial investigations were inspired by the entomological research of Professor Thomas Eisner in to the bombardier beetle's defensive spray system. A typical example of a bombardier beetle from Africa can be seen in Fig. 1. Computer simulations of the spray system, using computational fluid dynamics (CFD), then led to the understanding of the physical principles governing the beetle's unique spray facility, and the µMist® system is based upon these principles. When threatened, the bombardier beetle produces a hot liquid spray from its abdomen, which is ejected at its attacker. Eisner's work focused on how this spray was generated by the beetle. He found that at the rear of the beetle's abdomen was a complex series of glands and chambers in which an exothermic chemical reaction occurred [1]. This chemical reaction heated its constituent liquid to a very high temperature in a reaction chamber, linked directly to an exhaust orifice sealed by a biological pressure-triggered exhaust valve. When the liquid in the reaction chamber reached a certain pressure, this exhaust valve opened, causing the liquid in the reaction chamber to be ejected as a spray.



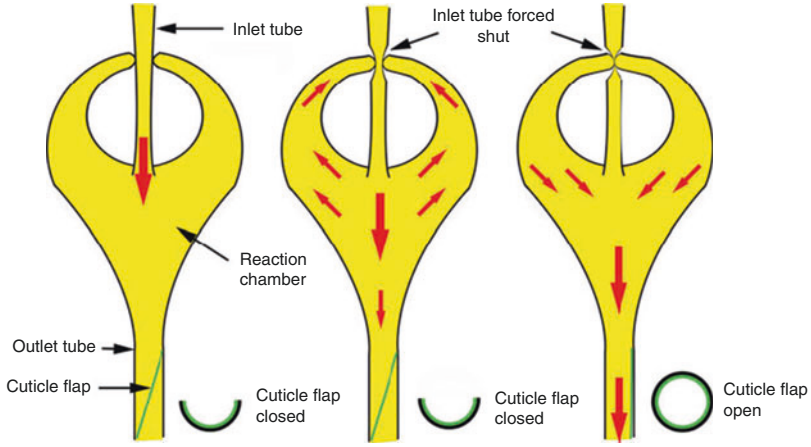
Spray Technologies Inspired by Bombardier Beetle, Fig. 2 An electron micrograph of the bombardier beetle's spray mechanism (Courtesy of Prof. T. Eisner, Cornell University)

Eisner's work inspired McIntosh and Beheshti [2–5], at Leeds University, to investigate the physical process in the bombardier beetles which generated this spray [2, 3] and found that it is a major advance on established liquid atomization methods. This initial work was funded by the Engineering and Physical Sciences Research Council (EPSRC) and also by Swedish Biomimetics 3000® Ltd. All further work mentioned in this article and all continuing work on the μMist® system has been and is currently being supported and advanced by Swedish Biomimetics 3000® Ltd under its V^2IO innovation acceleration model. Additional funding has been given by partners for particular projects, including Carbon Connections and the NIHR i4i program.

The initial investigation conducted by McIntosh and Beheshti consisted of a series of Computational Fluid Dynamics (CFD) simulations to better determine the thermodynamic and fluid dynamic processes which were occurring in the chamber of the bombardier beetle. These simulations did not include the associated chemical reaction seen in the bombardier beetle, as in this context it simply functions as a heat source to the chamber. The results of this study showed very good correlation between the simulated behavior of the bombardier beetle's spray mechanism and the behavior observed in the actual beetle [2, 3]. It was these results which inspired the ideas that led to the development of the initial physical μMist® system.

The Bombardier Beetle System and Simulation Work

The series of glands and chambers which make up the spray mechanism of the bombardier beetle are shown in Fig. 2, and the most important section is shown in diagram form in Fig. 3. The core of the beetle's spray mechanism is the chamber and the hard cuticle of the outlet tube, which acts as a pressure triggered exhaust valve. Initially the chamber is mostly empty, the cuticle is closed, blocking the outlet tube, and the inlet tube is open. When the beetle decides to spray, reactants are fed through the inlet tube into the chamber in the form of a dilute aqueous solution. These reactants are a toxic mixture of hydroquinone and hydrogen peroxide which react exothermically due to the addition of a catalyst believed to be in the inner surface collagen of the beetle's combustion chamber. This causes an increase of both temperature and pressure in the chamber. As the pressure in the chamber increases, it causes the inlet tube to be pinched shut by the chamber's extremities, forming a sealed volume of liquid (mainly water, plus some quinones from the products of the reaction). The ongoing exothermic reaction further heats the liquid in the chamber above its usual boiling point. Only a small fraction of the liquid is able to evaporate due to both the inlet and outlet tubes now being blocked. Eventually, the chamber pressure becomes high enough that it can force open a cuticle flap which holds the outlet tube shut. This opening occurs very suddenly and is entirely due to the buildup of pressure



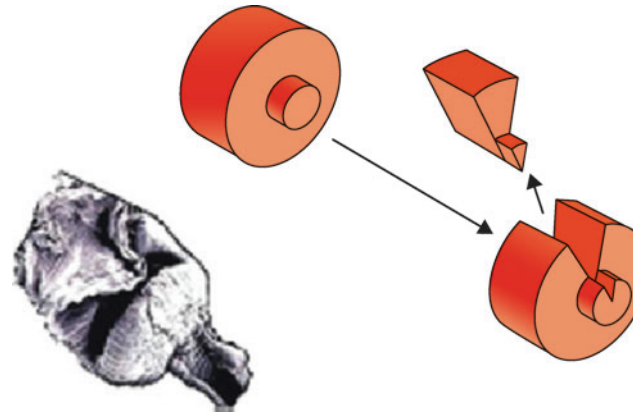
Spray Technologies Inspired by Bombardier Beetle, Fig. 3 A diagram of the spray mechanism of the bombardier beetle at three different stages in the ejection cycle

in the chamber. Thus, the cuticle flap is essentially a pressure relief exhaust valve which is passively triggered in the outlet tube. As the liquid in the chamber has been heated to a point high above its normal boiling point, and the opening of the exhaust valve is so sudden, a very rapid evaporation of the liquid occurs, known as flash evaporation. This flash evaporation does not affect the whole volume of liquid in the chamber, so liquid is caused to be forced out of the outlet tube by a rapidly expanding mass of vapor. The evaporation is rapid due to a vapor explosion, as sudden expansion takes place because of the change in bulk volume of water to vapor. This then shatters the remaining water into small droplets and the mixture is ejected with great force, with shear forces also contributing to the break up into a cascade of ever smaller droplets. The size of the droplets produced are inversely proportional to the magnitude of the forces exerted on and in the liquid volume, so larger forces produce smaller droplets as the surface tension and viscosity of the liquid are more easily overcome. As the liquid is ejected from the bombardier beetle's chamber, the chamber pressure drops, which causes chamber physiology of the outlet tube to now be closed again by the collapse of the cuticle valve. Since the chamber is now partially empty the inlet port is open once again and the process repeats itself. This is exactly the same principle used in modern pulse combustors.

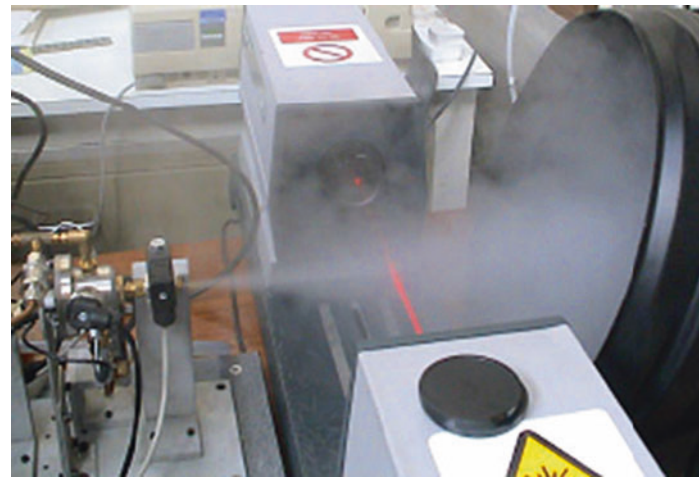
This process was modeled by using Computational Fluid Dynamics (CFD) starting the simulation from the moment at which the cuticle valve on the outlet tube is forced open by the chamber pressure. The chamber is full of liquid (mainly water) already at a suitably high temperature. The inlet tube was not included in the CFD model. The outlet tube however was included in the model and, at the beginning of the simulation, the liquid in the chamber is allowed to flash evaporate. A diagram showing the chamber model used in the simulation can be seen in (Fig. 4) – a slice of the cylindrical chamber is simulated (assuming an axial symmetry within the chamber). The chamber dimensions were such that the volume was similar to that of the bombardier beetle, around 0.1 mm^3. Good correlation was found between the results of the simulation work and the results of tests on the spray produced by the beetle [3] in terms of characteristics such as spray velocity, throw ratio (i.e. ejection distance divided by chamber length), ejection duration, and mass discharged per pulse of spray. This showed that the method of spray atomization used by the beetle was correctly modeled. With this natural process now understood, the research was taken further by building a scaled-up artificial chamber to develop and thereby assess the suitability of this mechanism as a practical approach to liquid atomization and to conduct further research into the process itself.

Overview of the Physical μMist® System

Based upon the simulation work conducted at Leeds University, an experimental demonstration



Spray Technologies Inspired by Bombardier Beetle, Fig. 4 An electron micrograph of a bombardier beetle's reaction chamber alongside diagram of simulated version used in CFD modeling (Electron micrograph courtesy of Prof. T. Eisner, Cornell University)



Spray Technologies Inspired by Bombardier Beetle, Fig. 5 The pre-prototype μMist® system in operation

facility (Fig. 5) was built, which implemented the principles of this liquid atomization method as seen in the bombardier beetle. This system is called μMist®. In a similar way to the bombardier beetle and simulation work, the core of this system is the chamber and valves, as well as a heat source which replaces the heat generated through the bombardier beetle's chemical reaction. Unlike the bombardier beetle and the early simulation work, the exhaust valve on the physical μMist® system is not opened due to a buildup of pressure in the chamber (that is, passively), but instead, all valves are activated electronically, resulting in an actively controlled pulsed spray. There are also other valves in the system, which control the refill flow into the chamber between ejections, similar to the chamber extremities which control flow into the beetle's chamber through an inlet tube. Opening or closing these valves and the exhaust valve then controls the flow of liquid as seen in the bombardier beetle and results in atomized spray generation. As with the beetle and CFD simulation work, this is due to the formation of a sealed volume of liquid at high temperature in the chamber just before the exhaust valve is opened. It is important for this high-temperature volume of liquid in the chamber to be sealed to allow the required temperature rise with minimal associated vaporization. This ensures that the chamber liquid temperature is allowed to rise above its boiling point and that a portion of liquid will flash evaporate once the exhaust valve is opened. This produces liquid atomization at much lower pressures than utilized in most liquid atomization devices

(which use pressure atomization – forcing the liquid through small holes at high pressure) and is energetically efficient. The fact that flash evaporation is used to achieve fast ejection of the water and steam, combined with appropriate control systems, means the system lends itself to a fine control of spray characteristics, particularly droplet size. This control is further improved by the decoupling of the exhaust valve from the chamber pressure. In an ideal system, there would be no vaporization of the chamber liquid during the heating phase; however, in practice, some vaporization will always occur, as governed by the equation for vapor pressure (Eq. 1). This is the major cause for the rise in chamber pressure seen in the bombardier beetle, which in that case triggers the opening of the passive pressure relief exhaust valve. Therefore in the bombardier beetle system, this limits the atomization level achieved, since the extent of flash evaporation will be roughly similar at each ejection. The liquid is heated a similar amount and exhausted at a similar time due to the correlation between chamber pressure and chamber temperature. This follows empirical equations [6] such as Eq. 1.

$$\ln P = C_1 + \frac{C_2}{T} + C_3 \ln T + C_4 T^{C5} \quad (1)$$

Where:

P = Vapor pressure of the liquid
T = Temperature of the liquid
C_1, C_2, C_3, C_4, C_5 = A series of constants depending on the liquid; values for these constants for a range of liquids can be found in [6].

As the exhaust valve on the physical μMist® system is not pressure triggered, but actively controlled electronically, the extent of flash evaporation can be controlled to a greater degree than that of the bombardier beetle.

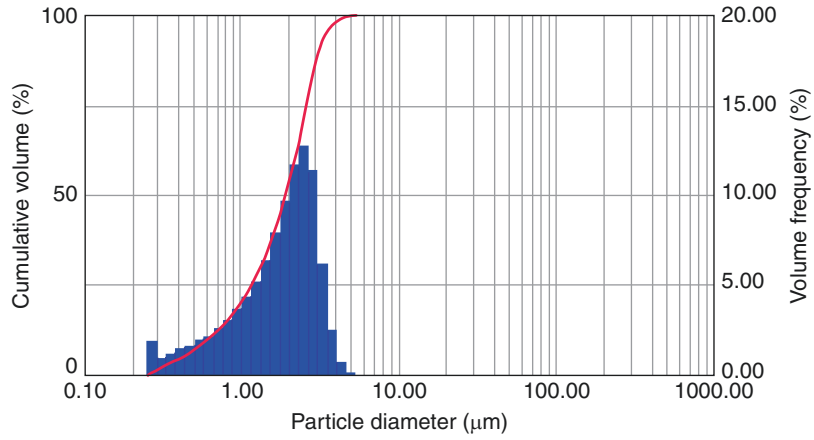
The physical scale of the first pre-prototype of the μMist® system is much larger than that of the system found in the bombardier beetle. Whereas the bombardier beetle system has a chamber of approximately 1×10^{-3} m in length, the first μMist® pre-prototype system has a much larger chamber, 0.02 m in length. This level of scaling is not limited to the chamber of the system and extends to other parameters, such as the diameter of the feed and outlet tubes. The process in the μMist® system however significantly increases the atomization of the liquid as it exits the chamber. The μMist® system can not only be used to atomize very small liquid volumes like those found in the bombardier beetle, but also much larger volumes with a wider range of practical uses. Some key performance parameters are also scaled with the increase in chamber size, such as the throw ratio. A bombardier beetle can throw liquid 0.2 m, using a chamber 1×10^{-3} m in length. This gives a throw ratio of 200 (Eq. 2). Comparatively, it was found that the chamber of 0.02 m in length on the μMist® system could also achieve a throw ratio of 200, throwing liquid a distance of up to 4 m [4, 5].

$$R_T = \frac{D_T}{L_C} \quad (2)$$

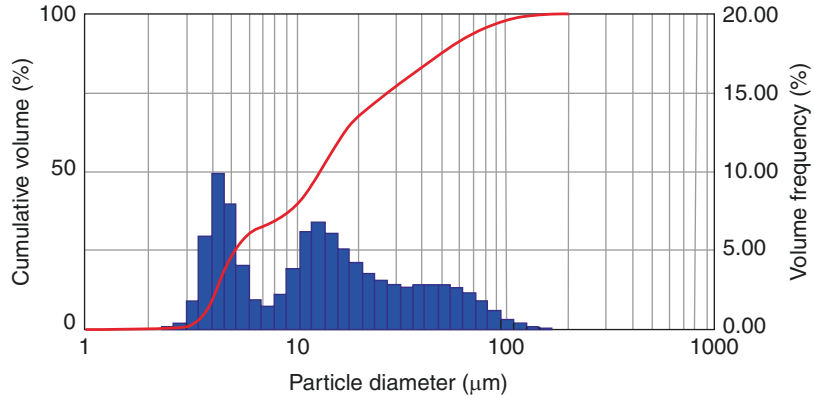
Where:

R_T = Throw Ratio (Dimensionless)
D_T = Distance of throw (m)
L_C = Characteristic length of Chamber (m)

In summary, despite the differences between the physical μMist® system and the liquid atomization seen in the bombardier beetle and the simulation work, the core atomization method is the same. A liquid, heated past its boiling point, and at a constant volume, is suddenly allowed to vaporize though the rapid opening of an exhaust valve. This causes a flash evaporation of a portion of the liquid, which generates a very large force, which then ejects vapor and liquid out through the valve. The flash evaporation is such that as the liquid is rapidly pushed out, large shear forces are generated on and within the volume of the liquid which repeatedly shatter it by overcoming its surface tension and viscosity in a cascade, producing a range of droplet sizes in the emanating spray. Many advantages of the liquid atomization seen in the bombardier beetle, such as high throw ratio, low chamber pressure, and overall energetic



Spray Technologies Inspired by Bombardier Beetle, Fig. 6 Typical droplet size distribution when pre-prototype μMist® system is producing its smallest droplets



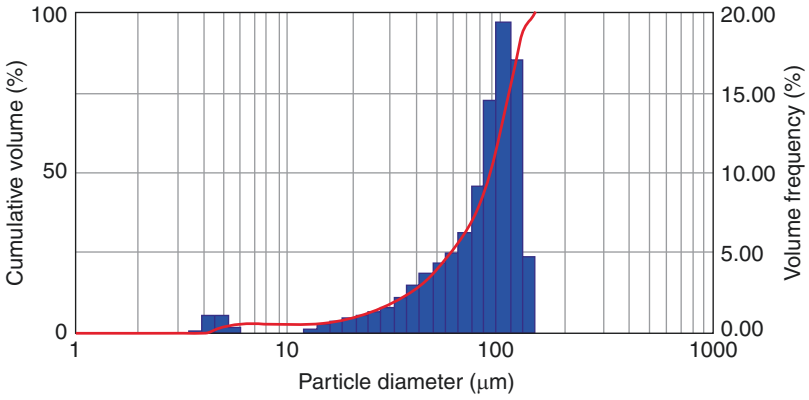
Spray Technologies Inspired by Bombardier Beetle, Fig. 7 Typical droplet size distribution when pre-prototype μMist® system is producing medium sized droplets

efficiency of the system are also seen in the physical μMist® system, despite it being on a much larger scale. The μMist® pre-prototype system also expands upon the liquid atomization method by decoupling the exhaust valve from the chamber pressure, allowing more control over an already versatile system.

Performance of the μMist® Spray System

One of the key features of the μMist® system as a liquid atomization system is that it can be controlled to produce large variation in the characteristics of the spray produced. The type of spray produced is controlled by a number of factors including pressure and temperature in the chamber. Consequently, a range of spray characteristics can be achieved which is much wider than possible with other liquid atomization systems. Experimental work on the μMist® system has shown that it can generate a very wide range of spray characteristics such as droplet size distribution, the ejection velocity of the spray, the mass ejection rate, and the temperature of the ejected spray.

One of the most significant characteristic for many spray applications is the droplet size distribution of the spray produced. The volume of droplets in a particular droplet size band is also important. This is usually categorized graphically, such as is shown in Figs. 6, 7, and 8. A usual indicator of size of droplets is given by calculating the droplet size below which the total volume of droplets account for 90 %, 50 %, and 10 % of the overall spray volume. These indicators are termed Dv (90), Dv (50), and Dv (10), respectively. The graphs in Figs. 6, 7, and 8 show typical droplet size distributions recorded during experimental



Spray Technologies Inspired by Bombardier Beetle, Fig. 8 Typical droplet size distribution when pre-prototype μMist® system is producing its largest droplets

work on the first μMist® pre-prototype system under different input conditions [4]. These results were recorded using a Malvern Spraytec system, which measures droplet sizes through laser diffraction. The graphs give an idea of the range of droplet sizes that can be achieved by altering basic input parameters of the system. Figure 7 is a good example of bimodality in the droplet size distribution, where the majority of the liquid volume is contained within two peaks, rather than one. This result is significant as some spray applications require two or more ranges of droplet sizes to make up the majority of the spray volume, each of which are designed to perform different tasks and when combined produce a more effective delivery than a unimodal distribution. A μMist® system would be particularly suited to these sorts of applications. The ability to generate such a wide range of droplet size distribution shown in Figs. 6, 7, and 8 shows that the μMist® system could potentially provide a single technology which can be used to satisfy a wide range of spray applications.

Applications of the μMist® Spray System

There are many potential applications for the μMist® system in the global market of sprays. This is primarily due to the wide range of spray characteristics which can be achieved using this technology and also due to it being much more environmentally friendly than current spray technologies. There are also related energy efficiencies inherent in the μMist® system, while using water can replace environmentally unfriendly substances currently in use for aerosols and some other spray applications. Some of the potential applications are currently being actively pursued in industrial development programs led by Swedish Biomimetics 3000® Ltd using its V^2 IO innovation acceleration business model. In this section, a few examples of potential applications for the type of droplet size distributions seen in Figs. 6, 7, and 8 will be examined, showing how the μMist® system is suited meet to a wide range of requirements.

Fuel Injectors

A very important application of the μMist® technology is to fuel injectors, and this is one of the potential applications currently being pursued by Swedish Biomimetics 3000® Ltd, with initial support from Carbon Connections. The droplet size distribution in Fig. 6 at the lower end of the range of the μMist® system shows why μMist® is a potential fuel injector technology. Here it can be seen that very small droplet sizes can be produced, resulting in droplets with a high surface area to volume ratio. This droplet size distribution is particularly suited to combustion applications, such as fuel injection, etc., as the larger overall surface area means that the fuel burns much more efficiently. Current fuel injectors primarily work on the principle of pressure atomization, where liquid is atomized by using high pressure to force it through a small opening to create a fine spray. It is this that generates the stresses on the surface of

the liquid which cause it to break down into small droplets [7], rather than the flash evaporation seen in the μMist® system. Generating these high pressures requires a significant amount of energy due to the large pressures used (over 1000 Bar) to ensure proper atomization of the liquid. The energy requirements are large even to produce relatively large droplets. The μMist® system would require substantially less energy to properly atomize the liquid, as all that is required is to heat the system to a relatively modest temperature. Additionally, typical fuel injectors produce droplet sizes in the region of 30–80 μm. It can be seen from Fig. 6 that a μMist® system is capable of producing droplet sizes far smaller than this, in the region of 1–3 μm with some fluids. This would be a significant improvement over current fuel injection technologies, as not only do the smaller droplets burn much more efficiently, increasing engine efficiency and decreasing fuel consumption, but the more complete burning of the fuel in the engine would also lead to fewer harmful emissions being released into the atmosphere. This is in addition to the much lower energy consumed to atomize the fuel compared to current technologies.

Drug Delivery Systems

Another potential application of the droplet size distribution shown in Fig. 6 – again being pursued by Swedish Biomimetics 3000® Ltd, with support from the NIHR i4i program – is in the design of next-generation drug delivery systems. For many illnesses, the preferred method of drug delivery is directly to the lungs. This is best achieved using small droplet sizes (typically less than 10 μm) that allow the liquid to travel deep into the lungs, where it can be most rapidly absorbed into the blood stream. The inherent efficacy of inhaled therapeutic drugs makes this a key development area. μMist® could potentially provide a generic drug delivery system, with one unit being capable of delivering a wide range of drugs, or a personalized medicine device, which is built to satisfy a particular patient's unmet medical needs. Current investigations are also being made to address the pharmaceutical industry's need for innovative drug delivery systems for the administration of novel compounds, including, but not limited to, peptides and oligonucleotides. This would be a significant step forward in drug delivery technology. Other potential drug delivery applications of the μMist® system include needleless injection and nasal drug delivery.

Consumer Aerosols

The μMist® system could also have significant impact as a consumer aerosol generator, primarily due to its technically advanced performance, but most importantly due to environmental benefits. Standard spray/aerosol cans, as well as most medical inhalers, generally use Volatile Organic Compounds (VOCs) such as propane and butane to generate the high pressures required to atomize the delivered liquid as it exits the can through the nozzle. It is thought that these VOCs cause environmentally damaging atmospheric compounds such as carbon dioxide and other atmosphere borne reactive species. They are also highly flammable and therefore are a considerable safety risk at the point of use. The μMist® system has the capability of delivering consumer spray aerosols as well or even better than the spray systems currently used without the need for possibly harmful dangerous VOCs. The only by-product of this system, besides the active ingredient itself, is water.

Fire Extinguishers and Fire Suppressants

The droplet size distribution seen in Fig. 8 primarily features large droplets (around 100 μm) with a wide overall droplet size distribution (down to around 4 or 5 μm). This sort of droplet size distribution would be well suited to a fire-fighting system as the different droplet sizes achieved both have a significant role in fighting fires. Large droplets, maybe ∼100 μm, are effective at cooling the fire to a level below its reaction temperature, whereas small droplets evaporate very quickly and move oxygen away from the source of the fire – fire suppression. This range of droplets being produced by one system would allow fires to be extinguished and suppressed efficiently as the fire progresses. Another significant feature of the μMist® system when producing droplets in this range is that it has a very high throw ratio,

even for small droplets, and thus targeting is possible. As mentioned previously, the current μMist® system has achieved a throw distance of up to 4 m under these conditions, which is a throw ratio with respect to the characteristic chamber length of 200 [4, 5]. With larger chamber volumes, it is envisaged that an even greater throw distance can be achieved, which would be particularly relevant to fire-fighting applications.

In summary, the wide range of spray characteristics and the significantly reduced environmental impact of the μMist® spray system allow it to be well suited to many applications, in particular fuel injection, drug delivery, consumer aerosol systems, fire-fighting systems, and fire-suppressant systems. In general, the μMist® spray system offers improvements in overall performance, efficiency, and environmental impact. Additionally, the system is economic, being very efficient in energy use – especially in comparison to pressure atomization systems.

Conclusions

The μMist® system is a versatile new liquid atomization technology inspired by the unique self-defense mechanism found in the bombardier beetle. The liquid atomization occurs when a sufficiently heated liquid in a sealed chamber is caused to flash evaporate due to the sudden release of an exhaust valve. The flash evaporation mechanism is a new method of providing the large forces necessary to rapidly eject the liquid from the chamber, with shear forces shattering it into a cascade of small droplets. Initial work on the understanding of the bombardier beetle's defense system as a general liquid atomization system was undertaken by McIntosh and Beheshti [2–5] at Leeds University, taking the form of a series of CFD simulations to better understand the key thermodynamic and fluid dynamic principles of the atomization process. This understanding of the thermodynamic and fluid dynamic principles of the atomization process was then used to inspire a design and then construct the first pre-prototype μMist® system as a base for experimental work and further research.

The experimental work on the μMist® system showed it to be highly versatile, producing a wide range of variation in the resulting spray characteristics. For example, it was shown experimentally that the system could be designed to produce droplets with diameters of anywhere between ~1 and 100 μm. Other spray characteristics which showed similar levels of variation due to the design of the μMist® system included the ejection velocity of the spray, modality cone angle or divergence, the mass ejection rate, and the temperature of the ejected spray.

The large variation in the spray characteristics seen in the experimental work on the μMist® system shows it to be adaptable to a wide range of applications, including fuel injection, drug delivery, consumer aerosols, fire fighting, and fire suppression. In some cases, the performance benefits of using the μMist® system would create significant efficiency savings and related environmental and economic benefits. In addition, further environmental benefits are possible as the μMist® system removes the need for harmful VOCs in those spray systems that currently use them, outputting only water, as well as requiring much less energy than standard pressure atomization systems.

Swedish Biomimetics has a worldwide exclusive licensing agreement with the University of Leeds to research, develop, and to commercialize the μMist® platform technology and its various potential applications. Patent pending publication Nos. US11/528,297 and WO2007/0342307 are in place for the μMist® platform technology and its application into fuel injection, respectively.

Cross-References

- ▶ Bioinspired CMOS Cochlea
- ▶ Biomimetic Flow Sensors
- ▶ Biomimetic Mosquito-like Microneedles
- ▶ Biomimetic Muscles and Actuators
- ▶ Biomimetics
- ▶ Biomimetics of Marine Adhesives
- ▶ Toward Bioreplicated Texturing of Solar-Cell Surfaces

References

1. Eisner, T.: For Love of Insects. Harvard University Press, Cambridge (2005)
2. Beheshti, N., McIntosh, A.C.: A biomimetic study of the explosive discharge of the bombardier beetle. J. Des. Nat. **1**(1), 61–69 (2007)
3. Beheshti, N., McIntosh, A.C.: The bombardier beetle and its use of a pressure relief valve system to deliver a periodic pulsed spray. Bioinspir. Biomim. **2**, 57–64 (2007)
4. Beheshti, N., McIntosh, A.C.: A novel spray system inspired by the bombardier beetle. In: Brebbia, C.A. (ed.), Presented Weds 25th June 2008 as Invited Presentation, Design and Nature IV, Proceedings of 4th International Conference on Design and Nature, Tivoli Almansor, Algarve, 24–26 June 2008. Design and Nature IV. WIT Transactions on Ecology and the Environment, vol. 114, pp. 13–21. WIT Press, Southampton/Boston (2008)
5. McIntosh, A.C., Beheshti, N.: Insect inspiration. Phys. World Inst. Phys. **21**(4), 29–31 (2008)
6. Perry, R.H., Green, D.W.: Perry's Chemical Engineers' Handbook. McGraw-Hill, New York (2008)
7. Lefebvre, A.: Atomization and Sprays. Taylor & Francis, London (1989)

SPS of Carbon Allotropes

► Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering

Sputtering

► Physical Vapor Deposition

Stable Clusters

► Prenucleation Clusters

Stable Pre-critical Clusters

► Prenucleation Clusters

Stereolithography

Hyundoo Hwang and Yoon-Kyoung Cho
School of Nano-Bioscience and Chemical Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Ulju-gun, Republic of Korea

Synonyms

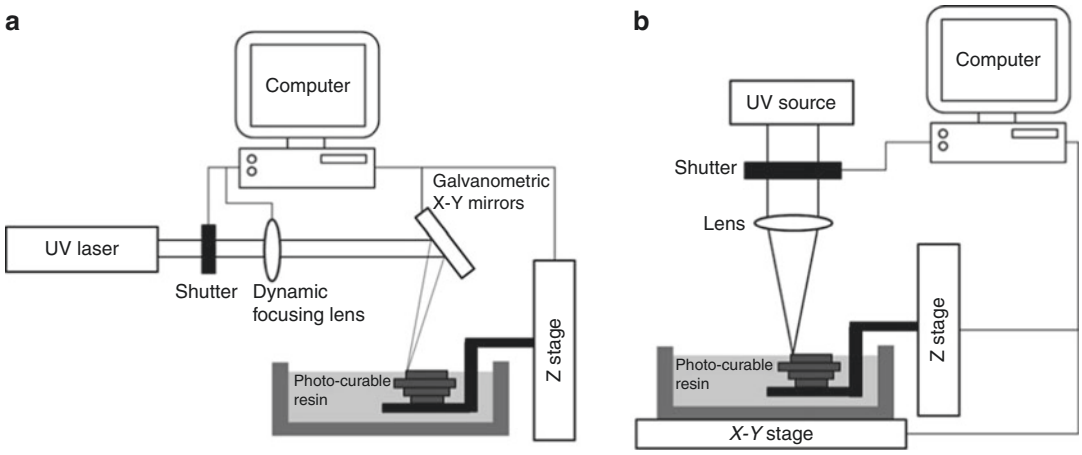
Microstereolithography

Definition

Stereolithography (SL) is an additive manufacturing process based on photopolymerization of a photo-curable polymer. The solid polymer part can be built via localized polymerization by its selective exposure to light in layer-by-layer format. SL is a kind of solid freeform fabrication (SFF) process for manufacturing solid objects by the sequential delivery of energy or material to specified points in space. SFF is sometimes referred to as rapid prototyping, rapid manufacturing, additive manufacturing, additive fabrication, or layered manufacturing. Microstereolithography or nanostereolithography simply refers to SL which possesses feature resolutions significantly higher than several tens of micrometers or nanometers precision, respectively.

Overview

Stereolithography (SL), which was first introduced in 1981 [1], is a three-dimensional (3D) manufacturing technology based on a layer-by-layer photopolymerization of a photo-curable polymer, so-called resin. A light irradiates or traces in a specific cross-sectional pattern on each layer of the photo-curable polymer, which is initially a liquid state, resulting in the solidification of the resin in the exposed area by the photopolymerization process. After every exposure step for a single layer, the other sliced



Stereolithography, Fig. 1 Schematic illustrations of (**a**) a classical stereolithography system using a dynamic focusing lens and (**b**) a microstereolithography system using a *X-Y* stage

two-dimensional (2D) pattern is exposed and polymerized on the below layer by moving the vertical position of the resin chamber or the light. This layer-by-layer 2D fabrication process is repetitively continued until the formation of 3D structure is completed. The performance of SL depends on various factors including the nature of resin, apparatus for fabrication process, strength of light energy, drawing speed of light beam, focusing angle of light beam, depth of focus, and constrained method of liquid-state resin.

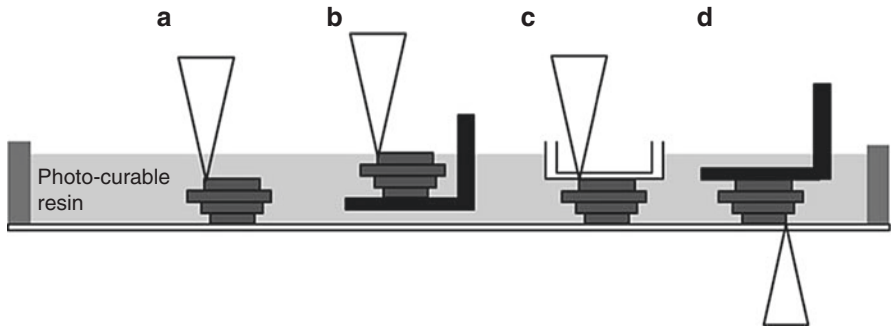
Due to the ardent wish of many researchers for reducing the minimum feature of structures that one can manufacture as small as possible, several kinds of stereolithographic techniques, which use much more tightly focused light to initiate polymerization and cure the part in micro- or nanoscale, have arisen. Microstereolithography (MSL), which allows much higher spatial resolution than conventional SL, has been one of the most promising technologies for manufacturing 3D structures in microscale due to its capability for direct fabrication of complicated 3D structures with high accuracy up to 1 or 2 μm. Most of the MSL techniques have been based on the conventional 2D photolithography using ultraviolet (UV) light source and UV curable polymer. Therein, UV exposure for hardening the liquid-state resin is repetitively conducted on each layer for finally forming 3D microstructures by stacking one layer at a time. Depending on the methods of light exposure, there are two big categories in SL for micro-/nanofabrication: (1) beam scanning-based [2–4]; and (2) image projection-based approaches [5–7]. The former uses a focused light beam tracing a specific 2D pattern over or within the liquid-state resin for each layer, while, in the latter, an image pattern generated by a dynamic pattern generator irradiates only one time for each layer.

Methodology

UV Beam Scanning Stereolithography

In the scanning-based SL process, a very tightly focused laser beam is scanned over or within the liquid-state polymer to produce a 3D solid object. There are several ways to construct a 3D solid part using this scanning-based SL scheme in the perspectives of (1) scanning component – a mirror or an *X-Y* stage (see Fig. 1); (2) light beam direction – from upper or bottom; (3) surface conditions of resin – free or constraint (see Fig. 2); and (4) number of light beam – single or multiple.

With the mirror scanning method, it is very important to maintain the focal point of the light beam over the planar surface of the liquid resin. For this purpose, a dynamic focusing lens



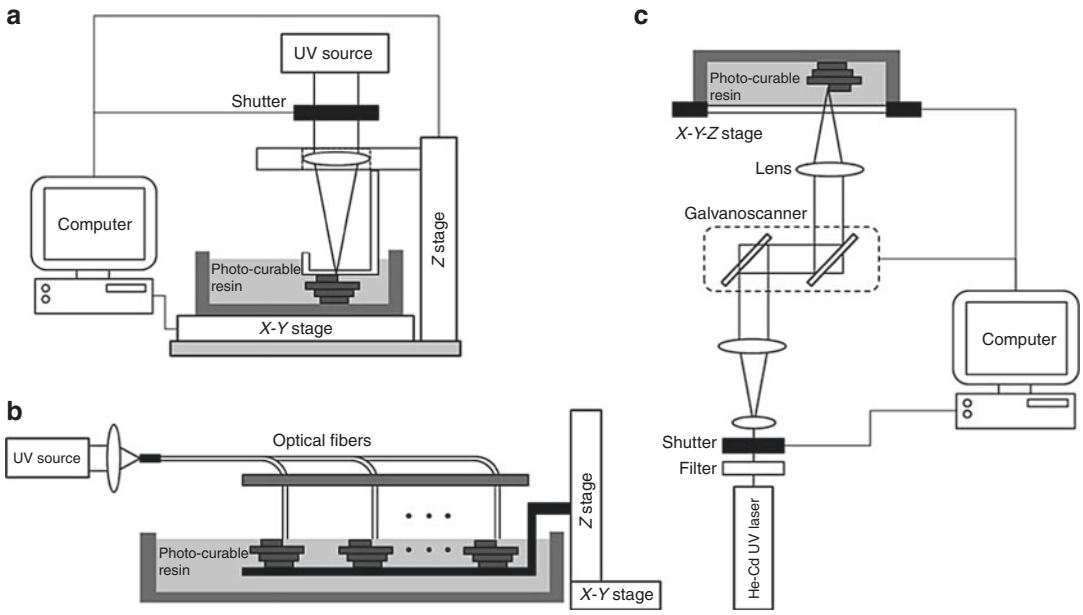
Stereolithography, Fig. 2 Schematic illustrations of four basic configurations for a stereolithography process. Free surface configurations with a light beam projected from upper direction (**a**) within and (**b**) over the surface of the liquid photo-curable resin. Constraint surface configuration with a light beam projected from (**c**) *upper* and (**d**) *bottom* directions

controlled by a computer should be required in a classical SL system (see Fig. 1a). However, it is very difficult to achieve high resolution using this dynamic light focusing. To deal with this problem, not only a Z-stage, but also an *X*-*Y* stage for scanning the resin chamber has been applied instead of the dynamic focusing lens (see Fig. 1b). While the chamber containing the liquid resin is translated in the *X*-*Y* plane, the focal point of the light beam keeps its position in the plane to achieve the tightest focus as possible. However, the maximum scanning speed of this *X*-*Y* stage scanning method is lower than that of the method using a rotating mirror.

Four types of SL system which can be distinguished on the basis of two criteria – the direction of light beam and the surface conditions of resin – are shown in Fig. 2. In the free surface configurations shown in Fig. 2a, b, the surface of the liquid-state resin is free, while the liquid surface is covered by the transparent plate, which allows UV beam passage with little energy loss, in the constraint surface configurations shown in Fig. 2c, d. The free surface configurations are relatively easy to construct, but it is difficult to control the thickness of each layer due to the viscosity of the liquid resin causing a relatively long time to be stabilized. On the other hand, the constraint surface configurations allow more accurate control of the layer thickness by controlling the height of the plate. However, the photopolymerized resin can be often adhered to the plate and broken in this constraint surface method.

In 1993, Ikuta and his colleagues have first demonstrated the MSL system using this constraint surface configuration as shown in Fig. 3a [2]. They scanned the resin chamber in the *X*-*Y* plane and the focused light beam in the *Z*-axis. They could achieve the minimum feature size of 5 μm in the *X*-*Y* plane and 3 μm in the *Z*-axis. However, it took about 30 min to produce a microstructure, the dimension of which is only 10 μm × 10 μm × 1,000 μm. To obtain a large number of 3D microstructures at a time using MSL, Ikuta and coworkers reported an advanced MSL system using multiple optical fibers in 1996 [3]. In this system, the light source was tightly focused onto a fiber-optic bundle, and each fiber end was placed in an ordered array for generating a series of exposure point sources in the liquid resin chamber, which is scanned in *X*-*Y*-*Z* directions (see Fig. 3b).

Despite these advances in MSL, there have still been some challenging issues caused by the basic scheme of original SL process, which is a layer-by-layer process. In the layer-by-layer process, the vertical resolution of the polymerized objects is limited by the thickness of the layer piled up. In addition, surface tension of viscous liquid resin can deform and destruct the solidified parts in microscale during the fabrication process. To overcome these limitations, Ikuta and coworkers have developed a more advanced MSL system,



Stereolithography, Fig. 3 Schematic illustrations of several types of microstereolithography (*MSL*) systems. (**a**) The first-generation MSL system based on constraint surface condition, a light from upper direction, and an *X-Y* stage. (**b**) Fiber-optic MSL system for massive manufacturing in large area. (**c**) High-resolution MSL system using a tightly focused UV laser source irradiating inside the photo-curable resin

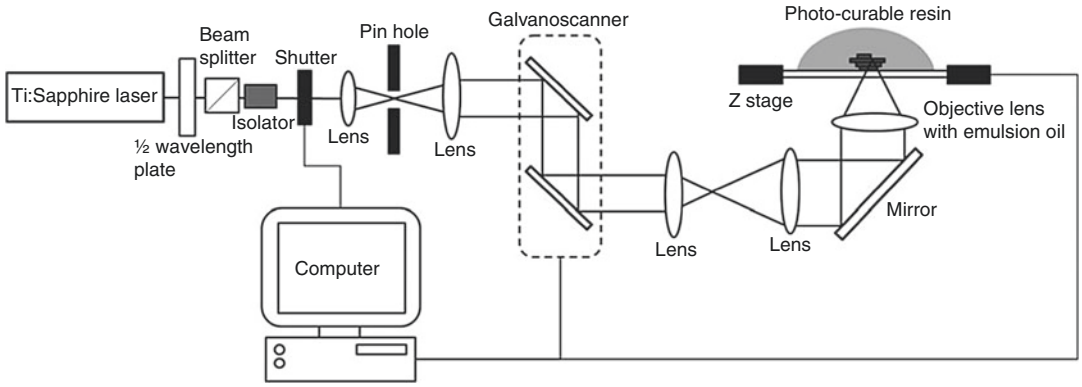
in which a UV laser source is tightly focused into the liquid resin, not over its surface, to induce the photopolymerization only at the focal spot of the beam, in 1998 (see Fig. 3c) [4]. Since the photopolymerization occurs only around the focal spot of the beam inside the liquid resin, not the surface, the problems related to the layer-by-layer process, which are mentioned above, can be reduced. They could produce freely moved 3D microstructures with 500 nm resolution using this technique. However, extensive polymerization in the defocused regions, where the beam passes through, is an unavoidable issue associated with this technique.

Two-Photon Stereolithography

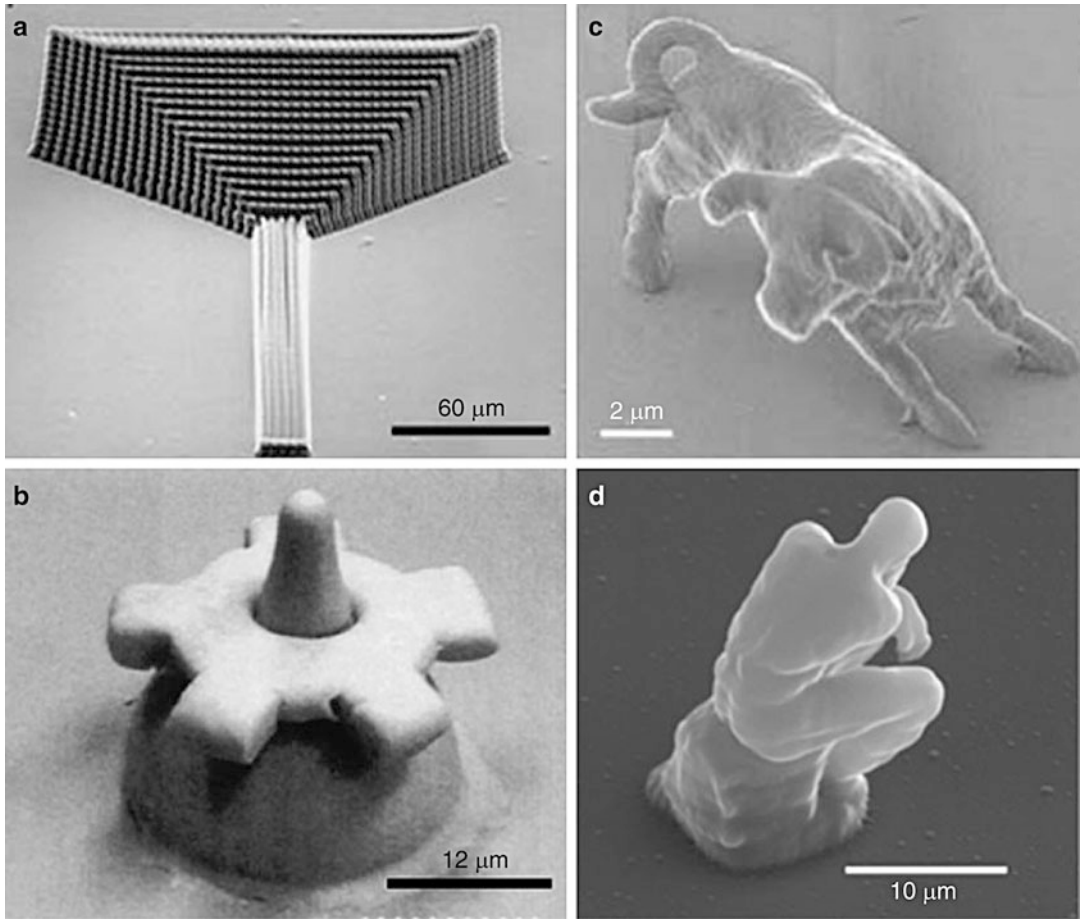
Two-photon stereolithography (TPS), which is based on two-photon induced photopolymerization, is one of typical SL techniques for the 3D micro-/nanofabrication [8–13]. The TPS technology is based on the polymerization of liquid-state polymer via two-photon absorption (TPA) within very local regions inside the focused high-intensity laser beam tracing in a specific cross-sectional pattern (see Fig. 4). TPA is the simultaneous absorption of two photons by a molecule resulting in its excitation from a low-energy electronic state to a higher energy electronic state. The TPA-initiated polymerization of liquid-state polymer occurs only at a very narrow focal spot of the beam where the light intensity is highest. Thus, the spatial resolution achievable by the TPS is almost 100 nm, which is smaller than the diffraction limit of the beam, while that of other SL systems based on single-photon polymerization is limited by the optical diffraction limit. Based on the fascinated advantages of TPS technology, SL could meet 3D nanofabrication since the 2000s. Not only photonic devices [9] and micromachines [10], but also microsculptures of bull [11] and The Thinker [12] with nanoscale precision have been fabricated using this technology as shown in Fig. 5. To date, the TPS systems have been able to produce the smallest objects based on MSL.

Image Projection Stereolithography

Another category of the approaches for MSL is the image projection-based one. In this method,

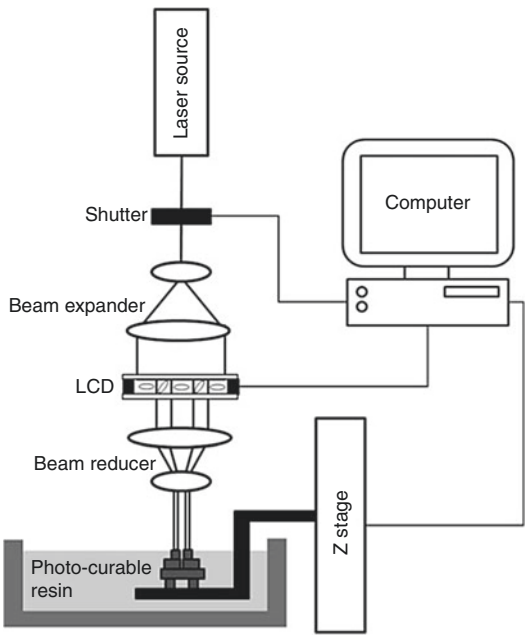


Stereolithography, Fig. 4 Schematic illustration of two-photon stereolithography system



Stereolithography, Fig. 5 Microstructures fabricated by two-photon stereolithography. Various structures including (**a**) a photonic device (Reprinted by permission from Macmillan Publishers Ltd: Nature [9], copyright (1999)) (**b**) a micromachine (Reprinted with permission from Ref. [10]. Copyright 2000, The Optical Society), and microsculptures of (**c**) bull (Reprinted by permission from Macmillan Publishers Ltd: Nature [11], copyright (2001)) and (**d**) The Thinker (Reprinted with permission from Ref. [12]. Copyright 2007, American Institute of Physics) with nanoscale precision have been fabricated

the photopolymerization and solidification of an object is carried out by irradiating the photo-curable resin with an image pattern in one time for one layer. The image pattern for each layer can be generated by a conventional photo-mask or a spatial light modulator (SLM), which is capable of generating and modulating dynamic image patterns in a continuous way. Typical types of the SLM include a liquid crystal display (LCD) and a digital micromirror device (DMD). Both display devices have a large number of pixels, which are individually controllable. The image pattern generated through the LCD or DMD, which are transmissive and reflective types, respectively, is focused by a lens onto the liquid resin surface. By continuously changing the image pattern using a computer, the resin chamber is scanned in the *Z*-axis to construct a 3D part (see Fig. 6). Bertsch and colleagues have first demonstrated the image projection MSL system using an LCD as the dynamic pattern generator in 1997 [5]. Since the mid-2000s, several systems using a DMD have been developed and commercialized because the DMD provides better performance – fill factor of each pixel and reflectivity – compared to the LCD [6, 7]. These image projection-based approaches take much shorter time to fabricate each layer because a large area of liquid resin can be simultaneously exposed in a pre-designed 2D pattern differently from the beam scanning methods, which scan all the area with a tightly focused light. However, these techniques still have the problems due to the layer-by-layer process mentioned above – vertical resolution dependent on the thickness of the layer stacked together.



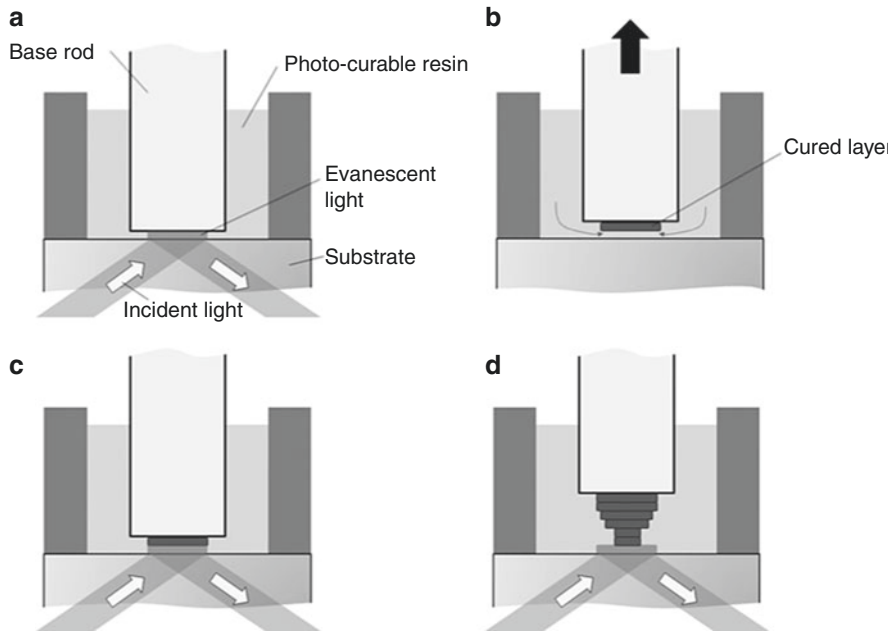
Stereolithography, Fig. 6 Schematic illustration of an image projection microstereolithography system using a dynamic image generated by a liquid crystal display (*LCD*)

Evanescent Light Stereolithography

Evanescent light stereolithography, which uses an evanescent light for the photopolymerization of the liquid-state polymer differently from the conventional stereolithography technologies based on a propagating light, has also been developed [14] (see Fig. 7). When light travels across materials with different refractive indices at an incident angle greater than the so-called critical angle, above which total internal reflection of the light occurs at the interface, the evanescent light is generated within a localized area from the boundary of the substrate and the photo-curable polymer. As a result, the liquid-state polymer is hardened by the evanescent light in the sub-micrometer-thickness layer. By repeating the processes, which includes refilling the boundary with the liquid-state polymer and its photopolymerization by the evanescent light, 3D structures with sub-micrometer scale in vertical axis are formed. However, the lateral resolution of the objects formed by this technique is not very fine yet.

Conclusion

SL has been a promising technology for the direct fabrication of complicated 3D structures in micro-/nanoscale. Although there has been a great deal of achievement in the stereolithographic technologies for several decades, there are still challenging issues, which should be explored for their practical uses: massive production, large-area fabrication, reliability of processes, and improvement of



Stereolithography, Fig. 7 Schematic illustration of evanescent light stereolithography processes. (**a**) First of all, an incident light exposes and cures the photo-curable polymer. (**b**) Next, the cured layer adhering on the base rod is lifted by raising the rod and filling the space with a new layer of the liquid-state polymer. (**c**) Then a light beam with different specific 2D patterns exposes and cures the next layer. (**d**) Finally, the desired object can be fabricated by repeating this loop continuously

photopolymerizable materials. Nevertheless, the SL has been one of the most trustworthy technologies for the fabrication of 3D nano-/microstructures and would take a place as a powerful 3D fabrication technology, which is applicable to diverse research fields.

References

1. Kodama, H.: Automatic method for fabricating a three-dimensional plastic model with photo-hardening polymer. Rev. Sci. Instrum. **52**, 1770–1773 (1981)
2. Ikuta, K., Hirowatari, K.: Real three dimensional micro fabrication using stereo lithography and metal molding. In: Proceedings of IEEE Micro Electro Mechanical Systems, pp. 42–47. Fort Lauderdale (1993)
3. Ikuta, K., Ogata, T., Tsubio M., Kojima, S.: Development of mass productive micro stereo lithography (Mass-IH process). In: Proceedings of IEEE Micro Electro Mechanical Systems, pp. 42–47. Fort Lauderdale (1993)
4. Ikuta, K., Maruo, S., Kojima, S.: New microstereolithography for freely movable 3D micro structure-super IH process with submicron resolution. In: Proceedings of IEEE Micro Electro Mechanical Systems, pp. 290–295. Heidelberg (1998)
5. Bertsch, A., Zissi, S., Jezequel, J.-Y., Corbel, S., Andre, J.C.: Microstereolithography using a liquid crystal display as dynamic mask-generator. Microsyst. Technol. **3**, 42–47 (1997)
6. Hadipoespito, G., Yang, Y., Choi, H., Ning, G., Li, X.: Digital micromirror device based microstereolithography for micro structures of transparent photopolymer and nanocomposites. In: Proceedings of the Solid Freeform Fabrication Symposium, pp. 13–24. Austin (2003)
7. EnvisionTEC.: http://www.envisiontec.de. Accessed 16 Mar 2011
8. Maruo, S., Nakamura, O., Kawata, S.: Three-dimensional microfabrication with two-photon-absorbed photopolymerization. Opt. Lett. **22**, 132–134 (1997)
9. Cumpston, B.H., Ananthavel, S.P., Barlow, S., Dyer, D.L., Ehrlich, J.E., Erskine, L.L., Heikal, A.A., Kuebler, S.M., Lee, I.-Y.S., McCord-Maughon, D., Qin, J., Rockel, H., Rumi, M., Wu, X.-L., Marder, S. R., Perry, J.W.: Two-photon polymerization initiators for three-dimensional optical data storage and microfabrication. Nature **398**, 51–54 (1999)
10. Sun, H.-B., Kawakami, T., Xu, Y., Ye, J.-Y., Matuso, S., Misawa, H., Miwa, M., Kaneko, R.: Real three-dimensional microstructures fabricated by

photopolymerization of resins through two-photon absorption. Opt. Lett. **25**, 1110–1112 (2000)
11. Kawata, S., Sun, H.-B., Tanaka, T., Takada, K.: Finer features for functional microdevices. Nature **412**, 697–698 (2001)
12. Yang, D.-Y., Park, S.H., Lim, T.W., Kong, H.-J., Yi, S.W., Yang, H.K., Lee, K.-S.: Ultraprecise microreproduction of a three-dimensional artistic sculpture by multipath scanning method in two-photon photopolymerization. Appl. Phys. Lett. **90**, 013113 (2007)
13. Park, S.-H., Yang, D.-Y., Lee, K.-S.: Two-photon stereolithography for realizing ultraprecise three-dimensional nano/microdevices. Laser Photonics Rev. **3**, 1–11 (2009)
14. Kajihara, Y., Inazuki, Y., Takahashi, S., Takamasu, K.: Study of nano-stereolithography using evanescent light. In: Proceedings of the American Society for Precision Engineering (ASPE) Annual Meeting, pp. 149–152. Orlando (2004)

Stimuli-Responsive Drug Delivery Microchips

Jian Chen[1], Jason Li[2], Michael Chu[2], Claudia R. Gordijo[2], Yu Sun[3] and Xiao Yu Wu[2]
[1]State Key Laboratory of Transducer Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing, People's Republic of China
[2]Advanced Pharmaceutics and Drug Delivery Laboratory, Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada
[3]Department of Mechanical and Industrial Engineering and Institute of Biomaterials and Biomedical Engineering and Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada

Synonyms

Intelligent drug delivery microchips; MEMS-based drug delivery devices; Smart drug delivery microchips

Definition

Stimuli-responsive drug delivery microchips are MEMS-based "smart" drug delivery devices composed of individually sealed drug reservoirs that can be opened selectively for complex drug release by various stimuli, targeting long-term implantation applications.

Overview

Overview of Working Mechanism

Advances in MEMS technology have enabled the precise fabrication of miniature biomedical devices with micrometer-sized features for implantable drug delivery. Drug delivery microchips contain small reservoirs that are loaded with drugs and separated from the outside environment by a drug release barrier. Examples of reported MEMS devices utilize various approaches including electrochemical dissolution, electrothermal activation, chemical degradation, or self-regulation to control temporal drug release or modulate the permeability of the drug release barrier for drug delivery on demand.

Advantages

These microreservoir devices are well suited for applications in chronotherapy due to their ability to achieve timed or regulated drug release with well-defined temporal profiles of drug depots and release mechanisms. In addition, microchip-based implantable drug delivery devices allow for localized delivery by direct placement of the device at the treatment site. Another benefit is that these devices contain no moving parts and are capable of delivering reservoir drugs in a solid, liquid, or gel formulation. The device design also protects the drug depot from the outside environment before release so that stability of the drug inside the reservoirs can be maintained more effectively in the controlled environment.

Potential Concerns

Implantable microchips require minor surgery for implantation and removal, and need to maintain their functions during the application. Therefore, these systems must be biologically inert to minimize inflammatory response and both chemically and physically stable to avoid premature device failure. While the implants need to be as

unobtrusive as possible, there is a fundamental limitation of device size imposed by the need for sufficient storage capacity for a chronic dosing regimen. Suitable candidate drugs will be potent and prepared in concentrate formulations and are stable for extended periods at body temperature in order to minimize implant size.

Methodology

Currently available drug delivery microchips can be divided into three main categories, based on drug release methods: active drug release (e.g., electrochemical dissolution and electrothermal activation approaches), passive drug release (e.g., chemical degradation), and self-regulated drug release (e.g., pH-responsive drug delivery).

Electrochemical Dissolution Approach

As shown in Fig. 1, the device is fabricated by the sequential processing of a silicon wafer using microelectronic processing techniques including UV photolithography, chemical vapor deposition, electron beam evaporation, and reactive ion etching. Samples of reservoirs are sealed at one end by a thin membrane of gold to serve as an anode in an electrochemical reaction. Electrodes are placed on the device to serve as a cathode.

When the microfabricated reservoir system is submerged in an electrolyte solution, ions form a soluble complex with the anode material in its ionic form. An applied electric potential oxidizes the anode membrane, forming a soluble complex with the electrolyte ions. The complex dissolves in the electrolyte, the membrane disappears, and the solution within the reservoir is released. The release time from each individual reservoir is determined by the time at which the reservoir's anode membrane is removed.

Release studies demonstrate that the activation of each reservoir can be controlled individually, creating a possibility for achieving many complex release patterns. Varying amounts of chemical substances in a solid, liquid, or gel form can be released into solution in a pulsatile or continuous manner, or a combination of both, either sequentially or simultaneously from a single device. Such a device has additional potential advantages including small size, quick response times, and low power consumption. In addition, all chemical substances to be released are stored in the reservoirs of the microchip itself, creating a possibility for the future development of autonomous devices.

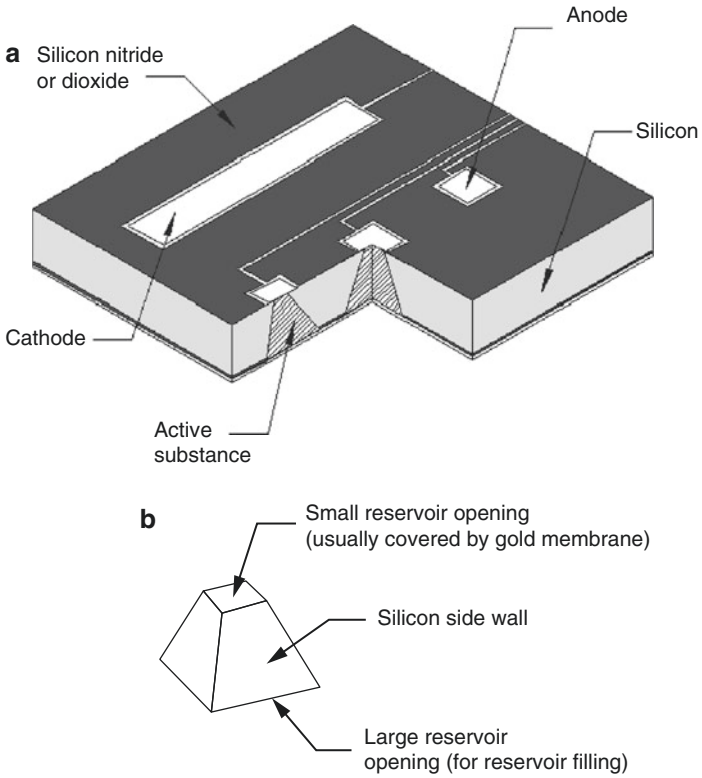
Electrothermal Activation Approach

The devices for this study were produced using standard microfabrication processes which contain an array of individually sealed and actuated reservoirs, each capped by a thin metal membrane comprised of either gold or multiple layers of titanium and platinum (see Fig. 2). The passage of a threshold level of electric current through the membrane causes it to disintegrate, thereby exposing the protected drugs of the reservoir to the surrounding environment. Compared to the electrochemical dissolution approach, electrothermal reservoir opening is more reliable and repeatable than the gradual opening achieved by the corrosion-based method.

In addition, reservoirs were aseptically sealed with spheres of indium-tin eutectic solder by thermocompression bonding. Filled and sealed microchips were electrically connected to the wireless communication hardware, power supply, and circuit boards of the in vivo implant, which were hermetically sealed inside a laser welded titanium case. Pulsatile release of a therapeutic polypeptide on demand, in response to telemetry between an external device controller and the implant, was demonstrated. Importantly, although a tissue capsule formed around the device (as expected), the capsule did not significantly affect the drug release profile during the 6-month implantation. This result addresses a major perceived risk of implanted drug delivery devices that a tissue barrier will compromise the effectiveness and control of release.

Chemical Degradation Approach

Biodegradable polymer version multireservoir drug delivery microchips have also been designed to investigate the feasibility of achieving multipulse drug release from a polymeric system over periods of several months without requiring



Stimuli-Responsive Drug Delivery Microchips, Fig. 1 Schematic illustration of the first implantable drug delivery microchip via electrochemical dissolution approach. (**a**) Fabrication of these microchips began by depositing low stress, silicon-rich nitride on both sides of prime grade silicon wafers using a vertical tube reactor. The silicon nitride layer on one side of the wafer was patterned by photolithography and electron cyclotron resonance-enhanced reactive ion etching to give a square device containing square reservoirs. The silicon nitride served as an etch mask for potassium hydroxide solution, which anisotropically etched square pyramidal reservoirs (**b**) into the silicon along the (111) crystal planes until the silicon nitride film on the opposite side of the wafer was reached. The newly fabricated silicon nitride membranes completely covered the *square* openings of the reservoir. Gold electrodes were deposited and patterned over the silicon nitride membranes by electron beam evaporation and liftoff. A layer of plasma-enhanced chemical vapor deposition silicon dioxide was deposited over the entire electrode-containing surface. The silicon dioxide located over portions of the anode, cathode, and bonding pads were etched with reactive ion etching to expose the underlying gold film; this technique was then used to remove the thin silicon nitride and chromium membranes located in the reservoir underneath the gold anode (Reproduced with permission from Ref. [1])

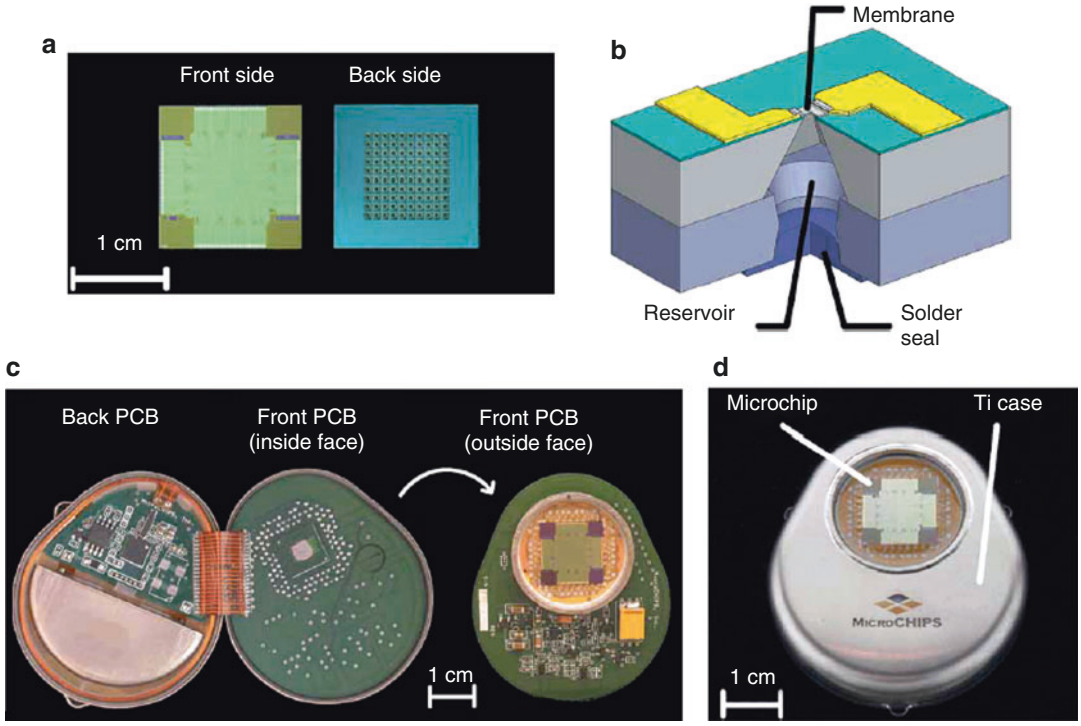
S

the application of a stimulus to trigger the drug release. A prototype device is shown in Fig. 3, from which drugs can be released at different times on the basis of the characteristics of the reservoir membranes that affect their degradation rate, such as the material used, the molecular mass, the composition, or the thickness.

Separation of the release formulation (the reservoir membranes) from the drug formulation (which is loaded into the reservoirs behind the membranes) might enable greater flexibility in adapting this system for a desired application than is currently achievable with existing methods. An advantage of biodegradable polymeric microchips is the elimination of a requirement for a second surgery to remove the device. In addition, the lack of electronics reduces size restrictions in terms of device manufacture.

Self-regulated Approach

Compared to microchips with release mechanisms of electrochemical dissolution, electrothermal

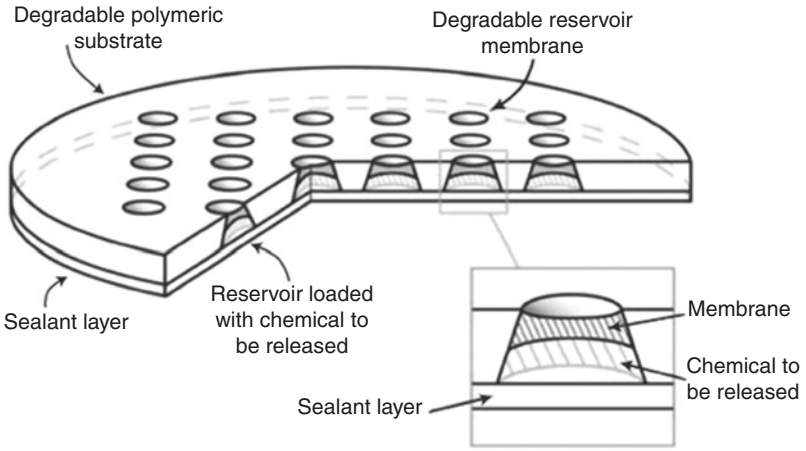


Stimuli-Responsive Drug Delivery Microchips, Fig. 2 Schematic illustration of the implantable drug delivery microdevice via electrothermal dissolution approach under remove control. (**a**) Front and back of the 100-reservoir microchip. (**b**) Representation of a single reservoir. (**c**) Electronic components on the printed circuit board (*PCB*) in the device package. (**d**) The assembled implantable device (Reproduced with permission from Ref. [2])

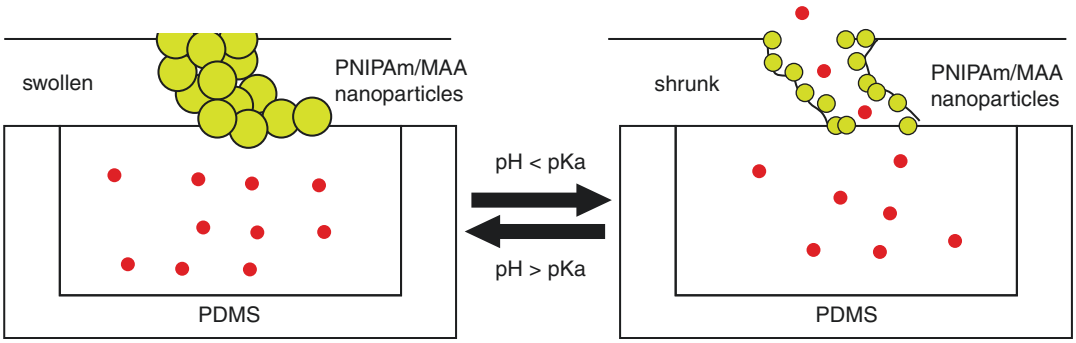
activation, and chemical degradation, which are not capable of regulating drug delivery rates in response to in vivo environmental conditions, a self-regulated microdevice enables drug release in response to local pH or glucose variations by integrating pH and/or glucose-responsive nanohydrogel embedded in composite membranes functioning as intelligent nanovalves [4–10].

As shown in Fig. 4, the patterned PDMS structure forms a drug reservoir and provides physical support for a thin nanohydrogel-embedded composite membrane. The hydrogel nanoparticles detect environmental pH changes and respond with corresponding volumetric swelling or shrinking, which regulates membrane permeability and thereby drug release rate. The polymeric microchips are monolithic without requiring peripheral control hardware or additional components for controlling drug release rates. By adjusting nanoparticle percentages, membrane rigidity, drug reservoir shape and size, and drug loading volume and concentration, well-controlled drug release profiles in response to local pH changes can be achieved, functioning as a platform technology for intelligent drug delivery.

Recent studies have employed this approach for glucose-responsive insulin release. Glucose sensitivity combined with pH responsiveness is achieved with the use of glucose oxidase, an enzyme that catalyzes the oxidation of glucose to produce gluconic acid which in turn reduces local pH. In vitro insulin release profiles demonstrate rapid and repeatable device response to normal and hyperglycemic conditions [6, 8, 9]. Implantation of microdevices into diabetic rats resulted in rapid return to euglycemia following glucose challenges and successfully maintained normal blood glucose levels for several days [8, 9].



Stimuli-Responsive Drug Delivery Microchips, Fig. 3 Schematic illustration of the first implantable drug delivery microdevice with biodegradable membranes. The main body of the device is composed of a reservoir-containing substrate that is fabricated from a degradable polymer. Truncated conical reservoirs in the substrate are loaded with the chemical to be released and sealed with polymeric degradable reservoir membranes on one end and a sealant layer (polyester tape) on the opposite end. *Inset*, close-up of a reservoir, reservoir membrane, sealant layer, and chemical to be released (Reproduced with permission from Ref. [3])



Stimuli-Responsive Drug Delivery Microchips, Fig. 4 Illustration of the mechanism for pH-responsive drug release out of the microdevice. *Left*: Nanoparticles are in the swollen state when the surrounding pH value is higher than pKa (acid dissociation constant) of the nanoparticles. *Right*: Nanoparticles are in the shrunk state when the surrounding pH value is lower than pKa. Resulting volumetric swelling and shrinking of the nanoparticles control drug release rates [5, 7]

S

Key Research Findings

Device Dimension and Shape

For versatility and acceptability, overall device size and shape are major considerations. To be commercially attractive, the device should be small enough to be implanted subcutaneously in a doctor's office, with local anesthesia and a profile that is not obtrusive or apparent. The shape and materials of the device must allow easy removal when the treatment is no longer needed or the device requires replacement.

MEMS-based implants must accommodate a sufficient quantity of drug doses in the device of minimum size, creating a design hurdle and constraint. This constraint requires the nondrug components to occupy minimal volume. Suitable candidate drugs should be potent, prepared in high-concentration formulations, and be stable for extended periods at the body temperature.

Materials

To minimize chronic inflammation and immune response, the tissue-contacting components must be constructed from biocompatible materials [11]. These materials must be stable in physiological fluids and tissues and nontoxic over the lifetime of the device. In addition, leachable from metal and polymeric materials should be minimal, nontoxic, and hypoallergenic. Silicon has traditionally been the most common material used for microdevice construction. More recent, device designs typically make use of biologically inert polymers such as polymethyl methacrylate (PMMA) and polydimethylsiloxane (PDMS) for device construction. These polymers have superior properties to silicon for BioMEMS applications with regard to cost and versatility of physical properties.

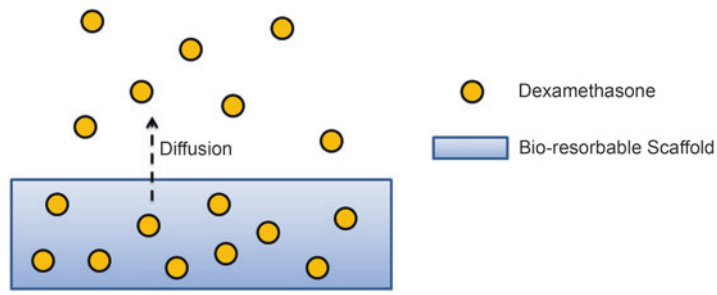
Inflammatory Response and Biocompatibility

Implantation of a foreign material within the body activates the host inflammatory response, which may influence device function and longevity through local production of reactive oxygen species, cell-mediated degradation, or formation of an impermeable fibrous capsule around the device [12]. The type and severity of the inflammatory response depend on the nature of the implant material, the geometry, surface topography, and various surface treatments of the device [10, 13, 14]. Thus, these parameters may be tuned to minimize the host inflammatory response and improve the overall biocompatibility of the device.

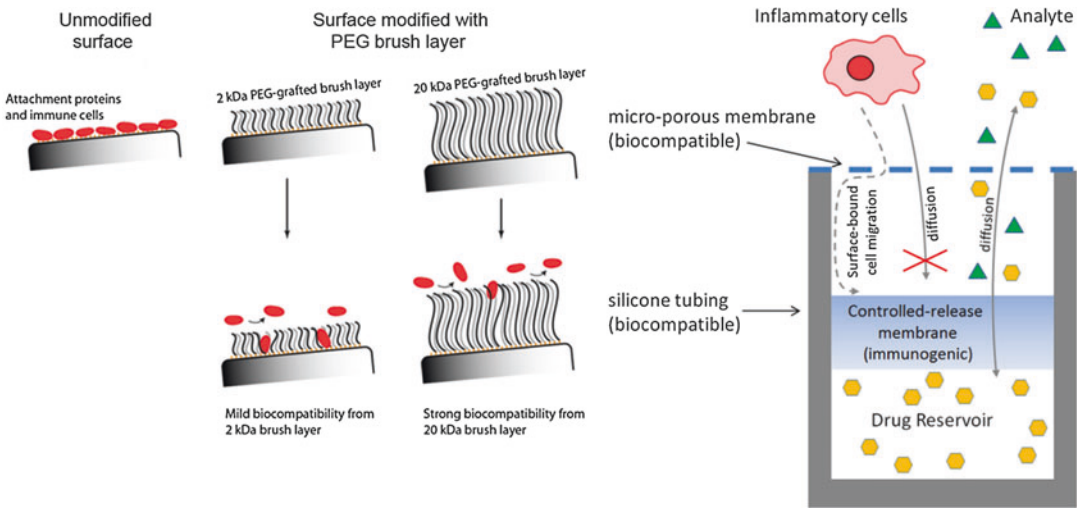
Biologically inert polymers can be used to mask non-biocompatible materials from the body in cases where their interaction with biological tissues is not required [15]. However, most stimuli-responsive drug delivery systems feature non-biocompatible active components such as metallic or semiconducting sensing electrodes, stimuli-responsive polymers, or enzymes that must contact and interact with the host tissue or biochemicals for proper device functions. For these components, several strategies, namely, active and passive strategies, have been devised for mitigating the immune response while simultaneously allowing efficient diffusion of molecules to and from the device [16].

Active strategies for attenuating the inflammatory response involve the sustained local release of anti-inflammatory agents such as dexamethasone [17], nitric oxide [18], sirolimus [19], or paclitaxel [20] to suppress the normal wound healing process (Fig. 5). Care must be taken when employing these agents as they may cause adverse side effects or complications.

Passive strategies reduce the inflammatory response by minimizing serum protein adsorption and immune cell adhesion onto the implant surface. Steric hindrance is generated using non-fouling self-assembled monolayers, polymer brushes, or hydrogel coatings made from hydrophilic polymers such as poly(ethylene glycol) (Fig. 6a) [10, 21, 22] or surface-supported phospholipid bilayers [23]. In doing so, protein adsorption onto the implant surface is minimized, and thus, downstream recruitment of inflammatory cells is prevented. Alternatively, the inflammatory response can be minimized by restricting access of immune cells themselves to the implant surface by covering the implant with nano-porous size-exclusion membranes [24] or with the use of PEGylated microporous membranes to geometrically impede cell migration to the implant surface (Fig. 6b) [25].



Stimuli-Responsive Drug Delivery Microchips, Fig. 5 Active strategy for mitigating the host inflammatory response involving the sustained release of anti-inflammatory agents from the implant surface



Stimuli-Responsive Drug Delivery Microchips, Fig. 6 Representative passive strategies for mitigating the host inflammatory response following device implantation. (**a**) Use of non-fouling self-assembled monolayers, polymer brushes, or hydrogel coatings to minimize protein adsorption onto implant surfaces (Reproduced with permission from Ref. [10]). (**b**) Use of microporous membranes and device geometry for hindering cell migration to the active implant surface [25]

Future Work

MEMS and miniaturization technologies have been used for producing novel stimuli-responsive drug delivery microchips that require minimal invasiveness of the medical procedures for implantation. Translation of this technology to the clinical setting will require improvements in implant reliability, safety, and lifetime. To achieve these goals, research efforts will focus on improving implant biocompatibility and development of stable and highly concentrated drug formulations.

Electrical systems will further require development of robust electrical components for selectively addressing individual reservoirs, biosensors for information acquisition, wireless communication hardware for remote control, and a long-term power source. These drug delivery implants will be tailored for personalized medicine and other unmet medical need [26].

References

1. Santini Jr., J.T., Cima, M.J., Langer, R.: A controlled-release microchip. Nature **397**, 335–338 (1999)
2. Prescott, J.H., Lipka, S., Baldwin, S., Sheppard Jr., N. F., Maloney, J.M., Coppeta, J., Yomtov, B., Staples, M.A., Santini Jr., J.T.: Chronic, programmed polypeptide delivery from an implanted, multireservoir microchip device. Nat. Biotechnol. **24**, 437–438 (2006)
3. Grayson, A.C.R., Choi, I.S., Tyler, B.M., Wang, P.P., Brem, H., Cima, M.J., Langer, R.: Multi-pulse drug delivery from a resorbable polymeric microchip device. Nat. Mater. **2**, 767–772 (2003)
4. Yam, F., Wu, X., Zhang, Q.: A novel composite membrane for temperature-and pH-responsive permeation. In: Park, K., Mrsny, R.J. (eds.) Controlled Drug Delivery: Designing Technologies for the Future. ACS Symposium Series, pp. 263–272. American Chemical Society, Washington, DC (2000)
5. Zhang, K., Wu, X.Y.: Temperature and pH-responsive polymeric composite membranes for controlled delivery of proteins and peptides. Biomaterials **25**, 5281–5291 (2004)
6. Zhang, K., Wu, X.Y.: Modulated insulin permeation across a glucose-sensitive polymeric composite membrane. J. Control. Release **80**, 169–178 (2002)
7. Chen, J., Chu, M., Koulajian, K., Wu, X.Y., Giacca, A., Sun, Y.: A monolithic polymeric microdevice for pH-responsive drug delivery. Biomed. Microdevices **11**, 1251–1257 (2009)
8. Gordijo, C.R., Shuhendler, A.J., Wu, X.Y.: Glucose-responsive bioinorganic nanohybrid membrane for self-regulated insulin release. Adv. Funct. Mater. **20**, 1404–1412 (2010)
9. Gordijo, C.R., Koulajian, K., Shuhendler, A.J., Bonifacio, L.D., Huang, H.Y., Chiang, S., Ozin, G. A., Giacca, A., Wu, X.Y.: Nanotechnology-enabled

closed loop insulin delivery device: in vitro and in vivo evaluation of glucose-regulated insulin release for diabetes control. Adv. Funct. Mater. **21**, 73–82 (2011)
10. Chu, M.K., Gordijo, C.R., Li, J., Abbasi, A.Z., Giacca, A., Plettenburg, O., Wu, X.Y.: In vivo performance and biocompatibility of a subcutaneous implant for real-time glucose-responsive insulin delivery. Diabetes Technol. Ther. **17**, 255–267 (2015)
11. Nichols, S.P., Koh, A., Storm, W.L., Shin, J.H., Schoenfisch, M.H.: Biocompatible materials for continuous glucose monitoring devices. Chem. Rev. **113**, 2528–2549 (2013)
12. Anderson, J.M., Rodriguez, A., Chang, D.T.: Foreign body reaction to biomaterials. Semin. Immunol. **20**, 86–100 (2008)
13. Parker, J.A., Walboomers, X.F., Von den Hoff, J.W., Maltha, J.C., Jansen, J.A.: The effect of bone anchoring and micro-grooves on the soft tissue reaction to implants. Biomaterials **23**, 3887–3896 (2002)
14. Moshayedi, P., Ng, G., Kwok, J.C.F., Yeo, G.S.H., Bryant, C.E., Fawcett, J.W., Franze, K., Guck, J.: The relationship between glial cell mechanosensitivity and foreign body reactions in the central nervous system. Biomaterials **35**, 3919–3925 (2014)
15. Muskovich, M., Bettinger, C.J.: Biomaterials-based electronics: polymers and interfaces for biology and medicine. Adv. Healthcare Mater. **1**, 248–266 (2012)
16. Bridges, A.W., García, A.J.: Anti-inflammatory polymeric coatings for implantable biomaterials and devices. J. Diabetes Sci. Technol. **2**, 984–994 (2008)
17. Blanco, E., Weinberg, B.D., Stowe, N.T., Anderson, J. M., Gao, J.: Local release of dexamethasone from polymer millirods effectively prevents fibrosis after radiofrequency ablation. J. Biomed. Mater. Res. A **76**, 174–182 (2006)
18. Hetrick, E.M., Prichard, H.L., Klitzman, B., Schoenfisch, M.H.: Reduced foreign body response at nitric oxide-releasing subcutaneous implants. Biomaterials **28**, 4571–4580 (2007)
19. Choi, J., Jang, B.N., Park, B.J., Joung, Y.K., Han, D.K.: Effect of solvent on drug release and a spray-coated matrix of a sirolimus-eluting stent coated with poly(lactic-co-glycolic acid). Langmuir **30**, 10098–10106 (2014)
20. Ren, K., Zhang, M., He, J., Wu, Y., Ni, P.: Preparation of polymeric prodrug paclitaxel-poly(lactic acid)-b-polyisobutylene and its application in coatings of drug eluting stent. ACS Appl. Mater. Interfaces (2015). doi:10.1021/acsami.5b01410
21. Lokanathan, A.R., Zhang, S., Regina, V.R., Cole, M.A., Ogaki, R., Dong, M., Besenbacher, F., Meyer, R.L., Kingshott, P.: Mixed poly (ethylene glycol) and oligo (ethylene glycol) layers on gold as nonfouling surfaces created by backfilling. Biointerphases **6**, 180–188 (2011)
22. Li, P., Poon, Y.F., Li, W., Zhu, H.-Y., Yeap, S.H., Cao, Y., Qi, X., Zhou, C., Lamrani, M., Beuerman, R.W., Kang, E.-T., Mu, Y., Li, C.M., Chang, M.W., Jan Leong, S.S., Chan-Park, M.B.: A polycationic antimicrobial and biocompatible hydrogel with microbe membrane suctioning ability. Nat. Mater. **10**, 149–156 (2011)
23. Glasmästar, K., Larsson, C., Höök, F., Kasemo, B.: Protein adsorption on supported phospholipid bilayers. J. Colloid Interface Sci. **246**, 40–47 (2002)
24. de Vos, P., Faas, M.M., Strand, B., Calafiore, R.: Alginate-based microcapsules for immunoisolation of pancreatic islets. Biomaterials **27**, 5603–5617 (2006)
25. Li, J., Chu, M.K., Gordijo, C.R., Abbasi, A.Z., Chen, K., Adissu, H.A., Lohn, M., Giacca, A., Plettenburg, O., Wu, X.Y.: Microfabricated microporous membranes reduce the host immune response and prolong the functional lifetime of a closed-loop insulin delivery implant in a type 1 diabetic rat model. Biomaterials **47**, 51–61 (2015)
26. Chertok, B., Webber, M.J., Succi, M.D., Langer, R.S.: Drug delivery interfaces in the 21st century: From science fiction ideas to viable technologies. Mol Pharm. **10**, 7 (2013).

Stimulus-Responsive Polymeric Hydrogels

► Smart Hydrogels

Stochastic Assembly

► Self-Assembly for Heterogeneous Integration of Microsystems

Strain Gradient Plasticity Theory

► Plasticity Theory at Small Scales

Structural Color in Animals

Mathias Kolle[1] and Ullrich Steiner[2]
[1]Harvard School of Engineering and Applied Sciences, Cambridge, MA, USA
[2]Department of Physics, Cavendish Laboratories, University of Cambridge, Cambridge, UK

Synonyms

Animal coloration; Biological photonic structures; Biological structural color; Bio-optics; Bio-photonics

Definition

Intense and bright colors result from the interaction of light with periodic micro- and nanostructures that cause color by interference, coherent scattering, or diffraction. These colors are termed structural colors, and structures that cause color by modulation of light are called photonic structures. Photonic structures are usually composed of regular lattices with periodicities on the order of the wavelength of light. Various organisms in nature are known to use intriguingly diverse photonic structures.

Introduction

Structural Colors and Photonic Structures

Structural colors in the animal kingdom have attracted increasing research interest in recent years. Biological organisms offer an enormous variety of periodic micro- and nanostructures that by specific interaction with light provide distinct coloration. This sometimes dynamic reflectivity is tailor-made for the organism's purpose within its natural illumination environment. Intriguing photonic structures have been identified on the wing cases and armors of beetles, the scales of butterflies, the feathers of birds, in the shells of marine animals, or even within the skin of mammals. Nature offers a huge choice of blueprints for novel artificial optical materials and photonic structures. The strongest color contrasts are achieved by a combination of different physical effects, including multilayer interference, diffraction, coherent scattering, and spatially confined absorption [1–3]. A common design concept in natural photonic systems is the complex interplay of structural regularity on the length scale of the wavelength of visible light combined with structural disorder on a larger scale [4]. In many cases, the complex interaction of these hierarchical structures with incident light lead to an outstanding, dynamic coloration, bright reflectivity that is perceivable in a wide angular range, brilliant whiteness, or enhanced transmission [5, 6]. While photonic structures can be made entirely from transparent materials, the incorporation of absorbing pigments deposited under or incorporated into a photonic structure is frequently used in nature to prevent spurious reflections. This improves the contrast leading to an enhancement of the color perceived from the photonic system [7]. Biological micro- and nanostructures that cause structural color are very diverse and often show periodicities on several length scales. This makes the optical characterization of a biological photonic system and the determination of the optical properties of its constituent materials (e.g., refractive index) very challenging. A set of useful techniques for the determination of the complex refractive indices of materials in biological photonic structures has recently been reported and successful attempts have been made to determine the complex refractive index of the organic cuticle material in the scales of butterflies and the armor of beetles [8].

Occurrence and Purpose of Structural Colors in Nature

Structural colors can be found in the feathers and skin of various birds [9–11], in the shells, spines, and scales of marine animals [12, 13], and in the skin of some mammalian species [14]. They are probably most abundant in species of the insect orders *Lepidoptera* and *Choleoptera* that comprise butterflies, moths, and beetles [15–19]. The different purposes of structural colors are as diverse as the organisms that use them [20]. Intense colors with stark contrast to their environment can serve in interspecies interaction including agonistic displays to confuse or scare away potential predators, while structural colors are also applied to induce a cryptic coloration for camouflage. For other animals, they are playing an import role for intraspecies communication such as competition between males of the same species for territory and/or females. Structural colors often provide sexual dichroism between males and females and are believed to have a function in sexual selection. They might also play a role in temperature regulation. While the physics and functioning of photonic structures found in many different organisms are mostly understood, it is frequently very challenging to clearly identify the specific benefit for the animal, which often remains mysterious.

The Physical Effects Underlying Structural Colors

Bright, pure, and intense colors arise from strong reflectivity in narrow spectral bands caused by highly ordered structures including multilayers, surface diffraction gratings, and photonic crystals. While many biological photonic structures are based on multilayer assemblies and two- or three-dimensional photonic crystals [1], diffraction gratings are rare in nature, possibly because they do not display a specific color but give rise to a range of colors depending on light incidence and observation direction [4]. The contrast of structural colors in their environment is often increased by the placement of absorbing elements spatially under or around the photonic structure. Intense blacks are achieved in nature by structurally assisted pigmentation as in the case of the butterfly *Papilio ulysses* [7]. In some avian species the pigments are incorporated directly into the photonic structure [9]. Brilliant white is achieved by multiple, incoherent scattering caused by highly disordered structures of random size, aperiodically arranged on the length scale of the wavelength of visible light [6]. Intense, spatially homogenous, angle-independent colors arise from coherent scattering caused by structures of well-defined size comparable to the wavelength of light, such as air pores of a narrow size distribution dispersed randomly in the volume of a material. This, for example, gives rise to the strong blue of the feathers of several different bird species [11]. High transparency is achieved by graded refractive index surfaces, involving arrays of conical protrusions found in moth eyes [5].

In the following, the structural designs, which give rise to strong coloration, ultrahigh blackness, or brilliant whites are discussed with emphasis on the interplay of different structural, hierarchically assembled elements. Specific natural organisms that apply these particular structural combinations are presented in an exemplary manner, without recounting all the animals known to apply the same or very similar concepts. First, static systems that induce structural colors based on "simple" multilayer elements are discussed before progressing to more and more sophisticated two- and three-dimensional structural arrangements. Attention is paid to combinations of different photonic elements and to the interplay of order and disorder in hierarchical structures on the nano- and microscale, a powerful concept occurring in natural structural colors.

Static Structural Colors in Animals

Multilayer Structures of Varying Complexity in Natural Photonic Systems

Thin film interference and multilayer interference are among the most common phenomena inducing structural color in animals. Light reflected from periodic stacks of multiple planar, optically distinct, transparent layers (often called Bragg mirrors) is colored, provided the thickness of the individual layers is comparable to the wavelength of light in the visible spectrum. The color, its intensity, and purity depend on the refractive index contrast of the multilayer materials, the individual layer thicknesses, and the number of layers in the stack [2]. Furthermore, the perceived color strongly varies with the angle of light incidence and the observation direction. The peak wavelength λ_{max} of light reflected from a multilayer in air composed of two materials with refractive index n_1, n_2 and film thicknesses d_1, d_2 upon incidence at an angle θ is given by the relation

$$m \cdot \lambda = 2 \cdot (n_1 d_1 \cos\theta_1 + n_2 d_2 \cos\theta_2), \qquad (1)$$

where m is a positive integer and the angles θ_1, θ_2 of the light path in the materials are given by Snell's law $n_{air} \sin\theta = n_1 \sin\theta_1 = n_2 \sin\theta_2$. In this case the rays reflected from the family of $n_1 - n_2$-interfaces interfere constructively with each other, if the light has the wavelength λ_{max}. The same holds for light rays reflected from the $n_2 - n_1$-interfaces in the stack. If a second relation given by

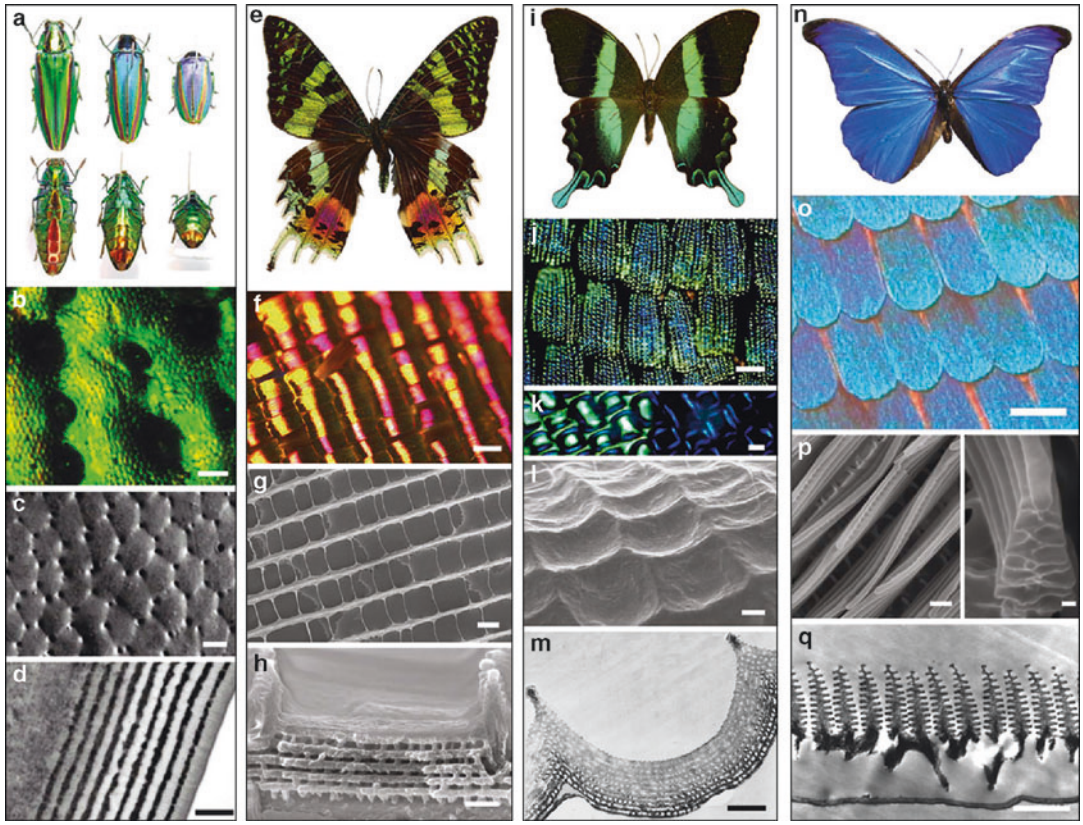
$$\left(m + \frac{1}{2}\right) \cdot \lambda = 2 n_1 d_1 \cos\theta_1 = 2 n_2 d_2 \cos\theta_2 \qquad (2)$$

is also satisfied, the multilayer stack is referred to as an ideal multilayer. This relation signifies the fact that in an ideal multilayer the light rays reflected from the $n_1 - n_2$-interfaces interfere constructively with the rays reflected from the $n_2 - n_1$-interfaces, which is not the case for a "nonideal" multilayer. The reflectance of a nonideal multilayer is therefore lower compared to an ideal multilayer made of the same materials, with an equal overall layer number and equal periodicity $d_1 + d_2$. Most multilayer arrangements found in nature are nonideal in the sense of this classification. In some cases this may even be beneficial for the organism since nonideal multilayers have a narrower bandwidth compared to ideal multilayers, which increases the purity of the reflected color.

The wood-boring Japanese jewel beetle *Chrysochroa fulgidissima* found in the forests of Japan during summer displays a vivid ventral and dorsal coloration (Fig. 1a). This color originates from interference of light in a multilayer stack of about 20 alternating layers (Fig. 1d) with refractive indices of 1.5 and 1.7 incorporated in the epicuticle of the beetle's wing cases [4]. The perceived coloration strongly depends on the orientation of the beetle with respect to the incident light and the observation direction, as expected for a multilayer structure (Fig. 1a). Nevertheless, the beetle's surface does not act like a colored mirror, as expected from a perfectly flat multilayer arrangement. Minute hexagonally arranged holes and indentations are distributed in the multilayer on the 10 μm length scale (Fig. 1c) and the areas bordered by the micro-holes vary in their inclination. On the 100 μm length scale, the surface shows a strongly irregular corrugation (Fig. 1b). This structural irregularity leads to the scattering of light reflected from the multilayer, giving the beetle its characteristic color in a larger angular range around the specular reflection direction thus increasing its visibility.

The scales of the Madagascan moth *Chrysiridia rhipheus* (also called sunset moth because of the two bright yellow-orange-violet spots on the lower wing pair) display intense, beautiful colors ranging from green to red, similar to the jewel beetle's hues (Fig. 1e, f). The different, similarly bright colors result from a multilayer incorporated into the body of the moth's wing scales (Fig. 1h). The multilayer is made from layers of one single material (cuticle) spaced by perpendicular struts, effectively creating a controlled air spacing between the cuticle layers. Including air as the second component in the multilayer stack leads to the maximization of the refractive index contrast, providing intense reflection with a much smaller number of layers compared to the Japanese jewel beetle, where the refractive index contrast is much lower. The air spacing varies in the differently colored regions of the moth's wing with the smallest spacing found in the green areas and the largest spacing seen in the red wing spots. The multilayered scales show a strong curvature along their long axis. This curvature superimposed onto the photonic multilayer structure leads to a high visibility of the colors from various directions and also introduces multiple reflection effects between adjacent scales, leading to an increased purity of the reflected color and to interesting polarization effects [21]. Overall, caused by the interplay of the highly reflective air-cuticle multilayer and the pronounced scale curvature, the moth wing displays bright colors with a velvety, textile-like shimmer.

Air-cuticle multilayers of even more complex shapes can be found on the wing scales of butterflies of the genus *Papilio*. The South-East Asian Emerald Swallowtail *Papilio palinurus* and the Green Swallowtail *Papilio blumei* (Fig. 1i) display bright green spots on their wings resulting from a concavely shaped multilayer architecture on the individual scales (Fig. 1l, m). The vivid green originates from a superposition of blue and yellow reflected from the edges and the centers of the multilayer concavities, respectively (Fig. 1j, k) [22]. Different colors result from distinct regions of the concavely shaped multilayer due to the spatially varying angle of light incidence on the edges and in the center of the concavities. Light reflected from the edges can undergo multiple reflections within a single concavity, inducing a change in polarization. The butterfly *Papilio ulysses* displays concavely shaped multilayer structures on its scales that have less curvature and smaller cuticle and air gap thicknesses, resulting



Structural Color in Animals, Fig. 1 Natural photonic systems based on multilayer arrangements. (**a**) The wing cases and the ventral side of the Japanese Jewel beetle *Chrysochroa fulgidissima* display bright and iridescent colors that show a strong angular color variation (**b**, **c**). The surface of the wing cases shows irregular corrugations on the 100 μm scale (**b**, scale bar ~100 μm) and more regular indentations on the 10 μm scale (**c**, scale bar 10 μm). (**d**) Cross-sectional transmission electron micrographs reveal the stack of alternating high and low refractive index layers that causes multilayer interference, resulting in the beetle's iridescent color, scale bar 400 nm (Pictures (**a**–**d**) reproduced with permission of S. Kinoshita and IOP Publishing © 2008 Kinoshita et al. [1]). (**e**) The Madagascan moth, *Chrysiridia rhipheus*, displays a range of bright shimmering colors on its wings. (**f**) Different wing regions are covered by patterns of colorful scales reflecting green, yellow, red, or violet light. The scales are highly curved so that the observer only perceives light reflected from part of the scales, which creates the impression of texture caused by the juxtaposition of bright and dark regions, scale bar 200 μm. (**g**) The scales are spanned by parallel micro-ribbed ridges, scale bar 2 μm. (**h**) Cuticle layers of well-defined thickness that are spaced by small struts extend between and under the ridges, forming a regular multilayer arrangement with air as the low refractive index material, scale bar 500 nm. (**i**) The butterfly *Papilio blumei* displays lucid green stripes on its upper and lower wing pairs. (**j**) At higher magnification, the green scales show regions of distinct blue or yellow color, scale bar 100 μm. (**k**) The yellow color from the centers of concavely shaped surface corrugations disappears when the scales are imaged between crossed polarizers while the blue from the edges of the concave shapes persists due to polarization rotation upon reflection, scale bar 5 μm. (**l**) Concave surface corrugations on the scales, scale bar 2 μm. (**m**) A cross-section through one of the concavities reveals cuticle layers of well-defined thickness and regular spacing forming a multilayer reflector, scale bar 1 μm (Figure (**m**) reproduced with permission of P. Vukusic and Springer Science + Business Media © 2009 Vukusic [19]). (**n**) The South American butterfly *Morpho rhetenor*. (**o**) Bright blue scales cover the wing membrane like tiles on a roof top, scale bar 100 μm. (**p**) Top and cross-sectional view of the ridges running along each scale, scale bars 500 and 100 nm. (**q**) A cross-section through a scale exposes the intricate design of the ridges, scale bar 1 μm. The horizontal extensions on each ridge act as a multilayer reflector tuned for the blue spectral range with air as the low refractive index medium (Figures (**o** and **q**) reproduced with permission of P. Vukusic and Springer Science + Business Media © 2009 Vukusic [19])

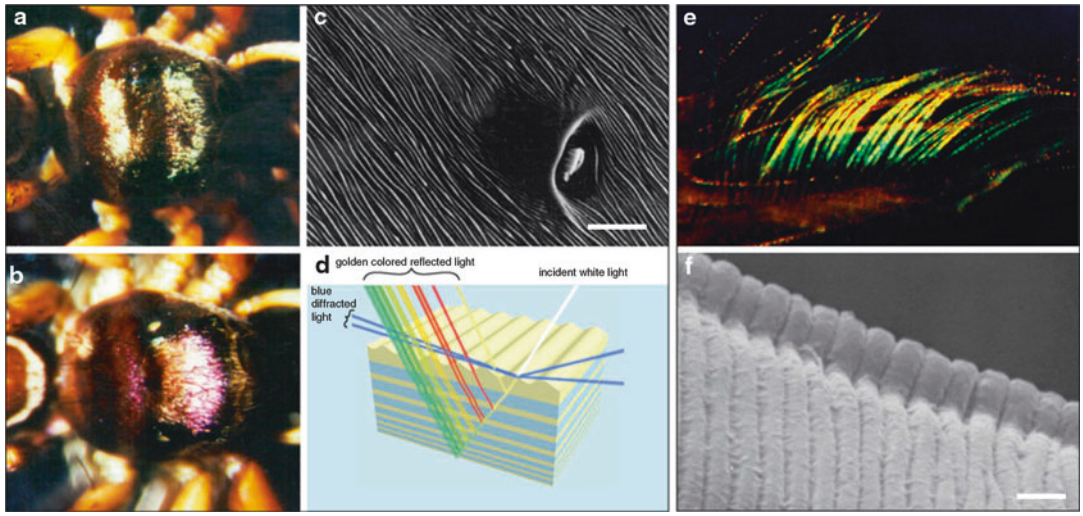
in a lucid blue color. While in the case of *Papilio palinurus* and *Papilio blumei* the higher local concavity wall curvature results in the juxtaposition of yellow and blue reflected from different regions forming green (a color that might have a purpose in camouflage), the shallower concavities of *Papilio ulysses* do not produce the same effect but rather diffuse the reflected light into a wide angular range for higher visibility [19].

The possibly most sophisticated and intensely studied multilayer structure giving rise to bright animal coloration is found on the wing scales of butterflies belonging to the genus *Morpho*. These butterflies that are native to South and Central America are known for their intense, widely visible coloration. Among them, *Morpho rhetenor* is the most striking example of vivid bright blue structural color (Fig. 1n, o). The intriguing hue of its wings results from the interference of light caused by ridge structures (Fig. 1p) on the wing scales that have a Christmas-tree like cross-section (Fig. 1p, q). The horizontal periodic protrusions of the ridges cause "quasi"-multilayer interference [4, 5]. The term "quasi" is used in this context, because light not only interferes when reflected from protrusions of the same ridge but also when reflected from protrusions on different adjacent ridges. As for the moth *Chrysiridia rhipheus* and the butterflies of the genus *Papilio*, the incorporation of air layers into the multilayer structure leads to a large refractive index contrast and consequently to a bright reflection in a wide spectral band. The ridges are spaced by $\lesssim 1$ μm on average, which in the past lead to the assumption that they also act as a diffraction grating. However, local variation between the ridges in height and distance and in the orientation of entire scales confines the diffraction to individual ridges and inhibits the manifestation of a pronounced diffraction grating effect in reflection [4]. Nevertheless, this randomization seems to have a beneficial side effect: despite its origin from multilayer reflection, the blue color of the *Morpho rhetenor* is very illumination and observation angle insensitive and perceivable from all points in space above the wing plane, only changing to violet at very high angles of light incidence or observation.

Diffractive Elements

Diffraction from surface gratings is believed to be rare in nature. This might be due to the fact that diffraction gratings with periodicities of 400 nm-2 μm need long-range order on the 60 μm scale (the spatial coherence of sun light) in order to be efficient, a criterium that is hard to meet in the natural context where irregularity is usually predominant at this length scale. During the past decade some animals, mostly invertebrates, have been shown to employ diffraction gratings, however [12]. As opposed to multilayer reflectors, which reflect light in a relatively narrow spectral band, diffraction gratings split incident light into its spectral components and redirect each color into a different direction. Consequently, diffraction gratings are less suitable for providing spectrally well-defined colors in a wide angular range (which is better achieved with a multilayer reflector combined with some irregularity on the micron-scale). While diffraction gratings made from transparent materials efficiently create vivid colors in transmission, they have to be made from (or backed by) reflective materials to produce strong colors in reflection, which seems to be less efficient for the organism.

Periodic grating-like surface structures on an underlying chirped broadband multilayer reflector have been found on the torso of spiders [23]. In a chirped multilayer the thicknesses of the individual layers decrease or increase gradually from the top to the bottom of the stack. Constructive interference occurs for the reflection of light of a particular color/wavelength range at a specific depth in the stack (Fig. 2d) where the layer thicknesses fulfill the condition discussed above (Eq. 1) thereby leading to a broadband reflection usually resulting in a silver or golden color. In the case of the spider *Cosmophasis thalassina* (Fig. 2a, b), the striations (Fig. 2c) mainly serve the purpose to disperse blue light before it interacts with the underlying reflector (Fig. 2d), thereby biasing the color of the spider's body toward a golden appearance, instead of the more silvery shine which it would have without the striations. Fourier analysis shows that there is no pronounced long-range order in these striations and



Structural Color in Animals, Fig. 2 Natural photonic systems based on diffraction. (**a**) The torso of the spider *Cosmophasis thalassina* shows two broad metallic golden-greenish stripes in air. (**b**) The color changes to purple-silvery when immersed in water. (**c**) The structure on the spider torso that causes these colors consists of a combination of surface striations superimposed on a chirped multilayer reflector (not shown here), scale bar 5 µm. (**d**) A schematic of the interaction of light with the spider's photonic structure. Blue light is very efficiently scattered by the striations while light of higher wavelength interferes with the multilayer stack. Red is specularly reflected in the top section of the chirped mirror where the layers are thicker while green is reflected by the thinner layers further down the stack. (**e**) The hairs of the first antenna of the seed shrimp *Azygocypridina lowryi* show strong iridescence caused by grating diffraction. (**f**) The diffraction grating on a single *setule* (hair) of the antenna, scale bar 1 µm (Images (**d** modified) reproduced with permission of A. R. Parker, IOP Publishing © 2003 Parker and Hegedus [23] and The Royal Society © 2005 Parker [13])

consequently the striations do not create strong grating diffraction. The striations are likely to disperse light reflected from the multilayer reflector into a wider angular range. This represents a good example of the interplay of different structures with varying length scales to achieve a specific optical response.

The use of periodic diffraction gratings as surface structures seems particularly beneficial on narrow cylindrical geometries where the implementation of multilayers might be impractical or impossible. The thin hairs (*setae*) of some Crustacean species including the antenna of the male seed shrimp, *Azygocypridina lowryi* (Fig. 2e), display vibrant colors resulting from periodic surface structures [12, 13]. In the antenna of *Azygocypridina lowryi* the grating is formed by regular undulations of the hair thickness with a periodicity of 600–700 nm (Fig. 2f).

A curious structure causing diffraction with a reverse angular color sequence was recently found on the wings of the male butterfly *Pierella luna* [24]. The ends of the scales in the central portions of this butterfly's forewings are curled upward, enabling regularly arranged, on average 440 nm spaced cross-ribs to act as an upward-directed diffraction grating in transmission, resulting in this intriguing phenomenon of "inverse" diffraction. As opposed to a conventional diffraction grating, these structures make the central forewing regions appear red for small angles of observation (measured from the surface normal), changing to yellow, green, and finally blue as the observation angle is increased.

Two-Dimensional Photonic Crystals

Two-dimensional photonic crystals that cause structural colors have been found in marine animals [13, 25, 26], birds [9, 10], and mammals [14]. This part of the review focuses on two-dimensional photonic systems found in nature that induce the striking color in some

animals. The *setae* of the polychaete worm *Pherusa* sp. (Fig. 3a, b) show a vivid play of colors originating from the periodic two-dimensional hexagonal arrangement of cylindrical channels (Fig. 3c, d) with a well-defined lattice constant within a single *seta* [26]. This system is strikingly similar to artificial photonic crystal fibers. A very similar arrangement was found earlier in the spines of the sea mouse *Aphrodita* sp. [25] (Fig. 3e–h).

The feathers of kingfishers, peacocks, ducks, pigeons, and trogons among various other birds display strong blue and green colors that reportedly have attracted the interest of scientists for more than 300 years [1, 4, 18]. The bright blues and greens result from well-ordered melanin granules in a dense keratin matrix [9] or a spongy keratin network with regular air pores [11] of varying dimensions and structural complexity within the birds' feather barbules. These structures have been classified and investigated in great detail in the last three decades [9]. The order of the layers of the solid or hollow melanin granules or the pores in the keratin matrix varies from species to species. Consequently, the physical effect causing the bright color was repeatedly identified as incoherent scattering for the less ordered randomly distributed structural elements with well-defined sizes or as coherent scattering (for instance, multilayer interference) in structures of quasi-ordered arrangements with higher spatial order. In many bird species not only the feather barbules but also the skin on different body parts show strong color. These colors are based on coherent scattering caused by two-dimensional photonic crystal structures, consisting of regular arrays of collagen fibers in the dermis of the birds (Fig. 3i–l). The dimensions of the collagen fibers are very well defined with a narrow distribution in fiber diameter for each species, while the extent of spatial order varies from species to species (compare Fig. 3j, l).

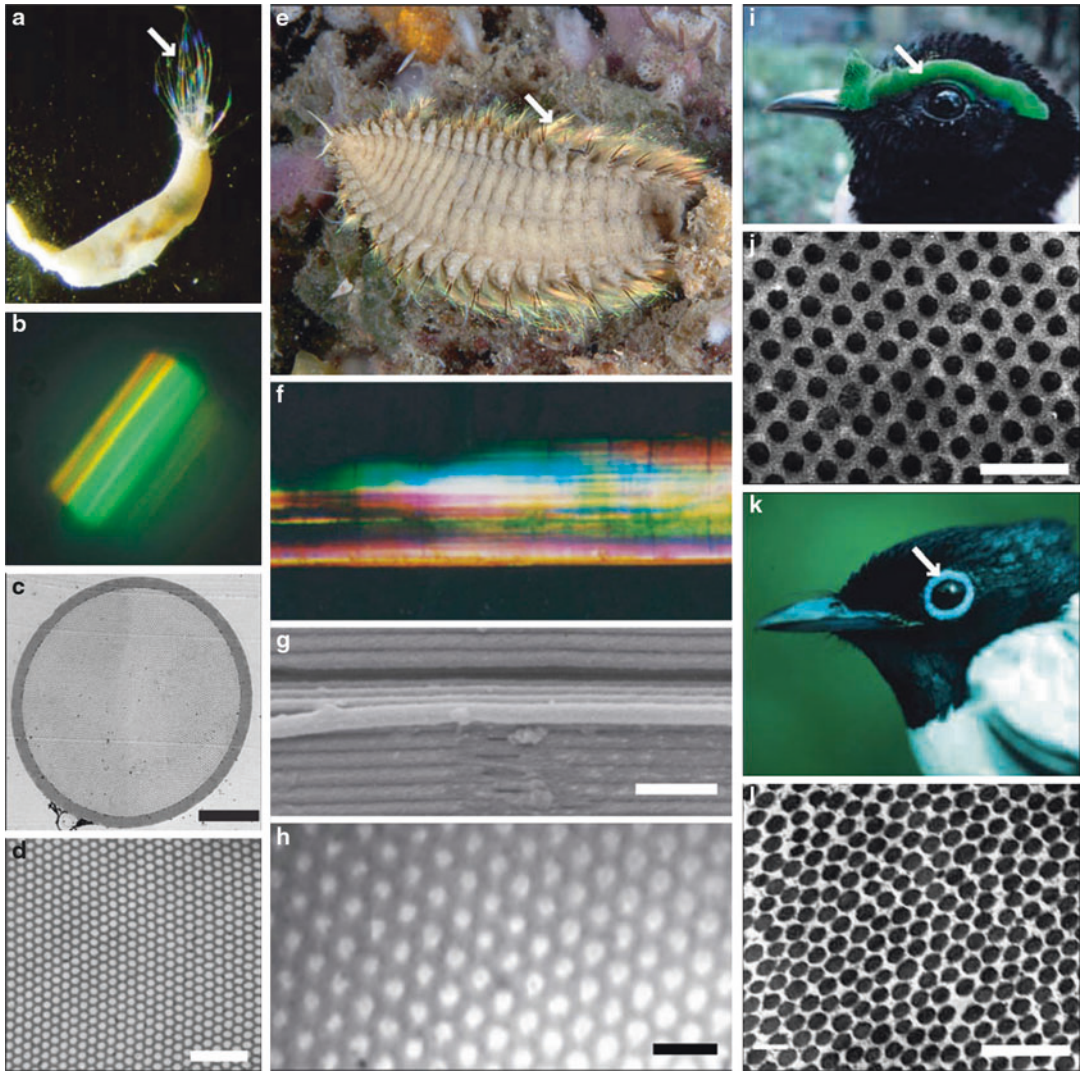
Three-Dimensional Photonic Crystals

The first discovered biological three-dimensional photonic crystal is a solid, hexagonally close-packed array of transparent spheres of 250 nm diameter in a matrix of hydrated chitinous material found inside the scales of the Australian weevil *Pachyrhynchus argus* [28] (Fig. 4a–c). Similar to the ordered arrangement of silica spheres in precious opal, this structure gives the beetle a metallic coloration visible in any direction. Recently, a similar structure was also observed in the Asian longhorn beetle *Pseudomyagrus waterhousei* (Fig. 4d–f).

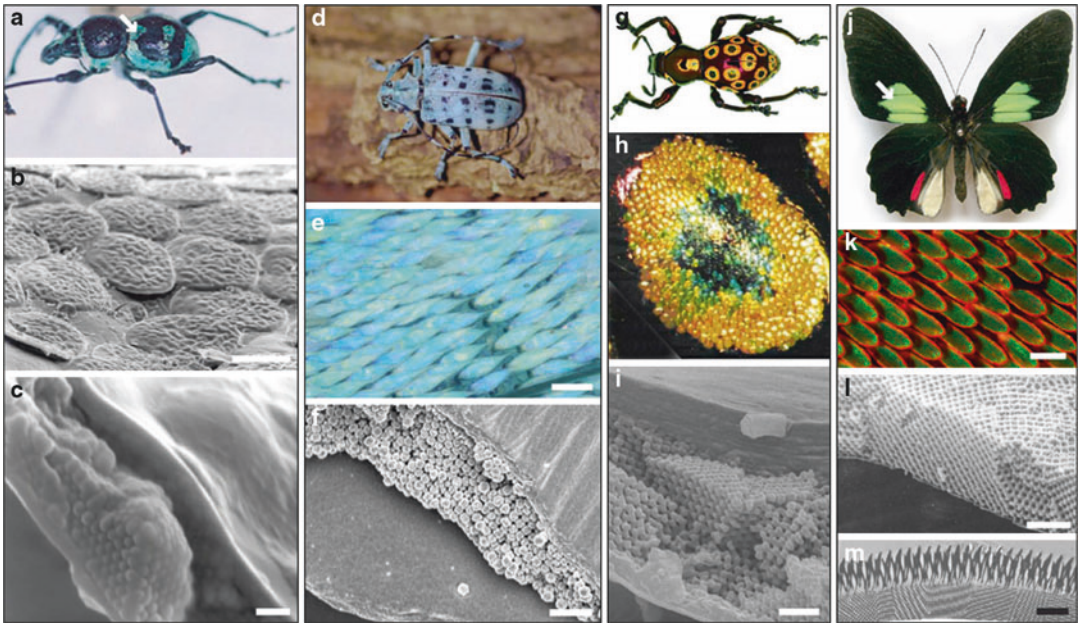
Inverse opal photonic structures have been identified in various butterfly and beetle species. The beetle *Pachyrrhynchus congestus pavonius* [30] displays vivid patches of orange color on its body (Fig. 4g, h) that originate from the interference of light within the beetle scales, which consist of a multilayered cortex surrounding an inverse opal photonic crystal (Fig. 4i). This is an example of the superposition of different structural geometries, in this case a multilayer mirror and a three-dimensional photonic crystal. The light reflected from a regular multilayer mirror results in an iridescent color while the interference of light with a polycrystalline three-dimensional photonic crystal usually leads to a uniform color over wide angular ranges. However, the precise influence of each structural component on the overall appearance of this particular beetle, *Pachyrrhynchus congestus pavonius*, remains to be investigated in detail. The wings of the butterfly *Parides sesostris* display patches of bright green angle-independent color (Fig. 4j, k), resulting from the interaction of light with a polycrystalline, three-dimensional photonic crystal structure that is buried under the superficial ridging in the body of the butterfly's wing scales (Fig. 4l, m). The individual crystallites were shown to have cubic symmetry and recent research suggests that they consist in fact of a bi-continuous regular gyroid network [31]. The different orientations of the photonic crystal domains and the light scattering induced by the superposed ridge structure ensure that the reflected color is independent of observation angle.

Structurally Assisted Blackness and Brilliant Whiteness

The natural photonic structures presented so far serve the organisms to display a distinct coloration. The contrast of a color is often enhanced in nature by placing it against a black background



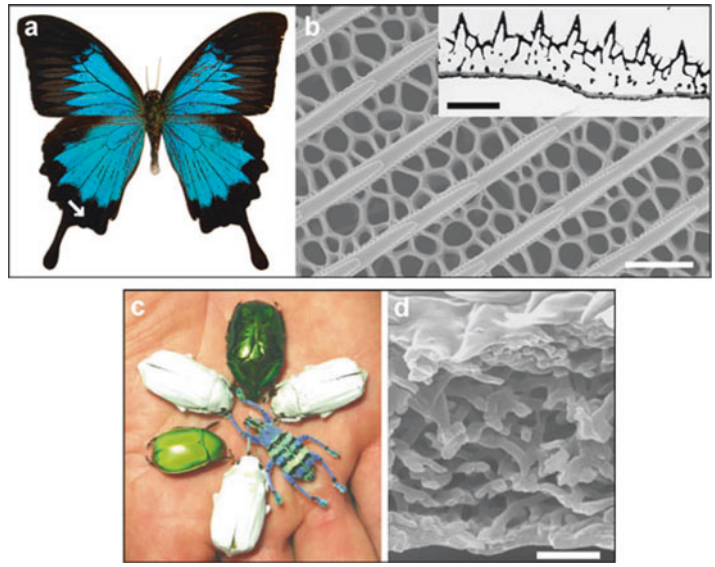
Structural Color in Animals, Fig. 3 Structural colors based on two-dimensional photonic crystals. (**a**) The *setae* of the polychaete worm *Pherusa* sp. show a remarkable structural color (*white arrow*). (**b**) Micrograph of a *seta* of the worm. (**c**) A cross-sectional transmission electron micrograph of a *seta*, scale bar 5 μm. (**d**) High-magnification image of the regular structure in the *seta*, scale bar 2 μm (Image (**a**) reproduced with permission of A. R. Parker, 1995. Images (**b**–**d**) reproduced with permission of P. Vukusic and The American Physical Society © 2009 Trzeciak and Vukusic [26]). (**e**) The colorful spines (*white arrow*) of the sea mouse *Aphrodita* sp. (Image courtesy of D. Harasti). (**f**) Optical micrograph of one of the colorful spines. (**g**, **h**) Side view and cross-section of the tubular structures in the spine, scale bars 2 and 1 μm (Images (**f**–**h**) reproduced with permission of A. R. Parker and The Royal Society © 2004 Parker [27]). (**i**) The bright green wattle (*white arrow*) of the Madagascan bird *Philepitta castanea* (velvet asity). (**j**) A transmission electron micrograph of a cross-section of the nanostructured arrays of dermal collagen fibers responsible for the wattle's green color, scale bar 500 nm. (**k**) The blue eye spot (*white arrow*) of the Madagascan bird *Terpsiphone mutata* (Madagascar paradise flycatcher). (**l**) A cross-section of the collagen fiber array that causes the blue color around the bird's eye, scale bar 500 nm (Images (**i**–**l**) adopted from Prum and Torres [10] with permission of R. O. Prum and T. Schulenberg (**k**))



Structural Color in Animals, Fig. 4 Structural colors based on three-dimensional photonic crystals. (**a**) Dorsal view of the Australian weevil *Pachyrhynchus argus* showing the green metallic spots (*white arrow*) on its back side. (**b**) Scanning electron micrograph of the beetles scales, scale bar 50 μm. (**c**) Cross-section of a scale revealing the opaline structure responsible for the beetle's metallic coloration, scale bar 1 μm (Images (**a–c**) reproduced with permission of A.R. Parker and The Royal Society © 2004 Parker [27]). (**d**) The South-East Asian bright blue longhorn beetle *Pseudomyagrus waterhousei*. (**e**) Micrograph of the iridescent drop-shaped scales on the beetle's body, scale bar 50 μm. (**f**) Scanning electron micrograph of a cross-section through a single scale revealing the origin of the iridescent blue color, regularly sized, closely packed spherules forming a direct opal, scale bar 1 μm (Images (**d–f**) reproduced with permission of J. P. Vigneron and The American Physical Society © 2011 Simonis and Vigneron [29]). (**g**) The beetle *Pachyrrhynchus congestus pavonius*. (**h**) Micrograph of the highly conspicuous annular spots on the beetle's thorax. The scales of different colors are clearly visible. (**i**) Scanning electron micrograph of the cross-section of a scale, revealing the photonic structure consisting of a combination of a multilayer reflector on top of a three-dimensional face-centered cubic photonic crystal, scale bar 1 μm (Images (**g–i**) reproduced with permission of J. P. Vigneron and The American Physical Society © 2007 Welch et al. [30]). (**j**) The South American butterfly *Parides sesostris*, commonly called Emerald-patched Cattleheart due to the bright green patches on its upper wing pairs (*white arrow*). (**k**) Scales in the green areas of the butterfly's wings, scale bar 100 μm. (**l**) SEM image of a cross-section of the three-dimensional photonic structure found within a single scale, scale bar 1 μm. (**m**) Transmission electron micrograph of the cross-section of a scale showing the superficial ridging and the domains of the underlying of photonic crystal, scale bar 2 μm (Images (**j, l, m**) reproduced with permission of P. Vukusic and Springer Science + Business Media B.V © 2009 Vukusic [19]. Image (**k**) courtesy of M. Doolittle)

and color purity is ensured by absorption of undesired spectral light components in regions beneath or around the color creating photonic elements. Consequently, optimized absorption plays an important role for natural structural colors. The scales of various butterflies that display structural colors are supported on a surface containing melanin pigments. The pigments absorb the light that passes through photonic structures of the scales, suppressing spurious reflection and consequently preventing the desaturation of the reflected color.

Very strong blackness can be achieved by an optimized interplay of pigment absorption and structure. A particularly good example is the butterfly *Papilio ulysses* [7]. The bright blue patches



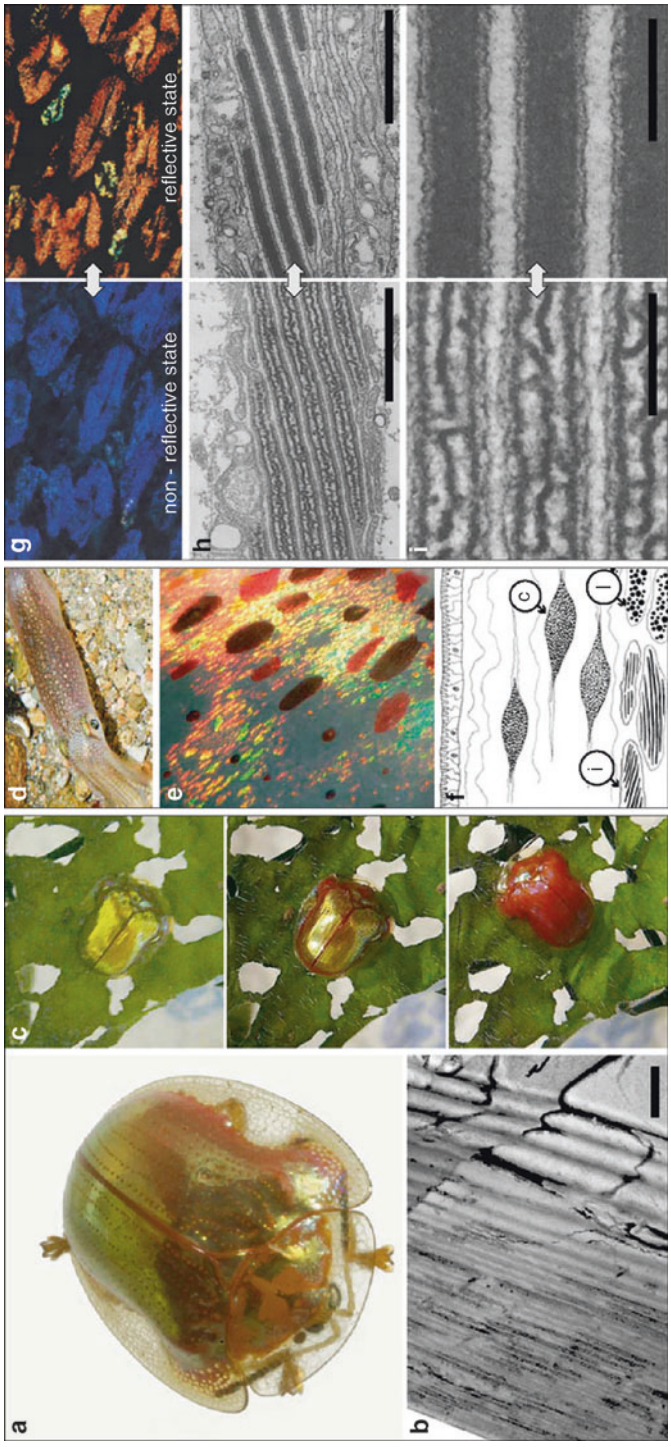
Structural Color in Animals, Fig. 5 Deep blackness and brilliant whiteness. (**a**) The butterfly *Papilio ulysses* shows regions of deep black (*white arrow*) surrounding the bright blue spots on its wings. (**b**) The scales in the black regions carry cuticle microstructures that trap light. The inset shows a cross-section of a scale revealing the dense network of pigment-loaded cuticle, scale bars 2 μm. (**c**) The brilliant white beetle *Cyphochilus* spp. compared to other beetles. (**d**) The intense white results from incoherent diffuse scattering from an interconnected network of filaments within the body of each beetle scale, scale bar 1 μm (Inset in (**b**), images (**c**) and (**d**) reproduced with permission of P. Vukusic, The Royal Society © 2004 Vukusic et al. [7] and The Optical Society of America © 2009 Hallam et al. [32])

on the butterfly's wings are surrounded by very dark scales that show strong light absorption induced by melanin-containing microstructures (Fig. 5a). These pigment-filled structures act as light traps that are tailored to prevent light from escaping once it enters the scales (Fig. 5b). This maximizes the interaction with the pigment, resulting in the extremely low reflectivity of the butterfly's black scales.

Materials with a high structural irregularity, a random distribution, orientation, and size of individual transparent or highly reflective scattering elements on the scale of the wavelength of light are known to equally scatter light of any color into a broad angular range, thereby creating an intense white appearance. An extra-ordinary example of brilliant structural whiteness has been discovered in the case of the *Cyphochilus* spp. beetle [6]. The insect's intense white (Fig. 5c) is based on the incoherent scattering of light from a random network of about 250 nm thick interconnected filaments that are randomly distributed within the ultrathin beetle scales (Fig. 5d). Despite the fact that the scales are only about 5 μm thick, an optimized void fraction, high aperiodicity, and the large refractive index contrast of about 0.56 ensure that light of all colors is very efficiently scattered in all directions.

Dynamically Variable Structural Colors in Animals

The most interesting structural color systems in nature are the ones that show aspects of intentional color tuning, like the photonic structure employed by the beetle *Charidotella egregia* (Fig. 6a). In situations of distress, for instance, caused by predator attacks [33], the beetle can change its appearance from a bright golden shine to a striking red color (Fig. 6c) within 2 min. The structure inducing this phenomenon is a multilayer buried in the exocuticle of the beetle's transparent armor (Fig. 6b). The multilayer is chirped



Structural Color in Animals, Fig. 6 Dynamically tunable structural color. (**a**) The golden beetle *Charidotella egregia*. (**b**) Transmission electron micrograph of the chirped multilayer reflector in the outer layer of the beetle's armor, showing the density variation of irregularly distributed porous and solid areas within the stack, scale bar 1 μm. (**c**) When threatened the beetle can change its golden color to a striking red within less than 2 min (Images (**a**–**c**)

which results in reflection of light in a broad wavelength range, leading to the golden appearance of the beetle. Internally, the multilayer structure shows a random distribution of porous regions and channels within the layer planes and perpendicular to the layers. In the default state, these irregularly distributed voids are filled with liquid leading to a homogenous refractive index within each layer thereby suppressing scattering and allowing interference of light in the chirped multilayer, which results in the golden metallic color. Beneath the multilayer resides a layer of bright red pigment. When under a potential threat the beetle resorts to an aposematic protection mechanism. It can withdraw the liquid from the porous regions in the multilayer which strongly increases the scattering thereby rendering the layer stack translucent and revealing the bright red color of the underlying pigments, which serves as a warning signal.

The impressive camouflage and signaling capacities of cephalopods (squid, cuttlefish, and octopus) result from an intricate interplay of mainly three different functional elements within the skin of the animals [35]. Chromatophores are small pigment-filled organs that can be stretched and compressed by radially attached muscle strands. The squid *Logilo pealeii* (Fig. 6d) can vary the size of its chromatophores from 1.5 mm in the expanded state to about 0.1 mm when retracted [34]. A distinct layer of iridophores buried in the skin beneath the chromatophores provide spectrally selective reflection (Fig. 6e, f). The iridophores are colorless cells, which contain stacks of thin platelets (iridosomes) that reflect light by multilayer interference. Among the iridophores, cuttlefish and octopus employ additional structural elements, so-called leucophores (Fig. 6f). These cells are made up of disordered spherical assemblies with particles ranging from 250 to 1,250 nm in diameter (leucosomes). They induce diffuse broadband scattering and are responsible for the white patterns on cuttlefish and octopus. While the leucophores are passive elements, cephalopods have a high physiological control over chromatophores and iridophores [35]. The chromatophores can be expanded or retracted within a fraction of a second. In the expanded state they display the pigment color (red, yellow/orange, or brown/black depending on the species) and hide the underlying iridophores and leucophores from interfering with the incident light. When the chromatophores are retracted iridophores and leucophores are revealed and determine the reflected color. Squids are able to change the iridescence of the iridophores with shifts of over 100 nm in the reflected wavelengths observed for some species. This reflectance change progresses much slower than the actuation of the chromatophores and can take several seconds to minutes. The change in iridophore reflection can result from two different processes. The platelets in the iridophores, which are made of a protein called reflectin, can change their refractive index by a change in state of the protein conformation [36] (Fig. 6g–i). Furthermore, the thickness of the plates can change to tune the reflection. In addition to the described passive optical elements, some cephalopod species make use of structurally very complex light-emitting photophores that can be highly directional in their emission. The emission of photophores is often enhanced by the incorporation of multilayer back-reflectors. More

Structural Color in Animals, Fig. 6 (continued) reproduced with permission of J.P. Vigneron and The American Physical Society © 2007 Vigneron et al. [33]). (**d**) The squid *Logilo pealeii*. (**e**) Microscope images of the squid's skin show the brown and reddish pigmented chromatophores and the underlying iridescent iridophores shimmering in colors from green to orange. (**f**) A sketch of the arrangement of chromatophores (**c**), iridophores (**i**), and leucophores (**l**) in the skin of cephalopods (Images (**d**–**f**) reproduced with permission of L. Mäthger, Springer Science + Business Media © 2007 Mäthger and Hanlon [34] and The Royal Society © 2009 Mäthger et al. [35]). (**g**) Optical microscope images, and (**h**, **i**) transmission electron micrographs of active iridophores of the squid *Lolliguncula brevis* reveal the variation in reflection and the ultrastructural changes in the iridophore platelets when switching between non-reflective (*left*) and reflective (*right*) states, scale bars 1 μm (**h**), 250 nm (**i**) (Images (**g**–**i**) reproduced with permission of R. T. Hanlon and Springer Science + Business Media © 1990 Cooper et al. [36])

complex photophores contain filters and light guides that help to channel and direct the emitted light. In summary, cephalopods use an extensive repertoire of actively tunable optical elements that rely on a range of physical effects in order to manipulate incident light, including absorption, light interference, bioluminescence, and scattering, which makes them the uncontested masters of color, light manipulation, and camouflage in nature.

Conclusion

In the course of evolution, various organisms in nature have developed a huge variety of photonic systems that by interference, diffraction, coherent, and incoherent scattering cause distinct color. While the bearers of such photonic structures come from very different animal orders, common design principles can be identified across the distinct taxonomic groups. Regular periodic multilayer arrangements and two- or three-dimensional photonic crystals build the base for strong color, while disordered arrays of particles and filaments cause bright whiteness. Pigment-loaded structures with corrugations on different length scales act as efficient light traps that render surfaces deep black. Strong blackness and brilliant whiteness are frequently employed in nature to provide contrast for intriguing patterns of color on wings and bodies of insects, scales and shells of marine animals, and feathers of birds. The coordinated interplay of regularity and irregularity on different length scales plays an important role in the function of many natural photonic systems [4, 9]. Well-defined, structural regularity and periodicity on the submicron scale ensures the reflection of strong bright colors, while irregularity on the scale of several microns inducing random scattering often mediates color stability and conspicuousness in a wide angular range.

By looking at organisms in nature new insight and knowledge can be gained for the design of materials that show specific and efficient interaction with light. This fact is widely acknowledged in the scientific community [15, 27]. Researchers show increased interest in the development of bio-inspired photonic systems. The current techniques and tools used in the industry, research, and everyday life for light harvesting, optical signaling, data transfer, and processing might soon benefit from a better understanding of the composition and functioning of biological photonic structures.

Cross-References

- ▶ Biomimetics of Optical Nanostructures
- ▶ Moth-Eye Antireflective Structures
- ▶ Nanostructures for Coloration (Organisms Other than Animals)
- ▶ Nanostructures for Photonics

References

1. Kinoshita, S., Yoshioka, S., Miyazaki, J.: Physics of structural colors. Rep. Prog. Phys. **71**, 076401 (2008)
2. Land, M.F.: The physics and biology of animal reflectors. Prog. Biophys. Mol. Biol. **24**, 75–106 (1972)
3. Parker, A.R., Martini, N.: Structural colour in animals – simple to complex optics. Opt. Laser Technol. **38**, 315–322 (2006)
4. Kinoshita, S., Yoshioka, S.: Structural colors in nature: the role of regularity and irregularity in the structure. Chemphyschem **6**, 1442–1459 (2005)
5. Vukusic, P., Sambles, J.R.: Photonic structures in biology. Nature **424**, 852–855 (2003)
6. Vukusic, P., Hallam, B., Noyes, J.: Brilliant whiteness in ultrathin beetle scales. Science **315**, 348 (2007)
7. Vukusic, P., Sambles, J.R., Lawrence, C.R.: Structurally assisted blackness in butterfly scales. Proc. R. Soc. B **271**, S237–S239 (2004)
8. Vukusic, P., Stavenga, D.G.: Physical methods for investigating structural colours in biological systems. J. R. Soc. Interface **6**, S133–S148 (2009)
9. Kinoshita, S.: Structural Colors in the Realm of Nature. World Scientific, Singapore (2008)
10. Prum, R.O., Torres, R.: Structural colouration of avian skin: convergent evolution of coherently scattering dermal collagen arrays. J. Exp. Biol. **206**, 2409–2429 (2003)
11. Prum, R.O., Torres, R.H., Williamson, S., Dyck, J.: Coherent light scattering by blue feather barbs. Nature **396**, 28–29 (1998)
12. Parker, A.R.: 515 million years of structural colour. J. Opt. A Pure Appl. Opt. **2**, R15–R28 (2000)
13. Parker, A.R.: A geological history of reflecting optics. J. R. Soc. Interface **2**, 1–17 (2005)
14. Prum, R.O., Torres, R.H.: Structural colouration of mammalian skin: convergent evolution of coherently scattering dermal collagen arrays. J. Exp. Biol. **207**, 2157–2172 (2004)

15. Biró, L.P., Vigneron, J.P.: Photonic nanoarchitectures in butterflies and beetles: valuable sources for bioinspiration. Laser Photonics Rev. **5**, 27–51 (2011)
16. Ghiradella, H.: Shining armor: structural colors in insects. Opt. Photonics News **10**, 46–48 (1999)
17. Seago, A.E., Brady, P., Vigneron, J.P., Schultz, T.D.: Gold bugs and beyond: a review of iridescence and structural colour mechanisms in beetles (*Coleoptera*). J. R. Soc. Interface **6**, S165–S184 (2009)
18. Srinivasarao, M.: Nano-optics in the biological world: beetles, butterflies, birds, and moths. Chem. Rev. **99**, 1935–1962 (1999)
19. Vukusic, P.: Advanced photonic systems on the wing-scales of *Lepidoptera*. In: Gorb, S.N. (ed.) Functional Surfaces in Biology: Little Structures with Big Effects. Springer, Dordrecht (2009)
20. Doucet, S., Meadows, M.: Iridescence: a functional perspective. J. R. Soc. Interface **6**, S115–S132 (2009)
21. Yoshioka, S., Kinoshita, S.: Polarization-sensitive color mixing in the wing of the Madagascan sunset moth. Opt. Express **15**, 2691–2701 (2007)
22. Vukusic, P., Sambles, J.R., Lawrence, C.R.: Structural colour: colour mixing in wing scales of a butterfly. Nature **404**, 457 (2000)
23. Parker, A.R., Hegedus, Z.: Diffractive optics in spiders. J. Opt. A Pure Appl. Opt. **5**, S111–S116 (2003)
24. Vigneron, J.P., et al.: Reverse color sequence in the diffraction of white light by the wing of the male butterfly *Pierella luna* (Nymphalidae: Satyrinae). Phys. Rev. E **82**, 021903 (2010)
25. Parker, A.R., McPhedran, R.C., McKenzie, D.R., Botten, L.C., Nicorovici, N.P.: Aphrodites iridescence. Nature **409**, 36–37 (2001)
26. Trzeciak, T.M., Vukusic, P.: Photonic crystal fiber in the polychaete worm *Pherusa* sp. Phys. Rev. E **80**, 061908 (2009)
27. Parker, A.R.: A vision for natural photonics. Philos. Trans. R. Soc. A **362**, 2709–2720 (2004)
28. Parker, A.R., Welch, V.L., Driver, D., Martini, N.: Structural colour: opal analogue discovered in a weevil. Nature **426**, 786–787 (2003)
29. Simonis, P., Vigneron, J.P.: Structural color produced by a three-dimensional photonic poly-crystal in the scales of a longhorn beetle: *Pseudomyagrus waterhousei* (Coleoptera: Cerambicidae). Phys. Rev. E **83**, 011908 (2011)
30. Welch, V., Lousse, V., Deparis, O., Parker, A.R., Vigneron, J.P.: Orange reflection from a three-dimensional photonic crystal in the scales of the weevil *Pachyrrhynchus congestus pavonius* (Curculionidae). Phys. Rev. E **75**, 041919 (2007)
31. Saranathana, V., et al.: Structure, function, and self-assembly of single network gyroid ($I4_132$) photonic crystals in butterfly wing scales. Proc. Natl. Acad. Sci. U. S. A. **107**, 11676–11681 (2010)
32. Hallam, B.T., Hiorns, A.G., Vukusic, P.: Developing optical efficiency through optimized coating structure: biomimetic inspiration from white beetles. Appl. Opt. **48**, 3243–3249 (2009)
33. Vigneron, J.P., et al.: Switchable reflector in the Panamanian tortoise beetle *Charidotella egregia* (Chrysomelidae: Cassidinae). Phys. Rev E **76**, 031907 (2007)
34. Mäthger, L.M., Hanlon, R.T.: Malleable skin coloration in cephalopods: selective reflectance, transmission and absorbance of light by chromatophores and iridophores. Cell Tissue Res. **329**, 179–186 (2007)
35. Mäthger, L.M., Denton, E.J., Marshall, N.J., Hanlon, R.T.: Mechanisms and behavioural functions of structural coloration in cephalopods. J. R. Soc. Interface **6**, S149–S163 (2009)
36. Cooper, K.M., Hanlon, R.T., Budelmann, B.U.: Physiological color change in squid iridophores. II. Ultrastructural mechanisms in *Lolliguncula brevis*. Cell Tissue Res. **259**, 15–24 (1990)

Structural Colors

▶ Nanostructures for Coloration (Organisms Other Than Animals)

Structural DNA Nanotechnology

▶ DNA Origami as Programmable Nanofabrication Tools

Structural Fluctuations

▶ DNA from First Principles

Structure and Stability of Protein Materials

Szu-Wen Wang
Chemical Engineering and Materials Science, The Henry Samueli School of Engineering, University of California, Irvine, CA, USA

Synonyms

Nanostructure; Thermostability

Definition

Protein-based materials are polymeric biomaterials comprising amino acid subunits that are connected together by peptide bonds. These materials are usually biomimetic, self-assemble into higher-order nanometer-scale architectures, and can interact with biological entities. Determination of their structure and stability is an important component of assessing their utility and function.

Protein-Based Nanomaterials

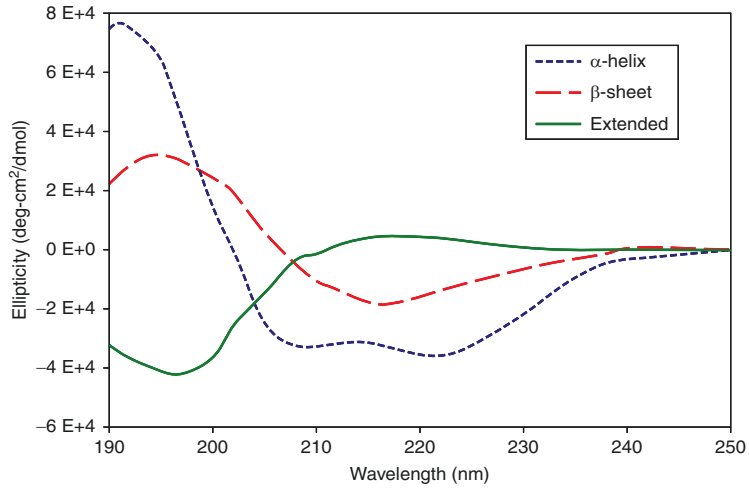
The control of architecture at the nanoscale is a challenge in which nature has been highly successful. Since genetic manipulation enables the definition of every monomer in a polymeric protein structure, giving far greater control than conventional chemical synthesis, one approach in material synthesis is the use of protein engineering to create biologically inspired materials [1]. By combining natural scaffolds, structural elements, and biologically reactive sites, materials with novel architectures and properties can be obtained. These materials are fabricated in microorganisms as recombinant proteins comprising amino acid units. Furthermore, proteins often self-assemble into complex, higher-order nanostructures, which include fibrous and spherical architectures [2, 3]. Protein-based materials have been evaluated for a wide variety of applications, such as tissue engineering scaffolds, templates for nanomaterial synthesis, and drug delivery carriers. The evaluation of structure and stability in these protein-based nanomaterials is primarily established using techniques developed for general proteins.

Structure

The structures of proteins span several length scales, and different methods are used depending on the type of structure being evaluated [4]. The *primary structure* of proteins describes the sequence of individual amino acids, or monomer residues, covalently coupled by peptide bonds to form a polypeptide chain. There are 20 known natural amino acids, each differing in their side chain identity, thereby imparting different physicochemical characteristics at each location along the polymer. One can also incorporate "unnatural" amino acids, which expands the functionalization of the protein-based material with useful chemical components that are not native [5, 6]. When the protein materials have been created in a recombinant system (such as in microorganisms), the DNA encodes for the protein sequence. Primary structure can also be determined by N-terminal peptide sequencing or proteolytic mass spectrometry.

The local primary structure affects the *secondary structure*, which are regions of local folding. Secondary structure of protein materials is commonly determined by circular dichroism (CD) [7]. Asymmetric molecules, such as polypeptides, will absorb left and right circularly polarized light to different degrees. Circular dichroism is this absorbance difference, and different types of secondary folding (such as α-helices, β-sheets, or extended helical conformations) have characteristic CD spectra (Fig. 1, [8]). The degree of folding can be used to monitor conformational changes of protein materials.

Tertiary structure is the three-dimensional conformation of the folded protein. The *quaternary structure* is the assembly of the individual polypeptide or protein subunits. Global size or assembly properties, such as hydrodynamic diameters or molecular weights of the protein complexes, can be evaluated by dynamic light scattering or ultracentrifugation. While these techniques can give useful information, they do not elucidate structural details. The most precise structural information can be obtained by techniques such as X-ray diffraction or nuclear magnetic resonance spectroscopy. However, these methods require highly pure protein, a crystalline form of the protein, and/or extensive data analysis, and these conditions are not usually available or feasible for protein-based materials used in nanotechnology applications. Therefore, alternative methods providing intermediate amounts of information are used to probe tertiary and quaternary structure.



Structure and Stability of Protein Materials, Fig. 1 Circular dichroism spectra of polypeptides with representative α-helix (*dotted line*), β-sheet (*dashed line*), and extended helical (*solid line*) conformations (Data replotted from Ref. [8]. Reprinted with permission, copyright 1969, American Chemical Society)

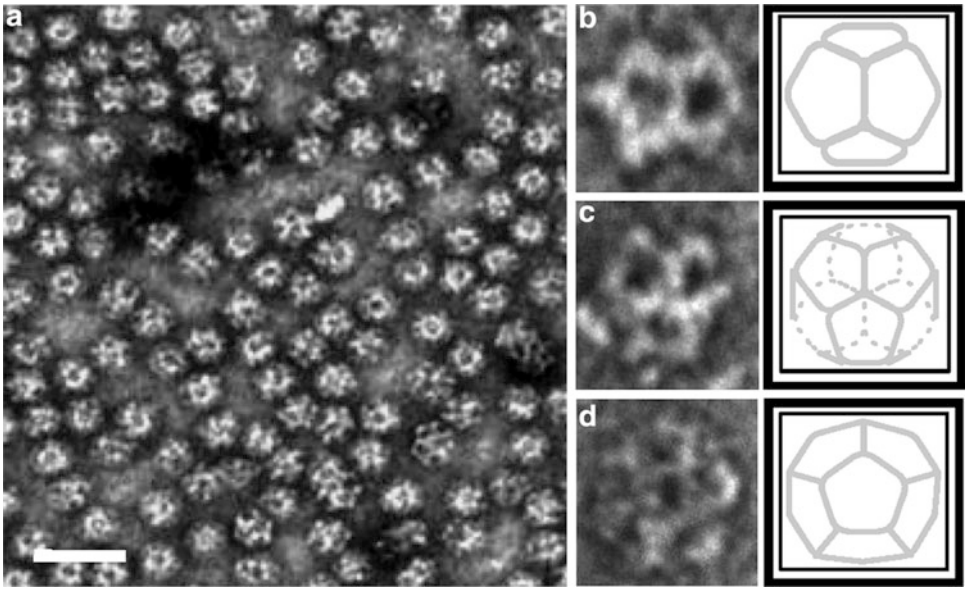
Microscopy strategies enable visualization of protein nanostructures ranging from nanometer-to-micron length scales. In transmission and scanning transmission electron microscopy (TEM and STEM, respectively), an electron beam is transmitted through a sample, and information regarding the scattered electrons is obtained [9]. Proteins are first deposited in a thin layer onto carbon-coated electron microscope grids. These grids are then frozen at liquid nitrogen temperatures in cryogenic preparation, or proteins can also be negatively stained with a heavy ion salt (e.g., uranyl acetate) to provide contrast. Due to the insulating nature of proteins, the resolution using TEM or STEM is low relative to conducting samples and is typically on the order of 5–10 nm (Fig. 2, [10]).

Alternatively, atomic force microscopy (AFM) can yield lateral resolutions of less than 1 nm for protein structures (Fig. 3, [9]). Proteins are adsorbed onto an atomically flat substrate (typically mica or graphite) and are imaged dried or hydrated in buffer. As the AFM probe is scanned over the surface of a sample, deflections (due to interactions between the probe tip and sample) are measured. Wet samples reflect native protein conditions, and therefore AFM potentially can yield more accurate structural information relative to electron microscopy. One challenge of AFM relative to electron microscopy, however, is the time-intensive nature of imaging.

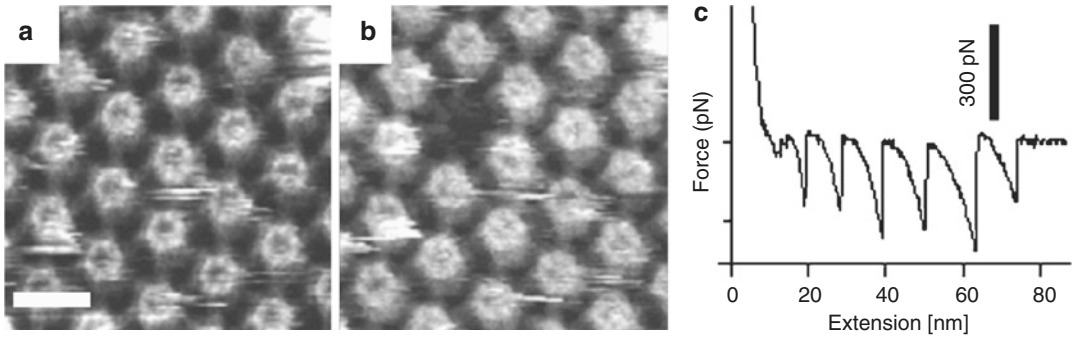
Stability and Mechanical Properties

The stability of protein materials can be evaluated as the structure and assembly are probed while stressing conditions (such as temperature, pH, denaturants, or ionic strength) are changed. For example, the thermostability of protein materials can be evaluated by CD measurement at a characteristic wavelength as the temperature of the sample is increased [11]. Using this data, thermodynamic and thermostability values for protein unfolding (e.g., enthalpy, entropy, free energy, midpoint of unfolding temperature, and onset of unfolding temperatures) can be calculated. Differential scanning calorimetry (DSC), which measures the heat capacity of a protein solution as a function of temperature, can also determine unfolding and thermodynamic parameters of protein-based materials [12]. DSC can also be used to gain insight into kinetic stability and conditions in which protein materials exist at a local energetic minimum with a nonideal structure [13]. Although more direct thermodynamic information can be obtained from DSC, advantages of CD over DSC include the significant lower amounts of protein required for analysis and greater accessibility to spectropolarimeters.

Applying force to single protein molecules gives insight into how the materials may be assembled and the stability of these interactions. Using single-molecule force spectroscopy, AFM can be used to pull out proteins within structured



Structure and Stability of Protein Materials, Fig. 2 Transmission electron microscopy image of dodecahedral protein complex. (**a**) Negatively stained protein assembly of the E2 subunit from pyruvate dehydrogenase. Scale bar is 50 nm. Projection views of (**b**) twofold, (**c**) threefold, and (**d**) fivefold axes of symmetry verify icosahedral assembly (Figure from Ref. [10]. Reprinted with permission, copyright 2008, Wiley)



Structure and Stability of Protein Materials, Fig. 3 Atomic force microscopy (*AFM*) analyses of protein structures. (**a**) Polymers of recombinant human collagen, with each strand corresponding to a trimeric assembly. Image size is 1 × 1 μm (Image provided by Richard Que, University of California, Irvine). (**b**) Extracellular hexagonally packed intermediate (*HPI*) monolayer. Scale bar is 20 nm. (**c**) Force-distance curve of a HPI monolayer exhibits six peaks, corresponding to successive unfolding of the linked HPI proteins. (Panels **b** and **c** are from Ref. [9]. Reprinted with permission, copyright 2011, Elsevier)

assemblies, and force-distance curves yield information about the forces required to disrupt folded proteins and protein complexes [9, 14]. Subsequent imaging by AFM can show the resulting effects on the complex (Fig. 3, [9]).

Protein-based materials often demonstrate both viscous and elastic properties, particularly if they exist at high concentrations or as hydrogels. The storage modulus (G') describes the elastic, solid-like properties, and the loss modulus (G'') measures its viscous, liquid-like character. These dynamic moduli can be determined experimentally with a rheometer or with particle-tracking microrheology [15]. The latter technique

S

measures the displacement of microparticles mixed with protein solution over time and is particularly useful when only small amounts of protein are available, which is often the case in early development of protein materials. Static mechanical properties, such as tensile stress and tensile strain, of bulk protein-based hydrogels can be measured using a conventional mechanical tester.

Cross-References

- ▶ AFM in Liquids
- ▶ AFM, Tapping Mode
- ▶ Atomic Force Microscopy
- ▶ Bioinspired Synthesis of Nanomaterials
- ▶ Biomimetic Synthesis of Nanomaterials
- ▶ Biomimetics
- ▶ Biomimetics of Marine Adhesives
- ▶ Mechanical Properties of Hierarchical Protein Materials
- ▶ Spider Silk
- ▶ Transmission Electron Microscopy

References

1. DiMarco, R.L., Heilshorn, S.C.: Multifunctional materials through modular protein engineering. Adv. Mater. **24**, 3923–3940 (2012)
2. Kluge, J.A., Rabotyagova, U., Leisk, G.G., Kaplan, D. L.: Spider silks and their applications. Trends Biotechnol. **26**, 244–251 (2008)
3. Molino, N.M., Wang, S.W.: Caged protein nanoparticles for drug delivery. Curr. Opin. Biotech. **28**, 75–82 (2014)
4. Creighton, T.E.: Proteins: Structures and Molecular Properties, 2nd edn. Freeman, New York (1993)
5. Xie, J.M., Schultz, P.G.: Innovation: a chemical toolkit for proteins – an expanded genetic code. Nat. Rev. Mol. Cell Biol. **7**, 775–782 (2006)
6. Connor, R.E., Tirrell, D.A.: Non-canonical amino acids in protein polymer design. Polym. Rev. **47**, 9–28 (2007)
7. Greenfield, N.J.: Using circular dichroism spectra to estimate protein secondary structure. Nat. Protoc. **1**, 2876–2890 (2006)
8. Greenfield, N.J., Fasman, G.D.: Computed circular dichroism spectra for evaluation of protein conformation. Biochemistry **8**, 4108–4116 (1969)
9. Muller, S.A., Muller, D.J., Engel, A.: Assessing the structure and function of single biomolecules with scanning transmission electron and atomic force microscopes. Micron **42**, 186–195 (2011)
10. Dalmau, M., Lim, S., Chen, H.C., Ruiz, C., Wang, S. W.: Thermostability and molecular encapsulation within an engineered caged protein scaffold. Biotechnol. Bioeng. **101**, 654–664 (2008)
11. Greenfield, N.J.: Using circular dichroism collected as a function of temperature to determine the thermodynamics of protein unfolding and binding interactions. Nat. Protoc. **1**, 2527–2535 (2006)
12. Ladbury, J.E., Doyle, M.L. (eds.): Biocalorimetry 2: Applications of Calorimetry in the Biological Sciences. Wiley, Hoboken (2004)
13. Sanchez-Ruiz, J.M.: Protein kinetic stability. Biophys. Chem. **148**, 1–15 (2010)
14. Muller, D.J., Dufrene, Y.F.: Atomic force microscopy as a multifunctional molecular toolbox in nanobiotechnology. Nat. Nanotechnol. **3**, 261–269 (2008)
15. Wirtz, D.: Particle-tracking microrheology of living cells: principles and applications. Annu. Rev. Biophys. **38**, 301–326 (2009)

Structure of Nanoparticles

- ▶ High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles

SU-8 Photoresist

Frederik Ceyssens and Robert Puers
Department ESAT-MICAS, KULeuven, Leuven, Belgium

Synonyms

Gamma-butyrolactone (GBL); Hardbake (HB); Lithographie, galvanoformung, abformung (LIGA); *N*-Methyl-2-pyrrolidone (NMP); Polyethylene carbonate (PEC); Polymethyl methacrylate (PMMA); Polypropylene carbonate (PPC); Postexposure bake (PEB); Propylene glycol methyl ether acetate (PGMEA)

Definition

SU-8 is a high aspect ratio epoxy-based negative photoresist commonly used as structural material in lithographic fabrication.

Introduction

SU-8 was developed by IBM as a thick negative photoresist targeted to the fabrication of molds for electroplating. The epoxy-based negative photoresist has some remarkable properties. First of all, it has a wide range of coating thicknesses: layers from several hundreds of nanometers up to several hundreds of microns can be deposited by a single standard spin coating step, using the appropriate dilution of the SU-8 resin. Even thicker layers can be deposited and photopatterned successfully as will be discussed later. Second, structures featuring almost straight sidewalls can be created in the SU-8 layer by simple UV exposure through a contact mask. Aspect ratios of about 15 are routinely obtained. Using optimized processes in layers over a few hundreds of microns thick, aspect ratios of over 100 are obtainable [1].

The fabrication of such polymer structures was up to that time only possible by the expensive LIGA technique which requires synchrotron radiation. Though LIGA is still superior in terms of achievable aspect ratio and wall straightness, SU-8 can be humoristically referred to as a "poor man's LIGA." This must not be taken pejoratively per se as "cheap" can be read as "opening new possibilities" just as well.

SU-8 has high chemical resistance and because of the high chemical functionality of the monomers its mechanical properties are good enough to render it useful as structural material in a large number of cases, as further discussed in the material properties section.

SU-8 is an insulator, but covered later it can be made conductive, for example, by metal deposition. Also, it can of course be used as an electroplating mold to create thick, high aspect ratio metal structures.

It is transparent, enabling easy inspection of the underlying structures and certain optical readout principles, for example, based on fluorescence. Furthermore, the processing of SU-8 does not require a large investment in equipment or consumables.

This all leads to the fact that SU-8 has undoubtedly become one of the most widely used structural polymer materials in microsystems. It is not only frequently used in work of the traditional micromechanics community but has also spread into groups working on labs-on-chip, integrated optics, packaging, and others. This is illustrated by the large number of research papers mentioning SU-8 in the title that have appeared over the last decade (Fig. 1).

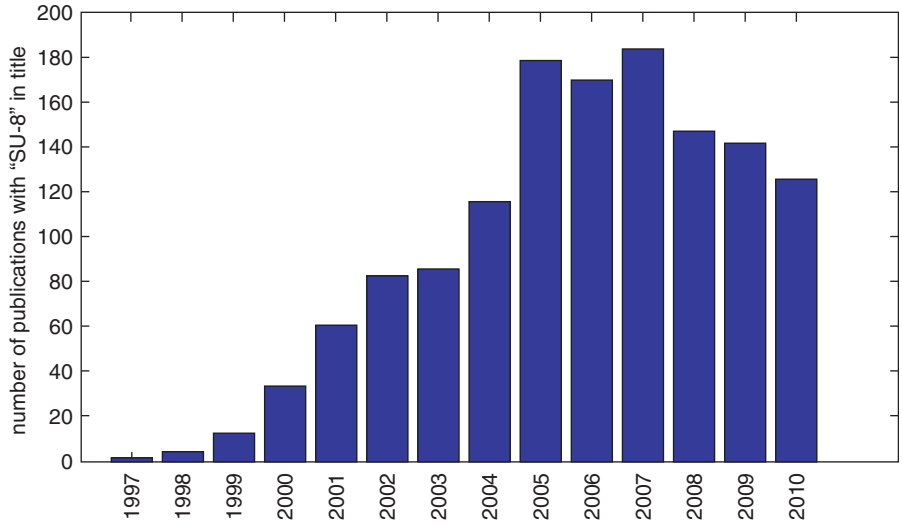
Commercially available SU-8 resist has a solid fraction consisting of SU-8 monomer (Fig. 2), which contains eight epoxy functional groups, and a mixture of triaryl sulfonium hexafluoroantimonate salts as photoinitiator (4.8 wt%).

The solvent added hereto is either gamma-butyrolactone (GBL) or cyclopentanone. Both types are available from Microchem in a wide range of dilutions under the names "SU-8" or "SU-8 2000," respectively. Another supplier is Gersteltec.

Material Properties

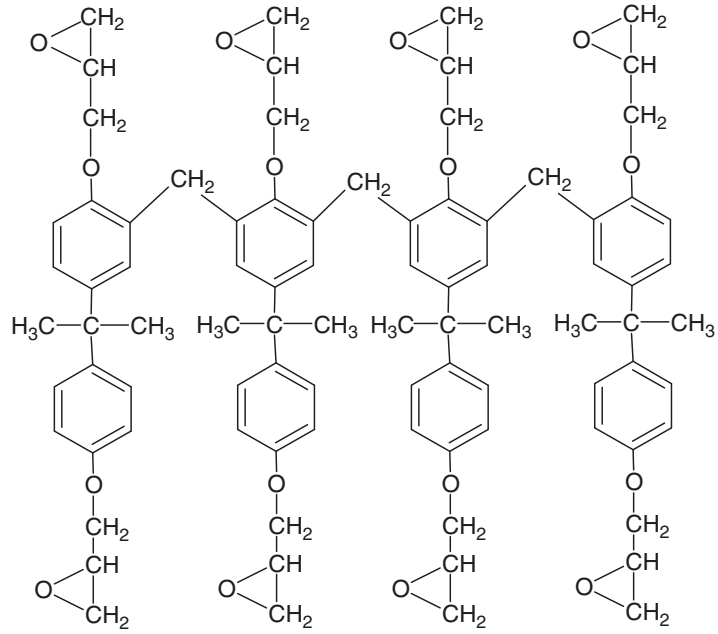
The mechanical and electrical properties of SU-8 are not unlike those of a typical epoxy. Some material data are summarized in Table 1. It can be seen that creep is rather low, which is a good indication of the suitability of the material for mechanical applications.

For microfluidic applications, the contact angle of the material is often important. Untreated SU-8 is rather hydrophobic, displaying a contact angle of 75° with water. The material can be made more hydrophilic by several means such as a short exposure to oxygen plasma. A treatment at 600 W for 8 s at a pressure of 27 Pa reduces the contact angle to 25.6°. After 4 min at 400 W, the contact angle is down to 3.2° but goes up again to 25° over a period of 40 days [4]. Another way is the addition of surfactants to the SU-8 itself. Bohl et al. [5] used 10 wt% trisiloxane alkoxylate to obtain a contact angle of 28°. Combined with



SU-8 Photoresist, Fig. 1 Number of research papers with "SU-8" in their title per calendar year according to Google Scholar on February 15, 2011. As SU-8 is becoming a mature and commonly used technology, it being mentioned in the title of research papers is expected to become less frequent again

SU-8 Photoresist, Fig. 2 SU-8 monomer



plasma activation this becomes a mere 10°. Mostly, adhesion is sufficient (Table 2), except when using glass substrates. An adhesion promoter based on the deposition of a titanium oxide monolayer (such as AP-300) can then be used.

Processing of Thick High Aspect Ratio Structures

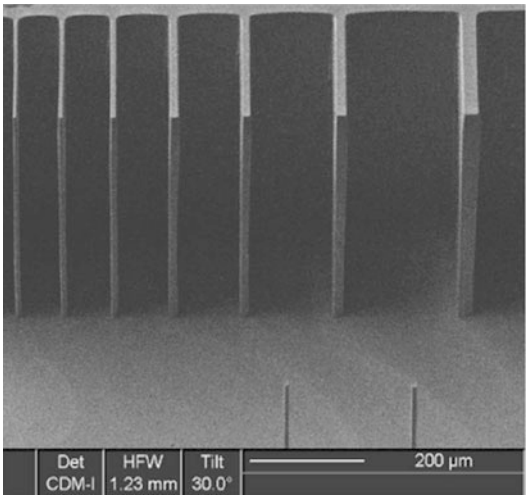
The procedure to deposit a single layer of SU-8 comprises seven major steps. In each of these steps, several parameters can be selected, each

SU-8 Photoresist, Table 1 Material properties of SU-8 [2]. Creep and fracture strength data from [3] who are using the following modified Voight-Kelvin model to model creep in SU-8: $\frac{\varepsilon(t)}{\sigma(\text{applied})} = D_0 + (D_e - D_0)\left(1 - e^{-(t/\tau)^m}\right)$. With ε as the mechanical strain and t time. The mechanical stress σ_{applied} used in these experiments was 13.2 MPa

Property	Value	Remarks
Young's modulus (GPa)	2.1 ± 0.2	HB @ 150 °C for 1 h
Maximum strain to break (%)	1.2 ± 0.5	HB @ 150 °C for 1 h
Fracture strength (MPa)	73.3 ± 9.2	HB @ 180 °C for 30′
Transient creep compliance De (GPa^{-1})	0.310 ± 2.5 %	HB @ 180 °C for 30′
Initial creep compliance D_0 (GPa^{-1})	0.251 ± 13.8 %	HB @ 180 °C for 30′
Creep time constant τ (s)	1.84×10^{10} ± 3.5 %	HB @ 180 °C for 30′
Creep coupling parameter m	0.208 ± 5.3 %	HB @ 180 °C for 30′
CTE (ppm/K)	52 std. 5.1	After PEB @ 95 °C
Thermal conductivity (W/m K)	0.3	
Breakdown strength (V/m)	200	Tested on a 3 μm thick layer
Refraction index	1.658	λ = 365 nm
	1.638	λ = 405 nm
Contact angle with water	75°	Untreated

SU-8 Photoresist, Table 2 Shear adhesion test for SU-8 2000 [6]. Test conditions: RCA clean for glass, plasma clean for other surfaces. Test tool: Dage 4,000 working on 100 × 100 × 50 μm SU-8 posts. SU-8 was hardbaked at 150 °C

Substrate	Adhesion (MPa)
Si	53
SiN	43
Ni	45
Au	29
Al/Cu (99-1)	23
Cu	38
Cu with AP-300	56
Glass	Poor
Glass with HDMS	Poor
Glass with AP-300	92
Quartz	61



SU-8 Photoresist, Fig. 3 SU-8 test structures, 700 um high, down to 10 um wide

having an effect on one or more properties of the final structure. Especially to attain thick, high aspect ratio features, such as shown in Fig. 3, careful processing is mandatory.

The main process sequence and its most significant parameters will be discussed below. Most important in process design is attaining a high ratio between the mechanical strength and the built-in mechanical stress in order to avoid deformation of the fabricated structures. This is done by controlling process parameters such as the solvent concentration during the softbake step and a few additions to the standard process such as the incorporation of relaxation steps and ramped heating [2].

Substrate Preparation

Standard cleaning procedures and a dehydration bake are recommended. On silicon, a HF dip improves adhesion and uniformity [7].

S

The substrate type has a large influence on uniformity and adhesion as well (Table 2).

Layer Deposition

The deposition of a uniform film of non-cross-linked SU-8 on the surface is of course of great importance for the successful outcome of the entire process. Not only is the uniformity of the height of the structures produced determined by this step, the lateral uniformity can be affected as well: an uneven distance between resist surface and mask changes the diffraction conditions during contact exposure, the most common method of UV exposing SU-8.

Spin coating is the most straightforward deposition technique for SU-8. During spinning, irregularities are introduced. The largest and always appearing irregularity is the so-called edge bead. This is a ring near the edge of the wafer where the resist is thicker. The edge bead can be negligible for thinner (<10 μm) SU-8 layers but can be over a centimeter for thick (>100 μm) layers. In the latter case, the edge bead is typically over 50 μm higher than the inner resist layer. Smaller irregularities disappear when the SU-8 layer is given enough time to reflow, preferably in a solvent-saturated atmosphere. The edge bead can be removed mechanically or by a directed spray of a solvent, a feature common in modern coating equipment.

Casting is another method commonly used to deposit SU-8 layers. During casting, a known weight or volume of low viscosity SU-8 is poured on a wafer resting on a leveled hot plate. Reflow and resist spreading in a solvent-rich environment (e.g., by putting a Petri dish over the wafer) is typically necessary. This way, for layers over a few hundreds of microns thick, superior thickness control and uniformity can be achieved. Also, no edge bead formation occurs.

In order to determine the final layer thickness, the shrinkage of the material due to solvent evaporation in the softbake step following the casting step has to be accounted for. A cast layer of SU-8 2007 was determined experimentally to shrink 44 % after a softbake of 6 h at 65 °C followed by 7 h at 95 °C. For SU-8 2010, this is about 32 % and for SU-8 2002, 61 % [2].

Exceptionally thick layers (>1 mm) can be deposited by casting too. However, in such layers a significant solvent concentration gradient was observed to remain after the subsequent softbake step. This tends to weaken fine structures. It was observed that depositing and softbaking several layers on top of each other does not alleviate this significantly, as solvent from an upper layer tends to diffuse into the dry baked lower layers. A better solution is to put a known mass of predried SU-8 flakes on the wafer surface, and melt it into a smooth layer [8].

Softbake

The purpose of the softbake is to lower the solvent concentration, prevent adhesion to a contact mask, and improve the lithographic performance, that is, the ability to produce high aspect ratio structures with straight sidewalls, by reducing the diffusion of the photoinitiator. The softbake temperature must be below 120 °C, which is the onset of thermally induced cross-linking.

A too high solvent content causes the photoacid generated during exposure to diffuse easily to unexposed regions, creating bulges or protrusions in the structure. A too low solvent concentration can cause cracks and peeling off of structures.

It is, therefore, advisable to tune softbake time and temperature to reach a certain solvent concentration. In thick layers, the solvent concentration can be easily monitored by weighing the wafer on a balance with milligram precision. An advised solvent concentration is 5–7 %. An even lower solvent content decreases cross-linking [9] and increases stress deformations in high aspect ratio features [2]. After softbaking, a relaxation period of several hours is advised.

Exposure

Typically, SU-8 is exposed using contact exposure on a mask aligner equipped with a standard mercury arc UV lamp. During exposure, the photoinitiator's triarylsulfonium hexafluoroantimonate salts are split, releasing Lewis acid [7]. The acid generated serves as a catalyst, allowing a two-step reaction to occur that links the epoxy groups of different SU-8 monomers

together. In practice, the latter two-step reaction occurs only very slowly at room temperature. To speed up the reaction rate, the SU-8 is heated up after exposure, during the postexposure bake (PEB) step.

For any practical application, both the exposure dose and the wavelength of the light used are important. From the critical dose up, parts that are unsolvable in developer will be formed. The critical dose for i-line (365 nm) exposure was determined to be 30 ± 0.5 mJ/cm^2 [1], for a PEB of 6′ at 65 °C and 3′ at 95 °C. This dose was observed to be by far insufficient for a good lithographic result: The *adhesion* of the structures formed to the substrate becomes only sufficiently strong to survive development at a much higher exposure dose. A second problem of a low dose is the permanent *deformation* of vulnerable and weakly cross-linked structures caused by internal stress during the development step. On the other hand, it is not possible to increase the exposure dose indefinitely as two problems arise with overexposure: the broadening of the features compared to their designed sizes and the appearance of effects caused by stress in larger parts. Careful choice of process parameters is therefore required. A typical i-line exposure dose is 200 mJ/cm^2.

Furthermore, the choice of exposure wavelength is important. Some UV wavelengths produced by a standard pressurized mercury arc lamp are absorbed too quickly and are not suitable to expose a relatively thick SU-8 layer evenly (Table 3: absorbance). The transmission of light down to a certain depth is illustrated in Fig. 2. This shows that for even exposure of a layer thicker than a few microns, the 313-nm line should always be filtered out. For layers thicker than a few hundred microns, filtering out all light below 400 nm is recommended. As the photoinitiator is less sensitive for 405-nm light, the recommended exposure dose is higher. A dose of 10 J/cm^2 suffices for most applications.

SU-8 Photoresist, Table 3 Absorbance (inverse of penetration depth) of unexposed SU-8 for typical wavelengths [10]

Wavelength (nm)	Absorbance (μm^{-1})
313	1.19×10^{-1}
334	5.85×10^{-2}
365	2.71×10^{-3}
405	1.41×10^{-4}
436	6.90×10^{-5}

When exposing layers of irregular thickness, the use of an index matching fluid such as glycerin to fill up the air gap between mask and wafer can be beneficial to reduce diffraction effects and increase resolution (Fig. 4) [24].

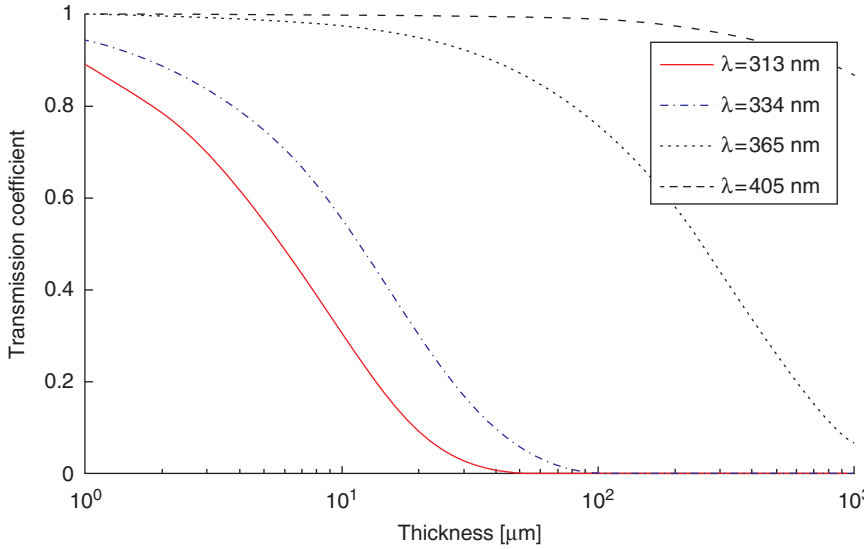
Postexposure Bake (PEB)

The baking step following the exposure of the resist causes a cross-linking reaction in the exposed parts, rendering those parts insoluble during the subsequent development step. Meanwhile, a few side effects occur. These effects are shrinking of the resist by the cross-linking reaction, the thermal expansion of the resist during baking, and the (dissimilar) thermal expansion of the substrate during baking. The situation is clearly different for the in-plane and the out-of-plane direction.

In the out-of-plane direction, the SU-8 layer is unconstrained and shrinking can occur freely. A shrinkage of approximately 4.5 % (std 0.25 %) was measured over a PEB of 1 h at 65 °C [2]. In-plane, although some relaxation occurs, an internal stress is built up (~4 MPa for a PEB temperature of 95 °C in a standard process [2]).

The combination of exposure energy and PEB time and temperature must cater for enough cross-linking for the structures to resist delamination and deformation during the subsequent development step. Typically, postexposure baking raises the glass transition temperature, that is, the temperature above which plastic deformation is possible, to a value slightly larger than the PEB temperature. After a PEB of 20 min, the glass transition temperature lies exactly at the PEB temperature. After that, saturation sets in quickly [11]. Therefore, it makes sense to use a minimum PEB time of 20 min. Typical PEB temperatures are between 80 °C and 100 °C.

After PEB, there is some stress relaxation over time. Hammacher et al. [12] notice a 7.5 % drop in stress during the first 8 h after the PEB.



SU-8 Photoresist, Fig. 4 Transmission of typical UV lines to the bottom of layer versus thickness of that layer, assuming no reflection of light and using the absorbances quoted in Table 3. transmission coefficient $= e^{\text{absorbance} \times \text{thickness}}$

After 4 days, there is an additional drop of 2 %. This indicates that it is beneficial to build a waiting period between PEB and development, diminishing adhesion loss and deformations caused by stress. Also, a ramp up and slow ramp down of the PEB temperature are beneficial. For example, a cool down step of 4 h in an oven can be employed [13].

Development

During development, the non-cross-linked parts of the SU-8 layer are removed in a solvent. Typically, propylene glycol methyl ether acetate (PGMEA) is employed, though ethyl lactate or diacetone alcohol can be used as healthier alternatives with lower vapor pressure [6]. Exposed structures will, even after postexposure baking, adsorb solvents and thus swell during development. Structures in the 10 μm range are typically saturated within a minute [9]. As the solvent evaporates again after development is over, this should not be a problem. However, swollen structures that are not cross-linked hard enough can become permanently deformed during development.

Development is dependent on agitation and on the design of the structures fabricated, and development rates as well as development times are therefore to be taken with caution. A development time of 8 min for a 50-mm-thick layer without agitation is typically sufficient. Ultrasonic agitation was found to increase the development rate by approximately eightfold. However, it can damage fragile structures. Megasonic actuation does not cause this problem. Another option is developing the wafer in an upside down position, which was found to decrease the development time by a factor 3.5 without inducing extra damage [2].

Hardbake

The hardbake is an optional extra step at the end of the SU-8 process. During the hardbake, the material is heated well above the threshold for thermally induced cross-linking. This causes, also in regions where no photoacid is present, additional cross-linking reactions to occur until saturation. Hardbaking is done for increasing the long-term stability of the material and to close small cracks that arise in some process sequences.

The additional cross-linking causes a higher chemical resistance as well: it was found impossible to remove hardbaked SU-8 structures from their substrates with solvents such as acetone or NMP (*N*-methyl-2-pyrrolidone). For this reason, a hardbake is often not advised when the SU-8 layer will have to be removed later in the fabrication process.

For the hardbake, typically a temperature between 150 °C and 250 °C is used. Following the observations of Hammacher et al. on the PEB, it can be assumed that most of the extra cross-linking is complete after 30 min. Hardbaking at 150 °C for 60 min increases internal stress to 12 MPa (tensile) in 45 % humidity air [9]. In the latter work, the built-in stress was shown to be strongly dependent on the humidity in the environment.

Removal of SU-8

The downside of the high chemical resistance of SU-8 is the difficulty to remove the material if needed. Nonhardbaked SU-8 can still be cracked up in strong solvents such as warm NMP. When this step is followed by rinsing or ultrasonification to remove the remaining flakes, a surface having a not too high topography can be clean again.

SU-8 can be cleanly removed in piranha acid, though this is aggressive toward many metals and polymers that may already be present on the wafer. Silicon and glass are not attacked at all. Also, platinum, tantalum, and gold are inert to piranha. Chromium is etched only very slowly, allowing the cleaning of contact masks in piranha after contamination with SU-8. Another application of piranha is for cleaning SU-8 off wafers for reuse.

Reactive ion etching (RIE) in oxygen plasma with small fluorine content is used as well, though traces of a nonvolatile antimony containing residue may be formed.

Typically, 4–5 % of fluorine-containing gas such as SF_6 or CF_4 is added to the reactor atmosphere to boost the etch rate significantly.

Other less widely used options are ashing in an air or oxygen atmosphere and molten salt removal. These need a rather high temperature, 450–500 °C and 350 °C, respectively.

Finally, one can consider downstream chemical etching (DCE). DCE is an organic removal technique in which reactants produced in plasma are blown on the wafer, which is not exposed to the plasma itself. With a wafer temperature of 225 °C and a gas composition of 98 % oxygen and 2 % CF_4, an etch rate of 6.8 μm/min can be obtained [14]. Given the reasonable working temperature, the etch speed, and the low metal etch rate, DCE is likely the most promising method for industrial use.

Alternative Exposure Methods

Next to classic UV lithography, a number of other methods to selectively cross-link SU-8 exist. Though much less commonly used, they open some extra possibilities and therefore, deserve consideration. A distinction can be made between direct write methods in which structures are written pixel by pixel or voxel by voxel, and wafer-scale methods using masks which are generally much faster as writing is done in parallel over a large surface.

The photoinitiator in SU-8 can be activated by the simultaneous absorption of two 800-nm photons, instead of by a single 400-nm photon [15]. The probability of such activation rises with the square of the intensity of the exposure. This is the basis for so-called *two-photon lithography*. The nonlinear dependence makes it possible, by focusing a laser to create a high-intensity spot that is small with respect to the layer thickness, to write true 3D structures such as photonic crystals directly in a single SU-8 layer.

Proton beam writing is another direct write methods found in literature. The penetration depth of the exposing proton beam is dependent on its energy. Thus, bridge-like structures can be written in a single SU-8 layer but not the more complicated structures that are possible with two-proton lithography.

X-ray exposure has been tested as well. Becnel et al. [16] used SU-8 as a more sensitive

alternative for the PMMA normally used as photoresist in LIGA processing. The exposure time needed was in the order of 10 min, two orders of magnitude smaller than for PMMA. The very limited diffraction and low absorption of the X-rays allow for a higher resolution in thick layers than that can be offered by UV-based processing and aspect ratios greater than 100.

Finally, *inclined light UV exposure* is worth mentioning. Some researchers have experimented with exposing SU-8 with inclined UV light through a contact mask. Thus, cylinders and beams with inclined orientation can be realized. Unless countermeasures are taken, the UV light will reflect from the substrate and a mirror image of the structure will be created by the light reflecting back from the substrate surface. At intersection of two cylinders, a ball-shaped structure forms that can be used as an in-plane microlens. Others use the technique to construct microsieves (Fig. 5). Due to the high refractive index of SU-8, it is hard to attain large inclinations as explained by Snell's law: Light entering at an oblique angle from the air will be bent toward the perpendicular axis of the higher refractive index material it enters. A solution for this is to immerse wafer and mask in a higher refractive index fluid such as glycerin. When X-rays are used instead of UV, this is not important as refractive indices do not vary much.

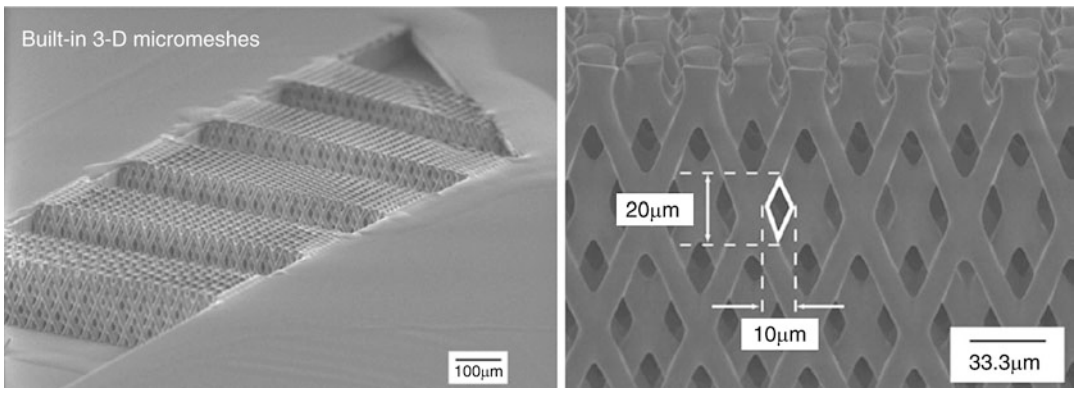
Multilayered, Freestanding Structures in SU-8

A few methods that can be used to fabricated freestanding structures such as cantilevers and microchannels out of SU-8 will be discussed here.

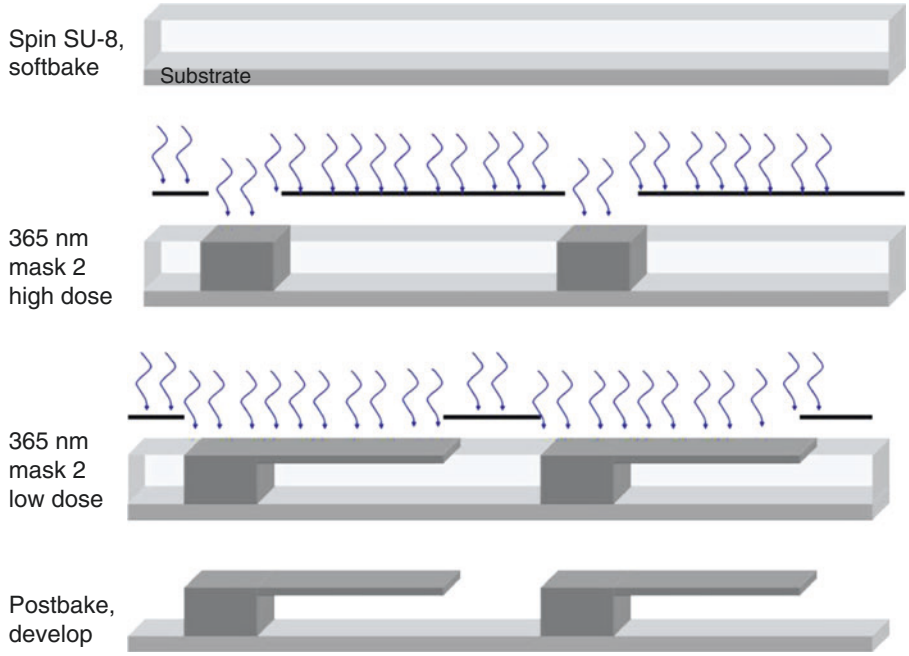
Modulation of Exposure Light

The most straightforward method is to modulate the exposure energy, that is, to use a high UV dose (i-line) to expose a single SU-8 layer from top to bottom where the anchors of the freestanding structures must come and a lower dose to define the freestanding structures (Fig. 6). However, as SU-8 is quite transparent the resulting layer thickness is very sensitive to small variations in the exposure energy and other process parameters. Also, the absorbance of SU-8 tends to increase during exposure.

Tight control of all the different parameters is required which makes it hard to achieve good tolerances. In practice, the method is only useful for fabricating thick freestanding structures with reasonable accuracy in relatively thick layers (>250 μm). Using an antireflective layer or a low-reflectivity substrate reduces the sensitivity of the process for exposure energy variations [18]. The minimum layer thickness is then reduced to 100 μm.



SU-8 Photoresist, Fig. 5 Microsieve structure in microchannel, fabricated in SU-8 using inclined UV exposure ([17], used with permission) *Left*: overview. *Right*: detail of sieve



SU-8 Photoresist, Fig. 6 Principle of the creation of freestanding structures by exposure time control

The above process can be improved by using different wavelengths in the two UV exposure steps. By using a wavelength that is quickly absorbed for the second exposure step, thinner freestanding structures can be made with increased repeatability. It is straightforward to use the 313-nm wavelength for this as it is already present in the spectrum of standard exposure tools. The practical thickness range that can be reached this way is between 6 and 25 μm (Fig. 7). Due to uneven exposure between the top and bottom parts of the freestanding structure, single-clamped beams were observed to bend down [2].

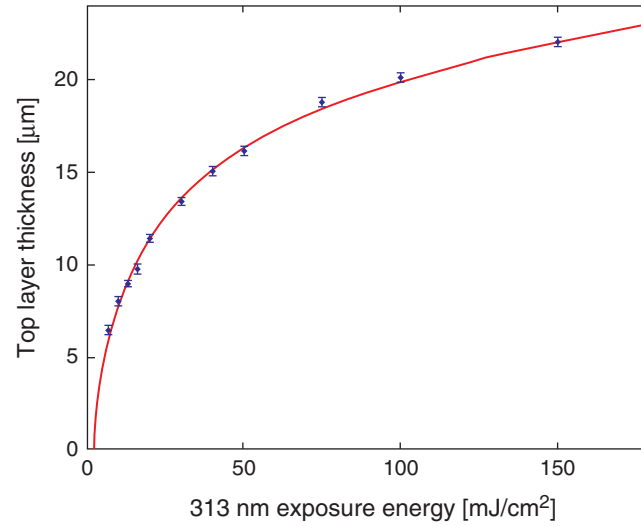
Buried Mask Process

In this process two SU-8 layers are used, as illustrated in Fig. 8. Thus, a larger thickness range and a better control of layer thicknesses are obtained at the cost of a higher process complexity. First, the lower layer is processed as usual up to the PEB step. Then, a UV-blocking layer is deposited. The purpose of this layer is to shield unexposed parts of the lower layer from the UV light used to expose the upper layer, later in the process. Then, the upper layer is spun on and processed. The two layers are developed together at the end of the process sequence. Halfway in the development step, the UV-blocking layer must be removed when it is not dissolving in the normal SU-8 developer, for example, when a metal UV-blocking layer is used. It is important to deposit the UV-blocking layer such that no excess UV or heat is produced, which would cross-link SU-8 unwantedly. Also, temperature must be kept at a minimum when processing the second SU-8 layer to avoid wrinkling [2]. Figure 9 shows some example structures fabricated with the process.

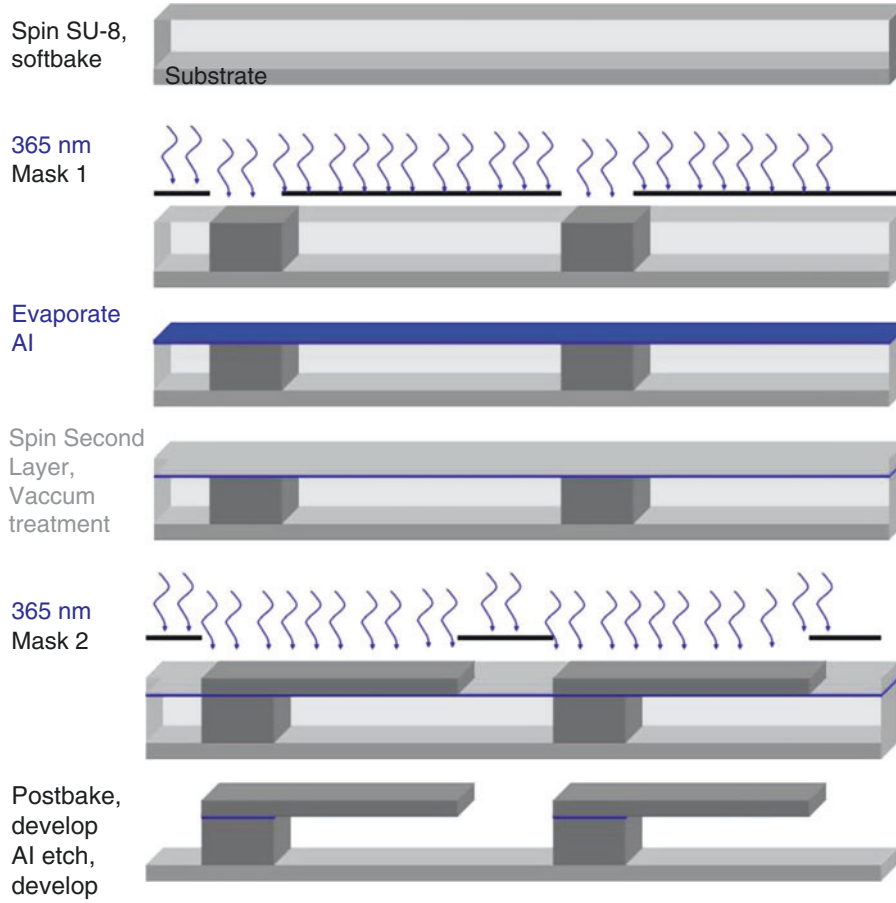
Evaporation by resistive heating of a low melting temperature metal such as aluminum, zinc, or magnesium is a suitable method. Another author uses a metal layer transfer process based on a gold-covered silicone stamp [19].

Sacrificial Layer-Based Processes

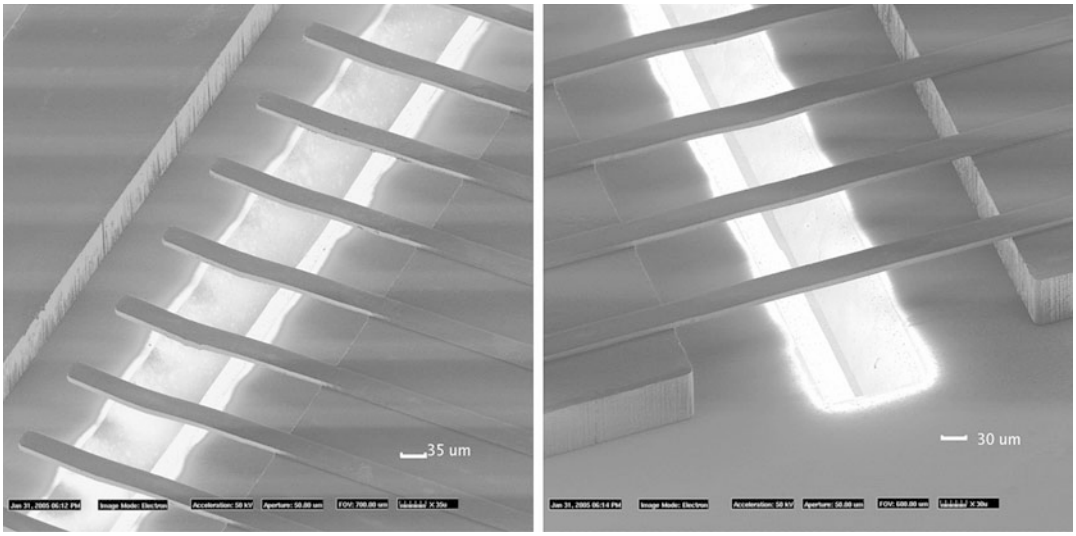
A third major category of processes used to fabricate freestanding SU-8 structures is based on the deposition of a sacrificial layer in another material. After the SU-8 layer is processed on top of the sacrificial layer the sacrificial layer is dissolved, creating freestanding SU-8 structures.



SU-8 Photoresist, Fig. 7 UV exposure energy at 313-nm wavelength versus resulting freestanding layer thickness showing LSE fit. Data points from measurement with (tiny) error *bars* showing standard deviation are plotted as well [2]



SU-8 Photoresist, Fig. 8 Buried mask process



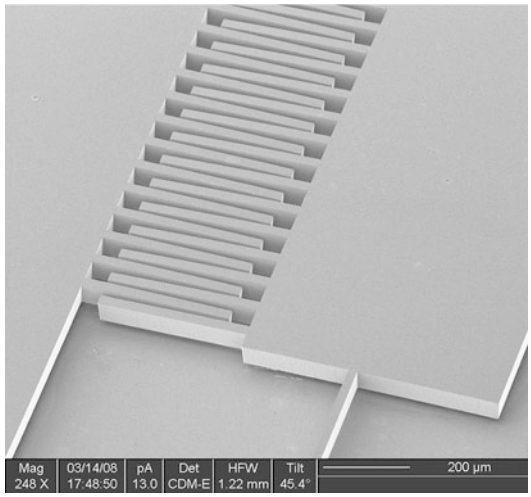
SU-8 Photoresist, Fig. 9 Single- and double-clamped freestanding beams, fabricated with the buried mask method, lying over a KOH-etched U-groove

The sacrificial layer can be patterned before SU-8 deposition to create anchors defining where the SU-8 structures that will be formed in the next step will be attached to the substrate. When no anchor sites are defined either a well-timed etch stop can define anchors or loose SU-8 components are created.

The main property a sacrificial layer must have is that there must be a way to remove it without damaging SU-8. Reversely, the solvent in SU-8 must not dissolve the sacrificial layer to prevent layer intermixing and unpredictable results. Another point of interest is the adhesion of SU-8 on the sacrificial layer, which must be sufficient to allow it to survive the processing. The other properties that discern the different sacrificial layers discussed here are their layer thickness range, processing cost and time, and achievable aspect ratio.

A metal sacrificial layer is a first option. Many metals common to micromachining, such as aluminum and copper, can be etched chemically with perfect selectivity with respect to SU-8. Also, the adhesion of SU-8 to metals is generally satisfying (Table 2). There are some disadvantages to the use of metals, though. The most straightforward way to apply a metal layer is by a thin film technique such as sputtering. Due to stress and relatively slow speed, these processes are limited to a thickness of a few microns. Furthermore, the attainable aspect ratio of the lower layer is determined by the etching process of the metal and is typically around one unless reactive ion etching is used. Both disadvantages can be overcome at the cost of an extra process step by the use of a sacrificial layer that is electroplated in a photoresist mold. Still, when using a metal sacrificial layer, the etch process of the metal may pose constraints on other metal layers that might be present and attacked as well by the release etch. A final point is that metal deposition methods are never self-planarizing, limiting the use on uneven surfaces.

A second candidate is polyimide sacrificial layers. Polyimides resist the solvent in SU-8 very well once properly cross-linked. However, their removal is a problem. Polydimethylglutarimide (PMGI), on the other hand, is etchable in alkaline solutions such as positive photoresist developer while still being resistant to solvents. Thus, it can be removed selectively with respect to cross-linked SU-8 and is easily patternable. Disadvantages are the limited lower layer thickness and the fact that the weak alkaline solution will still etch aluminum and, after an incubation time in which native oxide is removed, even (slowly)



SU-8 Photoresist, Fig. 10 Comb actuator structure fabricated in SU-8 using a PMGI sacrificial layer

silicon. In Ref. [2], comb actuator structures were fabricated based on this process (Fig. 10).

Positive photoresists can also be considered as sacrificial layer. They are inexpensive, can be quickly applied, and are available in a thickness range from below 1 µm to about 100 µm. Therefore, they would be ideally suited as sacrificial layer for a wide range of applications. The main problem hindering their applicability as sacrificial layer for SU-8 is that they dissolve in the solvent (cyclopentanone or GBL) present in SU-8. There are two ways to prevent this. The first is to deposit a thin metal layer on top of the positive resist. The second one is to strongly hardbake the positive photoresist. However, in the latter case resist reflow typically occurs, which severely limits the attainable aspect ratio of the sacrificial layer.

Other sacrificial polymers used by workers in the field are polystyrene (PS) and PMMA. Toluene can be used as a solvent for spincoating and release, as cured SU-8 is resistant to several hours of immersion in toluene.

It is possible to use RIE to pattern the polymers or to include photosensitizers in the PMMA, such as is done for deep-UV photoresists.

A final interesting type of sacrificial polymers is polymers that decompose into volatile components cleanly at a temperature SU-8 can withstand. Examples are polypropylene carbonate (PPC) and polyethylene carbonate (PEC) [20]. Dissolved in NMP (*N*-Methyl-2-pyrrolidone), they can be deposited by spin coating. Structuring is possible using oxygen plasma. PEC starts to decompose thermally around 220 °C in nitrogen at atmospheric pressure. For PPC this is around 240 °C. The most remarkable property of this process is that the volatile decomposition products can diffuse throughout SU-8. Thus, the sacrificial layer can even be removed from closed cavities, and long channels can be cleared in a time independent of channel length. For a 30-mm-thick SU-8 layer, the sacrificial layer removal process time is about 10 min. For very thick upper layers, this time might be considerably longer.

Layer Transfer Processes

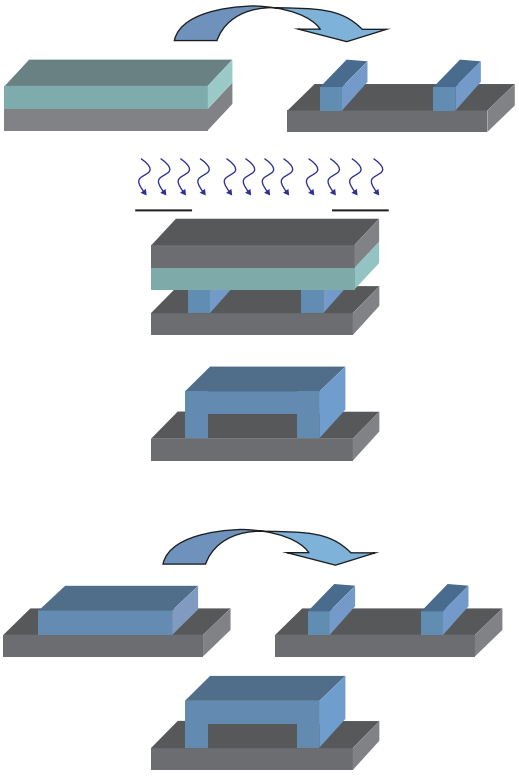
A final major approach to fabricating freestanding structures in SU-8 is to fabricate a single SU-8 layer on a carrier substrate and then transfer that layer to a second substrate on which another SU-8 layer is already present.

This has the advantage that closed cavities or long channels can be fabricated. As the sealed cavities may contain released micromechanical structures, these processes can be used for capping purposes as well.

The different variations of the processes can be further categorized according to the process step after which the layer transfer and optionally the removal of the carrier substrate is done (Fig. 11).

When the transfer is done after softbaking, typically by applying limited heat and pressure, the material is still thermoplastic. Therefore, in order to achieve good control over the shape of the fabricated structure care must be taken to select process parameters such that little flowing of the non-cross-linked layer occurs. A typical sign of too much reflow is the clogging of channels that were sealed by the process and rounded corners.

Another way to limit channel clogging is to use a micron-thin softbaked SU-8 layer to bond two substrates together. Bonding is done at 2 bars in a vacuum bonder [21]. However, the advantage that the upper layer can be readily patterned, for example, integrating vertical microfluidic connectors, is thereby lost.



SU-8 Photoresist, Fig. 11 Layer transfer processes. *Top*: cross-linking after transfer. Sometimes the carrier wafer is not removed, and SU-8 serves as an adhesive between the two substrates. *Bottom*: transfer of partially cross-linked layers

As the transferred SU-8 layer still must be cross-linked, either the carrier wafer must be UV transparent or it must be removed before exposure. After exposure and PEB, the carrier wafer must be removed in any case, unless the SU-8 layer is used as a simple adhesive. The removal is not straightforward with common hard substrates covered with a sacrificial material as this involves under etching of that sacrificial material for several centimeters. A better way is to use a foil as carrier, which can be laminated on with relatively low pressure, avoiding air entrapment, and channel clogging. Furthermore, a foil can be removed by simply peeling it off.

A polyimide foil should be peeled off before exposure as it is not UV transparent. Teflon foil, on the other hand, is transparent to UV.

Due to the rather rough nature of the carrier wafer removal process, it is best used when the top layer contains relatively large features. Examples are device capping and the sealing of microchannels.

Deformation of the transferred layer can be avoided altogether by cross-linking it first before transferring it. The transferred layer can then also be developed before transfer.

To make sure sufficient reactivity remains available to enable thermocompression bonding, the exposure energy and PEB must be scaled down. Arroyo et al. [22] determined the optimal parameters as 140 mJ/cm^2 for i-line exposure followed by a PEB of 4′ at 85 °C. Due to the decreased deformability of the SU-8, a much higher bonding force is needed during thermocompression bonding: good results (95 % yield) were achieved with a pressure 3.25 bar applied at a temperature of 88 °C for 12 min.

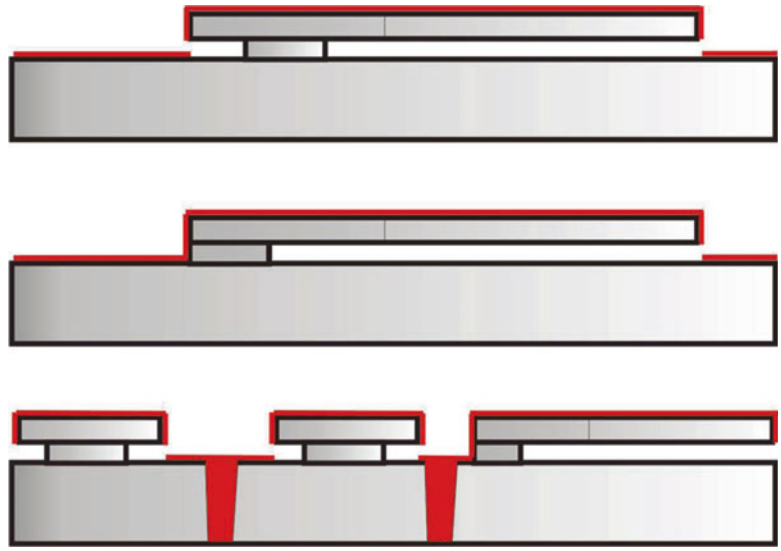
Furthermore, it is of course essential to have good thickness uniformity. The edge bead must certainly be removed, be it mechanically, chemically, or by designing the mask such that the rim of the wafer is not exposed. In this case too, it can be advantageous to use a Teflon or polyimide foil as carrier for the second layer as it can be easily pulled off.

Electrically Active Structures in SU-8

In this paragraph, a few strategies will be discussed to render SU-8 conductive, which is mandatory for many actuation and sensing applications such as comb drives and capacitive accelerometers.

Conductive polymers such as polyaniline (PANI) can be added to the SU-8 resin. However, it was shown that high PANI quantities (up to 50 %) are required in order to achieve useful conductance, a point at which many SU-8 properties such as adhesion degrade.

A different route to create conductive SU-8 is to use a silver nanoparticle-based filler. A silver content of 6 vol% is enough to reach to the percolation threshold, and to create a conductivity of over 10 S cm^{-1}, which can go up a factor 100 with



SU-8 Photoresist, Fig. 12 Metallization of SU-8 by sputter deposition of a thin metal layer (*red*). *Top*: isolated freestanding structure. *Middle*: connection to substrate. *Bottom*: isolation of bonding pad (*left*)

a higher silver content [23]. Such blends are commercially available. However, because of light reflected by the filler, the available aspect ratio and achievable thickness are limited to about 1 and 10 μm, respectively. Carbon nanotubes have been explored as filler material as well. However, conductivities are typically quite limited.

A straightforward method to achieve conductive SU-8 structures is to sputter deposit a thin metal layer. As during sputtering material gets deposited on top of as well as on sidewalls of structures, the mask layout must be designed such as not to cause short circuits, for example, by incorporating overhanging structures such as illustrated in Fig. 12 [2].

Conclusion

Concluding this oversight of SU-8 photoepoxy processing, there is little doubt that this material will find – and indeed already has found – its way into a wide range of applications that benefit from the unique combination of low cost wafer-scale fabrication and the high aspect ratios and very high layer thicknesses attainable. Examples of those are the fabrication of microchannels, labs-on-a-chip, integrated optics, ink jet heads, and electroplating molds for fabricating metal structures. For the latter, electroplating in SU-8 is a low cost alternative for the LIGA process, provided minimum features are several microns wide and more relaxed tolerances can be acceptable. As SU-8 is a low-temperature process, it even shows potential for the fabrication of sensors directly on CMOS wafers.

Cross-References

- ▶ Lab-on-a-Chip for Studies in *C. elegans*
- ▶ Microfluidic Whole-Cell Biosensor
- ▶ Plating
- ▶ Stereolithography

References

1. Zhang, J., Tan, K.L., Hong, G.D., Yang, L.J., Gong, H. Q.: Polymerization optimization of SU-8 photoresist and its applications in microfluidic systems and MEMS. J. Micromech. Microeng. **11**, 20–26 (2002)
2. Ceyssens, F.: Micromachining in polymers and glass: process development and applications. Ph.D. thesis, KULeuven, Leuven, Belgium (2009)
3. Schoeberle, B., Wendlandt, M., Hierold, C.: Long-term creep behavior of SU-8 membranes: application of the time-stress superposition principle to determine the master creep compliance curve. Sens. Actuators A Phys. **142**, 242–249 (2008)
4. Chung, C.K., Hong, Y.Z.: Surface modification of SU8 photoresist for shrinkage improvement in a

monolithic MEMS microstructure. J. Micromech. Microeng. **17**, 207–212 (2007)
5. Bohl, B., Steger, R., Zengerle, R., Koltav, P.: Multilayer SU-8 lift-off technology for microfluidic devices. J. Micromech. Microeng. **15**, 1125–1130 (2005)
6. MicroChem: SU-8 datasheet and adhesion results-shear analysis. www.microchem.com (2007). Accessed 18 Nov 2011
7. Teh, W.H., Dürig, U., Drechsler, U., Smith, C.G., Gntherodt, H.-J.: Effect of low numerical-aperture femtosecond two-photon absorption on SU-8 resist for ultrahigh-aspect-ratio microstereolithography. J. Appl. Phys. **97**, 4095 (2005)
8. Becnel, C., Desta, Y., Kelly, K.: Ultra-deep x-ray lithography of densely packed SU-8 features: I. An SU-8 casting procedure to obtain uniform solvent content with accompanying experimental results. J. Micromech. Microeng. **15**, 1242–1248 (2005)
9. Wouters, K., Robert, P.R.: Diffusing and swelling in SU-8: insight in material properties and processing. J. Micromech. Microeng. **20**, 095013 (2010)
10. Reznikova, E.F., Mohr, J., Hein, H.: Deep photolithography characterization of SU-8 resist layers. J. Microsyst. Technol. **11**, 282–291 (2005)
11. Feng, R., Farris, R.: Influence of processing conditions on the thermal and mechanical properties of SU8 negative photoresist coatings. J. Micromech. Microeng. **13**, 80–88 (2003)
12. Hammacher, J., Fuelle, A., Flaemig, J., Saupe, J., Loechel, B., Grimm, J.: Stress engineering and mechanical properties of SU-8-layers for mechanical applications. J. Microsyst. Technol. **14**, 1515–1523 (2007)
13. Williams, J.D., Wang, W.: Using megasonic development of SU-8 to yield ultrahigh aspect ratio microstructures with UV lithography. J. Microsyst. Technol. **10**, 694–698 (2004)
14. Dentinger, P.M., Miles, C., Goods, S.H.: Removal of SU-8 photoresist for thick film applications. Microelectron. Eng. **61–62**, 993–1000 (2002)
15. Witzgall, G., Vrijen, R., Yablonovitch, E., Doan, V., Schwartz, B.J.: Single-shot two-photon exposure of commercial photoresist for the production of threedimensional structures. Opt. Lett. **23**, 1745 (1998)
16. Becnel, C., Desta, Y., Kelly, K.: Ultra-deep x-ray lithography of densely packed SU-8 features: II. Process performance as a function of dose, feature height and post exposure bake temperature. J. Micromech. Microeng. **15**, 1249–1259 (2005)
17. Sato, H., Matsumura, H., Keino, S., Shoji, S.: An all SU-8 microfluidic chip with built-in 3D fine microstructures. J. Micromech. Microeng. **16**, 2318–2322 (2006)
18. Chuang, Y., Tseng, F., Cheng, J., Lin, W.: A novel fabrication method of embedded micro-channels by using SU-8 thick-film photoresists. Sens. Actuators A Phys. **103**, 64–69 (2003)
19. del Campo, A., Greiner, C.: SU-8: a photoresist for high-aspect-ratio and 3D submicron lithography. J. Micromech. Microeng. **17**, R81–R95 (2007)
20. Metz, S., Jiguet, S., Bertsch, A., Renaud, P.: Polyimide and SU-8 microfluidic devices manufactured by heat-depolymerizable sacrificial material technique. Lab Chip **4**, 114–120 (2004)
21. Carlier, J., Arscott, S., Thomy, V., Fourrier, J.C., Caron, F., Camart, J.C., Druon, C., Tabourier, P.: Integrated microfluidics based on multi-layered SU-8 for mass spectrometry analysis. J. Micromech. Microeng. **14**, 619–624 (2004)
22. Arroyo, M.T., Fernández, L.J., Agirregabiria, M., Ibañez, N., Aurrekoetxea, J., Blanco, F.J.: Novel all-polymer microfluidic devices monolithically integrated within metallic electrodes for SDS-CGE of proteins. J. Micromech. Microeng. **17**, 1289–1298 (2007)
23. Jiguet, S., Bertsch, A., Hofmann, H., Renaud, P.: Conductive SU8-silver composite photopolymer. In: Proceedings of Micro Electro Mechanical Systems, Maastricht, The Netherlands. pp. 125–128 (2004)
24. Yang, R., Wang, W.: A numerical and experimental study on gap compensation and wavelength selection in UV-lithography of ultra-high aspect ratio SU-8 microstructures. Sens. Actuators B **110**, 279–288 (2005)

Sub-retinal Implant

► Artificial Retina: Focus on Clinical and Fabrication Considerations

Subwavelength

► Moth-Eye Antireflective Structures

Sub-wavelength Waveguiding

► Light Localization for Nano-optical Devices

Superelasticity and the Shape Memory Effect

Xiaodong Han[1], Shengcheng Mao[1] and Ze Zhang[1,2]
[1]Institute of Microstructure and Property of Advanced Materials, Beijing University of Technology, Chaoyang District, Beijing, People's Republic of China
[2]State Key Laboratory of Silicon Materials and Department of Materials Science and Engineering, Zhejiang University, Hangzhou, China

Synonyms

Hyperelasticity; Pseudoelasticity; Theoretical elasticity; Ultralarge strain elasticity

Definition

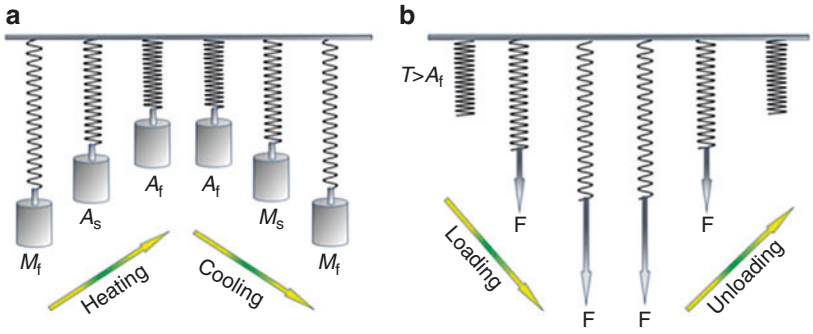
Superelasticity, or pseudoelasticity, is a unique property of shape memory alloys (SMAs), wherein up to 13 % deformation strain can be sustained and the material can recover its original shape after removing the stress. The shape memory effect occurs in SMA and is defined as when a material can remember its original shape upon heating or cooling.

Overview

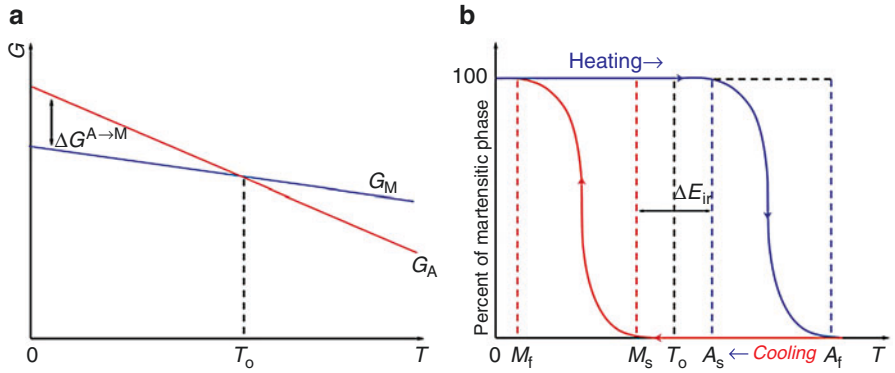
Superelasticity (SE), also called pseudoelasticity, normally refers to a phenomenon observed in shape memory alloys (SMAs), a kind of metal that can remember its original shape after heating or cooling [1–3]. SE is different from true elasticity in bulk metals, which is a reflection of the interatomic spacing variation within a material (Hooke's law). Regular elasticity in bulk metals is normally less than 0.5 %, whereas the elasticity of shape memory alloys can be up to 13 % in Fe-based SMA [4]. This SE occurs through reversible martensitic transformations from austenite to martensite and vice versa. Upon external mechanical loading or temperature change (or a mixture of the two), the SMA deforms by reversible martensitic phase transformations rather than by irreversible plastic dislocation glide. The martensitic transformation occurs through shear on an invariable plane (habit plane) along the shear direction. SE occurs when the metals are deformed in the austenite state, and the deformation shape is recoverable when the applied load is removed.

Large strain elasticity (LSE) can be achieved in nanomaterials due to their defect-free structure, particularly in nanowires. The true elasticity of metallic nanowires can even approach the theoretical elastic limit (8 %) [5], and this is typically associated to extremely high strength [6]. In theory and in computer simulations, some metallic nanowires, such as Cu nanowires, can reversibly deform up to 50 % by a combination of large strain elasticity and pseudoelasticity [7]. Some of these exceptional mechanical properties have already been experimentally demonstrated in metallic nanowires.

The *shape memory effect* (SME) also relates to SMAs, and it refers to the phenomenon where a metallic specimen changes shape in response to temperature changes [1–3]. Figure 1 shows a sketch comparing SME and SE. A deformed SMA can return to its original high-temperature shape upon heating or go back to the low-temperature shape upon cooling (see Fig. 1a). If a SMA is deformed by mechanical loading at a constant temperature higher than a critical temperature (the austenite finish temperature, defined below), the initial shape can be recovered by unloading, through a reverse martensitic transformation (Fig. 1b). The SME was first discovered by A. Olander in Au-Cd alloy in 1932. Investigations into SMA and their applications have been further promoted since the discovery of the SME in the near-equiatomic NiTi alloys, known as "Nitinol." The SME has also been found in other metallic alloys such as Cu-Al-Ni, Cu-Zn, Fe-Mn-Si, Ni-Mn-Ga, and Ti-Ni-Hf. Because of these unique properties,



Superelasticity and the Shape Memory Effect, Fig. 1 A schematic illustration of (**a**) the shape memory effect and (**b**) superelasticity



Superelasticity and the Shape Memory Effect, Fig. 2 Schematic illustration of the evolutions of (**a**) the Gibbs free energies of the martensite and austenite phases and (**b**) the martensitic fraction with temperature

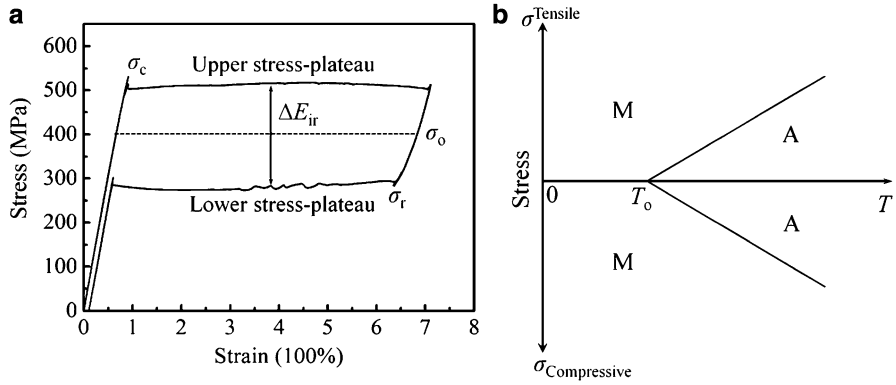
SMAs have been used for a variety of applications, including orthodontic wire, stents, microactuators, pipe coupling, eyeglass frames, microelectromechanical systems, and a variety of biomedical devices [8, 9]. *Superelasticity and the shape memory effect are* correlated with thermoelastic and stress/strain-induced martensitic transformations. In the following, we introduce some basic concepts of *martensitic transformations* in NiTi SMAs.

Martensitic Transformation

Thermodynamic Consideration of the Thermoelastic Martensitic Transformation

The martensitic transformation is a transformation between austenite (stable at high temperature) and martensite (stable at low temperature), and it is usually treated as a diffusionless first-order phase transformation. The Gibbs free energy, defined as $\Delta G = \Delta H - T\Delta S$ (where ΔH and $T\Delta S$ are the changes in chemical enthalpy and entropy), is the pertinent chemical potential of a system, which is at the minimum when reaches equilibrium at constant pressure and temperature. The martensite and austenite Gibbs free energies (ΔG^M and ΔG^A) gradually decrease, following different slopes, with increasing temperature (T), as illustrated in Fig. 2a. At temperature T_o, the Gibbs free energies of the two phases reach a thermodynamic equilibrium ($\Delta G^M = \Delta G^A$). To drive the forward or reverse martensitic transformation, the energy gap between the two phases ($\Delta G^{A\to M}$ or $\Delta G^{M\to A}$) needs to be overcome by either temperature or mechanical loading, as expressed in Eqs. 1 and 2: [10].

$$\begin{aligned}\Delta G^{A\to M} &= \Delta H^{A\to M} - T\Delta S^{A\to M} - \Delta E_{mech} \\ &= \Delta G^{A\to M} - T\Delta S^{A\to M} - \frac{\sigma\varepsilon}{\rho}\end{aligned} \quad (1)$$



Superelasticity and the Shape Memory Effect, Fig. 3 (**a**) A typical stress–strain curve of NiTi SMA showing superelasticity. (**b**) Schematic illustration of the stress–temperature relationship of an SMA. The stress can be both tensile and compressive stress

$$\Delta G^{M\to A} = \Delta H^{M\to A} - T\Delta S^{M\to A} + \frac{\sigma\varepsilon}{\rho} \qquad (2)$$

where ρ, σ, and ε are the density of materials, the external stress, and strain, respectively.

Figure 2b shows schematically the evolution of the martensite volume fraction with temperature during forward and reverse transformations. There are four characteristic temperatures defining a thermoelastic martensitic transformation: (1) The martensite start temperature, *M***s**, at which martensite starts to nucleate; (2) the martensite finish temperature, M_f, with $M_f < M_s$, at which the transformation has completed; (3) the austenite start temperature, A_s, at which the system starts moving from martensite to austenite upon heating; and (4) the austenite finish temperature, A_f, with $A_f > A_s$, at which the martensitic phase has fully transformed into the austenite phase.

Aside from the thermally induced martensitic transformation, martensite can also be obtained by applying external stress. Figure 3a shows a typical stress–strain curve of NiTi SMA showing superelasticity. There are two stress plateaus, defined as the upper stress plateau and the lower stress plateau, corresponding to the forward and reverse martensitic transformation, as indicated in Fig. 3a. The martensite starts to nucleate after reaching a critical transformation stress, σ_c. The martensitic transformation then proceeds at nearly constant stress upon further straining (the physical mechanisms are described below). As shown in Fig. 3a, the upper plateau stress is lower than the critical transformation stress. As a result, the growth energy is smaller than that of the nucleation energy of martensite. The austenite nearly fully transforms to martensite at the end of the upper stress plateau. Upon unloading to a critical reverse stress, σ_r, martensite starts transforming back to austenite. The martensite phase is nearly fully reverted to the austenite phase at the end of the lower stress plateau. According to Eq. 1, higher stress is required to trigger martensite transformations as the ambient temperature increases. The stress–temperature relationship in stress-induced martensitic transformations is schematically illustrated in Fig. 3b.

As shown in Figs. 2b and 3a, the forward and reverse deformation paths do not overlap, i.e., temperature and stress hysteresis exist during thermo- and stress-induced martensitic transformation. The austenite start temperature is higher than the martensite start temperature ($A_s > M_s$, $A_s > T_o$ and $M_s < T_o$), and the unloading stress plateau is lower than the loading stress plateau. This is because the generation of irreversible energies (ΔE^{ir}), including the frictional energy of the martensite–austenite phase interface propagation, the acoustic emission, and the production and motion of dislocations, is ineluctable during the martensitic transformation [11]. Therefore, additional chemical or mechanical energy is required to promote the forward and reverse martensitic transformations.

Crystallography of Stress-Induced Martensitic Transformation

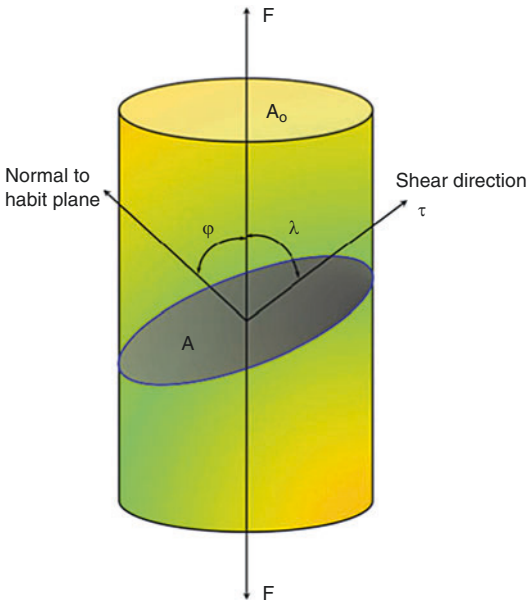
Critical Transformation Stress

Two important characteristics, the critical transformation stress and transformation strain, define the stress-induced martensitic transformation. It has been suggested that the selection of a martensitic variant under external stress (uniaxial tension and compression) comes from satisfying Schmid's law, as illustrated in Fig. 4 [12]. The stress component along the shear direction in the habit plane can be expressed as:

$$\tau_s \frac{F_s}{A_s} = \frac{F\cos\lambda}{A_o/\cos\varphi} = \frac{F}{A_o}\cos\lambda\cdot\cos\varphi = \sigma\cos\lambda\cdot\cos\varphi \tag{3}$$

where $m = \cos\lambda\cdot\cos\varphi$ is defined as the Schmid factor, λ is the angle between the loading axis and the shear direction, and φ is the angle between the loading axis and the normal of the habit plane.

According to Eq. 3, the variant with the highest Schmid factor and, consequently, the highest resolved shear stress will be triggered in one grain, which shows that the critical transformation stress of martensite is strongly related to the grain/crystal orientation.



Superelasticity and the Shape Memory Effect, Fig. 4 Schematic illustration of Schmid's law

Maximum Recoverable Strain

The two phases of NiTi alloys, austenite and martensite, are known to have cubic (B2) and monoclinic (B19') structures. Figure 5 shows a schematic illustration of the lattice distortion of the B2–B19' transformation in NiTi SMA. A martensitic unit cell is identified by the dashed line within the austenite lattice, as shown in Fig. 5a. The embryo may be transformed to the martensitic unit cell by a simple "bain distortion," as indicated in Fig. 5b, which can be expressed as an expansion in c, contractions in the a and b directions, and a change of the beta angle from 90° to 96.8°.

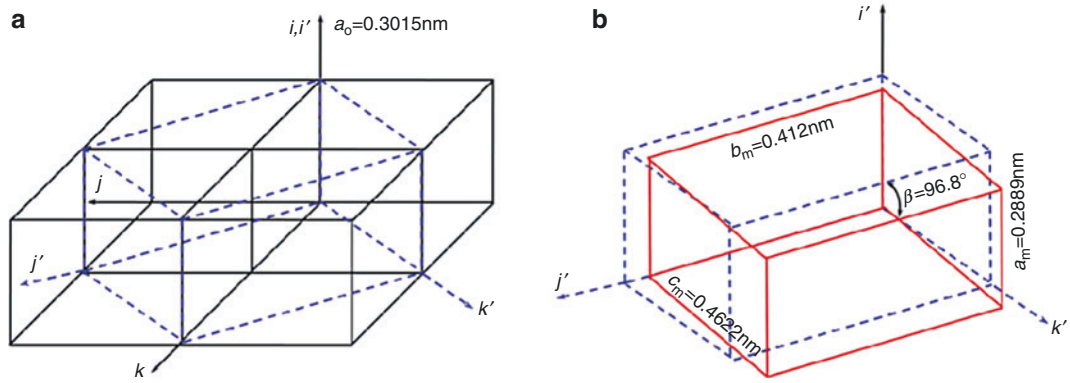
Such a transformation can happen in 12 equivalent lattice correspondence martensitic variants (LCMVs) in NiTi [13]. The lattice distortion shown in Fig. 5 corresponds to LCMV #1 with lattice correspondences of $[001]_m$–$[001]_a$, $[010]_m$–$[011]_a$, and $[001]_m$–$[0\text{–}11]_a$. The lattice deformation matrix M' in the coordinates of the martensite (i', j', k') can be expressed using the lattice constants of the parent and the product phases. For the martensitic transformation in near-equiatomic NiTi, the lattice constant of the austenite is $a_o = 0.3015$ nm and those of the martensite are $a_m = 0.2889$ nm, $b_m = 0.412$ nm, $c_m = 0.4622$ nm, and $\beta = 96.8°$ [14]; thus:

$$M' = \begin{bmatrix} \frac{a_m}{a_o} & 0 & \frac{c_m\cos\beta}{\sqrt{2a_o}} \\ 0 & \frac{b_m}{\sqrt{2a_o}} & 0 \\ 0 & 0 & \frac{c_m\sin\beta}{\sqrt{2a_o}} \end{bmatrix} = \begin{bmatrix} 0.9582 & 0 & 0.1283 \\ 0 & 0.9663 & 0 \\ 0 & 0 & 1.0763 \end{bmatrix} \tag{4}$$

The lattice deformation matrix M in the coordinates of austenite (i, j, k) can then be transformed from M' with a coordinate transformation matrix R from the martensite to the austenite via:

$$M = RM'R^{\mathrm{T}} \tag{5}$$

S



Superelasticity and the Shape Memory Effect, Fig. 5 Schematic illustration of the lattice distortion of the B2–B19′ martensitic transformation in NiTi: (**a**) atomic coordinate systems, with (i, j, k) representing the reference frame in the austenite and (i', j', k') representing the reference frame in the martensite, (**b**) lattice distortion from B2 (*dashed lattice*) to B19′ (*solid lattice*)

where R^{T} is the transpose of R. Using the lattice correspondence of austenite and martensite, the coordinate transformation matrix R for the variant shown in Fig. 5 is expressed as:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \tag{6}$$

The lattice deformation matrix M is then calculated to be:

$$M = RM'R^T = \begin{bmatrix} \alpha & -\beta & \beta \\ 0 & \omega & -\gamma \\ 0 & -\gamma & \omega \end{bmatrix} \tag{7}$$

where $\alpha = 0.9582$, $\beta = 0.0907$, $\omega = 1.0213$, and $\gamma = 0.0550$.

With this, a vector $\boldsymbol{x}$ in the austenite is transformed to $\mathbf{x}'$ upon the martensitic transformation by the following equation:

$$\mathrm{x}' = Mx \tag{8}$$

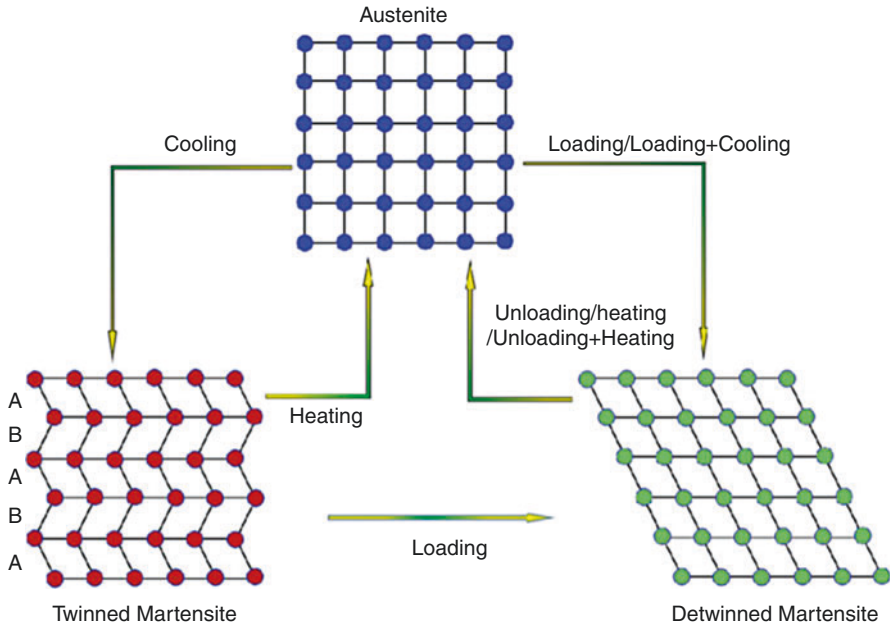
Consequently, the transformation strain can be calculated by:

$$\varepsilon = \frac{|\mathrm{x}'| - |\mathrm{x}|}{|\mathrm{x}|} \tag{9}$$

According to thermodynamic principles, the variant with the maximum strain in the direction of the applied load produces the maximum driving force for the stress-induced martensitic (SIM) transformation [15] and is triggered to form first.

Mechanisms of the Shape Memory Effect and Superelasticity

The unique properties, superelasticity and the shape memory effect, of SMA can be expressed by a simplified 2D geometric depiction, as illustrated in Fig. 6. Upon cooling to a temperature $T < M_{\mathrm{f}}$, two martensitic variants, defined as A and B, with the same crystal structure but different orientations were cooperatively triggered to accommodate and minimize the martensitic transformation strain. The interface between the thermally induced martensite, composed of the matrix and its twin, and the austenite is the habit plane. The martensite variant at the habit plane is then referred to as the habit plane variant. Under external stress ($T < M_{\mathrm{f}}$), martensitic variant A grows at the expense of variant B by the movement of the interface of variants A and B. The deformation process is referred to as detwinning or reorientation of the martensites. The twinned or detwinned martensitic variants can fully transform back to austenite when heating the specimen



Superelasticity and the Shape Memory Effect, Fig. 6 Schematic illustration of the martensitic structure upon loading, unloading, heating, and cooling

to a temperature above A_f corresponding to the recovery of the macroscopic SMA to its original shape. There are two types of SME: one-way and two-way. The difference is whether the SMA remembers its low-temperature shape. The critical stress to drive the reorientation of martensite is known to be very small (about 100 MPa); therefore, it is very easy to set the shape of SMA at low temperatures ($T < M_f$). For the one-way SME, the shape of SMA will remain stable until heated above the reverse transformation start temperature (A_s) and reverse to its original when the temperature is higher than A_f. Upon the second cooling to a temperature lower than M_f, no macroscopic shape change happens, indicating that only the high-temperature shape is "remembered."

The two-way SME is an effect where both the low-temperature and high-temperature shapes can be "remembered." Thus, when cooling the SMA to a temperature $T < M_f$, the high-temperature shape will change to the low-temperature shape; likewise, the low-temperature shape will transform back to the high-temperature shape on heating to $T > A_f$. The shape variations between the two shapes are stress free. It is an intrinsic property of SMA to "remember" the high-temperature shape; however, additional training is required for SMA to "remember" the low-temperature shape.

In addition to the SME, SMAs have another unique property known as superelasticity, or pseudoelasticity, which occurs at temperatures above A_f and distinguishes SMA from other metallic alloys. Under external mechanical loading, detwinned martensite nucleates in the austenite matrix after reaching the critical resolved shear stress. Martensite continuously propagates with increasing external strain in a Lüders-type manner under a nearly constant stress (see the upper stress plateau in Fig. 3). The nucleation of martensite is so concentrated that a region nearly fully transformed to martensite is the first to form in the SMA. There is a clear edge between the martensite and austenite, with a shear angle of about 55° to the loading axis [15]. The martensite band propagates by new nucleation of martensite ahead of the martensite–austenite edge [16]. The inhomogeneous martensitic transformation is similar to the stress-induced slip band in low-carbon steels. This transformation behavior was first

reported by Guillaume Piobert and W. Lüders in 1864, and the martensitic band is then referred to as a Lüders-like deformation band (LBD).

The stress-induced martensite is unstable at the testing temperature (*higher than Af*) and will transform back to austenite when the external stress is unloaded. The superelasticity of SMA is associated with the reverse martensitic transformation and recovery of the macroscopic deformation strain. Upon unloading to the lower stress plateau (see Fig. 3), the martensite plates inside the LDB gradually transform back to austenite. At the macroscopic scale, the reverse transformation is characterized by shrinkage of the LDB.

Cross-References

- ▶ Nanomechanical Properties of Nanostructures
- ▶ Plasticity Theory at Small Scales
- ▶ Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures
- ▶ Surface Tension Effects of Nanostructures

References

1. Otsuka, K., Wayman, C.M.: Mechanism of shape memory effect and superelasticity. In: Otsuka, K., Wayman, C.M. (eds.) Shape Memory Materials, pp. 27–48. Cambridge University Press, Cambridge (1998)
2. Saburi, T., Nenno, S.: The shape memory effect and related phenomena. In: Proceedings of the International Conference on Solid-Solid Phase Transitions, pp. 1455–1479. Pittsburg (1981)
3. Otsuka, K., Ren, X.: Physical metallurgy of Ti–Ni-based shape memory alloys. Prog. Mater. Sci. **50**, 511–678 (2005)
4. Tanaka, Y., Himuro, Y., Kainuma, R., Sutou, Y., Omori, T., Ishida, K.: Ferrous polycrystalline shape-memory alloy showing huge superelasticity. Science **327**, 1488–1490 (2010)
5. Yue, Y.H., Liu, P., Zhang, Z., Han, X.D., Ma, E.: Approaching the theoretical elastic limit in Cu nanowires. Nano Lett. **11**, 3151 (2011)
6. Wong, E.W., Sheehan, P.E., Lieber, C.M.: Nanobeam mechanics: elasticity, strength and toughness of nanorods and nanotubes. Science **277**, 1971 (1997)
7. Park, H., Gall, S.K., Zimmerman, J.A.: Shape memory and pseudoelasticity in metal nanowires. Phys. Rev. Lett. **95**, 255504 (2005)
8. Humbeeck, J.V.: Non-medical applications of shape memory alloys. Mater. Sci. Eng. A **273–275**, 134–148 (1999)
9. Duerig, T., Pelton, A., Stöckel, D.: An overview of nitinol medical applications. Mater. Sci. Eng. A **273–275**, 149–160 (1999)
10. Wollants, P., Roos, J.R., Delaey, L.: Thermally- and stress-induced thermoelastic martensitic trans- formations in the reference frame of equilibrium thermodynamics. Prog. Mater. Sci. **37**, 227–288 (1993)
11. Liu, Y., McCormick, P.G.: Thermodynamic analysis of the martensitic transformation in NiTi – I. Effect of heat treatment on transformation behaviour. Acta Metall. Mater. **42**, 2401–2406 (1994)
12. Gall, K., Sehitoglu, H.: The role of texture in tension–compression asymmetry in polycrystalline NiTi. Int. J. Plast. **15**, 69–92 (1999)
13. Matsumoto, O., Miyazaki, S., Otsuka, K., Tamura, H.: Crystallography of martensitic transformation in Ti-Ni single crystals. Acta Metall. **35**, 2137–2144 (1987)
14. Wollants, P., Bonte, M.D., Roos, J.R.: A thermodynamic analysis of the stress-induced martensitic transformation in a single crystal. Z. Metallkd. **70**, 113–117 (1979)
15. Shaw, J.A.: Thermomechanical simulations of localized thermo-mechanical behavior in a NiTi shape memory alloy. Int. J. Plast. **16**, 541–562 (2000)
16. Mao, S.C., Luo, J.F., Zhang, Z., Wu, M.H., Liu, Y., Han, X.D.: EBSD studies of the stress-induced B2-B19′ martensitic transformation in NiTi tubes under uniaxial tension and compression. Acta Mater. **58**, 3357–3366 (2010)

Superhydrophobicity

- ▶ Lotus Effect

Superoleophobicity of Fish Scales

Lei Jiang[1] and Ling Lin[2]
[1]Center of Molecular Sciences, Institute of Chemistry Chinese Academy of Sciences, Beijing, People's Republic of China
[2]Beijing National Laboratory for Molecular Sciences (BNLMS), Key Laboratory of Organic Solids, Institute of Chemistry Chinese Academy of Sciences, Beijing, People's Republic of China

Synonyms

Oil-repellency of fish scales

Definition

Superoleophobicity of fish scales is a wetting phenomenon in which fish scales show oil repellency in water, with a static oil contact angle (CA) higher than 150°. Generally, underwater superoleophobicity is defined as a static oil CA higher than 150° on a solid surface in an oil/water/solid three-phase system.

Chemical and Physical Principles

Wettability is a fundamental property of solid surfaces, which not only affects the behavior of the creatures in nature but also plays an important role in all aspects of our life. A direct expression of wetting behavior is a static CA of a liquid droplet sitting on a solid surface when the surface tensions at multiphase interfaces reach thermodynamic equilibrium.

For an ideal flat surface in a liquid/gas/solid system, the CA (/θ/) is given by the Young's equation [1]:

$$\cos\theta = \frac{\lambda sg + \lambda sl}{\lambda lg} \tag{1}$$

where λsg is the solid/gas interface tension, λsl is the solid/liquid interface tension, and λlg is the liquid/gas interface tension.

For a rough surface in the air atmosphere, the situation is more complex. The surface topographic structure has a great influence on the wettability. Two distinct models, Wenzel's model and Cassie-Baxter's model, are commonly used to explain the effect of roughness on the apparent CAs of liquid drops. As described by Wenzel's model [2], liquid completely penetrates into the rough structures, leading to an increase in the contact area of the solid/liquid interface. Therefore, the apparent CA (θ_w) can be described as

$$\cos\theta_w = r\cos\theta \tag{2}$$

where the surface roughness factor (r) is defined as the ratio of the actual contact area of solid/liquid interface to the projected contact area ($r \geq 1$) and θ is the intrinsic CA on the flat surface. As described by Cassie-Baxter's model [3], liquid suspends on the rough surface rather than completely penetrates. The rough surface therefore can be considered as a solid/gas composite surface, and the apparent CA (θ_c) can be described as

$$\cos\theta_c = f\cos\theta - (1 - f) \tag{3}$$

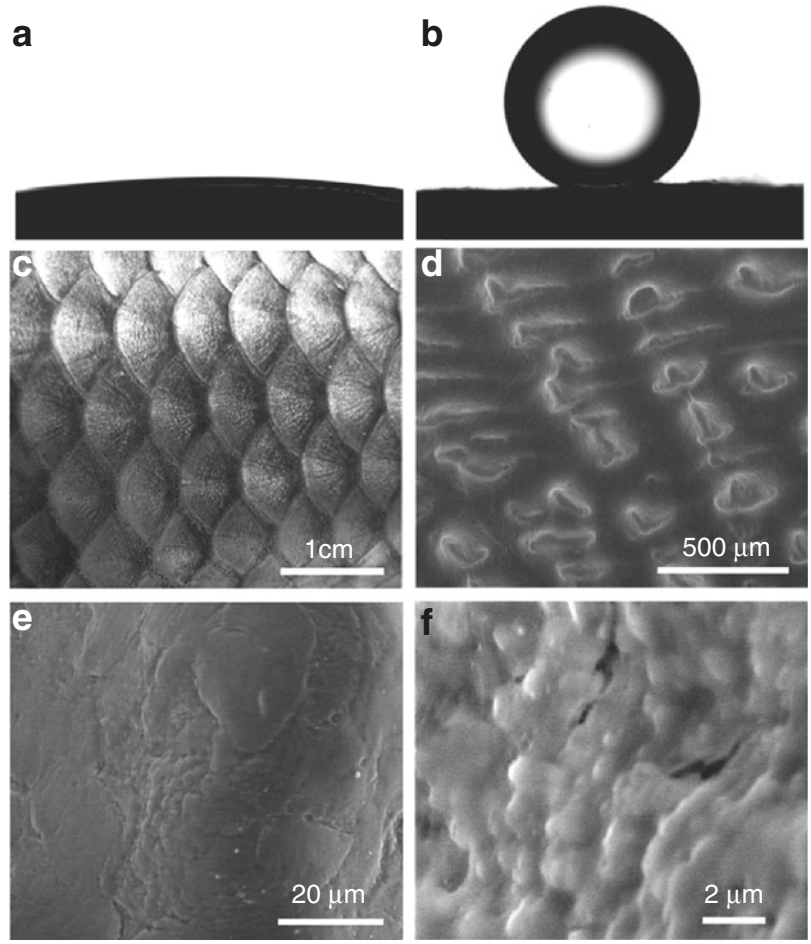
where f is the area fraction of the solid/liquid interface and $(1 - f)$ is that of the liquid/gas interface. Based on both theories, scientists can better understand different wetting phenomena in nature, which would help the design of artificial materials with functional surfaces [4–6].

Key Research Findings

Oil-Wetting Behaviors on Fish Scales with Micro-/Nanostructures

Nature abounds with mysterious living organisms which have special surface properties. A famous example is the lotus leaf, a typical superhydrophobic surface in nature. When a water droplet falls on the lotus leaf, it keeps its bead shape and rolls off the surface immediately taking away the adherent dirt particles. This so-called lotus effect has been understood as a result of the complementary roles of low-surface free energy and micro-/nanostructures on the leaf surface [8]. Inspired by this phenomenon, tremendous artificial superhydrophobic surfaces have been fabricated and have facilitated practical applications in broad fields [9–11].

In a similar manner to the lotus effect in the air atmosphere, fish can resist oil pollution in water, possessing an antifouling mechanism. This fascinating phenomenon shows great potential for many applications in an oil/water/solid system, such as marine antifouling, prevention of oil spills, microfluidic technology, and bioadhesion [12, 13]. Liu et al. first reported the oil-wetting behavior on fish scales in the water environment [7]. It is common knowledge that fish scales are covered by a thin layer of mucus that leads to their hydrophilic nature. Liu et al. revealed that the



Superoleophobicity of Fish Scales, Fig. 1 (**a**) Fish scales show superoleophilicity in air (1,2-dichloroethane (*DCE*) as a detecting oil, density 1.245 g cm^{-3}, surface tension at 25 °C 31.86 mN m^{-1}). (**b**) Fish scales become superoleophobic once they are immersed in water (droplet of DCE as in (**a**)). (**c**–**f**) SEM images of fish scales disclose the surface micro-/nanostructures with the increase of magnification (Reproduced with permission. Copyright Wiley-VCH Verlag GmbH & Co. KGaA (2009))

hydrophilic surface of fish scales showed superoleophilicity in air (Fig. 1a) [14]. However, the surface of fish scales turned to be superoleophobic once it was immersed in water, with an oil CA larger than 150° (Fig. 1b). To figure out the reason for oil-wetting reversion, they firstly observed surface structures on fish scales in detail. Figure 1c–f display typical images of fish scales (crucian carp, *Carassius carassius*) using scanning electronic microscopy (SEM). The fan-shaped fish scales with diameters of 4–5 mm are densely arranged (Fig. 1c). The magnified images disclose that there are oriented micropapillae on each fish scale. Each micropapilla is in the length of 100–300 µm and in width of 30–40 µm (Fig. 1d). In high-magnification SEM images (Fig. 1e, f), nanoscale roughness is clearly observed on the surface of micropapillae. It was suggested that these hierarchical structures could trap water and form a composite interface on fish scales to resist oil, which might play an important role in the oil-wetting reversion.

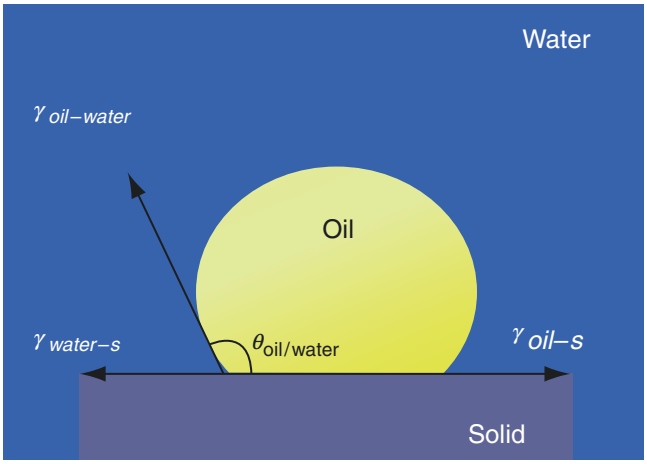
Mechanism of Wetting Behaviors on Fish Scales in Oil/Water/Solid System

To better understand the reversion of oil-wetting behavior on fish scales, Liu et al. chose three kinds of designed surfaces of silicon wafers as models: smooth surface, microstructured surface, and micro-/nanostructured surface [7]. Table 1 lists the contact angles of oil (DCE) or water on these three solid surfaces in the air or water environment, respectively. In air, the smooth surface is hydrophilic with a CA of 52.5 ± 1.4°, while microstructured and micro-/nanostructured

Superoleophobicity of Fish Scales, Table 1 Contact angles of oil (1,2-dichloroethane, DCE) or water on three solid surfaces with different surface structures in the air or water environment, respectively

Surfaces systems	Smooth silicon	Microstructured silicon	Micro-/nanostructured silicon
$C_2H_4Cl_2$ droplets (in air)	<5°	<5°	<5°
Water droplets (in air)	52.5 ± 1.4°	<5°	<5°
$C_2H_4Cl_2$ droplets (in water)	134.8 ± 1.6°	151.5 ± 1.8°	174.8 ± 2.3°

Source: Reproduced with permission. Copyright Wiley-VCH Verlag GmbH & Co. KGaA (2009)



Superoleophobicity of Fish Scales, Fig. 2 Schematic illustration of Young's equation in an oil/water/solid three-phase system, where a liquid droplet (*oil*) sits on a smooth surface (*solid*) in another liquid (*water*) phase

surfaces are both superhydrophilic (CAs < 5°). For the oil-wetting behaviors, all surfaces are superoleophilic, with the oil CAs smaller than 5°. However, in the water environment, all three kinds of surfaces become oleophobic or even superoleophobic. For the smooth surface, the oil CA is 134.8 ± 1.6°; for the microstructured surface, the oil CA is 151.5 ± 1.8°; and for the micro-/nanostructured surface, the oil CA is 174.8 ± 2.3°. The reversions of oil-wetting behaviors on these three surfaces are similar to that on fish scales. Comparing the different environments of two systems, the oil-wetting reversion is likely caused by the surrounding water. With these model silicon surfaces, the mechanism can be further analyzed through classical theories.

Although Young's equation was originally applied in a liquid/gas/solid three-phase system, it can be extended to an oil/water/solid system, in which an oil droplet sits on a solid surface in a water environment (Fig. 2). In this case, Young's equation is expressed as follows:

$$\cos\theta_{\mathrm{oil/water}} = \frac{\gamma_{\mathrm{water-s}} - \gamma_{\mathrm{oil-s}}}{\gamma_{\mathrm{oil-water}}} \tag{4}$$

where $\theta_{\mathrm{oil/water}}$ is an oil CA on solid surface in an oil/water/solid three-phase system, $\gamma_{\mathrm{water-s}}$ is the water/solid interface tension, $\gamma_{\mathrm{oil-s}}$ is the oil/solid interface tension, and $\gamma_{\mathrm{oil-water}}$ is the oil/water interface tension. Considering the liquid/gas/solid system, $\gamma_{\mathrm{oil-s}}$ and $\gamma_{\mathrm{water-s}}$ can be described as

$$\gamma_{\mathrm{s-g}} = \gamma_{\mathrm{oil-s}} + \gamma_{\mathrm{oil-g}} \cos\theta_{\mathrm{oil}} \tag{5}$$

$$\gamma_{\mathrm{s-g}} = \gamma_{\mathrm{water-s}} + \gamma_{\mathrm{water-g}} \cos\theta_{\mathrm{water}} \tag{6}$$

where θ_{oil}, θ_{water} is the oil or water CA in the air atmosphere, respectively, $\gamma_{\mathrm{oil-g}}$ is the oil/gas interface tension, and $\gamma_{\mathrm{water-g}}$ is the water/gas interface tension.

So Eq. 4 can also be expressed as follows:

$$\cos\theta_{oil/water} = \frac{\gamma_{oil-g}\cos\theta_{oil} - \gamma_{water-g}\cos\theta_{water}}{\gamma_{oil-water}} \quad (7)$$

Through Eq. 7, it can be easily understood why an oleophilic surface in air becomes oleophobic water. Taking DCE as an example, the surface tension ($\gamma_{oil\text{-}g}$) of DCE is 24.1 mN m^{-1}, the water surface tension ($\gamma_{water-g}$) is 73 mN m^{-1}, and the interfacial tension ($\gamma_{oil-water}$) of DCE/water is 28.1 mN m^{-1} [14]. As mentioned in the above experimental results, in air, the water CA (θ_{water}) on the smooth silicon surface is 52.5 ± 1.4° and the oil CA (θ_{oil}) nearly 0°. Therefore, it can be calculated that cos $\theta_{oil/water} = -0.72$ and $\theta_{oil/water} \approx 136°$. This result implies that the reversion of oil-wetting behavior certainly happens when different three-phase systems are involved, which is consistent with the experimental oil CA (134.8 ± 1.6°) in water.

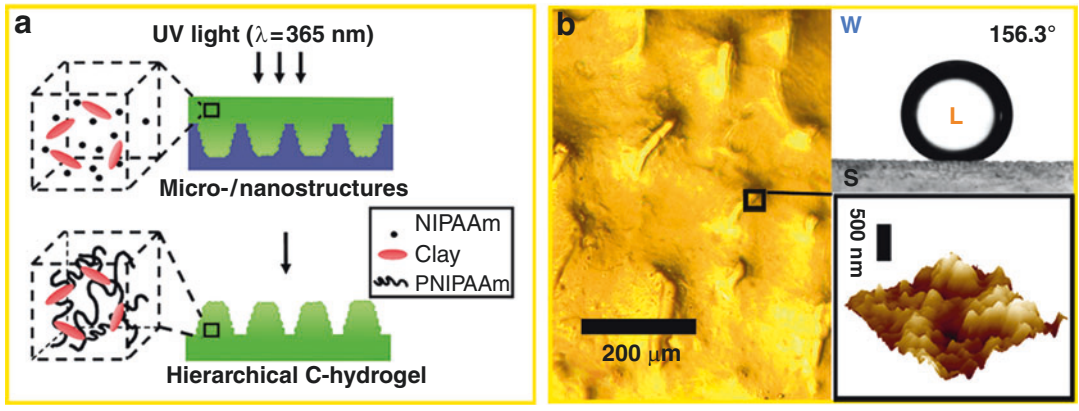
Moreover, when it comes to micro-/nanostructured surfaces, the oil CA is up to 174.8 ± 2.3° in the oil/water/solid system. This phenomenon is likely caused by the new forming interface on the micro-/nanostructured surface. According to the Cassie-Baxter's model, micro-/nanostructures could trap air forming a composite solid/gas surface in a liquid/gas/solid three-phase system. As to oil/water/solid three-phase system, hierarchical structures can also trap abundant water forming a composite water/solid interface; therefore, the apparent oil CA (cos $\theta'_{oil/water}$) in the water environment can be extended as

$$\cos\theta'_{oil/water} = f'\cos\theta_{oil/water} - (1 - f') \quad (8)$$

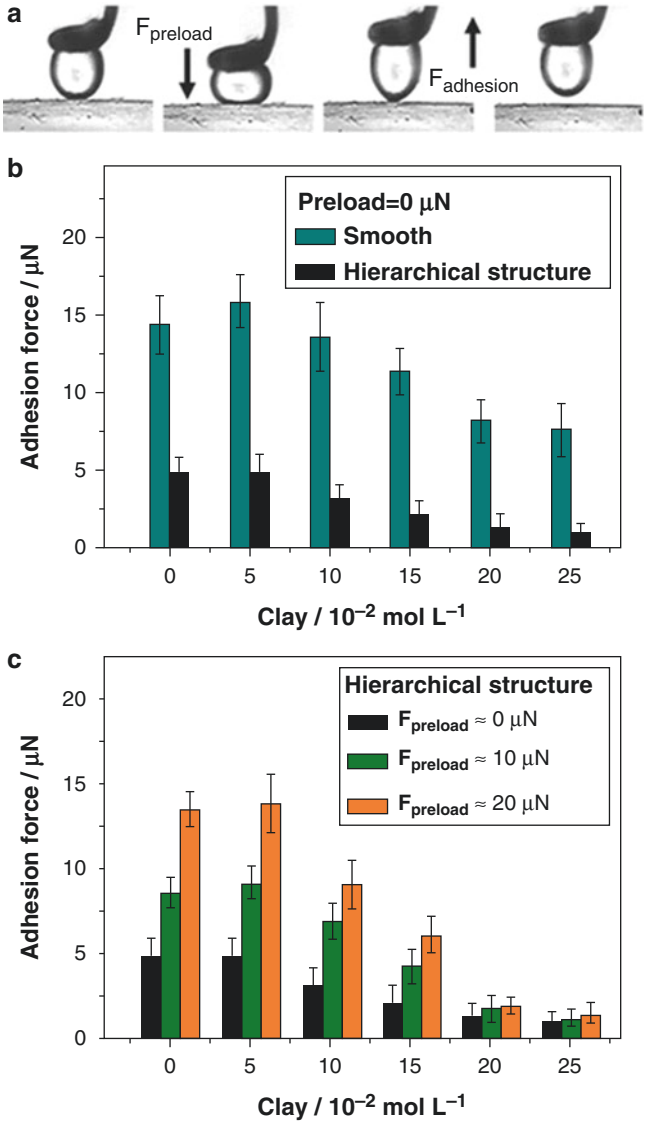
where f' is the area fraction of oil/solid interface, $(1 - f')$ is that of the oil/water interface, and $\theta_{oil/water}$ is the intrinsic CA of oil on a flat surface in oil/water/solid system. Once the composite surface forms, the oil droplet can rarely touch the solid surface, leading to the large increase of apparent oil CA. This effect endows the micro-/nanostructured silicon surface with superoleophobicity and as well contributes to the oil resistance of micro-/nanostructured fish scales.

Biomimetic Hydrogels for Robust Underwater Superoleophobicity

By understanding the fish oil-repellency nature, Lin et al. began to design a bionic artificial surface with robust underwater superoleophobicity [15]. From a practical perspective, the robustness of superoleophobic surface is crucial for underwater applications. Inspired by fish scales,



Superoleophobicity of Fish Scales, Fig. 3 Biomimetic design of PNIPAAm-nanoclay hydrogels (C-hydrogels) with hierarchical surface structures. (**a**) Schematic illustration of the fabricating process. (**b**) Optical and AFM images indicate micro-/nanostructured surface of fish scale replica which shows underwater superoleophilicity. S = hydrogel; L = 1, 2-dichloroethane (DCE); W = water (Reproduced with permission. Copyright Wiley-VCH Verlag GmbH & Co. KGaA (2010))



Superoleophobicity of Fish Scales, Fig. 4 Robustness of superoleophobicity on C-hydrogel surfaces measured by dynamic underwater oil-adhesion measurements (oil: DCE). (**a**) A measurement process. (**b**) Oil-adhesion forces on hierarchical surfaces are much lower than those on smooth surface (preload = 0). (**c**) As the preload increases, the oil-adhesion force increases only on the low clay content C-hydrogel, while C-hydrogel with high clay content retains ultralow oil adhesion (Reproduced with permission. Copyright Wiley-VCH Verlag GmbH & Co. KGaA (2010))

hydrophilic hydrogel, biophysically similar to mucus, was chosen to construct a fish scale-like surface with micro-/nanostructures. Meanwhile, hydrophilic clay (synthetic hectorite), a rigid nanolayered structure, was selected as a composite component to enhance mechanical strength. In the experiment, the mixture of /N/-isopropylacrylamide (NIPAAm) and nanoclay was poured onto a polydimethylsiloxane (PDMS) template which was firstly molded from dried fish scales (grass carp, *Ctenopharyngodon idella*) (Fig. 3a). Then, a photo-initiated in situ radical polymerization ($\gamma = 365$ nm, 40 min) was taken to fabricate hybrid PNIPAAm-nanoclay hydrogels (C-hydrogels) with hierarchical surface structures. Figure 3b shows that micro-/nanostructures are well molded on the surface of the fish scale replica. Furthermore, note that the surface exhibits superoleophobicity with a static oil (CA) of $156.3 \pm 1.4^\circ$ in an oil/water/solid three-phase system (DCE as the detecting oil). This result showed that hybrid hydrogels with hierarchical surface structures successfully mimicked the oil-repellent fish scales.

S

To further investigate the robustness of superoleophobicity in a complex water environment, underwater oil-adhesion experiments were dynamically measured by a high sensitivity microelectromechanical balance system [13]. In each process, an oil droplet (DCE) was squeezed against the surface with a constant preload and then released, during which the adhesion force between oil and surface was recorded (Fig. 4a). It was found that, in the case of no preload on surfaces, the adhesion forces between oil and hierarchical surfaces were much lower than those between oil and smooth surfaces (Fig. 4b). As the preload increased, the oil-adhesion behaviors on hierarchical surfaces changed. The adhesion force increased only on the surface of low clay content C-hydrogel, while the C-hydrogel surface with high clay content retained excellent low oil adhesion (Fig. 4c). These results were attributed to the synergetic effects of rigid nanoclays and flexible macromolecules. It is known that the PNIPAAm possesses hydrophilic macromolecular chains which can trap water to prevent oil adhesion. But this effect is weakened by the frangibility of hierarchical structures on the PNIPAAm surface. On the other hand, hybrid hydrogels with rigid nanoclays can enhance mechanical strength of surface micro-/nanostructures and, thus, keep the stability of trapped water on the surface. As a result, hybrid hydrogels with high clay content achieve robust underwater superoleophobicity. This study brings a new concept to the design and fabrication of underwater antifouling materials.

Future Directions for Research

The research on superoleophobic fish scales may open a new branch of the wettability field and make a strong impact on underwater applications. The studies in oil/water/solid three-phase systems have just begun, and many challenges remain on their way to development. First of all, the wetting theories in oil/water/solid system need to be further established. It is based on the tremendous wettability data of different materials with specific chemical components and surface structures. Secondly, inspired by fish scale's effect, underwater superoleophobic materials with environment-friendly, durable properties need be further explored, which show great potential for underwater antifouling application. Finally, in oil/water/solid system, intelligent responsive materials with switchable oil-wetting behavior will attract great attention due to the promising applications in artificial muscles, actuators, and sensors. In the future, learning from nature will be a primary principle to design biomimetic or bioinspired functional materials with special wettability.

Cross-References

- ▶ Biomimetics of Marine Adhesives
- ▶ Shark Skin Effect

References

1. Young, T.: An essay on the cohesion of fluids. Philos. Trans. R. Soc. Lond. A **95**, 65–87 (1805)
2. Wenzel, R.N.: Resistance of solid surfaces to wetting by water. Ind. Eng. Chem. **28**, 988–994 (1936)
3. Cassie, A.B..D., Baxter, S.: Wettability of porous surfaces. Trans. Faraday Soc. **40**, 0546–0550 (1944)
4. Öner, D., McCarthy, T.J.: Ultrahydrophobic surfaces. Effects of topography length scales on wettability. Langmuir **16**(20), 7777–7782 (2000)
5. Extrand, C.W.: Model for contact angles and hysteresis on rough and ultraphobic surfaces. Langmuir **18**(21), 7991–7999 (2002)
6. Quéré, D.: Wetting and roughness. Annu. Rev. Mater. Res. **38**, 71–99 (2008)
7. Liu, M.J., Wang, S.T., Wei, Z.X., Song, Y.L., Jiang, L.: Bioinspired design of a superoleophobic and low adhesive water/solid interface. Adv. Mater. **21**(6), 665–669 (2009)
8. Barthlott, W., Neinhuis, C.: Purity of the sacred lotus, or escape from contamination in biological surfaces. Planta **202**(1), 1–8 (1997)
9. Blossey, R.: Self-cleaning surfaces – virtual realities. Nat. Mater. **2**(5), 301–306 (2003)
10. Gennes, P.-G., Brochard-Wyart, F., Quéré, D.: Capillarity and Wetting Phenomena: Drops, Bubbles, Pearls, Waves. Springer, New York (2004)
11. Yao, X., Song, Y.L., Jiang, L.: Recent developments in bio-inspired special wettability. Adv. Mater. **23**, 719–734 (2011)
12. Nosonovsky, M., Bhushan, B.: Multiscale effects and capillary interactions in functional biomimetic

surfaces for energy conversion and green engineering. Philos. Trans. R. Soc. Lond. A **367**(1893), 1511–1539 (2009)
13. Liu, M.J., Zheng, Y.M., Zhai, J., Jiang, L.: Bioinspired super-antiwetting interfaces with special liquid–solid adhesion. Acc. Chem. Res. **43**(3), 368–377 (2010)
14. Lide, D.R.: CRC Handbook of Chemistry and Physics, 84th edn. CRC Press, Boca Raton (2003–2004)
15. Lin, L., Liu, M.J., Chen, L., Chen, P.P., Ma, J., Han, D., Jiang, L.: Bio-inspired hierarchical macromolecule-nanoclay hydrogels for robust underwater superoleophobicity. Adv. Mater. **22**(43), 4826–4830 (2010)

Superparamagnetism

▶ Magnetic Nanoparticles for Biomedical Applications

Support Loss

▶ Anchor Loss in MEMS/NEMS

Surface Dissipations in NEMS/MEMS

Jinling Yang
Institute of Semiconductors, Chinese Academy of Sciences, Beijing, People's Republic of China
State Key Laboratory of Transducer Technology, Shanghai, People's Republic of China

Synonyms

Surface loss in micromechanical/nanomechanical resonators; Surface loss in NEMS/MEMS; Mechanical energy dissipation

Definition

Surface dissipation is the mechanical energy loss caused by surface defects, such as dangling bonds, absorbates, and crystal termination defects. It becomes dominant as the dimensions of nanoelectromechanical systems (NEMS)/microelectromechanical systems (MEMS) resonators are reduced and the surface-to-volume ratio grows.

Overview

Nanoelectromechanical systems (NEMS)/microelectromechanical systems (MEMS) are systems integrating nanometer/micrometer-scale mechanical and electrical components. NEMS/MEMS resonators play an important role in viable commercial technologies and are becoming more and more prevalent in research applications; for example, micromechanical resonators are excellent transducers for force or mass detection [1, 2]. Advances in nanofabrication technology have enabled extreme miniaturization of resonant sensors. As tools for basic research, NEMS resonators have demonstrated extraordinary sensitivity to external forces, allowing the detection of electron spin flips [3], single molecules [4], and other fundamental phenomena [5, 6].

The fundamental characteristics of a mechanical resonator are determined by the resonance frequency and quality factor Q (inverse dissipation). Better device response is obtained by high frequency and high Q. Scaling down the dimensions is necessary for achieving high resonance frequencies. However, the expected gains in resonance frequencies from size reduction beyond the submicron scale have often been offset by degradation in quality factor. With the ever-increasing technical ability, it has become possible to miniaturize the resonator dimension to several tens of nanometer. Correspondingly, many different dissipation sources have been identified in the NEMS/MEMS device.

Dissipation effects appear in all mechanical systems and help define fundamental dynamical behavior. Understanding this issue is important, not only for improving the resonator mechanical properties but also for establishing their performance limits. Low dissipation is generally desirable, which makes a device more efficient and more sensitive and less susceptible to wear and mechanical noise [7]. Dissipation studies of

miniaturized resonators have been conducted for decades. The factors that degrade the Q value can be categorized as intrinsic or extrinsic. Extrinsic damping is attributed to interactions of the NEMS/MEMS with its surrounding environment, i.e., the gas molecules surrounding it, or the substrate on which it lies. Intrinsic damping results from flaws or defects in the NEMS/MEMS. There are essentially four major loss mechanisms for NEMS/MEMS. Extrinsic losses include loss in airflow and radiation of elastic wave at the support (support/clamping loss). Intrinsic losses are thermoelastic loss and surface loss [8].

Dissipation is a measure of energy lost per oscillation in the resonator. The Q-factor can be described either as the fullwidth at half-maximum of the measured resonance peak or the rate at which the NEMS/MEMS loses energy per vibrational period. The resonator Q is defined as $Q = 2\pi W_0/\Delta W$, where W_0 is the stored vibrational energy and ΔW is the total energy lost per cycle of vibration. ΔW can be written as $\Delta W = \Sigma_i \Delta W_i$, where ΔW_i represents the energy lost due to the various dissipation mechanisms. Every one of the abovementioned loss mechanism has an associated Q factor; the overall quality factor Q_{tot} can be found from [9]

$$\frac{1}{Q_{tot}} = \sum_i \frac{1}{Q_i}, \tag{1}$$

It is obvious that Q_{tot} cannot exceed the smallest Q_i.

As the resonators become thinner (or narrower), the surface-to-volume ratio grows. The surface contains a large amount of defects due to the lattice termination and surface impurities, which contribute significantly to damping in nanoresonators. Eventually, the surface properties play a dominant role in the dissipation over the bulk behaviors.

Key Principle

Surface Dissipation Mechanism

Cantilever is a typical NEMS/MEMS resonator; extensive studies have been done to the mechanical behavior of the cantilever [8, 10–12]. Hereafter, the NEMS/MEMS cantilever will be used to describe the general dissipation behavior of NEMS/MEMS resonator. When the cantilever thickness scales down, the surface-to-volume ratio increases, the surface loss becomes dominant. The surface loss is caused by absorbates or surface defects. The charge transfer between the surface and the absorbates, the Coulomb repulsion of the dipole moments associated with the adsorbate atoms, and the overlap of the wave function of absorbate atom orbits at a close distance would modify the surface stress and lead to surface loss [13–15].

The surface dissipation is modeled by considering the complex Young's modulus $E_c = E + iE_d$, where E_c and E_d are the complex value and the dissipation part of Young's modulus, respectively [12]. For a rectangular cantilever vibrating sinusoidally, the stored energy can be written as [16]

$$W_0 = \frac{1}{6} whE \int_0^l \varepsilon_{\max}^2(x)dx, \tag{2}$$

where l, w, and h are the beam length, width, and thickness, respectively, ε_{max} is the strain occurring on the top or bottom surface of beam during vibration. Considering the surface layer with thickness δ and complex modulus $E_{cs} = E_s + iE_{ds}$, the energy loss per cycle due to the surface layer is [12]

$$\Delta W_s = 2\pi\delta E_{ds}\left(w + \frac{h}{3}\right)\int_0^l \varepsilon_{\max}^2(x)dx, \tag{3}$$

where E_{cs}, E_s, and E_{ds} are the complex, conventional, and the dissipation values of the Young's modulus of the surface layer, respectively. Thus the surface loss related Q is

$$Q = \frac{wh}{2\delta(3w+h)}\frac{E}{E_{ds}}, \tag{4}$$

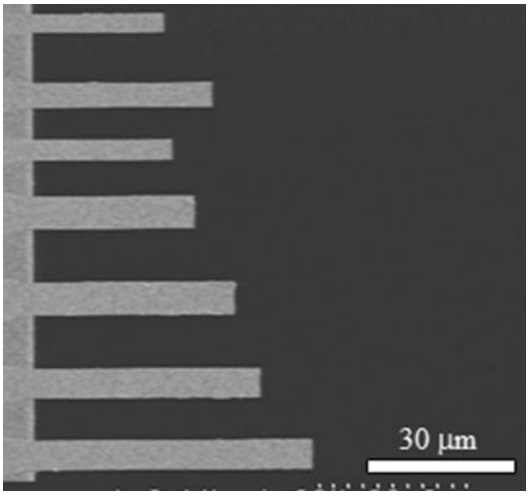
In Equation 4, E_{ds}, as a property of the absorbate layer and its defect, is closely related to the surface

stress and results in the surface loss. For a thin cantilever with $h << w$, the surface-loss related Q factor would be proportional to the thickness. Therefore, in the high surface-to-volume ratio structure, surface effect has become a crucial restriction for resonator miniaturization. To gain deep insight into this dissipation mechanism and its relationship with the dimensions of devices, the surface effect has been investigated via surface treatment in various conditions.

Key Research Findings

The cantilevers, with lengths of 5–120 μm, have been fabricated from (100)-oriented SIMOX (Separated by IMplanted OXygen) wafers with 60- (Fig. 1) and 170-nm-thick top Si layers, SOI (Silicon On Insulator) wafers with a 500-nm-thick top Si layer, and (110)-oriented SIMOX wafers with a 160-nm-thick top Si layer [10, 11]. Length to width ratio for all the cantilevers is about 10:1, as shown in Fig. 1. The cantilevers are actuated by piezo. The resonance of the cantilevers was measured by laser Doppler vibrometer. All measurements are performed at high vacuum ($<10^{-5}$ Torr) in order to avoid air damping.

Surface dissipation is proportional to the square of the peak strain integrated over the surface area of the beam. For the fundamental flexural mode of the cantilever, the Q factor would be independent of cantilever width, but proportional to the thickness because of the dominant surface dissipation [11]. Thus the surface effects can experimentally be quantified by changing the cantilever thickness. For very thin and long beams, surface dissipation is dominant until clamping losses increase sharply as the length is reduced. As shown in Fig. 2, for cantilevers with $l > 30$ μm, Q factor is almost proportional to the thickness, when the thickness of cantilevers is scaled up to 500 nm or down to 60 nm. But deviations from this relationship were found when $l < 30$ μm, suggesting that the support loss plays an important role in the short cantilevers [8]. The linear increase in the mechanical Q with the cantilever thickness was also observed in silicon–nitride cantilever of thicknesses 200 nm, 510 nm, 700 nm, and 1.2 μm [12].



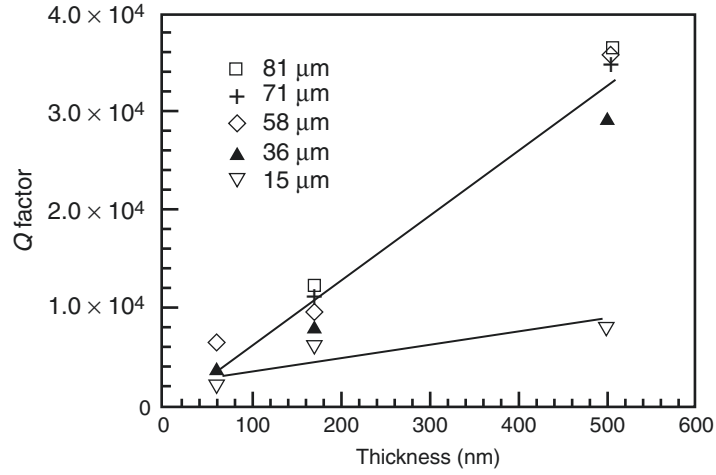
Surface Dissipations in NEMS/MEMS, Fig. 1 SEM image for cantilevers of 170 nm in thickness and of 10–100 μm in length

Surface Treatments at High Temperature

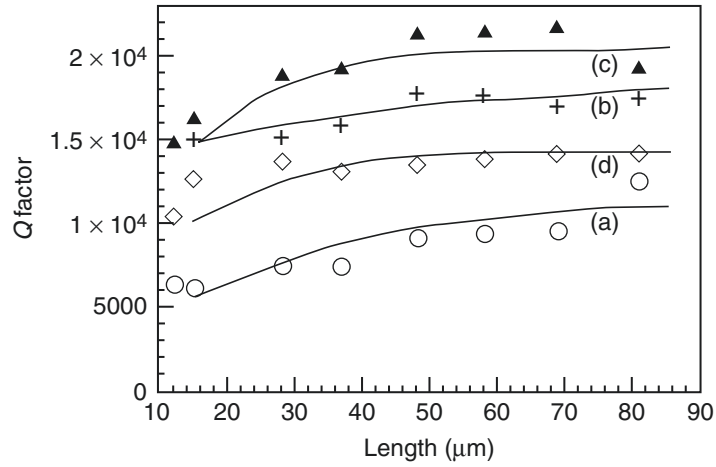
Surface treatments such as high-temperature annealing, which removes adsorbates and releases strain, or chemical treatments have significant effects on the quality factor, and improvements over one order of magnitude have been reported [8, 10–12].

It is well known that the clean Si(100) surface usually displays a 2 × 1 structure [17]. On these bulk-terminated surfaces, each Si atom has one or two dangling bonds (dbs), which determines the initial reactivity of the surface and a final surface structure. In addition, hydrogen plays an important role in silicon surface chemistry [17, 18]. Thus the surface stress could be modified by hydrogen dose under various conditions. And hydrogen-terminated surface is chemically inert and is not easily oxidized in atmosphere.

In atmosphere, the Si cantilever surface is covered by native SiO_2 and absorbates, which finally causes energy loss [10, 11]. The following treatment process was employed to modify the Si surface chemistry: firstly, annealing is done in the UHV chamber (1×10^{-10} mbar) to obtain a clean surface, that is, the cantilevers were outgassed at 600 °C for 30–60 min, then quickly



Surface Dissipations in NEMS/MEMS, Fig. 2 Q factors versus thickness for cantilevers with lengths of 15 μm, 36 μm, 58 μm, 71 μm, and 81 μm. All the measurements were done in the high vacuum system



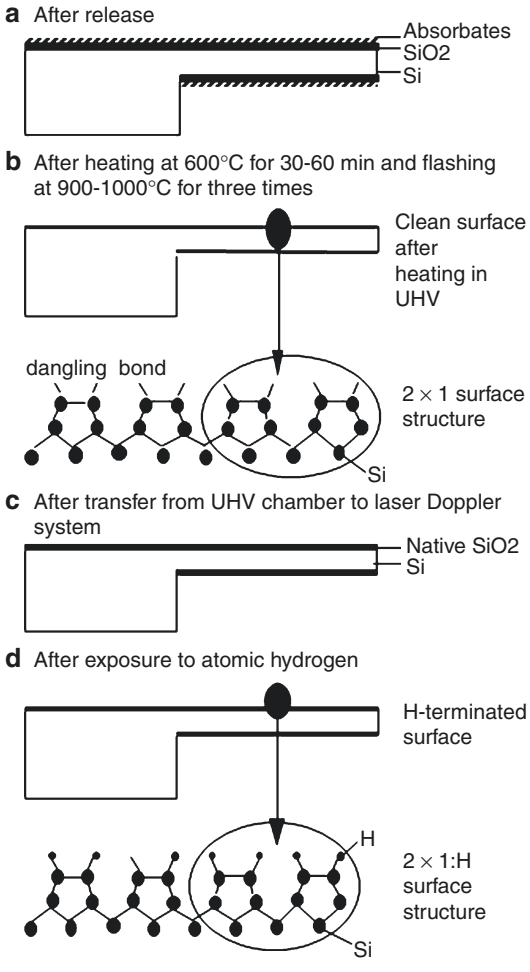
Surface Dissipations in NEMS/MEMS, Fig. 3 Length dependence of Q factors for 170-nm-thick cantilevers before (*a*) and after (*b*) heating at 600 °C for 1 h and flashing at 1000 °C for three times, after exposure to atomic hydrogen above 300 °C for 30 s (*c*) and at room temperature for 2 min (*d*). The surface treatment was performed in UHV setup, and all the measurements were done in the high vacuum system

flashed at 900°–1000 °C three times to remove SiO_2; secondly, exposure to atomic hydrogen is performed to passivate the clean surface, atomic hydrogen was produced by decomposition of molecule hydrogen on a 1500 °C tungsten filament. Exposure to atomic hydrogen with a pressure around 5×10^{-8} torr was accomplished at room temperature and above 300 °C for 30–120 s. After being cooled to room temperature in the UHV chamber, the cantilevers were transferred to a laser Doppler system for measurement. The time interval for this transfer was kept below 15 min to reduce the atmosphere effect on the surface.

As shown in Fig. 3, for 170-nm-thick Si (100) cantilever, after heating at 1000 °C, Q factors of all cantilevers clearly increase due to the removal of the SiO_2 layer and surface absorbates. After exposure to atomic hydrogen, the cantilever surface could further be modified, since hydrogen reacts with dbs on the Si dimer atoms, and the Si surface becomes inert, the growth of native SiO_2 and the absorption process during sample transfer slow down, thus, higher Q factors were achieved for the cantilevers (Fig. 3c). However, the extended exposure to atomic hydrogen for 2 min at room temperature results in serious etching of the cantilevers and deteriorates their Q factors very much (Fig. 3d). It is well established in [17, 18] that saturation exposure of a clean Si(100) 2 × 1 surface to atomic hydrogen at 300 °C results in the formation of a

monohydride surface, which preserves the Si dimer bonds and the associated 2 × 1 surface periodicity. Saturation exposure at lower temperature (room temperature) results in the 1 × 1 dihydride surface structure. The steric interaction between the neighboring dihydride units on the 1 × 1 surface produces strain which weakens the bonds within the units; as a result, these bonds are susceptible to further attack and etching by hydrogen atoms.

The surface modification process can be outlined as Fig. 4, annealing at 600 °C in UHV could remove some absorbates from the cantilever surface, e.g., H_2O and some organics. Subsequent flashing at 1000 °C for a short time will further clean the surface by desorbing the organics and SiO_2 layer and modify the surface structure; some of the surface structures may convert to Si(100) 2 × 1 with one dangling bond on each dimer atom. During transfer from the treatment chamber to the laser Doppler system, very thin native oxide and absorbate layer could be formed on the clean surface. Exposure to atomic hydrogen at above 300 °C is favorable for formation of Si(100) 2 × 1 : H monohydride surface structure with high stability and less stress. The considerable increase of the Q factor after exposure to atomic hydrogen could be a consequence of both surface modifications by hydrogen and the passivation effect.

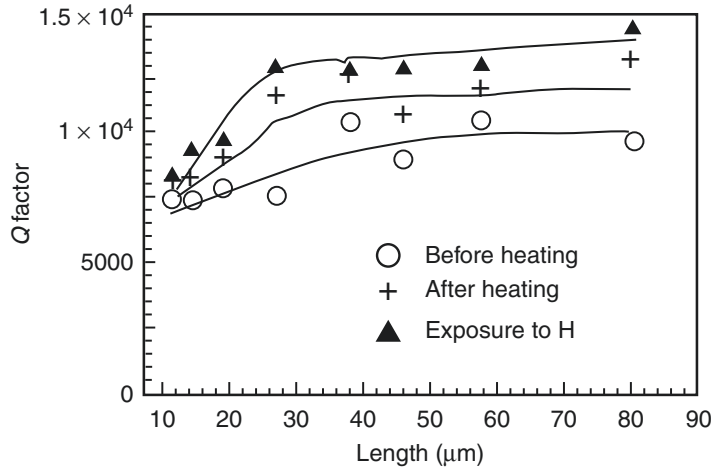


Surface Dissipations in NEMS/MEMS, Fig. 4 Schematic drawing of surface treatment for Si (100) resonators

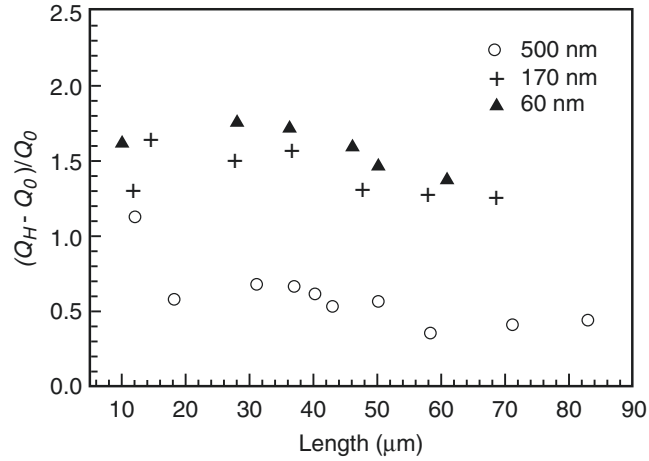
Surface dissipation is closely related to the crystallographic orientation with different surface defect densities. As shown in Fig. 5, the Q factors of the 160-nm-thick cantilevers made from (110)-oriented SOI wafer are comparable to the corresponding values of the 170-nm-thick cantilevers made from (100)-oriented SOI wafer. However, heating and exposure to atomic hydrogen increases the Q values of the 160-nm-thick cantilevers much less than the 170-nm-thick cantilevers. This distinction could be attributed to the different structures and defect density of two surfaces. After annealing at 1000 °C, the clean Si (110) surface could reconstruct into a 2 × 16 structure, and hydrogen exposure would form a bulk-like 2 × 16 : H surface structure [10]. The surface modification on Si(110) by annealing and exposure to atomic hydrogen is far different from the Si (100) structure.

Surface treatment has more influence on the Q factors of the thinner structures, as depicted in Fig. 6 for the relative change of Q values before and after exposure to atomic hydrogen $(Q_0\text{-}Q_H)/Q_0$ [10].

For clarifying how large the surface dissipation is in the ultrathin cantilever, the atmosphere effect must be completely eliminated. This is important not only for the improvement of the Q factor but also for a stable operation of the cantilever dynamically under a certain environment. Therefore, an UHV system with a preparation chamber for surface treatment and a laser Doppler chamber for in situ mechanical properties measurement



Surface Dissipations in NEMS/MEMS, Fig. 5 Q-factors versus length for (110)-oriented cantilevers before and after heating, and after exposure to atomic hydrogen

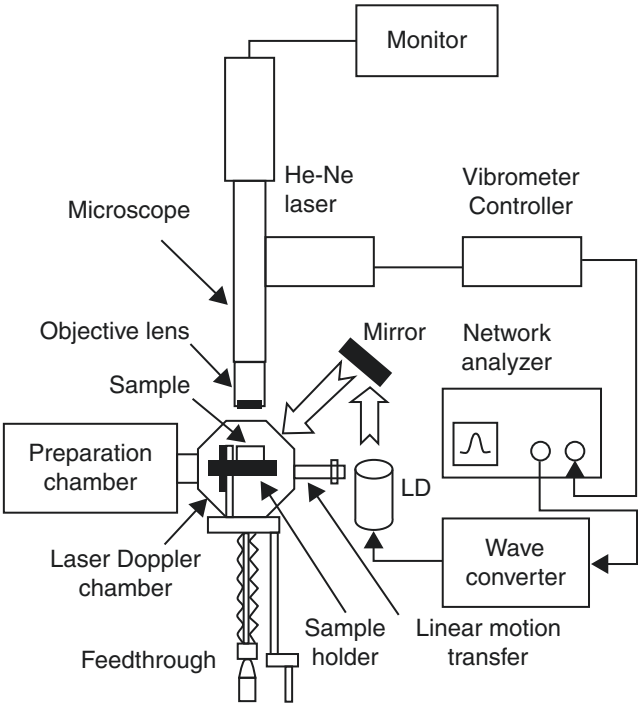


Surface Dissipations in NEMS/MEMS, Fig. 6 Length dependence of the relative change of Q factors before (Q_0) and after (Q_H) exposure to atomic hydrogen for cantilevers with thickness of 60 nm, 170 nm, and 500 nm. All the measurements were done in the high vacuum system

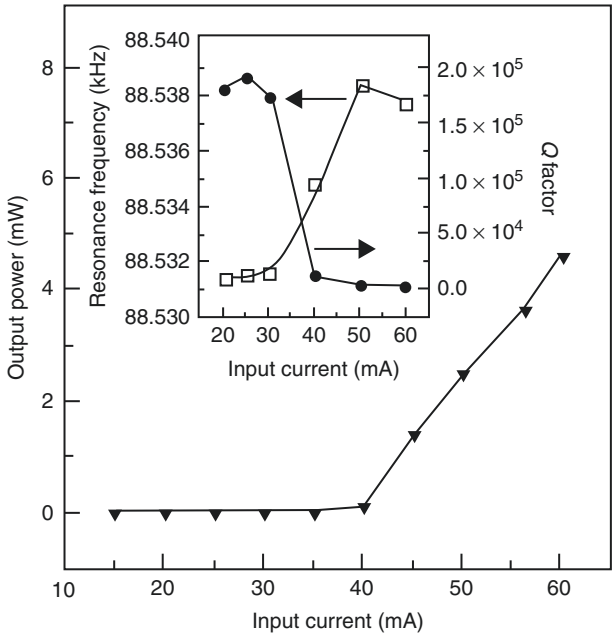
was employed [11], as shown in Fig. 7. The cantilever is optically actuated by a laser beam. The laser power dependence of the resonance frequency and the Q factor of a 48 μm long cantilever annealed at 1000 °C for 30 s was investigated and is shown in the inset of Fig. 8. A laser output power was limited to approximately 40 μW, which is enough to excite and maintain a stable resonant vibration for all cantilevers with different lengths and can ensure that the measurements were undisturbed by the "soft spring" effect.

The good vacuum condition in UHV system enables the clean surfaces of the cantilevers to be preserved during measurement. After annealing at 600 °C or 1000 °C in the preparation chamber, the sample was transferred to the laser Doppler chamber by a linear motion transfer. The mechanical resonance of the actuated cantilever is detected by the laser Doppler system and analyzed by a network analyzer, which also outputs a drive signal to modulate the pump frequency of the laser diode and thereby drives the cantilever into oscillation.

Figure 9 shows the length dependence of the Q factors for 170-nm-thick cantilevers under different treatments in the UHV chamber [11]. The freshly fabricated cantilevers of 30–90 μm long have the Q factors around 10^4, with a little dependence on their length. The shorter cantilevers ($L < 30$ μm) have the lower Q factors due to the obvious support loss [8]. Annealing at 600 °C



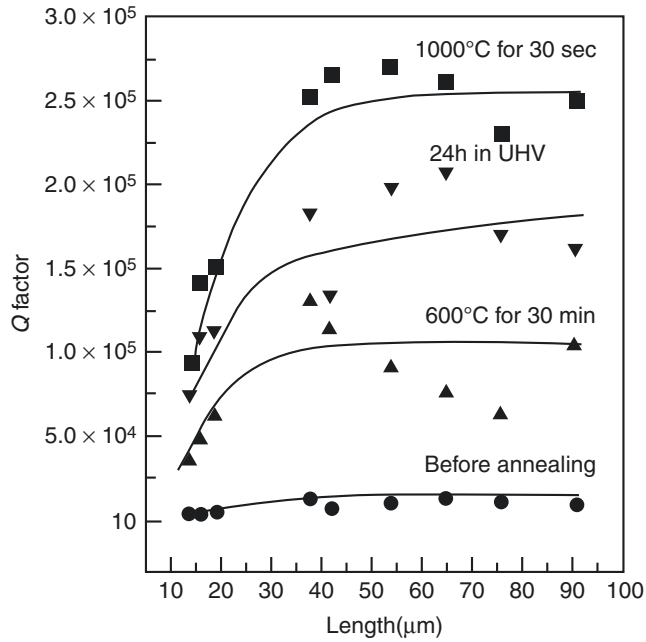
Surface Dissipations in NEMS/MEMS, Fig. 7 Schematic diagram of UHV system consisting of a preparation chamber, a measurement chamber, and a laser Doppler measurement loop. The cantilevers are actuated optically



Surface Dissipations in NEMS/MEMS, Fig. 8 Input current dependence of laser power before irradiation onto the UHV optic window. Inset is the measured resonance frequency and Q factor as a function of input current of LD for a cantilever of 48 μm long, 5 μm wide, and 170 nm thick. Soft spring effect induced sharp changes in the resonance frequency and Q factor can be clearly seen at 30–40 mA

for 30 min in UHV increases the Q factor by one order of magnitude (into the 10^5 range). Subsequent annealing at 1000 °C for 30 s causes a further increase of the Q values by a factor of 2–3. The improvement of Q factor was found to be associated with removing the absorbates and the deoxidization of the surface. Keeping the sample in UHV for about 24 h after annealing at 1000 °C obviously reduces the Q factors of all the cantilevers, which suggests that the change in

S



Surface Dissipations in NEMS/MEMS, Fig. 9 Q factors for a set of cantilevers fabricated on one chip, 170 nm in thickness and 13–90 μm in length. UHV annealing leads to significant increment on Q factors, up to $\sim 2.5 \times 10^5$ for cantilevers of 30–90 μm long annealed at 1000 °C for 30 s. All the annealing and measurements were done in the UHV system

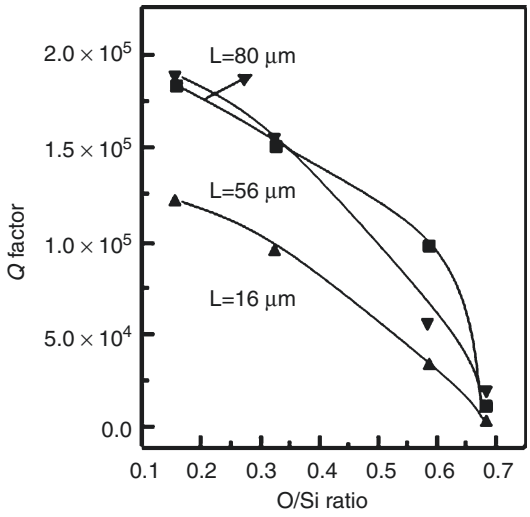
Q factor is caused by the surface modification rather than the annealing effect. These results indicate that the ultrathin cantilever is seriously subjected to the surface effect, that is, the surface treatment in UHV modifies the surface structure and thus the surface stress, which is related to the surface energy dissipation.

Moreover, in order to get further insight into the surface modification during annealing the samples, one large pattern of 400 × 400 μm^2 plus several microcantilevers of 170 nm thick and several tens of micrometers long were fabricated on one chip by the same process [11]. This patterned chip was annealed in UHV at 1000 °C for 15 s four times in a sequence. After each annealing, the surface of this square pattern was monitored by x-ray photoelectron spectroscopy (XPS) while the mechanical properties of the microcantilevers were examined. After annealing at 1000 °C for the first 15 s, no carbon peak could be detected, only photoelectron peaks of Si_{2S}, Si_{2P}, and O_{1S}, and Auger peak of O_{KLL} were observed. As shown in Fig. 10 for cantilevers of 16 μm, 56 μm, and 80 μm long, annealing at 1000 °C in UHV reduced oxygen concentration on the surface and simultaneously enhanced the Q values.

Chemical Passivation

Mechanical energy dissipation in micromechanical silicon structures is sensitive to the chemical state of the surface. By changing a single monolayer of molecules with less than 0.07 % of the total mass on the surface of a 5-μm-wide, 250-nm-thick Si(111) resonator, the resonator's quality factor can be improved by at least 70 %. When a single monolayer of hydrogen atoms is replaced by 13 Å of silicon oxide, corresponding to the oxidation of less than two silicon bilayers, which is standard for commercial silicon devices, mechanical energy dissipation increases significantly: at least 75 % of the energy in the oxide-terminated resonators is dissipated at the surface [19].

The mechanical properties of coated resonators are correlated with the chemical passivation properties of the monolayer. High-density monolayers (methyl groups) that resist the formation of electronic defects have the best performance. Resonators terminated with these passivation monolayers are also very stable in both vacuum and air due to their relatively low densities of electronic defects as well as the resistance to chemical attack [20].



Surface Dissipations in NEMS/MEMS, Fig. 10 Monotonic dependence of Q factors on O/Si ratio for three cantilevers of 170 nm thick, 16 μm, 56 μm, and 80 μm long, respectively, which are determined by in situ laser Doppler measurement and XPS after being annealed at 1000 °C for 15 s, 30 s, 45 s, and 60 s

Reducing Surface Dissipation

To improve device performance of NEMS/MEMS, methods have been developed to reduce surface dissipation. Here, we present two chemical methods that can have a profound influence on the dissipative nature of NEMS/MEMS structures. The first is a method where surface dissipation is reduced through annealing. The second is passivation of the surface with submonolayer.

Annealing at High Temperature

The atoms that lie at the surfaces of the NEMS/MEMS have fewer bonding neighbors than atoms in the bulk and are susceptible to surface contamination, which will reduce the quality factor of a NEMS/MEMS resonator. For silicon resonators, annealing in various conditions has proven to be effective for reducing the surface dissipation. Treatment at high temperature in UHV chamber was able to remove the oxide and absorbates on the resonator surface, modify the surface stress, and improve the Q values beyond 105 for 170-nm-thick cantilever. As expected, the thin structures are more sensitive to surface degradation and treatment than the thicker structures [10].

High-temperature annealing the 70 nm and 170-nm-thick single-crystal silicon cantilevers at 700 °C for 1 h in forming gas (Ar with 4.25 % H) and in a nitrogen atmosphere produced increases about a factor of two and three, respectively [12].

Passivating Resonator Surface

The clean Si(100) surface usually displays a 2 × 1 structure, each Si atom has one or two dangling bonds (dbs), which determines the initial reactivity of the surface and a final surface structure. The surface of the freshly fabricated resonators or the clean surface after annealing in UHV is easily subjected to surface contamination and energy loss in atmosphere, even in UHV the Q factors of the cantilevers degrades with time. Passivating surface atoms to make their bonding environments more bulk-like has resulted in higher Q factors [8, 19, 20].

Mechanical energy dissipation in a micromechanical silicon resonator is sensitive to submonolayer changes in surface chemistry. Hydrogen plays an important role in silicon surface chemistry and hydrogen-terminated surface is chemically inert. The effect of different alkenes and hydrogen surface termination on silicon resonators was studied [20]. The rate of mechanical energy dissipation and the density of electronically active defects on freshly prepared Si(111) resonators and functionalized surfaces follow the trend: silicon oxide >> long-chain alkyl > H > CH_3. After an extended period of air exposure, H-terminated surfaces degrade significantly, leading to silicon oxide > H > long-chain alkyl > CH_3. Si(111) resonators terminated by a single monolayer of methyl groups have significantly higher quality factors, and thus lower rates of mechanical energy dissipation, than those terminated with either long-chain alkyl monolayers ($-C_nH_{2n+1}$, n = 2–18) or hydrogen monolayers.

In summary, scaling down the size of the NEMS/MEMS resonators does not necessarily

mean the suppression of Q factor and that a higher Q factor and thus higher sensitivity are achievable if the proper surface treatment and passivation process are adopted for the NEMS/MEMS resonators.

Cross-References

- ▶ Microcantilever Chemical and Biological Sensors
- ▶ MicroElectroMechanical Systems
- ▶ NEMS
- ▶ NEMS Resonant Mass Sensors

References

1. Rugar, D., Zuger, O., Hoen, S., Yannoni, C.S., Vieth, H.M., Kendrick, R.D.: Force detection of nuclear magnetic resonance. Science **264**, 1560–1563 (1994)
2. Huang, X., Feng, X., Zorman, C., Mehregany, M., Roukes, M.: VHF, UHF and microwave frequency nanomechanical resonators. New J. Phys. **7**, 247 (2005)
3. Zolfagharkhani, G., Gaidarzhy, A., Degiovanni, P., Kettemann, S., Fulde, P., Mohanty, P.: Nanomechanical detection of itinerant electron spin flip. Nat. Nanotechnol. **3**, 720–723 (2008)
4. Naik, A., Hanay, M., Hiebert, W., Feng, X., Roukes, M.: Towards single-molecule nanomechanical mass spectrometry. Nat. Nanotechnol. **4**, 445–450 (2009)
5. Wu, G., Ji, H., Hansen, K., Thundat, T., Datar, R., Cote, R., Hagan, M., Chakraborty, A., Majumdar, A.: Origin of nanomechanical cantilever motion generated from biomolecular interactions. Proc. Natl. Acad. Sci. **98**, 1560–1564 (2001)
6. Montemagno, C., Bachand, G.: Constructing nanomechanical devices powered by biomolecular motors. Nanotechnology **10**, 225–231 (1999)
7. Cleland, A.: Themomechanical noise limits on parametric sensing with nanomechanical resonators. New J. Phys. **7**, 235 (2005)
8. Yang, J.L., Ono, T., Esashi, M.: Energy dissipation in submicrometer thick single–crystal silicon cantilevers. IEEE J. Microelectromech. Syst. **11**, 775–783 (2002)
9. Stemme, G.: Resonant silicon sensors. J. Micromech. Microeng. **1**, 113–125 (1991)
10. Yang, J.L., Ono, T., Esashi, M.: Investigating surface stress: surface loss in ultrathin single-crystal silicon cantilevers. J. Vac. Sci. Technol. B **19**, 551–556 (2001)
11. Yang, J.L., Ono, T., Esashi, M.: Surface effects and high quality factors in ultrathin single-crystal silicon cantilevers. Appl. Phys. Lett. **77**, 3860–3862 (2000)
12. Yasumura, K.Y., Stowe, T.D., Chow, E.M., Pfafman, T., Kenny, T.W., Stipe, B.C., Rugar, D.: Quality factors in micro- and submicron- thick cantilevers. J. Microelectromech. Syst. **9**, 117–125 (2000)
13. Ibach, H.: Adsorbate-induced surface stress. J. Vac. Sci. Technol. **A12**, 2240–2245 (1994)
14. Ibach, H.: The role of surface stress in reconstruction, epitaxial growth and stabilization of mesoscopic structures. Surf. Sci. Rep. **29**, 193–263 (1997)
15. Grossmann, A., Erley, W., Hannon, J.B., Ibach, H.: Giant surface stress in heteroepitaxial films: invalidation of a classical rule in epitaxy. Phys. Rev. Lett. **77**, 127–130 (1996)
16. Nowick, A.S., Berry, B.S.: Anelastic Relaxation in Crystalline Materials. Academic, New York (1972)
17. Boland, J.J.: Structure of H-saturated Si(100) surface. Phys. Rev. Lett. **65**, 3325–3328 (1990)
18. Boland, J.J.: Role of bond-strain in the chemistry of hydrogen on the Si(100) surface. Surf. Sci. **261**, 17–28 (1992)
19. Wang, Y., Henry, J., Sengupta, D., Hines, M.: Methyl monolayers suppress mechanical energy dissipation in micromechanical silicon resonators. Appl. Phys. Lett. **85**, 5736–5738 (2004)
20. Henry, J., Wang, Y., Sengupta, D., Hines, M.: Understanding the effects of surface chemistry on q: mechanical energy dissipation in alkyl-terminated (c1–c18) micromechanical silicon resonators. J. Phys. Chem. B **111**, 88–94 (2007)

Surface Electronic Structure

Regina Ragan
Chemical Engineering and Materials, The Henry Samueli School of Engineering, Science
University of California, Irvine, CA, USA

Synonyms

Local density of states

Definition

Surface electronic structure is defined by filled and empty electronic states of the system near the surface of a solid material. In both bulk systems and nanosystems, the type of atoms in the system, the atomic arrangement, and atomic order versus disorder influence electronic structure.

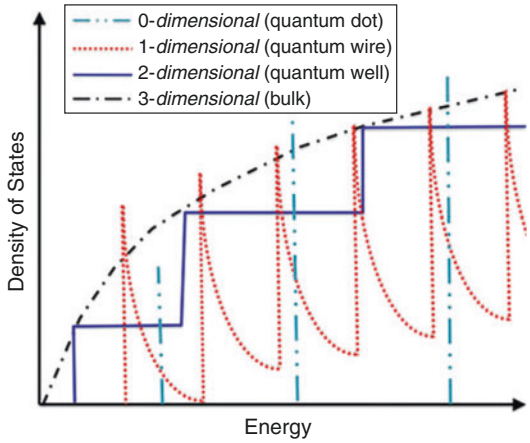
Tight Binding, an early method for calculating electronic structure using a single particle Hamiltonian, approximates the electronic wave functions in a crystal as a linear combination of the atomic wave functions of constituent atoms. Tight Binding analysis shows that electrons in solid materials exhibit collective behavior that deviates from discrete electronic states in isolated atoms. Coupling of atomic wave functions in crystals leads to a continuum of energy states, called energy bands. The electronic structure at the surface often differs from that in the bulk due to broken bonds and atomic surface reconstructions. If the surface to volume ratio is high, as in nanosystems, then surface effects can dominate observed electronic properties.

An important physical parameter to characterize electronic structure is the Fermi energy. The Fermi energy is often referred to as the highest occupied energy state since the probability of electron occupation of states having higher energy than the Fermi energy decays exponentially at equilibrium as determined from the Fermi-Dirac distribution function. The Fermi energy level is typically measured indirectly by the work function of the material. The work function is the energy needed to remove an electron from a material and is equal to the difference between the Fermi energy and the vacuum energy level. The position of the Fermi energy with respect to the energy bands determines whether the material is a metal, semiconductor, or insulator. Consider a simplified intuitive view. If atoms in the crystal have an odd number of valence electrons, then the energy bands that arise due to splitting of atomic energy levels are half filled. In this case, as found in metals, there is a continuum of available electronic states near the Fermi energy. A low resistance to electron motion is exhibited and thus metals are good conductors. In comparison, from this simple view, if there is an even number of valence electrons in constituent atoms of the crystal, then the (valence) energy bands are completely filled at zero Kelvin and the Fermi energy is in the energy bandgap. Near zero Kelvin, in both semiconductors and insulators, electrons do not occupy states in the conduction band and thus will exhibit negligible conductivity. Near room temperature, semiconductors have few electrons at energy levels available for conduction.

In the case of nanosystems, the number of atoms in the system also affects electronic structure due to quantum confinement of electrons. In metals the "space"an electron occupies in the crystal is defined by the Fermi wavelength that has an inverse relationship with the electron density. In semiconductors, where charge carriers are electron and holes, the Bohr exciton radius defines the "space" for an electron hole pair. The Bohr exciton radius is a function of the dielectric constant of the material and the effective mass of the charge carriers. It is typically much larger than the Fermi wavelength in magnitude. When dimensions of nanoscale systems decrease below the Fermi wavelength (metals) or Bohr exciton radius (semiconductors) unique physical properties not observed in bulk materials arise due to quantum-size effects.

Overview

Drude theory provides a simple, intuitive analysis of the collective behavior of electrons in metallic systems. After J.J. Thompson discovered the electron in 1897, Drude treated valence electrons in metals as an electron gas and applied the kinetic theory of gases to describe electrical and thermal conductivity in metals. In this model, valence electrons are assumed to move freely in the solid and the potential exerted on electrons in the crystal lattice is assumed to be uniform due to nondirectional bonding of atoms in the crystal (metals have high coordination numbers). Electrons are only assumed to interact with ion cores during finite scattering events. Thus, the model is called the free electron model and the approximation for finite scattering events is called the relaxation time approximation. Although Drude theory only works well for alkali metals, it provides a qualitative understanding of thermal and electrical conductivity in metals. Early on Sommerfeld addressed some of the shortcomings of Drude theory to evaluate thermal conductivity by using Fermi-Dirac statistics to model the velocity distributions of electrons in metals in the context of



Surface Electronic Structure, Fig. 1 Schematic of density of states for three-dimensional, bulk system (*black dot-dashed curve*), two-dimensional, quantum well (*blue solid curve*), one-dimensional, quantum wire (*red dashed curve*), zero-dimensional, quantum dot (*cyan double dot-dashed curve*)

Drude theory. Quantum mechanical corrections are needed to accurately estimate the number of electrons that contribute to conductivity in most metals. Despite the need for these corrections, the approach of the electron gas to model observed electrical, thermal, and optical properties in materials was pivotal in our understanding of experimental observations in metals and semiconductors. The density of states, that is the number of available states per unit energy at a particular energy level, is derived from the electron gas approach. A schematic of how the density of states varies between bulk systems and quantum-confined systems is shown in Fig. 1. Van Hove singularities (discontinuity in the density of states) can be seen in quantum-confined systems. The concept of an electron gas is still used and describes quantum-confined systems such as metallic, single-walled carbon nanotubes (*one-dimensional* electron gas) that exhibit Van Hove singularities [1] and graphene (*two-dimensional* electron gas) that exhibits the quantum hall effect where the Hall conductivity exhibits quantized values [2].

When analyzing systems of atoms that do not form metallic bonds, ionic, and covalent systems, the approximations that bonding is nondirectional and the potential is uniform is no longer valid. Kronig and Penny provided early intuition of energy levels and bands in crystals using both Bloch's theorem and a simple one-dimensional, periodic square wave potential that roughly approximates the potential of atoms in a periodic crystalline system. Via this simple analysis it is found that there are energy levels that yield nonphysical solutions for the electron wave functions and thus are not allowed for electrons in the system. These forbidden energy levels are defined as an energy bandgap. If the Fermi energy is in the bandgap, there are few electrons occupying energy states in the conduction band (above the Fermi energy) and thus conductivity is lower than in metals, hence the name semiconductor. As mentioned prior, the main difference between metals and semiconductors/insulators is that the Fermi energy sits within a band in a metal and in the energy bandgap in a semiconductor/insulator. The main difference between semiconductors and insulators is the magnitude of the bandgap. If greater than approximately 4 eV, then the probability of electrons occupying states in the conduction band is negligible at room temperature.

Basic Methodology

Computational Approaches

Significant advances in computational abilities coupled with important fundamental physical simplifications have allowed for first principles, ab initio, calculations of electronic structures of many body systems more closely modeling solid macroscopic and nanosystems and thus increasing accuracy. Since the Schrodinger equation cannot be solved analytically for these many electron systems, it was pivotal that Hohenberg and Kohn proved that the ground state energy of a many body system is a unique function of the charge density distribution [3]. Charge density distributions have a relationship with the atomic structure and periodicity in the crystal and thus can be evaluated by the types of atoms in the crystal and the crystal structure (atomic arrangement). Later, using an exchange correlation potential acquired from the theory of a homogeneous electron gas, Kohn and Sham determined how a many

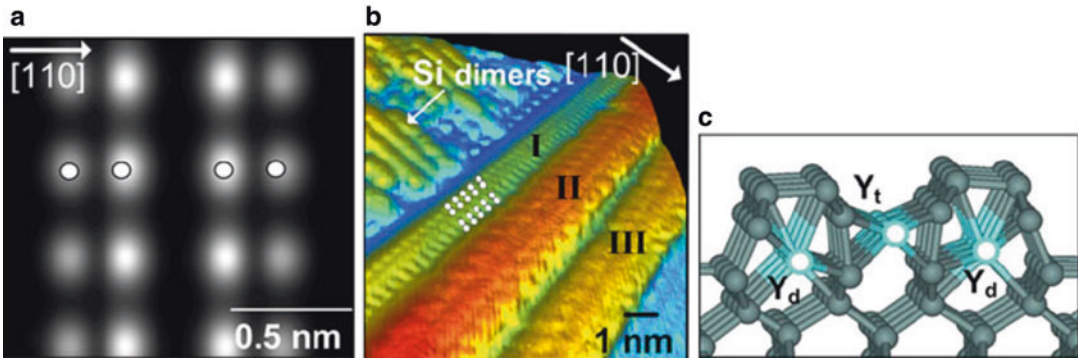
body system could be reduced to a single particle equation for input into the Schrodinger equation. This is referred to as the local density approximation [3].

Due to these critical advances in understanding of many body systems, density functional theory (DFT) is widely used to calculate electronic structure that can be directly compared to experimental systems for understanding of experimental observations or used to predict physical properties to guide experiments. For example, DFT provided fundamental understanding of low-dimensional magnetism such as giant magnetic moments found in two-dimensional systems, such as monolayers of Mn and Cr. Giant magnetic moments arise due to an increase in the density of electronic states near the Fermi energy [3]. The phenomenon of Giant Magnetoresistance allowed for an increase in storage density in magnetic hard disk drives. Furthermore, low-dimensional metallic nanowires and nanoparticles can exhibit ballistic transport or unique chemical activity, respectively, and typically require feature sizes smaller than achievable with lithography. Thus self-organization is needed for fabrication. Atomic arrangements and driving forces for self-organization of low-dimensional metallic systems can be determined using atomic scale imaging in conjunction with DFT [4, 5]. Figure 2 shows how the correlation between DFT simulations and scanning tunneling microscopy (STM) measurements allowed for an understanding of the atomic structure in disilicide nanowires that exhibit *one-dimensional* electron transport [4, 5]. The correct atomic arrangement is also critical for understanding the charge density distribution and in turn the surface electronic structure that is relevant, for example, to the development of nanocatalysts [6]. Improving efficiency and selectivity of catalysts will have significant economic benefit for the chemical industry. DFT has also been used to design electrode materials for lithium ion batteries. For example, Meng et al. have correlated structure with performance in electrodes using ab initio methods [7]. Materials design for batteries is crucial for development of plug-in electric vehicles that meet consumer performance requirements. Overall DFT allows for fundamental understanding of the relationship between atomic structure and electronic structure that is critical for understanding and utilizing physical properties of nanoscale systems.

Measurement Techniques

Photoelectron Spectroscopy

Photoelectron spectroscopy (PES) is a traditional surface analysis technique that uses a photon beam to provide energy for electrons to escape the potential of atoms or molecules near the surface. Kinetic energy distributions of emitted



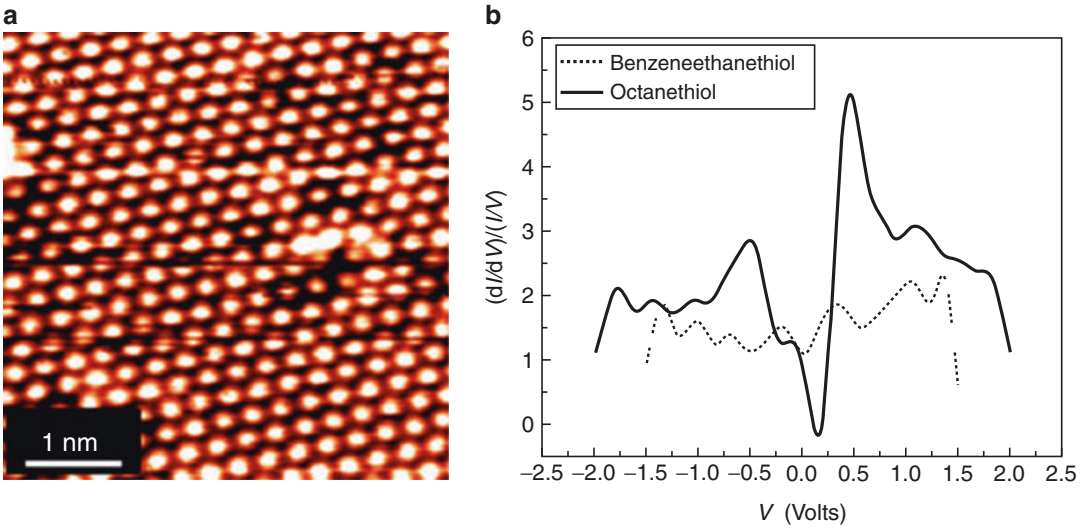
Surface Electronic Structure, Fig. 2 (**a**) DFT simulation of the calculated electronic structure on the surface of an yttrium disilicide nanowire having a width of 1.1 nm. (**b**) STM image of dysprosium disilicide nanowires on Si (001) substrate. Nanowire labeled I has a width of 1.1 nm, the calculated surface electronic structure matches the STM image as indicated by the white round circles. (**c**) Cross-sectional view of the relaxed atomic struture in the disilicide nanowires determined from DFT calculations and STM images (Printed with permission from Ref. [4])

photoelectrons provide information regarding ionization energies or work function of the surface that reflects the composition and electronic states on the surface. Typical methods are x-ray photoelectron spectroscopy (XPS) that uses soft X-rays with energy on the order of 200–2,000 eV and ultraviolet photoelectron spectroscopy (UPS) that uses ultraviolet light with energy in the range of 10–45 eV. The energy of the photon beam affects the type of electrons that are emitted from the sample. During XPS core electrons are emitted from the atoms and during UPS the energy is sufficient to emit only valence electrons. Angle resolved photoemission spectroscopy (ARPES) provides additional information about the momentum of the emitted electrons. Since momentum is conserved, one can obtain the energy-momentum relationship (called the dispersion relationship) of the electrons in the crystal that reflects the energy band structure as a function of crystallographic direction. ARPES has been an important tool to measure the electronic band structure in bulk materials and in quantum-confined systems. For example, Yeom et al. measured a charge density wave in linear chains of indium atoms on silicon (001) surfaces [8]. Charge density waves are a coupling between electrons and lattice vibrations that exhibit a periodic modulation of charge that can be observed on a *one-dimensional* metallic surface. In the same study, a measured temperature dependent metal to semiconductor transition for in atomic chains on silicon (001) was attributed to a Peierls transition that is another signature of *one-dimensional* quantum confinement.

Scanning Probe Microscopy

Scanning probe microscopy (SPM) includes techniques such as scanning tunneling microscopy/spectroscopy (STM/STS), atomic force microscopy (AFM), electrostatic force microscopy (EFM), and Kelvin probe force microscopy (KPFM) that can be used to measure atomic structure and/or surface electronic structure. STM, invented in 1981 by Binnig and Rohrer, probes both atomic and electronic structure on surfaces by measuring tunneling current from the probe tip to sample surface across a narrow vacuum (dielectric) gap. STM measurements typically achieve atomic scale spatial resolution. The AFM was invented shortly after the STM, 1986, to measure topography on nonconducting surfaces and is also capable of atomic and molecular resolution on surfaces. AFM measures van der Waals and electrostatic forces between cantilever tip and surface and thus does not require a conductive sample surface. KPFM is a variant of AFM in which conducting tips are used; KPFM allows for determination of the local surface potential with nanometer spatial resolution.

Tunneling current, as measured in STM, is extremely sensitive to both electron density and surface topography and convolutes the two properties. The tip-sample polarity affects whether electrons tunnel from sample to tip or from tip to sample and thus determines if the measurement probes filled or empty surface electronic states, respectively. A combination or empty and filled states imaging is often used in order to deconvolute the surface electronic structure from the atomic structure. STM can resolve corrugations heights on the sub-angstrom level, i.e., 0.2 Å corrugations between atoms on a clean platinum surface have been measured. The atomic arrangement on clean Pt(111) and how the surface structure evolves after depositing self-assembled monolayers on the surface have been measured using STM [9]. Figure 3a shows an STM image of Pt(111) where the hexagonal close packing of the atoms on the surface is easily observed. Scanning tunneling spectroscopy (STS) is a derivative of STM that measures the density of electronic states in the vicinity of the atomic scale STM tip, again yielding high spatial resolution. During STS, the current–voltage spectrum is measured between tip and surface while the voltage is swept across a specified range usually at constant tip-to-sample distance. Normalization of the differential conductance (dI/dV) with the conductance (I/V) can reflect the local density of states near the Fermi energy. Figure 3b shows normalized differential conductance data for benzene ethanethiol and octanethiol self-assembled monolayers on Pt (111) determined from STS data. Note that the Fermi energy is set at zero volts on the x-axis. The octanethiol/Pt(111) junction has zero

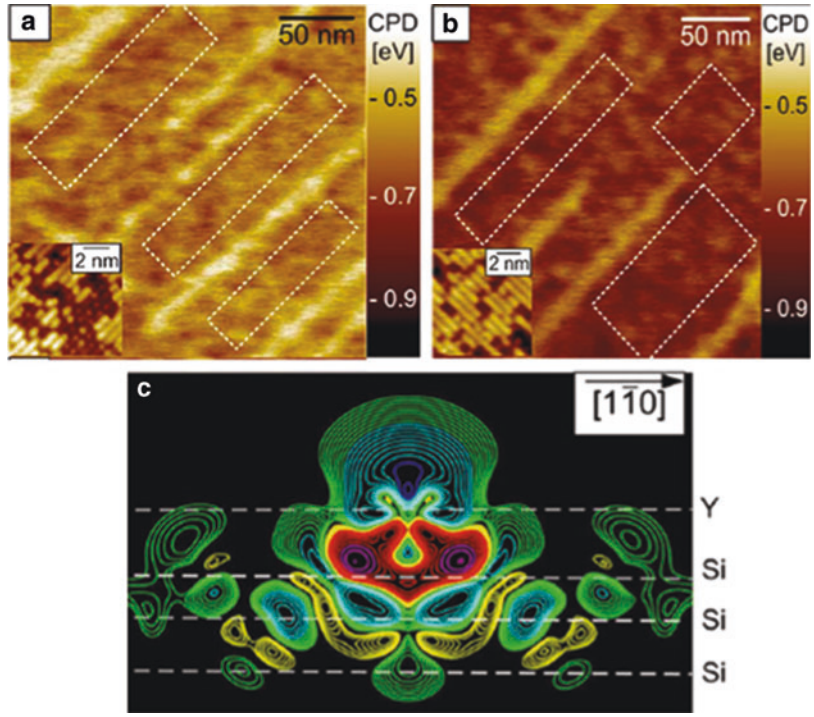


Surface Electronic Structure, Fig. 3 (**a**) STM image of Pt(111) surface. *R. Ragan unpublished results* (**b**) Normalized differential conductance of self-assembled monolayers of benzenethiol (*dashed curve*) and octanethiol (*solid curve*) on Pt(111) (Printed with permission from Ref. [10])

conductance at the Fermi energy and this represents an energy bandgap, i.e., no available states near the Fermi energy. For the benzene ethanethiol/Pt(111) junction, there is finite conductance at the Fermi energy and this is an indication of metallic behavior. A higher conductivity near the Fermi energy across benzene ethanethiol/Pt(111) junctions in comparison to octanethiol/Pt(111) junctions can be attributed to the fact that the benzene ethanethiol molecule has conjugated bonds in the molecule [9].

AFM measures the forces between AFM tip and surface by optically measuring the deflection of an AFM cantilever. With this basic mechanism, the type of signal feedback provides a wealth of information about the sample surface. Amplitude and frequency modulated AFM monitors the variation of the amplitude or frequency, respectively, of the AFM cantilever in response to forces between the tip and the surface and measures topography of the surface. In intermittent contact mode, the phase shift of the free cantilever resonance frequency provides nanometer scale information of the viscoelastic properties and adhesion force of the surface since it represents energy dissipation between tip and surface. For example, when imaging under a repulsive tip-sample condition, regions of the surface with the higher elastic modulus appears darker in a phase contrast AFM image. One can observe a contrast reversal when the tip changes from repulsive to attractive mode. Variations in local topography also induce a phase shift in the cantilever frequency and thus topography and phase images need to be analyzed in conjunction to understand the physical properties of the surface.

Derivatives of AFM can be used to measure electrical properties when using a conducting cantilever. EFM measures electrostatic forces between surface and cantilever and is modeled by treating the vacuum, air or any dielectric gap between tip and surface as a capacitor. The force is dependent on the tip-surface distance and the potential difference between tip and surface. In particular, KPFM directly measures the contact potential difference (difference between sample surface work function and tip work function) using a lock-in technique to null the electrostatic forces between tip and sample surface. Recently, Ragan and Wu et al. have demonstrated that measured values of work function obtained from KPFM compare quantitatively with DFT calculations and together provide information on the atomic arrangement and termination of atoms on

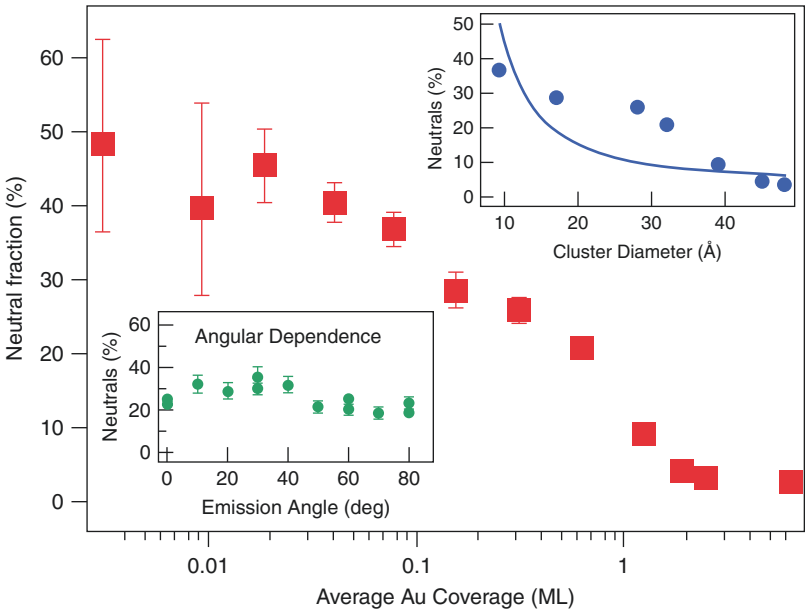


Surface Electronic Structure, Fig. 4 KPFM images of dysprosium nanowires on Si(001) that have been annealed post-growth at (**a**) 600 °C and (**b**) 680 °C. (**c**) Simulated charge density difference image for a single metal adatom on Si(001) with the cross section perpendicular to the surface. Charge accumulation increases from yellow to pink contour lines, whereas charge depletion increases from green to blue contour lines. The greatest charge depletion is seen at the metal adatom location and the greatest charge accumulation is seen in the region between the adatom and the subsurface Si atoms (Printed with permission from Ref. [11])

alloy surfaces [6]. Work function variations on surfaces reflect charge transfer [11, 12], quantum-size effects [13], and localized surface charge [14]. Figure 4 demonstrates how the work function on silicon changes when metal adatoms are on the semiconductor surface due to charge transfer between metal adatoms and substrate atoms. Figure 4a is a KPFM image after deposition of dysprosium on Si(001) and annealing the sample at 600 °C. Metallic disilicide nanowires form on the surface and appear as bright lines in the KPFM image since the disilicide nanowires have a lower work function than the Si(001) substrate. The Si(001) substrate in between nanowires is highlighted with white dashed boxes. A comparison of the substrate regions in Fig. 4a with those in Fig. 4b shows that the work function is lower on the substrate regions of the former. The STM images shown as insets in the lower left corner show that the Si(001) surface in Fig. 4a has more metal adatoms (dark regions in STM image) on the surface than the Si(001) surface of Fig. 4b due to the lower annealing temperature. The DFT simulation of Fig. 4c demonstrates that metal adatoms transfer charge to the Si atoms on the surface. This creates a dipole on the surface that lowers the work function [11]. Overall, KPFM measurements provide information about material behavior in devices; KPFM across heterojunctions provides information regarding band offsets, device performance of diodes and chemical sensitive field effect transistors and trapped charge at interfaces in high electron mobility transistors.

Ion Scattering

Low energy ion scattering (LEIS) is a method complimentary to scanning probe techniques that also provides information about surface composition, electronic structure, and atomic structure. LEIS has less stringent requirements on surface conditions than SPM as data can be acquired using LEIS from rough and contaminated materials. The energy of scattered ions depends on the ratio of the projectile and target masses, providing a measure of the atomic mass distribution on the surface. The degree of charge exchange that occurs during scattering is dependent on the surface electronic properties when using projectiles with low ionization energies, such as alkalis.



Surface Electronic Structure, Fig. 5 Neutral fractions (NF) of singly scattered 2.0 keV Na ions shown sas a function of the average Au coverage. The *right side inset* shows NF versus cluster diameter, with the symbols indicating experimental data and the solid line a theoretical fit. The *left side inset* shows NF for Na + scattered from a 0.15 ML Au coverage as a function of the emission angle with respect to the surface normal (Printed with permission from Ref. [16])

When an alkali-metal atomic particle is in the vicinity of a surface, its ionization level shifts up due to the image charge interaction, while it broadens due to overlap of the ion and surface wave functions. The measured neutral fraction depends on the ionization potential, the degree that the level shifts near the surface, and the work function at a point just above the scattering site. The charge exchange process is well described by a non-adiabatic resonant charge transfer model [15]. Since the interaction is local to the site where projectile atoms exit the surface, the neutralization of low energy alkali ions provides a unique method for measuring the local work function and quantum-size behavior in nanomaterials.

Yarmoff et al. has used LEIS to measure quantum-size effects in the electronic structure of gold nanoclusters on TiO_2 [16] to provide insight regarding how catalytic activity and surface electronic structure are correlated. Au or other heavy metal nanocrystals are ideal for ion scattering experiments, as the large mass of the cluster atoms enables a complete separation of the ions that impact the nanoclusters from those that impact the substrate. The integrated single scattering peaks in "Neutrals" and "Total Yield" spectra are divided to obtain the neutral fraction for scattering from the clusters. The neutral fraction for Na^+ ions scattered from Au nanocrystals as a function of Au coverage is directly correlated with the size of the Au nanocrystals. The neutral fraction goes from about 50 % for the smallest clusters down to about 3 % for the film [16]. The enhanced neutralization from small clusters is due to participation of quantum-confined states in the non-adiabatic resonant charge transfer process. Bulk Au has a relatively high work function (5.1 eV), so that the Fermi level is degenerated with the Na ionization level (also ~ 5.1 eV) and most of the scattered Na remains ionic. The neutral fraction for small Au clusters is considerably larger than for bulk Au because filled states associated with the clusters provide electrons that can tunnel to the outgoing projectile. The existence of such filled states is consistent with reports that small Au clusters are negatively charged. The additional filled states above the Fermi level depend on the size of the clusters, and thus provide a measure of the quantum-size behavior (Fig. 5).

Examples of Application

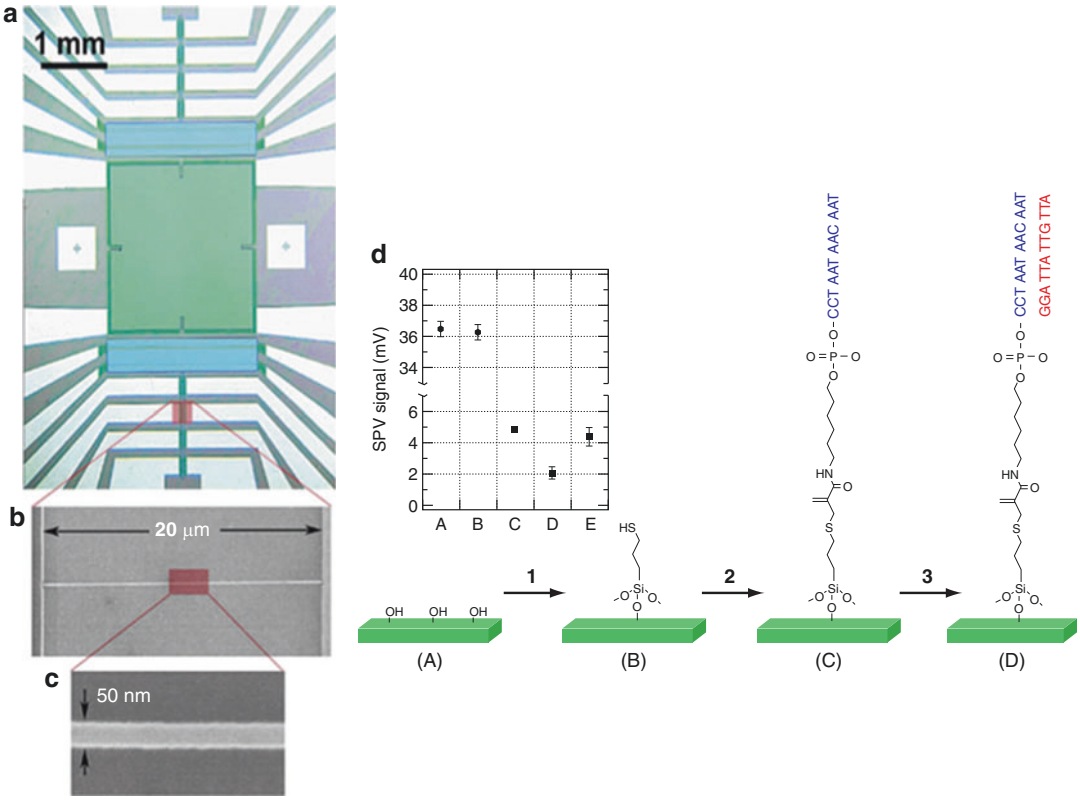
Nanowire Sensors

Chemical and biological nanowire sensors that use electronic transduction are an example of a

class of devices that uses changes in surface electronic structure to measure local molecular binding events or adsorption of molecules on surfaces. A chemical or biological sensor can be reduced to some basic components: a receptor (biological recognition element if selectivity is required), a transducer, and a means for processing whether or not target molecules of interest have bound to the surface and relaying this information to the user. Transducers can sense molecular interactions electronically (e.g., changes in localized charge) or optically (e.g., changes in dielectric constant). Changes in surface electronic structure can affect the conductance and this is easily measured in an electronic platform. A high surface to volume ratio in nanowires leads to high sensitivity for measuring changes in conductance due to surface binding events. If the surface of nanowires is functionalized with receptors that selectively bind to the target molecule of interest then selectivity can be engineered into the sensor platform. Since an early demonstration of semiconducting carbon nanotubes as chemical sensors, many different materials that form nanowires have exhibited the capability to sense molecules. In_2O_3 nanowires, SnO_2 nanoribbons, Si and ZnO nanowires have all been used as chemical and/or biological sensors. Si nanowires have been used for the detection of DNA hybridization, protein–protein interactions, cancer markers, and viruses. Some recent review articles that discuss nanowires-based sensor devices are referenced here [17, 18].

Nanowire sensors based on electronic transduction are commonly fabricated in a field effect transistor (FET) architecture with a back gate configuration. A single or multiple nanowires are connected to a source and drain and nanowires serve as the conducting channel. The conductance in the channel will change as molecules bind to the surface. Molecular binding events on semiconducting nanowire surfaces induce accumulation (depletion) of charge carriers that can be measured as an increase (decrease) lateral conductivity along nanowires in the FET architecture. The onset of accumulation or depletion of carriers depends on the surface charge induced and the carrier-type in the nanowire, *n-type* signifies that the majority carriers are negatively charged electrons and *p-type* signifies that the majority carriers are positively charged holes. Thus in a FET architecture, molecular binding events on surfaces can be monitored in an electronic circuit. For example, in a study by Cui et al. Si nanowires were functionalized with biotin receptors. When streptavidin, target molecule, was introduced into the system at a concentration of 250 nM an increase in conductance was observed. Streptavidin is well known to have a high binding affinity with biotin. The increase in conductance due to the introduction of strepavidin could be measured for concentrations as low as 25 pM [17, 18]. This early study demonstrated both selectivity and high sensitivity of Si nanowire sensors.

Si has some advantages as a sensing platform since microelectronic circuits are typically made from Si. This approach provides a strategy for fabricating high-density, high-quality nanoscale sensors that can be integrated with Si-based circuits. Si nanowire sensors have been fabricated by either "top-down" or "bottom-up" fabrication methods where the former uses lithographic methods and the latter uses self-organization routes. Previously, Li et al. reported an approach to configure FET device architectures using Si nanowires for DNA sensors using a standard "top-down" semiconductor process [19]. Reactive ion etching is used to fabricate Si nanowires on silicon-on-insulator substrates that have been patterned using electron beam lithography. Figure 6a is an optical image of the entire sensor system. Figure 6b, c are high-resolution SEM images showing a Si nanowire in the system. The chemical functionalization process to bind a single strand DNA receptors on nanowire surfaces is illustrated in the schematic of Fig. 6d. The chemical modification process can be monitored after each step by measuring changes in surface potential using the surface photovoltage technique. When solutions containing complementary strands of DNA with concentrations of 25 pmol are introduced to *p-type* Si nanowire surfaces, there is an increase in conductance due to accumulation of carriers. In the case of *n-type* Si nanowires, the DNA binding event leads to a



Surface Electronic Structure, Fig. 6 (**a**) Optical image of lab on chip system using Si nanowires for signal transduction. (**b**) SEM image of Si nanowire spanning electrical leads. (**c**) High-resolution SEM image of Si nanowire. (**d**) Chemical functionalization of Si nanowire surfaces, steps labeled (**a**–**d**), for sequence specific detection of DNA. The inset shows how the change in surface photovoltage (*SPV*) signal monitors each chemical functionalization step (Printed with permission from Ref. [19])

decrease in conductance due to depletion of carriers. Accumulation (*p-type*) or depletion (*n-type*) of carriers is associated with the negative charge on the backbone of DNA molecules. Quitoriano et al. introduced a different fabrication method for Si nanowire FET devices. This method is a combined "bottom-up" and "top-down" fabrication process utilizing the vapor–liquid–solid growth mechanism and optical lithography to fabricate FET devices [20, 21]. In this method, Au nanoparticles are deposited on the bottom surfaces of lithographically defined Si electrodes that overhang over a recessed trench of silicon dioxide. The growth of Si nanowires is guided from the bottom of the electrode at the site of the Au catalyst along the silicon dioxide layer to the opposite electrode. A benefit of combining "bottom-up" methods with "top-down" methods is to integrate nanowires into microelectronic circuits using high-throughput methods. In both cases, Si nanowires sensors can be connected directly to the adjoining circuitry for signal amplification and automated data acquisition.

Nanoscale Catalysts

Metallic and bimetallic surfaces and nanostructures with tunable physical and chemical properties have attracted particular attention in recent years due to their potential for use in a broad range of applications. The discovery by Haruta, et al. that gold nanoparticles on oxide supports exhibit surprisingly high catalytic activity for reactions such as CO oxidation and propylene epoxidation has inspired an enormous wave

of research in the quest for innovative nanocatalysts. Chemically active Au nanoparticles have been prepared on reducible (TiO_2, ZrO_2, NiO, or Fe_2O_3) oxides, and irreducible (SiO_2 or Al_2O_3) oxides in order to gain insight on the role of quantum-size effects, facets and steps at the edges of nanoclusters, and charge transfer to the substrate in chemical activity. A review of mechanisms involved in enhanced activity of Au nanoclusters can be found here [22]. Bimetallic surfaces of Pt and Pd, and trimetallic nanoparticles of Au/Pt/Rh were also found to be highly effective in promoting a variety of reactions, typically higher than corresponding monometallic nanoparticles. Thus there appears to be a variety of parameters affecting chemical activity.

Surface electronic structure has been identified as playing an important role in enhanced catalytic activity in nanoscale metallic systems. Electronic structure can be modified in nanosystems in many ways such as by charge transfer between nanocatalyst and substrate support or due to alloy effects in bimetallic systems. Both experimental and theoretical results show the strong coupling between electronic structure and catalytic activity. The Hammer-Nørskov model predicted that chemisorption energy correlates with the d-band center in transition metals and this is experimentally observed for oxygen and sulfur chemisorption energies on different metal surfaces [23]. The effect of electronic structure and its relationship to catalytic activity is also observed in experiments. For example, Au clusters on TiO_2 were measured by STS to undergo a metal to insulator transition at nanometer length scales. The nanocluster size were the metal to insulator transition occurs exhibited the highest turnover frequency for CO oxidation, while larger clusters having no band gap had lower activity [24]. It has also been shown that catalytic reaction rates, in the context of electrochemical catalysis, exhibit an exponential dependence on the catalyst work function and catalytic rate enhancements of up to a factor of 60 having been reported. Changes in work function as small as 200 meV can lead to an increase in rate enhancement by a factor of 10 [25]. The work function of Au nanowires can be varied by surface alloying [6] in order to optimize catalytic properties. Heterogeneous metal nanocatalysts with clusters that are a few nanometers in size hold great promise because of their large surface area to volume ratios, the availability of an enormous number of active sites and their enhanced resistance to poisoning from products of the reactions.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Atomic Force Microscopy
- ▶ Kelvin Probe Force Microscopy
- ▶ Nanomaterials for Electrical Energy Storage Devices
- ▶ Nanostructure Field Effect Transistor Biosensors
- ▶ Scanning Tunneling Microscopy
- ▶ Scanning Tunneling Spectroscopy
- ▶ Self-Assembly

References

1. Wildoer, J.W.G., Venema, L.C., Rinzler, A.G., Smalley, R.E., Dekker, C.: Electronic structure of atomically resolved carbon nanotubes. Nature **391**, 59 (1998)
2. Berger, C., Song, Z.M., Li, T.B., Li, X.B., Ogbazghi, A.Y., Feng, R., Dai, Z.T., Marchenkov, A.N., Conrad, E.H., First, P.N., de Heer, W.A.: Ultrathin epitaxial graphite: 2D electron gas properties and a route toward graphene-based nanoelectronics. J. Phys. Chem. B **108**, 19912 (2004)
3. Freeman, A.J., Wu, R.Q.: Electronic-structure theory of surface, interface and thin-film magnetism. J. Magn. Magn. Mater. **100**, 497 (1991)
4. Shinde, A., Wu, R., Ragan, R.: Thermodynamic driving forces governing assembly of disilicide nanowires. Surf. Sci. **604**, 1481 (2010)
5. Zeng, C., Kent, P.R.C., Kim, T., Li, A., Weitering, H. H.: Charge-order fluctuations in one-dimensional silicides. Nat. Mater. **7**, 539 (2008)
6. Ouyang, W., Shinde, A., Zhang, Y., Cao, J., Ragan, R., Wu, R.: Structural and chemical properties of gold rare earth disilicide core − shell nanowires. ACS Nano **5**, 477 (2011)
7. Meng, Y.S., Arroyo-de Dompablo, M.E.: First principles computational materials design for energy storage materials in lithium ion batteries. Energy Environ. Sci. **2**, 589 (2009)

8. Yeom, H.W., Takeda, S., Rotenberg, E., Matsuda, I., Horikoshi, K., Schaefer, J., Lee, C.M., Kevan, S.D., Ohta, T., Nagao, T., Hasegawa, S.: Instability and charge density wave of metallic quantum chains on a silicon surface. Phys. Rev. Lett. **82**, 4898 (1999)
9. Ragan, R., Ohlberg, D., Blackstock, J.J., Kim, S., Williams, R.S.: Atomic surface structure of UHV-prepared template- stripped platinum and single-crystal platinum(111). J. Phys. Chem. B **108**, 20187 (2004)
10. Lee, S., Park, J., Ragan, R., Kim, S., Lee, Z., Lim, D. K., Ohlberg, D.A.A., Williams, R.S.: Self-assembled monolayers on Pt(111): molecular packing structure and strain effects observed by scanning tunneling microscopy. J. Am. Chem. Soc. **128**, 5745 (2006)
11. Shinde, A., Cao, J.X., Lee, S.Y., Wu, R.Q., Ragan, R.: An atomistic view of structural and electronic properties of rare earth ensembles on Si(001) substrates. Chem. Phys. Lett. **466**, 159 (2008)
12. He, T., Ding, H.J., Peor, N., Lu, M., Corley, D.A., Chen, B., Ofir, Y., Gao, Y.L., Yitzchaik, S., Tour, J. M.: Silicon/molecule interfacial electronic modifications. J. Am. Chem. Soc. **130**, 1699 (2008)
13. Lee, S., Shinde, A., Ragan, R.: Morphological work function dependence of rare-earth disilicide metal nanostructures. Nanotechnology **20**, 6 (2009)
14. Rosenwaks, Y., Shikler, R., Glatzel, T., Sadewasser, S.: Kelvin probe force microscopy of semiconductor surface defects. Phys. Rev. B **70**, 085320 (2004)
15. Kimmel, G.A., Goodstein, D.M., Levine, Z.H., Cooper, B.H.: Local adsorbate-induced effects on dynamic charge-transfer in ion-surface interactions. Phys. Rev. B **43**, 9403 (1991)
16. Liu, G.F., Sroubek, Z., Yarmoff, J.A.: Detection of quantum confined states in Au nanoclusters by alkali ion scattering. Phys. Rev. Lett. **92**, 216801 (2004)
17. Patolsky, F., Lieber, C.M.: Nanowire nanosensor. Mater. Today **8**, 20 (2005)
18. Kolmakov, A., Moskovits, M.: Chemical sensing and catalysis by one-dimensional metal-oxide nanostructures. Annu. Rev. Mater. Res. **34**, 152 (2005)
19. Li, Z., Chen, Y., Li, X., Kamins, T.I., Nauka, K., Williams, R.S.: Sequence-specific label-free DNA sensors based on silicon nanowires. Nano Lett. **4**, 245 (2004)
20. Quitoriano, N.J., Kamins, T.I.: Integratable nanowire transistors. Nano Lett. **8**, 4410 (2008)
21. Quitoriano, N.J., Wu, W., Kamins, T.I.: Guiding vapor–liquid–solid nanowire growth using SiO_2. Nanotechnology **20**, 145303 (2009)
22. Min, B.K., Friend, C.M.: Heterogeneous gold-based catalysis for green chemistry: low-temperature CO oxidation and propene oxidation. Chem. Rev. **107**, 2709 (2007)
23. Greeley, J., Norskov, J.K., Mavrikakis, M.: Electronic structure and catalysis on metal surfaces. Annu. Rev. Phys. Chem. **53**, 319 (2002)
24. Valden, M., Lai, X., Goodman, D.W.: Onset of catalytic activity of gold clusters on titania with the appearance of nonmetallic properties. Science **281**, 1647 (1998)
25. Vayenas, C.G., Bebelis, S., Ladas, S.: Dependence of catalytic rates on catalyst work function. Nature **343**, 625 (1990)

Surface Energy and Chemical Potential at Nanoscale

Vanni Lughi
DI3 – Department of Industrial Engineering and Information Technology, University of Trieste, Trieste, Italy

Synonyms

Interfacial energy and chemical potential at nanoscale; Nanothermodynamics; Surface energy density and chemical potential at nanoscale; Surface free energy and chemical potential at nanoscale; Surface tension and chemical potential at nanoscale

Definition

Chemical potential is the energy increment of a system, associated to the addition of one single element of the substance – e.g., one atom or molecule. *Surface energy* is commonly defined as the energy increment associated to the formation of a unit area of a new surface. Although the latter is by far the most common and practical definition of surface energy and serves its purpose in most physical situations, it is subject to a number of ambiguities, which will be sorted out in the following section.

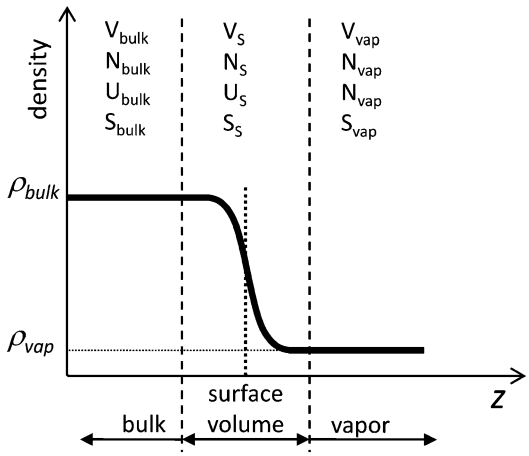
In nanoscale systems, a large portion of the atoms or molecules constituting the system is at or near surfaces. Knowledge of the surface properties, and in particular of surface energy, is therefore a key step in understanding and controlling nanostructures and nanostructured materials. On the other hand, surface geometry affects the chemical potential of the system, as shall be seen below,

and this effect becomes most important at the nanoscale. Chemical potential and surface energy are therefore not only key elements when it comes to describing systems at the nanoscale, but they are also somewhat interrelated and are therefore presented here together.

In most of current studies and applications, the subjects of surface energy and chemical potential can be properly discussed within a classical approach – and this is done in most of this entry. The more consistent and modern, but less agile, approach of nanothermodynamics is briefly introduced in the last paragraph.

Fundamental Considerations on Surface Energy

Before turning to the specifics of the subject, it is important to sort out some ambiguities that normally arise when discussing surface energy. In Gibbs' approach, the surface of a solid or liquid is considered as a transition volume between the bulk and its vapor, as depicted in Fig. 1. This model can in principle be generalized to multicomponent systems. All extensive thermodynamic quantities that characterize the system – such as volume, V; number of atomic or molecular constituents, N; internal energy, U; and entropy, S – can therefore be partitioned and a specific portion V_S, N_S, U_S, and S_S is assigned to the "surface volume" (Fig. 1). As these quantities represent extra amounts with respect to the bulk values, they are known as *excess* quantities. By combining these surface excesses, the Gibbs free energy $G_s = U_S - TS_S + PV_S$ *of the surface* and the Helmholtz free energy $F_S = U_S - TS_S$ *of the surface* can be defined – in complete analogy with the bulk. G_S and F_S are also excess quantities. Here T and P are the temperature and the pressure, respectively. Without loss of generality [1, 2], it is possible to assign the excess quantities that originally pertain to the surface volume to a representative, two-dimensional interface of area A (dotted line in Fig. 1). In this case, the surface energies can finally be defined:



Surface Energy and Chemical Potential at Nanoscale, Fig. 1 Gibbs model of a solid in equilibrium with its vapor (one-component system). The density, ρ, of the component is plotted as a function of the distance normal to the surface, z. The volume partition and the associated extensive quantities are indicated by the *dashed lines*

$$\text{Surface energy } u_s = \frac{U_S}{A} \tag{1}$$

$$\text{Gibbs surface free energy } g_s = \frac{G_S}{A} \tag{2}$$

$$\text{Helmhotlz surface free energy } f_s = \frac{F_S}{A} \tag{3}$$

Hence, strictly speaking, surface energy is defined as the internal energy *of the surface* per unit area, and the surface free energy is defined as the free energy (Helmholtz or Gibbs) *of the surface* per unit area, where in this case the "surface" is defined as the transition *volume* between the semi-infinite, bulk phases of the system as illustrated in Fig. 1.

In a more common approach, one can consider the entire free energy, G, of the generic one-component solid system characterized by the presence of a surface [2, 3] (the equation can be generalized to multicomponent systems by adding a $\mu_i N_i$ term for each ith component):

$$G = U - TS + PV = \mu N + \gamma A \tag{4}$$

Here μ and γ are the intensive quantities *chemical potential* and *surface tension*, which are therefore thermodynamically defined as:

$$\gamma = \left.\frac{\partial G}{\partial A}\right|_{T,P,N} \tag{5}$$

$$\mu = = \left.\frac{\partial G}{\partial A}\right|_{T,P,N} \tag{6}$$

Adsorption and Relationship Between Surface Tension and Surface Energy

Unfortunately, in most materials science, nanoscience, and nanotechnology literature, the distinction between surface energy, surface free energy, and surface tension is not always maintained. Although strictly speaking γ as defined in Eq. 5 is the surface tension and can be used as such for solids and liquids, the terms "surface energy" and "surface free energy" are also commonly used to identify the same quantity. In general, however, surface tension (Eq. 5) and surface (free) energy (Eqs. 1, 2, and 3) are different, and the difference depends on the surface excess of species: This can be in general the absorption of foreign species, or for the case of multicomponent systems simply the surface excess per unit area of the ith component of the system $\Gamma_i = N_{S,i}/A$. As shown by Gibbs [1], one finds:

$$\gamma = f_s - \sum_i \Gamma_i \mu_{S,i} \tag{7}$$

Equation 7 can be used to relate the concentration of a solute at the surface with a change of the surface energy. In pure one-component systems the summation is nil and $\gamma = f_S$, but this is a rather ideal case. In common systems, it is virtually impossible to obtain perfectly clean surfaces. In nanosystems foreign species are often added on purpose – such as in the case of colloidal nanoparticles, where capping agents are added to avoid the formation of clusters.

In the following, the term "surface energy" will be used as a generic term indicating the relevant one among "surface tension" (Eq. 5) and the different surface free energies defined in Eqs. 1, 2, and 3. These are all intensive quantities. Throughout this entry, the term "energy *of the surface*" will be used to indicate the relevant one among the extensive quantities γA, $g_S A$, or $f_S A$, where A is the surface area of the system. Note that some authors prefer to use the term "surface (free) energy *density*" when referring to these intensive quantities, in order to avoid confusion with the extensive quantities.

Surface Stress

In the definition of surface energy given above, the implicit mechanism for the formation of a new surface is cleaving of a bulk material. However, the area of an existing surface can be increased by other mechanisms, i.e., by applying a deformation (stretching, distortion). In this case a term needs to be added to Eq. 4:

$$G = U - TS + PV = \mu N + \gamma A + \xi A \tag{8}$$

Here ξ is the surface deformation energy density. For an elastic solid it is $\frac{1}{2}\sum_{i,j}\sigma_{ij}\varepsilon_{ij}$ where σ_{ij} and ε_{ij} are the tensor components of the *surface* stress and strain, respectively, and are reciprocally correlated by the compliance properties of the material. (Note that in this case σ_{ij} and ε_{ij} are defined as surface tensors in complete analogy with the bulk, and have units of force per unit length.) Some mathematical manipulation (see for example [2]) leads to:

$$A\,dy + S_S\,dT + V_S\,dP + N_S\,d\mu + A\sum_{i,j}\left(\gamma\delta_{ij} - \sigma_{ij}\right)d\varepsilon_{ij} = 0 \tag{9}$$

where S_S, V_S, and N_S are the excess quantities as defined previously and δ_{ij} is the Dirac delta function. Equation 9 is one of the key results in surface thermodynamics: On one hand, it can be regarded as a general form of Eq. 7 (some literature reports it as the *Gibbs adsorption equation*); moreover, it shows the relationship between chemical potential, surface energy, and surface stress (note that,

thanks to the Gibbs-Duhem equation $SdT - VdP + Nd\mu = 0$, which defines the general relationship between the intensive thermodynamic parameters, of the five variables in Eq. 9 γ, μ, ε, P, T, only three are independent). Finally, from Eq. 9 it can be shown [2] that:

$$\sigma_{ij} = \gamma\delta_{ij} + \left.\frac{\partial\gamma}{\partial\varepsilon_{ij}}\right|_T \tag{10}$$

Equation 10 shows that surface stress is rigorously equal to surface tension only when there is no dependence of the latter on the surface deformation. This is normally only true in liquids, where the surface atoms can rapidly rearrange in response to a deformation, whereas in solids surface stresses have to be relieved by opportune mechanisms, such as dislocations or buckling of the surface. In solid nanosystems, due to the predominance of surfaces, these effects can have a rather large impact.

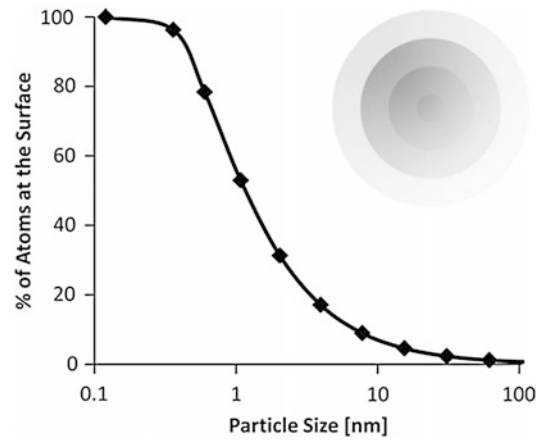
Typical Values

When a surface is created, a number of atomic bonds are broken. This consideration enables a practical estimation of the surface energy, since for a solid-vapor interface, the energy per unit area associated to the broken bonds is simply $\frac{1}{2}\varepsilon N_{bb}\rho_S$ where $\varepsilon/2$ is half of the bond energy, N_{bb} is the number of broken bonds per surface atom, and ρS is the planar atomic density at the surface (atoms per unit area). The bond energy can in turn be estimated – quite precisely in some instances such as the case of pure metals – from the latent heat of sublimation: $L_{sub} = \frac{1}{2}\varepsilon N_A N_{nm}$ where N_A is Avogadro constant and Nnn is the number of atom's nearest neighbors in the bulk [3]. L_{sub} is a known quantity that can be found in thermodynamic tables. The one described here, however, only provides a rough estimate. An entropic term should also be considered, and a number of structural rearrangements of the system, which shall be described in the following section, contribute to the actual, lower value of the surface energy.

Surface energy and surface tension are reported indistinctly with units of $J\ m^{-2}$ or $N\ m^{-1}$, although the latter is often preferred for expressing surface tension in liquids. Common values range from a few tens (in liquids, in polymers, and in ceramics) to a few thousands (in metals) of $mJ\ m^{-2}$.

Surface Energy at the Nanoscale

Nanosystems have very large surface extensions with respect to bulk systems. Based on geometrical considerations, it is easy to show that once the total amount of material is fixed, the total surface area of an ensemble of objects is inversely proportional to the linear size of the object, and so will be the contribution of the energy *of the surface* to the total energy system. For instance, an ensemble of nanoparticles with 10 nm diameter has 10^6 times more surface area (and energy of the surface) than the same amount of material organized in a single 1-cm particle. From a different standpoint, one can observe that the fraction of atoms that lie at the surface increases at the nanoscale. Figure 2 in particular shows a very sharp increase when reducing the size below ~20 nm. Materials properties that are strongly dependent on the energy of the system are expected to undergo dramatic changes in a similar size range. A number of such effects are briefly



Surface Energy and Chemical Potential at Nanoscale, Fig. 2 Geometrical approximation of the fraction of atoms at the surface of a spherical particle as a function of the sphere diameter. The *curve* is constructed by approximating the number of surface atoms with the volume of concentric spherical shells (*inset*)

discussed in the next paragraph dedicated to the chemical potential.

Minimization Mechanisms for the Energy of the Surface: Role in Nanostructures

As part of the total energy minimization process, a system will in general undergo a number of rearrangements at all scales in order to reduce the energy of the surface [4]. In nanosystems, this is particularly important for two reasons: Firstly, because the energy of the surface constitutes a large portion of the total system's energy. Moreover, even small rearrangements that would hardly affect a macroscopic system's morphology and/or properties can strongly impact structures with nanometric size. There are three categories of mechanisms that can contribute to reducing the total energy of the surface in a system, involving: (a) local surface phenomena, which act on reducing the surface energy of the structures; (b) the individual structure; (c) the overall system. These mechanisms have general validity, but will be discussed in the following from the standpoint of nanosystems, for which they are most relevant.

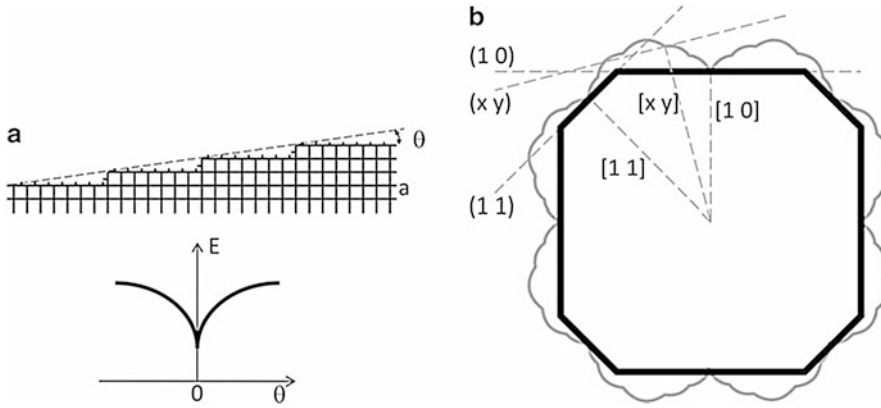
Local Surface Mechanisms. The first and unavoidable mechanism is *surface relaxation*: Atoms at the surface cannot maintain the position they would have in the bulk, as they would be subject to asymmetric forces, and are therefore bound to find a different equilibrium position. This normally results in an inward and/or lateral shifts of the near-surface atomic layers, depending on the crystal symmetry. Whereas in bulk materials the associated reduction of the average lattice constant is essentially negligible, it can be noticeable in some small nanoparticle systems. If there is more than one broken bond per atom, relaxation of the surface can occur by *reconstruction*, where new bonds form at the surface and the geometry of the surface changes. A classic case is the 7×7 reconstruction of clean (111) silicon surfaces. If foreign species are available, *physical or chemical adsorption* also reduces surface energy. In atmosphere, essentially all surfaces are covered by adsorbed species, most commonly hydrogen or hydroxyl groups. Finally, changes of the chemical composition at the surface, for instance by *segregation* of impurities, are another effective way to reduce surface energy. This mechanism can have a strong impact on the properties of nanoparticles and nanowires: In these cases segregation at the surface can fully deplete any impurity in the interior of the nanostructure. If on one hand this is one of the reasons for the high purity of nanocrystals, on the other it is a major, fundamental obstacle for achieving effective electronic doping in semiconducting nanostructures.

Individual Structure Mechanisms. An individual structure – e.g., nanoparticles, nanowires, etc. when considering the nanoscale – can reduce the energy of the surface by opportunely changing shape. In isotropic systems, such as liquids or amorphous solids, surface area, A and total energy of the surface, γA, are strictly proportional, therefore minimizing the former also minimizes the latter.

For anisotropic systems, such as crystals, the surface energy varies from one crystal plane to the other. The quantity that needs to be minimized is $\sum_i A_i \gamma_i$ (the index i refers to the different crystallographic planes in the crystal) so that minimization of the area is not in general the optimal solution. This problem is solved by using the *Wulff construction*, which predicts the equilibrium shape of crystals. As illustrated in Fig. 3a, a *vicinal surface* – i.e., a crystal plane forming a small angle θ with the close-packed plane – will have additional broken bonds with respect to a close-packed plane. The number of broken bonds on such a vicinal surface can be derived from simple geometrical considerations:

$$N_{bb} = \frac{cos\theta + sin\,|\theta|}{2a^2} \tag{11}$$

where a is the lattice parameter [2]. In proximity of a close-packed plane the surface energy will then have the shape shown in Fig. 3a, in accordance with Eq. 11: The close-packed plane, characterized in general by small Miller indexes, is associated to a local minimum. The surface energy for all crystal directions can be plotted in a polar diagram, as shown in Fig. 3b for a two-dimensional cross section. One can draw a radius vector for each crystal direction, and then,

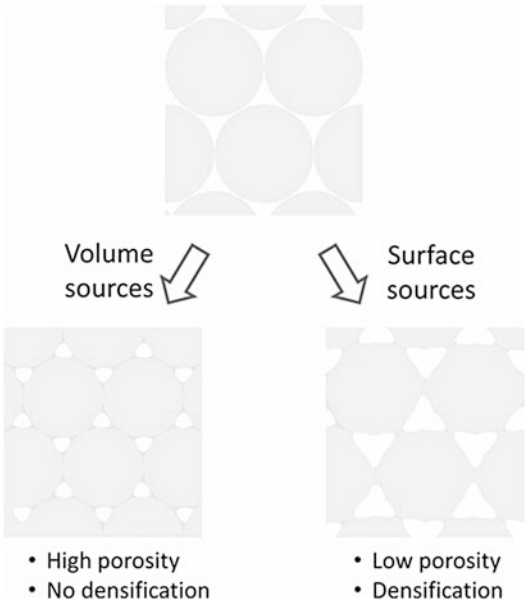


Surface Energy and Chemical Potential at Nanoscale, Fig. 3 (**a**) Schematic cross section of a vicinal surface forming an angle θ with a close-packed plane, and diagram of the surface energy as a function of the angle. (**b**) Example of a two-dimensional polar plot of the surface energy (*gray solid line*). The *dashed lines* indicate the Wulff construction of the planes perpendicular to selected crystallographic directions. The *black solid line* indicates the expected equilibrium shape of the crystal. Note that the generic (xy) does not contribute to the final equilibrium shape because of the high surface energy associated to it

at the intersection point, the plane that is perpendicular to such vector. Wulff demonstrated that the internal envelope of all such planes defines the equilibrium shape of a crystal, as it minimizes the quantity $\sum_i A_i \gamma_i$. While in macroscopic crystals the shape predicted by the Wulff construction is rarely observed because the kinetics to equilibrium are rather slow at the temperatures of interest, nanosystems often do reach equilibrium due to the short spatial scales involved, and the approach described here can be a useful tool for predicting the shape of nanocrystals.

Mechanisms Involving the Overall System. A system consisting of an ensemble of structures can reduce the overall energy of the surface by simply *aggregating* such structures to one another, thus reducing the surface area. An important case is that of suspensions colloidal particles, where stabilization mechanisms (such as electrostatic charging, or steric stabilization by coating the particles with organic ligands) are needed especially for nanosized particles. Lack of proper stabilization leads to aggregation of the nanoparticles with the formation of large clusters, and the benefits of the colloidal suspension as well as the nanoscale properties of the particles are lost.

An aggregate, e.g., of particles, can further reduce the energy of the surface by *sintering*. During sintering, atoms move toward concave surfaces driven by diffusion mechanisms, which in turn are controlled by the radius of curvature of the target surface: The lower the radius (negative values correspond to concave surfaces), the faster the diffusion toward that area. This results in the progressive merging of adjacent particles, as depicted in Fig. 4. The sources of atoms can be the interior volume of the particles, the surface of the particles, or the boundaries between particles (i.e., as sintering proceeds, the *grain boundaries*). When the source of atoms is the interior volume or a grain boundary, the particles get closer, and densification can occur. If properly conducted, sintering can lead to the formation of a fully dense solid from an aggregate of particles. When the source of atoms is the surface, then no densification can occur, and the final structure will have a high porosity – and therefore still a high surface area and energy of the surface. The dominant source of atoms strongly depends on the sintering conditions and can in principle be controlled, thus providing a tool for tailoring the material's morphology at the nanoscale and consequently engineering its properties: On one hand, obtaining a fully dense solid with grains of nanoscale size has generated quite a bit of interest because it leads to excellent mechanical properties and, in the case of materials subject to phase transformation, to enhanced phase stability (as illustrated in the



Surface Energy and Chemical Potential at Nanoscale, Fig. 4 Schematic of a sintering process and of the resulting microstructures. (*Left*) If the source of the diffusing species is the volume of the particles or the grain boundaries, densification occurs: The distance between the particle centers decreases, and the porosity is reduced. (*Right*) If the source of the diffusing species is the particle surface, no densification can occur: The material is just redistributed along the surfaces, favoring the necks that are now forming between particles, and the particle centers do not change position

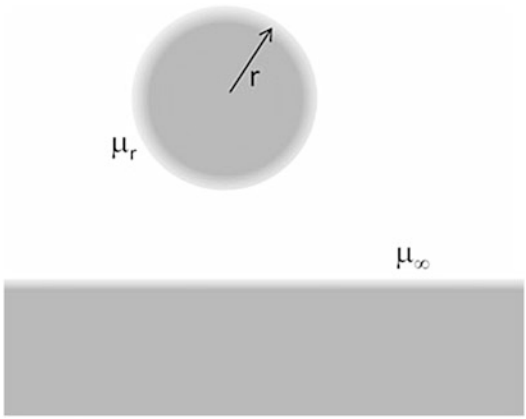
following paragraph). On the other hand, obtaining a highly porous material, where the pore size is at the nanoscale, is of interest for a number of situations including: high-performance gas sensors; separation of pollutants in environmental applications; catalysis and photocatalysis, including innovative routes for producing and storing hydrogen for energy applications; scaffolds in biological and biomedical applications; electrodes for fuel cells and batteries; high-performance thermal insulators. However, sintering in nanosystems can also be an unwanted effect. This is especially true when the nanosystem is subject to heat treatment, even at rather low temperatures: The reason is that the diffusion mechanisms that govern sintering become well active at temperatures close to 50–70 % of the melting temperature, which in turn drops dramatically at the nanoscale as mentioned in the following paragraph.

Finally, *Ostwald ripening* is another important mechanism where interaction between the elements of the entire system leads to reduction of the overall surface area and of the energy associated to it. In this case, larger particles grow at the expenses of smaller particles. The reason is that atoms at the surface of a small particle are less stable than those at the surface of a larger particle. (More details on this mechanism will be given in the following section.) Ostwald ripening is of major importance when dealing with nanostructures, where small radii of curvature are common. Important examples include sintering processes of nanopowders – in particular densification processes, where the ripening should be avoided as it conflicts with the need of keeping grain size small and uniformly distributed. (See for example the phenomenon of "exaggerated grain growth" typical of ceramic materials processing.) Also, Ostwald ripening is often observed during the synthesis of colloidal nanoparticles – where it can either be deleterious or, if opportunely controlled, lead to a better size distribution.

Chemical Potential at the Nanoscale and Relationship with Surface Energy

The definition of chemical potential was given above as the energy increment of a system associated to the addition of one atom or molecule, and was expressed quantitatively in Eq. 6. One requirement that is explicit in this equation is that the area of the system must remain constant. In general, this is not true, as addition of even one atom or molecule to the material does change, in principle, the surface area. This effect is only important when the size of the system is small. It can be shown [3, 4] that the difference between the chemical potential, μ_r, of a spherical particle of radius r and the chemical potential, μ_∞, of a semi-infinite bulk with a flat surface made of the same material (Fig. 5) is:

$$\mu_r - \mu_\infty = \frac{2\gamma\Omega}{\mathrm{r}} \tag{12}$$



Surface Energy and Chemical Potential at Nanoscale, Fig. 5 Model for calculating the chemical potential difference, $\mu_r - \mu_\infty$, between a particle of radius r and a semi-infinite bulk with a flat surface

where Ω is the atomic volume. This relationship, named after Gibbs and Thomson who first independently derived it, is obtained by realizing that the energy change due to the addition of a single atom to the particle, $\mu_r \partial N$, must be equal to the energy change associated to the change in surface area, $\gamma \partial A$, and mathematically manipulating the balance equation. The term $2\gamma/r$ is known as the Laplace pressure, and it can be shown that the pressure difference, Δp, between the material (solid or fluid) inside and outside of the curved surface is:

$$\Delta p = \frac{2\gamma}{\mathrm{r}} \tag{13}$$

Known as the *Young-Laplace equation*, this is also often regarded as a mechanical definition of the surface tension, and is one of the key equations when studying micro- and nanofluidics. Note that, in the case of a solid, a consequence of this relationship is that in a small crystal, there should be a contraction of the lattice parameter, an effect that is actually observed but only for extremely small particles (below about 1 nm). Equations 12 and 13 can be extended to any curved surface, by substituting $2/r$ with the generalized local curvature of the surface, $(1/r_1 + 1/r_2)$, where r_1 and r_2 are the principal radii of curvature.

The Gibbs-Thomson relationship shows the coupling between surface energy and chemical potential, and provides a solid framework for deriving their combined effects on mechanisms and properties of systems, especially at the nanoscale where $\mu_r - \mu_\infty$ can be very large and the effects are most evident. Some important examples are briefly described in the following [4].

First, the Gibbs-Thomson can be extended to the *vapor pressure* of the system. Assuming an ideal gas behavior for the vapor, one can show the chemical potential of an atom in the vapor phase is $\mu_v = \mu_\infty - kT \ln P_\infty$ where P_∞ is the equilibrium vapor pressure of a flat surface, T the temperature, and k Boltzmann's constant; an analogous relationship holds when a curved surface is involved and P_r is the equilibrium vapor pressure of a curved surface. It is therefore possible to calculate $\mu_r - \mu_\infty$ and to use Eq. 12 to find the *Kelvin equation*:

$$\ln \frac{P_r}{P_\infty} = \frac{2}{r} \frac{\gamma \Omega}{kT} \tag{14}$$

The vapor pressure increases as the inverse of the particle (or droplet in the case of a liquid) radius. Once again, the effect becomes sensible at the nanoscale. An identical relationship can be derived for the *solubility* of solids, S, revealing that nanoparticles are much more soluble than the bulk counterpart:

$$\ln \frac{S_r}{S_\infty} = \frac{2}{r} \frac{\gamma \Omega}{kT} \tag{15}$$

The solubility difference between particles of different radius is also strictly correlated with *Ostwald ripening*, a phenomenon that is observed in colloidal suspensions of nanocrystals or during sintering processes of ceramic nanopowders. The driving force for this mechanism is the difference in chemical potential between small and large particles. Considering for example a suspension of particles in a solvent, the equilibrium between precipitate and solute will initially be established when the chemical potentials of the solution, μ_{SOL}, and that of the *smallest* particle, $\mu_{r,small}$, are the same. However, precipitation will still

occur locally at the surface of the larger particles, since $\mu_{r,large} < \mu_{r,small}$ and $\mu_{r,small} = \mu_{SOL}$. This precipitation, however, will further reduce the chemical potential in the solution, so now $\mu_{SOL} < \mu_{r,small}$, leading to further dissolution of the smaller particles. The process continues, and smaller particles dissolve while large particles grow.

One of the most important effects of the coupling between size and chemical potential, especially because of its practical consequences in nanotechnology processes, is the dramatic drop of the melting temperature, *Tm*, with size. For a spherical particle, mathematical manipulation of the Gibbs-Thomson equation leads to the following relationship:

$$\frac{T_m(r)}{T_{m.\infty}} = \left(1 - \frac{2\gamma\Omega}{H_f r}\right) \tag{16}$$

where r is the particle radius, $T_{m,\infty}$ is the melting temperature of the bulk, H_f is the enthalpy of fusion. As mentioned, this has a major effect also on the sintering processes of nanopowders and in general of nanostructured materials.

The size of a system can also influence *thermodynamic phase stability* at a given temperature. This can again be derived by considering that the energy of the surface increases when reducing size. Consider a bulk material that is thermodynamically stable in a phase A at higher temperature and in a phase B at lower temperature. Considering now a spherical particle of radius r, the free energy change associated to the transformation from phase A to phase B must include surface terms as well as bulk terms:

$$\Delta G_{A\to B} = G_B - G_A = \left(\frac{4}{3}\pi r^3 \frac{\rho \mathrm{N}_A}{\mathrm{M}}\mu_\mathrm{B} + 4\pi r^2 \gamma_\mathrm{B}\right) - \left(\frac{4}{3}\pi r^3 \frac{\rho \mathrm{N}_A}{\mathrm{M}}\mu_\mathrm{A} + 4\pi r^2 \gamma_\mathrm{A}\right) \tag{17}$$

Here the factor $(4\pi r^3/3)(\rho N_\mathrm{A}/M)$ multiplying the chemical potential is simply the number N of atomic or molecular constituents in the particle, as defined before; M is the molar mass, ρ the density, and N_A the Avogadro number. First, one can observe that $\mu_\mathrm{B} < \mu_\mathrm{A}$ because in the bulk, below the transformation temperature, A will spontaneously transform to B. Then, by setting $\Delta G_{A\to B} = 0$, simple mathematical manipulation leads to the result that, if $\gamma_\mathrm{B} > \gamma_\mathrm{A}$, there exists a critical radius:

$$\mathrm{r_c} = 3\frac{M}{\rho N_\mathrm{A}}\frac{\gamma_\mathrm{B} - \gamma_\mathrm{A}}{\mu_\mathrm{A} - \mu_\mathrm{B}} \tag{18}$$

such that if $r < r_c$ then $\Delta G_{A\to B} > 0$. In other words, particles smaller than the critical radius are thermodynamically stable, and the transformation from A to B does not occur even if the system is below the bulk transformation temperature. Clearly, this effect is only important when the surface terms in Eq. 17 are of the same magnitude as the volume terms, and this can only happen for small radii. Considering typical values of γ and g, r_c is in the order of a few nanometers. A typical example [5] is the stabilization, obtained by reducing the grain size below the critical radius, of tetragonal zirconia at room temperature – where the thermodynamically stable phase for the bulk would be monoclinic. The critical size for pure zirconia powders is in the range of 5–10 nm, while in a polycrystalline zirconia (where stress effects arise) it is about 30 nm. However, for the practical use of zirconia, chemical stabilizers are commonly added, too, such that stabilization is achieved for larger grain sizes, compatible with common processing.

Nanothermodynamics

Although quite effective in achieving useful and realistic results, the approach outlined above for studying the thermodynamics of small systems – consisting in adding surface-related ad-hoc terms to the energy equations – has a certain degree of inconsistency: When dealing with macroscopic systems, energy is an extensive quantity, i.e., proportional to the amount of material that constitutes the system. This does not hold anymore in nanoscopic systems, since a large portion of the total energy of the system is

associated to the surface, which is dependent on the size of the objects that constitute the system. Hill has proposed a consistent thermodynamic approach [6, 7], overcoming these inconsistencies.

One can consider a macroscopic system comprising η equivalent and non-interacting, smaller subsystems (a good example would be a colloidal suspension of nanocrystals), characterized by the extensive variables U, S, V, N as defined before. Then, for the entire system $U_{tot} = \eta\, U$; $V_{tot} = \eta\, V$; $S_{tot} = \eta\, S$, $N_{tot} = \eta\, N$. Hill modifies the standard thermodynamic equation for the entire system as:

$$dU_{tot} = TdS_{tot} - PdV_{tot} + \mu dN_{tot} + Ed\eta \quad (19)$$

Essentially, he adds the term $Ed\eta$, where $E = \partial U_{tot}/\partial\eta|_{S_{tot},V_{tot},N_{tot}}$ is the *subdivision potential*. The procedure followed by Hill is analogous to what Gibbs had done by introducing the chemical potential. The analogy is even stronger, however: While the chemical potential determines how the energy of the system varies when changing the number N_{tot} of atoms or molecules that constitute the system, the subdivision potential determines the energy change upon a variation of the number of subsystems η – which in turn corresponds to a variation of the size of the subsystems if the total amount of material in the system is kept constant. In nanoscopic systems, changes in the number of subsystems imply large variations to the total energy, and the subdivision potential contributes significantly to Eq. 19. However, the subdivision potential will be essentially zero for macroscopic systems, as a change in the number of subsystems will have little effect on the total energy (considering for example the surface energy contribution, this will be negligible because in the surface area variations are small as long as the system size is not in the nanoscale regime). In this case, Eq. 19 reduces to the standard thermodynamic equation, so that in this view nanothermodynamics can be viewed as a generalization of standard thermodynamics. Without recurring to ad-hoc additions to the thermodynamic equations, the subdivision potential enables one to naturally and consistently treat all size-related phenomena, such as surface effects, edge effects, rotation and translation of the systems, etc., thus becoming particularly useful in nanosystems and especially when such phenomena occur together and possibly when coupled. A number of complex nanoscale problems have been successfully treated by using nanothermodynamics, most notably: open linear biological aggregates, local correlations in bulk magnetic materials, glassy systems, and metastable liquid droplets in vapor. More formulations of nanothermodynamics have been formulated over the past couple of decades, with some advantages in terms of the treatment of fluctuations – which might be important in nanosystems – or for the treatment of complex systems, but are essentially equivalent to Hill's.

Cross-References

- ▶ Applications of Nanofluidics
- ▶ Capillary Flow
- ▶ Lotus Effect
- ▶ Mechanical Properties of Nanocrystalline Metals
- ▶ Nanomechanical Properties of Nanostructures
- ▶ Nanoscale Properties of Solid–Liquid Interfaces
- ▶ Nanostructures for Surface Functionalization and Surface Properties
- ▶ Nanotribology
- ▶ Self-Assembled Monolayers for Nanotribology
- ▶ Self-Assembly
- ▶ Self-Assembly of Nanostructures
- ▶ Surface Tension Effects of Nanostructures
- ▶ Surface-Modified Microfluidics and Nanofluidics
- ▶ Wetting Transitions

References

1. Gibbs, J.W.: Collected Works, vol. 1. Yale University Press, Hew Haven (1948)
2. Zangwill, A.: Physics at Surfaces. Cambridge University Press, New York (1988)

3. Porter, D.A., Easterling, K.E.: Phase Transformations in Metals and Alloys. Van Nostrand Reinhold, Wokingham (1981)
4. Cao, G.: Nanostructures & Nanomaterials: Synthesis, Properties & Applications. Imperial College Press, London (2004)
5. Garvie, R.C.: Stabilization of the tetragonal structure in zirconia microcrystals. J. Phys. Chem. **82**, 218 (1978)
6. Hill, T.L.: A different approach to nanothermodynamics. Nano Lett. **1**, 273–275 (2001)
7. Hill, T.L.: Thermodynamics of Small Systems. Dover, New York (1994)

Surface Energy Density

► Surface Tension Effects of Nanostructures

Surface Energy Density and Chemical Potential at Nanoscale

► Surface Energy and Chemical Potential at Nanoscale

Surface Engineering, Tailored Wettability, and Applications

Solomon Adera, Jiansheng Feng and Evelyn N. Wang
Device Research Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Synonyms

Contact angle; Hydrophilic; Hydrophobic; Three-phase contact line

Definition

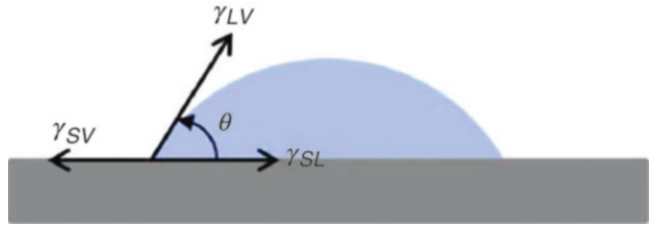
Wettability is a fundamental property of solid surfaces, which plays a role in all aspects of our lives. It is characterized by the contact angle that a liquid droplet makes when deposited on a solid surface. The contact angle is the angle subtended by the liquid-vapor and the solid-liquid interface from the liquid side at the three-phase contact line where the three phases (solid, liquid, and vapor) meet. The Young contact angle is the angle that a droplet makes when it comes in contact with an atomically flat, chemically homogeneous, nonreactive, rigid, and insoluble surface (Fig. 1). This angle is typically referred to as the intrinsic or equilibrium or static contact angle (θ) and can be obtained by balancing the interfacial forces at the three-phase contact line as given by the classical Young equation [1]

$$\cos\theta = \frac{\gamma_{SV} - \gamma_{SL}}{\gamma_{LV}}. \quad (1)$$

Here, γ is the interfacial energy per unit area (or equivalently the force per unit length) that acts along the interface indicated by the respective indices (S for solid, L for liquid, and V for vapor). For example, γ_{LV} is the force per unit length along the liquid-vapor interface. Thus, Eq. 1 shows that the equilibrium contact angle for a smooth surface is determined by the chemical nature of the different phases involved. For a water droplet, if this equilibrium contact angle is less than 90°, the surface is termed hydrophilic (wetting), and if the equilibrium contact angle is greater than 90°, the surface is termed hydrophobic (non-wetting).

Introduction

Surface texturing with micro-/nanostructures has been established as an effective way for achieving desirable fluidic (e.g., wetting, self-cleaning) [2], optical (e.g., antireflection) [3], mechanical (e.g., hydrodynamic drag reduction) [4], and thermal (e.g., heat transfer and phase change) [5] properties. These applications, however, require superior wetting characteristics which is a long standing challenge and current research interest in surface engineering. Two extreme limits are often desired: complete wetting and non-wetting behaviors. In the limit of complete wetting, a liquid droplet



Surface Engineering, Tailored Wettability, and Applications, Fig. 1 Schematic of a droplet deposited on a smooth solid surface showing interfacial forces acting at the three-phase contact line. The equilibrium contact angle is determined by balancing the interfacial forces at the contact line

spontaneously spreads with a near zero contact angle forming a thin liquid film. On the other hand, in the limit of complete non-wetting, a liquid droplet remains contained and maintains a spherical shape with high contact angle ($\theta > 150°$) and minimal solid-liquid contact area.

Micro-/nanofabrication is used to improve surface wettability. By introducing roughness, an intrinsically hydrophilic surface can be rendered more hydrophilic by reducing the apparent contact angle, while an intrinsically hydrophobic surface can be rendered more hydrophobic by increasing the apparent contact angle [6]. Consequently, micro-/nanostructuring offers exciting opportunities in tailoring the wetting characteristics of surfaces for various engineering applications ranging from liquid transport to heat transfer, where insight into the fundamental physics and mechanistic understanding of the wetting dynamics are essential, as explained in this entry.

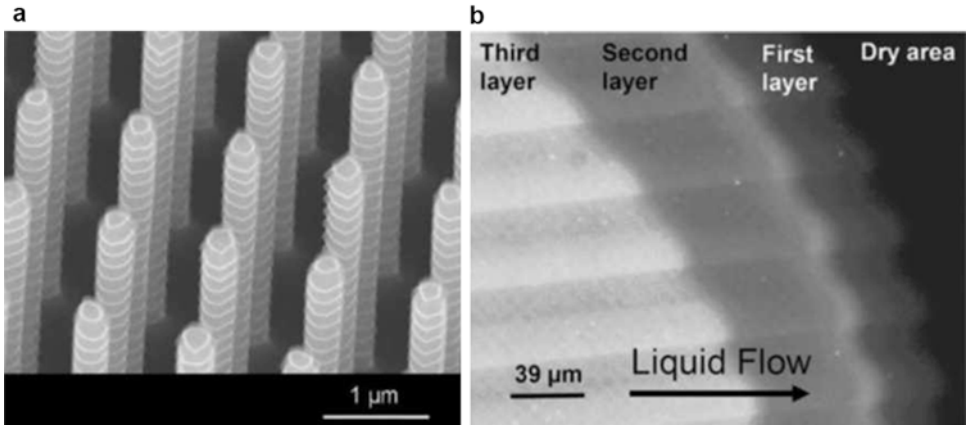
Liquid Spreading Dynamics

Symmetry of the boundary conditions of a liquid droplet deposited on a solid surface dictate that the spreading process as well as the final shape should be axisymmetric. However, chemical heterogeneity and surface roughness, which are typical of actual surfaces, can introduce local energy barriers that can cause contact line pinning resulting in asymmetric propagation dynamics. This asymmetry and liquid propagation on micro-/nanostructured surfaces has been extensively studied due to its applications in microfluidics and lab-on-a-chip devices. However, the liquid transport and wetting dynamics on these surfaces, which are influenced by surface tension forces due to the small length scale, are complex and not always well understood leaving room for speculation to the governing physics. In this section, we describe two examples where structure length scale and induced roughness can play important roles in unique liquid spreading behavior.

Multilayer Liquid Spreading in Micropillar Arrays

Recent experimental investigations show that silicon micropillar arrays with scallop features can cause a spreading or receding liquid front to separate into multiple layers with varying thickness [7] as shown in Fig. 2.

The macroscopic layer-by-layer liquid spreading is due to the presence of scallop features on the side walls of the pillars which result from the deep reactive ion etching process that is used to create the micropillars. The scallop features act as energy barriers during the spreading process by pinning the three-phase contact line at the sharp edges on the sides of the pillars. The propagating liquid overcomes these energy barriers in a stepwise manner. When the liquid overcomes the first energy barrier, it forms the first liquid layer that propagates forward and wets the surface. Subsequent layers follow as the liquid climbs up the scalloped tiers by sequentially overcoming the respective energy barriers. This results in subsequent layers of liquid propagating on top of the first layer resulting in multilayer liquid spreading. For pillars with the same diameter and spacing but in the absence of nonvisible scallop features, the liquid spreads uniformly in one layer across the



Surface Engineering, Tailored Wettability, and Applications, Fig. 2 Multilayer liquid spreading in micropillar arrays. (**a**) Scanning electron micrograph of a surface with nanostructured arrays of pillars with diameter of 500 nm and spacings of 800 nm. The scallops, which are a result of the Bosch deep reactive ion etching process to create deep trenches, have dimensions of approximately 100 nm. (**b**) Visualization of multilayer spreading on the corresponding geometry in (**a**). Liquid spreads from *left* to *right* where the dark area is dry. The differences in brightness indicate variations in liquid film thickness (Reproduced with permission from Xiao et al. [7])

entire micropillar array structure with no distinct layers.

This multilayer liquid spreading on silicon micropillar arrays can be explained using a surface energy-based thermodynamic model. The model treats the scallop features as tiered steps that present an energy barrier in the liquid propagation dynamics. The analytical model provides design guideline for creating micropillar arrays that promote multilayer liquid spreading and offers opportunities to control the propagation dynamics as well as the liquid film thickness on superhydrophilic surfaces.

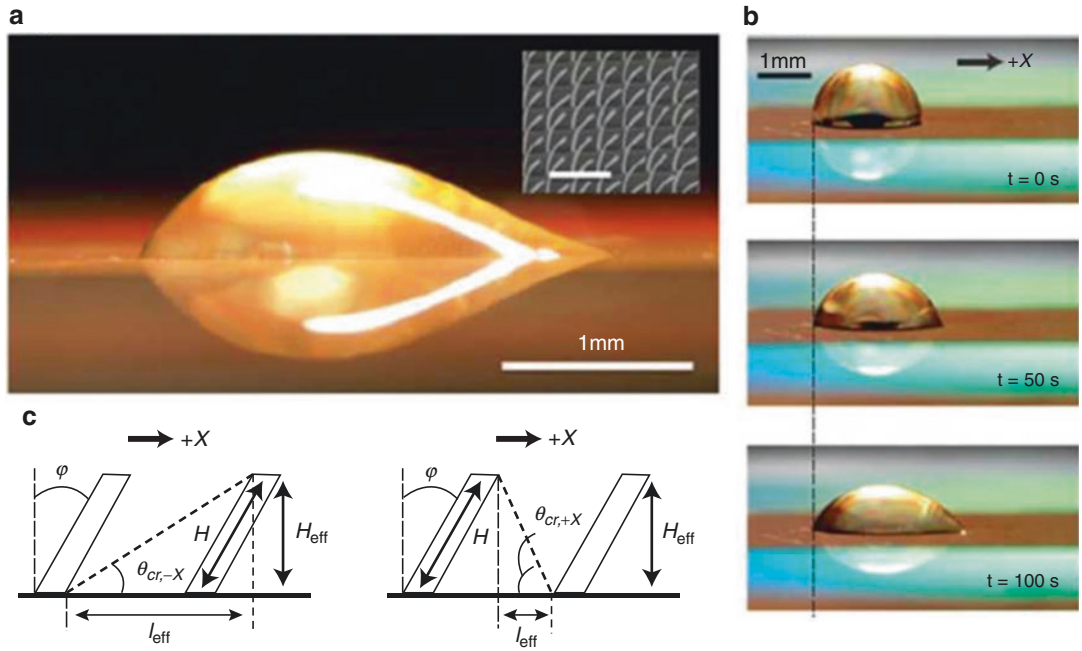
Unidirectional Liquid Spreading

Unidirectional spreading of liquids has been experimentally demonstrated on a macroscopically uniform surface which consists of arrays of asymmetric pillars which are collectively orientated in a single direction [8]. The slanted nanopillars allow the wetting liquid to propagate in one direction only (Fig. 3a, b), i.e., along the pillar-tilting direction (denoted as +X, Fig. 3c). In the opposite direction (−X), the three-phase contact line is pinned at the base of the nanopillars, and the liquid does not spread. To make such a surface, first arrays of upright silicon nanopillars were fabricated using deep reactive ion etching. This is followed by depositing a thin gold film on one side of the nanopillars using electron-beam evaporation. When cooled to room temperature, the residual thermal stress on the thin gold film caused the partially coated silicon nanopillars to deflect in one direction. The deflection angle is observed to be a function of the thickness of the gold film.

When a wetting droplet is deposited on these surfaces, the slanted arrays of nanopillars pull the liquid toward the pillar top. As the meniscus at the spreading front is pulled forward by one row of pillars, the contact line at the base of the substrate reaches the next row of pillars. Modeling results show that the liquid continues to spread if the intrinsic contact angle is below a critical angle that is determined from the pillar array geometry. In the opposite direction, however, the contact line is pinned at the base of the front row of nanopillars, and the liquid does not spread resulting in unidirectional spreading.

Wettability and Heat Transfer

In boiling heat transfer, textured hydrophilic surfaces have been demonstrated to enhance the thermal performance. Similarly, condensation heat



Surface Engineering, Tailored Wettability, and Applications, Fig. 3 Unidirectional liquid spreading on slanted pillar array. (**a**) The characteristics of a spreading droplet on asymmetric nanostructured surface (*inset*) at one instant in time. The scale bar in the inset is 10 μm. The diameter, spacing, height, and deflection angle of the nanopillars are 0.5 μm, 3.5 μm, 10 μm, and 12°, respectively. (**b**) Side view time-lapse images of unidirectional spreading of a 1 μL droplet of deionized water with 0.002 % by volume of surfactants (Triton X-100). The initial location of the droplet contact line in the −X direction is indicated by the *dashed line*. (**c**) Schematic diagrams explaining the geometries for the model to determine the critical angle in the −X (*left*) and +X (*right*) directions (Reproduced with permission from Chu et al. [8])

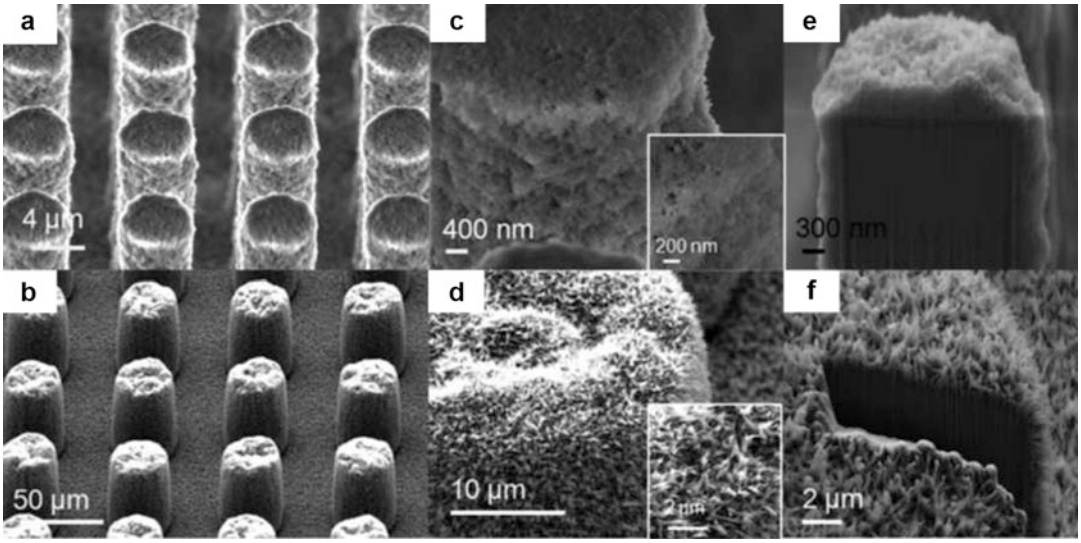
transfer has been improved by using textured hydrophobic surfaces. The enhancements in both cases are attributed to the induced roughness and improved wettability. In this section, we briefly discuss three examples where surface wettability affects the heat transfer and phase-change process, sometimes even in unexpected ways.

Hierarchical Surfaces for Pool Boiling Heat Transfer

Boiling heat transfer is used in various large-scale industrial applications including water purification, desalination, and power generation. Consequently, modest enhancements in boiling heat transfer would lower energy consumption and cost by making the system more efficient.

Boiling heat transfer is driven by microlayer and interline evaporation, transient conduction, and microconvection as bubbles nucleate, grow, and depart from the heated surface [9]. As the heat flux increases, however, a competition arises between the vapor generated at the nucleation site and the replenishing of the liquid. At a certain point, the system reaches an operational limit beyond which the boiling heat transfer performance decreases due to the formation of a thermally insulating vapor layer that separates the heated surface from the working fluid [10]. This operational limit is termed the critical heat flux (CHF), and it signifies the beginning of transition boiling.

Recent advances in micro-/nanofabrication techniques have produced high-surface-area superhydrophilic surfaces that are capable of

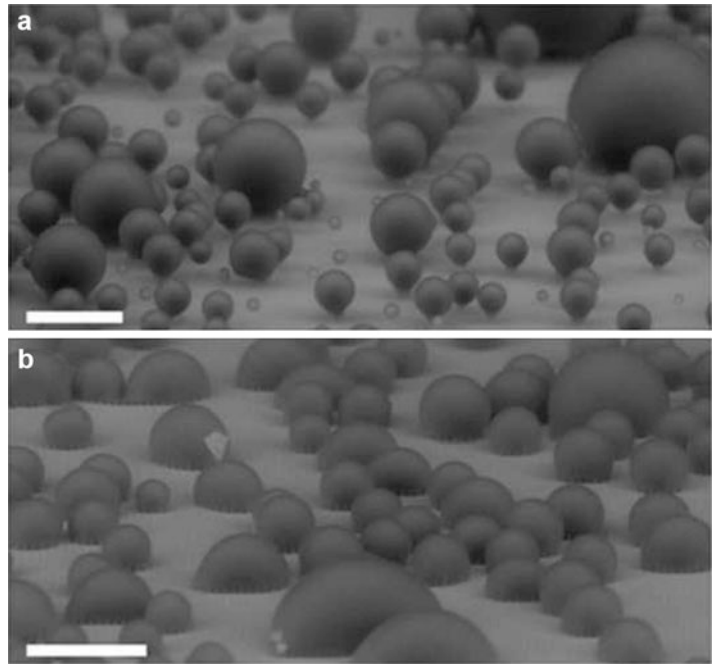


Surface Engineering, Tailored Wettability, and Applications, Fig. 4 Scanning electron micrograph of the representative, fabricated silica- and CuO-based hierarchical surfaces. (**a**) Electrophoretic deposition (EPD)-coated silica; micropillars in a square array with heights of 10 μm, diameters of 10 μm, and spacings of 5 μm; and (**b**) CuO micropillars in a square array with heights of 61 μm, diameters of 30 μm, and spacings of 30 μm. (**c**) Magnified view of the silica-based micropillar and EPD-coated SiO_2 nanoparticles (*inset)*. (**d**) Magnified view of the CuO micropillar and CuO nanostructures formed on the surfaces (inset). (**e**) Cross section view of the silica-based micropillar and (**f**) cross section view of the CuO micropillar obtained using focused ion beam milling (Reproduced with permission from Chu et al. [12])

increasing the CHF limit [11, 12]. Single length scale microstructured surfaces have increased the CHF of water to ≈150−210 W/cm^2. Chu et al. [12] further increased the CHF to ≈250 W/cm^2 using multiple length scale roughness or hierarchical surfaces which were fabricated using electroplated copper microstructures covered with copper oxide (CuO). The CuO hierarchical surfaces have a roughness factor, defined by the ratio of the overall area to the projected area, as high as 13.3. The large enhancements in CHF on the hierarchical surfaces (up to 200 % compared to smooth SiO_2 surfaces) were attributed to the increased roughness factor which provides an even larger surface force to counteract the momentum force from evaporation. Both experiments and modeling show that the roughness-amplified surface forces contribute significantly toward improving the CHF limit in pool boiling on micro-/nanostructured hierarchical surfaces (Fig. 4).

Superhydrophobic Surfaces for Enhanced Condensation

During condensation heat transfer, water droplets preferentially condense on rough solid surfaces (heterogeneous condensation) rather than directly from the vapor phase (homogeneous condensation) due to the smaller energy barrier or activation energy. Condensing vapor on typical industrial surfaces such as copper and stainless steel forms a thin liquid film because of the high surface energy of these metals. As a result, the condensing liquid film creates additional thermal resistance in the heat conduction path lowering the thermal performance and efficiency. This mode of condensation is called filmwise [13]. On the other hand, if the surface is chemically coated to lower the surface energy, the vapor condenses on the surface with discrete liquid droplets. This mode of condensation is called dropwise [14]. In dropwise condensation, gravity

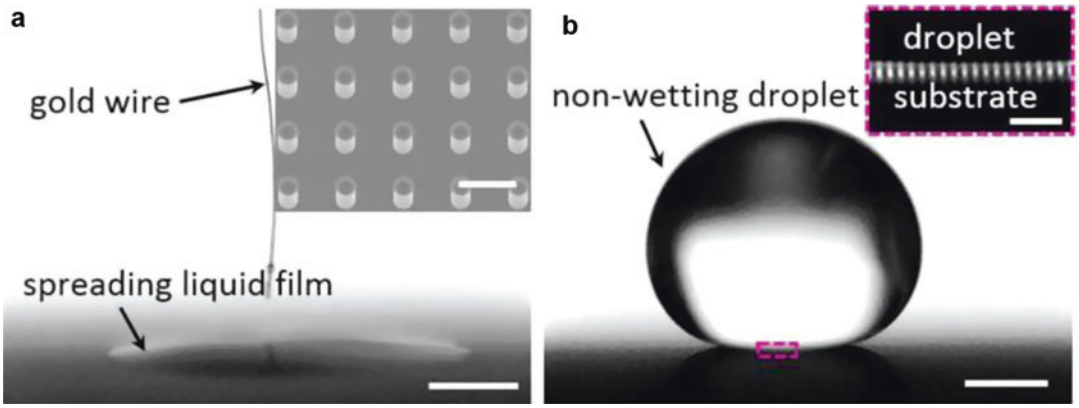


Surface Engineering, Tailored Wettability, and Applications, Fig. 5 Droplet morphology during condensation. Condensed droplet growth observed using environmental scanning electron microscope (ESEM) on Si nanopillars with (**a**) Cassie droplets with pitch spacing $l = 2$ μm and (**b**) Wenzel droplets with $l = 4$ μm. The diameter and height of the pillars were $d = 300$ nm and $h = 6.1$ μm, respectively. The surfaces were functionalized with a silane having an intrinsic advancing contact angle of $\theta_a \approx 22°$. Scale bars = 60 μm (Reproduced with permission from Enright et al. [18])

removes the condensing droplets when they reach the capillary length (≈2.7 mm for water) and refreshes the surface preparing it for re-nucleation. As a result, dropwise condensation has been demonstrated to improve the thermal performance by five to seven times when compared to filmwise condensation [15].

Recent advances in micro-/nanofabrication have enabled the development of micro-/nanostructured hydrophobic surfaces that allow greater mobility of condensing droplets (Fig. 5). If properly designed, the small length scale and solid fraction of these superhydrophobic surfaces enable the condensing droplets to move without significant resistance, promoting the merging of neighboring droplets. When droplets merge and form a bigger droplet, the excess surface energy created during coalescence is converted to kinetic energy enabling the merged droplet to jump off the surface. This mode of condensation is called jumping droplet condensation [16]. The coalescence-induced droplet ejection can greatly improve the overall efficiency of the condensation process on superhydrophobic surfaces [16]. Compared to state-of-the-art dropwise condensation on copper surfaces, jumping droplet condensation has demonstrated a 30 % increase in the heat transfer coefficient (≈92 kW/m^2-K) [17]. The initiation and growth of the condensate droplet depends on the surface-droplet and droplet-droplet interaction [18]. A unified model coupling individual droplet heat transfer, droplet size distribution, and wetting morphology suggests that a range of geometries from 0.5 to 2 μm promise a 190 % overall heat flux enhancement of jumping droplet condensation over conventional dropwise condensation on flat surfaces [17].

Jumping droplet condensation, however, cannot be sustained when the nucleation density is very high compared to structure density. In such cases, neighboring droplets coalesce and form discrete non-jumping droplets that strongly



Surface Engineering, Tailored Wettability, and Applications, Fig. 6 (**a**) A droplet deposited on a superhydrophilic surface at 120 °C spontaneously spreads into a thin film and wets the surface. The *inset* shows the scanning electron micrograph of the microstructured surface. Scale bars in the figure and the inset are 0.5 mm and 20 μm, respectively. (**b**) A similar-sized droplet at an elevated temperature (≈160 °C) did not wet the same surface; instead, it rested on top of the microstructured surface. The *inset* shows a magnified view of the boxed section near the droplet base indicating that the droplet remained in contact with the pillar tops. Scale bars in the figure and the inset are 0.5 mm and 100 μm, respectively (Reproduced with permission from Adera et al. [20])

adhere to the surface. This mode of condensation is called flooding [17]. The flooding creates an additional thermal resistance to the heat conduction path that degrades the thermal performance and reduces the heat transfer coefficient (≈44 kW/m^2-K) [17].

Evaporation-induced Non-wetting Water Droplets on Superhydrophilic Surfaces

Microstructuring along with chemical functionalization is extensively used in enhancing surface wettability at room temperature [6, 19]. However, the wetting behavior of droplets at room temperature does not necessarily apply to microstructured surfaces heated above the saturation temperature. Unlike their wetting behavior at room temperature, water droplets exhibit drastically different and yet potentially useful non-wetting behavior on nominally hot superhydrophilic surfaces [20].

A water droplet deposited on a rough hydrophilic surface at room temperature spreads spontaneously with vanishing contact angle due to the induced roughness (Fig. 6a). However, when the microstructured hydrophilic (i.e., superhydrophilic) surface is heated slightly above the saturation temperature, it can sustain non-wetting water droplets with contact angles as high as ≈160° (a superhydrophobic behavior) due to induced evaporation (Fig. 6b) [20]. The temperature at which this change in the wetting behavior occurs is significantly lower than the classical Leidenfrost temperature [10]. Moreover, the evaporation-induced non-wetting water droplets remain in contact with the pillar tops (inset Fig. 6b) making them distinctively different from Leidenfrost drops where solid-liquid contact is absent due to the presence of a thin (10–100 μm) [21] and yet stable vapor cushion.

The evaporation-induced non-wetting behavior of droplets on hot superhydrophilic surfaces is explained by using a model that analyzes the forces acting on the suspended droplet. Surface tension force acts to induce wetting due to the hydrophilicity of the surface. The heat conduction through the pillars causes the droplet to evaporate from its base. Due to the low permeability of the porous media, the radially escaping vapor underneath the droplet gives rise to a pressure gradient and an upward pushing force that counteracts wetting. By balancing the two competing forces (surface tension and pressure), the model predicts the superheat that is required for a water droplet to

reside on an arrays of silicon micropillars without wetting.

The presence of micropillar arrays has increased the effective thermal conductivity and decreased the vapor permeability of the porous media. As a result, water droplets deposited on nominally heated microstructured hydrophilic surfaces remain non-wetting. The mechanistic insight gained from this study provides design guidelines. The use of low thermal conductivity, high aspect ratio, and small-scale, highly porous structures is useful for applications requiring wetting droplets. Conversely, large-scale structures with high thermal conductivity should be used to reduce hydrodynamic drag on nominally heated microstructured hydrophilic surfaces.

Cross-References

► Wetting Transitions

References

1. Young, T.: An essay on the cohesion of fluids. Philos. Trans. R. Soc. Lond. **95**, 65–87 (1805)
2. Seemann, R., Brinkmann, M., Kramer, E.J., Lange, F. F., Lipowsky, R.: Wetting morphologies at microstructured surfaces. Proc. Natl. Acad. Sci. U. S. A. **102**, 1848–1852 (2005)
3. Huang, Y.-F., et al.: Improved broadband and quasi-omnidirectional anti-reflection properties with biomimetic silicon nanostructures. Nat. Nanotechnol. **2**, 770–774 (2007)
4. Vakarelski, I.U., Patankar, N.A., Marston, J.O., Chan, D.Y., Thoroddsen, S.T.: Stabilization of Leidenfrost vapour layer by textured superhydrophobic surfaces. Nature **489**, 274–277 (2012)
5. Launay, S., Fedorov, A., Joshi, Y., Cao, A., Ajayan, P.: Hybrid micro-nano structured thermal interfaces for pool boiling heat transfer enhancement. Microelectron. J. **37**, 1158–1164 (2006)
6. Quéré, D.: Wetting and roughness. Ann. Rev. Mater. Res. **38**, 71–99 (2008)
7. Xiao, R., Chu, K.-H., Wang, E.N.: Multilayer liquid spreading on superhydrophilic nanostructured surfaces. Appl. Phys. Lett. **94**, 193104 (2009)
8. Chu, K.-H., Xiao, R., Wang, E.N.: Uni-directional liquid spreading on asymmetric nanostructured surfaces. Nat. Mater. **9**, 413–417 (2010)
9. Kandlikar, S.G.: A theoretical model to predict pool boiling CHF incorporating effects of contact angle and orientation. J. Heat Transf. **123**, 1071–1079 (2001)
10. Leidenfrost, J.G.: On the fixation of water in diverse fire. Int. J. Heat Mass Transf. **9**, 1153–1166 (1966)
11. Chu, K.-H., Enright, R., Wang, E.N.: Structured surfaces for enhanced pool boiling heat transfer. Appl. Phys. Lett. **100**, 241603 (2012)
12. Chu, K.-H., Joung, Y.S., Enright, R., Buie, C.R., Wang, E.N.: Hierarchically structured surfaces for boiling critical heat flux enhancement. Appl. Phys. Lett. **102**, 151602 (2013)
13. Kutateladze, S., Gogonin, I.: Heat transfer in film condensation of slowly moving vapour. Int. J. Heat Mass Transf. **22**, 1593–1599 (1979)
14. Bonner, R.W.: 2010 14th International Heat Transfer Conference 221–226 (American Society of Mechanical Engineers)
15. Sikarwar, B.S., Khandekar, S., Agrawal, S., Kumar, S., Muralidhar, K.: Dropwise condensation studies on multiple scales. Heat Trans. Eng. **33**, 301–341 (2012)
16. Boreyko, J.B., Chen, C.-H.: Self-propelled dropwise condensate on superhydrophobic surfaces. Phys. Rev. Lett. **103**, 184501 (2009)
17. Miljkovic, N., Wang, E.N.: Condensation heat transfer on superhydrophobic surfaces. MRS Bull. **38**, 397–406 (2013)
18. Enright, R., Miljkovic, N., Al-Obeidi, A., Thompson, C.V., Wang, E.N.: Condensation on superhydrophobic surfaces: The role of local energy barriers and structure length scale. Langmuir **28**, 14424–14432 (2012)
19. Wenzel, R.N.: Resistance of solid surfaces to wetting by water. Ind. Eng. Chem. **28**, 988–994 (1936)
20. Adera, S., Raj, R., Enright, R., Wang, E.N.: Non-wetting droplets on hot superhydrophilic surfaces. Nat. Commun. **4**, 2518 (2013)
21. Biance, A.-L., Clanet, C., Quere, D.: Leidenfrost drops. Phys. Fluids **15**, 1632–1637 (2003)

Surface Figuring

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Surface Force Balance

► Surface Forces Apparatus

Surface Forces Apparatus

Carlos Drummond[1] and Marina Ruths[2]
[1]Centre de Recherche Paul Pascal, CNRS–Université Bordeaux 1, Pessac, France
[2]Department of Chemistry, University of Massachusetts Lowell, Lowell, MA, USA

Synonyms

Surface force balance

Definition

The surface forces apparatus (SFA) is an instrument for sensitive measurements of normal and lateral forces between two macroscopic surfaces in contact or separated by a thin film. The surface separation distance can be measured and independently controlled to 0.1 nm. The surfaces typically form a single-asperity contact where the substrates deform elastically, and time- and rate-dependent effects in the measured forces can be ascribed to phenomena in the thin films or adsorbed monolayers confined between the surfaces.

Basics of the SFA Technique

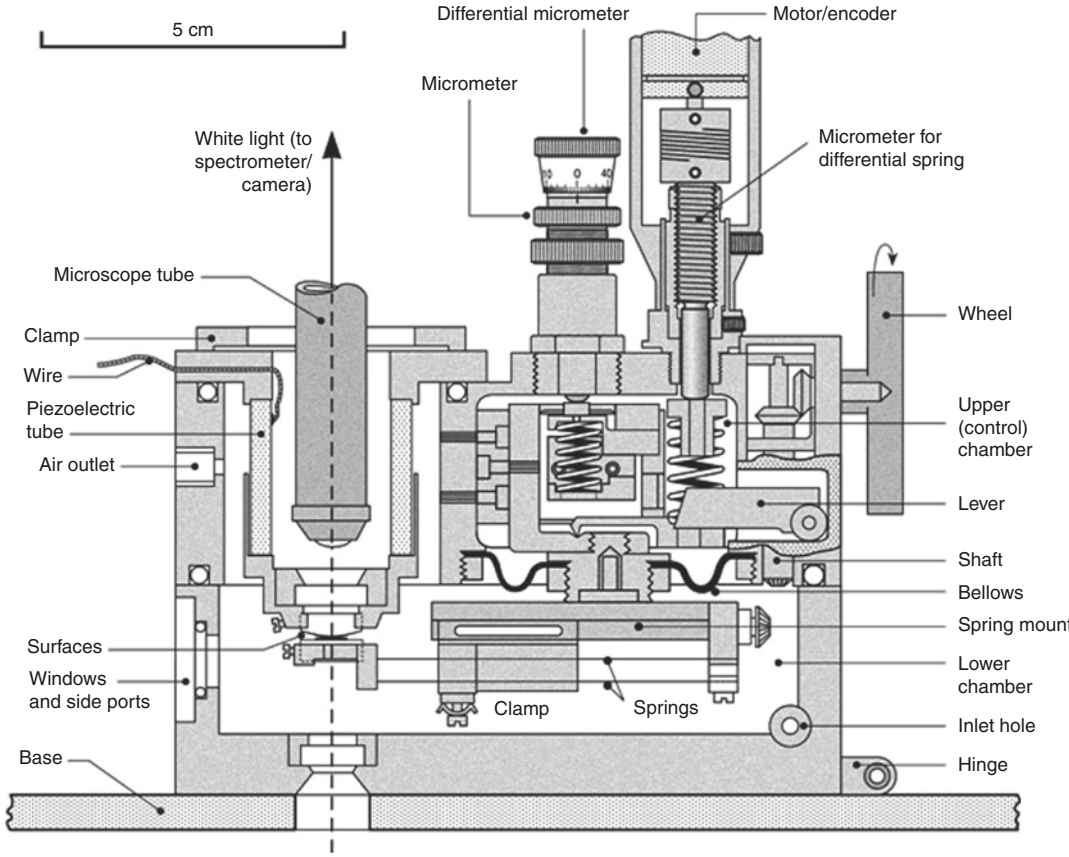
The Surfaces and the Apparatus

In SFA experiments, normal and lateral interaction forces are measured between two surfaces across air or a medium (a confined film). The most commonly used surface substrates are back-silvered, molecularly smooth muscovite mica sheets glued to half-cylindrical fused-silica disks. These half-cylindrical surfaces are mounted in the SFA in a crossed-cylinder configuration (cf. Figs. 1 and 2). At surface separations (distances D) much smaller than the radius of curvature (R) of the surfaces, this is equivalent to a sphere-on-flat geometry, which presents several advantages over that of two parallel plates: Alignment issues and edge effects are avoided, and different points of contact between the two surfaces can be easily investigated by displacing the disks laterally (e.g., to check the repeatability of measurements at different regions of a pair of surfaces or to move away from wear debris or contamination). In addition, the normal force F between cylindrical surfaces is related to the interaction energy between flat surfaces W according to the Derjaguin approximation [4], $F(D)/R = 2\pi W$, which provides a convenient means for comparison of experimental data with model predictions. For this reason, and to allow quantitative comparison of data from different experiments, normal force data are typically presented normalized by R.

Several different SFA setups have been developed for the measurements of normal and lateral forces. Early designs by Tabor, Winterton, and Israelachvili [2] were refined for measurements in liquids and vapors (the Mk II model) [4, 5]. More recent models are easier to assemble and clean (the Mk IV [6]) or have improved distance controls and many different attachments (the Mk III/SFA3 [1] in Figs. 1 and 2 and the SFA2000 [7]). The newer designs are also more user-friendly and have improved stability against thermal and mechanical drifts.

Normal Distance and Force Measurements

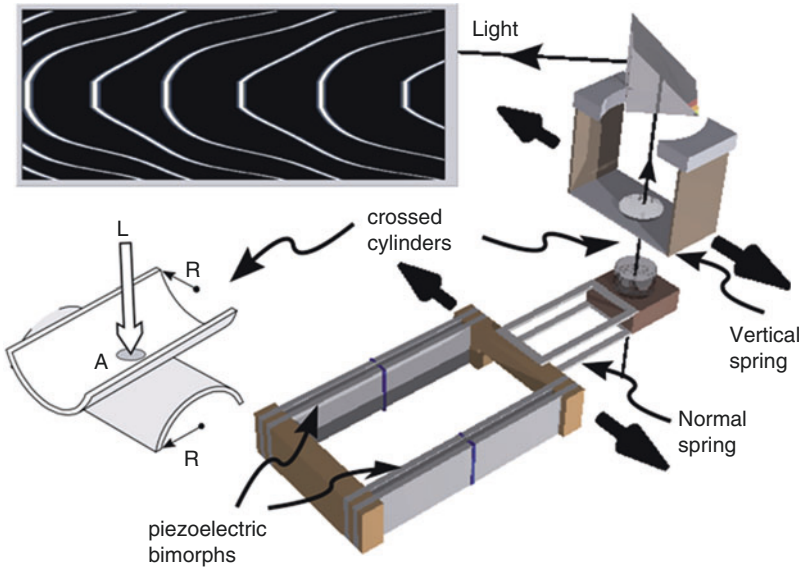
The interaction forces normal to the surfaces are determined from the deflection of a spring supporting the lower surface (Fig. 1). Typically, a double-cantilever spring is used to minimize tilting and/or sliding of the surfaces. Its spring constant, k, is determined by putting small weights on its end and measuring the resulting deflection with a traveling optical microscope. During an experiment to determine normal force, F, as a function of distance (surface separation), D, the spring deflection is found by monitoring the change in distance with multiple beam interferometry (MBI) [8] as the surfaces are moved



Surface Forces Apparatus, Fig. 1 Schematic drawing of the SFA3 (Mk III) configured for measurements of forces normal to the surfaces [1]. The lower surface is supported on a "force-measuring" double-cantilever spring with movable clamping, and the upper surface is mounted on a holder containing a piezoelectric tube for fine control of the separation distance between the surfaces, D. The mechanical distance control system (a system of weak and stiff springs reducing the movement of the manual and motor-driven micrometers) is separated from the main chamber of the instrument by Teflon bellows. The path of the light used for multiple beam interferometry (MBI) is indicated with an *arrow* (Reprinted with permission from Ref. [2])

toward or away from one another using motorized stages (for moving the base of the spring, the "spring mount" in Fig. 1) or a piezoelectric actuator (for moving the top surface). Any deviation in measured distance from the one expected from a calibration of the movement at large separations (where no forces act between the surfaces) represents a deflection of the spring, ΔD. The change in force when moving from one position to another is $\Delta F = k\Delta D$, according to Hooke's law.

Regions of a force versus distance curve where the gradient of the force, $\mathrm{d}F/\mathrm{d}D$, exceeds k are inaccessible to the technique due to mechanical instability, and the surfaces will spontaneously jump from one stable region to the next. The choice of spring stiffness is therefore of importance for the detection of different regions of the force curve, and several SFA designs incorporate springs whose length can be changed during the experiment (by moving a clamp along the spring, cf. Fig. 1). The use of force feedback to control the force applied to the surfaces independently of the displacement has also been suggested, to maintain a constant deflection of the spring. Implementations of this include the use of a magnetic force transducer and a bimorph deflection sensor [9] or a feedback system utilizing capacitive displacement transducers [10].



Surface Forces Apparatus, Fig. 2 Schematic illustration of the SFA tribometer design by Israelachvili and coworkers. As in Fig. 1, two back-silvered mica surfaces with a radius of curvature R are mounted in a crossed-cylinder configuration. The FECO arising from flattening of these surfaces over a contact area A at a given load L are shown in the *top* graph. (The alternating shapes of the fringes arise from odd and even orders of interference.) The lower surface is mounted on a double-cantilever "normal spring" (spring constant k) attached to a piezoelectric bimorph device allowing lateral movement of the lower surface in the direction of the arrows. The upper surface is mounted on a friction sensing device with vertical springs whose deflection is monitored with resistance strain gauges. The direction of incoming and emerging light is indicated with *arrows* (Adapted with permission from Ref. [3])

Multiple Beam Interferometry (MBI)

The use of multiple beam interferometry (MBI) [8, 11] makes it possible to determine the distance D between the surfaces with subnanometer resolution (0.1 nm) and to measure the refractive index of confined films to 0.01. MBI is a particular useful feature of the standard SFA setup, since it allows real-time, in situ monitoring of the geometry of the interacting surfaces and of the region of contact. Parameters such as the film thickness (including the thickness of individual molecular layers) and the diameter or radius of the contact area (to a lateral resolution of about 1 μm) can be readily measured. The occurrence of protrusions or regions of different refractive index can also be detected, which may indicate wear or accumulation of material in some region of the contact.

The mica surfaces, any material confined between them, and the semitransparent, highly reflective silver layers on their back sides form a built-in Fabry-Perot interferometer. White light is passed through the lower surface, and the wavelengths that interfere constructively after multiple reflections between the silver mirrors emerge through the top surface. This light is focused onto the slit of a grating spectrometer. The resulting pattern, fringes of equal chromatic order (FECO, cf. top left of Fig. 2), represents the shape and thickness of the space between the silver layers and depends on the optical path length through the materials in the interferometer (the thicknesses and refractive indexes of the mica sheets and any air gap or material confined between them). A measurement of the interfering wavelengths from surfaces in mica-mica contact in dry N_2 gas is used as the basis for calculation of distances (D) between the front sides of the mica surfaces (and thus of film thicknesses of confined materials), since an increase in optical path causes the fringes to shift to larger wavelengths with respect to this value. Relationships between the measured wavelengths, refractive indexes, and

D for a symmetric interferometer (two mica sheets of equal thickness) [8] have later been extended to more complicated arrangements such as asymmetric, absorbing, anisotropic, or more complicated multilayer systems [11].

SFA-Based Tribometers

Control of Film Thickness and Contact Area

Since the contact geometry in the SFA closely resembles a sphere-on-flat contact and the surfaces can be very smooth (atomically smooth in the case of bare mica surfaces), a single-asperity contact can be formed whose size, position, and deformation are easily monitored using MBI. This contact geometry is also advantageous for comparisons with contact mechanics models [4]. In nanotribological experiments, the film thickness and applied load (normal force) are typically regulated with the motor-driven distance controls. If the surfaces adhere or if the load is sufficiently high, the glue layers under the mica sheets deform elastically [2, 4] and a flat, circular contact area of uniform film thickness T and area of contact A is formed, as shown schematically in Fig. 2 (drawing on left). The corresponding FECO (Fig. 2, top left) show a characteristic flattening indicating a region of uniform film thickness. The diameter of the contact area can be measured directly (to about 1 μm) from this flattened region of the FECO (observations can be made in two orthogonal directions by rotating the light directed to the spectrometer with a dove prism, not shown here). Surfaces with a radius of curvature of $R \approx 2$ cm typically give rise to contacts with a diameter of a few μm to a few hundred micrometers, depending on the load and strength of adhesion. The resulting maximum pressure in the middle of the contact is typically at most a few tens of megapascals.

Any changes in shape of the surfaces or in thickness of the confined film can be observed directly on the FECO and measured with the same resolution as during a measurement of the normal force. Thickening and thinning of the confined film can be observed, as well as shear-induced hydrodynamic deformations. Furthermore, any damage of the surfaces can be easily detected as soon as it occurs, which enables one to identify sliding conditions that lead to wear and investigate wearless friction versus friction with wear.

Identification and accurate measurements of the contact area are very important for the analysis of nanotribological experiments. In many cases, the measured friction force mainly arises from the region of highest confinement, i.e., from the film in the flattened contact. Certain systems, especially ones where there is adhesion between the surfaces, give rise to a friction force that increases nonlinearly with load and appears to be proportional to this contact area [2]. Unlike many nonadhesive systems, the friction of adhesive contacts is thus not well described by the commonly used friction coefficient, μ (the slope of the friction force as a function of load). They are better characterized by their shear stress, σ, which is the friction force normalized by the contact area. In SFA experiments on a variety of thin films, the shear stress depended on the film structure and thickness and typically remained constant over the investigated range of loads. However, there are situations where the shear stress may vary with pressure, and it is important to recognize that because of the curved surfaces in the SFA, the pressure in the direction normal to the surfaces is not a constant over the flattened contact area but varies in a manner that can be described by contact mechanics models [4].

One can also envision a more complicated situation where regions of the surfaces outside the flattened area contribute to the friction response. This may occur, for example, if molecules are able to bridge a relatively large gap between the surfaces, and their "bonds" to the surfaces have to be broken and reformed during sliding. One possibility to define the contact area is to adopt a cutoff length and assume that the contributions to the frictional force are negligible over larger separation distances. Information on the extension of molecules from the surface can be obtained from measurements of the normal force, but it is still difficult to estimate how a region outside the contact will contribute to the total measured friction force. However, this problem

is shared by most experimental techniques in nanotribology, and the SFA coupled with the MBI technique provides the most direct way to directly observe the contact geometry during sliding.

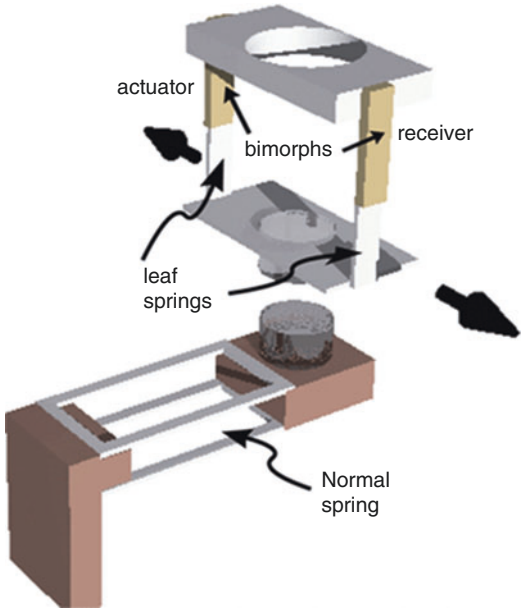
Shear and Friction Attachments and Measurements

Several types of shear and friction attachments for the SFA have been developed in the past 30 years, each with its own capabilities and limitations. These setups can be used to investigate different regimes of sliding velocity, magnitude of the friction forces, and sliding distance (amplitude). The most commonly used setups are discussed below. All of these setups have in common that the mechanical properties of the system (i.e., its compliance and inertial mass) will influence the results, and these factors have to be taken into account in order to obtain meaningful information from the measured signals. Because of the mechanical simplicity and easily characterized mechanical properties of the SFA, this can be done in a more straightforward manner than in many other devices used for nanotribological investigations. Detailed descriptions of the devices and of experimental data obtained on different systems can be found in the original publications.

Several devices for lateral shear and friction measurements have been developed by Israelachvili and coworkers [3, 7]. A schematic illustration of a current version is shown in Fig. 2. The lower surface is mounted on a double-cantilever spring (of known spring constant k) used to apply and measure the load (normal force) based on expected (calibrated) versus directly measured distance changes. This spring is attached to a "bimorph slider" consisting of sectored piezoelectric elements (electromechanical transducers) [3]. As a constant slope voltage ramp (triangular waveform) is applied over the electrodes of certain sectors of the bimorph elements, the lower surface is translated laterally in a linear manner. The maximum distance (amplitude) of the sliding motion is typically tens of micrometers, depending on the characteristics of the bimorph elements. The driving speed (sliding velocity of the lower surface) can be varied between a few Å/s to about 0.1 mm/s. The bimorph slider can also be used for nanorheological measurements if a constant frequency sinusoidal input is chosen instead of a triangular wave.

The detection of friction forces or viscoelastic responses is done with the device holding the upper surface [3, 7]. This "friction device" consists of a vertical double-cantilever metal spring (with a known spring constant K), whose deflection is measured using strain gauges forming the arms of a Wheatstone bridge. The force experienced by the upper surface due to the movement of the lower can thus be calculated from this spring deflection using Hooke's law. The detection limit depends on the stiffness of the spring and sensitivity of the strain gauges and is typically a few μN. In cases where larger displacements are desired than the ones available with the bimorph slider, the friction device can be used as both an actuator and detector: The base of the vertical double cantilever can be translated laterally (approx. ± 5 mm) using a reversible, variable speed motor-driven micrometer (not shown), and the deflection of the vertical springs recorded simultaneously. A combination of the bimorph slider and the friction device, or the friction device alone, has been used by Israelachvili and coworkers to study the shear or friction response of a wide range of systems such as highly confined simple liquids, polymer melts and solutions, self-assembled surfactant and polymer layers, and liquid crystals (see Refs. [2, 4]).

A different design, used to investigate smaller deformations and mainly the linear response of confined films, has been developed by Granick and coworkers [12]. Small deformations are advantageous for investigations of longtime relaxation processes that might be occurring in the contact region. Although mainly developed for the study of deformations of the order of the film thickness or less, the displacement range goes from less than 1 nm to a few hundred nanometers, with a reported force sensitivity of about 5 μN. This type of device is illustrated schematically in Fig. 3. The load (normal force) is regulated by adjusting the vertical position of the lower



Surface Forces Apparatus, Fig. 3 Schematic illustration of the SFA tribometer design by Granick and coworkers. Bending of one of the bimorph elements (the actuator) causes the upper surface to move laterally with a displacement detected with the other bimorph element (the receiver). The distance (film thickness) and load are controlled by moving the base of the spring holding the lower surface, as in the design in Fig. 1 (Adapted with permission from Ref [12])

surface, and this surface remains stationary in the lateral direction during shear experiments.

The upper surface is mounted on a double-cantilever device. Different from the design in Fig. 2, the double cantilever consists not of metal springs with strain gauges but of metal springs with attached piezoelectric bimorph elements. One element is used as an actuator and the other as a detector (receiver). In a typical experiment, a constant frequency sinusoidal input causes bending of the actuator bimorph. The output voltage from the detector bimorph is used to measure the actual displacement of the upper surface. The viscoelastic properties of the confined film is extracted by comparing these signals to those measured with the surfaces in strongly adhesive (mica-mica) contact and when separated, i.e., when the device is moving freely. The electromechanical characteristics of the system are modeled as a combination of effective masses, springs, and dashpots representing the different components of the apparatus [12]. In practice, the response in mica-mica contact (from the mica and glue layers) is typically purely elastic. This setup has been used by Granick and coworkers to study a variety of systems, including simple liquids, polymer melts, adsorbed layers, and solutions.

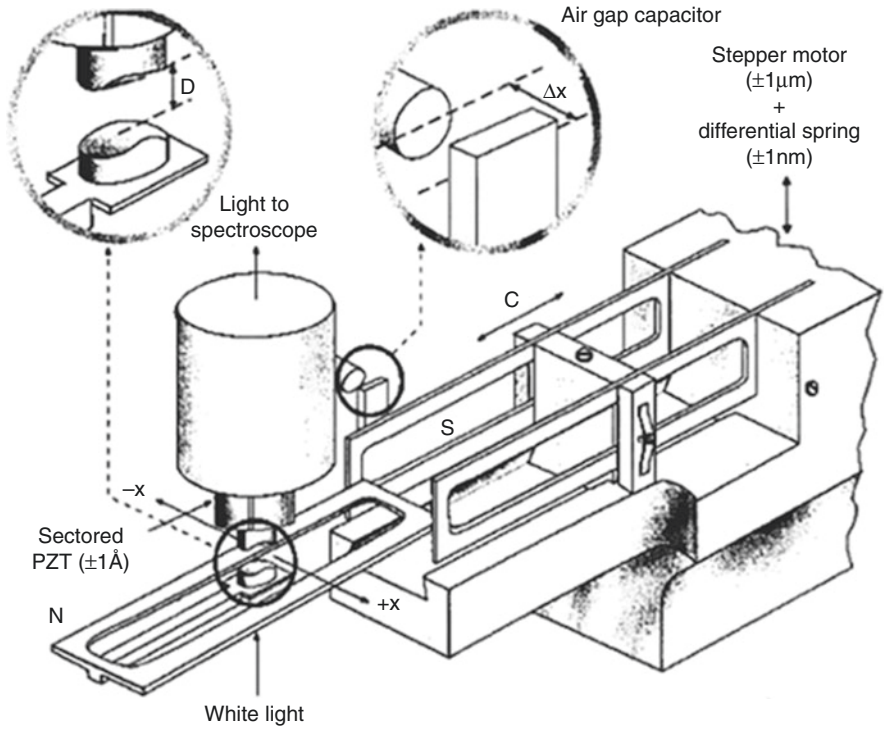
A third type of shear device, developed by Klein and coworkers [13], is shown schematically in Fig. 4. In this design, the displacement is induced by the upper surface, and the friction force is measured at the lower. The upper surface is attached to a sectored piezoelectric tube that is used to produce a normal or lateral displacement. The lower surface is mounted on a single-cantilever spring (for measurements of normal forces, N in Fig. 4), attached to a double-cantilever shear force spring (S in Fig. 4). The lateral displacement of this double-cantilever spring is measured using an air-gap capacitor, and the friction force calculated using Hooke's law. The reported sensitivity of this device is 50 nN, and the maximum displacement of the upper surface is a few tens of micrometers. In this device, the sensitivity to the measured friction forces is greatly improved with respect to the designs discussed above. The high sensitivity arises from the detection method and has proven to be very valuable for investigations by Klein and coworkers of films of water (where the friction can be very low) and simple liquids, as well as polymer melts and solutions.

Capacitance sensors have also been used in a two-dimensional friction measuring device with a reported sensitivity of 2 μN [14]. The friction detecting capabilities were shown for confined hexadecane with a film thickness of two and three molecular layers.

Recent Technique Developments

Advances in Distance and Force Measurements

Many different improvements to the original techniques and setups have been proposed and implemented in the past decades. Significant steps have been made to automate the normal



Surface Forces Apparatus, Fig. 4 Schematic drawing of the surface force balance (SFB) designed by Klein and coworkers and configured for nanotribological measurements. The distance between the surfaces and the lateral displacement of the upper surface is controlled with a sectored piezoelectric tube. The deflection of the shear force spring (S) is detected by an air-gap capacitor (Reprinted with permission from Ref. [13])

and lateral force measurement, and non-interferometric techniques have been introduced to measure the separation between the surfaces to enable the use of opaque substrates. Several designs for non-interferometric control and measurements of distance have incorporated piezoelectric bimorph elements [15, 16]. These devices have been used for thin-film viscosity measurements and studies of time- and rate-dependent adhesion in the laboratories of Israelachvili and Granick, respectively. Generally, piezoelectric elements are inadequate for long or quasi-static measurements because of their intrinsic drift and signal decay. The use of an ultrahigh impedance amplifier to lengthen the decay time of the bimorph sensor has to some extent helped overcome this difficulty. An entirely different method of distance detection is based on measuring the capacitance either between the silver layers on the back side of the mica sheets or between one plate of a capacitor attached to the chamber of the instrument and one attached to the moving surface [17]. These techniques allow fast and accurate measurements of the displacement of the surfaces and eliminate the constraint of having to use transparent surfaces. However, when opaque surfaces are used, it is not possible to obtain a real-time, in situ image of the contact region during shear, which is one of the major strengths of the standard SFA technique.

Local Structural Information: Combination of SFA with Other Techniques

The data gathered in a conventional SFA experiment is the result of an average response of the confined film to shear and compression. Information on the molecular-level film structure during these processes would add significantly to the interpretation of the measured forces. Obtaining such information is very challenging because of the nature of the confined region: The response arises from a relatively small number of

molecules, which implies that any characteristic spectroscopic signal from this region will be of low intensity. The confined film is surrounded by thicker layers of materials that will give rise to spectroscopic signals of their own, typically much stronger than that from the film. Despite these experimental difficulties, interesting observations on film structure and molecular alignment have been made by combining SFA with several different techniques, as briefly described below. Further development of techniques suitable for studying the molecular properties of highly confined films can be expected to significantly improve our understanding of friction phenomena.

Extensions of standard MBI have been used to obtain structural information on the orientation and intermolecular interactions of optically active molecules in confined films [18]. Since the light adsorption of optically active (dye) molecules is enhanced by the multiple reflections in the optical cavity, very thin films can be studied, and local information in the contact area can be obtained.

Combining the SFA with other optical techniques has been limited by the presence of the reflective silver layers needed for the standard MBI approach. These layers strongly reduce the illumination of confined films, limiting the performance of spectrometric techniques. This has been successfully overcome by replacing the silver layers with multilayer dielectric coatings [19], with transparency to different wavelengths, so that information on the structure of ultrathin films under shear can be obtained. For example, the SFA has been combined with fluorescence correlation spectroscopy to measure the molecular diffusion coefficient in thin films within spots of submicron size, obtaining spatially resolved measurements [19]. A drawback is that fluorescent molecules have to be added to the film to be investigated. A reduction in the diffusion coefficient by about two orders of magnitude was found with confinement of films of simple liquids, and the diffusion coefficient was found to decrease when going from the edges toward the center of the contact region. Results have also been reported from combining the SFA with confocal Raman spectroscopy [19], which avoids the problem of the bulk contributing to the scattered signal. Using multilayer reflective coatings, the geometry of the contact area could be monitored simultaneously with the Raman signal, showing the effects of shear on the orientations of molecules within the confined film.

The SFA has also been combined with x-ray diffraction measurements (the XSFA) [20]. Although this technique has thus far been limited to films thicker than 500 nm, the shear-induced alignment of liquid crystal molecules in confined films has been shown. Because of the size of the x-ray beam, the results represent an average over the contact area. Active research in this area is aiming at reducing the investigated region and film thickness, recently indicating that structural information on confined molecular films can be obtained with x-ray reflectivity [21].

The range of shear rates has been extended by incorporating a standard quartz crystal resonator as the lower surface in the SFA [22], giving a high oscillation frequency (MHz). A back-silvered mica piece was attached to the planar quartz crystal, and the top surface was mica glued to a half-spherical disk, so that a sphere-on-flat geometry could be obtained. Standard MBI was used to measure the diameter of the contact area and the radius of curvature of the sphere-on-flat contact geometry.

Mica and Beyond: Modified and Alternative Substrates

Muscovite mica is the most commonly used substrate in SFA experiments, because of its transparency and the relative ease by which large areas of atomically smooth, step-free sheets of uniform thickness (typically 2–5 μm) can be cleaved from larger crystals [5, 8]. The compressibility of mica is quite low, and it is chemically inert, i.e., in its native state, it does not expose functional groups to which chemical reactions can occur. On the one hand, this implies that its properties are quite well-known from one experiment to the other. On the other hand, a substantial effort is needed to modify mica for different purposes by physisorption or chemical reactions. The substrate plays a large role in the surface phenomena investigated by SFA, particularly in nanotribology, and unmodified (bare) mica is not necessarily

representative of many surfaces of engineering or biomedical interest. A large number of procedures have been developed for the purpose of modifying or replacing the mica surfaces with, for example, organic thin films, metals, or metal oxides to get surfaces with different chemical properties and surface energies (cf. Refs. [2, 4]).

Important considerations when modifying or replacing the mica substrates are the smoothness of the resulting surface and its transparency to light. An increase in roughness might change the standard single-asperity contact to a more complex, multi-asperity one. However, in many of the approaches described below, the roughness of the deposited layers can be controlled or modified at least to some extent, and in cases where rougher surfaces are formed, this allows investigations of the effects of surface roughness on the measured forces, which in itself is an important field of research. Electrochemical experiments in the SFA have shown that surface roughness affects the counterion distribution and thus the double-layer forces measured between charged surfaces [23] and have also demonstrated the possibility of electrochemical in situ modification.

Surface modifications by deposition of thin organic layers (often monolayers) can often be done so that the smoothness of the surface is close to that of the mica substrates themselves. Many different types of surface-active molecules (surfactants, lipids, block copolymers, proteins) have been deposited by self-assembly from solution or by Langmuir-Blodgett deposition [2, 4]. For example, in aqueous solution, the mica surface becomes negatively charged (because of solvation of K^+ ions), and positively charged species spontaneously adsorb onto it. The friction response of the modified surfaces depends strongly on the properties of the adsorbed layers such as their surface energy, morphology, and inherent stiffness [2, 4]. The attachment of organic layers to the mica surface can be enhanced by chemical modification of the mica itself, for example, by water vapor plasma treatment. This enables the chemical binding of chlorosilane species to form robust, molecularly smooth, hydrophobic surfaces.

Mica surfaces have also been used as substrates for deposition of various inorganic materials, including metals and dielectrics. In many of these cases, the inertness of the native mica increases the risk of dewetting of the deposited layer, and some pretreatment of the mica (e.g., plasma treatment or the deposition of a separate adhesion layer) is necessary.

The analysis of the resulting FECO becomes more complicated because of the larger number of optical layers in the interferometer, but algorithms for this are available [11]. Thin deposited layers of materials such as silver, gold, platinum, silica, alumina, and zirconia with various film thicknesses and smoothness have been investigated [2, 4].

The interface between a supporting mica sheet and a material deposited on it has a roughness similar to that of the mica sheet itself. Because of the inertness of the mica, the bond between the mica and the deposited material is weak, and this can be utilized to form so-called template-stripped surfaces for use in the SFA [24]. The material of interest, for example, gold, is deposited on a freshly cleaved mica sheet to a desired thickness; its exposed surface is attached to another substrate (e.g., by gluing), and the mica sheet is peeled off with tweezers or with adhesive tape to expose a very smooth gold substrate that can be further modified through chemical reactions.

Substrates not based on mica or on a sacrificial mica substrate have also been developed and successfully investigated (cf. Ref. [4] and references therein): Among these, the earliest example is sapphire (aluminum oxide) single crystals grown from the vapor phase. More recently, silicon nitride surfaces have been formed by plasma-enhanced chemical vapor deposition onto rigid silica disks coated with a reflective layer [25]. A more easily prepared substrate, sheets of silica (quartz glass) or borosilicate glass [26], collected from a bubble with a thickness of a few to 10 μm formed by standard glassblowing techniques, has been used in a couple of studies. These flexible glass substrates, which are quite robust and have a roughness of only a fraction of a nanometer due to the surface tension of the glass when molten, can be silvered on their back side and used in a similar

manner to mica sheets. Surface modifications developed for glass and silica substrates (e.g., chemical bonding of silanes) can be readily applied to these substrates. Thin sheets of polymers have also been successfully prepared by casting and stretching for use in adhesion studies [27].

Cross-References

- ▶ Atomic Force Microscopy
- ▶ Disjoining Pressure and Capillary Adhesion
- ▶ Friction Force Microscopy
- ▶ Nanotribology

References

1. Israelachvili, J.N., McGuiggan, P.M.: Adhesion and short-range forces between surfaces. Part 1: new apparatus for surface force measurements. J. Mater. Res. **5**, 2223–2231 (1990)
2. Ruths, M., Israelachvili, J.N.: Surface forces and nanorheology in molecularly thin films. In: Bhushan, B. (ed.) Springer Handbook of Nanotechnology, 3rd edn, pp. 857–922. Springer, Berlin/Heidelberg (2010), and references therein
3. Luengo, G., Schmitt, F.-J., Hill, R., Israelachvili, J.: Thin film rheology and tribology of confined polymer melts: contrasts with bulk properties. Macromolecules **30**, 2482–2494 (1997), and references therein
4. Israelachvili, J.N.: Intermolecular and Surface Forces, 3rd edn. Academic, Amsterdam (2011), and references therein
5. Israelachvili, J.N., Adams, G.E.: Measurements of forces between two mica surfaces in aqueous electrolyte solutions in the range 0-100 nm. J. Chem. Soc. Faraday Trans. I **74**, 975–1001 (1978)
6. Parker, J.L., Christenson, H.K., Ninham, B.W.: Device for measuring the force and separation between two surfaces down to molecular separations. Rev. Sci. Instrum. **60**, 3135–3138 (1989)
7. Israelachvili, J., Min, Y., Akbulut, M., Alig, A., Carver, C., Greene, W., Kristiansen, K., Meyer, E., Pesika, N., Rosenberg, K., Zeng, H.: Recent advances in the surface forces apparatus (SFA). Rep. Prog. Phys. **73**, 036601 (2010), and references therein
8. Israelachvili, J.N.: Thin film studies using multiple-beam interferometry. J. Colloid Interface Sci. **44**, 259–272 (1973)
9. Stewart, A.M., Parker, J.L.: Force feedback surface force apparatus: principles of operation. Rev. Sci. Instrum. **63**, 5626–5633 (1992)
10. Tonck, A., Georges, J.M., Loubet, J.L.: Measurements of intermolecular forces and the rheology of dodecane between alumina surfaces. J. Colloid Interface Sci. **126**, 150–163 (1988)
11. Heuberger, M.: The extended surface forces apparatus. Part I. Fast spectral correlation interferometry. Rev. Sci. Instrum. **72**, 1700–1707 (2001), and references therein
12. Peachey, J., Van Alsten, J., Granick, S.: Design of an apparatus to measure the shear response of ultrathin liquids films. Rev. Sci. Instrum. **62**, 463–473 (1991), and references therein
13. Raviv, U., Tadmor, R., Klein, J.: Shear and frictional interactions between adsorbed polymer layers in a good solvent. J. Phys. Chem. B **105**, 8125–8134 (2001), and references therein
14. Qian, L., Luengo, G., Douillet, D., Charlot, M., Dollat, X., Perez, E.: New two-dimensional friction force apparatus design for measuring shear forces at the nanometer scale. Rev. Sci. Instrum. **72**, 4171–4177 (2001)
15. Israelachvili, J.N., Kott, S.J., Fetters, L.J.: Measurements of dynamic interactions in thin films of polymer melts: the transition from simple to complex behavior. J. Polym. Sci. B **27**, 489–502 (1989)
16. Dhinojwala, A., Granick, S.: New approaches to measure interfacial rheology of confined films. J. Chem. Soc. Faraday Trans. **92**, 619–623 (1996)
17. Stewart, A.M.: Capacitance dilatometry attachment for a surface-force apparatus. Measure. Sci. Technol. **11**, 298–304 (2000)
18. Mächtle, P., Müller, C., Helm, C.A.: A thin absorbing layer at the center of a Fabry-Perot interferometer. J. Phys. II **4**, 481–500 (1994)
19. Bae, S.C., Wong, J.S., Kim, M., Jiang, S., Hong, L., Granick, S.: Using light to study boundary lubrication: spectroscopic study of confined films. Philos. Trans. A Math. Phys. Eng. Sci. **366**, 1443–1454 (2008), and references therein
20. Idziak, S.H.J., Koltover, I., Israelachvili, J.N., Safinya, C.R.: Structure in a confined smectic liquid crystal with competing surface and sample elasticities. Phys. Rev. Lett. **76**, 1477–1480 (1996)
21. Seeck, O.H., Kim, H., Lee, D.R., Shu, D., Kaendler, I. D., Basu, J.K., Sinha, S.K.: Observation of thickness quantization in liquid films confined to molecular dimension. Europhys. Lett. **60**, 376–382 (2002)
22. Berg, S., Ruths, M., Johannsmann, D.: High-frequency measurements of interfacial friction using quartz crystal resonators integrated into a surface forces apparatus. Phys. Rev. E **65**, 026119 (2002)
23. Valtiner, M., Banquy, X., Kristiansen, K., Greene, G. W., Israelachvili, J.N.: The electrochemical surface forces apparatus: the effect of surface roughness, electrostatic surface potentials, and anodic oxide growth on interaction forces, and friction between dissimilar surfaces in solution. Langmuir **28**, 13080–13093 (2012)

24. Chai, L., Klein, J.: Interaction between molecularly smooth gold and mica surfaces across aqueous solutions. Langmuir **25**, 11533–11540 (2009), and references therein
25. Golan, Y., Alcantar, N.A., Kuhl, T.L., Israelachvili, J.: Generic substrate for the surface forces apparatus: deposition and characterization of silicon nitride surfaces. Langmuir **16**, 6955–6960 (2000)
26. Horn, R.G., Smith, D.T., Haller, W.: Surface forces and viscosity of water measured between silica sheets. Chem. Phys. Lett. **162**, 404–408 (1989)
27. Merrill, W.W., Pocius, A.V., Thakker, B.V., Tirrell, M.: Direct measurement of molecular level adhesion forces between biaxially oriented solid polymer films. Langmuir **7**, 1975–1980 (1991)

Surface Free Energy and Chemical Potential at Nanoscale

▶ Surface Energy and Chemical Potential at Nanoscale

Surface Hopping

▶ Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Surface Loss in Micromechanical/Nanomechanical Resonators

▶ Surface Dissipations in NEMS/MEMS

Surface Loss in NEMS/MEMS

▶ Surface Dissipations in NEMS/MEMS

Surface Micro-patterning

▶ Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique

Surface Modeling of Ceramic Biomaterials

Marta Corno and Piero Ugliengo
Department of Chemistry, University of Torino, Torino, Italy

Synonyms

2D-slab approach; Computational studies; Simulations of solid interfaces

Definition

A surface of a generic solid material can be considered as the topmost boundary layer of the material. If a crystalline solid is theoretically described as the infinite repetition of a unit cell in the three dimensions of space, the surface can be considered as the greatest defect of the crystal itself, since it interrupts the periodicity, *though a real crystal is a macroscopic object made of finite atoms*. Indeed, physical and chemical phenomena are in a large number of cases occurring at the interface between a solid material and the environment, i.e., at its surface. To further highlight the relevance of surfaces in science, it is sufficient to mention that the 2007 Nobel Prize in Chemistry was awarded to Gerhard Ertl "for his studies of chemical processes on solid surfaces"[1]. To this respect in the last decades, computational methods have become relevant in the study of surface properties. Therefore, surface modeling has acquired a great interest, both theoretically and for practical applications in the chemical industry, e.g., in catalysis, adsorption, corrosion, and oxidation, and in the biomedical field. Especially in this last research area, complex mechanisms are involved in the interaction among material surfaces and body biological fluids. Surface modeling can then help in the progress of scientific and technological knowledge for the design of improved biomaterials to be integrated within the human body.

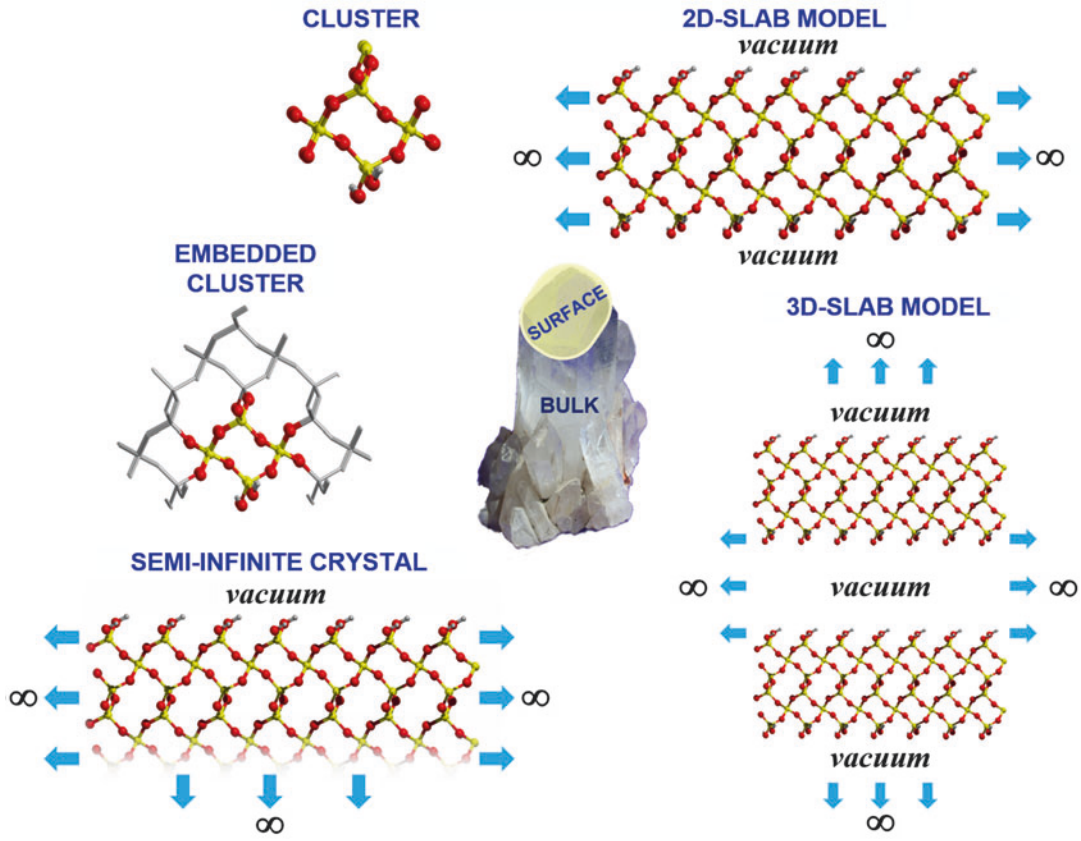
Introduction to Surface Modeling

Quantum mechanics is the fundamental physic framework for studying processes occurring at surfaces of solid materials, including both static and dynamic phenomena. For the latter, the treatment with a classical or semiclassical mechanics approach can be also mentioned, characterized by the development over the years of ad hoc force-fields. By means of the increasingly accurate computational techniques, experimental findings have become reproducible by simulations and often even clearer to be interpreted due to the synergic use of theory and experiments. In this sense, it is worth mentioning that Density Functional Theory (DFT) methods have been and currently are successfully coupled with experimental surface science techniques, as scanning tunneling microscopy (STM), temperature-programmed desorption (TPD), and X-ray diffraction, among others, to gain a deeper knowledge of surface properties for materials of different chemical nature (metals, oxides, nanoparticles, etc.) [2].

The complete and rigorous treatment of quantum-mechanical methods for computing surface properties is outside the scope of this essay. A more general overview of the most frequently used approaches to surface modeling will be exposed in the following sections.

Main Approaches to Surface Modeling

For modeling surfaces three main approaches are possible: cluster (or embedded cluster), semi-infinite crystal, and slab models (Fig. 1).



Surface Modeling of Ceramic Biomaterials, Fig. 1 Main approaches to surface modeling for the case of quartz. *Blue arrows* and the infinite symbol indicate periodicity; atoms are colored as follows: silicon in *yellow*, oxygen in *red* and hydrogen in *light gray*. In the embedded cluster model, atoms represented as gray sticks are those treated with a lower degree of accuracy with respect to the full colored ones

In the *molecular cluster* approach, a piece (cluster) of atoms belonging to the extended surface and representative of the specific site to be simulated is cut out from the surface itself. Obviously, the larger the size of the cluster the better the similarity with the original surface. Within this model no periodicity is considered, so the translational symmetry is lost and the point group to which the cluster belongs is the new reference for symmetry operations. Using this approach, caution has to be paid due to the introduction of spurious effects originated by the limited size and the need to add extra atoms to cap the unfilled valences at the cluster boundaries. Usually, hydrogen atoms are used to cover the cluster frontier. Because of the intrinsic limitations of the cluster model for representing surfaces, a number of factors must be taken into account to evaluate the reliability of the obtained results. First of all, the size has to be carefully chosen, by analyzing the dependence of the cluster physicochemical properties on the cluster size. Moreover, other fundamental requisites are electroneutrality and the same stoichiometry, average atomic coordination, and symmetry elements as in the bulk. To overcome the cluster size limitations, an alternative technique consists in embedding the cluster in an external field which mimics the missing bulk crystal environment. In this way the whole system is divided into subsystems, which are treated with different levels of theory and consequently accuracy (e.g., by adopting Green function techniques). The embedding technique is also commonly adopted when handling the simulation of defects in solids.

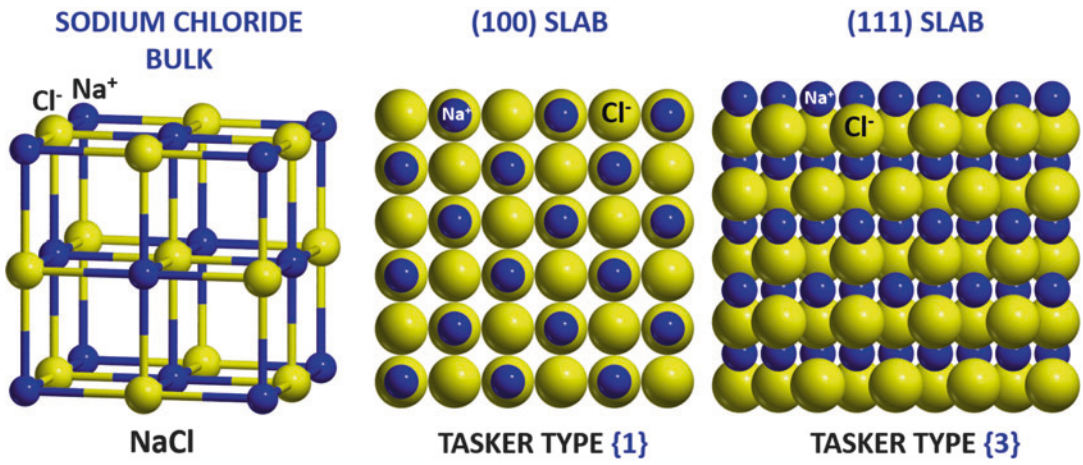
In the second approach to surface modeling, the *semi-infinite crystal* approach, two semi-infinite crystals are obtained by cutting the 3D crystal through a selected crystalline plane. The periodicity is maintained in 2D by the semicrystals, while is lost along the perpendicular z direction and the problem is reduced to an infinite + finite problem, that is the few layers close to the surface constitute a finite subsystem connected to the semicrystal with known electronic structure. This approach requires the theoretical framework of the Green functions and in the past has been successfully applied for alkali metals.

In the third and most common approach, the *slab model*, the bulk structure is cut to extract a layer of finite thickness exposing two surfaces. However, the first operation is the choice of the reference surface plane for cutting the crystal bulk. Conventionally, each surface is indicated by three integers, known as Miller indices (*hkl*), which identify the family of crystallographic planes to which the surface belongs. The most probable and stable surfaces are usually those characterized by the lowest Miller indices, as they envisage surfaces with the highest atomic density. Usually, the kind and extension of surfaces are derived from the experimental crystalline habit, or from the analysis of experimental images recorded by means of the transmission and scanning electron microscope.

The slab model can be implemented following two different schemes: (i) a 2D slab envisaging true vacuum space at top/bottom of the slab and extending through infinity, while keeping two-dimensional periodicity within the slab itself or (ii) a series of slabs is repeated along the direction perpendicular to the slab plane and separated by vacuum gaps, spaced apart enough to avoid spurious interactions among replicas. This latter case is indeed a pseudocrystal in which the unit cell to be repeated in 3D contains the slab of a given thickness plus the vacuum gap. The choice between the two schemes is associated to the choice of the adopted basis set functions used to represent the wave functions of the Schrödinger equation to be solved within the approximate quantum-mechanical method [3]. For instance, when plane waves functions are adopted (functional form $\phi_\kappa(r) = \exp^{ik\cdot r}$), only the 3D slab model is suitable, due to the inherent periodicity of plane waves. Conversely, computer programs based on localized Gaussian-type basis sets can adopt both models [4].

Surface Stability

When considering ionic crystalline materials, the stability of the corresponding modeled surfaces follows a classification proposed by Tasker, stating that three categories exist: (i) type 1 slab, with neutral planes or layers containing both anions and cations (same stoichiometry of the parent



Surface Modeling of Ceramic Biomaterials, Fig. 2 Example of Tasker type surface models for sodium chloride (NaCl): leftmost side, the structure of the corresponding sodium chloride bulk unit cell; center, lateral view of the NaCl (100) slab, characterized by six neutral layers, with both anions and cations in the two outermost layers (type 1); and right side, lateral view of the NaCl (111) slab, exhibiting ten alternating charged layers of anions and cations, with the two outermost conferring a net dipolar moment across the slab (type 3)

crystal); (ii) type 2 slab with charged planes or layers arranged symmetrically so that no net dipole moment is generated across the surface; and (iii) type 3 slab with charged planes or layers imparting a net dipole moment across the surface [5]. The definition of type 3 dipolar surface brings about an inherent instability due to the buildup of an infinite dipole moment with the increasing slab thickness ultimately breaking the system apart. Figure 2 reports an example of surfaces for sodium chloride, classified according to Taker types. One should be aware that the Tasker's classification assumes a truly ionic nature of the material in which no charge transfer or polarization is permitted. Obviously, this is not the case when the system is treated quantum mechanically and the nature of the material is not entirely ionic. In the following, the case of the dipolar (001) hydroxyapatite surface will be described, in which the role of polarization and charge transfer allow the existence of this surface up to thickness of hundredth of nanometers.

For both slab and multislab models, one crucial parameter to carefully evaluate is the number of layers forming the slab, i.e., its thickness. Since the slab is cut out from an infinite crystal, the thicker is the surface (or larger is the number of layers), the closer the slab will be to the "real" material. The energy needed to create the surface from the bulk (surface energy) is also dependent on the slab thickness. In principle, the correct thickness for a given slab implies a converged value of the surface energy with the slab thickness. In practice, one should compromise between accuracy of the converged surface energy (improving with the slab thickness) and cost of the calculation (increasing with the slab thickness).

The surface energy (E_s, γ or σ depending on the chosen formalism) is the energy required to cut the slab out of the bulk. Lower is the surface energy, more stable is the considered surface. The general definition of the surface energy is the following

$$E_{surf} = \frac{(E^{n}_{slab} - n \cdot E_{bulk})}{2A} \quad (1)$$

Where E^{n}_{slab} is the energy of an n-layered slab, E_{bulk} is the energy of a formula unit of the material in its bulk, n is the number of formula units in the slab model, A is the surface area, doubled due to the presence of two limiting surfaces. This definition is for symmetric slabs only, where the two limiting surfaces are identical. The same

definition is also adopted for nonsymmetrical cases, in which the E_{surf} is usually considered as an average surface energy value. The surface energy is a positive quantity, otherwise the bulk structure would exfoliate, and is expressed in units of J/m^2. By increasing the slab thickness ($n \rightarrow$ infinite), the surface energy should converge to a reference value and this check, as already mentioned, is essential for comparing the relative stability of different crystal surfaces, as the E_{surf} convergence depends strongly on the (*hkl*) plane. The definition of E_{surf} by Eq. 1 involves unit cell energies for systems of different dimensionality (2D vs 3D). Depending on the computational code, care should be taken to ensure that these energies are computed with the same accuracy for 2D and 3D cases. To mitigate the problem, the E_{surf} can be computed by alternatives formulas:

$$E_{surf} = \frac{\left[E_{slab}{}^{n} - \frac{n}{2}\left(E_{slab}{}^{n} - E_{slab}{}^{n-1}\right)\right]}{2A} \quad (2)$$

$$E_{surf} = \frac{\left(E_{slab}{}^{n} - E_{bulk}{}^{n}\right)}{2A} \quad (3)$$

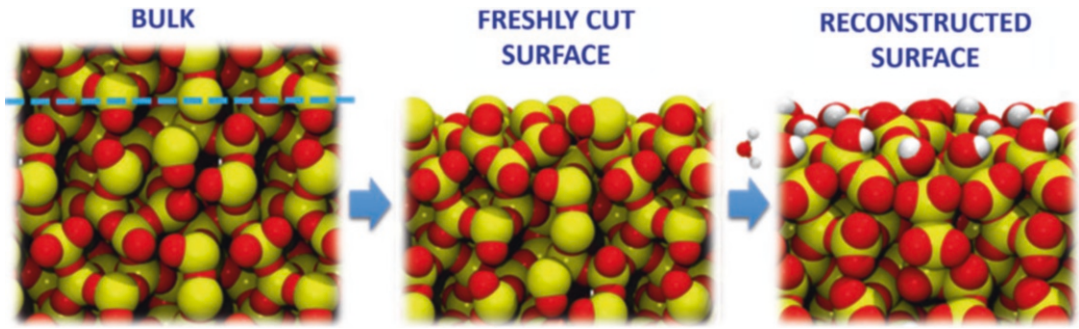
In both equations, the idea is to compare unit cell total energy values computed in the closest computational conditions. For Eq. 2, only energies from slab models are considered, while in Eq. 3 the unit cell of the bulk is built as a super cell of the reference cell to ensure that the crystal vector along the cut has the same module of the slab thickness.

Equations 1, 2, and 3 are only valid for chemical composition of the surface slabs which are an integer multiple n of the bulk unit cell content (stoichiometric surfaces). When dealing with nonstoichiometric surfaces formulas Equations 1, 2, and 3 are no longer applicable. The E_{surf} can be arrived at by a more complicated approach which takes into account the chemical potential of single ions or species needed to keep the system in thermodynamic equilibrium. In this way, E_{surf} will depend on the reference state with respect to which chemical potential values have been chosen.

Surface Relaxation and Reconstruction

According to the above definition, it is possible to list a ranking of stability for various surfaces of the same material, characterized by diverse Miller indices, and compare it to experimental data, if present. The ranking of surface relative stability can give information about the crystal morphology on a purely thermodynamic ground. Kinetic effects frequently dominate the final morphology of a real crystal and should be taken into account by different methods which will not be addressed here.

Nevertheless, the (*hkl*) indices only define the direction perpendicular to which the slab is cut; however, the specific position of the cut plane is somehow dictated by minimizing the number of chemical bonds to be cut at the surface. This process is not straightforward, because as a function of the chemical nature of the material (ionic, metallic, molecular, or covalent) different strategies are needed to cope with the unfilled valencies resulted by the bonds cut. For ionic and metallic surfaces, there is no need to add extra atoms provided that, for ionic crystals, the system is kept electroneutral. For molecular crystals, the cut maintains the molecular integrity of the crystal constituents resulting in a rather rough surface. For these cases, the geometry of the final slab should be fully relaxed by minimizing the forces acting on all atoms. This process is called "relaxation" and is indeed very crucial to compute a correct E_{surf}. Its importance is obvious: the most exposed atoms/ions or molecules will move inwards the slab to maximize the number of interactions with the bulk like atoms. For instance, surface cations in ionic crystals will move inwards to increase the favorable Coulombic attraction with the underneath anions. For molecular crystals, the outmost molecules will relax to establish the missed interactions due to the cut through new H-bonds and dispersive interactions with the nearby molecules. The most difficult case is for covalently bound crystals. In that case, there is no way to avoid unfilled valences at the newly created surface. For a homolitic cut simulating a process occurring in ultrahigh vacuum, the unpaired electrons will pair to reconstruct new chemical bonds. This is the case of the silicon



Surface Modeling of Ceramic Biomaterials, Fig. 3 Example of surface restructuring by water: the case of amorphous silica (silicon *yellow*, oxygen *red*, hydrogen *white*)

(100) surface. This kind of process causes large relaxation energy in the reconstructed surfaces. When covalent crystals are formed in presence of water moisture, or directly from water solution, the surface is chemically reconstructed by the reaction of water molecule and the unsaturated atoms at the surface, resulting in the formation of hydroxyl OH groups. A concrete example of modeling surface reconstruction by water is the case of silica, as for instance in the amorphous phase, as illustrated by Fig. 3. For this structure, defining a surface means cutting the Si-O bonds homolitically, leaving Si and O at the surface which will be converted to SiOH and OH groups by reacting with water molecules. Indeed, the –SiOH groups, silanols, have been detected experimentally at silica surfaces by a variety of spectroscopic techniques [6]. This is an example of surface reconstruction, in which, at variance with the silicon case, the chemical composition is changed with respect to the bulk. Other more extreme examples of surface reconstruction may imply the exchange of ions from surface top to bottom and vice versa, with the aim of stabilizing polar structures of Tasker type 3.

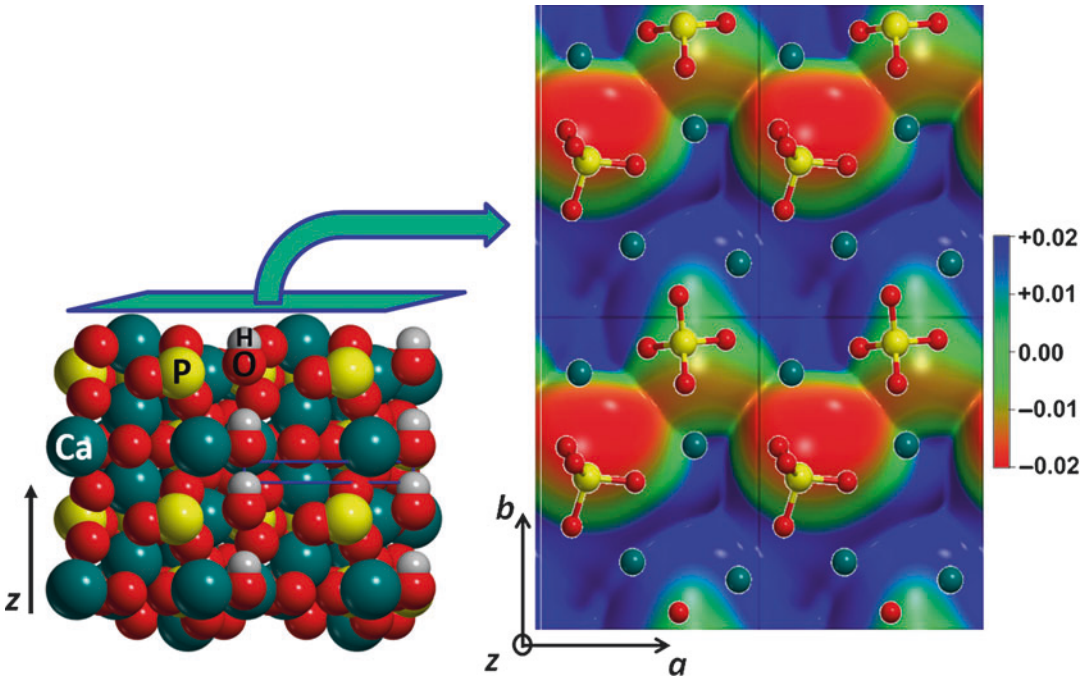
Both relaxation and reconstructions are essential steps in surface characterization. Relaxation can be dramatic if too thin slab models are adopted to simulate a given surface. Reconstruction always implies the need of a careful relaxation due to the change in chemical environment of the newly created functionalities.

Another degree of complexity in surface modeling is the definition of defective surfaces, since real surfaces are neither fully stoichiometric nor are perfectly flat. The first class of defects envisages substitution of chemically related elements (for instance Na by K) or atom vacancies. The second class of defects includes the presence of steps and kinks. Typically, vicinal surfaces with high Miller indices, which are those surfaces cut at a relatively small angle to one of the low index surfaces, are used to define stepped surfaces. The chemical reactivity of the resulting surface models is often studied towards molecules adsorption.

Simulation of Surface Adsorption Processes

Among physicochemical properties of surfaces which can be derived by quantum-mechanics calculations, one of the most informative and useful for further adsorption studies is the electrostatic potential at the surface. Furthermore, by mapping the electrostatic potential on the electron charge density, a clear picture of the distribution and presence of negative, positive, and neutral potential values is obtained. Indeed, as displayed in Fig. 4 for the (001) hydroxyapatite surface (*vide infra*), colored zones are visible on the maps, and the color codes identify negative potential ones as red, positive as blue, and neutral as green. By performing this kind of analysis, an immediate picture of the different electrostatic fields experienced by a molecule approaching the surface is gained.

The simulation of electrostatic properties at surfaces becomes of great importance for the design of proper models of adsorption processes happening on specific faces of the examined



Surface Modeling of Ceramic Biomaterials, Fig. 4 Electrostatic potential mapped on the electronic density for the (001) hydroxyapatite surface. *Left side*: side view of the hydroxyapatite slab (atoms as van der Waals spheres, colors: calcium *cyan*, phosphorous *yellow*, oxygen *red*, hydrogen *light gray*; *blue lines* identify the unit cell borders); *right side*: top view of the electrostatic potential mapped on the electron density isosurface (isovalue: $10^{-5}e$, color scale shown in the picture)

material, considering both physisorption and chemisorption. Indeed, the electrostatic complementarity principle can help to define a reasonable starting adsorption geometry. For example, acidic moieties (COOH, OH, etc.) of adsorbed molecules orient towards electrophilic zones of the surface and vice versa for basic moieties (CO, NH_2, etc.) in adsorbed molecules. The simulation of these processes can be conducted via the cluster model (the adsorbed molecule is part of the cluster system) or the slab model (in both 2D and multislab, though the first approach is more natural).

Usually, in ab initio static calculations, the starting geometry of an adsorbed molecule/surface structure is optimized with the chosen computational code and the optimized total energy is compared to the corresponding values for optimized isolated surface and molecule. So doing, the interaction energy per unit cell associated to the adsorption reaction of one molecule onto a surface can be computed as follows

$$\begin{aligned} \text{Slab} + \text{molecule} &\rightarrow \text{Slab} \\ &\quad - \text{molecule}\ \Delta E_{ads} \\ &= E_{SM} - (E_S + E_M) \end{aligned} \quad (4)$$

for a reference reaction in the gas phase where SM is the slab/molecule adduct, S and M the free slab and molecule, respectively. All three energy values are referred to optimized structures and are negative, implying a negative value for ΔE_{ads} for favorable adsorption processes.

Instead of the interaction energy, in studying adsorption, the binding energy (BE) can be considered, defined so that $BE = -\Delta E_{ads}$. The more negative the interaction energy ΔE_{ads} or the most positive the binding energy BE is, the stronger the interaction between the molecule and the surface. If the molecule interacts with its replica in the nearby unit cell, also the lateral interactions contribute to the adsorption energy and must be included in the ΔE_{ads} estimation.

In the 2D slab model approach with localized Gaussian functions, the basis set is never complete, so that the Basis Set Superposition Error (BSSE) biases the interaction energy. The BSSE derives by summing and subtracting energy values computed with different set of functions in the three cases of bare surface, isolated molecule and adsorbent-adsorbate complex. If not taken into account, the BSSE can heavily affect the ΔE_{ads} (by giving too negative values), leading to the misinterpretation of computed results for the adsorption phenomenon. A convenient way to correct the binding energy is the counterpoise method (CP), in which energy calculations are repeated for the slab and molecule by including the extra basis set functions of each other.

Simulation of Bioceramic Surfaces

In a very general definition, biomaterials are those natural or synthetic materials meant to be in contact and interact with biological systems, and eventually able to substitute, repair, or reconstruct damaged body tissues [7, 8]. Among the numerous classes of biomaterials, bioceramics are considered biocompatible and osteoconductive. Physical and chemical interactions between complex biological systems and inorganic bioceramics must be investigated to improve the knowledge for designing new improved prosthetic materials.

Surface modeling is a useful tool to predict and interpret the experimental measurements, especially in describing, at an atomistic level, the interaction between the ceramic biomaterials and biological molecules.

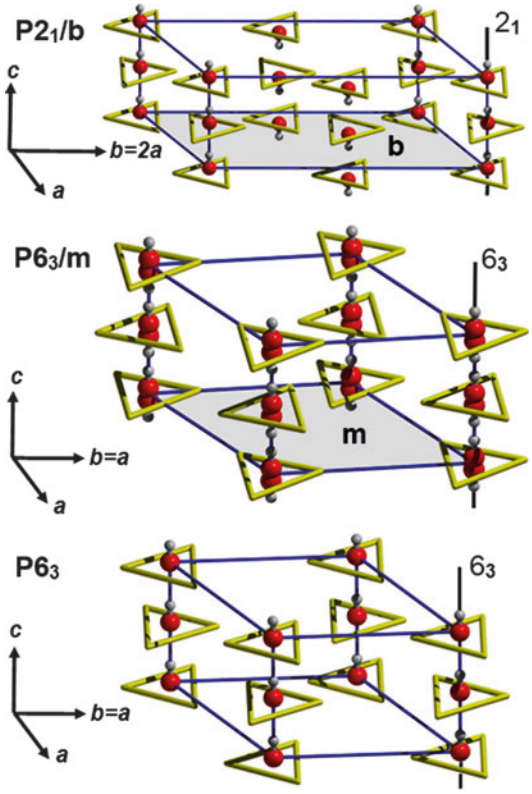
In the following section, an example of surface modeling applied to a common bioceramic-hydroxyapatite, will be reported. All the theoretical concepts and quantities discussed in the previous part will be used here in practical examples.

How to Model Hydroxyapatite Surfaces

Hydroxyapatite [HA, $Ca_{10}(PO_4)_6(OH)_2$] is the main constituent of the inorganic phase in bone and tooth tissues and has been widely studied as a biomaterial, mainly in combination with other materials for compensating its brittleness.

HA is also involved in the bioactivity mechanisms of other ceramic biomaterials, since according to the so-called Hench mechanism the crucial step for the integration of the inorganic prosthetic material within the human living tissues goes through the formation of a layer of carbonated hydroxyapatite. The role of this biomaterial's surface in the complex chemical and biological reactions of molecular and cellular recognition is of paramount importance. For that reason, a large number of experimental and theoretical studies have been performed, and the issue of surface modeling has become essential for the simulation and the understanding of the interaction processes, also when hydroxyapatite nanoparticles are concerned.

As described in the previous theoretical section, a surface is modeled starting from the deep knowledge of the corresponding bulk structure, irrespective of the chosen modeling technique (cluster, semi-infinite crystal, slab, or multislab model). In case of the hydroxyapatite crystal, the bulk structure, as revealed by X-ray and neutron scattering analysis, can be found in nature in two polymorphs, hexagonal and monoclinic. The distinctive feature of HA is the presence of columns or channels of hydroxyl groups (OH), aligned parallel to the crystallographic *c* axis. Indeed, these OH columns are inserted in the pattern where calcium and phosphate ions are disposed in a hexagonal fashion for both crystal forms. In the last 20 years, it has been established that the HA structure suffers of a large scale proton disorder, meaning that the orientation of the OH groups can vary, with the hydrogen atom point upwards or downwards with respect to the reference *c* axis. The difference among the monoclinic and the hexagonal phases is precisely how the OH groups are oriented with respect to the *c* axis orientation. In the monoclinic case, neighbor OH columns exhibit antiparallel orientations, while in the hexagonal one the OH orientation changes within the same column. In order to computationally treat the hexagonal form, which is the one mostly involved in biomineralization processes, the conventional way is to select a specific orientation of the OH



Surface Modeling of Ceramic Biomaterials, Fig. 5 Different HA structures: monoclinic $P2_1/b$ (*top*), experimental hexagonal $P6_3/m$ (*middle*), and theoretical adopted hexagonal $P6_3$ (*bottom*) unit cells. P, O, and Ca atoms not relevant for the scheme were omitted for clarity of representation. Ca ions are at the vertices of triangles around each hydroxyl group. Reproduced from Ref. 9 with permission of the PCCP Owner Societies

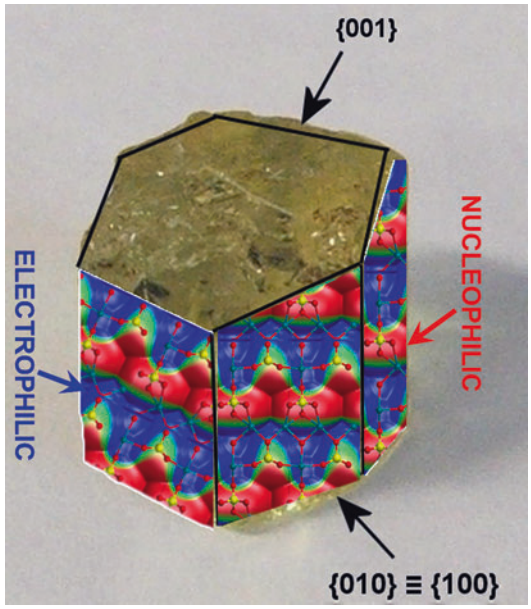
groups, so fixing the hydrogen positions inside the crystal unit cell, as shown in Fig. 5. In this way, the HA crystal structure is a feasible model for theoretically investigating its physicochemical properties, such as geometrical, electronic, mechanical, and vibrational properties, also with the intent to assess the model by comparing calculated values to experimental data.

(001) and (010) HA Surfaces

Once the hydroxyapatite crystal bulk model has been validated, the first step towards surface designing has required the inspection of the crystalline habit, which is informative of the type and extension of specific faces. In the HA hexagonal case, crystallites have needle-like morphology, and the most exposed face is the stoichiometric (010), which is equal to the (100) in the hexagonal system. Furthermore, it is known from the literature that the biomineralization process of bones and tooth enamel involves the (001) surface of hydroxyapatite, which is less exposed but directly interacting with the collagen fibers in bone and enamel growth [9]. In the framework of DFT calculations and taking as a reference the slab model techniques, several computational studies have concerned these two HA surfaces, not only per se but also in interaction with (bio)molecules [9–12].

The (001) plane represents the most natural cut of the bulk structure and does not require any covalent bond breaking. Nevertheless, the aligned OH columns build up a finite dipole moment across the (001) slab. As discussed, this could compromise the surface stability with the increasing thickness, i.e., the surface energy would increase instead of reaching a constant value at convergence with the slab thickness, causing the surface to collapse. The issue is quite complicated, since this slab has a chemical nature intrinsically different form the Tasker type 3 surfaces, where anions and cations were alternating in charged layers (*vide supra*) [5]. The careful inspection of the variation with thickness of a large number of properties, such as dipole moment along the cut direction, band gap, electrostatic potential at surface, surface formation energy, and others, proved only a moderate increase of the dipolar character, without signs of electronic instability. Recent calculations have revealed that for a slab thickness of about 40 nm, that is the real size of nanocrystals found in collagen, the surface is still stable (dipole moment lower than 1 Debye), without any sign of metallization typical of the classical type 3 surfaces, e.g., electronic band gap closure. This (001) HA surface indeed can be seen as an example of ferroelectric surface, whose stability is granted by a delicate balance between the field due to the OH groups and the counter-polarization due to calcium and phosphate ions, the latter free to rotate to minimize the dipole moment across the slab [13].

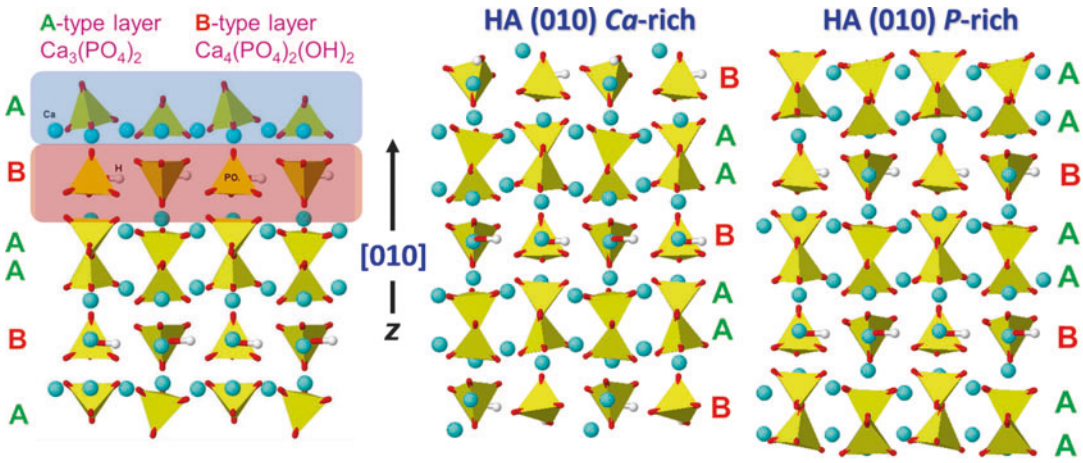
As for the most extended surface, the (010) one, the hydroxyl groups are aligned perpendicularly to the direction of the cut, and the surface is free from the problems affecting the (001) one.



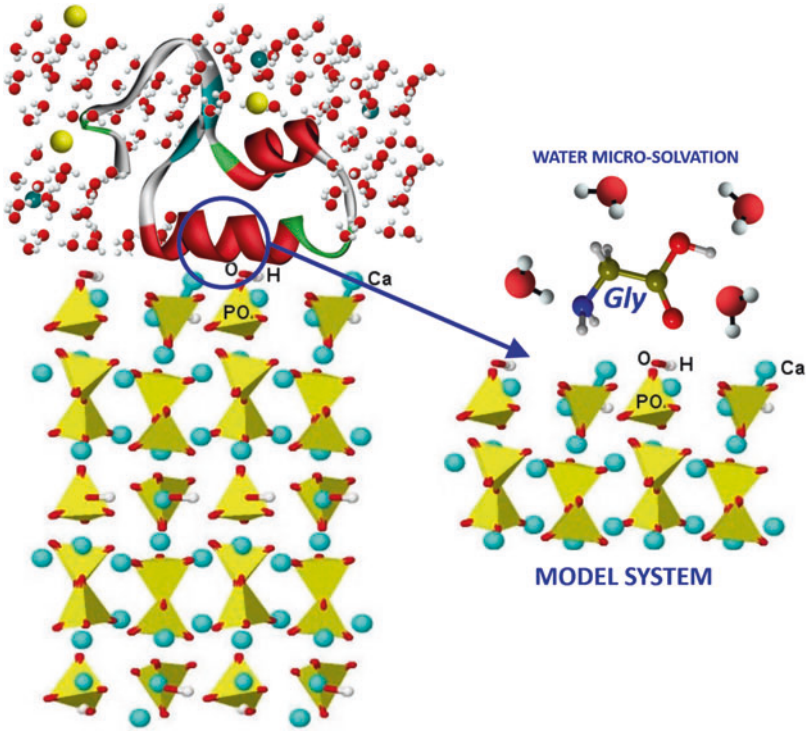
Surface Modeling of Ceramic Biomaterials, Fig. 6 Electrostatic potential mapped on the electronic density for the (010) hydroxyapatite faces, superimposed on a picture of the real HA crystal

Comparing the topology of the two surfaces, different exposed ions are available at surface for interacting with molecules, as it can be seen in Fig. 6 from the electrostatic potential mapped on electron density superimposed to a picture of the HA crystal for sake of clarity. Red zones correspond to negative values of the electrostatic potential and are due to the phosphate anions, while the blue zones are associated to calcium ions, i.e., positive values. By applying an electrostatic complementarity principle, adsorption processes at the two surfaces have been simulated for several molecules to probe specific surface physicochemical features and to improve the knowledge of reaction mechanisms within the body.

For hydroxyapatite, as for other materials, also nonstoichiometric surfaces are relevant, as experimentally shown for the (010) case [14]. Indeed, considering the [010] direction of the cut, two layers can be distinguished based on chemical composition: $Ca_3(PO_4)_2$, called A-layer and $Ca_4(PO_4)_2(OH)_2$, called B-layer. The stoichiometric bulk ratio is A_2B, and in Fig. 7 these layers are displayed, with the indication of the three different terminations for the (010) surface family: (1) stoichiometric surface (Ca/P = 1.67); (2) nonstoichiometric with a ratio Ca to P larger than the stoichiometric one, called Ca-rich surface (Ca/P = 1.71); and (3) nonstoichiometric with a



Surface Modeling of Ceramic Biomaterials, Fig. 7 Graphical representation of: (*left*) the A and B layers alternating along the [010] direction in the HA bulk structure (*middle*) the HA (010) nonstoichiometric Ca-rich surface and (*right*) the HA (010) nonstoichiometric P-rich surface



Surface Modeling of Ceramic Biomaterials, Fig. 8 Simplified approach to the study of solvated protein adsorption on hydroxyapatite surfaces (*left*): only a few water molecules microsolvate the simplest amino acid, glycine, on the top of a hydroxyapatite slab of finite thickness (*right*)

Ca/P ratio lower than the stoichiometric, called P-rich surface (Ca/P = 1.62). As abovementioned, the issue of computing surface energy for nonstoichiometric surfaces is not trivial. The thermodynamic [15, 16] treatment has been applied, by considering as a reservoir the chemical potential value of pure calcium phosphate and calcium hydroxide. Thus, by combining and varying the chemical potential values of the two reference solid phases, a comparison can be obtained among the two nonstoichiometric surfaces in term of surface energy, and, ultimately, their relative stability.

HA Surfaces in Interaction with Biomolecules

In this section, only a brief mention to two examples of adsorption processes on the above described HA surfaces will be reported.

The first example aims at unveiling if water molecules compete with biomolecules (i.e., proteins) for the interaction with hydroxyapatite surfaces. In principle, the study of water solvation should be addressed by means of ab initio molecular dynamics techniques, considering a large number of molecules and requiring extremely heavy computational resources. However, to reduce the complexity of the system and obtain accurate answers, a microsolvation approximation can be followed, that is modeling a few water molecules in interaction with specific functionalities of the selected biomolecule. In detail, the adsorption of the simplest amino acid, glycine (Gly), was simulated on (001), (010) – stoichiometric and nonstoichiometric – hydroxyapatite surfaces at first in the gas phase, to establish the most favorable interactions (see Fig. 8). For the case of (001) surface, water molecules were added on the most stable HA and Gly-HA structures and the results confirmed experimental findings that claimed a direct contact of the amino acid with the hydroxyapatite surface. Calculations were able to simulate the displacement of water by the incoming glycine molecules [17]. Obviously, the mentioned approach is characterized by drastic simplifications, so results must be taken with caution, but constitute a preliminary step towards a more complex design of the system, for instance with a larger number of molecules and a small peptide instead of a single amino acid.

The second example goes in this direction, since the interaction of a small oligopeptide, a polyglycine

of 12 Gly, was simulated on the surface of a monoclinic HA, in the gas phase. The goal of the computational study was to investigate if the interaction of a protein with HA surfaces could change the protein conformation, so influencing the biochemical processes occurring when an inorganic prosthetic material is inserted in the body. Interestingly, the oligopeptide had to be modified by substituting Gly residues in amino acids with a higher electrostatic affinity to the HA surface (lysine and glutamic acid). Without describing all the details, it is worth mentioning, as the most relevant result of this study, the stabilizing effect of HA surface for the α-helix conformation of the mutated oligopeptide, with respect to the folded conformation. In this sense, calculations would compare to a classical example of the statherin protein, which changes its role when in different conformations. Indeed, in the folded state, it prevents calcium phosphate to precipitate, while due to the contact with the HA surface of the teeth enamel, a partial unfolding has been observed [18].

Conclusion

Surface modeling has gained in the last decade great interest due to the increasing computational power, especially in the biomedical field. Understanding the atomistic details of reaction mechanisms at surface of inorganic biomaterials can be fundamental for an improved knowledge and design of the new materials. By means of slab modeling, the physicochemical properties of real 2D layers are computed and can be exploited for subsequent interaction with selected molecules. As a specific example, the simulation of hydroxyapatite surfaces represents a starting point for more complex studies of interaction with proteins towards the simulation of biomineralization processes.

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Computer Modeling and Simulation of Materials
- ▶ Molecular Dynamics Simulations of Nanobiomaterials

References

1. Ertl, G.: Reactions at surfaces: from atoms to complexity (Nobel lecture). Angew Chem Int Ed **47**, 3524–3535 (2008)
2. Sholl, D.S., Steckel, J.A.: Density Functional Theory. Wiley, Hoboken (2009)
3. Jensen, F.: Introduction to Computational Chemistry, 2nd edn. Wiley, Chichester, England (2006)
4. Dovesi, R., Civalleri, B., Orlando, R., Roetti, C., Saunders, V.R.: Ab initio quantum simulation in solid state chemistry. Rev Comp Chem **21**, 1–125 (2005)
5. Tasker, P.W.: The stability of ionic crystal surfaces. J Phys C Solid State Phys **12**, 4977–4984 (1979)
6. Rimola, A., Costa, D., Sodupe, M., Lambert, J.-F., Ugliengo, P.: Silica surface features and their role in the adsorption of biomolecules computational modeling and experiments. Chem Rev **113**, 4216–4313 (2013)
7. Williams, D.F.: Williams Dictionary of Biomaterials. Liverpool University Press, Liverpool (1999)
8. Ratner, B.D., Hoffman, A.F., Schoen, F.J., Lemons, J.E.: Biomaterials science. Elsevier, New York (2013)
9. Corno, M., Rimola, A., Bolis, V., Ugliengo, P.: Hydroxyapatite as a key biomaterial: quantum-mechanical simulation of its surfaces in interaction with biomolecules. Phys Chem Chem Phys **12**, 6309–6329 (2010)
10. Almora-Barrios, N., Austen, K.F., de Leeuw, N.H.: Density functional theory study of the binding of glycine, proline, and hydroxyproline to the hydroxyapatite (0001) and (0110) surfaces. Langmuir **25**, 5018–5025 (2009)
11. De Leeuw, N.H.: Computer simulations of structures and properties of the biomaterial hydroxyapatite. J Mater Chem **20**, 5376 (2010)
12. Almora-Barrios, N., de Leeuw, N.H.: A density functional theory study of the interaction of collagen peptides with hydroxyapatite surfaces. Langmuir **26**, 14535–14542 (2010)
13. Chiatti, F., Corno, M., Ugliengo, P.: Stability of the dipolar (001) surface of hydroxyapatite. J Phys Chem C **116**, 6108–6114 (2012)
14. Ospina, C.A., Terra, J., Ramirez, A.J., Farina, M., Ellis, D.E., Rossi, A.M.: Experimental evidence and structural modeling of nonstoichiometric (010) surfaces coexisting in hydroxyapatite nano-crystals. Colloids Surf B Biointerfaces **89**, 15–22 (2012)
15. Astala, R., Stott, M.: First-principles study of hydroxyapatite surfaces and water adsorption. Phys Rev B **78**, 075427 (2008)

16. Slepko, A., Demkov, A.: First principles study of hydroxyapatite surface. J Chem Phys **139**, 044714 (2013)
17. Rimola, A., Corno, M., Zicovich-Wilson, C.M., Ugliengo, P.: Ab initio modeling of protein/biomaterial i: competitive adsorption between glycine and water onto hydroxyapatite surfaces. Phys Chem Chem Phys **11**, 9005–9007 (2009)
18. Rimola, A., Aschi, M., Orlando, R., Ugliengo, P.: Does adsorption at hydroxyapatite surfaces induce peptide folding? Insights from large-scale B3LYP calculations. J Am Chem Soc **134**, 10899–10910 (2012)

Surface Patterning

▶ Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Surface Plasmon Enhanced Optical Bistability and Optical Switching

Weiqiang Mu and John B. Ketterson
Department of Physics and Astronomy, Northwestern University, Evanston, IL, USA

Definition

A surface plasmon is a collective oscillation involving the coupled motion of the electrons in a metal and the associated electromagnetic field in small structures and at interfaces. In small structures (of order a wavelength and smaller), the oscillations are localized, whereas at a planar interface separating a metal and a dielectric a propagating mode exists that decays exponentially on both sides of the interface. Plasmons can be excited by fast electrons or, in the right geometry, an incoming electromagnetic wave. Here, optical excitation is considered. An optically bistable system can have two different but stable outputs for the same input signal over some range. By changing the input beyond some threshold, the output of an optically bistable system can be switched between the two stable states, a feature that can be used to switch the output between two channels.

Optical Bistability and Optical Switching

The development of modern optical communications and optical signal processing requires ultrafast switching of optically encoded information. All-optical switching, which uses light to control light, is a very promising candidate for obtaining high operating speeds. One strategy to achieve all-optical switching is to exploit optical bistability. This phenomenon was first observed in sodium vapor in the 1970s [1]. Soon after that, many materials, gas, liquid, and solid, were shown to produce optical bistability [2–4].

To make a system bistable, both nonlinearity and feedback are required. For an optical system having these two properties, the response function, F, of the system can be defined as

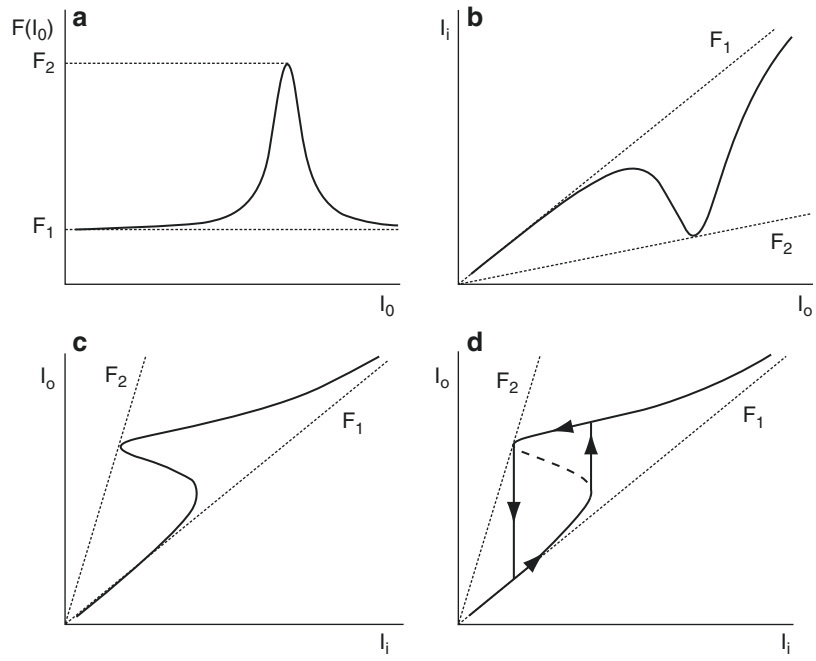
$$F(I_i, I_o) = \frac{I_o}{I_i} \tag{1}$$

where I_i and I_o are respectively the input and output signal. The response function can have various forms, depending on system parameters, which in turn affect the relation between the input and output.

As a simple but important example, consider the effect of the bell-shaped response function shown in Fig. 1a relative to the output signal. Then I_i and I_o are related as

$$I_i = \frac{I_o}{F(I_o)}. \tag{2}$$

Since $F(I_o)$ has a single valued bell-shaped behavior, it is a simple matter to plot I_i using I_o as the independent variable, as shown in Fig. 1b. The corresponding behavior of I_o as a function of I_i is then shown in Fig. 1c. The system is clearly nonlinear, but more importantly for a special range of the inputs the output has three possible



Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 1 (**a**) Typical response curve of an optically bistable system; (**b**) the corresponding I_i relative to I_o behavior; (**c**) the corresponding I_o relative to I_i behavior. (**d**) When I_i continuously increases or decreases, I_o shows a hysteresis; the *dashed line* denotes the unstable region

solutions, as can be seen from Fig. 1c. Those involving a negative slope are unstable and hence the system can be bistable in the region between the two points where the derivative diverges. When continuously varying the input, the output will select one of the stable branches, depending on its history; that is, a kind of hysteresis is achieved as shown in Fig. 1d. As noted, the dashed lines correspond to a region of unstable solutions. The sudden changes and the history dependence of the output can be utilized to make optical switches.

There are many ways to introduce feedback in an optical system. As examples, bistability has been demonstrated in devices based on Fabry–Perot etalons [5], microdisk resonators [6], photonic-crystal cavities [7], etc. The nonlinearities can arise if either refractive index or absorption is intensity dependent. Devices based on the former are said to be dispersive, while the latter are called dissipative. The intensity dependent nonlinearities, which arise from electronic contributions to the third order susceptibility, $\chi^{(3)}$, such as the Kerr effect, can respond in a picosecond or less. While such response times are fast enough for presently envisioned applications, the magnitude of these nonlinearities is very small; very high laser powers are then required to generate the desired optical bistability.

Surface Plasmons

The Dielectric Constant of Metals

When free electrons near a metal surface are electromagnetically excited, the charge fluctuations are typically confined within a Thomas-Fermi screening length of the metal surface (~1 Å). The metals used in plasmonic devices are typically gold, silver, copper, or aluminum. It is common to distinguish two types of surface plasmons: localized surface plasmons (LSP) and propagating surface plasmons (PSP); this primarily depends on the geometry and the distinction is not always sharp. The main difference is that the LSP does not propagate (although, as in the case of a strip resonator, the mode can be considered as counter propagating waves). On the other hand, the PSP transports energy present within the interface. Both originate from the special dielectric properties of metals: a negative real part and a relatively small imaginary part. In the very simplest model

the dielectric constant, ε_m, can be described by the Drude-Lorentz form:

$$\varepsilon(\omega) = 1 - \frac{{\omega_p}^2}{\omega^2 + i\omega\gamma} \tag{3}$$

here $\omega_p = 4\pi ne^2/m$ is the electron plasma frequency, with n and m being the electron density and mass respectively, while γ is the collision rate that governs the damping. The dielectric response is clearly strongly frequency dependent. In visible or near infrared, γ is generally much smaller than ω. In this limit, the real and imaginary parts of ε_m can be approximated as

$$\begin{aligned} \varepsilon_r(\omega) &= 1 - \frac{\omega_p^2}{\omega^2} \\ \varepsilon_i(\omega) &= \frac{\gamma\omega_p^2}{\omega^2} \end{aligned} \tag{4}$$

From Eq. 4, it is clear that when $\omega < \omega_p$, the real part, ε_r, is negative.

Equation 3 neglects the host dielectric response, which can be incorporated by replacing the one on the right hand side by a parameter ε_∞; more complicated forms, involving the addition of Lorentzian-like poles, have been used which give a much better overall representation of the measured optical response [8]. Most metals have a plasma frequency in the ultraviolet or visible. The real part of $\varepsilon(\omega)$ for the noble metals gold and silver is negative in the visible and infrared while the imaginary part is small. In addition, they are very stable chemically which makes them the best candidates for plasmonics. Besides metals, some doped semiconductors are good plasmonic materials in the infrared region.

Localized Surface Plasmons

For LSP, the simplest case is a spherical metal particle excited by an optical field $\mathbf{E}_0(\omega)$, where the sphere radius, a, is much smaller than the optical wavelength, λ. Using a simple electrostatic model the induced dipole moment is given by

$$\vec{p}\ (\omega) \propto \frac{\varepsilon_m - \varepsilon_d}{\varepsilon_m - 2\varepsilon_d}\vec{E}_0(\omega) \tag{5}$$

Here ε_d is the dielectric constant of the medium. For the model based on Eq. 2, it is easy to see that with $\varepsilon_d = 1$, the real part of the dominator in Eq. 5 will vanish for $\omega = \omega_p/\sqrt{3}$. Similar behavior happens with more complex models. Because of the relative small value of ε_i for Ag and Au, the induced dipole moment can be very large. The scattering from such metal particles is then strong. Clearly, the position of the resonance is determined by the real part of ε_m, but the magnitude and bandwidth are determined by the imaginary part of ε_m.

Propagating Surface Plasmons

As a surface wave, the propagating surface plasmon can only exist as a transverse-magnetic (TM) wave, as shown the inset of Fig. 2. The electric fields in the two media have the form [9]:

$$\begin{aligned} \vec{E}_d &= (E_{dx}, 0, E_{dx}) \cdot \exp\left[i \cdot \left(k_{sp}x - \omega t\right) - \gamma_d z\right] (z > 0) \\ \vec{E}_m &= (E_{mx}, 0, E_{mx}) \cdot \exp\left[i \cdot \left(k_{sp}x - \omega t\right) - \gamma_m z\right] (z > 0) \end{aligned} \tag{6}$$

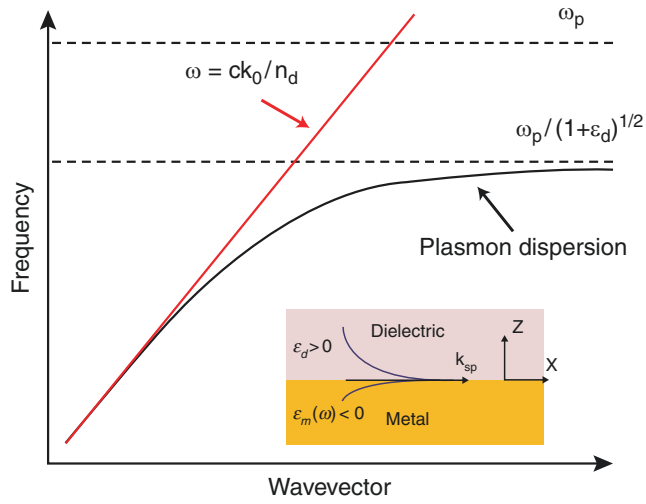
where k_{sp} is the surface plasmon wavevector, and γ_d and γ_m satisfy:

$$\gamma_d = \sqrt{k_{sp}^2 - \varepsilon_d k_0^2};\ \gamma_m = \sqrt{k_{sp}^2 - \varepsilon_m k_0^2} \tag{7}$$

where $k_0 = \omega/c$ is the free space wavevector. In order that the EM wave decays in both sides of the boundary (a requirement for a surface wave), the real part of γ_m and γ_d must be positive. From the conditions that $E_{//}$ and $D_\perp$ be continuous, yielding $E_{dx} = E_{mx}$, and $\varepsilon_d E_{dz} = \varepsilon_m E_{mz}$, the surface plasmon dispersion relation can be derived:

$$k_{sp}(\omega) = k_0\sqrt{\frac{\varepsilon(\omega) \cdot \varepsilon_d}{\varepsilon(\omega) + \varepsilon_d}} \tag{8}$$

The necessary condition for a propagating surface plasmon is that the real part of the quotient under the square root sign in Eq. 8 be positive. This imposes a restriction that $\varepsilon_r(\omega) < -\varepsilon_d$. The real part of k_{sp}, is the propagation wavevector of the



Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 2 Dispersion relation of propagating surface plasmons based on Eqs. 3 and 8 with $\varepsilon_d = 1$. The *red line* is the light dispersion in the dielectric. The plasmon dispersion asymptotically approaches the frequency $\omega_p/(1+\varepsilon_d)^{1/2}$. The *inset* is the schematic representation of surface plasmon at the metal-dielectric interface. The wave is TM wave, and the fields decay exponentially in both sides of the interface

surface plasmon. The characteristic plasmon propagation distance is the inverse of the imaginary part of k_{sp}. To achieve long propagation distances, it is necessary to have small ε_i or larger $|\varepsilon_r|$. In this entry, silver is selected as the plasmonic material. Figure 2 shows the dispersion relation for the surface plasmons propagating at the interface between silver and fused silica (n = 1.4607 [10]). By fitting the experimental values of the dielectric constant of silver to polynomials, the dielectric constant of silver in visible region can be well represented using the following equation [11]:

$$\begin{aligned}\varepsilon_r &= -255.32 + 198.63\omega - 60.79\omega^2 + 8.38\omega^3 - 0.43\omega^4 \\ \varepsilon_i &= 83.26 - 132.79\omega + 90.47\omega^3 + 6.66\omega^4 \\ &\quad - 0.71\omega^5 + 0.03\omega^6\end{aligned} \tag{9}$$

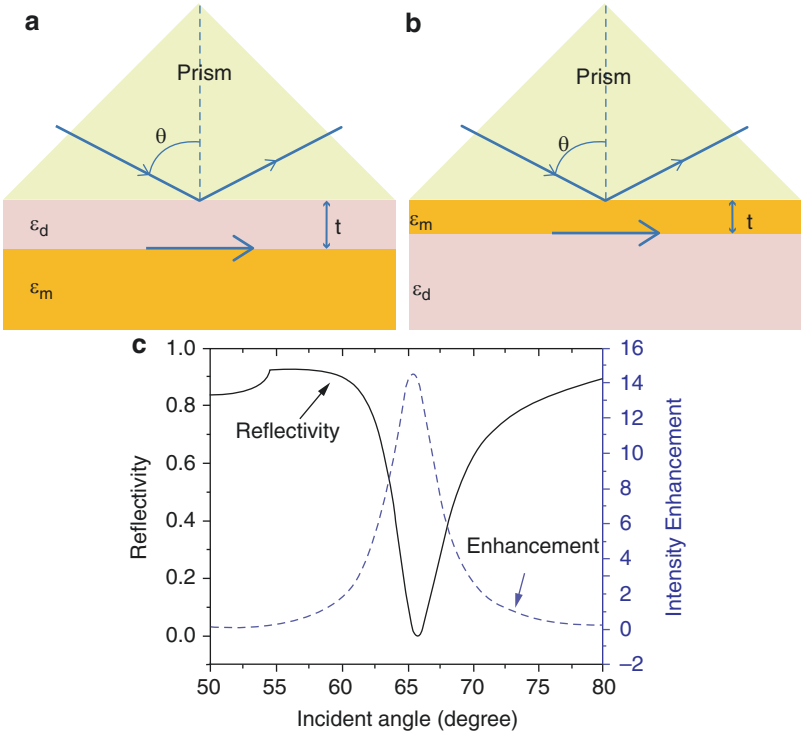
where $\omega = 2\pi \cdot c/\lambda \cdot 10^{-15} S^{-1}$.

Because the PSP wavevector is larger than that in the dielectric, the mode cannot be directly excited from the dielectric. Two popular methods to excite PSP are grating coupling and prism coupling. The grating coupling [12–14] utilizes a diffraction grating to provide the extra contribution to the wavevector of the photons propagating in the dielectric side to match the real part of k_{sp} (this is called an umklapp process in solid state physics). The grating coupling occurs when

$$k_{//} = k_{sp}^{r} + mK_g \, m = 0, \pm 1, \pm 2 \ldots \tag{10}$$

where $k_{//}$ is the in-plane wavevector of the incoming wave, k_{sp}^{r} is the real part of the surface plasmon wavevector, and K_g is the grating wavevector. The coupling efficiency depends on the parameters of the grating such as the ditch depth and ditch shape.

The two common prism coupling schemes, the Otto [15] and the Kretschmann [16] geometries (named after their originators), are shown in Fig. 3a, b. For both configurations, the coupling prisms must have a higher refractive index than the dielectric at the metal interface. By correctly choosing the incident beam parameters, such as the incident angle or the wavelength, the in-plane wavevector $k_{//}$ of TM mode incoming beams can be matched to k_{sp}^{r} for the surface plasmon between metal and the dielectric. If the coupling layer between the prism surface and the plasmon



Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 3 (**a**) The Otto geometry and (**b**) the Kretschmann configuration for coupling to a propagating surface plasmon mode with a TM optical wave by total internal reflection. (**c**) The reflectivity and the intensity enhancement of the surface plasmon with Kretschmann configuration

interface is not too thick, the evanescent field from the prism surface will excite the metal surface. In the Otto configuration, the coupling from the prism to the metal film is through a layer with dielectric constant smaller than that of the prism, as in Fig. 3a. In the Kretschmann configuration, the coupling occurs through the metal film itself.

When light is coupled into the PSP with prism coupling, the reflectivity, which occurs in a regime that would otherwise correspond to total internal reflection (TIR), will have a dip; this is called *attenuated total internal reflection* (ATR). This results from part of the energy of the incoming beam has been coupled into the PSP wave, and partially from dissipation by the ohmic losses in the metal. For both methods, the coupling efficiency, or the dip depth, depends on the thickness of the coupling layer. If the coupling layer is too thick, only a small portion of the incoming beam is coupled into surface plasmon; this is referred to as undercoupling. On the other hand, if the coupling layer is too thin, a large portion of the excited surface plasmon will be coupled back into the reflection beam and the buildup of the wave is limited; this is called overcoupling. To get the optimum coupling, which occurs where the reflected beam intensity vanishes and all the energy goes into the mode, the coupling layer has to have just the right thickness.

The reflectivity and the intensity enhancement associated with the PSP can be modeled with the multilayer Fresnel reflection method. The intensity enhancement is defined as the ratio of the intensity at the metal-dielectric boundary to the incoming intensity in the prism. Figure 3c shows simulation results with the Kretschmann configuration. In the simulation, the light wavelength was fixed at 532 nm, corresponding to the second harmonic frequency of a Nd:YAG laser, and the incident angle is scanned from 50° to 80°. The refractive indices of prism (SF11) and dielectric material (fused silica) are 1.7948 and 1.4607 respectively. The silver thickness has been optimized as 44.5 nm to obtain the best coupling.

Surface Plasmon–Enhanced Optical Bistability

Surface plasmons can be exploited to significantly enhance optical bistability. First, as shown in the

simulation results of Fig. 3c, the local optical field at the interface is greatly enhanced. This reduces the needed laser driving power. Secondly, the high sensitivity of surface plasmons to the change in the refractive index of the adjacent dielectric makes for a fast and sensitive switch. Surface plasmons (in many geometries) can then be utilized to produce optical bistability. In what follows, several approaches are described to achieve plasmon-based bistability.

The first approach that exploits plasmon induced optical bistability utilizes the Kretschmann configuration [17, 18]. If the dielectric material adjacent to the silver surface supporting the plasmons is replaced by a nonlinear Kerr medium, then the refractive index will depend on the local optical intensity. The refractive index n_k of Kerr material is written as (When the surface plasmon is excited, the actual optical intensity in the Kerr medium is exponentially decaying. In the simulations presented here, the value of n_k is taken to be that calculated with the intensity *at the interface* as a simplified model and neglect its change in the Kerr medium, as in Ref. [17]. In Ref. [18], it is shown that if the intensity distribution is considered, the bistability shape will be same, but the required optical intensity will be higher than the value predicted by this simple model)

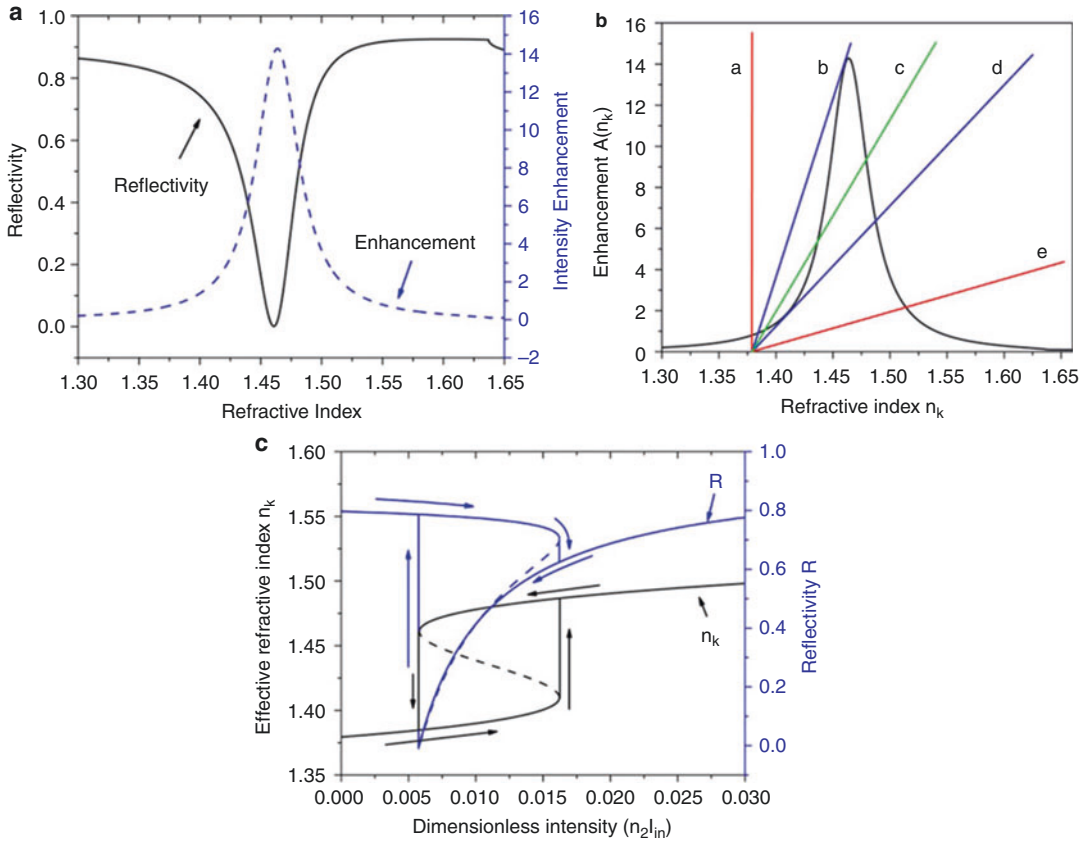
$$n_k = n_0 + n_2 \cdot I = n_0 + n_2 \cdot A \cdot I_{in}; \qquad (11)$$

here, n_0 and n_2 are the linear and second order nonlinear refractive index of the Kerr material, I the light intensity in the Kerr medium, I_{in} the input optical intensity, and A is the intensity enhancement due to the surface plasmon wave. The Kerr medium examined here is a mixture of nitrobenzene and ethanol. The values of n_0 and n_2 for nitrobenzene are 1.553 and 2.766×10^{-20} $(m/V)^2$ respectively [19]; the refractive index of ethanol is 1.362. By mixing the two liquids in the right ratio, the linear refractive index n_2 can be adjusted to lie between 1.362 and 1.553. The effective n_2 will be assumed to be proportional to the volume percentage of the nitrobenzene.

Using the same geometry as for the simulation of Fig. 3c but with the incident angle fixed at 65.81 ° (which corresponds to the best coupling in Fig. 3c), and scanning the refractive index of the dielectric medium from 1.3 to 1.65, the corresponding intensity enhancement factor A (n_k) is plotted as the dashed line in Fig. 4a. With the calculated A(n_k), the values of n_k based on Eq. 11 will exhibit bistability relative to the incident intensity, I_{in}. This is graphically depicted in Fig. 4b [20]. Equation 11 is first rewritten as:

$$A(n_k) = \frac{n_k - n_0}{n_2 I_{in}} = \frac{n_k - n_0}{I_d} \qquad (12)$$

here I_d is defined as the dimensionless input intensity. The left side of Eq. 12 A(n_k) is plotted as the black curve in Fig. 4b. With fixed value of n_0, the right side of Eq. 12 is straight line, the slope of which is proportional to the reciprocal of I_{in} (or I_d). The values of n_k are then determined by the intersections of the black curve and the straight line. For the simulation shown in Fig. 4b, the value of n_0 is selected to be 1.38. On the other hand, the resonance, which is the peak position of curve A(n_k), occurs at $n_k = 1.46$. The difference between these two values corresponds to three times the Δn associated with the full width at half maximum (FWHM) of the resonance. Note all of the lines, resulting from different I_{in}, intersect at $n_k = n_0$, as seen in Fig. 4b. As I_{in} increases from zero, the slope of the lines continuously decreases, from $a \rightarrow b \rightarrow c \rightarrow d \rightarrow e \rightarrow$. For the regions with very low and very high input intensity (the region between lines a and b and the region beyond line d), there is only one solution for n_k. Line b and line d correspond to critical points, where there are two solutions to the equation. Between line b and line d, there are three possible values of n_k for the same input, line c being an example. When there are multiple solutions, the physical system selects one, depending on its history. Figure 4c shows the effective index of refraction (n_k) and the reflectivity (R) as a function of the dimensionless input intensity; note the bistability in both n_k and R is clearly presented. The dashed parts of the line are the unstable solutions.



Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 4 (**a**) The reflectivity and intensity enhancement relative to the refractive index of the dielectric in the Kretschmann configuration. (**b**) Graphic solutions for n_k with different optical input intensities. The *black curve* is the intensity enhancement $A(n_k)$ relative to n_k; the different straight lines represent different input intensities; the intersections between the *straight lines* and the *enhancement curve* give the solution for n_k. (**c**) Solutions for n_k and the corresponding values of reflectivity for different input intensities. By continuously changing the input intensity, both n_k and the reflectivity exhibit bistability

Long-Range Surface Plasmon–Based Optical Bistability

For most of the nonlinear materials, the Kerr coefficient n_2 in Eq. 11 is very small. As in Fig. 4c, to achieve the bistability, the required dimensionless intensity needs to be at least 0.017, corresponding to an input optical intensity on the order of 10^{12} W/cm^2. Therefore, to have a strong optical bistability, it is better to have a sharper resonance, and with it a higher enhancement of the local field. Long-range surface plasmon polaritons (LRSPP), a mode that exists under special conditions, possesses these two characteristics. Of course, a narrower resonance comes at the cost of a longer response time.

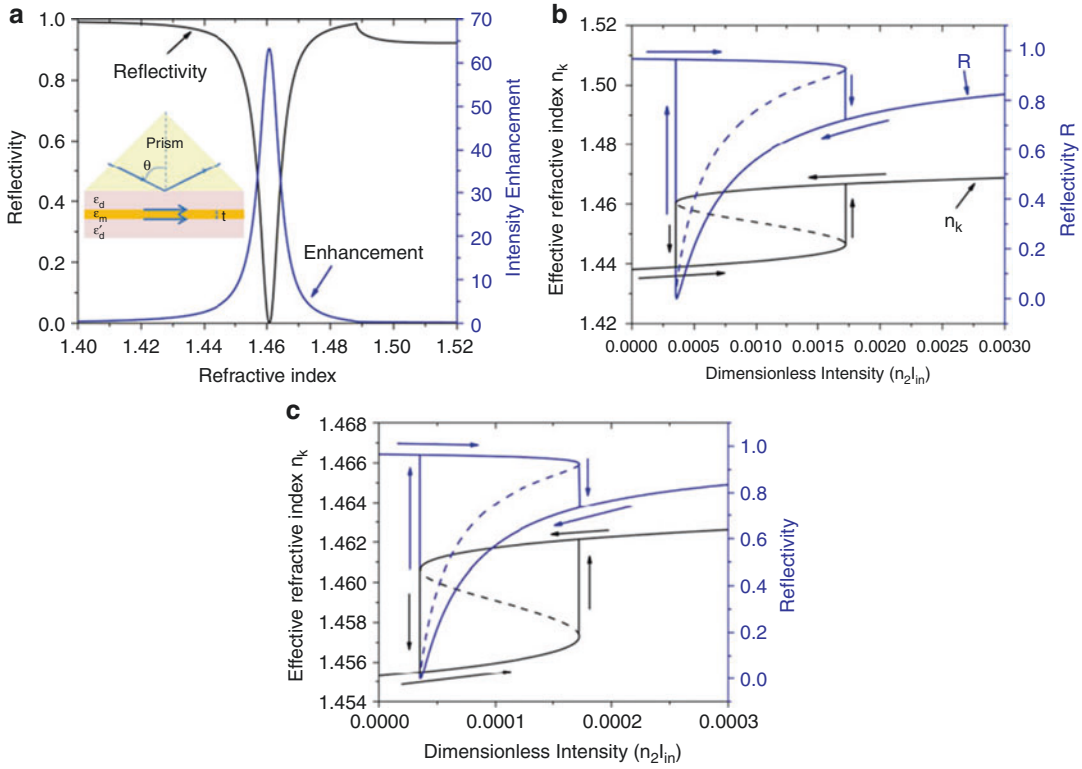
If a thin metal film is sandwiched between dielectric materials having the same optical properties, surface plasmon waves will exist on *both* sides of the metal film. When the metal film is thin enough, the two surface plasmons will interact with each other and split into two modes: a symmetric mode and an antisymmetric mode relative to the in-plane electric field at the film center. The electric fields on the two surfaces are in phase for symmetric mode and 180 ° out of phase for the antisymmetric mode. In a manner similar to that used to obtain the dispersion relation for SPP, and

after imposing the boundary conditions, the wavevectors for these two surface plasmon modes are as [21]:

$$\tanh\left(\frac{\gamma_m t}{2}\right) = \frac{\varepsilon_m \gamma_d}{\varepsilon_d \gamma_m} (\text{anti} - \text{symmetric mode})$$
$$\tanh\left(\frac{\gamma_m t}{2}\right) = \frac{\varepsilon_d \gamma_m}{\varepsilon_m \gamma_d} (\text{symmetric mode}) \quad (13)$$

where t is the thickness of the metal layer, and $\gamma_i = \sqrt{k_{sp}^2 - \varepsilon_i k_{0i}^2}$. For the symmetric mode, the optical fields constructively interfere inside the metal, which, in turn, results in an increase in the ohmic damping. Therefore, the plasmon lifetime (and with it the decay length) is relatively small; hence it is called a short-range surface plasmon polariton (SRSPP). For the antisymmetric mode, the electric fields from the two surfaces destructively interfere inside the metal, thereby decreasing the ohmic losses. The plasmon lifetime (and the decay length) then become much longer; this mode is called the long-range surface plasmon polariton (LRSPP). As in Eq. 13, the wavevectors of these surface plasmons depend on the metal thickness t. In the thin film limit, the damping of LRSPP is approximately proportional to t^2. Therefore, in principle, by continuously decreasing the metal thickness, we should have very small losses and hence very large local optical intensities.

In order to be able to compare with the optical bistability induced by conventional PSP shown in Fig. 4, the same calculations were repeated for the LRSPPs. The inset of Fig. 5a shows the geometry of the excitation of LRSPP using the ATR coupling method. When the silver thickness is 20 nm,



Surface Plasmon Enhanced Optical Bistability and Optical Switching, Fig. 5 (**a**) The reflectivity and intensity enhancement of the LRSPP when it is excited via the ATR method. The *inset* shows the geometry used to excite the LRSPP. (**b**) The bistability of refractive index of the Kerr material and the reflectivity relative to the dimensionless intensity for the case where the silver film thickness is 20 nm. (**c**) The bistability of refractive index of the Kerr material and the reflectivity relative to the dimensionless intensity when silver film thickness is 10 nm

perfect coupling (the absence of reflection) happens when the coupling layer, here a fused silica layer between the silver and the prism, is 555 nm, and the incident angle is 56.02°. Then the last layer of the dielectric is replaced with a material for which the refractive index ranges between 1.40 and 1.52. The reflectivity and the intensity enhancement are the black and blue curves respectively. Clearly, compared with Fig. 4a, the peak intensity enhancement resulting from the LRSPP is much larger, while the width of the bell-shape enhancement is much narrower. To directly compare the enhancement between the LRSPP and conventional PSP, the difference between n_0 and the resonance of intensity enhancement curve $A(n_k)$ in Fig. 5b is still set to be three time the Δn associated with the full width at half maximum (FWHM) of the resonance of $A(n_k)$. Figure 5b shows the resulting bistability of the refractive index n_k for the Kerr medium and the total reflectivity. Clearly, the peak laser intensity needed to produce the optical bistability here is much smaller than the value needed in Fig. 4c, decreasing by one order of magnitude (from 0.017 to 0.0015). By further decreasing the silver film thickness, the needed peak intensity to produce bistability will decreases even more; this is demonstrated in Fig. 5c for the case where the silver film is 10 nm. Here, the optimized coupling layer thickness is 965 nm and coupling angle 54.87°.

Conclusions

The resonant behavior exhibited by conventional propagating surface plasmon polaritons and especially the long-range propagating plasmon polaritons can be exploited to enhance the optical intensity in an adjacent dielectric media. When combined with a nonlinear Kerr medium, one can produce optical bistability and with it the ability to form an optical switch. The effect can be manifested as a change in the reflectivity or a phase delay from plasmons excited in an ATR geometry [22], a change in the scattering intensity from localized surface plasmons [23], a shift in the transmission or reflection coefficient in a grating-coupled surface plasmon geometry [24], and so on. To lower the pumping threshold for bistability, geometries leading to a sharper resonance, and with it a larger intensity enhancement, must be exploited.

Cross-References

- ▶ Active Plasmonic Devices
- ▶ Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures
- ▶ Surface Plasmon-Polariton-Based Detectors

References

1. Gibbs, H.M., McCall, S.L., Venkatesan, T.N.C.: Differential gain and bistability using a Sodium-filled Fabry–Perot interferometer. Phys. Rev. Lett. **36**, 1135–1138 (1976)
2. Abraham, E., Smith, S.D.: Optical bistability and related devices. Rep. Prog. Phys. **45**, 815–885 (1982)
3. Gibbs, H.M.: Optical Bistability: Controlling Light with Light. Academic, Orlando (1985)
4. Gibbs, H.M., McCall, S.L., Venkatesan, T.N., Gossard, A.C., Passner, A., Wiegmann, W.: Optical bistability in semiconductors. Appl. Phys. Lett. **36**, 451–453 (1979)
5. Felber, F.S., Marburger, J.H.: Theory of nonresonant multistable optical devices. Appl. Phys. Lett. **28**, 731–733 (1976)
6. Borselli, M., Johnson, T.J., Painter, O.: Beyond the Rayleigh scattering limit in high-Q silicon microdisks, theory and experiment. Opt. Express **13**, 1515–1530 (2005)
7. Notomi, M., Shinya, A., Mitsugi, S., Kira, G., Kuramochi, E., Tanabe, T.: Optical bistable switching action of Si high-Q photonic-crystal nanocavities. Opt. Express **13**, 2678–2687 (2005)
8. Sukharev, M., Sievert, P.R., Seideman, T., Ketterson, J.B.: Perfect coupling of light to surface plasmons with ultra-narrow linewidths. J. Chem. Phys. **131**, 034708 (2009)
9. Rather, H.: Surface Plasmons on Smooth and Rough Surfaces and on Gratings. Springer, Berlin/New York (1988)
10. Palik, E.D.: Handbook of Optical Constants of Solids. Academic, New York (1985)
11. Chen, Z., Hooper, I.R., Sambles, J.R.: Srongly coupled surface plasmons on thin shallow metallic gratings. Phys. Rev. B **77**, 161405 (2008)
12. Mu, W., Buchholz, D.B., Sukharev, M., Jang, J.I., Chang, R.P.H., Ketterson, J.B.: One-dimensional long-range plasmonic-photonic structures. Opt. Lett. **35**, 550–552 (2010)

13. Chen, Y.J., Koteles, E.S., Seymour, R.J., Sonek, G.J., Ballantyne, J.M.: Solid State Commun. **46**, 95 (1983)
14. Gruhlke, R.W., Holland, W.R., Hall, D.G.: Optical emission from coupled surface plasmons. Opt. Lett. **12**, 364–366 (1987)
15. Otto, A.: Excitation of nonradiative surface plasma waves in silver by method of frustrated total reflection. Z. Phys. **216**, 398–410 (1968)
16. Kretschmann, E.: Z. Phys. **241**, 313 (1971)
17. Wysin, G.M., Simon, H.J., Deck, R.T.: Optical bistability with surface plasmon. Opt. Lett. **6**, 30–32 (1981)
18. Hickernell, R.K., Sarid, D.: Optical bistability using prism-coupled, long-range surface plasmons. J. Opt. Soc. Am. B **3**, 1059–1069 (1986)
19. Boyd, R.W.: Nonlinear Optics. Academic, Boston (2003)
20. Saleh, B.E.A., Teich, M.C.: Fundamentals of Photonics. Wiley, New York (1991)
21. Sarid, D.: Long- range surface-plasma waves on very thin metal films. Phys. Rev. Lett. **47**, 1927–1931 (1981)
22. Nazvanov, V.F., Kovalenko, D.I.: Phase optical bistability in structures with surface plasmons. Tech. Phys. Lett. **24**, 650–651 (1998)
23. Leung, K.M.: Optical bistability in the scattering and absorption of light from nonlinear microparticles. Phys. Rev. A **33**, 2461–2464 (1986)
24. Wurtz, G.A., Pollard, R., Zayats, A.V.: Optical bistability in nonlinear surface-plasmon polaritonic crystals. Phys. Rev. Lett. **97**, 057402 (2006)

Surface Plasmon Nanophotonics

► Optical Properties of Metal Nanoparticles

Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices

Jing Feng and Hong-Bo Sun
State Key Laboratory on Integrated Optoelectronics, College of Electronic Science and Engineering, Jilin University, Changchun, China

Introduction

In organic optoelectronic devices, such as organic light-emitting devices (OLEDs) and organic solar cells (OSCs), high efficiency is one of the key issues for their applications, and much effort has been devoted to developing novel materials and device structures [1, 2]. It is well known that the majority of the generated light is trapped in OLEDs. Around 80 % of internally generated light is trapped in the form of waveguide (WG) modes in organic and indium tin oxide (ITO) anode layers and in surface plasmon polariton (SPP) modes associated with the metallic electrode/organic interface in OLEDs [3, 4]. In organic solar cells (OSCs), the active layer is generally less than 100 nm due to the short exciton diffusion length, which limits the efficiency of incident light absorption. A thicker active layer offers higher light absorption; however, it comes at the expense of lowered exciton harvesting [5, 6]. So far, there is still the greatest scope for significant improvements about efficiency in OLEDs and OSCs.

Microstructure with wavelength- to subwavelength-scale periodicity has played an important role in optical and optoelectronic devices, particularly in optical fibers, distributed feedback lasers, LEDs, and solar cells, through manipulating the generation and propagation of photons in materials. Introducing a microstructure onto the metallic electrode is especially crucial for recovering the power lost to the associated SPP modes in OLEDs by providing an additional momentum to couple the SPP modes [3, 7]. Plasmonic nanostructures have been introduced into solar cells for highly efficient light harvesting. Two types of plasmonic resonances, surface plasmonic resonances (SPRs) [8, 9] and localized plasmonic resonances (LPRs) [10, 11], can be used for enhancing light absorption in an organic solar cell without increasing the thickness of its active layers. This entry concerns recent progresses in fabrication of periodic microstructure in organic optoelectronic devices and their effects on improving the device performance. The SPP modes that were generally lost in conventional OLEDs have been successfully recovered by employing the periodic corrugations, and a much enhanced light extraction has been observed. The introduction of the microstructures into OSCs is effective in relieving the mismatch

between optical absorption length and charge transport scale, and enhanced light absorption in thinner organic films has been achieved.

OLEDs Integrated with Microstructures

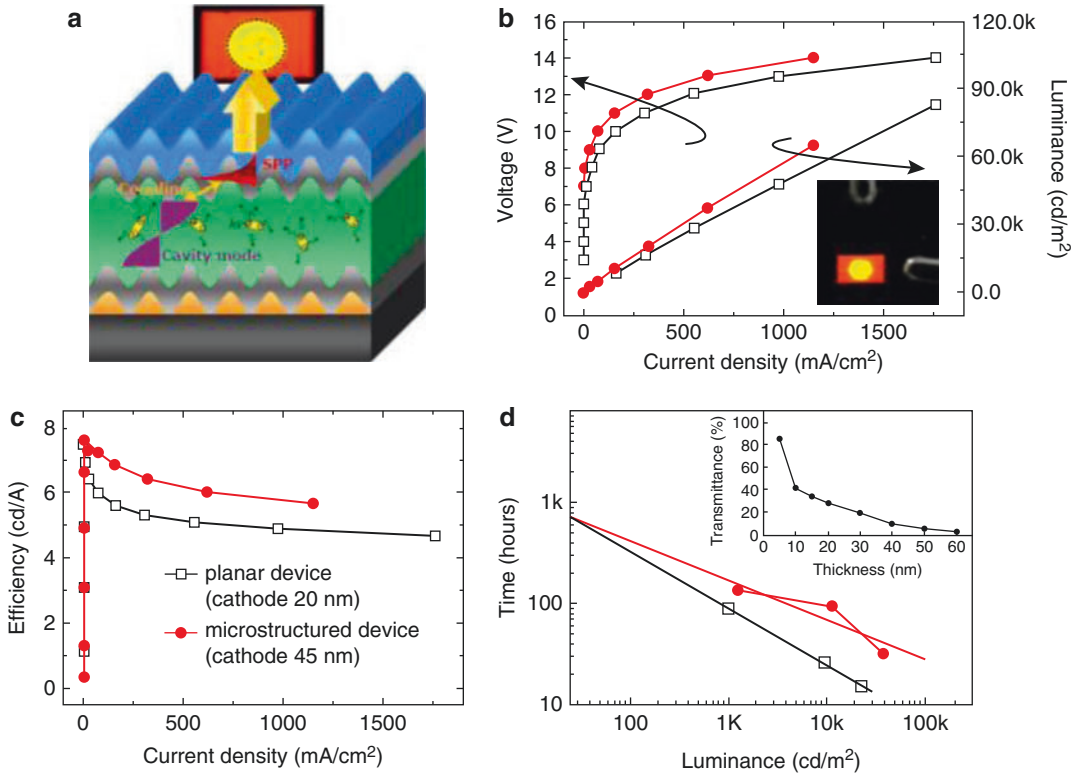
Solving Efficiency-Stability Tradeoff in TOLEDs by Employing Periodically Corrugated Metallic Cathode

In conventional top-emitting organic light-emitting devices (TOLEDs), there exists a fundamental tradeoff between device stability and efficiency, which is really a challenge for their further applications. The semitransparent top cathode of a conventional TOLED is usually composed of multiple functional layers to achieve both optical transmission and effective electron injection [12]. An ultrathin layer of a reactive, low work function metal (~1 nm) is used for electron injection, and a noble metal with high transparency is used to reduce the sheet resistance of the composite cathode and also as a protective layer for the underlying reactive metal and organic layers, for example, bilayer cathodes of Al/Ag, Ca/Ag, etc. Ag is a widely used protective layer because of its low optical absorption and high conductivity. The Ag layer should be less than 20 nm to ensure a high transparency; however, more pinholes are formed in the Ag film because of poor film continuity for such a thin film, which results in easy diffusion of water and oxygen from air into the devices and accelerates the degradation of the devices [13, 14]. Thicker Ag layers offer higher film continuity; however, the higher device stability comes at the expense of lower efficiency because of the decreased optical transmittance.

Employing a periodic corrugation in the metal cathode of the TOLEDs provides a possibility to optimally solve the tradeoff [15]. SPP-mediated emission from an OLED incorporating a periodic wavelength-scale microstructure has been observed [16, 17]. In the case of a thin metal film, SPP excitation on the opposite interface can be supported, and they can couple to each other by the familiar grating-coupling mechanism [18, 19]. This results in a much enhanced light transmission through the classically opaque metal film [20] and permits a relatively thick metal film of ~50 nm [21], which is much more effective in protecting the devices from exposure to atmosphere. In the case of the TOLEDs, the reflective anode and semitransparent cathode are parallel to each other and form a microcavity. The cavity length is fixed by the distance between the two parallel electrodes and defines its resonant wavelength. In this case, the cross coupling occurs between the SPP modes associated with the top surface of the cathode and the microcavity modes within the device, instead of the two SPPs, which also results in an enhanced light transmission. Taking into account that the SPP resonance can be tuned by adjusting the period of the corrugated metal film or the refractive index of the dielectric on the metal surface to coincide with the microcavity modes, the light transmission can be enhanced at a desired wavelength. Therefore, employing a periodically corrugated cathode opens a possible avenue to release the efficiency-stability tradeoff that persists in TOLEDs. Figure 1a demonstrates that the introduction of a periodic corrugation into TOLEDs to realize a corrugated metal cathode is effective in relieving the tradeoff between device stability and efficiency, through the cross coupling of the SPPs associated with the Ag cathode surface and the microcavity modes within the TOLEDs [4]. The thickness of the Ag cathode for the corrugated TOLEDs was increased from 20 to 45 nm, and both the device lifetime and efficiency are significantly improved, which has been examined by performing current density (J)-voltage (V)-luminance (L) and lifetime measurements of the electrically pumped TOLEDs as shown in Fig. 1b–d.

Surface Plasmon Polariton-Mediated Red Emission from OLEDs Based on Metallic Electrodes Integrated with Dual-Periodic Corrugation

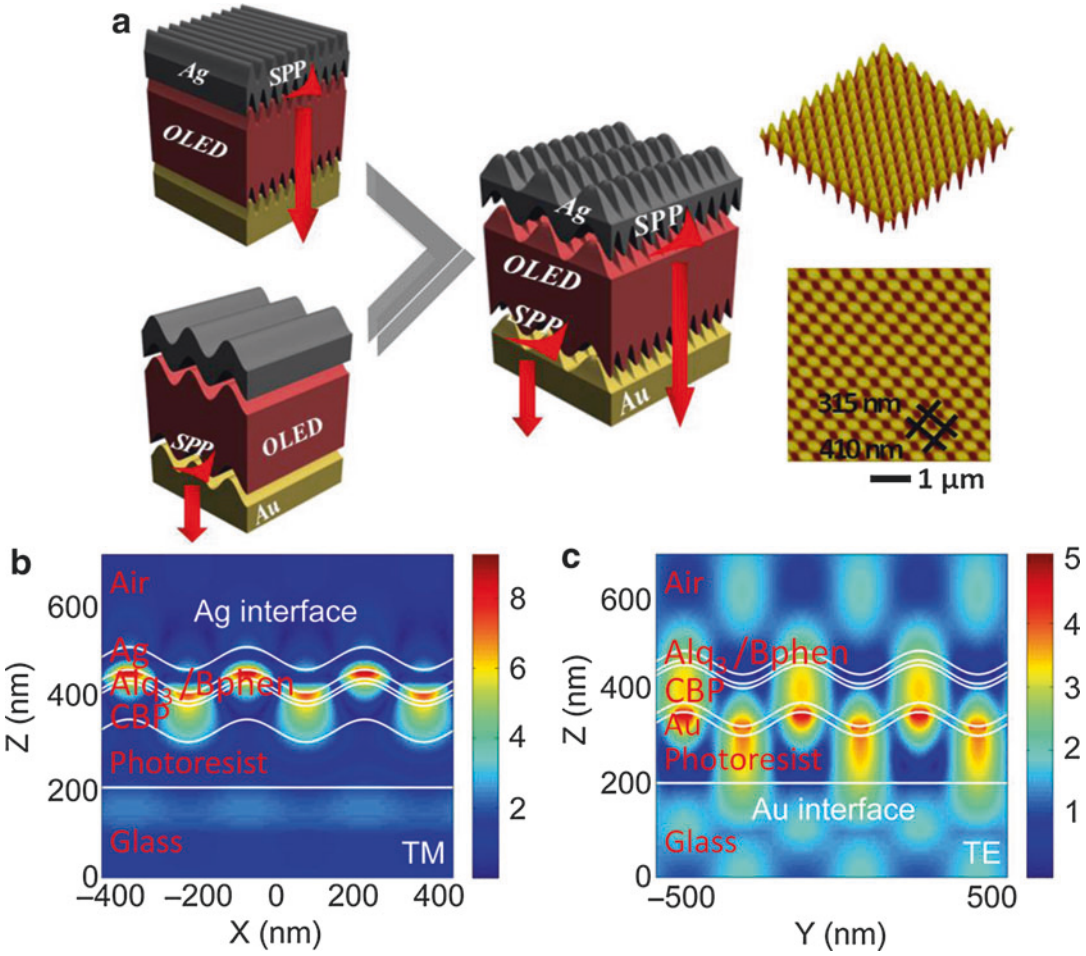
Using metallic anode in OLEDs, for example, thin metal film, metal grid, or metal nanowires with high optical transmission and electrical conductivity as a direct replacement for ITO, has a potential to recover the power lost to the WG modes in ITO [22, 23]. OLEDs with metallic films as both



Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 1 (**a**) Schematic cross section of the *red* TOLED with periodic microstructure. (**b**–**d**) EL performance of the corrugated and planar TOLEDs. Current density-voltage-luminance (**c**), current density efficiency (**d**) characteristics, and half-luminance lifetime (**e**), respectively. *Inset* in (**d**) shows the photograph of the operating corrugated and planar OLEDs at the same substrate and under same driving voltage. *Inset* in (**d**) shows the transmittance of the Ag film with various thicknesses from 5 to 60 nm. Reproduced with permission from Ref. [4]. Copyright 2012, Wiley-VCH

top cathode and bottom anode by employing semitransparent Au thin film as anode have shown superior device performance [3, 24]. In this case, the light trapped in the SPP modes existed at cathode/organic interface and anode/organic interface become the main power lost. Introducing a wavelength-scale microstructure onto the metallic electrode surface has exhibited its remarkable effect for excitation and outcoupling of the SPP modes by providing an additional momentum to couple the SPP modes into light [4, 25]. A periodic microstructure is suitable for specific narrow range of wavelength by satisfying the Bragg scattering condition. SPP resonant wavelength would be different for the cathode and anode interface due to the different metal materials used, and the excitation and outcoupling of the SPP modes in the desired wavelength region can be realized only at one of the electrode interfaces. Therefore, recovering the power lost to SPPs at both cathode and anode interfaces, simultaneously, is an important issue to further improve the efficiency of the OLEDs.

Efficient excitation and outcoupling of the SPP modes associated with both top and bottom electrode/organic interfaces have been realized by introducing a two-dimensional (2-D) grating with dual-periodic corrugation into the OLEDs (Fig. 2a) [26]. The 2-D grating consists two sets of corrugations with different periods. The SPP resonance at the cathode/organic and anode/organic interfaces can be tuned to the same



Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 2 (**a**) Schematic of excitation and outcoupling of the SPP modes associated with both cathode/organic and anode/organic interfaces in OLEDs by engaging dual-periodic corrugation and AFM images of surface morphologies of photoresist with 2-D dual-periodic corrugation. Distributions of the magnetic field intensity across the dual-periodic sample with the normal incident light at the wavelength of 605 nm with TM polarization associated with Ag/organic interface (**b**) and TE polarization associated with Au/organic interface (**c**). Reproduced with permission from Ref. [26]. Copyright 2014 Nature Publishing Group

wavelength to coincide with the electroluminescent (EL) peak of the OLEDs by adjusting the appropriate periods of the two sets of corrugations. As a result, the light trapped in the SPP modes associated with the two electrode/organic interfaces are both efficiently extracted from the OLEDs, respectively. In-house generated finite-difference time-domain (FDTD) code is applied to simulate the magnetic field intensity distributions across the corrugated devices as shown in Figs. 2b and 3c. The numerical simulations confirm the excitation and outcoupling of the SPP modes from both top and bottom metal/organic interfaces in the dual-periodic corrugated devices.

Broadband Light Extraction from White OLEDs by Employing Corrugated Metallic Electrodes with Dual Periodicity

Broadband extraction is important for the efficient outcoupling of the trapped photons associated with substrate mode, SPP, and WG modes,

especially for that trapped in the white organic light-emitting devices (WOLEDs) whose spectra covering the whole visible wavelength. It can be easily realized for broadband extraction of the substrate modes by attaching a microlens array on the outside of the substrate. While in the case of the SPP and WG modes inside of the OLEDs, structure modification has to be introduced inside the device structure, and broadband extraction is difficult to be obtained. Wavelength-scale periodic microstructures introduced into the OLEDs are suitable for specific narrow range of wavelength by satisfying the Bragg scattering condition and applicable only for monochromatic OLEDs [4, 27]. Other microstructures, such as spontaneously formed buckles or defective hexagonal close-packed arrays, have been used to effectively enhance light extraction for OLEDs [28], while their effect on broadband light extraction in WOLEDs has not yet been examined. So far, broadband light extraction from WOLEDs is still a challenge, which is an obstacle for its applications in both display and lighting.

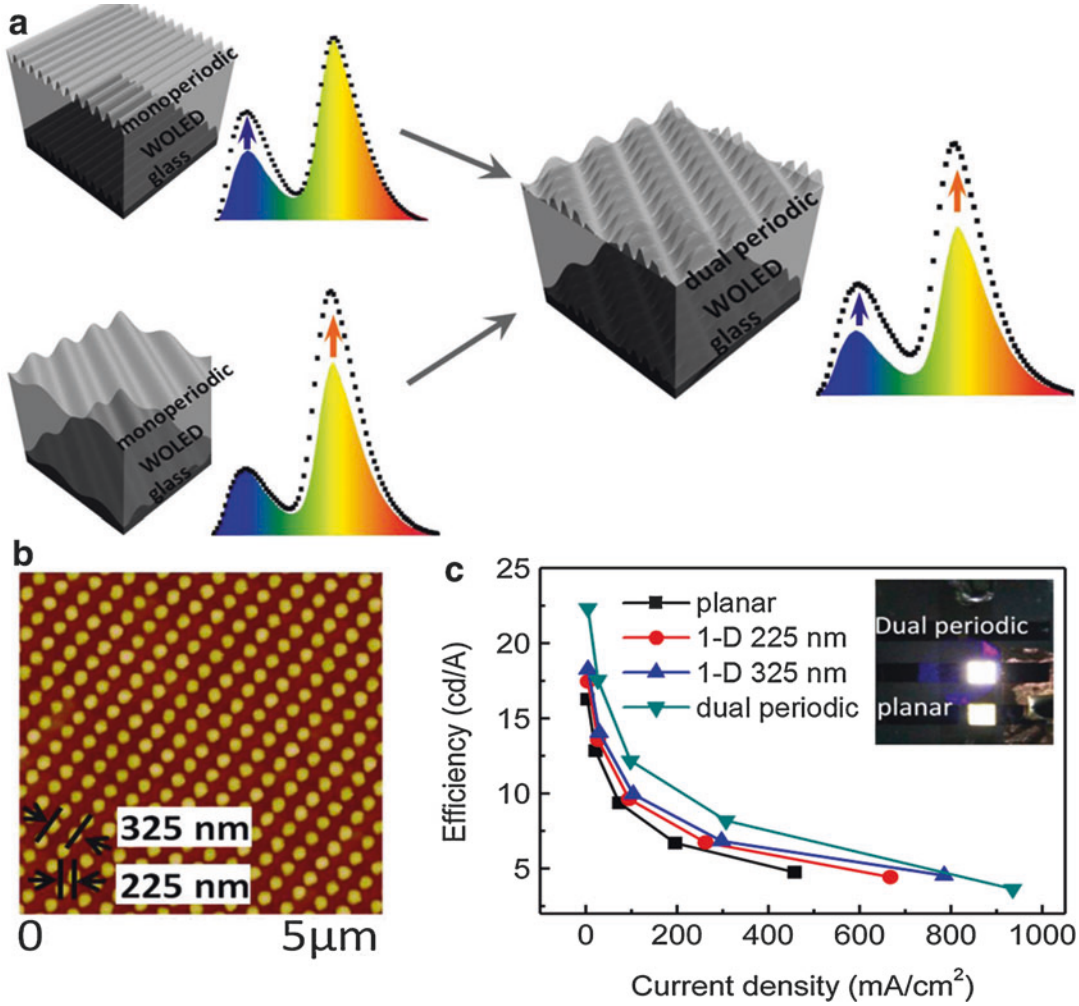
OLEDs with two metallic electrodes by employing metallic film with high optical transmission and electrical conductivity to replace ITO as anode could eliminate the power lost to the WG modes in ITO [3, 29]. In this case, light trapped in the SPP modes are the main power lost, and therefore, highly efficient light extraction could be expected by broadband excitation and outcoupling of the SPP modes in the WOLEDs. Broadband excitation and outcoupling of the SPP modes in the WOLEDs have been realized by introducing a 2-D grating with dual-periodic corrugation into the WOLEDs (Fig. 3a) [24]. The 2-D grating which consists two sets of corrugations with different periods can broaden the SPP resonance compared to that of the monoperiodic grating (Fig. 3b). The blue and orange emissions are both efficiently extracted from the WOLEDs based on the two complementary color strategies by adjusting appropriate periods of the dual-periodic corrugation. Both experimental and numerical results support the validity of the broadband light extraction, and a 37 % enhancement in current efficiency compared to those of the conventional planar devices has been obtained (Fig. 3c) [24].

OSCs Integrated with Microstructures

Surface Plasmon-Enhanced Absorption in OSCs by Employing a Periodically Corrugated Metallic Electrode

OSCs based on conjugated polymers and small molecules are an attractive alternative to Si-based solar cells due to their advantages of low cost, light weight, simple fabrication process, and flexibility [30]. However, the power conversion efficiency (PCE) of the OSCs is still low for commercial applications. The tradeoff between the efficiency of photon absorption and exciton harvesting is one of the main limitation factors. The active layer is generally less than 100 nm due to the short exciton diffusion length, which limits the efficiency of incident light absorption [31]. A thicker active layer offers higher light absorption; however, it comes at the expense of lowered exciton harvesting. Therefore, an exploration of device design to improve light absorption without increasing the physical thickness of the photovoltaic absorber layer is required.

A periodically nanopatterned metal film forms a particularly interesting class, since their periodicity can be tuned to adjust the SPP resonance and the metal film itself can be used as metallic electrode in the devices. Up to 50 % increment of the absorption in the active layers of the OSCs has been predicted by previous theory analysis [32]. Therefore, using the periodically nanopatterned metal to excite SPP may provide a possible way to release the tradeoff between photon absorption and exciton harvesting efficiency. Such a corrugated metallic electrode with wavelength-scale periodicity was integrated into the OSC structure to excite the SPP at the metal/organic interface with tunable resonance [33]. The SPP resonance has been tuned by changing the grating period, so that a coincidence between the SPP resonance and the absorption wavelength region of the absorber layer can be obtained to realize a maximum absorption enhancement. The short-circuit current (Jsc) and the PCE were

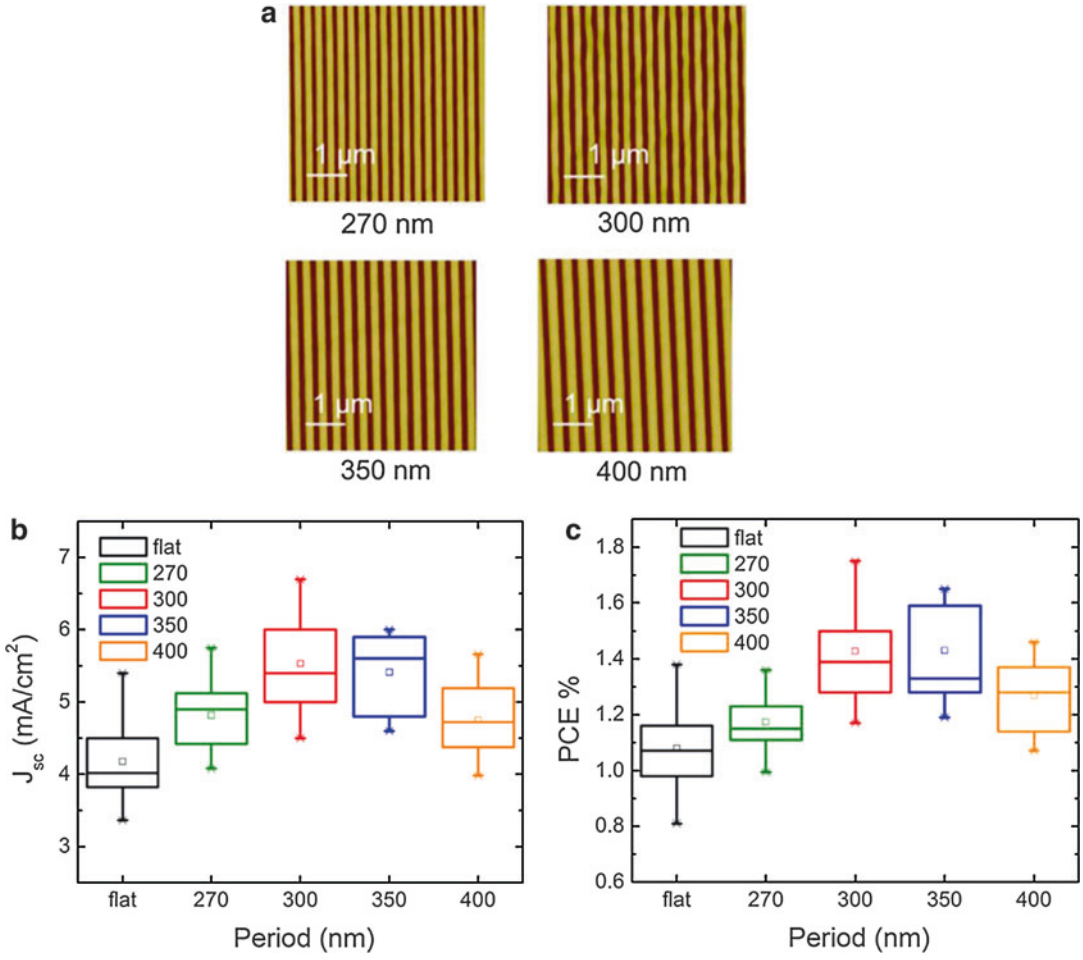


Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 3 (**a**) Schematic of the broadband light extraction by using the dual-periodic corrugation. (**b**) AFM images of surface morphologies of photoresist with 2-D dual-periodic corrugation. (**c**) EL performance of the corrugated and planar WOLEDs. The *inset* in (**c**) shows the photograph of the operating dual-periodic and planar WOLEDs at the same substrate and under the same driving voltage. Reproduced with permission from Ref. [24]. Copyright 2013, Wiley-VCH

significantly improved for the corrugated OSCs with appropriate period (Fig. 4).

Effective and Tunable Light Trapping in Bulk Heterojunction Organic Solar Cells by Employing Au-Ag Alloy Nanoparticles

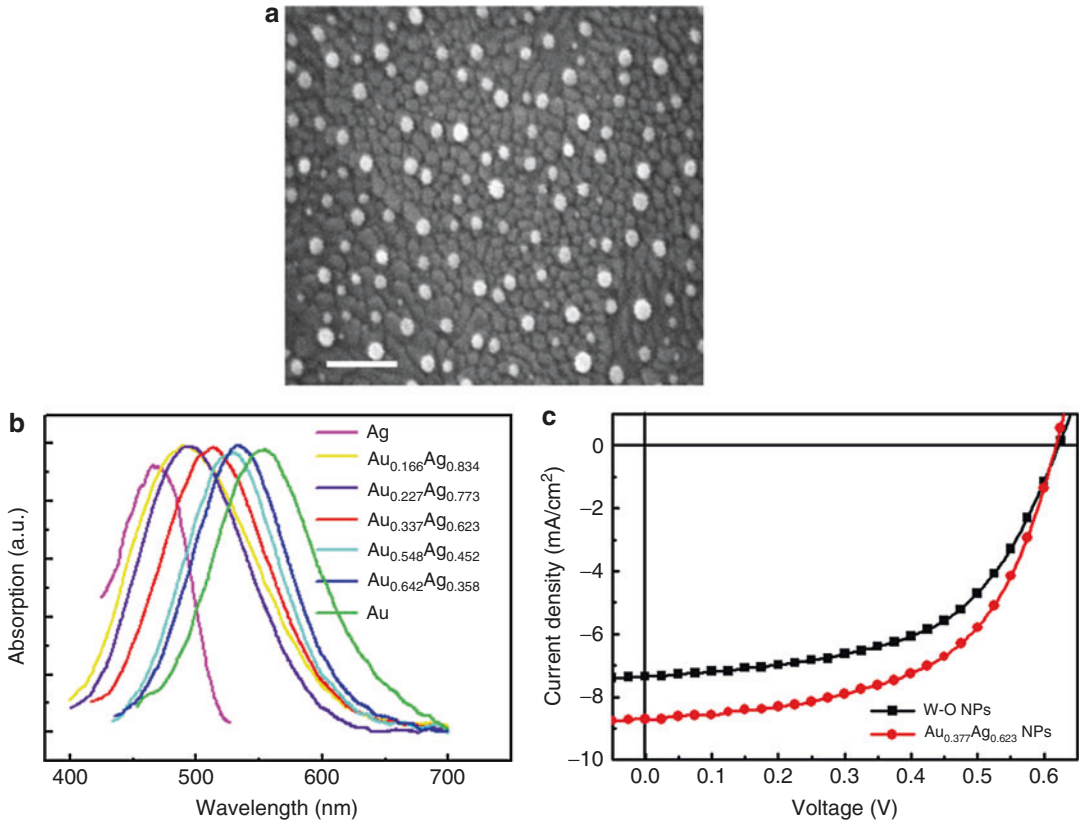
Various light trapping approaches have been explored to solve the conflict about the thickness of the active layer by increasing the absorption [33, 34]. The incorporation of metallic nanoparticles (NPs) into the OSC structures supports the amplification of the electric field near the particle surface through localized SPR (LSPR) and could result in enhanced light absorption owing to the near-field coupling. Au and Ag NPs have been studied in great detail in the OSCs due to their relative strong scattering efficiency, broad LSPR absorption band in the visible range, and chemical stability [35, 36]. Mixing of the Au and Ag NPs has also been investigated in OSCs, which result in larger efficiency enhancement due to a dual resonance enhancement [37].



Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 4 (**a**) AFM images of surface morphology of periodic corrugation with period of 270, 300, 350, and 400 nm. Multiple device statistical performance analysis of the short-circuit current (**b**) and the power conversion efficiency (**c**) produced in planar device and the corrugated devices with various periods. Reproduced with permission from Ref. [33]. Copyright 2012 AIP Publishing LLC

Compared to monometallic NPs and physical mixing of the Au and Ag NPs, the Au-Ag alloy NPs have more advantages due to the potential to combine the best of both metals in terms of plasmonic properties and exhibit unique electronic, optical, and catalytic properties that distinct from those of the corresponding monometallic particles [38, 39]. Moreover, Au and Ag easily form alloys for different compositions with very little surface segregation due to their similar lattice constants [40]. Different from the monometallic NPs, for which both the shape and size of the metallic NPs are key factors determining the LSPR wavelength [41], the LSPR can be tuned easily to the desired spectral range by varying the molar ratio of the Au-Ag alloy NPs. Therefore, the use of the Au-Ag alloy NPs in the OSCs offers a great potential for effective and tunable light trapping.

The thermal annealing method has been employed to fabricate the Au-Ag alloy NPs with tunable molar ratio, and its effect on light trapping in the OSCs has been investigated. The Au-Ag alloy NPs were obtained by a co-evaporation of Au and Ag onto the ITO substrate and followed by the thermally annealing process, so that the molar



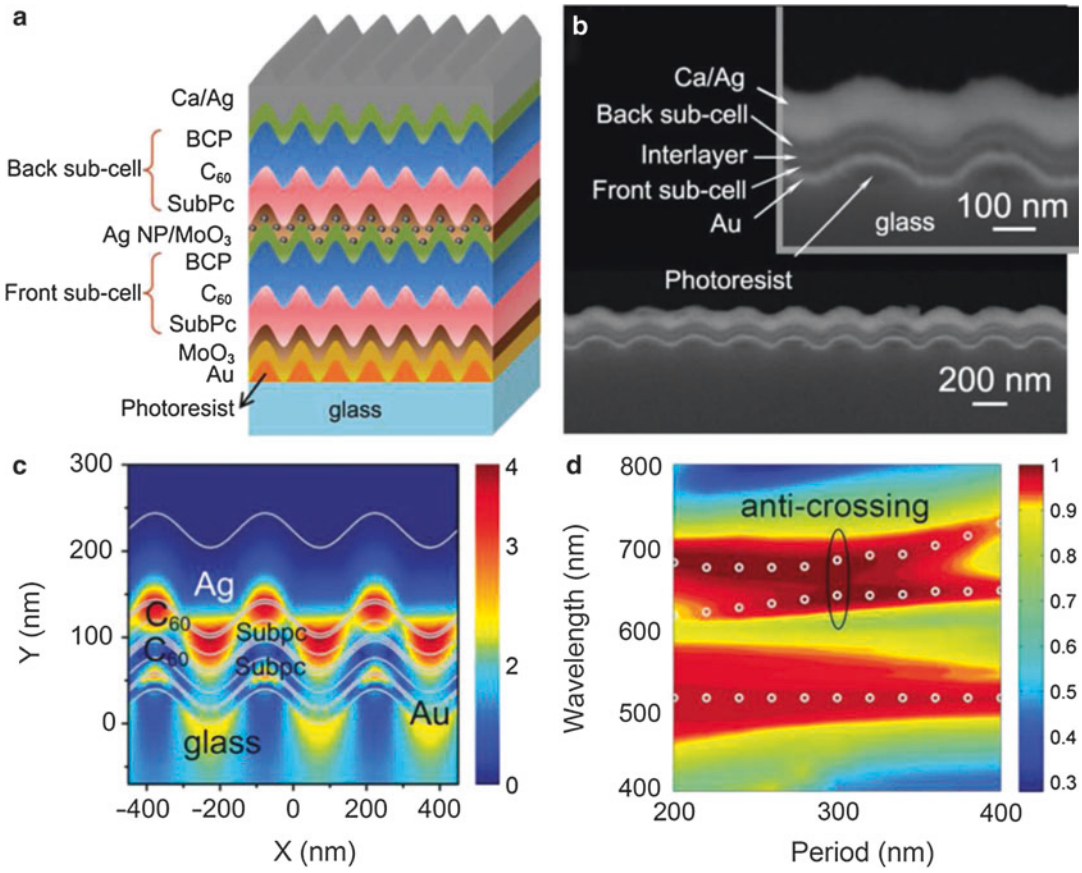
Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 5 (**a**) SEM images of the Au-Ag alloy NPs formed by vacuum annealing on the ITO substrates. (**b**) UV/vis absorption spectra of Au and Ag monometallic NPs and Au-Ag alloy NPs with various molar ratios. (**c**) J-V characteristics of the OSCs with the $Au_{0.377}$ $Ag_{0.623}$ alloy NPs and without the NPs. Reproduced with permission from Ref. [42]. Copyright 2014 AIP Publishing LLC

ratio of the alloy NPs can be easily and precisely tuned by controlling the evaporation rate of the Au and Ag individually (Fig. 5a) [42]. Light absorption enhancement in the active layer has been obtained by strong LSPR excited at the absorption wavelength region of the absorber by the alloy NPs with an appropriate molar ratio (Fig. 5b). Meanwhile, light scattering from the NPs elongates the optical path length which contributes to the light absorption enhancement at a shorter wavelength. The short-circuit photocurrent density (Jsc) of the alloy NP-based OSCs is improved from 7.37 to 8.74 mA/ cm^2, and the efficiency of 3.03 % is obtained, which corresponds to an enhancement factor of 19 % (Fig. 5c).

Matching Photocurrents of Sub-cells in Double-Junction Organic Solar Cells via Coupling Between Surface Plasmon Polaritons and Microcavity Modes

Tandem configurations, in which two or more single cells are stacked in series, have been accepted as a successful and universal strategy to increase the V_{oc} of the OSCs [43]. A record of up to 7 V for six-junction solar cells has been reported [44]. However, the photocurrent of the tandem cells is usually lower than that of the single-junction cells, because less light is absorbed by the back sub-cell compared to the front one in a typical tandem device, which creates an imbalance in the photocurrent generated by the

S



Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices, Fig. 6 (**a**) Schematic structure of the corrugated double-junction solar cells. (**b**) SEM cross section of the corrugated device. (**c**) Distribution of the magnetic field intensity in the corrugated double-junction cells with period of 300 nm at the wavelength of 650 nm. (**d**) Dispersion relationship for the wavelength versus the period of the corrugation of the corrugated double-junction cells. Reproduced with permission from Ref. [34]. Copyright 2013, Wiley-VCH

back and front sub-cells and limits the PCE of the overall device. Appropriate choice of materials with different bandgaps can overcome this problem to a certain degree [45]. However, this approach is difficult to apply to multi-junction organic solar cells, especially for small molecule-based cells, due to the limited choice of the small molecule donor materials that have significantly different absorption profiles. Careful design of the optical field and photovoltaic contributions from the sub-cells is another strategy to optimize the tandem cell. Surface plasmon polariton (SPP)-induced field enhancement has been believed to be a highly attractive solution to enhance optical absorption in an organic solar cell without increasing the thickness of its active layers [33, 46].

SPP resonance has been employed in the small molecule-based double-junction organic solar cells with two identical sub-cells to enhance the optical absorption of the back sub-cell by employing periodical corrugation in the metallic cathode (Fig. 6) [34]. Both the SPP and microcavity modes are supported by the corrugated device structure, and the SPP resonance is tuned by tuning the period of the corrugation, so that an anti-crossing between the SPP and microcavity modes within the device is realized. This anti-crossing plays an important role in

enhancing the absorption of the back sub-cell and thereafter achieving a balanced photocurrent of front and back sub-cells, which results in an improved PCE for the double-junction device. Both simulation and experimental results support the effect of the anti-crossing behavior on the absorption enhancement. For the double-junction devices with this periodical corrugation, 10.4 % enhancement in the photocurrent and 11.3 % enhancement in the PCE have been achieved.

Conclusions

Microstructures have been introduced into optoelectronic devices, and enhanced performances have been demonstrated. The introduction of periodic corrugations has allowed the observation of the emission originating from the SPP modes, which are usually trapped within planar OLEDs. By employing a periodic microstructure in TOLEDs, a much enhanced light transmission through a thick cathode has been observed through the grating-induced cross coupling between the SPPs associated with top interface of the cathode and microcavity modes within the device cavity. Dual-periodic corrugation has been introduced into WOLED metallic electrodes to realize broadband light extraction. Improved efficiency of OSCs by employing a periodically corrugated metallic electrode has been demonstrated. Light absorption enhancement in the active layer has been obtained by strong LSPR excited at the absorption wavelength region of the absorber by the alloy NPs with an appropriate molar ratio. By employing a periodic wavelength-scale corrugation into the device structure, the optical absorption of the back sub-cell in the double-junction organic solar cells has been enhanced by strong anti-crossing coupling between the microcavity and propagating SPP modes at the Ag/back sub-cell interface.

Cross-References

- ▶ Light Emitting Diodes
- ▶ Organic Photovoltaics: Basic Concepts and Device Physics

References

1. Park, S.H., et al.: Bulk heterojunction solar cells with internal quantum efficiency approaching 100%. Nat. Photon. **3**, 297–303 (2009)
2. D'Andrade, B.W., Forrest, S.R.: White organic light-emitting devices for solid-state lighting. Adv. Mater. **16**, 1585–1595 (2004)
3. Bi, Y.G., et al.: Enhanced efficiency of organic light-emitting devices with metallic electrodes by integrating periodically corrugated structure. Appl. Phys. Lett. **100**, 053304 (2012)
4. Jin, Y., et al.: Solving efficiency-stability tradeoff in top-emitting organic light-emitting devices by employing periodically corrugated metallic cathode. Adv. Mater. **24**, 1187–1191 (2012)
5. Liu, L., Stanchina, W.E., Li, G.: Effects of semiconducting and metallic single-walled carbon nanotubes on performance of bulk heterojunction organic solar cells. Appl. Phys. Lett. **94**, 233309 (2009)
6. Brabec, C.J., Sariciftci, N.S., Hummelen, J.C.: Plastic solar cells. Adv. Funct. Mater. **11**, 15–26 (2001)
7. Nalwa, K.S., Park, J.M., Ho, K.M., Chaudhary, S.: On realizing higher efficiency polymer solar cells using a textured substrate platform. Adv. Mater. **23**, 112 (2011)
8. Baba, A., Wakatsuki, K., Shinbo, K., Kato, K., Kaneko, F.: Increased short-circuit current in grating-coupled surface plasmon resonance field-enhanced dye-sensitized solar cells. J. Mater. Chem. **21**, 16436–16441 (2011)
9. Sha, W.E.I., Choy, W.C.H., Chew, W.C.: A comprehensive study for the plasmonic thin-film solar cell with periodic structure. Opt. Express **18**, 5993–6007 (2010)
10. Yang, J., et al.: Plasmonic polymer tandem solar cell. ACS Nano **5**, 6210–6217 (2011)
11. Wang, D.H., et al.: Enhanced power conversion efficiency in PCDTBT/PC70BM bulk heterojunction photovoltaic devices with embedded silver nanoparticle clusters. Adv. Energy Mater. **1**, 766–770 (2011)
12. Hung, L.S., Tang, C.W., Mason, M.G., Raychaudhuri, P., Madathil, J.: Application of an ultrathin LiF/Al bilayer in organic surface-emitting diodes. Appl. Phys. Lett. **78**, 544–546 (2001)
13. Lim, S.F., Ke, L., Wang, W., Chua, S.J.: Correlation between dark spot growth and pinhole size in organic light-emitting diodes. Appl. Phys. Lett. **78**, 2116–2118 (2001)
14. Kolosov, D., et al.: Direct observation of structural changes in organic light emitting devices during degradation. J. Appl. Phys. **90**, 3242–3247 (2001)
15. Atwater, H.A., Polman, A.: Plasmonics for improved photovoltaic devices. Nat. Mater. **9**, 205–213 (2010)
16. Hobson, P.A., Wedge, S., Wasey, J.A.E., Sage, I., Barnes, W.L.: Surface plasmon mediated emission from organic light-emitting diodes. Adv. Mater. **14**, 1393–1396 (2002)

17. Lupton, J.M., Matterson, B.J., Samuel, I.D.W., Jory, M.J., Barnes, W.L.: Bragg scattering from periodically microstructured light emitting diodes. Appl. Phys. Lett. **77**, 3340–3342 (2000)
18. Ebbesen, T.W., Lezec, H.J., Ghaemi, H.F., Thio, T., Wolff, P.A.: Extraordinary optical transmission through sub-wavelength hole arrays. Nature **391**, 667–669 (1998)
19. Hooper, I.R., Sambles, J.R.: Coupled surface plasmon polaritons on thin metal slabs corrugated on both surfaces. Phys. Rev. B **70**, 045421 (2004)
20. Gruhlke, R.W., Holland, W.R., Hall, D.G.: Surface plasmon cross coupling in molecular fluorescence near a corrugated thin metal film. Phys. Rev. Lett. **56**, 2838–2841 (1986)
21. Wedge, S., Hooper, I.R., Sage, I., Barnes, W.L.: Light emission through a corrugated metal film: the role of cross-coupled surface plasmon polaritons. Phys. Rev. B **69**, 245418 (2004)
22. Kuang, P., et al.: A new architecture for transparent electrodes: relieving the trade-off between electrical conductivity and optical transmittance. Adv. Mater. **23**, 2469–2473 (2011)
23. Chiappe, D., Toma, A., de Mongeot, F.B.: Transparent plasmonic nanowire electrodes via self-organised ion beam nanopatterning. Small **9**, 913–919 (2013)
24. Bi, Y.G., et al.: Broadband light extraction from white organic light-emitting devices by employing corrugated metallic electrodes with dual periodicity. Adv. Mater. **25**, 6969–6974 (2013)
25. Yates, C.J., Samuel, I.D.W., Burn, P.L., Wedge, S., Barnes, W.L.: Surface plasmon-polariton mediated emission from phosphorescent dendrimer light-emitting diodes. Appl. Phys. Lett. **88**, 161105 (2006)
26. Bi, Y.G., et al.: Surface plasmon-polariton mediated red emission from organic light-emitting devices based on metallic electrodes integrated with dual-periodic corrugation. Sci. Rep. **4**, 7108 (2014)
27. Bai, Y., et al.: Outcoupling of trapped optical modes in organic light-emitting devices with one-step fabricated periodic corrugation by laser ablation. Org. Electron. **12**, 1927–1935 (2011)
28. Koo, W.H., et al.: Light extraction of organic light emitting diodes by defective hexagonal-close-packed array. Adv. Funct. Mater. **22**, 3454–3459 (2012)
29. Helander, M.G., et al.: Oxidized gold thin films: an effective material for high-performance flexible organic optoelectronics. Adv. Mater. **22**, 2037–2040 (2010)
30. Peumans, P., Forrest, S.R.: Very-high-efficiency double-heterostructure copper phthalocyanine/C60 photovoltaic cells. Appl. Phys. Lett. **79**, 126–128 (2001)
31. Agrawal, M., Peumans, P.: Broadband optical absorption enhancement through coherent light trapping in thin-film photovoltaic cells. Opt. Express **16**, 5385–5396 (2008)
32. Min, C., et al.: Enhancement of optical absorption in thin-film organic solar cells through the excitation of plasmonic modes in metallic gratings. Appl. Phys. Lett. **96**, 133302 (2010)
33. Jin, Y., et al.: Surface-plasmon enhanced absorption in organic solar cells by employing a periodically corrugated metallic electrode. Appl. Phys. Lett. **101**, 163303 (2012)
34. Jin, Y., et al.: Matching photocurrents of sub-cells in double-junction organic solar cells via coupling between surface plasmon polaritons and microcavity modes. Adv. Opt. Mater. **1**, 809–813 (2013)
35. Rand, B.P., Peumans, P., Forrest, S.R.: Long-range absorption enhancement in organic tandem thin-film solar cells containing silver nanoclusters. J. Appl. Phys. **96**, 7519–7526 (2004)
36. Tong, S.W., et al.: Improvement in the hole collection of polymer solar cells by utilizing gold nanoparticle buffer layer. Chem. Phys. Lett. **453**, 73–76 (2008)
37. Lu, L., Luo, Z., Xu, T., Yu, L.: Cooperative plasmonic effect of Ag and Au nanoparticles on enhancing performance of polymer solar cells. Nano Lett. **13**, 59–64 (2013)
38. Qingbo, Z., Jim Yang, L., Jun, Y., Chris, B., Jixuan, Z.: Size and composition tunable Ag–Au alloy nanoparticles by replacement reactions. Nanotechnology **18**, 245605 (2007)
39. Kelly, K.L., Coronado, E., Zhao, L.L., Schatz, G.C.: The optical properties of metal nanoparticles: the influence of size, shape, and dielectric environment. J. Phys. Chem. B **107**, 668–677 (2003)
40. Li, Z.Y., et al.: Structures and optical properties of 4–5 nm bimetallic AgAu nanoparticles. Faraday Discuss. **138**, 363–373 (2008)
41. Catchpole, K.R., Polman, A.: Design principles for particle plasmon enhanced solar cells. Appl. Phys. Lett. **93**, 191113 (2008)
42. Xu, M., et al.: Effective and tunable light trapping in bulk heterojunction organic solar cells by employing Au-Ag alloy nanoparticles. Appl. Phys. Lett. **105**, 153303 (2014)
43. Cheyns, D., Rand, B.P., Heremans, P.: Organic tandem solar cells with complementary absorbing layers and a high open-circuit voltage. Appl. Phys. Lett. **97**, 033301 (2010)
44. Sullivan, P., et al.: Ultra-high voltage multijunction organic solar cells for low-power electronic applications. Adv. Energy Mater. **3**, 239–244 (2013)
45. Dou, L., et al.: Tandem polymer solar cells featuring a spectrally matched low-bandgap polymer. Nat. Photon. **6**, 180–185 (2012)
46. Duche, D., et al.: Improving light absorption in organic solar cells by plasmonic contribution. Solar Energy Mater. Solar Cells **93**, 1377–1382 (2009)

Surface Plasmon Resonance

► Quantum System Near Metallic Particle

Surface Plasmon-Coupled Emission

▶ Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons

Surface Plasmon-Polariton Photodetectors

▶ Surface Plasmon-Polariton-Based Detectors

Surface Plasmon-Polariton-Based Detectors

Pierre Berini
School of Information Technology and Engineering (SITE), University of Ottawa, Ottawa, ON, Canada

Synonyms

Surface plasmon-polariton photodetectors

Definition

A surface plasmon polariton is a transverse-magnetic optical surface wave propagating along the interface of a metal and a dielectric; it is a coupled excitation formed from electromagnetic fields coupled to a charge density wave in the metal. A surface plasmon-polariton detector is a device capable of detecting surface plasmons or involving surface plasmons in the detection process.

Notation

An $e^{+j\omega t}$ time-harmonic dependence is assumed. The relative permittivity is denoted ε_r, and is written for a metal in terms of real and imaginary parts as $\varepsilon_{r,m} = -\varepsilon_R - j\varepsilon_I$. $\boldsymbol{k}$ denotes a wavevector (in general). $k_0 = 2\pi/\lambda_0 = \omega/c_0$ is the wavenumber of plane waves in vacuum, λ_0 the wavelength in vacuum, c_0 the speed of light in vacuum, and $\omega = 2\pi v$ is the angular frequency.

Single-Interface Surface Plasmon-Polariton

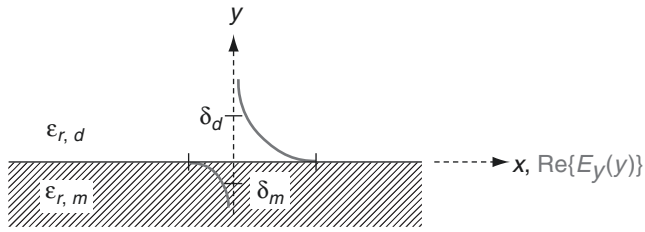
A surface plasmon-polariton (SPP) waveguide is a metallo-dielectric structure along which SPP modes are guided. The simplest structure supporting SPPs is an interface between an optically semi-infinite dielectric and an optically semi-infinite metal [1], as sketched in Fig. 1 (the relative permittivity of the metal is denoted $\varepsilon_{r,m}$ and that of the dielectric $\varepsilon_{r,d}$.) This structure is termed the single-interface.

Highly conductive metals and good dielectrics are used to implement SPP waveguides. Ag is often preferred because it has the lowest optical loss among the metals over a broad wavelength range, but it is reactive, so care must be taken during fabrication and use in order to avoid degradation (which can occur if Ag is exposed to, e.g., air or water). Au is also a good choice given its chemical stability and its good optical performance. A good metal for supporting SPPs normally satisfies $\varepsilon_R >> \varepsilon_I$.

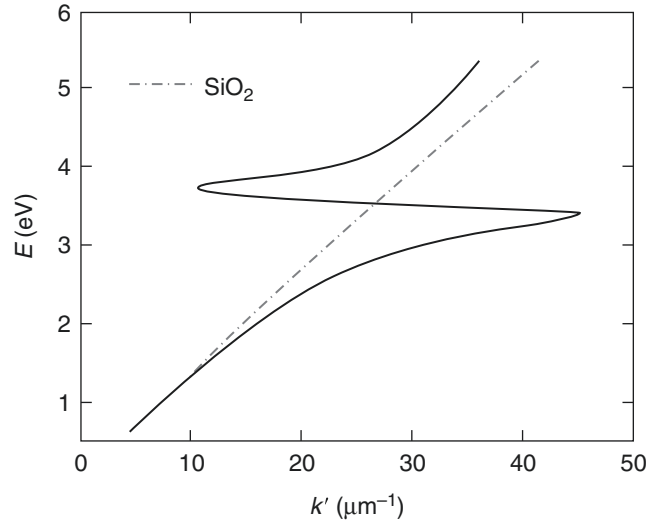
Metals are dispersive at optical wavelengths. Away from interband transitions, the Drude model for the permittivity captures the dispersive character of metals [1]:

$$\varepsilon_{r,m} = -\varepsilon_R - j\varepsilon_I = 1 - \frac{\omega_p^2}{\omega^2 + 1/\tau^2} - j\frac{\omega_p^2/\tau}{\omega(\omega^2 + 1/\tau^2)} \quad (1)$$

In the above, ω_p is the plasma frequency and τ the relaxation time. The "Drude region" corresponds to that portion of the electromagnetic spectrum where Eq. 1 holds. For many metals, this region spans the range from visible wavelengths to the infrared. The metal approaches a perfect electric conductor as the wavelength increases through the infrared and beyond.



Surface Plasmon-Polariton-Based Detectors, Fig. 1 Single-interface surface plasmon-polariton waveguide; the distribution of the main transverse electric field component (E_y) of the SPP is sketched on the structure as the *thick gray curve*



Surface Plasmon-Polariton-Based Detectors, Fig. 2 Dispersion of the SPP along a Ag/SiO_2 interface. The *light line* in SiO_2 is plotted as the *dash-dot curve*

The single-interface (Fig. 1) supports one purely bound (nonradiative) SPP mode. This mode is transverse magnetic (TM) and may propagate at any angle in the *x*-*z* plane (e.g., along + *z*). The SPP fields (E_y, E_z, and H_x) are confined along the *y* direction, peaking at the interface and decaying exponentially into both media. The distribution of the main transverse electric field component (E_y) of the SPP is sketched on the structure of Fig. 1 as the thick gray curve. The field penetration depth in the metal δ_m is much smaller than the field penetration depth in the dielectric δ_d. Confinement of the SPP arises because the metal and dielectric have Re{ε_r} of opposite sign at the wavelength of operation (indeed over a large wavelength range).

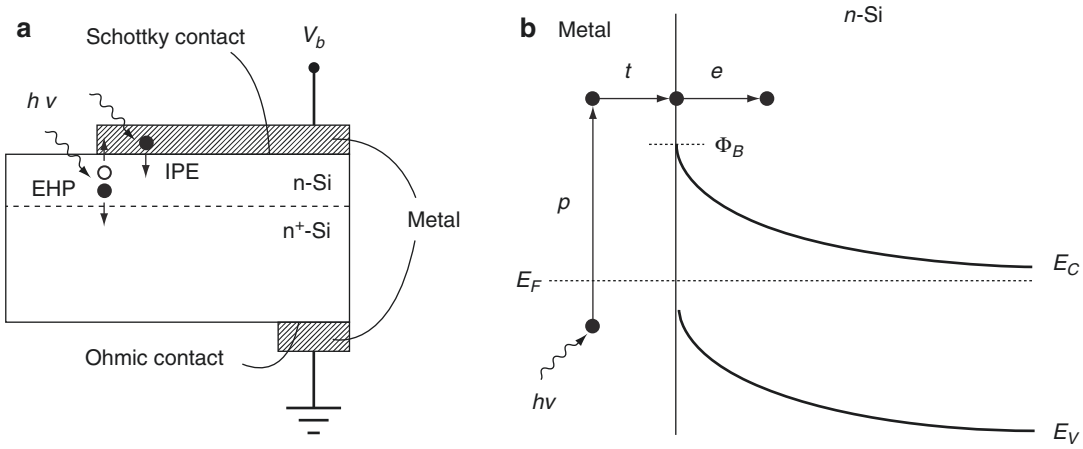
The wavenumber of the single-interface SPP is [1]:

$$k^{\mathrm{SPP}} = k_0 \left(\frac{\varepsilon_{r,m}\varepsilon_{r,d}}{\varepsilon_{r,m} + \varepsilon_{r,d}} \right)^{\frac{1}{2}} \tag{2}$$

For a lossless dielectric cladding (Im{$\varepsilon_{r,d}$} = 0), the above simplifies to the following approximate expressions for the real (phase) and imaginary (attenuation) parts of k^{SPP} [1]:

$$k' \cong k_0 \left(\frac{\varepsilon_R \varepsilon_{r,d}}{\varepsilon_R - \varepsilon_{r,d}} \right)^{\frac{1}{2}} \text{and}$$
$$\mathrm{k}'' \cong \mathrm{k}_0 \frac{\varepsilon_I}{2\varepsilon_R^2} \left(\frac{\varepsilon_{r,d}\varepsilon_R}{\varepsilon_R - \varepsilon_{r,d}} \right)^{\frac{3}{2}} \tag{3}$$

Figure 2 plots the dispersion curve of the SPP ($E = h\nu$ in eV versus k', h is Planck's constant) on a Ag/SiO_2 interface [2]; the light line in SiO_2 is also plotted as the dash-dot curve. As the frequency decreases (increasing wavelength), the metal approaches a PEC, the confinement of the



Surface Plasmon-Polariton-Based Detectors, Fig. 3 (**a**) Schottky diode on n-Si illuminated by light of photon energy hv. Reverse biasing ($V_b < 0$) is assumed. Electrons are depicted by filled circles and holes by unfilled ones. (**b**) Energy band diagram of a Schottky contact on n-Si and the three-step internal photoemission process: p photoexcitation, t transport, e emission. E_C and E_V are the conduction and valence band edges, respectively, E_F is the Fermi level, and Φ_B is the Schottky barrier height

SPP decreases, and its dispersion curve merges with the light line (SPPs are not supported at the interface between a smooth PEC and a semi-infinite dielectric). As the frequency increases, the SPP approaches an "energy asymptote," readily observed in Fig. 2 near $E \sim 3.4$ eV ($\lambda_0 \sim 360$ nm) where k' diverges. The group velocity decreases and the optical density-of-states increases as the asymptote is approached. The SPP bend-back is also observed for wavelengths shorter than the asymptote ($\lambda_0 < 360$ nm, $E > 3.4$ eV) and links the nonradiative SPP on the right of the light line to the radiative one on the left.

The nonradiative SPP is located to the right of the light line (Fig. 2) and so cannot be directly excited by an incident beam. An additional structure is required to increase the in-plane momentum of the beam to match that of the SPP. A prism or a corrugated grating is commonly used to accomplish this task; alternatively, the SPP can be excited by end-fire coupling [1].

Detection Mechanisms

Figure 3a gives a schematic of a conventional Schottky photodiode on n-type Si (n-Si) [3], adopted here to describe generic detection mechanisms that can also be used for the detection of SPPs. A rectifying Schottky contact is formed at the abrupt interface between the top metal and the semiconductor, and an Ohmic (non-rectifying) contact is formed at the interface between the bottom metal and the heavily doped body. The structure is reverse biased (for detection) by applying $V_b < 0$. The complementary structure, obtained by exchanging the dopants (n $\leftrightarrow$ p, $n^+ \leftrightarrow p^+$), is also of interest. The Schottky diode is a convenient structure with which to detect SPPs because SPPs may propagate along its metal surfaces.

Figure 3a shows two mechanisms used for photodetection. The first consists of the creation of electron–hole pairs (EHPs) in the semiconductor due to absorption therein of incident radiation of energy $h\nu$ greater than the bandgap energy of the semiconductor E_g. This mechanism, sketched in Fig. 3a and labeled EHP, involves three steps: optical absorption and creation of EHPs, separation of EHPs and transport across the absorption region (with or without gain) under reverse bias, and collection of EHPs into the photocurrent at the device contacts.

The second mechanism consists of the internal photoemission (IPE) of hot carriers created in the metal due to absorption therein of incident

radiation of energy $h\nu$. This mechanism is sketched in Fig. 3a, labeled IPE, and is described in greater detail via the energy band diagram of Fig. 3b. IPE is a three-step process consisting of the photoexcitation of hot (energetic) carriers in the metal by optical absorption (p), the transport and scattering of hot carriers toward the Schottky contact (t), and the emission of hot carriers over the Schottky barrier into the semiconductor (e) where they are collected under a reverse bias as the photocurrent. This mechanism requires that $h\nu$ be greater than the Schottky barrier energy Φ_B, and is useful when $\Phi_B < h\nu < E_g$; i.e., for detection at energies below the bandgap of the semiconductor (because the creation of EHPs in the semiconductor is more efficient and dominates over IPE when $h\nu > E_g$).

The internal quantum efficiency η_i (internal photoyield) is a useful measure to characterize and optimize the detection mechanism. It is defined as the number of carriers that contribute to the photocurrent I_p per absorbed photon per second:

$$\eta i = \frac{I_p/q}{S_{abs}/h\nu} \tag{4}$$

where S_{abs} is the absorbed optical power and q is the elemental charge. $\eta_i \sim 1$ for detection via the creation of EHPs in high-quality (defect-free) direct bandgap semiconductors (assuming absorption in the semiconductor only). In the case of detection via IPE, assuming absorption in the metal only near the Schottky contact, η_i is given approximately by [4]:

$$\eta_i = \frac{1}{2}\left(1 - \sqrt{\frac{\Phi_B}{h\nu}}\right)^2 \tag{5}$$

Typical Schottky barrier heights are $\Phi_B = 0.34$, 0.8, 0.58, and 0.72 eV for Au/p-Si, Au/n-Si, Al/p-Si, and Al/n-Si, respectively [3]; for detection at, e.g., $\lambda_0 = 1{,}310$ nm η_i ranges from about 0.3 % to 9 %. Metal silicides can also be used, providing lower Schottky barriers.

The external quantum efficiency η_e (external photoyield) and the responsivity R describe how well the detector performs when inserted into a system. η_e is defined similarly to η_i except that it depends on the incident optical power S_{inc}:

$$\eta_e = \frac{I_p/q}{S_{\text{inc}}/h\nu} \tag{6}$$

η_e and η_i are related by:

$$\eta_e = A\eta_i \tag{7}$$

where A is the optical absorptance, defined as:

$$A = \frac{S_{\text{abs}}}{S_{\text{inc}}} \tag{8}$$

The responsivity R is given by the ratio of the photocurrent to the incident optical power, and can be expressed in terms of η_e and η_i:

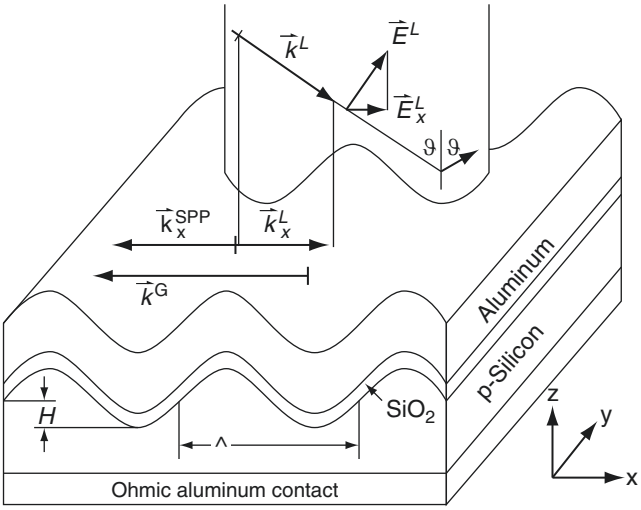
$$R = \frac{I_p}{S_{\text{inc}}} = \frac{\eta_e q}{h\nu} = \frac{A\eta_i q}{h\nu} \tag{9}$$

SPP detectors, or SPP-enhanced detectors, typically combine a metallic structure that supports SPPs with a detector structure such as that shown in Fig. 3.

Grating-Coupled Detectors

An SPP detector structure of interest combines a Schottky detector with a grating designed to couple incoming optical waves to SPPs, as shown in Fig. 4 [5]. The structure consists of a top Al electrode on a thin SiO_2 tunneling barrier on p-Si, formed as a sinusoidal corrugated grating of height H, period Λ, and grating vector $\boldsymbol{k}^G = \boldsymbol{x}2\pi/\Lambda$ ($\boldsymbol{x}$ is the x-directed unit vector). p-polarized light having a wavevector $\boldsymbol{k}^L$ and an electric field $\boldsymbol{E}^L$ is incident onto the grating at the angle ϑ. The incident light couples to the SPP propagating along the air/Al interface in the x direction when the following momentum conservation equation is satisfied:

$$k_x^L + mk^G = k_x^{\text{spp}} \tag{10}$$



Surface Plasmon-Polariton-Based Detectors, Fig. 4 SPP Schottky detector integrated with a grating coupler to excite *x*-propagating SPPs along the top surface of the top Al contact (Adapted from Ref. [5]. © 1989 Optical Society of America)

where k_x^L is the *x*-directed component of $\boldsymbol{k^L}$, *m* is an integer, and k_x^{SPP} is the SPP wavenumber (following the notation of [5]). Equation 10 holds for shallow gratings – *H* is a small perturbation to the surface. The Al Schottky contact is thin enough (35 nm) for the SPP to tunnel through and leak into the high refractive index p-Si layer for detection via the creation of EHPs at the wavelength of operation ($\lambda_0 = 646$ nm).

Grating-coupled SPP detectors are sensitive to the angle of incidence, polarization, and wavelength of the incident light through Eq. 10. The polarization sensitivity, for instance, can be exploited in an arrangement of two detectors to determine the polarization angle of linearly polarized normally incident ($\vartheta = 0$) light [5].

Grating-coupled SPPs have also been used to increase the absorptance in the metal contact of Schottky detectors based on IPE (for sub-bandgap detection) leading to increased responsivity. For example, a 30× increase in responsivity was observed for a Au/p-InP Schottky detector by exciting SPPs along the air/Au interface at $\lambda_0 = 1{,}150$ nm via an integrated corrugated grating, relative to the same structure without the grating [6].

Prism coupling, in the Otto arrangement [1] for instance, has also been used to excite SPPs on detector structures as an alternative to grating coupling. For example, a 10–20× increase in IPE was observed for Au/n-GaAs Schottky detectors by exciting SPPs along the Au/GaAs interface at $\lambda_0 = 1{,}150$ nm in an Otto arrangement implemented monolithically through the substrate [7].

Hole-Coupled Detectors

Optical transmission through a single sub-wavelength hole in a metal film can be significantly larger than predicted by Bethe's theory for a hole in a perfect conductor [8]. This larger (extraordinary) transmission is due to the excitation of SPPs on the metal surface near the hole that then propagate through the hole and couple to radiation on the other side; i.e., the transfer of energy through the hole is mediated by SPPs. Transmission at visible wavelengths is largest for a hole that is 150–300 nm in diameter in 200–300 nm thick Au or Ag films [8].

Structuring the metal film into a corrugation, e.g., around a sub-wavelength hole or slit leads to increased collection of incident light and to greater transmission at resonant wavelengths [8]. Applying this concept to detectors leads to devices that may offer improved performance [9–11].

Ishi et al. [9] reported a Schottky detector on n-Si where the metal forming the Schottky contact was structured into a circular corrugated grating surrounding a 300 nm diameter hole (and covered

by SiO_2). The contact metal was a Ag/Cr stack 200/10 nm thick. Detection was reported at 840 nm based on absorption in a 300 nm diameter n-Si mesa located directly below the hole, and the creation of EHPs therein. A prospective advantage of the structure is high-speed operation due to the small size of the detection volume while maintaining good responsivity due to the light collection ability of the grating/hole combination; thus the detector exhibits a favorable trade-off between these two parameters (compared to conventional detectors). It has also been predicted that the grating/hole (or slit) combination can lead to detectors having an improved signal-to-noise ratio [10, 11].

Arrays of holes in metal films also couple incident light to SPPs, in a manner analogous to a corrugated grating [8]. The momentum conservation equation for a two-dimensional (low perturbation) periodic structure on a metal surface is [8, 12]:

$$k^{SPP} = k_{||} + lG_x + mG_y \tag{11}$$

where $\boldsymbol{k}_{||}$ is the in-plane wavevector of the incident light, $\boldsymbol{G_x}$ and $\boldsymbol{G_y}$ are reciprocal lattice vectors of the periodic structure, and l, m are integers. This equation models approximately the coupling of light to SPPs on a two-dimensional hole array, although the holes are actually neglected along with their associated effects such as scattering and transmission. For example, in the case of a hexagonal or triangular lattice of holes of lattice constant a, illuminated at normal incidence ($\boldsymbol{k}_{||} = 0$) and neglecting all losses, the above yields the wavelengths λ_c for which strong coupling to SPPs on the array occurs:

$$\lambda_c = a\left[\frac{4}{3}\left(l^2 + lm + m^2\right)\right]^{\frac{-1}{2}}\left(\frac{\varepsilon_R \varepsilon_{r,d}}{\varepsilon_R - \varepsilon_{r,d}}\right)^{\frac{1}{2}} \tag{12}$$

In the above, $\varepsilon_{r,d}$ corresponds to the dielectric bounding the array on the illuminating side. If the structure is symmetric (same dielectric on both sides of the array), then these wavelengths also correspond to those at which transmission peaks occur. If the structure is asymmetric (different dielectrics bounding the array) then two sets of transmission peaks are observed.

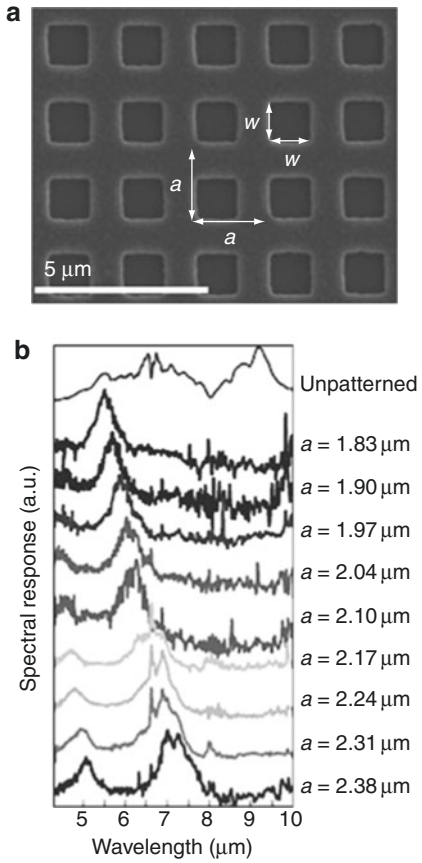
Hole arrays can be integrated with broadband detectors to introduce wavelength and polarization selectivity, and potentially, to enhance the responsivity [12, 13]. This is achieved by structuring the top or bottom contact of a detector into an array of sub-wavelength holes and then illuminating at normal incidence. The photocurrent is then largest at wavelengths for which extraordinary transmission occurs, which depend on the lattice constants of the array, e.g., via Eq. 12 for the case of a hexagonal array, and so are easily tunable. Polarization selectivity can be introduced by using elliptical or rectangular holes.

The responsivity can be enhanced for a thin detection layer by enhancing the absorptance (A) via mediation by SPPs. This is achieved by placing the detection layer close to the hole array such that SPPs propagating along the array overlap with the layer and are absorbed therein; thus the enhancement arises because SPPs propagate along the absorption region rather than through it perpendicularly.

Wavelength selectivity and responsivity enhancement are particularly attractive for mid-infrared detector arrays, which are typically broadband and have a low responsivity due to the low absorption of the detector materials used in this wavelength range. Applications such as multi-color (multi-spectral) night vision or medical imaging would benefit from inexpensive spectral filtering integrated at the pixel level. Figure 5a shows a square array (lattice constant a) of square holes (dimensions w) patterned onto the top contact (150-nm thick Ag) of an InAs mid-infrared quantum dot detector [13]. Figure 5b shows the measured spectral response of such detectors as a function of a, compared to an unpatterned control detector (upper trace). The peak wavelength response is observed to range from $\lambda_0 = 5.5$–7.2 μm as the lattice constant ranges from $a = 1.83$–2.38 μm.

Detectors Incorporating Nanoparticles

Small metal particles exhibit resonant responses under optical excitation, characteristic of particle



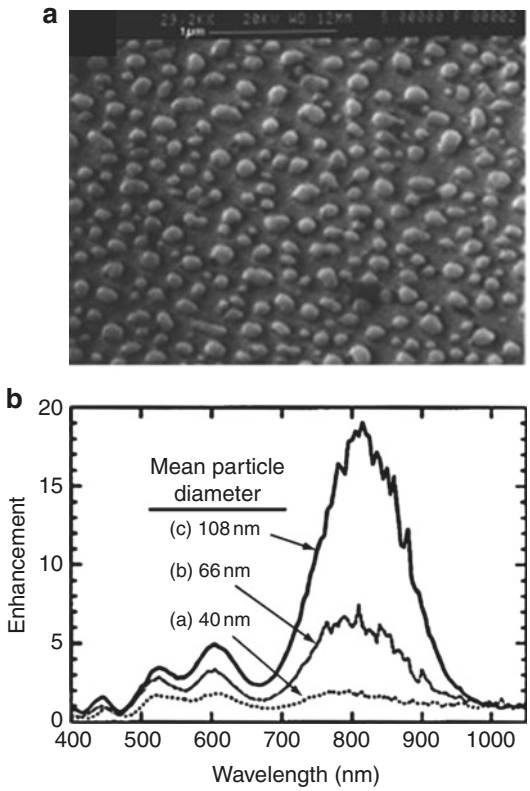
Surface Plasmon-Polariton-Based Detectors, Fig. 5 (**a**) Image of a square lattice of square holes; the lattice constant is a and the hole size is w. (**b**) Spectral response of quantum dot detectors each bearing a hole array of the indicated lattice constant; $w = 0.6a$ for all detectors (Adapted from Ref. [13]. © 2009 American Institute of Physics)

shape, size, and composition, the dielectric environment in which they find themselves, and the wavelength of illumination [14]. Resonances occur when the electron charge density oscillates coherently with the illuminating optical (electrical) field. The fundamental resonant mode of, e.g., a spherical metal nanoparticle (small compared to the illuminating wavelength) is dipolar with densities of opposite charge forming at opposite spherical caps of the particle along the polarization of the illuminating electric field. The wavelength of the dipolar resonance red-shifts as the radius of the particle increases or as the index of the background increases. If the particle is large enough then higher-order resonant modes can exist (e.g., quadrupolar). On resonance, the electric field near the particle (in the dielectric) is strongly enhanced compared to the illuminating field. Resonant features appear in the measured extinction and scattering spectra of metal nanoparticles, in correlation with the plasmon resonances that are excited thereon.

Metal nanoparticles can be integrated with a detector most readily via deposition onto the surface through which light penetrates into the detector [15, 16]. As light propagates through, resonances can be excited, and strong scattering can occur, leading to improved detector performance.

Figure 6a shows a scanning electron microscope image of Ag-island nanoparticles on a 30-nm thick LiF layer on a Si-on-insulator (SOI) detector [15]. These approximately hemispherical particles are 108 nm across, on average. They were formed by first depositing a thin Ag film, then annealing the film to induce islandization. This process has the advantage of being simple but it generates particles that are non-uniform in shape and size, albeit, with a controllable average size. Detectors were formed in the thin Si slab (165 nm thick) above the insulator but beneath the LiF (and the nanoparticles) as pn junction detectors. Figure 6b shows measured spectral responses of detectors coated with Ag islands similar to those of Fig. 6a as a function of average island size. The enhancement is defined as the ratio of the photocurrent from a detector with Ag islands to the photocurrent from a detector without islands. A strongly enhanced photocurrent ($\sim 20\times$) is noted at $\lambda_0 = 800$ nm in the case of islands that are 108 nm in average size. The enhancement is attributed to the large scattering cross-section of the islands, particularly to mediation by the islands whereby dipole resonances excited thereon radiate into guided modes of the thin underlying Si slab. These modes propagate longitudinally and thus are absorbed more effectively by the detector compared to perpendicularly incident light.

Chemically synthesized (colloidal) nanoparticles can also be deposited and attached to detector surfaces. Spectral responses of



Surface Plasmon-Polariton-Based Detectors, Fig. 6 (**a**) Scanning electron microscope image of Ag-island nanoparticles on LiF-coated Si-on-insulator (*SOI*) detector. The particles are 108 nm across, on average. (**b**) Spectral response of pn junction SOI photodetectors coated with Ag islands similar to those of (**a**) as a function of average island size. The enhancement is defined as the ratio of the photocurrent from a detector with Ag islands to the photocurrent from a detector without the islands (Adapted from Ref. [15]. © 1998 American Institute of Physics)

detectors coated with 50, 80, and 100 nm diameter spherical Au nanoparticles showed strong photocurrent enhancement at wavelengths where peaks were measured in the extinction spectra of the particles [16]. In contrast to the devices of [15], the enhanced photocurrent was attributed to enhanced absorption in the Si regions in contact with resonating nanoparticles, around which the resonating electric fields are significantly enhanced compared to the illuminating fields.

The effects induced by metal nanoparticles, i.e., increased scattering into detectors and increased detector absorption on resonance, are useful to increase the absorptance of detectors, particularly those based on a thin absorption layer, or at wavelengths where the absorption is low (e.g., near the bandgap of an indirect semiconductor). These effects can potentially improve the efficiency of, e.g., thin-film Si solar cells [17].

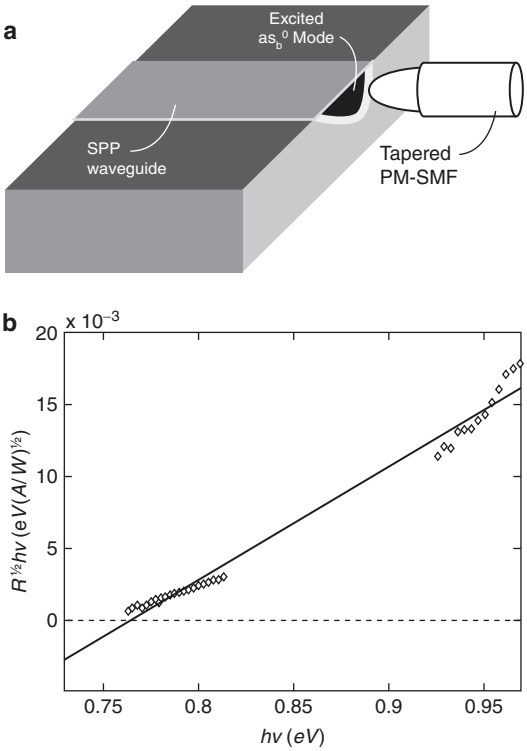
Waveguide Detectors

The single-interface (Fig. 1) is one example of an SPP waveguide; other popular 1-D geometries include the thin metal slab bounded by dielectrics and the thin dielectric slab bounded by metals [2]. The latter are often termed the IMI and MIM, respectively (I – insulator, M – metal). The IMI and MIM support super-modes formed from the coupling of single-interface SPPs through the thin intervening layer. The structures can be designed such that they support super-modes that are low loss but weakly confined (symmetric IMI) or strongly confined but high loss (MIM), and thus that are at opposite ends of the confinement-attenuation trade-off [2].

SPP waveguides can be formed into SPP detectors and conveniently integrated with other plasmonic or photonic waveguide structures. SPP waveguide detectors based on absorption in, e.g., organics [18], semiconductors [19], or metals [20] have been reported.

The detector described in [18] consists of a thick Ag film on which SPPs are excited by illuminating a slit in the film. The SPPs then propagate along the Ag film to the detector section, which consists of a pn junction fabricated from organic polymers directly on the Ag film. The pn junction is covered by another Ag film, thereby creating an MIM-type detector. SPPs incident from the Ag film onto this MIM detector excite SPPs therein, which are absorbed by the pn junction creating EHPs that are collected as the photocurrent. Current maps correlating the photocurrent with the excitation of SPPs at $\lambda_0 = 632.8$ nm along the Ag film are reported.

The structure described in [19] consists of an MIM (Au/HSQ/Au) waveguide with slits, fabricated on a GaAs wafer. A slit in the top Au layer is used to couple incident light to SPPs propagating



Surface Plasmon-Polariton-Based Detectors, Fig. 7 (**a**) Sketch of an SPP waveguide detector consisting of a metal stripe forming a Schottky contact on Si, excited (in the as_b^0 mode) via butt-coupling to a tapered polarization-maintaining single mode fiber (*PM-SMF*). (**b**) Spectral response of a Au on n-Si detector (as in (**a**)), plotted in Fowler form (Adapted from n. © 2010 Optical Society of America)

in the MIM. A slit in the bottom Au layer, placed a short distance away from the input, couples the SPP into radiation directed into the GaAs wafer, which is absorbed therein creating EHPs that are collected as the photocurrent. Spectral responses and photocurrent maps demonstrating the operation of the detector are reported.

Figure 7a gives a sketch of a SPP waveguide detector consisting of a thin narrow metal stripe on semiconductor (with air on top) [20]. The thin metal stripe is similar to the asymmetric IMI, but it supports SPPs with additional confinement along the lateral dimension (stripe width). Detectors have been fabricated as Au or Al stripes on n-Si forming Schottky contacts thereon [20]. The as_b^0 mode, localized to the bottom metal-Si interface, is excited via butt-coupling to a tapered polarization-maintaining single mode fiber (PM-SMF). This mode propagates along the stripe with strong absorption, creating hot carriers therein, some of which may cross the Schottky barrier and be collected as the photocurrent; thus detection occurs via IPE. Figure 7b gives the spectral response of a Au on n-Si detector, plotted in Fowler form. The intercept with the abscissa yields the cutoff photon energy (~0.765 eV), corresponding in this case to a cutoff wavelength of $\lambda_0 \sim 1{,}620$ nm. Responsivities up to 1 mA/W were reported with this detector scheme.

Concluding Remarks

SPP detectors, or SPP-enhanced detectors, typically combine a metallic structure that supports SPPs, such as a planar or grating structure, metallic nanoparticles, or holes in metal films, with a semiconductor detector structure such as a Schottky or a pn junction. Properties inherent to SPPs are exploited to convey additional characteristics to a detector such as polarization, angular or spectral selectivity, or to enhance the absorptance (A) of the detector. The involvement of SPPs has led to detectors with improved performance and greater functionality.

References

1. Maier, M.A.: Plasmonics: Fundamentals and Applications. Springer, New York (2007)
2. Berini, P.: Figures of merit for surface plasmon waveguides. Opt. Express **14**, 13030–13042 (2006)
3. Sze, S.M.: Physics of Semiconductor Devices. Wiley, New York (1981)
4. Scales, C., Berini, P.: Thin-film Schottky barrier photodetector models. IEEE J. Quantum Electron. **46**, 633–643 (2010)
5. Jestl, M., Maran, I., Kock, A., Beinstingl, W., Gornik, E.: Polarization-sensitive surface plasmon Schottky detectors. Opt. Lett. **14**, 719–721 (1989)
6. Brueck, S.R.J., Diadiuk, V., Jones, T., Lenth, W.: Enhanced quantum efficiency internal photoemission detectors by grating coupling to surface-plasma waves. Appl. Phys. Lett. **46**, 915–917 (1985)
7. Daboo, C., Baird, M.J., Hughes, H.P., Apsley, N., Emeny, M.T.: Improved surface plasmon enhanced photodetection at an Au-GaAs Schottky junction

using a novel molecular beam epitaxy grown Otto coupling structure. Thin Solid Films **201**, 9–27 (1991)
8. Genet, G., Ebbesen, T.W.: Light in tiny holes. Nature **445**, 39–46 (2007)
9. Ishi, T., Fujikata, J., Makita, K., Baba, T., Ohashi, K.: Si Nano-photodiode with a surface plasmon antanna. Jpn. J. Appl. Phys. **44**, L364–L366 (2005)
10. Yu, Z., Veronis, G., Fan, S., Brongersma, M.L.: Design of midinfrared photodetectors enhanced by surface plasmons on grating structures. Appl. Phys. Lett. **89**, 151116 (2006)
11. Bhat, R.D.R., Panoiu, N.C., Brueck, S.J.R., Osgood, R.M.: Enhancing the signal-to-noise ratio of an infrared photodetector with a circular metal grating. Opt. Express **16**, 4588–4596 (2008)
12. Chang, C.Y., Chang, H.Y., Chen, C.Y., Tsai, M.W., Chang, Y.T., Lee, S.C., Tang, S.F.: Wavelength selective quantum dot infrared photodetector with periodic metal hole arrays. Appl. Phys. Lett. **91**, 163107 (2007)
13. Rosenburg, J., Shenoi, R.V., Vandervelde, T.E., Krishna, S., Painter, O.: A multispectral and polarization-selective surface-plasmon resonant midinfrared detector. Appl. Phys. Lett. **95**, 161101 (2009)
14. Kelly, K.L., Coronado, E., Zhao, L.L., Schatz, G.C.: The optical properties of metal nanoparticles: the influence of size, shape, and dielectric environment. J. Phys. Chem. B **107**, 668–677 (2003)
15. Stuart, H.R., Hall, D.G.: Island size effects in nanoparticle-enhanced photodetectors. Appl. Phys. Lett. **73**, 3815–3817 (1998)
16. Schaadt, D.M., Feng, B., Yu, E.T.: Enhanced semiconductor optical absorption via surface plasmon excitation in metal nanoparticles. Appl. Phys. Lett. **86**, 063106 (2005)
17. Atwater, H.A., Polman, A.: Plasmonics for improved photovoltaic devices. Nat. Mater. **9**, 205–213 (2010)
18. Ditlbacher, H., Aussenegg, F.R., Krenn, J.R., Lamprecht, B., Jakopic, G., Leising, G.: Organic diodes as monolithocally integrated surface plasmon polariton detectors. Appl. Phys. Lett. **89**, 161101 (2006)
19. Neutens, P., Van Dorpe, P., De Vlaminck, I., Lagae, L., Borghs, G.: Electrical detection of confined gap plasmons in metal-insulator-metal waveguides. Nat. Photonics **3**, 283–286 (2009)
20. Akbari, A., Tait, R.N., Berini, P.: Surface plasmon waveguide Schottky detector. Opt. Express **18**, 8505–8514 (2010)

Surface Properties

► Nanostructures for Surface Functionalization and Surface Properties

Surface Science

► Influence of Defects in Photocatalysis

Surface Tension and Chemical Potential at Nanoscale

► Surface Energy and Chemical Potential at Nanoscale

Surface Tension Effects of Nanostructures

Ya-Pu Zhao and Feng-Chao Wang
State Key Laboratory of Nonlinear Mechanics (LNM), Institute of Mechanics, Chinese Academy of Sciences, Beijing, China

Synonyms

Interface excess free energy; Surface energy density

Definition

The surface tension is the reversible work per unit area needed to elastically stretch/compress a preexisting surface. In other words, it is a property of the surface that characterizes the resistance to the external force. Surface tension has the dimension of force per unit length or of energy per unit area. For nanostructures with a large surface to volume ratio, the surface tension effects dominate the size-dependent mechanical properties.

Overview

Nanostructures have a sizable surface to volume ratio as compared to bulk materials, which leads

its mechanical properties to be quite different from those of bulk materials [1]. The size-dependent mechanical properties of nanostructures are generally attributed to the surface effects, in which surface tension is one of the most predominant factors.

In the framework of the thermodynamics, the surface (interface) between two phases is first modeled as a bidimensional geometrical boundary of zero thickness, that is, the mathematical surface, as shown in Fig. 1a. The physical quantities between the two phases are discontinuous, and the surface (interfacial) tension is a jump in stress. This idealization was then extended by Gibbs [2], who proposed that the physical quantities should undergo a smooth transition at the surface (interface) while the surface is still modeled as an infinitesimal thin boundary layer, as shown in Fig. 1b. In order to preserve the total physical properties of the system, the excess physical properties have to be assigned to the geometrical surface. The surface tension, which is defined based on the interface excess free energy, can be written as

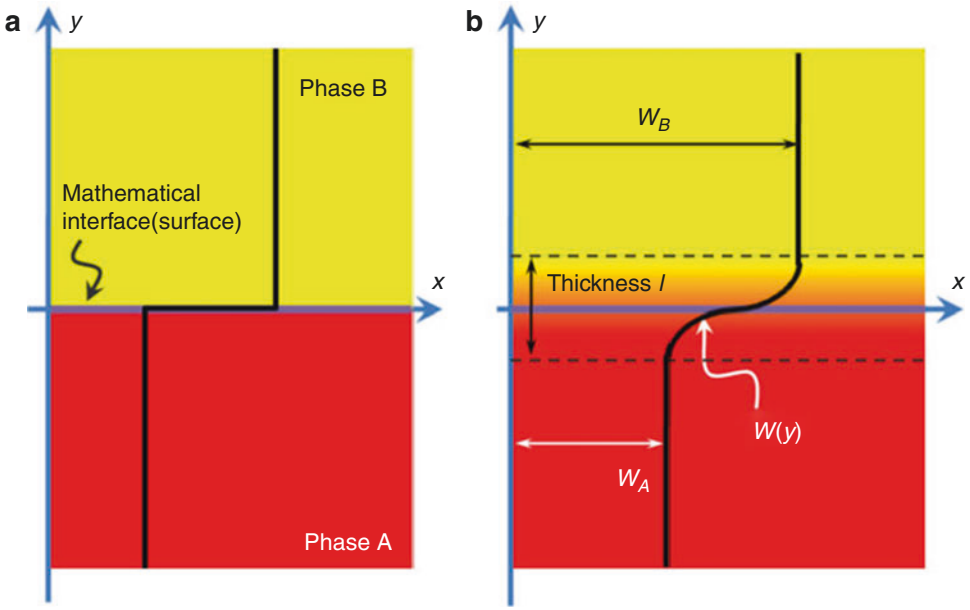
$$\gamma = \int_0^{\infty} [w(y) - W_A]dy + \int_{-\infty}^{0} [w(y) - W_B]dy \quad (1)$$

where w is the free energy distribution of the actual surface, w_A and w_B are free energy in the two phases of A and B.

In some other theoretical models, the surface is treated as an extended interfacial region with a nonzero thickness. According to Cahn–Hilliard theory [3], the surface tension can be derived as

$$\gamma = N_V \int_{-\infty}^{\infty} \left[w_0(c) + k(\mathrm{d}c/\mathrm{d}y)^2 - c\mu_B - (1-\mathrm{c})\mu_A \right] \mathrm{d}y, \quad (2)$$

in which N_V is the number of atoms per unit volume, c is one of the intensive scalar properties, such as composition or density, $w_0(c)$ is the free energy per atom of a solution of uniform composition c, k reflects the crystal symmetry, μ_A and μ_B are the chemical potentials per atom in the



Surface Tension Effects of Nanostructures, Fig. 1 An illustration for the surface (interface) models. (**a**) The mathematical surface of zero thickness, the physical quantities are discontinuous. (**b**) Gibbs's surface with an infinitesimal thickness, the physical quantities are continuous. Cahn–Hilliard model for surface with a finite thickness is also illustrated

A or B phase. The surface (interfacial) thickness can be obtained by

$$l = 2\Delta c_e \sqrt{\frac{k}{\Delta w_{\max}}}, \tag{3}$$

where $2\Delta c_e = c_B - c_A$ is the difference of the uniform composition in the two phases, $\Delta w_{\max}$ is the maximum of the free energy referred to a standard state of an equilibrium mixture. Cahn–Hilliard model gives a finite thickness through Eq. 3, which has been used in many applications [4].

From the standpoint of molecular theory, the surface tension effects arise due to the difference of the atomic interactions in the bulk and on the surface. Atoms are energetically favorable to be surrounded by others. At the surface, the atoms are only partially surrounded by others and the number of the adjacent atoms is smaller than in the bulk. Thus the atoms at the surface are energetically unfavorable. If an atom moves from the bulk to the surface, work has to be done. With this view, the surface tension can be interpreted as the energy required to bring atoms from the bulk to the surface [5]. Therefore the term "surface energy density" is often used to when the surface tension is referred to. For the surface of solid, Gibbs pointed out that surface tension and surface energy density are not identical. The surface tension is the reversible work per unit area needed to elastically stretch/compress a preexisting surface. The surface energy density is the reversible work per unit area needed to create a new surface. The surface tension can be positive or negative, while the surface energy density is usually positive. The surface tension for liquid surface is a property that characterizes its resistance to the external force, which is identical to the surface energy density.

Basic Methodology

Since the surface to volume ratio increases as the dimension scale decreases, surface tension effects of nanostructures can be overwhelming. In the absence of external loading, the surface tension effects would induce a residual stress field in bulk materials. The relations between surface stress and surface tension for small deformations can be described by the Shuttleworth-Herring equation [6, 7],

$$\sigma_{ij} = \gamma\delta_{ij} + \frac{\partial\gamma}{\partial\varepsilon_{ij}}, \tag{4}$$

where γ is the surface tension, δ_{ij} is the Kronecher delta, σ_{ij} and ε_{ij} are the surface stress tensor and the surface strain tensor, respectively. The Shuttleworth-Herring equation interprets that the difference between the surface stress and surface tension is equal to the variation of surface tension with respect to the elastic strain of the surface.

With the development of computational materials science, molecular dynamics (MD) simulations are wildly performed to investigate the surface tension effects on the mechanical properties of nanostructures. Especially for the surface elastic constant, atomistic simulation is almost the only way to get them up to now. According to the Gibbs's definition, the surface tension of a solid is given by $\gamma = (E_S - nE_B)/A_0$, where A_0 is the total area of the surface considered, E_S is the total energy of a n-layer slab and E_B is the bulk energy per layer of an infinite solid. In cases where the surfaces of the slab are polar, the electrostatic energy of the slab contains an energy contribution, E_{pol}, proportional to the substrate thickness and the surface energy needs corrections. Thus there is an alternative way to calculate the surface tension $\gamma = \left(E_S - nE_B - E_{\text{pol}}\right)/A_0$, which does not rely on an exact knowledge of the lattice or polar energy. MD simulations have identified that the surface tension induced surface relaxation is proved to be a dominate factor of the size-dependent mechanical properties. When the surface stress is negative, the surface relaxation is inward; otherwise, the relaxation is outward.

Surface stress has been used as an effective molecular recognition mechanism. Surface stresses due to DNA hybridization and receptor-ligand binding induce the deflection of a cantilever sensor [8]. The curvature of bending beam

under a surface stress is governed by Stoney's formula [9]. Stoney's formula serves as a cornerstone for curvature-based analysis and a technique for the measurement of surface stress, which is given as follows as a general form for a film/substrate system,

$$\sigma = \frac{Et_s^2 f}{6(1-v)} \tag{5}$$

in which E is the effective Young's modulus, v is Poisson's ratio of the sensor material, ts is the substrate thickness, $f = 3\Delta z/2\ L^2$ is the sensor curvature, L is the length and Δz is the deflection. The applicability of the above Stoney's formula relies on several assumptions, which are well summarized as the following six: (1) both the film and substrate thicknesses are small compared to the lateral dimensions; (2) the film thickness is much less than the substrate thickness; (3) the substrate material is homogeneous, isotropic, and linearly elastic, and the film material is isotropic; (4) edge effect near the periphery of the substrate are inconsequential and all physical quantities are invariant under change in position parallel to the interface; (5) all stress components in the thickness direction vanish throughout the material; and (6) the strains and rotations are infinitesimally small. However, the one or several of above six assumptions can be easily violated in reality, which is to say that the Stoney's formula needs to be revised to fit in the real applications.

To summarize for the solid cases, the behaviors of nanostructures can be affected significantly by either of the two distinct parameters, surface tension and surface stress. The relation between the two parameters can be obtained by the Shuttleworth-Herring equation. For the surface stress induced deflection of cantilever sensors, the curvature is by Stoney's formula. MD simulation is helpful to understand the surface effects of nanostructures and partial results are comparable to the experiments.

For liquid droplets in contact with the surface of a nanostructure, surface tension is responsible for the shape of liquid droplets in the equilibrium state, as well as the dynamics response in wetting and dewetting. There are several characteristic time which are related to the surface tension effects, listed in Table 1. Dimensionless number related to the surface tension effects are listed in Table 2. Surface tension is dependent on temperature T, concentration of surfactants c, and the electric field V. The gradient of surface tension caused by these factors can be described by [10]

$$\mathrm{d}\gamma = \frac{\partial \gamma}{\partial T} dT + \frac{\partial \gamma}{\partial c} dc + \frac{\partial \gamma}{\partial V} dV. \tag{6}$$

To the first order, the dependency of the surface tension on temperature is given by Guggenheim–Katayama formula with power index $n = 1$, $\gamma = \gamma_0(1 - T/T_c)$. Here γ_0 is a constant for each liquid and T_c is the critical temperature. For the dependency of the surface tension on concentration c, the surface tension can be expressed as a linear function of the concentration, $\gamma = \gamma_0[1 + \beta(c - c_0)]$. The solid–liquid surface tension can be changed by applying a voltage V, $\gamma_{SL} = \gamma_{SL}^0 - \frac{\varepsilon_0 \varepsilon_D}{2d} V^2$, where γ_{SL}^0 is the solid–liquid surface tension in the absence of the applied voltage, ε_0 is the permittivity of vacuum, ε_D is the relative permittivity of the dielectric layer

Surface Tension Effects of Nanostructures, Table 1 Characteristic time related to the surface tension effects

Name	Expression	Meaning
Capillary characteristic time	$t_c = \sqrt{m/\gamma_{LV}}$	Characteristic time derived from the droplet mass and the liquid–vapor surface tension
Lord Rayleigh's period	$t_p = \frac{\pi}{4}\sqrt{\rho d_0^3/\gamma_{LV}}$	The period of a free droplet in free oscillation
Lord Rayleigh's characteristic time	$t_{\widetilde{R}} \sqrt{\rho l^3/\gamma_{LV}}$	Characteristic time of droplet dynamics
Viscous characteristic time	$t_{\widetilde{vis}} \eta l/\gamma_{LV}$	Characteristic time related to the liquid viscosity

m mass of the droplet, d_0 diameter, ρ density, η viscosity, l characteristic length

Surface Tension Effects of Nanostructures, Table 2 Dimensionless number related to the surface tension effects

Name	Expression	Meaning
Adhesion number	$N_a = \cos\theta_a - \cos\theta_r$	θa and θ_r are the advancing and receding angle
Bond number	$\mathrm{Bo} = \rho g l^2/\gamma_{LV}$	The capillary length $l_c = \sqrt{\gamma_{LV}/(\rho g)}$ can be determined
Capillary number	$\mathrm{Ca} = \eta v/\gamma_{LV}$	$\mathrm{Ca} = \mathrm{On} \cdot \sqrt{\mathrm{We}}$; the capillary velocity $v = \gamma_{LV}/\eta$ can be obtained
Deborah number	$\mathrm{De} = t/t_R = t/\sqrt{\rho l^3/\gamma_{LV}}$	t is the relaxation time, Lord Rayleigh characteristic time can be derived
Elasto-capillary number	$\mathrm{Ec} = t\gamma_{LV}/(\eta l)$	t is the relaxation time, viscous characteristic time can be obtained
Electrowetting number	$\eta_e = \varepsilon_0\varepsilon_D V^2/(2d\gamma_{LV})$	The ratio of the electrostatic energy to the surface tension
Laplace number	$\mathrm{La} = (1/\mathrm{On})^2$	See Ohnersoge number
Marangoni number	$\mathrm{Ma} = \frac{d\gamma_{LV}}{dT}\frac{L\cdot\Delta T}{\eta\alpha}$	Thermal surface tension force divided by viscous force
Ohnersorge number	$\mathrm{On} = \eta/\sqrt{\rho l \gamma_{LV}}$	$\frac{\mathrm{Viscousforce}}{\sqrt{\mathrm{Inertiaforce}\times\mathrm{Surfacetension}}}$
Weber number	$\mathrm{We} = \rho \mathrm{v}^2 l/\gamma_{LV}$	The inertia force compared to its surface tension

g gravitational acceleration, α thermal diffusivity, ΔT temperature difference

with a thickness d separating the bottom electrode from the liquid.

The wetting properties of a solid surface can be described by the introduction of the contact angle, which is defined as the angle at which the liquid–vapor interface meets the solid surface. The contact angle is affected by various factors, including the surface tension, the line tension, the applied voltage as well as the molecular interactions between the liquid and solid surfaces. It is proposed that the dependence of the contact angle on these factors can be represented by the generalized Young's equation, which has a form of

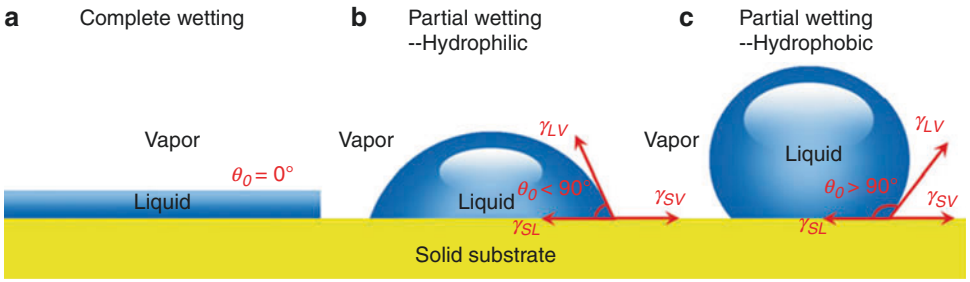
$$\cos\theta = \cos\theta_0 - \frac{\tau}{\gamma_{LV}R} + \frac{\varepsilon_0\varepsilon_D}{2d\gamma_{LV}}V^2 + \frac{A}{12\pi h^2\gamma_{LV}}, \quad (7)$$

where

$$\cos\theta_0 = \frac{\gamma_{SV} - \gamma_{SL}}{\gamma_{LV}} \quad (8)$$

is the classical Young's equation, θ_0 is the equilibrium contact angle (also the Young contact angle); The subscripts S, L, and V denote solid, liquid, and vapor, respectively. The second term on the right-hand side of Eq. 7 is related to the line tension. τ is the line tension and R is the radius of the contact area. For a more generalized case, R can be replaced by $1/\kappa$, in which κ is the geodesic curvature of the triple contact line. $\eta_e = \frac{\varepsilon_0\varepsilon_D}{2d_{LVd}}V^2$ is defined as the dimensionless electrowetting number. The last term in Eq. 7 measures the strength of the effective interaction energy compared to surface tension. The interaction energy is related to the disjoining pressure $\Pi(h)$, which can be obtained by $W(h) = \int_h^\infty \Pi(h')dh'$, where h is the film thickness and A is the Hamaker constant [11, 12]. As discussed in the following paragraphs, each term would be illustrated in detail.

If only the classical Young's equation is referred to, the wetting properties of the surface can be obtained directly if the three tensions are known. The spreading parameter S determines the type of spreading, which is defined as $S = \gamma_{SV} - (\gamma_{SL} + \gamma_{LV})$. If $S > 0$, the liquid spreads on the solid surface and forms a macroscopic liquid layer covers the whole solid surface; it is the case for complete wetting. If $S < 0$, the contact angle $0° < \theta_0 < 180°$, which means the liquid forms a droplet with a finite contact angle; it is the case for speak of partial wetting. The wettability of the solid surface could be distinguished by the contact angle θ_0, as shown in Fig. 2. If



Surface Tension Effects of Nanostructures, Fig. 2 Wetting properties of the surface

$0° \leq \theta_0 < 90°$, the solid substrate is hydrophilic. If $90° < \theta_0 \leq 180°$, the solid substrate is hydrophobic. Especially, if $150° < \theta_0 \leq 180°$, the solid substrate is superhydrophobic.

The physics behind wettability is that, the solid surfaces have been divided into high energy and low-energy types [11, 12]. The relative energy of a solid has to do with the bulk nature of the solid itself. (1) High-energy surfaces such as metals, glasses, and ceramics are bound by the strong chemical bonds, for example, covalent, ionic, or metallic, for which the chemical binding energy E_{binding} is of the order of 1 eV. The solid–liquid interface tension is given by $\gamma_{SV} \approx E_{binding}/a^2 \sim 0.5 - 5\text{N/m},$, in which a^2 is the effective area per molecule. Most high-energy surfaces are hydrophilic, some can permit complete wetting. (2) For low-energy surfaces, such as weak molecular crystals (bound by van der Waals forces or in some special cases, by hydrogen bonds), the chemical binding energy is of the order of k_BT. In this category, the surface tension is $\gamma_{SV} \approx k_BT/a^2 \sim 0.01 - 0.05\text{N/m}$. Depending on the type of liquid chosen, low-energy surfaces can be either hydrophobic or hydrophilic.

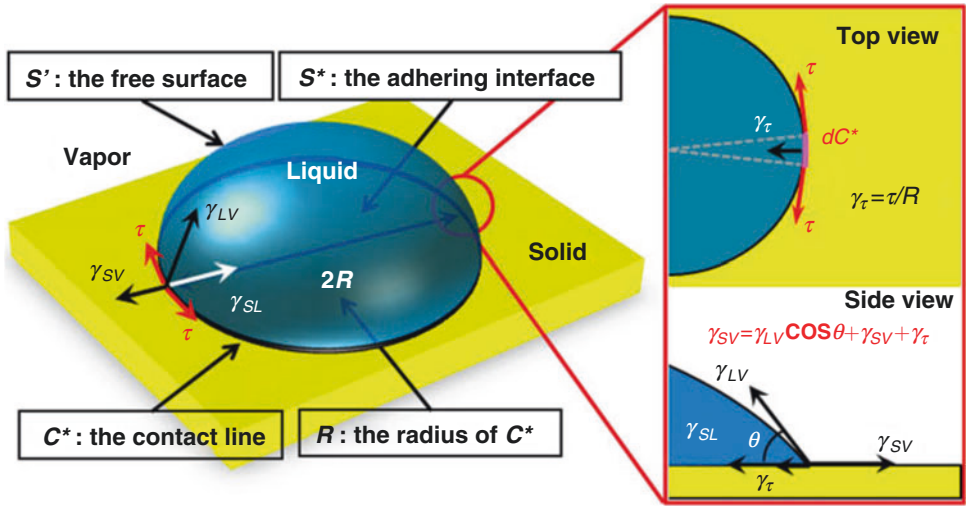
When the liquid droplet comes to the nanoscale, the classical Young's equation seems to be not applicable, since it has been derived for a triple line without consideration of the interactions near the triple contact line. The molecules close to the triple line experience a different set of interactions than at the interface [10]. To take into account this effect, the "line tension" term has been introduced in the generalized Young's equation. A sketch of line tension is shown in Fig. 3.

Line tension was first introduced by Gibbs [2]. The line tension τ was introduced as an analogue of the surface tension:

$$F = \gamma_{LV}\int_{S'} \mathrm{d}S' + (\gamma_{LV} - W)\int_{S'} \mathrm{d}S^* + \tau\int_{C*} \mathrm{d}C^*, \tag{9}$$

where F, S', S^*, and C^* are the free energy, the free surface, the adhering interface, and the contact line of the droplet. $w = \gamma_{LV} + \gamma_{SL} - \gamma_{SV}$ is the adhesion potential. The line tension, depending on the radius R of the contact line, should be connected to the classical Young's equation to include the interactions near the triple contact line. In the three-phase equilibrium systems when R decreases, the contact angle θ will increase at $\tau > 0$ and will decrease at $\tau < 0$. However, in accordance to the mechanical equilibrium stability conditions, while the surface tension can only be positive, the line tension can have either positive or negative values. The characteristic length of this problem $l = |\tau|/\gamma_{LV}$ ($|\tau| < 10^{-9}$ N, $\gamma_{LV} \sim 0.1$ N/m for water) ranges from 10^{-8} to 10^{-6} m, which means small droplet (with typical dimension of $|l|$) should appreciate the line tension effect. The line tension can be indirectly measured through the contact angle, $\tau \approx 4\delta\sqrt{\gamma_{SV}\gamma_{LV}}\cot\theta$, in which δ denotes the average distance between liquid and solid molecules. Thus the line tension is negative for an obtuse contact angle, while it is positive for an acute contact angle.

In the presence of a charged interface, which can be achieved by applying a direct or

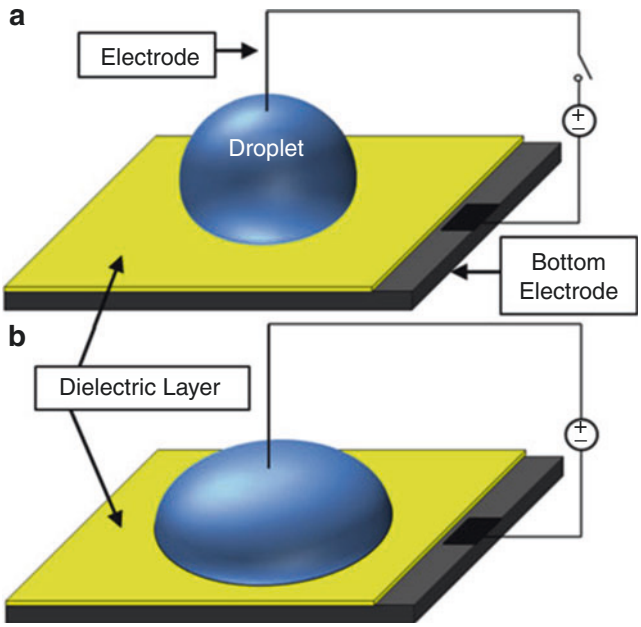


Surface Tension Effects of Nanostructures, Fig. 3 Illustration of the line tension τ

alternating-current electric field, the wetting properties of solid surface will be modified. The physics describing the electric forces on interfaces of conducting liquids and on triple contact lines is called "electrowetting" [10]. In the year of 1875, Gabriel Lippmann observed the capillary depression of mercury in contact with an electrolyte solution could be varied by applying a voltage between the mercury and electrolyte. This phenomenon is called electrocapillarity, which is the basis of modern electrowetting. Then, the idea was developed to isolate the liquid droplet from the substrate using a dielectric layer in order to avoid electrolysis. This concept has subsequently become known as electrowetting on dielectric (EWOD) and involves applying a voltage to modify the wetting behavior of a liquid in contact with a hydrophobic, insulated electrode. When an electric field was applied to the system (as shown in Fig. 4), electric charges gather at the interface between the conductive electrodes and the dielectric material; the surface becomes increasingly hydrophilic (wettable), the contact angle is reduced and the contact line moves. The change in contact angle over the buried electrodes can be evaluated by the Lippmann–Young equation. The parabolic variation of cos θ in the Lippmann–Young equation is $\cos\theta = \cos\theta_0 + \varepsilon_0\varepsilon_D V^2/(2d\gamma_{LV})$ only applicable when the applied voltage lies below a threshold value. If the voltage is increased above this threshold, then contact angle saturation starts to occur and cos θ eventually becomes independent of the applied voltage.

The physics of why a liquid film wets or dewets is found in the derivative of the effective interfacial potential $W = \gamma_{LV} + \gamma_{SL} - \gamma_{SV}$ with respect to film thickness h, called the disjoining pressure $\Pi(h)$ [11, 12]. $\Pi(h) = \mathrm{d}W(h)/\mathrm{d}h$, where the surface area remains constant in the derivative. The effective interface potential $W(h)$, which arises from the interaction energies of molecules in a film being different from that in the bulk, is the excess free energy per unit area of the film. If the interactions between the molecules in the film and the solid substrate are more attractive than the interactions between molecules in the bulk liquid, $W(h) > 0$. Consequently, a liquid film with a thickness in a range where $\Pi(h) > 0$ can lower its free energy by becoming thicker in some areas while thinning in others, that is, by dewetting. When $\Pi(h) < 0$, wetting or spreading occurs.

The van der Waals interaction $w(r) \propto 1/r^6$ includes all intermolecular dipole–dipole, dipole–induced dipole, and induced dipole–induced dipole interactions. Performing a volume integral over all molecules present in the



Surface Tension Effects of Nanostructures, Fig. 4 Illustration of EWOD. The external voltage is applied between a thin electrode (the Pt wire) and the bottom electrode (typically the indium-tin-oxide glass). Partially wetting liquid droplet in the absence of an applied voltage (**a**) and after the voltage is applied (**b**)

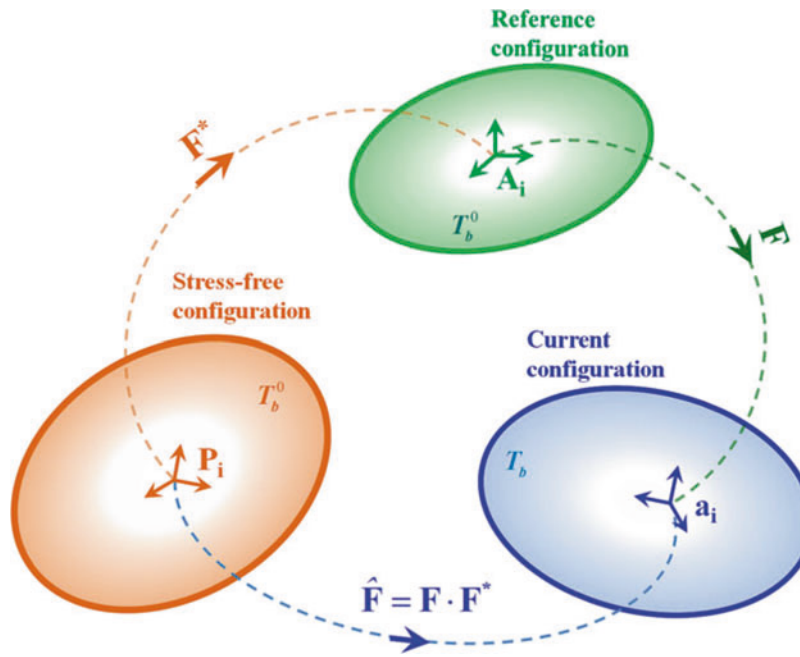
two half spaces bounding the film one finds a corresponding decay $\Pi(h) \sim A/6\pi h^3$ [11, 12], where the Hamaker constant A ($\sim 10^{-19}$ J) gives the amplitude of the interaction. In the "attractive" case, in which the layer tends to thin, $A < 0$. In the "repulsive" case, in which the layer tends to thicken, $A > 0$. Because of the disjoining pressure, there is a precursor film ahead of the nominal contact line of the liquid droplet [13]. The thickness of the precursor film can be defined by a molecular length [11, 12] $h_{PF} \sim \sqrt{A/6\pi\gamma_{LV}}$, which is on the order of several Å.

Key Research Findings

Elastic Models for the Nanostructures

Gurtin and Murdoch established the theoretical framework of the surface elasticity under the classical theory of membrane [14]. Recently, studies [15] have shown that, even in the case of infinitesimal deformations, one should distinguish between the reference and the current configurations; otherwise the out-plane terms of surface displacement gradient, associated with the surface tension, may sometimes be overlooked in the Eulerian descriptions, particularly for curved and rotated surfaces. By combining elastic models for surface and bulk, the size-dependent elastic and properties of nanomaterials have been investigated. Usually, in the absence of external mechanical or thermal loadings, the surfaces of a nanostructure will be subjected to residual surface stresses, and an elastic field in the bulk materials will be induced by such residual surface stresses induce from the point of view of equilibrium conditions. This self-equilibrium state without external loadings is usually chosen as the reference configuration, from which nanostructures will deform (see Fig. 5). That is to say, the bulk will deform from the residual stress states. However, in the prediction of elastic and thermoelastic properties of nanostructures, the elastic response of the bulk is usually described by classical Hooke's law, in which the aforementioned residual stress was neglected in the existing literatures.

Considering a bulk material with the surface properties aforementioned, the surface tension would induce a stress field in the bulk. According to Young–Laplace equation, the surface tension will result in a nonclassical boundary condition. The boundary condition together with the equations of classical elasticity forms a coupled system of field equations to determine the stress



Surface Tension Effects of Nanostructures, Fig. 5 The choice of the reference configuration: the stressed state

distribution. To solve a problem considering the surface properties, the surface model and the bulk model are established separately and using the Young–Laplace equation to bridge the two models together.

Here, an example named surface elasticity would be used to show how to establish a model combined the surface and bulk together. Assuming the surface to be isotropic and homogeneous, the constitutive relations of the surface in the Lagrangian description can be written as [15]

$$\mathbf{S}_s = \gamma_0^* \mathbf{I}_0 + \left(\gamma_0^* + \gamma_1^*\right)(\mathrm{tr}\mathbf{E}_s)\mathbf{I}_0 - \gamma_0^*\left(\overline{\nabla}_{os}\mathbf{u}_0\right) + \gamma_1 \mathbf{E}_s + \gamma_0^* \mathbf{F}_s^{(o)}, \qquad (10)$$

where $\mathbf{S}_s$ is the first kind Piola–Kirchhoff stress of the surface, $\mathbf{I}_0$ is the identity tensor on the tangent planes of the surface in the reference configuration; the constants γ_0^*, γ_1^*, and γ_1 are the surface tension and the surface Lame moduli; $\mathbf{E}_s$, $\overline{\nabla}_{os}\mathbf{u}_0$, and $\mathbf{F}_s^{(o)}$ denote, respectively, the surface strain tensor, the in-plane part of surface displacement gradient and the out-plane term of surface deformation gradient. In view of the importance of the linearization of the general constitutive equations, the linear elastic constitutive relations of the bulk with residual stresses can be written as follows:

$$\mathbf{S} = \mathbf{T}_R + \mathbf{u}\nabla \cdot \mathbf{T}_R + \lambda \mathrm{tr}(\mathbf{E})1 + 2\mu\mathbf{E}, \qquad (11)$$

where S is the first Piola–Kirchhoff stress, $\mathbf{T}_R$ is the residual stress in the reference configuration, $\mathbf{u}\nabla$ is the displacement gradient calculated from the reference configuration, $\mathbf{E}$ is the infinitesimal strain, and λ and μ are material elastic constants.

In the absence of external loading, surface tension will induce a compressive residual stress field in the bulk of the nanoplate and there may be self-equilibrium states which correspond to plate - self-buckling. The self-instability of nanoplates is investigated and the critical self-instability size of simplify supported rectangular nanoplates is proposed. The critical size for self-buckling is $b = \pi h\sqrt{\left(\alpha_A^{-2} + 1\right)Eh/\left[24\gamma_0^*(1-\nu^2)\right]}$, where b and h are the width and thickness of the nanoplate, respectively; E and ν denote the Young's modulus and Poisson's ratio, $\alpha_A = l/b$ is the aspect ratio [15].

Surface Tension Effects on the Mechanical Properties of Nanostructures

Many models are developed to relax one or some of the six above assumptions to extend Stoney's formula to a more generalized and realistic application. For example, the effects such as axial force, the damaged/nonideal interface effect and gradient stress, which violates one or several of the above assumptions, are analyzed and the extended/revised Stoney's formulas are given. Surface stress physically is a distributed one and this characteristic is not emphasized in many studies. The analysis by Zhang et al. shows that the Stoney's formula is obtained by assuming the influence of surface stress as a concentrated moment applied at the free end of a cantilever beam [16]. However, if the influence of surface stress is modeled as a distributed axial load and bending moment, the following nonlinear governing equation is obtained:

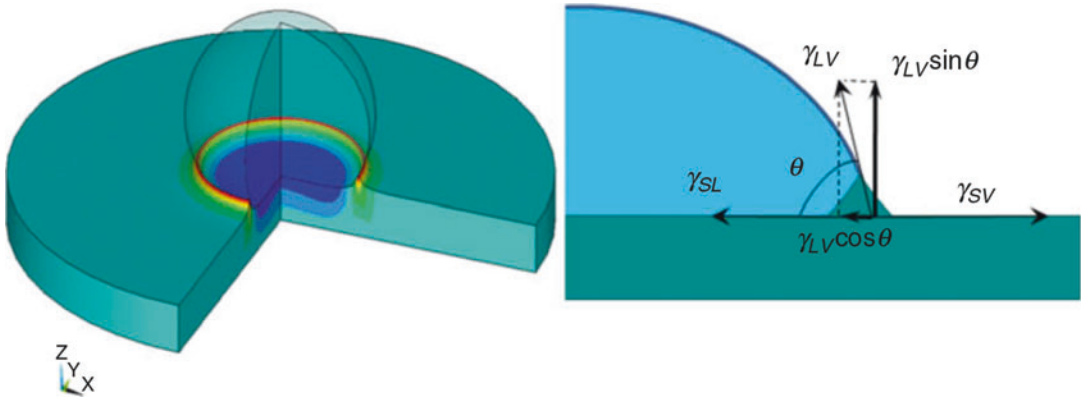
$$EI\frac{\mathrm{d}^4 z}{\mathrm{d}x^4} - \sigma b(L-x)\frac{\mathrm{d}^4 z}{\mathrm{d}x^2} + \sigma b\frac{\mathrm{d}z}{\mathrm{d}x} = 0, \qquad (12)$$

in which EI is the cantilever effective bending stiffness; b, L, and z are the beam width, length, and deflection, respectively. To solve the above nonlinear equation with the (given) boundary conditions at the two ends of the beam is a two-point-boundary value problem [16], which is rather difficult. The semi-analytical series solutions can more or less ease the difficulty of solving Eq. 12. One implication of the Stoney's formula is that because the beam curvature is constant, the beam deflection under a surface stress is an arc of a circle (or a parabola if the approximate curvature definition is used). There is no mechanism to guarantee such kind of deflection. In general the curvature of a beam under a surface stress is not a constant, which has been verified in the experiments.

The Stoney's formula has been used to explain recent experimental results, in which a hybrid device based on a microcantilever interfaced with bacteriorhodopsin (bR), undergoes controllable and reversible bending when the light-driven proton pump protein, bR, on the microcantilever surface is activated by visible light [17]. It should be pointed that the Young's modulus of a nanostructure is size-dependent, that is, it would be enhanced or softened with decreasing the size of the nanostructure, which is generally attributed to the surface effects. Surface tension is one of the most important factors that cause the size effects of the Young's modulus of a nanostructure. The surface tension can be introduced into mechanical model via energy method. Using the relation of energy equilibrium, the effective elastic modulus of nanobeams are dependent on the surface tension [18].

Surface Tension Effects Induced the Deformation of Nanostructures

The classical Young's equation describes the equilibrium of forces in the direction parallel to the solid surface, while the vertical component of liquid–vapor interfacial tension is ignored. That is, there is a net force $\gamma_{LV}\sin\theta$ acting normal to the smooth solid surface at the solid–liquid–vapor contact line. Due to the unbalance force, there will be a surface deformation, as shown in Fig. 6. Indeed, several decades ago, surface deformation of semi-infinite solid was theoretically analyzed with the physical assumption that the liquid–vapor has a finite thickness (maybe at the order of tens nanometers) and the liquid–vapor interfacial tension acts uniformly in this region. Their research suggests there is a wetting ridge at the three-phase contact line. Later, Shanahan and Carré used height dimensional analysis to characterize the maximum at the order of γ_{LV}/G, where G is the shear modulus of solid [19]. For the material widely used at that time were very rigid (at the order of at least 100 GPa), such a deformation is too small to be considered. However, to meet with the rapid development of microelectromechanical systems (MEMS) and nanoelectromechanical systems (NEMS), polydimethylsiloxane (PDMS) is widely fabricated to channels or membrane, which has at least one dimension on the order of submillimeters or even nanometers. The surface deformation might no longer be neglected. Moreover, it should be noted that whether the theoretical solution for semi-infinite case can be extended to the case of thin flexible membrane. Recently, Yu and Zhao considered the



Surface Tension Effects of Nanostructures, Fig. 6 Sketch of deformation of PDMS membrane induced by a water droplet

deformation of thin elastic membrane induced by sessile droplet and gave a theoretical solution correspondingly [20]. There are two important conclusions. The first is that there exists a saturated membrane thickness at the order of millimeter, if the solid is thicker than this, it can be taken regard as semi-infinite; otherwise, it is better to consider the effect of membrane thickness. The second is that if the membrane has a very low Young's modulus (for example, on the order of MPa or much less), the effect of membrane thickness will become significant.

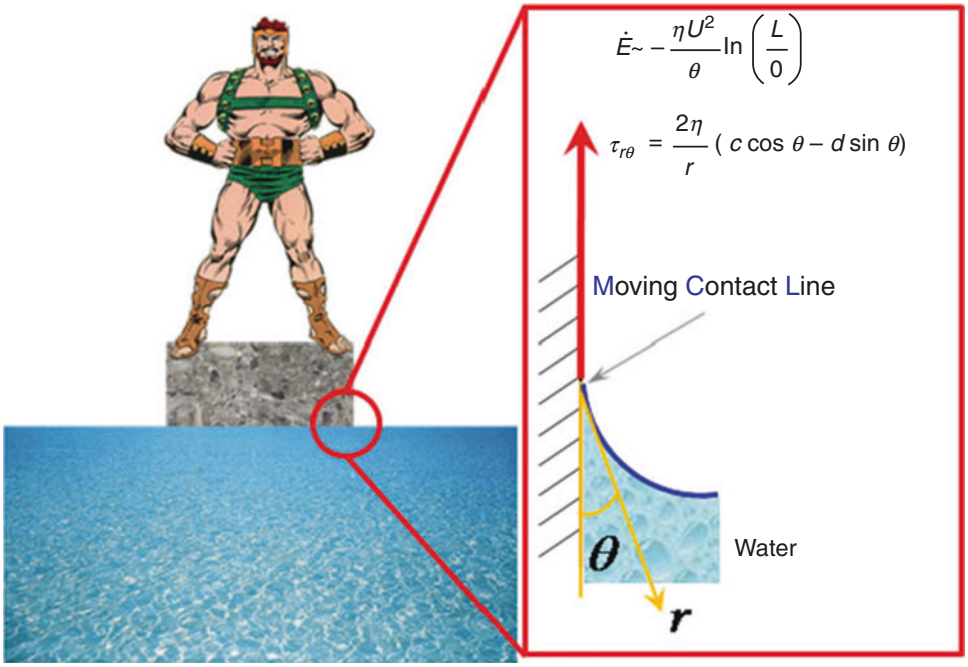
Apart from theoretical analysis, experimental investigations on surface deformation induced by droplet have also been reported. Because of the surface resolving ratio, it is difficult to get the detailed information of surface deformation at the contact line. Moreover, there might be a highly stressed zone near the contact line so that such a question should be studied further when necessary.

Surface Tension Effects of Wetting at the Three-Phase Contact Line

When considering the surface tension effects at the three-phase contact line, there is a famous paradox named Huh–Scriven paradox. It was first pointed out by Huh and Scriven [21] that there is a conflict between the moving contact line and the conventional no-slip boundary condition between a liquid and a solid. The interface meets the solid boundary at some finite contact angle θ. Owing to the no-slip condition, the fluid at the bottom moves with constant velocity U and viscosity η, while the flux through the cross section is zero. The energy dissipation per unit time and unit length of the contact line is obtained, $\dot{x} \sim -\eta U^2 \ln(L/0)$, where L is an outer length scale like the radius of the spreading droplet. Stresses are unbounded at the contact line, and the force exerted by the liquid on the solid becomes infinite. The energy dissipation is logarithmically diverging, "not even Herakles could sink a solid" (Fig. 7)

In reality, dynamic wetting occurs at a finite rate with changes in the wetted area and liquid shape. These processes are thermodynamically irreversible and therefore dissipative. But, the energy dissipation is finite. There are two typical theories in identifying the effective channel of energy dissipation for small Capillary and Reynolds number. One of the two approaches is the hydrodynamic theory emphasizes energy dissipation caused by viscous flow within the wedge of liquid near the moving contact line. The other is the molecular kinetic theory emphasizes energy dissipation caused by of attachment (or detachment) of fluid molecules to (or from) the solid surface [22].

The Huh–Scriven paradox is raised from four ideal assumptions summarized as: incompressible Newtonian fluid, smooth solid surface,



Surface Tension Effects of Nanostructures, Fig. 7 Huh and Scriven's paradox: not even Herakles could sink a solid

impenetrable liquid/solid interface, and no-slip boundary. Hence, the typical methods proposed to relieve the dynamical singularity near the contact line are the precursor film, surface roughness, diffuse interface, and nonlinear slip boundary, aiming the four assumptions respectively, as shown in Fig. 8.

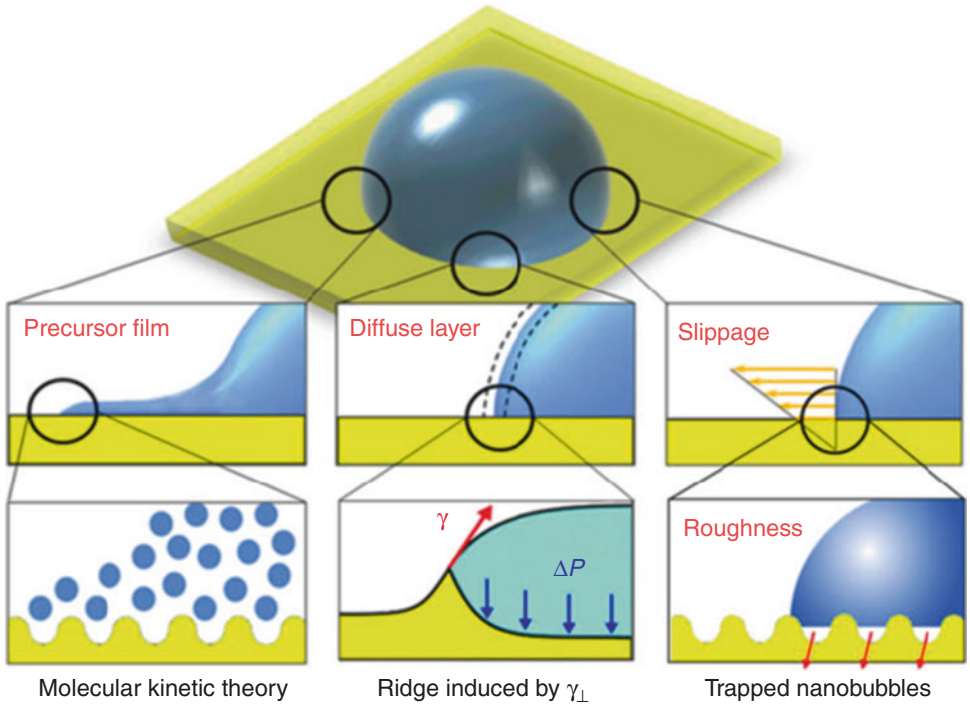
Examples of Application

Nanostructures of Silicon Used as the Anode Material of Lithium Ion Batteries

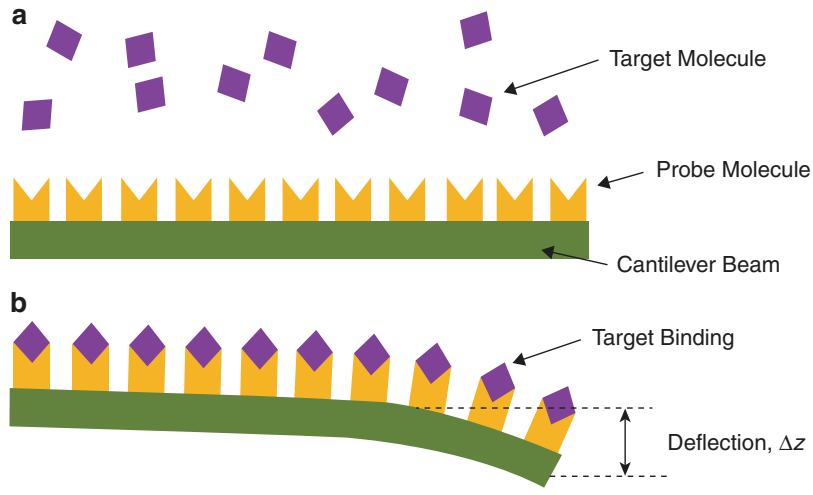
Rechargeable lithium ion batteries become the most suitable energy carrier for portable electro-equipments, electromobiles, and high performance computing, not only for the high-energy density and low cost, but also for the environmental needs for energy storage. Silicon is selected as a promising anode material of the lithium ion batteries due to the high-energy density (about 4,200 $mAhg^{-1}$). Nevertheless, a loss of electrical contact due to the fracture and crack of the bulk silicon induced by huge volume change (~400 %) in charging and discharging cycles hinders its applications. In recent years, nanostructured silicon is investigated as the material for electrode effectively circumvented the fragmentation, since surface tension effects of the nanostructured silicon on diffusion induced stresses would have much effect on the stress distribution [23].

Surface Tension Effects Induced the Deflection of the Cantilever Sensor

Surface stresses scale linearly with dimension and surface to volume ratio increases as micro/nanostructure scale decreases. Therefore, surface stress can be very important in microstructures in the size domain of MEMS/NEMS. Surface stress has been used as an effective molecular recognition mechanism. For a microcantilever used in a bioactuator, some biomaterials and biomolecules, which are immobilized on the microcantilever surface, are used to convert chemical energy into mechanical energy. When excitations such as DNA hybridization and receptor-ligand binding are applied, the microcantilever generates a



Surface Tension Effects of Nanostructures, Fig. 8 Possible mechanisms to solve the Huh and Scriven's paradox



Surface Tension Effects of Nanostructures, Fig. 9 Schematic illustration of the deflection of a cantilever sensor due to the surface stress induced by surface tension effects [8, 17]. (**a**) The cantilever is functionalized on one side with the probe molecules. (**b**) After the target binding, the cantilever bends and the deflection can be measured

nanomechanical deflection response due to the surface tension effects, as shown in Fig. 9.

Surface Tension Effects Used in the Application of Electrowetting

One particularly promising application area for electrowetting is the manipulation of individual droplets in digital microfluidic systems. Applications range from "lab-on-a-chip" devices to adjustable lenses and new kinds of electronic displays. Besides, electrowetting has been used for the application to displays showing video content since the switching speed is very high (only a few milliseconds) [10].

Summary

The surface tension effects become particularly predominant in nanostructures since the surface to volume ratio is sizable. In this entry, the surface tension effects that lead to the size-dependent mechanical properties of nanostructures are considered. Some key research findings related to this issue were listed and several examples of application were given, which are expected to be helpful to readers. Understanding and controlling these surface tension effects is a basic goal in the design and application of nanodevices.

Cross-References

- ▶ Disjoining Pressure and Capillary Adhesion
- ▶ Electrowetting
- ▶ Nanoscale Properties of Solid–Liquid Interfaces
- ▶ Surface Energy and Chemical Potential at Nanoscale
- ▶ Wetting Transitions

References

1. Bhushan, B. (ed.): Handbook of Nanotechnology. Springer, New York (2010)
2. Gibbs, J.W.: On the equilibrium of heterogeneous substances. In: Gibbs, J.W. (ed.) The Scientific Papers of J. Willard Gibbs. volume 1: Thermodynamics, pp. 55–353. Dover, New York (1961)
3. Cahn, J.W., Hilliard, J.E.: Free energy of a nonuniform system. I. Interfacial free energy. J. Chem. Phys. **28**, 258–267 (1957)
4. Lu, W., Suo, Z.: Dynamics of nanoscale pattern formation of an epitaxial monolayer. J. Mech. Phys. Solids **49**, 1937–1950 (2001)
5. Butt, H.J., Graf, K., Kappl, M.: Physics and Chemistry of Interface. Wiley-VCH, Weinheim (2003)
6. Shuttleworth, R.: The surface tension of solids. Proc. Phys. Soc. A **63**, 444–457 (1950)
7. Herring, C.: Surface tension as a motivation for sintering. In: Kingston, W.E. (ed.) The Physics of Powder Metallurgy, pp. 143–179. McGraw Hill, New York (1951)
8. Fritz, J., Baller, M.K., Lang, H.P., Rothuizen, H., Vettiger, P., Meyer, E., Guntherodt, H.J., Gerber, C., Gimzewski, J.K.: Translating biomolecular recognition into nanomechanics. Science **288**, 316–318 (2000)
9. Stoney, G.: The tension of metallic films deposited by electrolysis. Proc. R. Soc. Lond. A **82**, 172–175 (1909)
10. Berthier, J.: Microdrops and Digital Microfluidics. William Andrew, Norwich (2008)
11. De Gennes, P.G.: Wetting: statics and dynamics. Rev. Mod. Phys. **57**, 827–863 (1985)
12. De Gennes, P.G., Brochard-Wyart, F., Quere, D.: Capillarity and Wetting Phenomena. Springer, New York (2004)
13. Yuan, Q.Z., Zhao, Y.P.: Precursor film in dynamic wetting, electrowetting and electro-elasto-capillarity. Phys. Rev. Lett. **104**, 246101 (2010)
14. Gurtin, M.E., Murdoch, A.I.: A continuum theory of elastic material surfaces. Arch. Ration. Mech. Anal. **57**, 291–323 (1975)
15. Wang, Z.Q., Zhao, Y.P., Huang, Z.P.: The effects of surface tension on the elastic properties of nano structures. Int. J. Eng. Sci. **48**, 140–150 (2010)
16. Zhang, Y., Ren, Q., Zhao, Y.P.: Modelling analysis of surface stress on a rectangular cantilever beam. J. Phys. D Appl. Phys. **37**, 2140–2145 (2004)
17. Ren, Q., Zhao, Y.P.: A nanomechanical device based on light-driven proton pumps. Nanotechnology **17**, 1778–1785 (2006)
18. Guo, J.G., Zhao, Y.P.: The size-dependent bending elastic properties of nanobeams with surface effects. Nanotechnology **18**, 295701 (2007)
19. Shanahan, M.E.R., Carré, A.: Nanometric solid deformation of soft materials in capillary phenomena. In: Rosoff, M. (ed.) Nano-Surface Chemistry. CRC Press, New York (2001)
20. Yu, Y.S., Zhao, Y.P.: Elastic deformation of soft membrane with finite thickness induced by a sessile liquid droplet. J. Colloid Interface Sci. **339**, 489–494 (2009)
21. Huh, C., Scriven, L.: Hydrodynamic model of steady movement of a solid/liquid/fluid contact line. J. Colloid Interface Sci. **35**, 85–101 (1971)
22. Wang, F.C., Zhao, Y.P.: Slip boundary conditions based on molecular kinetic theory: the critical shear stress and the energy dissipation at the liquid–solid interface. Soft Matter **7**, 8628–8634 (2011)
23. Cheng, Y.T., Verbrugge, M.W.: Evolution of stress within a spherical insertion electrode particle under potentiostatic and galvanostatic operation. J. Power Sources **190**, 453–460 (2009)

Surface Tension–Driven Flow

- ▶ Capillary Flow

Surface Tension–Powered Self-Assembly

- ▶ Capillary Origami

Surface-Enhanced Raman Scattering for Imaging Biological Cells

Nicolas Pavillon[1], Katsumasa Fujita[2] and Nicholas Smith[1]
[1]Biophotonics Lab, Immunology Frontier Research Center (IFReC), Osaka University, Osaka, Japan
[2]Department of Applied Physics, Osaka University, Osaka, Japan

Definitions

Nanoparticles, as a term, generally refer to any type of particle that is composed of metallic, semiconductor, or even biological material. The particles sizes that fall under the term usually range from around 1 to 100 nm in diameter. Very small size nanoparticles, where the number and arrangement of atoms are well determined, are also termed nanoclusters (e.g., a 22-atom gold nanoparticle is usually referred to as a nanocluster). In the current article, the term "nanoparticles" is used to refer to 50 nm diameter gold nanoparticles, unless otherwise noted. Surface-enhanced Raman scattering (SERS) is a term that refers to the amplification of the wavelength shift in scattered light where the shift is caused by vibrations in a sample (similar to a Doppler shift), and the amplification is caused by the presence of a surface. "Biological cells" usually refer to living cells, either in tissue or cultured on a substrate. Whether mammalian or not, cells are the building blocks of organisms and are therefore primary targets of interest for these emerging nanoparticle-based analytic methods. SERS imaging in biological cells is then the use of these particles in cells and the methods of attempting to use the resulting spatial and spectral information for analyzing the cell state and in particular the molecular conditions around the particle. This can be either on the cell membrane surface or inside the cell if the particles can be moved to positions of interest in the cell.

Overview

Nanoparticle types suitable for biological SERS measurements: For biological applications where SERS is a key interest, the particle material is limited to compositions that have plasmonic properties over suitable light wavelength ranges. Usually the toxicity of nanoparticles themselves and/or the phototoxicity of the light which excites the SERS effect needs to be reduced to as low a level as possible [1]. Practically, this means that most applications use inert gold as the basis for the nanoparticles, since gold is relatively chemically inert [2] while still exhibiting significant plasmonic enhancement for suitable particle sizes. Even though gold is a precious metal, the cost is usually not a significant factor in the choice of metal; the amount of gold used in typical SERS imaging experiments using cultured cells in a dish is on the order of a few US dollars or less.

For a given metal, the optimum nanoparticle size is then also bounded by a number of physical limitations: the particle must be large enough to undergo plasmonic enhancement at excitation wavelengths compatible with intracellular experiments, and it must be small enough to allow uptake or injection into a cell. It must be of a size that does not significantly interfere with the homeostasis of the cell. Very small nanoparticles ($\sim$2 nm) are thought to intercalate with nucleic acids, while very large particles will either be excluded by the cell or may significantly disrupt normal cellular function if taken up by a cell. Taken together, these constraints usually result in the choice of metal nanoparticles of approximately 50 nm diameter and spheroidal shape, with gold as the metal. These can be synthesized in a number of different ways; however, since they are commercially available, explicit treatment of synthesis methods is not covered in this current work. There are a number of points to be aware of in the purchase and/or synthesis of such gold nanoparticles. For example, the surface coating can significantly change the uptake, aggregation, toxicity, and surface charge of the nanoparticles. Some nanoparticles in literature reports are

referred to as "naked nanoparticles," but this usually refers to a minimal surface coating that is not thought to influence the properties of the nanoparticles significantly. Truly naked (i.e., without any surface coating) nanoparticles are difficult to prepare as a colloidal solution since aggregation would make the solution unstable.

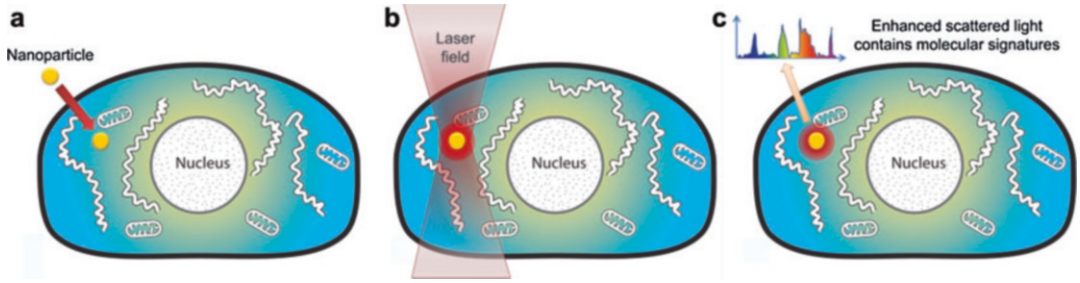
Non-spheroidal Nanoparticles

The plasmonic properties of metal nanoparticles depend on the type of metal and on the size and geometry of the particle. The most widely used particle shapes for intracellular measurements are 50 nm spherical gold particles; however, a very large range of complex particles shapes can be formed by so-called "bottom-up" methods [3]. In contrast to cutting shapes out by electron beam or other patterned methods, the shape is formed in bottom-up methods by controlling the fabrication parameters so that the desired shape emerges as a result of the fabrication reaction. These wide-ranging shaped nanoparticles have a number of applications. One commonly used technique is to use nanoparticles that are approximately cylindrical in shape; this then gives a long and short axis. Due to the plasmonic resonances, there are typically two resonance peaks associated with the cylindrical geometry, and the particles can be readily excited by wavelengths between the visible and near-infrared range. Another type of particle geometry involves using nanoshells where the outside shell of the particle is made of a plasmonic material (often gold) and the inner core can be made of a variety of different materials. This then allows much more sophisticated tuning of the plasmonic properties. Even for spherical-shaped nanoshells, the breaking of symmetry (whereby the core center is not located at the center of the particle) allows significant modification or properties.

Plasmonic enhancement is best treated in dedicated articles, which are numerous, but are only recently entering the public awareness [4]. For the purposes of this current work, nanoparticle plasmonic enhancement can be briefly stated to be the interaction (i.e., the coupling) between the "plasma" or electrons in the metal and an applied external light field. The electrons are affected by the external field and move in response as waves on the surface of the metal, which leads to the resonant properties of the effect and thereby the dependencies on both particle geometry and the wavelength of the external field.

Practical Use of Nanoparticles for Cellular Imaging

Even the use of the word "imaging" requires some clarification since particles are usually only present in low numbers and in relatively few regions of the cell. Starting from the idea of a cell without nanoparticles (Fig. 1a), and by some means the particles enter the cell, so that if a laser of the right wavelength is irradiated onto the region containing the nanoparticle (Fig. 1b), a local-enhancement occurs, and the enhanced Raman signal is then read out from the cell (Fig. 1c). Getting nanoparticles into cells is in itself often a challenge, and a number of different options are available to the experimentalist to induce the uptake. For some cell types (most notably macrophages), uptake of general debris is one of the biological functions of the cell, and for these types, nanoparticles are relatively quickly taken up. For other types, such as HeLa cells, which do not have specific uptake roles in the body, nanoparticles will still enter, but at a much lower rate. This is not necessarily a limitation but it should be taken into consideration when designing experiments. If nanoparticle-enhanced SERS signals are an integral part of the measurements, then long incubation times (as a rough estimate, on the order of 24 h but with wide variance) may be required to have a sufficient number of particles inside the cell in order to achieve meaningful results. Additionally, by modifying nanoparticle surface properties, or the culturing conditions, these uptake rates can be significantly changed.



Surface-Enhanced Raman Scattering for Imaging Biological Cells, Fig. 1 Overview of nanoparticle-cell measurements by surface-enhanced Raman scattering: Nanoparticles enter an otherwise normal cell (**a**), where an external laser field then excites a strong local field around the nanoparticle (**b**). With the possibility of additional chemical enhancement, the nanoparticle can be used to report the local Raman molecular signature in the cell (**c**)

Targeted Versus Tagged Versus Naked Nanoparticles

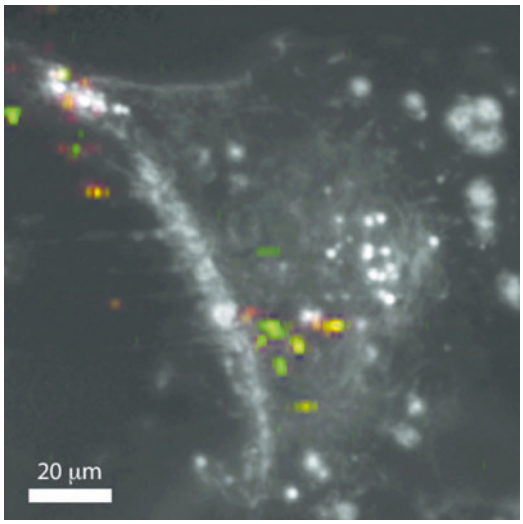
A major distinction between different types of intracellular SERS measurements is whether nanoparticles are modified to target them to particular intracellular molecules, tagged so that they contain their own labels [5] or left relatively untreated to attempt unbiased measurement of endogenous intracellular molecules. For targeting, the particles can be conjugated with a compound that binds with an intracellular target molecule. Tagging can be achieved by using a layer of Raman-active material which results in enhanced emission, often with an additional layer of inert material over the Raman-active material. These tagged nanoparticles can then be additionally targeted to intracellular molecules by conjugation with an appropriate molecule. Even in the case of the so-called "naked" nanoparticles, where particles are expected to enter the cell, provide a base for surface enhancement, and thereby act as a means to report the intracellular milieu, the idea of an unbiased measurement of intracellular molecules using gold nanoparticles to probe them is not strictly accurate. The reason is that there are a number of effects at play in the surface enhancement of molecules. Two main contributing roles are played by two fundamentally different physical mechanisms. Electromagnetic enhancement of the field occurs near the surface and will enhance both the incident field and the emitted SERS field, while a chemical enhancement also occurs due to chemical interactions between the surface and the molecule, increasing the probability of generating Raman-scattered photons. The exact role of these two modes in SERS was debated for a number of years in the literature, but what is important for the present discussion is not so much the relative role of each but rather that they both exist, and therefore two different molecules in the same enhanced field will experience a different probability of generating SERS, leading to a description of molecules being "SERS-active" or not. Other effects coming into play include the fact that the same molecule can interact with a surface in different ways, with different orientations, and therefore produce a difference not only in the intensity of the SERS signal but also in the nature of the signal provided. Some Raman bands will not be active in SERS, some will be enhanced, and some will show significant dependence on the exact nature of the chemical interaction with the substrate, incident field polarization, fluctuations in time with diffusion, and other parameters. This leads to issues which must not only be dealt with at the time of interpreting data but must also be considered when designing the experiment. As a rough example, lipids tend to be poor molecules for intracellular observation by SERS. If lipids were a key target, then conjugated nanoparticles targeted specifically for lipids would be a relevant choice (although practically conjugation possibilities for lipids are limited). If, instead, it was desired to attempt to measure the endogenous molecules responsible for endocytosis and subsequent transport, naked nanoparticles would be a reasonable choice since they assume nothing

about the molecules involved and may instead report the local molecules which transport the nanoparticle (subject to the inherent bias resulting from the difference in SERS activity for different molecules).

Measuring SERS

The next issue in intracellular SERS measurements would be the choice of an appropriate wavelength for the laser which excites the SERS signals. While the use of the nanoparticle imparts some plasmonic effect which is needed to boost the signal to a detectable level, it is often not straightforward to choose the laser wavelength. Intuitively, matching the laser wavelength to the plasmon resonance would seem to be an ideal choice, but practically, there are reasons to avoid this. When we are using gold nanoparticles of nominally 50 nm diameter (for reasons outlined above), the plasmon resonance is typically around 532 nm. Choosing an excitation laser wavelength of 532 nm (which happens to be one of the most common laser wavelengths) tends to excite luminescence from the gold, which is a broadband, red-shifted light output that easily overwhelms the relatively weak SERS signal. The excitation by 532 nm can give a precise image of the locations of nanoparticles due to this luminescence, but we usually want to observe the SERS signals. This then leads to a trade-off: at 532 nm we will have strong plasmonic enhancement but overwhelming luminescence. At longer wavelengths between say 640 and 800 nm, we will have significantly less luminescence but also reduced SERS signals. It is still practically beneficial to choose a wavelength between approximately 600 and 800 nm, and most SERS experiments inside cells use this wavelength range.

Detection is a complex issue for this type of experiment. Before determining how detection can be achieved, we need to consider how the particles might be located in the cell. Particles do not cover the entire cell, and SERS will only be observed in regions where particles exist (and often specifically at hotspots between particles). If we image the entire field of view using, for example, the excitation wavelength of 780 nm, then the resulting image would be detected through the spectrometer (details clarified below) and the SERS signals would appear in locations where particles existed. This is outlined in Fig. 2. Here, the cell was first imaged using a dark-field imaging mode which highlights strongly scattering objects (including nanoparticles as well as endogenous scattering molecules such as lipid vesicles and cellular structure), analogous to the way that dust in the air is often visible when a beam of early morning light strikes into a room through a curtain. Overlaid on the dark-field imaging mode is a color-mapped red channel that corresponds to the locations of nanoparticle in the cell, as determined from luminescence measurements, and also a green channel which shows "high-quality" SERS signals detected from the SERS mode (we will clarify what "high quality" means below). In areas that have both green and red, i.e., where the locations of particles correspond perfectly to the locations of "high-quality" SERS signals, the pixels should appear yellow. Note that very few of the pixels in the color



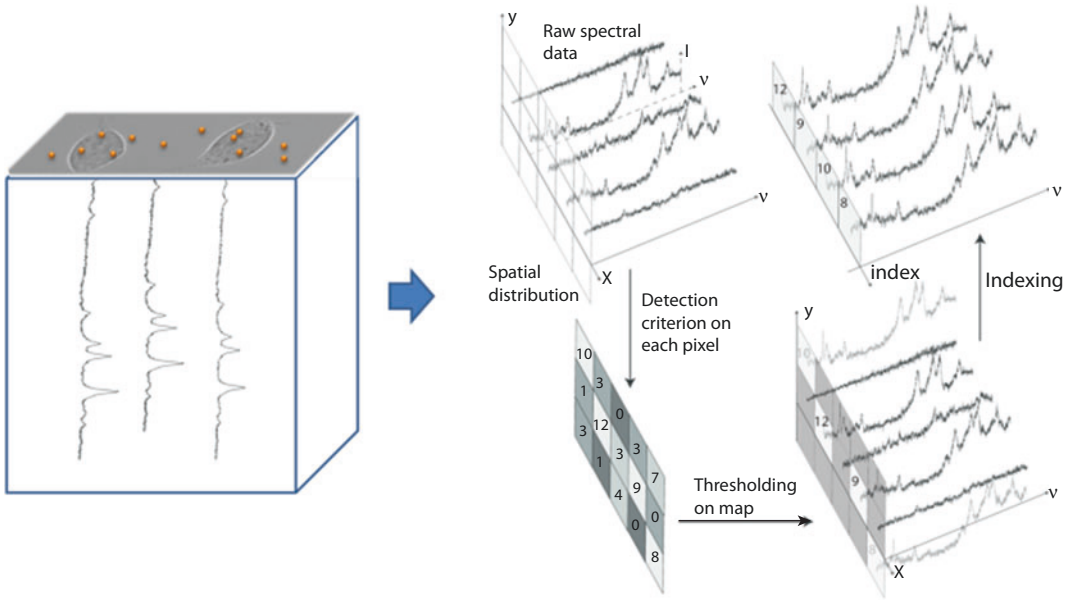
Surface-Enhanced Raman Scattering for Imaging Biological Cells, Fig. 2 Image of a HeLa cell with gold nanoparticles. In grayscale is shown a dark-field image taken before SERS measurement. The red channel shows the position of GNPs within the cell, and the green one presents the quality map for the spectra detection (See Pavillon et al. [6] for details)

overlay actually appear yellow. This gives us a strong hint that the detection of spatial arrays of SERS is actually a significantly complex issue that must be carefully considered before acquiring SERS measurements from a cell.

For labeled nanoparticles, the SERS measurements can be carried out in a manner analogous to fluorescence measurements, with the exception that rather than a fluorescence excitation process, the SERS process is generated by the excitation laser and emission is detected. With a full spectrum expected from each pixel, the simplest method is to raster scan the excitation beam and then to pass the emission light through a spectrometer, acquiring the emission spectrum for each point as rapidly as possible. The overall imaging rate will be limited by this spectral readout time for each spatial pixel. Since modern spectrometers are available with "imaging" capability and with a 2D detector, it is possible to excite an entire line of the sample and, for each point on the line, to resolve the spectra at the detector plane of the spectrometer. This means each full-frame readout of the 2D detector at the spectrometer provides one spatial line in the sample plus the corresponding spectral information for each pixel in that line. Such line-scanning SERS methodology is now commercially implemented and available, making frame rates of roughly one minute per frame for a frame size of 400 by 400 pixels, although it should be noted the frame rates will depend very significantly on a number of factors, including the sample type, exposure time, excitation wavelength, power, etc.

Obtaining Meaning from Large SERS Datasets

For a 2D image of cells containing nanoparticles, a key issue is that before measuring we often don't know (a) where the nanoparticles are in the cell and (b) whether those nanoparticles will provide any SERS signals. There are several approaches we can choose to accommodate this lack of knowledge. For tagged or labeled nanoparticles, the use of a Raman reporter can make the emission strong enough that the particles effectively behave like fluorescent tags (albeit with the possibility of multiplexing due to the discrimination ability of the additional spectral dimension). For experiments where the SERS signals result from molecules outside the nanoparticle, there is usually significant variation in the signals, depending on the local environment in the cell, as well as fluctuations in time. If we simply add nanoparticles to a cell culture, wait for the requisite time for particle uptake, and measure the resulting spatial distribution by SERS imaging, the resulting dataset is a 3-dimensional stack that will contain some SERS signals at some locations. Along with the SERS data, background signals including fluorescence, cosmic rays, and surface contamination of the nanoparticles will often appear. This means that after acquiring the SERS dataset, it is still challenging to determine which data is relevant for the SERS experiment. While nanoparticles are required to generate SERS signals, the SERS signals do not appear at all nanoparticle locations, due to a number of effects (such as gaps between particles, differences in molecule-surface interactions, and others described above). This is demonstrated in Fig. 3. The resulting dataset will be composed of mostly non-SERS-related pixels. Classification of SERS spectra can be done by algorithm, but careful choice and testing of the algorithm on known datasets should be done before application to new or unknown datasets. For example, a simple algorithm to separate SERS data from non-SERS data would be to select the pixels that contain the most spectral energy, under the assumption that the surface-enhanced regions should be brighter than the background. This is often not the case though for real data and is further complicated by the fact that the surface enhancement also applies to background signals and fluorescence. An algorithm that then classifies SERS signals based on spectral peaks, spectral energy and other properties specific to SERS can be developed to recognize SERS features. Pixels ranked high in the SERS detection algorithm can be loosely described as having "high-quality" SERS spectra. Once this information is known and the locations of those spectra are known, higher-order analysis can be applied whereby spectra are further sorted into clusters of similar



Surface-Enhanced Raman Scattering for Imaging Biological Cells, Fig. 3 HeLa cells incubated with gold nanoparticles, showing how nanoparticle distribution does not match detected SERS distributions. The large datasets resulting from imaging SERS contain a large number of pixels with background signals which do not contribute to the SERS information of interest. One approach to simplify the data is to design an algorithm to remove pixels which have a spectra that do not have SERS features, by applying a detection criteria based on known properties of SERS spectra and thresholding data that does not meet the criteria [6]

spectra, and questions such as "do all cells show the same variations in spectra?," "are all spectra completely randomly distributed within a cell?," or "are there correlations between variance in spectra and cell type or function?" can all be asked and answered from the data analysis [6].

A Tracking Approach to Nanoparticle Dynamics and SERS Measurement

The method described above measures the entire field of view and then subjects the data to analysis using customized algorithms that recognize SERS features. An alternative to measuring the entire field is to instead track the location of a single particle and follow its trajectory with the SERS excitation laser beam [7]. This allows the possibility of discrimination of very rapid changes in SERS signal in time. If the SERS excitation laser follows the particle, then the SERS detection scheme essentially reduces to 1-dimensional detection (the point is read out as a 1D spectral array). The minimum exposure time can then be reduced to an interval as short as several tens of milliseconds. SERS signals usually contain a significant amount of inherent temporal fluctuations, which can derive from the combination of molecular movement and the very high sensitivity of the SERS effect. Such dynamics are a source of information, particularly inside a cell where molecular motion may result from simple Brownian motion or by directed intracellular transport. By reducing the spatial dimensions to a point, i.e., with the loss of spatial information, there is the possibility of a significant gain in temporal information. Even with the SERS detection only occurring from a single point in the sample, the technique can still be considered a type of imaging: in order to track and follow the particle motion with the SERS laser, a secondary imaging mode is needed which provides an image of the particle location. This mode can be a simple imaging mode, such as dark-field microscopy, and due to the a priori knowledge of the nanoparticle shape, a centroid detection algorithm can be employed to estimate the center of the particle to a level much

finer than the overall microscope resolution. The particle trajectory can then be tracked with a resolution of tens of nanometers, which is a scale of particular importance to cell biology and active transport. The combination of the particle trajectory information with the SERS feedback allows the additional possibility of correlating the two sets of information. This is important because it allows the determination of which types of SERS signals correspond with which types of particle motion. The typical SERS spectra for Brownian motion can be distinguished from typical SERS spectra for directed transport within the cell, and this is a good starting point to use SERS-based analysis to help discover molecules involved in the cellular transport mechanism.

Other Approaches to SERS Measurements of Cells

With the possibilities of tracking single molecules, high resolution, and making new discoveries in the cell, it is not surprising that a number of approaches to SERS measurements in cells have been developed, not limited only to the ones described above. For example, nanoparticles can be created inside the cells by laser induced photoreduction, and then used to read out molecular signatures [8]. Additionally, SERS is not limited only to the interior of cell samples. It is often easier, in practice, to use SERS as a tool to probe membrane-based targets, since it does not require cellular uptake. Incubation of gold nanoparticles with cultured cells results in settling of some portion of the nanoparticles on the cell surface within several minutes. These can then be used to attempt to detect surface receptors or uptake pathways. It is also relatively common to use a SERS-active material as a substrate for cultured cells. This allows the investigation of the basal cell membrane, with a particular affinity for cell adhesion molecules, since they are the regions actually in contact with the substrate. A substrate approach has a number of key features: the detection range in the axial direction (i.e., away from the substrate) is extremely short and on the order of a few tens of nanometers. Additionally, the substrate itself should be relatively stable so that cell movements should be movements across the substrate and not dynamics of the substrate itself. This can simplify the analysis of SERS signals and negates the requirement to account for changes of the SERS sites themselves which can complicate analysis in intracellular nanoparticle-based SERS. It is also possible to fabricate gold nanoparticles that can be used for SERS directly in targeted locations inside a cell. This approach uses focused laser light to generate a photoreduction reaction of gold ions and form gold nanoparticles at the laser focus. This makes the task of finding the SERS locations much simpler since they appear at the laser focus itself. Currently, SERS measurements in cells are already demonstrated to be a method by which unique information on cell content can be determined. It is still an area under intense development, however, and very significant advances are to be expected over the next decade.

Cross-Reference

► Plasmonics

References

1. Sutariya, V.B., Yashwant, P.: Biointeractions of Nanomaterials. CRC Press, Boca Raton (2014)
2. Mohr, F. (ed.): Gold Chemistry: Applications and Future Directions in the Life Sciences. Wiley-VCH, Weinheim (2009)
3. Sau, T.K., Rogach, A.L. (eds.): Complex-Shaped Metal Nanoparticles. Wiley-VCH, Weinheim (2012)
4. Atwater, H.A.: The promise of plasmonics. Sci. Am. **296**, 56–62 (2007)
5. Chau, L.-K., Chang, H.-T.: From Bioimaging to Biosensors: Noble Metal Nanoparticles in Biodetection. Pan Stanford, Singapore (2013)
6. Pavillon, N., Bando, K., Fujita, K., Smith, N.I.: Feature-based recognition of surface-enhanced Raman spectra for biological targets. J. Biophotonics **6**, 587–597 (2013)
7. Ando, J., Fujita, K., Smith, N.I., Kawata, K.: Dynamic SERS imaging of cellular transport pathways with endocytosed gold nanoparticles. Nano Lett. **11**, 5344–5348 (2011)
8. Smith, N.I., Mochizuki, K., Niioka, H., Ichikawa, S., Pavillon, N., Hobro, A.J., Ando, J., Fujita, K., Kumagai, Y.: Laser-targeted photofabrication of gold nanoparticles inside cells. Nat. Comm. **5**, 5144 (2014)

Surface-Modified Microfluidics and Nanofluidics

Kaushik K. Rangharajan and Shaurya Prakash
Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, OH, USA

Definition

Fluid phenomena in micrometer (1–100 μm) or nanometer (1–100 nm)-sized channels that are governed by coupled principles from fluid mechanics, surface chemistry, electrochemistry, and electrostatics are generally referred to as microfluidic and nanofluidic phenomena [1]. The smaller length scales compared to traditional fluid mechanics provide several new and interesting phenomena due to significant enhancement of the surface-area-to-volume ratio, which makes surface-mediated flows important to the overall field of microfluidics and nanofluidics.

Introduction to Surfaces in Microfluidics and Nanofluidics

Microfluidic and nanofluidic devices and systems are characterized by high surface-area-to-volume (*SA/V*) ratio. For example, a rectangular cross-section channel with 100 μm width, 100 nm depth, and 1 cm length will have a *SA/V* ratio on the order of 10^6 m^{-1}. In fact, operational devices incorporating components with *SA/V* ratio on the order of 10^9 m^{-1} have already been reported [2]. Consequently, the influence of device and system walls in affecting phenomena within confined spaces can no longer be ignored since the walls (i.e., surfaces) of the devices and systems interact extensively and directly with the species contained within these devices. Therefore, the surface of interest here is defined as an interface or thin region in space (often only a few nm in physical extent) that influences transport and reaction phenomena in its vicinity. In this surface-region properties such as chemical composition, refractive index, mechanical strength, conductivity, and charge can significantly differ from the bulk material of the underlying substrate.

Scaling of Surface Forces in Microchannels and Nanochannels

One non-dimensional parameter often used to characterize flow regimes is the ratio of the inertial forces transferred by the velocity (or momentum) of the fluid to the viscous (or frictional forces), which is the Reynolds number, Re, and is given by

$$\mathrm{Re} = \frac{\rho V L_c}{\mu}, \tag{1}$$

where, ρ is the fluid density, μ is the fluid viscosity, and V is the fluid velocity.

As seen from Eq. 1, Re is directly proportional to the characteristic length, L_c of the flow. With L_c decreasing (and consequently *SA/V* increasing), the inertial forces decrease, and as a result the Re becomes smaller. The direct consequence is that for given flow parameters of fixed velocity and fluid type, a decreasing L_c implies increasing influence of viscous forces in contrast to the inertial forces.

In most microfluidics and nanofluidics applications, a particle (ion, colloid, biomolecule, AFM probe tip, etc.) will interact with the channel walls. Considering three important forces – electrostatic or Coulombic (F_{el}), van der Waals (F_{vdw}), and hydrodynamic forces (F_h) – will provide a better insight between surface-particle interactions and subsequent phenomena. For an infinite flat surface separated by a distance D from a particle of radius R, F_{el} is given by [3]

$$F_{el} = -\frac{2\pi R L_D}{\varepsilon\varepsilon_0}\left[2\sigma_S\sigma_P\exp(-D/\lambda_D) + \left(\sigma_S^2 + \sigma_P^2\right)\exp(-2D/\lambda_D)\right] \tag{2}$$

where, λ_D is the Debye length and σ_s and σ_p are the surface charge densities of surface and particle, respectively. F_{vdw} is estimated by [3]

$$F_{vdW} = \frac{A_H R}{6D^2}, \tag{3}$$

where A_H is the Hamaker constant, which serves as an indicator of the interaction between the particle and the sample surface, and F_h has been approximated by [3] the expression

$$F_h = b_s V = -f^* \frac{6\pi\mu R^2 V}{D}, \tag{4}$$

where V is the particle velocity, b_s is the hydrodynamic damping coefficient, and μ is the viscosity of the fluid. The coefficient f^* is related to the boundary slip properties. If the no-slip assumption holds true at the solid/water interface, then $f^* = 1$, and if slip exists, then $f^* < 1$. As an aside, it should be noted that Eq. 4 is often used in literature for determining slip lengths using a colloidal AFM probe tip. As the operational length scales (here, characteristic length scale is separation distance, D) decrease, it can be seen from Eqs. 2 to 4 that contributions for F_{el} display an exponential dependence, F_{vdw} display an inverse quadratic dependence, and F_h display an inverse linear dependence on the separation distance. Therefore, consideration of each force term at the microscale, and even more so at the nanoscale, is essential toward completely describing importance of the wall-species interactions.

Most microfluidic and nanofluidic devices with liquid flows are driven by electric fields and fall under the category of electrokinetic flows. The main reason for using applied voltages to generate electric fields for driving flows is the scaling of pressure forces. For example, for a circular nanochannel with laminar, incompressible, Poiseuille flow with water as the working fluid, the necessary pressure drop across a 100 μm long channel which is 1 nm in diameter for only an attoliter (10^{-18} m^3) per second would be greater than 3 GPa, which is impractical for any device. Electrokinetic flows can sustain higher flow rates through nanometer channels without excessive pressures [2].

Methods for Surface Modification

Surface modification methods can be divided in two broad categories: physical and chemical methods [4]. Physical methods, in most cases, do not change the chemical composition of the surface but may change the surface roughness, grain sizes and grain boundaries, and faceting. Physical methods often use lasers, plasmas, temperature changes, ion beams, ball milling, and polishing or grinding, to alter the surface state of a material of interest. While the main intent with physical modification methods is to not alter the chemical composition of the material, in some cases, physical surface modification methods can lead to changes in the chemical composition of the surface due to removal or addition of material or chemical reactions on surfaces, as in the case of selective or ion-beam sputtering, or by selective cross-linking in the presence of plasmas. Temperature gradients and thermal treatments have been used to change surface roughness, grain sizes, and grain boundaries and create nanoscale features, facets, textures, and nanoparticles on ceramics, metals, polymers, and semiconductors. Thermal treatments in the presence of gases such as oxygen or water vapor can cause creation of steps or induce other forms of nanostructures. For example, as was discussed in a recent review [4], crystalline α-Al_2O_3 surfaces heat treated at 1,500 °C in Ar/O_2 and $H_2/He/O_2$ led to step formation and roughening as quantified through AFM topology images.

Chemical methods introduce a change in the chemical composition at the surface by introducing chemical properties (e.g., surface charge density or surface energy) different from the bulk material. Among the methods for chemical modification formation of surface layers, either covalently bonded or physisorbed, has been most common. Other chemical methods include treatment with UV light and reactive plasmas. Modification schemes are governed by a wide range of parameters including sample type (polymers, metals, ceramics, etc.), stability to treatment conditions (e.g., thermal or structural), and eventual applications. For example, polymeric surfaces are often modified by photochemical methods of

which UV irradiation in air, other reactive atmospheres such as ozone combined with lasers, and grafting surface layers are fairly common. The use of reactive plasmas has been gaining popularity due the wide compatibility of materials and integration to microfabrication processes for device development. Many gas plasmas have been used such as air, oxygen, water vapor, ammonia, and argon for modification of polymer surfaces. Plasma modification processes generate new chemical species on polymer surfaces. The new chemical species can arise due to surface reactions with reactive gases or due to physical sputtering (such as with Ar plasma) caused by active gas-phase species. These new surface chemical species can provide an anchor for attaching a series of different molecules that display different properties from the underlying bulk polymer. For example, preferential hydroxylation of poly (methyl methacrylate) or PMMA surfaces by the use of a water-vapor plasma as opposed to an oxygen plasma has been used to activate surfaces toward trichlorosilane modification and subsequent "click" chemistries to form surface scaffolds of desired functionalities. The "click" chemistry method has also been used to modify glass channels for systematic control over electroosmotic flow velocity [5].

Applications

In this section, representative applications utilizing modified surfaces are presented along with short discussions of the underlying physics. Chemically modified surfaces are influenced by changes to surface charge density, as discussed above. Any charged surface in contact with a liquid forms an electric double layer (EDL). A schematic diagram illustrating the classic EDL structure is shown in Fig. 1, depicting the positions to which different characteristic surface charge-related potentials are referenced within the EDL.

For a given substrate, it is useful to quantify the relationship between the surface charge and the surface potential in the presence of EDL. Based on the Gouy-Chapman theory, Grahame's equation [6] shown as Eq. 5 addresses this relationship assuming overall electroneutrality (i.e., the sum of surface charge and the total ionic charge, sometimes also referred to as space charge within the device volume, must be zero):

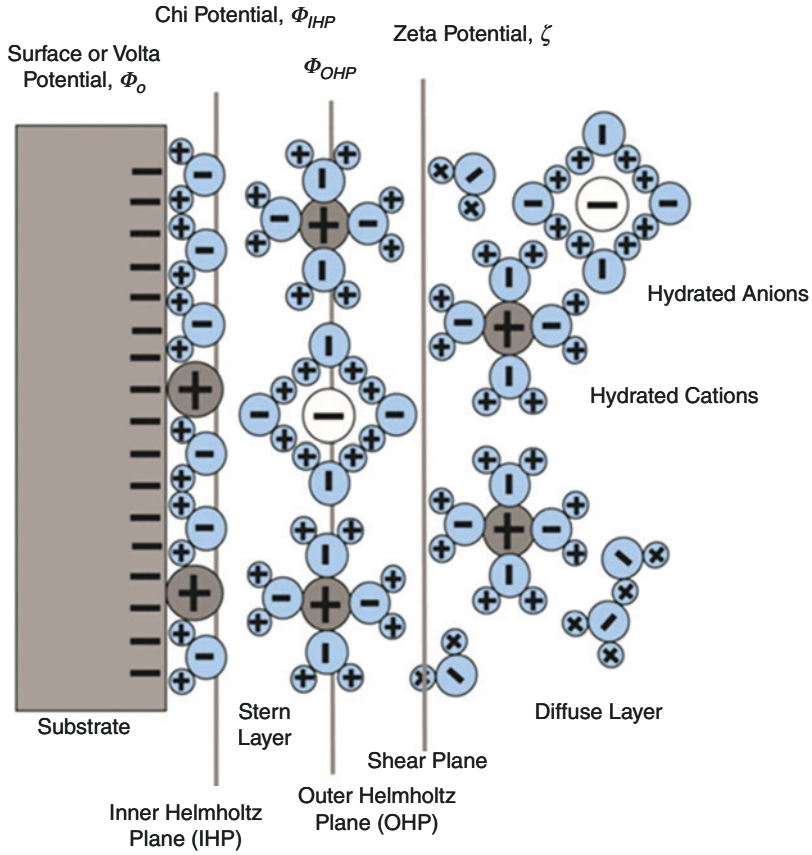
$$\sigma_s = \sqrt{8 C_0 \varepsilon \varepsilon_0 k_B T}\, Sinh\left(\frac{e\psi_0}{2k_B T}\right) \qquad (5)$$

where, σ_s is the surface charge, C_0 is the bulk concentration, ε is the dielectric permittivity, ε_0 is the permittivity of free space, T is temperature, e is the elementary charge, and ψ_0 is the surface potential. Recent results [7] show that actively controlling the surface potential (and consequently the excess surface charge) via embedded electrodes can alter flow phenomenon in nanofluidic channels. By making the electrode potential progressively negative, active flow pumping of aqueous electrolytes was observed, and making the electrode potential progressively positive resulted in reversal of both the bulk flow and electric current [7] for a native wall with a net negative surface charge density.

One direct consequence of surface modification is a change in the interaction forces between the surface and the surroundings. Specifically for microfluidics and nanofluidics, changes to surface charge density, surface roughness, and surface energy alter the electrostatic and van der Waals forces that determine flow (ionic and fluid) phenomena. Following the scaling of forces discussed above, electric fields and capillary forces are the two most common tools used to fill microchannels and nanochannels with liquids. For capillary forces, consider the Washburn equation, which provides one approach to quantify the rate of channel filling:

$$\frac{dl_f}{dt} = \frac{a\gamma \cos\theta_c}{4\mu l_f}, \qquad (6)$$

where l_f is the fill length, a is the radius (or characteristic length) of the channel, θ_c is the contact angle between the fluid wall and the fluid, γ is the interfacial (or surface) tension, and μ is the fluid viscosity. Equation 6 shows the dependence



Surface-Modified Microfluidics and Nanofluidics, Fig. 1 Schematic representation of the electric double layer. The surface charge and potentials are depicted based on the theories developed over the past century (Figure from Prakash and Yeom [1])

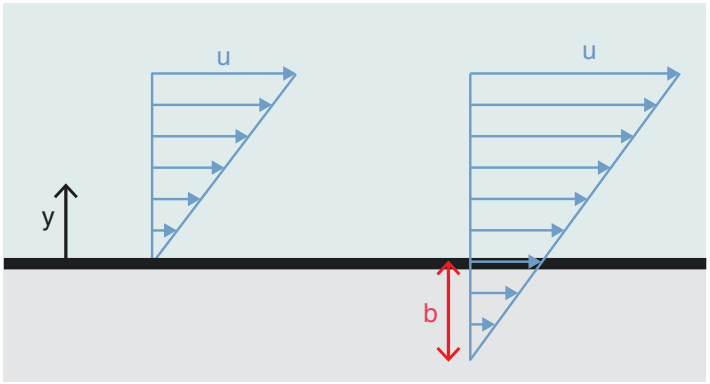
of the capillary (or micro-/nanochannel) filling as a function of surface properties governed by θ_c. Chemical surface modification can easily affect a change to the surface wettability and therefore change θ_c. Therefore, for a given liquid such as water in a micro- or nanochannel, the rate of filling decreases as the surface becomes progressively more hydrophobic.

Given the importance of walls for microfluidics and nanofluidics, a contrast to the boundary conditions with traditional fluid mechanics should be considered. For example, macroscale fluid mechanics assumes a no-slip condition at the wall, which implies that the velocity of the fluid layer in contact with the walls is the same as that of the walls. However, in the early 1800s, Navier postulated that fluid slip maybe possible and suggested the slip boundary condition with u_w, the velocity at the wall (tangential component), can be expressed by the relationship in Eq. 7:

$$u_w = b\frac{\partial u_b}{\partial y} \tag{7}$$

where, b is the slip length, u_b is the bulk velocity, and y is the axis perpendicular to the wall. Figure 2 shows a conceptual schematic for defining the physical effect of slip flow. Conceptually, one can imagine the slip length as the additional distance the wall must be extended to where the fluid velocity would be zero.

Literature has noted that there are three types of slip [8]. First is molecular or intrinsic slip, which occurs when molecular forces (e.g., van der Waals) are balanced by viscous forces. Intrinsic slip is usually observed at very high shear rates (e.g., shear rate of 10^{12} s^{-1} for water). Second is apparent slip in which there is a finite distance between the slip and no-slip planes. Electrokinetically driven microfluidics and nanofluidics follow this type of slip with the slip plane (location where ζ potential is defined in Fig. 1)



Surface-Modified Microfluidics and Nanofluidics, Fig. 2 Conceptual schematic for the slip boundary condition. The panel on the *left* shows the no-slip boundary condition for a stationary wall, with the wall velocity and the tangential component of fluid velocity being zero. The *right panel* shows a case for slip flow with a finite slip length (Figure based on [5, 6], and assistance from Mr. M. Hansen in drawing the figure is acknowledged)

being at a finite distance from the physical channel wall. Third is the effective slip which refers to the estimate of either intrinsic or apparent slip by averaging a measurement over the length scale of the experiment. The main factors that affect slip are roughness of the surface and wettability, which presents a measure of the surface energy of the surface. Other possible factors are gases trapped in the fluid, nanobubbles, shear rates of fluid at the surface, and adsorbates in the fluid [8–10], all of which can be influenced by surface modification.

Recently, it was also demonstrated that surface charge can alter the drainage velocity for confined fluids indicating possibly a role for surface charge in mediating slip flow [3]. Additional use of slip phenomena has been in developing drag reduction configurations both for laminar and turbulent flows, self-cleaning super hydrophobic surfaces [11], and enhanced mixing in microfluidic devices [12].

External pressure-driven flow of aqueous solutions in confined micro-/nanochannel configurations (with a finite surface charge) results in the generation of a current and potential and is termed streaming current and streaming potential, respectively [13]. Contrary to the typical no-slip boundary condition, for a given pressure gradient, fluid slip at the walls results in a higher flow rate of the electrolytes. Extrapolating this phenomenon to a nanofluidics battery, a higher flow rate is commensurate to an increased efficiency of streaming potential generation in the case of slip flows [14].

Over a decade back, research on imaging hydrophobic surfaces immersed in water using atomic force microscopy resulted in the discovery of soft gaseous regions on the surface [15]. Previous research over the last decade has shown that irrespective of the substrate, the height of these bubbles is at least an order of magnitude smaller than its diameter [16]. Among its many applications, the existence of these interfacial gas bubbles helps in increasing the capture efficiency in industrial froth flotation experiments, and subsequent degassing of the liquid resulted in substantial decrease in capture efficiency [17]. Also, formation of nanobubbles on coal surfaces using hydrodynamic cavitation has shown to enhance flotation recovery [18]. Electrochemical formation of nanobubbles at surfaces helps in protein removal and thus prevents surface fouling [19]. The presence of nanobubbles is believed to be one of the reasons for the existence of fluid slip at hydrophobic surfaces as the bubbles offer a low shear surface in comparison to the solid substrate and may therefore be pivotal in drag reduction applications in micro- and nanoscale flows [20]. Moreover, recent advances show that altering the surface hydrophobicity of borosilicate glass via surface modification can significantly alter

nanobubble morphology, density, and the contact line tension [21].

Continuing with a discussion of applications relevant to microfluidics and nanofluidics, valving (or metering of fluids) is one critical operation for sample manipulation for lab-on-chip devices. Valving by surface-mediated flows is a passive method due to a fixed surface state and can be achieved by the tendency of the fluid to favor a channel direction based on the wettability or the surface charge of the channel. Examples include the use of hydrophobic-hydrophilic interfaces to generate boundaries for directing water [22] and use of chemically modified surfaces for exploiting differences in surface charge for selectively driving electrokinetically driven flows in microchannels [23].

Summary

Microfluidic and nanofluidic phenomena are governed by large *SA/V* ratios. Therefore, interaction between fluids, species, and the channel walls are critical. Toward the development of devices and systems-enabling applications and related fluid control, the ability to systematically modify and control wall properties provides a powerful tool for scientists and engineers.

References

1. Prakash, S., Yeom, J.: Nanofluidics and Microfluidics. Elsevier (2014)
2. Prakash, S., et al.: Nanofluidics: systems and applications. IEEE Sensors J. **8**, 441–450 (2008)
3. Wu, Y., et al.: Dynamic response of AFM cantilevers to dissimilar functionalized silica surfaces in aqueous electrolyte solutions. Langmuir **26**(22), 16963–16972 (2010)
4. Prakash, S., Karacor, M.B., Banerjee, S.: Surface modification in microsystems and nanosystems. Surf. Sci. Rep. **64**(7), 233–254 (2009)
5. Prakash, S., et al.: 'Click' modification of silica surfaces and glass microfluidic channels. Anal. Chem. **79**(4), 1661–1667 (2007)
6. Grahame, D.C.: The electrical double layer and the theory of electrocapillarity. Chem. Rev. **41**(3), 441–501 (1947)
7. Pinti, M., et al.: Active surface potential control for artificial ion pumps. In: Technical Digest Solid-State Sensors, Actuators, and Microsystems Workshop. Hilton Head Island (2014)
8. Lauga, E., Stone, H.A., Brenner, M.P.: Microfluidics: the no-slip boundary condition. In: Tropea, C., Yarin, A., Foss, J.F. (eds.) Handbook of Experimental Fluid Dynamics, pp. 1219–1240. Springer, New York (2007)
9. Granick, S., Zhu, Y., Lee, H.: Slippery questions about complex fluids flowing past solids. Nat. Mater. **2**, 221 (2003)
10. Neto, C., et al.: Boundary slip in Newtonian liquids: a review of experimental studies. Rep. Prog. Phys. **68**, 2859–2897 (2005)
11. Bhushan, B., Jung, Y.C.: Natural and artificial surfaces for superhydrophobicity, self-cleaning, low adhesion, and drag reduction. Prog. Mater. Sci. **56**, 1–108 (2011)
12. Rothstein, J.P.: Slip on superhydrophobic surfaces. Annu. Rev. Fluid Mech. **42**, 89–109 (2010)
13. Daiguji, H., et al.: Electrochemomechanical energy conversion in nanofluidic channels. Nano Lett. **4**(12), 2315–2321 (2004)
14. Goswami, P., Chakraborty, S.: Energy transfer through streaming effects in time-periodic pressure-driven nanochannel flows with interfacial slip. Langmuir **26**(1), 581 (2009)
15. Tyrrell, J.W.G., Attard, P.: Images of nanobubbles on hydrophobic surfaces and their interactions. Phys. Rev. Lett. **87**(17), 176104 (2001)
16. Craig, V.S.J.: Very small bubbles at surfaces-the nanobubble puzzle. Soft Matter. **7**(1), 40 (2011)
17. Dai, Z., Fornasiero, D., Ralston, J.: Influence of dissolved gas on bubble-particle heterocoagulation. J. Chem. Soc. Faraday Trans. **94**(14), 1983–1987 (1998)
18. Zhou, Z.A., et al.: On the role of cavitation in particle collection in flotation – a critical review. II. Miner. Eng. **22**(5), 419 (2009)
19. Wu, Z., et al.: Cleaning using nanobubbles: Defouling by electrochemical generation of bubbles. J. Colloid Interface Sci. **328**(1), 10–14 (2008)
20. Wang, Y., Bhushan, B., Maali, A.: Atomic force microscopy measurement of boundary slip on hydrophilic, hydrophobic, and superhydrophobic surfaces. Chin. J. Vac. Sci. Technol. A: Vac. Surf. Films **27**(4), 754–760 (2009)
21. Rangharajan, K.K., Kwak, K.J., Conlisk, A.T., Wu, Y., Prakash, S.: Effect of surface modification on interfacial nanobubble morphology and contact line tension. Soft Matter, 11, 5214–5223 (2015).
22. Zhao, B., Moore, J.S., Beebe, D.J.: Surface-directed liquid flow inside microchannels. Science **291**(5506), 1023 (2001)
23. Prakash, S., Karacor, M.B.: Surface mediated flows in glass nanofluidic channels. In: Solid State Sensors, Actuators, and Microsystems Workshop. Transducer Research Foundation, Hilton Head Island (2010)

Surface-Plasmon-Enhanced Solar Energy Conversion

▶ Plasmonic Structures for Solar Energy Harvesting

Suspension Interaction with External AC Field Gradient

▶ AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation

Synchronization

▶ Nonlinear and Parametric NEMS Resonators

Synthesis of Carbon Nanotubes

Simon J. Henley, José V. Anguita and S. Ravi P. Silva
Nano Electronics Center, Advanced Technology Institute, University of Surrey, Guildford, Surrey, UK

Synonyms

Growth of carbon nanotubes; Growth of CNTs

Definition

Carbon nanotubes (CNTs) are allotropes of graphitic carbon with a cylindrical structure and diameters of <100 nm. A CNT can consist of one or many concentric graphene sheets rolled up as cylinders. CNTs are of interest for a wide variety of technological applications. In this entry, the various experimental methods to synthesize carbon nanotubes are introduced.

Introduction

Carbon nanotubes (CNTs) were first brought to worldwide attention by Iijima in 1991 [1] after he analyzed, by electron microscopy, the samples produced during an electrical arc discharge between carbon rods held in a helium atmosphere. He observed nanoscale hollow tubes, similar to those seen by Russian researchers in the 1950s [2]. Carbon nanotubes can be visualized as graphene sheets rolled up to form tubes; if one sheet is rolled up, a single-walled carbon nanotube (SWCNT) is formed. The structures formed when two or more concentric tubes are present are termed multiwalled carbon nanotubes (MWCNTs) [3–5]. The high temperatures formed during an electrical arc give an idea as to the extreme conditions that were thought to be required to form these structures. Since then, a wide variety of different growth technologies have been developed covering a broad range of different environmental conditions, from high-temperature laser vaporization down to catalytically enhanced growth in rarefied gas mixtures at much lower temperatures.

Initially, the arc discharge method [2, 6] was mainly employed to produce carbon nanotubes. Later, many more techniques such as laser ablation [7] or chemical vapor deposition (CVD) [8] were developed, researched, and optimized toward the large-scale synthesis of CNTs with repeatable and controllable morphologies. The synthesis of CNTs has now moved from a small-scale research activity to an industrial scale, with production facilities producing hundreds of tons per year. However, CNTs are not only synthesized in laboratories; they can be formed in flames produced by burning organic chemicals such as methane and benzene. They have been found also in soot in the air, likely from industrial or automotive processes, and have been generated by metallurgical process such as those used to make Damascus steel. These naturally occurring CNTs are

typically of low quality though. In this entry the main methods used to synthesize CNTs are introduced.

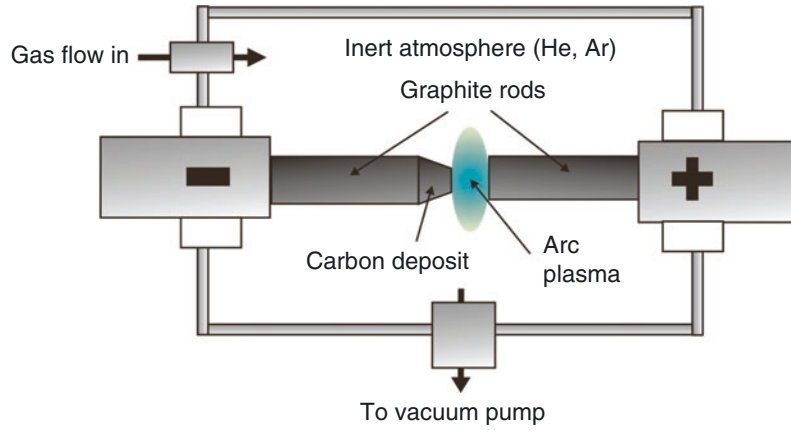
Synthesis Methods

Arc Discharge

The arc discharge technique (see Fig. 1) generally involves the use of two high-purity graphite electrodes. The electrodes are brought in close proximity, and an electric arc is struck between them by connection to a high-current, low-voltage power supply [3, 6, 9]. Similar electric arcs were used by Roger Bacon in the early 1960s to synthesize large carbon fibers. The synthesis is carried out in a controlled environment at a low pressure (50–700 millibar) of an inert gas, typically helium or argon.

The distance between the electrodes is controlled to keep a large current (50–150 A) flowing between them. The conditions in the region between the electrodes are extreme, and a high-temperature plasma is formed. The temperature is high enough that carbon evaporates (sublimes) from the anode. After the reaction time (typically 30 s to a couple of minutes) has expired and the system has cooled, the products can be collected. The products take different forms depending on the region of the reactor that they are collected from. Carbon soot is found to coat the chamber walls, and a more solid deposit is left on the cathode. The soot contains mainly fullerenes, while MWCNTs and graphitic nanoparticles are found in the solid deposit. The yield of the process can be quite high and the CNTs are of high quality, typically, with few defects. The rate of CNT synthesis can be high, producing around 20–100 mg/min [3]. The rate of synthesis has been shown to increase with increasing gas pressure. Although higher currents during the arc would be expected to produce more material, it has been shown that the deposit becomes fused at higher current, actually reducing the effective yield [3, 6]. As no metallic catalyst particles are required for the growth of MWCNTs, arc discharge tubes can be used without acidic treatment to remove the metal.

When metal catalysts (e.g., iron, nickel, or cobalt) are incorporated into the anode and evaporated along with the graphite, the nature of the deposits changes and SWCNTs can be synthesized. The deposit on the electrode is found to contain SWCNTs, MWCNTs, metal-filled MWCNTs, and other graphitic contaminants, while the soot contains MWNTs and SWNTs. In addition, in some situations, a "collar" is observed around the deposit containing mainly SWNTs (80 %) with diameters of around 1 nm. The exact nature of the growth products is determined strongly by the physical conditions during the arc such as the gas pressure, the carbon flux, the arc current, and the composition of the catalyst. Growth models have been proposed by a range of groups in order to try and explain the process [3, 6].



Synthesis of Carbon Nanotubes,
Fig. 1 Schematic diagram of an arc discharge system to synthesize CNTs

Laser Ablation

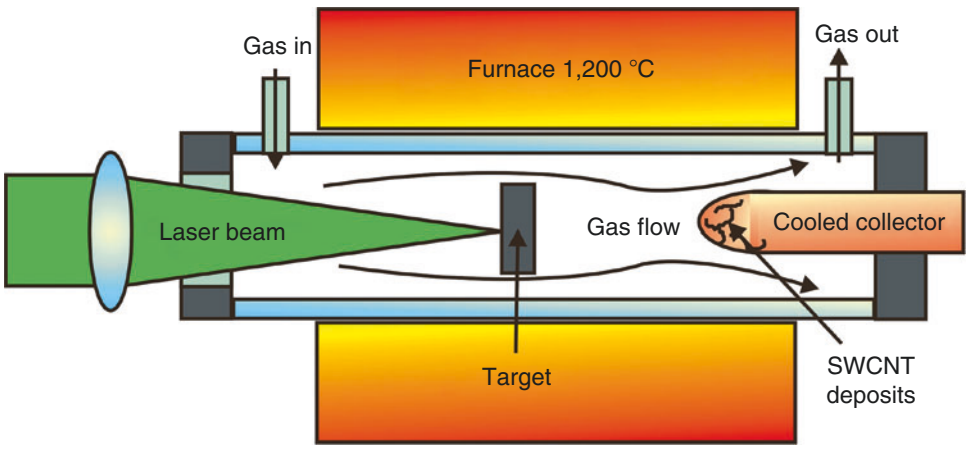
The high power densities produced by focusing a short-duration laser pulse onto a target lead to a very rapid heating of the material at the focal point. Subsequent thermal and other direct-sputtering processes ablate material from the target, and the interaction between the laser pulse and the ablated material forms a highly energetic, nonequilibrium, plasma plume [10]. The presence of ambient gas can have a dramatic effect on the expansion dynamics of these laser-ablated plasmas. At higher pressures plume confinement and plume heating occur. The high collision rates enhance conversion of the plume stream velocity into thermal energy, producing a confined, high-temperature plasma, with short mean-free paths for species. In this regime, the self-assembly of nanoclusters in the gas phase can occur. By ablating a target containing a mixture of the growth material and a suitable catalyst and by controlling the gas chemistry, nanomaterials such as CNTs can be self-assembled in the gas phase.

In the laser ablation CNT synthesis process (see Fig. 2), a pulsed laser vaporizes a graphite target held at a high temperature (typically 1,000–1,200 °C) inside a tube, within a furnace. An inert gas flows through the tube driving the ablation products toward one end. SWCNTs that grow in the gas phase are deposited in the cooler parts of the reactor as the vaporized carbon condenses. A water-cooled collector may be included in the system to collect the CNTs [3, 5, 9].

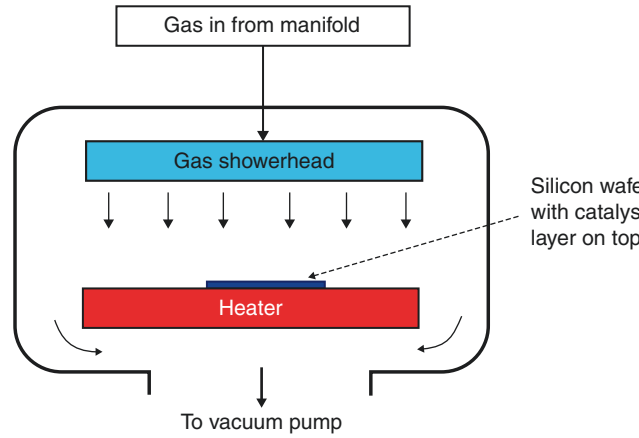
The laser ablation synthesis method was demonstrated by Smalley et al. [7, 11] where the growth of MWCNTs was demonstrated. Later this method was developed further to produce SWCNTs by ablating graphite targets mixed with metallic elements such as cobalt and nickel [3, 11–13]. The laser ablation method with a catalyst present has a yield around 70 % and produces primarily single-walled carbon nanotubes with a diameter determined by the temperature of the furnace. The quality of the SWCNTs is very high, although there is typically a significant quantity of catalyst nanoparticles present in the product, which may need to be removed by acid treatment producing additional defects.

Chemical Vapor Deposition

Chemical vapor deposition (CVD) is a vacuum deposition technique that allows the deposition of materials uniformly over large surface areas (typically flat surfaces) with high levels of purity. For this reason, CVD equipment (and its variants) is commonly found in semiconductor-manufacturing foundries and is used for depositing thin layers of materials on silicon wafers. For CNT growth, a catalyst is often used, in the form of a thin metal layer on top of the substrate, in



Synthesis of Carbon Nanotubes, Fig. 2 Schematic diagram of an arc discharge system to synthesize CNTs



Synthesis of Carbon Nanotubes, Fig. 3 Schematic of a CVD system

order to lower the temperature that is required for CNT growth [3, 8, 9]. This method is sometimes referred to as catalytic CVD (CCVD) in the literature due to the involvement of catalytic activity during the CVD growth process [3, 8]. In this method, the substrate plus catalyst is mounted on a heated stage, inside a vacuum chamber. A schematic of the CVD arrangement is shown in Fig. 3.

The temperature of the substrate is raised, typically between 400 and 1,000 °C in a low pressure of hydrogen gas. At elevated temperatures, the catalyst thin film breaks up spontaneously and forms nanoparticles on the surface of the substrate. A carbon-containing gas, typically a hydrocarbon such as acetylene or methane, is then bled into the chamber, together with a carrier gas, typically hydrogen. Hydrocarbon concentrations of typically between 1 % and 10 % are used. Upon contact with the catalyst nanoparticles, the hydrocarbon gas decomposes and releases carbon to the individual nanoparticles. The nanoparticles then extrude this carbon in the form of concentric graphitic sheets (the nanotube). A dynamic equilibrium is set up at the nanoparticles between the hydrocarbon decomposition rate and the growth speed of the CNTs, which ensures the continual growth of the CNTs. The exact growth mechanism and events that take place at the nanoparticles during growth are still the subject of debate [8].

Although commercial CVD equipment has been available for more than half a century for the deposition of traditional semiconductors and dielectrics, the development of systems solely for CNT growth is only recent, and commercial systems (e.g., Surrey NanoSystems' NanoGrowth system) have only been released to the market within the last 5 years. This is due to the more recent discovery of the techniques that allow for growth processes that are compatible with the already existing semiconductor infrastructure. In particular, techniques for reducing the growth temperatures of CNTs have been sought, as silicon processing only allows maximum operating temperatures of up to around 450 °C before degradation of the silicon diffusion layers starts to take place [14, 15].

Water-Assisted CVD

Another area that is still subject of debate is the mechanism by which CNT growth termination occurs when using CCVD techniques. It is noted that CNT growth stops abruptly sometime after growth initiation, typically after around 5–10 min. It is believed that this takes place due to the buildup of amorphous carbon at the surface of the catalyst particles, which effectively "poisons" the catalytic properties of the particles, inhibiting the dynamic equilibrium required for CNT growth. This theory was enlightened when samples of CNTs where growth had terminated were cleaned using an oxygen plasma to remove this amorphous carbon. It was noted that CNT growth could resume after performing this step.

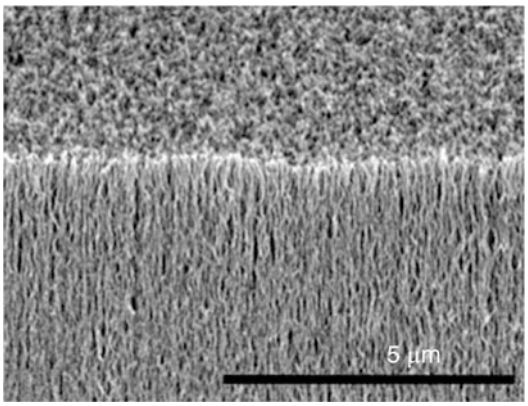
Despite this success, it was clear that introducing oxygen gas as a cleaning agent during CNT

growth (which involves hydrogen and hydrocarbon gas mix) would be unsafe. For this reason water vapor is currently used for this purpose, which although less effective than oxygen at removing amorphous carbon is compatible with the CNT growth gas mix. It is now well accepted that introducing a small amount of water vapor to the hydrogen and hydrocarbon gas mix reduces the growth termination effect (see Fig. 4). This allows growing CNTs in lengths up to several millimeters, reaching the centimeter scale, in a single run [16].

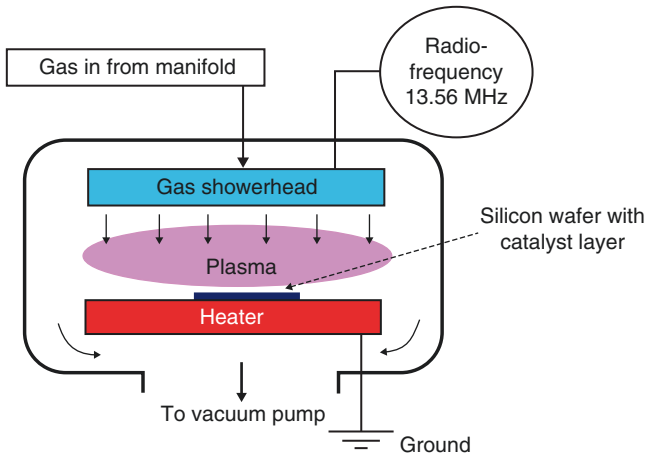
Plasma-Enhanced CVD

A further variation of the CVD process is plasma-enhanced CVD (PECVD) [17]. In this case, a plasma is struck between the gas showerhead and the substrate electrode (with the heater), by means of the application of a radio frequency to the showerhead electrode. This arrangement is depicted in Fig. 5. In this case, a fraction of the hydrocarbon gas decomposes in the plasma phase to produce smaller molecules and energetic species which provide additional energy at the catalytic nanoparticles, allowing a further reduction of the temperature of the substrate. Additionally, this plasma provides a strong electric field between the surface of the substrate and the plasma sheath (a region of the plasma that is located only a few millimeters above the substrate electrode). This electric field provides a vertically upward pull force on the nanoparticles, which results in a high degree of CNT verticality, as depicted in Fig. 6a. This allows for the formation of vertical CNT "forests" even when the CNT density is low. Although (PE)CVD can produce large quantities of CNTs, the quality of the material is lower than that produced by arc discharge and laser ablation, with many structural defects present in the tubes (see Fig. 6b).

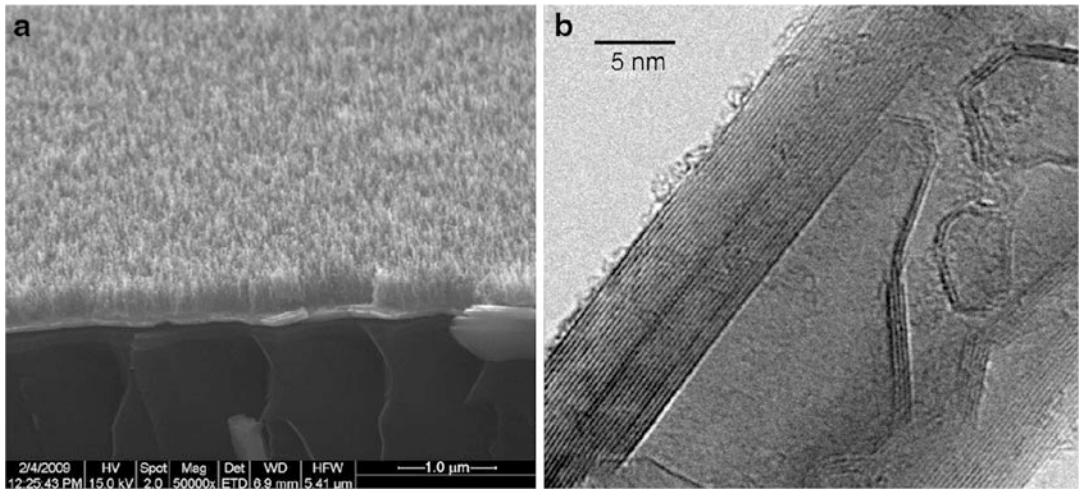
As for the case of CVD, PECVD process equipment is also common in silicon-manufacturing foundries and has been commercially available for many decades. However, specific equipment dedicated to CNT growth has only been available in recent years. This is partly because of specialized techniques that are involved in striking a stable plasma at the higher



Synthesis of Carbon Nanotubes, Fig. 4 SEM image of long CNTs grown by water-assisted CVD



Synthesis of Carbon Nanotubes, Fig. 5 Simplified schematic of a PECVD system



Synthesis of Carbon Nanotubes, Fig. 6 (**a**) SEM image of CNTs grown using the PECVD process, showing a high degree of verticality. (**b**) High-resolution TEM image of a CVD MWCNT showing internal structural defects and surface contaminants

gas pressures required for CNT growth, typically 5 Torr and above, up to several hundred Torr. These pressures are much higher than those typically associated with the common PECVD processes, of a few hundred millitorr.

Fluidized Bed CVD

The CVD process is a promising route for the bulk production of CNTs; however, producing truly industrial-scale volumes of material is challenging. Fluidized bed CVD (FBCVD) reactors are becoming one of the most widely targeted methods to achieve these scales. However, scale-up of the reactor is still a significant challenge, and this field has become an active area of research [18, 19]. In the FBCVD process, the CVD reaction occurs within a fluidized bed of catalyst particles. The setup consists of a reaction tube inside a cylindrical furnace. The contiguous mixing that occurs in the catalyst beds allows efficient utilization of all of the catalyst particles, which are all exposed to the feed gases. One such catalytic process, CoMoCAT®, is briefly introduced, but a detailed discussion of the mechanisms can be found in a recent review article [19].

The CoMoCAT® technique uses active cobalt stabilized in a nonmetallic state by interaction with molybdenum oxide (MoO_3) in a fluidized bed. When the mixture is exposed to carbon monoxide, the Co-Mo dual oxide is carburized, producing molybdenum carbide and small Co nanoparticles which are of very uniform sizes and well dispersed, allowing the growth of SWCNTs with small diameters and a high selectivity for certain chiral indices.

Floating Catalyst CVD

Similar to the fluidized bed CVD, a high degree of catalyst utilization can be achieved if the nanoparticles are kept suspended in the gas phase. If a narrow size distribution of small catalyst nanoparticles can be generated, growth of SWCNTs can be achieved. This idea leads to a simple method for production whereby the catalyst is introduced into a CVD reactor either by (1) a syringe process using the catalyst dissolved in a carbon source, (2) by sublimation of the catalyst at elevated temperatures, or (3) using a gaseous catalyst source, e.g., $Fe(CO)_5$. The catalyst and the carbon source are directly reacted in the gas phase. One such process that has attracted much attention is the HiPco technique, as it has been scaled to produce industrial quantities of SWCNTs. We discuss this process in more detail here.

The high-pressure carbon monoxide disproportionation process (HiPco) is a method for the production of SWCNTs using a continuous flow of high-pressure carbon monoxide as the carbon source and an iron carbonate $Fe(CO)_5$ as the source of catalyst nanoparticles [20]. This process was developed at Rice University by the group of the late R.E. Smalley. SWCNTs are produced by flowing CO, mixed with a small amount of $Fe(CO)_5$ through a heated reactor. The HiPco process produces SWCNTs with diameters of approximately 1.1 nm. The yield achieved is approximately 70 % SWCNTs with 97 % purity at rates of up to 450 mg/h [20].

Summary of Growth Mechanisms

In this entry the main routes to synthesize CNTs have been detailed. A summary of the important points is shown in Table 1.

Purification

The technical challenges in the production of usable CNTs often do not end when the material is first grown. After growth the yield and purity of the product is often not high enough for direct utilization; so, purification to remove unwanted contaminants such as metallic catalyst particles and non-CNT carbonaceous materials is required. The techniques employed typically are strong gaseous oxidation and chemical treatments such as acid refluxing, which will both have an effect (often detrimental) on the structure of the tubes [3]. Following the demonstration that carbon nanotubes could be attacked by oxidizing gases, Ebbessen et al. realized that the more defective carbon nanoparticle contaminants, such as amorphous carbon, would be oxidized more readily than the more perfect nanotubes. They found that a significant purification of CNTs could be achieved this way, but with a significant loss in yield [6]. For the chemical purification routes, a typical treatment method would involve (a) dispersal in organic solvent and filtration to remove large particulates, (b) treatment with concentrated acids to remove fullerenes and catalyst particles, (c) centrifugal separation, and (d) microfiltration. Such a chemical treatment can cause damage to the surface layer of MWCNTs and possible total destruction of SWCNTs, if too extreme.

Synthesis of Carbon Nanotubes, Table 1 Summary of the three main growth methods

Method	Arc discharge	Laser ablation	CVD
CNT type	*SWCNTs*: short nanotubes with diameters of around a nanometer	*SWCNTs*: formation of bundles of long (many microns) nanotubes, with diameters of a few nm	*SWCNTs*: growth of isolated tubes with diameters in the range of 1–4 nm
	MWCNTs: short nanotubes with inner diameters of a few nm and outer diameter of around 10 nm	*MWCNTs*: not typically grown by this method	*MWCNTs*: growth of long (many microns) nanotubes with diameters from a few nm up to hundreds of nm (carbon fibers)
Yield	30–90 %	Up to 70 %	Up to 100 %
Pros	Relatively simple equipment required and produces large quantities of SWCNTs or MWCNTs. CNTs produced are of a good quality, with few defects. Yield is reasonable	SWCNTs produced by this method have the highest quality of all the methods and diameter control is good. Yield is good	Can be grown over large areas and in large quantities (good for industrial scale-up). Diameters can be controlled relatively easily by size of catalyst. Low-temperature growth possible. Purity can be high
Cons	Size control is difficult and significant purification is often required	Expensive experimental setup (high-peak power lasers). Purification is often required	Nanotubes are highly defective typically

Cross-References

- ▶ Carbon Nanotube-Metal Contact
- ▶ Carbon Nanotubes
- ▶ Carbon Nanotubes for Chip Interconnections
- ▶ Functionalization of Carbon Nanotubes

References

1. Iijima, S.: Helical microtubules of graphitic carbon. Nature **354**, 56 (1991)
2. Radushkevich, L.V., Lukyanovich, V.M.: Zurn. Fisic. Chim. **26**, 88 (1952)
3. Harris, P.F.: Carbon Nanotube Science: Synthesis, Properties and Applications, 2nd edn. Cambridge University Press, Cambridge (2009)
4. Ebbesen, T.W. (ed.): Carbon Nanotubes, Preparation and Properties. CRC Press, Boca Raton (1996)
5. Saito, R., Dresselhaus, G., Dresselhaus, M.S. (eds.): Physical Properties of Carbon Nanotubes. World Scientific, Singapore (1998)
6. Ebbesen, T.W.: Carbon nanotubes. Annu. Rev. Mater. Sci. **24**, 235 (1994)
7. Guo, T., Nikolaev, P., Rinzler, A.G., Tomanek, D., Colbert, D.T., Smalley, R.E.: Self-assembly of tubular fullerenes. Phys. Chem. **99**, 10694–10697 (1995)
8. Kumar, M., Ando, Y.: Chemical vapor deposition of carbon nanotubes: a review on growth mechanism and mass production. J. Nanosci. Nanotechnol. **10**, 3739–3758 (2010)
9. Dresselhaus, M.S., Dresselhaus, G., Avouris, P. (eds.): Carbon Nanotubes: Synthesis, Structure, Properties, and Applications. Topics in Applied Physics, vol. 80. Springer, Berlin (2001)
10. Ashfold, M.N.R., Claeyssens, F., Fuge, G.M., Henley, S.J.: Pulsed laser ablation and deposition of thin films. Chem. Soc. Rev. **33**, 23–31 (2004)
11. Yakobson, B.I., Smalley, R.E.: Fullerene nanotubes: C-1,000,000 and beyond. Am. Sci. **85**, 324 (1997)
12. Guo, T., Nikolaev, P., Thess, A., Colbert, D.T., Smalley, R.E.: Catalytic growth of single-walled nanotubes by laser vaporization. Chem. Phys. Lett. **243**, 49–54 (1995)
13. Thess, A., Lee, R., Nikolaev, P., Dai, H., Petit, P., Robert, J., Xu, C., Lee, Y.H., Kim, S.G., Rinzler, A. G., Colbert, D.T., Scuseria, G.E., Tománek, D., Fischer, J.E., Smalley, R.E.: Crystalline ropes of metallic carbon nanotubes. Science **273**, 483 (1996)
14. Boskovic, B.O., Stolojan, V., Khan, R.U.A., Haq, S., Silva, S.R.P.: Large area synthesis of carbon nanofibres at room temperature. Nat. Mater. **1**, 165–168 (2002)
15. Chen, G., Jensen, B., Stolojan, V., Silva, S.R.P.: Growth of carbon nanotubes at temperatures compatible with integrated circuit technologies. Carbon **49**, 280–285 (2011)
16. Amama, P.B., Pint, C.L., McJilton, L., Kim, S.M., Stach, E.A., Murray, P.T., Hauge, R.H., Maruyama, B.: Role of water in super growth of single-walled carbon nanotube carpets. Nano Lett. **9**, 44–49 (2009)
17. Meyyappan, M., Delzeit, L., Cassell, A., Hash, D.: Carbon nanotube growth by PECVD: a review. Plasma Sources Sci. Technol. **12**, 205–216 (2003)
18. See, C.H., Harris, A.T.: A review of carbon nanotube synthesis via fluidized-bed chemical vapor deposition. Ind. Eng. Chem. Res. **46**, 997–1012 (2007)
19. MacKenzie, K.J., Dunens, O.M., Harris, A.T.: An updated review of synthesis parameters and growth mechanisms for carbon nanotubes in fluidized beds. Ind. Eng. Chem. Res. **49**, 5323–5338 (2010)
20. Nikolaev, P.: Gas-phase production of single-walled carbon nanotubes from carbon monoxide: a review of the hipco process. J. Nanosci. Nanotechnol. **4**, 307 (2004)

Synthesis of Functional Materials for Bone Regeneration

Oscar Castaño
Biomaterials for Regenerative Therapies, Institute for Bioengineering of Catalonia (IBEC), Barcelona, Spain

Synonyms

Bioactive materials; Osteogenic materials; Osteoinduction materials

Definitions

3D scaffold processing methods: These are techniques to implement 3D porous scaffolds in order to obtain proper mechanical properties, as well as suitable porosity, to be able to colonize the whole structure with cells. Examples include foaming, freeze-drying, sintering, salt leaching, rapid prototyping, electrospinning, etc., [1].

Angiogenesis: The creation of new blood vessels from preexisting ones. It is commonly accepted that cells cannot be further than 150–200 μm from the closest blood capillary to avoid inducing necrosis by hypoxia.

Bioactivity: The ability to trigger a specific biological response. In bone tissue regeneration, a material is usually considered to have bioactive ability when it is possible to carefully precipitate bone-like hydroxyapatite onto its surface after immersion in an aqueous serum-like solution. However, a serious controversy exists among the scientific community because this process leads to many false positive and false negative results [2].

Bioceramic: An inorganic ceramic material, a crystal, a glass or a glass-ceramic, that is intended to be in contact with living tissues. In bone regeneration, bioceramics are used to regenerate and repair living tissues damaged by disease or trauma. They play a key role in specific clinical applications in orthopedics and dentistry or drug delivery [3]. Typical bioceramics are hydroxyapatite, polymorphs α and β of tricalcium phosphate ($Ca_3(PO_4)_2$), and silicon-based (bioglass) and titanium-based bioactive glasses.

Biocompatibility: The ability of a material to be parenterally implanted with a suitable host response in developing a specific function.

Biodegradability: The ability to be naturally dissolved and transformed into new compounds by hydrolysis, oxidation, enzymes, bacteria, or osteoclasts [4].

Biodegradable biomaterial: A material that, once parenterally applied, can be naturally dissolved, or enzymatically degraded, and its byproducts released into the bloodstream. Ideally, a biodegradable material would also be eliminated by the human body by natural pathways and without toxicity. Examples are calcium phosphates, polylactic acid, and polyglycolic acid.

Bone: A strong and tough connective tissue that supports and protects the rest of the internal organs, allows the body to move, offers an ideal environment for blood cell formation, and acts as a store for salts (especially calcium and phosphates), among other functions [1]. The bone, as a hybrid natural construct, presents a complex and highly hierarchical organized structure [5]. Morphologically, it can be subdivided into two distinct types according to porosity and unit microstructure: the cortical bone (also known as compact or dense bone), which represents around 80 % of the total bone mass, and the trabecular bone (also called cancellous or spongy bone) with a high surface area and porosity enough for the bone marrow and blood vessels [6].

Bone is basically the combination of two main phases: An inorganic one, mainly formed by carbonated hydroxyapatite, which represents around 60 % of the total mass, and an organic constituent, mainly formed by collagen type I organized in fibril bundles and other minor proteoglycans, non-collagenous macromolecules. Finally, cells complete the organic composition, forming a highly hierarchical organized 3D network together.

Bone extracellular matrix: A complex collagen-based fibrous structure that includes proteins, proteoglycans, and other signals and structure molecules. It contains the necessary signals to guide bone cells in the restoration of structure and function of damaged or dysfunctional bone tissue by surface mineralization [7].

Chemotaxis: The ability to recruit a particular type of cells by chemical signals. In bone regeneration, special emphasis is given to the promotion of the transferring and participation of MSCs to the damaged site [8].

Confluency: Part of the culture surface that is covered by a monolayer of cells.

Extracellular matrix (ECM): In mammals, a collagen-based structure expressed by cells as one of their main functions that acts as guiding template support and biochemical signaling platform for the development of other functions such as cell adhesion, proliferation, cell-cell signaling, or differentiation. Composition is very variable depending on the mammal species and tissues, but in general, it is basically formed by collagen fibrils that provide other proteins (fibronectin, elastin, laminin, etc.), growth factors, glycosaminoglycans, and other physical and chemical signals.

Extracellular matrix signaling: A combination of chemical and physical stimuli related to the extracellular matrix, providing a three-dimensional extracellular microenvironment. Physical stimuli include topography; mechanical stimuli include stiffness or external loads; and chemical stimuli include molecules, usually proteins or steroids that can be plasma soluble or

non-soluble, usually related the extracellular matrix by some type of bond.

Growth factor: Signaling protein or hormone that is able to regulate cell response by matching with the proper receptor on the cell membrane.

Instructive material: An engineered device providing the required signals on the surface or in the body of a material to be able to interact and control cell tissue response in order to promote bone healing [7]. Signals can be chemical (chemical gradients, ion and biomolecule release, functional groups on the surface, etc.), physical (roughness, anisotropy, topography gradients, engineered nanofeatures), or mechanical (stiffness, strain, external active load, mechanical gradients, etc.).

Osteoblasts: Mononuclear cells whose origin is differentiation from mesenchymal progenitor cells. They are responsible for the nucleation, growth, and mineralization of a new bone by the control of the local extracellular matrix ionic calcium and phosphate concentrations. They modulate hydroxyapatite formation using a previously synthesized extracellular matrix, which acts as a template and is mainly formed by collagen I and several types of glycoproteins and chemical and physical signals. At the end of the process, osteoblasts auto-mineralize their membrane and are trapped alive in the neophyte bone, becoming an osteocyte [9].

Osteoclasts: Multinucleated cells whose main function is the degradation and resorption of the bone through local acidification by an enzymatic reaction.

Osteoconduction: The ability of a surface to act as template and allow the growth of bone tissue by facilitating attachment, proliferation, migration, and osteoblast differentiation to form new tissue [10].

Osteocyte: Mineralized and differentiated osteoblasts embedded within the bone matrix. They are smaller than osteoblasts and have lost many of their cytoplasmic organelles [6].

Osteogenesis: The ability to create new bone tissue.

Osteoinduction: The ideal osteoinductive material should be able to induce ectopic bone synthesis when implanted into extra-skeletal body places; in other words, it should be osteogenic in soft tissues [11].

Scaffold: A 3D structure that acts as a guidance template for bone regeneration and generally involves macroporosity ($Ø > 50$ μm) to allow cell migration and vascular colonization, micro- and nanoporosity for permeability and, if required, better degradation.

Surface modification: Engineering processes that involve determined change in the surface of a material in order to properly interact with the host tissue cells.

Vasculogenesis: The process that involves the creation of new blood vessels by the new formation of endothelial cells from mesoderm cell precursors [12].

Introduction

Bone tissue regeneration has been one of the driving forces in the development of biomaterials and surface functionalization for the production of instructive materials. The knowledge acquired has formed a basis for assays in other tissues.

Functional materials for bone tissue date from the 1980s, emerging particularly in second-generation biomaterials [13]. Functional materials for bone tissue substitution or regeneration require several features related to mechanical and bioactive characteristics whose main goal is to guide tissue growth. This has enabled the development of a third generation of biomaterials and has allowed the field to advance toward short- to mid-term expectations for smart biomaterials and instructive scaffolds to guide the growth of bone tissue, as well as other human tissues [14, 15].

The tendency nowadays in functional biomaterials is to mimic what is known as extracellular matrix (ECM), including its dynamic cell control characteristics and hierarchical organization [14]. Osteoid, for example, is the result of a process involving osteoblast mineralization of an ECM. Depending on the degree of mineralization, osteoid Young's modulus can range from ∼27 kPa to ∼1 GPa, this value being characteristic of mature bone [16]. Today, a controversy exists about the mechanical rigidity of functional biomaterials for bone regeneration: should the initial graft be as rigid as bone, or

should the final result be as stiff as bone, while the biomaterial mimics less rigid, osteoid characteristics?

What seems to be clear is that the design of the material should be tailored to the final application. There are many different bones within the body, and their architecture and composition, even when similar, depend on their specific role and function. This is why there is no universal bone substitute but rather particular designs for the development of specific functions.

Bone

The bone is a hard connective tissue that acts as a protector of other internal organs, contributes to movement support, stores calcium and phosphates, and provides a porous environment for blood cell dynamics, among other specific functions [17]. Bones can be morphologically divided into flat, long, irregular, and short bones [6]. In addition, based on their microstructure, they can also be divided into two main families: cortical or compact bone (dense, low porosity, and void spaces, representing about 80 % of the total bone mass normally located in the outer shell of bone structures for mechanical support) and trabecular, cancellous, or spongy bone [6]. Compact and cancellous bones have a similar hierarchical composition and organized structure [5]. It is mainly formed from a mineral phase (carbonated hydroxyapatite, among other minor compounds) and an organic phase (collagen, among others) both contributing to its exceptional mechanical properties that combine rigidity and load-bearing strength with elasticity and flexibility [5]. Parameters such as composition, crystallinity, constituent morphology, 3D conformation, and distribution vary depending on the bone and its specific function within the body.

3D Scaffolds to Control Cell Fate

It is well known that cells need a material template with the correct 3D chemical and physical guidance to develop their functions, and 3D scaffolds are the basis for tissue engineering approaches [18]. Such scaffolds' architecture and, in particular, their surface features are directly responsible for governing cell fate by affecting cell functions. Therefore, the design of 3D architectures at the nanoscale which are able to specifically interact with cells' receptors is of utmost importance. The strict biocompatibility requirements can rule out about 90 % of available materials. Once the biological mechanisms involved in the regenerative process are identified, the generated knowledge can lead to the design of a biomaterial specifically tailored toward the control of stem cell fate, thus producing the desired phenotype [19].

Scaffold Design

The design of scaffolds requires:

- The correct mechanical properties to maintain the scaffold's structure during its life.
- A shape that can easily adapt to the damaged cavity.
- Optimal biocompatibility.
- Degradation at the same rate as new bone tissue growth.
- Sterilizability, without undesirable degradation or cross-linking.
- Enough porosity to allow cell and blood vessel colonization, the supply of nutrients and oxygen, and waste elimination. Porosity is measured at three levels: macro-, meso-, and microporosity. Macroporosity refers to pores with a width larger than 50 nm, mesoporosity to those between 2 and 50 nm, and microporosity to those with a width smaller than 2 nm [20].
- Optimal mass and fluid transport through the scaffold and good interconnections between pores.
- Be sterilizable.

Macroporosity is a crucial factor that directly controls cell behavior, bone growth and vascularization, as well as the mechanical properties of scaffolds. The optimal pore size required for bone regeneration depends on the cell lineage, differing efficiency depending on the upper limit of functionality [21]. The porosity of the material is a

Synthesis of Functional Materials for Bone Regeneration, Table 1 List of different sequences with bioactivity effect on cells

Peptide sequence	Derived from	Proven function
RGD	Multiple proteins	Relevant to bone engineering; RGD enhances cell adhesion and differentiation into bone, cartilage, neural, and endothelial tissue
FHRRIKA	Heparin-binding domain	Increases osteoblast adhesion and mineralization
KRSR	Heparin-binding	Increases osteoblast adhesion and mineralization
YIGSR	Laminin	Increases human foreskin fibroblast adhesion
IKVAV	Laminin	Increases neurite extension
REDV	Fibronectin	Increases endothelial cell adhesion
KHIFSDDSSE	Neural cell adhesion molecules	Increases astrocyte adhesion
VPGIG	Elastin	Increases stiffness of synthetic matrices
SVSVGMKPSPRP	Phage display	High selectivity toward hydroxyapatite and tooth enamel
VTKHLNQISQSY	Phage display	High selectivity toward hydroxyapatite and bone-like mineral (BLM)
CRKRLDRNC	Phage display	Interacts with IL-4 receptor on endothelial cells, macrophages, and smooth muscle cells
THRTSTLDYFVI	Phage display	High selectivity toward polypyrrole that increases adhesion of PC12 cells
STFTKSP	Phage display	Homes to primitive hematopoietic progenitor cells in bone marrow
WYRGRL	Phage display	Binds to collagen IIα1
ASSLINA	Phage display	Binds to both skeletal and cardiac tissue
CAGALCY	Phage display	Targets brain tissue

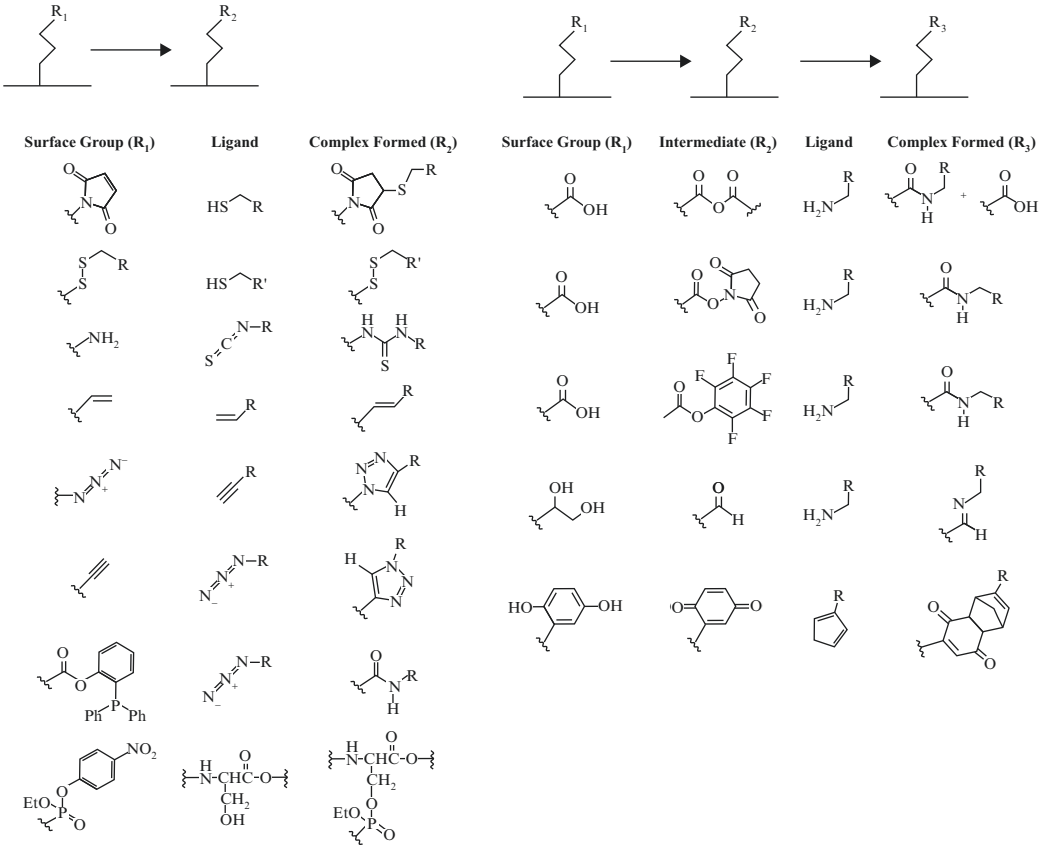
Adapted from [28] with kind permission from Springer Science and Business Media

delicate balance between biological behavior and mechanical properties, as the higher the void volume, the better the vascularization and the osteointegration, but the worse the scaffold's resistance to mechanical failure [22]. The interconnectivity of pores is essential for bone regeneration as it enables the infiltration of bone, the development of an efficient network of blood vessels, and the promotion of cell-cell interactions [23]. Specific surfaces also help to have a better understanding of the kinetic degradation of the construct, minimizing usually undesired diffusion control, and this is closely related to the degree of porosity.

On the other hand, the surface morphology also affects cellular response in the form of small pores, patterns, or roughness. Adhesion is one of the favored processes and depends on the design of surface [24]. The strength of the anchorage of cells further determines aspects such as migration, proliferation, or differentiation. Adhesion is one aspect that creates more controversy among scientists, as weak adhesion leads to cell apoptosis, and very strong adhesion leads to early differentiation before the cells proliferate. As a result, it seems that a medium adhesion strength is ideal, allowing proliferation until the point when a cell will differentiate, for example, and/or a dense confluence (another crucial factor to describe cell behavior is the cell-to-cell communication). The surface can be modulated by texturing and patterning using different methods [25]: blasting, electropolishing, chemical treatments [26], lithography, plasma treatment, and focused ion beam, among others. This can positively influence cell response in terms of adhesion detachment, spreading, migration, and proliferation, which are better at guiding cell behavior.

The nature of the functional groups present on the surface can also be relevant in modulating cell response. Aspects such as hydrophilicity, surface charge, and the surface energy are directly related to the chemical composition of the interface. On the other hand, bioactive biomolecules can also be

Synthesis of Functional Materials for Bone Regeneration, Table 2 Common strategies in surface covalent functionalization: (left) direct link; (right) link through intermediates



Adapted with permission from [29]. Copyright 2015 American Chemical Society

added on the surface in order to induce a biological response in cells. This mainly relies on the use of ECM proteins (collagen, fibronectin, laminin, etc.), peptide sequences that are known to affect specific functions (RGD, YIGSR, REDV, IKVAV, KRSR, etc.), growth factors (BMP, BMP2, VEGF, TNF-α, etc.), drugs (antibiotics, analgesics, antitumor agents, biphosphonates, etc.), dyes (rhodamine, fluorescein, etc.), or bioactive nanoparticles (hydroxyapatite, bioglasses, ormoglasses, etc.) [27]. The strength of the surface binding would depend on the chosen strategy; the biomolecule can remain on the surface once the cell attaches or, alternatively, be released to the media to be detected by the cell (Table 1).

The binding of the biomolecule to the surface of the scaffold can be weak or strong, by physiadsorption or chemiadsorption. In the second case, there are many options depending on the type of surface being functionalized: EDC/NHS-based chemistry, click chemistry, self-assembled monolayers, electrostatic union, etc. [29, 30] (Table 2).

Many processing methods have been developed to shape biomaterials depending on the application. No particular one is perfect, but all provide special feature that can be brought into play for specific roles. Foaming allows the possibility of obtaining a very high porosity, but pore window junctions can be an unsolved problem for

cell migration; sintering is an option only available for crystalline ceramics; salt leaching is an economical way of obtaining high porosity and good interconnectivity, but with poor mechanical properties and thick, slowly degrading walls; rapid prototyping also results in high porosity combined with an optimized mechanical resistance, but also has the problem of thick walls in the range of tens of micrometers; electrospinning is the cheapest method to mimic ECM and one of the methods preferred by cells, but with a poor macroporosity that makes colonization difficult.

Cross-References

- Angiogenesis
- Bioderived Smart Materials
- Bioinspired Synthesis of Nanomaterials
- Biomimetics
- Bone Remodeling
- Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- Electrospinning
- Nanoengineered Hydrogels for Cell Engineering
- Nanopatterned Surfaces for Exploring and Regulating Cell Behavior
- Nanostructured Functionalized Surfaces
- Nanostructures for Surface Functionalization and Surface Properties
- Nanotechnology
- Surface Modeling of Ceramic Biomaterials

References

1. Sachot, N., Engel, E., Castaño, O.: Hybrid organic–inorganic scaffolding biomaterials for regenerative therapies. Curr. Org. Chem. **18**(18), 2299–2314 (2014)
2. Bohner, M., Lemaitre, J.: Can bioactivity be tested in vitro with SBF solution? Biomaterials **30**, 2175–2179 (2009). doi:10.1016/j.biomaterials.2009.01.008
3. Vallet-Regi, M.: Bio-Ceramics with Clinical Applications. Southern Gate, Chichester, West Sussex, UK (2014)
4. Bandyopadhyay, A., Bernard, S., Xue, W., Böse, S.: Calcium phosphate-based resorbable ceramics: influence of MgO, ZnO, and SiO 2 dopants. J. Am. Ceram. Soc. **89**, 2675–2688 (2006). doi:10.1111/j.1551-2916.2006.01207.x
5. Clarke, B.: Normal bone anatomy and physiology. Clin. J. Am. Soc. Nephrol. **3**, 131–139 (2008). doi:10.2215/CJN.04151206
6. Pérez-Amodio, S., Engel, E.: Bone biology and regeneration. In: Bio-Ceramics with Clinical Applications, pp. 315–42. Southern Gate, Chichester, West Sussex, UK (2014). doi:10.1002/9781118406748.ch11
7. Lutolf, M.P., Hubbell, J.A.: Synthetic biomaterials as instructive extracellular microenvironments for morphogenesis in tissue engineering. Nat. Biotechnol. **23**, 47–55 (2005)
8. Sordi, V.: Mesenchymal stem cell homing capacity. Transplantation **87**, S42–S45 (2009). doi:10.1097/TP.0b013e3181a28533. [pii] r00007890-200905151-00004
9. Bose, S., Tarafder, S.: Calcium phosphate ceramic systems in growth factor and drug delivery for bone tissue engineering: a review. Acta Biomater. **8**, 1401–1421 (2012). doi:10.1016/j.actbio.2011.11.017
10. Vacanti, C.A., Pietrzak, W.S.: Musculoskeletal Tissue Regeneration: Biological Materials and Methods. Humana Press, Totowa, New York (2008)
11. Miron, R.J., Zhang, Y.F.: Osteoinduction: a review of old concepts with new standards. J. Dent. Res. **91**, 736–744 (2012). doi:10.1177/0022034511435260
12. Risau, W., Flamme, I.: Vasculogenesis. Annu. Rev. Cell Dev. Biol. **11**, 73–91 (1995). doi:10.1146/annurev.cb.11.110195.000445
13. Dorozhkin, S.V.: Calcium orthophosphates and human beings: a historical perspective from the 1770s until 1940. Biomatter **2**, 53–70 (2012)
14. Holzapfel, B.M., Reichert, J.C., Schantz, J.-T., Gbureck, U., Rackwitz, L., Nöth, U., et al.: How smart do biomaterials need to be? A translational science and clinical point of view. Adv. Drug Deliv. Rev. **65**, 581–603 (2013). doi:10.1016/j.addr.2012.07.009
15. Castano, O., Sachot, N., Xuriguera, E., Engel, E., Planell, J.A., Park, J.-H., et al.: Angiogenesis in bone regeneration: tailored calcium release in hybrid fibrous scaffolds. ACS Appl. Mater. Interfaces (2014). doi:10.1021/am500885v
16. Engler, A.J., Sen, S., Sweeney, H.L., Discher, D.E.: Matrix elasticity directs stem cell lineage specification. Cell **126**, 677–689 (2006)
17. Bandyopadhyay-ghosh, S.: Bone as a collagen-hydroxyapatite composite and its repair. Trends Biomater. Artif. Organs **22**, 116–124 (2008)
18. Stapor, P.C., Azimi, M.S., Ahsan, T., Murfee, W.L.: An angiogenesis model for investigating multi-cellular interactions across intact microvascular networks.

Am. J. Physiol. Heart Circ. Physiol. **304**, H235–H245 (2013). doi:10.1152/ajpheart.00552.2012
19. Provenzano, P.P., Keely, P.J.: Mechanical signaling through the cytoskeleton regulates cell proliferation by coordinated focal adhesion and Rho GTPase signaling. J. Cell Sci. **124**, 1195–1205 (2011). doi:10.1242/jcs.067009
20. Rouquerol, J., Avnir, D., Fairbridge, C.W., Everett, D. H., Haynes, J.M., Pernicone, N., et al.: Recommendations for the characterization of porous solids. Pure Appl. Chem. **66**, 1739–1758 (1994)
21. O'Brien, F.J., Harley, B.A., Yannas, I.V., Gibson, L.J.: The effect of pore size on cell adhesion in collagen-GAG scaffolds. Biomaterials **26**, 433–441 (2005). doi:10.1016/j.biomaterials.2004.02.052
22. Hollister, S.J.: Porous scaffold design for tissue engineering. Nat. Mater. **4**, 518–524 (2005). doi:10.1038/nmat1421
23. Valerio, P., Guimaráes, M.H.R., Pereira, M.M., Leite, M.F., Goes, A.M.: Primary osteoblast cell response to sol–gel derived bioactive glass foams. J. Mater. Sci. Mater. Med. **16**, 851–856 (2005). doi:10.1007/s10856-005-3582-5
24. Dalby, M.J., Gadegaard, N., Tare, R., Andar, A., Riehle, M.O., Herzyk, P., et al.: The control of human mesenchymal cell differentiation using nanoscale symmetry and disorder. Nat. Mater. **6**, 997–1003 (2007). doi:10.1038/nmat2013
25. Gadegaard, N., Dalby, M.J., Martines, E., Seunarine, K., Riehle, M.O., Curtis, A.S.G., et al.: Nano patterned surfaces for biomaterial applications. Adv. Sci. Technol. **53**, 107–115 (2006)
26. Montanaro, L., Arciola, C.R., Campoccia, D., Cervellati, M.: In vitro effects on MG63 osteoblast-like cells following contact with two roughness-differing fluorohydroxyapatite-coated titanium alloys. Biomaterials **23**, 3651–3659 (2002)
27. Sachot, N., Castaño, O., Mateos-timoneda, M.A., Engel, E., Planell, J.A.: Hierarchically engineered fibrous scaffolds for bone regeneration hierarchically engineered fibrous scaffolds for bone regeneration. J. R. Soc. Interface **10**, 1–5 (2013)
28. Segvich, S., Kohn, D.: Phage display as a strategy for designing organic/inorganic biomaterials. In: Puleo, D. A., Bizios, R. (eds.) Biological Interactions on Materials Surfaces SE – 6, pp. 115–32. Springer, New York (2009). doi: 10.1007/978-0-387-98161-1_6
29. Love, J.C., Estroff, L.A., Kriebel, J.K., Nuzzo, R.G., Whitesides, G.M.: Self-assembled monolayers of thiolates on metals as a form of nanotechnology. Chem. Rev. **105**, 1103–1170 (2005). doi:10.1021/cr0300789
30. Conde, J., Dias, J.T., Grazú, V., Moros, M., Baptista, P. V., de la Fuente, J.M.: Revisiting 30 years of biofunctionalization and surface chemistry of inorganic nanoparticles for nanomedicine. Front. Chem. **2**, 48 (2014). doi:10.3389/fchem.2014.00048

Synthesis of Gold Nanoparticles

Munish Chanana[1], Cintia Mateo[1], Verónica Salgueirino[2] and Miguel A. Correa-Duarte[1]
[1]Departamento de Química Física, Universidade de Vigo, Vigo, Spain
[2]Departamento de Física Aplicada, Universidade de Vigo, Vigo, Spain

Synonyms

Production of gold nanoparticles

Definition

Description of different synthetic approaches for the fabrication of gold nanoparticles.

Overview

Since the beginning of recorded history, gold has always held the majestic throne among all the noble metals [1, 2]. It has been known by artisans since the Chalcolithic (Copper) Age and has become a highly coveted metal for coinage, jewelry, and other arts since that time. The first extraction of gold has been dated back to the fifth millennium B.C. in Bulgaria, but "soluble" gold (colloidal gold) first appeared around the fifth century B.C. in Egypt and China [1, 2]. Colloidal gold was used to make ruby glass and to color ceramics. Perhaps the most famous examples are the Lycurgus Cup (manufactured around fifth/fourth century B.C., exhibited in British Museum) and the pigment "Purple of Cassius" (invented by Andreas Cassius, seventeenth century) [1, 2]. But scientific research on gold sol started with Michael Faraday. In 1857, Faraday reported the formation of deep red solutions of colloidal gold by the reduction of an aqueous solution of chloroaurate ($HAuCl_4$) using phosphorus in CS_2 (a two-phase system) [1, 2].

The most relevant properties of gold colloids are based on the presence of a strong absorption band in the visible-NIR, which is the origin of the observed brilliant red/purple colors of certain gold nanoparticles in solution. This absorption band results from the collective oscillation of the conduction band electrons in resonance with the frequency of the incident electromagnetic field and is known as surface plasmon resonance (SPR) absorption. The SPR frequency and thus the color of the gold NP mainly depends on the particle size, shape, the nature of the surrounding medium, and the interparticle distance. The influence of shape and interparticle distance is in general even greater than that of size [1, 2]. While a single absorption band is present for spherically symmetric gold particles, multiple absorption bands correlated with their various axes appear for nonspherical ones. Such structures can support both propagating and localized surface plasmons. For instance, gold nanorods possess two different resonance modes due to electron oscillation across and along the long axis of the nanorod and are commonly labeled the transverse and longitudinal modes, respectively, the latter of which is extremely sensitive to the aspect ratio of the rod [3]. Gold nanotriangles, for instance, exhibit three different resonances in the UV–Vis absorption spectra, one SPR absorption out of plane and two in-plane SPR absorptions at longer wavelengths [4, 5]. In the case of branched, platonic, or platelet nanostructures, the plasmon resonance shifts to longer wavelengths with respect to the common resonance of spherical nanoparticles at 520 nm [4–6]. Special attention requires the case of gold nanoshells which, usually due to their synthetic approach, are composed of a dielectric core coated with a thin metallic layer. Although these structures have similar properties to spherical gold nanoparticles in terms that they only present a single SPR absorption, they offer, however, the ability of tuning the SPR over a full range of wavelengths from the visible to the infrared region [4–6].

Since particle shape has a tremendous effect on the physical, chemical, optical, electronic, and catalytic properties of nanoparticles, gold nanoparticles have been synthesized in various shapes, including rods, cubes, plates, polyhedrons, and wires following different and ingenious techniques [1–8]. In the literature, there are multiple excellent reviews [1–6] and books [7, 8] regarding the synthesis of gold nanoparticles of different sizes and shapes, where their different physical properties (e.g., optical, electronic, or mechanical) are also discussed. Therefore, in this chapter, we provide a comprehensive overview on the major wet-chemistry synthetic approaches for the preparation of gold nanoparticles attending to the particle morphology. Hence, the main criterion on which this chapter is structured is the shape of the nanoparticles, predominantly according to the isotropy of the particles.

Isotropic Gold Nanoparticles

Spheres

Sphere is the most thermodynamically favored shape for all kinds of colloidal particles synthesized in bulk solutions. Spherical gold nanoparticles can be synthesized in solution in various sizes, ranging from 1 nm up to several hundreds of nanometers, and with different capping agents. The capping agent mainly dictates the chemical and physicochemical behavior of the particles, and this in turn determines the employment of the particles [1, 2].

One of the most popular synthetic methods for the preparation of gold nanospheres is based on the reduction of $HAuCl_4$ by citrate in water, which was first described in 1951 by Turkevich [1, 2]. In this method, citrate serves as both a reducing agent and an anionic stabilizer. It yields 15 nm nanoparticles with a fairly narrow size distribution. The Turkevich method has been modified by a number of groups to produce spherical gold nanoparticles with diameters ranging from 15 to 150 nm by either controlling the ratio of citrate to $HAuCl_4$ or employing a γ-radiation method. However, nanoparticles larger than 20 nm synthesized by this method usually lack isotropy and size uniformity [1, 2].

Other reduction methods have also been developed to achieve a better control over the size and

monodispersity, including the "Schmid method" published in 1981 and the "Brust-Schiffrin method" reported in 1994 [1, 2]. The Schmid's cluster $[Au_{55}(PPh_3)_{12}Cl_6]$ remained unique for a long period of time with its narrow dispersity (1.4 ± 0.4 nm) for the study of a quantum-dot nanomaterial, despite its arduous synthesis. The synthesis requires rigorously anaerobic conditions and diborane gas as a reducing agent. As a result, phosphine-stabilized gold nanoparticles have lost favor since the development of a more convenient, scalable synthesis of thiol-stabilized air-stable gold nanoparticles of reduced polydispersity and controlled size by the method of Brust et al. [1, 2].

In the Brust-Schiffrin method, the synthesis is carried out in a two-phase system and thiol ligands that strongly bind to gold $AuCl_4^-$ are transferred to toluene using tetraoctylammonium bromide as the phase-transfer reagent and reduced by $NaBH_4$ in the presence of dodecanethiol. The organic phase changes color from orange to deep brown within a few seconds upon addition of $NaBH_4$. The particle size ranges between 1.5 and 5.2 nm. Additionally, these gold NPs can be repeatedly isolated and dispersed in common organic solvents without any aggregation or decomposition. Using the principle of this synthesis method, modified procedures have been developed to synthesize gold nanoparticles in the size range of 4–10 nm capped with thiol-functionalized molecules (small molecules or polymers) in one phase system using methanol as the solvent [1, 2].

The use of microemulsions, copolymer micelles, and inverse micelles is a significant research field for the synthesis of stabilized gold NPs. Typically, these syntheses involve a two-phase system with a surfactant that causes the formation of the microemulsion or the micelle maintaining a suitable microenvironment of controlled dimensions. The surfactants create small pockets of a water phase in an organic solvent or vice versa where the surfactant faces the aqueous phase with its polar group and the tail faces the organic phase. Varying the ratio of the two solvents affects the dimensions of the micelles, which allows for tuning the size of the resulting nanoparticles with good monodispersity [1, 2].

All of the methods described above lead to uniform nanoparticles, however, only in the size range of below 20 nm. For the synthesis of larger spherical NPs (Fig. 1), the seeded-growth procedure is the most popular technique that has been intensively studied in recent years [1, 2]. A strong chemical reducing agent (usually sodium borohydride) is used to generate small, generally spherical, nanoparticles (commonly denoted as *seeds*), which are then added to a growth solution composed of surfactants and Au metal ions to induce particle growth. The growth solution employs a weaker reducing agent (often ascorbic acid) to reduce the metal salt to an intermediate state so that only catalyzed reduction on the nanoparticle surface is allowed, avoiding secondary nucleation. During the growth process, usually a small amount of undesired nonspherical particles is also generated, which can be removed through a CTAB-assisted shape-selective separation method [1, 2, 4]. The gold nanospheres obtained after such purification can be further grown by means of the same procedure described here up to a diameter of 200 nm or more. The final size of the nanospheres synthesized by this method depends strongly on the size of the employed seeds and the ratio between the seeds and the gold salt. To achieve predetermined final sizes of the nanospheres, the concentrations can be adjusted using the following equation [9]:

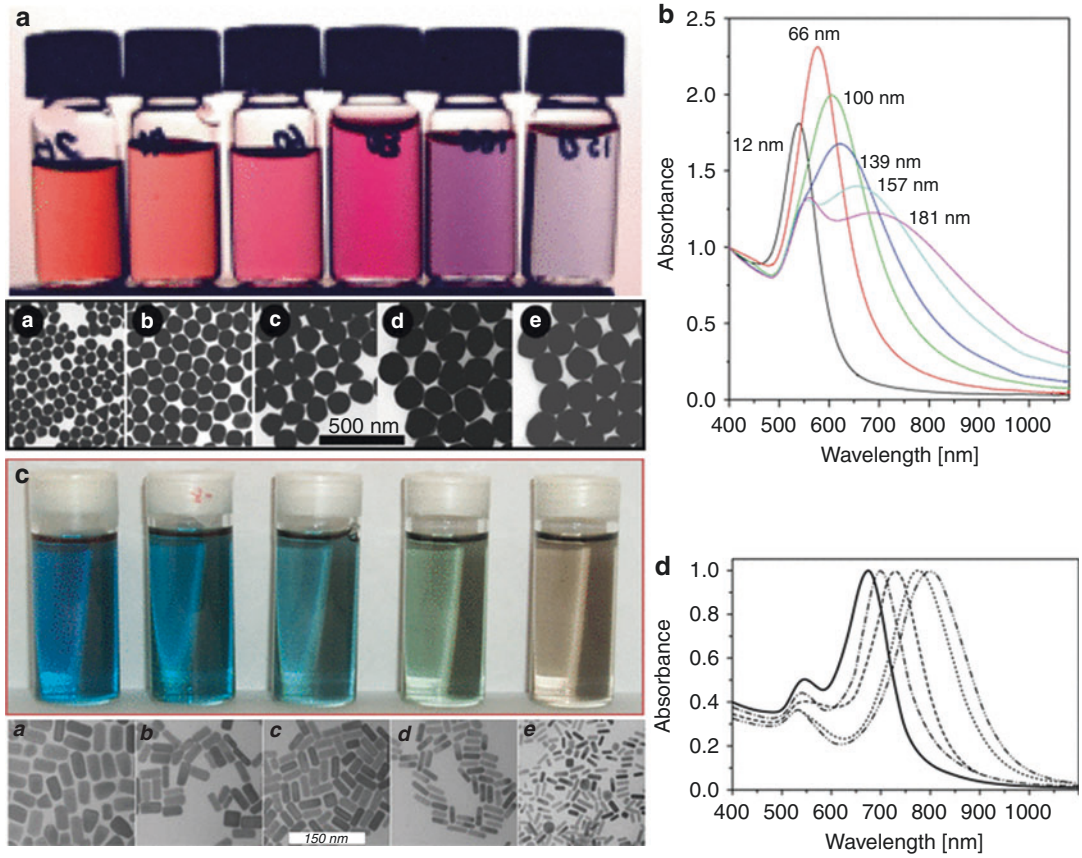
$$R_{\text{final}} = R_{\text{seeds}} \left(\frac{[Au^{3+}]_{\text{final}} + [Au_{\text{seeds}}]_{\text{final}}}{[Au_{\text{seeds}}]_{\text{final}}} \right)^{\frac{1}{3}}$$

where $[Au_{\text{seeds}}]_{\text{final}}$ is the value to be calculated:

$$[Au_{\text{seeds}}]_{\text{final}} = -\frac{[Au^{3+}]_{\text{final}}}{1 - (R_{\text{final}}/R_{\text{seeds}})^3}$$

Shells

Nanoshells have similar properties to spherical gold nanoparticles in the sense that they also exhibit a single surface plasmon resonance (SPR) absorption. However, the SPR absorption band of nanoshells can be tuned across the visible and infrared region over a range of wavelengths spanning hundreds of nanometers, far exceeding



Synthesis of Gold Nanoparticles, Fig. 1 (**a**) Representative TEM micrographs of Au spheres (*bottom*) obtained after subsequent growth steps with their respective color solution (*top*). Average diameters are 66 (*a*), 100 (*b*), 139 (*c*), 157 (*d*), and 181 (*e*) nm. (**b**) UV–Vis spectra of Au spheres with various average diameters (*right*) (Reprinted with permission from Ref. [24]. Copyright 2006 American Chemical Society.) (**c**) Photographs (*top*) and the respective TEM micrographs (*bottom*) of gold nanorod solutions with average aspect ratios of 1.94 (*a*), 2.35 (*b*), 2.48 (*c*), 3.08 (*d*), and 3.21 (*e*). (**d**) UV–Vis absorption spectra of different gold nanorod samples stabilized with CTAB with increasing aspect ratio from *left to right* (Reprinted with permission from Ref. [3, 25]. Copyright 2004 Elsevier B.V.)

the spectral range of spherical particles. This has a huge advantage for SERS-based applications, because the plasmon resonance can be tuned to the excitation of common laser radiation sources optimizing the electromagnetic enhancement mechanism [5, 10, 11]. Apart from their optical properties, these nanostructures also present inert biological activities which make them appropriate for biomedical applications, such as targeted drug delivery, photothermal therapeutic applications, and molecular imaging, as contrast agents for optical coherence tomography (OCT).

Nanoshells are usually composed of dielectric cores (e.g., silica particles) coated with a single or multiple thin metal layers [10]. The growth of metal nanoshells on core particles combines techniques of molecular self-assembly with the reduction chemistry of metal colloid synthesis. This approach is general and can potentially be adapted to a variety of core and shell materials. The common synthesis route of gold nanoshells involves the synthesis of the dielectric core material, typically silica or polystyrene, and their surface functionalization with terminal amine groups to facilitate attachment of small colloidal gold, which are usually formed in a separate process and act as nucleation sites for the subsequent gold salt reduction and shell growth. The absorbed

seed colloids increase in size as reduction ensues, followed by the coalescence of gold particles on the surface, until finally the apparent formation of a polycrystalline continuous metallic nanoshell occurs [10].

Although this common two-step process of seeding, i.e., formation of small colloidal gold followed by attachment to the core, is effective, it is also laborious and costly, especially in terms of time. Another alternative and also very effective approach is the deposition-precipitation (DP) method, which employs the in situ formation of gold seeds directly on the support. The functionalized core particles are placed in contact with a basic solution of gold (III) chloride, and upon heating, oxidic precursor particles ($Au[OH]_3$) are formed on the support. These precursor seed particles then act as nucleation sites for the formation of gold nanoshells. The size and density of the seeds on the substrate is easily controlled by the concentration of gold (III) chloride, support surface functionalization, reaction temperature and time, and the pH of reaction. This DP method is relatively easy to handle and is currently used for producing commercial gold catalysts in a highly controlled manner [11].

Anisotropic Gold Nanoparticles

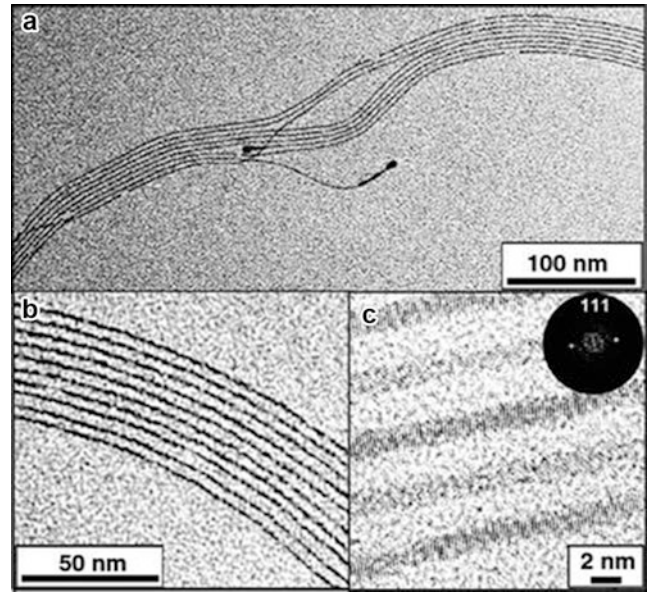
Since the last century, colloid chemists have gained excellent control over the synthesis of spherical gold particles in terms of size and uniformity. However, synthetically controlling particle shape, particularly in solutions, has been a bigger challenge and has met somewhat limited success. Nevertheless, especially in the last 10 years, various methods have been developed to synthesize a variety of anisotropic gold nanoparticles with somewhat controllable shape and morphology. Anisotropic gold nanoparticles have been synthesized in symmetric shapes, such as cylindrical, cubic, polyhedral, and plate-like shapes, but more interestingly, there are methods that produce predominately asymmetric branched morphologies such as nanostars. In the following, the most representative synthesis techniques for various types of anisotropic gold nanoparticles will be described in ascending order of shape complexity [3–6].

Nanorods

In the field of anisotropic nanoparticles, the synthesis of gold nanorods is perhaps the most established protocol in terms of the degree of control of the size, shape, and monodispersity [3, 4]. There are three main methods used to produce gold rods through wet chemistry, namely, the template method, electrochemical method, and seeded-growth method [3]. The template method for the synthesis of gold nanorods is based on the electrochemical deposition of Au within the pores of nanoporous polycarbonate or alumina template membranes. For this, Ag or Cu is first sputtered onto the alumina template membrane to provide a conductive film for electrodeposition, i.e., the growth of Au nanorods. Subsequently, both the template membrane and the copper or silver film are selectively dissolved in the presence of a polymeric stabilizer such as poy(vinylpyrrolidone) (PVP), which allows for dispersing the rods either in water or in organic solvents by means of sonication or agitation. The dimensions of these synthesized nanorods coincide with the pore diameter of the template membranes which can be tuned by controlling the pore sizes [3].

The electrochemical synthesis of gold nanorods is conducted within a simple two-electrode-type electrochemical cell, where a gold metal plate is used as a sacrificial anode while the cathode is a platinum plate with similar dimensions. Both electrodes are immersed in an electrolytic solution containing a mixture of cationic surfactants, hexadecyltrimethylammonium bromide (CTAB), tetradodecylammonium bromide (TDDAB), acetone, and cyclohexane The electrolytic cell containing the mixture is then placed inside an ultrasonic bath at 36 °C and controlled-current electrolysis performed with a typical current of 3 mA for a typical electrolysis time of 30 min. The particle sizes can be tuned either via electrolysis time or by addition of silver ions (silver plate) [3].

The seeded-growth method is perhaps the most facile and most applied method to synthesize gold nanorods. During the early attempts of



Synthesis of Gold Nanoparticles, Fig. 2 TEM images of Au nanowires at different magnifications (**a**, **b**) and an HRTEM image (**c**) showing the single-crystalline structure (Reprinted with permission from Ref. [13]. Copyright 2008 American Chemical Society)

seed-mediated approaches of growing presynthesized spherical gold nanoparticles in solution, formation of a distinct population (5–10%) of colloidal gold rods was observed. In the beginning of this century, it was found that by controlling the growth conditions in aqueous surfactant media, it was possible to synthesize gold nanorods with tunable aspect ratio (Fig. 1). Moreover, addition of $AgNO_3$ influences not only the yield and aspect ratio control of the gold nanorods but also the mechanism for gold nanorod formation and correspondingly its crystal structure morphology and optical properties. A fine-tuning of the aspect ratio of the nanorods can be achieved by simply varying the pH in silver-free syntheses or by adjusting one of the parameters in the silver-mediated procedures, namely, the concentration of silver ions, gold ions, seeds, or reducing agent (ascorbic acid), while keeping the other parameters constant [3, 4].

Nanowires

In comparison to nanorods, nanowires exhibit much larger aspect ratios, and hence larger anisotropy, which makes them ideal candidates for many applications, such as biosensors, logic circuits, field-effect transistors, and nonvolatile memory elements [4, 5]. Being conductive materials, metal nanowires have driven a multitude of theoretical studies on their conductance behavior in the fabrication of nanoelectronics. Inspired from the synthesis of gold nanorods, various methods have been developed to synthesize gold nanowires. These include the pore-template method, block-polymer or DNA template methods, patterned lithography, and chemical reduction methods in bulk solutions. Seed-mediated growth in aqueous CTAB solutions [12] and seedless growth in oleylamine solutions of organic solvents are prime examples of the chemical synthesis of gold nanowires in solutions (Fig. 2).

Depending on the synthesis method and the growth mechanism, gold nanowires can feature different dimensions (diameter, length, and aspect ratio), crystallinity (monocrystalline or polycrystalline or twinned), and shape (straight or twisted). Inspired by the seeded-growth method, Kim et al. reported an acidic solution route to synthesize gold nanowires at room temperature with diameters tunable between 16 and 66 nm and lengths up to 10 μm with an aspect ratio larger than 200 [12]. On the other hand, the method described by Giersig and coworkers, [13] ensures

the synthesis of thin gold nanowires with diameters of ~1.6 nm and lengths ranging from 10 nm to ≥3.5 μm. Control over Au wire lengths is realized by tuning the oleylamine/$HAuCl_4$ volume ratio, the reaction time, and the addition of a second solvent.

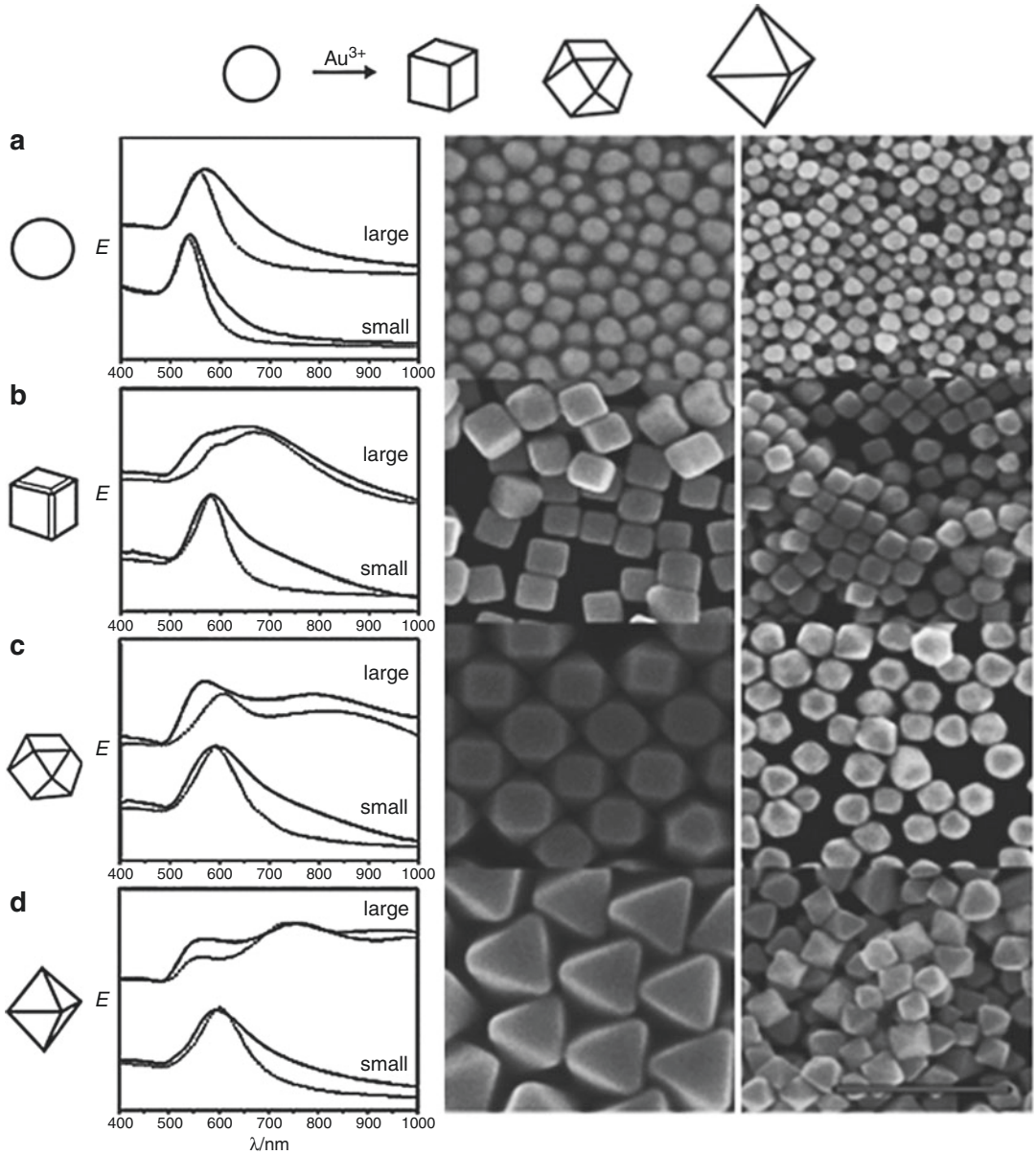
Platonic Nanoparticles (Polyhedrons)

Platonic and quasiplatonic nanocrystals with multiple faces and vertices have been successfully synthesized in relative high yields using both aqueous and nonaqueous approaches [4, 5, 14, 15]. In the method developed by Sau et al. [14] platonic nanoparticles are synthesized in a seeded-growth procedure in aqueous solutions at room temperature. A fine control of nanoparticle shape is achieved by systematic variation of experimental parameters. The experimental procedures involve addition of an appropriate quantity of the presynthesized gold seeds to aqueous growth solutions containing desired quantities of CTAB, $HAuCl_4$, ascorbic acid, and in some cases a small quantity of silver ions. The morphology and dimension of the gold nanoparticles depend on the concentrations of the seed particles and CTAB, in addition to the reactants (Au^{3+} and ascorbic acid).

The synthesis of platonic gold nanoparticles in nonaqueous media is mainly based on the so-called polyol process. This process involves the thermal reduction of a gold salt in an organic solvent with a relatively high boiling point such as poly(ethylene glycol) or *N*,*N*-dimethylformamide (DMF) in the presence of a polymeric stabilizer, usually poly(vinyl pyrrolidone). The ratio of gold salt to PVP is an important parameter to control the shape of the polyhedral particles. Tetrahedral particles are formed at higher gold/PVP ratios and icosahedrons at lower ratios. However, the presence of silver ions induces the formation of uniform gold hexahedrons, i.e., nanocubes. Octahedral, truncated octahedral, cubo-octahedral, and cubic particles (Fig. 3) were synthesized by Seo et al. by conversion of cubo-octahedral particles into different-shaped particles with the aid of silver ions, which seems to be related to selective blocking of the growth along certain crystallographic facets [15]. Also, gold nanorods, synthesized in aqueous media, can be completely reshaped into perfect, single-crystalline octahedrons in a DMF-PVP solution [4]. Besides reshaping and interconversion methods, faceted seeds can be also employed to grow polyhedral nanoparticles, as in the case of synthesizing very regular and nearly monodisperse decahedrons from preformed pentatwinned gold nanoparticle seeds. The synthesis methods for different platonic nanoparticles along with their chemical and physical properties are summarized in Table 1.

Gold Nanoplates

Other very interesting types of anisotropic nanoparticles are triangular or hexagonal gold nanoplates (Fig. 4), which are highly attractive for a number of applications such as optical biosensing and surface-enhanced Raman spectroscopy (SERS), due to their sharp edges, which lead to high local electric field gradients under illumination [4, 5]. The generation of gold nanoplates is often observed during modifications employed for the synthesis of gold nanorods via the seeded-growth method. Small changes in the synthetic procedure, such as an increase in the pH or surfactant concentration, or the addition of extra halide ions, can lead to the formation of planar nanostructures. To date, all current procedures result in rather low yields (40–65%) compared to the synthesis of gold nanorods (~99 %); hence, purification steps are needed in order to eliminate isotropic and nonplanar structures. In the last few years, other techniques rather than seeded-growth or polyol methods have been employed to synthesize nanoplates. Very recently, a more facile synthesis procedure for the fabrication of gold nano- and microplates was reported by Lin et al. [16]. They used a single tree-type multiple-head surfactant, bis (amidoethyl-carbamoylethyl) octadecylamine (C18N3), which functions as both the reducing and capping agent in the reaction system. The triangle and hexagonal plate, polyhedron, and sphere morphology of the gold nanoparticles could be easily controlled simply by changing the molar ratio of $C_{18}N_3/HAuCl_4$ (Fig. 4).



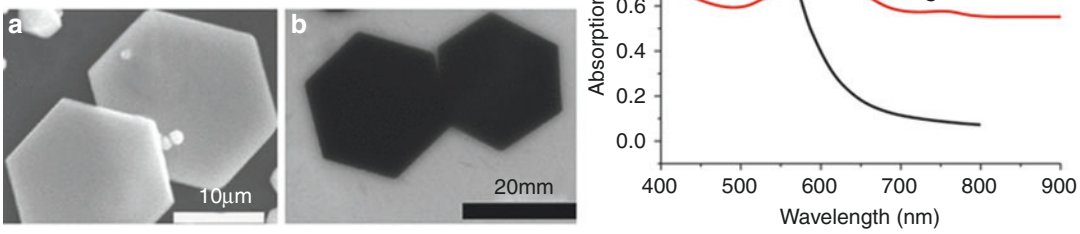
Synthesis of Gold Nanoparticles, Fig. 3 Size and shape control of Au polyhedral nanocrystals from spherical seeds. UV–Vis extinction spectra (*solid lines*) and results of DDA calculations (*dotted lines*) of large (**a**) and small spherical seeds (**a**), large (edge length 116 nm) and small (67 nm) cubes (**b**), large (edge length 122 nm) and small (54 nm) cuboctahedrons (**c**), and large (edge length 236 nm) and small (88 nm) octahedrons (**d**). SEM images (on the *right*, from *top to bottom*) of spherical seeds (818 and 495 nm), cubes (11,611 and 676 nm), cuboctahedrons (12,213 and 548 nm), and octahedrons (23,619 and 8,810 nm). The *bar* represents 500 nm (Reprinted with permission from Ref. [15]. Copyright 2004 Wiley-VCH)

Branched NPs

Coming from symmetric isotropic and anisotropic gold nanoparticles, another important group of gold particles are asymmetric nanoparticles, the so-called branched nanoparticles or nanostars (Fig. 5) [6]. Although these obviously more complex structures are not yet well understood, they generate interest because of their sharp edges and

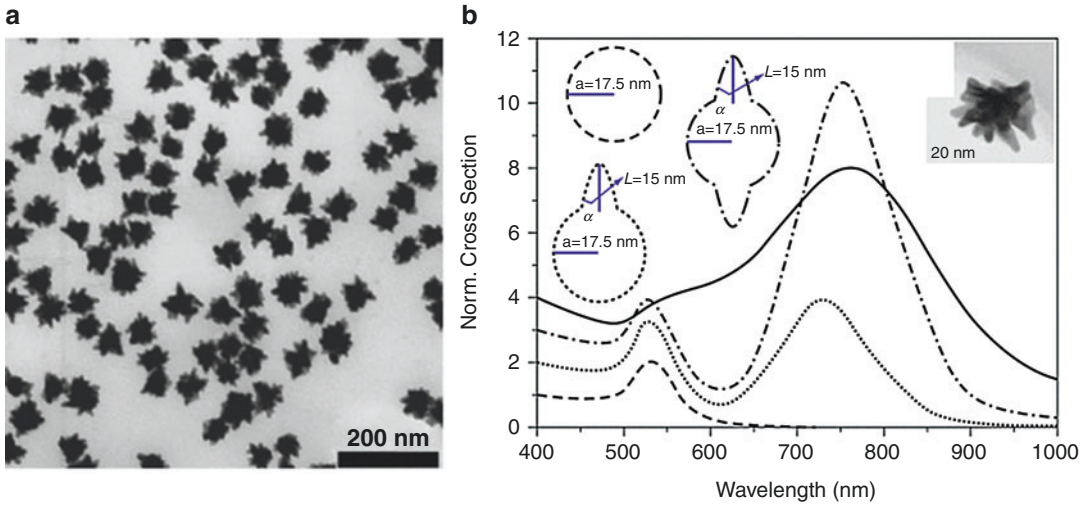
Synthesis of Gold Nanoparticles, Table 1 An overview of various synthesis methods of anisotropic gold nanoparticles, with their physical and chemical properties

Particle shape	Synthesis method	Max. leng. (nm)	SPR max. (nm)	Synthesis medium	Reduction agent	Capping agent
Rods [18]	Seeded growth	≤500 nm	λ_{max}tr: 520–530 λ_{max}long: 750–800	H_2O	Ascorbic acid	CTAB
Wires [12]	Seeded growth	≤10,000	λ_{max}tr: ~540 λ_{max}long: >1,200 nm	H_2O	Ascorbic acid	CTAB
Wires [13]	Seedless growth	≥3,500	λ_{max}tr: – λ_{max}long: > 1,200 nm	Oleylamine	Oleylamine	Oleylamine
Tetrahedra [19]	Polyol	~210	~626 and ~ 950	Ethylene glycol	Ethylene glycol	PVP
Cubes [14]	Seeded growth	~90	~555	H_2O	Ascorbic acid	CTAB
Cubes [19]	Polyol	~150	~621	Ethylene glycol	Ethylene glycol	PVP
Octahedra [20]	Polyol	20–400	530–1,100	Ethylene glycol	Ethylene glycol	PDDA
Decahedra [21]	Seeded growth	36–65	600–700	H_2O/DMF	PVP	PVP
Decahedra [22]	Polyol	48–88	600–700	Diethylene glycol	Diethylene glycol	PVP
Icosahedra [19]	Polyol	230	~613 and ~950	Ethylene glycol	Ethylene glycol	PVP
Triangles [23]	Seeded growth	90–210	750–1,300	H_2O/DMF	Ascorbic acid	CTAB
Hexagons [16]	Oligoamine surfactant	20–15,000	550–900 broad	H_2O	Oligoamine surfactant	Oligoamine surfactant
Stars [14]	Seeded growth	~350	Broad	H_2O	Ascorbic acid	CTAB
Stars [17]	Nonaqueous seeded growth	60–80	Broad	DMF	PVP	PVP



Synthesis of Gold Nanoparticles, Fig. 4 Influence of concentration of C18N3 on the size of gold nanoplates in standard conditions. SEM and TEM images of gold nano- and microplates obtained at different concentrations of C18N3, 2 mM (**a**) and 0.1 mM (**b**). On the *right* are shown the UV–Vis absorption spectra of gold nanospheres (*A*), gold decahedrons (*B*), and gold microplates (*C*) (Reprinted with permission from Ref. [16]. Copyright 2010 American Chemical Society)

S



Synthesis of Gold Nanoparticles, Fig. 5 Representative transmission electron microscopy (*TEM*) image of Au nanostars (**a**), experimental (*solid line*) and calculated (*dashed line*) absorption spectra for gold nanostars; the inset shows a TEM image of a single nanostar (**b**) (Reprinted with permission from Ref. [17]. Copyright IOP Publishing 2008)

the correspondingly high localization of surface plasmon modes. This makes them potential candidates for a number of applications, including SERS-based detection and analysis, as well as the case of nanoplates. Sau et al. proposed the synthesis of branched nanoparticles via the seeded-growth method in the presence of silver ions by varying the seed-to-gold salt ratio and the amount of reducing agent in order to increase the rate of gold ion reduction and thereby induce branching (see Table 1). However, an easier way to synthesize gold nanostars comprises the use of a nonaqueous approach [6, 17]. Branched "nanostars" were synthesized in a concentrated solution of PVP in DMF at room temperature, where PVP acts as a stabilizer and reductant. Nanostars were not only obtained in a seed-mediated growth process but could also be synthesized in a seedless growth process. The fascinating aspect of these particles is that each tip is a single crystal, so the challenge remains to understand the underlying growth mechanism of the selective tip growth on the seed particle's surface.

In summary, the most relevant synthetic approaches mentioned above for the preparation of different anisotropic nanoparticles are presented in Table 1. This table offers the opportunity to take a look at the wide range of synthetic possibilities for the production of anisotropic nanoparticles but at the same time gives a good idea of the physical and chemical properties as well as limitations of these synthetic approaches for the production of particular morphologies with specific sizes. That is why further synthetic research efforts are still needed.

Cross-References

- ▶ Cylindrical Gold Nanoparticles
- ▶ Gold Nanorods
- ▶ Hollow Gold Nanoshells
- ▶ Hollow Gold Nanospheres
- ▶ LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design
- ▶ Nanobiosensors
- ▶ Nanocluster
- ▶ Nanocrystalline Materials
- ▶ Nanoencapsulation
- ▶ Nanofabrication
- ▶ Nanoparticle
- ▶ Nanoparticles
- ▶ Nanophotonic Structures for Biosensing

- ▶ Nanoplasmonics Involved in Localized Photopolymerization
- ▶ Nanorods
- ▶ Nano-Sized Particle

References

1. Eustis, S.E.-S., El-Sayed, M.A.: Why gold nanoparticles are more precious than pretty gold: noble metal surface plasmon resonance and its enhancement of the radiative and nonradiative properties of nanocrystals of different shapes. Chem. Soc. Rev. **35**(3), 209–217 (2006)
2. Astruc, D., Daniel, M.C.: Gold nanoparticles: assembly, supramolecular chemistry, quantum-size-related properties, and applications toward biology, catalysis, and nanotechnology. Chem. Rev. **104**(1), 293–346 (2004)
3. Perez-Juste, J., et al.: Gold nanorods: synthesis, characterization and applications. Coord. Chem. Rev. **249** (17–18), 1870–1901 (2005)
4. Grzelczak, M., et al.: Shape control in gold nanoparticle synthesis. Chem. Soc. Rev. **37**(9), 1783–1791 (2008)
5. Treguer-Delapierre, M., et al.: Synthesis of non-spherical gold nanoparticles. Gold Bull. **41**(2), 195–207 (2008)
6. Guerrero-Martínez, A., et al.: Nanostars shine bright for you: colloidal synthesis, properties and applications of branched metallic nanoparticles. Curr. Opin. Colloid Interface Sci. **16**(2), 118–127 (2011)
7. Chow, P.E. (ed.): Gold Nanoparticles: Properties, Characterization and Fabrication Nanotechnology Science and Technology. Nova, New York (2010)
8. Feldheim, D.L. (ed.): Metal Nanoparticles: Synthesis Characterization and Applications. CRC Press, Boca Raton (2001)
9. Jana, N.R., Gearheart, L., Murphy, C.J.: Evidence for seed-mediated nucleation in the chemical reduction of gold salts to gold nanoparticles. Chem. Mater. **13**(7), 2313–2322 (2001)
10. Oldenburg, S.J., et al.: Nanoengineering of optical resonances. Chem. Phys. Lett. **288**(2–4), 243–247 (1998)
11. Kah, J.C.Y., et al.: Synthesis of gold nanoshells based on the deposition-precipitation process. Gold Bull. **41**(1), 23–36 (2008)
12. Kim, F., et al.: Chemical synthesis of gold nanowires in acidic solutions. J. Am. Chem. Soc. **130**(44), 14442–14443 (2008)
13. Pazos-Perez, N., et al.: Synthesis of flexible, ultrathin gold nanowires in organic media. Langmuir **24**(17), 9855–9860 (2008)
14. Sau, T.K., Murphy, C.J.: Room temperature, high-yield synthesis of multiple shapes of gold nanoparticles in aqueous solution. J. Am. Chem. Soc. **126**(28), 8648–8649 (2004)
15. Seo, D., et al.: Directed surface overgrowth and morphology control of polyhedral gold nanocrystals. Angew. Chem. Int. Ed. **47**(4), 763–767 (2008)
16. Lin, G.H., et al.: A simple synthesis method for gold nano- and microplate fabrication using a tree-type multiple-amine head surfactant. Cryst. Growth Des. **10**(3), 1118–1123 (2010)
17. Kumar, P.S., et al.: High-yield synthesis and optical response of gold nanostars. Nanotechnology **19**(1), 015606 (2008)
18. Liu, M.Z., Guyot-Sionnest, P.: Mechanism of silver(I)-assisted growth of gold nanorods and bipyramids. J. Phys. Chem. B **109**(47), 22192–22200 (2005)
19. Kim, F., et al.: Platonic gold nanocrystals. Angew. Chem. Int. Ed. **43**(28), 3673–3677 (2004)
20. Li, C.C., et al.: A facile polyol route to uniform gold octahedra with tailorable size and their optical properties. ACS Nano **2**(9), 1760–1769 (2008)
21. Sanchez-Iglesias, A., et al.: Synthesis and optical properties of gold nanodecahedra with size control. Adv. Mater. **18**(19), 2529–2534 (2006)
22. Seo, D., et al.: Shape adjustment between multiply twinned and single-crystalline polyhedral gold nanocrystals: decahedra, icosahedra, and truncated tetrahedra. J. Phys. Chem. C **112**(7), 2469–2475 (2008)
23. Millstone, J.E., et al.: Observation of a quadrupole plasmon mode for a colloidal solution of gold nanoprisms. J. Am. Chem. Soc. **127**(15), 5312–5313 (2005)
24. Rodriguez-Fernandez, J., et al.: Seeded growth of submicron Au colloids with quadrupole plasmon resonance modes. Langmuir **22**(16), 7007–7010 (2006)
25. Perez-Juste, J., Correa-Duarte, M.A., Liz-Marzan, L. M.: Silica gels with tailored, gold nanorod-driven optical functionalities. Appl. Surf. Sci. **226**(1–3), 137–143 (2004)

Synthesis of Graphene

Swastik Kar[1] and Saikat Talapatra[2]
[1]Department of Physics, Northeastern University, Boston, MA, USA
[2]Department of Physics, Southern Illinois University Carbondale, Carbondale, IL, USA

Synonyms

Fabrication of graphene or graphene fabrication; Graphene synthesis; How to synthesize/make/manufacture/fabricate/produce graphene; Making graphene; Manufacturing graphene or graphene

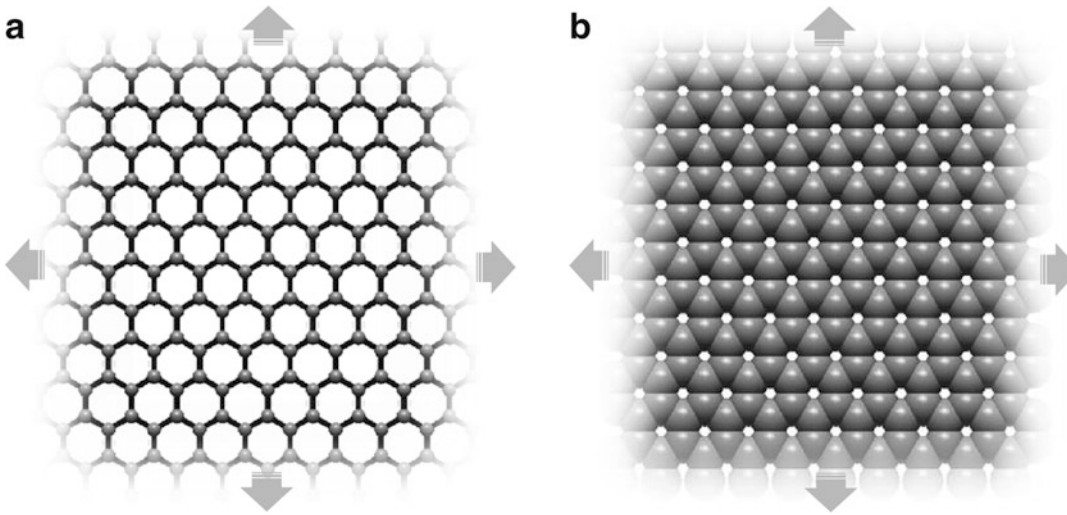
manufacturing; Production of graphene or graphene production

Definition

► Graphene (or monolayer graphene or single-layer graphene) is a single-atom-thick, quasi-infinite, sp^2-hybridized allotrope of carbon in which the atoms are packed in a planar honeycomb crystal lattice (see Fig. 1). It can be visualized as a single sheet of graphite and its lattice structure is related to those of ► fullerenes for drug delivery and ► carbon nanotubes. Few-layered graphene, multilayered graphene, or multigraphene refers to a few layers of graphene stacked (with weak interlayer attraction) in a manner similar to graphite. Figure 1a, b is a schematic representation of a sheet of single-layer graphene.

Overview

Compiling a review on the diverse methods available for the synthesis of graphene is a daunting task, chiefly due to the explosive intensity at which this field has grown over the past few years. Numerous techniques for graphene synthesis have been developed, based on both conventional and innovative techniques, and the graphene thus produced can have vastly different properties. Quite often, this diversity stems from a need for the production of graphene that is tailored for a broad variety of experiment-specific and application-specific research. It is quite impossible to enlist all these methods into a few categories, and this entry is not an attempt to individually address all of them. In general, this work attempts to track and identify some of the most popular, novel, and/or promising methods for graphene synthesis. An attempt has been made to categorize these methods into groups. Since the search for new and/or improved methods for graphene synthesis is in its ascent, the material presented here will not reflect published reports beyond May of 2011. In most cases, generic descriptions of the synthesis conditions have been presented. The aim is to provide the reader with a starting point rather than a complete recipe in each case. It should be noted that experimental conditions and parameters presented here represent typical synthesis conditions as reported in the literature, since what constitutes the "best practice" is both subjective and continuously evolving. It should also be noted that this entry does not deal with the many electronic, optical, thermal, mechanical, or



Synthesis of Graphene, Fig. 1 Schematic representation of carbon atoms in a graphene lattice: (**a**) a ball-and-stick model and (**b**) a space-filling model. The arrows indicate that the sheets are quasi-infinitely extended in all directions. The edge carbon atoms are usually hydrogen terminated

other fascinating properties of graphene, for which excellent review articles already exist in the literature.

Historical Perspective

The concept of graphene has been around for quite a long time, and the descriptive presence of this material appears in a variety of reports in a number of languages. For example, "very thin layers of graphite" including "single foils of carbon" were reported as early as 1962, prepared through a partial reduction of graphite oxide [1]. Monolayer graphite/carbon was identified to form at elevated temperatures on surfaces of Pt in 1969 [2], C-doped Ni (111) crystals in 1979 [3], and other transition metal surfaces [4]. The term "graphene" was already in use in the early 1990s [5], to describe not only monolayer graphite but also the structure of carbon fibers and fullerenes. An earnest endeavor to obtain isolated high-quality graphene from graphite started in 1999, using mechanical exfoliation of patterned graphite pillars [6, 7]. This approach matured in 2004 [8], the same year that synthesis of graphene was also reported by a high-temperature thermal annealing of SiC [9]. These two methods opened up the floodgate of graphene research and catapulted it into becoming the millennium material. In the following paragraphs, a flavor of these and some of the other popular graphene synthesis methods that have been developed since then has been captured.

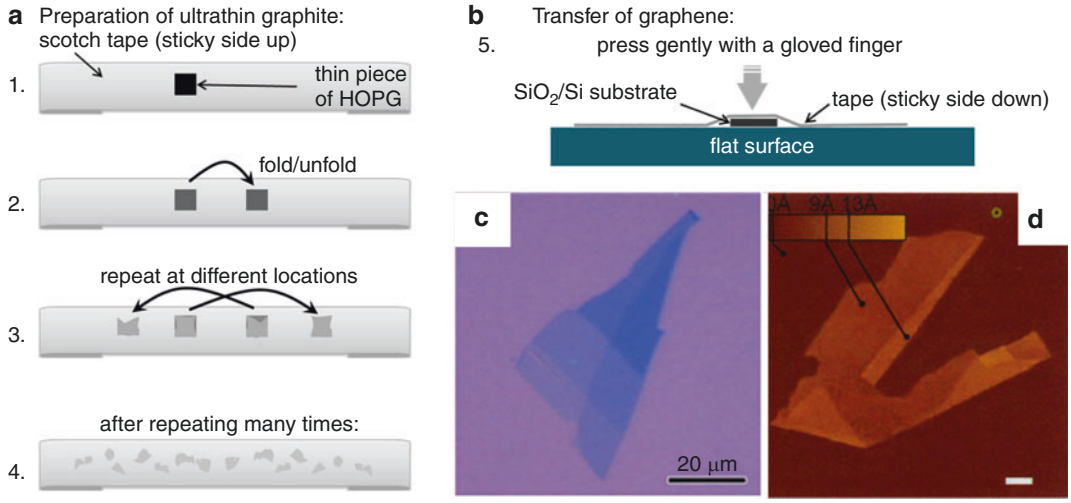
Recent Advances in Graphene Synthesis

Micromechanical Exfoliation of Graphene from Graphite

Strictly not a synthesis process, the micromechanical exfoliation of graphene (both monolayer and few-layer graphene) from highly oriented pyrolytic graphite (HOPG), as reported by Novoselov et al., enabled, for the first time, a direct measurement of its unique 2D electronic properties. This ended the decades-long debate whether pure monatomic two-dimensional crystals can indeed be stable in an isolated state [10, 11]. The few tens-of-microns-sized graphene and few-layered graphene flakes could be easily transferred onto SiO_2/Si and other substrates and enabled researchers to perform a wide variety of electronic, optical, thermal, and mechanical property measurements. In their first report [8], Novoselov et al. described an elaborate process that started from millimeter-thick platelets of HOPG and involved patterning, etching, photoresist, an adhesive-tape peel-off, and a subsequent ultrasonic rinse. A simplified prescription for obtaining graphite was reported by the same group a year later, whereby a freshly cleaved HOPG surface was mechanically rubbed against any solid surface. This invariably left behind debris of crystallites on the surface, and among them, one can usually find single- and few-layered graphene [12]. Owing to the simplicity of this method, a variety of alternatives have become popular, and Fig. 2a, b describes a protocol that simply uses an adhesive tape, which is favored by many. This method has become known as the scotch tape method. As a receiving substrate, it is helpful to use a Si wafer that has ~300-nm-thick oxide layer: graphene itself is nearly transparent, but using this oxide thickness enables one to "see" graphene even using an optical microscope due to the suitable amount of interferometric phase contrasting, as seen in Fig. 2c [13]. Figure 2d is an ▶ atomic force microscopy (AFM) image of a typical flake of graphene, showing the graphene layer thicknesses in flat and folded regions [12].

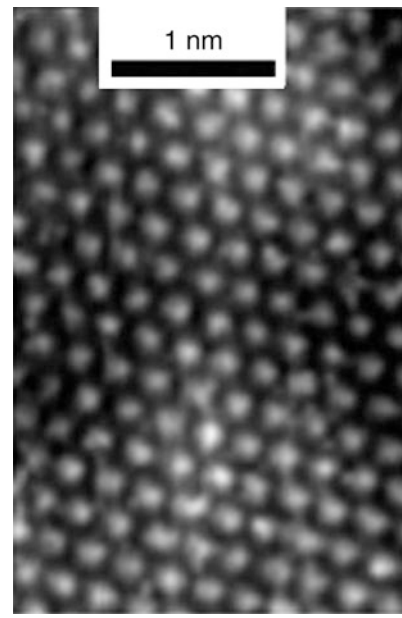
Epitaxial Growth of Graphene on Silicon Carbide

In 2004, another technique for graphene synthesis was reported by Berger et al. [9]. In this method, graphene crystals were produced on the Si-terminated (0001) face of single-crystal 6H-SiC by thermal desorption of Si. The SiC surfaces were initially prepared by oxidation or H_2 etching. The oxide was then removed by electron bombardment-assisted heating in ultrahigh vacuum (~10^{-10} Torr) to 1,000 °C (repeating the oxidation/deoxidation cycle improves the surface quality, and the best initial surface quality was obtained with H_2 etching). The deoxidized



Synthesis of Graphene, Fig. 2 A simple recipe for obtaining single- and few-layered graphene crystals using the micromechanical cleavage of graphite. (**a**) *1* Attach a very thin piece of HOPG onto a scotch tape. *2–4* Fold and unfold as shown to obtain ultrathin (almost invisible) flakes of graphite on the tape. (**b**) Select a region with very thin flakes and transfer on a suitable substrate (e.g., SiO_2/Si) as shown. Carefully remove the tape from the substrate. (**c**) Typical image of a few-layered flake with monolayer extensions as seen under an optical microscope (Reprinted with permission from Ref. [13]; copyright © 2010 Elsevier). (**d**) A typical AFM image of graphene flakes (scale bar = 1 μm) (Reprinted with permission from Ref. [12]; copyright © 2005 National Academy of Sciences, U.S.A.)

samples were heated to temperatures ranging from 1,250 °C to 1,450 °C for 1–20 min, during which time, thin graphite layers were formed. Figure 3 shows a scanning tunneling microscope (STM) image of graphene produced this way by heating SiC for 8 min at 1,400 °C. Over the years, this process has been fine-tuned so as to enable one to controllably produce monolayer graphene on SiC substrates [14]. It turns out that graphenes grown from the SiC(000$$ \bar{1} $$) (C-terminated) surface are of higher quality than those grown on SiC(0001) [15]. Growth on the C face renders graphene with domain sizes more than three times larger than those grown on the Si face and with significantly reduced disorder. In 2009, it was reported that an ex situ graphitization of Si-terminated SiC (0001) in an argon atmosphere of about 1 bar produces monolayer graphene films with much larger domain sizes than those previously attained. This method for the direct synthesis of large-domain graphene (∼10 s of microns in length, with ∼ micron-sized widths that were limited by the step sizes of the underlying substrate)



Synthesis of Graphene, Fig. 3 Atomically resolved STM image of a graphene sample grown on SiC(0001) at 1,400 °C for 8 min (Reprinted with permission from Ref. [9]; copyright © 2004 American Chemical Society)

on an insulating underlayer formed an important step toward realizing graphene-based nanoelectronics on a semiconductor-processable substrate.

In the initial years, most samples were produced using one of the two abovementioned methods, which produced high-quality samples of graphene appropriate for fundamental studies such as electric- and magnetic-field-modulated electronic transport, ARPES studies, Raman studies, and a variety of related experiments. At the same time, as it became evident that there are several other important properties of graphene including their gigantic specific surface area, extreme mechanical strength values, and other thermal and optical properties, several other methods have been developed for large-scale production of graphene. The large-scale production can be categorized into two directions. One route is to develop methods for mass production, often using liquid-phase exfoliation of graphene from graphite, while the other is headed toward wafer-scale production of single- or few-layered graphene, targeted toward electrical, electromechanical, and optoelectronic applications.

Liquid-Phase Exfoliation

Reduced Graphene Oxide and Chemically Modified Graphene

A significant amount of effort has gone into obtaining graphene from the exfoliation of graphite oxide into graphene oxide (GO), followed by a reduction step [16]. The resultant material is more appropriately termed reduced graphene oxide (rGO), and it belongs to the broader category of chemically modified graphenes. Typically, graphite is oxidized using techniques based on Hummer's method [17]. GO, as produced by Hummers' method, is composed of functionalized graphene sheets decorated by strongly bound oxidative debris, which acts as a surfactant to stabilize aqueous GO suspensions [18]. The oxide can hence be dispersed in water using an ultrasonic bath till it becomes a clear-colored liquid with no visible particulates. The parent graphite oxide and the dispersed graphene oxide are both insulators and hence are unsuitable for any applications where electrical conductivity is important. One approach to overcome this problem is a liquid- or gas-phase reduction of the oxidized carbon. Among reducing agents, the use of hydrazine hydrate has become quite popular, and this can be added to the aqueous GO dispersion and heated in an oil bath at 100 °C for a day, during which time, the reduced graphene oxide gradually precipitates out as a black solid. This precipitate is a high-surface-area material which can be separated out using a simple vacuum filtration technique that leaves behind a thick film of rGO. Such films can be transferred onto other substrates in the form of large-area ultrathin films of reduced graphene oxide, which are useful for the development of transparent and flexible electronic materials [19] (see also ▶ Flexible Electronics) and other applications [20, 21].

Although the reduction process restores its conductivity by several orders of magnitude compared to that of GO [22], the oxidation of graphene seems to be quite robust and can only be partially removed. Solid-state ^{13}C NMR spectroscopy of these materials suggests that the sp^2-bonded carbon network of graphite is strongly disrupted, and a significant fraction of this carbon network is bonded to hydroxyl groups or participates in epoxide groups [23], with some carboxylic or carbonyl groups present at the edges. Further works have shown the presence of five- and six-membered ring lactols [24]. As a result, this material has a poor electrical conductivity compared to that of pristine graphene and does not display most of the exciting quantum properties of pure graphene. These groups can be significantly removed by using a sodium borohydride and sulfuric acid treatment, followed by thermal annealing [24], which is effective in restoring the π-conjugated structure and to a significant extent the conductive nature of the graphene materials. While the effort to restore the electronic properties of rGO to a level similar to that of pristine graphene has yet to be reported, rGO-related materials can have several important properties of their own, including large specific surface area and other mechanical properties that enable them to be used as various kinds of energy storage devices (▶ Nanomaterials for Electrical Energy

Storage Devices), sensors, and composite structures.

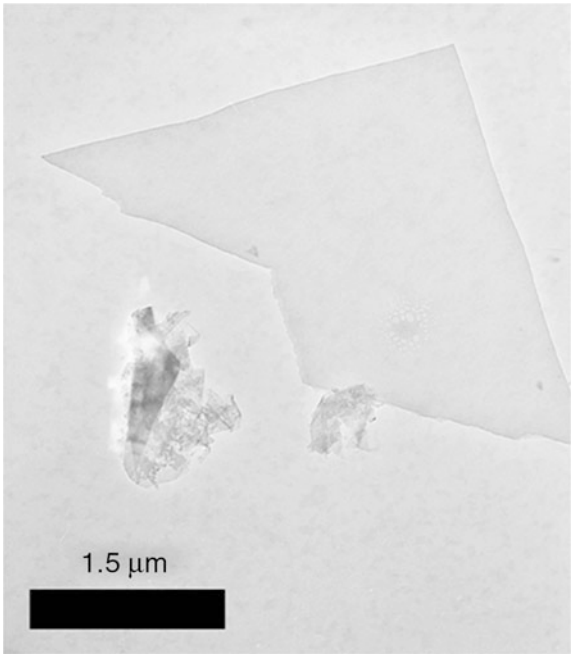
Liquid-Phase Exfoliation of Graphene Directly from Graphite

In some applications, where it is important to have large volumes of *chemically unmodified* graphene, a different approach can be adopted, involving liquid-phase exfoliation of graphene directly from graphite. The idea of liquid-phase exfoliation is essentially a combination of the two methods – mechanical exfoliation from graphite and the liquid-phase environment that is usually applied to GO. In case of GO, the process is pretty straightforward, owing to the weaker interplanar interactions and the strong hydrophilic nature of the individual graphene sheets, and simple ultrasonication of GO causes it to easily disperse stably in water. In contrast, graphene sheets in graphite interact more strongly and, being hydrophobic, do not disperse readily into aqueous media.

Early efforts to exfoliate graphite were mostly through the intercalation of graphite using a range of compounds [25]. With the resurging interest in graphene, this idea has been taken forward to produce large-scale exfoliation of graphene from commercially available expandable graphite [26]. Expandable graphite can be prepared by chemical intercalation of sulfuric acid and nitric acid. Upon heating, the volatile gaseous species released from the intercalant help exfoliate graphite very rapidly into multilayer graphene. This product is then reintercalated with oleum (fuming sulfuric acid with 20 % free SO_3), and tetrabutylammonium hydroxide (TBA, 40 % solution in water) is inserted into the oleum-intercalated graphite. The TBA-inserted oleum-intercalated graphite is sonicated in a *N, N*-dimethylformamide (DMF) solution of 1,2-distearoyl-sn-glycero-3-phosphoethanolamine-*N*-[methoxy(polyethyleneglycol)-5000](DSPE-mPEG) for 60 min to form a homogeneous suspension. A process of centrifugation removes large pieces of material from the supernatant and results in large amounts of graphene sheets suspended in DMF. The exfoliation-reintercalation-expansion of graphite can produce high-quality single-layer graphene sheets stably suspended in organic solvents and possessing high electrical conductance and can be collected as large transparent conducting films by Langmuir-Blodgett assembly in a layer-by-layer manner.

In a more simple approach, stable dispersions up to ~0.01 mg/ml can be directly exfoliated from graphite by ultrasonication in organic solvents such as *N*-methyl pyrrolidone [27]. This method is physicochemically possible because the energy required to exfoliate graphene is balanced by the solvent-graphene interaction for solvents whose surface energies match that of graphene. Typically, graphite is dispersed in a relevant solvent by sonicating in a low-power ultrasonic bath for 30 min. The resultant dispersion is then centrifuged for 90 min at 500 rpm and decanted by pipetting off the top half of the dispersion. The decanted product usually contains a mixture of single- and few-layer graphene flakes that can be made into thin films by passing them through nanoporous membranes and drying them off.

A different method of obtaining stable dispersions in a liquid medium was reported by An et al. [28]. In this method, 1-pyrenecarboxylic acid (PCA) was used as a method to "wedge" out graphene sheets from graphite in suitably polarity-controlled combination of media to give rise to noncovalently functionalized graphene sheets that were stably dispersed in water. In a controlled medium, initially, PCA serves as a "molecular wedge" that cleaves the individual graphene flakes from the parent graphite pieces and then forms stable polar functional groups on the graphene surface via a noncovalent π-π stacking mechanism that does not disrupt its sp^2 hybridization. Figure 4 shows a typical TEM (▶ Transmission Electron Microscopy) image of graphene obtained this way. The hydrophilic –COOH group of PCA facilitates the formation of stable aqueous dispersions of graphene, in a manner similar to that of graphene oxide, but without degrading the sp^2 structure. The advantage of this method lies in the fact that the final product is stable in water – opening up the possibilities of experiments where an aqueous medium could be important, such as interactions with



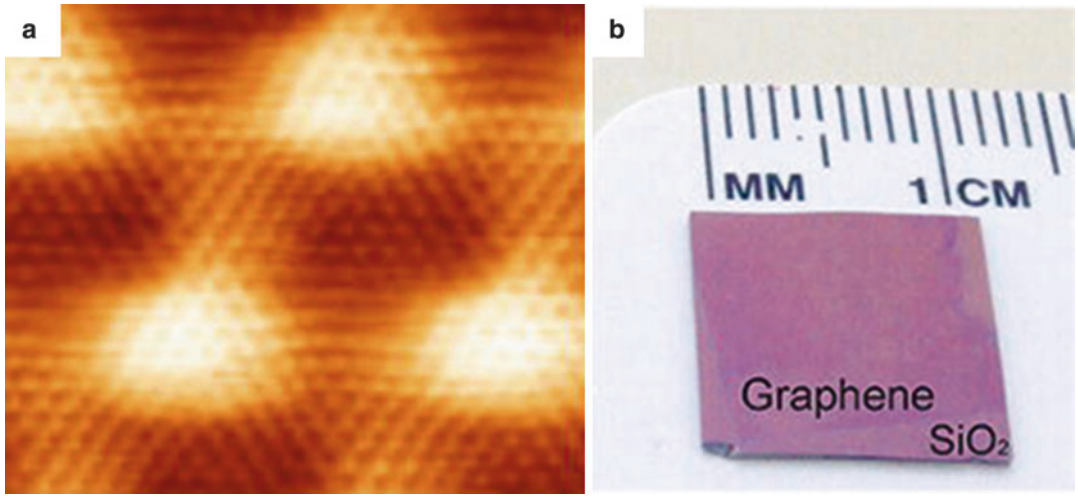
Synthesis of Graphene, Fig. 4 (**a**) A TEM image of a typical graphene flake exfoliated from graphite using a molecular wedging technique [28]

biological samples. At the same time, a simple methanol wash will remove the excess PCA from the surface and the pure graphene sheet precipitates out with time – to be utilized for experiments where pure, unfunctionalized graphene sheets are required. Similar methods have been demonstrated by other groups to obtain dispersions of graphene from graphite in surfactant-water solutions [29], through which significant levels of exfoliation and monolayer formation can be achieved.

Epitaxial and Large-Area Graphene on Metal Surfaces

High-temperature annealing of Pt single crystals produces an atomically thin entity on their surfaces, and this layer has been identified as monolayer graphite back in 1969 through low-energy electron diffraction (LEED) experiments [2]. Since then, it became clear that graphene in this morphology appears either due to the surface desegregation of carbon or through decomposition of hydrocarbons. Graphene has been reported to grow on a variety of metals such as Co, Ni, Ru, Rh, Pd, Ir, and Pt, with varying degrees of grain size, epitaxial alignment with the underlying metal, defect densities, and surface morphologies [30]. For example, it was shown in 2006 that ethylene, absorbed at room temperature on Ir (111) planes, could be desegregated at 1,450 K to produce graphene flakes of ~100 nm size that covers about 30 % of the surface [31]. More recently, the desegregation method was successfully utilized to graphene on Ru (0001) surfaces, where macroscopic single-crystalline domains with linear dimensions exceeding 200 µm could be achieved [32, 33]. Figure 5a shows a high-resolution STM (▶ Scanning Tunneling Microscopy) image of graphene grown on Ru surface. By controlling the chamber pressure, growth time, and temperature, millimeter-scale graphene was reported on Ru surfaces using a controlled desegregation process in 2009 [34].

In 2008, the desegregation method received a boost when a low-pressure ▶ chemical vapor deposition (CVD) technique was utilized to synthesize graphene on an Ir (111) surface in a controlled manner over a range of temperatures from 1,120 to 1,320 K [35]. Scanning tunneling microscopy of the grown structures reveals that graphene prepared using this technique exhibits a continuity of its carbon rows over terraces and



Synthesis of Graphene, Fig. 5 (**a**) Atomically resolved STM image of the graphene layer grown on Ru (0001) (Reprinted with permission from Ref. [32]; copyright © (2007) by the American Physical Society). (**b**) A digital photograph of a graphene film grown on a Cu foil using a CVD method and transferred onto a SiO_2/Si substrate (Reprinted with permission from Ref. [38]; copyright © (2009) by the American Association for the Advancement of Science)

step edges, enabling the possibility of large-scale growth that is not limited by surface features of metals. The thermodynamics of carbon segregation and graphene formation involves a complex interplay of carbon diffusion from the bulk to the surface of the host metal, followed by surface diffusion of C atoms that lead to nucleation and growth. In a typical CVD process, a metal substrate is placed in a flow-type furnace which is heated to a temperature between 750 °C and 1,000 °C in the presence of a mixture of Ar and H_2 at a fixed chamber pressure P. Once the target temperature stabilizes, the flow is maintained for some time (30–60 min) to clean and prepare the surface of the metal. At the end of this preparation time, the carbon source, typically ethylene or methane, is introduced for a fixed time, at the end of which the furnace is allowed to cool down. Once cooled, the surface is found to be covered with single- and multilayered graphene. Factors that appear to control layer thickness and domain sizes of graphene obtained such as temperature, feed-gas concentration, chamber pressure, and time appear to be vastly different for different metals. For example, Yu et al. used CH_4:H_2:Ar = 0.15:1:2 with a total gas flow rate of 315 sccm and found that the type of graphene obtained on Ni foils depended strongly on the cooling rate, with multilayer graphene obtained only when the cooling rate was controlled to 10 °C/s [36]. A great benefit of this technique is that thin layers of pre-patterned Ni could be utilized to fabricated large-scale patterned growth of graphene films that could be transferred to arbitrary substrates [37].

At elevated temperatures, the solubility of C atoms in Ni is much higher compared to that in Cu, and the resulting graphene formation can have different morphologies for these two metals. Centimeter-scale graphene films were shown to grow rather easily on Cu substrates in 2009 through a similar CVD process [38], with more than 95 % of the surface found to be covered with single-layer graphene that are continuous across copper surface steps and grain boundaries. Figure 5b shows an example of a macroscopic sheet of graphene grown on a Cu foil and transferred onto a SiO_2/Si substrate. In case of Cu, it seems that the cooling rate is unimportant, most probably due to the low solubility of carbon in copper which helps to make the growth process self-limiting. There were no discernible

differences in the graphene morphology when the cooling rate was varied from >300 °C/min to about 40 °C/min. The ability to grow high-quality large-scale graphene on a low-cost material (Cu) at such ease has been rapidly taken advantage of to produce graphene at a roll-to-roll capability [39]. This, along with a wet chemical doping and layer-by-layer stacking process, can be used to devise many-layer macroscopic graphene films with sheet resistance at values as low as ~30 Ω $\square^{-1}$ at ~90 % transparency, which is superior to commercial transparent electrodes such as indium tin oxides and opens up several new possibilities in the fields of optics, optoelectronics, and photovoltaics.

Other Novel Advances in Graphene Synthesis

In addition to the categories described above, there are numerous innovative methods being reported on a regular basis that are either extensions of the abovementioned categories or are completely new methods that can become significantly useful and/or popular in future. In addition to graphene, there is a whole array of related materials, including doped and functionalized graphene and graphene nanoribbons, each of which is unique and important by its own rights. It is impossible to either enlist all or describe these methods individually within the scope of this entry. Some of the more popular and/or innovative ones are enlisted in Table 1 in order to provide the reader with a starting point. It should be borne in mind that a keyword search of existing databases through available searching tools was employed to obtain these references, and omission of any significant report(s) is purely accidental and inadvertent.

Synthesis of Graphene, Table 1 Some recent, novel, and uncategorized graphene synthesis techniques

Method	Publication details
Solvothermal synthesis	Choucair et al. Nat. Nanotechnol. **4**, 30–33 (2009)
Substrate-free synthesis	Dato et al. Nano Lett. **8**, 2012–2016 (2008)
Electrochemical synthesis	1. Liu et al. Adv Func. Mat. **18**, 1518–1525 (2008)
	2. Guo et al. ACS Nano **3**, 2653–2659 (2009)
Microwave plasma enhanced CVD	Malesevic et al. Nanotech. **19**, 305604.1–305604.6 (2008)
Arc discharge exfoliation	Wu et al. ACS Nano **3**, 411–417 (2009)
Microwave-solvothermal synthesis	Murugan et al. Chem. Mat. **21**, 5004–5006 (2009)
Doped graphene	1. Panchokarla et al. Adv Mat **21**, 4726–4730 (2009)
	2. Ci et al. Nat. Mater. **9**, 430–435 (2010)
	3. Li et al. Carbon **48**, 255–259 (2010)
Graphite exfoliation in ionic liquids	Lu et al. ACS Nano **3**, 2367–2375 (2009)
Hydrophilic and organophilic graphene	Wang et al. Carbon **47**, 1359–1364 (2009)
Graphene nanoribbons	1. Han et al. Phys. Rev. Lett. **98**, 206805.1–206805.4 (2007)
	2. Li et al. Science **319**, 1229–1232 (2008)
	3. Jiao et al. Nat. Nanotech. 5, 321–325 (2010)
Reducing sugar	Zhu et al. ACS Nano **4**, 2429–2437 (2010)

Conclusion

It is evident that within a very short time frame of a few years, the scientific community has witnessed a tremendous advancement in terms of various techniques of graphene synthesis. This advancement has progressed hand-in-hand with a number of extremely important fundamental discoveries related to this novel 2D crystalline material and has also opened up several avenues for groundbreaking technological directions. Despite the rapid success that this field has experienced, the interest in developing better and more cost-effective synthesis methods keeps growing. In the near term, production of high-quality graphene with reproducible electrical as well as mechanical properties will be a challenge. Similarly, techniques and parameters for size-specific, controlled synthesis of graphene such as graphene nanoribbons, graphene quantum dots, etc., need to be well understood. Processes for large-scale synthesis of monolayer graphene (e.g., through chemical exfoliation) need to perfected in order to

utilize these materials in a variety of bulk applications such as electrodes in electrical energy storage systems, fillers in composite, etc. Overall, the current level of progress as well as the ongoing research efforts in graphene synthesis has been quite promising for certain niche area applications, and the scientific community is quite optimistic that in the future, graphene will eventually emerge as the next ultimate engineering material.

Cross-References

- Atomic Force Microscopy
- Carbon Nanotubes
- Chemical Vapor Deposition (CVD)
- Flexible Electronics
- Fullerenes for Drug Delivery
- Graphene
- Nanomaterials for Electrical Energy Storage Devices
- Scanning Tunneling Microscopy
- Transmission Electron Microscopy

References

1. Boehm, H.P., Clauss, A., Fischer, G.O., Hofmann, U.: Das Adsorptions verhalten sehr dünner Kohlenstoffolien. Z. Anorg. Allg. Chem. **316**, 119–127 (1962)
2. May, J.W.: Platinum surface LEED rings. Surf. Sci. **17**, 267–270 (1969)
3. Eizenberg, M., Blakely, J.M.: Carbon monolayer phase condensation on Ni(111). Surf. Sci. **82**, 228–236 (1979)
4. Aizawa, T., Souda, R., Otani, S., Ishizawa, Y., Oshima, C.: Anomalous bond of monolayer graphite on transition-metal carbide surfaces. Phys. Rev. Lett. **64**, 768–771 (1990)
5. Dresselhaus, M.S., Dresselhaus, G., Saito, R.: Carbon fibers based on C60 and their symmetry. Phys Rev. B**45**, 6234–6242 (1992). Aizawa, T., Hwang, Y., Hayami, W., Souda, R., Otani, S., Ishizawa, Y.: Phonon dispersion of monolayer graphite on Pt(111) and NbC surfaces: bond softening and interface structures. Surf Sci. **260**, 311–318 (1992)
6. Lu, X., Yu, M., Huang, H., Ruoff, R.S.: Tailoring graphite with the goal of achieving single sheets. Nanotechnology **10**, 269–272 (1999)
7. Lu, X., Huang, H., Nemchuk, N., Ruoff, R.S.: Patterning of highly oriented pyrolytic graphite by oxygen plasma etching. Appl. Phys. Lett. **75**, 193–195 (1999)
8. Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Zhang, Y., Dubonos, S.V., Grigorieva, I.V., Firsov, A.A.: Electric field effect in atomically thin carbon films. Science **306**, 666–669 (2004)
9. Berger, C., Song, Z., Li, T., Li, X., Ogbazghi, A.Y., Feng, R., Dai, Z., Marchenkov, A.N., Conrad, E.H., First, P.N., de Heer, W.A.: Ultrathin epitaxial graphite: 2D electron gas properties and a route toward graphene-based nanoelectronics. J. Phys. Chem. B **108**, 19912–19916 (2004)
10. Peierls, R.E.: Quelques proprieties typiques des corpses solides. Ann. Inst. Henri. Poincare **5**, 177–222 (1935)
11. Landau, L.D.: Zur Theorie der phasenumwandlungen II. Phys. Z. Sowjetunion **11**, 26–35 (1937)
12. Novoselov, K.S., Jiang, D., Schedin, F., Booth, T.J., Khotkevich, V.V., Morozov, S.V., Geim, A.K., Rice, T. M.: Two-dimensional atomic crystals. Proc. Natl. Acad. Sci. U. S. A. **102**, 10451–10453 (2005)
13. Soldano, C., Mahmood, A., Dujardin, E.: Production, properties and potential of graphene. Carbon **48**, 2127–2150 (2010)
14. Rollings, E., Gweon, G.-H., Zhou, S.Y., Mun, B.S., McChesney, J.L., Hussain, B.S., Fedorov, A.N., First, P.N., de Heer, W.A., Lanzara, A.: Synthesis and characterization of atomically thin graphite films on a silicon carbide substrate. J. Phys. Chem. Solids **67**, 2172–2177 (2006)
15. Hass, J., Feng, R., Li, T., Li, X., Zong, Z., de Heer, W. A., First, P.N., Conrad, E.H., Jeffrey, C.A., Berger, C.: Highly ordered graphene for two dimensional electronics. Appl. Phys. Lett. **89**, 143106–143108 (2006)
16. Stankovich, S., Dikin, D.A., Piner, R.D., Kohlhaas, K. A., Kleinhammes, A., Jia, Y., Wu, Y., Nguyen, S.T., Ruoff, R.S.: Synthesis of graphene-based nanosheets via chemical reduction of exfoliated graphite oxide. Carbon **45**, 1558–1565 (2007)
17. Hummers Jr., W.S., Offeman, R.E.: Preparation of graphitic oxide. J. Am. Chem. Soc. **80**, 1339–1339 (1958)
18. Rourke, J.P., Pandey, P.A., Moore, J.J., Bates, M., Kinloch, I.A., Young, R.J., Wilson, N.R.: The real graphene oxide revealed: stripping the oxidative debris from the graphene-like sheets. Angew. Chem. Int. Ed. **50**, 3173–3177 (2011)
19. Eda, G., Fanchini, G., Manish Chhowalla, M.: Large-area ultrathin films of reduced graphene oxide as a transparent and flexible electronic material. Nat. Nanotechnol. **3**, 270–274 (2008)
20. Gilje, S., Han, S., Wang, M., Wang, K.L., Kaner, R.B.: A chemical route to graphene for device applications. Nano Lett. **7**, 3394–3398 (2007)
21. Tung, V.C., Allen, M.J., Yang, Y., Kaner, R.B.: High-throughput solution processing of large-scale graphene. Nat. Nanotechnol. **4**, 25–29 (2009)
22. Si, Y., Samulski, E.T.: Synthesis of water soluble graphene. Nano Lett. **8**, 1679–1682 (2008)

23. Park, S., Ruoff, R.S.: Chemical methods for the production of graphenes. Nat. Nanotechnol. **4**, 217–224 (2009)
24. Gao, W., Alemany, L.B., Ci, L., Ajayan, P.M.: New insights into the structure and reduction of graphite oxide. Nat. Chem. **1**, 403–408 (2009)
25. Dresselhaus, M.S., Dresselhaus, G.: Intercalation compounds of graphite. Adv. Phys. **30**, 139–326 (1981)
26. Li, X., Zhang, G., Bai, X., Sun, X., Wang, X., Wang, E., Dai, H.: Highly conducting graphene sheets and Langmuir-Blodgett films. Nat. Nanotechnol. **3**, 538–542 (2008)
27. Hernandez, Y., Nicolosi, V., Lotya, M., Blighe, F.M., Sun, Z., De, S., McGovern, I.T., Holland, B., Byrne, M., Gun'Ko, Y.K., Boland, J.J., Niraj, P., Duesberg, G., Krishnamurthy, S., Goodhue, R., Hutchison, J., Scardaci, V., Ferrari, A.C., Coleman, J.N.: High-yield production of graphene by liquid-phase exfoliation of graphite. Nat. Nanotechnol. **3**, 563–568 (2008)
28. An, X., Simmons, T., Shah, R., Wolfe, C., Lewis, K. M., Washington, M., Nayak, S.K., Talapatra, S., Kar, S.: Stable aqueous dispersions of noncovalently functionalized graphene from graphite and their multifunctional high-performance applications. Nano Lett. **10**, 4295–4301 (2010)
29. Lotya, M., Hernandez, Y., King, P.J., Smith, R.J., Nicolosi, V., Karlsson, L.S., Blighe, F.M., De, S., Wang, Z., McGovern, I.T., Duesberg, G.S., Coleman, J.N.: Liquid phase production of graphene by exfoliation of graphite in surfactant/water solutions. J. Am. Chem. Soc. **131**, 3611–3620 (2009)
30. Wintterlin, J., Bocquet, M.-L.: Graphene on metal surfaces. Surf. Sci. **603**, 1841–1852 (2009)
31. N'Diaye, A.T., Bleikamp, S., Feibelman, P.J., Michely, T.: Two-dimensional Ir cluster lattice on a graphene moire on Ir(111). Phys. Rev. Lett. **97**, 215501.1–215501.4 (2006)
32. Marchini, S., Gunther, S., Wintterlin, J.: Scanning tunneling microscopy of graphene on Ru(0001). Phys. Rev. B **76**, 075429.1–075429.9 (2007)
33. Sutter, P.W., Flege, J.-I., Sutter, E.A.: Epitaxial graphene on ruthenium. Nat. Mater. **7**, 406–411 (2008)
34. Pan, Y., Zhang, H.G., Shi, D.X., Sun, J., Du, S., Liu, F., Gao, H.-J.: Highly ordered, millimeter-scale, continuous, single-crystalline graphene monolayer formed on Ru (0001). Adv. Mater. **21**, 2777–2780 (2009)
35. Coraux, J., N'Diaye, A.T., Busse, C., Michely, T.: Structural coherency of graphene on Ir(111). Nano Lett. **8**, 565–570 (2008)
36. Yu, Q., Lian, J., Siriponglert, S., Li, H., Chen, Y.P., Pei, S.-S.: Graphene segregated on Ni surfaces and transferred to insulators. Appl. Phys. Lett. **93**, 113103–113105 (2008)
37. Kim, K.S., Zhao, Y., Jang, H., Lee, S.Y., Kim, J.M., Kim, K.S., Ahn, J.-H., Kim, P., Choi, J.-Y., Hong, B. H.: Large-scale pattern growth of graphene films for stretchable transparent electrodes. Nature **457**, 706–710 (2009)
38. Li, X., Cai, W., An, J., Kim, S., Nah, J., Yang, D., Piner, R., Velamakanni, A., Jung, I., Tutuc, E., Banerjee, S.K., Colombo, L., Ruoff, R.S.: Large-area synthesis of high-quality and uniform graphene films on copper foils. Science **324**, 1312–1314 (2009)
39. Bae, S., Kim, H., Lee, Y., Xu, X., Park, J.-S., Zheng, Y., Balakrishnan, J., Lei, T., Kim, H.R., Song, Y.I., Kim, Y.-J., Kim, K.S., Özyilmaz, B., Ahn, J.-H., Hong, B.H., Iijima, S.: Roll-to-roll production of 30-inch graphene films for transparent electrodes. Nat. Nanotechnol. **5**, 574–578 (2010)

Synthesis of Multicationic 1-D Oxide Nanostructures

▶ Growth of 1-D Oxide Nanostructures

Synthesis of Nanoparticles

▶ Nanoparticles
▶ Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods

Synthesis of Subnanometric Metal Nanoparticles

Javier Calvo Fuentes[1], José Rivas[2] and M. Arturo López-Quintela[2]
[1]Nanogap Sub-Nm-Powder S.A., Milladoiro – Ames (A Coruña), Spain
[2]Laboratory of Magnetism and Nanotechnology, Institute for Technological Research, University of Santiago de Compostela, Santiago de Compostela, Spain

Synonyms

Atomic cluster; Cluster; Nanocluster; Quantum cluster; Quantum dot

Definition

Metal atomic clusters consist of groups of atoms with well-defined compositions and one or very few stable geometric structures. They represent the most elemental building blocks in nature – after atoms – and are characterized by their size, comparable to the Fermi wavelength of an electron, which makes them a bridge between atoms and nanoparticles or bulk metals, with properties very different from both of them.

Introduction

Typical metal nanoparticles with dimensions from two to several tens of nanometers show smoothly size-dependent properties. However, when particle size becomes comparable to the Fermi wavelength of an electron (~0.52 nm for gold and silver), properties of metal clusters are dramatically different from what should be expected if they were due only to their high surface-to-volume ratio. In these subnanometric species, quantum effects are responsible for totally new chemical, optical, and electronic properties such as, for example, magnetism, photoluminescence, or catalytic activity.

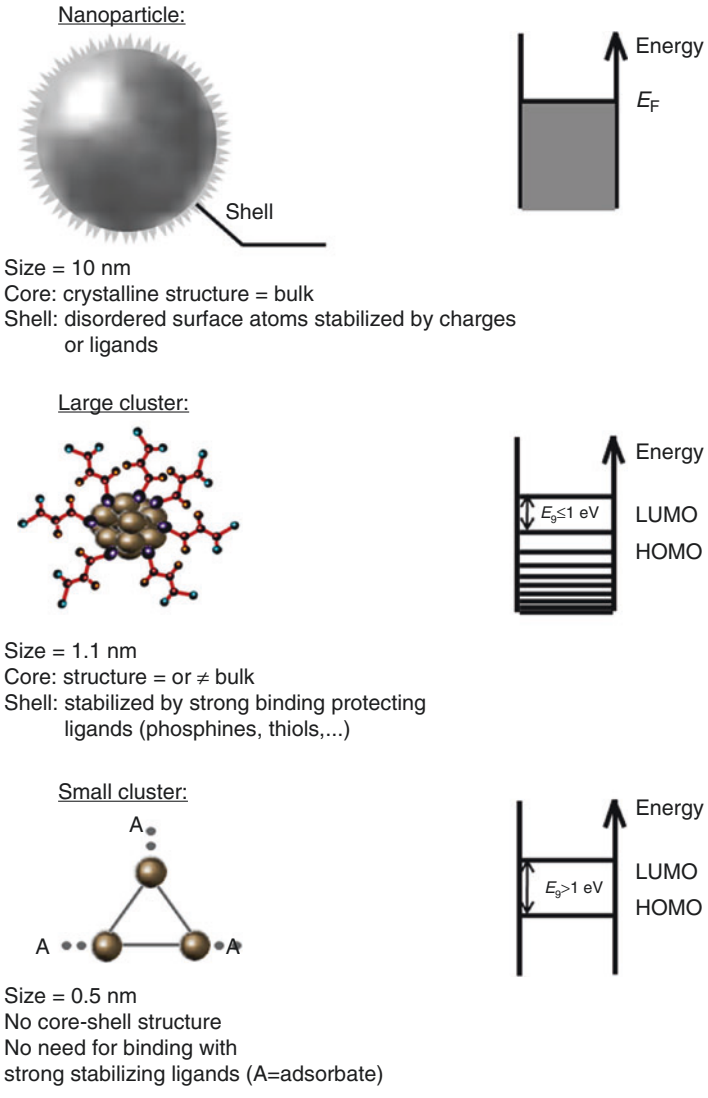
The term *nanoparticle* usually refers to any particle of bulk metal with dimensions in the nanoscale. Nanoparticles usually present a core-shell structure with a core of bulk metal surrounded by a shell of disordered atoms, and their enhanced properties are due mainly to their high surface-to-volume ratio. The term *cluster* is used in reference to *subnanometric species* consisting of well-defined structures of metal atoms stabilized by different types of protecting ligands, with sizes below approximately 1–2 nm. In general, clusters can be divided into (1) large clusters, consisting of a core formed by a number of metal atoms in the range ≈10–20 to 100–200 and a protecting shell of strong ligands such as phosphines or thiols; and (2) small clusters formed by a reduced number of atoms (≈2 to 10–20), which do not need any strong stabilizing ligand and have almost all their atoms on the surface. Due to quantum effects, both kinds of clusters – large and small – present discrete energy levels and an increasing bandgap with decreasing size (Fig. 1).

Because of this splitting of energies at the Fermi level, clusters show different properties to those of nanoparticles. Cluster properties are highly dependent on their sizes. As an example, the typical surface plasmon band observed in Ag-metal nanoparticles (Fig. 2a) disappears for clusters below approximately 1–2 nm (Fig. 2b, c corresponding to large and small clusters, respectively), indicating that all the conducting electrons are now *frozen* and the *metal* silver loses its typical *metallic character*. At the same time, a clear difference between large and small clusters can be observed: large clusters show a continuous decrease of the absorption band with some small bumps, similar to the absorption displayed by semiconductors (Fig. 2b), and small clusters display well-defined absorption bands indicating a *molecular-like* behavior (Fig. 2c).

The presence of such size-dependent properties can be well represented on a 3D periodic table of elements – schematically depicted in Fig. 3 – where, between the atomic and the bulk state of every element, a whole range of new materials with their different size-dependent properties appear.

Cluster Structure

It is very difficult to do precise calculations of the electronic structure of metal clusters, particularly when they are larger than just a small number of atoms, and they are in solution stabilized by some capping molecules. Such difficulty can be avoided using simple models, such as the Jellium model [1] firstly developed for clusters in the gas phase. In this model, the real cluster is replaced by an electronic shell structure consisting of a uniform, positively charged *jellium* sphere surrounded by valence electrons. The electrons are considered to be like moving in a mean-field potential occupying energy levels according to the Aufbau principle ($1S^2 \mid 1P^6 \mid 1D^{10} \mid 2S^2\ 1F^{14} \mid 2P^6\ 1G^{18} \mid \ldots$) as represented in Fig. 4. This model gives a



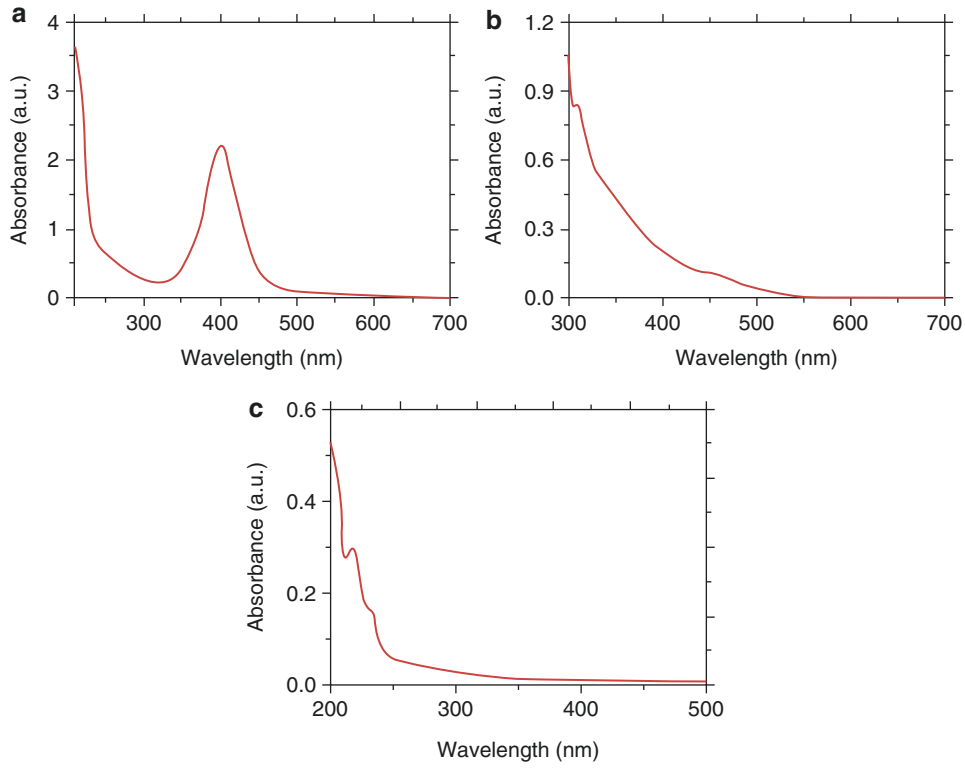
Synthesis of Subnanometric Metal Nanoparticles, Fig. 1 Schematic properties of nanoparticles and subnanometric clusters

considerably good approximation, preserving many of the physicochemical characteristics of clusters.

The total energy as a function of cluster size, calculated by this approximation, has discontinuities corresponding to electronic shell closures that are consistent with the positions of the peaks appearing in the mass spectrum of the vaporized metal. These peaks are associated with the abundance of exceptionally stable clusters with certain number of atoms: 8, 18, 20, 34, 40, 58, 92, 138, etc., usually referred to as *magic numbers*.

As it was mentioned above, one simple consequence of this quantum-size regime with the presence of discrete states in metal clusters is the appearance of a sizable HOMO-LUMO bandgap similar to that of semiconductors. Such semiconductor-like behavior is particularly significant for smaller clusters (i.e., those with a low number of atoms) with bandgaps widely exceeding 1 eV, as it is shown in Fig. 5 for some selected Ag and Cu clusters.

Some efforts have been made in the last years to find prediction models for ligand-protected



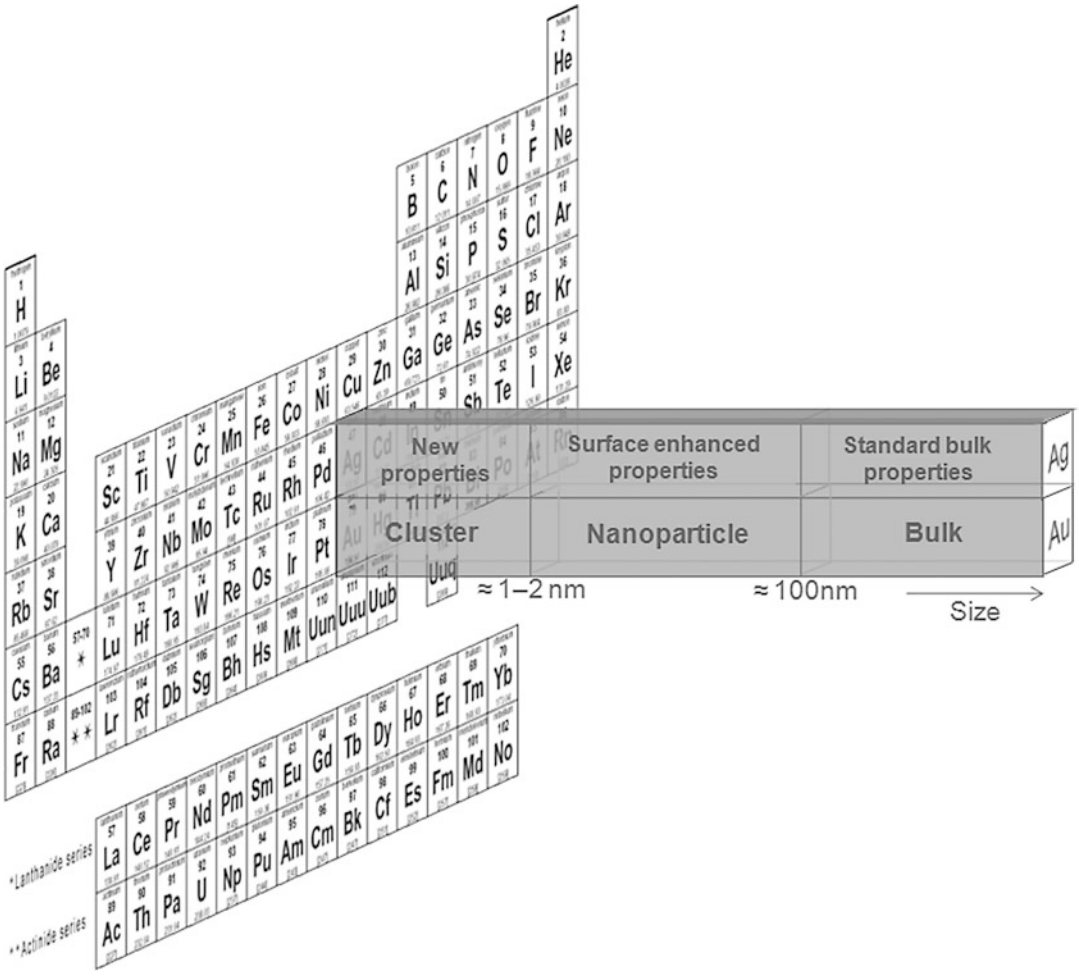
Synthesis of Subnanometric Metal Nanoparticles, Fig. 2 Absorption spectra of (**a**) silver nanoparticles displaying the typical plasmon band at ≈400 nm, (**b**) surfactant-protected large silver clusters showing a typical continuum decrease of the absorption, and (**c**) *strong-ligand-free* small silver clusters with well-defined *molecular-like* absorption bands

gold-cluster species in solution. In such models, clusters are assumed to be electronically stabilized by ligands that either withdraw electrons from the metal core or are attached as weak Lewis base ligands coordinated to the core surface by dative bonds. In this case, the requirement for an electronically closed shell [3], formulated as $(L_S \cdot A_N X_M)^z$, is

$$n^* = N\nu_A - M - z \qquad (1)$$

where n^* represents the number of electrons for shell closing of the metallic core, which has to match one of the magic numbers corresponding to strong electron shell closures in an anharmonic mean-field potential, giving rise to stable clusters; N stands for the number of core metal atoms (A); ν_A is the atomic valence; M is the number of electron-localizing (or electron-withdrawing) ligands X, assuming a withdrawal of one electron per ligand molecule; and z represents the overall charge of the complex. The weak ligands represented by L_S may be needed for completing the steric surface protection of the cluster core.

Trying to combine at the same time (1) the mentioned magic numbers required for electronic stabilization of the cluster, (2) the requirements of the Eq. 1, and (3) the fact that solution-phase clusters require a sterically complete protective ligand shell compatible with a compact atomic shell structure for the metallic core complicates enormously the calculations of the structures satisfying such conditions. Furthermore, in cases such as some gold and silver thiolate clusters, the identity of the actual X groups is not clear due to the undefined nature of the surface chemical bonds. There have been only relatively few studies so far to resolve the cluster structures by ab



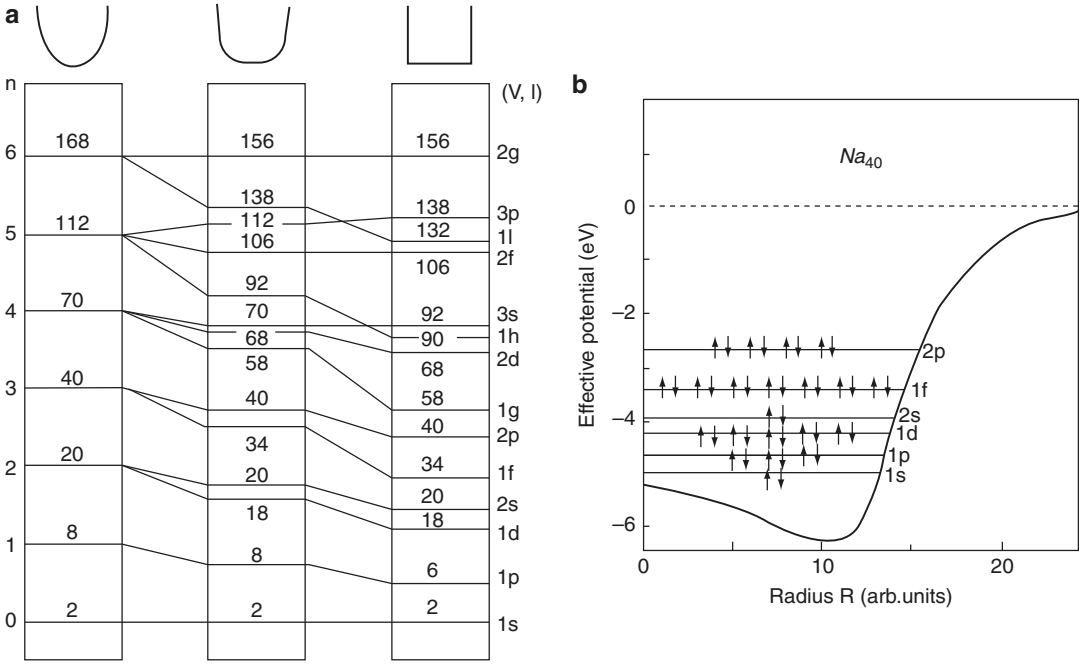
Synthesis of Subnanometric Metal Nanoparticles, Fig. 3 Size-dependent 3D periodic table of elements

initio calculations. Very recently, Häkkinen's group managed to determine, through large-scale density functional theory (DFT) calculations, [3] the electronic structure of $Au_{102}(p\text{-MBA})_{44}$ (*p*-MBA, *para*-mercaptobenzoic acid, $SC_7O_2H_5$), an all-thiolate-protected cluster formed by 102 gold atoms, as well as some smaller phosphine-halide- and phosphine-thiolate-protected clusters, as it is shown in Fig. 6.

In all cases, the characteristic bandgap of the clusters is clearly observed. In Table 1, it is shown how this bandgap changes for different cluster sizes and ligands. For comparison purposes, the bandgap calculated from the simple spherical Jellium model is included (for clusters with sizes $\approx\geq$25 atoms, a correction of −0.4 eV was used because of the anharmonicity observed in large clusters) [4]. It can be observed that, in general, bandgap increases when the cluster size decreases, being a nice agreement between the DFT-calculated bandgap values and those predicted by the simple Jellium model, showing at the same time that ligands play a relatively minor role.

Properties of Metal Clusters

Among the properties that make metal clusters unique and useful in many applications, the following ones can be highlighted here as examples:



Synthesis of Subnanometric Metal Nanoparticles, Fig. 4 (**a**) Energy-level occupations for spherical three-dimensional, harmonic, intermediate, and square-well potentials. (**b**) Self-consistent effective potential of a Jellium sphere corresponding to Na_{40} with the electron occupation of the energy levels (Reprinted with permission from Ref. [2]. Copyright 1993 by the American Physical Society)

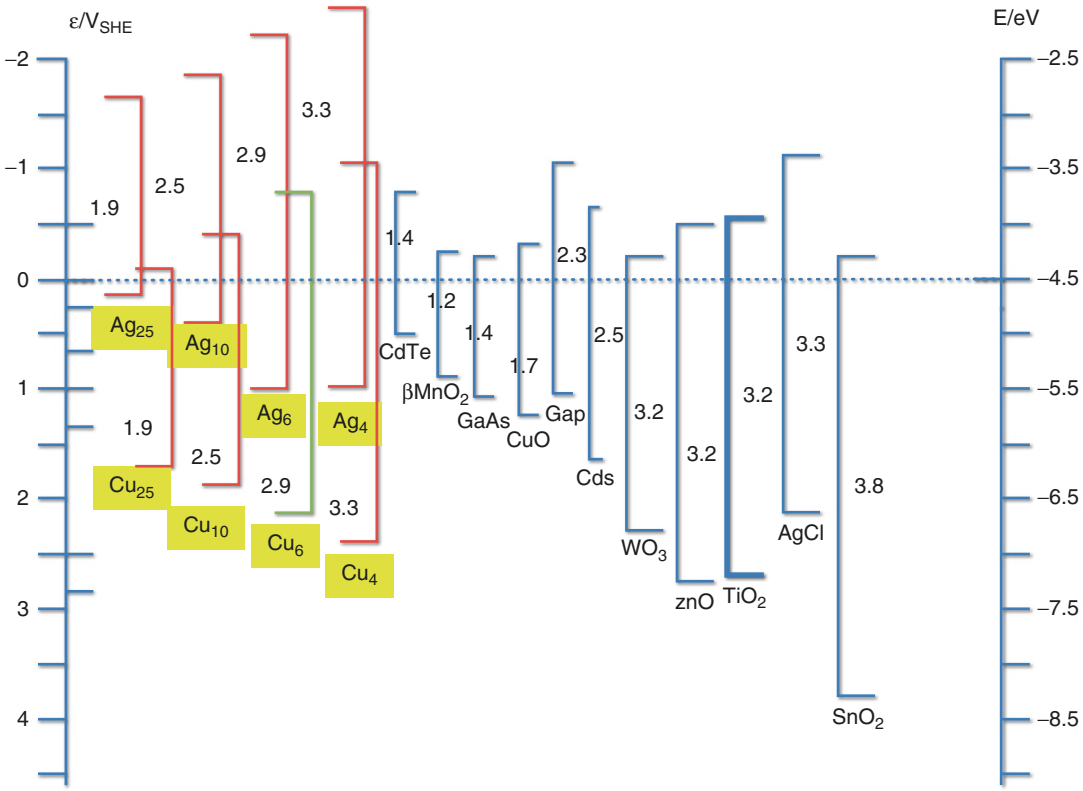
Photoluminescence

The molecular-like behavior of few-atom noble metal clusters, and the mentioned bandgap similar to that of semiconductors, allows the possibility of size-tunable electronic intraband transitions [5].

The appearance of these optical transitions with energies ranging from the UV-visible to the infrared region makes noble metal quantum dots ideal potential candidates for applications such as fluorescent labeling in biology or light-emitting sources in nanoelectronics. As an example, Fig. 7 shows the fluorescence of aqueous solutions of Cu_N clusters ($N < 14$) excited at 296 nm with quantum yields over 10 %, which are stable for more than 2 years.

Even though, in cases where the quantum yield is lower than that cited in previous example, fluorescent noble metal clusters are a very interesting alternative to semiconductor quantum dots as biological labels, due not only to their small hydrodynamic size and inert nature but especially because of their biocompatibility and also because they present reduced photobleaching compared to organic fluorophores. Water-soluble gold clusters have been, therefore, explored as fluorescent labels in biological experiments. Within this approach, Parak's group described [7] the use of stable gold clusters capped with dihydrolipoic acid (DHLA) and conjugated through EDC chemistry to biologically relevant molecules such as PEG, BSA, avidin, or streptavidin for biolabeling. Unconjugated clusters can also be nonspecifically uptaken by living cells (Fig. 8).

The last contributions in this field exploit a large Stokes shift displayed by a new type of specially synthesized clusters which are particularly interesting for biosensing applications since they present an intense fluorescent emission in the red spectra (highly desired due to its large signal-to-noise ratio). While the origin of this large Stokes shift – usually associated to a metal-ligand charge transfer – for this new class of materials is not totally understood, the technical development is going fast due to the facile synthesis procedures

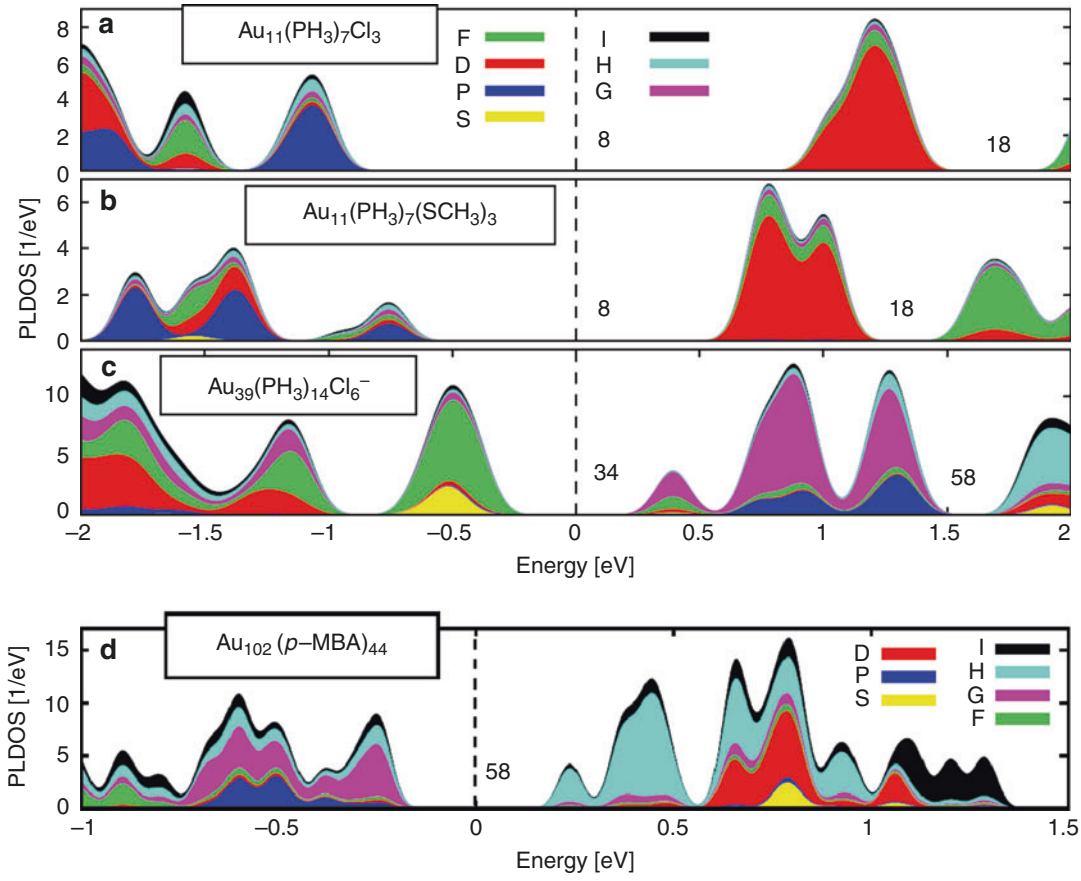


Synthesis of Subnanometric Metal Nanoparticles, Fig. 5 Schematic comparison between bandgaps of some silver and copper clusters (M_N, N = number of atoms) and those of well-known semiconductors. Bandgaps (E_g) were calculated from the spherical Jellium model ($E_g = E_F/N^{1/3}$; E_F = Fermi level), and the position of the conduction band, E_{CB}, was estimated by the formula $E_{CB} = \chi - E_F - \frac{1}{2} E_g$ (χ = electronegativity)

already developed [8, 9]. For example, Biang et al. [8] prepared histidine (His)- and 11-mercaptoundecanoic acid (MUA)-capped gold clusters, ($Au_{17}MUAH_4Hiss_{22}$), with strong fluorescence around 600 (nm) and a large lifetime ~7μs (see Fig. 9). The as prepared clusters were tested in HeLa cell lines and a marginal toxicity was observed. Fluorescence microscopy showed that such clusters were internalized into the cell cytoplasm by endocytosis. These novel clusters are ideal for deep tissue penetration and biological imaging by fluorescence microscopy. Kong et al. [9] reported another interesting example of highly fluorescent Au clusters using bovine pancreatic ribonuclease A (RNase A) as bio-template. The RNase A-encapsulated Au clusters show a strong fluorescence (quantum yield of 12 %) in the near-infrared region (682 nm) and exhibit a large Stokes shift (210 nm) and a single dominant fluorescence lifetime of ~1.5 μs. By coupling these Au clusters with vitamin B_{12}, a multifunctional platform was developed, suitable for simultaneous targeting and imaging of cancer at the cellular level using Caco-2 cell lines as in vitro model.

Catalysis

Small clusters have been found to own a catalytic activity not observed in their bulk analogs or even nanoparticles, which makes them very attractive as new catalytic materials. Quantum chemical calculations indicate that such high reactivity is due to undercoordination of the metal atoms forming the cluster [10]. Several families of metal clusters have been found to possess surprisingly high and selective catalytic activity when immobilized on a support. For example, platinum clusters with 8–10 atoms can be used as catalysts for the oxidative dehydrogenation of



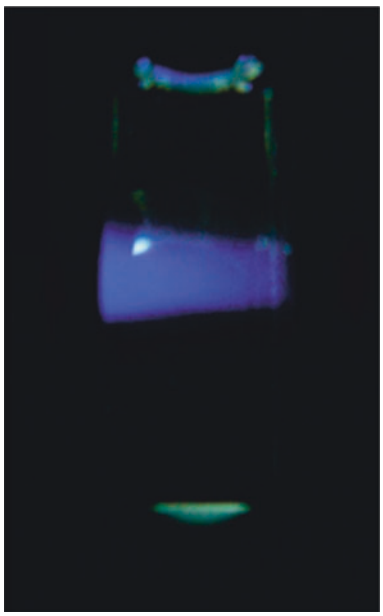
Synthesis of Subnanometric Metal Nanoparticles, Fig. 6 Electronic structure analysis of gold clusters. The angular-momentum-projected local electron density of states (PLDOS) (projection up to the l symmetry, i.e., $l = 6$) for the gold core in $Au_{11}(PH_3)_7Cl_3$ (**a**), $Au_{11}(PH_3)_7(SCH_3)_3$ (**b**), $Au_{39}(PH_3)_{14}Cl_6^-$ (**c**), and $Au_{102}(p\text{-MBA})_{44}$ (**d**). The zero energy corresponds to the middle of the HOMO-LUMO gap (*dashed line*). For plotting, each individual electron state is displayed by a Gaussian smoothing of 0.07 eV (0.03 eV in **d**). Shell-closing number is indicated for each case (For more details see Ref. [3]. Copyright 2008 National Academy of Sciences, USA)

Synthesis of Subnanometric Metal Nanoparticles, Table 1 Comparison between: (1) experimentally determined bandgaps for free gas phase gold cluster anions from photoelectron spectroscopy; (2) theoretical (DFT) values for HOMO-LUMO gaps of passivated gold cluster compounds that correspond to 8, 34, and 58 conduction-electron shell closings; and (3) bandgaps calculated through the Jellium model for clusters with the same number of gold atoms

	Experiment		Theory		Jellium model
Shell closing	Cluster	Gap (eV)	Cluster compound	Gap (eV)	$E_g = 5.32/N^{1/3}$ (eV)
8e ($1S^21P^6$)			$Au_{11}(PH_3)_7(SMe)_3$	1.5	2.4
8e			$Au_{11}(PH_3)_7Cl_3$	2.1	2.4
8e			$Au_{13}(PH_3)_{10}Cl_2^{3+}$	1.8	2.3
8e			$Au_{25}(SMe)_{18}^-$	1.2	1.3[a]
34e (8e + $1D^{10}2S^21F^{14}$)	$Au34^-$	1.0	$Au_{39}Cl_6(PH_3)_{14}^-$	0.8	1.0[a]
58e (34e + $2P^61G^{18}$)	$Au58^-$	0.6	Au_{102}(p-MBA)$_{44}$	0.5	0.6[a]
58e			$Au_{102}(SMe)_{44}$	0.5	0.6[a]

Adapted from Ref. [3]. Copyright 2008 National Academy of Sciences, USA
[a]For these clusters, a correction of −0.4 eV was used – see text

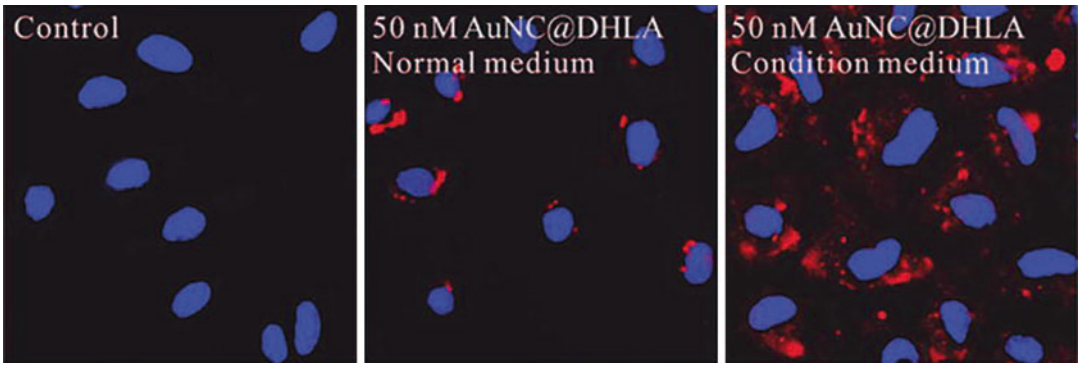
propane [11], while gold clusters with 6–10 atoms have been shown to be highly active for catalyzing propene epoxidation [12]. Recently, Harding et al. [13] studied the control and tunability of the catalytic oxidation of CO by Au_{20} clusters deposited on MgO surfaces, finding that the active site on the cluster is characterized by enhanced electron density, which activates the adsorbed O_2 molecule and promotes the bonding of CO.



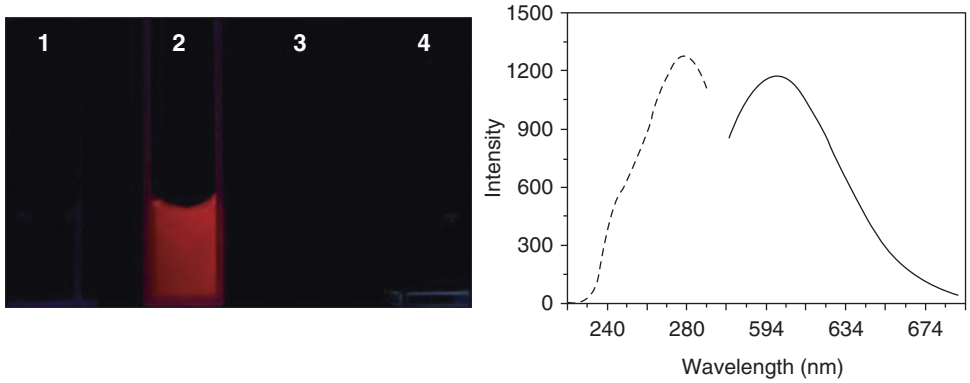
Synthesis of Subnanometric Metal Nanoparticles, Fig. 7 Blue emission observed for an aqueous solution of copper clusters ($N \leq 14$ atoms) under an excitation wavelength of 296 nm (Reprinted with permission from Ref. [6]. Copyright 2010 American Chemical Society)

Small metal clusters have been also found to display electrocatalytic activities different from the material in bulk or as nanoparticles [14]. Although this is a field not too much explored so far, these electrocatalytic properties make metal clusters to be promising materials in fuel-cell applications. As another interesting example of such properties, a recent report of their use to prevent pathologies, such as fetal alcohol syndrome, [15] is highlighted here: It has been proved that some clusters are able to electrocatalyze the oxidation of alcohols and prevent its cytotoxicity under physiological conditions and at very low potentials, like those found on living cells, as it is schematically represented in Fig. 10.

Differently from NPs and bulk surfaces, clusters present an optimum window for their size-dependent catalytic activity which is different for each type of reaction. Recently, Corma et al. carried on a deep theoretical and experimental study to unravel the exceptional catalytic activity shown by small Au clusters (between 5 and 10 atoms) for the aerobic thiophenol oxidation, comparable to sulfhydryl oxidase enzymes [16]. Under this approach, the existence of this optimum window size is qualitatively explained in terms of the position of the HOMO-LUMO bands of gold clusters and the bonding to the



Synthesis of Subnanometric Metal Nanoparticles, Fig. 8 Nonspecific uptake of unconjugated fluorescent DHLA-capped Au clusters by human endothelial cells. Cell nuclei were stained to yield *blue* fluorescence. *Red* fluorescence corresponds to the clusters. (see Ref. [7] for more information. Copyright 2009 American Chemical Society. Reprinted with permission)



Synthesis of Subnanometric Metal Nanoparticles, Fig. 9 A large Stokes shift displayed by specifically synthesized clusters allows to obtain intense fluorescent emission in the *red* spectra which, due to its high signal-to-noise ratio, is highly desired for biosensing applications (see Ref. [8] for a complete description. Reprinted with permission of The Royal Society of Chemistry)

thiols. A weak bonding to thiols is necessary for the catalysis to take place, thus activating the thiol, but a strong bonding inhibits the reaction. While big clusters –with a low HOMO band- and nanoparticles strongly bond to thiols, very small clusters (<5 atoms) –with a high HOMO band- cannot bond to them. This leads to the observed fact that only medium-size clusters, between 5 and 10 atoms, have the weak bonding necessary to activate the thiol.

Synthesis of Metal Clusters

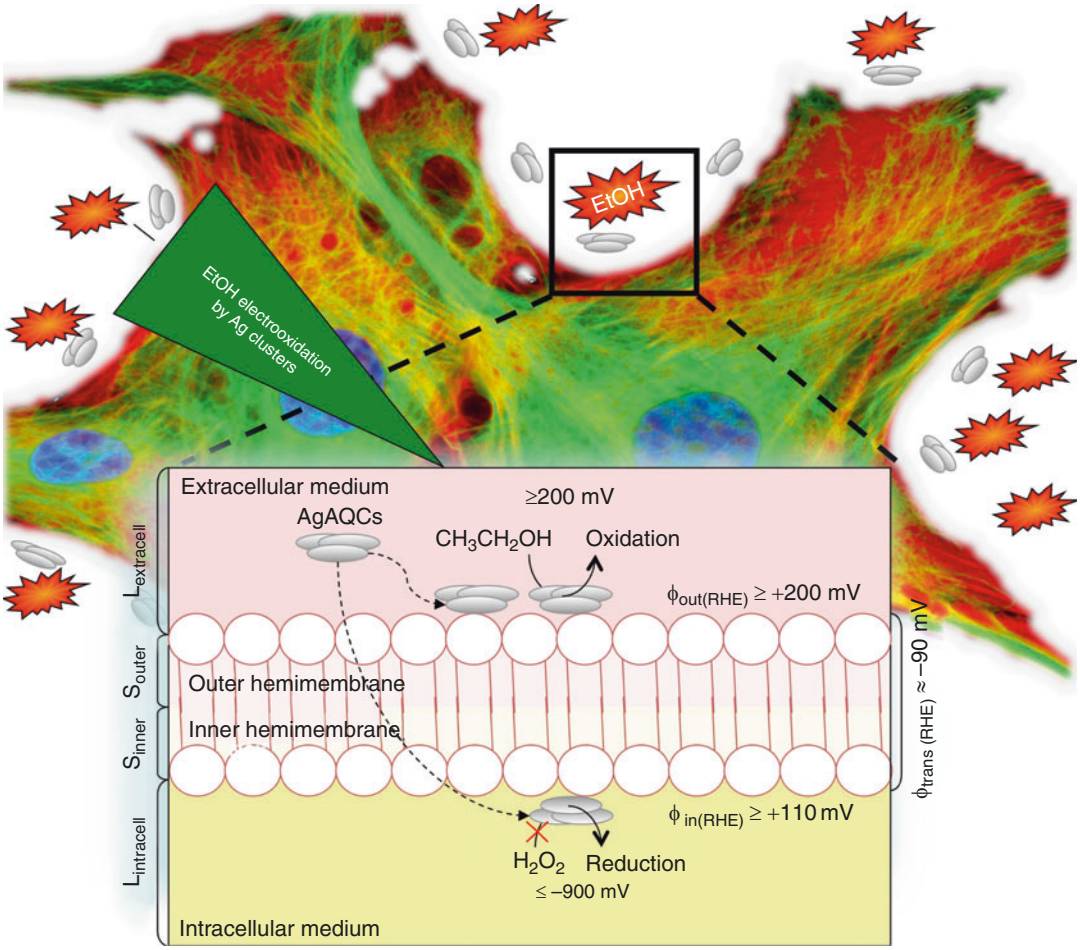
As it also occurs in the case of nanoparticles, there are two main approaches to the synthesis of metal clusters: top-down and bottom-up.

Bottom-Up

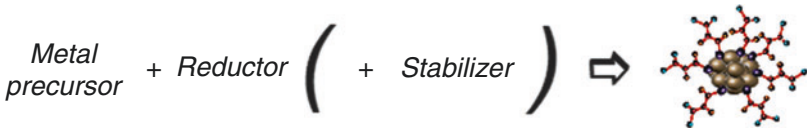
In bottom-up approaches, clusters are built from smaller structures at atomic level such as single atoms or ions (see Fig. 11). In this case, the most commonly used techniques differ depending on the size of the clusters to be synthesized. Wet chemical preparation of clusters includes the chemical reduction of metal salts in the presence of strong binding ligands [4, 17, 18] or cages [19] and by kinetic control using, e.g., microemulsions [20–22], electrochemical methods [6, 14, 23], etc.

For the synthesis of large clusters, a large variety of strong, protecting ligands can be used to control the growth of the primarily formed clusters and to stabilize them. Within this approach, we can mention the size-controlled synthesis of glutathione-capped gold clusters [24, 25]. The key of this beautiful design is based on the use of a reducing agent, gaseous CO, which establishes a mild reaction environment for a slow and well-controlled growth of Au clusters. The fine-tune kinetics is achieved by precisely controlling the pH, leading to the formation of a monodisperse set of products: $Au_{10\text{-}12}GSH_{10–12}$, $Au_{15}GSH_{13}$, $Au_{18}GSH_{14}$, and $Au_{25}GSH_{18}$ (Fig. 12).

For small clusters, the kinetic control, based on slow reaction rates provided either by soft reducing agents or low concentrations (as e.g., those provided by microemulsions, electrochemical methods, etc.), is the key point to their synthesis. It is generally believed that the synthesis of small clusters is extremely difficult to achieve because of the highly precise control of experimental conditions required to stop and then isolate the clusters as soon as they are formed. That belief is based on the theory of nucleation and growth (NGT), according to which, nuclei formed during the first steps of the chemical synthesis are only stable when they grow over a specific size called critical nucleus. Below that size, nuclei dissolve because of their large Laplace pressure. Above that size, they continue growing in order to reduce their surface energy by different mechanisms like autocatalysis, Ostwald ripening, etc., until the growth



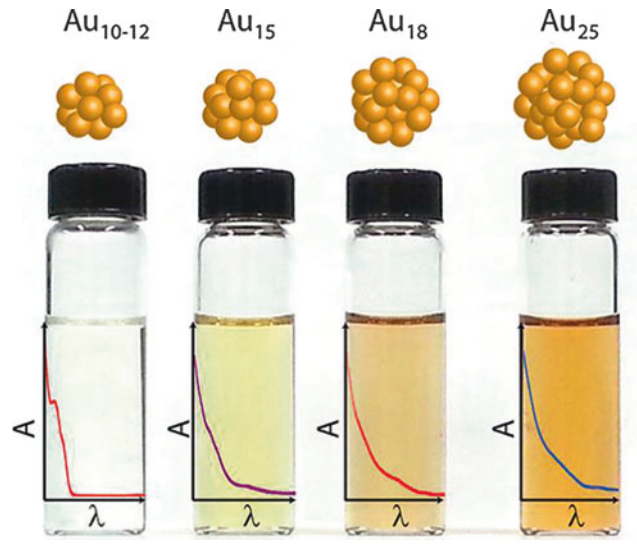
Synthesis of Subnanometric Metal Nanoparticles, Fig. 10 Schematic representation of the electrocatalysis of the oxidation of ethanol by silver clusters on the cellular membrane, preventing the alcohol toxicity in living cells (For more details, see Ref. [15]. Copyright 2010 American Chemical Society. Reprinted with permission)



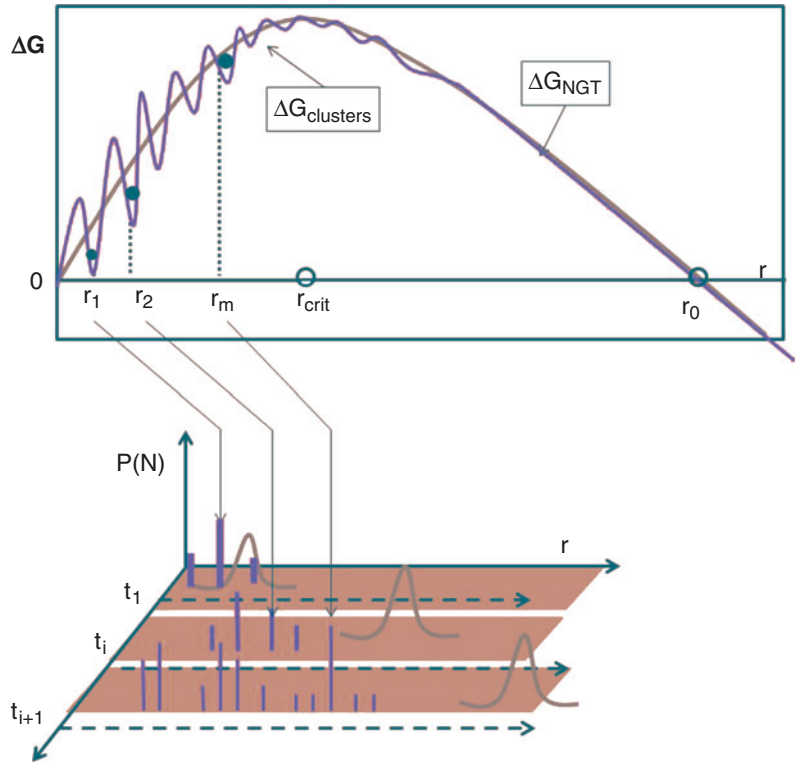
Synthesis of Subnanometric Metal Nanoparticles, Fig. 11 Schematic representation of bottom-up synthesis

is stopped by capping agents or other templates. However, such arguments, which correctly can be applied to fast reaction rates or high-temperature synthesis of nanoparticles and larger clusters having the same structure of the bulk material, are not correct when they apply to very small clusters, as those produced by kinetic control. As it was previously mentioned, numerous theoretical and experimental reports indicate that clusters can be especially stable because of their particular electronic and geometric structures different from the bulk.

In such a case, the use of macroscopic thermodynamic arguments, like those used in the theory



Synthesis of Subnanometric Metal Nanoparticles, Fig. 12 Glutathione-capped gold clusters, synthesized with a simple procedure (Reprinted from Ref. [25] with permission. Copyright 2014 American Chemical Society)



Synthesis of Subnanometric Metal Nanoparticles, Fig. 13 Schematic representation of the variation of free energy during the chemical synthesis of metal nanoparticles following the classical NGT compared to the one modified including the cluster effect in the NGT. Size distribution for cluster formation becomes a discrete magnitude that reflects the large stability of magic number clusters (Reprinted from Ref. [26] with permission from Elsevier)

of nucleation and growth, cannot be applied, as it was recently published in a revision of the NGT that takes into account the existence of stable clusters [26]. Figure 13 shows a comparison between the free energy evolution throughout a reaction of NP formation in solution, obtained using the classical NGT, and the modified one including clusters' existence. The cost in free energy for the growth of clusters evolves through a stepwise stability wells which correspond to the enhanced stability of stable clusters with magic numbers. Under this approach, clusters do not obey the

rule, assumed in the classical NGTs, of being continuously formed and dissolved and correspondingly the size distribution, P(N), which for classical NGT evolves in time as a continuous function of size becomes for the clusters' formation a discrete magnitude that reproduces the large stability of magic number clusters.

As said before, in a typical synthesis of nanoparticles, a fast reduction reaction is used, which means that the driving force (which can be represented by the difference in the electrochemical potentials of the metal to be synthesized and the reducing agent) is enough to drive the system directly to the first stable nucleus (r_{crit}) having the same structure of the bulk. However, when a small reaction rate is used, stable clusters like r_1, r_2, etc., are produced, and with an adequate control of the reaction kinetics, it is possible to stop the reaction at different times and, in this way, to obtain stable metal clusters with the desired size.

Following this scheme, silver clusters with less than 10 atoms have been synthesized, for example, in water-in-oil microemulsions consisting of a mixture of sodium bis(2-ethylhexyl) sulfosuccinate, isooctane, and water [21]. Synthesis conditions necessary for the appearance and isolation of stable clusters were achieved using small concentrations and a mild reduction agent as sodium hypophosphite monohydrate. Silver clusters produced in this way were found not only to be stable for years, but also to have a large bandgap (≈2.3 eV) and to present molecular-like paramagnetic properties and an intense photoluminescence.

In the same way, using electrochemical techniques, very small clusters can be synthesized, providing a good control of the cluster size. For example, stable poly(*N*-vinylpyrrolidone), PVP-protected gold clusters with only 2 or 3 atoms, could be synthesized through a simple electrochemical method based on the anodic dissolution of a gold electrode in the presence of the homopolymer PVP and subsequent electroreduction of the Au-PVP complexes [23]. The resulting clusters (Au_2 and Au_3) are the smallest ones that can be prepared and they display fluorescent (Au_2) and paramagnetic (Au_3) properties. By a similar procedure, highly fluorescent copper clusters stabilized with tetrabutylammonium nitrate, with less than 14 atoms, have also been reported [6]. These clusters can be dispersed in different solvents (both polar and apolar), are stable for several years, and display also similar fluorescent properties than those synthesized in microemulsions [22].

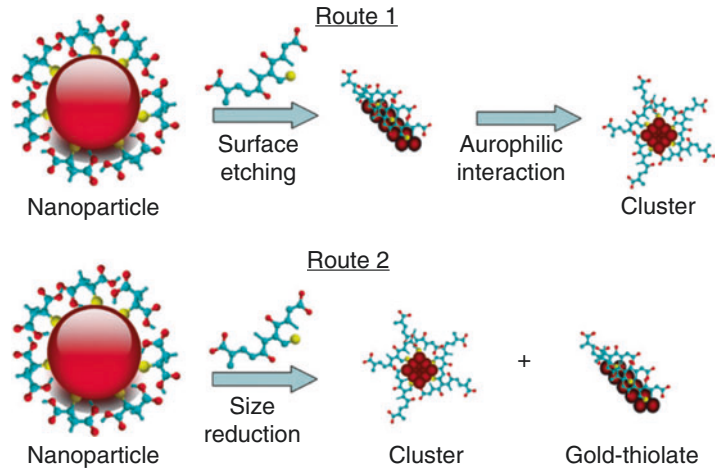
Finally, it is worth mentioning the use of hard cages to inhibit the growth and synthesize small clusters, as it was demonstrated very recently by Royon et al. [19], showing that silver-containing glasses can be used for preparing fluorescent materials, reducing the Ag ions in the matrix with laser pulses. In this way, they were able to prepare perennial high-capacity 3D-optical recording media.

Top-Down

In top-down approaches, clusters are synthesized from larger precursors such as nanoparticles or bulk metal. The most common top-down technique used for the synthesis of clusters is etching of nanoparticles (Fig. 14). In this technique, clusters are synthesized using the etching capacity of some ligands (like thiols), which remove the surface atoms of metallic nanoparticles or break them into smaller pieces leading to stable quantum clusters characterized by shell-closing magic numbers. Figure 13 can also help to clarify and understand this important top-down process in energy terms. Nanoparticles with sizes close to the critical nucleus ($r_{crit} < r < r_0$) can be etched, yielding stable clusters whose potential wells are located below the DG of the etched NPs ($DG > 0$, i.e., for sizes in the metastable region), making the process energetically favorable. Experimentally, etching is widely used as a top-down technique.

For example, etching of mercaptosuccinic acid-protected gold nanoparticles with excess glutathione has been used to yield small photoluminescent gold clusters [27]. In such work, clusters, either with 8 or 25 gold atoms, were produced by etching of previously synthesized nanoparticles with 4–5 nm core diameter. Selection of the final cluster size was possible by only varying the etching pH from ~3 (for 25 atom clusters) to 7–8 (for 8 atom clusters).

Multivalent coordinating polymers such as polyethylenimine can also be used to etch



Synthesis of Subnanometric Metal Nanoparticles, Fig. 14 Schematic representation of two possible routes for the formation of gold clusters by etching of mercaptosuccinic acid-capped gold nanoparticles (Reprinted from Ref. [27] with kind permission from Springer Science + Business Media B.V)

preformed colloidal gold nanocrystals, producing highly fluorescent water-soluble nanoclusters formed by 8 gold atoms [28].

In summary, it can be said that by appropriate use of the typical techniques developed for the synthesis of nanoparticles, clusters of different sizes down to 2 atoms can be produced, isolated, and used as stable materials with novel properties, which differ very strongly from the bulk materials and nanoparticles due to their totally different electronic and geometrical structure.

Cross-References

- ▶ Dry Etching Processes
- ▶ Nanoparticles
- ▶ Optical and Electronic Properties

References

1. Chou, M.Y., Cleland, A., Cohen, M.L.: Total energies, abundances, and electronic shell structure of lithium, sodium, and potassium clusters. Solid State Commun. **52**, 645–648 (1984)
2. de Heer, W.A.: The physics of simple metal clusters: experimental aspects and simple models. Rev. Mod. Phys. **65**, 611–676 (1993)
3. Walter, M., Akola, J., Lopez-Acevedo, O., Jadzinsky, P.D., Calero, G., Ackerson, C.J., Whetten, R.L., Grönbeck, H., Häkkinen, H.: A unified view of ligand-protected gold clusters as superatom complexes. Proc. Natl. Acad. Sci. **105**, 9157–9162 (2008)
4. Zheng, J., Zhang, C., Dickson, R.M.: Highly fluorescent, water-soluble, size-tunable gold quantum dots. Phys. Rev. Lett. **93**, 077402 (2004)
5. Zheng, J., Nikovich, P.R., Dickson, R.M.: Highly fluorescent noble metal quantum dots. Annu. Rev. Phys. Chem. **58**, 409–431 (2007)
6. Vilar-Vidal, N., Blanco, M.C., López-Quintela, M.A., Rivas, J., Serra, C.: Electrochemical synthesis of very stable photoluminescent copper clusters. J. Phys. Chem. C **114**, 15924–15930 (2010)
7. Lin, C.-A.J., Yang, T.-Y., Lee, C.-H., Huang, S.H., Sperling, R.A., Zanella, M., Li, J.K., Shen, J.-L., Wang, H.-H., Yeh, H.-I., Parak, W.J., Chang, W.H.: Synthesis, characterization, and bioconjugation of fluorescent gold nanoclusters toward biological labeling applications. ACS Nano **3**, 395–401 (2009)
8. Biang, P., Zhou, P., Liu, Y., Ma, Z.: One-Step fabrication of intense red fluorescent gold nanoclusters and their application in cancer cell imaging. Nanoscale **5**, 6161–6166 (2013)
9. Kong, Y., Chen, J., Gao, F., Brydson, R., Johnson, B., Heth, G., Zhang, Y., Wu, L., Zhou, D.: Near-infrared fluorescent ribonuclease-A-encapsulated gold nanoclusters. Nanoscale **5**, 1009–10017 (2013)
10. Heiz, U., Landman, U.: Nanocatalysis (Nanoscience and Technology). Springer, Berlin (2007)
11. Vajda, S., Pellin, M.J., Greeley, J.P., Marshall, C.L., Curtiss, L.A., Ballentine, G.A., Elam, J.W., Catillon-Mucherie, S., Redfern, P.C., Mehmood, F., Zapol, P.: Subnanometre platinum clusters as highly active and selective catalysts for the oxidative dehydrogenation of propane. Nat. Mater. **8**, 213–216 (2009)
12. Lee, S., Molina, L.M., López, M.J., Alonso, J.A., Hammer, B., Lee, B., Seifert, S., Winans, R.E., Elam, J.W., Pellin, M.J., Vajda, S.: Selective propene epoxidation on immobilized Au_{6-10}clusters: the effect of hydrogen and water on activity and selectivity. Angew. Chem. Int. Ed. **48**, 1467–1471 (2009)
13. Harding, C., Habibpour, V., Kunz, S., Farnbacher, A. N.-S., Heiz, U., Yoon, B., Landman, U.: Control and

manipulation of gold nanocatalysis: effects of metal oxide support thickness and composition. J. Am. Chem. Soc. **131**, 538–548 (2009)
14. Rodríguez-Vázquez, M.J., Blanco, M.C., Lourido, R., Vázquez-Vázquez, C., Pastor, E., Planes, G.A., Rivas, J., López-Quintela, M.A.: Synthesis of atomic gold clusters with strong electrocatalytic activities. Langmuir **24**, 12690–12694 (2008)
15. Selva, J., Martínez, S.E., Buceta, D., Rodríguez-Vázquez, M.J., Blanco, M.C., López-Quintela, M.A., Egea, G.: Silver sub-nanoclusters electrocatalyze ethanol oxidation and provide protection against ethanol toxicity in cultured mammalian cells. J. Am. Chem. Soc. **132**, 6947–6954 (2010)
16. Corma, A., Concepción, P., Boronat, M., Sabater, M. J., Navas, J., Yacamán, M.J., Larios, E., Posadas, A., López-Quintela, M.A., Buceta, D., Mendoza, E., Guilera, G., Mayoral, A.: Exceptional oxidation activity with size-controlled supported gold clusters of low atomicity. Nat. Chem. **5**, 775–781 (2013)
17. Schaeffer, N., Tan, B., Dickinson, C., Rosseinsky, M.J., Laromaine, A., McComb, D.W., Stevens, M.M., Wang, Y., Petit, L., Barentin, C., Spiller, D.G., Cooper, A.I., Lévy, R.: Fluorescent or not? Size-dependent fluorescence switching for polymer-stabilized gold clusters in the 1.1–1.7 nm size range. Chem. Commun. **34**, 3986–3988 (2008)
18. Negishi, Y., Nobusada, K., Tsukuda, T.: Glutathione-protected gold clusters revisited: bridging the gap between gold(I)-thiolate complexes and thiolate-protected gold nanocrystals. J. Am. Chem. Soc. **127**, 5261–5270 (2005)
19. Royon, A., Bourhis, K., Bellec, M., Papon, G., Bousquet, B., Deshayes, Y., Cardinal, T., Canioni, L.: Silver clusters embedded in glass as a perennial high capacity optical recording medium. Adv. Mater. **22**, 5282–5286 (2010)
20. López-Quintela, M.A.: Synthesis of nanomaterials in microemulsions: formation mechanisms and growth control. Curr. Opin. Colloid Interface Sci. **8**, 137–144 (2003)
21. Ledo-Suárez, A., Rivas, J., Rodríguez-Abreu, C.F., Rodríguez, M.J., Pastor, E., Hernández-Creus, A., Oseroff, S.B., López-Quintela, M.A.: Facile synthesis of stable subnanosized silver clusters in microemulsions. Angew. Chem. Int. Ed. **46**, 8823–8827 (2007)
22. Vázquez-Vázquez, C., Bañobre-López, M., Mitra, A., López-Quintela, M.A., Rivas, J.: Synthesis of small atomic copper clusters in microemulsions. Langmuir **25**, 8208–8216 (2009)
23. Santiago González, B., Rodríguez, M.J., Blanco, C., Rivas, J., López-Quintela, M.A., Martinho, J.M.G.: One step synthesis of the smallest photoluminescent and paramagnetic PVP-protected gold atomic clusters. Nano Lett. **10**, 4217–4221 (2010)
24. Yu, Y., Chen, X., Yao, Q., Yu, Y., Yan, N., Xie, J.: Scalable and Precise Synthesis of Thiolated Au_{10-12}, Au_{15}, Au_{18}, and Au_{25} Nanoclosters via pH controlled CO Reduction. Chem. Mater. **25**, 946–952 (2013)
25. Stamplecoskie, K.G., Kamat, P.V.: Size-dependent excited state behavior of glutathione –capped gold clusters and their light-harvesting capacity. J. Am. Chem. Soc. **136**, 11093–11109 (2014)
26. Piñeiro Redondo, Y., Buceta, D., Huseyinova, S., Cuerva, M., Perez Mariño, A., Dominguez, B., Calvo, J., López-Quintela, M. A.:Large stability and high catalytic activites of sub-nm metal (0) clusters: implications into the nucleation and growth theory. J. Colloid Interface Sci. **449**, 279–285 (2015)
27. Habeeb Muhammed, M.A., Ramesh, S., Sinha, S.S., Pal, S.K., Pradeep, T.: Two distinct fluorescent quantum clusters of gold starting from metallic nanoparticles by pH-dependent ligand etching. Nano Res. **1**, 333–340 (2008)
28. Duan, H., Nie, S.: Etching colloidal gold nanocrystals with hyperbranched and multivalent polymers: a new route to fluorescent and water soluble atomic clusters. J. Am. Chem. Soc. **129**, 2412–2413 (2007)

Synthesized Conductance Injection

► Dynamic Clamp

Synthesized Ionic Conductance

► Dynamic Clamp

Synthesized Synaptic Conductance

► Dynamic Clamp

Synthetic Biology

Soichiro Tsuda
School of Chemistry, University of Glasgow, Glasgow, UK

Synonyms

Synthetic genomics

Definition

Synthetic biology is the design and construction of biological components (e.g., enzymes, gene circuits, and whole cells) from scratch or from standardized parts.

Overview

The term "synthetic biology" was first used in 1980 by Barbara Hobom to describe genetically engineered bacteria using recombinant DNA technology. In 2000, the term reappeared again by Eric Kool and others at the annual meeting of the American Chemical Society to refer to synthesis of unnatural chemical products (e.g., organic molecules) as an extension of synthetic chemistry [1].

Although the scope of synthetic biology has explosively expanded in the last decade, synthetic biology can be best described as "efforts to redesign life," that is, the design and construction of biological components (e.g., enzymes, gene circuits, and whole cells) using living cells. Similar approaches have already been taken long before the advent of synthetic biology, for example, by genetic engineering, which is one of the roots of synthetic biology. However, what makes synthetic biology distinctive from other preceding biological research fields is its strong engineering perspective. While genetic engineering focused on modifying individual genes and pathways, synthetic biology does on whole systems of genes and gene products. It aims at designing and constructing the behavior of organisms to perform new tasks [2]. Computer engineering analogy is often used to explain the goal of synthetic biology: In order to build a computer, it involves several hierarchical layers. The most fundamental layer of computers is the electrical components, such as resistor, transistor, capacitor, etc. Based on them, a higher-level structure, logic gates (AND, OR, NOT, XOR, etc.) are built, which are further integrated into even upper layer, integrated circuits, and so on. Biological systems can be described similarly: DNA, RNA, and proteins are the most fundamental components of biological cells. These components form gene circuit networks, which are parts of intracellular signal pathways, and so on. Thus, synthetic biology research involves the developments of biological components in each layer for engineering new biological functions: standardized fundamental biological parts, gene circuit construction, and whole biological cells, and biological and chemical products synthesized by living cells.

Standardized Biological Parts

To engineer complex systems, the standardization of components used in the systems (in this case biological systems) is necessary. Tom Knight and Drew Endy proposed the BioBrick standard biological parts as basic units for synthetic biological systems and set up an online registry of BioBrick parts where anyone can contribute in an open-source manner [3–5]. It provides a collection of well-characterized and standardized building blocks that can be used for the "plug-and-play" design and construction of unnatural biological systems. Using BioBrick standard parts, a biological engineer can program a living cell, as an electrical engineer can program an electrical circuit using electrical components. BioBrick parts are mostly DNA sequences with functionality, such as promoters, plasmids, and ribosome binding sites. Not only natural components, biologically inspired molecules can be BioBrick parts, for example, DNA analog PNA, in which the sugar-phosphate backbone is altered by *N*-(2-aminoethyl)-glycine units linked by peptide bonds [6].

Gene Circuit Design and Construction

Having set the standards for biological components, the next step is to develop minimum functional components from the standard parts. They are typically gene circuits that express particular proteins. These single circuits are connected together to construct a gene regulatory network. Network motifs, such as cascades, feedfoward, and feedback, are commonly used for the design of artificial regulatory network (and they are also

commonly found in natural biological systems) [7]. Cascades are series-connected gene circuits that output (protein) of an upstream gene circuit regulates expression of its immediately downstream gene. Feedforward motif involves a master regulatory gene that influences downstream genes through noncircular pathways. When the feedforward loop is introduced in a network, for example, it is possible to construct a network that shows a nonlinear non-monotonic change in the output. For example, a sigmoid function with a steep transition can be implemented with two gene circuits connected by negative (inhibitory) feedforward motif [8]. Basu and coworkers implemented a pulse-generator system using the motif [9]. When the input stimulus is monotonously changed from low to high, the output of the pulse-generator network changes from low to high and then back to low. In addition, the pulse amplitude and delay reflect not only the input concentration but also its rate of change. Feedback is a common biological regulatory motif not only for intracellular gene networks but also for intercellular control (e.g., negative feedback regulation of hormonal control). Feedback loop allows various complex regulations over expression of genes: noise reduction [10], expression level control [11], bistable gene toggle switch [12], autonomous genetic oscillator [13], etc. In principle, any artificial regulatory scheme can be implemented with the combination of the above network motifs.

Although standardization and other engineering concepts, such as abstraction, modularity, and reliability, greatly contribute to the speed and tractability of gene circuit design, one should note the liquid nature of biological systems and components. In contrast to solid electrical components, biological components, DNA, RNA, and proteins, all operate in the liquid solution which provide an appropriate "context" (e.g., pH, temperature, materials, energy, etc.) for the systems to function. This cellular context dependence and sensitivity make it difficult for biological components to be modular and interchangeable. A gene circuit from naturally occurring system is likely to be optimized to a specific cellular context through evolution. Thus, it may not function when it is placed in an artificial context. Ron Weiss pointed out the need to understand "how the function of a module or an entire biological system can be derived from the function of its component parts "and establish" the biological rules of composition" in order to build biological components and modules [2].

Minimal Life

Craig Venter and his coworkers took the synthetic gene design to its extreme. They proposed to synthesize a complete genome from scratch and create "artificial life." In 2003, they have synthesized the whole genome of Phi-X 174, a 5386 base pair bacteriophage, from synthetic oligonucleotides, which took them only 14 days to complete the whole process [14]. The synthesized genome was injected into *E. coli* bacteria and confirmed to be infectious. After the success in synthesized virus, they took a further step to create a synthesized organism that has a minimal genome. A bacterial genome of *Mycoplasma genitalium*, which is one of the smallest known genomes in any living organisms (517 genes, 580,000 DNA base pairs), was chosen as a base for the synthesis. They removed genes in the genome that are considered redundant and injected into a *M. capricolum* cell in which the original genome was removed in order to test if the cell with the modified genome can be alive. In 2010, they reported that the size of the genome was eventually reduced to approximately one mega base pair that are minimal and sufficient to be considered alive. They called the cell *Mycoplasma laboratorium* which is the first artificial life, of which parent is a digital computer [15].

Another important aspect of minimal life in synthetic biology is "cells as chassis." The context dependence of biological components requires the constant maintenance of appropriate cellular context. Artificially synthesized biological components and modules are not exception and therefore require cellular context and its maintenance to perform tasks. Apart from the synthetic genome mentioned above, thus, all the other parts of protein expression process (transcription, RNA processing, translation, protein folding,

amino-acid modifications) can be subjects of synthetic biology research [16]. The construction of a minimal cell from scratch, as a basic unit for proving cellular context, is also an active area in this sub-research field. Lipid vesicles (liposomes) are commonly used as minimal bioreactors to simulate biological cells surrounded by lipid bilayer membrane. Functions of biological cells, such as membrane division, fusion, protein expression, and self-replication, have been attempted to replicate inside vesicles [17]. In recent years, microfluidic techniques, called digital microfluidics, have been applied to the production of lipid vesicles [18]. These techniques allow high-throughput production and screening of protein expression in the microfluidic devices and are expected to facilitate the more precise control of experimental conditions.

Applied Biological and Chemical Synthesis

Practical applications of synthetic biology are expected to be quite enormous because one of ultimate goals in synthetic biology is to create synthetic organisms that perform commercially useful tasks. For example, Craig Venter's group developed the synthetic genome described above so that it could be served as a platform for future synthetic organisms on which new functions (i.e., gene circuits) can be added. Since 2009, Venter's company, Synthetic Genomics Incorporated (SGI), has been working with Exxon Mobil Corporation to develop synthetic algae that produce next-generation biofuel, such as ethanol or hydrogen [19]. They also claim it would be possible to synthesize organisms which can fix carbon dioxide using photosynthesis to mitigate climate change [20].

Another example is medical and pharmaceutical applications of synthetic biology. Microbes, typically bacteria and yeast, are engineered and used as "chemical factory" to produce chemical products that are difficult to synthesize artificially. This process inevitably involves metabolic pathway engineering of living cells and therefore requires management of complex and emergent metabolic processes [21]. Keasling and coworkers developed synthetic bacteria for the antimalarial drug production [1, 22]. The drug, artemisinin, extracted from *Artemisia annua L* (sweet wormwood) is known to be the most effective drug for the treatment of malaria, but expensive to produce because of short supply of the wood. The engineered bacteria incorporate several genes originally from bacteria *E. coli* and yeast *S. cerevisiae* to produce artemisinic acid, a biogenetic precursor of artemisinin. First, MevT operon encodes enzymes that transform acetyl-CoA into mevalonate. Enzymes encoded by MBIS operon then transform mevalonate into farmesyl pyrophosphate (FPP). Additionally, genes for amorphadiene synthase (transforming FPP into amorphadiene) and oxidase and redox partners (amorphadiene into artemisinic acid) were introduced. Although *E. coli* has a native pathway, 1-deoxy-D-xylulose 5-phosphate (DXP) pathway, to produce FPP, it has been found that the engineered mevalonate pathway produces more amorphadiene than the native pathway. Artemisinic acid purified from bacterial extracts is transformed to artemisinin using established chemistry.

Cross-References

- ▶ Computational Systems Bioinformatics for RNAi
- ▶ Liposomes
- ▶ Nanotechnology Applications in Polymerase Chain Reaction (PCR)
- ▶ Structure and Stability of Protein Materials

References

1. Benner, S.A., Sismour, A.M.: Synthetic biology. Nat. Rev. Genet. **6**(7), 533–543 (2005)
2. Andrianantoandro, E., Basu, S., Karig, D.K., Weiss, R.: Synthetic biology: new engineering rules for an emerging discipline. Mol. Syst. Biol. **2**(1), 2006.0028 (2006)
3. The BioBricks Foundation. http://bbf.openwetware.org/
4. Knight, T.: Idempotent vector design for standard assembly of biobricks standard biobrick sequence

interface. MIT Synthetic Biology Working Group Report, MIT Artificial Intelligence Laboratory. 1–11 (2003)
5. Endy, D.: Foundations for engineering biology. Nature **438**(7067), 449–453 (2005)
6. Nielsen, P.E., Egholm, M.: An introduction to peptide nucleic acid. Curr. Issues Mol. Biol. **1**(1–2), 89–104 (1999)
7. McDaniel, R., Weiss, R.: Advances in synthetic biology: on the path from prototypes to applications. Curr. Opin. Biotechnol. **16**(4), 476–483 (2005)
8. Hooshangi, S., Thiberge, S., Weiss, R.: Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. Proc. Natl. Acad. Sci. U. S. A. **102**(10), 3581–3586 (2005)
9. Basu, S., Mehreja, R., Thiberge, S., Chen, M.-T., Weiss, R.: Spatiotemporal control of gene expression with pulse-generating networks. Proc. Natl. Acad. Sci. U. S. A. **101**(17), 6355–6360 (2004)
10. Becskei, A., Serrano, L.: Engineering stability in gene networks by autoregulation. Nature **405**, 590–593 (2000)
11. Nevozhay, D., Adams, R.M., Murphy, K.F., Josic, K., Balzsi, G.: Negative autoregulation linearizes the dose–response and suppresses the heterogeneity of gene expression. Proc. Natl. Acad. Sci. U. S. A. **106**(13), 5123–5128 (2009)
12. Gardner, T.S., Cantor, C.R., Collins, J.J.: Construction of a genetic toggle switch in *Escherichia coli*. Nature **403**(6767), 339–342 (2000)
13. Judd, E.M., Laub, M.T., McAdams, H.H.: Toggles and oscillators: new genetic circuit designs. Bioessays **22**(6), 507–509 (2000)
14. Smith, H.O., Hutchison, C.A., Pfannkoch, C., Venter, J.C.: Generating a synthetic genome by whole genome assembly: x174 bacteriophage from synthetic oligonucleotides. Proc. Natl. Acad. Sci. U. S. A. **100**(26), 15440–15445 (2003)
15. Gibson, D.G., Glass, J.I., Lartigue, C., Noskov, V.N., Chuang, R.-Y., Algire, M.A., Benders, G.A., Montague, M.G., Ma, L., Moodie, M.M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E.A., Young, L., Qi, Z.-Q., Segall-Shapiro, T.H., Calvey, C.H., Parmar, P.P., Hutchison, C.A., Smith, H.O., Venter, J.C.: Creation of a bacterial cell controlled by a chemically synthesized genome. Science **329**(May), 52–56 (2010)
16. Forster, A.C., Church, G.M.: Towards synthesis of a minimal cell. Mol. Syst. Biol. **2**(45), 45 (2006)
17. Luisi, P.L., Ferri, F., Stano, P.: Approaches to semi-synthetic minimal cells: a review. Naturwissenschaften **93**(1), 1–13 (2006)
18. Theberge, A.B.., Courtois, F., Schaerli, Y., Fischlechner, M., Abell, C., Hollfelder, F., Huck, W. T.S.: Microdroplets in microfluidics: an evolving platform for discoveries in chemistry and biology. Angew. Chem. Int. Ed. **49**(34), 5846–5868 (2010)
19. Synthetic Genomics Inc. Press Release http://www.syntheticgenomics.com/media/press/71409.html. Accessed 14 July 2009
20. ETC group report, Extreme Genetic Engineering: An Introduction to Synthetic Biology, http://www.etcgroup.org/content/extreme-genetic-engineering-introduction-synthetic-biology (2007) Accessed 7 May 2015
21. Khosla, C., Keasling, J.D.: Metabolic engineering for drug discovery and development. Nat. Rev. Drug Discov. **2**(12), 1019–1025 (2003)
22. Keasling, J.: Synthetic biology for synthetic chemistry. ACS Chem. Biol. **3**(1), 64–76 (2008)

Synthetic Genomics

► Synthetic Biology

Synthetic Lubricants

► Boundary Lubrication

Systems Level Data Mining for RNAi

► Computational Systems Bioinformatics for RNAi

T

TEM

► Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells

Terahertz

► Terahertz Technology for Nano Applications

Terahertz and Infrared Plasmonics with Unconventional Materials

F. D'Apuzzo[1], M. Autore[2], P. Di Pietro[3], F. Giorgianni[2], O. Limaj[4], I. Lo Vecchio[2], A. Starace[2] and S. Lupi[5]
[1]Istituto Italiano di Tecnologia and Sapienza University of Rome, Rome, Italy
[2]Department of Physics, Sapienza University of Rome, Rome, Italy
[3]INSTM Udr Trieste-ST and Sincrotrone Trieste, Trieste, Italy
[4]EPFL IBI-STI-BIOS, Lausanne, Switzerland
[5]IOM-CNR and Department of Physics, Sapienza University of Rome, Rome, Italy

Synonyms

Infrared plasmonics beyond noble metals; Low carrier density materials for infrared plasmonics

Definition

Collective electronic excitations confined at metal surfaces play a key role in a wide spectrum of science, ranging from physics and materials science to biology and biomedicine [1]. These excitations, called surface plasmons (SPs), are determined by the long-range nature of the Coulomb interaction among electrons in metals and show a well-defined frequency/wave vector dispersion described by Eq. 1 which is represented in Fig. 1.

$$v(K_x) = \frac{c}{2\pi}\sqrt{\frac{\varepsilon_1^m + \varepsilon_1^d}{\varepsilon_1^m \varepsilon_1^d}}K_x \qquad (1)$$

Here K_x is the propagation wavevector along a direction x at the metal surface, and ε_1^d and ε_1^m the real parts of the dielectric functions of the dielectric medium and metal, respectively. Surface plasmons cannot be directly coupled to propagating electromagnetic radiation because the conservation of momentum is not satisfied in the photon absorption process. One way to overcome this limitation is to break the continuous translational symmetry of the system, thus providing the necessary extra wavevector by a patterned 2D surface with a periodic subwavelength structure [2].

Due to their intrinsic two-dimensional nature, SPs exhibit a strong confinement along the direction perpendicular to $\boldsymbol{K}$ (see inset in Fig. 1) and, as

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Terahertz and Infrared Plasmonics with Unconventional Materials,

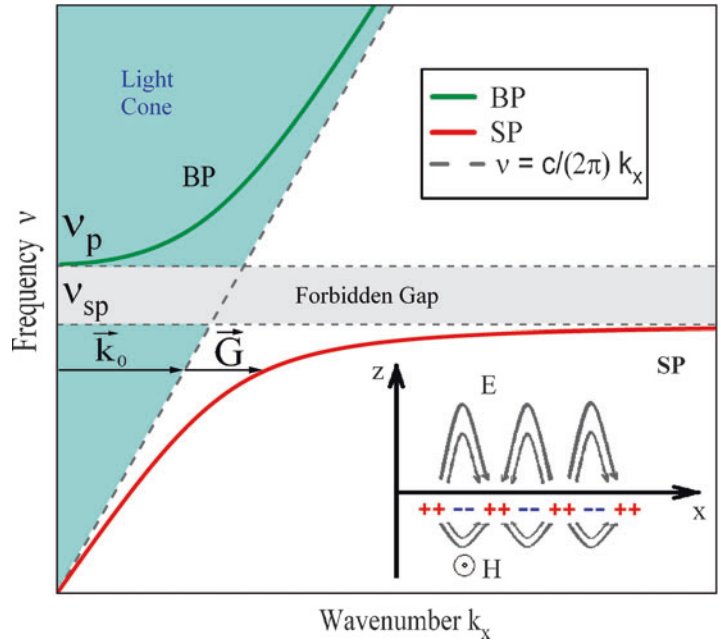


Fig. 1 Dispersion relation for the surface plasmon. The *dashed line* represents the light dispersion $v = ck/2\pi$. In the *inset*, the SP propagation along a metal-dielectric interface is shown. SP electromagnetic field intensity is highest at the surface and decays exponentially away from the interface (see *inset*). An SP can be viewed as a special type of light wave propagating along the metal surface

a consequence, a strong light confinement. This implies that also in the longitudinal direction, the SP wavelength, λ_{SP}, is always smaller than the electromagnetic excitation wavelength, λ_0. The strong confinement makes SPs extremely sensitive to any variation of the dielectric medium at the interface, and this property has opened the path to a variety of bio-sensing applications [1].

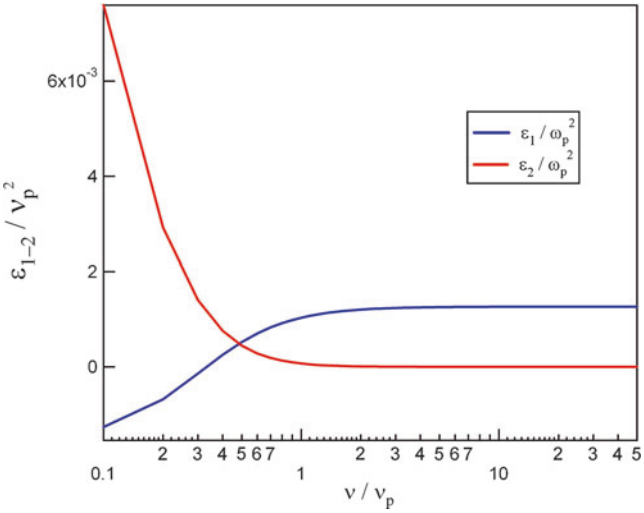
In particular, significant efforts have been made to achieve SP-based biosensors working in the visible spectrum: several devices with and without nanoscale patterns have been designed, exploring a wide variety of geometries and exploiting different configurations for the SP coupling with electromagnetic radiation. Current applications of visible plasmonics include subwavelength waveguiding, near-field optical microscopy, tip-enhanced Raman scattering, surface-enhanced Raman scattering, waveguides and modulators in integrated optics, and surface plasmon resonance analysis for the label-free detection of biological molecules [3]. Because of this high sensitivity, those techniques have exploited SPs as probes for detecting small amounts of target substances [5] as gases [4], environmental pollutants [4], picomoles of proteins [6], or dangerous pathogens [7].

The behavior of surface plasmons in the near-infrared and the visible spectral region is largely dominated by loss in the metal and by the electron lifetime. Indeed, the optical (and plasmonic) properties of metals can be described in terms of the Drude model [8] where the (complex) permittivity can be written as

$$\begin{aligned}\tilde{\varepsilon}(v) &= \varepsilon_1(v) + i\varepsilon_2(v) \\ &= \varepsilon_\infty - \frac{v_p^2}{v^2+\Gamma^2} + i\frac{v_p^2\Gamma}{(v^2+\Gamma^2)v}\end{aligned} \quad (2)$$

where v_p is the plasma frequency ($v^2{}_p \sim n$ where n is the electron concentration in the metal), Γ is the Drude scattering rate ($\Gamma \sim 1/\tau$ where τ is the electron lifetime), and ε_∞ is the high-frequency contribution (background permittivity).

The frequency dependence of ε_1 and ε_2 normalized to $v_p{}^2$ is represented in Fig. 2. Here, one has $\varepsilon_\infty = 5$ and $\Gamma = 3$ THz. One can see that ε_1 is negative below v_p reaching ε_∞ at higher frequency. The real and the imaginary part of the permittivity play a crucial role in the performance of plasmonic devices (e.g., a rule of thumb in plasmonics is $\varepsilon^m{}_1 = -\varepsilon^d{}_1$), and all characteristics of SP, such as dispersion relation, propagation



Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 2 The real ε_1 and imaginary part ε_2 of the dielectric permittivity for a Drude metal (see Eq. 2 in the text). The frequency axis is related to the plasma frequency v_p, while ε_1 and ε_2 to v_p^2

length, and plasmonic wavelength, depend on the permittivity function of the metal [2]. All plasmonic figures of merit (FOM) like the ratio between the plasmon propagation length and its wavelength ($\mathrm{FOM}_{\mathrm{prop}} = L_{\mathrm{prop}}/\lambda_{\mathrm{SP}}$) and the ratio between the free space wavelength λ_0 and the plasmon lateral decay length L_z ($\mathrm{FOM}_{\mathrm{conf}} = \lambda_0/L_z$) strongly depend on metal losses [2]. Actually, metal losses depend on Γ and scale with ε_2, i.e., with v_p^2, which is proportional to the concentration of carriers. Hence, a small ε_2 can be achieved by either reducing Γ or by a smaller carrier concentration or preferably both.

Probably, the confinement $\mathrm{FOM}_{\mathrm{conf}}$ is the most important figure of merit in plasmonics. This ratio sets the quality of confinement of the free-space electromagnetic field in small volumes which gives rise to a strong enhancement of the electric field around plasmonic structures. The lateral decay length naturally increases proportionally to the wavelength, thus hindering confinement (and all its applications) when going from the visible to the infrared and THz spectral region. As reported in standard plasmonic literature [2, 9], the decay length L_z can be computed as in the following:

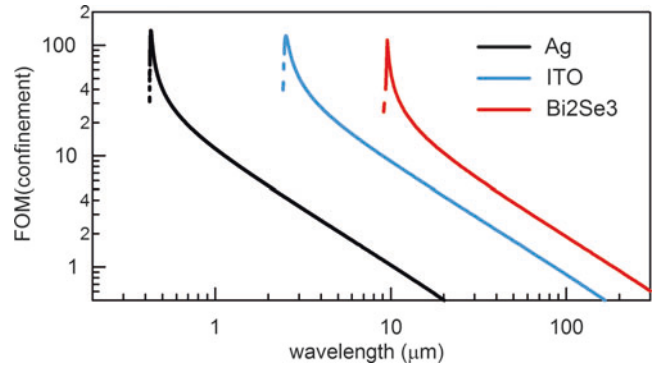
$$L_Z = \frac{1}{2\sqrt{\varepsilon_d\left(\frac{2\pi\nu}{c}\right)^2 - K_x^2}} \tag{3}$$

where K_x can be computed from Eq. 1. From the above relation, it is evident that as the SP dispersion drifts away from the light line $\left(\varepsilon_d\left(\frac{2\pi\nu}{c}\right)^2\right.$ term in Eq. 3), as evident in Fig. 1 at the frequency v_{SP}, there is a consequent decrease of decay length, i.e., higher energy confinement to the metal-dielectric interface.

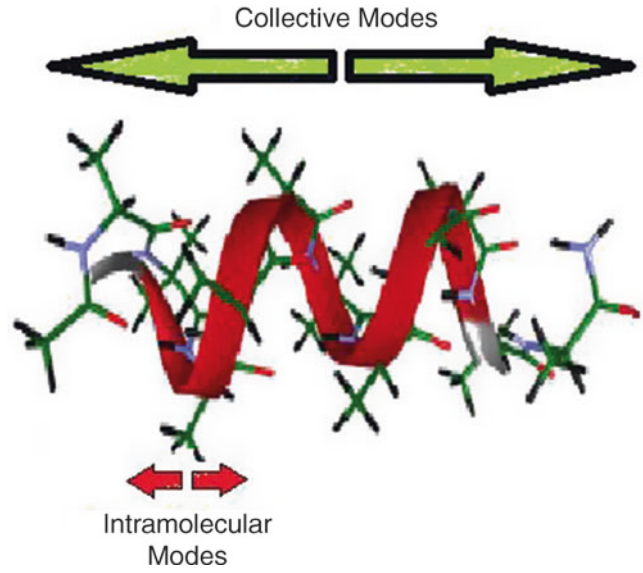
The $\mathrm{FOM}_{\mathrm{conf}}$ for Ag (representing the class of noble metals) is reported as a function of wavelength in Fig. 3. A value of the $\mathrm{FOM}_{\mathrm{conf}}(\mathrm{Ag}) \sim 100$ accounts for an energy localization down to the nanoscale. In other terms in the visible spectral range, the electromagnetic energy is localized in 1/100th of its wavelength.

However, the confinement for silver worsens toward infrared and THz wavelengths, reaching two orders of magnitude lower than its peak value. Nevertheless, the implementation of new materials with a lower carrier concentration can retrieve nanoscale confinement in the mid-infrared (such as ITO) and in the THz frequency range (topological insulator), as reported in Fig. 3. Another need for new plasmonic materials comes from the fact that many applications, such as transformation optics devices, require that the imaginary part of a metal's dielectric function be small.

A loss reduction cannot be easily obtained in standard metals where the carrier concentration $n \sim 10^{22} - 10^{23}/\mathrm{cm}^3$ sets the plasma frequency



Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 3 The plasmonic confinement figure of merit FOM_{conf} versus frequency from THz to visible Ag (standard metal), ITO (transparent metal oxide), and Bi_2Se_3 (topological insulator)



Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 4 Vibrational collective and intramolecular mode fingerprints can be observed in the THz and mid-IR parts of the electromagnetic spectrum

toward the near-IR/visible range. Moreover, such a high carrier density cannot be changed much with the application of electric fields, optical fields, or temperature, etc. Then, implementation of new materials beyond standard metals may both extend, plasmonic properties like strong matter-radiation interaction, light confinement, and electric field enhancement toward the mid-infrared (mid-IR) and the terahertz (THz) region of the electromagnetic spectrum and provide dynamical tunability of the plasmonic devices. Indeed, plasmonic devices working in those spectral regions are particularly important as intramolecular vibrations associated to specific chemical groups like $C = O$, CH_2, CH_3, included in the macromolecular structures, absorb radiation in the mid-IR (Fig. 4). Intermediate and collective macromolecule vibrational modes, instead, are associated with THz resonances, and these modes are particularly sensitive to macromolecule conformation (Fig. 4).

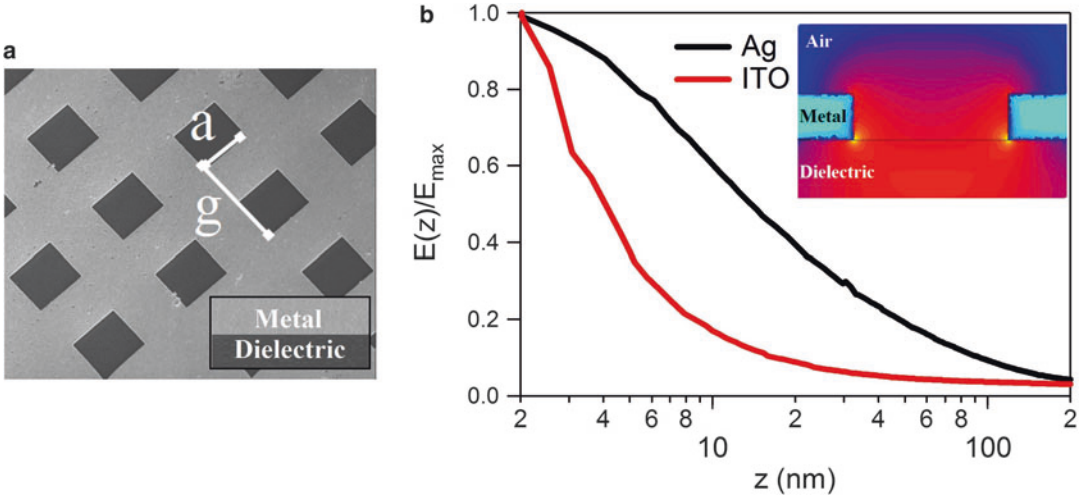
In the following, we will present different examples of unconventional materials that can be used for plasmonic applications both in the infrared and terahertz regions. In particular, we will discuss performances of transparent metallic oxides (TMO) in the mid-IR and in the THz range the properties of two-dimensional electron gas in topological insulator materials and graphene, and superconducting electrons in high-Tc superconductors.

Mid-Infrared Plasmonics on Transparent Metallic Oxides

Oxide semiconductors such as zinc oxide, cadmium oxide, and indium oxide can be made metallic through doping [10]. As the plasma frequency can be varied in a broad range from mid-IR to near-IR and the pristine semiconductors have a large electronic gap, they are transparent in the visible range and metallic at lower frequencies. These materials are then known as transparent metallic oxides (TMOs) and are particularly used for applications in display panels and photovoltaics. The most popular TMO is indium tin oxide (ITO), which shows, if highly doped, a high DC conductivity (sheet resistance $<10\ \Omega$/square). Moreover, its physical/chemical properties can be easily integrated in standard lithographic fabrication, providing a useful platform for infrared plasmonics [11].

As illustrated in Fig. 1, when trying to excite SP at low frequencies, the dispersion relation is tangent to the light dispersion; thus, the SP mode acquires a strong light-like character. This means that most of the energy of the mode is in the dielectric, with only small contribution of the electron plasma oscillation, thus resulting in weak confinement. In this regime, diffraction theory sufficiently accounts for optical behavior of the device; the metal can be approximately described with a perfect conductor with $\varepsilon_2^m = 0$ and $\varepsilon_1^m = -\infty$. If a periodic perturbation is exploited for SP coupling mechanism (see Fig. 4), then the device will behave as a diffraction grating with weak or no plasmonic effects. Noble metals working at infrared frequencies are in this regime. The substitution of noble metals with TMOs, or with the other materials mentioned in this entry, will recover plasmonic behavior at mid-infrared and THz frequencies. One signature of a plasmonic behavior is the previously mentioned energy confinement quantifiable through the decay length L_z in the dielectric medium and the corresponding $\mathrm{FOM}_{\mathrm{conf}}$. The $\mathrm{FOM}_{\mathrm{conf}}$ as calculated for ITO is represented in Fig. 3. For instance, at $\lambda = 10$ μm in the mid-IR ITO shows a confinement of about 10 against a value of about 1 (practically no confinements) in silver.

As an example of the plasmonic behavior of ITO, one reports the result of electromagnetic simulations performed on a plasmonic structure which support propagating SP along the patterned surface (Fig. 5a). Here, one has a periodic array of subwavelength square apertures on ITO film on Si



Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 5 (**a**) Image of a 2D metallic grating for coupling infrared radiation to SP modes. (**b**) Simulated electric field profile decaying in the z-direction (*inset*) for ITO grating (*red*) and silver (*black*). A cross section of the structure (**a**) with a simulated field distribution in the *inset*

T

where the periodicity is symmetric in both in-plane directions. The electromagnetic simulation of the electric field profile along the z-direction (perpendicular to the surface) demonstrates that for ITO, mid-IR light can be squeezed in volumes of about 10 nm^3 (see Fig. 5b). For comparison, the same structure was simulated after substituting a noble metal (Ag) in place of ITO. The field profile now exhibits a decay length of around 100 nm, which is almost an order of magnitude larger. This comparison is consistent with the relative values of FOM_{conf} for the two materials represented in Fig. 3.

This difference in electromagnetic confinement clearly depends on the reduced plasma frequency for ITO with respect to silver (or other noble metals) and represents the most important result for using unconventional materials like TMOs and highly doped semiconductors in place of standard metals for mid-IR applications.

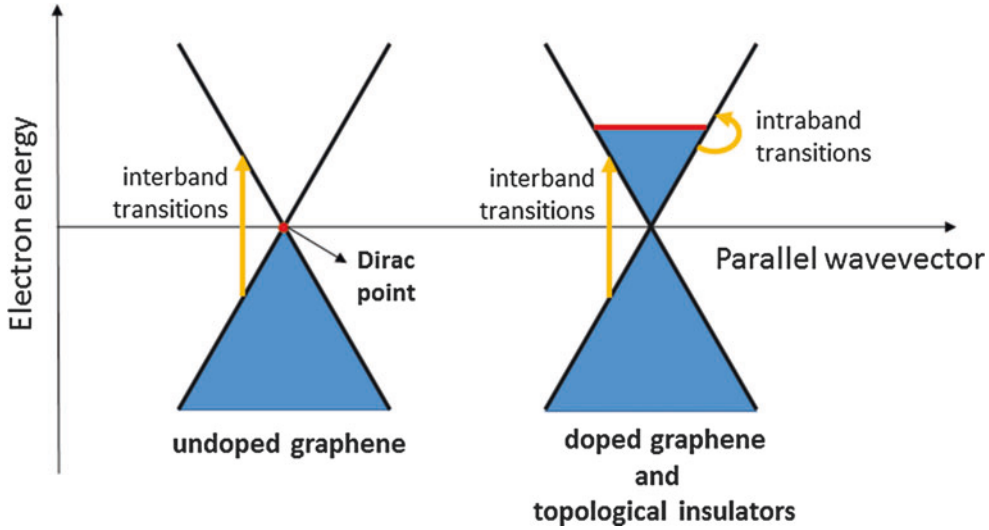
However, in the case of Ag, the decay length becomes comparable with the nanostructure size and is probably effected by it. Geometry-induced confinement is indeed an interesting way to enhance electromagnetic energy in small volumes (even with noble metals) as exploited in plasmonic rod antennas separated by nanogaps [12]. Nevertheless, intrinsic (material dependent) nanoconfinement like in ITO (and in graphene and topological insulators, see the next paragraph) permits to circumvent extremely demanding fabrication and could be useful for micro and nanofluidics [13].

Other TMOs are subject of current research, especially as shortage of indium demands a material alternative. Research is ongoing to substitute indium-containing ITO with compounds constituted of more common elements such as Al ZnO and Ge ZnO (AZO, GZO) [11].

THz Plasmonics on Topological Insulators and Graphene

Currently, THz plasmonic is based on conventional metals such as gold and silver [14, 15]. However, as pointed out in the introduction, those metals, due to the high carrier concentrations ($n \sim 10^{22} - 10^{23}/cm^3$), show a very limited tunability of their optical properties. Moreover, as discussed in the previous paragraph, electromagnetic confinements for metals having a high-plasma frequency are strongly reduced in mid-IR and THz regions. In the past, therefore, researches have been focalized on the metallic and optical properties of 2D electron gases (2DEG), whose plasmonic response can be more easily adapted to the THz range and, further, to be made tunable through the application of external stimuli. In particular, the investigations of plasmonic properties of 2DEGs have received a strong impulse in recent years after the discovery of graphene [16] and topological insulators [17].

Graphene is a 2D structure of carbon in which the atoms are located in a honeycomb lattice [18] and shows very specific mechanical, electric, magnetic, and thermal properties having a multitude of exciting applications in fundamental research and industry. Topological insulators (TIs) instead are quantum materials which show exotic/peculiar electronic properties: an insulating gap in the bulk, due to a strong spin-orbit coupling, and metallic surface states at the interface with the vacuum or another dielectric [19]. Despite the enormous difference in the physical properties of these two innovative materials, both graphene and TIs share a common characteristic, i.e., their electric transport properties are based on a Dirac fermion 2DEG, i.e., a gas composed of electrons having a linear dispersion frequency/wavevector $v(k) = \frac{1}{2\pi} v_F k$, where $v_F \sim 10^7 cm/s$ is the Fermi velocity. The high surface concentration of charge carriers reached in both materials ($n_{2DEG} \sim 10^{11} - 10^{14}/cm^2$), which provides a plasma frequency which varies from sub-THz to tenths of THz, and the very strong matter-light interaction, which arises from the small effective electron mass of Dirac carriers with respect to conventional 2DEG in semiconductors [20], have been shown to be very useful for several plasmonic applications [20]. In particular, the optical and plasmonic properties of graphene and topological insulators can be easily modulated through temperature, electric field, magnetic field, and optical pulses, opening the road for tunable devices working in the THz range [20, 21].



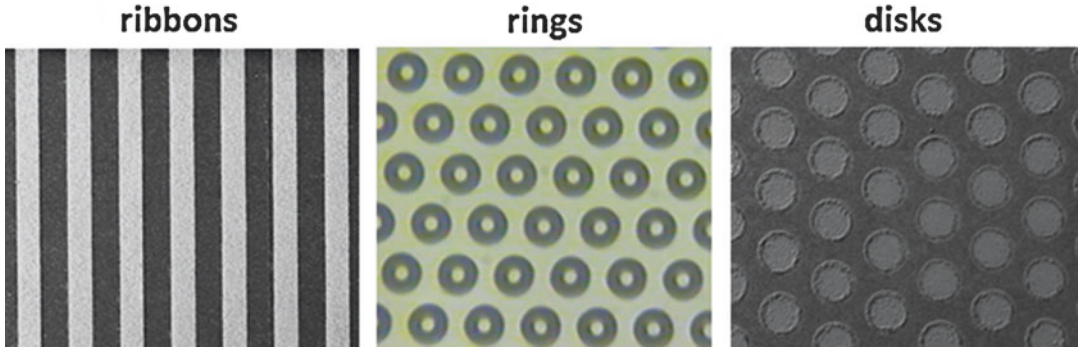
Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 6 The Dirac cone for the undoped graphene (*left*) and for doped graphene and topological insulators (*right*). The interband and intraband electronic transitions are indicated by *yellow arrows*

The electronic band structure of graphene near E_F can be described in terms of two inequivalent cones which touch at the so-called Dirac points. At zero doping, the Fermi energy coincides with the Dirac points and the pure graphene is a zero-gap semiconductor (see Fig. 6). Chemical and/or electrostatic gating injects charge carriers in the cones moving away the Fermi level from the Dirac points and generating a metal. At variance with graphene, TIs are intrinsically doped Dirac systems, which show two Dirac cones, one for each metallic surface, and a Fermi energy around 0.5 eV. Then, this means that TIs naturally support plasmon excitations. In graphene, instead, in order to have plasmons, some carriers should be injected in the Dirac cones. Plasmonics in TI, for instance, can be competitive with respect to standard metals as shown in Fig. 3 by the FOM_{conf} for Bi_2Se_3.

As specified in Fig. 1, SPs cannot be directly excited by electromagnetic radiation because their dispersion law is such as to prevent the conservation of wavevector in the photon absorption process. The necessary extra wavevector can be obtained through a patterning of the material surface with a variety of subwavelength structures. In Fig. 7, one shows an example of these structures fabricated on thin films of Bi_2Se_3 topological insulators. From left to right, one has parallel ribbons, disks, and rings with an average size in the micrometric range. The frequency/wavevector dispersion of microribbon arrays having a transversal size of W and period of 2 W has been calculated for a 2DEG of Dirac carriers [22, 23] and results in the low-k limit:

$$v_{\mathrm{Dirac}}(k) = \sqrt{k}\left[\frac{e^2}{4\pi\varepsilon_0\varepsilon\hbar}\sqrt{2\pi n_{2\mathrm{DEG}} g_s g_v}\right]^{1/2} \quad (4)$$

where $k = \pi/W$ is the wavevector, e the electron charge, the dielectric permittivity $\varepsilon = \varepsilon_1 + \varepsilon_2/2$ takes into account the dielectric environment by averaging the dielectric permittivities of materials between which graphene (topological insulator) is placed, n_{2DEG} is the surface carrier concentration, and g_s and g_v the valley and spin degeneracies, respectively. In particular, $g_s = g_v = 2$ in graphene and 1 in TI. Equation 4 shows the characteristic $n_{2\mathrm{DEG}}^{1/4}$ dependence of the Dirac plasmon energy with the carrier density which is qualitatively different with respect to the $n_{2DEG}^{1/2}$ law appearing in conventional semiconductors [22, 23]. The plasmonic absorption in TI parallel



Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 7 From *left* to *right*: ribbon, disk, and ring microstructures have been fabricated on Bi_2Se_3 topological insulator films. These structures provide the necessary extra wavevector for plasmonic absorption of THz radiation

ribbon arrays has been measured in Ref. [21] and is shown in Fig. 8a for ribbon size of $W = 2, 4, 8,$ and 20 μm. In this particular patterning geometry, the plasmon wavevector is $K = 2\pi/W$. In Fig. 8a, the hardening of the plasmon frequency for decreasing W (geometrical tuning) can be easily appreciated. In Fig. 8b, one shows the corresponding frequency/wavevector dispersion. Here, the experimental data (Fig. 8a) as measured by THz spectroscopy (open circles) are compared to Eq. 4 (as calculated for $n_{2DEG} = 3 \times 10^{13}\,cm^{-2}$ and $\varepsilon = 11.5$, red dashed line). The agreement between data and calculation is very good, showing that plasmon excitations in Dirac 2DEG can be well described in terms of classical density-density charge fluctuations [22, 23].

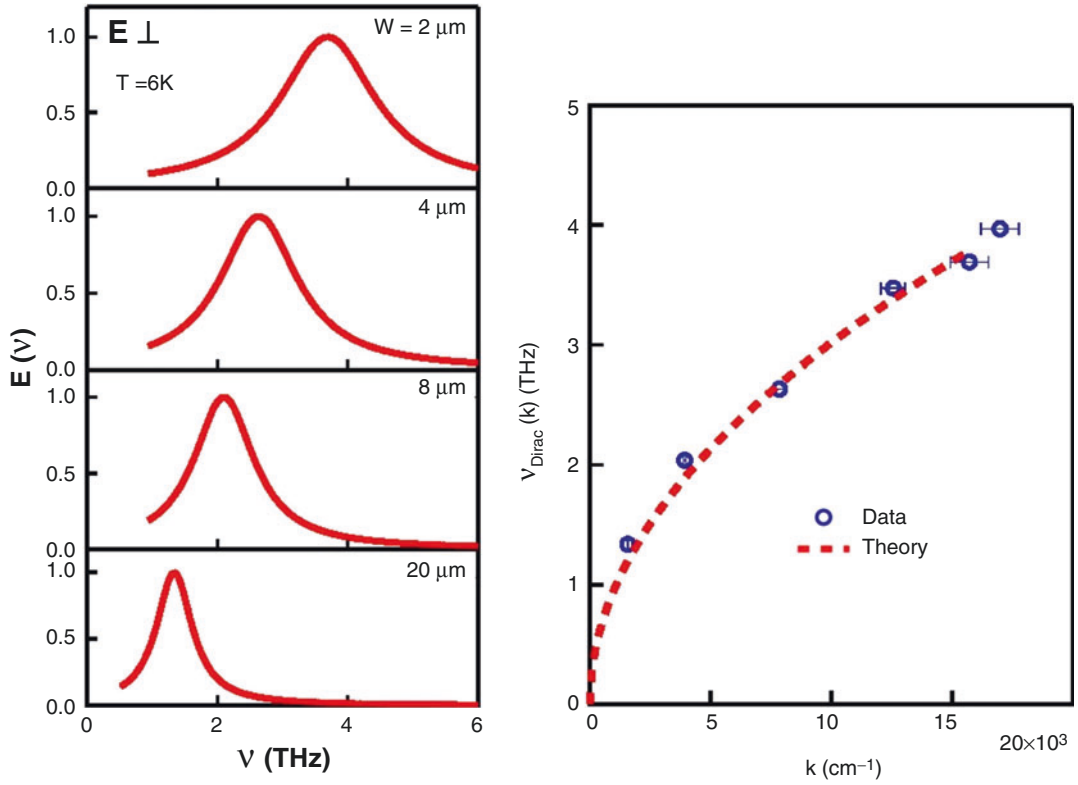
A ribbon array of graphene is represented in Fig. 9a for $W = 1, 2,$ and 4 μm [24]. In Fig. 9b, the plasmon absorption frequency clearly hardens for decreasing W like observed in TI (see Fig. 8).

The tunability of Dirac plasmons can be achieved through an electrostatic control of the n_{2DEG} density [20] and/or a photodoping generated through femtosecond laser pulses. An example of the electrostatic doping case is reported in Fig. 10. Here, one shows the plasmon absorption in a graphene disk with a diameter of 50 nm versus the Fermi energy which is tuned by an electrostatic voltage. In Fig. 10a, one shows the scheme of the devices used in Ref. [25], consisting of an ITO-coated silica substrate (4), showing on an array of graphene nanodisks (3), which is covered with ion gel (2). A bias potential is applied between the ITO and a gold contact (1) to electrically dope the disks through the ion gel. In Fig. 10b, one observes a shift to high frequency of the plasmon absorption in agreement with the dependence of the Dirac plasmon frequency on E_F: $v_{Dirac} \approx E_F^{1/2} \approx n_{2DEG}^{1/4}$ and its enhanced intensity due to the increase of the density of Dirac fermions involved in the optical transitions.

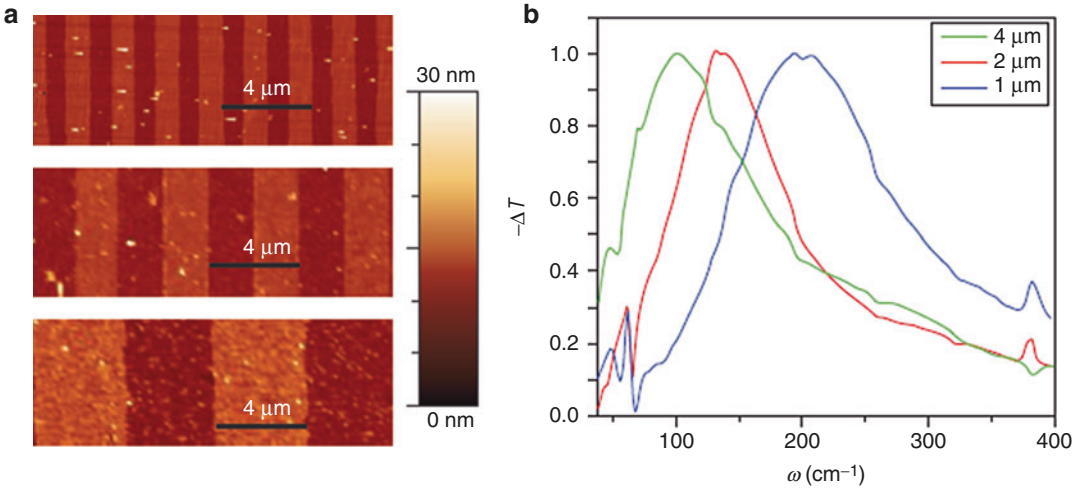
THz Plasmonics on Superconductors

As pointed out in the introduction, losses and scarce tunability of plasmonic devices are the major issues in THz plasmonics. The use of 2D electron gases, as discussed in the previous paragraph, is one of the possible key in developing tunable THz devices. Recently, superconductors emerge as another intriguing class of materials for THz plasmonic applications due to the fact that they bring two unique advantages, namely: low loss and intrinsic nonlinearity, tuning, and switching behavior [26, 27].

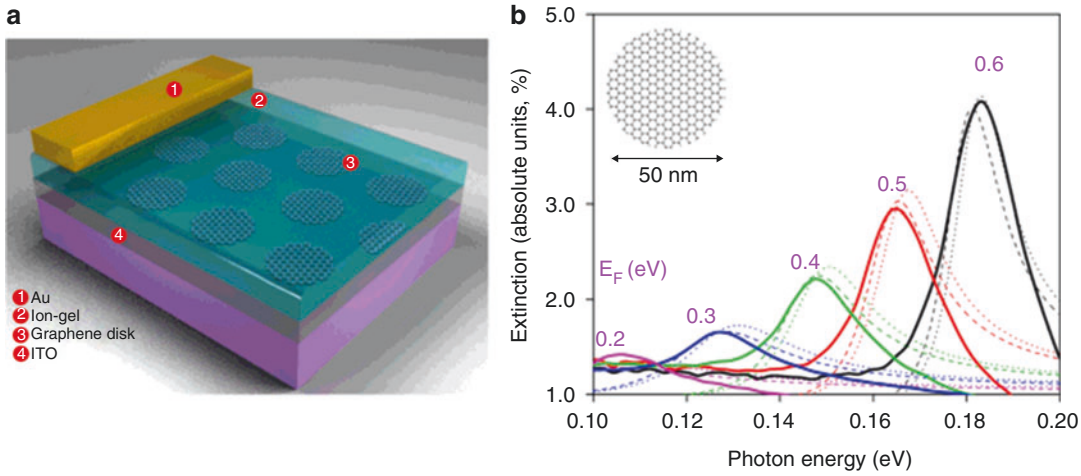
First, we briefly review the main aspects of superconductivity. Superconductivity is a macroscopic quantum state of matter arising from a many-body correlated electron state associated with pairing of electrons (Cooper pairs). The pairing energy is 2Δ where Δ is called the superconducting gap which opens at the Fermi energy in the single particle excitations. The



Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 8 (**a**) Plasmonic absorption spectra for different ribbon arrays of Bi_2Se_3 topological insulator samples fabricated through electron beam lithography [21] versus frequency and (**b**) the corresponding frequency/wavevector dispersion. The experimental data follow very well the low-k limit analytical law (Eq. 4) for Dirac plasmons

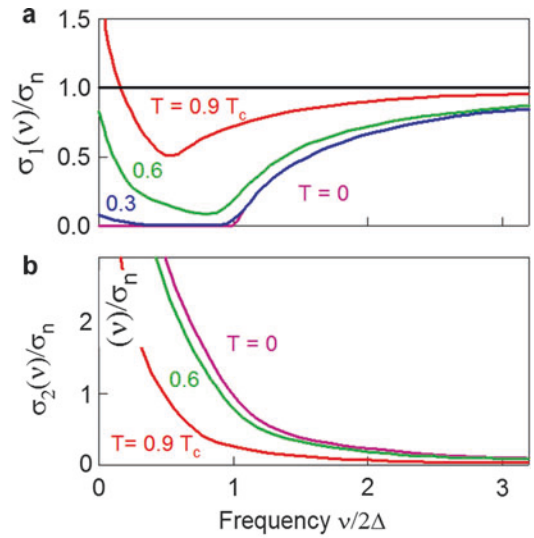


Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 9 (**a**) Three different ribbon arrays of graphene with a ribbon size W = 1, 2, and 4 μm and a period 2 W. In the panel (**b**) the THz plasmonic absorption hardens with decreasing W (Adapted from Ref. [24])

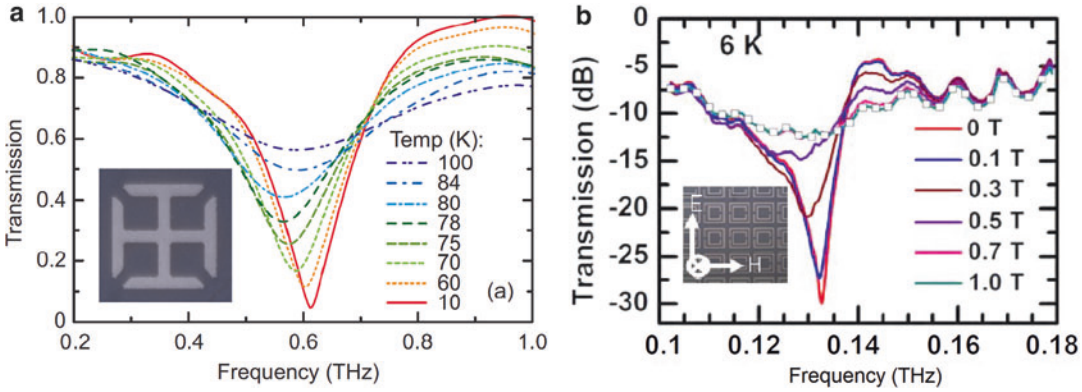


Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 10 Absorption from a graphene disk as a function of an electrostatic doping. An increase in the surface charge carriers n_{2DEG} corresponds to an enhanced Fermi energy ($v_D \approx E_F^{1/2} \approx n_{2DEG}^{1/4}$) and a shift toward high frequency of the plasmon excitation (Adapted from Ref. [25])

superconducting state establishes for $T \leq Tc$ where Tc is called the critical temperature, and, in many cases, it can be described in terms of a macroscopic coherent wavefunction having a well-defined magnitude and phase. The superconducting wavefunction plays the role of a complex order parameter in the Ginzburg-Landau treatment of the superconducting phase transition. The order parameter magnitude is related to the density of superconducting electrons n_s, while its phase is related to the supercurrent. This last quantity corresponds to the dissipation-free direct current $Jc(T)$ which a superconductor may sustain before reverting to the normal state (for $J > Jc$) at temperatures $T < Tc$. The electromagnetic properties of a superconductor can be described in terms of its complex optical conductivity $\widetilde{\sigma}(v) = \sigma_1(v) + i\sigma_2(v)$ (Fig. 11), which is related to the dielectric permittivity through the simple equation: $\widetilde{\varepsilon}(v) = 1 - \frac{4\pi}{v}\widetilde{\sigma}(v)$ (for the response of a standard metal, see Fig. 2). Through $\widetilde{\sigma}(v)$, one can calculate the current J induced by the application of an electric field E as $J = \widetilde{\sigma}(v)E$. At sufficiently low temperatures



Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 11 (**a**) The real σ_1 and the (**b**) imaginary σ_2 part of the optical conductivity in the superconducting state at various temperatures versus frequency. σ_1 and σ_2 are normalized to their values in the normal (metallic) state. The frequency axis is normalized with respect to the superconducting gap Δ value. When T decreases below Tc, a gap develops in σ_1, while σ_2 shows an inductive response with a $1/v$ frequency dependence

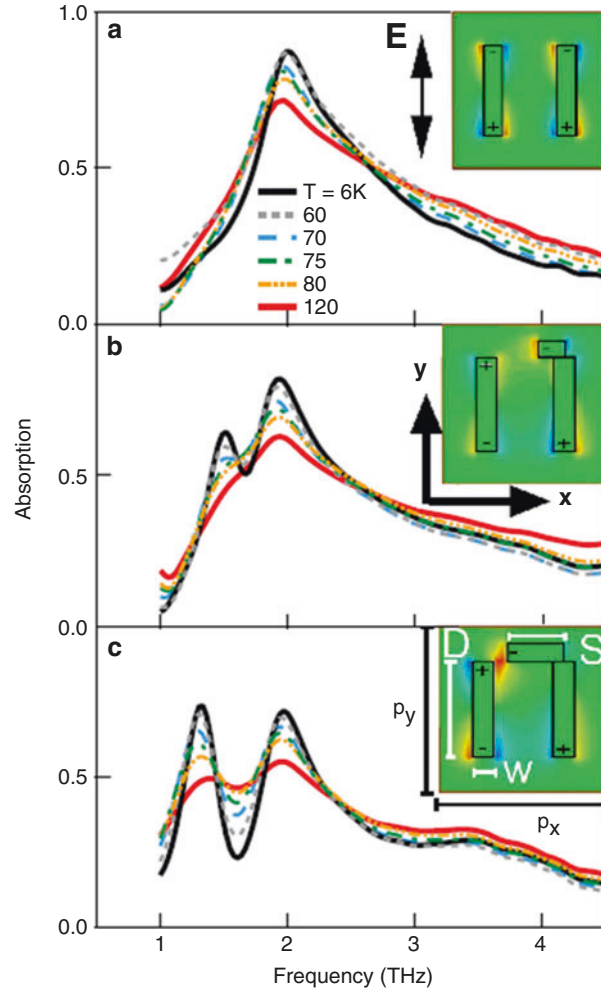


Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 12 (**a**) A plasmonic resonance is hardly visible around 0.6 THz in the transmittance of HTCS YBCO cuprate in its normal state (at 100 K). When T is reduced with respect to the critical temperature $Tc = 90$ K, the minimum in the transmittance starts to deepen due to the strongly temperature-dependent electromagnetic response in the superconducting state. In the inset is represented the plasmonic structure (Adapted from Ref. [28]). (**b**) In superconducting Nb ($T = 6$ K), the plasmonic resonance at 130 GHz can be modulated through a magnetic field. For fields above the critical field $Hc = 0.7$ T, the resonance is practically destroyed (Adapted from Ref. [29])

(T << Tc) and frequencies, one actually has $\sigma_2(v)/\sigma_1(v) >> 1$ (see Fig. 11), and then superconductors show an electrodynamic response that is mainly inductive in character ($\sigma_2 > 0$). Moreover, superconductors are lossy (i.e., they have a nonzero σ_1) at finite frequency and temperature due to the existence of nonsuperconducting electrons, whose abundance (and then losses) systematically increases with temperature. The frequency limitation of superconducting electromagnetic response is set by the gap frequency 2Δ (T)/h, where h is Planck's constant [26]. It can be shown [3] that temperature (and other external stimuli like magnetic field, electric field, and optical pulses) strongly affects the electromagnetic response of a superconductor for frequencies up to $\sim 4 - 6\Delta(T)/h$, while, for higher frequencies, the electrodynamic properties approach those of the material in the normal state (see Fig. 11).

In most of low-Tc superconductors, the frequency corresponding to $4 - 6\Delta(T)/h$ is located in the microwave and sub-THz regions [3], so tunable plasmonic applications are only restricted in this spectral region. On this ground, high critical temperature cuprate superconductors (HTCS) are of particular interest for two reasons: (i) they exhibit a *Tc* up to 140 K, hence well above liquid nitrogen temperature (77 K), which makes them very convenient for technological applications, and (ii) they provide much larger superconducting gaps which extends the tunability of plasmonic devices beyond 5 THz [26].

In Fig. 12, one reports on the transmittance at THz frequencies of an array of superconducting plasmonic structures. In the left panel, the structures have been achieved through standard photolithographic procedure on an epitaxial film of HTCS $YBa_2Cu_3O_7$ (YBCO) cuprate (deposited on sapphire), having a critical temperature $Tc = 90$ K [28] and are represented in the inset of Fig. 12a. In the main figure. one reports the transmittance at various *T* below and above *Tc*. For decreasing *T* below *Tc*, the plasmonic resonance around 0.6 THz, which is present also in the normal state (at 100 K), starts to deepen due to the strong inductive electromagnetic response of the superconducting state. This shows the tunability of such a plasmonic device by controlling the temperature. In the right panel, a similar resonance has been measured at lower frequency (around 130 GHz) and $T = 6$ K in a patterned Nb film on Si, having a $Tc = 9$ K [29]. Here, the plasmonic absorption has been measured as a function of a magnetic field (from 0 to 1 T). In particular, the plasmonic resonance is practically canceled out for fields higher than the critical field Hc (which is



Terahertz and Infrared Plasmonics with Unconventional Materials, Fig. 13 A transparency window is present in the THz response of a plasmonic dolmen structure fabricated on YBCO films. (**a**) A symmetric structure presents an absorption which is scarcely modulated by the superconducting transition. (**b**, **c**) When the EIT effect is induced by the geometrical broken symmetry along the x-axis, superconducting transition strongly affects the ETI response generating a modulation in the transparency of about 50 %

around 0.7 T in this case). Indeed, at 1 T the transmittance nearly coincides with that of 26 K (green empty squares), i.e., that in the normal state of Nb.

Another exciting application of the electromagnetic response of superconductors concerns the possibility to modulate the THz transparency through the superconducting transition [26, 30]. In general, electromagnetically induced transparency (EIT) is a quantum interference effect that permits light propagation in an otherwise opaque medium. The plasmonic analogue of EIT has been recently demonstrated in many systems [30, and references therein] from THz to the visible. In these examples, EIT is generated through the interaction between a plasmonic bright superradiant mode (of dipolar nature) with a plasmonic dark subradiant mode (quadrupolar or a multipolar mode). The dark mode can also be either a molecular vibration [30] or a phonon excitation [21]. In most cases, the EIT transparency window is set by the geometrical parameters of the plasmonic structure or material properties, with no possibility for active control of the optical response. This control can be obtained by fabricating the plasmonic EIT structure on HTCS [30]. In Fig. 13, indeed, we show a plasmonic dolmen structure which can sustain both a dipolar (superradiant excitation) and a quadrupolar

plasmonic mode (subradiant excitation) and where the interaction between the modes can be geometrically tuned. In Fig. 13a (completely symmetric structure), the absorption, for light polarization along the y-axis, shows a broad maximum related to a plasmonic dipolar excitation. As the x-axis symmetry of the dipolar structure is broken by the addition of the horizontal arm in samples b and c, a transparency window opens at 1.8 THz close to the peak absorption of the dipolar band. These minima are strongly modulated through the superconducting transition, exhibiting a modulation of up to 50 % activated by superconductivity.

Conclusions

Mid-IR and THz plasmonics may strongly benefit from use of unconventional metallic materials which provide a stronger confinement of light in small volumes and a major tunability of plasmonic resonances with respect to standard metals. In this entry, we have described some plasmonic applications of unconventional systems like transparent metallic oxides, graphene, topological insulators, and superconductors. These materials start to be used routinely in plasmonics and their utilization has opened the road for a further investigation of electromagnetic properties beyond those of standard metals.

Cross-References

- ▶ Plasmonic Waveguides
- ▶ Plasmonics
- ▶ Nano-optics, Nanoplasmonics

References

1. Pitarke, J.M., et al.: Theory of surface plasmons and surface-plasmon polaritons. Rep. Prog. Phys. **70**, 1–87 (2007)
2. Maier, S.A.: Plasmonics: Fundamentals and Applications. Springer, New York (2007)
3. Pattnaik, P.: Surface plasmon resonance: applications in understanding receptor-ligand interaction. Appl. Biochem. Biotechnol. **126**, 79–92 (2005)
4. Herminjard, S., Sirigu, L., Herzig, H.P., Studemann, E., Crottini, A., Pellaux, J.P., Gresch, T., Fischer, M., Faist, J.: Surface Plasmon Resonance sensor showing enhanced sensitivity for CO2 detection in the mid-infrared range. Opt. Express **17**, 293 (2009)
5. Limaj, O., Lupi, S., Mattioli, F., Leoni, R., Ortolani, M.: Midinfrared surface plasmon sensor based on a substrateless metal mesh. Appl. Phys. Lett. **98**, 091902 (2011)
6. Adato, R., Yanik, A.A., Amsden, J.J., Kaplan, D.L., Omenetto, F.G., Hong, M.K., Erramilli, S., Altug, H.: Ultra-sensitive vibrational spectroscopy of protein monolayers with plasmonic nanoantenna arrays. Proc. Natl. Acad. Sci. U. S. A. **106**, 19227 (2010)
7. Yanik, A.A., Huang, M., Kamohara, O., Artar, A., Geisbert, T.W., Connor, J.H., Altug, H.: An optofluidic nanoplasmonic biosensor for direct detection of live viruses from biological media. Nano Lett. **10**, 4962 (2010)
8. Dressel, M., Gruner, G.: Electrodynamics of Solids. Cambridge University Press, Cambridge (2002)
9. Dionne, J.A., Sweatlock, L.A., Atwater, H.A., Polman, A.: Planar metal plasmon waveguides: frequency-dependent dispersion, propagation, localization, and loss beyond the free electron model. Phys. Rev. B **72**, 075405 (2005)
10. Minami, T.: Present status of transparent conducting oxide thin-film development for Indium-Tin-Oxide (ITO) substitutes. Thin Solid Films **516**, 5822–5828 (2008)
11. Kim, J., Naik, G.V., Emani, N.K., Boltasseva, A.: Plasmonic resonances in nanostructured transparent conducting oxide films. IEEE J. Sel. Top. Quantum Electron. **19**, 3 (2013)
12. Huck, C., Neubrech, F., Vogt, J., Toma, A., Gerbert, D., Katzmann, J., Hartling, T., Pucci, A.: Surface-enhanced infrared spectroscopy using nanometer-sized gaps. ACS Nano **8**, 4908 (2014)
13. Eftekhari, F., Escobedo, C., Ferreira, J., Duan, X., Girotto, E.M., Brolo, A.G., Gordon, R., Sinton, D.: nanoholes as nanochannels: flow-through plasmonic sensing. Anal. Chem. **81**, 4308–4311 (2009)
14. D'Apuzzo, F., Candeloro, P., Domenici, F., Autore, M., Di Pietro, P., Perucchi, A., Roy, P., Sennato, S., Bordi, F., Di Fabrizio, E.M., Lupi Plasmonics, S.: Resonating Terahertz Response of Periodic Arrays of Subwavelength Apertures, Plasmonics (2014). DOI: 10.1007/s11468-014-9775-3
15. Razzari, L., Toma, A., Shalaby, M., Clerici, M., Proietti Zaccaria, R., Liberale, C., Marras, S., Ibraheem, A., Al-Naib, I., Das, G., De Angelis, F., Peccianti, M., Falqui, A., Ozaki, T., Morandotti, R., Di Fabrizio, E.: Extremely large extinction efficiency

and field enhancement in terahertz resonant dipole nanoantennas. Opt. Express **19**, 27 (2011)
16. Novoselov, K.S., Geim, A.K., Morozov, S.V., Jiang, D., Katsnelson, M.I., Grigorieva, I.V., Dubonos, S.V., Firsov, A.A.: Two-dimensional gas of massless Dirac fermions in graphene. Nature **438**, 197–200 (2005)
17. Hasan, M.Z., Kane, C.L.: Colloquium: Topological insulators. Rev. Mod. Phys. **82**, 3045 (2010)
18. Grigorenko, A.N., Polini, M., Novoselov, K.S.: Graphene plasmonics. Nat. Photonics **6**, 749–758 (2012)
19. Moore, J.E.: The birth of topological insulators. Nature **464**, 194–198 (2010)
20. Javier Garcìa de Abajo, F.: Graphene plasmonics: challenges and opportunities. ACS Photonics **1**(3), 135–152 (2014)
21. Di Pietro, P., Ortolani, M., Limaj, O., Di Gaspare, A., Giliberti, V., Giorgianni, F., Brahlek, M., Bansal, N., Koirala, N., Oh, S., Calvani, P., Lupi, S.: Observation of Dirac plasmons in a topological insulator. Nat. Nanotechnol. **8**, 556–560 (2013)
22. Polini, M., Asgari, R., Borghi, G., Barlas, Y., Pereg-Barnea, T., MacDonald, A.H.: Plasmons and the spectral function of graphene. Phys. Rev. B **77**, 081411(R) (2008)
23. Das Sarma, S., Hwang, E.H.: Collective modes of the massless Dirac plasma. Phys. Rev. Lett. **102**, 206412 (2009)
24. Ju, L., Geng, B., Horng, J., Girit, C., Martin, M., Hao, Z., Bechtel, H.A., Liang, X., Zettl, A., Ron Shen, Y., Wang, F.: Graphene plasmonics for tunable terahertz metamaterials. Nat. Nanotechnol. **6**, 630–634 (2011)
25. Fang, Z., Thongrattanasiri, S., Schlather, A., Liu, Z., Ma, L., Wang, Y., Ajayan, P.M., Nordlander, P., Halas, N.J., Javier Garcıa de Abajo, F.: Gated tunability and hybridization of localized plasmons in nanostructured graphene. ACS Nano **3**, 2388–2395 (2013)
26. Anlage, S.M.: The physics and applications of superconducting metamaterials. J. Opt. **13**, 0240001–0240010 (2011)
27. Tsiatmas, A., Fedotov, V.A., Javier Garcıa de Abajo, F., Zheludev, N.I.: Low-loss terahertz superconducting plasmonics. New J. Phys. **14**, 115006–115015 (2012)
28. Chen, H.-T., Yang, H., Singh, R., O'Hara, J.F., Azad, A.K., Trugman, S.A., Jia, Q.X., Taylor, A.J.: Tuning the resonance in high-temperature superconducting terahertz metamaterials. Phys. Rev. Lett. **105**, 247402 (2010)
29. Wu, J., Biaobing, J., Caihong, Z., Engelbrecht, S., Pimenov, A., Xu, Q., Cao, C., Chen, J., Xu, W., Kang, L., Wu, P.: Tuning of superconducting niobium nitride terahertz metamaterials. Opt. Express **19**, 12021–12026 (2011)
30. Limaj, O., Giorgianni, F., Di Gaspare, A., Giliberti, V., de Marzi, G., Roy, P., Ortolani, M., Lupi, S.: Superconductivity-induced transparency in terahertz metamaterials. ACS Photonics **1**, 570–575 (2014)

Terahertz Technology for Nano Applications

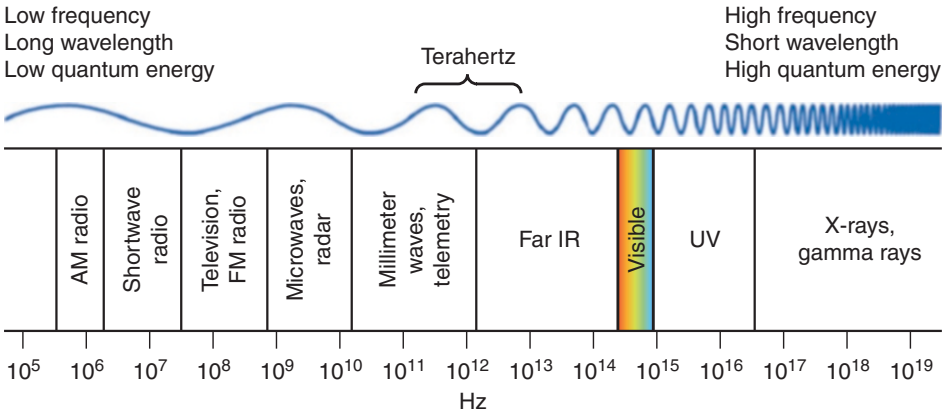
Nezih Pala and Ahmad Nabil Abbas
Department of Electrical and Computer Engineering, Florida International University, Miami, FL, USA

Synonyms

Terahertz; THz; T-rays

Definition

The terahertz (THz) region of the electromagnetic spectrum is generally defined as the frequency range of 0.1–10 THz (10^{12} cycles per second) corresponding to quantum energy of 0.4 meV–0.4 eV (see Fig. 1). THz electromagnetic waves (also known as T-rays) have several properties that could promote their use as sensing and imaging tool. There is no ionization hazard for biological tissue and Rayleigh scattering of electromagnetic radiation is many orders of magnitude less for THz wavelengths than for the neighboring infrared and optical regions of the spectrum. THz radiation can also penetrate nonmetallic materials such as fabric, leather, and plastic which makes it useful in security screening for concealed weapons. The THz frequencies correspond to energy levels of molecular rotations and vibrations of DNA and proteins, as well as explosives, and these may provide characteristic fingerprints to differentiate biological tissues in a region of the spectrum not previously explored for medical use or detect and identify trace amount of explosives. THz wavelengths are particularly sensitive to water and exhibit absorption peaks which makes the technique very sensitive to hydration state and can indicate tissue condition. THz radiation has also been used in the characterization of semiconductor materials, and in testing and failure analysis of VLSI circuits. THz techniques also allowed art historians to see murals hidden



Terahertz Technology for Nano Applications, Fig. 1 A schematic showing the THz region within the electromagnetic spectrum. Although there is no strict definition of what qualify as THz waves, it is commonly considered the region between 100 GHz and 10 THz

beneath coats of plaster or paint in centuries-old building, without harming the artwork.

The lack of efficient sources and detectors in THz frequency range compared to the relatively well-developed lower-frequency RF/microwave and higher-frequency infrared (IR)/far infrared (FIR) ranges was referred as "THz gap" in the scientific community about a decade ago. However, the unique position of THz frequencies in the electromagnetic spectrum has allowed an intense multidisciplinary approach in the development of THz emission and detection techniques as well as their applications, closing the "THz gap" from both sides. Especially rapidly evolving nanosciences and nanotechnology immensely contribute to the advancement of THz technology. These advancements lead to creation of many unique applications some of which in turn contributes to nanotechnology. Therefore, the interaction between THz and nanotechnologies is reciprocal.

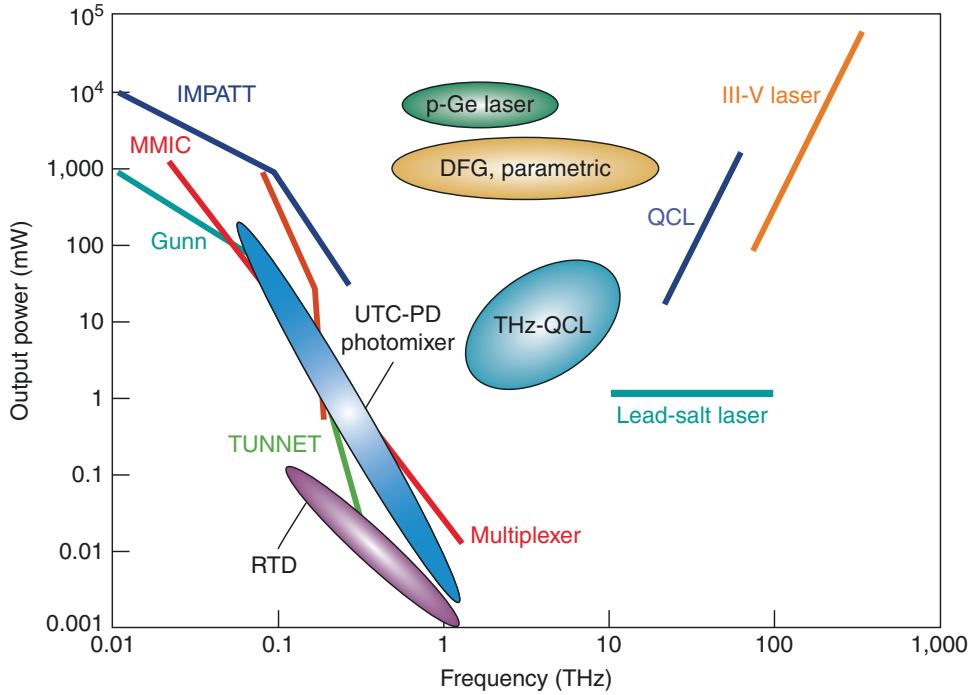
Principles of THz Technology

Terahertz Sources

Taking advantage the above-mentioned unique spectral position THz frequencies, four major approaches have been adopted to develop THz sources. The first approach is pushing the operation frequencies of the existing microwave and millimeter-wave devices into the THz range. The second one is the free-electron-based sources like free-electron lasers and backward wave oscillators (BWOs). The third is the utilization of optical methods which has been the major technique for the demonstration of many THz applications. And the fourth one is the newly developed Quantum Cascade Lasers (QCLs). THz emission power of selected sources are shown in Fig. 2 as a function of frequency. Despite the tremendous research and development efforts, currently available THz source technologies, particularly the tunable emitters, such as free-electron lasers, backward wave oscillators, gas lasers, heterodyne photomixers, optical parametric converters are usually large, complex, require high power, and hence costly. Semiconductor electronic devices such as Gunn oscillators or Schottky diodes need frequency multipliers which severely impact the output power. Quantum cascade lasers which are the true compact solid-state THz emitters are still in their infancy and operate only at cryogenic temperatures.

Solid-State Electronic THz Sources

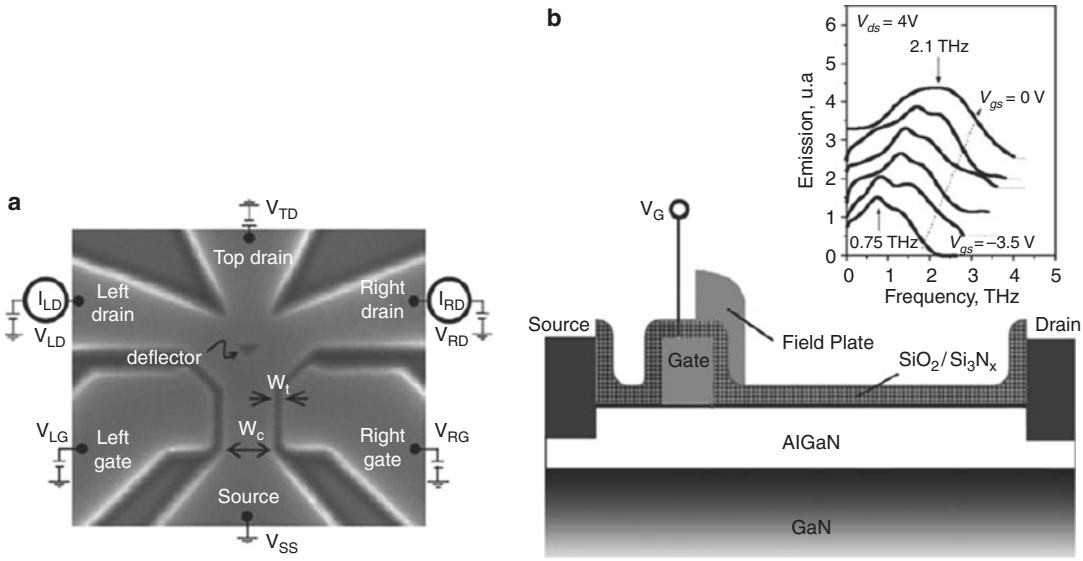
Solid-state electronic emitters are limited in frequency due to the transient time of carriers which results in high-frequency roll-off. This intrinsic limitation prevented the development of a single



Terahertz Technology for Nano Applications, Fig. 2 THz-emission power as a function of frequency. Solid lines are for the conventional THz sources; IMPATT diode stands for impact ionization avalanche transit-time diode, MMIC stands for microwave monolithic integrated circuit, TUNNET stands for tunnel injection transit time, and the multiplexer is an SBD frequency multiplier. Ovals denote recent THz sources. The values of the last two are indicated by peak power; others are by c.w. power (After Ref. [1])

electronic device which could oscillate in the entire THz bandwidth of 0.1–10 THz. Gunn diodes (also known as transferred electron devices), IMPATT (Impact Avalanche Transit Time) diodes, and TUNNETT (Tunnel Injection Transit Time) diodes are well-developed high-frequency sources. They take advantage of negative differential resistance (NDR) in current voltage characteristics. When a DC bias large enough to drive such a device into NDR is applied, unstable state triggers oscillations with frequencies up into the mm-wave range. Specific mechanism resulting the NDR determines the upper limit of the oscillations frequency. Current Gunn diodes and IMPATT diodes can reach oscillating frequencies of 400–500 GHz. InP Gunn devices with graded doping profiles generated RF output powers of 283 μW at 412 GHz, 203 μW at 429 GHz and the highest third-harmonic frequency of 455 GHz with an output power of 23 μW. GaAs TUNNETs, on the other hand, produced 10 mW at 202 GHz. Resonant tunneling diodes (RTDs) are semiconductor devices with double barrier structures which forms a quantum well in between. Electrons in this quantum well are at quantized discrete energy levels. At a particular DC bias, the energy levels at the bottom of the conduction band on one side align with the lowest energy level in the well, resulting in a resonant condition in which electrons can tunnel through the barriers. As the bias increases further, the resonant condition disappears and the current falls. This NDR process combined with the inherently fast tunneling mechanism allows RTDs working at high operation frequencies. RTDs based on InGaAs/AlAs material system have reached the fundamental oscillation frequency of 915 GHz with, however, quite low output level of a few tens of nanowatts.

Recently, in parallel to the advancements in material- and device-processing technologies,



Terahertz Technology for Nano Applications, Fig. 3 (**a**) SEM Image of top view of ballistic deflection transistors (After Ref. [2]). (**b**) Schematics of the GaN/AlGaN plasmonic HEMT with gate covered by a field plate reported in Ref. [3]. The inset shows the emission spectra of the device

transistors with very high cutoff frequencies (f_T) have also been reported. Northrop Grumman announced an InP high electron mobility transistor (HEMT) with extrapolated $f_T > 1$ THz. This device had sub-50 nm gate length, 25 % increase in electron mobility, and reduced contact resistance. Similarly, heterojunction bipolar transistors (HBTs) based on InP materials systems achieved a cutoff frequency f_T of 765 GHz at room temperature and 855 GHz at 218 K.

As a newly emerged alternative approach, ballistic deflection transistors based on InGaAs-InAlAs heterostructure on an InP substrate were developed at University of Rochester with a theoretical maximum f_T of approximately 1.02 THz (Fig. 3a). Another novel approach is the plasma wave instabilities in the FET channels with high sheet carrier concentrations. Such instabilities were predicted by Dyakonov and Shur in 1991. Recently, a group of researchers from Japan and Europe reported room temperature generation of radiation at 0.75 and 2.1 THz from AlGaN/GaN HEMTs (Fig. 3b).

Due to the limited frequency performance of solid-state emitters, they are often used in combination with multiplication circuits to reach THz frequencies. A multiplier consists of a nonlinear electronic device, such as a Schottky varactor diode placed between an input and an output-matching network. Unfortunately, the output power is much lower than that of the input, which is a serious drawback for the THz frequency range. An input with a power of 200–300 mW at 100 GHz can be produced by HEMT amplifiers, but a multiplier with a high-order of multiplication from 100 GHz up to 1–3 THz is not feasible due to the very high losses. Much lower losses are achievable only in multipliers with a low-order of multiplication, i.e., doublers (2×) and triplers (3×), so that a THz multiplier could consist of a sequence of doublers and triplers of the frequency up to the desired THz frequency.

Free-Electron-Based Sources

Free-electron lasers have quite different working principles than more traditional optical lasers such as gas lasers. They can generate either CW or pulsed high-power THz radiation, but they are very costly and have very large dimensions, functioning in large rooms containing many additional facilities which limits their availability to only

limited number of facilities around the world. However, backward wave oscillators (BWO) (also called carcinotron) are based on the same principles as an electron laser and are able to deliver a few milliwatts in a tunable range of 0.3–1.3 THz with a high sweeping rate. Although BWO requires a water-cooling system due to its high bias voltages of 1–6 kV their size and weight is much smaller (∼15 kg) which makes them convenient and affordable for many research labs [4].

Recently, there has been an intensifying research effort to develop micromachined Vacuum Electronic Devices, or "micro-VEDs (μVEDs)." Vacuum technology offers a feasible solution for efficient THz transmitters but requires a significant invention to overcome the problem of complex and difficult size scaling necessary to achieve THz operation which calls for advanced fabrication methods and stringent tolerances for interaction structures. Critical developments include the fabrication of the interaction circuit, a high current density integratable cathode structure, and achieving stable high-power electron beam transport. Fortunately, today's microfabrication methods such as Deep Reactive Ion Etching (DRIE) and LIGA are now capable of manufacturing slow-wave interaction structures at these frequencies with the required resolution and surface roughness characteristics. Further, significant progress has been made in the development of new cathode materials and structures.

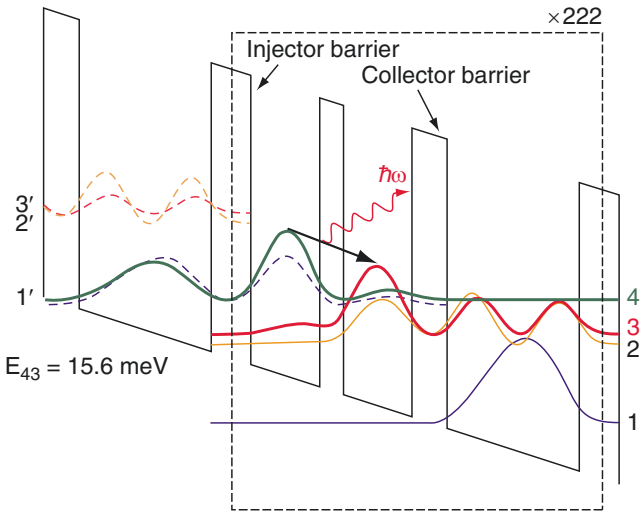
Folded waveguide (FWG) approach has emerged as promising approach to design micro-VED high-power amplifiers (HPAs) and THz-regime oscillators. FWG offers sufficient gain and bandwidth to achieve 60 mW at 670 GHz and 10 mW at 1 THz and 56 mW at 560 GHz as THz Oscillator. Another approach is the Extended Interaction Klystron (EIK) approach which applies the concept of multiple interaction structures at offset frequencies to simultaneously achieve high gain from the superposition of multiple structures interacting with the signal and high bandwidth from the gradual offset design of the same structures. This concept is expected to allow design of an HPA that would deliver output power >20 dBm and gain >20 dB at 1.03 THz. The EIK approach requires machining optical quality surfaces over macroscopic areas because the skin depth in the metal at 1 THz is comparable to the previously achievable surface roughness and will result in significant ohmic losses. The issue is complicated by the severe thermal constraints placed on the interaction structure which limits material choices.

Optical THz Emission Techniques

Gas lasers are probably the most conventional CW THz sources in the frequency range of 0.9–3 THz and with output powers in the range of 1–30 mW. They are typically pumped by a carbon dioxide laser, and the output frequency is determined by the gas in the cavity (CH_4, N_2, etc.). Although gas lasers are not tunable, they have several discrete emission lines.

Photomixing, also known as optical heterodyne down-conversion, is a technique used to generate CW THz radiation by using ultrafast lasers. In its most commonly used version, photoconductive emitters biased at ∼5 kV/cm are excited by Ti-sapphire lasers, which typically deliver pulse lengths of 50–100 fs, and can be as short as 10 fs with the repetition rate of the order of 50–100 MHz and with the average power in the range of 0.2–2 W. Low-temperature grown (LT), GaAs is usually the material of choice due to its photocarrier lifetime (<0.5 ps) bandgap of 1.42 eV, allowing absorption of the 800 nm radiation from Ti-sapphire lasers. When a laser pulse hits on a semiconductor, the absorbed photons generate photocarriers. These photocarriers are accelerated by the bias field, while simultaneously their density changes under the varying laser intensity. As a result, ultrashort high-peak currents are generated in the semiconductor, which radiate into free space at broad range of THz frequencies.

Another use of the ultrafast lasers for THz generation is in optical rectification. Optical rectification is the result of the transient polarization which occurs when a short, high-intensity laser pulse interacts with the electrooptic medium. The THz power generated is proportional to the square of the optical power, is determined by the second-order optical nonlinear coefficient ($\chi^{(2)}$), and varies with the relative orientation of the laser



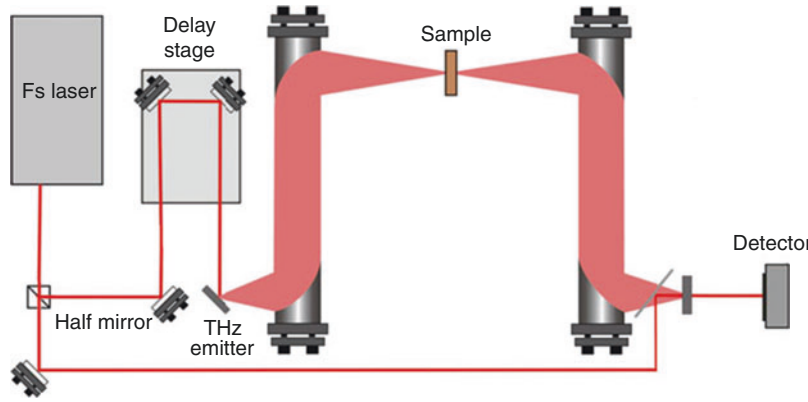
Terahertz Technology for Nano Applications, Fig. 4 Conduction band diagram of a diagonal design quantum cascade laser with the lasing power of approximately 5 mW at 180 K. 1′ is the injector level from the preceding module. The figure shows that the upper- and lower-state wave functions are localized in separate wells with little spatial overlap and the radiative transition is from 4 → 3 ($f = E_{43}/h \cong 3.8$ THz). Therefore, this scheme is called diagonal design (Adapted from Ref. [6])

polarization and the crystallographic axes. Typical THz bandwidths obtained from optical rectification in EO crystals are 0.1–3 THz, with a total average THz power of a few milliwatts. However, much broader THz emission has been achieved. In both ZnTe and GaSe bandwidths of up to 40 THz have been demonstrated.

Difference frequency generation (DFG) is another second-order nonlinear optical process which is used to produce THz electromagnetic waves by mixing two optical beams in a nonlinear crystal, such that the output frequency is the difference between the two input frequencies. DFG method allowed the demonstration of THz peak powers of the order of hundreds of milliwatts in the frequency range, including frequencies above 3 THz, with the ability to produce a tunable narrow-line single frequency. Tuning over 0.5–7 THz has been obtained in a GaP crystal, with peak THz powers of around 100 mW. In an organic DAST crystal, continuous tuning between 2 and 20 THz has been achieved, with peak powers of >10 W.

Quantum Cascade Lasers

One of the most exciting approaches to generate tunable THz radiation is the quantum cascade lasers (QCLs). QCL idea was proposed in 1971 as a FIR radiation source and experimentally demonstrated in 1994. In a QCL, the light produced by one carrier transition between two levels is amplified due to photon-assisted tunneling of a single type of carriers in a sequence of coupled quantum wells (superlattice) that has a staircase-like band energy. Therefore, it is a unipolar laser where the carriers can be either electrons or holes. The number of amplification stages determines the output power. The discreteness of energy levels, named sub-bands, inside the same band is a result of the spatial confinement of carriers inside the heterostructure, and the radiation frequency is determined by the energy difference of sub-bands between which radiative/lasing transitions occur. The first QCL working in the THz range was reported in 2002 [5]. This laser delivers about 2 mW power at 4.4 THz and operates at 50 K. The output power decreases dramatically with increasing temperature and becomes nearly zero at room temperature. The realization of a QCL at THz frequencies encounters a series of difficulties and limitations due to the very large values of the wavelength. Among them are very large free-carrier absorption losses and the necessity of growing a very thick heterostructure. In 2009, a resonant-phonon terahertz QCL operating up to a record braking heat-sink temperature of 186 K was demonstrated (see Fig. 4). At the lasing frequency of 3.9 THz, 63 mW of peak optical power was measured at 5 K, and approximately 5 mW could still be detected at 180 K. All these



Terahertz Technology for Nano Applications, Fig. 5 Simplified schematic of a typical THz-TDS system

THz cascade lasers are based on n-type carriers (electrons), and the photon emission is parallel to the heterostructure plane (edge-emission).

Terahertz Detection Systems

Present imaging techniques in THz applications are divided between Pulsed Time Domain (PTD) and Continuous Wave (CW) modalities. THz time-domain spectroscopy (THz-TDS) has been the primary PTD method for demonstration of the potential application in THz technology. A typical THz-TDS system, as shown in Fig. 5, includes an ultrafast laser (e.g., Ti-Sapphire laser) which produces optical pulses in fs duration. Each pulse is separated into two optical paths. One path travel through a time delay stage and hits the emitter such as a photoconductive antenna or a nonlinear crystal in which the optical pulses are converted into ultrashort electromagnetic pulses. The generated EM pulses are focused onto the sample under test and collimated again to be focused onto the detector. The other part of the pulse is also delivered onto the detector directly. Amplitude of the electromagnetic waves are measured by the detector to calculate the change in the waveform due to the sample. By changing delay time between the two beams, it is possible to scan the THz pulse and construct its electric field as a function of time. Subsequently, a Fourier transform is used to extract the frequency spectrum from the time-domain data. Comparison of the waveforms with and without the sample allows the estimation of the complex refractive index of the material, which results other parameters, such as the dielectric constant, conductivity and surface impedance. Although they have been widely used in research facilities, their size and complexity prevent widespread deployment for out-of-lab applications.

On the other hand, CW imaging systems can be passive, where the terahertz or sub-terahertz energy analyzed is present in the environment or emitted by the object itself (i.e., the human body), or active, where the energy is supplied by a dedicated source. In active systems, a major concern is the presently only limited available terahertz sources. This limitation places significant demands on detector capabilities. For CW systems, two primary modalities exist for detection of signals in the terahertz range. One involves the reduction of the terahertz signal to sufficiently low frequencies to allow amplification and signal processing. This (heterodyne receiver) approach typically uses nonlinear diodes to mix the terahertz signal with a local oscillator reference. Heterodyne detection imposes many restrictions on the system design and makes it complex and expensive. These restrictions may be difficult or impossible to fulfill. The second approach is direct detection, where the terahertz signal impinging upon a detecting device results in a measurable response. Available direct detection technologies with key parameters are summarized in Table 1.

As it is shown on Table 1, today the most sensitive THz detectors are Golay cells, pyroelectric detectors, bolometers, and Schottky diodes. However, they are not portable, not tunable or,

Terahertz Technology for Nano Applications, Table 1 Comparison of THz direct detection technologies

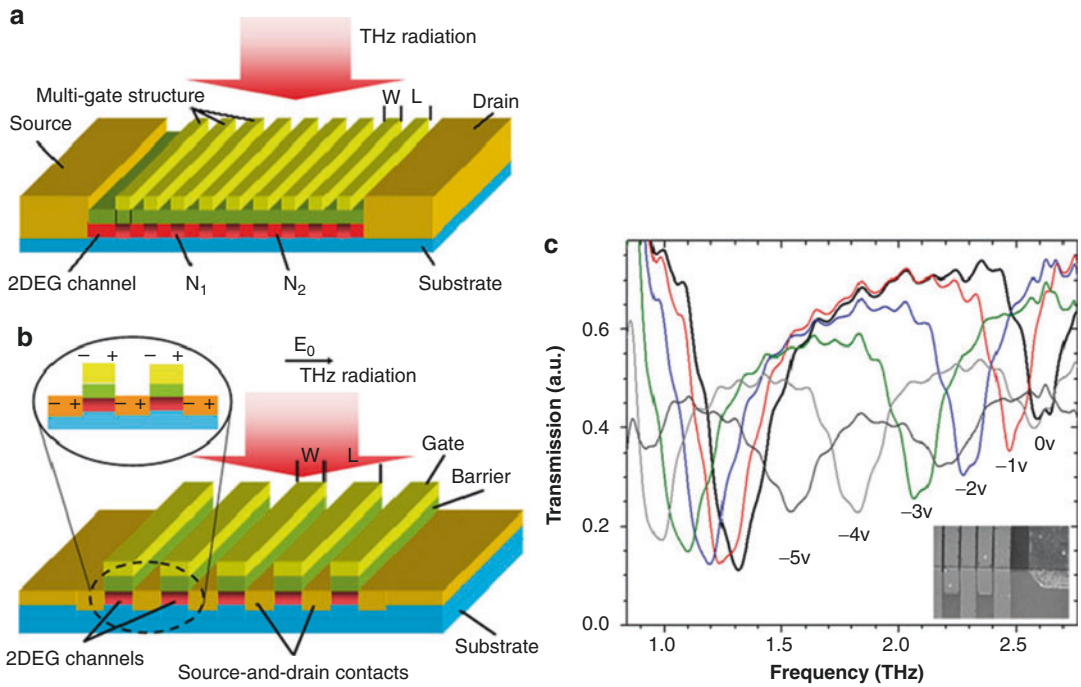
Detector type	NEP (W/Hz$^{1/2}$)	Responsivity (V/W)	Resp. time (s)	Op. tmp. (K)	Size	Tunable
Bolometers	10^{-16} to 10^{-13}	10^5–10^7	10^{-3} to 10^{-2}	≤4.2	Large	No
Hot-e microbolometers	10^{-19} to 10^{-17}	10^9	10^{-8}	≤0.3	Compact	No
STJ detectors	10^{-16}	10^9	10^{-3}	≤0.8	Compact	No
Golay cells	10^{-10}	10^5	10^{-2}	300	Large	No
Pyroelectric detectors	10^{-10}	10^5	10^{-2}	240–350	Medium	No
Schottky diodes	10^{-12}	10^3	10^{-12}	10–420	Compact	No
Plasmonic HEMTs	10^{-10a}	10^{3a}	10^{-10a}	10–420	Compact	Yes

[a]Expected

most of the time, are very slow. Schottky diodes, for instance, can reach very low noise equivalent power (NEP) and high responsivity at millimeter waves. However, their responsivity drops orders of magnitude (to $<10^3$ V/W) at frequencies greater than 1 THz and they are not tunable.

Dyakonov and Shur proposed a novel idea of using plasmonic resonances in the two-dimensional electron gas (2DEG) for tunable emission and detection of terahertz radiation, which are being explored and proven experimentally, culminating in the recent demonstration of resonant detection of the THz radiation at room temperature. The theory [7, 8] predicts that such detectors should have very high sensitivities (10^7 V/W at 77 K and 10^4 V/W at 300 K). The plasmonic detectors are capable for operating even at zero bias current, thus minimizing shot noise and allowing unprecedented sensitivity. GaN-based detectors with noise equivalent power (NEP) smaller than 10^{-8} W/Hz$^{1/2}$ have been demonstrated. This value is slightly higher than for such commercial detectors as Golay cell, pyroelectric detectors, and Schottky diodes, with the potential advantage of operation at very high sampling frequencies. Indeed, theoretical analysis of the detection of modulated sub-terahertz and terahertz radiation by a short channel plasmonic detectors predicts a very high upper limit for modulation frequency up to 100 GHz or even higher. These detectors can detect not only the intensity but also polarization and direction of a terahertz beam. Most intriguing advantage of all, the plasmonic terahertz detectors are tunable by DC bias over a wide frequency range.

Plasmonic detection of THz radiation is based on the excitation of plasma oscillations in the two-dimensional electron gas (2DEG) channels of FET structures by the incoming radiation. Such FETs support two different types of plasma oscillations. The first is the plasma oscillations in the gated region of the 2DEG channel. If an infinite perfectly conductive plane is located at distance d from the infinite 2D electron sheet and the conducting plane is close enough to 2D electron sheet (i.e., $kd \ll 1$), then the dispersion relation is given by $\omega_p = k\sqrt{(e^2Nd = m^*\varepsilon_0\varepsilon_1)}$, where ω_p and k are the frequency and wave vector of plasma wave, respectively, N is the sheet electron density, e and m^* are the charge and effective mass of electron, ε_0 is the dielectric permittivity, and ε_1 is the dielectric constant of the insulator separating the 2D electron layer from the perfectly conductive plane. The second type of plasma oscillations exist in ungated regions. For negligible electron scattering in 2D electron gas, the dispersion relation for ungated plasma oscillations in an infinite homogenous 2D electron sheet is $\omega_p = \sqrt{k}\sqrt{(e^2Nd = m^*\varepsilon_0\overline{\varepsilon})}$ where $\overline{\varepsilon}$ is the effective dielectric function which depends on the geometry of the structure. In both cases, the resonant frequency ω_p depends on the sheet carrier concentration N which can be easily controlled by an applied gate bias V_g. Encouraging results have been reported for THz detection by plasmonic nanoscale single-gate FETs. However, response of a single FET plasma detector is limited by its small area compared to the beam cross section. Also, gated plasmons in a single-gate FET are



Terahertz Technology for Nano Applications, Fig. 6 Schematics of the plasmonic devices with (**a**) multiple gates (grating gate) on a common channel, (**b**) multichannel FETs having an array of separate channels with combined intrinsic source and drain contacts, (**c**) transmission spectra of a AlGaN/GaN single-channel multi-gate plasmonic device at different gate voltages. The data shows the possibility of tunable THz detection. The inset shows the close-up SEM image of the gate elements of the measured device (Adapted from Ref. [9])

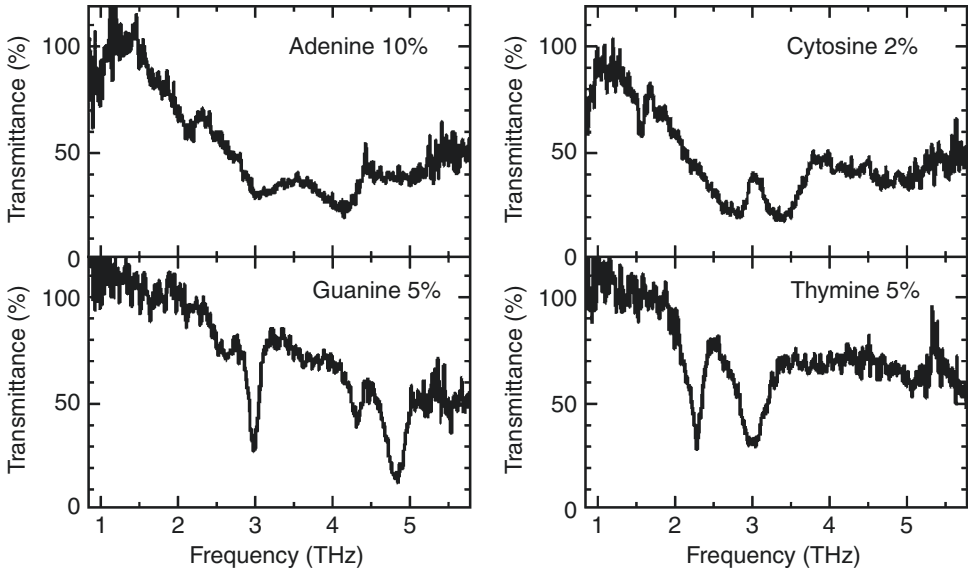
weakly coupled to terahertz radiation because of the strong screening of the gate plasmons by the metal gate electrode and their vanishingly small net dipole moment due to their acoustic nature. Coupling efficiency can be greatly enhanced by utilizing multiple gates (grating gate) on a common channel (Fig. 6a) or multichannel FETs having an array of separate channels with combined intrinsic source and drain contacts (Fig. 6b). Resonant coupling of THz radiation into the plasmons in 2DEG of a single-channel multi-gate plasmonic device have been demonstrated which could be used for precise tuning by the applied gate voltage (Fig. 6c).

THz Applications

Distinctive position of THz frequencies in the electromagnetic spectrum with their lower quantum energy compared to IR and higher frequency compared to MW range calls for many potential applications unique to them. The former characteristic allows detection and identification of substances based on their spectral signatures due to molecular vibration and rotation, while the latter one allows wide bandwidth for wireless communication. Indeed, spectroscopic analysis, sensing, and imaging represent the major application category of today's THz technology. Near-field imaging for subwavelength-scale small features is a notably fast-advancing field in this category. Communication, although still at infancy, is also promising application for THz frequencies.

Spectroscopic Sensing and Imaging

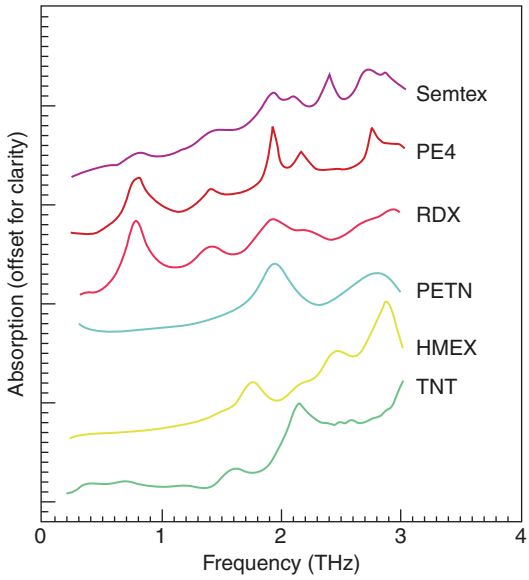
The unique advantage of THz waves for many applications can be attributed to their low frequency (or low quantum energy) compared to the commonly used parts of the optical spectrum. The low energy of terahertz photons can perhaps be better appreciated when compared to average



Terahertz Technology for Nano Applications, Fig. 7 Transmission spectra of the nucleobases (adenine, cytosine, guanine, thymine) at THz frequencies identify their spectral signatures (After Ref. [12])

thermal energy, 1 THz corresponds to just 4 meV, or about one sixth of the energy in one degree of freedom at 300 K. Therefore, terahertz radiation is useful for studying very low-frequency phenomena such as vibrations and rotations in molecules and soft lattice vibrations in dielectrics. Since materials, such as polymers, clothes, plastic, wood, and soil are nearly transparent to terahertz rays, while water, window-glass, and most metals are opaque; many novel imaging applications are available with T-rays. To date, THz waves have been used to read letters in closed envelopes, to detect hidden explosives and weaponry, pharmaceutical quality control, and to detect tumors in breast and epidermal tissue.

THz frequencies are particularly important for biosensing applications since numerous characteristic vibrational modes of macromolecules like proteins and DNA are located in this spectral range. This correspondence opens up the possibility of label-free biosensing. Although millimeter wave spectroscopy has been known since the 1950s, the cost was prohibitive and studies did not extend to large biological proteins such as DNA. In 1995, Hu and Nuss demonstrated the first THz imaging system based on THz-TDS which has become the method of choice for demonstration of many sensing and imaging applications [10]. Shortly after, Woolard et al. proposed to use millimeter-THz spectrocopy for detecting DNA mutagenesis. In this first attempt, they successfully demonstrated the lesion-induced vibrational modes in DNA observed from 80 to 1,000 GHz. The observed modes were associated with localized defects of the DNA polymers [11]. Low-frequency collective vibrational modes of lyophilized powder samples of calf thymus DNA, bovine serum albumin (BSA), and collagen in the 0.06–2.00 THz frequency range have been studied, showing that a large number of the low-frequency collective modes for these systems are active in this range. Further studies demonstrated the feasibility of THz-TDS to (a) identify biomolecular species, (b) identify the conformational state, and (c) identify the mutation of biomolecules. The understanding of conformational change and flexibility is particularly critical for bioengineering of biomolecules for drug discovery and therapeutic purposes. Transmission spectra of the nucleobases (adenine, cytosine, guanine, thymine) have also been measured to identify the spectral signatures at THz frequencies (see Fig. 7).

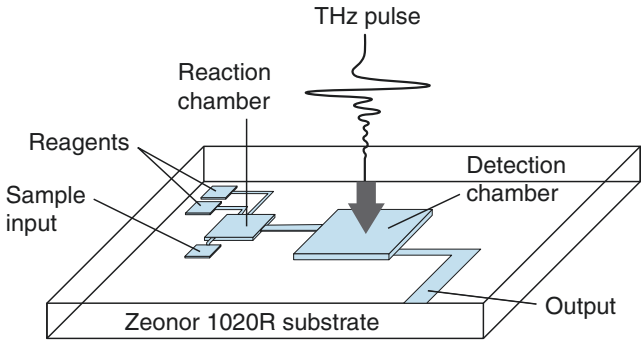


Terahertz Technology for Nano Applications, Fig. 8 Terahertz transmission spectra of the raw explosive materials TNT, HMX, PETN, and RDX together with the spectra of the compound explosives PE4 and Semtex H [13]

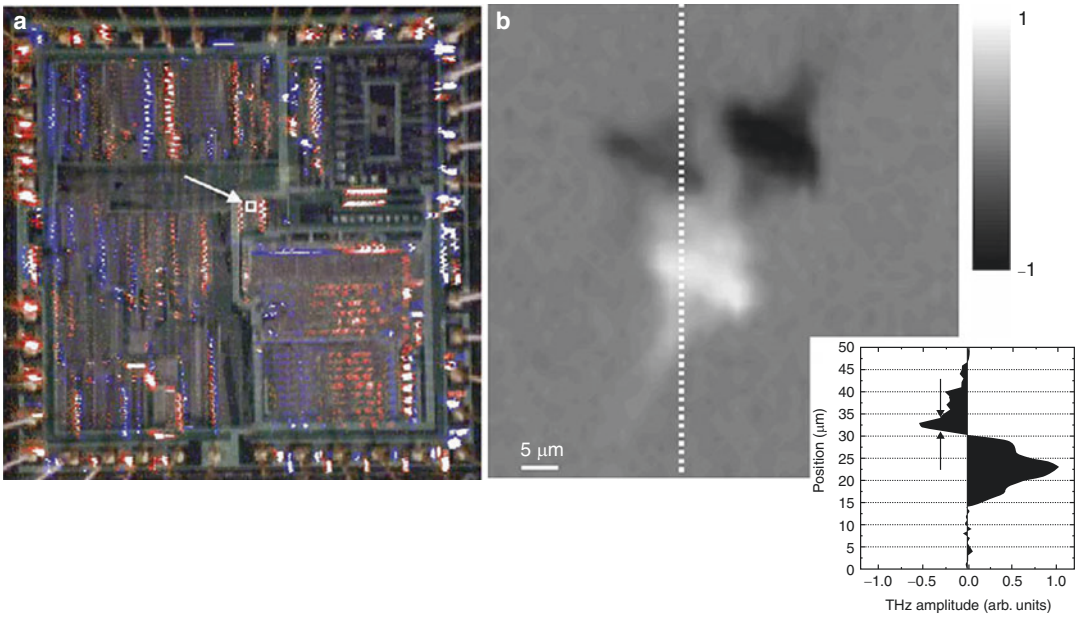
In the security side of the THz sensing applications, absorption spectra of explosives and related compounds (ERCs) by using THz-TDS has also been measured. The obtained absorption spectra in the range of 0.1–3 THz show that most of the ERCs have THz fingerprints which are caused by both the intramolecular and intermolecular vibrational modes of these materials (Fig. 8). Since commonly used nonpolar dielectric materials (e.g., fabric, leather) are almost transparent in the THz range, THz fingerprints makes THz technology a competitive technique for detecting hidden explosives. In most of these studies, the samples under investigation were either simply deposited on a THz-transparent (e.g., polyethylene) substrate, placed in a cell consisting of the same or, in rare cases, formed into durable free standing films.

By the time, more sophisticated micro-/nanoscale devices have been developed for advanced THz sensing applications. Nagel et al. developed functionalized planar waveguide sensors to detect hybridization of unknown DNA molecules to complementary immobilized probe molecules with a known base sequence. Thin-film microstrip (TFMS) line waveguides building a ring resonator have resonance frequencies depending on the effective permittivity of the TFMS line. In order to use the ring resonator as a sensor for DNA molecule detection, its metallic signal line surface is chemically functionalized with single-stranded probe molecules. Target molecules can hybridize to these capture probes consisting of complementary single-stranded DNA oligonucleotides. At frequencies f_r, where the resonance condition is fulfilled, the number of transmission events through the sample is approximately multiplied by the quality factor $Q = f_r/\Delta f_r$ of the resonator. The ring resonator amplifies the interaction between the THz signal and the probed sample in a controlled and efficient way. The complex effective permittivity of the TFMS lines and with it the characteristic parameters of the resonator, such as resonance frequency or quality factor, is altered in dependence on the dielectric properties and amount of sample material brought to the resonator. Such sensing systems allows DNA detection without complex and costly labeling (tagging) procedure. Researchers in Cornell University developed a microfluidic chip for terahertz spectroscopy of biomolecules in aqueous solutions. The chip is fabricated out of Zeonor 1020R, a plastic material that is both mechanically rigid and optically transparent with near-zero dispersion in the terahertz frequency range. Microfluidic channel devices could make possible the real-time THz spectroscopy of chemical reactions and allow the use of low-power THz-TDS systems by avoiding excess water absorption and to enable the spectroscopy of picomole quantities of biomolecules. Moreover, such devices can easily be integrated with photonic components to realize multifunctional spectroscopy platforms. A schematic of a THz sensing platform based on microfluidic channels is shown in Fig. 9.

One notable application which might be particularly important for nanotechnology is the THz imaging of integrated circuits (ICs) for quality control and testing. Laser terahertz emission microscope (LTEM) for noncontact, nondestructive inspection of electrical failures in circuits has been proposed and demonstrated that it could



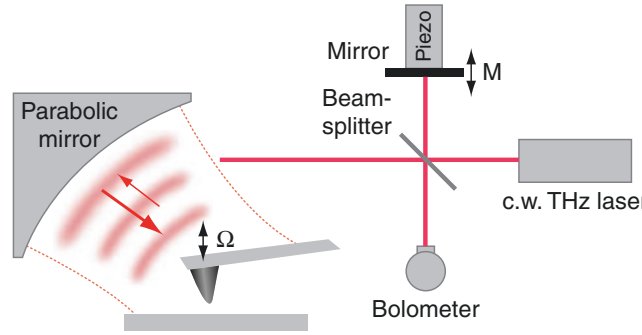
Terahertz Technology for Nano Applications, Fig. 9 A system for on-chip THz sensing. The device contains sample and reagent inputs, a reaction chamber, and a detection chamber for THz, as well as optical and IR spectroscopy (After Ref. [14])



Terahertz Technology for Nano Applications, Fig. 10 (**a**) The THz emission image an LSI 8-bit microprocessor chip. The red and blue regions indicate THz emissions with positive and negative amplitudes, respectively, which correspond to the direction of the electric field. (**b**) The magnified THz emission image of the area indicated by the white square in Fig. 10a. The inset shows the cross-sectional distribution of the THz emission image at the dotted line (After Ref. [15])

localize a broken line in an operational amplifier. LTEM is based on the measurement of the THz emission in free space from the ultrafast transient photocurrent in a circuit by scanning it with femtosecond (fs) laser pulses. The THz waves can be emitted from areas in the circuits such as photoconductive switch structures with external bias voltage, various unbiased interfaces with built-in electric fields – such as p-n junctions, Schottky contacts (metal/semiconductor interfaces) – or just the surface of semiconductor. The THz emission amplitude is proportional to the local electric field in the photo-excited area. By comparing the THz emission images of a normal and damaged chip, the abnormal electrical field resulting from an electrical failure in the circuit can be visualized (Fig. 10). However, the spatial resolution of LTEM was strongly affected by the long focal length of the focusing lens and the aperture size in the system, limiting it to about 20 μm.

T



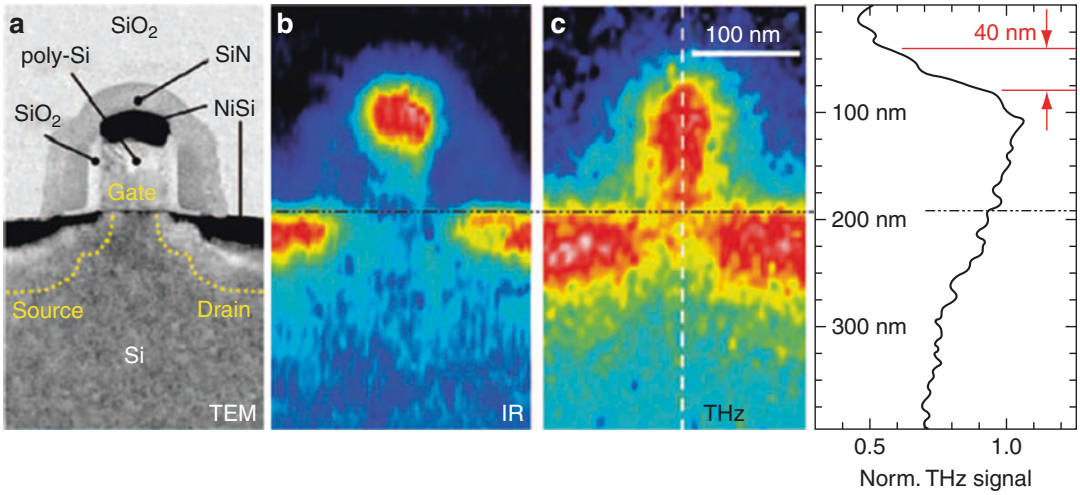
Terahertz Technology for Nano Applications, Fig. 11 Schematic of apertureless THz scanning near-field optical microscope (THz-SNOM) (After Ref. [17])

Therefore, noncontact testing of ICs using THz radiation required near-field imaging techniques which will be discussed in the next section.

THz Near-Field Imaging and THz Microscopy

Low-energy photons of THz radiation make a rich variety of light-matter interactions possible by exciting molecular vibrations and phonons as well as plasmons and electrons of nonmetallic conductors. Therefore, T-rays offer new ways for material and device characterization. However, diffraction limits the spatial resolution to the order of wavelength and prevents the use of THz waves for imaging beyond ~100 μm. Conventional imaging methods cannot be employed for THz mapping of micro- or nanoelectronic devices, low-dimensional semiconductor nanostructures, cellular entities, or single molecules. This problem can be circumvented by invoking the near-field scanning optical microscopy technique for THz imaging applications. One way to overcome the diffraction limit is to limit the detection area with an aperture. It has been shown that using a detector with an integrated aperture, near-field images can be made with a spatial resolution determined by the aperture size, and not by the THz wavelength. This method, however, suffered from microfabrication complexity and waveguide effects which lead to strong attenuation for long wavelength components. A widely explored alternative method is the use of sharp metal tips as local field enhancers. The sharp metal tip is held in close contact to the sample under study, while light is scattered by the tip. Acting as antennas, the wires capture incident THz waves and convert them into strongly confined near fields at the wire tip apex which has submicrometer dimensions. When this confined field becomes modified by a close-by scanned sample, the scattered radiation carries information on the local dielectric properties of the sample. This apertureless near-field imaging, which was originally developed at visible and mid-infrared wavelengths, has recently been demonstrated at terahertz wavelengths. THz images with a subwavelength resolution can be obtained by recording the scattered radiation by a distant THz receiver. A big advantage of this imaging technique is that it does not suffer from a waveguide cutoff, which limits the resolution in aperture techniques. Attempts of realizing such THz-scattering near-field optical microscopy (THz-SNOM), however, suffer from weak signals and faint material contrasts owing to strong background scattering.

In the first demonstration of THz near-field imaging at THz frequencies, Hunsche et al. used ultrashort, broadband pulses which were focused into a tapered metal tip with a small exit aperture and scanned a sample in the near field of this aperture. Their method allowed spatial resolution better than $\lambda/4$ for the corresponding average wavelength of 220 mm [16]. Huber et al. demonstrated high-resolution THz near-field microscopy using the experimental setup shown in Fig. 11 [17]. The THz signal was incident through a parabolic mirror with an angle to allow total internal reflection over the sample surface. On top of the sample, a cantilever of 20 μm length holding a metallic tip of 30 nm radius vibrated to modulate the incident wave in an On and Off manner. The cantilever



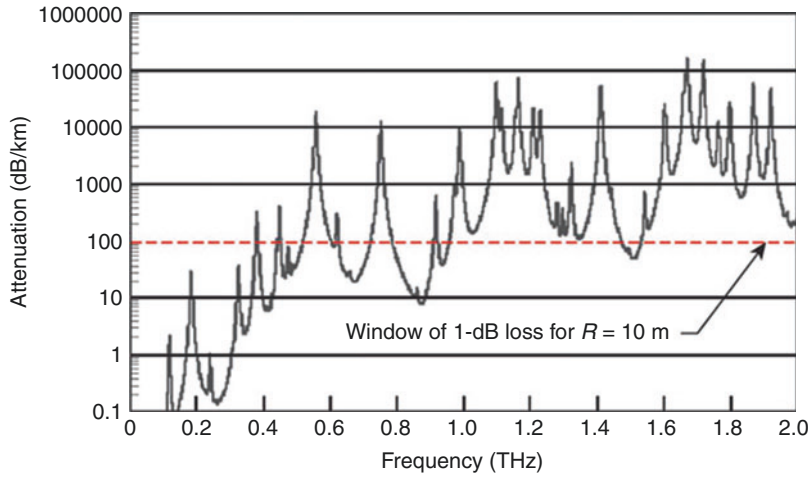
Terahertz Technology for Nano Applications, Fig. 12 (**a**) TEM image of a single transistor. The highly doped regions below the source and drain NiSi contacts are marked by dashed yellow lines. (**b**) Infrared image of the single transistor 12a ($\lambda \approx 11$ μm). (**c**) High-resolution THz image of the single showing all essential parts of the transistor: source, drain, and gate. The THz profile extracted along the dashed white line (averaged over a width of 12 nm and normalized to the signal obtained on the metallic NiSi gate contact) allows the estimation of a spatial resolution of about 40 nm, from the strong signal change at the SiO_2/SiN/NiSi transition [17]

mechanically oscillated at the mechanical resonance frequency of 35 kHz. A Michelson interferometer detector and hot electron bolometer were used to measure backscattered THz radiation. This technique allowed 30–150 nm spatial resolution imaging, corresponding to λ/4,000 for 2.54 THz. Moreover, near-field THz contrast between materials with different dielectric properties allowed imaging of a polished cut through nanoscale transistor structure and simultaneous recognition of materials and concentration of mobile carriers in the range of 10^{16}–10^{19} cm^{-3}. Since the spatial resolution of 40 nm infers that the volume probed by the THz near-field is about $(40\ nm)^3$, it can be concluded that an average of less than 100 electrons in the probed volume suffices to evoke significant THz contrast. This opens the possibility of THz studies of single electrons by improving the used setup and, in conjunction with ultrafast techniques, even their dynamics. The reported results are encouraging to use THz near-field microscopy for studying other charged particles and quasiparticles in condensed matter, for example, in superconductors, low-dimensional electron systems, or conducting biopolymers, which possess intrinsic excitations at THz quantum energies and thus should exhibit resonantly enhanced THz contrast (Fig. 12).

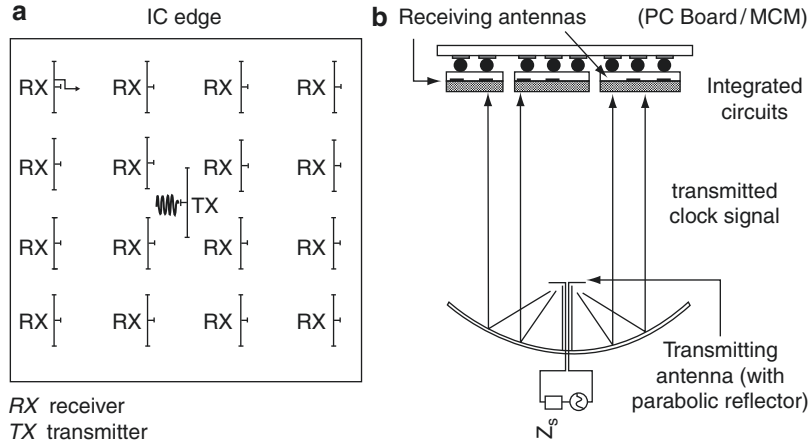
THz Communication

T-rays are very attractive for communication applications for the large bandwidth at these frequencies which could provide very high data transmission rates. Moreover, frequencies above 300 GHz are currently not allocated by the Federal Communications Commission. On the down side, strong atmospheric absorption due to the water vapor together with the low available power of the current THz sources limits the range of communication. A 10 mW source and 1 pW detection sensitivity would be required for 1 km communication range considering the typical attenuation rate of ∼100 dB/km rate in the atmospheric absorption windows (Fig. 13).

The strong atmospheric attenuation, however, does not pose a real threat for a number of communication applications. Space communication, for instance, can benefit from the high data transmission rates and smaller antenna size required for THz frequencies which would make possible the development of smaller, lighter therefore satellites with longer lifetime. Large bandwidth of THz



Terahertz Technology for Nano Applications, Fig. 13 Atmospheric attenuation in the THz frequency range [18]



Terahertz Technology for Nano Applications, Fig. 14 Conceptual system illustrations of (**a**) intra-chip and (**b**) inter-chip wireless interconnect systems for clock signal distribution (After Ref. [19])

frequencies can also be used to for short-range indoor communication exceeding Gigabit/s transmission rates. Rayleigh scattering of electromagnetic radiation is many orders of magnitude less for THz wavelengths than for the alternative infrared frequencies.

One unique and promising potential for THz frequencies is the on-chip and communication, especially systems-on-chip (SoC) applications. With the ever-increasing complexity of integrated circuits, larger and larger number of functional and storage cores are squeezed into a single die. Extrapolating from the current CMOS scaling trends, conventional on-chip interconnect technologies have been projected to reach their limits to meet the performance needs of the next-generation multi-core and SoC systems. Interconnect delays do not scale as well as local wires with respect to gate delays, and global interconnects have an increasing impact on the performance of overall SoCs. A radical alternative to the existing metal/dielectric interconnect infrastructures is to use transmission of signals via wireless interconnects. On-chip wireless interconnects were demonstrated first for clock signal distribution at 15 GHz carrier frequency [19]. If the THz frequencies are employed for on-chip communication, corresponding antenna sizes decrease occupying much less chip real estate. Nanowires and carbon nanotubes (CNTs) can serve as antenna structures in such applications. CNT bundles are predicted to enhance performance of

antenna modules by up to 40 dB in radiation efficiency and provide excellent directional properties in far field patterns and can achieve a bandwidth of 500 GHz and hence can achieve much higher data rates (Fig. 14).

Cross-References

- ▶ Biosensors
- ▶ Nanostructure Field Effect Transistor Biosensors

References

1. Tonouchi, M.: Cutting-edge terahertz technology. Nat. Photonics **1**, 97–105 (2007)
2. Diduck, Q., Irie, H., Margala, M.: A room temperature ballistic deflection transistor for high performance applications. Int. J. High Speed Electron. Syst. **19**, 23–31 (2009)
3. El Fatimy, A., Dyakonova, N., Meziani, Y., Otsuji, T., Knap, W., Vandenbrouk, S., Madjour, K., Théron, D., Gaquiere, C., Poisson, M.A., Delage, S., Prystawko, P., Skierbiszewski, C.: AlGaN/GaN high electron mobility transistors as a voltage-tunable room temperature terahertz sources. J. Appl. Phys. **107**, 024504 (2010)
4. Dragoman, D., Dragoman, M.: Terahertz fields and applications. Prog. Quantum Electron. **28**, 1–66 (2004)
5. Kohler, R., Tredicucci, A., Beltram, F., Beere, H.E., Linfield, E.H., Gilles Davies, A., Ritchie, D.A., Iotti, R.C., Rossi, F.: Terahertz semiconductor-heterostructure laser. Nature **417**, 156–159 (2002)
6. Kumar, S., Hu, Q., Reno, J.L.: 186 K operation of terahertz quantum-cascade lasers based on a diagonal design. Appl. Phys. Lett. **94**, 131105 (2009)
7. Dyakonov, M.I., Shur, M.S.: Shallow water analogy for a ballistic field effect transistor: new mechanism of plasma wave generation by dc current. Phys. Rev. Lett. **71**, 2465 (1993)
8. Dyakonov, M.I., Shur, M.S.: Detection, mixing, and frequency multiplication of terahertz radiation by two dimensional electronic fluid. IEEE Trans. Electron Devices **43**, 380 (1996)
9. Muravjov, A.V., Veksler, D.B., Popov, V.V., Polischuk, O.V., Pala, N., Hu, X., Gaska, R., Saxena, H., Peale, R.E., Shur, M.S.: Temperature dependence of plasmonic terahertz absorption in grating-gate GaN HEMT structures. Appl. Phys. Lett. **96**, 042105 (2010)
10. Hu, B.B., Nuss, M.C.: Imaging with terahertz waves. Opt. Lett. **20**, 1716–1718 (1995)
11. Woolard, D.L., Koscica, T., Rhodes, D.L., Cuj, H.L., Pastore, R.A., Jensen, J.O., Jensen, J.L., Loerop, W.R., Jacobsen, R.H., Mittleman, D., Nuss, M.C.: Millimeter wave-induced vibrational modes in DNA as a possible alternative to animal tests to probe for carcinogenic mutations. J. Appl. Toxicol. **17**, 243–246 (1997)
12. Nishizawa, J.: Development of THz wave oscillation and its application to molecular sciences. Proc. Jpn. Acad. B **80**, 74 (2004)
13. Kemp, M.C., Taday, P.F., Cole, B.E., Cluff, J.A., Fitzgerald, A.J., Tribe, W.R.: Security applications of terahertz technology. In: Hwu, R.J., Woolard, D.L. (eds.) Terahertz for Military and Security Applications, vol. 5070, pp. 44–52. SPIE, Bellingham (2003)
14. George, P.A., Hui, W., Rana, F., Hawkins, B.G., Smith, A.E., Kirby, B.J.: Microfluidic devices for terahertz spectroscopy of biomolecules. Opt. Express **16**, 1577 (2008)
15. Yamashita, M., Kawase, K., Otani, C.: Imaging of large-scale integrated circuits using laser terahertz emission microscopy. Opt. Express **13**, 115 (2005)
16. Hunsche, S., Koch, M., Brener, I., Nuss, M.C.: THz near-field imaging. Opt. Commun. **150**, 22 (1998)
17. Huber, A.J., Keilmann, F., Wittborn, J., Aizpurua, J., Hillenbrand, R.: Terahertz near-field nanoscopy of mobile carriers in single semiconductor nanodevices. Nano Lett. **8**, 3766–3770 (2008)
18. Fitch, M.J., Osiander, R.: Terahertz waves for communications and sensing. Johns Hopkins APL Tech. Dig. **25**, 348 (2004)
19. Floyd, B.A., Hung, C.M., Kenneth, K.O.: Intra-chip wireless interconnect for clock distribution implemented with integrated antennas, receivers and transmitters. IEEE J. Solid State Circuits **37**, 543 (2002)

Teratogenic

▶ Use of Nanotechnology in Pregnancy

Theoretical Elasticity

▶ Superelasticity and the Shape Memory Effect

Theory of Artificial Electromagnetic Materials

▶ Theory of Optical Metamaterials

Theory of Nonadiabatic Electron Dynamics in Nanomaterials

Alexey V. Akimov and Oleg V. Prezhdo
Department of Chemistry, University of South California, Los Angeles, CA, USA

Synonyms

Charge transfer; Electron transfer; Molecular dynamics; Nonadiabatic dynamics; Photocatalytic; Photovoltaic; Quantum dynamics; Surface hopping

Definition

Nonadiabatic dynamics is an evolution of a coupled quantum-classical electron–nuclear system involving changes of electronic state in response to nuclear trajectory.

Basic Grounds of Nonadiabatic Dynamics

Electron dynamics is a process involving changes of the quantum state of an electronic subsystem over time. The quantum state is defined by the system wave function, $\Psi(r, R, t)$. The wave function depends on electronic, r, and nuclear, R, coordinates of all particles belonging to the system and on time, t. The evolution of the wave function is given by the time-dependent Schrodinger equation (TD-SE):

$$i\hbar\frac{\partial\Psi(r,R,t)}{\partial t}=\hat{H}(r,R,t)\Psi(r,R,t), \qquad (1)$$

where $\hat{H}(r, R, t)$ is the total Hamiltonian of the system. The total Hamiltonian is a sum of the kinetic energy of nuclei, $\hat{T}_{nucl}$, and the electronic Hamiltonian, $\hat{H}_{el}(r, R, t)$:

$$\hat{H}(r,R,t)=\hat{T}_{nucl}+\hat{H}_{el}(r,R,t), \qquad (2a)$$

$$\hat{T}_{nucl}=-\sum_{\alpha}\frac{\hbar^2}{2M_\alpha}\nabla_\alpha^2. \qquad (2b)$$

Here, M_i is the mass of the i-th nucleus, $\vec{\nabla}_\alpha\equiv\left(\frac{\partial}{\partial X_\alpha},\frac{\partial}{\partial Y_\alpha},\frac{\partial}{\partial Z_\alpha}\right)^T$ is the gradient operator, and $\hbar$ is the reduced Planck's constant. The electronic Hamiltonian, $\hat{H}_{el}(r, R, t)$, incorporates kinetic energy of the electrons, electron–nuclear, electron–electron, and nuclear–nuclear interactions. The latter is merely an additive constant depending only on the geometry of the system. The electronic Hamiltonian may also include explicit dependence on time, e.g., via field–matter interaction terms.

Solution of Eq. 1 constitutes the main goal of electronic dynamics studies. For simple low-dimensional models, the equation can be solved numerically on a grid. Such solution is considered exact, provided that appropriate numerical parameters are used, namely, sufficiently small integration time step, dense grid, and stable and accurate integration scheme.

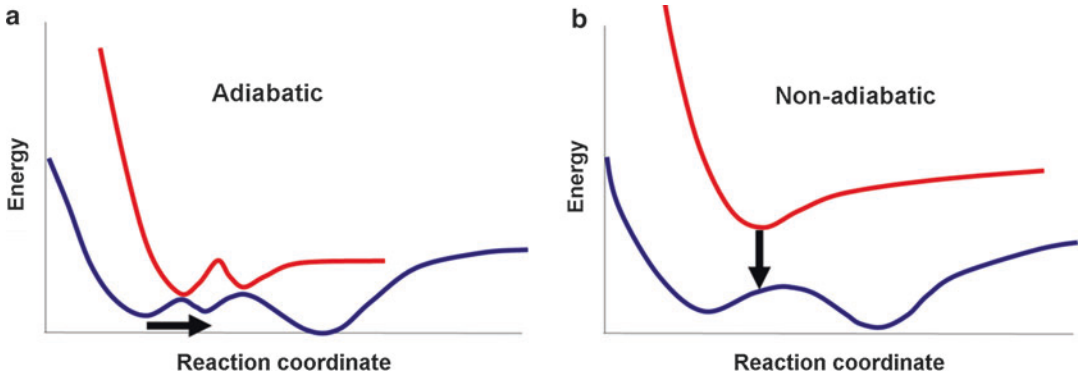
In most practical cases, the TD-SE is solved by representing the total wave function as a linear combination of products of electronic, $\Phi_i(r; R(t))$, and nuclear, $\chi_i(t, R(t))$, wave functions:

$$\Psi(r,R,t)=\sum_i\chi_i(t,R(t))\Phi_i(r;R(t)). \qquad (3)$$

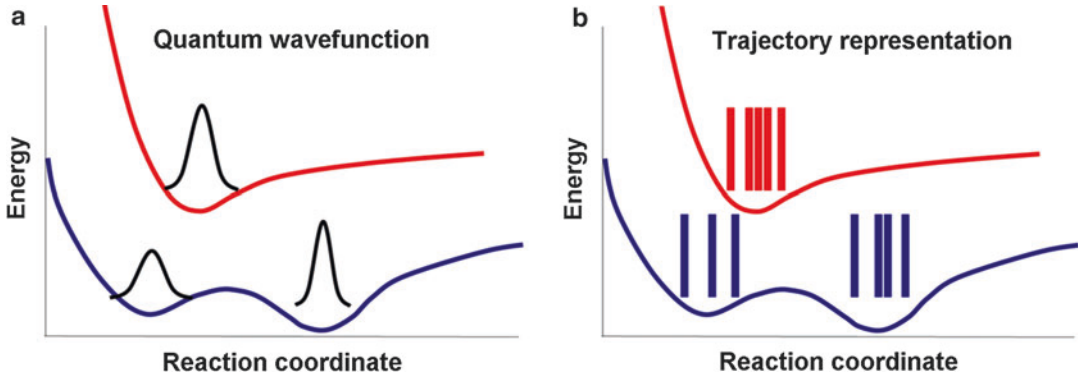
The ansatz, Eq. 3, reflects the principle of separation of variables. The principle originates from the fact that nuclear masses are significantly larger than the electronic mass. Hence, electrons and nuclei move on disparate timescales.

Depending on the number of terms in the sum, Eq. 3, one may distinguish two main types of quantum dynamics – adiabatic and nonadiabatic (Fig. 1). If only one term in Eq. 3 is adequate for an accurate description of quantum dynamics, the dynamics is called adiabatic (Fig. 1a). In this case, electron dynamics occurs by delocalization of the electronic wave function, i.e., its diffusion. The diffusion is induced by nuclear motions and leads to adiabatic electron or energy transfer.

In the case when it is necessary to retain many terms in Eq. 3 for achieving reasonable accuracy



Theory of Nonadiabatic Electron Dynamics in Nanomaterials, Fig. 1 Adiabatic (**a**) and nonadiabatic (**b**) electron dynamics. The *blue* and *red* potential energy surfaces correspond to different electronic states



Theory of Nonadiabatic Electron Dynamics in Nanomaterials, Fig. 2 Electron–nuclear wave function in different representations: (**a**) quantum-mechanical wave packet description, (**b**) semiclassical ensemble of trajectories representation

and qualitatively correct picture, the electron dynamics is called nonadiabatic (Fig. 1b). Nonadiabatic dynamics is realized when many electronic wave functions are notably involved. This is the case of near or complete degeneracy, which is very common for photoinduced isomerization, bond breaking, intersystem crossing, and long-distance electron transfer. A vast majority processes occurring in nanomaterials proceed via the nonadiabatic mechanism. Hence, theoretical foundation for simulating such processes is of great importance.

In the ansatz, Eq. 3, the stationary electronic states, $\Phi_i(r; R(t))$, are typically considered basis functions. They are obtained by solving the stationary Schrodinger equation:

$$\hat{H}_{el}(r,R,t)\Phi_i(r;R(t),t) = E_i(R(t),t)\Phi_i(r;R(t),t). \quad (4)$$

The stationary state wave functions depend on nuclear coordinates only parametrically. The time dependence of the stationary wave functions and energy eigenvalues is incorporated implicitly via time-dependent nuclear coordinates, $R(t)$, although an explicit dependence can be present as well. The stationary electronic states are obtained with the standard methods of quantum chemistry, such as ab initio or density functional theory (DFT) approaches. The wave functions can be chosen in many ways, depending on the nature of electronic states and the approximations used to study dynamics and to solve the stationary SE (Fig. 2).

Most often, they are associated with N-electron Slater determinants, although more accurate choices are based on linear combinations of such determinants or on their spin- or symmetry-adapted versions. In simplistic approximations, the electronic basis states can be chosen as Kohn–Sham (KS) or Hartree–Fock 1-electron molecular orbitals (MOs).

Once the electronic basis states are determined, the goal of solving the TD-SE, Eq. 1, transforms into finding the expansion coefficients, $\chi_i(t, R(t))$, as functions of time. These coefficients are essentially the nuclear wave functions corresponding to each of the electronic states. The squares of their amplitudes, $|\chi_i(t, R(t))|^2$, determine the probability of finding the system in the electronic state $\Phi_i(r; R(t))$ at time t and point $R(t)$ of the nuclear phase space. The nuclear wave functions are multidimensional and can have complex non-monotonous dependence on their arguments, especially when they describe highly excided vibrational states.

Substituting the ansatz, Eq. 3, into the TD-SE, Eq. 1, and projecting the result onto electronic state i give the TD-SE for the nuclear wave functions:

$$i\hbar\frac{\partial\chi_i(t,R(t))}{\partial t}=\sum_j\left\{\left(\hat{T}_{nucl}+E_j(R)\right)\delta_{ij}-\hbar^2\sum_\alpha\frac{\vec{d}^{(1)}_{ij,\alpha}}{M_\alpha}\vec{\nabla}_\alpha-\hbar^2\sum_\alpha\frac{d^{(2)}_{ij,\alpha}}{2M_\alpha}\right\}\chi_j(t,R(t)). \tag{5}$$

The product $\vec{d}^{(1)}_{ij,\alpha}\vec{\nabla}_\alpha$ is a scalar product of two vectors, $\left(\vec{d}^{(1)}_{ij,\alpha}\right)^T\vec{\nabla}_\alpha$, $d^{(n)}_{ij,\alpha}$ are the nth order nonadiabatic couplings between the electronic states i and j, induced by the nuclear degree of freedom α:

$$\vec{d}^{(1)}_{ij,\alpha}=\left\langle\Phi_i(r;R(t))\middle|\vec{\nabla}_\alpha\Phi_j(r;R(t))\right\rangle, \tag{6a}$$

$$d^{(2)}_{ij,\alpha}=\left\langle\Phi_i(r;R(t))\middle|\nabla^2_\alpha\Phi_j(r;R(t))\right\rangle. \tag{6b}$$

For convenience, Eq. 5 can be recast into a vector–matrix form:

$$i\hbar\dot{\chi}=\hat{H}_{vib}\chi, \tag{7}$$

where $\chi=(\chi_1\quad\ldots\quad\chi_N)^T$ is the vector of nuclear wave functions corresponding to all electronic states up to state N and $\hat{H}_{vib}$ is the quantum-mechanical vibronic Hamiltonian operator, whose matrix elements are given by

$$H_{vib,ij}\equiv\langle\Phi_i|\hat{H}_{vib}|\Phi_j\rangle=\left(\hat{T}_{nucl}+E_j(R)\right)\delta_{ij}-\hbar^2\sum_\alpha\frac{\vec{d}^{(1)}_{ij,\alpha}}{M_\alpha}\vec{\nabla}_\alpha-\hbar^2\sum_\alpha\frac{d^{(2)}_{ij,\alpha}}{2M_\alpha}. \tag{8}$$

The TD-SE, Eq. 5, is fully quantum mechanical and is exact. The only approximation made so far concerns separation of variables. However, with the choice of sufficiently large basis, the errors introduced by such an approximation can be minimized to negligible. Equation 5 is, therefore, the main equation of nonadiabatic electron dynamics. It can be solved directly or can be used as the starting ground for derivation of approximate but computationally more efficient simulation schemes applicable to large systems, for studying nonadiabatic electron dynamics in nanomaterials.

Methods for Nonadiabatic Electron Dynamics in Nanomaterials

Practical simulation of electron dynamics in large nanoscale systems necessitates a simple and efficient methodology. Typically, a fully quantum-mechanical treatment is avoided because it is very costly. Simpler theories and approximations are utilized instead [1].

Semiclassical and quantum-classical methods give rise to the most popular group of approximations. These methods mimic an intrinsically quantum-mechanical object, such as a wave packet, by a set of fully classical objects – classical trajectories that undergo a modified evolution. Approximate treatments start by

neglecting quantum description of nuclei and treating them classically. This is equivalent to taking a classical limit of quantum-mechanical operators acting on nuclear degrees of freedom:

$$-\hbar^2 \frac{\vec{d}^{(1)}_{ij,\alpha}}{M_\alpha} \vec{\nabla}_\alpha \rightarrow -i\hbar \frac{\vec{d}^{(1)}_{ij,\alpha}}{M_\alpha} \vec{p}_\alpha, \tag{9a}$$

$$\hat{T}_{nucl} = -\sum_\alpha \frac{\hbar^2}{2M_\alpha} \nabla^2_\alpha \rightarrow \sum_\alpha \frac{p^2_\alpha}{2M_\alpha} = T_{nucl}. \tag{9b}$$

The approximations are accurate when quantum effects are weak – at high temperatures and in sufficiently large systems, in which quantization effects become negligible. The limits, Eqs. 9a and b, allow one to avoid nonlocal nuclear effects, which would couple phase-space points with different nuclear positions and momenta. The nonlocal effects originate because of the derivative operators, ∇_α and ∇^2_α, acting on the nuclear component of the overall wave function. When these derivative operators are replaced by the classical phase-space variables, the quantum vibronic Hamiltonian, Eq. 8, reduces to the semiclassical Hamiltonian:

$$H_{vib,ij} = \left(\left(\sum_\alpha \frac{p^2_\alpha}{2M_\alpha} \right) + E_j(R) \right) \delta_{ij} - i\hbar \sum_\alpha \frac{\vec{p}^T_\alpha \vec{d}^{(1)}_{ij,\alpha}}{M_\alpha} - \hbar^2 \sum_\alpha \frac{d^{(2)}_{ij,\alpha}}{2M_\alpha}. \tag{10}$$

The system of partial differential equations of the second order, Eq. 5, simplifies to the system of ordinary differential equations of the first order. The latter are much simpler to solve, resulting in faster calculations of the effective nuclear wave packet dynamics.

Electron–Nuclear Coupling, Electron Back-Reaction

The quantum-mechanical TD-SE, Eq. 7, with the corresponding quantum-mechanical vibronic Hamiltonian, Eq. 8, provides a rigorous way to the coupled electron–nuclear dynamics, in which both nuclei and electrons are treated on the same footing. The semiclassical approximation, Eq. 9, breaks this homogeneity. As a result, a special treatment of electron–nuclear coupling is required. This special treatment is based on approximating nuclei classically and electrons quantum-mechanically via trajectory-based methods. The original Hilbert space for quantum dynamics of vibronic states is thus factorized into a direct product of the Hilbert space of electronic basis states and the phase space of nuclear positions and momenta.

The trajectory-based semiclassical schemes constitute a de facto standard for simulating dynamics with many electronic states. Following the separation of classical and electronic variables, generalized to many states, one can consider electron dynamics occurring along classical nuclear trajectories. This is called a classical path approximation (CPA) in a broad sense.

An ensemble of trajectories (Fig. 2) represents the quantum-mechanical wave function within the CPA methods. Each trajectory follows stochastic evolution in the space of electronic states, nuclear positions, and velocities. A swarm of trajectories represents a combined electron–nuclear wave function. An intrinsically quantum object, such as wave packet, is mimicked by a set of fully classical objects – classical trajectories that undergo stochastic jumps in their electronic state labels. This philosophy is a distinguishing feature of the quantum-classical and semiclassical methods.

The electronic dynamics depends on the nuclear evolution through the parametric dependence of the electronic Hamiltonian on nuclear coordinates. In turn, the evolution of nuclei themselves is affected by the dynamics of electronic state. When the electronic state changes – suddenly or continuously – the nuclear trajectory has to be adjusted accordingly. This effect is called electron back-reaction. Generally, there are three types of description of electron–nuclear coupling and associated electron back-reaction:

- Mean field (MF) or Ehrenfest
- Trajectory surface hopping (TSH)
- Neglect of back-reaction (NBR), which is often called CPA, but in a narrow sense

Theory of Nonadiabatic Electron Dynamics in Nanomaterials, Table 1 Comparative scheme of three groups of methods: mean field, trajectory surface hopping, and neglect of back-reaction approximation

Method	Nuclei affect electrons	Electrons affect nuclei	Nuclear dynamics	Electronic dynamics
MF	Yes	Yes	Continuous	Continuous
TSH	Yes	Yes	Stochastic	Stochastic
NBR	Yes	No	Continuous	Stochastic

The methods comparison is summarized in Table 1.

Below we provide more detail on the methodologies and emphasize their strong and weak points.

Mean-Field or Ehrenfest Dynamics

Both electronic and nuclear degrees of freedom vary continuously and deterministically in the MF scheme. The electronic state variables are represented by the wave functions, χ, in the basis set expansion (Eq. 3). Various integration schemes (integrators) can be used to propagate the wave function of a system, by solving Eq. 7 with the semiclassical vibronic Hamiltonian (Eq. 10).

The formal solution of Eq. 7 is given via the matrix exponent:

$$\chi(t) = \exp\left(-\frac{i}{\hbar}\int_{t}^{t_0} H_{vib}(t')dt'\right)\chi(t_0). \qquad (11)$$

For numerical purposes, the time interval $t - t_0$ can always be split into sufficiently small subintervals, such that the Hamiltonian matrix can be assumed constant during each subinterval and the integral in the exponent can be approximated by $H_{vib} \cdot (t - t_0)$. The constant H_{vib} can be chosen as the midpoint value, for example. Then, the matrix exponent can be computed via diagonalization of the Hamiltonian matrix and the Taylor series representation of the exponential function, generalized to matrices.

The solution, Eq. 11, assures that the evolution of wave functions is time reversible and that the norm of wave functions is always conserved to unity. This allows one to utilize sufficiently large integration time steps and integrate for long time intervals without pathological divergences and numerical instabilities. However, the need for diagonalization of the vibronic Hamiltonian matrix leads to a cubic scaling of the required CPU time, $O(N^3)$, with the number of basis states, N. Such calculations are therefore practical only for relatively small problems – with no more than a few hundreds of basis states. In many cases, the dimensionality of the space of vibronic states can be very large, up to millions of states. More efficient schemes are required under these conditions.

An alternative approach to solving Eq. 7 with the semiclassical Hamiltonian, Eq. 8, is based on the Liouville formalism followed by the Trotter factorization of the evolution operator. The solution of equations of motion, Eq. 7, can be given in terms of the Liouville operator, iL:

$$\chi(t) = \exp(iL \cdot (t - t_0))\chi(t_0), \qquad (12)$$

which is essentially the differential form of Eq. 11. The exponential operator $\exp(iL \cdot (t - t_0))$ is called the evolution operator or propagator. It is closely related to the nonequilibrium Green's functions (NEGFs).

The Liouville operator corresponding to Eq. 7 is given by

$$iL = \sum_{i=0}^{N-1}\dot{\chi}_i\frac{\partial}{\partial\chi_i} = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left(-\frac{i}{\hbar}H_{ij}\right)\chi_j\frac{\partial}{\partial\chi_i} = \sum_{i=0}^{N-1}\left(-\frac{i}{\hbar}H_{ii}\right)\chi_i\frac{\partial}{\partial\chi_i} + \sum_{i=0}^{N-1}\sum_{j>i}^{N-1} -\frac{i}{\hbar}\left(H_{ij}\chi_j\frac{\partial}{\partial\chi_i} + H_{ji}\chi_i\frac{\partial}{\partial\chi_j}\right)$$
$$= \sum_{i=0}^{N-1} iL_i^{(1)} + \sum_{i=0}^{N-1}\sum_{j>i}^{N-1} iL_{ij}^{(2)} \qquad (13)$$

Similar to integration of Eq. 11, the solution, Eq. 12, can be obtained by splitting the integration interval $t - t_0$ into smaller integration subintervals, dt. The action of the propagator on each subinterval can be obtained by the second-order symmetric factorization:

$$\begin{aligned}\exp(iL\cdot dt) = {} & \exp\left(\sum_{i=0}^{N-1} iL_i^{(1)}\cdot\frac{dt}{2}\right)\exp\left(iL_{01}^{(2)}\cdot\frac{dt}{2}\right)\exp\left(iL_{02}^{(2)}\cdot\frac{dt}{2}\right)\ldots\exp\left(iL_{0(N-1)}^{(2)}\cdot\frac{dt}{2}\right)\\ & \exp\left(iL_{12}^{(2)}\cdot\frac{dt}{2}\right)\ldots\exp\left(iL_{1(N-1)}^{(2)}\cdot\frac{dt}{2}\right)\ldots\exp\left(iL_{(N-2)(N-1)}^{(2)}\cdot\frac{dt}{2}\right)\exp\left(iL_{(N-2)(N-1)}^{(2)}\cdot\frac{dt}{2}\right)\ldots\\ & \exp\left(iL_{1(N-1)}^{(2)}\cdot\frac{dt}{2}\right)\ldots\exp\left(iL_{12}^{(2)}\cdot\frac{dt}{2}\right)\exp\left(iL_{0(N-1)}^{(2)}\cdot\frac{dt}{2}\right)\ldots\exp\left(iL_{02}^{(2)}\cdot\frac{dt}{2}\right)\exp\left(iL_{01}^{(2)}\cdot\frac{dt}{2}\right)\\ & \exp\left(\sum_{i=0}^{N-1} iL_i^{(1)}\cdot\frac{dt}{2}\right)+O(dt^2).\end{aligned} \tag{14}$$

The factorization, Eq. 14, contains the operators of two types: $\exp(iL_i^{(1)}dt)$ and $\exp(iL_{ij}^{(2)}dt)$. Their action is given by the unitary and time-reversible maps:

$$\exp\left(iL_i^{(1)}dt\right) = \exp\left(\alpha_i\cdot dt\cdot\chi_i\frac{\partial}{\partial\chi_i}\right):\chi_i\to\exp(\alpha_i\cdot dt)\chi_i = \left[\cos\left(\frac{dt}{\hbar}\varepsilon_i\right) - i\sin\left(\frac{dt}{\hbar}\varepsilon_i\right)\right]\chi_i. \tag{15a}$$

$$\exp\left(iL_{ij}^{(2)}dt\right) = \exp\left(dt\cdot\beta_{ij}\left(\chi_j\frac{\partial}{\partial\chi_i} - \chi_i\frac{\partial}{\partial\chi_j}\right)\right):\begin{pmatrix}\chi_i\\ \chi_j\end{pmatrix}\to\begin{pmatrix}\cos(dt\cdot\beta_{ij}) & \sin(dt\cdot\beta_{ij})\\ -\sin(dt\cdot\beta_{ij}) & \cos(dt\cdot\beta_{ij})\end{pmatrix}\begin{pmatrix}\chi_i\\ \chi_j\end{pmatrix}, \tag{15b}$$

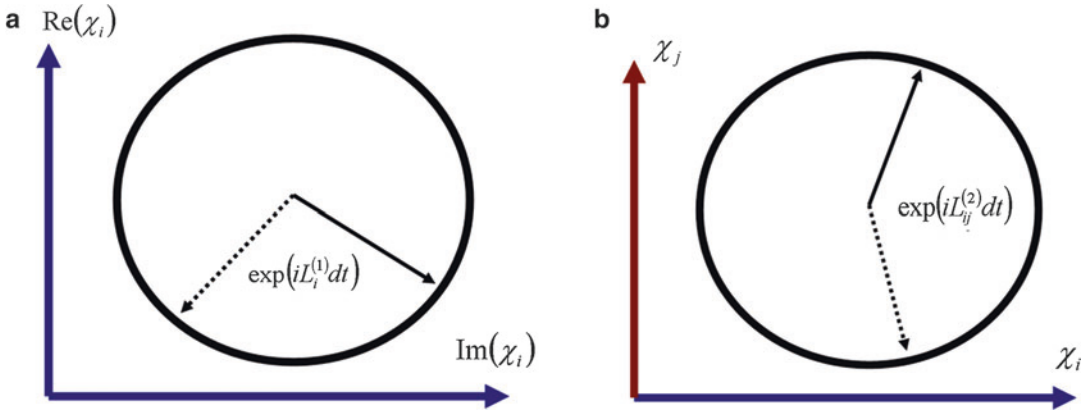
$$\alpha_i = -\frac{i}{\hbar}H_{ii}, \tag{15c}$$

$$\beta_{ij} = -\frac{i}{\hbar}H_{ij}. \tag{15d}$$

The maps, Eqs. 15, carry a transparent interpretation: the operators of the first type, $\exp(iL_i^{(1)}dt)$, propagate the phase of each electronic state, while the operators of the second type, $\exp(iL_{ij}^{(2)}dt)$, account for population transfer between each pair of electronic states. Both maps have a form of rotation. The first one, Eq. 15a, performs a rotation of a single state vector on the plane spanned by its real and imaginary coordinates (Fig. 3a). The map of the second type, Eq. 15b, rotates two state vectors on the complex hyperplane spanned by their original directions (Fig. 3b).

The overall vibronic dynamics that is generated by Eq. 7 with the Hamiltonian, Eq. 8, can be though as a sequence of rotations of different kind. Because each of the involved operators is unitary, their action conserves the sum of the norms of the wave functions they act on. The norm of the wave function is conserved exactly. The symmetric splitting scheme, Eq. 13, assures that the propagator is time reversible. Most importantly, the integration based on the operator splitting scales quadratically $O(N^2)$, which allows one to include more basis states the dynamics.

T



Theory of Nonadiabatic Electron Dynamics in Nanomaterials, Fig. 3 Illustration of action of the operators of first and second type (Eq. 15): (**a**) the operator $\exp(iL_i^{(1)}dt)$ propagates the phase of a given state χ_i, rotation in the plane spanned by real and imaginary components of that complex number; (**b**) the operator $\exp(iL_{ij}^{(2)}dt)$ transfers the population between the two states χ_i and χ_j, rotation in the complex plane spanned by these two complex numbers

In the MF method, the dynamics of nuclei is governed by classical Newton's laws, with the forces given by the averages of forces on all electronic potential energy surface (PES), weighted by the squares of the corresponding vibronic amplitude, $|\chi_i|^2$. The off-diagonal elements of the vibronic Hamiltonian may be nonzero in a general basis and may contribute to the overall mean-field force acting on the particles:

$$\vec{F}_{MF,\alpha} = \sum_i |\chi_i|^2 \vec{F}_{i,\alpha} + \sum_{i,j} \chi_i^* \chi_j (E_i - E_j) d_{\alpha,ij}^{(1)}. \tag{16}$$

The force corresponding to the adiabatic electronic state $|i\rangle$ is computed according to:

$$\vec{F}_{i,\alpha} = -\langle i| \frac{\partial V}{\partial \vec{R}_\alpha} |i\rangle, \tag{17}$$

The Ehrenfest method is mostly applicable in the regions where the coupling between electronic states is very strong, such that the weights of different electronic states are comparable to one another. Under such conditions, the TSH would produce frequent surface hops between coupled potential energy surfaces and therefore would effectively be similar to the MF treatment.

The trajectories generated in the Ehrenfest dynamics are fully deterministic and continuous. There are no stochastic hops between energy surfaces. Moreover, the Ehrenfest dynamics can be given a fully Hamiltonian formulation, in the sense of theoretical mechanics. Such an approach provides a formally classical formulation of the intrinsically quantum dynamics of electronic states. The Hamiltonian formulation is simple and transparent and enhances the analysis of instabilities and other dynamical properties of the electronic subsystem.

The fully Hamiltonian formulation restricts the dynamics to closed systems. System–bath interactions, which induce electronic decoherence via wave function collapse, cannot be described within such a framework. Instead, the wave function always exists as a superposition of pure states, even in regions with negligible coupling between the states. Such state mixing is unphysical. In addition, the method excessively preserves coherences – the effect that typically accelerates electronic transitions. Because of the MF nature, the Ehrenfest approach is unable to account for wave packet splitting – branching and related tunneling effects. Finally, the Ehrenfest dynamics is unable to achieve thermal equilibrium in mixed quantum-classical systems. In particular, it misrepresents energy flow from the

electronic to the nuclear subsystem and violates the principle of detailed balance.

Recently, electronic decoherence was introduced in the MF framework via coherence penalty functional (CPF) [2]. The formulation starts with the effective mapped Hamiltonian and introduces new terms that penalize the development of excessive coherences during the coupled electron–nuclear evolution. The terms serve to include quasi-random fluctuations in the rotation angles describing evolution of the electronic subsystem (e.g., see Fig. 3). The formulation remains Hamiltonian, but it affects the correlation of the electronic states and introduces phenomenological decoherence corrections. As a result, the overall electron transfer rates decrease – the effect consistent with experiment and the results obtained by TSH-based decoherence schemes.

Trajectory Surface Hopping

Developed by Tully in 1990 [3], fewest switches surface hopping (FSSH) is most popular among various TSH formulations. The seminal work is cited over 1,000 times. The FSSH method builds on the MF. Unlike the MF formulation, there are two distinctive features: (i) nuclei are propagated on a single PES, which may vary in time stochastically; (ii) the evolution of electronic variables is stochastic, although it is based on the auxiliary semiclassical TD-SE solutions, χ_i, obtained in a manner similar to MF, except for the condition (i).

The FSSH assigns a probability for the transition from the current electronic state $|i\rangle$ to a new state $|j\rangle$ during a sufficiently small time interval Δt according to the rule

$$g_{i\to j}(t) = \max\left(0, P_{i\to j}(t)\right). \tag{18}$$

The raw hopping probabilities, $P_{i\to j}(t)$, are derived from the requirement that the probability, $P_i(t)\Delta t$, to leave the active state i, currently occupied by the system, to any other state during the time interval Δt is consistent with the quantum-mechanical change of electronic populations:

$$P_i(t)\Delta t = -\frac{\dot{\rho}_{ii}\Delta t}{\rho_{ii}}, \tag{19}$$

where ρ_{ii} is the diagonal element of the vibronic density matrix:

$$\rho_{ii} = \chi_i^*\chi_i. \tag{20}$$

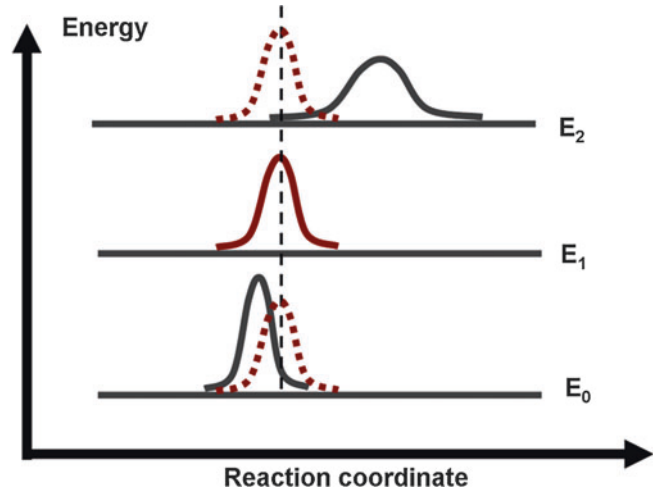
The total probability, $P_i(t)\Delta t$, can be split into probabilities to hop into different final states j. The splitting is not unique, in general. The most commonly used choice is:

$$\begin{aligned} P_{i\to j}(t) &= 2\frac{\int_t^{t+\Delta t} \mathrm{Re}\left(\chi_i^*(t')\chi_j(t')\sum_\alpha \frac{\vec{p}_\alpha(t')}{M_\alpha}\vec{d}^{(1)}_{\alpha,ij}(t')\right)dt'}{\chi_i^*(t)\chi_i(t)} \\ &\approx 2\frac{\mathrm{Re}\left(\chi_i^*(t)\chi_j(t)\sum_\alpha \frac{\vec{p}_\alpha(t)}{M_\alpha}\vec{d}^{(1)}_{\alpha,ij}(t)\right)\Delta t}{\chi_i^*(t)\chi_i(t)}. \end{aligned} \tag{21}$$

The probabilities, Eq. 21, are compared in the FSSH algorithm with a uniformly distributed random number in order to determine if the system should remain in the current electronic state or if it should move (hop) to a new state. If the electronic transition is selected, the nuclear velocities should be rescaled to conserve the total electron–nuclear, quantum-classical energy. If such rescaling is not possible, the transition is rejected. The rejected (also called frustrated) hops are necessary to obey the detailed balance in quantum-classical systems.

The FSSH method is very accurate in most cases. The use of multiple trajectories and stochastic events allows the quantum-mechanical branching – the trajectories may reside in different parts of the nuclear phase space and may be labeled with different discrete electronic state labels (e.g., see Fig. 2). This effect is typically not achievable in the plain MF formulation. Yet, the FSSH method has several difficulties.

Based on the MF theory, the FSSH method inherits the overcoherence problem. This effect originates from the CPA, which states that for each individual trajectory, the nuclear coordinates for all electronic states are restricted to be the same. In the fully quantum-mechanical formulation, the wave function may spread over different



Theory of Nonadiabatic Electron Dynamics in Nanomaterials,
Fig. 4 Illustration of the overcoherence problem in the CPA-based TSH algorithms, such as FSSH. In the CPA methods, the matrix elements of the vibronic Hamiltonian are evaluated at the same position (*red distribution*), while the positions may be different in the fully quantum treatment

nuclear coordinate regions for different electronic states (Fig. 4).

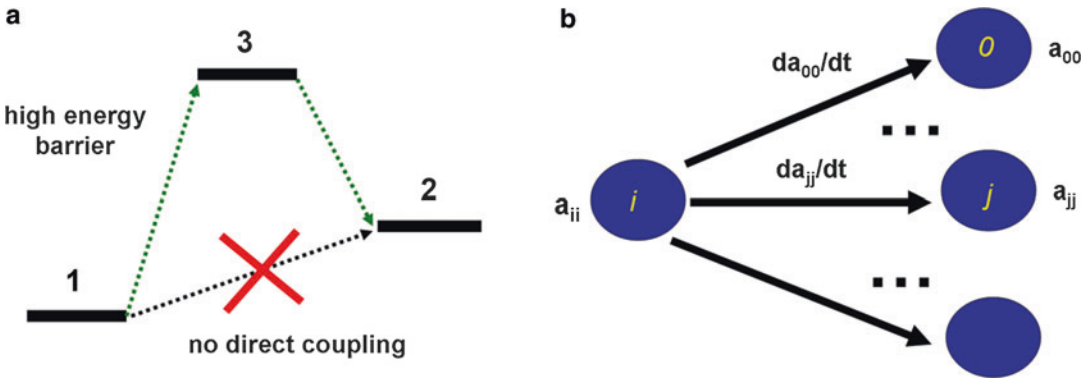
In the TSH methods, the nuclear component of the vibronic states χ_i is essentially a delta distribution, $\prod_\alpha \delta\left(\vec{p}_\alpha - \left\langle \vec{p}_\alpha \right\rangle\right)\delta\left(\vec{R}_\alpha - \left\langle \vec{R}_\alpha \right\rangle\right)$. It can be regarded as a zero-width limit of the products of Gaussian distributions. In the case of finite-width Gaussian distributions, the evaluation of the NACs involves overlaps of nuclear wave packets. All the wave packets are positioned at the same center under the CPA, and the overlap is unitary at all times. In the fully quantum dynamics, the wave packets can diverge, leading to a decay of the overlap magnitude. The NACs are decreased accordingly. This leads to slower electronic transitions – the effect caused by the loss of electronic coherence.

To account for effects of coherences, many decoherence schemes were developed. Schemes based on the stochastic Schrodinger equation [4] achieve branching due to decoherence, obviating the need for ad hoc expressions, such as Eq. 21. These frameworks imply non-unitary wave function evolution, which is needed for the description of open systems and system–bath interactions. Phenomenological approaches are based on the manual wave function collapse that occurs on the timescales of electronic decoherence. For example, in the recently developed decoherence-induced surface hopping (DISH) scheme [5], the decoherence rates, r_{ij}, are computed using the optical response theory or wave packet dynamics results. These quantities are used to compute time-dependent decoherence interval time, τ_i, via:

$$\frac{1}{\tau_i} = \sum_{j \neq i} |\chi_j|^2 r_{ij}. \tag{22}$$

Surface hops are attempted when the time since the last attempt has past the decoherence interval. State i diverges from all other states on the time-scale of decoherence interval, τ_i. The state is "measured/observed" by the environment (bath), and, as a consequence, the wave function should collapse onto either the diverged state or a superposition of all other states. The DISH method used the same velocity rescaling procedure as in the original FSSH, to satisfy detailed balance.

Yet another difficulty of the FSSH method concerns its formulation in a diabatic (site) basis. Namely, the phenomenon called superexchange cannot be modeled. The problem arises when there are states 1 and 2 that are not directly coupled to each other, and the population transfer is mediated by a higher-lying state 3 (Fig. 5a). The overall population transfer between the states 1 and 2 is inhibited in the FSSH by the Boltzmann factor associated with the intermediate population transfer from the state 1 to the state 3. On the contrary, the semiclassical TD-SE does not



Theory of Nonadiabatic Electron Dynamics in Nanomaterials, Fig. 5 Superexchange problem. (**a**) Electronic states 1 and 2 are not directly coupled; thus population transfer is only possible via the intermediate state 3. The high-energy barrier for transition from 1 to 3 inhibits the overall transition 1 → 2. In the MF description, the population transfer is not affected by the energy barrier. (**b**) Probabilities to leave the active state i via different channels j are proportional to the rates of change of the state j population

include any factors inhibiting the transfer of electronic population between the states 1 and 3. Consequentially, the true electron transfer rate between the states 1 and 2 can be significantly larger than predicted by the FSSH, especially when the temperature is low.

To circumvent the difficulties of the FSSH in describing the superexchange, the global-flux surface hopping (GFSH) scheme has been developed recently [6]. The method modifies the surface hopping probabilities. The total probability to leave the active state i is still given by the expression Eq. 19; however, the relative probabilities to take different routes, i → j, are changed. In contrast to the splitting scheme based on Eq. 21, the weights of the available channels are taken in proportion to the rates of change of the populations of final states (Fig. 5b):

$$P_{i\to j}(t) \sim \dot{a}_{jj}(t). \tag{23}$$

This construction avoids the Boltzmann factor associated with the artificial barrier. It also ensures that the change of the number of trajectories associated with different electronic states is consistent with the solution of the TD-SE.

Neglect of Back-Reaction

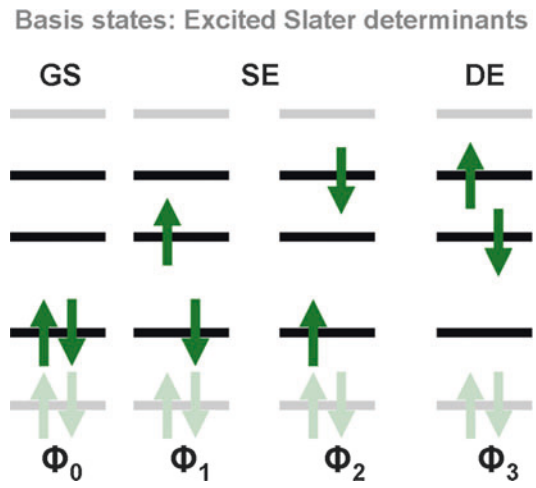
The evolution of nuclear degrees of freedom is affected in the MF and TSH approaches by either continuous (MF) or stochastic (TSH) change of the electronic state. This change affects the future of the electron–nuclear system. Statistically meaningful results can be obtained with a sufficient number of stochastic realizations. The number of such trajectories depends on the system, but may be quite large, hundreds or thousands. Propagation of each trajectory involves solving the stationary SE that produces instantaneous electronic basis states for every nuclear configuration along the trajectories. The calculation can become expensive in large nanoscale materials, creating a major bottleneck. Yet, it is common in nanomaterials that the electron back-reaction is weak and leads to small changes of nuclear coordinates and velocities. Under these conditions, a further simplification of either MF or TSH methods can be made – neglect of back-reaction.

In the NBR approximation, the classical trajectory is independent of electronic dynamics, while the electronic dynamics still depends on the classical coordinates. The NBR is valid if the energy of the nuclei is large, such that the electron–nuclear energy exchange does not affect the nuclear evolution appreciably. The NBR is also valid if the PES associated with different electronic states differs only slightly, compared to the amplitude of the thermal fluctuations in the nuclear coordinates, for instance.

The NBR approximation allows significant reduction of computational efforts. Unlike the TSH and MF approaches, the NBR approximation can use a single nuclear trajectory. The trajectory is then employed to solve the TD-SE along the classical path only once, whereas a large number of initial conditions and stochastic realizations of the TSH algorithm can be used afterward to obtain statistically accurate results. This methodology greatly increases the calculation speed, thus allowing computations on large systems. In order to represent the hop rejection of the original FSSH, the transition probabilities $g_{i\to j}(t)$, Eq. 18, are rescaled by the Boltzmann factor, taking into account detailed balance:

$$g_{i\to j}(t) \to g_{i\to j}(t) b_{i\to j}(t), \tag{24}$$

$$b_{i\to j}(t) = \begin{cases} \exp\left(-\dfrac{E_j - E_i}{k_B T}\right), E_j > E_i \\ 1, E_j \le E_i \end{cases}. \tag{25}$$



Theory of Nonadiabatic Electron Dynamics in Nanomaterials, Fig. 6 Excited Slater determinants are used as the electronic basis states for nonadiabatic molecular dynamics calculations. Excitations are only performed within the active space and include configurations of different kind: ground state (*GS*), singly excited (*SE*), and doubly excited (*DE*) states

Basis States

Construction of the electronic wave function basis, $\{\Phi_i\}, i = 0, \ldots, N_b - 1$, which is used for description of time-evolving wave function of the system, deserves a special discussion. Rigorous calculation of electronic excited states typically requires correlated electronic structure methods, such as configuration interaction (CI) or coupled-cluster (CC) theory, often combined with multireference (MR) treatments. However, such computations are extremely expensive and cannot be routinely applied to large systems.

An alternative approach employs a single Slater determinant constructed from the molecular orbitals (MO), produced at a specific level of theory:

$$\Phi_i(1, 2, \ldots, N) = \hat{A}\left[\varphi_{i_1}(1)\varphi_{i_2}(2)\ldots\varphi_{i_N}(N)\right]. \tag{26}$$

Here, $\varphi_i(n)$ is the adiabatic 1-electron spin orbital ($i = 1,\ldots,$ N) dependent on the coordinates of the nth electron, N is the number of the MOs included in the nonadiabatic molecular dynamics model, and $\hat{A}$ is the antisymmetrization operator. The electronic state basis is formed by excitations of the ground-state Slater determinant within the range of active MO levels (Fig. 6). The resulting multi-electron states can include single, double, and higher excitations of electrons and holes. The doubly and other highly excited configurations are important for modeling processes such as Auger recombination, multiple exciton generation, and singlet fission.

The formulation based on Kohn–Sham (KS) density functional theory (DFT) has been successfully used [7]. In it, the 1-electron orbitals in Eq. 26 are associated with KS spin orbitals. The formulation based on DFT accounts for dynamic electronic correlation via corresponding correlation functional, yet it is computationally favorable. The non-dynamic correlation effects are accounted for to a different degree by the use of multiple excited Slater determinant configurations. The latter constitute an effective diabatic basis. The method was widely used to model electron and hole transfer in a range of nanomaterials, represented by large molecular and solid-state systems.

Another approach is based on semiempirical or tight-binding calculations, such as extended Hückel theory (EHT) or density functional tight-

binding (DFTB) methods. These methods are significantly faster than DFT or ab initio schemes. Semiempirical methods often provide a clear physical interpretation of the interatomic and orbital interactions and are parameterized to reproduce experimental data. The parameterization partially accounts for exchange and correlation effects. The accuracy of such models can be comparable or even higher than that of the standard DFT or ab initio methods. The semiempirical methods allow studying large molecular systems. They have been used to study electron transfer dynamics in interfacial systems, such as dye-sensitized solar cell materials, and in nanoscale materials.

The Hamiltonian matrix elements between many-electron configurations are determined based on the propertied of MOs. The excitation energies are represented by the differences in the energies of the MOs occupied in the ground $\{\varepsilon_{i_k} : i_k \in GS\}$ and excited state $\{\varepsilon_{i_k} : i_k \in ES\}$ configurations. The diagonal elements of the Hamiltonian are:

$$E_i \equiv H_{ii} = \sum_{i_k \in ES} \varepsilon_{i_k} - \sum_{i_k \in GS} \varepsilon_{i_k}. \tag{27}$$

The nonadiabatic couplings determine the off-diagonal elements of the many-electron Hamiltonian. In particular, the coupling between the many-electronic states i and j, represented by Slater determinants Φ_i and Φ_j, is given by:

$$H_{ij} \equiv -i\hbar \langle \Phi_i | \frac{\partial}{\partial t} | \Phi_j \rangle. \tag{28}$$

Nonadiabatic couplings, defined according to Eq. 6, are essential for nonadiabatic dynamics simulations. To be more rigorous, the vectors (Eq. 6) are called the derivative couplings, while the nonadiabatic couplings are defined via time derivatives:

$$\vec{d}_{ij}^{(1)} = \left\langle \Phi_i(r; R(t)) \right| \frac{\partial}{\partial t} \Phi_j(r; R(t)) \Big\rangle = \sum_\alpha \frac{\vec{p}_\alpha \vec{d}_{ij,\alpha}^{(1)}}{M_\alpha}. \tag{29}$$

One can observe that nonadiabatic couplings, Eq. 10, do not necessarily include explicit derivative couplings. Derivative couplings are needed in the TSH algorithms, since nuclear velocities are rescaled along directions of derivative couplings. The nonadiabatic couplings of Eq. 29 suffice for NBR calculations, providing additional computational savings.

Exiting methods for nonadiabatic electron dynamics, especially those based on the NBR, compute nonadiabatic couplings using numerical time differentiation of electronic wave functions along the trajectory:

$$\langle \Phi_i | \frac{\partial}{\partial t} | \Phi_j \rangle = \frac{\langle \Phi_i(t) | \Phi_j(t+dt) \rangle - \langle \Phi_i(t+dt) | \Phi_j(t) \rangle}{2 \cdot dt}. \tag{30}$$

The utilization of Eq. 30 helps eliminate divergences of nonadiabatic couplings for nearly degenerate energy levels (e.g., conical intersection or weakly avoided crossing regions). Equation 30 is computationally very simple: it requires only wave functions at adjacent time steps. Note the minus sign. It arises because the $\frac{\partial}{\partial t}$ operator is anti-Hermitian or, equivalently, because the $i\hbar \frac{\partial}{\partial t}$ and momentum, $-i\hbar \frac{\partial}{\partial R}$, operators are Hermitian.

Software and Implementation

A number of programs capable of performing nonadiabatic dynamics simulations of atomistic systems exist. Examples include Newton-X, MOLPRO, MNDO, CPMD, Octopus, and Fireball – each with its own set of advantages and limitations. Typically, these programs provide access to high-level or semiempirical description of electronic excited states, analytic derivative and nonadiabatic couplings, and the TSH or MF algorithms. As such, the methods can mostly be applied to relatively small systems. Moreover, the schemes for decoherence corrections are rarely available.

Recently, the PYXAID [8, 9] (PYthon eXtension for Ab Initio Dynamics) program for nonadiabatic molecular dynamics has been reported. It takes advantage of DFT and semiempirical methods of electronic structure calculation. It uses the Slater determinant basis for time propagation of the wave function. Further, PYXAID implements the NBR and efficient TD-SE

integration schemes. All these factors allow calculations on large systems to be performed in a reasonable time. In addition, the package includes various decoherence schemes, such as DISH or CPF, to allow more realistic estimates of the timescales of electron and energy transfer processes. The package implements the GFSH methodology, to allow accurate simulation of dynamics involving superexchange pathways: charge transfer, conductivity, Auger recombination, multiple exciton generation, and singlet fission.

Application of Nonadiabatic Dynamics to Studies of Electron Transfer in Nanomaterials

In this section, we summarize some of the key principles of charge and energy transfer dynamics in nanoscale materials. In many cases, these results are obtained from theoretical analysis and computational modeling of prototype systems. In other situations, the initial insights are obtained experimentally and simulations are used to confirm the findings and to provide a better understanding of the underlying mechanisms and interrelations. The theory of nonadiabatic molecular dynamics has been widely used by many research groups to study charge and energy transfer processes in various materials. As examples, we consider electron dynamics in several types of systems, including interfaces, bulk, and soft materials, as well as quantum-confined systems.

Sensitized interfaces are very common in solar energy materials, since most semiconductors are not efficient light absorbers by themselves – their extinction coefficients are relatively low, and the energy gaps are usually large, corresponding to UV photons (>3 eV). The latter are scarce in the solar spectrum and transmit only a small fraction of the incoming solar energy. The common practice to improve the efficiency of solar harvesting is to utilize organic, inorganic, or combined sensitizers, such as small organic and metal–organic molecules or quantum dots, which have large extinction coefficients and energy gaps that correspond to the most abundant visible photons in the solar spectrum (e.g., 1.5–2 eV). The sensitizers are adsorbed at the semiconductor surfaces. When they interact with visible light, the photoexcited electrons are produced to have enough energy to transfer to lower-energy states localized on the substrate (conduction band of semiconductor). A photogenerated hole, initially localized on the sensitizer group, can be transferred elsewhere, e.g., to the counter electrode via a shuttle system (in Grätzel-type photovoltaic cells, such system is often represented by the iodide/triiodide pair I^-/I_3^- dissolved in a nonaqueous solvent). As an overall result, the energy of the solar photon is transformed into work performed to separate the charge carriers (electron and hole). There are two major pathways of further utilization of the charges: (a) to generate an electric current (photovoltaic applications) and later perform external work or (b) to utilize the energy of the separated charge carriers to drive chemical reactions (photocatalysis). In the latter case, the energy is eventually stored in the form of chemical bonds of generated fuel.

A deeper analysis of the operation schemes just outlined can reveal many factors that can interfere with the ability to gain maximal efficiency from the solar harvesting and transformation setup. This is because there are many unproductive processes the system may undergo, all of them competing with the productive ones. For example, when the sensitizer is excited, electrons may be transferred to solvent molecules, surface trap states, or just recombine with the nearby holes. When the electron is transferred to a substrate, it still has to move over a long distance and can be easily trapped in defect states present at interfaces and in the bulk material. When the electron is not far from the interface, it may also return to the adsorbed sensitizer (e.g., to recombine with the hole), or, again, it can be captured by solvent molecules. Thus, to gain a better understanding of the ways to improve the efficiency of solar energy materials (photovoltaic and photocatalytic), a detailed understanding of the underlying processes, their mechanisms, and kinetics must be obtained.

Over the past few decades, the methodology of nonadiabatic molecular dynamics has provided many useful insights into dynamics of charge

and energy transfer in solar energy materials [1]. The main groups of factors under investigation include (a) the nature of sensitizer and substrate, (b) the orientation and alignment of sensitizer, (c) the separation of donor and acceptor groups, and (d) the thermal motion and type of vibrational modes that may be responsible for charge and energy evolution. In addition, there are other factors of extremely high importance to electron transfer in nanomaterials. These include (a) surface structure, (b) quantum confinement, (c) intersystem crossing, and (d) electronic decoherence.

The Nature of Sensitizer and Substrate

The nature of sensitizer and substrate comes into play mostly via the energy level alignment of the potential donor and acceptor states. This is a standard criterion of thermodynamic feasibility of any process. In the zeroth-order approximation, it is enough to compare the potential energies. However, in many situations other, somewhat overlooked, factors must be taken into account – the entropic factors. The latter turn out to be key in the processes of singlet fission and multiple exciton generation in organic crystals and semiconducting quantum dots [10].

In singlet fission, a singlet electronic state undergoes transformation into a pair of coupled triplet states. This process is followed by decoherence that transforms the initial coherent mixture into a pair of uncorrelated triplet excitations. Since many more microstates exist for the noncoherent mixture compared to the coherent excitation, the process increases the entropy. The entropic gain turns out to be essential to driving singlet fission in many materials. Nonadiabatic molecular dynamics describes singlet fission at the atomistic level and naturally includes the entropic effect [11].

Another example demonstrating the role of the entropic effects is the thermally activated electron transfer at the alizarin/TiO_2 interface [12]. In the energy minimum, the donor energy levels of the alizarin are positioned inside the energy gap of TiO_2, rendering the electron transfer inefficient. However, once thermal fluctuations are taken into account, the situation changes qualitatively. Periodically, random fluctuations bring donor energy levels above the acceptor energy levels. During these moments, the electron transfer reaction can proceed efficiently, explaining the experimentally observed photo-properties of the system.

The nature of participating components also comes into play via the density of states, most commonly the density of acceptor states. Often, this factor correlates with the energy level alignment of the donor state (e.g., LUMO of a photosensitizer molecule) and the lowest energy level of acceptor states (e.g., conduction band minimum of a semiconductor substrate). The bigger this difference, the more likely it is that the density of the acceptor states at the donor energy level will be higher. This hypothesis was verified in various experimental and theoretical works using multiple combinations of substrate and sensitizer. High state density strongly affects rapid nonadiabatic dynamics [11, 12].

The Orientation and Alignment of Sensitizer

The relative configuration of the donor and acceptor sites becomes especially important for the interfacial charge transfer. The term implies four factors: (a) the type of chemical bonds with which the donor and acceptor sites are anchored, (b) the relative geometric orientation of the donor species and substrate, (c) the distance between the donor and acceptor sites, and (d) surface structure and its modification due to defects, solvent, etc.

Many works have shown that covalent bonds between the anchor group of a sensitizer and the substrate make a remarkable difference [1]. The bonds increase the couplings by several orders of magnitude, compared to physisorbed molecules, and lead to more efficient charge injection into the substrate. The extensive computational work tested a large number of anchor groups [13]. The authors ranked different groups and outlined the basic chemical principles for the design of efficient anchors. Specifically, the presence of conjugated bonds (double or triple) and anchor structure that favors localization of the LUMO orbitals into the substrate can strongly enhance the donor–acceptor coupling and increase the electron transfer rate. Certain heteroatoms, such

as N or P included in the structure of the anchor groups, were found to be helpful in this regard. Explicit time-domain nonadiabatic molecular dynamics studies confirm these findings [12].

Orientation of the sensitizer group with respect to the surface was found to be important to a lesser extent than the presence of covalent chemical bonds. Appropriate sensitizer orientation can increase the nonadiabatic coupling between donor and acceptor. For example, if the dye molecule contains a planar organic moiety, then parallel dye-surface orientation may be a natural way to increase the π–π or dative interactions. The orientation of donor and acceptor cites may be even more important in crystalline systems. Depending on the packing and the angle between planes of acene molecules in crystal structure, the nonadiabatic couplings can show periodic behavior as a function of displacement of one molecule with respect to the other [14]. This effect directly translates into triplet yield in singlet fission, for instance [11, 15].

The Separation of Donor and Acceptor Groups

The question of donor–acceptor separation attracted interest of many experimentalists and theorists. The main result on this front is the existence of approximately exponential dependence of the ET rates on the distance. Such dependence is characteristic to electron tunneling rates in arbitrary systems, as can be shown in a standard quantum-mechanical problem of a particle in a step potential. Delocalized electrons of conjugated or aromatic groups represent low tunneling barriers. On the contrary, single bonds create high barriers for electron tunneling between donor and acceptor species. Further, nonadiabatic couplings between states are proportional to orbital overlaps, which also decay exponentially with atomic separation.

Surface Structure

Surface structure and its modification may seem rather minor compared to other discussed factors, but it is not by far. In particular, the surface is the best place to find defects of all kinds – vacancies, kinks, steps, partially oxidized/reduced sites, adsorbates, etc. Discussing electron transfer, we are mostly concerned about the electronic structure of the conduction band of the substrate material. Surface defect sites create states that are lower in energy than the conduction band minimum of an ideal material. Unlike the delocalized conduction band orbitals, the trap states are localized. The low energy and strong localization of the trap states prevent the injected electrons from further transport. Traps can also attract holes and, hence, serve as electron-recombination centers.

By controlling the nature and properties of surface states, one can rationally favor or disfavor particular processes. For example, in a series of experiments, a set of CdS nanoparticles was investigated [16]. The initial stoichiometric CdS core structure was coated sequentially with the excess of either Cd or S atoms, which, respectively, resulted in quenching of photoluminescence or its reappearance. The nonadiabatic simulations combined with electronic structure computations provided explanation for this interesting result: S atoms, prone to directional, covalent bonding, created numerous defects in nonstoichiometric structures and the defects quenched fluorescence. On the contrary, reintroduction of Cd atoms healed the CdS surface, since metallic-type bonds of Cd atoms are much more flexible than covalent bonds of S. Lowering the surface density of defect states minimized nonradiative relaxation and led to reappearance of the photoluminescence signal. A separate study [17] investigated modification of semiconductor surface used in dye-sensitized solar cells with adsorbate molecules with bulky outer groups. The modification reduced transfer of the injected electron back to the dye and reduced its recombination with the oxidized electrolyte (shuttle). The overall effect of such surface modification was the increased efficiency of photon-to-energy conversion.

Quantum Confinement

Additional factors controlling the efficiency of solar energy harvesting may appear depending on the system. Nanoscale materials with so-called quantum confinement are of special interest. Quantum dots, carbon nanotubes,

graphene, and metal dichalcogenides are examples of systems with quantum confinement in zero-, one- and two-dimensions, respectively. Important for nonadiabatic dynamics, quantum confinement modifies electronic structure and electron–phonon coupling. The energy levels in quantum-confined systems are separated by notable energy gaps, and these gaps are usually larger for smaller systems (e.g., recall the energy spectrum in a particle-in-a-box problem). The nonradiative energy transformation from electronic excitations to nuclear vibrations is the major mechanism of energy dissipation and conversion into heat. The efficiency of this process depends on the type of vibrations existing in a system and on the electronic energy gap of the material used. The ratio of energy gap to the phonon energy determines the number of phonons that must be excited in a short period of time, for the energy to be lost into heat. The larger this number, the higher order is the ensuing process and the smaller is the probability for such process to occur. Based on this simple idea, the phenomenon of quantum bottleneck was proposed experimentally and later confirmed with the nonadiabatic molecular dynamics simulations [18]. The effect originates because of the special electronic structure (large, finite gaps) in such systems, as opposed to bulk materials, in which the electronic structure is organized into continuous bands.

Intersystem Crossing

Intersystem crossing is yet another important aspect of electron transfer in solar energy materials. Typical singlet electronic excitations are of the same spin symmetry as the original ground state, as well as numerous trap and defect states. Thus, electron–hole recombination and nonradiative energy decay proceed efficiently, reducing the yield of charge carrier separation. In photocatalytic processes, an ultrashort lifetime of the separated (free) charge carriers or excitations contradicts the necessity of having enough time span for these species to react with other molecules, the process that becomes rate limiting. As an example, nonadiabatic simulations demonstrated an ultrashort lifetime (ca. 100 fs) of hot holes in a GaN material used for water oxidation [19]. The energy of the initially generated photohole, favorably localized on OH groups, is quickly lost into heat in this case. Hence, the productive water oxidation process has negligible efficiency.

To overcome the above limitations, many scientists focus on generation and study of long-lived reactive species and charge carriers. Intersystem crossing from the initially created singlet state into long-lived triplets and higher multiplicity states provides a straightforward route to enhance the lifetime. Nonadiabatic molecular dynamics simulation involving intersystem crossing requires calculation of spin–orbit coupling, which should be combined with nonadiabatic coupling to account simultaneously for spin crossover and energy relaxation [20].

Electronic Decoherence

Finally, we want to emphasize the importance of quantum coherence in energy and electron transfer in nanoscale materials. Relaxation of singlet excited states into same-symmetry ground state requires on the order of hundreds picoseconds to over nanoseconds. However, it would have proceeded much faster, on a picosecond or sub-picosecond timescale, if the transitions had involved fully coherent states. Curiously, the FSSH method discussed in the theory section assumes a fully coherent description of electronic excited-state dynamics. Hence, it is unlikely that the FSSH would produce correct timescales, except for ultrafast processes. Decoherence effects should be taken into account in order to obtain correct long-time dynamics. To illustrate the importance of these effects, we can consider relaxation of low-lying excited states in solid pentacene and solid C_{60} [8]. The relaxation of hot electrons occurs on the order of 500 fs in the first system, and FSSH works perfectly well. On the contrary, the timescale of nonradiative relaxation in C_{60} is greatly underestimated by plain FSSH, while introducing a decoherence correction increases the timescale by an order of magnitude, in good agreement with experiment.

Departing from the methodological consequences of decoherence, we want to emphasize

two other points. First, since decoherence is largely defined by system's environment (although some degrees of freedom of the system can also be a part of that environment), the lifetimes of excitations or separated charge carriers can be controlled by a proper choice of the environment. This type of manipulations is known and utilized in quantum optics and quantum computing. Second, the idea of controlled decoherence is already utilized by nature. In this case, light-harvesting complexes are embedded into protein environments that are capable of maintaining long-lived coherences. As a result, the dynamics of energy transfer becomes very efficient, and a nearly 100 % yield is obtained for energy transfer from site to site over long spatial distances. Proper theoretical description of nonadiabatic electron dynamics in nanoscale materials requires accurate treatment of quantum coherence [5, 9, 18].

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Car–Parrinello Molecular Dynamics
- ▶ Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
- ▶ Electronic Transport in Carbon Nanomaterials
- ▶ Molecular Dynamics Simulations of Nanobiomaterials
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanostructures for Energy

References

1. Akimov, A.V., Neukirch, A.J., Prezhdo, O.V.: Theoretical insights into photoinduced charge transfer and catalysis at oxide interfaces. Chem. Rev. **113**, 4496–4565 (2013)
2. Akimov, A.V., Long, R., Prezhdo, O.V.: Coherence penalty functional: a simple method for adding decoherence in ehrenfest dynamics. J. Chem. Phys. **140**, 194107 (2014)
3. Tully, J.C.: Molecular dynamics with electronic transitions. J. Chem. Phys. **93**, 1061–1071 (1990)
4. Prezhdo, O.V.: Mean field approximation for the stochastic schrödinger equation. J. Chem. Phys. **111**, 8366 (1999)
5. Jaeger, H.M., Fischer, S., Prezhdo, O.V.: Decoherence-induced surface hopping. J. Chem. Phys. **137**, 22A545 (2012)
6. Wang, L., Trivedi, D., Prezhdo, O.V.: Global flux surface hopping approach for mixed quantum-classical dynamics. J. Chem. Theory Comput. **10**, 3598–3605 (2014)
7. Craig, C., Duncan, W., Prezhdo, O.: Trajectory surface hopping in the time-dependent kohn-sham approach for electron-nuclear dynamics. Phys. Rev. Lett. **95**, 163001 (2005)
8. Akimov, A.V., Prezhdo, O.V.: The PYXAID program for non-adiabatic molecular dynamics in condensed matter systems. J. Chem. Theory Comput. **9**, 4959–4972 (2013)
9. Akimov, A.V., Prezhdo, O.V.: Advanced capabilities of the PYXAID program: integration schemes, decoherence effects, multiexcitonic states, and field-matter interaction. J. Chem. Theory Comput. **10**, 789–804 (2014)
10. Smith, M.B., Michl, J.: Singlet fission. Chem. Rev. **110**, 6891–6936 (2010)
11. Akimov, A.V., Prezhdo, O.V.: Nonadiabatic dynamics of charge transfer and singlet fission at the pentacene/C_{60} interface. J. Am. Chem. Soc. **136**, 1599–1608 (2014)
12. Prezhdo, O.V., Duncan, W.R., Prezhdo, V.V.: Photoinduced electron dynamics at the chromophore–semiconductor Interface: a time-domain Ab initio perspective. Prog. Surf. Sci. **84**, 30–68 (2009)
13. Martsinovich, N., Troisi, A.: High-throughput computational screening of chromophores for dye-sensitized solar cells. J. Phys. Chem. C **115**, 11781–11792 (2011)
14. Coropceanu, V., Cornil, J., da Silva Filho, D.A., Olivier, Y., Silbey, R., Bredas, J.-L.: Charge transport in organic semiconductors. Chem. Rev. **107**, 926–952 (2007)
15. Wang, L., Olivier, Y., Prezhdo, O.V., Beljonne, D.: Maximizing singlet fission by intermolecular packing. J. Phys. Chem. Lett. **5**, 3345–3353 (2014)
16. Wei, H.H.-Y., Evans, C.M., Swartz, B.D., Neukirch, A.J., Young, J., Prezhdo, O.V., Krauss, T.D.: Colloidal semiconductor quantum dots with tunable surface composition. Nano Lett. **12**, 4465–4471 (2012)
17. Wang, C.-L., Lan, C.-M., Hong, S.-H., Wang, Y.-F., Pan, T.-Y., Chang, C.-W., Kuo, H.-H., Kuo, M.-Y., Diau, E.W.-G., Lin, C.-Y.: Enveloping porphyrins for efficient dye-sensitized solar cells. Energy Environ. Sci. **5**, 6933 (2012)
18. Kilina, S.V., Neukirch, A.J., Habenicht, B.F., Kilin, D. S., Prezhdo, O.V.: Quantum zeno effect rationalizes the phonon bottleneck in semiconductor quantum dots. Phys. Rev. Lett. **110**, 180404 (2013)
19. Akimov, A.V., Muckerman, J.T., Prezhdo, O.V.: Nonadiabatic dynamics of positive charge during photocatalytic water splitting on GaN(10–10) surface: charge localization governs splitting efficiency. J. Am. Chem. Soc. **135**, 8682–8691 (2013)

20. Habenicht, B.F., Prezhdo, O.V.: Ab initio time-domain study of the triplet state in a semiconducting carbon nanotube: intersystem crossing, phosphorescence time, and line width. J. Am. Chem. Soc. **134**, 15648–15651 (2012)

Theory of Optical Metamaterials

Carsten Rockstuhl, Christoph Menzel, Stefan Mühlig and Falk Lederer
Institute of Condensed Matter Theory and Solid State Optics, Abbe Center of Photonics, Friedrich-Schiller-Universität Jena, Jena, Germany

Synonyms

Composite materials; Effective media; Left-handed materials; Theory of artificial electromagnetic materials

Definition

Optical metamaterials can be understood as artificial media consisting of suitably assembled micro- or nanostructured materials such as dielectrics or metals and acquiring their essential properties from the structure rather than from the underlying material. The purpose of metamaterials is to affect the characteristics of light propagation in a manner inaccessible with natural available media. Their theoretical description evokes various concepts which are outlined in this essay.

Introduction

Optical metamaterials (MMs) can be understood as artificial media consisting of suitably assembled micro- or nanostructured materials such as dielectrics or metals [1–3] and acquiring their essential properties from the structure rather than from the underlying material. The purpose of MMs is to affect the characteristics of light propagation in a manner inaccessible with natural available media; hence the term material was linked with the prefix *meta*, being derived from Greek and literally meaning beyond or after. This rather generic and broad definition comprises many kinds of structured materials. Examples thereof would be photonic crystals or plasmonic structures [4, 5]. In a stricter sense, as the definition is further refined here, MMs are thought to be made from unit cells a few times or much smaller than the wavelengths in the spectral domain of interest. Here restriction is made specifically to the latter case where the propagating light does not probe for the fine details of the unit cells [6]. Its propagation can rather be fully understood by replacing the complex-shaped MM with a homogenous medium to which effective properties are assigned [7]. In the alternative case where the unit cell is only a few times less than the relevant wavelength, that is, the metamaterials represent a mesoscopic system, nonlocal effects come into play and the assignment of both an effective permittivity and permeability ceases to be meaningful. Alternative approaches for describing light propagation have to be used which are beyond the scope of this entry.

Depending on which approach is chosen to discuss MMs, the innovation comes with the deliberation from restrictions imposed by nature. The linear optical properties of natural materials at optical frequencies are entirely described by considering the permittivity function only, that is, a complex frequency-dependent function characterizing the response of matter to an electric field. The permeability function, which characterizes the response of matter to a magnetic field, can be safely disregarded since it marginally deviates from that of vacuum. For any frequency considered both permeability and permittivity represent material parameters and are linked to wave parameters, as for example, the wave vector or the refractive index, by the dispersion relation of the normal- or eigenmodes which is of pivotal importance to describe the propagation of electromagnetic fields in any structure. These eigenmodes are self-consistent solutions to Maxwell's equations in the respective medium. The propagation of an arbitrary field in this medium can then be written

as a linear superposition of such eigenmodes. The eigenmodes of a homogenous, dispersive achiral and isotropic medium are elliptically polarized plane waves characterized by wave parameters, such as impedance and refractive index. For natural materials both quantities are linked by inversion. Hence, if one parameter is determined in a given material, the other is equally fixed. Thus, independent of whether material or wave properties are considered, there is only a single degree of freedom [8, 9].

Metamaterials have the potential to circumvent this limitation and to provide independent control over permittivity and permeability, or refractive index and impedance. Both sets of parameters are connected by defining suitable constitutive relations. They have to be introduced into Maxwell's equations to specify the matter responses to an electric or magnetic field. Having materials at hand in which these material or wave parameters can be freely adjusted, many future applications can be envisaged which at first glance contradict the daily experience of how light propagates. Examples are the perfect lens which reconstructs an object in the image plane with unprecedented resolution or a cloaking device that allows concealing objects from external observers. It requires precise guidelines for designing MMs to achieve the necessary effective properties for bringing such applications to reality. This entry documents such approaches.

However, it has to be underlined again that any description of MMs in terms of effective material parameters is necessarily an approximation. It may serve as a model which is valid under certain constraints and where a finite number of free parameters mimic the complex structure. This greatly facilitates the description of current MMs and allows considering them in a subsequent design process for functional devices. The model is valid if the derived effective parameters are such that they do not violate the assumptions made in deriving them (for details, see Ref. [10]). A simple canonical example which illustrates the case is the retrieval of isotropic material properties for a MM whose unit cell fulfills all geometrical requirements to be considered as isotropic in the quasi-static (long wavelength) limit. Obviously, while being isotropic such material parameters must not depend on the angle of incidence. Unfortunately, this may be observed in the spectral domain where the dispersive effects of interest occur. Then, obviously, the material cannot be strictly considered as isotropic, though the deviations might be miniscule for suitably designed MMs and might be acceptable for a given application [10]. What is acceptable or not cannot be rigorously defined and it will depend on the explicit application and purpose of the effective properties, and it is moreover subject to a constantly heated debate. It is therefore not the intention of this entry to discuss the details of effective properties of various MMs but rather to explain and summarize the framework and the means to derive effective properties of optical MMs on theoretical grounds. Nevertheless, to illustrate representative unit cells for MMs and to provide an idea on how such structures look like, Fig. 1 shows examples of canonical unit cells for optical MMs along with an indication of which effective properties are affected.

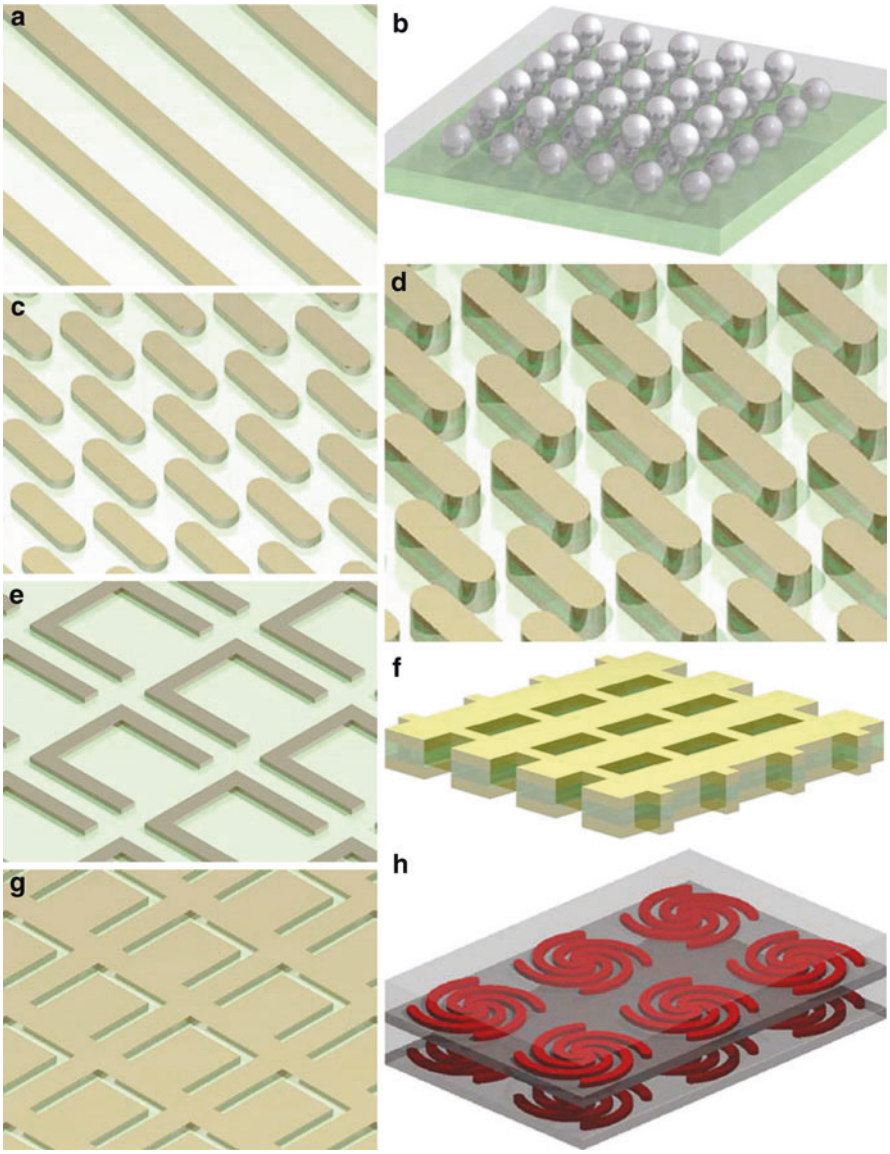
The entry is structured as follows. In the second section, the idea of assigning constitutive relations to a MM is outlined. In the third section, an overview of theoretical approaches is provided that can be used to discuss the properties of optical MM on theoretical grounds. The limitations of such descriptions are finally discussed and eventually some conclusions are provided.

Constitutive Relations and Effective Properties

Any theoretical description of optical MMs has necessarily to start with Maxwell's equations which read in time domain as

$$\mathbf{rot}\,\mathbf{E}(\mathbf{r},t) = \frac{\partial \mathbf{B}(\mathbf{r},t)}{\partial t}, \qquad \mathrm{div}\,\mathbf{D}(\mathbf{r},t) = 0,$$
$$\mathbf{rot}\,\mathbf{H}(\mathbf{r},t) = \frac{\partial \mathbf{D}(\mathbf{r},t)}{\partial t} + \mathbf{j}(\mathbf{r},t), \qquad \mathrm{div}\,\mathbf{B}(\mathbf{r},t) = 0,$$

where $\mathbf{E}(\mathbf{r},t)$ and $\mathbf{H}(\mathbf{r},t)$ are the electric and magnetic fields; $\mathbf{D}(\mathbf{r},t)$ and $\mathbf{B}(\mathbf{r},t)$ are the electric displacement and the magnetic induction,



Theory of Optical Metamaterials, Fig. 1 Characteristic optical MMs consisting of different unit cells. (**a**) Metallic wire grid which acts as a diluted metal with a Drude-type effective permittivity; (**b**) and (**c**) MMs made from small metallic nanoparticles which will induce a Lorentzian-type effective permittivity around the resonance frequency of the localized plasmon polariton; (**c**) indicates that the resonance frequency can be tuned by modifying the geometry of the nanoparticles; (**d**) cut-wire pair where two metallic particles separated by a dielectric are strongly coupled, leading to a symmetric and an anti-symmetric resonance where either the permittivity or the permeability exhibits a Lorentzian resonance; (**e**) split-ring resonator supports complex resonances depending on the illumination direction and the polarization, but being best known as a unit cell possessing a magnetic resonance; (**f**) fishnet structure where a cut-wire pair is combined with a metallic wire, and the resulting medium may show both a dielectric and a magnetic resonance in the same spectral range leading to a negative refractive index; (**g**) complementary structures, as the complementary split-ring resonator, are equally important; (**h**) gammadion structure which acts as an effective chiral medium

respectively; and $\mathbf{j}(\mathbf{r},t)$ is the conductive current. The latter depend both, in general, on the electric and magnetic field in a complicated manner, specified by the respective constitutive relations. It has to be stressed that in this special form of Maxwell's equations the external charge carrier density is assumed to be zero. Assuming the realm of linear optics to be valid, it is convenient to write Maxwell's equations in temporal Fourier space by assuming a time harmonic oscillation of all fields. This leads to

$$\begin{aligned} \mathbf{rot}\,\overline{\mathbf{E}}(\mathbf{r},\omega) &= \mathbf{i\omega}\overline{\mathbf{B}}(\mathbf{r},\omega), & \operatorname{div}\overline{\mathbf{D}}(\mathbf{r},\omega) &= 0, \\ \mathbf{rot}\,\overline{\mathbf{H}}(\mathbf{r},\omega) &= -\mathbf{i\omega}\overline{\mathbf{D}}(\mathbf{r},\omega) + \overline{\mathbf{j}}(\mathbf{r},\omega), & \operatorname{div}\overline{\mathbf{H}}(\mathbf{r},\omega) &= 0, \end{aligned}$$

Arbitrary solutions to Maxwell's equations in time domain are then written as a superposition of those time harmonic solutions weighted with appropriate amplitudes. Quantum mechanical or phenomenological models allow the derivation of frequency-dependent permittivity and permeability functions or tensors, where usually the electrical conductivity is absorbed into the complex permittivity which includes contributions from bound states and free electrons. Whereas the general dependency of $\mathbf{D}(\mathbf{E},\mathbf{H})$, $\mathbf{B}(\mathbf{E},\mathbf{H})$, and $\mathbf{j}(\mathbf{E},\mathbf{H})$ can be arbitrary complex, in most cases there is a linear relationship between these quantities mediated by permittivity, permeability, and electromagnetic coupling functions. The most general constitutive relation that is yet amenable to a theoretical treatment is that of a bi-anisotropic material. Assuming from now on an infinite homogenous space and dropping the dependency on the spatial coordinate, the resulting local constitutive relations read as [11] provided that the complex conductivity is absorbed into the complex permittivity:

$$\overline{\mathbf{D}}(\omega) = \varepsilon_0\overline{\overline{\varepsilon}}(\omega)\overline{\mathrm{E}}(\omega) + \frac{1}{c}\overline{\overline{\xi}}(\omega)\overline{\mathbf{H}}(\omega)$$

$$\overline{\mathbf{B}}(\omega) = \frac{1}{c}\overline{\overline{\zeta}}(\omega)\overline{\mathrm{E}}(\omega) + \mu_0\overline{\overline{\mu}}(\omega)\overline{\mathbf{H}}(\omega).$$

The four tensorial material functions are the permittivity $\overline{\overline{\varepsilon}}(\omega)$, the permeability $\overline{\overline{\mu}}(\omega)$, and two further tensors $\overline{\overline{\xi}}(\omega)$ and $\overline{\overline{\zeta}}(\omega)$ which mediate the electromagnetic coupling. Since the magnetic (electric) field depends on the curl of the electric (magnetic) field, the equations can also be understood as a Taylor series of a more complex functional dependency of the electric (magnetic) displacement on the electric (magnetic) fields. In this expansion only the lowest order coefficients are retained. For a medium with sufficiently high symmetry, that is, unit cells composed of three mirror-symmetric orthogonal planes, electromagnetic coupling vanishes. Consequently, in this Taylor expansion a next higher order term also needs to be considered, that is, media possessing a weak nonlocal response (weak spatial dispersion). If this holds, terms would appear where the displacement fields depend on the second derivative of the fields. Under appropriate conditions, which are assumed here to be valid, this term in the electric displacement field can be transformed as to appear in the dispersive permeability $\overline{\overline{\mu}}(\omega)$ in the magnetic constitutive relation and does not need to be considered explicitly. Such transformation can be understood as a gauge transformation since physical observable fields are not affected. Moreover, such transformation is useful to explain how it is possible to observe artificial magnetism at optical frequencies for which all available materials possess usually a nondispersive permeability. An artificial medium which supports a weak nonlocal electric response, that is, a nonlocal current appearing in mesoscopic metallic structures, is hence required to observe artificial magnetism, leading to the above constitutive relations [10, 12].

At this point it has to be stressed that the constitutive relations are very general. If they can be furthermore assumed to be reciprocal, they reduce to

$$\overline{\mathbf{D}}(\omega) = \varepsilon_0\overline{\overline{\varepsilon}}(\omega)\,\overline{\mathbf{E}}(\omega) + i\frac{1}{c}\overline{\overline{k}}(\omega)\overline{\mathbf{H}}(\omega),$$

$$\overline{\mathbf{B}}(\omega) = -i\frac{1}{c}\overline{\overline{\kappa}}^{T}(\omega)\overline{\mathbf{E}}(\omega) + \mu_0\overline{\overline{\mu}}(\omega)\overline{\mathbf{H}}(\omega).$$

If, additionally, electromagnetic coupling effects may be excluded, it is possible to arrive at the relation for anisotropic crystal.

$$\overline{\mathbf{D}}(\omega) = \varepsilon_0 \overline{\overline{\varepsilon}}(\omega)\, \overline{\mathbf{E}}(\omega), \quad \overline{\mathbf{B}}(\omega) = \mu_0 \overline{\overline{\mu}}(\omega) \overline{\mathbf{H}}(\omega).$$

If isotropy holds the simplest form of the constitutive relations is achieved as

$$\begin{aligned} \overline{\mathbf{D}}(\omega) &= \varepsilon_0 \varepsilon(\omega)\, \overline{\mathbf{E}}(\omega), \\ \overline{\mathbf{B}}(\omega) &= \mu_0 \mu(\omega) \overline{\mathbf{H}}(\omega) \end{aligned}$$

It has to be noted that the introduction of more complex constitutive relations, for example, explicitly containing strong spatial dispersion, when compared to the departing point is rather pointless, though they can be written down phenomenologically. The boundary conditions that need to be imposed onto the fields at the interface between two such media are quite involved (additional boundary conditions). They have to be applied in addition to ordinary Maxwell's boundary conditions and are the subject of ongoing research in the field of MMs [13].

These material parameters can be related to wave parameters describing the propagation of an eigenmode in a MM that is characterized by any of the constitutive relations above. Assuming the simplest scenario of a homogenous isotropic medium, the wave parameters of the eigenmodes are found by plugging the ansatz for an elliptically polarized plane wave $\mathbf{E}(\mathbf{r}, t) = \tilde{\mathbf{E}}(\omega) e^{i[k(\omega) r - \omega t]}$ as the possibly simplest solution into Maxwell's equations where the propagation vector $\mathbf{k}$ is one of the wave parameters. The dispersion relation relates it to the frequency ω which appears in the material parameters

$$\mathbf{k}^2 = k_x^2 + k_y^2 + k_z^2 = \frac{\omega^2}{c^2} \varepsilon(\omega) \mu(\omega)$$

The dispersion relation defines the set of plane waves that are allowed to propagate. The second wave parameter derives from the ratio of the moduli of electric and magnetic fields and is the impedance Z

$$Z = \sqrt{\frac{\mu(\omega)}{\varepsilon(\omega)}}$$

Whereas the propagation vector describes essentially the wave evolution of the normal mode in the respective medium, the impedance determines the reflection and transmission properties at interfaces to other media. For media with more complicated constitutive relations the corresponding dispersion relations can be equally derived on analytical grounds and can be found in literature.

With these governing equations the light propagation through an arbitrary complex material can be described exploiting a set of simple quantities. The challenge to theoretically describe optical MMs is to link a MM made of mesoscopic unit cells to a homogenized medium in which the fine details of the structures may be neglected. It is therefore required that at least the fields propagating either in the actual MM or in an effectively homogeneous material possess the same wave parameters. Both materials would then be indistinguishable since their optical action is fully preserved while the complexity is significantly reduced. This link has to be established first. By assuming furthermore feasible constitutive relations, these wave parameters can be linked to effective material parameters. Established approaches for exactly this purpose are documented in the following entry.

Retrieval of Effective Properties

The retrieval of effective properties of optical MMs is usually based on an analysis of the interaction of light with the structure under investigation; therefore, it is not possible to define the properties in an absolute sense but they are rather probed for a particular illumination scenario. In theory, the interaction and the propagation of such probing fields with the MM is numerically analyzed by solving rigorously Maxwell's equations and taking into account the fine details of the spatially distributed materials forming the unit cells, usually given by $\varepsilon(\mathbf{r}, \omega)$. The spatial distribution of the permeability does not need to be considered since the intrinsic permeability of natural materials is negligible. Various numerical means exist that can be applied for this purpose such as the Finite-Difference Time-Domain (FDTD) method, Finite-Element Method (FEM),

or plane wave expansion techniques [14]. Generally, depending on whether the unit cells are periodically or randomly arranged in space, various strategies can be applied to retrieve the effective properties. In the following between these two categories will be distinguished.

Periodic Structures

Most of current MMs are fabricated by top-down technologies such as electron beam lithography or direct laser beam writing. These technologies allow fabricating unit cells with critical dimensions in the order of a few tenths of nanometers, in most cases as planar nanostructured thin films on substrates. Repeating these techniques in a sequential or parallel process allows manufacturing MMs made of periodically arranged unit cells in all spatial dimensions. To eventually observe effective properties beyond those achievable by a mere spatial average of the intrinsic material properties of the involved constituents, resonances in the unit cells are often exploited. The required sub-wavelength size of the unit cells usually requires the coupling of the electromagnetic field to the free electrons in metals; hence plasmonic materials are often used in the design of optical MMs. In most cases gold is preferred since it is chemically stable, but also silver or aluminum can be used. Such MMs are therefore comparable to metallic photonic crystals, however, with much smaller periods, and consequently the language used to discuss their properties is intimately linked to it. Most notably, light propagation in bulk MMs is characterized by a dispersion relation as it will be discussed at first. The dispersion relation only allows determining the wave vector $\mathbf{k}$ that characterizes the evolution of a Bloch mode (for a definition see below), from which an effective index can be derived. As argued below, from the dispersion relation further quantities can be derived that are important to describe the propagation of finite beams (either spatially or temporally confined) through the bulk MM. From the tangential fields of an eigenmode the impedance linked to it can be derived. With these two quantities the entire set of wave parameters would be at hand. Moreover, for a deviating geometry and by relying on simulated reflection and transmission coefficients of an incident plane wave at a MM slab with finite thickness, both wave parameters can be equally retrieved. Such a procedure is part of a second succeeding section. Eventually, only if the wave parameters as retrieved by both approaches are identical the MM can be considered homogenous.

Retrieval Using the Dispersion Relation

An appropriate description of bulk properties of MMs made of periodically arranged unit cells is based on the dispersion relation. The periodic structure is characterized by the dielectric function

$$\varepsilon(\mathbf{r},\omega) = \varepsilon(\mathbf{r} + m\Lambda,\omega)$$

with Λ and m being the period of the structure and an integer, respectively. For simplicity it has been assumed that the MM's unit cells are made from isotropic materials. Since in the infinite medium two points displaced by one period in space are indistinguishable, the eigenmodes have to feature identical properties, that is, it is said they are Bloch periodic

$$\mathbf{u}_\mathrm{k}(\mathbf{r},\omega) = \mathbf{u}_\mathrm{k}(\mathbf{r} + m\Lambda,\omega)$$

The Bloch modes amplitude is periodic with the lattice and its phase evolves like a plane wave

$$\mathbf{E}_\mathrm{k}(\mathbf{r},\omega) = \mathbf{u}_\mathrm{k}(\mathbf{r},\omega)\mathrm{e}^{i[\mathrm{k}(\omega)\mathrm{r}-\omega t]}$$

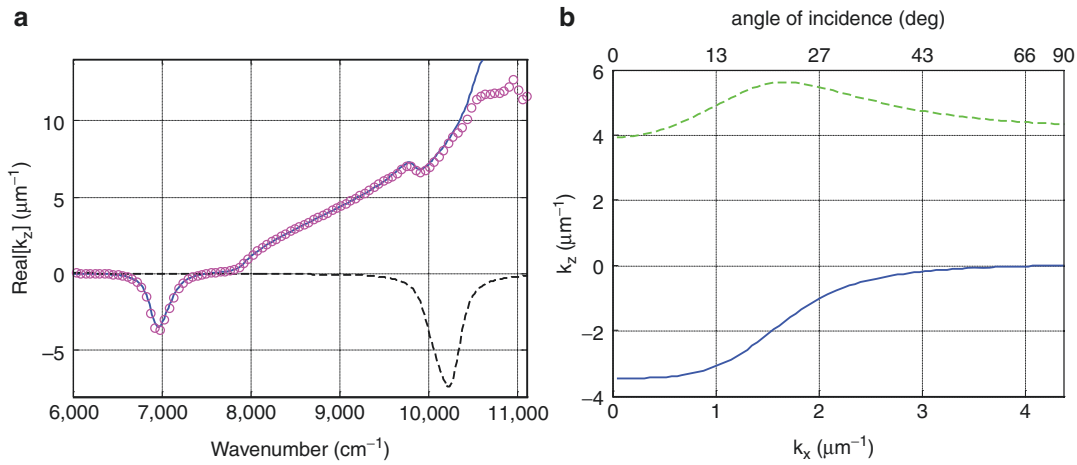
As for any normal mode the Bloch wave vector $\mathbf{k}$ and the frequency ω are interrelated, that is, they have to obey a specific dispersion relation. This dispersion relation to be solved requires in general numerical means. There are a discrete number of Bloch modes which can be excited in the bulk MM. Nevertheless, from a conceptual point of view the approach is identical to that of a homogeneous, dispersive medium, that is, the governing dispersion relation is a single equation with four interrelated parameters. It suggests that three of them can be chosen freely. Alternatively, fixing three parameters by external constraints only excites Bloch modes with a given fourth

parameter. In dealing with photonic crystals it is common to fix all components of the wave vector and then solve for the frequency at which the solutions are allowed to oscillate. However, a more rational approach is preferred here that is linked to an experiment. Therefore, the frequency ω (imposed by selecting a light source) and the tangential wave vector components at interfaces *kx* and *ky* are fixed. This is regarded as the referential plane where the evolution of the Bloch modes in the bulk starts (imposed by selecting the angle of incidence for an external plane wave since the tangential wave vector components are conserved). The dispersion relation is solved for the longitudinal wave vector component k_z. This allows to reveal by numerical means the entire functional dependency of k_z on all parameters and allows to disclose the entire dispersion relation $k_z = k_z(k_x, k_y, \omega)$ [15].

The number of modes that can be solved for is generally infinite. For an empty sub-wavelength lattice, however, only a single mode has a real valued propagation constant whereas all the others are evanescent. The same holds for a MM which is expected to be homogenizable, that is, there is a single mode with an imaginary part much less than those of all higher order modes. The imaginary part of this dominating mode is not strictly zero as soon as materials are incorporated into the unit cells that are intrinsically absorptive (e.g., metals). It furthermore has to be said that the eigenmodes appear always pairwise with opposite sign corresponding to forward and backward propagating modes, respectively. The proper branch has to be chosen in accordance with considerations imposed by the finiteness of the amplitudes, that is, by considering as the principal propagation the positive *z*-axis, only those eigenmodes can be excited which decay exponentially upon propagation. The corresponding eigenmode with the negative propagation constant has to be taken into account for opposite propagation direction. A MM supporting a backward wave is easily identified in such analysis by unraveling the parameter space where the imaginary part of the propagation constant is opposite to the real part. In this spectral domain the MM is said to have a negative index, since the index is defined as the length of the wave vector in the MM normalized to the length of the wave vector in vacuum.

From the dispersion relation of the modes in the center of the Brillouin zone, that is, $k_x = k_y = 0$, and while considering additionally the symmetry of the unit cell, many conclusions can be already drawn on the polarization properties of the MM. For example, a C_4 symmetric unit cell (identical under rotation by 90° with respect to the z-axis) leading to a pair of identical propagation constants will show a polarization independent response. The field of the two modes will then be dominated by a component either polarized in the *x*- or in the *y*-direction and an arbitrary superposition of both modes is an eigenmode, in particular, also circularly polarized light. If the two propagation constants are different while still being C_4 symmetric, the medium is chiral and the chirality constant κ can be extracted from the difference of both propagation constants. A complete discussion of such properties is beyond the scope of this entry but can be found in literature.

The question on whether the medium is isotropic or anisotropic (biaxial or arbitrary) can only be answered while disclosing the functional dependency of k_z on k_x and k_y at a fixed frequency of interest. As argued before, this iso-frequency curve has a fixed shape for a given medium, that is, it is spherical for an isotropic medium, ellipsoidal for a biaxial anisotropic medium, but in most cases it possesses a complexity beyond those simple shapes and the medium is anisotropic in the literal sense, that it is just said to be non-isotropic. An example of such a dispersion relation depending on the frequency for a fixed tangential wave vector and another one depending on a tangential wave vector for a fixed frequency are shown in Fig. 2 for the fishnet MM. As can be seen in Fig. 2, the fishnet MM is neither isotropic nor biaxial anisotropic, but the iso-frequency curve shows a much more complicated functional dependency on the frequency where the relevant dispersive effects occur, that is, where the effective index takes negative values for the real part. This has dramatic consequences on the ability of the structure to refract or diffract light. While the direction into which light is refracted is linked to the refractive index only for an isotropic material,



Theory of Optical Metamaterials, Fig. 2 Dispersion relation of a referential optical metamaterial, i.e., the fishnet as shown in Fig. 1f. (**a**) Shows the real part of the propagation constant of the two lowest order modes (*blue solid line* and *black dashed line*) as a function of the frequency at $k_x = k_y = 0$. The *dotted points* show for comparison results as retrieved from the finite structure which requires a retrieval using the complex reflection and transmission coefficient as discussed in the entry further below. (**b**) Shows the real (*blue solid line*) and the imaginary part (*green dashed line*) of the lowest order mode as a function of the tangential wavevector component k_x at $k_y = 0$. The frequency was fixed as to correspond to the spectral domain where the propagation constant is the largest negative, i.e., 7,000 cm^{-1}

it is given in such more complicated mesoscopic structures by the normal vector of the iso-frequency curve. This is an extension of a well-known concept for biaxial anisotropic materials, where the direction of the Poynting vector, that is, the direction into which energy flows or a bundle of rays is refracted, is no more parallel to the wave vector but normal of the iso-frequency curve for the respective wave vector.

Similarly, the next higher order derivative is linked to the diffraction coefficient, being a measure how strong a finite bundle spreads upon propagation (normal diffraction – positive curvature of the iso-frequency curve) or whether the opposite holds and a diffractively broadened beam can be re-focused while propagating through a suitably designed MM (anomalous diffraction – negative curvature of the iso-frequency curve). Such compensation of the normal diffraction of free space with a suitable anomalous diffraction explains how it is possible to use a MM slab as an imaging device.

The conclusion can be drawn that in order to describe the propagation of pulsed beams in an arbitrarily structured bulk medium it suffices to know the dispersion relation of the normal modes. Effects like diffraction and dispersive spreading have their physical origin in the different phase evolution in propagation direction for different frequencies and transverse wave vector components. A simple example for this statement is pulse propagation in optical fibers. There diffraction is arrested (only one transverse wave component) and the first term of the Taylor expansion of the dispersion relation provides the group velocity whereas the second one provides group velocity dispersion (dispersive spreading). And similarly, just as dispersion managed fiber (alternating segments of normal and anomalous group velocity dispersion) allows for a complete pulse restoration, a MM slab that compensates the diffraction in a homogeneous medium may act as a perfect lens that enables to restore a spatially confined input field in some output plane.

However, the dispersion holds only in an infinite bulk medium. It contains no information on the transition between different media (conventional medium-MM, MM-MM). The missing link is the field impedance which is a measure for the coupling strength of normal

modes propagating in different media. Various definitions with different complexity are feasible. Here a definition that is based on the coupling efficiency of an input to a homogenizable MM (field propagation may be described by only one Bloch mode) at a certain interface is used. This is the so-called tangential impedance which is defined as the ratio of the spatial average of the tangential electric to the tangential magnetic fields across a given interface. Assuming a mirror-symmetric unit cell the average of the x-component of the electric field vanishes and the MM impedance reads as (the superscript hints to a Bloch mode)

$$Z = \frac{\left\langle E_y^B \right\rangle}{\left\langle H_x^B \right\rangle}.$$

To sum up to only describe light transition between media and light propagation in an arbitrary, but homogenizable, MM it suffices to know the dispersion relation and the impedance of the normal modes, which represent so-called wave parameters. In general, it is impossible, but also not required, to assign effective material parameters to MMs. However, if this is desired the consistent validity of the respective constitutive relations is required. This aspect will be discussed in the succeeding section when the parameter retrieval is performed using the reflection and transmission coefficients of a plane wave at a MM slab. This method, known as NRW-method, is particularly popular since it allows for exploiting measured reflection/transmission data for the retrieval thus not requiring propagation simulations [16]. But, since both complex coefficients need to be known with high precision, most notably in the measured phase, the retrieval using solely experimental results still constitutes a major challenge.

Retrieval Using Reflection and Transmission Data from a Slab

The appealing aspect of the parameter retrieval using reflection/transmission data relies on its simplicity. For a MM characterized by two parameters (propagation constant and impedance), two independent sets of measurements are required. Since most MMs are fabricated as slabs with a finite thickness, their explicit consideration is attractive. To render such simple approach to be applicable, it has to be assumed that the MM does not generate depolarized light, that is, electromagnetic coupling should not occur, and the light propagating through the slab has to remain in the plane of incident while traversing the structure. If both conditions are valid, the complex reflection and transmission coefficients can be computed under the assumption that the field propagating in the MM is characterized by a single Bloch mode (homogenization), that has a characteristic dispersion relation, and the coupling from a surrounding media occurs equally only to that single eigenmode, that can be expressed using the tangential impedance. Then reflection and transmission coefficients of a MM slab (labeled by F) with thickness d, sandwiched between a substrate (labeled by S) and a cladding (labeled by C) material where the dispersion relation for isotropic media holds, read as

$$T = \frac{2k_S\xi A}{\xi(k_S + k_C)\cos(kd) - i\left(\xi^2 + k_S k_C\right)\sin(kd)},$$
$$R = \frac{\xi(k_S + k_C)\cos(kd) - i\left(\xi^2 + k_S k_C\right)\sin(kd)}{\xi(k_S + k_C)\cos(kd) - i\left(\xi^2 + k_S k_C\right)\sin(kd)}.$$

Here it has been assumed that the z-coordinate is normal to the slab and that only the x-component of the tangential wave vector is nonzero. Moreover, some abbreviations have been introduced to facilitate the notation, that is (for details, see Ref. [10]),

$$k_{S,C} = \alpha^{S,C}k_z^{S,C}, k = k_z^F, \xi = \alpha^F k_z^F, \alpha_{TE}^i = 1/\mu^i, \alpha_{TM}^i = 1/\varepsilon^i A_{TE} = 1, A_{TM} = \sqrt{\frac{\mu^C\varepsilon^S}{\varepsilon^C\mu^S}}$$

Here a TE and TM polarization is introduced where the electric field is normal to or in the plane of incidence, respectively. Here a restriction is imposed to the TE case. The equations can be uniquely inverted as to provide the desired wave vector component k in longitudinal direction

inside the MM slab and the tangential impedance ξ. These equations read as

$$k(k_x,\omega) = \pm\cos^{-1}\left\{\frac{k_z^S(1-R^2)+k_z^C T^2}{T\left[k_z^S(1-R)+k_z^C(1+R)\right]}\right\}+2m\pi,$$

$$\xi(k_x,\omega) = \pm\sqrt{\frac{\left(k_z^S\right)^2(1-R)^2-k_z^C T^2}{\left[(1+R)^2-T^2\right]}}$$

The different signs account for forward and backward propagation. They have to be chosen such that the waves decay upon propagation (imaginary part of k) and the tangential impedance takes positive values at normal incidence where it is linked to the impedance via $\xi_{TE} = \frac{\omega}{c}\frac{1}{Z}$ and $\xi_{TM} = Z\frac{\omega}{c}$. The branch number m occurring in the expression for the longitudinal wave number reflects that the field may exhibit a different number of nodes in a finite slab. The correct determination of the proper branch number can be achieved by various means. For example, it is possible to solve for the longitudinal wave number at marginally small frequencies where independent of the absolute thickness of the slab a field node does not appear because of the long wavelength. An alternative approach is to solve for the longitudinal wave for varying thickness. A consistent solution must not depend on the thickness. Another popular approach, the authors prefer too, is to solve the problem at the frequency of interest where all geometrical MM dimensions are scaled sufficiently down and the branch number can then be unambiguously fixed as to be $m = 0$. Up-scaling the geometry hitherto and requiring continuity of the longitudinal wave number while the scaling is performed requires the permanent adjustment of the branch number. The process is repeated until the geometry of interest is finally reached.

The wave parameters as retrieved by considering the reflection and transmission data are only comparable to those retrieved from the dispersion relation if the light propagation in the MM is dominated by a single, dominating Bloch mode and if the external illumination excites only this mode. If both conditions are met, the MM is said to be homogenizable since all the complex details are projected on two quantities, the wave number and the tangential impedance. If MMs are homogenizable their properties will be independent of the MM thickness. Since the tangential impedance is linked to the field distributions in a particular plane of the unit cell, it is furthermore required that the unit cell is symmetric for both propagation directions. Extensions of the parameter retrieval taking other properties into account, for example, electromagnetic coupling, are reported in literature and shall be not discussed here [17].

Crucial to an effective description of MMs is the fact that the tangential impedance and the longitudinal wave vector component will depend explicitly on the tangential wave vectors and the frequency. The frequency dependence is usually not an issue since any known material possesses a frequency dependent response, that is, it is said to be dispersive. To further reduce the complexity of the wave parameters and to remove the dependency on the tangential wave vectors, the wave parameters as retrieved can be linked in some cases to material parameters. This requires, however, the strict validity of the constitutive relations. For example, if the material is assumed to be isotropic, the scalar and only frequency dependent material properties are linked to the wave number and the impedance for the scenario above for TE polarization as

$$\mu(\omega) = \frac{k(k_x,\omega)}{\xi(k_x,\omega)} \text{ and } \varepsilon(\omega) = \frac{n^2(\omega)}{\mu(\omega)}$$

and for TM polarization as

$$\varepsilon(\omega) = \frac{k(k_x,\omega)}{\xi(k_x,\omega)} \text{ and } \mu(\omega) = \frac{n^2(\omega)}{\varepsilon(\omega)}$$

Here the effective index of the MM defined as $n^2(\omega) = \frac{k_x^2+k^2}{(\omega/c)^2}$ has been introduced.

Please note, the assumption on the validity of the isotropy enters this framework of equation by assuming that the iso-frequency curve derived from the dispersion relation is a circle in the 2D case and the index will not depend explicitly on the tangential wave vector. Moreover, the functional dependency of k and ξ on k_x is also

intimately linked and it will cancel out for an isotropic MM, that is, it will provide always the same material parameter independent of the tangential wave vector component. Of course, this does not necessarily need to be valid and deviations can be taken as a clear signature that the MM does not act as an isotropic material.

Similar procedures exist for biaxial materials, though then the response of the matter at two tangential wave vector components needs to be considered to disclose some selected entries of the material tensors. Other illumination scenarios are then furthermore required (i.e., modifying k_y or the considered polarization) to disclose all entries. Also, retrieval procedures that reveal the entire tensors taking into account also electromagnetic coupling effects exist in literature [16]. More complicated constitutive relations cannot be considered in most cases and in the case of contradictions, that is, different material properties are revealed while considering waves with different tangential wave vector components, a description in terms of material properties is not suggested and instead considerations should be restricted to the wave parameters.

Amorphous Structures

Whereas in the approach discussed in the previous section it was assumed that the unit cells are periodically arranged and a discussion in terms of Bloch modes could be performed, MMs can be also fabricated using bottom-up technologies that rely on self-organization mechanisms, for example, based on colloidal nanochemistry. Then, control over the geometry of the individual unit cell might be possible but not over the way they arrange. Likely, the structures will be randomly or amorphously arranged. For MMs of such geometry various mixing rules have been established in the past that allow to assign effective properties while considering only the scattering properties of the individual unit cells and their volume filling fraction. Such approaches are discussed in this section. Since the importance of such bottom-up MMs is still less when compared to MMs fabricated with top-down technologies, this section is kept rather concise to reflect its weight.

Retrieving Properties of Individual Meta-atoms

The properties of individual meta-atoms that eventually form a bulk MM are best accessed by investigating the scattered field generated by these meta-atoms. This scattered field can be expanded into multipoles which can be written either in a Cartesian or a spherical base. The idea of such an expansion is best understood in analogy to a Fourier series, where an arbitrary wave field can be written as a superposition of plane waves. Here the fields are decomposed into terms that are not spatially delocalized as plane waves, but they are linked to a definite origin where the scattered field emanates from. Since the multipoles are orthonormal functions, the expansion is unique and characteristic for a certain meta-atom. An expansion in spherical coordinates is well established in other fields of photonics, that is, in Mie theory. There the field scattered by a sphere is decomposed into spherical harmonics [18]. The expansion of a field scattered from an arbitrary object using such spherical harmonics reads as

$$\mathbf{E}_{\mathrm{Sca}}(r,\theta,\varphi) = \sum_{n=1}^{\infty}\sum_{m=-n}^{n} k^2 E_{nm}[a_{nm}\mathbf{N}_{nm}(r,\theta,\varphi) + b_{nm}\mathbf{M}_{nm}(r,\theta,\varphi)]$$

where k is the free space wave number of the surrounding medium, E_{nm} are some normalizing amplitudes, $\mathbf{N}_{nm}$ and $\mathbf{M}_{nm}$ are the vector spherical harmonics which are defined in literature, and a_{nm} and b_{nm} are the complex amplitudes at which they are excited to generate a particular scattered field. It has to be stressed that any scattered field can be expanded around a chosen origin in this way. Occasionally, the center of the coordinate system might be chosen inappropriate, leading to some ambiguity in the expansion. Therefore, the coordinate system should be appropriately chosen such that the smallest number of scattering coefficients attains nonzero values. In most practical cases this is either the geometrical center of the meta-atom or the center of gravity. Moreover, the vector spherical harmonics have a physical meaning. Roughly speaking, the vector spherical harmonics $\mathbf{N}_{nm}$ describe the scattered electric field in terms of contributions evoked by electric

multipole situated in the origin [19]. The first index n designates the respective Cartesian multipole. The electric field generated by their magnetic counterparts is associated to the coefficients $\mathbf{M}_{nm}$, respectively. Assembling the scattering properties of a single meta-atom is best performed by establishing the so-called *T*-matrix. The entries of this T-matrix suggest that illuminating the meta-atom with an incident field corresponding to a purely spherical harmonic of the order *n'm'* will generate scattered multipoles of all the orders with the index *nm* and their respective amplitude strengths are the entries of one column of the T-matrix.

The retrieval of the contributions to the scattered field generated by a given meta-atom is performed by computing its scattered field $\mathbf{E}(\mathbf{r}, \omega)$ with a rigorous method of choice (e.g., FDTD or FEM) upon illumination with a suitably chosen illumination and solving for the mode contents of the scattered field. Computing the overlap integral over a sphere at a certain radius a yields all coefficients of the multipole expansion

$$a_{nm} = \frac{\int_0^{2\pi}\int_0^{\pi} \mathbf{E}(r=a)\mathbf{N}^*_{nm}(r=a)\sin\theta d\theta d\varphi}{\int_0^{2\pi}\int_0^{\pi} \mathbf{N}_{nm}(r=a)\mathbf{N}^*_{nm}(r=a)\sin\theta d\theta d\varphi},$$

$$b_{nm} = \frac{\int_0^{2\pi}\int_0^{\pi} \mathbf{E}(r=a)\mathbf{M}^*_{nm}(r=a)\sin\theta d\theta d\varphi}{\int_0^{2\pi}\int_0^{\pi} \mathbf{M}_{nm}(r=a)\mathbf{M}^*_{nm}(r=a)\sin\theta d\theta d\varphi},$$

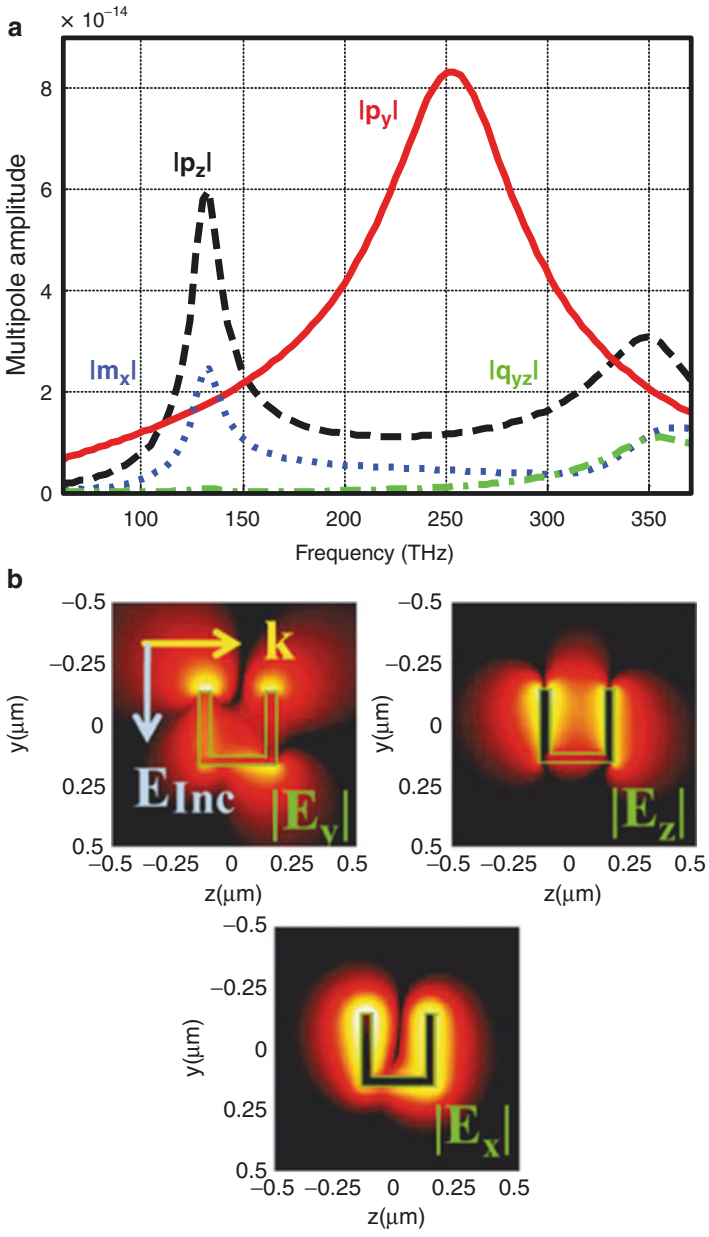
where the asterisks designate complex conjugated fields.

From these coefficients it can be seen whether the structure indeed causes a magnetic response, which likely will induce a resonance in the magnetic dipole coefficients linked to b_{1m}, or to an electric dipole coefficient linked to a_{1m}. It also allows revealing the possible contribution of higher order multipoles, the appearance of which prevents the assignment of effective material parameters. Nevertheless, with such an analysis all properties of interest can be evaluated.

An example where the scattered field of a split-ring resonator is expanded into multipoles at optical frequencies can be seen in Fig. 3. Since the symmetry of the split-ring resonator suggests the usage of a Cartesian coordinate system, the spherical multipoles are transferred into Cartesian ones. Only multipoles with significant nonzero amplitude are shown. It can be seen that the lowest order resonance is linked to a magnetic dipole oriented normal to the plane of the split-ring resonator and to an electric dipole oriented in this plane. The second order resonance is linked entirely to an electric dipole where currents oscillate in the two arms of the split-ring resonator in phase.

Mixing Rules

By assuming that the MM is finally composed of complex meta-atoms which generate upon the excitation by an external field secondary electric or magnetic dipole, effective properties can be assigned to such materials on the base of suitably chosen mixing rules [20]. Then, the meta-atoms attain the position of artificial polarizable entities and an averaging procedure can be applied, similar to the transition of microscopic to macroscopic Maxwell's equations, just with the only noticeable distinction that the transition is now made from mesoscopic to macroscopic Maxwell's equations where this mesoscopic system is now man-made. Many of such mixing rules exist and usually they have been developed for a particular MM geometry. For example, for MMs made of interleaved domains the Bruggemann formula was used; a MM made of a sequence of slabs requires for the introduction of an effective property either to take the geometrical mean of the intrinsic material properties (in the direction where the electric field is parallel to the slab, i.e., TE polarization) or the reciprocal of the geometrical mean of the reciprocal material properties (in the direction where the electric field is perpendicular to the slab, i.e., TM polarization). However, all these approaches represent approximations which have to be verified by suitable experiments. Thus, the application of



Theory of Optical Metamaterials, Fig. 3 (**a**) Amplitude of the dominant electromagnetic multipole contributions to the scattered field if a split ring is illuminated with a plane wave propagating in the plane of the split-ring resonator and the electric field is also polarized in the plane of the split-ring resonator, as shown in the *top* of (**b**). (**b**) Amplitude of the scattered electric field from an individual meta-atom in the y–z plane at 130 THz

particular mixing rules might be suitable but occasionally it might also be inappropriate. An exhaustive overview over such mixing rules can be found in the respective literature.

In the context of effective properties as discussed here where it is assumed that the MM is made of strongly scattering localized meta-atoms for which the scattered field can be expanded in terms of electromagnetic multipoles, the introduction of effective properties using the Clausius-Mossotti formula is most appropriate. In its simplest formulation it assumes that the MM consists of scattering entities characterized by both an electric and magnetic polarizability $\alpha_{E,M}$ which can be obtained by normalizing the electric and magnetic dipole moments to the dipole coefficients of the incident fields. By assuming that all meta-atoms are arranged in the same direction, the

T

Cartesian polarizabilities can be used to retrieve the particular entries of the respective material tensors which are valid by assuming the constitutive relations of an anisotropic crystal. If the surrounding medium is vacuum the effective parameters are given by

$$\varepsilon_{ii,\,\mathrm{Eff}}(\omega) = \frac{3 + 2V\alpha_{ii,\,\mathrm{E}}(\omega)}{3 - V_{\alpha ii,\,\mathrm{E}}(\omega)} \quad \text{and}$$
$$\mu_{ii,\,\mathrm{Eff}}(\omega) = \frac{3 + 2V\alpha_{ii,\,\mathrm{M}}(\omega)}{3 - V_{\alpha ii,\,\mathrm{M}}(\omega)}$$

Here V enters the equation as the volume filling fraction which is required to smaller, or even significantly smaller, than the percolation limit. For a completely statistical orientation of meta-atoms, an appropriate averaging procedure over the polarizability for all possible configurations has to be performed at first to extract the isotropic material properties emerging in this case. By this final step, effective material properties can eventually be assigned to a medium made from such strongly scattering entities.

However, the aspects last discussed here are not yet fully explored since the availability of suitable MMs made from strongly scattering elements is currently the subject of ongoing research. The approach discussed above has been extensively verified for MMs made from small metallic nanoparticles where all the assumptions imposed here are valid. But in any case the outlined procedure can be used to assign effective properties to such MM once sufficient information concerning their structure and scattering properties is available.

Conclusion

In conclusion, an introduction has been presented here that will facilitate the understanding of optical metamaterials. The entry is written from a perspective of a theoretician who may basically pursue two different aims. On the one hand he starts from a medium at hand that exhibits particular constitutive relations. Whereas for most optical materials available in nature, the complexity is limited by a dispersive tensorial permittivity, metamaterials may provide more complex relations which include dispersive tensorial permittivity and permeability as well as electromagnetic coupling. Then a genuine theoretical goal consists in the investigation of exciting implications for light propagation. Optics investigated under such assumption with the addition of a spatial dependent material is now frequently termed transformation optics and bears a great potential for innovative applications in optical systems.

On the other hand, at present it is feasible to fabricate almost arbitrarily nanostructured unit cells with an ever increasing complexity. The complex details tend to be important if light interacts resonantly with the unit cells, leading to unusual and extraordinary effects which allow for the control of light propagation in MMs formed from these unit cells. In this respect it is required that the transition from/to the environment and the propagation within the MM may be described by a simple set of wave parameters as the propagation vector and the impedance of the respective normal modes. Under certain conditions it might be possible to relate these parameters to effective material parameters the introduction of which facilitates appreciably the description of optical effect in MMs. In this case a very close connection to the first issue can be established. The complete understanding of the second issue is currently an important aim in theoretical optics.

The entry is actually devoted to shed some light on these issues and to provide a first guide of orientation to discuss them. Many more questions could likely be raised and many aspects necessarily had to be kept quite concise. Many details just mentioned can be found in the literature referred to here. The more detailed references that have been included hopefully will serve the valuable reader as a valuable source for further information beyond the basic ideas discussed in here.

Cross-References

► Active Plasmonic Devices
► Light Localization for Nano-optical Devices

- ▶ Nanostructures for Photonics
- ▶ Optical Properties of Metal Nanoparticles
- ▶ Optical Techniques for Nanostructure Characterization

References

1. Engheta, N., Ziolkowski, R.W.: Electromagnetic Metamaterials: Physics and Engineering Explorations. Wiley-IEEE Press, Piscataway (2006)
2. Cai, W., Shalaev, S.: Optical Metamaterials: Fundamentals and Applications. Springer, New York (2009)
3. Marqués, R., Martín, F., Sorolla, M.: Metamaterials with Negative Parameters: Theory, Design and Microwave Applications. Wiley, Hoboken (2008)
4. Sakoda, K.: Optical Properties of Photonic Crystals. Springer, Berlin (2004)
5. Maier, S.A.: Plasmonics: Fundamentals and Applications. Springer, New York (2010)
6. Sihvola, A.: Metamaterials in electromagnetics. Metamaterials **1**, 2–11 (2007)
7. Simovski, C.A.: On electromagnetic characterization and homogenization of nanostructured metamaterials. J. Opt. **13**, 013001 (2010)
8. Jackson, J.D.: Classical Electrodynamics. Wiley, New York (1998)
9. Born, M., Wolf, E.: Principles of Optics. Cambridge University Press, Cambridge (1999)
10. Menzel, C., Paul, T., Rockstuhl, C., Pertsch, T., Tretyakov, S., Lederer, F.: Validity of effective material parameters for optical fishnet metamaterials. Phys. Rev. B **81**, 035320 (2010)
11. Tretyakov, S.: Analytical Modeling in Applied Electromagnetics. Artech House, London (2003)
12. Serdyukov, A., Semchenko, I., Tretyakov, S., Sihvola, A.: Electromagnetics of Bi-anisotropic Materials–Theory and Applications. Gordon and Breach, Amsterdam (2001)
13. Agranovich, V.M., Yudson, V.I.: Boundary conditions in media with spatial dispersion. Opt. Commun. **7**, 121–124 (1973)
14. Taflove, A., Hagness, S.C.: Computational Electrodynamics: The Finite-Difference Time-Domain Method. Artech House, Norwood (2005)
15. Rockstuhl, C., Menzel, C., Paul, T., Pertsch, T., Lederer, F.: Light propagation in a fishnet metamaterial. Phys. Rev. B **78**, 155102 (2008)
16. Morits, D., Simovski, C.: Electromagnetic characterization of planar and bulk metamaterials: a theoretical study. Phys. Rev. B **82**, 165114 (2010)
17. Kwon, D.-H., Werner, D.H., Kildishev, A.V., Shalaev, V.M.: Material parameter retrieval procedure for general bi-isotropic metamaterials and its application to optical chiral negative-index metamaterial design. Opt. Express **16**, 11822–11829 (2008)
18. van de Hulst, H.C.: Light Scattering by Small Particles. Dover, New York (1981)
19. Raab, R.E., De Lange, O.L.: Multipole Theory in Electromagnetism. Oxford University Press, New York (2005)
20. Sihvola, A.E.: Electromagnetic Mixing Formulas and Applications. The Institution of Electrical Engineers, London (1999)

Thermal Actuators

Joseph J. Brown and Victor M. Bright
Department of Mechanical Engineering, University of Colorado, Boulder, CO, USA

Synonyms

Electrothermal actuators; Electrothermomechanical actuators; Heatuators; Thermomechanical actuators

Definition

Thermal actuators are mechanical systems that use the thermally induced expansion and contraction of materials as a mechanism for the creation of motion. These devices are compliant structures, using elastic deformation and mechanical constraints, that frequently are designed to amplify the motion generated by thermal expansion or contraction. Temperature changes that result in thermal actuation are most commonly provided by environmental changes or by Joule heating from electrical current flow. In the context of nanotechnology, thermal actuators refer to microscale and nanoscale devices used to mechanically interact with nanoscale structures, with motion generated by the thermally induced expansion and contraction of materials.

Key Principles, Concepts, and Phenomena

Thermal actuators are useful for applications where low voltage, small footprint, and high force output are desirable device characteristics.

In the context of microscale and nanoscale engineering, thermal actuators can be used as microelectromechanical systems (MEMS) capable of translating electrical signals into controllable displacements with resolution smaller than 1 μm. Displacement for thermal actuators can be designed as small as several nanometers [1]. For the discussions below, it should be noted that simple calculations discussed in this essay have been provided in order to enable quick initial evaluation of possible actuator designs. For more accurate understanding and operation of these actuators, multiphysics simulations and calculations, and experimental characterization of constructed devices are recommended.

Surface Contact and Connections in MEMS/NEMS Design

At size scales of hundreds of microns or smaller, contacting surfaces can have significant friction and stiction capable of inhibiting device motion. Stiction is the phenomenon where static surfaces contact each other, preventing motion and separation of the surfaces. Stiction may be caused by electrostatic and van der Waals forces. This can cause devices to pull into contact with underlying surfaces and substrates. In contact, these surfaces experience friction, the resistance of contacted surfaces to sliding motion, and can be strong enough to impede all motion. In pinned joints, contact surfaces are smaller, but friction can lead to significant wear, resulting in eventual failure of a device. For design of solid microscale and nanoscale actuators, contact surfaces are minimized to only those surfaces where contact is absolutely necessary for device function. For maximum reliability, actuators themselves are typically designed as suspended structures with clamped end conditions. The fabrication of pinned beam connections is typically imprecise using common microfabrication techniques. Additionally, they present regions subject to wear and friction which may decrease the reliability of actuator motion.

Compliant Structures

Compliant structures are arrangements of elastic bodies with constrained joints that generate or transmit motion. Because pinned joints are not optimal for microscale and most nanoscale thermal actuator designs, these designs employ elastic deformation to allow motion. As such, thermal actuators are a subset of compliant mechanisms used for the generation of motion. Compliant structures and compliant mechanisms are a significant area of study and are frequently used in conjunction with thermal actuators to achieve additional displacement amplification or other effects such as motion sensing, thermal management of moving parts, or motion stabilization [2–11]. Repeatable operation of the thermal actuator requires operation within the elastic limits of the actuator material. Excessive heating and mechanical strain can ultimately lead to failure of the thermal actuator mechanism. In particular, it is possible to "burn out" the thermal actuator due to excessive input power, which leads to creep, softening, and eventual melting of the actuator material.

Thermal Expansion

Thermal expansion and contraction of materials is the fundamental mechanism by which motion is generated in thermal actuators. Thermal expansion in solids originates from the nature of atomic or molecular bonding within solids. A rise in temperature increases the bond lengths between atoms or molecules in a solid due to increased energy of the populations of phonons and electrons within the solid.

Thermal expansion can be treated as a continuum behavior in classical solid mechanics. The local three-dimensional strain tensor ε_{ij} is given by Eq. 1 for isotropic materials, where E is the Young's Modulus, ν is the Poisson's ratio, α is the linear thermal expansion coefficient, σ_{ij} is the local stress tensor, σ_{kk} is the sum of the three normal stresses ($\sigma_{kk} = \sigma_{11} + \sigma_{22} + \sigma_{33}$), the term $(T - T_0)$ defines the temperature rise above a reference temperature T_0, and δ_{ij} is the Kronecker delta [12]. In quasi-steady-state behavior, the actuator establishes a thermal equilibrium with its environment, which develops a constant temperature profile throughout the actuator. The corresponding thermal strain distribution in the

actuator drives motion to a mechanical equilibrium.

$$\varepsilon_{ij} = \frac{1+v}{E}\sigma_{ij} - \frac{v}{E}\sigma_{kk}\delta_{ij} + \alpha(T - T_0)\delta_{ij} \quad (1)$$

Joule Heating and Power Dissipation

The flow of electricity through an electrically resistive material dissipates power according to Ohm's law. Microscopically, the local density of power dissipation $\dot{q}$ is given by the local resistivity ρ, current density $\vec{J}$, and electric field $\vec{E}$ according to Eq. 2.

$$\dot{q} = \vec{J} \cdot \vec{E} = \frac{1}{\rho}\left|\vec{E}\right|^2 = \rho\left|\vec{J}\right|^2 \quad (2)$$

Equation 2 is valid for DC currents and the instantaneous power dissipation from time-varying currents and electric fields. For average power dissipation from time-varying fields, the instantaneous power dissipation is averaged in time t', according to Eq. 3a, where t_0' may, for instance, be defined as the length of time for one period of oscillation. In practice, the result is Eq. 3b for harmonic, time-dependent electric fields and currents [13]. Note that although most thermal actuators rely on heating due to controlled electrical currents, other heating mechanisms exist such as absorbance of electromagnetic waves which may create similar thermal responses.

$$\dot{q}_{\text{ave}} = \frac{1}{t_0'}\int_0^{t_0'} \dot{q}dt' = \frac{1}{t_0'}\int_0^{t_0'} \frac{1}{\rho}\left|\vec{E}\right|^2 dt'$$

$$= \frac{1}{t_0'}\int_0^{t_0'} \rho\left|\vec{J}\right|^2 dt' \quad (3a)$$

$$\dot{q}_\omega = \frac{1}{2\rho}\left|\vec{E}\right|^2 = \frac{1}{2}\rho\left|\vec{J}\right|^2 \quad (3b)$$

Integrating the local power density $\dot{q}$ through the volume of the actuator gives the lumped form of Ohm's law, Eq. 4, where total power dissipation P is readily calculated from measurement of the current flow I, voltage drop V, and total electrical resistance R in the actuator.

$$P = IV = \frac{V^2}{R} = I^2R \quad (4)$$

Frequently the thermal actuator is constructed from materials whose resistivity changes with temperature and strain (i.e., due to piezoresistive effects), such as silicon. In this case, the local resistivity ρ cannot be assumed to be constant for all output displacements and levels of mechanical loading, nor can the total device resistance R. For this reason, power dissipation in thermal actuators should be derived from measurements of current and voltage without assuming a constant resistivity or resistance. Control based on current or voltage alone rather than total power dissipation should be validated experimentally before reliance on these simpler mechanisms to control the thermal actuator mechanical output.

Heat Transfer: Fundamental Concepts

The temperature rise in the actuator is determined by the heat generation and heat transfer throughout the thermal actuator. This behavior is governed by Eq. 5, the heat equation for stationary materials, and Eq. 6, the Fourier heat law [14]. The heat equation is properly written with the full time derivative, $\frac{DT}{Dt'} = \frac{\partial T}{\partial t'} + \vec{v} \cdot \nabla T$, where $\vec{v}$ is the local velocity and t' is time. From a macroscopic view there is no convective mass transport within the solid structures, therefore the velocity components $\vec{v}$ in the full derivative are reduced to zero, and the full derivative reduces to the partial time derivative.

Understanding of the temperature profile of the thermal actuator is developed by solving Eqs. 5 and 6 for the boundary conditions appropriate to specific thermal actuator designs. The local heat generation $\dot{q}$ was defined in Eq. 2 for Joule heating. The average temperature rise ($\Delta T = T - T_a$) in a beam element can be found using Eq. 7.

$$\eta c\frac{\partial T}{\partial t'} = k_\theta \nabla^2 T + \dot{q} \quad (5)$$

$$\vec{q} = -k_\theta \nabla T \quad (6)$$

T

$$\Delta T = \frac{1}{L_i}\int_0^{L_i} T dx' - T_a \tag{7}$$

T Temperature (K)
T_a Initial temperature (K)
ΔT Average temperature change (K)
k_θ Thermal conductivity (W/m-K)
$\vec{q}$ Heat flow vector (W/m^2)
H Density (kg/m^3)
C Heat capacity (J/kg-K)
x' Local distance coordinate along the length of the beam (m)
L_i Length of beam element over which the average is calculated (m)

Heat Transfer: Environment Effects on Actuation

By modifying the boundary conditions for heat transfer from the thermal actuator, the ambient environment surrounding a thermal actuator affects its output motion. Heat transfer mechanisms modify the temperature profile and corresponding thermal strain of the thermal actuator. In vacuum, heat transfer from the thermal actuator, for MEMS dimensions, occurs primarily through conduction through the actuator clamps. Radiative heat transfer is a secondary mechanism for heat loss from the thermal actuator, with energy transferred from the actuator to its substrate and surrounding environment. For initial design considerations, analysis of conductive heat transfer in the solid actuator material in vacuum allows estimation of the order of magnitude of actuator power requirements and temperatures, and is one step to understanding of the thermal behavior of the actuator.

In nonvacuum environments, heat transfer to a surrounding fluid can significantly alter the temperature profile of the thermal actuator. Because of the small dimensions of most thermal actuators, conduction usually dominates over convection in air, and published models generally only account for conduction through air [15–17]. Operation of thermal actuators in environments such as air may require significantly more power than used in a vacuum environment in order to obtain the same output displacement [15]. As an example, 1.4 mA applied to one design in vacuum may lead to the same temperature rise as 5.0 mA applied to the same device in air [15]. This result implies that more than 12 times the power consumption may be needed in air as in vacuum in order to achieve similar output from some thermal actuator designs.

A lower bound on the time constant for response of the thermal actuating system in vacuum can be found by considering a lumped model of the heat capacity and the thermal resistance. Equation 8 demonstrates applications of such a lumped model to a suspended beam with heat flow out the ends of the beam. The time constant τ_θ is calculated from the lumped heat capacity C and thermal resistance R_θ of the actuator beam that experiences the maximum current density. From this time constant calculation it may be expected that the thermal actuator will not move faster than a frequency of f_θ.

$$R_\theta = \frac{L_i}{k_\theta A_i} \tag{8a}$$

$$C = c\eta L_i A_i \tag{8b}$$

$$\tau_\theta = R_\theta C \tag{8c}$$

$$f_\theta = \frac{1}{2\pi\tau_\theta} = \frac{1}{2\pi R_\theta C} \tag{8d}$$

A_i Beam cross-section area (m^2)
f_θ Thermal bandwidth maximum frequency (Hz)
τ_θ Thermal time constant (s)
C Lumped heat capacity (J/K)
R_θ Lumped thermal resistance of a beam (K/W)

Statically Indeterminate Structures

The output motion of thermal actuators derives from statically indeterminate structures, typically beams, subjected to mechanical constraints. Design of these constraints can determine whether motion will occur in-plane or out-of-plane. A number of texts are available that further discuss the concepts of indeterminate structures [18]. Several detailed models have been developed for the output motion of constrained beams within thermal actuators [4, 19]. Some

of these analyses will be examined in the examples below.

Displacement Output and Force Output

Experimental measurement of an output displacement curve recorded versus input power, voltage, or current can provide a fit curve that allows prediction of the freely moving thermal actuator displacement d_0 without requiring detailed analysis of the thermomechanical system. When the thermal actuator is used to provide an output force, this force may be estimated using the spring constant k calculated for the beams of the thermal actuator. The output force F provides a perturbation δd to the free-moving displacement curve, giving actual displacement d. Force is estimated Eq. 9 according to Hooke's law applied to this perturbation from the free-moving displacement d_0 [5]. Simulations indicate that for some device mechanisms the output force may differ significantly from these predicted values in cases when displacement is constrained to nearly zero [20]. Experimental measurement of this effect and calibration of the output forces of thermal actuators remains an ongoing active area of research.

$$F = k\delta d = k\,(d_0 - d) \tag{9}$$

The spring constant k_i of many individual thermal actuator beams can be estimated Eq. 10 according to the solid mechanics of a beam with one end clamped and one end "guided" (prevented from rotation but allowed to displace in one direction), where L_i is the beam length and I_z is its moment of inertia. For beams with width w and thickness t, I_z is given by Eq. 11, and k_i is estimated with Eq. 12. The total spring constant is found Eq. 13 by multiplying the number of beams N by the spring constant k_i for each beam.

$$k_i = \frac{12EI_z}{L_i^3} \tag{10}$$

$$I_z = \frac{tw^3}{12} \tag{11}$$

$$k_i = \frac{Etw^3}{L_i^3} \tag{12}$$

$$k = Nk_i \tag{13}$$

Using k and the mass m calculated based on the density η and volume of the whole actuator assembly (e.g., $m = N\eta L_i wt$), a lower bound to the natural frequency f_0 can be found Eq. 14.

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{14}$$

For operation in laboratory environments, it is important that the natural frequency not be too low, otherwise the actuator may be significantly perturbed by environmentally induced vibrations. A good rule of thumb is $f_0 > 20$ kHz, which implies attenuation of most sounds and other laboratory disturbances.

Applications

Thermal actuators are one of the key mechanisms for effecting motion in microscale and nanoscale solid structures. Other common microelectromechanical system (MEMS) mechanisms typically use electrostatic forces, hydrostatic forces, magnetic forces, or piezoactuation [16]. The advantages of thermal actuators over these other mechanisms lie in low operating voltage, high output force, and compact footprint. Disadvantages include comparatively high power consumption, comparatively slow response time, and dependence of actuation output on ambient environment.

Because thermal actuators dissipate heat, thermal management of the environment surrounding the actuator may be an important design consideration. The substrate supporting a thermal actuator can experience temperature changes due to heating from the actuator. In vacuum the substrate to which the actuator is anchored becomes the primary mechanism for dissipation of heat from the actuator. Material selection for the substrate should allow for temperature changes due to

heating and adequate thermal engineering to prevent overheating. Furthermore, placement of temperature-sensitive materials near the actuator should be carefully planned.

Thermal actuators also face high temperature and low temperature limits to operation. At high temperatures, material creep and failure may occur. At low temperatures such as cryogenic environments, heating of the environment due to the actuator may disrupt the environment. Furthermore, some materials such as semiconductors may be insufficiently conductive to allow much current flow. Other materials such as many metals may develop superconductivity at cryogenic temperatures, which leads to insufficient power dissipation to create significant temperature changes for thermal expansion and actuation.

In the examples discussed below, the V-type actuator design has an advantage over the hot-arm, U-type design in that all beams contribute to actuation motion in the V-type design. In the U-type design, motion is generated by the hot arms but resisted by the cold arms. As a result, a V-type design requires a smaller footprint to achieve the same output as a given U-type design. Furthermore, the inherently linear motion output of the V-type designs makes them easier to incorporate in device designs.

History

The hot-arm, U-type actuators were originally demonstrated in the early 1990s and quickly became a staple for generation of mechanical motion in solid MEMS structures. In the late 1990s, V-type actuators were initially demonstrated, and this design has been subject to much further development in the subsequent years. In particular, V-type actuators have been explored for driving compliant mechanisms, microgripper devices, and micropositioning devices. Bimorph thermal actuators (Example 3 below) have been in use for submillimeter-scale applications since the late 1980s. Extensive review of the history of microfabricated thermal actuators may be found in Ref. [16].

By the early 2000s, the fundamental approaches to thermal actuators, discussed below as examples, had become well-established actuation mechanisms. Designs have advanced at a fundamental level in approaches to thermal management and the application of computational tools in optimization of designs [7, 8, 21].

Applications for thermal actuators have been demonstrated in many areas. These include microscale and nanoscale tensile testing; microhandling, nanopositioning, and nanomanipulation; optical and electrical switching, and micromirror positioning; microrobot locomotion; and additional compliant designs [4, 5, 9–11, 16, 20, 22–28]. Thermal actuators were seen in many reports as the drivers of mechanical motion in many geared and pivoting MEMS systems, particularly in the time period of 1994–2002. Such systems are fundamentally limited by frictional wear, and consequently they have not been the subject of much development in more recent years. Newer applications focus more on structures for manipulating and interacting with objects at the microscale and nanoscale.

In the 2010s, the implementation of new materials in microfabrication processing resulted in a diversification of thermal actuators into many material systems. These include the following: conducting polymers and polymer composites; graphene, carbon nanotubes, and other fullerenes; nanocrystalline diamond; Al-W bimetallic systems; and Ni metal [29–33]. Furthermore, the availability of an expanded range of materials in fabrication has enabled both the computational optimization of thermal actuators based on composite materials and a new approach to thermal actuation [34–36]. In this new approach, material membranes are subjected to constraints on biaxial strain and therefore move in an out-of-plane direction as the constituent materials expand upon heating. One demonstration used conducting membranes that heated themselves [35]. A second demonstration used high thermal expansion polymers supported on microfabricated heaters [36].

Examples

In application, three main mechanisms have been implemented as the most common thermal actuators: V-type, hot arm/U-type, and bimorph thermal

Thermal Actuators, Table 1 Some material properties of PolyMUMPS polycrystalline silicon, approximated as an isotropic material [4, 5, 15, 38]

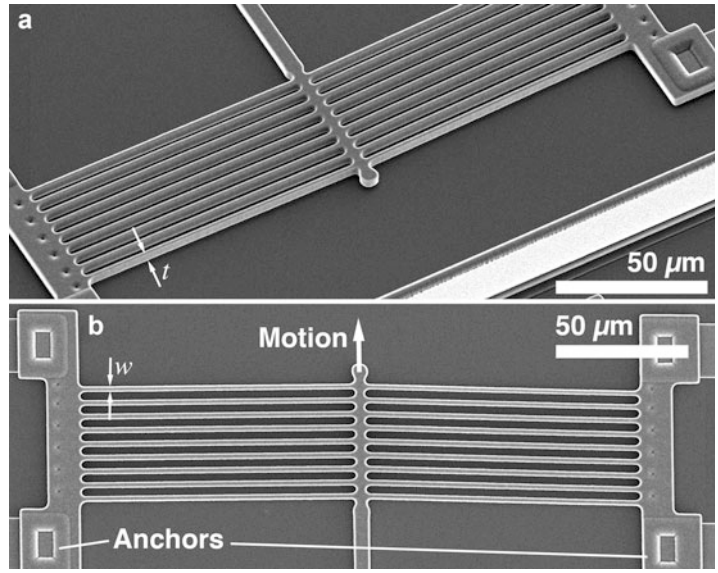
Property	Symbol	Value	Units
Electrical resistivity			
Constant	ρ	1.7×10^{-5} to 4×10^{-4}	Ω-m
Temperature dependent	$\rho = \rho_0[1 + \xi(T - T_0)]$		
		$\rho_0 = 34 \times 10^{-6}$	Ω-m
		$\xi = 1.25 \times 10^{-3}$	K^{-1}
		$T_0 = 293$	K
Thermal conductivity			
Constant value, polycrystalline Si	k_θ	29–34	W/m-K
Constant value, single crystal Si	k_θ	150	W/m-K
Density	η	2.33×10^3	kg/m^3
Heat capacity			
Constant value	c	705	J/kg-K
Temperature dependent	$c = 1.976 \times 10^{-6}T^3 - 3.767 \times 10^{-3}T^2 + 2.623T + 215.0$		J/kg-K
Young's modulus	E	162 ± 14	GPa
Poisson's ratio	v	0.29	
Linear coefficient of thermal expansion			
Constant value (at 293 K)	α	2.7×10^{-6}	K^{-1}
Temperature dependent	$\alpha = \left[3.725\left(1 - e^{-5.88\times10^{-3}(T-125)}\right) + 5.548 \times 10^{-4}T\right] \times 10^{-6}$		K^{-1}

actuators. Calculations are provided with these examples in order to walk through the basic design of common thermal actuators. These calculations are intended to help the designer to understand the range of actuator performance and the effect of device geometry on the output behavior. Final design should use multiphysics simulations and experimental characterization of fabricated devices in order to understand the real performance of any particular design.

In order to facilitate design calculations, Table 1 provides some published values on the properties of polycrystalline silicon (or "polysilicon"), a common material for thermal actuators. Because it is a semiconductor, polycrystalline silicon has electrical resistivity that varies with dopant concentration, temperature, and applied strain. The values reported in this table for resistivity and other properties reflect the range of values that have been used in research literature as inputs to simulations of polysilicon layers generated by the PolyMUMPS foundry service (MEMSCAP Inc., NC, USA). In general, many of the properties of polycrystalline silicon are process dependent and can vary widely [37]. Therefore, designs should ultimately be based upon measurements of the properties of material created according to a specific, defined process, and validation of device models against measured experimental output.

Example 1: V-Type Actuators

The first common thermal actuator mechanism has been called, variously, V-type, chevron, bent-beam, or symmetric thermal actuators. In this mechanism (Fig. 1), suspended, opposing beams are connected to a suspended shuttle structure. The ends of the actuator beams are anchored to a substrate that provides electrical current through the beams. Power dissipation leads to expansion of the beams, which experience a buckling motion. When the beams are placed symmetrically, linear motion is achieved in the shuttle. Engineering of the beams guides the buckling motion that they experience. The aspect ratio of the beam is defined asymmetrically in order to confine beam bending motion to one plane, typically chosen as motion in the plane of fabrication. If the beam has a cross section with thickness t, defined by the thickness of the film used for



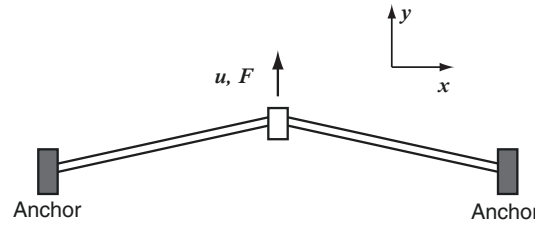
Thermal Actuators, Fig. 1 Examples of V-type "chevron" thermal actuators. (**a**) Tilted, oblique view in scanning electron microscope (*SEM*) image. (**b**) SEM image of the top view of a fabricated V-type thermal actuator. Beams are angled 1° off of perpendicular to the direction of motion of the central shuttle (Image credit: J.J. Brown and V.M. Bright)

fabrication of the beam, and a width w, defined by patterning perpendicular to the film, in-plane motion is achieved when $t > w$. In this case, the beam has a moment of inertia that is smaller for in-plane motion than for out-of-plane motion, so in-plane motion is more probable as buckling is induced. Another technique used to guide the buckling is to angle the beams at their attachment point to the shuttle. Typically the beams are angled a few degrees off of perpendicular to the shuttle, and the device is built with mirror symmetry across the shuttle in order to cancel bending moments that would cause deviations from uniaxial motion. With elongation of the actuator beams, buckling and shuttle motion is achieved in a predictable direction. The choice of angle is made dependent on whether displacement or output force is more important. Smaller angles from perpendicular provide more output displacement for a given actuator power; larger angles provide more output force.

Electrical Calculation

The goals of electrical calculations associated with thermal actuator design are the following:

1. Find the total resistance R of the actuator system in order to allow design of an appropriate power supply.

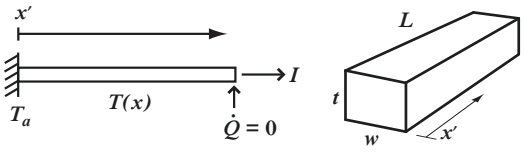


Thermal Actuators, Fig. 2 Fundamental unit of the V-type actuator, a pair of angled beams. Motion u and output force F occur in direction y. A loading force F may oppose motion in the y direction

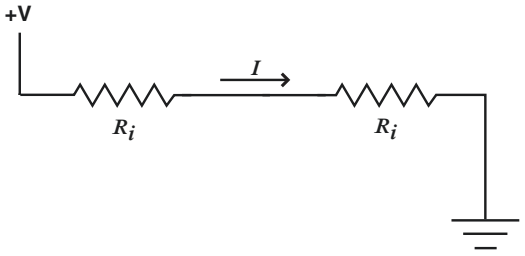
2. Understand how R relates to the actuator geometry.
3. Allow the use of I^2R to determine the total power generation and by extension the thermal power that must be managed by the anchors and other structures supporting the thermal actuator.

Regarding the third goal, the packaging environment surrounding the actuator must be able to remove the steady-state power generated by the actuator in order to establish stable operation.

The fundamental unit of a V-type thermal actuator is a symmetric beam pair as seen in Fig. 2. For simplicity of this example calculation, power consumption in the central shuttle will be omitted from the calculations. Because the actuator is



Thermal Actuators, Fig. 3 Basic geometry of one of the V-type actuator beams, for electrical and thermal calculations. The distance coordinate along the length of the angled beam is defined by x'



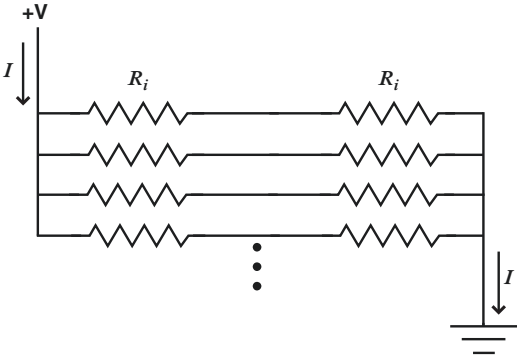
Thermal Actuators, Fig. 4 Electrical circuit for the pair of beams in the thermal actuator. Resistance in the central shuttle is omitted for simplicity

formed from symmetric beams, the calculation only necessitates examination of the power consumption in one beam. The geometry of the beam is defined as in Fig. 3. The beam has an electrical resistivity ρ, which is a temperature-dependent property. For the purposes of this initial calculation, ρ will be held constant, and the material of the beam will be defined as polycrystalline silicon, using the property values defined in Table 1. The beam resistance is calculated according to Eq. 15, using example dimensions for one beam: $L_i = 100\ \mu m$, $t_i = 3.5\ \mu m$, $w_i = 2\ \mu m$, $A_i = w_i t_i = 7\ \mu m^2$.

$$R_i = \frac{\rho L_i}{A_i} \tag{15a}$$

$$R_i = \frac{\left(34 \times 10^{-6}\ \Omega m\right)(100\ \mu m)}{(7\ \mu m^2)} = 490\ \Omega \tag{15b}$$

In Fig. 4, the electrical model of the pair of beams in the actuator indicates connection in series, so the resistance of the pair $R_p = 2R_i$. For an actuator built from N_p beam pairs, the beam pair resistance



Thermal Actuators, Fig. 5 Electrical circuit diagram for an assembly of beam pairs such as that seen in Fig. 1. The resistances of each beam pair add in parallel. The central shuttle makes electrical connections in the center of this diagram, but it is omitted from the diagram because there is assumed to be an identical voltage drop in each beam. Therefore, there should be no current flow from one beam pair to another through the central shuttle

adds in parallel, and the total resistance R of the actuator may be found according to Eq. 16. For $N_p = 10$ and the result from Eq. 15b, $R = 98\ \Omega$ (Fig. 5).

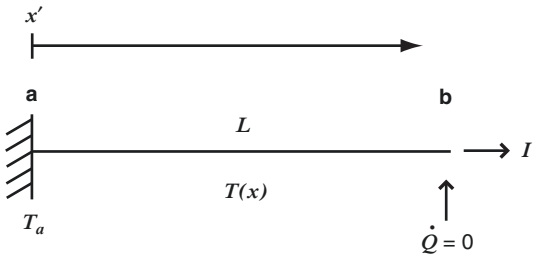
$$R = \frac{R_p}{N_p} = \frac{2R_i}{N_p} \tag{16}$$

Thermal Calculation

The goals of thermal calculations for the V-type actuator are the following:

1. Find an average temperature rise ΔT at a given input current in order to estimate the thermal strain in subsequent mechanical calculations.
2. Find maximum temperature as a function of input current, in order to understand the performance limits to the actuator.
3. Find boundaries on the thermal time constant in order to understand the limit to the actuator response time.

The calculation of the temperature profile in the thermal actuator can focus on the profile within a single beam, similar to the electrical calculation. As discussed earlier, for initial design purposes, it is necessary only to examine thermal conduction. Other heat transfer effects should be



Thermal Actuators, Fig. 6 One-dimensional model of the symmetric thermal system for the V-type actuator

considered for a complete understanding of the actuator thermal behavior. Simulations based on the coupled, temperature-dependent electrical, thermal, and mechanical behavior show that V-type thermal actuators may develop nonlinear output force behavior when subjected to displacement constraints [20]. However, for basic understanding, conductive heat transfer generally suffices. As with the electrical calculation, thermal conductivity k_θ and specific heat capacity c are temperature dependent, but for initial design they are considered to be constant.

The calculation can approximate the beam as a one-dimensional element with a temperature varying along position x. The boundary conditions are defined in two locations: the anchored end maintains a constant temperature due to significant heat transfer from the anchor to the supporting substrate. The suspended end of the beam has no heat flow, due to the symmetry of the suspended structure. Consider the diagram in Fig. 3, where T_a is the temperature at point **a**, the beam has a temperature profile that varies along the beam with position x', heat leaves the beam only through point **a**, and a total current I flows through the beam. This diagram can be simplified into a one-dimensional model (Fig. 6) with the following boundary conditions, where $\dot{Q}$ is one-dimensional heat flow with units of watts (W):

$$T(0) = T_a \, \dot{Q}(L_i) = 0$$

In this case, all variables in the thermal model are considered constant across the beam cross section at any given x' position. Using Eq. 17 to set up the one-dimensional model and assuming steady-state operation, the heat equation Eq. 5 can be rewritten as Eqs. 18a and 18b, given for cases without and with air conduction, respectively. In Eq. 18a, constant thermal conductivity is also assumed. For the model with conduction through air to a nearby substrate, S is a shape factor, z_g is the distance between the suspended device and the substrate, $k_{\theta,\text{air}}$ is the thermal conductivity of air, and Z is a thermal insulation term defined by the layers through which heat flows (with thickness z_i, thermal conductivity $k_{\theta,I}$, and total number p) from the suspended beam to a substrate held at a constant temperature [15–17, 38].

$$k_\theta A_i \frac{\partial T}{\partial x'} = -\dot{Q} \tag{17a}$$

$$\frac{\partial \dot{Q}}{\partial x'} = \dot{q} A_i \tag{17b}$$

$$\dot{q} = \rho \left| \vec{J} \right|^2 = \rho \frac{I^2}{A_i^2} \tag{17c}$$

$$S = \frac{t}{w}\left(\frac{2z_g}{t} + 1\right) + 1 \tag{17d}$$

$$Z = \sum_{j=1}^{p} \frac{z_i}{k_{\theta,\,i}} = \frac{z_g}{k_{\theta,\,\text{air}}} \tag{17e}$$

$$0 = k_\theta A_i \frac{\partial^2 T}{\partial x'^2} + \dot{q} A_i = k_\theta A_i \frac{\partial^2 T}{\partial x'^2} + \rho \frac{I^2}{A_i} \tag{18a}$$

$$0 = \frac{\partial}{\partial x'}\left(k_\theta \frac{\partial T}{\partial x'}\right) + \rho \frac{I^2}{A_i^2} - \frac{S(T - T_a)}{tZ} \tag{18b}$$

The temperature profile in the beam in vacuum, Eq. 19, is found by integrating Eq. 18a twice and applying the boundary conditions to determine the constants of integration. A more precise analytical solution to Eq. 18a may be found when k_θ and ρ are allowed to vary with temperature. Reference [17] provides a solution for Eq. 18b for constant k_θ and temperature-dependent ρ. Elimination of air conduction terms in that solution provides a solution for Eq. 18a with temperature-dependent ρ.

The maximum temperature T_{max} is found at the location where the first derivative of temperature is zero, which is where heat flow $\dot{Q}$ is zero. In this case, heat flow is zero at $x' = L_i$, and T_{max} is given by Eq. 20. The average temperature rise in the beam can be calculated by the application of Eq. 19 to Eq. 7, yielding Eq. 21.

$$T = T_a + \rho \frac{I^2}{A_i^2 k_\theta} \left(L_i x' - \frac{x'^2}{2} \right) \tag{19}$$

$$T_{max} = T_a + \rho \frac{I^2 L_i^2}{2A_i^2 k_\theta} \tag{20}$$

$$\Delta T = \frac{\rho I^2 L_i^2}{3A_i^2 k_\theta} \tag{21}$$

For an upper bound on the thermal time constant, the lumped thermal resistance R_θ can be calculated using the beam length L_i and assuming that the substrate is an infinitely large thermal reservoir that can adequately remove the generated thermal power P from the actuator without undergoing significant changes in temperature, thereby maintaining T_a at the beam anchor point. Equations 22a and 22b show the resulting time constant and the minimum bandwidth for operation in vacuum, derived from Eq. 8, polysilicon properties from Table 1, and example dimensions for one beam: $L_i = 100$ μm, $t_i = 3.5$ μm, $w_i = 2$ μm, $A_i = w_i t_i = 7$ μm^2. In air, the overall thermal resistance is reduced, thereby decreasing τ_θ and increasing the bandwidth.

$$\begin{aligned} \tau_\theta = R_\theta C &= \left(\frac{L_i}{k_\theta A_i} \right) (c\eta L_i A_i) = \frac{c\eta L_i^2}{k_\theta} \\ &= \frac{(705)(2330)\left(10^{-4}\right)^2}{30} \\ &= 5.5 \times 10^{-4} \text{s} \end{aligned} \tag{22a}$$

$$f_\theta = \frac{1}{2\pi\tau_\theta} = 290 \text{ Hz} \tag{22b}$$

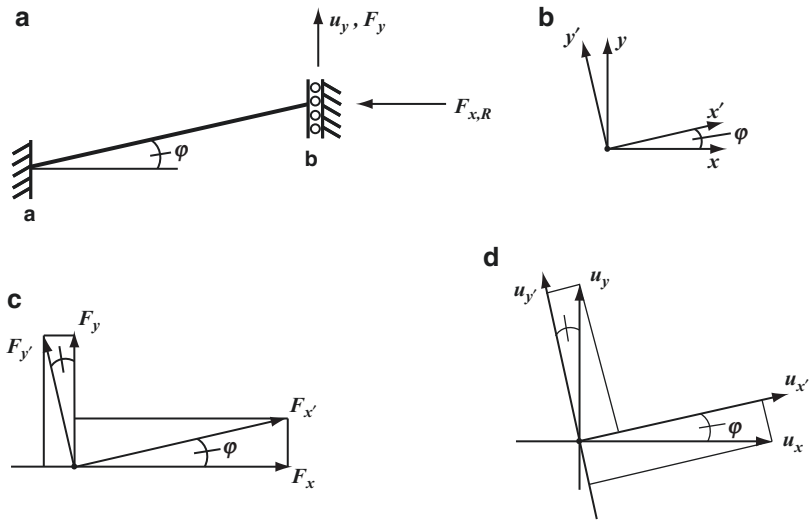
Analysis of Mechanical Behavior

For basic design of a thermal actuator, modeling of the mechanical behavior can help the designer to understand the maximum force and range of displacements that are available from a given actuator design. Mechanical analysis can be performed with the following goals:

1. Find an expression for the maximum force output available from the actuator
2. Find an expression for the maximum displacement from the actuator
3. Develop an understanding of how device geometry impacts the force and displacement

Several contrasting mechanical analyses of V-type actuators have been published [3, 4, 6, 19]. The analysis presented here follows that of Ref. [4]. As with the electrical and thermal analyses, the symmetric device geometry allows simplification of the mechanical analysis to an examination of only one beam. Begin by simplifying the actuator design in Fig. 2 to the free body diagram seen in Fig. 7a, and consider the coordinate system presented in Fig. 7b. At point **a**, the beam is anchored and restricted from displacement or rotation. At point **b**, the end of the beam is restricted from rotation or motion in the x direction, but it can move in the y direction. A reaction force $F_{x,R}$ opposes motion in the x direction. Figure 7b depicts two Cartesian coordinate systems. The xy system rests in the frame of the substrate and actuator support. The actuator output motion u_y is defined in the y direction of the xy coordinate system. The rotated $x'y'$ coordinate system is a local system oriented to the reference frame of the angled actuator beam. Consequentially, the $x'y'$ system is rotated an angle φ from the xy system.

An external force F_i may be applied to point **b** in the y direction, affecting the u_y displacement. In the reference frame of the beam, motion at point **b** can be decomposed to motion $u_{x'}$ in the x' direction and motion $u_{y'}$ in the y' direction (Fig. 7d). Motion in the x' direction is pure elongation, described by Eq. 23a to include thermal strain. Motion in the y' direction is pure bending, described by Eq. 23b for the case where rotation at the beam ends is not allowed.



Thermal Actuators, Fig. 7 Diagrams for mechanical analysis of V-type actuators. (**a**) Free body diagram. (**b**) Coordinate systems for output motion in *xy* and motion in the reference frame of the beam $x'y'$. (**c**) Bending and extensional forces in the beam can be projected into forces in the *xy* coordinate system. (**d**) Motion in the *xy* coordinates is projected into motion in the $x'y'$ coordinates

$$F_{x'} = EA\left(\frac{u_{x'}}{L} - \alpha\Delta T\right) = \frac{EA}{L}u_{x'} - \alpha\Delta TEA \tag{23a}$$

$$F_{y'} = \frac{12EI_z}{L^3}u_{y'} \tag{23b}$$

The forces $F_{x'}$ and $F_{y'}$ can be projected (Fig. 7c) to forces in x and y according to Eq. 24 for elongation and Eq. 25 for bending.

$$F_{x,e} = F_{x'}\cos\varphi \tag{24a}$$

$$F_{y,e} = F_{x'}\sin\varphi \tag{24b}$$

$$F_{x,b} = -F_{y'}\sin\varphi \tag{25a}$$

$$F_{y,b} = F_{y'}\cos\varphi \tag{25b}$$

Summing the F_x forces and F_y forces yields Eq. 26.

$$-F_{x,R} + F_{x,b} + F_{x,e} = -F_{x,R} - F_{y'}\sin\varphi + F_{x'}\cos\varphi = 0 \tag{26a}$$

$$-F_{x,R} - \sin\varphi\frac{12EI_z}{L^3}u_{y'} + \left(\frac{EA}{L}u_{x'} - \alpha\Delta TEA\right)\cos\varphi = 0 \tag{26b}$$

$$F_i + F_{y,b} + F_{y,e} = F_i + F_{y'}\cos\varphi + F_{x'}\sin\varphi = 0 \tag{26c}$$

$$F_i + \cos\varphi\frac{12EI_z}{L^3}u_{y'} + \left(\frac{EA}{L}u_{x'} - \alpha\Delta TEA\right)\sin\varphi = 0 \tag{26d}$$

The displacements $u_{x'}$ and $u_{y'}$ can be written in terms of u_x and u_y by projecting u_x and u_y onto the rotated coordinate system (Fig. 7d).

$$u_{x'} = u_y\sin\varphi - u_x\cos\varphi \tag{27a}$$

$$u_{y'} = u_y\cos\varphi - u_x\sin\varphi \tag{27b}$$

However, $u_x = 0$ at point **b**, so Eq. 27 simplifies to Eq. 28.

$$u_{x'} = u_y\sin\varphi \tag{28a}$$

$$u_{y'} = u_y\cos\varphi \tag{28b}$$

With substitution of Eq. 28 into Eq. 26d and some algebraic manipulation, a very useful final expression Eq. 29 can be derived that relates the average actuator temperature rise ΔT, the actuator output displacement u_y, and the force F_i that may be produced by the actuator and which may constrain the displacement. Substitution of Eq. 28 into

Eq. 26b would yield an expression for the x direction reaction force, if desired. Reference [4] appeared to show reasonable agreement between this model and measured unconstrained experimental displacement output. However, an exploration of force and displacement behavior performed in Ref. [20] demonstrated that V-type actuators undergo asymmetric buckling which complicates the relationship between force and displacement at different power or temperature levels. Therefore, Eq. 29 may be useful for basic design, but it underrepresents the interaction between F_i and u_y when u_y is constrained near zero or when large force is applied.

$$u_y = \frac{\alpha \Delta T E A \sin\varphi}{\left(\frac{EA}{L}\sin^2\varphi + \frac{12EI_z}{L^3}\cos^2\varphi\right)} + \frac{-F_i}{\left(\frac{EA}{L}\sin^2\varphi + \frac{12EI_z}{L^3}\cos^2\varphi\right)} \tag{29}$$

If the force constraint F_i is left at zero, Eq. 29 reduces to an expression for the free-moving actuator displacement. If instead the displacement u_y is constrained to zero, the magnitude of the maximum force generated by the beam Eq. 30 is seen to be the result of plane strain in the beams. A detailed thermomechanical derivation of stresses due to plane strain and thermal expansion applied to Eq. 1 may be found in Ref. [12]. The result is that, under the normal strain condition $\varepsilon_{33} = 0$ and no external traction or internal body forces, no stresses are present except for the compressive normal stress $\sigma_{33} = -\alpha E \Delta T$.

$$F_{i,max} = \alpha \Delta T E A \sin\varphi \tag{30}$$

If the thermal strain $\alpha\Delta T$ is held to zero, then a spring constant k_i for the beam Eq. 31 can be determined using Hooke's Law. This spring constant reduces to Eq. 10 in the small-angle approximation where $\varphi \to 0$. For actuators that contain N angled beams, the total spring constant k can be found according to Eq. 13 by multiplying k_i by N, and the maximum output force from the actuator would be found by multiplying Eq. 30 by N.

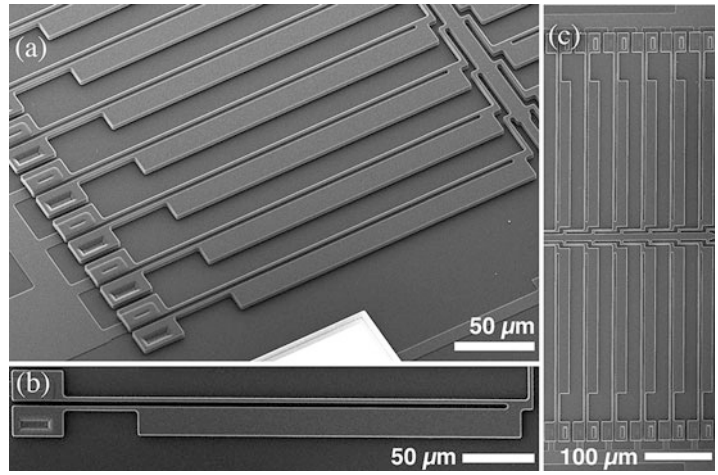
$$k_i = \frac{-F_i}{u_y} = \frac{EA}{L}\sin^2\varphi + \frac{12EI_z}{L^3}\cos^2\varphi \tag{31}$$

A final comment about the mechanical analysis of V-type actuators is that it may sometimes be necessary to include a temperature-dependent coefficient of thermal expansion in the average thermal strain calculated as an input to Eq. 29. Because the mechanical analysis presented here only concerns motion at the end of the actuator beam, the thermal strain can be determined as an average over the length of the beam using a modified version of Eq. 7, as shown here in Eq. 32.

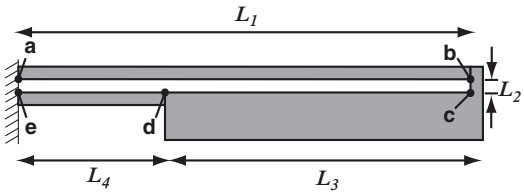
$$\overline{\alpha \Delta T} = \frac{1}{L}\int_0^L \alpha(T - T_a)dx' \tag{32}$$

Example 2: Hot-Arm Actuators

The second common thermal actuation mechanism may be referred to by any of the following names: pseudobimorph, U-type, asymmetric, hot-arm, or flexure electrothermal actuators. In these structures (Fig. 8), two parallel beams are provided. These beams are fixed to a substrate at one end and linked to each other at the opposite end. The beams are designed to experience differential thermal expansion such that one beam expands further than the other beam. These actuators can be designed for motion out of a plane of manufacture, but they are more commonly seen for motion within planes. For in-plane motion with one common design, a "hot arm" and a "cold arm" are defined. The cold arm typically consists of a narrow flexure near the attachment point of the arm, and a wider distal portion. The hot arm is a narrow beam of uniform cross section. Electric current flows through this structure. In the wide portion of the cold arm, the current density and thermal resistance are lower than in the other portions of this actuator due to a larger cross-section area. This region is thus subject to less temperature change and less thermal expansion than the other regions. With the flow of electric current, the cold arm elongates less than the hot arm. Because the two arms are constrained to move together, the two arms sweep an arc of motion due to the difference in thermal expansion



Thermal Actuators, Fig. 8 SEM images of U-type "hot arm" thermal actuators. (**a**) Tilted, oblique view of an actuator array. (**b**) Magnified view of one of the actuators in this array. The actuator is anchored at *left,* and greatest motion occurs in the parts at *right*. (**c**) *Top* view of an array of fabricated U-type thermal actuators (Image credit: J.J. Brown and V.M. Bright)



Thermal Actuators, Fig. 9 Schematic of one hot-arm actuator

between them. By symmetrically placing these U-type actuators around a central shuttle or yoke, and linking them to the shuttle with compliant structures, linear motion can be generated.

Electrical Analysis

Electrical and thermal analysis of these actuators can be performed by dividing the actuator into four segments, as seen in Fig. 9. Ignoring small discrepancies at the boundaries to the segments, the actuator can be converted into a one-dimensional model similar to that of Example 1 but constructed from a series connection of the four distinct segments, as seen in the electrical circuit diagram, Fig. 10.

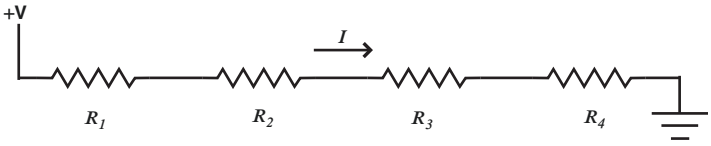
The total electrical resistance can be calculated, Eq. 33, by summing in series the electrical resistance for each segment and assuming constant resistivity. The resistance for each segment can be found according to Eq. 15a. In most cases, the thickness will be identical for all segments due to fabrication from a surface-micromachined material layer, so the thickness term can be moved outside of the summation along with the resistivity. For a polycrystalline silicon device with typical dimensions of $L_1 = 100$ μm, $L_2 = 5$ μm, $L_3 = 80$ μm, $L_4 = 20$ μm, $w_1 = w_4 = 2$ μm, $w_2 = 5$ μm, $w_3 = 20$ μm, and $t = 3.5$ μm, and with $\rho = 34 \times 10^{-6}$ Ωm, the total resistance R is 632 Ω.

$$R = \sum_{i=1}^{n} R_i = \sum_{i=1}^{n} \frac{\rho L_i}{A_i} = \rho \sum_{i=1}^{n} \frac{L_i}{w_i t_i} \tag{33}$$

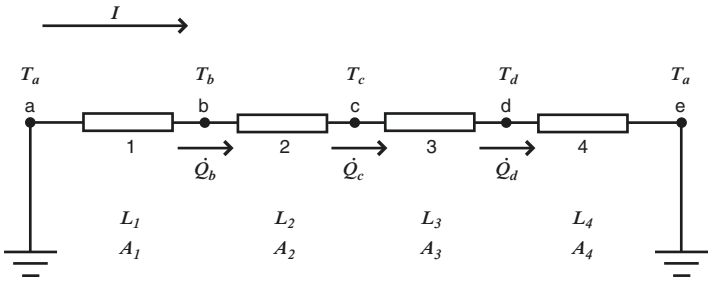
Thermal Analysis

The thermal analysis of the hot-arm actuator can proceed similarly to the electrical analysis, dividing the actuator into four segments according to the schematic in Fig. 11. The four segments are coupled together with temperature T_i and heat flow $\dot{Q}_i$ in each segment i, and with temperature and heat flow boundary conditions described in Table 2. The mechanical anchor points are maintained at constant temperature T_a. Uniform current flow I and constant electrical resistivity are assumed through the actuator.

Solution of Eq. 18a following a similar approach as in Example 1 leads to equations for temperature and heat flow in each segment, given by Eq. 34, with constants C_{1i} and C_{2i} determined for each ith segment by algebraic manipulation of the boundary conditions. The constants are



Thermal Actuators, Fig. 10 Simplified electrical schematic for a hot-arm actuator



Thermal Actuators, Fig. 11 Schematic for thermal analysis of the hot-arm actuator

Thermal Actuators, Table 2 Boundary conditions for the hot-arm actuator thermal model

Temperature boundary conditions	Heat flow boundary conditions
$T_1(0) = T_a$ $T_1(L_1) = T_b = T_2(0)$ $T_2(L_2) = T_c = T_3(0)$ $T_3(L_3) = T_d = T_4(0)$ $T_4(0) = T_a$	$\dot{Q}_1(L_1) = \dot{Q}_2(0)$ $\dot{Q}_2(L_2) = \dot{Q}_3(0)$ $\dot{Q}_3(L_3) = \dot{Q}_4(0)$

reported in Table 3. This analysis is similar to the approach of Ref. [17], with the difference that Ref. [16] also examines heat transfer to air in solving the heat equation for the hot-arm actuator.

$$T_i = -\frac{\rho I^2}{A_i^2 k_\theta}\frac{x_i^2}{2} + C_{1i}x_i + C_{2i} \tag{34a}$$

$$\dot{Q}_i = -k_\theta A_i \frac{\partial T}{\partial x_i} = -k_\theta A_i C_{1i} + \frac{\rho I^2}{A_i} x_i \tag{34b}$$

The maximum temperature $T_{\max}$ and its location $x_{\max}$ in the hot arm segment 1 can be found by setting the first derivative of temperature equal to zero. The first derivative of the temperature is proportional to the heat flow, given in Eq. 34b using C_{11}. The resulting temperature and position are given in Eqs. 35 and 36, respectively.

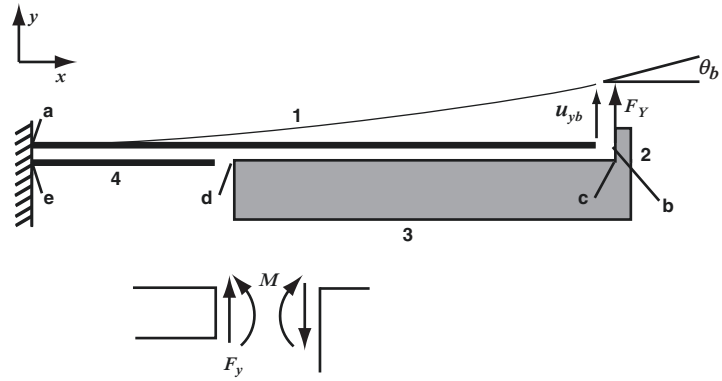
$$T_{\max} = T_1(x_{\max}) = T_a + \frac{I^2\rho}{2k_\theta}\frac{x_{\max}^2}{A_1^2} = T_a + \frac{I^2\rho}{2k_\theta}\left(\frac{1}{4}\right)\left(\frac{L_1}{A_1} + \frac{L_2}{A_2} + \frac{L_3}{A_3} + \frac{L_4}{A_4}\right)^2 \tag{35}$$

$$x_{\max} = \frac{A_1}{2}\left(\frac{L_1}{A_1} + \frac{L_2}{A_2} + \frac{L_3}{A_3} + \frac{L_4}{A_4}\right) \tag{36}$$

The average temperature rise ΔT and the thermal time constant can be calculated as described earlier, subject to the discretion of the designer in choosing the complexity of the system to be represented. One option is to focus solely on the hot arm because this is the segment that experiences the highest temperatures and the most significant thermal strain. In the hot arm the thermal behavior is most important for the actuator operation. In this case, the thermal capacitance C is calculated using the volume of the hot arm $V_m = L_1A_1$. The thermal resistance can be calculated using the location $x_{\max}$ of maximum temperature or the total length L_1 of the hot arm and the hot arm cross-section area A_1. For the average temperature rise ΔT, the average is taken over the length L_1 of the hot arm only, giving Eq. 37.

Thermal Actuators, Table 3 Algebraic constants for Eq. 34 in each segment of the hot-arm actuator

C_{1i}	C_{2i}
$C_{11} = \frac{I^2\rho}{2k_\theta A_1}\left(\frac{L_1}{A_1}+\frac{L_2}{A_2}+\frac{L_3}{A_3}+\frac{L_4}{A_4}\right)$	$C_{21} = T_a$
$C_{12} = \frac{I^2\rho}{2k_\theta A_2}\left(-\frac{L_1}{A_1}+\frac{L_2}{A_2}+\frac{L_3}{A_3}+\frac{L_4}{A_4}\right)$	$C_{22} = T_a + \frac{I^2\rho}{2k_\theta}\left(\frac{L_1}{A_1}\right)\left(\frac{L_2}{A_2}+\frac{L_3}{A_3}+\frac{L_4}{A_4}\right)$
$C_{13} = \frac{I^2\rho}{2k_\theta A_3}\left(-\frac{L_1}{A_1}-\frac{L_2}{A_2}+\frac{L_3}{A_3}+\frac{L_4}{A_4}\right)$	$C_{22} = T_a + \frac{I^2\rho}{2k_\theta}\left(\frac{L_1}{A_1}+\frac{L_2}{A_2}\right)\left(\frac{L_3}{A_3}+\frac{L_4}{A_4}\right)$
$C_{14} = \frac{I^2\rho}{2k_\theta A_4}\left(-\frac{L_1}{A_1}-\frac{L_2}{A_2}-\frac{L_3}{A_3}+\frac{L_4}{A_4}\right)$	$C_{24} = T_a + \frac{I^2\rho}{2k_\theta}\left(\frac{L_1}{A_1}+\frac{L_2}{A_2}+\frac{L_3}{A_3}\right)\left(\frac{L_4}{A_4}\right)$



Thermal Actuators, Fig. 12 Diagram for mechanical analysis of the hot-arm actuator. The lower part of this diagram illustrates the matching of moments and forces at points *b* and *d* where the rigid frame is cut from the two cantilevers

$$\Delta T = \frac{1}{L_1}\int_0^{L_1} T_1 dx - T_a = \frac{I^2\rho}{4k_\theta}\left(\frac{L_1}{A_1}\right)\left(\frac{L_1}{3A_1}+\frac{L_2}{A_2}+\frac{L_3}{A_3}+\frac{L_4}{A_4}\right) \quad (37)$$

Analysis of Mechanical Behavior

Few mechanical analyses of hot-arm actuators have been published. The results of analyses based on energy methods may be found in Refs. [17, 19]. The analysis presented here uses a more straightforward technique based on superposition. Begin by dividing the device depicted in Fig. 9 into three sections: two cantilevers and one rigid frame. As seen in Fig. 12, section 1 becomes a cantilever with clamped attachment at left and applied force F_{1y} and moment M_1 at right. Similarly, section 4 becomes a cantilever with shorter length and with clamped attachment at left and applied force F_{4y} and moment M_4 at right. Sections 2 and 3 become an L-shaped rigid frame that links the two cantilevers together with moment and force reactions at the joining surfaces. Additionally, an external force F_Y is applied at point **b** of this L-frame, and the frame also experiences a bending moment M due to thermal stresses in the cantilevers. In order to maintain simplicity in calculating the coupled bending behavior of the cantilevers for this example, this thermal bending moment is calculated separately rather than combining the axial deformation behavior of the cantilevers with the rotation of the rigid frame and the bending of the cantilevers.

To calculate M from the thermal strain in cantilever 1, use the sum of forces in the x direction, the sum of moments at point **c**, and Hooke's law applied to elongation of the two cantilevers in order to generate Eq. 38. For simplicity, thermal strain is assumed only to be present in cantilever 1. Algebraic manipulation leads to an expression for the displacement of the rigid frame, Eq. 39a, and an expression, Eq. 39b, for the thermal bending moment that drives operation of the hot-arm actuator.

$$F_{xb} + F_{xd} = 0 \quad (38a)$$

$$M - F_{xb}L_2 = 0 \quad (38b)$$

$$F_{xb} = k_1 e_1 = \frac{A_1 E}{L_1}(u_x - \alpha \Delta T L_1) \tag{38c}$$

$$F_{xd} = k_4 e_4 = \frac{A_4 E u_x}{L_4} \tag{38d}$$

$$u_x = \frac{\alpha \Delta T A_1}{\left(\frac{A_1}{L_1} + \frac{A_4}{L_4}\right)} \tag{39a}$$

$$M = \frac{-\alpha \Delta T A_1 L_2 E}{\left(1 + \frac{A_1 L_4}{A_4 L_1}\right)} \tag{39b}$$

In order to calculate the actuator motion, begin by examining the rigid frame and recording a set of force and moment equilibrium equations and compatibility equations that describe the relationship between the rotation and displacements of the cantilever ends Eq. 40. Note that the small-angle approximation is used to simplify Eq. 40d to the relation in Eq. 40e.

$$F_Y - F_{1y} - F_{4y} = 0 \tag{40a}$$

$$-M_4 - M_1 + M + L_3(F_Y - F_1) = 0 \tag{40b}$$

$$\theta_b = \theta_d = \theta \tag{40c}$$

$$u_{yd} + L_3 \sin\theta = u_{yb} \tag{40d}$$

$$u_{yd} = u_{yb} - L_3 \theta \tag{40e}$$

The bending of cantilevers is a well-characterized problem, and nodal equations may be written that relate the force and moment applied to the free end of the cantilever to the displacement and rotation that are present at that end [18]. For the cantilevers of sections 1 and 4 of the hot-arm actuator, these nodal equations are provided in Eq. 41.

$$F_{1y} = \frac{EI_1}{L_1^3}\left(12u_{yb} - 6L_1\theta\right) \tag{41a}$$

$$M_1 = \frac{EI_1}{L_1^3}\left(-6L_1 u_{yb} + 4L_1^2\theta\right) \tag{41b}$$

$$F_{4y} = \frac{EI_4}{L_4^3}\left(12u_{yd} - 6L_4\theta\right) \tag{41c}$$

$$M_4 = \frac{EI_4}{L_4^3}\left(-6L_4 u_{yd} + 4L_4^2\theta\right) \tag{41d}$$

By combining Eqs. 40 and 41, expressions may be derived that relate the force and moment applied in the rigid frame to the displacement and rotation experienced by the end of the actuator. These relations, Eq. 42, may be further manipulated to provide an expression for the actuator displacement as a function of applied force and applied moment, Eq. 43. Using the equation for M found in Eq. 39b, the expression in Eq. 43 relates input temperature and output force and displacement for the hot-arm actuator. Similar to Eq. 29 in Example 1, Eq. 43 may be further manipulated to find a spring constant for the actuator and a maximum output force.

$$\frac{F_Y}{E} = 12u_{yb}\left(\frac{I_1}{L_1^3} + \frac{I_4}{L_4^3}\right) - 6\theta\left(\frac{I_1}{L_1^2} + \frac{2L_3 I_4}{L_4^3} + \frac{I_4}{L_4^2}\right) \tag{42a}$$

$$\frac{M}{E} = -6u_{yb}\left(\frac{I_1}{L_1^2} + \frac{2L_3 I_4}{L_4^3} + \frac{I_4}{L_4^2}\right) + 2\theta\left[2\left(\frac{I_1}{L_1} + \frac{I_4}{L_4}\right) + 6I_4\left(\frac{L_3}{L_4^2} + \frac{L_3^2}{L_4^3}\right)\right] \tag{42b}$$

$$u_{yb} = \frac{\frac{M}{2E}\left(\frac{I_1}{L_1^2} + \frac{I_4}{L_4^2} + \frac{2L_3 I_4}{L_4^3}\right) + \frac{F_Y}{6E}\left[2\left(\frac{I_1}{L_1} + \frac{I_4}{L_4}\right) + 6I_4\left(\frac{L_3}{L_4^2} + \frac{L_3^2}{L_4^3}\right)\right]}{\left(\frac{I_1}{L_1^2} + \frac{I_4}{L_4^2}\right)^2 + \frac{4I_1 I_4}{L_1 L_4}\left(\frac{1}{L_1} - \frac{1}{L_4}\right)^2 + \frac{12L_3 I_1 I_4}{L_1^2 L_4^2}\left(\frac{1}{L_1} - \frac{1}{L_4} + \frac{L_3}{L_1 L_4}\right)} \tag{43}$$

Experimental verification of this or related hot-arm actuator models has not been clearly performed, as most hot-arm actuator experiments have recorded data in air environments, which require different thermal analysis from that performed in this example. In Ref. [17], experimental displacement measurements appear to correspond fairly well with a curve predicted from input current but with ΔT calculated to include the effects of heat loss to air. A final point about this mechanical analysis of hot- and cold-arm actuators is that variations of these actuators exist, but similar analytic approaches can apply to these variations. For instance, if L_3 is set equal to zero, then Eq. 43 describes the motion of a two-arm actuator that has no wide region but does have different length cantilevers and different temperature profiles in each cantilever due to the length difference [39]. The out-of-plane pseudobimorph actuator could be analyzed similarly [16].

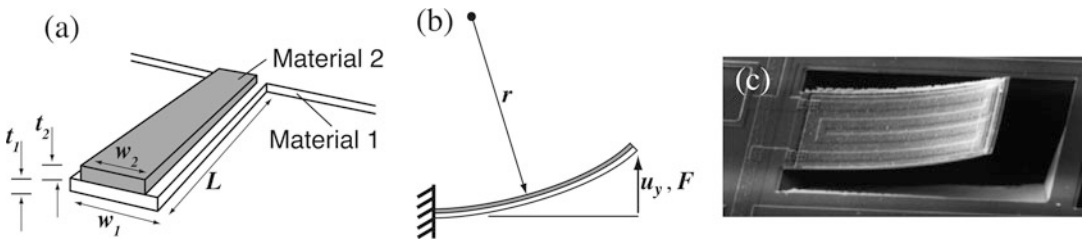
Example 3: Thermal Bimorph Actuators

The bimorph thermal actuator predates the designs explored in the previous two examples. The principle of operation of this device is the generation of bending due to a differential in thermal expansion between two materials stacked together to form the actuator. The bimorph actuator is typically constructed as shown in Fig. 13, as a cantilevered beam composed of two (bimorph) or more (multimorph) dissimilar materials and with a resistive heater patterned on top of the beam or formed from one of the beam materials. Advantages of bimorph structures are high force, simple design, and low operating voltage. The major disadvantages of these structures are relatively high electrical power consumption, curling motion, low operating frequency, and laminated construction. Many examples of implementation of bimorph thermal actuators including more discussion of simulation and analytical models may be found in Ref. [16].

Power consumption and average temperature rise can be calculated by applying concepts from Example 1. Others have extensively described the mechanical behavior of bimorph thermal actuators [16, 19, 40]. This example will review the relevant mechanical concepts and results. According to Fig. 13, assume the lengths of the two layers in the actuator are identical to length L. Using the small-angle approximation, deflection due to uniform curvature in the cantilever is given by Eq. 44, where r is a radius of curvature in the cantilever that is generated due to the thermal expansion mismatch of the two materials. Superimposed on this displacement is the deflection due to force applied at the end of a cantilever, Eq. 45, where $\overline{EI}_i$ is the average flexural rigidity of the composite cantilever. Together, these give a total displacement described by Eq. 46, analogous to Eqs. 29 and 43 presented earlier.

$$u_{y1} = \frac{L^2}{2r} \tag{44}$$

$$u_{y2} = \frac{FL^3}{3\overline{EI}_i} \tag{45}$$



Thermal Actuators, Fig. 13 Illustration of bimorph thermal actuators. (**a**) Bimorph cantilever consisting of two materials with identical length but different thicknesses, widths, and composition. (**b**) Diagram of radius of curvature r and cantilever tip displacement u_y used in analysis of the output motion of the bimorph actuator. (**c**) SEM image of a bimorph actuator ~300 μm length, with a heater trace patterned on its *top* surface (Image credit: B. Read and V.M. Bright, Air Force Institute of Technology, 1994)

$$u_y = \frac{L^2}{2r} + \frac{FL^3}{3\overline{EI_i}} \tag{46}$$

The radius of curvature r is determined Eq. 47 using force and moment equilibrium in the composite beam and strain compatibility at the bonded surface between the two materials [19, 40]. The combined flexural rigidity $\overline{EI_i}$ is found for the nonhomogeneous beam according to Eq. 48 after the moments I_1 and I_2 are first calculated for the different layers in relation to the neutral surface of the beam [18, 19]. The result is Eq. 49. In Ref. [40], Eqs. 47 and 44 are derived and shown to lead to agreement with previously reported experimental data. Equations 45, 46, and 49 are derived through continuum solid mechanics analysis [18].

$$r = \frac{\left(E_1 w_1 t_1^2\right)^2 + \left(E_2 w_2 t_2^2\right)^2 + 2E_1 w_1 t_1 E_2 w_2 t_2 \left(2t_1^2 + 3t_1 t_2 + 2t_2^2\right)}{6E_1 w_1 t_1 E_2 w_2 t_2 (\alpha_1 - \alpha_2)\Delta T} \tag{47}$$

$$\overline{EI_i} = E_1 I_1 + E_2 I_2 \tag{48}$$

$$\overline{EI_i} = \frac{\left(E_1 w_1 t_1^2\right)^2 + \left(E_2 w_2 t_2^2\right)^2 + 2E_1 w_1 t_1 E_2 w_2 t_2 \left(2t_1^2 + 3t_1 t_2 + 2t_2^2\right)}{12(E_1 w_1 t_1 + E_2 w_2 t_2)} \tag{49}$$

Future Directions

Materials Development, Optimization, and Miniaturization

Recent years have seen large advancement in materials technologies and computational tools, and it is likely that these advances will continue to impact thermal actuator systems for a significant amount of time. The principles described above are valid for any mechanical system with materials that experience thermal expansion; therefore, thermal actuators provide a pathway to immediate device fabrication using newly demonstrated materials. Furthermore, the phenomena of electrical power dissipation, heat flow, and thermal expansion are valid concepts at submicron size scales. Continued miniaturization may require consideration of additional phenomena such as the effects of dimensional constraints on the motion of phonons and electrons, increased importance of surface forces with smaller sizes, and the Thomson effect in systems where electrical current is flowing [41]. It may be expected that thermal actuator designs will be demonstrated in submicron sizes as fabrication technology improves. Ultimately, thermal actuator designs formed from engineered macromolecular structures may be possible. As molecular-scale thermal actuators are developed, such devices may not be subject to the same friction conditions as current microscale technology, and thus pinned connections and sliding surfaces may be within the range of designs allowed at the nanoscale.

Environmental Robustness

As discussed earlier, the temperature profile of the thermal actuator, and the corresponding output displacement or force, can vary due to the environment in which the actuator is operated. In vacuum environments at ambient temperature, this may be a small effect, but operation in ambient air and gaseous environments can require significant modification to the actuator power. For reliable and predictable operation across a range of environments, improved control systems for thermal actuators are needed. Additionally, many possible studies of environment effects comparing experiment with simulations and analytical models remain to be explored for thermal actuators. These may, for instance, include effects of

vacuum, liquid, high temperature, high radiative heat transfer, low temperature, packaging design, and environmental transients.

Device Control and Sensor Integration

One approach to improved displacement control and robustness would be implementation of closed-loop control systems that rely on data from a displacement sensor integrated to the actuator. In practice, integration of accurate in-plane displacement sensors for actuator control presents a significant design and fabrication challenge. Capacitive sensors have been integrated to thermal actuators, but these sensors typically require a large footprint to achieve useful sensitivity at small displacements [4]. Operation of capacitive sensors must be driven by sophisticated electronics placed physically close to the sensor in order to avoid signal losses due to parasitic capacitance. Piezoresistive sensors require smaller device footprints than capacitive sensors but may be subject to distorted readings due to temperature variation resulting from heat dissipation from the thermal actuator. Furthermore, readings from piezoresistive sensors can be distorted by the electrical operation of the thermal actuator if the sensor is monolithically linked and thus electrically connected to the thermal actuator. Strategies for electrical isolation of piezoresistive sensors are needed. Optical sensors provide the most accurate options for displacement measurement, but they are not yet easily integrated with other chip scale devices for in-plane displacement measurements.

Application Development

As one of the key mechanisms for generating motion in microscale and nanoscale electromechanical systems (MEMS and NEMS), it may be expected that a major thrust of thermal actuator development will track the overall development and use of MEMS and NEMS. Thermal actuators are an important part of the toolkit available to microsystem designers. As new designs proliferate, it may be expected that a certain portion of these designs incorporate thermal actuators. As discussed earlier, current usages include optical and electrical switches, microgrippers, nanomanipulation structures, and tensile test devices. As these tools move into broader use, thermal actuators are likely to remain key components of these devices. Other key opportunities for new application development using thermal actuators include control of optical gratings, lock and security systems, and controlled motion in microscale positioning stages. Many MEMS and NEMS applications may yet be conceived requiring low actuation voltage, compact footprint, and high force output, and thermal actuators provide a useful design option for all such applications.

Force Output Measurement and Calibration

The use of thermal actuators in applications where force is applied to microscale and nanoscale objects raises the question "Exactly how much force is applied?" Answering this question requires calibration and measurement of the force applied with the actuator. No standardized, traceable approaches exist for microscale force calibration, and current best practices rely on measurement of the displacement of a well-characterized spring. Such an approach relies on the characterization of the spring, which itself is not a traceable, standardized practice for microscale and nanoscale devices. Furthermore, many analytical and finite element mechanical models have been presented in literature, but more experimental verification of these models is needed, as the observations of Ref. [20] show. As thermal actuators and other actuators are employed for generation of motion in microscale and nanoscale devices, further strategies are needed for understanding the magnitude of forces within these devices.

Cross-References

- ▶ Basic MEMS Actuators
- ▶ Compliant Mechanisms
- ▶ Organic Actuators
- ▶ Thermally Actuated Silicon Resonators

References

1. Lu, S., Dikin, D.A., Zhang, S., Fisher, F.T., Lee, J., Ruoff, R.S.: Realization of nanoscale resolution with a micromachined thermally-actuated testing stage. Rev. Sci. Instrum. **75**, 2154–2162 (2004)
2. Howell, L.L.: Compliant Mechanisms. Wiley, New York (2001)
3. Que, L., Park, J.-S., Gianchandani, Y.: Bent-beam electrothermal actuators – part I. Single beam and cascaded devices. J. Microelectromech. Syst. **10**, 247–254 (2001)
4. Zhu, Y., Corigliano, A., Espinosa, H.D.: A thermal actuator for nanoscale in situ microscopy testing: design and characterization. J. Micromech. Microeng. **16**, 242–253 (2006)
5. Brown, J.J., Suk, J.W., Singh, G., Baca, A.I., Dikin, D. A., Ruoff, R.S., Bright, V.M.: Microsystem for nanofiber electromechanical measurements. Sensors Actuators A Phys. **155**, 1–7 (2009)
6. Gianchandani, Y.B., Najafi, K.: Bent-beam strain sensors. J. Microelectromech. Syst. **5**, 52–58 (1996)
7. Lee, S.-W., Hierold, C.: Electrical and thermal insulation via an oxidized, rough contact interface for the electro-thermal actuation of carbon nanotubes. Sensors Actuators A Phys. **210**, 10–17 (2014)
8. Qin, Q., Zhu, Y.: Temperature control in thermal microactuators with applications to in-situ nanomechanical testing. Appl. Phys. Lett. **102**, 013101 (2013)
9. Alamin Dow, A.B.., Jazizadeh, B., Kherani, N.P., Rangelow, I.: Development and modeling of an electrothermally MEMS microactuator with an integrated microgripper. J. Micromech. Microeng. **21**, 125026 (2011)
10. Kim, D.H., Park, Y.C., Park, S.: Design and fabrication of twisting-type thermal actuation mechanism for micromirrors. Sensors Actuators A Phys. **159**, 79–87 (2010)
11. Erdem, E.Y., Chen, Y.-M., Mohebbi, M., Suh, J.W., Kovacs, G.T.A., Darling, R.B., Böhringer, K.F.: Thermally actuated omnidirectional walking microrobot. J. Microelectromech. Syst. **19**, 433–442 (2010)
12. Fung, Y.C., Tong, P.: Classical and Computational Solid Mechanics. Advanced Series in Engineering, vol. 1. World Scientific Publishing Co. Pte. Ltd, Singapore (2001)
13. Jackson, J.D.: Classical Electrodynamics, 2nd edn. Wiley, New York (1975)
14. Mills, A.F.: Heat Transfer, 2nd edn. Prentice Hall, Upper Saddle River (1999)
15. Lott, C.D., McLain, T.W., Harb, J.N., Howell, L.L.: Modeling the thermal behavior of a surface-micromachined linear-displacement thermomechanical microactuator. Sensors Actuators A Phys. **101**, 239–250 (2002)
16. Geisberger, A.A., Sarkar, N.: Techniques in MEMS microthermal actuators and their applications. In: Leondes, C.T. (ed.) MEMS/NEMS Handbook: Techniques and Applications. Sensors and Actuators, vol. 4, pp. 201–261. Springer, New York (2006)
17. Huang, Q.-A., Lee, N.K.S.: Analysis and design of polysilicon thermal flexure actuator. J. Micromech. Microeng. **9**, 64–70 (1999)
18. Craig Jr., R.R.: Mechanics of Materials, 2nd edn. Wiley, Hoboken (2000)
19. Lobontiu, N., Garcia, E.: Mechanics of Microelectromechanical Systems. Kluwer, New York (2005)
20. Wittwer, J.W., Baker, M.S., Howell, L.L.: Simulation, measurement, and asymmetric buckling of thermal microactuators. Sensors Actuators A Phys. **128**, 395–401 (2006)
21. Kwan, A.M.H., Song, S., Lu, X., Lu, L., Teh, Y.-K., Teh, Y.-F., Chong, E.W.C., Gao, Y., Hau, W., Zeng, F., Wong, M., Huang, C., Taniyama, A., Makino, Y., Nishino, S., Tsuchiya, T., Tabata, O.: Improved designs for an electrothermal in-plane microactuator. J. Microelectromech. Syst. **21**, 586–595 (2012)
22. Hazra, S.S., Baker, M.S., Beuth, J.S., de Boer, M.P.: Compact on-chip microtensile tester with prehensile grip mechanism. J. Microelectromech. Syst. **20**, 1043–1053 (2011)
23. Sahu, B., Taylor, C.R., Leang, K.K.: Emerging challenges of microactuators for nanoscale positioning, assembly, and manipulation. J. Manuf. Sci. Eng. **132**, 030917 (2010)
24. Brown, J.J., Dikin, D.A., Ruoff, R.S., Bright, V.M.: Interchangeable stage and probe mechanisms for microscale universal mechanical tester. J. Microelectromech. Syst. **21**, 458–466 (2012)
25. Arthur, C., Ellerington, N., Hubbard, T., Kujath, M.: MEMS earthworm: a thermally actuated peristaltic linear micromotor. J. Micromech. Microeng. **21**, 035022 (2011)
26. Ouyang, J., Zhu, Y.: Z-shaped MEMS thermal actuators: piezoresistive self-sensing and preliminary results for feedback control. J. Microelectromech. Syst. **21**, 596–604 (2012)
27. Gerratt, A.P., Bergbreiter, S.: Microfabrication of compliant all-polymer MEMS thermal actuators. Sensors Actuators A Phys. **177**, 16–22 (2012)
28. Kim, Y.-S., Dagalakis, N.G., Gupta, S.K.: Design of MEMS based three-axis motion stage by incorporating a nested structure. J. Micromech. Microeng. **24**, 075009 (2014)
29. Li, J., Vadahanambi, S., Kee, C.-D., Oh, I.-K.: Electrospun fullerenol-cellulose biocompatible actuators. Biomacromolecules **12**, 2048–2054 (2011)
30. Liang, J., Huang, L., Li, N., Huang, Y., Wu, Y., Fang, S., Oh, J., Kozlov, M., Ma, Y., Li, F., Baughman, R., Chen, Y.: Electromechanical actuator with controllable motion, fast response rate, and high-frequency resonance based on graphene and polydiacetylene. ACS Nano **6**, 4508–4519 (2012)
31. Buja, F., Sumant, A.V., Kokorian, J., van Spengen, W. M.: Electrically conducting ultrananocrystalline diamond for the development of a next generation of

micro-actuators. Sensors Actuators A Phys. **214**, 259–266 (2014)
32. Pal, S., Xie, H.: Fabrication of robust electrothermal MEMS devices using aluminum-tungsten bimorphs and polyimide thermal insulation. J. Micromech. Microeng. **22**, 115036 (2012)
33. Khazaai, J.J., Qu, H.: Electro-thermal MEMS switch with latching mechanism: design and characterization. IEEE Sensors J. **12**, 2830–2838 (2012)
34. Xu, Y., Li, G.: Thermal actuation using nanocomposites: a computational analysis. J. Heat Transf. **134**, 112401 (2012)
35. Liu, Q., Liu, L., Kuang, J., Dai, Z., Han, J., Zhang, Z.: Nanostructured carbon materials based electrothermal air pump actuators. Nanoscale **6**, 6932–6938 (2014)
36. Brown, K.A., Eichelsdorfer, D.J., Mirkin, C.A.: Cantilever-free thermal interfacing. J. Vac. Sci. Technol. **B31**, 06F201 (2013)
37. Sharpe Jr., W.N.: Mechanical properties of MEMS materials. In: Gad-el-Hak, M. (ed.) MEMS: Introduction and Fundamentals, 2nd edn. CRC/Taylor & Francis, Boca Raton (2006)
38. Howell, L.L., McLain, T.W., Baker, M.S., Lott, C.D.: Techniques in the design of thermomechanical actuators. In: Leondes, C.T. (ed.) MEMS/NEMS Handbook: Techniques and Applications. Sensors and Actuators, vol. 4, pp. 187–200. Springer, New York (2006)
39. Huang, Q.-A., Lee, N.K.S.: Analytical modeling and optimization for a laterally-driven polysilicon thermal actuator. Microsyst. Technol. **5**, 133–137 (1999)
40. Chu, W.-H., Mehregany, M., Mullen, R.L.: Analysis of tip deflection and force of a bimetallic cantilever microactuator. J. Micromech. Microeng. **3**, 4–7 (1993)
41. Jungen, A., Pfenniger, M., Tonteling, M., Stampfer, C., Hierold, C.: Electrothermal effects at the microscale and their consequences on system design. J. Micromech. Microeng. **16**, 1633–1638 (2006)

Thermal Cancer Ablation Therapies Using Nanoparticles

Steven Curley[1] and Charles B. Barker[2]
[1]Department of Surgical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, TX, USA
[2]Department of Mechanical Engineering and Materials Science, Rice University, Houston, TX, USA

Synonyms

Heat; Magnetic electromagnetic radiation; Near infrared; Radiofrequency

Definition

Nanoparticle-sized materials have interesting physical, chemical, and electronic properties. Among these is the release of heat upon exposure to various types of electromagnetic radiation. For example, iron and other magnetic nanoparticles can be injected directly into malignant tumors and then release heat in response to frictional heating in an alternating magnetic field. The intrinsic plasmon resonance of gold has been utilized to produce near-infrared laser-induced thermal activation of gold nanoparticles to treat malignant lesions on the skin or mucosal surfaces. Finally, several types of conducting and semiconducting nanoparticles have been conjugated successfully to antibodies or peptides to target the nanoparticles to malignant tumor cells or to the tumor neovasculature. The malignant cells are then exposed to a noninvasive shortwave radiofrequency field which results in intracellular heating and thermal cytotoxicity to the malignant cells. Cancer cells generally are more sensitive to thermal injury compared to normal cells; thus, nanoparticle-based thermal tumor ablation has the potential to be an important and hopefully less toxic addition to current anticancer treatments.

Overview

Hyperthermic cancer therapy was first described in Egyptian papyri over 4000 years ago by applying hot oil or cautery to tumors. Modern hyperthermic treatments are now applied to both premalignant and malignant tumors in various locations such as the peritoneal cavity, liver, kidneys, lungs, and prostate. Hyperthermia also potentiates the effects of cytotoxic chemotherapeutic agents and improves the response of tumors to ionizing radiation. Thus, regional or systemic hyperthermia has been induced as part of a treatment regimen in patients with certain types of malignancies. Hyperthermic isolated limb perfusion with melphalan with or without dactinomycin is performed to treat in-transit melanoma metastases and is being investigated as a treatment for advanced extremity soft tissue sarcoma [1]. In the last several years,

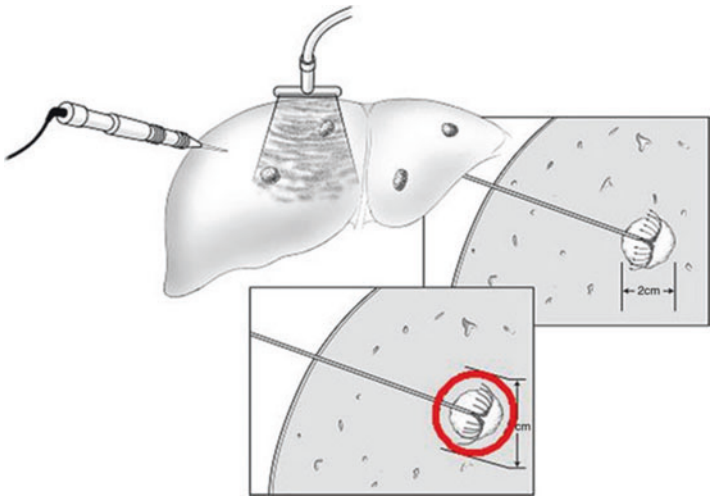
there have been a growing number of reports on hyperthermic intraperitoneal chemotherapy combined with cytoreductive surgery for patients with diffuse peritoneal spread of colorectal cancer, gastric cancer, mesothelioma, or pseudomyxoma peritoneii [2].

Although effective at treating several cancer types, recent studies have shown that some mammalian cancer cells are sensitive, whereas others are significantly more resistant to hyperthermia compared to healthy cells [3]. The reasons for the variable sensitivity to hyperthermia in cancer are multifactorial, but are primarily due to the role of heat shock proteins (HSPs) in cancer's resistance to the physical stress of increasing temperature [3]. Mammalian cancer cells that are heat sensitive frequently have mutated or underexpressed HSPs, whereas deregulation in some malignant cells leads to increased expression of HSPs which confers increased thermal protection and reduces apoptosis following thermal stress.

Many hyperthermic devices have been designed to produce low-level total-body hyperthermia (42–46 °C) and have been used to treat a wide variety of advanced malignancies combined with systemic cytotoxic agents. Both regional and systemic hyperthermic treatment systems elevate tissue temperatures from 3 °C to 7 °C above normal over a 1–12-h period. Some populations of malignant cells are sensitive to relatively long durations of exposures to low-level hyperthermia, but results to date suggest that these regional and systemic hyperthermic treatments have produced little benefit in overall patient outcomes at the expense of major regional and systemic toxicities and side effects. Interestingly, exposure of mammalian cells to temperatures of 55–60 °C for periods of a few seconds to less than a minute has been shown to produce apoptosis and necrosis, so a treatment system that can achieve such brief periods of heating in cancer cells would be highly desirable.

Invasive Thermal Ablation Therapies

Thermal tumor ablation is routinely practiced, particularly for malignant liver tumors, through direct intratumoral insertion of radiofrequency (RF) needle electrodes, microwave-emitting probes, or end- and side-firing laser fibers [4]. RF ablation therapy is an invasive treatment that is implemented by inserting needle electrodes directly into the tumor(s) to be treated and applying RF current through the wire into the tumor, resulting in local tumor and adjacent tissue necrosis (Fig. 1). The basis for the observed heating phenomena is understood by the rotational response of the charged ions and proteins in the tissue under the RF field. This rapid molecular rotation results in local frictional heating of the aqueous environment of the tumor and death of



Thermal Cancer Ablation Therapies Using Nanoparticles, Fig. 1 Schematic illustration demonstrating invasive radiofrequency ablation of malignant liver tumors. A multiple array needle electrode is inserted directly into the tumor to be treated using ultrasound guidance to assure proper placement of the needle. An alternating current is then passed across the array resulting in frictional ionic heating in the tumor and tissue surrounding the array

tumor cells through protein denaturation, melting of lipid bilayers, and destruction of nuclear DNA.

Although invasive RFA has demonstrated efficacy in cancer therapy, there are several limitations to this approach. First, clinical studies have shown incomplete tumor destruction in 5–40 % of the treated lesions, allowing significant probability of local cancer recurrence and metastasis. Another limitation of RFA is that the treatment is not targeted (i.e., nonspecific). Placement and insertion of the needle electrode is usually guided by an imaging system with poor spatial resolution (e.g., ultrasonography ~ mm), and therefore, this method is dependent on the experience of the physician that implements RFA. Furthermore, RFA treatments result in significant complications in approximately 10 % of patients due to thermally induced necrosis of healthy noncancerous tissues. The final and the most problematic limiting factor with respect to wire-based RFA is that it is a clinical treatment that is indicated for just a few select organ sites and specific tumor types (e.g., liver, kidney, breast, lung). On the other hand, tissue penetration by noninvasive RF energy fields is known to be excellent. Theoretically, noninvasive RF treatment of malignant tumors at any site in the body should be possible if agents that preferentially respond to RF energy with intracellular release of heat can be delivered to the malignant cells.

Roles for Nanotechnology in Cancer Diagnosis and Therapy

Nanotechnology is a term used to describe the engineering of materials that utilize the unique physical properties of nanometer scale matter. Nanomaterials have been used across all the major scientific disciplines including chemistry, physics, and medicine. It is in the medical realm that nanotechnology is widely heralded as one of the most promising and important approaches to diagnose and treat cancer [5]. Some nanoscale materials are already FDA approved and are used clinically. For example, iron oxide nanoparticles (Feridex™) have been shown to enhance the diagnostic capability of magnetic resonance imaging (MRI) to detect cancer [6]. Conjugating iron nanoparticles to antibodies that target proteins expressed on the surface of human cancer cells may further enhance the accuracy of MRI to diagnose early stage cancer [6].

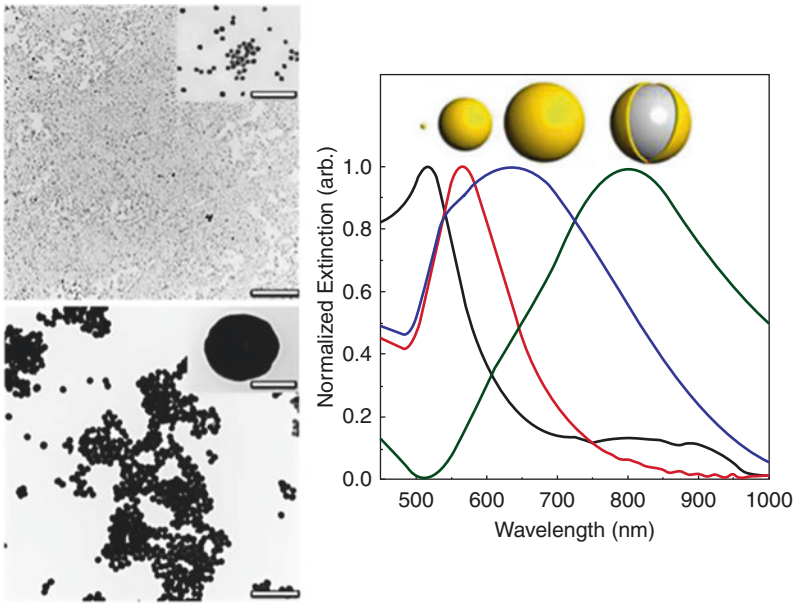
Other nanoparticles such as carbon or polymeric nanoparticles labeled with fluorine-18 deoxyglucose have been studied in preclinical models to enhance tumor diagnosis and detection rates using positron emission tomography. Surface modification of quantum dots, which are semiconducting nanocrystals that fluoresce upon absorption of visible light, is being investigated to improve detection of lymph nodes and other sites of metastases during surgical procedures [7]. Additionally, conjugation of quantum dots with tumor-specific peptides or antibodies may improve targeting of cancer cells and thus improve the diagnostic accuracy of this technique. Optimal imaging techniques using fluorescent nanoparticles targeted to a variety of types of human cancer with immunoconjugation to antibodies are being studied to permit in vivo localization of malignant cells. It is hoped that such imaging techniques will improve the diagnostic accuracy in numerous types of imaging modalities used to detect and follow patients with cancer. It is possible that these techniques may also allow earlier detection of cancer in high-risk populations and guide the duration and type of therapy in patients with more advanced stages of malignant disease. Finally, immunocomplexes consisting of gold nanoparticles and labeled antibodies have been demonstrated to improve the detection of several known serum tumor markers, including carcinoembryonic antigen, carcinoma antigen 125, and carbohydrate antigen 19–9, in a more rapid and accurate fashion than currently available techniques.

Nanoparticles are being used as intravascular carriers to enhance delivery of anticancer agents to malignant cells. Given that most anticancer drugs have a narrow therapeutic index with significant acute and cumulative toxicities in normal tissues, a variety of types of nanoparticles are being studied in preclinical and in vitro experiments to increase the delivery of cytotoxic chemotherapy drugs to cancer cells while reducing

toxicity by limiting exposure of the drug to normal tissues. For example, nanoparticles loaded with hydroxycamptothecin, 5-fluorouracil, docetaxel, and gemcitabine have been used in preclinical studies to improve chemotherapy-induced cytotoxicity in lung, colon, squamous, and pancreas cancers. Colloidal gold nanoparticles bearing tumor necrosis factor-α (TNF-α) molecules are being used in early phase human clinical trials to treat several types of cancer. Preclinical studies demonstrated that delivery of TNF-α to malignant tumors was enhanced using the nanoparticle delivery system, while avoiding the systemic toxicities that usually limit the clinical utility of this biologic agent. Nanoparticles conjugated with targeting agents are also being investigated to deliver gene therapy payloads to malignant cells. Carbon nanotubes have been used to deliver genes or proteins through nonspecific endocytosis in cancer cells. Similar to some cytotoxic chemotherapy drugs, gold or other metal nanoparticles have been shown to improve the therapeutic efficacy of external beam ionizing radiation in preclinical models. In addition to being used to deliver pharmaceutical agents, nanoparticles may be used as intrinsic chemotherapeutic agents. Gold nanoparticles 5–10 nm in diameter have been shown to have antiangiogenic properties in solid tumors and appear to induce apoptosis in chronic lymphocytic B cell leukemias through mechanisms not elucidated entirely.

Nanoparticle-Based Thermal Tumor Ablation Strategies

Recent studies have shown that nanoscale materials are capable of enhancing thermal deposition in cancer cells by absorbing optical energy. Gold-silica nanoshells 150 nm in diameter have a particular plasmon resonance in near-infrared (NIR) wavelengths (650–950 nm) of light and release significant heat when exposed to NIR laser sources (Fig. 2) [8]. In a murine colon carcinoma



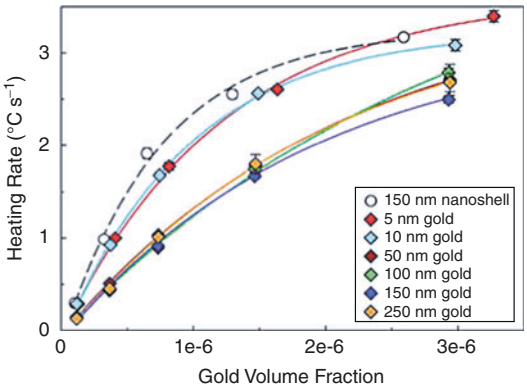
Thermal Cancer Ablation Therapies Using Nanoparticles, Fig. 2 *Left column* – Transmission electron micrographs of 10-nm diameter solid gold nanoparticles (*top*) and 150-nm diameter gold nanoshells (*bottom*). TEM images are ×20,000 magnification and ×500,000 (*inset*) magnification, and scale bars = 1000 nm (main) and 100 nm (*inset*). *Right column* – Diameter-dependent visible spectra of solid gold nanoparticles (*black* = 10 nm, *red* = 100 nm, *blue* = 150 nm) and the red-shifted gold nanoshell spectrum (*green* = 150 nm). Nanoparticle illustrations above each corresponding spectra are drawn to scale. Note, spectra have been normalized, and background subtracted to ease viewing of relative peak positions

subcutaneous tumor model, it was demonstrated that intravenous injection of polyethylene glycol-coded gold-silica nanoshells followed 6 h later by illumination of the tumor with a diode laser (8.0 nm, 4 W/cm^2, 3 min) led to complete destruction of the malignant tumors. Animals treated with laser treatment alone without prior intravenous injection of the nanoshells continued to grow at a rate similar to control animals who received neither nanoshell injection or laser treatment. The authors point out that this is a nontargeted delivery of the gold-silica nanoshells to tumors, and the size of the nanoparticles (130–150 nm) corresponds to the size of the enlarged pores and fenestrations in the neovasculature of tumors. The gold-silica nanoshells circulate and a significant number arrest in these pores, thus allowing thermal destruction of the vasculature of the tumor under laser irradiation. A subsequent study of prostate cancer cells using 110-nm gold nanoshells injected intravenously demonstrated that there was a nanoshell dose-related response. The tumors were again treated with a near-infrared laser at 4 W/cm^2 for 3 min. Tumors treated with laser therapy alone continued to grow at a rate similar to untreated control tumors, while tumors treated at the highest dose of gold nanoshells were over 93 % necrotic following a single laser treatment. A clinical trial using gold-silica nanoshell hyperthermia following NIR light exposure has been initiated for patients with oropharyngeal malignancies. Thermal therapy with NIR energy is limited to the treatment of superficial tumors (less than 1–2-cm deep) due to the significant attenuation of NIR light by biologic tissues.

Iron oxide nanoparticles are also being evaluated to produce alternating current magnetic field-based hyperthermia in tumors, but the limited thermal enhancement of these nanoparticles in a magnetic field requires extremely high concentrations of iron oxide [9]. Human breast carcinomas implanted into mice underwent intratumoral injection of colloidal suspensions of iron oxide particles 10–200 nm in diameter. Twenty minutes after the injection, the animals were exposed to an alternating current magnetic field (6.5kA/m; frequency, 400 kHz) for 4 min. Real-time temperature monitoring revealed that temperatures increased from 12 °C to 73 °C at various areas in the tumor. Subsequent analysis of the tumor showed some areas of coagulative necrosis, but other areas of viable tumor remained. Subsequent studies have attempted to address the issue of variable distribution of the superparamagnetic iron oxide nanoparticles through different regions of the tumor. Tumor cells near blood vessels are more difficult to treat with the magnetic nanoparticles because of the heat-sink phenomena of heat being carried away by blood flow in adjacent vessels coupled with the diffusion of nanoparticles in the tumor microenvironment into the blood vessels. One of the principle issues and limitations is that the iron oxide particles must be injected directly into the tumor. While this allows high concentrations of the magnetic nanoparticles for subsequent magnetic field thermal treatment, it is an invasive technique and allows for only treatment of tumors that can be identified through imaging techniques or during an operation. The magnetic particles can be used as imaging vehicles for tumors, but there are problems with diffusion of the nanoparticles out of the tumor and into the surrounding normal tissues. These high concentrations and the difficulty targeting the iron oxide nanoparticles to only cancer cells lead to destruction of normal (nonmalignant) cells surrounding the tumor, and diffusion of nanoparticles out of some areas of the tumor may cause insufficient heating to produce thermal cytotoxicity.

It was recently demonstrated that gold or carbon nanoparticles and a confined, noninvasive shortwave RF field produces significant local heat release that can be used to produce thermal cytotoxicity in cancer cells in vitro and in vivo with RF field treatments of 2 min or less [10, 11]. The heat release from gold nanoparticles exposed to shortwave RF fields is particularly impressive and is dependent on both the concentration and diameter of gold nanoparticles (Fig. 3). Solid gold nanoparticles smaller in diameter (5–10 nm) heated at a more rapid rate than gold nanoparticles 50–250 nm in diameter.

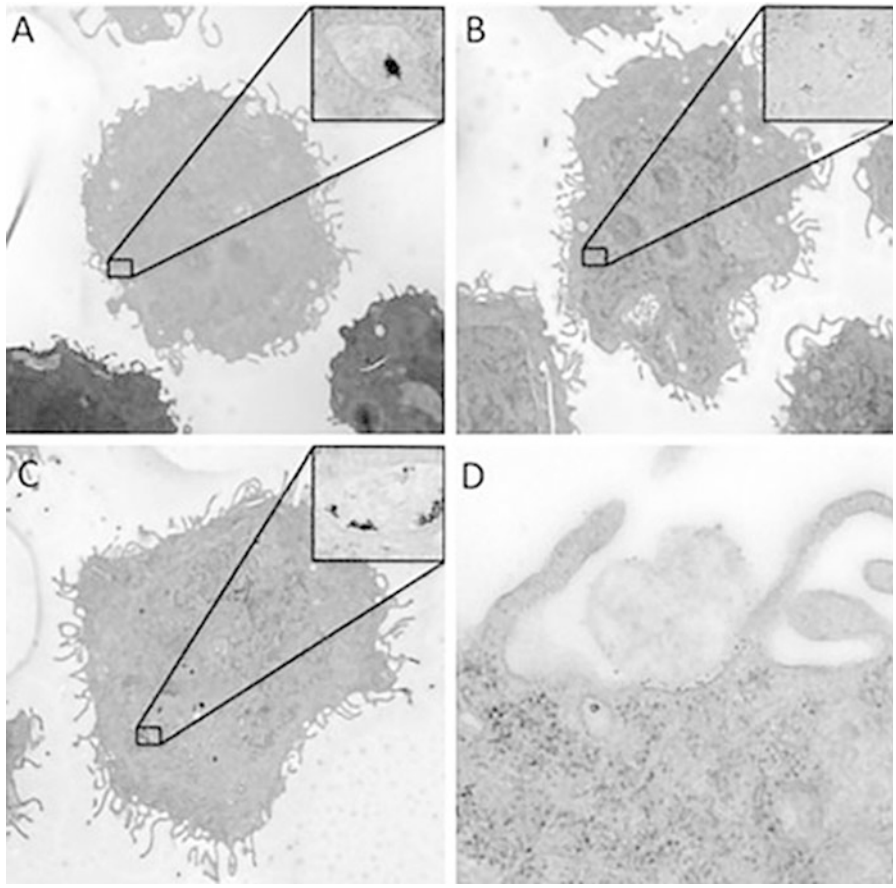
In addition, this method does not suffer the limitations of other nanoparticle-based



Thermal Cancer Ablation Therapies Using Nanoparticles, Fig. 3 At gold nanoparticle concentrations less than 10 ppm, the heating rate of deionized water in a 13.56 MHz RF field was almost 3 °C/s. This remarkable heating rate at very low concentrations of gold nanoparticles represents a heating efficiency of > 30,000 W/g of nanoparticles and indicates the generation of elevation temperatures in the nanoscale environment surrounding the gold nanoparticles sufficient to denature proteins, melt lipid bilayers, and produce irreparable damage to intracellular structures and organelles

hyperthermia systems given that shortwave radiofrequency (RF) energy has excellent deep tissue penetration, low tissue specific absorption rates (SAR), and documented safety for brief exposures of humans to this form of electromagnetic radiation [12]. Concerns have been expressed about possible mutagenic alterations in cells exposed to nonionizing electromagnetic radiation in the GHz range. However, studies of shortwave RF from 10 to 100 MHz has not been associated with any cytogenetic or cytotoxic damage [13]. Exposure of volunteer test subjects to prolonged periods of shortwave RF irradiation can produce minimal thermoregulatory effects (vasodilation, sweating) but has not been found to cause any acute or long-term damage to normal cells and tissues. Thus, the finding that charged gold nanoparticles release heat in response to shortwave RF exposure for durations of seconds to a few minutes has important implications. The physical mechanism of this phenomenon is being investigated, and data suggests that conductive polarizations of metal nanoparticles may explain the high thermal efficiency and tremendous heat release at low concentrations of nanoparticles exposed to the RF field. Importantly, since tissue penetration by shortwave RF fields is excellent with low attenuation of the signal, it should be possible to treat tumors at almost any site in the body. The crucial component of this treatment will be successful delivery of adequate numbers of gold nanoparticles to cancer cells to produce intracellular temperatures of at least 55 °C following an RF field exposure of no more than a few minutes duration, while ideally limiting uptake of the nanoparticles by normal cells.

It has shown that by directly conjugating 5-nm gold nanoparticles to cetuximab to target epidermal growth factor receptors (EGFRs) expressed on cancer cells, there is enhanced uptake of the gold nanoparticles (Fig. 4). In contrast, cells expressing little or no EGFR should have no enhanced uptake of cetuximab-conjugated gold nanoparticles and, thus, should not be significantly affected by treatment in the RF field. Targeting gold nanoparticles approximately 5–10 nm in diameter to cancer cells to create RF-induced hyperthermic cytotoxicity has several advantages: gold nanoparticles are simple and inexpensive to synthesize; they are easily characterized due to the presence of a characteristic surface plasmon resonance band; their surface chemistry permits manipulation of charge and shape relatively easily; attaching cancer cell targeting molecules, including antibodies, peptides, or pharmacologic agents, is easily achieved; and they are biocompatible and not associated with any acute or chronic toxicities in preclinical studies [14]. In addition to producing significant heat upon exposure to shortwave RF fields, 5–10-nm gold charged nanoparticles easily penetrate through pores and fenestrations in the neovasculature of solid tumors. This method may be promising for focused local or regional application of a noninvasive RF field to treat primary or established metastatic malignant tumors overexpressing EGFR, but would be problematic for whole-body RF therapy to treat diffuse micrometastatic disease. Constitutive expression of EGFR at normal levels occurs throughout many tissues in the body, and the results in SN12PM6 renal cancer cells (low-level expression of EGFR) show that even with low levels of EGFR



Thermal Cancer Ablation Therapies Using Nanoparticles, Fig. 4 (**a**) Transmission electron microscope (*TEM*) image of Panc-1 human pancreatic adenocarcinoma cells (×7500 magnification) that were incubated for 30 min with unconjugated, nontargeted 5-nm gold nanoparticles (control group). Almost no gold nanoparticles are seen within the cytoplasm of the cell, and it is only at a magnification of ×100,000 (*inset upper right*) that an intracytoplasmic vesicle containing a small number of gold nanoparticles is seen. Incubation of Panc-1 cells for 60 or 120 min with the unconjugated gold nanoparticles also produced only a few scattered vesicles in the cytoplasm containing gold nanoparticles. (**b**) TEM image of Panc-1 cells incubated for 30 min with nonspecific IgG-conjugated gold nanoparticles (control group). Similar to the unconjugated gold nanoparticles, the nonspecific IgG-conjugated gold nanoparticles were not readily taken into the Panc-1 cells. The larger image (×7500 magnification) demonstrates few or no obvious intracytoplasmic vesicles containing gold nanoparticles. At higher magnification (×100,000, *inset upper right*), an occasional intracytoplasmic vesicle with a few gold nanoparticles is evident. (**c**) TEM image of Panc-1 pancreatic adenocarcinoma cells incubated for 30 min with C225-conjugated 5-nm gold nanoparticles. At 30 min, a number of intracytoplasmic vesicles (dense *black* vesicles) containing gold nanoparticles are evident in the larger image (×7500 magnification). At a higher magnification (×100,000), numerous vesicles containing dense collections of gold nanoparticles are found throughout the cytoplasm. At 60 and 120 min of incubation with the C225-targeted gold nanoparticles, the number of intracytoplasmic vesicles containing gold nanoparticles was seen to increase. (**d**) A higher magnification TEM image (×100,000) showing details of the cell surface of Panc-1 cells incubated for 30 min with C225-conjugated gold nanoparticles. Numerous gold nanoparticles are seen at the surface of the cell consistent with binding of the C225 to epidermal growth factor receptor (*EGFR*). Pretreatment of the cells with C225 alone completely blocked any binding of targeted gold nanoparticles to the cell surface and led to no gold nanoparticles appearing in intracytoplasmic vesicles 30, 60, or 120 min after incubation

expression, there is enough uptake of cetuximab-targeted gold nanoparticles to produce low levels of thermal apoptosis following exposure to the RF field. It may be possible to modulate or minimize effects on normal tissues expressing a target molecule like EGFR by use of brief duration pulsed RF therapy or repeated short duration exposures, but these effects must be studied in greater detail to confirm this theoretical consideration. Nonetheless, these results indicate that it is possible to use overexpression of a specific cell surface moiety, e.g., EGFR, to enhance delivery and uptake of gold nanoparticles to produce thermal cytotoxicity in the malignant cells following brief exposure to a noninvasive RF field.

There are numerous other possibilities for targeted delivery of nanoparticles into malignant cells. Radiolabeled carbon nanotubes have been conjugated with rituximab and lintuzumab to diagnose and potentially treat lymphoma cells [15]. Recently, aptamers, nucleic acid ligands much smaller than monoclonal antibodies, have been shown to target specific molecules in the neovasculature of tumors or on the surface of different types of cancer cells [16]. These aptamers have been conjugated to gold nanoparticles initially for use as an improved diagnostic technique. However, co-conjugation of cytotoxic agents, or use of our noninvasive RF field treatment technique, could be used as an actual anticancer therapy using this targeting technique. So-called cell-penetrating peptides, generally less than 100 amino acids in length, also have been shown to target certain types of cancer cells. The 86-amino acid HIV-1 Tat protein has been conjugated to gold nanoparticles, with subsequent rapid intracellular uptake of the nanoparticles and localization to the nucleus [17, 18]. The monoclonal antibody-conjugated gold nanoparticles used by our group are found in the cytoplasm of cells within 30 min of adding these nanoparticle conjugates to cancer cells in culture. Localization of the nanoparticles to the nucleus with thermal destruction of DNA following RF field exposure must be investigated as a technique to enhance cancer cell destruction. Chlorotoxin is another small peptide (36 amino acids) cancer cell agent that can be easily conjugated with gold nanoparticles and is currently being investigated as a cancer cell targeting agent in several human malignant cell lines [19]. Identification of cancer-specific ligands not expressed on normal cells will permit targeting of gold nanoparticles to only malignant cells for treatment in our noninvasive RF field generator and should allow treatment of measurable tumors and micrometastatic disease with minimal therapy-related toxicities. Targeting ligands, like EGFR, that are overexpressed but not unique to cancer cells are feasible but will require careful evaluation to determine acute and long-term thermal effects in normal cells and organs following RF field therapy.

Nanoscale thermal therapy of targeted cancer cells is a promising new weapon in the battle against cancer. By combining nonionizing radiation and targeted metallic nanoparticles, noninvasive targeted thermal therapy has the promise of benefiting cancer patients that need a nontoxic alternative to the debilitating effects of conventional chemotherapeutics. Using noninvasive RF fields, magnetic fields, or laser irradiation to heat nanoscale materials that are internalized into cancer cells or the tumor microvasculature has the promise of reducing toxicity and maximizing therapeutic benefit for cancer patients.

References

1. Beasley, G.M., Ross, M.I., Tyler, D.S.: Future directions in regional treatment strategies for melanoma and sarcoma. Int. J. Hyperthermia **24**, 301–309 (2008)
2. Stewart, J.H.T., Shen, P., Russell, G., Fenstermaker, J., McWilliams, L., Coldrun, F.M., Levine, K.E., Jones, B.T., Levine, E.A.: A phase I trial of oxaliplatin for intraperitoneal hyperthermic chemoperfusion for the treatment of peritoneal surface dissemination from colorectal and appendiceal cancers. Ann. Surg. Oncol. **15**, 2137–2145 (2008)
3. Kampinga, H.H.: Cell biological effects of hyperthermia alone or combined with radiation or drugs: a short introduction to newcomers in the field. Int. J. Hyperthermia **22**, 191–196 (2006)
4. Arciero, C.A., Sigurdson, E.R.: Diagnosis and treatment of metastatic disease to the liver. Semin. Oncol. **35**, 147–159 (2008)
5. Hartman, K.B., Wilson, L.J., Rosenblum, M.G.: Detecting and treating cancer with nanotechnology. Mol. Diagn. Ther. **12**, 1–14 (2008)

6. Neumaier, C.E., Baio, G., Ferrini, S., Corte, G., Daga, A.: MR and iron magnetic nanoparticles. Imaging opportunities in preclinical and translational research. Tumori **94**, 226–233 (2008)
7. Misra, R.D.: Quantum dots for tumor-targeted drug delivery and cell imaging. Nanomedicine **3**, 271–274 (2008)
8. Gobin, A.M., Lee, M.H., Halas, N.J., James, W.D., Drezek, R.A., West, J.L.: Near-infrared resonant nanoshells for combined optical imaging and photothermal cancer therapy. Nano Lett. **7**, 1929–1934 (2007)
9. Kalambur, V.S., Longmire, E.K., Bischof, J.C.: Cellular level loading and heating of superparamagnetic iron oxide nanoparticles. Langmuir **23**, 12329–12336 (2007)
10. Gannon, C.J., Cherukuri, P., Yakobson, B.I., Cognet, L., Kanzius, J.S., Kittrell, C., Weisman, R.B., Pasquali, M., Schmidt, H.K., Smalley, R.E., Curley, S.A.: Carbon nanotube-enhanced thermal destruction of cancer cells in a noninvasive radiofrequency field. Cancer **110**, 2654–2665 (2007)
11. Gannon, C.J., Patra, C.R., Bhattacharya, R., Mukherjee, P., Curley, S.A.: Intracellular gold nanoparticles enhance non-invasive radiofrequency thermal destruction of human gastrointestinal cancer cells. J. Nanobiotechnol. **6**, 2 (2008)
12. Adair, E.R., Blick, D.W., Allen, S.J., Mylacraine, K. S., Ziriax, J.M., Scholl, D.M.: Thermophysiological responses of human volunteers to whole body RF exposure at 220 MHz. Bioelectromagnetics **26**, 448–461 (2005)
13. Klima, J., Scehovic, R.: The field strength measurement and SAR experience related to human exposure in 110 MHz to 40 GHz. Meas. Sci. Rev. **6**, 40–44 (2006)
14. Daniel, M.C., Astruc, D.: Gold nanoparticles: assembly, supramolecular chemistry, quantum-size-related properties, and applications toward biology, catalysis, and nanotechnology. Chem. Rev. **104**, 293–346 (2004)
15. McDevitt, M.R., Chattopadhyay, D., Kappel, B.J., Jaggi, J.S., Schiffman, S.R., Antczak, C., Njardarson, J.T., Brentjens, R., Scheinberg, D.A.: Tumor targeting with antibody-functionalized, radiolabeled carbon nanotubes. J. Nucl. Med. **48**, 1180–1189 (2007)
16. Javier, D.J., Nitin, N., Levy, M., Ellington, A., Richards-Kortum, R.: Aptamer-targeted gold nanoparticles as molecular-specific contrast agents for reflectance imaging. Bioconjug. Chem. **19**, 1309–1312 (2008)
17. Berry, C.C.: Intracellular delivery of nanoparticles via the HIV-1 tat peptide. Nanomed. **3**, 357–365 (2008)
18. Berry, C.C., de la Fuente, J.M., Mullin, M., Chu, S.W., Curtis, A.S.: Nuclear localization of HIV-1 tat functionalized gold nanoparticles. IEEE Trans. Nanobioscience **6**, 262–269 (2007)
19. Mamelak, A.N., Jacoby, D.B.: Targeted delivery of antitumoral therapy to glioma and other malignancies with synthetic chlorotoxin (TM-601). Expert Opin. Drug Deliv. **4**, 175–186 (2007)

Thermal Chemical Vapor Deposition

▶ Chemical Vapor Deposition (CVD)

Thermal Conductance

▶ Thermal Conductivity and Phonon Transport

Thermal Conductivity

▶ Heat Transfer in Semiconductor Nanostructures

Thermal Conductivity and Phonon Transport

Yee Kan Koh
Department of Mechanical Engineering, National University of Singapore, Singapore, Singapore

Synonyms

Heat conduction; Heat conductivity; Nanoscale heat transport; Thermal conductance; Thermal resistance; Thermal resistivity; Thermal transport

Definition

Thermal conductivity (symbols: κ or Λ) is an empirical material property that relates local heat flux (J) across a real or imaginary surface to the temperature gradient ($\partial T/\partial x$) at the same point in direction x normal to the surface. Normally, the concept of thermal conductivity is used for heat conduction not mediated by massive advection of atoms and molecules, or propagation of electromagnetic waves; such heat transport is traditionally referred to as convective and radiative heat transfer, respectively.

Thermal conductivity is defined according to the Fourier's law of heat conduction given below,

$$J = -\kappa \frac{\partial T}{\partial x}$$

The negative sign indicates that heat flow against temperature gradient, i.e., from the hot side to the cold side. The units commonly used for thermal conductivity include watts per meter-Kelvin (W m^{-1} K^{-1}) and watts per centimeter-Kelvin (W cm^{-1} K^{-1}). The reciprocal of thermal conductivity is called the thermal resistivity.

Overview

As summarized by the second law of thermodynamics, Nature seeks to approach equilibrium in an isolated system, be it a mechanical, chemical, or thermal system. In the context of heat transfer, heat flows from the hot side to the cold side whenever the system is not in thermal equilibrium (i.e., temperature is not uniform), until all temperature difference is ultimately eliminated. Although the second law of thermodynamics is very general and applies to any thermal system, it does not however stipulate how equilibrium is reached. In other words, although heat transfer occurs whenever temperature difference exists according to the second law of thermodynamics, the rate of heat transfer differs from systems to systems depending on the properties of materials or mediums in the thermal system.

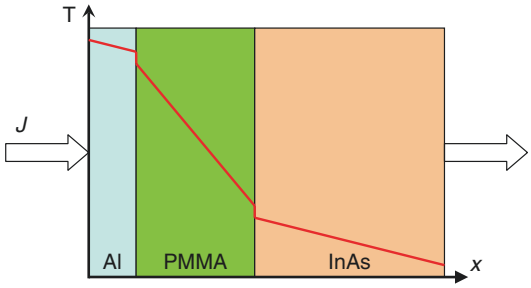
Historically, heat transfer is classified as conduction, convection, and radiation, depending on the mediums involved and the mechanisms in which heat is transported across the mediums. Conduction usually refers to heat transfer in solids, in which atoms and molecules are practically immobile and diffusion of atoms and molecules is relatively slow (as long as the solids are not at highly elevated temperatures). In such cases, heat is predominantly carried by vibration of atoms, or other quasiparticles in the crystals such as electrons and holes (see more discussions on quasiparticles below). Also, in still fluids, heat transfer is due to vibration and collisions of atoms and molecules, and thus is also categorized as conduction. On the other hand, the term convection is mainly used for heat transfer from a solid surface to fluids, mediated by diffusion or advection (macroscopic motion) of molecules and atoms. In convection, fluids are driven either by external means such as pumps (forced convection) or by buoyancy forces (natural convection). Lastly, when heat is carried by electromagnetic waves, the heat transfer is called radiation. Unlike conduction and convection, no mediums are required for radiation, since electromagnetic waves can travel through vacuum. For example, heat is transported from the Sun to the Earth by visible or infrared light through radiation. Of course, radiation could also occur through transparent mediums (e.g., glass, air etc.).

For heat conduction, the rate of thermal energy transfer per unit time and unit area is proportional to temperature gradient in the respective direction. This empirical relationship is known as the Fourier's law of heat conduction. Mathematically,

$$J = -\kappa \frac{\partial T}{\partial x}$$

where J is the heat flux across a real or imaginary interface (SI unit: W m^{-2}) and $\partial T/\partial x$ is the local temperature gradient (SI unit: K m^{-1}). Thermal conductivity (κ) is defined as the proportional constant in the Fourier's law of heat conduction and has the SI unit of W m^{-1} K^{-1}. Thermal conductivity for fully condensed solids spans for more than five orders of magnitude, from $\kappa > 1{,}000$ W m^{-1} K^{-1} for carbon allotropes [1, 2], to $\kappa < 0.01$ W m^{-1} K^{-1} for disordered crystals [3]. Theoretically, thermal conductivity is an intrinsic material property and thus should not depend on the geometry and size of the materials. This theoretical definition of thermal conductivity is somewhat relaxed recently for the thermal conductivity of nanostructures. Thermal conductivity of nanostructures depends on the characteristic lengths of the nanostructures, see the discussions below.

The effect of thermal conductivity on heat conduction in solids is illustrated in Fig. 1. For a



Thermal Conductivity and Phonon Transport, Fig. 1 Schematics of a hypothetical multilayered solid of aluminum (Al) with $\kappa > 200$ W m^{-1} K^{-1}, amorphous polymer PMMA with $\kappa \approx 0.2$ W m^{-1} K^{-1}, and semiconductor indium arsenide (InAs) with $\kappa \approx 27$ W m^{-1} K^{-1}, being heated uniformly with a constant heat flux of J. The red solid line is the temperature profile in the hypothetical solid. Different slopes of the temperature profile are due to different thermal conductivity of each layer. Temperature drops at the interfaces are due to finite Kapitza thermal conductance of the interfaces

constant heat flux applied on a hypothetical multilayered solid as in Fig. 1, temperature drop across each layer is inversely proportional to the thermal conductivity of the layer. For a high thermal conductivity material (e.g., aluminum), the temperature drop is minimal since heat could be easily transported across the material. In other words, thermal energy could easily spread across high thermal conductivity materials, resulting in a near-uniform temperature distribution. On the other hand, low thermal conductivity impedes heat transport, and thus higher temperature difference could be maintained in the materials. This simple understanding of thermal conductivity is crucial for many modern applications (such as thermal barrier coatings and thermoelectrics), in which high temperature differences enhance the performance of the devices. Researchers in these fields seek to modify the existing materials, often through nanostructuring, to reduce the thermal conductivity so that a higher temperature difference could be maintained in the devices.

While the thermal conductivity is a bulk property, interfacial or Kapitza thermal conductance [4] (symbol: G) is a property of interfaces. As heat flows across an interface, a finite temperature drop is observed at the interface, as illustrated in Fig. 1, due to resistance of the interface to thermal transport. Kapitza thermal conductance (G) relates this temperature drop (ΔT) to heat flux across the interface; $J = -G\,\Delta T$. Typically, thermal conductance ranges from $G > 1$ GW m^{-2} K^{-1} for metal/metal interfaces [5] to $G \approx 10$ MW m^{-2} K^{-1} for interfaces with highly dissimilar materials [6]. It is important to realize that the Kapitza thermal conductance is not solely due to imperfections of interfaces (e.g., imperfect contact as a result of surface roughness, diffuse interfaces due to intermixing, etc.). Even for atomically perfect interfaces, heat flow is still impeded by finite thermal conductance because differences in the properties (e.g., density of phonon states) of two mediums reduce the probability for heat carriers to transmit across the interfaces. In other words, since interfaces break the reflection and translational symmetry, even perfect interfaces are imperfections by definition. Thus, heat transport is retarded at interfaces due to this break of symmetry engendered by the interfacial imperfections.

Heat Transport by Phonons in Nonmetallic Solids

In a microscopic view, heat conduction in solids is due to quasiparticles that carry heat across the solids. In simple words, quasiparticles (especially those important for heat conduction in solids) are quantized excitation modes resulting from quantum-mechanical interaction of atoms and electrons in the periodic lattices. Strictly speaking, some of these quasiparticles are only theoretically defined for perfect crystals, but this restriction is usually relaxed for imperfect crystals. Quasiparticles behave like real particles with properties and behavior modified from their real particle counterparts.

Common quasiparticles that are important for heat conduction in solids include phonons, electron quasiparticles and holes. Phonons are quantized lattice vibration modes due to interaction of sound waves in solids and are responsible for heat conduction in most nonmetallic solids (dielectrics and semiconductors). Electron quasiparticles are quantized excitation modes due to interaction between electrons and periodic electrical potential

in the lattices. Electron quasiparticles are the main carriers for both heat and charges in metals, and as a result of electron quasiparticles being the common carriers, the thermal conductivity and electrical conductivity of metals are related by a fundamental relationship called the Wiedemann–Franz law. Electron quasiparticles also contribute significantly to heat conduction in highly *n*-doped semiconductors and semimetals. Holes, on the other hand, are empty states available in the crystals that are not filled by electrons. Holes behave similar to electron quasiparticles, except that they carry positive charges and have different effective masses. Holes can contribute significantly to heat conduction in highly *p*-doped semiconductors and semimetals. Note that electron quasiparticles have different behavior (e.g., different effective masses) from stand-alone electrons, even though both carry negative charges and both are fermions. Consistent with the literature and for simplicity, electron quasiparticles are thereafter referred to as "electrons" throughout the text.

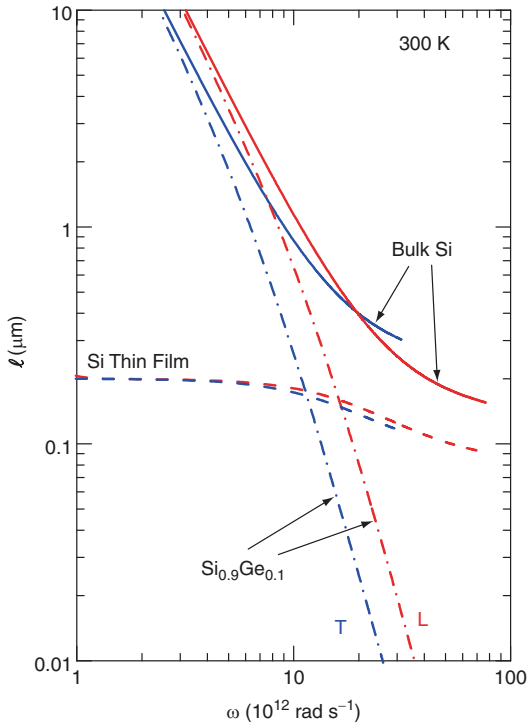
To understand heat conduction by phonons in crystalline dielectrics and semiconductors, it is important to realize that phonons do not carry same amounts of heat due to different polarizations (longitudinal or transverse) and a broad distribution of frequencies (ω) and lifetimes (τ) of phonons. Heat conduction by phonons depends on three properties of phonons:

1. The volumetric heat capacity (C) of phonon modes, which affects the amount of heat carried by phonons. The volumetric heat capacity is a function of phonon frequency, polarization, and temperature. At high temperatures, the heat capacity of each phonon mode approaches a classical value of k_B, where $k_B = 1.38 \times 10^{-23}$ J K^{-1} is the Boltzmann constant.
2. The group velocity (υ) of phonons, which determines how fast phonons propagate in the materials and is defined as

$$\upsilon = \frac{\partial \omega}{\partial k}$$

where k is the wavenumber of phonons. The group velocity depends on the slope of phonon dispersion and thus is a function of phonon frequency (ω) and polarization. At low frequencies, the group velocity of acoustic phonons approach the speed of sounds of the materials, while at high frequencies, the group velocity of acoustic phonons approach zero due to interference of reflected waves near the Brillouin zone boundaries.
3. The mean-free-path (ℓ) of phonons, which is the average distance phonons propagate without being scattered. In crystals, phonons are scattered by imperfections including impurities, anharmonicity, and grain boundaries. This scattering of phonons is the source of retardation for heat conduction in crystals, since perfect crystals have infinitely high thermal conductivity. Normally, the scattering process is quantified by the relaxation time (τ) or the mean-free-path (ℓ) of phonons; the relaxation time and mean-free-path are related by $\ell = \upsilon\tau$, where υ is the group velocity of phonons. Both the relaxation time and the mean-free-path strongly depend on the frequency of phonon modes.

At high temperatures, most phonon modes are excited, but heat is carried mainly by a small fraction of these excited modes (long-wavelength acoustic phonons). Acoustic phonons near the edge and the center of the Brillouin zone, and optical phonons are inefficient in carrying heat due to their low group velocity or heat capacity. The mean-free-paths of these heat-carrying acoustic phonons, however, are not constant but span more than an order of magnitude in most crystals depending on phonon frequency. For example, calculations [7] show that ≈ 90 % of heat could be carried by phonons with $0.1 \leq \ell(\omega) \leq 1$ μm in Si. The mean-free-paths of phonons are also sensitive to the purity and harmonicity of materials, since imperfections, such as impurities, point defects and anharmonicity, scatter phonons. In fact, the disparity in the thermal conductivity of different materials is mainly due to differences in the mean-free-paths (instead of the heat capacity or the group velocity) of phonons in these



Thermal Conductivity and Phonon Transport, Fig. 2 Mean-free-path (ℓ) of longitudinal (*red*) and transverse (*blue*) acoustic phonons in bulk Si (*solid lines*), a 200 nm Si thin film (*dashed lines*) and bulk $Si_{0.9}Ge_{0.1}$ alloy (*dash-dotted lines*) calculated from a Callaway-type model [8]. ω is the frequency of phonons. The dominant scattering mechanisms in bulk Si, the SiGe alloy, and the Si thin film are three-phonon process (anharmonicity), Rayleigh scattering, and boundary scattering, respectively

materials, since the heat capacity and phonon velocity of different materials do not vary by more than an order of magnitude.

Despite the importance, understanding of the mean-free-paths of phonons is still incomplete, mainly due to lack of a direct method to measure the mean-free-paths of phonons [8]. The mean-free-paths of phonons are usually inferred from systematic measurements of the thermal conductivity by comparing the experimental results to theoretical calculations. Figure 2 shows an example of calculations of the mean-free-paths of acoustic phonons in bulk silicon, a 200 nm silicon thin film and bulk $Si_{0.1}Ge_{0.9}$ alloy, as a function of phonon frequency. Note that different imperfections are efficient in scattering phonons of different frequency. At room temperature, phonons in bulk Si are predominantly scattered by anharmonicity, which is effective to scatter high-frequency phonons but not low-frequency phonons, see Fig. 2. Similarly, Rayleigh scattering, which is the dominant scattering mechanism for phonons in alloys, is more effective to scatter high-frequency phonons than low-frequency phonons. On the other hand, boundary scattering is effective to scatter low- and medium-frequency phonons. Over the past decade, nanostructures, such as thin films, superlattices, and nanowires, have been widely used to scatter the low-energy phonons to further reduce the thermal conductivity. It is stressed that the scattering rate and mean-free-paths of phonons in most materials are *not* known – calculations as in Fig. 2 are at best just an approximation. In fact, polarization of the dominant heat carriers in silicon was determined as recent as a decade ago [9].

Thermal Conductivity Reduction in Nanostructures

Over the past decade, researchers advanced the understanding of heat conduction in nanostructures through systematic measurements of the thermal conductivity of nanostructures. This understanding on nanoscale heat transport facilitates the design of more efficient thermoelectric materials with reduced thermal conductivity. In this section, prior experimental studies on heat conduction by phonons in different nanostructures, including crystalline thin films, superlattices, nanowires, and semiconductor embedded with nanoscale precipitates, are briefly and selectively review.

Goodson and coworkers [9, 10] reported the in-plane thermal conductivity of Si thin films over a wide film thickness (74–1.6 μm) and temperature (20–320 K) range. The in-plane thermal conductivity was measured by the 3ω method [11], which is discussed below. Heat was previously perceived to be carried by transverse acoustic phonons in Si, but Ju and Goodson [9] find from the in-plane thermal conductivity measurements of Si thin films that longitudinal acoustic phonons

with mean-free-path $\ell \approx 300$ nm are the dominant carriers in Si at 300 K.

Yao [12] first reported the cross-plane thermal conductivity of AlAs/GaAs superlattices, measured using an ac calorimetric method. He finds that the thermal conductivity of a $(AlAs)_{5\ nm}(GaAs)_{5\ nm}$ superlattice approaches the thermal conductivity of AlGaAs alloys at room temperature. The observation that the thermal conductivity of superlattices approaches the thermal conductivity of the alloys is not universal. For example, Lee and coworkers [13] reported the thermal conductivity of fully strained Si/Ge superlattices, measured by the 3ω method. They find that for period <6 nm, the thermal conductivity decreases with decreasing period and is governed by the thermal conductance of interfaces. For longer period of >10 nm, however, the defect density is so high that the thermal conductivity approaches the amorphous limit. Further measurements [14, 15] on other Si/Ge nanostructures are consistent with measurements by Lee et al.

Venkatasubramanian [16] reported the thermal conductivity of Bi_2Te_3/Sb_2Te_3 superlattices as a function of superlattice period and find that the thermal conductivity of short-period superlattice is minimum when the period is ≈ 4 nm. Subsequent measurements on similar samples by Touzelbaev and coworkers [17], however, do not present similar minimum thermal conductivity. The minimum thermal conductivity is also not observed on AlN/GaN superlattices [18], in which the AlN/GaN interfaces are atomically sharp and chemically abrupt.

Li and coworkers [19] first measured the thermal conductivity of Si nanowires using a microfabricated device. Li et al. find that the thermal conductivity of nanowires is reduced by more than an order of magnitude when the diameter of the nanowire is 22 nm. The approach has subsequently been applied to study the thermal conductivity of other nanostructures, including rough Si nanowires [20]. The reduction of thermal conductivity, in most cases, is due to enhanced boundary scattering.

Harman and coworkers [21] estimated the thermal conductivity of PbTe/PbSe nanodot superlattices from their thermoelectric devices. They found that the thermal conductivity is 0.33 W m^{-1} K^{-1}, a factor of 3 lower than the thermal conductivity of the corresponding alloys. More complete and systematic measurements [22] on similar materials, however, indicate that the thermal conductivity of PbTe/PbSe nanodot superlattices is on the order of 1 W m^{-1} K^{-1}, comparable with the thermal conductivity of PbTeSe alloys.

The thermal conductivity of InGaAs with ErAs nanoparticles randomly distributed in the InGaAs matrix was first reported by Kim and coworkers [23]. The thermal conductivity of the nanostructured materials is found to be reduced from the thermal conductivity of the alloys by up to a factor of 2, even when the concentration of the ErAs nanoparticles is relatively small (0.3 %). This reduction of thermal conductivity is explained by enhanced Rayleigh scattering when the size of the nanoparticles better match the wavelength of phonons. Subsequent measurements [24] on InGaAs with higher concentration of ErAs nanoparticles indicate that the thermal conductivity could be further reduced with higher concentration of ErAs nanoparticles.

Modern Techniques for Measurements of the Thermal Conductivity of Nanostructures

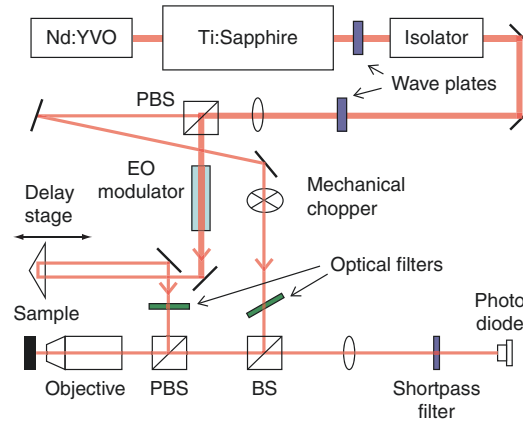
For measurements of heat conduction on submicron or nanometer length scales, temperature profile has to be monitored within the similar characteristic length scales (~1 μm) or equivalently the temperature decay profile in nanosecond time scales. Conventional approaches to measure the thermal conductivity are incapable to reach these small length and time scales, and thus are not suitable for measurements of the thermal conductivity of nanostructures. In this section, two important techniques that are capable of characterizing the thermal properties of nanostructures, namely, the 3ω method and the time-domain thermoreflectance (TDTR), are introduced.

The 3ω method [11] is widely used to measure the thermal conductivity of bulk solids and thin

films. In 3ω measurements, a narrow metal line (often Au or Pt with Cr or Ti as the adhesion layer) of ~400 nm thickness and ~30 μm width is patterned on a sample. Electrical current $\boldsymbol{i}$ of frequency ω is applied to the metal line with electrical resistance $\boldsymbol{r}$, generating joule heating of $\boldsymbol{i}^2\boldsymbol{r}$ within the metal line with a frequency component at 2ω. As a result of this oscillating heat source, a temperature oscillation and a corresponding resistance oscillation at frequency 2ω are induced in the metal line. Hence, a component of the voltage oscillation ($\boldsymbol{v} = \boldsymbol{ir}$) across the metal line contains a third harmonic, 3ω. The thermal conductivity of the sample can be deduced from this 3ω voltage oscillation.

The cross-plane thermal conductivity of thin films can be measured by the differential 3ω method [16]. A reference sample without the thin film of interest is prepared simultaneously with the sample containing the film of interest such that the metal line patterned on both samples has the same thickness and width. Both the thin film sample and the reference are measured using similar heating power and the same range of heater frequencies. The temperature drop across the thin film ΔT_f can then be derived from the difference in the amplitude of the temperature oscillation of the sample and the reference. Assuming one-dimensional (1D) heat conduction, the cross-plane thermal conductivity of the thin film is derived from ΔT_f. If the thin film is semiconducting, a thin dielectrics layer is required to electrically insulate the metal lines from the thin films. For semiconducting crystalline thin films, the differential 3ω method is usually only suitable to measure the thermal conductivity of the films if the films are >1 μm thick [24].

Another important technique for measurements of heat conduction on nanometer length scale is the time-domain thermoreflectance (TDTR) [25]. Figure 3 shows a schematic diagram [26] of the TDTR setup. Before TDTR measurements, the samples are usually coated with a thin layer (~80 nm) of metal with high thermoreflectance (e.g., Al) by magnetron sputtering or thermal evaporation. The metal film serves as a transducer to absorb the heating pump beam and to convert the temperature excursions at the surface into changes in the intensity of the reflected probe beam. Usually, Al is used as the metal transducer due to the high thermoreflectance. However, for high temperature measurements, Al is not suitable due to its low melting point; Pt, Ta, or AuPd could be used for high temperature measurements, see [27] for further discussion. The thickness of the metal transducer layer is normally simultaneously determined during TDTR measurements by picosecond acoustics.



Thermal Conductivity and Phonon Transport, Fig. 3 A schematic [26] of the TDTR setup. The optical filters and the shortpass filter are optional and can eliminate the artifacts due to diffusely scattered pump beam leaked into the Si photodiode during TDTR measurements of rough samples (Reprinted with permission from Ref. [26]. Copyright 2008, American Institute of Physics)

In TDTR measurements, the output of a mode-locked Ti:sapphire laser oscillator is split into a pump beam and a probe beam, with the relative delay time between the pump and probe pulses being adjusted via a mechanical stage. Samples are heated by the pump beam, which is modulated by an electro-optic modulator at frequency f, $0.1 < f < 10$ MHz. Cooling of the surface after being heated by pump pulses is then monitored through changes in the intensity of the reflected probe beam using a Si photodiode and a radio-frequency (rf) lock-in amplifier. If the samples are rough, TDTR measurements should be performed using a two-tint configuration [26], see Fig. 3. Normally, heat waves only penetrate <1 μm in TDTR measurements, while the $1/e^2$ radius of the

laser beam is usually >5 μm. Hence, heat flow in TDTR is predominantly one dimensional.

Data analysis in TDTR is considerably complicated. The changes in reflected intensity at frequency f have both an in-phase V_{in} and out-of-phase V_{out} components. Normally, the ratio V_{in}/V_{out} is used for the analysis (instead of just the *Vin* component) to make use of the additional information in the out-of-phase signal and eliminate artifacts created by unintended variations in the diameter or position of the pump beam created by the optical delay line. Measurements of the ratios V_{in}/V_{out} as a function of delay time are compared to numerical solutions of the diffusion equation in cylindrical coordinates using a thermal model [28]. The thermal model normally has two free parameters: the thermal conductance of the Al/sample interface and Λ of the sample. For most cases, these two parameters can be separated from the fitting of the model calculations to the measurements [29].

Future Directions for Research

Detailed understanding of heat transport in crystalline solids, especially on how phonons are scattered in nanostructures, is crucial for the design of new thermoelectric materials and thermal management of electronic devices. In this regard, future research will focus on experimental and theoretically studies on the details of how phonons carry heat across different kinds of nanostructures. One example is recent efforts by researchers in the field to investigate how heat is transported along [1, 2] and across [30] monolayer and few-layer graphenes. By carefully comparing the experimental results to theoretical calculations, these studies advance our knowledge on how phonons are scattered by impurities and nanostructures and thus facilitate future design of more efficient devices.

Cross-References

- ▶ Finite Element Methods for Computational Nano-optics
- ▶ Nanostructured Thermoelectric Materials
- ▶ Nanostructures for Energy

References

1. Balandin, A.A., et al.: Nano Lett. **8**, 902 (2008)
2. Seol, J.H., et al.: Science **328**, 213 (2010)
3. Chiritescu, C., et al.: Science **315**, 351 (2007)
4. Swartz, E.T., Pohl, R.O.: Rev. Mod. Phys. **61**, 605 (1989)
5. Gundrum, B.C., Cahill, D.G., Averback, R.S.: Phys. Rev. B **72**, 245426 (2005)
6. Lyeo, H.-K., Cahill, D.G.: Phys. Rev. B **73**, 144301 (2006)
7. Koh, Y.K.: Ph.D. thesis, University of Illinois at Urbana-Champaign (2010)
8. Koh, Y.K., Cahill, D.G.: Phys. Rev. B **76**, 075207 (2007)
9. Ju, Y.S., Goodson, K.E.: Appl. Phys. Lett. **74**, 3005 (1999)
10. Asheghi, M., Leung, Y.K., Wong, S.S., Goodson, K. E.: Appl. Phys. Lett. **71**, 1798 (1997)
11. Cahill, D.G., Pohl, R.O.: Phys. Rev. B **35**, 4067 (1987)
12. Yao, T.: Appl. Phys. Lett. **51**, 1798 (1987)
13. Lee, S.-M., Cahill, D.G., Venkatasubramanian, R.: Appl. Phys. Lett. **70**, 2957 (1997)
14. Huxtable, S.T., Abramson, A.R., Tien, C.-L., Majumdar, A., LaBounty, C., Fan, X., Zeng, G., Bowers, J.E., Shakouri, A., Croke, E.T.: Appl. Phys. Lett. **80**, 1737 (2002)
15. Yang, B., Liu, J.L., Wang, K.L., Chen, G.: Appl. Phys. Lett. **80**, 1758 (2002)
16. Venkatasubramanian, R.: Phys. Rev. B **61**, 3091 (2000)
17. Touzelbaev, M.N., Zhou, P., Venkatasubramanian, R., Goodson, K.E.: J. Appl. Phys. **90**, 763 (2001)
18. Koh, Y.K., Cao, Y., Cahill, D.G., Jena, D.: Adv. Funct. Mater. **19**, 610 (2009)
19. Li, D., Wu, Y., Kim, P., Shi, L., Yang, P., Majumdar, A.: Appl. Phys. Lett. **83**, 2934 (2003)
20. Hochbaum, A.I., Chen, R., Delgado, R.D., Liang, W., Garnett, E.C., Najarian, M., Majumdar, A., Yang, P.: Nature **451**, 163 (2008)
21. Harman, T.C., Taylor, P.J., Walsh, M.P., LaForge, B. E.: Science **297**, 2229 (2002)
22. Koh, Y.K., Vineis, C.J., Calawa, S.D., Walsh, M.P., Cahill, D.G.: Appl. Phys. Lett. **94**, 153101 (2009)
23. Kim, W., Zide, J., Gossard, A., Klenov, D., Stemmer, S., Shakouri, A., Majumdar, A.: Phys. Rev. Lett. **96**, 045901 (2006)
24. Koh, Y.K., Singer, S.L., Kim, W., Zide, J.M.O., Lu, H., Cahill, D.G., Majumdar, A., Gossard, A.C.: J. Appl. Phys. **105**, 054303 (2009)
25. Paddock, C.A., Eesley, G.L.: J. Appl. Phys. **60**, 285 (1986)
26. Kang, K., Koh, Y.K., Chiritescu, C., Zheng, X., Cahill, D.G.: Rev. Sci. Instrum. **79**, 114901 (2008)
27. Wang, Y., Park, J.-Y., Koh, Y.K., Cahill, D.G.: J. Appl. Phys. **108**, 043507 (2010). Accepted
28. Cahill, D.G.: Rev. Sci. Instrum. **75**, 5119 (2004)

29. Costescu, R.M., Wall, M.A., Cahill, D.G.: Phys. Rev. B **67**, 054302 (2003)
30. Koh, Y.K., Bae, M.-H., Cahill, D.G., Pop, E.: Nano Lett. **10**, 4363 (2010)

Thermal Evaporation

▶ Physical Vapor Deposition

Thermal Infrared Detector

▶ Insect Infrared Sensors

Thermal Management Materials

▶ Modeling Thermal Properties of Carbon Nanostructure Composites

Thermal Resistance

▶ Thermal Conductivity and Phonon Transport

Thermal Resistivity

▶ Thermal Conductivity and Phonon Transport

Thermal Transport

▶ Thermal Conductivity and Phonon Transport

Thermally Actuated MEMS Resonators

▶ Thermally Actuated Silicon Resonators

Thermally Actuated Micromechanical Resonators

▶ Thermally Actuated Silicon Resonators

Thermally Actuated Nanoelectromechanical Resonators

▶ Thermally Actuated Silicon Resonators

Thermally Actuated Nanomechanical Resonators

▶ Thermally Actuated Silicon Resonators

Thermally Actuated Resonators

▶ Thermally Actuated Silicon Resonators

Thermally Actuated Silicon Resonators

Siavash Pourkamali
Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, USA

Synonyms

Thermally actuated MEMS resonators; Thermally actuated micromechanical resonators; Thermally actuated nanoelectromechanical resonators; Thermally actuated nanomechanical resonators; Thermally actuated resonators; Thermal-piezoresistive resonators

Definition

Thermally actuated electromechanical resonators are micro- to nanoscale structures whose specific mechanical resonant modes can be actuated by periodic thermal expansion forces resulting from joule heating in parts of their structure.

Introduction

Thermal actuation based on joule heating-induced thermal expansion of solids is a well-known mechanism that can be conveniently implemented at micro- and nanoscale. An electrothermal actuator simply consists of a conductive element that heats up as a result of electrical current flowing through it and therefore can be implemented without any fabrication challenges or the need for material integration. Electrothermal actuators also have great properties such as large actuation force, low operating voltage, and simplicity of design and integration. On the downside, their considerable power consumption makes them an undesirable choice for many applications. Furthermore, thermal actuators are generally considered slow actuators only suitable for DC or very-low-frequency applications. This is mainly based on the general perception emanating from the more familiar macro-world where thermal phenomena have relatively large time constants in the few seconds or even minutes and hours range. However, as will be shown in this section, thermal actuation along with piezoresistive sensing provides an interesting option for transduction of high-frequency resonant modes of micro- and nanostructures.

Micro- and nanoscale electromechanical resonant devices have received a lot of attention over the past decade as a potential emerging technology for next-generation integrated frequency references, electronic filters, and resonant sensors. As a result, a wide variety of high-frequency microelectromechanical resonator technologies have been developed over the past few years, most of which use piezoelectric [1, 2] or electrostatic (capacitive) [3–5] transduction. Despite significant improvements in both piezoelectric and capacitive microresonator technologies, each approach still has certain challenges to overcome. Piezoelectric microresonators require integration of piezoelectric and metallic thin films generally resulting in lower quality factors and frequency and quality control issues. In the case of air-gap capacitive resonators, the weak electromechanical coupling necessitates deep submicron transduction gaps leading to fabrication and reliability challenges, power handling limitations, and excessive squeezed film damping when operating in air. Solid dielectric resonators [5] use electrostatic forces in dielectric thin films (preferably a high-K dielectric) to take advantage of the larger dielectric coefficients and improve the electromechanical coupling strength. The fully solid structure of such device relaxes several problems associated with aggressive reduction of the air gaps such as stiction and pull-in issues and allows reduction of the dielectric gap sizes to a few nanometers. However, similar to the piezoelectric resonators, such devices require integration and physical contact of dielectric thin films and conductive electrodes with the resonant structures which defeats some of the main purposes of moving away from piezoelectric devices. Moreover, large dielectric constants of the very thin dielectric films result in large parasitic capacitors to be associated with such devices.

Frequency tuning and compensation of the temperature-induced frequency drift are among other formidable challenges for both piezoelectric and capacitive resonators. Finally, in order to achieve a strong enough electromechanical transduction and consequently a large signal to noise ratio, both capacitive and piezoelectric resonators require relatively large electrode areas (thousands of μm^2) consuming costly real estate on the chip and limiting the maximum number of devices that can be integrated on a single chip.

Electrothermal actuation presents a third option for actuation of resonant modes of micro- and nanoscale structures that has been neglected to some extent so far, especially for high-frequency applications. Different versions of lower frequency thermally actuated micromechanical resonant devices have been utilized as mass sensors for chemical [6] or

particulate [7] sensing. Utilization of thermal actuation for high-frequency resonant modes has been gaining more attention lately. A 10-MHz MEMS oscillator was demonstrated in [8] via electrothermal actuation of resonant modes in a suspended polysilicon membrane. In [9], thermal actuation and piezoresistive detection of flexural resonant modes of a silicon-carbide nanobeam with frequencies up to 1.1 GHz has been demonstrated. Recent studies and experiments have shown promising results and plenty of unexplored potentials for thermally actuated high-frequency resonators especially as the structural dimensions are scaled down into the nanoscale [10]. Both modeling and experimental measurements suggest that as opposed to the general presumption, thermal actuation becomes a more efficient actuation approach for higher-frequency resonant modes provided that higher frequency is achieved by shrinking the structural dimensions down. Unlike electrostatic and piezoelectric resonators, thermal-piezoresistive resonators perform better as their size is shrunk down and therefore could be a much stronger candidate for realization of highly integrated nanomechanical arrays of resonant signal processors. Self-sustained oscillation capability [11, 12] and robustness and design flexibility for sensory applications [7] are among other unique properties that thermal-piezoresistive transduction offers.

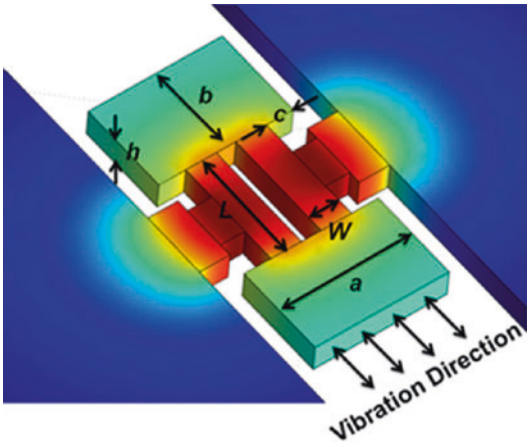
High-Frequency Thermal-Piezoresistive Resonators: Principles of Operation and Modeling

High-Frequency Thermal Actuation: As mentioned earlier, based on the perception from the macro-world, it is hard to imagine a thermal system with small enough thermal time constants to respond to high-frequency excitations. Similar to electrical RC circuits, thermal time constant of a thermal actuator, for example, a suspended conducting beam, is the product of its thermal capacitance and thermal resistance: $\tau_{th} = R_{th} C_{th}$. Thermal resistance can be calculated using a very similar equation to electrical resistance, that is, $R_{th} = \sigma_{th}^{-1} L/A$, where σ_{th} is the thermal conductivity of the structural material and L and A are the length and cross-sectional area of the element. Similar to electrical resistance, thermal resistance increases proportionally as all the dimensions of the element are scaled down. This implies that if a thermal actuator is scaled by a factor of S, its thermal resistance changes by a factor of S^{-1}. On the other hand, thermal capacitance is proportional to the mass and consequently the volume of the element and scales by a factor of S^3. Hence, overall, the thermal time constant of the element scales by a factor of S^2.

On the other hand, if a mechanical structure is scaled by a factor S, its mechanical resonant frequency (ω_m) changes by a factor of S^{-1}, that is, its mechanical time constant changes by a factor of S. The overall conclusion is that if the dimensions of a thermally actuated resonant structure are scaled down, its thermal time constant shrinks faster than its mechanical time constant. To be more specific, both thermal and mechanical responses of the system become faster upon scaling the dimensions down; however, the increase in the speed of the thermal response is sharper (proportional to S^{-2}) than the increase in the mechanical resonant frequency (proportional to S^{-1}). Therefore, as the structural dimensions are scaled down, the performance of thermally actuated mechanical resonators is expected to improve (more actuation force for the same actuation power) as the thermal response of the system can catch up with the mechanical vibrations more effectively.

Thermal-Piezoresistive Transduction: Piezoresistive sensing is the most convenient sensing mechanism to be integrated and utilized along with thermal actuation. Similar to thermal actuation, piezoresistive sensing requires only a conductive element preferably with a high piezoresistive coefficient. To add to the simplicity, the same elements can be used for both thermal actuation and piezoresistive sensing simultaneously.

Thermal-piezoresistive resonators can come in different shapes and sizes. Figure 1 shows the 3D schematic view of a resonant structure, known as I-shaped bulk acoustic wave resonator or IBAR (also known as dog-bone resonator), which is very suitable for thermal-piezoresistive transduction



Thermally Actuated Silicon Resonators, Fig. 1 3D schematic view of a thermally actuated IBAR (dog-bone resonator) showing the qualitative distribution of temperature fluctuation amplitude when subject to a fluctuating current flow (*red* shows the maximum and *blue* shows the minimum)

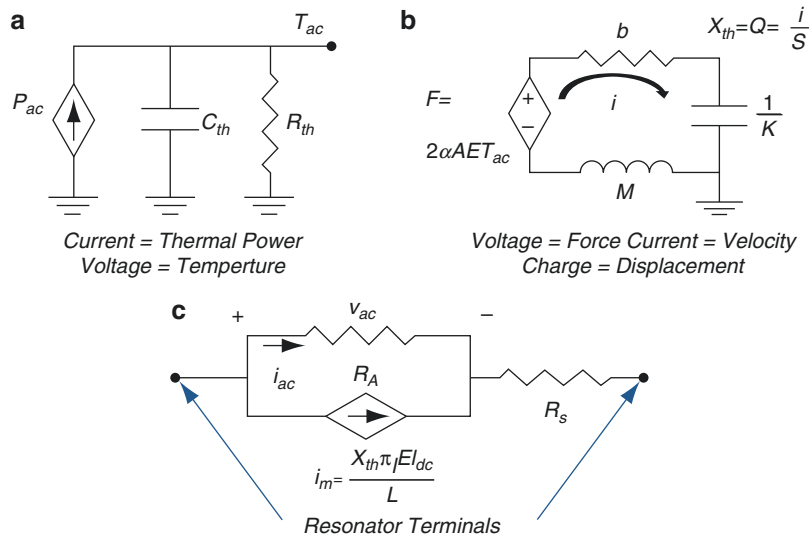
while being capable of providing relatively high resonance frequencies. In the in-plane extensional resonance mode, the two blocks hanging on the two sides of the structure move back and forth in opposite directions subjecting the two beams connecting them together to periodic tensile and compressive stress. Such resonance mode can simply be actuated by applying the actuation voltage (or current) between the two support pads on the two sides of the structure. The ohmic loss caused by the current flow is maximized in the narrower parts of its path, that is, the extensional pillars in the middle of the structure (the two beams connecting the two blocks to each other). Upon application of a fluctuating actuation voltage, the fluctuating ohmic power generation results in a fluctuating temperature gradient and therefore periodic thermal expansion of the extensional actuator beams. The alternating extensional force resulting from the fluctuating temperature in the beams can actuate the resonator in its in-plane extensional resonance mode. If the ohmic power and consequently the resulting temperature fluctuations have the same frequency as the mechanical resonance frequency of the structure, the mechanical vibration amplitude is amplified by the mechanical quality factor (Q) of the resonator. The amplified alternating stress in the pillars leads to considerable fluctuations in their electrical resistance due to the piezoresistive effect. When biased with a DC voltage, such resistance fluctuations modulate the current passing through the structure resulting in an AC component known as the motional current.

Thermo-electromechanical Modeling: Operation of a thermal-piezoresistive resonator involves phenomena in three different domains of thermal, mechanical, and electrical nature. Ohmic loss turns the electrical input voltage into a temperature, thermal expansion turns the temperature into a mechanical force that causes a mechanical displacement, and finally the piezoresistive effect turns the mechanical displacement back into an electrical signal (motional current).

Figure 2 shows the equivalent electrical circuit models for the three domains. In the thermal domain, the actuator beams act like a low-pass RC combination. The equivalent circuit consists of a current source representing the alternating heating power along with a capacitance (C_{th}) and a resistance (R_{th}) representing the effective thermal capacitance and thermal resistance of the thermal actuators, respectively (Fig. 2a). The resulting voltage across the parallel RC combination (T_{ac}) represents the temperature fluctuation amplitude.

Typically, the fluctuating actuation voltage is provided by a combination of a DC and an AC voltage component. Due to the square relationship between ohmic power generation and electrical voltage (or current), application of only an AC actuation voltage with frequency of f_a results in an ohmic power component with frequency of $2f_a$. In order to have the same frequency as the input AC actuation voltage for the thermal actuation force, a combination of AC and DC voltages (V_{dc} and v_{ac}) needs to be applied between the two terminals of the resonators. The resulting power loss component at the same frequency as v_{ac} in this case will be $P_{ac} = \frac{2V_{dc}v_{ac}}{R_A}$, where R_A is the electrical resistance of the actuator elements (extensional beams in IBARs) and V_{dc} and v_{ac} are the applied DC and AC actuation voltages, respectively.

The RLC combination of Fig. 2b represents the mechanical resonance behavior of the structure.



Thermally Actuated Silicon Resonators, Fig. 2 Equivalent electrical circuit for a thermal-piezoresistive resonator: (**a**) the thermal domain; (**b**) the mechanical domain; (**c**) the electrical domain

The voltage source represents the thermally generated mechanical force which is the product of thermal expansion coefficient (α) of the structural material, thermal actuator cross-sectional area (A), and Young's modulus (E) of the thermal actuator along its length. The coefficient "2" has been added since there are two actuator beams in each resonator contributing to the actuation force. The inductor, capacitor, and resistor represent the effective mechanical mass (M), stiffness (K), and damping (b) of the resonator, respectively. Consequently, current represents velocity, and electrical charge ($Q = \int i.dt$) represents displacement (X_{th}), which is the elongation of the thermal actuators (resonator vibration amplitude). The vibration amplitude, which is the output of the mechanical subsystem, acts as an input to the electrical subsystem (Fig. 2c) causing a change in the resistance of the thermal actuators due to the piezoresistive effect. This change in resistance modulates the current flow through the resonator which is shown as a current source in the equivalent circuit. R_s in Fig. 2c represents the parasitic electrical resistances in series with the actuator resistance which mainly includes resistance of resonator support beams.

The overall transfer function that relates the AC motional current (output current) to the AC input voltage is the product of the three transfer functions from the thermal, mechanical, and electrical domains. The simplified transfer function at resonance frequency is

$$H_T|s = j\omega_0 \quad = g_m = 4\alpha E^2 \pi_l Q \frac{AI_{dc}^2}{KLC_{th}\omega_0} \qquad (1)$$

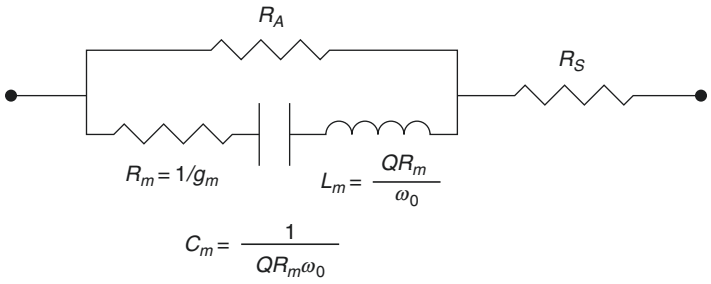
where π_l is the longitudinal piezoresistive coefficient of the structural material, L is the thermal actuator length, and Q and ω_0 are the quality factor and angular resonant frequency of the resonator.

For IBAR structures, knowing that $K = 2EA/L$, Eq. 1 can be further simplified to

$$g_m = 2\alpha E \pi_l Q \frac{I_{dc}^2}{C_{th}\omega_0}. \qquad (2)$$

A more detailed version of the modeling and derivations above is available in [10]. The transfer function in Eqs. 1 and 2 can be referred to as small signal voltage to current gain, or motional conductance (g_m) of the resonator at its resonance frequency, which is one of the most important parameters of a thermal-piezoresistive resonator when utilized as an electronic circuit component.

Resonator Equivalent Electrical Circuit: For the resonators discussed here, where the extensional beams acting as thermal actuators are also utilized for piezoresistive readout, the physical resistance of the resonator connects the input and



Thermally Actuated Silicon Resonators, Fig. 3 Compact overall equivalent electrical circuit for one-port thermally actuated resonators with piezoresistive readout

output of the devices. The equivalent electrical circuit in this case includes a resistance $R_A + R_s$ connected between the two terminals of the resonator, which is the overall resistance in the current path between the two pads of the resonator. In parallel to the static resistance of the actuators, there is a series RLC combination that represents the mechanical resonant behavior of the structure. The value of R_m in the RLC has to be set so that at resonance, a motional current of $i_m = g_m \cdot v_{ac} = v_{ac}/R_m$ is added to the feedthrough current passing through R_A. Therefore, $R_m = {g_m}^{-1}$ and L_m and C_m values can be calculated based on the value of R_m as shown in Fig. 3 according to the resonance frequency and quality factor of the resonator.

To calculate the motional conductance (g_m) of a thermal-piezoresistive IBAR, which is the most important parameter to be extracted from the model using Eq. 2, all the parameters except the effective thermal capacitance of the actuators (C_{th}) are known for every set of resonator dimensions and structural material. Due to the distributed nature of the thermal parameters including thermal generator (electrical resistance of the structure), thermal capacitance, and thermal conductance, analytical derivations or finite element analysis should be used to find an accurate value for C_{th}. In [10, 13], finite element analysis has been used to simulate the transient thermal response of the actuators and calculate the fluctuating temperature amplitude (T_{ac}) at different points along the length of the actuator beams. The mean value of the small signal temperature amplitudes at different points along the actuators is the effective temperature fluctuation amplitude for the actuator from which the effective thermal capacitance value can be extracted. As shown in [13], the effective thermal capacitance for a thermal actuator beam is equal to 0.92 of the static lumped element thermal capacitance of the beam with a good approximation.

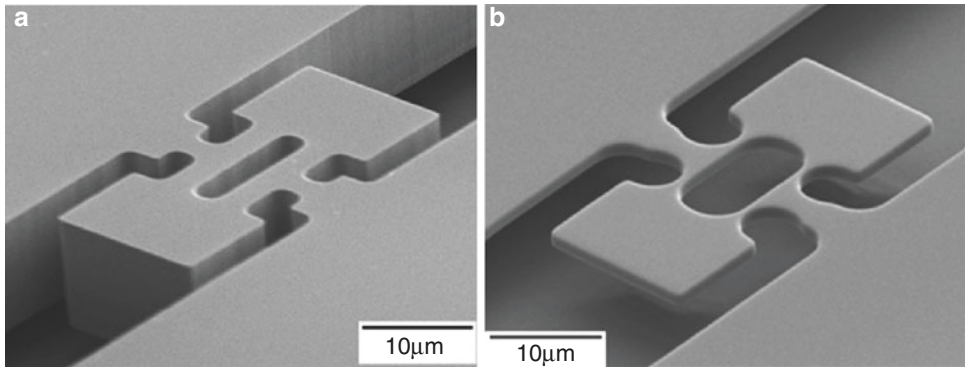
Resonator Scaling Behavior and Optimization

As an electronic component, it is desirable to maximize the resonator motional conductance (g_m) (for improved signal to noise ratio) while minimizing power consumption. Therefore, a figure of merit for the resonators can be defined as the ratio of the resonator g_m to the overall DC power consumption:

$$F.M. = \frac{g_m}{P_{DC}} = \frac{2\alpha E \pi_l Q}{C_{th}\omega_0(R_A + R_s)} \qquad (3)$$

where $P_{DC} = {I_{DC}}^2 \cdot (R_A + R_S)$ is the static power consumption of the resonator. Different parameters that can be used to maximize *F.M.* are as follows:

Actuator Beam Dimensions: Generally, smaller actuator dimensions lead to smaller C_{th} improving resonator figure of merit. However, the effect of actuator dimensions on its electrical resistance (R_A) should also be taken into account. Since the extensional stiffness of the actuator beams defines the resonance frequency of the IBAR, in order to maintain the same resonance frequency for an IBAR while reducing its actuator thermal capacitance, both length and width of the actuator should be scaled down simultaneously. Such scaling does not affect the actuator electrical resistance (R_A). Therefore, scaling down both the length and width



Thermally Actuated Silicon Resonators, Fig. 4 SEM view of single crystalline silicon thermal-piezoresistive IBARs (**a**) a 61-MHz 15-μm-thick IBAR fabricated on a low-resistivity p-type SOI substrate; (**b**) a 41-MHz 3-μm-thick IBAR fabricated on a low-resistivity n-type SOI substrate followed by thermal oxidation to thin down its thermal actuator beams

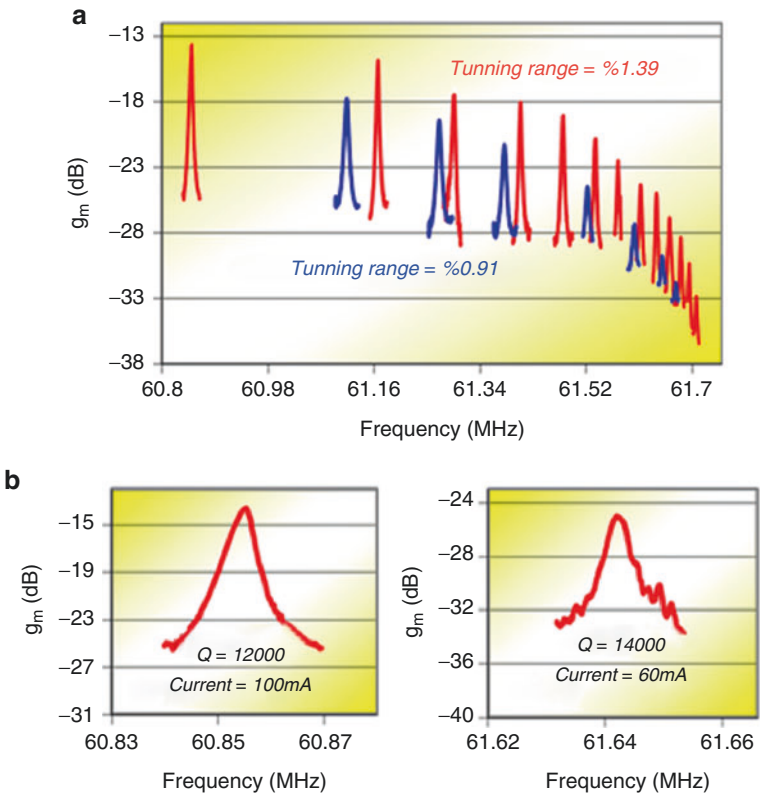
of the actuator beams by a scale S_a results in an improvement in the resonator *F.M.* by a factor of S_a^2 while maintaining an almost constant resonance frequency for the device.

Resonance Frequency and Resonator Scaling: At a first glance at Eq. 3, higher resonant frequencies seem to have a deteriorating effect on resonator figure of merit. However, if higher resonant frequencies are achieved by shrinking the resonator size, at the same time, C_{th} will be shrinking sharply. If all resonator dimensions are scaled down proportionally by a factor S, $C_{th} \propto S^{-3}$, $\omega_0 \propto S$, and $(R_A + R_S) \propto S$. Therefore, *F.M.* $\propto S$, that is, *F.M.* increases proportionally if the resonator dimensions are scaled down, while its resonance frequency increases at the same time. This confirms the validity of the explanation provided previously claiming suitability of thermal actuation for high-frequency applications. It can be concluded from the last two discussions that the capability to fabricate resonator actuator beams with very small (nanoscale) dimensions is the key to achieving very high performance levels for thermal-piezoresistive resonators, both at low and high frequencies.

Electrical Resistivity and Other Material Properties: One of the most important parameters that can also be controlled for some of the most popular resonator structural materials (e.g., Si or SiC) is the electrical resistivity of the structural material. Lower electrical resistivity of the structural material improves the transduction figure of merit by lowering $R_A + R_s$. This can be explained by the fact that, with lower electrical resistivity, the same amount of DC bias current and therefore the same motional conductance can be maintained while burning less ohmic power in the structure.

A number of other structural material properties also affect the figure of merit for thermal-piezoresistive resonators. Higher thermal expansion coefficient and lower specific heat capacity improve the thermal actuation resulting in proportionally higher *F.M.* Higher piezoresistive coefficient improves the piezoresistive readout having the same effect on *F.M.* One interesting and promising fact is the giant piezoresistive effect found in single crystalline silicon nanowires that has been observed and reported by different researchers [14]. Therefore, as the resonators are scaled down to reach higher resonant frequencies in the hundreds of MHz and GHz range, it is expected that the actuator beams with deep submicron width exhibit such strong piezoresistivity significantly improving resonator performances. Finally, using structural materials with higher Young's modulus (E) can also improve the resonator figure of merit while increasing its resonant frequency.

Figure 4 shows the scanning electron micrograph (SEM) of two different microscale single crystalline silicon IBAR structures. Figure 5 shows the motional conductance frequency plots

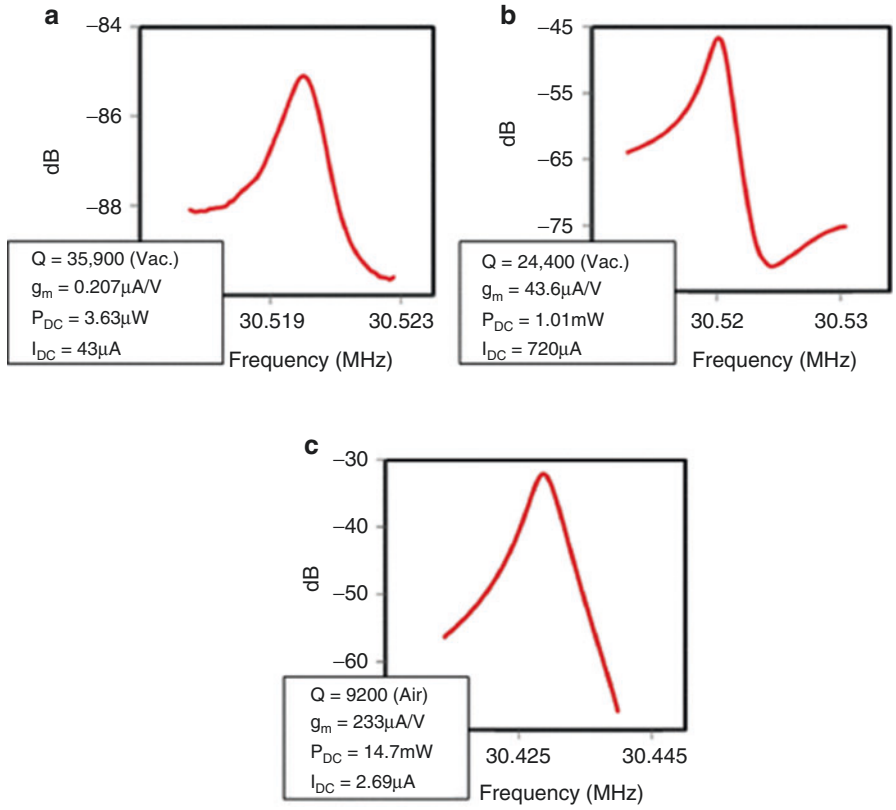


Thermally Actuated Silicon Resonators, Fig. 5 Measured frequency response of the thermally actuated 61-MHz resonator of Fig. 4a with different bias currents. The vertical axis in the frequency response plots shows the motional conductance of the resonators (g_m) in dB with respect to 50-Ω network analyzer terminations. *Red* and *blue* plots refer to vacuum and air testing conditions, respectively. Current range is 45–100 mA in vacuum and 55–100 mA in air

extracted from the measured frequency response of the 61-MHz resonator in Fig. 4a. The resonator quality factor ranges from 12,000 to 14,000 under vacuum and drops to 6,000–8,000 under atmospheric pressure. A large-frequency tuning range is demonstrated for this resonator by changing its DC bias current. This is mainly due to the raising temperature of the resonating body, especially its extensional beams that provide most of the structural stiffness in the in-plane extensional mode. As expected, as the DC bias current and consequently resonator motional conductance at resonance (g_m) increase, due to higher static temperature and softening of the structural material, the resonance frequency decreases. Despite having high-quality factors and relatively large motional conductance, the large current requirement and power consumption for the resonator in Fig. 5 could be a major drawback against using such electronic components.

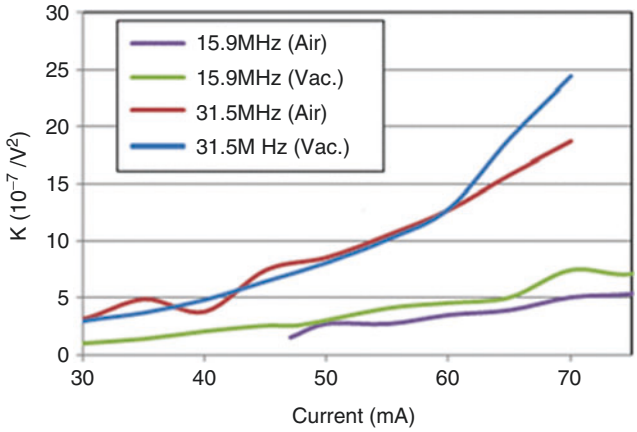
Figure 6 shows a few frequency response plots measured for a 30.5-MHz resonator with narrowed down actuator beams (~750 nm wide) similar to the structure in Fig. 4b. Submicron actuator beams can be achieved by controlled oxidation of the silicon structures followed by oxide removal leaving behind a narrower beam. Because of its much thinner actuators, a resonance peak can be detected for this resonator with currents as low as 43 μA translating into DC power consumptions as low as a 3.6 μW. At the DC bias current of 720 μA (DC power of 1 mW), motional conductance of 44 μA/V was measured for this resonator (Fig. 6b). Maximum currents tolerable for such narrow actuator beams are in the few mA range leading to motional conductances as high as 0.23 mA/V in air with Q of 9,200 (Fig. 6c). According to the derived model, much higher motional conductances can be achieved by consuming the same amount of power using lower-resistivity structural material allowing higher bias currents.

To verify better suitability of thermal actuation for higher-frequency resonators with smaller



Thermally Actuated Silicon Resonators, Fig. 6 Measured motional conductance frequency plots for a 3-μm-thick 30.5-MHz IBAR. The vertical axis in the plots shows the motional conductance of the resonators (g_m) in dB with respect to 50-Ω network analyzer terminations. Due to its very narrow (750 nm wide) actuator beams, clear resonance peaks with motional conductance in the tens of μA/V have been measured for this resonator with sub-mA DC bias currents (sub-mW power consumption)

dimensions, Fig. 7 shows the measured transduction figure of merit for two thermal-piezoresistive resonators with the effect of quality factor taken out $\left(K = \frac{F.M.}{Q}\right)$ under different bias currents. The two resonators are fabricated on the same substrate and have similar horizontal dimension aspect ratios with the 32-MHz resonator being a 2X-scaled-down version of the 16-MHz resonator. As expected based on the previously discussed scaling trend and the derived model, the higher-frequency resonator has a higher transduction figure of merit. This further confirms that thermal actuation is not only a potentially suitable actuation mechanism for high-frequency applications, but also it can provide higher performances at higher rather than lower frequencies. Therefore, much higher transduction efficiencies are expected to be achievable for higher-frequency versions of such devices with resonant frequencies in the hundreds of megahertz and even GHz range, provided that the dimensions can be scaled down proportionally into the nanoscale. The derived resonator model backed up by measurement results shows that with sub-100-nm actuator width and submicron actuator length, g_m as high as 1 mA/V can be achieved for thermal-piezoresistive silicon IBARs with resonant frequencies in the gigahertz range with power consumptions as low as a few microwatts [10]. This is not including the giant piezoresistive effect in crystalline silicon that could further improve this by over an order of magnitude [14].

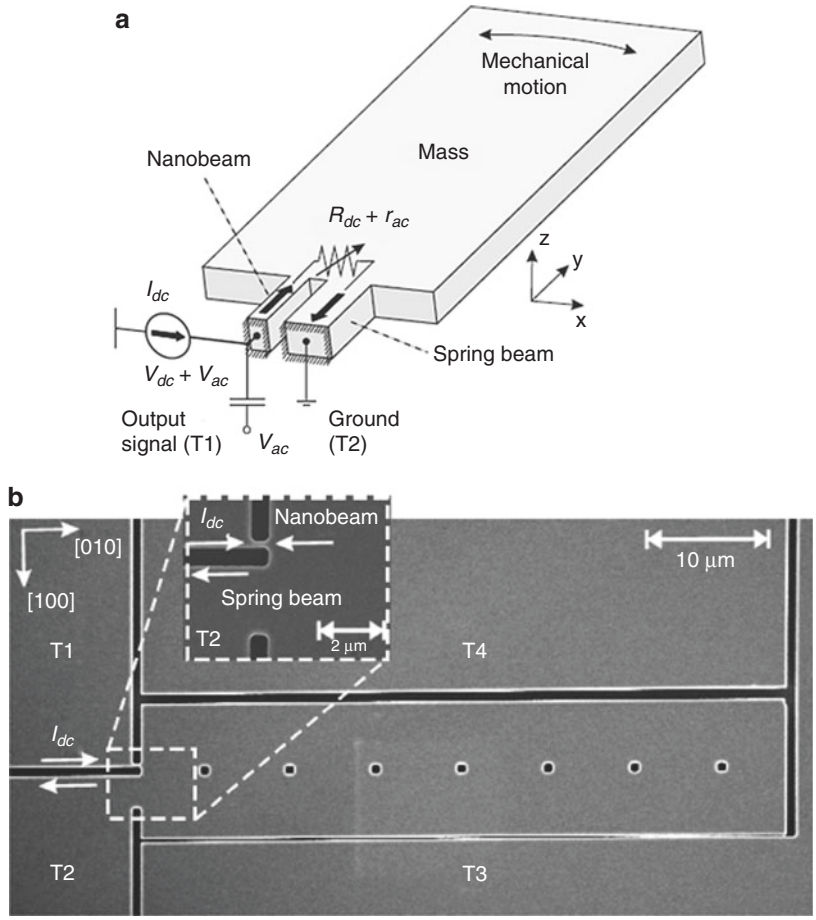


Thermally Actuated Silicon Resonators, Fig. 7 Extracted transduction figure of merit (K) for two IBARs versus resonator bias current. The higher-frequency (31.5 MHz) resonator, which is a 2X-scaled-down version of the 15.9-MHz resonator, has a higher transduction strength. The increase in the transduction figure of merit with bias current, which is not directly predicted by the derivations, is expected to be a result of the effect of increased static temperature at higher currents on different material properties, for example, increased thermal expansion coefficient at higher temperatures

Self-Sustained Thermal-Piezoresistive Oscillators

Generally, in order to implement an electronic oscillator with a mechanical resonator as its frequency reference, the resonator needs to be engaged in a positive feedback loop consisting of amplifying circuitry with appropriate phase shift. One of the interesting aspects of the thermal-piezoresistive mechanical resonators that use the same elements as both thermal actuators and piezoresistive sensors is the internal feedback mechanism formed between thermal actuation force and piezoresistivity. This feedback results from the fact that the piezoresistive motional current of the structure that passes through the thermal actuators also acts as an excitation for the resonator by affecting the ohmic loss in the actuator beams. Depending on the sign of the structural material piezoresistive coefficient, electrical terminations of the resonator, and the phase shift resulting from thermal delays, this could act as a positive feedback stimulating mechanical vibrations or a negative feedback limiting vibration amplitude of the structure or reducing Brownian motion of the structure at that specific frequency (refrigeration [11]). For example, ohmic power generation in a thermal actuator biased with a constant current (terminated with a large electrical resistance) decreases if the actuator is subjected to tensile stress, due to the decreased electrical resistance. As part of a resonant structure, when such actuator is fully stretched and is about to start to compress, its decreased ohmic power and resulting cooling assist the actuator going into compression. In the same manner, after compression of the actuator, increased ohmic power generation helps its expansion via raising its temperature. Mechanical and thermal time delays play a key role here in determining whether the feedback has an amplifying or attenuating effect.

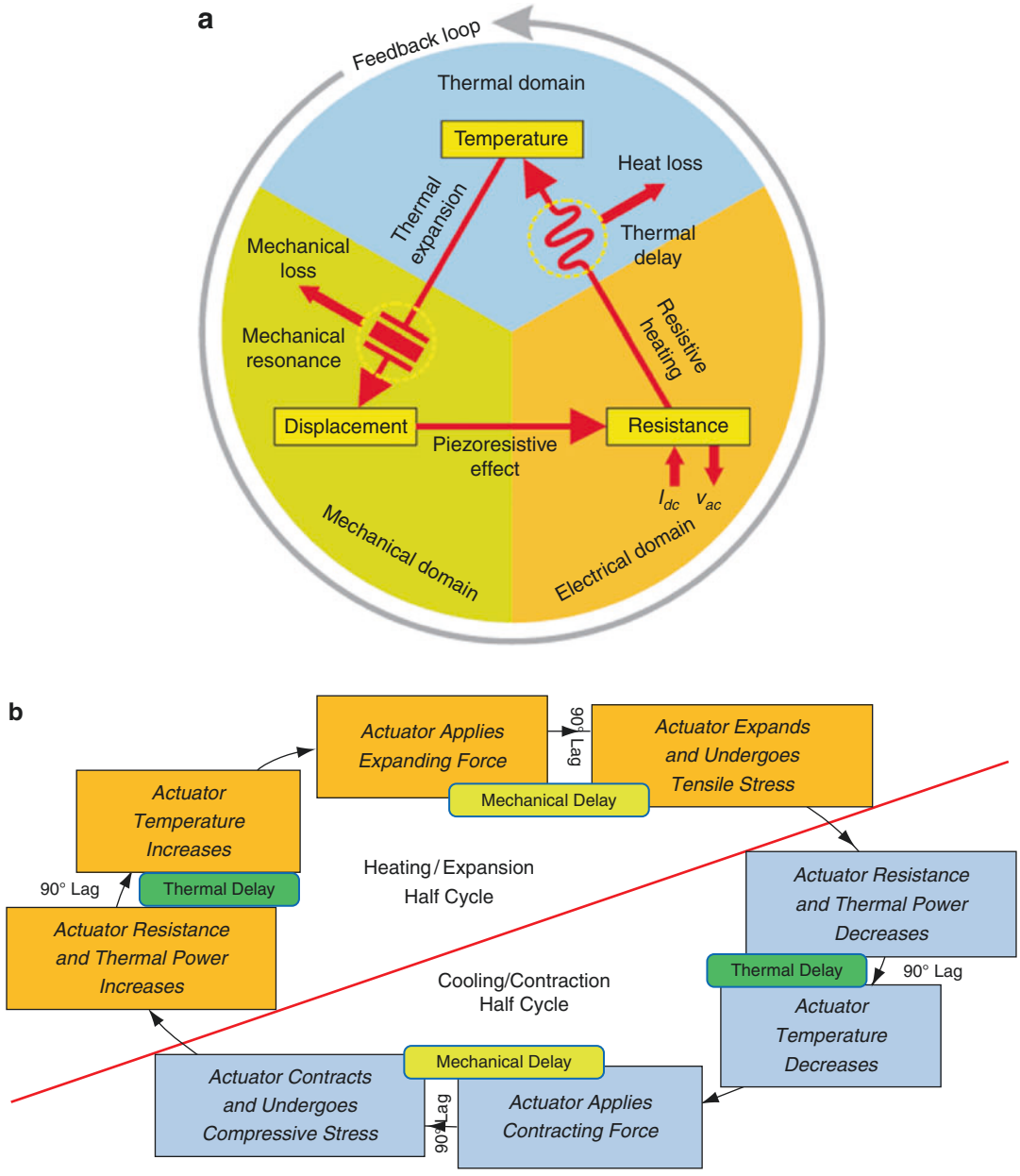
It was demonstrated by researchers at NXP Semiconductors [11] that such positive feedback can be strong enough to initiate and maintain mechanical oscillations upon providing a DC electrical power source to the structure without the need for any external AC stimulation. Figure 8 shows the scanning electron micrograph of the suspended N-type single crystalline silicon structure for which such effect was demonstrated for the first time [11]. The structure is made of N-type silicon because N-type silicon has a negative longitudinal piezoresistive coefficient, which is required for such oscillation under constant current conditions.



Thermally Actuated Silicon Resonators, Fig. 8 Schematic view and SEM top view of a fully micromechanical crystalline silicon oscillator. To operate the oscillator, a DC is applied between the terminals of the device, and the output AC voltage resulting from the resonator's spontaneous mechanical vibrations is measured. The inset shows a zoomed-in view of the nanobeam actuator and the wide beam providing the flexural stiffness of the structure. Output electrical signal amplitude of 27 mV with frequency of 1.26 MHz was measured for this resonator when biased with a current of 1.2 mA (1.2-mW power consumption) under vacuum

Figure 9 schematically shows the feedback loop and the transduction mechanisms connecting the three physical parameters: displacement (x), temperature (T_{ac}), and voltage (v_{ac}) leading to self-sustained oscillation [11]. As shown in Fig. 9b, physically, this can be explained as follows: upon application of a DC bias, the thermal actuator heats up and expands, pushing the mass to the right side. Due to the mass (inertia) of the plate, the nanobeam actuator experiences an overexpansion after pushing the mass and undergoes tensile stress. The negative piezoresistive coefficient turns the tensile stress into reduced electrical resistance for the nanobeam. When biased with a constant DC bias current, this means reduced ohmic power in the actuator forcing it to contract. After structural overcontraction (due to the mass of the plate), the actuator will be under compressive stress and will have increased electrical resistance. This translates into a higher thermal power forcing the structure to expand again. If the resulting driving force (due to heating and cooling) in each cycle is large enough to compensate for mechanical losses of the structure, the same sequence is repeated over and over in a self-sustained manner, and the vibration amplitude keeps increasing until it is limited by nonlinearities. The 90° phase lags (Fig. 9b) resulting from the mechanical and thermal response delays lead to a perfect timing for the abovementioned sequence of physical phenomena to continue.

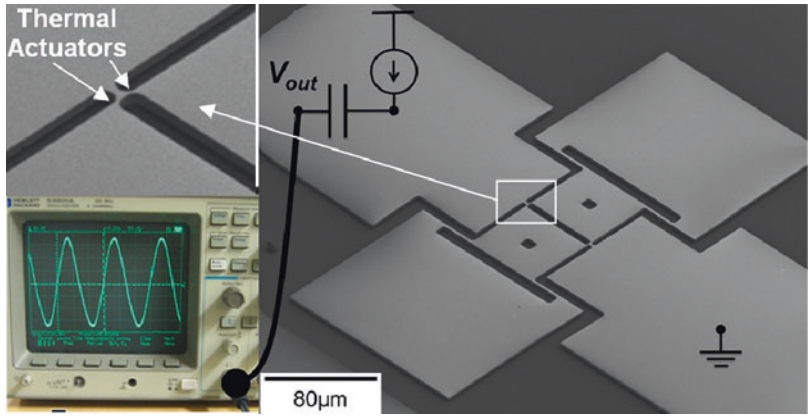
Figure 10 shows the SEM view and time response of an extensional mode 6.7-MHz self-sustained thermal-piezoresistive oscillator operating based on the same principle. Such structure is



Thermally Actuated Silicon Resonators, Fig. 9 (**a**) Schematic demonstration of the feedback loop and the transduction mechanisms linking the thermal, mechanical, and electrical parameters to each other. Thermal delay and mechanical resonance provide the required phase shift for self-sustained oscillation. (**b**) Sequence of phenomena causing an internal positive feedback loop in thermal-piezoresistive resonators biased with a constant current that lead to self-sustained oscillation

a derivative of the IBAR structures discussed in the previous sections with relatively larger plates. It operates in an in-plane resonant mode with its two plates moving back and forth in opposite directions and the extensional actuator beams being stretched and compressed periodically. Extensional resonant modes can provide higher frequencies as well as higher-quality factors both

T



Thermally Actuated Silicon Resonators, Fig. 10 SEM view and output signal of a 6.6-MHz single crystalline silicon thermal-piezoresistive dual-plate resonator capable of self-sustained oscillation. Only a DC bias current of 3.0 mA is being applied to the resonator, and the resonator is operating in air. The power consumption of the oscillator is about 20 mW, and its output voltage amplitude is 70 mV

under vacuum and in air. Dual plate oscillators with frequencies in the few megahertz range, power consumption of a few to tens of milliwatts, and operating under both vacuum and atmospheric pressure have been demonstrated [12]. The output voltage amplitudes range from tens to hundreds of millivolts.

It can be shown that the requirement for self-sustained oscillation in such devices is that the absolute value of the motional conductance of the resonator becomes larger than the physical static conductance of its actuators. The minimum DC power required to achieve this condition is

$$P_{dc_{\mathrm{Min}}} = \frac{KLC_{th}\omega_m}{4\alpha E^2|\pi_l|QA}. \tag{4}$$

Again, the key to having low power consumption is scaling down the resonator dimensions that reduce the power requirement with a square relationship leading to significantly lower power consumption for scaled-down versions of such devices while having higher oscillation frequencies.

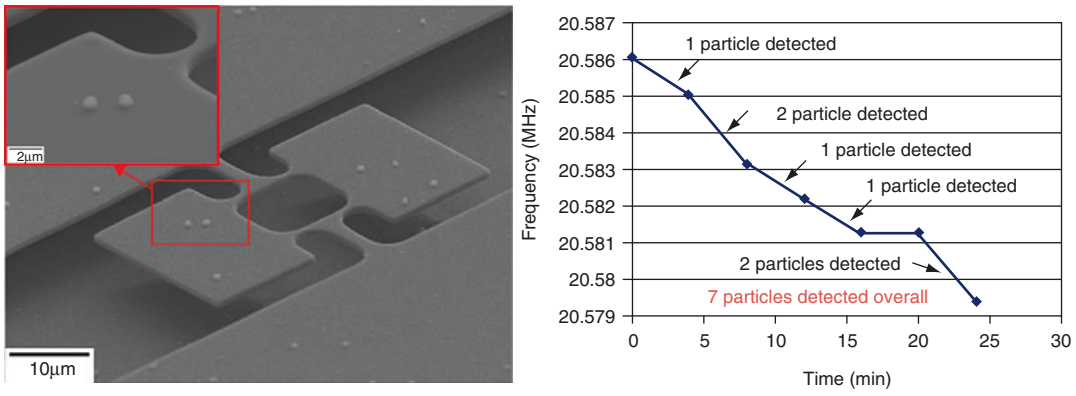
Another area where such devices need significant improvement is the output voltage amplitude, which is a small fraction of the DC voltage across the resonator (typically 0.01–0.1). What limits the oscillation amplitude of the resonators is nonlinearities (most likely mechanical nonlinearities) limiting the feedback loop gain to unity after reaching a certain amplitude. Therefore, a thorough study on minimization of the mechanical nonlinearities can possibly lead to fully micromechanical oscillators with higher output amplitudes. Piezoresistive coefficient of the structural material is another parameter that has a direct effect on the output voltage amplitude of such oscillators. With the same mechanical vibration amplitude and bias current, higher piezoresistive coefficient leads to proportionally higher output voltage amplitude. This further shows the significance of fabricating such devices with sub-100-nm actuator width taking advantage of the huge piezoresistivity observed in silicon nanowires [14] to increase the output amplitude.

Elimination of the sustaining amplifier by the self-sustained fully micromechanical oscillators described here might not be a significant improvement as electronics can be cheap and small and readily necessary to support other functions within the systems employing the electromechanical oscillators. However, for applications where large oscillator arrays are required, for example, sensors arrays, elimination of the amplifiers and the associated interconnections could be highly beneficial.

Thermal-Piezoresistive Resonant Mass Sensors

Electromechanical resonators have great potential as highly sensitive mass sensors for gas, biomolecular, or particulate sensing applications. For such sensory applications where the resonator has to be in contact with the surrounding environment, resonator robustness becomes a critical factor. This makes thermal-piezoresistive resonators made of a single monolithic structural material a great candidate for mass sensing applications. Especially for particulate sensing for air quality monitoring or environmental research, thermal-piezoresistive resonators seem to be the most suitable choice [7, 15]. Air-gap capacitive resonators are extremely vulnerable to contaminants or particulates entering and getting stuck in their nanogaps making it practically impossible to use them as particulate sensors. For piezoelectric resonators, it is very hard (if not impossible) to have a uniform mass sensitivity all over the resonator sensing surface or at least a large portion of it. This is due to the fact that such devices generally operate in their bulk resonance modes resulting in different vibration amplitudes and therefore different effective resonator mass at different locations on the surface. Uniform mass sensitivity is a necessity when targeting mass measurement and counting of individual particles. Furthermore, metallic and piezoelectric thin films in a piezoelectric resonant structure are typically less robust than a monolithic crystalline structure, especially when subjected to impact by high-speed airborne particles.

Thermal-piezoresistive resonators can continue to work flawlessly with tens to thousands of particles deposited on them [7]. For the IBAR and dual-plate structures, the vibrating counter masses (plates) act as rigid bodies having almost the same vibration amplitude on their whole surface. This allows measuring the mass of individual deposited particles as they arrive providing the overall count and mass distribution of particles present in air samples. There is still the possibility of particles landing on the actuator beams and experiencing different mass sensitivities. However, the surface area of the actuator beams is typically much smaller than that of the plates making that probability small enough for the resulting error to be within the acceptable range for most applications. Furthermore, the miniaturization capability of thermal-piezoresistive resonators into the nanoscale allows much higher mass sensitivities to be achievable enabling detection and mass measurement of individual nanoparticles not feasible using the larger capacitive or piezoelectric resonant devices. Figure 11 shows the SEM view of an IBAR exposed to a flow of artificially generated airborne particles



Thermally Actuated Silicon Resonators, Fig. 11 SEM view of a 20.5-MHz IBAR with seven artificially generated 1-μm diameter airborne particles deposited on it. The changes in the measured resonance frequency versus time confirm the arrival of seven particles. There are five frequency reduction steps, of which two have a slope twice that of the rest which show that two particles have been deposited on the device during those periods

with the same size and composition along with its measured resonance frequency during the deposition [15]. The effect of arrival of each particle on the resonance frequency is clearly detectable and the same for all the particles. This confirms uniform mass sensitivity on the resonator sensing plates.

Cross-References

- ▶ Laterally Vibrating Piezoelectric Resonators
- ▶ NEMS Resonant Chemical Sensors
- ▶ Optomechanical Resonators
- ▶ Piezoresistivity
- ▶ Thermal Actuators

References

1. Piazza, G.: Integrated aluminum nitride piezoelectric microelectromechanical system for radio front ends. J. Vac. Sci. Technol. A **27**(4), 776–784 (2009)
2. Rinaldi, M., Zuniga, C., Chengjie, Z., Piazza, G.: Super-high-frequency two-port AlN contour-mode resonators for RF applications. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **57**(1), 38–45 (2010)
3. Nguyen, C.T.C.: MEMS technology for timing and frequency control. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **54**(2), 251–270 (2007)
4. Pourkamali, S., Ho, G.K., Ayazi, F.: Low-impedance VHF and UHF capacitive silicon bulk acoustic wave resonators. IEEE Trans. Electron Devices **54**(8), 2017–2023 (2007)
5. Weinstein, D., Bhave, S.A.: Internal dielectric transduction in bulk mode resonators. J. Microelectromech. Syst. **18**(6), 1401–1408 (2009)
6. Seo, J.H., Brand, O.: High Q-factor in-plane-mode resonant microsensor platform for gaseous/liquid environment. J. Microelectromech. Syst. **17**(2), 483–493 (2008)
7. Hajjam, A., Wilson, J.C., Rahafrooz, A., Pourkamali, S.: Fabrication and characterization of thermally actuated micromechanical resonators for airborne particle mass sensing, part II: device fabrication and characterization. J. Micromech. Microeng. **20**, 125019 (2010)
8. Reichenbach, R.B., Zalalutdinov, M., Parpia, J.M., Craighead, H.G.: RF MEMS oscillator with integrated resistive transduction. IEEE. Electron. Device. Lett. **27**, 805–807 (2006)
9. Bargatin, I., Kozinsky, I., Roukes, M.L.: Efficient electrothermal actuation of multiple modes of high-frequency nanoelectromechanical resonators. Appl. Phys. Lett. **90**, 093116 (2007)
10. Rahafrooz, A., Pourkamali, S.: High frequency thermally actuated electromechanical resonators with piezoresistive readout. IEEE. Trans. Electron. Devices. **58**(4), 1205–1214 (2011)
11. Steeneken, P.G., Phan, K.L., Goossens, M.J., Koops, G.E.J., Verheijden, G.J.A.M., van Beek, J.T.M.: Piezoresistive heat engine and refrigerator. Nat. Phys. **7**, 354 (2012). doi:10.1038/nphys1871
12. Rahafrooz, A., Pourkamali, S.: Fully micromechanical piezo-thermal oscillators. In: Proceedings of the IEEE international electron device meeting (IEDM), San Francisco. doi: 10.1109/IEDM.2010.5703314 (2010)
13. Rahafrooz, A., Pourkamali, S.: Fabrication and characterization of thermally actuated micromechanical resonators for airborne particle mass sensing, part I: resonator design and modeling. J. Micromech. Microeng. **20**, 125018 (2010)
14. He, R., Yang, P.: Giant piezoresistance effect in silicon nanowires. Nat. Nanotechnol. **1**(1), 42–46 (2006)
15. Hajjam, A., Wilson, J.C., Pourkamali, S.: Individual air-borne particle mass measurement using high frequency micromechanical resonators. IEEE Sensors J. **11**(11), 2883–2890 (2011)

Thermal-Piezoresistive Resonators

- ▶ Thermally Actuated Silicon Resonators

Thermodynamics of Small Systems

- ▶ Nanocalorimetry

Thermoelectric Nanomaterials and Nanostructured Thermoelectrics

- ▶ Nanostructured Thermoelectric Materials

Thermomechanical Actuators

- ▶ Thermal Actuators

Thermometry

► Nanocalorimetry

Thermostability

► Structure and Stability of Protein Materials

Thickness

► Interfacial Investigation of Protein Films Using Acoustic Waves

Thiols

► Nanostructures for Surface Functionalization and Surface Properties

Third-Generation Vectors

► Multistage Vectors

3D Micro-/Nanomanipulation with Force Spectroscopy

H. Xie[1] and Stéphane Régnier[2]
[1]The State Key Laboratory of Robotics and Systems, Harbin Institute of Technology, Harbin, China
[2]Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie, Paris, France

Synonyms

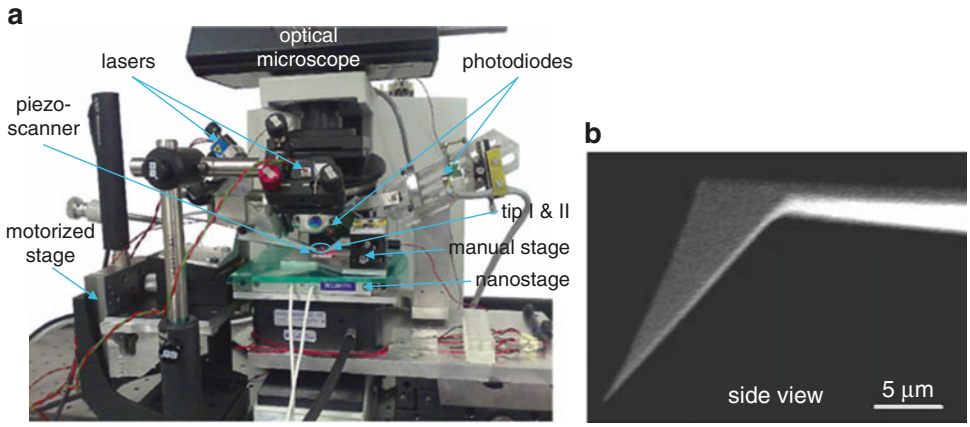
Force spectroscopy; Nano manipulation; Nanorobotics

Definition

A micromanipulation system is a robotic device which is used to physically interact with a sample under a microscope, where a level of precision of movement is necessary that cannot be achieved by the unaided human hand.

It is well known that pick-and-place is very important for 3-D microstructure fabrication since it is an indispensable step in the bottom-up building process. The main difficulty in sufficiently completing such pick-and-place manipulation at this scale lies in fabricating a very sharp end-effector that is capable of smoothly releasing microobjects deposited on the substrate. Moreover, this end-effector has to provide enough grasping force to overcome strong adhesion forces [1–3] from the substrate as well as being capable of sensing and controlling interactions with the microobjects. Furthermore, compared with the manipulation of larger microobjects under an optical microscope, visual feedback at several microns more suffers from the shorter depth of focus and the narrower field of view of lenses with high magnifications, although different schemes or algorithms have been introduced on techniques of autofocusing [4, 5] and extending focus depth [7]. Compared with vision-based automated 2-D micromanipulation, automated 3-D micromanipulation at the scale of several microns to submicron scale is more challenging because of optical microscope's resolution limit (typically 200 nm). Moreover, additional manipulation feedback is needed that is beyond the capability of optical vision, such as in the cases of vertical contact detection along the optical axis or manipulation obstructed by opaque components. Therefore, multi-feedback is of vital importance to achieve such accurate and stable 3-D micromanipulation at the scale of several microns to submicron scale.

Pick-and-place nanomanipulation is also a promising technique in 3D nanostructure fabrication since it is an indispensable step in the bottom-up building process. It can overcome limitations of bottom-up and top-down methods of nanomanufacturing and further combine advantages of these two methods to build complex 3D



3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 1 (**a**) A photo of the AFM-FRS on the configuration for nanoscale pick-and-place manipulation. (**b**) A SEM image of the cantilever fabricated with a protruding tip

nanostructures. In literature, nanostructures have been manipulated, assembled, and characterized by integrating nanomanipulators or nanogrippers into scanning electron microscopes (SEM) and transmission electron microscopes (TEM) [12–16]. Both the SEM and the TEM provide a vacuum environment where the van der Waals force is the main force to be overcome during the manipulation. 3D nanomanipulation could be also achieved with optical tweezers in liquid, where the adhesion forces are greatly reduced [17–19]. However, the pick-and-place nanomanipulation in air is still a great challenge due to the presence of strong adhesion forces, including van der Waals, electrostatic and capillary forces [20]. In this case, the main difficulties in achieving the 3D nanomanipulation are fabricating sharp end-effectors with enough grasping force, as well as capabilities of force sensing while controlling interactions between the nanoobject and the tool or the substrate.

In this entry, a concept is presented to make a prototype of an AFM-based flexible robotic system (FRS) which is equipped with two collaborative cantilevers. By reconfiguring the modular hardware and software of the AFM-FRS, different configurations can be obtained. With different manipulation strategies and protocols, two of the configurations can be respectively used to complete pick-and-place manipulations from the microscale of several microns to the nanoscale that are still challenges.

This entry is organized as follows. Section I introduces the prototype and experimental setup of the AFM-FRS. In subsection II and III, pick-and-place manipulation of microspheres and nanowires for building 3-D micro/nano structures are performed using the AFM-FRS.

AFM-Based Flexible Robotic System

AFM-FRS Setup

Figure 1a shows the AFM-FRS system setup, which is in the configuration for nanoscale pick-and-place. The system is equipped with an optical microscope and two sets of modules commonly used in a conventional AFM, mainly including two AFM cantilevers (namely, tip I and tip II, ATEC-FM Nanosensors, see Fig. 1b), two sets of nanopositioning devices and optical levers. The motion modules include an open-loop *X-Y-Z* piezoscanner (PI P-153.10H), an *X-Y-Z* closed-loop nanostage (MCL Nano-Bio2M on the *X*- and 7-axis, PI P-732. ZC on the Z-axis), an *X-Y-Z* motorized stage, and an *X-Y-Z* manual stage. Detailed specifications of the motion modules are summarized in Table 1. Figure 1c shows an optical microscope image of the collaborating tips in microsphere grasping mode.

A data acquisition (NI 6289) card is used for high-speed (500 ~ 800 Hz of sampling frequency for force and 600 kHz for amplitude) capture of the photodiode voltage output to estimate deflections on both tips induced by force loading or resonant oscillation. A multi-thread planning and control system based on the C^{++} is developed for AFM image scan and two-tip coordination control during manipulation. This control system enables programming of complex tasks on the highly distributed reconfigurable system.

Force Sensing During Pick-and-Place

Figure 2 shows a schematic diagram of the nanotip gripper for the micro/nanoscale pick-and-place operation that has a clamping angle $\theta \approx 44^\circ$ micro/nanoscale pickup manipulation, for instance, interactive forces applied on tip I include repulsive forces F_{r1}, friction forces F_{f1}, and adhesive forces F_{a1}.

The forces applied on tip I can be resolved into two components on the *X*-axis and the *Z*-axis in the defined frame, namely, F_{x1} and F_{z1}, respectively. F_{x1} is the clamping force that holds the micro/nanoobject. F_{z1} is the pickup force that balances adhesion forces from the substrate. To sense the pickup force, it is necessary to know the normal deflection on both cantilevers. The normal deflection ζ_{n1} associated with the normal voltage output of the optical lever on tip I is given by:

$$\zeta_{n1} = \frac{F_{z1}\cos\gamma + F_{x1}\sin\gamma}{k_n} + \frac{F_{z1}\sin\gamma + F_{x1}\cos\gamma}{k_{xz}} \quad (1)$$

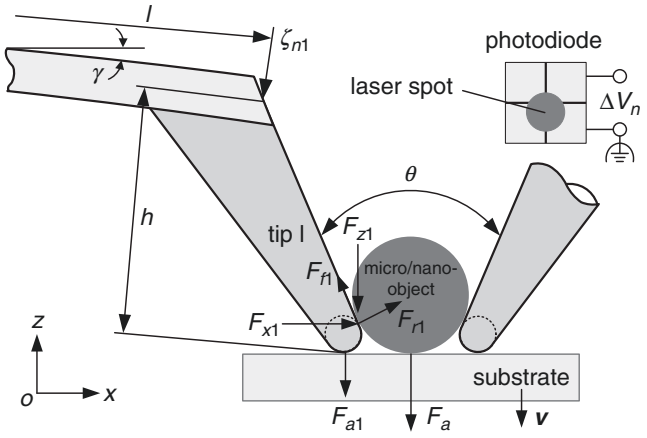
where γ is the mounting angle of the cantilever, $kxz = 2lk_n/3\ h$ is the bending stiffness due to the moment applied on the tip end, where h is the tip height. Assuming the magnitude of F_{z1} and F_{x1} are of the same order, contributions from the Fx_1 to the normal deflection of tip I are relatively very small since $k_{xz} \gg k_n$ and $\gamma = 5^\circ$. Therefore, the normal deflection induced from F_{z1} is only considered in the following calculations of the adhesion force F_a applied on the micro/nanoobject. Thus, F_{z1} can be simplified by estimating the normal voltage output ΔV_{n1} from the tip I:

$$F_{z1} = \beta_1 \times \Delta V_{nl} \quad (2)$$

where β_1 is the normal force sensitivity of the optical lever. A similar pickup force F_{z2} can be also obtained on Tip II. Before the gripper pulls

3D Micro-/Nanomanipulation with Force Spectroscopy, Table 1 Specifications of each motion module

Actuator	Travel range	Resolution
X-Y-Z piezoscanner	10 × 10 × 10(μm)	Sub-nm
X-Y-Z nanostage	50 × 50 × 10(μm)	0.1 nm
X-Y-Z motorized stage	25 × 25 × 25(mm)	50 nm
X-Y-Z manual stage	5 × 5 × 5(mm)	0.5 μm



3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 2 Schematic diagram of the nanotip gripper and force simulation during the pickup manipulation

T

off the substrate, the adhesion force F_a can be estimated as:

$$F_a = F_{z1} + F_{z2} = \beta_1 \Delta V_{n1} + \beta_2 \Delta V_{n2} - (F_{a1} + F_{a2}) \quad (3)$$

where β_2, ΔV_{n2}, and F_{a2} are respectively the normal force sensitivity, normal voltage output, and adhesive force on tip II. Once the gripper pulls off the substrate, e.g., in the case of nanowire/tube pick-and-place, the adhesion force F_a is estimated as:

$$F_a = F_{z1} + F_{z2} = \beta_1 \Delta V_{n1} + \beta_2 \Delta V_{n2}. \quad (4)$$

Experimental Results

3-D Micromanipulation Robotic System

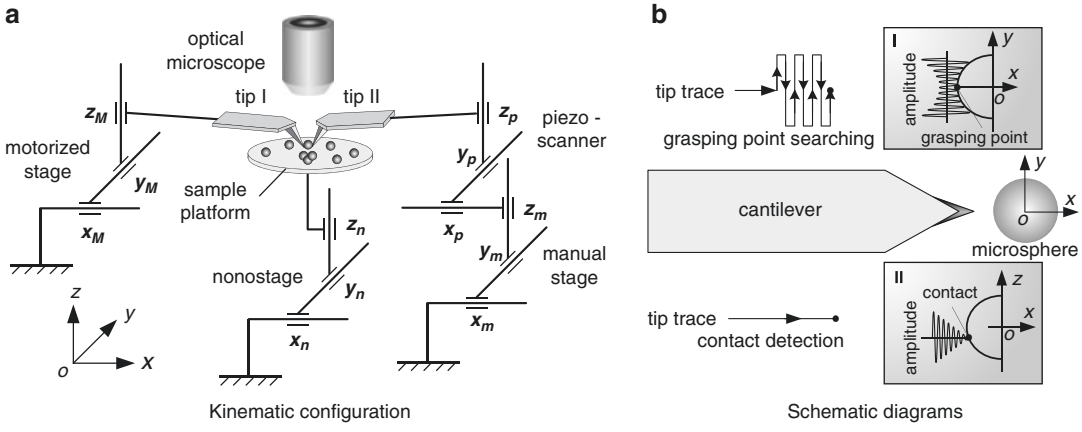
System Configuration for 3-D Micromanipulation

As the size of microobjects is reduced to several microns or submicrons, problems will arise with these conventional grippers: (1) Sticking phenomena becomes more severe due to the relatively larger contact area between the gripper and the microobject. (2) The tip diameters of the microfabricated clamping jaws are comparable in size to the microobjects to be grasped. Conventional grippers are not geometrically sharp enough to pick up microobjects of several micrometers deposited on the substrate. Fortunately, the AFM tip has a very tiny apex (typically ~10 nm in radius) with respect to the size of the microobject to be manipulated. Thus, the nanotip gripper can be used to achieve pick-and-place at the scale of several microns since the contact area of the gripper-microobject is much smaller than the microobject-substrate contact. Moreover, real-time force sensing makes the manipulation more controllable.

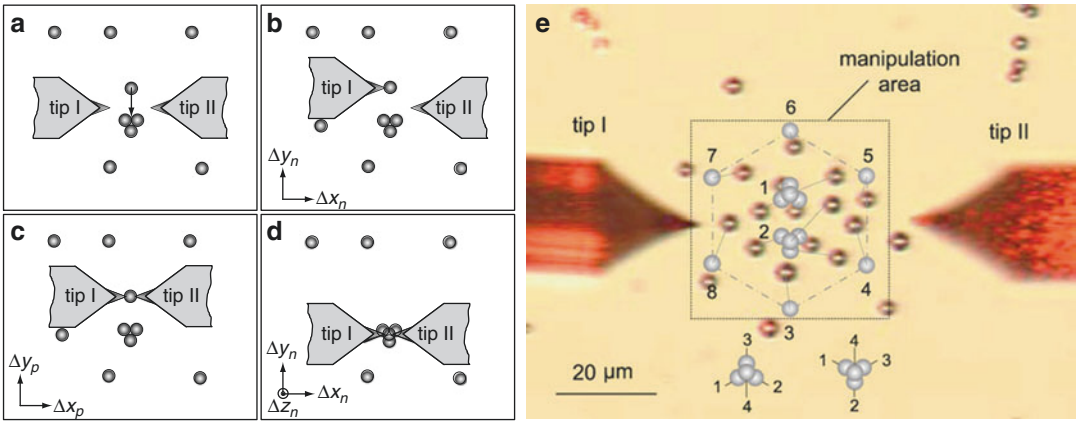
As shown in Fig. 3, the system configuration for 3-D micromanipulation is reconfigured as follows:

1. For a large manipulation travel range, the nanostage here is used to support the sample platform and transport the microobject during the manipulation.
2. Tip I, immovable during the pick-and-place micromanipulation, is fixed on the motorized stage for coarse positioning.
3. Tip II is actuated by the piezoscanner for gripper opening and closing operations. The piezoscanner is supported by the manual stage for coarse positioning.

Benefiting from AFM-based accurate and stable amplitude feedback of a dithering cantilever, the grasping state can be successfully achieved by the amplitude feedback, with very weak



3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 3 A kinematic configuration and schematic diagrams for 3-D micromanipulation at a scale of several micrometers or grasping point searching (*inset* I) and contact detection (*inset* II) with amplitude feedback of the dithering cantilever



3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 4 Microsphere manipulation protocol. (**a**) Task planning. (**b**) Tip I and the microsphere are in contact. (**c**) The nanotip gripper is formed. (**d**) Pick up and release the microsphere to its target position. (**e**) Task descriptions of the microsphere assembly

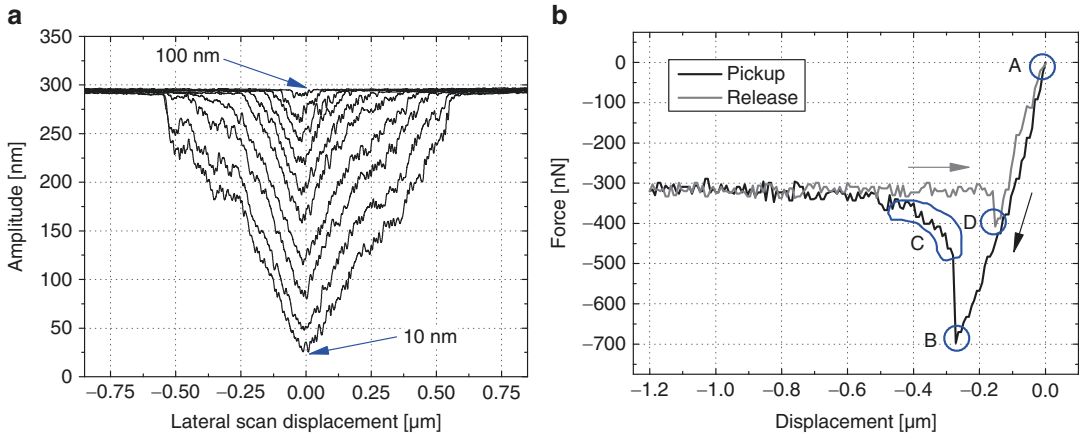
interaction at the nano-Newton scale, protecting the fragile tips and the microobjects from damage during manipulation. As shown in inset I of Fig. 3, the dithering cantilever with its first resonant mode is used to locate grasping points and detect contact. When approaching the microsphere with a separation between the tip and the substrate (typically 500 nm), the tip laterally sweeps the microsphere over the lower part of the microsphere. By this means, the grasping point can be accurately found by locating the minimum amplitude response of each single scan.

From the scheme depicted in inset II of Fig. 3, the amplitude feedback is also used for contact detection. Tip-microsphere contact is detected as the amplitude reduces to a steady value close to zero. As shown in Fig. 4, a protocol for pick-and-place microspheres mainly consists of four steps:

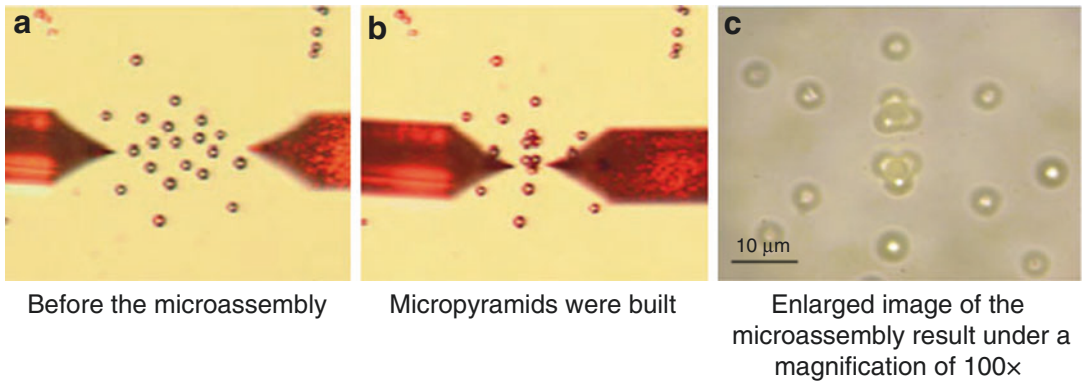
- System initialization and task planning: Each axis of the nanostage and the piezoscanner are set in a proper position, supplying the manipulation with enough travel range on each axis. Then the task is planned in Fig. 4a with a global view of the manipulation area that provides coarse positions of the microspheres and tips.
- Making tip I-microsphere in contact: In Fig. 4b, tip I has started to approach the microsphere by moving the nanostage with amplitude feedback to search for the grasping point and detect contact.
- Forming the gripper: Similarly, tip II approaches the microsphere by moving the piezoscanner. Once tip II and the microsphere are in contact, a nanotip gripper is configured in Fig. 4c for a manipulation.
- Pick-and-place micromanipulation: In Fig. 4d, the microsphere is picked up, transported, and released by moving the nanostage with a proper displacement on each axis that depends on the diameter of the microsphere and its destination. The whole process of 3-D micromanipulation is monitored by real-time force sensing.

3-D Microsphere Assembly

Nylon microspheres with diameter of about 3 ~ 4 μm were manipulated to build 3-D microstructures in experiments. The microspheres were deposited on a freshly cleaned glass slide and then an area of interest was selected under the optical microscope. Figure 4 shows a plan view of the selected area, in which 14 microspheres separated in a 50 μm square frame are going to be manipulated to build two 3-D micropyramids and a regular 2-D hexagon labeled by assembly sequences from 1 to 8. Each pyramid is constructed from four microspheres with two layers and the



3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 5 Experimental results

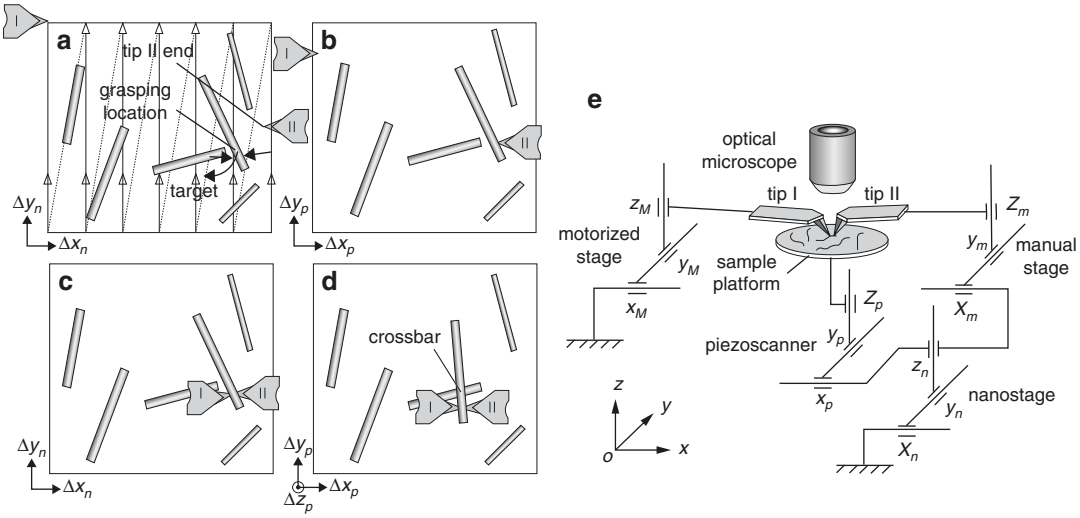


3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 6 A microassembly result

assembly sequences are shown in the bottom insets for two different arrangements of the pyramids.

Figure 5 shows a result of grasping point searching, in which the dithering tip II laterally sweeps the microsphere within a range of 1.75 μm on the *y*-axis and with a free oscillating amplitude of about 285 nm. Ten different distances to the microsphere were tested from 100 to 10 nm with an interval of 10 nm and, consequently, the grasping point is well located with an accuracy of ±10 nm. Figure 5 shows a full force spectroscopy curve during the pick-and-place of a microsphere deposited on a glass slide with an ambient temperature of 20 °C and relative humidity of 40 %. In this curve, point A represents the start of the pick-and-place, point B the pull-off location of the microsphere-substrate contact, point C nonlinear force restitution due to the tip-microsphere frictions, and point D the snap-in point between the microsphere and the substrate. The force spectroscopy curve is synthesized from force responses on tip I and tip II.

Figure 6 shows an automated microassembly result consisting of two 3-D micropyramids and a 2-D pattern of a regular hexagon. The whole



3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 7 A protocol for nanowire pick-and-place (**a**) Image scan for task planning. (**b**) Tip II is in contact with the nanowire. (**c**) Tip I is in contact with the nanowire, forming a nanotip gripper. (**d**) Pick up and release a nanowire to its target position (**e**) A kinematic configuration of the AFM-RFS for nanoscale pick-and-place

manipulation process was completed in 11 min, so the average manipulation time for each microsphere is about 47 s, which mainly breaks down into about 20 s for microsphere grasping including the grasping point search and contact detection processes using amplitude feedback, 10 ~ 35 s for microsphere release and, the remaining time for transport.

Assembly of the fourth is the key to success in building a micropyramid. During pick-and-place of the fourth microsphere, microscopic vision was firstly used for coarse positioning it target, then the normal force feedback of the gripper was used to detect the vertical contact between the fourth microsphere and other three microspheres on the base. When the contact is established, a small vertical force was applied on the fourth microsphere by moving the nanostage upward and it will adjust to contact with all the base microspheres.

3-D Nanomanipulation Robotic System

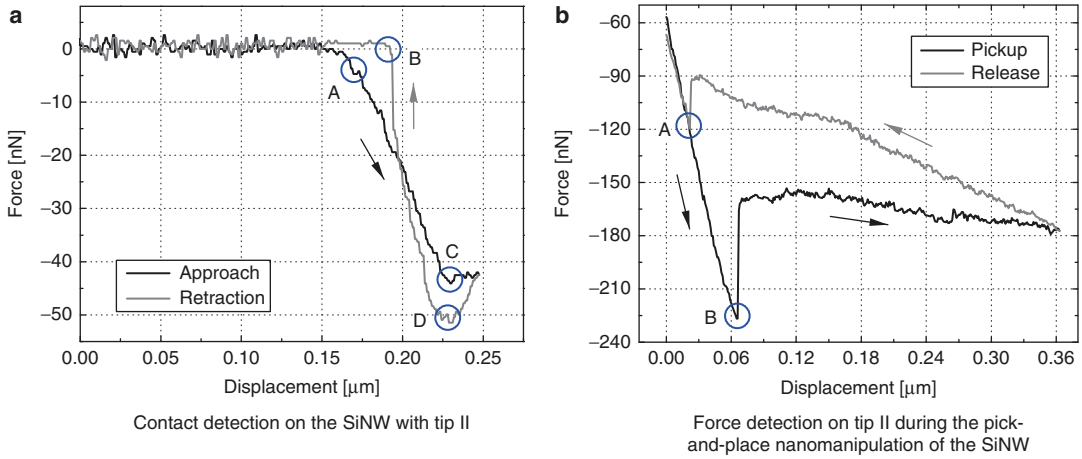
System Configuration for 3-D Nanomanipulation

Compared with the 3-D micromanipulation, tip alignment precision is the key factor in succeeding the nanoscale pick-and-place. Therefore, the closed-loop nanostage is considered in this configuration for accurate tip alignment. As shown in Fig. 7, the system configuration for 3-D nanomanipulation is reconfigured as follows:

1. The nanostage is used for image scan with tip I.
2. Nanoobjects are supported and transported by the piezoscanner.
3. Tip I, fixed on the motorized stage for coarse positioning, is immovable during the pick-and-place micromanipulation. Before manipulation, cantilever I acts as an image sensor for nanoobject positioning.
4. For accurate gripper alignment between Tip I and Tip II, Tip II is fixed on the nanostage rather than the piezoscanner. Tip II is supported by the manual stage for coarse positioning.

Nanowires and nanotubes are being intensively investigated. Thus, a protocol is developed here for nanowire or nanotube pick-and-place. However, applications can easily be extended to, for example, pick-and-place of nano-rods or nanoparticles dispersed on a substrate.

Once the manipulation area is selected under the optical microscope, both the tips are aligned as



3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 8 Contact and force detection

a quasi-gripper above the center of the manipulation area. Each axis of the nanostage and the piezoscanner is initialized at an appropriate position to allow for enough manipulation motion travel.

In this step, tip I is used to fully scan the relevant area obtaining a topographic image that contains nanoobjects to be manipulated and the end of tip II. Figure 7a shows a simulated image that contains the topography of two nanowires and the end of tip II. The image provides the following pick-and-place with relative positions between tip I, tip II and the nanoobjects to be manipulated. However, after a long period image scan, relocating tip II is recommenced to eliminate system's thermal drift.

As shown in Fig. 7b, tip II approaches the nanowire to make contact by moving the *X*-axis of the piezoscanner. A gap (typically ~20 nm above the snap-in boundary) between Tip II and the substrate should be maintained during the approach to enable a negative deflection response in the form of a tiny force applied on tip II, and hence, sensitive detection of the tip-nanowire contact.

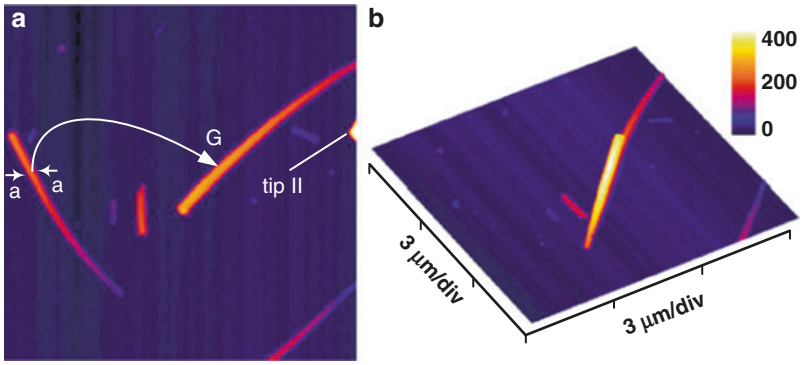
Similarly, in Fig. 7c, once tip I is in contact with the nanowire, a nanotip gripper is configured for pick-and-place manipulation of the nanowire.

The nanotip gripper in this step is used to pick up, transport, and release the nanowire to its target position by moving the piezoscanner on the *X*, *Y*, or *Z*-axis. The displacement on each axis depends on the dimensions of the nanowire and the location of the destination. Figure 7d shows a simulated post-manipulation image, in which a nanowire crossbar is built. The complete pick-and-place procedure is monitored by force sensing.

Pick-and-Place Nanomanipulation

In experiments, silicon nanowires (SiNWs) were deposited on a freshly cleaned silicon wafer coated with 300 nm silicon dioxide. AFM images show that the SiNWs have a taper shape and have diameters of 25 nm (top), ~200 nm (root), and lengths of about 4 ~ 7 μm.

Figure 8 shows an example of the contact detection with tip II: Point A and point C are where the tip contacts with the SiNW and the Si substrate, respectively; Point B and point D are where the tip breaks the contact with the Si substrate and the SiNW, respectively. Figure 8 shows a curve of the peeling force spectroscopy on tip II for the pick-and-place manipulation of the SiNW: point A and point B are where the tip snaps in and pulls off the Si substrate, respectively. The shape of curve of the force responses on tip I are similar except for the force magnitude due to different force sensitivities on each tip and uneven grasping due to asymmetric alignment of the SiNW relative to the grasping direction. This force spectroscopy during the pickup operation shows stable grasping for further SiNW transport.



3D Micro-/Nanomanipulation with Force Spectroscopy, Fig. 9 Pick-and-place results of the SiNWs. (**a**) A pre-scanned image. In which a–a and G are the grasping location and the target position, respectively. (**b**) A post-manipulation image verifies that the manipulated SiNW is piled upon another nanowire deposited on the Si substrate

Figure 9 shows an experimental result of 3-D SiNW manipulation. A prescanned image (9 × 9 µm) is shown in Fig. 9a, which includes the topographic image of SiNWs and the local image of tip II. A grasping location of the nanowire to be manipulated is marked with a–a, where the SiNW has a height of 160 nm. Figure 9b is a post-manipulation image. It can be seen that the SiNW has been successfully transported and piled onto another SiNW. The manipulation procedure is described as follows. Once the SiNW was reliably grasped, the piezoscanner moved down 560 nm at a velocity of 80 nm/s. In this step, the SiNW was transported a distance of 4.4 µm along the *X*-axis at a velocity of 120 nm/s and 0.12 µm along the *Y*-axis at a lower velocity of 3.3 nm/s. In the releasing step, the piezoscanner moved up at a velocity of 100 nm/s. As tip II was slightly bent upward leading to a positive response of 0.015 V, tip I and tip II were separated by moving both the nanostage and the piezoscanner on the *X*-axis to release the SiNW from the nanotip gripper.

Conclusion and Future Directions

A flexible robotic system developed for multi-scale manipulation and assembly from nanoscale to microscale is presented. This system is based on the principle of atomic force microscopy and comprises two individually functionalized cantilevers. After reconfiguration, the robotic system could be used for pick-and-place manipulation from nanoscale to the scale of several micrometers. Flexibilities and manipulation capabilities of the developed system are validated by pick-and-place manipulation of microspheres and silicon nanowires to build three-dimensional micro/nanostructures in ambient conditions. Complicated micro/nano manipulation and assembly can be reliably and efficiently performed using the proposed flexible robotic system. 3-D nanomanipulation methods are indispensible for heterogeneous integration of complex nanodevices. Serial nanomanipulation systems could only enable low-volume prototyping applications. However, high-volume and high-speed nanomanipulation systems are indispensable for future nanotechnology products. Therefore, autonomous and massively parallel AFM systems are required.

Cross-References

- ▶ AFM
- ▶ Atomic Force Microscopy
- ▶ Force Modulation in Atomic Force Microscopy
- ▶ Friction Force Microscopy
- ▶ Manipulating
- ▶ Nanogrippers
- ▶ Nanorobotic Assembly
- ▶ Nanorobotics
- ▶ Self-Assembly of Nanostructures

References

1. Menciassi, A., Eisinberg, A., Izzo, I., Dario, P.: From "macro" to "micro" manipulation: models and experiments. IEEE-ASME Trans. Mechatronics. **9**(2), 311–320 (2004)
2. Lambert, P., Régnier, S.: Surface and contact forces models within the framework of microassembly. J. Micromechatron. **3**(2), 123–157 (2006)
3. Sitti, M.: Microscale and nanoscale robotics systems-characteristics, state of the art, and grand challenges. IEEE Robot. Autom. Mag. **14**(1), 53–60 (2007)
4. Xie, H., Rong, W.B., Sun, L.N.: Construction and evaluation of a wavelet-based focus measure for microscopy imaging. Microsc. Res. Tech. **70**, 987–995 (2007)
5. Liu, X.Y., Wang, W.H., Sun, Y.: Dynamic evaluation of autofocusing for automated microscopic analysis of blood smear and pap smear. J. Microsc. **227**, 15–223 (2007)
6. Xie, H., Rong, W.B., Sun, L.N., Chen, W.: Image fusion and 3-D surface reconstruction of microparts using complex valued wavelet transforms. In: IEEE International Conference on Image Process, Atlanta, pp. 2137–2140 (2006)
7. Xie, H., Vitard, J., Haliyo, S., Régnier, S.: Optical lever calibration in atomic force microscope with a mechanical lever. Rev. Sci. Instrum. **79**, 096101 (2008)
8. Xie, H., Rakotondrabe, M., Régnier, S.: Characterizing piezoscanner hysteresis and creep using optical levers and a reference nanopositioning stage. Rev. Sci. Instrum. **80**, 046102 (2009)
9. Vögeli, B., von Knel, H.: AFM-study of sticking effects for microparts handling. Wear **238**, 20–24 (2000)
10. Varenberg, M., Etsion, I., Halperin, G.: An improved wedge calibration method for lateral force in atomic force microscopy. Rev. Sci. Instrum. **74**, 3362–3367 (2003)
11. Xie, H., Vitard, J., Haliyo, S., Régnier, S.: Enhanced accuracy of force application for AFM nanomanipulation using nonlinear calibration of optical levers. IEEE Sens. J. **8**(8), 1478–1485 (2008)
12. Fukuda, T., Arai, F., Dong, L.X.: Assembly of nanodevices with carbon nanotubes through nanorobotic manipulations. Proc. IEEE **91**, 1803–1818 (2003)
13. Dong, L.X., Arai, F., Fukuda, T.: Electron-beam-induced deposition with carbon nanotube emitters. Appl. Phys. Lett. **81**, 1919–1921 (2002)
14. Dong, L.X., Arai, F., Fukuda, T.: Destructive constructions of nanostructures with carbon nanotubes through nano-robotic manipulation. IEEE/ASME Trans. Mechatron. **9**, 350–357 (2004)
15. Molhave, K., Wich, T., Kortschack, A., Boggild, P.: Pick-and-place nanomanipulation using microfabricated grippers. Nanotechnology **17**, 2434–2441 (2006)
16. Dong, L.X., Tao, X.Y., Zhang, L., Nelson, B.J., Zhang, X.B.: Nanorobotic spot welding: controlled metal deposition with attogram precision from copper-filled carbon nanotubes. Nano Lett. **7**, 58–63 (2007)
17. Leach, J., Sinclair, G., Jordan, P., Courtial, J., Padgett, M.J., Cooper, J., Laczik, Z.J.: 3D manipulation of particles into crystal structures using holographic optical tweezers. Opt. Express. **12**, 220–226 (2004)
18. Yu, T., Cheong, F.C., Sow, C.H.: The manipulation and assembly of CuO nanorods with line optical tweezers. Nanotechnology **15**, 1732–1736 (2004)
19. Bosanac, L., Aabo, T., Bendix, P.M., Oddershede, L. B.: Efficient optical trapping and visualization of silver nanoparticles. Nano Lett. **8**, 1486–1491 (2008)
20. Rollot, Y., Régnier, S., Guinot, J.C.: Simulation of micro-manipulations: adhesion forces and specific dynamic models. Int. J. Adhes. Adhes. **19**, 35–48 (1999)

Three-Dimensional Imaging of Human Tissues

► Imaging the Human Body Down to the Molecular Level

Three-Phase Contact Line

► Surface Engineering, Tailored Wettability, and Applications

THz

► Terahertz Technology for Nano Applications

Tight-Binding Simulations of Nanowires

Alberto García-Cristóbal[1] and Alejandro Molina-Sánchez[2]
[1]Universidad de Valencia, Valencia, Spain
[2]University of Luxembourg, Luxembourg, Luxembourg

Synonyms

Electronic properties; Electronic structure calculations; Nanodevices; Nanostructures; Nanowires; Optical properties; Semiempirical methods; Transport properties

Definition

The tight-binding (TB) method [1] refers to both a theoretical formalism and a practical procedure for the calculation of the single-particle electronic structure of solid-state systems. The essence of the method is the representation of the energy eigenvalue problem in a basis of atomic-like states. Although it can be formulated as an exact method, its most extended use is the so-called semiempirical version. This version is based on (a) several plausible and well-controlled approximations, (b) the judicious simplification of the problem taking into account the symmetry of the system, and (c) finally, the consideration of the remaining Hamiltonian matrix elements as parameters to be fitted against reliable experimental or theoretical data. Although designed originally for bulk systems, it can be easily adapted to the calculation of the electronic structure of nanostructures. The TB calculations permit to investigate phenomena at an atomic scale hardly accessible by other techniques, thus emerging as a fundamental tool for addressing open issues regarding nanostructured systems. Among them, the class of systems called nanowires (NWs) [2] is particularly interesting. For the purposes of this essay, nanowires must be understood as one-dimensional (1D) free-standing single-crystal structures formed by bottom-up synthesis, with diameters of a few to tens of nanometers and lengths of tens of nanometers to tens of microns. They have received much attention since, by virtue of its large surface-to-volume ratio, they exhibit properties that are interesting for the applications in fields such as electronics, photonics, sensors, and the life sciences.

Foundations of the Tight-Binding Method

The tight-binding (TB) methodology provides a systematic approach to the calculation of the electronic structure of multiatom systems, from molecules to macroscopic bulk solids, passing by intermediate systems, such as atomic or molecular wires, nanostructures, and in particular nanowires. In general terms, the goal is to solve for the single-particle energy spectrum determined by the mean-field potential $V(\boldsymbol{r})$ associated to a system of atoms located at positions $\boldsymbol{R}_I$. In order to do so, one must solve the Schrödinger equation:

$$H\,\Psi(\boldsymbol{r}) \;=\; [\,T \;+V(\boldsymbol{r})\,]\,\Psi(\boldsymbol{r}) = \;E\,\Psi(\boldsymbol{r}). \quad (1)$$

The driving concept behind the TB method, introduced by Felix Bloch already in 1928 [3], is the attempt at solving the eigenvalue problem (1) by expanding $\Psi(\boldsymbol{r})$ as a linear combination in terms of a complete basis of orbitals $\psi_{\boldsymbol{R}_I,a}(\boldsymbol{r})$ localized around the atomic sites $\boldsymbol{R}_I$. These orbitals can correspond to some reference atomic problem or simply have been chosen by its convenient mathematical form. In the following, we call them atomic-like orbitals irrespective if they carry or not any physical meaning. The label a contains the necessary information to fully specify the orbital, and in particular, it specifies its symmetry properties.

The potential V in Eq. 1 contains the interaction of the electron with the system of nuclei and the best possible mean-field account of the electron-electron interaction effects. One reasonable starting point would be the so-called ab initio GW approximation. In practice, this complicated unknown potential is further approximated at various levels, leading, e.g., to Kohn–Sham, Hartree–Fock, and Hartree potentials. In principle, these potentials and the corresponding eigenfunctions must be solved self-consistently.

Rigorously speaking, one should take a complete infinite set of orbitals on every atomic site starting from the lowest one. However, the single-particle states relevant for most physical and chemical properties are those with energies in the neighborhood of the Fermi energy or around the fundamental bandgap. Therefore, many approaches to the electronic structure introduce the so-called frozen-core approximation, in which the core (nucleus together with inner closed electronic shells) is represented by a pseudopotential. In the context of the TB method, this allows to select a truncated minimal set of orbitals on each atomic site, which is expected to be sufficient for an

accurate description of the highest occupied and lowest unoccupied states of the system. In virtually all implementations of the TB method, one works within the frozen-core approximation with a finite basis of atomic-like orbitals.

Since most of the applications of the TB method to nanostructures are based on their bulk counterpart, it is convenient to first review the TB method for infinite crystalline systems. In this case, the method proceeds in a special way to take advantage of the Bloch theorem [3, 4]. First, in a crystal, the atomic sites must be relabeled as $\boldsymbol{R}_l \rightarrow \boldsymbol{R}_n + \boldsymbol{d}_\alpha$, where $\boldsymbol{R}_n$ points to the lattice site labeled by integer numbers abbreviated as n and $\boldsymbol{d}_\alpha$ is the position of atom of type α within the primitive unit cell. Next, it is convenient to construct, for each orbital a associated to atom α, the following function, called Bloch sum:

$$\chi_{\alpha,a}^{(\boldsymbol{k})}(\boldsymbol{r}) = \sqrt{\frac{\Omega_c}{(2\pi)^3}} \sum_{\boldsymbol{R}_n} e^{i\boldsymbol{k}\cdot(\boldsymbol{R}_n + \boldsymbol{d}_\alpha)}\, \psi_{\boldsymbol{R}_n,\alpha,a}(\boldsymbol{r}), \tag{2}$$

where the wavevector $\boldsymbol{k}$ runs over the Brillouin zone and Ω_c is the volume of the unit cell. The Bloch sums satisfy the Bloch theorem and therefore are best suited for an expansion of the crystal states. It is convenient to first introduce the composite label $j = (\alpha, a)$ and then write the energy eigenfunction $\Psi_{\boldsymbol{k}}(\boldsymbol{r})$ as a linear combination of all the Bloch sums (2), with coefficients $C_j^{(\boldsymbol{k})}$ to be determined:

$$\Psi_{\boldsymbol{k}}(\boldsymbol{r}) = \sum_j C_j^{(\boldsymbol{k})}\, \chi_j^{(\boldsymbol{k})}(\boldsymbol{r}). \tag{3}$$

The insertion of ansatz (3) into the original eigenvalue problem (1) leads to the generalized matrix eigenvalue problem:

$$\sum_{j'} \left[H_{jj'}^{(\boldsymbol{k})} - E(\boldsymbol{k}) S_{jj'}^{(\boldsymbol{k})} \right] C_{j'}^{(\boldsymbol{k})} = 0, \tag{4}$$

where the Hamiltonian matrix elements are given by

$$H_{jj'}^{(\boldsymbol{k})} = \sum_{\boldsymbol{R}_{n'}} e^{i\boldsymbol{k}\cdot(\boldsymbol{R}_{n'} + \boldsymbol{d}_{\alpha'} - \boldsymbol{d}_\alpha)}\, H_{0,j;\boldsymbol{R}_{n'},j'}, \tag{5a}$$

$$H_{\boldsymbol{R}_n,j;\boldsymbol{R}_{n'},j'} = \int \psi_{\boldsymbol{R}_n,j}^*(\boldsymbol{r})\, H\, \psi_{\boldsymbol{R}_{n'},j'}(\boldsymbol{r})\, d^3\boldsymbol{r}, \tag{5b}$$

and the overlap matrix elements are given by

$$S_{jj'}^{(\boldsymbol{k})} = \sum_{\boldsymbol{R}_{n'}} e^{i\boldsymbol{k}\cdot(\boldsymbol{R}_{n'} + \boldsymbol{d}_{\alpha'} - \boldsymbol{d}_\alpha)}\, S_{0,j;\boldsymbol{R}_{n'},j'}, \tag{6a}$$

$$S_{\boldsymbol{R}_n,j;\boldsymbol{R}_{n'},j'} = \int \psi_{\boldsymbol{R}_n,j}^*(\boldsymbol{r})\, \psi_{\boldsymbol{R}_{n'},j'}(\boldsymbol{r})\, d^3\boldsymbol{r}. \tag{6b}$$

In some cases, it is necessary to include into the Hamiltonian H the effects of the spin–orbit interaction. This is almost always done by (i) changing to a basis of spin-orbitals $\psi_{\boldsymbol{R}_n,j} X_\sigma$, where X_σ $(\sigma = \uparrow, \downarrow)$ are the spin up and down spinors, and (ii) adding an atomic spin–orbit interaction energy term $H^{(\mathrm{SO})}$ so that the Hamiltonian matrix elements become [5]

$$H_{\boldsymbol{R}_n,j;\boldsymbol{R}_{n'},j'} \rightarrow \delta_{\sigma\sigma'} H_{\boldsymbol{R}_n,j;\boldsymbol{R}_{n'},j'} + \left(1 - \delta_{\sigma\sigma'}\right) \delta_{\boldsymbol{R}_n\boldsymbol{R}_{n'}} \delta_{\alpha\alpha'} H_{a,\sigma;a',\sigma'}^{(\mathrm{SO})}.$$

The necessary input to solve Eq. 4 are the matrix elements $H_{\boldsymbol{R}_n,j;\boldsymbol{R}_{n'},j'}$ and $S_{\boldsymbol{R}_n,j;\boldsymbol{R}_{n'},j'}$ between the atomic-like orbitals. The orbitals associated to a given atomic site are easily forced to be orthonormal, but in general there will be nonzero overlaps $S_{\boldsymbol{R}_n,j;\boldsymbol{R}_{n'},j'}$ between orbitals taken on different sites, therefore leading to a generalized eigenvalue problem. This is amenable to numerical solution, but the computational burden is heavier than for standard diagonalization problems. Fortunately, there exist formal procedures that allow, starting from the nonorthogonal orbitals on a given crystalline structure, $\psi_{\boldsymbol{R}_n,j}$, to obtain a set of orthonormal orbitals $\varphi_{\boldsymbol{R}_n,j}$ (Löwdin orbitals) with the same site symmetry as the original ones [6, 7]. The Löwdin orbitals are still localized in the vicinity of the atomic sites, but they happen to have a larger spatial extent than the original ones. When the TB method is

formulated in terms of the Löwdin orthonormal basis, then $S_{\boldsymbol{R}_n,j;\boldsymbol{R}_{n'},j'} \rightarrow \delta_{\boldsymbol{R}_n,\boldsymbol{R}_{n'}}\delta_{j,j'}$ and Eq. 4 reduces to a more manageable standard eigenvalue problem.

The procedure explained contains the core of the TB methodology as applied to bulk systems. In principle, one can obtain at this point, for each $\boldsymbol{k}$, the energies $E(\boldsymbol{k})$ and eigenfunctions $\Psi_{\boldsymbol{k}}(\boldsymbol{r})$ by solving the matrix eigenvalue problem (4). However, this requires the prior knowledge of all the Hamiltonian (and overlap) matrix elements in the basis of atomic-like orbitals, which is still a formidable problem. Therefore, before attempting the solution of Eq. 4, the calculation must be made affordable by incorporating several considerations and approximations that we summarize below [4]:

1) The summations in Eqs. 5a and 6a can be organized as an expansion in neighbor shells of increasing distance. Since it is reasonable to expect that the integrals in Eqs. 5b and 6b will decrease as the interatomic distance increases, a first natural simplification is to truncate that expansion after a certain number of shells. One can then use only nearest neighbors or only nearest and second nearest, etc.
2) Since the system potential V can always be decomposed as a sum involving contributions $V_\alpha(\boldsymbol{r} - \boldsymbol{d}_\alpha)$ centered on the atomic sites, then the Hamiltonian matrix elements are formally classified into one-center, two-center, and also three-center integrals. The three-center integrals are certainly smaller than the two-center integrals, but they are more numerous. Thus, its importance is not clear a priori. Notwithstanding, a popular choice at this point is the so-called two-center approximation, in which all three-center integrals are neglected.
3) In general terms, many of the matrix elements Eq. 5b and Eq. 6b are related through demands of crystal symmetry. These relations can be elucidated with the help of group theory and lead to a reduction of the number of independent matrix elements to consider. This simplification procedure is particularly transparent and effective when the atomic-site potentials are approximated as spherically symmetric, i.e., $V_\alpha(|\boldsymbol{r} - \boldsymbol{d}_\alpha|)$, and the orbitals are chosen to have definite angular momentum. Then the treatment of the multicenter integrals becomes equivalent to the one developed for molecular structure calculations. Slater and Koster [4] showed how to reduce all the two-center integrals, under the above assumptions, to a minimum set of independent matrix elements, which depend solely on the interatomic distance. This represents the maximum possible simplification of the problem.

At this point, one has to decide how to deal with the resulting matrix elements. Their explicit calculation requires the knowledge of the exact potential of the system. As said before, this can be obtained, in a self-consistent manner, from ab initio calculations. This way of proceeding leads to the so-called ab initio tight-binding approach [8, 9]. Alternatively, Slater and Koster proposed in his pioneering work [4] to use the rigorous TB formalism rather as a simple interpolation method. This means that neither the system potential nor the atomic-like orbitals are explicitly calculated, but the few independent matrix elements are instead considered as disposable parameters which are fitted to appropriate targets (experimental data or reliable ab initio calculations). This is the essence of the so-called Slater–Koster semiempirical (or empirical) TB method. It represents a great simplification that leads to an extremely useful approach to understanding electrons in materials. It has been said that "Of all the methods for treating electrons, the Slater-Koster approach provides the simplest, most illuminating picture of electronic states" [1]. The semiempirical TB method has become very popular, and many variants have been devised for the analysis of various properties in many crystals. However, one should remember that, being of semiempirical nature, studies aimed at different aspects of the electronic structure may require different parametrizations even for the same material. One important figure in any implementation of the TB method is the order of the Hamiltonian matrix, $N_H = N_S \times N_{\text{atom}} \times N_{\text{orbital}}$, where N_{atom} is the number of atoms per unit cell, N_{orbital} is the number of

T

orbitals per atom, and N_S is 2 or 1, depending whether spin-orbit is taken into account or not. The TB problem then reduces, for each $\boldsymbol{k}$, to the solution of a matrix eigenvalue problem of order N_H.

A great variety of specific implementations of the TB method are available in the literature. One important feature distinguishing them is the choice of the basis of atomic-like orbitals. Here is a short summary of the most common models found in the literature:

- sp^3 model ($N_{orbital} = 4$): In this model one considers a small basis containing one s orbital and three p orbitals (p_x, p_y, p_z) located on each atom. This is appropriate for a fair description of the upper valence bands and the bandgap energy of direct gap semiconductors. However, they fail to reproduce the conduction band minima of indirect gap semiconductors.
- sp^3s^* model ($N_{orbital} = 5$): It has been found that an effective solution to cure the above deficiencies is to enlarge the sp^3 model by the addition of an auxiliary s^* orbital. The model is now flexible enough to accommodate the possibility of an indirect bandgap but does not reproduce other important features of the conduction band.
- $sp^3d^5s^*$ model ($N_{orbital} = 10$): The inclusion of the full shell of empty d state allows at last a nearly perfect modeling of the relevant part of the band structure, including the lowest conduction bands.

As a summary, a given TB model can thus be characterized by the choice of the basis set of atomic-like orbitals, by its orthogonality or not, by the range of the considered matrix elements (i.e., whether they are kept up to first- , second- , or third-nearest neighbors) and by the enforcement or not of the two-center approximation.

Specifics of Nanowires

The TB method lies at the center of the catalogue of nanostructure simulation methods in both computational cost and reliability. Its main advantages are its atomistic and full Brillouin zone representations. The application of the TB method to nanostructures and in particular to one-dimensional (1D) ones, like nanowires, is relatively straightforward. We concentrate here on free-standing NWs surrounded by vacuum. The procedure to adapt the bulk TB methodology introduced above for the modeling of nanostructures requires first to build the nanostructure geometry. Although there are some treatments of genuine three-dimensional elongated structures, one usually focuses on the study of idealized infinitely long nanowires. The NW axis is conventionally taken as the Z axis. Then one only needs to define the minimal supercell that contains the relevant cross section of the system, since the infinite NW is generated by translating the supercell along the Z axis by the axial lattice vectors $\boldsymbol{Z}_n = n a_z \hat{\mathbf{z}}$, where n is an integer and a_z is the axial lattice parameter. The atoms in the supercell are denoted by μ and located at positions $\boldsymbol{D}_\mu$. A set of atomic-like orbitals $\psi_{\boldsymbol{Z}_n,\mu,a}(\boldsymbol{r})$ is taken on each of these atoms.

On account of the 1D lattice periodicity, the procedure runs parallel to the one presented for bulk systems. First, one has to construct the NW Bloch functions:

$$\chi^{(k_z)}_{\mu,a}(\boldsymbol{r}) = \sqrt{\frac{a_z}{(2\pi)}} \sum_{\boldsymbol{Z}_n} e^{i\,k_z\cdot(\boldsymbol{Z}_n+\boldsymbol{D}_\mu)}\,\psi_{\boldsymbol{Z}_n,\mu,a}(\boldsymbol{r}),$$

where the wavevector $\boldsymbol{k}_z = k_z\hat{\mathbf{z}}$ runs over the one-dimensional Brillouin zone, i.e., $k_z \in \left]-\frac{\pi}{a_z}, +\frac{\pi}{a_z}\right]$. It is convenient to first introduce here the composite label $l = (\mu, a)$ and then write the energy eigenfunction $\Psi_{k_z}(\boldsymbol{r})$ as a linear combination of Bloch sums, with coefficients $c_l^{(k_z)}$ to be determined:

$$\Psi_{k_z}(\boldsymbol{r}) = \sum_l c_l^{(k_z)}\,\chi_l^{(k_z)}(\boldsymbol{r}). \tag{7}$$

When Eq. 7 is inserted into the original eigenvalue problem (1), the following generalized matrix eigenvalue problem results:

$$\sum_{l'}\left[\, h_{ll'}^{(\boldsymbol{k}_z)} \;-\; \epsilon(\boldsymbol{k}_z)\, s_{ll'}^{(\boldsymbol{k}_z)}\right] c_{l'}^{(\boldsymbol{k}_z)} = 0, \qquad (8)$$

where it is seen that the electron energy spectrum is organized into subbands as a function of the wave vector $\boldsymbol{k}_z$. The matrix elements are now given by:

$$h_{ll'}^{(\boldsymbol{k}_z)} = \sum_{\boldsymbol{Z}_{n'}} e^{i\boldsymbol{k}_z\cdot\left(\boldsymbol{Z}_{n'}+\boldsymbol{D}_{\mu'}-\boldsymbol{D}_{\mu}\right)} h_{0,l;\boldsymbol{Z}_{n'},l'}, \qquad (9a)$$

$$h_{\boldsymbol{Z}_n,l;\boldsymbol{Z}_{n'},l'} = \int \psi^*_{\boldsymbol{Z}_n,l}(\boldsymbol{r})\, H\, \psi_{\boldsymbol{Z}_{n'},l'}(\boldsymbol{r})\, d^3\boldsymbol{r}. \qquad (9b)$$

$$s_{ll'}^{(\boldsymbol{k}_z)} = \sum_{\boldsymbol{Z}_{n'}} e^{i\boldsymbol{k}_z\cdot\left(\boldsymbol{Z}_{n'}+\boldsymbol{D}_{\mu'}-\boldsymbol{D}_{\mu}\right)} s_{0,l;\boldsymbol{Z}_{n'},l'}, \qquad (10a)$$

$$s_{\boldsymbol{Z}_n,l;\boldsymbol{Z}_{n'},l'} = \int \psi^*_{\boldsymbol{Z}_n,l}(\boldsymbol{r})\, \psi_{\boldsymbol{Z}_{n'},l'}(\boldsymbol{r})\, d^3\boldsymbol{r}. \qquad (10b)$$

If the interatomic topology and the crystalline structure are retained in the NW structure, the positions of the atoms in the supercell can still be labeled with reference to the underlying bulk lattice. In other words, one can write $\boldsymbol{D}_\mu \rightarrow \boldsymbol{\rho}_m + \boldsymbol{d}_\alpha$, where $\boldsymbol{\rho}_m$ is an in-plane lattice vector pointing to the unit cells in the cross section and m is an integer index, and α indicates the type of the atoms in the unit cell.

The label l can then be rewritten as $l = (\mu, a) = (\boldsymbol{\rho}_m, \alpha, a) \equiv (\boldsymbol{\rho}_m, j)$, and the atomic-like functions and matrix elements become:

$$\psi_{\boldsymbol{Z}_n,l}(\boldsymbol{r}) \rightarrow \psi_{\boldsymbol{Z}_n,\boldsymbol{\rho}_m,j}(\boldsymbol{r}) = \psi_{\boldsymbol{Z}_n+\boldsymbol{\rho}_m,j}(\boldsymbol{r}), \qquad (11)$$

$$h_{\boldsymbol{Z}_n,l;\boldsymbol{Z}_{n'},l'} = \int \psi^*_{\boldsymbol{Z}_n+\boldsymbol{\rho}_m,j}(\boldsymbol{r})\, H\, \psi_{\boldsymbol{Z}_{n'}+\boldsymbol{\rho}_{m'},j'}(\boldsymbol{r})\, d^3\boldsymbol{r}, \qquad (12a)$$

$$s_{\boldsymbol{Z}_n,l;\boldsymbol{Z}_{n'},l'} = \int \psi^*_{\boldsymbol{Z}_n+\boldsymbol{\rho}_m,j}(\boldsymbol{r})\, \psi_{\boldsymbol{Z}_{n'}+\boldsymbol{\rho}_{m'},j'}(\boldsymbol{r})\, d^3\boldsymbol{r}, \qquad (12b)$$

which turn out to be formally similar to the bulk matrix elements, since $\boldsymbol{Z}_n + \boldsymbol{\rho}_m$ is nothing but a three-dimensional lattice vector. However, H corresponds here to the nanowire potential, which is different from the bulk potential, due to the termination of the system at the surface and the possible atomic rearrangements induced by it. A rigorous account of that difference can only be made with ab initio calculations. Although there are a few ab initio tight-binding treatments of small nanowires [10], the application of the TB method to nanowires is still dominated by the semiempirical approach. The use of this approach implies therefore the neglect of the difference between the nanowire and bulk and the assumption that the implicit basis set is the same used in bulk calculations. Under this assumptions, the matrix elements (12) coincide with those in Eq. 5b and Eq. 6b, and the same parametrization of the constituent bulk materials can be used in the modeling of NWs.

Nevertheless, the presence of free surfaces implies the existence of dangling bonds for the surface atoms. The most common procedure to deal with this problem is to passivate the dangling bonds at the surface by adding pseudo-hydrogen atoms [11]. The passivating pseudo-hydrogen atoms are supposed to interact only through their s orbital with the NW atoms. This is somewhat equivalent to the hard-wall boundary condition imposed in the models based on the effective mass approximation. This approach, though crudely representing realistic surface effects, has already proved to be highly successful in describing intrinsic quantum confinement.

If the crystalline structure of the NW is distorted by a moderate change of the bond lengths and angles, as happens, for example, under mechanical loading or when one takes into account the rearrangement of the atoms due to the surface reconstruction or during TBMD simulation (see below), the TB method can still be used, but the TB matrix elements must be recalculated for the new configuration with the help of some scaling law that gives their variation with the interatomic distance.

Finally, once the bulk TB model (with associated order N_H has been chosen, the NW TB problem reduces to the numerical solution, for each k_z, of a matrix eigenvalue problem of order $N_{wire} = N_{xy} \times N_H + N_{surface}$, where N_{xy} is the

number of unit cells in the supercell cross section and $N_{surface}$ is the number of orbitals on the surface passivating pseudo-hydrogen atoms. Due to the local interactions of the model, the resulting matrices are sparse. The standard techniques that have been applied are the Lanczos and iterative Jacobi–Davidson algorithms. In this way, it is possible to compute a finite number of conduction and valence subbands around the NW bandgap.

Applications of the TB Method to Nanowire Modeling

The TB methodology has been applied in a range of studies on NW systems, touching upon all the properties that ultimately depend on the electronic structure. These studies embrace all types (semiconducting, metallic, oxide, and other) of nanowires and nanotubes. Among semiconductor NWs, the whole range of materials, with diamond, zincblende, wurtzite, and other crystalline structures, have been investigated. Most of the studies deal with the calculation of the subband electronic structure of technological semiconductors and its dependence on geometry, composition, orientation, etc. This is so because this information translates immediately into the associated optical and transport properties. Some examples of this kind of calculations are reviewed in the following, and later on other less explored topics will be mentioned.

Subband Electronic Structure of Semiconductor Nanowires

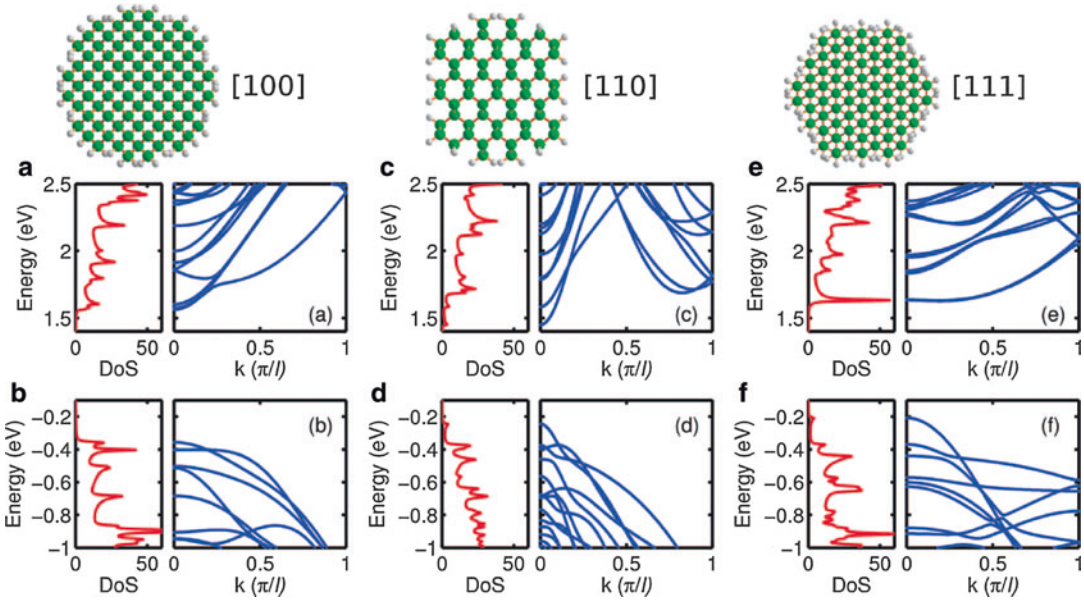
If the nanowire cross section has a size comparable to the de Broglie wavelength, there appear significant quantization effects in the electronic structure that are strongly dependent on the material composition and wire geometry, and this influence needs to be well understood in order to design new devices. At the lowest end of this nanometer scale, the application of the effective mass approximation for the calculation of the electronic structure is highly questionable. The TB methodology stands here as an optimum choice.

Figure 1 displays a comparative study of the subband structures and densities of states (DoS) of a Si NW with radius R = 1 nm for three different orientations, namely, along the [100], [110], and [111] crystal directions [12]. An orthogonal second-nearest neighbor three-center sp^3 TB model, without including the spin–orbit coupling, has been used. The Van Hove singularities typical of 1D subbands are clearly visible in the DoS. It is apparent that the electronic structure obtained is very sensitive to the wire orientation. The position and curvature of the conduction band (CB) minimum is particularly sensitive, exhibiting a different character of direct, quasi-indirect, and indirect gap as the orientation is changed. The general trends can be qualitatively understood in terms of the zone folding of the bulk band structure. While the valence band (VB) maximum always lies at $k_z = 0$, there is significant mixing and splitting of the various valence subbands. In [110]- and [111]-oriented Si NWs, the highest valence subband is rather dispersive and well separated from the other bands. In [100]-oriented Si NWs, the upper valence subbands are closer to each other and much flatter. From these results, one can anticipate that the [110] direction will be the best orientation for electron transport, while the [110] and [111] directions will be the best orientations for hole transport.

The TB methodology has also been applied to NW heterostructures made of different materials, such as core-shell NWs and quantum disks and tunnel barriers embedded in NWs, where strain and piezoelectric fields may be important. Other problems that have been treated are the effects on the electronic structure due to the presence of impurities, alloy disorder, interface roughness, and surface treatment and the importance of the screening induced by the dielectric mismatch between the NW and the surrounding environment.

Anisotropic Optical Absorption of Semiconductor Nanowires

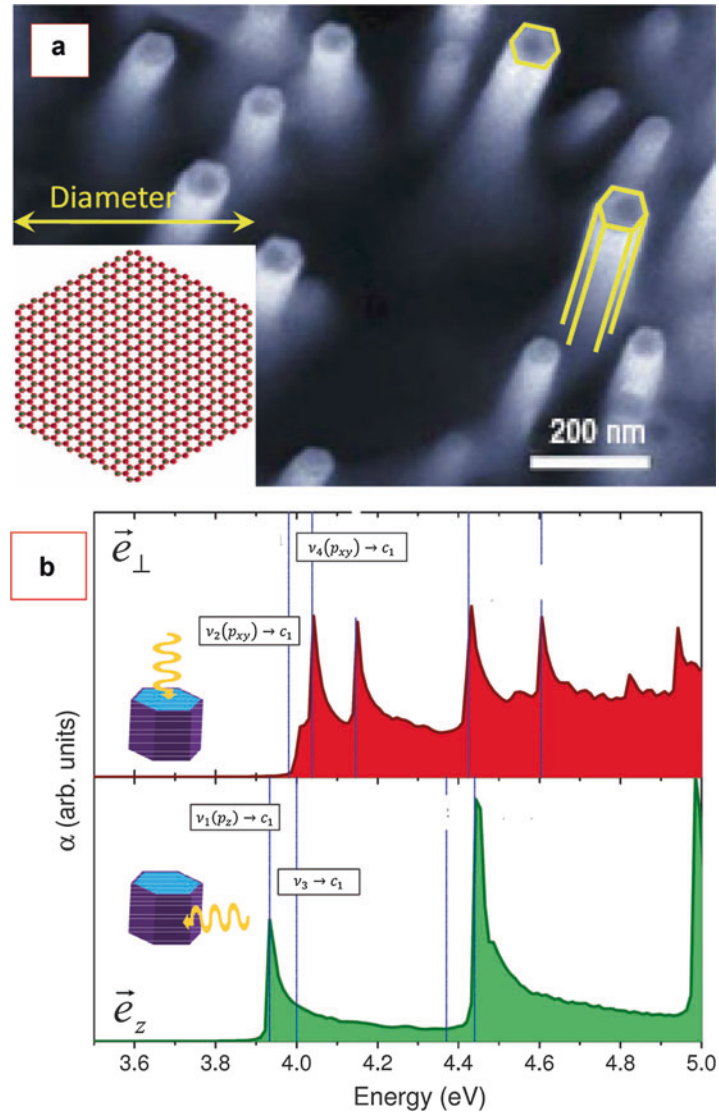
The detailed knowledge of the electronic structure provided by the TB calculations, as illustrated above, gives access to other related properties.



Tight-Binding Simulations of Nanowires, Fig. 1 *Top panel*: Cross sections of [100]- , [110]- , and [111]-oriented Si nanowires with radius $R = 1$ nm. *Middle* and *bottom panels*: Associated conduction and valence subband structures and density of states (DoS) (after [12], with permission)

The most immediate ones are the optical properties that are fully determined by the single-particle states. The easiest to compute is the spectrum of the optical absorption coefficient, α(E).

An example of such calculations is reviewed here. The anisotropic optical absorption of wurtzite GaN NWs is calculated with an orthogonal nearest neighbor two-center sp^3 TB model, including the spin–orbit interaction [11]. A sketch of the NW under study is shown in Fig. 2a. The NWs are oriented along the wurtzite c-axis. For the purposes of modeling, the NW is considered to be infinite and to have a hexagonal cross section. The electronic structure for a nanowire with diameter 3.1 nm (not shown here) exhibits a set of conduction and valence subbands whose edges are denoted as c_1, c_2 and v_1, v_2. The fundamental NW bandgap $E_{c_1} - E_{v_1}$ is located at 3.92 eV. The TB method allows a detailed analysis of the symmetry of the corresponding wavefunctions, which can thus be characterized by the dominant contribution of the atomic-like orbitals (either s or $p_{xy} = (p_x, p_y)$ or p_z) and by the overall symmetry of the envelope function. The states c_i have dominant s character. On the other hand, due to the interplay of confinements and the anisotropy of the band structure of wurtzite GaN, the state v_1 has dominant p_z character, whereas in state v_2, the content of p_{xy} dominates. Moreover, c_1 and v_1 have the same envelope symmetry, which is different from that of v_2. Therefore, it is necessary to distinguish the absorption spectra for light polarized along the c-axis ($\boldsymbol{e}_z$) or polarized in-plane ($\boldsymbol{e}_\perp$). Figure 2 presents the absorption spectrum for both polarizations, corresponding to different possible experimental geometries as indicated in the inset. The characteristic 1D density of states is visible in the whole energy range for both polarizations. The lowest transitions are also indicated. The edge of the $\boldsymbol{e}_\perp$ spectrum is blue shifted with respect to the NW bandgap, since it is due to the excited transitions v_3, $v_4 \rightarrow c_1$ at ∼4.05 eV. The optical transition $v_1 \rightarrow c_1$ is optically inactive due to the dominant p_z character of the state v_1, whereas $v_2 \rightarrow c_1$ is inactive due to the difference in the symmetry of the corresponding envelope functions. On the other hand, the absorption edge of the $\boldsymbol{e}_z$ spectrum coincides with the NW bandgap, since the transition $v_1 \rightarrow c_1$ is now optically active.

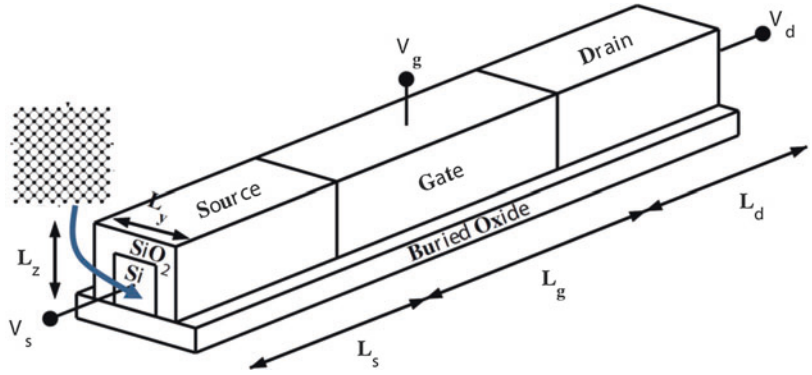


Tight-Binding Simulations of Nanowires, Fig. 2 (**a**) Scanning electron microscopy image of GaN NWs with a sketch of the modeled NW. A ball-and-stick representation of the NW cross section is also displayed as an *inset*. (**b**) Optical absorption spectrum for a GaN NW of diameter 3.1 nm. *Top*, $\boldsymbol{e}_\perp$-polarized light; *bottom*, $\boldsymbol{e}_z$-polarized light. The *dotted lines* indicate relevant electronic transitions (after [11], with permission)

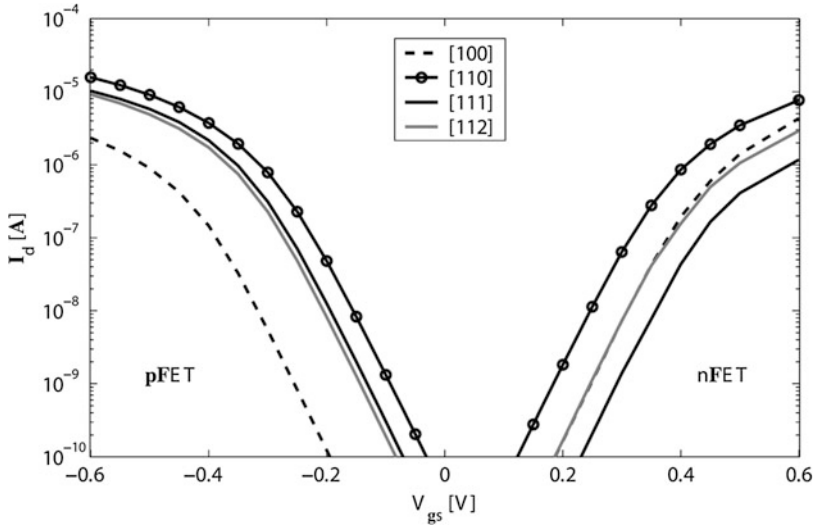
Atomistic Simulations of Nanowire Transistors

The constant request for performance improvement of electronic devices is mainly satisfied by reducing the size of their active components, the transistors. Thirty years of aggressive scaling have pushed the dimensions of the modern devices to the nm range, thus opening the new research field called nanoelectronics. Semiconductor NWs may play an important role in the future of this field, since they can act both as wire connectors and active devices. One typical case is the NW field-effect transistors (FET) that use NWs as active channel. Modeling the current flow in these devices requires to include quantum-mechanical phenomena and atomic granularity. The TB method is well suited to study these problems, and quantitative estimates of the elastic mean free paths, charge mobilities, and localization lengths in semiconductor NWs are routinely calculated [13].

In recent years, there have been reports of state-of-the-art full-band transport simulation tools based on the TB methodology that allow to simulate the operation of realistic devices. As an example, illustrative results obtained with a 3D quantum transport simulator based on the



Tight-Binding Simulations of Nanowires, Fig. 3 Schematic view of a triple-gate nanowire transistor, composed of a square 2.1 nm × 2.1 nm Si NW channel, laterally coated by an oxide layer of thickness 1 nm and electrically coupled to source and drain contacts and to a surrounding gate. The device is deposited on a buried oxide. The source, the drain, and the gate measure $L_s = 9.7$ nm, $L_d = 9.7$ nm, and $L_g = 13$ nm. The oxide layers do not participate to the transport calculation, but they are included to obtain the 3D electrostatic potential (after [14], with permission)



Tight-Binding Simulations of Nanowires, Fig. 4 Full-band transfer characteristics of triple-gate nanowire transistors with a gate length of $L_g = 13$ nm. The *left part* of the plot compares the I_d (V_{gs}) characteristic (with $V_{ds} = -0.4$ V) of *p*-FETs with transport directions along [100], [110], [111], and [112]. In the *right part* of the figure, the I_d (V_{gs}) (with $V_{ds} = +0.4$ V) of *n*-FETs with the same channel orientations are compared (after [14], with permission)

orthogonal nearest neighbor two-center $sp^3d^5s^*$ TB model, neglecting the spin–orbit coupling, will be reviewed here [14]. The simulation code implements open boundary conditions to simulate the action of the gates and is self-consistently coupled to the three-dimensional computation of the electrostatic potential in the device. This simulator allows the analysis of the performance of Si NW FETs with different channel orientations, gate lengths, etc. Figure 3 shows a sketch of the simulated NW FET.

The transfer characteristics of the Si NW FET sketched in Fig. 3 are simulated in the subthreshold regime by applying a variable potential V_{gs} to the gate, with $V_s = 0$ V on the source and $V_{ds} = \pm 0.4$ V on the drain. Figure 4 shows

T

the I_d (V_{gs}) characteristic of *n*-FETs (right) and *p*-FETs (left) with channel orientation along the [100], [110], [111], and [112] crystal axes. The channel orientation [110] offers the highest ON-current in the case of *n*-FETs. For *p*-FETs, the trends are qualitatively the same. The prominence of the [110] channel for both *n*-FETs and *p*-FETs is a consequence of the good compromise between the electron velocity (proportional to the inverse of the effective mass) and the number of available states (proportional to the effective mass) that is reached by the lowest subbands of the [110] wire. The lowest [111] conduction subbands, for example, have too heavy electrons that cannot flow efficiently from source to drain.

Although the above simulations correspond to the ballistic transport regime, it is also possible to incorporate the effects of elastic scattering (due to interface disorder and alloy disorder) as well the inelastic scattering due to the electron–phonon interaction [15], which may play a considerable role in the performance of NW transistors.

Other Applications

The applications of the TB method can be extended to other less studied properties of NWs that also depend directly of the electronic structure. Here are two further representative examples:

1. Thermoelectric properties: By using the electronic structure obtained from an atomistic $sp^3d^5s^*$ TB model with spin–orbit coupling and the associated scattering rates due to surface roughness, impurity, and electron–phonon scattering mechanisms, as input to the linearized Boltzmann equation, it has been possible to calculate the electrical conductivity, the Seebeck coefficient, and the thermoelectric power factor of narrow Si NWs with diameters from 12 nm down to 3 nm [16]. The results reflect an enhancement of the *ZT* figure of merit and may provide guidance in the design of efficient thermoelectric nanodevices.
2. Magnetic properties: The TB method is also a useful method to study the interplay of nanostructure and magnetic properties as manifested, e.g., in ultrathin metallic nanowires. Thus, the electronic and magnetic properties of the rhodium nanowires have been calculated using a spin-polarized *spd* TB Hamiltonian in the unrestricted Hartree–Fock approximation [17]. It has been found that, even though bulk rhodium is a nonmagnetic solid, all the NWs studied are ferromagnetic. The magnetic properties of rhodium nanowires are related not only to their size but also to their atomic structure, with the small diameter center-atom core wires (i.e., the centered pentagonal and hexagonal wires) being more favorable to a larger magnetization.

Since the TB is an electronic structure method, it does not give directly information on lattice-related properties, such as mechanical, thermal, or vibrational properties. These properties are usually accessed from simulations based on molecular dynamics (MD), either expensive ab initio MD or oversimplified empirical classical MD. However, in the past two decades, an intermediate scheme called tight-binding molecular dynamics (TBMD) has been developed. In the TBMD, the ground state total energy is approximated as a sum of two terms, $E_{GS} = U_{REP} + E_{BS}$ [8]. U_{REP} is an ion-ion repulsive potential energy, which in most cases is modeled simply as a sum of short-ranged two-body repulsive interatomic potentials. E_{BS} is the band structure energy, determined by the sum of the energies of the set of occupied single-particle states of the system. The essence of the TBMD method consists in taking advantage of the TB method for the calculation of E_{BS} as a function of the instantaneous configuration of the system. The interatomic forces are finally evaluated in a straightforward way based on the Hellmann–Feynman theorem, and the rest of the procedure is almost identical to the standard MD simulation, but now the quantum-mechanical many-body nature of the interatomic forces is taken into account naturally by calculating the quantum state of the electrons at each MD time step. Since the TBMD scheme uses a minimal basis set for the electronic structure calculations and the Hamiltonian matrix elements are parametrized or tabulated, it allows to study the lattice-related properties of much larger systems than the

ab initio method while maintaining better accuracy than empirical MD. The scheme has been demonstrated to be a powerful method for studying various structural, dynamical, and electronic properties of covalent as well as metallic nanostructures. In the field of NWs, it has been successfully used, e.g., for the calculation of the mechanical properties of wurtzite ZnO unsaturated NWs [18] and the study of the vibrational spectrum of Si NWs [19].

Cross-References

- ▶ Ab Initio DFT Simulations of Nanostructures
- ▶ Computer Modeling and Simulation of Materials
- ▶ Heat Transfer in Semiconductor Nanostructures
- ▶ Nanostructure
- ▶ Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism
- ▶ Nanowires
- ▶ Optical and Electronic Properties

References

1. Martin, R.M.: Electronic Structure – Basic Theory and Practical Methods. Cambridge University Press, Cambridge, UK (2008)
2. Meyyappan, M., Sunkara, M.K.: Inorganic Nanowires: Applications, Properties, and Characterization. CRC Press, Boca Raton (2009)
3. Bloch, F.: Über die Quantenmechanik der Elektronen in Kristallgittern. Z. Phys. **52**, 555 (1928)
4. Slater, J.C., Koster, G.F.: Simplified LCAO method for the periodic potential problem. Phys. Rev. **94**, 1498 (1954)
5. Chadi, D.J.: Spin-orbit splitting in crystalline and compositionally disordered semiconductors. Phys. Rev. B **16**, 790 (1977)
6. Löwdin, P.O.: Quantum theory of cohesive properties of solids. Adv. Phys. **5**, 1 (1956)
7. von Boehm, J., Calais, J.-L.: Completely orthonormalised symmetry-adapted atomic orbitals for solid state calculations. J. Phys. C **17**, 249 (1984)
8. Foulkes, W.M.C., Haydock, R.: Tight-binding models and density-functional theory. Phys. Rev. B **39**, 12520 (1989)
9. Elstner, M., Porezag, D., Jungnickel, G., Elsner, J., Haugk, M., Frauenheim, T., Suhai, S., Seifert, G.: Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. Phys. Rev. B **58**, 7260 (1998)
10. Sarkar, S., Pal, S., Sarkar, P.: Electronic structure and band gap engineering of CdTe semiconductor nanowires. J. Mater. Chem. **22**, 10716 (2012)
11. Molina-Sanchez, A., García-Cristóbal, A.: Anisotropic optical response of GaN and AlN nanowires. J. Phys. Condens. Matter **24**, 295301 (2012)
12. Persson, M.P., Lherbier, A., Niquet, Y.-M., Triozon, F., Roche, S.: Orientational dependence of charge transport in disordered silicon nanowires. Nano Lett. **8**, 4146 (2008)
13. Yamada, Y., Tsuchiya, H., Ogawa, M.: Atomistic modeling of electron-phonon interaction and electron mobility in Si nanowires. J. Appl. Phys. **111**, 063720 (2012)
14. Luisier, M., Schenk, A.: Atomistic simulation of nanowire transistors. J. Comput. Theor. Nanosci. **5**, 1031 (2008)
15. Luisier, M., Klimeck, G.: Atomistic full-band simulations of silicon nanowire transistors: effects of electron-phonon scattering. Phys. Rev. B **80**, 155430 (2009)
16. Neophytou, N., Kosina, H.: Effects of confinement and orientation on the thermoelectric power factor of silicon nanowires. Phys. Rev. B **83**, 245305 (2011)
17. Wang, B., Wang, G., Ren, Y., Sun, H., Chen, X., Zhao, J.: Electronic and magnetic properties of ultrathin rhodium nanowires. J. Phys. Condens. Matter **15**, 2327 (2003)
18. Tuoc, V.N.: The self-consistent charge density functional tight binding study on wurtzite nanowire. Comput. Mater. Sci. **49**, S161 (2010)
19. Zörgiebel, F., Kunstmann, J., Nozaki, D., Cuniberti, G.: Computing Raman and infrared wavenumbers of nanostructures: application to silicon nanowires. J. Raman Spectr. **43**, 1214 (2012) 83 (2004)

TiO_2 Nanotube Arrays: Growth and Application

Karthik Shankar
Department of Electrical and Computer Engineering, W2-083 ECERF, University of Alberta, Edmonton, AB, Canada

Definition

TiO_2 nanotube arrays (TNAs) denote self-organized tubular nanostructures of titanium dioxide grown by the process of electrochemical anodization. When it is titanium metal which is

subjected to electrochemical anodization, the TNAs grow directly in a single step without the assistance of a template. When it is aluminum metal which is subjected to anodization, the formation of TNAs is said to be templated and requires multiple steps including (1) the formation of the nanoporous alumina template, (2) a process to coat the walls of the template with TiO_2 or infiltrate a TiO_2 precursor into the template followed by heating, etc. (3) release of the alumina template by selective chemical etching to form TNAs. TNAs produced by the one-step process shall henceforth be referred to as anodic TiO_2 nanotube arrays and TNAs produced by the multistep process as templated TiO_2 nanotube arrays. In either process, the titanium dioxide nanotubes grow in arrays with their long axis aligned orthogonal to the substrate, due to which they are commonly referred to as vertically oriented TiO_2 nanotube arrays.

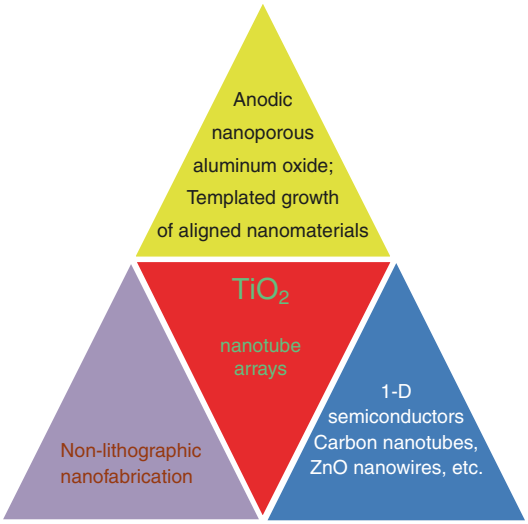
Introduction

The formation of anodic porous titania by using fluoride-based electrochemistry was first reported by Zwilling in 1999 [1]. However, it was Gong et al. [2] in 2001 that first formed nanoporous TiO_2 with a clearly identifiable nanotubular structure. Pioneering contributions to the field of TNAs were made by the research groups of Craig Grimes [3] at Penn State University and Patrik Schmuki [4] at the University of Erlangen. Initially, anodic TiO_2 nanotube arrays could only be grown on titanium foil substrates. In 2005, Mor et al. reported the successful formation of anodic TNAs on Ti-coated glass substrates. Since then, anodic TNAs have been formed on silicon wafers and on transparent conducting oxide (TCO) coated glass substrates. Correspondingly, the first report on templated TNAs emerged in 1996 [5], but more detailed studies commenced in late 2001 [6].

TiO_2 is a highly versatile compound with applications as a pigment for paints, nontoxic biocompatible material for orthopedic implants, gas sensor, germicide, photocatalyst, n-type semiconductor scaffold for sensitized solar cells, optical coating, structural ceramic, electrical circuit varistor, spacer material for magnetic spin-valve systems, and more. TiO_2 is also relatively inexpensive owing to its stability and the plentiful abundance of Ti and O in the earth's crust. TiO_2 occurs in three allotropes: anatase, rutile, and the much less common brookite. The metastable anatase phase has superior electronic properties. The key motivation behind the development of TNAs is to enhance and exploit the versatility inherent in TiO_2 using the specific nanostructured morphology of TNAs. The oriented and aligned one-dimensional TNA architecture offers access to the anatase phase, a high surface area and directed electron percolation pathways for electronic and optoelectronic applications. For other applications of TNAs such as their use in stem-cell differentiators, antibiotic elution and energy transfer-based light harvesting, the specific properties of TiO_2 are less important; these applications derive instead from size selectivity, confinement in nanochannels, and other geometrical attributes of the nanoscale TNA architecture.

The research area of TNAs has witnessed an explosion of interest in the last decade with more than 1,000 research papers on this topic. Apart from the versatility of TiO_2 referenced above, another reason for the intense interest is the location of self-organized vertically oriented TiO_2 nanotube arrays at the intersection of three major trends in nanotechnology research over the last decade as shown in the schematic of Fig. 1. One such trend has been the drive to form one-dimensional structures such as nanorods and nanotubes in semiconductors, inspired by the unique properties of carbon nanotubes. Another trend, which is also one of the driving motivations underlying nanotechnology, is the exploitation of self-assembly and self-organization to form complex structures from the "bottom-up" as an alternative to patterning and fabricating structures from the "top down." The third trend has been the use of ordered anodic nanoporous alumina as a template to form a variety of oriented and aligned nanomaterials.

The use of electrochemical anodization to form nanoporous alumina has been known since the 1950s. However nanoporous alumina is an



TiO_2 Nanotube Arrays: Growth and Application, Fig. 1 TNAs sit at the intersection of three major nanotechnology trends

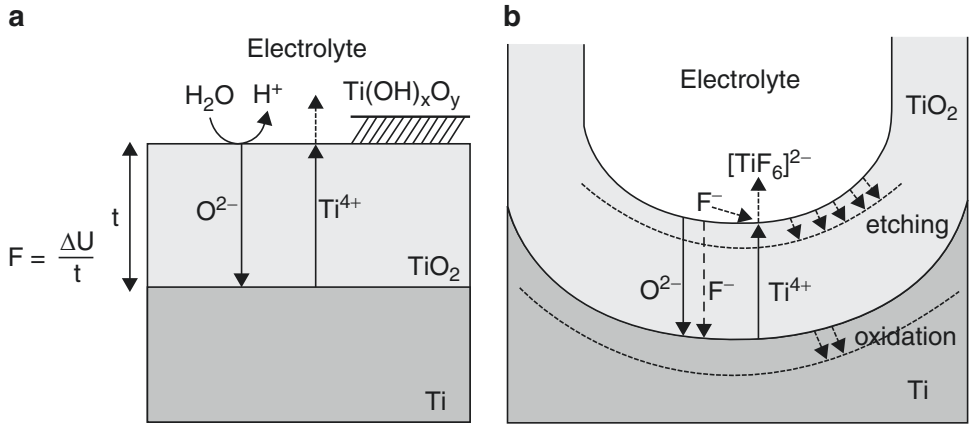
insulator which limits its device applications. Anodic aluminum oxide (AAO) is widely used as a template to form functional nanomaterials. Since crystalline TiO_2 is a semiconductor, anodically formed TiO_2 nanotube arrays can be directly used as a functional material after the induction of crystallinity. A major difference between AAO and TNAs is that a wider range of morphologies are achievable in TNAs. Ordered AAO has been formed in a close-packed hexagonal pore architecture and in square and triangular lattices. In TNAs, the anodization electrolyte has a decisive role in determining the morphology of the resulting structure. By anodization in ethylene glycol-based electrolytes, TNAs may be formed in a close-packed hexagonal architecture similar to AAO. At the same time, nanotube arrays with a distinctive cylindrical cross section may be achieved in other electrolytes. The wall thickness can be varied to change the porosity of the architecture. In comparison with nanochannels in AAO or in track-etched polycarbonate membranes, the nanotubular structure has a significantly higher surface area due to the availability of both the inner and outer surfaces of the nanotube walls.

In the course of the last 10 years 2001–2010, a high degree of control has been achieved over the geometrical dimensions of the anodic TNA architecture by optimizing the anodization process. Potentiostatic anodization is most commonly used either by direct application of a constant anodization voltage or though a linear sweep of the potential to a final constant value. The key parameters involved in the fabrication process are the composition of the anodization electrolyte, the anodization voltage applied, the temperature of the anodization bath, and the anodization duration. For applications involving increased pattern order of the TNAs and for TNA growth on nonnative substrates such as glass, Si, etc., the nature of the Ti film or Ti foil specifically its density, roughness, grain size, and crystallographic texture (if any) are also critical. By tuning the various anodization parameters listed above, control over the geometrical dimensions of TNAs has been demonstrated over remarkably broad length scales [4, 7–9]; the diameter of nanotubes has been varied from 16 to 900 nm, the tube length from 250 to 1 mm, the wall thickness from 5 to 45 nm and the inter-tube spacing from 0 (close packed nanotubes) to several micrometers (widely separated nanotubes). Of these geometrical features, only the diameter and the tube length are independently controllable to date. The wall thickness and inter-tube spacing depend on the diameter and the tube length.

Growth of Anodic TiO_2 Nanotube Arrays

When a Ti foil or a Ti film coated electrode is subjected to anodization in highly conductive electrolytes, the anodization current experiences a steep drop in the first several seconds of the process. The drop in current occurs due to the rapid formation of a passivating oxide layer (called the barrier layer) through which electronic conduction can only occur through tunneling. In the simplest case, the anodization current density J for Fowler-Nordheim tunneling through the oxide barrier layer (BL) may be expressed as:

$$J = C\left(\frac{\Delta U}{t}\right)^2 e^{-\frac{\beta_{FN} t}{\Delta U}} \quad (1)$$



TiO_2 Nanotube Arrays: Growth and Application, Fig. 2 Schematic representation of the Ti anodization (**a**) in absence of fluorides (results in flat layers), and (**b**) in presence of fluorides (results in the tube growth) [4] (Reprinted with permission from Elsevier)

where ΔU is the portion of the anodization voltage which drops across the barrier layer, t is the oxide thickness and C, β_{FN} are constants. Equation 1 shows that the tunneling current reduces exponentially with increasing BL thickness. As the passivating layer increases in thickness, tunneling weakens and the formation of nascent gases at the anode due to electron transfer reactions is suppressed. Visually, this is indicated by the cessation of bubbling activity at the anode. At this point, current flowing through the oxide consists almost entirely of ions. The solid state migration of ions through the oxide is described by the high-field transport model in which the anodization current density J is given by:

$$J = A_e^{\frac{\beta_{HF}\Delta U}{t}} \tag{2}$$

where A and β_{HF} are constants.

Later in the anodization process, the growth process is under mass-transport control as the anodization current is controlled by diffusion of fluoride ions through the tubes to the pore bottoms and results in an inverse dependence of the anodization current on the tube length [4]. The current density is given as:

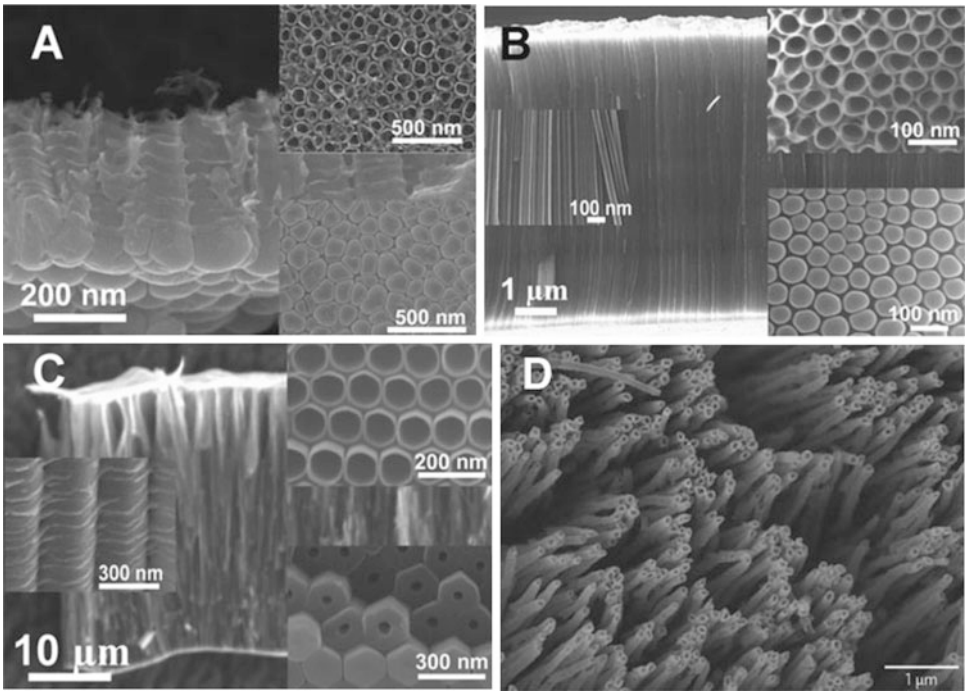
$$J = -pD\frac{\partial c}{\partial x}(0 \le x \le L) \tag{3}$$

where D, c, x, and L are the diffusion coefficient and concentration of the ionic species, thickness of the diffusion layer, and length of the nanotubes, respectively, and p is the ratio of the weight of dissolved oxide to the weight of produced oxide.

The anodic formation of TNAs is a highly nonequilibrium process which occurs at high overpotentials as a result of the interplay between three competing and essential processes (see Fig. 2): Field-assisted oxide dissolution and cation migration, field-assisted oxidation of Ti and chemical etching. It is chemical etching which is responsible for the development of nanotubes as opposed to mere nanopores. In contrast, the formation of nanoporous alumina is controlled mainly by the field-assisted processes. The field-assisted reactions occur on either side of the barrier layer at the bottom of the nanotubes and are responsible for driving the Ti/TiO_2 interface deeper into the Ti foil or Ti film, a process which increases the length of the nanotubes. Chemical etching shortens the length of the nanotubes. The reactions involved in the form ation of TNAs are:
Field assisted migration:

$$Ti + 2H_2O \rightarrow TiO_2 + 4H^+ + 4e^- \tag{4}$$

Field assisted migration:



TiO_2 Nanotube Arrays: Growth and Application, Fig. 3 Scanning electron micrographs showing the effect of anodization electrolyte on the morphology of TNAs: (**a**) Typical morphology obtained in aqueous acidic fluoride or HF electrolytes, (**b**) glycerol/fluoride electrolytes, (**c**) ethylene glycol/fluoride electrolytes, and (**d**) DMSO/fluoride electrolytes (Images (**a**) through (**c**) were reprinted from [4] with permission from Elsevier. Image (**d**) was reprinted from [9] with permission from Nature)

$$Ti^4 + 6F^- \rightarrow [TiF_6]^{2-} \quad (5)$$

Field assisted migration:

$$TiO_2 + 6F^- + 4H^+ \rightarrow [TiF_6]^{2-} + 2H_2O \quad (6)$$

Chemical dissolution:

$$TiO_2 + 6HF \rightarrow [TiF_6]^{2-} + 2H_2O + 2H^+ \quad (7)$$

The as-anodized nanotube arrays are amorphous and non-insulating. Manifestation of the semiconductor properties of TiO_2 requires induction of crystallinity, which is typically accomplished by annealing the nanotube arrays in air or in an oxygen-rih ambient at elevated temperatures of 350–580 °C. For annealing temperatures of up to 500 °C, the nanotubes are transformed into polycrystalline anatase alone; at higher temperature, increasing amounts of rutile are present as well. Annealing at temperatures higher than 580 °C results in the destruction of the nanotubular architecture.

The evolution of different fabrication recipes for anodic TNAs has occurred in four generations to date [3, 7, 8]. The first three generations were defined by the maximum length of nanotubes that was achieved and they all used fluoride ions to effect the necessary chemical etching. The fourth generation was defined by its nonuse of fluoride ions. First-generation nanotubes were up to 500 nm long and were grown by anodization in strongly acidic aqueous electrolytes containing 0.1–0.5 wt% HF (pH $\leq$1). Other additives such as acetic acid were sometimes added to improve the mechanical strength and quality of the nanotubes. The tube length in first-generation TNAs was limited by the strong chemical etching by HF. Typical morphology obtained in first-generation nanotubes is shown in Fig. 3a. The steady-state TNA film thickness (corresponding

to the tube length) occurred when the rate of movement of the Ti/TiO_2 interface deeper into the metal was exactly balanced by the chemical etching at the mouths of the nanotubes. In the second generation, fluoride ion-bearing aqueous electrolytes were still used but the rate of chemical etching was tightly controlled by adjusting the pH by adding buffers to the anodization electrolyte. Typically, trisodium citrate and sodium dihydrogen sulfate were used as buffering agents in the electrolyte and salts such as ammonium fluoride, sodium fluoride, and potassium fluoride were used as sources for fluoride ions. As the pH was varied from 1 upward toward 5, the nanotubes grew correspondingly longer due to the reduction in chemical etching. The maximum nanotube length in aqueous electrolytes of 6.6 μm was achieved at a pH of 5. The maximum growth rate was ~0.25 μm/h. When the pH was increased beyond 5, nanotube formation did not occur. Nanotube formation in both the first and second generations occurred in aqueous electrolytes in a narrow window of anodization voltages ranging from 10 to 27 V, which restricted the inner diameters of the nanotubes to the range 22–100 nm. In the third generation (Fig. 3b–d), fluoride-bearing organic electrolytes were used to vastly expand the morphological parameter space achievable. Nanotubes as long as 1 mm were grown by complete conversion of the underling Ti foil to TNAs. Thus, the nanotube growth is limited in length only by the thickness of the underlying substrate and in principle can be extended even longer by choosing thicker substrates. The inner diameter of the nanotubes could be varied from 17 to 900 nm. When chemical etching is minimized, the slow solid state ionic transport of reactants through the barrier layer is the rate-limiting step preventing further increase in tube length. From Eq. 7, it is clear that the high-field ionic current is increased by a thinner barrier layer. This key insight motivated the use of organic electrolytes in the third generation which provided more reducing conditions than aqueous electrolytes and ensured the formation of a thinner barrier layer. The thinner barrier layer allows growth rates as high as 15 μm/h to be achieved. Third generation electrolytes consist of a fluoride-ion bearing salt dissolved in an amphiprotic organic solvent of moderate to high viscosity such as formamide, *N*-methyl formamide, ethylene glycol, glycerol, etc. About 1–5 % of water is typically added to form organic electrolytes in order to provide oxygen-bearing hydroxyl ions during anodization for oxide formation. Nanotubes do not grow well when low-viscosity protic solvents such as methanol or ethanol are used as the base solvent to form the organic electrolyte (although mixtures with water and HF have been found to result in anodic TNAs). The same is true of aprotic solvents such as propylene carbonate and acetonitrile. An exception is dimethylsulfoxide, an aprotic solvent which has been used to form high-quality TNAs. By detaching the nanotube arrays from the foil substrate, TNA membranes similar to nanoporous alumina membranes may be produced. The nanotube arrays need to be at least 10 μm long for the detached film to be mechanically robust enough to a form free-standing membrane. Some of the methods used to detach the TNAs from the native Ti foil substrate include the use of natural stresses in the thick TNA films through ultrasonication, selective etching of the Ti substrate by immersion in a solution of CH_3OH/Br_2, and the use of induced stresses by double-step anodization. The membranes have a strong tendency to curl up due to capillary forces. The membranes can be kept planar by processing with low surface tension solvents or by supercritical CO_2 drying. Fourth-generation nanotubes were formed in fluoride-free electrolytes (Table 1).

Growth of Templated TiO_2 Nanotube Arrays

Both hard templates such as anodic nanoporous aluminum oxide (AAO) and soft templates such as polycarbonate membrane filters have been used to form TNAs. Soft templates have two major disadvantages: (1) an elevated temperature (~450–500 °C) thermal annealing step is frequently to induce or improve the crystallinity of the templated TNAs. However, most polymers are not able to withstand temperatures above 200 °C.

TiO_2 Nanotube Arrays: Growth and Application, Table 1 Common recipes for TNAs and the resulting morphological parameters

Electrolyte	Anodic potential (V)	Inner diameter (nm)	Length (μm)
Aqueous 0.5 % HF	10	22	0.2
Aqueous 0.5 % HF	10–23 (nanocones)		0.4
Aqueous buffered 0.1 M KF (pH ~ 5)	25	110	6.6
0.1 M NH_4F in 1:1 DMF:H_2O	15 porous		
0.15 M NH_4F + 5 % H_2O in formamide	10	12	3.6
0.15 M NH_4F + 5 % H_2O in formamide	15	29	5.4
0.15 M NH_4F + 5 % H_2O in formamide	20	70	14.4
0.22 M NH_4F + 5 % H_2O in formamide	35	180	30
0.27 M NH_4F + 5 % H_2O in formamide	20	65	19.6
0.3 % NH_4F in EG + 1 % H_2O	30	45	9
0.3 % NH_4F in EG + 1 % H_2O	40	70	12
0.3 % NH_4F in EG + 1 % H_2O	65	135	30

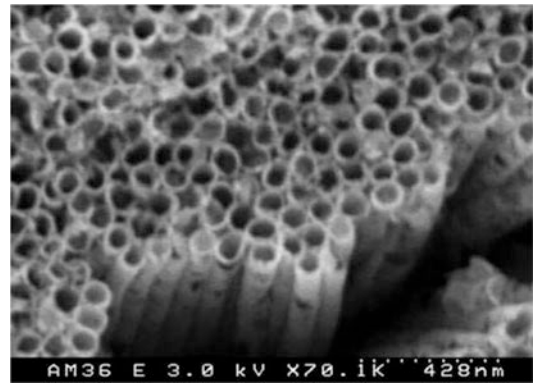
Reprinted with permission from [7]

Consequently, the polymer template pyrolyzes before the phase transition in TiO_2 thus eliminating the mechanical support provided by the template and/or result in the undesirable incorporation of carbonaceous materials in the TNAs; (2) the pore densities are smaller and the nanochannels in polymeric membrane filters are frequently tilted from the surface normal and sometimes intersect within the membrane [10]. Therefore, discussion shall be restricted to AAO-templated TNAs.

AAO templates are formed by electrochemical anodization of aluminum metal in three common electrolytes based on sulfuric acid, oxalic acid, and phosphoric acid. The choice of electrolyte is determined by the desired pore size. The Al metal maybe a native substrate such as aluminum foil or could be a film deposited by vacuum evaporation or sputtering on nonnative substrates such as glass and silicon. High-ordered templates approaching ideal hexagonal, square, or triangular lattices may be prepared by techniques such as two-step anodization and various imprinting methods to guide pore formation [11].

Methods to synthesize TNAs in AAO templates can be classified into liquid phase and gas-phase methods. Sol-gel hydrolysis, electrochemical hydrolysis and electrophoretic deposition of nanoparticles are examples of liquid phase methods. Atomic layer deposition (ALD) and chemical vapor deposition (CVD) are gas-phase methods. The growth of nanotubes as opposed to nanorods in nanoporous alumina requires preferential formation of TiO_2 on the walls of the template. This is accomplished by preferential wetting of the template walls by TiO_2 precursors in sol-gel methods, by preferential reduction or oxidation at the walls in electrochemical methods, by preferential concentration of the electric field lines at the pore walls in electrophoretic deposition, and by the non-line-of-sight conformal deposition on the walls in gas-phase methods. ALD is a self-limiting film growth process in which alternate saturative surface reactions are made to occur by pulsing the vapors of two or more precursors in the reaction chamber alternately [12]. ALD is capable of producing highly conformal coatings even on substrates with complex features and offers exceptional control in tuning the wall thickness of the nanotubes. Consequently, ALD is the most widely used method to form the templated growth of TNAs. Typically, a flow reactor configuration is employed and the precursors used to grow TiO_2 films are a titanium alkoxide, whose generic formula is $Ti(OR)_4$ and water (H_2O). The deposition temperatures used in ALD of TiO_2 films for templated TNA growth vary widely, from 100 °C to 325 °C. The exposure times of the template to each precursor pulse are typically

limited to a few seconds and the chamber is evacuated prior to the introduction of the next pulse. In addition, in some recipes, an inert gas is used to purge the chamber before the next pulse. The main disadvantages of ALD are the high cost and complexity of ALD equipment and the extremely slow film growth rate. A brief overview of other processes used to form templated TNAs is now provided. The operating principle of the sol-gel method is the use of capillary forces to drive a TiO_2 sol into the nanochannels of the template which transforms into TNAs upon thermal annealing. A sol-gel process based on TiF_4 hydrolysis with hydroxypropylcellulose added as a gelling agent, has been used to form 200 nm diameter TNAs in an AAO template. However, the process was highly nonuniform over large areas with irregular deposition of the TiO_2 and poor penetration of the sol particles into the template nanochannels. Much better results were obtained using the pressure-assisted infiltration of liquid titanium isopropoxide (TiO_2 precursor) into the AAO nanochannels followed by thermal decomposition method of the precursor at 500 °C in an oxygen stream to yield TiO_2 nanotube arrays [6]. In the electrophoretic sol-gel method, an electric field is used to provide significantly larger forces than mere capillary action to drive the charged particles created by sol-gel processing into the nanochannels [13]. A triple step templating process has been used to form TiO_2 by the anodic oxidative hydrolysis of $TiCl_3$ [5]. In this process, the AAO template morphology was transferred to a polymer mold which contained a negative replica of the template pattern. Gold nanorods were formed by electroless deposition on to the polymer mold. Then the gold nanorods served as a conducting template to form TNAs (Fig. 4).

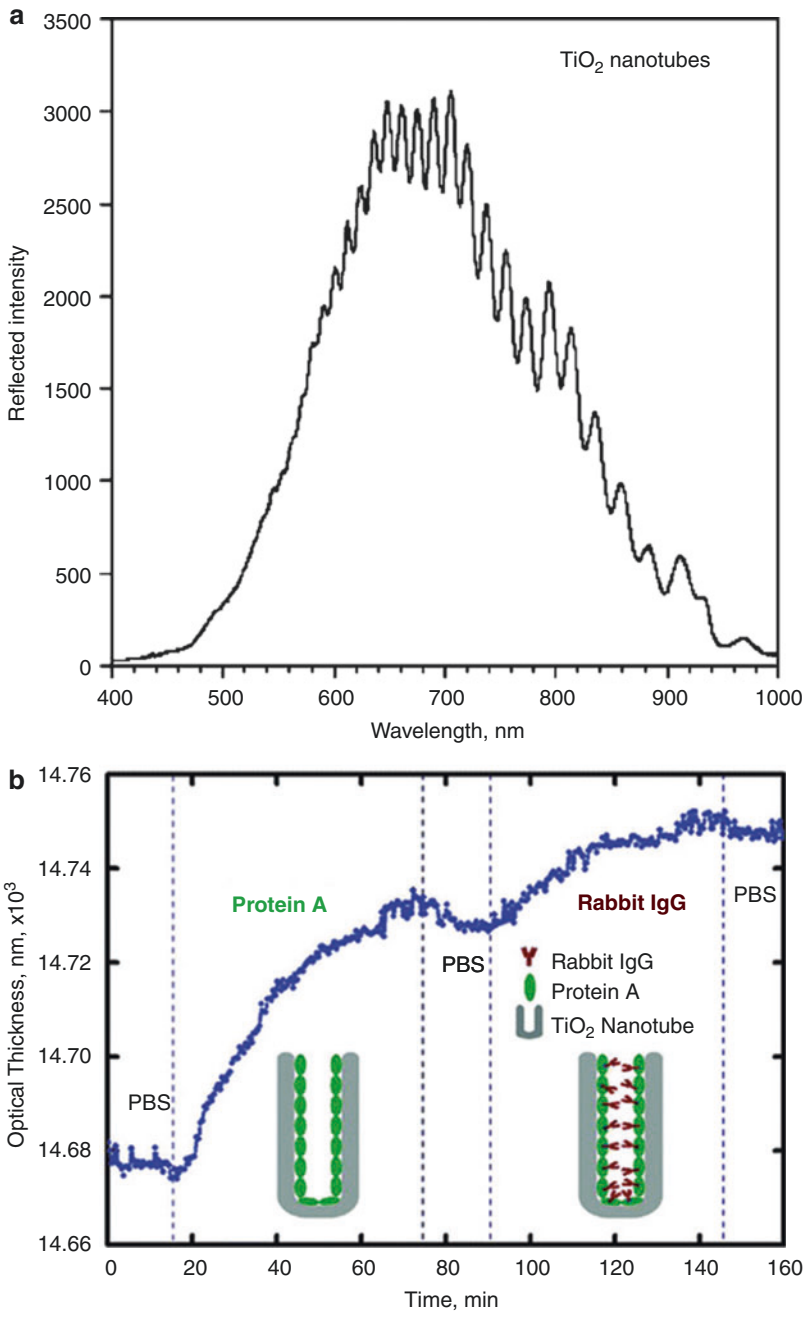


TiO_2 Nanotube Arrays: Growth and Application, Fig. 4 Scanning electron micrograph of templated TNAs formed by sequential infiltration and oxidation of titanium isopropoxide, followed by removal of the alumina matrix in which the tubes were formed (Reprinted with permission from [6])

Biological Applications of TiO_2 Nanotube Arrays

Biosensing Applications

Biosensors are devices that combine a biological component (a recognition layer) and a physicochemical detector component (a transducer). The majority of biosensors are targeted to proteins but biosensors for cells and for metabolically important small molecules are also being investigated. Molecular labels have enabled the construction of biosensors of high sensitivity and specificity. However, labels often disrupt the accurate measurement of kinetic constants, particularly binding equilibria, and problems, such as antibody cross reaction or impure solutions, can occur. Due to these drawbacks, label-free biosensors are highly desired for a variety of biotechnology applications such as disease diagnostics, drug discovery, environmental monitoring, and food safety. For label-free optical biosensors, the transduction methods can be divided into two categories: optical interferometric and surface plasmon methods [14]. In interferometric sensors, sensing is achieved by extracting the refractive index of the porous matrix from the Fabry-Perot interference fringes in the reflectance spectrum. Some of the prerequisites for an effective optical interferometric sensing material are a large internal surface area, ease of fabrication, and a chemically modifiable surface. TiO_2 nanotube arrays satisfy the above requirements but in addition possess a high refractive index ($n = 2.5$ for TNAs) and superior chemical stability over a wide pH range in comparison with other candidate materials such as porous



TiO_2 Nanotube Arrays: Growth and Application, Fig. 5 TNA-based interferometric biosensors: (**a**) representative reflectivity spectrum of anodic TNAs showing the characteristic interference fringes and (**b**) time-dependent optical thickness measurements showing sequential binding of Protein A and rabbit IgG within TiO_2 nanotubes (Reprinted with permission from [14])

silicon and nanoporous alumina. The larger index contrast between the porous host and the aqueous matrix in which the biomolecular binding measurement is carried out is expected to provide greater contrast in the interferometric spectrum, leading to lower noise and higher sensitivity [14]. Mun et al. [14] demonstrated the highly specific label-free interferometric sensing of rabbit immunoglobulin G (IgG) using protein A (derived from *Staphylococcus aureus* bacteria) immobilized on the walls on the nanotubes as the capture probe (Fig. 5).

T

Orthopedic Implant Applications

Because of their mechanical strength, chemical inertness, and biocompatibility, TiO_2 and its alloys (particularly Ti-6Al-4 V) have been used extensively in orthopedic and dental implants since 1970 [15]. Osteogenesis is the process of formation of new bone tissue by cells called osteoblasts. Ideal implants are those whose surfaces will promote in vitro and in vivo osteogenesis. However, the influence of surface micro- and nanotopography on osteogenesis is still not well understood. High surface area porous materials are used in implants to mimic bone, which is a naturally occurring porous ceramic material composed of nano-sized organic and mineral phases that form a large macrostructure. Proteins in extracellular bone matrix and calcium phosphate, important constituents of the bone matrix, are nanostructured but the porosities in human bone are predominantly in the range 1–100 μm [15]. Spark anodization is a process which leads to the formation of a disordered oxide structure with porosities in the same regime as bone and is therefore commonly used to increase the biocompatibility of titanium and its alloys. However, recent studies using nanoporous alumina and nanotubular titania have revealed that their much smaller pores (20–200 nm) allow bone ingrowth. TNAs, in particular, have been demonstrated to be favorable templates for bone cell growth and differentiation, and also shown to significantly enhance osteoblast activity [8, 15]. The nanoporous architecture of TNAs provides a framework in which osteoblasts produce new bone. The anatase phase of TiO_2 is much more efficient for nucleation and growth of bone than rutile. Another advantage of TNAs is their strong adherence to the underlying Ti substrate. In cell culture experiments, increased adhesion to TNAs of cells which produce and maintain cartilage (chondrocytes) was observed and in animal trials, chronic inflammation was found to be absent in TNAs implanted below the skin in rats. Furthermore, the calcium and phosphorous concentrations were 50 % higher on these surfaces, suggesting that deposition of the bone matrix was accelerated on TNAs [8, 15]. SEM images have been used to show that the microspike extensions (filopodia) of migrating osteoblast cells go into the nanotube pores and in general progagate much faster on TNA surfaces. It has been speculated that the nanotube pore and the gap between the nanotubes serve as a useful pathway for continuous supply of ions, nutrients, and proteins, which is likely to positively contribute to the health of the growing cells [8].

Drug Eluting Coatings for Medical Implants

Bacterial infection and acute inflammation at the site of the implant are the most common postsurgical complications of bone implants, dental implants, and vascular stents. To reduce the chance of complications and to prevent early stage implant failure, antibiotic therapy is indicated 6–8 weeks after surgery [15]. Systemic administration of drugs such as antibiotics to avoid or treat the infection is often ineffective at reaching infected tissues near the implant and in addition, can cause toxicity and other adverse side effects. On the other hand, localized drug delivery from the site of an implant has more efficient therapeutic effects and fewer side effects. In general, an ideal drug eluting coating must have the ability to incorporate a drug, preserve it, and deliver it gradually over the time to a specific target site [16]. TiO_2 nanotube arrays are being actively researched for localized drug delivery for implants because they combine the appropriate biointegration and biocompatibility with the ability to controllably release drugs (a process known as elution) confined in their channels to their surroundings. Drug loading is performed through capillary action by either immersing the templates in the concentrated drug solution or dropping the solution slowly on the template surfaces. In addition, simulated body fluid (SBF)-assisted physical adsorption of antibiotics and anti-inflammatory agents was found to be a superior method of drug loading which also resulted in improved prolonged release of both types of drugs [16]. The elution of gentamicin (a powerful aminoglycoside antibiotic used against gram-negative bacteria) from titania nanotubes has been studied [8]. In TNAs of 80 nm pore diameter

and 400 nm length loaded with gentamicin (70–85 % loading efficiency), drug release kinetics were found to be dependent on initial loading; and there was a sustained release, for 45, 90, and 150 min for loadings of 200, 400, and 600 μg, respectively. In addition, it was found that bacterial adhesion on the surface of gentamicin-loaded titania nanotubes was reduced significantly (in comparison with titanium surface and unloaded nanotubes), whereas normal osteoblast adhesion and proliferation is retained [8, 15]. It has also been shown that TiO_2 nanotubes can control small molecule delivery in the order of weeks, and larger molecules in the order of months. The total drug loading as well as elution was mostly affected by nanotube length, and maximum small molecule elution was reached at ~2 weeks. TNA-based drug eluting coatings are superior to the conventional strategy of drug elution through a drug-loaded polymer coating, because even though the polymer coating is effective in delivering a drug for long periods, it has been known to induce an inflammatory response due to eventual degradation of the polymer. Such inflammation is implicated in late-stage implant failure.

TiO_2 Nanotube Arrays for Stem-Cell Differentiation

Recent studies on cell interactions with TNAs have demonstrated that stem cell fate is dictated by the diameter of the nanotubes. Using amorphous (unannealed) TNAs, Schmuki et al. found the adhesion, proliferation, migration, and differentiation of rat bone marrow mesenchymal stem cells to be highest on nanotubes 15 nm in diameter and dramatically decreased on 70- and 100-nm nanotubes. In contrast, Jin et al. found using crystalline (annealed) TNAs and human mesenchymal stem cells that the optimum length scale for cell vitality and differentiation was 100 nm. More recent studies seem to indicate that nanotube size plays a more important role than crystallinity and nanotube chemical composition although more research is necessary to clarify the nature of topographical cues for cell behavior in this emerging field.

Applications of TiO_2 Nanotube Arrays in Light Harvesting Devices

One of the factors limiting maximum achievable efficiencies in conventional light harvesting devices including photovoltaic cells, photoelectrochemical cells, and photocatalytic films is the ever-present trade-off between light absorption and charge generation on the one hand, and charge separation and carrier collection on the other. This trade-off is best illustrated by considering the case of a planar p-n junction solar cell where only electron-hole pairs generated close to the depletion region are swept apart by the inbuilt electric field and efficiently collected. On the one hand, the light absorbing semiconductor layer needs to be thick enough to absorb a majority of the incident photons and on the other hand, the large thickness means that a portion of the minority charge carriers generated away from the junction are lost to recombination as they diffuse toward the junction. The vertically oriented TiO_2 nanotube array architecture overcomes this trade-off by orthogonalizing the processes of light absorption/charge generation and charge separation/carrier collection so that the trade-off is no longer as severe. The process of charge separation now occurs in the radial direction while light absorption occurs in the longitudinal direction along the axis of the nanotubes. To facilitate charge separation, the diameter of the nanotubes can be made smaller or the nanotube walls may be made thinner while simultaneously making the nanotubes longer to maximize light absorption. The porosity of the ordered structure allows the incident photons to be more effectively absorbed, due to scattering effects, than on a flat electrode. For these reasons, the vertically oriented nanotube array architecture has been posited as nearly ideal for light harvesting applications.

TNAs for the Photocatalytic Reduction of CO_2 to Methane

The free energy for overall methane formation through the chemical reaction is about 801 kJ/mol, a process requiring eight photons:

$$CO_2 + 2H_2O(g) \rightarrow CH_4 + 2O_2 \quad (8)$$

Hydrogen formation from water involves a free energy change (ΔG^0) of 237 kJ/mol and an enthalpy change (ΔH^0) of 285 kJ/mol; the corresponding values for CO formation from CO_2 are 257 and 283 kJ/mol at 25 °C (1 atm.). Hence, the minimum energy required for water and CO_2 splitting processes are, respectively, 1.229 and 1.33 eV (per photon). In theory, the bandgap of a photocatalyst used for co-splitting of CO_2 and water should be at least 1.33 eV, which corresponds to absorption of solar photons of wavelengths below about 930 nm. Considering the energy loss associated with entropy change (about 87 J/mol · K) and other losses involved in the CO_2 splitting process (forming CO and O_2), a bandgap between 2 and 2.4 eV is optimal [17]. Using nitrogen-doped TiO_2 nanotube arrays, with a wall thickness low enough to facilitate effective carrier transfer to the adsorbing species, surface-loaded with nano-sized islands of platinum and copper cocatalysts, efficient sunlight-driven conversion of CO_2 into CH_4 and other hydrocarbons was achieved, with hydrocarbon production rates as high as 160 μL/(g h).

The outstanding nature of TNAs as a platform to perform the photocatalytic reduction of CO_2 is shown by the fact that the above rate of CO_2 to hydrocarbon production obtained under outdoor sunlight is at least 20 times higher than previous published reports [8, 17].

TNAs for Oxidative Photochemistry

Due to the large oxidizing power of photogenerated valence band holes in titania, TiO_2 nanotube arrays are excellent catalysts for oxidative photochemistry. OH radicals and superoxide (O_{2-}) anions produced by photogenerated charge carriers in TiO_2 can oxidize essentially all organic molecules present in the solution into carbon dioxide and water [4]. It has been shown that crystalline TiO_2 nanotubes show considerably higher efficiency for the decomposition of organic small molecules than a compacted Degussa P25 layer (20–30 nm diameter nanopowder composed of anatase and rutile) under comparable conditions [4]. The TNA architecture is also exceptionally stable with no degradation observed in the samples with exposure to UV/visible light over a course of a several months. For sunlight-driven water photolysis, the titania nanotube array architecture results in a large effective surface area in close proximity with the electrolyte thus enabling diffusive transport of photogenerated holes to oxidizable species in the electrolyte. In the TNA architecture, most minority carriers generated within a "retrieval" length from the material surface, i.e., a distance from the surface equal to the sum of the depletion layer width and the diffusion length, and such carriers escape recombination and reach the electrolyte redox species. Consequently, bulk recombination is greatly reduced and the quantum yield (incident photon-to-electron current efficiency) is close to unity. Under visible light illumination, the efficiency of TNAs is limited by the large bandgap of TiO_2 (3.0–3.2 eV), which renders the TNAs sensitive only to ultraviolet photons. A photoconversion efficiency for photoelectrolysis of 12.25 % was obtained under UV (320–400 nm) illumination using second-generation nanotubes 6 μm long annealed at 580 °C. High quantum yields >80 % were obtained indicating that the incident UV photons were effectively utilized by the nanotube arrays for charge carrier generation. The corresponding hydrogen evolution rate was 76 mL/h W. The water-splitting reaction was confirmed by the 2:1 ratio of evolved hydrogen to oxygen. Using third-generation nanotubes ~200 nm in diameter and 30 μm long formed by anodization at 35 V in a formamide-based electrolyte, the water-splitting efficiency improved to 16.25 % [7].

TiO_2 Nanotube Arrays as Scaffolds in Excitonic Solar Cells

In excitonic solar cells, photons of energy larger than the absorption threshold of the excitonic absorber do not create free electron-hole pairs as in a conventional inorganic semiconductor. Instead excitons (electron-hole pairs bound by Coulombic attraction) are generated. Unless these excitons are dissociated at a suitable interface, they are lost to recombination. Therefore, the trade-off in excitonic solar cells is between light

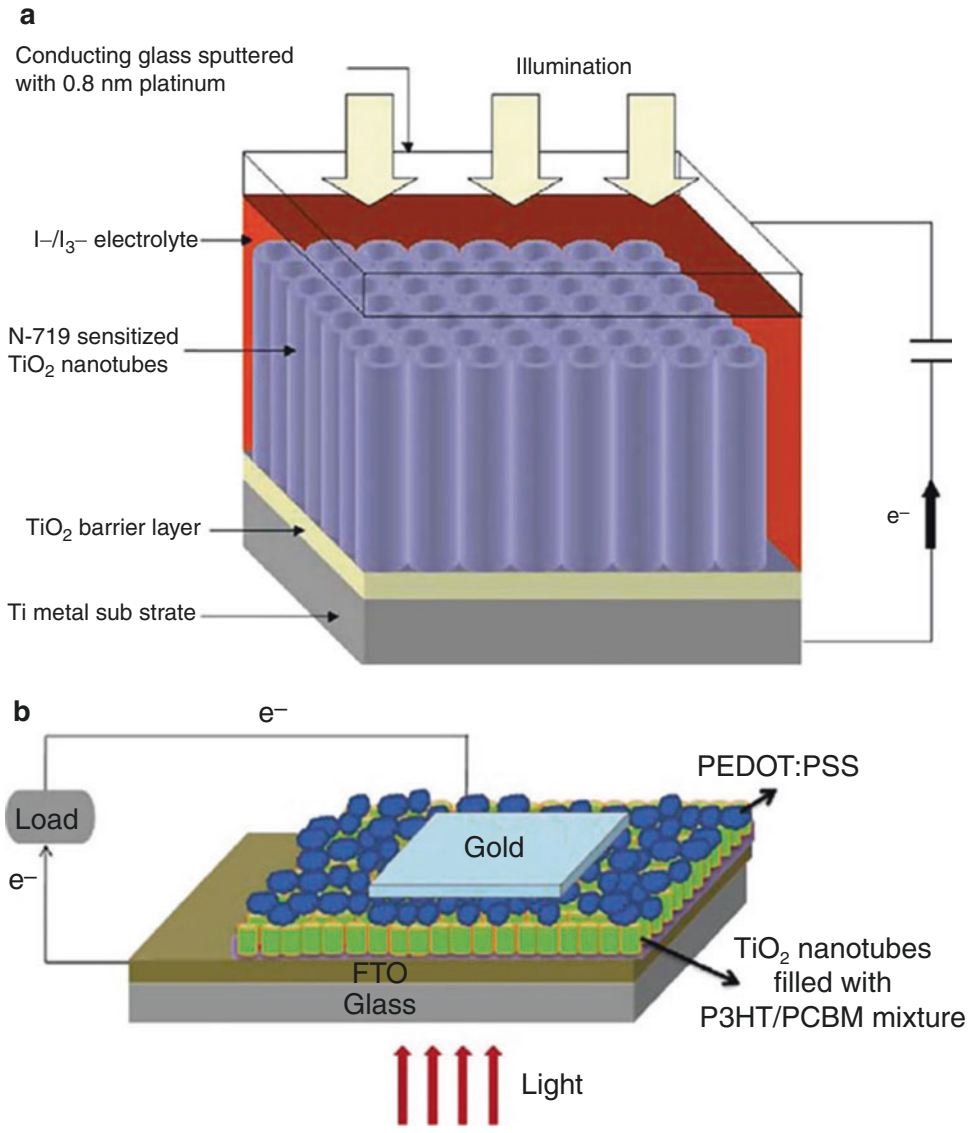
absorption and exciton diffusion. As alluded to before, vertically oriented TNAs orthogonalize the processes of exciton generation and exciton diffusion overcoming the trade-off. Also, TNAs provide electron percolation pathways for the directed (nonrandom) movement of electrons in the structure. TNAs provide flexibility in device design since the excitonic absorbers may be coated on the nanotube walls thus exploiting the large internal surface area of the TNA architecture or the absorbers may be volumetrically filled inside the hollow cylindrical spaces of the nanotubes. Due to these reasons, TNAs are very attractive materials for use as n-type semiconducting scaffolds for excitonic solar cells. TNAs are used in two different kinds of excitonic photovoltaic devices: dye sensitized solar cells, where the absorber is a broadly absorbing dye molecule anchored to the surface of the TNAs and ordered bulk heterojunction solar cells, where the absorber is filled inside the nanotubes. Templated TNAs formed by a sol-gel method were found to have very large roughness factors which enabled them to adsorb >500 nmol cm^{-2} of N-719. Using 17.6 μm long transparent anodic TNAs grown on conductive glass substrates, Varghese et al. obtained N-719 sensitized solar cells with an efficiency of 6.9 % under AM 1.5 1 sun illumination and quantum yields higher than 65 % in the spectral range 400–670 nm [9]. Using a free-standing membrane consisting of 63 μm long TNAs transferred to a FTO:glass substrate, dye-sensitized solar cells subjected to 1 sun AM.15 illumination demonstrated an open circuit photovoltage of 0.77 V, a short-circuit photocurrent of 18.5 mA cm^{-2} and a fill factor of 0.64, resulting in an overall efficiency of 9.1 % [18]. There is also increasing research interest in sensitized solar cells where the sensitizers are quantum dots of CdS or CdSe or CdTe instead of molecular dyes. Quantum yields as high as 55 % have been demonstrated in quantum dot-sensitized TNA solar cells [19]. An ordered double heterojunction solar cell was constructed by infiltrating a blend of regioregular poly (3-hexylthiophene) (RR-P3HT) and a methanofullerene (PCBM) into transparent anodic TNAs on FTO coated glass. The configuration used is termed a double heterojunction because of the availability of two interfaces for charge separation, namely, the TiO_2/P3HT and PCBM/P3HT interfaces. The resulting solar cell had an overall conversion efficiency under 1 sun AM 1.5 illumination of 4.1 % [8] (Fig. 6).

Chemical Sensing Applications of TiO_2 Nanotube Arrays

Chemical sensors often suffer from unwanted contamination, or poisoning which limits the sensor's operational lifetime. Volatile organic compounds are a frequent source of sensor contamination. In this respect, an important advantage of TNA-based chemical sensors is their self-cleaning ability in the presence of ultraviolet radiation [8]. The strong oxidizing power of photogenerated holes in crystalline TiO_2 oxidizes organic contaminants and results in self-cleaning.

TNA-Based Ultrasensitive H_2 Gas Sensors

One of the first applications of TNAs was in hydrogen gas sensing by measuring a change in the electrical resistance of the nanotubes upon exposure to hydrogen. A metal-semiconductor-metal (MSM) structure was employed for resistive H_2 gas sensing wherein Ti metal underlying the nanotube arrays was used as the ohmic contact and a high work function metal (mainly platinum) was used as the Schottky contact. Initial work used first-generation nanotubes at high operating temperatures (e.g., 290 °C) to obtain a three-order change in resistivity when exposed to a few ppm of hydrogen. By sputtering a thin overlayer of Pd on the nanotubes to enable facile dissociation of molecular hydrogen into atomic form, the operating temperature of the resistive gas sensors were brought down to room temperature. By using second-generation nanotubes anodized in aqueous electrolytes at pH 4.5 in combination with Pd sputtering, an unprecedented high sensitivity of 8.7 orders of magnitude was achieved. Room temperature operation coupled with the high sensitivities ushered in applications such as transcutaneous hydrogen sensors. The quantification of



TiO_2 Nanotube Arrays: Growth and Application, Fig. 6 Schematics of TNA-based solar cells (**a**) Liquid junction dye-sensitized solar cells and (**b**) ordered bulk heterojunction solar cells (Reprinted with permission from [8])

ppm-ppb hydrogen gas concentrations escaping from blood through the skin has medical relevance as an indicator of lactose intolerance, fructose malabsorption, microbial activity, bacterial growth, fibromyalgia, diabetic gastroparesis, and neonatal necrotizing enterocolitis. The prevailing explanation for the remarkable change in resistivity is that chemisorption of hydrogen atoms on the nanotube surface results in partial charge transfer to the TiO_2 resulting in an electron accumulation layer on the nanotube surface which drastically improves the electrical conductance of the MSM structure. In addition, hydrogen adsorption reduces the work functions of Pd and Pt. Consequently, the height of the potential barriers at the TiO_2/Pt and TiO_2/Pd interfaces are lowered, thus again increasing the conductance of the MSM structure.

TNA-Based H_2O_2 Sensors

TNAs on which horseradish peroxidise (HRP) and thionine were co-immobilized have been used as electrochemical sensors to detect and quantify hydrogen peroxide [8]. Due to the high electrocatalytic activity of the immobilized HRP for thionine oxidation in the presence of H_2O_2, a large increase of reduction current occurs for the Th/HRP/TiO_2 electrode due to peroxide. The detection limit for H_2O_2 was estimated to be as low as 1.2 μM.

TNA-Based Amine Sensors

Amines are one of several chemical pollutants released in oil and gas spills which contaminate fresh and seawater. In the light of the fact that many amines are proven or suspected to be carcinogenic and have been implicated in inducing bladder cancer, environmental monitoring of amine levels assumes significance. In this regard, TNAs coated with an electrostatically adsorbed layer of ruthenium tris(bipyridinium) ions [$Ru(bpy)_3^{2+}$] have been used in an electrochemiluminescence sensor to detect and quantify the tripropylamine concentrations in water as low as 1 nM [20].

Cross-References

- ▶ Active Carbon Nanotube-Polymer Composites
- ▶ Atomic Layer Deposition
- ▶ Biosensors
- ▶ Hybrid Solar Cells
- ▶ Microfabricated Probe Technology
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanostructured Materials for Sensing

References

1. Zwilling, V., Aucouturier, M., Darque-Ceretti, E.: Anodic oxidation of titanium and TA6V alloy in chromic media. An electrochemical approach. Electrochim. Acta **45**(6), 921–929 (1999)
2. Gong, D., Grimes, C.A., Varghese, O.K., Hu, W.C., Singh, R.S., Chen, Z., Dickey, E.C.: Titanium oxide nanotube arrays prepared by anodic oxidation. J. Mater. Res. **16**(12), 3331–3334 (2001)
3. Mor, G.K., Varghese, O.K., Paulose, M., Shankar, K., Grimes, C.A.: A review on highly ordered, vertically oriented TiO_2 nanotube arrays: fabrication, material properties, and solar energy applications. Sol. Energy Mater. Sol. Cells **90**(14), 2011–2075 (2006). doi:10.1016/j.solmat.2006.04.007
4. Macak, J.M., Tsuchiya, H., Ghicov, A., Yasuda, K., Hahn, R., Bauer, S., Schmuki, P.: TiO_2 nanotubes: self-organized electrochemical formation, properties and applications. Curr. Opin. Solid State Mater. Sci. **11**(1–2), 3–18 (2007). doi:10.1016/j.cossms.2007.08.004
5. Hoyer, P.: Formation of a titanium dioxide nanotube array. Langmuir **12**(6), 1411–1413 (1996)
6. Michailowski, A., AlMawlawi, D., Cheng, G.S., Moskovits, M.: Highly regular anatase nanotubule arrays fabricated in porous anodic templates. Chem. Phys. Lett. **349**(1–2), 1–5 (2001)
7. Shankar, K., Basham, J.I., Allam, N.K., Varghese, O.K., Mor, G.K., Feng, X.J., Paulose, M., Seabold, J.A., Choi, K.S., Grimes, C.A.: Recent advances in the use of TiO_2 nanotube and nanowire arrays for oxidative photoelectrochemistry. J. Phys. Chem. C **113**(16), 6327–6359 (2009). doi:10.1021/jp809385x
8. Grimes, C.A., Mor, G.K.: TiO_2 Nanotube Arrays: Synthesis, Properties, and Applications. Springer, New York (2009)
9. Varghese, O.K., Paulose, M., Grimes, C.A.: Long vertically aligned titania nanotubes on transparent conducting oxide for highly efficient solar cells. Nat. Nanotechnol. **4**(9), 592–597 (2009). doi:http://www.nature.com/nnano/journal/v4/n9/suppinfo/nnano.2009.226_S1.html
10. Wu, X.J., Zhu, F., Mu, C., Liang, Y.Q., Xu, L.F., Chen, Q.W., Chen, R.Z., Xu, D.S.: Electrochemical synthesis and applications of oriented and hierarchically quasi-1D semiconducting nanostructures. Coord. Chem. Rev. **254**(9–10), 1135–1150 (2010). doi:10.1016/j.ccr.2010.02.014
11. Masuda, H.: Highly ordered nanohole arrays in anodic porous alumina. In: Wehrspohn, R.B. (ed.) Ordered Porous Nanostructures and Applications, pp. 37–56. Springer, New York (2005)
12. Kemell, M., Pore, V., Tupala, J., Ritala, M., Leskela, M.: Atomic layer deposition of nanostructured TiO_2 photocatalysts via template approach. Chem. Mater. **19**(7), 1816–1820 (2007). doi:10.1021/cm062576e
13. Ren, X., Gershon, T., Iza, D.C., Munoz-Rojas, D., Musselman, K., MacManus-Driscoll, J.L.: The selective fabrication of large-area highly ordered TiO_2 nanorod and nanotube arrays on conductive transparent substrates via sol–gel electrophoresis. Nanotechnology **20**(36), 365604 (2009). doi:10.1088/0957-4484/20/36/365604
14. Mun, K.S., Alvarez, S.D., Choi, W.Y., Sailor, M.J.: A stable, label-free optical interferometric biosensor based on TiO_2 nanotube arrays. ACS Nano **4**(4), 2070–2076 (2010). doi:10.1021/nn901312f

15. Losic, D., Simovic, S.: Self-ordered nanopore and nanotube platforms for drug delivery applications. Expert Opin. Drug Deliv. **6**(12), 1363–1381 (2009). doi:10.1517/17425240903300857
16. Gultepe, E., Nagesha, D., Sridhar, S., Amiji, M.: Nanoporous inorganic membranes or coatings for sustained drug delivery in implantable devices. Adv. Drug Deliv. Rev. **62**(3), 305–315 (2010). doi:10.1016/j.addr.2009.11.003
17. Roy, S.C., Varghese, O.K., Paulose, M., Grimes, C.A.: Toward solar fuels: photocatalytic conversion of carbon dioxide to hydrocarbons. ACS Nano **4**(3), 1259–1278 (2010). doi:10.1021/nn9015423
18. Lin, C.J., Yu, W.Y., Chien, S.H.: Transparent electrodes of ordered opened-end TiO_2-nanotube arrays for highly efficient dye-sensitized solar cells. J. Mater. Chem. **20**(6), 1073–1077 (2010). doi:10.1039/b917886d
19. Baker, D.R., Kamat, P.V.: Photosensitization of TiO_2 nanostructures with CdS quantum dots: particulate versus tubular support architectures. Adv. Funct. Mater. **19**(5), 805–811 (2009). doi:10.1002/adfm.200801173
20. Xu, Z.H., Yu, J.G.: A novel solid-state electrochemiluminescence sensor based on $[Ru(bpy)_3^{2+}]$ immobilization on TiO_2 nanotube arrays and its application for detection of amines in water. Nanotechnology **21**(24), 245501 (2010). doi:10.1088/0957-4484/21/24/245501

Titanium Dioxide

► Fate of Manufactured Nanoparticles in Aqueous Environment

Total Internal Reflection (Fluorescence) Velocimetry

► Evanescent-Wave Particle Visualization in Nano- and Microchannels

Toward Bioreplicated Texturing of Solar-Cell Anodes

► Toward Bioreplicated Texturing of Solar-Cell Surfaces

Toward Bioreplicated Texturing of Solar-Cell Surfaces

Aditi Risbud[1], Akhlesh Lakhtakia[2] and Michael H. Bartl[3]
[1]Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2]Department of Engineering Science and Mechanics, Pennsylvania State University, University Park, PA, USA
[3]Department of Chemistry, University of Utah, Salt Lake City, UT, USA

Synonyms

Toward bioreplicated texturing of solar-cell anodes

Definition

Biological species have developed myriad photonic structures to efficiently interact with light. The results of these interactions include large angular fields of view, reduced surface reflection, Bragg diffraction, and multiple scattering. Bioreplicated texturing of solar cell surfaces and photoanodes takes advantage of these excellent optical properties in biology. The aim is to convert biological structures into exact replicas made out of semiconductors, metals, and polymers and incorporate these replicas into solar cells to enhance light harvesting and light–matter interactions.

Introduction

The development of efficient, inexpensive, and environmentally benign energy sources is one of the biggest technological challenges of the twenty-first century. Even the most conservative predictions suggest worldwide demand for energy will triple within the next 40 years [1]. This rapidly growing energy demand, coupled with economic and ecologic impacts of fossil fuel–based

energy sources, has sparked research into low-cost and environment-friendly alternatives. Given the immense amount of energy the sun delivers to earth daily, solar power constitutes by far the most promising and attractive alternative to fossil fuels. However, to become the primary energy source for future consumption, highly efficient yet inexpensive strategies for direct conversion of solar power into thermal energy, electricity and chemical fuels are needed. A central challenge in solar energy conversion and photocatalysis is to optimize "light management" at the nanometer scale – from capture of sunlight to transfer and conversion of solar light into energy, fuels, or drivers of chemical reactions.

A prime example of highly evolved light management is photosynthesis – one of the most fundamental processes in biology. Calculations show photosynthesis produces approximately 90 TW of energy annually, approximately six times the current worldwide total energy consumption per year [1]. In addition, the light-independent phase of photosynthesis fixes carbon dioxide from the atmosphere, providing a potential mechanism for carbon dioxide sequestration. Photosynthetic organisms capture sunlight and channel solar energy from light-harvesting outer units to a reaction center by transferring energy along multiple functional units and converting it to fuel. It is this elaborate, integrated energy conversion system consisting of controlled and cooperative interaction between functional nanoscale components that results in efficient capture, transfer and use of solar energy. In contrast, artificial solar conversion units are still far from achieving the sophistication of their counterparts in nature. The goal is thus to examine simple models of biological systems, to learn from nature's complex energy conversion systems and to unlock and mimic its mechanisms [2–4].

For example, in biological materials, such as the wings of a butterfly or the exoskeleton of an insect, chitin and air are organized in ordered structures to produce optical interference leading to structural colors, camouflage and the control of solar radiative heat intake for thermal regulation. An early study calculated the effects of total solar radiation absorption and emissivity in the forms of beam radiation, which travels directly from the sun to a surface; diffuse radiation, which is a uniform distribution of light on a surface from all possible angles; and reflected radiation, which is light reflected from the ground on an insect thin film [5]. The outcome of this study suggests minute changes in the thickness of these biological thin films, typically 100 nm thick, strongly influence intake of energy from solar radiation. For example, changes in butterfly wing scale film thickness could significantly shift absorption levels (by up to 25 %) over a span of 80 nm, while diffuse solar absorption and emissivity remain relatively constant. Furthermore, iridescence has been found to be strongly correlated with higher absorption of the solar radiation [6].

In this entry, several novel concepts inspired by biological systems are introduced to enhance the first steps in solar energy conversion: the capture of sunlight through bioreplicated surfaces and light–matter interactions driving the conversion process. In the first section, strategies to increase the light-harvesting capability of conventional solar cells are discussed. These strategies are based on the excellent optical properties of insect eyes. In these structures, called compound eyes, intricate microscale and nanoscale features result in significantly enhanced angular fields of view, while minimizing surface reflectivity. Since these parameters are of utmost importance for optimizing light harvesting in conventional solid-state junction solar cells, researchers have recently started to develop methods for mass-replicating these biological structures as polymer films for deposition onto the surface of solar cells [7, 8]. In the second section, focus will be directed on methods to increase light–matter interactions in a new type of solar cell, the Grätzel-type electrochemical cell [9]. The heart of this low-cost alternative to solid-state junction devices is a nanocrystalline titania photoanode sensitized with visible-light absorbing dyes. Interestingly, coupling this photoanode to photonic crystal structures can increase its optical path, and therefore results in higher solar conversion efficiency [10, 11]. Since some of the most effective photonic crystal structures are found in biological systems, such as colored beetles and butterflies

[12], researchers are devising ways to replicate these remarkable architectures into nanocrystalline titania [13]. These efforts are reviewed and several examples of attempts to generate bioreplicated photoanodes are given.

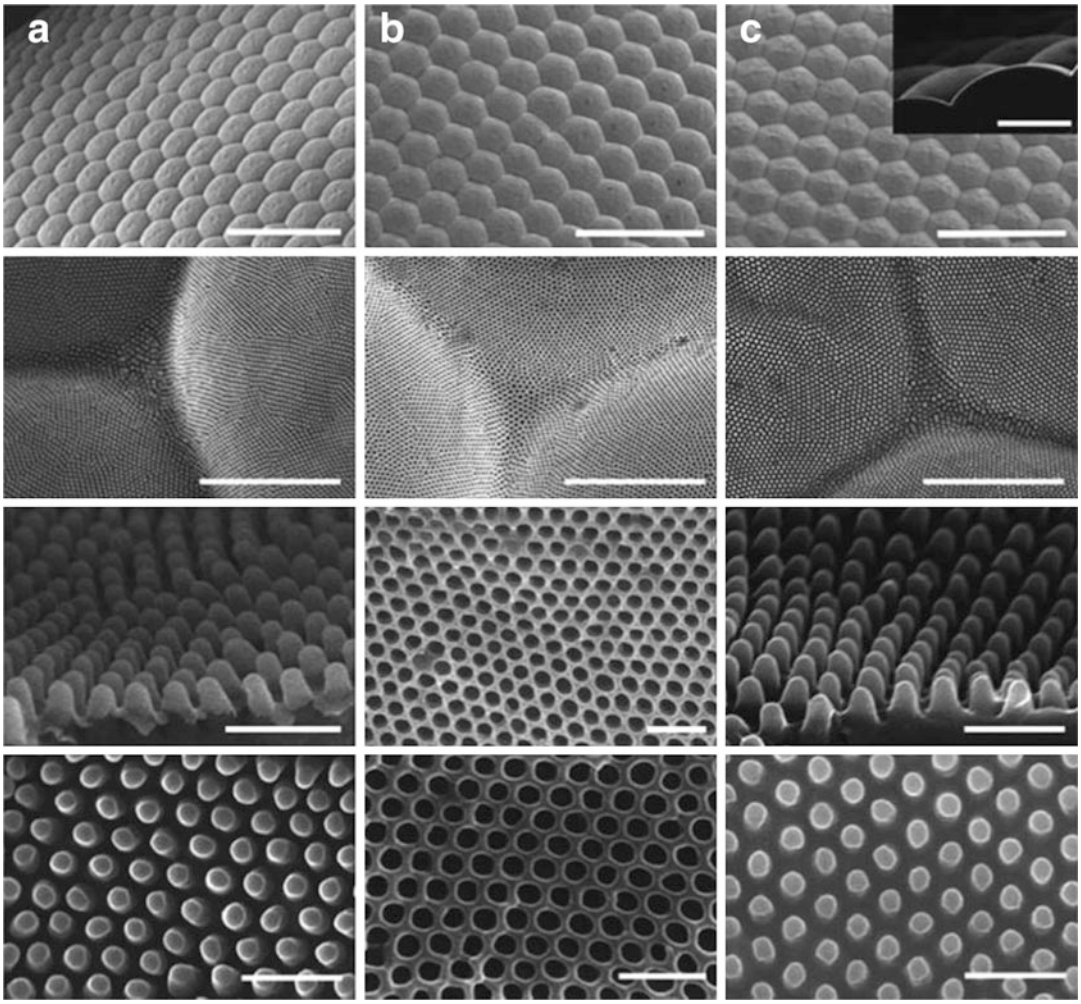
Surface Structuring of Solar Cells

A dominant factor in the performance of any solar conversion device is the amount of incident sunlight coupled into the active module of a given solar cell. The key considerations for efficient coupling of light into solar cells are to maximize the angular field of view of the solar absorption module and to minimize the overall reflection of light at the cell's surface. Increasing the angular field of view aims to harvest as much light as possible, in terms of direct illumination from the sun, as well as diffuse light. Harvesting the direct illumination of sunlight is highest when it hits the solar cell surface at a 90° angle. While, in principle, this can be achieved simply by installing a mechanical device that rotates the solar cell surface in response to the sun's position, such motion devices increase the cost of solar cells, require maintenance and use up part of the cell's energy. In addition, these motion devices would only minimally increase harvesting of diffuse light – scattered and reflected light that hits the solar cell from any possible angle.

During the course of millennia, biological structures have evolved to capture light from a wide range of directions. Indeed, houseflies have optimized their angular field of view through eye placement and underlying structure such that these insects can see 270° around them from a horizontal plane [2]. This remarkable ability stems from a compound eye structure, consisting of an array of microscale cylindrical lenses arranged on a curved surface. These tiny lenses not only capture light along the axis of the cylinder, but also can absorb light propagating in other directions through the cylinder's sidewalls. In addition, compound eyes can have nanometer-scale surface structures, which serve both as an antireflective [7] and a superhydrophobic surface [14]. These properties are the result of a gradual change in refractive index between the air and the eye. While superhydrophobicity keeps the eye clean, reduced reflection from insect eyes protects them from being visible to predators, and it also significantly enhances their photosensitivity in dim environments. As a result, these surfaces are an important enhancement strategy that could be used to boost efficiency in a solar cell.

Although these biological structures have complex nano-to-microscale features, recent advances in engineered biomimicry now allow scientists to replicate these structures using materials optimal for harvesting light for solar cells. Both physical and chemical nanofabrication techniques such as atomic layer deposition, conformal-evaporated-film-by-rotation technique, chemical vapor deposition, soft lithography and nanocasting are now employed by researchers to generate replicas of various biological structures in technologically relevant materials. For example, a research group developed a soft lithographic patterning technique to replicate a nanometer-scale antireflective coating found on the lens of a moth's compound eye [7]. A three-dimensional mold of this coating was designed with perfluoropolyether, an elastomeric polymer material capable of capturing nanometer-scale surface features. This mold was then used to create polyurethane replicas with antireflective behavior. As shown in Fig. 1, this approach results in high-fidelity replicas preserving topographical features as small as 100 nm. In addition, the group successfully used this technique to emboss antireflective nanoscale patterns in a widely used polymer blend for the photoactive region of an organic photovoltaic cell.

In an attempt to scale up compound eye bioreplication methods to mass-produce antireflective/large-field-of-view solar cell components, researchers developed a method to create reusable master negatives [8]. Specifically, corneas of the common blowfly were used as templates and replicated into metallic nickel. To achieve both a high-fidelity imprint of the hierarchically organized cornea surface and a stable and reusable master negative, the replication process was conducted in two steps. First, a modified version of the conformal-evaporated-film-by-rotation method was used to deposit a thin layer

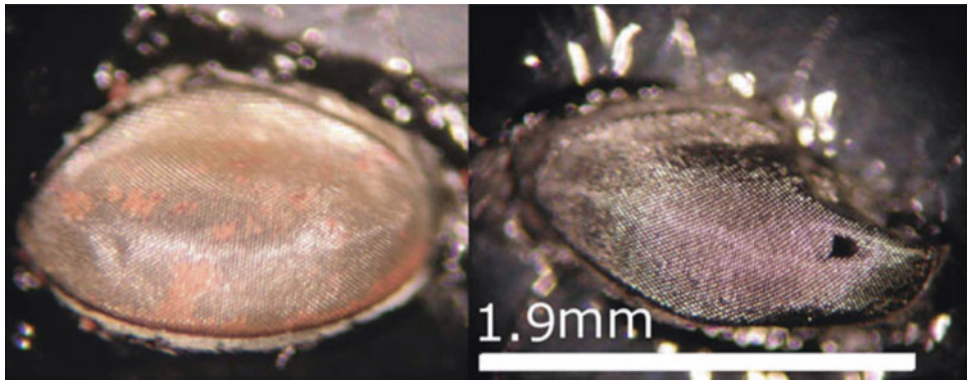


Toward Bioreplicated Texturing of Solar-Cell Surfaces, Fig. 1 Scanning electron micrographs of (**a**) the *Attacus atlas* moth eye structure, (**b**) the negative polymer replica, (**c**) the positive replica made by polymerization in the negative mold. *Inset* of (**c**) shows curved surface of the replica. Scale bars are 100 μm for the top images (*inset*: 20 μm), 5 μm for the second row images, and 500 nm for the third and fourth row images (Reproduced from Ref. [7] with permission of The Royal Society of Chemistry. doi:10.1039/C1SM05302G)

of nickel onto the surface relief of the cornea sample. In a second step, additional nickel was deposited by electroforming to strengthen the initial thin film replica. Finally, after removal of the biotemplate, a free-standing negative copy of the original cornea was obtained (Fig. 2). This group also showed this negative master could be successfully used (and reused) to create polymer replicas by a casting technique to faithfully reproduce micrometer-sized surface features. In addition, it was suggested by using stamping instead of casting, nanometer-sized features could be transferred into mass-produced polymer replicas.

Biotemplating of Photoanodes

Along with light-harvesting techniques, engineered biomimicry is a promising approach to enhance light–matter interactions in solar conversion systems. This is a particularly important area of research in next-generation, low-cost

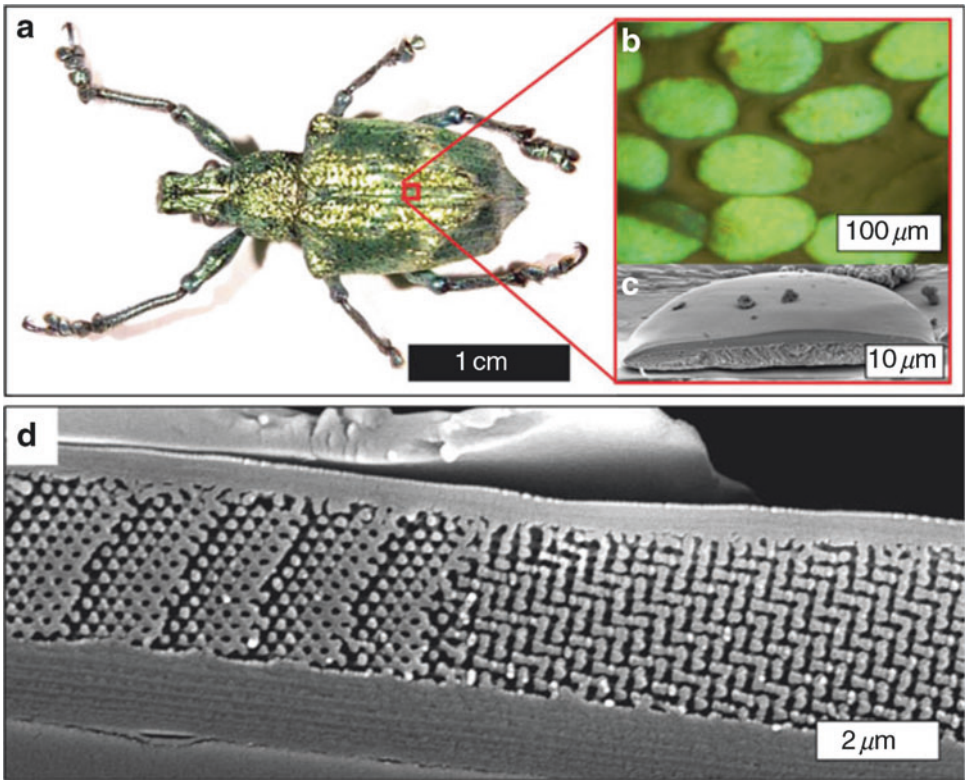


Toward Bioreplicated Texturing of Solar-Cell Surfaces, Fig. 2 Photographs of (*left*) a blowfly's cornea and (*right*) a nickel replica after removal of the biotemplate. The nickel structure is a positive replica when viewed from above (as shown), but a negative replica when viewed from below (Reproduced with permission from Ref. [8]. Copyright 2010 IOP Publishing Ltd.)

devices based on nanostructured semiconductors, conducting polymers and various hybrid device architectures. For example, theoretical studies recently showed coupling the photoanode of a Grätzel-type solar cell to a photonic crystal could significantly enhance its efficiency [10, 11]. Photonic crystals are periodically ordered structures with lattice parameters comparable to the optical wavelength of interest [15]. Due to this periodic variation in the refractive index, light within a photonic crystal experiences a combination of Bragg diffraction and multiple scattering events. As a consequence, light of certain frequencies undergoes multiple reflections as it slowly percolates through this structure. This results in a greatly increased nominal optical path length, an increase in light–photoanode interactions and significantly enhances the probability of light-to-charge conversion.

Interestingly, millions of years before we "discovered" photonic crystals, biological systems created such structures to produce vibrant colors, control solar absorption and enhance bio-fluorescence for a variety of reasons, ranging from camouflaging to frightening predators and attracting mates [12, 13]. Biological photonic structures thus were optimized to produce a wide range of optical effects and function under various illumination conditions. Not surprisingly, the range of photonic structures found in biological species is virtually limitless, ranging from simple one-dimensional multilayers, to two-dimensional lattices found in skeletons and hairs of sea animals and bird feathers and to various three-dimensional architectures within colored cuticle scales of many species of beetles and butterflies. For example, scientists discovered certain beetles obtain their color from diamond-based photonic crystals – one of the most effective lattice structures [16]. What's more, such photonic structures (as an example, Fig. 3 shows the green Brazilian beetle *Lamprocyphus augustus*) are still beyond our engineering capabilities. Similarly, photonic structures found in many butterfly wing scales have an astonishing hierarchy, consisting of ridges, lamellae and ribs, as well as highly periodic honeycomb and gyroid structures [12]. These elaborate hierarchical-scale structures act as optical components interacting with sunlight, such as multilayer reflectors, diffraction gratings and photonic crystals.

Only recently have scientists begun to use this invaluable portfolio of biological photonic crystal lattices for engineered biomimicry approaches. The method of choice is bioreplication: here, a biological photonic structure is used as a template to create positive or negative copies out of photocatalytically active compounds. An example of this approach is shown in Fig. 4 for conversion of the diamond-based photonic structure of the beetle *L. augustus*. Replication is achieved by a sol–gel double-imprint method using a sacrificial

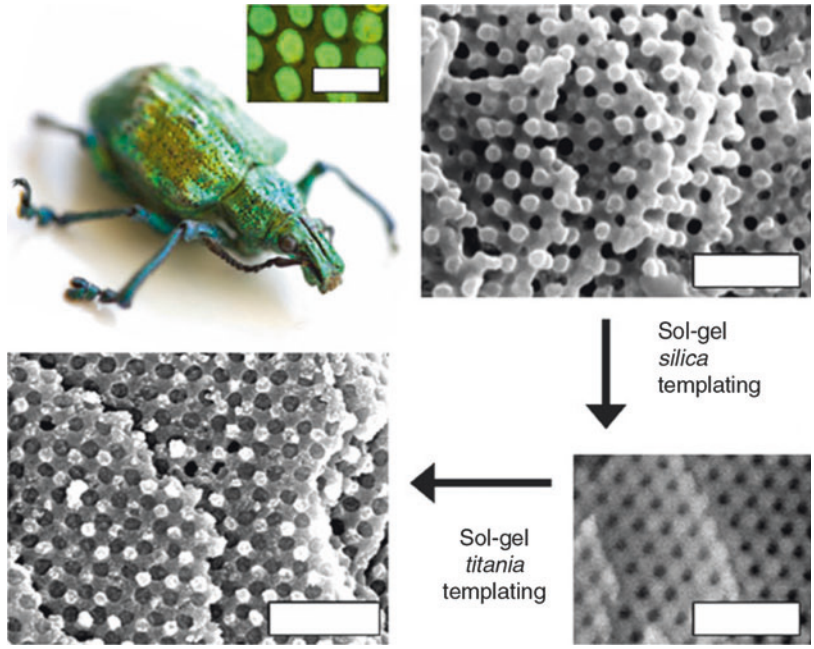


Toward Bioreplicated Texturing of Solar-Cell Surfaces, Fig. 3 Photographs of (**a**) the weevil *L. augustus* and (**b**) individual scales attached to its exoskeleton. Cross-sectional scanning electron microscopy images of (**c**) a single scale and (**d**) a region of a scale at higher magnification (Reproduced with permission from Ref. [16]. Copyright 2008, American Physical Society)

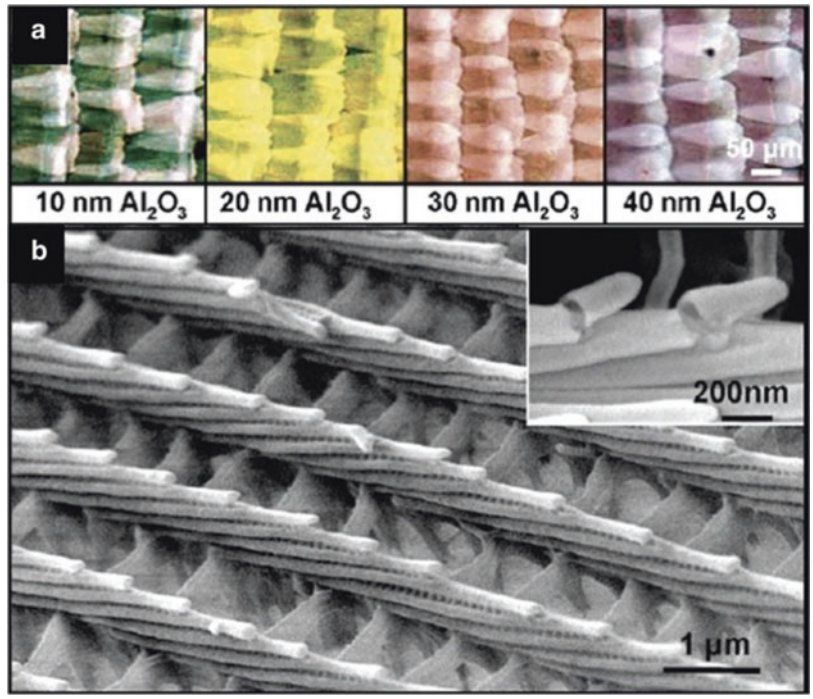
silica glass intermediary [17]. Sol–gel templating is quite attractive due to its simplicity and low cost. The original biological templates were backfilled with liquid precursors of the desired inorganic compound (e.g., silica, titania, alumina). After solidification of the inorganic compound, the original template was selectively removed by etching or heat treatment. This approach enabled a positive copy ("inverse-of-the-inverse") of the original structure composed of photocatalytically active nanocrystalline titania.

Along with sol–gel chemistry, other physical and chemical deposition methods can be used for replication of biological structures – as long as they operate under mild deposition conditions and temperatures. Successful methods include atomic layer deposition [18] and conformal-evaporated-film-by-rotation techniques [19]. In general, these techniques can be applied under noncorrosive reaction environments, mild pH conditions, and relatively low deposition temperatures (below 100–200 °C). Since these replicas are formed by deposition from a gas phase, these methods are highly efficient in infiltrating even complex three-dimensional frameworks as long as the internal structure is fully accessible. Furthermore, the deposition parameters can readily be controlled, enabling the amount of deposited material to be carefully controlled. As shown in Fig. 5, control of atomic layer deposition can be used to create shell-like replicas of the original biological structures with finely tuned replica shell thickness and, as a result, well-controlled optical properties.

Recently, a group of scientists took such bioreplication approaches one step closer to solar conversion applications by devising strategies to construct a Grätzel-type titania photoanode

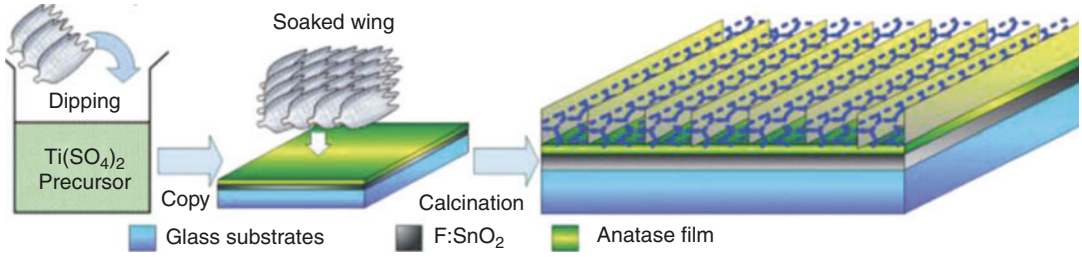


Toward Bioreplicated Texturing of Solar-Cell Surfaces, Fig. 4 Sol–gel bioreplication route of photonic weevil scales into titania replicas (*top left*). Scanning electron microscopy images of the original biopolymeric photonic structure (*top right*), the negative replica made of silica (*bottom right*), and the positive replica made of titania (*bottom left*). Scale bars are 200 μm (*top left*) and 1 μm (*all others*) (Reproduced with permission from Ref. [17]. Copyright Wiley-VCH Verlag GmbH & Co. KGaA)



Toward Bioreplicated Texturing of Solar-Cell Surfaces, Fig. 5 (**a**) Optical microscope image of the alumina-coated *Morpho peleides* butterfly wing scales; color changes from original blue to pink. (**b**) Higher magnification scanning electron microscopy image of alumina replicas of butterfly wing scales and two broken rib tips (*inset*) (Reproduced with permission from Ref. [18]. Copyright 2006 American Chemical Society)

coupled to arrays of replicated butterfly scales [20]. In this approach, 1 × 1 cm^2 samples of butterfly wings were soaked in a titania sol–gel precursor and arranged on top of a Grätzel-type photoanode (Fig. 6). This composite structure was then heat-treated to allow the top butterfly-wing replica layer to bond to the bottom titania layer. This heat treatment (500 °C in air) also induced



Toward Bioreplicated Texturing of Solar-Cell Surfaces, Fig. 6 Schematic illustration of creating bioreplicated titania photoelectrodes. Step 1, soaking the butterfly wings in the titania precursor; step 2, calcination of soaked butterfly wings on top of an anatase titania film (Reproduced with permission from Ref. [20]. Copyright 2009 American Chemical Society)

titania crystallization into polycrystalline anatase and removed the biopolymeric template. Despite structural shrinkage during the heat-based framework densification-crystallization treatment, these replicated materials displayed improved light absorption behavior as compared to anodes without the replicated butterfly wings. While this outcome is an important step toward bioreplicated solar cells, these researchers also note major challenges remaining in generating such biotemplated anodes, optimizing their electrical conductivity and integrating these structures into functional solar conversion devices.

Concluding Remarks

Engineered biomimicry has arrived as a serious contender in materials research and engineering. While a few years ago the concepts of supervision from bug eyes, ultrastrong fabrics from spiders, vertically climbing gecko-like robots, or optical devices from beetles was more the stuff of Hollywood science fiction, these ideas have found their way into research laboratories and inspire the development of new devices in electronics, optics and mechanics. Indeed, although the pure beauty of natural iridescence in the form of opal gem stones, jewel beetles and feather ornaments has fascinated and inspired mankind for thousands of years, today, these materials and optical concepts offer revolutionary new ways to manipulate light. In addition, these materials hold great potential for more efficient solar cells and photocatalysts, new optical reflectors and sensors and ultrafast switches.

An exciting area in engineered biomimicry is bioreplication, in which complex biological structures are converted into replicas made of semiconductors, metals or polymers. In this entry, a few examples of solar-related bioreplication strategies to increase light-harvesting ability and light–matter interaction efficiency in various types of solar cells were discussed. While many of these approaches are promising and first successful steps have been achieved, it should be emphasized there are also several limitations and challenges in bioreplication. For example, biological samples have an inherent size limit given by the dimensions of the eyes, scales, or feathers used as templates. Direct use and integration of biological samples, therefore, have limited applications, suggesting more efficient mass-duplication routes have to be developed. In addition, most natural structures have a much higher defect density than what is tolerated by current engineering standards, and no two natural structures are exactly the same. For example, the structural parameters of colored scales taken from different parts of a beetle, or even within a single scale, can vary by as much as 20 %.

Consequently, bioreplication must be viewed in its correct and just perspective: Although materials prepared by bioreplication will perhaps never achieve the precision, uniformity and reproducibility of their counterparts engineered by micro- or nano-fabrication methods, bioreplication opens the door to the full portfolio of highly complex

hierarchical structures found in nature – structures still far from our engineering capabilities. It will therefore be of vital importance to devise strategies to use and integrate materials created by engineered biomimicry with micro- or nanofabrication methods and combine the appealing aspects of two worlds – biology and materials science – in a concerted approach. Engineered biomimicry has found its way from science fiction into the laboratories; whether it makes the next step from fundamental research to wide-scale applications will depend on scientific creativity and imagination.

Cross-References

- ▶ Biomimetics of Optical Nanostructures
- ▶ BioPatterning
- ▶ Hybrid Solar Cells
- ▶ Moth-Eye Antireflective Structures
- ▶ Nanomaterials for Excitonic Solar Cells
- ▶ Nanostructures for Energy
- ▶ Nanostructures for Photonics
- ▶ Structural Color in Animals

References

1. Hoffert, M.I., Caldeira, K., Jain, A.K., Haites, E.F., Harvey, L.D.D., Potter, S.D., Schlesinger, M.E., Schneider, S.H., Watts, R.G., Wigley, T.M.L., Wuebbles, D.J.: Energy implications of future stabilization of atmospheric CO_2 content. Nature **395**, 881 (1998)
2. Chiadini, F., Fiumara, V., Scaglione, A., Pulsifer, D.P., Martín-Palma, R.J., Pantano, C.G., Lakhtakia, A.: Insect eyes inspire improved solar cells. Opt. Photonics News **22**(4), 38 (2011)
3. Paris, O., Burgert, I., Fratzl, P.: Biomimetics and biotemplating of natural materials. MRS Bull. **35**, 219 (2010)
4. Jin, L., Zhai, J., Heng, L., Wei, T., Wen, L., Jiang, L., Zhao, X., Zhang, X.: Bio-inspired multi-scale structures in dye-sensitized solar cell. J. Photochem. Photobiol. C **10**, 149 (2009)
5. Heilman, B.D., Miaoulis, I.N.: Insect thin film as solar collectors. Appl. Optics **33**, 6642 (1994)
6. Bosi, S.G., Hayes, J., Large, M.C.J., Poladian, L.: Color, iridescence, and thermoregulation in Lepidoptera. Appl. Opt. **47**, 5235 (2008)
7. Ko, D.-H., Tumbleston, J.R., Henderson, K.J., Euliss, L.E., DeSimone, J.M., Lopez, R., Samulski, E.T.: Biomimetic microlens array with antireflective "moth-eye" surface. Soft Matter **7**, 6404 (2011). doi:10.1039/C1SM05302G
8. Pulsifer, D.P., Lakhtakia, A., Martín-Palma, R.J., Pantano, C.G.: Mass fabrication technique for polymeric replicas of arrays of insect corneas. Bioinsp. Biomim. **5**, 036001 (2010)
9. Grätzel, M.: Photoelectrochemical cells. Nature **414**, 338 (2001)
10. Mihi, A., Miguez, H.: Origin of light-harvesting enhancement in colloidal-photonic-crystal-based dye-sensitized solar cells. J. Phys. Chem. B **109**, 15968 (2005)
11. Halaoui, L.I., Abrams, N.M., Mallouk, T.E.: Increasing the conversion efficiency of dye-sensitized TiO_2 photoelectrochemical cells by coupling to photonic crystals. J. Phys. Chem. B **109**, 6334 (2005)
12. Srinivasarao, M.: Nano-optics in the biological world: beetles, butterflies, birds, and moths. Chem. Rev. **99**, 1935 (1999)
13. Jorgensen, M.R., Bartl, M.H.: Biotemplating routes to three-dimensional photonic crystals. J. Mater. Chem. **21**, 10583 (2011). doi:10.1039/C1JM11037C
14. Koch, K., Bhushan, B., Barthlott, W.: Diversity of structure, morphology and wetting of plant surfaces. Soft Matter **4**, 1943 (2008)
15. Joannopoulos, J.D., Villeneuve, P.R., Fan, S.H.: Photonic crystals: putting a new twist on light. Nature **386**, 143 (1997)
16. Galusha, J.W., Richey, L.R., Gardner, J.S., Cha, J.N., Bartl, M.H.: Discovery of a diamond-based photonic crystal structure in beetle scales. Phys. Rev. E **77**, 050904 (2008)
17. Galusha, J.W., Jorgensen, M.R., Bartl, M.H.: Diamond-structured titania photonic bandgap crystals from biological templates. Adv. Mater. **22**, 107 (2010)
18. Huang, J., Wang, X., Wang, Z.L.: Controlled replication of butterfly wings for achieving tunable photonic properties. Nano Lett. **6**, 2325 (2006)
19. Martín-Palma, R.J., Pantano, C.G., Lakhtakia, A.: Biomimetization of butterfly wings by the conformal-evaporated-film-by-rotation technique for photonics. Appl. Phys. Lett. **93**, 083901 (2008)
20. Zhang, W., Zhang, D., Fan, T., Gu, J., Ding, J., Wang, H., Guo, Q., Ogawa, H.: Novel photoanode structure templated from butterfly wing scales. Chem. Mater. **21**, 33 (2009)

Toxic Mechanisms

▶ Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment

Toxicity

► In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles

Toxicity of Metal and Metal Oxide Nanoparticles on Prokaryotes (Bacteria), Unicellular and Invertebrate Eukaryotes

► Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates

Toxicity/Ecotoxicity

► Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms

Toxicology: Plants and Nanoparticles

Marie Carrière and Camille Larue
Laboratoire Lésions des Acides Nucléiques, Commissariat à l'Energie Atomique, SCIB, UMR-E 3 CEA/UJF-Grenoble 1, INAC, Grenoble, France

Synonyms

Ecotoxicity; Nanoparticle; Phytotoxicity

Definition

Impact of engineered nanoparticles (ENP) on plant growth, development, physiology, and morphology.

Key Research Findings

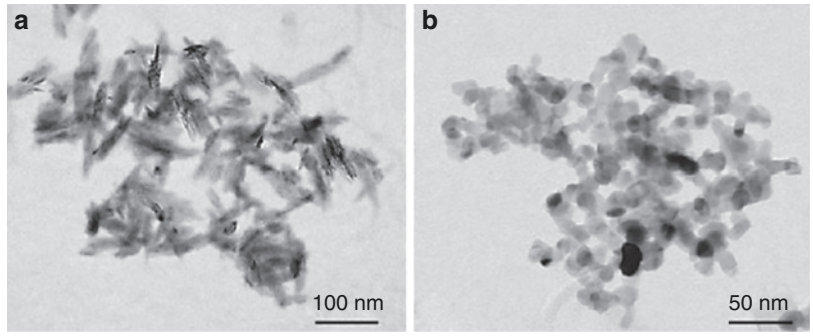
Three research areas are currently being investigated to define engineered nanoparticle (ENP) impacts on plant development (seed germination, root and shoot elongation, etc.), on plant physiology, and ENP uptake and translocation through plants.

Impact on Plant Development

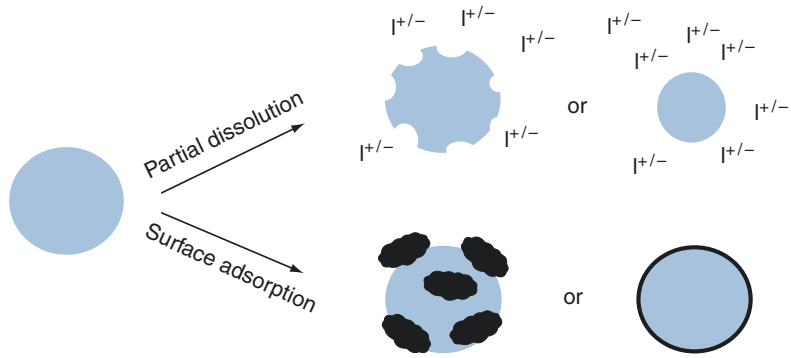
In this research area, regulatory tests are used, such as the ones described in EPA guidelines [1]. These guidelines have been developed for toxicity testing of chemical pollutants, and their relevance for toxicity testing of ENPs is under debate, with several groups working on their implementation to render them accurate. Among these tests, seed germination, root and shoots elongation assays have been intensively used to assess ENP impact on plant development, and the obtained results are either positive, or negative, or inconsequential. For instance multi-walled carbon nanotubes (CNT) have been shown to cause plant biomass decrease, particularly on *Cucurbita pepo* [2], while they showed no impact on rye grass, lettuce, rape, radish, corn, and cucumber [3]. The same is true for TiO_2 nanoparticles, which were shown in some studies to alter leaf growth and transpiration on maize [4], while they did not cause any alteration in the growth of willow tree plantlets [5].

Described results are difficult to compare, since model plant species are often different, and the considered nanoparticles are often different as well: Multi-walled CNTs may differ in their production process, post-synthesis surface treatment, and in the preparation of dispersed suspensions. TiO_2 nanoparticles may differ in their primary diameter, specific surface area, shape (round, elongated, see Fig. 1), crystalline structure, etc. These conflicting effects are easily explained by the broad panel of tested ENPs, and their very different physicochemical properties.

Moreover indirect effects such as ENP dissolution or impact of the solvent (or other environmental pollutants) which may be adsorbed on ENP surface have to be considered (Fig. 2).



Toxicology: Plants and Nanoparticles, Fig. 1 Titanium dioxide nanoparticles, rutile and elongated-shaped (**a**) or anatase and round-shaped (**b**)



Toxicology: Plants and Nanoparticles, Fig. 2 Schematic representation of potential ENP evolution in the environment: ENPs may partially dissolve, releasing free ions ($I^{+/-}$). They may also adsorb environmental pollutants or solvent on their surface

The best examples of the impact of ENP dissolution on plant development was demonstrated with Cu and ZnO nanoparticles. Even if the impact of dissolved ions did not explain all the toxic effects observed in the following studies, they greatly contributed to them. Indeed, Cu NPs have been demonstrated to cause inhibition of plant growth and to accumulate in mung beans and wheat [6], but also have some influence on the germination of lettuce [7]. Among all the ENPs tested for their impact on rye grass, lettuce, rape, radish, corn, and cucumber, only Zn and ZnO NPs reduced germination of ryegrass and corn, respectively. These two ENPs also reduced root elongation [3]. ZnO-NP dissolution was then considered by the same team the year after [8].

Uptake and Translocation

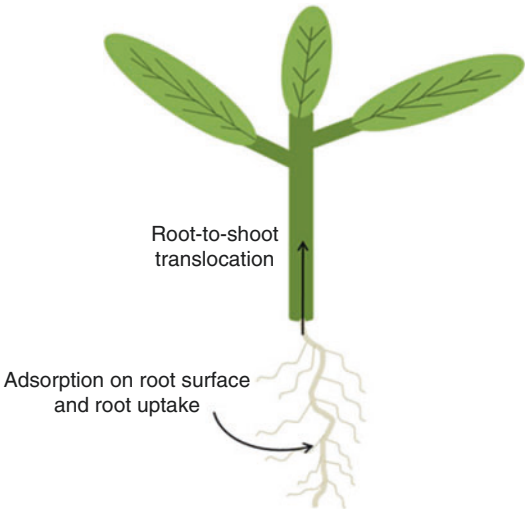
Despite the relative impermeability of plant roots to particulate matter, due to their small size, ENPs have been demonstrated to penetrate plant root wall, even if most of the ENPs remain adsorbed on root surface. It has for instance been demonstrated for Fe_3O_4 NPs on pumpkins [9], for TiO_2-alizarin red nanoconjugates on *Arabidopsis thaliana* [10], and CeO_2 nanoparticles on alfalfa, cucumber, corn, and tomato [11]. The consequence of such result is that ENPs may then be transferred to the food chain. Note that the ENP content in these plants is very low: ENP uptake is not an efficient process; consequently the amount of ENPs reaching the food chain would also be low.

Several techniques are available to demonstrate ENP uptake by plants. First, electron microscopy, due to its very high spatial resolution (typically 1 nm or below), is the method of choice to identify the presence of ENPs in biological samples, and particularly in plants. Scanning electron microscopy (SEM) permits to observe ENPs on the surface of plant tissues (roots or leaves), but does not give access to the internalized ENPs. Conversely transmission electron microscopy (TEM), achieved on dehydrated plant samples included in a resin and cut in ultrathin sections, permits the detection of nanosized materials in

plant cells or extracellular compartments. TEM has been successfully used to prove that carbon-based ENPs were accumulated in rice plantlets [12]. Coupling these two techniques with energy-dispersive spectroscopy (EDS) allows a local chemical identification of the electron-dense spots observed by TEM or SEM in plant samples, proving that they contain the chemical element constituting the ENP (when EDS analysis is performed, then spatial resolution ranges from 30 to 50 nm).

Another valuable technique is micro-X-ray-fluorescence (μ-XRF), achieved either on a laboratory source or using synchrotron radiation. This technique allows local identification of chemical composition, and basically the identification of hot spots of one particular element constituting the ENP. With this technique, spatial resolution can be as high as 100 nm on some synchrotron beam lines such as ID22 (ESRF, Grenoble, France). On some synchrotron beam lines, this technique is coupled to X-ray absorption spectroscopy (XAS), which allows describing the local environment of the analyzed element, giving the absolute proof that the element is in the nanoparticulate form, and not present as dissolved ions. Moreover it informs on ENP crystal structure, and sometimes on their size (it is the case for TiO_2-NPs). For instance this technique has been used to evidence that alfalfa, cucumber, corn, and tomato plants accumulated CeO_2-NPs in their roots [11].

Regarding the detection of CNTs, TEM enables the observation of single or clusters of NTCs. However, images obtained by TEM can be misinterpreted, since the shape of CNT is close to the structure of natural vegetal compounds. In order to avoid any misinterpretation, very specific techniques such as photothermal and photoacoustic detection have been developed and used for the identification of their accumulation and translocation in tomato plantlets [13]. These techniques are highly sensitive and specific, allowing detecting single MWCNT in plant tissues. Raman spectroscopy is also a very specific technique, which has been used to identify the presence of CNTs inside plant seeds thanks to their G band (1568 cm^{-1}) [14].



Toxicology: Plants and Nanoparticles, Fig. 3 Schematic representation of ENP adsorption on root surface, root uptake, and transfer to the shoot

Another point is that, when taken up in plant roots, ENPs are in some cases transferred from the roots to the shoots (Fig. 3). This phenomenon, also called translocation, was demonstrated for instance with Au-NPs in tobacco [15] and Fe_3O_4 NPs on pumpkins [9]. Carbon-based NPs have also been demonstrated to translocate from the roots to the shoots and from the shoots to the seeds. The newly developed plantlets, resulting from the germination of these contaminated seeds, also contained carbon-based NPs, meaning that ENPs were transmitted to the next generation [12].

Very recently, thanks to these techniques, gold NPs were shown to be transferred to higher trophic levels: hornworms fed with Au-NP-exposed tobacco were shown to be contaminated with Au-NPs [15]. In this particular case, gold maximal concentration in tobacco reached less than 100 mg Au kg^{-1} dry tobacco.

Impact on Plant Physiology

The impact of ENP exposure on plant physiology and metabolism is poorly described, underlining that the area of research of ENP phytotoxicity is at its infancy. A very recent study describes the impact of NTC on plant gene expression [13]. Many stress-related genes are activated

upon exposure to NTC, and among them the gene encoding the protein constituting water channels (aquaporins) [13]. This channel regulates water permeability, and therefore is significantly implicated in seed germination and plant growth. In one study TiO_2-NPs have also been demonstrated to inhibit root growth and transpiration by interference with water transport [4]. Actually, ENPs blocked apoplastic water flow in plant roots (within a plant, the apoplast is the free diffusional space outside the plasma membrane), certainly by obstructing nanosized cell wall pores, but also by reducing the exclusion diameter of these pores. However TiO_2-NPs are likely to agglomerate when dispersed in a saline, neutral aqueous solution. Therefore, exposing plant roots to TiO_2-NPs dispersed in a $CaCl_2$ solution at near-neutral pH, even if it is more physiologically relevant, means that plants were exposed to highly agglomerated ENPs. It is thus not surprising that water flow through the plant was blocked, and it rather derives from root exposure to a high concentration (0.3 or 1 gl^{-1}) of agglomerated material, than by a specific effect of ENPs.

Several studies reported that ENPs impacted plant photosynthetic activity. In a series of articles, TiO_2-NPs were shown to induce an increase in spinach germination, growth, nitrogen metabolism, and photosynthesis, and protected plants from oxidative damage induced by UV radiation [16]. Indeed TiO_2-NPs promote energy transfer from chlorophyll b and carotenoids to chlorophyll a, but also to the P680 chlorophyll a of D1/D2/Cyt 559 complex, in vitro. The authors conclude that nanosized TiO_2 would enter spinach cells and chloroplasts more easily due to their small size, then bind to the D1/D2/Cyt 559 complex and improve absorption, transfer, and conversion of light energy in chloroplasts [16].

Lastly, genotoxic effects of ENPs are currently investigated on several plant species. Root exposure to Ag-NPs reduce the mitotic index of exposed cells in *Allium cepa*, and more generally impair the cell division process, resulting in chromatin bridges, stickiness, disturbed metaphase, chromosomal breaks, and cell disintegration [17]. In the same test (*A. cepa* test is a standardized assay) TiO_2-NPs show no genotoxic impact, neither chromosomal aberration nor micronuclei [18]. Conversely DNA damage is detected in *A. cepa* and *Nicotiana tabaccum* exposed to TiO_2-NPs, by using comet assay and DNA laddering experiment. This genotoxic effect is correlated with the induction of malondialdehyde production, which is a marker of lipid peroxidation, as a result of oxidative stress [19].

Concluding Remarks

As a conclusion, the impact of ENPs on plants is still poorly documented; however some effects have been reported. Some ENPs, particularly those which partly dissolve, impact plant developments, i.e., seed germination, root and shoot elongation, and plant biomass. Some ENPs are taken up by plant roots, sometimes translocated to the shoots, and even transmitted to the next generation via their transfer to seeds. When accumulated in plant leaves, some ENPs can be transferred to primary consumer, for instance hornworms, meaning that they reach the food chain. The genotoxic potential of some ENPs has been demonstrated, although other articles relate opposite results. Some ENPs would modulate photosynthesis by impacting photosystem activity. Note that only some ENPs have these properties and exert these effects, the results described herein definitely cannot be generalized to all ENPs. Note also that almost all the results were obtained on plants exposed to ENPs in vitro, but only one publication describes the impact of ENPs-amended soil on plant development. Data thus need to be collected on environmentally relevant systems, i.e., plants cultivated on natural soils and contaminated with low doses of ENPs since massive contamination of the environment with ENPs would improbably occur.

Cross-References

- ▶ Cellular Mechanisms of Nanoparticle Toxicity
- ▶ Effect of Surface Modification on Toxicity of Nanoparticles

References

1. U.S. Environmental Protection Agency: Ecological effects test guidelines (OPPTS 850.1200): seed germination/root elongation toxicity test. http://www.epa.gov/opptsfrs/publications/OPPTS_Harmonized/850_Ecological_Effects_Test_Guidelines/Drafts/850-4200.pdf (1996)
2. Stampoulis, D., Sinha, S.K., White, J.C.: Assay-dependent phytotoxicity of nanoparticles to plants. Environ. Sci. Technol. **43**(24), 9473–9479 (2009)
3. Lin, D., Xing, B.: Phytotoxicity of nanoparticles: inhibition of seed germination and root growth. Environ. Pollut. **150**(2), 243–250 (2007)
4. Asli, S., Neumann, P.M.: Colloidal suspensions of clay or titanium dioxide nanoparticles can inhibit leaf growth and transpiration via physical effects on root water transport. Plant Cell Environ. **32**(5), 577–584 (2009)
5. Seeger, E., et al.: Insignificant acute toxicity of TiO_2 nanoparticles to willow trees. J. Soils Sediments **9**(1), 46–53 (2009)
6. Lee, W.M., et al.: Toxicity and bioavailability of copper nanoparticles to the terrestrial plants mung bean (*Phaseolus radiatus*) and wheat (*Triticum aestivum*): plant agar test for water-insoluble nanoparticles. Environ. Toxicol. Chem. **27**(9), 1915–1921 (2008)
7. Shah, V., Belozerova, I.: Influence of metal nanoparticles on the soil microbial community and germination of lettuce seeds. Water Air Soil Pollut. **197**(1–4), 143–148 (2009)
8. Lin, D., Xing, B.: Root uptake and phytotoxicity of ZnO nanoparticles. Environ. Sci. Technol. **42**(15), 5580–5585 (2008)
9. Zhu, H., et al.: Uptake, translocation, and accumulation of manufactured iron oxide nanoparticles by pumpkin plants. J. Environ. Monit. **10**(6), 713–717 (2008)
10. Kurepa, J., et al.: Uptake and distribution of ultrasmall anatase TiO_2 alizarin red S nanoconjugates in *Arabidopsis thaliana*. Nano Lett. **10**(7), 2296–2302 (2010)
11. Lopez-Moreno, M.L., et al.: X-ray absorption spectroscopy (XAS) corroboration of the uptake and storage of CeO_2 nanoparticles and assessment of their differential toxicity in four edible plant species. J. Agric. Food Chem. **58**(6), 3689–3693 (2010)
12. Lin, S., et al.: Uptake, translocation, and transmission of carbon nanomaterials in rice plants. Small **5**(10), 1128–1132 (2009)
13. Khodakovskaya, M.V., et al.: Complex genetic, photothermal, and photoacoustic analysis of nanoparticle-plant interactions. Proc. Natl. Acad. Sci. U. S. A. **108**(3), 1028–1033 (2011)
14. Khodakovskaya, M., et al.: Carbon nanotubes are able to penetrate plant seed coat and dramatically affect seed germination and plant growth. ACS Nano **3**(10), 3221–3227 (2009)
15. Judy, J.D., Unrine, J.M., Bertsch, P.M.: Evidence for biomagnification of gold nanoparticles within a terrestrial food chain. Environ. Sci. Technol. **45**(2), 776–781 (2011)
16. Mingyu, S., et al.: Promotion of nano-anatase TiO_2 on the spectral responses and photochemical activities of D1/D2/Cyt b559 complex of spinach. Spectrochim. Acta A **72**, 1112–1116 (2009)
17. Kumari, M., Mukherjee, A., Chandrasekaran, N.: Genotoxicity of silver nanoparticles in *Allium cepa*. Sci. Total Environ. **407**(19), 5243–5246 (2009)
18. Klancnik, K., et al.: Use of a modified *Allium* test with nanoTiO_2. Ecotoxicol. Environ. Saf. **74**(1), 85–92 (2011)
19. Ghosh, M., Bandyopadhyay, M., Mukherjee, A.: Genotoxicity of titanium dioxide (TiO_2) nanoparticles at two trophic levels plant and human lymphocytes. Chemosphere **81**(10), 1253–1262 (2010)

Transdisciplinary Approaches

► Biornametics: Architecture Defined by Natural Patterns

Transduction

► Micromachined 3-D Force Sensors

Transfer

► Fate of Manufactured Nanoparticles in Aqueous Environment

Translational Movement of Polynucleotides in a Nonuniform Electric Field

► Dielectrophoresis of Nucleic Acids

Transmission Electron Microscopy

Helmut Kohl
Physikalisches Institut, Westfälische Wilhelms-Universität Münster, Münster, Germany

Definition

Transmission electron microscopy is a method to image small thin objects by the use of the transmitted electrons.

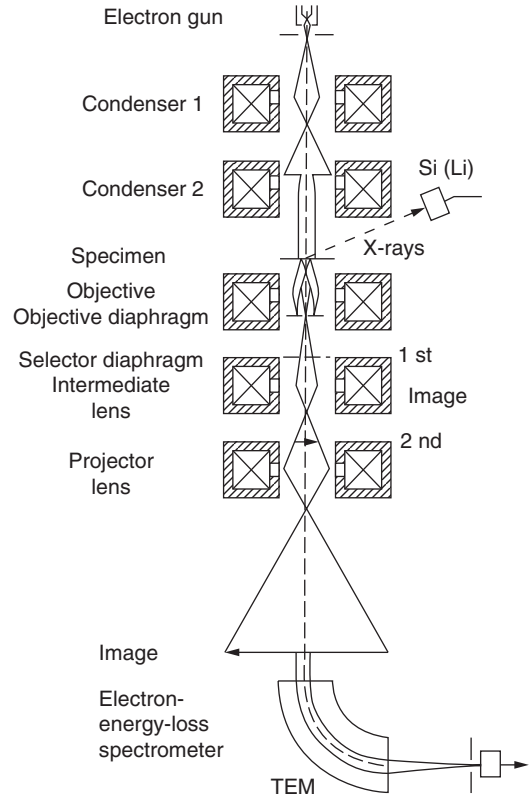
The Transmission Electron Microscope (TEM)

Similar to the general scheme of a light microscope, a transmission electron microscope [1, 2] consists of an electron source, a condenser system, an objective lens, and a projector system as shown in Fig. 1. Many transmission electron microscopes have additional instruments attached to it, such as an X-ray detector and/or an energy loss spectrometer in order to be able to perform elemental analyses. These techniques will be explained later on.

The electron source, commonly called "gun," is composed of a cathode and an accelerating system. There are different types of cathodes being used:

- Thermal cathodes
- Schottky emitters, often misleadingly called (hot) field emission cathodes
- (Cold) field emission cathodes

Thermal cathodes make use of thermal emission of electrons at high temperatures. Most often they use tungsten or LaB_6 as emitting material. The emission barrier can be lowered by applying a strong electrostatic field to a tip (Schottky effect). This is the basic principle of a Schottky emitter. Using an even stronger electrostatic field at a sharp tip, one can extract electrons by means of the quantum mechanical tunnel effect. The resulting field emission cathodes require ultrahigh vacuum to operate, because otherwise positively charged ions will be accelerated toward the tip eventually leading to its destruction. All of these cathodes are followed by an accelerating system ending with the anode. Most transmission electron microscopes usually operate at voltages of 100–300 kV, even though there are special low-voltage TEMs operating at voltages from 20 kV upward and high-voltage TEMs operating at voltages up to 1.2 MeV.



Transmission Electron Microscopy, Fig. 1 Schematic ray paths for a transmission electron microscope (*TEM*) equipped with an additional X-ray spectrometer and an energy loss spectrometer [1]

The role of the condenser system is to transfer the electrons from the gun onto the specimen. It is composed of two to five electron lenses to allow to independently vary the illumination angle and the size of the illuminated area. The transmitted electrons are then focused to a magnified intermediate image which in turn is magnified and transferred into the final image plane.

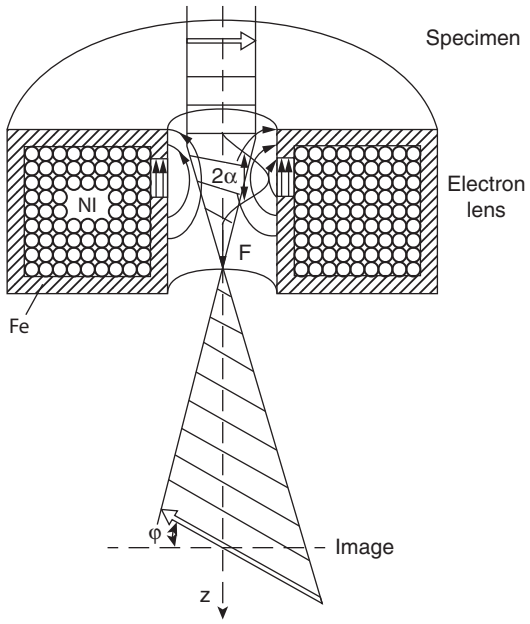
Magnetic Electron Lenses

To focus an electron beam, we need lenses operating similar to glass lenses in light optics. These can be constructed by coils encapsulated by a ferromagnetic material as shown in Fig. 2. To understand their working principle, let us consider the movement of an electron in the homogeneous magnetic field of a long coil. The Lorentz force on the electron is given by

$$\boldsymbol{F} = -e\boldsymbol{v} \times \boldsymbol{B} \tag{1}$$

The velocity $\boldsymbol{v}$ can be separated into a component parallel v_{II} to the magnetic induction $\boldsymbol{B}$ and one perpendicular component $v_{\perp}$. As there is no force operating on the parallel component, v_{II} remains constant. Perpendicular to the magnetic induction $\boldsymbol{B}$, the electron performs a circular motion with a radius r that is determined by the equivalence of the Lorentz force to the centrifugal force:

$$\frac{mv_{\perp}^2}{r} = ev_{\perp}B \tag{2}$$



Transmission Electron Microscopy, Fig. 2 Schematic diagram of an electromagnetic lens [1]

or

$$r = \frac{mv_{\perp}}{eB} \tag{3}$$

An electron starting from the optic axis will cross it again after it has completed one circle – in other words, the electrons starting from the initial plane are all imaged in the latter plane. For a homogeneous magnetic field, it can be shown that the focal length f is determined by

$$\frac{1}{f} = \frac{e}{8m\phi^*}B^2L \tag{4}$$

where $\phi^* = \varphi(1 + e\varphi/(2E_0))$ is the (relativistically corrected) potential, φ the electrostatic potential, $E_0 = 511$ keV the rest energy of the electron, and L the length of the coil. As can be seen from Eq. 4, the focal length depends strongly on the magnetic induction $\boldsymbol{B}$. This general fact is also true for the commonly used "short magnetic lenses" that use an inhomogeneous magnetic field. The focal length of a magnetic electron lens in an electron microscope can therefore be easily modified by adjusting the current through the coil, thus changing the magnetic field.

Unfortunately, electron lenses suffer from strong spherical and chromatic aberrations [3, 4] that limit the spatial resolution limit for conventional electron microscopes to about 0.15 nm. Just in the very last years, it has become possible to correct for aberrations so that with the most advanced instruments one can achieve a resolution limit of about 0.05 nm.

As the focal length of every lens can be easily varied by changing the lens current, the ray paths in a transmission electron microscope can be chosen within a wide range. In particular, it is easily possible to change the illumination from a large area to a small spot mode. Whereas the illumination of a large area is used for conventional imaging, the small spot mode allows the investigation of small areas whose size may well approach the 0.1 nm level. By changing the excitations of the projector lenses, it is easily possible to vary the magnification over a large range.

Image Formation in the Transmission Electron Microscope

Figure 3 shows schematically the image formation process for a periodic specimen. The incident electron beam has to be treated quantum mechanically as a wave that is diffracted by the periodic object. Constructing the ray paths in the usual way, one finds that all the rays diffracted in a particular direction are recombined in one point in the back focal plane of the objective lens. Rays emitted parallel to the optic axis are focused in the focal point. In other words, in the back focal plane, the electrons are selected according to the scattering angles or, to put it differently again, one obtains a diffraction pattern in the back focal plane. Therefore, the back focal plane is often called the diffraction plane.

For a general, nonperiodic specimen, the diffraction pattern consists of a continuous distribution rather than of individual spikes. Its intensity distribution is proportional to the squared modulus of the Fourier transform of the electron wave just below the specimen.

In the case of ideal imaging, all rays originating in one point of the specimen will be refocused into one point in the image plane. As in light optics, the image formation in an electron microscope can be viewed as a sequence of two Fourier transforms – the first one between the specimen and the back focal plane and the second one between the back focal and the image plane. Consequently, any manipulation in the back focal plane will influence the final image.

One example is the insertion of a circular aperture which prevents electrons scattered by an angle larger than α_0 from entering into the image. If the first diffraction angle of a periodic specimen is larger than α_0, only the undiffracted beam will enter into the image. In that case, the diffraction amplitude beyond the aperture is identical to the amplitude without any specimen, and thus one obtains an image without any structure. This example demonstrates that the resolution is then limited by the maximum angle passing through the objective aperture. Using Bragg's law

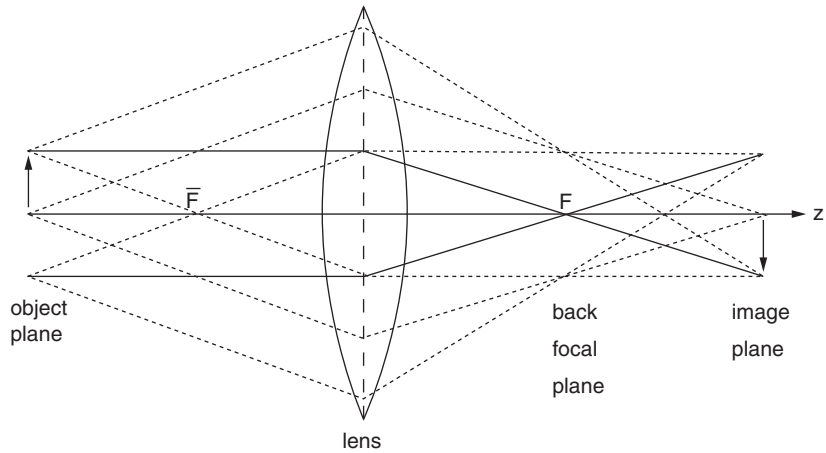
$$2a \sin\theta = n\lambda \tag{5}$$

where a is the lattice spacing and θ the scattering angle for nth order diffraction, one obtains for the lattice resolution limit d

$$d \approx \frac{\lambda}{2 \sin\alpha_0} \tag{6}$$

Contrast Mechanisms in Transmission Electron Microscopy

As the fast incident electrons pass through the thin specimens used for transmission electron microscopy, there is practically no absorption. Therefore, an "ideal" image (without any aperture or any other optical element in the back focal plane)



Transmission Electron Microscopy, Fig. 3 Ray paths showing the imaging by a lens. All rays emitted by the specimen in a given direction are focused to an off-axis point in the back focal plane

does not exhibit any contrast. In order to see anything in the image, the electron wave in the back focal plane has to be influenced in order to discern the structure of the specimen. This can be done in several ways leading to different types of contrast.

Scattering Absorption Contrast

A circular diaphragm, inserted into the back focal plane, intercepts all the electrons scattered at angles larger than $\alpha_0 = r/f$, where r is the radius of the hole in the diaphragm and f the focal length of the objective lens. Strong scatterers, who scatter a substantial fraction of the incident electrons into angles larger than α_0, appear darker in an otherwise bright image. It has to be stressed that no electrons are absorbed in the specimen. The scattering takes place in the specimen, the absorption in the diaphragm in the back focal plane. The contrast is strongest for small aperture angles, because then a large fraction of the scattered electrons will be removed from the beam. Equation 6 shows that for small scattering angles, the resolution deteriorates. Even though this method is very simple, it can only be used for low- and medium-resolution images (coarse structures). For practical applications, a suitable compromise has to be found between contrast and resolution.

Dark-Field Imaging

To avoid the conflict between resolution and contrast, one can use a completely different way to obtain contrast. One major problem of scattering absorption contrast is the fact that in the image one sees only a minor decrease of intensity, which is hard to discern, in an otherwise bright image. Therefore, it is of great interest to visualize strong scatterers as bright spots. To form such an image, the unscattered electrons have to be removed so that only the scattered electrons remain in the image. This can be done by the use of a beam stop in the back focal plane intercepting all unscattered electrons that have left the specimen in a direction parallel to the optic axis. One way of doing this is to mount the beam stop on a thin wire and place it in the diaphragm in the back focal plane. It is difficult, however, to properly adjust the beam stop.

The essential requirement for the dark-field imaging mode is that the unscattered electrons are removed in the back focal plane. As in light microscopy, this can also be achieved by inserting a circular diaphragm in the back focal plane and using a hollow-cone illumination whose angle is larger than the objective aperture angle α_0. In electron microscopy, such an illumination can be most easily obtained by using deflection coils in the condenser system to tilt the incident beam and rotate the direction of incidence by properly addressing the deflection coils.

A dark-field image exhibits bright spots at the positions where scattering is strong on an otherwise dark background. The resolution is determined by the aperture angle α_0. Usually, only a small fraction of the incident electrons is scattered. Therefore, the intensity in dark-field images is rather low.

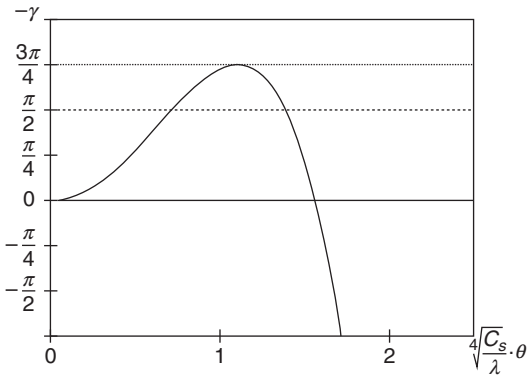
Phase Contrast Imaging

In quantum mechanics, the influence of a thin specimen on the incident electron wave can be described by a spatially varying phase shift ϕ, which is proportional to the projected potential $V(\boldsymbol{\rho}, z)$

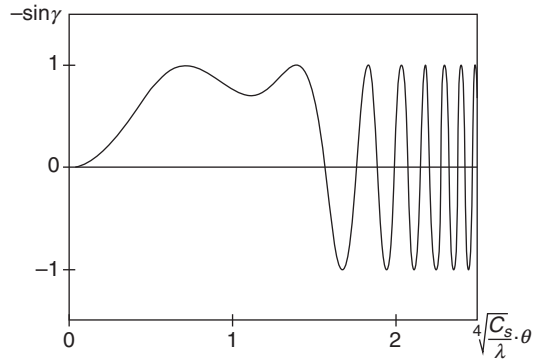
$$\varphi(\boldsymbol{\rho}) = -\frac{i}{\hbar v}\int_{-\infty}^{\infty} V(\boldsymbol{\rho}, z)\, dz$$

where $\boldsymbol{\rho}$ is a two-dimensional vector describing the lateral position in the object plane, v is the speed of the electron, z is the direction of incidence, and $\hbar$ is Planck's constant.

Is it possible to obtain an image whose intensity is proportional to the phase shift? In light optics this can be achieved using Zernike's phase contrast method using a phase plate in the back focal plane to introduce a phase shift of $\pi/2$ between the unscattered and the scattered beam. To achieve this, one has to fabricate a glass plate



Transmission Electron Microscopy, Fig. 4 Phase shift γ introduced by the spherical aberration of the objective lens



Transmission Electron Microscopy, Fig. 5 Transfer function $L(\theta) = -\sin\gamma(\theta)$ for phase contrast imaging

of the proper thickness. How can this effect be utilized in electron microscopy?

As mentioned earlier, electron lenses suffer from severe spherical aberrations that correspond to a phase shift increasing with scattering angle. Including the effect of a small defocus Δf, this phase shift γ is given by

$$\gamma(\theta) = k\left\{\frac{C_s}{4}\theta^4 - \Delta f\frac{\theta^2}{2}\right\}$$

where C_S is the spherical aberration constant, $k = 2\pi/\lambda$ the wave number of the electrons, and θ the scattering angle. By a proper choice of the defocus, one can obtain a phase shift of approximately $\pi/2$ over a large range of scattering angles. This is shown in Fig. 4. To optimize phase contrast, the defocus is chosen so that γ stays close to $\pi/2$ for a large range of scattering angles.

The performance of this method can best be described by the use of a transfer function. Such transfer functions are well known in electronics, where they determine to what extent an electric signal of a given frequency f is transmitted through a circuit. To describe the transmission of a general electric signal, the latter is decomposed into its frequencies – in other words, it is Fourier transformed. The concept of a transfer function can readily be generalized to two-dimensional spatial signals. Due to the circular symmetry of an aligned microscope, the transfer function depends only on the modulus of the spatial frequency θ/λ. For phase contrast imaging, the "phase contrast transfer function" is given by the sine of γ:

$$L(\theta) = -\sin\gamma(\theta)$$

Figure 5 shows its behavior for the optimum defocus, the so-called Scherzer focus:

$$\Delta f \approx 1.2\sqrt{C_s\lambda}$$

The first zero of the phase contrast transfer function determines the resolution limit d, which is given by $d = 0.67\sqrt[4]{C_s\lambda^3}$. It can be shown that spherical aberration is unavoidable for electric–magnetic round lenses, provided they are time independent and that there is no electrostatic charge within the lens. Typical values of the spherical aberration constant C_S are about 1 mm. Inserting $C_S = 1$ mm and assuming an accelerating voltage of 300 kV, corresponding to a wave length $\lambda = 2$ pm, one obtains a resolution limit of $d = 0.2$ nm. One can thus obtain high-resolution images with good contrast.

High-Resolution Electron Microscopy

So far the discussion has been confined to thin objects, where multiple scattering can be neglected. Frequently, high-resolution electron

microscopy is applied to crystals oriented so that the incident electron beam is parallel to a low index axis [5]. Often the resolution is then sufficient to discern atomic columns. In such crystals, however, multiple elastic scattering plays an important role. These processes strongly influence the intensity distribution in the image. It is therefore no longer possible to interpret high-resolution images of crystals directly. Instead, one has to simulate an image of a model crystal, compare the result with the experimental data, modify the model, recalculate, compare, etc., until a good match between the simulated and the experimental image has been achieved.

In recent years it has become possible to determine the amplitude as well as the phase of the electron wave function at the exit surface of the specimen either by a focal-series reconstruction or by electron holography. A comparison of the exit wave function with a corresponding simulation provides an even more rigorous test for the model structure.

Correctors

Due to the spherical aberration of the objective lenses, the resolution limit of electron microscopes has stayed at values well above 0.1 nm for decades. The correction for this aberration has been a challenging task, which has been achieved only very recently by using lenses without rotational symmetry [4]. Nowadays, several correctors using sextupole or quadrupole and octupole lenses are commercially available, some of which do not only correct for the spherical aberration but also for chromatic aberration. In this way, a resolution limit of 0.05 nm can be achieved.

Scanning Transmission Electron Microscopy

It is also possible to obtain images by focusing the beam to a small spot that is scanned in a raster-like manner over the specimen. The number of counts, e.g., of the scattered electrons, is then stored as intensity value for that pixel. The setup of a scanning transmission electron microscope (STEM) is schematically shown in Fig. 6. To obtain an image in a STEM, one commonly uses an annular detector subtending angles outside the cone of the incident rays. This produces a dark-field image; hence, the detector is called an annular dark field (ADF) or, for a large inner angle, a high-angle ADF (HAADF) detector. A major advantage of scanning transmission electron microscopy is the possibility to use other secondary signals to determine the intensity in the corresponding pixel of the image. In particular, the use of an element-specific signal (e.g., X-rays of a particular transition) permits to acquire images showing the distribution of that element within the specimen. Whereas dedicated STEMs are not very widely spread, many modern TEMs have a scanning attachment permitting to perform scanning transmission electron microscopy in addition to the conventional operating modes.

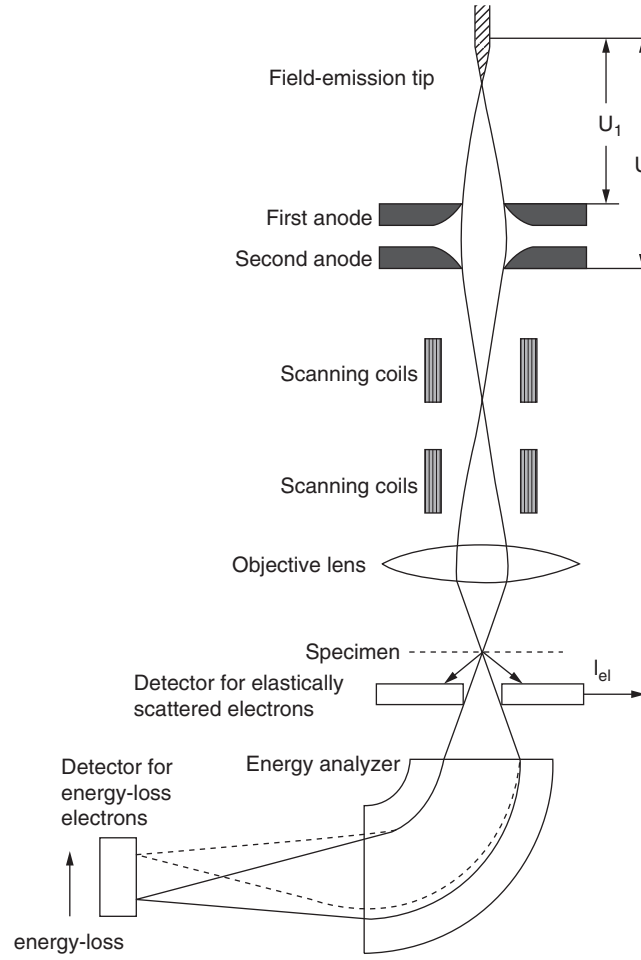
Electron Diffraction

In the previous entries, the various possibilities to obtain an image in an electron microscope have been discussed. By adjusting the intermediate lens, which is the first lens in the projector system, it is easily possible to switch to the diffraction mode, where the back focal plane, i.e., the diffraction pattern, is imaged on the final screen [6, 7]. Essentially in this mode the second and further intermediate image planes and diffraction planes are interchanged as can be seen from Fig. 7.

Using diffraction mode, one can obtain a wealth of information on the structure of a crystallite with any standard TEM. To obtain the position of the diffraction spots, one can use Laue's equation which states that the difference between the wave vector $\mathbf{k}_0$ of the incident and $\mathbf{k}_f$ of the diffracted beam must be equal to a reciprocal lattice vector $\mathbf{g}$:

$$\mathbf{k}_f - \mathbf{k}_0 = \mathbf{g}$$

Every spot in a diffraction pattern corresponds to a reciprocal lattice vector. Electron diffraction thus permits to measure the reciprocal lattice, which in

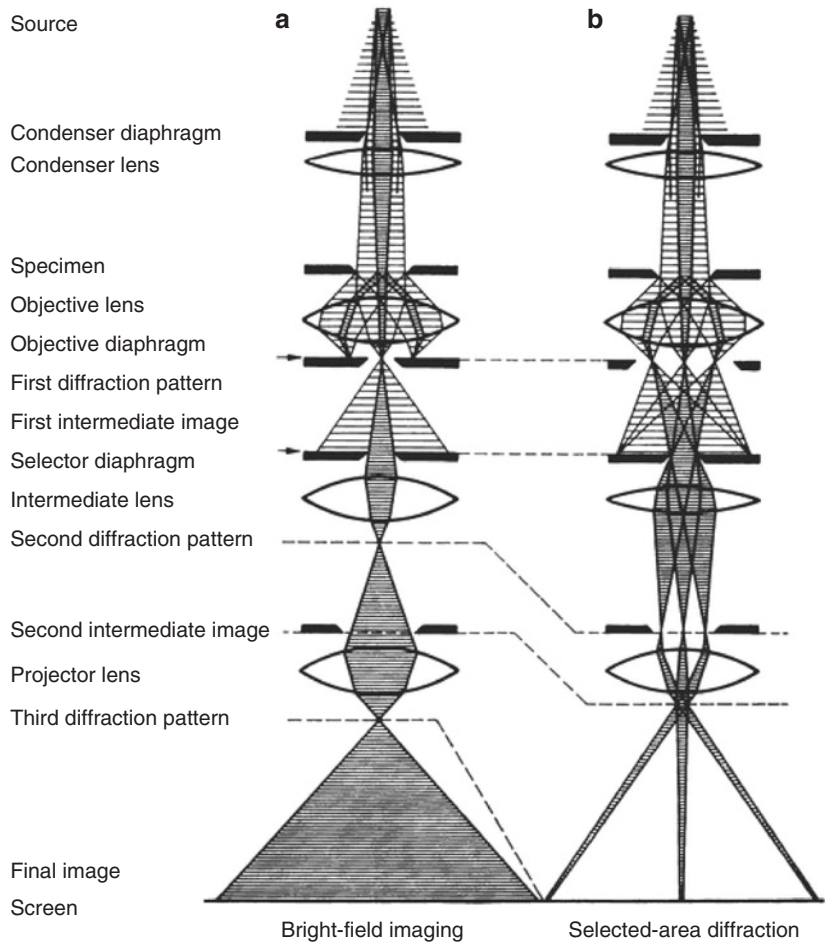


Transmission Electron Microscopy, Fig. 6 Field emission scanning transmission electron microscope equipped with an energy loss spectrometer [1]

turn allows determining the Bravais lattice of the illuminated crystallite. The position of the spots can most easily be visualized using Ewald's construction (Fig. 8). There are two conditions for the appearance of a diffraction spot:

1. The energy of the electrons, and therefore the magnitude of their wave vector, remains unchanged during the scattering process (elastic scattering). This can be visualized by drawing a sphere around the origin of the wave vector of the incident electron with a radius corresponding to its length. The wave vectors of both the incident and the scattered electron must end on the surface of this so-called Ewald sphere.
2. Drawing the wave vector of the incident electron as to end in the origin of the reciprocal lattice, the wave vector of the scattered electron has to end in another reciprocal lattice point.

Putting these two conditions together, one finds that electrons can only be diffracted if the Ewald sphere crosses a reciprocal lattice point. This condition is equally valid for the diffraction of X-rays, neutrons, and electrons. Fast electrons with energies of the order of 100 keV or more have wavelengths of a few pm – much smaller than the lattice constant of a crystal (about 1 nm). Consequently, the length of the wave vector of the electron beam is two or three orders of magnitudes larger than the length of a reciprocal lattice vector.

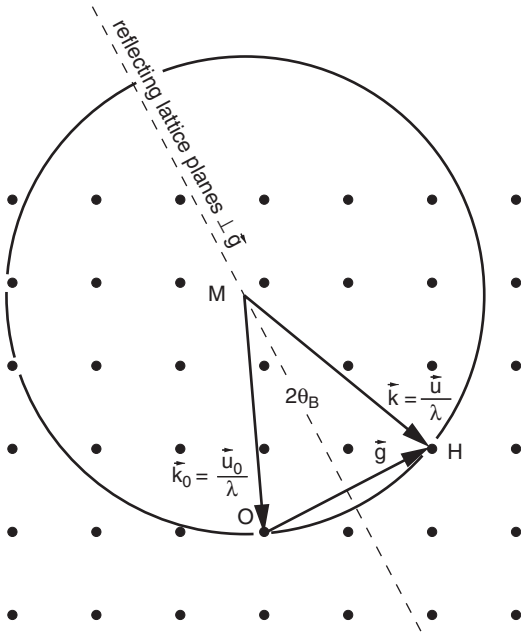


Transmission Electron Microscopy, Fig. 7 Ray diagram for a TEM for (**a**) bright-field imaging and (**b**) diffraction [1]

In the area near the origin of the reciprocal lattice, the Ewald sphere can then be approximated by its tangential plane. Thus, Ewald's construction states that in electron diffraction near the central spot, one observes a planar cut through the reciprocal lattice. The plane through the origin is called the zero-order Laue zone. As an example, Fig. 9 shows the arrangements of spots for a face-centered cubic lattice illuminated along different zone axes. For larger scattering angles, this planar approximation is no longer valid, and the Ewald sphere cuts through reciprocal lattice vectors of planes parallel to the zero order Laue zone, which are denoted as first-order, second-order, etc., Laue zones.

So far electron diffraction has been discussed for using a parallel beam for illumination. This method results in a spot pattern giving information of the illuminated crystallite. If there are many randomly oriented crystallites in the illuminated area, one obtains Debye–Scherrer rings just like in powder diffraction patterns. The determination of the crystal potential, however, is more complicated due to the fact that in electron diffraction, multiple scattering effects cannot be neglected. These depend strongly on the thickness of the crystal and on the exact direction of the incident beam. Instead of varying the incident beam direction with time, one can illuminate the crystal with a range of incident beam directions all at once. This technique is called convergent beam electron diffraction (CBED) (Fig. 10). Instead of the diffraction spots for parallel illumination, one then observes circles that exhibit an internal

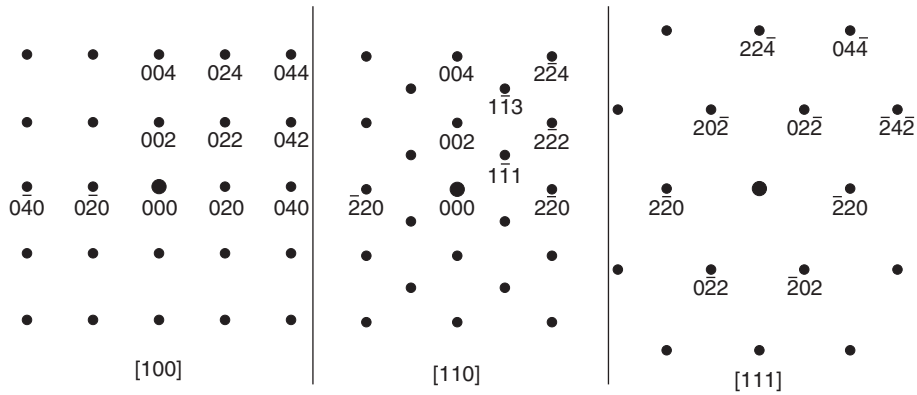
structure. Comparing the intensity distribution within these circles with the results of simulations, one can determine the crystal potential with high accuracy. As long as the illumination angle is smaller than the Bragg angle, the circles do not overlap and each one corresponds to a specific reciprocal lattice vector.



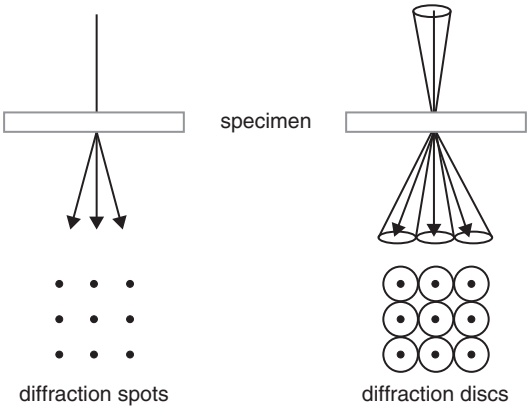
Transmission Electron Microscopy, Fig. 8 Ewald sphere construction to determine the directions of diffracted radiation [1]

Analytical Electron Microscopy

Taking a broader view of the scattering of electrons in a thin specimen, one can see that in addition to the elastic scattering processes discussed up to now, many electrons experience inelastic scattering [1, 2, 6]. In doing so, they lose part of their energy. The various processes are shown schematically in Fig. 11. In order to excite an inner-shell electron into an unoccupied state, the incident electron has to transfer an energy corresponding to at least the binding energy of the original inner-shell electron (Fig. 11a). As these binding energies are characteristic for the element, they can be used as a signature of that particular element. Measuring the energy loss of the scattered electrons thus permits to determine the chemical composition of the illuminated area of the specimen. This is the basis of a technique called EELS for electron energy loss spectroscopy. Shortly after its excitation, the atomic electron decays into its ground state releasing the energy via an X-ray photon (Fig. 11b) or an Auger electron (Fig. 11c). Both of these secondary processes can in principle be used to determine the chemical composition of the specimen. Due to the small mean free path length, the emission of Auger electrons is very surface sensitive. As their detection requires an UHV environment, they are frequently used in surface science, but not in transmission electron microscopy. X-ray



Transmission Electron Microscopy, Fig. 9 Diffraction pattern of a face-centered cubic crystal illuminated along the [100], [110], and [111] zone axes near the central spot [1]

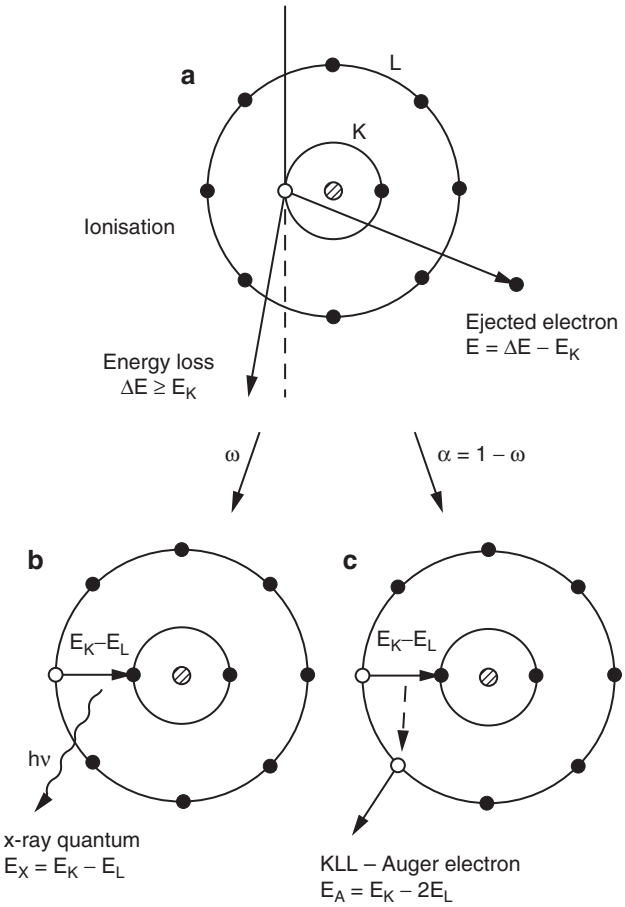


Transmission Electron Microscopy, Fig. 10 Schematic diagram demonstrating the effect of a converging beam on the diffraction pattern (CBED)

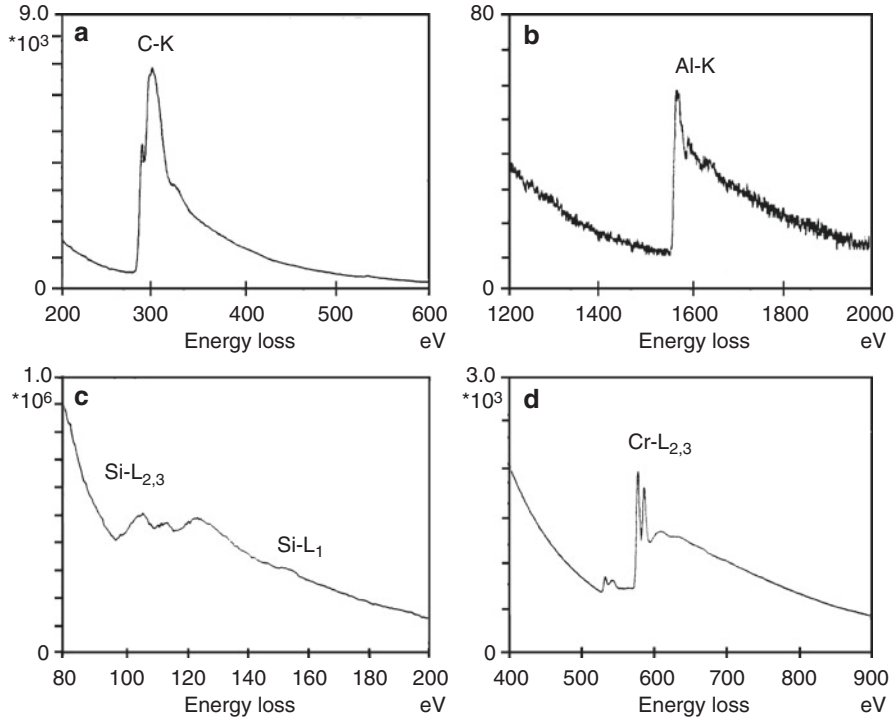
spectroscopy, however, is a very common technique in electron microscopy.

Electron Energy Loss Spectroscopy

Using a monoenergetic incident beam and measuring its energy distribution after it has passed through a thin specimen, one obtains an energy loss spectrum [8]. A few examples are shown in Fig. 12. At characteristic energies, corresponding to the binding energy of an inner shell, the signal shows a marked increase followed by a slow decay. The latter is due to the fact that it is also possible to excite an electron into any state above the Fermi level. The excitation probability, however, decreases for final states with relatively large



Transmission Electron Microscopy, Fig. 11 Schematic representation of (**a**) the ionization process, (**b**) the X-ray emission, and (**c**) the Auger emission [1]



Transmission Electron Microscopy, Fig. 12 Examples of energy loss spectra for the K-edges of (**a**) carbon and (**b**) aluminum and the L23-edges of (**c**) silicon and (**d**) chromium [1]

energies. In order to obtain a quantitative measure of the concentration of an element in the specimen, the background stemming from the tails of edges with lower energies has to be extrapolated. This is most often done by using a power law model for the background intensity I_b as a function of the energy loss E:

$$I_b = AE^{-r} \tag{12}$$

where A and r are free parameters that need to be fitted to experimental data. The signal intensity remaining after background subtraction is proportional to the area density of the element N_A, the intensity of the incident beam I_0, and the excitation cross section σ_A (α_0, ΔE), which depends on the acceptance angle α_0 and the energy loss window ΔE. For two elements A and B, one obtains the two equations:

$$I_A = N_A I_0 \sigma_A(\alpha_0, \Delta E) \tag{13}$$

and

$$I_B = N_B I_0 \sigma_B(\alpha_0, \Delta E) \tag{14}$$

Dividing Eq. 13 by Eq. 14, one obtains

$$\frac{I_A}{I_B} = \frac{N_A \sigma_A(\alpha_0, \Delta E)}{N_B \sigma_B(\alpha_0, \Delta E)} \tag{15}$$

and for the concentration ratio

$$\frac{N_A}{N_B} = \frac{I_A \sigma_B(\alpha_0, \Delta E)}{I_B \sigma_A(\alpha_0, \Delta E)} \tag{16}$$

The concentration can thus easily be determined from the background-corrected intensities above the energy loss edges multiplied by the inverse of the ratio of the cross sections. To measure an energy loss spectrum, it is most convenient to attach a magnetic sector field spectrometer underneath the image plane of a transmission electron microscope. Modern commercial instruments

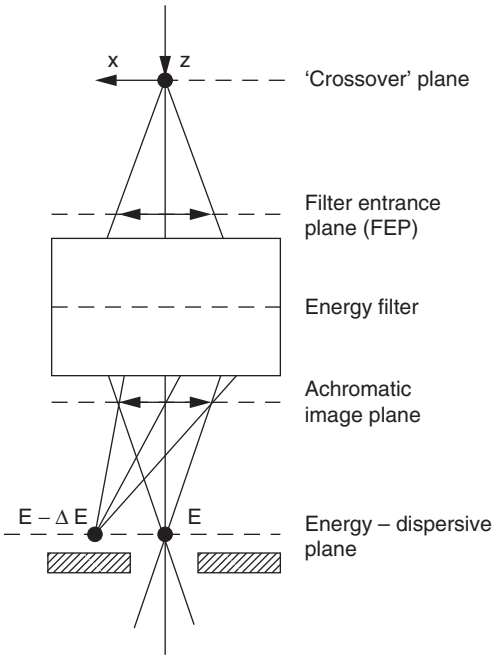
allow measuring a large part of the spectrum in parallel with a diode array.

Using such a spectrometer in a scanning transmission electron microscope, it is possible to obtain the spatial distribution of the elements in the specimen for each pixel. This technique results in images called elemental maps with a resolution at the atomic level in favorable cases.

Alternatively, one can obtain elemental maps by using an imaging energy filter, which turns the TEM into an energy-filtered transmission electron microscope (EFTEM). Such a filter is inserted in the ray path after the first part of the projective system and intercepts all electrons except for those lying within a well-defined energy window. Ideally, such an energy filter should provide an image without any chromatic aberration, but at the same time, it needs an energy dispersion to be able to select the desired energy range. These two conditions are met by a setup which images an intermediate image plane – the filter entrance plane – into an achromatic image plane and images the corresponding diffraction plane – the crossover plane – into the energy-dispersive plane (Fig. 13). In the achromatic image plane, the energy loss is "coded" by a tilt proportional to the energy loss resulting in a spatial displacement in the corresponding diffraction plane. The energy selection takes place in this energy-dispersive plane.

To obtain an elemental map in an EFTEM, one takes a series of at least three images at different energy losses – two below and one above the edge. With the images below the edge, one determines the parameters A and r on a pixel-by-pixel basis; from these, the background intensity is computed and subtracted from the image above the edge (Fig. 14).

The total information that can be obtained via EELS in a STEM or in an EFTEM is conveniently visualized as a data cube with two spatial coordinates and one energy loss coordinate (Fig. 15). In a STEM many energy channels corresponding to one pixel are measured at the same time (Fig. 15a), whereas in an EFTEM one measures the entire picture for one energy channel (Fig. 15b). The efficiency of both methods depends on the number of pixels and energy channels to be acquired.



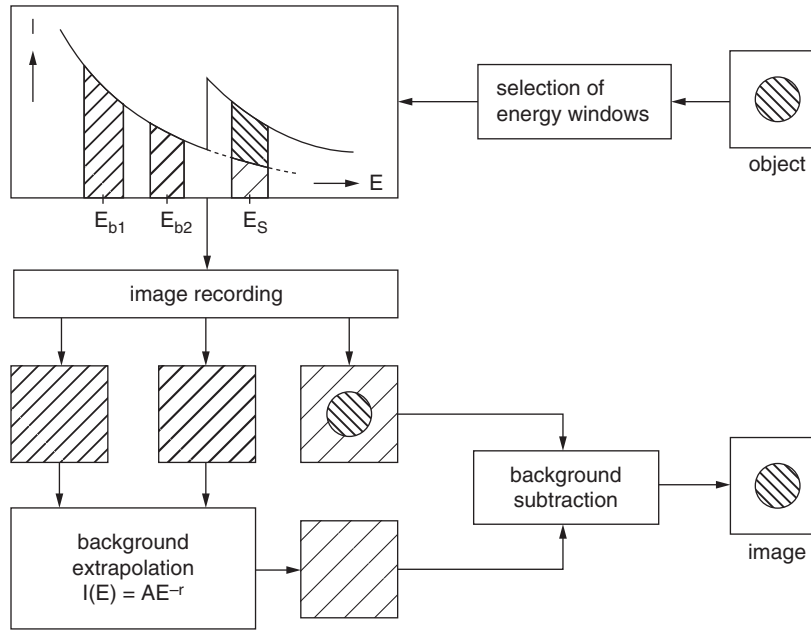
Transmission Electron Microscopy, Fig. 13 Schematic diagram showing the ray paths and the relevant planes in an imaging energy filter [1]

Taking a closer look at a spectrum just above an edge, one finds fine details in the structure that depend on the chemical environment. These details can be used to determine the bonding of the element. The energy loss near-edge structure (ELNES) is similar to the X-ray absorption near-edge structure (XANES).

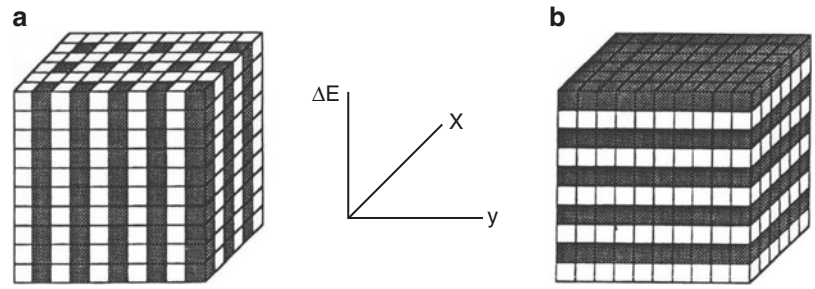
At lower energy losses (<50 eV), EELS can be used to determine the dielectric function of a specimen [8]. Recently, it has been demonstrated that lattice vibrations in the meV regime can be detected by using a spectrometer with an extremely high energy resolution.

Energy-Dispersive X-Ray Spectroscopy

Rather than detecting the inelastically scattered electrons, it is also possible to use secondary processes for the determination of the elemental distribution [6]. This is the basis of X-ray



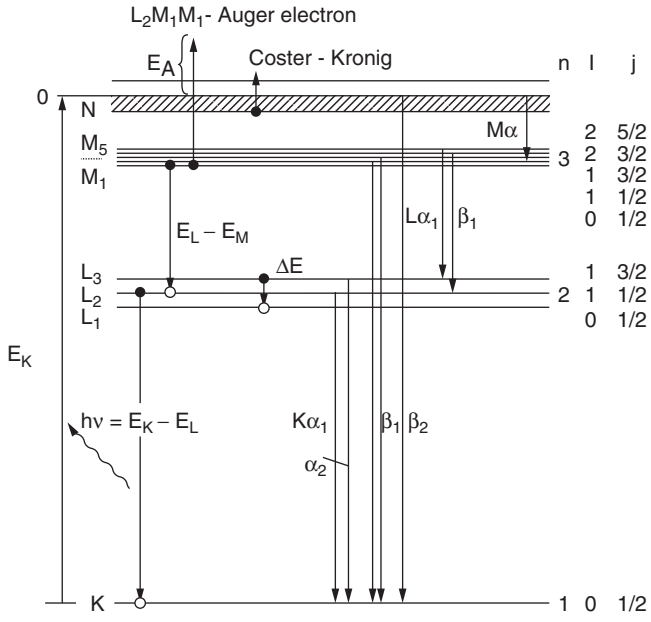
Transmission Electron Microscopy, Fig. 14 The steps necessary to acquire an elemental map



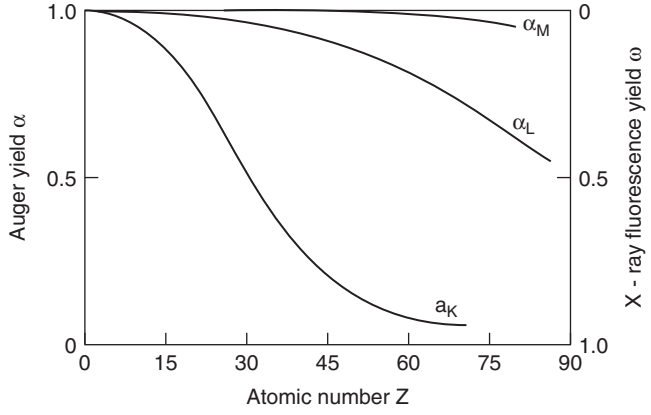
Transmission Electron Microscopy, Fig. 15 The data cube depicting the complete information that can be explored in different ways. (**a**) A STEM acquires spectra point by point. (**b**) In EFTEM, the information is obtained energy slice by energy slice

spectroscopy for elemental analysis. The inner-shell levels of an atom and the corresponding transitions are shown schematically in Fig. 16. The energy difference between the excited and the ground state can be released either by emission of an X-ray or an Auger electron. The probability for the emission of an Auger electron, called Auger yield, and the corresponding fluorescence yield are shown in Fig. 17 for K-, L-, and M-shell transitions. From these graphs one can clearly see that for large atomic numbers the fluorescence yield dominates.

To determine the chemical composition of the specimen, one needs to measure the energy spectrum of the emitted X-rays. This can be done either by using Bragg reflection from a crystal spectrometer to obtain the wavelength or by measuring the number of electron–hole pairs created after absorption of the X-ray photon in a semiconductor detector. Whereas the use of the first method called wavelength-dispersive X-ray spectroscopy (WDX or WDS) is confined to electron microprobes and scanning electron microscopes, the latter method named energy-dispersive X-ray spectroscopy (EDX or EDS) is used in transmission electron microscopes as well as in scanning electron microscopes. An example of an EDX spectrum, taken with a high-purity germanium detector, is shown in Fig. 18. The position of the lines reveals the elements present in the illuminated area of the specimen. To determine the chemical composition, one has to establish a



Transmission Electron Microscopy, Fig. 16 Energy levels of the atomic subshells with their quantum numbers n, l, j and possible transitions to fill a hole in the K-shell including the nomenclature of the emitted X-ray lines [1]



Transmission Electron Microscopy, Fig. 17 Auger electron yield a and fluorescence yield $\omega = 1 - a$ as a function of atomic number [1]

relation between the intensity of the X-ray line and the number of corresponding atoms. Even though there are many factors entering into this relation (angular range subtended by the detector, fluorescence yield, detection efficiency, etc.), it is obvious that the number of counts is proportional to the concentration of the element. For two elements A and B, one can lump all these unknown factors into the two parameters k_A and k_B, thus obtaining for the counts I_A and I_B:

$$I_A = k_A c_A \quad \text{and} \tag{17}$$

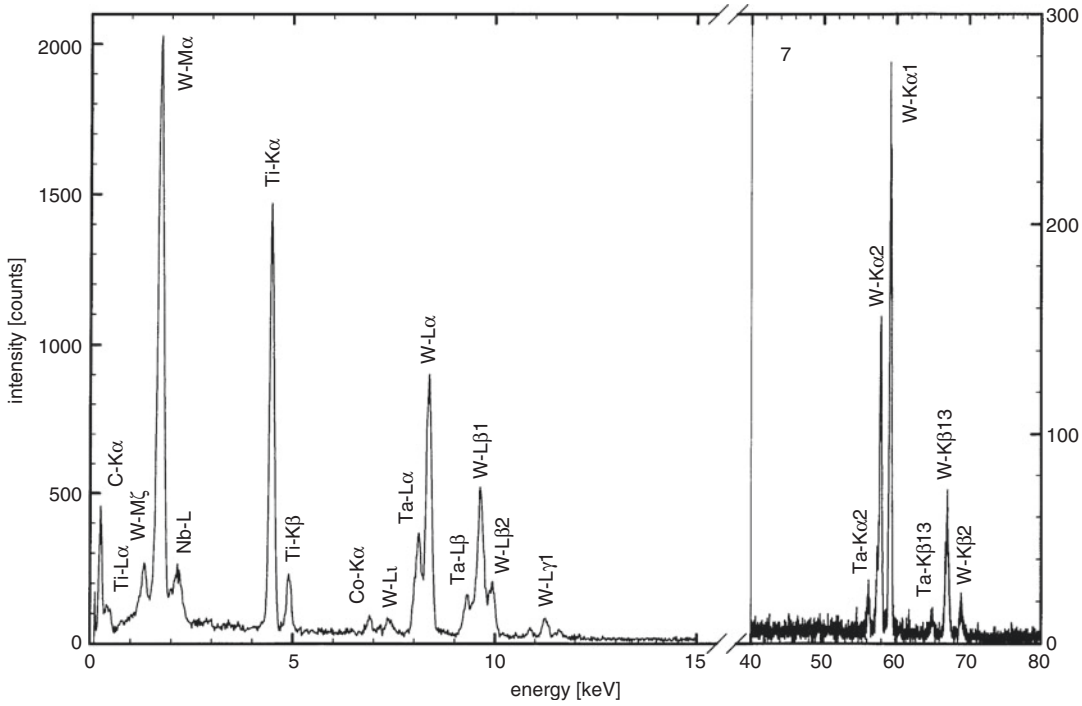
$$I_B = k_B c_B. \tag{18}$$

Dividing Eq. 17 by Eq. 18, one obtains

$$\frac{C_A}{C_B} = k_{AB} \frac{I_A}{I_B} \tag{19}$$

where the "k-factor" $k_{AB} = k_B/k_A$ is a parameter that needs to be determined. This can be done experimentally by the use of well-defined standard specimens. In practice, one can measure the

T



Transmission Electron Microscopy, Fig. 18 Energy-dispersive X-ray spectrum of a hard metal (WC-TaC-TiC-Co) recorded with a high-purity germanium detector [1]

k-factors with respect to one element, e.g., Fe or Si, and use the relation

$$k_{AB} = k_{AC}/k_{BC}.$$

This should be done once for every instrument and can then be used as long as the conditions remain the same.

Comparing the limits of EELS and EDX for the detection of an element, one finds that the dependence of the fluorescence yield on the atomic number favors EELS for light elements. The relatively large background in EELS measurements as compared to an X-ray line indicates that the detection of low concentrations is easier with EDX than with EELS. On the other hand, the fine structure in an EELS spectrum yields additional information about the chemical bonding not accessible in an EDX spectrum. Thus, both methods have their merits and their drawbacks; and the choice strongly depends on the specimen and the information to be gained from it. In favorable cases both methods allow the acquisition of atomically resolved elemental maps.

Cross-References

- ▶ Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
- ▶ Field Electron Emission from Nanomaterials
- ▶ Scanning Electron Microscopy

References

1. Reimer, L., Kohl, H.: Transmission Electron Microscopy. Springer, New York (2008)
2. Williams, D., Carter, B.: Transmission Electron Microscopy. Springer, New York (2009)
3. Hawkes, P., Kasper, E.: Principles of Electron Optics, vol. I & II. Academic, London (1989)
4. Rose, H.: Geometrical Charged-Particle Optics. Springer, Berlin (2013)
5. Spence, J.: High-Resolution Electron Microscopy. Oxford University Press, Oxford (2013)
6. Loretto, M.H.: Electron Beam Analysis of Materials. Chapman & Hall, London (1994)

7. Morniroli, J.P.: Large-Angle Convergent Beam Electron Diffraction (LACBED). Société Française des Microscopies, Paris (2002)
8. Egerton, R.: Electron Energy-Loss Spectroscopy in the Transmission Electron Microscope. Plenum, New York (2011)

Transport Properties

▶ Tight-Binding Simulations of Nanowires

T-Rays

▶ Terahertz Technology for Nano Applications

Tunable Filter

▶ Tunable RF MEMS Filters: A Review

Tunable RF MEMS Filters: A Review

Feng Lin and Mina Rais-Zadeh
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA

Synonyms

Band pass filters; Bandpass tunable filter; LC filters; RF filters; RF MEMS reconfigurable filter; Tunable filter

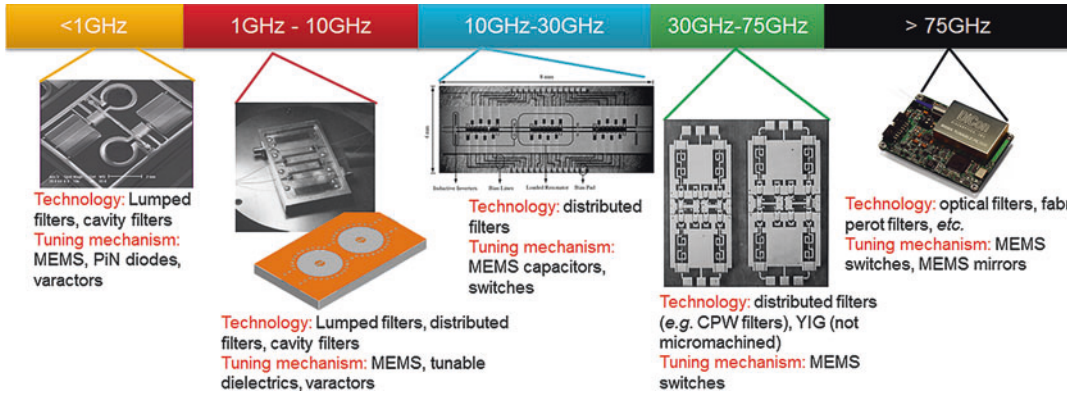
Definition

Tunable RF MEMS filters are defined as microelectromechanical filters for radio frequency (RF) applications that (a) are capable of tuning or reconfiguring and (b) may use microelectromechanical systems (MEMS) switches or capacitors for realizing the tuning, e.g., the frequency or bandwidth.

Introduction

The need for tunable or reconfigurable filters is becoming more apparent as the radio technologies advance and the radio spectrum becomes more crowded. Bandpass filters, specifically, can improve the radio performance by improving the signal to noise ratio, avoiding any saturation due to out-of-band noise contributions [1]. Low-loss tunable filters can replace several surface acoustic wave (SAW) filters currently used in the radios operating in the ultra high frequency (UHF) band, defined as frequencies between 300 MHz and 3 GHz by the International Telecommunication Union (ITU). The UHF band is also extensively used for satellite communication and broadcasting, in paging systems, and by fourth-generation (4G) wireless and long-term evolution (LTE) services. To support all these applications, advances in radio technologies demand reconfigurable RF front-end filters that can support different frequencies and communication standards [2]. Systems employing tunable filters can offer smaller size at a lower cost compared to conventional radios employing multiple single-frequency filters. Some of the important performance specifications of tunable filters are insertion loss, tuning range, tuning speed, power consumption, size, power handling limit, linearity, and cost. Designing a tunable filter that meets the application specified criteria in all these aspects is very challenging, leading to extensive research that has been carried out for years on developing tunable bandpass filters.

Based on their resonator type, tunable filters can be classified as cavity, planar transmission line, or lumped *LC* filters (see Fig. 1). Among them, cavity filters can offer the highest unloaded quality factor (Q_u) resulting in low insertion loss, especially for narrow-band application. But they suffer from large volume, complicated fabrication, and difficulty of integration with other system components. The planar transmission line (microstrip, coplanar waveguide or coplanar



Tunable RF MEMS Filters: A Review, Fig. 1 Tunable filter architectures at different frequency ranges of operation. Common among these schemes is the MEMS technology used as the tuning mechanisms. Figures are reprinted from [3–8]

stripline) filters have moderate Q_u and can be readily integrated with other components and circuits. In the sub-gigahertz frequency regime of UHF band, lumped *LC* filters offer the smallest size compared to the two aforementioned alternative implementations. However, the realization of integrated *LC* tunable filters remains a great challenge at low UHF frequencies, as in this range, the ohmic loss is dominant and the quality factor (Q) of the integrated passives is very low (e.g. inductor Q is less than 30). The key challenge in designing lumped *LC* filters is to implement high-Q inductors and capacitors.

The tuning elements in all types of filters can be implemented electronically – by semiconductor varactors or switches, RF microelectromechanical systems (MEMS) switches/varactors, variable dielectric capacitors, magnetically – using yttrium-iron-garnet (YIG), or thermally – using the more recently developed phase change RF switches [9]. In either case, the most common type of tuning element is to employ either switches or variable capacitors [1]. The semiconductor varactors (PIN diodes or gallium arsenide (GaAs) Schottky diodes), which feature low price, low tuning voltage, and high tuning range, have been demonstrated in many tunable planar filters [10–12]. However, the major drawbacks of semiconductor varactors are their low Q of less than 150 and non-linear performance at high RF power levels. Ferroelectrics with voltage tunable permittivity can also be used to implement variable capacitors. The advantage of this technique is the low voltage requirement to achieve significant capacitive tuning [13]. The main challenge, on the other hand, lies in the deposition process. Most ferroelectrics, such as barium strontium titanate (BST), are sputter deposited on small two inch wafers. The uniformity and quality of the deposited film suffer significantly by going to larger size wafers. Also, these films can only be sputtered in high quality on specific metals and substrates, such as platinum on sapphire [14]. RF MEMS technology is a promising candidate for providing tuning elements. Because of their high Qs, RF MEMS varactors and switched capacitors have been utilized to design tunable filters with excellent tuning performances. In addition, RF MEMS technology can be used to implement monolithically integrated high-Q capacitors and inductors, meeting most – if not all – the requirements of wireless systems. There exist several implementations of tunable filters based on MEMS tuning elements (Fig. 1). In this paper, we present a brief review of classical tunable filters designed in the UHF or higher frequency C band (4–8 GHz) or X band (7–11.2 GHz) with emphasis on the UHF filters. In providing the review, we also compare the advantages and challenges of each technology.

Tunable Filter Technologies

Tunable Filters Using Cavity Resonator

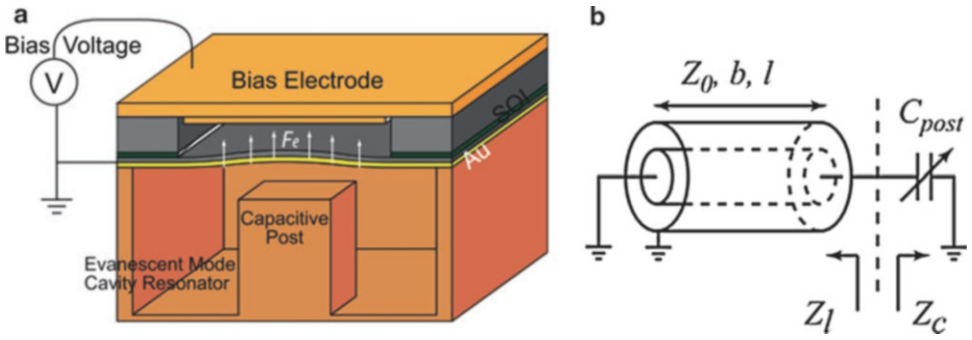
Three dimensional (3D) cavity resonators, including evanescent-mode (EVA) resonators, dielectric resonators (DR), and substrate integrated waveguide (SIW) resonators, are mostly used to design narrow-band tunable filter. The cavity resonators exhibit high Q_u of more than 1000; however, to be tunable, they require complicated assembly of a tuning element (e.g. a semiconductor or MEMS varactor) inside the cavity, which results in significant degradation of their Q_u.

The EVA resonator is composed of a rectangular or circular cavity with an electrically shorted center post. Fig. 2a, b show a schematic of an evanescent-mode cavity resonator and its equivalent circuit [15]. The gap between the post and waveguide wall represents a shunt capacitance, C_{post}. Due to this loaded capacitor, the size of the EVA resonator can be largely reduced (at the cost of lower Q). The resonant frequency can be tuned by changing the value of C_{post}. The most common tuning method is to tune the gap between the metal post and the cavity ceiling (or cap). A cavity filter can be implemented by coupling two or more cavity resonators. For example, in [15], a second-order filter was designed by coupling two cavity resonators. First, the low-pass Butterworth prototype was designed with elements g_i, i = 0, 1, 2, 3. Given the specifications including the frequency tuning range and the 3-dB fractional bandwidths (BW_{3dB} or Δ), the desired values of external quality factors Q_{ext} and the coupling coefficients k_{12} of the cavity resonators were obtained using Equation (1). The required coupling coefficient k_{12} was obtained by changing the distance between the two capacitive posts and the opening width of the coupling iris. The required Q_{ext} was realized by changing the length of the shorted CPW feed lines in the input and output feed structure.

$$Q_{ext} = \frac{g_1 g_2}{\Delta}, \quad K_{12} = \frac{\Delta}{\sqrt{g_1 g_2}} \tag{1}$$

Using a similar principle, two EVA resonators based filters with MEMS tuners, electrostatic diaphragm actuator and piezoelectric disk actuator, achieved Q_u of 300–650 at 5.0–1.9 GHz [15] and Q_u of 310–225 at 0.8–1.43 GHz [16], respectively.

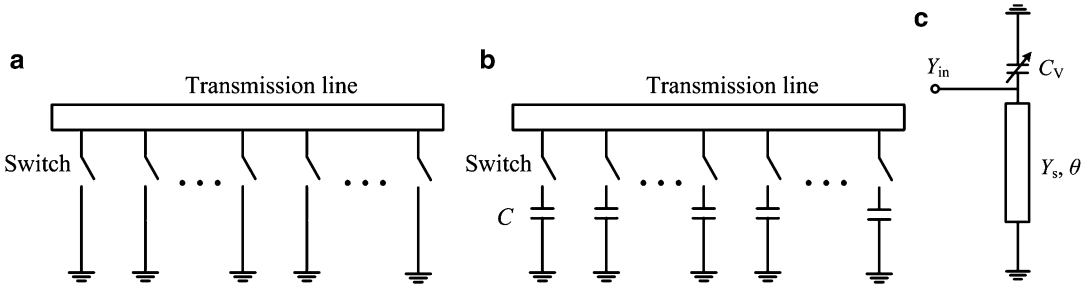
The same tuning technology was also extended to realize tunable helical filters [17], which showed a Q_u of 100–542 at 257–365 MHz. In addition, by inserting the RF MEMS switching network inside the EVA, a tunable filter operating from 4.07 to 5.58 GHz with Q_u of 300–500 was demonstrated in [18]. However, these technologies attain high Q_u at the cost of a complex fabrication process needed for precise assembly of the tuners, and large volume needed to store the EM energy. Also they require high actuation voltages (>100 V) as small gaps (in the order of few micrometer) cannot be easily and precisely defined between the capacitance tuner and the post using current technologies.



Tunable RF MEMS Filters: A Review, Fig. 2 The (**a**) structure, and (**b**) equivalent circuit model of an evanescent-mode (*EVA*) resonator [15]

Tunable RF MEMS Filters: A Review, Table 1 Performance of tunable cavity filters

	Ref. [15]	Ref. [17]	Ref. [19]	Ref. [20]	Ref. [21]
Frequency Range (MHZ)	1900–5000	257–365	500–1100	1200–1600	2393–2503
Technology	EVA+ MEMS	Helical + Piezo	EVA+ varactors	SIW+ MEMS	EVA+ Piezo
Q_u	300–650	100–542	90–214	93–132	374–1300
IIP_3 (dBm)	–	–	17–30	>65	–
Size (cm^3)	3.4	21.4	7.5	29	27
Actuation Voltage (V)	0–140	0–90	0–30	34	90



Tunable RF MEMS Filters: A Review, Fig. 3 Different tuning techniques of transmission line resonators by (**a**) switching physical length, (**b**) switching loading capacitors, and (**c**) loading varactors

To achieve a high Q_u using a simple process, 3D cavities integrated with commercially available surface mount tuning components are alternative solutions. In [19], the center metal post in the EVA resonator is extended to be on the same plane as the ceiling/cap, where solid-state varactors are placed across the gap between the post and the cap. This tunable resonator has a 0.5–1.1 GHz tuning range with Q_u of 90–214. In [20], packaged RF MEMS switches mounted on a SIW are used to achieve Q_u of 132–93 at 1.6–1.2 GHz. RF MEMS capacitor banks mounted on a combline cavity resonator resulted in Q_u of 1,300–374 at 2.50–2.39 GHz [21]. Table 1 summarizes the performance of RF tunable cavity filters, with different tuning technologies.

Tunable Filters Using Transmission Line Resonators

Planar transmission line resonators, such as microstrip, stripline, and coplanar waveguide (CPW) resonators based RF/Microwave filters, are widely used in wireless communication systems due to their compact size (at super high frequencies (SHF)) and low manufacturing cost. The tuning arrangement is mainly implemented by loading the planar resonators with tuning elements (switches or capacitors) to change their effective electrical length. Figure 3 shows the three main tuning methods. The first method is to change the physical length of the resonator by using switches. Switching different loading capacitors can also realize the frequency tuning as done in the second method. However, these two methods suffer from limited number of tuning states. The third approach is to tune the capacitance value in a varactor-loaded resonator. This approach is mostly used in the design of wide-range continuously or digitally tunable filters. The challenging aspect in the design process of such resonators is implementing a wide tuning range varactor. Reducing the length of transmission line can extend the frequency tuning range to some extent, but it also reduces the Q_u of the resonator. Thus, a simple and effective design method for wide tuning planar resonator is an ongoing challenge.

The design procedure for planar filter is similar to the one described earlier for cavity filters. For example in [22], first, the second-order low-pass is design in 0.01-dB Chebyshev prototype

configuration. Given the specifications including the tuning frequency range and the BW_{3dB}, the desired values of Q_{ext} and the coupling coefficients k_{12} of the combline resonators were obtained using Eq. 1. Next, the dimensions of transmission line resonators were determined, according to the required frequency tuning range and the capacitance ratio of the varactors. Finally, the space between the resonators was fine tuned to achieve the needed coupling coefficient k_{12} that satisfied the bandwidth requirement. The shorted coupling line structure was adopted to satisfy the requirement of Q_{ext} over the tuning range.

The tuning elements in planar resonators have been implemented by semiconductor varactors, or RF MEMS switches/varactors [10, 11, 22]. The major drawbacks of designs with semiconductor varactors are their low Q_u (<150) and poor linearity. Thus, they are limited to applications such as radar communication systems. On the other hand, high-Q and low-loss RF MEMS varactors and switches have been utilized to design tunable filters with excellent tuning performances, with low loss and high linearity [23, 24]. The major disadvantage of MEMS tunable planar filters is their limited tuning speed.

Tunable Filters Using Lumped LC Resonators

At the lower UHF band (<1 GHz), tunable passive filters have been implemented mostly using lumped LC elements as the wavelength becomes substantially large in this range of frequency. The size of LC filter is more compact, allowing their low-cost and on-chip integration with silicon circuitry. The Q of a lumped LC resonator or a so-called LC tank is directly related to the Q of the lumped components by $1/Q = 1/Q_{ind} + 1/Q_{cap}$. The Q of the integrated passives is mainly limited by the ohmic loss and is in the range of 50–100 [25]. For wide-band filters, high-Q inductors can be soldered to the filter chip and still a reasonable insertion loss can be obtained [26]. However, size reduction will become an issue when using off-chip elements. To achieve both size reduction and high performance, monolithic integration of LC components using MEMS technology on a single substrate is the most promising approach.

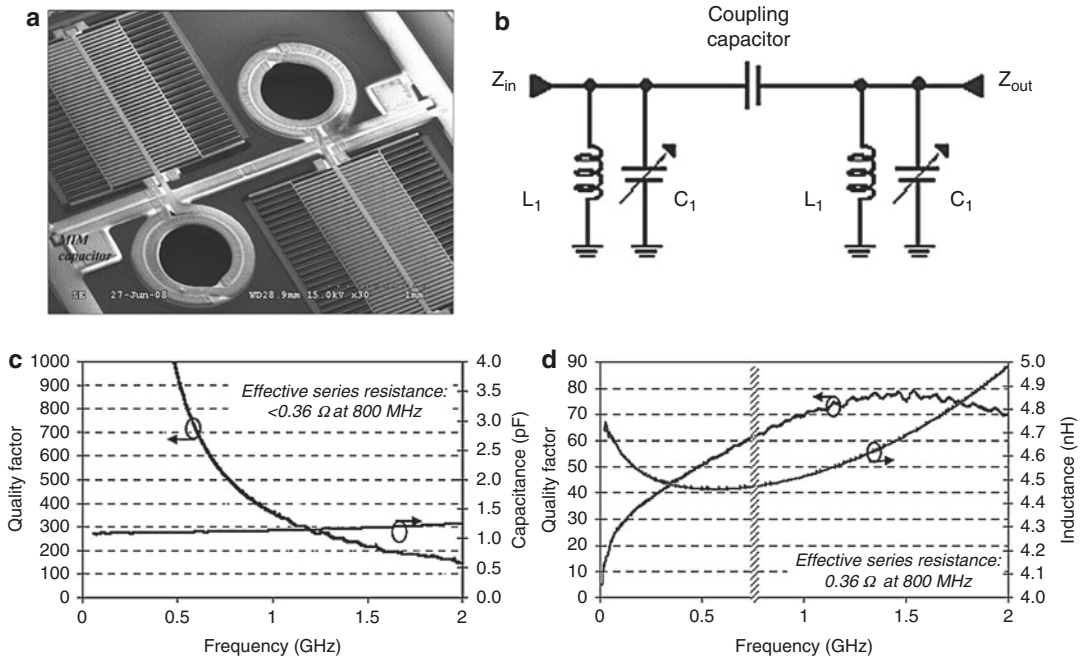
A second-order fully integrated tunable lumped filter on silicon using a low-temperature silver micromachining process is shown in [3] (Fig. 4a, b). The thick silver micromachining technique results in significant reduction of metal loss, which ensures high-Q inductors and tunable capacitors. To achieve frequency tuning, a MEMS varactor was employed in each tank. The two LC tank circuits were coupled using a fixed coupling capacitor, resulting in bandwidth variation as the filter center frequency was tuned.

The measured characteristics of the tunable capacitor and on-chip inductors are shown in Fig. 4c, d. The measured filter has an insertion loss of 4–6 dB with a BW_{3dB} of 5–6 % over the 786–836 MHz tuning range. Using the same technique, an insertion loss of better than 3 dB has been achieved [27] for a second-order filters across 220–640 MHz with unloaded Q as high as 50. The wide tuning range was realized by forming a three-way filter array fully integrated on silicon.

Y. Shim et al. at the University of Michigan developed a continuously tunable three-pole filter on a glass substrate using on-chip inductors and MEMS varactors (Capacitance ratio $C_r = 5.91{:}1$, $C = 220$ fF-1.3 pF) [28]. The filter design procedure was as following: First, the lowpass prototype in Chebychev configuration was designed. The lowpass filter was then converted to a coupled resonator bandpass filter configuration using admittance inverters. With this configuration, values of the lumped components are easily realizable using MEMS surface micromachining technologies. The inductive networks were converted into mutually coupled inductors (Fig. 5a). To achieve frequency tuning, a tunable capacitor bank was employed in each resonator section, which consisted of one fixed capacitor (MIM capacitor), one continuously-tunable capacitor, and three capacitive switches. The miniature filter showed a tuning range of 602–1011 MHz with a fractional bandwidth of 13–14 % and an insertion loss of 3.0–3.6 dB. The measured IIP_3 is greater than 20 dBm. The temperature stability of the center frequency from −50 ° C to 50 ° C is better than 2 %. The measured tuning speed of the filter is better than 80 μs.

A similar filter was packaged using a low-temperature packaging process first described in [29]. The packaged filter showed

T



Tunable RF MEMS Filters: A Review, Fig. 4 (**a**) SEM view of the silver tunable lumped *LC* coupled-resonator filter and (**b**) its equivalent circuit. Measured embedded characteristic of the tunable (**c**) silver capacitor and (**d**) on-chip silver inductors [27]

negligible degradation in performance compared to the unpackaged filter (Fig. 5).

Tunable Filters Based on The Frequency Range

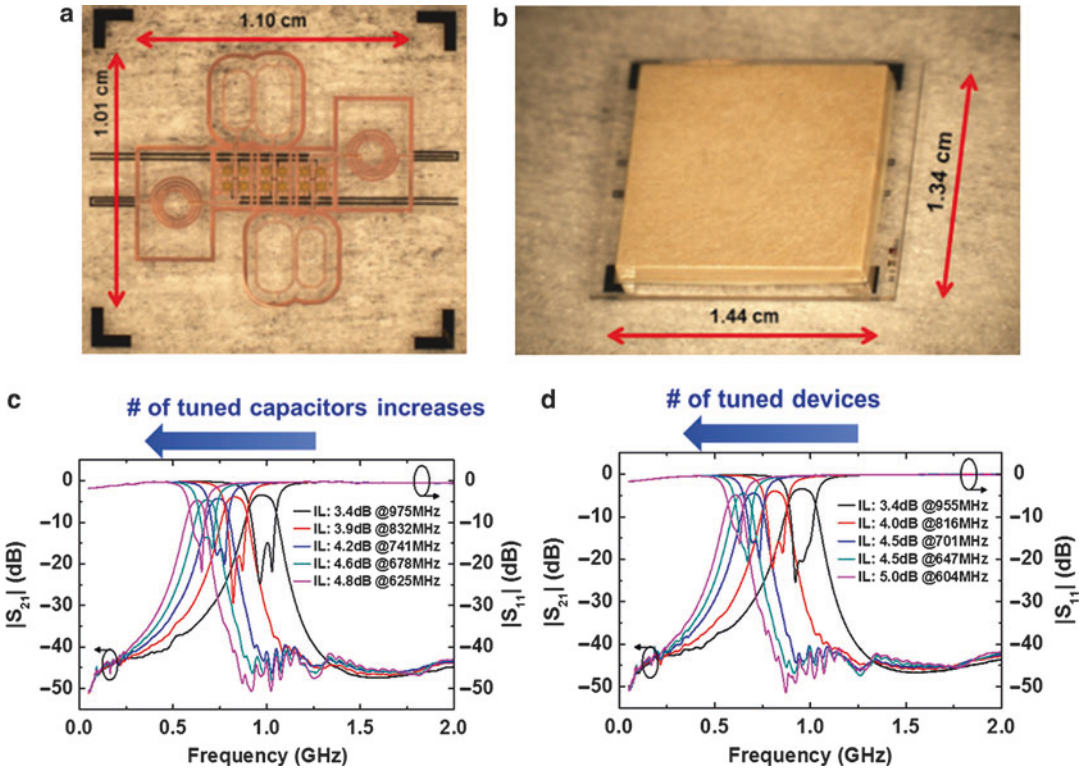
Another way to classify tunable filters is by the frequency range of operation. Most tunable filters are implemented at frequencies above 200 MHz. There have been many reports on MEMS reconfigurable filters centered at 12 GHz and above. Such filters are mostly implemented using distributed filter architectures (Fig. 1) and are not the focus of this paper. In the two subsections below, state of the art tunable filters in UHF and SHF bands are compared.

UHF Filters

Since most cellular communication bands fall in the UHF range, there is a great demand for reconfigurable UHF filters. Table 2 summarizes the performance of reported MEMS tunable filters in the UHF range, built with different technologies. The first two works [4, 30] describe integrated MEMS tunable capacitors on a PCB. Due to the large size of RF circuits on PCB, this type of integration has larger size than complete integration of MEMS technology [18]. The customized multilayer PCB technology [32] can provide much smaller size than other technologies, but the filter implemented using this technique does not show a proper response across the entire tuned spectrum mainly due to the low Q of passive components embedded in the lower PCB layers. The listed filters can be also classified into two categories based on their fractional bandwidths: wideband filters with more than 10 % of BW_{3dB} and narrow-band filters with ~5 % of BW_{3dB}. Monolithic integration of RF components using micromachining technologies is clearly more beneficial for the narrow-band filter implementations, which require passive components with high Q values.

SHF Filters in C and X Bands

Table 3 summarizes the performance of RF tunable filters centered in the C-band and X-band



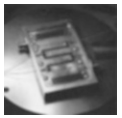
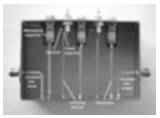
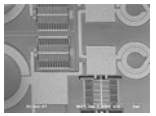
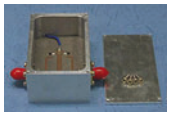
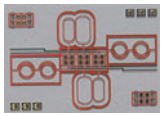
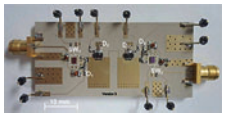
Tunable RF MEMS Filters: A Review, Fig. 5 (**a**) Micrograph of a fully integrated *LC* tunable filter before (**a**) and after (**b**) packaging. Measured S_{21} of the filter at different tuned state (**c**) before and (**d**) after packaging [2]

ranges. References [35–38] describe planar transmission line filters whereas references [16, 18], and [39] implemented tunable cavity filters. Among them, cavity filters offer the lowest insertion loss, especially for narrow-band ($BW_{3dB} < 1$ %) applications, owning to their high Q_u. However, as mentioned earlier, cavity type filters suffer from large volume and complicated fabrication processes. The frequency tuning in these designs are realized using ferroelectric (BST) capacitor [36], magnetic material (YIG) [37], varactor diodes [16], and MEMS varactor [18, 35, 38, 39]. Among them, the MEMS varactors can provide the lowest loss. By inserting the RF MEMS switching network inside the dielectric resonator, a tunable filter tuning from 5.20 to 5.02 GHz with Q_u from 800 to 550 is demonstrated in [39], showing one of the highest unloaded *Q*s for tunable resonators in the SHF range.

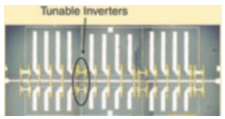
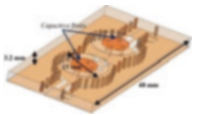
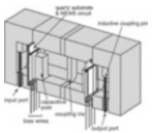
Discussion

Despite advances in microfabrication technologies, design and implementation of high-*Q* reconfigurable filters with fast tuning speed is an ongoing challenge. MEMS components have slower tuning speed compared to other switch technologies, limiting the application of MEMS tunable filters. The main advantage of MEMS filters, besides the small size, that has been the biggest push behind their use is that MEMS filters are very linear. The reliability and packaging of MEMS needs further improvement. MEMS switches with remarkable reliability has been in production for years but a more complex and complete system product using such components does not currently exist and highly needed. New technologies, such as GaN on silicon RF switches [40] and fast switches using phase change materials [41, 42], are now being developed and it is

Tunable RF MEMS Filters: A Review, Table 2 Comparison of reported tunable filters in the UHF range

	Brown'00 [4]	Borwick'03 [30]	S.-Renedo'05 [31]	Lee'09 [32]	Rais-Zadeh'09 [33]	Zhang'10 [11]	Shim'12 [28]	Cho'14 [34]
Picture								
f_c (MHz)	700–1330	225–400	470–862	510–910	786–836	680–1000	602–1011	730–1030
Filter Architecture	5th order discretely tunable combline filter	2nd order filter with off-chip inductors	3rd order continuously tunable combline filter	2nd order filter with off-chip varactor	2nd order continuously tunable lumped filter	2nd order continuously tunable combline filter	3rd order continuously tunable lumped filter	Reconfigurable 2 pole-4 pole filter
Insertion Loss (dB)	2.0–6.0	4.7–6.2	3.0–6.0	1.8–2.5	4.0–6.0	1.1–1.5	3.0–3.6	5.1–7.6
BW_{3dB} (% of f_c)	8–22	4	1–3	20	5–6	8–12	13–14	~10
BW_{30dB}/BW_{3dB}	2.0–3.0	5.0–6.0	3.0–5.0	4.5–6.5	>7.0	5.0–6.0	3.2–4.7	<4
Tuning Speed	N/A	<600 μs	<1 μs	<1 μs	<1 ms	<1 μs	<80 μs	N/A
IIP_3 (dBm)	18–24	30–38	32.8	N/A	N/A	13	>20	20–27
Technology	PCB (microstrip)	PCB (SMT) + MEMS	PCB + CMOS varactor	PCB (multilayer)	MEMS (single chip)	PCB + CMOS varactor	MEMS (single chip)	Duroid + MEMS switches and CMOS varactors
Size (mm^2)	<31 × 41	30.0 × 44.5	<50 × 65	4.4 × 3.4	<5 × 6	<30 × 30	11 × 15	<20 × 40

Tunable RF MEMS Filters: A Review, Table 3 Comparison of reported tunable filters in the C-band and X-band range

	Kraus'04 [35]	Courrèges'09 [36]	Zhu'09 [37]	Joshi'09 [16]	Park'10 [18]	Cheng'11 [38]	Huang'11 [39]
Picture							
f_c (GHz)	6–10	8.35–8.85	8–12	3–5.6	4.07–5.57	3.7–5.95	5.02–5.2
Q	N/A	N/A	N/A	500–750	300–500	115–250	550–800
Filter Architecture	3rd order 2-bits tunable End-coupled filter	4th order microstrip quasi-elliptic filter	Bandpass filter with tunable bandstop filters	2nd order filter with off-chip varactor	2nd order tunable cavity filter	2nd order tunable stripline filter	2nd order tunable dielectric resonator filter
Insertion Loss (dB)	3.3–3.8	3.5–5.7	5.3	2.5–2.8	3.2–4.9	1.9–2.5	<2.8
BW_{3dB} (% of f_c)	15	5.5–7.3	10.1–13.9	0.64–0.75	0.79–1.08	2.5–5.1	0.5–0.57
IIP_3 (dBm)	N/A	N/A	N/A	<36	>40	>35	N/A
Technology	PCB + MEMS varactor	PCB (sapphire) + BST capacitor	PCB + YIG	Evanescent-mode resonators + CMOS varactor	Cavity + MEMS varactor	Suspended Stripline resonator + MEMS varactor	Dielectic resonator + MEMS varactor
Volume (mm^3)	N/A	5.7 × 5.5 × 0.43	<40 × 15 × 0.25	15 × 48 × 3.2	1.5×10^3	>21 × 21 × 2	$2.5 \times 4.5 \times 1.5 \times 10^3$

envisioned that they will be soon integrated with other components to implement tunable RF modules, such as filters. From the cost perspective, implementation of MEMS filters mentioned above requires only a small number of photolithography masks, much smaller than those needed in a standard semiconductor fabrication process. Therefore, improved reliability and yield of MEMS processes should result in reduced cost for tunable RF modules and would make them more attractive [9].

Conclusion

The design challenges for developing tunable MEMS bandpass filters were discussed. For applications in which the size is not the major concern but low insertion losses and narrowband are needed, cavity resonators can be employed. While the transmission line resonators can be employed for applications requiring easy fabrication and convenient integration with other planar components and circuits. For low UHF band applications, lumped *LC* filters offer the smallest size compared to cavity and transmission line resonator implementations.

Cross-References

- ▶ Capacitive MEMS Switches
- ▶ Electrostatic RF MEMS Switches
- ▶ MEMS Fabricated Microwave Devices
- ▶ RF MEMS Switches

References

1. Lucyszyn, S.: Review of radio frequency microelectromechanical systems technology. IEE Proc.-Sci. Meas. Technol. **151**(2), 93–103 (2004)
2. Shim, Y.: Fully integrated high-performance MEMS lumped element filters for reconfigurable radios. Ph.-D. thesis, 2013, University of Michigan (2013)
3. Rais-Zadeh, M., Lavasani, H.M., Kapoor, A., Ayazi, F.: An integrated 800-MHz coupled-resonator tunable bandpass filter in silver with a constant bandwidth. J. Microelectromech. Syst. **18**(4), 942–949 (2009)
4. Brown, A.R., Rebeiz, G.M.: A varactor-tuned RF filter. IEEE Trans. Microwave Theory Tech. **48**(7), 1157–1160 (2000)
5. Sigmarsson, H. H. et al.: In-situ control of tunable evanescent-mode cavity filters using differential mode monitoring. In: IEEE MTT-S International Microwave Symposium, Boston, MA, pp. 633–636 (2009)
6. Kim, H.-T., Park, J.-H., Kim, Y.-K., Kwon, Y.: Low-loss and compact V-band MEMS-based analog tunable bandpass filters. IEEE Microw. Wireless Compon. Lett. **12**(11), 432–434 (2002)
7. "MEMS Tunable Filter (Type B)," DiCon [online]. http://www.diconfiberoptics.com/products/?prod=0172&menu=ftr&sub=0. Accessed 7 Oct 2014
8. Entesari, K., Rebeiz, G.: A 12–18-GHz three-pole RF MEMS tunable filter. IEEE Trans. Microw. Theory Tech. **53**(8), 2566–2571 (2005)
9. Rais-Zadeh, M., Fox, J., Wentzloff, D., Gianchandani, Y.: Reconfigurable radios: a possible solution to reduce entry costs in wireless phones. In: Proceedings of IEEE, submitted, 2014
10. Park, S.-J., Rebeiz, G.M.: Low-loss two-pole tunable filters with three different predefined bandwidth characteristics. IEEE Trans. Microwave Theory Tech. **56**(5), 1137–1148 (2008)
11. Zhang, X.Y., Xue, Q., Chan, C.H., Hu, B.-J.: Low-loss frequency-agile bandpass filters with controllable bandwidth and suppressed second harmonic. IEEE Trans. Microwave Theory Tech. **58**(6), 1557–1564 (2010)
12. Wang, X.-G., Cho, Y.-H., Yun, S.-W.: A tunable combline bandpass filter loaded with series resonator. IEEE Trans. Microwave Theory Tech. **60**(6), 1569–1576 (2012)
13. Tombak, A., et al.: Voltage-controlled RF filters employing thin-film barium-strontium-titanate tunable capacitors. IEEE Trans. Microwave Theory Tech. **51**(2), 462–467 (2003)
14. York, R.: Tunable dielectrics for RF circuits. In: Steer, M. (ed.) Chapter 6 of Multifunctional Adaptive Microwave Circuits and Systems. Raleigh, NC, Wiley (2009)
15. Liu, X., Katehi, L.P.B., Chappell, W.J., Peroulis, D.: High-tunable microwave cavity resonators and filters using SOI-based RF MEMS tuners. J. Microelectromech. Syst. **19**(4), 774–784 (2010)
16. Joshi, H., Sigmarsson, H.H., Moon, S., Peroulis, D., Chappell, W.J.: High-Q fully reconfigurable tunable band pass filters. IEEE Trans. Microwave Theory Tech. **57**(12), 3525–3533 (2009)
17. Psychogiou, D., Peroulis, D.: Tunable VHF miniaturized helical filters. IEEE Trans. Microwave Theory Tech. **62**(2), 282–289 (2014)
18. Park, S.J., Reines, I., Patel, C., Rebeiz, G.M.: High-Q RF-MEMS 4–6 GHz tunable evanescent-mode cavity filter. IEEE Trans. Microwave Theory Tech. **58**(2), 381–389 (2010)
19. Anand, A., Small, J., Peroulis, D., Liu, X.: Theory and design of octave tunable filters with lumped tuning elements. IEEE Trans. Microwave Theory Tech. **61**(12), 4353–4364 (2013)

20. Sekar, V., Armendariz, M., Entesari, K.: A 1.2–1.6-GHz substrate-integrated-waveguide RF MEMS tunable filter. IEEE Trans. Microwave Theory Tech. **59**(4), 866–876 (2011)
21. Fouladi, S., et al.: High-Q narrowband tunable combline bandpass filters using MEMS capacitor banks and piezomotors. IEEE Trans. Microwave Theory Tech. **61**(1), 393–402 (2013)
22. Xu-Guang, W., Young-Ho, C., Sang-Won, Y.: A tunable combline bandpass filter loaded with series resonator. IEEE Trans. Microwave Theory Tech. **60**(6), 1569–1576 (2012)
23. El-Tanani, M.A., Rebeiz, G.M.: High performance 1.5–2.5 GHz RF MEMS tunable filters for wireless applications. IEEE Trans. Microwave Theory Tech. **58**(6), 1629–1637 (2010)
24. Reines, I., Brown, A., El-Tanani, M., Grichener, A., Rebeiz, G.: 1.6–2.4 GHz RF MEMS tunable 3-pole suspended combline filter. In: IEEE MTT-S International Microwave Symposium, Atlanta, GA, pp. 133–136 (2008)
25. Rais-Zadeh, M.: Design and fabrication considerations in developing high-Q MEMS capacitors and inductors. In: IEEE 11th Topical Meeting on Silicon Monolithic Integrated Circuits in RF Systems (SiRF'11), Phoenix, pp. 161–164 (2011)
26. Palego, C., et al.: A two-pole lumped-element programmable filter with MEMS pseudodigital capacitor banks. IEEE Trans. Microwave Theory Tech. **56**(3), 729–735 (2008)
27. Shim, Y., Tabrizian, R., Ayazi, F., Rais-Zadeh, M.: Low-loss MEMS filters with improved out-of-band rejection by exploiting inductive parasitics. In: IEEE International Electron Device Meeting (IEDM'09), Baltimore, pp. 1–4 (2009)
28. Shim, Y., Wu, Z., Rais-Zadeh, M.: A high-performance continuously tunable MEMS bandpass filter at 1 GHz. IEEE Trans. Microwave Theory Tech. **60**(8), 2439–2447 (2012)
29. Ozkeskin, F.M., Choi, S., Sarabandi, K., Gianchandani, Y.B.: Batch-fabricated high-power RF microrelays with direct on-PCB packages. J. Microelectromech. Syst. **21**(4), 990–1001 (2012)
30. Borwick, R.L., Stupar, P.A., DeNatale, J.F., Anderson, R., Erlandson, R.: Variable MEMS capacitors implemented into RF filter systems. IEEE Trans. Microwave Theory Tech. **51**(1), 315–319 (2003)
31. Sanchez-Renedo, M., Gomez-Garcia, R., Alonso, J.I., Briso-Rodriguez, C.: Tunable combline filter with continuous control of center frequency and bandwidth. IEEE Trans. Microwave Theory Tech. **53**(1), 191–199 (2005)
32. Lee, T. C., Park, J. Y.: Compact PCB embedded tunable filter for UHF TV broadcasting. In: IEEE MTT-S International Microwave Symposium, Boston, pp. 505–508 (2009)
33. Rais-Zadeh, M., Ayazi, F.: Small-bandwidth integrated tunable bandpass filters for GSM applications. In: IEEE International Conference on Microelectromechanical Systems (MEMS), Tucson, pp. 1032–1035 (2008)
34. Cho, Y.-H., Rebeiz, G.: 0.73–1.03-GHz tunable bandpass filter with a reconfigurable 2/3/4-pole response. IEEE Trans. Microwave Theory Tech. **62**(2), 290–296 (2014)
35. Kraus, G. M., Goldsmith, C. L., Nordquist, C. D., Dyck, C. W., Finnegan, P. S., Austin, F., IV, Muyshondt, A., Sullivan, C. T.: A widely tunable RF MEMS end-coupled filter. In: IEEE MTT-S International Microwave Symposium, Dallas, pp. 429–432 (2004)
36. Courrèges, S., Li, Y., Zhao, Z., Choi, K., Hunt, A., Papapolymerou, J.: A low-loss X-band quasi-elliptic ferroelectric tunable filter. IEEE Microw. Wireless Compon. Lett. **19**(4), 203–205 (2009)
37. Zhu, Y., Qiu, G., Chi, K.H., Wang, B.T., Tsai, C.S.: A tunable X-band band-pass filter module using YIG/GGG layer on RT/duroid substrate. IEEE Trans. Magn. **45**(10), 4195–4199 (2009)
38. Cheng, C.-C., Rebeiz, G.M.: High- 4–6-GHz suspended stripline RF MEMS tunable filter with bandwidth control. IEEE Trans. Microwave Theory Tech. **59**(10), 2469–2476 (2011)
39. Fengxi, H., Fouladi, S., Mansour, R.R.: High-Q tunable dielectric resonator filters using MEMS technology. IEEE Trans. Microwave Theory Tech. **59**(12), 3401–3409 (2011)
40. Yu, M., Ward, R. J., Hegazi, G. M.: High power RF switch MMICs development in GaN-on-Si HFET technology. In: IEEE Radio Wireless Symposium, Orlando, FL, pp. 855–858 (2008)
41. El-Hinnawy, N., et al.: Low-loss latching microwave switch using thermally pulsed non-volatile chalcogenide phase change materials. Appl. Phys. Lett. **105**, 013501 (2014) (4 pp)
42. Shim, Y., Hummel, G., Rais-Zadeh, M.: RF switches using phase change materials. In: IEEE International Conference on Microelectromechanical Systems (MEMS), Taipei, pp. 237–240 (2013)

Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering

Keqin Yang, Dale Hitchcock, Jian He and Apparao M. Rao
Department of Physics and Astronomy, Clemson University, Clemson, SC, USA

Definition

The spark plasma sintering technique utilizes highly localized intensive Joule heating and

discharging generated by a self-adjusted high-density current flow to alter the micromorphology of carbon nanotubes.

Introduction

Carbon comes in many forms, ranging from *sp*-bonded metastable carbene chains, sp^2-bonded graphite, graphene, single-wall and multiwall nanotubes, fullerenes, and amorphous soot to sp^3-bonded diamond. In particular, carbon nanotubes (CNTs) exhibit unique electrical, magnetic, thermal, chemical, mechanical, and optical properties that stem from their tubular topology and reduced dimensionality. Topologically, a single-wall nanotube (SWNT) can be viewed as one sp^2-bonded graphene sheet rolled into a seamless tube. A sibling of the SWNT is the multiwalled nanotube (MWNT), made of a few to tens of concentric shells with random chiral registries, spaced by $\sim$ 0.34 nm, and coupled through van der Waals interactions as in graphite. In the past few decades, CNTs have been extensively studied from fundamental research and technological points of view. With a nm scale diameter and a large aspect ratio, CNTs are one of the closest material realizations of a 1-D system, offering unprecedented opportunities to study 1-D physics. Furthermore, CNTs are prime candidates for molecular electronics, one of the most promising directions of nanotechnology.

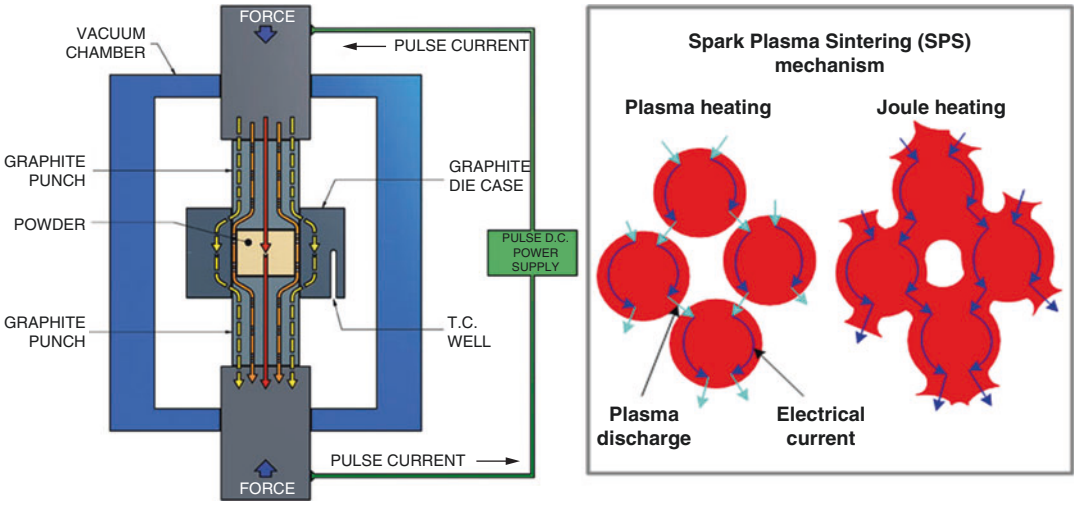
The electrical properties of CNTs are sensitive not only to intrinsic parameters such as the number of shells, tube diameter, and chirality but also to extrinsic parameters such as defects, dopants, and the environment in which they reside [1]. For example, devices based on macroscopic forms of CNTs such as arrays, yarns, and films are gaining intensive interest. In these devices, the primary factor governing the macroscopic electrical transport is the micromorphology, comprised of three basic aspects: (1) the crystal chemistry of an individual CNT such as chirality, defects, and dopants, (2) the topology of the interconnected CNTs, and (3) the density and nature of inter-tube junctions. It is often the case that aspects (2) and (3) matter more than (1) in determining the electrical properties of a macroscopic CNT network. The control of these intrinsic and extrinsic parameters remains challenging and relies on techniques capable of tailoring CNT structures on the atomic, nm, and μm scales.

Tailoring CNT Morphology to Realize New Functions

Many methods have been used to alter the structure of CNTs, from simply hot pressing to more advanced techniques such as electron or ion irradiation, which gives rise to a seemingly endless number of new phenomena and potential functions. In an early attempt to densify CNTs, Ma et al. employed hot pressing at 2,000 °C and 25 MPa in an argon atmosphere to form a "soft sintered" CNT network without altering the tubular morphology of CNTs [2]. In addition, Rubio et al. showed that it is possible to cut and introduce defects into CNTs with very high precision using an STM tip, which has the salient advantage of real-time monitoring of the change of structure [3].

Using Irradiation to Tailor CNT Morphology

More recent research has focused on altering the structure of CNTs using electron or ion irradiation. How would CNTs respond if atoms were removed from the hexagonal lattice continuously [4]? At first glance, irradiating CNTs should lead to a detrimental effect on their crystal structure and thereby properties; however, it has been demonstrated that proper irradiation conditions can have beneficial effects on the properties of CNTs. This is due mainly to the CNT's innate ability to self-heal. The lattice of CNTs can be rapidly rearranged to accommodate the formation of defects and restore a new local coherent structure from a state that is driven far away from equilibrium during the irradiation process. For example, vacancies in SWNTs lead to the formation of pentagons and heptagons, while in MWNTs, they result in inter-shell bonds. This extraordinary innate ability renders a way to tailor the nanostructure of CNTs and facilitates a



Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering, Fig. 1 (*Left*) The schematic of a SPS system and (*right*) two major heating mechanisms in the SPS process

controlled method of self-assembly [5]. The main advantage of using irradiation is to tailor the CNT structure on a nm scale, while the focused electron beam of an aberration-corrected scanning transmission electron microscope can even reach Å scale resolution. The major drawback of this approach is the low throughput of beam-modified CNTs. Lastly, it should be mentioned that when dangling bonds are formed in adjacent tubes, it is often energetically favorable to form a junction between the tubes rather than the reconstruction of individual tubes. The goal of this chapter is to introduce the spark plasma sintering (SPS) process as an alternate approach for controlling inter- and intratubular bonding in CNT networks, at nm and μm scale lengths. In electron beam irradiation, it is known that only a small portion of the impinging electron energy can be transferred to and displace nuclei due to conservation of momentum. As such, a very high electron energy (on the order of 10–100 keV) is needed to alter the structure of CNTs. The SPS process utilizes a low-voltage high-density current and the accompanied localized intensive joule heating or discharging to modify CNTs. The SPS process can control the structure of CNTs in much the same way as irradiation but using macroscopic control rather than the microscopic control of irradiation.

Mechanisms of Spark Plasma Sintering

The SPS technique has been used widely in powder metallurgy, and densification of nanomaterials and nanocomposites [6]. The basic idea dates back to 1930s but was not put into wide use until recently, primarily due to the lack of applications and the high cost of equipment. SPS is a pulsed plasma discharge process that can generate intensive heating (up to 2,000 °C and above) in a short time (few minutes). The voltage is merely on the order of 1–10 V, while the current density is typically on the order 10^2-10^3 A/cm^2 and is highly concentrated at the intergranular contact or interfacial boundaries. The SPS process features a high thermal efficiency because of the direct heating of the sintering mold and sintered materials, and it results in a homogeneously sintered and densified sample because of the "self-adjusting" uniform heating mechanism [7], surface decontamination, and surface activation. The densification mechanism is related to plastic deformation and grain rotation coalescence and sliding, aided by the premelting of grain boundaries. A schematic diagram of the SPS system and the primary heating mechanisms are shown in Fig. 1.

The control parameters of SPS process are macroscopic: they are the sintering pressure, current density, ON/OFF ratio, and sintering time.

It should be noted that heating and current flow are interdependent but are regarded as separate mechanisms in the SPS process. The sintered materials are under a combination of one or more of the following effects [6]: spark plasma, spark impact pressure, Joule heating, and electric field-induced diffusion. In particular, the surface of sintered materials is more easily degassed and chemically activated, promoting strong mass transport at atomic, nano-, and macroscopic levels at a relatively lower temperature and in a shorter time than conventional processes such as hot pressing. Mass transport in SPS is commonly believed to be driven by electric field diffusion. In view of the self-healing ability of CNTs, the current density in the SPS process can be varied to control the deviation from equilibrium, yielding a controllable number of permanent defects. At present, a comprehensive detailed model of the impact of SPS process on CNTs is lacking so that most ongoing work is experimental and exploratory in nature.

Using SPS to Tailor CNT Morphology

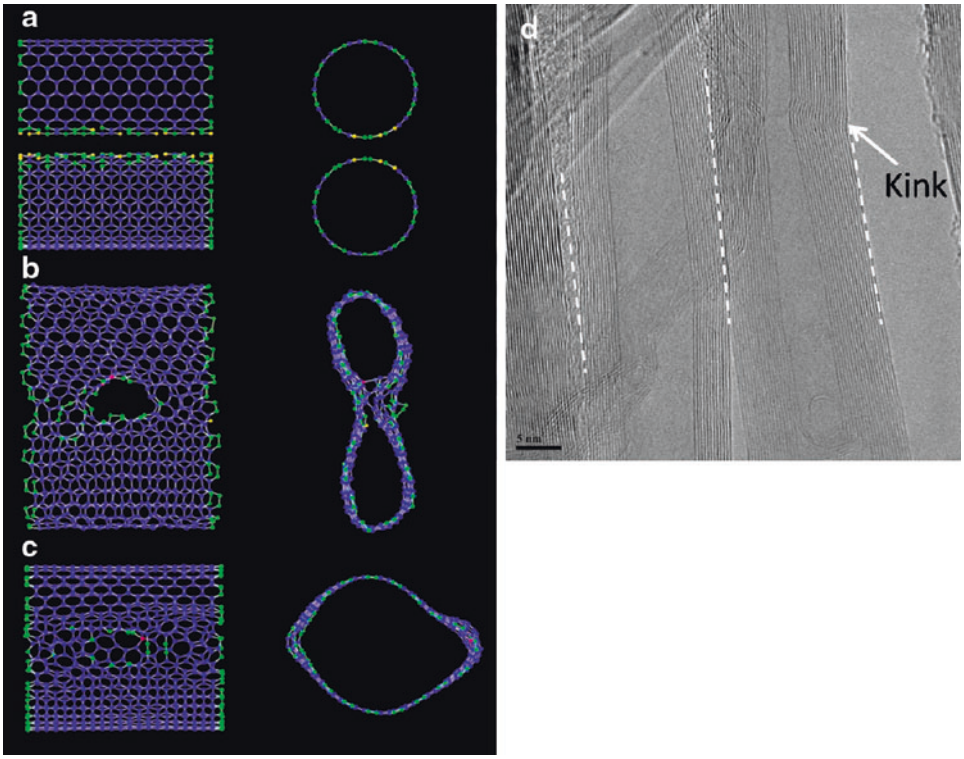
It has been shown that the SPS process may cause the bonding in CNTs to transform from van der Waals in a rope or yarn, to robust sp^2, and then to sp^3 in micro-diamonds [8] or lead to graphitization and crystalline–amorphous heterojunctions in amorphous carbon fibers [9]. Depending on whether CNTs survive the SPS process, the SPS process can be *loosely* categorized into two regimes. The first regime is the mid-power range, where most research is done. In this regime, the tubular morphology is preserved, and there is enough energy to introduce intratubular restructuring (including defects and doping) and create coherent intertubular bonding (i.e., junctions). At present, all experimental evidence suggests that the intra-tube restructuring and inter-tube bonding occur simultaneously during the SPS process. The process of densification also occurs in the mid-power regime. Despite the complexity of inter-tube interactions, the inter-tube spacing and relative orientation of CNTs are the two most important factors which determine the nature of the resulting inter-tube junctions and the packing density. Densification of CNTs is of practical importance in a number of applications such as the electrodes for supercapacitors, biosensors, and artificial bone scaffolds in which one takes advantage of the high surface-to-volume ratio and the outstanding mechanical and electronic properties of CNTs.

The other regime is the high-power regime where the temperature and current density are high enough to destroy tubes, forming other carbon allotropes. The boundary between the mid-power range and high-power range is not clear-cut. Preliminary studies suggest a boundary temperature of 1,000–1,200 °C for MWNTs, while the boundary temperature is lower for SWNTs and depends more on the perfectness of CNT structure and environment.

This section mainly addresses the effects of SPS in the mid-power range, focusing on the SPS-mediated (1) intra-tube restructuring (defects and doping) and (2) formation of inter-tube and inter-shell junctions. These modifications are crucial for controlling the electrical and thermal properties of CNT networks. When a high-density current is pulsed though a CNT assembly, the contact resistance between tubes is much higher than the intrinsic resistance of the tubes, which leads to a highly localized intensive Joule heating or discharging, and the temperature can be sufficiently high to cause rearrangement of atoms/bonds.

Inter-tube and Inter-shell Junctions

It is known that inter-tube junctions dominate the electrical properties of a CNT network. An important objective of ongoing research is to study to what extent and in what aspects the transport properties of a SPSed network are inherited from individual CNTs. Therefore, the manipulation and formation of inter-tube junctions, especially coherent bonding, is of utmost importance in nanoelectronics (molecular electronics) and transparent conductors [10]. A direct impact of the formation of coherent inter-tube bonding is the change of the effective dimensionality of a CNT assembly. In particular, phonons and electrons may have different effective dimensionality in a SPSed CNT network. With coherent inter-tube

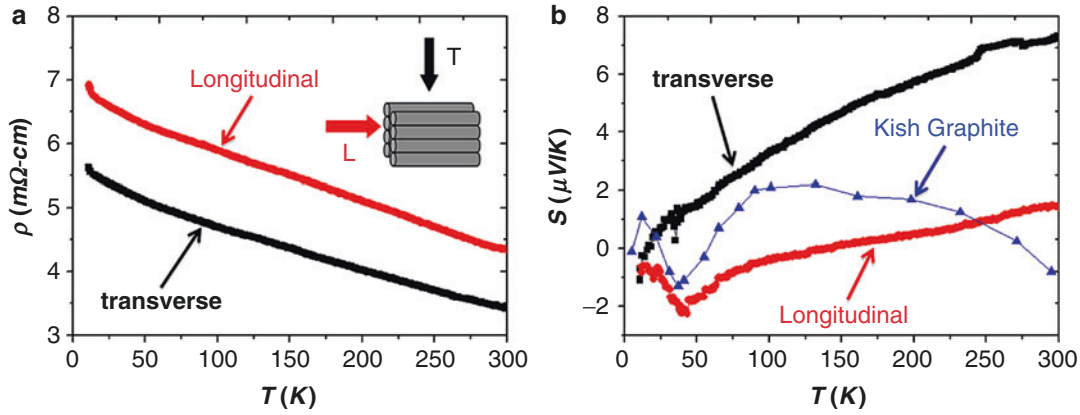


Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering, Fig. 2 (**a**–**c**) Three stages from a theoretical model of the coalescence of two SWNTs. (**d**) Shows that several MWNTs are welded with some degree of corrugations in the sidewalls

junctions, the CNT assembly offers a flexible network architecture, opening up opportunities for several applications.

The modification of inter-tube bonding from van der Waals type to robust sp^2 inter-tube bonding is a nontrivial task, since the topologically complicated arrangement of sp^2 bonds in the inter-tube bonding is not energetically favored. For this reason, high-energy nonequilibrium processes are needed. One example of inter-tube junctions is the coalescence of single- or double-wall CNTs to form a single SWNT of larger diameter [4]. Charlier et al. have proposed a mechanism for the coalescence of tubes using molecular dynamics simulations. The model begins with unzipping of a nanotube to form a carbene chain, and if the metastable carbene chain comes in contact with a neighboring nanotube, it will initiate a reaction to minimize its energy and form a coalesced nanotube. Three steps in Charlier's model are shown in Fig. 2a–c. In spark-plasma-sintered aligned MWNTs, Yang et al. have observed that coherent inter-tube bonding can be formed at a sintering temperature of 1,500 °C for 7 min under 7 MPa axial pressure. The current is applied in the direction perpendicular to the tube direction. As shown in Fig. 2d, the SPS process welds part of two parallel MWNTs together [11].

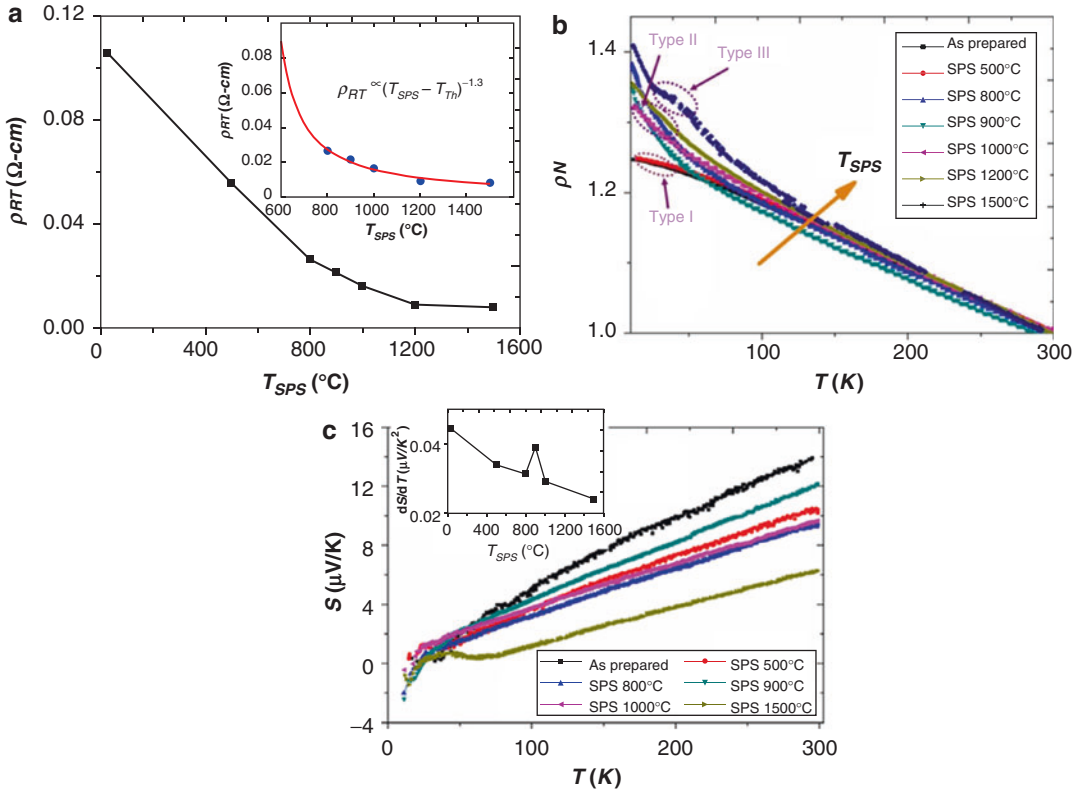
Several interesting properties of SPSed MWNTs have been uncovered through electrical and thermal transport measurements [11]. First, the transverse resistivity (ρ_T) of a low-dimensional electron system such as CNTs is expected to be much higher than the longitudinal resistivity (ρ_L), as is the case in thin films made of aligned CNTs [12]. The MWNTs SPSed at 1,500 °C, however, showed lower resistivity in the transverse direction than the longitudinal direction (Fig. 3a). This surprising observation



Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering, Fig. 3 MWNTs SPSed at 1,500 °C. (**a**) The resistivity is lower in the transverse direction than the longitudinal direction. (**b**) Thermopower in the longitudinal direction shows a strong phonon drag peak below 50 K, which is similar to Kish graphite

may be due to the following reasons: the SPS process creates coherent inter-tube bonds that account for the lower ρ_T, the SPS-induced defects within the tubes may lower ρ_L, and finally lower ρ_T may be due to the larger cross section of the conduction surface in the transverse direction than in the longitudinal direction due to the hollow core and inter-tube spacing of CNTs. Second, although MWNTs are typically metallic, the resistivity in both directions shows nonmetallic behavior ($d\rho/dT < 0$) as in aligned CNT think films, suggesting that inter-tube junctions dominate the overall electrical conductivity of the CNT network. The temperature dependence of the resistivity is also unusual and cannot be satisfactorily interpreted in the frameworks of the Luttinger liquid model, variable-range hopping, or weak localization theory. Nonetheless, the temperature dependence of resistivity is reminiscent of the results of individual highly disordered MWNTs in which the $T^{1/2}$ dependence has been attributed to electron–electron scattering in the presence of random impurities [11]. Third, the thermopower is not only of different magnitudes in the transverse and longitudinal directions but also shows different temperature dependences, with the longitudinal thermopower showing a pronounced phonon drag peak below 50 K very similar to that of Kish graphite (Fig. 3b). In view of electron–phonon coupling, the phonon drag features in the thermopower are due to carriers being preferentially scattered by the phonons in the direction of heat flow, when the mean free paths of both electrons and phonons are long enough. For individual CNTs, the quantized electronic band structure and phonon dispersion relations lead to a reduced phase space for electron–phonon coupling. Hence, the observed phonon drag feature supports the argument that the SPS process leads to an increased effective dimensionality of the CNT assembly.

Subsequently, Yang et al. [13] showed that the electrical connectivity of a network made of randomly oriented MWNTs can be effectively tuned by varying the SPS temperature. An increase in the SPS temperature from 500 °C to 1,600 °C leads to an increase in the packing density from ~ 0.6 to ~ 1.25 g/cm^3, reflecting a change in the inter-tube spacing. The concomitant changes in the magnitude (Fig. 4a) and temperature dependence (Fig. 4b) of resistivity strongly suggest a change in the nature of the inter-tube junctions and thereby the electrical connectivity of the network. As shown in Fig. 4a, the room temperature resistivity as a function of SPS temperature follows a power-law T-dependence, suggesting a percolation behavior. Percolation phenomena are important for the operation of electronic devices based on CNT conducting networks. In Fig. 4b, it is noted that the normalized resistivity $\rho_N(T)$ below 100 K can



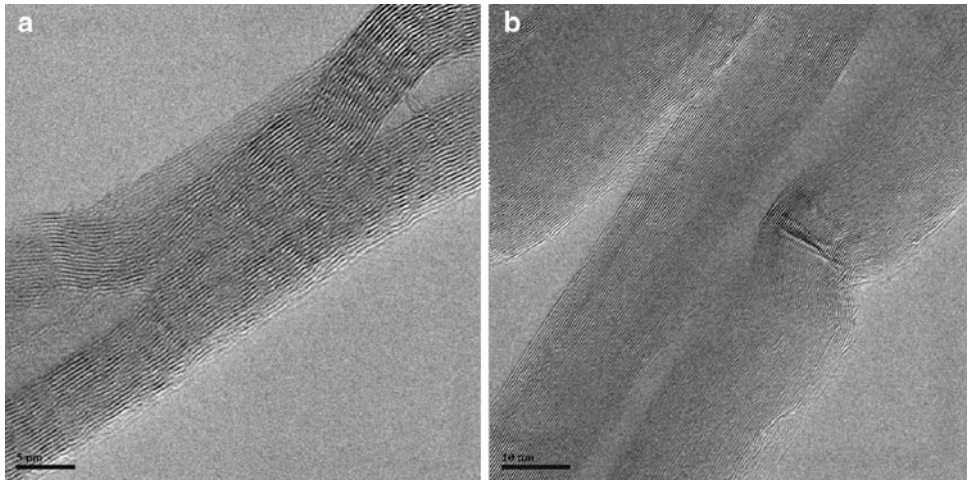
Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering, Fig. 4 (**a**) Room temperature resistivity value as a function of SPS temperature, the inset shows the power-law fitting to a percolative model. (**b**) The crossover in the temperature dependence of normalized resistivity, types I, II, and III, reflects the change of the nature of inter-tube junctions, and the normalization is performed by dividing the $\rho(T)$ by its corresponding room temperature ρ_{RT}. (**c**) The temperature dependence of thermopower of the sample SPSed at 1,500 °C shows a strong phonon drag peak below 50 K [13]

be categorized into three characteristic *T*-dependencies. The as-prepared (AP) sample and the sample SPSed at 500 °C (S500) exhibit a sublinear *T*-dependence, which is labeled as type I. In traces grouped into type II (S800, S1000, and S1200) and type III (S1500) dependencies, there is an upturn below 60 K. This upturn may be associated with the formation of coherent inter-tube junctions. In addition, a plateau is evident below 50 K in sample S1500. The "S" shape anomaly in the thermopower for the S1500 sample (Fig. 4c) resembles the thermopower behavior of Kish graphite (Fig. 3b), as noted above. It is interesting to note that this feature in the thermopower data occurs at the same temperature as a plateau in the resistivity data for the SPS1500 sample, which suggests that there is a single unified mechanism governing both phenomena. The possibility of a Kondo effect has been ruled out [13].

SPS can also induce coherent bonding between CNTs and neighboring graphitic particles. In a CNT bundle where graphitic particles are not bonded to CNTs, Zhang et al. have shown that though the graphitic particles have lower intrinsic resistivity than CNTs, their contribution to the overall conduction will be small [14]. Without bonding to CNTs, the conduction of graphitic particles is restricted to variable-range hopping (VRH) conduction, which will be very small as long as the concentration of graphitic particles is small.

Intra-tube Restructuring

It is also possible to introduce intra-tube defects by the SPS process, including creation of vacancies,



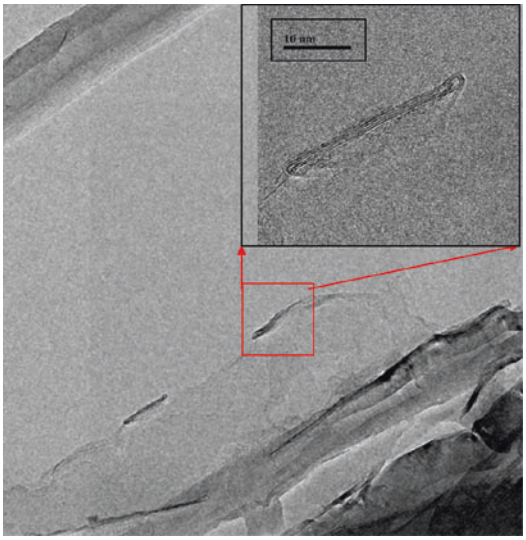
Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering, Fig. 5 The SPS-induced defects in MWNTs (**a**) the collapse of a MWNT (**b**) the kink formed on the wall of a MWNT [11]

collapsing/zipping tubes (Fig. 5a), and kinks on the tube wall (Fig. 5b) [11]. The local curvature change of CNTs often indicates the removal of carbon atoms from the lattice and the subsequent reconstruction of a new coherent structure containing non-hexagonal rings and other types of defects. Defects may profoundly affect the electronic properties of CNTs by creating new localized states near the Fermi level. Electrical transport is also affected in view of the quasi-1-D molecular nature of CNTs. Finally, these SPS-induced crystal lattice imperfections, especially vacancies, serve as a good platform for doping CNTs since the dopant atoms tend to lodge themselves at these energetically favorable positions.

Doping CNTs

Doping is a powerful tool for controlling the electrical properties of CNT networks because it offers a way to vary the carrier concentration and thereby tune the electron–electron and electron–phonon coupling that basically govern the electrical properties of CNTs. It has been shown that through doping the majority carrier conduction in a CNT network can be varied from p-type to n-type in a controllable way [15]. However, doping CNTs is technically challenging for several reasons. First, it is difficult to attain a uniform doping level, due to the finite length effect, the subtle differences in capping topology, the amount and distribution of defects, and the proximity environment. This situation is worsened by the low solubility limit of dopants in carbon allotropes (e.g., the solubility of B in graphite is limited to $\sim$ 4 at.% at 1,200 °C). Second, in small-diameter CNTs, substitutional doping inevitably compromises the integrity of the hexagonal lattice. Third, electric field doping does not work well in MWNTs and thick films. Apparently, controlling crystal chemistry in the synthesis stage is *not* adequate for controlling the doping of CNTs, some post-synthesis treatment is needed. Ion implantation or low-energy ion bombardment has been often used to dope carbon nanostructures [5, 16], but the application is limited by expensive equipment. It is also possible that by varying the doping conditions, the nature of the doping could be varied from low-energy endohedral (filling) or exohedral (intercalation) doping to higher energy substitutional doping. The SPS process is a promising candidate for post-growth doping of CNTs.

Recently, Hitchcock et al. have shown that a post-growth treatment of SWNT bundles with boric acid solution followed by SPS process is a viable method for doping CNTs and controlling their carrier concentration. Using boric acid concentration as the control parameter, they were able



Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering, Fig. 6 Graphitic nanostructures formed in SPS by unzipping MWNTs [11]

to alter the carrier concentration of the SPSed CNT networks. As a result, the resistivity, thermopower, and Hall coefficient measurements show a doping-induced shift of the Fermi level and the accompanied changes in electron–phonon coupling [17].

Altering the Morphology of Tubes (Beyond the Soft Sintering)

At high SPS temperatures (>1,500 °C), it is possible to alter the morphology of CNTs. In an extreme case, it is possible to "unzip" CNTs to form graphitic nanostructures (Fig. 6) [18, 19]. This has been proposed as a novel method for generating large amounts of graphene in a short amount of time. Under the proper conditions, SPS can also change the sp^2 bonds of CNTs to sp^3 bonds, forming diamond [8]. Shen et al. [20] have shown that using the SPS, it is possible to grow diamond particles up to 100 μm in diameter at very low pressure (80 MPa). It is noteworthy that SPS under a moderate pressure of 10–100 MPa leads to the conversion of CNTs to diamonds, whereas a relatively higher pressure in the GPa range is usually required for diamond formation by other methods. It is proposed that the nucleation of diamond is from the core of carbon nano-onions, which is an intermediate product from the breakage of carbon–carbon bonds and unzipping CNTs [8]. The nano-onions behave as compression cells ("self-compression") to help nucleate and grow diamonds. Hence, the SPS process provides for a new path for diamond synthesis, which could offer many applications in industry due to the exceptional hardness and thermal conductivity of diamond.

Conclusion

Since their discovery, carbon nanotubes have been a topic of much research due to their amazing physical properties. However, in order to take CNTs from the lab to applications, scalable techniques must be developed to tune their properties. While it is possible to alter the structure and morphology of CNTs with very high precision using an irradiation technique, this method involves microscopic control parameters and the throughput is very low. In this regard, the SPS technique utilizes high-density current without the need to focus or align the electron beam, the control parameters of SPS are macroscopic, and the throughput is very high. One can envisage a combination of these two techniques to give some control of the structure and properties of CNTs at multiple length scales, which is a very promising approach to move CNTs from materials used solely in research to viable materials with real-world applications.

Cross-References

► Carbon Nanotubes (CNTs)

References

1. Saito, R., Dresselhaus, G., Dresselhaus, M.S.: Physical Properties of Carbon Nanotubes. Imperial College Press, London (1998)
2. Ma, R.Z., Xu, C.L., Wei, B.Q., Liang, J., Wu, D.H., Li, D.J.: Electrical conductivity and field emission characteristics of hot-pressed sintered carbon nanotubes. Mater. Res. Bull. **34**, 741–747 (1999)
3. Rubio, A., Apell, S.P., Venema, L.C., Dekker, C.: A mechanism for cutting carbon nanotubes with a scanning tunneling microscope. Eur. J. Phys. B. **17**, 301–308 (2000)

4. Charlier, J.C.: Defects in carbon nanotubes. Acc. Chem. Res. **35**, 1063 (2002)
5. Krasheninnikov, A.V., Banhart, F.: Engineering of nanostructured carbon materials with electron or ion beams. Nat. Mater. **6**, 723 (2007)
6. Munir, Z.A., Anselmi-Tamburini, U., Ohyanagi, M.: The effect of electric field and pressure on the synthesis and consolidation of materials: a review of the spark plasma sintering method. J. Mater. Sci. **41**, 763 (2006)
7. Song, X., Liu, X., Zhang, J.: Neck formation and self-adjusting mechanism of neck growth of conducting powders in spark plasma sintering. J. Am. Ceram. Soc. **89**, 494–500 (2006)
8. Zhang, F., Shen, J., Sun, J., Zhu, Y.-Q., Wang, G., McCartney, G.: Conversion of carbon nanotubes to diamond by spark plasma sintering. Carbon **43**, 1254 (2005)
9. Qi, X., Bao, Q., Li, C.M., Gan, Y., Song, Q., Pan, C., Tang, D.Y.: Spark plasma sintering-fabricated one-dimensional nanoscale "crystalline-amorphous" carbon heterojunction. Appl. Phys. Lett. **92**, 113113 (2008)
10. Gruner, G.: Carbon nanotube films for transparent and plastic electronics. J. Mater. Chem. **16**, 3533 (2006)
11. Yang, K., He, J., Su, Z., Reppert, J.B., Skove, M., Tritt, T.M., Rao, A.M.: Inter-tube bonding, graphene formation and anisotropic transport properties in spark plasma sintered multi-wall carbon nanotube arrays. Carbon **48**, 756 (2010)
12. Hone, J., Llaguno, M.C., Nemes, N.M., Johnson, A.T., Fischer, J.E., Walters, D.A., Casavant, M.J., Schmidt, M., Smalley, R.E.: Electrical and thermal transport properties of magnetically aligned single wall carbon nanotube films. Appl. Phys. Lett. **77**, 666–668 (2000)
13. Yang, K., He, J., Puneet, P., Su, Z., Skove, M.J., Gaillard, J., Tritt, T.M., Rao, A.M.: Tuning electrical and thermal connectivity in multiwall carbon nanotube buckypaper. J. Phys. Condens. Matter **22**, 334215 (2010)
14. Zhang, H.-L., Li, J.-F., Zhang, B.-P., Yao, K.-F., Liu, W.-S., Wang, H.: Electrical and thermal properties of carbon nanotube bulk materials: experimental studies for the 328–958 K temperature range. Phys. Rev. B **75**, 205407 (2007)
15. Duclaux, L.: Review of doping of carbon nanotubes (multiwalled and single-walled). Carbon **40**, 1751–1764 (2002)
16. Bangert, U., Bleloch, A., Gass, M.H., Seepujal, A., van der Berg, J.: Doping of few layered graphene and carbon nanotubes using ion implantation. Phys. Rev. B **81**, 245423 (2010)
17. Hitchcock, D., Yang, K., He, J., Rao, A.M.: Electrical transport properties of single-walled carbon nanotube bundles treated with boric acid. Nano: Brief Rep. Rev. **6**(4), 337–341 (2011)
18. Gutirrez, H.R., Kim, U.J., Kim, J.P., Eklund, P.C.: Thermal conversion of bundled carbon nanotubes into graphitic ribbons. Nano Lett. **5**(11), 2195–2201 (2005)
19. Kim, W.S., Moon, S.Y., Bang, S.Y., Choi, B.G., Ham, H., Sekino, T., Shim, K.B.: Fabrication of graphene layers from multiwalled carbon nanotubes using high dc pulse. Appl. Phys. Lett. **95**, 083103 (2009)
20. Shen, J., Zhang, F.M., Sun, J.F., Zhu, Y.Q., McCartney, D.G.: Spark plasma sintering assisted diamond formation from carbon nanotubes at very low pressure. Nanotechnology **17**, 2187–2191 (2006)

Turfs

▶ Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

2D-Slab Approach

▶ Surface Modeling of Ceramic Biomaterials

248-nm Lithography

▶ DUV Photolithography and Materials

Two-Photon Lithography

▶ Laser Beam Machining

U

uFOM

► Microfluidic Optomechanics

Ultrahigh Vacuum Chemical Vapor Deposition (UHVCVD)

► Chemical Vapor Deposition (CVD)

Ultralarge Strain Elasticity

► Superelasticity and the Shape Memory Effect

Ultra-precision Finishing

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Ultra-precision Machining

► Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Ultraprecision Machining (UPM)

Suhas S. Joshi
Department of Mechanical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

Synonyms

Nanomachining; Nano-mechanical machining

Definition

Ultraprecision machining refers to the ultimate ability of a manufacturing process wherein processing of a material at its lowest scale that is, at the atomic scale, is achieved. It is known that the lattice distances between two atoms are of the order of 0.2–0.4 nm; therefore, the ultraprecision machining refers to processing or removal actions of a manufacturing process in the vicinity of 1 nm. The process is also referred as "atomic bit" processing. To remove or process atomic bits, extremely large energy density is required, which is equivalent to the atomic bonding energy. The conventional cutting tools neither have high strength to sustain high specific cutting energy nor have hardness to sustain the tool wear. Therefore, ultraprecision machining refers to use of single crystal diamond (SCD) tools for ultrafine cutting

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

or very fine abrasives for lapping or polishing. It may also refer to use of high-energy elementary particles like photons, electrons, ions, and reactive atoms to undertake removal [1].

Historical Perspective

In early 1970s, Japanese researcher N. Taniguchi illustrated historical evolution of precision in manufacturing through "Taniguchi curves" [1]. It was realized that it takes about 20 years to improve precision by one decimal point; see Table 1 depicting the progression of precision.

He extrapolated the curves further to postulate that the resolution of machining processes would reach to 1 nm in the year 2000 [1]. True to the expectations, today, a number of ultraprecision machining processes are available.

Applications

Ultraprecision processing is required in the manufacture of high-precision block gages, diamond indenters and tools, 3D metallic mirrors, etc., in the mechanical domain; Si wafers, ICs memory, thin film, ULSI devices in electronics field; and optical flats, diffraction gratings, mirrors, and aspherical lenses in the optical field.

Types of UPM Processes

The ultraprecision machining processes can be broadly classified into three categories [1]:

1. Nanomechanical processing: These processes use either SCD tools to perform diamond turning or very fine abrasives in the bonded or loose form to perform nanogrinding, nanolapping, or nanopolishing. The other processes in this category include progressive mechanochemical polishing for Si wafers, pitch polishing for aspherical lenses.
2. Nanophysical processing: It involves use of high-energy elementary particles like photons, electrons, ions, and reactive atoms to perform direct ablation of substrate or carry out lithography.
3. Nanochemical or electrochemical processing: These processes involve use of chemical or electrochemical principles to effect material removal. They are chemical milling, photochemical machining, and other processes.

In the following sections, principles of nanometric removal in ultraprecision machining processes are elucidated along with their applications.

Nanomechanical Processing

In these processes, the processing unit, defined as the size of chip generated in a single stroke of tool, should be of the order of single or subatomic magnitude.

Principle of Ultraprecision Removal

The removal of atomic bits in nanomechanical processes requires specific cutting energy of varying degree as the scale of the process decreases and cutting tools with certain sharpness and hardness as discussed below:

Size-Effect Implications

As the processing scale reduces from simple tension testing to grinding, where submicrometric chips are generated, the specific shear or breaking energy becomes extremely large. This effect is called as *size effect*. See Table 2 showing processing scales and shear energies [2].

Extending this trend further, when the machining scale reaches an atomic-bit, the specific

Ultraprecision Machining (UPM), Table 1 Evolution of ultraprecision

Year	1900	1920	1940	1960	1980	2000
Resolution (μm)	>10	5	0.5	0.05	0.005	0.001

shear energy reaches atomic bonding energy [1] or the theoretical shear strength of a defect-free material given by

$$\tau_{th} = \frac{G}{2\pi} \quad (1)$$

For carbon steel with the modulus of rigidity, G = 82 GPa, the theoretical shear strength is given by

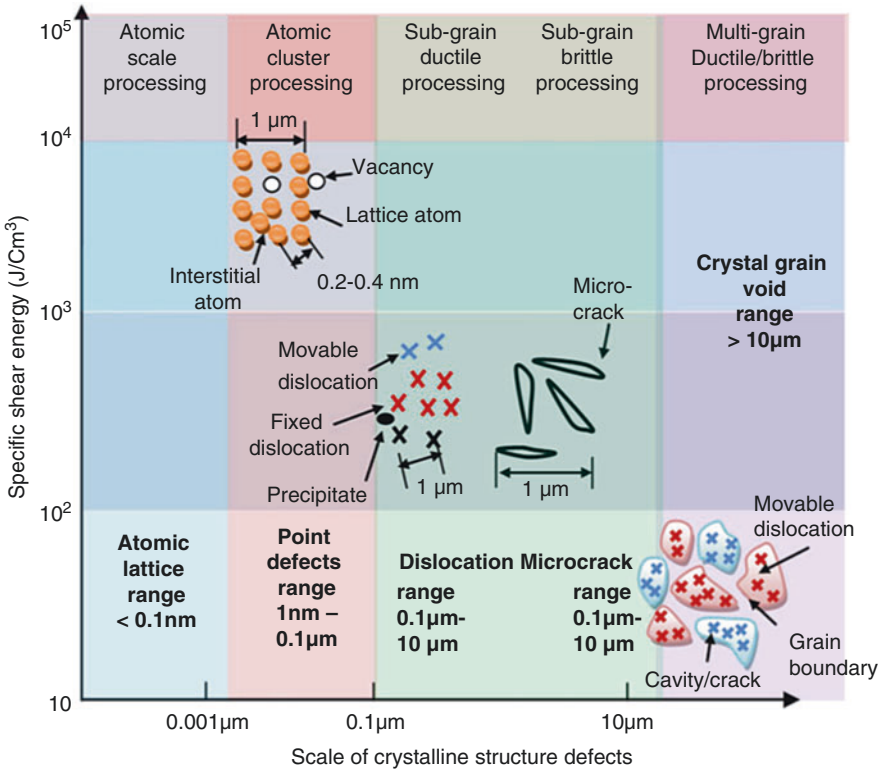
$$\pi_{thCarbonSteel} = 13GPa \quad (2)$$

Ultraprecision Machining (UPM), Table 2 Size-effect in processes as a function of chip thickness

Process	Process scale	Chip thickness (μm)	Resisting shear stress (N/mm^2)
Tension test	Multicrystal grain	300–500	300
Turning	Subcrystal grain	40–50	500
Precision milling	Subcrystal grain	5–10	1,000
Grinding	Atomic cluster region	0.5–1	10,000

The reason for increasing the specific shear energy is attributed to the kind of defects available in the crystalline structure at that processing scale. At the atomic bit scale (0.001 μm), the energy required is comparable to the theoretical shear strength (10^5 J/cm^3) [1] (Fig. 1). At the atomic cluster scale, deformation of a material can take place only with the help of point defects (Fig. 1) and requires energy of the order of 10^3–10^4 J/cm^3 [1].

Nevertheless, at some higher scale (0.1–10 μm), called as subcrystal grain size, 2D or 3D defects in the crystalline structure cause a significant reduction in the specific shear energy to 10^2–10^3 J/cm^3



Ultraprecision Machining (UPM), Fig. 1 Defect distribution and scale of specific shear energy required

(Fig. 1) [1]. In the crystal grain size range, (>10 μm), all kinds of dislocations and defects at grain boundaries lower the processing energy further to 10^1–10^2 J/cm^3 [1]. Unlike the case of the ductile (metallic) materials discussed above, the brittle materials have a network of microcracks in place of the network of dislocations (Fig. 1). In the ductile materials, below the point defect scale (1 nm to 0.1 μm) (Fig. 1), and in the brittle materials below the microcracks scale (0.1–10 μm) (Fig. 1), the specific shearing energy is equal to the theoretical shear strength of the work material.

Cutting Tool Material and Geometry Implications

The cutting tool edge in ultraprecision machining being subjected to extremely high cutting pressure should withstand not only the pressure but also the wear. In addition, the edge should have high degree of sharpness to perform atomic bit removal. The SCD tools, possessing the desired properties, are the most suitable tool materials for ultraprecision machining. However, their limitation stems from the achievable cutting edge sharpness, which normally is limited to hundreds of nanometers. Therefore, in UPM, several atomic layers or atomic clusters are processed than atomic bits, primarily on softer metallic materials like Al and Cu. On the other hand, on hardened steels, and other brittle materials, SCD tools undergo rapid wear, thereby necessitating the use of diamond abrasive process like nanogrinding.

Ultraprecision Machining Systems

It uses time-tested principles of precision engineering along with (1) the recent advancements in the control, feedback, and drive systems, (2) CAD/FEM for designing, followed by (3) fine-tuned assembly as well as system integration. This combination results in a machine that is thermally stable, highly reliable, and flexible and at the same time faster in response [3]. Several features of an ultraprecision machine are summarized in Table 3.

Ultraprecision Machines and Processes

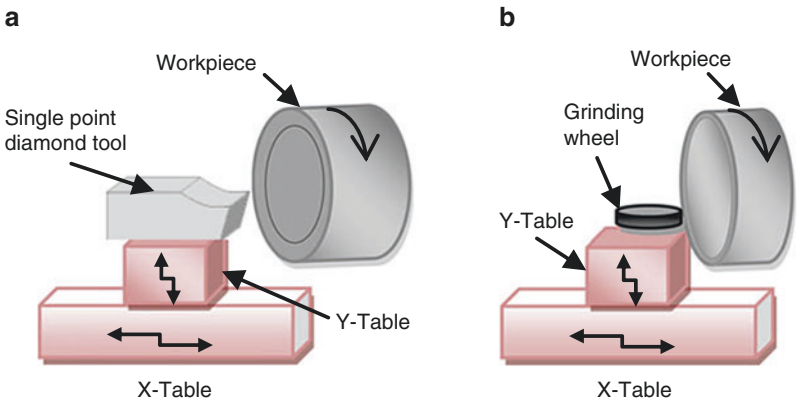
A number of ultraprecision machines and processes have evolved using nanomechanical removal principles. They are discussed below.

Single Point Diamond Turning

It is a nanomechanical processing machine that was developed primarily for 3D shaping and surface finishing of soft metals such as copper and aluminum and polymers. The machine with the features listed in Table 3 uses very sharp SCD tools to perform ultraprecision cutting. The chips formed during the process are extremely thin,

Ultraprecision Machining (UPM), Table 3 Salient features of an ultraprecision machining system

Machine elements/ purpose/types	Main features
Machine base	
• Provides thermal and mechanical stability, damping characteristics	• Made of cast iron, natural or epoxy granite, polymer concrete
Work spindle	
• Spindle motion errors significantly affect surface quality and accuracy of machined features • Use aerostatic or hydrostatic, recent grooved air bearings	• Both the spindle types have high rotational accuracy and rotational speeds • Aerostatic spindles are for low/medium loads, hydrostatic bearings take heavy loads
Drives	
• Slide drives provide stiffness, acceleration, speed, smoothness of motion, accuracy, and repeatability • Spindle drives are usually AC/DC motors • Slides are usually provided with linear motor or friction drives	• Servo drives are used in contouring operations • Small and precise motions of tools for tool positioning and fine motion are achieved by piezoelectric actuators
Controls	
• Controls are required for linear and rotary drives, limiting, position, and time switches, sensors • They also control thermal, geometrical, and tool setting errors	• A multiaxes CNC controllers are used • PC-based controls are used more recently • Feedback controls have a resolution of nm or sub-nm
Measurement and inspection systems	
• Provides rapid and accurate positioning of cutting tool towards work surface • It also monitors the tool wear condition	• Online measurement and error compensation • Laser interferometer for tool position control



Ultraprecision Machining (UPM), Fig. 2 Typical schematics of (**a**) Single-point diamond turning (**b**) Diamond grinding machine

which indicates ductile deformation of the material during machining. At the same time, the process leaves extremely thin degenerated layer of the work surface as the process initiates shear slip using dislocations (Fig. 1).

Applications of the process include fabrication of metallic mirrors, precision component fabrication in laser, and space and optics fields. A typical diamond turning lathe has T-axis configuration (Fig. 2a). The main machine slide moves along x-axis, and the cutting tool is moves along Z-direction.

Nanogrinding

Single-point diamond turning machine is usually provided with add-on grinding attachment (Fig. 2b) for nanogrinding. This way, harder work materials can also be processed on the machine using the grinding attachment. However, it increases the stiffness requirement of the machine. In nanogrinding process, shear slip is generated due to very low (tens of nm) depth of cut (Fig. 1). This causes crackless ductile failure of hard/brittle materials under the abrasive grains in the diamond grinding wheels. Since the working stress on the abrasives is extremely high, only fine-grained diamond abrasives can be used for this process. The process can achieve mirror like finish on hard and brittle materials like glasses and ceramics. Its other applications involve grinding of mould inserts, glass lenses, and aspherical lenses.

Electrolytic In-process Dressing (ELID) Grinding

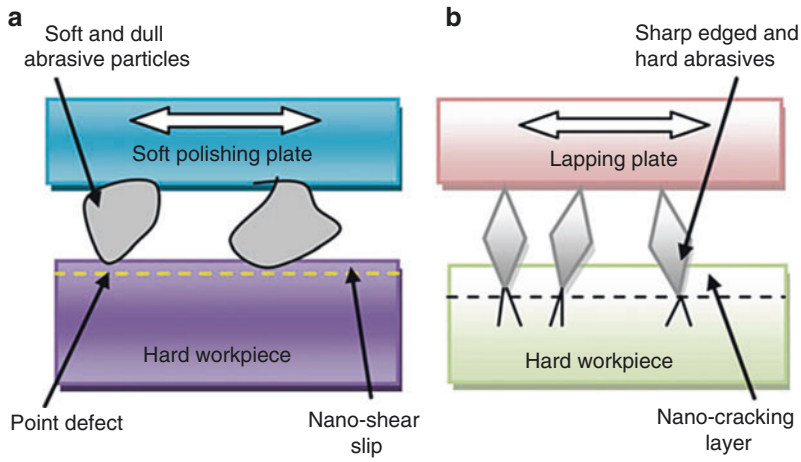
It is a type of nanogrinding process in which metal-bonded diamond grinding wheels are used. The process uses an electrolytic solution as a grinding liquid. Upon application of voltage, anodic dissolution of bonding metal from the grinding wheel occurs. This ensures that worn abrasive particles on the grinding wheel are removed quickly (a wheel dressing action), thereby ensuring continuous availability of sharp particles for the grinding process. The sharp particles initiate shear slip at low depth of cut (Fig. 3) causing crackless material removal. The process therefore is capable of achieving mirror like surface finish on glasses and ceramics.

Nanolapping and Nanopolishing

Unlike nanogrinding, which uses fixed abrasives, these processes use free abrasives to achieve mirror-like finish. The processes involve atomic cluster processing.

Nanopolishing involves sliding of soft abrasives between a soft polishing pad and work surface (Fig. 3a). The abrasives are usually soft and dull; they include Fe_2O_3, Cr_2O_3, CeO_2, or MgO [1]. The abrasives get embedded in the soft pad and originate shear slip at the point defects to effect removal [1] (Fig. 3a). This ensures smooth polishing of hard metallic surfaces.

The nanolapping on the other hand uses a medium to hard pad besides extremely sharp and hard abrasives such as diamond, CBN, SiC, SiO_2, or B_4C [1]. The fine and hard abrasives initiate cracking at the point defects to effect removal [1] (Fig. 3b). This helps generate mirror-like finishing on hard and brittle materials.

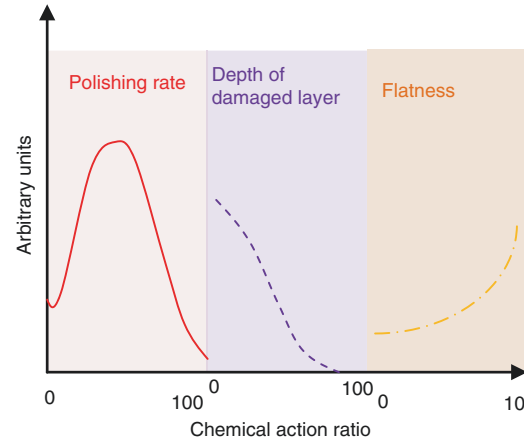


Ultraprecision Machining (UPM), Fig. 3 (**a**, **b**) Nanopolishing and nanolapping

P-MAC

The progressive mechanical–chemical (P-MAC) polishing is a hybrid nanopolishing process that uses mechanical as well as chemical action to effect removal to achieve mirror-like finish and damage-free polishing of Si wafers. The process uses ultrafine SiO_2 abrasives (size ~10 nm) suspended in an alkaline (pH ~10) solution along with an artificial leather foam to perform polishing in four stages. The abrasive particle size and the polishing pressure and the stock removal are gradually reduced after each stage [1].

In P-MAC, the greater the chemical action, the higher is the polishing efficiency and the lower is the damaged layer [1] (Fig. 4). However, the increased chemical action reduces control over accuracy (flatness), Fig. 4 [1].



Ultraprecision Machining (UPM), Fig. 4 General processing characteristics of P-MAC

Aspherical Lens Polishing

Aspherical profiles are required at the edges of the lenses to ensure that light beams from the lens edges converge at the focal point. Therefore, the aspherical lens polishing involves polishing over free-form surfaces. The process is often required for polishing of highly accurate X-ray optics and lenses in commercial cameras.

It involves gradual removal using a pitch polishing tool that has very small contact area as compared to the dimensions of the workpiece. The polishing rate is governed by Preston's rule of thumb [1].

$$h = \alpha \cdot v \cdot \Delta t \qquad (3)$$

The pitch polishing tool acts as a pad and supports the abrasives. It remains solid at room temperature but flows with time when heated or under pressure. The composition of the pitch material is usually proprietary of various manufacturers but contains materials like tar, oil, wood, paraffin wax, shellac, and so on. The properties of pitch material include [1] viscosity in the range of 10^7–10^9 Pa · s), softening point, (55–7 °C), penetration hardness (60–80 by Shore D). In this process, the surface roughness tends to improve as the penetration number of the pitch increases, i.e., as the pitch becomes softer. It is seen that CeO_2 and ZrO_2 give better surface finish than super hard diamond abrasives.

Nanophysical Processes

The principle of removal in these processes involves photothermal or photochemical interactions with the work surface. The photothermal interaction causes melting and vaporization of work material due to energy of photon or laser beams. On the other hand, the photochemical interaction involves use of photon energy to break bond between work elements [1]. It is understood that usually laser beams with metallic materials perform removal action by the photothermal principle. However, when the polymeric materials interact with laser beams, photochemical reaction takes place. With the ceramics, the interaction is partly by the photothermal and partly by photochemical reactions.

The photon beams owing to their wavelength of the order of 10–0.1 μm and the photon energy of the order of 0.1–10 eV cause thermal energy transfer that is smaller than the atomic bonding energy of materials. Hence, the photon beams are not suitable for nanometric material removal. Similarly, an electron beam due to small size (2.8×10^{-6} nm), mass (9×10^{-31} kg), and low energy of several hundred kilovolts penetrate deep (several micrometers) into the work surface. Therefore, the electron beams are also not suitable for nanometric material removal.

Unlike the electrons, most of the ions interfere with the surface atoms of a workpiece because their mean diameter is ~0.1 nm and the mean atomic distance is ~0.3 nm. Consequently, the projected energized ions frequently collide with the nuclei of atoms of the workpiece and knock out or sputter the surface atoms. Hence, ion beams are most suitable for atomic bit removal. The process is also called as ion sputter machining. In the process, the electrically accelerated inert-gas ions such as Ar ions with average energy of 10 keV (which corresponds to a speed of ~200 km/s) are unidirectionally oriented and projected on to the workpiece surface in a high vacuum (1.3×10^{-4} Pa) [1]. Hence, they adhere firmly on the target surface. A beam of focused ions can also be used for cutting of extremely hard materials like diamond or sharpening of SCD tools.

Nanochemical or Electrochemical Processing

The chemical reactions inherently involve atomic bit processing. A chemical reaction is a change in the atomic combination of reacting molecules, in which the atomic bonding of a reacting molecule is broken and a new molecule is generated.

In the process, a chemically reactive gas or liquid is applied to a specified position on the solid surface of the workpiece. The reacted molecules are removed or diffused into the surrounding reacting gas or liquid. If the reacted molecules are insoluble or not in vapor form, chemically reactive deposition occurs on the workpiece surface. But if the reacted or reagent molecules diffuse into the surface layers of the workpiece and react with the atoms or molecules there, they perform a chemically reactive surface treatment [1]. The dimensional accuracy obtainable in chemical reactions is in the nanometer range when the processing conditions are stable. In addition, the following aspects are necessary for nanometric processing [1]:

1. In-process measurement and feedback control of position of the processing point and control of the processed volume (area) are necessary. However, it is difficult to realize these in practice. In the photoresist method, control of the processing point position or area is achieved by the patterned mask.
2. The control of the processing volume or depth can be done only by adjusting the processing time and flow rate of etchants.

In the electrochemical process, the basic reaction is the same, but activation energy for the reaction is given by electric field potential and differs from the ordinary activation energy of the chemical reactions based on the thermal potential energy.

Summary

Ultraprecision machining uses mechanical, physical, chemical, or electrochemical sources of energy for effecting material removal to the nanometric scale. Grain size, load applied on the grains, and

the type of work material govern removal in nanomechanical processes. Size of the energy beam particle governs the processing resolution in nanophysical processes. In-process measurement and feedback control are essential for the nanochemical or electrochemical processes.

Cross-References

- ▶ Electrochemical Machining (ECM)
- ▶ Nanochannels for Nanofluidics: Fabrication Aspects
- ▶ Nanotechnology

References

1. Taniguchi, N. (ed.): Nanotechnology – Integrated Processing Systems for Ultra-precision and Ultra-fine Products. Oxford University Press, New Delhi (2008)
2. Baker, W.R., Marshall, E.R., Shaw, M.C.: The size effect in metal cutting. Trans. ASME **74**, 61 (1952)
3. Lou, X., Cheng, K., Webb, D., Wardle, F.: Design of ultra-precision machine tools with applications to manufacture of miniature and micro components. J. Mater. Process. Technol. **167**, 515–528 (2005)

Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies

Thomas Arnold[1], Thomas Franz[2], Frank Frost[1] and Axel Schindler[2]
[1]Leibniz-Institute of Surface Modification, Leipzig, Germany
[2]NTG Neue Technologien GmbH & Co. KG, Gelnhausen, Germany

Synonyms

Chemical-assisted ion beam etching (CAIBE); Chemical vapor machining (CVM); Ion beam etching (IBE); Ion beam figuring (IBF); Ion beam smoothing (IBS); Plasma-assisted chemical etching (PACE); Plasma jet machining (PJM); Rapid atomic processing (RAP); Reactive ion beam etching (RIBE); Surface figuring; Surface patterning; Ultra-precision finishing; Ultra-precision machining

Definition

Ultra-precision surfaces and structures with nanometer accuracy comprise surface shapes and/or surface structures with the highest achievable processing accuracy of nanometer or even picometer rms range. Root mean square (rms) is a value calculated from differences of the surface measuring data and the desired (designed) surface shape data. Its value is a measure of how far on average the error is from zero. The ultimate accuracy of surface shapes and structures is a prerequisite for their physical functioning in many of today's fields of science, technology, and products like electronics, microsystems, and mainly in advanced optics. Examples range from small millimeter size microscope lenses for deep ultraviolet (DUV) light to large telescope mirrors of up to meters in size. In high-performance optics, the extremely low rms values have to be realized for part or the complete surface spatial wavelength spectrum from the full aperture down to the micrometer. In special cases, the processing accuracy must be within few atomic layers for spatial wavelength up to some micrometers. In general, ultra-precision surface machining requires and is done in repeated loops in turns with adequate surface measurements. Advanced measuring equipment of different kinds for the coverage of more than eight orders of magnitude of spatial wavelength is necessary for it.

Ultra-precision surfaces are fabricated by multistep processing chains starting from coarse shaping like grinding up to final fine shaping like polishing. Ion beam and plasma jet techniques besides others, such as magnetorheological finishing (MRF), chemical vapor machining (CVM), and elastic emission machining (EEM), are highly deterministic. They belong to the class of finishing tools that are favorable to achieve the ultimate accuracy.

Ion beams and plasma jets or plasma-assisted techniques, respectively, base upon physical

and/or chemical interaction on atomic level between atomic particles out of the ion and plasma tool, respectively, with surfaces. Generally, such processes are adaptable for different machining in wide application ranges. The capability to set appropriate parameters together with their tight controllability adequate to the required processing condition qualify them as key tools to achieve the desired ultra-precision of surfaces and structures.

High-vacuum ion beam processing uses accelerated mostly inert ions of some hundred eV to about 2 keV (i) to remove atoms from a surface by atomic impact – the so-called ion or atom sputtering; (ii) to deposit films by secondary ion beam sputter deposition or to assist film deposition; (iii) to create nanostructures by complex ion-induced self-assembling effects; or (iv) to smooth surfaces. Combined with additional chemical active species by adding respective gases to the ion source, chemical reaction with surface atoms is exploited to enhance removal rates and/or etching selectivity between two materials like an adhesive mask and the underlying glass substrate. This is called reactive ion beam etching (RIBE). A second version is chemical-assisted ion beam etching (CAIBE). In this case the source is operated by an inert gas and the reactive gas is directed to the surface processed.

Plasma jets or torches are free-burning electrically excited gas discharges under atmospheric or rough vacuum conditions using mostly inert gases like argon. By admixture of adequate gases as precursor, chemical reactive atomic particles like radicals are produced in the plasma. Depending on the application, different gas types are used. They can (i) chemically react with surface atoms to form volatile compounds, thus removing material by dry etching, or (ii) dissociate to deposit thin films. Just the transfers of energy out of the plasma to the surface can (iii) smooth rough surfaces.

Ion beam techniques especially ion beam figuring (IBF) is well established in optics production today, whereas plasma jet techniques like plasma jet machining (PJM), plasma-assisted chemical etching (PACE), and rapid atomic processing (RAP) are at the doorstep to enter in production. Nevertheless, presently both principles are improved further by intense research and development for higher performance, productivity, and new applications.

Overview

N. Taniguchi introduced the term "nanotechnology" in a keynote paper of the International Conference on Production Engineering (ICPE) in Tokyo in 1974. This time nanotechnology was initially proposed to provide precise target accuracy for fabrication processes that involve ultra-precision surface finishing [1, 2]. In the frame of developments of high technologies like semiconductor, laser, x-ray, synchrotron radiation, space telescopes, and advanced military technologies beginning about 1980 worldwide, big efforts had been started for the development of deterministic machining technologies to meet the steadily growing demands of extreme high-quality surfaces mainly for optics. In high-performance optics, the extremely low rms values have to be realized for the complete surface spatial wavelength spectrum from the full aperture down to micrometer or for parts of it. Only these extreme accuracies near the physical limits for processing of (i) macroscopic surface shapes like lenses for electronic circuit lithography or meter-sized giant space telescope mirrors and of (ii) sub-nanometer uniformity of thin film thickness, for example, of mass storage hard disks, enable the functionality of those components in modern equipment and products. The same applies for dimensional control of micro- or nanostructures on wafer surfaces for electronics or micromechanics and micro-optics as used, for example, in miniaturized camera systems for smartphones.

For ultra-precision surfaces, there is a wide variety of shapes from flat to spherical and aspherical or recently even so-called free forms on a similarly wide diversity of work sizes. Materials range from monocrystalline silicon over many types of glass or metal up to very hard and brittle ceramics. Binary and 3D micro- and nanostructures on flat and more and more on curved surfaces are needed, for example, for diffractive lenses or antireflection enhancement.

Different surface shaping and finishing technologies targeting ultra-precision quality have been developed. Besides ion beam and plasma jet techniques described here, there are SPDT (single-point diamond turning), CCP (computer-controlled polishing), chemical mechanical polishing (CMP), MRF, RAP, CVM, EEM, and some others. Summarizing overviews on advanced deterministic surface machining and finishing technologies have been published very early at the beginning of the ongoing development and recently [3–7].

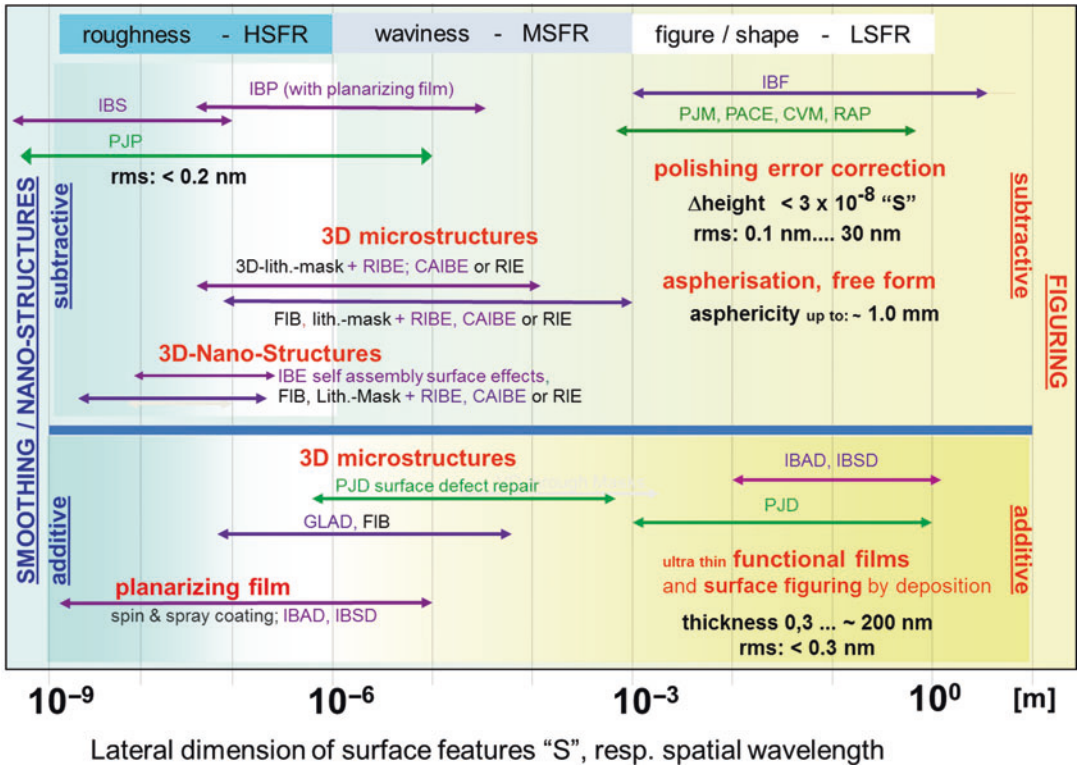
Ion beams and plasmas including recently developed atmospheric plasma jets with their different but well-defined physical and/or chemical characteristics for their interaction with material surfaces are the main tools to achieve ultra-precision accuracies. The heavy atomic particles involved in these technologies are ions, atoms, molecules, and radicals. The capability to tightly control their parameters like mass, kinetic energy, excitation, density, charge state, incidence angle to the surface, chemical reactivity, and others in wide ranges allows tailoring of respective particle tools for the individual processing of surfaces to achieve the demanding surface quality. Automation, long-term stability, and flexibility of ion beam and plasma jet sources combined with advanced surface processing algorithms and reliable machinery including computer control make these technologies efficient for production up to high volumes. Further, in high-end optics fabrication ion beam figuring (IBF) and ion beam smoothing (IBS) are able to overcome physical constraints of the conventional full lap and small tool abrasive polishing processes. For ultra-precision ion beam and plasma jet processing, there are two basic principles: (i) material removal as sputtering and/or dry etching and (ii) material deposition.

Surface patterning includes techniques for preparing, patterning, and structuring surfaces, from the arrangement of single molecules to the macroscale, for example, through etching, electrochemical patterning, and film deposition. Patterning and structuring in most cases are a combination of a mask, binary, or 3D that is defining the pattern and one or more subsequent etching steps to transfer the pattern onto the substrate or a film. Mask structuring is performed by lithography. Photolithography, e-beam lithography, laser writing, laser holographic interference lithography, and ion beam lithography are used. Main etching techniques are the ion-based ones, RIE (reactive ion etching) or RIBE/CAIBE, due to their most important characteristic of being anisotropic (no mask under-etching) and highly selective for different materials (e.g., mask and substrate). Both techniques comprise physical sputtering and chemical reaction between reactive species and surface atoms to remove material.

Not included in this survey of ultra-precision surface processing of ion beams and plasma jets is the special field of focused ion beam techniques. The physical and or chemical ion-induced processes are more or less the same, but the technique similar to electron beam microscopy is very different. Applications are in the field of fabricating prototype nano- and microstructures and mainly on mask repair, basic research, and sample preparation for surface and structure analytics and depth profiling as well. Recently reviews have been published [8–10].

Figure 1 shows schematically the usage of different ion beam and plasma jet techniques with respect to the spatial wavelength over more than eight orders of magnitude.

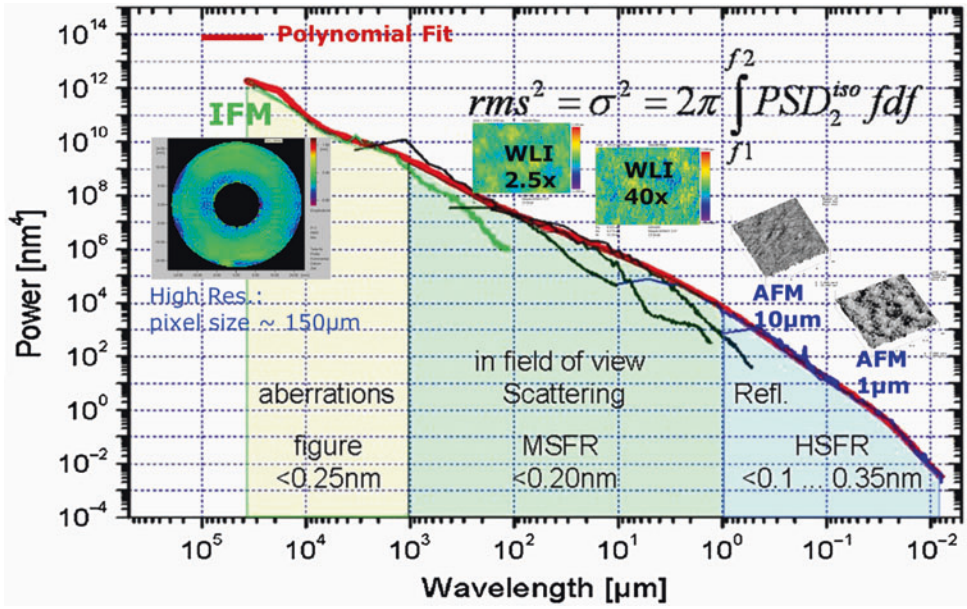
The main technology driver in this field was and is still lithography for semiconductor technology. Advanced deep and extreme UV lithography (DUVL and EUVL) use 193 nm light wavelength and EUV soft x-rays of 13.4 nm, respectively [11]. Advanced high end aspheric lenses and mirrors riqire demanding surface precision in the sub-nanometer range of rms surface shape errors across the entire spatial wavelength range with respect to the mathematical design. This has to be realized from the full aperture range down to micrometer spatial roughness in series production on all surfaces of projection optics for lithography for instance. Figure 2 shows measured power spectral density (PSD) curves of an EUV mirror substrate surface. PSD function represents the surface error amplitude distribution over the spatial wavelength. The measurements were performed over the entire spatial wavelength



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 1 Scheme of summarizing of ultraprecision ion beam and plasma jet methods for large surface processing represented with respect to the spatial wavelength. Some typical target parameters are given (*black*). "Subtractive" means material removal, "additive" material deposition; *HSFR*, *MSFR*, and *LSFR* high, mid, and low spatial frequency roughness, respectively, *IBE*, *IBF*, *IBP*, *IBS*, *IBAD*, and IBSD ion beam etching, ion beam figuring, ion beam planarization, ion beam smoothing, ion beam-assisted deposition, and ion beam secondary deposition, respectively, *PJM*, *PJP*, and *PJD* plasma jet machining, plasma jet polishing, and plasma jet deposition, respectively, *PACE* plasma-assisted chemical etching, *CVM* chemical vapor machining, *RAP* rapid atomic processing, *RIBE* reactive ion beam etching, *CAIBE* chemical-assisted ion beam etching, *RIE* reactive ion etching, *FIB* focused ion beam, *GLAD* glancing angle deposition (by IBSD)

using different equipment according to their appropriate spatial resolutions. From nanometer to micrometer spatial range, atomic force microscopy (AFM, blue curves) from micrometer to millimeter white light interference microscopy (WLI, black curves) and from millimeter to full aperture (green curve) high-resolution interferometry (IFM) was used. The red curve is a polynomial fit comprising the individual PSDs. The three colored areas represent the impact of the different spatial wavelength surface error ranges on the optical performance of the imaging EUV optic for the 13.4 nm light wavelength. The figures are the rms value specs for EUV optics. Shape errors from full aperture up to 1 mm wavelength (LSFR) cause image aberrations. MSFR from 1 mm to 1 μm leads to field of view light scattering resulting in flare or blurring of the image, and HSFR (<1 μm spatial wavelength) causes wide-angle light scatter lowering the image contrast. The overall rms value of the surface σ is related to the integral of the PSD over a considered part or the entire spectral wavelength (formula insert).

As Fig. 2 exemplarily shows, surface topology measurement is the crucial point for final ultraprecision accuracy. This is necessary not only for the final steps but for all phases of the processing chain. Because generally surfaces



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 2 Power spectral density functions (*PSD*) of a well-finished EUV optic surface calculated from individual 3D surface topology measurements [12]. For HSFR, AFM measurements (*blue curves*) with two scan areas, for MSFR white light interference microscopy (*black curves*) with two magnifications and for figure large area interferometry (*green curve*) have been used. The *red curve* is the polynomial fit of all measured data. The three colored areas "figure," "MSFR," and "HSFR" represent the different impact of surface errors of the spatial wavelength ranges on the optical performance of the surface. The indicated figures are the error specs for EUVL optics (Courtesy of U. Dinger, Carl Zeiss SMT GmbH, Oberkochen, Germany)

must be measured for all parts of the spatial wavelength spectrum, different measuring equipment is necessary since each measuring principle is bandwidth limited with respect to the spatial wavelength spectrum as shown in Fig. 2. Therefore, the development of ultra-precision surface processing is characterized by alternating stimulation between measurement development and technology performance, which is still going on [13, 14]. In Table 1 major requirements for surface shape and profile measurements together with some status quo figures for advanced optics are given.

Ion Beam Technologies

Ion Beam Figuring

At the beginning of the 1980s, worldwide efforts had been started for the development of deterministic ultra-precision machining technologies to meet the steadily growing demands of extreme high-quality optics surfaces. This time digital interferometry was in an early stage of its development with very limited performance especially concerning its spatial resolution because CCD cameras had pixel figures of only 32 × 32. Nevertheless, ion beam figuring for polishing error correction of lenses could be demonstrated. Besides scientists from Arizona State University, USA, German researchers in Leipzig in cooperation with Carl Zeiss from Jena obtained early IBF result in 1989 as shown in Fig. 3.

In the middle of the 1990s, IBF brought the breakthrough for efficient volume production in DUV lithography optics due to its high processing determinism connected with the capability to overcome physical constraints of the conventional full lap and small tool abrasive polishing processes as it is their unstable tool behavior at the edges of the optic.

Using multistep scanning schemes combined with dwell time algorithms, processing by

sub-aperture ion beams of different sizes results in an efficient correction of low and part of mid-spatial frequency surface errors down to the sub-nanometer amplitude level with almost no or minimum surface or subsurface damage, respectively, due to the gentle beam surface interaction. Figuring involves specified scanning paths, like meander, spiral, and cycloid, with a variation of the scan speed according to the calculated dwell times. In a standard procedure the pathway of the ion beam spot on the surfaces follows a spherical bowed meander along the lines of latitude. The path velocity of the beam spot on the work surface results from a dwell time distribution proportional to the desired material removal depth distribution. Typically IBF processing software involve the calculation of the dwell times (deconvolution between the error 3D topology to be removed and the 3D removal function of the ion beam tool) for a given topology to be removed and the simulation of the removal process (convolution). Main standard procedures for both are fast Fourier methods. At given local speeds, dwell times t, and the material removal function r, the dwell time procedure can be expressed as follows:

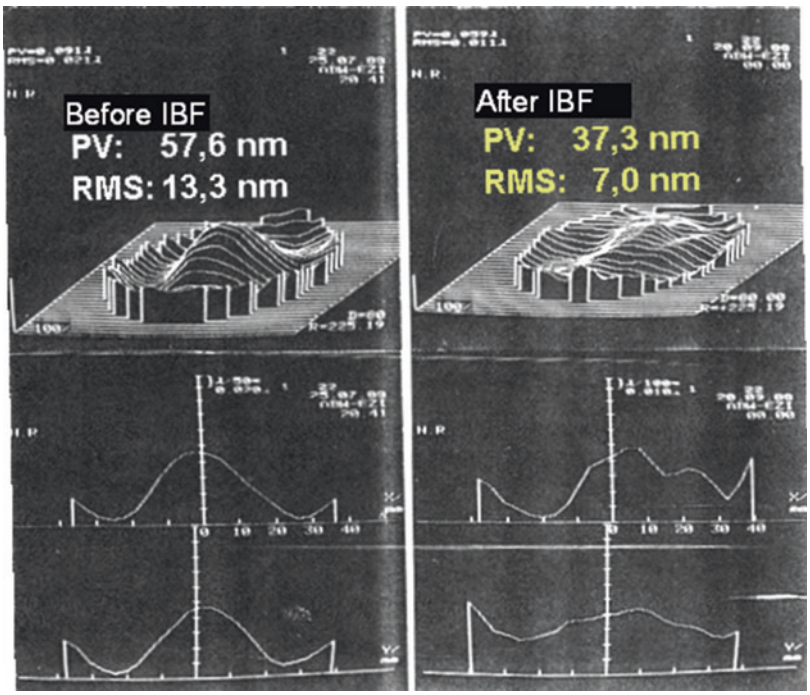
$$h = t * r + \varepsilon \tag{1}$$

where h stands for the heights of the removed material and the symbol * denotes the convolution operator. Matrix notation yields

$$h_{ij} = \sum_k \sum_l t_{i-k,j-l} r_{kl} + \varepsilon_{ij}, \tag{2}$$

Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Table 1 Surface shape and profile measurement requirements for ultra-precision surface processing mainly for optics

Requirement	Influenced by	Current specification rms [nm]
Repeatability	Temperature fluctuations, coherent noise	0.2 . . . 0.6
Reproducibility	Handling	0.5 . . . 1.0
Accuracy	Calibration	0.7 . . . 1.5
Precision	Deviation from: (i) an ideal surface or (ii) a given reference	



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 3 Historical early result on IBF: Interferograms (taken with a 32 × 32 pixel CCD camera) of an 80 mm diameter convex lens before and after IBF obtained by researchers in Leipzig, Germany 1989

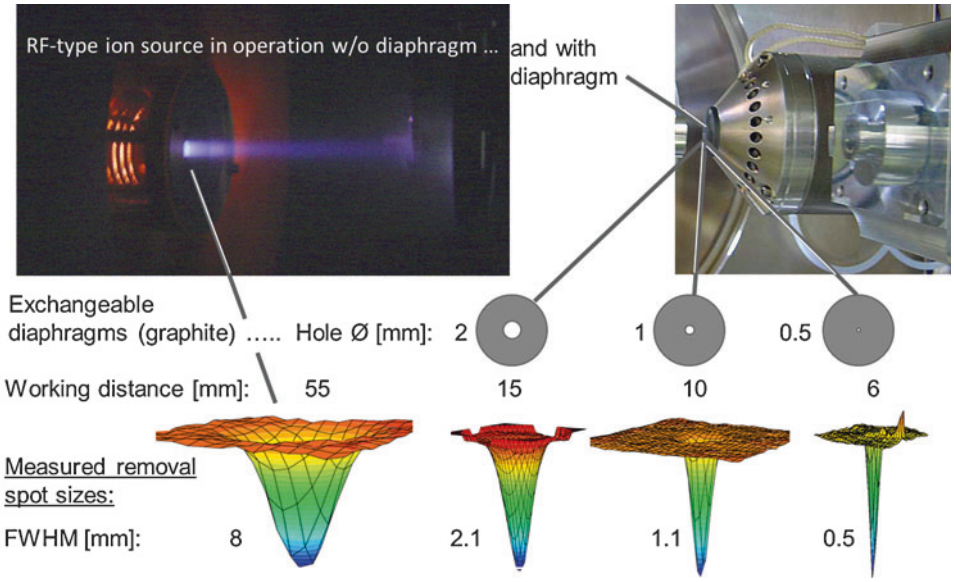
where ε_{ij} denotes a certain amount of error. The summation extends over all values of the material removal function r_{kl}, which are assumed to be nonzero. The amount ε_{ij} is composed of errors resulting from height measurements of the topology, fluctuations in the removal function, and errors coming from speed and location of the motion system. The calculation of the dwell times and the arising final topology are based for planar and spherical surfaces on modified iteration methods selectable according to Van Cittert, Lucy-Richardson, Wiener, and regularized filter [15] and for aspherical surfaces or local dependency removal function on enhanced Gold's method [15]. The relevance of the dwell times with respect to acceleration and speed limits of the motion system is examined, and if necessary, it is immediately corrected by additional smoothing procedure calculations in order to get a final topology with highest quality at a minimum in etching time. The iteration is stopped if (i) the rms value falls below the targeting data of the residual error topology, (ii) the maximum acceleration of the motion system is exceeded despite the abovementioned numerical smoothing procedure, and (iii) the maximum number of iterations is achieved or no better dwell times leading to smaller rms values are calculable. The material removal function of the ion beam can either be generated as a data file of the etching rate matrix or by parameters of superimposed two-dimensional Gaussians. The parameters of the material removal function are externally foreseeable by a corresponding nonlinear parameter fit of the superimposed function to the measured values of a footprint, grove, or meander etching.

The assumption of modeling the removal process by using an isotropic removal function and etching proportional to the dwell times enables the application independent of the physical removal process. The removal function of the ion beam is estimated experimentally by etching the same material as for the figuring and measuring the etched topology by interferometry together with the etching time. Ion beam figuring methods have been reviewed recently [16]. Recently IBF technology development is aimed to extend surface correction of millimeter spatial size range down to below nanometer dimension in height to meet the demanding requirements of the MSFR especially for aspheres or free forms for lithography DUV, EUV, and synchrotron optics. High spatial resolution in deterministic figuring of such tiny surface error features tight technical specs reqirements have to be fulfilled: (i) for the ion beam tool with beam size, stability, and linear etching behavior with time, (ii) for the surface figure measuring with the sufficient spatial resolution, (iii) for the adjustment of the beam tool with respect to the work, and (iv) for the adjustment of the measured data matrix to the physical work surface. Improvements of all these items have been realized already and are still under development to make them mature for production.

For the correction of smaller spatial error features, for instance, an 8 mm FWHM high-current-density beam of a compact 13.56 MHz RF source is reduced by diaphragms of different sizes to 2 mm, 1 mm, and 0.5 mm FWHM (Fig. 4). Figure 5 shows exemplarily the polishing error correction of small spherical optics down to rms <1 nm.

Ion Beam Figuring Plant

IBF high-vacuum plants for workpiece sizes from millimeter up to 2–3 m in diameter are today available on the market. The main components of the systems are (i) a stainless steel processing chamber with a base vacuum of 10^{-4} Pa, (ii) a computer-controlled precision 5-axis (x, y, z + orthogonal rotational axes A and B) or 3-axis (x, y z) system for the scanning path movement of the ion source, (iii) an RF-ion source with an automated Faraday cup system for measuring beam profile and beam position, (iv) a workpiece handling system, (v) a vacuum load-lock chamber for small- and medium-sized optics up to about 70 cm in diameter enabling uninterrupted operation of the high-vacuum and fast workpiece change, (vi) power supply, and (vii) a process gas handling system. The machines are computer controlled and operate fully automatic including workpiece data handling and process modeling. 3-axis systems are preferred as they are more stable in their dynamic behavior compared to the 5-axis systems. In contrast motion to 5-axis in the case of



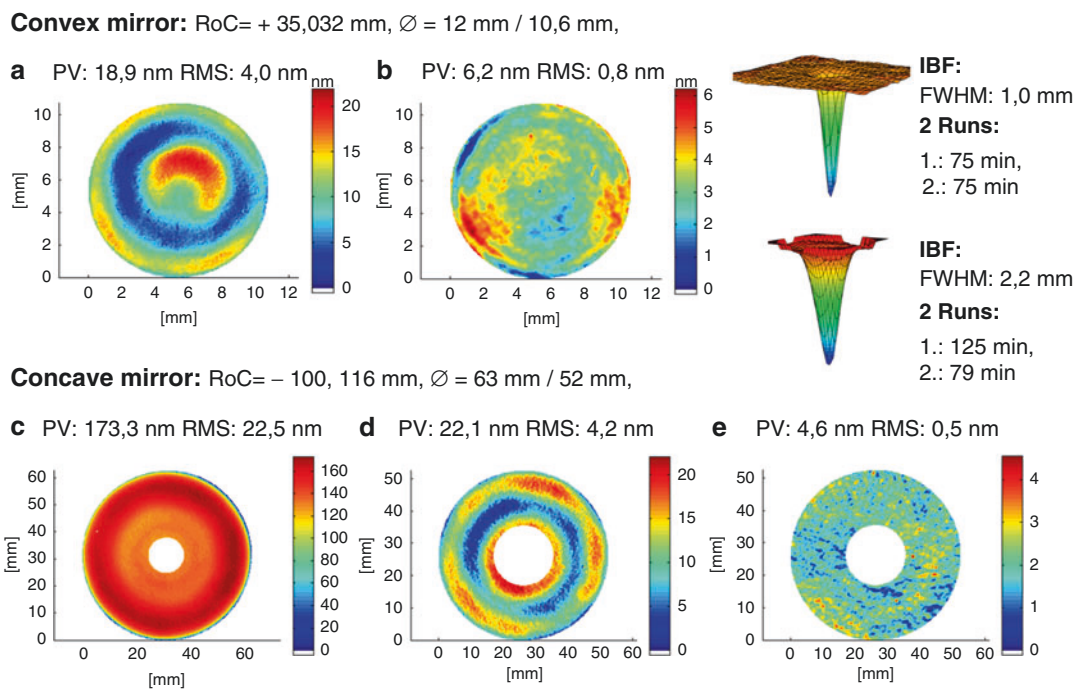
Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 4 RF-type ion source (13.56 MHz) for IBF; *left upper corner*: ion source in operation with an Ar ion beam; *right upper corner*: same ion source equipped with a frame for beam shaping diaphragms made from graphite and hosting inside a hot filament electron emitting neutralizer for avoiding charging up insulating workpieces by the positively charged ions. In the *middle* row, three diaphragms with different hole diameters are shown and below interferometrically measured removal spot sizes (footprints) of the beam without and with the beam shaping apertures, respectively. Also given are the respective work distance between diaphragm front side and workpiece surface and the beam sizes at full width at half maximum (*FWHM*)

3-axis motion system, the ion beam centerline cannot be directed perpendicular to each point of the curved workpiece surface; instead, the beam direction is parallel to the z-axis and the source in this case is maintained in a constant distance to all surface points. Thus, the removal spot size on the one hand and the atomic sputter yield on the other change depending on the curvature of the surface causing variation of the incidence angle of the ion beam. This behavior is now fully included in the IBF processing software for calculating the dwell time matrix to control the ion beam scan path motion. Figure 6 shows state-of-the-art IBF plants for high-end optics production.

The IBF processing simulation module on the base of the measured ion beam removal function and the measured initial surface figure date calculates a dwell time of the ion gun for each point of the surface to get the desired final surface figure. After the checking of the simulation results by the operator, they are fully automatically converted into motion data for the multi-axis system and the ion gun, respectively.

Ion Beam Sources

Ion energies for ion beam etching and figuring are in the range of about 400 eV up to 2 keV. Different types of ion sources hot filament (Kaufman type), RF type (inductively coupled 13.56 MHz), and ECR type (ECR: electron cyclotron resonance) energized by 2.45 GHz microwave are in use. The ion beam is extracted by a high-voltage-powered multigrid ion extraction system forming a compact ion beam out of the individual beamlets from the grid holes. Different grid designs allow variable beam shapes, divergent, parallel, and convergent. With the help of the grid voltages, the beam shape can be adjusted within some limits. For IBF technology mainly Kaufman-type and RF-type sources are used. The RF-type ion source (Fig. 4) is preferred for IBF production purposes due to lower machine downtime by



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 5 Interferograms of IBF with small spot ion beam for ULE® mirror substrates of a Schwarzschild optic for EUV application; (**a**) convex mirror initial surface, (**b**) result after 2 IBF runs using a 1.0 mm FWHM beam size, (**c**) concave mirror, initial surface, (**d**) after 1st run and (**e**) final result after 2nd run. For the concave mirror, a 2.2 mm FWHM beam size was used. The IBF machining times are given together with the interferometrically measured shape of both beam sizes. *RoC* radius of curvature, the second diameter value is the optically active area

maintenance (no hot filaments exchange after their breaking). Sources are available for different beam sizes in the range of about 5 mm to 200 mm FWHM of the near-Gaussian beam shape with reasonable working distances by means of different extraction grid designs. Additional diaphragms mounted in front of the source are used to reduce beam sizes down to ~0.5 mm FWHM. Hot filament or so-called plasma bridge neutralizer mounted outside of the beam area ejects electrons and compensates the positive charge of the ion beam. This is necessary for a stable processing of insulating materials.

Ion Beam Erosion Self-Assembled Nanostructures and Smoothing

Stringently deterministic IBF finishing technologies with macroscopic beam tool sizes as described above are now well established in production for shape correction and figuring of high-performance optics. In contrast, ion beam-induced smoothing of spatially micrometer and nanometer features is strongly coupled to atomistic processes that are characteristic for the much smaller spatial length scales and are less deterministic. Low-energy ion beam techniques (ion energy <2 keV) have been developed to improve the microscopic surface roughness of solid surfaces on a nanometer to some ten micrometer spatial scale to achieve ultrasmooth surfaces with rms roughness values <0.2 nm. Two processing schemes *ion beam direct smoothing* and *smoothing with planarization or sacrificial layers* are demonstrated [17]. In ion beam direct smoothing (IBS), favorable relaxation mechanisms in the development of surface topography during the ion beam erosion are exploited. For the second method, the smoothing effect results from

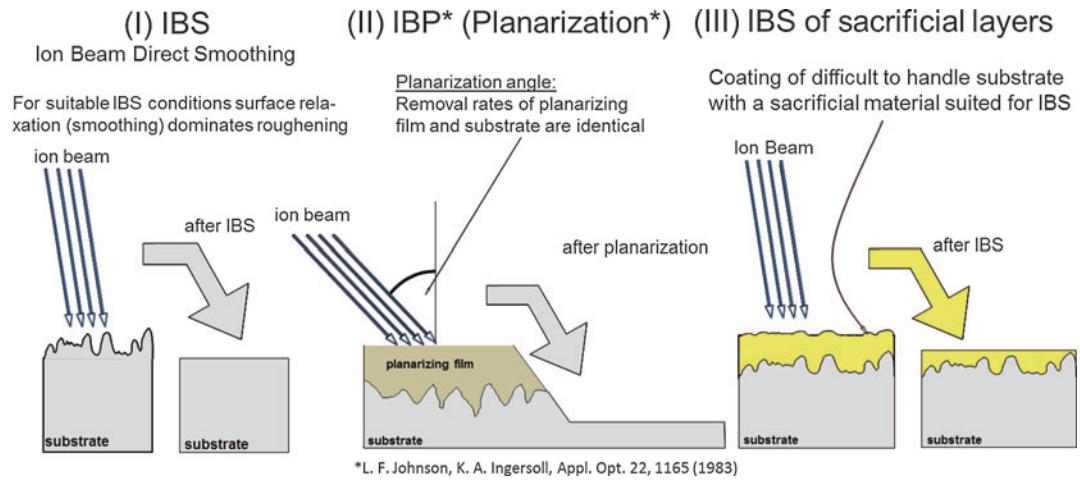


Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 6 State-of-the-art ion beam plants for IBF and partly for IBS (5-axis motion system necessary, as in plant (**a**)); (**a**) plant is for high-throughput production of workpieces up to 100 mm in diameter and (**b**) IBF plant for large mirrors up to 1.5 m in diameter, up to 520 mm thickness including mechanical frame of the assembly of large mirrors, and up to 1000 kg weight. On the *left* of the photograph, the mechanical interface including slide system for mounting the mirrors is shown (Courtesy NTG GmbH, Gelnhausen, Germany; www.ntg.de)

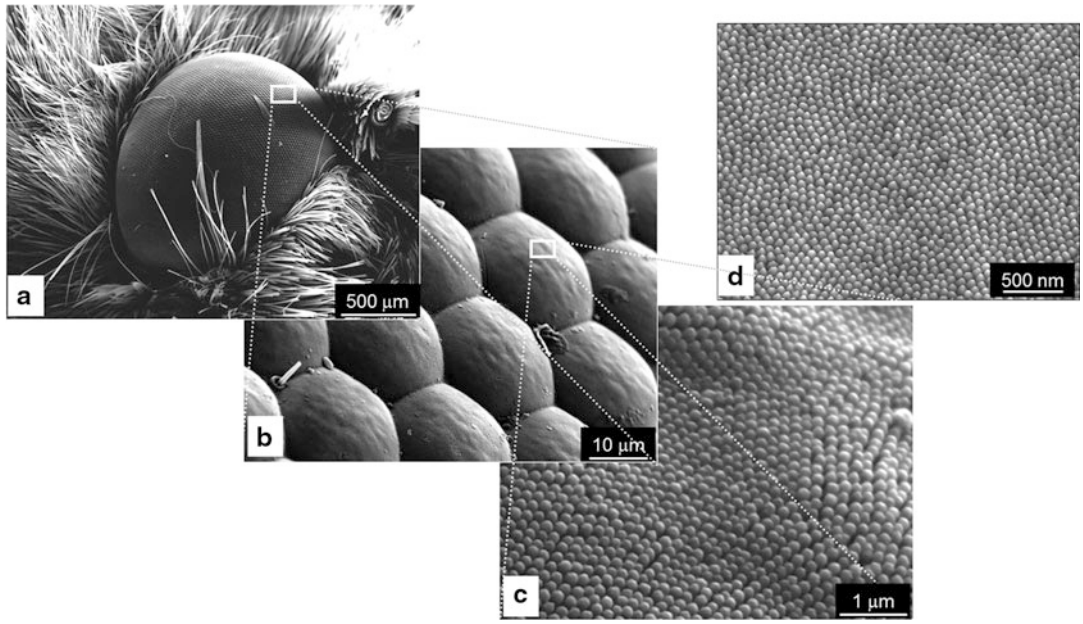
the removal of a planarizing layer deposited on the substrate (IBP – ion beam planarization). Both smoothing techniques were investigated for a wide range of various materials. In Fig. 7 ion beam smoothing working principles are shown schematically.

Generally, the surface topography evolution on the nanometer and micrometer spatial scale during low-energy ion beam sputtering is rather complex. A multitude of topographies can result from the surface erosion due to different roughening and smoothing mechanisms. Besides the actual removal of material, the surface erosion process often gives rise to a pronounced topography evolution, generally accomplished by a kinetic roughening of the surface and the generation of so-called self-assembling nanostructures [17]. Figure 8 shows 3D structures in the range of 100 nm spatial dimensions: a moth eye as a result of the evolution of life (Fig. 8a–c) and fabricated by self-assembling effects by Ar ion beam erosion of an InP semiconductor wafer surface (Fig. 8d) [19].

One goal of ongoing intensive research and development is to get insight in the physics of the surface atomic processes responsible for such structures. The second one follows the vision to exploit the ion-induced self-assembling surface effects to develop techniques to generate nanostructured surfaces efficiently without costly mask techniques. Theoretical and experimental research during the last several years brought much progress for the understanding of the ion-induced nanoscale surface topography evolution by self-assembling effects. Several atomic-level effects have been identified. They are all connected with stimulated motion of surface atoms along the surface. Surfaces can be smoothed via thermally driven surface diffusion at sufficient temperatures under ion bombardment-related defect production. Ion-induced directed or random fluxes of recoil atoms moving parallel to the surface are able to compensate the curvature-dependent sputtering or can contribute to ballistic surface diffusion process. For amorphous materials or ion beam amorphized surface regions, defect-mediated ion-enhanced viscous flow may also occur. The individual conditions of material, ion species, energy, angle of incidence, and surface temperature or the participation of third-party



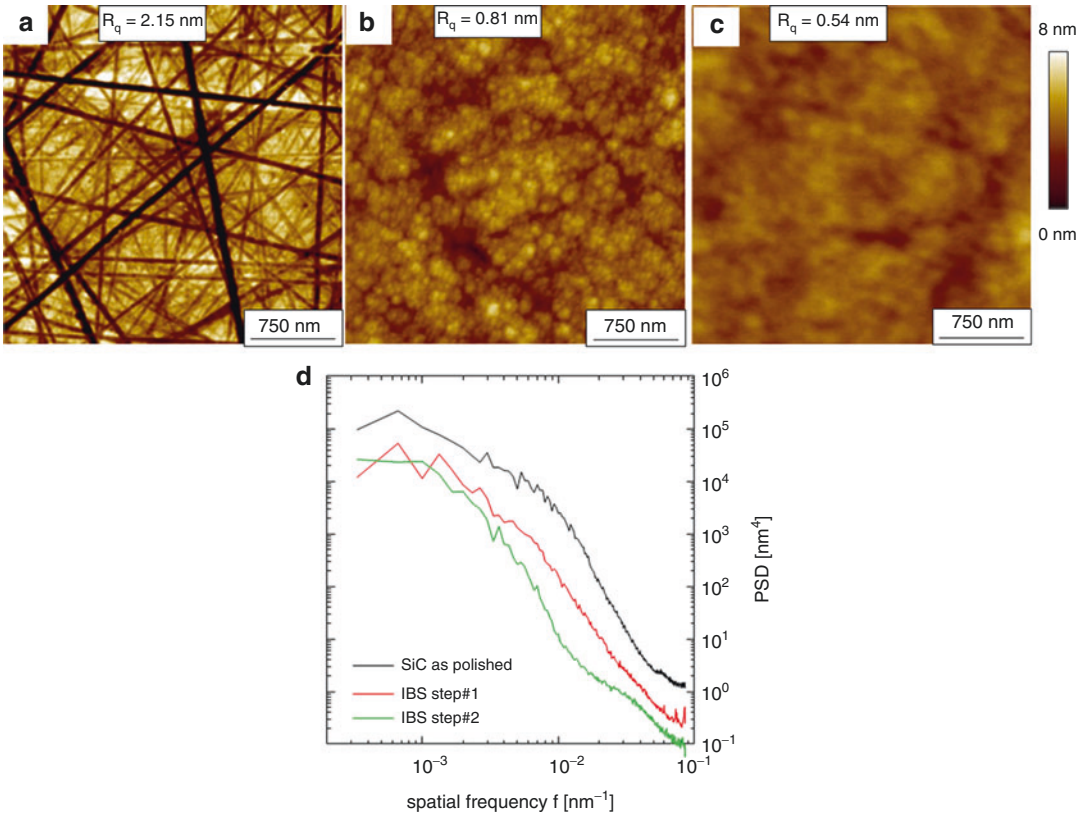
Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 7 Ion beam-assisted methods for surface smoothing. The planarization technique was described first by L. F. Johnson and K. A. Ingersoll [18]



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 8 Scanning electron microscopy pictures of nanometer-sized 3D structured surfaces; (**a**–**c**) a moth eye with three different magnifications; (**c**) nanostructures are generated by the evolution of life to enhance the antireflection of the eye surface for improved night vision of the insect; (**d**) similar 3D nano-patterned Si wafer surface fabricated by atomic-scale self-assembling due to Ar ion beam erosion [19]

particles (intentionally or unintentionally added to the ion beam processing), to mention the most important ones, determines the resulting surface structure.

However, dependent on the same physical effects, also favorable ion beam erosion conditions can be realized where surface smoothing dominates the topography evolution on the



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 9 Example for ion beam direct smoothing of a polished SiC surface (Ar^+, E_{ion} = 800 eV, j_{ion} = 250 μAcm^{-2}, simultaneous sample rotation). AFM images show (**a**) the initial as polished surface, (**b**) after IBS step1 (ion incidence angle 70°) and (**c**) after IBS step2 (ion incidence angle 0°). The rms roughness was reduced from R_q = 2.15 nm to R_q = 0.54 nm. (**d**) shows the calculated power spectral density curves from the three AFM data [20]

microscopic spatial scale. This is illustrated in Fig. 9 for ion beam direct smoothing of hard-to-polish polycrystalline SiC. Starting with a mechanically polished surface (Rq = 2.15 nm, Fig. 9a) after Ar^+ IBS (E_{ion} = 800 eV, j_{ion} = 250 μA cm^{-2}) at 70° incidence under simultaneous sample rotation (Fig. 9b), 0.7 nm rms was achieved, and after a second IBS step at 0°, 0.54 nm rms (Fig. 9c). The PSD functions in Fig. 9d show for the first IBS step a uniform reduction of the roughness over the whole spatial frequency spectrum, whereas for spatial frequencies <0.002 nm^{-1}, the smoothing drops down. This indicates a long spatial wavelength limit for efficient ion beam direct smoothing owing to the limited spatial range of the related atomic surface transport processes mentioned above, which are responsible also for this kind of smoothing.

Therefore, ion beam direct smoothing is less efficient for spatially larger roughness features. Alternatively, a modified planarization technique can be used for the reduction of surface roughness at lower spatial frequencies. In ion beam planarization (IBP), the rough surface is spin or spray coated by a suited sacrificial layer with a low viscosity (e.g., photoresist) to level out the surface roughness. Subsequently the sacrificial layer is removed by ion beam etching at the planarization angle. This angle is defined as the local ion beam incidence angle where the removal rates (sputter yields) of the sacrificial layer and the underlying substrate are nearly identical.

Thus, the smooth surface of the sacrificial layer is transferred into the substrate. Using this technique, roughness components with spatial wavelength up to some tens of microns can be reduced. In Fig. 10 NiP surface smoothing by IBP combined with IBS steps after single-point diamond turning is shown. The efficient removal of the diamond tool marks is evident from the AFM pictures in Fig. 10a for different resolutions. Also shown in the right upper corner are the etch rates dependent on the ion incidence angle for NiP and resist type AZ1450J used for the planarizing film, indicating ~30° as the planarization angle in this case. The resist was spray coated with a thickness of ~80 nm. The PSD in Fig. 10b comprises measurements by AFM and optical profilometry and clarifies the range of spatial wavelength where smoothing operates is up to some 10 μm.

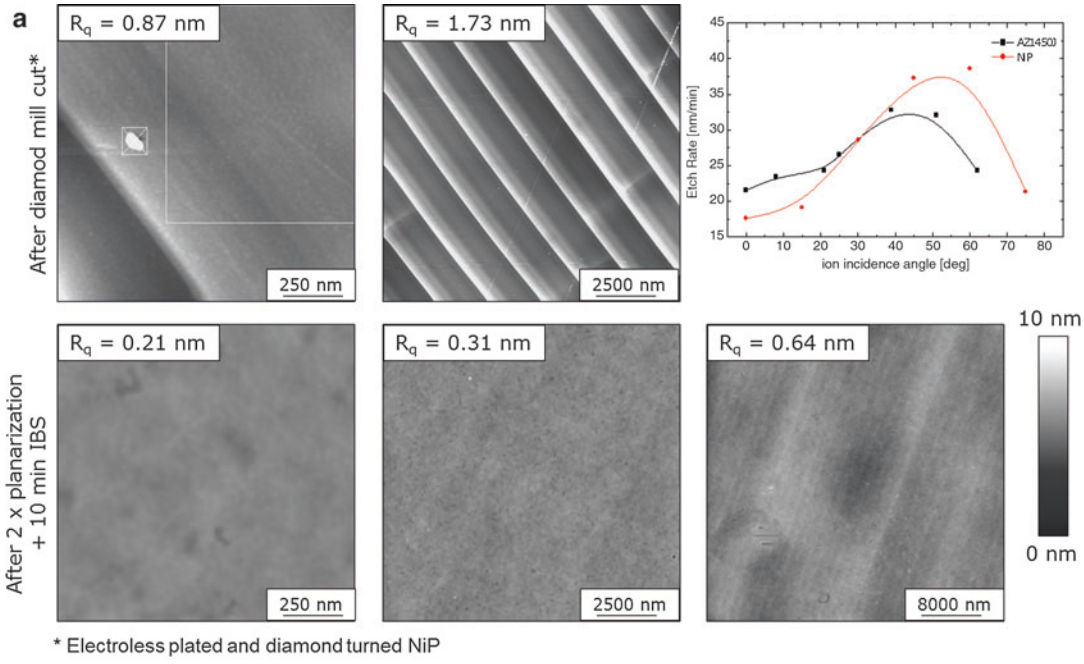
In Fig. 11 a sequence of ion beam smoothing steps and thin film Si deposition to reduce the ion beam etched-related roughness enhancement resulting from IBF for polishing error correction of Zerodur® is shown. First, an IBF shape correction was performed. Zerodur® as a glassy ceramics in contrast to glass, fused silica, or Si has a nonuniform chemical and structural composition leading to different ion sputter removal of the material constituents. Therefore, during lBF surface roughness is enhanced above the EUVL spec of the HSFR (<0.15 nm rms) (left in Fig. 11). In a second step, the substrate is coated with a thin a-Si layer by IBSD. The surface roughness of the Si film surface is reduced (middle in Fig. 11). Next an IBP step and finally an IBS step were applied (right in Fig. 11). After optimization loops were performed, an HSFR of <0.15 nm rms could be achieved. One has to exercise care in doing such technical trick not to remove the Si layer completely at any point. This would cause again roughness enhancement on Zerodur. The procedure all at all does not destroy the initially achieved shape accuracy by IBF due to the minimum film thicknesses applied and hence the minimum and deterministic material removal for the smoothing. The error values given correspond to the standard deviations obtained from at least ten measurements on a two-inch substrate. The PSD graph shows the effective surface smoothing for spatial frequencies that are $>f = 1 \times 10^{-4}\,nm^{-1}$ corresponding to a long wavelength limit in the range of some 10 μm.

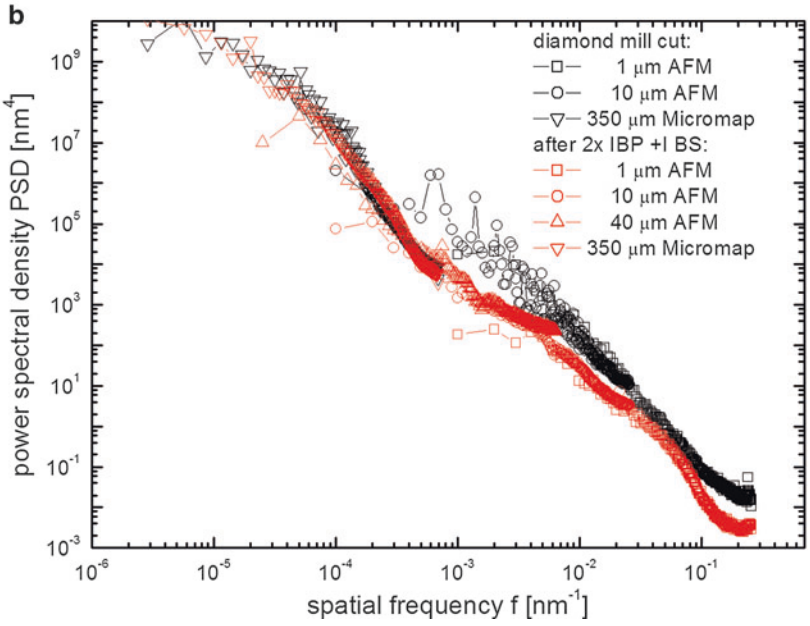
Three examples of ion beam-related surface smoothing of very different materials und surface conditions shown here demonstrate conclusively the high application potential for fabrication of ultrasmooth surfaces. Meanwhile, the effects were demonstrated for a wide variety of materials and different applications. At present developments are going on to qualify ion beam equipment and processing schemes for making ion beam smoothing ready for an efficient production technology. It will definitely supplement existing technologies for such cases where they come to technical or performance limits.

Finally, a further machining method with high smoothing potential, the gas cluster ion beam technique [22], exists. Here clusters of up to some 1000 atoms are formed when high-pressure gas flows through a tight nozzle. Then they are ionized and accelerated up to some 10 keV. The gas cluster ions break when they hit a surface, and every individual atom gets only a small part of the kinetic energy in the range of some 10 eV. Their preferred motion is parallel to the surface; thus, they transfer momentum to top most surface atoms and smooth, and additionally they generate only minimum or no subsurface damage.

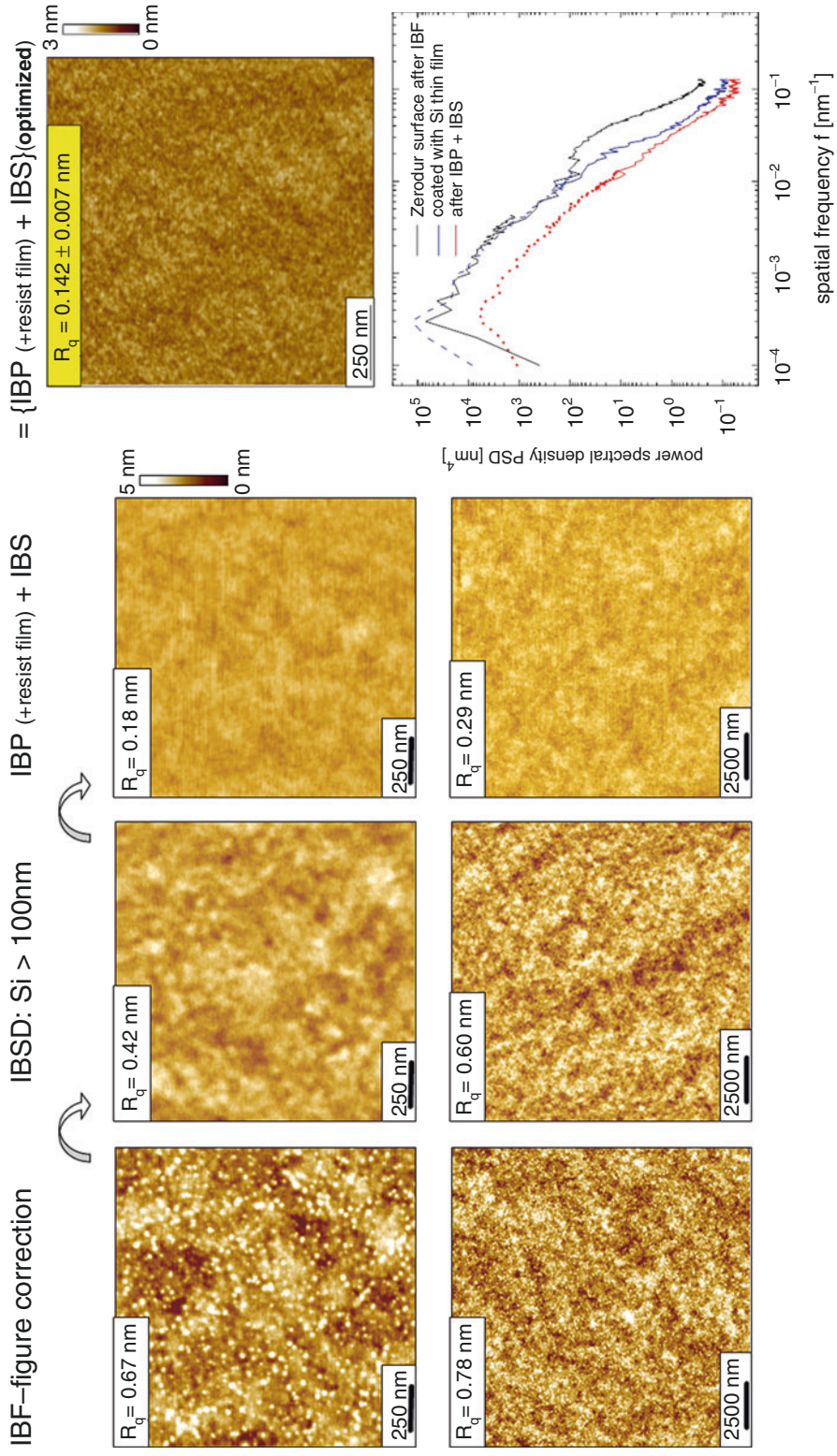
RIBE Pattern Transfer

In modern microsystems and micro-optics technologies, there is a growing need of micro- and nano-patterned surfaces. Besides binary structures, 3D-shaped complex patterns for a variety of materials are required, e.g., refractive micro lens arrays or diffractive blazed Fresnel lenses. Advanced 3D mask techniques like dose dependent e-beam lithography, laser direct writing lithography, grey tone mask, and holographic interference lithography have been developed. The majority of substrates are plane wafers but also thick curved quartz glass substrates, for example, for molds to fabricate polymer imaging blazed diffraction gratings by embossing must be processed. The standard procedure for micro- and





Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 10 Removal of tool marks and surface smoothing on NiP surfaces by ion beam planarization. (**a**) Sequence of AFM measurements using different scan areas after single-point diamond turning (*upper row*) and after 2× IBP + 1× IBS (*lower row*), (**b**) PSD calculated from the shown AFM and additional *white* light interference microscope measurements indicating efficient smoothing for spatial frequencies $>10^{-4}$ nm^{-1} (<10 µm) spatial wavelength [21]



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 11 Demonstration of an ion beam processing sequence for surface finishing of Zerodur® used as substrate for EUVL optical element. Details are explained in the text [21]

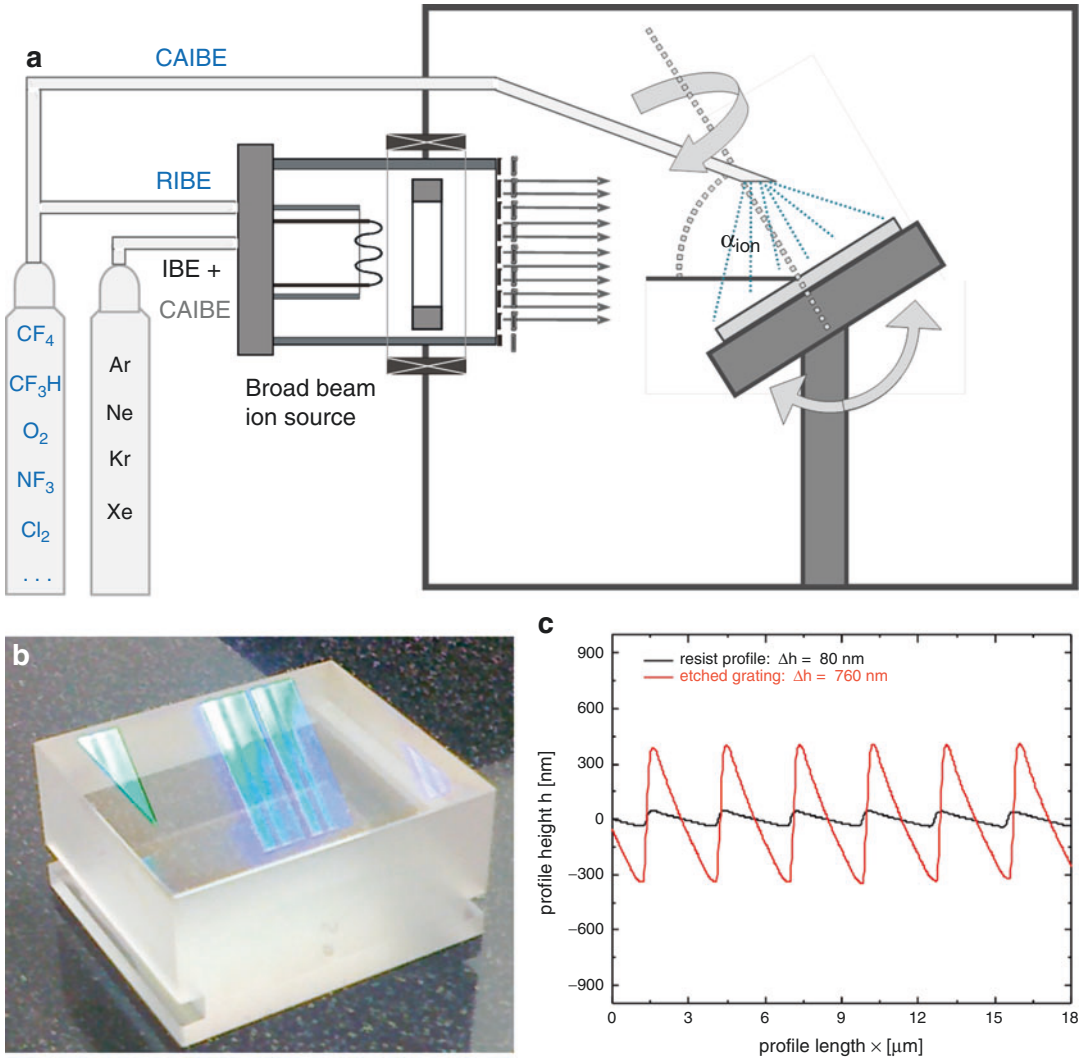
nano-patterning is the transfer of masks by reactive plasma or ion-based etching. In 3D case the mask is sacrificial and the transfer can be either proportional or altering the shape. Reactive ion etching (RIE) as inductively coupled plasma etching (ICPE) or capacitively coupled plasma (CCP) is the mainly used process due to its high performance for etch rates and material selectivities [23, 24]. Ion beam techniques IBE, RIBE, and CAIBE offer the advantage of separating the workpiece from the plasma. This allows additional parameters with respect to RIE. Thus, the transfer etching behavior can be adjusted beneficially: angle of particle incidence, independent control of process parameters, sample manipulation during processing (rotation), and co-deposition of other materials featuring, e.g., to etch profiles of predetermined sidewall angles. Figure 12a shows the schemes for IBE, RIBE, and CAIBE, Fig. 12b shows a RIBE-etched blazed grating on a thick quartz substrate, and Fig. 12c shows AFM measurements of profiles of the resist mask and the blazed grating RIBE transferred into fused silica. Figure 13 shows examples of tailoring of profile structure during RIBE by variation of ion beam incidence angle and process gas mixture.

Plasma Jet Technologies

PJM: Deep Aspherization and Polishing Error Correction

Early developments of plasma-assisted chemical etching (PACE) originate from Bollinger and Zarowin [25]. A variety of atmospheric plasma jet (APJ) tools with different removal rates and tool sizes have been developed covering different applications by material removal but also including thin film deposition. In advanced high-tech optics, there is a growing demand for precision aspherical and free-form optical elements with large deformations from standard flat, cylindrical, or spherical shapes. Plasma Jet machining (PJM) is a noncontact, locally dry-etching method that can be beneficially used in the fabrication of such elements. The working principle is pure chemical reaction between reactive species originated within the plasma and the surface atoms. In the case of silicon or silicon compounds, fluorine radicals or atoms are used to build together with Si atoms the volatile molecule SiF_4. This escapes and the etching proceeds. Figure 14 shows the plasma jet source operating scheme and the simplified chemical surface reaction of F with Si and fused silica together with an atmospheric plasma jet source in operation machining a concave fused silica cylinder lens.

Preferred precursors admixed to the plasma inert source feed gas (e.g., He or Ar) are the standard process gases usually used in silicon wafer technology like CF_4, SF_6, and the like. For fused silica and ULE®, material removal rates of up to 50 mm^3/min are easily achieved. Besides such high-rate PJM with removal rate performances comparable to grinding, very low removal rates can be achieved as in the case of IBF by reducing RF power, reactive gas concentration, or gas flow. Also wide changes in the size of the removal function are possible by changing the diameter of the nozzle of the plasma jet source. Thus, large parts can be as efficiently processed to achieve deep aspheres as the correction of spatial millimeter size error features with nanometer deviation can be performed on up to large surfaces. The processing algorithm is mostly the same dwell time principle as described above for IBF. Primary advantages of the plasma-assisted chemical methods are the following: (i) there are nearly no limitations arising from local curvatures of the parts to be machined, and most importantly (ii) no subsurface damages of the material are generated during the pure chemical machining process. Therefore, plasma-assisted chemical etching leads to machined surfaces with unchanged material characteristics equal to the volume. This is in beneficial contrast to the subsurface damaged layers originating from mechanical or chemomechanical machining (e.g., grinding, lapping, polishing, and others) where the processing modifies up to some 10 μm thick surface layers by mechanical forces and/or the incorporation of contaminants from polishing slurry. Even the bombardment of low-energy heavy ions during IBF leads to comparable low material modification within some 10 nm from the surface. PJM can also completely remove all such

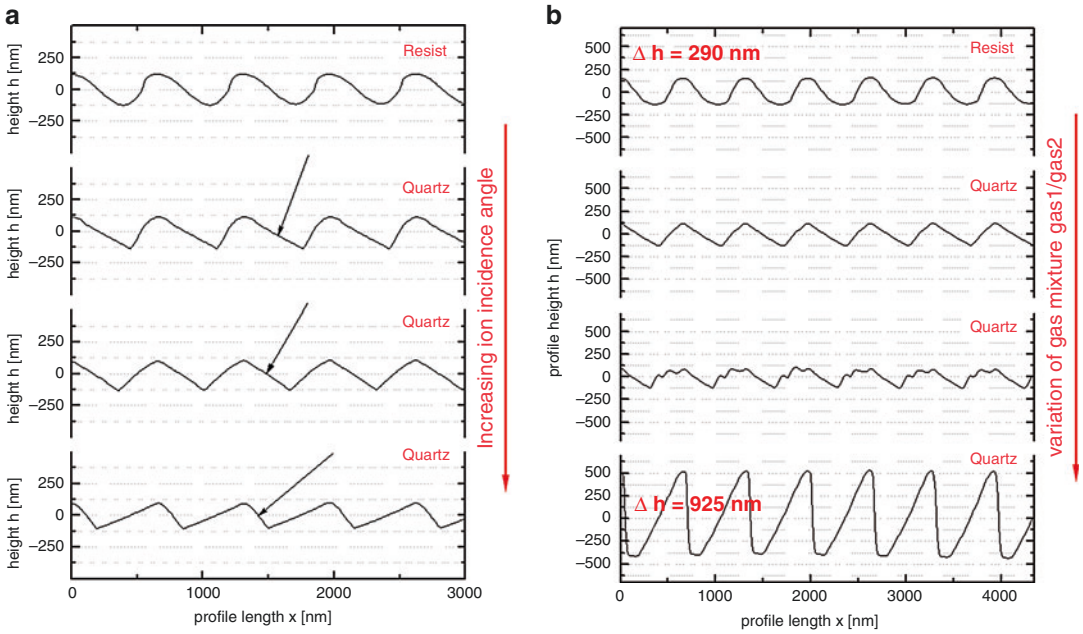


Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 12 Schemes of IBE, RIBE, and CAIBE with variation of ion beam incidence angle and workpiece rotation (**a**), RIBE-etched blazed grating on a thick quartz substrate (**b**), and measured profiles of the resist mask and the blazed grating RIBE transferred into fused silica (**c**), (Fig. 12b) (Courtesy Carl Zeiss Jena GmbH)

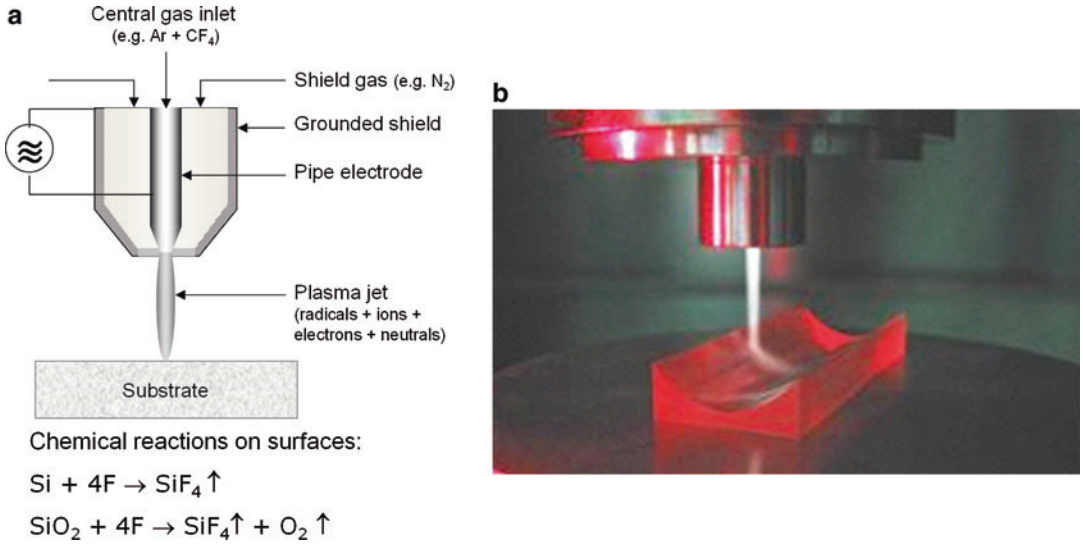
damage layers. This reduces the polishing effort afterwards largely. Shape conserving post polishing after PJM is usually necessary for ultra-precision surfaces in cases with several micrometer plasma jet removals because the spontaneous chemical etching, which starts preferred at the statistical distributed microscopic surface defects with their weakened atomic binding forces, results in a slight roughness enhancement.

Mid-spatial surface error features are typical for modern CNC coarse machining (grinding and lapping), and if existing on a surface, they are hard to remove. They originate mainly from unavoidable residual noise or vibrations from the machine mechanics transferred to the tool that is in more or less hard contact with the workpiece surface. Plasma-assisted high-rate removal technology is contactless as is ion beam, and therefore, it can a priori avoid or at least minimize such hard to remove features.

On the other hand, there are some critical points for at least high-rate plasma jet machining that



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 13 Resist grating pattern transfer into quartz by CHF_3/O_2/Ar-RIBE: tunable blaze, anti-blaze angle, and length ratio of a sawtooth structure by variation of the ion beam incidence angle but with constant profile depth; (**a**) only Ar-IBE (selectivity quartz/resist ~1) and (**b**) CHF_3/O_2-RIBE with variation of the flow rate ratio of process gases CHF_3/O_2 which allows to alter the profile depth in situ (selectivity ~3)



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 14 (**a**) Scheme of an atmospheric plasma jet source with a coaxial electrode configuration typical for the use of 2.45 GHz microwave or standard RF frequencies plasma excitation, indicated are simplified chemical reactions for Si and SiO_2 etching by plasma generated F radicals with surface atoms; (**b**) plasma jet source machining of a fused silica concave cylinder lens, in this case an additional shielding aperture in front of the plasma source prevents the contact of the microwave excited primary plasma with the surface in favor of a radical beam [26]

should be considered and that indeed slows down the progress to implement PJM into production: (i) the etching process is very sensitive to surface structural defects and contaminant chemical impurities, which means that perfect cleaning of surfaces to remove any traces of grease or other contaminating materials that will inhibit chemical reaction by the plasma-generated reactive species is crucial for plasma-assisted chemical etching; (ii) in contrast to semiconductor industry, hazardous chemicals like F radicals that are produced in the plasma jet are not common in standard optics fabrication; and (iii) plasma etching rate is proportional to two independent parameters: (a) linear to the flow density of the reactive species on the surface and (b) exponential to the surface temperature in the reaction zone, which causes principally a nonlinear etching rate. The chemical reaction is characterized by an activation energy that causes an exponential Arrhenius behavior of the etching rate on temperature. The plasma jet delivers both the flow of radicals and the thermal energy for the reaction. Therefore, in the usual case of dwell time controlled figuring, the removal is nonlinear with respect to the dwell time in contrast to IBF. Especially for materials with rather low thermal conductivity as fused silica (quartz glass), the local surface temperature in the contact point of the jet and the extension of the reaction zone around this point are strongly dependent on the scan velocity of the jet. Additionally, superimposed to this is the heating up of the whole workpiece with proceeding machining time. For a reliable technology, measures must be taken to linearize the removal rate with dwell time. Recently a first attempt was made to compensate for the nonlinearity of the fast local heating effects from the dwell time variation and of the slowly changing temperature distribution within the whole workpiece. The procedure calculates the dynamical heat flow from the jet throughout the workpiece including the geometrical and thermal data of the sample and the relevant data from the jet and its dwell time-related motion across the surface in a FEM model [27]. In the simple case of a flat quartz cylinder workpiece used to demonstrate the feasibility of the procedure and to check it experimentally, an accuracy improvement from 10 % residual error from the design shape without nonlinearity compensation to 1 % with the compensation was achieved.

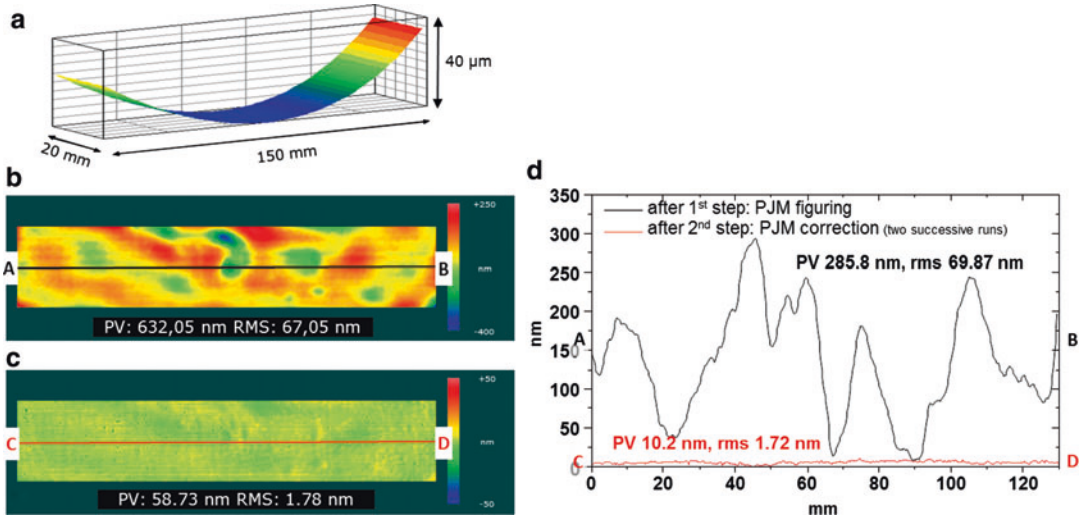
In addition to deep aspherization, also deterministic surface machining up to sub-mm spatial wavelength range with nanometer depth accuracy can be performed by PJM [25]. In such cases main PJM source processing parameters, namely, electrical power and reactive precursor gas flow, will be downsized to reduce the etching rate by several orders of magnitude. The size of the removal function can be reduced up to about 0.5 mm FWHM by using appropriate diameter pipe electrodes. Figure 15 shows results of the manufacturing of a fused silica cylindrical off-axis asphere using a two-step PJM, (i) a high-rate PJM for aspherization and (ii) a low-rate high spatial resolution figure error correction. For the low etching rate regime of PJM, nonlinear effects as described do not occur.

PJP: Plasma Jet Polishing Demonstration

A third atmospheric plasma jet (APJ) processing technique in optical surface fabrication is smoothing of, e.g., ground surfaces with nearly no material removal and hence with minimum changing of the surface figure. Furthermore, due to the possible small size of the plasma jet tool, all APJ techniques can be performed very locally. This gives the opportunity either to polish small subareas or small-sized and strongly curved parts. Now, the idea for the future is to establish a cost-efficient two-step process chain consisting of precision grinding and plasma jet polishing for the manufacturing of aspheres and free forms. If higher shape accuracy is necessary, the plasma jet-based surface error correction should be included just by switching the plasma jet process from smoothing to etching. Figure 16 shows examples of plasma jet polishing of fine grounded fused silica surfaces [28].

Outlook

Ion beam etching-based surface finishing is a versatile technology with a high degree of

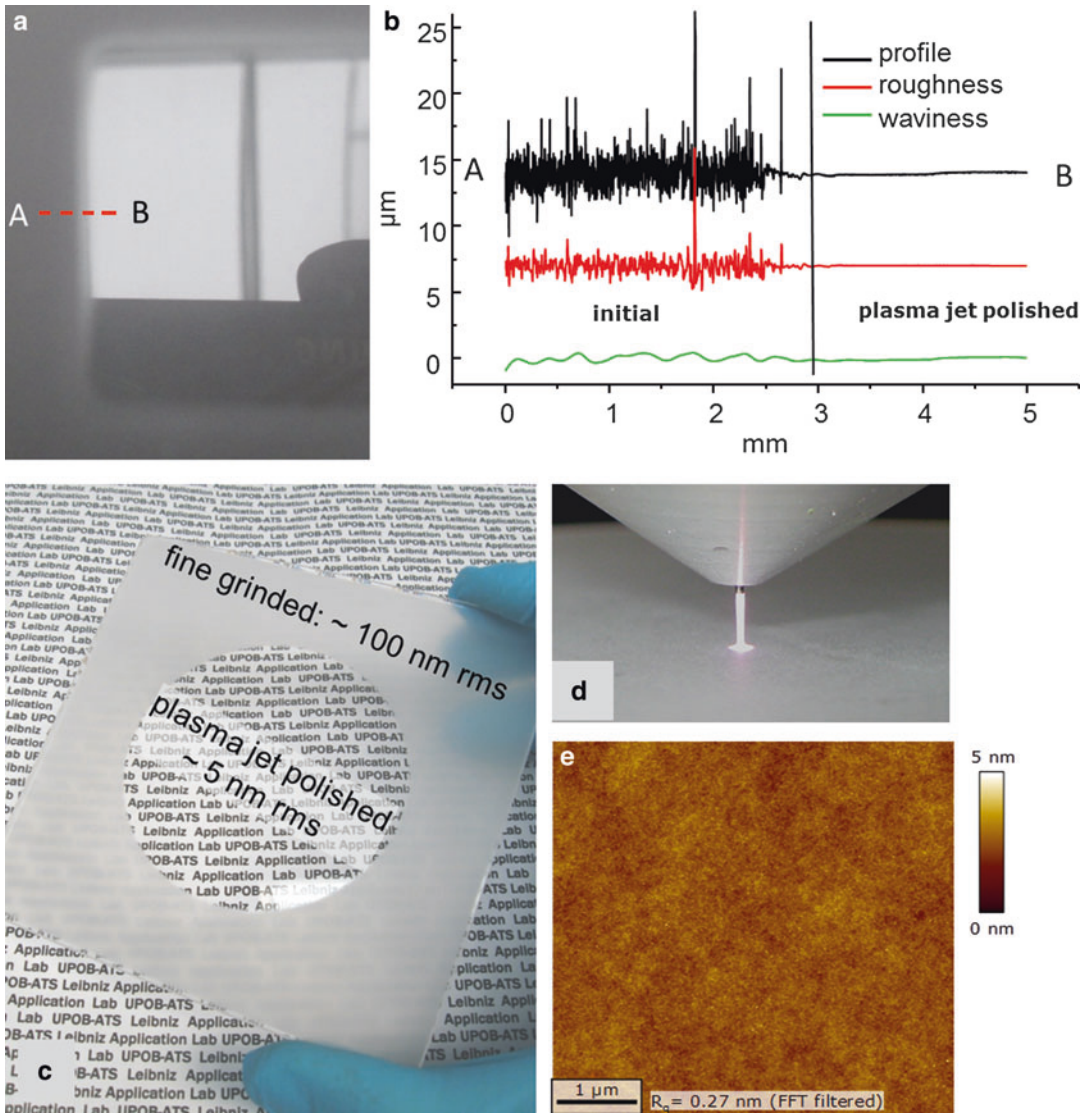


Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 15 PJM – nanometer figure error correction of a fused silica plano-elliptical mirror blank for synchrotron radiation optics, (**a**) 2D surface design profile of the optic, (**b**) and (**c**) stitching interferometer topology measurements, (**b**) initial surface after high-rate PJM figuring and post polishing of clear aperture of 130 × 25 mm^2, (**b**) after two plasma jet correction cycles with a 1 mm FWHM PJ-tool, (**d**) center profile lines of the measurement of (**b**) and (**c**) [26]

predictability due to the high stability of state-of-the-art ion sources and the acquired knowledge of the physics of beam surface interaction. The independent control of the ion energy and the ion current density over wide ranges and the possible additional use of chemical reactive species in combination with physical sputter removal allow solving tasks in a wide variety of applications of ultra-precision surface figure and patterns.

The developed techniques, the productive machines, and the finishing results show that ion beam etching technologies are advanced and powerful tools for present and future surface processing and finishing in ultra-precision technology. Ongoing developments on ion beam finishing technology will concentrate on further improvement of accuracy, reliability, new applications, and materials. Presently higher spatial resolution figuring is a challenging target just as the introduction of the ion beam smoothing techniques in production chains. New supplementary approaches will possibly further improve the performance of the technologies. One example is pulse width modulation ion sources combined with standard continuous beam to extend the dynamic range of the dwell times by a controlled reduction of the average ion beam power. Thus, two orders of magnitude and more dwell time reduction should be possible with an apparent benefit for the removal of very small amounts of material as in cases of small beam sizes. Further, locally controlled beam switching on and off in situ during dwell time processing allows figuring of separate areas of a surface, leaving other parts unaffected. This allows to generate different shapes on one surface, enabling multifunctional ultra-precision surfaces. Another new approach is ion beam figuring by dwell time-controlled deposition of materials using ion beam sputter deposition [29] instead of removal. Targets of ongoing development of pattern transfer are experimental data-backed process modeling for future adaptive RIBE processing to enhance versatility and reliability on the one hand and sub-aperture RIBE by scanning large surfaces like IBF technique on the other.

During the last years new research fields with high technological potential for precision surface processing emerged in ion beam and in plasma jet techniques. Related keywords are ion



Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies, Fig. 16 Demonstration of the efficient polishing effect by a plasma jet. (**a**) Plasma polished part of Yb-doped SiO_2. (**b**) Laser profiler measurement data along track "A –B" shown in **a**) Initial: Ra = 443.0 nm; plasma jet polished: Ra = 1.3 nm (white light interferometer, magnification 20×); *black curve*: raw data; red and green curve, low pass filtered data. (**c**) Plasma polishing of a fine ground fused silica wafer (~100 nm rms), polished area Ø 60 mm (~5 nm rms). (**d**) Microwave (2.45 GHz)-powered atmospheric plasma jet source operating in cw mode for plasma polishing and (**e**) AFM measurement of a plasma jet polished fine ground fused silica surface showing a roughness of 0.3 nm rms on a measuring area of 5 µm × 5 µm [28]

beam-induced self-organized nanostructures, ion beam-assisted growth of sculptured thin films, damage-free plasma jet figuring, and plasma jet sacrificial oxidation [30], to mention some of the activities. Several research groups in the world from academia and industry are focused on these fields on both basic and technology-oriented research and development for further progress in atomic beam-related technologies for ultra-precision surface processing.

Cross-References

- Atomic Force Microscopy
- Dry Etching Processes
- EUV Lithography
- Nanotechnology
- Optical Techniques for Nanostructure Characterization
- Self-Assembly of Nanostructures
- Ultraprecision Machining (UPM)

References

1. Taniguchi, N.: On the basic concept of 'nano-technology. Proceedings of the International Conference on Production Engineering, p. 18. Japan Society of Precision Engineering, Tokyo, Part II (1974)
2. Taniguchi, N. (ed.): Nanotechnology: Integrated Processing Systems for Ultra-Precision and Ultra-Fine Products. Oxford University Press, Oxford (1996)
3. Stowers, I. F., Komanduri , R., Baird, E. D.: Review of precision surface generating processes and their potential application to the fabrication of large optical components, Proc. SPIE 0966, Advances in Fabrication and Metrology for Optics and Large Optics, **62** (1989); doi:10.1117/12.948050; http://dx.doi.org/10.1117/12.948050
4. Jacobs, S.D.: International innovations in optical finishing. Proc. SPIE **5533**, 264 (2004)
5. Allen, D.M., Shore, P., Evans, R.W., Fanara, C., O'Brien, W., Marson, S., O'Neill, W.: Ion beam, focused ion beam, and plasma discharge machining. CIRP Ann. Manuf. Technol. **58**(2), 647–662 (2009). doi:10.1016/j.cirp.2009.09.007
6. Brinksmeier, E.: Ultraprecision machining. In: Laperrière, L.., Reinhart, G. (eds.) CIRP Encyclopedia of Production Engineering: SpringerReference (www.springerreference.com). Springer, Berlin/Heidelberg (2013)
7. Shore, P., Morantz, P.: Ultra-precision: enabling our future. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **370**(1973), 3993–4014 (2012). doi:10.1098/rsta.2011.0638
8. Fang, F., Xu, Z.: State-of-the-Art for nanomanufacturing using ion beam technology. In: Nee, A. (ed.) Handbook of Manufacturing Engineering and Technology: SpringerReference (www.springerreference.com). Springer, Berlin/Heidelberg (2013)
9. Li, W., Gu, C.: Ion beam instruments used for nanomanufacturing. In: Nee, A. (ed.) Handbook of Manufacturing Engineering and Technology: SpringerReference (www.springerreference.com). Springer, Berlin/Heidelberg (2013)
10. Xu, Z., Fang, F., Zeng, G.: Focused ion beam nanofabrication technology. In: Nee, A. (ed.) Handbook of Manufacturing Engineering and Technology: SpringerReference (www.springerreference.com). Springer, Berlin/Heidelberg (2013)
11. Taylor, John S., Sommargren, Gary E., Sweeney, Donald W., Hudyma, Russell M.: Fabrication and testing of optics for EUV projection lithography, Proc. SPIE 3331, Emerging Lithographic Technologies II, **580** (1998); doi:10.1117/12.309619; http://dx.doi.org/10.1117/12.309619
12. Dinger, U., Seitz, G., Schulte, S., Eisert, F., Muenster, Ch., Burkart, S.S. et al.: Fabrication and metrology of diffraction limited soft x-ray optics for the EUV microlithography. In: Ali, M., Khounsary, A.M., Dinger, U., Ota, K. (eds.) Proceedings of SPIE 5193, Advances in Mirror Technology for X-Ray, EUV Lithography, Laser, and Other Applications, pp. 18–28. (2004). doi:10.1117/12.511489
13. Hansen, H.N., Carneiro, K., Haitjema, H., De Chiffre, L.: Dimensional micro and nano metrology. CIRP Ann. Manuf. Technol. **55**(2), 721–743 (2006). doi:10.1016/j.cirp.2006.10.005
14. Leach, R. (ed.): Optical Measurement of Surface Topography. Springer, Berlin/Heidelberg. (2011). doi:10.1007/978-3-642-12012-1
15. Xu, C., Aissaoui, I., Jacquey, S.: Algebraic analysis of the Van Cittert iterative method of deconvolution with a general relaxation factor. J. Opt. Soc. Am. A **11**(11), 2804 (1994). doi:10.1364/JOSAA.11.002804. https://en.wikipedia.org/wiki/Richardson%E2%80%93Lucy_deconvolution; https://en.wikipedia.org/wiki/Wiener_deconvolution
16. Xie, X., Li, S.: Ion beam figuring technology. In: Nee, A. (ed.) Handbook of Manufacturing Engineering and Technology: SpringerReference (www.springerreference.com). Springer, Berlin/Heidelberg (2013)
17. Baglin, J.E.E.: Ion beam nanoscale fabrication and lithography – a review. Appl. Surf. Sci. **258**(9), 4103–4111 (2012). doi:10.1016/j.apsusc.2011.11.074
18. Johnson, L.F., Ingersoll, K.A.: Appl. Opt. **22**, 1165 (1983). doi:10.1364/AO.22.001165
19. Frost, F., Ziberi, B., Schindler, A., Bigl, F.: Surface engineering with ion beams: from self-organized nanostructures to ultra-smooth surfaces. Appl. Phys. A 91: Mater. Sci. Process. **551**(2008). doi:10.1007/s00339-008-4516-0
20. Arnold, T., Boehm, G., Fechner, R., Meister, J., Nickel, A., Frost, F., Haensel, T., Schindler, A.: Ultra-precision surface finishing by ion beam and plasma jet techniques-status and outlook. Nucl. Inst. Methods Phys. Res. Sect. A **616**(2–3), 147–156 (2010). doi:10.1016/j.nima.2009.11.013
21. Frost, F., Fechner, R., Ziberi, B., Voellner, J., Flamm, D., Schindle, A.: Large area smoothing of surfaces by ion bombardment: fundamentals and applications. J. Phys. Condens. Matter **21**(22) (2009). doi:10.1088/0953-8984/21/22/224026

22. Yamada, I.: Historical milestones and future prospects of cluster ion beam technology. Appl. Surf. Sci. **310**, 77–88 (2014). doi:10.1016/j.apsusc.2014.03.147
23. Laermer, F., Franssila, S., Sainiemi, L., Kolari, K.: Deep reactive Ion etching. In: Handbook of Silicon Based MEMS Materials and Technologies, pp. 349–374. Elsevier (2010). http://linkinghub.elsevier.com/retrieve/pii/B9780815515944000231
24. Vawter, G.A.: Ion beam etching of compound semiconductors. In: Shul, R.J., Pearton, S.J. (eds.) Handbook of Advanced Plasma Processing Techniques, p. 507 ff. Springer Science & Business Media (2000)
25. Bollinger, L.D., Zarowin, Ch.B.: Rapid, nonmechanical, damage-free figuring of optical surfaces using plasma-assisted chemical etching (PACE): part I experimental results. In: Arnold, J.B., Parks, R.E. (eds.)Proceedings of SPIE, Advances in Fabrication and Metrology for Optics and Large Optics, vol. 966, pp. 82–90. (1989). doi:10.1117/12.948052
26. Arnold, T., Böhm, G., Paetzelt, H.: Ultra-precision surface machining with reactive plasma jets. Contrib. Plasma Phys. **54**(2), 145–154 (2014). doi:10.1002/ctpp.201310058
27. Meister, J., Arnold, T.: New process simulation procedure for high-rate plasma jet machining. Plasma Chem. Plasma Process **31**, 91–107 (2011)
28. Arnold, Th., Böhm, G., Paetzelt, H.: Plasma jet polishing of rough fused silica surfaces. In: Conference Proceedings of the 13th International Conference of the EUSPEN V2, pp. 19–22 (2013)
29. Liao, W., Dai, Y., Xie, X., Zhou, L.: Combined figuring technology for high-precision optical surfaces using a deterministic ion beam material adding and removal method. Opt. Eng. **52**(1), 010503 (2013). doi:10.1117/1.OE.52.1.010503
30. Yamamura, K., Sano, Y.: Plasma-based nanomanufacturing under atmospheric pressure. In: Nee, A. (ed.) Handbook of Manufacturing Engineering and Technology: SpringerReference (www.springerreference.com). Springer, Berlin/Heidelberg (2013)

Ultrashort Carbon Nanotubes

Lesa A. Tran and Lon J. Wilson
Department of Chemistry, The Richard E. Smalley Institute for Nanoscale Science and Technology, Rice University, Houston, TX, USA

Synonyms

US-tubes

Definition

Ultrashort carbon nanotubes (US-tubes) are 20–100 nm segments of single-walled carbon nanotubes (SWNTs).

Overview

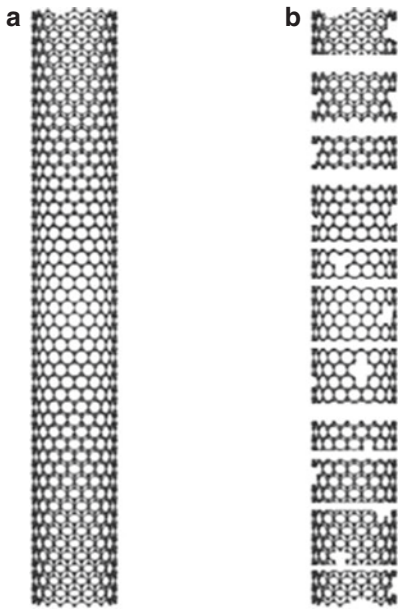
In recent years, single-walled carbon nanotubes (SWNTs) have been extensively studied for their use in biomedical applications. Owing to their peculiar physical and chemical properties, SWNTs have become a widely used platform for the design of medical therapeutic and diagnostic agents [1, 2]. Nevertheless, the biocompatibility and biodistribution of SWNTs still remain uncertain and are under current investigation. The ideal length of SWNTs for biomedical applications has yet to be determined, but it has been suggested that discrete, individualized SWNTs of shorter (<300 nm) lengths are more ideal for in vivo use [3]. Therefore, ultrashort carbon nanotubes (US-tubes), which are 20–100 nm segments of SWNTs, may be best suited for biomedical applications (Fig. 1).

Methodology

Synthesis

Chemical Cutting

Generally, the chemical cutting process of SWNTs into US-tubes can be viewed as a two-step procedure: (1) the modification of the SWNT sidewalls with functional groups or defect sites, and (2) the removal of or the cutting at these sidewall modifications. While the predominant mechanism of chemical SWNT cutting is debatable, there are two proposed methods under consideration: (1) the "fuse burning" mechanism, with carbon etching occurring at the bond-strained tube ends, and (2) defect site propagation, with cutting occurring at the random defect sites on the SWNT walls. Both oxidizing and nonoxidizing methods to chemically cut SWNTs have been explored.



Ultrashort Carbon Nanotubes, Fig. 1 Illustration of (**a**) a single-walled carbon nanotube and (**b**) a collection of ultrashort carbon nanotubes (with sidewall defects) derived from a single-walled carbon nanotube

Various oxidative processes using different strong oxidizing agents (HNO_3, H_2SO_4, and H_2O_2) have been proposed for the cutting of SWNTs. However, the efficiency and length distributions of these synthetic methods vary with the type and concentration of the oxidizing agents used, the reaction temperature and time, the method of dispersion during the reaction (refluxing or sonication), and the length and diameter of the SWNTs being cut. Although more severe reaction conditions result in more reaction exposure of SWNTs and, thus, shorter lengths of resulting US-tubes, less severe reaction conditions are preferred in order to minimize amorphous carbon production.

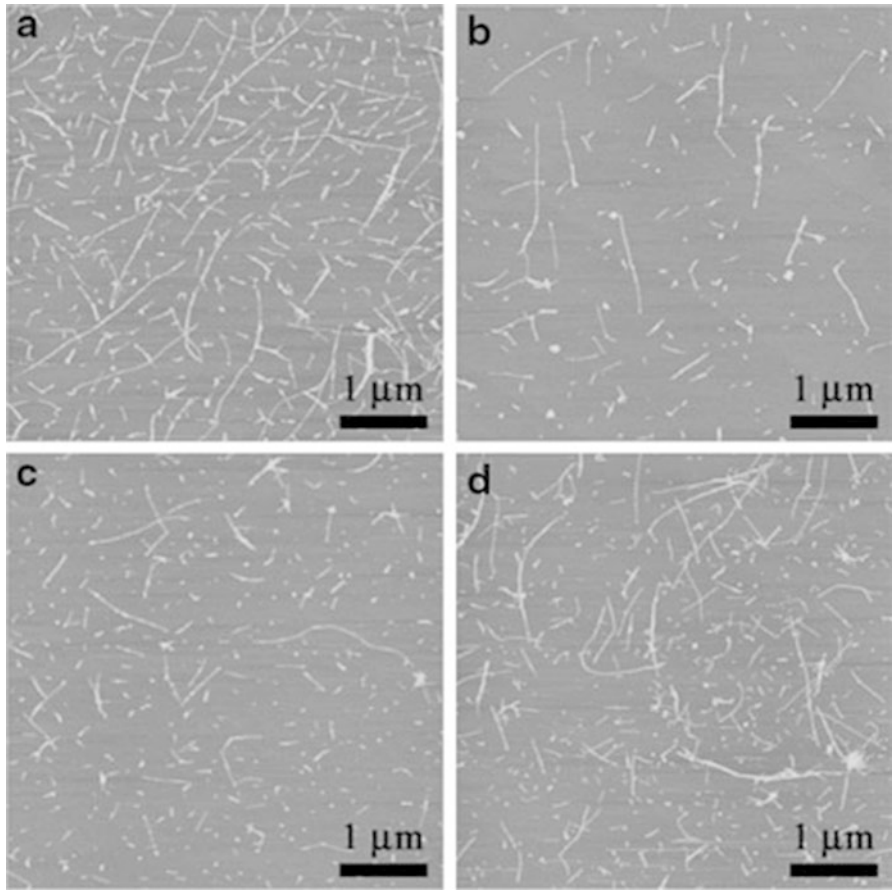
For instance, it has been shown that 4:1 (vol/vol 96 % H_2SO_4/30 % H_2O_2) piranha solutions are able to attack existing damage sites on the nanotube sidewalls to chemically shorten SWNTs [4]. While hot (70 °C) piranha solutions can induce more damage sites, room-temperature (22 °C) solutions can react with existing sidewall defects but are incapable of initiating further damage sites, resulting in minimal carbon loss and slow etch rates (Fig. 2). On the other hand, high-temperature conditions produce increasingly shorter US-tubes with increasing reaction time, with approximately half of the starting material destroyed due to chemical etching. Therefore, room-temperature piranha solutions offer the ability to exploit active damage sites in the sidewalls of the nanotube in a controlled manner without the counterproductive destruction of the nanotubes.

The aggressive oxidation of SWNTs with oleum (100 % H_2SO_4 with 20 % SO_3) and nitric acid can simultaneously shorten and carboxylate the SWNTs, rendering water-soluble US-tubes less than 60 nm in length [5]. Purified SWNTs are first dispersed in oleum to disentangle individual SWNTs from their rope-like bundles. Next, the SWNT-oleum dispersion is extensively oxidized with strong HNO_3 at elevated temperatures to cut the SWNTs into US-tubes. The oxidation of the SWNTs also forms carboxylic acid groups at the sidewall defects of the SWNTs, making US-tubes that are up to 2 wt% soluble in polar organic solvents, acids, and water.

One nonoxidative process to synthesize US-tubes involves the chemical shortening of SWNTs to lengths of less than 50 nm by a fluorination/pyrolysis treatment [6]. Briefly, purified SWNTs are exposed to fluorine gas, during which fluorine atoms tend to arrange around the circumference of the nanotube to form band-like regions along the nanotube sidewalls. Upon pyrolysis of the fluorinated SWNTs at 1000 °C in an inert atmosphere, the chemically bound fluorine atoms are driven off in the form of CF_4 and COF_2. This leaves behind chemically cut US-tubes with partial sidewall defects along the nanotube axis.

Mechanical Cutting

There have been various documented methods to mechanically cut SWNTs into shorter segments, including ultrasonication, grinding, and ball milling. However, these techniques lack length control; such mechanical methods create extensive sidewall

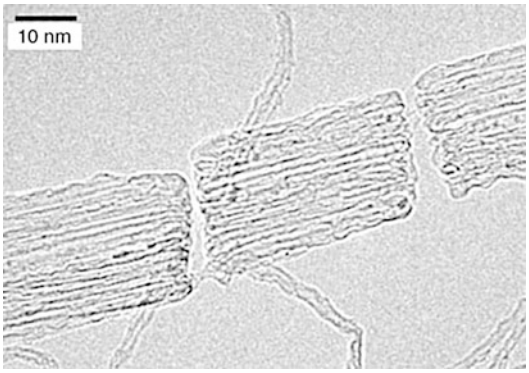


Ultrashort Carbon Nanotubes, Fig. 2 Atomic force microscopy (AFM) images of (**a**) purified SWNTs and SWNTs after piranha treatment at 22 °C for (**b**) 1 h, (**c**) 3 h, and (**d**) 9 h (Reproduced with permission from Ref. [4])

damage and are not able to achieve sub-100 nm lengths. Therefore, more precise mechanical cutting techniques have been developed.

Microlithography can be used to cut SWNTs into US-tubes [7]. This technique involves a protective photoresist polymer layered on top of a SWNT-covered solid substrate in a patterned manner to protect portions of nanotubes from lithographic damage and modification. Reactive ion etching using an oxygen plasma is then used to remove the unprotected sections of the SWNTs. After etching, the newly formed nanotube ends remain, with dangling carboxylic acid and ether groups. The photoresist polymer can then be removed, leaving cut segments of SWNTs. Regardless of whether the SWNTs are previously aligned or properly dispersed prior to cutting, microlithography offers a straightforward method to shorten SWNTs with a narrow and selective length distribution.

Electron beam etching has also been used to cut SWNTs with nanometer precision [8]. Using electron energy beams exceeding 100 keV, which is made possible by currently available electron microscopes, partial and complete cuts into SWNTs can be made with great control without the need of covering portions of SWNTs with a protective layer. Similar methods have been developed using high voltage-bearing scanning tunneling microscope (STM) tips. Interestingly, upon cutting into a SWNT bundle, the newly formed tube ends immediately form hemispherical



Ultrashort Carbon Nanotubes, Fig. 3 Transmission electron microscopy (TEM) image of the complete cutting of a SWNT bundle into US-tubes using electron beam etching (Reproduced with permission from Ref. [8])

caps, enhancing the stability of the cut US-tubes (Fig. 3). Another peculiar observation is that forming a second cut within 10–50 nm from a preexisting cut proves more difficult to perform. This increase in stability is thought to be a result of the generation and annealing of vacancy-interstitial atom pairs.

More recently, an ultramicrotome has been used to cut frozen layers of magnetically aligned SWNT membranes, or Buckypapers [9]. Enough mechanical control was obtained to create open-ended US-tubes with minimal sidewall damage, typically seen after chemical oxidative methods. When the Buckypaper was cut into 50 nm segments, the mean length was about 87 nm, with about 60 % of the tubes ranging from 50 to 150 nm. However, the large pressure exerted by the diamond knife tip in its proximity can cause deformation and cracking of the nanotubes.

Individualization of US-tubes

Like their SWNT precursors, US-tubes have a large tendency to bundle together, which renders it difficult to form single-tube suspensions. Although it is possible for full-length SWNTs to disperse in solution with the aid of polymers and surfactants, US-tubes cannot disperse as readily in these same conditions [10]. Because of their high aspect ratios, SWNTs are relatively flexible and can "peel" more readily from each other upon introduction to surfactant wrapping. On the other hand, US-tubes have a much lower aspect ratio, therefore making them more rigid and more difficult for surfactant molecules to insert between individual US-tubes. Therefore, chemical means to debundle and individualize US-tubes have been explored to better disperse US-tubes.

Reduction of US-tubes

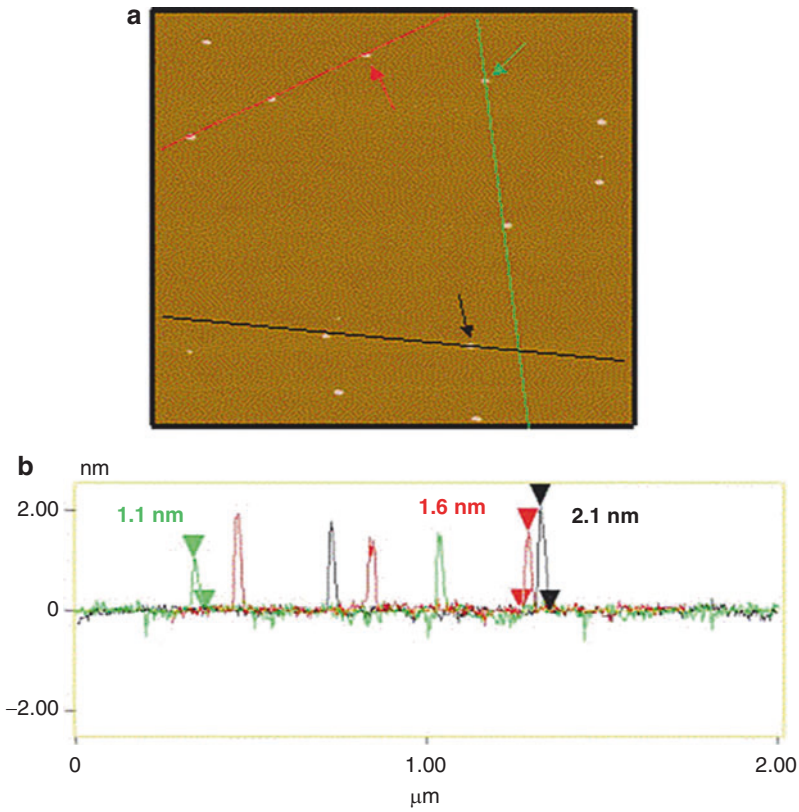
Chemical reduction can be used to individualize, or debundle, US-tubes [10]. Upon sonication with K° metal in THF, the negative charge imparted on the US-tubes electrostatically overcomes the bundling forces to yield individualized US-tubes (Fig. 4). It is thought that the individualization of US-tubes via chemical reduction increases US-tube dispersion in solution, thus allowing for more even chemical functionalization.

Functionalization of US-tubes

The chemical functionalization of US-tubes is thought to occur at carboxylic acid end groups of the sidewall defects. This allows for the attachment of various chemical groups that would render the US-tubes more soluble in different solvent systems. One functionalization method utilizes an in situ Bingel-Hirsch (CBr_4/DBU) cyclopropanation to functionalize US-tubes with the water-soluble malonic acid bis-(3-tert-butoxycarbonylaminopropyl)ester to yield four to five adducts per nm [10]. Another documented functionalization method involves cyclopropanation using diazoacetic ester in the presence of $Rh_6(CO)_{16}$ [11]. This allows for the US-tube to be readily reactive with primary amines, thus making possible the functionalization of US-tubes with a number of amino acid derivatives, such as water-soluble serine, with 3–19 groups per nm. Other biologically relevant moieties, such as antioxidants [12] and oligonucleotides [13], have been successfully attached onto US-tubes and are described below. Alkylation-based reduction can also be performed on US-tubes to improve nanotube dispersion in nonpolar environments [14].

Loading of US-Tubes

In addition to being able to chemically functionalize the sidewalls of US-tubes, the



Ultrashort Carbon Nanotubes, Fig. 4 AFM image of individualized US-tubes after chemical reduction (Reproduced with permission from Ref. [10])

hollow interior of the US-tubes can be loaded or filled with various ions and small molecules, such as Gd^{3+} ions [15], molecular I_2 [16], and ^{211}AtCl radionuclides [17]. This property is especially relevant for drug design, as the carbon structure of the US-tube serves as an encapsulating sheath to protect potentially toxic moieties from their surrounding environment. Although the loading mechanism is not completely clear, it is likely that ions and small molecules load into the US-tubes through the sidewall defects.

Phenomena

Toxicology

Before US-tubes are considered for biomedical applications, their toxicological profiles must be known. A recent study documented the toxicology as well as the acute and subchronic effects of high doses (50–1000 g/kg bodyweight) of Tween-suspended US-tubes in Swiss mice [3]. Oral administration of US-tubes did not result in death or any behavioral abnormalities for up to 14 days. Intraperitoneal administration allowed for the US-tubes to reach systemic circulation and various tissues through the lymphatic system. As opposed to unfunctionalized SWNTs, well-individualized US-tubes had the ability to pass through the reticuloendothelial system, excreting via the kidneys and bile ducts. High doses of US-tubes can induce strongly adherent granuloma formation on and inside organs due to large (>10 μm) nanotube aggregation into fiber-like structures. However, small (<2 μm) aggregates of US-tubes can readily enter various cell types without granuloma formation or any sign of toxicity. Neither death nor growth or behavioral abnormalities were observed for the animals used in the experiment.

Applications

Diagnostic Agent Design

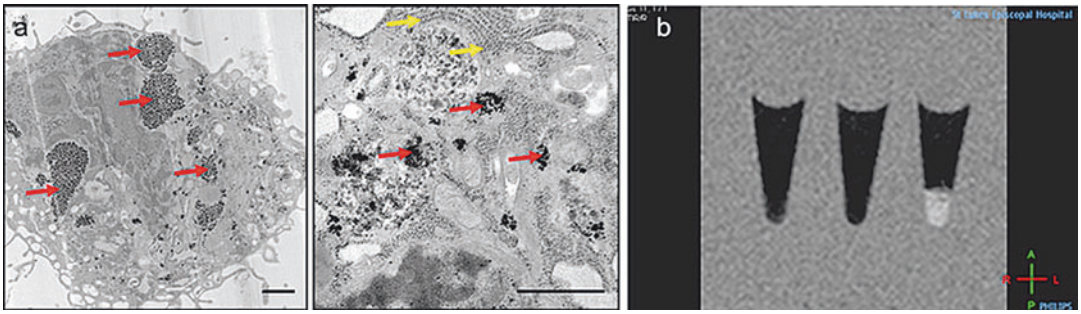
Magnetic Resonance Imaging (MRI) Contrast Agents

Gadonanotubes, or US-tubes loaded with Gd^{3+} ions, have been shown to outperform currently used clinical T_1-weighted MRI contrast agents by approximately 40-fold [15]. Although the exact loading mechanism is unclear, it is likely that the Gd^{3+} ions enter the US-tubes through the sidewall defect sites, where they form nanoscale clusters. Because of the lipophilicity of the carbon sidewalls and their short lengths, Gadonanotubes can also readily translocate across cell membrane, allowing for Gadonanotubes to become an effective T_1-weighted cellular MRI contrast agent. It has already been shown that Gadonanotubes can be readily internalized and imaged inside a variety of cells, including mesenchymal stem cells (Fig. 5) [18].

In the absence of Gd^{3+} ions, the US-tubes alone can perform as a T_2-weighted MRI contrast agent [19]. The US-tubes are approximately three times more efficacious than Ferumoxtran®, a clinically used Fe_3O_4-based contrast agent. Although the US-tubes contain much less (<1 %) iron catalyst than unpurified SWNTs, the US-tubes exhibit a much higher *r2* performance, suggesting that the carbon nanotube material itself is contributing to the superparamagnetic nature and that the iron particles are not the only contributing factor to the superparamagnetic signature of the US-tubes. The shorter lengths and the sidewall defects of the Gadonanotubes may also contribute to the high performance. This also implies that Gadonanotubes can concurrently perform as both a T_1- and T_2-weighted MRI contrast agents. Gadonanotubes have previously been used as a T_2-weighted contrast agent in vivo, with r_2 relaxivities as high as 578 mM^{-1} s^{-1} at 7 T [20]. This suggests that the Gadonanotubes are useful T_1 positive agents at lower magnetic fields and T_2 negative agents at higher fields.

Computed Tomography (CT) Agents

Molecular I_2 has also been loaded into US-tubes for CT X-ray imaging contrast agent design [16]. After being functionalized with serinol amide groups via the Bingel-Hirsch reaction, the US-tubes were filled with gaseous I_2 at 100 °C, then washed with ethanol and reduced with NaH to remove excess I_2 from the nanotube exterior. A 10 wt% loading of I_2 was observed for the US-tubes. Using XPS and X-ray-induced Auger emission spectroscopy, it was confirmed that the molecular I_2 was sequestered within the US-tubes. Micro-CT images were obtained to show that the iodine-loaded US-tubes were more X-ray opaque. These serinol amide-functionalized, I_2-loaded US-tubes are the first water-soluble CT contrast agent derived from SWNT materials.



Ultrashort Carbon Nanotubes, Fig. 5 (**a**) TEM images of a Gadonanotube-labeled mesenchymal stem cell (MSC). *Red arrows* point to the Gadonanotube aggregates. *Yellow arrow* points to the ribosomes of the endoplasmic reticulum. (**b**) T_1-weighted MR images at 1.5 T and 25 C of (*left* to *right*) unlabeled MSCs, Gd-DTPA-labeled MSCs, and Gadonanotube-labeled MSCs at T_i = 500 ms. Scale bar = 1 μm (Adapted with permission from Ref. [18])

Therapeutic Agent Design

Radionuclide Containment and Delivery

Similar to the loading of Gd^{3+} ions and I_2 molecules into US-tubes, $^{211}At^-$ ions can be loaded and oxidized into $^{211}AtCl$ within US-tubes to contain and deliver the alpha-emitting radionuclides in the nanotube capsule [17]. Upon oxidizing the $^{211}At^-$ ions with chloramine-T or *N*-chlorosuccinimide, the AtCl molecules labeled the US-tubes to a great extent, with 91.3 % by aqueous loading. Subsequent PBS washing and serum exposure removed excess AtCl residing on the exterior or at the sidewall defects of the nanotubes, but the US-tubes exhibit reasonable stability toward physiological challenge.

Bone Tissue Scaffold Composites

US-tubes have also been used to dope the highly porous (poly(propylene fumarate)) polymer for bone tissue scaffold engineering [14]. These composites are injectable, thermally cross-linkable, and cytocompatible, making them ideal for trabecular bone tissue scaffolding biomaterials. While high (75–90 vol.%) porosity is preferred to allow for cell adhesion, proliferation, and vascularized tissue ingrowth, the mechanical properties of highly porous polymer scaffolds become compromised. Therefore, US-tubes have been added to PPF polymer scaffolds to improve the mechanical properties. It was found that dodecylated US-tubes functionalized with dodecyl groups better dispersed into PPF, improving their reinforcing effects. While there was no significant difference in porosity, pore size, and pore interconnectivity among US-tube-doped PPF and PPF alone, the compressive modulus, offset yield strength, and compressive strength of the dodecylated US-tube nanocomposites were higher than or similar to that of PPF alone. The US-tube/PPF nanocomposite has also been shown to be osteoconductive, allowing marrow stromal cells to attach and proliferate readily on the scaffold.

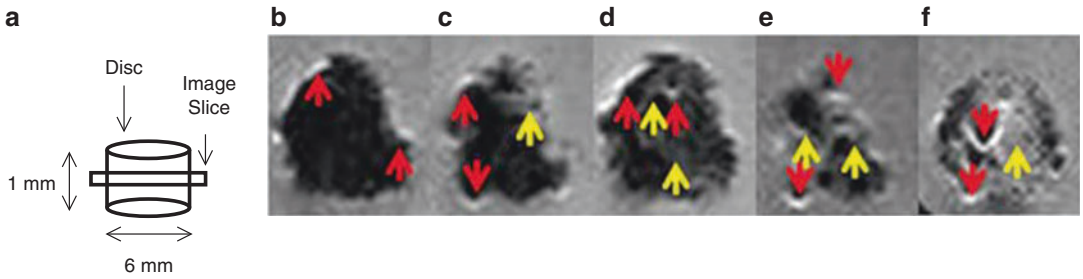
Gadonanotubes have also been incorporated into poly(lactic-co-glycolic acid) (PLGA) polymer scaffolds in order to explore the degradation process of the scaffold and the biodistribution of the Gadonanotubes upon their release from the polymer matrix in vivo [20]. T_2-weighted MR imaging of the Gadonanotube-PLGA nanocomposites and their degradation process were obtained in vitro, showing the strong influence of the Gadonanotubes on surrounding water proton relaxation (Fig. 6). In addition to providing mechanical reinforcement from the US-tube structure, the Gadonanotube-reinforced PLGA scaffolds may elucidate the mechanism of the biodegradation of the polymer and the release and biodistribution of the Gadonanotubes in vivo.

Free Radical Scavengers

US-tubes have also been functionalized as a scaffold for the phenolic antioxidant, butylated hydroxytoluene (BHT) using DCC/DMAP coupling [12]. As expected, the higher amount of BHT functionalization (either via covalent or noncovalent interactions) to the carboxylic acid groups on the oxidized US-tubes corresponded to higher overall antioxidant activity, with performances more than 200 times better than that of the radioprotective dendritic fullerene, DF-1. Strikingly, PEGylated US-tubes alone without any BHT functional groups are extremely effective antioxidants, performing nearly 40 times greater than DF-1 in oxygen scavenging ability. Even though the US-tubes have heavily compromised sidewalls from the oxidative cutting process, US-tubes are still able to sequester oxygen radicals due to the sp^2-hybridized carbon nanotube framework.

Oligonucleotide Delivery

US-tubes have also served as delivery vehicles for oligonucleotide (ODN) decoys, which are synthetic ODNs containing the DNA-binding sequence of a transcription factor [13]. Because US-tubes can inherently translocate into cells, US-tubes can efficiently facilitate the intracellular delivery of ODN decoys into target cells so that ODN decoys can better target transcription factors of interest for gene therapy. Specifically, US-tubes have successfully delivered a decoy ODN against nuclear factor-κB (NF-κB), which is a transcription factor that regulates genes important in immune and inflammatory responses. Using carbodiimide chemistry, the amino groups at the



Ultrashort Carbon Nanotubes, Fig. 6 (**a**) Schematic of the nanocomposites sample arrangement within the MR-I. Representation two-dimensional images through a nanocomposite disc after (**b**) 2 h, (**c**) 24 h, (**d**) 3 days, (**e**) 5 days, and (**f**) 7 days. The higher (*white*) pixel intensities within the dark disc represent regions of higher water concentration. The high intensity surrounding the disc is the agarose gel (TE = 15 ms) (Reproduced with permission from Ref. [20])

5′ end of the double-stranded ODN was covalently coupled to the carboxylic acid groups of the oxidized US-tubes. To test the performance of the US-tubes as a delivery vehicle, NF-κB-dependent gene expression was significantly downregulated in monocyte-derived human macrophages incubated with ODN-functionalized US-tubes, as compared to macrophages incubated with US-tubes functionalized with nonspecific ODN decoys.

Future Directions

While there have been significant strides made in the development and application of US-tubes, many challenges still must be addressed. In order to produce more uniform US-tubes from SWNTs, better size control during the cutting procedure and size sorting after synthesis must be achieved. Additionally, improving the solubility and dispersion of US-tubes in biological media must be achieved to fully realize their potential in biomedical applications.

Taking advantage of the exterior functionalization and the interior loading capabilities of US-tubes, a nearly limitless number of possibilities abound for advanced US-tube-based therapeutic and diagnostic agent design. US-tubes can either be used as a nanocapsule, sequestering biologically relevant ions or molecules that may be inherently toxic if not sequestered, or as a nanovector, delivering various drug payloads or peptide sequences to target cells. Nevertheless, much work still needs to be completed to better characterize the acute effects, long-term toxicity, and biodistribution of US-tubes, as well as other types of carbon nanostructures, in living systems.

Cross-References

▶ Carbon Nanotubes
▶ Fullerenes for Drug Delivery

References

1. Kostarelos, K., Bianco, A., Prato, M.: Promises, facts and challenges for carbon nanotubes in imaging and therapeutics. Nat. Nanotechnol. **4**, 627–633 (2009)
2. Liu, Z., Tabakman, S., Welsher, K., Dai, H.: Carbon nanotubes in biology and medicine: in vitro and in vivo detection, imaging and drug delivery. Nano Res. **2**, 85–120 (2009)
3. Kolosnjaj-Tabi, J., Hartman, K.B., Boudjemaa, S., Ananta, J.S., Morgant, G., Szwarc, H., Wilson, L.J., Moussa, F.: In vivo behavior of large doses of ultrashort and full-length single-walled carbon nanotubes after oral and intraperitoneal administration to Swiss mice. ACS Nano **4**(3), 1481–1492 (2010)
4. Ziegler, K.J., Gu, Z., Peng, H., Flor, E.L., Hauge, R. H., Smalley, R.E.: Controlled oxidative cutting of single-walled carbon nanotubes. J. Am. Chem. Soc. **127**(5), 1541–1547 (2005)
5. Chen, Z., Kobashi, K., Rauwald, U., Booker, R., Fan, H., Hwang, W.-F., Tour, J.M.: Soluble ultra-short single-walled carbon nanotubes. J. Am. Chem. Soc. **128**(32), 10568–10571 (2006)
6. Gu, Z., Peng, H., Hauge, R.H., Smalley, R.E., Margrave, J.L.: Cutting single-wall carbon nanotubes through fluorination. Nano Lett. **2**(9), 1009–1013 (2002)

7. Lustig, S.R., Boyes, E.D., French, R.H., Gierke, T.D., Harmer, M.A., Hietpas, P.B., Jagota, A., McLean, R. S., Mitchell, G.P., Onoa, G.B., Sams, K.D.: Lithographically cut single-walled carbon nanotubes: controlling length distribution and introducing end-group functionality. Nano Lett. **3**(8), 1007–1012 (2003)
8. Banhart, F., Li, J., Terrones, M.: Cutting single-walled carbon nanotubes with an electron beam: evidence for atom migration inside nanotubes. Small **1**(10), 953–956 (2005)
9. Wang, S., Liang, Z., Wang, B., Zhang, C., Rahman, Z.: Precise cutting of single-walled carbon nanotubes. Nanotechnology **18**(5), 055301 (2007)
10. Ashcroft, J.M., Hartman, K.B., Mackeyev, Y., Hofmann, C., Pheasant, S., Alemany, L.B., Wilson, L.J.: Functionalization of individual ultra-short single-walled carbon nanotubes. Nanotechnology **17**, 5033–5037 (2006)
11. Mackeyev, Y., Hartman, K.B., Ananta, J.S., Lee, A.V., Wilson, L.J.: Catalytic synthesis of amino acid and peptide derivatized gadonanotubes. J. Am. Chem. Soc. **131**, 8342–8343 (2009)
12. Lucente-Schultz, R.M., Moore, V.C., Leonard, A.D., Price, B.K., Kosynkin, D.V., Lu, M., Partha, R., Conyers, J.L., Wilson, L.J.: Antioxidant single-walled carbon nanotubes. J. Am. Chem. Soc. **131**, 3934–3941 (2009)
13. Crinelli, R., Carloni, E., Menotta, M., Giacomini, E., Bianchi, M., Ambrosi, G., Giorgi, L., Magnani, M.: Oxidized ultrashort nanotubes as carbon scaffolds for the construction of cell-penetrating NF-κB decoy molecules. ACS Nano **4**(5), 2791–2803 (2010)
14. Shi, X., Sitharaman, B., Pham, Q.P., Liang, F., Wu, K., Billups, W.E., Wilson, L.J., Mikos, A.G.: Fabrication of porous ultra-short single-walled carbon nanotube nanocomposite scaffold for bone tissue engineering. Biomaterials **28**, 4078–4090 (2007)
15. Sitharaman, B., Kissell, K.R., Hartman, K.B., Tran, L. A., Baikalov, A., Rusakova, I., Sun, Y., Khant, H.A., Ludtke, S.J., Chiu, W., Laus, S., Tóth, É., Helm, L., Merbach, A.E., Wilson, L.J.: Superparamagnetic gadonanotubes are high-performance MRI contrast agents. Chem. Commun. **31**, 3915–3917 (2005)
16. Ashcroft, J.M., Hartman, K.B., Kissell, K.R., Mackeyev, Y., Pheasant, S., Young, S., Van der Heide, P.A.W., Mikos, A.G., Wilson, L.J.: Single-molecule I_2@US-tube nanocapsules: a new x-ray contrast-agent design. Adv. Mater. **19**, 573–576 (2007)
17. Hartman, K.B., Hamlin, D.K., Wilbur, S., Wilson, L. J.: ^{211}AtCl@US-tube nanocapsules: a new concept in radiotherapeutic-agent design. Small **3**(9), 1496–1499 (2007)
18. Tran, L.A., Krishnamurthy, R., Muthupillai, R., Cabreira-Hansen, M.G., Willerson, J.T., Perin, E.C., Wilson, L.J.: Gadonanotubes as magnetic nanolabels for stem cell detection. Biomaterials **31**(36), 9482–9491 (2010)
19. Ananta, J.S., Matson, M.L., Tang, A.M., Mandal, T., Lin, S., Wong, K., Wong, S.T., Wilson, L.J.: Single-walled carbon nanotube materials as *T2*-weighted MRI contrast agents. J. Phys. Chem. C **113**, 19369–19372 (2009)
20. Sitharaman, B., Van Der Zande, M., Ananta, J.S., Shi, X., Veltien, A., Walboomers, X.F., Wilson, L.J., Mikos, A.G., Heerschap, A., Jansen, J.A.: Magnetic resonance imaging studies on gadonantube-reinforced biodegradable polymer nanocomposites. J. Biomed. Mater. Res. A **93A**(4), 1454–1462 (2010)

Ultrasonic Atomization

▶ Acoustic Nanoparticle Synthesis for Applications in Nanomedicine

Ultrasonic Force-Assisted Nanomachining

▶ Ultrasonic Vibration-Assisted Nanomachining

Ultrasonic Force-Regulated Nanomachining

▶ Ultrasonic Vibration-Assisted Nanomachining

Ultrasonic Machining

Suhas S. Joshi
Department of Mechanical Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India

Synonyms

Non-conventional machining; USM; Vibration assisted machining

Definition

Ultrasonic machining involves imparting ultrasonic vibrations (of frequency ~20 kHz) to a tool to effect material removal. The process is more effective on materials that have hardness more than RC 40, but it is used on almost all including metallic and nonmetallic materials such as glasses, ceramics, and composites.

Process Variants

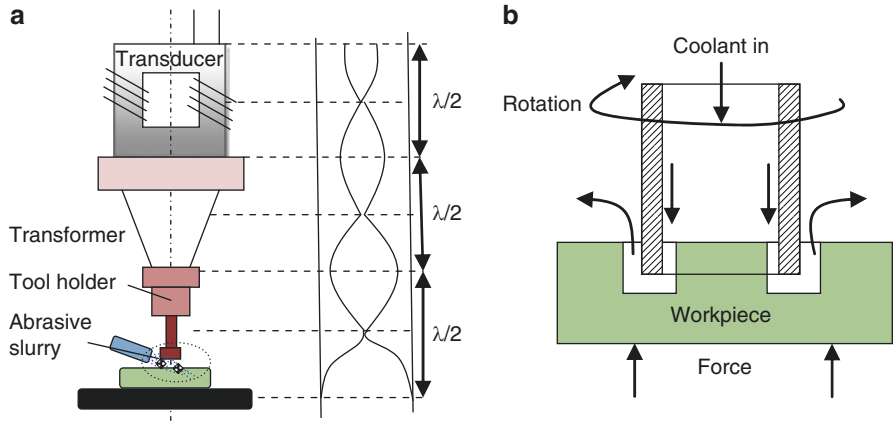
The process is configured into two main types (Fig. 1a–b) depending upon how the force is transferred from the vibrating tool to the work surface. If the transfer of vibration energy is achieved by introducing a liquid containing abrasive grits in the gap between the tool and work surface, the process is called "ultrasonic impact grinding" or simply "ultrasonic machining." On the other hand, if the abrasive grits are not present in the carrier liquid, then the tool is rotated besides being vibrated, and the process is called as "rotary ultrasonic machining" [1, 2].

Basic differences in these two processes are illustrated in Table 1.

Basic Equipment

Equipment for ultrasonic machining is available in various forms, like dedicated machines or as an attachment to other machines as well as table-top machines. Typical machine elements for performing both the ultrasonic machining (USM) and rotary ultrasonic machining (RUM) are identical. They include [1, 2]:

1. *Power Supply*: a power supply that provides electrical output at ultrasonic frequency
2. *Transducer*: a transducer that converts high-frequency electrical supply to tool displacement at ultrasonic frequency. The transducers work on magnetostrictive or piezoelectric principles. The magnetostrictive effect involves application of alternating current to a ferromagnetic coil, which induces vibrations at the applied frequency in the coil. The phenomenon allows transfer of vibrations over wider frequency band of (17–23 kHz). However, it also causes high electrical losses (e.g., eddy current loss) and has low energy conversion efficiency (~50 %). Most of the energy losses appear as heat; hence, these transducers require air or water cooling. Consequently, such systems are bulky. On the other hand, the piezoelectric transducers comprise of a stack of ceramic discs placed between a high-density material base and low-density material radiating face. The energy conversion efficiency can be in the range of (~90–96 %). Therefore, they do not require cooling.
3. *Horn*: the horn is referred with various names such as tool holder, sonotrode, stub, concentrator, or acoustic coupler. With appropriate design of horn, the vibration amplitude can be amplified by as high as 600 % over its initial value at the transducer surface of ~0.001–0.1 μm. The increase in the vibration amplitude is inversely proportional to the reduction in the area ratio between top and bottom faces of the horn. Usually, the horn is made from Monel, titanium 6–4, stainless steel, aluminum, or aluminum bronze.
4. *Tool*: the tool is shaped such that it is the inverse of the cavity to be made. The tool is attached to the horn either by brazing or soldering. This helps avoid possible fatigue failure or self-loosening of the threaded fasteners. The tool is designed such that the amplitude of vibrations is maximum at its free end. Mass and length of the tool are also important. While the bulky tools absorb ultrasonic vibrations, very long tools will flutter during the process. Therefore, tools with slenderness ratio of less than 20:1 are recommended. Often when the tools are of large cross-sectional area, the tool center faces slurry starvation. Therefore, the tools with larger perimeter for same cross-sectional area are preferred. The tools should be designed to have flutes to aid slurry flow away from the cut. Sometimes the tools are relieved



Ultrasonic Machining, Fig. 1 (**a**) Ultrasonic machining (**b**) Rotary ultrasonic machining

Ultrasonic Machining, Table 1 Comparison between the two variants of USM

Parameters	USM	RUM
Tool motions	Reciprocation	Reciprocation and rotation
Tool-work interaction	No contact	Contact
Tool materials	Tough and ductile metals, mild steel, stainless steel	Diamond tools
Tool shape	Mirror image of the desired cavity	Core drill type tool
Abrasives	Diamond, CBN, boron carbide, silicon carbide, aluminum oxide	Diamond particles impregnated in the tool
Carrier liquid functions	Carrying abrasive grits and cooling	No abrasive grits hence only cooling
Force transfer	Tool vibrations force abrasive grits to impinge on work surface	Vibrating diamond tool contacts work surface to cause removal
Removal mechanism	Indentation and chipping	Indentation and abrasion
Shapes generated	Mirrors tool shape	Holes and other axisymmetric shapes
Size of features generated	About 90 mm diameter and 64 mm deep	Maximum diameter 38–50 mm

behind the face. In the case of longer tools, availability of center holes for feeding abrasive slurry is of great advantage as it avoids sidewall friction. The tool materials should be tough and ductile and should have high wear and fatigue resistance. The tool materials usually include mild steel, stainless steel, brass, Monel, bearing steel, and molybdenum. Softer materials like aluminum or brass may face significant tool wear.

5. *Abrasives*: the abrasives transfer the force of vibrations to the work surface and impinge on the surface to perform removal operation. The abrasives material should be harder than the work material, and the size of the abrasive grits should be the same as that of the amplitude of vibrations being used. Abrasives are normally carried by slurry containing 50 % volume of abrasives and 50 % volume of water. The rate of penetration of tool in the work surface increases with increasing abrasive concentration but reaches a maximum, beyond which the jamming of the abrasives takes place, thereby the tool penetration reduces. The abrasives used for the USM process include diamond, cubic boron nitride, boron carbide, silicon carbide, and aluminum oxide. Boron carbide is the most widely used abrasive material. It is often used in the processing of tungsten carbide, ceramics, minerals, metals, and precious and semiprecious stones. Silicon carbide is used for low-density ceramics, glasses, Si, Ge, and mineral stones. Aluminum oxide is used for machining of glasses and sintered or hard powder components.

Mechanism of Removal

Typical sequence of events that occurs in ultrasonic machining includes (Fig. 2) [3]:

1. Indentation by abrasive particles leading to generation of an inelastic deformation zone around the particle.
2. At some threshold, deformation induced flow develops into a microcrack, termed as a "median" or "lateral" crack.
3. An increase in load further causes steady growth of the median crack.
4. During unloading action, the median crack begins to close inducing formation of lateral vents.
5. Upon complete unloading, the lateral vents continue their extension towards specimen surface and lead to chipping or forming a fragmented section on the work surface.

The critical load for the initiation of the median crack is given by [4]

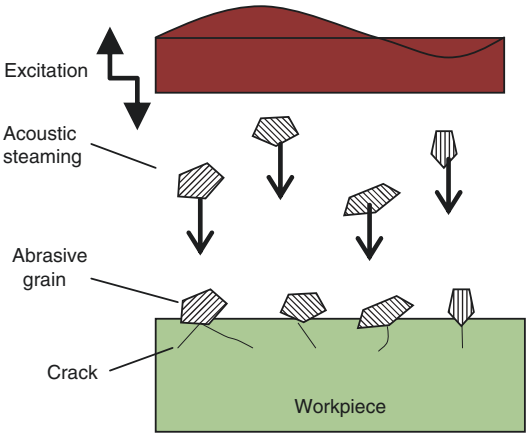
$$P_c = \alpha \frac{K_{IC}^4}{H_v^4} \tag{1}$$

where, α is a dimensionless factor. The size of the mean crack (C_s) is given by [5]

$$(C_s)^m = K \cdot P \tag{2}$$

where $m = 1$–1.5 and k is a coefficient. This indicates that the size of the median crack grows with an increase in the load and a decrease in the fracture toughness of the work material.

A simple model to evaluate material removal rate (MRR) in ultrasonic machining, which gives contribution of direct impact of the abrasive particles on the work surface in total MRR, was developed by Shaw [6]. The model is primarily based on the assumption that the work volume removed is proportional to the number of particles making impact per cycle and the volume of work material removed per particle impact. The abrasive grains are assumed to be identical and spherical in shape. The MRR therefore is given by [6]



Ultrasonic Machining, Fig. 2 Schematic of crack formation due to impact in USM

$$Z_w = k \cdot n \cdot f \left(d_g \cdot \delta\right)^{\frac{3}{2}} \tag{3}$$

where δ is the depth of indentation, n is the number of impacting particles per cycle, f frequency of ultrasonic vibrations, and k coefficient.

In the above equation, the depth of indentation is given by

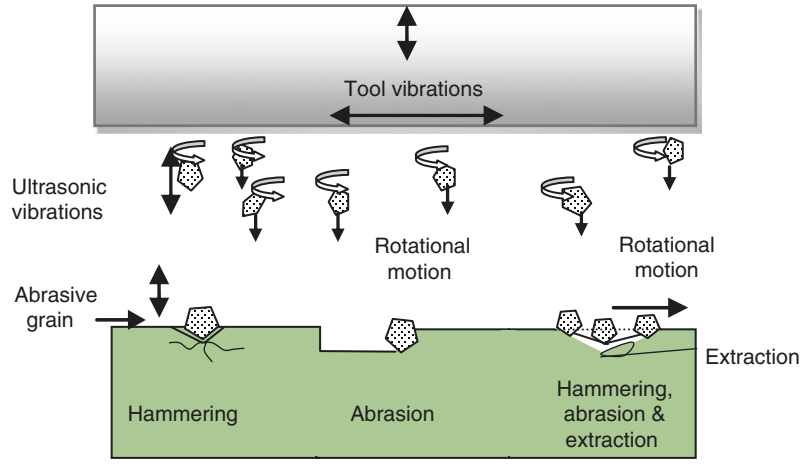
$$\delta = \left[\frac{8F_s A d_g}{\pi k_1 H_v C(1+q)}\right]^{\frac{1}{2}} \tag{4}$$

where F_s is static force, A is amplitude of ultrasonic vibration, k_1 is the constant of proportionality, H_v is the hardness of workpiece, C concentration of slurry, and q ratio of the hardness of the workpiece to that of the tool.

Therefore, substituting Eq. 3 in Eq. 4, we get, MRR for ultrasonic machining as

$$Z_w = k\,nf\left(d_g\right)^{\frac{3}{2}} \left(\frac{8F_s A d_g}{\pi k_1 H_v C(1+q)}\right)^{\frac{3}{4}} \tag{5}$$

The Shaw's model is the most widely used because of its simplicity, despite its limitations. It predicts monotonous increase in the MRR with an increase in static force, F_s. However, in practice, this is not so. After a certain increase in the static force, the abrasive particles tend to crack, thereby reducing the MRR.



Ultrasonic Machining, Fig. 3 Removal mechanism in RUM

The above mechanism prevails mainly during machining of brittle work materials that are more susceptible to fracture. However, when the work materials are ductile, such as metals, the abrasives have tendency to get embed in the surface. The embedded particles work harden the surface, and upon successive impacts, chipping removal from the work surface occurs.

Thus, though USM can be used on both brittle and ductile work materials, it is more effective in the case of former than the latter.

In the RUM, the removal mechanism involves linear as well as rotational motion to the abrasives that are bonded to the tools, (Fig. 3). Therefore, the actions involved include hammering followed by abrasion and extraction [7].

Extensive work on the mechanism of removal shows that besides indentation, multiple mechanisms operate in USM process; these include [8]:

1. Mechanical abrasion by hammering of abrasive particles
2. Microchipping by impact of free abrasives
3. Cavitation effects from abrasive slurries
4. Chemical action associated with the fluid employed

Process Parameters

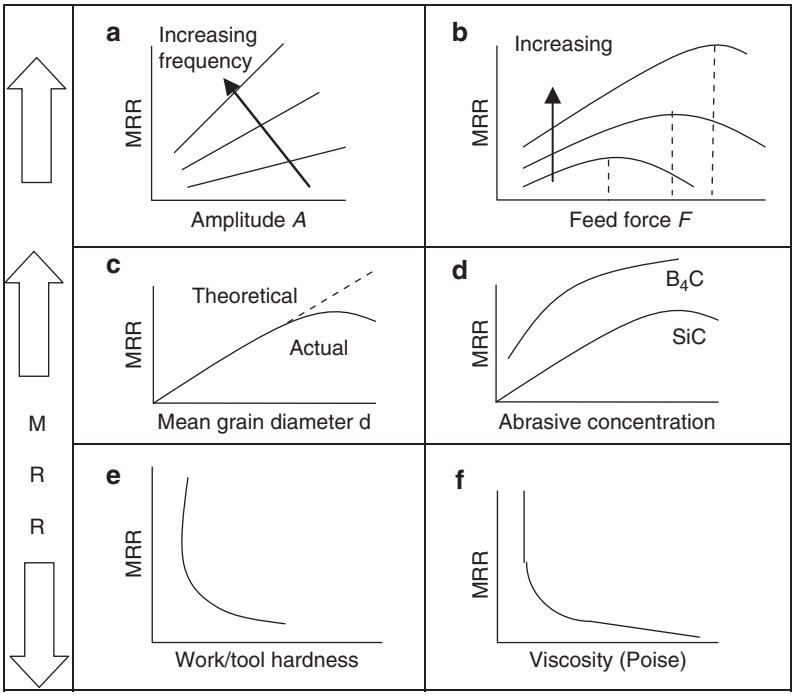
The theoretical as well as experimental investigations on the process have shown that the MRR in the ultrasonic machining is a function of a large number of parameters such as [9] (1) frequency, (2) amplitude, (3) static force, (4) hardness ratio (q) of the work to tool, (5) grain size, and (6) concentration of abrasive slurry. A pictorial summary of these effects is presented in Fig. 4a–f.

It is understood that the MRR increases with an increase in the frequency and the amplitude of ultrasonic vibrations (Fig. 4a). The MRR increases with an increase in the static force, but after a certain increase in the static force (Fig. 4b), the abrasive particles tend to break thereby reducing the MRR. A decrease in the hardness of workpiece in the ratio (q) causes a rapid reduction in MRR (Fig. 4e). This indicates that the brittle work materials are machined rapidly by USM. A table showing relative removal rate of USM in machining of glass by considering it as 100 is presented in Table 2.

The MRR increases in proportion to the increase in the grain size (Fig. 4c) until the grain size becomes more than the amplitude of vibrations, that is, where the particles tend to fracture.

The MRR increases proportional to the one-fourth power of slurry concentration, i.e., $C^{1/4}$ (Fig. 4d). However, the increasing trend continues till concentration reaches 30 %, beyond which an increase in the concentration does not help. The MRR decreases significantly with a decrease in the viscosity of the carrier liquid (Fig. 4f).

The machined surface roughness using USM depends upon grain size of abrasives [8]. The smaller the grain size, lower the roughness.



Ultrasonic Machining, Fig. 4 (**a**–**f**) Variation of MRR with process parameters in USM

In addition, surface roughness is proportional to {hardness (H)/elastic modulus (E)}^{n} as depicted in Fig. 5a. The work materials that have low H/E ratio are found to give lower surface roughness.

Again, in most of the cases, the RUM process is found to give better surface finish than the USM (Fig. 5a). Nevertheless, the values as low as 0.4 μmRa for surface roughness are obtainable by using the process.

The holes made by USM are oversized and have form errors [8]. The oversize is maximum at the entry the of hole, and it increases as the length of the hole increases. Usually, the upper limit on the oversize is equal to the maximum size of the abrasives used. The oversize and out of roundness are found to reduce with increasing the static load, as it suppresses the lateral vibrations of the tool. Again, the work materials with lower H/E ratio again show lower out of roundness (Fig. 5b).

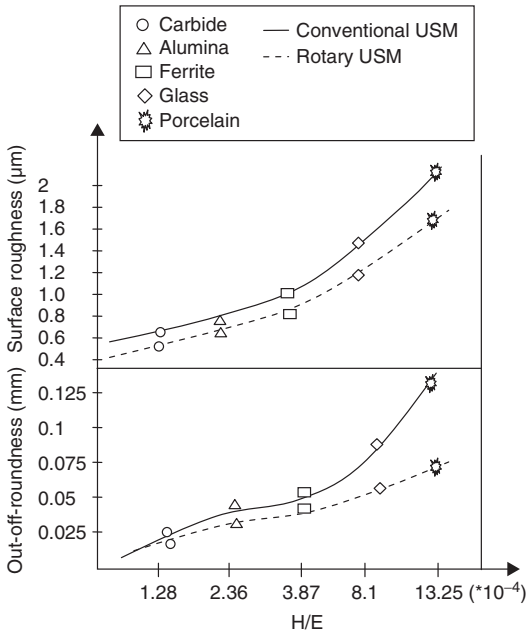
Tool Wear

Tool wear is an important variable in USM that affects both MRR and hole accuracy [8].

Ultrasonic Machining, Table 2 Relative removal rates in USM (For $f = 16.3$ kHz, $A = 12.3$ μm, mesh size of abrasives = 100) [9]

Work material	Relative removal rate
Glass	100.0
Brass	6.6
Tungsten	4.8
Titanium	4.0
Steel	3.9
Chrome steel	1.4

In general, the tool wear occurs along the length (W_L) as well as along the diameter (W_D) of the tool. It is understood that the tool wear is maximum when the static load is maximum. Also, at the maximum static load, the MRR is also maximum. Therefore, the maximum static load may be considered as the optimum point for the tool wear. The tool wear tends to increase with an increase in the hardness and size of the abrasives. W_L is more influenced by the product of hardness (H) and impact strength (K_i) of the tool material. On the other hand, the W_D is more influenced by the hardness (H) of the tool materials. Therefore, Nimonic 80A, thoriated tungsten, or silver steel



Ultrasonic Machining, Fig. 5 Effect of *H*/*E* on (**a**) surface roughness for various materials under USM and RUM (**b**) out of roundness [8]

is highly recommended as tool materials. The tool wear also increases with an increase in the depth of hole drilled and machining time.

A number of times, fatigue failure of the tools is observed in the USM process. Therefore, the tool materials should have high fatigue strength too.

Recent Trends

The recent developments in ultrasonic machining include use of various tools and techniques to improve performance of the USM process and the application of ultrasonic vibrations in various other manufacturing processes to improve their productivity.

These developments include (a) use of automatic and intelligent machining control systems, (b) use of elliptical vibrations during ultrasonic machining to improve MRR, (c) use of numerical tools to design the horn for efficient conversion of longitudinal vibrations to mixed lateral and longitudinal vibrations thereby maximizing the MRR, and (d) combining the ultrasonic vibrations with other machining processes, for example, in electric discharge machining, where the vibrations help better debris disposal.

Summary

- USM is a nonthermal process and is the most preferred method for machining of hard and brittle materials with hardness above 40 HRC.
- Of the two process variants, USM and RUM, the RUM gives better machining performance than USM; however, the process application is limited to axisymmetric shapes.
- The removal mechanism in the process involves mainly indentation, impact, and chipping leading to generation of lateral crack and followed by the materials removal on brittle work surfaces.
- On the other hand, on ductile work surfaces, removal involves impact induced work hardening followed by chipping.
- In USM, the MRR increases mainly with an increase in the amplitude, frequency, and the static force.

Cross-References

► Ultraprecision Machining (UPM)

References

1. Tyrrell, W.R.: Ultrasonic Machining, Machining, ASM Metals Handbook, vol. 16. ASM International, Metals Park (1989)
2. Drozda, T.: Tool and manufacturing engineers handbook, Chapter 14. In: Randy, G. (ed.) Ultrasonic Machining, vol. 1, 4th edn. SME, Michigan (1983)
3. Evans, A.G.: Fracture mechanics determinations. In: Bradt, R.C., et al. (eds.) Fracture Mechanics of Ceramics, vol. 1. Plenum, New York (1974)
4. Lawn, B.R., Evans, A.G.: A model for crack initiation in elastic/plastic indentation fields. J. Mater. Sci. **12**, 2195–2199 (1977)
5. Lawn, B.R., Evans, A.G., Marshall, D.B.: Elastic/Plastic indentation damage in ceramics: the median/radial crack system. J. Am. Ceram. Soc. **63**, 574 (1980)

6. Shaw, M.C.: Ultrasonic grinding. Microtechnic **10**(6), 257–265 (1956)
7. Pei, Z.J., Ferreira, P.M., Kapoor, S.G., Haselkorn, M.: Rotary ultrasonic machining for face milling of ceramics. Int. J. Mach. Tool Manufact. **35**(7), 1033–1046 (1995)
8. Thoe, T.B., Aspinwall, D.K., Wise, M.L.H.: Review on ultrasonic machining. Int. J. Mach. Tool Manufact. **38**(4), 239–255 (1998)
9. Ghosh, A., Mallik, A.K.: Ultrasonic Machining (USM), Manufacturing Science. Affiliated East–west Press Private Limited, New Delhi (1985)

Ultrasonic Vibration-Assisted Nanomachining

Jingyan Dong
Edward P. Fitts Department of Industrial and System Engineering, North Carolina State University, Raleigh, NC, USA

Synonyms

Ultrasonic force-assisted nanomachining; Ultrasonic force-regulated nanomachining

Definition

Ultrasonic vibration-assisted nanomachining is a mechanical machining-based nanofabrication approach using sharp cantilever tips, in which controlled ultrasonic vibration between the tip and the sample is utilized to regulate the dimension of the fabricated features and to improve the performance and speed of nanolithography.

Introduction

Ultrasonic vibration-assisted nanomachining is one of the tip-based nanolithography techniques that enable high-rate tunable nanofabrication on polymeric and metallic substrates. The sharp tip of the microcantilever provides capabilities for low-cost fabrication of polymeric or metallic nanostructures in the sub-50 nm regime and has potentials for fabricating 3D nanostructures. As a mechanical modification approach with an atomic force microscopy (AFM) tip, the ultrasonic vibration-assisted nanomachining overcomes many limitations from traditional nanomachining approaches (e.g., nano-scratching and dynamic plowing lithography). One major disadvantage of traditional mechanical nanomachining is their low efficiency, since the fabricated feature width is mainly determined by the AFM tip size. Multi-scratching steps have to be taken to fabricate a large feature. Moreover, poor process controllability is the other disadvantage of the traditional nanomachining. In nano-scratching, the sharp tip modifies the sample surface by plastic deformation and has been successfully used to produce sub-50 nm features on many samples (e.g., polymers, silicon, aluminum, etc.) [1–8]. A large normal force is generally required to push the tip into the sample for mechanical modification. This normal force depends heavily on properties of the sample materials and the cantilever used. Control of feature dimension in fabrication for different samples and features is very difficult. In dynamic plowing lithography [9, 10], the cantilever tip operating at the tapping mode is used in fabrication. The capability to regulate the machining behavior is limited by the large tip-sample interaction at the resonant frequency of the cantilever, since the cantilever oscillation is very sensitive to the tip-sample gap and set-point force.

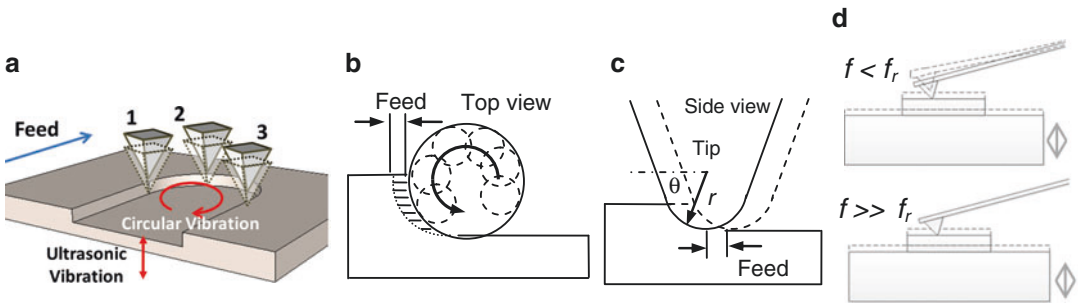
Ultrasonic vibration-assisted nanomachining provides high-rate tunable fabrication for polymeric or metallic nano-features. The ultrasonic tip-sample z-vibration can reliably regulate the fabrication depth. A high-frequency in-plane circular tip-sample vibration controls the width of the fabricated features in one scan and significantly improves the speed of the nanolithography. The ultrasonic tip-sample vibration also reduces the normal force and friction during nanomachining and improves the issue of tip wear in traditional tip-based nanomachining.

Schematic of Ultrasonic Vibration-Assisted Nanomachining

In ultrasonic vibration-assisted nanomachining, controlled high-frequency vibration between the cantilever tip and the sample is generated to increase the speed and controllability of the nanomachining process. The tip-sample vibration includes an ultrasonic z-vibration and in-plane circular vibration. The ultrasonic z-vibration is utilized to regulate fabrication depth and reduce friction in nanomachining. The high-frequency in-plane circular vibration is used to control the width of the fabricated features and to improve the speed of nanolithography. The schematic illustration of ultrasonic vibration-assisted nanomachining is shown in Fig. 1.

Ultrasonic tip-sample vibration in z-direction effectively increases the dynamic stiffness of the cantilever and improves the machining performance. Here the ultrasonic vibration frequency needs to be much larger than the resonant frequency of the cantilever. When the sample is brought into contact with the cantilever and vibrated with a frequency f that is below the resonant frequency f_r of the cantilever ($f < f_r$), the cantilever follows the sample vibration. When the ultrasonic vibration is applied with its frequency much higher than the resonant frequency of the cantilever, the cantilever cannot follow the sample vibration due to its inertia, which makes the tip immerse into the sample surface on average. With the vibration amplitude larger than the initial deflection of the cantilever, in one vibration circle, the tip experiences a contact phase and a detach phase. During contact, the sharp tip is indented into the sample since the cantilever is dynamically frozen at ultrasonic frequency. The ultrasonic force, as an additional force from the tip-sample interaction under static condition, has been estimated in ultrasonic force microscopy (UFM) [11–14], which is the average of the repulsive force that depends on the surface elasticity and vibration amplitude, and is insensitive to the oscillation frequency and set-point force. Such ultrasound excitation has been used in Ultrasonic Force Microscopy (UFM) as an imaging technique to obtain elastic or subsurface images of the sample. In ultrasonic vibration-assisted nanomachining, the ultrasonic tip-sample z-vibration is used to control the machining depth. For a chosen sample material, the depth of indentation mostly depends on the vibration amplitude and is insensitive to the oscillation frequency, which provides a reliable method to control the machining depth.

In ultrasonic vibration-assisted nanomachining, besides the ultrasonic z-vibration, a high-frequency circular tip-sample motion in the xy-plane is utilized to achieve better control of the fabrication process and to overcome many limitations from direct mechanical scratching, such as large tip-sample interaction and slow machining speed. The feature dimension, such as line width, can be regulated by changing the amplitude of the circular vibration/motion. As shown in Fig. 1b, the outer edge of the circular path can be viewed as the edge



Ultrasonic Vibration-Assisted Nanomachining, Fig. 1 (**a**) Schematic illustration of vibration-assisted nanomachining. (**b**, **c**) The top view and side view show the tip-sample engagement of the virtual tool (when the tip is vibrated with an in-plane circular path) with the sample during one rotation cycle. (**d**) At low frequency, cantilever follows the sample vibration. At high frequency ($f>>f_r$), cantilever begins to indent into the sample

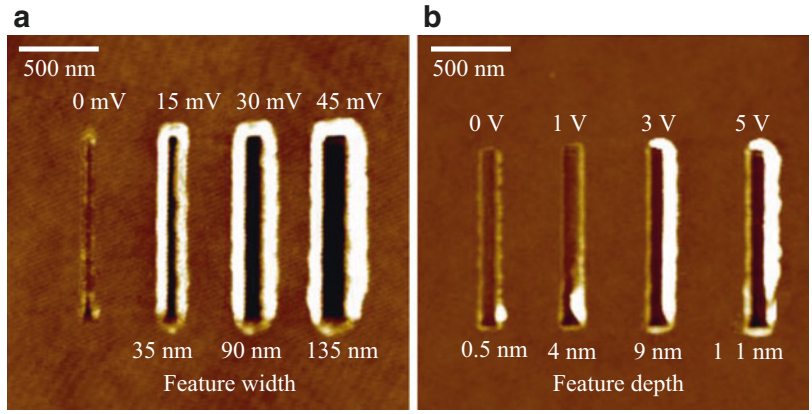
of a virtual tool with a single cutting tooth, and the effective cutting edge is the edge of the tip that is tangent to the virtual tool at that moment. The diameter of the virtual tool is controlled by the amplitude of vibration. This machining scheme is similar to the conventional-scale single-tooth fly-cutting, in which the rotation of the virtual tool is obtained by the circular motion of the tip.

By rotating the tip in *xy*-plane with high frequency, the tip-sample interaction can be distributed to each rotation cycle. During each cycle, only a thin slice of material defined by feed per rotation is removed. Thus, tip-sample interaction force can be significantly reduced and the machining speed can be significantly increased. Given the feed per rotation f and the rotary speed of the tool N, the machining speed can be expressed as $V = fN$. Even with a small feed per rotation to reduce the tip-sample interaction in each cycle, if a high rotary speed is used, a large machining speed can still be achieved. For example, assuming f is 10 nm and N is 10 kHz, the lithography speed can be as large as 100 μm/s, which is very difficult to be obtained by other mechanical modification approaches with an AFM. Moreover, by choosing different feed per rotation, the machining force during each rotational cycle can be controlled. The process has better controllability to improve the overall performance (e.g., better finish/topography uniformity) and to achieve more efficient material removal.

The ultrasonic *z*-vibration and in-plane rotation for ultrasonic vibration-assisted nanomachining is produced by a high-speed *xyz* nanopositioner as a nano-vibrator [15, 16], which is mounted onto the AFM scanner to provide the high-frequency-controlled vibration in *xyz* directions. The vibration amplitude of the nano-vibrator in *xy* directions ranges from sub-nm to a few microns depending on the command voltage to the piezo actuator, which provides good tunability for the nanomachining process.

The experiments of the ultrasonic vibration-assisted nanomachining were performed on a 200 nm thick PMMA film, which is spin-coated on a cleaned silicon substrate. A tapping mode cantilever was used in the nanomachining experiment. Compared with direct scratching and dynamic plowing, the ultrasonic vibration-assisted nanomachining process provides a faster machining speed and greater material removal rate. In Zhang and Dong [15], straight line patterns are machined on the PMMA film under different vibrational conditions to characterize the capability of ultrasonic vibration in regulating both machining depths. After an *xy*-circular vibration is excited with a frequency of 10 kHz and machining speed of 20 μm/s, even with a small set-point force of 100 nN, a noticeable cutting depth about 1 nm can be achieved, compared with direct scratching. With circular xy-vibration, tunable line width can be achieved in one machining path by just changing the *xy*-vibration amplitude, which is controlled by the command voltages given to the *xy*-piezo actuators, as shown in Fig. 2. From the fabrication results in Fig. 2a, clearly a large vibrational command produces a wider trench feature, and a small vibration signal corresponds to a narrow trench. The line width ranges from 35 to 135 nm, as the vibrational commands increase from 15 to 45 mV. Since the cutting load is distributed to each small cutting cycles, good surface finish is obtained; surface roughness in the trench along the machining direction is less than 1 nm. The machined trench width follows a nearly linear relationship with the *xy*-piezo driving signal amplitude.

Set-point z-force has been widely used to control the cutting depth in tip-based nanomachining. However, large z-force can bring large friction and reduce tip life. It was found that ultrasonic z-vibration is able to control the fabrication depth and, at the same time, to reduce friction [15, 16]. The effect of ultrasonic z-vibration is shown in Fig. 2b. The machining speed in the tests is set to be 20 μm/s. Without ultrasonic z-vibration, using a small set-point force of 100 nN, the cantilever can barely make any modification on the sample. After the sample is vibrated at ultrasonic frequency (2 MHz), noticeable trenches were fabricated on the sample surface. The depth of the trench depends on the vibration amplitude in z-direction. The trench depth is directly controllable by the vibration amplitude in the z-direction, as shown in Fig. 2b. With a larger command to z-piezo, a large trench depth is obtained. With or



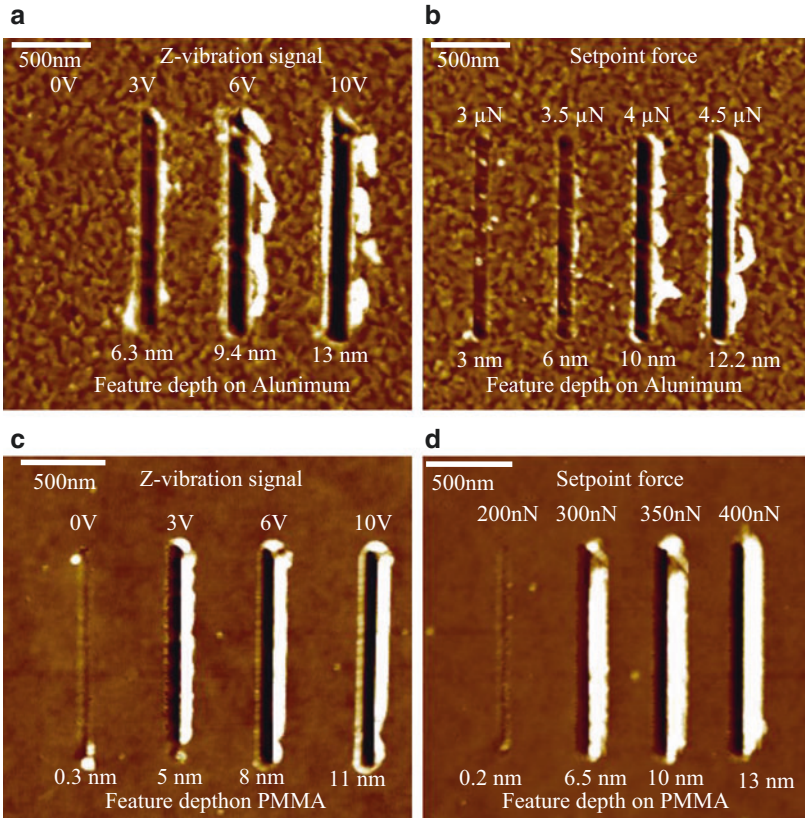
Ultrasonic Vibration-Assisted Nanomachining, Fig. 2 (**a**) Effect of *xy*-circular vibration on nanomachining. Larger *xy*-vibration amplitude results in large trench width with a nearly linear relationship. (**b**) Effect of ultrasonic z-vibration on nanomachining. Z-vibration amplitude and trench depth follow a nearly linear relationship

without *xy*-circular vibration, the trench depth follows an approximately linear relationship with the z-vibration amplitude, which provides a reliable method to regulate the cutting depth. By using larger z-vibration amplitude, thicker slice of materials (as thick as 80 nm) can be machined in one scan with the same machining speed.

Robust Material-Insensitive Feature Depth Regulation by Ultrasonic Vibration

It was demonstrated by the nanomachining experiments on aluminum and PMMA in Zhang et al. [16] that when the set-point force is used to control fabrication depth (i.e., without ultrasonic z-vibration), a much larger normal set-point force is required to fabricate features with similar depth on aluminum, compared with the required set-point force for PMMA (a few micro-Newtons for aluminum vs. a few hundred nN for PMMA). These results are very easy to understand, since aluminum is a much stiffer material than PMMA. However, when using ultrasonic z-vibration, with the same small set-point force (200 nN in Fig. 3) and the same vibration amplitude, the ultrasonic vibration-assisted nanomachining produces similar feature depth on both PMMA and aluminum, even though these two materials have very different mechanical properties. The fabrication depth is predominantly controlled by the amplitude of the ultrasonic z-vibration and is insensitive to substrate materials and other process conditions, which indicate the great potential of using ultrasonic z-vibration in regulating the feature depth.

The depth regulation by ultrasonic vibration can be explained by the increased dynamic stiffness of the vibrating cantilever at the ultrasonic frequencies. When the sample vibration frequency is much higher than the resonant frequency of the cantilever, the cantilever cannot follow the sample vibration due to its inertia. In other words, the cantilever is dynamically frozen at the ultrasonic frequency, due to its extremely large dynamic stiffness. For a rectangular beam cantilever, its static stiffness can be expressed by $K_s = 3EI/L^3$, where E is Young's modulus, I is the moment of inertia, and L is the cantilever length. At an ultrasonic frequency, the effective dynamic stiffness of a cantilever can be expressed by $K_d = 3EI/b^3\lambda^3$, where λ is the wavelength of ultrasonic flexural wave and b is the constant of order unity. By making $K_s = K_d$ at the resonant frequency of the cantilever, the constant b can be derived, and the dynamic stiffness can be expressed as $K_d = K_s\,(L/0.298\,\lambda)^3 = K_s\,(\omega/\omega_0)^{3/2}$, where ω_0 is the angular resonant frequency of the cantilever. As an example, for a tapping mode cantilever with a static stiffness of 48 N/m and resonant frequency of 190 kHz, when it is excited at the ultrasonic frequency of 3 MHz, the effective dynamic stiffness of 1,350 N/m can be obtained. Compared with the extremely large dynamic stiffness of the cantilever, the sample stiffness is



Ultrasonic Vibration-Assisted Nanomachining, Fig. 3 Comparison of nanomachining with and without z-vibration on aluminum (**a**, **b**) and PMMA (**c**, **d**). (**a**, **c**) Feature depth controlled by ultrasonic z-vibration with a small set-point force of 200nN. (**b**, **d**) Feature depth controlled directly by set-point force

relatively much smaller for many materials, including aluminum and PMMA. Using Hertzian approximation, the sample stiffness can be estimated as $k_s = (6RE^{*2}F)^{1/3}$, where F is the tip-sample contact force, E^* is the effective elasticity which can be expressed as

$$\frac{1}{E^*} = \frac{1-v_t^2}{E_t} + \frac{1-v_s^2}{E_s},$$

and E_t, v_t and E_s, v_s are Young's modulus and Poisson's ratio of the tip and sample, respectively. If we assume an elastic contact between the tip and sample surface and a 10 nN contact force, the effective sample stiffness of PMMA and aluminum is 24.5 and 153 N/m, respectively. The dynamic stiffness of the cantilever at ultrasonic frequency is 10 times larger than the sample stiffness of aluminum and more than 50 times larger than PMMA. Moreover, the above sample stiffness is based on the full contact between the cantilever and the tip. In ultrasonic vibration-assisted nanomachining scheme, since the vibration in the z-direction (MHz) is much faster than in the xy-plane (~10 KHz), the tip-sample contact in each z-vibration cycle is far less than the full sphere-plane contact, which will further decrease the effective sample stiffness.

Simply speaking, at ultrasonic frequency, the cantilever is much stiffer than the sample. The ultrasonic vibration, especially its vibration amplitude, is not affected by the tip-sample interaction. During vibration, the tip can easily indent into relatively soft samples for mechanical modification without much resistance. The tip-sample engagement or cutting depth is mainly controlled by the vibration amplitude. Therefore, even at a small set-point normal force, using the same vibration amplitude, ultrasonic vibration-assisted nanomachining produces similar feature depth on PMMA and aluminum samples. The fabrication depth is predominantly controlled by the

amplitude of tip-sample ultrasonic z-vibration and is insensitive to substrate materials.

Reduction of Friction and Machining Force by Ultrasonic Vibration

After applying ultrasonic tip-sample vibration in the z-direction, a significant reduction of the normal force and lateral force (friction) was observed during nanomachining [16]. To compare the effect of ultrasonic z-vibration on machining forces, a set of linear patterns were fabricated with depth controlled by the amplitude of the ultrasonic z-vibration and by the set-point z-force without vibration. Force data is compared for features with similar depth and width. The normal force and friction were acquired by measuring the cantilever normal deflection (A–B from the photodiode detector) and torsion deflection (C–D from the photodiode detector).

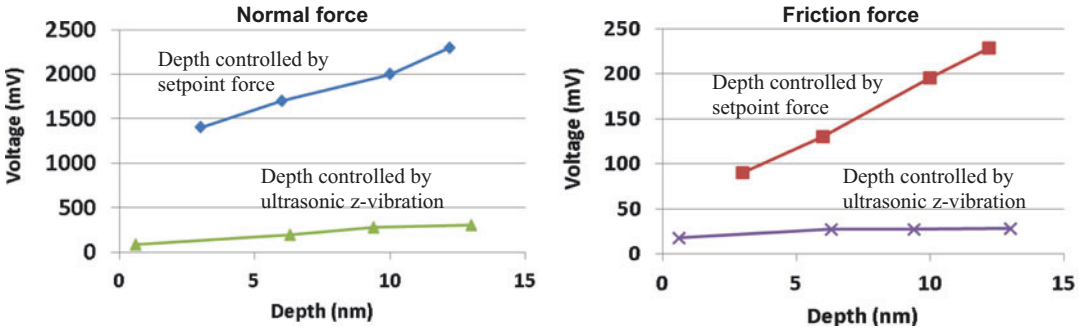
For the linear trench features machined on aluminum with depth controlled by ultrasonic z-vibration and by set-point z-force, the acquired normal and lateral force signals were plotted in Fig. 4. When the feature depth is controlled by the set-point force (no ultrasonic z-vibration), large normal forces have to be applied to fabricate features with the same depth. Ultrasonic tip-sample vibration in vertical direction significantly reduces the normal and lateral force in nanomachining, and the machining force only has a weak dependence on the fabricated feature depth. For ultrasonic vibration-assisted nanomachining, for feature depth up to about 12 nm, normal force and lateral force are kept at a very small value with normal force less than 400 mV (compared with 1,400 to 2,300 mV without vibration) and lateral friction force less than 30 mV (compared with 90 to 228 mV without vibration). For example, to fabricate a trench with a depth of 12 nm, 10 V z-vibration is required for ultrasonic vibration-assisted nanomachining, or a set-point force of 4,500 nN has to be used for nanomachining without z-vibration. When using the ultrasonic z-vibration, the normal force is 400 mV, about 1/6 of 2,300 mV for nanomachining without ultrasonic z-vibration, and the lateral force is 28 mV, only 1/8 of 230 mV for nanomachining without ultrasonic z-vibration.

The decrease in machining forces by ultrasonic z-vibration can be attributed to the small normal force required in nanomachining and intermittent tip-sample contact from ultrasonic z-vibration. The reduced normal force improves the applicability of ultrasonic vibration-assisted nanomachining to a broad set of materials, especially for hard materials. For traditional nanomachining of hard materials, custom-designed cantilevers with extremely large stiffness have to be used to provide the required huge normal force. Furthermore, the reduced machining force by ultrasonic vibration can potentially reduce tip wear and elongate tip life for the nanomachining process.

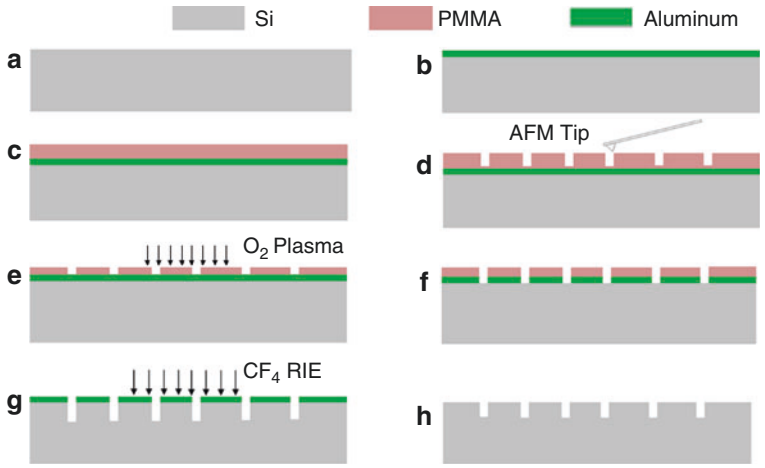
Ultrasonic Vibration-Assisted Nanomachining for Nanolithography

Besides fabricating patterns directly on many substrates, ultrasonic vibration-assisted nanomachining can be used as a nanolithography process to fabricating features on polymer films and then transfer to other functional substrates. A typical nanomachining-based nanolithography process [15] is illustrated in Fig. 5. A silicon substrate is evaporated with a thin aluminum layer and then spin-coated with PMMA film. Then, nanomachining is applied to make features on the PMMA film. To protect the cantilever tip from severe wearing and to elongate its lifetime, the machining depth is set to slightly less than the PMMA thickness to prevent the tip from direct contact with the aluminum layer. The residue PMMA can be removed by oxygen plasma asher to expose the underneath aluminum layer. After wet aluminum etching, the pattern is transferred from the PMMA to the aluminum layer. The RIE process can be further used to transfer the pattern from the aluminum mask to silicon.

Using nanomachining-based nanolithography process, features with width of 120 nm are machined on a 35 nm thick PMMA, as shown in Fig. 6. The depth of feature on the PMMA film is about 25 nm. The roughness at the bottom of the



Ultrasonic Vibration-Assisted Nanomachining, Fig. 4 Comparison of normal force and friction for nanomachining on aluminum with depth controlled by z-vibration and by set-point force

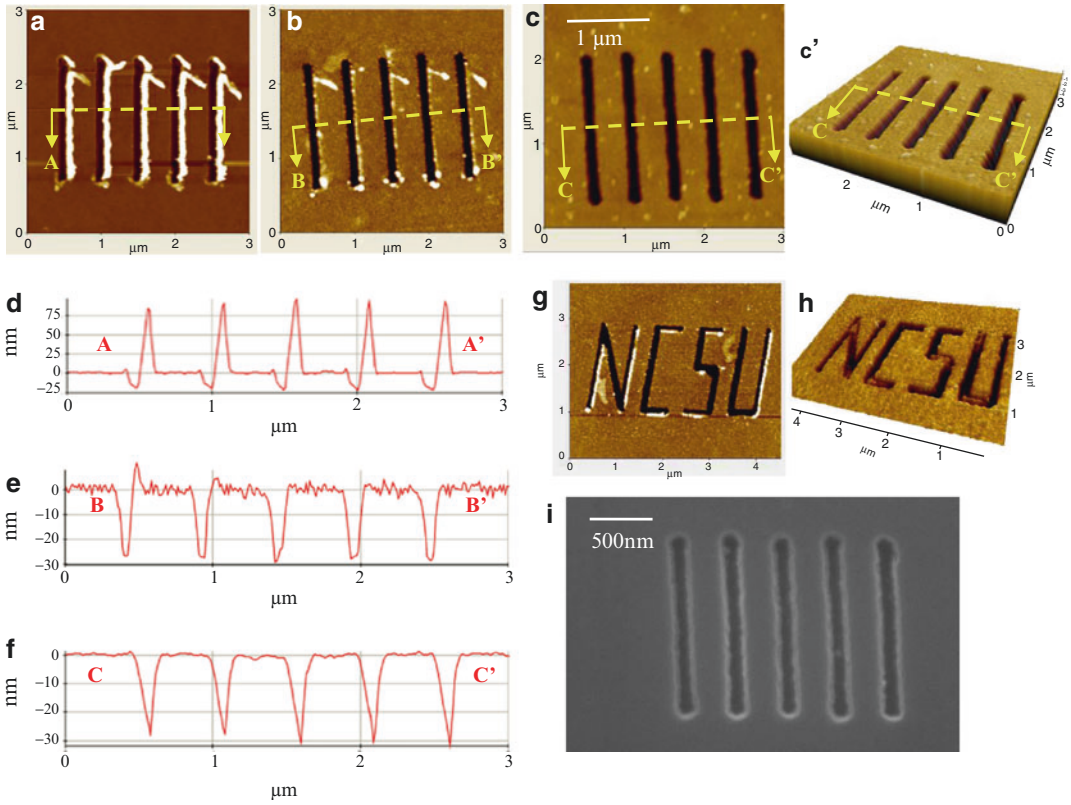


Ultrasonic Vibration-Assisted Nanomachining, Fig. 5 Fabrication sequences of nanomachining-based lithography process. (**a**) Starting from the cleaned silicon substrate; (**b**) aluminum deposition; (**c**) PMMA spin-coating; (**d**) machining patterns on PMMA; (**e**) PMMA etching in O2 plasma till the aluminum layer surface is exposed; (**f**) aluminum etching; (**g**) RIE etching of silicon; (**h**) aluminum striping and sample cleaning (Reprinted from Zhang and Dong [15] with permission)

slots is ∼1 nm. The remaining PMMA in the features after nanomachining is removed in oxygen asher. After that, the sample is placed in aluminum etchant to transfer the slot patterns from the PMMA to the aluminum layer. After aluminum etching, the roughness at the bottom of slots is about 0.7 nm, indicating the exposed silicon surface. After RIE etching and aluminum mask striping, the AFM image shows a cut depth of ∼30 nm in silicon substrate as shown in Fig. 6f and h. The SEM image of the fabricated line patterns is shown in Fig. 6h. The patterns can be reliably transferred from features that are machining on PMMA film to silicon substrate.

Future Directions

The ultrasonic vibration-assisted nanomachining has demonstrated promising potentials for improving the speed and material removal rate of nanomachining, reducing the machining force and tip wear, and increasing the controllability of the nanomachining processes. Future work aimed at modeling the nanomachining process to predict the feature produced and machining force at different process parameters will be important to make ultrasonic vibration-assisted nanomachining a viable and reliable process for nanofabrication. The application of ultrasonic



Ultrasonic Vibration-Assisted Nanomachining, Fig. 6 Images of results at different fabrication steps. (**a**) PMMA surface after nanomachining; (**b**) PMMA surface after aluminum mask forming etching; (c, c′) silicon substrate surface after RIE etching; (**d**) topography at A–A′; (**e**) topography at B–B′; (**f**) topography at C–C′. (**g**) Patterned image for letters "NCSU" after nanomachining and aluminum etching and (**h**) after RIE etching. (**i**) SEM images of the slot features on silicon substrate (Figure is reprinted from Zhang and Dong [15] with permission)

vibration-assisted nanomachining in the fabrication of 3D nanostructures can also be expected, which is actually a unique capability of mechanical machining-based nanofabrication processes.

References

1. Wendel, M., Lorenz, H., Kotthaus, J.P.: Sharpened electron beam deposited tips for high resolution atomic force microscope lithography and imaging. Appl. Phys. Lett. **67**, 3732–3734 (1995)
2. Yan, Y., Sun, T., Liang, Y., Dong, S.: Investigation on AFM-based micro/nano-CNC machining system. Int. J. Mach. Tool Manuf. **47**, 1651–1659 (2007)
3. Göbel, H., Von Blanckenhagen, P.: Atomic force microscope as a tool for metal surface modification. J. Vacuum Sci. Technol. B. **13**, 1247–1251 (1995)
4. Li, X., Wang, X., Xiong, Q., Eklund, P.C.: Top-down structure and device fabrication using in situ nanomachining. Appl. Phys. Lett. **87**, 233113 (2005)
5. Yan, Y., Hu, Z., Zhao, X., Sun, T., Dong, S., Li, X.: Top-down nanomechanical machining of three-dimensional nanostructures by atomic force microscopy. Small **6**, 724–728 (2010)
6. Hu, S., Altmeyer, S., Hamidi, A., Spangenberg, B., Kurz, H.: Novel approach to atomic force lithography. J. Vacuum Sci. Technol. B. **16**, 1983–1986 (1998)
7. Gozen, B.A., Ozdoganlar, O.B.: A rotating-tip-based mechanical nano-manufacturing process. Nanoscale Res. Lett. **5**, 1403–1407 (2010)
8. Schmid, S.R., Hector, L.G.: Simulation of asperity plowing in an atomic force microscope. Part II – plowing of aluminum alloys. Wear **215**, 257–266 (1998)
9. Heyde, M., Rademann, K., Cappella, B., Geuss, M., Sturm, H., Spangenberg, T., Niehus, H.: Dynamic plowing nanolithography on polymethylmethacrylate

using an atomic force microscope. Rev. Sci. Instrum. **72**, 136–141 (2001)
10. Klehn, B., Kunze, U.: Nanolithography with an atomic force microscope by means of vector-scan controlled dynamic plowing. J. Appl. Phys. **85**, 3897–3903 (1999)
11. Kolosov, O., Yamanaka, K.: Nonlinear detection of ultrasonic vibrations in an atomic force microscope. Jpn. J. Appl. Phys. **32**, 1095–1098 (1993)
12. Yamanaka, K., Ogiso, H., Kolosov, O.: Ultrasonic force microscopy for nanometer resolution subsurface imaging. Appl. Phys. Lett. **64**, 178–180 (1994)
13. Shekhawat, G.S., Dravid, V.P.: Nanoscale imaging of buried structures via scanning near-field ultrasound holography. Science **310**, 89–92 (2005)
14. Cuberes, M.T., Briggs, G.A.D., Kolosov, O.: Nonlinear detection of ultrasonic vibration of AFM cantilevers in and out of contact with the sample. Nanotechnology **12**, 53–59 (2001)
15. Zhang, L., Dong, J.: High-rate tunable ultrasonic force regulated nanomachining lithography with an atomic force microscope. Nanotechnology **23**(8), 085303 (2012)
16. Zhang, L., Dong, J., Cohen, P.H.: Material-insensitive feature depth control and machining force reduction by ultrasonic vibration in AFM-based nanomachining. IEEE Trans. Nanotechnol. **12**(5), 743–750 (2013)

Unconventional Computing

► Molecular Computing

Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics

Shadi Rohani, Marta Quintanilla, Rafik Naccache, Roberto Morandotti, Luca Razzari and Fiorenzo Vetrone
Institut National de la Recherche Scientifique – Énergie Matériaux et Télécommunications, Université du Québec, Varennes, QC, Canada

Synonyms

Anti-stokes luminescence; Emission enhancement; Nanoplasmonic enhancement

Definition

Upconversion is a process in which the sequential absorption of two (or more) low-energy photons leads to the emission of one photon with higher energy. The upconversion process can occur in materials such as lanthanide-doped nanoparticles and can be enhanced by the strong local electric field found in the proximity of metallic (plasmonic) nanostructures.

Overview

Nanomaterials have attracted considerable attention due to their size-dependent properties, allowing many interesting and sometimes novel features to be observed [1]. Of particular interest are luminescent nanoparticles as they have been touted to find widespread integration in applications ranging from the realization of display devices to bioimaging [2]. One particular class of luminescent nanomaterials that has garnered considerable interest is the lanthanide (Ln^{3+})-doped upconverting nanoparticles (UCNPs). These UCNPs are particularly attractive as they possess the capability of converting low-energy (long λ) light such as near infrared (NIR) into high-energy (short λ) emission spanning the UV, visible, and even the NIR, in a process known as upconversion [3]. The ability to generate light covering the UV-NIR regions of the spectrum via NIR excitation implies that this material can be useful in optical applications such as the realization of lighting materials, anti-counterfeiting measures, displays, and imaging to name a few [4]. However, since upconversion is a multiphoton process, it suffers from low efficiency (or low quantum yield which is defined as the ratio between emitted photons and absorbed photons) [5] in comparison to other direct excitation approaches (i.e., standard fluorescence). Additionally, alternative depopulation routes of the excited states of the lanthanides result in loss of energy that otherwise could be used to promote upconversion emission. Consequently, scientists have looked at ways to enhance the upconversion emission via several routes including the

development of novel low-energy phonon hosts, higher crystallinity materials, the preparation of core/shell architectures, where a shell passivates the luminescent core, and, finally, the exploitation of the synergistic effects that arise due to the combination of plasmonic nanomaterials with UCNPs. The latter is the subject of this work and is discussed in detail below.

Upconverting Nanoparticles

Prior to discussing the synergy between UCNPs and nanoplasmonic materials, we define upconversion in greater detail. Upconversion is a physical process that involves the conversion of lower-energy photons into higher-energy photons using the real (intermediate) states of an atom. This phenomenon has been observed in transition metals and actinides, but mainly in the lanthanide elements. In their trivalent state, Ln^{3+} ions have valence electrons in an inner shell with the $4f^n$ electron configuration that is protected from the environment by outer energy levels (shells). The abundant and unique energy level structures remain, thus, largely unmodified by a change of environment, and for this reason, they are widely used as dopants in transparent materials to add narrow and well-defined emission and absorption bands in the UV, visible, and NIR ranges [6]. When the lanthanide ions are exposed to a light source with photons of energy equal to the energy distance between two levels, i.e., the ground and an excited state of the ion, the electrons can be promoted from the lower to this higher-energy state. Absorption of another photon from the excitation light, and/or energy from electrons in neighbor ions, can result in the excitation to higher-energy levels. If the excited electrons then release their energy and return to the ground state, the result is the emission of photons featured by shorter wavelengths (such as UV or visible).

Processes that also account for NIR to visible light transformation, such as two-photon absorption, can be observed in other materials. However, it is noteworthy to mention that the excitation of Ln^{3+}-doped UCNPs occurs via real electronic states featured by defined (relatively long) lifetimes, in contrast to the case of two-photon absorption. As a consequence, high-power, ultrafast, expensive laser setups are not required for the efficient excitation of Ln^{3+}-doped UCNPs, which can be achieved by using low-power, readily available solid-state laser diodes.

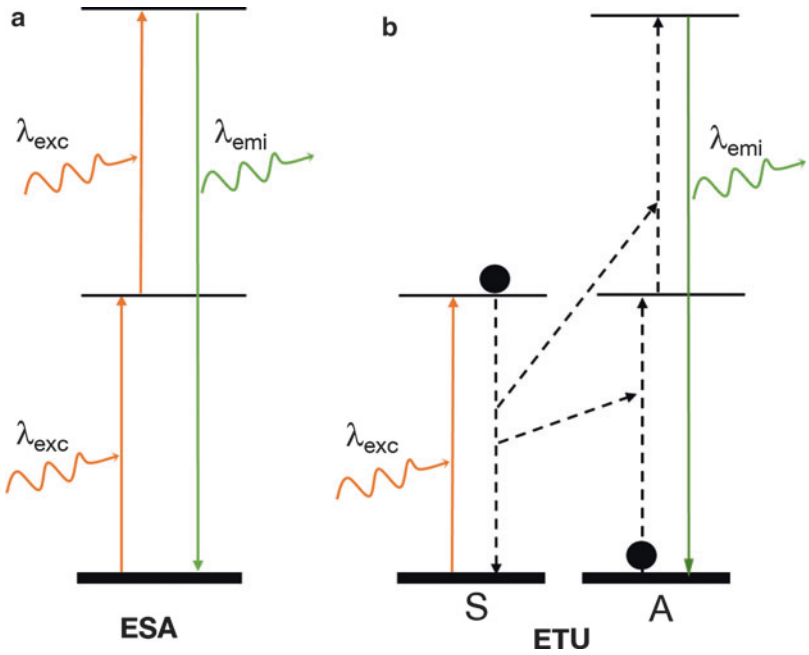
Upconversion is known to occur via three distinct mechanisms, namely, excited state absorption (ESA), energy transfer upconversion (ETU), and photon avalanche (PA). The latter is not commonly observed in the cases covered by this text and will not be discussed below.

Excited State Absorption

In an ESA mechanism, electrons of the ground state absorb one or more photons, sequentially, from the excitation source. When one electron in the ground state absorbs one photon, it is excited into an intermediate state. Then, after sequentially absorbing a second photon, it can be promoted to a higher excited state, and, upon returning to the ground state, it can emit higher-energy photons (Fig. 1a). Although in the simplest case, the energy levels are equidistant; ESA may be observed in systems with non-equal energy levels through phonon-mediated processes. Moreover, since this is a single ion process, it is generally observed in materials with low Ln^{3+} dopant concentrations where the Ln^{3+}-Ln^{3+} distances are too far for any interactions between ions to occur.

Energy Transfer Upconversion

Energy transfer upconversion is known to be the most efficient upconversion process in Ln^{3+}-doped nanomaterials [6]. It occurs between two different ions with nearly equally spaced energy levels, with another requirement being the nearby localizations of the ions involved in the process (distance typically of the order of one nanometer) so that the transfer of energy can easily occur from one ion to its neighbor. These two ions are normally called the sensitizer (S) (or donor) and the activator (A) (or acceptor). The sensitizer is an ion with a high-absorption cross section and that can absorb the photons from the excitation source efficiently, so electrons are promoted from their ground to an excited state. Subsequently, the sensitizer transfers this absorbed energy to the



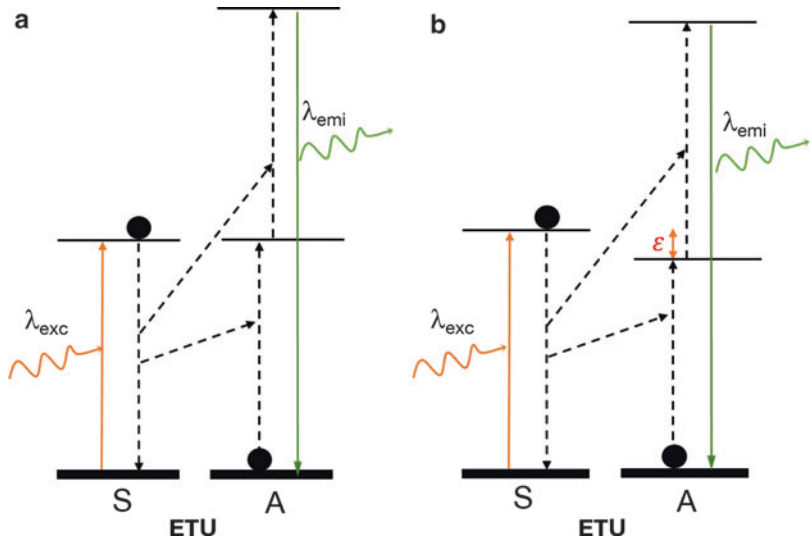
Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics, Fig. 1 Schematic diagram of two common upconversion mechanisms: (**a**) excited state absorption and (**b**) energy transfer upconversion

electrons of the activator, allowing them (after several steps) to reach higher-energy states, from which the electrons will decay by emitting the upconverted luminescence (Fig. 1b).

In the energy transfer process between sensitizer and acceptor ions, two situations, namely, resonant non-radiative transfer and phonon-assisted non-radiative transfer, can occur. In resonant non-radiative transfer, the energy levels of both S and A are equally spaced, while, in the phonon-assisted non-radiative transfer, an energy mismatch exists between the energy levels of the sensitizer and acceptor ions, so energy from the lattice (quantized in the form of phonons) is necessary to fulfill energy conservancy constrains. Figure 2 shows schematic diagrams of the ETU processes between two ions. In these images, we observe resonant non-radiative transfer (Fig. 2a) and phonon-assisted non-radiative transfer (Fig. 2b). Theoretically, most lanthanide ions can undergo the NIR to visible upconversion process; however, relatively efficient upconversion is only possible with a few trivalent lanthanide ions (e.g., Er^{3+} and Tm^{3+}) under low pump densities. To date, the Yb^{3+}/Er^{3+} or Yb^{3+}/Tm^{3+} dopant ion combinations have been greatly employed in co-doped nanomaterials. This is due to the fact that their energy levels are reasonably matched allowing for efficient energy transfer following NIR excitation.

Interactions Between Metallic Nanostructures and Upconverting Nanoparticles

As it was mentioned in the introduction, upconversion processes are intrinsically limited in terms of quantum yield due to their nonlinear nature. As such, in a two-step process in which two NIR photons are absorbed to emit one green photon, as it happens in the very common upconversion of co-doped Er^{3+}/Yb^{3+} ions, the quantum yield cannot be higher than 50 %. In the same way, blue emission through the upconversion of Tm^{3+}/Yb^{3+}, a three-step process, reduces the theoretical maximum quantum yield to 33 %. Moreover, as there are additional depopulation routes besides upconversion that can affect every energy state involved in the process, the



Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics, Fig. 2 Schematic diagram of a resonant (**a**) and a phonon-assisted (**b**) upconversion process

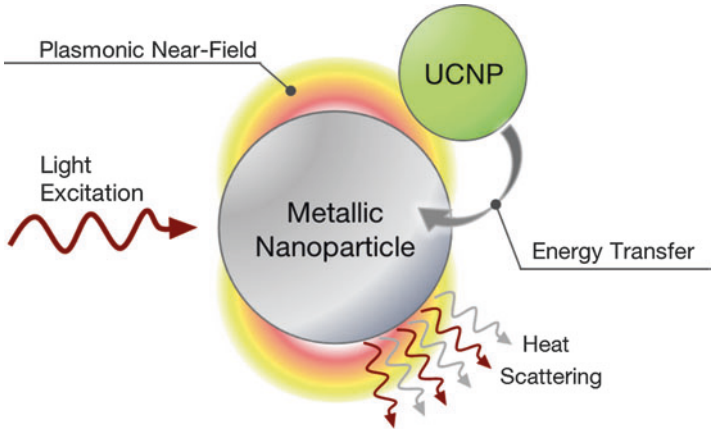
quantum yield is typically significantly lower than the predicted maximum. The list of alternative depopulation processes includes radiative and non-radiative transitions to lower-energy states, as well as additional energy transfer mechanisms, including the quenching triggered by molecules adsorbed on the surface of nanoparticles, defects, etc. Altogether, the upconversion quantum yield is normally below 10 % and in nanoparticles often below 1 % (depending on the concentrations of the sensitizer and the activator, their nature, the host material, the size of the nanoparticles, the excitation power density, etc.) [5].

It has been demonstrated that the plasmonic properties of metallic nanostructures can be applied to enhance the emission probabilities of emitters (atoms, molecules, or quantum dots) in their close proximity [7]. With regard to upconversion, the fact that the resultant emissions are due to multistep processes amplifies the possibility for enhancement upon coupling to suitable plasmonic nanostructures. To fully describe the interaction between metallic nanostructures and UCNPs, two different physical phenomena must be considered (Fig. 3). One is the influence of the plasmonic near-field, and the second is the non-radiative energy transfer between the UCNP and the metallic nanostructures [8, 9]. Prior to delving into the precise nature of the interactions of luminescent and plasmonic nanostructures, we introduce the relevant concept of surface plasmon resonance.

Plasmonic Nanomaterials and Surface Plasmon Resonance

It has been long known that the properties of materials change once they enter nanoscopic dimensions (in the 1–100 nm range). This can be as a result of confinement (such as limitations imposed on the movement of electrons), or alternatively of an increased surface-to-volume ratio. In fact, optical and magnetic properties are dependent on the electron motion characteristics of a system, while, surface properties can gain importance compared to volume properties in a nanoscale scenario [10]. In the case of metallic nanostructures, one of the best examples is related to the excitation of localized surface plasmons, resulting from the interaction between the free electrons of the metal and the optical electromagnetic field.

A metallic nanoparticle, for example, can be described as a lattice of ionic cores with conduction electrons moving almost freely inside the nanoparticle (the Fermi sea). When it is illuminated by a light source, the corresponding electromagnetic field applies a force on the conduction electrons at the nanoparticle surface. This results



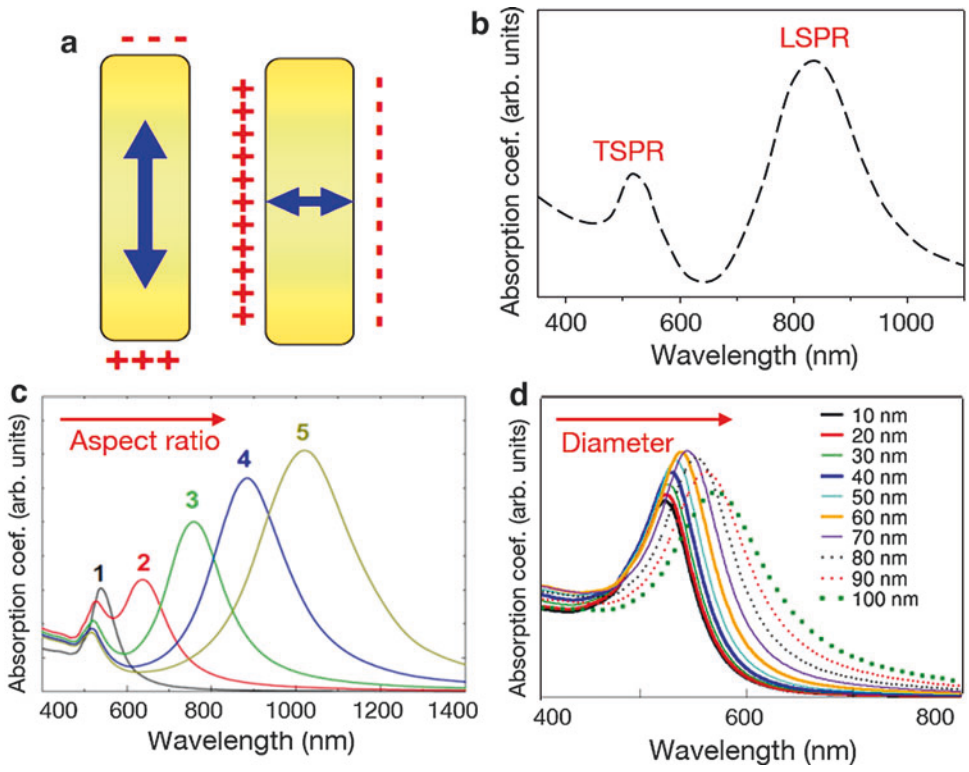
Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics, Fig. 3 The interaction between metallic nanostructures (a nanoparticle in this case) and upconverting nanoparticles (*UCNPs*) is mainly driven by the plasmonic near-field generated by the metallic particle and by nonresonant energy transfer between both. Additionally, the metallic nanostructure can scatter light or release heat

in a negative charge accumulated on one side of the nanoparticle, leaving the other side positively charged, thus creating an electric dipole. This dipole gives rise to an electric field with opposite direction to that of the illuminating light and, thus, drives the electrons into returning to the equilibrium position. The displaced electrons will naturally tend to oscillate with a certain frequency that is known as surface plasmon frequency. If the frequency of the excitation light matches the frequency of the nanoparticle surface plasmon, a resonant effect occurs (called localized surface plasmon resonance – localized SPR), which involves the collective oscillations of the electrons in the nanoparticles. This in turn causes an increase in the kinetic and electrostatic energies associated with the electric fields of the dipole which is supplemented by the exciting radiation [10]. As a result, the incident light is partially extinguished when exciting SPRs.

Noble metal nanostructures have the ability to confine the radiation well beyond the diffraction limit. This phenomenon happens when the incident light is at the SPR of the nanostructures. The SPR is strongly affected by the particle shape and size. This is clearly exemplified by gold nanorods (GNRs). In this case, since rods are geometrically anisotropic, the electron oscillation can occur differently along the rod's two main axes, depending on the polarization of the incident light. The charge accumulations will be different for the electrons oscillating along the rod long axis (longitudinal plasmons) and those moving along the perpendicular direction (transversal plasmons) (Fig. 4a). Accordingly, GNRs show two surface plasmon resonances, one strong longitudinal SPR (called LSPR) which is usually located in the NIR region and one transverse surface plasmon resonance (TSPR), which lies in the visible range (Fig. 4b). The LSPR of GNRs strongly depends on the aspect ratio (length/width) of the GNRs, that is, the higher the aspect ratio of the rods, the greater is the red shift associated to the LSPR, and vice versa (Fig. 4c). The strong LSPR of GNRs in the NIR region offers the possibility to use excitation light of larger wavelength, which is preferable for biological applications as these wavelengths are less harmful to the biological tissues. This contrasts to other geometries such as nanospheres where only one SPR is observed and its location can vary depending on the particle radius (Fig. 4d).

Effect of the Plasmonic Near-Field on Upconversion

The relaxation rate of each excited state of a lanthanide ion, incorporated as a dopant in a



Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics, Fig. 4 (**a**) Schematic of electron accumulation on the surface of a GNR (reproduced with permission from [10]), (**b**) the absorption spectra for GNRs showing the TSPR and LSPR, (**c**) the absorption spectra for GNRs with different aspect ratios from 1 to 5 (reproduced with permission from [10]), and (**d**) the absorption spectra for spherical nanoparticles with different diameters from 10 to 100 nm (reproduced from [11])

certain host, is characterized by a radiative emission probability (when energy is released in the form of light, i.e., photons), W_{rad}, and a non-radiative emission probability (when energy is released in the form of heat, i.e., phonons), W_{nr}. Typically, their sum is inversely related to the experimental lifetime of the excited state, τ_{exp}. However, we note that a third term can be added when energy transfer is possible from this state. In this case, τ_{exp} can be defined as:

$$\frac{1}{\tau_{exp}} = W_{rad} + W_{nr} + W_{ET} \qquad (1)$$

where W_{ET} is the probability of energy transfer. In general, upconversion processes benefit from long lifetimes in the intermediate states and large radiative emission probabilities in the final emitting state. Accordingly, these are the parameters to consider in order to improve the upconversion efficiency. Normally, these values depend on both the lanthanide and the host (via the crystal field, the phononic modes, and the presence of other lanthanides). Thus, in order to modify these parameters, a change of the host structure is often necessary. The so-called Purcell's effect states that the radiative properties of an emitter are dependent on its environment and, therefore, offers an alternative to optimize the upconversion. In the presence of a metallic nanostructure, the plasmonic interaction induces an increase in the density of states which, according to Fermi's golden rule, will consequently increase the radiative transition rate of the ions [7].

Equation 1 considers the different depopulation routes of an energy state. Naturally, absorption must take place prior to depopulation. Therefore, its control constitutes an additional option for improving the quantum yield, as the plasmonic electromagnetic field (which is connected with an inhomogeneous redistribution of the photon energy density in the nanostructure surroundings) clearly affects this important parameter.

In order to promote absorption enhancement, the plasmon must be resonant with one (or more) among the absorption lines of the lanthanides, which can then absorb its energy. In upconversion processes, the emission intensity, I, is strongly dependent on the excitation power density, P, due to the nonlinearity of this excitation scheme. This dependence follows a power function, $I \propto P^n$, where n is a positive value related to the number of excitation photons needed to complete the upconversion process, i.e., the number of upconversion steps [12]. Thus, it is clear that the high-photon energy densities achieved in the plasmonic field can strongly affect the intensity emitted by the UCNPs, through the enhancement of the absorption process. In this case, both the absorption processes originating from an excited state (e.g., ESA), as well as stimulated-type emissions, should be considered toward an optimized process and this in addition to the standard absorption from the ground state [7, 8].

The different ways in which the plasmonic field can affect all the steps in an upconversion process make the final enhancement larger than in a standard fluorescent particle [13]. In that regard, one further difference between upconversion and standard fluorescence is that, in upconversion, absorption and emission wavelengths are well separated, while this is not normally the case in standard fluorescence. This offers an additional route to engineer the upconversion emission, since the anisotropy of certain metallic nanostructures, allowing for more than one SPR, can be used to have enhancement both at the excitation and at the emission wavelength of the UCNPs [14]. The overall enhancement is not described by a general rule, as it depends on each particular upconversion process through parameters such as donor-acceptor distances, nature of the process (ETU vs ESA), and the number of upconversion steps.

The Non-Radiative Energy Transfer

When an excited UCNP is in the close environment of a metallic nanostructure, there is a certain probability of having a non-radiative energy transfer from the UCNP to the metallic counterpart, resulting in some energy loss and, as a consequence, quenching of the emission. In such cases, the distance between the emitter or sensitizer (the UCNP, S) and the acceptor (the metallic nanostructure, A), R_{SA}, is a critical parameter in defining the process probability. Note that UCNPs are by definition nanoparticles of transparent crystals with lanthanides distributed as dopants in the lattice. Consequently, the characteristics of the particle play a role in determining R_{SA}, which is thus not described by a single value, but rather by a probability distribution that should consider the dimensions, morphology and dopants concentration of the nanoparticle.

The non-radiative energy transfer can happen through three different processes: exchange (overlap of the electronic wave functions), multipolar electric, and multipolar magnetic interactions. Due to the nature of the first process, Ln^{3+} ions must be very close to the metallic nanostructure (few Å) for it to happen, and thus, (although it can be indeed important at short distances) it is often negligible in real experimental settings. Magnetic interaction is also relevant at very short distances, but only when exchange and electric interaction are otherwise forbidden by the relevant selection rules. Consequently, it is the multipolar electric interaction, defined by the Coulomb law, which more often accounts for energy transfer [15].

In many cases, the electric interaction can be described via a dipole-dipole approximation where the Förster formalism can be applied [16]. When this approximation is valid, the probability of the non-radiative energy transfer is related to R_{SA} through the relation $W_{ET} = C_{SA}/R_{SA}^6$, where C_{SA} is known as the energy transfer microparameter. This relationship clearly shows the strength of the dependence of the degree of quenching

on the interparticle distance, so that the process has a particular relevance when the surface of the UCNP is in contact with the metallic nanostructures (in which case, additional interactions and multipolar electric terms might be relevant).

However, R_{SA} is not the only quantity that affects the probability of the process, since the microparameter C_{SA} contains a dependency on the refractive index of the environment as well as on the emission and absorption energies of sensitizer and activator. Following Förster's model, C_{SA} can be defined as

$$C_{SA} = \frac{3\hbar^4 c^4}{4\pi n^4} \frac{Q_A}{\tau_0} \int \frac{L_S(E) \cdot L_A(E)}{E^4} dE \qquad (2)$$

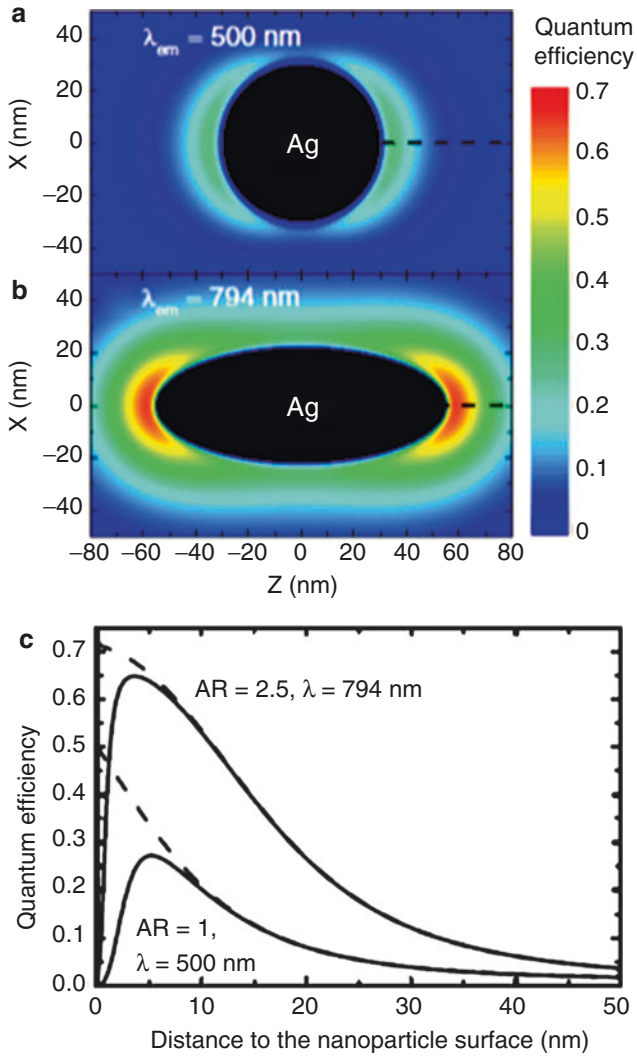
where the subindex S refers again to the sensitizer, i.e., the ion in the UCNP that provides energy in the transfer process and A refers to the activator, i.e., the receptor of the energy, which in the present case is the metallic nanostructure. Accordingly, L_S and L_A are the normalized line shapes of the sensitizer's emission and the activator's absorption, respectively, so the integral represents the overlap between both transitions. Q_A is the integrated absorption cross section of the activator, ħ is the Planck constant, c is the speed of light in vacuum, n is the refractive index of the medium, and, finally, τ_0 is the radiative lifetime of the sensitizer in the absence of the activator. Looking at this definition, it becomes clear that for a metallic nanostructure with a high-absorption cross section and in resonance with at least one emission line of the lanthanides, the quenching related to the energy transfer can become very relevant.

Controlling Enhancement in a Real System: The Key Parameters

The physical phenomena explained thus far are the main reasons that account for the modification of the upconversion intensity. Although the emission enhancement of the UCNPs in close proximity to the plasmonic nanostructures is now well established, studies on this matter have reported a wide range of enhancement factors or even quenching in some cases. A comparison between different approaches and efforts generated to date shows that upconversion emission enhancement can be strongly dependent several features that can affect any of the two main physical phenomena, plasmon mediated enhancement and energy transfer, behind the metallic nanostructure-UCNP interaction. Here, we describe a few of these features to illustrate which parameters are more relevant.

Impact of the Nanostructure Architecture

Plasmon-induced electromagnetic fields are strongly dependent on the shape, size, and composition of the metallic nanostructure. Thus, wide ranges of luminescence enhancement factors were found in previous works due to the interplay of UCNPs and metallic nanostructures in different configurations. For example, regarding the choice of material, it has been shown that silver nanoparticles induce a higher emission enhancement in comparison to gold nanoparticles with the same size, since the former possesses stronger plasmonic resonances [17]. On the other hand, for the same composition but different morphology, recent theoretical studies show that prolate-shaped metallic nanoparticles can lead to much higher enhancements than the corresponding spherical nanoparticles (Fig. 5). It was demonstrated by the work of Mertens and Polman that the emission enhancement in the proximity of a prolate geometry can be up to two times higher in comparison to the case in which the nanoparticle is a sphere with the same volume [18]. This significant improvement is due to two effects. Firstly, the coupling between the UCNP emitter and the plasmon mode is more effective in the prolate geometry due to the higher shape-induced field enhancement near the sharp tips [18]. Secondly, the drop in quantum efficiency close to the prolate-shaped nanoparticle surface sets in at smaller separations. For example, the distance to the metal surface at which the quantum efficiency reaches half the maximum value is 2.2 nm for a spherical nanoparticle and only 0.8 nm for its prolate-shaped counterpart. Quenching due to coupling with higher-order plasmon modes is thus less effective for prolate-shaped nanoparticles [18]. In a different work by Mertens et al. [19],



Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics, Fig. 5 Luminescence quantum efficiency of a dipole emitter as a function of its position in a plane through the nanoparticle's major axis for (**a**) a 60-nm-diameter Ag sphere and (**b**) an Ag prolate spheroid-shaped nanoparticle with an aspect ratio of 2.5 and a volume that is equal to the volume of the sphere. The orientation of the source dipole is set as parallel to the *z*-axis, and the refractive index of the embedding medium is 1.5. The emission wavelengths thus correspond to the maximum radiative decay rate enhancement: 500 nm for the sphere and 794 nm for the prolate-shaped nanoparticle. (**c**) Line traces of the luminescence quantum efficiency along the *dashed lines* indicated in (**a**) and (**b**), taking into account coupling to either all multipole modes up to $l = 100$ (*solid lines*) or to the dipole mode only (*dashed lines*) (Reproduced with permission from Mertens et al. [18]. DOI 10.1063/1.3078108. Copyright [2009], AIP publishing LLC)

the composition and shape of the metal nanoparticle are kept constant to address another possible contribution: the effect of the size of spherical plasmonic nanoparticles. It was proven that there is an optimal diameter in the range of 30–110 nm (depending on the material properties) for which maximum enhancement takes place, defined by the competition between two different mechanisms: emitter-plasmon coupling, which is more effective for small spheres, and coupling of

plasmons into radiation, which is more efficient for large spheres [19].

Emission/Absorption Enhancement in Upconversion

As described earlier, at wavelengths resonant with the metallic nanostructure plasmon frequency, the interaction between the incident light and the metal nanostructure results in local electric field strengths that are much higher than the incident electric field. This field enhancement is the main reason for the luminescence enhancement in UCNPs. However, it should be noted that the plasmon resonance of the metallic nanostructure could be designed to overlap with either the absorption wavelength of the UCNPs (in the NIR region), with the emission wavelength of the UCNPs (visible region), or both. To date, most works have focused on the resonant emission enhancement. For example, in the work by Saboktakin et al. [17], metal nanoparticles of gold or silver were fabricated to match the plasmon resonance with the emission wavelength of the UCNPs (the SPR of gold nanospheres was at 534 nm, which is in resonance with the green emission of Er^{3+} ions centered at 545 nm), and the enhancement was found to be 5.2-fold in proximity of Au nanospheres. However, the use of nanostructures such as GNRs that can possess two surface plasmons resonant with both the absorption and emission wavelength is expected to give rise to even higher enhancement values. GNRs can be tailored to have their longitudinal SPR overlapping with the excitation wavelength of the UCNPs and their transverse SPR overlapping with the upconversion emission of UCNPs. In such structures, the overall enhancement mechanism can be a combination of resonant enhancement in both absorption and emission, which can lead to an overall increase of the total enhancement.

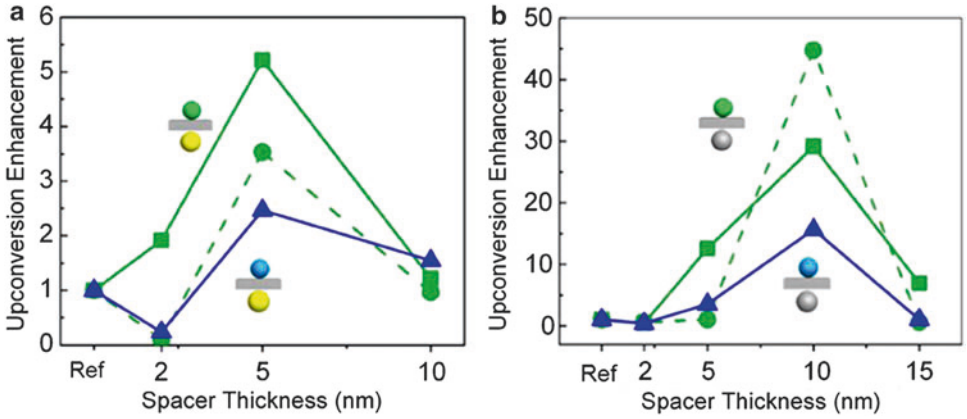
Effect of the Excitation Pump Power

As previously discussed, upconversion efficiency itself is a function of the excitation power density, $I \propto P^n$, and, thus, the enhancement factors from the same sample could vary widely depending on the excitation power. In this equation, n is presumed as the number of photons needed to excite the emitting state, also known as the order of the upconversion process. For example, for a two-photon upconversion process, the value of n should be close or equal to 2. However, it is theoretically proven that this relation strictly refers to the case of small upconversion rates at low excitation powers, and an efficient upconversion system will exhibit an intensity-versus-power dependence with lower n [12].

Due to the characteristics of the upconversion emission, a dependence of the plasmonic enhancement on the excitation power becomes intuitive. Its effect on the different contributions to upconversion enhancement (absorption, emission, and Förster energy transfer between Ln^{3+} ions) has been studied in the experimental work by Lu et al. [20]. In this work, a silver nanograting coated with a spacer of Si_3N_4 and featured by a plasmon resonance at the excitation wavelength of the UCNPs was probed at power densities between 1 and 100 kW/cm^2. It was found that the described general behavior of the upconverted luminescence was preserved in the presence of the surface plasmon, meaning that the upconverted luminescence intensity depends quadratically on the excitation power for a weak excitation, and linearly in the strong excitation limit. Yet, the intensity-versus-power dependency curves were shifted to lower-power densities, revealing an enhanced local electromagnetic energy density due to the surface plasmons. Additionally, through a numerical analysis based on rate equations, it was concluded that the enhancement of the upconversion luminescence results from an increase in absorption and energy transfer rates. The absorption enhancement was found to play a dominant role at high-power densities, while both absorption and energy transfer enhancement contributed in the weak excitation regime (here, the absorption enhancement was described by Maxwell's equations and energy transfer by Förster's theory) [20].

Impact of the UCNP-Metallic Nanostructure Interparticle Distance

As described earlier, the local field enhancement can lead to an increased upconversion emission,



Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics, Fig. 6 Upconversion luminescence enhancement of two types of UCNPs ($NaYF_4$:Yb^{3+}, Er^{3+} and $NaYF_4$:Yb^{3+}, Tm^{3+}) in close proximity of Au (**a**) and Ag (**b**) spherical nanoparticles with a diameter of 5 nm as a function of the spacer thickness between the two nanostructures. The *green solid lines* correspond to the 540 nm emission band of $NaYF_4$:Yb^{3+}, Er^{3+}. The *green dashed lines* correspond to the 650 nm emission band of $NaYF_4$:Yb^{3+}, Er^{3+}. The *blue solid lines* correspond to the 475 nm emission band of $NaYF_4$:Yb^{3+}, Tm^{3+} (Reproduced with permission from Saboktakin et al. [17])

which may impose the idea that the closer the UCNPs are to the plasmonic nanostructures, the better is the enhancement. However, the non-radiative energy transfer from the UCNPs to the plasmonic nanostructures can happen in very close proximity, leading to a decrease of the luminescence quantum yield (quenching). The combination of the plasmonic field enhancement of the upconversion intensity and the non-radiative energy transfer from UCNPs to plasmonic nanostructures is such that the luminescence intensity can be either enhanced or quenched depending on their relative distance [14]. Consequently, this results in the existence of an optimal distance (d_{opt}) for luminescence enhancement. Anger et al. investigated, both theoretically and experimentally, the effect of a gold nanoparticle on a fluorescent molecule [9]. They illustrated that a simple change in the relative distance can lead to a luminescence emission widely ranging from enhancement to quenching, thus proving the existence of a well-defined d_{opt}. In addition, they showed that d_{opt} depends on the size of the plasmonic nanoparticle, where, for larger nanospheres, d_{opt} was found to be smaller. However, there are reports of either fluorescence enhancement or fluorescence quenching of the upconversion in the presence of metallic structures. While this can appear somehow confusing, it can be readily explained by the lack of control over the interparticle distance in the experimental design. In recent years, the study of the effect of interparticle distance on the luminescence enhancement has been extended to upconversion. For instance, in an experimental work by Saboktakin et al., the effect of the distance between UCNPs and two different types of metallic nanospheres (gold and silver) with 5 nm diameters was studied. Preliminary series of control experiments in this work showed that, in a random configuration (without a spacer between the two types of materials), the luminescence intensity was greatly reduced. However, by using a nanometer-scale oxide layer as the spacer between the nanostructures, the existence of an optimum distance was revealed and found to be dependent on the metallic nanosphere composition. For gold nanospheres, d_{opt} was equal to 5 nm, while, for silver nanospheres with the same diameter, d_{opt} was 10 nm (Fig. 6). This could be explained as the result of a higher per volume extinction cross section of silver in comparison to gold.

Note that the extinction coefficient in metallic nanostructures arises both from absorption and scattering processes; although following Mie's theory, it is dominated by absorption for small diameter nanoparticles (5 nm in this case) [17]. Therefore, the higher absorption cross section of Ag nanoparticles generates a stronger non-radiative energy transfer (Eq. 2) followed by a stronger quenching effect. Accordingly, a larger optimal distance for emission enhancement is observed in this case.

Cross-References

- ▶ Gold Nanorods
- ▶ Local Surface Plasmon Resonance (LSPR)
- ▶ LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design
- ▶ Metal Nanoparticles from First Principles
- ▶ Plasmon Resonance Energy Transfer Nanospectroscopy
- ▶ Plasmonic Structures for Solar Energy Harvesting
- ▶ Synthesis of Gold Nanoparticles

References

1. Cao, G.: Nanostructures and Nanomaterials: Synthesis, Properties and Applications, p. 581. Imperial College Press, London (2004)
2. Chen, X., Liu, Y., Tu, D.: Lanthanide-Doped Luminescent Nanomaterials: From Fundamentals to Bioapplications (Nanomedicine and Nanotoxicology). Springer, Berlin/Heidelberg (2014)
3. Auzel, F.: Upconversion and anti-stokes processes with f and d ions in solids. Chem. Rev. **104**, 139 (2004)
4. Liu, Y., Ai, K., Lu, L.: Designing lanthanide-doped nanocrystals with both up- and down-conversion luminescence for anti-counterfeiting. Nanoscale **3**, 4804 (2011)
5. Boyer, J.-C., van Veggel, F.C.: Absolute quantum yield measurements of colloidal $NaYF_4:Er^{3+}/Yb^{3+}$ upconverting nanoparticles. Nanoscale **2**, 1417 (2010)
6. Liu, G., Jacquier, B. (eds.): Spectroscopic Properties of Rare Earths in Optical Materials. Springer, Tsinghua (2005)
7. Shvets, G., Tsukerman, I. (eds.): Plasmonics and Plasmonic Metamaterials. Analysis and Applications, vol. 4. World Scientific, Singapore (2012)
8. Fisher, S., Hallermann, F., Eichellkraut, T., Plessen, G. V., Kramer, K.W., Biner, D., Steinkemper, H., Hermle, M., Goldschmidth, J.C.: Plasmon enhanced upconversion luminescence near gold nanoparticles-simulation and analysis of the interactions. Opt. Express **20**, 271 (2012)
9. Anger, P., Bharadwaj, P., Novotny, L.: Enhancement and quenching of single-molecule fluorescence. Phys. Rev. Lett. **96**, 113002 (2006)
10. Garcia, M.A.: Surface plasmons in metallic nanoparticles: fundamentals and applications. J. Phys. D Appl. Phys. **44**, 283001 (2011)
11. http://nanocomposix.com/
12. Pollnau, M., Gamelin, D.R., Lüthi, S.R., Güdel, H.U.: Power dependence of upconversion luminescence in lanthanide and transition-metal-ion systems. Phys. Rev. B: Condens. Matter **61**, 3337 (2000)
13. Esteban, R., Laroche, M., Greffet, J.J.: Influence of metallic nanoparticles on upconversion processes. J. Appl. Phys. **105**, 033107 (2009)
14. Rohani, S.: Coupling of gold nanorods with lanthanide-doped upconverting nanoparticles for luminescence enhancement and nanothermometry. Master's Thesis, INRS-EMT (2015)
15. Dexter, D.L.: A theory of sensitized luminescence in solids. J. Chem. Phys. **21**, 836 (1953)
16. Forster, T.: Zwischenmolekulare energiewanderung und fluoreszenz. Ann. Phys. **437**, 55 (1948)
17. Saboktakin, M., Ye, X., Oh, S.J., Hong, S.-H., Fafarman, A.T., Chettiar, U.K., Engheta, N., Murray, C.B., Kagan, C.R.: Metal-enhanced upconversion luminescence tunable through metal nanoparticle-nanophosphor separation. ACS Nano **6**, 8758 (2012)
18. Mertens, H., Polman, A.: Strong luminescence quantum efficiency enhancement near prolate metal nanoparticles: dipolar versus higher-order modes. J. Appl. Phys. **105**, 044302 (2009)
19. Mertens, H., Koenderink, A.F., Polman, A.: Plasmon-enhanced luminescence near noble-metal nanospheres: comparison of exact theory and an improved Gersten and Nitzan model. Phys. Rev. B **76**, 115123 (2007)
20. Lu, D., Cho, S.K., Ahn, S., Brun, L., Summers, C.J., Park, W.: Plasmon enhancement mechanism for the upconversion processes in NaYF4:Yb(3+), Er(3+) nanoparticles: Maxwell versus Forster. ACS Nano **8**, 7780 (2014)

Uptake/Internalization/Sequestration/Biodistribution

- ▶ Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms

Use of Nanotechnology in Pregnancy

Achini K. Vidanapathirana
Australian Research Council (ARC) Centre for Nanoscale Biophotonics, University of Adelaide, Adelaide, SA, Australia
Heart Health, South Australian Health and Medical Research Institute, Adelaide, SA, Australia

Synonyms

Chemotherapeutic; Contraception; Drug delivery; Embryotoxic; Fetus; Nanoparticles; Nanoscale technologies; Nanotechnology; Nanotoxicity; Placenta; Pregnancy; Teratogenic

Definition

This entry examines the applications of nanotechnology during pregnancy (current, research, and potential applications), the unique pregnancy-related characteristics that relate in the application of nanoscale materials, and the concerns on maternal/fetal well-being.

Overview

Nanotechnology is being progressively researched, and nanoscale products are manufactured for various engineering devices, consumer products, diagnostic, and therapeutic applications to be utilized during different human life stages including pregnancy. Pregnancy is a life stage combined with the unique physiological changes and characteristic remodeling of maternal tissues supporting the growth and development of the fetus(es). As a result of these adaptive mechanisms to support fetal growth, the drug metabolism, particle distribution, and elimination kinetics during pregnancy are significantly different from both males and nonpregnant females. The gestational transformations are primarily dependent on hormonal changes and are initiated in the preimplantation stage, progressing to its maximum in the third trimester. These physiological changes evolve into the lactation phase and are partially reversed postpartum. Such alterations, when occurring at the molecular level in conjunction with the changes in tissue permeability, become significant particularly when designing, developing, and applying nanoscale technologies during pregnancy. Three main entities, the mother, the fetus, and the placenta along with their physical/functional connections and interdependency, are the key considerations for nanotechnology use during pregnancy.

The development and application of nanotechnology becomes crucial during the pregnancy conditions along two different aspects. The positive aspect is that the products incorporating nanoscale technologies can be used from a biomedical perspective and also in consumer nano-products during all stages of pregnancy. The negative aspect and implications are the reported and postulated detrimental effects of nanoparticle exposure during this vulnerable, transforming life stage.

Constructive Applications of Nanotechnology in Pregnancy

The properties of nanoparticles make them resemble biological molecules and are modifiable, making nanotechnology utilizable during the very early stages of pregnancy. Commercially available home pregnancy test is one example where nanoparticles are currently in use to confirm gestation during early pregnancy. These diagnostic concepts can be further extended to nanoelectromechanical systems (lab-on-a-chip devices) for detailed diagnostics during the different stages of pregnancy in both the mother and fetus. These include early detection of fetal congenital abnormalities by analyzing maternal serum, amniotic fluid, and cord blood. The relatively higher sensitivity of the nanoscale analyzing techniques can be very useful in early pregnancy where minute volumes (nano- to microscale) of biological materials and tissues are involved. Nanoscale technologies are considered for utilizing in contraception

by incorporating various binding, transporting, release, and strength-related properties. Research is currently ongoing for designing intrauterine contraceptive devices (IUCD) containing nanoparticles to replace the conventional copper IUCD [1]. The applicability of nanotechnology for further strengthening the barrier contraceptive devices is also currently under investigation.

The ability to functionalize nanoparticles and tailor them to support individual tissue-specific needs makes it a very versatile technology especially in the mid- and later gestational stages. In circumstances such as in severe preeclampsia and cancer treatment, the benefits of maternal drug/nanoparticle administration outweigh the risks of minute fetal exposure to nanomaterials [2]. These nanoparticulate drug delivery systems offer a potential avenue to deliver diagnostic, therapeutic, or theranostic agents to tissues that are not directly in contact with the fetus and can be separated by the placental barrier properties. One example that is currently under investigation is the treatment of ectopic placental tissue or placental tumors by targeted nanoparticle delivery of chemotherapeutic agents such as doxorubicin. In these studies, EnGeneIC Delivery Vehicles (EDVs) which are nanocells that can promote tissue-specific delivery of drugs were loaded with the chemotherapeutic doxorubicin. These nanocells were targeted to the epidermal growth factor receptor (EGFR) that is very highly expressed on the placental surface and was reported to be able to regress placental cells *in vitro*, *ex vivo*, and *in vivo* [3].

Potential Negative Consequences of Using Nanotechnology in Pregnancy

The occupational, therapeutic, and general environmental exposure to nanoparticles during the preconceptional, gestational, and lactation periods may impose unknown risks affecting the healthy progression and outcome of pregnancies. Considering the currently available distribution profiles of different nanoparticles, these adverse effects can be manifested as alterations in pulmonary, cardiovascular, hepatic, renal, and immunological functions of the maternal side. The emerging evidence on selective placental transfer of nanoparticles raises concerns on their utility in obstetrics and on the possible harmful effects to fetal development. Similarly, detrimental effects of nanoparticle exposure as postulated with animal studies can influence conception, implantation, placentation, embryogenesis, organogenesis, growth, and development in human pregnancies. These toxicity concerns and their poorly understood mechanisms are controversial and may at times contribute to decelerate the application of nanotechnology during reproductive life stages.

Key Research Findings

Relatively less number of studies is reported in literature on the application of nanoparticles, nanoparticle distribution, and toxicity during pregnancy; hence, literature reviews on nanotechnology and nanotoxicity relevant to gestational stages are scarce. Almost all the *in vivo* studies are carried out in animal models including zebra fish, mice, and rats.

Distribution of Nanoparticles During Pregnancy and Placental Transfer

Research and the in-depth understanding of the distribution kinetics of nanoparticles are critical as such knowledge is essential in both therapeutics and toxico-pharmacology. Several studies have focused on the distribution of different types of nanoparticles during pregnancy. The studies on intravenously administered nonfunctionalized radiolabeled C_{60} fullerenes formulated in polyvinylpyrrolidone (PVP) have reported distribution (24 and 48 h post-administration) in both maternal and fetal tissues in pregnant rats when administered on the fifteenth day of gestation. Radioactivity (~6 %) was also distributed to the reproductive tract, placenta, and fetuses of the pregnant rats. The relative tissue distribution patterns varied between the pregnant and lactating female rats that may be dependent on the physiological changes related to gestation, permeability differences, and structural properties of various maternal tissues. While the majority of the C_{60} were

distributed to the maternal liver, there was a limited distribution to the maternal lungs during pregnancy. In contrast, this study reported a higher distribution in the lungs in the lactating stage [4].

Relatively larger and longer carbon-based nanoparticles such as multiwalled carbon nanotubes (MWCNT) also cross the placental barrier as revealed by a study using isotopic tracers in mice. Oxidized MWCNT had crossed the placental barrier and were detectable in the fetal tissues [5]. Similarly, in a study where rats were periconceptionally exposed with gavages of zinc oxide nanoparticles, the distribution of zinc to mammary tissue in the pregnant rats and the liver and kidney of the pups was reported [6]. Metallic silver nanoparticles, when gavaged to pregnant and lactating rats, were also detected in the fetal tissues and breast milk [7].

Placenta is the key organ undergoing remodeling throughout the pregnancy, contributing to the systemic changes in pregnancy while facilitating nutrient, antibody, and other material particle transfers between maternal and fetal tissues. It also performs an important barrier function protecting the fetus from the maternal insults. The nanoparticle transfer across the placenta and the permeability characteristics are emerging as research foci in the field of reproductive nanomedicine and nanotoxicology [8]. Studies in pregnant mice exposed to different functionalized gold nanoparticles have reported placental and fetal accumulation of maternally administered nanomaterials suggesting the placental transfer of these materials. This particle transfer across the placenta is influenced by both stage of embryonic or placental maturation and nanoparticle surface composition [9]. The initial *in vivo* and *ex vivo* studies demonstrate that the ability of the nanoparticles to cross the placental barrier and passage to the fetus is dependent on the size and the surface coating of nanoparticles as well as on the experimental model [10].

The abovementioned studies were done in animal models, and the basic structural differences and deviations from the human placental function have to be considered in interpreting these findings as relevant to human exposures. Placenta is a species-specific organ, and the speculation of these findings particularly in relation to particles with less than 100 nm size range needs to be done with caution as the differences at cellular/molecular level may have a significant influence on these inferences. Studying the distribution of nanoparticles in the human placenta is an ethical challenge, and the only available research tools are the *ex vivo* perfusion of the expelled placenta and *in vitro* assessments of placental cells [8]. Previously reported studies in the human placenta have not observed any transfer of gold nanoparticles to the fetal side of the placental circulation [11]. Another study which used the *ex vivo* human placental perfusion model to investigate whether nanoparticles can cross the placental barrier has demonstrated that fluorescently labeled polystyrene beads with diameters up 240 nm were taken up by the placenta and were able to cross the placental barrier without affecting the viability of the placental explants [12].

Nanotoxicological Studies During Pregnancy

Most of the researches on pregnancy and nanotechnology have focused on the adverse effects of occupational, environmental, and medical-related nanoparticle exposure during pregnancy. The available literature on toxicological studies on pregnancy have focused primarily on pulmonary, intravenous, and oral exposure to different nanomaterials including silver, gold, silica, iron, cobalt, titanium, cerium oxide, carbon-based nanoparticles, and quantum dots. These studies have been done in different animal models in varying doses (usually higher doses), sizes, and formulations of nanoparticles. These *in vivo* exposures in mice have resulted in adverse effects in both fetal and maternal components. Increased acute cellular inflammation, inhibition of maternal weight gain, and changes in pregnancy hormones were reported in pregnant dams [5, 11]. The fetal outcomes include reduced body weights of pups, changes in gene expression related to development or function of central nervous system in male pups, increased susceptibility to allergy in pups, reduced number of Sertoli cells, and alterations in epididymal sperm motility in pups.

C_{60}, quantum dots, silica, and titanium nanoparticles are reported to be associated with embryological abnormalities, placental abnormalities, and early pregnancy loss [11, 13].

The changes observed in these studies in the fetal weight and the placental structure were dependent on the type, size, and surface chemistry on the nanoparticles [13]. The oxidized MWCNT which cross the placenta induced gestational age and parity-dependent abortion rates and also resulted in narrowed blood vessels and decreased the number of blood vessels in the placenta as assessed by histology [5]. Quantum dots are potential nanoparticles for imaging which can be used in pregnancy-related diagnostic, imaging, and monitoring progression of diseases/treatments. Studies in rats exposed with CdSe/ZnS or CdTe quantum dots on the 6th, 13th, and 18th days of embryogenesis did not report any direct embryotoxic or teratogenic effects. However, they reported adverse effects on the maternal organism as they induced necrosis in the tissues of the peritoneal cavity. In the same study, spectroscopic and microscopic tissue examination revealed the accumulation of quantum dots in the placenta in the absence of penetration to the embryonic tissues [14].

Gold nanoparticles are potential drug carriers that can be coated or formulated differently to make them biocompatible. In a study with intravenously administered gold nanoparticles with different surface modifications (ferritin, PEG, and citrate) at 5.5–15.5 days of gestation in mice, it was demonstrated that prior to gestation day 11.5, all tested nanoparticles could be visualized and detected in fetal tissues in significant amounts. Subsequently, fetal gold levels declined dramatically after gestation day 11.5. In contrast, gold nanoparticle accumulation in the extra embryonic tissues increased 6–15-fold with gestational age. Fetal and extra embryonic accumulation of ferritin- and PEG-modified gold nanoparticles was considerably greater than citrate-capped nanoparticles [9]. These and other studies on placental and fetal transfer of nanoparticles suggest that fetal exposure to nanoparticles and their toxicity in pregnancy is influenced by both stage of embryonic/placental maturation (gestational stage) and physicochemical properties of nanoparticle.

The changes in the fetal growth and abnormities induced by nanoparticle exposure are speculated to be due to several mechanisms. The first mechanism is explained by the placental transfer of specific nanomaterials and their directed effects on the fetal tissues. The interruptions of the placental and uterine blood supply and the contribution of the maternal cardiovascular system during pregnancy in mediating the effects of nanoparticles are combined as the second mechanism and are emerging as a significant area in research. The changes in maternal and fetal microvascular function and the contribution of these have been studied with exposure to titanium nanoparticles [15]. In addition, studies using C_{60} and multiwalled carbon nanotubes have reported vascular bed, suspension medium, and route of exposure-dependent changes in the vascular tissue responses and fetal growth during pregnancy [16, 17]. The inflammatory responses in the maternal side (generally following a pulmonary exposure) are another mechanism that can contribute to both maternal and fetal responses following gestational exposure.

Future Directions for Research

Pregnancy and nanotechnology in combination initiate a wide spectrum of research with infinite possibilities of applications that can be studied. These range from nano-engineering, basic molecular mechanisms, preclinical studies in animal models or *in vitro* systems, clinical trials, and toxicological (teratogenic and embryotoxic) assessments.

Interactions of the Nanoparticles with the Placenta

The main obstacle in applying nanotechnology during pregnancy is the lack of proper understanding on the precise nanoparticle interactions with the placenta. Limited studies focusing on this topic suggest a considerable uncertainty and

ambiguity regarding the transplacental deposition and passage of nanoparticles [2]. Our understanding on the criteria that determine placental transferability of nanoparticles is inadequate. Designing different nanoscale particles and manipulating their functional properties can potentially control the placental function as a barrier, filter, or a target to nanoparticles, provided the properties of placenta that contribute to these functions are properly understood. These are the first-line focus areas for future research in this field. In order to adequately understand nanomaterial-biokinetics and placental toxicity in early gestation (particularly the first trimester where teratogenic effects are critical), the special architecture of the placenta, the hypoxic condition, the bilayer of villous trophoblast, the plugging of spiral arteries, the contribution of intrauterine glands to nutrition, and the delicate immunologic situation at the implantation site need to be considered [18]. The physicochemical properties, functionalizing potential, targeting mechanisms, and radiolabeling ability of nanoparticles can be used as a research tool to study placental transport mechanism in further studies.

Almost all the current reported studies on nanotechnology in pregnancy are in animal models, and the speculations of these results to human populations are limited by the species-dependent differences in the placental structure, gestational periods, litter size, and particle distribution. Descriptive, retrospective studies on exposed human populations and further *ex vivo* placental perfusion studies on expelled human placentae can be utilized to fill this knowledge gap. The emerging knowledge and applications of nanoscale biophotonics expand the possibilities of applying these fine imaging techniques to study the human placental molecular mechanisms that govern the placental functions including their derangements in conditions such as preeclampsia.

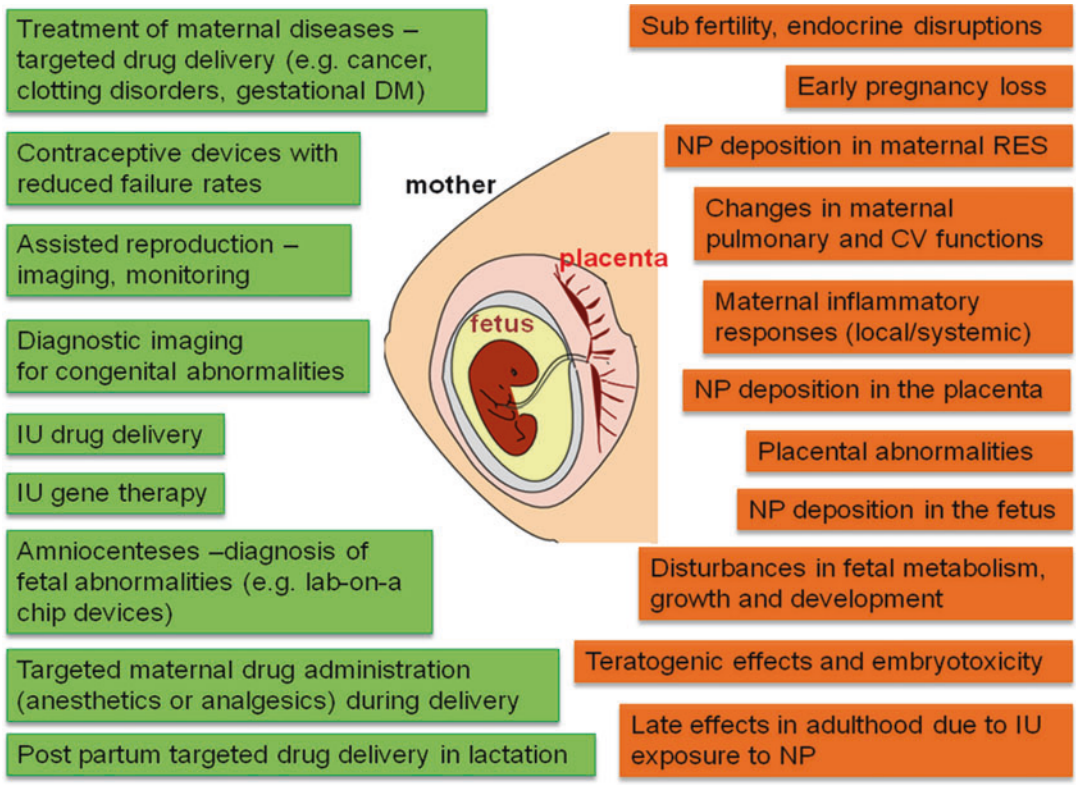
Utilizing Nanotechnology During Pregnancy

One of the significant and ambitious research questions is whether these nanoparticles can be used as modes of drug delivery during pregnancy. This can be in the form of targeted drug delivery to specific maternal tissues while minimizing the fetal adverse effects during pregnancy. The existing knowledge on targeted drug delivery for cancer treatment can be extended and applied to the pregnant state to selectively deliver drugs to the placenta and the fetus. These technologies can be then utilized to diagnose and treat congenital conditions during the intrauterine life. Intrauterine gene therapy and immunomodulation are few examples of the potential applications that could be used during pregnancy. Alternatively, these applications can be studied and applied in treating maternal disease conditions such as progressive cancers, diabetes mellitus, coagulopathies, transplant rejection, preeclampsia, and immunological disorders (e.g., asthma, rheumatoid arthritis, systemic lupus erythematosus) during pregnancy. These applications include imaging, diagnostics, monitoring, and theranostics for maternal disease treatments without affecting the fetal growth [8]. Minimizing fetal side effects using the permeability and transfer properties of the placenta and engineering the nanoparticles to suit these barrier mechanisms will lay the foundations for such applications.

Further areas of research for pregnancy-related nanotechnology include the management of subfertility and diagnosis of preeclampsia using serum or placental markers. Novel fine nanoscale sensing mechanisms such as nanoscale biophotonics can be developed for procedures and functions that need nanoscale sensitivity. These include sensors that can be applied in subfertility and in planned pregnancies. Assisted reproductive procedures such as artificial insemination, *in vitro* fertilization, and implantation would benefit from such devices.

Refined Nanotoxicological Studies

The current nanotoxicological studies are limited largely to nonfunctionalized, higher doses of nanoparticles and are aimed at more mechanistic approaches. The actual formulations of nanoparticles that can be utilized in nanomedicine are different and involve tissue-specific



Use of Nanotechnology in Pregnancy, Fig. 1 The future of nanotechnology use during pregnancy. Pregnancy consists of three critical entities (the mother, fetus, and placenta) and is a unique stage where nanotechnology can be used for various applications. The fine balance between these constructive applications (in *green boxes*) and potential maternal or fetal adverse effects (in *orange boxes*) needs to be considered with caution before incorporating nanotechnology in pregnancy (*CV* cardiovascular, *DM* diabetes mellitus, *IU* intrauterine, *NP* nanoparticles, *RES* reticuloendothelial system)

functionalization in realistic doses and formulations. In addition, the adverse effects of nanoparticles during the first trimester are relatively understudied, but critical [18] and should be a main research focus in the future. This also includes the prepregnancy ovum production (starting as early as the intrauterine stage of the female), puberty, fertilization, and implantation stages when the molecular mechanisms can be disrupted or manipulated by using nanotechnology. The contribution of nanoparticles on miscarriages, intrauterine growth restriction, preterm deliveries, and preeclampsia also need to be further studied. These conditions are reported with other toxins and inflammatory conditions. Such adverse consequences need to be excluded before nanotechnology can be applied to the pregnant status. Limited/narrow-scale research and anecdotal evidence on pregnancy-related complications are reported in pregnant mothers working in nanoparticle production facilities. These need to be further confirmed with extensive research on exposed populations.

The unique properties of nanoparticles that are technically constructive such as high surface-to-volume ratio, protein corona, and its evolution within the biological systems also increase their potential for toxicity. These properties make it difficult to assess the risk of nanomedical or nanotechnological products using the current risk

assessment methods particularly during pregnancy in which the maternal, fetal, and placental components are functioning in a complex environment. Pregnancy may be one scenario where the nanotechnology itself can be applied to assess the toxicity of different nanomaterials and other drugs.

Nanoscale technologies are very versatile with many advantageous applications during pregnancy if used with caution. The fine balance between the type, dose, frequency, formulation of nanoparticles, and stage of gestation with the relevant physiological changes will guide to the effective use of nanotechnology in pregnancy (Fig. 1). The effect of nanoparticle exposure on pregnancy stages need to be addressed before releasing any nano-combined diagnostic, therapeutic, or theranostic applications. As with any other cosmetic, consumer, or pharmacological product, teratologic risk assessment is essential with nanomedical applications. These include engineering, mechanistic, preclinical, clinical, and post-marketing monitoring strategies that need to be meticulously carried out prior to applying nanotechnology during pregnancy.

Cross-References

- ▶ Chemotherapeutic
- ▶ Contraception
- ▶ Drug Delivery
- ▶ Embryotoxic
- ▶ Fetus
- ▶ Nanoparticles
- ▶ Nanoscale Technologies
- ▶ Nanotechnology
- ▶ Nanotoxicity
- ▶ Placenta
- ▶ Pregnancy
- ▶ Teratogenic

References

1. Yu, J., et al.: Comparative study on contraceptive efficacy and clinical performance of the copper/low-density polyethylene nanocomposite IUD and the copper T220C IUD. Contraception **78**, 319–323 (2008)
2. Menezes, V., Malek, A., Keelan, J.A.: Nanoparticulate drug delivery in pregnancy: placental passage and fetal exposure. Curr. Pharm. Biotechnol. **12**, 731–742 (2011)
3. Kaitu'u-Lino, T.J., et al.: Targeted nanoparticle delivery of doxorubicin into placental tissues to treat ectopic pregnancies. Endocrinology **154**, 911–919 (2013)
4. Sumner, S.C., Fennell, T.R., Snyder, R.W., Taylor, G. F., Lewin, A.H.: Distribution of carbon-14 labeled C_{60} ($[^{14}C]C_{60}$) in the pregnant and in the lactating dam and the effect of C_{60} exposure on the biochemical profile of urine. J. Appl. Toxicol. **30**, 354–360 (2010)
5. Qi, W., et al.: Damaging effects of multi-walled carbon nanotubes on pregnant mice with different pregnancy times. Sci. Rep. **4**, 4352 (2014)
6. Jo, E., et al.: Exposure to zinc oxide nanoparticles affects reproductive development and biodistribution in offspring rats. J. Toxicol. Sci. **38**, 525–530 (2013)
7. Melnik, E.A., et al.: Transfer of silver nanoparticles through the placenta and breast milk during *in vivo* experiments on rats. Acta Nat. **5**, 107–115 (2013)
8. Buerki-Thurnherr, T., von Mandach, U., Wick, P.: Knocking at the door of the unborn child: engineered nanoparticles at the human placental barrier. Swiss Med. Wkly. **142**, 3559 (2012)
9. Yang, H., et al.: Effects of gestational age and surface modification on materno-fetal transfer of nanoparticles in murine pregnancy. Sci. Rep. **2**, 847 (2012)
10. Kulvietis, V., Zalgeviciene, V., Didziapetriene, J., Rotomskis, R.: Transport of nanoparticles through the placental barrier. Tohoku J. Exp. Med. **225**, 225–234 (2011)
11. Ema, M., Kobayashi, N., Naya, M., Hanai, S., Nakanishi, J.: Reproductive and developmental toxicity studies of manufactured nanomaterials. Reprod. Toxicol. **30**, 343–352 (2010)
12. Wick, P., et al.: Barrier capacity of human placenta for nanosized materials. Environ. Health Perspect. **118**, 432–436 (2010)
13. Yamashita, K., et al.: Silica and titanium dioxide nanoparticles cause pregnancy complications in mice. Nat. Nanotechnol. **6**, 321–328 (2011)
14. Zalgeviciene, V., Kulvietis, V., Bulotiene, D., Didziapetriene, J., Rotomskis, R.: The effect of nanoparticles in rats during critical periods of pregnancy. Medicina (Kaunas) **48**, 256–264 (2012)
15. Stapleton, P.A., et al.: Maternal engineered nanomaterial exposure and fetal microvascular function: does the Barker hypothesis apply? Am. J. Obstet. Gynecol. **209**, 227 (2013)
16. Vidanapathirana, A.K., et al.: PVP formulated fullerene (C_{60}) increases Rho-kinase dependent vascular tissue contractility in pregnant Sprague Dawley rats. Reprod. Toxicol. **49**, 86–100 (2014)
17. Vidanapathirana, A.K., et al.: Vascular tissue contractility changes following late gestational exposure to multi-walled carbon nanotubes or their dispersing

vehicle in Sprague Dawley rats. J. Nanomed. Nanotech. **5**(3), 1000201 (2014)
18. Juch, H., Nikitina, L., Debbage, P., Dohr, G., Gauster, M.: Nanomaterial interference with early human placenta: sophisticated matter meets sophisticated tissues. Reprod. Toxicol. **41**, 73–79 (2013)

USM

► Ultrasonic Machining

US-Tubes

► Ultrashort Carbon Nanotubes

UV Lithography

► Photolithography

V

Vertically Aligned Carbon Nanotubes (VACNTs)

▶ Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior

Siddhartha Pathak[1] and Shelby B. Hutchens[2]
[1]Chemical and Materials Engineering, University of Nevada, Reno, Reno, NV, USA
[2]Mechanical Science and Engineering, University of Illinois Urbana-Champaign, Champaign, IL, USA

Synonyms

Bundles; Brushes; Carbon Nanotube (CNT) arrays; Foams; Forests; Mats; Turfs; Vertically Aligned Carbon Nanotubes (VACNTs)

Definition

Vertically aligned carbon nanotubes are complex, hierarchical structures of intertwined tubes arrayed in a nominally vertical alignment due to their perpendicular growth from a stiff substrate. They are a unique class of materials having many of the desirable thermal, electrical, and mechanical properties of individual carbon nanotubes, while exhibiting these properties through the collective interaction of thousands of tubes on a macroscopic scale.

Introduction

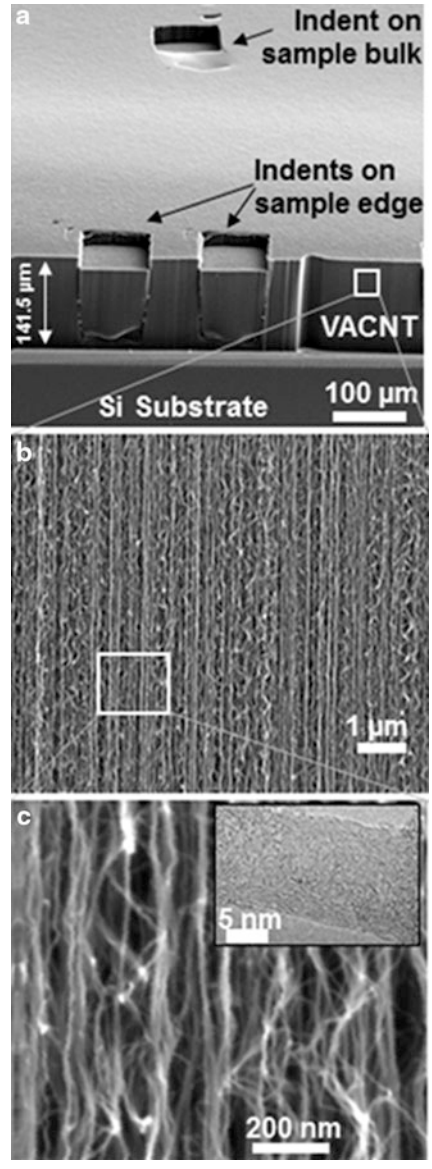
While individual CNTs have been announced as the strongest material known and have shown extremely high strength and Young's modulus in tensile tests on individual tubes (see ▶ Carbon Nanotubes), VACNTs are more likely to find use in applications requiring large compliance and deformability. Examples of these include microelectromechanical systems (MEMS) and impact mitigation/energy absorption, where they are promising candidates for their multifunctional nature, wide-ranging thermal stability, well-defined large surface area, and relative ease of manufacture. A proper understanding of the collective mechanical behavior of these structures, especially instabilities leading to buckling and inhomogeneities which weaken mechanical performance, is thus of great importance for their design and success in these and other future applications.

VACNTs are distinct from other CNT structures (see ▶ Nanomaterials) in that the tubes that make up the material grow perpendicularly to the support substrate, making them *nominally*

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

vertically aligned. An important characteristic all VACNTs share is that the tubes themselves are long enough to become intertwined with each other during the growth process, leading to a highly complex and hierarchical microstructure with distinct organizational details at different magnification levels. This is in contrast to arrays of vertically aligned CNTs that are short and/or sparse enough that each CNT stands alone [1, 2]. Figure 1 shows the complex hierarchical nature of the VACNT microstructure with their distinct organizational details across multiple length scales. Thus, while the VACNT appears as a continuous film at lower magnifications of 260× (Fig. 1a), the nominally vertical alignment of CNTs growing perpendicularly to the support substrate in the VACNT bundle becomes visible at a higher magnification of 30,000× (Fig. 1b). Increased magnification of 240,000× reveals significant intertwining in the long, curved lengths of individual CNTs, revealing the isotropic CNT network [3]. As a result, in their as-grown states, certain segments of the CNT forest appear to be pre-buckled/pre-bent, and the favorable contact energy between the tubes (van der Waals) is thought to balance the bending strain energy of their arrangement, resulting in a stable low-energy configuration. At still higher magnifications (nanometer length scales; see TEM image in Fig. 1c inset), individual, discrete CNTs dominate the mechanical performance with details like CNT diameter, number of walls, etc., governing their deformation. Such hierarchy in the VACNT microstructure governs its overall mechanical behavior, which is a result of not only the properties of individual CNTs but also of their complex mutual interactions and distribution throughout the array.

The method by which these structures are synthesized is the primary factor affecting their complex, hierarchical morphology. This microstructure, in turn, affects their mechanical behavior, in particular the modulus, buckling strength, and recoverability. Synthesis techniques for VACNTs can be divided into two main categories:



Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Fig. 1 SEM images reveal the hierarchical morphology of the (**a**) 141.5 μm-thick VACNT films (magnification 260×), which consist of (**b**) nominally vertical aligned CNTs visible at a lower magnification of 30k× and (**c**) a complex intertwined network seen at higher magnifications of 240k×. SEM pictures are taken at a 60 tilt angle. (**c**, *Inset*) Individual-multiwalled CNTs of outer diameter 8.8 ± 2.1 nm (average ± standard deviation) are visible in the TEM image ([4] reproduced with permission)

the chemical vapor deposition (CVD) synthesis method and the carbide-derived carbon (CDC) synthesis method – which relate to the top-down versus bottom-up growth processes, respectively. In the CVD process, the VACNT film is coated onto an existing substrate. This is accomplished by depositing a thin layer of catalyst (e.g., Fe) on the substrate (typically Si or Quartz) and flowing a carbon source (e.g., ethylene) over the substrate at atmospheric pressure and temperatures typically around 750 °C. In the CDC method [5], on the other hand, carbon is formed by selective extraction of the metal or metalloid atoms in the carbide (e.g., silicon carbide) at high temperatures (>1,600 °C), transforming the carbide structure into pure carbon. Since the CNT layer is formed by inward growth, this usually retains the original shape and volume of the precursor.

VACNTs grown by these different techniques demonstrate very different structure and mechanical properties. Even within materials grown via CVD, control of the growth conditions, such as the atmosphere, catalyst activity, and pressure, is known to significantly affect the repeatability of the VACNT's morphology and hence the consistency of mechanical properties [6]. For example, using "floating" versus "fixed" catalysts [7] in the CVD syntheses has been shown to result in vastly different VACNT morphologies. Differences in the growth processes used are revealed in the widely varying VACNT information reported in literature, as seen in the large range in properties in Table 1. This is not to mention variations in tube diameter (from 2–3 nm [8] to 20–50 nm [3] to greater than 100 nm [9]), number of walls in the CNT structure, and degree of tube alignment, properties which are sometimes neither measured nor reported. Further, the different stages of the CNT growth process can result in a height-dependent inhomogeneity [10]. This manifests as a gradient in both the density and the alignment of the tubes within the same VACNT structure. As discussed later in this entry, such a structure gradient may lead to a corresponding strength and stiffness gradient along the VACNT height. As an extreme example of the microstructure-property relationship, CDC-VACNTs are known to have a considerably higher average density (roughly ten times higher than typical VACNTs), due to the conformal transformation of the carbide into carbon. This in turn leads to significantly larger values for the elastic modulus and yield stress in CDC-VACNTs [8] (see Table 1). These promising characteristics, however, are unavailable for applications requiring macroscopic films as currently only VACNTs grown via CVD can reach macroscopic heights (~mm). Growth of CDC-VACNTs remains limited to only a few micrometers in height. Only CVD-VACNTs when used in combination with conformal coating methods, where the CNTs are coated with nanoscale coatings of ceramics like Al_2O_3 [11], amorphous SiC [12], etc. (Table 1), are able to reach a similar level of mechanical performances as compared to the CDC-VACNTs.

Figure 2 highlights four literature examples of the differences in the mechanical response of various CVD-VACNT micro-pillars subjected to compression. Note in particular the higher stiffness and strength of the VACNTs grown using the floating catalyst technique in Fig. 2a [13], as compared to the ones grown using the fixed catalyst method (Figs. 2b [7], 2c [17], and 2d [3]). Other differences between these nominally identical VACNT samples are the ability of some of them to recover almost completely after large compressions (Fig. 2a, c), [13, 17] while others deform permanently even at modest strains (Fig. 2b, d), [3, 21–31].

Figure 2 also demonstrates the similarities in the deformation characteristics of the various VACNT systems. All of the four VACNT systems show three distinct regimes in their stress-strain response – elastic, plateau, and densification – similar to open-cell foams. Unlike foams, however, the plateau region in the VACNTs generally has a positive slope, which can vary significantly between VACNT samples. The slopes of the plateau region were calculated to be 11, 5, and 0.6 MPa for Figs. 2a, c, d, respectively. Note also that all VACNT systems shown in Fig. 2 exhibit a bottom-to-top sequence of buckling.

Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Table 1 Summary of reported VACNT elastic modulus and yield strength values. *White rows* denote values for CVD-synthesized VACNTs, while *gray rows* are for CDC-VACNTs

VACNT details (Density, porosity etc.)	Measurement Method	Modulus	Yield/ Buckling Strength	Reference
CVD-VACNTs				
87% porosity	compression	50 MPa	12 MPa	[13]
/	compression	< 2 MPa	/	[14]
10^{10} tubes/cm^2 [a]	compression	0.22-0.25 MPa	/	[15]
97% porosity	compression	818 MPa	14.1 MPa	[16]
0.08 g/cm^3	compression	177±11 MPa	2.69±0.12 MPa	[17]
/	compression	9-31 MPa	0.15 – 0.3 MPa	[18]
0.008 g/cm^3 (fixed catalyst, argon anneal) 0.015 g/cm^3 (fixed catalyst, hydrogen anneal) 0.2 g/cm^3 (floating catalyst)	compression	1.86 MPa (range 1.3-3.26) 7.44 MPa (range 5.2-13.04) 138.5 MPa	0.021 MPa 0.117 MPa 9.997 MPa	[7]
0.018 g/cm^3 0.114 g/cm^3 (After post-growth CVD treatment)	compression	/	0.12 MPa 5.5 MPa (After post-growth CVD treatment)	[19]
0.12 g/cm^3	compression	0.55 MPa (parallel to the	/	[20]

(*continued*)

Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Table 1 (continued)

		nanotube axis) 1.09 MPa (normal to CNT axis)		
3×10^{10} / cm^2 (number density) 1 g/cm^3 (after Al_2O_3 coating)	compression	14 MPa 20 GPa (after Al_2O_3 coating)	compressive strength (0.8 MPa to 0.16 GPa)	[11]
10^{10} / cm^2 (number density)	compression	200 MPa 125 GPa (after amorphous SiC coating)	compressive strength <1 MPa (uncoated VACNTs) 1.8 GPa (after amorphous SiC coating)	[12]
/	nanoindentation – Berkovich uniaxial compression	15 MPa	0.2-4.3 MPa	[21]
0.08 g/cm^3	Flat punch indentation	120-175 MPa	1.49-1.75 MPa	[4]
0.13 gm/cm^3 0.06 gm/cm^3	Flat punch indentation	7-33 MPa 100-350 MPa	0.44±0.04 MPa 0.9 MPa	[22]
/	nanoindentation – Berkovich	58 MPa	/	[23]
/	nanoindentation - Berkovich	50±25 MPa	/	[24]

(*continued*)

V

Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Table 1 (continued)

/	DMA(Dynamic Mechanical Analysis)	~ 50 MPa	/	[25]
0.009 g/cm^3	DMA	~ 1 MPa	/	[26]
CDC-VACNTs				
0.95 g/cm^3	compression	30 GPa	800 MPa	[27]
0.95 g/cm^3	nanoindentation – Spherical	18-20 GPa	90-590 MPa	[8]

Here the first buckle generally nucleates close to the substrate, and each subsequent lateral collapse event initiates only after the preceding one was completed, thus sequentially collapsing the entire structure [3, 7, 13, 17]. A recent report by [18] has suggested that such sequential bottom-to-top buckling and hardening in stress-strain response are observed in samples with smaller relative density at the bottom. When the density gradient was insubstantial or reversed, a different buckling sequence was observed where the bottom regions were always found to buckle last and a flat stress plateau was obtained.

Experimental Techniques for Studying VACNT Mechanics

Instrumented indentation, using a variety of tip geometries such as flat punch, spherical, Berkovich, and cube corner, has been the most common method for studying mechanical properties of VACNTs [9]. Each geometry has its own strengths and weaknesses. While maintaining parallel contact between the indenter and the sample is a major concern for flat punch indentation, it does allow for more uniaxial-like, compressive loading. The sharper Berkovich or cube corner geometries can cause the CNTs to bend away from the indenter, testing a slightly different mode of behavior [1, 2, 9]. In both cases, modulus and hardness are measured from the unloading portion of the test. On the other hand, spherical indentation is advantageous in that it allows indentation stress-strain curves to be extracted from the raw load-displacement data, which enables resolution of the evolution of the mechanical response in the VACNT array: from initial elasticity to the initiation of buckling to post-buckling behavior at finite strains [8]. Though indentation is a relatively simple test to perform, analyses of the results, especially nonlinear elastic behavior, are difficult due to their highly localized stress fields. It is also limited in total strain.

Another testing geometry, that of uniaxial compression, eliminates these localized, applied stress fields and helps reveal the existence of a localized deformation mechanism in VACNTs. In these tests, the samples are either large (~1 mm tall) bulk films and compressed between two platens [13] or microscale (up to tens of microns in height) cylinders and compressed using a flat punch indenter [3, 7].

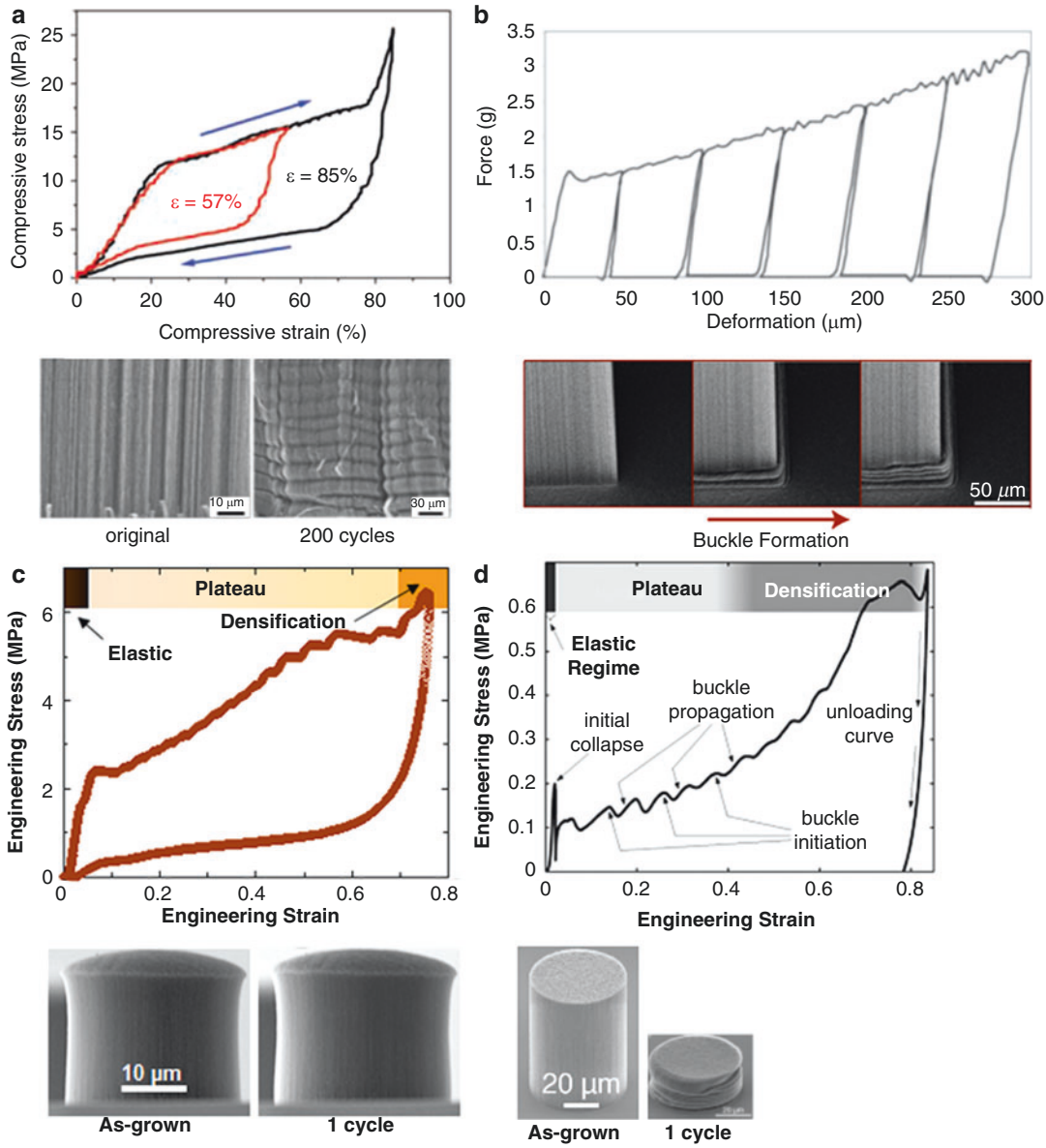
Both indentation and compression tests can be performed in situ, where the load-displacement data is gathered simultaneously with micrographs in an optical or scanning electron microscope (SEM). These tests can offer valuable insights on

the morphological evolution in the VACNTs during deformation and are discussed in detail in the next sections.

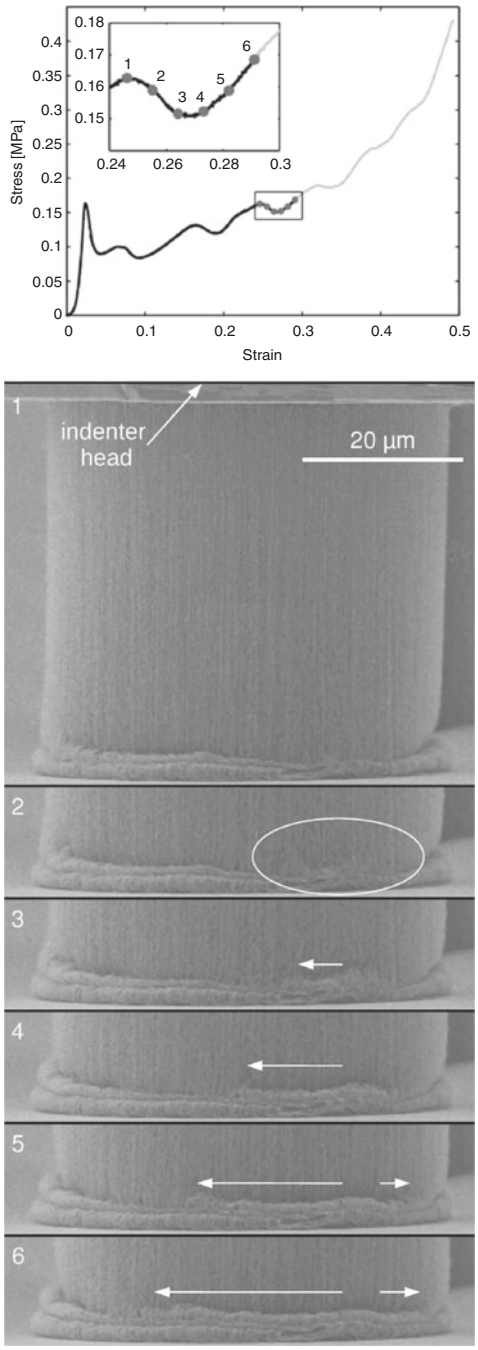
Compression Response Under Large Strain

The deformation of VACNTs is governed by their hierarchical microstructure, collective intertube interactions, and inherent property gradient. Taken as a whole, their highly porous nature gives them an overall foamlike response. The idealized stress-strain response of traditional foams under compression is characterized by three distinct loading regimes: an initial elastic loading at low strain, followed by a plateau in the stress during which the struts bend and buckle, and finally a densification regime in which the space between struts has been nearly eliminated and the material begins to approach behavior intrinsic to the struts themselves. For such a response, it is the intermediate plateau regime that is responsible for the bulk of energy absorption in the material, since the area under this region of the stress-strain curve, corresponding to the work done on the material, is the largest. This foamlike response of a bulk VACNT film is apparent in the three distinctly differently sloped regions of the stress-strain responses shown in Fig. 2. Locally, however, the response of VACNTs is quite unlike that of traditional foams. In VACNTs, the accommodation of strain during uniaxial compression is accomplished entirely through the formation of folds or buckles of small regions of the structure, while the remaining portion remains nearly undeformed. This is in contrast to traditional foams, where cell-edge bending and cell collapse are primarily responsible for the elastic-plastic foam response [32, 33]. The superposition of an overall foamlike response with localized strain accommodation is the key characteristic of VACNT deformation.

Experimental characterization of this buckle formation yields several interesting qualitative results. In their early study, [13] compressed relatively large structures (area, 0.5–2 cm^2; height, 860 μm–1.2 mm) and observed that the buckles formed near the bottom of the structure (the end from which the CNTs grow perpendicularly to the substrate) are more deformed than those that formed near the top (Fig. 2a). Motivated by this observation, they hypothesized that the bottom buckles form first. A reversal of the loading direction, by flipping the sample upside down, resulted in the same deformed morphology, with the tightest buckles forming at what was the end of the sample attached to the growth substrate. These observations point to the idea of an inherent, axial property gradient being responsible for the sequential nature of the buckling. Note that each individual buckle is on the order of 12–25 μm in size (depending on sample height) so that several tens of buckles form during deformation. The sequential, localized buckling phenomenon was later observed in much smaller samples by other researchers (Fig. 2b–d) illustrating the universality of this response in VACNTs. Buckles in these microscopic studies were 12 μm [28] and 7 μm [3] in wavelength (measured from the unbuckled conformation) for cylindrical samples with diameters of 30–300 μm [28] and 50 μm [3], respectively. The bottom-first buckling mechanism was visually verified by [3] through in situ experiments that further revealed the mechanism by which a single, localized buckle evolves. As shown in their study, each individual buckle does not form all at once, but rather nucleates at one point and then propagates laterally across the cylindrical sample until completion [3]. In addition, each subsequent buckle begins only after the previous buckle has completely formed, i.e., the crease had spread across the entire cylinder [3, 7, 13, 17]. Figure 3 illustrates the evolution and localization of deformation for a uniaxially loaded VACNT cylinder. Cross-correlation of the in situ images with the accompanying stress-strain data reveals that the localized buckle formation and propagation correspond to undulations in the nominal stress-strain curve. Specifically, softening corresponds to the first appearance of the buckle and the subsequent hardening coincides with the lateral propagation of the buckle. This localized response overlays the aforementioned overall foamlike behavior seen in these materials.



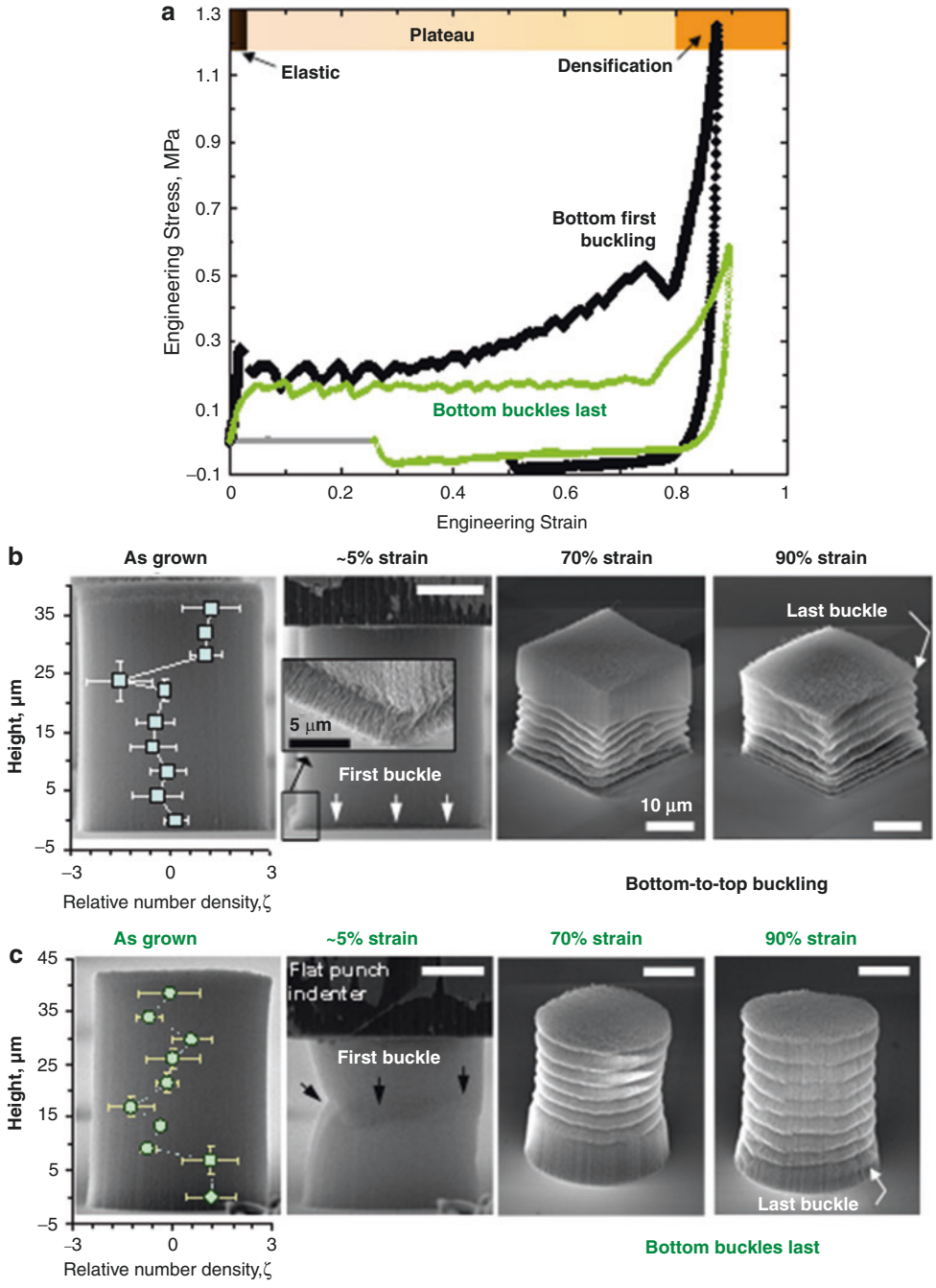
Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Fig. 2 VACNT behavior under compression for (**a**) 860 μm-thick VACNT array (Reprinted from [13] with permission from AAAS), (**b**) 400 μm-diameter CNT column [7] (Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission), (**c**) 30 × 30 μm (diameter × height) VACNT micro-pillar (Adapted with permission from [17] Copyright 2012 American Chemical Society), and (**d**) 50 × 60 μm (diameter × height) VACNT micro-pillar [3] (Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission). All of these VACNT systems show three distinct regimes in their stress-strain response – elastic, plateau, and densification – similar to open-cell foams. Unlike foams, the plateau region in the VACNTs generally has a strong positive slope. All VACNT systems shown here also exhibit a bottom-first sequence of buckling. These VACNT systems differ widely in their ability to recover from large deformations. Thus while both VACNT systems in (**a**) and (**c**) show an almost complete recovery, even after multiple cycles, the VACNT systems in (**b**) and (**d**) do not exhibit any appreciable recovery



Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Fig. 3 In situ video micrographs and accompanying nominal stress-strain data illustrating bottom-first buckling and buckle initiation and propagation events [3] (Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission)

Notable differences between the classic foamlike stress-strain behavior and the overall response of VACNTs in uniaxial compression studies of VACNT structures have also been observed in both macroscopic [13–15] and microscopic [3, 21] samples. First, the plateau regime is highly sloped. In typical foams, a sloped plateau indicates some homogeneous variation in strut buckling stress due either to random strut alignment, a distribution of strut sizes (aspect ratio, diameter, etc.), or both. While both kinds of inhomogeneities exist in VACNTs, the plateau regime is too highly sloped to be explained by these small fluctuations. Moreover, the nonlocal nature of the buckling points to an alternate cause. In fact, the sloped plateau regime is suggestive of the presence of an axial property gradient in VACNTs. As discussed previously, a gradient in tube density arises in VACNTs as a result of the CVD growth process. This gradient can be such that there is a lower tube density at the bottom of the structure (i.e., the point at which the substrate attaches) than the top. It follows that such a tube density gradient would result in a corresponding strength and stiffness gradient. This property gradient is evident in the stress-strain responses seen in Figs. 2 and 3 in the progressive increase in peak stress values for undulations in the plateau region. Throughout this plateau, buckles are known to form sequentially, bottom to top, and therefore each subsequent buckle forms at a higher (and more dense) location within the cylinder than the previous buckle, requiring a larger stress in order to form a new fold.

A recent study by [18] has reported a different buckling sequence in VACNTs. In addition to the commonly reported bottom-to-top buckling pattern (where the bottom buckle forms first), these authors also observed VACNT samples where the bottom buckle was instead the *last* buckle to form (Fig. 4c). These experiments were performed on two sets of VACNT micro-pillars, both grown by the same chemical vapor deposition (CVD) synthesis on the same Si substrate but located on different regions of the substrate. In situ uniaxial compression experiments conveyed that the



Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Fig. 4 (**a**) Comparison of the stress-strain responses of two VACNT pillar systems grown on the same substrate: one showing the commonly reported bottom-to-top buckling sequence where the bottom buckle is the first one to form (**b**) versus the ones where the bottom buckle is the last buckle to form (**c**). (**b**) and (**c**) also show the relationship between the relative tube number density ζ and the sequence of buckle formation in these pillars (Adapted with permission from [18] Copyright 2013 American Chemical Society)

sequence in the localized fold formation was unique for each sample type (Fig. 4b, c). In the first set (Fig. 4b), the first buckling-like instability, which corresponded to the transition from elastic loading to plateau in the stress-strain data, always formed at the bottom of the sample. After initiation, the fold propagated laterally until it fully spanned the pillar width. Bottom-to-top buckling occurred in succession, with each subsequent buckle initiating only after the previous one, located below it, had completed, similar to the observations in [3, 13]. Unloading from maximum compression of ~70 % strain left the top third of the pillar relatively undeformed (Fig. 4b, third panel from left), and the buckle closest to the top always formed last (Fig. 4b, rightmost panel). In contrast, post-compression SEM images of the second set of VACNT pillars, unloaded from the same maximum strain of 70 %, showed that it was the *bottom* third of the pillar that remained relatively undeformed (Fig. 4c, third panel from left). The first instability in these pillars always formed somewhere at their mid-height (Fig. 4c, second panel from left), and the buckle closest to the substrate always formed last (Fig. 4c rightmost panel).

The authors in [18] have suggested that the substantial differences in the deformation responses between the two pillar types shown in Fig. 4 were likely caused by their differing microstructural features and density gradients (the shapes of VACNT pillar cross sections – square versus circular – did not appear to influence their deformation characteristics). In order to capture the density gradients in these two pillar types, the authors used edge detection analysis of systematically obtained SEM images to calculate the average relative number density, ζ, of tubes in each SEM image. Image analysis revealed that the two types of pillars exhibited opposite trends in relative density, which can be used to predict the location of initial buckling instability in a VACNT matrix (Fig. 4b, c). For example, it is reasonable to expect the first folding/buckling event during compressions of VACNTs to originate close to the substrate because of its rigid constraint [21]. Pillars showing a bottom-to-top buckling sequence corroborated this notion (Fig. 4b). These VACNT pillars had highest ζ in the top 10 μm, and their density profiles resembled a step function, where $\zeta \approx 1$ for all locations above the height of ~23 μm and $\zeta \leq 0$ at all locales below. In contrast, the incipient instability in pillars where the bottom buckles formed last occurred somewhere in the middle of pillar height, at the locales with the lowest relative CNT density (Fig. 4c). Maximum ζ in these pillars was located in the bottom ~10 μm, close to the substrate, while the minimum ζ was in the midsection of the pillars. It appears that the lower relative density trumps the constraining effects of the substrate in driving the location of the buckling instability in these pillars. The last buckles in both sets of pillars occurred in the regions with the highest relative CNT densities: near the top in the samples showing bottom-to-top buckling sequence and at the base in the ones where the bottom buckles formed last.

In addition, these two pillar types also show some key differences in their stress-strain response (Fig. 4a). The stress plateau in the pillars showing bottom-to-top buckling had two separate slopes: a relatively flat section up to 30 % strain and a steeper region, with the slope of 0.65 MPa up to unloading at 75 % strain. In contrast the pillar set where the bottom buckles formed last had a nearly flat plateau (slope ~ 0.02 MPa) up to a strain of 75 %. There was also a noticeable difference in the amount of recovery between the two sample types when unloaded from post-densification regime: pillars showing bottom-to-top buckling ($R \approx 57.9 \pm 0.9$ %) showed an almost 45 % higher recovery than those where the bottom buckles formed last ($R \approx 39.8 \pm 3.9$ %).

It has been suggested that such differences in the compressive responses of VACNTs can be predicted based on their local variations in density along the VACNT sample heights. Such property dependence was initially modeled by [3] in a viscoplastic finite element framework which showed a direct correspondence between the presence and extent of an applied property gradient and the amount of hardening within the plateau.

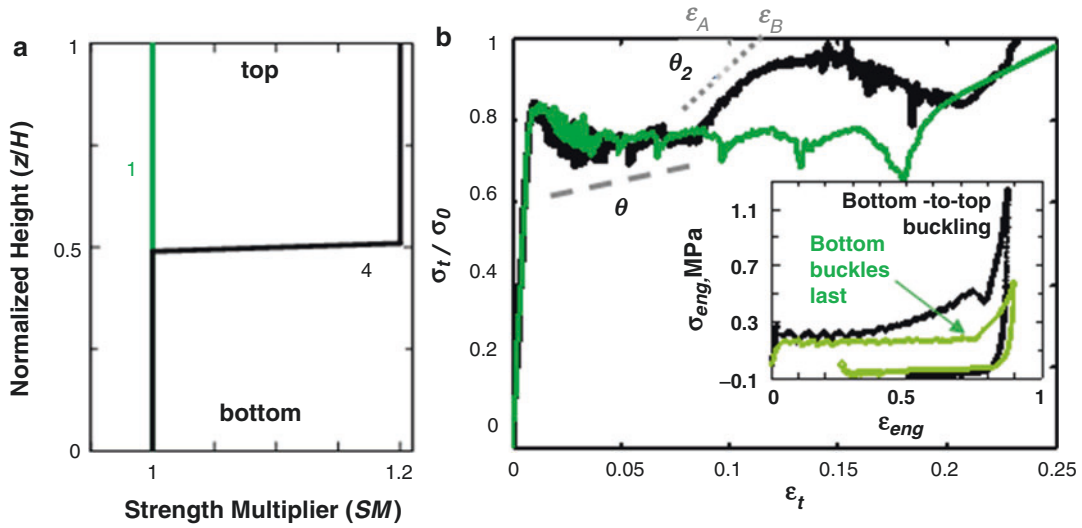
Details of this and other efforts to capture VACNT deformation mechanisms through modeling are summarized in the next section.

Modeling of VACNT Deformation

Only a few preliminary models exist to describe the mechanical deformation of VACNTs. Motivated by both the morphology of the VACNTs (a series of nominally vertical struts) and the observed buckling behavior, many researchers utilize an Euler buckling framework to mechanistically describe their findings. In Euler buckling, an ideal column (perfectly straight, homogeneous, and free from initial stress) is determined to buckle when the applied load reaches some critical value that causes the column to be in a state of unstable equilibrium. Up to the point of buckling, the column is assumed to be perfectly elastic. Scaling calculations by [13] utilize this critical stress to estimate the transition stress value for departure from linear elastic behavior, i.e., the buckling stress, given a reasonable estimate of the tube number density. Similarly, [25] included an additional intertube contact energy to account for favorable van der Waals interactions between tubes in their energetic treatment of VACNTs. Another Euler-based model by [21], motivated by the formation of buckles they observed, proposes piecewise buckling that assumes the top of the pillar deforms via simple shear while the bottom undergoes collapse. Using this framework, they predict that the buckling stress increases asymptotically with decreasing pillar height (for similar aspect ratio pillars). Notably, however, none of these models make predictions concerning the length scale of the buckles formed or their dependence on material parameters such as density, stiffness, tube alignment, or size as the parameter space is difficult to grow in these materials.

Another subset of mechanical analyses utilizes an alternating hardening-softening-hardening local constitutive relation and is the only theories that attempt to capture material response beyond the initial buckling event. The first, a hierarchical bistable spring model, captures the *quantitative* stress-strain response of VACNTs in compression [33]. A subsequent finite element viscoplastic solid with a positive-negative-positive sloped flow strength captures the *qualitative* sequential periodic buckle morphology. The bistable spring model consists of mesoscale elements characterized by elastic-plateau-densification and, most importantly, hysteresis in the unloading curve. These mesoscale elements were shown to be the limiting case of infinitely many bistable spring elements in series. Briefly, a bistable spring consists of two thermodynamically stable elastic loading sections separated by an unstable, negative stiffness region of the stress-strain curve across which the material snaps, similarly to a phase transition. When placed in series, these mesoscale elements capture the hysteretic unloading response seen by [13] very closely. Thus, this model can be utilized to characterize the energy dissipation in VACNTs. In a similar vein, [34] postulate a positive-negative-positive stress-strain relation reminiscent of a bistable spring, but rather than being elastic, it is used as a plastic flow strength function governing the local deformation of an element post-elastic loading. The latter analyses find that this constitutive relation is capable of producing sequential, periodic buckles in an axisymmetric, circular cylindrical mesh with fixed boundary conditions at the base, identical to those in VACNTs. In addition, an axial gradient in the strength is not necessary to initiate bottom-first buckling, as seen in experiments; rather, the fixed boundary conditions are sufficient. However, a reversal of buckle initiation, top to bottom, can be achieved for a sufficient inverse axial gradient (having lower strength at the top of the pillar than the bottom). Both of these analyses capture essential elements of VACNT deformation, but do so for two very different sets of experimental observations: recoverable deformation and plastic deformation.

An example of the approach proposed by [34] is shown in Fig. 5. Here the same model was utilized to predict the stress and strain responses of the two VACNT pillar types shown earlier in Fig. 4, assuming the stiffness of VACNTs to be linearly related to its density, $E \sim \zeta$.

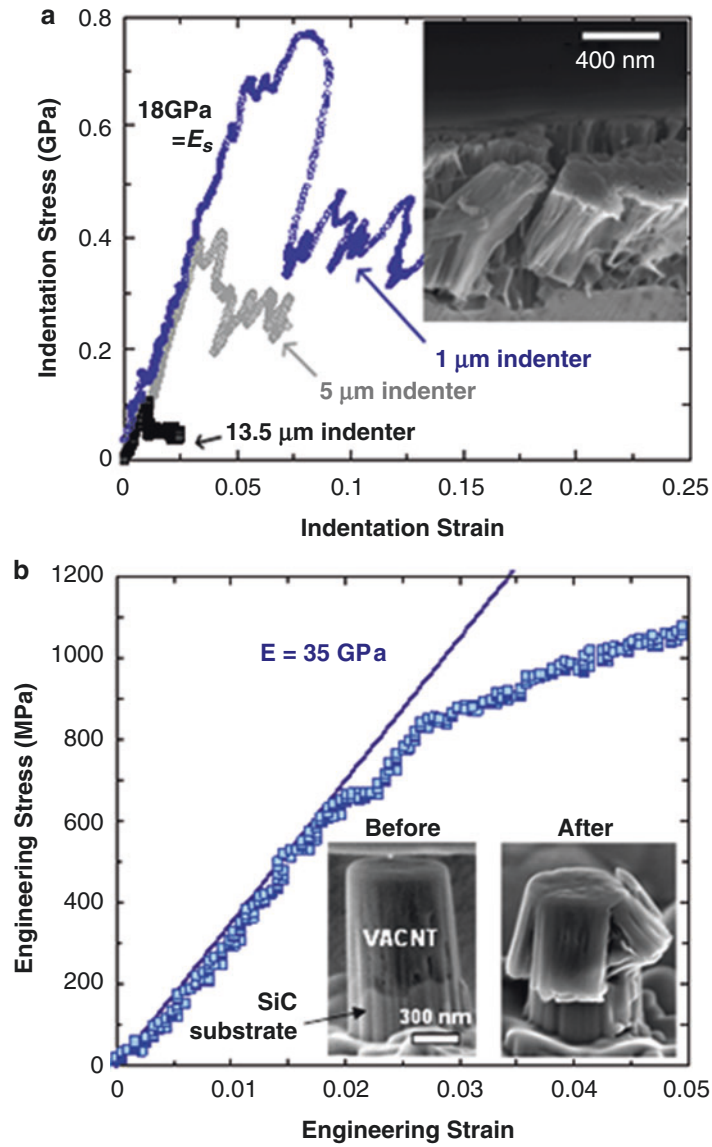


Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Fig. 5 (**a**) Two variations of the *SM* (strength multiplier) function used for simulations: Case 1 (constant *SM* function, *green curve*) versus Case 4 (step variation of the *SM* function over the pillar height, *black curve*). (**b**) Corresponding normalized stress-strain curves for these two cases. Note that the plateau region in Case 4 can be further subdivided by two distinct slopes, θ_1 and θ_2. (**b**, *Inset*) Experimental uniaxial stress-strain responses of the two VACNT pillars shown earlier in Fig. 4 showing the similarity in their responses as compared to the two simulation cases in (**a**) (Adapted with permission from [18] Copyright 2013 American Chemical Society)

The mechanical property gradient was represented by a strength multiplier, *SM*, which is a multiplier giving the spatial variations of E (Young's modulus) and σ_0 where $SM = E/E_{\text{bottom}}$ and similarly for σ_0. Two particular cases are of interest (which matches the experiments most closely): Case 1, *SM* is constant throughout the pillar height, and Case 4, *SM* has a step increase at the midpoint of the pillar height, i.e., at $z/H = 0.5$, where z is the coordinate along the pillar height. Strength gradients over the height in terms of *SM* for the two cases considered are shown in Fig. 5a, and normalized stress-strain data are shown in Fig. 5b. Of special note is the similarity between the experimental and simulated stress versus strain relationships (Fig. 5b inset) in these two cases – such as the flat plateau region for a constant *SM* function (Case 1), similar to the experimental case where the bottom buckles form last, as compared to the two different hardening slopes observed for the step function in *SM* (Case 4), similar to the experimental case of bottom-to-top buckling.

Carbide-Derived Carbons (CDC) VACNTs

While most of the previous sections in this chapter have focused on CVD-VACNTs, VACNTs prepared by the carbide-derived carbon (CDC) synthesis method also deserve special mention. As discussed earlier, CDC-VACNTs refer to a bottom-up growth process, where carbon is formed by selective extraction of the metal or metalloid atoms in the carbide (e.g., silicon carbide) at high temperatures (>1,600 °C), transforming the carbide structure into pure carbon. It has been shown [35] that high-temperature decomposition of SiC by the reactions SiC → Si (g) + C, SiC + 1/2O_2 (g) → SiO (g) + C, leads to the formation of CNTs growing normal to the carbon-terminated $(000\bar{1})$ C-face of hexagonal SiC with primarily zigzag chirality [36] and graphite growth on the Si-terminated (0001) Si-face. These carbide-derived carbon (CDC) nanotube brushes have been shown to have a density close to 0.95 g/cm^3 [37], which is significantly (ten times or more) higher than in catalytic



Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Fig. 6 (**a**) Indentation stress-strain response on 1.3 μm-thick CDC-VACNT brush as a function of indenter size (radius) showing an initial elastic behavior followed by the buckling instability (Reproduced from [8] with permission from Elsevier). (**b**) Uniaxial compression of a CDC-VACNT micro-pillar of diameter ~600 nm. The *inset* images show the VACNT micro-pillar before and after compression (Reproduced from [27] with permission from Elsevier)

CVD growth of any kind of nanotubes. This higher density is generally thought to be due to a conformal transformation of SiC into carbon. These dense CDC CNT brushes consist of small-diameter (1–3 nm outer diameter, one to four walled) non-catalytic CNTs with double-walled CNTs being the most common, as determined from transmission electron microscopy (TEM) and a strong RBM mode in Raman spectra [37]. Assuming an average outer diameter of 3 nm and 0.35 nm as the intertube distance, this would correspond to an aerial density of ~100,000 tubes per μm^{-2} for a hexagonal arrangement of CNTs and ~89,000 tubes per μm^{-2} for a square arrangement. The actual aerial density of the CNTs packed randomly in a dense brush is likely to be somewhere in between.

The higher density of these CDC-VACNT brushes is evident in the SEM image in Fig. 6a inset of the CNT brush grown at 1,700 °C, where no apparent porosity is visible. Such dense VACNTs are ideal for indentation testing using

smaller spherical indenters, something that is not possible for the wider spaced CVD-VACNTs (where the CNTs tend to simply bend away from the indenter). The use of spherical indenters is also advantageous in that, by using appropriate data analysis protocols [27, 38], the extracted load-displacement data from these tests can be transformed into meaningful indentation stress-strain curves. The indentation stress-strain curves in turn enable one to follow the entire evolution of the mechanical response in the VACNT array, from initial elasticity to the initiation of buckling to post-buckling behavior at finite plastic strains. In addition, the indentation stress-strain curves also allow one to estimate the Young's modulus and the stress at buckling in the indentation experiment. For example, using a 1 μm spherical indenter, the Young's modulus of 200 nm-thick CDC-VACNT brushes was estimated to be ~17 GPa, and the critical buckling stress was estimated as ~0.3 GPa at a load of 0.02 mN [8].

The authors in [8] also used indentation stress-strain curves from different indenter radii to explore the effects of indentation zone sizes and the material defect density on the VACNT buckling stress. This is shown in Fig. 6a where the indentation response on a thicker VACNT brush (thickness 1.2–1.4 μm; see Fig. 6a inset) is shown as a function of three different indenter radii. In general, all three indenter sizes show a similar trend in the VACNT response: an initial elastic behavior (Young's modulus ~18 GPa), followed by a sharp drop at a critical stress corresponding to CNT buckling. Note however that the values of buckling stresses vary significantly between the three different indenters, where indentation with the smaller 1 μm indenter shows the highest buckling stress, followed by the 5 μm indenter, while buckling with the largest 13.5 μm indenter occurs at a significantly lower indentation stress.

Such differences in buckling stresses can be understood by a closer analysis of the indentation zone sizes between the three indenter radii used. The indentation zone sizes at buckling for the larger 13.5 μm indenter (~3.58 μm) extend well beyond the VACNT brush thickness (1.2–1.4 μm), while only a limited thickness of the VACNT brush (~0.39 μm) is exposed to indentation stress for the smaller 1 μm indenter. The corresponding differences in buckling stress are thus largely a consequence of the effective buckling length available in each case; with smaller indenters, there is only a smaller buckling length and the material is able to withstand higher buckling stresses. Interestingly the authors in [8] also report a larger variation in the values of the buckling stress for the small 1 μm radii indenter as compared to the larger indenters. These values seem to indicate that defect density varies from one location to another in one VACNT brush. Thus the smaller volume sampled by the 1 μm indenter is more sensitive to the variation in defect density when different regions of the sample are probed and causes a larger spread of the buckling stress values when using this indenter. The larger indenters, on the other hand, encounter a bigger volume, and hence they mostly register a combined VACNT-defect response leading to a lower buckling stress and a smaller spread in the buckling stress.

The above results have been validated by conducting compression tests on VACNT micro-pillars machined using a focused ion beam (FIB) on the same sample. This is shown in Fig. 6b which shows the uniaxial compression data for a ~600 nm-diameter VACNT pillar. As seen from this figure, the values measured from indentation and compression tests are complementary to one another. The increase in the Young's modulus values in the micro-pillar compression tests (~30 GPa as compared to ~17–18 GPa in indentation) is due to ion beam irradiation during the micromachining process using FIB. This modifies the structure of the outer rim of CNTs in the VACNT pillar, which potentially results in intertube bridging between the densely packed CNTs and hence an increase in their Young's modulus [39].

These numbers suggest that these CDC-VACNT brushes perform significantly better and exhibit considerably higher mechanical properties compared to CVD-VACNT turfs; CDC-VACNTs have values of Young's modulus one to two orders of magnitude higher [25] and buckling strengths several orders of magnitude higher [1, 2] compared to a CVD-VACNT turf – a difference which is explained by the much higher density of the

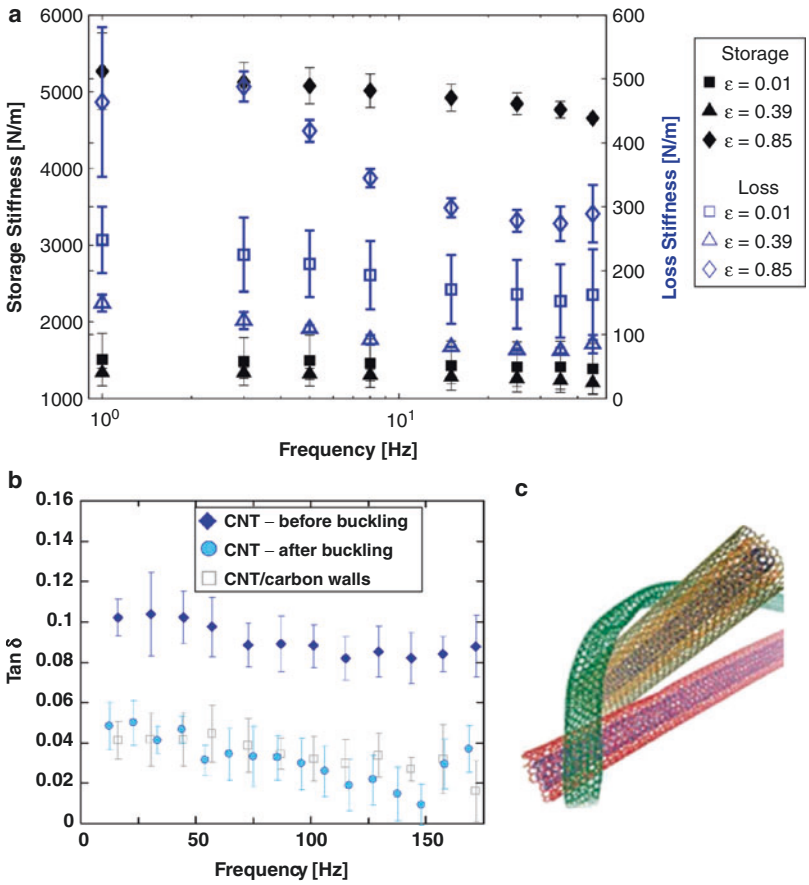
tubes per unit area in the CDC-VACNT resulting in considerably higher mechanical properties. These higher properties are of extreme importance for making selective CNT membranes for gas or liquid filtration/separation or CNT coatings for tribological applications.

Viscoelasticity

In addition to their distinctive buckling behavior and their ability to recover from large deformations, VACNTs have also been reported to demonstrate another case of extreme mechanical performance – a unique viscoelastic response that spans a truly wide temperature range from −196 to 1,000 °C – something no other material has shown so far. Viscoelastic materials exhibit both viscous and elastic characteristics when subjected to load. Thus, a viscoelastic material is able to both dissipate energy through viscous behavior (as in honey), while storing energy through elasticity (as in rubber band). The stress-strain response of viscoelastic materials is typified by hysteresis in the loading-unloading cycle.

Viscoelastic behavior in a material is generally characterized in terms of its loss (E'') or storage modulus (E') (or by the ratio, known as $\tan\delta = E''/E'$, of these two moduli). E'' relates to the amount of energy dissipated, while E' represents the stored energy in the material. The angle δ is the phase lag between the oscillatory load and displacement responses under a sine wave load. In a typical experiment, the material is loaded to a predetermined strain and the mechanical probe is oscillated across a range of frequencies. By measuring the resultant load amplitude, displacement amplitude, and the phase lag during the test, the values of loss modulus, storage modulus, and tan δ are determined. Similarly, viscoelasticity can also be quantified by the memory or hysteresis effects during load-unload cycles under deformation, where the energy dissipated is given by the area of the hysteresis loop. These viscoelastic effects have recently been documented in the highly intertwined random networks of VACNTs. Viscoelasticity in CVD-VACNTs has been demonstrated by [3], where both storage and loss stiffness were studied in a frequency range from 1 to 45 Hz at different strain levels (Fig. 7a). Note that these stiffness values are proportional to moduli given a known Poisson's ratio which is lacking for VACNTs. The elastic response is clearly frequency independent, indicating that the VACNT's elastic deformation is likely due to the same mechanism (likely tube bending) over the range of timescales tested. As shown in Fig. 7a, more energy is dissipated (higher values of loss stiffness) at larger strain levels when a higher fraction of the VACNT pillar has buckled. Interestingly, for the more dense CDC-VACNTs, the opposite seems to be true (Fig. 7b). For the highly dense CDC-VACNTs, [8] have shown a significant drop in the tan δ values of the CNTs after buckling, i.e., when the material was highly compacted. These observations indicate that while an increase in density can significantly increase the viscoelastic behavior of CNTs, there appears to be a cutoff beyond which the contacting CNTs became increasingly bundled resulting in a decrease in their ability to dissipate energy.

A recent report by [26] also suggests a rubber-like viscoelastic behavior in a random network of long, interconnected, and tangled CNTs, similar in many ways to VACNTs but significantly less aligned, over a very wide temperature range – from −196 to 1,000 °C – with a possibility of extending this behavior beyond 1,500 °C. While the oxidizing nature of carbon may limit the application of these materials to only vacuum or protective (reducing) environments, CNTs are nevertheless the only known solids to demonstrate such behavior at extremely low or very high temperatures. The authors attribute the reversible dissipation of energy in CNTs over this remarkable range to the zipping and unzipping of the CNTs upon contact (see Fig. 7c) caused by the van der Waals interactions. In this instance, "reversible" means that beyond a critical strain, the zipping/unzipping process was no longer possible everywhere as more and more tubes become permanently entangled at higher strain. This would lead to a loss in their viscoelasticity, similar to the observations in the highly dense CDC-VACNTs (Fig. 7b).



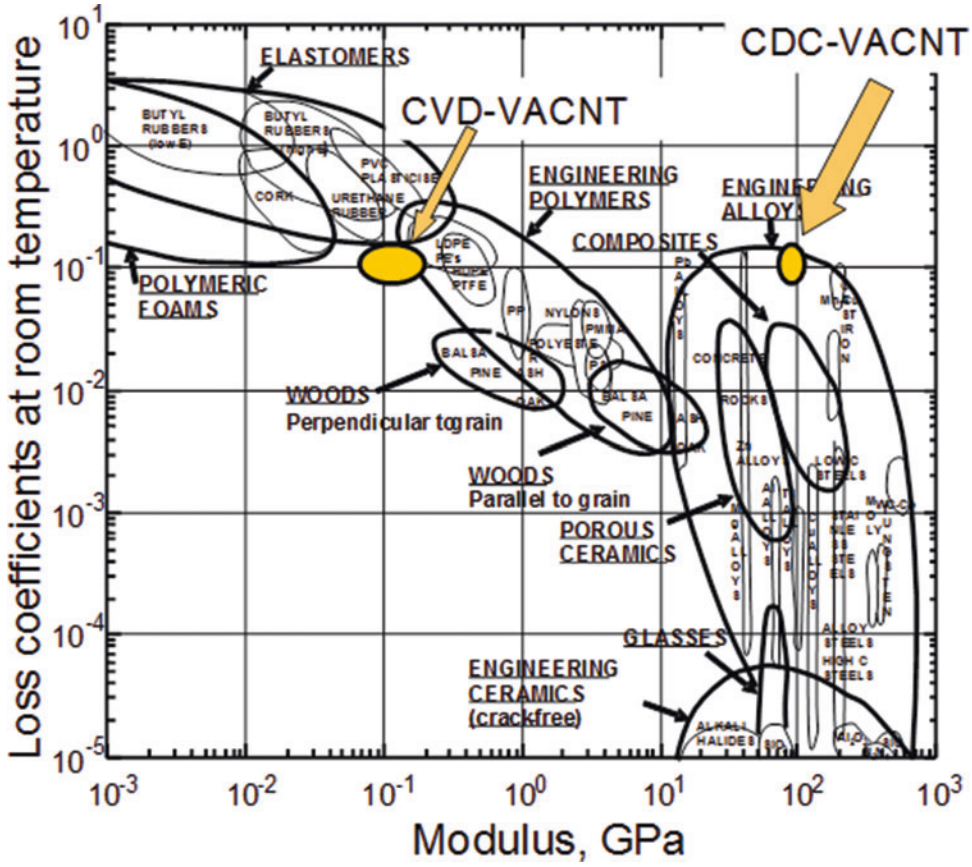
Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior,

Fig. 7 Viscoelastic behavior as seen in (**a**) CVD-VACNTs showing an increase in the loss stiffness with increasing strains [3] (Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission). (**b**) CDC-VACNTs showing a significant decrease in values of tan δ after buckling (Reproduced from [8] with permission from Elsevier). (**c**) Atomistic modeling image of a possible entangled arrangement of single-, double-, and triple-walled CNTs leading to their rubberlike viscoelastic behavior (Reprinted from [41] with permission from AAAS)

The unique combination of superior mechanical properties and the ability to dissipate energy during deformation are expected to have significant implications in damping applications utilizing VACNTs. These, combined with the wide temperature range of its viscoelastic behavior, make VACNT-based materials a promising choice for use in mechanical applications under extreme temperatures or temperature gradients. Possible environments range from the cold of interstellar space to down-to-earth viscoelastic applications, such as MEMS devices to high-temperature vacuum furnaces.

Applications

Multiwall carbon nanotubes are well known for a variety of exceptional properties. These include a high tensile modulus, on the order of 1 TPa; high strength, on the order of tens of GPa; high thermal conductivity; recoverability after large bending angle deformation; and a range of electrical properties depending on the chirality of the graphene walls. For these reasons, they are currently the subject of a wide range of research. In the form of VACNTs, these individual tube properties may be significantly altered by the collective interaction of thousands of tubes. Several applications of VACNT structures currently under study rely on the mechanical behavior described earlier in this chapter. A remarkable but incomplete list includes components of highly compliant thermal contacts for microelectromechanical systems (MEMS) and microelectronics [6, 21], dry adhesives, thermally robust energy-dissipating rubber [26, 40], and energy absorption or impact mitigation [41]. Other applications, such as optical coatings and cold cathode arrays, may rely less on the



Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior, Fig. 8 Ashby plot of loss coefficient versus modulus showing the performance of VACNTs

mechanical properties for optimization and design, but understanding is still necessary for evaluation of robustness and in-use lifetime analysis.

We go through each application briefly to highlight the VACNT-specific mechanical behavior of interest. First, thermal contacts for delicate electronic devices have two requirements. They must be highly thermally conductive and make conformal contact (to increase heat transfer) while avoiding damage to the components they are transferring heat to and from. CNT foams are ideal for this application due to the high compliance that comes from the reduced load capacity of the CNT struts as they buckle. Dry adhesive applications of VACNTs take advantage of the hierarchical structure which can, as desired in MEMS switches, make conformal contact to surfaces at a variety of roughness length scales, thereby increasing the attractive interactions between the tubes and the surface. Third, energy-dissipating and energy-absorbing applications require more in-depth knowledge of the deformation mechanism. As mentioned in the previous section, energy dissipation is currently thought to be due to tube zipping/unzipping or tube-to-tube sliding and rearrangement [8, 14, 25, 26, 40], while energy absorption is certainly maximized during buckle formation. Future applications may be able to take advantage of yet undiscovered traits of the incompletely characterized VACNTs.

Summary and Outlook

Although the mechanisms governing VACNTs collective mechanical behavior are still largely uncharacterized, the special properties of multiwall carbon nanotubes combined with the complex interactions that arise between them in the hierarchical VACNT microstructure have generated significant research interest. Their wide range of properties, mechanical, electrical, and thermal, make them ideal candidates as multifunctional materials, particularly for applications in which soft materials (such as polymers) have traditionally dominated. Notably, VACNTs occupy a unique niche among engineering materials as shown in the Ashby property chart in Fig. 8 with the ability to span a wide range of properties given a tunable microstructure. There remain two major hurdles to rationally designing VACNTs for any application: control of the CNT growth process and understanding of the relationship between the microstructure and mechanical response. These hurdles are interrelated through the fact that the design space cannot be systematically probed until the microstructure can be systematically controlled. Mastery of these unknowns will not only further VACNT's place as novel materials in the current applications under study, but may reveal previously undiscovered behaviors in microstructures yet to be created or characterized.

Acknowledgments This work was performed, in part, at the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy (DOE) Office of Science. Los Alamos National Laboratory, an affirmative action equal opportunity employer, is operated by Los Alamos National Security, LLC, for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. SP gratefully acknowledges funding from the Los Alamos National Laboratory Director's Postdoctoral Fellowship during the writing of this chapter.

References

1. Waters, J.F., Guduru, P.R., Jouzi, M., Xu, J.M., Hanlon, T., Suresh, S.: Shell buckling of individual multiwalled carbon nanotubes using nanoindentation. Appl. Phys. Lett. **87**, 103109-1-3 (2005)
2. Waters, J.F., Riester, L., Jouzi, M., Guduru, P.R., Xu, J.M.: Buckling instabilities in multiwalled carbon nanotubes under uniaxial compression. Appl. Phys. Lett. **85**(10), 1787–1789 (2004)
3. Hutchens, S.B., Hall, L.J., Greer, J.R.: In situ mechanical testing reveals periodic buckle nucleation and propagation in carbon nanotube bundles. Adv. Funct. Mater. **20**(14), 2338–2346 (2010)
4. Pathak, S., Mohan, N., Pour Shahid Saeed Abadi, P., Graham, S., Cola, B.A., Greer, J.R.: Compressive response of vertically aligned carbon nanotube films gleaned from *in-situ* flat punch indentations. J. Mater. Res.: Focus. Issue De Novo Carbon Nanomater. **28**(7), 984–997 (2013). DOI: 10.1557/jmr.2012.366(2013) pp 984–997
5. Presser, V., Heon, M., Gogotsi, Y.: Carbide-derived carbons – from porous networks to nanotubes and graphene. Adv. Funct. Mater. **21**(5), 810–833 (2011)
6. McCarter, C.M., Richards, R.F., Mesarovic, S.D., Richards, C.D., Bahr, D.F., McClain, D., Jiao, J.: Mechanical compliance of photolithographically defined vertically aligned carbon nanotube turf. J. Mater. Sci. **41**(23), 7872–7878 (2006)
7. Yaglioglu, O., Cao, A., Hart, A.J., Martens, R., Slocum, A.H.: Wide range control of microstructure and mechanical properties of carbon nanotube forests: a comparison between fixed and floating catalyst CVD techniques. Adv. Funct. Mater. **22**(23), 5028–5037 (2012)
8. Pathak, S., Cambaz, Z.G., Kalidindi, S.R., Swadener, J.G., Gogotsi, Y.: Viscoelasticity and high buckling stress of dense carbon nanotube brushes. Carbon **47**(8), 1969–1976 (2009)
9. Qi, H.J., Teo, K.B.K., Lau, K.K.S., Boyce, M.C., Milne, W.I., Robertson, J., Gleason, K.K.: Determination of mechanical properties of carbon nanotubes and vertically aligned carbon nanotube forests using nanoindentation. J. Mechanics Phys. Solids **51**, 2213–2237 (2003)
10. Bedewy, M., Meshot, E.R., Guo, H.C., Verploegen, E.A., Lu, W., Hart, A.J.: Collective mechanism for the evolution and self-termination of vertically aligned carbon nanotube growth. J. Phys. Chem. C **113**(48), 20576–20582 (2009)
11. Brieland-Shoultz, A., Tawfick, S., Park, S.J., Bedewy, M., Maschmann, M.R., Baur, J.W., Hart, A.J.: Scaling the stiffness, strength, and toughness of ceramic-coated nanotube foams into the structural regime. Adv. Funct. Mater. **24**(36), 5728–5735 (2014)
12. Poelma, R.H., Morana, B., Vollebregt, S., Schlangen, E., van Zeijl, H.W., Fan, X., Zhang, G.Q.: Tailoring the mechanical properties of high-aspect-ratio carbon nanotube arrays using amorphous silicon carbide coatings. Adv. Funct. Mater. **24**(36), 5737–5744 (2014)
13. Cao, A.Y., Dickrell, P.L., Sawyer, W.G., Ghasemi-Nejhad, M.N., Ajayan, P.M.: Super-compressible foamlike carbon nanotube films. Science **310**(5752), 1307–1310 (2005)

14. Suhr, J., Victor, P., Sreekala, L.C.S., Zhang, X., Nalamasu, O., Ajayan, P.M.: Fatigue resistance of aligned carbon nanotube arrays under cyclic compression. Nat. Nanotechnol. **2**(7), 417–421 (2007)
15. Tong, T., Zhao, Y., Delzeit, L., Kashani, A., Meyyappan, M., Majumdar, A.: Height independent compressive modulus of vertically aligned carbon nanotube arrays. Nano Lett. **8**(2), 511–515 (2008)
16. Deck, C.P., Flowers, J., McKee, G.S.B., and Vecchio, K.: Mechanical behavior of ultralong multiwalled carbon nanotube mats. J. Appl. Phys. 101(2), 23512-1-9 (2007)
17. Pathak, S., Lim, E.J., Pour Shahid Saeed Abadi, P., Graham, S., Cola, B.A., Greer, J.R.: Higher recovery and better energy dissipation at faster strain rates in carbon nanotube bundles: an in-situ study. ACS Nano **6**(3), 2189–2197 (2012)
18. Pathak, S., Mohan, N., Decolvenaere, E., Needleman, A., Bedewy, M., Hart, A.J., Greer, J.R.: Local relative density modulates failure and strength in vertically aligned carbon nanotubes. ACS Nano **7**(10), 8593–8604 (2013)
19. Bradford, P.D., Wang, X., Zhao, H., Zhu, Y.T.: Tuning the compressive mechanical properties of carbon nanotube foam. Carbon (Article in Press 2011)
20. Ci, L., Suhr, J., Pushparaj, V., Zhang, X., Ajayan, P.M.: Continuous carbon nanotube reinforced composites. Nano Lett. **8**(9), 2762–2766 (2008)
21. Zbib, A.A., Mesarovic, S.D., Lilleodden, E.T., McClain, D., Jiao, J., Bahr, D.F.: The coordinated buckling of carbon nanotube turfs under uniform compression. Nanotechnology **19**(17), 175704-1-7 (2008)
22. Pathak, S., Raney, J.R., Daraio, C.: Effect of morphology on the strain recovery of vertically aligned carbon nanotube arrays: an in situ study. Carbon **63**, 303–316 (2013)
23. Zhang, Q., Lu, Y.C., Du, F., Dai, L., Baur, J., Foster, D.C.: Viscoelastic creep of vertically aligned carbon nanotubes. J. Phys. D-Appl. Phys. 43(31) (2010)
24. Qiu, A., Bahr, D.F., Zbib, A.A., Bellou, A., Mesarovic, S.D., McClain, D., Hudson, W., Jiao, J., Kiener, D., Cordill, M.J.: Local and non-local behavior and coordinated buckling of CNT turfs. Carbon **49**(4), 1430–1438 (2011)
25. Mesarovic, S.D., McCarter, C.M., Bahr, D.F., Radhakrishnan, H., Richards, R.F., Richards, C.D., McClain, D., Jiao, J.: Mechanical behavior of a carbon nanotube turf. Scr. Mater. **56**(2), 157–160 (2007)
26. Xu, M., Futaba, D.N., Yamada, T., Yumura, M., Hata, K.: Carbon nanotubes with temperature-invariant viscoelasticity from-196 degrees to 1000 degrees C. Science **330**(6009), 1364–1368 (2010)
27. Pathak, S., Kalidindi, S.R.: Spherical nanoindentation stress–strain curves. J. Mater. Sci. Eng.: R: Rep. **91**, 1–36 (2015)
28. Zhang, Q., Lu, Y.C., Du, F., Dai, L., Baur, J., Foster, D.C.: Viscoelastic creep of vertically aligned carbon nanotubes. J. Phys D-Appl Phys 43(31), 315401-1-7 (2010)
29. Cao, C., Reiner, A., Chung, C., Chang, S.-H., Kao, I., Kukta, R.V., Korach, C.S.: Buckling initiation and displacement dependence in compression of vertically aligned carbon nanotube arrays. Carbon **49**(10), 3190–3199 (2011)
30. Yaglioglu, O.: Carbon nanotube based electromechanical probes – PhD Thesis, in Dept. of Mechanical Engineering. Massachusetts Institute of Technology (2007)
31. Gibson, L.J.: Mechanical behavior of metallic foams. Annu. Rev. Mater. Sci. **30**(Compendex), 191–227 (2000)
32. Gibson, L.J., Ashby, M.F.: Cellular Solids: Structure and Properties. Cambridge University Press, Cambridge, UK (1999)
33. Fraternali, F., Blesgen, T., Amendola, A., Daraio, C.: Multiscale mass-spring models of carbon nanotube foams. J. Mechanics Phys. Solids **59**(1), 89–102 (2011)
34. Hutchens, S.B., Needleman, A., Greer, J.R.: Analysis of uniaxial compression of vertically aligned carbon nanotubes. J. Mechanics Phys. Solids **59**(10), 2227–2237 (2011)
35. Kusunoki, M., Rokkaku, M., Suzuki, T.: Epitaxial carbon nanotube film self-organized by sublimation decomposition of silicon carbide. Appl. Phys. Lett. **71**(18), 2620–2622 (1997)
36. Kusunoki, M., Suzuki, T., Honjo, C., Hirayama, T., Shibata, N.: Selective synthesis of zigzag-type aligned carbon nanotubes on SiC (000-1) wafers. Chem. Phys. Lett. **366**(5–6), 458–462 (2002)
37. Cambaz, Z.G., Yushin, G., Osswald, S., Mochalin, V., Gogotsi, Y.: Noncatalytic synthesis of carbon nanotubes graphene and graphite on SiC. Carbon **46**(6), 841–849 (2008)
38. Kalidindi, S.R., Pathak, S.: Determination of the effective zero-point and the extraction of spherical nanoindentation stress–strain curves. Acta Mater. **56**(14), 3523–3532 (2008)
39. Kis, A., Csanyi, G., Salvetat, J.P., Lee, T.-N., Couteau, E., Kulik, A.J., Benoit, W., Brugger, J., Forro, L.: Reinforcement of single-walled carbon nanotube bundles by intertube bridging. Nat. Mater. **3**(3), 153–157 (2004)
40. Gogotsi, Y.: High-temperature rubber made from carbon nanotubes. Science **330**(6009), 1332–1333 (2010)
41. Misra, A.A.G., Julia, R., Daraio, C.: Strain rate effects in the mechanical response of polymer-anchored carbon nanotube foams. Adv. Mater. **20**, 1–5 (2008)

Vertically-Aligned Carbon Nanotubes

► Chemical Vapor Deposition (CVD)

Vibration Assisted Machining

► Ultrasonic Machining

Viscosity

► Interfacial Investigation of Protein Films Using Acoustic Waves

Visual Prosthesis

► Artificial Retina: Focus on Clinical and Fabrication Considerations

Visual Servoing for SEM

► Robot-Based Automation on the Nanoscale

W

Wavefront Deformation Particle Image Tracking

► Astigmatic Microparticle Imaging

Waypoint Detection

Richard P. Mann
Centre for Interdisciplinary Mathematics, Uppsala University, Uppsala, Sweden

Synonyms

Landmark detection; Key point detection

Definition

Animals are frequently required to navigate accurately from one location to another. The means an animal uses to perform a navigational task depends on the available environmental cues, their specific sensory systems, and particularly whether they are in a familiar or unfamiliar area. A *familiar area* can be defined as one of which the animal has extensive experience, leading to detailed knowledge of the environment and available navigational cues.

Waypoints are geostationary locations used by an animal to navigate in a familiar area. Specifically they are small-scale regions of space that the animal returns to as part of the navigational task. A waypoint may be associated with a memorized *landmark* which allows the animal to identify the waypoint from a distance and direct itself toward it. Navigation toward an eventual goal may involve visiting an ordered succession of waypoints finishing at the final location – this mechanism is generally referred to as *pilotage*. Alternatively waypoints may be a by-product of navigational constraints imposed by the environment; if, for example, the goal can be reached only by passing through a narrow passage, that passage will act as a waypoint.

While waypoints may be associated with landmarks – recognizable sensory features such as prominent visual cues or memorable odors – the task of waypoint detection is not in general synonymous with landmark detection. Landmarks may be identified and used by the animal at a distance, without necessarily visiting the landmark's true location. Conversely, as noted above, a waypoint may exist independently of any memorable feature. Instead, waypoints are more accurately defined as importance regions of space and segments of movement paths, whose salience may be the result of a variety of cues and constraints, which cause discernable effects upon the animal's movement.

Waypoint detection is the task of identifying waypoints from observations of the animal's movements. Since the definition of a waypoint requires the animal to return to the specific area,

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

this therefore constitutes a classification task, identifying the segments of the animal's movement trajectory, or *path*, that correspond to waypoint locations. This entry gives an overview introduction to the purpose and methods of waypoint detection, examining three methods for isolating particularly salient elements of movement paths.

Purpose

The principle purpose of detecting waypoints is to understand what drives and facilitates navigation. It is a way of identifying important areas which may contain salient information the animal uses to recall a memorized path or important resources the animal must utilize. Examples of this might include visually striking features, high food concentrations, or unseen constraints on the animal's movement (strong wind patterns acting on flying creatures, for instance). Waypoint detection can be useful for biomimicry, allowing both an understanding of how animals use key features to navigate and replication of this process in artificial navigating agents such as robots.

In the literature, waypoint detection has been used as the basis for understanding which features animals attend to in the landscape [6, 7, 10], determining the extent of an animal's knowledge of the area [5] and determining where animals forage [1, 4]. The technique of isolating important regions of the path for further analysis is a basis for linking animal movement behaviors and the surrounding environment, such as the landscape, sensory cues, and other animals, by determining what the environmental conditions are in the spatial and temporal vicinity of the detected waypoints.

Three Methods for Detecting Waypoints

Specific identification of waypoints from recorded animal movements is a novel task, and there are consequently few formalized examples in the scientific literature that deal explicitly with this problem. However, the likely existence of waypoints in specified regions of space is implicit in several analyses of movement paths. The three principle methods for identifying these spatial regions can be broadly classified as:

- Localized variability *between* multiple movement paths – *Path similarity*
- Changes in speed and direction *within* a single path – *Path complexity*
- Information content of a path segment – *Path predictability*

This entry describes these methods largely in the context of experiments on homing pigeons, where the use of global positioning satellite (GPS) devices has created a huge amount of data calling for novel analysis. This has led to the development of methods for path analysis, which are generalizable and applicable to movement paths from other species.

Path Similarity

Path similarity presents an intuitive method for waypoint detection. The principle property of a waypoint is that it is a *restricted* region of space which the animal will pass through on its way to the objective. A waypoint therefore is likely to present itself as a relatively small region of space to which the animal repeatedly returns. In this case it is likely, though not guaranteed, that the movement paths of the animal will be significantly more similar in the vicinity of the waypoint, since their freedom to vary is constrained, than outside these regions where the animal's movements may vary more widely. Increasing path similarity over time has been linked to the emergence of detailed visual memories of the landscape and the use of pilotage as a navigational strategy [2, 3, 12].

To use path similarity objectively requires an algorithmic measure of the variability between different paths. To identify specific regions of space demands a *localized* measure – one that varies through space – rather than a *global* measure which relates the overall similarity of the paths. Here several options present themselves. None can be considered the sole correct way to

measure variability, nor are these options exhaustive. Different measures of variability are necessarily highly correlated but the specific choice depends on the exact nature of the analysis. The basic principle is to measure the spatial separation between paths within some localized region. For this purpose it is helpful to define the concept of the *nearest-neighbor distance* (NND). This is the distance from a point on one path to the closest point on another. Using this distance allows the path separation to be determined in a purely spatial manner. The alternative approach of measuring the distance between paths *at the same point in time* measures a spatiotemporal separation. Ordinarily it is purely spatial separation that best describes the expected effect of waypoints.

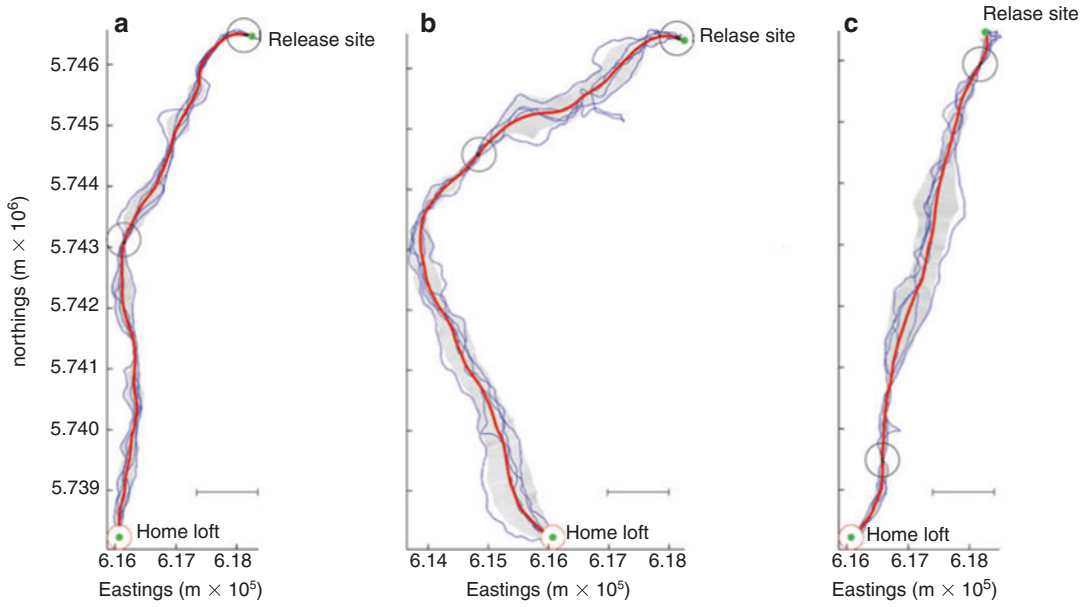
A simple means of calculating path separation is to measure the NND from every point on a path to every other path. The value of the separation can then be taken as the mean, median, minimum or maximum of these values as desired, giving a value for separation at every position along the path. This provides a measure of separation along each path, but not a single description of the overall separation of all the paths as a single variable, which may make it difficult to identify single points where variation is low. This may be ameliorated by the fact that areas where the separation is low will also be areas where the paths are very close, so regions of low variability will be spatially restricted.

It is also possible to measure a single value of separation, using variation around a constructed *mean path*. A mean path is a single path that is representative of the full set of paths. As with path separation there is no absolute way to construct such a path. In analogy with the mean of a set of scalar variables, a mean path can be constructed to minimize a particular measure of separation relative to the original paths, much as the scalar mean minimizes the square distance between itself and the original variables. This approach is used, for example, by Freeman et al. [5], where a mean path is generated which minimizes the total nearest-neighbor separation from itself to the true paths. Constructing such a path may require an iterated approximation method rather than a direct calculation. See Fig. 1 for examples of constructed mean paths and mean path similarity in relation to the original paths. Similarly to the case where path similarity is calculated along each route, the mean path will be most representative of the true paths in regions of peak similarity, which correspond to waypoint locations. Therefore, having calculated similarity along the mean path, one can be sure that the waypoints identified from this similarity will be in approximately the correct spatial location.

A related approach is that used by Lipp et al. [8], for investigating the propensity of pigeons to follow roads and other linear features. Here the spatial area containing the paths is segmented into a two-dimensional grid. The number of paths passing through each grid element is recorded and compared with a null hypothesis for the distribution of paths (see [8] for an example of generating an appropriate null hypothesis). In the original paper, the analysis demonstrates that grid elements containing roads also contain significantly more paths than expected. The converse argument can also be applied – grid elements containing more paths than expected (judged by some significance test) are likely to contain waypoints. The disadvantage of this method is the difficulty in striking a balance between the size of the grid and the density of paths. If the grid elements are too small there will be few paths in any single element and it will be impossible to distinguish from the null hypothesis. If the grid elements are too large, the resolution will be limited to the size of the element, making it difficult to pinpoint the location of the waypoint accurately.

Path Complexity

Less intuitive than path variability, but applicable even when only a single path is observed, path complexity segments the movement path into regions of high and low localized complexity. Increased complexity can be an indicator of waypoint location since waypoints may be associated with decision points, where the animal must first adjust its heading to reach the waypoint and subsequently readjust again to direct itself toward the

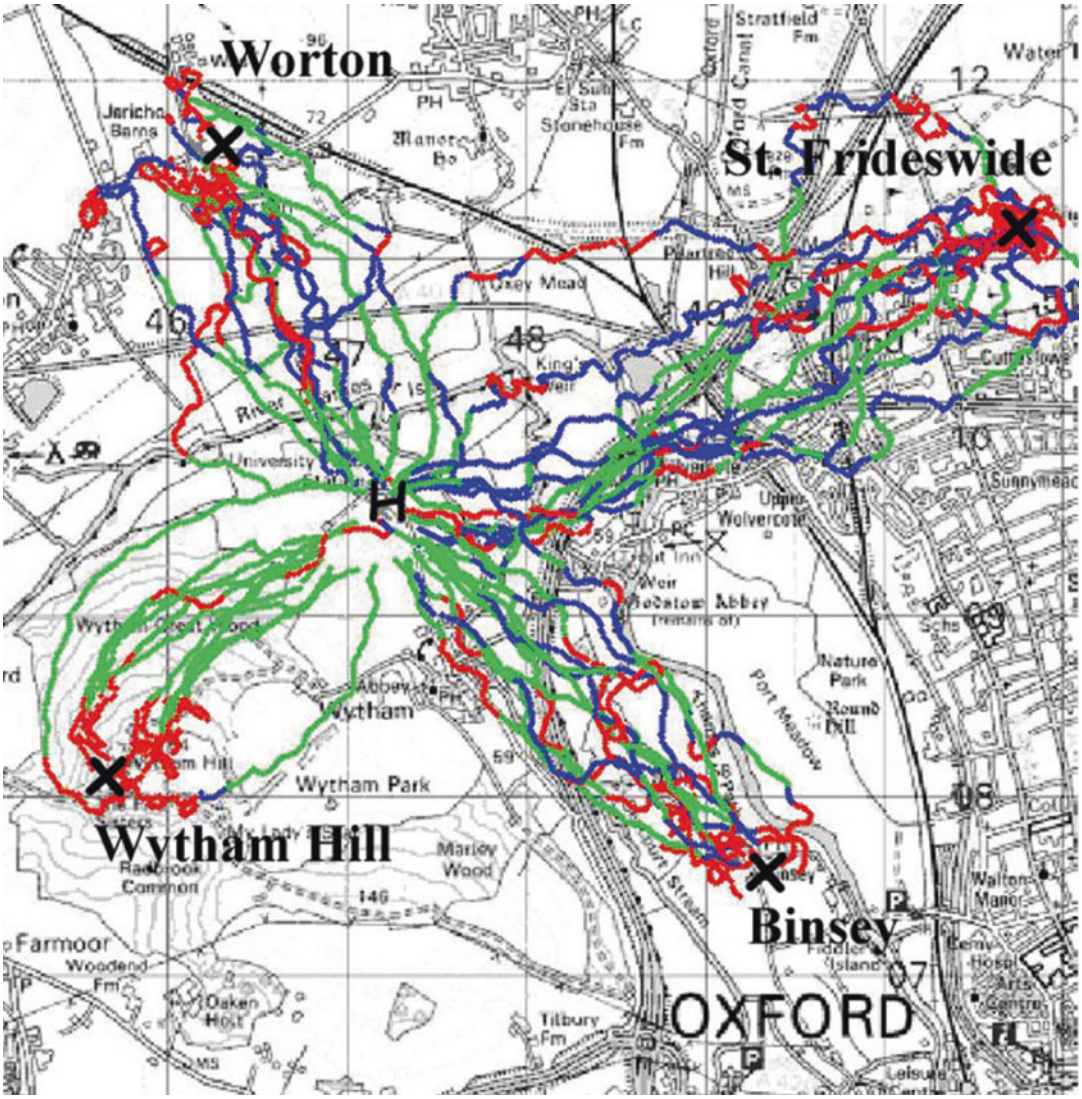


Waypoint Detection, Fig. 1 Five flights each from three pigeons (*light blue*) showing the calculated mean path (*bold red*), variance around the mean path (*grey band*), and location of areas of peak route similarity (*black circles*). The peak similarity regions correspond to variance below a threshold value. Each peak similarity region can contain multiple successive points, but for clarity only one circle is shown per region (Reprinted from Ref. [5], Copyright [2010], Creative Commons open access license)

next objective. Conversely when the animal is far from any known locations, it would be expected to follow a smooth path since there is no external cue available to correct the heading and any other variation would lead to unnecessary energy loss.

The complexity of a path segment can be viewed as a measure of how much information is required to describe it. A straight line or a very smooth curve can be described simply by stating the end points and the degree of curvature. A segment whose curvature is constantly changing, which may change direction discontinuously or which returns back to a previously covered point (looping) will require far more information to adequately describe. To understand the relationship between complexity and information, imagine having to describe a path segment to someone who is unable to see it, well enough that they could draw it themselves. The more details that are required to convey that path segment, the more complex it is.

As with path similarity, path complexity can be measured in a variety of ways, which are similarly highly correlated with each other. These range in abstractness from *tortuosity*, the local degree of inefficiency in the path, to the highly abstract concept of spatiotemporal positional entropy introduced by Roberts et al. [13]. Figure 2 shows a plot of a number of recorded movement paths, color-coded by the local degree of complexity, to demonstrate how high complexity and low complexity path segments present themselves visually. Each path is segmented into low-, medium-, and high-complexity regions, using positional entropy as a measure of complexity. In the figure, it can be clearly seen that low complexity (green) regions are smooth, with low curvature and few if any sharp changes of direction. High-complexity (red) regions often include loops, sharp turns, and many changes of direction. The medium-complexity (blue) regions lie between these extremes. The figure also shows that high-complexity regions are often short, indicating



Waypoint Detection, Fig. 2 Identified regions of low (*green*), medium (*blue*), and high (*red*) complexity in a variety of homing pigeon flight paths, as determined by positional entropy (Reprinted from Ref. [6], Copyright [2004], with permission from Elsevier)

some local effect on the path, and occur particularly around the release point, when the pigeons are deciding on a heading to take.

Another similar measure that has been used to detect foraging within movement paths is *passage time*. This is amount of time taken by an animal from the moment when it enters a (typically circular) region of space to the moment when it leaves. This can be measured as the first passage time [4, 11] – the time from first entering to first leaving the region or total passage time [1] – the time from first entering to the final time leaving. This is strongly related to path complexity, since highly complex, inefficient path segments with lots of changes of direction will have high passage times. The use of passage time to detect foraging within directed movement is broadly analogous to waypoint detection, since foraging stops within a directed path will form waypoints by virtue of the concentration of the food source.

Complexity measures have associated spatio-temporal scales which must be appropriate for the detection task. Tortuosity is measured as the inefficiency of a path segment, the length must be selected. Likewise positional entropy measures the complexity of a moving segment of the path. Passage time is measured with respect to a spatial region whose size must be chosen. These scales must bear in mind the classic trade-off between resolution and signal-to-noise ratio; the smaller the scale, the more fine structure will be detected and the more tightly determined will be the position of identified waypoints. Longer scales improve the signal-to-noise ratio by filtering out high-frequency noise from the behavior but also reduce the resolution of the measurement.

Variation in complexity measures has been suggested to correlate with changes in the animal's behavioral state [1, 2, 4, 6, 11] and to segment the path into distinct behavioral phases [1, 6, 13] supporting the use of complexity to identify key path segments that may correspond to waypoint locations. Path complexity has also been linked to the surrounding visual environment [7], indicating that the waypoints identified may correspond to visual landmarks, as expected under the pilotage hypothesis.

Path Predictability

Path predictability differs distinctly from the previous two approaches since it uses complete movement paths in a global analysis, rather than examining only localized properties of the paths. Waypoints are selected as those segments of the path that can be used to provide the best prediction of the complete path and other paths (where these exist). The logic of this approach depends on the hypothesis of navigation by *pilotage*. This hypothesis claims that the animal makes it way from its original position to its final objective by successively visiting an ordered sequence of fixed waypoints, as per the definition above. The path predictability method argues that if this hypothesis is accurate, the path(s) of the animal must be defined by the locations of those waypoints, the segments between waypoints being subject to constraints such as energy minimization and random variation from environmental and internal factors. These waypoints are thus analogous to the fixed positions in a curve-fitting exercise, with the constraints given above acting similarly to the constraint of curvature minimization in spline fitting. The upshot of this is that those segments of the path(s) close to the waypoints should provide the maximum amount of information about the rest of path(s).

Similarly to path complexity, path predictability is intimately linked to information. Consider trying to convey to someone not just a path segment but instead a set of complete paths. Now imagine that the only information you are allowed to send is the position of the paths at a small set of times. If your friend subsequently "joins the dots" to link these points, how well will they approximate the true paths? The most informative, or predictive, locations are those that allow the best estimate of the real paths using this technique.

Technically, the information contained in a path can be defined how it affects the probability of another path. If path B is informative about path A then the probability of path A will increase as a result of knowing path B. Mathematically, the information contained in path B about path A is related to the ratio of probability of path A, conditional on knowing path B through some model M, compared to the probability of path A when B is unknown. Information is defined as the log of this probability ratio:

$$\mathrm{Information(A|B)/bits} = \log_2 \mathrm{p(path\,A \mid path\,B, M)} - \log_2 \mathrm{p(path\,A \mid M)} \quad (1)$$

The probability P(path A | path B, M) depends on the choice of model. A simple model might be that each point on path A should be the same as on path B at the equivalent point in time, give or take some amount of Gaussian noise to account for natural variation. More realistic models should take account of the correlations over time that lead to the spatial structure of the path. Mann et al. [10]

provide a more sophisticated model choice implemented via the use of Gaussian processes, allowing for autocorrelation within the path over time as well as the correlations between paths. In this implementation, all paths are assumed to be samples from a distribution centered on a mean path – similar to the mean path defined as part of the path similarity measure. Constructing a model to describe the probability distribution of movement paths is a complicated task which is not readily reducible to a simpler scheme. This task is necessary, however, to specify the mutual information contained between paths and between elements of the same path and thus to be able to use path predictability as a measure. See Mann et al. [10] for guidance on model construction.

Having defined a model and therefore a probability distribution from one path to another, the task of identifying the waypoints can be clarified as follows. Path B will be a recorded path, with positions recorded at a set of time points, t. Let t' be a subset of those timesteps such that path B' is the recorded positions at times t'. The information in those positions is calculated as

$$\begin{aligned} \mathrm{Information}(t') &= \mathrm{Information}(A \mid B') \\ &= \log_2 p(\mathrm{path}\, A \mid \mathrm{path}\, B', M) \end{aligned} \tag{2}$$

Having hypothesized that waypoints correspond to the most informative segments of the path, the task is now rephrased as finding the subset of t, t*, that maximizes Eq. 2,

$$t^* = \mathrm{argmax_}t' \mathrm{Information}(t') \tag{3}$$

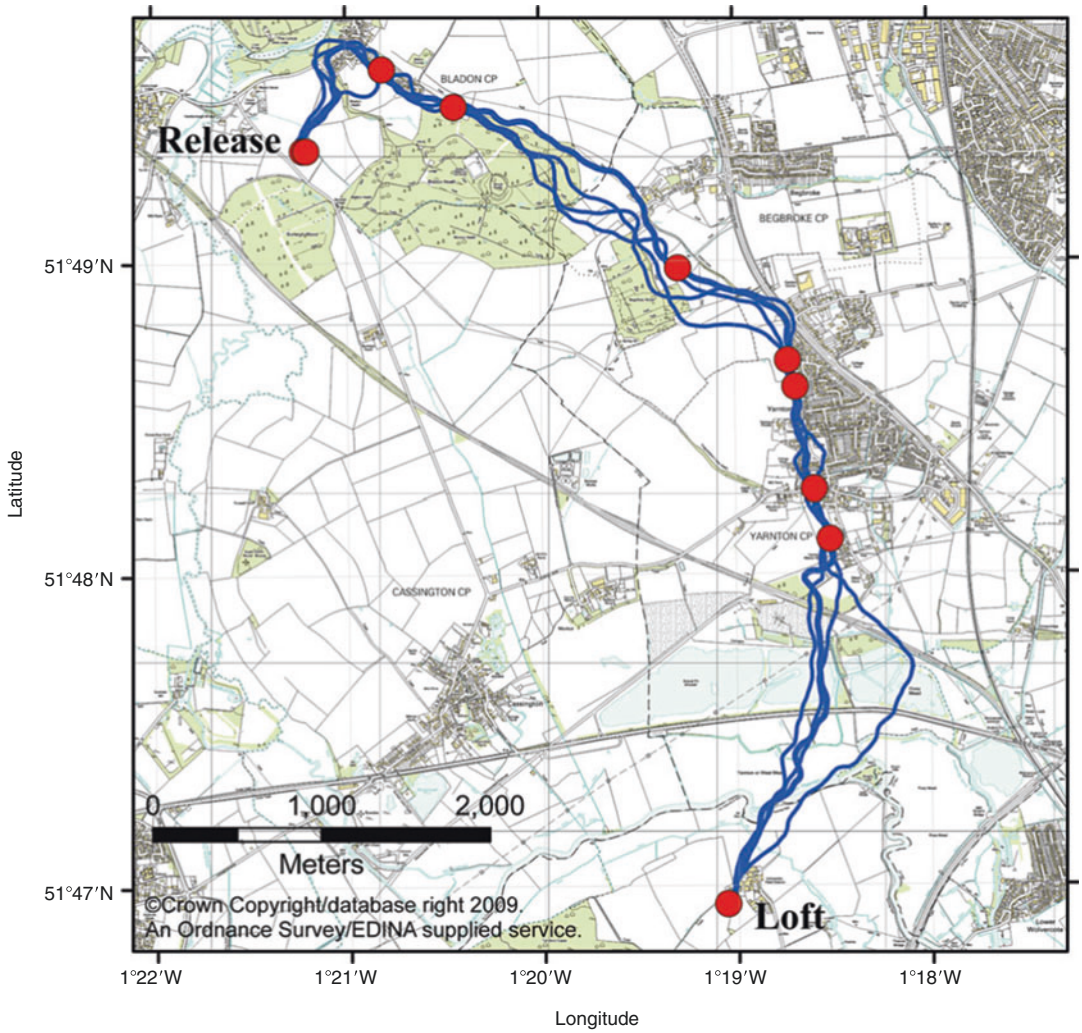
Finding a limited number, m, of waypoints can be accomplished simply by searching for the t* restricted to subsets of size m. Optimization of Eq. 2 can be accomplished by "greedy" forward selection, gradient ascent, Monte Carlo sampling, or other established optimization techniques. Standard texts on learning algorithms, for example, [9], can provide comprehensive details of optimization algorithms.

Predictability between different paths tends to be maximal under two circumstances. The first is in regions where the two paths are very similar. Here knowing one path gives good information about the likely location of the other. This then is similar to the path similarity criterion. The other case is where the location of the path is highly unpredictable either a priori, or conditional on the other chosen waypoints. Locally this bears a resemblance to the path complexity measure, but also takes account of the global unpredictability – for example finding path segments that are locally smooth but globally far from the straight-line path between start and finish. This corresponds to an intuitive analysis of the path. If the bird consistently flies to a location far from the most efficient route this is unlikely to be due to chance; there must be a waypoint at this location.

Figure 3 shows an example of waypoints selected by maximizing the path predictability, taken from Mann et al. [10]. In this demonstration, the number of selected waypoints is ten and the information in Eq. 2 is maximized by forward selection. It can be seen from this example that the selected waypoints each exhibit at least one of the three properties that maximize information content. Many of the waypoints are located where the paths coalesce into narrow corridors, exhibiting high path similarity. Others are located where the paths turn sharply, which correspond to high path complexity. Finally there are waypoints located in the regions furthest from the most efficient beeline path, at the apex of the curve, which is the most unlikely area for the pigeons to visit a priori.

Summary

Detection of waypoints in recorded movement paths is a novel problem that has potential future applicability in studies of animal movement and biomimicry. The methods described in this entry constitute a basis for identifying important regions of the movement path based on hypotheses about the likely effects of waypoints upon the animal's movement. These methods span a range of mathematical complexity and intuitive applicability. More recent developments such as the path



Waypoint Detection, Fig. 3 Ten waypoints identified from five flight paths from a single homing pigeon, using path predictability as a criterion. The waypoints are identified preferentially in flight segments that a similar between paths, where the path exhibits high complexity and where the path is far from the most efficient straight "beeline" between start and finish (Reprinted from Ref. [10], Copyright [2010], Creative Commons open access license)

predictability tend toward higher levels of abstraction. The methods explored in this entry are not mutually exclusive and combinations should be explored to achieve the best detection results. With animal movement tracking still developing as an experimental field, methods for the analysis of recorded data are being created in parallel and the development of further methods for identifying salient regions from movement paths can be expected in the near future.

Cross-References

- ▶ Biomimetics
- ▶ Nanoparticle Tracking Analysis

References

1. Barraquand, F., Benhamou, S.: Animal movements in heterogeneous landscapes: identifying profitable places and homogeneous movement bouts. Ecology **89**, 3336–3348 (2008)

2. Biro, D., Guilford, T., Dell'Omo, G., Lipp, H.-P.: How the viewing of familiar landscapes prior to release allows pigeons to home faster. J. Exp. Biol. **205**, 3833–3844 (2002)
3. Biro, D., Meade, J., Guilford, T.: Familiar route loyalty implies visual pilotage in the homing pigeon. Proc. Natl. Acad. Sci. U. S. A. **101**(50), 17440–17443 (2004)
4. Fauchald, P., Tveraa, T.: Using first-passage time in the analysis of area-restricted search and habitat selection. Ecology **84**, 282–288 (2003)
5. Freeman, R., Mann, R., Guilford, T., Biro, D.: Group decisions and individual differences: route fidelity predicts flight leadership in homing pigeons (Columba livia). Biol. Lett. **7**(1), 63–66 (2010). Advance online publication
6. Guilford, T., Roberts, S., Biro, D., Rezek, I.: Positional entropy during pigeon homing II: navigational interpretation of Bayesian latent state model. J. Theor. Biol. **227**(1), 25–38 (2004)
7. Lau, K.-K., Roberts, S., Biro, D., Freeman, R., Meade, J., Guilford, T.: An edge-detection approach to investigating pigeon navigation. J. Theor. Biol. **239**(1), 71–78 (2006)
8. Lipp, H.-P., Vyssotski, A.L., Wolfer, D.P., Renaudineau, S., Savini, M., Tröster, M., Dell'Omo, G.: Pigeon homing along highways and exits. Curr. Biol. **14**, 1239–1249 (2004)
9. Mackay, D.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge (2003)
10. Mann, R., Freeman, R., Osborne, M., Garnett, R., Armstrong, C., Meade, J., Biro, D., Guilford, T., Roberts, S.: Objectively identifying landmark use and predicting flight trajectories of the homing pigeon using Gaussian processes. J. Roy. Soc. Interface. **8**, 210–219 (2011)
11. McKensie, H.W., Lewis, M.A., Merrill, E.H.: First Passage time analysis of animal movement and insights into the functional response. Bull. Math. Biol. **71**(1), 107–129 (2009)
12. Meade, J., Biro, D., Guilford, T.: Homing pigeons develop local route stereotypy. Proc. R. Soc. B. **272**(1558), 17–23 (2005)
13. Roberts, S., Rezek, I., Guilford, T., Biro, D.: Positional entropy during pigeon homing I: application of Bayesian latent state modelling. J. Theor. Biol. **227**(1), 39–50 (2004)

Wear

► Nanotribology

Wet Adhesion in Tree and Torrent Frogs

Thomas Endlein and W. Jon. P. Barnes
University of Glasgow, Institute of Molecular Cell and Systems Biology, Centre for Cell Engineering, Joseph Black Building, University Avenue, Glasgow, Scotland, UK

Overview

Biomimetics of Animal Adhesion

Our understanding of the adhesive mechanisms of climbing animals has increased rapidly in recent years, in no small way due to advances in materials science providing both the tools and the theoretical background. This research has shown that, as a result of millions of years of evolution, the adhesive mechanisms of climbing animals are highly dynamic, showing many features that are the envy of materials scientists. Unlike most man-made adhesives, they (i) cope well with rough and anti-adhesive substrates, (ii) have self-cleaning mechanisms and so recover quickly following contamination, (iii) can control attachment so that it only occurs when required, and (iv) possess mechanisms for effortless detachment. Finally (v), since animals attach and detach their adhesive pads every time they take a step, sticking ability is not lost with repeated application, this being mainly due to their resistance to wear and contamination. To date, most research effort has been directed at investigating the adhesive mechanisms of geckos (gecko adhesion) and insects (adhesion in insects), but tree and torrent frogs, adapted to living in habitats with a high rainfall, offer solutions to different problems, namely, adhesion under wet conditions.

Climbing and Adhesion

It goes without saying that climbing animals, in addition to adaptations for climbing, need mechanisms for maintaining a grip and thus avoiding falling. Taking, for example, an animal on an

inclined tree trunk, the force of gravity will cause this animal to fall unless it can generate a reaction force between itself and the trunk that is equal and opposite to the force generated by the gravitational pull. Such forces can be generated in a number of ways, namely, mechanical interlocking, friction, and bonding. *Interlocking* involves interlocking the surface of the animal with that of the support. This is how claws, present in many climbing animals (e.g., squirrels and members of the cat family), work. Claws may either catch on preexisting surface irregularities or be pushed into the surface if it is soft enough. Many animals without claws (e.g., monkeys and humans) utilize a simple *friction* grip. It only works, however, when hands or feet can span almost the entire branch diameter – i.e., a vertical wall cannot be climbed by friction alone. Many specialist climbers, however, possess specialized adhesive pads on their feet and adhere by *bonding* (bioadhesion). Such bonds need to be reversible; otherwise, they would only be good for permanent attachment. The main mechanisms of bonding are dry and wet adhesion. In *dry adhesion*, the adhesive pads are able to get so close to the substrate that adhesion is primarily by van der Waals forces. Such forces can hold molecules together in liquids and solids and only operate at distances of a few nanometers. In *wet adhesion*, on the other hand, there is a fluid joint, secreted liquid between the pad and substrate, and adhesion is thought to be mainly by capillarity. It is generally accepted that dry adhesion is found in lizards such as geckos [1], while wet adhesion is found in insects [2], in amphibians [3] and, among mammals, in arboreal possums [4] and at least one bat species [5].

Two rather different types of adhesive pads – smooth and hairy – have evolved that allow animals to climb and adhere to vertical and overhanging surfaces. Hairy adhesive pads are found in many insects, spiders, and lizards. In geckos, they consist of millions of tiny branching hairs (setae), each branch ending in a flattened spatula no more than 200 nm across. This allows the tips of the hairs to form extremely close contact to rough surfaces, enabling intermolecular van der Waals forces (as described above) to provide strong adhesion [1]. In contrast, the smooth adhesive pads of tree frogs, arboreal salamanders, and insects such as ants secrete a fluid so that contact is maintained by wet adhesion [6]. Interestingly, the hairy pads of insects also secrete tiny amounts of fluid, so that, in contrast to geckos, adhesion is also likely to be by wet adhesion. Until recently, spiders were thought to use dry adhesion, but there is new evidence of a fluid secretion from the adhesive setae [7], so it is possible that they, like insects, use wet adhesion.

Adhesion in Tree and Torrent/Stream Frogs

Among the Amphibia, adaptations for climbing are found in neotropical salamanders of the genus *Bolitoglossa* and at least ten families of frogs. Where the habitat is vegetation (herbs, shrubs, trees), the frogs are known as tree frogs, while those found clambering on wet rock in the vicinity of streams and waterfalls are known as rock and torrent frogs, respectively. All are characterized by the presence of adhesive pads on their digits. In spite of the fact that these pads have evolved independently in different families, their structures are very similar, consisting of polygonal (mostly hexagonal) flat-topped cells separated by narrow channels. Interestingly, similar structures are also found on the smooth adhesive pads of some insects (e.g., grasshoppers) [8]. The development of adhesive pads is thus an excellent example of convergent evolution. This points toward a favorable design (in frogs at least), and this is important for biomimetics. Most frog adhesion research has been carried out on tree frogs of the families Hylidae and Rhacophoridae, while work on torrent frogs of the family Ranidae is still at an early stage.

Physical Principles

Although the dominant force in tree frog adhesion is believed to be capillarity, hydrodynamic forces based on viscosity are also considered to play a role. Additionally, since the thickness of the fluid layer was measured to be less than 35 nm in some areas of the contact zone, it is probable that there is direct contact between nanostructural features of the pad surface and the substrate [9]. Thus,

van der Waals forces may also be involved, but to what extent is unknown. The physical principles underlying capillarity and viscous forces are outlined below.

Wet Adhesion: Capillary Forces

A common model that has been used for quantifying capillary forces consists of a sphere on a plane surface, connected by a drop of liquid (Fig. 1). In this model, the surface tension of the meniscus results in a pressure difference, where the pressure inside the liquid is lower than it is outside, as long as the meniscus is concave. This pressure difference, the Laplace pressure, will resist the separation of the two surfaces. The attractive Laplace force (F_L) for a macroscopic, perfectly smooth, and homogeneous sphere in contact with a plane is given by

$$F_L = -2\pi R\gamma(\cos\theta_1 + \cos\theta_2) \quad (1)$$

where R is the radius of the sphere, γ is the surface tension of the liquid, and θ_1 and θ_2 are the contact angles of the liquid with the plane surface and sphere, respectively. If the liquid completely wets both surfaces (i.e., $\theta_1 = \theta_2 = 0$), Eq. 1 simplifies to

$$F_L = -4\pi R\gamma \quad (2)$$

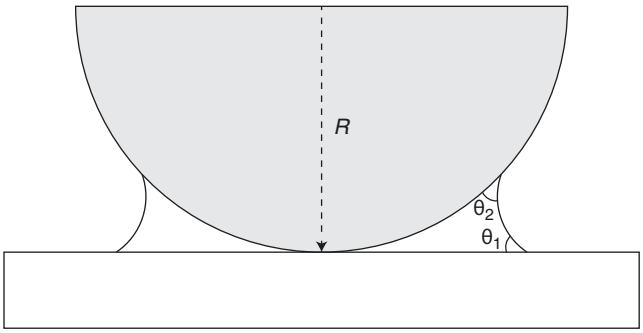
An alternative model examines capillary forces between two rigid plates, separated by a thin layer of fluid (Fig. 2). This model produces substantial forces if the radius (r) of the area of contact is large and the separation distance of the plates (h) is small, according to the equation:

$$F_L = \pi r^2\gamma\left[r^{-1} - (\cos\theta_1 + \cos\theta_2)h^{-1}\right] \quad (3)$$

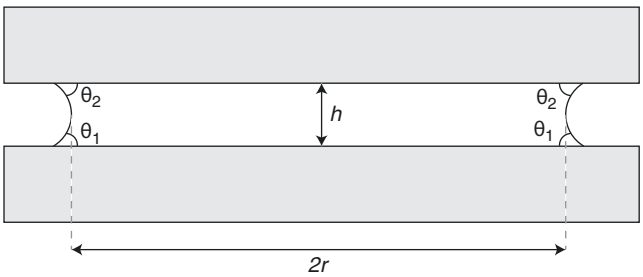
In a large flat meniscus where $r \gg h$ and $\theta_1 = \theta_2 = 0$, Eq. 3 approximates to Eq. 4, namely,

$$F_L \approx -2\pi r^2\gamma/h \quad (4)$$

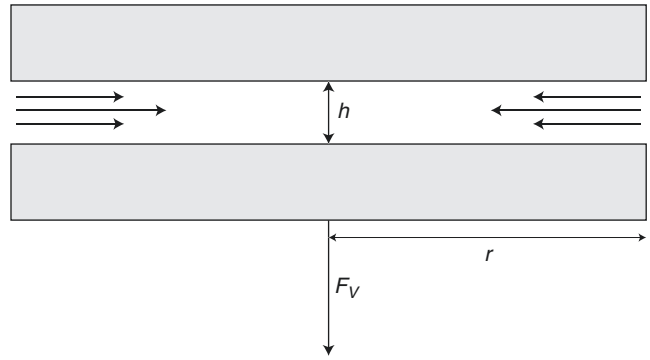
A second component of the adhesive force is the tensile force (F_T) of the meniscus. In the rigid plate model, this surface tension force is



Wet Adhesion in Tree and Torrent Frogs, Fig. 1 Capillarity model of a sphere on a plane surface connected by a drop of fluid



Wet Adhesion in Tree and Torrent Frogs, Fig. 2 Capillarity model of two flat, rigid surfaces separated by a small volume of fluid



Wet Adhesion in Tree and Torrent Frogs, Fig. 3 Stefan adhesion model of two flat rigid plates, fully immersed in a fluid, subject to an external force

$$F_T = -2\pi r\gamma \sin\theta \tag{5}$$

The total meniscal force (F_M) is the sum of F_L and F_T:

$$F_M = F_L + F_T \tag{6}$$

F_T is negligible in any single large meniscus (e.g., ones with a radius of 0.5–3 mm as occur in tree frog toe pads), but would contribute to a larger extent to the adhesive forces in a fly's adhesive pad where you have large numbers of microscopic menisci (radii of ca. 1 μm).

In a recent study, Eq. 2 has been extended from hard, undeformable surfaces and spheres to soft, elastic materials, such as the toe pad of a tree frog. The mathematical equations relate the capillary attraction between the bodies to their elastic repulsion. Although they do not lend themselves to any easy calculation of the adhesive force, they are of interest as they predict that F_M scaling will gradually change from length scaling to area scaling with increasing R, this change occurring more rapidly for materials with a lower effective elastic modulus (E_{eff}). For instance, in Eq. 7 which deals with the Laplace pressure component of adhesion, the first term is proportional to R and is identical to the force with a hard sphere (Eq. 2), while the second term is proportional to R^2. It is negligible for high Young's moduli, but becomes significant for soft materials.

$$F_L = -4\pi R\gamma - \left(\frac{\pi\gamma}{2r'}\right)^3 \cdot \frac{2R^2}{3\left(E_{eff}^2\right)} \tag{7}$$

where r' is the radius of curvature of the meniscus.

Wet Adhesion: Viscous Forces

The second component of wet adhesion is provided by time-dependent viscous forces (such as Stefan adhesion). Consider two rigid plates fully submersed in a fluid (Fig. 3). Separating the plates involves fluid flowing into the gap between them, so that separation is resisted by a viscous force until the fluid movements are complete. This force (F_V) will be greater for more viscous fluids and for smaller values of h. As Eq. 8 indicates, F_V scales with area squared.

$$F_V = -\frac{dh}{dt} \cdot \frac{3\pi\eta r^4}{2h^3} \tag{8}$$

where η is the viscosity of the liquid and t is the time needed to separate the two plates. For completeness, the equation for the hydrodynamic force between a sphere and a plate (Eq. 9) is also included.

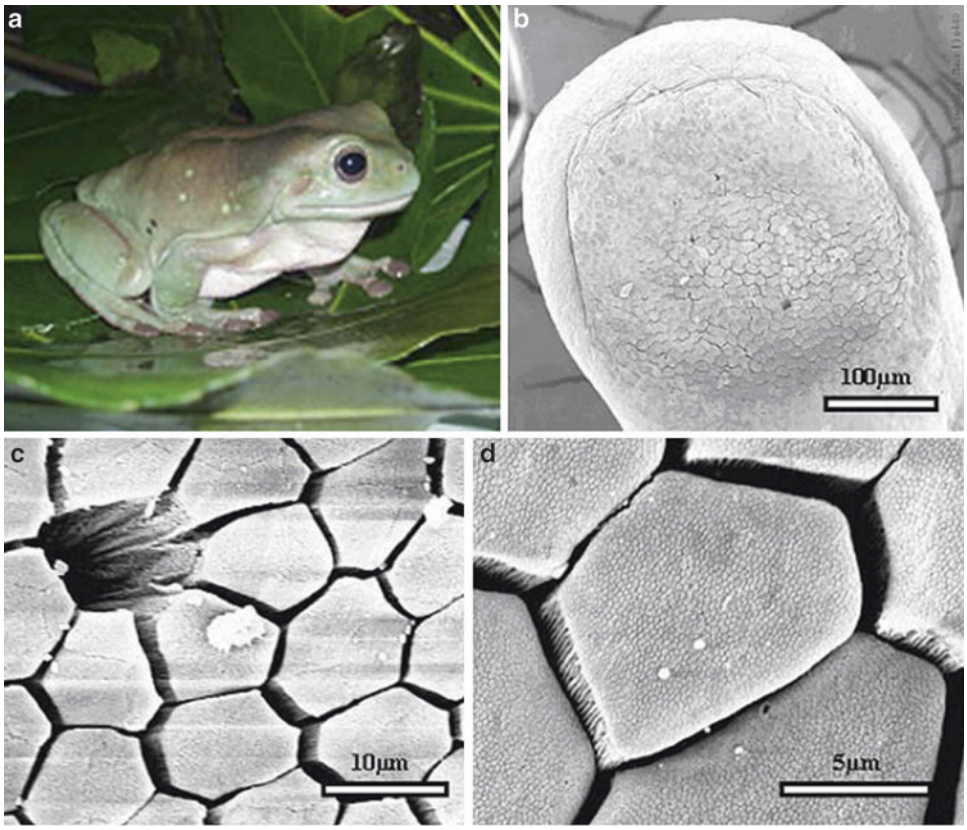
$$F_V = -6\pi\eta r \frac{dh}{dt}\left(1 + \frac{r}{h}\right) \tag{9}$$

When the sphere touches the plate, as in Fig. 1, $h = 0$. There is a need to extend such studies to soft, deformable materials.

Key Research Findings

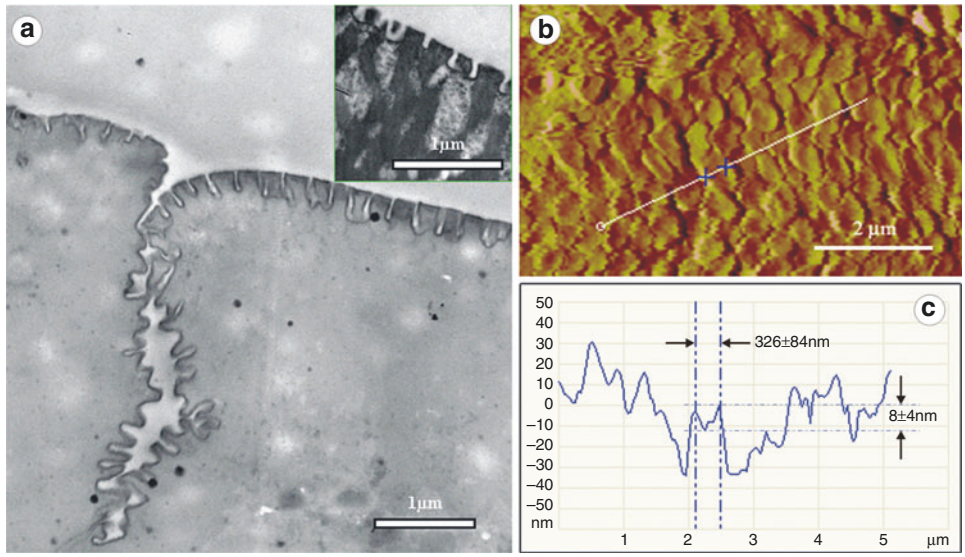
Toe Pad Micro- and Nanostructure

In both tree and torrent frogs, soft domed pads occur on the ventral surface of the tip of each digit, the specialized pad epithelium being delineated from normal skin by distinct grooves (Fig. 4).



Wet Adhesion in Tree and Torrent Frogs, Fig. 4 (**a**) White's tree frog, *Litoria caerulea*. (**b**–**d**) Scanning electron microscopy of toe pad epithelium. (**b**) Low-power micrograph of whole pad. (**c**) Micrograph showing a mucus pore and (largely) hexagonal epithelial cells separated from each other at their distal ends by channels. (**d**) Higher-power micrograph indicating the presence of nanostructuring on the "flat" surface of the epithelial cells (From Scholz et al. [15])

There are also much smaller adhesive areas located elsewhere on the ventral surface of the feet, in particular the subarticular tubercles located more proximally on the digits. Tree frog adhesive pads have a stratified columnar epithelium, the cells being separated from each other at their apices. Scanning electron microscopic studies (scanning electron microscopy) show that most of these cells are hexagonal, but, as Fig. 4 shows, some are pentagonal and a few heptagonal. Pores of mucous glands open into the channels between the cells. The toe pad epithelium thus consists of an array of flat-topped cells (approx. 10–15 µm in diameter each) separated by mucous-filled grooves (approx. 1 µm wide) [3, 10–12]. In functional terms, having the cells separated at their tips enables the pads to conform to the shape of surface irregularities to which the frog is adhering. The mucous glands are necessary to produce the watery secretion that forms an essential part of the adhesive mechanism of the pad. The hexagonal array of channels that surround each epithelial cell presumably functions to spread mucus evenly over the pad surface and, under wet conditions (most tree frog species live in rainforests), remove surplus water. Finally, the presence of grooves could aid adhesion by reduction of crack propagation (peeling) [13]. Pull-off stress is spread between a larger number of hexagons rather than being concentrated at the edge of the contact zone. Such features have been incorporated into bioinspired artificial patterned surfaces to increase their adhesion [14]. In torrent frogs, the epithelial cells have



Wet Adhesion in Tree and Torrent Frogs, Fig. 5 White's tree frog, *Litoria caerulea*, (**a**) Transmission electron micrograph of a toe pad showing a section of one of the channels that separate adjacent epithelial cells and the surface nanostructures, which are themselves separated from each other by narrow channels. The *inset* shows similar nanostructures on a toe pad of the hylid tree frog, *Scinax ruber*. Here the nanostructures are associated with filaments running at right angles to the cell surface. (**b**) Atomic force microscope deflection image showing the shape of the dense array of peg-like nanopillars that constitute the adhesive surface of the pad. (**c**) Height profile of line in **b**. The two crosses delineate a columnar nanopillar that lay precisely on the profile line, showing that these peg-like structures have a small dimple at their center. The figures are means ± standard deviations of nanopillar width and dimple depth. (NB The height of the nanopillars cannot be measured by the atomic force microscope as the cantilever cannot reach the base of the channels between them) (Redrawn from Scholz et al. [15])

become elongated along the pad's proximal-distal axis, which results in the channels between them being straighter and shorter in this direction, presumably an adaptation for efficient drainage of excess water.

At the nanometer scale, the pad surface is not smooth but is covered by a dense array of nanopillars, 300–500 nm in diameter and 200–300 nm in height. Under the atomic force microscope, they can be clearly seen to have small dimples at their tips (Fig. 5). In many species, these nanostructures are the ends of fiber bundles lying within the cytoplasm of the epithelial cells that run at right angles to their external surface. Similarly, oriented fibers are found in the smooth adhesive pads of insects. As in tree frogs, they form the cytoskeleton of the pads. The functions of the nanopillars that constitute the so-called "flat" surface of these cells remain a matter for speculation. The following is a list of obvious possibilities, none of which are mutually exclusive. (1) Like the epithelial cells and the channels that surround them, the nanopillar array may allow close conformation to surface irregularities, but on a much smaller length scale (nanometers rather than micrometers) than applies to whole epithelial cells. (2) The narrow channels between them could serve to absorb excess water (like a sponge), much as sipes (fine-scale grooves) do on a wet weather car tire [13]. This would allow rapid optimization of the thickness of the intervening fluid layer (as thin as possible, but without any air pockets). (3) The nanopillars could be very important in the generation of friction forces, in that their tips will be in actual contact with the surface (see later). (4) It cannot be excluded that the dimples give rise to a suction effect.

Neotropical salamanders of the genus *Bolitoglossa* are also arboreal. Some species have toe pads, whereas in others, digits are much reduced, and there is extensive digital webbing

[16]. In these cases, the whole foot appears to function as a single smooth pad. In smaller species, adhesion is thought to be by wet adhesion, but in some larger species, there is circumstantial evidence that suction may play a role [16]. In most species, the surfaces of the cells are smooth, although in *Bolitoglossa odonelli*, the surface is covered by nanopillars that appear very similar to those found in tree frogs. Many newts can also climb smooth vertical surfaces but do not show the complexity of structures found on (tree and torrent) frog pads.

Physical Properties of Toe Pads

The elastic modulus (or in this case the effective elastic modulus, as Poisson's ratio for toe pads is not known) provides a quantitative measure of material stiffness, an important parameter for adhesion. Stiffness is estimated from force-distance curves, resulting from indentation of the toe pads, using equations based on, for instance, the Johnson-Kendall-Roberts (JKR) theory. Such equations will also allow calculation of the work of adhesion and the pull-off force. Micro-indentation usually involves specialized microtribometers, while, at the nano-level, either a nanoindenter or an atomic force microscope can be used.

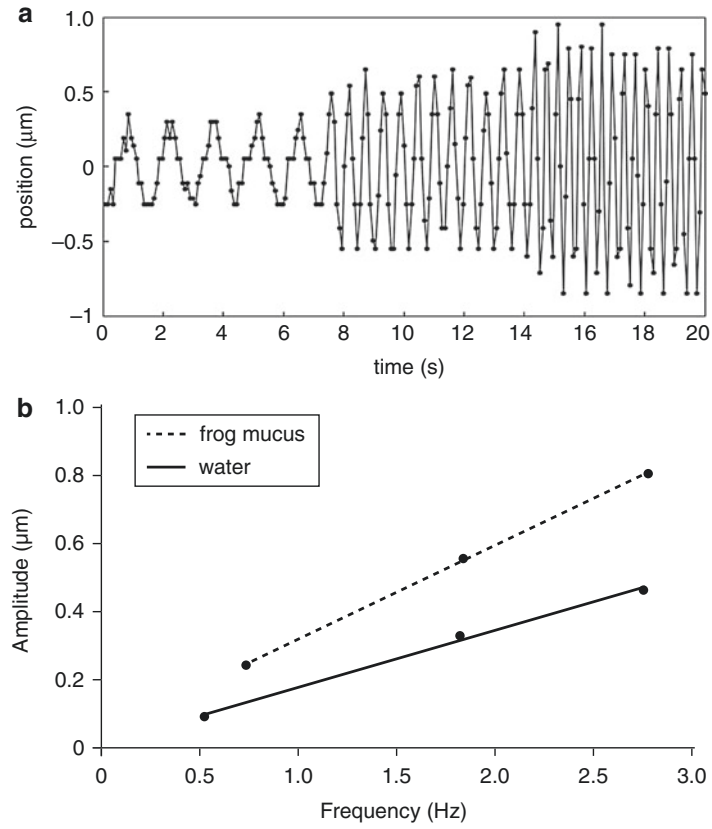
Use of atomic force microscope in nanoindenter mode indicates that the outer keratinized layer of tree frog toe pads, which is about 15 μm thick, has an effective elastic modulus (E_{eff}) of 5–15 MPa, equivalent to that of silicone rubber [15]. However, it gives no information on the physical properties of deeper structures. To measure the stiffness and elasticity of the pad as a whole, micro-indentation is used. It demonstrates that tree frog toe pads are among the softest of biological tissues and that they show a gradient of stiffness, being stiffest on the outside due to increasing keratinization of the outer cell layers. Estimates of E_{eff} lie in the range 4–25 kPa [17]. Although not as soft as adipose tissue ($E_{\text{eff}} \approx 3$ kPa), values are much lower than for aorta ($E_{\text{eff}} \approx 500$ kPa), cartilage ($E_{\text{eff}} \approx 20$ MPa), or the adhesive pads of locusts ($E_{\text{eff}} = 250$–750 kPa). At the opposite extreme, tooth enamel has an E_{eff} value of 60,000 MPa. Functionally, the stiffer outer surface provides a degree of resistance against wear, while the softer interior provides an increased ability to form intimate contact with various surface profiles. Since the epithelium as a whole is a living, growing tissue, the long-term solution to the wear problem is a simple one; the outer cell layer is shed at intervals, presumably after the cell layer below it has become keratinized.

A dense network of blood capillaries lies below the pad epithelium. This may well have a mechanical function, acting as a shock absorber. This could be particularly important when a tree frog lands on a vertical surface after a jump. Having soft inelastic material beneath the pad allows the pad to mold to the surface the frog has landed on without any tendency for elastic recoil.

Viscosity and Thickness of Fluid Layer

Microrheometry of toe pad mucus has been undertaken using a laser tweezer technique (optical tweezers) [18]. Latex beads (2.8 μm in diameter) were mixed with droplets of toe pad mucus (ca. 5 μL in volume) on a coverslip on a motorized stage of a microscope equipped with a laser tweezer. The laser tweezer traps individual beads in its focus. The stage was then oscillated sinusoidally at a number of different frequencies. The fluid movement exerts a drag on the beads, which is dependent on the velocity and also on the viscosity of the fluid. Graphs are plotted of the amplitude of the bead's lateral displacement against oscillation frequency. The viscosity of the mucus was calculated by comparing the slopes of the regression lines obtained for frog mucus and for pure water. Such measurements (Fig. 6) show that the viscosity of toe pad mucus is approximately 1.43 mPa · s (i.e., approx. 1.4 × the viscosity of water). No evidence for non-Newtonian properties was found, but it is possible that the situation changes for more concentrated mucus.

The thickness of the mucus layer beneath the frog's toe pad has been investigated by a technique called interference reflection microscopy. This enables the calculation of the fluid layer thickness from an optical interference pattern of reflected wave fronts [18]. Monochromatic light was reflected both from the surface of the pad and



Wet Adhesion in Tree and Torrent Frogs, Fig. 6 Microrheology of mucus from White's tree frog, *Litoria caerulea*, using a laser tweezer technique. (**a**) Amplitude of bead displacement elicited by sinusoidal movement of the microscope stage (stimulus amplitude ≈ 10 µm) at three different frequencies. (**b**) Relationship of bead displacement amplitude to the stimulus frequency, measured for both toe pad mucus and pure water. The difference between the slopes of the two lines indicates that this sample of toe pad mucus was 1.65 times more viscous than water (Redrawn from Federle [18])

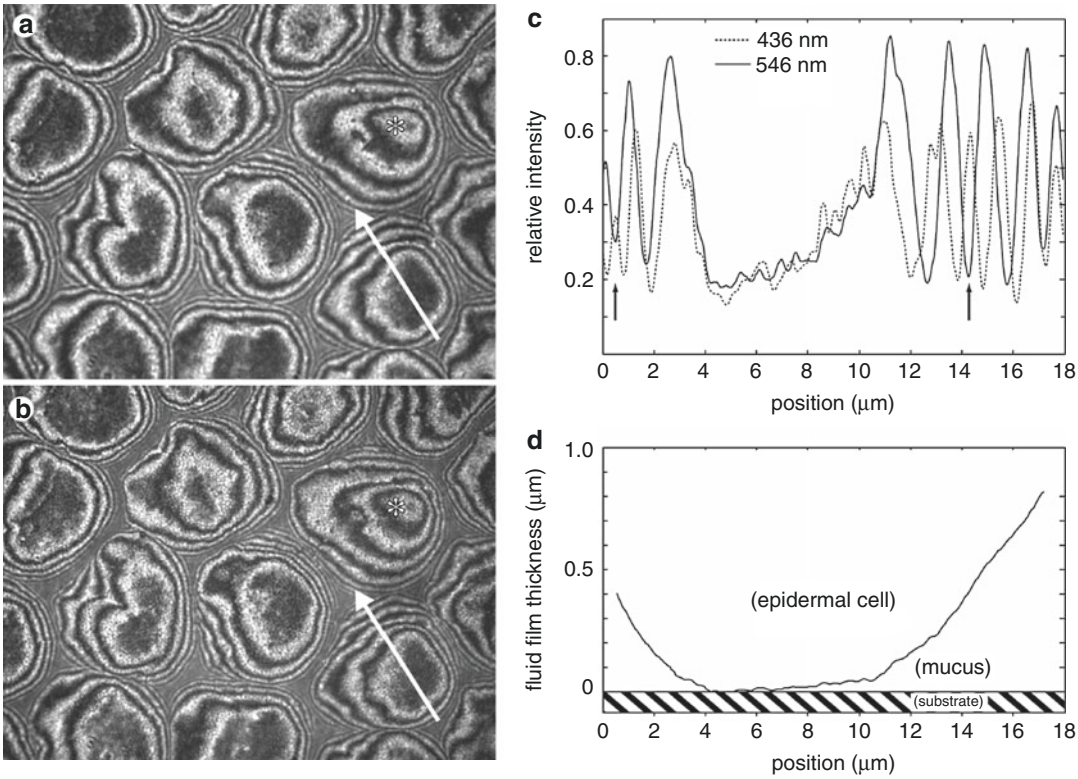
from the top surface of the glass plate on which the frog was sitting (a large coverslip). This leads to interference effects, and the intensity of the reflection allows estimation of the thickness of the fluid layer. Different interference fringes are distinguished by comparing image pairs obtained by light of different wavelengths (Fig. 7).

Examination of images shows that most of the toe pad epithelial cells had dark zones in the center, which correspond to zero-order interference minima (i.e., to a mucus film thickness of less than 100 nm). Further analysis of the intensity of these central areas in relation to interference maxima and minima leads to the conclusion that the film thickness ranges from 0 to 35 nm and is indistinguishable from zero in many epithelial cells (median 6.0 nm). Given that there will be fluid in the narrow channels between the nanopillars, it cannot be excluded that the tops of many nanopillars will be in actual contact with the substrate in an attached toe pad. Additionally, in some species, an increase in the area of closest contact is observed as frogs are tilted from the horizontal toward the vertical.

Adhesion and Friction Forces

Tree frogs are good climbers, many tropical species being found high in the canopy of rain forests. When supports such as twigs are available, they are grasped, and frogs climb by extending limbs and pulling like human climbers. Toe pads come into their own on smooth substrates such as leaves. Unlike geckos, tree frogs cannot run across a ceiling, but they can climb vertical and overhanging surfaces, and a small tree frog can hang on to an inverted glass plate using just its toe pads. At rest, thigh and belly skin aid adhesion.

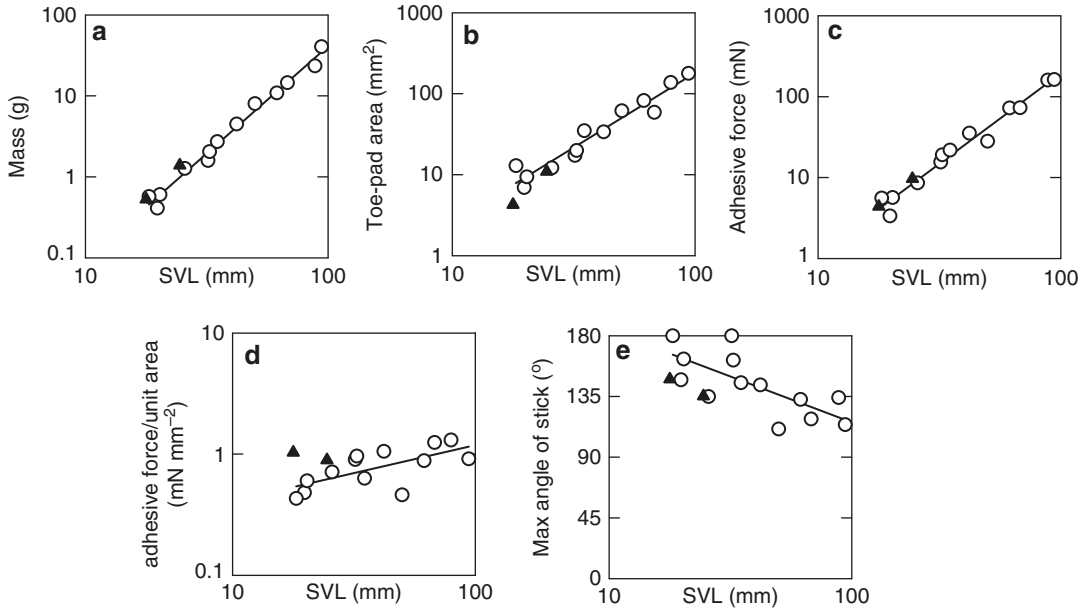
Quantitative measurements of the adhesive (normal) and friction (parallel) forces that can be generated by tree frogs have been made by a simple procedure first developed by Emerson and Diehl [11]. Immediately after being weighed,



Wet Adhesion in Tree and Torrent Frogs, Fig. 7 In vivo analysis of White's tree frog, *Litoria caerulea*, toe pad contact with glass using interference reflection microscopy. (**a**, **b**) Images of the same area of the contact zone using different wavelengths (**a**, 436 nm; **b**, 546 nm) using an illuminating numerical aperture of 0.27. *Asterisks* indicate an epithelial cell that is not in close contact (fluid film thickness > 40 nm). (**c**) Intensity profile along the *arrow* shown in **a** and **b**. *Arrows* indicate that the position of the second-order "green" minimum coincides with the second-order "blue" maximum. From this, it can be deduced that the central dark areas of most epithelial cells are zero-order minima. (**d**) Reconstruction of the fluid film thickness along the *arrow* shown in **a** and **b** (Redrawn from Federle [18])

each frog was placed "head up" on a smooth surface on a rotation platform that is rotated slowly from 0° (horizontal) through 90° (vertical) to 180° (upside down). The angles at which the frog first slips on the surface (slip angle) and falls from the surface (fall angle) were recorded. These slip and fall angles are used to calculate maximum shear and adhesive forces, respectively, by simple trigonometry. If the total toe pad area is also measured, these values can be converted to the force per unit area that toe pads can generate. In all species studied, toe pad adhesive force per unit area lies in the range 0.7–2.0mNmm^{-2} (0.7–2.0 kPa) [19].

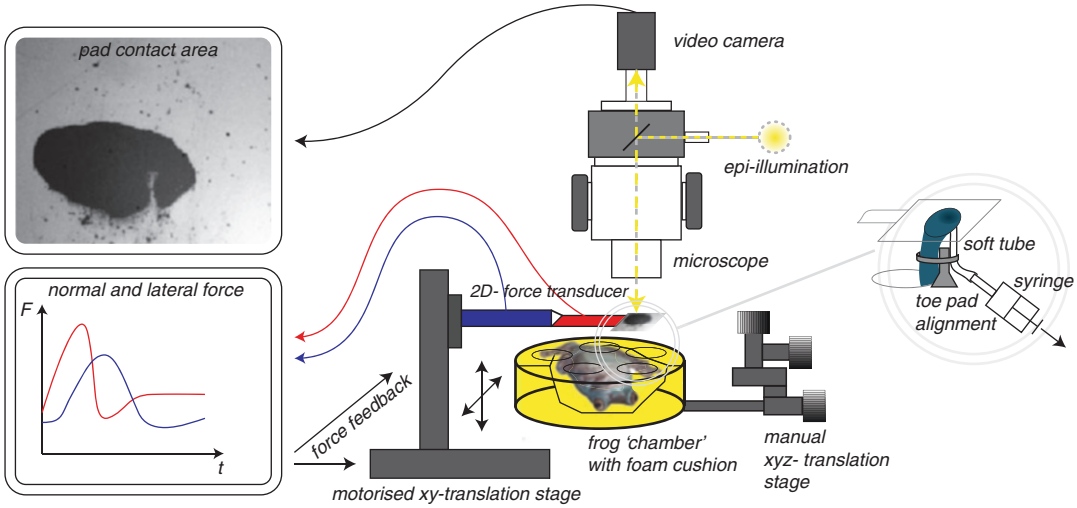
Frogs are good models for allometric studies, as general body shape is maintained from metamorphosis to full adult size, i.e., frogs grow approximately isometrically. By plotting mass, toe pad area, and adhesive force against frog length on log-log coordinates, both between and within species, it can be seen that adhesive force (F_A) scales approximately with toe pad area or length squared ($F_A \propto a$ or $F_A \propto l^2$) (Fig. 8; [11, 12, 20]. This presents a problem for larger frogs, since mass is proportional to the cube of linear dimensions, while adhesive force, depending upon area, would only be proportional to the square of the linear dimension. A within-species study of hylid tree frogs in Trinidad [12] shows that toe pad area increases as length squared, consistent with area dependence, but mass increases by a factor rather less than the expected



Wet Adhesion in Tree and Torrent Frogs, Fig. 8 Allometric relationships of 13 hylid species (*open circles*) and two nonhylids (*filled triangles*) plotted against snout-vent length (SVL). (**a**, **b**) Morphometric relationships between frog mass, toe pad area, and length (log-log plots); mass $\propto$ SVL$^{2.65}$, significantly different from expected value of 3.0 ($p < 0.01$), and area $\propto$ SVL$^{1.85}$, not significantly different from expected value of 2.0. (**c**, **d**) Adhesive force and force per unit area plotted against length (log-log plots); force $\propto$ SVL$^{2.24}$, not significantly different from 2.0, the value expected from area scaling, and unit force $\propto$ SVL$^{0.46}$, a significant positive relationship ($p < 0.05$); (**e**) fall angle against length (log-linear plot). The regression line shows a significant negative relationship ($p < 0.01$) (Redrawn from Barnes et al. [22])

length cubed. Also, toe pad efficiency (force per unit area) increases with age, perhaps associated with the increased structural complexity of adult toe pads. However, in the majority of the species studied, adult frogs did not completely overcome the allometry problem, for they fell from the rotation platform at smaller angles than younger frogs. In ecological terms, this disadvantage is less than one might expect, in that frogs can use their digits to grasp objects such as twigs, as described above. There is even a positive correlation between adult size and degree of arboreality, in that the largest species are often found high in the canopy, while smaller species are most commonly found in shrubs no more than a meter or so above the ground. Finally, the way in which forces scale with size can provide insights into the underlying mechanisms of adhesion. Classic formulae for wet adhesion, based on flat metal disks (e.g., Eq. 4), show capillarity forces scaling with toe pad area and have therefore been used as evidence for the importance of capillarity in tree frog adhesion. However, at the small values of h measured by interference reflection microscopy, they predict adhesive forces several orders of magnitude greater than toe pads produce. Indeed, the softness of the pad material makes this model clearly inapplicable to tree frogs. Equation 2 predicts forces in the right range, but scaling with length and not area. Peeling models, applicable to the forces that prevent peeling in adhesive tape, are also worthy of consideration because of the way in which toe pads detach (see below). However, they also have force scaling with length, the forces being concentrated in a very narrow zone along the peeling line and thus dependent on the width of the tape. Probably, the solution lies in the recent analysis of Butt et al. [21] for adhesion using the model illustrated in Fig. 1. This analysis predicts a change from length to area scaling with increasing r and decreasing E_{eff}.

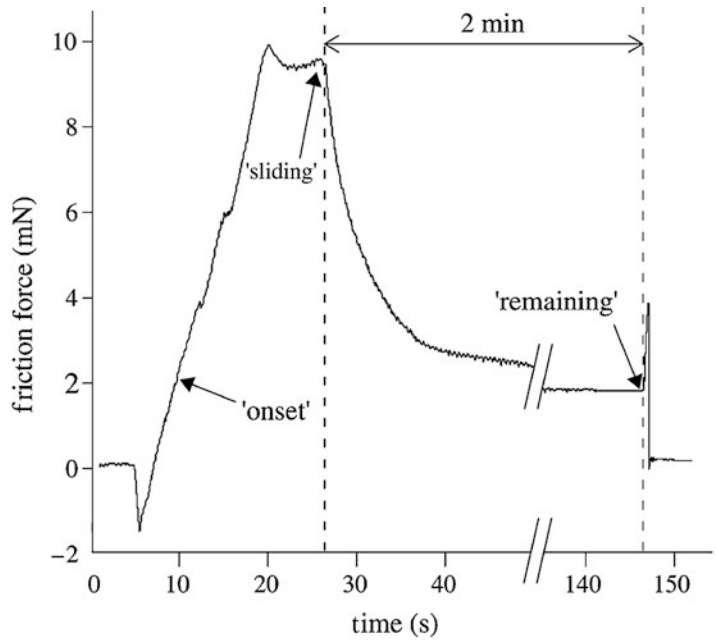
A common misconception is that the subdivision of the surface by the channels between the



Wet Adhesion in Tree and Torrent Frogs, Fig. 9 Experimental setup for measuring adhesion and friction forces from single toe pads of tree frogs. Frogs are held stationary in a small chamber (foam padding not shown), with a toe pad protruding through a hole in its base. Using a micromanipulator, a toe pad is brought into contact with a glass force plate attached to a 2D bending beam force transducer for measuring friction (shear) and adhesion (normal) forces. A computer-controlled XZ translational stage moves the frog in relation to the bending beam. Normal forces can be controlled by a feedback mechanism when required. Contact area is imaged from below using reflected light. Toe pad alignment is achieved by holding the backside of the toe by gentle suction coming from a soft rubber tubing attached to a syringe (From Crawford et al. [24])

cells increases adhesion according to the principle of contact splitting. This theory states that adhesive force is proportional to the length of the contact; therefore, by splitting up the contact zone into many small areas of contact, the total adhesive force can be increased by the square root of the density of these small areas [23]. This principle clearly applies to wet adhesion. However, as the major force component of wet adhesion is capillarity, it is necessary that an air-water interface (meniscus) should surround each small area of contact. This is true of the hairy pads of insects, but is not applicable to tree frogs, where the meniscus surrounds the whole toe pad, not the individual epithelial cells.

In recent years, the development of a technique to measure adhesion and friction forces from single toe pads (Fig. 9) has resulted in a number of interesting findings. The frog is held in a container with holes in its base through which individual digits can protrude. Toe pad forces are measured by a two-axis force plate, manipulated so that it makes contact with the protruding toe. As well as recording forces, the plate has a feedback facility so that, for instance, load can be kept constant while recording friction forces. This technique has provided good evidence for the presence of static friction in tree frog toe pads [18]. This is demonstrated both by the buildup of friction force at the onset of sliding and also from the presence of a remaining shear force 2 min after the sliding motion had stopped (Fig. 10). This is strongly indicative of actual contact of the nanopillars with the substrate, in accord with the findings of interference reflection microscopy (see above). By making measurements from pads of different sizes, one can also do scaling experiments with single pads. Such experiments demonstrate that friction forces are independent of load but correlate well with the area of contact. This is contrary to what one would expect of classic (Coulomb) friction, but in line with rubber friction. This conclusion fits well with the results of indentation experiments (see above), since the low elastic modulus of tree frog toe pads allows close contact between pad and substrate. Adhesion forces of single pads scale with pad area when pads are pulled off the force plate at a low angle



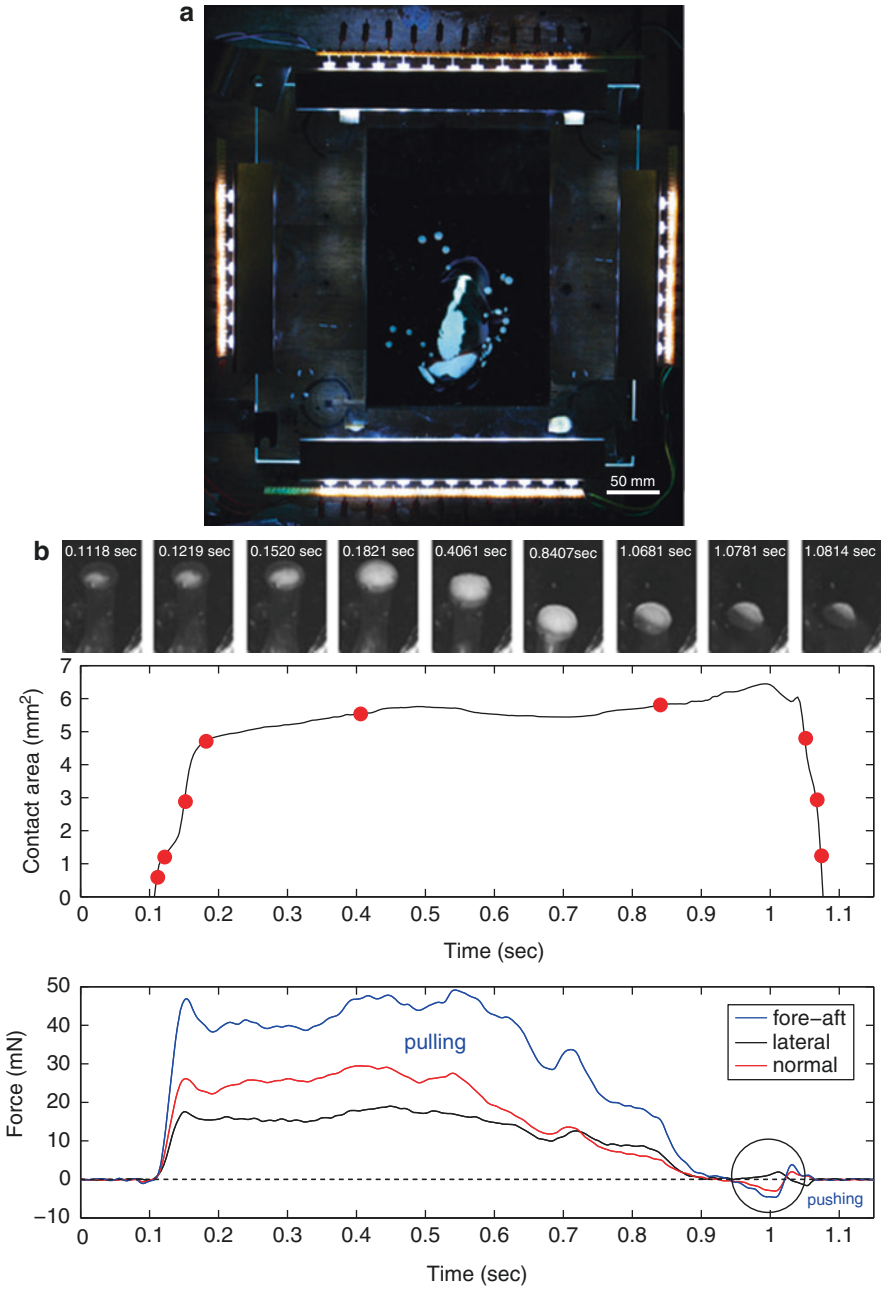
Wet Adhesion in Tree and Torrent Frogs, Fig. 10 Shear force measurement in single toe pads of White's tree frog *Litoria caerulea*. The recording shows the result of a 20 s sliding movement toward the body (500 μms^{-1}), followed by a 2 min period when no movement occurred. The *arrows* indicate the force when the pad began to slide ("onset"), the force when sliding ended ("sliding"), and the force when the experiment was terminated after the 2 min rest ("remaining") (Redrawn from Federle [18])

(20–50°), in line with whole animal measurements. However, they scale with length when the pull-off is at right angles to the surface. This has significance for toe pad detachment mechanisms (see below).

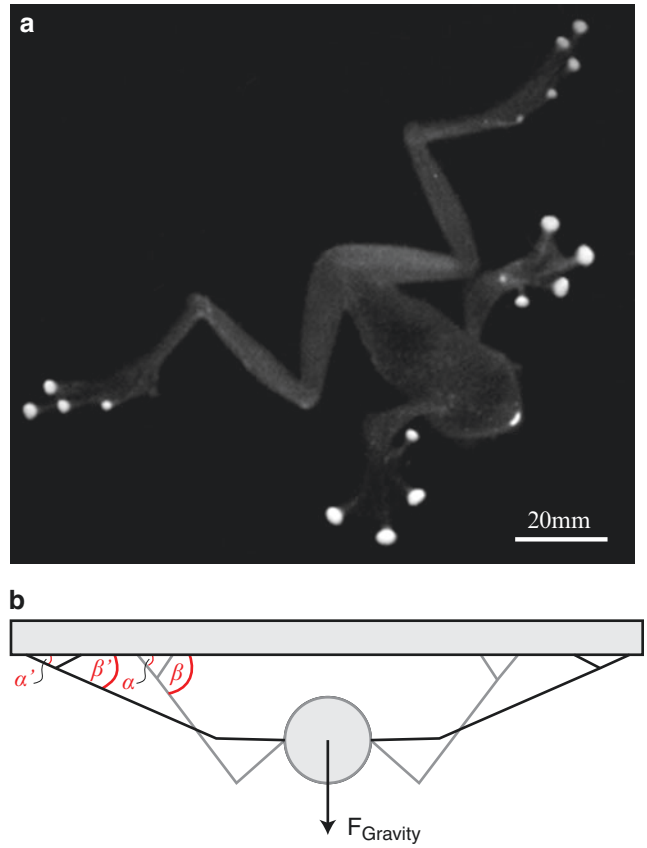
Finally, such single pad experiments provide evidence for the dominant role of capillarity in tree frog adhesion, since toe pad adhesive forces fall to low levels when the meniscus surrounding the pad is removed by submerging the whole pad in a drop of water. However, recent results from studies on rock frogs of the genus *Staurois* demonstrate that they are exceptionally good (in comparison to similarly sized tree frogs) at adhering to rough surfaces when they are completely flooded with water [25]. This is made possible by the frogs using large areas of their ventral surface (especially belly and thighs) in addition to their toe pads. This experiment clearly demonstrates a role for hydrodynamic forces (or possibly van der Waals forces resulting from direct contact), as the depth of the water will have abolished most of the air-water interfaces needed for capillarity.

Toe Pad Detachment

During locomotion, toe pads are detached by peeling, a mechanism that obviates the need for the frogs to pull their feet from the substrate every time they wish to jump or take a step [3]. During forward walking, climbing, or jumping, toe pads are peeled from the rear forwards, i.e., in a proximal to distal direction. Indeed, use of a force platform to measure directly the forces generated by the feet during locomotion shows that, during forward locomotion, this peeling is not accompanied by any significant detachment forces. It appears to occur as a natural consequence of the way in which toes are lifted off surfaces from the rear forwards during forward locomotion. This is supported both by observations of peeling in single toe pads of anesthetized frogs and by the inability of frogs to adhere to vertical surfaces in a head-down orientation. Indeed, frogs on a rotating vertical surface have been observed to adjust their orientations back toward a "head-up" orientation whenever their deviation from this orientation reached $85 \pm 21°$. The peeling process is seen most clearly when the pad is viewed from the ventral side using high-speed video imaging (Fig. 11b), recording both toe pad contact area and ground reaction forces. For instance, recordings during climbing on a vertical surface show that fore-aft forces (along the longitudinal snout-to-vent axis of the frog) drop to zero during the



Wet Adhesion in Tree and Torrent Frogs, Fig. 11 Toe pad dynamics of the tree frog *Litoria caerulea* climbing a vertical surface. (**a**) A glass plate is illuminated from the side by arrays of light-emitting diodes that direct the light within the thickness of the plate, the light only escaping (and therefore visible to the camera) when an object, here the frog, contacts the glass. (**b**) Change of a (*right*) forelimb pad contact area during a complete step (data acquisition rate 1,613 Hz, video rate 300 Hz) showing the attachment and detachment process by peeling (*red dots* correspond to the time stamps of the image sequence). Ground reaction forces (corrected for the alignment of the toe to the sensor) show a sign change at the beginning of the detachment process (encircled area). Unpublished data kindly provided by Diana Samuel, University of Glasgow



Wet Adhesion in Tree and Torrent Frogs, Fig. 12 (**a**) Video image of a ventral view of the tree frog (*Rhacophorus pardalis*) clinging to an inverted translucent surface, illuminated with internally reflected light (see Fig. 11a). Note how both fore- and hind limbs stretched out sideways. *Bright spots* are areas in contact with the surface. (**b**) Two-legged explanatory model. Both toes and legs are extended (from *gray* to *black* figure) to reduce the angles α (toe/substrate angle) and β (leg/substrate angle) to α' and β' and thus reduce the chances of peeling

detachment process and may even change in sign, indicating that the frog actively pushes outward to release any shear strain on the pads and thus facilitates peeling.

When frogs are induced to walk backward, peeling occurs in the opposite direction, from the front of the pad rearward in a distal to proximal direction. In this situation, peeling is an active process involving the distal tendons of the toes. Clear detachment forces can be recorded, as also occur when belly skin inadvertently comes into contact with the force platform.

Behavioral Strategies of Adhering Frogs

Interestingly, peeling has consequences for frog behavior. This is because pads can produce strong adhesion and friction when the angle of force application is low, but peel easily off the surface when the pull-off angle approaches 90°. On vertical and overhanging surfaces, frogs must therefore adjust their posture so that gravity acts on the pads at angles where they can withstand the force and not fall from the surface. On a rotation platform, the first response to the changing angle occurs at angles between 60° and 100°, heavier frogs tending to respond first. Any frog not already in a head-up position will turn to face head up, so that gravity will not be acting to pull the pads off the surface. At still higher angles, both tree and torrent frogs spread their legs out sideways (Fig. 12). This is an adaptation to prevent peeling, since, as the figure shows, spread out limbs have a much smaller pad-substrate angle than limbs held close to the body. Gravity, of course, will tend to pull the limbs inward, but this is resisted by pad friction forces. In this way, friction and adhesion both play roles in helping frogs to maintain their grip under challenging conditions. The term frictional adhesion has been coined to describe adhesion in geckos. It is unclear whether this term should also be applied to tree frog adhesion, but there is certainly

interplay between friction and adhesion forces in frog toe pads too.

A recent study by Endlein et al. [26] has explained the frog's spreading behavior by studying the underlying ground reaction forces during this behavior. By using an array of multiple force transducers embedded in a walkway, the authors were able to record the ground reaction forces of a frog clinging on to the platform at varying slopes. When the incline of the platform reached an overhanging position (at inclines past the vertical), the frogs spread their legs out sideward as described above. By using opposing legs, frogs pulled their legs inward which resulted in a dramatic increase in the lateral forces (forces perpendicular to the animals' longitudinal snout-to-vent axis). These large lateral forces caused the overall resulting force vector (the vectorial sum of all three force components) to adopt a very shallow angle toward the surface. As explained above, this can be compared to the peeling of an adhesive tape where a low angle of peel makes it very difficult to remove the tape from the surface, thus enhancing the adhesive capability of the frog. However, as the lateral forces increased, legs tended to slide slowly inward (toward the body). This in turn increased the risk of peeling (larger peeling angle), and frogs compensated for this by intermittently placing their feet outward again.

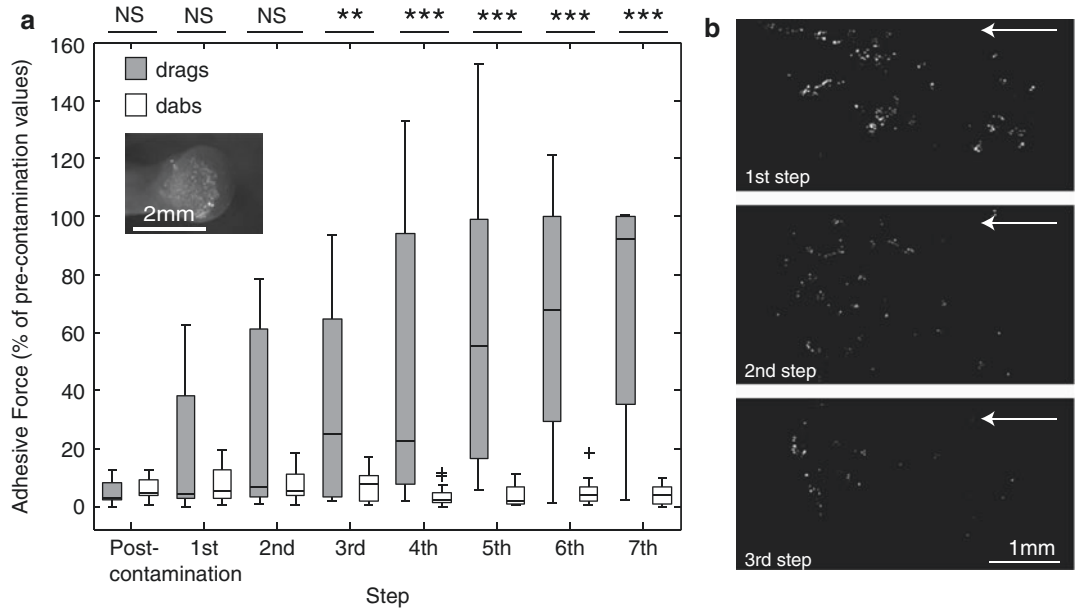
Self-Cleaning of Toe Pads

For any organism which uses adhesive pads, keeping them clean and contaminant-free is key for efficient climbing, as the presence of particles on the pad interface can have a detrimental effect on adhesion. However, the natural world is full of loose particles like dust, sand grains, or even plant coatings which have evolved to break off in order to clog up the adhesive pads and thus prevent the attachment of climbing animals. It was shown by Crawford et al. [24] that a self-cleaning mechanism (which works passively as the frog is climbing, rather than requiring grooming) exists in the toe pads of tree frogs which allows pads to become decontaminated quickly and therefore regain adhesive ability. This was studied by firstly challenging the frogs to cling on to increasingly steeper inclines using a tilting platform. After testing the performance of uncontaminated, "clean" toe pads, the pads were artificially contaminated using 50 μm glass beads and tested again for recovery of sticking ability. It was found that frogs which moved and used their pads more while climbing removed more contaminant and therefore recovered their sticking ability quicker. In order to understand the mechanism behind this recovery of adhesion, the authors measured the adhesive forces produced by a single toe pad using a force transducer mounted to a motorized stage. This way, controlled movements of the pads ("simulated steps") could be performed while measuring their adhesion. It was found that the inclusion of a short shear drag of the pad against the surface ("drag" vs. "dab"; Fig. 13a) before detachment drastically improved the detachment of the contaminant particles by rolling off. This was also aided by the adhesive fluid left behind on the surface with each step where contaminants would gather (Fig. 13b). Like geckos and insects, tree frogs' self-cleaning toe pads provide a novel trait of dynamic adhesives in nature which would be highly useful in man-made adhesives.

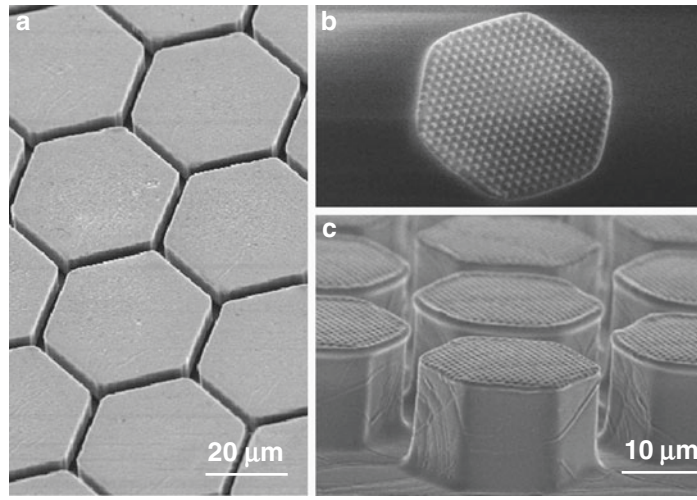
Biomimetic Advances

As far as tree frog adhesion is concerned, the process of micro- or nanofabrication of toe pad replicas, where the effect of the individual design parameters can be analyzed and quantified precisely, is in its infancy. An early attempt at copying the hexagonal cell structure of tree frog toe pads in polydimethylsiloxane (PDMS) is illustrated in Fig. 14a, whereas a recent mimic of structure of a tree frog's toe pad epithelium that replicates both nano- and microstructures, from del Campo's research group at the Max Planck Institute for Polymer Research in Mainz, Germany, is shown in Fig. 14b, c.

However, the role of surface patterning in increasing adhesion is well established. Adhesive tape detaches by the spreading of cracks along the interface from the point of peeling. When all the energy can be concentrated at a single crack, then peeling occurs readily, but micropatterning can



Wet Adhesion in Tree and Torrent Frogs, Fig. 13 Recovery of adhesive forces after contamination. (**a**) Adhesive force (normal component of pull-off force) as a percentage of the mean pre-contamination values for both the "drag" (*filled boxes*) and "dab" (*open boxes*) forms of the experiment. The *boxes* represent consecutive simulated "steps" taken by the toe pad on the force plate following pad contamination. Statistical tests were performed on each pair, with the results for each shown above the plots; *NS* not significant, *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$. *Inset*: contaminated toe pad. (**b**) Example of deposited beads after three consecutive steps (including a drag movement). Beads are deposited in characteristic "footprints" and frequently toward the end of the "drag" in clusters caused by the mucus. The *arrows* indicate the direction of the drag (Modified from Crawford et al. [24])



Wet Adhesion in Tree and Torrent Frogs, Fig. 14 (**a**) Polydimethylsiloxane (PDMS) replica of a tree frog toe pad, copying the hexagonal pattern of epithelial cells (Image: Centre for Cell Engineering, University of Glasgow). (**b**, **c**) Hierarchical PDMS replica of both micro- and nanostructuring on a tree frog toe pad. The PDMS was made hydrophilic by treatment in a plasma chamber and has an elastic modulus in the range 1–5 MPa (Image: Max Planck Institute for Polymer Research, Mainz (Germany))

increase the force required to produce peeling by up to a factor of three by the principle of contact separation. Cracks form wherever there is a groove in the pattern, so that the energy is spread between many cracks, resulting in an increased force being required to produce separation [14]. Recently, this principle has been taken a step further, and the role of subsurface structures (air- or oil-filled micro-channels) has been investigated. These have crack-arresting properties just like the transverse grooves of a patterned surface, but the effect is much more dramatic, for adhesion can be increased up to 30-fold. The adhesive itself remains elastic, which renders it reusable with no reduction in adhesive efficiency [27].

Turning from patterned dry adhesion to wet adhesion, some of the published work is insect rather than frog inspired. Of particular interest are studies of the frictional properties of a structure like that in Fig. 14a. Although the addition of fluid lowered friction as would be expected, the micropatterned surface was able to generate significantly higher frictional forces during sliding than an otherwise identical smooth one [28, 29], presumably by preventing aquaplaning. It has also been demonstrated that hydrophobic hexagonal patterns can be beneficial for adhesion under wet conditions, depending on the topographic design. For instance, mushroom-shaped pillars can produce a tenfold increase in adhesion underwater against flat controls [28, 30]. Whether this increase is due to a suction effect remains to be confirmed.

The effect of hydrodynamic forces on toe pad-inspired mimics was investigated recently by Shahsavan and Zhao [31] and Gupta and Fréchette [32]. Both studies confirmed a reduction of hydrodynamic force for squeezing out the fluid in the presence of channels, thus facilitating the close contact of a structured toe pad to the substrate.

Examples of Possible Applications

Since adhesive tapes inspired by the dry adhesive mechanisms of gecko toe pad setae are now reaching a stage where commercialization is imminent, it is appropriate to consider whether the rather different wet adhesion mechanism of tree and torrent frogs might also have biomimetic relevance. Improved wet weather tires [13, 33, 34], nonslip footwear, plasters for surgery able to adhere to tissue, and holding devices for neurosurgery or MEMS devices are obvious examples of the many uses to which these toe pad analogues might be applied. However, as biomimetics is as much to do with inspiration as copying, a beetle-inspired device providing reversible adhesion based on capillarity is worthy of mention as it utilizes wet adhesion [35]. It generates an adhesive force through the formation of a large number of liquid bridges between two plates, which can be quickly made or broken by a low-voltage pulse that drives electroosmotic flow. There is good potential for the development of a grab-and-release device that could carry significant loads.

A recent study by Tsipenyuk and Varenberg [36] showed the advantage of using structured surfaces (hexagonal patterns inspired by tree frogs and some insects) for disposable safety razors. In their study, the authors replaced the standard stretching unit of the razors with bioinspired ones and reported significantly enhanced friction forces against human skin. This was due to more effective channeling of the lubricant fluid out of the contact zone.

Cross-References

- ▶ Adhesion in Insects
- ▶ Atomic Force Microscopy
- ▶ Bioadhesion
- ▶ Biomimetics
- ▶ Gecko Adhesion
- ▶ Optical Tweezers
- ▶ Scanning Electron Microscopy

References

1. Autumn, K., Sitti, M., Liang, Y.A., Peattie, A.M., Hansen, W.R., Sponberg, S., Kenny, T.W., Fearing, R., Israelachvili, J.N., Full, R.J.: Evidence for van der Waals adhesion in gecko setae. Proc. Natl. Acad. Sci. U. S. A. **99**(19), 12252–12256 (2002). doi:10.1073/pnas.192252799

2. Federle, W., Riehle, M., Curtis, A., Full, R.: An integrative study of insect adhesion: mechanics and wet adhesion of pretarsal pads in ants. Integr. Comp. Biol. **42**, 1100–1106 (2002). doi:10.1093/icb/42.6.1100
3. Hanna, G., Barnes, W.J.P.: Adhesion and detachment of the toe pads of tree frogs. J. Exp. Biol. **155**, 103–125 (1991)
4. Rosenberg, H.I., Rose, R.: Volar adhesive pads of the feathertail glider, *Acrobates pygmaeus* (Marsupialia; Acrobatidae). Can. J. Zool. **77**, 1–16 (1999). doi:10.1139/z98-212
5. Riskin, D.K., Racey, P.A.: How do sucker-footed bats hold on, and why do they roost head-up? Biol. J. Linn. Soc. **99**, 233–240 (2010). doi:10.1111/j.1095-8312.2009.01362.x
6. Barnes, W.J.P.: Functional morphology and design constraints of smooth adhesive pads. MRS Bull. **32**, 479–485 (2007). doi:10.1557/mrs2007.81
7. Peattie, A.M., Dirks, J.-H., Henriques, S., Federle, W.: Arachnids secrete a fluid over their adhesive pads. PLoS One **6**(5), e20485 (2011). doi:10.1371/journal.pone.0020485
8. Jiao, Y., Gorb, S., Scherge, M.: Adhesion measured on the attachment pads of *Tettigonia viridissima* (Orthoptera, Insecta). J. Exp. Biol. **203**, 1887–1895 (2000)
9. Federle, W., Barnes, W.J.P., Baumgartner, W., Drechsler, P., Smith, J.: Wet but not slippery: boundary friction in tree frog adhesive toe pads. J. R. Soc. Interface **3**(10), 689–697 (2006). doi:10.1098/rsif.2006.0135
10. Green, D.M.: Treefrog toe pads: comparative surface morphology using scanning electron microscopy. Can. J. Zool. **57**, 2033–2046 (1979). doi:10.1139/z79-268
11. Emerson, S., Diehl, D.: Toe pad morphology and mechanisms of sticking in frogs. Biol. J. Linn. Soc. **13**(3), 199–216 (1980). doi:10.1111/j.1095-8312.1980.tb00082.x
12. Smith, J.M., Barnes, W.J.P., Downie, J.R., Ruxton, G. D.: Structural correlates of increased adhesive efficiency with adult size in the toe pads of hylid tree frogs. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. **192**(11), 1193–1204 (2006). doi:10.1007/ s00359-006-0151-4
13. Persson, B.N.J.: Wet adhesion with application to tree frog adhesive toe pads and tires. J. Phys. Condens. Matter **19**(37), 376110 (2007). doi:10.1088/0953-8984/19/37/376110
14. Ghatak, A., Mahadevan, L., Chung, J., Chaudhury, M., Shenoy, V.: Peeling from a biomimetically patterned thin elastic film. Proc. R. Soc. A Math. Phys. Eng. Sci. **460**(2049), 2725–2735 (2004). doi:10.1098/rspa.2004.1313
15. Scholz, I., Barnes, W.J.P., Smith, J.M., Baumgartner, W.: Ultrastructure and physical properties of an adhesive surface, the toe pad epithelium of the tree frog, *Litoria caerulea* White. J. Exp. Biol. **212**(2), 155–162 (2009). doi:10.1242/jeb.019232
16. Green, D.M., Alberch, P.: Interdigital webbing and skin morphology in the neotropical salamander genus *Bolitoglossa*. J. Morphol. **170**, 273–282 (1981). doi:10.1002/jmor.1051700302
17. Barnes, W.J.P., Goodwyn, P., Nokhbatolfoghahai, M., Gorb, S.: Elastic modulus of tree frog adhesive toe pads. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. **197**, 969–978 (2011). doi:10.1007/s00359-011-0658-1
18. Federle, W.: Why are so many adhesive pads hairy? J. Exp. Biol. **209**(14), 2611–2621 (2006). doi:10.1242/jeb.02323
19. Smith, J.M., Barnes, W.J.P., Downie, J.R., Ruxton, G. D.: Adhesion and allometry from metamorphosis to maturation in hylid tree frogs: a sticky problem. J. Zool. **270**(2), 372–383 (2006). doi:10.1111/j.1469-7998.2006.00145.x
20. Barnes, W.J.P.: Dynamic adhesion in animals: mechanisms and biomimetic implications. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. **192**(11), 1165–1168 (2006). doi:10.1007/s00359-006-0144-3
21. Butt, H.-J., Barnes, W.J.P., del Campo, A., Kappl, M., Schonfeld, F.: Capillary forces between soft, elastic spheres. Soft Matter **6**, 5930–5936 (2010). doi:10.1039/C0SM00455C
22. Barnes, W.J.P., Oines, C., Smith, J.M.: Whole animal measurements of shear and adhesive forces in adult tree frogs: insights into underlying mechanisms of adhesion obtained from studying the effects of size and scale. J. Comp. Physiol. A Neuroethol. Sens. Neural Behav. Physiol. **192**(11), 1179–1191 (2006). doi:10.1007/s00359-006-0146-1
23. Arzt, E., Gorb, S., Spolenak, R.: From micro to nano contacts in biological attachment devices. Proc. Natl. Acad. Sci. **100**(19), 10603–10606 (2003). doi:10.1073/pnas.1534701100
24. Crawford, N., Endlein, T., Barnes, W.J.P.: Self-cleaning in tree frog toe pads; a mechanism for recovering from contamination without the need for grooming. J. Exp. Biol. **215**, 3965–3972 (2012). doi:10.1242/jeb.073809
25. Endlein, T., Barnes, W.J.P., Samuel, D.S., Crawford, N.A., Biaw, A.B.., Grafe, U.: Sticking under wet conditions: the remarkable attachment abilities of the torrent frog, *Staurois guttatus*. PLoS One **8**(9), e73810 (2013). doi:10.1371/journal.pone.0073810
26. Endlein, T., Ji, A., Samuel, D., Yao, N., Wang, Z., Barnes, W.J.P., Federle, W., Kappl, M., Dai, Z.: Sticking like sticky tape: tree frogs use friction forces to enhance attachment on overhanging surfaces. J. R. Soc. Interface **10**(80), 20120838 (2013). doi:10.1098/rsif.2012.0838
27. Majumder, A., Ghatak, A., Sharma, A.: Microfluidic adhesion induced by subsurface microstructures. Science **318**(5848), 258–261 (2007). doi:10.1126/science.1145839
28. Drotlef, D.-M., Stepien, L., Kappl, M., Barnes, W.J.P., Butt, H.-J., del Campo, A.: Insights into the adhesive mechanisms of tree-frogs using artificial mimics. Adv.

Funct. Mater. **9**, 1094 (2012). doi:10.1002/adfm.201202024
29. Varenberg, M., Gorb, S.N.: Hexagonal surface micropattern for dry and wet friction. Adv. Mater. **21**(4), 483–486 (2009). doi:10.1002/adma.200802734
30. Varenberg, M., Gorb, S.: Shearing of fibrillar adhesive microstructure: friction and shear-related changes in pull-off force. J. R. Soc. Interface **4**(15), 721–725 (2007). doi:10.1098/rsif.2007. 0222
31. Shahsavan, H., Zhao, B.: Bioinspired functionally graded adhesive materials: synergetic interplay of top viscous–elastic layers with base micropillars. Macromolecules **47**(1), 353–364 (2014). doi:10.1021/ma4018718
32. Gupta, R., Fréchette, J.: Measurement and scaling of hydrodynamic interactions in the presence of draining channels. Langmuir **28**(41), 14703–14712 (2012). doi:10.1021/la303508x
33. Barnes, W.J.P.: Tree frogs and tire technology. Tire. Technol. Int. **99**, 42–47 (1999)
34. Barnes, J., Smith, J., Oines, C., Mundl, R.: Bionics and wet grip. Tire. Technol. Int. 56–60 (2002)
35. Vogel, M.J., Steen, P.H.: Capillarity-based switchable adhesion. Proc. Natl. Acad. Sci. **107**(8), 3377–3381 (2010). doi:10.1073/pnas.0914720107
36. Tsipenyuk, A., Varenberg, M.: Use of biomimetic hexagonal surface texture in friction against lubricated skin. J. R. Soc. Interface. **11**(94), (2014). doi: 10.1098/rsif.2014.0113

Wet Chemical and Electrochemical Etching Processes

Mathew Ombaba[1], Salman B. Inayat[1] and M. Saif Islam[1,2]
[1]University of California, Davis, CA, USA
[2]Electrical and Computer Engineering, University of California – Davis Integrated Nanodevices and Nanosystems Lab, Davis, CA, USA

Isotropy and Anisotropy

When a material is attacked by a liquid or vapor etchant, the material will be removed isotropically (uniformly in all directions) or through anisotropic etching (uniformity in vertical direction). The difference between isotropic etching and anisotropic etching is shown in Fig. 1. Material removal rate for wet etching is usually faster than the rates for many dry etching processes and can easily be changed by varying temperature or the concentration of active species.

Wet Etch

Synonyms

Chemical etching; Liquid etching

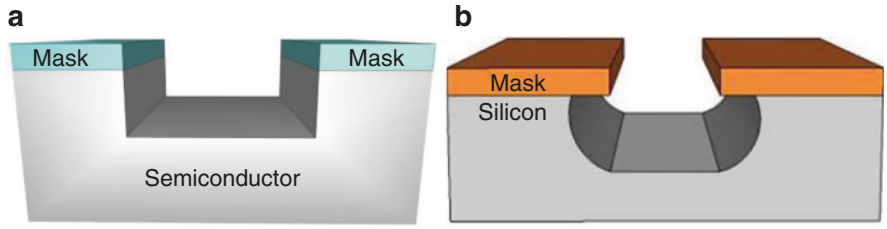
Definition

Wet etching is a material removal process that uses liquid chemicals or etchants to remove materials from a wafer. The specific patters are defined by photoresist masks on the wafer. Materials that are not protected by this mask are etched away by liquid chemicals. These masks are deposited on the wafer in an earlier fabrication step called lithography.

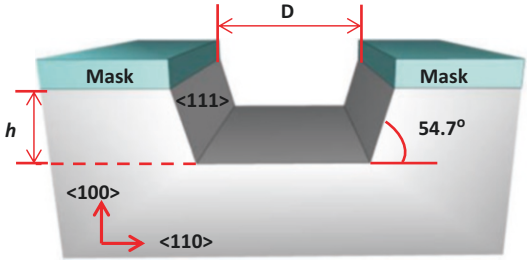
A wet etching process involves multiple chemical reactions that consume the original reactants and produce new reactants. The wet etch process can be described by three basic steps (1) diffusion of the liquid etchant to the structure that is to be removed and (2) the reaction between the liquid etchant and the material being etched away. A reduction-oxidation (redox) reaction usually occurs. This reaction entails the oxidation of the material then dissolving the oxidized material. (3) Diffusion of the byproducts in the reaction from the reacted surface.

Anisotropic Wet Etching

Liquid etchants etch crystalline materials at different rates depending upon which crystal face is exposed to the etchant. There is a large difference in the etch rate depending on the silicon crystalline plane. In materials such as silicon, this effect can allow for very high anisotropy [1]. Some of the anisotropic wet etching agents for silicon are potassium hydroxide (KOH), ethylenediamine pyrocatechol (EDP), or tetramethylammonium hydroxide (TMAH). Etching a (100) silicon wafer would result in a pyramid-shaped etch pit as shown in Fig. 2. The etched wall will be flat and



Wet Chemical and Electrochemical Etching Processes, Fig. 1 (**a**) Completely anisotropic etching of a semiconductor (**b**) Isotropic etching where the semiconductor is etched all around uniformly if not masked



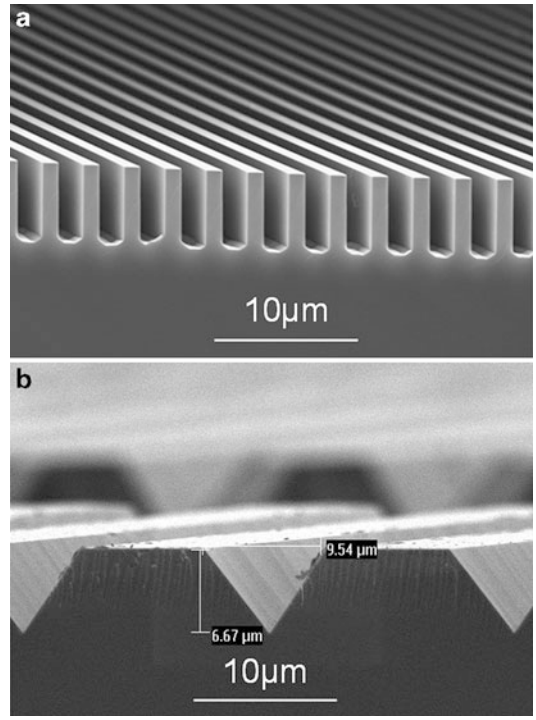
Wet Chemical and Electrochemical Etching Processes, Fig. 2 Example of an etch profile in anisotropic of (100) silicon wafer. The etchant in this case is aqueous KOH that etches the unmasked areas of the silicon wafer till it encounters the <111> plane

angled. The angle to the surface of the wafer is 54.7°.

The relationship between mask dimensions, etch depth and the floor width is given in Eq. 1.

$$d = D - \left(\frac{2h}{\tan(54.7^\circ)}\right) \qquad (1)$$

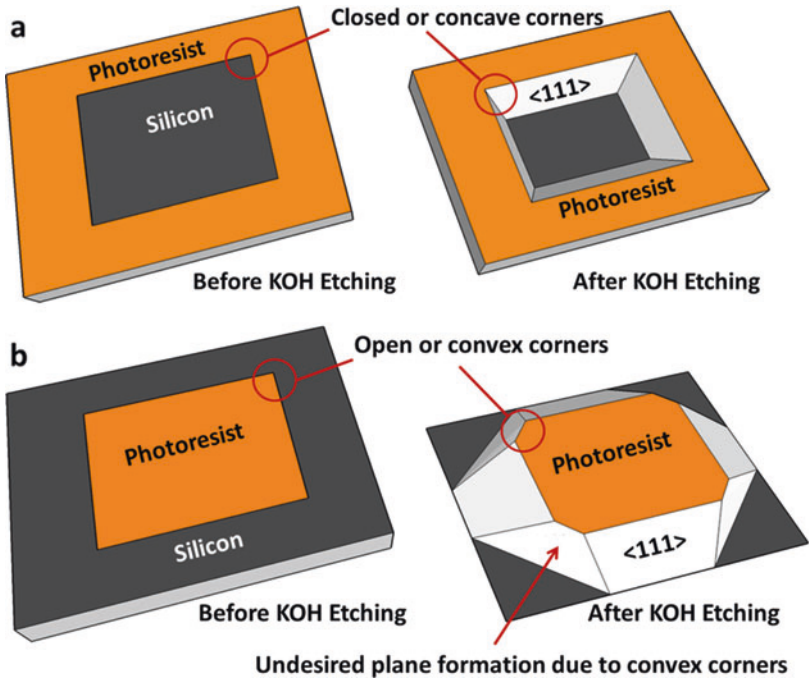
On the other hand, the orientation of the 111 plane on <110> oriented silicon lies horizontally to its surface at exactly 90°. Therefore, KOH etching of the <110> oriented silicon wafer results in perfectly anisotropic trenches as depicted in the scanning electrode microscopy picture in Fig. 3a unlike.



Wet Chemical and Electrochemical Etching Processes, Fig. 3 Anisotropic etching of (**a**) <110> and (**b**) <100> oriented Si in KOH solution

Problems Related to Anisotropic KOH Etching

Structures with open corners commonly termed as convex corners loose the integrity of the corners during KOH etching. The widely accepted understanding describes that the higher-order planes confining these corners tend to etch faster resulting in undesirable corner formation [2]. In order to compensate for the corner undercutting for open ends in KOH, compensation structures are added to these open corners. However, the space consumed by the compensation structures is a major impediment toward the use of the corner compensation concept. In extreme cases, compensation structures twice as wide as the etch depth targeted with KOH etching may be required [3]. For through wafer KOH etching, the use of larger

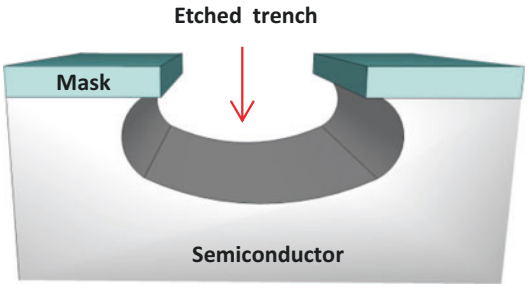


Wet Chemical and Electrochemical Etching Processes, Fig. 4 Convex corner undercutting with KOH etching

and more efficient compensation structures is needed. Design and fabrication of space-efficient compensation structures to provide sharp resultant corners after KOH etching is an actively pursued research topic in the field of silicon processing [4] (Fig. 4).

Isotropic Wet Etching

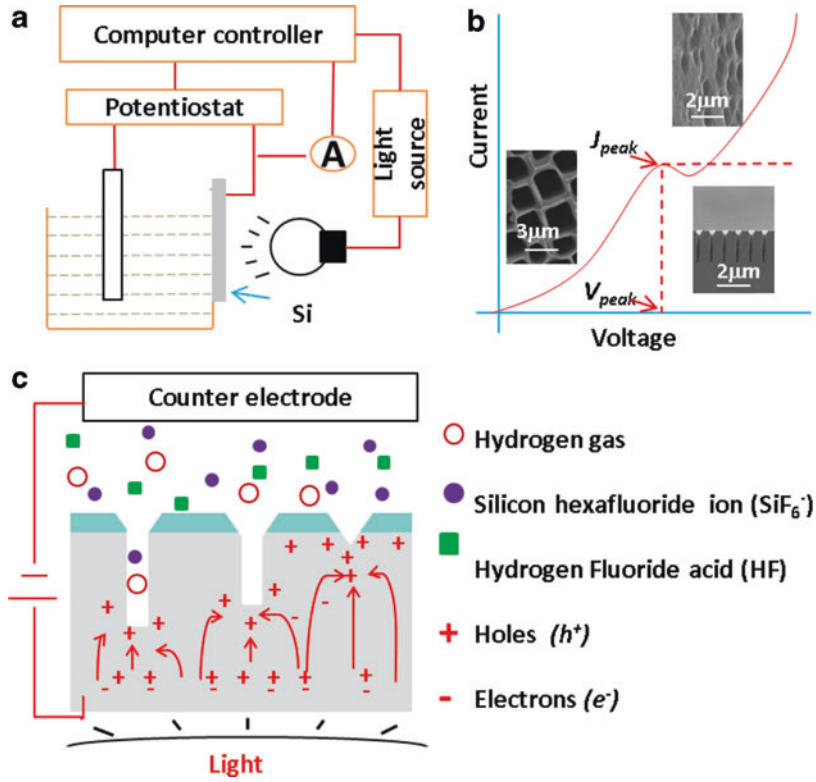
For isotropic wet etching, a mixture of hydrofluoric acid, nitric acid, and acetic acid (HNA) is the most common etchant solvent for silicon. The concentrations of each etchant determines the etch rate. Silicon dioxide or silicon nitride is usually used as a masking material against HNA. As the reaction takes place, the material is removed laterally at a rate similar to the speed of etching downward. This lateral and downward etching process takes places even with isotropic dry etching which is described in the dry etch section. The isotropic wet etch is shown in Fig. 5. Wet chemical etching is generally isotropic even though a mask is present since the liquid etchant can penetrate underneath the mask. Since directionality is usually important for high-resolution pattern transfer, wet chemical etching is normally not used.



Wet Chemical and Electrochemical Etching Processes, Fig. 5 Isotropic etching of masked semiconductor

Electrochemical Etching Techniques

Apart from chemical approaches, semiconductors and even other materials such as metals can be etched by application of an external electrical current. Here, a potentiostat or galvanostat provides the electrical charge required to perform the etching [5]. It further provides a means of



Wet Chemical and Electrochemical Etching Processes, Fig. 6 (**a**) The setup for photoelectrochemical of *n*-type semiconductors such as silicon. For *p*-type silicon, the light source and lamp is not included. (**b**) The current versus voltage characteristics of a *p*-type and strongly illuminated n-type silicon wafer under anodic biasing. The SEM images show etching characteristics of the semiconductor surface at various potentials. Notice that at voltages above *Vpeak* and current bellow J_{peak}, good control of anisotropic etching can be achieved. (**c**) The etching mechanism of an *n*-type silicon. Holes are generated at the backside of the wafer. Due to the applied potential, an electric field created therein causes the holes to travel to the surface of the wafer which is in contact to the electrolyte where a reaction leads to dissolution of silicon

controlling the etching process and profiles of the features to be etched through increasing or decreasing either the etching current or potential as desired. In a typical setup for such an etching process, a semiconductor to be etched is arraigned in an electrochemical cell with the surface of the semiconductor to be etched exposed to an electrolyte as shown in Fig. 6a. The electrolyte is composed of a liquid and ions that are needed to conduct electrical charge as well as dissolve the semiconductor upon application of the required electrical current. The back of the semiconductor is connected to the anode of the power supply [6]. Successful electrochemical etching is dependent on several factors such as:

(i) Type of semiconductor
(ii) Doping density (resistivity) of the semiconductor
(iii) Type of doping (*n* or *p*) doped

Although theoretically any semiconductor can be anodically dissolved in an acidic electrolyte which contains F^- ions if holes (h^+) are available, this condition is not met for *n*-doped semiconductors since it does not have holes. This can be achieved by illuminating the sample from the back as illustrated in Fig. 6a. So far, most developed recipes have been using silicon wafers whose resistivity lies in the range of 1–30 Ω.cm. The etching mechanism for dissolution of silicon

has been established. In the presence of holes at a surface, hydrofluoric acid dissolves silicon according to the reaction as shown in Eq. 2 below [5].

$$Si_{(s)} + 6HF_{(aq)} + 2h^{+} -> SiF_6{}^{2-}{}_{(aq)} + 4H^{+}{}_{(l)} + H_{2\,(g)} \tag{2}$$

where e^{-} and h^{+} denote an electron and hole respectively. The current voltage relations of either *p*-type or illuminated n-type silicon under anodic biasing are shown in Fig. 6b. For current densities above J_{peak}, electropolishing (uniform surface etching) occurs, while below that, the semiconductor surface forms a porous layer. However, at voltages higher than V_{peak} and current density below J_{peak}, successful control of anisotropicity can be achieved [7].

For anisotropic dissolution of *n*-type silicon to form deep trenches or holes, a silicon wafer is first patterned with KOH to form reverse pyramids or "V" grooves on the surface of silicon like those depicted in Fig. 3b. Upon illumination of the wafer from the back while under anodic biasing, generated holes travel to the opposite surface of the wafer. The holes preferentially accumulate at the trench depth due to the electric field generated from the applied potential. Once at the trench depth, the reaction depicted in Eq. 2 occurs. The number of holes generated can be controlled by varying the intensity of the light. This can have an effect of modulating the reaction rate as well as the anisotropicity of the generated holes or trenches. This is depicted in Fig. 6c. The same process can occur in *p*-type wafer since the charge carriers therein are holes [8]. However, light has no effect in that case. Holes with uniform and nonuniform diameters can be easily generated by maintaining the current/voltage ratio over time as shown in Fig. 7a. However, by modulating the applied current over time, trench or hole diameters can be changed as depicted in Fig. 7b, c.

Rather than appling potential anodically on a semiconductor to be etched, holding the anodizing electrode next to the electrode can catalyze the dissolution of the semiconductor just below the electrode [9]. As long as this contact is maintained, etching proceeds. Anisotropy is achievable by applying pressure on the anodizing electrode. The electrolyte to be used is dependent on the semiconductor to be used. For silicon, an aqueous 30 M HF electrolytic solution was used. Anisotropic etching of SiC on the other hand uses sulfuric acid in combination with hydrofluoric acid [10]. Just like the noncontact electrochemical etching, dissolution of Si via contact anodization is achieved using a similar mechanism as the noncontact dissolution. Hence, Eq. 3 is still valid. However, the holes are injected into the semiconductor from the electrode as depicted in Fig. 8.

Metal Assisted Chemical Etching

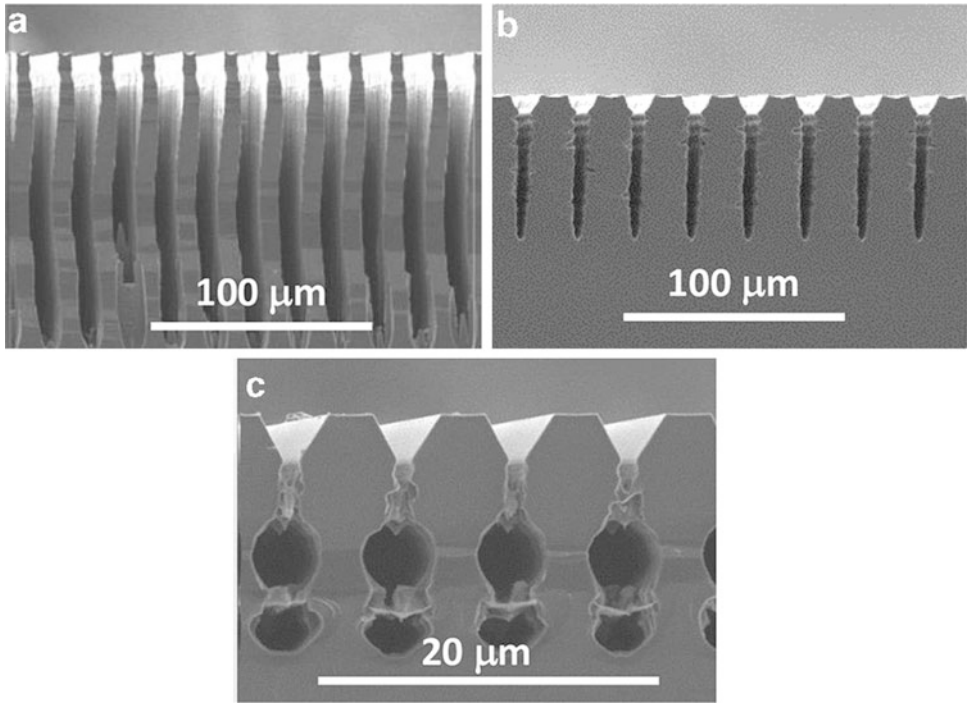
In this technique, a semiconductor substrate is selectively etched in an aqueous solution of HF with the help of noble metal in the presence of a strong oxidizing agent like peroxide in the normal plane. The reaction mechanism can be summarized in Eqs. 3 and 4. Here, the oxidant (hydrogen peroxide) first oxidizes the metal catalyst (Eq. 3). The oxidized metal then reacts with the semiconductor (silicon) in the presence of the acidic ions to produce soluble silicon hexafluoride ion while the metal is reduced to its metallic state as shown in Eq. 4 [11, 12].

$$H_2O_{2(aq)} + 2H^{+}{}_{(aq)} + Ag_{(s)} <- > 2H_2O_{(l)} + 2Ag^{+}{}_{(Aq)} \tag{3}$$

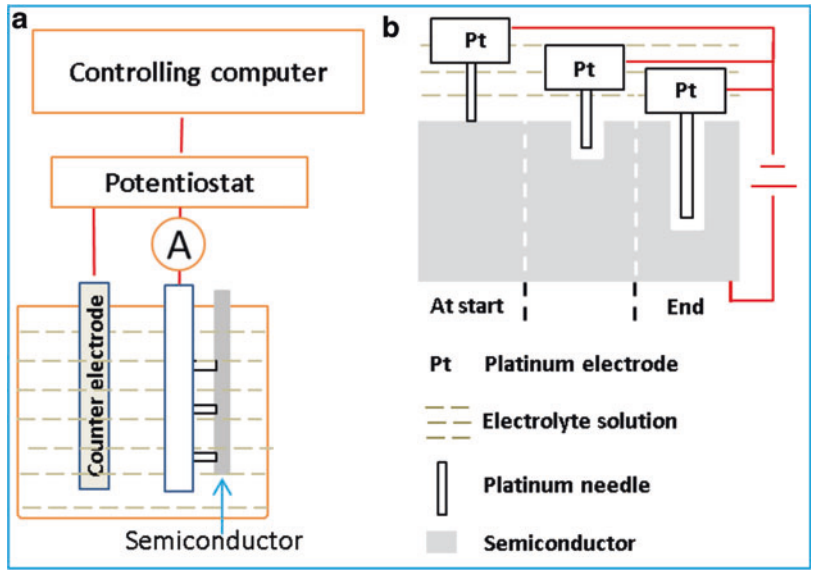
$$4Ag^{+}{}_{(aq)} + Si + 6F^{-}{}_{(aq)} <- > 4Ag_{(s)} + SiF_6{}^{2}{}_{-(aq)} \tag{4}$$

The first reaction is known to be thermodynamically favorable and stops with the depletion of the peroxide around the metallic surfaces which effectively decreases the etching rate for *n*-type silicon given that Ag^{+} is pertinent in Si etching.

High aspect ratio and uniform trenches and holes with diameters as low as 20 nm can be generated. Etching progresses as long as the catalytic film is in contact with the semiconductor. This is achieved gravitationally. Overly wide catalyst films tend to slow the etching process and sometimes may cause nonuniform features. Some of the large structures obtained via this method in our laboratory are depicted in Fig. 9.



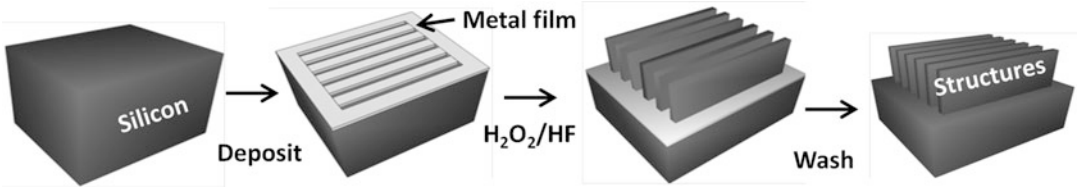
Wet Chemical and Electrochemical Etching Processes, Fig. 7 Deep trenches on *n*-type silicon obtained by photoelectrochemical etching under (**a**) constant voltage and potential, (**b**, **c**) modulated current



Wet Chemical and Electrochemical Etching Processes, Fig. 8 (**a**) The setup for contact anodic electrochemical etching of semiconductors. The anodizing potential is applied directly on the surface of the semiconductor directly with an electrode. (**b**) A schematic depicting the progress of the etching process. Over time, the anodizing electrode burrows into the semiconductor

Magnetically Guided Metal Assisted Chemical Etching

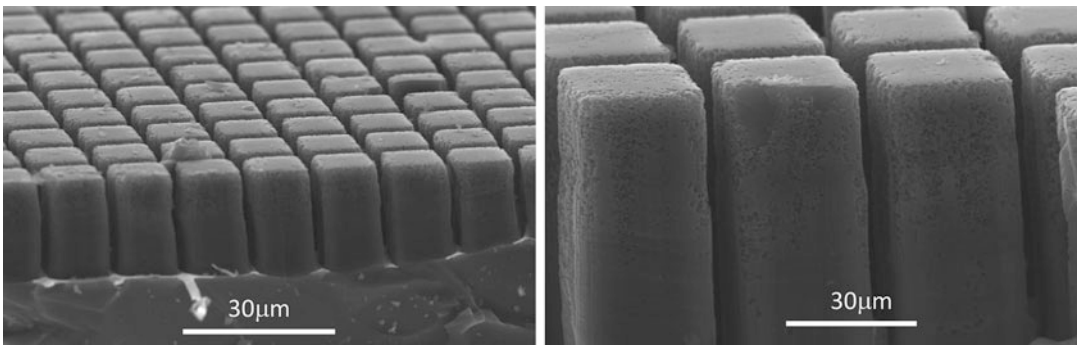
Here, an additional magnetic metal is incorporated into the etchant metal film such that it can respond to external magnetic stimuli so as to guide and manipulate the etching process [13]. The etching mechanism as described in Eqs. 3 and 4 is still valid. The additional magnetic field can



Wet Chemical and Electrochemical Etching Processes, Fig. 9 Schematic of metal-assisted chemical etching process. In the magnetically assisted version, a ferrous metal is sandwiched inside the catalytic noble metal film

Wet Chemical and Electrochemical Etching Processes, Table 1 Comparison of semiconductor etching systems other than the traditional CMOS adapted ones

	Method	Advantages
Photoelectrochemical etching of n-type silicon	Strong source of light N-doped silicon (1–30 Ω.com) 2–5 % HF solution Potentiostat Platinum reference and counterelectrode required	High aspect ratio structures achieved Good control of the anisotropic etch Low acidity used High etch rate (up to 10 μm/min) Accommodated wider range of substrate resistivity (1–30 Ω.com)
Electrochemical etching of p-type silicon	Strong source light N-doped silicon (20 Ω.com) 2–5 % HF solution Potentiostat Platinum reference and counterelectrode required	High aspect ratio structures achieve Low acidity used No light requirement High etch rate (up to 10 μm/min)
Contact anodical etching of semiconductor	28 % HF solution Nobel metal to be in contact with substrate	No light requirement Etch rate of up to 3 μm/min Can produce structures at acute angles
Metal-assisted electroless etching	5 M HF, 0.5 M H_2O_2electrolyte required Nobel metal pattering High etch rate	High aspect ratio structure achieve No light requirement Etch rate approaching 2 μm/min
Magnetically controlled metal-assisted electroless etching	5 M HF, 0.5 M HF, H_2O_2electrolye required Nobel metal and ferrous metal sandwich patterning Magnet manipulation require Low etch rate	Infinite possibilities on 3-dimensional etch profiles The etch rate can be modulated by the strength of the magnet Can produce structures at any angle Has no restriction on the crystal orientation of the wafer



Wet Chemical and Electrochemical Etching Processes, Fig. 10 SEM images silicon micropillars prepared by metal-assisted chemical etching

potentially be manipulated in such a way that intricate three-dimensional features that cannot be obtained via any other etching method are realized. A key impediment however lies with the incorporation of the ferrous metal onto the metallic etchant which involved tandem deposition of metallic trilayers in which the latter sandwiches the former for protection from the strong reducing etching solution. Additionally, unlike electrical field which is straight, the shape of the magnetic field often changes with distance hence it is difficult to exclusively manipulate the metallic etching system. The resolution of the resultant three-dimensional features is therefore a little low. Table 1 is a summary of these etching techniques (Fig. 10).

References

1. Radjenovic, B., Radmilovic-Radjenovic, M., Mitric, M.: Level set approach to anisotropic wet etching of silicon. Sensors (Basel) **10**, 4950–4967 (2010)
2. Elwenspoek, N.: Silicon Micromachining. Cambridge University Press, Cambridge, UK (1998)
3. Mayer, G.K., Offereins, H.L., Sandmaier, H., Kühl, K.: Fabrication of non-underetched convex corners in anisotropic etching of (100)-silicon in aqueous KOH with respect to novel micromechanic elements. J. Electrochem. Soc. **137**, 3947–3951 (1990)
4. Wacogne, B., Sadani, Z., Gharbi, T.: Compensation structures for V-grooves connected to square apertures in KOH-etched (1 0 0) silicon: theory, simulation and experimentation. Sensors Actuators A Phys. **112**, 328–339 (2004)
5. Lehmann, V., Föll, H.: Formation mechanism and properties of electrochemically etched trenches in n-type silicon. J. Electrochem. Soc. **137**, 653–659 (1990)
6. Bassu, M., Surdo, S., Strambini, L.M., Barillaro, G.: Electrochemical micromachining as an enabling technology for advanced silicon microstructuring. Adv. Funct. Mater. **22**, 1222–1228 (2012)
7. Lehmann, V.: The physics of macropore formation in low doped n-type silicon. J. Electrochem. Soc. **140**, 2836–2843 (1993)
8. Janshoff, A., Dancil, K.-P.S., Steinem, C., Greiner, D. P., Lin, V.S.Y., Gurtner, C., et al.: Macroporous p-type silicon Fabry – Perot layers. Fabrication, characterization, and applications in biosensing. J. Am. Chem. Soc. **120**, 12108–12116 (1998). 1998/11/01
9. Lee, C.-L., Kanda, Y., Hirai, T., Ikeda, S., Matsumura, M.: Electrochemical grooving of Si wafers using catalytic wire electrodes in HF solution. J. Electrochem. Soc. **156**, H134–H137 (2009)
10. Sugita, T., Hiramatsu, K., Ikeda, S., Matsumura, M.: Fabrication of pores in a silicon carbide wafer by electrochemical etching with a glassy-carbon needle electrode. ACS Appl. Mater. Interfaces **5**, 2580–2584 (2013)
11. Peng, K.Q., Xu, Y., Wu, Y., Yan, Y.J., Lee, S.T., Zhu, J.: Aligned single-crystalline Si nanowire arrays for photovoltaic applications. Small **1**, 1062–1067 (2005)
12. Peng, K.Q., Yan, Y.J., Gao, S.P., Zhu, J.: Dendrite-assisted growth of silicon nanowires in electroless metal deposition. Adv. Funct. Mater. **13**, 127–132 (2003)
13. Oh, Y., Choi, C., Hong, D., Kong, S.D., Jin, S.: Magnetically guided nano-micro shaping and slicing of silicon. Nano Lett. **12**, 2045–2050 (2012)

Wet Chemical Processing

▶ Sol–Gel Method

Wet Chemical Synthesis

▶ Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods

Wetting Transitions

Edward Bormashenko
Laboratory of Polymers, Physics Department, Ariel University, Ariel, Israel

Definition

The wetting transition is an abrupt change, spontaneous or induced by external stimulus, in the wetting properties of a flat or rough solid surface.

Wetting of Rough Surfaces and Wetting States

Wetting is a spreading of liquid on a solid or liquid substrate. It is ubiquitous and of primary importance in chemical, automobile, and food industries and medicine, soil, and climate sciences. Various technological applications call for surfaces with a

controlled wettability, in particular water- and oil-repellent surfaces. Rapid progress of *nanotechnologies* has allowed the fabrication of biomimetic surfaces demonstrating pronounced water repellence (the so-called lotus effect [1]). A variety of natural objects, such as birds' and butterflies' wings and legs of water striders, demonstrate extreme water repellency (superhydrophobicity, which is one of the *surface tension effects of nanostructures*) [2–4]. Various sophisticated experimental techniques, including UV lithography and plasma etching, were successfully applied for manufacturing lotus-like biomimetic surfaces [4, 5]. Superhydrophobic surfaces demonstrate a potential for a variety of green technologies, including energy conversion and conservation, and environment-friendly self-cleaning underwater surfaces [5].

The wetting of flat and rough surfaces is characterized by a contact angle, which is defined as the angle between the tangent to the liquid–fluid interface and the tangent to the solid surface at the contact line between the three phases (solid, liquid, and vapor) [6, 7]. The Young (or equilibrium) contact angle θ is the angle that a liquid makes with an ideal (smooth, rigid, chemically homogenous, insoluble, and nonreactive) surface. The Young angle θ is given by the well-known Young equation:

$$\cos\theta = \frac{\gamma_{SA} - \gamma_{SL}}{\gamma} \tag{1}$$

where γ, γ_{SL}, and γ_{SA} are the surface tensions at the liquid/air (vapor), solid/liquid, and solid/air (vapor) interfaces, respectively. The contact angle of a droplet deposited on a solid depends on external parameters, such as temperature. Change in temperature may stimulate the transition from the partial wetting to the complete wetting of the solid substrate. In this case, one observes a wetting transition (WT) on a smooth solid surface [8].

The Young equation supplies the sole value of a contact angle for a certain combination of solid, liquid, and gaseous phases. However, the experimental situation is much more complicated; even on atomically flat solid surfaces, a diversity of contact angles is observed. This is due to the long-range interaction between molecules forming the triple (three-phase) line of the droplet and molecules forming the solid substrate. The maximal possible contact angle observed on the surface is called the advancing angle θ_{adv}; the minimal one is called the receding angle θ_{rec} [6, 7]. The difference between advancing and receding contact angles $\theta_{\mathrm{adv}} - \theta_{\mathrm{rec}}$ is called the contact angle hysteresis [6, 7]. Experimental establishment of advancing and receding angles is a challenging task, and it should be mentioned that the reported contact angles are sensitive to the experimental technique used for their measurement. Both advancing and receding are equilibrium (though metastable) contact angles [7].

The wetting of rough and chemically heterogeneous surfaces is described by the apparent contact angle θ^* (APCA), which is different from the local contact angle given by Young's relation [6, 7]. The APCA is an equilibrium contact angle measured macroscopically. The detailed topography of a rough surface cannot be viewed with regular optical means; therefore, this contact angle is defined as the angle between the tangent to the liquid–vapor interface and the apparent solid surface, as macroscopically observed [7]. Superhydrophobic surfaces are characterized by high values of APCA (usually $>150^\circ$). However, high APCA does not necessarily guarantee the true superhydrophobicity and self-cleaning properties observed in nature. Moreover, high-stick surfaces demonstrating a high APCA were reported recently. Low contact angle hysteresis resulting in low sliding angles and high stability of "lotus-like" wetting state, as well as high APCA, supply water-repellent and self-cleaning properties to the surface. The design of such surfaces, allowing emerging green applications, remains a challenging scientific and technological task.

Chemical heterogeneities and roughness strengthen the hysteresis of a contact angle. The wetting of flat, chemically heterogeneous surfaces is characterized by the APCA θ^*, predicted by the Cassie–Baxter wetting model [9]. Consider wetting of a composite flat surface comprised of several materials. Each material is characterized by its own surface tension coefficients $\gamma_{i,\,SL}$ and $\gamma_{i,\,SA}$,

and by the fraction f_i in the substrate surface, $f_1 + f_2 + \ldots + f_n = 1$. The APCA in this case is supplied by the Cassie–Baxter equation:

$$\cos\theta^* = \frac{1}{\gamma}\sum_1^n f_i(\gamma_{i,SA} - \gamma_{i,SL}) = \sum_1^n f_i \cos\theta_i \quad (2)$$

where θ_i are equilibrium contact angles for the i^{th} material. The Cassie–Baxter equation can be applied to the solid surface comprised of pores (the contact angle for pores is π and $\cos\theta = -1$; see Fig. 1a). In this case, the Cassie–Baxter equation yields

$$\cos\theta^* = -1 + f_S(\cos\theta + 1) \quad (3)$$

where f_S and $1-f_S$ are the relative fractions of the solid and air fractions underneath the drop, respectively. The Cassie air-trapping wetting state is also called the "fakir state." The Cassie-like air-trapping wetting provides easy sliding of water droplets. When the relief is hierarchical, the Cassie–Baxter equation obtains more complicated forms, taking into account the interrelation between the scales constituting the topography of the relief.

The wetting of rough, chemically homogenous surfaces is governed by the Wenzel model [10]. According to the Wenzel model, surface roughness r, defined as the ratio of the real surface in contact with liquid to its projection onto the horizontal plane, always magnifies the underlying wetting properties (see Fig. 1b). Both hydrophilic and hydrophobic properties are strengthened by surface textures. The apparent Wenzel contact angle is given by Eq. 4:

$$\cos\theta^* = r\cos\theta \quad (4)$$

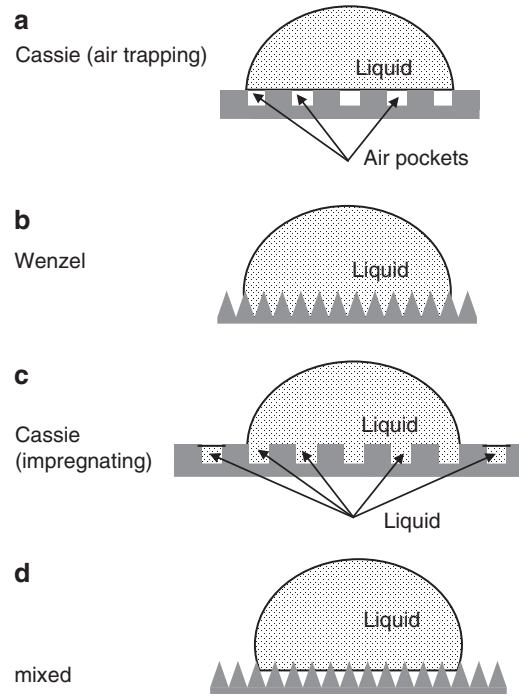
The Cassie wetting is featured by low contact angle hysteresis, whereas Wenzel-wetted surfaces demonstrate high hysteresis of the contact angle. Actually, pure Cassie and Wenzel wetting situations are rare in occurrence, and a mixed wetting state is often observed [11]. In the mixed wetting state, a droplet partially wets the surface and partially sits on air pockets, as described in Fig. 1d. The APCA θ^* is supplied in this case by Eq. 5:

$$\cos\theta^* = r_f f\cos\theta + f - 1 \quad (5)$$

In this equation, f is the fraction of the projected area of the solid surface that is wetted by the liquid, and r_f is the roughness ratio of the wet area. When $f = 1$, $r_f = r$ and Eq. 5 turn into the Wenzel Eq. 4; when $r_f = 1, f = f_S$ and Eq. 5 turn into the Cassie–Baxter Eq. 3.

One more wetting state is possible. This is the Cassie impregnating state, depicted in Fig. 1c. In this case, liquid penetrates into grooves of the solid, and the drop finds itself on a substrate viewed as a patchwork of solid and liquid (solid "islands" ahead of the drop are dry, as shown in Fig. 1c) [6]. The APCA is expressed as

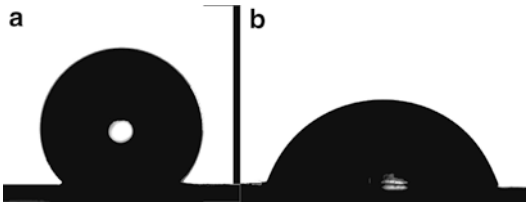
$$\cos\theta^* = 1 - f_S + f_S\cos\theta \quad (6)$$



Wetting Transitions, Fig. 1 Various wetting states occurring on rough surfaces: (**a**) Cassie air-trapping state, (**b**) Wenzel state, (**c**) Cassie impregnating wetting state, and (**d**) mixed wetting state

It should be stressed that Eqs. 3, 4, 5, and 6 could be applied when the radius of the droplet is much larger than the characteristic scale of the surface heterogeneities. The rigorous thermodynamic derivation of Eqs. 3, 4, 5, and 6 may be obtained within the general variational approach [12].

The physical mechanisms of WT on flat and rough surfaces are quite different. WTs on rough surfaces are related to the penetration of liquid into the grooves constituting the relief and result in the change of the APCA (see Fig. 2) [13–20]. Wetting transitions were observed under various experimental techniques utilizing a diversity of factors: gravity, pressure, bouncing and evaporation of droplets, electric field in the electrowetting experiments, and vibration of droplets [13–18]. WT was observed with the use of various experimental techniques, including reflection interference contrast microscopy and environmental scanning electron microscopy [20].



Wetting Transitions, Fig. 2 Wetting transitions observed by vibration of 15 μl waterdrop deposited on the micrometrically rough PDMS surface: (**a**) the initial Cassie state and (**b**) the Cassie impregnating state induced by vibrations

Energetic Approach to Wetting Transitions

Various wetting states featured by very different APCA can coexist on the same heterogeneous surface. The diversity of APCA could be easily understood if one takes into account that the Gibbs energy curve for a droplet on a real surface is characterized by multiple minima points. It could be shown that the Wenzel state (Fig. 1b) is energetically favorable compared to the pure Cassie air-trapping one, shown in Fig. 1a, when the inequality Eq. 7 holds:

$$\cos\theta > \frac{f_S - 1}{r - f_S} \tag{7}$$

The Wenzel state will be energetically favorable when compared to the mixed trapping state depicted in Fig. 1d, when the inequality Eq. 7a holds:

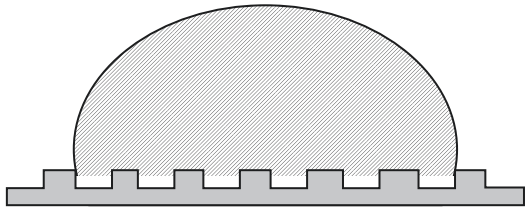
$$\cos\theta > \frac{f - 1}{r - r_f f} \tag{7a}$$

Obviously for hydrophilic materials, the Wenzel state is always favorable in relation to the Cassie one, whereas for the hydrophobic materials, the situation is more complicated. When the hydrophobic material is very rough ($r >> fs$), the Cassie state will be favorable for any obtuse θ. For single-scale hydrophobic materials, the Cassie air-trapping wetting state usually corresponds to the highest of the multiple minima of the Gibbs energy of the droplet deposited on a rough surface. Thus, for the WT, the energy barrier should be surmounted [18]. It was supposed that this energy barrier is equal to the surface energy variation between the Cassie state and the hypothetical *composite* state, with the almost complete filling of surface asperities by liquid, keeping the liquid–air interface under the droplet and the contact angle constant, as shown in Fig. 3.

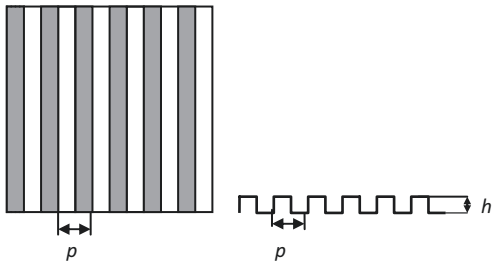
Wetting of hydrophobic side surfaces of asperities constituting the relief is energetically unfavorable; this gives rise to the energetic barrier. The collapse of the Cassie state will occur when the liquid filling the grooves will eventually touch the bottoms of the substrate pores, and the liquid–air interface will disappear. Contrasting to the equilibrium mixed wetting state, the composite state is metastable. For the simple topography depicted in Fig. 4, the energy barrier could be calculated as follows:

$$\begin{aligned} E_{\text{trans}} &= 2\pi R^2 h(\gamma_{SL} - \gamma_{SA})/p \\ &= -2\pi R^2 h\gamma\cos\theta/p \end{aligned} \tag{8}$$

where h and p are the geometrical parameters of the relief, shown in Fig. 4, and R is the radius of the contact area. The numerical estimation of the energetic barrier according to the Formula (8) with the parameters $p = h = 20$ μm, $R = 1$ mm, $\theta = 105°$ (corresponding to low-density



Wetting Transitions, Fig. 3 The composite wetting state



Wetting Transitions, Fig. 4 Geometrical parameters of the model relief illustrating the calculation of the Cassie–Wenzel transition energetic barrier

polyethylene (LDPE)), and $\gamma = 72\ \mathrm{mJ{\cdot}m^{-2}}$ gives a value of $E_{\mathrm{trans}} = 120$ nJ. It should be stressed that according to (8), the energy barrier scales as $E_{\mathrm{trans}} \sim R^2$. The validity of this assumption will be discussed below. The energetic barrier is extremely large compared to thermal fluctuations $\frac{E_{\mathrm{trans}}}{kT} \approx \left(\frac{R}{a}\right)^2 >> 1$, where a is an atomic scale. On the other hand, it is much less than the energy of evaporation of the droplet $Q: kT << E_{\mathrm{trans}} << Q$. Actually, this interrelation between characteristic energies makes wetting transitions possible. If it were not the case, a droplet exposed to external stimuli might evaporate before the wetting transition.

The energy barrier height from the side of the metastable (higher energy) wetting state is always much lower than that from the side of the stable one. The mentioned energy dependence is weak for large apparent angles and strong for low ones; this leads to a high asymmetry of the energy barrier, i.e., to a large difference between barrier heights. As a result, wetting transitions on rough surfaces are irreversible for both hydrophobic and hydrophilic materials.

Single- and two-stage pathways of WT were observed, including Cassie (air trapping)–Wenzel–Cassie (impregnating), Wenzel–Cassie (impregnating), and Cassie (air trapping)–Cassie (impregnating) transitions.

The Concept of the Critical Pressure

Various experimental methods used for the study of WT supplied the close values of the pressure necessary for the Cassie–Wenzel transition, which is of the order of magnitude of 100–300 Pa for 10 μl droplets deposited on micrometrically scaled rough surfaces. It is noteworthy that the Cassie air-trapping wetting regime observed on natural objects (birds' wings) is much more stable when compared to that on artificial surfaces.

The important parameter characterizing WT is the critical pressure under which the transition takes place [19]. For a single-scaled pillar-based biomimetic surface, with pillar width a and groove width b, the critical pressure equals

$$p_c = -\frac{\gamma f_s \cos\theta}{(1-f_s)\lambda} \tag{9}$$

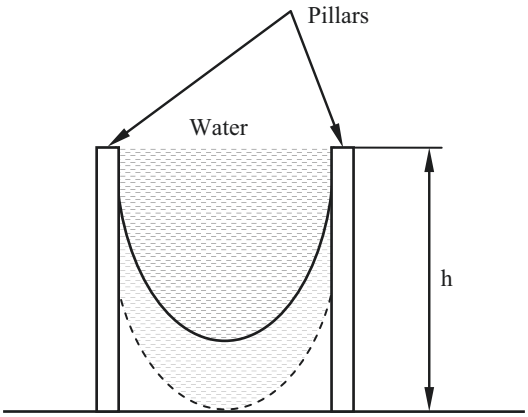
where $\lambda = \frac{A}{L}$, and A and L are the pillar cross-sectional area and perimeter, respectively. For the Teflon surface, $\theta = 114^\circ$, $a = 50$ μm, $b = 100$ μm yield $p_c = 296$ Pa, in agreement with experimental results, reported by various groups [15, 17]. Recalling that the dynamic pressure of rain droplets may be as high as 10^4–10^5 Pa, which is much larger than $p_c \approx 300$ Pa, it could be inferred that creating biomimetic reliefs with very high critical pressure is of practical importance. The concept of critical pressure leads to the conclusion that reducing the microstructural scales (e.g., the pillar diameters and spacing) is the most efficient means to enlarge the critical pressure [19].

The mechanism of WT based on the concept of critical pressure was proposed. It was supposed that as the pressure applied to the droplet increased, the meniscus will move toward the flat substrate, as shown in Fig. 5. The meniscus will eventually touch the substrate; this will cause the collapse of the Cassie wetting and consequently lead to a Cassie–Wenzel transition.

Wetting Transitions on Hydrophilic Surfaces

The Cassie wetting was observed on inherently hydrophilic surfaces. For the hydrophilic materials, the difference $\gamma_{SL} - \gamma_{SA}$ is negative, and liquid penetration into pores might have to proceed spontaneously, without the energy barrier. For the explanation of the roughness-induced superhydrophobicity of inherently hydrophilic materials, it was supposed that air is entrapped by cavities constituting the topography of the surface. The simple mechanism of "geometrical" trapping could be explained as follows: consider a hydrophilic surface ($\theta < 90°$) comprised of pores as depicted in Fig. 6. It is seen that air trapping is possible only if $\theta > \varphi_0$, where φ_0 is the angle between the tangent in the highest point of the pattern and the horizontal symmetry axis O_1O. Indeed, when the liquid level is descending, the angle φ is growing (see Fig. 6), and if the condition $\theta > \varphi_0$ is violated, the equilibrium $\theta = \varphi$ will be impossible.

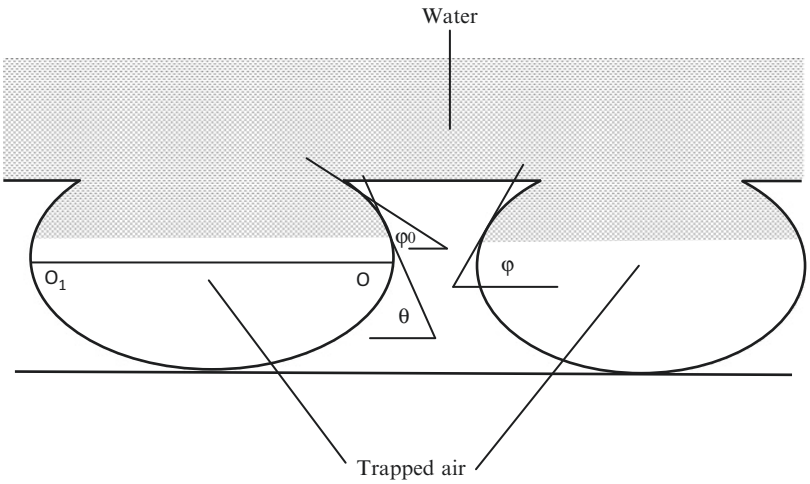
In the equilibrium position, small fluctuations of the contact angle lead to the appearance of curvature on the plane air–water interface that is energetically unfavorable. Below the central plane $O - O_1$ where $\varphi > 90°$, the equilibrium is impossible in the case of $\theta < \frac{\pi}{2}$. For a large pore, small fluctuations of θ can lead to touching the pore bottom near its center by the curved air–liquid interface, followed by filling the pore and the consequent collapse of the Cassie air-trapping wetting regime.



Wetting Transitions, Fig. 5 Pressure-induced displacement of the waterfront, leading to the collapse of the Cassie wetting state

The Dynamics of Wetting Transitions

The dynamics of WT is not clearly understood. The characteristic time of Cassie–Wenzel transition was established with reflection interference contrast microscopy as less than 20 milliseconds [20]. Two regimes of droplet front displacement were observed: zipping and non-zipping. In the zipping regime, the velocity of the front in one direction (to advance to the next row of pillars) is much smaller than the velocity in the other direction (liquid filling up one row of micro-pillars). It was established that the velocity of the wetting front increases with increasing gap size,



Wetting Transitions, Fig. 6 Geometrical air trapping on the hydrophilic reliefs

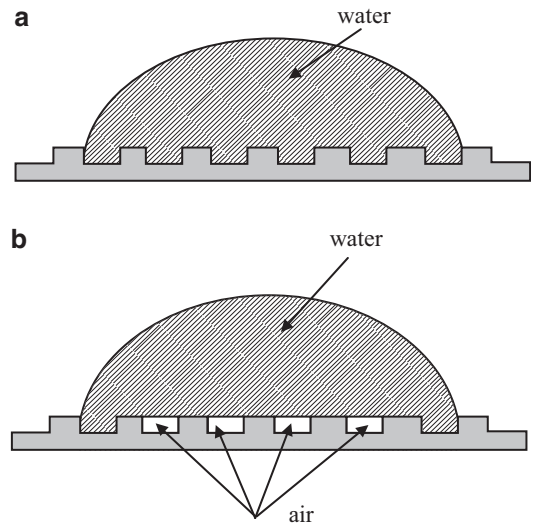
decreasing pillar height, or decreasing contact angle. A velocity of the wetting front as high as 1.5 m/s was registered.

The "Dimension" of Wetting Transitions

One of the unsolved problems in the field of WT is the problem of the "dimension" of the transitions or, in other words, whether all pores underneath the droplet should be filled by liquid (the "2D scenario") or, perhaps, only the pores adjacent to the three-phase (triple) line are filled under external stimuli such as pressure, vibrations, or impact (the "1D scenario"). Indeed, the APCA is dictated by the area adjacent to the triple line and not by the total area underneath the droplet. 1D and 2D scenarios of WT are illustrated by Fig. 7. The experiments carried out with vibrated drops supported the 1D scenario of wetting transitions. It has been established that the transition occurs when the condition $F_{cr} = \text{const}$ is fulfilled, where F_{cr} is the critical force acting on the unit length of the triple line, and the transition is caused by de-pinning of the triple line [17]. The critical value of the de-pinning force F_{cr} has been established experimentally for various microscopically structured surfaces as $F_{cr} \approx 200 - 350\ \text{mN} \cdot \text{m}^{-1}$ [17]. The energy barrier E_{trans} to be surmounted for the elementary displacement of the triple line δr could be estimated as

$$E_{trans} \approx 2\pi R F_{cr} \delta r, \qquad (10)$$

which scales as R. This scaling law corresponds to results obtained with vibrated drops, but contradicts the scaling law given by expression (8). The electrowetting stimulated WT also supported the 1D mechanism of transitions. De-pinning of the triple line was observed directly with the use of reflection interference contrast microscopy [20]. The potential barrier E_{trans}, according to expression (10) for a drop with a radius of $R \approx 1$ mm deposited on the LDPE relief, presented in Fig. 4 ($F_{cr} \approx 350\ \text{mN} \cdot \text{m}^{-1}$, the elementary displacement $\delta r \approx p/2 \sim 10^{-5}\text{m}$) equals $E_{trans} \approx 20$ nJ, smaller than the value predicted by formula (10), but still much larger than thermal fluctuations.



Wetting Transitions, Fig. 7 (**a**) 2D mechanism of WT (all pores are filled). (**b**) 1D scenario of WT (only pores adjacent to the triple line are filled)

It has been also suggested that the Cassie–Wenzel transition occurs via a nucleation mechanism starting from the drop center. In spite of intensive theoretical and experimental effort expended for the study of wetting transitions, the numerous aspects of these phenomena, including the dynamics of transitions and their "dimension," remain unclear and call for future investigations.

Cross-References

- ▶ Lotus Effect
- ▶ Nanotechnology
- ▶ Surface Tension Effects of Nanostructures

References

1. Koch, K., Bohn, H.F., Barthlott, W.: Hierarchically sculptured plant surfaces and superhydrophobicity. Langmuir **25**, 14116–14120 (2009)
2. Bormashenko, E., Bormashenko, Y., Stein, T., Whyman, G., Bormashenko, E.: Why do pigeon feathers repel water? Hydrophobicity of pennae, Cassie–Baxter wetting hypothesis and Cassie–Wenzel capillarity-induced wetting transition. J. Colloid Interface Sci. **311**, 212–216 (2007)

3. Zheng, Y., Gao, X., Jiang, L.: Directional adhesion of superhydrophobic butterfly wings. Soft Matter **3**, 178–182 (2007)
4. Quéré, D., Reyssat, M.: Non-adhesive lotus and other hydrophobic materials. Philos. Transact. R. Soc. A **366**, 1539–1556 (2008)
5. Nosonovsky, M., Bhushan, B.: Superhydrophobic surfaces and emerging applications: Non-adhesion, energy, green engineering. Curr. Opin. Colloid Interface Sci **14**, 270–280 (2009)
6. de Gennes, P.G., Brochard-Wyart, F., Quéré, D.: Capillarity and Wetting Phenomena. Springer, New York (2004)
7. Marmur, A.: A guide to the equilibrium contact angle maze. In: Mittal, K.L. (ed.) Contact Angle Wettability and Adhesion, vol. 6, pp. 3–18. VSP, Leiden (2009)
8. Bonn, D., Ross, D.: Wetting transitions. Rep. Prog. Phys. **64**, 1085–1163 (2001)
9. Cassie, A.B..D., Baxter, S.: Wettability of porous surfaces. Trans. Faraday Soc. **40**, 546–551 (1944)
10. Wenzel, R.N.: Resistance of solid surfaces to wetting by water. Ind. Eng. Chem. **28**, 988–994 (1936)
11. Marmur, A.: Wetting on hydrophobic rough surfaces: to be heterogeneous or not to be? Langmuir **19**, 8343–8348 (2003)
12. Bormashenko, E.: Young, Boruvka–Neumann, Wenzel and Cassie–Baxter equations as the transversality conditions for the variational problem of wetting. Colloids Surf. A **345**, 163–165 (2009)
13. Bartolo, D., Bouamrirene, F., Verneuil, E., Buguin, A., Silberzan, B., Moulinet, S.: Bouncing of sticky droplets: impalement transitions on superhydrophobic micropatterned surfaces. Europhys. Lett. **74**, 299–305 (2006)
14. Lafuma, A., Quéré, D.: Superhydrophobic states. Nat. Mater. **2**, 457–460 (2003)
15. Jung, Y.C., Bhushan, B.: Dynamic effects induced transition of droplets on biomimetic superhydrophobic surfaces. Langmuir **25**, 9208–9218 (2009)
16. McHale, G., Aqil, S., Shirtcliffe, N.J., Newton, G.M. I., Erbil, H.Y.: Analysis of droplet evaporation on a superhydrophobic surface. Langmuir **21**, 11053–11060 (2005)
17. Bormashenko, E., Pogreb, R., Whyman, G., Erlich, M.: Cassie-Wenzel wetting transition in vibrating drops deposited on rough surfaces: is dynamic Cassie-Wenzel wetting transition a 2D or 1D affair? Langmuir **23**, 6501–6503 (2007)
18. Barbieri, L., Wagner, E., Hoffmann, P.: Water wetting transition parameters of perfluorinated substrates with periodically distributed flat-top microscale obstacles. Langmuir **23**, 1723–1734 (2007)
19. Zheng, Q.-S., Yu, Y., Zhao, Z.-H.: Effects of hydraulic pressure on the stability and transition of wetting modes of superhydrophobic surfaces. Langmuir **21**, 12207–12212 (2005)
20. Moulinet, S., Bartolo, D.: Life and death of a fakir droplet: impalement transitions on superhydrophobic surfaces. Eur. Phys. J. E **24**, 251–260 (2007)

Whispering Gallery Mode Biosensor

▶ Whispering Gallery Mode Resonator Biosensors

Whispering Gallery Mode Resonator Biosensors

Kerry Allan Wilson[1] and Frank Vollmer[2]
[1]London Centre for Nanotechnology, University College London, London, UK
[2]Laboratory of Biophotonics and Biosensing, Max Planck Institute for the Science of Light, Erlangen, Germany

Synonyms

Microcavity biosensor; Monolithic total internal reflection biosensor; Optical cavity biosensor; Optical resonance biosensor; Optical ring resonator biosensor; Resonant evanescent wave biosensor; Whispering gallery mode biosensor

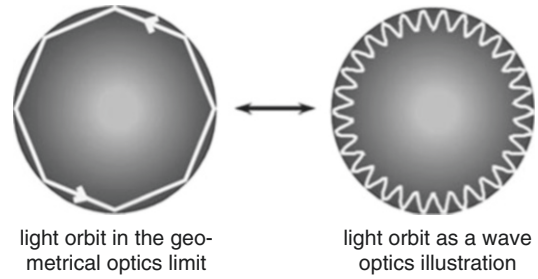
Definition

A whispering gallery mode resonator biosensor monitors optical resonances in microcavities for label-free detection of molecules and their interactions.

Overview

Whispering Gallery mode (WGM) resonators are named for the "whispering gallery" of St. Paul's Cathedral in London where sound waves propagate by reflection around the curved conduit of the gallery such that a word whispered at the front end of the gallery can be easily heard by a listener at the far end of the acoustic gallery. Similarly, optical Whispering gallery modes occur when light waves confined by total internal reflection in a dielectric microstructure are reflected back on

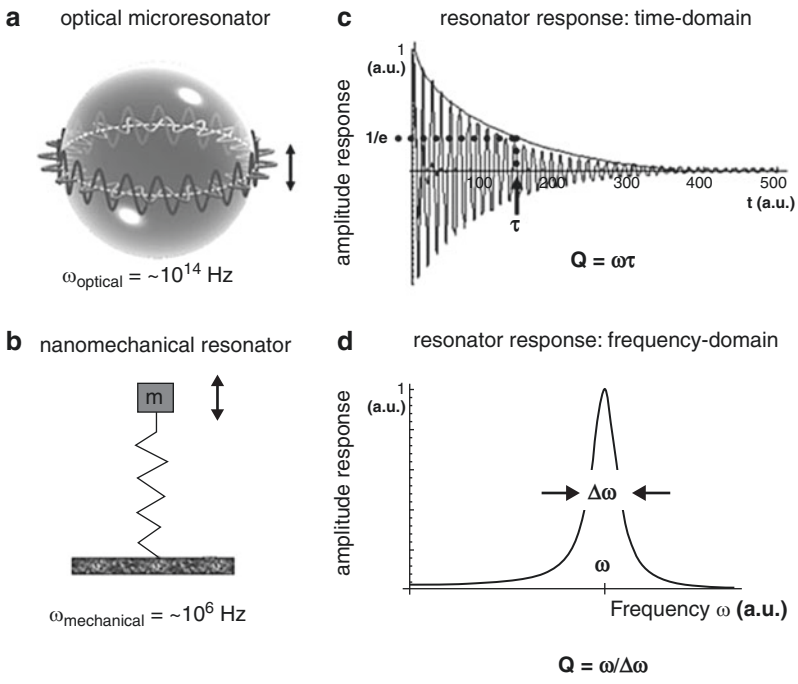
the same optical path where they interfere constructively. Interference generates the optical resonance signal, which occurs whenever exactly an integer number of wavelengths fit on the confined (circular) light path (Fig. 1). Molecules binding to the microresonator and interacting with the evanescent field generated at the sensor surface are detected from a change in optical resonance frequency. The frequency response is rendered specific for molecular binding events by functionalization of the sensor surface with biorecognition elements such as antibodies or oligonucleotides. WGM biosensors achieve single particle, single virus, and potentially single molecule detection capability by monitoring changes in the optical resonance frequencies in microspheres, microrings, microcapillaries, and microtoroids. Glass microspheres and microtoroids are particularly sensitive WGM biosensors because of their high Quality (Q) factor and their small modal volume (V). However, other types of optical resonators (microcavities) of different shapes and made of different materials are similarly suitable for biosensing applications, examples of these are ring resonators (made of silica, silicon, or transparent polymers) and photonic crystals (fabricated in silicon-on-insulator or in silicon-nitrate wafers). WGM biosensors and optical resonator derivatives have emerged as the one of the most sensitive microsystem biodetection technology that does not require chemical amplification or labelling of the analyte.



Whispering Gallery Mode Resonator Biosensors, Fig. 1 Schematic representation of light confined inside a resonant microcavity by total internal reflectance

Optical Resonator Physics: Confining Light at High Q-Factor

A high-quality (Q) factor is necessary for achieving sensitive detection using optical resonance in microcavities. Optical resonance is created by confining coherent light inside a miniature dielectric structure such that the returning light wave interferes constructively [1–3]. Figure 1 illustrates the example of an optical resonance confined by total internal reflection in a glass microsphere. Ideally, optical resonators (also referred to as cavities or microcavities) would confine light indefinitely. Real-world divergence from this condition due to damping losses results in a finite resonance lifetime τ. This finite resonance lifetime can be measured, i.e., by charging the microcavity with light, terminating the light source, and then measuring the time it takes for the intensity of the confined light wave to drop to 1/e of its initial value. This exponentially decaying intensity-amplitude of the optical resonator is analogous to the decreasing amplitude of a (nano) mechanical resonator, here illustrated as mass-on-a-spring (Fig. 2b). To compare optical, mechanical, and other resonant systems which can exhibit very different resonance lifetimes, one defines a time-independent and dimensionless figure of merit, the quality (Q)-factor. The Q-factor is proportional to the total number of oscillations stored in the resonator, and this definition is independent from the type of resonator (optical, mechanical, etc.) and the nature of its oscillation. For example, in an optical resonator, the value of the Q-factor is 2π times the number of electromagnetic oscillations before amplitude of the confined light wave drops below 1/e of its initial value. Similarly, the value of the mechanical Q-factor is just 2π times the number of (Pendulum-)swings before that same limit is reached. For any resonant system, the Q-factor is thus defined as $Q = \omega\,\tau$, where τ is the resonator lifetime and where ω is 2π times the resonance frequency (Fig. 2). The resonance frequency for optical microresonators ranges from a few TeraHertz (10^{12} Hz) to one PetaHertz (10^{15} Hz), whereas nanomechanical resonators operate at a few MHz at most. Although resonance lifetimes of very large mechanical resonator



Whispering Gallery Mode Resonator Biosensors, Fig. 2 (**a**) and (**b**) Comparison of optical and nanomechanical resonators. (**c**) Definition of Q-value in the time domain. (**d**) Definition of Q-value in the frequency domain

(i.e., Feaucault Pendulum) exceed those of optical resonators (many hours compared to microseconds), the optical microresonators exhibit much higher Q values as compared to their large mechanical and nanomechanical counterparts: in fact, microcavities exhibit the highest measured Q-factor of almost $\sim 10^{10}$, whereas a good nanomechanical resonator reaches a Q value of a few million at best and a Pendulum operated in vacuum reaches a Q of about 100,000. In fact, the "ultimate" Q-factor has been measured in a microsphere cavity where an optical resonance (a whispering gallery mode, WGM) is efficiently confined by total internal reflection – with only residual damping due to absorption in the glass material.

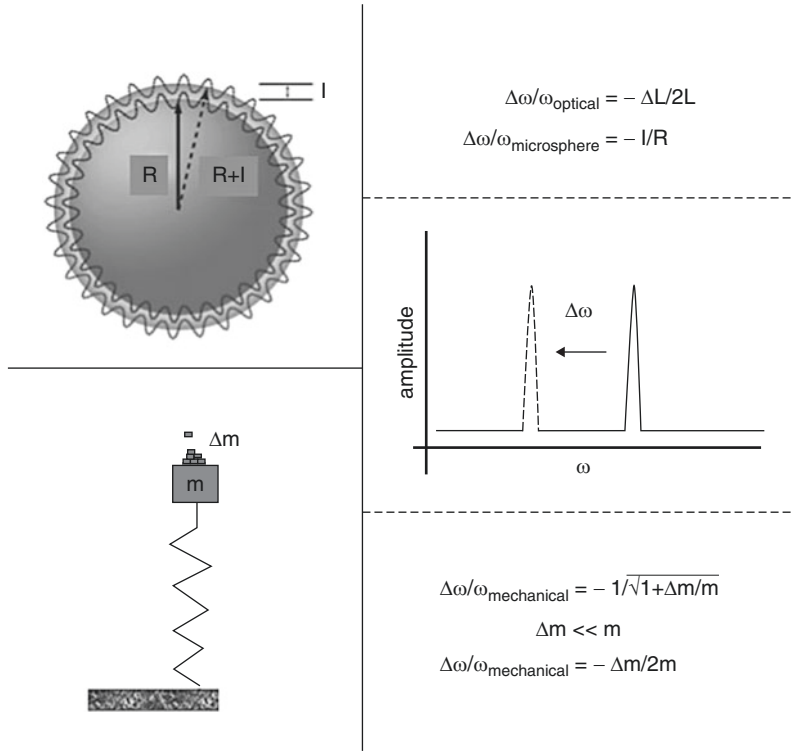
Apart from the time-domain response, a resonator also has a frequency-domain response (Fig. 2c) and it is important to realize that a large Q-factor is associated with a very narrow spectral linewidth where $\Delta\omega_{\text{line}} = \omega/\text{Q}$ (Fig. 2d). A narrow linewidth allows for precise measurements of resonance frequency and changes thereof, which is of paramount importance for achieving high sensitivity in biosensing applications. Ultimate-Q microsphere resonators are therefore the ideal choice to build a sensitive WGM biosensor [4]. The direct relationship between a high-Q factor and a narrow linewidth can be intuitively understood from the fact that the frequency-domain response of the resonant system (Fig. 2d) is just the Fourier-transform of the time-domain response (Fig. 2c). The Q-factor is not only related to the time-domain and the frequency-domain signal of a resonator, but also to the energy stored in a resonant system. The Q-factor can be calculated from the total energy W_{total} stored in the resonator divided by the energy ΔW lost per oscillation (multiplied by 2 Pi), $\text{Q} = 2\pi\ W_{\text{total}}/\Delta W_{\text{per oscillation}}$. For example, the lifetime and Q-factor of an ultimate WGM microsphere resonator is limited only by residual absorption and calculated as $\text{Q} = 2\pi\ (W_{\text{total}}/\Delta W)_{\text{per oscillation}} = 2\pi\ n/\lambda\alpha$, where α is the extinction coefficient of glass, about 7 dB/km, λ is the wavelength (here 633 nm), and n the refractive index of the glass (here $\sim$1.45). Table 1 gives an overview of Q-factors of resonant optical, electrical, and mechanical systems.

Principles of Resonant WGM Biosensing

An optical resonator transduces the binding of molecules into a resonance frequency shift signal which can be detected with great precision due to the high Q-factor of the optical detection system [5]. Optical resonators have no moving parts and maintain a high Q-factor even in a liquid biosensing environment [4] where any mechanical system would be heavily damped (low Q). The change in resonance frequency of an optical resonator occurs in response to the binding of molecules which increase the round-trip optical path length. This is different from a mechanical resonator where the frequency shift would occur due to mass loading by the added molecules (Fig. 3). Such a frequency shift, i.e., for a nanomechanical resonator can be easily calculated: Additional molecular mass loading, Δm, on a system with a nominal mass, m, will cause a resonance frequency shift of $\Delta\omega_{\text{mechanical}} = -\omega_{\text{mechanical}}\ \Delta m/2m$ (Fig. 3), where only nanomechanical systems are useful for such measurements due to their small mass. For an optical resonator instead, the binding of molecules add additional optical path length ΔL causing the resonance frequency ω_{optical} to shift by $\Delta\omega_{\text{optical}} = -\omega_{\text{optical}}\ \Delta L/2L$, where $2L$ is the round-trip optical path length in the microcavity (Fig. 3). For the special case of a spherical WGM optical resonator where bound molecules form an optically contiguous layer of thickness l (Fig. 3), the resonance frequency shift is calculated as $\Delta\omega_{\text{microsphere}} = -\omega_{\text{microsphere}}\ l/R$, where l is the thickness of the adsorbed molecular layer and R is the radius of the microsphere. This estimate assumes an identical refractive index for

Whispering Gallery Mode Resonator Biosensors, Table 1 Q-factors of various resonant systems

Integrated circuit filter	20
Mechanical wristwatch	100–300
Surface acoustic waves	2,000
Foucault pendulum (in vacuum)	10^4–10^5
Quartz crystal	10^6
Microsphere optical resonator	10^{10}



Whispering Gallery Mode Resonator Biosensors, Fig. 3 Reactive sensing principle for an optical resonator and for a nanomechanical resonator

microsphere and adsorbed molecular layer, $n_{glass} = n_{molecules}$, a good approximation for the case of proteins adsorbing to a glass microsphere ($n_{glass} = n_{protein} \sim 1.45$). Further assuming the ability to detect the resonance frequency shift (Fig. 3) with a resolution of S = 1/50 of the linewidth, and assuming a modest Q of 10^7, which can be readily obtained with a 100 100 μm mu;m diameter glass microsphere in a liquid biosensing environment, the smallest detectable thickness l of a molecular layer bound to a WGM microsphere cavity is calculated as l = S/Q * R = (0.02/10^7)* 50 μm = 10^{-7} = 0.1 pm! Biomolecules are several nm in diameter (~3.5 nm for a typical protein such as BSA), so detection of a small fraction of a monolayer coverage is an easy task for a WGM biosensor. Achieving such a feat with a (nano)-mechanical system is impossible due to the heavy damping (low Q) of any mechanical resonator after immersion in a liquid biosensing environment. In contrast, high-Q optical resonators have no moving parts and the Q-factor (and corresponding resonance lifetime) is mainly limited by the absorption of the light in the resonator material and in the surrounding medium (ultimate Q). By choosing a wavelength with minimal absorption losses, and by optimizing the material and size of a microsphere resonator, it has been estimated that unprecedented, label-free single molecule detection capability is achievable [6]. This possibility combined with the great variety of materials and geometries available for the design of high-Q optical systems has spurred the tremendous interest in the development and application of microcavity biosensors.

WGM Biosensing: Reactive Sensing Principle

The frequency shift $\Delta\omega$ of an optical resonance due to binding of any kind of molecule to any kind of resonator can be precisely calculated from first-order perturbation theory. Assuming that bound molecules do not change the light field, then the fractional change in resonance frequency $\Delta\omega/\omega$ (and corresponding fractional wavelength $\Delta\lambda/\lambda$) is calculated [6] by comparing the energy it takes to polarize the molecule bound to the microcavity surface at position $\vec{r}_0$, $\Delta W_{particle} = h\varpi = \frac{1}{2}\alpha \vec{E}\left(\vec{r}_o\right)^2$, to the total electromagnetic energy confined in the microcavity $W_{total} = \int \varepsilon \vec{E}\left(\vec{r}\right)^2 dV$:

$$\frac{\Delta\lambda}{\lambda} = \frac{\Delta W_{particle}}{W_{total}} = \frac{\alpha \vec{E}\left(\vec{r}_0\right)^2}{2\int \varepsilon \vec{E}\left(\vec{r}\right)^2 dV} \quad (1)$$

This so-called reactive sensing principle can be applied to any microcavity structure and to any biomolecule. Furthermore, a large frequency shift and therefore a high sensitivity is obtained for a small denominator in Eq. 1, which is achieved by confining the resonant light field within a small modal volume defined as $V = \int \varepsilon \vec{E}\left(\vec{r}\right)^2 / \vec{E}\left(\vec{r}\right)^2_{max} dV$, where $\vec{E}\left(\vec{r}\right)_{max}$ is the highest field intensity measured in the cavity. For high-sensitivity biosensing applications, it is therefore important not only to operate at a high Q value, enabling precise measurement of resonance frequency, but equally important to use a miniature optical resonator with small modal volume V, producing a large frequency shift signal due to the binding of molecules. Biosensing applications, therefore, require optimizing the Q/V value. WGM microsphere biosensors typically exhibit modal volumes ~500 um^3 at Q ~ 10^7 and therefore achieve Q/V ~ 2 × 10^4 (1/μm^3). Photonic crystal microcavity structures may further boost current Q/V values since these resonators can confine light in record-low modal volumes of V < $(\lambda)^3$/n ~ 0.2 μm^3 (where n is the refractive index of the resonator), a volume which is less than a wavelength cubed! This strong spatial localization of light in a photonic crystal cavity is possible due to multiple Bragg scattering from many lattice elements, different from a WGM resonator where the light field is confined by internal reflection. Although photonic crystals can confine light in record low modal volumes, Q-factors of photonic crystal cavities are typically lower as compared to WGM

resonators and reach only ~one to a few million. Here, the Q-factor is mainly limited by the finite precision of the nanoscale photonic crystal fabrication process where residual surface roughness causes additional damping of the resonant light field. With further improvements of the nanoscale fabrication processes, however, photonic crystal microcavities may surpass current record Q/V's of WGM biosensors and are therefore an interesting resonator structure that can be potentially mass produced by deep UV lithography.

WGM biosensors to date are more sensitive as compared to any other commercialized label-free biosensor technology such as surface plasmon resonance sensors (SPR) and quartz crystal microbalances (QCM). WGM biosensors routinely achieve sensitivity far below 1 pg/mm^2 molecular mass loading, approaching the specific detection of the binding of molecules from fM (and less) solution concentration levels. Impressively, WGM technology was used to demonstrate the detection of a single Influenza A virus particle from fM solution concentration [7], a noteworthy feat considering that a single Influenza A virus particle measures only ~50 nm in radius and weighs less than ~500 attograms.

WGM Biosensing: Label-free Analysis of Homogenous Molecular Binding Events

For WGM microsphere resonators, Eq. 1 can be used to derive an analytic equation for the observed fractional frequency shift $\Delta\omega/\omega$ due to homogenous binding of nanoscale molecules or particles at random location on the microsphere sensor surface [4, 6]. Then, the frequency shift $\Delta\omega/\omega$ and the corresponding wavelength shift $\Delta\lambda/\lambda$ is calculated from (Fig. 4):

$$\frac{\Delta\lambda}{\lambda} = -\frac{\Delta\omega}{\omega} = \frac{\sigma\alpha}{\varepsilon_0\left(n_s^2 - n_m^2\right)R} \qquad (2)$$

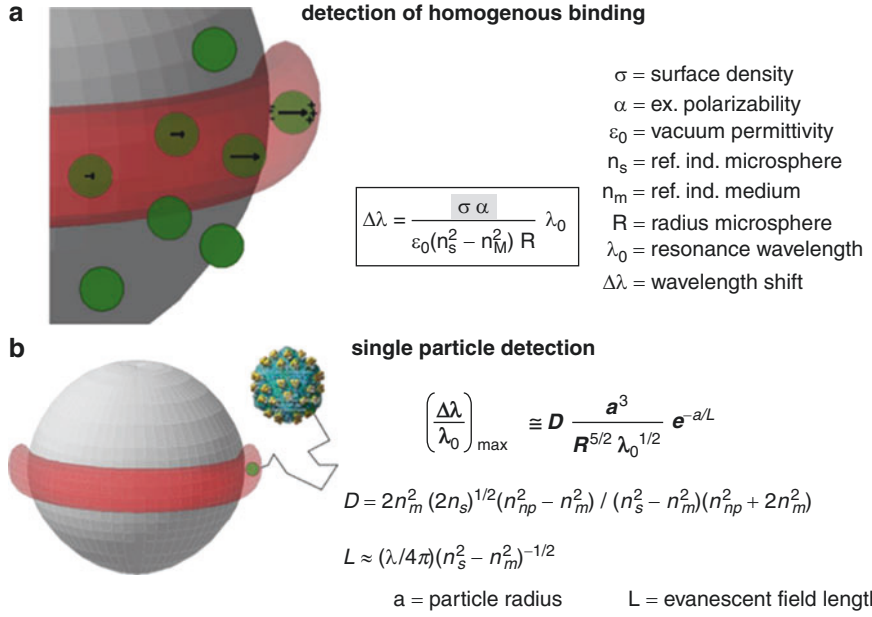
where σ is the surface density of bound molecules, α is their excess polarizability (in excess to the displaced water), ε_0 is the vacuum permittivity, n_s and n_m are the refractive indices of the glass microsphere and the surrounding water, respectively, and R is the radius of the microsphere. The negative sign for $\Delta\omega/\omega$ results from the fact that a shift to a longer resonance wavelength corresponds to a shift to lower resonance frequency. The excess polarizability α can be estimated from tabulated "dn/dc" values for each biomolecule [8] measured in aqueous solution. The refractive index increment "dn/dc" is defined as the increase Δn of the refractive index n measured for the increase Δc of the biomolecule solution concentration c in the limit for c - > 0: $dn/dc = \lim(\Delta n/\Delta c)c_{->0}$. The excess polarizability α is then calculated from dn/dc values according to,

$$\alpha = \varepsilon_0 2n\frac{dn}{dc}m \qquad (3)$$

where n is the solvent refractive index (water), and m is the mass of the single biomolecule. Proteins such as bovine serum albumin (BSA) have a typical dn/dc value of about 0.183 cm^3/g. The excess polarizability of BSA is then calculated as $\alpha_{BSA} = 4\ \pi\ \varepsilon_0\ 3.85\ *\ 10^{-21}$ cm^3, where the mass of a single BSA molecule is ~66,432 (g/mol)/6.02 10^{23} (1/mol) = 1.1 10^{-19} g. The excess polarizability α is proportional to the molecular mass, m, of a protein and may vary only slightly between different proteins of similar molecular weight but of different amino acid sequence, such as for BSA and HSA (human serum albumin). Similar arguments can be made for many other biopolymers [8]. Table 2 lists typical dn/dc values of various biopolymers and biomolecules, in different solvent conditions.

Note that dn/dc values depend on many additional parameters such as temperature, solvent, ionic strength, and probing wavelength (Table 2). Interestingly, one can also calculate the wavelength dependence of polarizability values from Kramers-Kronig transformations by integration over the respective absorption spectra [8].

In biosensing experiments, one often measures the resonance wavelength shift $\Delta\lambda$ instead of the frequency shift: $\Delta\lambda = \lambda_{res,\ after} - \lambda_{res,before}$. The wavelength shift $\Delta\lambda$ can be converted in the associated frequency shift Δf by:



Whispering Gallery Mode Resonator Biosensors, Fig. 4 (**a**) Detection of homogenous binding of biomolecules in the evanescent field of a WGM. (**b**) Detection of single particle interacting with WGM

$\Delta f = c_{\text{light}}\,(1/\lambda_{\text{res, after}} - 1/\lambda_{\text{res, before}}) = -\,c_{\text{light}}\,\Delta\lambda/\lambda^2$, where λ is the initial resonance wavelength and c_{light} is the speed of light. In all WGM biosensing experiments, the wavelength shift, $\Delta\lambda$, or frequency shift, Δf, reports on the surface density σ of molecules bound to the microsphere resonator, at a specific time point. Depending on specific cavity material or surface functionalization, the saturation surface density of bound molecules may vary between 0 and $10^{15}/\text{cm}^2$. To achieve a surface coverage of more than $10^{15}/\text{cm}^2$ requires molecules to form multiple layers or to bind within less than a nm distance, certainly a physical limit as the size of an atom is, ~0.1 nm = 1 Å. High-Q microcavities are invaluable tools for studying the binding of molecules to the sensor surface since precise measurements of high-surface densities can reveal interactions between closely spaced molecules, and precise measurements at low surface densities $<10^9/\text{cm}^2$ can reveal intricate mechanisms of adsorption on non-fouling surfaces.

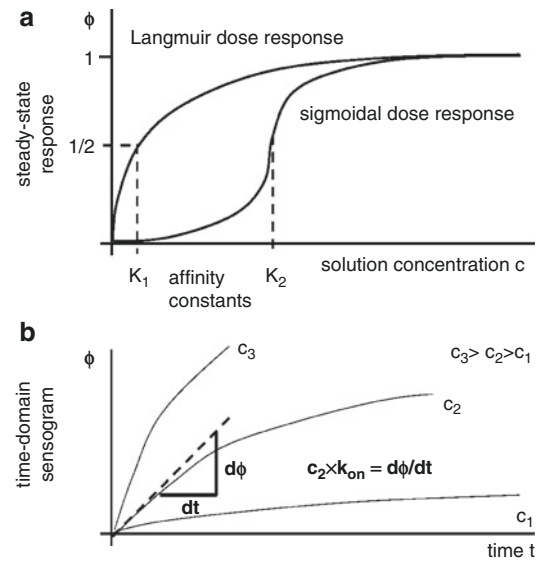
The time scale of resonance wavelength shift depends on the diffusion of molecules and on the time scales associated with their interactions at the

Whispering Gallery Mode Resonator Biosensors, Table 2 Typical dn/dc values of biopolymers and molecules

Biomolecule or biopolymer	~dn/dc (cm^3/g) [at ~ 633 nm wavelength]	Measurement parameter notes
Bovine serum albumin	0.183	Water at 23 °C
Deoxyribonucleic acid from calf thymus	0.166	0.1N tris buffer pH = 8.4 at 23 °C
a-chymotrypsinogen	0.171	0.15M NaCl; 0.1 pH buffer, pH = 7.2 at 23 °C
Dextran	0.147	Water at 23 °C
Gelatin	0.163	0.1N NaCl at 23 °C
Lignin sulfonate – Na	0.188	0.1M $NaNO_3$; pH = 6.6 at 26 °C
Mucopolysaccharides	0.110	0.1 % NaCl at 23 °C
Polydimethylsiloxane	−0.091	Toluene at 23 °C
Polyvinyl alcohol 98% hydrolyzed	0.150	0.05M $NaNO_3$
RNA	0.160	Water at 23 °C

resonator surface. The simplest steady-state kinetic model which can be applied to the binding of molecules to a surface is given by the Langmuir-like model [9] which assumes equilibrium binding between molecules and a fixed number of surface binding sites, further assuming equal affinity and no interaction between the sites. To determine a Langmuir binding isotherm, one first calculates the fractional surface coverage Φ by dividing the equilibrium frequency shift by the shift at saturation (measured at high solution concentration). Then, one plots the measurement of fractional surface coverage Φ versus the different solution concentration levels (Fig. 5a). The rate equation for the Langmuir model is given by $d\Phi/dt = k_{on}*c*(1 - \Phi) - k_{off}*\Phi$, so that in steady state $d\Phi/dt = 0$ and $\Phi/(1 - \Phi) = k_{on}/k_{off}*c$. For half-maximal surface coverage at $\Phi = 1/2$, the solution concentration is just equal to the equilibrium (dissociation) constant $c = K = k_{off}/k_{on}$, which can be conveniently determined from the Langmuir isotherm (Fig. 5a). The affinity (dissociation) constant is, e.g., measured at ~1 nM for BSA proteins binding to an aminosilanized microsphere [4]. For more complex binding and adsorption processes, different kinetic models can be applied such as the random sequential adsorption (RSA) model, which is derived from the Langmuir model by taking into account the additional exclusion volume of bound molecules. Another kinetic model considers conformational changes by modeling bound proteins as discs with radii that may change in a post-binding step. The reader is referred to standard pharmacokinetic textbooks for further analysis of dose–response curves and their importance in analytical and clinical chemistry. Obtaining such dose–response curves is essential for relating the solution concentration of analyte molecules to the measured surface concentrations as measured by the resonator shift signal.

Instead of analyzing molecular recognition and binding in the steady-state regime, the interaction of biomolecules at a resonator surface can also be analyzed in the time domain by plotting the resonance frequency or wavelength shift signal versus time (sensogram, Fig. 5b). The sensogram and its derivative report on the kinetics and the rate of



Whispering Gallery Mode Resonator Biosensors, Fig. 5 (**a**) Steady-state response of WGM biosensor. (**b**) Time-domain response (sensogram) of a WGM biosensor

adsorption or desorption, respectively. It is common to extract the derivative (rate of adsorption or desorption) at the onset of the experiment where $d\Phi/dt = k_{on} * c$ in a binding experiment started at $c > 0$ and $\Phi = 0$, and $d\Phi/dt = k_{off}$ for an unbinding experiment started at $c = 0$ and $\Phi > 1$. Since time scales for binding can be fast, it is often necessary to start the experiment rapidly using microfluidic techniques that allow a rapid change of solution concentration. In principle, optical resonators follow such a fast response without problem since the time resolution of a microcavity is principally limited only by the resonator photon lifetime τ, which is on the order of micro to nanoseconds, and dependent on the Q-value.

WGM microsphere cavities have a further advantage as compared to SPR and QCM techniques in that they exhibit multiple optical resonances signals, which are useful for detection. In fact, optical resonances appear throughout an optical resonator spectrum since the resonance condition states that a multiple number of resonance wavelengths have to fit the optical path. The precise calculation of the resonances requires Maxwell's equations and has many solutions,

and a more versatile sensor uses more than one resonance shift signal to analyze the binding of molecules to the resonator surface. For example, differently polarized (TE and TM) optical resonances can be used to determine the orientation of molecules at the sensor surface [5]. Each one of the two resonances (TE and TM) probe the surface of the microsphere in a different direction, and the specific arrangement of molecules in a monolayer and changes thereof can be measured by comparing the TE and TM wavelength shifts. Different resonance wavelength shifts have already been used to determine conformational changes in photosensitive bacteriorhodopsin membrane and to measure the orientation of the photochrome retinal within a lipid bilayer [5]. It is also possible to compare the resonance wavelength shifts for resonances excited at different wavelength (e.g., at 763 nm and 1,310 nm) in parallel. Using this approach, it is possible to directly determine the thickness and the permittivity (refractive index) of an adsorbed nanolayer, and changes thereof [10].

$$\frac{\left(\frac{\partial\lambda}{\lambda}\right)_1}{\left(\frac{\partial\lambda}{\lambda}\right)_2} \approx \frac{L_1[1-\exp(-l/L_1)]}{L_2[1-\exp(1-l/L_2)]} \tag{4}$$

Where l is the thickness of the adsorbed layer and L_1 and L_2 are the evanescent field length at probing wavelength λ_1 and λ_2, with $L \approx (\lambda/4\pi)\left(n_s^2 - n_m^2\right)^{-1/2}$.

WGM Biosensing: Label-free Single Particle Analysis

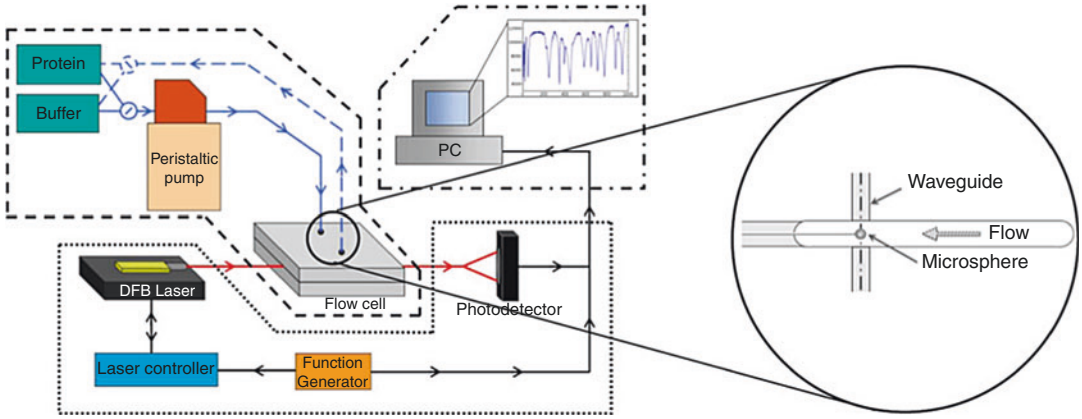
On the single particle level, the frequency shift (or rather "step") associated with a single binding event depends on the location of the particle on the resonator surface. For a microsphere WGM biosensor, the light field is confined along the equator of the sphere (Fig. 4) and the largest frequency shift is recorded for an equatorial binding event. The frequency shift associated with the equatorial binding event can be calculated from the reactive sensing principle [7] Eq. 1, where the fractional wavelength shift now depends on particle size, a, and the evanescent field length, L (Fig. 4b). Single bioparticle detection was demonstrated for a single Influenza A virion, and analysis of the discrete frequency shift signals (Fig. 4b) was used to determine the size of the particle (~50 nm radius) as well as its weight (~500 attograms) [7]. The size of a bound nanoparticle can also be determined independently of binding location by measuring resonance frequency shift as well as changes in Q-factor in parallel [11]. In general, single particle detection opens up the possibility to determine size and shape of nanoparticles by monitoring frequency shifts of different optical resonances (TE, TM, etc.) and associated Q-factors in parallel. Even without binding, but already within the evanescent field, the Brownian motion of a single particle causes resonance wavelength fluctuations [5], a signal that can be used to extract single particle diffusion constants. Furthermore, the frequency shift signal associated with a single particle trapped by the optical field of the resonator can be used to determine a chemical surface potential: If a particle of radius a is trapped in the optical field of the resonator, the resonance wavelength shift reports on the radial position of the particle measured as the distance h from the microsphere surface [12]:

$$\frac{\Delta\lambda_r(h+a)}{\Delta\lambda_r(a)} = \frac{\exp[-(h+a)/L]}{\exp[-a/L]} = \exp[-h/L] \tag{5}$$

From this measurement, a histogram of equilibrium particle positions can be replotted as a Boltzmann-distribution in kT units (k = Boltzmann constant, T = Temperature), revealing the trapping potential consisting of a chemical potential on the resonator surface and the optical potential of the evanescent field.

Implementation and Operation of a WGM Resonator Biosensor

In its simplest implementation (Fig. 6), the WGM biosensor device is build by coupling light from a tapered optical fiber to a microsphere optical

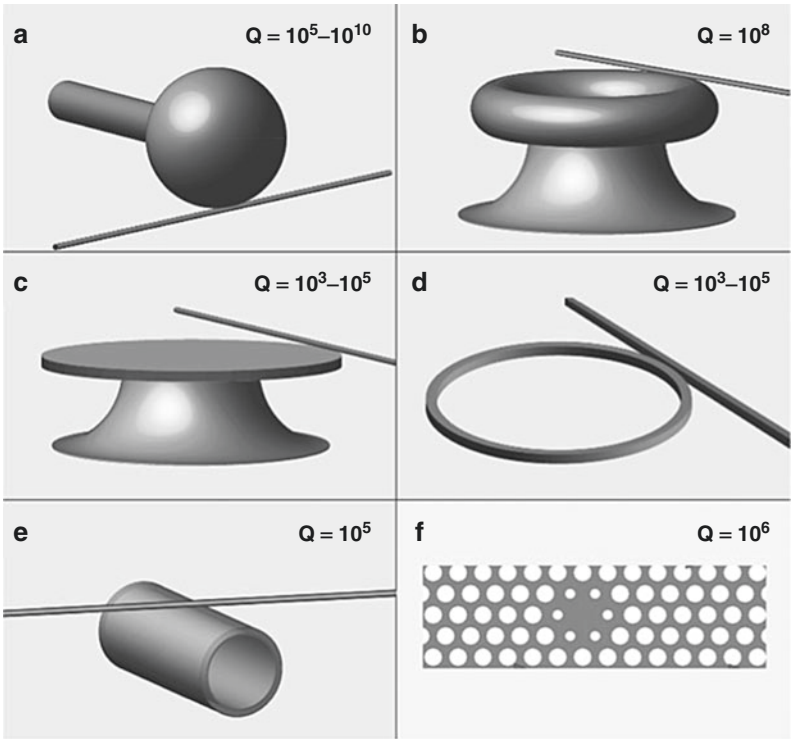


Whispering Gallery Mode Resonator Biosensors, Fig. 6 Schematic representation of WGM biosensor experimental setup comprised of fluidics component (*dashed lines*), excitation source and detector (*dotted line*), and data acquisition software (*dash-dot line*), and sphere-on-stem resonator housed in a flow cell (*inset*)

resonator [4]. The microsphere itself is fabricated from an optical fiber piece where one end is melted in a hot flame, forming a sphere on-a-stem structure. The sphere is then positioned adjacent to the tapered fiber region, where the light can couple from the optical fiber to the resonator. Light from a tunable laser is coupled into one end of the optical fiber from where it is delivered to this tapered coupling region. The output end of the optical fiber is connected to a photodetector, which records the transmitted intensity. As the laser wavelength is tuned to match the resonance wavelength of the microsphere, light starts to couple from the tapered fiber region to the resonator where it remains confined as WGM optical resonance. Due to interference in the coupling region, the light then no longer reaches the photodetector, which now records a minimum in the transmitted intensity. A computer identifies and tracks the minimum in the recorded transmission spectrum and plots the resonance wavelength versus time, at ms to μs time resolution, typically limited by the speed of the acquisition system, as well as by the speed of the laser control system. The sensogram is analyzed either *post hoc* or in real time and contains information on the binding and interaction of molecules at the WGM biosensor surface.

Several other experimental configurations exist for coupling light into WGM biosensors and for detecting resonances. Earlier studies of WGM resonances even employed "free beam" coupling methods in which the incident laser is guided through external optics and focused onto the microcavity to excite the resonance. Only a small fraction of the incident light, however, can couple radiatively to the microcavity. It is also possible to use light sources other than lasers, e.g., light sources with a broader emission spectrum, such as an ASE (amplified spontaneous emission) light source operating in the telecom band. In this case, however, a high-resolution monochrometer is necessary to resolve the narrow linewidth resonances, certainly a limitation when using a high Q cavity. In other implementations, it is possible to use spheres labeled with a fluorophore and resolve resonances in the emission spectrum, which is collected on a bead-by-bead basis. This implementation limits the operation of the biosensor to lower Q factors, but allows the use of freestanding fluorescent spheres, such as 10 um-diameter polystyrene spheres soaked with rhodamine 6G fluorophore [13].

Most common experimental setups utilize tapered fiber optic waveguides or planar waveguides to deliver the light to various resonator geometries such as discs, rings, toroids, or spheres (Fig. 7). Light sources typically consist of a tunable distributed feedback laser (DFB) with 300 kHz linewidth available at various nominal wavelengths (760–1,550 nm) and which can be tuned by a fraction of a nm simply by ramping the



Whispering Gallery Mode Resonator Biosensors, Fig. 7 Schematic illustration of different resonator geometries. (**a**) Sphere, (**b**) Toroid, (**c**) Disk, (**d**) Ring, (**e**) Cylinder, (**f**) Photonic crystal

operating (laser diode) current [4]. A simple (silicon or InGaAs) photodetector is sufficient to detect the intensity, and a standard computer can record the spectrum and locate the resonance wavelength. Biosensing experiments often require the integration of the resonator with a microfluidic flow system. Figure 6 shows the integration of a complete microsphere WGM biosensor with a simple flow cell.

Resonator Geometry

One of the primary benefits of WGM biosensors is the unprecedented sensitivity provided by high Q-factors in optical microcavities. High Q-factors are found for a variety of resonator shapes and fabricated by different procedures (Fig. 7). The decision for particular sensor geometry depends on the required sensitivity, cost, multiplexing capability, and need for integration with microfluidics and other photonic or electronic components. Often, the most important factor remains sensitivity and to date the highest Q-factor was reported for a microsphere resonator at $Q \sim 10^{10}$ which is thus the ideal choice for building highly sensitive WGM biosensors. The ultimate sensitivity obtained in ultrahigh-Q resonators enables sensing of single virus particle and molecules. However, more modest Q factors ($Q = 1 \times 10^4$ to 1×10^7) observed in other resonator geometries and materials (Fig. 7) are sufficient in most biosensing applications and can already achieve detection limits superior to SPR sensors (<1 pg/mm^2). As pointed out in previous sections, reducing the modal volume, V, of the resonator, thus optimizing the Q/V value, further enhances the sensitivity. High Q/V values are observed in photonic crystal cavities, which are an interesting alternative to micospherical and microtoroidal optical resonators (Fig. 7f).

High-Q Resonators

Disk and Ring Resonators

Ring and disk resonators (Fig. 7c, d) are perhaps the most commonly implemented resonator geometry due to their ease of fabrication

[14]. The 2-D and quasi 1-D geometries can be manufactured with photolithographic techniques and in different material systems. For example, rings and disks can be fabricated in SiO_2, Si, nanocrystalline diamond, and organic polymers. Photolithography allows for fabrication of large sensor arrays with multiplexed detection capability [15]. Within such an array, individual sensors are used as internal references to eliminate noise from thermal fluctuations and nonspecific molecular binding. This is a significant advantage over systems utilizing only one resonator and greatly simplifies experimentation and data analysis. Due to their ease of fabrication using standard "CMOS techniques," ring and disk resonators are ideal candidates for integration into standardized platforms compatible with portable electronic devices and commercial bench-top systems. Microfabricated ring/disk resonator arrays are also quite easily incorporated with microfluidics due the planar nature of the substrates on which they are fabricated. This feature of chip-based resonators is critical for achieving sensitivities and signal-to-noise ratios that surpass those of currently available systems, as well as allowing handling of the increasingly small samples found in lab-on-chip applications. Commercial systems based on silicon ring-resonator arrays are currently in development and should be available in the near future.

Ring and disk resonators typically possess a low modal volume, which as discussed previously is key to ensure a large resonant shift for detection of molecules. However, rings and disks tend to exhibit relatively low Q-factors, especially after immersion in water ($Q \approx 10^{3-4}$), which makes determination of resonant frequency less precise and limits the ultimate sensitivity of resonators with these geometries. While such Q-factors are sufficient for many biosensing applications, these geometries may not be suitable for single virus or single molecule detection. For the time being, however, ring and disk resonators provide the most attractive option as sensing elements for commercially available bench-top and portable systems.

Liquid Core Optical Ring Resonators (LCORRs)

A relatively new class of microcavity resonators is the liquid core optical ring resonator (LCORR) [16]. This class of resonator utilizes the cylindrical geometry of microcapillaries to excite WGM resonances about the inner and outer circumference of the thin microcavity wall (Fig. 7e). Microcapillaries are fabricated from quartz or glass preforms by drawing under high heat to a final outer capillary diameter of ~100 μm, and a wall thickness of a few microns. The LCORR can then be mounted on a waveguide or tapered fiber, which is used to couple light to the LCORR and excite the WGM resonances. The microcapillary sensors are sensitive to molecules binding to the inside of the capillary and one of the main advantages of this approach is the ease of fluid handling. In the capillary configuration of WGM biosensors, it is not necessary to create a completely separate flow cell, or mount fluidic channels on top of a wafer that holds the resonator structures. Instead, LCORRs can be arranged in parallel on waveguides to create sensor array already integrated with microfluidics. By employing multiple waveguides along a single capillary, it is possible to interrogate the region about each waveguide as an independent sensor. This linear configuration is well suited to such applications as capillary electrophoresis, where sensing regions at defined intervals can be used to monitor the speed of a profile of analyte passing a region. Additionally, LCORRs support high-Q resonances (~10^5), which provide them with the necessary sensitivity to compete with established biosensing methods such as SPR and QCM. Investigation into applications of LCORRs is on going and remains an intriguing alternative to other geometries.

Ultrahigh-Q Resonators

Toroids

Silica toroidal resonators (Fig. 7b) exhibit the highest Q-values measured in water [17] in excess of $Q = 10^8$. These types of resonators are fabricated from layers of silica (glass) grown on crystalline silicon substrate, etched into discs with hydrofluoric acid, and then undercut with, i.e., xenon difluoride etch to form a free standing silica discs held on a silicon post. Reflow of the silica disk is achieved in a self-terminating process by heating the silica with a CO_2 laser results in a

toroidal (donut) shape. Toroidal resonators exhibit modal volumes lower than those of microspheres of comparable diameter, and exhibit ultrahigh Q-factors due to an atomically smooth surface produced by reflow of glass. Thus, the toroid geometry boasts a high sensitivity and is a good candidate for single molecule detection. To date, only high-Q toroidal resonators made from silica material in the sensing region have been reported, as successful fabrication relies on the reflow of glass under high heat. Therefore, these geometries are difficult to fabricate using higher refractive index materials, such as sapphire, diamond, and alkali halides. Silica-based toroidal resonators are amenable to chemical functionalization by standard silane chemistry as well as other means of covalent and noncovalent modification and are thus useful as ligand-specific biosensors. This resonator geometry can be potentially integrated into on-chip arrays, similarly as described for disks and rings. However, integration with waveguides or optical fiber requires secondary fabrication steps and still remains a challenge. Whether or not practical implementation of microtoroid arrays with integrated waveguides will be achieved still remains a topic of research. As such, they are excellent for bench-top applications that require ultrahigh sensitivity.

Spheres and Spheroids

Next to rings, spherical resonators are perhaps the most common resonator geometry employed. Spherical silica resonators are very easy to fabricate by melting the end of an optical fiber and exhibit ultimate Q-factors of up to 10^{10}. Light can be coupled to a sphere, or sphere-on-stem by tapered fiber, waveguide, prism, or free beam coupling allowing a wide range of possibilities in implementation. Additionally, free standing spheres can be fabricated using a variety of alternative procedures and materials. Silica microspheres are most common, but reports have demonstrated the utility of spheres made from or coated with high refractive index polymers such as polystyrene. Free standing spheres can be used in top-down fabrication schemes that allow a great deal of flexibility in experimental configuration. In addition, the symmetry of a sphere allows for easy analysis and quantification of the WGM resonance signals. The inherent flexibility and simplicity of the spherical geometry makes it perhaps the most promising for future implementations of WGM biosensors. Integration with microfluidics and multiplexing remains challenging, although there are proposals for generating large arrays of microspheres, which can be read out in parallel. Microsphere resonators are currently mainly used in bench-top proof-of-principle experimental applications. However, as the technology develops, the spherical resonator will find applications to commercial instruments as well.

Applications

Protein Detection

Many applications of WGM biosensors are focused on the label-free detection of proteins from solution [18]. In 2002, Vollmer and coworkers [4] demonstrated the quantification of the thickness and surface density of a layer of bovine serum albumin (BSA) adsorbed on a microsphere functionalized with the silane aminopropyltrimethoxysilane (APTES). This result provided the proof-of-principle for using resonance frequency shifts in WGM for real-time biosensing applications. Furthermore, this experiment demonstrated that the glass surface resonators could be functionalized with various recognition elements such as proteins and antibodies, and showed the utility of the reactive sensing principle (Eqs. 1 and 2).

Due to the remarkable sensitivity of WGM biosensors, they hold great promise for the detection of low copy numbers of biomarkers that are indicative of early stage disease. Andrea and coworkers published work demonstrating the detection of molecule binding events for Interleukin-2 (IL-2), a molecule indicative of immune response [17]. This combination of sensitivity and the ability to create highly specific sensor surfaces makes WGM sensors ideal candidates for sensor elements in point of care devices and fieldable instrumentation. A similar approach was also demonstrated by Washburn and coworkers in 2009 who reported the detection of the cancer marker carcinoembryonic

antigen (CEA) from clinical samples containing serum [19]. This was the first demonstration of detection of a disease marker from a complex biological sample by WGM frequency shift.

DNA Detection

In the post-genomic era, the ability to detect and sequence DNA from low copy number is critical for clinically relevant applications. Despite advances in polymerase chain reaction (PCR) methods, it is still desirable to be able to detect and discriminate DNA sequences directly from unlabelled samples. In this respect, WGM biosensors offer the potential to detect down to single-molecule binding events for DNA using resonators functionalized with complementary strands of DNA. Vollmer and coworkers [20] demonstrated the sequence-based discrimination of DNA binding to complementary sequences immobilized on a glass microsphere [20]. In these experiments, two microspheres functionalized with different complementary sequences were coupled to a single waveguide. Solutions with target DNA sequences were added and the response from the microspheres was simultaneously measured. It was shown that a single base-pair mismatch could be discriminated. Thus, WGM is a highly suitable technique for the sequence-based detection and discrimination of DNA molecules.

Virus Particle Detection

In addition to the detection of protein and DNA molecules, it is very desirable to be able to detect single particles such as viruses. As a single virus particle has the potential to infect a cell and replicate itself billions of times in a relatively short period of time, it is obviously worthwhile to have sensors capable of detecting virions on the ultimate single particle level. In 2008, Vollmer and coworkers [7] demonstrated the ability to detect the binding and unbinding of single particle of the Influenza A (InfA) virus to an unmodified microsphere.

Sensor Arrays

Due to the ease of fabrication of these types of sensing elements, it is possible to incorporate resonators of tailored geometries into large arrays that can be individually addressed. This is perhaps one of the most promising aspects of WGM biosensors, in that they not only provide sensitivity and specificity better than currently available optical methods, but can also apply these benefits to potentially massive arrays of sensing elements. Washburn and coworkers have taken a first step in this direction by applying arrays of ring resonators for the detection of CEA from complex biological samples [19]. Here it was shown that an array of 16 individually addressable resonators could be interrogated in real time for the detection of their target antigen. Furthermore, due to the redundancy provided by an array of sensors, it was possible to utilize a subset of the resonators as internal controls for compensating for fluctuations in temperature.

Future Directions

The future is bright for the application of WGM biosensors. With the myriad benefits associated with the method and the proof-of-concept work performed to date, much rests on the ability to control surface functionalization and engineering of the associated sensor components (laser source, detectors, electronics, sample handling, etc.) to miniaturize the systems to a truly portable handheld format. Similar as electronic devices that have become smaller and more portable, it is inevitable that the components of WGM biosensors will one day be packaged in a form small enough for a primary health care giver or technician to use devices of this type on a regular basis at either the bed side of a patient, or in the field testing for biological agents or environmental contaminants. Further refinement of quantification methods and sensor design (geometry, materials, etc.) will provide unmatched sensitivity and accuracy in point-of-care devices.

Cross-References

- ▶ Biosensors
- ▶ Nanostructured Materials for Sensing
- ▶ Nanostructures for Photonics
- ▶ Organic Sensors

References

1. Heebner, J., Grover, R., Ibrahim, T.: Optical Microresonators: Theory, Fabrication, and Applications, 1st edn. Springer, London (2007)
2. Bloch, S.C.: Introduction to Classical and Quantum Harmonic Oscillators, Har/Dsk ed. Wiley-Interscience, Hoboken (1997)
3. Vahala, K.: Optical Microcavities. World Scientific, Singapore (2005)
4. Vollmer, F., Braun, D., Libchaber, A., Khoshsima, M.: Appl. Phys. Lett. **80**, 4057 (2002)
5. Vollmer, F., Arnold, S.: Nat. Methods **5**, 591–596 (2008)
6. Arnold, S., Khoshsima, M., Teraoka, I., Holler, S., Vollmer, F.: Opt. Lett. **28**, 272–274 (2003)
7. Vollmer, F., Arnold, S., Keng, D.: Proc. Natl. Acad. Sci. U. S. A. **105**(52), 20701–20704 (2008)
8. Hinchcliffe, A., Munn, R.W.: Molecular Electromagnetism. Wiley, Chichester (1985)
9. Hill, T.: An Introduction to Statistical Thermodynamics. Dover Publications, New York (1987)
10. Noto, M., Vollmer, F., Keng, D., Teraoka, I., Arnold, S.: Opt. Lett. **30**(5), 510–512 (2005)
11. Zhu, J., Ozdemir, S.K., Yun-Feng, X., Li, L., He, L., Chen, D.-R., Yang, L.: Nat. Photon. **4**, 4 (2010)
12. Arnold, S., Keng, D., Shopova, S., Holler, S., Zurawsky, W., Vollmer, F.: Opt. Express **17**, 6230–6238 (2009)
13. Francois, A., Himmelhaus, M.: Sensors **9**(9), 6836–6852 (2009)
14. De Vos, K., Bartolozzi, I., Schacht, E., Bienstman, P.: Opt. Express **15**, 7610–7615 (2007)
15. Bailey, R.: Future Sci. Bioanalysis **1**(6), 1043–1047 (2009)
16. Fan, X., White, I., Shopova, S., Zhu, H., Suter, J.: Anal. Chim. Acta **620**, 8–26 (2008)
17. Armani, A., Kulkarni, R., Fraser, S.: Science **317**, 783 (2007)
18. Matsko, A., Savchenkov, A., Strekalov, D.: IPN progress report (2005)
19. Washburn, A.L., Gunn, L.C., Bailey, R.C.: Anal. Chem. **81**(22), 9499–9506 (2009)
20. Vollmer, F., Arnold, S., Braun, D., Teraoka, I., Libchaber, A.: Biophys. J. **85**(3), 1974–1979 (2003)

Wine-Glass Mode Resonators

► Lamé Mode MEMS Resonators

Wireless Propulsion of Helical Nanobelt Swimmers

► Remotely Powered Propulsion of Helical Nanobelts

Worm Chip

► Lab-on-a-Chip for Studies in *C. elegans*

X

X-Ray Diffraction with Micron Sized X-Ray Beams

► Selected Synchrotron Radiation Techniques

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Y

Young's Modulus

► Nanomechanical Properties of Nanostructures

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Z

Zinc Oxide Nanoparticles (ZnO-NPs)

► Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

List of Entries

A

Ab Initio DFT Simulations of Nanostructures
Ab Initio Modeling and Simulation of Metal Nanoparticles
Ab Initio Molecular Dynamics
AC Dielectrophoresis and Dipolar Interactions for Particle Manipulation
AC Electrokinetic Manipulation
AC Electrokinetics
AC Electrokinetics of Colloidal Particles
AC Electrokinetics of Nanoparticles
AC Electrokinetics of Sub-micrometer Particles
AC Electroosmosis: Basics and Lab-on-a-Chip Applications
ac-Calorimetry
Accumulator
Acoustic Contrast Factor
Acoustic Enrichment
Acoustic Nanoparticle Synthesis for Applications in Nanomedicine
Acoustic Particle Agglomeration
Acoustic Trapping
Acoustic Tweezers
Acoustophoresis
Active Carbon Nanotube-Polymer Composites
Active Nanoantenna System
Active Plasmonic Devices
Adhesion
Adhesion in Insects
Adsorbate Adhesion
Aerosol-Assisted Chemical Vapor Deposition (AACVD)
AFM
AFM Force Sensors
AFM in Liquids
AFM Probes
AFM Tips
AFM, Noncontact Mode
AFM, Tapping Mode
Aging
ALD
Alteration
Alum
Aluminum Salts
Aluminum Vaccine Adjuvant
Aluminum-Based Nano-adjuvants
Amphibian Larvae
Amplitude-Phase-Distance (APD) Curves
Anamorphic Particle Image Tracking
Anchor Loss in MEMS/NEMS
Anderson Localization
Angiogenesis
Animal Coloration
Anodic Arc Deposition
Antifogging Properties in Mosquito Eyes
Anti-icing
Anti-Stokes Luminescence
Aperture Trap
Applications of Nanofluidics
Arc Discharge
Arrays
Arthropod Strain Sensors
Artificial Muscles
Artificial Retina: Focus on Clinical and Fabrication Considerations
Artificial Synapse
Artificial Vision
Arts-Based Research
Astigmatic Microparticle Imaging
Atmospheric Pressure Chemical Vapor Deposition (APCVD)
Atomic Cluster

© Springer Science+Business Media Dordrecht 2016
B. Bhushan (ed.), *Encyclopedia of Nanotechnology*,
DOI 10.1007/978-94-017-9780-1

Atomic Force Microscopy
Atomic Force Microscopy in Liquids
Atomic Layer Chemical Vapor Deposition
Atomic Layer Chemical Vapor Deposition (ALCVD)
Atomic Layer CVD (AL-CVD)
Atomic Layer Deposition
Atomic Layer Deposition (ALD)
Atomic Layer Epitaxial (ALE)
Atomic Layer Epitaxy (ALE)
Atomistic and Multiscale Modeling of Metal Nanoparticles
Atomistic Simulations
Attocalorimetry
Automatic Data Analysis Workflow for RNAi

B

Bacterial Electrical Conduction
Band Alignment
Band Pass Filters
Bandpass Tunable Filter
Basic MEMS Actuators
Batteries
Bendable Electronics
Bending Strength
Bilayer Graphene
Binary Systems
Bioaccessibility/Bioavailability
Bioactive Materials
Bioadhesion
Bio-adhesion
Bioadhesives
Biocalorimetry
Biocharacterization
Bioderived Smart Materials
Biodistribution
Bio-FET
Biofilms in Microfluidic Devices
Biognosis
Bioimpedance Sensing
Bio-inspiration
Bioinspired and Bacterial Nanorobotics
Bioinspired CMOS Cochlea
Bioinspired Microneedles
Bioinspired Synthesis of Nanomaterials
Biological Breadboard Platform for Studies of Cellular Dynamics
Biological Nano-crystallization
Biological Photonic Structures
Biological Platform
Biological Sensors
Biological Structural Color
Biologicalcell Manipulation
Biomedical Applications
BioMEMS/NEMS
Biomimetic
Biomimetic Energy Conversion Devices
Biomimetic Flow Sensors
Biomimetic Infrared Detector
Biomimetic Mosquito-Like Microneedles
Biomimetic Muscles and Actuators
Biomimetic Muscles and Actuators Using Electroactive Polymers (EAP)
Biomimetic Synthesis
Biomimetic Synthesis of Nanomaterials
Biomimetics
Biomimetics of Marine Adhesives
Biomimetics of Optical Nanostructures
Biomimicked Microneedles
Biomimicry
Biomolecular Mechanics
Biomolecules
Bionic Ear
Bionic Eye
Bionic Microneedles
Bionics
Bio-optics
Biopatterning
Bio-photonics
Biophysics
Bioprobes
Biornametics: Architecture Defined by Natural Patterns
Biosensing
Biosensors
Bond-Order Potential
Bone Remodeling
Bone Turnover
Boron- and/or Nitrogen-Doped Carbon Nanotubes
Boron Nitride Nanotubes (BNNTs)
Bottom-Up Fabrication of Plasmonic Materials
Bottom-Up Nanofabrication
Bottom-Up Processing
Boundary Lubrication
Brain Implants
Brushes

Buckminsterfullerene
Bulk Acoustic Wave MEMS Resonators
Bundles

C

C_{60}
Cancer
Cancer Modeling
Capacitive MEMS Switches
Capillarity Induced Folding
Capillary Flow
Capillary Origami
Carbon MEMS
Carbon Microelectromechanical Systems
Carbon Nanotube (CNT) Arrays
Carbon Nanotube Materials
Carbon Nanotube NEMS
Carbon Nanotube-Metal Contact
Carbon Nanotube-Metal Interface
Carbon Nanotubes
Carbon Nanotubes (CNTs)
Carbon Nanotubes for Chip Interconnections
Carbon Nanotubes for Interconnects in Integrated Circuits
Carbon Nanotubes for Interconnects in Microprocessors
Carbon Nanowalls
Carbon-Nanotubes
Car–Parrinello Molecular Dynamics
Carrier-Free Electrophoresis
Catalyst
Catalytic Bimetallic Nanorods
Catalytic Chemical Vapor Deposition (CCVD)
Catalytic Janus Particle
Cathodic Arc Deposition
Cavity Optomechanics
Cell Adhesion
Cell Adhesion and Detachment
Cell Micro-patterning
Cell Migration
Cell Patterning
Cellular and Molecular Toxicity of Nanoparticles
Cellular Dynamics
Cellular Electronic Energy Transfer
Cellular Imaging
Cellular Mechanisms of Nanoparticle Toxicity
Cellular Toxicity
Chaos
Characterizations of Zinc Oxide Nanowires for Nanoelectronic Applications
Charge Transfer
Charge Transfer on Self-Assembled Monolayer Molecules
Charge Transport in Carbon-Based Nanoscaled Materials
Charge Transport in Self-Assembled Monolayers
Chem-FET
Chemical Beam Epitaxial (CBE)
Chemical Blankening
Chemical Dry Etching
Chemical Milling and Photochemical Milling
Chemical Modification
Chemical Solution Deposition
Chemical Vapor Deposition (CVD)
Chemical Vapor Machining (CVM)
Chemical-Assisted Ion Beam Etching (CAIBE)
Chemistry of Carbon Nanotubes
Chemotherapeutic
Chitosan
Chitosan Nanoparticles
Clamping Loss
Clinical Adhesives
Cluster
C-MEMS
CMOS (Complementary Metal-Oxide-Semiconductor)
CMOS MEMS Biosensors
CMOS MEMS Fabrication Technologies
CMOS-CNT Integration
CMOS-MEMS
CMOS-MEMS Resonators
CNT assembly
CNT Biosensor
CNT Handling and Integration
CNT NEMS
CNT Resonator
CNT-FET
Coarse-Grained and Hybrid Simulations of Nanostructures
Coarse-Grained Molecular Dynamics
Cochlea Implant
Cold Field Electron Emission from Nanostructured Materials
Cold-Wall Thermal Chemical Vapor Deposition
Compliant Mechanisms
Compliant Systems

Composite Materials
Computational Chemistry for Drug Discovery
Computational Micro-/Nanofluidics: Unifier of Physical and Natural Sciences and Engineering
Computational Studies
Computational Study of Nanomaterials: From Large-Scale Atomistic Simulations to Mesoscopic Modeling
Computational Systems Bioinformatics for RNAi
Computer Modeling and Simulation of Materials
Computer Modeling of Electrochemical Systems
Concentration Polarization
Concentration Polarization at Micro-/Nanofluidic Interfaces
Conductance Injection
Conduction Mechanisms in Organic Semiconductors
Conductivity of Metal Nanowires Studied by Infrared Plasmon-Polariton Spectroscopy
Confocal Laser Scanning Microscopy
Confocal Scanning Optical Microscopy (CSOM)
Conformal Electronics
Contact Angle
Contour-Mode Resonators
Contraception
Contrast Enhancement
Cooling of Electronic Components
Core/Shell Nanostructures
Coupled Mode Theory
Coupling
Coupling Clamp
CPMD
Creep
Crystallite
Cuticle
Cylindrical Gold Nanoparticles

D

Decoherence and Dissipation in Nanoelectronic Devices
Decoration of Carbon Nanotubes
Defects
Deicing
Densification
Density
Density-Matrix Theory
Deoxyribonucleic Acid
Deposited Metal Nanostructures
Dermal and Transdermal Delivery
Detection of Nanoparticle Biomarkers in Blood
Detection of Nanoparticulate Biomarkers in Blood
Device Modeling
DFT-Based Molecular Dynamics
DFT-Based Simulations of Electrified Interfaces
Diamond Formation in Graphene Nanoplatelets, Carbon Nanotubes, and Fullerenes Under Spark Plasma Sintering
Diamond Synthesis
Dielectric Force
Dielectrophoresis
Dielectrophoresis Directed Nanocolloidal and Supramolecular Assembly
Dielectrophoresis of Nucleic Acids
Dielectrophoretic Assembly
Dielectrophoretic Nanoassembly of Nanotubes onto Nanoelectrodes
Dielectrophoretic Phenomena
Differential Scanning Calorimetry
Digital Microfluidics
Dipolar Particle Chaining
Dipole–Dipole Particle Interaction
Dip-Pen Nanolithography
Direct Laser Writing
Directed Assembly
Directed Self-Assembly
Disjoining Pressure and Capillary Adhesion
Dislocation Dynamics
Disordered Plasmonics
Dispersion
Dissipative Cavity
Dissipative Particle Dynamics, Overview
Dissociated Adsorption
DNA Computing
DNA FET
DNA from First Principles
DNA Manipulation
DNA Manipulation Based on Nanotweezers
DNA Origami as Programmable Nanofabrication Tools
DOLLOP (Dynamically Ordered Liquid-Like Oxyanion Polymer)
Doping in Organic Semiconductors
Doping of Organic Semiconductors
Double-Walled Carbon Nanotubes (DWCNTs)
Droplet Microfluidics

Drug Delivery
Drug Delivery and Encapsulation
Drug Delivery System
Drug Design
Drug Targeting
Dry Adhesion
Dry Etching
Dry Etching Processes
Dual-Beam Piezoelectric Switches
Duffing
DUV Lithography
DUV Photolithography and Materials
Dye Sensitized Solar Cells
Dye-Doped Nanoparticles in Biomedical Diagnostics
Dye-Doped Nanospheres in Biomedical Diagnostics
Dynamic Clamp
Dynamic Force Measurements on Modified Surfaces with AFM
Dynamic Force Spectroscopy

E

E-Beam Lithography (EBL)
ECM
Ecotoxicity
Ecotoxicity of Inorganic Nanoparticles: From Unicellular Organisms to Invertebrates
Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment
Ecotoxicology of Carbon Nanotubes Toward Amphibian Larvae
Ecotribology
EC-STM
Effect of Surface Modification on Toxicity of Nanoparticles
Effective Media
Elastic and Plastic Behavior of Materials with Nanometric Dimensions
Elastic Modulus Tester
Elasto-capillary Folding
Electric Cell-Substrate Impedance Sensing (ECIS)
Electric Double Layer Capacitor
Electric Field–Directed Assembly of Bioderivatized Nanoparticles
Electrical Conductivity
Electrical Impedance Cytometry
Electrical Impedance Spectroscopy
Electrical Impedance Tomography for Single-Cell Imaging
Electric-Field-Assisted Deterministic Nanowire Assembly
Electrocapillarity
Electrochemical Energy Storage and Conversion
Electrochemical Interfaces for Energy Storage and Conversion
Electrochemical Machining (ECM)
Electrochemical Scanning Tunneling Microscopy
Electrode–Organic Interface Physics
Electrodics from First Principles
Electrohydrodynamic Forming
Electrokinetic Fluid Flow in Nanostructures
Electrokinetic Manipulation
Electrokinetics
Electron Beam Evaporation
Electron Beam Lithography (EBL)
Electron Beam Physical Vapor Deposition (EBPVD)
Electron Microscopy of Interactions Between Engineered Nanomaterials and Cells
Electron Transfer
Electron Transport in Carbon Nanotubes and Graphene Nanoribbons
Electron-Beam-Induced Chemical Vapor Deposition
Electron-Beam-Induced Decomposition and Growth
Electron-Beam-Induced Deposition
Electronic and Optical Properties Calculations
Electronic and Optical Properties of Oxides Nanostructures by First-Principles Approaches
Electronic Contact Testing Cards
Electronic Expression
Electronic Pharmacology
Electronic Properties
Electronic Quantum Transport
Electronic Structure Calculations
Electronic Structure of Nucleic Acids
Electronic Transport in Carbon Nanomaterials
Electronic Visual Prosthesis
Electroplating
Electroresponsive Polymers
Electrospinning
Electrostatic Actuation of Droplets
Electrostatic MEMS Microphones
Electrostatic RF MEMS Switches

Electrothermal Actuators
Electrothermomechanical Actuators
Electrowetting
Electrowetting-on-Dielectric (EWOD)
Embryotoxic
Emission Enhancement
Energy-Level Alignment
Engineered Nanoparticles
Environmental Impact
Environmental Impact Assessment
Environmental Toxicology
Epiretinal Implant
Epiretinal Prosthesis
EUV Lithography
Evanescent-Wave Particle Visualization in Nano- and Microchannels
Excitonic Solar Cell
Exoskeleton
Exposure and Toxicity of Metal and Oxide Nanoparticles to Earthworms
Extreme Ultraviolet Lithography (EUVL)

F

Fabrication of Graphene or Graphene Fabrication
Fate of Manufactured Nanoparticles in Aqueous Environment
Fatigue Strength
FDTD
Femtocalorimetry
Fetus
Fiber Reinforced Composite
Fibrillar Adhesion
FIB-SEM
Field Electron Emission from Nanomaterials
Field-Driven Particle Assembly
Filling of Carbon Nanotubes
Finite Element Methods for Computational Nano-optics
Finite-Difference Frequency-Domain Technique
Finite-Difference Time-Domain Technique
Finite-Element Analysis (FEA)
First Principles Calculations
First Principles Molecular Dynamics
First-Principle Calculations
First-Principles Approaches
First-Principles Quantum Monte Carlo
Flapping-Wing Flight
Flash Evaporation Liquid Atomization
Flexible Electronics
Flexible Sensors
Flexure Mechanisms
Flexures
Floccules
Fluctuating Hydrodynamics
Foams
Focused-Ion-Beam Chemical-Vapor-Deposition (FIB-CVD)
Focused-Ion-Beam-Induced Deposition (FIB-ID)
Force Measurement
Force Modulation in Atomic Force Microscopy
Force Spectroscopy
Force-Detected Magnetic Resonance
Forests
Fracture Toughness
Friction
Friction Force Microscopy
Friction-Reducing Sandfish Skin
Fuel Cell
Fullerenes for Drug Delivery
Functionalization
Functionalization of Carbon Nanotubes
Functionalized Nanomaterials
Fundamental Properties of Zinc Oxide Nanowires

G

Gallium Arsenide (GaAs)
Gamma-Butyrolactone (GBL)
Gammatone Filters
Gas Etching
Gas-Phase Molecular Beam Epitaxy (Gas-Phase MBE)
Gas-Phase Nanoparticle Formation
Gecko Adhesion
Gecko Effect
Gecko Feet
Gel Chemical Synthesis
Genotoxicity of Nanoparticles
Gold Nanorods
Graphene
Graphene Laser Writing
Graphene Nanostructures for Memristive Devices
Graphene NEMS
Graphene Synthesis
Graphite
Greases
Green Tribology and Nanoscience

Green's Function Integral Equation Method
Growth of 1-D Oxide Nanostructures
Growth of Carbon Nanotubes
Growth of CNTs
Growth of Oxide Nanowires
Growth of Silica Nanowires
Growth of Silicon Dioxide Nanowires
GW and BSE Approaches
Gyroscopes

H

Hardbake (HB)
Hardness
Hard-Tip Soft-Spring Lithography
Harnessing Disorder at the Nanoscale
Heat
Heat Capacity
Heat Conduction
Heat Conductivity
Heat Removal
Heat Transfer Enhancement
Heat Transfer in Semiconductor Nanostructures
Heatuators
Helical Nanobelt
Heterogeneous Catalysts
Heterogeneous Integration
HGN
Hierarchical Mechanics
High Energy Synchrotron Radiation and Its Impact on Characterizing Nanoparticles
High Index Fluid Immersion
High-Density Testing Cards
High-Energy X-Rays
Higher-Order Plasticity Theory
High-Pressure Carbon Monoxide (HiPCO)
HIV Quantification
Hollow Gold Nanoshells
Hollow Gold Nanospheres
Hot Filament Chemical Vapor Deposition (HFCVD)
Hot-Wall Thermal Chemical Vapor Deposition
How to Synthesize/Make/Manufacture/Fabricate/Produce Graphene
Hybrid Network Method
Hybrid Optoelectric Manipulation
Hybrid Opto-electric Technique
Hybrid Photovoltaics
Hybrid Solar Cells
Hydrophilic
Hydrophobic
Hyperelasticity
Hyperthermia Treatment

I

Icephobicity
Imaging the Human Body Down to the Molecular Level
Immersion Lithography
Immersion Lithography Materials
Impedance Cytometry
Impedance Measurement of Single Cells
Impedance Sensing Techniques for Biomedical Microdevices
Impedance Spectroscopy for Single-Cell Analysis
Implementing Alternating Nanolaminates in Trenched Energy Storage Systems
Imprint Lithography
In Situ STM
In Situ Transmission Electron Microscopy: Mechanical Testing
In Vitro and In Vivo Toxicity and Pharmacokinetics of Silver Nanoparticles
In Vitro Toxicity
In Vivo Microscopy Analysis
In Vivo Toxicity of Carbon Nanotubes
In Vivo Toxicity of Titanium Dioxide and Gold Nanoparticles
Indentation
Induced Charge Electroosmosis
Inductively Coupled-Plasma Chemical Vapor deposition (ICP-CVD)
Inertial Sensors
Influence of Defects in Photocatalysis
Influence of Surface Engineering on Toxicity of Nanoparticles
Influence of Surface Functionalization on Toxicity of Nanoparticles
Infrared Plasmonics Beyond Noble Metals
Injectable Hydrogel
Injectable In Situ Forming Gel
Inorganic Nanoparticle Radioenhancers for Oncology
Inorganic Nanoparticle Radiosensitizers for Cancer Therapies
Inorganic Nanoparticle Radiosensitizers for Oncology

Insect Adhesion
Insect Flight and Micro Air Vehicles (MAVs)
Insect Infrared Sensors
Insect-Inspired Visual Systems and Visually Guided Behavior
Insects
Integrated Approach for the Rational Design of Nanoparticles
Integrated Biosensors
Integrated Micro-acoustic Devices
Integration
Integration of Nanostructures Within Microfluidic Devices
Intelligent Drug Delivery Microchips
Interaction of the Immune System with Nanoparticles
Interface Excess Free Energy
Interfacial Charging Effect
Interfacial Energy and Chemical Potential at Nanoscale
Interfacial Investigation of Protein Films Using Acoustic Waves
Interference Modulator (IMOD) Technology
Intravital Microscopy Analysis
Ion Beam
Ion Beam Etching (IBE)
Ion Beam Figuring (IBF)
Ion Beam Smoothing (IBS)
Iridescent Colors (Organisms Other Than Animals)
Isothermal Calorimetry

K

Kelvin Probe Force Microscopy
Key Point Detection

L

Lab-on-a-ChipLab-on-a-Chip (LOC)
Lab-on-a-Chip for Studies in *C. elegans*
Lab-on-a-Chip Whole-Cell Biosensor
Lamb-Wave Resonators
Lamé Mode MEMS Resonators
Lamé Mode Resonators
Laminates: Multilayered Structures
Landmark Detection
Large Unilamellar Vesicle (LUV)
Laser Beam Machining
Laser Cutting
Laser Interactions with Particles in Solution
Laser Machining
Laser Micromachining
Laser Scanning Confocal Microscopy
Laser Scribing
Laser Tweezers
Laser Tweezers Using Nanoapertures in Metal Films
Laser Writing
Laser-Assisted Tip-Based Heating and Fabrication
Laser-Assisted Tip–Substrate Interaction
Laser-Reduced Graphene Oxide
Laser-Scribed Graphene
Lateral Force Microscopy (LFM)
Lateral Shear Mode Resonators
Lateral Thinking
Laterally Vibrating Piezoelectric Resonators
Lattice Boltzmann Methods for Nanofluidics
LC Filters
Left-Handed Materials
Lesson from Nature
Life Cycle
Ligand-Directed Gold-Phage Nanosystems
Light Emitting Diodes
Light Localization for Nano-optical Devices
Light-Element Nanotubes and Related Structures
Lipid Vesicles
Liposomes
Lipospheres
Lippmann-Schwinger Integral Equation Method
Liquid Atomization Through Vapor Explosion
Liquid Condensation
Liquid Film
Liquid Immersion Lithography
Liquid Meniscus
Lithographie, Galvanoformung, Abformung (LIGA)
LM10
Local Density of States
Local Surface Plasmon Resonance (LSPR)
Localized Plasmon Polaritons in the Infrared
Log-Domain
Logic-Embedded Vectors
Lotus Effect
Low Carrier Density Materials for Infrared Plasmonics
Low Fluid Drag Surface

Low-Power
Low-Pressure Chemical Vapor Deposition (LPCVD)
LSPR in Plasmonic Nanostructures: Theoretical Study with Application to Sensor Design
LSPR-Based Sensing
Lubrication
Lumbricidae/Oligochaeta/Earthworms

M

Macromolecular Crystallization Using Nano-volumes
Magnetic Assembly
Magnetic Dipole Emitters
Magnetic Dipole Transitions
Magnetic Electromagnetic Radiation
Magnetic Nanoparticles
Magnetic Nanoparticles for Biomedical Applications
Magnetic Nanostructures and Spintronics
Magnetic Particle Imaging
Magnetic Resonance Force Microscopy
Magnetic Resonance Imaging
Magnetic Self-Assembly
Magnetic Separation
Magnetic-Field-Based Self-Assembly
Magneto-optics and Plasmonics
Magnetoplasmonics
Magnetron Sputtering
Making Graphene
Manipulating
Manufacturing Graphene or Graphene Manufacturing
Manufacturing Silica Nanowires
Marine Bioinspired Adhesives
Marine Organisms
Massive
Materiomics
Mats
Maxwell–Bloch Simulations or Equations or Studies
Maxwell–Schrödinger Simulations (or Equations)
Maxwell–Wagner Effect
Mean Field Approaches to Many-Body Systems
Mechanical Behavior of Nanostructured Metals
Mechanical Characterization of Nanostructures
Mechanical Energy Dissipation
Mechanical Properties of Biological Materials
Mechanical Properties of Hierarchical Protein Materials
Mechanical Properties of Nanocrystalline Metals
Mechanical Properties of Nanostructures
Mechanism for Assembling Arrays of Rotary Nanoelectromechanical Devices
Mechanisms of Laser Ablation Synthesis of Nanoparticles in Liquid Solution
Mechanisms of Nanoparticle Generation by Laser Ablation in Liquids
Mechanoreceptors
Medical adhesives
Medical Nanorobotics
MEM Devices
Memory Resistors
Memristors
MEMS
MEMS (Micro-Electro-Mechanical Systems)
MEMS = Microelectromechanical Systems
MEMS Adaptive Optics
MEMS Capacitive Microphone
MEMS Condenser Microphone
MEMS Display Based on Interferometric Modulation Technology
MEMS Encapsulation
MEMS Fabricated Microwave Devices
MEMS High-Density Probe Cards
MEMS Microphone
MEMS Neural Probes
MEMS on Flexible Substrates
MEMS Packaging
MEMs Pumps
MEMS Resonant Infrared Detectors
MEMS Resonant Infrared Sensors
MEMS Resonant Thermal Detectors
MEMS Resonant Thermal Sensors
MEMS Resonators Integrated with Circuits
MEMS Vacuum Electronics
MEMS/IC Integrated Resonator Circuits
MEMS-Based Drug Delivery Devices
Meniscus Adhesion
Mesoscopic Modeling
Mesoscopic Simulation
Mesoscopic Simulations
Metal Alloy Nanoparticles
Metal Nanoparticles from First Principles
Metal Plating
Metallic Waveguides

Metallofullerene
Metalorganic Chemical Vapor Deposition (MOCVD)
Metal–Organic Interfaces
Metal-Organic Molecular Beam Epitaxy (MOMBE)
Metal-Oxide Nanostructures for Surface Enhanced Raman Spectroscopy
Micro Air Vehicles (MAVs)
Micro- and Nanofluidic Devices for Medical Diagnostics
Micro- and Nanomanipulation for Nanomanufacturing
Micro Coulter
Micro FACS
Micro Flow Cytometry
Micro Free-Flow Electrophoresis (μFFE)
Micro/Nano Flow Characterization Techniques
Micro/Nano Transport in Microbial Energy Harvesting
Micro-/Nanodevices
Micro-/Nanofluidic devices
Micro-/Nanomachines
Micro-/Nanostructured Icephobic Materials
Micro-/Nanosystems
Microactuators
Microarray
Microarrays/Microfluidics for Protein Crystallization
Microbial Aggregations
Microbial Electrolysis Cell
Microbial Fuel Cell
Microcalorimetry
Microcantilever Chemical and Biological Sensors
Microcavity Biosensor
Microchannel Flows as Heat Sinks
Microcontact Printing
Microcontact Transfer
Microelectrodes
Microelectromechanical Systems
Microelectromechanical Systems (MEMS)
Microfabricated MVED (Microwave Vacuum Electron Devices)
Microfabricated Preconcentrator-Focuser in Micro Gas Analyzers
Microfabricated Probe Technology
Microfluidic Device
Microfluidic Devices
Microfluidic Optomechanics
Microfluidic Pumps
Microfluidic Whole-Cell Biosensor
Microfluidics
Microinjection Molding
Micromachined 3-D Force Sensors
Micromachined Acoustic Devices
Micromachines
Micromachining
Micromechanical Resonators Fabricated Using CMOS-MEMS Technologies
Micromechanical Switches
μMist®
Micro-optomechanical Systems (MOMS)
Micropreconcentrator Technology for Portable Gas Chromatography System
Micropreconcentrators for Micro Gas Chromatography System
Micropumps
Microrobotics
Microrororobotics
Microscale Fluid Mechanics
Microscopic Modeling of Solid-State Quantum Devices
Microstereolithography
Microsystem
Microsystems
Microthermal Analysis
Microtrap
μVED (Micro Vacuum Electronic Devices)
Mini Pumps
mirasol® Displays
Modeling Nanoscale Heat Transport
Modeling of Batteries and Fuel Cells
Modeling Thermal Properties of Carbon Nanostructure Composites
Models for Tumor Growth
Modification of Carbon Nanotubes
Modulus
MOEMS/NOEMS
Molding
Molecular Beam Epitaxy (MBE)
Molecular Computation
Molecular Computing
Molecular Dynamics

Molecular Dynamics Method
Molecular Dynamics Simulations of Interactions Between Biological Molecules and Nanomaterials
Molecular Dynamics Simulations of Nano-biomaterials
Molecular Encapsulation
Molecular Layer Deposition (MLD)
Molecular Modeling
Molecular Modeling on Artificial Molecular Motors
Molecular Patterning
Molecular Plasmonics
Molecular Printing
Molecularly Thick Layers
Monolayer Lubrication
Monolithic Integration of Carbon Nanotubes
Monolithic Total Internal Reflection Biosensor
Mosquito Fascicle Inspired Microneedles
Moth-Eye Antireflective Structures
Motion Sensors
Mucoadhesion
Multifunctional Particles
Multilamellar Vesicle (MLV)
Multiphoton Fabrication
Multiscale Fluidics
Multiscale Modeling
Multiscale Simulations
Multistage Delivery System
Multistage Vectors
Multiwalled Carbon Nanotubes (MWCNTs)

N

Nano (Evanescent-Wave)-Particle Image Velocimetry
Nano Manipulation
Nanoalloy Simulation
Nanoantennas
Nanoassembly
Nanobelts
Nanobiosensors
Nanobonding
Nanocalorimeter
Nanocalorimetric Sensors
Nanocalorimetry
Nanocarriers
Nanoceramics by Spark Plasma Sintering
Nanochannels for Nanofluidics: Fabrication Aspects
Nanocluster
Nanocoil
Nanocombs
Nanocomposite Materials
Nanocomposites
Nano-concrete
Nanocrystalline Functional Materials in Bulk Form with Grain Size Below 50 nm
Nanocrystalline Materials
Nanodentistry
Nanodevices
Nano-ecotoxicology
Nano-effects
Nanoelectrodes
Nanoelectromechanical Systems
Nanoelectromechanical Systems (NEMS)
Nanoencapsulation
Nanoengineered Concrete
Nanoengineered Hydrogels for Cell Engineering
Nanofabrication
Nanofatigue Tester
Nano-FET
Nanofibrous Materials and Composites
Nanofluidic Channels
Nanofluidics
Nanofocusing of Light
Nanogap Biosensors
Nanograppers
Nanohardness
Nanohardness Tester
Nanohelix
Nanoimmunology
Nanoimprint Lithography
Nanoimprinting
Nanoindentation
Nanointerconnection
Nanojoining
Nanolayered Coatings
Nanolayers
Nanomachining
Nanomagnetism
Nanomanipulation for Cell Analysis
Nanomanipulation of Biocells
Nanomanufacturing with Magnetically Recorded Nanotemplates and Directed Self-Assembly

Nanomaterials
Nanomaterials and Nanoproducts
Nanomaterials for Electrical Energy Storage Devices
Nanomaterials for Excitonic Solar Cells
Nanomaterials for Sensors
Nanomechanical
Nano-mechanical Machining
Nanomechanical Properties of Nanostructures
Nanomechanical Resonant Sensors and Fluid Interactions
Nanomechanical Resonators Based on III–V Semiconductors
Nanomechanics
Nanomedicine
Nanometric
Nanomotors
Nano-optical Biosensors
Nano-optics Biosensing
Nano-optics, Nanoplasmonics
Nano-optomechanical Systems (NOMS)
Nanoparticle
Nanoparticle Cytotoxicity
Nano-particle Image Velocimetry
Nanoparticle Self-Assembly
Nanoparticle Tracking Analysis
Nanoparticle–Adsorbome
Nanoparticle–Biocorona
Nanoparticle–Biomolecular Corona
Nanoparticle–Protein Corona
Nanoparticles
Nanoparticles Toxicity Examined In Vivo in Animals
Nanoparticulate Materials and Core/Shell Structures Derived from Wet Chemistry Methods
Nanopatterned Substrata
Nanopatterned Surfaces for Exploring and Regulating Cell Behavior
Nano-patterning
Nanophotonic Fabrication
Nanophotonic Structures for Biosensing
Nano-piezoelectricity
Nano-pillars
Nanoplasmonic Enhancement
Nanoplasmonic Sensing
Nanoplasmonics Involved in Localized Photopolymerization
Nanopolaritonics
Nanopores
Nanopowders
Nanoprinting
Nanoribbons
Nanorobotic Assembly
Nanorobotic Manipulation of Biological Cells
Nanorobotic Platforms for Biological Studies
Nanorobotic Spot Welding
Nanorobotics
Nanorobotics for Bioengineering
Nanorobotics for Bioengineering and Medical Interventions
Nanorobotics for MEMS and NEMS
Nanorobotics for NEMS Using Helical Nanostructures
Nanorobotics for Synchrotron Radiation Applications
Nanorods
Nanoscaffold
Nanoscale Drug Vector
Nanoscale Fluid Dynamics
Nanoscale Fluid Mechanics
Nanoscale Heat Transport
Nanoscale Mass Transport
Nanoscale Metal Alloys
Nanoscale Metals
Nanoscale Particle
Nanoscale Printing
Nanoscale Properties of Solid–Liquid Interfaces
Nano-scale Structuring
Nanoscale Technologies
Nanoscale Thermal Analysis
Nanoscale Water Phase Diagram
Nanoscratch Tester
Nano-sized
Nanosized Nanocrystalline and Nanotwinned Metals
Nano-sized Particle
Nanoslits
Nanosoldering
Nanospring
Nanostructure
Nanostructure Field Effect Transistor Biosensors
Nanostructured Antireflective Surfaces
Nanostructured Functionalized Surfaces
Nanostructured Hydrogels
Nanostructured Materials
Nanostructured Materials for Sensing

Nanostructured SERS-Active Metal Oxides
Nanostructured Solar Cells
Nanostructured Thermoelectric Materials
Nanostructures
Nanostructures and Characteristics of Carbon Nanofibers
Nanostructures and Characteristics of Graphene and Graphene Oxides
Nanostructures Based on Porous Silicon
Nanostructures by Electrospinning
Nanostructures for Coloration (Organisms Other Than Animals)
Nanostructures for Energy
Nanostructures for Photonics
Nanostructures for Surface Functionalization and Surface Properties
Nanostructures: Nanodevices
Nanotechnology
Nanotechnology Applications in Polymerase Chain Reaction (PCR)
Nanotechnology in Cardiovascular Diseases
Nanotechnology in Dental Medicine
Nanotechnology in Pulmonary Disease
Nanothermodynamics
Nano-TiO_2 Life-Cycle Assessment Perspective
Nanotoxicity
Nanotoxicology
Nanotransfer Printing
Nanotribology
Nanotribology Applications in Microprojector Technology
Nanotube Adsorption Sensor
Nanotube Force Sensor
Nanotube Mass Sensor
Nanotube Resonant Sensor
Nanotube Resonator
Nanotubes
Nanotweezers
Nano-twinned Materials
Nanovaccinology
Nanowire FET Biosensor
Nanowire FET Simulations Based on the Nonequilibrium Green's Function Formalism
Nanowire-Based Field Effect Transistor (NWFET)
Nanowires
Nature Inspired Microneedle
Near Infrared
Near-Field Focusing and Heating
Near-Field Molecular Optics
Near-Field Optics
Near-Field Scanning Optical Microscopy (NSOM)
NEMS
NEMS Gravimetric Chemical Sensors
NEMS Gravimetric Sensors
NEMS Mass Sensors
NEMS Piezoelectric Switches
NEMS Resonant Chemical Sensors
NEMS Resonant Mass Sensors
Neural Activity-Dependent Closed Loop
Neural Interfaces
N-Methyl-2-pyrrolidone (NMP)
Nonadiabatic dynamics
Non-conventional Machining
Nonlinear and Parametric NEMS Resonators
Nonlinear Electrokinetic Transport
Nonlinear Optical Absorption and Induced Thermal Scattering Studies in Organic and Inorganic Nanostructures
Nonlinear Parametric MEMS
Nonlinearity
NS300
NS500
NTA
Nucleation of Silica Nanowires
Nucleic Acid Amplification

O

Officinal Skink
Oil-Repellency of Fish Scales
Oligonucleotide Amplification
Oligonucleotide Synthesis
OMFR
On-Chip Flow Cytometry
193-nm Lithography
Optical and Electronic Properties
Optical and Mechanical Nanostructured Coatings for Future Large-Scale Manufacturing
Optical Cavity Biosensor
Optical Frequency Magnetic Dipole Transitions
Optical Lithography
Optical Lithography by Liquid Immersion
Optical Properties
Optical Properties of Metal Nanoparticles
Optical Reflectance Spectroscopy of Ancient Paper
Optical Resonance Biosensor

Optical Response of Metal Particles
Optical Ring Resonator Biosensor
Optical Spectroscopy for Ancient Paper Diagnostic
Optical Techniques for Nanostructure Characterization
Optical Trap
Optical Tweezers
Optimization of Nanoparticles
Optimization of Plasmonic Substrates for SERS Application
Optoelectrically Enabled Multiscale Manipulation
Optoelectrofluidics
Optoelectrokinetic Technique
Optoelectronic Properties
Optofluidic Cavities
Optofluidic Resonators and Sensors
Optofluidics
Optomechanical Resonators
Optomechanics
Opto-mechano-fluidics
Organic Actuators
Organic Bioelectronics
Organic Bioelectronics (OBOE)
Organic Films
Organic Photovoltaics: Basic Concepts and Device Physics
Organic Sensors
Organic Solar cells
Organic/Inorganic Interactions
Organic–Inorganic Heterojunctions
Organic–Inorganic Photovoltaics
Organosilanes
Orientation Sensors
Oscillator
Osteogenic Materials
Osteoinduction Materials

P

Parameter-Free Study of Electrochemical Interfaces
Parametric
Parametrically Excited Nonlinear MEMS Resonators
Passive Pumping
Patterned Hydrogels
Patterning
PBPK Modeling
Perfluorocarbon Nanoparticles
Perfluoropolyethers
Petal Effect
Petroleum Lubricants
PFC Nanoparticles
PFOB Nanoparticles
Pharmacokinetics
Phase Transformation of Carbon to Diamond
Phonon Transport
Phononic Crystals
Photocatalysis
Photocatalytic
Photoetching
Photofabrication
Photolithography
Photomilling
Photonic Crystal Nanobeam Cavities
Photonics
Photopolymerization Induced by Localized Surface Plasmon Resonance
Photovoltaic
Photovoltaic Devices
Physical Colors
Physical Dry Etching
Physical Modification
Physical Vapor Deposition
Physical-Chemical Etching
Physicochemical Properties of Nanoparticles in Relation with Toxicity
Physics in Laser Near-Field Nanomanufacturing: Fundamental Understanding and Novel Probing
Phytotoxicity
Piezoelectric Effect at Nanoscale
Piezoelectric MEMS Switch
Piezoelectric Switches
Piezoresistance
Piezoresistive Effect
Piezoresistivity
Piezoresponse Force Microscopy and Spectroscopy
Placenta
Plasma
Plasma Etching
Plasma Grafting
Plasma Jet Machining (PJM)
Plasma Micro-contact Patterning
Plasma Patterning

Plasma Polymerization
Plasma-Assisted Chemical Etching (PACE)
Plasma-Enhanced Chemical Vapor Deposition (PECVD)
Plasmon Resonance Energy Transfer from Metallic Nanoparticles to Biomolecules
Plasmon Resonance Energy Transfer Nanospectroscopy
Plasmon-Assisted Two-Photon Polymerization
Plasmon-Enhanced Fluorescence
Plasmonic Amplification for Fluorescence Bioassays Utilizing Propagating Surface Plasmons
Plasmonic Amplification of Fluorescence
Plasmonic Fluorescence Biosensor
Plasmonic Photothermal Therapy (PPTT)
Plasmonic Photothermal Therapy with Gold Nanorods/Reduced Graphene Oxide Core/Shell Nanocomposites
Plasmonic Structures for Solar Energy Harvesting
Plasmonic Waveguides
Plasmonics
Plasmonics for Photovoltaics
Plasmonics for Solar Cells
Plasmonics in Self-Organized Media
Plasticity of Nanostructured Solids
Plasticity Theory at Small Scales
Plating
PmCP
Polarization-Induced Effects in Heterostructures
Polyethylene Carbonate (PEC)
Polymer Coatings
Polymer Nanocapsules
Polymer Pen Lithography
Polymer Sensors
Polymer-Based 3D Micro-/Nanofabrication by Laser Direct Writing
Polymeric Surface Modifications
Polymer–Metal Oxide Solar Cells
Polymethyl Methacrylate (PMMA)
Polypropylene Carbonate (PPC)
Postexposure Bake (PEB)
Precise Biopatterning with Plasma: The Plasma Micro-contact Patterning (PμCP) Technique
Pregnancy
Prenucleation Clusters
Preparation of Complex Oxide Unidimensional Nanostructures
PRINT®
Probe Technology of Scanning Probe Microscopy (SPM) and Its Fabrication Technology
Product of Random Matrices
Production of Gold Nanoparticles
Production of Graphene or Graphene Production
Properties Related to Applied Stress or Strain
Propylene Glycol Methyl Ether Acetate (PGMEA)
Prostheses
Protein Nano-crystallization
Pseudoelasticity
PuCP
Pulsed Focused Ultrasound
Pulsed HIFU
Pulsed High-Intensity Focused Ultrasound
Pulsed Laser Deposition (PLD)
Pyroelectricity
PμCP

Q

Quantum Cluster
Quantum Dot
Quantum Dot Nanophotonic Integrated Circuits
Quantum Dot Solar Cells
Quantum Dot Toxicity
Quantum Dynamics
Quantum Kinetics
Quantum Nanoplasmonics
Quantum System Near Metallic Particle

R

Radiofrequency
Random Matrix Theory
Rapid Atomic Processing (RAP)
Rapid Electrokinetic Patterning and Its Applications
Rate Sensors
Reactive Current Clamp
Reactive Empirical Bond-Order Potentials
Reactive Ion Beam Etching (RIBE)
Realtime Green's Function Approach
REBO
Reduced Graphene Oxide
Refractometric Sensing Using Plasmonic Nanoparticles
Relaxation
Relaxation Calorimetry

Reliability of Nanostructures
Remotely Powered Propulsion of Helical Nanobelts
Replica Molding
Resistive Switching
Resonant Evanescent Wave Biosensor
Retina Implant
Retina Silicon Chip
Retinal Prosthesis
Reversible Adhesion
RF Filters
RF MEMS Reconfigurable Filter
RF MEMS Switches
RF MEMS/NEMS
RF-MEMS/NEMS
Rheology
Riblets
Rigorous Maxwell Solver
RNA Interference (RNAi)
RNAi in Biomedicine and Drug Delivery
Robot-Based Automation on the Nanoscale
Robotic Insects
Robotic Platforms for Nanomanipulations
Role of Cavitation Bubble Dynamics on Nanoparticle Generation
Role of Plasma Confinement on Nanoparticle Generation
Rolled-Up Nanostructure
Rolled-Up Nanotechnology
Rose Petal Effect
Rotary Nanoelectromechanical Systems (NEMS)
Rotary Nanomachines
Rotary Nanomotors
Rotation Sensors

S

Sand Skink
SANS
SAXS
Scanning Electron Microscopy
Scanning Electron Microscopy and Secondary-Electron Imaging Microscopy
Scanning Force Microscopy in Liquids
Scanning Kelvin Probe Force Microscopy
Scanning Near-Field Optical Microscopy
Scanning Probe Microscopy
Scanning Surface Potential Microscopy
Scanning Thermal Microscopy
Scanning Thermal Profiler
Scanning Tunneling Microscopy
Scanning Tunneling Spectroscopy
Scanning X-Ray Diffraction Microscopy (SXDM)
Scanning-Probe Lithography
Scincus officinalis
Scincus scincus
Scrolled Nanostructure
Selected Synchrotron Radiation Techniques
Self-Assembled Monolayers
Self-Assembled Monolayers for Nanotribology
Self-Assembled Nanostructures
Self-Assembled Protein Layers
Self-Assembly
Self-Assembly for Heterogeneous Integration of Microsystems
Self-Assembly of Nanostructures
Self-Cleaning
Self-Energy and Excitonic Calculations in Many-Body Systems
Self-Healing Materials
Self-Organized Layers
Self-Organized Nanostructures
Self-Regeneration
Self-Repairing Materials
Self-Repairing Photoelectrochemical Complexes Based on Nanoscale Synthetic and Biological Components
SEM
Semiconductor Nanocrystals
Semiconductor Piezoresistance
Semiempirical Methods
Sense Organ
Sensors
Shark Denticles
Shark Skin Drag Reduction
Shark Skin Effect
Shark Skin Separation Control
Short (Low Aspect Ratio) Gold Nanowires
Short-Interfering RNA (siRNA)
Shrinkable and Stretchable Nanomanufacturing
Si Nanotubes
Silent Flight of Owls
Silent Owl Wings
Silica Gel Processing
Silicon Microphone
Silicon Thin Films

Silks of Spiders as Model Bio-polymers
Silver (Ag)
Simulating Nanoscale Heat Transport
Simulation of Supported Metal Clusters
Simulations of Bulk Nanostructured Solids
Simulations of Oxide/Polymer Hybrids
Simulations of Solid Interfaces
Single Cell Analysis
Single Cell Impedance Spectroscopy
Single-Cell Electrical Impedance Spectroscopy
Single-Cell Impedance Spectroscopy
Single-Walled Carbon Nanotubes (SWCNTs)
Sintered
siRNA Delivery
Size-Dependent Plasticity of Single Crystalline Metallic Nanostructures
Small Angle X-Ray Scattering in Grazing Incidence Geometry
Small Unilamellar Vesicle (SUV)
Small-Angle Neutron Scattering
Small-Angle Scattering
Small-Angle Scattering of Nanostructures and Nanomaterials
Small-Angle X-Ray Scattering
Smart Carbon Nanotube-Polymer Composites
Smart Drug Delivery Microchips
Smart Hydrogels
Soft Actuators
Soft Lithography
Soft Matter
Soft X-Ray Lithography
Soft X-Ray Microscopy
Soil/Terrestrial Ecosystem/Terrestrial Compartment
Solar Cells
Sol–Gel Method
Solid Lipid Nanocarriers
Solid Lipid Nanoparticles (SLN)
Solid Lipid Nanospheres
Solid Lipid-Based Nanoparticles
Solid–Liquid Interfaces
Solid-State Heat Convertors
Sound Propagation in Fluids
Spectromicroscopy
Spectroscopic Techniques
Spectroscopy of Ancient Documents
Spider Silk
Spiders
Spintronic Devices
Spontaneous Polarization
Spray Technologies Inspired by Bombardier Beetle
SPS of Carbon Allotropes
Sputtering
Stable Clusters
Stable Pre-critical Clusters
Stereolithography
Stimuli-Responsive Drug Delivery Microchips
Stimulus-Responsive Polymeric Hydrogels
Stochastic Assembly
Strain Gradient Plasticity Theory
Structural Color in Animals
Structural Colors
Structural DNA Nanotechnology
Structural Fluctuations
Structure and Stability of Protein Materials
Structure of Nanoparticles
SU-8 Photoresist
Sub-retinal Implant
Subwavelength
Sub-wavelength Waveguiding
Superelasticity and the Shape Memory Effect
Superhydrophobicity
Superoleophobicity of Fish Scales
Superparamagnetism
Support Loss
Surface Dissipations in NEMS/MEMS
Surface Electronic Structure
Surface Energy and Chemical Potential at Nanoscale
Surface Energy Density
Surface Energy Density and Chemical Potential at Nanoscale
Surface Engineering, Tailored Wettability, and Applications
Surface Figuring
Surface Force Balance
Surface Forces Apparatus
Surface Free Energy and Chemical Potential at Nanoscale
Surface Hopping
Surface Loss in Micromechanical/Nanomechanical Resonators
Surface Loss in NEMS/MEMS
Surface Micro-patterning
Surface Modeling of Ceramic Biomaterials

Surface Patterning
Surface Plasmon Enhanced Optical Bistability and Optical Switching
Surface Plasmon Nanophotonics
Surface Plasmon Polariton-Enabled High-Performance Organic Optoelectronic Devices
Surface Plasmon Resonance
Surface Plasmon-Coupled Emission
Surface Plasmon-Polariton Photodetectors
Surface Plasmon-Polariton-Based Detectors
Surface Properties
Surface Science
Surface Tension and Chemical Potential at Nanoscale
Surface Tension Effects of Nanostructures
Surface Tension–Driven Flow
Surface Tension–Powered Self-Assembly
Surface-Enhanced Raman Scattering for Imaging Biological Cells
Surface-Modified Microfluidics and Nanofluidics
Surface-Plasmon-Enhanced Solar Energy Conversion
Suspension Interaction with External AC Field Gradient
Synchronization
Synthesis of Carbon Nanotubes
Synthesis of Functional Materials for Bone Regeneration
Synthesis of Gold Nanoparticles
Synthesis of Graphene
Synthesis of Multicationic 1-D Oxide Nanostructures
Synthesis of Nanoparticles
Synthesis of Subnanometric Metal Nanoparticles
Synthesized Conductance Injection
Synthesized Ionic Conductance
Synthesized Synaptic Conductance
Synthetic Biology
Synthetic Genomics
Synthetic Lubricants
Systems Level Data Mining for RNAi

T

TEM
Terahertz
Terahertz and Infrared Plasmonics with Unconventional Materials
Terahertz Technology for Nano Applications
Teratogenic
Theoretical Elasticity
Theory of Artificial Electromagnetic Materials
Theory of Nonadiabatic Electron Dynamics in Nanomaterials
Theory of Optical Metamaterials
Thermal Actuators
Thermal Cancer Ablation Therapies Using Nanoparticles
Thermal Chemical Vapor Deposition
Thermal Conductance
Thermal Conductivity
Thermal Conductivity and Phonon Transport
Thermal Evaporation
Thermal Infrared Detector
Thermal Management Materials
Thermal Resistance
Thermal Resistivity
Thermal Transport
Thermally Actuated MEMS Resonators
Thermally Actuated Micromechanical Resonators
Thermally Actuated Nanoelectromechanical Resonators
Thermally Actuated Nanomechanical Resonators
Thermally Actuated Resonators
Thermally Actuated Silicon Resonators
Thermal-Piezoresistive Resonators
Thermodynamics of Small Systems
Thermoelectric Nanomaterials and Nanostructured Thermoelectrics
Thermomechanical Actuators
Thermometry
Thermostability
Thickness
Thiols
Third-Generation Vectors
3D Micro-/Nanomanipulation with Force Spectroscopy
Three-Dimensional Imaging of Human Tissues
Three-Phase Contact Line
THz
Tight-Binding Simulations of Nanowires
TiO_2 Nanotube Arrays: Growth and Application
Titanium Dioxide
Total Internal Reflection (Fluorescence) Velocimetry

Toward Bioreplicated Texturing of Solar-Cell Anodes
Toward Bioreplicated Texturing of Solar-Cell Surfaces
Toxic Mechanisms
Toxicity
Toxicity of Metal and Metal Oxide Nanoparticles on Prokaryotes (Bacteria), Unicellular and Invertebrate Eukaryotes
Toxicity/Ecotoxicity
Toxicology: Plants and Nanoparticles
Transdisciplinary Approaches
Transduction
Transfer
Translational Movement of Polynucleotides in a Nonuniform Electric Field
Transmission Electron Microscopy
Transport Properties
T-Rays
Tunable Filter
Tunable RF MEMS Filters: A Review
Tuning Electrical Properties of Carbon Nanotubes via Spark Plasma Sintering
Turfs
2D-Slab Approach
248-nm Lithography
Two-Photon Lithography

U

uFOM
Ultrahigh Vacuum Chemical Vapor Deposition (UHVCVD)
Ultralarge Strain Elasticity
Ultra-precision Finishing
Ultra-precision Machining
Ultraprecision Machining (UPM)
Ultraprecision Surfaces and Structures with Nanometer Accuracy by Ion Beam and Plasma Jet Technologies
Ultrashort Carbon Nanotubes
Ultrasonic Atomization
Ultrasonic Force-Assisted Nanomachining
Ultrasonic Force-Regulated Nanomachining
Ultrasonic Machining
Ultrasonic Vibration-Assisted Nanomachining
Unconventional Computing
Upconversion Enhancement in Lanthanide-Doped Nanoparticles Using Nanoplasmonics
Uptake/Internalization/Sequestration/Biodistribution
Use of Nanotechnology in Pregnancy
USM
US-Tubes
UV Lithography

V

Vertically Aligned Carbon Nanotubes (VACNTs)
Vertically Aligned Carbon Nanotubes, Collective Mechanical Behavior
Vertically-Aligned Carbon Nanotubes
Vibration Assisted Machining
Viscosity
Visual Prosthesis
Visual Servoing for SEM

W

Wavefront Deformation Particle Image Tracking
Waypoint Detection
Wear
Wet Adhesion in Tree and Torrent Frogs
Wet Chemical and Electrochemical Etching Processes
Wet Chemical Processing
Wet Chemical Synthesis
Wetting Transitions
Whispering Gallery Mode Biosensor
Whispering Gallery Mode Resonator Biosensors
Wine-Glass Mode Resonators
Wireless Propulsion of Helical Nanobelt Swimmers
Worm Chip

X

X-Ray Diffraction with Micron Sized X-Ray Beams

Y

Young's Modulus

Z

Zinc Oxide Nanoparticles (ZnO-NPs)